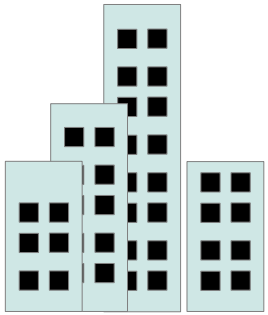
Paralelización de Tareas en Clústeres de Ordenadores: Scripts en SGE

Francisco J. Romero Campero
http://www.cs.us.es/~fran/

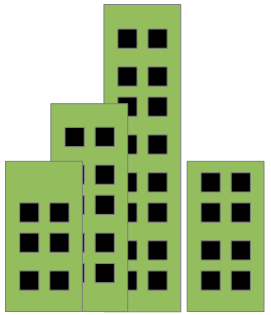
Dpt. de Ciencias de la Computación e Inteligencia Artificial
Universidad de Sevilla

Paralelización

- Una tarea es **paralelizable** cuando puede dividirse en **subtareas** que pueden realizarse de forma **independiente** y **simultánea** de tal forma que la solución final es la **combinación** de la solución de cada subtarea.



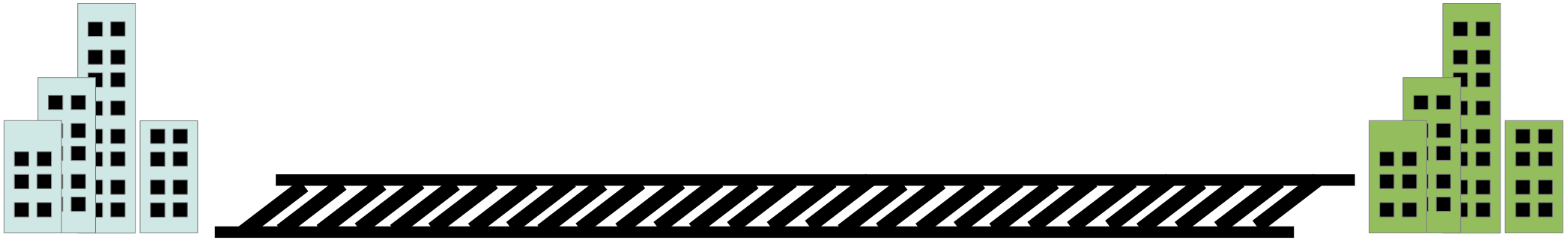
Ciudad 1



Ciudad 2

Paralelización

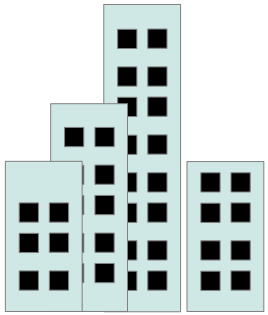
- Una tarea es **paralelizable** cuando puede dividirse en **subtareas** que pueden realizarse de forma **independiente** y **simultánea** de tal forma que la solución final es la **combinación** de la solución de cada subtarea.



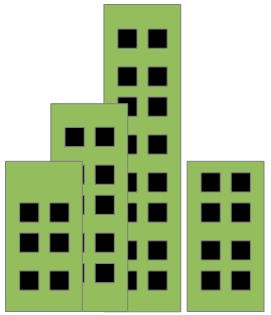
Ciudad 1 **Ciudad 2**

Paralelización

- Una tarea es **paralelizable** cuando puede dividirse en **subtareas** que pueden realizarse de forma **independiente** y **simultánea** de tal forma que la solución final es la **combinación** de la solución de cada subtarea.



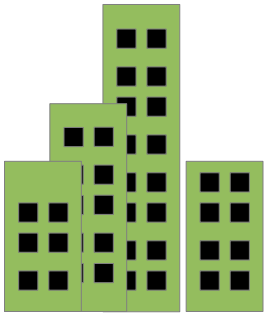
Ciudad 1



Ciudad 2

Paralelización

- Una tarea es **paralelizable** cuando puede dividirse en **subtareas** que pueden realizarse de forma **independiente** y **simultánea** de tal forma que la solución final es la **combinación** de la solución de cada subtarea.

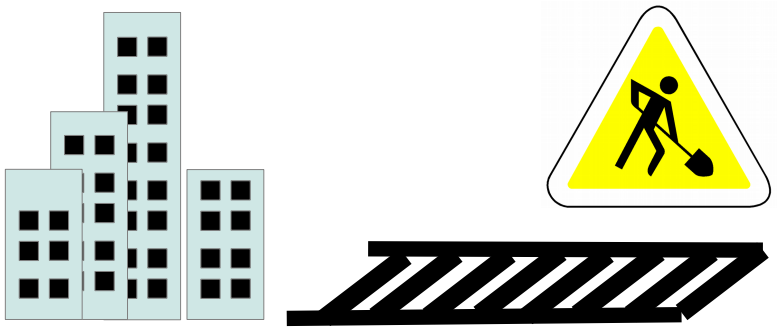


Ciudad 1

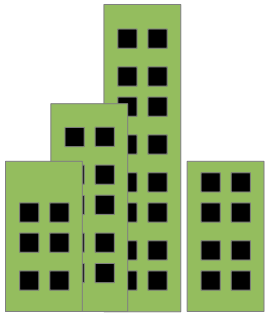
Ciudad 2

Paralelización

- Una tarea es **paralelizable** cuando puede dividirse en **subtareas** que pueden realizarse de forma **independiente** y **simultánea** de tal forma que la solución final es la **combinación** de la solución de cada subtarea.



Ciudad 1



Ciudad 2

Paralelización

- Una tarea es **paralelizable** cuando puede dividirse en **subtareas** que pueden realizarse de forma **independiente** y **simultánea** de tal forma que la solución final es la **combinación** de la solución de cada subtarea.

Ciudad 1 **Ciudad 2**

Paralelización

- Una tarea es **paralelizable** cuando puede dividirse en **subtareas** que pueden realizarse de forma **independiente** y **simultánea** de tal forma que la solución final es la **combinación** de la solución de cada subtarea.

Ciudad 1 **Ciudad 2**

Paralelización

- Una tarea es **paralelizable** cuando puede dividirse en **subtareas** que pueden realizarse de forma **independiente** y **simultánea** de tal forma que la solución final es la **combinación** de la solución de cada subtarea.

Ciudad 1 **Ciudad 2**

Paralelización

- Una tarea es **paralelizable** cuando puede dividirse en **subtareas** que pueden realizarse de forma **independiente** y **simultánea** de tal forma que la solución final es la **combinación** de la solución de cada subtarea.

Ciudad 1 **Ciudad 2**

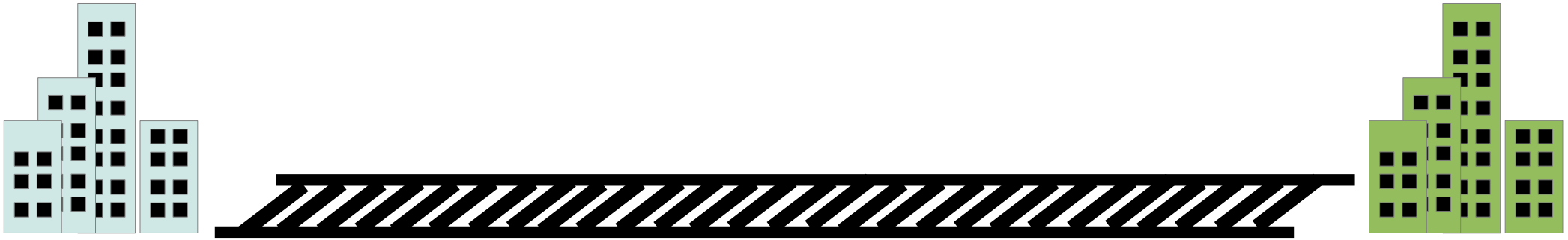
Paralelización

- Una tarea es **paralelizable** cuando puede dividirse en **subtareas** que pueden realizarse de forma **independiente** y **simultánea** de tal forma que la solución final es la **combinación** de la solución de cada subtarea.

Ciudad 1 **Ciudad 2**

Paralelización

- Una tarea es **paralelizable** cuando puede dividirse en **subtareas** que pueden realizarse de forma **independiente** y **simultánea** de tal forma que la solución final es la **combinación** de la solución de cada subtarea.

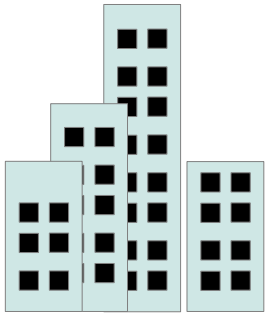


Ciudad 1 **Ciudad 2**

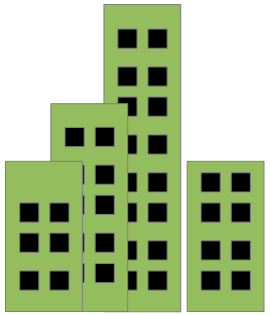
Siete unidades de tiempo en terminar la tarea

Paralelización

- Una tarea es **paralelizable** cuando puede dividirse en **subtareas** que pueden realizarse de forma **independiente** y **simultánea** de tal forma que la solución final es la **combinación** de la solución de cada subtarea.



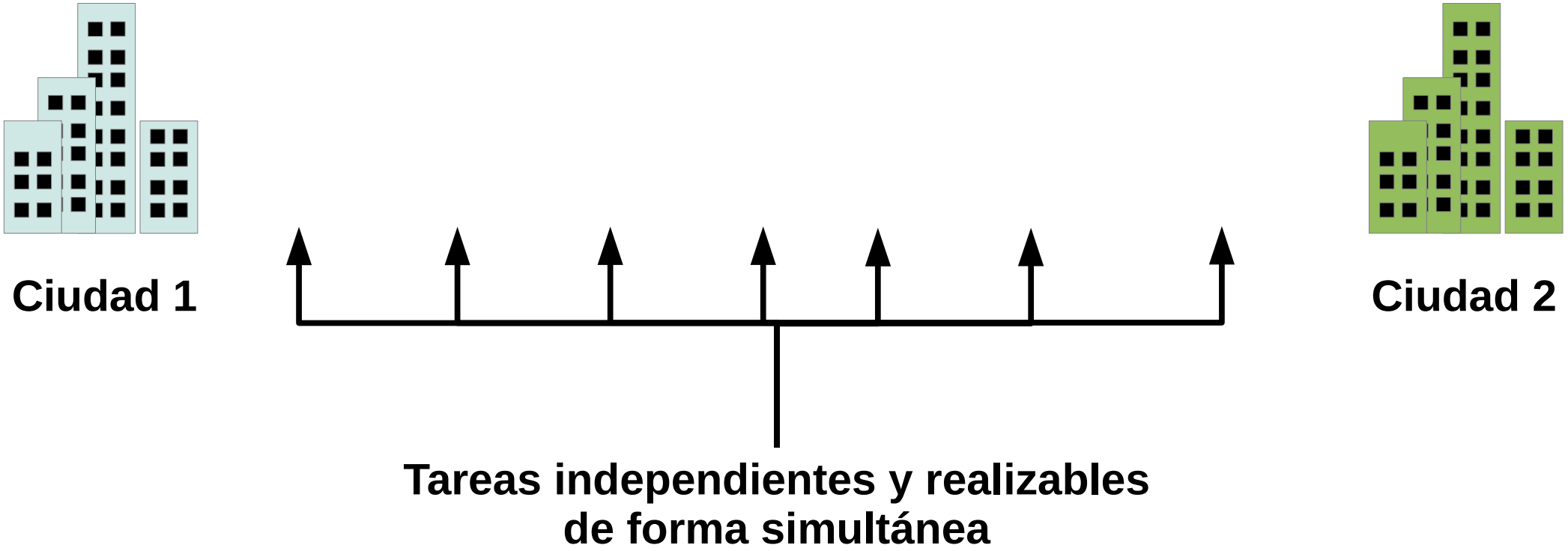
Ciudad 1



Ciudad 2

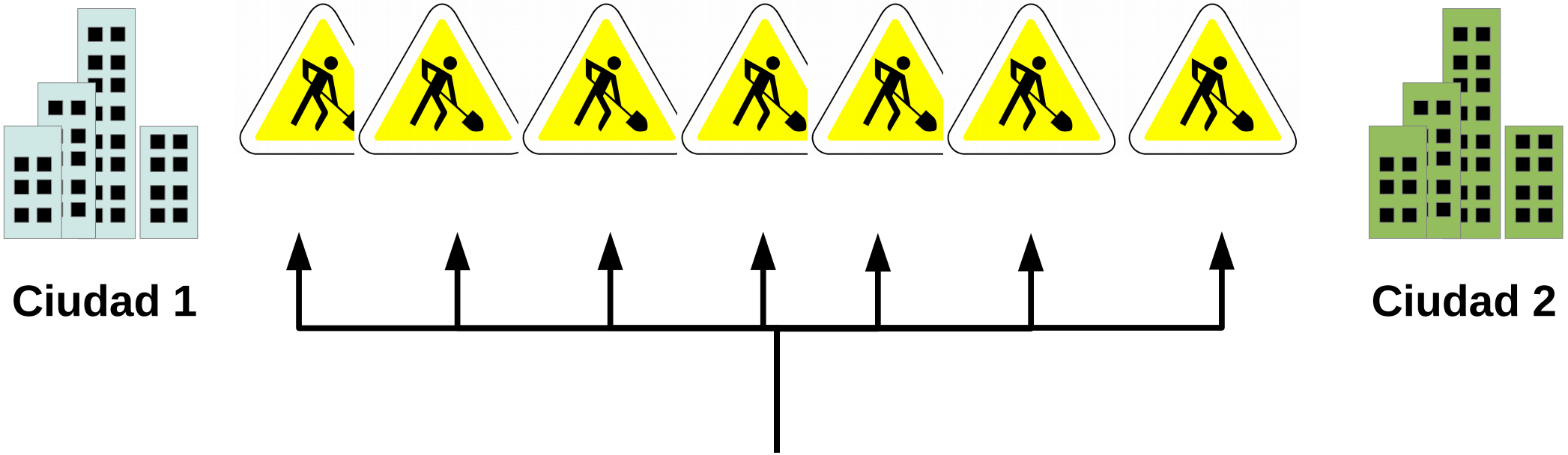
Paralelización

- Una tarea es **paralelizable** cuando puede dividirse en **subtareas** que pueden realizarse de forma **independiente** y **simultánea** de tal forma que la solución final es la **combinación** de la solución de cada subtarea.



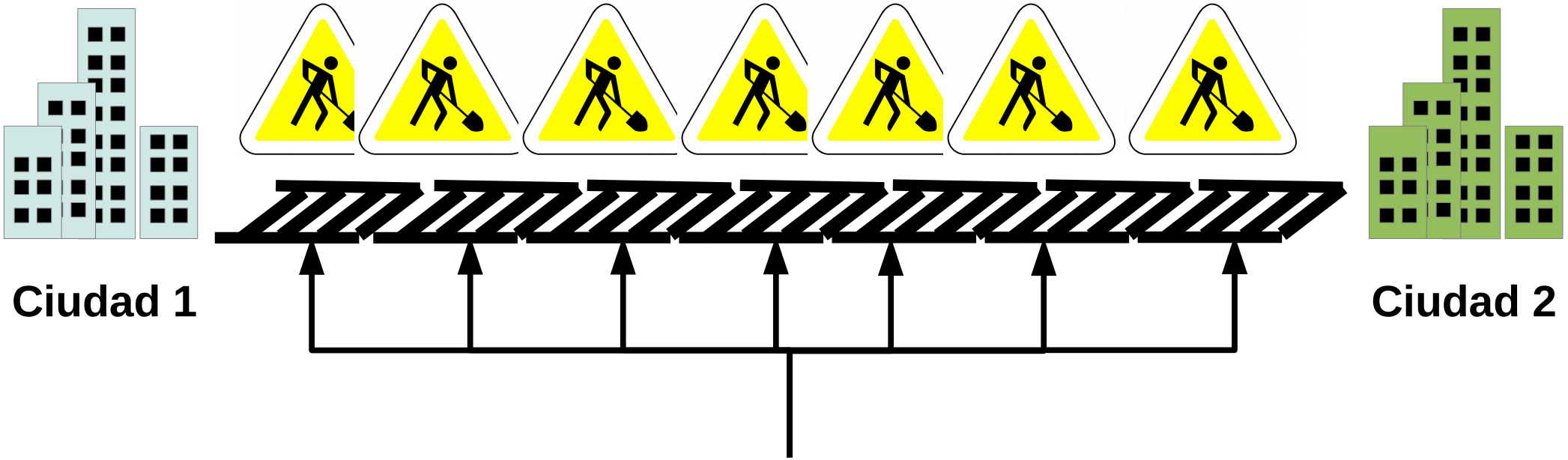
Paralelización

- Una tarea es **paralelizable** cuando puede dividirse en **subtareas** que pueden realizarse de forma **independiente** y **simultánea** de tal forma que la solución final es la **combinación** de la solución de cada subtarea.



Paralelización

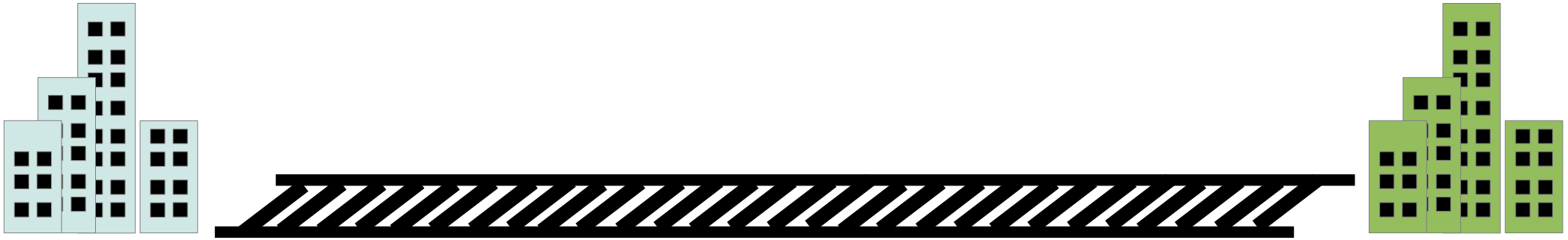
- Una tarea es **paralelizable** cuando puede dividirse en **subtareas** que pueden realizarse de forma **independiente** y **simultánea** de tal forma que la solución final es la **combinación** de la solución de cada subtarea.



Tareas independientes y realizables de forma simultánea

Paralelización

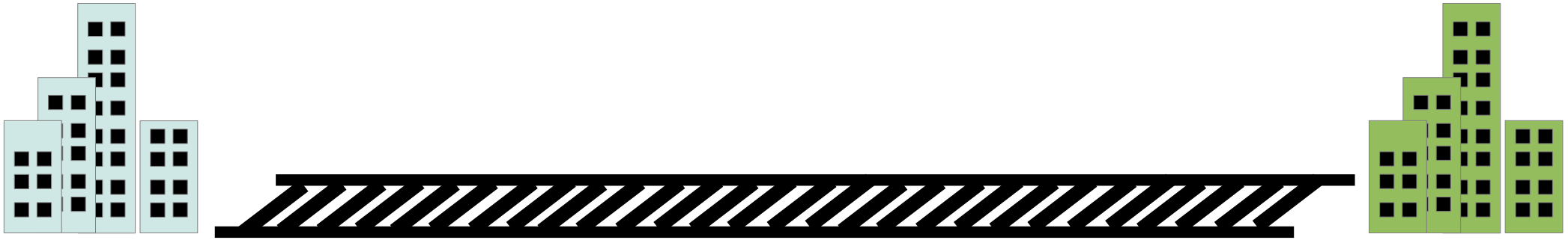
- Una tarea es **paralelizable** cuando puede dividirse en **subtareas** que pueden realizarse de forma **independiente** y **simultánea** de tal forma que la solución final es la **combinación** de la solución de cada subtarea.



Ciudad 1 **Ciudad 2**

Paralelización

- Una tarea es **paralelizable** cuando puede dividirse en **subtareas** que pueden realizarse de forma **independiente** y **simultánea** de tal forma que la solución final es la **combinación** de la solución de cada subtarea.

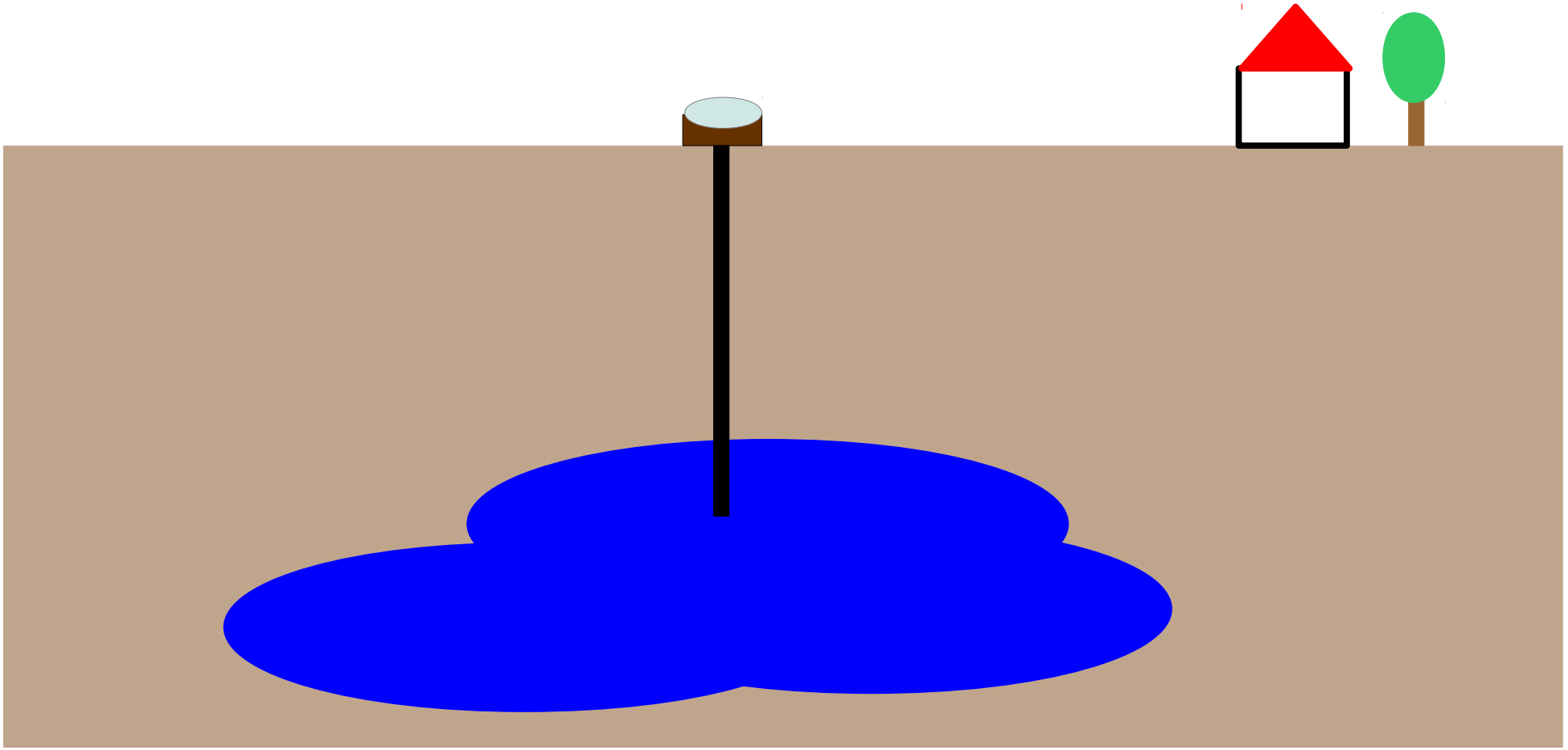


Ciudad 1 **Ciudad 2**

Una unidad de tiempo en terminar la tarea

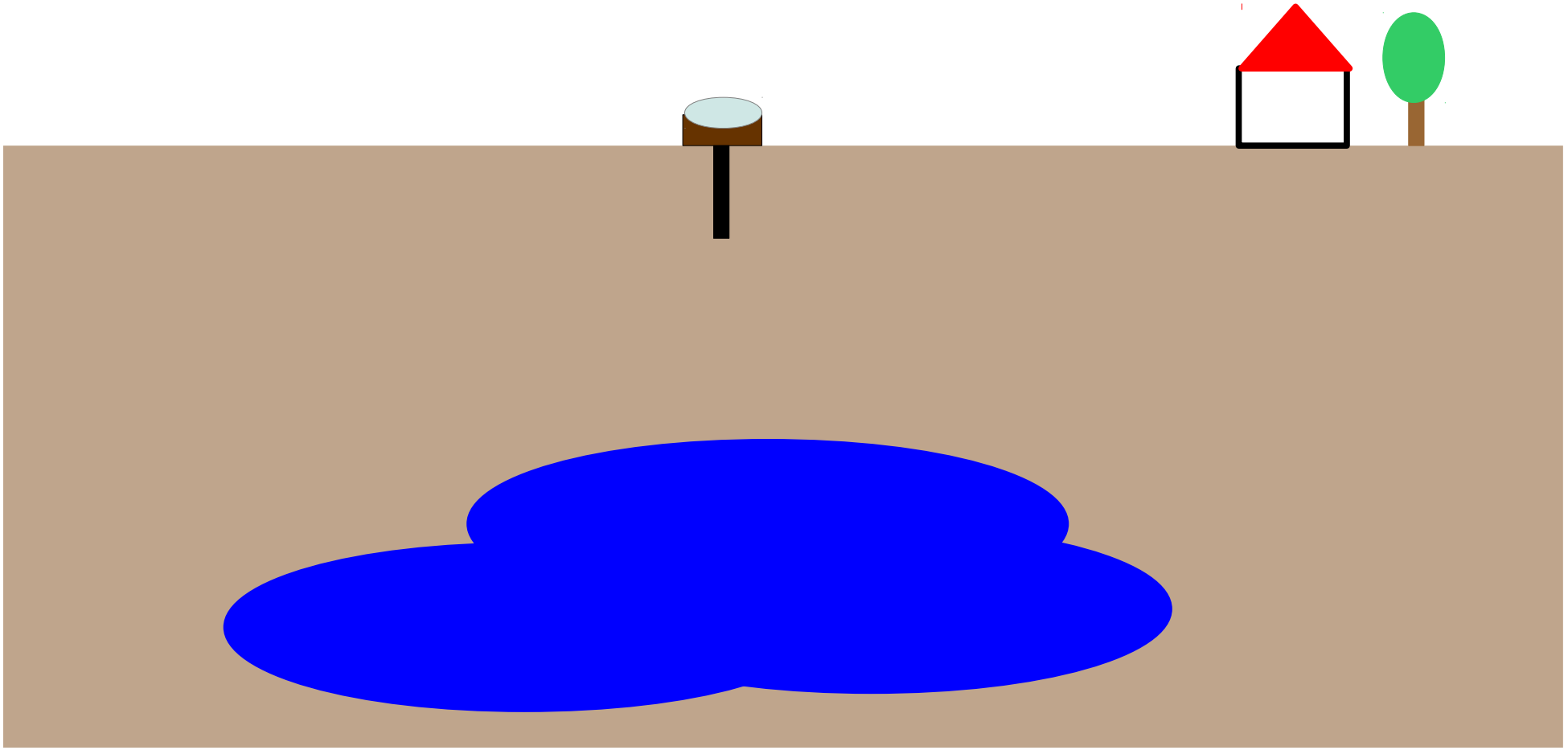
Paralelización

- Una tarea es **paralelizable** cuando puede dividirse en **subtareas** que pueden realizarse de forma **independiente** y **simultánea** de tal forma que la solución final es la **combinación** de la solución de cada subtarea.



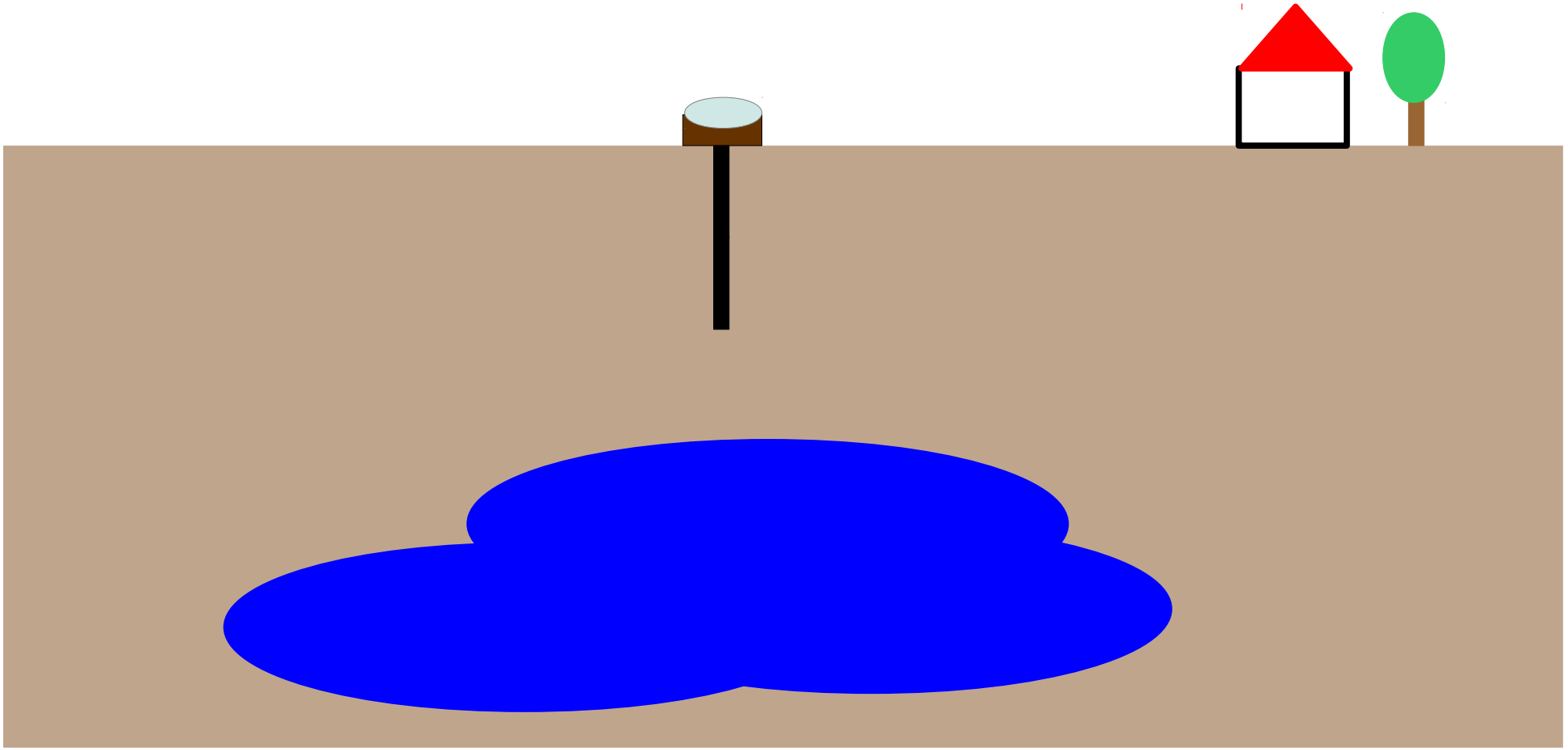
Paralelización

- Una tarea es **paralelizable** cuando puede dividirse en **subtareas** que pueden realizarse de forma **independiente** y **simultánea** de tal forma que la solución final es la **combinación** de la solución de cada subtarea.



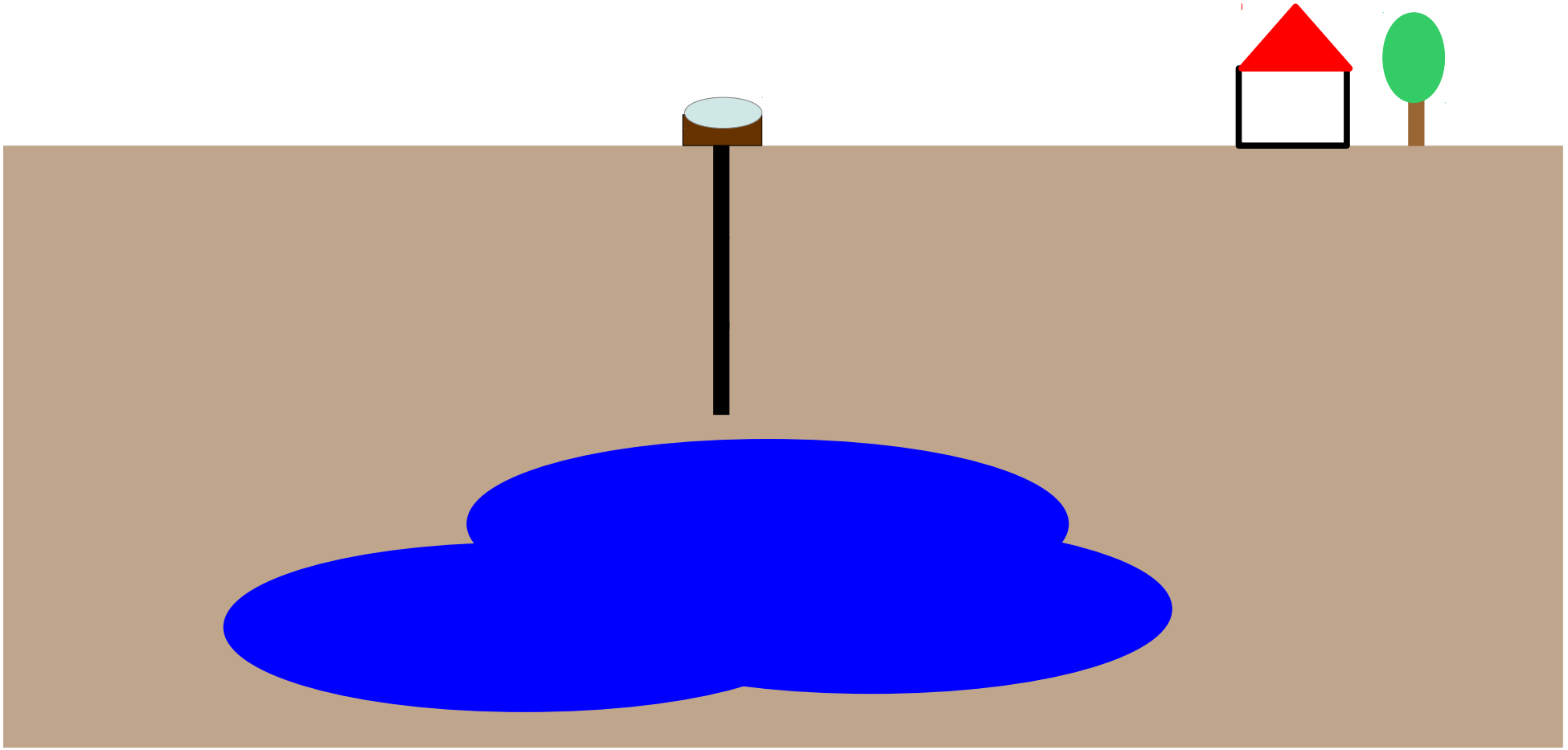
Paralelización

- Una tarea es **paralelizable** cuando puede dividirse en **subtareas** que pueden realizarse de forma **independiente** y **simultánea** de tal forma que la solución final es la **combinación** de la solución de cada subtarea.



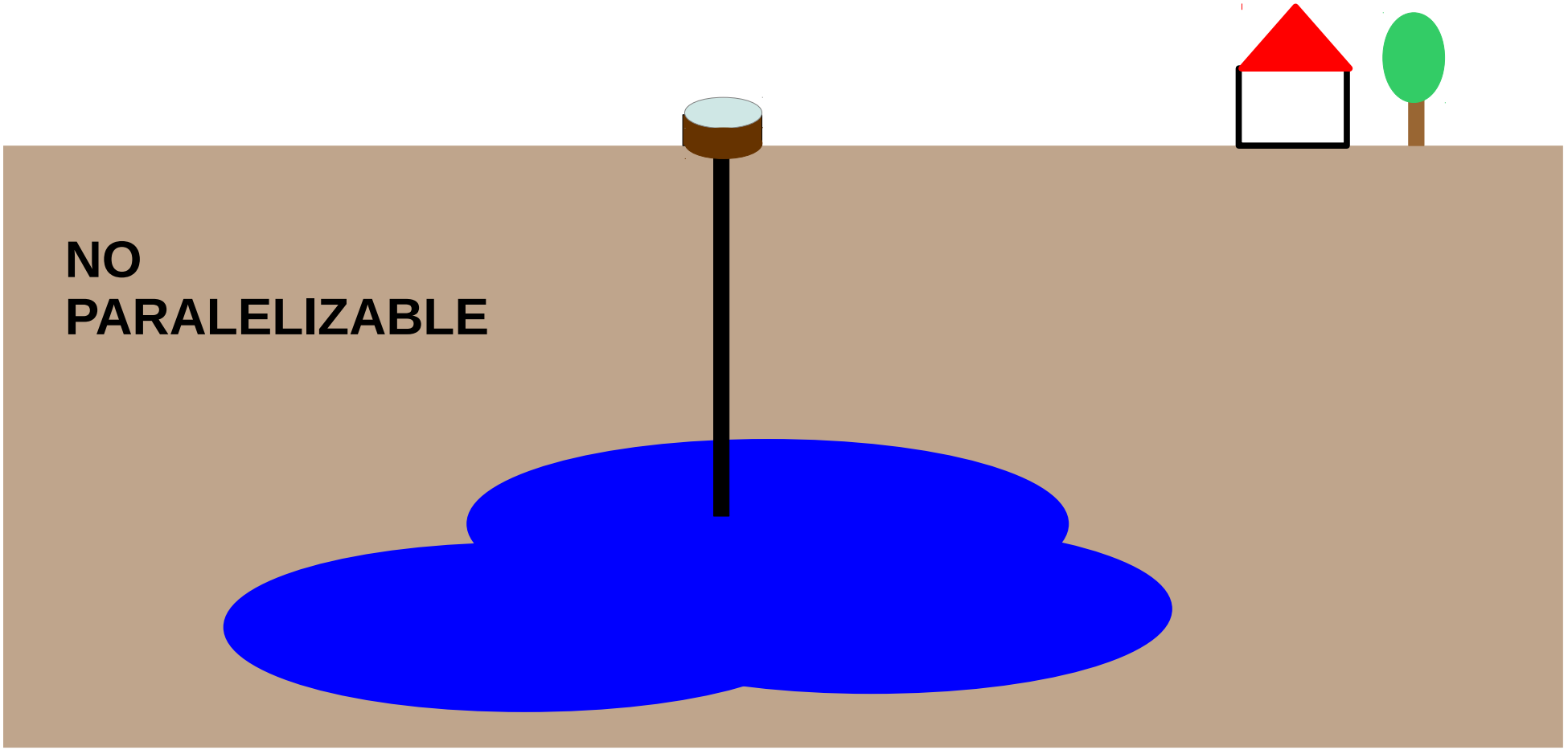
Paralelización

- Una tarea es **paralelizable** cuando puede dividirse en **subtareas** que pueden realizarse de forma **independiente** y **simultánea** de tal forma que la solución final es la **combinación** de la solución de cada subtarea.



Paralelización

- Una tarea es **paralelizable** cuando puede dividirse en **subtareas** que pueden realizarse de forma **independiente** y **simultánea** de tal forma que la solución final es la **combinación** de la solución de cada subtarea.



Pipeline para el Análisis de datos de RNA-seq

Pipeline para el Análisis de datos de RNA-seq

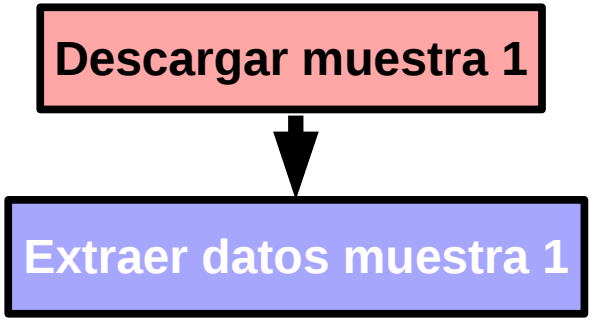
Descargar muestra 1

Pipeline para el Análisis de datos de RNA-seq

Descargar muestra 1

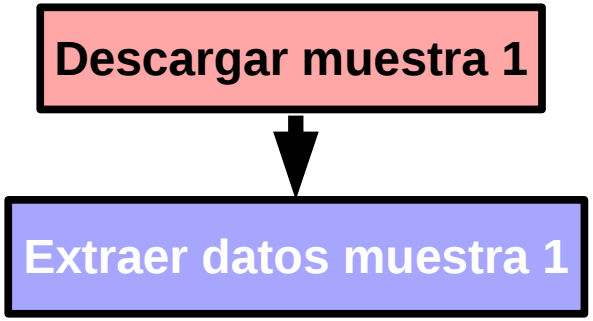
wget

Pipeline para el Análisis de datos de RNA-seq



wget

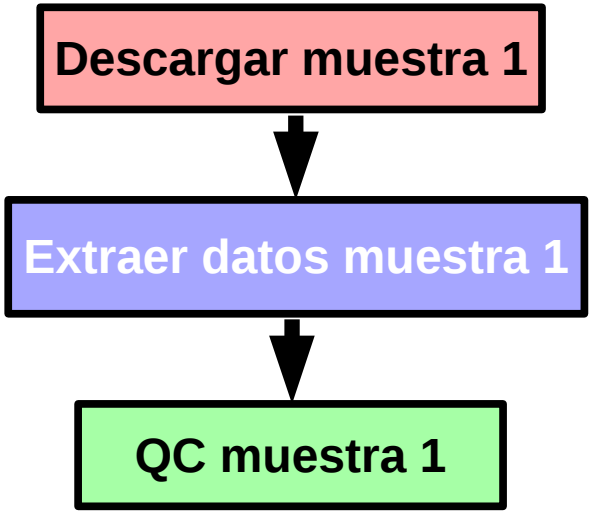
Pipeline para el Análisis de datos de RNA-seq



wget

fastq-dump
--split-files

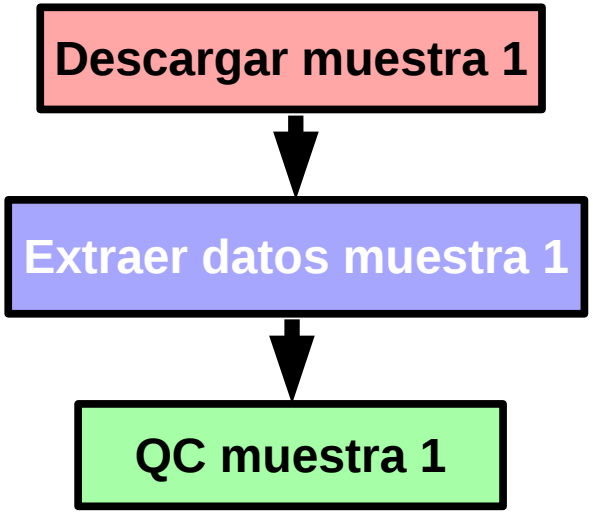
Pipeline para el Análisis de datos de RNA-seq



wget

fastq-dump
--split-files

Pipeline para el Análisis de datos de RNA-seq

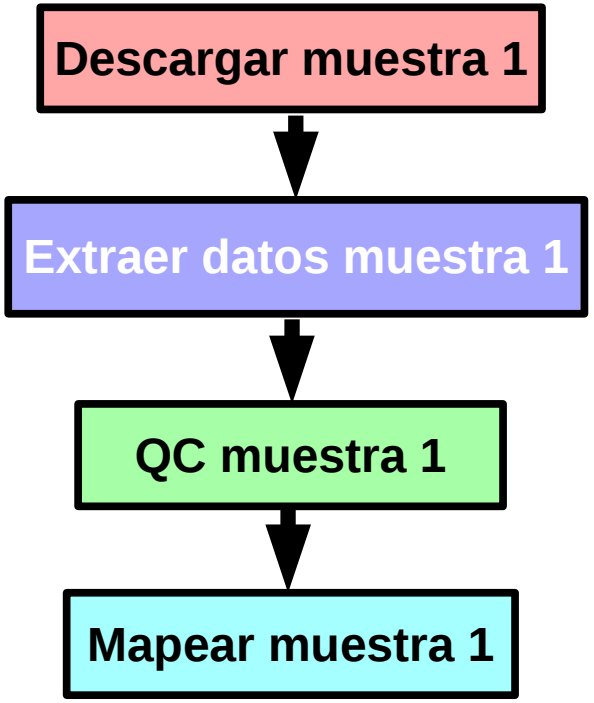


wget

fastq-dump
--split-files

fastqc

Pipeline para el Análisis de datos de RNA-seq

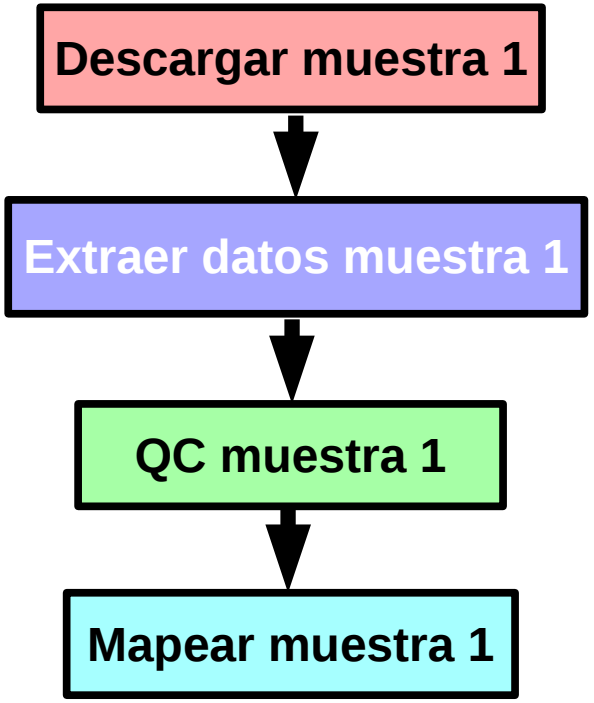


wget

fastq-dump
--split-files

fastqc

Pipeline para el Análisis de datos de RNA-seq



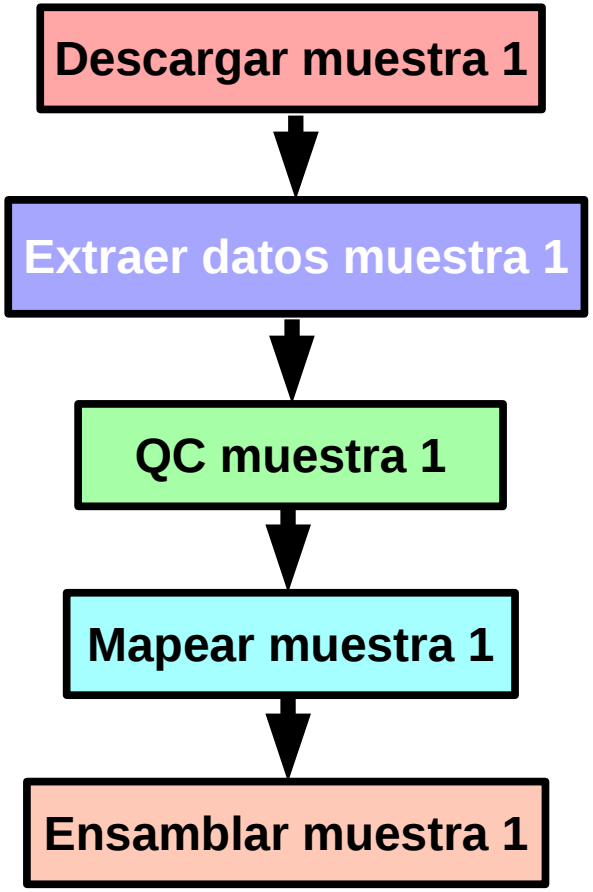
wget

fastq-dump --split-files

fastqc

hisat2

Pipeline para el Análisis de datos de RNA-seq



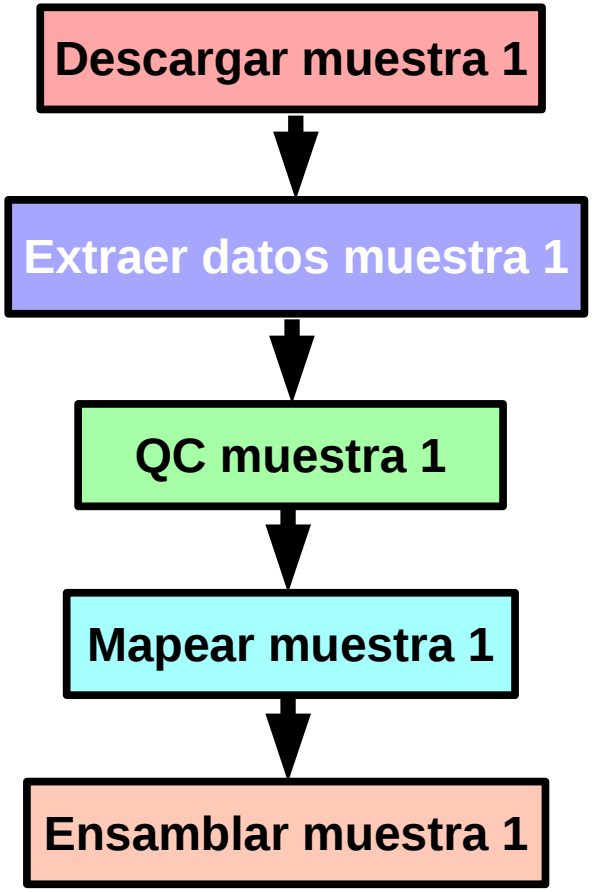
wget

fastq-dump --split-files

fastqc

hisat2

Pipeline para el Análisis de datos de RNA-seq



wget

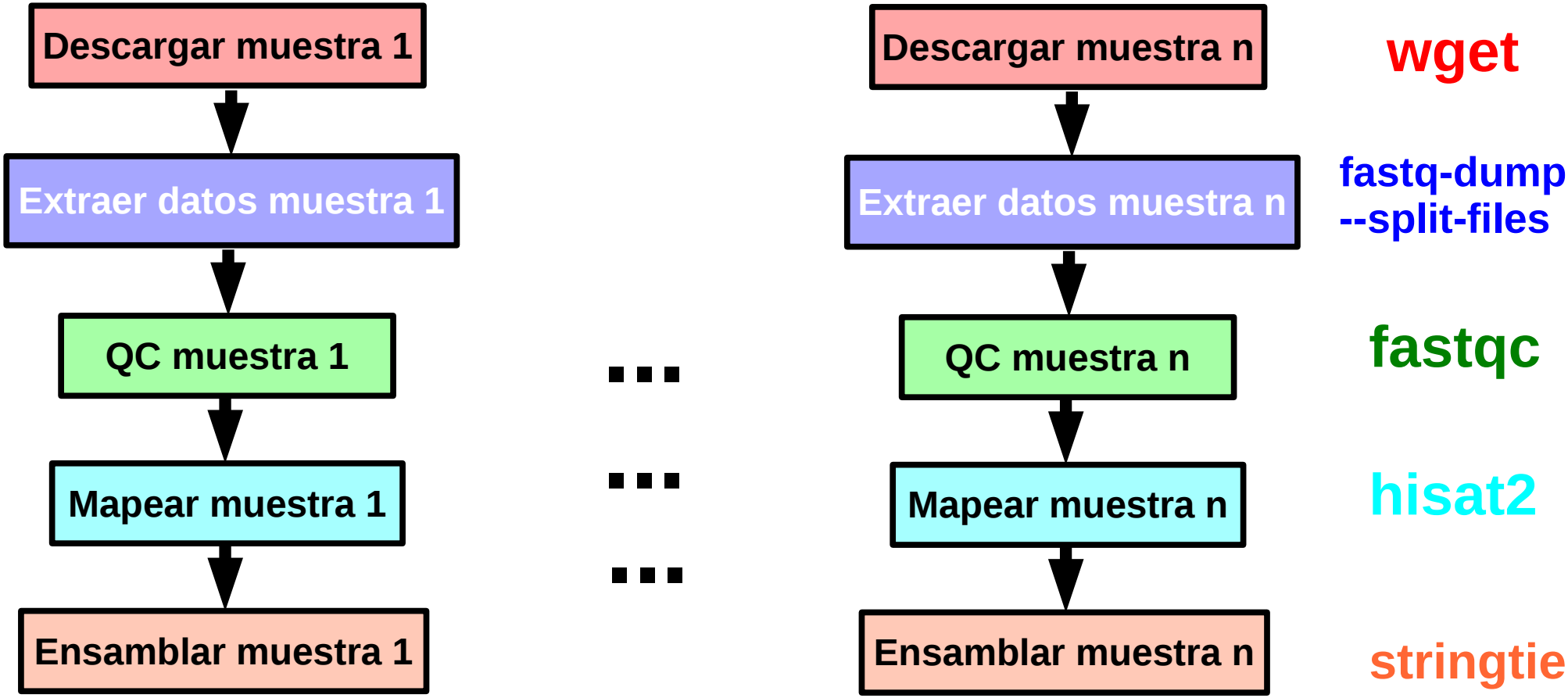
fastq-dump
--split-files

fastqc

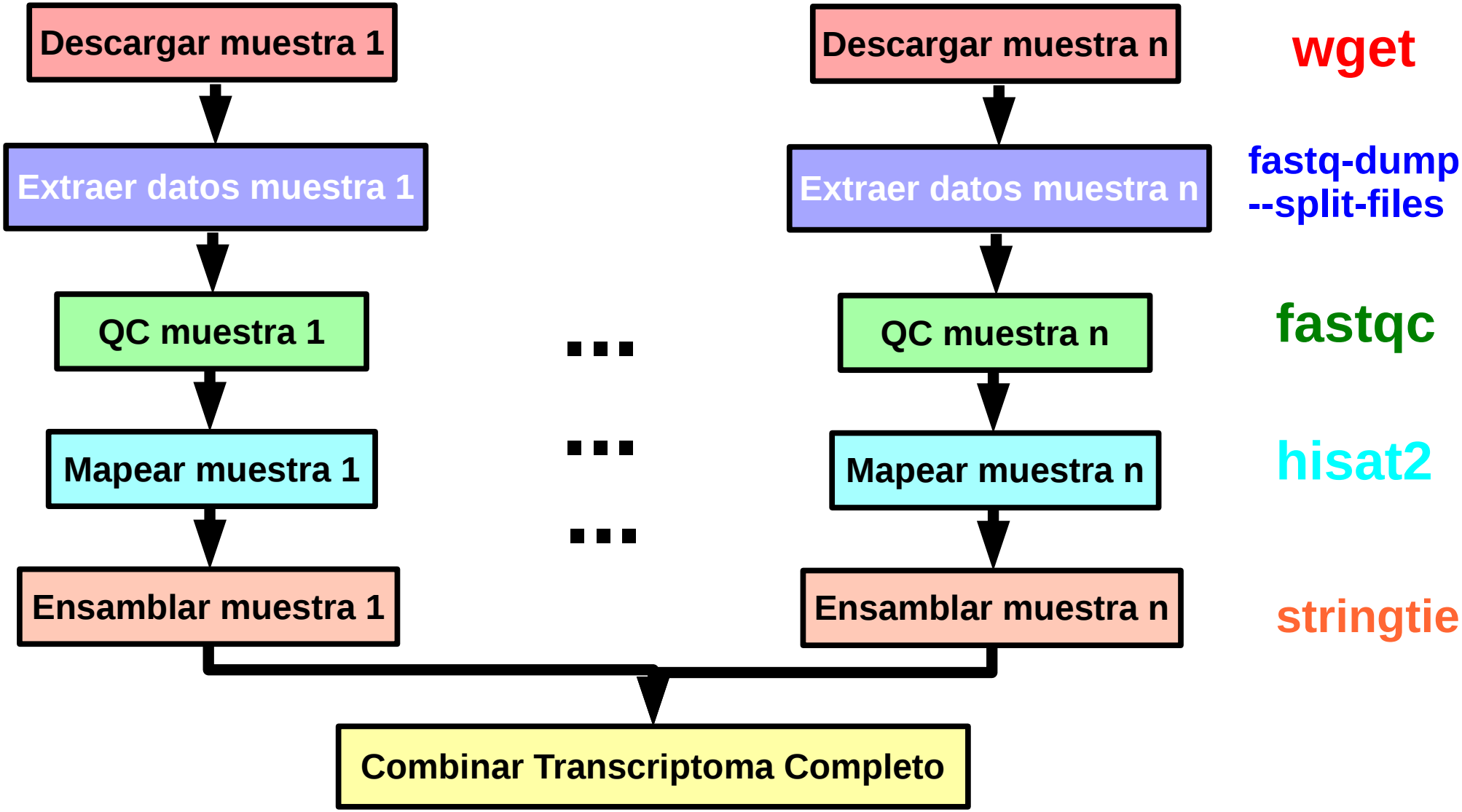
hisat2

stringtie

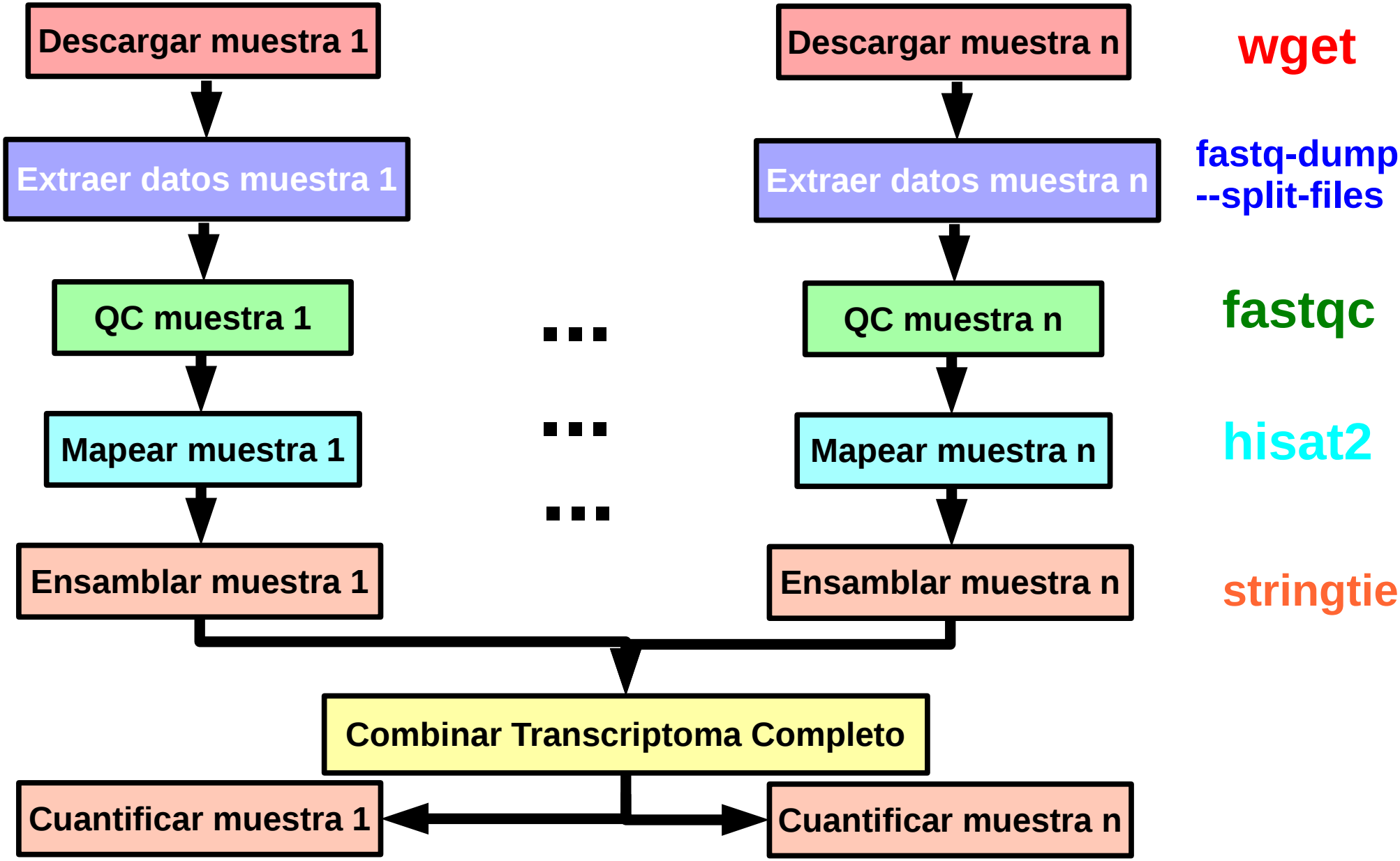
Pipeline para el Análisis de datos de RNA-seq



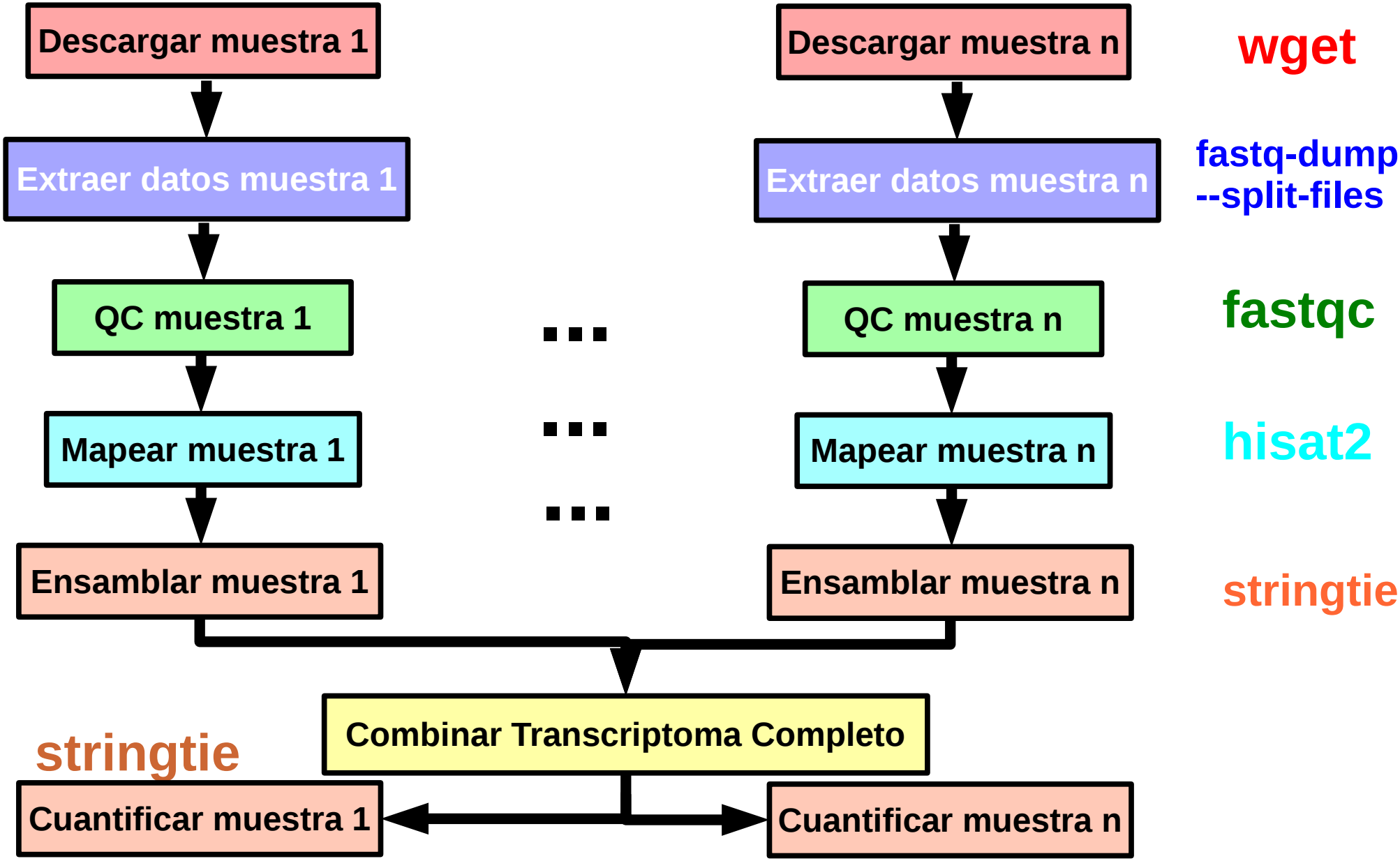
Pipeline para el Análisis de datos de RNA-seq



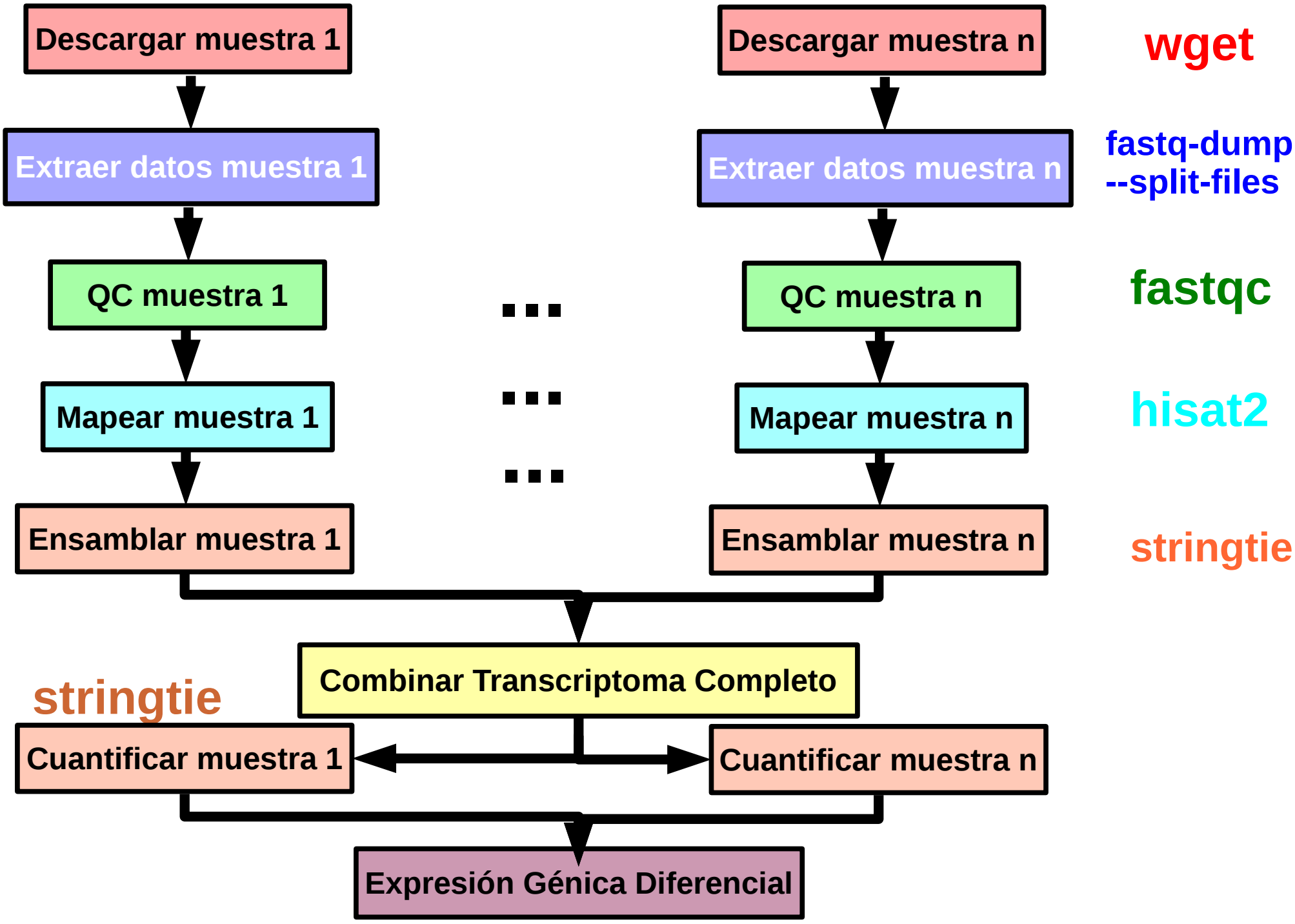
Pipeline para el Análisis de datos de RNA-seq



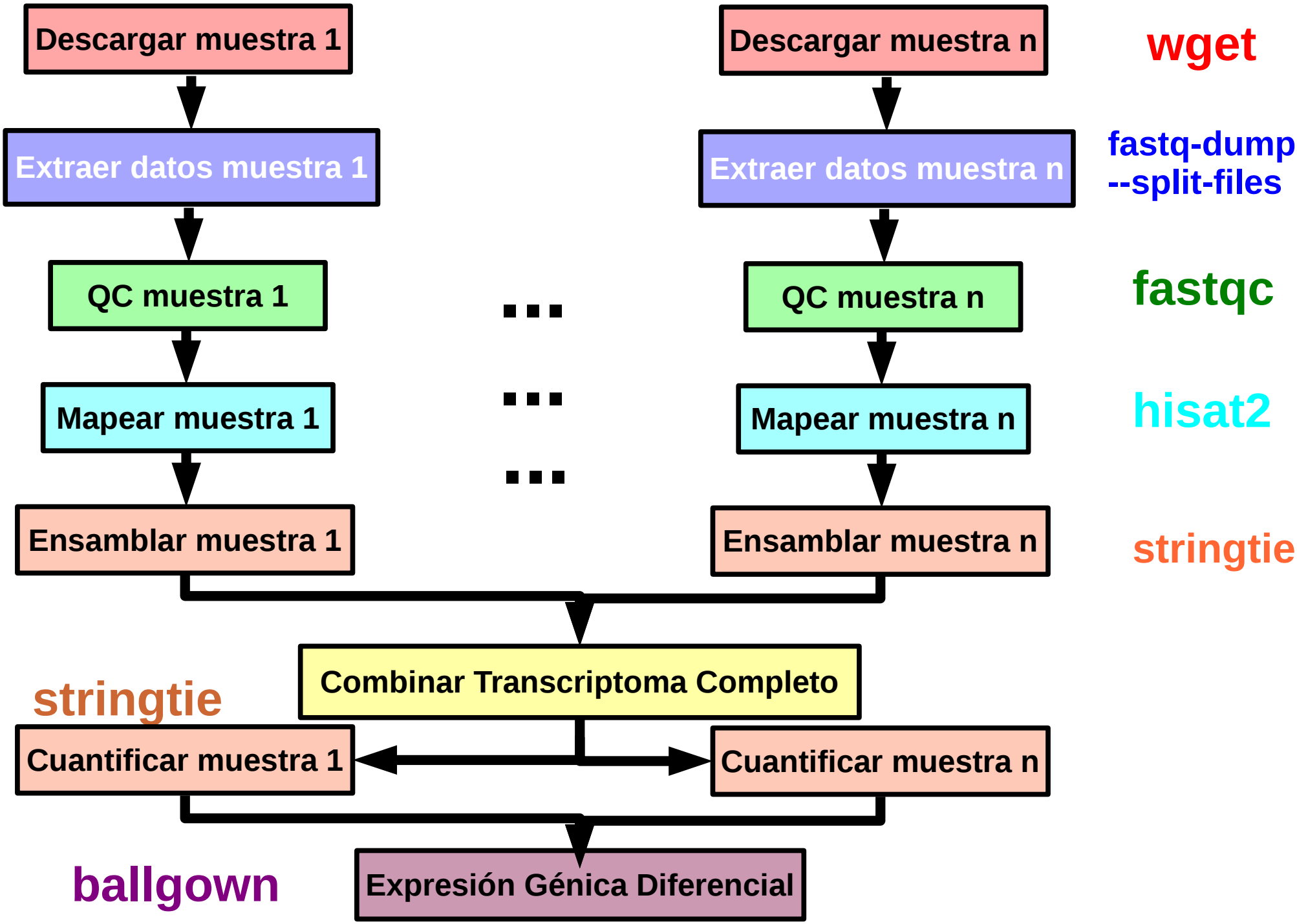
Pipeline para el Análisis de datos de RNA-seq



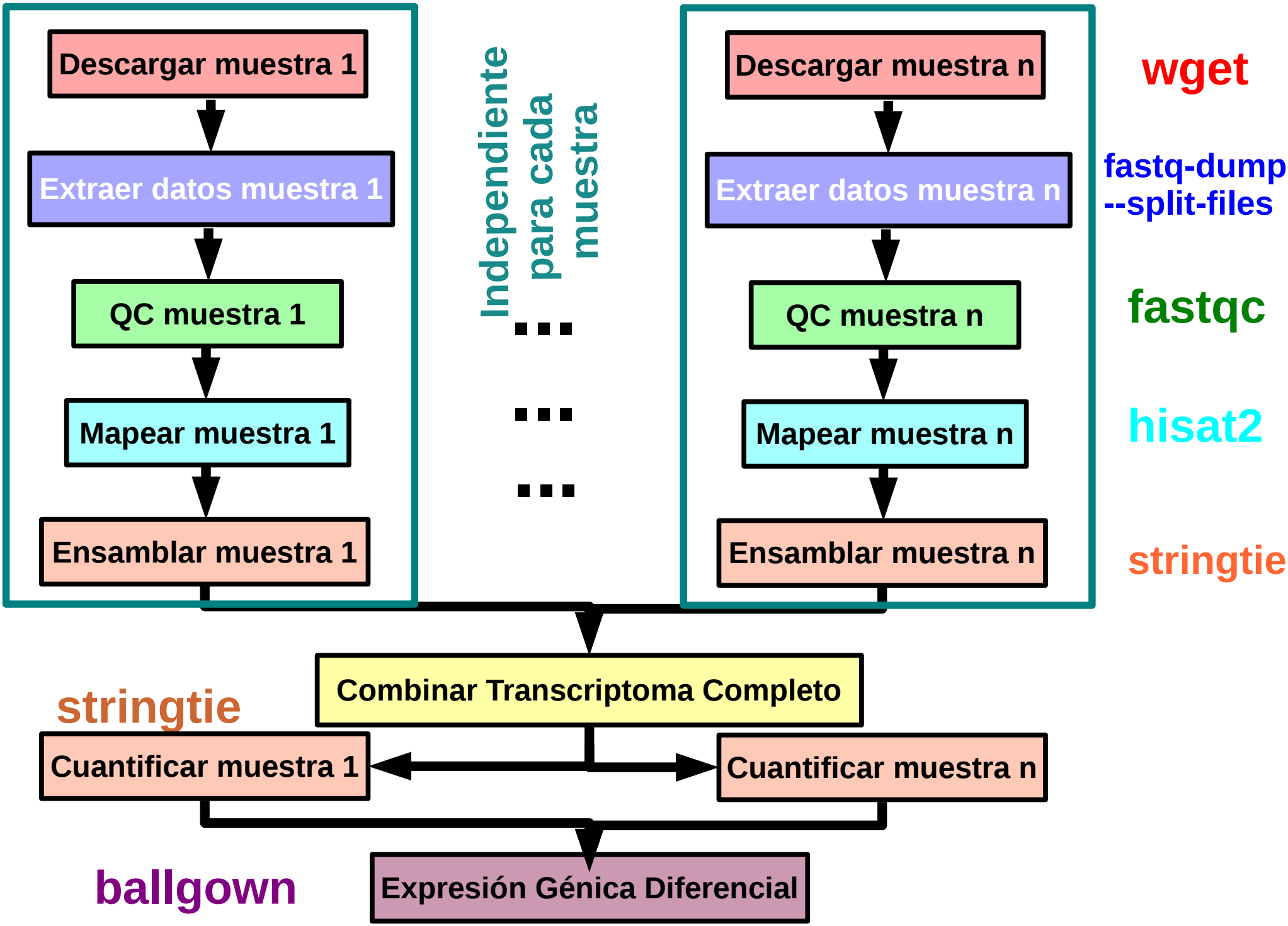
Pipeline para el Análisis de datos de RNA-seq



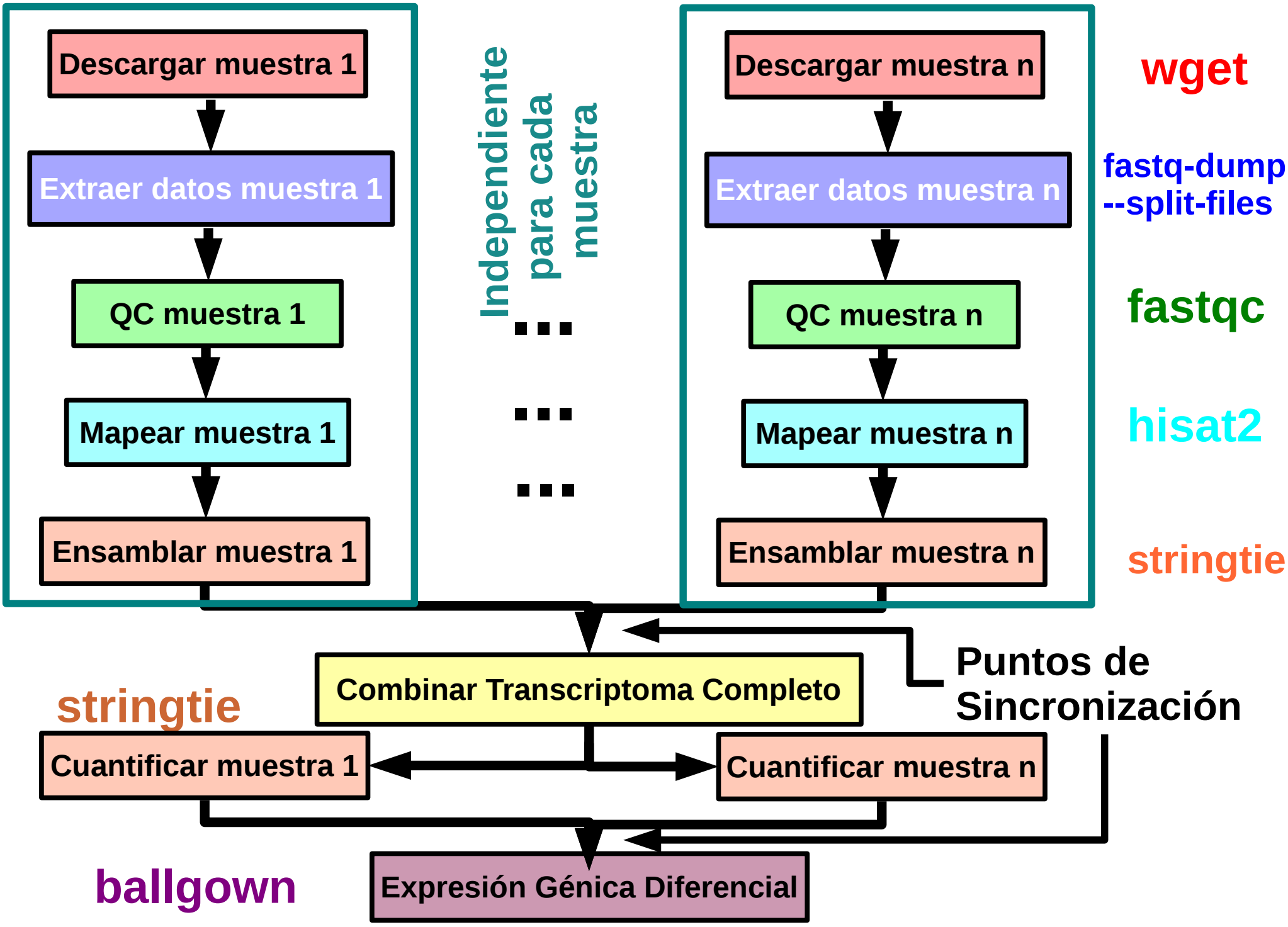
Pipeline para el Análisis de datos de RNA-seq



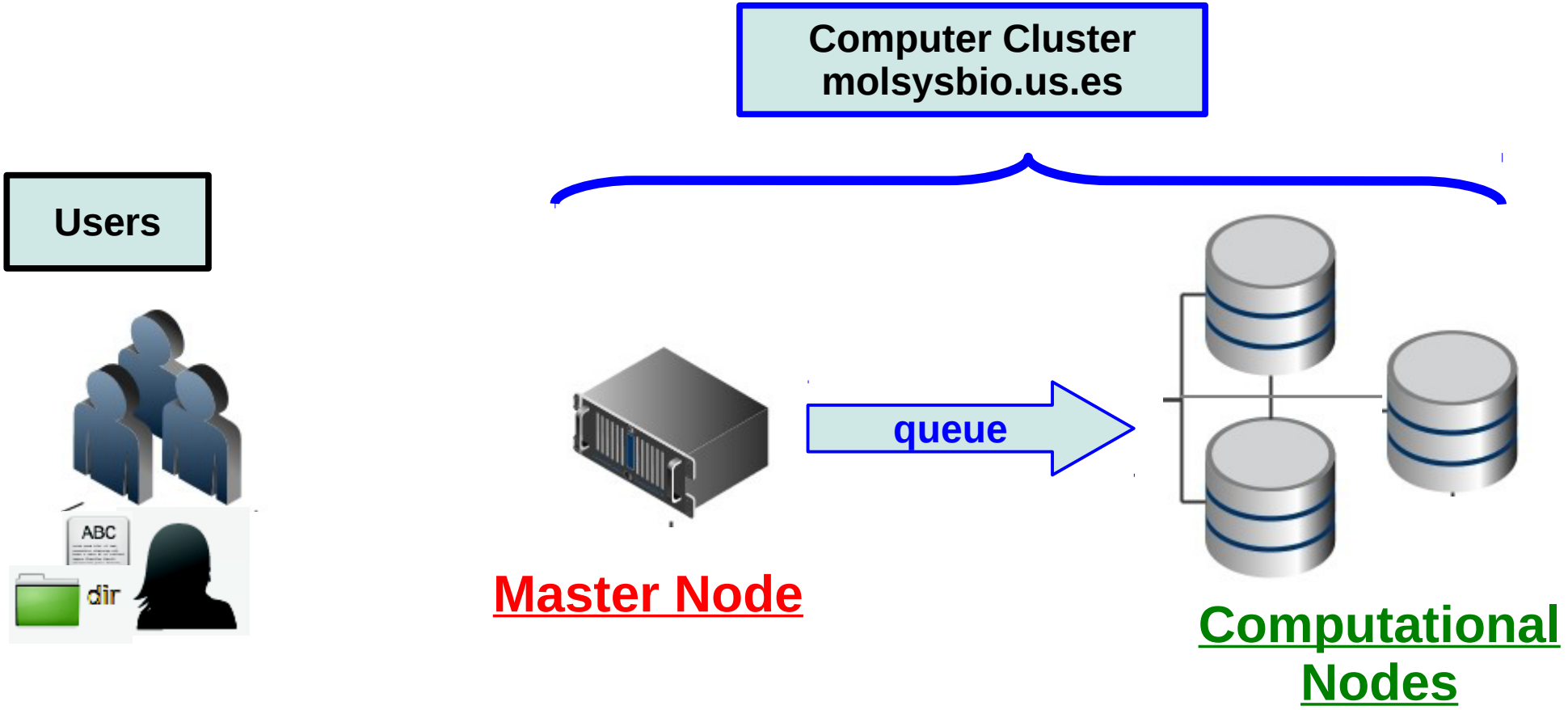
Pipeline para el Análisis de datos de RNA-seq



Pipeline para el Análisis de datos de RNA-seq



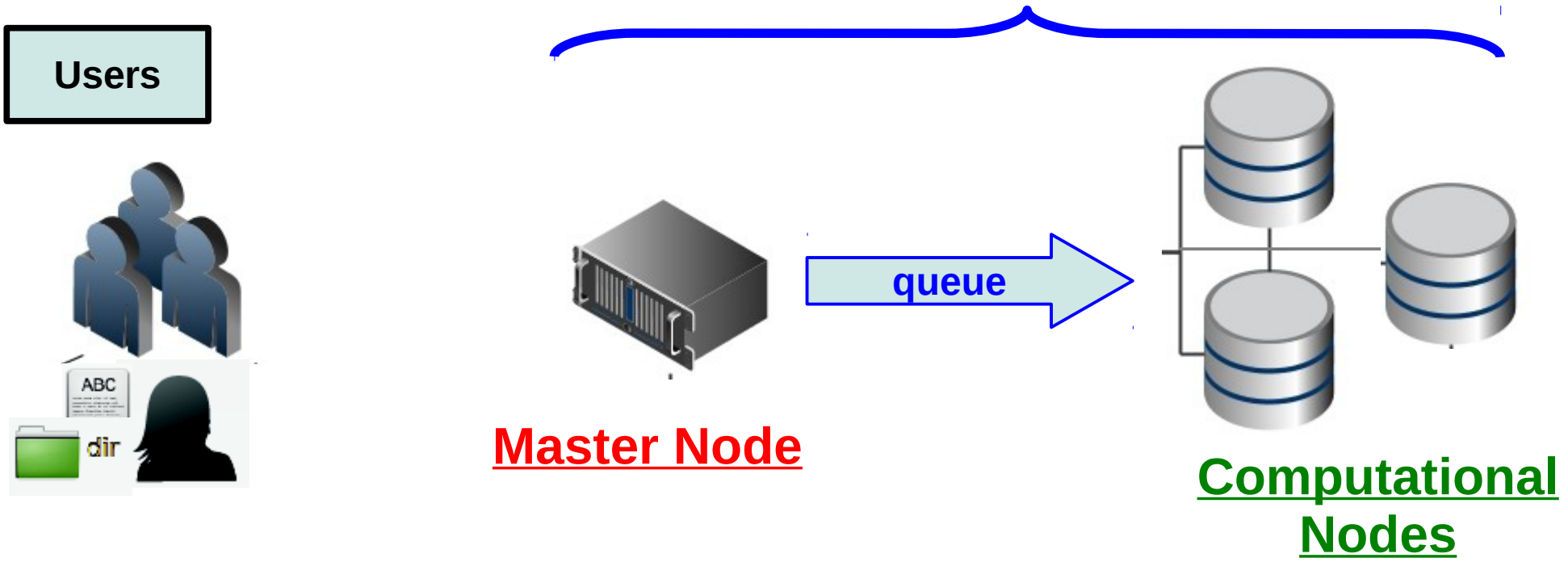
Sun Grid Engine (SGE)



- **Computer clusters** are the *classical* **High Performance Computing (HPC)** platforms used in bioinformatics.

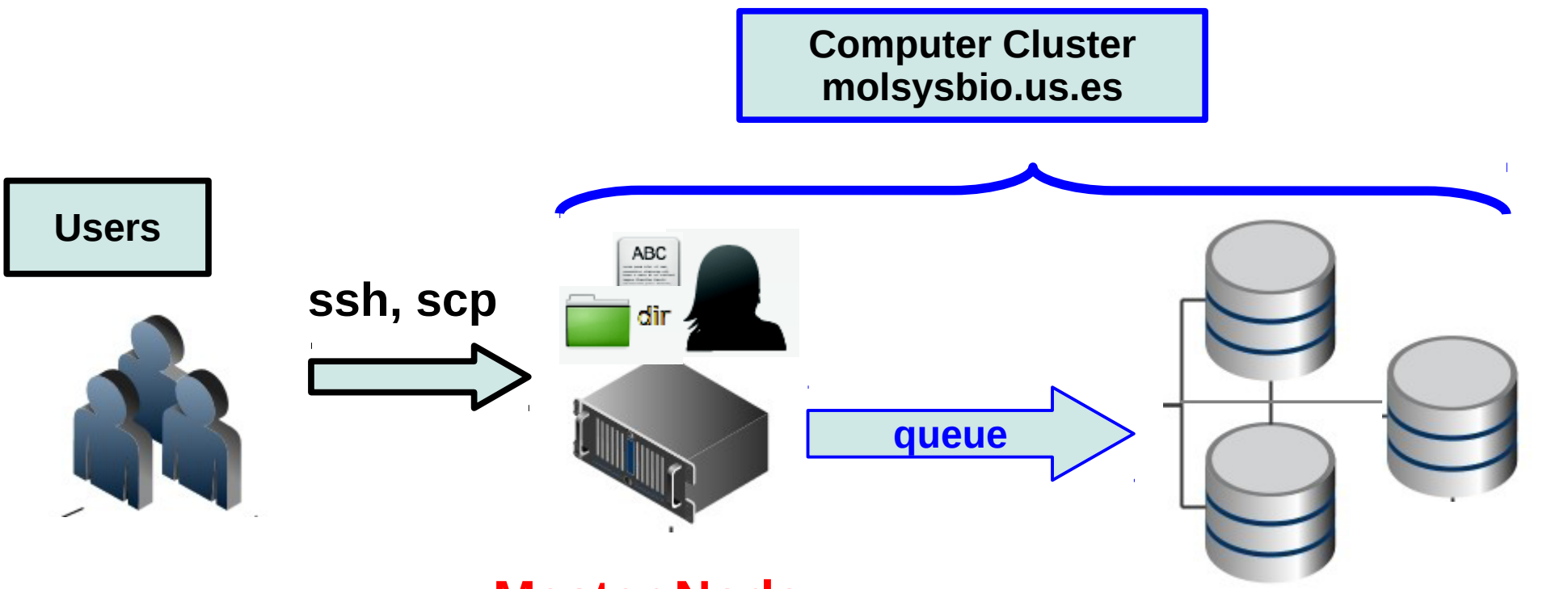
- Computer clusters are composed of different computers, processors or nodes. Typically, a specific node called the **master node** performs the administration of the jobs that are run on the rest of the nodes called **computational nodes**.

Sun Grid Engine (SGE)



- **Sun Grid Engine (SGE)** is one of the most commonly used HPC clusters managing open-source tools.
- It is responsible for accepting, scheduling, dispaching and managing the distributed execution of different user jobs.
- **SLURM**, is a popular alternative to SGE that is getting popular.

Sun Grid Engine (SGE)

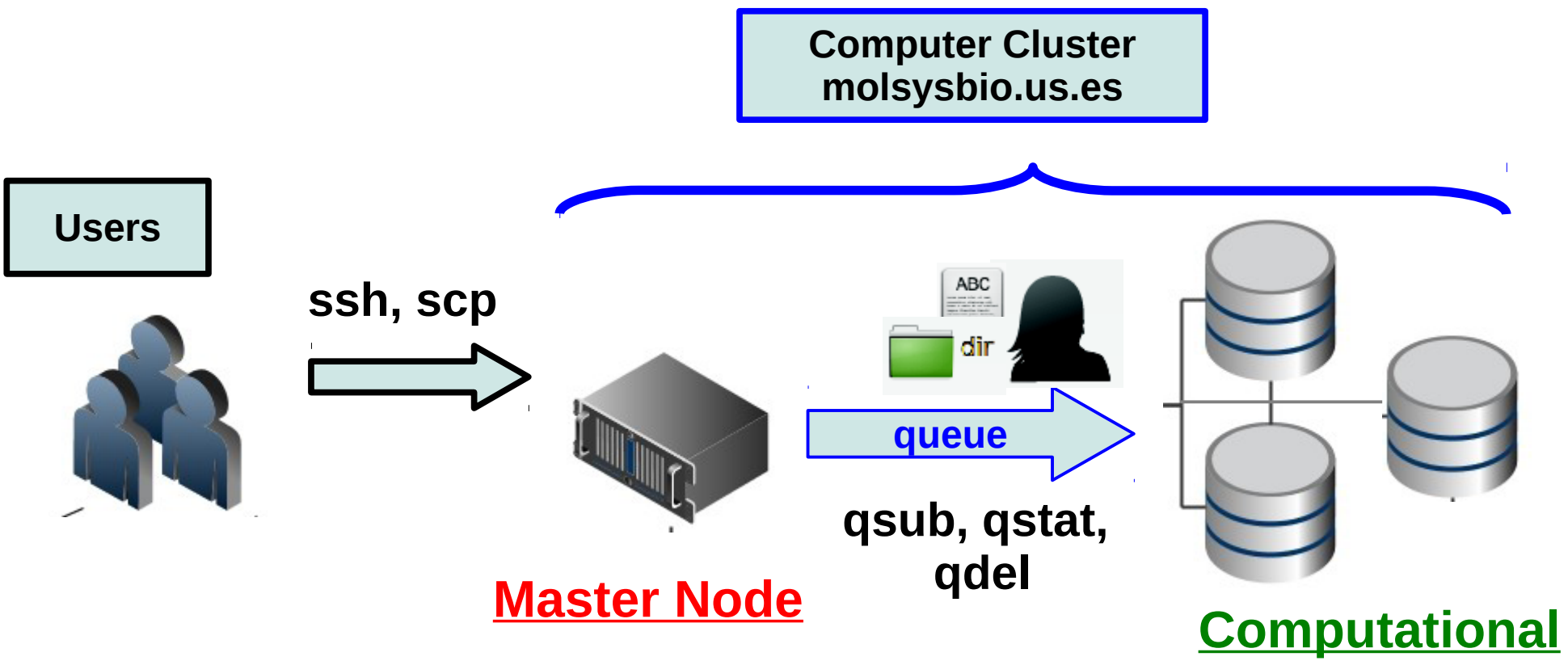


Users can access the master node:
ssh user_name@computer_name

Users can upload files to the system:
scp file_name user_name@computer_name:

Users can upload directories to the system:
scp -r directory_name user_name@computer_name:

Sun Grid Engine (SGE)

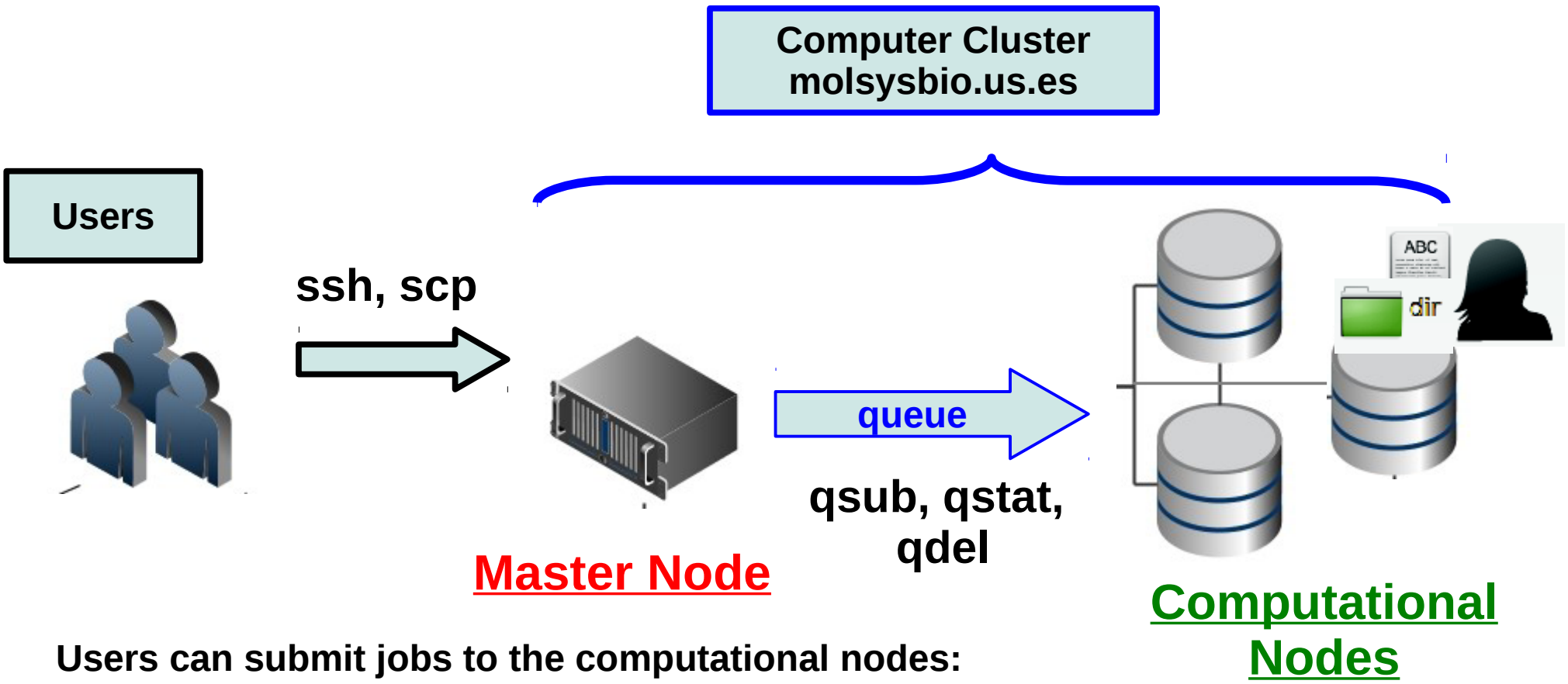


Users can submit jobs to the computational nodes:
qsub script.sh

Users can check the queue status:
qstat -u '*'

Users can remove jobs from the queue:
qdel job_id

Sun Grid Engine (SGE)

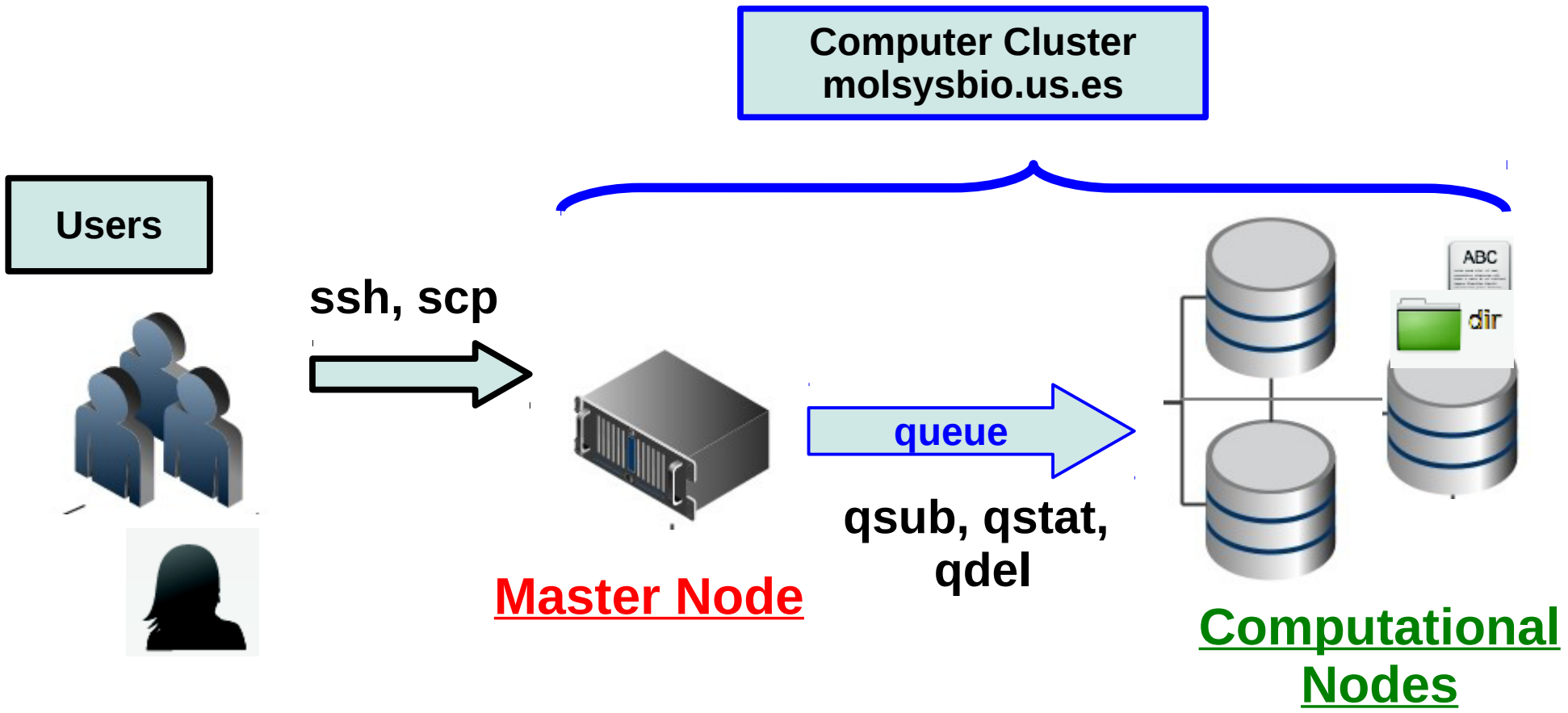


Users can submit jobs to the computational nodes:
qsub script.sh

Users can check the queue status:
qstat -u '*'

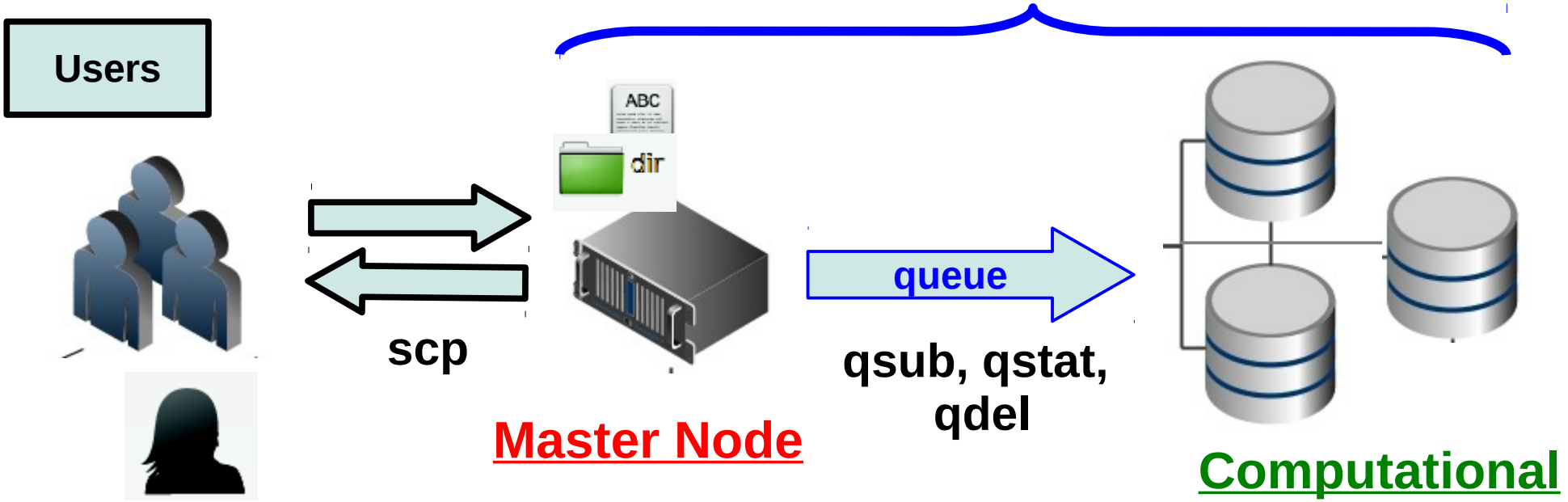
Users can remove jobs from the queue:
qdel job_id

Sun Grid Engine (SGE)



Users can exit the system:
exit

Sun Grid Engine (SGE)



Once the results are available users can download them from their computers:

scp user_name@computer_name:path_to_file/file .
scp -r user_name@computer_name:path_to_directory .

Sun Grid Engine (SGE)

Computer Cluster
molsysbio.us.es

Users

scp

Master Node

queue

qsub, qstat, qdel

Computational Nodes

Once the results are available users can download them from their computers:

scp user_name@computer_name:path_to_file/file .
scp -r user_name@computer_name:path_to_directory .

Sun Grid Engine (SGE)

Computer Cluster
molsysbio.us.es

Users

ssh, scp

queue

Master Node

Computational Nodes

Sun Grid Engine (SGE)

Computer Cluster
molsysbio.us.es

Users

ssh, scp

Master Node

queue

qsub, qstat, qdel

Computational Nodes

Sun Grid Engine (SGE)

Computer Cluster
molsysbio.us.es

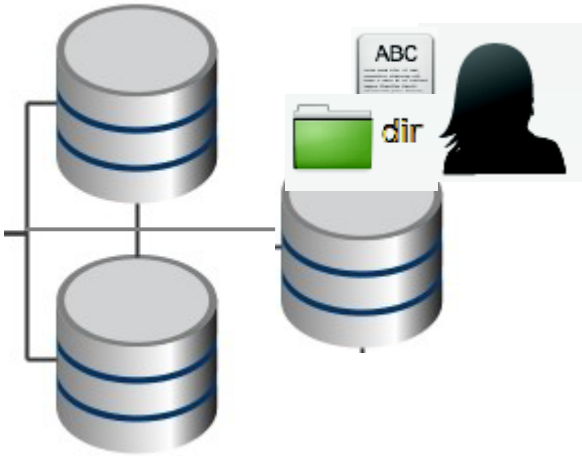
Users

ssh, scp

Master Node

queue

qsub, qstat, qdel



Computational Nodes

Sun Grid Engine (SGE)

Computer Cluster
molsysbio.us.es

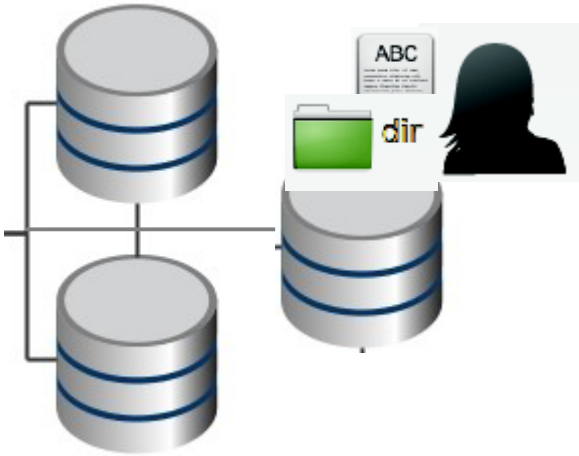
Users

ssh, scp

Master Node

queue

qsub, qstat, qdel



Computational Nodes

Sun Grid Engine (SGE)

Computer Cluster
molsysbio.us.es

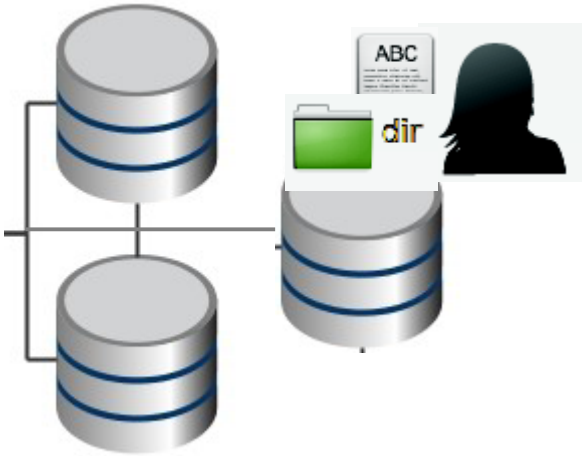
Users

ssh, scp

Master Node

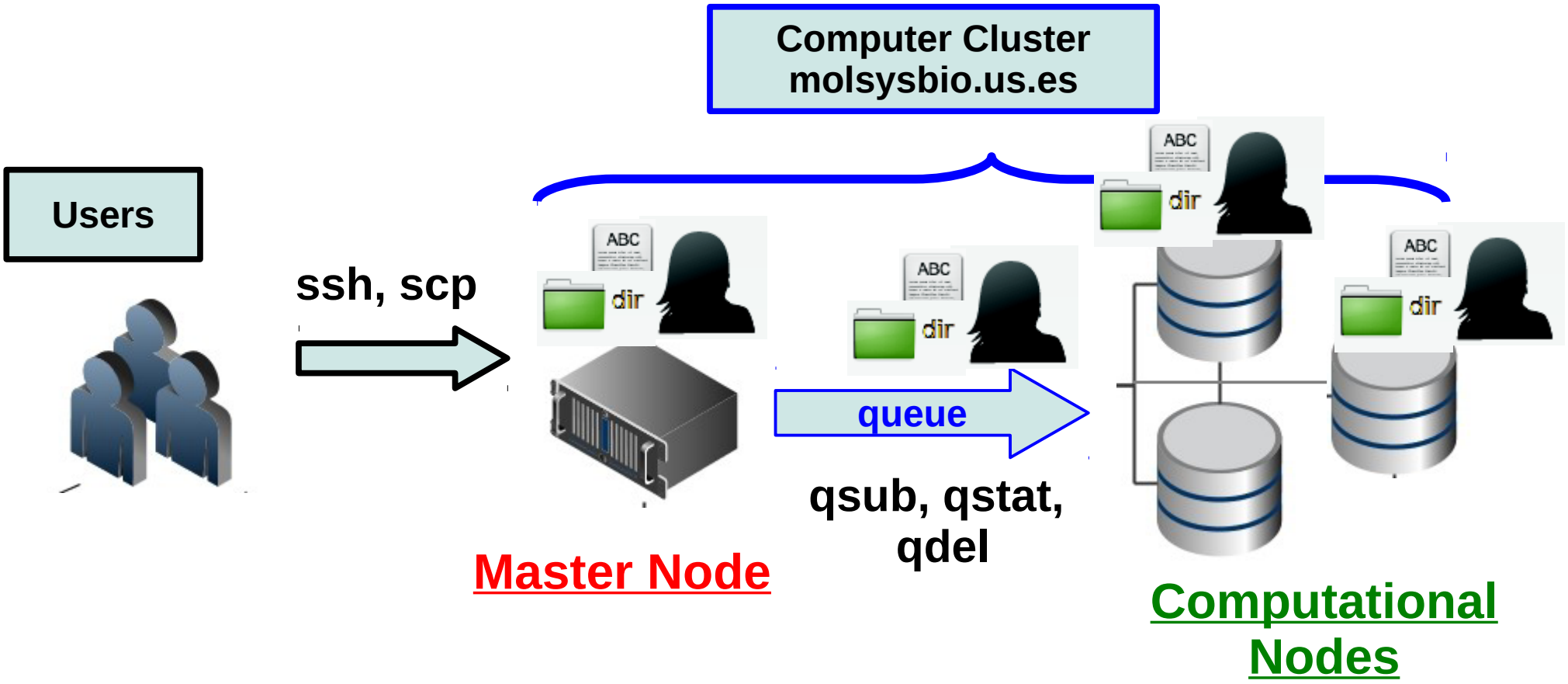
queue

qsub, qstat, qdel

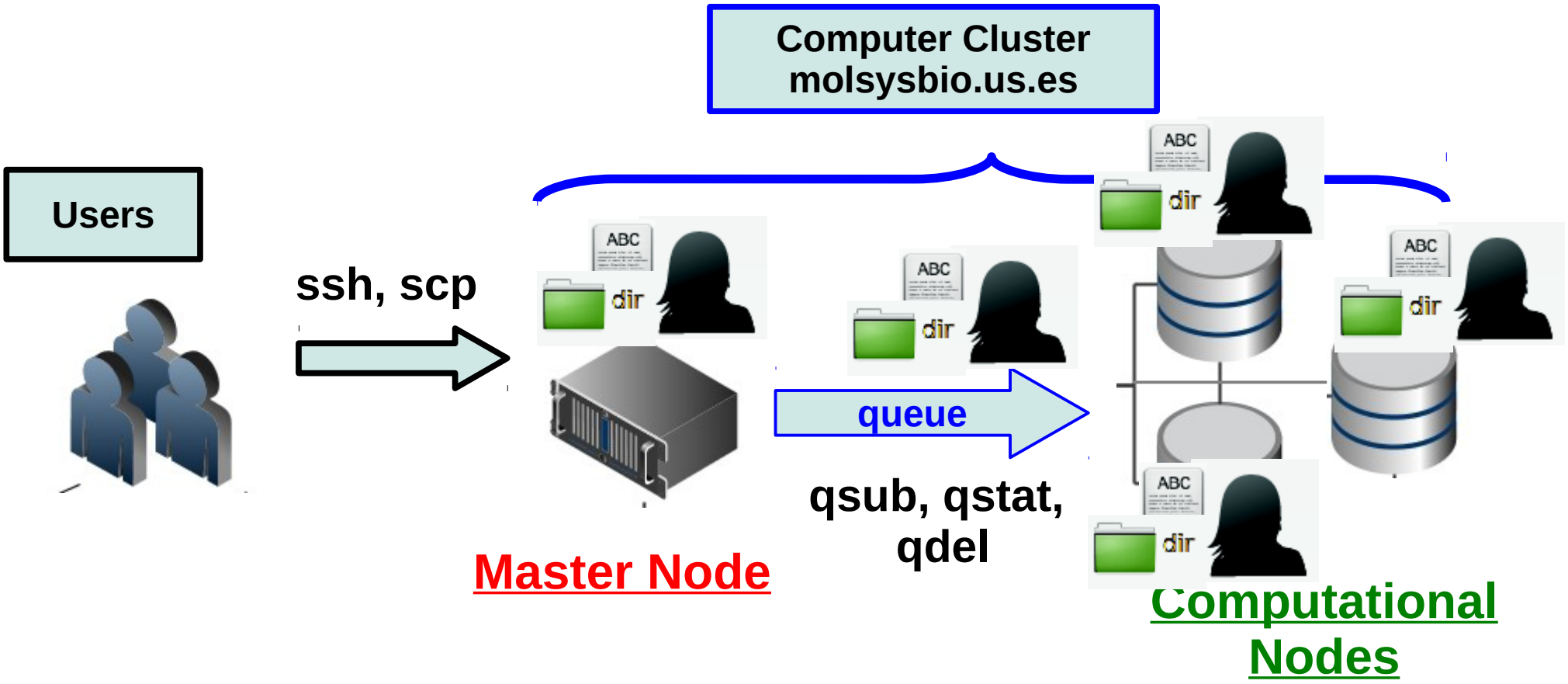


Computational Nodes

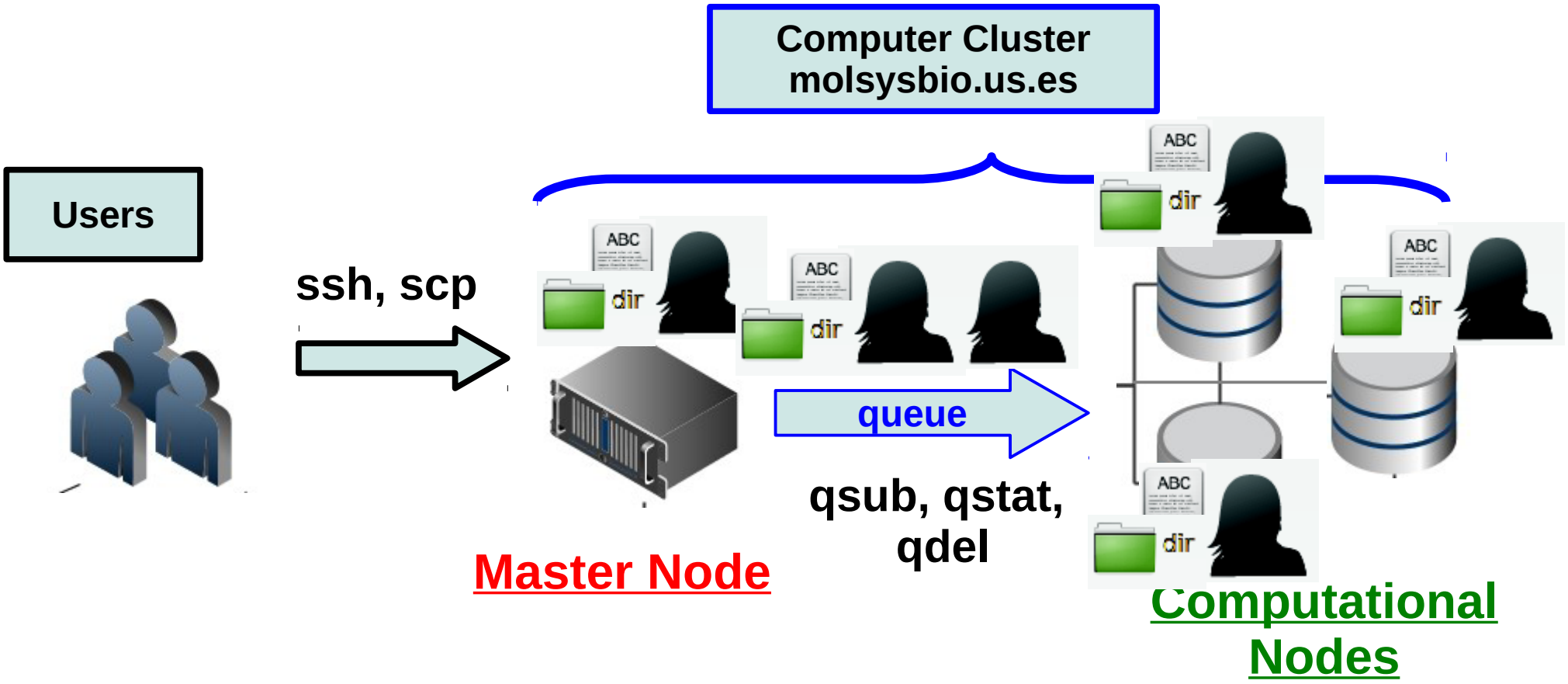
Sun Grid Engine (SGE)



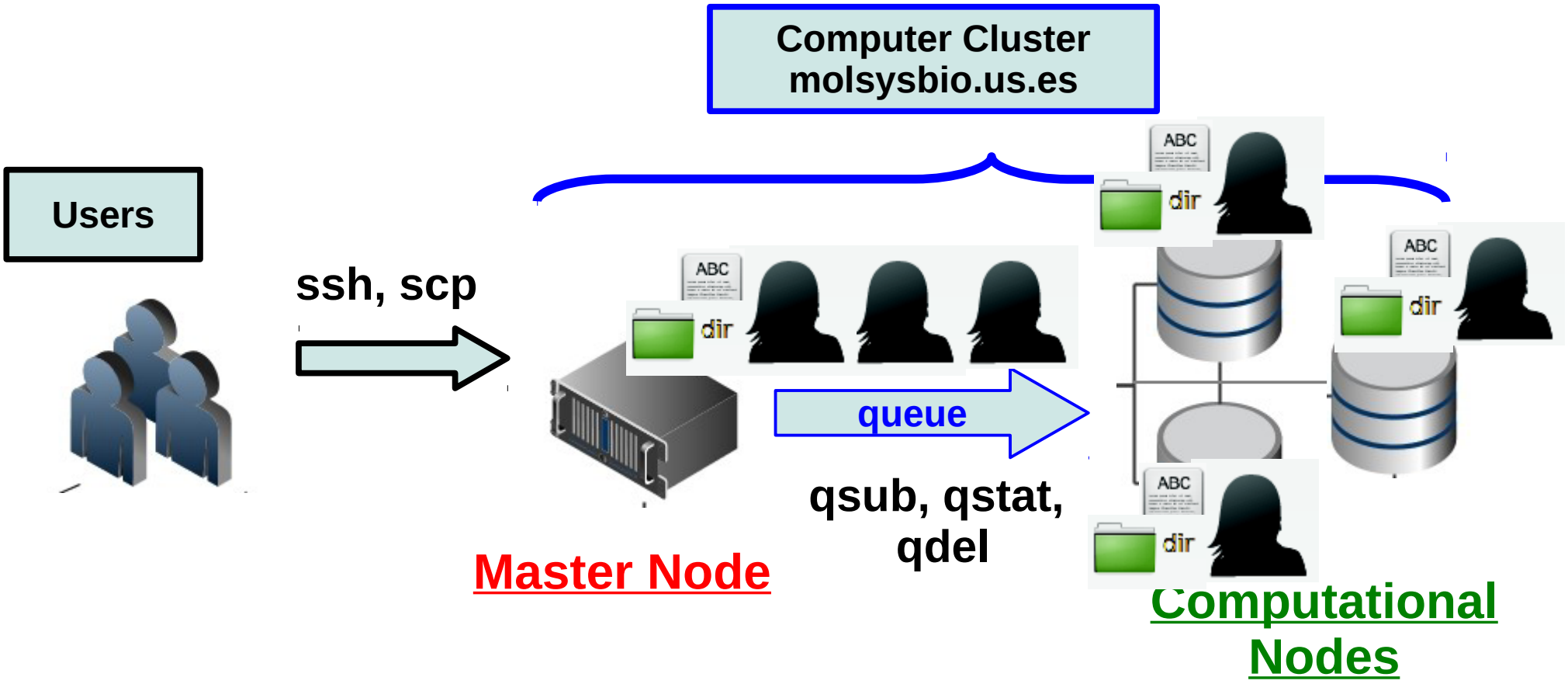
Sun Grid Engine (SGE)

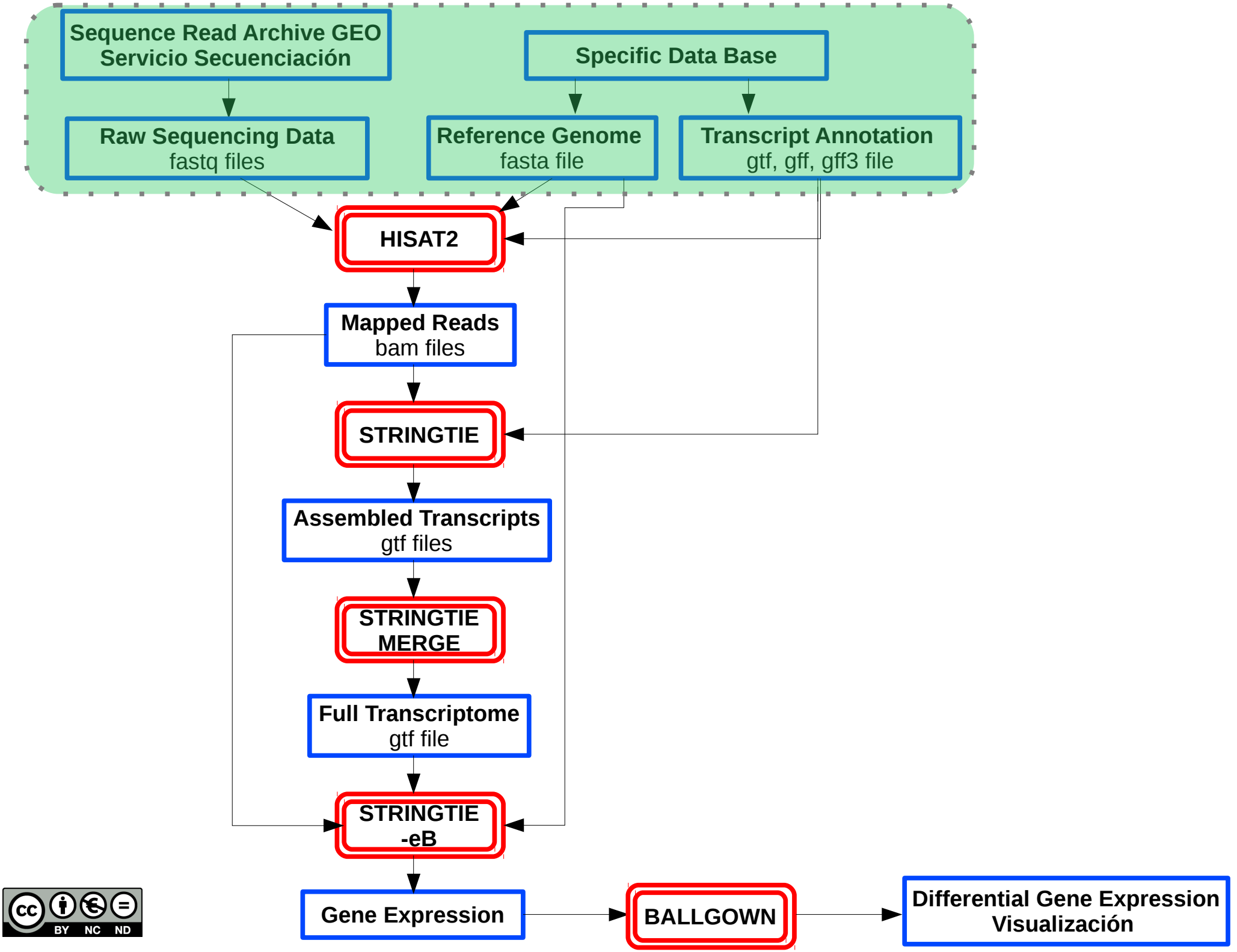


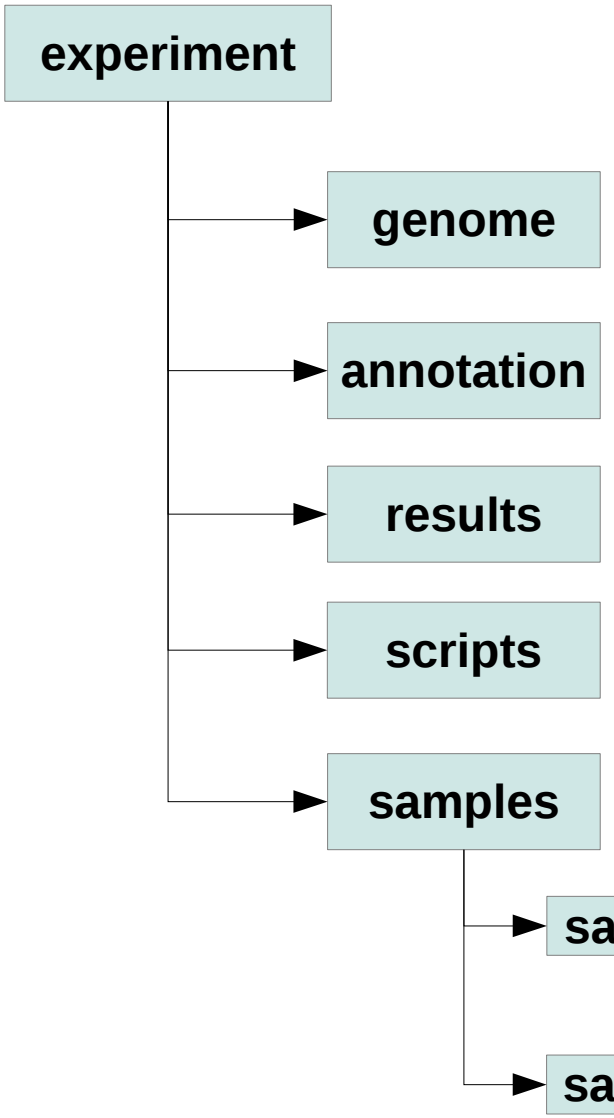
Sun Grid Engine (SGE)



Sun Grid Engine (SGE)




Sequence Read Archive GEO
Servicio Secuenciación
Specific Data Base
Raw Sequencing Data
fastq files
Reference Genome
fasta file
Transcript Annotation
gtf, gff, gff3 file
HISAT2
Mapped Reads
bam files
STRINGTIE
Assembled Transcripts
gtf files
STRINGTIE
MERGE
Full Transcriptome
gtf file
STRINGTIE
-eB
Gene Expression
BALLGOWN
Differential Gene Expression
Visualización



Contiene fichero fasta del *genoma de referencia* y el *índice* construído con HISAT2. Disponible en phytozome o Ensembl.

Contiene fichero gff3 con la anotación de la localización de genes, transcritos y exones en el *genoma de referencia.* Disponible en phytozome o Ensembl.

Almacenará los resultados del análisis de expresión génica diferencial.

Contiene los *scripts sge* para realizar el análisis de datos de RNA-seq.

Contiene en subcarpetas los ficheros fastq con las lecturas cortas obtenidas durante la secuenciación del RNA en cada muestra.

- Es crítico **mantener un espacio de trabajo ordenado**. Se recomienda la siguiente distribución:
  - Abrir un terminal con **Ctrl+Alt+t**
  - Para cada experimento a analizar crear una carpeta con el nombre apropiado. **mkdir nombre_carpeta**
  - Para ver el contenido de una carpeta escribir **ls**
  - Acceder a la carpeta principal: **cd nombre_carpeta**
  - Dentro de la carpeta principal crear cuatro subcarpetas:

    mkdir genome annotation samples results scripts
  - Acceder a la carpeta samples
    cd samples
  - Crear una subcarpeta para cada muestra

    mkdir sample_1 sample_2 sample_3 sample_4

Ejemplo de índice y anotacion (build_index.sh)

```
#$ -S /bin/bash
#$ -N index
#$ -V
#$ -cwd
#$ -j yes
#$ -o genome_index

# Acceder a la carpeta donde se guarda el genoma de referencia
cd /home/<grupo>/<exp>/genome
# Descargar usando el enlace de ENSEMBL el genoma de referencia
wget <enlace al genoma de referencia obtenido desde ENSEMBL>
# Descomprimir el genoma de referencia
gunzip <nombre_fichero_fasta>.fa.gz

# Acceder a la carpeta donde se guarda la anotacion
cd /home/<grupo>/<exp>/annotation
# Descargar usando el enlace de ENSEMBL la anotacion
wget <enlace a la anotacion obtenido desde ENSEMBL>
# Descomprimir la anotacion
gunzip <nombre_fichero_gtf>.gz

## Construir el indice del genoma de referencia
extract_splice_sites.py <annotation.gtf> > splice.ss
extract_exons.py <annotation.gtf> > annot_exons.exon
cd /home/<grupo>/<exp>/genome
hisat2-build --ss ../annotation/splice.ss --exon ../annotation/annot_exons.exon <genome.fa> <genome>
```

Ejemplo de índice y anotacion (build_

```
#$ -S /bin/bash
#$ -N index
#$ -V
#$ -cwd
#$ -j yes
#$ -o genome_index
```

Programa de ejecución del script

Nombre de la tarea

Fichero de salida

```
# Acceder a la carpeta donde se guarda el genoma
cd /home/<grupo>/<exp>/genome
# Descargar usando el enlace de ENSEMBL el genoma de referencia
wget <enlace al genoma de referencia obtenido desde ENSEMBL>
# Descomprimir el genoma de referencia
gunzip <nombre_fichero_fasta>.fa.gz

# Acceder a la carpeta donde se guarda la anotacion
cd /home/<grupo>/<exp>/annotation
# Descargar usando el enlace de ENSEMBL la anotacion
wget <enlace a la anotacion obtenido desde ENSEMBL>
# Descomprimir la anotacion
gunzip <nombre_fichero_gtf>.gz

## Construir el indice del genoma de referencia
extract_splice_sites.py <annotation.gtf> > splice.ss
extract_exons.py <annotation.gtf> > annot_exons.exon
cd /home/<grupo>/<exp>/genome
hisat2-build --ss ../annotation/splice.ss --exon ../annotation/annot_exons.exon <genome.fa> <genome>
```

Ejemplo de índice y anotacion (build_index.sh)

```
#$ -S /bin/bash
#$ -N index
#$ -V
#$ -cwd
#$ -j yes
#$ -o genome_index

# Acceder a la carpeta donde se guarda el genoma de referencia
cd /home/<grupo>/<exp>/genome
# Descargar usando el enlace de ENSEMBL el genoma de referencia
wget <enlace al genoma de referencia obtenido desde ENSEMBL>
# Descomprimir el genoma de referencia
gunzip <nombre_fichero_fasta>.fa.gz

# Acceder a la carpeta donde se guarda la anotacion
cd /home/<grupo>/<exp>/annotation
# Descargar usando el enlace de ENSEMBL la anotacion
wget <enlace a la anotacion obtenido desde ENSEMBL>
# Descomprimir la anotacion
gunzip <nombre_fichero_gtf>.gz

## Construir el indice del genoma de referencia
extract_splice_sites.py <annotation.gtf> > splice.ss
extract_exons.py <annotation.gtf> > annot_exons.exon
cd /home/<grupo>/<exp>/genome
hisat2-build --ss ../annotation/splice.ss --exon ../annotation/annot_exons.exon <genome.fa> <genome>
```

Ejemplo de índice y anotacion (build_index.sh)

```
#$ -S /bin/bash
#$ -N index
#$ -V
#$ -cwd
#$ -j yes
#$ -o genome_index

# Acceder a la carpeta donde se guarda el genoma de referencia
cd /home/<grupo>/<exp>/genome
# Descargar usando el enlace de ENSEMBL el genoma de referencia
wget <enlace al genoma de referencia obtenido desde ENSEMBL>
# Descomprimir el genoma de referencia
gunzip <nombre_fichero_fasta>.fa.gz

# Acceder a la carpeta donde se guarda la anotacion
cd /home/<grupo>/<exp>/annotation
# Descargar usando el enlace de ENSEMBL la anotacion
wget <enlace a la anotacion obtenido desde ENSEMBL>
# Descomprimir la anotacion
gunzip <nombre_fichero_gtf>.gz

## Construir el indice del genoma de referencia
extract_splice_sites.py <annotation.gtf> > splice.ss
extract_exons.py <annotation.gtf> > annot_exons.exon
cd /home/<grupo>/<exp>/genome
hisat2-build --ss ../annotation/splice.ss --exon ../annotation/annot_exons.exon <genome.fa> <genome>
```

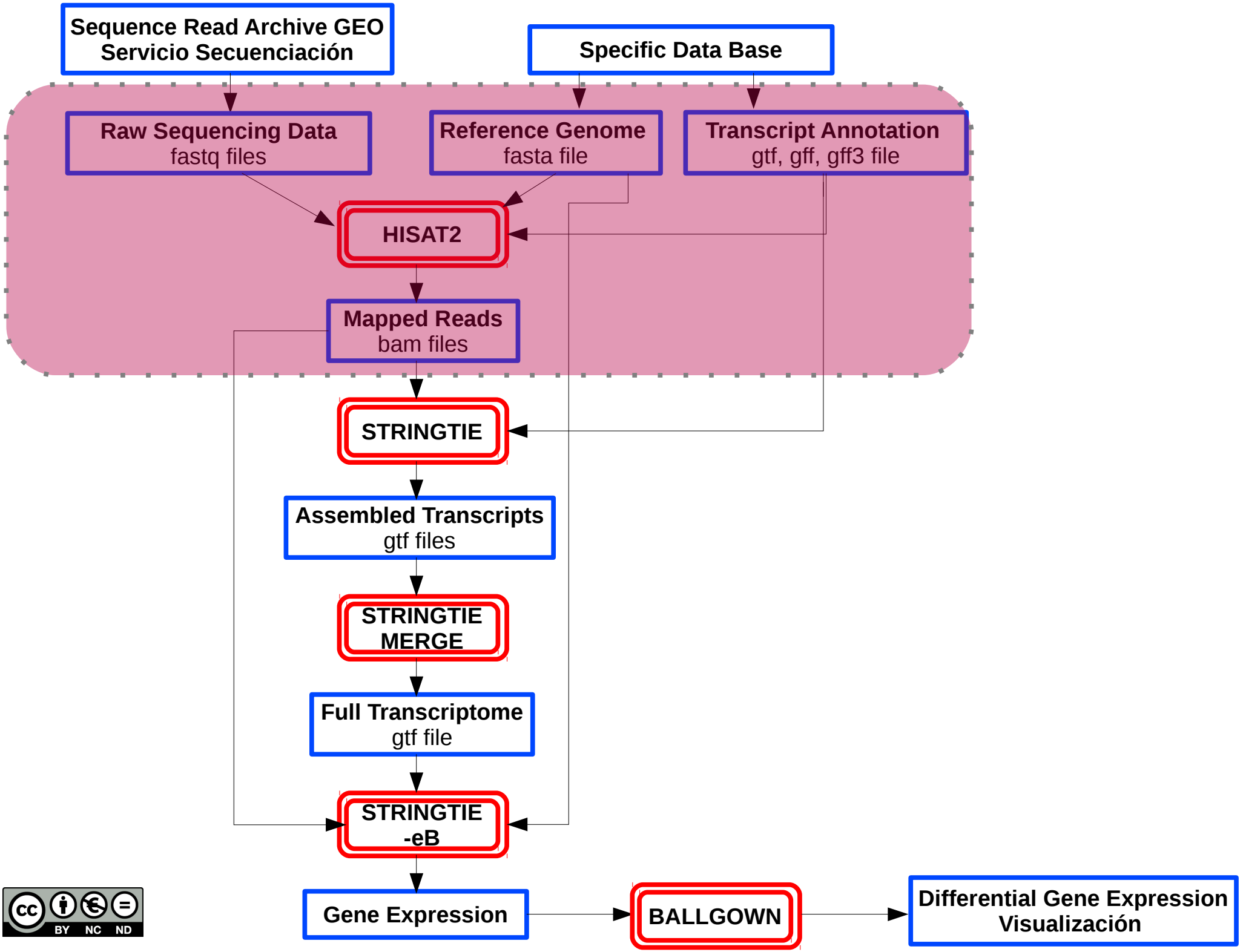
Ejemplo de índice y anotacion (build_index.sh)

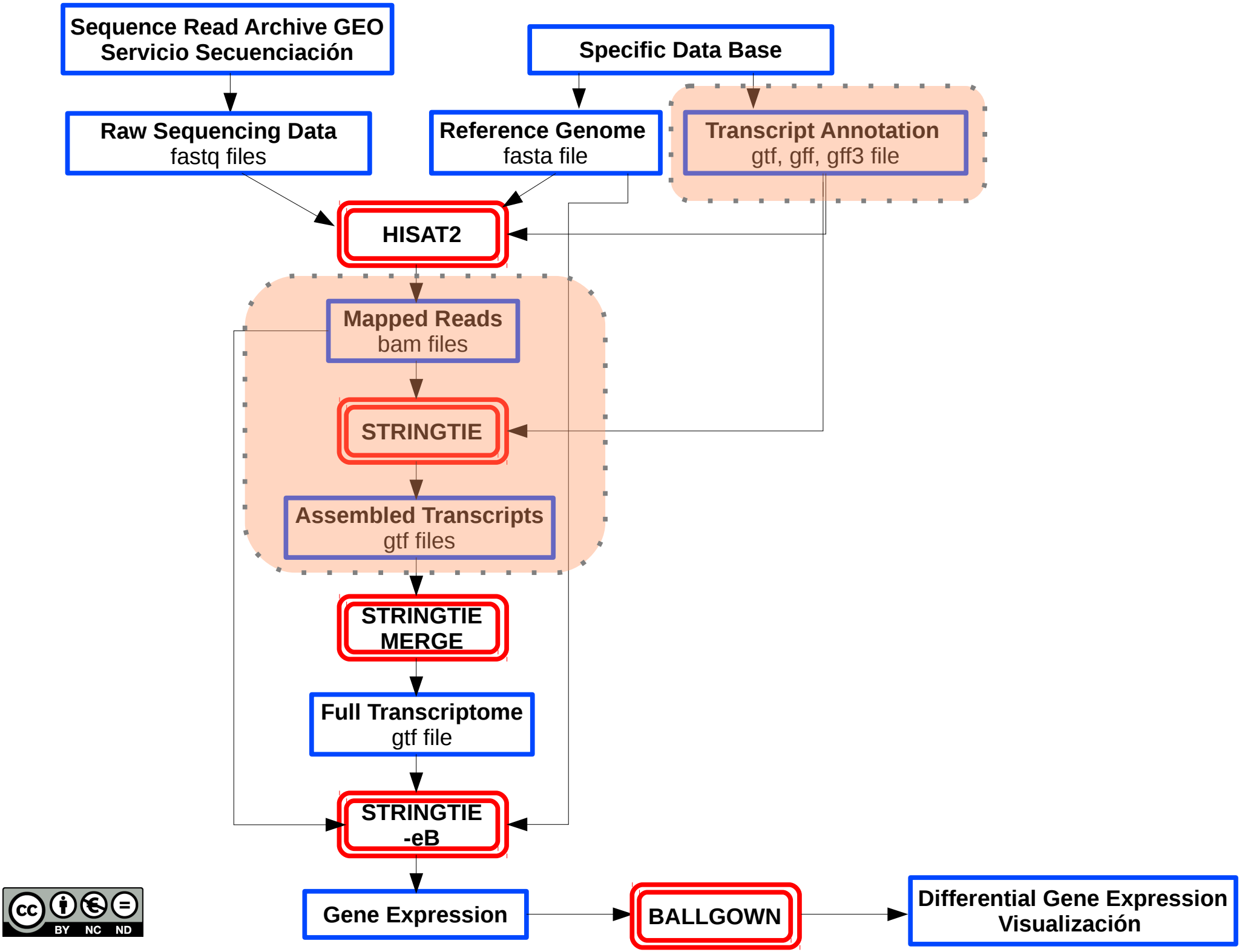
```
#$ -S /bin/bash
#$ -N index
#$ -V
#$ -cwd
#$ -j yes
#$ -o genome_index

# Acceder a la carpeta donde se guarda el genoma de referencia
cd /home/<grupo>/<exp>/genome
# Descargar usando el enlace de ENSEMBL el genoma de referencia
wget <enlace al genoma de referencia obtenido desde ENSEMBL>
# Descomprimir el genoma de referencia
gunzip <nombre_fichero_fasta>.fa.gz

# Acceder a la carpeta donde se guarda la anotacion
cd /home/<grupo>/<exp>/annotation
# Descargar usando el enlace de ENSEMBL la anotacion
wget <enlace a la anotacion obtenido desde ENSEMBL>
# Descomprimir la anotacion
gunzip <nombre_fichero_gtf>.gz
```

```
## Construir el indice del genoma de referencia
extract_splice_sites.py <annotation.gtf> > splice.ss
extract_exons.py <annotation.gtf> > annot_exons.exon
cd /home/<grupo>/<exp>/genome
hisat2-build --ss ../annotation/splice.ss --exon ../annotation/annot_exons.exon <genome.fa> <genome>
```


Sequence Read Archive GEO
Servicio Secuenciación
Specific Data Base
Raw Sequencing Data
fastq files
Reference Genome
fasta file
Transcript Annotation
gtf, gff, gff3 file
HISAT2
Mapped Reads
bam files
STRINGTIE
Assembled Transcripts
gtf files
STRINGTIE
MERGE
Full Transcriptome
gtf file
STRINGTIE
-eB
Gene Expression
BALLGOWN
Differential Gene Expression
Visualización


Sequence Read Archive GEO
Servicio Secuenciación
Specific Data Base
Raw Sequencing Data
fastq files
Reference Genome
fasta file
Transcript Annotation
gtf, gff, gff3 file
HISAT2
Mapped Reads
bam files
STRINGTIE
Assembled Transcripts
gtf files
STRINGTIE
MERGE
Full Transcriptome
gtf file
STRINGTIE
-eB
Gene Expression
BALLGOWN
Differential Gene Expression
Visualización

Script de alineamiento y ensamblado

```
#$ -S /bin/bash
#$ -N sample_<N>
#$ -V
#$ -cwd
#$ -j yes
#$ -o sample_<N>

# Acceder a la carpeta de la muestra N
cd /home/<grupo>/<exp>/samples/sample_<N>
# Descargar el fichero sra usando el enlace de GEO
wget <enlace_de_GEO_al_sra_de_la_muestra>

# Extraer el fichero fastq, borrar el fichero sra y realizar el control de la calidad
fastq-dump [--split-files] <fichero_muestra>.sra
rm<fichero_muestra>.sra
fastqc <fichero_muestra>.fastq

# Mapeo de las lecturas cortas al genoma de referencia
hisat2 --dta -x /home/<grupo>/<exp>/genome/<prefijo_index> -U <muestra>.fastq -S
<muestra.sam>
samtools sort -o <muestra.bam> <muestra.sam>

# Ensamblado de transcritos
stringtie -G /home/<grupo>/<exp>/annotation/<fichero_gtf> -o <sample.gtf> -l <sample>
<sample.bam>
```

Script de alineamiento y ensamblado

```
#$ -S /bin/bash
#$ -N sample_<N>
#$ -V
#$ -cwd
#$ -j yes
#$ -o sample_<N>
```

```
# Acceder a la carpeta de la muestra N
cd /home/<grupo>/<exp>/samples/sample_<N>
# Descargar el fichero sra usando el enlace de GEO
wget <enlace_de_GEO_al_sra_de_la_muestra>

# Extraer el fichero fastq, borrar el fichero sra y realizar el control de la calidad
fastq-dump [--split-files] <fichero_muestra>.sra
rm<fichero_muestra>.sra
fastqc <fichero_muestra>.fastq

# Mapeo de las lecturas cortas al genoma de referencia
hisat2 --dta -x /home/<grupo>/<exp>/genome/<prefijo_index> -U <muestra>.fastq -S
<muestra.sam>
samtools sort -o <muestra.bam> <muestra.sam>

# Ensamblado de transcritos
stringtie -G /home/<grupo>/<exp>/annotation/<fichero_gtf> -o <sample.gtf> -l <sample>
<sample.bam>
```

Script de alineamiento y ensamblado

```
#$ -S /bin/bash
#$ -N sample_<N>
#$ -V
#$ -cwd
#$ -j yes
#$ -o sample_<N>

# Acceder a la carpeta de la muestra N
cd /home/<grupo>/<exp>/samples/sample_<N>
# Descargar el fichero sra usando el enlace de GEO
wget <enlace_de_GEO_al_sra_de_la_muestra>

# Extraer el fichero fastq, borrar el fichero sra y realizar el control de la calidad
fastq-dump [--split-files] <fichero_muestra>.sra
rm<fichero_muestra>.sra
fastqc <fichero_muestra>.fastq

# Mapeo de las lecturas cortas al genoma de referencia
hisat2 --dta -x /home/<grupo>/<exp>/genome/<prefijo_index> -U <muestra>.fastq -S
<muestra.sam>
samtools sort -o <muestra.bam> <muestra.sam>

# Ensamblado de transcritos
stringtie -G /home/<grupo>/<exp>/annotation/<fichero_gtf> -o <sample.gtf> -l <sample>
<sample.bam>
```

Script de alineamiento y ensamblado

```
#$ -S /bin/bash
#$ -N sample_<N>
#$ -V
#$ -cwd
#$ -j yes
#$ -o sample_<N>

# Acceder a la carpeta de la muestra N
cd /home/<grupo>/<exp>/samples/sample_<N>
# Descargar el fichero sra usando el enlace de GEO
wget <enlace_de_GEO_al_sra_de_la_muestra>

# Extraer el fichero fastq, borrar el fichero sra y realizar el control de la calidad
fastq-dump [--split-files] <fichero_muestra>.sra
rm<fichero_muestra>.sra
fastqc <fichero_muestra>.fastq

# Mapeo de las lecturas cortas al genoma de referencia
hisat2 --dta -x /home/<grupo>/<exp>/genome/<prefijo_index> -U <muestra>.fastq -S
<muestra.sam>
samtools sort -o <muestra.bam> <muestra.sam>

# Ensamblado de transcritos
stringtie -G /home/<grupo>/<exp>/annotation/<fichero_gtf> -o <sample.gtf> -l <sample>
<sample.bam>
```

Script de alineamiento y ensamblado

```
#$ -S /bin/bash
#$ -N sample_<N>
#$ -V
#$ -cwd
#$ -j yes
#$ -o sample_<N>

# Acceder a la carpeta de la muestra N
cd /home/<grupo>/<exp>/samples/sample_<N>
# Descargar el fichero sra usando el enlace de GEO
wget <enlace_de_GEO_al_sra_de_la_muestra>

# Extraer el fichero fastq, borrar el fichero sra y realizar el control de la calidad
fastq-dump [--split-files] <fichero_muestra>.sra
rm<fichero_muestra>.sra
fastqc <fichero_muestra>.fastq

# Mapeo de las lecturas cortas al genoma de referencia
hisat2 --dta -x /home/<grupo>/<exp>/genome/<prefijo_index> -U <muestra>.fastq -S
<muestra.sam>
samtools sort -o <muestra.bam> <muestra.sam>

# Ensamblado de transcritos
stringtie -G /home/<grupo>/<exp>/annotation/<fichero_gtf> -o <sample.gtf> -l <sample>
<sample.bam>
```

Script de alineamiento y ensamblado

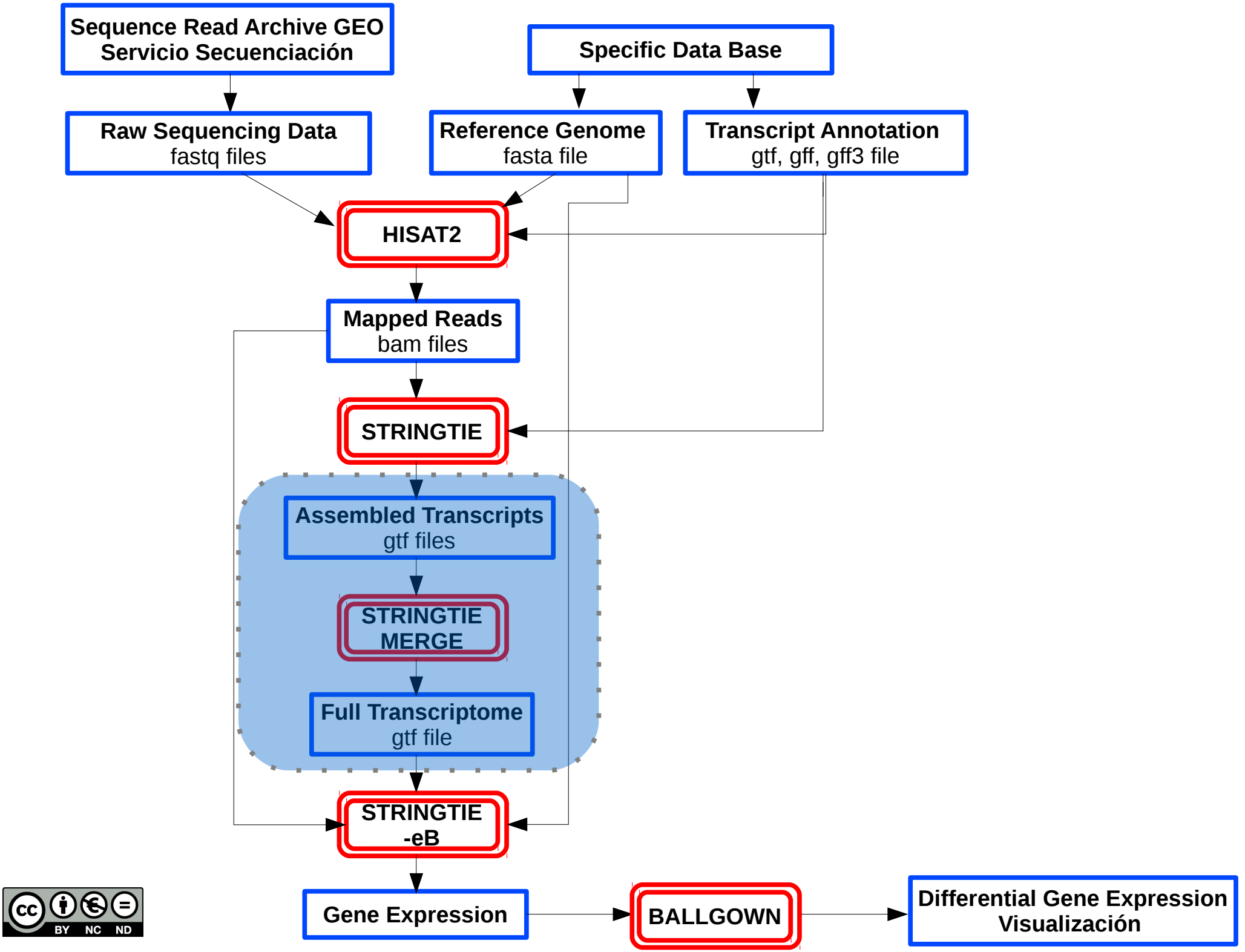
```
#$ -S /bin/bash
#$ -N sample_<N>
#$ -V
#$ -cwd
#$ -j yes
#$ -o sample_<N>

# Acceder a la carpeta de la muestra N
cd /home/<grupo>/<exp>/samples/sample_<N>
# Descargar el fichero sra usando el enlace de GEO
wget <enlace_de_GEO_al_sra_de_la_muestra>

# Extraer el fichero fastq, borrar el fichero sra y realizar el control de la calidad
fastq-dump [--split-files] <fichero_muestra>.sra
rm<fichero_muestra>.sra
fastqc <fichero_muestra>.fastq

# Mapeo de las lecturas cortas al genoma de referencia
hisat2 --dta -x /home/<grupo>/<exp>/genome/<prefijo_index> -U <muestra>.fastq -S
<muestra.sam>
samtools sort -o <muestra.bam> <muestra.sam>

# Ensamblado de transcritos
stringtie -G /home/<grupo>/<exp>/annotation/<fichero_gtf> -o <sample.gtf> -l <sample>
<sample.bam>
```


Sequence Read Archive GEO
Servicio Secuenciación
Specific Data Base
Raw Sequencing Data
fastq files
Reference Genome
fasta file
Transcript Annotation
gtf, gff, gff3 file
HISAT2
Mapped Reads
bam files
STRINGTIE
Assembled Transcripts
gtf files
STRINGTIE
MERGE
Full Transcriptome
gtf file
STRINGTIE
-eB
Gene Expression
BALLGOWN
Differential Gene Expression
Visualización

#$ -S /bin/bash
#$ -N merge
#$ -V
#$ -cwd
#$ -j yes
#$ -o merge

Generación del transcriptoma de estudio y comparación con la anotación

Acceder a la carpeta results
cd /home/<grupo>/<exp>/results/

Generar el fichero mergelist.txt con las rutas a los transcripts.gtf
echo /home/usuario/exp/samples/sample_1/sample_1.gtf > mergelist.txt
echo /home/usuario/exp/samples/sample_2/sample_2.gtf >> mergelist.txt
...

echo /home/usuario/exp/samples/sample_<M>/sample_<M>.gtf >> mergelist.txt

Combinar los trasncriptomas parciales de cada muestra
stringtie --merge -G /home/<grupo>/<exp>/annotation/<fichero_gtf> -o stringtie_merged.gtf mergelist.txt

Comparar transcriptoma completo con la referencia
gffcompare -r /home/<grupo>/<exp>/annotation/<fichero.gtf> -G -o merged mergelist.txt

Generación del transcriptoma de estudio y comparación con la anotación

```
#$ -S /bin/bash
#$ -N merge
#$ -V
#$ -cwd
#$ -j yes
#$ -o merge
```

Directrices generales

```
# Acceder a la carpeta results
cd /home/<grupo>/<exp>/results/


# Generar el fichero mergelist.txt con las rutas a los transcripts.gtf
echo /home/usuario/exp/samples/sample_1/sample_1.gtf > mergelist.txt
echo /home/usuario/exp/samples/sample_2/sample_2.gtf >> mergelist.txt
                                                    ...
echo /home/usuario/exp/samples/sample_<M>/sample_<M>.gtf >> mergelist.txt

# Combinar los trasncriptomas parciales de cada muestra
stringtie --merge -G /home/<grupo>/<exp>/annotation/<fichero_gtf> -o
stringtie_merged.gtf mergelist.txt

# Comparar transcriptoma completo con la referencia
gffcompare -r /home/<grupo>/<exp>/annotation/<fichero.gtf> -G -o merged mergelist.txt
```

```
#$ -S /bin/bash
#$ -N merge
#$ -V
#$ -cwd
#$ -j yes
#$ -o merge

# Acceder a la carpeta results
cd /home/<grupo>/<exp>/results/

# Generar el fichero mergelist.txt con las rutas a los transcripts.gtf
echo /home/usuario/exp/samples/sample_1/sample_1.gtf > mergelist.txt
echo /home/usuario/exp/samples/sample_2/sample_2.gtf >> mergelist.txt
                                                    ...
echo /home/usuario/exp/samples/sample_<M>/sample_<M>.gtf >> mergelist.txt

# Combinar los trasncriptomas parciales de cada muestra
stringtie --merge -G /home/<grupo>/<exp>/annotation/<fichero_gtf> -o
stringtie_merged.gtf mergelist.txt

# Comparar transcriptoma completo con la referencia
gffcompare -r /home/<grupo>/<exp>/annotation/<fichero.gtf> -G -o merged mergelist.txt
```

Generación del transcriptoma de estudio y expresión génica diferencial

```
#$ -S /bin/bash
#$ -N merge
#$ -V
#$ -cwd
#$ -j yes
#$ -o merge


# Acceder a la carpeta results
cd /home/<grupo>/<exp>/results/

# Generar el fichero mergelist.txt con las rutas a los transcripts.gtf
echo /home/usuario/exp/samples/sample_1/sample_1.gtf > mergelist.txt
echo /home/usuario/exp/samples/sample_2/sample_2.gtf >> mergelist.txt
                                                ...
echo /home/usuario/exp/samples/sample_<M>/sample_<M>.gtf >> mergelist.txt

# Combinar los trasncriptomas parciales de cada muestra
stringtie --merge -G /home/<grupo>/<exp>/annotation/<fichero_gtf> -o
stringtie_merged.gtf mergelist.txt

# Comparar transcriptoma completo con la referencia
gffcompare -r /home/<grupo>/<exp>/annotation/<fichero.gtf> -G -o merged mergelist.txt
```

Generación del transcriptoma de estudio y expresión génica diferencial

```
#$ -S /bin/bash
#$ -N merge
#$ -V
#$ -cwd
#$ -j yes
#$ -o merge


# Acceder a la carpeta results
cd /home/<grupo>/<exp>/results/


# Generar el fichero mergelist.txt con las rutas a los transcripts.gtf
echo /home/usuario/exp/samples/sample_1/sample_1.gtf > mergelist.txt
echo /home/usuario/exp/samples/sample_2/sample_2.gtf >> mergelist.txt
                                                    ...
echo /home/usuario/exp/samples/sample_<M>/sample_<M>.gtf >> mergelist.txt

# Combinar los trasncriptomas parciales de cada muestra
stringtie --merge -G /home/<grupo>/<exp>/annotation/<fichero_gtf> -o
stringtie_merged.gtf mergelist.txt

# Comparar transcriptoma completo con la referencia
gffcompare -r /home/<grupo>/<exp>/annotation/<fichero.gtf> -G -o merged mergelist.txt
```

Generación del transcriptoma de estudio y expresión génica diferencial

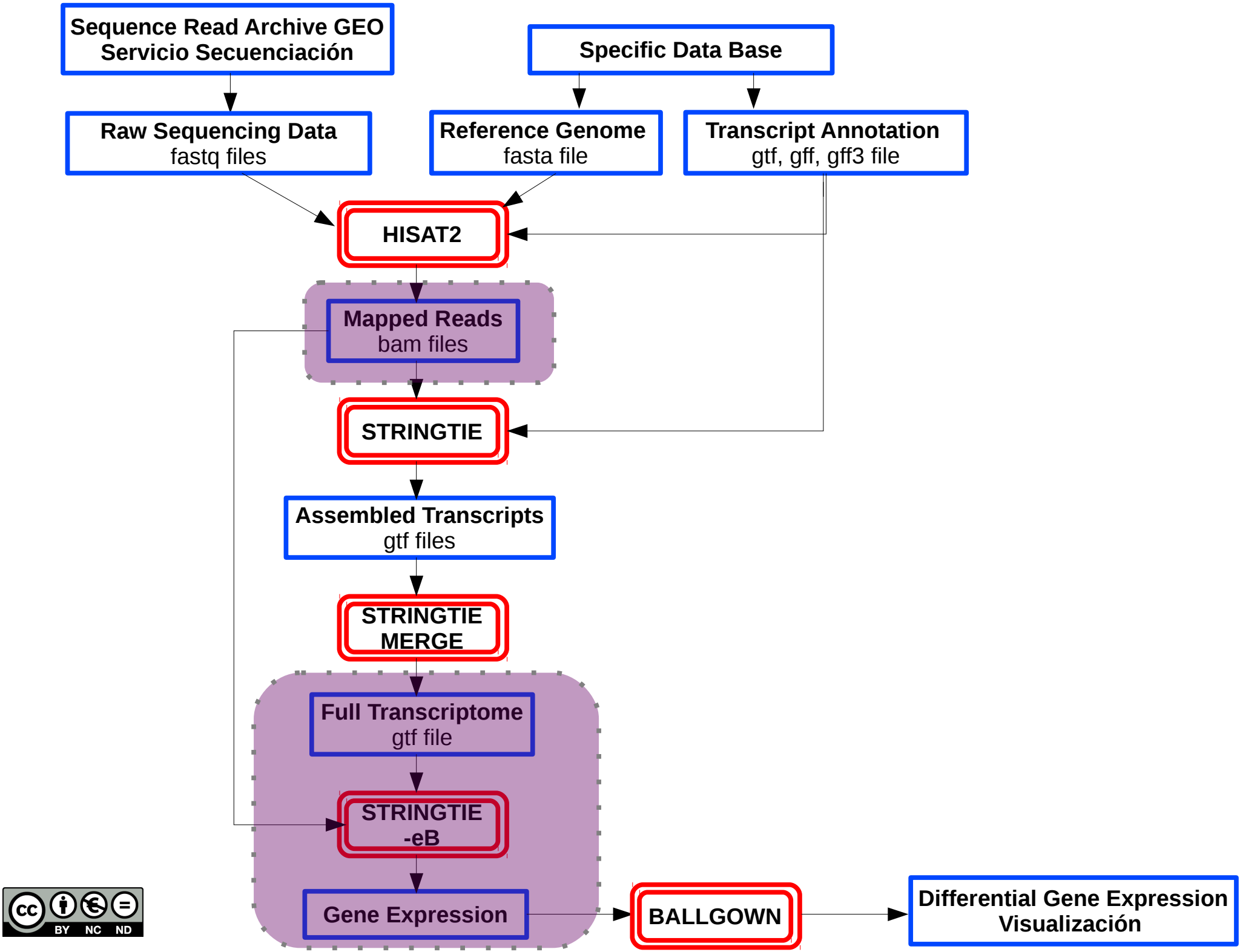
```
#$ -S /bin/bash
#$ -N merge
#$ -V
#$ -cwd
#$ -j yes
#$ -o merge


# Acceder a la carpeta results
cd /home/<grupo>/<exp>/results/


# Generar el fichero mergelist.txt con las rutas a los transcripts.gtf
echo /home/usuario/exp/samples/sample_1/sample_1.gtf > mergelist.txt
echo /home/usuario/exp/samples/sample_2/sample_2.gtf >> mergelist.txt
                                                    ...
echo /home/usuario/exp/samples/sample_<M>/sample_<M>.gtf >> mergelist.txt

# Combinar los trasncriptomas parciales de cada muestra
stringtie --merge -G /home/<grupo>/<exp>/annotation/<fichero_gtf> -o
stringtie_merged.gtf mergelist.txt
```

```
# Comparar transcriptoma completo con la referencia
gffcompare -r /home/<grupo>/<exp>/annotation/<fichero.gtf> -G -o merged mergelist.txt
```


Sequence Read Archive GEO
Servicio Secuenciación
Specific Data Base
Raw Sequencing Data
fastq files
Reference Genome
fasta file
Transcript Annotation
gtf, gff, gff3 file
HISAT2
Mapped Reads
bam files
STRINGTIE
Assembled Transcripts
gtf files
STRINGTIE
MERGE
Full Transcriptome
gtf file
STRINGTIE
-eB
Gene Expression
BALLGOWN
Differential Gene Expression
Visualización

Script de cuantificación de los niveles de expresión génicos

```
#$ -S /bin/bash
#$ -N quant_<N>
#$ -V
#$ -cwd
#$ -j yes
#$ -o quant_<N>

# Acceder a la carpeta de la muestra N
cd /home/<grupo>/<exp>/samples/sample_<N>

# Cuantificación de los niveles de expresión
stringtie -e -B  -G /home/<grupo>/<exp>/results/stringtie_merged.gtf -o <sample.gtf>
<sample.bam>

# Eliminar ficheros sam y fastq
rm <sample.sam>
rm *.fastq
```

Descargar resultados del fichero .db

- Una vez obtenidos los resultados nos los bajamos a nuestro ordenador para analizarlos con el paquete ballgown.
- Desde nuestro portátil y con una muy buena conexión a internet ejecutar la siguiente instrucción:

scp -r grupo@molsysbio.us.es:<exp>/samples .

Descargar resultados del fichero .db

- Una vez obtenidos los resultados nos los bajamos a nuestro ordenador para analizarlos con el paquete ballgown.
- Desde nuestro portátil y con una muy buena conexión a internet ejecutar la siguiente instrucción:

scp -r grupo@molsysbio.us.es:<exp>/samples .