

# Herramientas de Ensamblaje de Genomas

# Ensamblaje de Genomas

La genómica se caracteriza por el estudio de los genomas y esto se logra con un gran número de genes obtenidos con herramientas que coleccionan estos datos.

El ensamblado de genomas es parte de estas herramientas lo cual busca encontrar una representación del genoma en estudio.

Esta tarea no es nada fácil, por lo que se han creado diferentes algoritmos para la reconstrucción de las secuencias de genes.

Estos algoritmos automatizados son los encargados de realizar lecturas cortas de una secuencia. Actualmente existen herramientas automatizadas que tenemos la facilidad de instalar en nuestras computadoras para poder realizar diferentes estudios entre ellas el ensamblaje de genomas.

Se mostrará paso a paso la implementación de ciertas herramientas que son necesarias para realizar el ensamblaje de genomas desde nuestras computadoras.

# Herramientas a implementar

La implementación de las herramientas de este tutorial será en el sistema operativo ubuntu.

Las herramientas que se mostrará como instalar son:

- Minia.
- KmerGenie
- QUAST
- SSPACE.

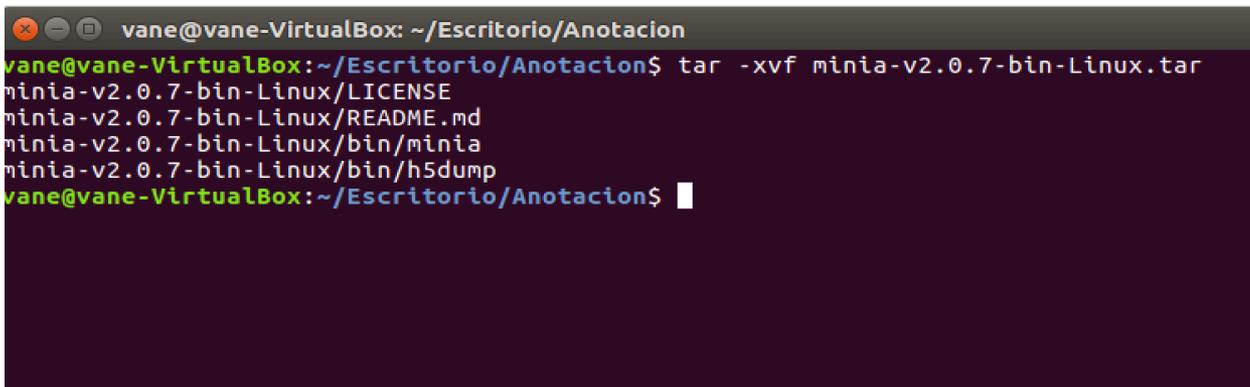
Hay pre-requisitos para estas herramientas que generalmente se encuentran instaladas de forma predeterminada en los sistemas operativos como ubuntu o biolinux que son:

- python
- R
- RStudio
- perl

# MINIA

---

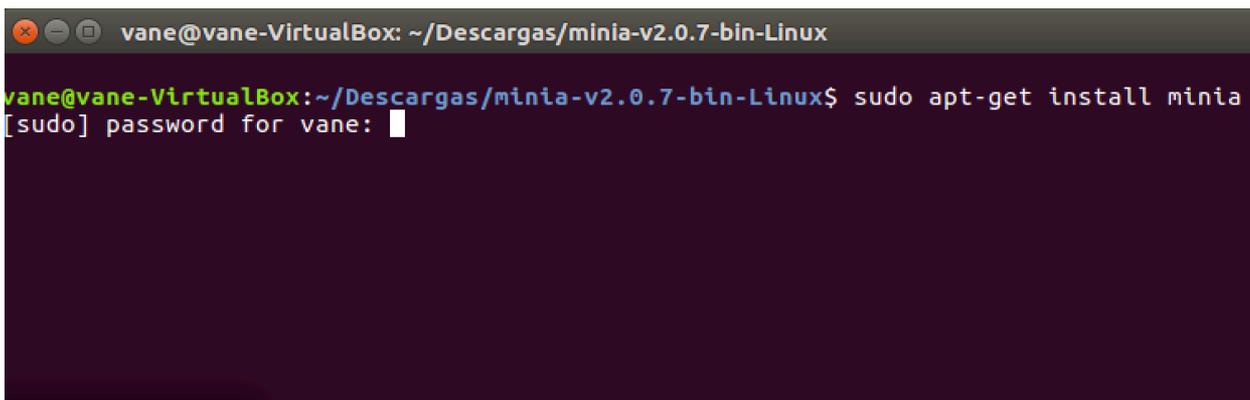
Minia es un ensamblador muy potente, basado en el algoritmo del gráfico de Bruijn. Minia produce resultados de contiguidad y precisión.



```
vane@vane-VirtualBox: ~/Escritorio/Anotacion
vane@vane-VirtualBox:~/Escritorio/Anotacion$ tar -xvf minia-v2.0.7-bin-Linux.tar
minia-v2.0.7-bin-Linux/LICENSE
minia-v2.0.7-bin-Linux/README.md
minia-v2.0.7-bin-Linux/bin/minia
minia-v2.0.7-bin-Linux/bin/h5dump
vane@vane-VirtualBox:~/Escritorio/Anotacion$
```

En la siguiente dirección <http://minia.genouest.org/> se podrá descargar el paquete que contiene el programa.

Una vez descargado se desempaquetará el programa con el siguiente comando mostrado en la imagen.



```
vane@vane-VirtualBox: ~/Descargas/minia-v2.0.7-bin-Linux
vane@vane-VirtualBox:~/Descargas/minia-v2.0.7-bin-Linux$ sudo apt-get install minia
[sudo] password for vane:
```

Una vez desempquetado, vamos a compilar Minia con el siguiente comando visto en la imagen, donde “**sudo**” pide permiso administrador por lo que es importante poner la contraseña del usuario de sesión.

```
vane@vane-VirtualBox: ~/Descargas/minia-v2.0.7-bin-Linux
vane@vane-VirtualBox:~/Descargas/minia-v2.0.7-bin-Linux$ minia
usage:
 minia input_file kmer_size min_abundance estimated_genome_size prefix
hints:
 min_abundance ~ 3
 estimated_genome_size is in bp, does not need to be accurate, only controls memory usage
 prefix is any name you want the results to start with
vane@vane-VirtualBox:~/Descargas/minia-v2.0.7-bin-Linux$
```

Una vez realizada la instalación se digitará en la terminal “**minia**” y el nos mostrará , una pequeña descripción de uso de la herramienta.

# KMERGENIE

---

Es una herramienta complementaria de Minia que permite realizar predicciones de genomas con la longitud del tamaño del genoma que se necesite denominado  $K$ , también permite optimizar más fácilmente los parámetros  $K$  utilizados y generar histogramas de abundancia que permite al usuario tomar decisiones.

```
vane@vane-VirtualBox: ~/Descargas
vane@vane-VirtualBox:~/Descargas$ tar -xzvf kmergenie-1.7016.tar.gz
kmergenie-1.7016/
kmergenie-1.7016/third_party/
kmergenie-1.7016/third_party/docopt.pyc
kmergenie-1.7016/third_party/__init__.pyc
kmergenie-1.7016/third_party/docopt.py
kmergenie-1.7016/third_party/__init__.py
kmergenie-1.7016/minia/
kmergenie-1.7016/minia/MultiConsumer.cpp
kmergenie-1.7016/minia/LargeInt.cpp
kmergenie-1.7016/minia/Pool.cpp
kmergenie-1.7016/minia/Hashing.h
kmergenie-1.7016/minia/Kmer.h
kmergenie-1.7016/minia/Utils.cpp
kmergenie-1.7016/minia/Hash16.h
kmergenie-1.7016/minia/Utils.h
kmergenie-1.7016/minia/LinearCounter.h
kmergenie-1.7016/minia/Kmer.cpp
kmergenie-1.7016/minia/Bank.h
kmergenie-1.7016/minia/Pool.h
kmergenie-1.7016/minia/Bloom.cpp
kmergenie-1.7016/minia/LinearCounter.cpp
kmergenie-1.7016/minia/Hash16.cpp
kmergenie-1.7016/minia/Bank.cpp
```

En la siguiente dirección se podrá descargar la herramienta <http://kmergenie.bx.psu.edu/>.

Se navegará a la carpeta donde esta ubicada la herramienta con el comando “**cd**”.y se descomprime para poder compilar KmerGenie.

```
vane@vane-VirtualBox: ~/Descargas/kmergenie-1.7016
vane@vane-VirtualBox:~/Descargas/kmergenie-1.7016$ make
g++ -o minia/Pool.o -c minia/Pool.cpp -O4 -pthread -D_largeint -DKMER_PRECISION=4
g++ -o minia/Bank.o -c minia/Bank.cpp -O4 -pthread -D_largeint -DKMER_PRECISION=4
minia/Bank.cpp: In member function 'void Bank::init(char**, int)':
minia/Bank.cpp:279:44: warning: ignoring return value of 'char* strerror_r(int, char*, si
ze_t)', declared with attribute warn_unused_result [-Wunused-result]
    strerror_r( errno, buffer, BUFSIZ ); // get string message from errno
    ^
minia/Bank.cpp: In member function 'void BinaryBank::open(bool)':
minia/Bank.cpp:717:44: warning: ignoring return value of 'char* strerror_r(int, char*, si
ze_t)', declared with attribute warn_unused_result [-Wunused-result]
    strerror_r( errno, buffer, BUFSIZ ); // get string message from errno
    ^
minia/Bank.cpp: In member function 'void BinaryReads::open(bool)':
minia/Bank.cpp:802:44: warning: ignoring return value of 'char* strerror_r(int, char*, si
ze_t)', declared with attribute warn_unused_result [-Wunused-result]
    strerror_r( errno, buffer, BUFSIZ ); // get string message from errno
    ^
minia/Bank.cpp: In member function 'int KmersBuffer::readkmers()':
minia/Bank.cpp:992:65: warning: ignoring return value of 'size_t fread(void*, size_t, siz
e_t, FILE*)', declared with attribute warn_unused_result [-Wunused-result]
    fread(buffer,sizeof( char),block_size, binary_read_file); // read a block of seq
    ^
g++ -o specialk minia/Pool.o minia/Bank.o minia/Hash16.o minia/Bloom.o minia/Kmer.o minia
/Uutils.o minia/LinearCounter.o minia/LargeInt.o minia/MultiConsumer.o minia/Hashing.o mai
n.cpp -O4 -pthread -D_largeint -DKMER_PRECISION=4 -lz -DSVN_REV=1.7016
scripts/test_install
Testing presence of specialk...
OK
Testing presence of Rscript...
R scripting front-end version 3.2.3 (2015-12-10)
OK
Testing basic Rscript functionality...
Rscript --no-init-file -e 'rnorm(1)'
[1] "rnorm(1)"
OK
Testing a simple KmerGenie example...

OK
Test successful. KmerGenie is ready, type `./kmergenie`.
```

Una vez descomprimido, se ingresa a la carpeta de Kmergenie con el comando “**cd**”, se coloca el comando “**make**” que es el que compilará la herramienta y se deberá ver como en la pantalla.

Al digitar en la terminal “./kmergenie” se despliega diferentes opciones de uso de la herramienta kmergenie. Y estará lista para funcionar.

```
vane@vane-VirtualBox: ~/Descargas/kmergenie-1.7016
vane@vane-VirtualBox:~/Descargas/kmergenie-1.7016$ ./kmergenie
KmerGenie

Usage:
  kmergenie <read_file> [options]

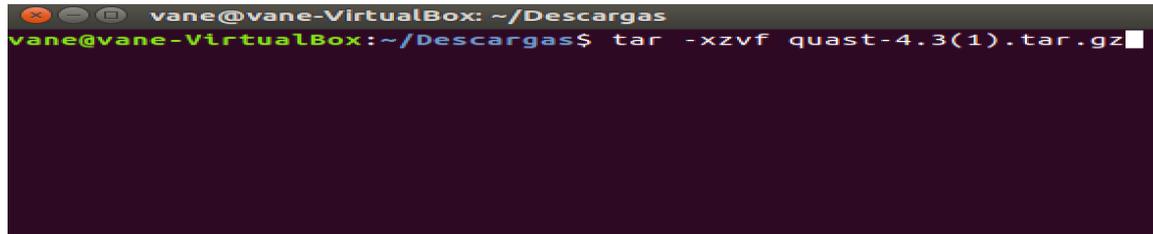
Options:
  --diploid      use the diploid model (default: haploid model)
  --one-pass     skip the second pass to estimate k at 2 bp resolution (default: two passes)
  -k <value>    largest k-mer size to consider (default: 121)
  -l <value>    smallest k-mer size to consider (default: 15)
  -s <value>    interval between consecutive kmer sizes (default: 10)
  -e <value>    k-mer sampling value (default: auto-detected to use ~200 MB memory/thread)
  -t <value>    number of threads (default: number of cores minus one)
  -o <prefix>   prefix of the output files (default: histograms)
  --debug       developer output of R scripts
vane@vane-VirtualBox:~/Descargas/kmergenie-1.7016$
```

# QUAST

---

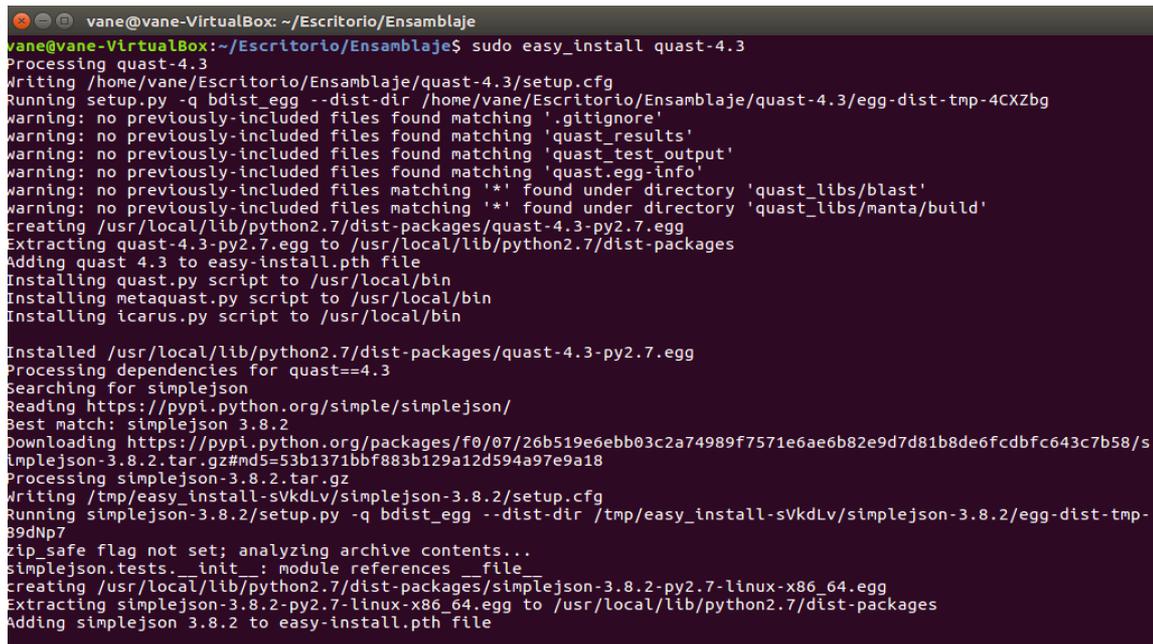
QUAST es una herramienta para evaluar y comparar el ensamblaje del genomas ya sea con o sin genomas de referencia. QUAST produce tablas y gráficos interactivos y fácil de interpretar.

En la siguiente dirección <http://bioinf.spbau.ru/quast> se descargará el paquete que contiene el programa y se descomprime como se ve en la imagen.



```
vane@vane-VirtualBox: ~/Descargas
vane@vane-VirtualBox:~/Descargas$ tar -xzf quast-4.3(1).tar.gz
```

Se ingresa con el comando “**cd**” a la carpeta QUAST y se escribirá en la terminal “**./setup.py install**” que es la instalación básica, o ya bien se podría la instalación completa “**./setup.py install\_full**” pero hay que tener presente que esta instalación completa pesa cerca de 800MB por lo que deben asegurarse de que tienen la memoria suficiente.



```
vane@vane-VirtualBox: ~/Escritorio/Ensamblaje
vane@vane-VirtualBox:~/Escritorio/Ensamblaje$ sudo easy_install quast-4.3
Processing quast-4.3
Writing /home/vane/Escritorio/Ensamblaje/quast-4.3/setup.cfg
Running setup.py -q bdist_egg --dist-dir /home/vane/Escritorio/Ensamblaje/quast-4.3/egg-dist-tmp-4CXZbg
warning: no previously-included files found matching '.gitignore'
warning: no previously-included files found matching 'quast_results'
warning: no previously-included files found matching 'quast_test_output'
warning: no previously-included files found matching 'quast.egg-info'
warning: no previously-included files matching '*' found under directory 'quast_libs/blast'
warning: no previously-included files matching '*' found under directory 'quast_libs/manta/build'
creating /usr/local/lib/python2.7/dist-packages/quast-4.3-py2.7.egg
Extracting quast-4.3-py2.7.egg to /usr/local/lib/python2.7/dist-packages
Adding quast 4.3 to easy-install.pth file
Installing quast.py script to /usr/local/bin
Installing metaquast.py script to /usr/local/bin
Installing icarus.py script to /usr/local/bin

Installed /usr/local/lib/python2.7/dist-packages/quast-4.3-py2.7.egg
Processing dependencies for quast==4.3
searching for simplejson
Reading https://pypi.python.org/simple/simplejson/
Best match: simplejson 3.8.2
Downloading https://pypi.python.org/packages/f0/07/26b519e6ebb03c2a74989f7571e6ae6b82e9d7d81b8de6fcd9c643c7b58/s
implejson-3.8.2.tar.gz#md5=53b1371bbf883b129a12d594a97e9a18
Processing simplejson-3.8.2.tar.gz
Writing /tmp/easy_install-sVkdLv/simplejson-3.8.2/setup.cfg
Running simplejson-3.8.2/setup.py -q bdist_egg --dist-dir /tmp/easy_install-sVkdLv/simplejson-3.8.2/egg-dist-tmp-
89dNp7
zip_safe flag not set; analyzing archive contents...
simplejson.tests._init : module references __file__
creating /usr/local/lib/python2.7/dist-packages/simplejson-3.8.2-py2.7-linux-x86_64.egg
Extracting simplejson-3.8.2-py2.7-linux-x86_64.egg to /usr/local/lib/python2.7/dist-packages
Adding simplejson 3.8.2 to easy-install.pth file
```

# SSPACE

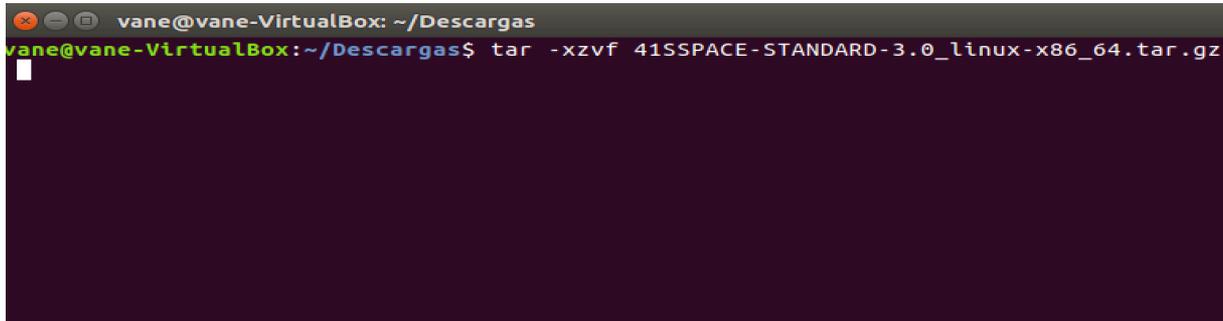
---

Es un programa para scaffold (super conjuntos de clones pre ensamblados) que mejora el tiempo de ejecución y muestra resultados prometedores donde la cantidad de conjuntos de clones iniciales se reducen considerablemente.

Sspace es capaz de evaluar el orden, la distancia y la orientación de los conjuntos de clones y combinar los super conjunto de clones, además utiliza formato FASTA y NGS (secuenciación de próxima generación).

<http://www.baseclear.com/genomics/bioinformatics/basetools/SSPACE> es la página oficial para descargar SSPACE hay que llenar un pequeño formulario, una vez completando se podrá descargar el paquete.

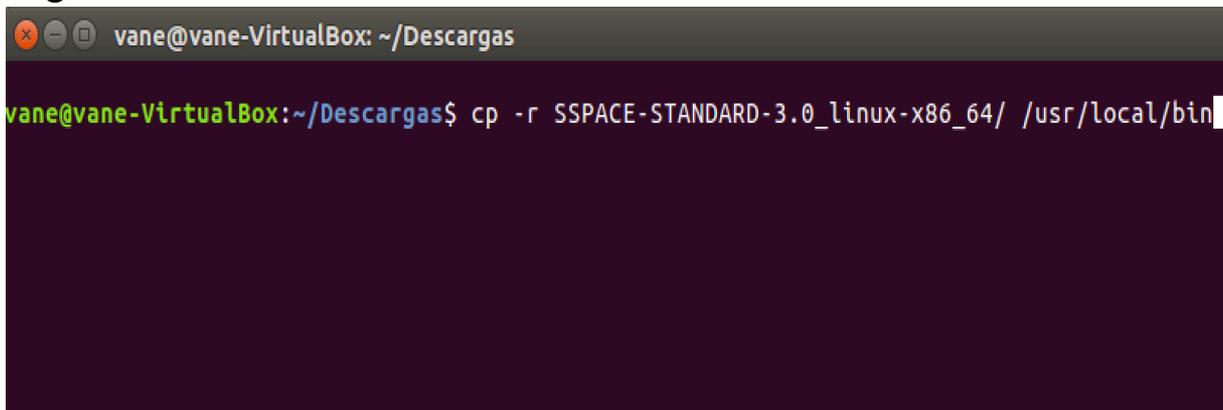
Como se ha venido realizando se desempaquetará el programa con el comando "**tar -xzvf**".



```
vane@vane-VirtualBox: ~/Descargas
vane@vane-VirtualBox:~/Descargas$ tar -xzvf 41SSPACE-STANDARD-3.0_linux-x86_64.tar.gz
```

Esta herramienta necesita una librería que en algunos casos no se encuentra instalada por lo que se recomienda ejecutar el siguiente comando "**sudo cpan Perl4::CoreLibs**" antes de iniciar la instalación de SSPACE.

Ahora, se utiliza el comando "**cp -r**" que servirá para copiar la carpeta SSPACE a /usr/local/bin. El sistema de archivos /usr es donde normalmente se encuentran instalados todos los programas.



```
vane@vane-VirtualBox: ~/Descargas
vane@vane-VirtualBox:~/Descargas$ cp -r SSPACE-STANDARD-3.0_linux-x86_64/ /usr/local/bin
```

```
vane@vane-VirtualBox: /usr/local/bin
vane@vane-VirtualBox:~/Descargas$ cd /usr/local/bin
vane@vane-VirtualBox:/usr/local/bin$ ls
config_data icarus.py metaquast.py quast.py SSPACE SSPACE-STANDARD-3.0_linux-x86_64
vane@vane-VirtualBox:/usr/local/bin$ ln -s SSPACE-STANDARD-3.0_linux-x86_64/ SSPACE
```

El primer comando “**cd**” navegará a la carpeta donde se encuentra SSPACE. “**ln -s**” realizará una versión independiente de la carpeta original llamada “SSPACE-STANDARD-3.0\_linux-x86\_64” y la llamé SSPACE.

```
vane@vane-VirtualBox: /usr/local/bin/SSPACE
vane@vane-VirtualBox:/usr/local/bin$ cd SSPACE
vane@vane-VirtualBox:/usr/local/bin/SSPACE$ ls
BaseTools License agreement_2014-03-31.pdf F132-03 SSPACE Standard User manual v3.0.pdf
bin F132-04 SSPACE Standard Tutorial v3.0.pdf
bowtie README
bwa SSPACE_Standard_v3.0.pl
dotlib tools
example
vane@vane-VirtualBox:/usr/local/bin/SSPACE$ ./SSPACE_Standard_v3.0.pl
```

Con el comando “**cd**” se ingresó a la nueva carpeta llamada SSPACE, y con “**ls**” se desplegará el listado con todos los paquetes e información que contiene SSPACE.

El comando “**SSPACE\_Standard\_v3.0.pl**” ejecutará el programa y deberá aparecer una pantalla muy similar a la que se muestra.

```
vane@vane-VirtualBox: /usr/local/bin/SSPACE
===== General Parameters =====
-l Library file containing two mate mate files with insert size, error and either mate p
air or paired end indication.
-s Fasta file containing contig sequences used for extension. Inserted pairs are mapped
to extended and non-extended contigs (REQUIRED)
-x Indicate whether to extend the contigs of -s using paired reads in -l. (-x 1=extensio
n, -x 0=no extension, default -x 0)
===== Extension Parameters =====
-m Minimum number of overlapping bases with the seed/contig during overhang consensus bu
ild up (default -m 32)
-o Minimum number of reads needed to call a base during an extension (default -o 20)
===== Scaffolding Parameters =====
-z Minimum contig length used for scaffolding. Filters out contigs that are below -z (de
fault -z 0 (no filtering), optional).
-k Minimum number of links (read pairs) to compute scaffold (default -k 5, optional)
-a Maximum link ratio between two best contig pairs *higher values lead to least accurat
e scaffolding* (default -a 0.7, optional)
-n Minimum overlap required between contigs to merge adjacent contigs in a scaffold (def
ault -n 15, optional)
===== Bowtie Parameters =====
-g Maximum number of allowed gaps during mapping with Bowtie. Corresponds to the -v opti
on in Bowtie. *higher number of allowed gaps can lead to least accurate scaffolding* (def
ault -v 0, optional)
===== Additional Parameters =====
-T Specify the number of threads to run SSPACE, used both for reading the input readfile
s and mapping the reads against the contigs. For reading in the files, multiple files are
read-in simultaneously. With read-mapping, the readmapper is called multiple times with
1 million reads per calls (default -T 1, optional)
-S Skip the processing of the reads. Meaning that SSPACE was already run, but user now w
ants to use different extension/scaffold parameters.
-b Base name for your output files (optional)
-v Runs the scaffolding process in verbose mode (-v 1=yes, -v 0=no, default -v 0, option
```

Las herramientas están listas para funcionar.!