

Software & Requirements

Javier Perales-Patón
jperales@cniio.es



Translational Bioinformatics Unit
CNIO. Madrid, Spain.

Fátima Al-Shahrour
[\[falshahrour@cniio.es\]](mailto:falshahrour@cniio.es)
Elena Piñeiro-Yáñez
[\[epineiro@cniio.es\]](mailto:epineiro@cniio.es)
Pedro Fernandes
[\[pfern@igc.gulbenkian.pt\]](mailto:pfern@igc.gulbenkian.pt)



Computational Facilities

- The **genome reference** must be **indexed** only once in your computer, but it takes a long time (2-3 days). The sizes of the indexes are huge (~100 Gb). See “Before running a variant analysis for first time” at:
<http://rubioseq.bioinfo.cnio.es/sites/default/files/PDF/RUbioSeq-book.pdf>
- You need a huge **hard disk capacity**:
 - **Human genome reference** (including indexes) : 150 Gb.
 - **Bundle of files for Variant Calling Analysis** (GATK) : 14 Gb.
 - **For each sample**:
 - Raw data : 7 - 15 Gb.
 - Intermediate files for each sample : 15 - 20 Gb (whole-exome seq).

WorkStation:

- **Minimal requirements**:
 - 16 Gb RAM
 - 500 Gb hard disk
 - 8 threads ~ 8 cores.
- **Recommended requirements (multi-sample analysis, samples storage)**:
 - 25 Gb RAM
 - 1 Tb hard disk (e.g. Exome-seq analysis from 50 samples → 1.5 Tb).
 - 16 threads.

Installing RUBioSeq+ in your computer



<http://rubioseq.bioinfo.cnio.es/>

Download and Run RUBioSeq+

RUBioSeq+ LiveDVD

RUBioSeq+ sources

RUBioSeq+ Docker



LINUX



LINUX

(Virtual machine, Ubuntu system)

1. First, install Docker client (depending on your Operating System):

Follow the corresponding guidelines:

- Ubuntu (<https://docs.docker.com/installation/ubuntu/linux/>)
- Mac OS X (<https://docs.docker.com/installation/mac/>)
- Windows (<https://docs.docker.com/installation/windows/>)

2. Second, launch the Docker Image (it will download it from the Internet):

- Windows:
(Start Menu): Program Files > Boot2Docker

- Unix (in terminal):

```
docker run -ti -p 0.0.0.0:8080:8080 --name RUBioSeq ubio/rubioseq:latest /bin/bash
```

3. Try it out:

```
bash /opt/RUBioSeq3.7/RUBioSeq.pl --help
```

More info: http://rubioseq.bioinfo.cnio.es/rubioseq_docker

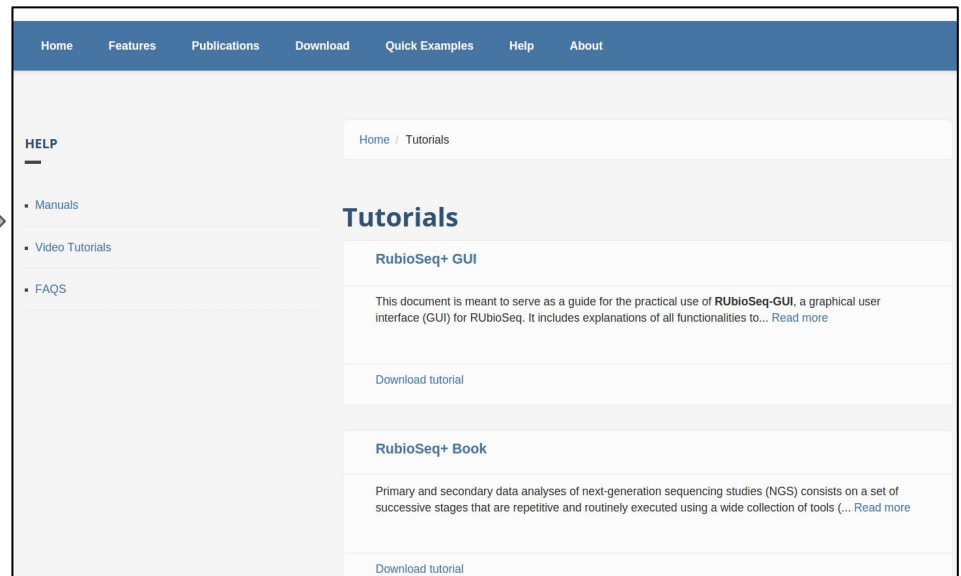
Ask for some help

RUBioSeq is under **development** by the Bioinformatics Unit at CNIO (Madrid, Spain):

- Miriam Rubio-Camarillo (mrubioc@cnio.es).
- José María Fernández (jmfernandez@cnio.es).
- Gonzalo Gómez-López (ggomez@cnio.es).

<http://rubioseq.bioinfo.cnio.es/tutorials>

- Video-tutorials.
- Example exercises.



[RUBioSeq: a suite of parallelized pipelines to automate exome variation and bisulfite-seq analyses](#)

Miriam Rubio-Camarillo*, Gonzalo Gómez-López, José M. Fernández, Alfonso Valencia and David G. Pisano.

2013, Bioinformatics, 29 (13): 1687-1689

GATK's bundle from FTP repository (14 Gb)



ftp://ftp.broadinstitute.org/bundle/2.8/hg19/

Index of /bundle/2.8/hg19/

Name	Size	Date Modified
[parent directory]		
1000G_omni2.5.hg19.sites.vcf.gz	49.4 MB	12/8/13, 1:00:00 AM
1000G_omni2.5.hg19.sites.vcf.gz.md5	97 B	12/8/13, 1:00:00 AM
1000G_omni2.5.hg19.sites.vcf.idx.gz	464 kB	12/8/13, 1:00:00 AM
1000G_omni2.5.hg19.sites.vcf.idx.gz.md5	101 B	12/8/13, 1:00:00 AM
1000G_phase1.indels.hg19.sites.vcf.gz	42.9 MB	12/8/13, 1:00:00 AM
1000G_phase1.indels.hg19.sites.vcf.gz.md5	103 B	12/8/13, 1:00:00 AM
1000G_phase1.indels.hg19.sites.vcf.idx.gz	326 kB	12/8/13, 1:00:00 AM
1000G_phase1.indels.hg19.sites.vcf.idx.gz.md5	107 B	12/8/13, 1:00:00 AM
1000G_phase1.snps.high_confidence.hg19.sites.vcf.gz	1.7 GB	12/8/13, 1:00:00 AM
1000G_phase1.snps.high_confidence.hg19.sites.vcf.gz.md5	117 B	12/8/13, 1:00:00 AM
1000G_phase1.snps.high_confidence.hg19.sites.vcf.idx.gz	3.4 MB	12/8/13, 1:00:00 AM
1000G_phase1.snps.high_confidence.hg19.sites.vcf.idx.gz.md5	121 B	12/8/13, 1:00:00 AM
CEUTrio.HiSeq.WGS.b37.bestPractices.hg19.vcf.gz	407 MB	12/8/13, 1:00:00 AM
CEUTrio.HiSeq.WGS.b37.bestPractices.hg19.vcf.gz.md5	119 B	12/8/13, 1:00:00 AM
CEUTrio.HiSeq.WGS.b37.bestPractices.hg19.vcf.idx.gz	3.2 MB	12/8/13, 1:00:00 AM
CEUTrio.HiSeq.WGS.b37.bestPractices.hg19.vcf.idx.gz.md5	123 B	12/8/13, 1:00:00 AM
Mills_and_1000G_gold_standard.indels.hg19.sites.vcf.gz	19.1 MB	12/8/13, 1:00:00 AM
Mills_and_1000G_gold_standard.indels.hg19.sites.vcf.gz.md5	120 B	12/8/13, 1:00:00 AM
Mills_and_1000G_gold_standard.indels.hg19.sites.vcf.idx.gz	426 kB	12/8/13, 1:00:00 AM
Mills_and_1000G_gold_standard.indels.hg19.sites.vcf.idx.gz.md5	124 B	12/8/13, 1:00:00 AM
dbSNP_138.hg19.excluding_sites_after_129.vcf.gz	334 MB	12/8/13, 1:00:00 AM
dbSNP_138.hg19.excluding_sites_after_129.vcf.gz.md5	119 B	12/8/13, 1:00:00 AM
dbSNP_138.hg19.excluding_sites_after_129.vcf.idx.gz	3.6 MB	12/8/13, 1:00:00 AM
dbSNP_138.hg19.excluding_sites_after_129.vcf.idx.gz.md5	123 B	12/8/13, 1:00:00 AM
dbSNP_138.hg19.vcf.gz	1.4 GB	12/8/13, 1:00:00 AM
dbSNP_138.hg19.vcf.gz.md5	93 B	12/8/13, 1:00:00 AM
dbSNP_138.hg19.vcf.idx.gz	3.8 MB	12/8/13, 1:00:00 AM
dbSNP_138.hg19.vcf.idx.gz.md5	97 B	12/8/13, 1:00:00 AM
hapmap_3.3.hg19.sites.vcf.gz	58.0 MB	12/8/13, 1:00:00 AM
hapmap_3.3.hg19.sites.vcf.gz.md5	94 B	12/8/13, 1:00:00 AM
hapmap_3.3.hg19.sites.vcf.idx.gz	807 kB	12/8/13, 1:00:00 AM
hapmap_3.3.hg19.sites.vcf.idx.gz.md5	98 B	12/8/13, 1:00:00 AM

You can download the different bundles from GATK's FTP (Broad Institute) visiting this URL with your Internet Browser:

<ftp://gsapubftp-anonymous@ftp.broadinstitute.org/bundle/2.8/>

1. **Genome Reference** (standard 1000 Genomes, fasta).
2. List of **Target beats or intervals** of genomic regions sequenced by the Library protocol.
3. **dbSNP** (VCF file) for a recent dbSNP release (build 138, it includes the 1000 Genomes).
4. HapMap genotypes and sites VCFs
5. **OMNI 2.5 genotypes for 1000 Genomes samples** (VCF).
6. The current best set of **known indels** to be used for local realignment); use both files:
 - 1000G_phase1.indels.b37.vcf (currently from the 1000 Genomes Phase I indel calls)
 - Mills_and_1000G_gold_standard.indels.b37.sites.vcf

 Tip for home: The following UNIX command downloads the whole bundle for **hg19** in one step (~hrs) :

```
$ wget -r -nH --cut-dirs=3 --reject-regex "NA12878|CEUTrio" \  
-P /path/to/your_directory/ ftp://gsapubftp-anonymous@ftp.broadinstitute.org/bundle/2.8/hg19/*
```

Dependencies (Ubuntu system)



1. Install **java 1.7** (<http://www.webupd8.org/2012/01/install-oracle-java-jdk-7-in-ubuntu-via.html>):

```
sudo add-apt-repository ppa:webupd8team/java
sudo apt-get update
sudo apt-get install oracle-java7-installer
```

2. Install **python**, then the following **python** modules:

```
sudo apt-get install python-dev
sudo pip install Numpy
sudo pip install Scipy
```

3. Install **perl** and these **perl** modules:

(For Bioperl, see “Installing using CPAN” at http://www.bioperl.org/wiki/Installing_BioPerl_on_Unix)

```
sudo cpan DBI DBD::mysql XML::LibXML Carp FindBin File::Basename
sudo cpan File::Spec File::Copy Getopt::long Class::Inspector
```

4. Library **libmysqlclient**:

(For Bioperl, see “Installing using CPAN” at http://www.bioperl.org/wiki/Installing_BioPerl_on_Unix)

```
sudo apt-get install libmysqlclient-dev
```

Download the installer:

1. Open  **Firefox**, and visit the RUBioSeq's Home page (www.rubioseq.bioinfo.cnio.es):

2. Click on **Download** from the menu options.

3. Navigate through the directory structure and download the automatic installer following:
Installation_bungle > RUBioSeq+ > rubioseq-bundle-installer.sh

Home

Name	Modified
Installation_bundle	2015-01-14
SourceVersions	2014-12-17
LiveDVDs	2015-12-17
LatestManual	2014-12-17
HG19_bundle	2014-12-05
Required_programs	2014-12-05
Tests_Data	2014-12-05

Totals: 7 Items

Home / Installation_bundle

Name	Modified
Parent folder	
RUBioSeq+	2015-02-19

Totals: 1 Item

Home / Installation_bundle / RUBioSeq+

Name	Modified	Size	Downloads / Week
Parent folder			
rubioseq-bundle-installer.sh	2015-02-19	45.4 MB	1

Totals: 1 Item 45.4 MB 1

4. Select **Save File**.

Locate your Home.

Create Folder in your Home called **Software**
(/home/USERNAME/Software)

Save the file in said directory.

Enter name of file to save to...

Name: rubioseq-bundle-installer.sh

Save in folder: jperales Soft Create Folder


Places: Search Recently Used jperales Desktop File System

Name Size Modified

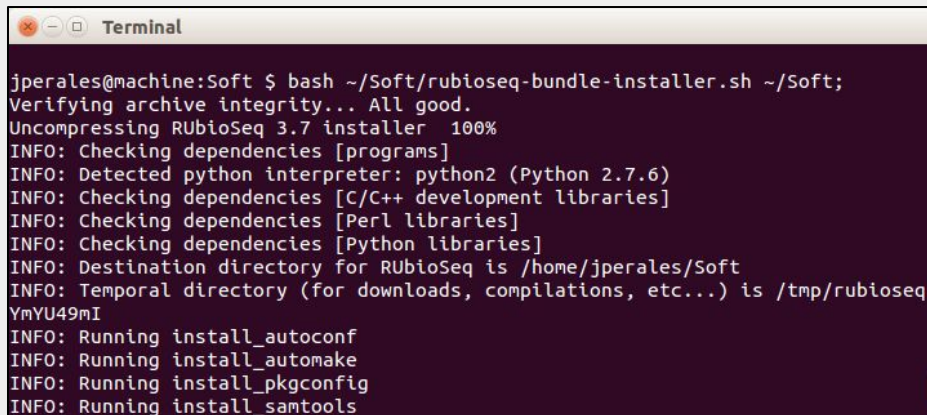
shell script

Cancel Save

Run the installer:

1. Open a terminal 
2. Execute a bash **command**:

```
bash /home/YOUR_USER/Software/rubioseq-bundle-installer.sh  
/home/YOUR_USER/Software/;
```



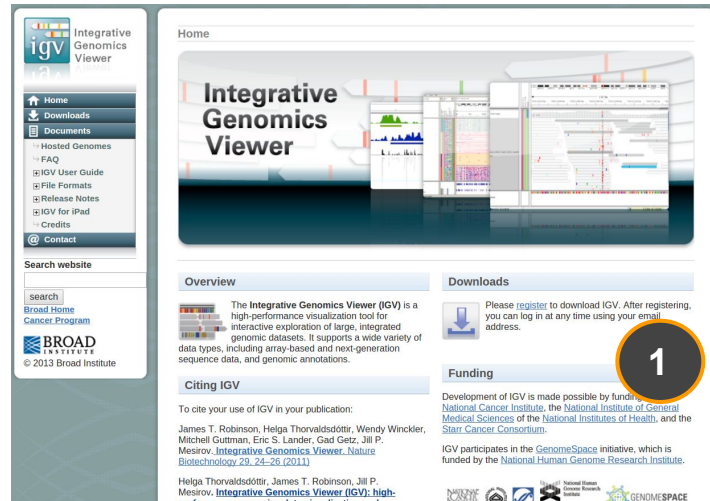
```
Terminal  
jperales@machine:Soft $ bash ~/Soft/rubioseq-bundle-installer.sh ~/Soft;  
Verifying archive integrity... All good.  
Uncompressing RUBioSeq 3.7 installer 100%  
INFO: Checking dependencies [programs]  
INFO: Detected python interpreter: python2 (Python 2.7.6)  
INFO: Checking dependencies [C/C++ development libraries]  
INFO: Checking dependencies [Perl libraries]  
INFO: Checking dependencies [Python libraries]  
INFO: Destination directory for RUBioSeq is /home/jperales/Soft  
INFO: Temporal directory (for downloads, compilations, etc...) is /tmp/rubioseq.  
YmYU49mI  
INFO: Running install_autoconf  
INFO: Running install_automake  
INFO: Running install_pkgconfig  
INFO: Running install_samtools
```

This will also configure the **PATHS** to the programs.

3. When it finishes, you can try it out using:

```
bash /home/YOUR_USER/Software/RUBioSeq3.7/RUBioSeq.pl --help
```

Integrative Genomics Viewer (IGV)



Open Firefox.

<https://www.broadinstitute.org/igv/>

1. Download section: Register and Fill out the form.

2. Download the Binary distribution: file.zip
Alternatively, you can use this link:

http://data.broadinstitute.org/igv/projects/downloads/IGV_2.3.66.zip

3. Binary Distribution

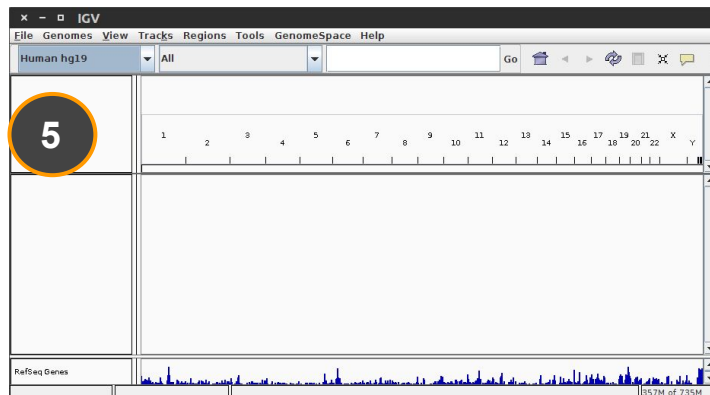
Download and unzip the binary distribution archive in a folder of your choosing. IGV is launched from a command prompt -- follow instructions in the "readme" file. To launch igv on Mac or Linux platforms use the shell script "igv.sh". On Windows use "igv.bat".

Download
Binary Distribution

2

3. Extract the Zip file.

4. Go to the new Directory. Click on igv.jar



IGV_2.3.66.zip



IGV_2.3.66



4



