

How to use getSequenceInfo (English version)

This is a user manual showing students and researchers how to use the getSequenceInfo software tool (<https://github.com/dcouvin/getSequenceInfo>) provided as both a command line interface (CLI) and a graphical user interface (GUI).

Keywords: Bioinformatics, DNA sequence, GenBank, RefSeq, European Nucleotide Archive (ENA), Perl programming language, BioPerl, Annotation, FASTA, FASTQ, GC-content, Statistics, NucleScore

[Introduction](#)

[Installation](#)

[Download or clone getSequenceInfo](#)

[Launching the installation on Unix or Windows systems](#)

[Content of the getSequenceInfo archive](#)

[How to use the tool](#)

[Examples of use \(GUI\)](#)

[Some command lines](#)

Introduction

getSequenceInfo is a Perl script allowing to easily download sequence data from public repositories such as the NCBI's GenBank and RefSeq, as well as the European Nucleotide Archive or ENA (hosted at the European Bioinformatics Institute). This Perl software programme allows users to download sequence data and statistics in various formats (FASTA, FASTQ, Genbank full format, XLS, TSV, HTML). The getSequenceInfo software tool can either be used as a command line programme or as a graphical user interface (GUI).

Installation

Perl is usually already installed on Unix systems (such as Linux and MacOS). However, for people using the Windows operating system, the language can be installed with Strawberry Perl (<http://strawberryperl.com/>).

[This video](#) shows you how to install Strawberry Perl:

If necessary, please see information on [how to launch](#) or [how to use](#) the Command Prompt in Windows. When using a Unix OS (Linux or Mac), Perl is generally already installed. But if it is not the case, you can see [this page](#) for its installation. You can follow this [wiki page](#) for information about the Shell Prompt. You can then check the installation by typing the following command:

```
perl -v
```

Users can then install required Perl modules (please note that these modules are already available within the provided installation files “**installer_Unix.sh**” and “**installer_Windows.bat**”):

- Tk
- BioPerl
- Date::Calc
- Bio::SeqIO
- LWP::Simple
- Data::Dumper
- IO::Uncompress::Gunzip
- IO::File
- Getopt::Long
- Net::FTP

Each Perl module can be installed using the **cpan** or the **cpanm** command as follows:

- `cpan -f -i <Module::Name>`
- `cpanm <Module::Name>`

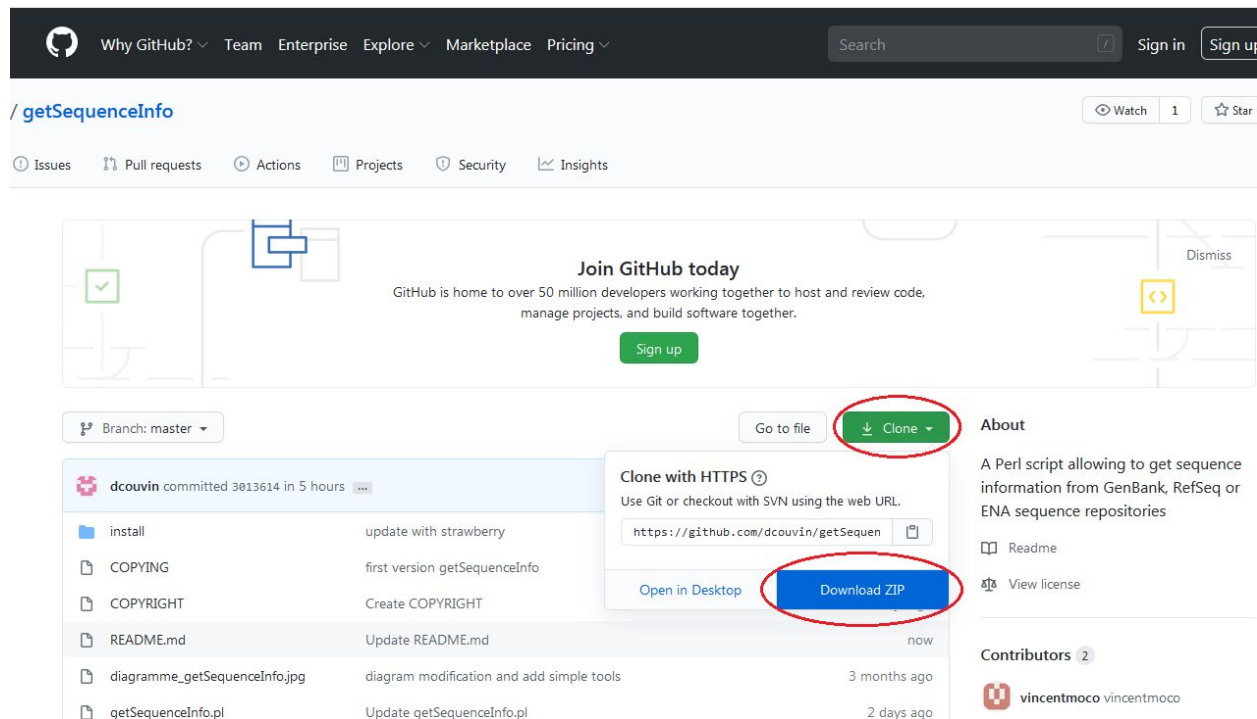
Please note that <Module::Name> should be replaced by the required Perl module (e.g. `cpan -f -i Date::Calc` or `cpanm Date::Calc`).

Further details concerning the installation of Perl modules are available at the webpage [How to install CPAN modules](#).

Download or clone getSequenceInfo

The tool can be downloaded or cloned from the git repository (<https://github.com/dcouvin/getSequenceInfo>).

The “Clone” button then the “Download ZIP” button can be used to download the repository:



Snapshot of the GitHub page allowing you to download the tool.

Otherwise, you can clone the tool using the “git clone” command. Please note that git (<https://git-scm.com/>) must be installed in your system to do this:

```
git clone https://github.com/dcouvin/getSequenceInfo.git
```

Once the archive has been cloned or downloaded, then unzipped in the place of your choice, you can go to the getSequenceInfo repository by typing the following command:

```
cd getSequenceInfo
```

Users can also navigate through classic windows to access the tool.

Launching the installation on Unix or Windows systems

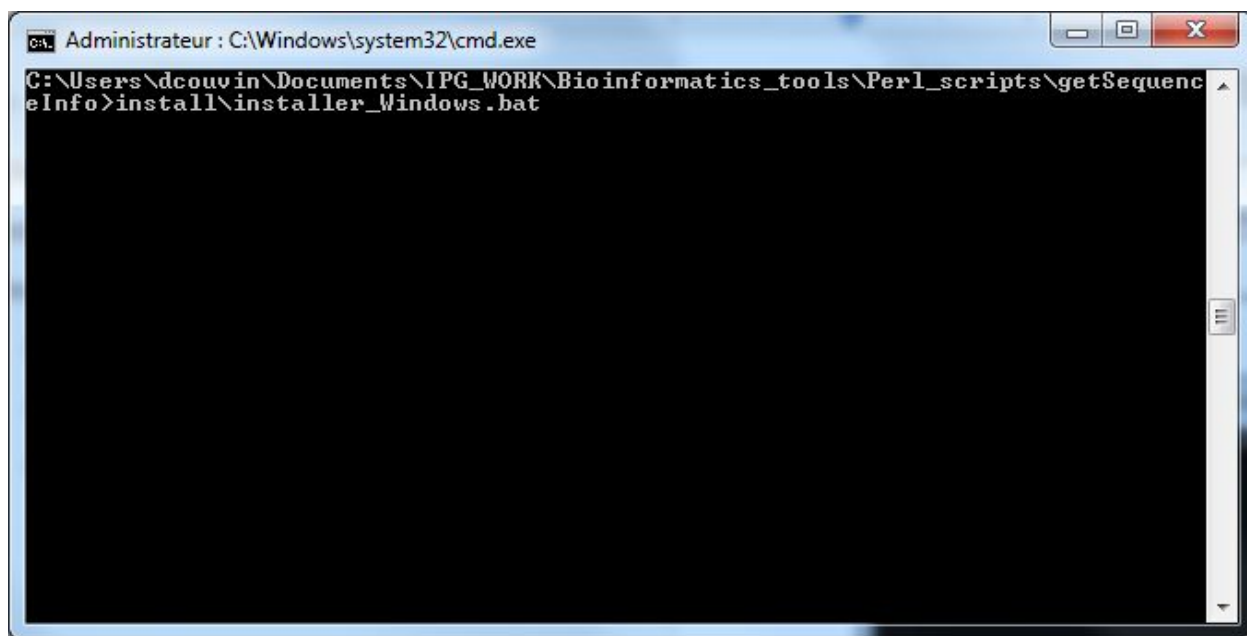
In order to install the getSequenceInfo software tool, the user must be placed into the corresponding uncompressed archive repository (getSequenceInfo).

Installation on Linux or MacOS (Unix system) Shell Prompt

```
bash install/installer_Unix.sh
```

Installation on Microsoft Windows system Command Prompt. Users can also install the tool by running the installer_Windows.bat file (double-click)

```
install\installer_Windows.bat
```



Snapshot of the Windows installation using the command line.

Installation instructions can also be provided when running the getSequenceInfo.pl Perl script.

Content of the getSequenceInfo archive

The archive contains the following files and folders:

- **install** (a folder containing the installation files “**installer_Unix.sh**” and “**installer_Windows.bat**”)
- **simple_tools** (a folder containing other Perl scripts dedicated to the execution of specific tasks)
- **COPYING** and **COPYRIGHT** (GPLv3 licence files)
- **README.md** (a simplified README file)
- **workflow.png** (a diagram representing the main functionalities of the software tool)
- **getSequenceInfo.pl** (main Perl program)
- **getSequenceInfoGUI.pl** (Graphical User Interface (GUI) version of the program)
- **launcher_Windows.bat** (executable file allowing to launch the GUI from Windows)
- **logo_getSequenceInfo.png** (a logo representing the tool)
- **User_manual.pdf** (this user manual)

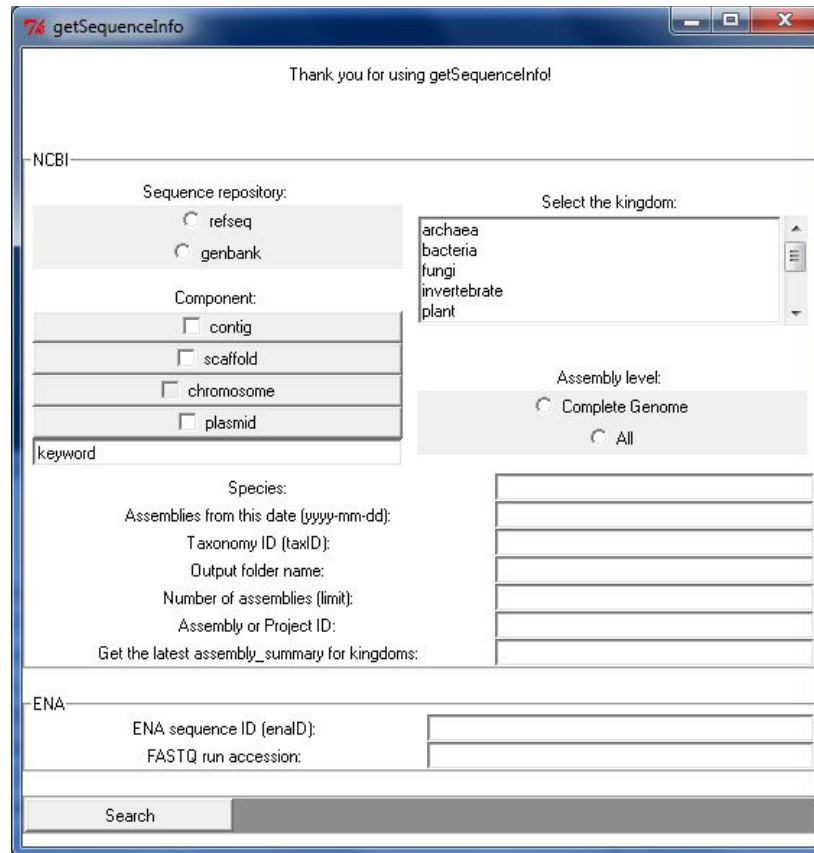
How to use the tool

As mentioned earlier, the tool can be used from the command line or with a graphical user interface (GUI). The same usage semantics are shared between the two interfaces. Please note that both Perl files (`getSequenceInfo.pl` and `getSequenceInfoGUI.pl`) should be localized in the same directory.

The user can launch the GUI version of the tool (`getSequenceInfoGUI.pl`) either by executing it (double click) or by typing the following command:

```
perl getSequenceInfoGUI.pl
```

The following window should appear in front of a command prompt:



Snapshot of the GUI window.

Examples of use (GUI)

There are many ways to use this tool to download assemblies and their related information.

The simplest way is to use the name of the species you want to search and its associated kingdom.

Suppose we want to search for information about the coronavirus.

We can use the following command in order to retrieve 12 genome assemblies available from the date 2019/12/01:

```
perl getSequenceInfo.pl -s coronavirus -k viral -date 2019-12-01 -n 12
```

getSequenceInfo

Thank you for using getSequenceInfo!

NCBI

Sequence repository:

- ☒ refseq
- ☒ genbank

Component:

- ☐ plasmid
- ☐ scaffold
- ☐ contig
- ☐ chromosome

keyword

Select the kingdom:

- plant
- protozoa
- vertebrate_mammalian
- vertebrate_other
- viral

Assembly level:

- ☒ Complete Genome
- ☐ All

Species:

coronavirus

Assemblies from this date (yyyy-mm-dd):

Taxonomy ID (taxID):

Output folder name:

Number of assemblies (limit): 12

Assembly or Project ID:

Get the latest assembly_summary for kingdoms:

ENA

ENA sequence ID (enaID):

FASTQ run accession:

Search

Snapshot of the GUI launched from Linux Ubuntu system to get sequence information regarding "coronavirus"

Another example to get sequence information regarding *Escherichia coli* bacterial species (limiting the number of assemblies to 4):

```
perl getSequenceInfo.pl -k "bacteria" -s "Escherichia coli" -d "genbank" -n 4
```

The screenshot shows a Windows application window titled "getSequenceInfo". The window has a blue title bar with standard Windows window controls. The main content area is divided into several sections. At the top, it says "Thank you for using getSequenceInfo!". Below this, there are two tabs: "NCBI" and "ENA". The "NCBI" tab is selected. Under the "NCBI" tab, there are several input fields and checkboxes. On the left, there is a "Sequence repository:" section with two radio buttons: "refseq" and "genbank" (which is selected). Below this is a "Component:" section with four checkboxes: "contig", "scaffold", "plasmid", and "chromosome". To the right of these is a "Select the kingdom:" section with a list box containing "archaea", "bacteria" (which is selected and highlighted in blue), "fungi", "invertebrate", and "plant". Below the list box is an "Assembly level:" section with two radio buttons: "Complete Genome" and "All". At the bottom of the "NCBI" section is a "keyword" text box. Below the "keyword" text box are several input fields for "Species:", "Assemblies from this date (yyyy-mm-dd):", "Taxonomy ID (taxID):", "Output folder name:", "Number of assemblies (limit):" (with the value "4" entered), "Assembly or Project ID:", and "Get the latest assembly_summary for kingdoms:". The "ENA" tab is also visible, with input fields for "ENA sequence ID (enaID):" and "FASTQ run accession:". At the bottom of the window is a "Search" button and a progress bar.

Thank you for using getSequenceInfo!

NCBI

Sequence repository:

☐ refseq

☒ genbank

Component:

☐ contig

☐ scaffold

☐ plasmid

☐ chromosome

keyword

Select the kingdom:

archaea

bacteria

fungi

invertebrate

plant

Assembly level:

☐ Complete Genome

☐ All

Species:

Escherichia coli

Assemblies from this date (yyyy-mm-dd):

Taxonomy ID (taxID):

Output folder name:

Number of assemblies (limit):

4

Assembly or Project ID:

Get the latest assembly_summary for kingdoms:

ENA

ENA sequence ID (enaID):

FASTQ run accession:

Search

Snapshot of a search using a Microsoft Windows Operating System

Another example to get sequences from *Naegleria fowleri* species (limiting the number of assemblies to 5):

```
perl getSequenceInfo.pl -k "protozoa" -s "Naegleria fowleri" -d "genbank" -n 5
```

The screenshot shows a graphical user interface for the `getSequenceInfo` application. The window title is `getSequenceInfo`. A message at the top says "Thank you for using getSequenceInfo!". The interface is divided into two main sections: **NCBI** and **ENA**.

NCBI Section:

- Sequence repository:** Radio buttons for `refseq` and `genbank` (selected).
- Component:** Checkboxes for `contig`, `chromosome`, `plasmid`, and `scaffold` (all unselected).
- Select the kingdom:** A list box containing `invertebrate`, `plant`, `protozoa` (selected), `vertebrate_mammalian`, and `vertebrate_other`.
- Assembly level:** Radio buttons for `Complete Genome` and `All` (selected).
- Species:** A text field containing `Naegleria fowleri`.
- Assemblies from this date (yyyy-mm-dd):** An empty text field.
- Taxonomy ID (taxID):** An empty text field.
- Output folder name:** An empty text field.
- Number of assemblies (limit):** A text field containing `5`.
- Assembly or Project ID:** An empty text field.
- Get the latest assembly_summary for kingdoms:** An empty text field.
- keyword:** An empty text field.

ENA Section:

- ENA sequence ID (enaID):** An empty text field.
- FASTQ run accession:** An empty text field.

At the bottom, there is a **Search** button and a progress bar with several blue segments.

Snapshot regarding the search for *Naegleria fowleri* assemblies

Some command lines

We can type the following command to display the help message:

```
perl getSequenceInfo.pl -h
```

The following Help message will appear:

```
Name:
    getSequenceInfo.pl

Synopsis:
    A Perl script allowing to get sequence information from GenBank RefSeq or ENA repositories.

Usage:
    perl getSequenceInfo.pl [options]
examples:
    perl getSequenceInfo.pl -k bacteria -s "Helicobacter pylori" -l "Complete Genome" -date 2019-06-01
    perl getSequenceInfo.pl -k viruses -n 5 -date 2019-06-01
    perl getSequenceInfo.pl -k "bacteria" -taxid 9,24 -n 10 -c plasmid -dir genbank -o Results
    perl getSequenceInfo.pl -ena BN000065
    perl getSequenceInfo.pl -fastq ERR818002
    perl getSequenceInfo.pl -fastq ERR818002,ERR818004

Kingdoms:
    archaea
    bacteria
    fungi
    invertebrate
    plant
    protozoa
    vertebrate_mammalian
    vertebrate_other
    viral

Assembly levels:
    "Complete Genome"
    Chromosome
    Scaffold
    Contig

General:
    -help or -h                displays this help
    -version or -v             displays the current version of the program

Options ([XXX] represents the expected value):
    -directory or -dir [XXX]   allows to indicate the NCBI's nucleotide sequences repository (default: genbank)
    -get or -getSummaries [XXX] allows to obtain a new assembly summary file in function of given kingdoms
                                (bacteria, fungi, protozoa...)
    -k or -kingdom [XXX]       allows to indicate kingdom of the organism (see the examples above)
    -s or -species [XXX]       allows to indicate the species (must be combined with -k option)
    -taxid [XXX]               allows to indicate a specific taxid (must be combined with -k option)
    -assembly_or_project [XXX] allows to indicate a specific assembly accession or bioproject (must be combined with
    -k option)
    -date [XXX]                indicates the release date (with format yyyy-mm-dd) from which sequence information
    are available
    -l or -level [XXX]         allows to select a specific assembly level (e.g. "Complete Genome")
    -o or -output [XXX]        allows users to name the output result folder
    -n or -number [XXX]        allows to limit the total number of assemblies to be downloaded
    -c or -components [XXX]    allows to select specific components of the assembly (e.g. plasmid, chromosome, ...)
    -ena [XXX]                 allows to download report and fasta file given a ENA sequence ID
    -fastq [XXX]               allows to download FASTQ sequences from ENA given a run accession
    (https://ena-docs.readthedocs.io/en/latest/faq/archive-generated-files.html)
    -log                        allows to create a log file
```