

# 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>`

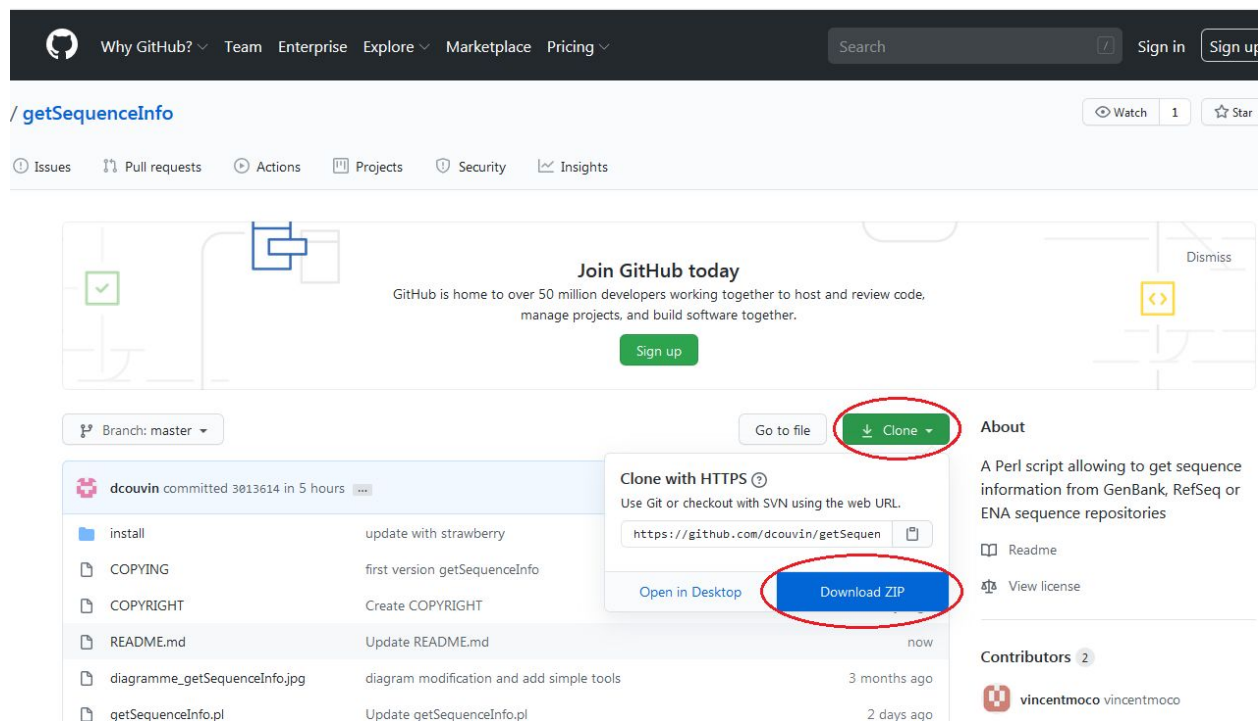
Please note that <Module::Name> should be replaced by the required Perl module (e.g. `cpan -f -i Date::Calc` or `cpanm Date::Calc`).

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/dcouverin/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/dcouverin/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

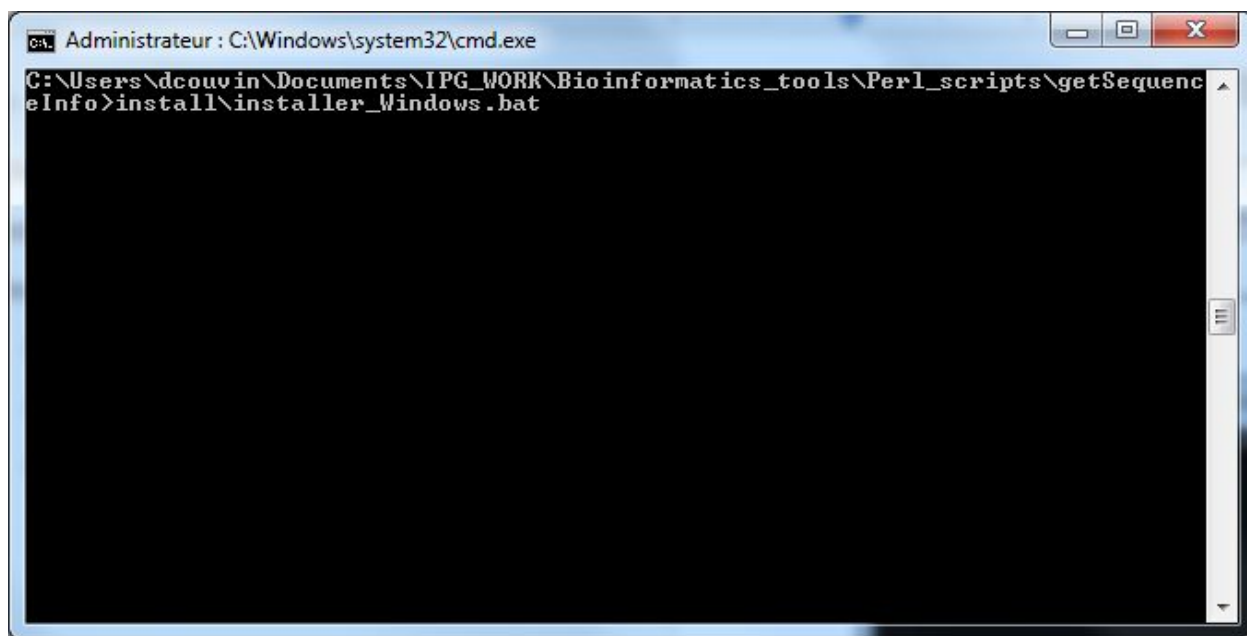
In order to install the getSequenceInfo software tool, the user must be placed into the corresponding uncompressed archive repository (getSequenceInfo).

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)

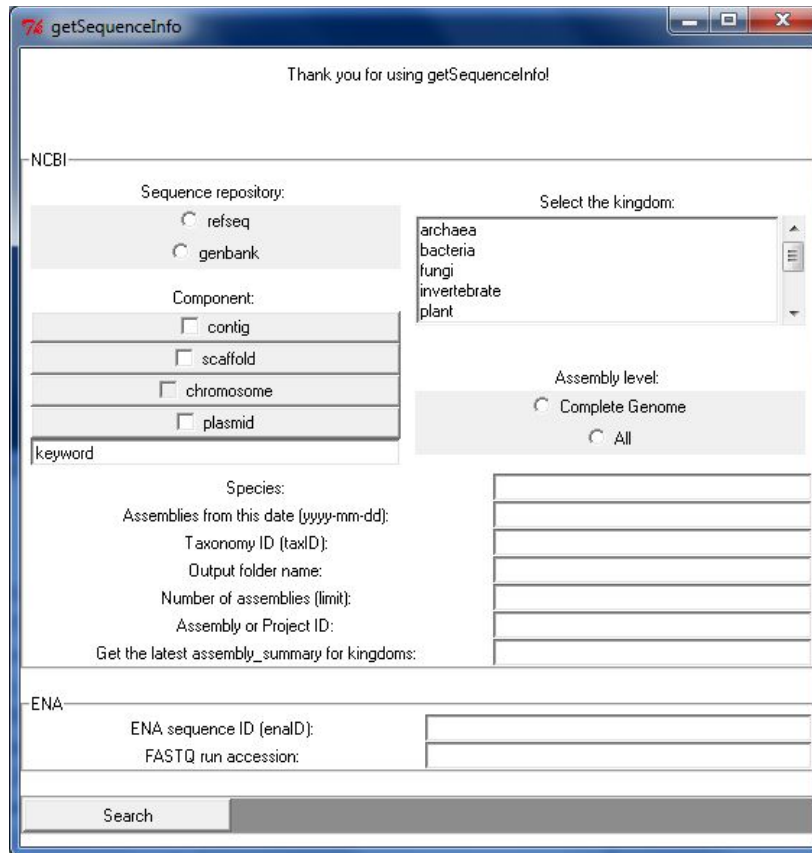
## How to use the tool

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 white and contains a form for searching sequence information. At the top, it says "Thank you for using getSequenceInfo!". Below this, there are two tabs: "NCBI" and "ENA". The "NCBI" tab is selected. The form is divided into several sections. On the left, under "Sequence repository:", there are two radio buttons: "refseq" and "genbank", with "genbank" selected. Below this, under "Component:", there are four checkboxes: "contig", "scaffold", "plasmid", and "chromosome", all of which are unchecked. To the right of these, under "Select the kingdom:", there is a list box with the following options: "archaea", "bacteria" (which is highlighted in blue), "fungi", "invertebrate", and "plant". Below the list box, under "Assembly level:", there are two radio buttons: "Complete Genome" and "All", both of which are unchecked. At the bottom of the form, there are several text input fields. The first is labeled "Species:" and contains the text "Escherichia coli". Below this are four empty text input fields. To the left of these fields, there are labels for "Assemblies from this date (yyyy-mm-dd):", "Taxonomy ID (taxID):", "Output folder name:", "Number of assemblies (limit):", and "Assembly or Project ID:". The "Number of assemblies (limit):" field contains the number "4". Below these fields, there is a checkbox labeled "Get the latest assembly\_summary for kingdoms:". At the bottom of the window, there is a "Search" button and a progress bar with several blue segments.

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 (enalD):

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 sections for NCBI and ENA. Under the NCBI section, there are fields for "Sequence repository:" (with radio buttons for `refseq` and `genbank`, where `genbank` is selected), "Component:" (with checkboxes for `contig`, `chromosome`, `plasmid`, and `scaffold`, all of which are unchecked), and a "keyword" field. There is also a "Select the kingdom:" list box containing `invertebrate`, `plant`, `protozoa` (which is selected), `vertebrate_mammalian`, and `vertebrate_other`. Below this is an "Assembly level:" section with radio buttons for `Complete Genome` (selected) and `All`. Further down, there are input fields for "Species:" (containing `Naegleria fowleri`), "Assemblies from this date (yyyy-mm-dd):", "Taxonomy ID (taxID):", "Output folder name:", "Number of assemblies (limit):" (containing the number `5`), "Assembly or Project ID:", and a checkbox for "Get the latest assembly\_summary for kingdoms:". The ENA section at the bottom has fields for "ENA sequence ID (enaID):" and "FASTQ run accession:". At the very bottom, there is a "Search" button followed by a progress bar consisting of 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
```

# Guide d'utilisation getSequenceInfo (Version française)

---

Ceci est une notice ayant pour but de montrer aux étudiants ainsi qu'aux chercheurs comment utiliser l'outil logiciel getSequenceInfo (<https://github.com/dcouvin/getSequenceInfo>) offrant une interface graphique ainsi qu'un invite de commande

**Mots-clés:** Bio-informatique, Sequence ADN, GenBank, RefSeq, European Nucleotide Archive (ENA), langage de programmation Perl, BioPerl, Annotation, FASTA, FASTQ, GC-content, Statistics, NucleScore

## [Introduction](#)

## [Installation](#)

[Télécharger ou cloner getSequenceInfo](#)

[Lancer l'installation sur un système Unix ou Windows](#)

[Contenu de l'archive getSequenceInfo](#)

## [Comment utiliser l'outil](#)

[Exemples d'utilisation de l'interface graphique \(ou GUI\)](#)

[Quelques lignes de commandes](#)

## Introduction

getSequenceInfo est un script Perl qui permet de facilement télécharger des séquences de données d'un répertoire public tel que NCBI's GenBank et RefSeq, en plus de l'European

Nucleotide Archive ou ENA (situé à l'Institut Européen de Bio-Informatique). Ce programme informatique Perl permet aux utilisateurs de télécharger des séquences de données et des statistiques dans différents formats (FASTA, FASTQ, Genbank full format, XLS, TSV, HTML). L'outil logiciel getSequenceInfo peut être utilisé grâce à l'invite de commande ou à l'interface graphique.

## Installation

En général, **Perl** est déjà installé sur les systèmes Unix (tels que Linux et MacOS). Cependant, pour les personnes utilisant le système d'exploitation Windows, le langage peut être installé avec Strawberry Perl (<http://strawberryperl.com/>).

[Cette vidéo](#) montre comment installer Strawberry Perl:

Si nécessaire, vous pouvez regarder des informations sur [comment lancer](#) ou [comment utiliser](#) l'invite de commande sous Windows. Quand vous utilisez un OS Unix (Linux ou Mac), Perl est généralement déjà installé. Mais si ce n'est pas le cas, vous pouvez regarder [cette page](#) pour l'installer. Vous pouvez suivre ce [wiki](#) pour des informations à propos du Shell. Vous pouvez vérifier l'installation en tapant la commande suivante:

```
perl -v
```

Les utilisateurs peuvent alors installer les modules Perl nécessaire (veuillez noter que ses modules sont déjà disponibles à l'intérieur des fichiers d'installations fourni "**installer\_Unix.sh**" et "**installer\_Windows.bat**"):

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

Chaque module Perl peut être installé en utilisant la commande **cpan** ou **cpanm** comme suit:

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

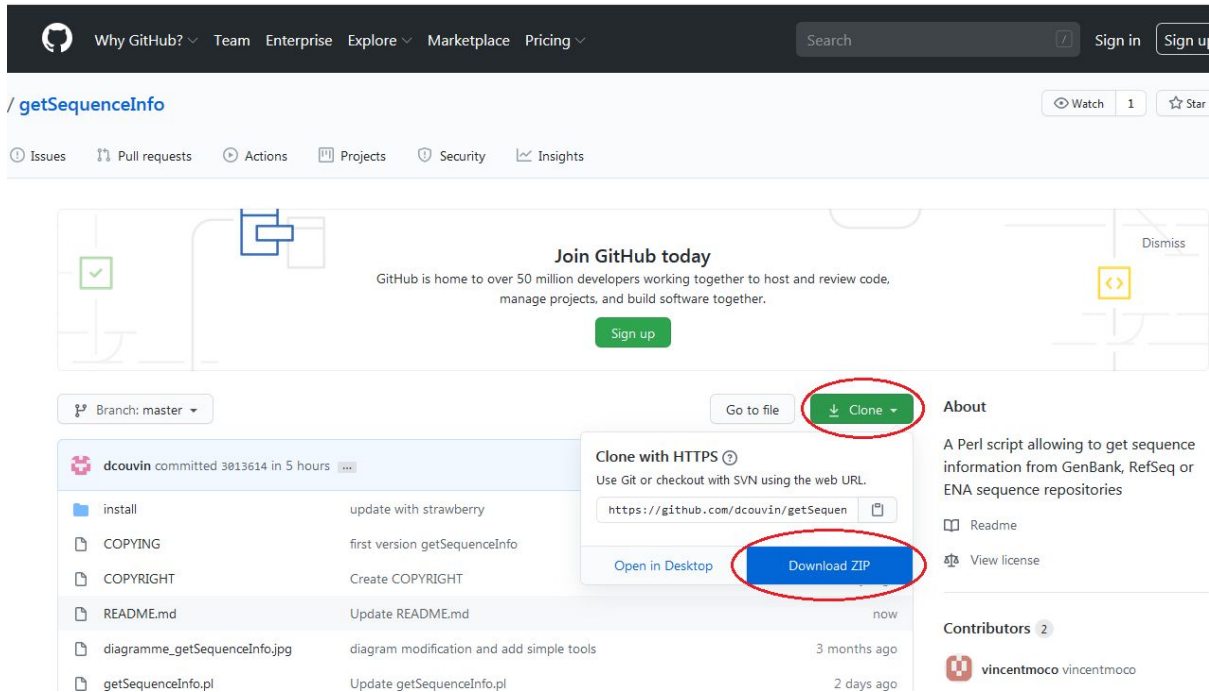
Veuillez noter que <Module::Name> devrait être remplacé par le module exigé (**exemple:** `cpan -f -i Date::Calc` or `cpanm Date::Calc`).

Plus de détails concernant l'installation de modules Perl sont disponibles sur [cette page](#).

## Télécharger ou cloner getSequenceInfo

L'outil peut-être téléchargé ou cloné depuis le [répertoire git](#).

Le bouton "Clone" puis le bouton "Download ZIP" peut-être utilisé pour télécharger le répertoire:



Capture d'écran de la page GitHub permettant de télécharger l'outil.

Autrement, vous pouvez cloner l'outil en utilisant la commande "git clone". Veuillez noter que [git](#) devrait être installé dans votre système pour ça soit possible:

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

Une fois que cette archive à été cloné ou téléchargé, puis dézippé à l'endroit que vous avez choisi, vous pouvez vous rendre dans le répertoire getSequenceInfo en tapant la commande suivante:

```
cd getSequenceInfo
```

Les utilisateurs peuvent aussi naviguer à travers les dossiers pour accéder à l'outil.

## Lancer l'installation sur un système Unix ou Windows

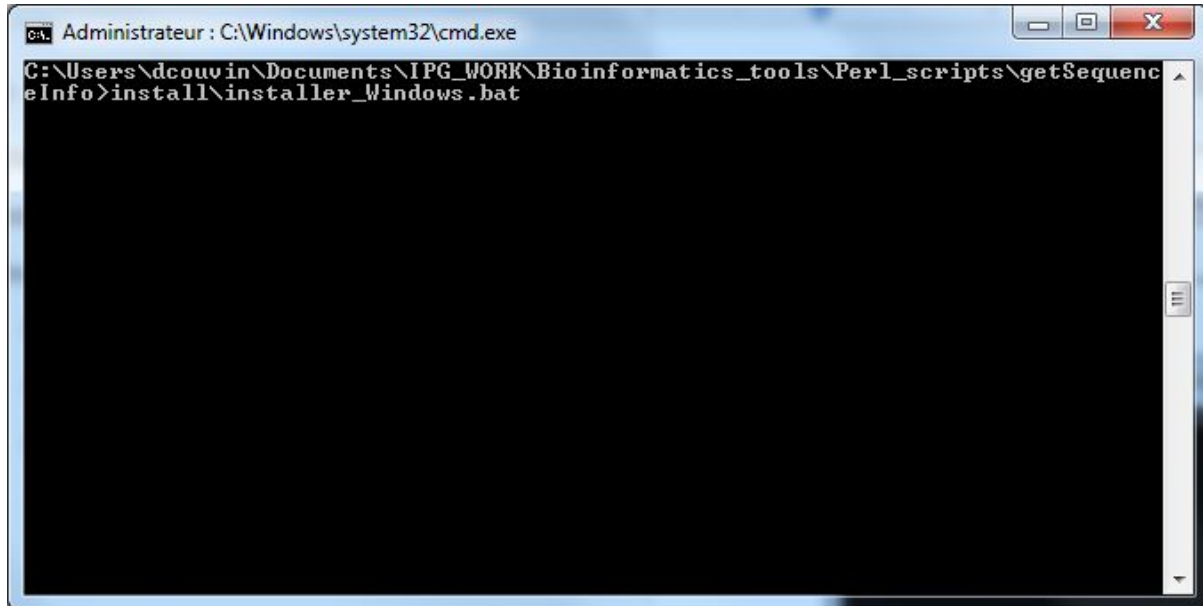
Afin d'installer l'outil logiciel getSequenceInfo, l'utilisateur devrait être placé dans l'archive du répertoire décompressé (getSequenceInfo).

Installation sur Linux ou MacOS (Système Unix) Shell Prompt

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

Installation sur un système Microsoft Windows invite de commande. Les utilisateurs peuvent également installer l'outil en exécutant le fichier `install_Windows.bat` (double-click)

```
install\installer_Windows.bat
```



Capture d'écran de l'installation sur Windows en utilisant l'invite de commande

Les instructions d'installation peuvent aussi être fournies en exécutant le script Perl `getSequenceInfo.pl`

## Contenu de l'archive getSequenceInfo

L'archive contient les fichiers et dossiers suivants:

- **install** (Un dossier contenant les fichiers d'installation "**installer\_Unix.sh**" et "**installer\_Windows.bat**")
- **simple\_tools** (un dossier contenant d'autres scripts Perl dédiés à l'exécution de tâches spécifiques.)
- **COPYING** et **COPYRIGHT** (fichiers de licence GPLv3)
- **README.md** (un simple fichier README)
- **workflow.png** (un diagramme représentant les principales fonctionnalités de l'outil logiciel)
- **getSequenceInfo.pl** (le principal programme Perl)

- **getSequenceInfoGUI.pl** (Interface graphique du programme)
- **launcher\_Windows.bat** (exécutable permettant de lancer l'interface graphique depuis Windows)
- **logo\_getSequenceInfo.png** (un logo représentant l'outil)
- **User\_manual.pdf** (Manuel utilisateur)

## Comment utiliser l'outil

Comme mentionné plus tôt, l'outil peut être utilisé depuis l'invite de commande ou avec l'invite de commande. Le même usage sémantique est partagé entre les deux interfaces. Veuillez noter que les deux fichiers Perl (getSequenceInfo.pl et getSequenceInfoGUI.pl) devraient être localisés dans le même répertoire.

L'utilisateur peut lancer l'interface graphique de l'outil (getSequenceInfoGUI.pl) soit en l'exécutant (double click) ou en tapant la commande suivante :

```
perl getSequenceInfoGUI.pl
```

La fenêtre suivante devrait apparaître devant un invite de commande:

The screenshot shows a graphical user interface for the getSequenceInfo tool. The window is titled 'getSequenceInfo' and displays a message: 'Thank you for using getSequenceInfo!'. The interface is organized into several sections:

- NCBI Section:**
  - Sequence repository:** Two radio buttons, 'refseq' and 'genbank'.
  - Component:** Four checkboxes, 'contig', 'scaffold', 'chromosome', and 'plasmid'.
  - keyword:** A text input field.
  - Select the kingdom:** A list box containing 'archaea', 'bacteria', 'fungi', 'invertebrate', and 'plant'.
  - Assembly level:** Two radio buttons, 'Complete Genome' and 'All'.
- Species Section:**
  - Species:** A text input field.
  - Assemblies from this date (yyyy-mm-dd):** A text input field.
  - Taxonomy ID (taxID):** A text input field.
  - Output folder name:** A text input field.
  - Number of assemblies (limit):** A text input field.
  - Assembly or Project ID:** A text input field.
  - Get the latest assembly\_summary for kingdoms:** A checkbox.
- ENA Section:**
  - ENA sequence ID (enaID):** A text input field.
  - FASTQ run accession:** A text input field.
- Search:** A button at the bottom left of the window.

Capture d'écran de l'interface graphique sous Windows.

## Exemples d'utilisation de l'interface graphique (ou GUI)

Il existe plusieurs manières d'utiliser cet outil pour télécharger un assemblage et les informations qui y sont apparentées.

Le moyen le plus simple est d'utiliser le nom de l'espèce que vous souhaitez chercher et y associer le règne correspondant.

Supposons que nous voulons chercher des informations sur le coronavirus.

Nous pouvons utiliser la commande suivante afin de récupérer 12 assemblages de génome disponible depuis le 2019/12/01:

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

The screenshot shows a graphical user interface titled "getSequenceInfo". At the top, it says "Thank you for using getSequenceInfo!". The interface is divided into two main sections: "NCBI" and "ENA".

**NCBI Section:**

- Sequence repository:** Two radio buttons, "refseq" (selected) and "genbank".
- Component:** Four checkboxes: "plasmid", "scaffold", "contig", and "chromosome".
- Select the kingdom:** A list box containing "plant", "protozoa", "vertebrate\_mammalian", "vertebrate\_other", and "viral" (selected).
- Assembly level:** Two radio buttons, "Complete Genome" (selected) and "All".
- Species:** A text input field containing "coronavirus".
- Assemblies from this date (yyyy-mm-dd):** An empty text input field.
- Taxonomy ID (taxID):** An empty text input field.
- Output folder name:** An empty text input field.
- Number of assemblies (limit):** A text input field containing "12".
- Assembly or Project ID:** An empty text input field.
- Get the latest assembly\_summary for kingdoms:** An empty text input field.

**ENA Section:**

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

At the bottom, there is a "Search" button followed by a row of seven blue square buttons and a grey square button.

Capture d'écran de l'interface graphique lancé depuis un système Linux Ubuntu afin de récupérer les séquences concernant le "coronavirus"



Un autre exemple pour récupérer des informations concernant l'espèce bactérienne *Escherichia coli* (en limitant le nombre d'assemblages à 4):

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

The screenshot shows a Windows application window titled "getSequenceInfo". The main content area has a header "Thank you for using getSequenceInfo!". Below this, there are two tabs: "NCBI" (selected) and "ENA". Under the "NCBI" tab, the "Sequence repository:" section has radio buttons for "refseq" and "genbank" (selected). The "Component:" section has checkboxes for "contig", "scaffold", "plasmid", and "chromosome". A "keyword" input field is present. The "Select the kingdom:" dropdown menu is open, showing options: "archaea", "bacteria" (selected), "fungi", "invertebrate", and "plant". The "Assembly level:" section has radio buttons for "Complete Genome" and "All". The "Species:" input field contains "Escherichia coli". Below this, there are input fields for "Assemblies from this date (yyyy-mm-dd):", "Taxonomy ID (taxID):", "Output folder name:", "Number of assemblies (limit):" (containing "4"), and "Assembly or Project ID:". A checkbox for "Get the latest assembly\_summary for kingdoms:" is also present. The "ENA" tab has input fields for "ENA sequence ID (enalD):" and "FASTQ run accession:". At the bottom, there is a "Search" button and a progress bar.

Capture d'écran d'une recherche faite avec l'interface graphique sous Windows

Un autre exemple pour récupérer des séquences de l'espèce *Naegleria fowleri* (en limitant le nombre d'assemblages à 5):

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

**getSequenceInfo**

Thank you for using getSequenceInfo!

---

**NCBI**

**Sequence repository:**

☒ refseq  
☒ genbank

**Component:**

☐ contig  
☐ chromosome  
☐ plasmid  
☐ scaffold

keyword

**Select the kingdom:**

invertebrate  
plant  
protozoa  
vertebrate\_mammalian  
vertebrate\_other

**Assembly level:**

☒ Complete Genome  
☐ All

**Species:**

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

**Taxonomy ID (taxID):**

**Output folder name:**

**Number of assemblies (limit):**

**Assembly or Project ID:**

**Get the latest assembly\_summary for kingdoms:**

Naegleria fowleri

5

---

**ENA**

**ENA sequence ID (enaID):**

**FASTQ run accession:**

**Search**

Capture d'écran concernant la recherche pour les assemblages de *Naegleria fowleri*

Quelques lignes de commandes

Nous pouvons entrer la commande suivante pour afficher le message d'aide:

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

Ce message d'aide devrait apparaître:

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