IntegronFinder Documentation

Release 1.5

Jean Cury, Bertrand Néron, Eduardo PC Rocha

November 14, 2016

Contents

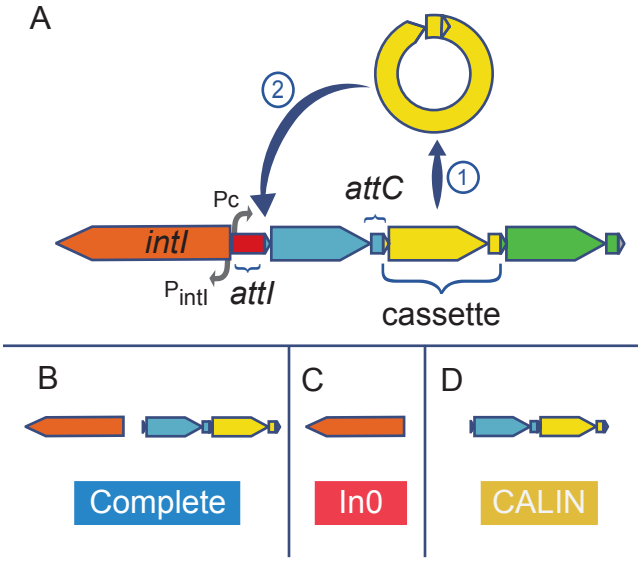
IntegronFinder is a program that detects integrons in DNA sequences. The program is available on a webserver *(Mobyle)*, or by command line (IntegronFinder on github).

- You already read the *paper* and want to install it ? Click *here*
- You did not read the paper (yet) but you would like to have rapid introduction to integrons and the program? click *here*

CHAPTER 1

Introduction

Integrons are major genetic element, notorious for their major implication in the spread of antibiotic resistance genes. More generally, integrons are gene-capturing platform, whose broader evolutionary role remains poorly understood. IntegronFinder is able to detect with high accuracy integron in DNA sequences. It is accurate because it combines the use of HMM profiles for the detection of the essential protein, the site-specific integron integrase, and the use of Covariance Models for the detection of the recombination site, the *attC* site.



How does it work ?

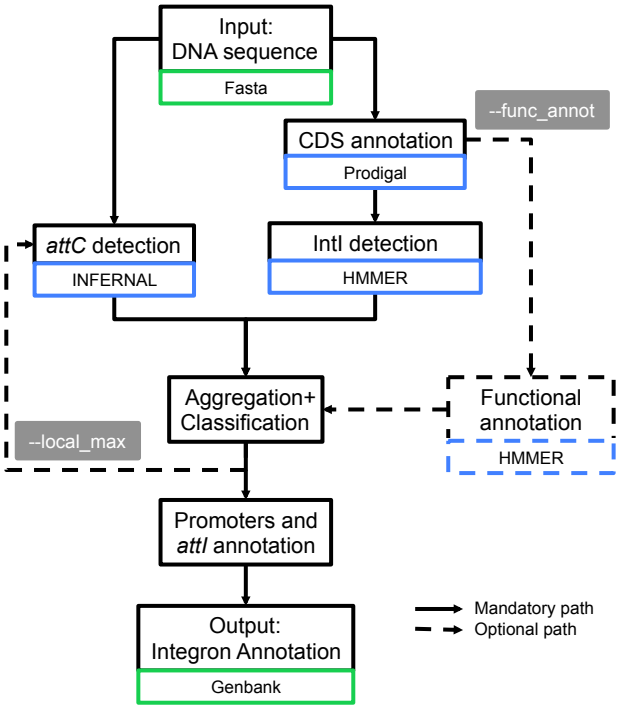
- First, IntegronFinder annotates the DNA sequence's CDS with Prodigal.
- Second, IntegronFinder detects independently integron integrase and *attC* recombination sites. The Integron integrase is detected by using the intersection of two HMM profiles:
  - one specific of tyrosine-recombinase (PF00589)
  - one specific of the integron integrase, near the patch III domain of tyrosine recombinases.

The *attC* recombination site is detected with a covariance model (CM), which models the secondary structure in addition to the few conserved sequence positions.

- Third, the results are integrated, and IntegronFinder distinguishes 3 types of elements:
  - **complete integron (panel B above)** Integron with integron integrase nearby *attC* site(s)
  - **In0 element (panel C above)** Integron integrase only, without any *attC* site nearby
  - **CALIN element (panel D above)** *attC* sites only, without integron integrase nearby. A rule of thumb to avoid false positive is to filter out singleton of *attC* site.

IntegronFinder can also annotate gene cassettes (CDS nearby *attC* sites) using Resfams, a database of HMM profiles aiming at annotating antibiotic resistance genes. This database is provided but the user can add any other HMM profiles database of its own interest.

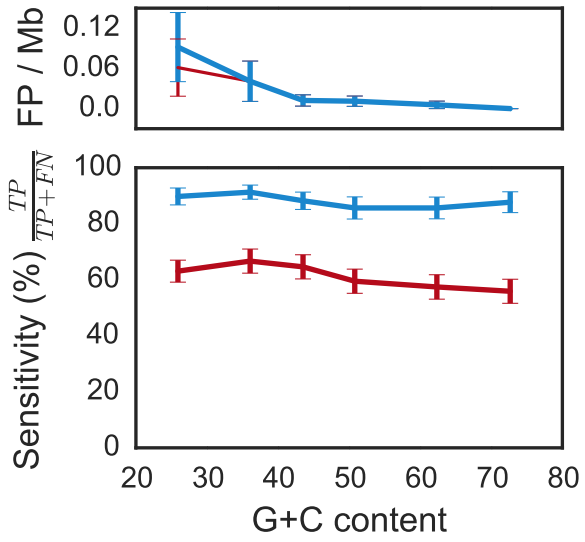
When available, IntegronFinder annotates the promoters and attI sites by pattern matching.



Does it work ?

Yes! The estimated sensitivity is 61% on average with the default option and goes up to 88% with the `--local_max` option. The missing *attC* sites are usually at the end of the array. The False positive rate with the `--local_max` option is estimated between 0.03 False Positive per Megabases (FP/Mb) to 0.72 FP/Mb. This leads to a probability of finding 2 consecutive *attC* sites within 4kb between 4.10^-6 and 7.10^-9. Finally, this parameters do not depend on the G+C percent of the given replicon.

	Default	local_max
Sensitivity	61.20%	88.03%
FP rate	0.02 FP/Mb	0.03 FP/Mb
Mean time	2.59 s	86.59 s



The time in the table correspond to the average time per run with a pseudogenome having attC sites on a Mac Pro, 2 x 2.4 GHz 6-Core Intel Xeon, 16 Gb RAM, with options –cpu 20 and –no-proteins.

Note: The time does not vary depending of the mode (default or local_max), and is about a couple of second, if the replicon does not contain any *attC* site.

CHAPTER 2

Installation

2.1 IntegronFinder dependencies

IntegronFinder is built with Python 2.7, and a few libraries are needed:

- **Python 2.7**
  - Pandas (>=0.18.0)
  - Numpy (>=1.9.1)
  - Biopython (>=1.65)
  - Matplotlib (>=1.4.3)

If you're not at ease with Python, see *here on how to install Python and its libraries*

In addition, IntegronFinder has external dependencies which have to be installed prior the use of the program (click to access the corresponding website).

- HMMER 3.1b1
- INFERNAL 1.1
- Prodigal V2.6.2

After installation of these programs, they should be in your `$PATH` (*i.e.* you can type in a terminal `hmmsearch`, `cmsearch`, or `prodigal` and a `command not found` shall not be displayed). If you have them installed somewhere else, please refers to the parameters to give complete path to IntegronFinder.

2.2 Installation procedure

1. Download the latest release

2. In a shell (*e.g.* a terminal), start installation with:

```
(sudo) python pip install integron_finder-x.x.tar.gz
```

Note: Super-user privileges (*i.e.*, `sudo`) are necesserary if you want to install the program in the general file architecture.

Note: If you do not have the privileges, or if you do not want to install IntegronFinder in the Python libraries of your system, you can install IntegronFinder in a virtual environment. See virtualenv or if you're using Canopy, see Canopy CLI

Warning: When installing a new version of IntegronFinder, do not forget to uninstall the previous version installed !

Warning: The installer does not work with pure setuptools procedure, it does not work in egg. Unless you disable egg by using the `--root` option. `python setup.py install --root /prefix/where/to/install/integron_finder`

2.3 Uninstallation procedure

To uninstall integron_finder, run in the following command:

```
(sudo) pip uninstall integron_finder
```

It will uninstall integron_finder executable

2.4 How to install Python

The purpose of this section is to provide some help about installing python dependencies for IntegronFinder if you never installed any python package.

As IntegronFinder has not been test on Windows, we assume Unix-based operating system. For Windows users, the best would be to install a unix virtual machine on your computer.

Usually a python distribution is already installed on your machine. However, if you don't know how to install libraries, we recommend to re-install it from a distribution which contains pre-compiled libraries. There are two main distributions (click to access website):

- Enthought Canopy

- Anaconda

Download version 2.7 which correspond to your machine, then make sure that python from these distributions is the default one (you can possibly choose that in the preference and/or during installation). They both come with all the needed packages but Biopython. If you have a **student email adress** from a university-delivering degree, you can request an academic licence to *Enthough Canopy* (see Canopy for Academics) which will allow you to download additional packages including Biopython.

Otherwise, you will have to install Biopython manually. `pip` is recommended as a python packages installer. It works as follow:

```
(sudo) pip install Biopython==1.65
```

To install version 1.65 of Biopython (recommended for IntegronFinder).

Note: If you don't manage to install all the packages, try googling the error, or don't hesisate to ask a question on stackoverflow.

CHAPTER 3

Tutorial

We assume here that the program is *installed*.

3.1 Basic use

Note: The different options will be shown separately, but they can be used alltogether unless otherwise stated.

You can see all available options with:

```
integron_finder -h
```

You can go to directory containing your sequence, or specify the path to that sequence and call:

```
integron_finder mysequence.fst
```

or:

```
integron_finder path/to/mysequence.fst
```

It will perform a search, and outputs the results in a directory called `Results_Integron_Finder_mysequence`. Within this directory, you can find:

- **mysequence.integrons** A tabular file with the annotations of the different elements
- **mysequence.gbk** A GenBank file with the sequence annotated with the same annotations from the previous file.
- **mysequence_X.pdf** For each complete integron, a simple graphic of the region is depicted
- **other** A folder containing outputs of the different step in the program. It includes notably the protein file in fasta (mysequence.prt).

3.1.1 Thorough local detection

This option allows a more sensitive search. It will be slower if integrons are found, but will be as fast if nothing is detected:

```
integron_finder mysequence.fst --local_max
```

3.1.2 Functional annotation

This option allows to annotate cassettes given HMM profiles. As Resfams database is distributed, to annotate antibiotic resistance genes, just use:

```
integron_finder mysequence.fst --func_annot
```

IntegronFinder will look in the directory `Integron_Finder-x.x/data/Functional_annotation` and use all `.hmm` files available to annotate. By default, there is only `Resfams.hmm`, but one can add any other HMM file here. Alternatively, if one wants to use a database which is present elsewhere on the user's computer without copying it into that directory, one can specify the following option:

```
integron_finder mysequence.fst --path_func_annot bank_hmm
```

where `bank_hmm` is a file containing one absolute path to a hmm file per line, and you can comment out a line:

```
~/Downloads/Integron_Finder-x.x/data/Functional_annotation/Resfams.hmm
~/Documents/Data/Pfam-A.hmm
# ~/Documents/Data/Pfam-B.hmm
```

Here, annotation will be made using Pfam-A et Resfams, but not Pfam-B. If a protein is hit by 2 different profiles, the one with the best e-value will be kept.

3.1.3 Parallelization

The time limiting part are HMMER and INFERNAL. So IntegronFinder does not have parallel implementation (yet?), but the user can set the number of CPU used by HMMER and INFERNAL:

```
integron_finder mysequence.fst --cpu 4
```

Default is 1.

3.1.4 Circularity

By default, IntegronFinder assumes your replicon to be circular. However, if they aren't, or if it's PCR fragments or contigs, you can specify that it's a linear fragment:

```
integron_finder mylinearsequence.fst --linear
```

However, if `--linear` is not used and the replicon is smaller than `4 x dt` (where `dt` is the distance threshold, so 4kb by default), the replicon is considered linear to avoid clustering problem

3.2 Advanced options

3.2.1 Clustering of elements

attC sites are clustered together if they are on the same strand and if they are less than 4 kb apart. To cluster an array of *attC* sites and an integron integrase, they also must be less than 4 kb apart. This value has been empirically estimated and is consistent with previous observations showing that biggest gene cassettes are about 2 kb long. This value of 4 kb can be modify though:

```
integron_finder mysequence.fst --distance_thresh 10000
```

or, equivalently:

```
integron_finder mysequence.fst -dt 10000
```

This sets the threshold for clustering to 10 kb.

Note: The option `--outdir` allows you to chose the location of the Results folder (`Results_Integron_Finder_mysequence`). If this folder already exists, IntegronFinder will not re-run analyses already done, except functional annotation. It allows you to re-run rapidly IntegronFinder with a different `--distance_threshold` value. Functional annotation needs to re-run each time because depending on the aggregation parameters, the proteins associated with an integron might change.

3.2.2 *attC* evalue

The default evalue is 1. Sometimes, degenerated *attC* sites can have a evalue above 1 and one may want to increase this value to have a better sensitivity, to the cost of a much higher false positive rate.

```
integron_finder mysequence.fst --evalue_attc 5
```

3.2.3 Palindromes

attC sites are more or less palindromic sequences, and sometimes, a single *attC* site can be detected on the 2 strands. By default, the one with the highest evalue is discarded, but you can choose to

keep them with the following option:

```
integron_finder mysequence.fst --keep_palindromes
```

CHAPTER 4

Mobyle

You can access IntegronFinder online, on the Mobyle server of the Pasteur institute

4.1 How to use it

1. Copy your sequence or upload it in the appropriate field.
2. Select the options you want
3. Click on Run

If you want more options:

3. Click on advanced options (instead of Run)
4. Select the options you want
5. Click on Run

You can see the role of the different functions in the *tutorial* page, or by clicking on the  in the corresponding field.

After submitting your job, you may need to enter your email.

4.2 Results

Once the job is finished, you have a result page, which contains:

- **integron_finder.out:** Log of the run. It tells you how many integrons have been found for each types along with the number of *attC* sites per type.

- **Schema of complete integron(s)** [replicon_X.pdf] Simple representation of one or more complete integrons found. The representation is very basic and a better representation can be obtained from the GenBank file and a software (eg Geneious) to represent it.
- **annotated sequence** [replicon.gbk] The GenBank file of the input sequence with the annotation corresponding to the elements found (integrase, *attC*, promoter, attI, etc...).
- **putative integrons** [replicon.integrons] A tabular file listing all the elements and their caracteristics.

Finally, you have your initial sequence of the replicon and the command line used.

For each of the aforementionned files, you can save them by clicking on the save button 

.

CHAPTER 5

References

If you use this software, please cite:

- Identification and analysis of integrons and cassette arrays in bacterial genomes

Jean Cury; Thomas Jove; Marie Touchon; Bertrand Neron; Eduardo PC Rocha. **Nucleic Acids Research**, 2016; doi: 10.1093/nar/gkw319

Please cite also the following articles:

- Nawrocki, E.P. and Eddy, S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. **Bioinformatics**, 29, 2933-2935.
- Eddy, S.R. (2011) Accelerated Profile HMM Searches. **PLoS Comput Biol**, 7, e1002195.
- Hyatt, D., Chen, G.L., Locascio, P.F., Land, M.L., Larimer, F.W. and Hauser, L.J. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. **BMC Bioinformatics**, 11, 119.

and if you use ResFams, cite the corresponding articles:

- Gibson, M.K., Forsberg, K.J. and Dantas, G. (2015) Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. **ISME J**, 9, 207-216.