rna_cd Documentation

Release 0.2.0

Sander Bollen, Leiden University Medical Center

Jun 21, 2019

Contents:

Human DNA-seq experiments may be contaminated with RNA. This usually negatively influences downstream analysis. For example, it can introduce false positive variants around splice sites, suppress alternative alleles due to allele-specific expression, and alter coverage patterns which may then influence CNV calling. Moreover, it is typically hard to detect.

Modern Illumina sequencers unfortunately are more prone to such contamination, as the very large capacity of novaseq sequencers typically means multiple projects are sequenced on the same flowcells. Combined with the increased risk of index hopping in novaseq sequencers, this alltogether means that cross-contamination of samples is more likely - including contamination of RNA into DNA-seq experiments.

rna_cd is a python package and command line tool designed to detect such RNA contamination of DNA-seq experiments. It uses the altered coverage and softclip patterns in contaminated samples to train a Support Vector Machine that can classify BAM files into contaminated ("positive") and uncontaminated ("negative") groups.

rna_cd stands for *RNA contamination detector*.

CHAPTER 1

User documentation

1.1 Installation

`rna_cd` is now available on PyPI! Simply install the package in your favorite python environment with:

```
$ pip install rna-cd
```

This will install both the `rna_cd` python package, and install two command line tools:

1. `rna_cd-train`: For training a model using BAM files.
2. `rna_cd-classify`: For classifying new BAM files.

1.1.1 Supported python versions

We only support the following python versions:

- python 3.5
- python 3.6
- python 3.7

> **Note:** Python 2 is **not** supported in any way. Even attempting to install the package on python 2 will result in failures.

1.2 Training

To train the support vector machine, you need a set of contaminated ("positive") and a set of uncontaminated ("negative") BAM files. For each category, you can organize your BAM files in two distinct ways:

1. Place all BAM files of the category in the same directory.
2. Make a flat text file, where each line points to a path of a BAM file.

Note: Your BAM files must be indexed.

Once you have this in place, you need to choose the contig in your BAM file that you want to collect metrics for, and the chunksize. rna_cd will split your contig of interest in chunks with a maximum size `chunksize`, and collect a number of metrics for each chunk. When you later use the model for classifications you **must** use the same contig and chunksize as you used during the training step.

In our hands, the mitochondrial contig, at a chunksize of 100 bases, gives enough information to be trainable. When choosing the mitochondrial contig, one also benefits from its small size, which makes the training step fast.

Training can work in multicore mode. When using multiple cores, you will process multiple BAM files simultaneously. This can drastically speed up the metric collection for large numbers of BAM files.

Lastly, you have to set the amount of fold cross validations. By default this is 3, but you may set it to any positive integer.

The created model will saved to disk as a JSON file. The JSON file contains the pickled model.

Optionally, you can save a plot of the top two principal components of the training samples to disk.

1.2.1 Examples

Directory method, chrM, chunksize = 100, cores = 3

```
rna_cd-train -c chrM -pd positives_dir -nd negatives_dir -j 3 \
--chunksize 100 -o model.json
```

List method, chrM, chunksize = 100, cores = 3

```
rna_cd-train -c chrM -pl positives.list -nl negatives.list -j 3 \
--chunksize 100 -o model.json
```

List method, chrM, chunksize = 100, cores = 3, with plot

```
rna_cd-train -c chrM -pl positives.list -nl negatives.list -j 3 \
--chunksize 100 -o model.json --plot-out pca.png
```

1.2.2 Usage

rna_cd-train

```
rna_cd-train [OPTIONS]
```

Options

--chunksize <chunksize>
Chunksize in bases. Default = 100

-c, --contig <contig>
Name of mitochrondrial contig in your BAM files. Default = chrM

-pd, --positives-dir <positives_dir>
Path to directory containing positive BAM files. Mutually exclusive with –positives-list

-nd, --negatives-dir <negatives_dir>
Path to directory containing negative BAM files. Mutually exlusive with –negatives-list

-pl, --positives-list <positives_list>
Path to file containing a list of paths to positive BAM files. Mutually exclusive with –positives-dir

-nl, --negatives-list <negatives_list>
Path to file containing a list of paths to negative BAM files. Mutuallly exclusive with –negatives-dir

--cross-validations <cross_validations>
Number of folds for cross validation run. Default = 3

--verbosity <verbosity>
Verbosity value for cross validation step. Default = 1

-j, --cores <cores>
Number of cores to use for processing of BAM files and cross validations. Default = 1

--plot-out <plot_out>
Optional path to PCA plot.

-o, --model-out <model_out>
Path where model will be stored. [required]

1.3 Classification

As with the training step, you can organize your BAM files in two distinct ways:

1. Place all BAM files of the category in the same directory.
2. Make a flat text file, where each line points to a path of a BAM file.

This time, there are no separate categories, as all BAM files are a-priori unknown.

Note: Your BAM files must be indexed.

Warning: As mentioned before, you **must** use the **exact** same contig and chunksize settings in this step as were used during the training step.

As with the training step, metric collection can run in multicore mode during classifcation as well.

Once you have prepared your BAM files, and chosen your parameters, you will use the model you generated during the training step to classify your BAM files into contaminated (“positive”) and uncontamined (“negative”) groups.

Additionally, there is an optional third category with label “unknown”. This represents samples for which we are unsure to which category they belong to. By default, samples are categorized as “unknown” when the class probability

of the most likely category is below 0.75. You can override this parameter, but it must be higher than 0.5 and lower than 1.0.

The classifications will be stored to disk in a three-column tab-delimited text file, with the following columns:

1. Name of the BAM file that was classified.
2. Assigned category ("pos" or "neg" for positive and negative classifications, respectively, and "unknown" for the unknown category).
3. The probability of the assigned category.

E.g. an example output file could look like

```
filename    predicted_class class_probability
a.bam   neg 0.95
b.bam   neg 0.88
c.bam   pos 0.75
d.bam   unknown 0.55
```

1.3.1 Examples

Directory method, chrM, chunksize = 100, cores = 3

```
rna_cd-classify -m model.json -d bams_dir -j 3 -c chrM \
--chunksize 100 -o classifications.out
```

List method, chrM, chunksize = 100, cores = 3

```
rna_cd-classify -m model.json -l bams.list -j 3 -c chrM \
--chunksize 100 -o classifications.out
```

1.3.2 Usage

rna_cd-classify

```
rna_cd-classify [OPTIONS]
```

Options

--chunksize <chunksize>
Chunksize in bases. Default = 100

-c, --contig <contig>
Name of mitochrondrial contig in your BAM files. Default = chrM

-j, --cores <cores>
Number of cores to use for processing of BAM files. Default = 1

-d, --directory <directory>
Path to directory with BAM files to be tested. Mutually exclusive with –list-items

-l, --list-items <list_items>
Path to file containing list of paths to BAM files to be tested. Mutually exclusive with –directory

-m, --model <model>
Path to model. [required]

-o, --output <output>
Path to output file containing classifications. [required]

-t, --unknown-threshold <unknown_threshold>
Threshold of most likely probability below which sampleswll be assinged as 'unknown'. Default = 0.75

1.4 Changelog

1.4.1 0.2.0-dev

- Add some more columns to the classification output

1.4.2 0.1.0

- Make tool
- Add tests
- Add sphinx documentation
- Add optional 'unknown' category.

1.5 LICENSE

```
                    GNU AFFERO GENERAL PUBLIC LICENSE
                       Version 3, 19 November 2007

 Copyright (C) 2007 Free Software Foundation, Inc. <http://fsf.org/>
 Everyone is permitted to copy and distribute verbatim copies
 of this license document, but changing it is not allowed.

                            Preamble

  The GNU Affero General Public License is a free, copyleft license for
software and other kinds of works, specifically designed to ensure
cooperation with the community in the case of network server software.

  The licenses for most software and other practical works are designed
to take away your freedom to share and change the works.  By contrast,
our General Public Licenses are intended to guarantee your freedom to
share and change all versions of a program--to make sure it remains free
software for all its users.

  When we speak of free software, we are referring to freedom, not
price.  Our General Public Licenses are designed to make sure that you
have the freedom to distribute copies of free software (and charge for
them if you wish), that you receive source code or can get it if you
```

(continues on next page)

(continued from previous page)

```
want it, that you can change the software or use pieces of it in new
free programs, and that you know you can do these things.

  Developers that use our General Public Licenses protect your rights
with two steps: (1) assert copyright on the software, and (2) offer
you this License which gives you legal permission to copy, distribute
and/or modify the software.

  A secondary benefit of defending all users' freedom is that
improvements made in alternate versions of the program, if they
receive widespread use, become available for other developers to
incorporate.  Many developers of free software are heartened and
encouraged by the resulting cooperation.  However, in the case of
software used on network servers, this result may fail to come about.
The GNU General Public License permits making a modified version and
letting the public access it on a server without ever releasing its
source code to the public.

  The GNU Affero General Public License is designed specifically to
ensure that, in such cases, the modified source code becomes available
to the community.  It requires the operator of a network server to
provide the source code of the modified version running there to the
users of that server.  Therefore, public use of a modified version, on
a publicly accessible server, gives the public access to the source
code of the modified version.

  An older license, called the Affero General Public License and
published by Affero, was designed to accomplish similar goals.  This is
a different license, not a version of the Affero GPL, but Affero has
released a new version of the Affero GPL which permits relicensing under
this license.

  The precise terms and conditions for copying, distribution and
modification follow.

                       TERMS AND CONDITIONS

  0. Definitions.

  "This License" refers to version 3 of the GNU Affero General Public License.

  "Copyright" also means copyright-like laws that apply to other kinds of
works, such as semiconductor masks.

  "The Program" refers to any copyrightable work licensed under this
License.  Each licensee is addressed as "you".  "Licensees" and
"recipients" may be individuals or organizations.

  To "modify" a work means to copy from or adapt all or part of the work
in a fashion requiring copyright permission, other than the making of an
exact copy.  The resulting work is called a "modified version" of the
earlier work or a work "based on" the earlier work.

  A "covered work" means either the unmodified Program or a work based
on the Program.

  To "propagate" a work means to do anything with it that, without
```

(continues on next page)

(continued from previous page)

```
permission, would make you directly or secondarily liable for
infringement under applicable copyright law, except executing it on a
computer or modifying a private copy.  Propagation includes copying,
distribution (with or without modification), making available to the
public, and in some countries other activities as well.

  To "convey" a work means any kind of propagation that enables other
parties to make or receive copies.  Mere interaction with a user through
a computer network, with no transfer of a copy, is not conveying.

  An interactive user interface displays "Appropriate Legal Notices"
to the extent that it includes a convenient and prominently visible
feature that (1) displays an appropriate copyright notice, and (2)
tells the user that there is no warranty for the work (except to the
extent that warranties are provided), that licensees may convey the
work under this License, and how to view a copy of this License.  If
the interface presents a list of user commands or options, such as a
menu, a prominent item in the list meets this criterion.

  1. Source Code.

  The "source code" for a work means the preferred form of the work
for making modifications to it.  "Object code" means any non-source
form of a work.

  A "Standard Interface" means an interface that either is an official
standard defined by a recognized standards body, or, in the case of
interfaces specified for a particular programming language, one that
is widely used among developers working in that language.

  The "System Libraries" of an executable work include anything, other
than the work as a whole, that (a) is included in the normal form of
packaging a Major Component, but which is not part of that Major
Component, and (b) serves only to enable use of the work with that
Major Component, or to implement a Standard Interface for which an
implementation is available to the public in source code form.  A
"Major Component", in this context, means a major essential component
(kernel, window system, and so on) of the specific operating system
(if any) on which the executable work runs, or a compiler used to
produce the work, or an object code interpreter used to run it.

  The "Corresponding Source" for a work in object code form means all
the source code needed to generate, install, and (for an executable
work) run the object code and to modify the work, including scripts to
control those activities.  However, it does not include the work's
System Libraries, or general-purpose tools or generally available free
programs which are used unmodified in performing those activities but
which are not part of the work.  For example, Corresponding Source
includes interface definition files associated with source files for
the work, and the source code for shared libraries and dynamically
linked subprograms that the work is specifically designed to require,
such as by intimate data communication or control flow between those
subprograms and other parts of the work.

  The Corresponding Source need not include anything that users
can regenerate automatically from other parts of the Corresponding
Source.
```

(continues on next page)

(continued from previous page)

```
  The Corresponding Source for a work in source code form is that
same work.

  2. Basic Permissions.

  All rights granted under this License are granted for the term of
copyright on the Program, and are irrevocable provided the stated
conditions are met.  This License explicitly affirms your unlimited
permission to run the unmodified Program.  The output from running a
covered work is covered by this License only if the output, given its
content, constitutes a covered work.  This License acknowledges your
rights of fair use or other equivalent, as provided by copyright law.

  You may make, run and propagate covered works that you do not
convey, without conditions so long as your license otherwise remains
in force.  You may convey covered works to others for the sole purpose
of having them make modifications exclusively for you, or provide you
with facilities for running those works, provided that you comply with
the terms of this License in conveying all material for which you do
not control copyright.  Those thus making or running the covered works
for you must do so exclusively on your behalf, under your direction
and control, on terms that prohibit them from making any copies of
your copyrighted material outside their relationship with you.

  Conveying under any other circumstances is permitted solely under
the conditions stated below.  Sublicensing is not allowed; section 10
makes it unnecessary.

  3. Protecting Users' Legal Rights From Anti-Circumvention Law.

  No covered work shall be deemed part of an effective technological
measure under any applicable law fulfilling obligations under article
11 of the WIPO copyright treaty adopted on 20 December 1996, or
similar laws prohibiting or restricting circumvention of such
measures.

  When you convey a covered work, you waive any legal power to forbid
circumvention of technological measures to the extent such circumvention
is effected by exercising rights under this License with respect to
the covered work, and you disclaim any intention to limit operation or
modification of the work as a means of enforcing, against the work's
users, your or third parties' legal rights to forbid circumvention of
technological measures.

  4. Conveying Verbatim Copies.

  You may convey verbatim copies of the Program's source code as you
receive it, in any medium, provided that you conspicuously and
appropriately publish on each copy an appropriate copyright notice;
keep intact all notices stating that this License and any
non-permissive terms added in accord with section 7 apply to the code;
keep intact all notices of the absence of any warranty; and give all
recipients a copy of this License along with the Program.

  You may charge any price or no price for each copy that you convey,
and you may offer support or warranty protection for a fee.
```

(continues on next page)

(continued from previous page)

```
  5. Conveying Modified Source Versions.

  You may convey a work based on the Program, or the modifications to
produce it from the Program, in the form of source code under the
terms of section 4, provided that you also meet all of these conditions:

    a) The work must carry prominent notices stating that you modified
    it, and giving a relevant date.

    b) The work must carry prominent notices stating that it is
    released under this License and any conditions added under section
    7.  This requirement modifies the requirement in section 4 to
    "keep intact all notices".

    c) You must license the entire work, as a whole, under this
    License to anyone who comes into possession of a copy.  This
    License will therefore apply, along with any applicable section 7
    additional terms, to the whole of the work, and all its parts,
    regardless of how they are packaged.  This License gives no
    permission to license the work in any other way, but it does not
    invalidate such permission if you have separately received it.

    d) If the work has interactive user interfaces, each must display
    Appropriate Legal Notices; however, if the Program has interactive
    interfaces that do not display Appropriate Legal Notices, your
    work need not make them do so.

  A compilation of a covered work with other separate and independent
works, which are not by their nature extensions of the covered work,
and which are not combined with it such as to form a larger program,
in or on a volume of a storage or distribution medium, is called an
"aggregate" if the compilation and its resulting copyright are not
used to limit the access or legal rights of the compilation's users
beyond what the individual works permit.  Inclusion of a covered work
in an aggregate does not cause this License to apply to the other
parts of the aggregate.

  6. Conveying Non-Source Forms.

  You may convey a covered work in object code form under the terms
of sections 4 and 5, provided that you also convey the
machine-readable Corresponding Source under the terms of this License,
in one of these ways:

    a) Convey the object code in, or embodied in, a physical product
    (including a physical distribution medium), accompanied by the
    Corresponding Source fixed on a durable physical medium
    customarily used for software interchange.

    b) Convey the object code in, or embodied in, a physical product
    (including a physical distribution medium), accompanied by a
    written offer, valid for at least three years and valid for as
    long as you offer spare parts or customer support for that product
    model, to give anyone who possesses the object code either (1) a
    copy of the Corresponding Source for all the software in the
    product that is covered by this License, on a durable physical
```

(continues on next page)

(continued from previous page)

```
    medium customarily used for software interchange, for a price no
    more than your reasonable cost of physically performing this
    conveying of source, or (2) access to copy the
    Corresponding Source from a network server at no charge.

    c) Convey individual copies of the object code with a copy of the
    written offer to provide the Corresponding Source.  This
    alternative is allowed only occasionally and noncommercially, and
    only if you received the object code with such an offer, in accord
    with subsection 6b.

    d) Convey the object code by offering access from a designated
    place (gratis or for a charge), and offer equivalent access to the
    Corresponding Source in the same way through the same place at no
    further charge.  You need not require recipients to copy the
    Corresponding Source along with the object code.  If the place to
    copy the object code is a network server, the Corresponding Source
    may be on a different server (operated by you or a third party)
    that supports equivalent copying facilities, provided you maintain
    clear directions next to the object code saying where to find the
    Corresponding Source.  Regardless of what server hosts the
    Corresponding Source, you remain obligated to ensure that it is
    available for as long as needed to satisfy these requirements.

    e) Convey the object code using peer-to-peer transmission, provided
    you inform other peers where the object code and Corresponding
    Source of the work are being offered to the general public at no
    charge under subsection 6d.

  A separable portion of the object code, whose source code is excluded
from the Corresponding Source as a System Library, need not be
included in conveying the object code work.

  A "User Product" is either (1) a "consumer product", which means any
tangible personal property which is normally used for personal, family,
or household purposes, or (2) anything designed or sold for incorporation
into a dwelling.  In determining whether a product is a consumer product,
doubtful cases shall be resolved in favor of coverage.  For a particular
product received by a particular user, "normally used" refers to a
typical or common use of that class of product, regardless of the status
of the particular user or of the way in which the particular user
actually uses, or expects or is expected to use, the product.  A product
is a consumer product regardless of whether the product has substantial
commercial, industrial or non-consumer uses, unless such uses represent
the only significant mode of use of the product.

  "Installation Information" for a User Product means any methods,
procedures, authorization keys, or other information required to install
and execute modified versions of a covered work in that User Product from
a modified version of its Corresponding Source.  The information must
suffice to ensure that the continued functioning of the modified object
code is in no case prevented or interfered with solely because
modification has been made.

  If you convey an object code work under this section in, or with, or
specifically for use in, a User Product, and the conveying occurs as
part of a transaction in which the right of possession and use of the
```

(continues on next page)

(continued from previous page)

```
User Product is transferred to the recipient in perpetuity or for a
fixed term (regardless of how the transaction is characterized), the
Corresponding Source conveyed under this section must be accompanied
by the Installation Information.  But this requirement does not apply
if neither you nor any third party retains the ability to install
modified object code on the User Product (for example, the work has
been installed in ROM).

  The requirement to provide Installation Information does not include a
requirement to continue to provide support service, warranty, or updates
for a work that has been modified or installed by the recipient, or for
the User Product in which it has been modified or installed.  Access to a
network may be denied when the modification itself materially and
adversely affects the operation of the network or violates the rules and
protocols for communication across the network.

  Corresponding Source conveyed, and Installation Information provided,
in accord with this section must be in a format that is publicly
documented (and with an implementation available to the public in
source code form), and must require no special password or key for
unpacking, reading or copying.

  7. Additional Terms.

  "Additional permissions" are terms that supplement the terms of this
License by making exceptions from one or more of its conditions.
Additional permissions that are applicable to the entire Program shall
be treated as though they were included in this License, to the extent
that they are valid under applicable law.  If additional permissions
apply only to part of the Program, that part may be used separately
under those permissions, but the entire Program remains governed by
this License without regard to the additional permissions.

  When you convey a copy of a covered work, you may at your option
remove any additional permissions from that copy, or from any part of
it.  (Additional permissions may be written to require their own
removal in certain cases when you modify the work.)  You may place
additional permissions on material, added by you to a covered work,
for which you have or can give appropriate copyright permission.

  Notwithstanding any other provision of this License, for material you
add to a covered work, you may (if authorized by the copyright holders of
that material) supplement the terms of this License with terms:

    a) Disclaiming warranty or limiting liability differently from the
    terms of sections 15 and 16 of this License; or

    b) Requiring preservation of specified reasonable legal notices or
    author attributions in that material or in the Appropriate Legal
    Notices displayed by works containing it; or

    c) Prohibiting misrepresentation of the origin of that material, or
    requiring that modified versions of such material be marked in
    reasonable ways as different from the original version; or

    d) Limiting the use for publicity purposes of names of licensors or
    authors of the material; or
```

(continues on next page)

(continued from previous page)

```
    e) Declining to grant rights under trademark law for use of some
    trade names, trademarks, or service marks; or

    f) Requiring indemnification of licensors and authors of that
    material by anyone who conveys the material (or modified versions of
    it) with contractual assumptions of liability to the recipient, for
    any liability that these contractual assumptions directly impose on
    those licensors and authors.

  All other non-permissive additional terms are considered "further
restrictions" within the meaning of section 10.  If the Program as you
received it, or any part of it, contains a notice stating that it is
governed by this License along with a term that is a further
restriction, you may remove that term.  If a license document contains
a further restriction but permits relicensing or conveying under this
License, you may add to a covered work material governed by the terms
of that license document, provided that the further restriction does
not survive such relicensing or conveying.

  If you add terms to a covered work in accord with this section, you
must place, in the relevant source files, a statement of the
additional terms that apply to those files, or a notice indicating
where to find the applicable terms.

  Additional terms, permissive or non-permissive, may be stated in the
form of a separately written license, or stated as exceptions;
the above requirements apply either way.

  8. Termination.

  You may not propagate or modify a covered work except as expressly
provided under this License.  Any attempt otherwise to propagate or
modify it is void, and will automatically terminate your rights under
this License (including any patent licenses granted under the third
paragraph of section 11).

  However, if you cease all violation of this License, then your
license from a particular copyright holder is reinstated (a)
provisionally, unless and until the copyright holder explicitly and
finally terminates your license, and (b) permanently, if the copyright
holder fails to notify you of the violation by some reasonable means
prior to 60 days after the cessation.

  Moreover, your license from a particular copyright holder is
reinstated permanently if the copyright holder notifies you of the
violation by some reasonable means, this is the first time you have
received notice of violation of this License (for any work) from that
copyright holder, and you cure the violation prior to 30 days after
your receipt of the notice.

  Termination of your rights under this section does not terminate the
licenses of parties who have received copies or rights from you under
this License.  If your rights have been terminated and not permanently
reinstated, you do not qualify to receive new licenses for the same
material under section 10.
```

(continues on next page)

(continued from previous page)

```
  9. Acceptance Not Required for Having Copies.

  You are not required to accept this License in order to receive or
run a copy of the Program.  Ancillary propagation of a covered work
occurring solely as a consequence of using peer-to-peer transmission
to receive a copy likewise does not require acceptance.  However,
nothing other than this License grants you permission to propagate or
modify any covered work.  These actions infringe copyright if you do
not accept this License.  Therefore, by modifying or propagating a
covered work, you indicate your acceptance of this License to do so.

  10. Automatic Licensing of Downstream Recipients.

  Each time you convey a covered work, the recipient automatically
receives a license from the original licensors, to run, modify and
propagate that work, subject to this License.  You are not responsible
for enforcing compliance by third parties with this License.

  An "entity transaction" is a transaction transferring control of an
organization, or substantially all assets of one, or subdividing an
organization, or merging organizations.  If propagation of a covered
work results from an entity transaction, each party to that
transaction who receives a copy of the work also receives whatever
licenses to the work the party's predecessor in interest had or could
give under the previous paragraph, plus a right to possession of the
Corresponding Source of the work from the predecessor in interest, if
the predecessor has it or can get it with reasonable efforts.

  You may not impose any further restrictions on the exercise of the
rights granted or affirmed under this License.  For example, you may
not impose a license fee, royalty, or other charge for exercise of
rights granted under this License, and you may not initiate litigation
(including a cross-claim or counterclaim in a lawsuit) alleging that
any patent claim is infringed by making, using, selling, offering for
sale, or importing the Program or any portion of it.

  11. Patents.

  A "contributor" is a copyright holder who authorizes use under this
License of the Program or a work on which the Program is based.  The
work thus licensed is called the contributor's "contributor version".

  A contributor's "essential patent claims" are all patent claims
owned or controlled by the contributor, whether already acquired or
hereafter acquired, that would be infringed by some manner, permitted
by this License, of making, using, or selling its contributor version,
but do not include claims that would be infringed only as a
consequence of further modification of the contributor version.  For
purposes of this definition, "control" includes the right to grant
patent sublicenses in a manner consistent with the requirements of
this License.

  Each contributor grants you a non-exclusive, worldwide, royalty-free
patent license under the contributor's essential patent claims, to
make, use, sell, offer for sale, import and otherwise run, modify and
propagate the contents of its contributor version.
```

(continues on next page)

(continued from previous page)

```
  In the following three paragraphs, a "patent license" is any express
agreement or commitment, however denominated, not to enforce a patent
(such as an express permission to practice a patent or covenant not to
sue for patent infringement).  To "grant" such a patent license to a
party means to make such an agreement or commitment not to enforce a
patent against the party.

  If you convey a covered work, knowingly relying on a patent license,
and the Corresponding Source of the work is not available for anyone
to copy, free of charge and under the terms of this License, through a
publicly available network server or other readily accessible means,
then you must either (1) cause the Corresponding Source to be so
available, or (2) arrange to deprive yourself of the benefit of the
patent license for this particular work, or (3) arrange, in a manner
consistent with the requirements of this License, to extend the patent
license to downstream recipients.  "Knowingly relying" means you have
actual knowledge that, but for the patent license, your conveying the
covered work in a country, or your recipient's use of the covered work
in a country, would infringe one or more identifiable patents in that
country that you have reason to believe are valid.

  If, pursuant to or in connection with a single transaction or
arrangement, you convey, or propagate by procuring conveyance of, a
covered work, and grant a patent license to some of the parties
receiving the covered work authorizing them to use, propagate, modify
or convey a specific copy of the covered work, then the patent license
you grant is automatically extended to all recipients of the covered
work and works based on it.

  A patent license is "discriminatory" if it does not include within
the scope of its coverage, prohibits the exercise of, or is
conditioned on the non-exercise of one or more of the rights that are
specifically granted under this License.  You may not convey a covered
work if you are a party to an arrangement with a third party that is
in the business of distributing software, under which you make payment
to the third party based on the extent of your activity of conveying
the work, and under which the third party grants, to any of the
parties who would receive the covered work from you, a discriminatory
patent license (a) in connection with copies of the covered work
conveyed by you (or copies made from those copies), or (b) primarily
for and in connection with specific products or compilations that
contain the covered work, unless you entered into that arrangement,
or that patent license was granted, prior to 28 March 2007.

  Nothing in this License shall be construed as excluding or limiting
any implied license or other defenses to infringement that may
otherwise be available to you under applicable patent law.

  12. No Surrender of Others' Freedom.

  If conditions are imposed on you (whether by court order, agreement or
otherwise) that contradict the conditions of this License, they do not
excuse you from the conditions of this License.  If you cannot convey a
covered work so as to satisfy simultaneously your obligations under this
License and any other pertinent obligations, then as a consequence you may
not convey it at all.  For example, if you agree to terms that obligate you
to collect a royalty for further conveying from those to whom you convey
```

(continues on next page)

(continued from previous page)

```
the Program, the only way you could satisfy both those terms and this
License would be to refrain entirely from conveying the Program.

  13. Remote Network Interaction; Use with the GNU General Public License.

  Notwithstanding any other provision of this License, if you modify the
Program, your modified version must prominently offer all users
interacting with it remotely through a computer network (if your version
supports such interaction) an opportunity to receive the Corresponding
Source of your version by providing access to the Corresponding Source
from a network server at no charge, through some standard or customary
means of facilitating copying of software.  This Corresponding Source
shall include the Corresponding Source for any work covered by version 3
of the GNU General Public License that is incorporated pursuant to the
following paragraph.

  Notwithstanding any other provision of this License, you have
permission to link or combine any covered work with a work licensed
under version 3 of the GNU General Public License into a single
combined work, and to convey the resulting work.  The terms of this
License will continue to apply to the part which is the covered work,
but the work with which it is combined will remain governed by version
3 of the GNU General Public License.

  14. Revised Versions of this License.

  The Free Software Foundation may publish revised and/or new versions of
the GNU Affero General Public License from time to time.  Such new versions
will be similar in spirit to the present version, but may differ in detail to
address new problems or concerns.

  Each version is given a distinguishing version number.  If the
Program specifies that a certain numbered version of the GNU Affero General
Public License "or any later version" applies to it, you have the
option of following the terms and conditions either of that numbered
version or of any later version published by the Free Software
Foundation.  If the Program does not specify a version number of the
GNU Affero General Public License, you may choose any version ever published
by the Free Software Foundation.

  If the Program specifies that a proxy can decide which future
versions of the GNU Affero General Public License can be used, that proxy's
public statement of acceptance of a version permanently authorizes you
to choose that version for the Program.

  Later license versions may give you additional or different
permissions.  However, no additional obligations are imposed on any
author or copyright holder as a result of your choosing to follow a
later version.

  15. Disclaimer of Warranty.

  THERE IS NO WARRANTY FOR THE PROGRAM, TO THE EXTENT PERMITTED BY
APPLICABLE LAW.  EXCEPT WHEN OTHERWISE STATED IN WRITING THE COPYRIGHT
HOLDERS AND/OR OTHER PARTIES PROVIDE THE PROGRAM "AS IS" WITHOUT WARRANTY
OF ANY KIND, EITHER EXPRESSED OR IMPLIED, INCLUDING, BUT NOT LIMITED TO,
THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR
```

(continues on next page)

(continued from previous page)

```
PURPOSE.  THE ENTIRE RISK AS TO THE QUALITY AND PERFORMANCE OF THE PROGRAM
IS WITH YOU.  SHOULD THE PROGRAM PROVE DEFECTIVE, YOU ASSUME THE COST OF
ALL NECESSARY SERVICING, REPAIR OR CORRECTION.

  16. Limitation of Liability.

  IN NO EVENT UNLESS REQUIRED BY APPLICABLE LAW OR AGREED TO IN WRITING
WILL ANY COPYRIGHT HOLDER, OR ANY OTHER PARTY WHO MODIFIES AND/OR CONVEYS
THE PROGRAM AS PERMITTED ABOVE, BE LIABLE TO YOU FOR DAMAGES, INCLUDING ANY
GENERAL, SPECIAL, INCIDENTAL OR CONSEQUENTIAL DAMAGES ARISING OUT OF THE
USE OR INABILITY TO USE THE PROGRAM (INCLUDING BUT NOT LIMITED TO LOSS OF
DATA OR DATA BEING RENDERED INACCURATE OR LOSSES SUSTAINED BY YOU OR THIRD
PARTIES OR A FAILURE OF THE PROGRAM TO OPERATE WITH ANY OTHER PROGRAMS),
EVEN IF SUCH HOLDER OR OTHER PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF
SUCH DAMAGES.

  17. Interpretation of Sections 15 and 16.

  If the disclaimer of warranty and limitation of liability provided
above cannot be given local legal effect according to their terms,
reviewing courts shall apply local law that most closely approximates
an absolute waiver of all civil liability in connection with the
Program, unless a warranty or assumption of liability accompanies a
copy of the Program in return for a fee.

                     END OF TERMS AND CONDITIONS

            How to Apply These Terms to Your New Programs

  If you develop a new program, and you want it to be of the greatest
possible use to the public, the best way to achieve this is to make it
free software which everyone can redistribute and change under these terms.

  To do so, attach the following notices to the program.  It is safest
to attach them to the start of each source file to most effectively
state the exclusion of warranty; and each file should have at least
the "copyright" line and a pointer to where the full notice is found.

    rna_cd
    Copyright (C) 2018  Leiden University Medical Center

    This program is free software: you can redistribute it and/or modify
    it under the terms of the GNU Affero General Public License as published
    by the Free Software Foundation, either version 3 of the License, or
    (at your option) any later version.

    This program is distributed in the hope that it will be useful,
    but WITHOUT ANY WARRANTY; without even the implied warranty of
    MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.  See the
    GNU Affero General Public License for more details.

    You should have received a copy of the GNU Affero General Public License
    along with this program.  If not, see <http://www.gnu.org/licenses/>.

Also add information on how to contact you by electronic and paper mail.

  If your software can interact with users remotely through a computer
```

(continues on next page)

(continued from previous page)

```
network, you should also make sure that it provides a way for users to
get its source.  For example, if your program is a web application, its
interface could display a "Source" link that leads users to an archive
of the code.  There are many ways you could offer source, and different
solutions will be better for different programs; see section 13 for the
specific requirements.

  You should also get your employer (if you work as a programmer) or school,
if any, to sign a "copyright disclaimer" for the program, if necessary.
For more information on this, and how to apply and follow the GNU AGPL, see
<http://www.gnu.org/licenses/>.
```

CHAPTER 2

API documentation

2.1 API documentation

2.1.1 bam_process

rna_cd.bam_process.**chop_contig**(*size: int*, *chunksize: int*) → Iterator[Tuple[int, int]]
For a contig of given size, generate regions maximally chunksize long. We use _0_ based indexing

rna_cd.bam_process.**softclip_bases**(*reader: pysam.libcalignmentfile.AlignmentFile, contig: str, region: Tuple[int, int]*) → int
Calculate amount of softclip bases for a region

rna_cd.bam_process.**coverage**(*reader: pysam.libcalignmentfile.AlignmentFile, contig: str, region: Tuple[int, int], method: Callable = <function mean>*) → float
Calculate average/median/etc coverage for a region

rna_cd.bam_process.**process_bam**(*path: pathlib.Path*, *chunksize: int = 100*, *contig: str = 'chrM'*) → numpy.ndarray
Process bam file to an ndarray

Returns numpy ndarray of shape (n_features,)

rna_cd.bam_process.**make_array_set**(*bam_files: List[pathlib.Path], labels: List[Any], chunksize: int = 100, contig: str = 'chrM', cores: int = 1*) → Tuple[numpy.ndarray, numpy.ndarray]
Make set of numpy arrays corresponding to data and labels. I.e. train/testX and train/testY in scikit-learn parlance.

Parameters

- **bam_files** – List of paths to bam files
- **labels** – list of labels.
- **cores** – number of cores to use for processing

Returns tuple of X and Y numpy arrays. X has shape (n_files, n_features). Y has shape (n_files,).

2.1.2 cli

rna_cd.cli.**directory_callback**(*ctx*, *param*, *value*)
Click callback function for getting bam/cram files from a directory.

rna_cd.cli.**list_callback**(*ctx*, *param*, *value*)
Click callback function for getting bam/cram files from a list file.

rna_cd.cli.**path_callback**(*ctx*, *param*, *value*)
Generic str to path callback. To be used for click.Path types that ought to return pathlib.Path

rna_cd.cli.**unknown_threshold_callback**(*ctx*, *param*, *value*)
Click callback function for threshold that has to be between 0.5 and 1.0

2.1.3 models

class rna_cd.models.**PredClass**
An enumeration.

rna_cd.models.**plot_pca**(*searcher: sklearn.model_selection._search.GridSearchCV*, *arr_X: numpy.ndarray*, *arr_Y: numpy.ndarray*, *img_out: pathlib.Path*) → None
Plot PCA with training samples of pipeline.

rna_cd.models.**predict_labels_and_prob**(*model*, *bam_files: List[pathlib.Path]*, *chunksize: int = 100*, *contig: str = 'chrM'*, *cores: int = 1*, *unknown_threshold: float = 0.75*) → List[rna_cd.models.Prediction]
Predict labels and probabilities for a list of bam files.

Parameters **unknown_threshold** – The probability threshold below which samples are considered to be 'unknown'. Must be between 0.5 and 1.0

Returns list of Prediction classes

rna_cd.models.**train_svm_model**(*positive_bams: List[pathlib.Path]*, *negative_bams: List[pathlib.Path]*, *chunksize: int = 100*, *contig: str = 'chrM'*, *cross_validations: int = 3*, *verbosity: int = 1*, *cores: int = 1*, *plot_out: Optional[pathlib.Path] = None*) → sklearn.model_selection._search.GridSearchCV
Run SVM training on a list of positive BAM files (i.e. _with_ contamination) and a list of negative BAM files (i.e. _without_ contamination).

For all bam files features are collected over one contig. This contig is binned, and for each bin two different metrics of coverage are collected, in addition to the softclip rate.

These features are then fed to a sklearn pipeline with three steps:

1. A scaling step using StandardScaler
2. A dimensional reduction step using PCA.
3. A classification step using an SVM.

Hyperparameters are tuned using a grid search with cross validations.

Optionally saves a plot of the top two PCA components with the training samples.

Parameters

- **positive_bams** – List of BAM files with contaminations
- **negative_bams** – List of BAM files without contaminations.

- **chunksize** – The size in bases for each chunk (bin)
- **contig** – The name of the contig.
- **cross_validations** – The amount of cross validations
- **verbosity** – Verbosity parameter of sklearn. Increase to see more messages.
- **cores** – Amount of cores to use for both metric collection and training.
- **plot_out** – Optional path for PCA plot.

Returns GridSearchCV object containing tuned pipeline.

2.1.4 utils

rna_cd.utils.**dir_to_bam_list**(*path: pathlib.Path*) → List[pathlib.Path]
Load a directory containing bam or cram files

rna_cd.utils.**echo**(*msg: str*)
Wrapper around click.secho to include datetime

rna_cd.utils.**load_list_file**(*path: pathlib.Path*) → List[pathlib.Path]
Load a file containing containing a list of files

rna_cd.utils.**load_sklearn_object_from_disk**(*path: pathlib.Path*) → Any
Load a JSON-serialized object from disk

rna_cd.utils.**save_sklearn_object_to_disk**(*obj: Any*, *path: pathlib.Path*)
Save an object with some metadata to disk as serialized JSON

CHAPTER 3

Indices and tables

- genindex
- modindex
- search

Python Module Index

r

Index

Symbols

C

D

E

L

M

P

R

S

T

U