RNA2DNAlign manual 1.0.11

https://github.com/HorvathLab/NGS/tree/master/RNA2DNAlign

Feb 17, 2016

Contents

GETTING STARTED

RNA2DNAlign evaluates evidence for asymmetric allele distribution in next-gen sequencing reads of DNA and RNA samples from the same individual. RNA2DNAlign requires, as input: genome aligned reads and SNV loci to analyze. Reads from each analysis type and sample must be aligned to the same version of the human genome reference. SNVs may be derived from the reads directly, using, for example, Samtools, or they may be derived from independent sources, such as lists of known annotated variants. Variant positions must correspond to genomic coordinates of the reference genome used for the alignment.

RNA2DNAlign is available as a self-contained binary package for 64-bit Linux systems and as Python source. The pysam package, plus a variety of common third-party python packages including numpy and scipy must be installed to use in Python source form. The self-contained binary package is appropriate for most Linux users.

INSTALLATION

RNA2DNAlign is available in two forms, a self-contained packaged binary for 64-bit Linux systems, and as Python source. We recommend the self-contained packaged binary for Linux systems.

RNA2DNAlign for 64-bit Linux

1. Unpack the downloaded release:

   ```
   tar xzf RNA2DNAlign-*.tgz

   ln -s RNA2DNAlign-* RNA2DNAlign
   ```

2. The RNA2DNAlign program is located in the bin subdirectory:

   ```
   ./RNA2DNAlign/bin/RNA2DNAlign -h

   ./RNA2DNAlign/bin/RNA2DNAlign
   ```

3. Test the install using the provided example data:

   ```
   cd RNA2DNAlign

   ./bin/RNA2DNAlign -r "data/example-*.bam" -s "data/example-SNV.tsv" -o testing
   ```

Python 2.7 RNA2DNAlign

1. Unpack the downloaded release:

   ```
   tar xzf RNA2DNAlign-*.tgz

   ln -s RNA2DNAlign-* RNA2DNAlign
   ```

2. Locate your Python binary and ensure it is version 2.7:

   ```
   python --version

   /path/to/python2.7 --version
   ```

   We refer to the Python binary as `python` below, please substitute whatever path and version numbers are required to run Python 2.7 on your system. We recommend the Enthought Python Distribution (EPD) which pre-installs all but the pysam third-party dependencies needed by RNA2DNAlign.

3. Ensure the necessary third-party Python modules are installed. pysam version >= 0.8.1 is required.

   ```
   pysam, numpy, scipy
   ```

   For the configuration and execution GUI (optional):

   ```
   wx
   ```

   For Excel format SNV input files (optional):

   ```
   xlrd, openpyxl
   ```

   The existence of required modules can be tested as follows (demonstrated here for `scipy`):

   ```
     python2.7 -c "from scipy import __version__; print __version__"
   ```

4. The RNA2DNAlign program is located in the src subdirectory:

   ```
   python ./RNA2DNAlign/src/RNA2DNAlign.py -h

   python ./RNA2DNAlign/src/RNA2DNAlign.py
   ```

5. Test the installation using the provided example data:

   ```
   cd RNA2DNAlign python

   ./src/RNA2DNAlign.py -r "data/example-*.bam" -s "data/example-SNV.tsv" -o testing
   ```

RNA2DNAlign USAGE

Synopsis

Graphical User Interface:

```
RNA2DNAlign

RNA2DNAlign.py
```

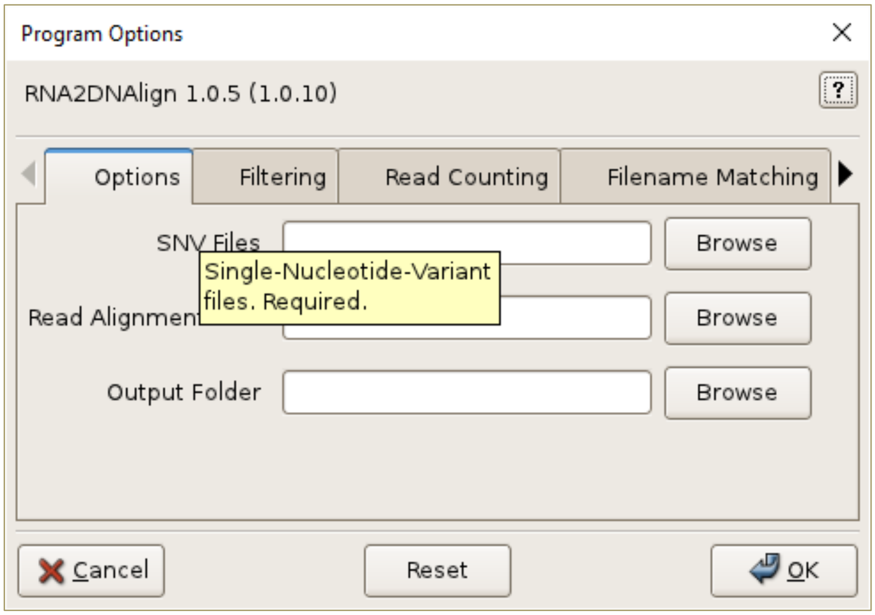
Command-line:

```
RNA2DNAlign [options]

RNA2DNAlign.py [options]
```

Graphical User Interface

Click the help icon (question mark) at the top right of the GUI and then an input field for help. Multiple files can be selected in the file-chooser using Ctrl-Click or Shift-Click. Fields can be reset to their default values using the Reset button. Click OK to execute RNA2DNAlign.



Single-Nucleotide-Variant files. Required.

Additional GUI option tabs are documented below.

Options

SNVs, -s SNVS, --snvs=SNVS
Single-nucleotide-polymophisms (SNVs). Tabular and VCF format SNVs are supported. Multiple files are specified inside quotes, separated by spaces, and by using file globbing. See Input Files for more information. Required.

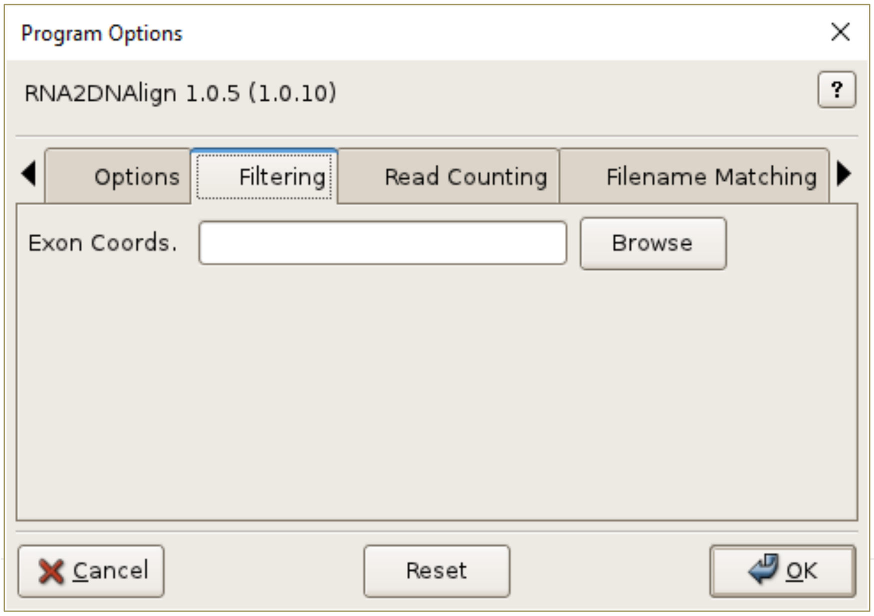
Read Alignment Files, -r ALIGNMENTS, --readalignments=ALIGNMENTS
Read alignments files in indexed BAM format, with extension `.bam`. BAM index with extension `.bam.bai` must be located in the same directory. Multiple files are specified inside quotes, separated by spaces, and by using file globbing. See Input Files for more information. Required.

Output Folder, -o OUTPUT, --output=OUTPUT
Output directory. Will be created if necessary. Files inside this directory will be overwritten by program output. See Output Files for more information on output files. Required.

--version
Show program's version number and exit.

-h, --help
Show program help and exit.

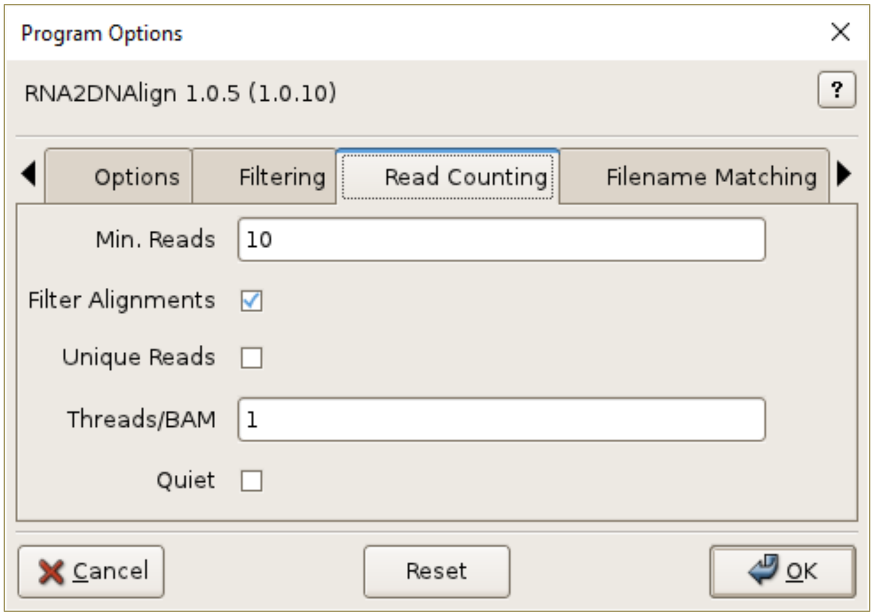
Filtering



Exon Coords.

Exon Coords., -e EXONCOORDS, --exoncoords=EXONCOORDS
Exon coordinates to filter out non-exonic SNVs. Use of exon coordinates to filter the SNVs is strongly recommended for semantic and performance reasons for transcriptome-to-exome types of analyses. See Annotation Files for format and download information. Optional.

Read Counting



RNA2DNAlign 1.0.5 (1.0.10)

Min. Reads: 10

Filter Alignments: ☑

Unique Reads: □

Threads/BAM: 1

Quiet: □

Min. Reads, -m MINREADS, --minreads=MINREADS
Minimum number of good reads at each SNV locus per alignment file. Default=10.

Filter Alignments, -f, --alignmentfilter
(Turn off) alignment filtering by length, edits, etc.

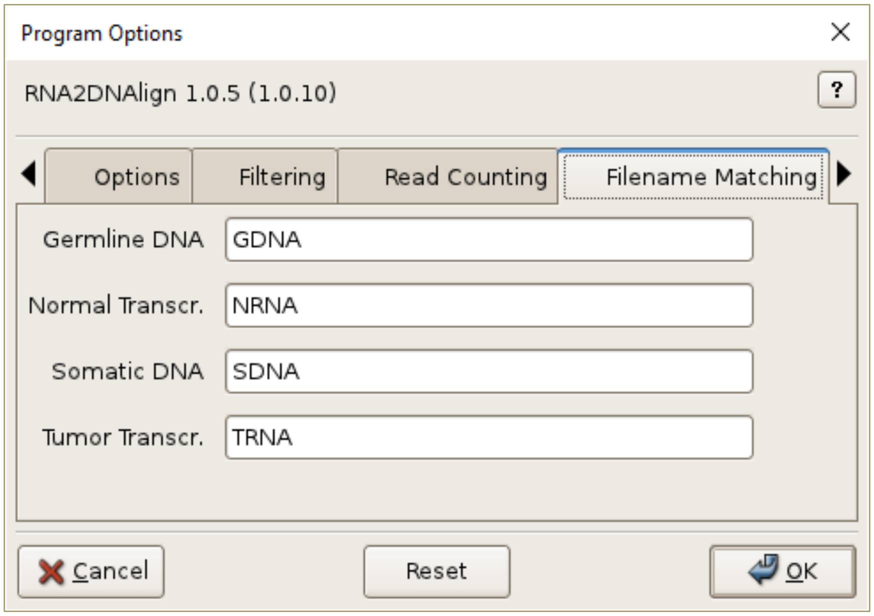
Unique Reads, -U, --uniquereads
Consider only distinct reads.

Threads/BAM, -t TPB, --threadsperbam=TPB
Worker threads per alignment file. Indicate no threading with 0. Default=1.

Quiet, -q, --quiet
Do not show readCounts progress.

Filename Matching

Unique string of 4 characters from the name of the corresponding file.



Program Options

RNA2DNAlign 1.0.5 (1.0.10)

Germline DNA: GDNA

Normal Transcr.: NRNA

Somatic DNA: SDNA

Tumor Transcr.: TRNA

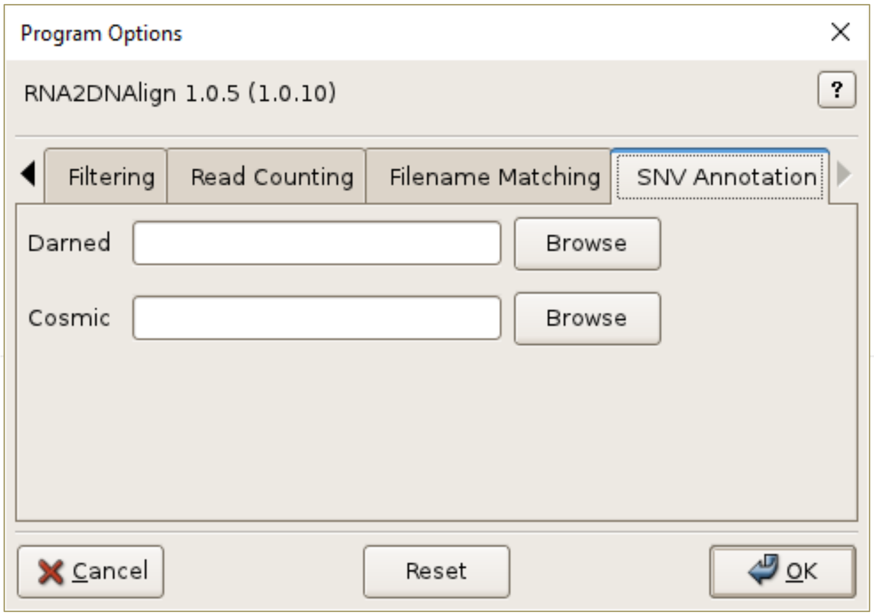
Germline DNA, --normaldnare=NORMALDNARE
Germline/Normal DNA filename regular expression. Default: GDNA.

Normal Transcr., --normaltransre=NORMALTRANSRE
Normal transcriptome filename regular expression. Default: NRNA.

Somatic DNA, --tumordnare=TUMORDNARE
Somatic/Tumor DNA filename regular expression. Default: SDNA.

Tumor Transcr., --tumortransre=TUMORTRANSRE
Tumor transcriptome filename regular expression. Default: TRNA.

Annotation



Darned, -d DARNED, --darned=DARNED
DARNED Annotations. See Annotation Files for format and download information. Optional.

Cosmic, -c COSMIC, --cosmic=COSMIC
COSMIC Annotations. See Annotation Files for format and download information. Optional.

Input Files

All input files - SNV loci, read alignments, and annotation files - must indicate genomic position with respect to the same specific release of a common reference genome.

SNVs

Single-nucleotide-variants (SNVs) in tabular or VCF format. Tabular formats and their required extensions include whitespace separated text files (`.txt`), tab-separated values files (`.tsv`), comma-separated values files (`.csv`), Excel (`.xlsx`), and Excel 2003 (`.xls`).
Text files must have four white-space separated columns representing the chromosome (CHROM), locus (POS), wild-type allele nucleotide (REF), and SNV nucleotide (ALT). Other tabular formats must provide CHROM, POS, REF, ALT headings.

Alignments

Read alignment files in indexed BAM format. Filename extension `.bam` expected with `.bam.bai` index files **in the same location/folder.** RNA2DNAlign will execute fastest if all BAM files are sorted and indexed in a consistent manner. Alignments produced using BWA, Bowtie, TopHat and STAR have been tested as inputs for RNA2DNAlign.

Output Files

RNA2DNAlign output files are created in the directory specified. The folder will be created if necessary. Existing files will be overwritten.

Summary File

The `summary_result.txt` file summarizes the count of each type of event observed.
See example output files in the `RNA2DNAlign/data` directory.

Event Files

Each execution of RNA2DNAlign will create (up to) eight tab-separated value event files representing the following events: RNA editing (`Events_RNAed.tsv`), tumor-specific RNA editing (`Events_T-RNAed.tsv`), variant-specific expression (`Events_VSE.tsv`) or loss (`Events_VSL.tsv`), tumor-specific variant expression (`Events_T-VSE.tsv`) or loss (`Events_T-VSL.tsv`), somatic mutagenesis (`Events_SOM.tsv`), and loss of heterozygosity (`Events_LOH.tsv`).
See example output files in the `RNA2DNAlign/data/example-output` directory.

Event File Fields

Each Event file contains the SNV positions with allele dispersion matching the corresponding event. The information for each SNV is contained in consecutive rows corresponding to the analyzed matching datasets (4 rows in the case of normal/tumor/exome/transcriptome analyses). Event files contain the following fields:

AlignedReads
Name of the aligned reads file for the following read counts.

CHROM
Chromosome identifier

POS
Chromosome position of the variant

REF
Reference allele nucleotide

ALT
Variant allele nucleotide

SNPCountForward
Number of forward oriented variant reads in the paired end alignment.

SNPCountReverse

Number of reverse oriented variant reads in the paired end alignment.

RefCountForward
Number of forward oriented reference reads in the paired end alignment.

RefCountReverse
Number of reverse oriented reference reads in the paired end alignment.

SNPCount
Total number of variant reads.

RefCount
Total number of reference reads.

R
Proportion of variant reads.

HomoVarSc
Score of locus as homozygous variant.

HetSc
Score of locus as heterozygous reference and variant.

HomoRefSc
Score of locus as homozygous reference.

VarDomSc
Score of locus as dominant for the variant allele.

RefDomSc
Score of locus as dominant for the reference allele.

NotHomoVarpV
p-Value of read counts with respect to homozygous variant null model.

NotHomoRefpV
p-Value of read counts with respect to homozygous reference null model.

NotHetpV
p-Value of read counts with respect to heterozygous reference and variant null model.

VarDompV
p-Value of increased variant read counts with respect to heterozygous reference and variant null model.

RefDompV

p-Value of increased reference read counts with respect to heterozygous reference and variant null model.

NotHomoVarFDR
Multiple-test corrected FDR significance of read counts with respect to homozygous variant null model.

NotHomoRefFDR
Multiple-test corrected FDR significance of read counts with respect to homozygous reference null model.

NotHetFDR
Multiple-test corrected FDR significance of read counts with respect to heterozygous reference and variant null model.

VarDomFDR
Multiple-test corrected FDR significance of increased variant read counts with respect to heterozygous reference and variant null model.

RefDomFDR
Multiple-test corrected FDR significance of increased reference read counts with respect to heterozygous reference and variant null model.

Read Counts

A tab-separated values file consisting of the computed read-counts is also provided (`readCounts.tsv`). This file contains the read counts for each SNV locus in each BAM file and computes the various statistical tests described above, in "Event File Fields". The read counts file can be used to investigate the computed values for expected events that didn't pass filtering, significance, or scoring thresholds.

Annotation Files

All annotation files, SNV loci, and read alignments must indicate genomic position with respect to the same specific release of a common reference genome.
The following files and instructions are for hg19/grch37/NCBI37.

RefSeq Exon Coordinates (UCSC)

Use of exon coordinates for transcriptome-to-exome comparisons is strongly recommended.

1. RefSeq exon coordinates are downloaded from the UCSC genome browser and provided in the RNA2DNAlign/data directory in the file: `UCSC_Human_hg19_RefSeq_CDS_exon_coordinates.txt`. The exon coordinates file can be used as provided.

2. RefSeq exon coordinates can be recreated as follows:

   ```
   cd data
   ./dlexons.sh hg19 > UCSC_Human_hg19_RefSeq_CDS_exon_coordinates.txt
   ```

   Exon coordinates should be tab-separated and sorted by chromosome number (1,2,3,...,X,Y), start position, end position, in increasing order.

COSMIC

COSMIC mutations are downloaded from the COSMIC website and provided in the RNA2DNAlign/data directory in the file: `CosmicMutantExport_hg19.tsv.gz`. The annotation file can be used as provided.

COSMIC mutations can be downloaded as follows:

- Register with COSMIC

  ```
  https://cancer.sanger.ac.uk/cosmic/register
  ```

- Download the COSMIC mutants:

  ```
  sftp "login"@sftp-
  cancer.sanger.ac.uk:/cosmic/grch37/cosmic/v75/CosmicMutantExport.tsv.gz
  ```

- COSMIC annotations can be used in its downloaded format.

DARNED

1. DARNED loci are downloaded form the DARNED website and provided in the RNA2DNAlign/data directory in the file:`DARNED_hg19.txt`. The annotation file can be used as provided.

2. DARNED loci can be downloaded for another assembly as follows:

   - Download the DARNED loci:

   ```
   http://darned.ucc.ie/static/downloads/hg19.txt
   ```

   - DARNED loci can be used in its downloaded format.

Examples

Command-line

Example 1: BAM files and single SNV file in TSV format.

```
cd RNA2DNAlign/data

../bin/RNA2DNAlign -r "example-*.bam" -s "example-SNV.tsv" -o example1
```

Example 2: BAM and VCF files for each dataset, exonic SNV filtering, and DARNED and COSMIC annotations using the supplied annotation files.

```
cd RNA2DNAlign/data

../bin/RNA2DNAlign -r "example-GDNA.bam example-NRNA.bam example-SDNA.bam example-TRNA.bam"
-s "example-*.vcf" -o example2 -e UCSC_Human_hg19_RefSeq_CDS_exon_coordinates.txt -d
DARNED_hg19.txt -c CosmicMutantExport_hg19.tsv.gz
```

Result files corresponding to this analysis are available in the `RNA2DNAlign/data` directory in the `example-output` directory.

Example 3: BAM and VCF files for each dataset, exonic SNV filtering, and minimum reads per loci in each dataset of 3.

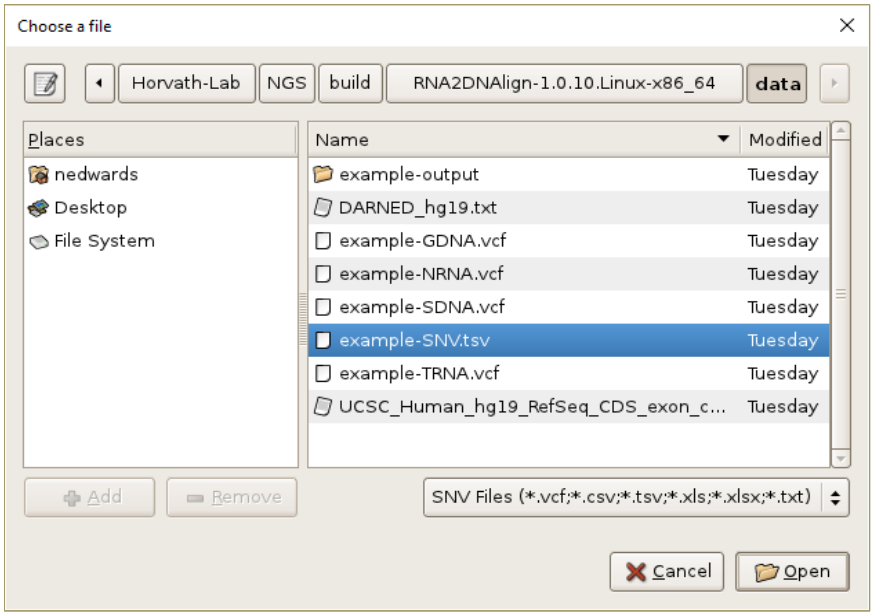
```
cd RNA2DNAlign/data

../bin/RNA2DNAlign -r "example-*.bam" -s "example-*.vcf" -o example3 -e
UCSC_Human_hg19_RefSeq_CDS_exon_coordinates.txt -m 3
```

Examples: Graphical User Interface

Example 1: BAM files and single SNV file in TSV format.

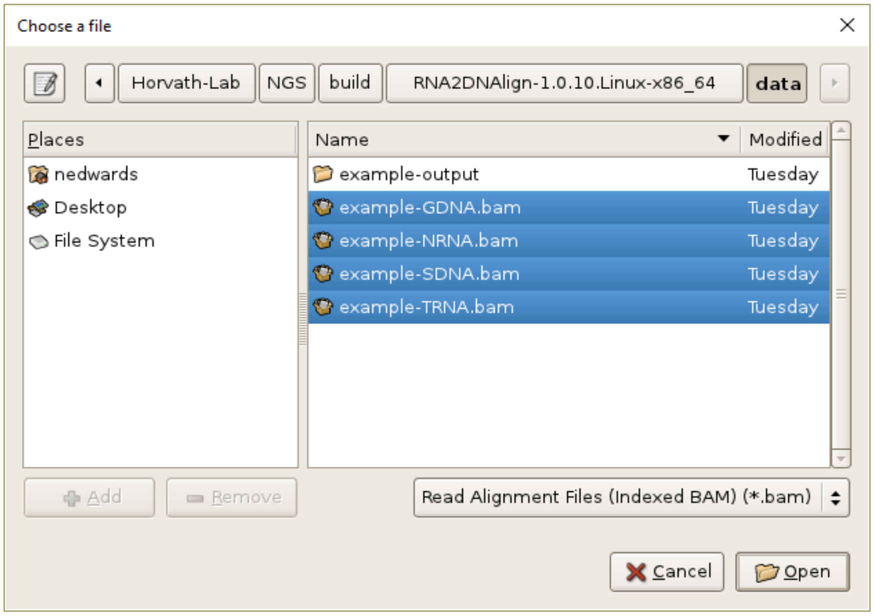
1. Select the SNV file by clicking on the `Browse` button, navigating to `RNA2DNAlign/data`, selecting `example-SNV.tsv`, and clicking `OK`.



Name	Modified
example-output	Tuesday
DARNED_hg19.txt	Tuesday
example-GDNA.vcf	Tuesday
example-NRNA.vcf	Tuesday
example-SDNA.vcf	Tuesday
example-SNV.tsv	Tuesday
example-TRNA.vcf	Tuesday
UCSC_Human_hg19_RefSeq_CDS_exon_c...	Tuesday

SNV Files (*.vcf;*.csv;*.tsv;*.xls;*.xlsx;*.txt)

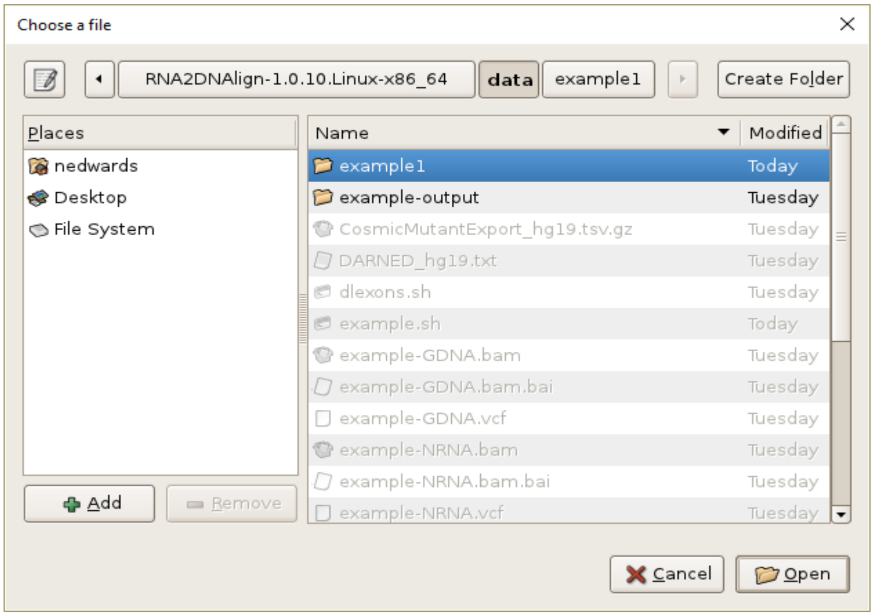
2. Select the BAM files by clicking on the `Browse` button, navigating to `RNA2DNAlign/data`, selecting all the BAM files, using shift-click or control-click as needed, and clicking `OK`.



Name	Modified
example-output	Tuesday
example-GDNA.bam	Tuesday
example-NRNA.bam	Tuesday
example-SDNA.bam	Tuesday
example-TRNA.bam	Tuesday

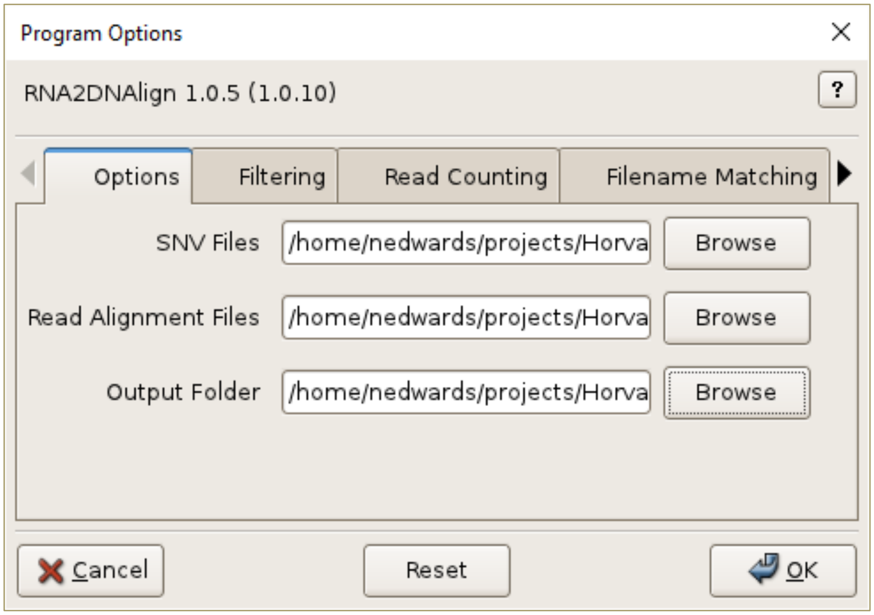
Read Alignment Files (Indexed BAM) (*.bam)

3. Specify the output directory by clicking on the `Browse` button, navigating to `RNA2DNAlign/data`, clicking `Create Folder`, entering "example1", and clicking `Open`.



Name	Modified
example1	Today
example-output	Tuesday
CosmicMutantExport_hg19.tsv.gz	Tuesday
DARNED_hg19.txt	Tuesday
dlexons.sh	Tuesday
example.sh	Today
example-GDNA.bam	Tuesday
example-GDNA.bam.bai	Tuesday
example-GDNA.vcf	Tuesday
example-NRNA.bam	Tuesday
example-NRNA.bam.bai	Tuesday
example-NRNA.vcf	Tuesday

4. Click `OK` to execute the program.



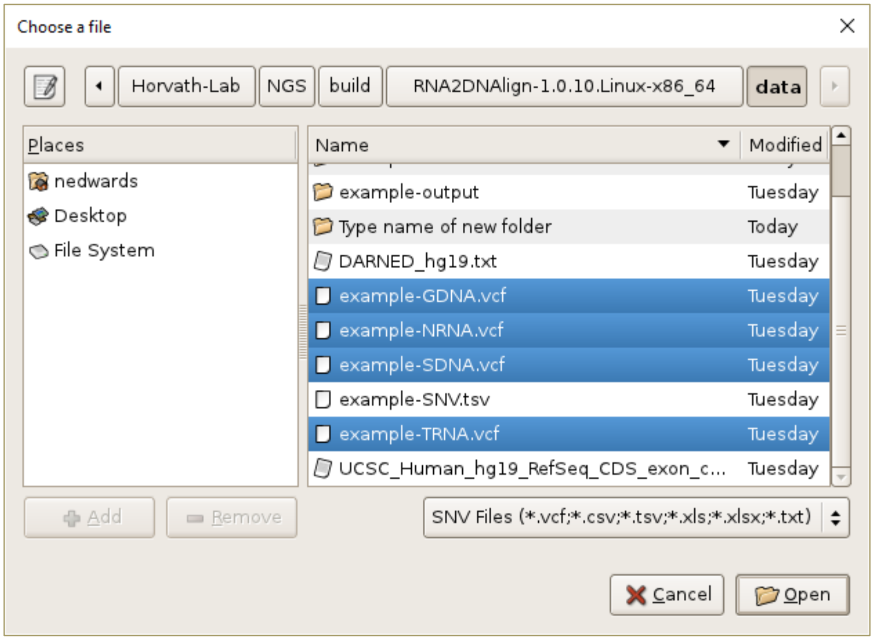
SNV Files /home/nedwards/projects/Horva

Read Alignment Files /home/nedwards/projects/Horva

Output Folder /home/nedwards/projects/Horva

Example 2: BAM and VCF files for each dataset, exonic SNV filtering, and DARNED and COSMIC annotations using the supplied annotation files.

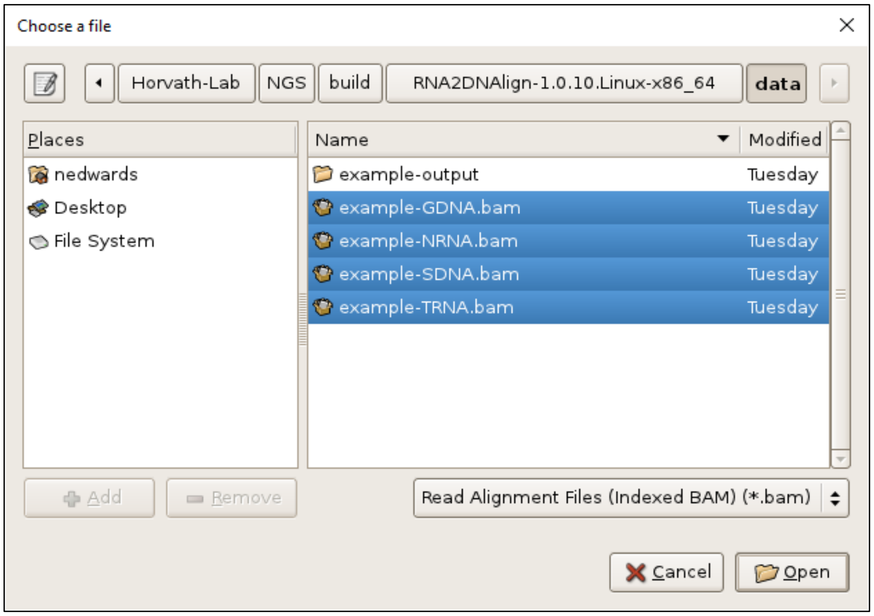
1. Select the VCF files by clicking on the `Browse` button, navigating to `RNA2DNAlign/data`, selecting all the VCF files, using shift-click or control-click as needed, and clicking `OK`.



Name	Modified
example-output	Tuesday
Type name of new folder	Today
DARNED_hg19.txt	Tuesday
example-GDNA.vcf	Tuesday
example-NRNA.vcf	Tuesday
example-SDNA.vcf	Tuesday
example-SNV.tsv	Tuesday
example-TRNA.vcf	Tuesday
UCSC_Human_hg19_RefSeq_CDS_exon_c...	Tuesday

SNV Files (*.vcf;*.csv;*.tsv;*.xls;*.xlsx;*.txt)

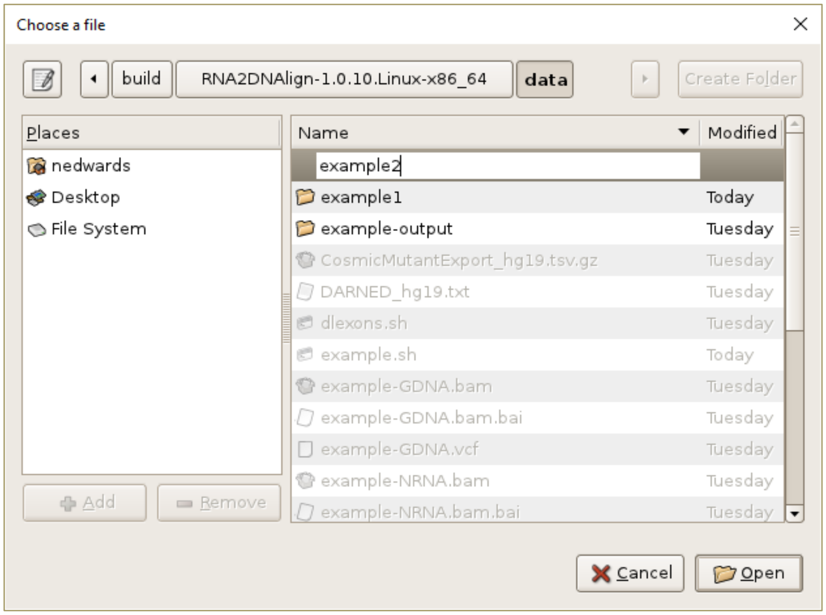
2. Select the BAM files by clicking on the `Browse` button, navigating to `RNA2DNAlign/data`, selecting all the BAM files, using shift-click or control-click as needed, and clicking `OK`.



Name	Modified
example-output	Tuesday
example-GDNA.bam	Tuesday
example-NRNA.bam	Tuesday
example-SDNA.bam	Tuesday
example-TRNA.bam	Tuesday

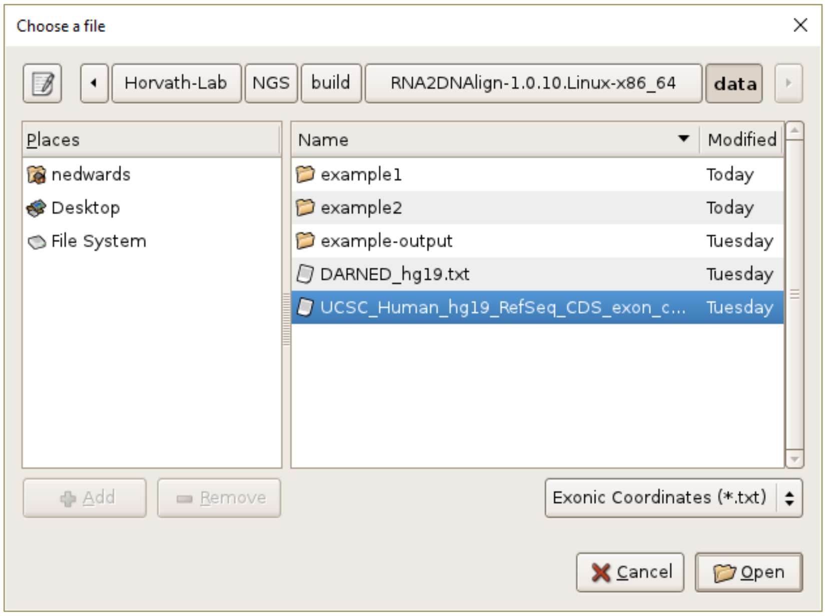
Read Alignment Files (Indexed BAM) (*.bam)

3 Specify the output directory by clicking on the `Browse` button, navigating to `RNA2DNAlign/data`, clicking `Create Folder`, entering "example2" and clicking `Open`.



Name	Modified
example2	
example1	Today
example-output	Tuesday
CosmicMutantExport_hg19.tsv.gz	Tuesday
DARNED_hg19.txt	Tuesday
dlexons.sh	Tuesday
example.sh	Today
example-GDNA.bam	Tuesday
example-GDNA.bam.bai	Tuesday
example-GDNA.vcf	Tuesday
example-NRNA.bam	Tuesday
example-NRNA.bam.bai	Tuesday

4 Specify exonic SNV filtering by selecting the `Filtering` tab, clicking on the `Browse` button, navigating to`RNA2DNAlign/data`, selecting "UCSC_Human_hg19_RefSeq_CDS_exon_coordinates.txt", and clicking `OK`.



Name	Modified
example1	Today
example2	Today
example-output	Tuesday
DARNED_hg19.txt	Tuesday
UCSC_Human_hg19_RefSeq_CDS_exon_c...	Tuesday

Exonic Coordinates (*.txt)

5 Specify DARNED and COSMIC annotation of SNP events on the `SNV Annotation` tab, selecting the files`DARNED_hg19.txt` and `CosmicMutantExport_hg19.tsv.gz`.

6 Click `OK` to execute the program. Result files corresponding to this analysis are available in the RNA2DNAlign/data directory in the example-output directory.