iAnnotateSV Documentation

Release 0.0.1

Ronak Hasmukh Shah

March 04, 2015

Author Ronak H Shah

Contact rons.shah@gmail.com

Source code http://github.com/rhshah/iAnnotateSV

License Apache License 2.0

iAnnotateSV is a Python library and command-line software toolkit to annotate and visualize structural variants detected from Next Generation DNA sequencing data. This works for majority is just a re-writing of a tool called dRanger_annotate written in matlab by Mike Lawrence at Broad Institue. But it also has some additional functionality and control over the annotation w.r.t the what transcripts to be used for annotation. It is designed for use with hybrid capture, including both whole-exome and custom target panels, and short-read sequencing platforms such as Illumina.

Contents:

IANNOTATESV

1.1 iAnnotateSV package

1.1.1 Module `iAnnotateSV` contents

1.1.2 Submodules

1.1.3 `AnnotateEachBreakpoint` module

- **This module will annotate each breakpoint taking in:**
    - **chr** : chromosome for the event,
    - **pos** : position in the chromosome for the event,
    - **str** : direction of the reads for the event[either 0 or 1],
    - **referenceDataframe** : a pandas data-frame that will store reference information

        Example `AnnotateEachBreakpoint(chr1,pos1,str1,refDF)`

1.1.4 `PredictFunction` module

- This module will predict the function of each annotated breakpoint
- **It takes two pandas series which has following information:**
    - **gene** : Gene for the event,
    - **transcript** : Transcript used for the event,
    - **site** : Explanation for site where the event occurs,
    - **zone** : Where does the event occur [1=exon, 2=intron, 3=3'-UTR, 4=5'-UTR, 5=promoter],
    - **strand** : Direction of the transcript,
    - **str** : Direction of the read,
    - **intronnum** : Which intron the event occurs if the event is in intron,
    - **intronframe** : What is the frame of the intron where the event is occuring.

        Example `ann1S = pandas.Series([gene1,transcript1,site1,zone1,strand1,str1,intron` `'transcript1', 'site1', 'zone1', 'txstrand1', 'readstrand1',` `'intronnum1','intronframe1'])`

```
ann2S = pandas.Series([gene2,transcript2,site2,zone2,strand2,str2,intronnum2,
'transcript2', 'site2', 'zone2', 'txstrand2', 'readstrand2',
'intronnum2','intronframe2'])
```

So **ann1S** & **ann2S** are series that will go to PredictFuntionForSV()

```
PredictFunctionForSV(ann1S,ann2S)
```

1.1.5 `helper` module

- **This module has multiple submodules**

  1. **ReadFile()**

     – This will read a tab-delimited file into a panadas dataframe

  2. **ExtendPromoterRegion()**

     – This will extend the promoter region to a given length

  3. **bp2str()**

     – This will convert base pair information to string information

1.1.6 `iAnnotateSV` module

- This module is the driver module, it takes user information and runs all other module to produce proper structural variant annotation

usage: iAnnotateSV.py [options]

Annotate SV based on a specific human reference

optional arguments:

> **-h, --help**   show this help message and exit
>
> **-v, --verbose**   make lots of noise [default]
>
> **-r hg19, --refFileVersion hg19**   Which human reference file to be used, hg18,hg19 or hg38

-o outfile, –outputFile out.txt   Full path with for the output file

-i inputSVfile, –svFile svfile.txt   Location of the structural variants file to annotate

-d distance, –distance 3000   Distance used to extend the promoter region

> **-a, --autoSelect**   Auto Select which transcript to be used[default]

-c canonicalExonsFile, –canonicalTranscripts canonicalExons.txt   Location of canonical transcript list for each gene. Use only if you want the output for specific transcripts for each gene.

CITATION

We are in the process of publishing a manuscript describing iAnnotateSV as part of the Structural Variant Detection framework. If you use this software in a publication, for now, please cite our website iAnnotateSV.

THREE

INDICES AND TABLES

- *genindex*
- *modindex*
- *search*