The Journal of Open Source Software

Finding and removing introns from RNA-Seq de novo assemblies with IntronSeeker

Sarah Maman [1¶], **Philippe Bardou** [1], **Emilien Lasguignes**[2], **Faustine Oudin**[2], **Floréal Cabanettes**[2], **and Christophe Klopp** [2]

1 Sigenae, GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326, Castanet Tolosan, France. **2** Université Fédérale de Toulouse, INRAE, BioinfOmics, GenoToul Bioinformatics facility, Sigenae, 31326, Castanet-Tolosan, France ¶ Corresponding author

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Kelly Rowland

Reviewers:

- @CFGrote
- @tkchafin

Submitted: 21 November 2023
Published: unpublished

License
Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License (CC BY 4.0).

Summary

intronSeeker identifies potentially retained introns in *de novo* RNA-seq assembly in order to quantify and remove them. To use it you have to provide a set of contigs resulting of a *de novo* transcriptome assembly and a set of RNA-Seq reads. The reads will be aligned on the contigs, splices sites will be searched and tested to check if they correspond to valid intron retention events. It includes two types of RNA-seq data simulation strategies to validate the detection process and measure the false positive detection rate. intronSeeker provides a list of potential intron candidates, it filters them and outputs a clean reference without introns in Fasta format.

Introduction

Short read RNA sequencing (RNA-Seq) is now routinely used to study gene expression. When a reference genome is available, RNA-Seq reads can be splice-aligned to the assembly and gene abundances can be measured by counting alignments found in each gene location. When no reference genome assembly is available, reads are usually assemble to build a reference transcriptome contig set. In this de novo approach reads are then aligned to the contigs without using a splice-aware aligner because they originate from mature transcripts.

In order not to sequence very abundant ribosomal RNAs, commonly used protocols include oligo(dT) transcript enrichment. Transcript poly-adhenylation and splicing take place in the nucleus before transfer in the cytoplasm. PolyA tail selection retrieves mainly mature spliced transcripts. Introns spanning reads are still found in most RNA-Seq sets as shown by (Ameur et al., 2011).

Intron retention is also a known biological gene regulation or alternative splicing mechanism. In plants it has been shown to increase the number of transcript splicing forms (Jia et al., 2020). (Braunschweig et al., 2014) have shown that, even in mammals, their is widespread intron retention linked with gene expression regulation.

Reads located on intron/exon boundaries harbor specific splice motifs. These splice motifs are di-nucleotides located at both intron ends. They can be canonical (GT/AG) or not. Canonical motifs are found in most site of most species.

Intron retention can be seen as an artefact or a biologically relevant mechanism and in both cases it is interesting to monitor. This is easy with a reference genome assembly because one can quantify reads aligned in introns. This is more complicated in the de novo approach in which the assembler can produce contigs with and without intron for the same transcript.

DRAFT

Contigs with introns are more difficult to annotated because introns are splitting coding sequences (CDS) and possibly generating several shorter protein alignments rather than a unique long match.

Canonical splice motifs presence, number of reads splicing at the same location, the minimum number of splice events and overlaps can be used to reduce faulty detection.

Hereafter we present IntronSeeker a software package enabling to search and remove introns from de novo assembled RNA-Seq contigs.

Statement of need

RNA-Seq de novo assemblies are widely used to study transcription in species lacking a reference genome. The classical extraction protocol collects RNA fragments using their PolyA tails and there-fore should only gather mature RNAs. Unfortunately part of the extracted RNA molecules still com-prise retained introns which can therefore be present in some assembled contigs. These introns generate transcript frameshifts and thus render annotation more difficult. IntronSeeker searches introns by splice-realigning reads on contigs. Introns candidates usually show canonical intron/exon boundaries supported by several sequences. IntronSeeker found between 0.02 and 11.96% of putative introns in twenty publicly available RNA-Seq de novo assembled samples. IntronSeeker produces an intron cleaned contig fasta file as well as a cleaning report. IntronSeeeker can be downloaded fromhttps://github.com/genotoul-bioinfo/intronseeker.

Searching retained introns in public datasets

The twenty public TSA contigs sets processed by IntronSeeker are classified in three Fungi, six Plantae, nine Animalia kingdoms and two Eukaryote superkingdoms. The number of contigs ranged from thirty thousand to three hundred thousand. The number of reads ranged from six millions to three hundred thirty millions. The proportion of retained intron candidates ranged from 0.02 to 11.96%. The figures are presented in Table 1.

Table 1: Rates of public contigs with introns, data from NCBI Transcript Shotgun Archive (TSA) and Short Read Archive (SRA).

Species	(super)king-dom	TSA id	Nb contigs	SRA and link to HTML intronSeeker report	Nb reads.	% cwi
Salvia m.	Plantae	GJJN01	69 705	SRR15718805	23 086 599	**11.96%**
Platichthys s.	Animalia	GAPK01	30 630	SRR1023744	516 791 904	**10,71%**
Rigido-porus m.	Fungi	GDMN01	34 441	SRR2187438	75 600 628	**5,97%**
Isatis t.	Plantae	GARR01	33 238	SRR1051997	113 134 348	**5,41%**
Go-niomonas a.	Eukaryota	GGUN01	48 844	SRR7601946	82 416 944	**5,27%**
Vriesea c.	Plantae	GHCB01	41 228	SRR500874	85 726 288	**3,94%**
Graminella n.	Animalia	GAQX01	37 537	SRR857257	43 693 708	**2,88%**

DRAFT

Species	(super)kingdom	TSA id	Nb contigs	SRA and link to HTML intronSeeker report	Nb reads.	% cwi
Carassius g.	Animalia	GJKR01	109 966	SRR12596368	21 661 960	**2.72%**
Rhizopus a.	Fungi	GDUK01	30 601	SRR2104505	64 801 576	**2,51%**
Diplonema p.	Eukaryota	GJNJ01	114 037	SRR14933372	63 775 926	**2.35%**
Azolla f.	Plantae	GBTV01	36 091	SRR1618559	122 059 452	**1,97%**
Tripidium r.	Plantae	GJDA01	106 494	SRR14143372	15 357 748	**1.50%**
Cimex l.	Animalia	GBYH01	39 124	SRR1660436	329 875 624	**1,24%**
Trichoplax sp. H2	Animalia	GFSF01	43 376	SRR5819939	128 665 904	**1,24%**
Phytophthora c.	Plantae	GBGX01	21 662	SRR1206033	14 346 946	**1.15%**
Piromyces sp.	Fungi	GGXH01	124 096	SRR7819335	15 615 535	**0.50%**
Sander l.	Animalia	GJIW01	56 196	SRR15372351	48 694 199	**0.36%**
Rhodnius n.	Animalia	GJJI01	67 217	SRR15602387	6 775 534	**0.11%**
Choromytilus c.	Animalia	GJJD01	106 298	SRR15058678	10 490 833	**0.10%**
Bombus t.	Animalia	GHFS01	48 241	SRR6148374	15 547 444	**0.02%**

NCBI taxonomy browser, superkingdom used when kingdom not provided.

intronSeeker files are available in "master" branch : intronSeeker / data / REAL_DATA

Fraction of contigs with introns (cwi) = Nb of contigs PASS / Nb total of contigs

Implementation

General overview

IntronSeeker is a python script which includes three steps enabling to align read on contigs, find and remove retained introns and produce an html report. Figure 1. presents theses steps with input and output file formats.

IntronSeeker is open source (GNU General Public License) and can be download from http://github.com/

DRAFT

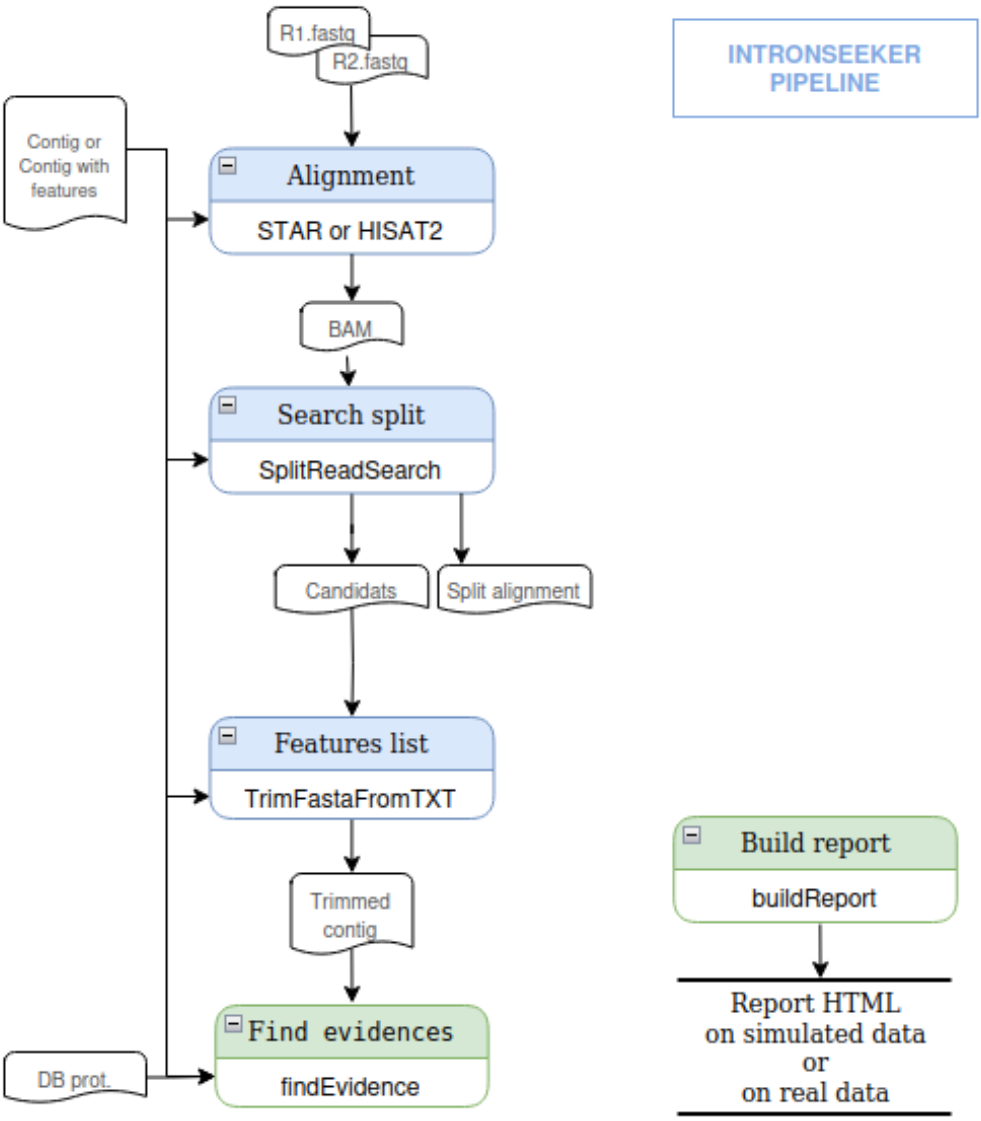

Figure 1: IntronSeeker steps diagram

Conda based installation procedure

To ease installation, intronSeeker includes an installation script (setup.sh) which run the installation of all the dependencies (1) but one (grinder) using conda and then installs grinder in the conda environment. Conda has to be installed beforehand. Installation can be checked using intronSeeker checkInstall program which will verify the presence and version of all dependencies. To run intronSeeker one has first to load the ISeeker_environment using conda 'source activate ISeeker_environment' command. An example data set named reduced_real_dataset is also provided in the data directory. It includes the result files which will enable manual comparison of the reference and produced intron files after complying the test. The installation procedure and the test command line can be found on the main page of the intronSeeker GITHUB WEB-page.

(1) IntronSeeker dependencies include seven software packages : grinder (Angly et al., 2012),

gffread (Pertea, 2019), hisat2 (Kim et al., 2015), STAR (Dobin et al., 2013), samtools (Li et al., 2009), TransDecoder (Haas et al., 2013), diamond (Buchfink et al., 2015).

Versions required before installation: - Python version 3.6 or above. - Miniconda 23.3.3 or above.

```
#Clone intronSeeker code from Git:
$ git clone https://forgemia.inra.fr/emilien.lasguignes/intronSeeker.git

#Load our miniconda environment and use libmamba
$ conda activate
$ conda update -n base conda
$ conda install -n base --override-channels -c conda-forge mamba 'python_abi=*=*cp*'

# Set up intronSeeker
$ cd intronSeeker/
$ CONDA_SOLVER="libmamba" /bin/bash ./setup.sh

#Activate ISeeker_environment and check installation
$ conda activate ISeeker_environment
$ intronSeeker checkInstall

# Command line help
$ intronSeeker -h
```

Setting default detection parameters

Detected introns are filtered according to thresholds (number of reads overlapping an intron and maximum length), canonical junction (GT_AG or CT_AC), and complexity (too long or overlapping introns).

Different GBS simulations where performed to find the best possible thresholds for minimum number of splice events, maximum intron spanning size, presence of canonical splice sites and minimum overlap size to call a splice event an intron retention candidate.

DRAFT

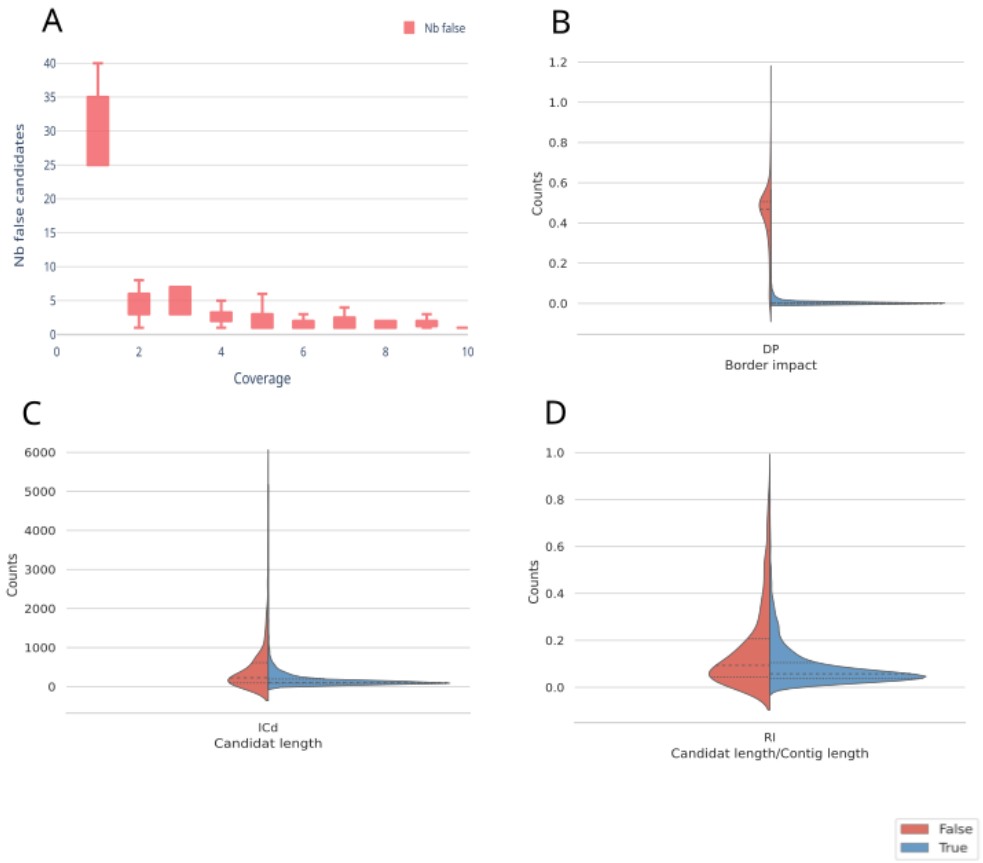

Figure 2: GBS parameters impacts

IntronSeeker parameters impacts. Graphics have been built on 10 samples of Arabidopsis thaliana data with powerlow abundance (1.2). Details are available in data/ directory.

A Impact of coverage on detection. Increased coverage means that false candidates are quickly lost, with very limited impact on the number of true candidates (tested with Arabidopsis thaliana data and powerlow abundance).

B Filtering candidates on the border enhances detection. DPratio is calculated as follows :

$$\frac{DP_{in}}{(DP_{before} + DP_{after})}$$

With DPbefore corresponding to the mean DP for 10bp before the candidate, DPin to the mean DP of candidate and DPafter to the mean DP for 10bp after the candidate.

C Filter candidates on candidat length.

D Filter candidates on retained intron ratio (candidat length / contig length).

Acknowledgement of financial support

The non permanent positions of Floréal Cabanette and Emilien Lasguignes were financed by projet France Génomique n° 31000523 and projet France Génomique n° 15000079.

References

Ameur, A., Zaghlool, A., Halvardson, J., Wetterbom, A., Gyllensten, U., Cavelier, L., & Feuk, L. (2011). Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. *Nature Structural Molecular Biology*, *18*(12), 1435–1440. https://doi.org/10.1038/nsmb.2143

Angly, F. E., Willner, D., Rohwer, F., Hugenholtz, P., & Tyson, G. W. (2012). Grinder: A versatile amplicon and shotgun sequence simulator. *Nucleic Acids Research*, *40*(12), e94–e94. https://doi.org/10.1093/nar/gks251

Braunschweig, U., Barbosa-Morais, N. L., Pan, Q., Nachman, E. N., Alipanahi, B., Gonatopoulos-Pournatzis, T., Frey, B., Irimia, M., & Blencowe, B. J. (2014). Widespread intron retention in mammals functionally tunes transcriptomes. *Genome Research*, *24*(11), 1774–1786. https://doi.org/10.1101/gr.177790.114

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, *12*(1), 59. https://doi.org/10.1038/nmeth.3176

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Lieber, M., & others. (2013). De novo transcript sequence reconstruction from RNA-seq using the trinity platform for reference generation and analysis. *Nature Protocols*, *8*(8), 1494.

Jia, J., Long, Y., Zhang, H., Li, Z., Liu, Z., Zhao, Y., Lu, D., Jin, X., Deng, X., Xia, R., Cao, X., & Zhai, J. (2020). Post-transcriptional splicing of nascent RNA contributes to widespread intron retention in plants. *Nature Plants*, *6*(7), 780–788. https://doi.org/10.1038/s41477-020-0688-1

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: A fast spliced aligner with low memory requirements. *Nature Methods*, *12*(4), 357. https://doi.org/10.1038/nmeth.3317

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Pertea, G. (2019). *GFF/GTF parsing utility providing format conversions, region filtering, FASTA sequence extraction and more*. https://github.com/gpertea/gffread/.

DRAFT