Transcript comprehensive comparison

IsoComp team

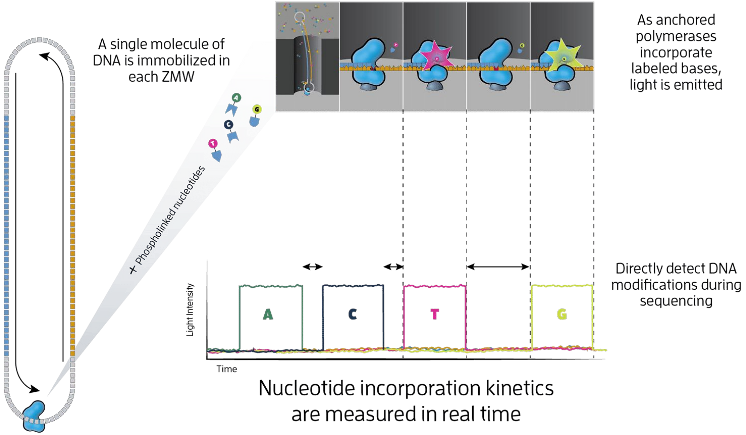
ISOCOMP

Introduction

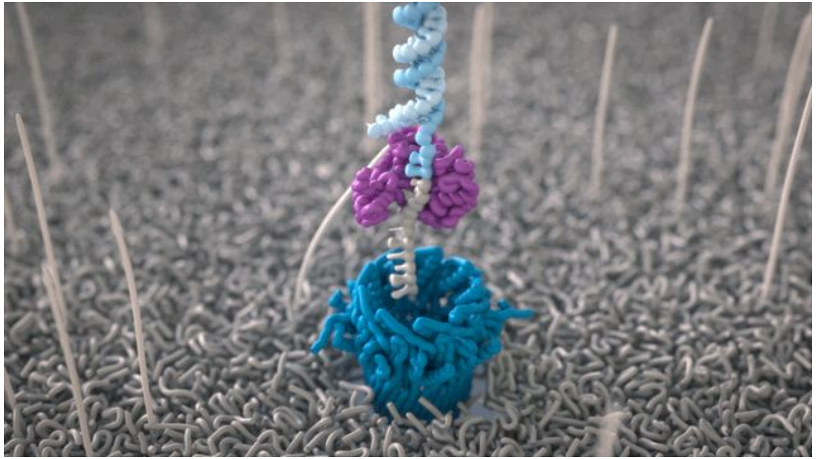
- 20,000 protein coding genes, they produce at least 100,000 splice isoforms
    - Alternative splicing is crucial regulator of gene expression and a key contributor to both normal developmental processes and disease states
- Long-read able bypass the transcript reconstruction challenges of short reads
- Integration of transcript similarity is as important as well as their genome coordinates

Why long reads?

PacBio


A single molecule of DNA is immobilized in each ZMW
+ Phospholinked nucleotides
As anchored polymerases incorporate labeled bases, light is emitted
Directly detect DNA modifications during sequencing
Light Intensity
Time
A
C
T
G
Nucleotide incorporation kinetics are measured in real time

ONT



-Sequence DNA in real time

-~ 20kb

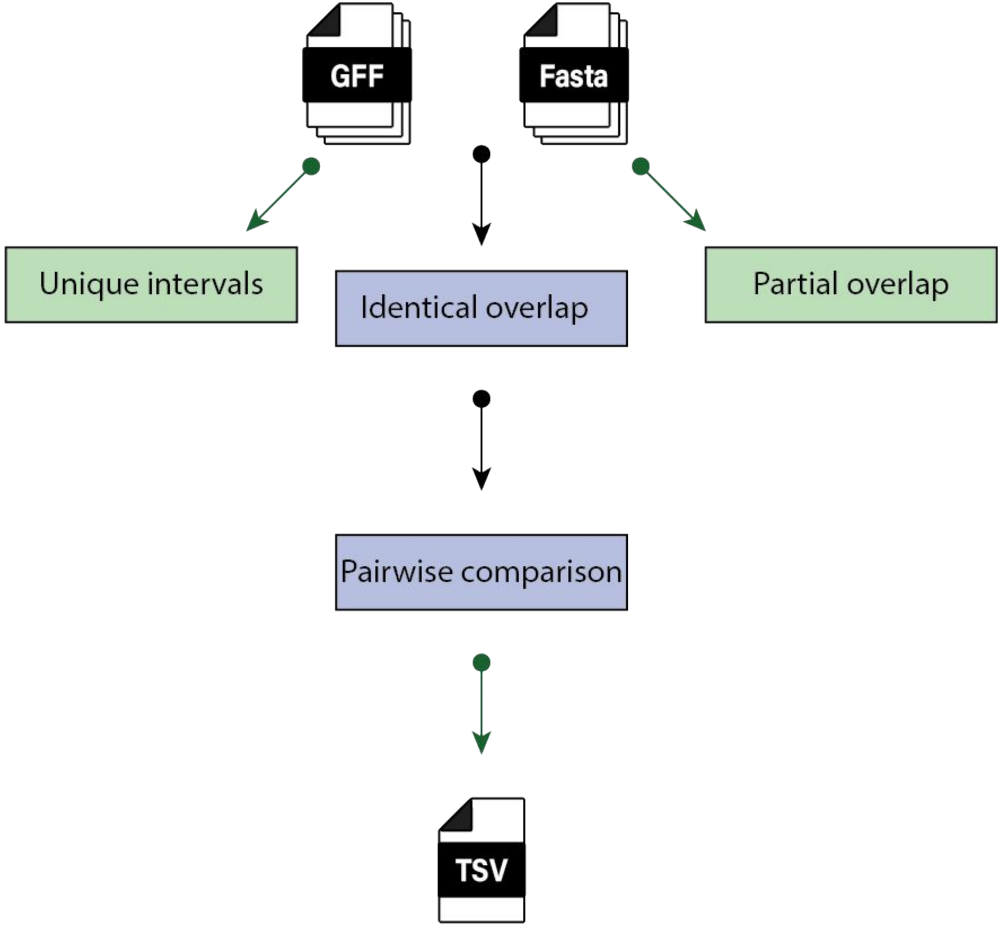
-Low error rate (<1% PacBio HiFi and <4% ONT)

Gene Annotation Formats

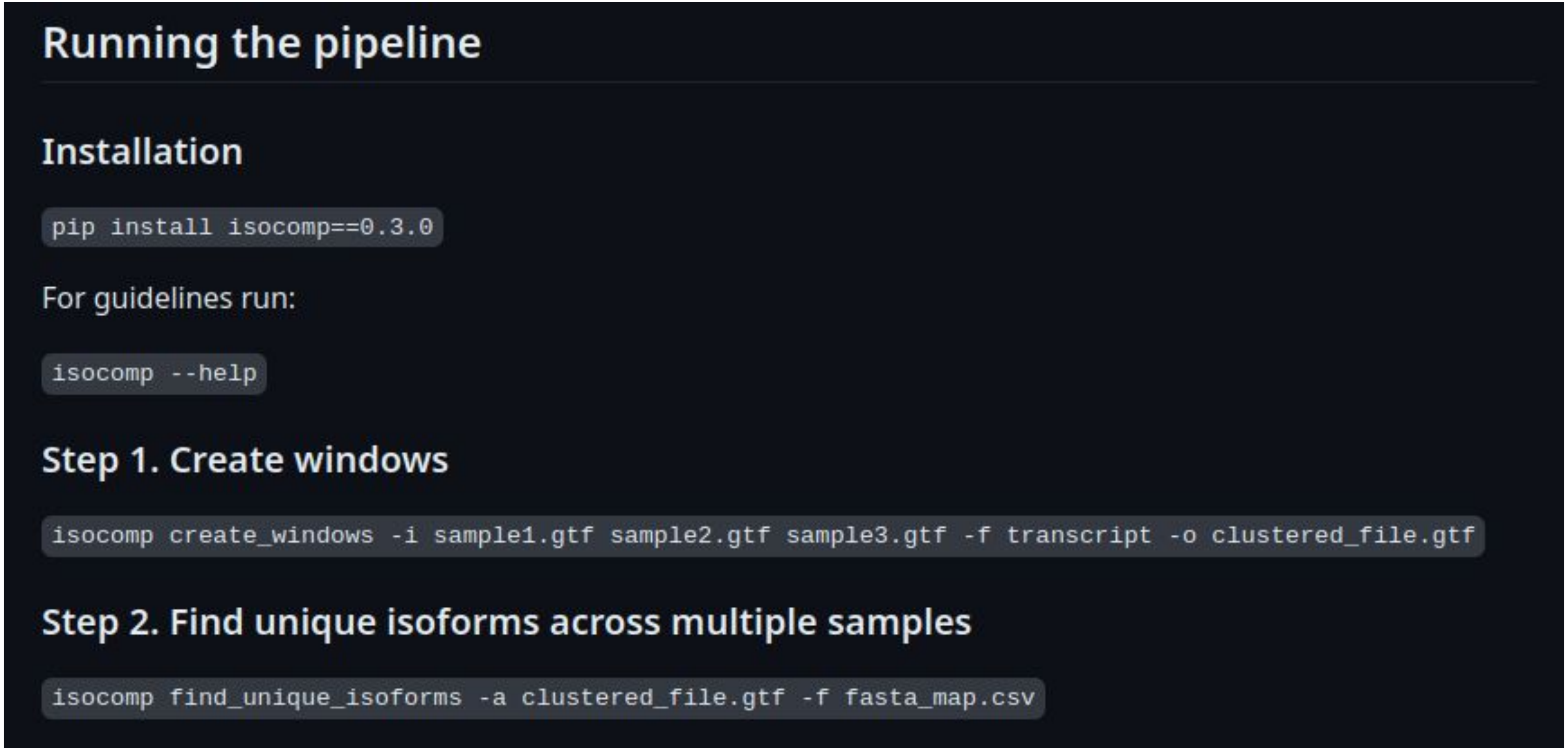
- GTF (Gene Transfer Format) and GFF (General Feature Format)
- GffRead, GffCompare, AGAT, bedtools, and parse eval are sources for manipulate GTF/GFF files
- Comparison is based on **transcript coordinates** only
- Transcripts can **exhibit variability** due to genetic variants or RNA editing processes
- Thus, there is a demand for a comprehensive **Transcript Comparison**

Our approach

- N number of samples
- Agnostic input (ONT or PacBio)
- Coordinate ***and*** Sequence level comparison
- We utilized samples HG002 (3 replicates), HG004, and HG005



Implementation



Running the pipeline

Installation

```
pip install isocomp==0.3.0
```

For guidelines run:

```
isocomp --help
```

Step 1. Create windows

```
isocomp create_windows -i sample1.gtf sample2.gtf sample3.gtf -f transcript -o clustered_file.gtf
```

Step 2. Find unique isoforms across multiple samples

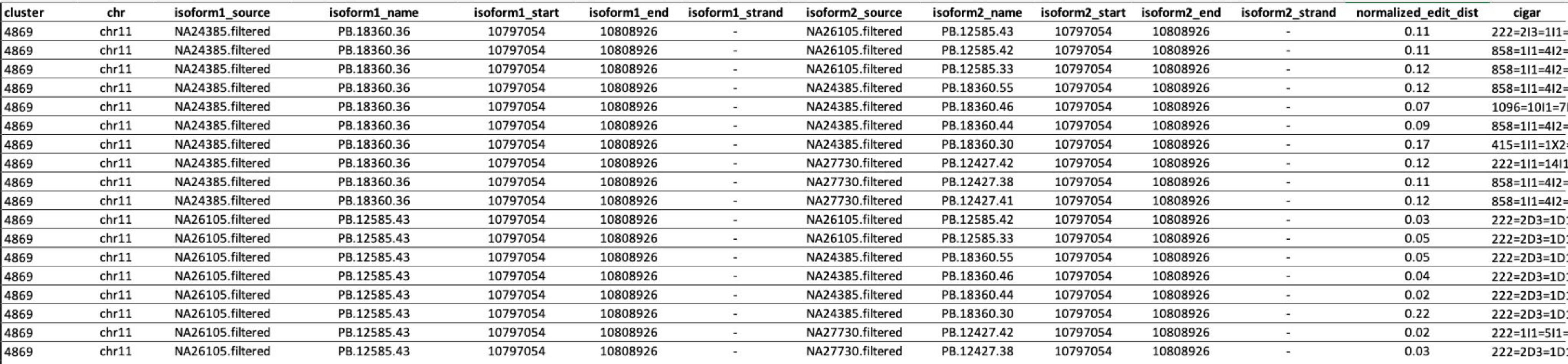
```
isocomp find_unique_isoforms -a clustered_file.gtf -f fasta_map.csv
```

Output (intermediate)

- Run time and CPUs
  - ~ 15 minute
  - 16 CPUs (DNAnexus)
  - ~ 8G RAM

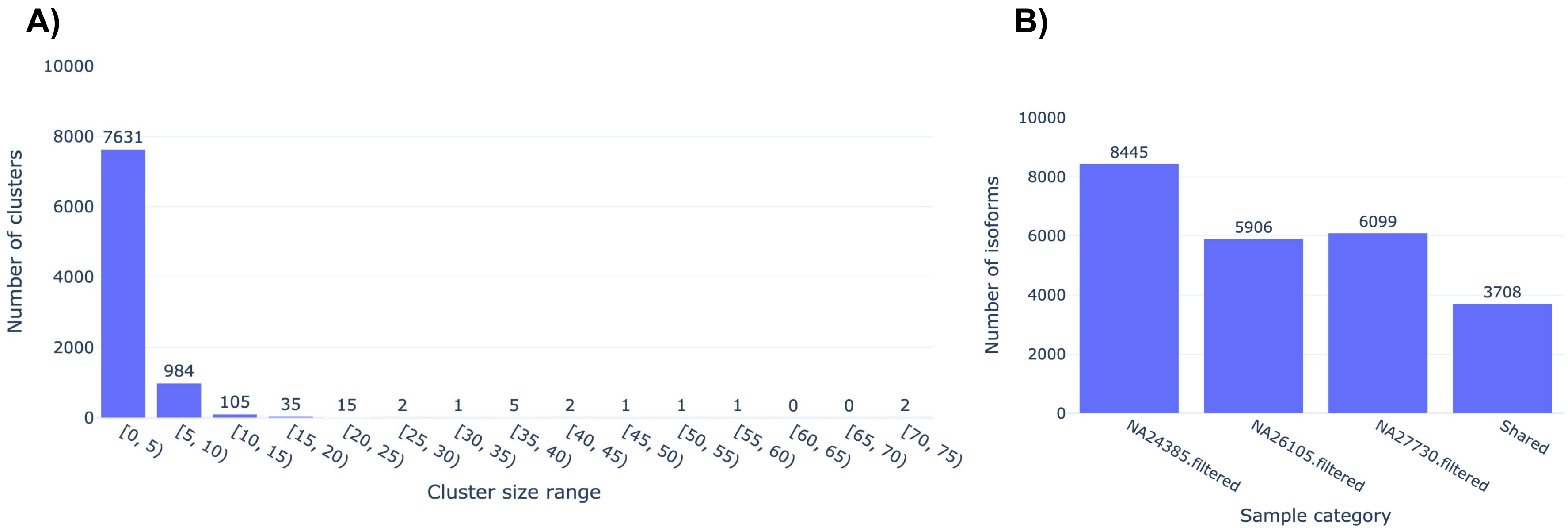
- Functionality
  - Find intervals
  - Compare intervals
  - Multithreading
  - Easy installation
  - Convenient TSV output (coming soon…)

Awesome job!!

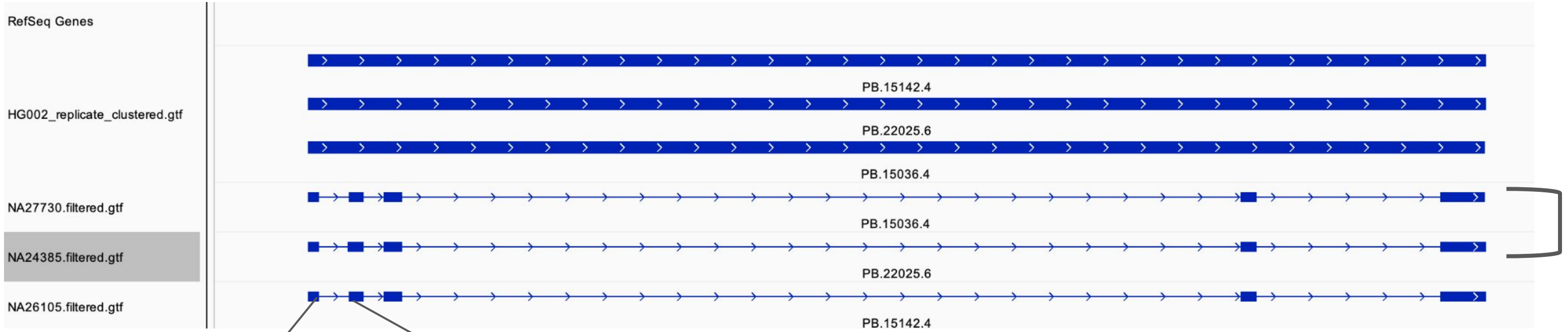


cluster	chr	isoform1_source	isoform1_name	isoform1_start	isoform1_end	isoform1_strand	isoform2_source	isoform2_name	isoform2_start	isoform2_end	isoform2_strand	normalized_edit_dist	cigar
4869	chr11	NA24385.filtered	PB.18360.36	10797054	10808926	-	NA26105.filtered	PB.12585.43	10797054	10808926	-	0.11	222=2I3=1I1=
4869	chr11	NA24385.filtered	PB.18360.36	10797054	10808926	-	NA26105.filtered	PB.12585.42	10797054	10808926	-	0.11	858=1I1=4I2=
4869	chr11	NA24385.filtered	PB.18360.36	10797054	10808926	-	NA26105.filtered	PB.12585.33	10797054	10808926	-	0.12	858=1I1=4I2=
4869	chr11	NA24385.filtered	PB.18360.36	10797054	10808926	-	NA24385.filtered	PB.18360.55	10797054	10808926	-	0.12	858=1I1=4I2=
4869	chr11	NA24385.filtered	PB.18360.36	10797054	10808926	-	NA24385.filtered	PB.18360.46	10797054	10808926	-	0.07	1096=10I1=7
4869	chr11	NA24385.filtered	PB.18360.36	10797054	10808926	-	NA24385.filtered	PB.18360.44	10797054	10808926	-	0.09	858=1I1=4I2=
4869	chr11	NA24385.filtered	PB.18360.36	10797054	10808926	-	NA24385.filtered	PB.18360.30	10797054	10808926	-	0.17	415=1I1=1X2=
4869	chr11	NA24385.filtered	PB.18360.36	10797054	10808926	-	NA27730.filtered	PB.12427.42	10797054	10808926	-	0.12	222=1I1=14I1
4869	chr11	NA24385.filtered	PB.18360.36	10797054	10808926	-	NA27730.filtered	PB.12427.38	10797054	10808926	-	0.11	858=1I1=4I2=
4869	chr11	NA24385.filtered	PB.18360.36	10797054	10808926	-	NA27730.filtered	PB.12427.41	10797054	10808926	-	0.12	858=1I1=4I2=
4869	chr11	NA26105.filtered	PB.12585.43	10797054	10808926	-	NA26105.filtered	PB.12585.42	10797054	10808926	-	0.03	222=2D3=1D
4869	chr11	NA26105.filtered	PB.12585.43	10797054	10808926	-	NA26105.filtered	PB.12585.33	10797054	10808926	-	0.05	222=2D3=1D
4869	chr11	NA26105.filtered	PB.12585.43	10797054	10808926	-	NA24385.filtered	PB.18360.55	10797054	10808926	-	0.05	222=2D3=1D
4869	chr11	NA26105.filtered	PB.12585.43	10797054	10808926	-	NA24385.filtered	PB.18360.46	10797054	10808926	-	0.04	222=2D3=1D
4869	chr11	NA26105.filtered	PB.12585.43	10797054	10808926	-	NA24385.filtered	PB.18360.44	10797054	10808926	-	0.02	222=2D3=1D
4869	chr11	NA26105.filtered	PB.12585.43	10797054	10808926	-	NA24385.filtered	PB.18360.30	10797054	10808926	-	0.22	222=2D3=1D
4869	chr11	NA26105.filtered	PB.12585.43	10797054	10808926	-	NA27730.filtered	PB.12427.42	10797054	10808926	-	0.02	222=1I1=5I1=
4869	chr11	NA26105.filtered	PB.12585.43	10797054	10808926	-	NA27730.filtered	PB.12427.38	10797054	10808926	-	0.03	222=2D3=1D

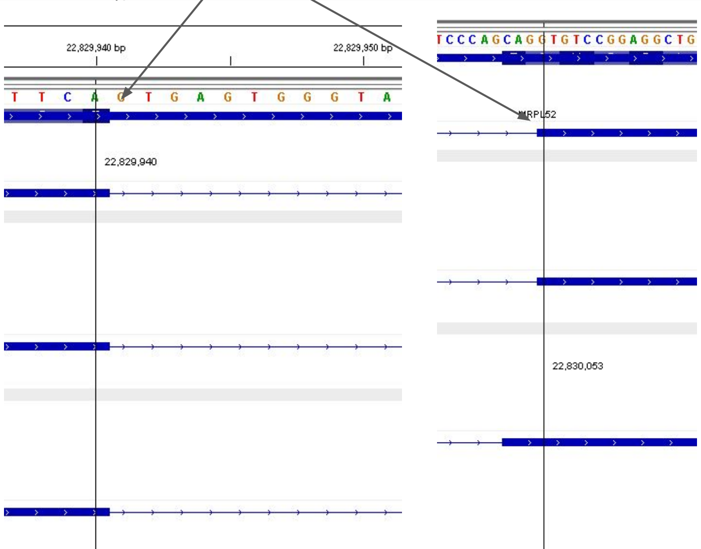
Results: Most Isoforms are unique by TSS and/or TTS

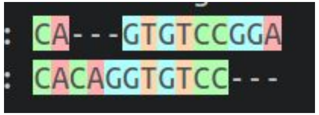


Results: Sequence comparison alone is difficult to classify



Same transcript



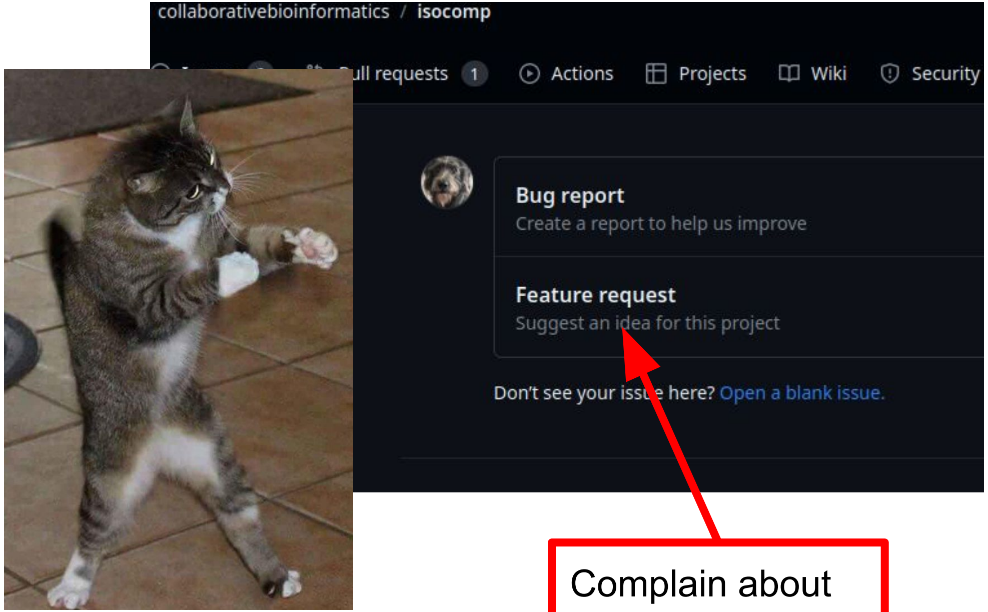


```
: CA---GTGTCCGGA
: CACAGGTGTCC---
```

This illustrates how difficult it is to detect and parse out from a cigar string an alternate splice site. This can *easily* be detected using coordinates. Sequence comparison should *only* be over the exactly overlapping sequences

Current State

- Proof of concept: There exist isoforms equivalent by coordinate which possess sequence variants



Bug report
Create a report to help us improve

Feature request
Suggest an idea for this project

Don't see your issue here? Open a blank issue.

Do you really want to change?

Complain about output format here

Future Directions

- Other tools possess more sophisticated methods of comparing intervals
    - Implement interval/comparison with interval tree
        - Using this, we can both classify isoform diversity with labels such as differential start/end site, exon usage, splice site variants etc.
        - And, for those transcripts which appear identical by coordinate, resolve any differences at the sequence level
- Output format – Current results table represents an intermediate step. The results still must be refined into a usable format

Acknowledge

Sej
Modha

Chase
Mateusiak

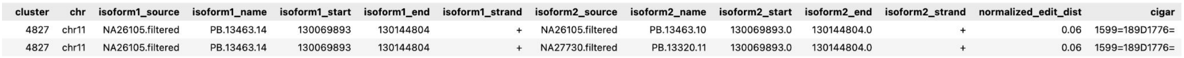
Trinh
Tat

Jędrzej
Kubica

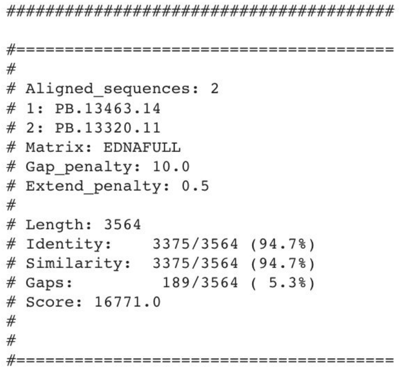
Medhat
Mahmoud

- Muhammad Sohail Raza
- Rupesh Kesharwani
- Sejal Modha
- Umran Yaman
- Yutong Qiu
- Abdullah Al Nahid
- Bida Gu
- Chia Sin Liew
- Evan Biederstedt
- Luis Fernández

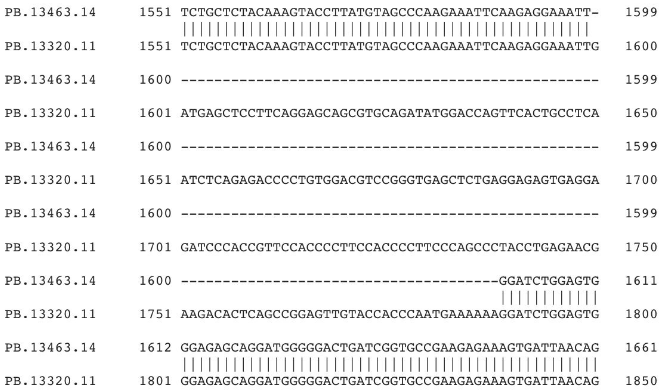
Example 2



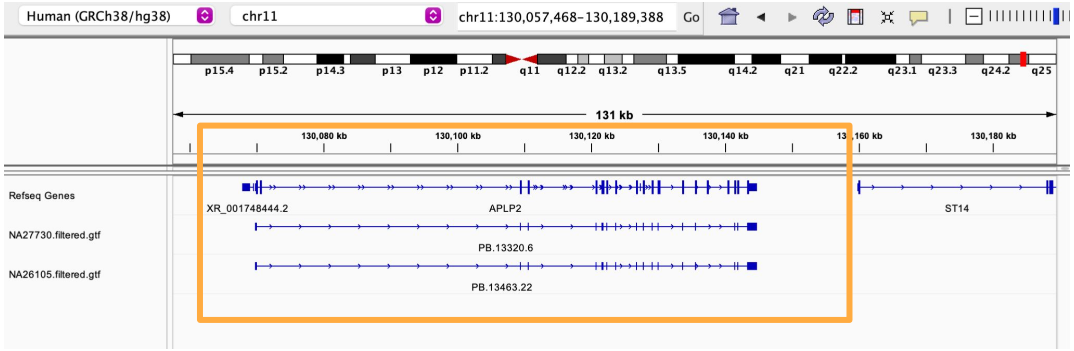
cluster	chr	isoform1_source	isoform1_name	isoform1_start	isoform1_end	isoform1_strand	isoform2_source	isoform2_name	isoform2_start	isoform2_end	isoform2_strand	normalized_edit_dist	cigar
4827	chr11	NA26105.filtered	PB.13463.14	130069893	130144804	+	NA26105.filtered	PB.13463.10	130069893.0	130144804.0	+	0.06	1599=189D1776=
4827	chr11	NA26105.filtered	PB.13463.14	130069893	130144804	+	NA27730.filtered	PB.13320.11	130069893.0	130144804.0	+	0.06	1599=189D1776=



```
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: PB.13463.14
# 2: PB.13320.11
# Matrix: EDNAFULL
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 3564
# Identity:    3375/3564 (94.7%)
# Similarity:  3375/3564 (94.7%)
# Gaps:         189/3564 ( 5.3%)
# Score: 16771.0
#
#
#=======================================
```



```
PB.13463.14       1551 TCTGCTCTACAAAGTACCTTATGTAGCCCAAGAAATTCAAGAGGAAATT-   1599
                       |||||||||||||||||||||||||||||||||||||||||||||||||
PB.13320.11       1551 TCTGCTCTACAAAGTACCTTATGTAGCCCAAGAAATTCAAGAGGAAATTG   1600

PB.13463.14       1600 --------------------------------------------------   1599

PB.13320.11       1601 ATGAGCTCCTTCAGGAGCAGCGTGCAGATATGGACCAGTTCACTGCCTCA   1650

PB.13463.14       1600 --------------------------------------------------   1599

PB.13320.11       1651 ATCTCAGAGACCCCTGTGGACGTCCGGGTGAGCTCTGAGGAGAGTGAGGA   1700

PB.13463.14       1600 --------------------------------------------------   1599

PB.13320.11       1701 GATCCCACCGTTCCACCCCTTCCACCCCTTCCCAGCCCTACCTGAGAACG   1750

PB.13463.14       1600 ---------------------------------------GGATCTGGAGTG   1611
                                                              ||||||||||||
PB.13320.11       1751 AAGACACTCAGCCGGAGTTGTACCACCCAATGAAAAAAGGATCTGGAGTG   1800

PB.13463.14       1612 GGAGAGCAGGATGGGGGACTGATCGGTGCCGAAGAGAAAGTGATTAACAG   1661
                       ||||||||||||||||||||||||||||||||||||||||||||||||||
PB.13320.11       1801 GGAGAGCAGGATGGGGGACTGATCGGTGCCGAAGAGAAAGTGATTAACAG   1850
```


Human (GRCh38/hg38)
chr11
chr11:130,057,468-130,189,388
Go
p15.4
p15.2
p14.3
p13
p12
p11.2
q11
q12.2
q13.2
q13.5
q14.2
q21
q22.2
q23.1
q23.3
q24.2
q25
131 kb
130,080 kb
130,100 kb
130,120 kb
130,140 kb
13 ,160 kb
130,180 kb
Refseq Genes
XR_001748444.2
APLP2
ST14
NA27730.filtered.gtf
PB.13320.6
NA26105.filtered.gtf
PB.13463.22