Tutorial

De Novo Transcript Discovery using Long and Short Reads

December 4, 2018

Sample to Insight

QIAGEN Aarhus · Silkeborgvej 2 · Prismet · 8000 Aarhus C · Denmark
Telephone: +45 70 22 32 44 · www.qiagenbioinformatics.com · ts-bioinformatics@qiagen.com

De Novo Transcript Discovery using Long and Short Reads

The Wild Strawberry *Fragaria vesca* genome (240 Mb) was re-annotated by Li et al., 2018. The new annotation makes use of Illumina and PacBio CCS reads. In this tutorial we will, in just a few steps, annotate a large fraction of the genome using their PacBio data and a subset of one of their 90 Illumina RNA-seq libraries.

The features demonstrated in this tutorial include:

- Using Large Gap Read Mapping with default parameters on the PacBio reads.
- Running a de novo Transcript Discovery with the PacBio Large Gap Read Mapping and no annotation tracks.
- Running Transcript Discovery again using the PacBio predictions as known gene and known transcript tracks, and the sampled Illumina read mapping as input.
- Comparing our results with those of Li et al., 2018 in a Track List.

Prerequisites For this tutorial, you must be working with CLC Genomics Workbench 12.0 or higher, and have installed the Transcript Discovery plugin. How to install plugins is described here: `http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Install.html`.

Download and import data First, download and save the archive set from our website: `http://resources.qiagenbioinformatics.com/testdata/transcript_discovery_tutorial.zip`

Start the workbench, and create a new folder named, for example, "Transcript discovery tutorial", where we will keep the data and the analysis results.

To import the archive, use the Standard Import tool:

File | **Import** () | **Standard Import** ()

Select the zip file, keep the default option **Automatic import** checked, and press **Next** to choose a folder where the result will be saved.

This will produce a number of new objects in your navigation area, as shown in figure 1.

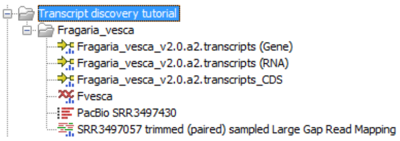

Transcript discovery tutorial
Fragaria_vesca
Fragaria_vesca_v2.0.a2.transcripts (Gene)
Fragaria_vesca_v2.0.a2.transcripts (RNA)
Fragaria_vesca_v2.0.a2.transcripts_CDS
Fvesca
PacBio SRR3497430
SRR3497057 trimmed (paired) sampled Large Gap Read Mapping

Figure 1: *New objects in the Navigation Area after import.*

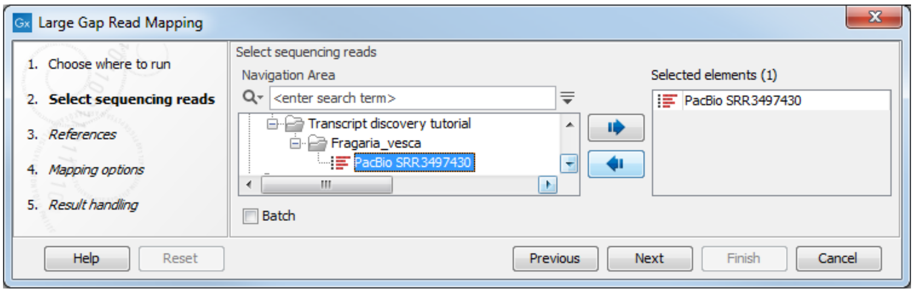
You can now see:

- PacBio reads
- The reference genome `Fvesca`
- Annotations (CDS, Gene and RNA) produced from Li et al., 2018
- A Large Gap Read Mapping performed on a subset of Illumina reads: the Illumina data was down-sampled to 10% of one of the 90 libraries present in the original paper.

Large Gap Read Mapping of PacBio reads

While the tutorial data set already includes a Large Gap Read Mapping for the Illumina reads, we will see here how to perform the one needed for the PacBio reads.

1. Start the **Large Gap Read Mapping** tool, and select the PacBio reads as shown in figure 2.



Large Gap Read Mapping

1. Choose where to run
2. **Select sequencing reads**
3. *References*
4. *Mapping options*
5. *Result handling*

Select sequencing reads

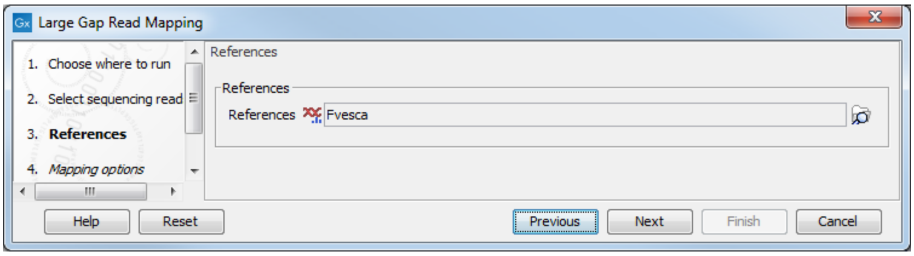
Navigation Area: <enter search term>
- Transcript discovery tutorial
  - Fragaria_vesca
    - PacBio SRR3497430

Selected elements (1): PacBio SRR3497430

☐ Batch

Figure 2: *Select the PacBio reads.*

2. In the next dialog, specify the reference sequence for Fragaria vesca (figure 3).



Large Gap Read Mapping

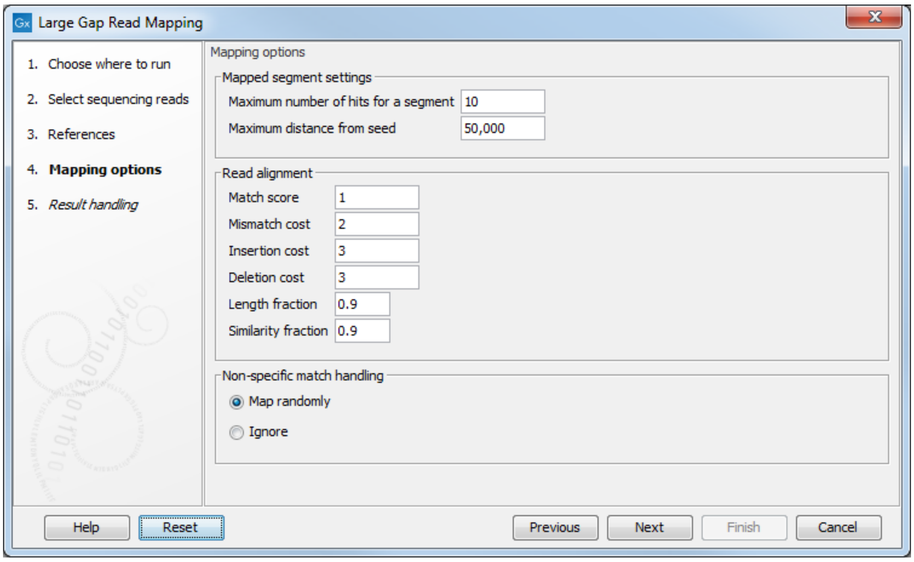
1. Choose where to run
2. Select sequencing read
3. **References**
4. *Mapping options*

References

References: Fvesca

Figure 3: *Specify the Fragaria vesca genome sequence.*

3. Leave the Large Gap Read Mapping options as they are set by default (figure 4).



Large Gap Read Mapping

1. Choose where to run
2. Select sequencing reads
3. References
4. **Mapping options**
5. *Result handling*

Mapping options

Mapped segment settings

Maximum number of hits for a segment: 10
Maximum distance from seed: 50,000

Read alignment

Match score: 1
Mismatch cost: 2
Insertion cost: 3
Deletion cost: 3
Length fraction: 0.9
Similarity fraction: 0.9

Non-specific match handling

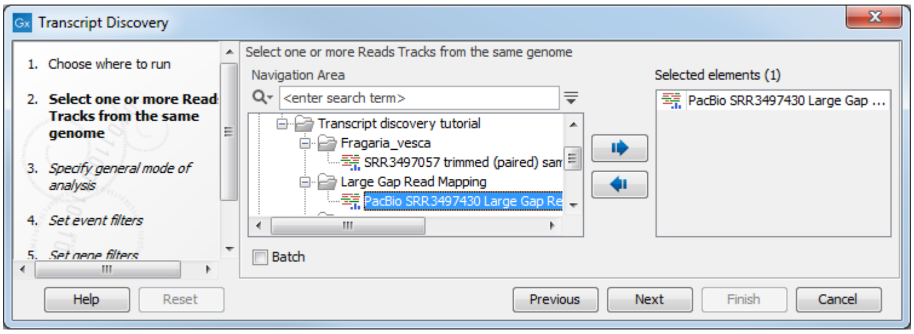
● Map randomly
○ Ignore

Figure 4: *Large Gap Read Mapping options.*

4. Finally, choose to **Save** the read mapping in the folder you created for this tutorial.

First run of Transcript Discovery

1. Start the **Transcript Discovery** tool, and select the PacBio Large Gap Read Mapping as shown in figure 5.



Transcript Discovery

1. Choose where to run
2. **Select one or more Read Tracks from the same genome**
3. *Specify general mode of analysis*
4. *Set event filters*
5. *Set gene filters*

Select one or more Reads Tracks from the same genome

Navigation Area

<enter search term>

- Transcript discovery tutorial
  - Fragaria_vesca
    - SRR3497057 trimmed (paired) sam
  - Large Gap Read Mapping
    - PacBio SRR3497430 Large Gap Re

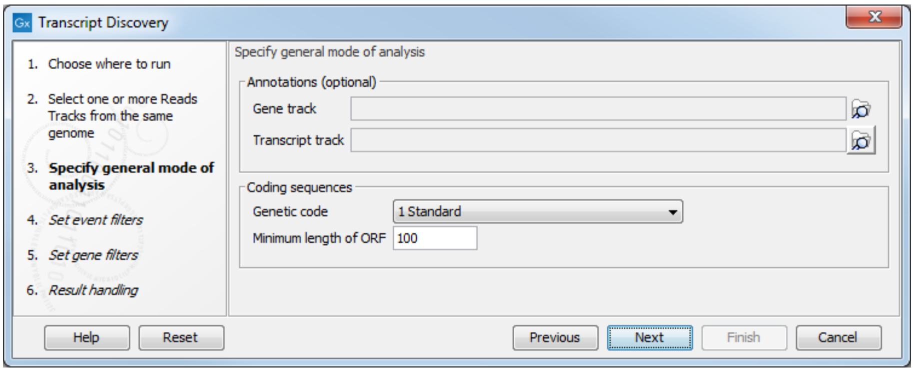
Selected elements (1)

PacBio SRR3497430 Large Gap ...

□ Batch

Figure 5: *Select the PacBio Large Gap Read Mapping.*

2. As you can see in figure 6, we do not specify any annotations (we are performing de novo discovery) and leave the parameters as they are set by default.



Transcript Discovery

1. Choose where to run
2. Select one or more Reads Tracks from the same genome
3. **Specify general mode of analysis**
4. *Set event filters*
5. *Set gene filters*
6. *Result handling*

Specify general mode of analysis

Annotations (optional)

Gene track

Transcript track

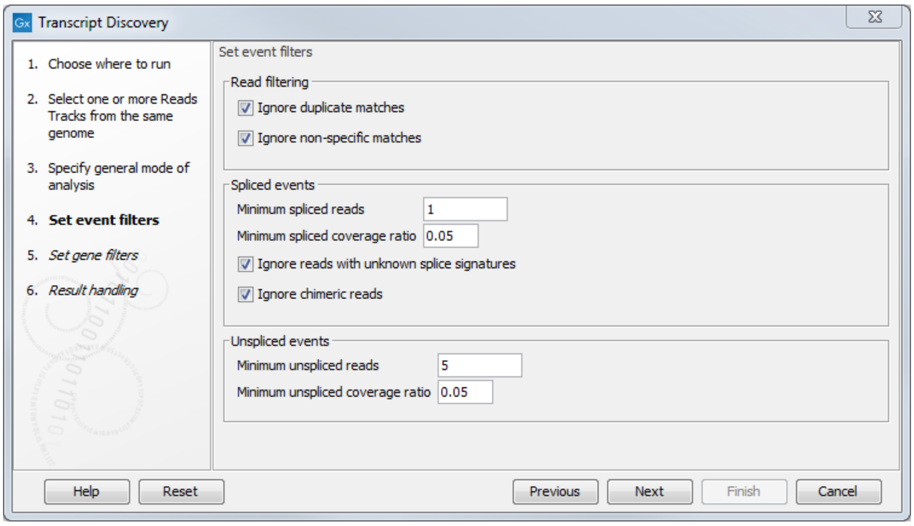
Coding sequences

Genetic code: 1 Standard

Minimum length of ORF: 100

Figure 6: *Select the Fragaria vesca Gene and RNA annotation tracks.*

3. In the event filters (figure 7), leave the parameters as they are set by default.



Transcript Discovery

1. Choose where to run
2. Select one or more Reads Tracks from the same genome
3. Specify general mode of analysis
4. **Set event filters**
5. *Set gene filters*
6. *Result handling*

Set event filters

Read filtering

☑ Ignore duplicate matches

☑ Ignore non-specific matches

Spliced events

Minimum spliced reads: 1

Minimum spliced coverage ratio: 0.05

☑ Ignore reads with unknown splice signatures

☑ Ignore chimeric reads

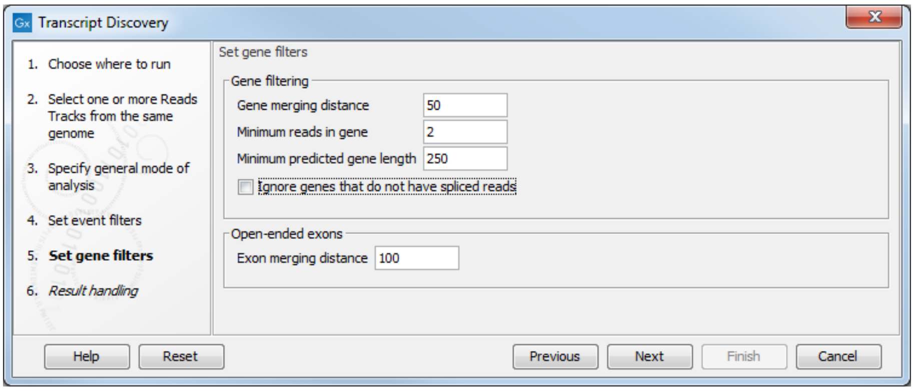
Unspliced events

Minimum unspliced reads: 5

Minimum unspliced coverage ratio: 0.05

Figure 7: *The event filters can be left as they are configured by default.*

4. In the gene filters (figure 8), set the *Minimum reads in gene* down from 10 to 2: the PacBio data is low coverage and supposedly full-length, so we are more likely to trust a gene with two independent pieces of evidence for a transcript than we would for shorter reads. Also uncheck *Ignore genes that do not have spliced reads*. It is checked by default to avoid getting a lot of noise in the output from regions where a handful of the many millions of reads might have mapped at random, but this is unlikely to occur with PacBio reads.



Transcript Discovery

1. Choose where to run
2. Select one or more Reads Tracks from the same genome
3. Specify general mode of analysis
4. Set event filters
5. **Set gene filters**
6. *Result handling*

Set gene filters

Gene filtering

Gene merging distance: 50
Minimum reads in gene: 2
Minimum predicted gene length: 250
□ Ignore genes that do not have spliced reads

Open-ended exons

Exon merging distance: 100

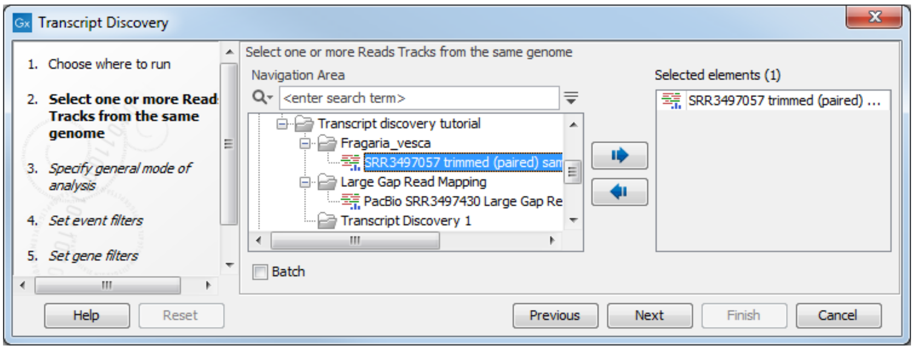
Figure 8: *Configuration of the gene filters when working with a long reads mapping.*

5. Finally, choose to **Save** the detected gene table and the report in a folder called Transcript Discovery 1.

Second run of Transcript Discovery

We will now run the Transcript Discovery tool a second time, using as input the Large Gap Read Mapping generated from a subset of the Illumina reads, and the annotations tracks generated in the previous section.

1. Start the **Transcript Discovery** tool, and select the Illumina Large Gap Read Mapping as shown in figure 9.



Transcript Discovery

1. Choose where to run
2. **Select one or more Read Tracks from the same genome**
3. *Specify general mode of analysis*
4. *Set event filters*
5. *Set gene filters*

Select one or more Reads Tracks from the same genome

Navigation Area

<enter search term>

- Transcript discovery tutorial
  - Fragaria_vesca
    - SRR3497057 trimmed (paired) sam
  - Large Gap Read Mapping
    - PacBio SRR3497430 Large Gap Re
  - Transcript Discovery 1

Selected elements (1)

SRR3497057 trimmed (paired) ...

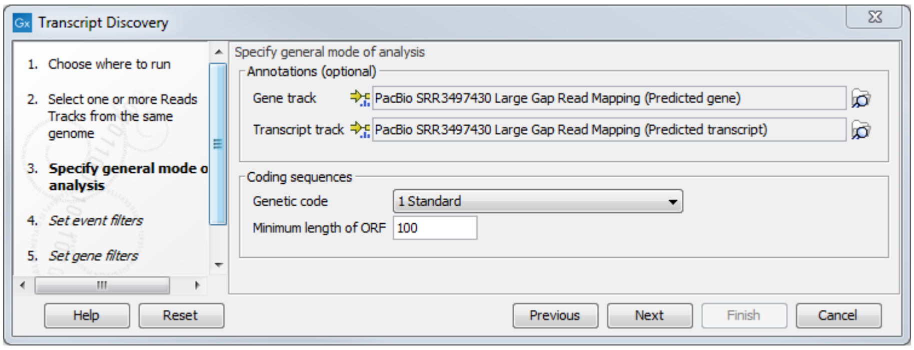
□ Batch

Figure 9: *Select the Illumina Large Gap Read Mapping.*

2. Specify the Gene and RNA annotation tracks generated during the first run of the tool (figure 10). Leave the parameters as they are set by default.

3. In the Event filters and Gene filters steps, click on **Reset** to configure all parameters as they are set by default (figure 11).

4. Finally, choose to **Save** the detected gene table and the report in a folder called Transcript Discovery 2.



1. Choose where to run
2. Select one or more Reads Tracks from the same genome
3. **Specify general mode of analysis**
4. *Set event filters*
5. *Set gene filters*

Specify general mode of analysis

Annotations (optional)

Gene track: PacBio SRR3497430 Large Gap Read Mapping (Predicted gene)

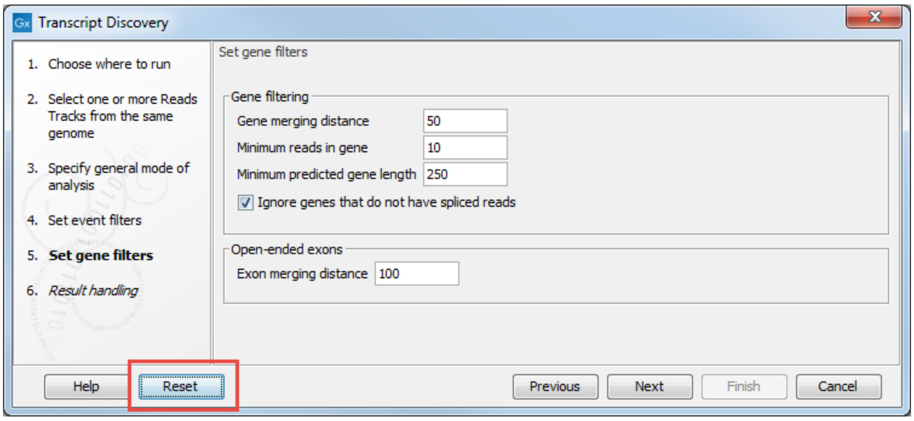
Transcript track: PacBio SRR3497430 Large Gap Read Mapping (Predicted transcript)

Coding sequences

Genetic code: 1 Standard

Minimum length of ORF: 100

Figure 10: *Select the Gene and RNA annotation tracks generated using the PacBio Large Gap Read Mapping.*



1. Choose where to run
2. Select one or more Reads Tracks from the same genome
3. Specify general mode of analysis
4. Set event filters
5. **Set gene filters**
6. *Result handling*

Set gene filters

Gene filtering

Gene merging distance: 50

Minimum reads in gene: 10

Minimum predicted gene length: 250

☑ Ignore genes that do not have spliced reads

Open-ended exons

Exon merging distance: 100

Figure 11: *Click on Reset to return all values to their default setting.*

Using a Track List to compare results

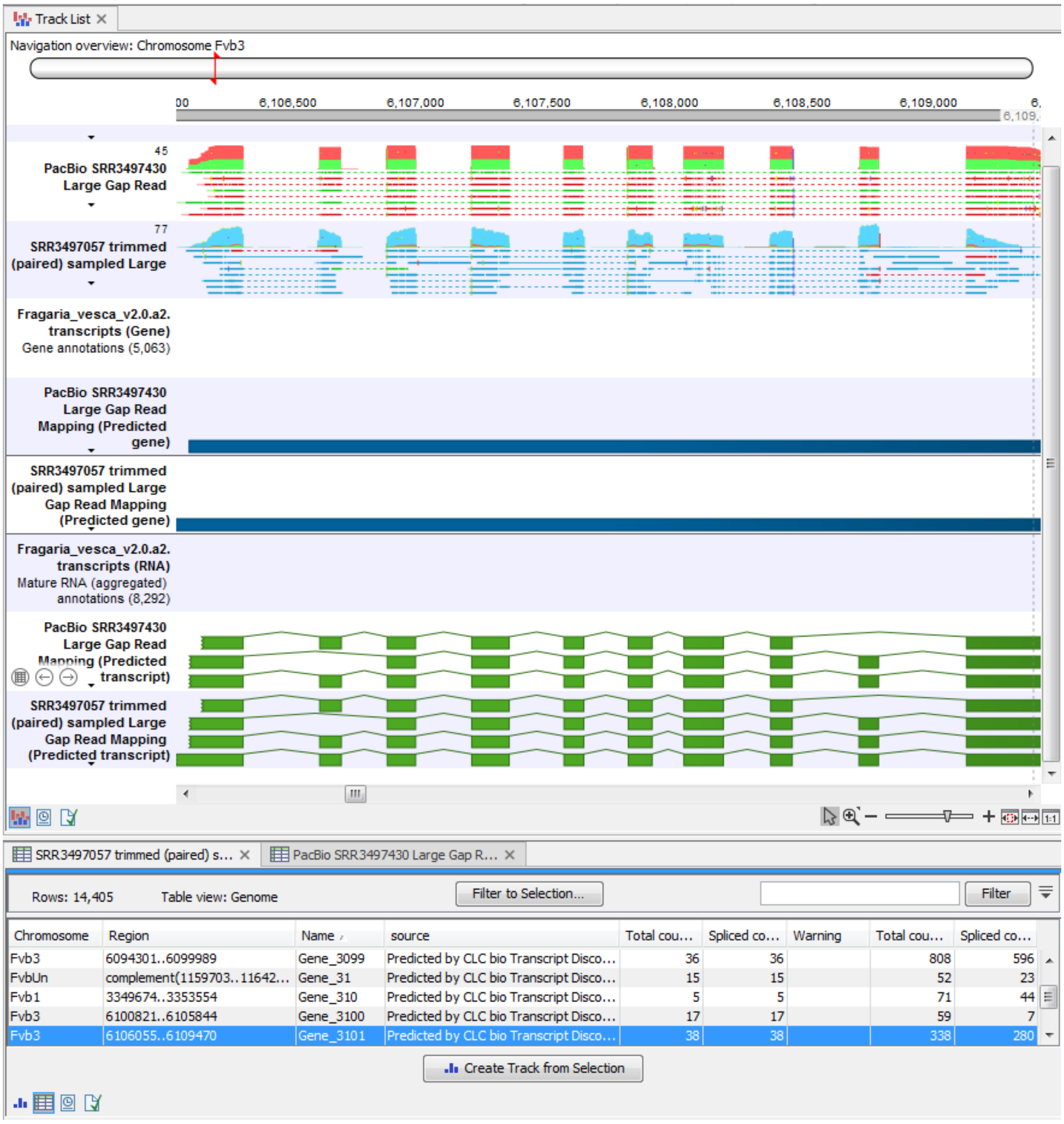
To have a better look at the results, we create a Track List including the following tracks:

- Fragaria vesca genome sequence
- Gene annotation tracks from the publication `v2.0.a2`, from the first run `PacaBio SRR3497430` and from the second run `SRR3497057`
- Transcript annotation tracks from the publication `v2.0.a2`, from the first run `PacBio SRR3497430` and from the second run `SRR3497057`
- Read Mappings `PacBio SRR3497430` and `SRR3497057`

Click on New | Track List and select the tracks listed above before clicking **Finish**. Once the Track List is open in the View Area, click on the little table icon under the `SRR3497057 Predicted Gene` track name to open these annotations as a table in split view.

In their paper, Li et al., 2018 annotated approximately 33,000 genes, which is many more than we annotated in this tutorial (14,000). However, this difference is expected, as we are only using 10% of a single Illumina library (as opposed to 90 libraries in the publication). The Illumina data still added 4,000 genes to the ones predicted by the PacBio data in the first run of Transcript Discovery.

Despite predicting fewer genes than in the initial paper, we were able to find new genes and transcripts that were not previously found. For example, figure 12 shows the annotation tracks at a location in Fvb3. We can see that our analysis has discovered a full-length CDS corresponding to a 346 amino acid protein likely to be - according to a BLAST search - a 3-hydroxyisobutyrate dehydrogenase. The top BLAST hit, XP_004293980, is actually identical to our predicted protein, and comes from a previous annotation of Fragaria Vesca.

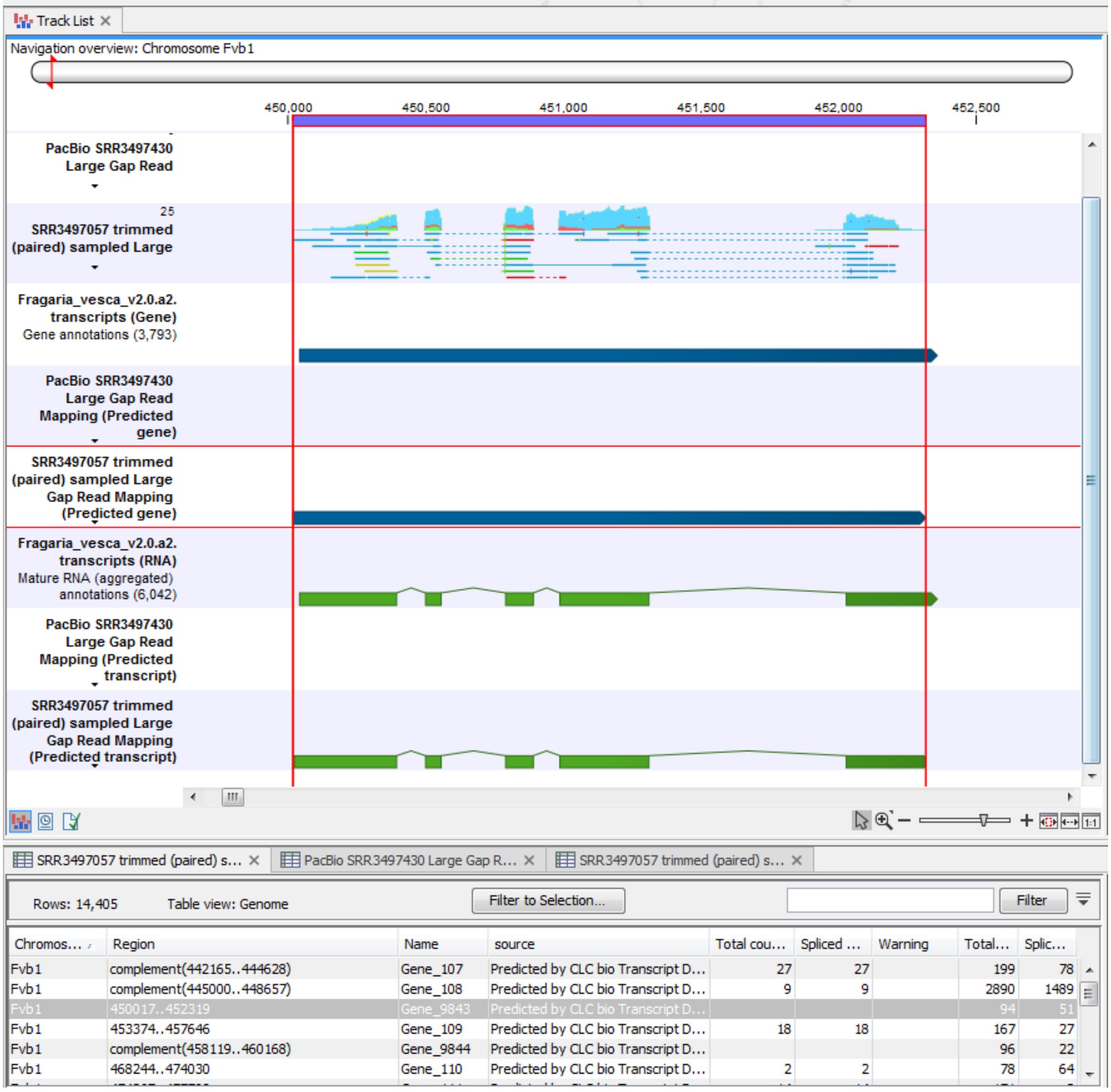

Track List
Navigation overview: Chromosome Fvb3
00
6,106,500
6,107,000
6,107,500
6,108,000
6,108,500
6,109,000
6,109,
45
PacBio SRR3497430 Large Gap Read
77
SRR3497057 trimmed (paired) sampled Large
Fragaria_vesca_v2.0.a2. transcripts (Gene)
Gene annotations (5,063)
PacBio SRR3497430 Large Gap Read Mapping (Predicted gene)
SRR3497057 trimmed (paired) sampled Large Gap Read Mapping (Predicted gene)
Fragaria_vesca_v2.0.a2. transcripts (RNA)
Mature RNA (aggregated) annotations (8,292)
PacBio SRR3497430 Large Gap Read Mapping (Predicted transcript)
SRR3497057 trimmed (paired) sampled Large Gap Read Mapping (Predicted transcript)
SRR3497057 trimmed (paired) s...
PacBio SRR3497430 Large Gap R...
Filter to Selection...
Filter
Create Track from Selection

Rows: 14,405 Table view: Genome

Chromosome	Region	Name	source	Total cou...	Spliced co...	Warning	Total cou...	Spliced co...
Fvb3	6094301..6099989	Gene_3099	Predicted by CLC bio Transcript Disco...	36	36		808	596
FvbUn	complement(1159703..11642...	Gene_31	Predicted by CLC bio Transcript Disco...	15	15		52	23
Fvb1	3349674..3353554	Gene_310	Predicted by CLC bio Transcript Disco...	5	5		71	44
Fvb3	6100821..6105844	Gene_3100	Predicted by CLC bio Transcript Disco...	17	17		59	7
Fvb3	6106055..6109470	Gene_3101	Predicted by CLC bio Transcript Disco...	38	38		338	280

Figure 12: *Annotation tracks seen in a Track List, with the SRR3497057 Predicted Gene annotations opened as a table in split view.*

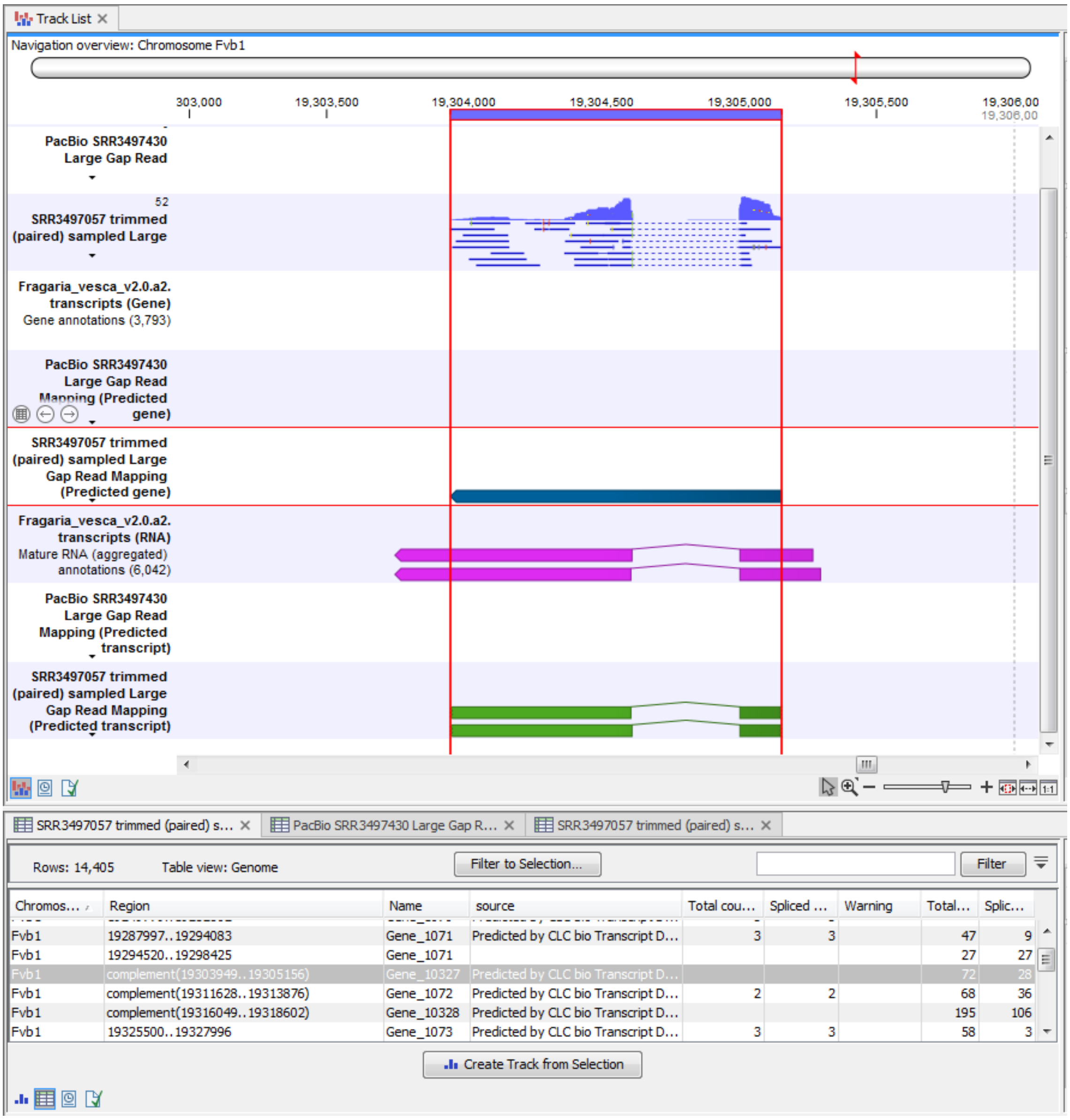
Figure 13 and figure 14 show examples of transcripts being discovered only based on Illumina reads. In the case of figure 14, our analysis was able to detect a long non coding RNA.


Track List
Navigation overview: Chromosome Fvb1
450,000
450,500
451,000
451,500
452,000
452,500
PacBio SRR3497430 Large Gap Read
25
SRR3497057 trimmed (paired) sampled Large
Fragaria_vesca_v2.0.a2. transcripts (Gene)
Gene annotations (3,793)
PacBio SRR3497430 Large Gap Read Mapping (Predicted gene)
SRR3497057 trimmed (paired) sampled Large Gap Read Mapping (Predicted gene)
Fragaria_vesca_v2.0.a2. transcripts (RNA)
Mature RNA (aggregated) annotations (6,042)
PacBio SRR3497430 Large Gap Read Mapping (Predicted transcript)
SRR3497057 trimmed (paired) sampled Large Gap Read Mapping (Predicted transcript)
SRR3497057 trimmed (paired) s...
PacBio SRR3497430 Large Gap R...
SRR3497057 trimmed (paired) s...
Filter to Selection...
Filter

Rows: 14,405 Table view: Genome

Chromos...	Region	Name	source	Total cou...	Spliced ...	Warning	Total...	Splic...
Fvb1	complement(442165..444628)	Gene_107	Predicted by CLC bio Transcript D...	27	27		199	78
Fvb1	complement(445000..448657)	Gene_108	Predicted by CLC bio Transcript D...	9	9		2890	1489
Fvb1	450017..452319	Gene_9843	Predicted by CLC bio Transcript D...				94	51
Fvb1	453374..457646	Gene_109	Predicted by CLC bio Transcript D...	18	18		167	27
Fvb1	complement(458119..460168)	Gene_9844	Predicted by CLC bio Transcript D...				96	22
Fvb1	468244..474030	Gene_110	Predicted by CLC bio Transcript D...	2	2		78	64

Figure 13: *An example of a gene and transcript that were discovered using Illumina reads only.*

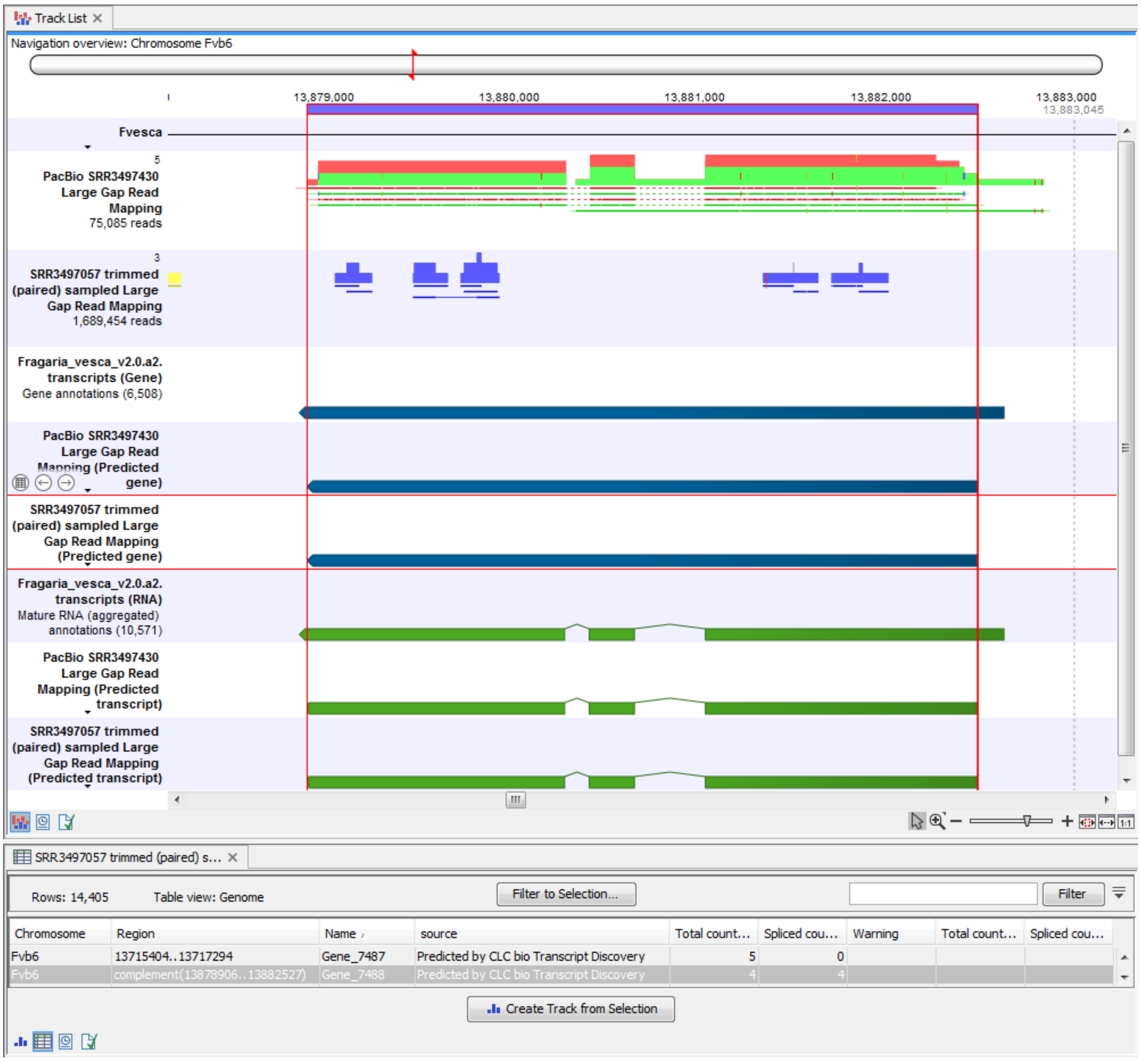

Track List
Navigation overview: Chromosome Fvb1
303,000
19,303,500
19,304,000
19,304,500
19,305,000
19,305,500
19,306,00
PacBio SRR3497430 Large Gap Read
52
SRR3497057 trimmed (paired) sampled Large
Fragaria_vesca_v2.0.a2. transcripts (Gene)
Gene annotations (3,793)
PacBio SRR3497430 Large Gap Read Mapping (Predicted gene)
SRR3497057 trimmed (paired) sampled Large Gap Read Mapping (Predicted gene)
Fragaria_vesca_v2.0.a2. transcripts (RNA)
Mature RNA (aggregated) annotations (6,042)
PacBio SRR3497430 Large Gap Read Mapping (Predicted transcript)
SRR3497057 trimmed (paired) sampled Large Gap Read Mapping (Predicted transcript)
SRR3497057 trimmed (paired) s...
PacBio SRR3497430 Large Gap R...
SRR3497057 trimmed (paired) s...
Filter to Selection...
Filter
Create Track from Selection

Rows: 14,405 Table view: Genome

Chromos...	Region	Name	source	Total cou...	Spliced ...	Warning	Total...	Splic...
Fvb1	19287997..19294083	Gene_1071	Predicted by CLC bio Transcript D...	3	3		47	9
Fvb1	19294520..19298425	Gene_1071					27	27
Fvb1	complement(19303949..19305156)	Gene_10327	Predicted by CLC bio Transcript D...				72	28
Fvb1	complement(19311628..19313876)	Gene_1072	Predicted by CLC bio Transcript D...	2	2		68	36
Fvb1	complement(19316049..19318602)	Gene_10328	Predicted by CLC bio Transcript D...				195	106
Fvb1	19325500..19327996	Gene_1073	Predicted by CLC bio Transcript D...	3	3		58	3

Figure 14: *A long non coding RNA is discovered using Illumina reads.*

On the other hand, because for Illumina reads we chose to "Ignore genes that do not have spliced reads", there are many genes without splice junctions in their transcripts that are only predicted because of the PacBio data. There are also a few examples where there was insufficient Illumina coverage to predict a gene, as in Gene_7488, which was found based on PacBio reads (figure 15).


Track List
Navigation overview: Chromosome Fvb6
13,879,000
13,880,000
13,881,000
13,882,000
13,883,000
13,883,045
Fvesca
5
PacBio SRR3497430 Large Gap Read Mapping
75,085 reads
3
SRR3497057 trimmed (paired) sampled Large Gap Read Mapping
1,689,454 reads
Fragaria_vesca_v2.0.a2. transcripts (Gene)
Gene annotations (6,508)
PacBio SRR3497430 Large Gap Read Mapping (Predicted gene)
SRR3497057 trimmed (paired) sampled Large Gap Read Mapping (Predicted gene)
Fragaria_vesca_v2.0.a2. transcripts (RNA)
Mature RNA (aggregated) annotations (10,571)
PacBio SRR3497430 Large Gap Read Mapping (Predicted transcript)
SRR3497057 trimmed (paired) sampled Large Gap Read Mapping (Predicted transcript)
SRR3497057 trimmed (paired) s...
Filter to Selection...
Filter
Create Track from Selection

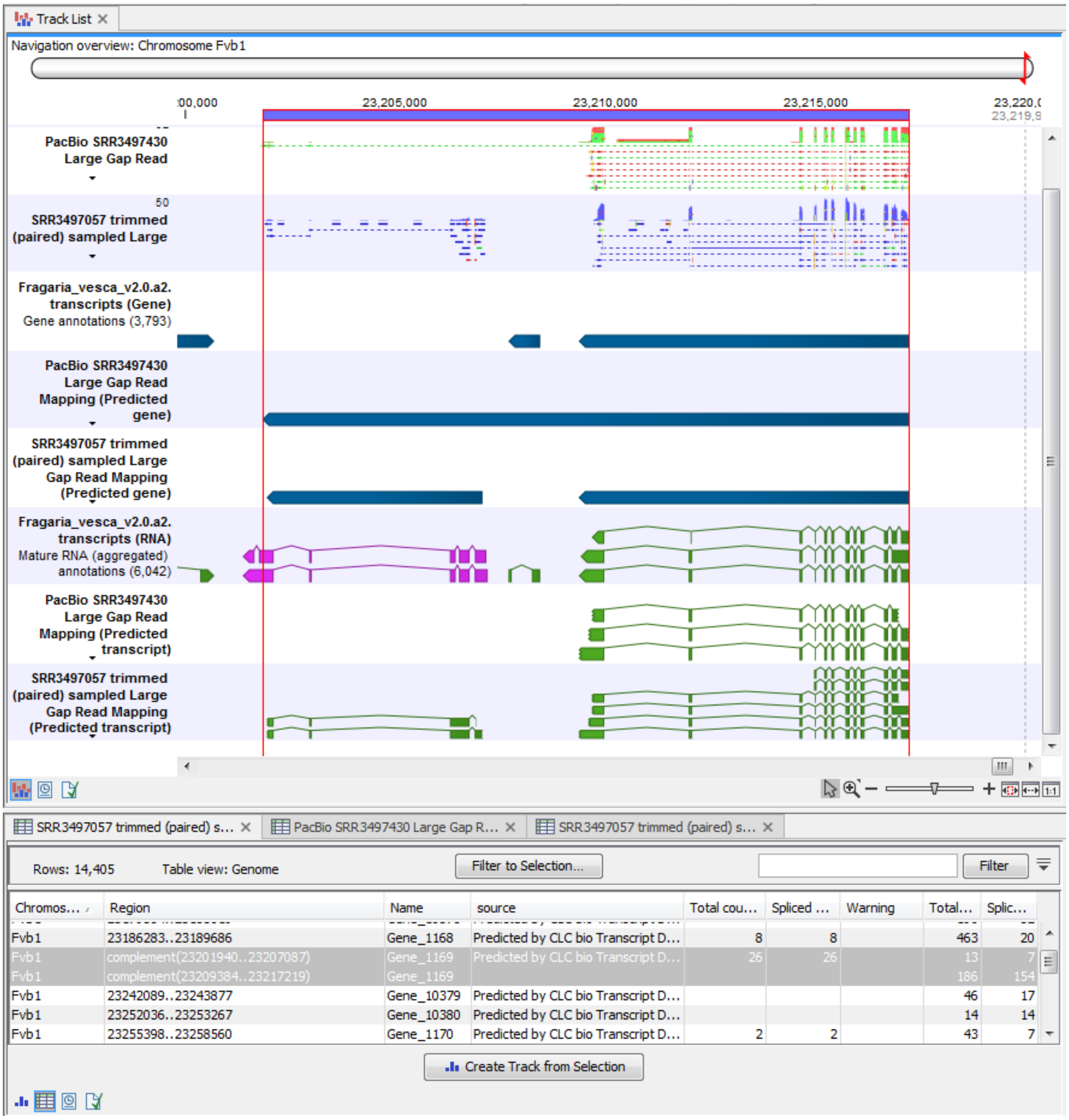
Rows: 14,405 Table view: Genome

Chromosome	Region	Name	source	Total count...	Spliced cou...	Warning	Total count...	Spliced cou...
Fvb6	13715404..13717294	Gene_7487	Predicted by CLC bio Transcript Discovery	5	0			
Fvb6	complement(13878906..13882527)	Gene_7488	Predicted by CLC bio Transcript Discovery	4	4			

Figure 15: *Gene and transcript can be discovered using Illumina reads only.*

Figure 16 illustrates the complementarity of long reads and Illumina reads. In this example, an overly long gene predicted from the PacBio data is later corrected using the Illumina data. Note that the Transcript Discovery tool will never change an input transcript, but it will correct input genes. Here we see a case where the PacBio data includes a single read that splices across two genes (in green at the top left of the screenshot). This is an example of transcriptional read-through, or possibly a technical artifact [Yuan et al., 2017]. The option "Ignore chimeric reads" attempts to detect and remove these reads, but it was not successful in this case.

Since the Illumina data has no read-through, the overly long PacBio gene is split in two in the Illumina annoation track: a gene for a long non-coding RNA on the left of the figure, and a protein-coding gene on the right.


Track List
Navigation overview: Chromosome Fvb1
:00,000
23,205,000
23,210,000
23,215,000
23,220,(
23,219,9
PacBio SRR3497430 Large Gap Read
50
SRR3497057 trimmed (paired) sampled Large
Fragaria_vesca_v2.0.a2. transcripts (Gene)
Gene annotations (3,793)
PacBio SRR3497430 Large Gap Read Mapping (Predicted gene)
SRR3497057 trimmed (paired) sampled Large Gap Read Mapping (Predicted gene)
Fragaria_vesca_v2.0.a2. transcripts (RNA)
Mature RNA (aggregated) annotations (6,042)
PacBio SRR3497430 Large Gap Read Mapping (Predicted transcript)
SRR3497057 trimmed (paired) sampled Large Gap Read Mapping (Predicted transcript)
SRR3497057 trimmed (paired) s...
PacBio SRR3497430 Large Gap R...
SRR3497057 trimmed (paired) s...
Filter to Selection...
Filter
Create Track from Selection

Rows: 14,405 Table view: Genome

Chromos...	Region	Name	source	Total cou...	Spliced ...	Warning	Total...	Splic...
Fvb1	23186283..23189686	Gene_1168	Predicted by CLC bio Transcript D...	8	8		463	20
Fvb1	complement(23201940..23207087)	Gene_1169	Predicted by CLC bio Transcript D...	26	26		13	7
Fvb1	complement(23209384..23217219)	Gene_1169					186	154
Fvb1	23242089..23243877	Gene_10379	Predicted by CLC bio Transcript D...				46	17
Fvb1	23252036..23253267	Gene_10380	Predicted by CLC bio Transcript D...				14	14
Fvb1	23255398..23258560	Gene_1170	Predicted by CLC bio Transcript D...	2	2		43	7

Figure 16: *A long non gene is being discovered using PacBio long reads.*

Using the Transcript Discovery plugin, we quickly annotated de novo the *Fragaria vesca* genome, and obtained results comparable to those published by Li et al., 2018. This demonstrates the value of our seamless solution as a means to validate a more complex workflow, or even as a replacement for a series of tools not specifically designed to work with each other. This tutorial also emphasizes the value of working with both long and short reads when performing a de novo annotation.

Bibliography

[Li et al., 2018] Li, Y., Wei, W., Feng, J., Luo, H., Pi, M., Liu, Z., and Kang, C. (2018). Genome re-annotation of the wild strawberry fragaria vesca using extensive illumina- and smrt-based rna-seq datasets. *DNA Research*, 25(1):61–70.

[Yuan et al., 2017] Yuan, C., Han, Y., Zellmer, L., Yang, W., Guan, Z., Yu, W., Huang, H., and Liao, D. J. (2017). It is imperative to establish a pellucid definition of chimeric rna and to clear up a lot of confusion in the relevant research. *International Journal of Molecular Sciences*, 4:714.