Finding nearest genomic features

In this section, you will learn how to find genomic features (genes) that are near the IP-enriched regions of the data. You will also learn how to classify the peak locations by gene section (5' UTR, 3' UTR, Promoter, exon, intron).

Finding the nearest genomic features

- Select **p-value_filtered** from the spreadsheet tree
- Select **Find Nearest Genomic Feature** from the *Peak Analysis* section of the *ChIP-Seq* workflow

The *Output Overlapping Features* dialog will open (Figure 1).

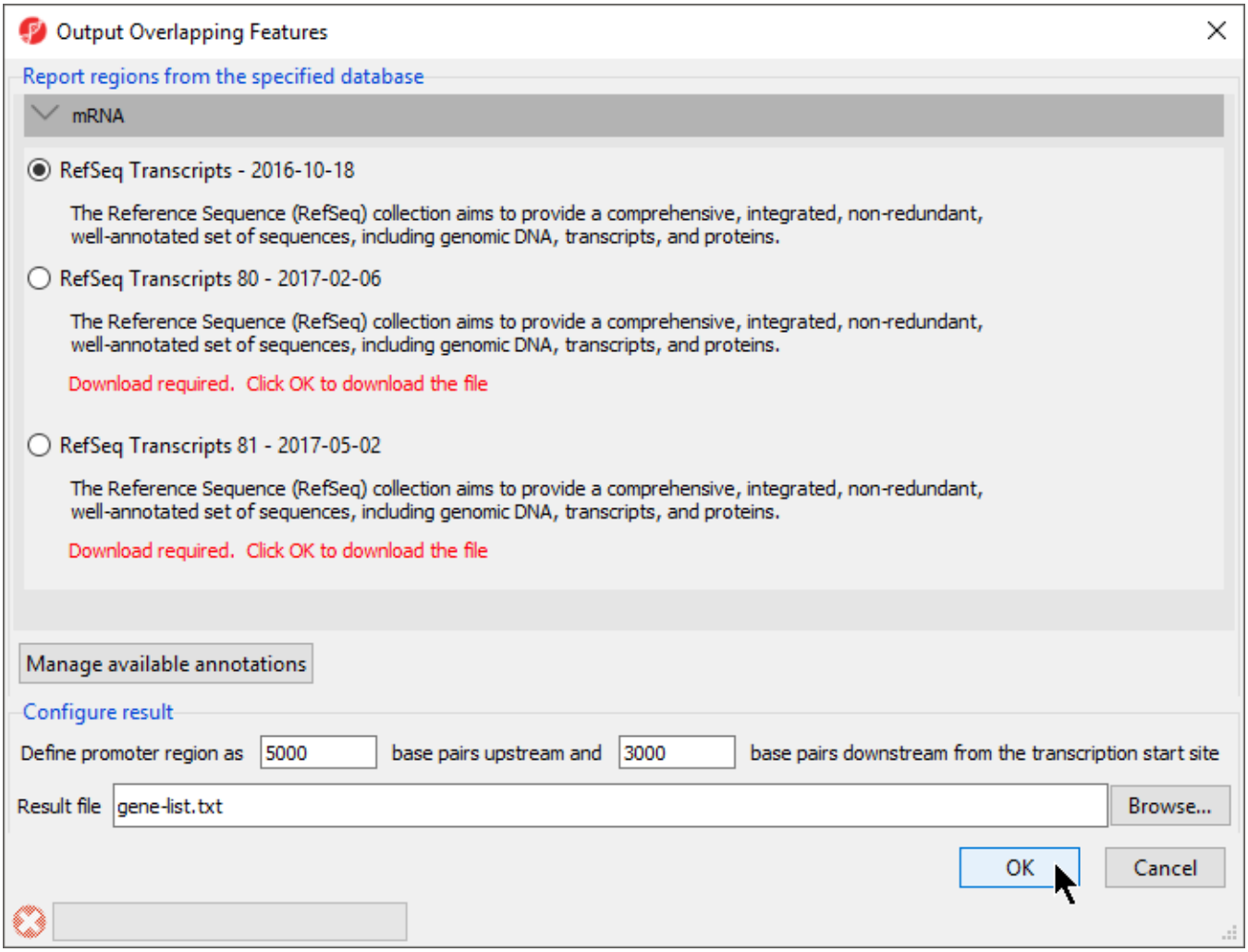


Figure 4. Selecting a database for overlapping features

With this dialog, you can specify the reference database.

- Select *RefSeq Transcripts 81 - 2017-08-02* or your preferred annotation database

The promoter region can also be defined. The default settings are appropriate in this case.

- Select **OK**

The resulting spreadsheet, *gene-list,* is a child of the *p-value_filtered* spreadsheet (Figure 2). Each row represents a transcript.

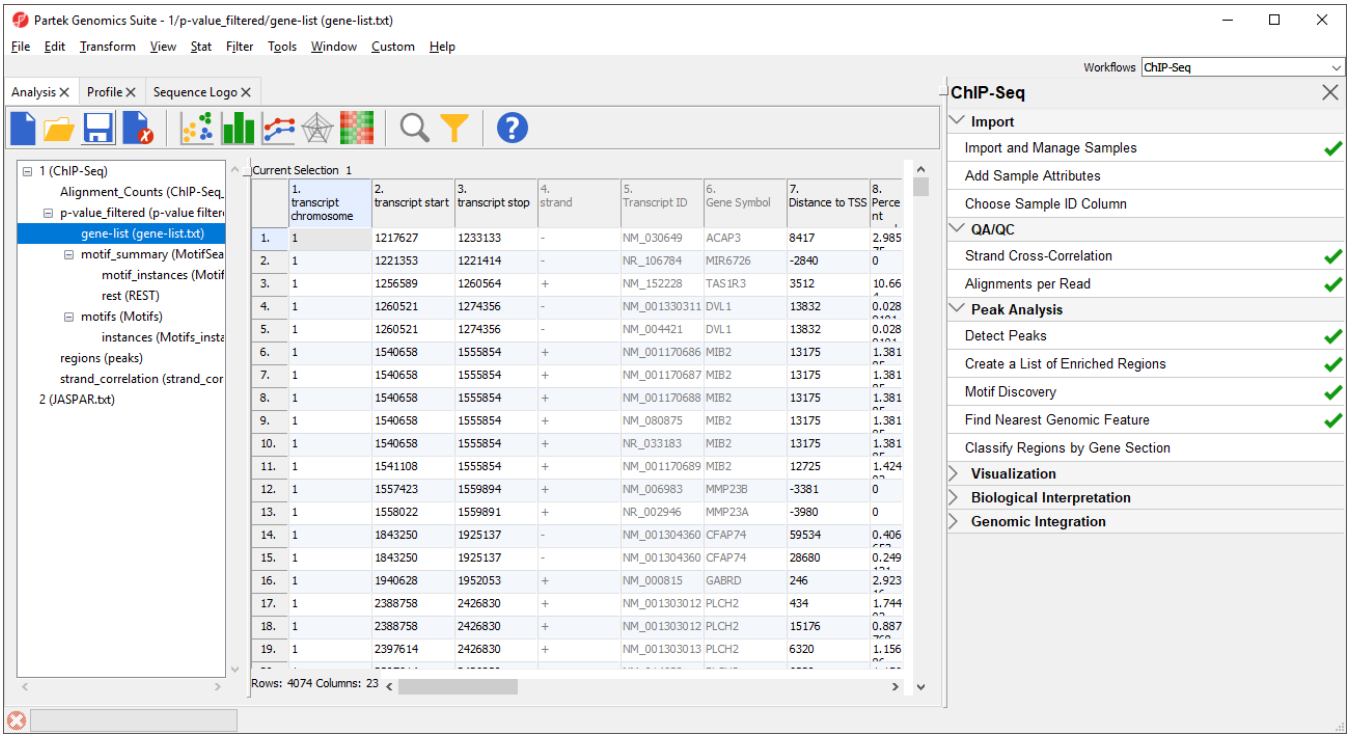


Figure 5. Viewing genes overlapped by regions

Column *1. transcript chromosome* gives the chromosome location of transcript

Column *2. transcript start* gives the start of transcript (inclusive)

Column *3. transcript stop* gives the end of transcript (exclusive)

Column *4. strand* gives the strand of the transcript

Column *5. Transcript ID* gives the identify of the transcript

Column *6. Gene Symbol* gives the identity of the gene

Column *7. Distance to TSS* gives the distance of each enriched region to the transcription start site in base pairs with positive indicating downstream and negative indicating upstream

Column 8. Percent overlap with gene gives the percent of the gene that overlaps with the region

Column *9. Percent overlap with region* gives the percent of the region that overlaps with the gene

Column *10.-23.* These columns are detailed in Detecting peaks and enriched regions in ChIP-Seq data

Percent overlap with gene is more likely to close to 1 in cases where one region covers several genes, in histone studies, for example. *Percent overlap with region* is likely to be close to 1 in cases where a region is relatively small and is found completely within a gene, in transcription factor binding studies, for example. If both columns are close to 1, then the gene and the region have nearly the same start and stop sites. If both columns are close to 0, then the region does not overlap with the gene directly and likely covers only the promoter region.

Classifying regions by gene section

Another way to interpret the genomic location of peaks is to use *Classify regions by gene selection*.

- Select **p-value_filtered** from the spreadsheet tree
- Select **Classify regions by gene selection** from the *Peak Analysis* section of the *ChIP-Seq* workflow

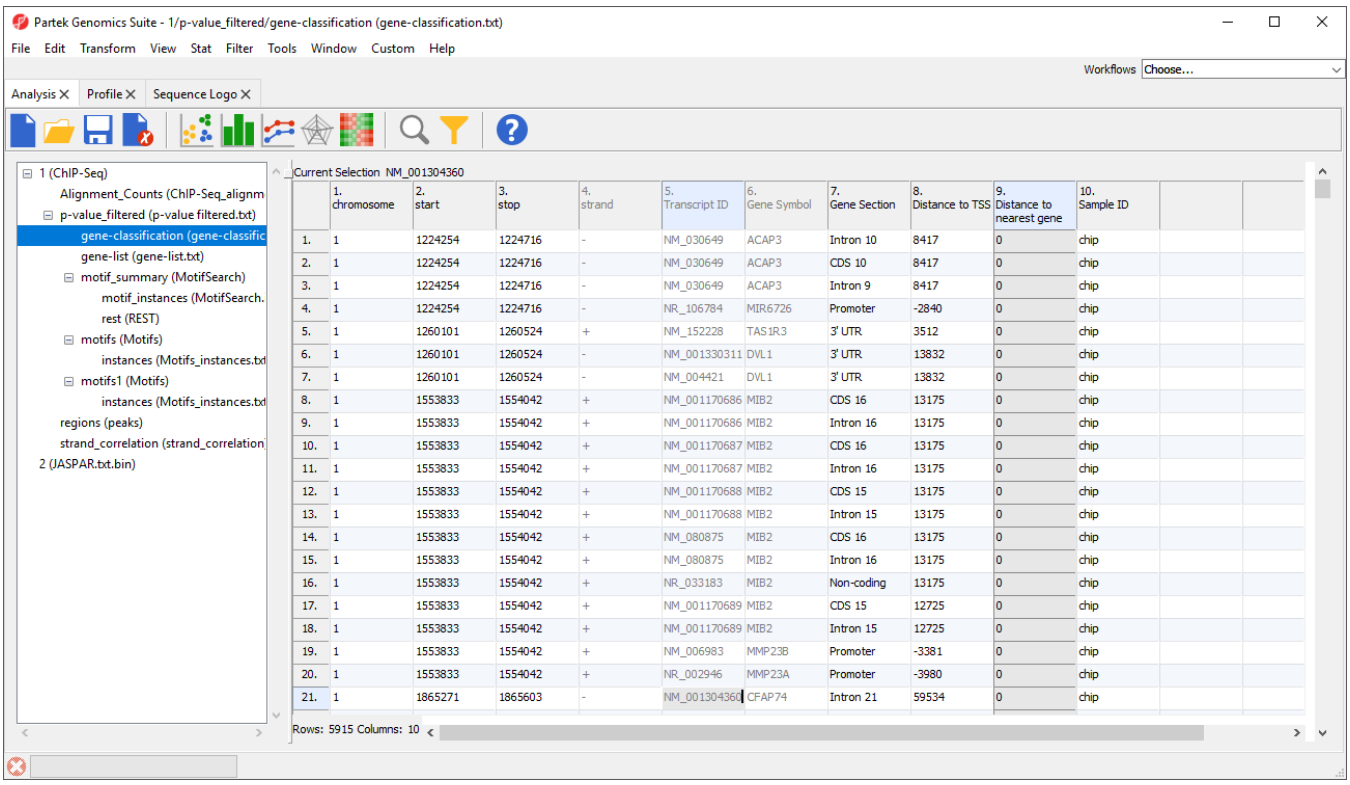
The *Output Overlapping Features* dialog will open.

- Select *RefSeq Transcripts 81 - 2017-08-02* or your preferred annotation database

The promoter region can also be defined. The default settings are appropriate in this case. The results can be further configured to give one result per detected region or one result per genomic feature. The default setting, one result per detected region, is appropriate in this case.

- Select **OK**

A new spreadsheet, *gene-classification* will be generated (Figure 3).



	1. chromosome	2. start	3. stop	4. strand	5. Transcript ID	6. Gene Symbol	7. Gene Section	8. Distance to TSS	9. Distance to nearest gene	10. Sample ID
1.	1	1224254	1224716	-	NM_030649	ACAP3	Intron 10	8417	0	chip
2.	1	1224254	1224716	-	NM_030649	ACAP3	CDS 10	8417	0	chip
3.	1	1224254	1224716	-	NM_030649	ACAP3	Intron 9	8417	0	chip
4.	1	1224254	1224716	-	NR_106784	MIR6726	Promoter	-2840	0	chip
5.	1	1260101	1260524	+	NM_152228	TAS1R3	3' UTR	3512	0	chip
6.	1	1260101	1260524	-	NM_001330311	DVL1	3' UTR	13832	0	chip
7.	1	1260101	1260524	-	NM_004421	DVL1	3' UTR	13832	0	chip
8.	1	1553833	1554042	+	NM_001170686	MIB2	CDS 16	13175	0	chip
9.	1	1553833	1554042	+	NM_001170686	MIB2	Intron 16	13175	0	chip
10.	1	1553833	1554042	+	NM_001170687	MIB2	CDS 16	13175	0	chip
11.	1	1553833	1554042	+	NM_001170687	MIB2	Intron 16	13175	0	chip
12.	1	1553833	1554042	+	NM_001170688	MIB2	CDS 15	13175	0	chip
13.	1	1553833	1554042	+	NM_001170688	MIB2	Intron 15	13175	0	chip
14.	1	1553833	1554042	+	NM_080875	MIB2	CDS 16	13175	0	chip
15.	1	1553833	1554042	+	NM_080875	MIB2	Intron 16	13175	0	chip
16.	1	1553833	1554042	+	NR_033183	MIB2	Non-coding	13175	0	chip
17.	1	1553833	1554042	+	NM_001170689	MIB2	CDS 15	12725	0	chip
18.	1	1553833	1554042	+	NM_001170689	MIB2	Intron 15	12725	0	chip
19.	1	1553833	1554042	+	NM_006983	MMP23B	Promoter	-3381	0	chip
20.	1	1553833	1554042	+	NR_002946	MMP23A	Promoter	-3980	0	chip
21.	1	1865271	1865603	-	NM_001304360	CFAP74	Intron 21	59534	0	chip

Figure 6. Classifying regions by gene section

Columns 1-6 have the same contents we saw in *gene-list*.

Column *7. Gene Section* gives the section of the gene that overlaps with the region

Column *8. Distance to TSS* gives the distance of each enriched region to the transcription start site in base pairs with positive indicates downstream and negative indicating upstream

Column *9. Distance to nearest gene* gives the distance of each enriched region to the nearest gene in base pairs with positive indicating downstream and negative indicating upstream

Column *10. Sample ID* gives the sample in which the region is enriched

« Identifying novel and known motifs Visualizing reads and enriched regions »

Additional Assistance

If you need additional assistance, please visit our support page to submit a help ticket or find phone numbers for regional support.

Your Rating: Results: 36 rates