Finding nearest genomic features

- · Finding the nearest genomic features
- Classifying regions by gene section

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).

Ø Output Overlapping Features	×
Report regions from the specified database	
∼ mRNA	
RefSeq Transcripts - 2016-10-18	
The Reference Sequence (RefSeq) collection aims to provide a comprehensive, integrated, non-redundant, well-annotated set of sequences, including genomic DNA, transcripts, and proteins.	
O RefSeq Transcripts 80 - 2017-02-06	
The Reference Sequence (RefSeq) collection aims to provide a comprehensive, integrated, non-redundant, well-annotated set of sequences, including genomic DNA, transcripts, and proteins.	
Download required. Click OK to download the file	
O RefSeq Transcripts 81 - 2017-05-02	
The Reference Sequence (RefSeq) collection aims to provide a comprehensive, integrated, non-redundant, well-annotated set of sequences, including genomic DNA, transcripts, and proteins.	
Download required. Click OK to download the file	
Manage available annotations	
Configure result	
Define promoter region as 5000 base pairs upstream and 3000 base pairs downstream from the transcription start	site
Result file gene-list.txt Browse	ż
OK Cance	!

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.

Partek Genomics Suite - 1/p-value_filtered/g File Edit Transform View State Filter T	jene-list (gene-lis	it.txt)							-	
The Fair Transform Mew Star Lifes in			,						Workflows ChIP-Seq	~
Analysis × Profile × Sequence Logo ×									-ChIP-Seg	×
			0						∑ Import	
📕 🦰 🗖 💽 🔛 🖉	Import and Manage Samples									
🗉 1 (ChIP-Seq)	ent Selection 1							^	Add Sample Attributes	
Alignment_Counts (ChIP-Seq_	1.	nscript romosome	3. transcript stop	4. strand	5. Transcript ID	6. Gene Symbol	7. Distance to TSS	8. Dorce	Chasse Correla ID Column	
p-value_filtered (p-value filtered)	chromosome							nt	Choose Sample ID Column	
gene-list (gene-list.txt) 1.	1	1217627	1233133	-	NM_030649	ACAP3	8417	2.985	∼ qa/qc	
 motif_summary (MotifSea 2. 	1	1221353	1221414	-	NR_106784	MIR6726	-2840	0	Strand Cross-Correlation	 ✓
motif_instances (Motif 3.	1	1256589	1260564	+	NM_152228	TAS1R3	3512	10.66	Alignments per Read	 Image: A second s
rest (KEST) 4.	1	1260521	1274356	-	NM_001330311	DVL1	13832	0.028	✓ Peak Analysis	
instances (Motifs insta	1	1260521	1274356	-	NM_004421	DVL1	13832	0.028	Detect Peaks	
regions (peaks) 6.	1	1540658	1555854	+	NM_001170686	MIB2	13175	1.381	Create a List of Enriched Pegians	
strand_correlation (strand_cor 7.	1	1540658	1555854	+	NM_001170687	MIB2	13175	1.381		
2 (JASPAR.txt) 8.	1	1540658	1555854	+	NM_001170688	MIB2	13175	1.381	Motif Discovery	 ✓
9.	1	1540658	1555854	+	NM_080875	MIB2	13175	1.381	Find Nearest Genomic Feature	 Image: A second s
10.	. 1	1540658	1555854	+	NR_033183	MIB2	13175	1.381	Classify Regions by Gene Section	
11.	. 1	1541108	1555854	+	NM_001170689	MIB2	12725	1.424	> Visualization	
12.	. 1	1557423	1559894	+	NM_006983	MMP23B	-3381	0	Biological Interpretation	
13.	. 1	1558022	1559891	+	NR_002946	MMP23A	-3980	0	Genomic Integration	
14.	. 1	1843250	1925137	-	NM_001304360	CFAP74	59534	0.406	,	
15.	. 1	1843250	1925137	-	NM_001304360	CFAP74	28680	0.249		
16.	. 1	1940628	1952053	+	NM_000815	GABRD	246	2.923		
17.	. 1	2388758	2426830	+	NM_001303012	PLCH2	434	1.744		
18.	. 1	2388758	2426830	+	NM_001303012	PLCH2	15176	0.887		
19.	. 1	2397614	2426830	+	NM_001303013	PLCH2	6320	1.156		
· · · · · · · · · · · · · · · · · · ·	4074 Columnau									
< > Kows	S. TO/T COlumns: .	23 <						> ~		

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).

analysis X Profile X Sequence I ono X										Workflows Choose	
) 🗁 🗔 🔥 🔛 🛃 🖄		Q Y	0								
1 (ChIP-Seq)	Current Selection NM	001304360									
Alignment_Counts (ChIP-Seq_alignm p-value_filtered (p-value filtered.txt)	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	
gene-classification (gene-classific	1. 1	1224254	1224716	-	NM_030649	ACAP3	Intron 10	8417	0	chip	
gene-list (gene-list.txt)	2. 1	1224254	1224716	-	NM_030649	ACAP3	CDS 10	8417	0	chip	
 motif_summary (MotifSearch) motif_instances (MotifSearch) 	3. 1	1224254	1224716	-	NM_030649	ACAP3	Intron 9	8417	0	chip	
rest (REST)	4. 1	1224254	1224716	-	NR_106784	MIR6726	Promoter	-2840	0	chip	
motifs (Motifs)	5. 1	1260101	1260524	+	NM_152228	TAS1R3	3' UTR	3512	0	chip	
instances (Motifs_instances.txl	6. 1	1260101	1260524	-	NM_001330311	DVL1	3' UTR	13832	0	chip	
 motifs1 (Motifs) 	7. 1	1260101	1260524	-	NM_004421	DVL1	3' UTR	13832	0	chip	
instances (Motifs_instances.txl	8. 1	1553833	1554042	+	NM_001170686	MIB2	CDS 16	13175	0	chip	
regions (peaks)	9. 1	1553833	1554042	+	NM_001170686	MIB2	Intron 16	13175	0	chip	
strand_correlation (strand_correlation	10. 1	1553833	1554042	+	NM_001170687	MIB2	CDS 16	13175	0	chip	
2 (JASPAR.txt.bin)	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.



Copyright © 2018 by Partek Incorporated. All Rights Reserved. Reproduction of this material without express written consent from Partek Incorporated is strictly prohibited.