Finding genes with copy number variation

With a list of amplified or deleted regions in our cohort in hand, one of the more interesting questions to ask is what genes have recurrent amplifications or deletions in the data set. To address this question, we can use the *Find overlapping genes* function to either add a column to our region list with the genes present in each region or create a new list of genes that overlap the regions.

Here, we will create a new spreadsheet with genes that overlap the regions in the *amplified_or_deleted* spreadsheet.

- Select the **amplified_or_deleted** spreadsheet in the spreadsheet tree
- Select **Find Overlapping Genes** from the *Copy Number Analysis* section of the workflow
- Select **Create a New Spreadsheet with Genes that Overlap the Regions** from the *Find Overlapping Genes* dialog (Figure 1)
- Select **OK**

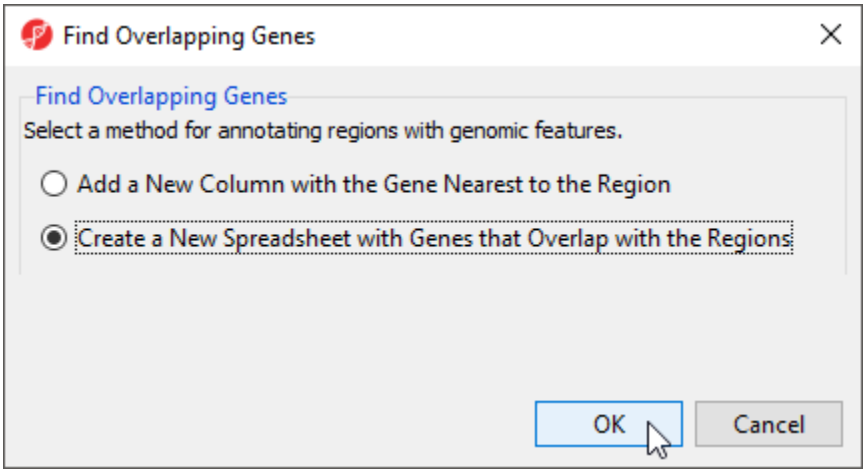


Figure 4. Options in Find Overlapping Genes dialog

To determine what regions in the genome correspond to genes, we need to select an annotation database (Figure 2).

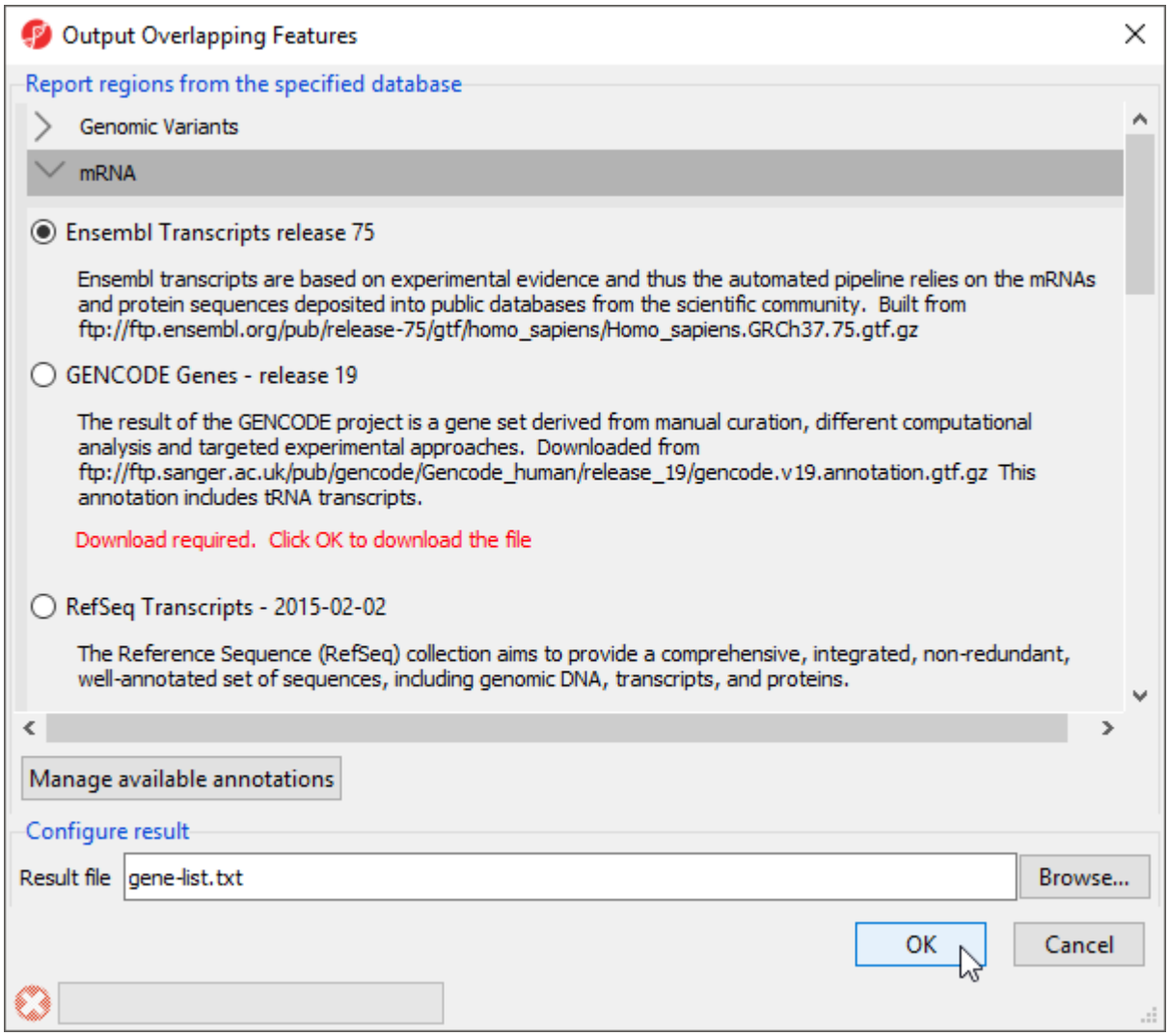
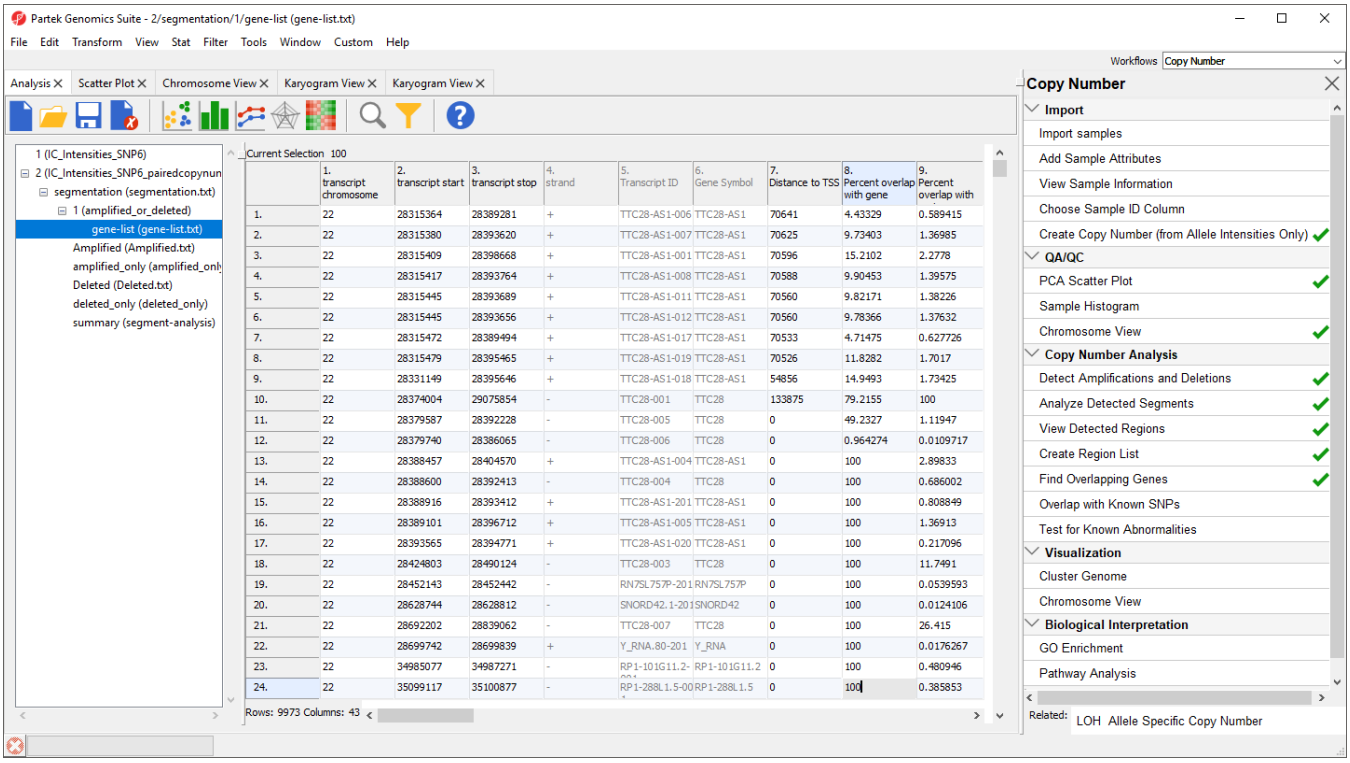


Figure 5. Viewing the Output Overlapping Features dialog. Database files not present on the computer display Download required in red

Partek Genomics Suite offers a variety of possibilities including RefSeq, Ensembl, and GENCODE; however, custom annotations can also be used. If the database file has not been downloaded, *Download required. Click OK to download the file*, will be listed in red beneath the annotation. Selecting *OK* will automatically download the file and then run the task.

- Select **Ensembl Transcripts release 75**
- Select **OK**

A new spreadsheet, *gene-list*, is created as a child spreadsheet of *amplified_or_deleted* (Figure 3).



	1. transcript chromosome	2. transcript start	3. transcript stop	4. strand	5. Transcript ID	6. Gene Symbol	7. Distance to TSS	8. Percent overlap with gene	9. Percent overlap with
1.	22	28315364	28389281	+	TTC28-AS1-006	TTC28-AS1	70641	4.43329	0.589415
2.	22	28315380	28393620	+	TTC28-AS1-007	TTC28-AS1	70625	9.73403	1.36985
3.	22	28315409	28398668	+	TTC28-AS1-001	TTC28-AS1	70596	15.2102	2.2778
4.	22	28315417	28393764	+	TTC28-AS1-008	TTC28-AS1	70588	9.90453	1.39575
5.	22	28315445	28393689	+	TTC28-AS1-011	TTC28-AS1	70560	9.82171	1.38226
6.	22	28315445	28393656	+	TTC28-AS1-012	TTC28-AS1	70560	9.78366	1.37632
7.	22	28315472	28389494	+	TTC28-AS1-017	TTC28-AS1	70533	4.71475	0.627726
8.	22	28315479	28395465	+	TTC28-AS1-019	TTC28-AS1	70526	11.8282	1.7017
9.	22	28331149	28395646	+	TTC28-AS1-018	TTC28-AS1	54856	14.9493	1.73425
10.	22	28374004	29075854	-	TTC28-001	TTC28	133875	79.2155	100
11.	22	28379587	28392228	-	TTC28-005	TTC28	0	49.2327	1.11947
12.	22	28379740	28386065	-	TTC28-006	TTC28	0	0.964274	0.0109717
13.	22	28388457	28404570	+	TTC28-AS1-004	TTC28-AS1	0	100	2.89833
14.	22	28388600	28392413	-	TTC28-004	TTC28	0	100	0.686002
15.	22	28388916	28393412	+	TTC28-AS1-201	TTC28-AS1	0	100	0.808849
16.	22	28389101	28396712	+	TTC28-AS1-005	TTC28-AS1	0	100	1.36913
17.	22	28393565	28394771	+	TTC28-AS1-020	TTC28-AS1	0	100	0.217096
18.	22	28424803	28490124	-	TTC28-003	TTC28	0	100	11.7491
19.	22	28452143	28452442	-	RN7SL757P-201	RN7SL757P	0	100	0.0539593
20.	22	28628744	28628812	-	SNORD42.1-201	SNORD42	0	100	0.0124106
21.	22	28692202	28839062	-	TTC28-007	TTC28	0	100	26.415
22.	22	28699742	28699839	+	Y_RNA.80-201	Y_RNA	0	100	0.0176267
23.	22	34985077	34987271	-	RP1-101G11.2-	RP1-101G11.2	0	100	0.480946
24.	22	35099117	35100877	-	RP1-288L1.5-00	RP1-288L1.5	0	100	0.385853

Figure 6. Viewing the gene-list spreadsheet, a result of overlapping genes with regions of copy number changes. Each row of the table represents one Ensembl transcript

Each row corresponds to a transcript and the columns are as follows:

1. Genomic coordinates of the transcript

4. Coding strand

5. Transcript ID

6. Gene Symbol

7. Minimum distance of the region to the transcription start site with positive values indicating downstream and negative values indicating upstream

8. Percent overlap with gene indicates how much of the transcript sequence overlaps the region

9. Percent overlap with region indicates how much of the region is overlapped by the transcript

10. + Correspond to the columns 1+ in the *segment-analysis* spreadsheet

This *gene-list* spreadsheet is gene-centric and enables genomic integration. For example, GO and Pathway enrichment can be directly invoked on the *gene -list* spreadsheet to detect functional groups affected by copy number changes. While not detailed in this tutorial, please feel free to explore these options on your own. For rmore information on enrichment analysis, you can consult the Gene Ontology Enrichment tutorial.

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: 34 rates