**Clustering Script Output Columns**

Once clustering is complete, the script outputs a single CSV with the following columns (plus any expanded window columns for C/T or G content). Below is a description of each:

1. **Start**
   * **Definition**: The starting genomic coordinate of the identified cluster region.
2. **Center**
   * **Definition**: The midpoint coordinate of the cluster region ((Start + End) / 2, rounded down).
3. **End**
   * **Definition**: The ending genomic coordinate of the cluster region.
4. **Strand**
   * **Definition**: + for forward or - for reverse. This is inferred based on whether the cluster originates from the forward or reverse log2 FC data.
5. **Split/Merged**
   * **Definition**: Indicates if the cluster was split or merged with others during processing. “Yes” if multiple pseudo-clusters were merged; “No” otherwise.
6. **Max\_WT\_Ratio**
   * **Definition**: The highest observed WT (wild-type) ratio within the cluster.
7. **Max\_WT\_Ratio\_Coord**
   * **Definition**: The genomic coordinate where the **maximum WT ratio** was observed. If it spans multiple positions, this value represents the center of that contiguous region.
8. **Average\_WT\_Ratio**
   * **Definition**: The average WT ratio across all nucleotides that fall within the cluster boundaries.
9. **Max\_log2\_FC**
   * **Definition**: The maximum observed log2 fold change within the cluster.
10. **Max\_log2\_FC\_Coord**
    * **Definition**: The genomic coordinate where the **maximum log2 fold change** was observed. If it spans multiple positions, this is the center of that contiguous region.
11. **Average\_log2\_FC**
    * **Definition**: The average log2 fold change across all nucleotides within the cluster.
12. **Upstream\_Seq**
    * **Definition**: The 100-nucleotide sequence **upstream** of the cluster’s **center** coordinate.
    * **Note**: For the forward strand, it’s the direct upstream sequence; for the reverse strand, it’s reverse-complemented as needed.
13. **Downstream\_Seq**
    * **Definition**: The 100-nucleotide sequence **downstream** of the cluster’s **center** coordinate.
    * **Note**: Reverse-complemented if the strand is -.
14. **Upstream\_From\_Start\_Seq**
    * **Definition**: The 100-nucleotide sequence **upstream** of the cluster’s **Start** coordinate.
15. **Downstream\_From\_Start\_Seq**
    * **Definition**: The 100-nucleotide sequence **downstream** of the cluster’s **Start** coordinate.
16. **Gene**
    * **Definition**: The gene name associated with the cluster region, if overlapping any non-pseudogene. If it’s fully intergenic or overlapping a pseudogene, this will say “Intergenic”.
17. **Sense\_or\_Anti**
    * **Definition**: Indicates if the cluster is on the **sense** or **antisense** strand relative to the overlapping gene (if any). Empty if it’s intergenic.
18. **Intergenic\_or\_AKA\_Code**
    * **Definition**: If intergenic, lists the names of the nearest upstream and downstream genes in the format UpstreamGene,DownstreamGene. Otherwise, left empty.
19. **Cluster\_Length**
    * **Definition**: The total length of the cluster in nucleotides (End - Start).
20. **CT\_Upstream\_Content**
    * **Definition**: The fraction/percentage of cytosine (C) plus thymine (T) bases in the **upstream** sequence (100 nt from the cluster’s Start coordinate).
    * **Range**: 0.0 to 1.0 (i.e., 0% to 100%).
21. **CT\_Downstream\_Content**
    * **Definition**: The fraction/percentage of C + T in the **downstream** sequence (100 nt from the cluster’s Startcoordinate).
22. **Closest\_TSS**
    * **Definition**: The coordinate of the **transcription start site** (TSS) closest to the cluster (based on TSS.csv).
    * **Note**: For forward clusters, we look at TSS entries with TSS\_direction = '+'; for reverse clusters, TSS\_direction = '-'.
23. **Closest\_ATG**
    * **Definition**: The genomic position of the **nearest ATG** (translation start) on the same strand as the cluster.
    * **Note**: We look through all CDS features in the GFF, extracting start or end based on strand.
24. **Highest\_CT\_Content\_Upstream\_Coord**
    * **Definition**: The coordinate (relative to the genome) of the **sliding window (30 nt)** in the **upstream** sequence that has the **highest CT content**.
    * The script calculates a sliding window of 30 nt over Upstream\_From Start\_Seq, identifies the window with the max CT content, and computes its center in genomic coordinates.
25. **Highest\_CT\_Content\_Downstream\_Coord**
    * **Definition**: Same as above, but in the **downstream** sequence.
26. **Lowest\_G\_Content\_Upstream\_Coord**
    * **Definition**: The coordinate of the **sliding window (30 nt)** in the **upstream** sequence that has the **lowest G content**.
27. **Lowest\_G\_Content\_Downstream\_Coord**
    * **Definition**: The coordinate of the **sliding window (30 nt)** in the **downstream** sequence that has the **lowest G content**.
28. **CT\_Upstream\_Window**\***\*** and **CT\_Downstream\_Window**\***\***
    * **Definition**: Expanded columns showing the C+T content for **each** position in a 30-nt sliding window within Upstream\_Seq or Downstream\_Seq.
    * These appear as multiple columns (e.g., CT\_Upstream\_Window1, CT\_Upstream\_Window2, ...) if you have enough sequence length for each window step.
29. **G\_Upstream\_Window**\***\*** and **G\_Downstream\_Window**\***\***
    * **Definition**: Similar expanded columns for **G** content in each 30-nt sliding window.

**Note**: The exact number of window columns depends on your upstream/downstream region length (100 nt) minus the window size (30 nt). Thus, you might see 70 columns for each type of content (CT or G) and region (Upstream or Downstream).