**HIRN Splice Event Data Guide**

**Overview**

The following R dataframes (.rds) are contained within the directory:

1. ***a3ss\_data.rds***: contains response and predictor data for all measured A3SS-type splice events
2. ***a3ss\_metadata.rds***: contains associated metadata for all measured A3SS-type splice events
3. ***a5ss\_data.rds***: contains response and predictor data for all measured A5SS-type splice events
4. ***a5ss\_metadata.rds***: contains associated metadata for all measured A5SS-type splice events
5. ***mxe\_data.rds***: contains response and predictor data for all measured MXE-type splice events
6. ***mxe\_metadata.rds***: contains associated metadata for all measured MXE-type splice events
7. ***ri\_data.rds***: contains response and predictor data for all measured RI-type splice events
8. ***ri\_metadata.rds***: contains associated metadata for all measured RI-type splice events
9. ***se\_data.rds***: contains response and predictor data for all measured SE-type splice events
10. ***se\_metadata.rds***: contains associated metadata for all measured SE-type splice events

A3SS refers to an alternative 3’ splice junction being used in the alternative splicing; A5SS to an alternative 5’ splice junction; MXE denotes a mutually exclusive exon event; RI a retained intron event; and SE a skipped exon event.

**Dataset details**

All data/metadata .rds pairs are formatted the same and contain largely the same set of variables, only specific to the corresponding splicing event. Nonetheless, descriptions of all of the contents of each dataset are subsequently provided.

1. ***a3ss\_data.rds***
   1. 8894 rows: Each corresponds to a unique A3SS splice event
   2. 265 columns:
      1. **Status\_TRAIN\_1 – Status\_TRAIN\_24**: columns containing the response data (i.e. case-control status) for each of the 24 samples in the training cohort.
      2. **MergeID**: Column containing the unique identifier for each splice event. This variable is used to merge a3ss\_data.rds and a3ss\_metadata.rds.
      3. **Status\_TEST\_1 – Status\_TEST\_24**: columns containing the response data (i.e. case-control status) for each of the 24 samples in the testing cohort.
      4. **IncLevel\_TRAIN\_1 – IncLevel\_TRAIN\_24**: columns containing inclusion level predictor data for each of the 24 training samples.
      5. **IncLevel\_TEST\_1 – IncLevel\_TEST\_24**: columns containing inclusion level predictor data for each of the 24 test samples, with missing values not imputed.
      6. **IncLevel\_TRAIN\_imputed\_1 – IncLevel\_TRAIN\_imputed\_24**: columns containing inclusion level predictor data for each of the 24 test samples, with missing values imputed based on the average of the observed training sample data.
      7. **IJC\_TRAIN\_1 – IJC\_TRAIN\_24**: columns containing inclusion junction count predictor data for each of the 24 training samples.
      8. **IJC\_TEST\_1 – IJC\_TEST\_24**: columns containing inclusion junction count predictor data for each of the 24 test samples, with missing values not imputed.
      9. **IJC\_TRAIN\_imputed\_1 – IJC\_TRAIN\_imputed\_24**: columns containing inclusion junction count predictor data for each of the 24 test samples, with missing values imputed based on the average of the observed training sample data.
      10. **SJC\_TRAIN\_1 – SJC\_TRAIN\_24**: columns containing skipped junction count predictor data for each of the 24 training samples.
      11. **SJC\_TEST\_1 – SJC\_TEST\_24**: columns containing skipped junction count predictor data for each of the 24 test samples, with missing values not imputed.
      12. **SJC\_TRAIN\_imputed\_1 – SJC\_TRAIN\_imputed\_24**: columns containing skipped junction count predictor data for each of the 24 test samples, with missing values imputed based on the average of the observed training sample data.
2. ***a3ss\_metadata.rds***
   1. 8894 rows: Each corresponds to a unique A3SS splice event
   2. 21 columns:
      1. **MergeID**: Column containing the unique identifier for each splice event. This variable is used to merge a3ss\_data.rds and a3ss\_metadata.rds.
      2. **GeneID – flankingEE**: unique characteristics of splice event; used to construct MergeID.
      3. **IncFormLen – GENENAME**: other metadata
      4. **PValue\_TRAIN**: p-values based on an inclusion level (IncLevel) comparison between cases and controls in the training cohort
      5. **FDR\_TRAIN**: Estimated local false discovery rate for inclusion level (IncLevel) comparison between cases and controls in the training cohort
      6. **Filter.Count\_TRAIN**: indicator for whether the data contain at least 10 counts per group for the specific event
      7. **PValue\_TEST**: p-values based on an inclusion level (IncLevel) comparison between cases and controls in the test cohort
      8. **FDR\_TEST**: Estimated local false discovery rate for inclusion level (IncLevel) comparison between cases and controls in the test cohort
      9. **Filter.Count\_TEST**: indicator for whether the data contain at least 10 counts per group for the specific event
3. ***a5ss\_data.rds***
   1. 5794 rows: Each corresponds to a unique A5SS splice event
   2. 265 columns:
      1. **Status\_TRAIN\_1 – Status\_TRAIN\_24**: columns containing the response data (i.e. case-control status) for each of the 24 samples in the training cohort.
      2. **MergeID**: Column containing the unique identifier for each splice event. This variable is used to merge a5ss\_data.rds and a5ss\_metadata.rds.
      3. **Status\_TEST\_1 – Status\_TEST\_24**: columns containing the response data (i.e. case-control status) for each of the 24 samples in the testing cohort.
      4. **IncLevel\_TRAIN\_1 – IncLevel\_TRAIN\_24**: columns containing inclusion level predictor data for each of the 24 training samples.
      5. **IncLevel\_TEST\_1 – IncLevel\_TEST\_24**: columns containing inclusion level predictor data for each of the 24 test samples, with missing values not imputed.
      6. **IncLevel\_TRAIN\_imputed\_1 – IncLevel\_TRAIN\_imputed\_24**: columns containing inclusion level predictor data for each of the 24 test samples, with missing values imputed based on the average of the observed training sample data.
      7. **IJC\_TRAIN\_1 – IJC\_TRAIN\_24**: columns containing inclusion junction count predictor data for each of the 24 training samples.
      8. **IJC\_TEST\_1 – IJC\_TEST\_24**: columns containing inclusion junction count predictor data for each of the 24 test samples, with missing values not imputed.
      9. **IJC\_TRAIN\_imputed\_1 – IJC\_TRAIN\_imputed\_24**: columns containing inclusion junction count predictor data for each of the 24 test samples, with missing values imputed based on the average of the observed training sample data.
      10. **SJC\_TRAIN\_1 – SJC\_TRAIN\_24**: columns containing skipped junction count predictor data for each of the 24 training samples.
      11. **SJC\_TEST\_1 – SJC\_TEST\_24**: columns containing skipped junction count predictor data for each of the 24 test samples, with missing values not imputed.
      12. **SJC\_TRAIN\_imputed\_1 – SJC\_TRAIN\_imputed\_24**: columns containing skipped junction count predictor data for each of the 24 test samples, with missing values imputed based on the average of the observed training sample data.
4. ***a5ss\_metadata.rds***
   1. 5794 rows: Each corresponds to a unique A5SS splice event
   2. 21 columns:
      1. **MergeID**: Column containing the unique identifier for each splice event. This variable is used to merge a5ss\_data.rds and a5ss\_metadata.rds.
      2. **GeneID – flankingEE**: unique characteristics of splice event; used to construct MergeID.
      3. **IncFormLen – GENENAME**: other metadata
      4. **PValue\_TRAIN**: p-values based on an inclusion level (IncLevel) comparison between cases and controls in the training cohort
      5. **FDR\_TRAIN**: Estimated local false discovery rate for inclusion level (IncLevel) comparison between cases and controls in the training cohort
      6. **Filter.Count\_TRAIN**: indicator for whether the data contain at least 10 counts per group for the specific event
      7. **PValue\_TEST**: p-values based on an inclusion level (IncLevel) comparison between cases and controls in the test cohort
      8. **FDR\_TEST**: Estimated local false discovery rate for inclusion level (IncLevel) comparison between cases and controls in the test cohort
      9. **Filter.Count\_TEST**: indicator for whether the data contain at least 10 counts per group for the specific event
5. ***mxe\_data.rds***
   1. 24398 rows: Each corresponds to a unique MXE splice event
   2. 265 columns:
      1. **Status\_TRAIN\_1 – Status\_TRAIN\_24**: columns containing the response data (i.e. case-control status) for each of the 24 samples in the training cohort.
      2. **MergeID**: Column containing the unique identifier for each splice event. This variable is used to merge mxe\_data.rds and mxe\_metadata.rds.
      3. **Status\_TEST\_1 – Status\_TEST\_24**: columns containing the response data (i.e. case-control status) for each of the 24 samples in the testing cohort.
      4. **IncLevel\_TRAIN\_1 – IncLevel\_TRAIN\_24**: columns containing inclusion level predictor data for each of the 24 training samples.
      5. **IncLevel\_TEST\_1 – IncLevel\_TEST\_24**: columns containing inclusion level predictor data for each of the 24 test samples, with missing values not imputed.
      6. **IncLevel\_TRAIN\_imputed\_1 – IncLevel\_TRAIN\_imputed\_24**: columns containing inclusion level predictor data for each of the 24 test samples, with missing values imputed based on the average of the observed training sample data.
      7. **IJC\_TRAIN\_1 – IJC\_TRAIN\_24**: columns containing inclusion junction count predictor data for each of the 24 training samples.
      8. **IJC\_TEST\_1 – IJC\_TEST\_24**: columns containing inclusion junction count predictor data for each of the 24 test samples, with missing values not imputed.
      9. **IJC\_TRAIN\_imputed\_1 – IJC\_TRAIN\_imputed\_24**: columns containing inclusion junction count predictor data for each of the 24 test samples, with missing values imputed based on the average of the observed training sample data.
      10. **SJC\_TRAIN\_1 – SJC\_TRAIN\_24**: columns containing skipped junction count predictor data for each of the 24 training samples.
      11. **SJC\_TEST\_1 – SJC\_TEST\_24**: columns containing skipped junction count predictor data for each of the 24 test samples, with missing values not imputed.
      12. **SJC\_TRAIN\_imputed\_1 – SJC\_TRAIN\_imputed\_24**: columns containing skipped junction count predictor data for each of the 24 test samples, with missing values imputed based on the average of the observed training sample data.
6. ***mxe\_metadata.rds***
   1. 24398 rows: Each corresponds to a unique MXE splice event
   2. 23 columns:
      1. **MergeID**: Column containing the unique identifier for each splice event. This variable is used to merge mxe\_data.rds and mxe\_metadata.rds.
      2. **GeneID – downstreamEE**: unique characteristics of splice event; used to construct MergeID.
      3. **IncFormLen – GENENAME**: other metadata
      4. **PValue\_TRAIN**: p-values based on an inclusion level (IncLevel) comparison between cases and controls in the training cohort
      5. **FDR\_TRAIN**: Estimated local false discovery rate for inclusion level (IncLevel) comparison between cases and controls in the training cohort
      6. **Filter.Count\_TRAIN**: indicator for whether the data contain at least 10 counts per group for the specific event
      7. **PValue\_TEST**: p-values based on an inclusion level (IncLevel) comparison between cases and controls in the test cohort
      8. **FDR\_TEST**: Estimated local false discovery rate for inclusion level (IncLevel) comparison between cases and controls in the test cohort
      9. **Filter.Count\_TEST**: indicator for whether the data contain at least 10 counts per group for the specific event
7. ***ri\_data.rds***
   1. 4838 rows: Each corresponds to a unique RI splice event
   2. 265 columns:
      1. **Status\_TRAIN\_1 – Status\_TRAIN\_24**: columns containing the response data (i.e. case-control status) for each of the 24 samples in the training cohort.
      2. **MergeID**: Column containing the unique identifier for each splice event. This variable is used to merge ri\_data.rds and ri\_metadata.rds.
      3. **Status\_TEST\_1 – Status\_TEST\_24**: columns containing the response data (i.e. case-control status) for each of the 24 samples in the testing cohort.
      4. **IncLevel\_TRAIN\_1 – IncLevel\_TRAIN\_24**: columns containing inclusion level predictor data for each of the 24 training samples.
      5. **IncLevel\_TEST\_1 – IncLevel\_TEST\_24**: columns containing inclusion level predictor data for each of the 24 test samples, with missing values not imputed.
      6. **IncLevel\_TRAIN\_imputed\_1 – IncLevel\_TRAIN\_imputed\_24**: columns containing inclusion level predictor data for each of the 24 test samples, with missing values imputed based on the average of the observed training sample data.
      7. **IJC\_TRAIN\_1 – IJC\_TRAIN\_24**: columns containing inclusion junction count predictor data for each of the 24 training samples.
      8. **IJC\_TEST\_1 – IJC\_TEST\_24**: columns containing inclusion junction count predictor data for each of the 24 test samples, with missing values not imputed.
      9. **IJC\_TRAIN\_imputed\_1 – IJC\_TRAIN\_imputed\_24**: columns containing inclusion junction count predictor data for each of the 24 test samples, with missing values imputed based on the average of the observed training sample data.
      10. **SJC\_TRAIN\_1 – SJC\_TRAIN\_24**: columns containing skipped junction count predictor data for each of the 24 training samples.
      11. **SJC\_TEST\_1 – SJC\_TEST\_24**: columns containing skipped junction count predictor data for each of the 24 test samples, with missing values not imputed.
      12. **SJC\_TRAIN\_imputed\_1 – SJC\_TRAIN\_imputed\_24**: columns containing skipped junction count predictor data for each of the 24 test samples, with missing values imputed based on the average of the observed training sample data.
8. ***ri\_metadata.rds***
   1. 4838 rows: Each corresponds to a unique RI splice event
   2. 21 columns:
      1. **MergeID**: Column containing the unique identifier for each splice event. This variable is used to merge ri\_data.rds and ri\_metadata.rds.
      2. **GeneID – downstreamEE**: unique characteristics of splice event; used to construct MergeID.
      3. **IncFormLen – GENENAME**: other metadata
      4. **PValue\_TRAIN**: p-values based on an inclusion level (IncLevel) comparison between cases and controls in the training cohort
      5. **FDR\_TRAIN**: Estimated local false discovery rate for inclusion level (IncLevel) comparison between cases and controls in the training cohort
      6. **Filter.Count\_TRAIN**: indicator for whether the data contain at least 10 counts per group for the specific event
      7. **PValue\_TEST**: p-values based on an inclusion level (IncLevel) comparison between cases and controls in the test cohort
      8. **FDR\_TEST**: Estimated local false discovery rate for inclusion level (IncLevel) comparison between cases and controls in the test cohort
      9. **Filter.Count\_TEST**: indicator for whether the data contain at least 10 counts per group for the specific event
9. ***se\_data.rds***
   1. 117657 rows: Each corresponds to a unique SE splice event
   2. 265 columns:
      1. **Status\_TRAIN\_1 – Status\_TRAIN\_24**: columns containing the response data (i.e. case-control status) for each of the 24 samples in the training cohort.
      2. **MergeID**: Column containing the unique identifier for each splice event. This variable is used to merge se\_data.rds and se\_metadata.rds.
      3. **Status\_TEST\_1 – Status\_TEST\_24**: columns containing the response data (i.e. case-control status) for each of the 24 samples in the testing cohort.
      4. **IncLevel\_TRAIN\_1 – IncLevel\_TRAIN\_24**: columns containing inclusion level predictor data for each of the 24 training samples.
      5. **IncLevel\_TEST\_1 – IncLevel\_TEST\_24**: columns containing inclusion level predictor data for each of the 24 test samples, with missing values not imputed.
      6. **IncLevel\_TRAIN\_imputed\_1 – IncLevel\_TRAIN\_imputed\_24**: columns containing inclusion level predictor data for each of the 24 test samples, with missing values imputed based on the average of the observed training sample data.
      7. **IJC\_TRAIN\_1 – IJC\_TRAIN\_24**: columns containing inclusion junction count predictor data for each of the 24 training samples.
      8. **IJC\_TEST\_1 – IJC\_TEST\_24**: columns containing inclusion junction count predictor data for each of the 24 test samples, with missing values not imputed.
      9. **IJC\_TRAIN\_imputed\_1 – IJC\_TRAIN\_imputed\_24**: columns containing inclusion junction count predictor data for each of the 24 test samples, with missing values imputed based on the average of the observed training sample data.
      10. **SJC\_TRAIN\_1 – SJC\_TRAIN\_24**: columns containing skipped junction count predictor data for each of the 24 training samples.
      11. **SJC\_TEST\_1 – SJC\_TEST\_24**: columns containing skipped junction count predictor data for each of the 24 test samples, with missing values not imputed.
      12. **SJC\_TRAIN\_imputed\_1 – SJC\_TRAIN\_imputed\_24**: columns containing skipped junction count predictor data for each of the 24 test samples, with missing values imputed based on the average of the observed training sample data.
10. ***se\_metadata.rds***
    1. 117657 rows: Each corresponds to a unique SE splice event
    2. 21 columns:
       1. **MergeID**: Column containing the unique identifier for each splice event. This variable is used to merge se\_data.rds and se\_metadata.rds.
       2. **GeneID – downstreamEE**: unique characteristics of splice event; used to construct MergeID.
       3. **IncFormLen – GENENAME**: other metadata
       4. **PValue\_TRAIN**: p-values based on an inclusion level (IncLevel) comparison between cases and controls in the training cohort
       5. **FDR\_TRAIN**: Estimated local false discovery rate for inclusion level (IncLevel) comparison between cases and controls in the training cohort
       6. **Filter.Count\_TRAIN**: indicator for whether the data contain at least 10 counts per group for the specific event
       7. **PValue\_TEST**: p-values based on an inclusion level (IncLevel) comparison between cases and controls in the test cohort
       8. **FDR\_TEST**: Estimated local false discovery rate for inclusion level (IncLevel) comparison between cases and controls in the test cohort
       9. **Filter.Count\_TEST**: indicator for whether the data contain at least 10 counts per group for the specific event