Supplementary material

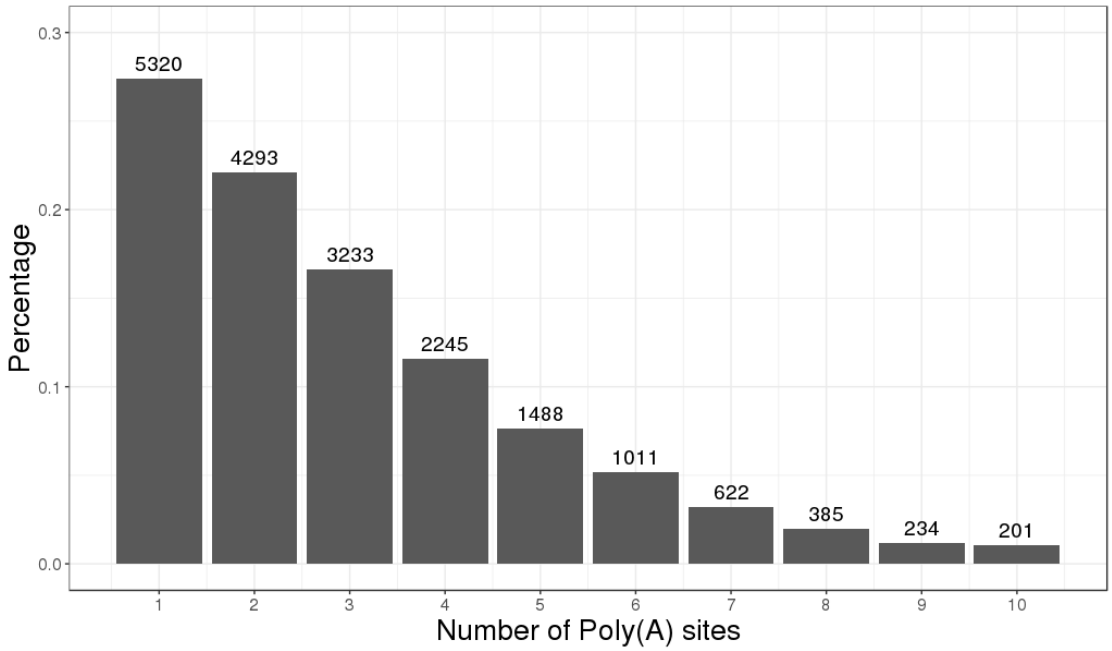

Figure S1. **Histogram of APA isoforms in Ensembl annotation GRCh38.94.** Poly(A) sites more than 10 are not shown here, and they constitute 2.1% of genes.

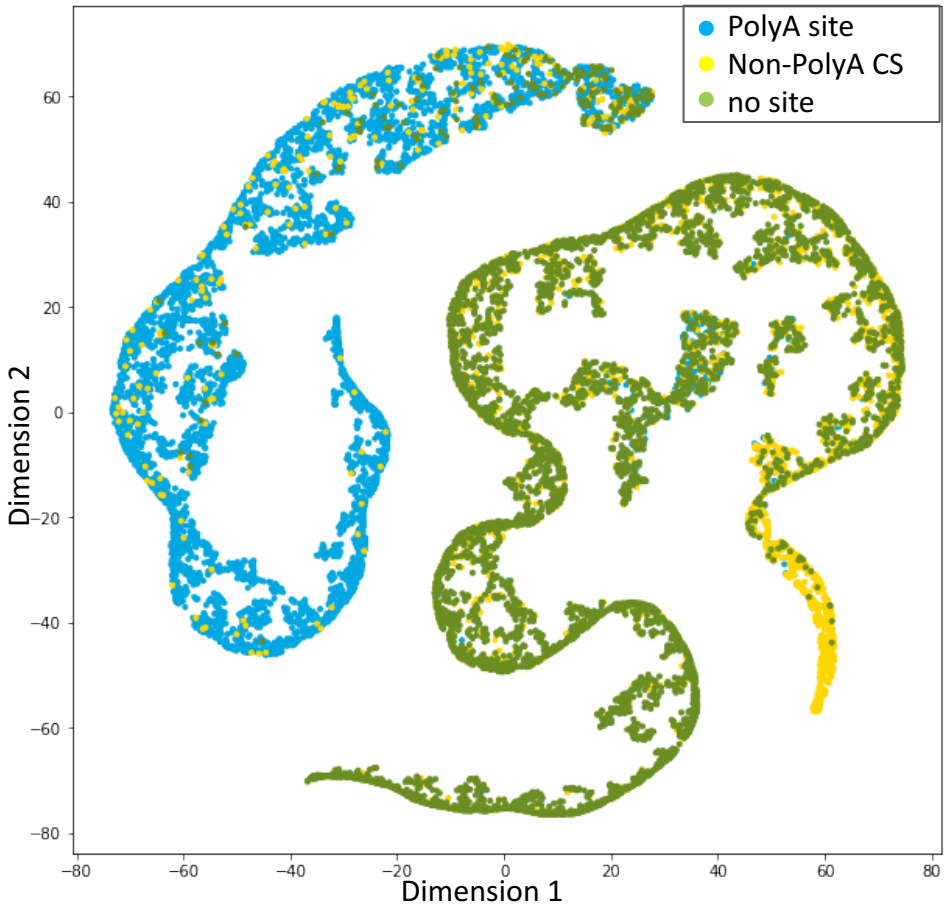

Figure S2. **T-sne plot visualizing the separation of three classes.** Each dot represents a test sequence of length 200 nt (100 nt upstream + 100 nt downstream). Sequences are projected into t-SNE space based on the weights of the last hidden layer from Termin$(A)_n$tor, with the first two components plotted as the axes of the plot. Cluster assignments of sequences are based on the their real class.

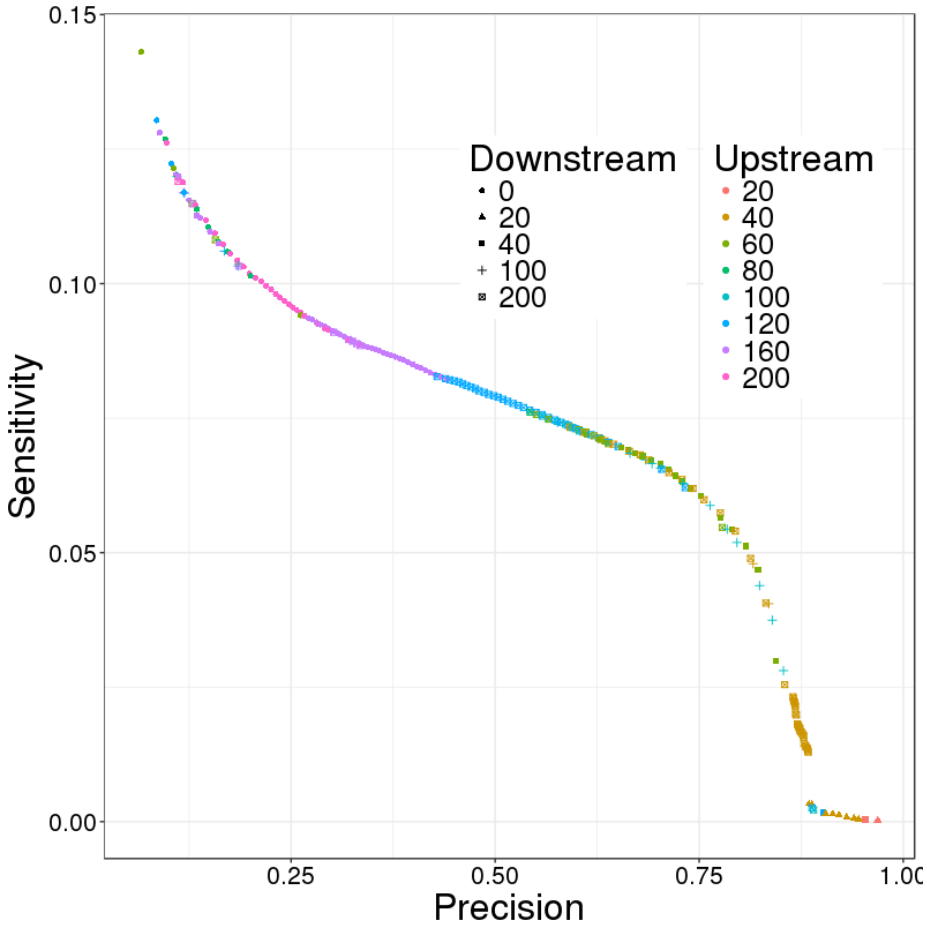

Figure S3. **Performance of 40 models with different length combinations on UHR sample.** Only the models whose performance lie on the Pareto Frontier are plotted in the figure. Each model was trained with the same set of sequences but of different up/down-stream lengths, and was applied on the same set of candidate sequences extracted from the UHR sample.

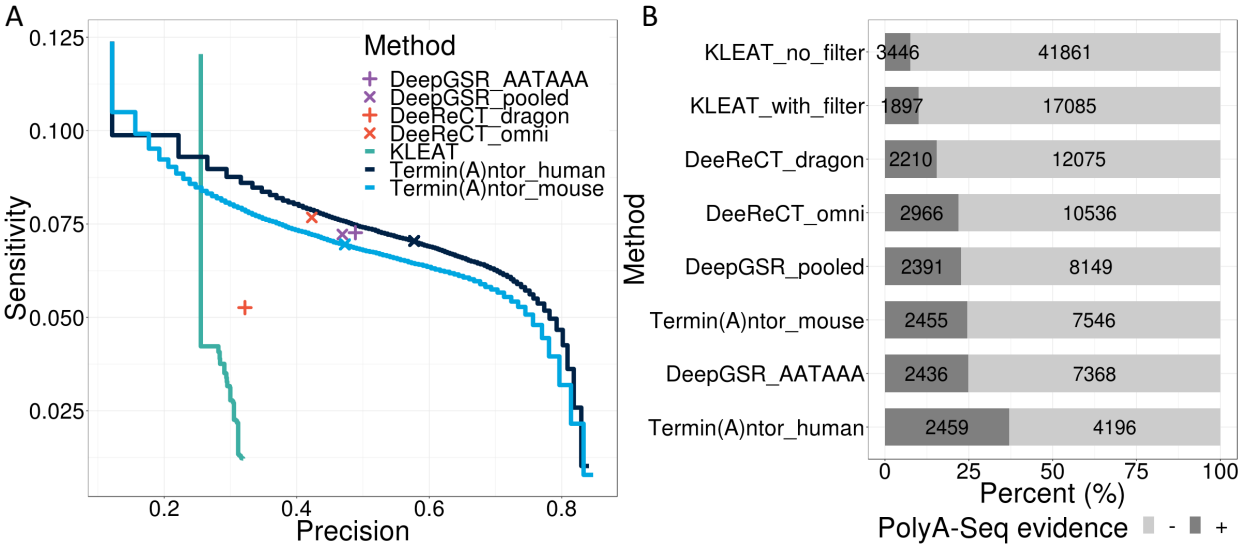

Figure S4. **Performance comparison on HBR sample.** (A) Sensitivity and specificity of poly(A) sites predicted by the 4 tools when compared to all Ensembl annotated poly(A) sites. Two pre-trained models of DeepGSR are the one with sequences containing only the hexamer AATAAA, and the one with sequences containing all hexamers pooled together. Two pre-trained models of DeeReCT-PolyA are the one with dragon data set and the one with omni data set. Two pre-trained Termin(A)$_n$tor models are the one with human data set and the one with mouse data set. The navy /blue crosses on Termin(A)$_n$tor human/mouse model represent probability = 0.5 cutoff, respectively. (B) Poly(A) sites that are missing from Ensembl annotation were compared to the ones predicted by PolyA-Seq.

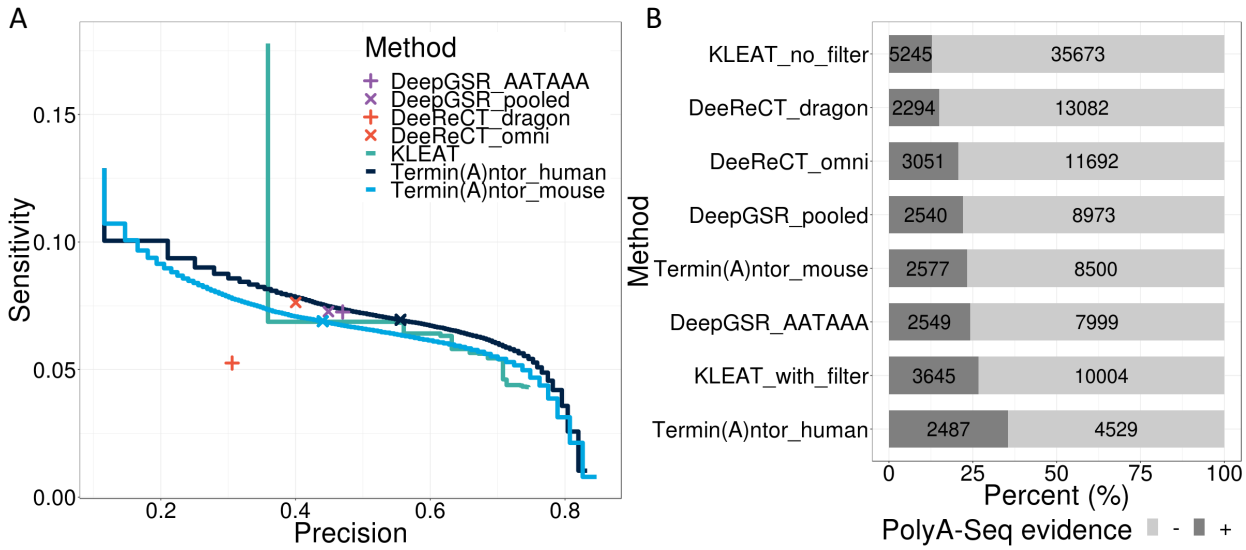

Figure S4. **Performance comparison on HBR sample.** (A) Sensitivity and specificity of poly(A) sites predicted by the 4 tools when compared to all Ensembl annotated poly(A) sites. Two pre-trained models of DeepGSR are the one with sequences containing only the hexamer AATAAA, and the one with sequences containing all hexamers pooled together. Two pre-trained models of DeeReCT-PolyA are the one with dragon data set and the one with omni data set. Two pre-trained Termin(A)$_n$tor odels are the one with human data set and the one with mouse data set. The navy /blue crosses on Termin(A)$_n$tor human/mouse model represent probability = 0.5 cutoff, respectively. (B) Poly(A) sites that are missing from Ensembl annotation were compared to the ones predicted by PolyA-Seq.

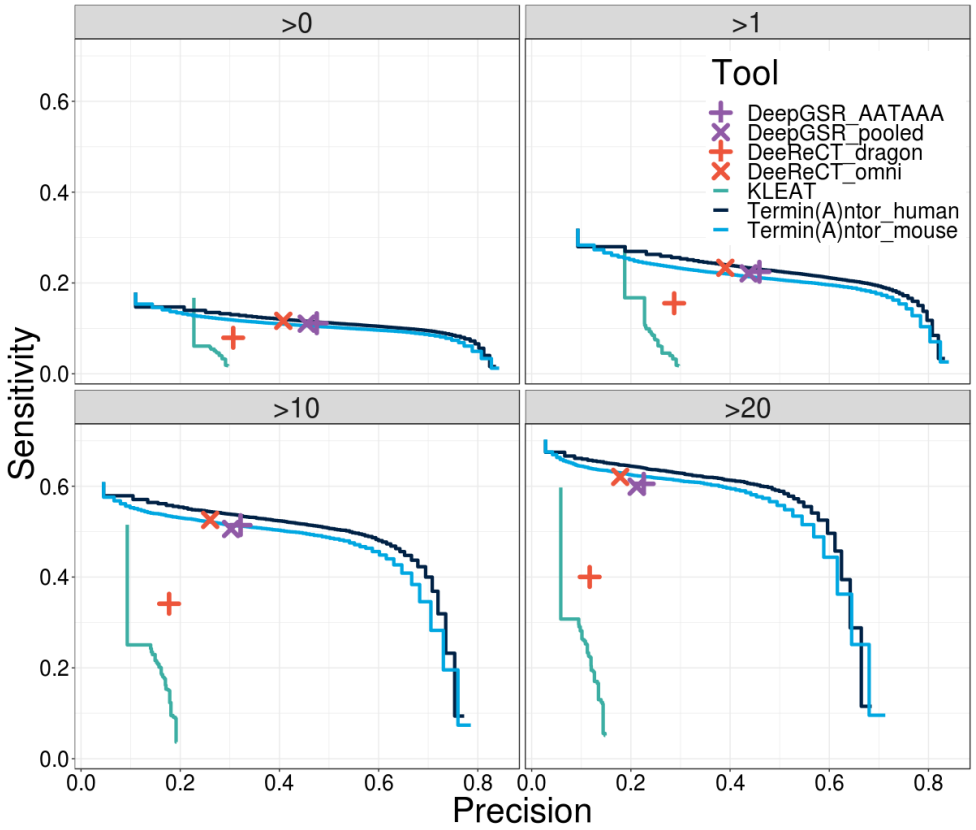

Figure S5. **Performance comparison on UHR sample with different expression level cutoffs.** Sensitivity and precision of poly(A) sites predicted by 4 tools (7 models) when compared to poly(A) sites of annotated transcripts with different expression levels. 4 facet plots represent the comparison with all expressed transcripts, expressed transcript with transcript per million (TPM) > 1, > 10, and > 20.

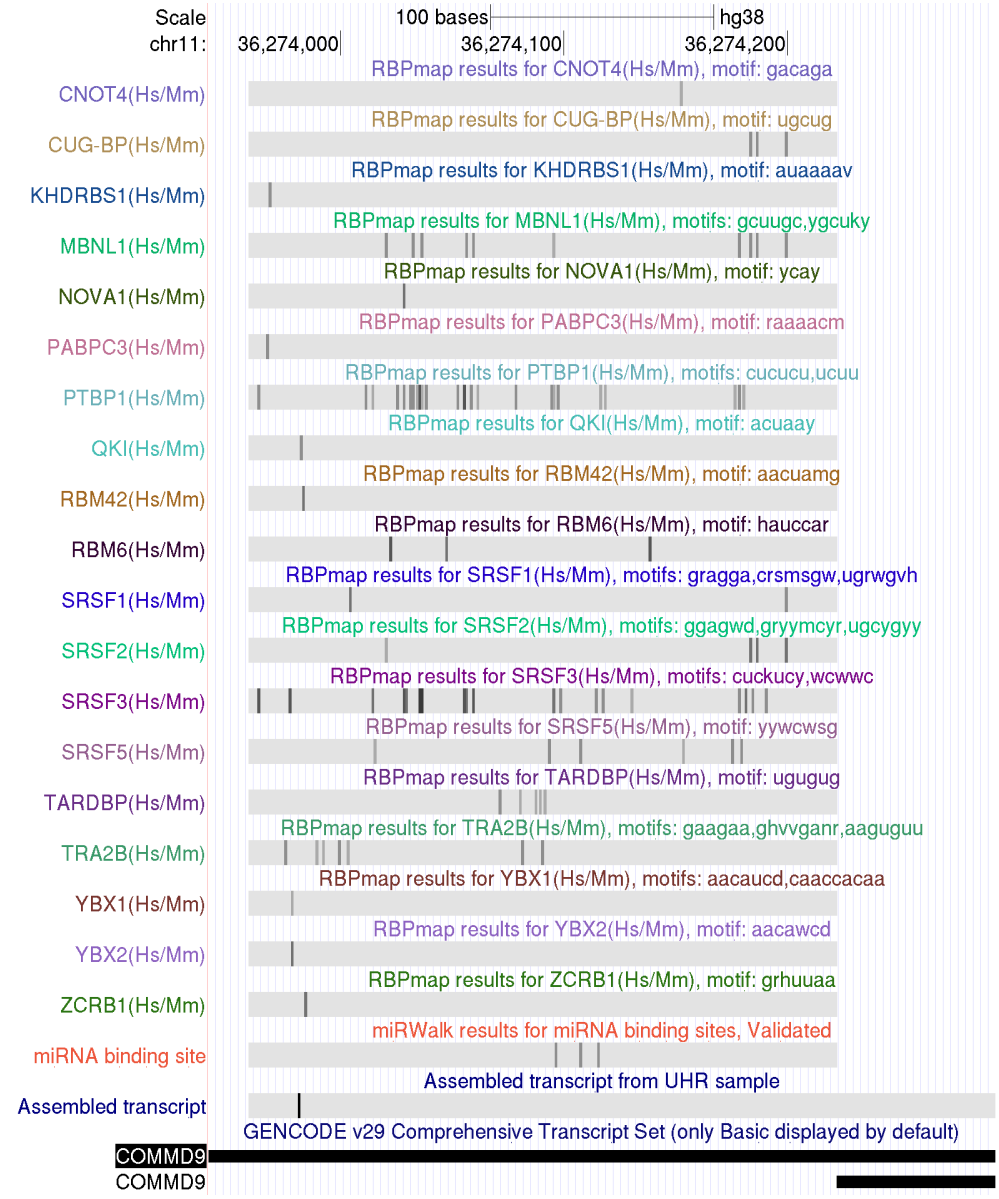

Figure S6. **MiRNA binding sites and RNA binding protein (RBP) sites on the 3' UTR region of COMMD9 with respect to the newly discovered poly(A) site.** The GENCODE track shows two annotated transcripts with different poly(A) sites and the assembled transcript track shows the assembled transcript till it's 3' end. The black tick indicates the C to A mutation. In the miRNA binding site track, 3 miRNA binding sites validated by MiRTarBase are shown from the end of the shorter poly(A) site till the end of the newly discovered one. All the rest tracks are RBP binding sites predicted by RBPmap.

Figure S7A. **RNA binding protein (RBP) sites on the extended 3' UTR region of LRRK1 with respect to the destroyed poly(A) site.** The GENCODE track and RefSeq track show two annotated transcripts with different poly(A) sites and the assembled transcript track shows the assembled transcript with the C to A mutation as the black tick. All the rest tracks are RBP binding sites predicted by RBPmap.

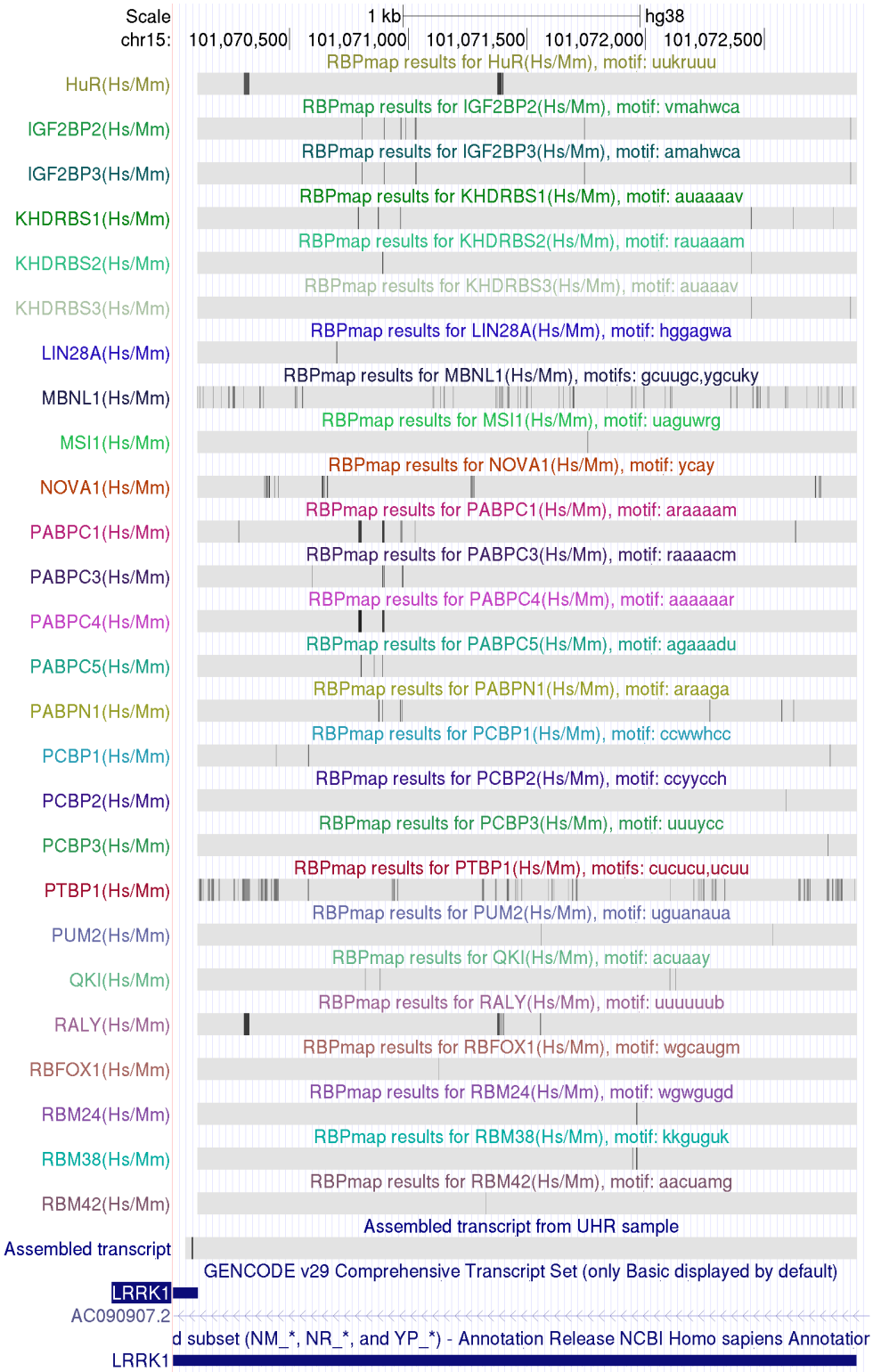

Figure S7A. **RNA binding protein (RBP) sites on the extended 3' UTR region of LRRK1 with respect to the destroyed poly(A) site.** The GENCODE track and RefSeq track show two annotated transcripts with different poly(A) sites and the assembled transcript track shows the assembled transcript with the C to A mutation as the black tick. All the rest tracks are RBP binding sites predicted by RBPmap.

Figure S7C. **MiRNA binding sites and RNA binding protein (RBP) sites on the extended 3' UTR region of LRRK1 with respect to the destroyed poly(A) site.** The GENCODE track and RefSeq track show two annotated transcripts with different poly(A) sites and the assembled transcript track shows the assembled transcript with the C to A mutation as the black tick. The miRNA binding site track shows all the miRNA binding sites predicted by TargetScan or miRDB. All the rest tracks are RBP binding sites predicted by RBPmap.

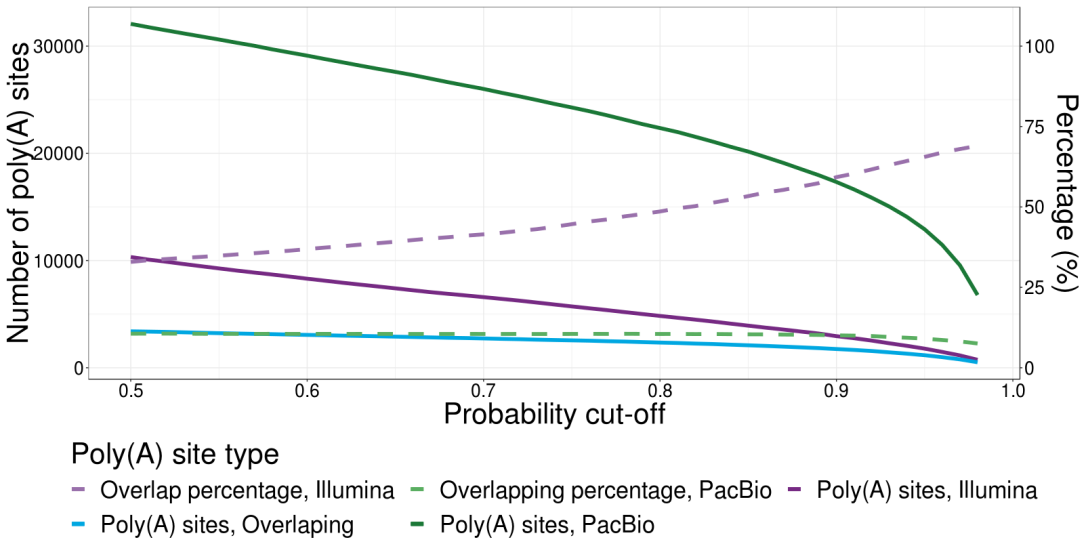

Figure S8. **Overlapping poly(A) sites identified in PacBio library and Illumina library.**