**Supplementary Materials**

**S1. Ψ site prediction by using Ψ-WHISTLE**

It is well-known that there are so far only three bioinformatics predictors relevant to Ψ, involving PPUS [1] which can predict the pseudouridine synthase-specific Ψ sites, and the other two predictors PseUI [2] and iRNA-PseU [3] which predict general Ψ sites from given DNA/RNA sequences. Recently, research has proved that sequence-derived features alone are insufficient to reach high-accuracy Ψ site prediction. The genomic features, such as conservation, gene annotation and miRNA binding, are also required and believed to significantly improve the performance of the prediction following WHISTLE approach, which was created for RNA m6A site prediction and achieved remarkable improvement in accuracy [4]. Similarity, we built a highly accurate Ψ site classifier, named Ψ-WHISTLE. Based on the combination of the two types of features, it can predict potential associations between pseudouridine and single nucleotide variants by comparing the pseudouridylation status of the original and mutated sequences.

**Training and testing data**

The data used for building models came from four different base-resolution Ψ profiling techniques, including Ψ-Seq, RBS-Seq, CeU-Seq and Pseudo-Seq (see **Table S2**). The Ψ sites were divided into two classes, one is directly downloaded from Gene Expression Omnibus (GEO) as positive sites, the other unmodified T sites located on the same transcripts of Ψ sites were randomly selected as the negative sites. Considering that positive data are small, large negative sites were split into 10 groups and respectively combined with positive sites to establish 10 separate predictors, then the performance of the classifier was accessed by averaging the results. During the Ψ site prediction, leave-one-out technique was used for evaluation [4], where 4 samples from datasets H1-H5 were used as training and the left one was used as the independent testing data. Finally, the classifier built by datasets H1-H5 was evaluated by the dataset H6, which was extracted from an independent technology (Pseudo-Seq).

**Features selection for Ψ site prediction**

***Sequence-derived features.***

According to the chemical properties of the nucleotide, we still followed the nucleotide encoding method previously used in PPUS, iRNA-PseU and PseU and generated sequence-derived information on 41 bp sequences where Ψ/non-Ψ sites reside at central. Given the three chemical properties of the nucleotides, the first property is the special structure of nucleotides, where A and G have two rings, while C and U only have one ring. The second property is functional groups. A and C contain amino group, whereas G and U contain the keto group. The last property is the number of hydrogen bonds formed. A and U can form 2 hydrogen bonds during hybridization, whereas G and C can form 3 hydrogen bonds. Based on the three structural chemical properties, the ith nucleotide on the sequence may be encoded by a vector. For example, A, C, G, U can be encoded as vectors (1,1,1), (0,1,0), (1,0,0) and (0,0,1), respectively. For the nucleotide density, it calculated the cumulative nucleotide frequency of each nucleotide position in the sequence. The density of nucleotide in ith position refers to the occurrences of the nucleotide before that position, the formula is defined as follows: . The sequence features based on the PseKNC were created by PseKNC-general [9] package.

***Genome-derived features.***

Comparing the original WHISTE approach [4], which only took 35 genomic features into account for the prediction of m6A RNA methylation sites, there were additionally 6 new genomic features used for building our prediction model. Thus, the total number of the genomic features was 42 and their detailed information was summarized at Table S3. Among all the genome-derived features, top 16 Genomic Features are dummy variable features, which indicates whether the uridine sites fallen within the transcript regions satisfy certain topological properties. Genomic Features 17-20 describe the relative position of the transcript regions, including 3’UTR, 5’UTR, CDS and whole transcript. If the values are set to zero, it suggests the site does not belong to the region. Genomic features 21-25 define the length of the transcript region where modification site resides and zero value suggests the site does not belong to this region. Genomic Features 26-29 means distances from the adenine sites to the 5’end or 3’end of the splicing junctions. Genomic Features 30-31 are the clustering statuses of uridine sites with 100 bp and 1000 bp flanking windows. Genomic Features 32-35 assess the evolutionary conservation score of the uridine sites and its flanking regions are measured by Phast-Cons [10] score and the fitness consequence scores [11]. All the features described above were generated by the GenomicFeatures R/Bioconductor package using the transcript annotations hg19 package [12].

**Table S2 Base-resolution dataset used for Ψ site prediction**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Species | Dataset ID | Cell line | Treatment | Technique | Source |
| Human | H1 | HEK293 |  | Ψ-Seq | [5] |
| H2 | Hela |  | RBS-Seq | [6] |
| H3 | HEK293T |  | CeU-Seq | [7] |
| H4 | HEK293T | H2O2 |
| H5 | HEK293T | Heat Shock (HS) |
| H6 | Hela |  | Pseudo-Seq | [8] |

**Table S3 Genome-derived features used for mammalian Ψ site prediction**

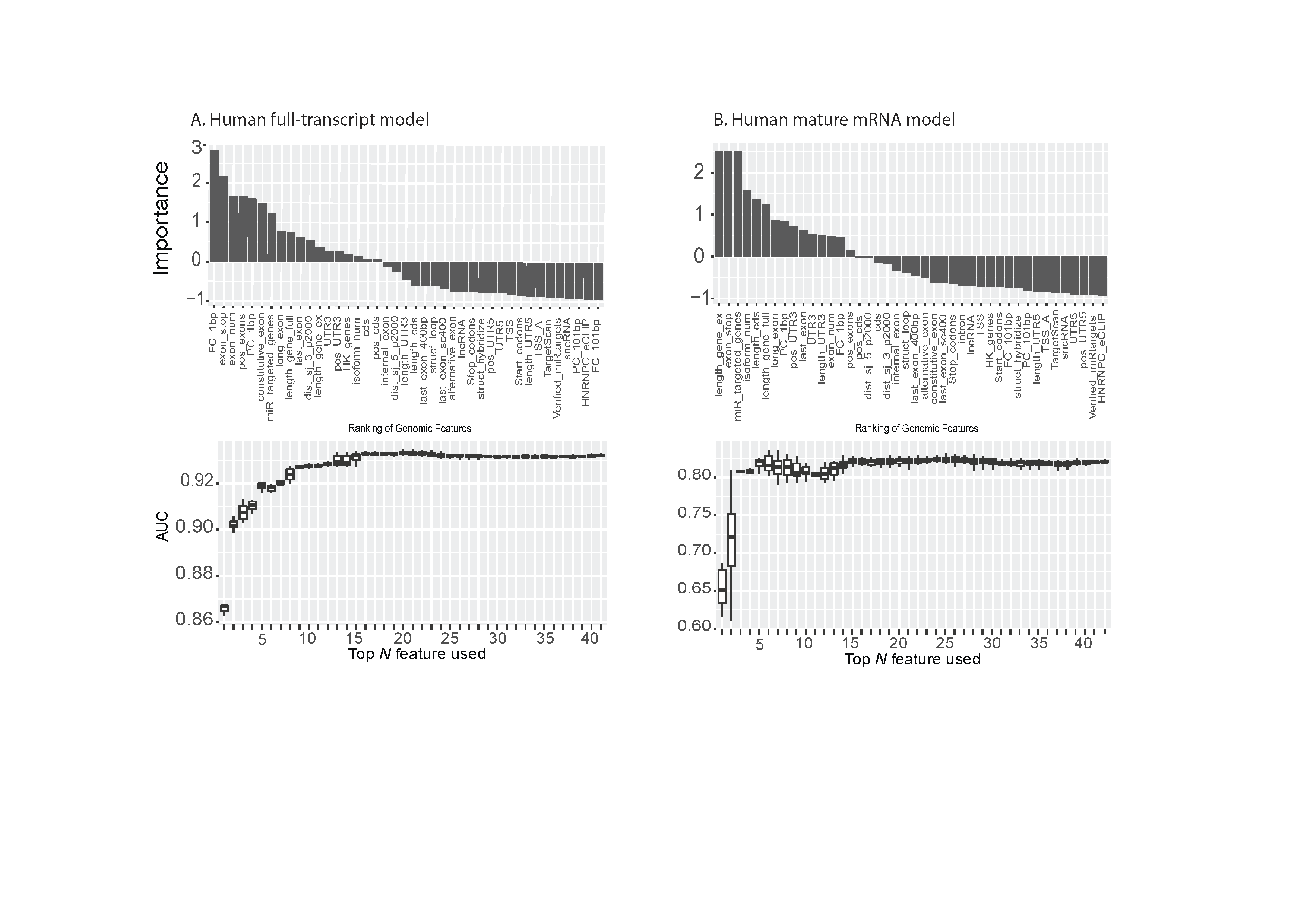
|  |  |  |  |
| --- | --- | --- | --- |
| ID | Name | Description | Note |
| 1 | UTR5 | 5' UTR | Dummy variables indicating whether the site is overlapped to the topological region on the major RNA transcript. |
| 2 | UTR3 | 3' UTR |
| 3 | cds | Coding sequence |
| 4 | Stop\_codons | stop codons flanked by 100bp |
| 5 | Start\_codons | start codons flanked by 100bp |
| 6 | TSS | downstream 100bp of TSS |
| 7 | TSS\_A | downstream 100bp of TSS on A |
| 8 | exon\_stop | exons containing stop codons |
| 9 | alternative\_exon | alternative exons |
| 10 | constitutive\_exon | constitutive exons |
| 11 | internal\_exon | Internal exons |
| 12 | long\_exon | long exons (exon length >= 400bp) |
| 13\* | last\_exon | 5’ last\_exon |
| 14 | last\_exon\_400bp | 5’ 400bp of the last exons |
| 15 | last\_exon\_sc400 | 5’ 400bp of the last exons containing stop codons |
| 16\* | intron | intron |
| 17 | pos\_cds | relative position on coding sequence | Relative position on the region |
| 18 | pos\_UTR5 | relative position on 5'UTR |
| 19 | pos\_UTR3 | relative position on 3'UTR |
| 20 | pos\_exons | relative position on exon |
| 21 | length\_UTR5 | 5'UTR length | The region length in bp. |
| 22 | length\_UTR3 | 3'UTR length |
| 23 | length\_gene\_ex | mature transcript length |
| 24 | length\_cds | Coding sequence length |
| 25 | length\_gene\_full | full transcript length |
| 26 | dist\_sj\_5\_p2000 | distance to the 5' splicing junction | Nucleotide distances toward the splicing junctions or the nearest neighboring sites. |
| 27 | dist\_sj\_3\_p2000 | distance to the 3' splicing junction |
| 28 | PC\_1bp | phastCons scores of the nucleotide | Scores related to evolutionary conservation |
| 29 | PC\_101bp | average phastCons scores within the flanking 50bp |
| 30 | FC\_1bp | fitCons scores of the nucleotide |
| 31 | FC\_101bp | average fitCons scores within the flanking 50bp region |
| 32 | struc\_hybridize | Predicted RNA hybridized region | RNA secondary structures |
| 33 | struc\_loop | Predicted RNA loop region |
| 34 | sncRNA | sncRNA | Genomic properties |
| 35 | lncRNA | lncRNA |
| 36 | HK\_genes | housekeeping genes |
| 37 | isoform\_num | Number of isoform |
| 38 | exon\_num | Number of exon |
| 39 | HNRNPC\_eCLIP | eCLIP data of HNRNPC RNA binding sites | Attributes of the genes or transcripts |
| 40 | Verified\_miRtargets | miRNA targeted sites verified by experiment |
| 41 | TargetScan | Predicted miRNA targeted sites by TargetScan |
| 42 | miR\_targeted\_genes | miRNA targeted genes |

**Machine learning and Performance evaluation**

SVM (Support Vector Machine) is one of the classical machine learning algorithms that have been widely applied in computational biology, such as microRNA target prediction [9], protein phosphorylation prediction [10] and m6A RNA methylation sites prediction [11], as most of some show outstanding performance. In this project, LIBSVM [12] was also used to establish predictors with a radial basis function as kernel. In performance evaluation, in order to avoid the overlap of training and testing datasets and allow the evaluation result directly reflect the capability of the algorithm in the identification of previously unknown Ψ sites, we only used the Ψ sites that were not used as training data previously. Here, the ROC (receiver operating characteristic) curve (sensitivity against 1-specificity) was used for performance evaluation and the area under ROC curve (AUROC) was considered as the major performance evaluation metric. Considering the fact that existing datasets overwhelmingly relied on polyA selection in RNA library preparation, and intronic Ψ sites are likely to be under-represented in the data, the performance was assessed in two modes: full transcript and mature mRNA modes. In the mature mRNA mode, only positive and negative Ψ sites located on mature mRNA transcripts are considered, as previously described [4].

**RESULTS**

Before building the classifier, feature selection was applied to extract the most significant subset of features for prediction, so as to avoid potential over-fitting. The Perturb method [13] was implemented to measure the relative importance of each genome-derived feature. According to the rank of importance, the different top *N* most important features were selected respectively to make prediction and were evaluated with 5-fold cross-validation. Remarkably, only the datasets H2-H5 were used as training data for prediction at the section of feature selection. (see **Figure S1**). After that, the performance of the classifier with the established different feature sets was evaluated. As shown at the result, the newly developed Ψ site predictor Ψ-WHISTLE substantially outperformed other approaches on independent datasets (**Table S4**) or benchmarked by an independent technique (**Table S5**).



**Figure S1**. Feature selection of the genome-derived features for Ψ site prediction. Different top rank genomic features were used in further prediction under human full transcript model (**Figure S1A**) and mature mRNA model (**Figure S1B**). The importance of each feature under different models was assessed by the Perturb method. The datasets H2-H5 were used as the training data and the performance (AUC) was evaluated using 5-fold cross-validation.

**Table S4 Human Ψ site prediction using different feature sets**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Testing  method | Predictor | Base-resolution Technique and Dataset ID | | | | | | Average |
| Ψ-Seq | RBS-Seq | CeU-Seq | | | |
| H1 | H2 | H3 | H4 | H5 | Average |
| Full  Transcript | Cross  validation | Ψ-WHISTLE | 0.966 | 0.953 | 0.965 | 0.965 | 0.971 | 0.967 | 0.964 |
| iRNA-PseU | 0.814 | 0.767 | 0.814 | 0.807 | 0.795 | 0.805 | 0.799 |
| PPUS | 0.809 | 0.756 | 0.807 | 0.802 | 0.807 | 0.805 | 0.796 |
| PesUI | 0.765 | 0.737 | 0.774 | 0.765 | 0.758 | 0.766 | 0.760 |
| Independent  Dataset | Ψ-WHISTLE | 0.957 | 0.978 | 0.977 | 0.972 | 0.794 | 0.914 | 0.936 |
| iRNA-PseU | 0.697 | 0.727 | 0.794 | 0.736 | 0.633 | 0.721 | 0.717 |
| PPUS | 0.700 | 0.721 | 0.796 | 0.740 | 0.636 | 0.724 | 0.719 |
| PesUI | 0.631 | 0.710 | 0.637 | 0.625 | 0.569 | 0.610 | 0.634 |
| mature  mRNA | Cross  validation | Ψ-WHISTLE | 0.866 | 0.870 | 0.852 | 0.866 | 0.864 | 0.861 | 0.863 |
| iRNA-PseU | 0.772 | 0.776 | 0.751 | 0.771 | 0.771 | 0.764 | 0.768 |
| PPUS | 0.776 | 0.779 | 0.755 | 0.775 | 0.775 | 0.768 | 0.772 |
| PesUI | 0.700 | 0.708 | 0.682 | 0.701 | 0.700 | 0.694 | 0.698 |
| Independent  Dataset | Ψ-WHISTLE | 0.859 | 0.770 | 0.856 | 0.867 | 0.868 | 0.864 | 0.844 |
| iRNA-PseU | 0.753 | 0.582 | 0.759 | 0.761 | 0.761 | 0.760 | 0.723 |
| PPUS | 0.749 | 0.575 | 0.756 | 0.757 | 0.758 | 0.757 | 0.719 |
| PesUI | 0.666 | 0.651 | 0.649 | 0.648 | 0.660 | 0.652 | 0.655 |

**Table S5. Human Ψ site prediction evaluated on an independent technique**

|  |  |  |
| --- | --- | --- |
|  | Full transcript model | Mature mRNA model |
| Ψ-WHISTLE | 0.972 | 0.857 |
| iRNA-PseU | 0.708 | 0.751 |
| PPUS | 0.705 | 0.748 |
| PesUI | 0.585 | 0.639 |

**Note**: The feature selection and training were performed on data generated from Ψ-Seq, RBS-Seq and CeU-Seq, while the performance was evaluated on dataset generated from a different technique (Pseudo-seq).

In addition, to evaluate the accuracy of the results of Ψ-WHISTLE more rigorously, the substrates of three mRNA dependent pseudouridine synthases (TruB1, PSU7 and TruB2) [14] were used in the independent testing (**Table S6**) and our newly developed method still shows substantial improvements compared with existing methods.

**S2. Mutations, regulations and diseases associated with Ψ**

To identify mutation events that may affect pseudouridylation, here we called the Ψ associated genetic variants as Ψ-SNP following the previous idea [15, 16]. Ψ-SNP is believed to has the potential capability to alter pseudouridylation status on an RNA transcript by two scenarios. One is that mutation directly alters T to another amino acid, leading to the loss of a Ψ site, or alters another nucleotide to T so as to gain an identified Ψ site. The other one is that mutation can significantly increase or decrease in pseudouridylation probability through altering the nucleotide within the 41 bp flanking window of a Ψ site (but not directly alter the putative Ψ site itself). To uncover the potential association between pseudouridylation and various diseases, we integrate with the GWAS catalog, the Johnson and O’Donnell data, and the ClinVar databases as annotation, to find the disease-associated genetic mutations that may cause the gain or loss of a Ψ site (or Ψ-SNP).

The association level (AL) between SNP and Ψ is calculated as follows:



where,  and  represent the likelihood of mutated and wild type sequences, respectively.

The impact of mutations on pseudouridylation status is evaluated using the newly developed Ψ-WHISTLE predictor. Here, the AL needs to be greater than 0.7 and it is recognized in as a Ψ-SNP.

**S3. Ψ-associated disease site prediction**

****

**Figure S2.** Feature selection of the all features for Ψ site prediction for Non-PU learning. Top 20 features were used in next prediction under disease (**Figure S2A**).and cancer model (**Figure S2B**). The importance of each feature under different models was assessed by the Perturb method. The performance (AUC) was evaluated using 5-fold cross-validation.



**Figure S3.** Feature selection of the all features for cancer-related Ψ site prediction in all disease-associated site for PU learning. Top 20 features were used in next prediction under cancer model. The importance of each feature under different models was assessed by the Perturb method. The performance (AUC) was evaluated using 5-fold cross-validation.

# S4. Pseudouridylation distribution and conversion on transcriptome

**REFERENCE**

[1] Y. H. Li, G. Zhang, and Q. Cui, "PPUS: a web server to predict PUS-specific pseudouridine sites," (in eng), *Bioinformatics,* vol. 31, no. 20, pp. 3362-4, Oct 15 2015.

[2] J. He, T. Fang, Z. Zhang, B. Huang, X. Zhu, and Y. Xiong, "PseUI: Pseudouridine sites identification based on RNA sequence information," (in eng), *BMC Bioinformatics,* vol. 19, no. 1, p. 306, Aug 29 2018.

[3] W. Chen, H. Tang, J. Ye, H. Lin, and K. C. Chou, "iRNA-PseU: Identifying RNA pseudouridine sites," (in eng), *Mol Ther Nucleic Acids,* vol. 5, p. e332, 2016.

[4] K. Chen *et al.*, "WHISTLE: a high-accuracy map of the human N6-methyladenosine (m6A) epitranscriptome predicted using a machine learning approach," *Nucleic Acids Research,* 2019.

[5] S. Schwartz *et al.*, "Transcriptome-wide mapping reveals widespread dynamic-regulated pseudouridylation of ncRNA and mRNA," *Cell,* vol. 159, no. 1, pp. 148-162, Sep 25 2014.

[6] V. Khoddami, A. Yerra, T. L. Mosbruger, A. M. Fleming, C. J. Burrows, and B. R. Cairns, "Transcriptome-wide profiling of multiple RNA modifications simultaneously at single-base resolution," (in eng), *Proc Natl Acad Sci U S A,* vol. 116, no. 14, pp. 6784-6789, Apr 2 2019.

[7] X. Li *et al.*, "Chemical pulldown reveals dynamic pseudouridylation of the mammalian transcriptome," (in eng), *Nat Chem Biol,* vol. 11, no. 8, pp. 592-7, Aug 2015.

[8] T. M. Carlile, M. F. Rojas-Duran, B. Zinshteyn, H. Shin, K. M. Bartoli, and W. V. Gilbert, "Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells," *Nature,* Letter vol. 515, no. 7525, pp. 143-6, Nov 06 2014.

[9] H. Liu, D. Yue, Y. Chen, S. J. Gao, and Y. Huang, "Improving performance of mammalian microRNA target prediction," (in eng), *BMC Bioinformatics,* vol. 11, p. 476, Sep 22 2010.

[10] Y.-H. Wong *et al.*, "KinasePhos 2.0: a web server for identifying protein kinase-specific phosphorylation sites based on sequences and coupling patterns," *Nucleic Acids Research,* vol. 35, no. suppl\_2, pp. W588-W594, 2007.

[11] W. Chen, H. Tang, and H. Lin, "MethyRNA: a web server for identification of N6-methyladenosine sites," *Journal of Biomolecular Structure and Dynamics,* vol. 35, no. 3, pp. 683-687, 2017.

[12] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.,* vol. 2, no. 3, pp. 1-27, 2011.

[13] M. Gevrey, I. Dimopoulos, and S. Lek, "Review and comparison of methods to study the contribution of variables in artificial neural network models," *Ecological Modelling,* vol. 160, no. 3, pp. 249-264, 2003/02/15/ 2003.

[14] M. Safra, R. Nir, D. Farouq, I. Vainberg Slutskin, and S. Schwartz, "TRUB1 is the predominant pseudouridine synthase acting on mammalian mRNA via a predictable and conserved code," (in eng), *Genome Res,* vol. 27, no. 3, pp. 393-406, Mar 2017.

[15] S. Jiang *et al.*, "m6ASNP: a tool for annotating genetic variants by m6A function," (in eng), *Gigascience,* vol. 7, no. 5, May 1 2018.

[16] Y. Zheng *et al.*, "m6AVar: a database of functional variants involved in m6A modification," (in eng), *Nucleic Acids Res,* vol. 46, no. D1, pp. D139-d145, Jan 4 2018.