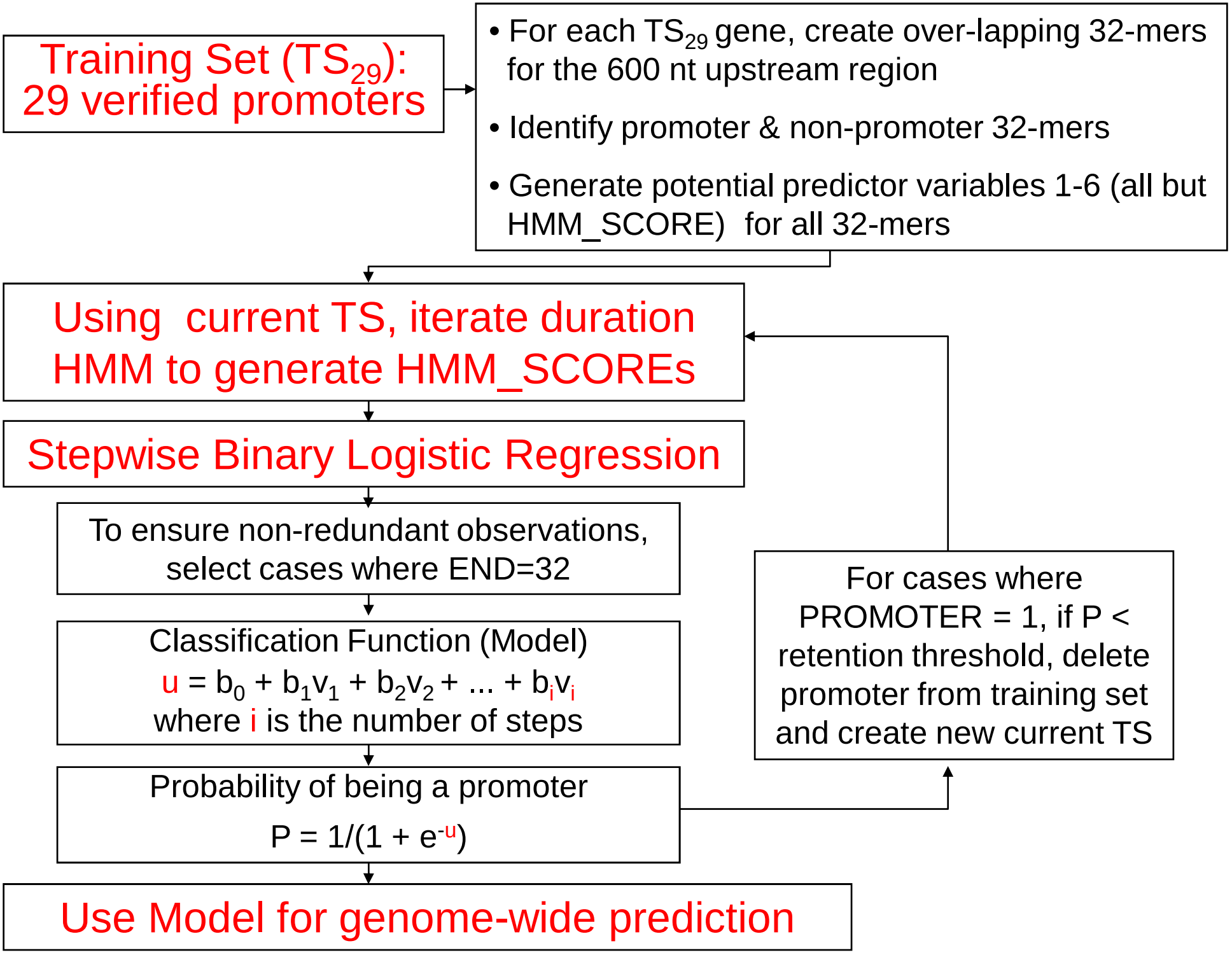
Training Set (TS_{29}): 29 verified promoters
• For each TS_{29} gene, create over-lapping 32-mers for the 600 nt upstream region
• Identify promoter & non-promoter 32-mers
• Generate potential predictor variables 1-6 (all but HMM_SCORE) for all 32-mers
Using current TS, iterate duration HMM to generate HMM_SCOREs
Stepwise Binary Logistic Regression
To ensure non-redundant observations, select cases where END=32
Classification Function (Model)
$u = b_0 + b_1v_1 + b_2v_2 + ... + b_iv_i$
where i is the number of steps
Probability of being a promoter
$P = 1/(1 + e^{-u})$
For cases where PROMOTER = 1, if P < retention threshold, delete promoter from training set and create new current TS
Use Model for genome-wide prediction