Efficient Selection of Optimal Time Points Over Biological Time-Series Data

1 Supplementary Results

1.1 *TempSelect* Pseudocode

Pseudocode of *TempSelect* is in Algorithm 1.

Algorithm 1 *TempSelect*: Iterative k-point selection

procedure ITERATIVE–TEMPORAL–SELECTION
  C_0 = select initial k time points by absolute difference sorting
  e_0 = error of remaining points by fitting splines to C_0
  $i = 0$
  do
    for each pair $(t_a, t_b) \in (T^- \setminus C_i) \times C_i$ **do**
      $C^* = C_i \cup \{t_a\} \setminus \{t_b\}$
      e^* = estimate error by fitting smoothing spline to C^* where regularization parameter is estimated by LOOCV
      if $e^* < e_i$ **then**
        $C_{i+1} = C^*$
        $e_{i+1} = e^*$
      end if
      $i = i + 1$
    end for
  while $e_{i+1} < e_i$
  Output C_i and e_i
end procedure

1.2 miRNA Clusters Are Enriched For Several Biological Processes

Detailed analysis of miRNA dataset shows clustered expression profiles of miRNAs as in Figure 1. We identified 8 stable miRNA clusters by k-means algorithm [4] where the number of clusters is selected by Bayesian Information Criteria [5]. We find clusters to change more frequently than mRNA data as

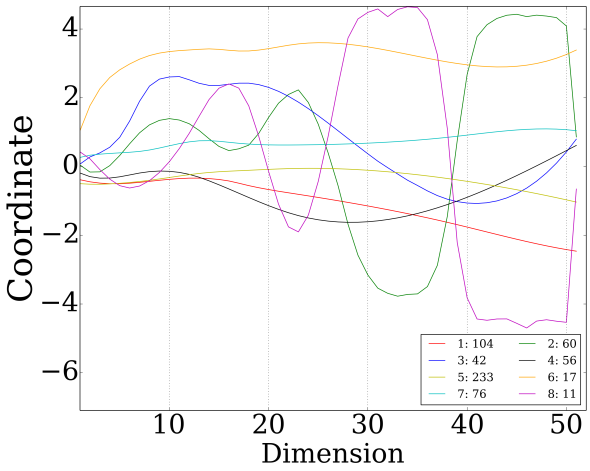


Figure 1: 8 stable clusters

miRNA is noisier than mRNA data which clusters are in Figure 2. After mapping each miRNA to the set of corresponding genes by TargetScan [1], we run gene-enrichment analysis by FuncAssociate [3]. We find clusters to be enriched for several Gene Ontology biological processes [2]. For instance, cluster 4 is enriched for single-organism cellular process, positive regulation of biological process, regulation of metabolic process, etc.

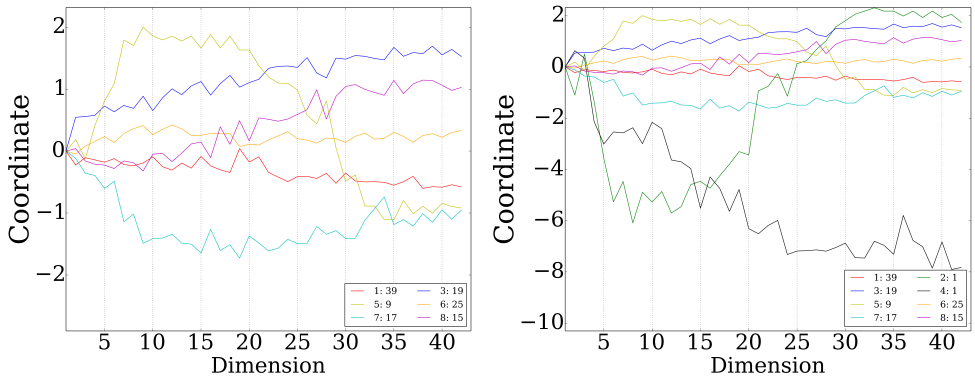
1.3 Selecting time points for Methylation analysis

2 Supplementary Figures

2.1 *TempSelect* identifies subset of important time points across multiple genes

To understand whether gene-expression profiles over time has a simple trend, we also compare the reconstruction performance of *TempSelect* with fitting piecewise linear curves between initial and middle time points and between middle and last time points. The reconstruction error by *TempSelect* is significantly better than the piecewise linear reconstruction for 102 genes out of 126 genes. We have plotted the comparison of reconstruction for several of these genes as in Supplementary Figure 5. The distribution of error difference between these methods looks significantly different than normal distribution ($p < 0.0001$ by Shapiro-Wilk test).

3 Supplementary Tables



(a) 6 stable clusters

(b) Centroids

Figure 2

Gene	Number of loci		Gene	Number of loci
Cdh11	14		Zfp536	16
Src	11		Igfbp3	34
Sox9	16		Wif1	21
Dnmt3a	41		Vegfa	20
Eln	20		Tnc	4
Foxf2	41		Lox	17
Akt1	11			

Table 1: Summary of methylation dataset

References

[1] Vikram Agarwal, George W Bell, Jin-Wu Nam, and David P Bartel. Predicting effective microrna target sites in mammalian mrnas. *eLife*, 4:e05005, 2015.

[2] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[3] Gabriel F Berriz, Oliver D King, Barbara Bryant, Chris Sander, and Frederick P Roth. Characterizing gene sets with funcassociate. *Bioinformatics*, 19(18):2502–2504, 2003.

[4] John A Hartigan. *Clustering algorithms*. John Wiley & Sons, Inc., 1975.

[5] Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

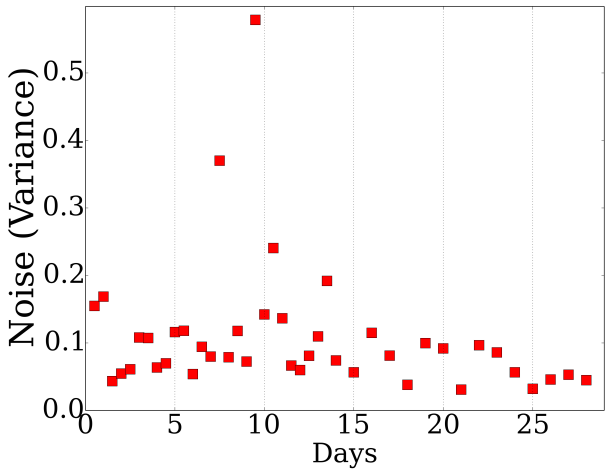


Figure 1: Average noise in each time point

Gene	r		Gene	r
Cdh11	-0.65		Lox	0.88
Src	-0.92		Igfbp3	-0.75
Sox9	-0.74		Wif1	0.54
Dnmt3a	-0.876		Vegfa	-0.81
Eln	0.72		Tnc	-0.70
Foxf2	-0.84			
Akt1	-0.91			

Table 2: Pearson correlation r between expression and methylation datasets over 8 time points for each gene.

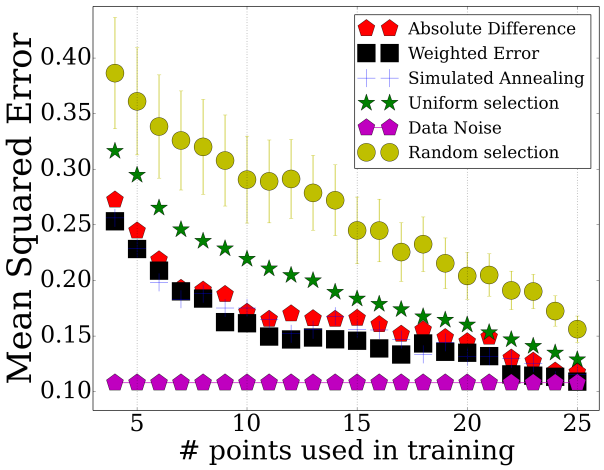
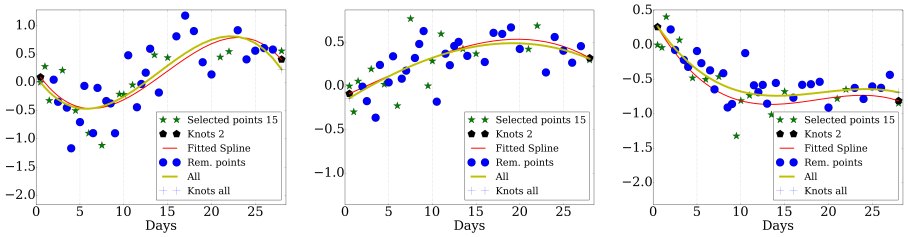
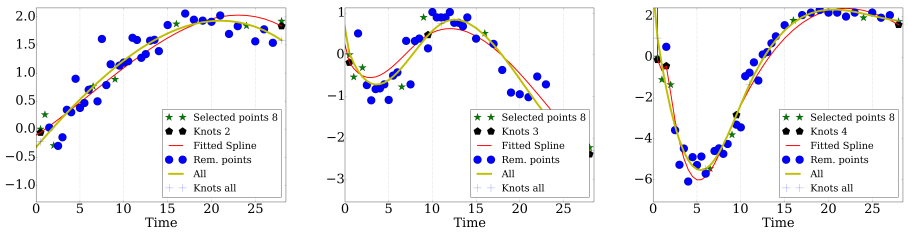


Figure 2: Performance of *TempSelect* by increasing number of selected points



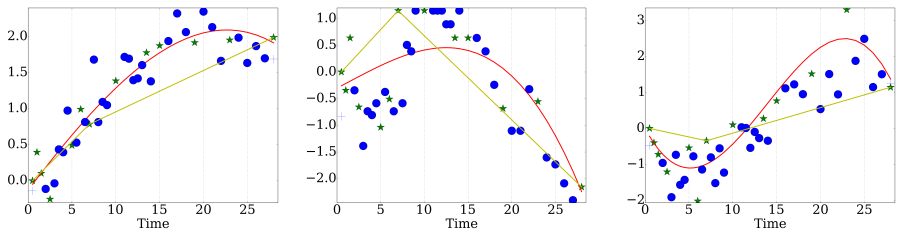
(a) ERB (b) NME3 (c) POL2RA

Figure 3: Expression profiles over several genes a) ERB, b) NME3, c) POL2RA



(a) PDGFRA (b) ELN (c) INMT

Figure 4: Reconstructed expression profiles by 8 points over genes a) PDGFRA, b) ELN, c) INMT



(a) PDGFRA (b) ELN (c) LRAT

Figure 5: Comparison of *TempSelect* and piecewise linear fitting over genes a) PDGFRA, b) ELN, c) LRAT

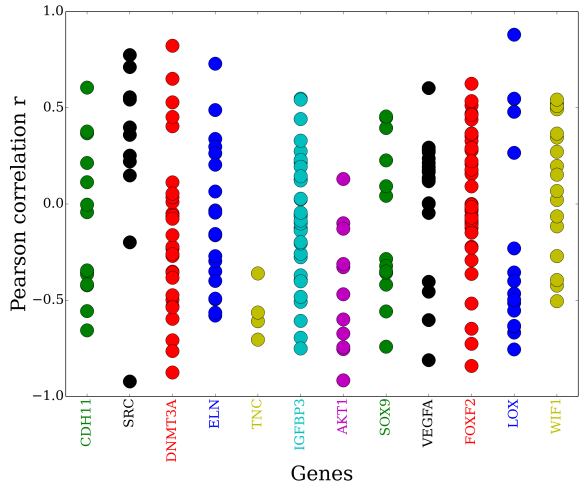


Figure 6: Distribution of gene expression correlation for loci of each gene