Determining sampling rates for time series high throughput studies

Contents

List of Figures

List of Tables

1 Supplementary Methods

1.1 *TPS* Algorithm

TPS is summarized in Algorithm 1.

1.2 Fitting Smoothing Spline

Regularized smoothing spline satisfies the piecewise cubic polynomial $\mu(t) = a_i + b_i(t-t_i) + c_i(t-t_i)^2 + d_i(t-t_i)^3$ for $t \in [t_i, t_{i+1})$, $i \in 1, \ldots, T-1$ as shown in [10]. Then, according to [8, 4], regularized smoothing spline objective can also be expressed as in:

$$\min (y-a)^{'}(y-a) + \lambda c^{'} Rc \tag{1}$$

where $a = (a_1, a_2, \ldots, a_T)$, $c = (c_2, c_3, \ldots, c_{T-1})$, and R is a $(n-2)^2$ tridiagonal symmetric matrix with entries $r_{i,i} = \frac{2(h_i+h_{i+1})}{3}$, $r_{i,i+1} = \frac{h_{i+1}}{3}$ where $h_i = t_{i+1} - t_i$. The continuity restrictions imply that:

$$Rc = Q^{'} a \tag{2}$$

where Q is an $n \times (n-2)$ tridiagonal matrix with entries $q_{i,i+1} = \frac{1}{h_{i+1}}$, $q_{i+1,i} = \frac{1}{h_{i+1}}$ and $q_{i,i} = -(\frac{1}{h_i} + \frac{1}{h_{i+1}})$. Thus, we may write Eq. 1 as:

$$\min (y-a)^{'}(y-a) + \lambda a^{'} Q R^{-1} Q^{'} a \tag{3}$$

Algorithm 1 *TPS*: 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

where a can be derived as in:

$$a = (I + \lambda Q R^{-1} Q^{'})^{-1} y \tag{4}$$

Once a is estimated, b, c, d are estimated by corresponding Equations in [8].

2 Supplementary Results

2.1 Settings for *TPS* testing

To test *TPS*, we perfromed the following analysis. First we fixed a set of points in advance (first (0.5'th day) and last (28'th day), which are required for any setting and day 7 which was previously determined to be of importance to lung development (see Supplementary Results for other settings). In addition, we have asked *TPS* to further select 10 more points (for a total of 13). For this setting, the method selected the following points: 0.5, 1.0, 1.5, 2.5, 4, 5, 7, 10, 13.5, 15, 19, 23, 28 out of 40 points. While we do not know the ground truth, the larger focus on the earlier time points determined by the method (with 7 of the 13 points for the first 7 days) makes sense in this context as several aspects of lung differentiation are determined in this early phase [5]. The other 3 weeks were more or less uniformly sampled by our method. This highlights the usefulness of an unbiased approach to sampling time points rather than just uniformly sampling through the time window.

2.2 *TPS* 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 *TPS* with fitting piecewise linear curves between initial and middle time points and between middle and last time points. The reconstruction error by *TPS* 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 Figure 7. The distribution of error difference between these methods looks significantly different than normal distribution ($p < 0.0001$ by Shapiro-Wilk test).

2.3 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 [6] where the number of clusters is selected by Bayesian Information Criteria [9]. We find clusters to change more frequently than mRNA data as 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.

2.4 miRNA Reconstruction

Figure 8 presents the reconstructed and measured expression values for a few miRNAs based on time points identified using the mRNA dataset. Several of these miRNAs are known to be involved in regulation of lung development. For example, mmu-miR-100 is known to regulate Fgfr3 and Igf1r, mmu-miR-136 targets Tgfb2, mmu-miR-152 targets Meox2, Robo1, Fbn1, Nfya [7].

2.5 Analysis of Methylation Data

Methylation data has 3 repeats for time points 0.5, 1.5, 2.5, 5, 10, 15, 19, 26 for 266 loci belonging to 13 genes. Among these genes all of them except Zfp536 also exist in mRNA dataset. Table 1 summarizes the number of loci for each gene in methylation dataset. We used shifted percentage of methylation at each time point in our analysis which is obtained by subtracting the median percentage of methylation at initial time point (baseline) from all data points for each gene.

Supplementary Figures

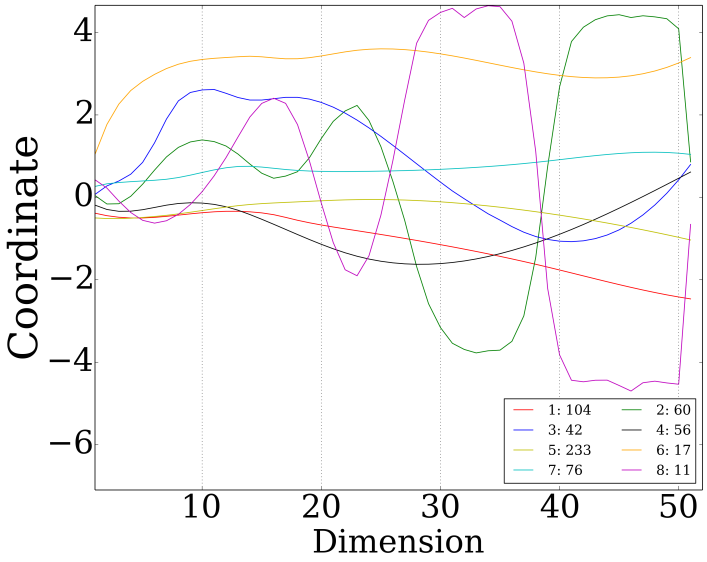
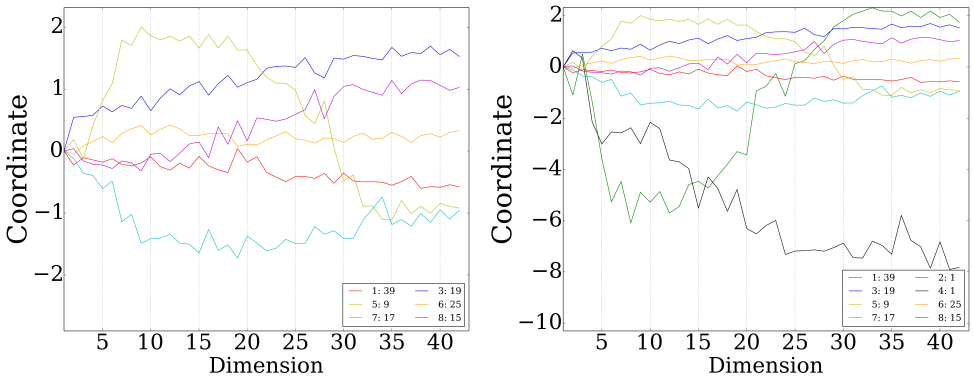


Figure 1: 8 stable clusters



(a) 6 stable clusters

(b) Centroids

Figure 2: Cluster analysis of mRNA data

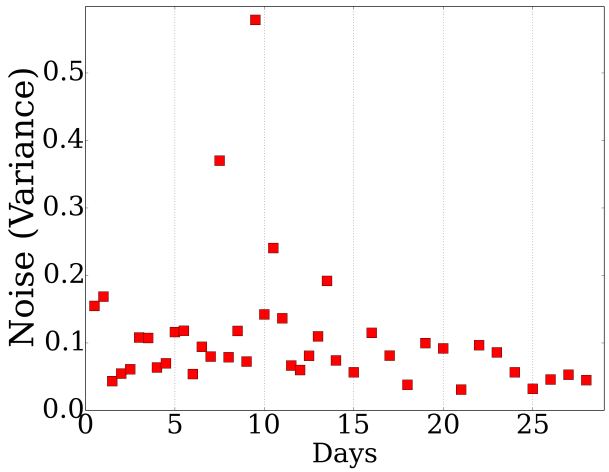


Figure 3: Average noise in each time point

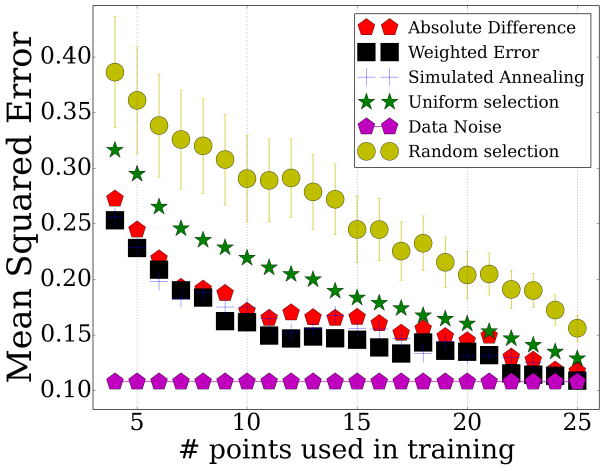
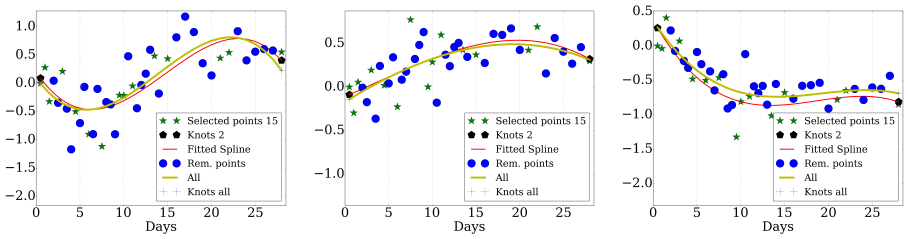
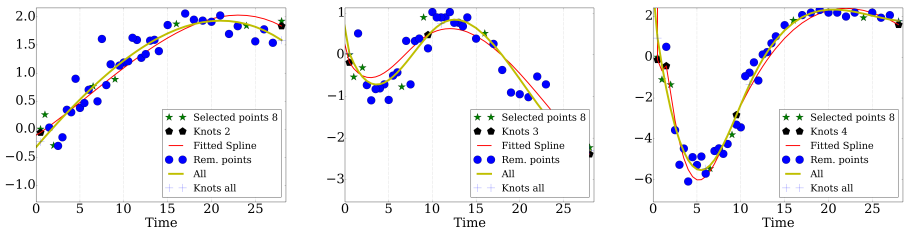


Figure 4: Performance of *TPS* by increasing number of selected points



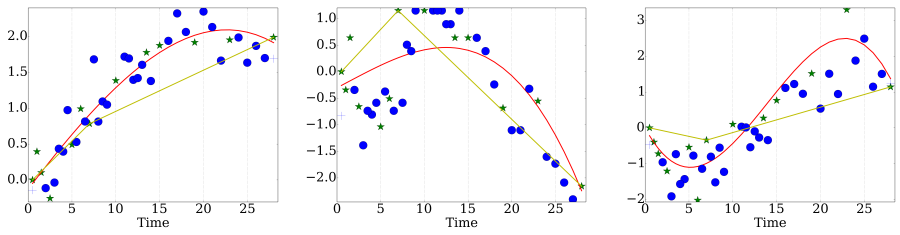
(a) ERB (b) NME3 (c) POL2RA

Figure 5: Expression profiles over several genes a) ERB, b) NME3, c) POL2RA



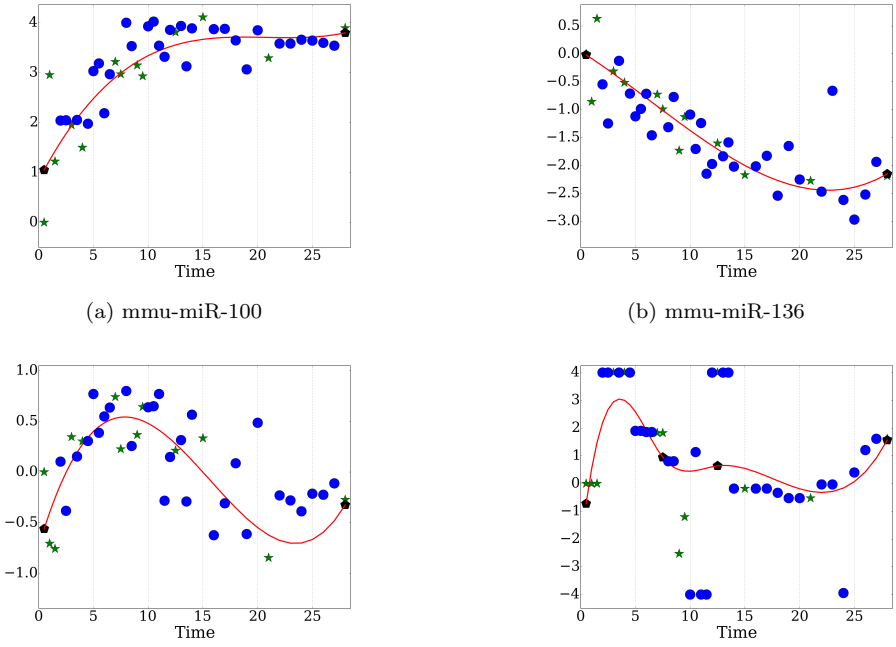
(a) PDGFRA (b) ELN (c) INMT

Figure 6: Reconstructed expression profiles by 8 points over genes a) PDGFRA, b) ELN, c) INMT



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

Figure 7: Comparison of *TPS* and piecewise linear fitting over genes a) PDGFRA, b) ELN, c) LRAT



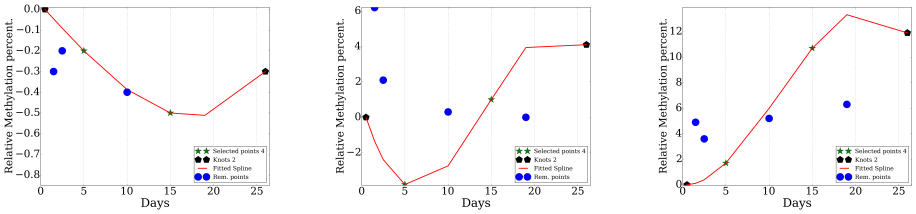
(a) mmu-miR-100

(b) mmu-miR-136

(c) mmu-miR-152

(d) mmu-miR-219

Figure 8: Predicted expression profiles of miRNAs a) mmu-miR-100, b) mmu-miR-136, c) mmu-miR-152, d) mmu-miR-219.



(a) Chromo 2, 157423995 (SRC)

(b) Chrom. 5, 134721315 (ELN)

(c) Chrom. 12, 112657170 (AKT1)

Figure 9: Reconstructed methylation profiles over several loci (chromosome, position) with corresponding genes.

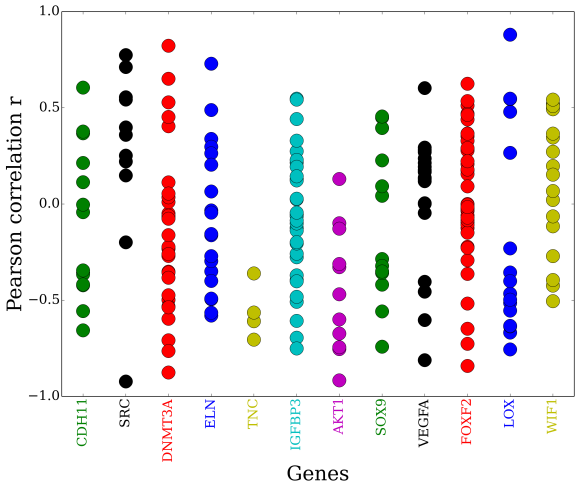


Figure 10: Distribution of gene expression correlation for loci of each gene

Supplementary Tables

Gene	Number of loci		Gene	Number of loci
Cdh11	14		Zfp536	16
Src	11		Igfbp3	34
Sox9	16		Wif1	21
Dnmt3a	41	if1	21	
Dnmt3a	41		Vegfa	20
Eln	20		Tnc	4
Foxf2	41		Lox	17
Akt1	11			

Table 1: Summary of methylation dataset

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.

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] Carl De Boor. A practical guide to splines. *Mathematics of Computation*, 1978.

[5] Martin Guilliams, Ismé De Kleer, Sandrine Henri, Sijranke Post, Leen Vanhoutte, Sofie De Prijck, Kim Deswarte, Bernard Malissen, Hamida Hammad, and Bart N Lambrecht. Alveolar macrophages develop from fetal monocytes that differentiate into long-lived cells in the first week of life via gm-csf. *The Journal of experimental medicine*, 210(10):1977–1992, 2013.

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

[7] Antonia P Popova, J Kelley Bentley, Tracy X Cui, Michelle N Richardson, Marisa J Linn, Jing Lei, Qiang Chen, Adam M Goldsmith, Gloria S Pryhuber, and Marc B Hershenson. Reduced platelet-derived growth factor receptor expression is a primary feature of human bronchopulmonary dysplasia. *American Journal of Physiology-Lung Cellular and Molecular Physiology*, 307(3):L231–L239, 2014.

[8] Christian H Reinsch. Smoothing by spline functions. *Numerische mathematik*, 10(3):177–183, 1967.

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

[10] Grace Wahba. *Spline models for observational data*, volume 59. Siam, 1990.