RESEARCH

Circadian Gene Selection for Time-to-event Phenotype by Integrating CNV and RNAseq Data

Arnab Kumar Maity[1†], Sang Chan Lee[2], Linhan Hu[2], Deborah Bell-Pedersen[3], Bani K. Mallick[2] and Tapasree Roy Sarkar[3,2*]

[*]Correspondence:
tsarkar@bio.tamu.edu
[3]Department of Biology, Texas A&M University, 3258 TAMU, 77843 College Station, USA
Full list of author information is available at the end of the article
[†]Equal contributor

Abstract

Background: The endogenous circadian clock, which controls daily rhythms in the expression of at least half of the mammalian genome, has a major influence on cell physiology. Consequently, disruption of the circadian system is associated with wide range of diseases including cancer. While several circadian clock genes have been associated with cancer progression, little is known about the survival when two or more platforms are considered together. Our goal was to determine if survival outcomes are associated with circadian clock function. To accomplish this goal, we developed a Bayesian hierarchical survival model coupled with the global local shrinkage prior and applied this model to available RNASeq and Copy Number Variation data to select significant circadian genes associates with cancer progression.

Results: Using a Bayesian shrinkage approach with the Bayesian accelerated failure time (AFT) model we showed the circadian clock associated gene DEC1 is positively correlated to survival outcome in breast cancer patients. The R package circgene implementing the methodology is available at https://github.com/MAITYA02/circgene.

Conclusions: The proposed Bayesian hierarchical model is the first shrinkage prior based model in its kind which integrates two omics platforms to identify the significant circadian gene for cancer survival.

Keywords: Bayesian survival regression; Bayesian hierarchical modeling; breast cancer; circadian genes; data integration; gene selection; global local shrinkage prior; TCGA

1 Background

The molecular circadian clock, regulates daily rhythms in the expression of at least half of all protein-coding genes [1]. Thus, it is not surprising that disruption of circadian rhythmicity is associated with significant disease, including metabolic disorder and cancer [2]. Increased cancer incidence and progression are linked to disruption of the molecular mechanism of the circadian clock [3]. At the core of the mammalian circadian clock system is a cellular circadian oscillator, which functions in most tissues at the single cell level [4], and is comprised of clock genes that form a core feedback loop. In this core loop, a heterodimer of the transcription factors CLOCK and BMAL1 activate the expression of Per and Cry genes, whose protein products negatively feedback on their own expression by inhibiting CLOCK/BMAL1 activity. Several additional loops contribute to the robustness of this core clock loop, including inhibition of CLOCK/BMAL activity by the basic helix–loop–helix

(bHLH) transcription factors DEC1 and DEC2 [5–7], which are themselves rhythmically activated by CLOCK/BMAL [8]. Both DEC1 and DEC2 have been associated with tumor progression in human cancers, an increased or decreased expression of DEC1 and DEC2 has been shown to regulate tumor progression [9]. However, the mechanisms for this regulation are not fully understood.

A recent study compared clock gene expression from human tumor and non-tumor samples from a range of cancer types that are publicly available in the Cancer Genome Atlas (TCGA) and NCBI Gene Expression Omnibus (GEO). By comparing human tumor and non-tumor samples from a range of cancer types, this study showed that clock gene co-expression is consistently deregulated in tumors [10]. This study supports the use of publicly available human datasets in understanding the role of the circadian clock in cancer development, progression, prognosis.

In the current era of precision medicine each subject is targeted for treatment modeled on individual healthcare data. Accurate prognostic prediction using molecular profiles is critical to develop precision medicine. However, cancer studies focus on one-dimensional omics data have provided limited information regarding the etiology of oncogenesis and tumor progression [11] To overcome this problem, recent work has focused on integrating multi-platform data in cancer research; as for example see [12] and references therein. Currently, TCGA is the largest collection of genomic data, which also includes parallel transcriptomics, and proteomics and patient demographic information. One primary aim of TCGA is to have more accurate stratification and prognosis of the disease by analyzing and interpreting molecular profiles for hundreds of clinical tumors representing various tumor types and their subtypes[13], at the DNA, RNA, protein and epigenetic level [8] To improve therapeutic response which may be evident from the phenotypical measures such as survival of the cancer patients, genomic alterations across these platforms has been identified. The presence of hundreds of genetic alterations inside of a genome provides a complementary view of the underlying complex biological process and thus an integrative analysis of multiple platform is required to achieve the overreaching goal of cancer studies.

To determine that the circadian gene which plays an important role in breast cancer progression and patient survival, we developed a Bayesian shrinkage approach, coupled with a Bayesian accelerated failure time (AFT) model, for integrative analysis of multiple platform of omics data. We use DNA copy number variation and RNAseq data-sources to predict patient survival. Using this approach, the clock gene and tumor suppressor DEC1 emerged as a significant gene associated with survival outcomes. While the concept of integration is very broad, and several Bayesian models exist [14–17], we believe this is the first model to include a shrinkage approach under the Bayesian regime to predict patient survival considering the circadian gene effects on the tumor progression.

Limited works have been reported on shrinkage prior in the survival settings [18, 19]. Both work specified the horseshoe shrinkage prior [20] on the regression coefficients in order to select the relevant biomarkers in the data. Additionally, they worked with the parametric models, [18] assumed a Weibull distribution for the survival model and [19] assumed a log normal distribution for the same, these works did not deal with integration among multi-platform omics data. In this article,

we propose a Bayesian log normal regression model for the survival outcome and exploit the local shrinkage parameter specification to achieve the desired variable selection.

In Section 2, we provide the detail description of the model specification, the global local Horseshoe prior specification on the regression model parameters and how this prior specification helps recovering the significant genes which are reflected via both RNASeq and CNV platforms. Additionally, the Markov Chain Monte Carlo (MCMC) scheme to generated posterior samples is developed. Section 3 describes the entire TCGA data analysis using our proposed method. Section 4 presents some simulation scenarios validating the model development. In Section 5 we provide our concluding remarks.

2 Methods

2.1 AFT Regression

We make use of the Accelerated Failure Time (AFT) model which regresses the survival time on the predictors. The AFT model is given by,

$$\log t_i = \sum_{j=1}^{p} x_{1ij}\beta_{1j} + \sum_{j=1}^{p} x_{2ij}\beta_{2j} + \epsilon_i, i = 1, \ldots, n, \quad j = 1, \ldots, p, \tag{1}$$

where i denotes the patient, j denotes copy number change or change in gene expression. Likewise, t_i is the survival time of i-th subject, x_{1ij} is the corresponding p-th copy number change in the data, and x_{2ij} is the corresponding p-th mRNA expression measured by the RNAseq technology. $\boldsymbol{\beta}_1 = (\beta_{11}, \ldots, \beta_{1p})$ is the vector of regression coefficients corresponding to the copy number changes, similarly, $\boldsymbol{\beta}_2 = (\beta_{21}, \ldots, \beta_{2p})$ is the vector of regression coefficients corresponding to the RNAseq; and $\boldsymbol{\epsilon}$ is the error vector. Assumption of $\epsilon \sim N(0, \sigma^2 I)$ gives raise to the log normal AFT model.

Letting c_i be the censoring time, the observed time may be denoted by $t_i^* = \min(t_i, c_i)$; the corresponding observed censored indicator is $\delta_i = I\{t_i \leq c_i\}$, $I\{.\}$ being the censoring indicator. Since the response is right censored, we follow the data augmentation approach of [21] to impute the censored data w_{ik} (see also [22]), $w_i = \log t_i^*$, if t_i is event time; and $w_i > \log t_i^*$, if t_i is right censored.

2.2 Shrinkage Prior

We adopted the Bayesian shrinkage approach using the horseshoe prior [20] on the regression coefficients. In the shrinkage framework, a scale-mixture representation of the global local priors allows parameters to be updated in blocks via an automatic Gibbs sampler [23] which makes it convenient for large scale problems.

Horseshoe prior in its original setting offers to recover the significant variables by specifying the same number of local shrinkage parameter as the number of regression parameters. In essence, there are shrinkage parameters for each of the regression coefficients such that the amount of shrinkage of each the regression coefficients is controlled by the corresponding local shrinkage parameter. In our setting, when there are two platforms – CNV and RNASeq expression data available for each circadian gene, a convenient way is to specify a local shrinkage parameter for two

regression parameters, one is for the CNV platform and the other is for the RNASeq platform.

In addition, when we assume the log normal distribution for the underlying time-to-event distribution the posterior samples generation can be carried out using convenient Gibbs sampling [24] or the variant. Often the presence of censored observation makes the posterior distribution more complex, however, in this setting, a remedy is to impute the right censored observation using the data augmentation scheme of [21], a successful application of which in time-to-event data has been shown in [25]. In what follows, the hierarchical Horseshoe representation for the log normal accelerated failure time (AFT) model is:

$$\begin{aligned}
\log t_i|\beta_{1j},\beta_{2j},\sigma^2 &\sim N\left(\alpha+\textstyle\sum_{j=1}^{p} x_{1ij}\beta_{1j}+\sum_{j=1}^{p} x_{2ij}\beta_{2j},\sigma^2\right)\\
\beta_{kj}|\lambda_j,\tau,\sigma^2 &\sim N(0,\lambda_j^2\tau^2\sigma^2),\quad k=1,2\\
\lambda_j &\sim C^+(0,1),\tau\sim C^+(0,1),\\
\alpha &\sim N(0,\sigma_\alpha^2\sigma^2),\sigma^2\sim \pi(\sigma^2)=1/\sigma^2
\end{aligned} \tag{2}$$

where, $C^+(0,1)$ is the truncated Cauchy density given by $f(x)=1/\{\pi(1+x^2)\}, x>0$.

2.3 Conditional Distributions and Posterior Computation

In our AFT model for group correlation structure, most of the conditional distributions are available explicitly, hence we can employ Gibbs sampling [24] technique to explore the posterior distribution. In particular, the complete conditional distributions of $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, and σ^2 are given by:

$$\begin{aligned}
\boldsymbol{\beta}_k|\boldsymbol{w},\alpha,\lambda,\tau,\sigma^2 &\sim N(B^{-1}X^T(\boldsymbol{w}-\alpha\mathbf{1}),\sigma^2A^{-1}), B=(\boldsymbol{X}^T\boldsymbol{X}+D^{-1}),\boldsymbol{X}=[\boldsymbol{X}_1,\boldsymbol{X}_2]\\
\sigma^2|\boldsymbol{w},\alpha,\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\lambda,\tau &\sim \text{Inverse Gamma}\Bigg(\text{shape}=\frac{n+p+1}{2},\\
&\qquad \text{scale}=\frac{1}{2}(\boldsymbol{w}-\alpha-\boldsymbol{X}_1\boldsymbol{\beta}_1-\boldsymbol{X}_2\boldsymbol{\beta}_2)^T(\boldsymbol{w}-\alpha-\boldsymbol{X}_1\boldsymbol{\beta}_1-\boldsymbol{X}_2\boldsymbol{\beta}_2)+\\
&\qquad \boldsymbol{\beta}_1^TD^{-1}\boldsymbol{\beta}_1+\boldsymbol{\beta}_2^TD^{-1}\boldsymbol{\beta}_2)\Bigg)\\
\alpha|\boldsymbol{w},\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\sigma^2 &\sim N(A^{-1}\mathbf{1}_n^T(\boldsymbol{w}-\boldsymbol{X}_1\boldsymbol{\beta}_1-\boldsymbol{X}_2\boldsymbol{\beta}_2),\sigma^2A^{-1}),\quad A=(\mathbf{1}^T\mathbf{1}+\sigma_\alpha^2)),
\end{aligned}$$

where, $D=\tau^2\text{diag}\left(\lambda_1^2,\ldots,\lambda_p^2,\lambda_1^2,\ldots,\lambda_p^2\right)$.

Due to the nature of the prior on λ and τ, a straightforward Gibbs sampling approach may not be possible. An alternative approach, which is based on the idea of slice sampling [26], has been discussed in the online supplement of [27]. It follows that,

$$\pi(\lambda_j|\beta_{1j},\beta_{2j},\tau,\sigma^2)\propto\frac{1}{\lambda_j}\exp\left(-\frac{1}{2}\frac{\beta_{1j}^2+\beta_{2j}^2}{\lambda_j^2\tau^2\sigma^2}\right)\frac{1}{1+\lambda_j^2}I(\lambda_j>0).$$

Defining $\phi_j = 1/\lambda_j^2$ and introducing a latent parameter u_j, the conditional posterior distribution looks like,

$$\pi(u_j, \phi_j | \beta_{1j}, \beta_{2j}, \tau, \sigma^2) \propto \exp\left(-\frac{1}{2}\frac{\phi_j(\beta_{1j}^2 + \beta_{2j}^2)}{\tau^2\sigma^2}\right) I(0 < u_j < \frac{1}{1+\phi_j}) I(\phi_j > 0).$$

Then the following scheme is used to sample the posterior distribution of λ:

1 Sample $u_j|\phi_j \sim U\{0, 1/(1+\phi_j)\}$.
2 Sample $\phi_j|u_j, \beta_{1j}, \beta_{2j}, \tau, \sigma^2 \sim$ truncated Exponential$\{(\beta_{1j}^2 + \beta_{2j}^2)/(2\lambda_j^2\tau^2\sigma^2)\} I\{0, 1/(u_j - 1)\}$.
3 Compute in $\lambda_j = 1/\sqrt{\phi_j}$.

Updating τ can be carried out in the similar fashion. We introduce a latent variable v and let $\xi = 1/\tau^2$ to yield desired posterior samples:

1 Sample $v|\xi \sim U(0, \{1/(1+\xi)\}$.
2 Sample $\xi|v, \boldsymbol{\beta}, \boldsymbol{\lambda}, \sigma^2 \sim$ truncated Gamma$\{(p+1)/2, (1/2\sigma^2)(\sum_{j=1}^p \beta_{1j}^2/\lambda_j^2 + \sum_{j=1}^p \beta_{2j}^2/\lambda_j^2)\} I\{0, (1/v - 1)\}$.
3 Compute in $\tau = 1/\sqrt{\xi}$.

Finally, we update the censored responses from $w_i \sim N\left(\alpha + \sum_{j=1}^p x_{1ij}\beta_{1j} + \sum_{j=1}^p x_{2ij}\beta_{2j}, \sigma^2\right)$ lower truncated at $\log t_i^*$.

We have written the posterior sampling strategy in an R package **circgene** format and make it available on github at the address https://github.com/MAITYA02/circgene.

2.4 Posterior Analysis

The goal is to identify potential common genes that affects survival rates using copy number change and changes in mRNA expression. Frequentist procedures such as lasso [28] or other extensions of lasso are designed to provide a sparse solution of the parameter vector. A Bayesian method, however, provides the posterior distribution of the parameter from which a posterior summary is extracted to make inferences. Motivated by this, researchers seek a unified proposal for obtaining good choice of the posterior summary which in turn recovers important features in high dimensional settings. Recently, [29] proposed a k-means clustering on the posterior space a successful application has been achieved in [22]. When a shrinkage prior such as the Horseshoe is used, even though the posterior estimate of $\boldsymbol{\beta}$ are not exactly zero, the MCMC sample obtained from posterior distribution of $\boldsymbol{\beta}$ is expected to produce two subsets – one set will be clustered around zero corresponding to noise variables and the other one will be away from zero corresponding to signals. Hence, fitting a k-means algorithm with $k = 2$, makes sense to determine the cluster of significant predictors i.e. the cluster with smaller size.

By construction, each gene is related to the survival by copy number variation and RNASeq data. Hence, there are two regression coefficient parameters for each gene corresponding to CNV data and RNASeq data respectively. However, in order to carry out an integrative analysis and to recover the common genes which are significant for both CNV data and RNASeq data a single set of $\boldsymbol{\lambda}$ is specified. Recall that, λ_j controls the shrinkage of the j-th gene, and specification of one λ_j for both β_{1j} and β_{2j} is the key to accomplish recovering the common genes. As a consequence, toward the goal of recovering common genes we fit a 2-means clustering algorithm on the posterior mean of $\boldsymbol{\lambda}$; the cluster which will have smaller size can be mapped to the corresponding genes which are significant for copy number change and mRNA data because of the structure of our model formulation.

3 Results in TCGA Data

The main goals of this study were to develop a Bayesian shrinkage coupled with Bayesian accelerated failure time model to carry out an integrative analysis and to select circadian genes that play important roles in cancer progression. We focused on transcriptome data because of its wide availability. For datasets of human cancer, we analyzed breast cancer data from TCGA. We collected the desired dataset from a version hosted by Broad Institute using the R package **TCGA2STAT** [30]. Overall, we obtained 366 samples with very high censoring rate (84.4%). Multiple clock genes are reported to play role in cancer progression [31]. From the known reports we have selected 10 genes [32] to investigate their importance in cancer progression. Earlier studies showed how the expression pattern of circadian genes are altered in different cancer as well as how different circadian genes get mutated in different types of cancer [33].

To investigate the direction of the association of the gene expression we first divide the expressions as measured by the CNV at their median point and classified the genes into two groups viz. "high" and "low" groups. The high values refer to the higher expression values of genes and the low values refer to the lower expression values of those genes. Thus, when the average survival times of the high group is higher than the low group that implies that the survival is positively associated with the expressions of that gene. Similarly, it is said that the survival time is negatively associated with the expressions of a gene when the average high expression values are lower than the average low expression values that is as expressions tend to lower then the survival times tend to go higher. To confirm this, we produce their survival times summarized by boxplots in Figures 1 and 2, with red boxes denoting high group and green boxes denoting the low group which means that their expression measurements are lower than the other group. From this visualisation it is evident that if the circadian genes are positively associated (Figure 1) or negatively associated (Figure 2). A similar plot was obtained using RNASeq data.

To identify clock genes associated with breast cancer survival, we used Bayesian shrinkage coupled with Bayesian accelerated failure time (AFT) on the TCGA data set. We obtain the posterior estimates averaging over 100,000 markov Chain Monte carlo (MCMC) samples after 10,000 samples as burnin. Using the method described in Section 2.4 on the posterior samples, we found that NPAS2, PER1, PER2, CRY2, CRY1, CSNK1E, and DEC1 are positively correlated with patients survival for breast cancer patients as shown by Figure 1, where as PER3, TIMELESS and MT2 are found to be negatively correlated with patient survivals (Figure 2). Finally, we integrated CNV and RNAseq expression data together to find out DEC1 is positively correlated with breast cancer patients survival. To see this we refer the readers to the posterior 95% Bayesian credible intervals for each gene reported in Table 1. One can note that the credible intervals cover the point 0 for all genes including DEC1, however, the intervals due to DEC1 include a bigger length of the positive part of the real line than that of other intervals which make the gene DEC1 significant.

It is known that DEC1 regulates the expression of factors associated with tumor growth and apoptosis, and is therefore linked to tumor progression. DEC1 is known to regulate breast cancer cell proliferation by stabilizing cyclin E protein, which

Table 1: 95% Bayesian credible intervals for coefficients

Gene	Interval for CNV data	Interval for RNASeq data
NPAS2	(-1.06, 1.36)	(-0.01, 0.16)
PER1	(-0.35, 1.50)	(-0.03, 0.07)
PER2	(-0.72, 1.53)	(-0.13, 0.05)
PER3	(-1.85, 0.44)	(-0.06, 0.10)
CRY2	(-0.78, 1.28)	(-0.10, 0.09)
CRY1	(-0.63, 1.57)	(-0.15, 0.11)
TIMELESS	(-1.67, 0.79)	(-0.05, 0.02)
CSNK1E	(-0.22, 1.33)	(-0.02, 0.03)
DEC1	(-1.00, 5.57)	(-1.98, 8.48)
MT2	(-0.83, 0.16)	(-0.00, 0.00)

Table 2: Simulation results based on 100 simulated datasets. All results are in proportion.

Censoring Rate	True Model Size	True Model Selection	Estimated Model Size	False Positive Rate	False Negative Rate
28%	2	1.00	2.00	0.00	0.00
35%	5	0.46	4.98	0.53	0.53
88%	2	1.00	2.00	0.00	0.00
82%	5	0.39	4.61	0.32	0.40

delays the progression of cell cycle S phase [34]. Our analysis using TCGA tumor samples supports a key role for DEC1 in tumor progression and suggests that DEC1 expression levels can be used to predict survival rates in breast cancer patients.

To asses the convergence in the MCMC chain we provide the trace plots of β_{11} and β_{19} in Figures 3 and 4 respectively. We notice good mixing in both posterior samples. The corresponding Gelman-Rubin convergence diagnostics are 1 and 1, which are less than 1.1 confirming that a convergence has been achieved. Similar trace plots can be obtained for other parameters which we skip here for brevity.

4 Results in Simulated Data

In this Section we provide some simulation studies with a similar setting as in the breast cancer data example in Section 3. We consider two design matrices – X_1 and X_2 both with dimension 300×10, assuming X_1 and X_2 correspond to CNV and RNASeq respectively, which replicates similar settings of the breast cancer data in TCGA. For the sake of simplicity the columns (genes) of the matrices are generated from uncorrelated Gaussian distribution with unit variance covariance matrix. We consider three true scenarios for the coefficient vector $\boldsymbol{\beta}$ – two of the genes for each X_1 and X_2 are significant and five of the genes for each X_1 and X_2 are significant. The true values of the significant coefficients are generated from an Uniform distribution with parameters (-1.5, 1.5). Then the survival times of the subjects are generated according to equation (1) in log scale with $\sigma^2 = 1$. The censoring rate is induced assuming a Gamma distribution. The censor rate in a particular example can be created by appropriately setting the parameters of the Gamma distribution (see also [19]). In this way we consider three censored situation depending on how many subjects are censored. In each setting we produce 100 simulated datasets and fit our proposed model developed in Section 2. In Table 2 we report the results averaged over 100 datasets.

To obtain the posterior operating characteristics, for each simulation we consider 10,000 MCMC samples after discarding 1,000 burnin samples, no thinning was

considered. After fitting our developed method (see Section 2) to the simulated datasets we processed the posterior samples using the method described in Section 2.4 to compute four different matrices to assess the performance of the developed method –

- **True Model Selection:** Proportion of times the true model or the data-generating model is identified by our method.
- **Estimated Model Size:** Average model size of the model identified by our method.
- **False Positive Rate:** Proportion of coefficients identified as significant by our method when in fact they were not present in the true model. Any method with lower false positive rate is preferred.
- **False Negative Rate:** Proportion of coefficients which were not identified as significant by our method when in fact they were present in the true model. Any method with lower false negative rate is preferred.

The simulation results are summarized in Table 2. One can note that the performance of the proposed method is well in terms of selecting the true model. The good performance is especially evident when the data generating model or the true model is sparse. For instance, when the 28% samples are right censored, the true model size is 2, the true model is recovered in each of the 100 simulated datasets. As a consequence, the false positive rate and false negative rate are 0. The good performance continues even when 88% data is censored, which is similar to the Breast cancer data (84.4% censored data). However, the performance degrades when the true model is not sparse; this can be justified because Horseshoe prior is widely known to produce parsimonious solution.

5 Discussion and Conclusion

To the best of our knowledge, this study is the first in its kind to analyze breast tumor samples data from TCGA for integrating omics data and selecting the circadian gene important in cancer progression. In this work we have exploited shrinkage nature of the global local Horseshoe prior in order to integrate the two platforms – copy number variation and RNAseq data to uncover the important genes associated with the patient survivals. By virtue of the unique specification of the global parameters and local parameters of the Horseshoe prior the analysis made it possible to identify the common genes which are important via the both types of the measurements of the gene expressions. The TCGA data of circadian gene measurements for the brain cancer patients discover the DEC1 as the associated gene with the patient survival which has already been known in the literature for its role in cancer patients.

In Section 2.3 the MCMC chain is constructed on the conditional distributions of the primary parameters α, $\boldsymbol{\beta}$, and σ^2 by virtue of which straightforward Gibbs sampling is carried out. Nonetheless, this construction is only possible if one has the full data likelihood computed. In a censored data scenario the data likelihood is typically consists of some censored observations which preclude to carry out Gibbs sampling. To mitigate this issue we augmented the full data and sampled the censored observations from truncated space. As we have noted above, the posterior sampling achieved good convergence in the application we have considered. Alternatively, a pseudo marginal sampling approach [35] can also be explored if good

mixing is obtained. For instance, the likelihood can be estimated unbiasedly using Poisson estimator [36] or the difference estimator [37]. Then the regression coefficients α and $\boldsymbol{\lambda}$ and the variance parameter σ^2 can be updated using the Metropolis Hastings scheme using a suitable proposal distribution. However, $\boldsymbol{\lambda}$ and τ can still be updated using the method described in Section 2.3. In this implementation if $n >> p$ one can use two step Metropolis Hastings approach, developed in [38] when updating the regression parameters to avoid observing a long MCMC chain for convergence.

List of abbreviations

AFT: Accelerated Failure Rate Models; CNV: Copy Number Variation; TCGA: The Cancer Genome Atlas; NCBI: National Center for Biotechnology Information; DNA: Deoxyribonucleic acid; RNA: Ribonucleic acid; MCMC: Markov chain monte carlo.

Declarations

Ethics approval and consent to participate
No ethics approval was required for the study.

Consent for Publication
Not applicable.

Availability of data and material
The data used for this study is available in TCGA.

Competing interests
The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
Research reported in this publication was partially supported by National Cancer Institute of the National Institutes of Health under award number R01CA194391 and NSF CCF-1934904 (PI: Dr B.K.M.). T.R.S. was supported through the NIH T32 Training grant (PI: Dr Raymond J Carroll);

Authors' contributions
AKM, DBP, BKM, and TRS designed the study. AKM, SCL, LH, and TRS collected and analyzed the data. AKM, DBP, and TRS wrote the manuscript. All authors have read and approved the manuscript

Acknowledgements
We are grateful to the Editor and the Referees for their helpful comments which substantially improved this article.

Author details

[1]Early Clinical Development Oncology Statistics, Pfizer Inc., 10777 Science Center Drive, 92121 San Diego, USA. [2]Department of Statistics, Texas A&M University, 3143 TAMU, 77843 College Station, USA. [3]Department of Biology, Texas A&M University, 3258 TAMU, 77843 College Station, USA.

References

1. Andreani TS, Itoh TQ, Yildirim E, Hwangbo DS, Allada R. Genetics of circadian rhythms. Sleep Medicine Clinics. 2015;10(4):413–421.
2. Sahar S, Sassone-Corsi P. Metabolism and cancer: the circadian clock connection. Nature Reviews Cancer. 2009;9(12):886.
3. Fu L, Kettner NM. The circadian clock in cancer development and therapy. In: Progress in Molecular Biology and Translational Science. vol. 119. Elsevier; 2013. p. 221–282.
4. Lowrey PL, Takahashi JS. Genetics of circadian rhythms in Mammalian model organisms. In: Advances in Genetics. vol. 74. Elsevier; 2011. p. 175–230.
5. Nakashima A, Kawamoto T, Honda KK, Ueshima T, Noshiro M, Iwata T, et al. DEC1 modulates the circadian phase of clock gene expression. Molecular and cellular biology. 2008;28(12):4080–4092.
6. Fujimoto K, Hamaguchi H, Hashiba T, Nakamura T, Kawamoto T, Sato F, et al. Transcriptional repression by the basic helix-loop-helix protein Dec2: multiple mechanisms through E-box elements. International Journal of Molecular Medicine. 2007;19(6):925–932.
7. Kondo J, Sato F, Fujimoto K, Kusumi T, Imanaka T, Kawamoto T, et al. 57Arg in the bHLH transcription factor DEC2 is essential for the suppression of CLOCK/BMAL2-mediated transactivation. International Journal of Molecular Medicine. 2006;17(6):1053–1056.
8. Kawamoto T, Noshiro M, Sato F, Maemura K, Takeda N, Nagai R, et al. A novel autofeedback loop of Dec1 transcription involved in circadian rhythm regulation. Biochemical and Biophysical Research Communications. 2004;313(1):117–124.

9. Sato F, Bhawal UK, Kawamoto T, Fujimoto K, Imaizumi T, Imanaka T, et al. Basic-helix-loop-helix (bHLH) transcription factor DEC2 negatively regulates vascular endothelial growth factor expression. Genes to Cells. 2008;13(2):131–144.
10. Shilts J, Chen G, Hughey JJ. Evidence for widespread dysregulation of circadian clock progression in human cancer. PeerJ. 2018;6:e4327.
11. Huang S, Chaudhary K, Garmire LX. More is better: recent progress in multi-omics data integration methods. Frontiers in Genetics. 2017;8:84.
12. Maity AK, Lee SC, Mallick BK, Sarkar TR. Bayesian Structural Equation Modeling in Multiple Omics Data with Application to Circadian Genes. Bioinformatics. 2020;.
13. Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, et al. The cancer genome atlas pan-cancer analysis project. Nature Genetics. 2013;45(10):1113.
14. Wang C, Parmigiani G, Dominici F. Bayesian effect estimation accounting for adjustment uncertainty. Biometrics. 2012;68(3):661–671.
15. Wong CH, Siah KW, Lo AW. Estimation of clinical trial success rates and related parameters. Biostatistics. 2019;20(2):273–286.
16. Stoolmiller M, Snyder J. Modeling heterogeneity in social interaction processes using multilevel survival analysis. Psychological Methods. 2006;11(2):164.
17. Sinding-Larsen R, Xu J. Bayesian discovery process modeling of the lower and middle Jurassic Play of the Halten Terrace, Offshore Norway, as compared with the previous modeling. Natural Resources Research. 2005;14(3):235.
18. Peltola T, Havulinna AS, Salomaa V, Vehtari A. Hierarchical Bayesian Survival Analysis and Projective Covariate Selection in Cardiovascular Event Risk Prediction. In: BMA@ UAI. Citeseer; 2014. p. 79–88.
19. Maity AK, Bhattacharya A, Mallick BK, Baladandayuthapani V. Bayesian Data Integration and Variable Selection for Pan-Cancer Survival Prediction using Protein Expression Data. Biometrics. 2019;.
20. Carvalho CM, Polson NG, Scott JG. The horseshoe estimator for sparse signals. Biometrika. 2010;97(2):465–480.
21. Tanner MA, Wong WH. The calculation of posterior distributions by data augmentation. Journal of the American Statistical Association. 1987;82(398):528–540.
22. Maity AK, Carroll RJ, Mallick BK. Integration of survival and binary data for variable selection and prediction: a Bayesian approach. Journal of the Royal Statistical Society: Series C (Applied Statistics). 2019;68(5):1577–1595.
23. Bhattacharya A, Chakraborty A, Mallick BK. Fast sampling with Gaussian scale mixture priors in high-dimensional regression. Biometrika. 2016;103(4):985–991.
24. Gelfand AE, Hills SE, Racine-Poon A, Smith AF. Illustration of Bayesian inference in normal data models using Gibbs sampling. Journal of the American Statistical Association. 1990;85(412):972–985.
25. Bonato V, Baladandayuthapani V, Broom BM, Sulman EP, Aldape KD, Do KA. Bayesian ensemble methods for survival prediction in gene expression data. Bioinformatics. 2010;27(3):359–367.
26. Neal RM. Slice sampling. The Annals of Statistics. 2003;31(3):705–767.
27. Polson NG, Scott JG, Windle J. The Bayesian bridge. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2014;76(4):713–733.
28. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 1996;58(1):267–288.
29. Li H, Pati D. Variable selection using shrinkage priors. Computational Statistics & Data Analysis. 2017;107:107–119.
30. Wan YW, Allen GI, Anderson ML, Liu Z. TCGA2STAT: Simple TCGA Data Access for Integrated Statistical Analysis in R; 2015. R package version 1.2.
31. Gery S, Koeffler H. The role of circadian regulation in cancer. In: Cold Spring Harbor symposia on quantitative biology. vol. 72. Cold Spring Harbor Laboratory Press; 2007. p. 459–464.
32. Yang N, Williams J, Pekovic-Vaughan V, Wang P, Olabi S, McConnell J, et al. Cellular mechano-environment regulates the mammary circadian clock. Nature Communications. 2017;8:14287.
33. Li HX. The role of circadian clock genes in tumors. OncoTargets and Therapy. 2019;12:3645.
34. Bi H, Li S, Qu X, Wang M, Bai X, Xu Z, et al. DEC1 regulates breast cancer cell proliferation by stabilizing cyclin E protein and delays the progression of cell cycle S phase. Cell Death & Disease. 2015;6(9):e1891.
35. Andrieu C, Roberts GO, et al. The pseudo-marginal approach for efficient Monte Carlo computations. The Annals of Statistics. 2009;37(2):697–725.
36. Wagner W. Unbiased multi-step estimators for the Monte Carlo evaluation of certain functional integrals. Journal of Computational Physics. 1988;79(2):336–352.
37. Quiroz M, Kohn R, Villani M, Tran MN. Speeding up MCMC by efficient data subsampling. Journal of the American Statistical Association. 2018;.
38. Payne RD, Mallick BK. Two-stage Metropolis-Hastings for tall data. Journal of Classification. 2018;35(1):29–51.

Figures

Additional Files

Additional file 1 — Sample additional file title

Additional file descriptions text (including details of how to view the file, if it is in a non-standard format or the file extension). This might refer to a multi-page table or a figure.

Additional file 2 — Sample additional file title

Additional file descriptions text.

Figure 1: Displayed boxes show that NPAS2, PER1, CRY2, CRY1, CSNK1E, and DEC1 are positively associated with patients survival. Right panel boxes show that PER3, TIMELESS and MT2 are negatively associated with patients survival. The left (green) boxes belong to the lower CNV measurements which means that the measurements are lower than the median point. The right (red) boxes belong to the higher CNV measurements which means that the measurements are more than the median point. The plotted boxes are the survival times of the individuals.

Figure 2: Displayed boxes show that PER3, TIMELESS and MT2 are negatively associated with patients survival. The left (green) boxes belong to the lower CNV measurements which means that the measurements are lower than the median point. The right (red) boxes belong to the higher CNV measurements which means that the measurements are more than the median point. The plotted boxes are the survival times of the individuals.

Figure 3: Trace plots of β_{11}.

Figure 4: Trace plots of β_{19}.