ESCI upf.
School of International Studies

MEMÒRIA DEL TREBALL DE FI DE GRAU DEL GRAU (ESCI-UPF)

A Bayesian non parametric integration approach for functional regulatory network inference in hematopoietic development from multi-omics data

AUTOR/A: Laia Carolina Alcalà Plazas	**NIA:** 106418
GRAU: Bachelor's degree in Bioinformatics	
CURS ACADÈMIC: 2022 - 2023	
DATA: 21/06/2023	
TUTOR/S: Anas A Rana (a.a.rana@bham.ac.uk)	

A Bayesian non parametric integration approach for functional regulatory network inference in hematopoietic development from multi-omics data

LAIA C. ALCALÀ[1] **AND ANAS A. RANA**[2]

[1] *School of International Studies, Pompeu Fabra Univeristy, Pg. de Pujades, 1, 08003 Barcelona, ES*
[2] *Centre for Computational Biology, University of Birmingham, Haworth Building, Edgbaston, B15 2TT Birmingham, UK*

ABSTRACT – ***Motivation:*** **During hematopoiesis, Sp1 and Sp3 transcription factors cooperate for the correct development. The overall structural similarity and ubiquitous expression patterns shared by Sp1 and Sp3 suggest that these two proteins are functionally equivalent. However, they show a different effect in their absence and the interaction of these TF at genomic level in each step of the process is not fully known.**
Results: **We have developed Bayesian non parametric model called Monet (Multi-Omics NETwork inference) in order to infer the TF-gene network interaction and the TF activity strength without knowing a priori the number of participating TF in each stage of a molecular process. Statistical inference is performed in a Bayesian setting using an implementation of Hamiltonian Monte Carlo (HMC)-derived No-U Turn Sampler (NUTS) in Stan. Here, we show the Monet's capability to extract these parameters and we apply it to mouse hematopoietic data to briefly analyze the changes in Sp1 knockdown during the different cell stages. Our model allows us to make predictions on the importance of specific TF over time and under different perturbation settings.**
Supplementary material and code**: Available in GitHub (https://github.com/laiaalcala/FDP_Laia_Alcala)**

1. INTRODUCTION

Hematopoiesis is the process of formation of blood cellular component which occurs during embryonic development and adulthood to produce and replenish the blood system [1]. During embryonic development definitive blood cells including hematopoietic stem cells arise from a common progenitor of hematopoietic cells and endothelial cells (hemangioblast) derived from the mesoderm in the dorsal aorta [2]. Through the different stages of the development, various Transcription Factors (TF) work together in order to generate or restrict different responses and regulate transcription of stage-specific genes forming a dynamic network and playing a major role in the overall development of the hematopoietic lineage [3].

For the correct development, both tissue-specific and ubiquitously TF such as Specificity 1 (Sp1) and Specificity 3 (Sp3) proteins are needed [4]. These two TF are closely related family members, ubiquitously expressed and both recognize G-rich DNA motifs [5]. They have highly similar structures, with a 90 percent of sequence homology in their DNA binding domains. A study [4] showed that Sp3 cooperates with Sp1 to enable hematopoiesis and, although both TFs recognize the same DNA sequence element and have similar binding affinity, they show a different effect in their absence. The Sp1 gene knockout or the removal of the Sp1 DNA binding domain, which is a highly conserved region, causes multiple heterogeneous phenotypic abnormalities and die in utero in mice [6]. Sp3 acts as a repressor of Sp1 transactivation and Sp3 knockout also showed evidence of defect in hematopoiesis and cardiac development, finally causing the postnatal death of mice [7, 8]. In addition, mice heterozygous for both TFs also caused its death during gestation with a wide range of phenotypes [9].

In conclusion, both Sp1 and Sp3 Transcription Factors, despite having a very similar DNA sequence,

they have different functions in the cell and are required for a robust hematopoietic specification. Studying hematopoiesis can help to understand better the molecular mechanisms behind blood disorders and cancers, and this could lead to new treatments. However, the interaction of these TF at genomic level in each step of the process is not fully known. The effect of perturbation of the cells trajectory in a process such as changes in speed, the interaction of the different Transcription Factors and genes, the changes in the pathway of the cell or its final state can be a challenging task, specially when working with time courses and bulk data.

In order to understand the role of Sp1 in the adult hematopoietic system, a Bayesian non parametric model called MONET (Multi-Omics NETwork inference) has been developed to infer the TF-gene network interaction and the TF activity strength without knowing a priori the number of participating TF in each cell stage.

A. Experiment

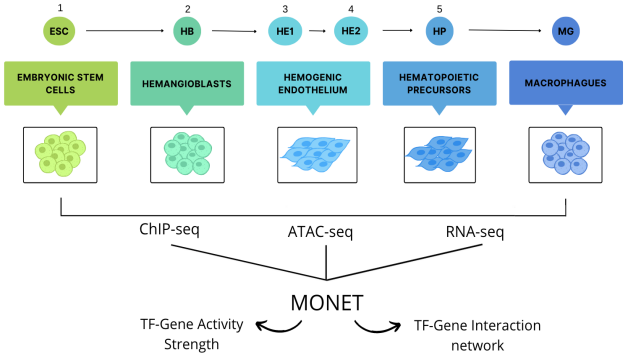


Fig. 1. Experiment scheme.

This experiment [4] was performed in vitro to avoid the problems the problems of embryonic mice death and was based on obtaining data on different stages of the differentiation process of mouse embryonic stem cells (ESC) during hematopoiesis for both Wild Type (WT) Sp1 and the TF Sp1 knockdown ($Sp1^{\triangle DBD/\triangle DBD}$). They obtained samples for the five stages of development; Embryonic stem cells (ESC), Hemangioblasts (HB), Hemogenic endothelium (HE1 and HE2) and Hematopoietic precursors (HP).

The resulting data was a complete WT time course as well as a complete time course of Sp1 knockdown. Multiple omics level data was generated, including RNA-seq data, ChIP-seq data and ATAC-seq data (see Figure 1). RNA-seq data encodes gene activity for any given cell, ATAC-seq data encodes chromatin accessibility state and ChIP-Seq identifies the binding sites of the different TF during hematopoiesis. The interplay between transcriptomics and epigenomics data is a compelling need to understand gene regulatory mechanisms [10]. Chromatin accessibility to the transcription machinery regulates gene expression and *vice versa*, some non-coding RNAs can also affect local chromatin states [11]. By integrating ATAC-seq and RNA-seq data, which are longitudinal data sets containing all time-points of interest , a novel approach to perform regulatory network inference has been developed. The proposed non-parametric Bayesian model does not need any pre-existing network, so this approach is entirely data-driven.

Having a cost-effective and data-driven tool that makes predictions on the most important TF interacting in a certain time in the different stages of a biological process can be a starting point for future studies and therapeutic approaches. Furthermore, there is a drawback on using a model that incorporates ChIP-seq data, which is the restricted availability of TF and time-points as it is only available for some TF and sometimes for only one time-point.

A similar approach was done in order to infer a functional regulatory network (FRN) using data integration from which this project was inspired by [12]. It is also Bayesian integration model called CRNET which infers from a ChIP-Seq data set at one time-point and a RNA-Seq time-course. Specifically, this model integrates prior binding information from ChIP-seq with time-course gene expression data from RNA-seq. However, CRNET is not useful for what we are interested in, as it is more focused in one-time point or steady state. Our interest is in understanding perturbation in a development time-course.

B. The Bayesian non parametric model

This non parametric Bayesian model infers the TF-network interaction and the number of participating TF without knowing a priori how many transcription factors will interact. We can write the model in the following way:

$$y_g = y_0 + a_g \circ z_g + \epsilon_g \quad \textbf{(1)}$$

Where y_g is the gene expression time-course matrix

measurement of gene g (where the rows are the genes and the columns are the time points), y_0 is the gene expression vector of G genes at time $t = 0$, the initial measurement and $\epsilon_g \sim N(0, \sigma^2)$ is the error variable. a_g is a binary matrix defining chromatin accessibility states for the g-th gene and z_g is also a matrix which denotes the gene-specific latent trajectories. Then, the time-course of each gene is modeled as a combination of the initial gene expression measurement and the gene-specific latent trajectory, which is modeled as:

$$z_g = \sum_{k=1}^{\infty} w_k b_{g,k} x_k \tag{2}$$

Where w_k is a potentially infinite vector of weights and x_k is the shared latent transcription factor activities modulated by a gene-specific transcription factor interaction network ($b_{g,k}$).

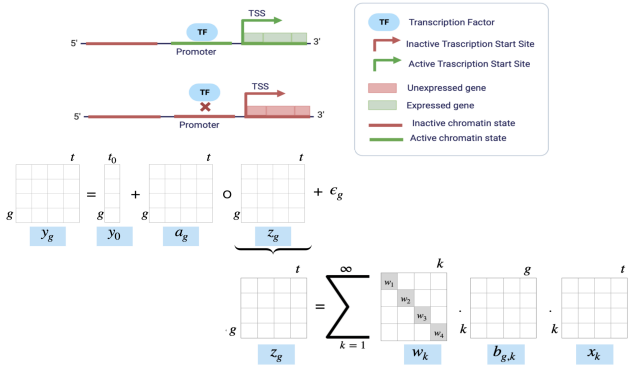


Fig. 2. Model scheme

The priors chosen for the model are the following:

$$w_k \sim DP(alpha) \tag{3}$$

$$b_{g,k} \sim Laplace(0, \beta) \tag{4}$$

$$x_k \sim TruncatedNormal(0, \sigma_x^2) \tag{5}$$

In this point is where the Dirichlet Process is applied as the prior of w_k , the infinite vector of weights in Eq. 2, is given by a DP. The Laplace distribution has been chosen as the prior of the gene-specific TF interaction network instead of the Normal distribution in order to impose sparsity. The Laplace distribution has a peak around zero so there is a greater distribution mass, while Normal distribution is more diffuse around zero, so non-zero values have greater probability mass. For the shared latent transcription factor activities the truncated normal distribution is used, which is the probability distribution derived from that of a normally distributed random variable by limiting value within a specific scope. This range sets boundaries on the possible values the variable can take, either below, above, or both. With this, the complexity of the inference is reduced and flexibility across genes is allowed [13].

C. Data

We will apply the model to real data from Peter Keane, Christopher Middleton and others who performed an expression and a genome binding/occupancy profiling by high throughput sequencing of the hematopoietic development in *Mus musculus* in the Institute for Cancer and Genomic Sciences. The resulting data has information for 12.000 genes and 10TF, and was published on June 2019. It can be found at the NCBI gene expression omnibus under accession numbers GSE126497 and GSE126501.

The data generated by ATAC-seq, RNA-seq and ChIP-seq for the 5 time-points must be prepared before using it in the model. The binarized measure of chromatin state over time data is generated by ATAC-seq (where 0 represents a closed status and 1 represents an open status for a given promoter site) and the gene expression over time is generated data by RNA-seq. Both data sets are longitudinal and are the input of MONET, corresponding to the y_g and a_g parameters of the model (Eq. 1) respectively. We also have the ChIP-seq data, which identifies the binding sites of specific TF over time and will be used for the model validation. All the data need to be preprocessed and analyzed using different tools in order to clean and keep the highest quality data.

D. Objectives and contribution

The objective of this project is (1) to demonstrate MONET's capability in extracting time-dependent TF-gene interaction networks, (2) to apply the model to evaluate the significance of the most important TF and their interaction in the different stages of hematopoiesis and (3) to analyze the changes of those interactions with the Sp1 knockdown perturbation in this process.

My contribution to this project is to help with the STAN implementation of the model, simulate the appropriate data to check the performance of the model, fit the model to real data applied to hematopoietic development and create a visualization of the results

to briefly analyze them.

2. MATERIALS AND METHODS

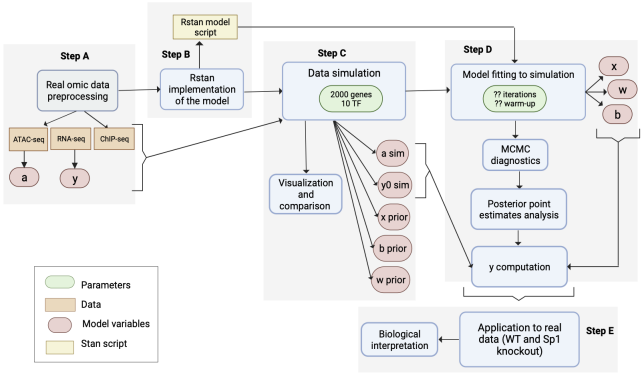


Fig. 3. Pipeline followed to develop and applicate the model to hematopoietic data.

A. Data preprocessing and analysis

The data preprocessing and analysis has been already done and the resulting data can be found in the GitHub repository. For ATAC-seq data this workflow has been used: sequencing reads quality assessment and cleaning, alignment and comparison with replicates. FastQC has been used to asses quality on sequencing data, where sequencing adapter contamination was observed and it was removed with Cutadapt, as well as bad quality reads. Then, an alignment with the mouse genome with Bowtie2 was done and the reads not uniquely aligned were also removed. Finally, the peaks, belonging to open chromatin regions, were compared with ATAC-seq replicates and merged using Bedtools. Peaks from both ATAC-seq and ChIP-seq were be annotated and classified; as within the promoter if they were within 2kb of the Transcription Start Site (TSS) or as distal otherwise. Promoter regions were determined if the ATAC-seq or ChIP-Seq peak was within 1.5 Kb of the TSS. Then, peaks annotations were transformed into a binary matrix, and a value of 1 was assigned if the gene TSS was found to have a peak associated with it and a value of 0 otherwise.

For RNA-seq data these steps has been followed: sequencing reads quality assessment and cleaning, alignment, gene expression quantification. As before, raw sequencing reads were assessed with FASTQC and adapter contamination and low sequencing reads were removed with trimmomatic. Then, reads were aligned to the mouse genome with the Hit2 tool. Only mapped reads with FPKM values greater than 1 in at least one sample (measured with stringtie) were used for further analysis. Finally, raw FPKM values were normalized and log2-transformed with R.

B. Rstan implementation

To do the inference this model has been implemented in STAN, a probabilistic programming language which does full Bayesian statistical inference with MCMC sampling from the posterior. It has been implemented it inside an R package called "monet" (available on request - https://github.com/anasrana/monet) using the R Stan interface rstan [14]. Monet uses sampling based on an Hamiltonian Monte Carlo (HMC)-derived No-U Turn Sampler (NUTS) implementation, which is the standard MCMC algorithm used in Stan. NUTS requires no tuning parameters and can efficiently sample from posterior distributions with correlated parameters [15, 16].

The Stan code is divided in four blocks; the data block where known parameters are declared, the parameters block where the parameters we want to estimate are declared (x_k, $b_{g,k}$ and w_k), the transformed parameters block where the model equations are defined and the model block, where the non parametric Bayesian model is defined and prior distributions and the likelihood are specified to define the posterior [17].

The model implementation is available in the GitHub link.

C. Data simulation

We have simulated some data similar to what we observe in our Sp1 Wild Type real data using the Monet package in order to show the accuracy of the estimation.

We have simulated 2000 genes with 30 TF where we have sampled y_0 from a normal distribution from our real RNA-seq data using it as a baseline, x_g from a truncated normal distribution from our real ATAC-seq data and $b_{g,k}$ from a Laplace distribution from our real ChIP-seq data. ϵ_g has been sampled from a normally distributed noise and the w_k component, belonging to the sum of z_g, has not been used for the simulation of the data because, on the one hand, including it in would lead to a synthetic data very close to the real one and, on the other hand, we specify the number of TF in the simulation.

Then, our simulation is based on

$$y_g = y_0 + a_g \circ z_g + \epsilon_g \quad (6)$$

where

$$z_g = \sum_{k=1}^{\infty} b_{g,k} x_k \quad (7)$$

The resulting Monet simulation function called takes as input the ChIP-seq, ATAC-seq and RNA-seq data paths, the number of genes, the number TF and the number of time points we want to simulate.

From the ChIP-seq data it samples a matrix with 2000 genes from the real data. For the simulation of $b_{g,k}$ it randomly generates 20000 values (2000 genes x 30 TF) from a a Laplace distribution with mean equal to 0 and standard deviation equal to 2. Values near 0.01 are changed to 0 to introduce more sparsity to the distribution. For the simulation of x_k it randomly samples from a truncated normal distribution with lower bound and mean equal to 0 and standard deviation equal to 2. Then, z_g is simulated by multiplying both simulations $b_{g,k}$ and x_k.For the initial gene expression y_0 it samples 2000 genes from the first cell stage column of the data.

For the ATAC-seq data simulation a_g it samples a matrix with 2000 genes from the real data. y_g is obtained by combining y_0, a_g, z_g and a normally distributed noise following the equation 6.

Finally, it generates a list containing all the results; the number of genes, the number of TF, the number of time points and the simulations of y_0, $b_{g,k}$, x_k, y_0 and a_g.

Once the simulation has been completed we have standardized the data and done a statistical comparison with the processed real data to check its similarity.

D. Model fitting to simulated data

We have fitted the model to the simulated data using the monet STAN fit model function we created in order to estimate the posterior of $b_{g,k}$, x_k and w_k.

The output comes from the MCMC algorithm sampler it is applied in the STAN fit (run for 6000 iterations after 2000 warm-up iterations). We first have checked the performance of the MCMC sampling with a trace plot, as it is important to make sure that the Markov Chain converges to a stationary distribution and that there is not any issue with our MCMC sample [18]. Later, we have extracted the posterior point estimation of these parameters (see Figure 10). Having $b_{g,k}$, x_k and w_k from the posterior point estimations and y_0 and a_g from the simulated data, we have been able to compute y_g, which contains the estimated gene expression of the different genes from time points 2 to 5 and we have standardized the resulting data set.

Then, we have filtered the genes that appear in both y_g (simulated and estimated) and we have done a comparison of both in order to see how well the prediction of the expression is using the estimated parameters.

E. Application to real data

Once we have checked the good performance of the monet model in the simulated data set, we have applied it to the observed data. We have repeated the following steps twice, for WT and Sp1 hypomorph time-course data.

Before fitting the model, we have removed genes that show constant expression as we are interested in genes with expression fluctuations. For the first run we have kept genes whose expression standard deviation >0.5, and whose expression standard deviation >2 for the second run. The reason for running this twice with different standard deviations is to explore how the model behaves by removing more or less constant genes. Then, we have filtered the RNA-seq data and ATAC-seq data to verify that both data sets contain the same genes and to create a new subset of 1000 genes of the data. The gene selection has been based on the difference in expression between WT and Sp1hyp and on high signal to noise in gene expression. Next, we have prepared the data in the correct input for the Rstan model and we have fitted it for the 1000 genes and 20TF.

3. RESULTS AND DISCUSSION

A. Data simulation

We have achieved a very similar omic data to the real one (y_g and a_g) and accurate priors for the parameters we want to estimate (x_k, $b_{g,k}$ and w_k). In Figure 7 we show the density distribution of the total gene expression for the observed genes and the simulated ones, where it can be seen that the data distribution shapes are the same. We have also computed the densities for each cell stage of both data sets to visualize the distribution similarity shapes of the data. We did

not expect to replicate the exact data, but rather we wanted data that is similar (8a). We have also plotted the trajectories of 10 random genes of the simulated data set (see Figure 8b), where we can observe that all the simulated trajectories are different from each other.

Regarding the chromatin state (ChIP-seq data) we have also obtained a very accurate result. Both data sets have a very similar pattern in the heat maps (see Figure 9), where red colour represents a closed chromatin state and green an open chromatin state. Of course, there is a smooth difference due to the different number of genes between the simulated (2000 genes) and the observed (12000 genes) data sets.

B. Model fitting to simulated data

In Figure 4 there is the MCMC log posterior trace plot of the Monet fitted model of the simulated data (run for 2000 genes, 10TF, 6000 iterations and 2000 warm up), where we can see that the model has converged. Note that the first 2000 iterations (warm up) are removed in the plot. This result means that the effective sample size and the number of iterations and chains we have used for the simulation has been large enough for the MCMC to approximate the target distribution.

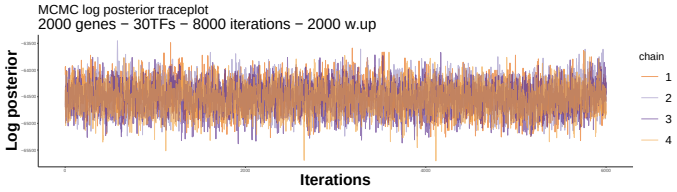


Fig. 4. Trace plot of the MCMC log posterior of the model fitting of the simulated data. This plot shows the parameter value at time t against the iteration number, where colors represent the different chains.

We have obtained the posterior point estimation for $b_{g,k}$, x_k and w_k (see Figure 10) and their density distributions (see Figures 11, 12). Note that all values of w_k in figure 10, which is the Dirichlet mixture component, are near 0 and sum to 1.

We also show the estimated trajectories (y_g) of 10 random genes based on the point estimation compared to the simulated ones (see Figure 5), where we can see an accurate reproduction of the estimated gene expressions, following the pattern of the same simulated gene expressions. Some deviation is observed, which is expected when fitting data.

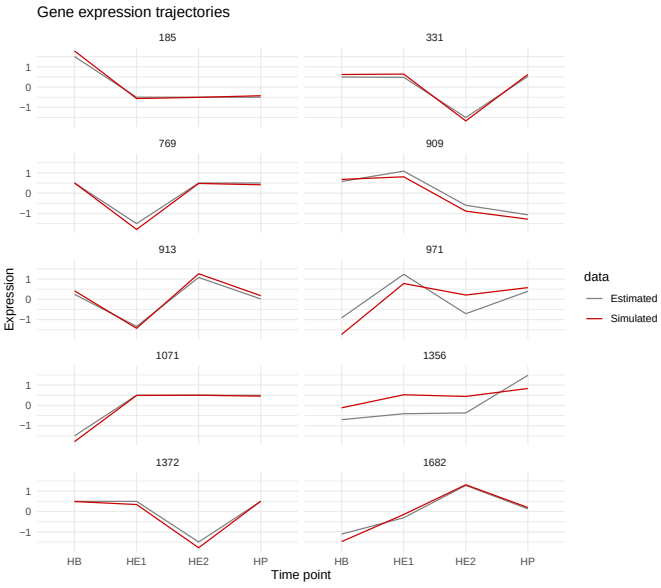


Fig. 5. Trajectories of 10 random genes of the simulated (grey) and estimated data (red) with the Monet point estimation.

C. Application to the real data

For our observed data, we can also see the convergence of the fitted model in the MCMC log posterior trace plots for both data sets (see Figure 13). In both cases, we can see a good mixture between chains and convergence in individual chains. We have obtained R-hats from 0.99 to 1.05 for WT and from 0.99 to 1.04 for Sp1 perturbation, which are lower than 1.1 so we can reaffirm that we can trust the sample for being fully representative of the posterior [19].

We have extracted the posterior point estimates of the parameters for both runs (where the standard deviation of gene expression across time is 0.5 and 2 respectively) and for both data sets. In the first case (see Figures 14 and 15) we observe a smooth difference in the Transcription Factors interaction plot between WT Sp1 and Sp1 perturbation. The second column, belonging to the HE1 cell stage is darker for $Sp1^{\triangle DBD/\triangle DBD}$, suggesting that some interactions are lost at this time point. If we change the gene list based on genes that have a higher level of variance (sd > 2) we see much more differences in the TF interaction, especially at time point 2 (see Figures 16 and 17). The interaction at this specific point for WT Sp1 remains very low up to 10 TF compared to the other time points, while in Sp1 perturbation is much higher. The likely source of that is that this run included genes with more variant gene expression.

For a better interpretation of the changes in the TF interaction we have computed the differences of the x_k matrices between WT Sp1 and Sp1 knockdown (see Figure 6). We still see a significant change in the second column (HE1 cell stage), where some TF loose activity in this time point in absence of Sp1 (i.e TF 9 in Figure 6a and TF 15 in Figure 6b) and then their activity starts increasing again in the following cell stages. There are other TF gain interactions in absence of Sp1 in the same time point (i.e TF 13 in Figure 6a and TF 6 in Figure 6b) and then decreases again. These results suggest that Sp1 TF exhibits predominant activity at the HE1 cell stage of the hematopoiesis development, and that the impact of the lack of Sp1 TF activity on TF interaction network accumulate over the following developmental stages [20]. In order to identify which are these relevant TF, we will need to use the ChIP-seq data.

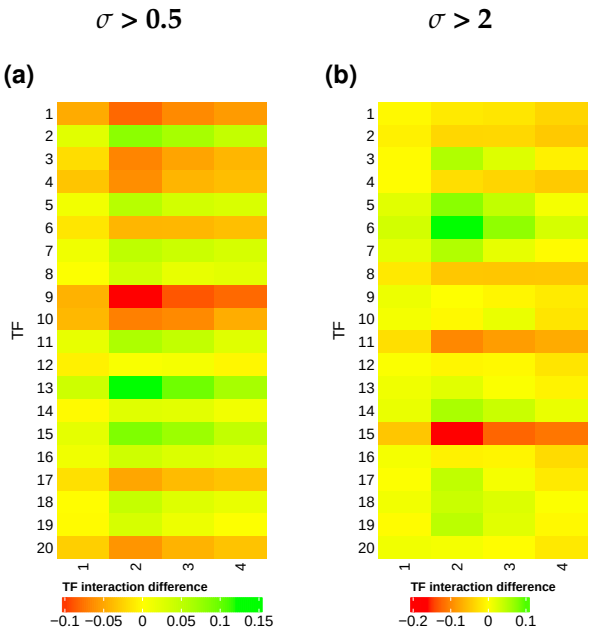


Fig. 6. x_k Sp1 WT and x_k Sp1 knockdown matrices difference for both gene lists (sd > 0.5 and sd > 2).

4. CONCLUSION

In this project, we demonstrate that the proposed Bayesian non parametric regression model designed for inferring regulatory networks and identifying temporally dependent drivers in biological systems during development achieves reasonable convergence time and accurately reproduces parameters. Apart from the fitting model algorithm, we have created a novel algorithm to create synthenic data that utilizes a mixed approach based on sampling from real data sampling for distributions. With this data, we have tested the performance of the model, and were also able to briefly analyse the results from the inference. In application to real data we still need to finalise best practices for the interpretation of results. In brief, we have seen that in the absence of Sp1 due to its knockdown there is a significant reduction in interactions between TF and genes during the process of HE1 formation from HB. Monet also allows us to reproduce gene expression trajectories with different temporal properties.

MONET is primarily designed for analyzing time course data, and having a larger number of time points in the data set improves the accuracy of estimations. However, it is common for data sets to have a limited number of time points due to constraints such as time or budget. Nevertheless, Monet is capable of providing valuable estimates with a minimum of 4 time points, where one of them represents y0, which corresponds to the gene-expression at the initial time point. By separating the regulatory network from the temporally dependent drive, Monet offers a valuable tool for performing perturbations and studying their effects on the developmental process.

Future directions

There is more work to be done in this project, such as finding a way to quantify how accurately the ATAC-seq data is reproduced apart from observing the similarity in the heat maps.Also, it needs to be settled what cut-off parameters to use in the model for the real input data to ensure that the output results can be interpretable and that TF behavior can be compared to others based on ChIP-seq data. After that, more work is also required in analyzing with more detail the differences observed with Sp1 and extracting important biological information for further investigation and ultimately therapeutic approaches.

Abbreviations

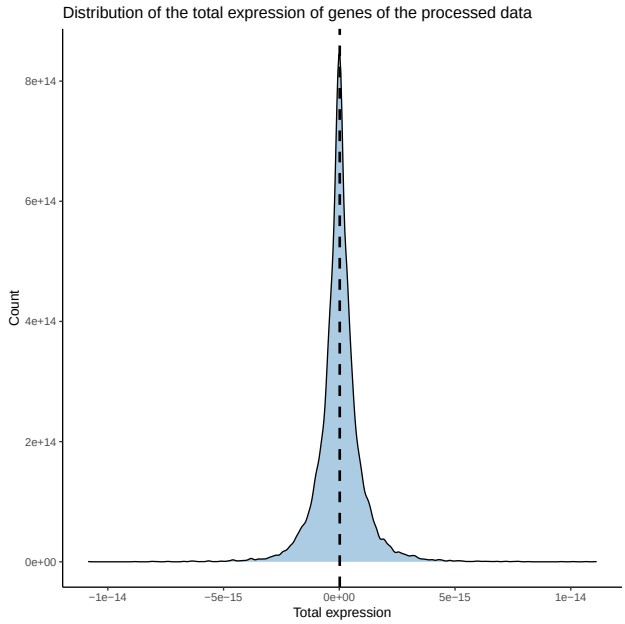
MONET, Multi-Omics Network Inference; WT, Wild Type; TF, Transcription Factor; ESC, Embryonic Stem Cells; HB, Hemangioblasts; HE1, Hemogenic Endothelium 1; HE2, Hemogenic Endothelium 2; HP, Hematopoietic precursors; TSS, Transcriptio Start Site; FRN, Functional Regulatory Network; FPKM, Fragments Per Kilobase Million

Acknowledgements

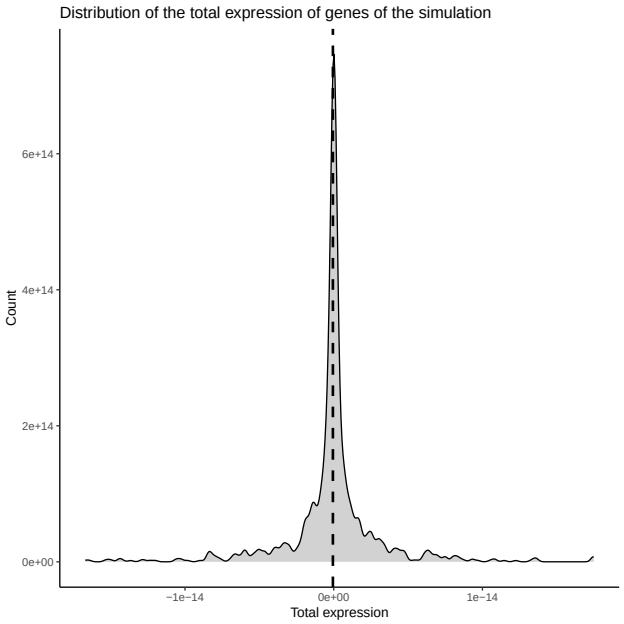
I would like to thank Professor Jean-Baptiste Crazier for letting me join the Centre for Computational Biology in the University of Birmingham these 5 months, as well as Anas A Rana for trusting in me and letting me help in his project. Thanks to both for the support since the first day and for making this stay a very enriching experience.

5. REFERENCES

1. Madhumita Jagannathan-Bogdan and Leonard I Zon. Hematopoiesis. *Development (Cambridge, England)*, Jun 2013.
2. Kiyomi Tsuji-Tamura and Minetaro Ogawa. Morphology regulation in vascular endothelial cells - inflammation and regeneration. *BioMed Central*, Sep 2018.
3. Michael G Daniel, Katrina Rapp, Christoph Schaniel, and Kateri A Moore. Induction of developmental hematopoiesis mediated by transcription factors and the hematopoietic microenvironment. *Annals of the New York Academy of Sciences*, Apr 2020.
4. Jane Gilmour, Leigh O'Connor, Christopher P. Middleton, Peter Keane, Nynke Gillemans, Jean-Baptiste Cazier, Sjaak Philipsen, and Constanze Bonifer. Robust hematopoietic specification requires the ubiquitous sp1 and sp3 transcription factors - epigenetics amp; chromatin. *BioMed Central*, Jun 2019.
5. Marjolein Meinders, Divine I. Kulu, Harmen J. G. van de Werken, Mark Hoogenboezem, Hans Janssen, Rutger W. W. Brouwer, Wilfred F. J. van Ijcken, Erik-Jan Rijkers, Jeroen A. A. Demmers, Imme Krüger, and et al. Sp1/sp3 transcription factors regulate hallmarks of megakaryocyte maturation and platelet formation and function. *American Society of Hematology*, Mar 2015.
6. Pim Visser Frank Grossveld Sjaak Philipsen Marisol Marin, Alar Karis. Transcription factor sp1 is essential for early embryonic development but dispensable for cell growth and differentiation. *Cell Press*, 89:619–628, May 1997.
7. Pieter Fokko van Loo, Peter Bouwman, Kam-Wing Ling, Sabine Middendorp, Guntram Suske, Frank Grosveld, Elaine Dzierzak, Sjaak Philipsen, and Rudolf W. Hendriks. Impaired hematopoiesis in mice lacking the transcription factor sp3. *American Society of Hematology*, 102, Aug 2003.
8. Pieter Fokko van Loo, Edris A. F. Mahtab, Lambertus J. Wisse, Jun Hou, Frank Grosveld, Guntram Suske, Sjaak Philipsen, and Adriana C. Gittenberger de Groot. Transcription factor sp3 knockout mice display serious cardiac malformations. *Molecular and Cellular Biology*, 27(24):8571–8582, 2007. PMID: 17923686.
9. Krüger I;Vollmer M;Simmons DG;Elsässer HP;Philipsen S;Suske G;. Sp1/sp3 compound heterozygous mice are not viable: Impaired erythropoiesis and severe placental defects. *Developmental dynamics: an official publication of the American Association of Anatomists*, Aug 2007.
10. Salvo Danilo Lombardo, Ivan Fernando Wangsaputra, Jörg Menche, and Adam Stevens. Network approaches for charting the transcriptomic and epigenetic landscape of the developmental origins of health and disease. *Genes*, Apr 2022.
11. Yuanyuan Sun, Nan Miao, and Tao Sun. Detect accessible chromatin using atac-sequencing, from principle to applications - hereditas. *BioMed Central*, Aug 2019.
12. Xi Chen, Jinghua Gu, Xiao Wang, Jin-Gyoung Jung, and Leena Hilakivi-Clarke. *CRNET: an efficient sampling approach to infer functional regulatory networks by integrating large-scale ChIP-seq and time-course RNA-seq data*, 34:1733–1740, May 2018.
13. Siqi Chen and Wenhao Gui. Estimation of unknown parameters of truncated normal distribution under adaptive progressive type ii censoring scheme. *Mathematics*, 9(1), 2021.
14. Stan Development Team. *RStan: The R interface to Stan*, Jan 2023.
15. Jeffrey Annis, Brent J. Miller, and Thomas J. Palmeri. Bayesian inference with stan: A tutorial on adding custom distributions - behavior research methods. *SpringerLink*, Jun 2016.
16. Matthew D. Hoffman and Andrew Gelman. The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *arXiv.org*, Nov 2011.
17. Robust statistical workflow with rstan.
18. Marco Taboga. Markov chain monte carlo (mcmc) diagnostics. *Statlect*, 2021.
19. Stan Development Team. Runtime warnings and convergence problems, 2022.
20. Jane Gilmour, Salam Assi, Ulrike Jaegle, Divine Kulu, Harmen van de Werken, Deborah Clarke, David Westhead, Sjaak Philipsen, and Constanze Bonifer. A crucial role for the ubiquitously expressed transcription factor sp1 at early stages of hematopoietic specification. *Development*, 141:2391–2401, 06 2014.

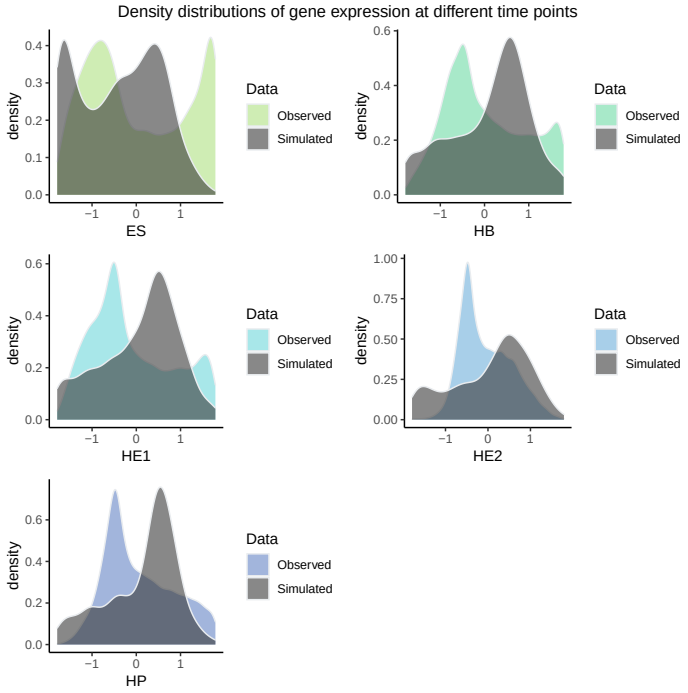


(a) Density distribution of the total expression of genes of the processed data

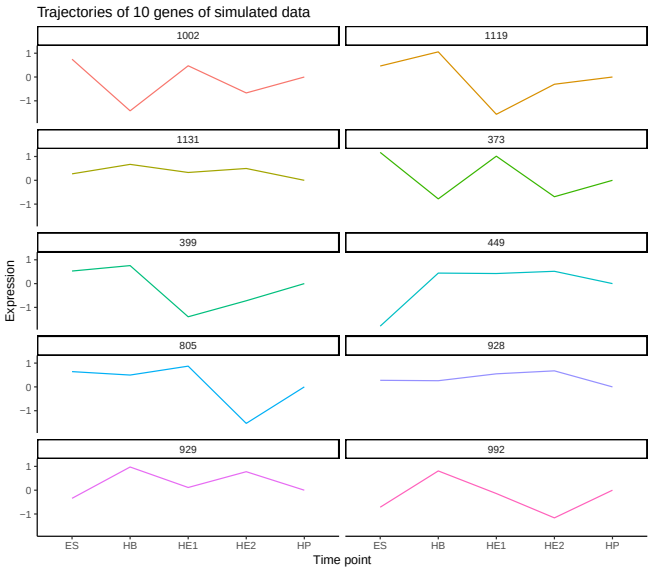


(b) Density distribution of the total expression of genes of the simulated data

Fig. 7. Comparison of the density distribution of total expression of genes of both data sets.

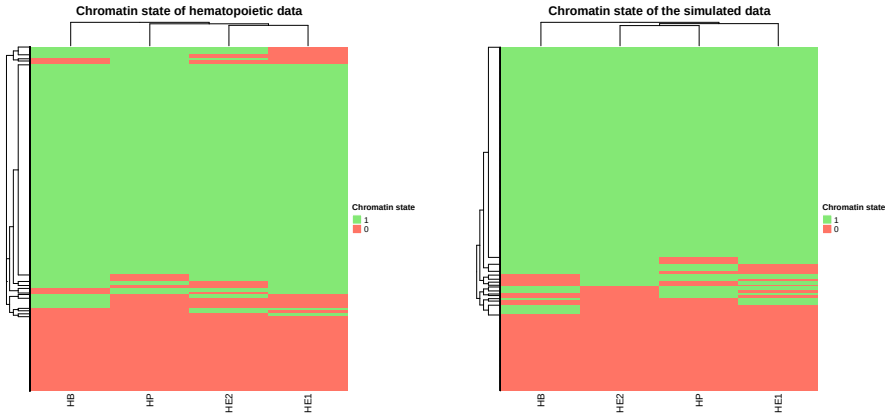


(a) Density distributions of the gene expression at the different cell stages of the observed and simulated data sets.



(b) Trajectories of 10 random genes across time points of the observed data set.

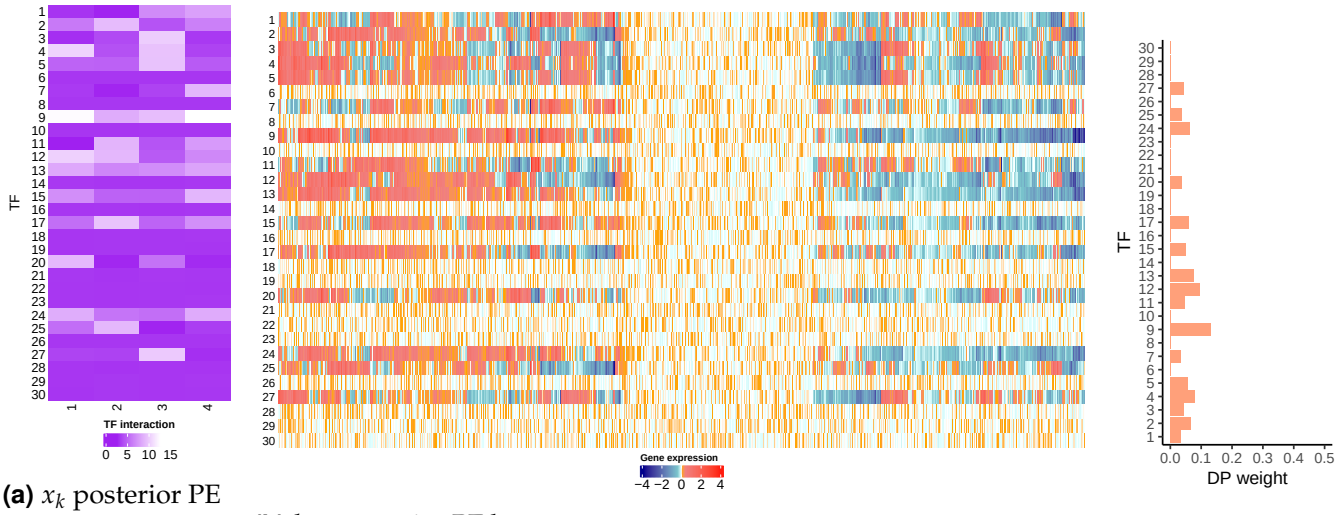
Fig. 8. These plots show the gene expression data of the synthenic data set. The first plot compares the density distributions of the gene expression at the different time points between our preprocessed and simulated data sets. The second one shows the trajectories of 10 random genes of the synthenic data, where it can be seen that all are different from each other.



(a) Observed chromatin state of genes at the different cell stages.

(b) Simulated chromatin state of genes at the different cell stages.

Fig. 9. Comparison of the heatmaps of the chromatin state of the different genes during the different time points for the simulated and observed data sets.

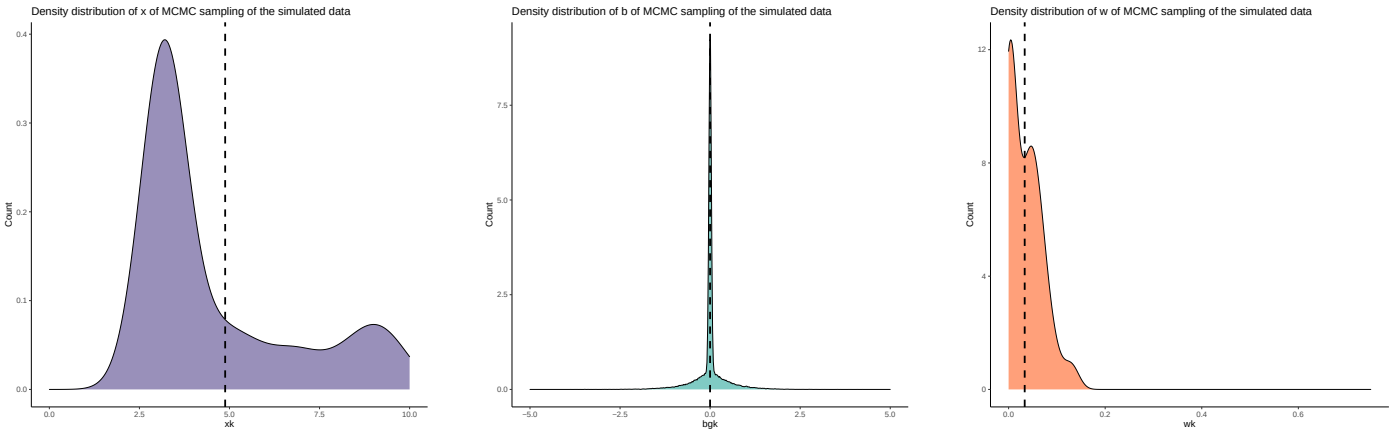


(a) x_k posterior PE heatmap

(b) $b_{g,k}$ posterior PE heatmap

(c) w_k posterior PE

Fig. 10. These plots show the posterior point estimation for x_k, $b_{g,k}$ and w_k of the simulated data set.



(a) Density distribution of x point estimate

(b) Density distribution of b point estimate

(c) Density distribution of w point estimate.

Fig. 11. Total density distributions of the posterior point estimates of w_k, $b_{g,k}$ and x_k

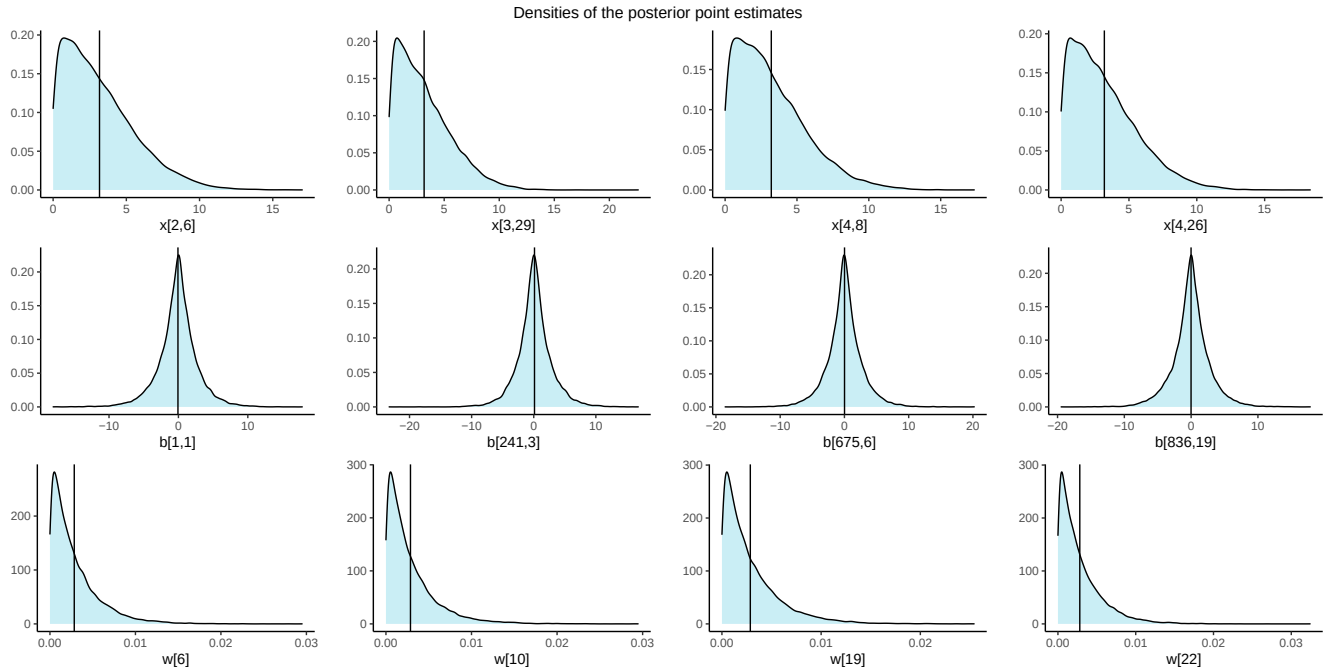
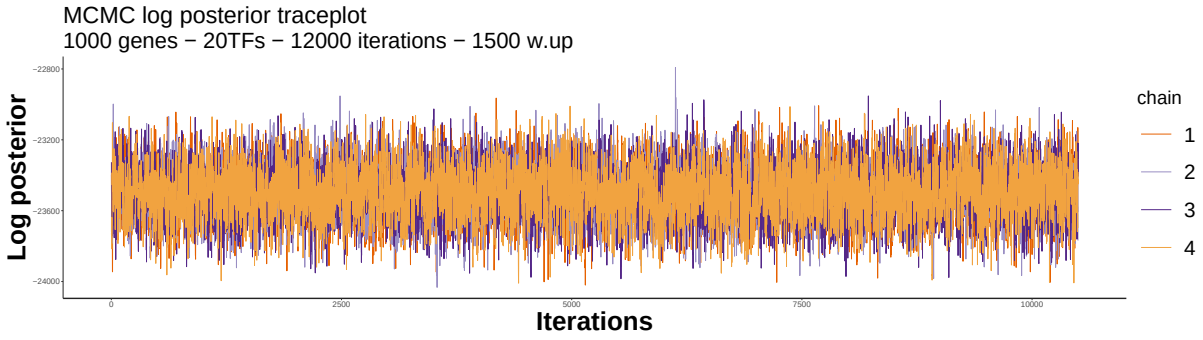
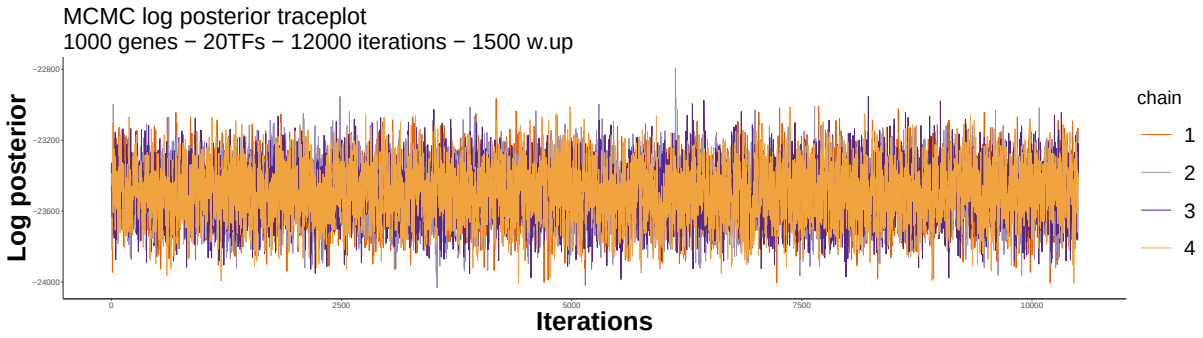


Fig. 12. Density estimates of random values of x_k, $b_{g,k}$ and w_k of the MCMC sampling.



(a) Traceplot of the MCMC log posterior of the model fitting of the WT hematopoietic data.



(b) Traceplot of the MCMC log posterior of the model fitting of the hematopoietic data with Sp1 perturbation.

Fig. 13. These plots show the parameter value at time t against the iteration number of the MCMC log posterior of the model fitting of the WT hematopoietic data and with Sp1 perturbation, where colors represent the different chains.

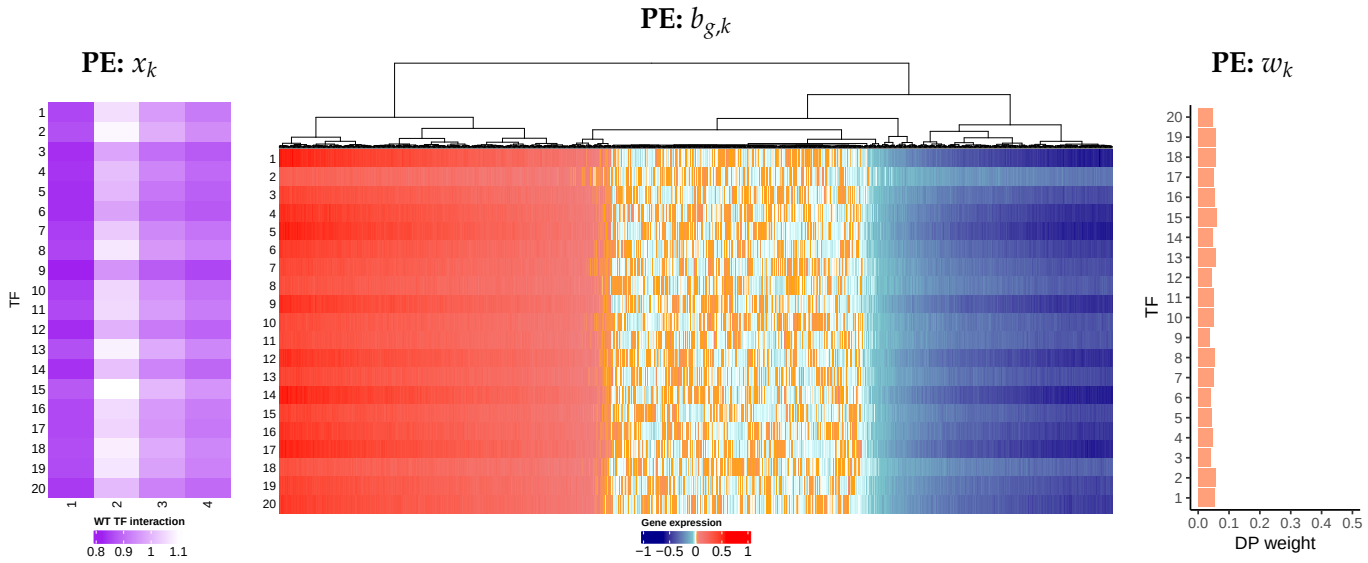


Fig. 14. These plots show the posterior point estimation for x_k, $b_{g,k}$ and w_k of the observed **Wild Type Sp1** hematopoietic data. It is a subset of 1000 genes whose gene expression standard deviation across time **> 0.5** .

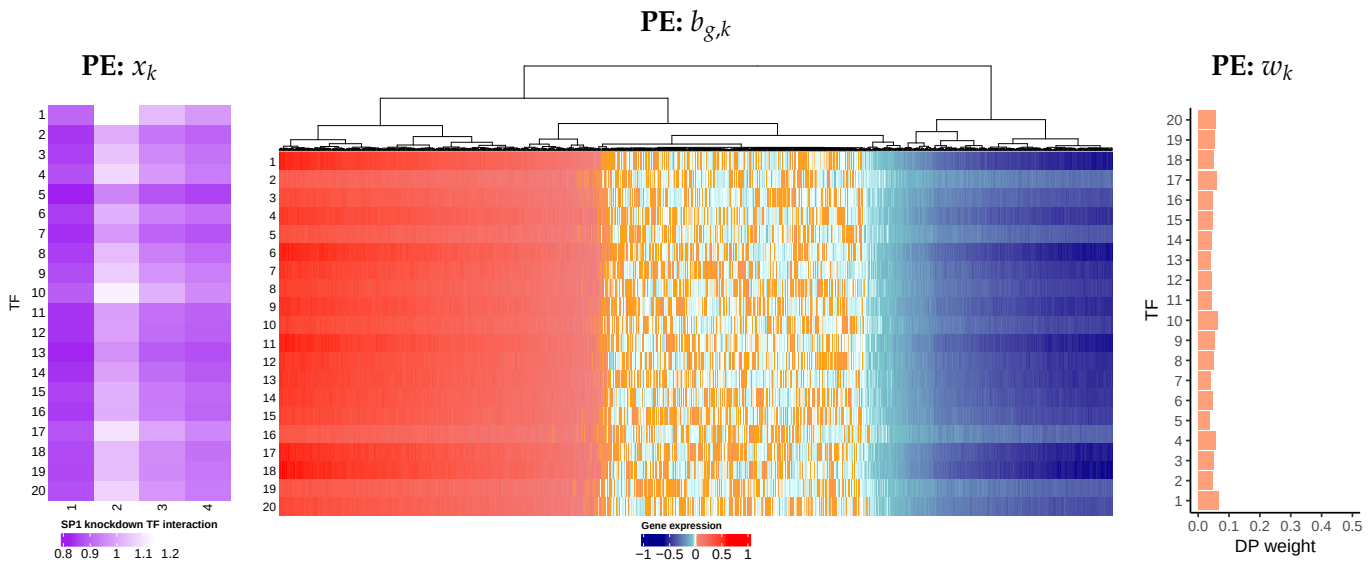


Fig. 15. These plots show the posterior point estimation for x_k, $b_{g,k}$ and w_k of the observed **Sp1 perturbation** hematopoietic data. It is a subset of 1000 genes whose gene expression standard deviation across time **> 0.5** .

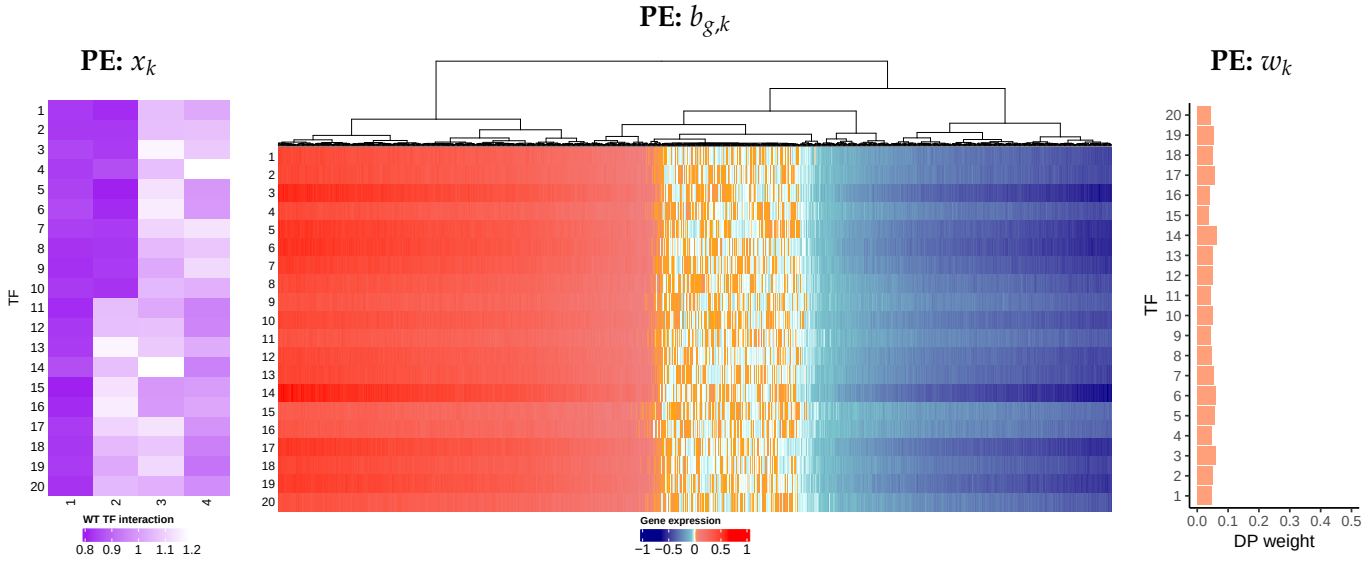


Fig. 16. These plots show the posterior point estimation for x_k, $b_{g,k}$ and w_k of the observed **Wild Type Sp1** hematopoietic data. It is a subset of 1000 genes whose gene expression standard deviation across time **> 2** .

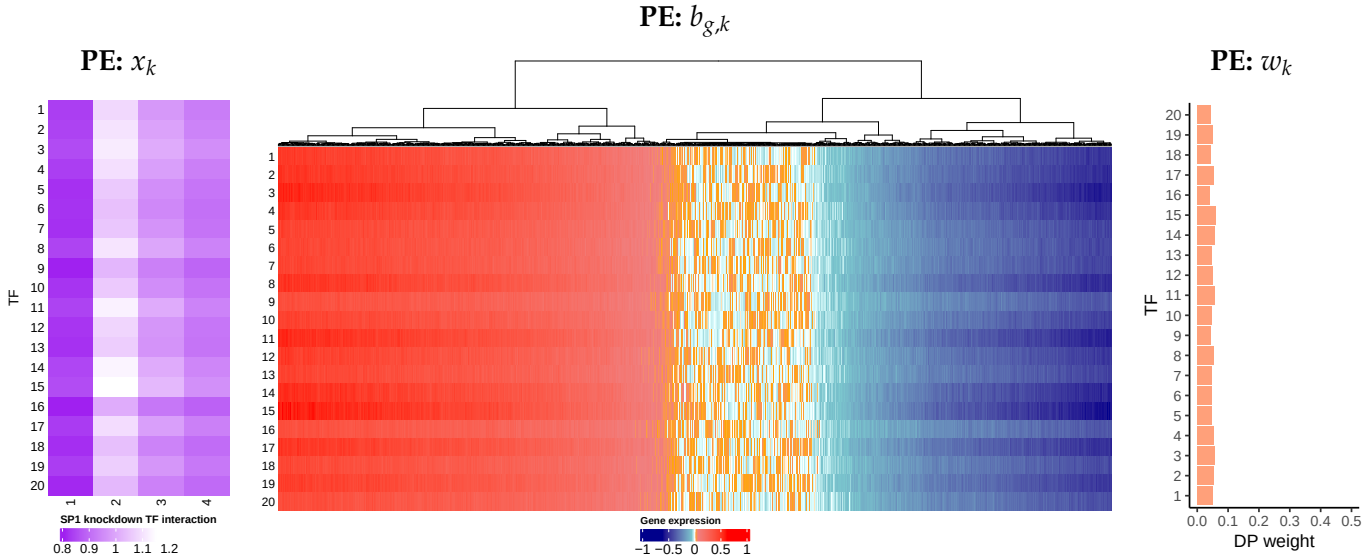


Fig. 17. These plots show the posterior point estimation for x_k, $b_{g,k}$ and w_k of the observed **Sp1 perturbation** hematopoietic data. It is a subset of 1000 genes whose gene expression standard deviation across time **> 2** .