Bioinformatics, 32(21), 2016, 3306–3313
doi: 10.1093/bioinformatics/btw395
Advance Access Publication Date: 10 July 2016
Original Paper

Systems biology

A subpopulation model to analyze heterogeneous cell differentiation dynamics

Yat Hin Chan[1,†], Jukka Intosalmi[1,†,*], Sini Rautio[1] and Harri Lähdesmäki[1,2,*]

[1]Department of Computer Science, Aalto University, 00076 Aalto, Finland and [2]Turku Centre for Biotechnology, University of Turku and Åbo Akademi, 20521 Turku, Finland

*To whom correspondence should be addressed.

Associate Editor: Janet Kelso

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Received on December 15, 2015; revised on April 27, 2016; accepted on June 17, 2016

Abstract

Motivation: Cell differentiation is steered by extracellular signals that activate a cell type specific transcriptional program. Molecular mechanisms that drive the differentiation can be analyzed by combining mathematical modeling with population average data. For standard mathematical models, the population average data is informative only if the measurements come from a homogeneous cell culture. In practice, however, the differentiation efficiencies are always imperfect. Consequently, cell cultures are inherently mixtures of several cell types, which have different molecular mechanisms and exhibit quantitatively different dynamics. There is an urgent need for data-driven mathematical modeling approaches that can detect possible heterogeneity and, further, recover the molecular mechanisms from heterogeneous data.

Results: We develop a novel method that models a heterogeneous population using homogeneous subpopulations that evolve in parallel. Different subpopulations can represent different cell types and each subpopulation can have cell type specific molecular mechanisms. We present statistical methodology that can be used to quantify the effect of heterogeneity and to infer the subpopulation specific molecular interactions. After a proof of principle study with simulated data, we apply our methodology to analyze the differentiation of human Th17 cells using time-course RNA sequencing data. We construct putative molecular networks driving the T cell activation and Th17 differentiation and allow the cell populations to be split into two subpopulations in the case of heterogeneous samples. Our analysis shows that the heterogeneity indeed has a statistically significant effect on observed dynamics and, furthermore, our statistical methodology can infer both the subpopulation specific molecular mechanisms and the effect of heterogeneity.

Availability and Implementation: An implementation of the method is available at http://research.ics.aalto.fi/csb/software/subpop/.

Contact: jukka.intosalmi@aalto.fi or harri.lahdesmaki@aalto.fi

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The capability of a cell to transform from one cell type to several more specialized cell types is crucial for the development of any multicellular organism. Besides developmental processes, cell specification is essential, e.g. for the immune system, particularly for its adaptive arm, where T and B cells respond and differentiate upon external signals. Cell differentiation processes are typically guided by extracellular signals which activate and control intracellular

mechanisms in a cell type specific way and, eventually, change the cell's gene expression pattern towards the pattern of a more specialized cell type. Given the central role of cell differentiation in biology, there is a keen interest to achieve a detailed understanding of the molecular mechanisms that drive differentiation processes and, especially, to understand the mechanisms that steer the activation of cell type specific regulatory programs. Among other approaches, mathematical modeling has been used to learn and predict molecular dynamics of cell differentiation processes (Intosalmi *et al.*, 2015; Schulz *et al.*, 2009).

Mathematical modeling can be especially useful when it is combined with time-course data by means of statistical techniques (Intosalmi *et al.*, 2015; Schulz *et al.*, 2009; Xu *et al.*, 2010). Using this approach, different hypotheses about the underlying molecular mechanisms and interactions are quantitatively expressed in the form of mathematical models and, further, the models can be objectively evaluated with respect to experimental data using well-defined statistical methods. In a typical setting, the models are informed using population average measurements, which are informative about the mechanisms of interest only if they come from a homogeneous cell culture. If the underlying cell culture turns out to be heterogeneous, the standard approaches can fail or provide unreliable results. In the context of cell differentiation, it is rather common that the cell culture of interest is a mixture of several cell types. For instance, if we consider a population of Type A cells which are either activated to Type B cells or stimulated to differentiate into Type C cells, it might be that only a fraction of the cells respond to the stimulation and the resulting cell culture is a mixture of Type B and C cells. When these kinds of data are analyzed, the possible heterogeneity needs to be taken into account properly.

The importance of heterogeneity modeling has been acknowledged also in other studies (Hasenauer *et al.*, 2011, 2014). In their recent study, Hasenauer *et al.* (2014) developed ordinary differential equation (ODE) constrained mixture models that can be used to detect dynamically distinct subpopulations using population snapshot data, such as FACS measurements. We also aim to detect dynamically distinct subpopulations and use ODEs to model molecular mechanisms. However, our approach is different in that we inform our models using population average sequencing data and, along with subpopulation detection, we explicitly infer the subpopulation specific regulatory mechanisms (i.e. network structures) from data. Further, we outline advanced statistical tools that can be used to quantitatively assess the significance of the modeling results.

In summary, we develop a novel modeling approach that can be applied to analyze cell differentiation dynamics in the presence of several co-existing cell types. We show how this approach can be used to detect the heterogeneity of the underlying cell culture and to infer cell type specific molecular interactions. We derive our model for naive human $CD4^+$ helper T (Th) cells that are induced to polarize towards Th17 lineage but, nevertheless, our approach is fully general. In our Th17 cell differentiation application, we construct alternative models for the core regulatory network driving the Th17 polarization in the form of ordinary differential equations and combine the alternative models with different hypotheses about the possible heterogeneity of underlying cell culture. To carry out quantitative model and parameter inference, we combine these models with time-course RNA sequencing (RNA-seq) data using a recently published statistical framework that is specifically designed for sequencing count data. Our results show that the proposed modeling approach works well for both simulated and experimental data. Further, to the best of our knowledge, we are the first ones to analyze the regulatory mechanisms that steer human Th17 lineage specification by means of data-driven mathematical modeling.

2 Materials and methods

2.1 Cell type specific subpopulations

We exemplify the cell type specific subpopulation model in the context of differentiation of naive human $CD4^+$T cells into Th17 lineage that can be induced by the cytokines transforming growth factor β (TGFβ), interleukin 6 (IL6) and interleukin 1 β (IL1β) (Korn *et al.*, 2009). These cytokines are crucial for the initiation of the differentiation process as well as for lineage maintenance (Korn *et al.*, 2009). In an ideal experimental setup, all naive $CD4^+$cells are equally exposed to these cytokine signals and go through the differentiation process as a homogeneous cell population as illustrated in Figure 1a. In practice, however, the differentiation efficiency can be notably lower than 100% as some fraction of cells typically responds only to the activation signal without actively going through the differentiation program (Fig. 1b). That is, a cell culture which is treated with the cytokines can thus consist of subpopulations of Th0 and Th17 cells. In model design, this can be taken into account by considering two cell types, Th0 cells that are activated but not exposed to the inducing cytokines and Th17 cells which are actively going through the differentiation program. Consequently, if we observe some intracellular factor x through population average data in the presence of heterogeneity, we actually observe a weighted average of the two subpopulations evolving in parallel. Formally, this can be expressed by writing

$$x_{\text{average}}(t) = (1-\alpha)x_{\text{Th0}}(t) + \alpha x_{\text{Th17}}(t) \tag{1}$$

where $x_{\text{Th0}}(t)$ and $x_{\text{Th17}}(t)$ are the abundances of the intracellular molecule in Th0 and Th17 subpopulations, respectively, at time t and $\alpha \in [0,1]$ is the fraction of cells in Th17 subpopulation. This subpopulation approach is very convenient as there is typically a control experiment in which the cells are activated in the absence of cytokines and, consequently, data on the pure Th0 cell population dynamics is directly available. Nevertheless, our method is fully general and such control experiments are not necessary for our method to work.

2.2 Mathematical models for Th17 and Th0 cell types

The most crucial transcriptional regulators driving the human Th17 cell differentiation are signal transducer and activator of transcription 3 (STAT3) and the retinoic acid receptor-related orphan receptor gamma t (RORγt) (Korn *et al.*, 2009). These two genes are known to exhibit strong dynamics during the early phase of

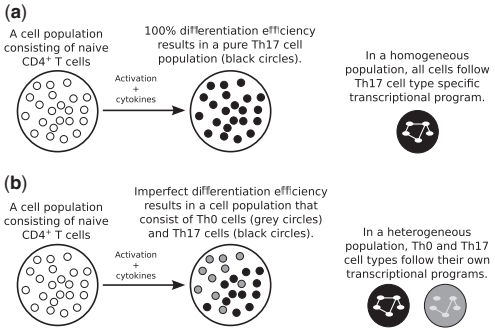

Fig. 1. Illustration of (**a**) homogeneous and (**b**) heterogeneous cell cultures going through Th17 cell differentiation process

polarization (Ciofani *et al.*, 2012; Yosef *et al.*, 2013) and, in the following, we derive a mechanistic description for the core Th17 network consisting of activation signaling, Th17 lineage specific cytokine signaling, as well as messenger RNA (mRNA) and protein levels for STAT3 and RORγt.

In the presence of activation, Th17 cell differentiation process is initiated by cytokine signals that induce STAT3 (Korn *et al.*, 2009). For modeling purposes, we assume that the joint effect of the activation and cytokine induction can be described using an input signal of the form

$$\xi(t,\theta_0,\theta_1,\theta_2) = \theta_0(1+\theta_1)\exp(-\theta_2 t), \tag{2}$$

which directly drives the STAT3 transcription. Here, θ_0 represents the activation effect, θ_1 is the relative amplification of the signal due to cytokines and θ_2 determines the time scale. When the input signal is combined with basal transcription, the possible autoactivation of STAT3 and the activation of STAT3 through RORγt derived signals, the STAT3 mRNA ($\mathrm{STAT3_{mRNA}}$) dynamics can be described using the differential equation

$$\frac{[\mathrm{STAT3_{mRNA}}]}{dt} = \xi(t,\theta_0,\theta_1,\theta_2) + \theta_3 + \theta_4[\mathrm{STAT3_{prot}}] + \theta_5[\mathrm{ROR\gamma t_{prot}}] - \theta_6[\mathrm{STAT3_{mRNA}}], \tag{3}$$

where θ_3, θ_4, θ_5 and θ_6 are unknown rate parameters (see Supplementary Table S1). We assume linear translation dynamics and describe the STAT3 protein ($\mathrm{STAT3_{prot}}$) level using the equation

$$\frac{[\mathrm{STAT3_{prot}}]}{dt} = \theta_7[\mathrm{STAT3_{mRNA}}] - \theta_8[\mathrm{STAT3_{prot}}], \tag{4}$$

where θ_7 and θ_8 are unknown rate parameters (see Supplementary Table S1). In our model, the RORγt activation is solely driven by STAT3 derived signals and RORγt autoactivation and, thus, the dynamics of RORγt mRNA and protein ($\mathrm{ROR\gamma t_{mRNA}}$ and $\mathrm{ROR\gamma t_{prot}}$, respectively) levels can be described using the differential equations

$$\frac{[\mathrm{ROR\gamma t_{mRNA}}]}{dt} = \theta_9[\mathrm{STAT3_{prot}}] + \theta_{10}[\mathrm{ROR\gamma t_{prot}}] - \theta_{11}[\mathrm{ROR\gamma t_{mRNA}}] \tag{5}$$

$$\frac{[\mathrm{ROR\gamma t_{prot}}]}{dt} = \theta_{12}[\mathrm{ROR\gamma t_{mRNA}}] - \theta_{13}[\mathrm{ROR\gamma t_{prot}}], \tag{6}$$

where θ_9, θ_{10}, θ_{11}, θ_{12} and θ_{13} are unknown rate parameters (see Supplementary Table S1). By combining the activation mechanisms and interactions introduced above, we obtain a putative core network driving the Th17 lineage specification and the corresponding ODE system describing the time-evolution of the components. Because we do not have prior information on the parameter values, we assume an uninformative, standard normal prior distribution in logarithmic scale for all parameters except the translation rates $\theta_7 = 1$ and $\theta_{12} = 1$ that are fixed. A schematic illustration of the core network is shown in Figure 2.

To model Th0 cells, which receive only the activation signal, we use the same mechanisms as for Th17 cells. However, we assume that Th0 cells are not capable of producing the active RORγt protein and, further, in these cells, the Th17 specific activating feedback mechanisms do not exist (dashed lines in Fig. 2). The regulatory mechanisms that are active in Th0 cells are highlighted in Figure 2 using black color. Another difference between the cell type specific models is that the input signal for Th0 cells is of the form $\xi(t,\theta_0,\theta_2) = \theta_0\exp(-\theta_2 t)$ because Th0 cells are assumed to be affected only by activation. Without losing any generality, the input signal can be scaled by a constant (40 in this study) in both cell types. The mechanisms that are active in both cell types are assumed to have the same rate parameters regardless of cell type. As a result, we have two parallel ODE systems

$$\dot{\mathbf{x}}_{M^{\mathrm{Th0}}}(t,\boldsymbol{\theta}) = f_{M^{\mathrm{Th0}}}(t,\mathbf{x}_{M^{\mathrm{Th0}}}(t,\boldsymbol{\theta})) \tag{7}$$

$$\dot{\mathbf{x}}_{M^{\mathrm{Th17}}}(t,\boldsymbol{\theta}) = f_{M^{\mathrm{Th17}}}(t,\mathbf{x}_{M^{\mathrm{Th17}}}(t,\boldsymbol{\theta})) \tag{8}$$

for Th0 and Th17 cell types, respectively. Here, $\mathbf{x}_{M^{\mathrm{Th0}}}(t,\boldsymbol{\theta})$ and $\mathbf{x}_{M^{\mathrm{Th17}}}(t,\boldsymbol{\theta})$ are the solutions of the systems, $\boldsymbol{\theta}$ is a vector containing the parameters and M^{Th0} and M^{Th17} specify the models (i.e. structures of the ODE systems). Due to the linearity of the systems, $\mathbf{x}_{M^{\mathrm{Th0}}}(t,\boldsymbol{\theta})$ and $\mathbf{x}_{M^{\mathrm{Th17}}}(t,\boldsymbol{\theta})$ can be solved using analytical expressions (for details, see Supplementary Material, Section 1).

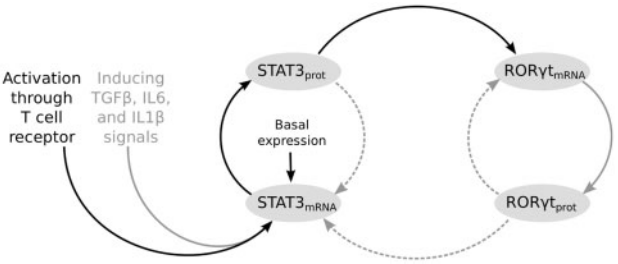

Fig. 2. Schematic illustration of putative Th17 core network. The black connectors represent mechanisms that are active in both Th0 and Th17 cell types. The gray connectors represent mechanisms that are active only in Th17 cell type. The dashed connectors denote the mechanisms that are considered when alternative models are constructed (Table 1)

Table 1. Alternative regulatory mechanisms for Th17 specific network

Mechanism	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8
RORγt autoactivation	–	–	–	–	×	×	×	×
STAT3 autoactivation	–	–	×	×	–	–	×	×
RORγt → STAT3	–	×	–	×	–	×	–	×

Symbols × and – denote active and inactive mechanisms, respectively.

2.3 Alternative models

We construct alternative models M^{Th17} by considering different regulatory mechanisms (also called wirings) for the Th17 cell type specific network. More precisely, we consider different combinations of the activating feedback mechanisms (dashed connectors in Fig. 2) and, as a result, we have altogether eight alternative wirings for the network (Table 1). The model for Th0 cell type (M^{Th0}) is kept fixed throughout this study. Further, in our application, the data contains independent replicates in two experimental conditions and we can form three biologically reasonable alternative hypotheses H considering the heterogeneity of the population; H_1: the cell culture treated with cytokines is pure and consists only of Th17 cells (Fig. 3, upper panel), H_2: the cell culture treated with cytokines is heterogeneous and fractions of subpopulations in independent experiments are the same (Fig. 3, middle panel), and H_3: the cell culture treated with cytokines is heterogeneous and fractions of subpopulations in independent experiments vary (Fig. 3, lower panel). The cell culture in control experiment (activation only) is always taken to be a pure culture of Th0 cells (Fig. 3, left column). Each of these hypotheses about the population structures results in different expressions for the observable average behavior.

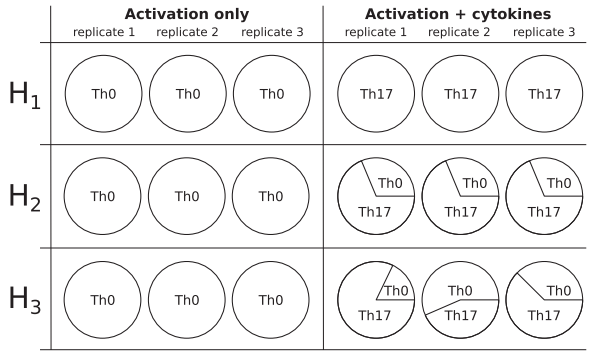

Fig. 3. Illustration of three hypothetical structures of the underlying cell culture. We assume that when cytokines are not added the resulting cell culture consists only of Th0 cells. In the presence of cytokines, the culture can be a homogeneous Th17 culture (H_1) or a mixture of Th0 and Th17 cell types (H_2 and H_3). In addition, the fractions of the cell types can be the same between independent cell cultures (H_2) or they may vary (H_3)

If we denote the predicted population average responses in control experiment (activation only) and the differentiation experiment (activation + cytokines) by $\boldsymbol{\phi}_k^0$ and $\boldsymbol{\phi}_k^1$, respectively, the hypothesis H_1 results in the expressions

$$\boldsymbol{\phi}_k^0\left(t,\boldsymbol{\theta},M^{\mathrm{Th0}}\right)=\mathbf{x}_{M^{\mathrm{Th0}}}(t,\boldsymbol{\theta}) \tag{9}$$

$$\boldsymbol{\phi}_k^1\left(t,\boldsymbol{\theta},M^{\mathrm{Th17}}\right)=\mathbf{x}_{M^{\mathrm{Th17}}}(t,\boldsymbol{\theta}), \tag{10}$$

which indicate that pure Th0 and Th17 populations are observed in all independent experiments ($k=1$, 2, 3 denotes the index of the replicate). On the other hand, under the hypothesis H_2 the expression for $\boldsymbol{\phi}_k^0$ remains the same but the cell population which is treated with cytokines consists of Th0 and Th17 subpopulations thus indicating

$$\boldsymbol{\phi}_k^1\left(t,\boldsymbol{\theta},M^{\mathrm{Th0}},M^{\mathrm{Th17}},\alpha\right)=(1-\alpha)\mathbf{x}_{M^{\mathrm{Th0}}}(t,\boldsymbol{\theta})+\alpha\mathbf{x}_{M^{\mathrm{Th17}}}(t,\boldsymbol{\theta}). \tag{11}$$

In this expression, the subpopulation fractions are determined by the replicate independent parameter α. Similarly, we can write the predicted average response under the hypothesis H_3 in the form

$$\boldsymbol{\phi}_k^1\left(t,\boldsymbol{\theta},M^{\mathrm{Th0}},M^{\mathrm{Th17}},\alpha\right)=(1-\alpha_k)\mathbf{x}_{M^{\mathrm{Th0}}}(t,\boldsymbol{\theta})+\alpha_k\mathbf{x}_{M^{\mathrm{Th17}}}(t,\boldsymbol{\theta}), \tag{12}$$

where $\boldsymbol{\alpha}=(\alpha_1,\alpha_2,\alpha_3)$ are the parameters representing the replicate dependent fractions of subpopulations. We treat α,α_1,α_2 and α_3 as free parameters with the uniform prior distribution on $[0,1]$.

For notational convenience, we merge α,α_1,α_2 and α_3 with $\boldsymbol{\theta}$ and denote the predicted population average responses in control and actual differentiation experiments by $\boldsymbol{\phi}_k^0(t,\boldsymbol{\theta})$ and $\boldsymbol{\phi}_k^1(t,\boldsymbol{\theta},M,H)$, respectively. Here, we denote $M=M^{\mathrm{Th17}}$ and omit M^{Th0} from the notation because it remains fixed. Further, we add the population structure hypothesis H explicitly into the arguments of $\boldsymbol{\phi}_k^1$. Collectively, when the hypotheses (H) regarding the heterogeneity of the underlying cell culture are combined with the alternative wirings of the Th17 core network (M), we can specify $8\times3=24$ alternative pairs (M, H) which represent our alternative models. We also note that the dimension of the parameter vector $\boldsymbol{\theta}$ depends on the pair (M, H) but we omit the subindex for clarity.

2.4 Statistical framework for RNA-seq data

To combine mathematical modeling with discrete, read count RNA-seq data, we use a recently published statistical framework which has been specifically designed for this data type (Intosalmi *et al.*, 2015). The framework is based on the negative binomial (NB) distribution which has been found particularly suitable for RNA-seq data (Robinson *et al.*, 2010).

We collect time-course RNA-seq data from both experimental conditions in three dimensional matrices D so that the element D_{ijk}, $(i=1,\ldots,n;\ j=1,\ldots,m;\ k=1,\ldots,l)$ represents the read count of gene i at the jth time point t_j in the kth replicate and use the NB distribution to model each element in the matrix. In other words,

$$D_{ijk}^e\sim\mathrm{NB}(C_{ijk}^e p_{ijk},\phi_{ij}), \tag{13}$$

where $C_{ijk}^e=L_iN_{jk}^e10^{-9}$ (L_i is the gene length, N_{jk}^e is the total number of mapped reads and e denotes the experimental condition), ϕ_{ij} is the gene specific, time-dependent dispersion parameter and p_{ijk} is the relative mRNA abundance of gene i at time t_j in the kth replicate. In this study, we make use of genome-wide data to obtain reliable estimates for the dispersion parameters and parameterize the sampling distribution accordingly. In general, however, if the genome-wide information is not available, the dispersion parameters can also be estimated along with the rate parameters using time-course data on model components. Further, we denote the data matrices corresponding to control and actual differentiation experiments by D^0 and D^1, respectively, and, thus, the likelihood of reproducing the data $D=\{D^0,D^1\}$ using the model (M, H) can be written in the form

$$\begin{aligned}p(D|\boldsymbol{\theta},M,H)=&\prod_{i=1}^{n}\prod_{j=1}^{m}\prod_{k=1}^{l}\mathrm{NB}(D_{ijk}^0;C_{ijk}^0\varphi_{ijk}^0(\boldsymbol{\theta}),\phi_{ij})\\&\times\prod_{i=1}^{n}\prod_{j=1}^{m}\prod_{k=1}^{l}\mathrm{NB}(D_{ijk}^1;C_{ijk}^1\varphi_{ijk}^1(\boldsymbol{\theta},M,H),\phi_{ij}),\end{aligned} \tag{14}$$

where $\varphi_{ijk}^0(\boldsymbol{\theta})$ is the ith component of $\boldsymbol{\phi}_k^0(t_j,\boldsymbol{\theta})$ and $\varphi_{ijk}^1(\boldsymbol{\theta},M,H)$ is the ith component of $\boldsymbol{\phi}_k^1(t_j,\boldsymbol{\theta},M,H)$ (confer, Intosalmi *et al.*, 2015). According to Bayes' theorem, the parameter posterior of model (M, H) can now be expressed in the form $p(\boldsymbol{\theta}|D,M,H)\propto p(D|\boldsymbol{\theta},M,H)p(\boldsymbol{\theta}|M,H)$, where $p(\boldsymbol{\theta}|M,H)$ is the prior distribution for the parameters of model (M, H) (for details about Bayesian methodology, see e.g. Gelman *et al.*, 2013). Further, the posterior distribution over alternative models can be written in the form $p(M,H|D)\propto p(D|M,H)p(M,H)$, where $p(M,H)$ is a prior distribution over alternative models (the uniform distribution in this study) and

$$p(D|M,H)=\int_{\boldsymbol{\theta}}p(D|\boldsymbol{\theta},M,H)p(\boldsymbol{\theta}|M,H)d\boldsymbol{\theta}, \tag{15}$$

is the marginal likelihood. In the context of mathematical models, the marginal likelihood can only rarely be solved analytically and different kinds of approximative and numerical approaches need to be used (Vyshemirsky and Girolami, 2008). The posterior predictive distribution over model dynamics can be estimated using the average library size and an interpolated version of time-dependent dispersion parameters as described in Intosalmi *et al.* (2015).

2.5 Population-based Markov chain Monte Carlo sampling and thermodynamic integration

The posterior distributions of the models derived above might end up being multimodal and contain complex dependencies, and it might be problematic to carry out the posterior analysis using standard Markov chain Monte Carlo (MCMC) methods. Consequently, we utilize the population-based MCMC sampling which is known

to perform well even with complex target distributions (Jasra *et al.*, 2007). Population-based MCMC sampler can be constructed by considering a product form of the target density

$$p^*(\boldsymbol{\theta}_{\beta_1}, \boldsymbol{\theta}_{\beta_2}, \ldots, \boldsymbol{\theta}_{\beta_{N_\beta}} | D, M, H) = \prod_{i=1}^{N_\beta} p_{\beta_i}(\boldsymbol{\theta}_{\beta_i} | D, M, H), \quad (16)$$

where $p_{\beta_i}(\boldsymbol{\theta}|D, M, H) \propto p(D|\boldsymbol{\theta}, M, H)^{\beta_i} p(\boldsymbol{\theta}|M, H)$ is the power posterior for fixed $0 = \beta_1 < \ldots < \beta_{N_\beta} = 1$ (confer, e.g. Calderhead and Girolami, 2009). The distributions p_{β_i}, including the posterior distribution $p(D|\boldsymbol{\theta}, M, H)p(\boldsymbol{\theta}|M, H)$, are marginal distributions of the product form of the target density and, by means of population-based MCMC sampling, we draw samples from each of these distributions in parallel. The sampling in each individual distribution p_{β_i} can be carried out by using standard MCMC techniques and, besides the local exploration of the distributions, the population-based MCMC sampling allows global moves between the distributions which notably improves the mixing properties of the chain. Further, the samples obtained from the population-based MCMC sampler can be directly used to estimate the marginal likelihood via thermodynamic integration (Friel and Pettitt, 2008; Friel *et al.*, 2014). The derivation of the thermodynamic integration can be found in Supplementary Material.

2.6 RNA-seq data and data pre-processing

The samples were prepared as previously described in Tuomela *et al.* (2016). In brief, the samples originate from $CD4^+$ cells that were isolated from human umbilical cord blood. To induce Th17 polarization, the naive cells were activated and simultaneously treated with IL6, IL1β and TGFβ in the presence of neutralizing anti-interferon γ and anti-IL4. For the control experiment, the cells were activated and cultured in the presence of neutralizing antibodies. In both experiments, the samples were collected in triplicates at the indicated time-points and the expression profiles were quantified by means of RNA sequencing. The data used in this study can be found in the Supplementary Table S2.

Sequence reads were mapped using Tophat (version 1.3.2) with default parameters to the GRCh37 human reference genome and Ensembl human transcriptome (release 63). Expression values of Ensembl genes were calculated using Python package HTSeq (Anders *et al.*, 2015) (version 0.5.3p3) with parameters '–type = exon –idattr = gene_id –stranded = no'. Bioconductor package edgeR (Robinson *et al.*, 2010) was used to estimate the time-dependent dispersion parameters (Supplementary Table S3) and the relative mRNA abundances were presented using reads per kilo base per million (RPKM) values. The data is accessible through GEO Series accession number GSE52260.

2.7 Computational implementation

The mathematical models and the sampling algorithm were implemented in Matlab (The MathWorks Inc., Natick, MA, USA). The initial values for mRNA levels are taken directly from data (under hypotheses H_1 and H_2, we use the average expression levels). The abundances of active proteins are assumed to be negligible in the beginning of the experiment. Population-based MCMC sampling was carried out using 10 different temperatures that were specified by $\beta_i = ((i-1)/(10-1))^5$, $i = 1, \ldots, 10$ [confer, e.g. Calderhead and Girolami (2009)]. Before the actual run, the proposal distributions were tuned adaptively. After the adaption and burn-in period, the sampler was run using fixed proposal distributions and every 1000th sample was collected until 1000 independent samples were obtained. This procedure was repeated at least three times for each model and, based on the resulting independent MCMC chains, the convergence was monitored using the potential scale reduction factor (Gelman *et al.*, 2013). Posterior analysis for each model is based on at least 3000 independent samples.

3 Results

3.1 Benchmarking the method using simulated data

In order to test the feasibility of our modeling approach and to evaluate the performance of our sampling algorithm, we run the inference first for simulated data. We generate the data using the wiring models $M_1, \ldots, M_8$ in combination with all three alternative hypotheses about the structure of the underlying population (H_1, H_2, and H_3). We simulate two realizations from each model and, as a result, obtain $2 \times 24 = 48$ simulated datasets. Each dataset consists of the same number of replicates and time-points in two different conditions as our experimental data and, in addition, we use the dispersion levels obtained from the real RNA-seq data to set the variation in the simulated datasets at a realistic level. For details about data simulation and visualizations of simulated datasets, see Supplementary Material (Section 3 and Fig. S1). For each of the 48 datasets, we carry out posterior analysis over all 24 models, and conclude that, in general, the posterior model distributions show high probabilities for the models that were used to simulated the data (Supplementary Fig. S2).

Even though the results are in a reasonably good agreement with the ground truth, the posterior model probabilities clearly vary from one realization to another. To assess how strong this variability is, we carry out a more extensive testing of our method's model ranking performance by considering a subset of the wiring models (i.e. the wiring models $M_5, \ldots, M_8$) in combination with all three hypotheses H_1, H_2 and H_3. Here, only a subset of models is considered to keep the computational burden of the experiment feasible. We simulate ten independent realizations from each model and carry out posterior analysis for all $10 \times 12 = 120$ simulated datasets. The resulting posterior model distributions are summarized in Figure 4. The posterior analysis results show robust model ranking performance in the case of models M_5 and M_7 in combination with all three hypotheses H_1, H_2 and H_3 (Fig. 4, the first and third row). When models M_6 and M_8 are considered, the posterior model probabilities seem to depend more on the underlying realization and there is more scatter in the distributions (Fig. 4, the second and fourth row). However, also in these cases a notable amount of probability mass is concentrated on the correct model and, consequently, the inference is likely to provide useful information in a wide range of practical settings.

An important aspect in our experiments with simulated data is that we simulate the data using fixed parameters for all mechanisms of interest and once a mechanism is removed from the model, we simply remove the corresponding parameter. The model ranking results may, at least to some extent, depend on the selected parameter values. To assess how sensitive our results are for the selection of the population fraction parameters, we simulate some further data using the well behaving model M_5 under the hypotheses H_2 and H_3 so that the population fraction parameters for each realization are drawn from the uniform distribution. We generate ten realizations using each combination (M_5H_2 or M_5H_3) and carry out posterior analysis for these data. The posterior analysis results are summarized in Supplementary Figure S3 and show that the model ranking performance is not sensitive to the population fraction parameters.

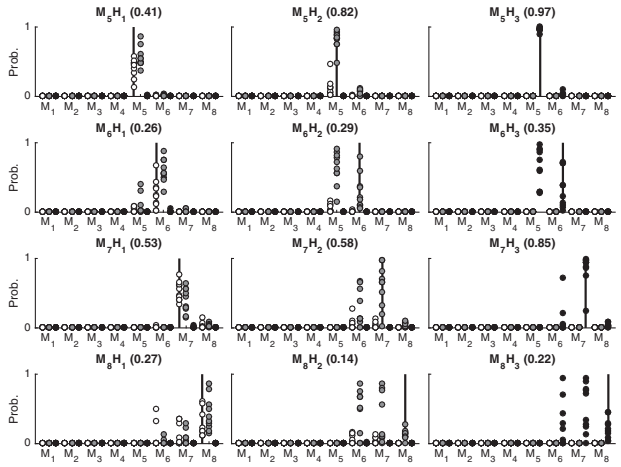

Fig. 4. Model discrimination results for 120 simulated datasets that are generated using the wiring models M_5, M_6, M_7, M_8 in combination with three population structure hypotheses. The model which is used to simulate the data is given on top of each subplot (indicated also by the vertical line) and the average posterior probability of the correct model is given in parentheses. The posterior probabilities of alternative models are plotted using circles and the fill color of the circle indicates the underlying population structure hypothesis (H_1, H_2 and H_3 are color coded using white, gray and black circles, respectively)

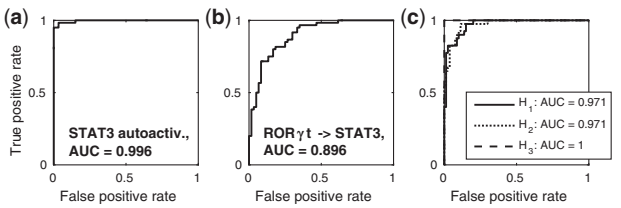

Fig. 5. ROC curves summarizing the method's performance in recovering specific mechanisms and population structure hypotheses from simulated data. The curves are computed based on the 120 simulated datasets for which the posterior model distributions are illustrated in Figure 4. The ROC curves are computed for recovering (**a**) STAT3 autoactivation, (**b**) RORγt → STAT3 and (**c**) the underlying population structure. The performance is also summarized by computing the area under a ROC curve (AUC)

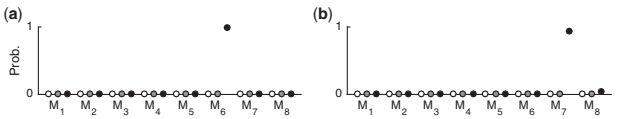

Fig. 6. Model discrimination results for the RNA-seq data. (**a**) The estimated posterior distribution over all combinations of wiring models and population structure hypotheses. (**b**) The same as in (a) but the RNA-seq data is complemented with additional qualitative data. The color coding for hypotheses H_1, H_2 and H_3 is the same as in Figure 4. The logarithmic marginal likelihood values that have been used to compute the posterior probabilities as well as convergence diagnostics of these runs can be found in Supplementary Table S4 and Figures S6–S7

Along with the analysis of the model posterior distributions, we study how well the proposed method performs in recovering specific mechanisms or population structure from simulated data. For this purpose, we compute the receiver operating characteristic (ROC) curves with respect to methods capability of discriminating if a mechanism or population structure is present or not. The ROC analysis is carried out using the simulated 120 datasets introduced above (for details about the ROC curve computation, see Supplementary Material, Section 4). The results of ROC analysis are summarized in Figure 5. The recovery of STAT3 autoactivation is almost perfect (Fig. 5a) and the activation of STAT3 by RORγt can be recovered successfully in most cases (Fig. 5b). Our method also recovers the underlying population structure successfully (Fig. 5c).

In addition to the good performance in the model discrimination tests, our method provides us with the estimates of the parameter posterior distributions. In all experiments, the parameter posterior distributions are notably updated from the prior distributions which indicate good model identifiability (for an example, see, Supplementary Fig. S4). Further, successful posterior analysis can be carried out even in the case of challenging posterior parameter distributions that exhibit multimodality and include complex dependencies between the parameters (for an example, see, Supplementary Fig. S5). The good sampling properties originate from the use of the population-based MCMC approach which is known to perform well even with complex distributions.

3.2 Analysis of Th17 RNA-seq data reveals heterogeneity and the most likely core network

We initiate the analysis of real RNA-seq data by estimating the time-dependent dispersion parameters (Supplementary Table S3). The resulting gene-wise dispersion values are very similar to the common dispersion values (data not shown) and, consequently, we can use the common dispersion in our inference without causing noticeable bias.

We run the inference for the real data using all eight alternative models for the network wiring in combination with the three alternative heterogeneity scenarios about the cell population. The estimated posterior distribution over the alternative models values are shown in Figure 6a. The wiring model M6 combined with the hypothesis H3 has clearly the highest evidence amongst the alternative scenarios (the estimated parameter posterior is illustrated in Supplementary Fig. S8). This suggests that the underlying cell culture is indeed heterogeneous and the fractions of cells in different cell types vary between independent experiments. Further, the result suggests that RORγt autoactivation and STAT3 activation through RORγt derived signals are actively affecting the differentiation process. On the other hand, these findings suggest that STAT3 autoactivation is not necessarily needed.

The above listed mechanisms inferred by our analysis have been discussed also in the recent literature related to murine Th17 lineage specification (see, for instance, Ciofani *et al.*, 2012). Autocrine mechanisms that result in positive feedback loops around RORγt have been reported in several studies (Murphy and Stockinger, 2010) and, thus, our findings based on human data are in agreement with these results. The inferred STAT3 activation through RORγt derived signals is a more controversial finding. In fact, a recent study shows that this kind of feedback mechanisms from RORγt to STAT3 should not exist, at least when mouse system is considered (Ciofani *et al.*, 2012). Further, mouse studies support the view that there should be many RORγt independent feedback mechanisms regulating the STAT3 expression and, as explained above, our inference does not support the existence of this kind of feedbacks (in our models, these feedback mechanisms are implicitly modeled through the STAT3 autoactivation). Our conclusion here is that our RNA-seq data alone is not sufficient to draw persuasive conclusions about the STAT3 regulation by RORγt.

To perform a more comprehensive inference about the wiring mechanisms, we complement the time-course RNA-seq data with qualitative information from murine studies. By means of RORγt

knock-out experiments, Ciofani *et al.* (2012) have shown that STAT3 expression level at time 48 h does not depend on RORγt expression (even though RORγt binds the vicinity of STAT3 gene). We incorporate this qualitative information into our likelihood function by requiring that the difference between STAT3 levels at time 48 h in wild-type and RORγt knock-out experiments follows a normal distribution with a relatively small variance (for details, see Supplementary Material). The whole inference is then repeated for all 24 alternative scenarios using this extended setting.

Even though the qualitative information is introduced only at one time point (48 h), it has a notable effect on the posterior probabilities over alternative scenarios (Fig. 6b). When the qualitative information is added, the inference still supports the population structure hypothesis H_3 but prefers the wiring model M_7 (the estimated parameter posterior is illustrated in Supplementary Fig. S9). The wiring model M_7 includes exactly the mechanisms that are reported in murine studies, that is, RORγt and STAT3 autoactivation but, on the other hand, does not support the controversial feedback from RORγt to STAT3. Our view is that the extended inference produces reasonable predictions that are in agreement with the existing results from murine studies and could be relevant also in the context of human system.

The time-course data as well as the posterior predictive distribution of the best ranking model (M_7,H_3) are illustrated in Figure 7. All in all, the predictive distributions are in a good agreement with the data. The estimated means of the parameters $\alpha_i, i = 1, 2, 3$, in the model (M_7,H_3) indicate that, on average, the subpopulation fractions in the three independent experiments are 84, 94 and 87%. It is noteworthy that the variation of the fractions of subpopulations differs between the replicates and this explains why evidence for H_3 is higher than the evidence for H_2.

4 Discussion and conclusions

Understanding dynamic gene regulatory mechanisms is a central challenge in the field of systems biology. Mathematical modeling inevitably has an essential role in this endeavor. In this study, we introduce a general methodology for constructing mechanistic models that are based on subpopulations which evolve in parallel. The subpopulations represent distinct cell types that are steered by different molecular mechanisms and exhibit different dynamic behaviors within a heterogeneous cell population. We demonstrate how the derived models can be calibrated in a data-driven manner and, most importantly, show how the significance of the modeling results can be assessed by means of rigorous statistical testing. Our integration of subpopulation modeling with population average data is unique and complements well the existing models for population snapshot data (Hasenauer *et al.*, 2011, 2014). In general, we believe that taking possible heterogeneity of experimental samples into account can be advantageous in many applications.

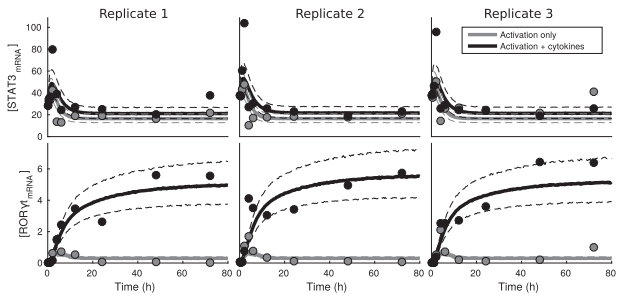

Fig. 7. Marginal posterior predictive distributions for model structure (M_7,H_3) plotted against the experimental data from the control experiment (activation, gray circles) and the actual Th17 differentiation experiment (activation + cytokines, black circles). The solid lines are the medians and the dashed lines show the 5 and 95% quantiles of the distribution. The posterior predictive distributions for all alternative wiring models under the hypothesis H_3 are shown in Supplementary Figure S10

The computational implementation that we present for our method relies heavily on advanced Markov chain Monte Carlo (MCMC) techniques. More specifically, we use population-based MCMC sampling (Jasra *et al.*, 2007) to carry out posterior analysis and estimate marginal likelihoods using thermodynamic integration (Friel and Pettitt, 2008). Excellent performance of our statistical implementation is for a great part due to these methods that can handle also complex posterior distributions. Due to the good performance of the sampling algorithm, our approach presumably also scales up to moderate sized ODE models if a feasible amount of data is available.

We apply our method to RNA-seq data that consists of activated $CD4^+$ cells and cells polarizing towards Th17 lineage to analyze the regulatory interactions between the core genes driving the Th17 cell differentiation. The interactions of these core genes have been studied by means of mechanistic models also in earlier studies in the context of murine data (Intosalmi *et al.*, 2015). However, to the best of our knowledge, the analysis that we present here is the first one that has been carried out for human data. Our results are in a good agreement with the current understanding about the molecular underpinnings of the Th17 regulatory mechanisms. Further, the predictions that we provide about the possible regulatory network topologies may turn out to be useful in future studies.

Like mentioned earlier, the applicability of our method is not restricted to the application presented in this study. As a matter of fact, the method is fully general and can be extended to cover also more than two subpopulations. In T cell biology, there are numerous interesting applications that could benefit from the kind of modeling we present. For instance, the differentiation from naive $CD4^+$ cells to the Th17 and regulatory T (Treg) cell lineages is reciprocal in nature (Bettelli *et al.*, 2006) and it would be natural to construct a model consisting of Th0, Th17 and Treg subpopulations and to study the regulatory mechanisms guiding the differentiation into these lineages. However, it is noteworthy that the amount of required data increases when the number of subpopulations becomes greater and, consequently, the models need to be carefully benchmarked using simulated data prior to application to real data. In a similar manner, our method could be used to analyze regulatory mechanisms of other subsets of T helper cells such as the subsets of Th1 and Th2 cells. One further strength of our approach is that it allows also the modeling of possible cross-talk between cell types which is central, for instance, for achieving a proper balance between T cell subsets and, thereby, proper regulation of the immune response.

In summary, we present a novel modeling approach that can be used to analyze cell differentiation processes within heterogeneous cell populations and show how it can be applied in practice. We test the approach first using simulated data and then apply it to analyze the gene regulatory network steering the Th17 lineage specification.

Acknowledgements

The authors acknowledge the computational resources provided by the Aalto Science-IT project and thank Prof. Riitta Lahesmaa and her research group at Turku Centre for Biotechnology, Turku, Finland for providing the experimental data and Henrik Mannerström for helpful discussions.

Funding

This work has been supported by the Academy of Finland [Centre of Excellence in Molecular Systems Immunology and Physiology Research (2012–2017) as well as the project 275537], EU FP7 grant [EC-FP7-SYBILLA-201106], EU ERASysBio ERA-NET.

References

Anders,S. *et al*. (2015) HTSeq – a Python framework to work with high-throughput sequencing data. *Bioinformatics*, **31**, 166–169.

Bettelli,E. *et al*. (2006) Reciprocal developmental pathways for the generation of pathogenic effector TH17 and regulatory T cells. *Nature*, **441**, 235–238.

Calderhead,B. and Girolami,M. (2009) Estimating Bayes factors via thermodynamic integration and population MCMC. *Comput. Stat. Data Anal.*, **53**, 4028–4045.

Ciofani,M. *et al*. (2012) A validated regulatory network for Th17 cell specification. *Cell*, **151**, 289–303.

Friel,N. *et al*. (2014) Improving power posterior estimation of statistical evidence. *Stat. Comput.*, **24**, 709–723.

Friel,N. and Pettitt,A.N. (2008) Marginal likelihood estimation via power posteriors. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **70**, 589–607.

Gelman,A. *et al*. (2013). *Bayesian Data Analysis*. 3rd edn. Chapman & Hall/CRC Texts in Statistical Science, Boca Raton, FL.

Hasenauer,J. *et al*. (2011) Identification of models of heterogeneous cell populations from population snapshot data. *BMC Bioinformatics*, **12**, 125.

Hasenauer,J. *et al*. (2014) ODE constrained mixture modelling: a method for unraveling subpopulation structures and dynamics. *PLoS Comput. Biol.*, **10**, e1003686.

Intosalmi,J. *et al*. (2015) Analyzing Th17 cell differentiation dynamics using a novel integrative modeling framework for time-course RNA sequencing data. *BMC Syst. Biol.*, **9**, 81.

Jasra,A. *et al*. (2007) On population-based simulation for static inference. *Stat. Comput.*, **17**, 263–279.

Korn,T. *et al*. (2009) IL-17 and Th17 Cells. *Annu. Rev. Immunol.*, **27**, 485–517.

Murphy,K.M. and Stockinger,B. (2010) Effector T cell plasticity: flexibility in the face of changing circumstances. *Nat. Immunol.*, **11**, 674–680.

Robinson,M. *et al*. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Schulz,E.G. *et al*. (2009) Sequential polarization and imprinting of type 1 T helper lymphocytes by interferon-gamma and interleukin-12. *Immunity*, **30**, 673–683.

Tuomela,S. *et al*. (2016) Comparative analysis of human and mouse transcriptomes of Th17 cell priming. *Oncotarget*, **7**, 13416–13428.

Vyshemirsky,V. and Girolami,M.A. (2008) Bayesian ranking of biochemical system models. *Bioinformatics*, **24**, 833–839.

Xu,T.R.R. *et al*. (2010) Inferring signaling pathway topologies from multiple perturbation measurements of specific biochemical species. *Sci. Sig.*, **3**, 134.

Yosef,N. *et al*. (2013) Dynamic regulatory network controlling TH17 cell differentiation. *Nature*, **496**, 461–468.