博士論文

EM algorithms for optimization of Wright Fisher model and reconstruction of spatiotemporal gene expression patterns

(EM アルゴリズムによる Wright-Fisher モデルの最適化と遺伝子発現の時空間再構築)

小嶋　泰弘

Abstract

Probabilistic models play important roles in extracting knowledge from biological data. In particular, a likelihood of the probabilistic models given the data are maximized over the parameters of the models, and the optimized parameters are regarded as some quantity reflecting the biological processes behind the data. For example, Sandberg *et al.* estimated the kinetics parameters of transcriptional regulation from single cell RNA-seq (scRNA-seq) data utilizing this methodology. However, the optimization of the probabilistic models often faces an obstacle that some variables constituting the models are unobserved. EM algorithms overcame this obstacle in many situations of the biology by marginalizing the log likelihood over the unobserved variables. In this thesis, we describe two studies where EM algorithms are utilized for efficient optimization of a mathematical model of population genetics, Wright-Fisher model (WF), and the reconstruction of time-continuous spatiotemporal gene expression patterns during embryogenesis.

Estimation of population genetic parameters using an EM algorithm and sequence data from experimental evolution populations

Introduction

Lower cost of genome sequencing has enabled us to derive time course sequence data, which provides us the genome wide dynamics of biological processes. In particular, a new experimental method, Evolve and Resequence, provides us time course allele frequency data of numerous single nucleotide polymorphisms (SNPs) during an experimental evolution induced by artificial selective pressure on specific phenotype. Many of researches utilizing this method mainly focused on detecting SNPs which are related to the phenotype selected in the experimental evolution. On the other hand, time course of allele frequencies has been extensively investigated by mathematical models in the field of population genetics. In particular, the dynamics of allele frequencies is characterized by the strength

and dominance of selection in WF model. These parameters optimized for E&R data are expected to help us derive new insight for the evolutionary dynamics behind the experimental evolution, since they provide us quantitative information of the evolutionary selection across numerous SNPs. However, it is difficult to estimate these parameters from E&R data due to the huge number of SNPs. Recently, several methods are developed for the optimization of WF models to E&R data by improving the efficiency of the estimation, while these methods can not evaluate the confidence of the estimation. Since the simultaneous estimation of s and h is often intrinsically unidentifiable, the estimates which is far away from true values can affect the post analysis of the estimated values. Here, we proposed a new method to efficiently estimate WF parameters from E&R data, utilizing an EM algorithm for continuous time Markov chain Furthermore, our method evaluates estimation confidence by calculating confidence intervals based on empirical Fisher information matrix.

Results and discussion

First, we validated the estimation of our method and compare it with other existing methods using simulated ER data. We found that the values estimated by our method distributed around true values used in simulation. On the other hand, we confirmed that the filtering process based on estimated confidence intervals selectively excludes the estimated values which is far away from true values. In our comparison with other methods, dominance estimation of our method was the most accurate after the filtering process, while the estimation time of our method was about one third of that of the most efficient other method. Next, we applied our method to a real E&R data where Drosophila population was bred in a new thermal condition, which is expected to induce selective pressure on phenotype related to thermal adaptation. As a result, we found that there was a specific peak in the distribution of the selection strength parameters of SNPs within a cosmopolitan inversion, In(3R)P, region, which suggested that many SNPs within In(3R)P region share the common allele frequency dynamics during this experimental evolution. On the other hand, the estimated dominance parameters are distributed around 1, which indicated that many of deleterious alleles in this experiment are recessive. This observation is consistent with the observation that many of deleterious alleles are recessive in natural population of Arabidopsis.

Reconstructing spatiotemporal gene expression patterns during embryogenesis

Introduction

Embryogenesis produces a specific shape of each species from a single fertilized egg by repeated cell division and cell movement. To achieve this complicated process, the behavior of cells in each spatiotemporal context must be accurately specified. One large element determining the cell behavior is the expression of numerous genes that reaches tens of thousands. Hence, the localization of gene expression has been an important clue for revealing the mechanism of each developmental events. Recently, several methods have enabled us to derive spatial expression patterns of numerous genes at one time. These methods enable us to explore spatial gene expression domains and gene modules sharing spatial expression patterns. On the other hand, these are conducted with one time point or sparse time course presumably due to the high cost of the experiments. Hence, the developmental events captured by this method is limited. Furthermore, the dynamics of these spatial gene expression patterns have not been explored at transcriptome wide manner, since the spatial patterns at each stage are analyzed separately presumably due to extensive cell migration during embryogenesis. Such molecular dynamics is now well captured by single cell RNA-seq (scRNA-seq). In particular, the application of this technology to 12 time points of zebrafish embryogenesis revealed molecular differentiation to divergent cell types. On the other hand, each cell of scRNA-seq has lost their spatial information, which obscure the spatiotemporal dynamics of these differentiation process. Here, we integrated transcriptome data containing spatial information and single cell RNA-seq data with cell movement data, which is derived by advanced imaging technology, and estimate time continuous spatiotemporal gene expression patterns for numerous genes to explore the spatiotemporal dynamics of cell type differentiation. In our method, we reconstructed original spatial position of scRNA-seq cells and spatial gene expression patterns at the observation time of the transcriptome data utilizing MAP-EM algorithm. From the estimated spatial gene expression patterns at the observation time, we predicted the spatial gene expression patterns at any time points involved in the cell movement data.

Results and discussion

In simulation experiments for validation of our method, we evaluated the estimation performance of scRNA-seq cell position and spatiotemporal gene expression patterns. We found that 75 %

of scRNA-seq cells are located at the position closer to true position than 20 % of the embryo diameter, while the Pearson's correlation coefficient between estimated and true spatiotemporal gene expression patterns exceeds 0.8 for 84 % of all genes. In the application to real data, we confirmed that the predicted spatial gene expression patterns are consistent with spatial patterns observed from real embryo for many genes. The success in the reconstruction was confirmed at two embryonic stages regardless of the existence of spatial gene expression profiles at the corresponding embryonic stages. Furthermore, we explored the spatiotemporal dynamics of cell type differentiation, and found that expression profiles at midbrain and hindbrain boundary (MHB) is closer to that at later embryonic stage than other regions of midbrain and hindbrain. This is presumably because the differentiation process is most progressed at the region close to MHB due to signaling from MHB, which induces the differentiation process into midbrain and hindbrain in real embryo.

Acknowledgment

At first, I would like to express my great appreciate to Prof. Hisanori Kiryu, whose instructions about the scientific research made up my basis as a researcher. The reviews and valuable comments on my thesis by Prof. Shinichi Morishita, Prof. Shinya Kuroda, Prof. Yutaka Suzuki and Prof. Haruka Ozaki made great improvements on my thesis. I am also thankful for Prof. Kiyosi Asai and the members of Asai Lab, who gave me insightful advises during my doctoral course. I am glad to have published my first research with Dr. Hirotaka Matsumoto. I am also grateful that Mr. Toshiyuki Yokoyama realized the marvelous visualization of the spatiotemporal gene expression patterns reconstructed in my research. The members of Kiryu lab gave me constructive advises and joyful discussions. The fellowship from JSPS was essential for conducting the research during my doctoral course. Finally, I would like to express my great appreciate to my family, fiance and friends for making me happy even when my research faced some difficulties.

Contents

Chapter 1

General introduction

Recent comprehensive molecular observation has enabled us to observe biological metrics across huge numbers of biological units. For example, expression levels of tens of thousands of genes are now observed at the same time [Stark et al., 2019], while allele frequencies of numerous SNPs in a specific population are also quantified at the same time [Schlötterer et al., 2014]. This situation provides us an unique opportunity to aggregate them and derive whole picture of biological process behind the data. On the other hand, each observation extracts only a few aspects of the biological processes despite the huge abundance of the produced data, e.g. RNA-seq data directly provides not the kinetics parameters of transcriptional regulation but the expression level of each gene, which results from the transcriptional regulation. Like this example, some interesting metrics of the biological processes are not directly accessible, even if other metrics of the same processes are extensively observed.

One of the effective methods to overcome this challenge is relating the data with the interesting biological metrics through a probabilistic model of the biological process. When we maximize the generation probability of the data over the metrics, we can derive maximum likelihood estimates, which are unbiased estimates of the true metrics values given abundant data [Sorensen and Gianola, 2007]. Utilizing this method, Sandberg *et al.* estimated the kinetics parameters of transcriptional regulation from single cell RNA-seq (scRNA-seq) data [Sandberg et al., 2018]. On the other hand, such generation probability of the data is often impossible to derive even if there is strong interaction between the data and the metrics. One potential cause of this problem is the lack of observation over some hidden variables, which are additionally required for deriving the generation probability of the data. EM algorithms, which are extensively utilized in the field of biology [Do and Batzoglou, 2008], enable us to maximize the generation probability of the data including the hidden variables

and derive the maximum likelihood estimates of the metrics. In this thesis, we describe two studies where EM algorithms are utilized for efficient estimation of the parameters of a mathematical model of population genetics, Wright-Fisher model (WF), and the reconstruction of time-continuous spatiotemporal gene expression patterns during embryogenesis in Chapter 2 and 3 respectively.

The study described in Chapter 2 are aimed at quantifying evolutionary selection on each single nucleotide polymorphisms (SNPs) during experimental evolution as WF parameters. This estimation was conducted from time course allele frequency data of numerous SNPs during experimental evolution, which is derived by an experimental method Evolve and Resequence (E&R) [Turner et al., 2011]. The optimization of WF parameters to E&R data is challenging because of the abundance of SNPs, for each of which we must estimate the WF parameters. Furthermore, WF parameter estimation faces intrinsic unidentifiability in some conditions, while existing WF optimization methods for E&R data can not evaluate the confidence of the estimates [Iranmehr et al., 2017, Taus et al., 2017]. Hence, there is a risk that the inaccurate estimates, which cannot be excluded by these methods, could affect the results of the post analysis. We constructed a probabilistic model of E&R data generation process, and developed an EM-algorithm to optimize this probabilistic model for E&R data, which enabled efficient estimation of WF parameters and evaluation of confidence for each estimates. We applied our newly developed algorithm to existing E&R data [OROZCO-terWENGEL et al., 2012], and aggregated the WF parameters estimated for numerous SNPs to derive new biological insights. As a result, we found the allele frequency dynamics shared by SNPs within a cosmopolitan inversion region.

In the study described in Chapter 3, we integrated the scRNA-seq data [Farrell et al., 2018] and spatially annotated transcriptome data [Junker et al., 2014] with a spatiotemporal cell movement data during zebrafish embryogenesis [Keller et al., 2008], and estimated the gene expression levels at arbitrary spatiotemporal points within the cell movement data. The unknown cell position of scRNA-seq prevent us from formulating generation probability of all the data. Hence, we constructed a probabilistic model of data generation process with scRNA-seq cell positions as hidden variables. We developed a MAP-EM algorithm which estimates scRNA-seq cell positions and spatial gene expression patterns at the observation time of the transcriptome data. From spatial gene expression patterns at the observation time, we predicted gene expression level at arbitrary spatiotemporal points using Gaussian process regression. In the application of our method to a real zebrafish embryo, we confirmed that the predicted expression patterns well reflected the anatomical structures of the real embryo. Furthermore, the analysis of cell differentiation based on the predicted

spatiotemporal gene expression patterns suggested that the boundary region between midbrain and hindbrain leads the differentiation processes into both of them.

In both of the studies described in this thesis, we connected the biologically interesting metrics with the observed data through the probabilistic models with the unobserved hidden variables. We developed EM algorithms for these probabilistic models to derive either of the maximum likelihood estimates or the maximum a posteriori estimates. We estimated them across a numerous number of SNPs or genes, and aggregated them to derive new insights for each biological processes.

Chapter 2

Estimation of population genetic parameters using an EM algorithm and sequence data from experimental evolution populations

2.1 Introduction

Recent advancements in DNA sequencing technologies have enabled the study of genome-wide changes throughout the process of evolution. For example, a method called evolve and resequence (E&R) has been developed to find associations between genotypes and phenotypes by combining experimental evolution and the resequencing of population genomes (Pool-seq; [Burke et al., 2010, Futschik and Schlötterer, 2010, Turner et al., 2011]). In E&R studies, populations of model organisms, such as *D. melanogaster*, are bred under artificial selective pressures associated with objective traits, e.g. body size [Turner et al., 2011], courtship song [Turner and Miller, 2012] and thermal conditions [OROZCO-terWENGEL et al., 2012, Tobler et al., 2014]. In some studies, allele frequency changes across all SNPs are observed by resequencing the population genomes sampled at multiple generations [OROZCO-terWENGEL et al., 2012, Tobler et al., 2014]. Hence, E&R data is expected to have the potential to provide information about selection on each SNP such as statistical significance, strength and dominance. Such information can help clarify not only which SNPs are under selection but also various quantitative properties of selection.

On the other hand, many studies have focused on only the detection of SNPs selected in evolving populations. For this detection, researchers have used general-purpose statistical tests and machine learning methods such as the Cochran–Mantel–Haenszel (CMH) test [OROZCO-terWENGEL et al., 2012] and BBGP-based test [Topa et al., 2015], which quantify only the statistical significance of selection. The CMH test detects significant allele frequency changes occurring over the course of artificial selection from sequenced allele counts. The BBGP-based test uses a Gaussian process to detect potential upward or downward temporal trends in allele frequencies by comparing inferred likelihoods with those from a time-invariant model. These approaches are limited in what they can reveal about the quantitative properties of selection, as they provide no information about the strength and dominance of selection.

The strength and dominance of selection are represented as parameters in the theoretical Wright–Fisher (WF) model. The WF model is a standard probabilistic model of theoretical population genetics that describes stochastic properties of lineage trees and genetic drift [Fisher, 1930, Wright, 1931, Ewens, 2004]. In this model, the transition density of allele frequencies at each generation is defined by population genetic parameters such as the effective population size, selection coefficient and dominance parameter. This means that we can calculate the likelihood of WF parameters given E&R data. Hence, the properties of selection on each SNP can be quantified as the maximum likelihood estimates of the WF parameters.

There are various methods to estimate WF parameters from temporal allele frequency changes. In many of them, model parameters were estimated using time-consuming methods such as MCMC sampling in which computing the likelihood is costly [Bank et al., 2014]. Hence, these methods are difficult to apply to *Drosophila* E&R data, which includes about one million SNPs. Other methods include computationally efficient algorithms based on approximate Bayesian computation (ABC) [Foll et al., 2015] and expectation maximization (EM) algorithms [Mathieson and McVean, 2013]. However, these methods are still expected to require 200 or more hours for E&R data analysis.

Recently, two methods focused on the estimation of WF parameters from E&R data have achieved sufficient efficiency for estimation based on one million SNPs. [Iranmehr et al., 2017] combined a grid search approach and efficient operation vectorization, while [Taus et al., 2017] estimated WF parameters based on a deterministic relationship between averaged allele frequency paths and WF parameters. However, the method developed by [Iranmehr et al., 2017] does not provide a reliability score for each estimate, while that developed by [Taus et al., 2017] computes

confidence intervals too slowly to be applied to a large number of SNPs. Hence, it is difficult to use these methods to analyse quantitative properties of evolutionary processes when estimated parameters tend to be inaccurate owing to a paucity of replicates as in most E&R studies, which manifests itself especially when selection coefficients and dominance parameters are simultaneously estimated.

In this study, we developed an EM algorithm of the WF model for E&R data (EMWER). Using the EM algorithm for continuous-time Markov chain (CTMC) developed by [Kiryu, 2011], EMWER does not need to calculate sufficient statistics at each generation, a process which has negative effects on the estimation speed of an existing EM algorithm [Mathieson and McVean, 2013]. Furthermore, EMWER can provide a confidence interval (CI) for each estimated parameter based on an empirical Fisher information matrix (FIM). We show that our method yields estimates distributed around the true values for realistic parameter ranges, and some aspects of the estimation performance are superior to the performance of other existing WF parameter estimation methods after excluding estimates with large confidence intervals, which are likely to be inaccurate. Furthermore, our method and another WF optimization methods [Iranmehr et al., 2017] detected selected SNPs more accurately than other statistical and machine learning methods applied to E&R data. We applied our method to an E&R study in which the authors examined the adaptation of *D. melanogaster* to a thermally fluctuating environment [OROZCO-terWENGEL et al., 2012]. From the estimated selection coefficients, we found a common selection affecting allele frequencies of many SNPs within the cosmopolitan *In(3R)P* inversion, which has been inferred to have a role in climate change adaptation of natural *D. melanogaster* populations [Rane et al., 2015]. Furthermore, our estimation of dominance parameters suggests that many of the positively selected alleles in this E&R experiment are dominant. This observation highlights the abundance of deleterious recessive alleles.

2.2 Methods

Fig. 2.1 shows our probabilistic model for analysing E&R data. Suppose we cannot observe the true allele frequencies x at various sequencing time points and the continuous allele frequency path Y between the observation time points in experimental evolution populations, but we can obtain read depths d and read counts α of the variant allele, which are sampled from the true allele frequencies x at a series of time points. We assume the dynamics of the allele frequencies follows the WF diffusion process, which is described by a kind of Kolmogorov forward equation (KE). This KE contains three kinds of parameters, selection coefficient s, dominance parameter h and effective population size

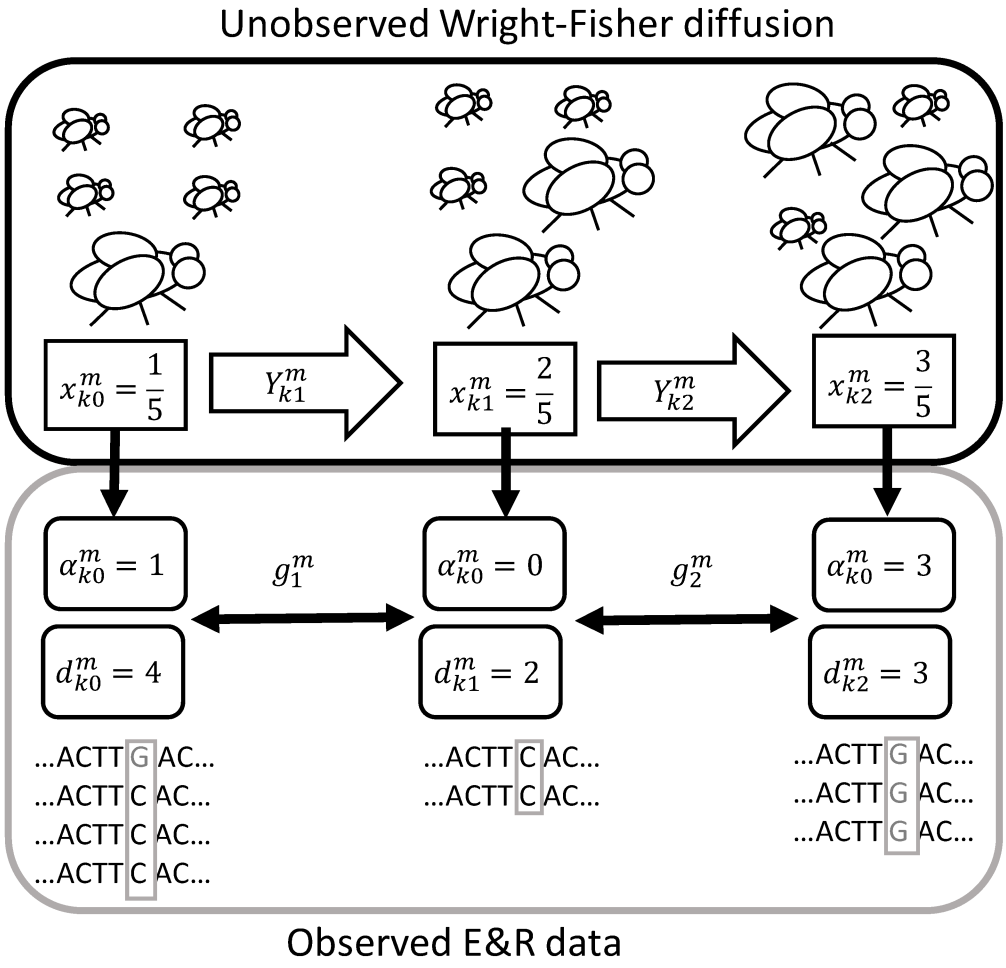


Figure 2.1: Our probabilistic model of E&R data for the k-th SNP and replicate m ($k = 1 \ldots K; m = 1 \ldots M$). d^m_{kl} and α^m_{kl} represent read depth and read count of the variant allele, respectively. They are observed at interval g^m_l from the previous time point and sampled from the unobserved true allele frequencies in the population $x^m_{k,l}$. We assume that the dynamics of the true allele frequencies follows Wright–Fisher diffusion and that the true allele frequency changes according to the trajectory $Y^k_{m,l+1}$ until the next time point $l+1$ at which sequencing occurs.

N_e, and our objective is to develop a method to estimate the most plausible values of s, h and N_e from time-course read counts d and variant allele counts α.

In the following text, we first describe the KE associated with WF diffusion and its conversion into discrete-state CTMC. Then, we introduce our new EM algorithm for WF parameter inference from E&R data, which is based on an EM algorithm for discrete-state CTMC [Kiryu, 2011]. We also describe a likelihood ratio test to detect selected SNPs and a method to compute confidence

intervals based on empirical FIM to exclude inaccurate estimates.

2.2.1 KE of Wright-Fisher model

The Wright–Fisher (WF) model describes statistical properties of ancestral lineage structures in a population with a constant effective population size N_e. When it is applied to allele frequency changes in a diploid population, it includes the strength s and dominance h of selection for variant alleles as parameters. The mean and variance of variant allele frequencies change at each generation as described below:

$$\mu_\theta(x) = \frac{sx(1-x)(x+h(1-2x))}{1+sx(x+2h(1-x))} \tag{2.1}$$

$$\phi_\theta^2(x) = \frac{1}{2N_\mathrm{e}}x(1-x). \tag{2.2}$$

where x represents the variant allele frequency. It is well known that allele frequencies under WF models can be approximated by a continuous-time diffusion process $dX_t = \mu_\theta(X_t)dt + \phi_\theta^2(X_t)dW_t$ where the unit time is one generation, X_t is the allele frequency at time t, $\mu_\theta(X_t)$ is a drift coefficient, $\phi_\theta^2(X_t)$ is a diffusion coefficient, W_t is a standard Brownian process and θ represents all WF parameters, i.e. s, h, N_e, at once [Ewens, 2004]. The time evolution of the probability density of X_t is given by the following KE:

$$\frac{\partial}{\partial t}p_\theta(x|z,t) = -\frac{\partial}{\partial x}\mu_\theta(x)p_\theta(x|z,t) + \frac{\partial^2}{\partial x^2}\phi_\theta^2(x)p_\theta(x|z,t) \tag{2.3}$$

where $p_\theta(x|z,t)$ is the probability density of $X_t = x$ given $X_0 = z$. Then, this can be formed as below:

$$\frac{\partial}{\partial t}p_\theta(x|z) = -\frac{\partial}{\partial x}\frac{1}{r_\theta(x)}\frac{\partial}{\partial x}\frac{1}{q_\theta(x)}p_\theta(x|z,t,\theta)$$

where

$$\begin{aligned} r_\theta(x) &= \exp\left(-2\int^x dx\frac{\mu_\theta(x)}{\phi_\theta^2(x)}\right) \\ &= \exp\left(-2N_\mathrm{e}\log(sx(x+2h(1-x))+1)\right) \\ &= (sx(x+2h(1-x))+1)^{-2N_\mathrm{e}} \\ q_\theta(x) &= \frac{1}{r_\theta(x)\,\phi_\theta^2(x)} \\ &= \frac{2N_\mathrm{e}\left(sx(x+2h(1-x))+1\right)^{2N_\mathrm{e}}}{x(1-x)}. \end{aligned} \tag{2.4}$$

2.2.2 Discretized KE

As in previous studies that have utilized the WF diffusion [Song and Steinrücken, 2012, Ferrer-Admetlla et al., 2016] for estimating WF parameters, we discretize the KE into D discrete states by considering a mapping $f : \mathbb{R} \to \mathbb{Z}$ from a continuous state $x \in [0, 1]$ to a discrete state variable $i \in \{0 \dots D-1\}$ such that $f(x) = i$ for $(i\delta < x < (i+1)\delta)$ where a small positive number $\delta = \frac{1}{D-1}$ is the grid width of discretization. Utilizing the discretization, we convert the KE (**Eq.2.3**) to the differential equation of the transition probabilities between the discretized states. Here, the discretized transition probability from the state $f(X_0) = j$ to the state $f(X_t) = i$ given the duration t can be formulated as below:

$$P_\theta(i|j,t) = \int_{i\delta}^{(i+1)\delta} dx \int_{j\delta}^{(j+1)\delta} dz p_\theta(x|z,t)\frac{1}{\delta} = \int_{i\delta}^{(i+1)\delta} dx p_\theta(x|j\delta,t) + \mathcal{O}(\delta^2). \tag{2.5}$$

As a first step of the conversion, we integrate the equation (**Eq.2.3**) for x over a micro interval $[i\delta, (i+1)\delta]$ as below:

$$\frac{\partial}{\partial t}P_\theta(i|j,t) = \frac{1}{r_\theta(i\delta)}\frac{\partial}{\partial x}\frac{1}{q_\theta(x)}p_\theta(x|j\delta,t)|_{x=i\delta} - \frac{1}{r_\theta((i+1)\delta)}\frac{\partial}{\partial x}\frac{1}{q_\theta(x)}p_\theta(x|j\delta,t)|_{x=(i+1)\delta} + \mathcal{O}(\delta^2) \tag{2.6}$$

For the accomplishment of the conversion, the representations of $\frac{\partial}{\partial x}p_\theta(x|j\delta,t)\big|_{x=i\delta}$ and $p_\theta(i\delta|j\delta,t)$ by $P_\theta(i|j,t)$ are needed because

$$\frac{\partial}{\partial x}\frac{1}{q_\theta(x)}p_\theta(x|j\delta,t)\bigg|_{x=i\delta} = \frac{\partial}{\partial x}\frac{1}{q_\theta(x)}\bigg|_{x=i\delta} p_\theta(i\delta|j\delta,t) + \frac{1}{q_\theta(i\delta)}\frac{\partial}{\partial x}p_\theta(x|j\delta,t)\bigg|_{x=i\delta}. \tag{2.7}$$

To derive the relationship between $P_\theta(i|j,t)$ and $\frac{\partial}{\partial x}p_\theta(x|z,t)|_{x=i\delta,z=j\delta}$ and $p_\theta(i\delta|j\delta,t)$, we approximate $p_\theta(x|j\delta,t)$ by the first two terms of the Taylor series $p_\theta^{(i)}(x|j\delta,t) = p_\theta(i\delta|j\delta,t) + (x - i\delta)\frac{\partial}{\partial x}p_\theta(x|j\delta,t)|_{x=i\delta}$. We can approximately derive $P_\theta(i|j,t)$ and $P_\theta(i-1|j,t)$ as below:

$$\begin{aligned}
P_\theta(i|j,t) &= \int_{i\delta}^{(i+1)\delta} dx p_\theta^{(i)}(x|j\delta,t) + \mathcal{O}(\delta^3) \\
&= \delta p_\theta(i\delta|j\delta) + \frac{1}{2}\delta^2 \frac{\partial}{\partial x}p_\theta(x|j\delta,t)\bigg|_{x=i\delta} + \mathcal{O}(\delta^3) \\
P_\theta(i-1|j,t) &= \int_{(i-1)\delta}^{i\delta} dx p_\theta^{(i)}(x|j\delta,t) + \mathcal{O}(\delta^3) \\
&= \delta p_\theta(i\delta|j\delta) - \frac{1}{2}\delta^2 \frac{\partial}{\partial x}p_\theta(x|j\delta,t)\bigg|_{x=i\delta} + \mathcal{O}(\delta^3).
\end{aligned}$$

Transforming these equations, we can derive

$$\begin{aligned}
p_\theta(i\delta|j\delta,t) &= \frac{1}{2\delta}\left(P_\theta(i|j,t) + P_\theta(i-1|j,t)\right) + \mathcal{O}(\delta^2) \\
\frac{\partial}{\partial x}p_\theta(x|j\delta,t)|_{x=i\delta} &= \frac{1}{\delta^2}\left(P_\theta(i|j,t) - P_\theta(i-1|j,t)\right) + \mathcal{O}(\delta).
\end{aligned} \tag{2.8}$$

Substituting (**Eq.2.7**) and (**Eq.2.8**) into (**Eq.2.6**), we can derive the discretized KE as below:

$$\begin{aligned}
\frac{\partial}{\partial t}P_\theta(i|j,t) &= (R_{\theta,i,i+1}P_\theta(i+1|j,t) + R_{\theta,i,i-1}P_\theta(i-1|j,t) + R_{\theta,ii}P_\theta(i|j,t)) + \mathcal{O}(\delta) \\
\Leftrightarrow \frac{\partial}{\partial t}\boldsymbol{P_\theta}(j,t) &= R_\theta \boldsymbol{P_\theta}(j,t) + \mathcal{O}(\delta) \qquad (2.9)
\end{aligned}$$

where $\boldsymbol{P_\theta}(j,t) \in \mathbb{R}^D$ is the probability distribution of $f(X_t)$ from $f(X_0) = j$ given duration t, $R_\theta \in \mathbb{R}^{D\times D}$ is the transition rate matrix,

$$[\boldsymbol{P_\theta}(j,t)]_i = P_\theta(i|j,t)$$

$$\begin{aligned}
R_{\theta,i,j} &= \begin{cases}
\frac{1}{\delta r_{\theta,i}}\left(\frac{1}{\delta q_{\theta,i}} - \frac{1}{2q'_{\theta,i}}\right) & (j = i-1) \\
\frac{1}{\delta r_{\theta,i}}\left(\frac{1}{\delta q_{\theta,i+1}} + \frac{1}{2q'_{\theta,i+1}}\right) & (j = i-1) \\
-(R_{\theta,i-1,i} + R_{\theta,i+1,i}) & (j = i) \\
0 & (i \neq j, j\pm 1)
\end{cases} \\
r_{\theta,i} &= r_\theta(i\delta) \\
q_{\theta,i} &= q_\theta(i\delta) \\
q'_{\theta,i} &= \frac{\partial}{\partial x}\frac{1}{q_\theta(x)}\bigg|_{x=i\delta}. \qquad (2.10)
\end{aligned}$$

The equation (**Eq.2.9**) can be regarded as the time evolution equation for a discrete-state CTMC with the transition rate matrix R_θ. Hence, when the allele frequency X_t is discretized by the discretizing function $f(x) = i \in \mathbb{Z}$ such that $i\delta < x < (i+1)\delta$, the stochastic process for $\tilde{X}_t = f(X_t)$ can be regarded as a discrete-state CTMC. Hereafter, we assume that allele frequencies x and z have already been discretized into D discrete states using function f.

2.2.3 EM algorithm for KE

We utilized the EM algorithm developed for a discrete-state CTMC [Kiryu, 2011] and derived our algorithm for the optimization of WF parameters from E&R data. In order to introduce the sufficient statistics and basic concepts utilized in our algorithm, we describe the details of the EM algorithm developed by [Kiryu, 2011] in this section before describing our algorithm for E&R data. Here, we naively applied the algorithm to the optimization of the parameters of KE, which can be converted into a discrete state CTMC as described in the above section.

2.2.3.1 Computing likelihood for KE

Suppose we have observed a set of independent observation $\{(x_a, z_a, t_a)|a = 1, \dots, A\}$ from a stochastic process X_t following KE (**Eq.2.3**), where each transition from the initial point z_a to the final point x_a occurs over an observation interval t_a. We want to infer θ that generated the observations. For each data point a, we divide the observation interval t_a into $N >> 1$ equal subintervals. Then, utilizing the finite difference of (**Eq.2.9**), the probability of transition from state j to state i in the short interval $\Delta t_a = t_a/N$ can be approximated as below:

$$P_\theta(i|j, \Delta t_a) = \left[Q^{(a)}\right]_{ij} + \mathcal{O}(\Delta t_a^2)$$

where

$$Q^{(a)} = I + \Delta t_a R_\theta.$$

We define $Y_a = \{y_{a,n}|n = 1...N-1\}$ as the $(N-1)$-step path of the discretized stochastic process $\tilde{X}_t = f(X_t)$ between the initial point z_a and the end point x_a. Because the transition from z_a to x_a through Y_a follows the KE, we can form $P_\theta(x_a, Y_a|z_a, t_a)$ as below:

$$\begin{aligned} P_\theta(x_a, Y_a|z_a, t_a) &= P_\theta(x_a|y_{a,N-1}, \Delta t_a) \prod_n P_\theta(y_{a,n}|y_{a,n-1}, \Delta t_a) P_\theta(y_{a,1}|z_a, \Delta t_a) \\ &= Q^{(a)}_{x_a, y_{a,N-1}} \prod_n Q^{(a)}_{y_{a,n}, y_{a,n-1}} Q^{(a)}_{y_{a,1}, z_a}. \end{aligned} \tag{2.11}$$

Here, we assume that R_θ can be diagonalized (**See details in Section 2.3.1**) and formulate the diagonalization as below:

$$R_\theta = U \Lambda U^{-1}$$

where $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots)$ is a diagonal matrix whose elements are the eigenvalues of R_θ, and U is an invertible matrix whose columns are corresponding eigenvectors. We numerically calculated $\{\lambda_i\}$ and U using "EigenSolver" implemented in "Eigen" library of "C++" language [Guennebaud et al., 2010]. The transition probability $P_\theta(x_a|z_a, t_a)$ without a middle path Y_a is equal to summation of $P_\theta(x_a, Y_a|z_a, t_a)$ for all possible Y_a regardless of the number of the steps N. Then, we can formulate

the $P_\theta(x_a|z_a,t_a)$ as below:

$$\begin{aligned}
P_\theta(x_a|z_a,t_a) &= \lim_{N\to\infty}\sum_{Y_a} P_\theta(x_a,Y_a|z_a,t_a) \\
&= \lim_{N\to\infty}\sum_{Y_a} Q^{(a)}_{x_a,y_{a,N-1}}\prod_n Q^{(a)}_{y_{a,n},y_{a,n-1}}Q^{(a)}_{y_{a,1},z_a} \\
&= \lim_{N\to\infty}\left[Q^{(a)N}\right]_{x_a,z_a} \\
&= \lim_{N\to\infty}\left[\{U(I+\Delta t_a\Lambda)U^{-1}\}^N\right]_{x_a,z_a} \\
&= \lim_{N\to\infty}\left[U(I+\Delta t_a\Lambda)^N U^{-1}\right]_{x_a,z_a} \\
&= \left[U\Lambda'^{(a)}U^{-1}\right]_{x_a,z_a} \qquad (2.12)
\end{aligned}$$

where $\Lambda'^{(a)} = \mathrm{diag}(\exp(t_a\lambda_1),\exp(t_a\lambda_2),\dots)$ is a diagonal matrix.

2.2.3.2 Sufficient statistics for KE

From the terms related to R_θ, we can form the equation (**Eq.2.11**) as below:

$$\begin{aligned}
P_\theta(x_a,Y_a|z_a,t_a) &= \lim_{N\to\infty} Q^{(a)}_{x_a,y_{a,N-1}}\prod_n Q^{(a)}_{y_{a,n},y_{a,n-1}}Q^{(a)}_{y_{a,1},z_a} \\
&= \lim_{N\to\infty}\prod_{i,j} Q^{(a)N_{a,ij}} \\
&= \lim_{N\to\infty}\prod_i\left((I+\Delta t_a R_{\theta,ii})^{N\frac{N_{a,ii}}{N}}\prod_{j=i\pm1}\Delta t_a R_{\theta,ij}^{N_{a,ij}}\right) \\
&= \exp\left(\sum_i\left(F_{a,i}t_a R_{\theta,ii}+\sum_{j=i\pm1}N_{a,ij}\log\left(\Delta t_a R_{\theta,ij}\right)\right)\right) \\
&= \exp\left(\sum_i\left(F_{a,i}t_a R_{\theta,ii}+\sum_{j=i\pm1}N_{a,ij}\log R_{\theta,ij}\right)+\text{const.}\right) \qquad (2.13)
\end{aligned}$$

where $N_{a,ij}$ is the number of transitions from state j to state i in the path from z_a to x_a through Y_a, and $F_{a,i}$ is the fraction of the time duration for which $\tilde{X}_t$ remained at state i during in same path. Then, we can regard $F_{a,i}$ and $N_{a,ij}$ as sufficient statistics for the statistics (x_a,Y_a,z_a) under parameter θ. Hence, calculating the Q-function of the EM algorithm, which is the expected value of the log likelihood, only requires the calculation of the expected values of $F_{a,i}$ and $N_{a,ij}$ for the set of hidden variables Y_a.

2.2.3.3 Expected values of sufficient statistics

The expected values of sufficient statistics $F_{a,i}$ and $N_{a,ij}$ can be derived from the transition probability $P_\theta(x_a|z_a,t_a)$ as below:

$$\begin{aligned}
\frac{1}{t_a P_\theta(x_a|z_a,t_a)}\frac{\partial}{\partial R_{\theta,ii}}P_\theta(x_a|z_a,t_a) &= \frac{1}{t_a P_\theta(x_a|z_a,t_a)}\frac{\partial}{\partial R_{\theta,ii}}\sum_{Y_a}P_\theta(x_a,Y_a|z_a,t_a) \\
&= \frac{1}{t_a P_\theta(x_a|z_a,t_a)}\frac{\partial}{\partial R_{\theta,ii}}\sum_{Y_a}\exp\left(F_{a,i}t_a R_{\theta,ii}+\text{const.}\right) \\
&= \frac{1}{t_a P_\theta(x_a|z_a,t_a)}\sum_{Y_a}F_{a,i}t_a P_\theta(x_a,Y_a|z_a,t_a) \\
&= \sum_{Y_a}F_{a,i}P_\theta(Y_a|x_a,z_a,t_a) \\
&= \overline{F}_{\theta,a,i} \qquad (2.14)
\end{aligned}$$

and when $j=i\pm 1$

$$\begin{aligned}
\frac{R_{\theta,ij}}{P_\theta(x_a|z_a,t_a)}\frac{\partial}{\partial R_{\theta,ij}}P_\theta(x_a|z_a,t_a) &= \frac{R_{\theta,ij}}{P_\theta(x_a|z_a,t_a)}\frac{\partial}{\partial R_{\theta,ij}}\sum_{Y_a}P_\theta(x_a,Y_a|z_a,t_a) \\
&= \frac{R_{\theta,ij}}{P_\theta(x_a|z_a,t_a)}\frac{\partial}{\partial R_{\theta,ij}}\sum_{Y_a}\exp\left(N_{a,ij}\log R_{\theta,ij}+\text{const.}\right) \\
&= \frac{R_{\theta,ij}}{P_\theta(x_a|z_a,t_a)}\sum_{Y_a}\frac{1}{R_{\theta,ij}}N_{a,ij}P_\theta(x_a,Y_a|z_a,t_a) \\
&= \sum_{Y_a}N_{a,ij}P_\theta(Y_a|x_a,z_a,t_a) \\
&= \overline{N}_{\theta,a,ij}. \qquad (2.15)
\end{aligned}$$

Furthermore, $\frac{\partial}{\partial R_{\theta,ij}}P_\theta(x_a|z_a,t_a)$ can be numerically calculated as below:

$$\begin{aligned}
\frac{\partial}{\partial R_{\theta,ij}}P_\theta(x_a|z_a,t_a) &\approx \frac{\partial}{\partial R_{\theta,ij}}\left[\lim_{N\to\infty}\left(I+\Delta t_a R_\theta\right)^N\right]_{x_a,z_a} \\
&= \lim_{N\to\infty}\sum_{n,j,j'}\left[\left(I+\Delta t_a R_\theta\right)^n\right]_{x_a,j}\left[\left(I+\Delta t_a R_\theta\right)^{N-(n+1)}\right]_{j',z_a}\frac{\partial}{\partial R_{\theta,ij}}\left[I+\Delta t_a R_\theta\right]_{jj'} \\
&= \lim_{N\to\infty}\sum_{n}\frac{t_a}{N}\left[\left(I+\Delta t_a R_\theta\right)^{N\frac{n}{N}}\right]_{x_a,i}\left[\left(I+\Delta t_a R_\theta\right)^{N(1-\frac{n+1}{N})}\right]_{j,z_a} \\
&= \int_0^1 t_a\left[\left(U\Lambda'^{(a)}U^{-1}\right)^w\right]_{x_a,i}\left[\left(U\Lambda'^{(a)}U^{-1}\right)^{1-w}\right]_{j,z_a}dw \\
&= \int_0^1 t_a\left[U\Lambda'^{(a)w}U^{-1}\right]_{x_a,i}\left[U\Lambda'^{(a)1-w}U^{-1}\right]_{j,z_a}dw \\
&= \int_0^1 t_a\sum_{u,v}U_{x_a,u}U^{-1}_{u,i}\exp(wt_a\lambda_u)U_{j,v}U^{-1}_{v,z_a}\exp((1-w)t_a\lambda_v)dw \\
&= t_a\sum_{u,v}U_{x_a,u}U^{-1}_{u,i}U_{j,v}U^{-1}_{v,z_a}K_{u,v,a} \qquad (2.16)
\end{aligned}$$

where

$$\begin{aligned} K_{u,v,a} &= \int_0^1 \exp(wt_a\lambda_u)\exp((1-w)t_a\lambda_v)dw \\ &= \int_0^1 \exp(wt_a(\lambda_u-\lambda_v)+t_a\lambda_v)dw \\ &= \begin{cases} \frac{\exp(t_a\lambda_u)-\exp(t_a\lambda_u)}{t_a(\lambda_u-\lambda_v)} & (u \neq v) \\ \exp(t_a\lambda_u) & (v=u) \end{cases}. \end{aligned}$$

2.2.3.4 EM algorithm Q-function

Using the equations (**Eq.2.14**), (**Eq.2.15**) and (**Eq.2.16**), we can calculate the expected values of the sufficient statistics $\overline{F}_{\theta,a,i}$ and $\overline{N}_{\theta,a,ij}$. Then, using the expected values and (**Eq.2.13**), we can calculate the Q-function of the EM algorithm and its gradient values as below:

$$\begin{aligned} Q(\theta|\theta') &= \sum_{a,Y_a} P_{\theta'}(Y_a|x_a,z_a)\log P_\theta(x_a,Y_a|z_a) \\ &= \sum_{a,Y_a} P_{\theta'}(Y_a|x_a,z_a)\left(\sum_i\left(F_{a,i}t_aR_{\theta,ii}+\sum_{j=i\pm1}N_{a,ij}\log R_{\theta,ij}\right)+\text{ const.}\right) \\ &= \sum_{a,i}\left(\overline{F}_{\theta',a,i}t_aR_{\theta,ii}+\sum_{j=i\pm1}\overline{N}_{\theta',a,ij}\log R_{\theta,ij}\right)+\text{const.} \end{aligned}$$

$$\frac{\partial}{\partial\theta}Q(\theta|\theta') = \sum_{a,i}\left(\overline{F}_{\theta',a,i}t_a\frac{\partial}{\partial\theta}R_{\theta,ii}+\sum_{j=i\pm1}\overline{N}_{\theta',a,ij}\frac{1}{R_{\theta,ij}}\frac{\partial}{\partial\theta}R_{\theta,ij}\right).$$

We can derive the local maximum likelihood estimates by iterating the calculation of the expected sufficient statistics and the maximization of the Q-function.

2.2.4 Application to ER data

The application of the algorithm of [Kiryu, 2011] to KE enables us to estimate WF parameters because WF diffusion is represented as KE. However, it requires true allele frequencies in a population during experimental evolution, which follow WF diffusion, while E&R data do not provide them. In this section, we describe an EM algorithm for estimation of WF parameters from E&R data (EMWER) by extending the algorithm of [Kiryu, 2011] described in the previous section.

2.2.4.1 Likelihood function for E&R data

Like previous E&R studies [Topa et al., 2015, Terhorst et al., 2015], we used a hidden Markov model as a probabilistic model for E&R data. Suppose we have M replicates of breeding populations and each population genome is sequenced at $(L+1)$ time points including the initial and final time points. We denote N_{e} as the common effective population size, s_k and h_k as the selection coefficient and the dominance parameter, respectively, for each SNP $k \in \{1, \ldots, K\}$ in the genome and θ as all these parameters combined. Then, let g_l^m be the number of generations between the $(l-1)$-th and l-th Pool-seq experiments for replicate m ($m = 1, \ldots, M, l = 1, \ldots, L$), and let $d_{k,l}^m$ and $\alpha_{k,l}^m$ respectively be the read depth and the read count of variant alleles of SNP k for replicate m at the l-th time point ($k = 1, \ldots, K, m = 1, \ldots, M, l = 0, \ldots, L$). Further, let $x_{k,l}^m$ be the unobserved true allele frequencies at the l-th time point and $Y_{k,l}^m$ be the stochastic trajectory of unobserved true allele frequencies between $x_{k,l-1}^m$ and $x_{k,l}^m$. Then, the joint probability of generating the observed and unobserved data is given by the following:

$$P_\theta(\alpha, x, Y | d, g, \theta) = \prod_{k,m} \left[\begin{array}{l} \left(\prod_l P(\alpha_{k,l}^m | x_{k,l}^m, d_{k,l}^m) P_{\theta_k}(x_{k,l}^m, Y_{k,l}^m | x_{k,l-1}^m, g_l^m) \right) \\ \times P(\alpha_{k,0}^m | x_{k,0}^m, d_{k,0}^m) P_0(x_{k,0}^m) \end{array} \right] \tag{2.17}$$

where $\alpha = \{\alpha_{k,l}^m\}$, $Y = \{Y_{k,l}^m\}$, $x = \{x_{k,l}^m\}$, $d = \{d_{k,l}^m\}$, $g = \{g_l^m\}$, and $\theta_k = \{N_{\mathrm{e}}, s_k, h_k\}$.

Since all probabilistic variables are independent of the variables with either different replicate or different locus indexes in our probabilistic model, we can separately consider sufficient statistics for each replicate and locus. Hence, until the later section detailing Q-function of our EM algorithm (**Section 2.2.4.3**), we consider only a single replicate and locus for simplicity, where $x_l = x_{k,l}^m$, $Y_l = Y_{k,l}^m$, $\alpha_l = \alpha_{k,l}^m$, $d_l = d_{k,l}^m$, $g_l = g_l^m$, $\theta = \theta_k$. Here, we assume that the number of observed variant alleles α_l follows the binomial distribution $P(\alpha_l | x_l, d_l) = \mathrm{Binom}(\alpha_l | x_l, d_l)$ and that the initial allele frequency of base populations x_0 follows a multinomial distribution $P_0(x_0) = q_{x_0}$ where $q = (q_0, \ldots, q_{D-1})^T$ is given. On the other hand, $P_\theta(x_l, Y_l | x_{l-1}, g_l)$ is the generation probability of the allele frequency trajectory following KE, which is described in **Section 2.2.3.1**. Then, we can

describe the likelihood of the E&R data without the hidden variables as below:

$$\begin{aligned} P_\theta(\alpha|d,t,\theta) &= \sum_{x,Y} P_\theta(\alpha,x,Y|d,t,\theta) \\ &= \sum_{x_*} \left[\begin{array}{l} \left(\prod_l P(\alpha_l|x_l,d_l) \sum_{Y_l} P_\theta(x_l,Y_l|x_{l-1},g_l)\right) \\ \times P(\alpha_0|x_0,d_0)P_0(x_0) \end{array} \right] \\ &= \sum_{x_*} \left(\prod_l [A_l]_{x_l,x_{l-1}} \right) [\boldsymbol{b}]_{x_0} \\ &= \mathbf{1}^T \left(\prod_l A_l \right) \boldsymbol{b} \end{aligned}$$

where $\sum_{x_*}$ represents the summation over all possible $\{x_l | l = 0, \ldots, L\}$, $\mathbf{b} \in \mathbb{R}^D$, $A_l \in \mathbb{R}^{D \times D}$

$$\begin{aligned} [\boldsymbol{b}]_i &= P(\alpha_0|i,d_0)P_0(i) \\ [A_l]_{i,j} &= P(\alpha_l|i,d_l)P_\theta(i|j,g_l) \\ \mathbf{1}^T &= (1,1,\ldots,1). \end{aligned}$$

Using (**Eq.2.12**), we can numerically calculate the likelihood of E&R data.

2.2.4.2 Expected values of sufficient statistics for E&R data

From the terms related to the WF model parameters, (**Eq.2.17**) can be formed as below:

$$\begin{aligned} \log P_\theta(\alpha,Y,x|d,t,\theta) &= \log \left[\begin{array}{l} (\prod_l P(\alpha_l|x_l,d_l)P_\theta(x_l,Y_l|x_{l-1},g_l)) \\ \times P(\alpha_0|x_0,d_0)P_0(x_0) \end{array} \right] \\ &= \sum_l \log P_\theta(x_l,Y_l|x_{l-1}) + \text{const.} \\ &= \sum_{l,i} \left(F_i^{(l)} g_l R_{\theta,ii} + \sum_{j=i\pm1} N_{ij}^{(l)} \log R_{\theta,ij} \right) + \text{const.} \end{aligned} \tag{2.18}$$

where $R_{\theta,ij}$ is an element of the transition rate matrix defined in (**Eq.2.10**). As in (**Eq.2.13**), $N_{ij}^{(l)}$ is the number of transitions from state j to state i among the transitions of allele frequencies from x_{l-1} to x_l through Y_l, and $F_i^{(l)}$ is the fraction of time for which the allele frequency remained at state i. Then, we can regard $F_i^{(l)}$ and $N_{ij}^{(l)}$ as sufficient statistics of the random path x_l, Y_l, and x_{l-1} for parameter θ. Now, we consider a statistic $J^{(l)} = J(x_l, Y_l, x_{l-1})$ that is dependent only on

x_l, Y_l, and x_{l-1} for some l. Then, the expected value of $J^{(l)}$ is given by

$$
\begin{aligned}
E_{x,Y|\alpha}[J^{(l)}] &= \sum_{x,Y} P_\theta(x, Y|\alpha) J^{(l)} \\
&= \sum_{x,Y} \frac{P_\theta(x, Y, \alpha)}{P_\theta(\alpha)} J^{(l)} \\
&= \frac{1}{P_\theta(\alpha)} \sum_{x_*, Y_*} \left[\begin{array}{l} (\prod_l P(\alpha_l|x_l, d_l) P_\theta(x_l, Y_l|x_{l-1}, g_l)) \\ \times P(\alpha_0|x_0, d_0) P_0(x_0) J^{(l)} \end{array} \right] \\
&= \frac{1}{P_\theta(\alpha)} \sum_{x_*} \left[\begin{array}{l} \left\{ \prod_{l'(\neq l)} [A_{l'}]_{x_{l'}, x_{l'-1}} \right\} [\boldsymbol{b}]_{x_0} \\ \times P(\alpha_l|x_l, d_l) \sum_{Y_l} P_\theta(x_l, Y_l|x_{l-1}, g_l) J^{(l)} \end{array} \right] \\
&= \frac{1}{P_\theta(\alpha)} \sum_{x_l, x_{l-1}} \left[\begin{array}{l} \left[\mathbf{1}^T \prod_{l'=l+1}^{L} A_{l'} \right]_{x_l} \left[\left(\prod_{l'=1}^{l-1} A_{l'} \boldsymbol{b} \right)^T \right]_{x_{l-1}} \\ \times P(\alpha_l|x_l, d_l) P_\theta(x_l|x_{l-1}, g_l) \sum_{Y_l} P_\theta(Y_l|x_l, x_{l-1}, g_l) J^{(l)} \end{array} \right] \\
&= \frac{1}{P_\theta(\alpha)} \sum_{x_l, x_{l-1}} \left[\left[\mathbf{1}^T \prod_{l'=l+1} A_{l'} \right]_{x_l} \left[\left(\prod_{l'=1}^{l-1} A_{l'} \boldsymbol{b} \right)^T \right]_{x_{l-1}} [A_l]_{x_l, x_{l-1}} \overline{J}_\theta^{(l)} \right]
\end{aligned}
\tag{2.19}
$$

where $\alpha = \{\alpha_l | l = 0, \ldots, L\}$, $x_* = \{x_l | l = 0, \ldots, L\}$, $Y_* = \{Y_l | l = 0, \ldots, L\}$, and

$$
\overline{J}_\theta^{(l)} = \sum_{Y_l} P_\theta(Y_l|x_l, x_{l-1}, g_l) J^{(l)} = E_{Y_l|x_l, x_{l-1}, g_l}[J^{(l)}].
$$

Thus, we can obtain the expected value of $J^{(l)}$ if we can compute $\overline{J}_\theta^{(l)}$. When $J^{(l)} \in \{F_i^{(l)}, N_{ij}^{(l)}\}$, $\overline{J}_\theta^{(l)}$ can be computed using (**Eq.2.14**), (**Eq.2.15**) and (**Eq.2.16**). Thus, we obtain $E_{x,Y|\alpha}[F_i^{(l)}]$ and $E_{x,Y|\alpha}[N_{ij}^{(l)}]$ using the above equations. Below, these expected values are denoted by

$$
\widetilde{F}_{\theta,i}^{(l)} = E_{x,Y|\alpha,\theta}[F_i^{(l)}] \tag{2.20}
$$

$$
\widetilde{N}_{\theta,ij}^{(l)} = E_{x,Y|\alpha,\theta}[N_{ij}^{(l)}]. \tag{2.21}
$$

2.2.4.3 EM algorithm Q-function for E&R data

Using the expected values of the sufficient statistics (**Eq.2.20**), (**Eq.2.21**) and the joint generation probability of observed and unobserved data (**Eq.2.18**), we can calculate the Q-function of the EM

algorithm and gradient as below:

$$
\begin{aligned}
Q(\theta|\theta') &= \sum_{x,Y} P_{\theta'}(x,Y|\alpha)\log P_\theta(x,Y,\alpha) \\
&= \sum_{k,l,m,x,Y} P_{\theta'}(x,Y|\alpha)\left(\sum_i F_i^{(k,l,m)} t_{k,l}^m R_{\theta_k,ii} + \sum_{j=i\pm1} N_{ij}^{(k,l,m)} \log R_{\theta_k,ij} + \text{const.}\right) \\
&= \sum_{k,l,m,i}\left(\widetilde{F}_{\theta'_k,i}^{(k,l,m)} g_l R_{\theta_k,ii} + \sum_{j=i\pm1} \widetilde{N}_{\theta'_k,ij}^{(k,l,m)} \log R_{\theta_k,ij}\right) + \text{const.}
\end{aligned}
$$

$$
\frac{\partial}{\partial\theta}Q(\theta|\theta') = \sum_{l,m,i}\left(\widetilde{F}_{\theta'_k,i}^{(k,l,m)} g_l \frac{\partial}{\partial\theta} R_{\theta_k,ii} + \sum_{j=i\pm1} \widetilde{N}_{\theta'_k,ij}^{(k,l,m)} \frac{1}{R_{\theta_k,ij}} \frac{\partial}{\partial\theta} R_{\theta_k,ij}\right) \tag{2.22}
$$

where $\widetilde{F}_{\theta'_k,i}^{(k,l,m)} = \widetilde{F}_{\theta',i}^{(l)}$ and $\widetilde{N}_{\theta'_k,ij}^{(k,l,m)} = \widetilde{N}_{\theta',ij}^{(l)}$ for locus k and replicate m. We can analytically derive the transition matrix R_{θ_k} and its derivative from (**Eq.2.4**) and (**Eq.2.10**). Then, we maximize $Q(\theta|\theta')$ using the gradient descent method LBFGS-B, whose boundary is determined so that all the transition rates between adjacent states $R_{\theta_k,ij}(j = i \pm 1)$ are positive (**Eq.2.10**). By iterating the calculation of $\widetilde{F}_{\theta_k,i}^{(k,l,m)}$ and $\widetilde{N}_{\theta_k,ij}^{(k,l,m)}$ (E-step) and maximizing $Q(\theta|\theta')$ for θ (M-step), we obtain maximum likelihood estimates of s_k and h_k. On the other hand, we compute the maximum likelihood estiamtes for N_e by grid search, because the flat likelihood surface for N_e causes slow convergence of the EM algorithm.

2.2.4.4 Adaptive discretization grids

We observed that the estimation of large selection coefficients often failed when we set the number of discretization grids to a small number (e.g. 20). This is because high selection coefficients produce strong fluxes of WF diffusion and make some transition rates between adjacent states $R_{\theta_k,ij}(j = i\pm1)$ negative. This issue is diminished when we set the large number of the grids (e.g. 100). However, a large number of the grids D increases computational costs, which follow $O(D^3)$ time complexity, while many of SNPs in E&R data are expected to be neutral ($s = 0$) and thus do not need such large grid sizes. Hence, we adopt an adaptive grid strategy; during preprocessing, EMWER determines an appropriate parameter range for each grid size (i.e. $D = 20, 30, \ldots, D_\mathrm{max}$). In the parameter inference, the grid size is firstly set to 20 and incremented by 10 each time estimated parameters exceed the range of the current grid size. In this study, we used $D_\mathrm{max} = 100$ when s or N_e are estimated solely and $D_\mathrm{max} = 200$ when s and h are estimated simultaneously.

2.2.4.5 Likelihood ratio test for selection

a large number of non selected SNPs makes it difficult to analyze the estimated WF parameters of fewer selected SNPs. Hence, the method to efficiently detect the selected SNPs is crucial in our analysis. In this section, we describe the likelihood ratio test for this purpose. After determining the effective population size N_{e} by using randomly sampled SNP sites, we can derive maximum likelihood estimates for selection coefficients $\hat{s}_k$ for the k-th SNP and the corresponding log likelihood $L_k(\theta) = \sum_{m,l} \log(P(\alpha_{kl}^m | d_{kl}^m, \hat{s}_k, h, N_{\mathrm{e}}))$, assuming additive selection ($h = 0.5$). A likelihood ratio is defined by $D_k = -2(L_k(\theta_0) - L_k(\hat{\theta}))$, where $\hat{\theta} = (\hat{s}, 0.5, N_e), \theta_0 = (0, 0.5, N_e)$. If a SNP is neutral ($s = 0$), its likelihood ratio follows a χ^2 distribution with one degree of freedom in the asymptotic limit. Hence, EMWER calculates statistic D_k as the significance index of selection at the k-th SNP. However, because E&R data are expected not to reach the asymptotic limit owing to their small numbers of replicates, we calculate empirical P-values from a comparison of D_k between real data and simulated data of neutral SNPs ($s = 0$), using the R package "qvalue". Furthermore, we calculate Q-values from empirical P-values using the R package "qvalue" and set the FDR threshold to 0.05 for real data analyses.

2.2.4.6 Confidence intervals of estimated values

It is desirable that inaccurate estimates are excluded because they obscure analyses of estimated parameter values. For that reason, we calculate confidence intervals (CIs) for each estimated parameter value and exclude inaccurate estimates whose confidence intervals are large. First, using the first-order derivative of $Q_{\mathrm{EM}}(\theta|\theta')$, which is equal to the first-order derivative of the likelihood [Oakes, 1999], and finite difference, we calculate an approximately empirical Fisher-information matrix (FIM) as below:

$$\begin{aligned} I(\hat{\theta})_{i,j} &= \frac{\partial}{\partial \theta_i} \frac{\partial}{\partial \theta_j} L(\theta) \Big|_{\theta=\hat{\theta}} \\ &\approx \frac{1}{\epsilon} \left(\frac{\partial}{\partial \theta_i} Q_{\mathrm{EM}}(\theta|\theta) \Big|_{\theta=\theta'=\hat{\theta}+\epsilon^{(j)}} - \frac{\partial}{\partial \theta_i} Q_{\mathrm{EM}}(\theta|\theta) \Big|_{\theta=\theta'=\hat{\theta}} \right) \end{aligned}$$

where $\epsilon << 1$ and $\epsilon_k^{(j)} = 0 (k \neq j), \epsilon (k = j)$. It is known that, at the asymptotic limit, an empirical FIM can be associated with the CI of maximum likelihood estimates $\hat{\theta}$; $CI_{\theta_i} = 1.96\sqrt{[I(\hat{\theta})^{-1}]_{ii}}$, where $[\hat{\theta}_i - CI_{\theta_i}, \hat{\theta}_i + CI_{\theta_i}]$ is the 95% CI for $\hat{\theta}_i$ [Oakes, 1999].

2.2.5 Simulated datasets

In this study, we generated artificial E&R data using the forward simulation tool MimicrEE [Kofler and Schlotterer, 2014]. This application evolves population genomes from a given set of diplotypes under specified selection intensities acting on SNPs. At each generation, crossover events are introduced based on a specified recombination rate, and mating among diploid pairs is performed with a success rate based on selection acting on diploid genotypes. Although MimicrEE is a simulation tool for the evolution of entire population genomes, we used it to iteratively simulate only a single SNP site (with 1000 iteration unless otherwise noted). This is because our method is a site-independent model, like most other methods, and further, the accuracies for the simulation data of entire genomes by MimicrEE exhibit high variability owing to strong correlations among sites and do not converge to a stable result with our available computational resources. Unless otherwise noted, the dominance parameter h was set to 0.5, the number of individuals N was set to 250, the depths of reads d_{kl}^{m} were sampled from the Poisson distribution with a mean of 50, and the initial allele frequencies x_{k0} were sampled from a uniform distribution between 0.05 and 0.95. We also used more realistic initial allele frequency distributions derived from the histogram of initial allele frequencies in a real E&R dataset [OROZCO-terWENGEL et al., 2012] for our comparison of the accuracies for detecting selected SNPs among several programs.

2.2.6 Comparison with other methods for inferring WF parameters

In order to study the accuracy and speed of the selection estimation, we performed a comparison between EMWER and five other WF parameter estimation methods by [Taus et al., 2017] , [Iranmehr et al., 2017], [Foll et al., 2015], [Mathieson and McVean, 2013] and [Ferrer-Admetlla et al., 2016]. All these methods can be used to estimate selection coefficients from time-series allele count data. For an optimization strategy, [Ferrer-Admetlla et al., 2016] used time-consuming sampling methods (MCMC). On the other hand, [Taus et al., 2017] , [Iranmehr et al., 2017], [Foll et al., 2015] and [Mathieson and McVean, 2013] used more efficient optimization techniques such as linear regression, operation vectorization, approximate Bayesian computation (ABC) and EM. As approximation strategies, [Taus et al., 2017] and [Ferrer-Admetlla et al., 2016] use the diffusion approximation as EMWER does. In contrast, [Iranmehr et al., 2017] and [Foll et al., 2015] use a discrete WF model without approximation. Using simulated E&R data with selection coefficients randomly sampled from the uniform distribution $s \sim [0, 0.3]$, we evaluated the accuracy as summarized by mean absolute errors (MAE). For comparisons with the methods that return sampled parameters [Foll et al.,

2015,Ferrer-Admetlla et al., 2016], we used the average value of sampled parameters as the estimate. In order to measure the estimation speed, we calculated the time required to estimate selection coefficients corresponding to one million SNPs with quad-core computing. These values were derived by appropriately scaling the measured execution time for 100 SNPs, since some programs require too much time to estimate the selection coefficients of one million SNPs. Allele frequency changes of the SNPs were sampled from real data [OROZCO-terWENGEL et al., 2012]. However, since the efficiency of operation vectorization manifests when the number of SNPs is very large, we also measured the execution time for estimating selection coefficients of 50,000 SNPs for the [Iranmehr et al., 2017] method. The implementations of each of these methods were downloaded from their respective websites; the [Taus et al., 2017] method at `https://github.com/ThomasTaus/poolSeq`, the [Iranmehr et al., 2017] method at `https://github.com/airanmehr/CLEAR`, the [Foll et al., 2015] method at `http://jjensenlab.org/software`, the [Mathieson and McVean, 2013] method at `https://github.com/mathii/s_lattice` and the [Ferrer-Admetlla et al., 2016] method at `https://bitbucket.org/phaentu/approxwf/src`.

2.2.7 Comparison with other methods for detecting selected SNPs

To evaluate the accuracy of calling selected SNPs, we sampled sequenced alleles α_{kl}^m under selection coefficients of $s = 0$ and $s = 0.1$ and investigated the discriminability of selected SNPs with $s = 0.1$ from neutral SNPs (i.e. $s = 0$). Since we are generally interested in a small number of highly significant SNPs among a large number of SNPs throughout the genome, we used the true positive rate (TPR) at a 0.01 false positive rate (FPR) on the receiver operating characteristic (ROC) curve as the accuracy measure. We compared the detection accuracy of our method to those of the CMH test [Agresti, 2002] and BBGP-based test [Topa et al., 2015]. The CMH test is a general statistical test for multiple 2×2 contingency tables. In E&R applications, the CMH test is used to detect a significant change of the relative proportion of major and minor allele counts between the initial and final conditions. The CMH test is available in the R statistical software environment as the "mantelhaen.test" function. In the BBGP-based test, allele counts of each population are assumed to be sampled from the same allele frequency for all the populations, and their evolution over time follows a Gaussian process. The BBGP-based test decides whether a SNP is significantly selected or not by comparing the likelihood of the time-evolving model to that of a time-invariant model. A BBGP-based test implementation is available at `https://github.com/handetopa/BBGP`.

2.3 Results

2.3.1 Diagonalizability of the transition rate matrix of WF diffusion

While we assume that R_θ is diagonalized as $U\Lambda U^{-1}$, the sufficient and necessary condition for diagonalizability of R_θ is the invertibility of U. Hence, the diagonalization may not be effective in an exceptional situation. In order to evaluate the numerical calculation of the diagonalized form $U\Lambda U^{-1}$ of R_θ, we numerically compared the diagonalized form and the original form. This comparison showed that $U\Lambda U^{-1}$ with small values of N_e and s were almost identical to R_θ, while the deviations between them were magnified for large values of N_e and s (**Fig. 2.2-a,b,c**). On the other hand, the diagonalized forms appears only as the exponential form of them $U\exp(t\Lambda)U^{-1}$ in our algorithm. We observed that the deviations between the exponential forms, $U\exp(t\Lambda)U^{-1}$ and $\left(I+\frac{t}{N}R\right)^N$ $(t=30, N=400)$, were relatively smaller even when N_e and s were large (**Fig. 2.2-d,e,f**). This is presumably because the calculation of the pairs of eingenvectors and engenvalues with large eigenvalues, which become dominant by exponential transformation, are relatively accurate.

2.3.2 Accuracy of estimates from simulation data

To evaluate the estimation precision of our algorithm, we estimated selection coefficients s, population sizes N_e and dominance parameters h from sampled allele counts at five time points (0, 10, 30, 50 and 60 generations). In this experiment, we estimated only one of s, N_e, and h at a time with the other two parameters fixed to correct values. Our estimates of s are distributed around true values respectively for $s = 0, 0.05, 0.1, 0.15, 0.2$ (**Fig. 2.3-a**). As the interval time g increases, the accuracy generally improves (**Fig. 2.4-a**). This is because it is easier to observe allele frequency changes that have occurred over longer intervals. However, these progressive improvements eventually saturate, especially for stronger selection coefficients, which implies many of the samples reached fixation around the saturation time points, after which allele frequencies cannot change any further.

Similarly, estimated effective population sizes N_e are distributed near the true values for the various number of individuals $N = 100, 200, 250, 300, 400, 500$. The median of the relative errors has maximum 3.8% at $N = 300$ when the initial allele frequencies were sampled from the uniform distribution between 0.2 and 0.8 (**Fig. 2.3-b**). When we used SNPs whose initial allele frequencies (IAF) are close to the boundaries (0 or 1), we observed that the estimated effective population sizes had small positive bias (**Fig. 2.5**). The small positive bias of the estimated N_e is presumably caused by the diffusion approximation to the discrete Wright-Fisher model, whose error is larger

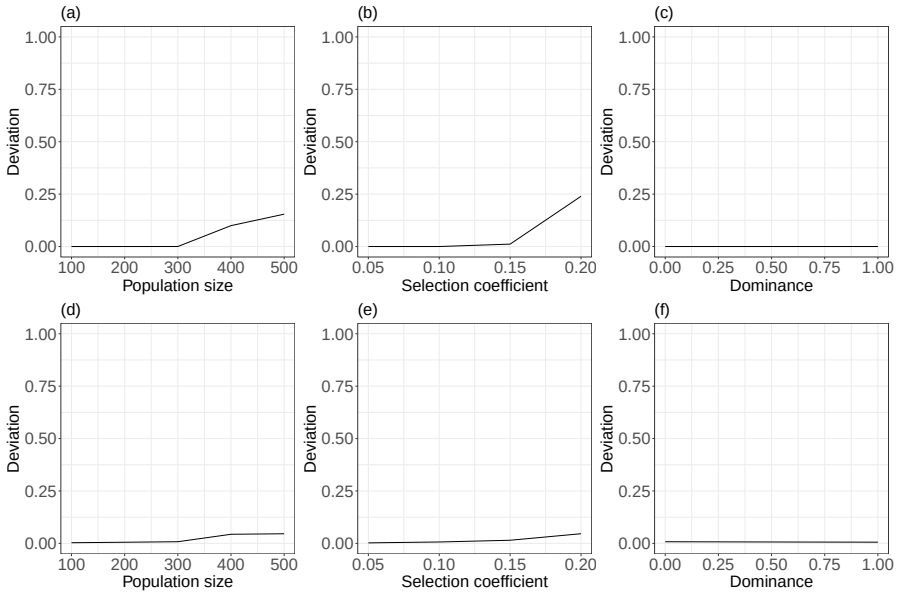


Figure 2.2: Errors of diagonalization. Deviations between the diagonalized form $\tilde{R_\theta} = U\Lambda U^{-1}$ and the original forms R_θ of the transition rate matrix are displayed in (a), (b) and (c). The deviations between exponential forms, $\exp t\tilde{R_\theta} = U\exp(t\Lambda)U^{-1}$ and $\exp tR_\theta = \left(I + \frac{t}{N}R\right)^N$ $(t = 30, N = 400)$, are displayed in (d), (e) and (f). The deviations are quantified as $\frac{\sum_{i,j}|\tilde{R}_{\theta ij} - R_{\theta,ij}|}{\sum_{i,j}|R_{\theta,ij}|}$ in (a), (b) and (c), and $\frac{\sum_{i,j}|\exp(t\tilde{R_\theta})_{ij} - \exp(tR_\theta)_{ij}|}{\sum_{i,j}|\exp(tR_\theta)_{ij}|}$ in (d), (e) and (f). We varied N_e in (a) and (c), s in (b) and (e), and h in (c) and (f).

for the allele frequency close to the boundaries. When we restricted the IAF to the interval 0.2 and 0.8, the decrements of positive bias was saturated (**Fig. 2.5**). Hence, we estimate N_e, using SNPs whose initial allele frequencies are between 0.2 and 0.8 in this study. The variance of the estimation was larger for large N, since in such cases genetic drift is weak and accuracy is affected by the finite sampling noise of sequenced reads. In these cases, estimation accuracy is considerably improved by increasing read coverage (**Fig. 2.4-b**). Additionally, we confirmed that estimated dominance parameters are distributed around true values for varying parameter values $h = 0, 0.2, 0.5, 0.8, 1$ (**Fig. 2.3-c**).

2.3.2.1 Simultaneous estimation for dominance and selection

In the dotted lines of **Fig. 2.6**, we illustrates the 2D density of estimated parameters $(\hat{s}, \hat{h})$, which are simultaneously estimated from samples generated under a condition $(s = 0.1, h = 0)$. They

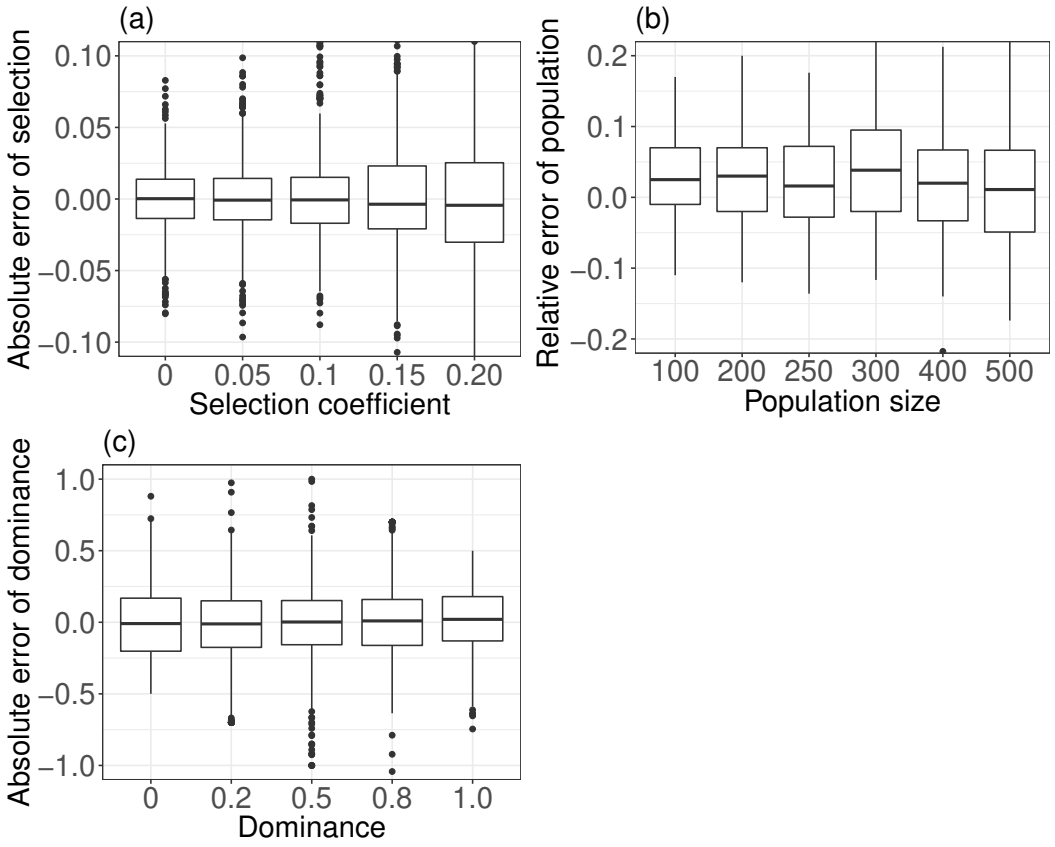


Figure 2.3: Distributions of errors of estimated selection coefficients (a), effective population sizes (b) and dominance parameters (c) for various parameter values. The y-axes represent absolute errors in (a) and (c) and relative errors in (b) from the true values.

show that the 2D densities have modes near the true value. When the number of replicates is small ($M = 3$, left panel), the mode slightly shifted from the true position toward positive dominance. As the number of replicates is increased ($M = 20$, right panel), the mode approaches very close to the true position. This result is consistent with the fact that maximum likelihood estimates are asymptotically unbiased, though there is currently no E&R experiments that use 20 or more replicates. The figure also shows that a considerable fraction of the estimates are much larger than the true value for both parameters (**Fig. 2.6**). This postive deviation of estimated parameters are due to the flat ridge of likelihood landscape of the WF probability model (**Fig. 2.8**). The solid lines of **Fig. 2.6** show that the filtering by confidence intervals efficiently removes such inaccurate estimates. We also note that dominance parameter is intrinsically difficult to estimate for weakly selected sites (**Fig. 2.4-c**) (**Fig. 2.7**), because the differences in relative fitness between genotypes

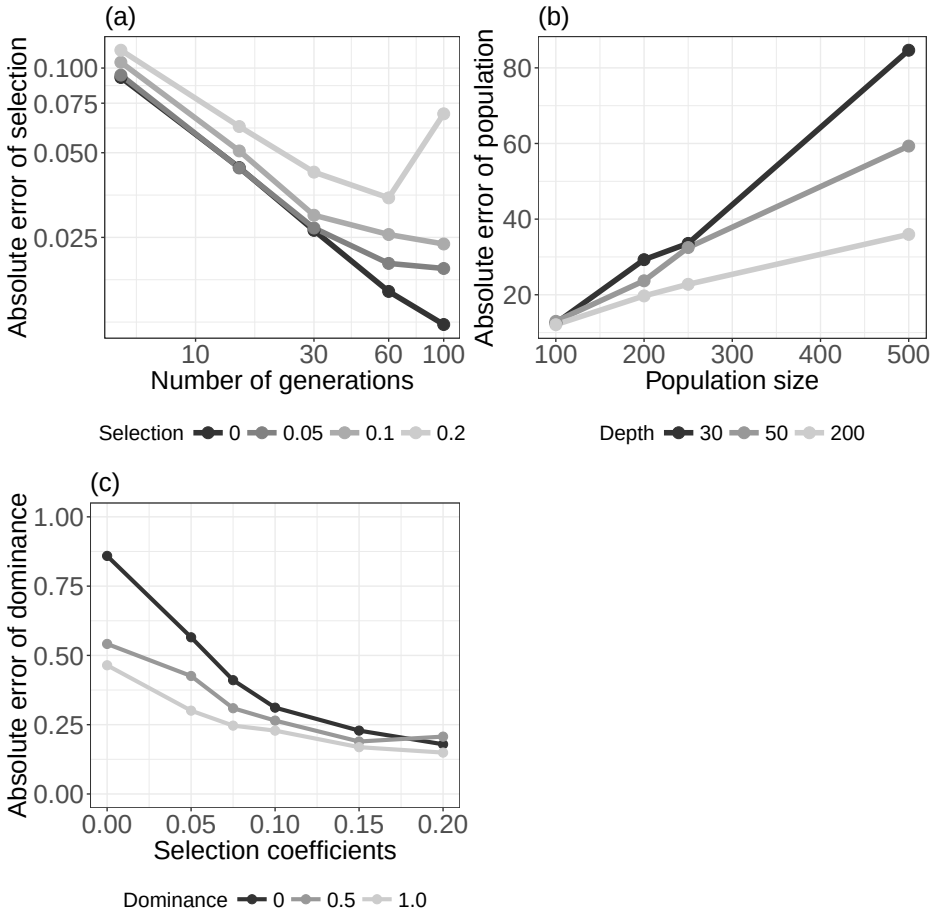


Figure 2.4: Dependencies of the accuracy of estimated parameters on various experimental conditions. (a) The absolute error of selection coefficients (y-axis) versus the interval lengths between two observations (x-axis, log scale) for varying selection coefficients. (b) The absolute error of estimated effective population size (y-axis) versus true population size (x-axis) for varying read depths. (c) The absolute error of estimated dominance parameters (y-axis) versus selection coefficients (x-axis) for varying levels of dominance.

are small regardless of dominance parameter values.

2.3.2.2 Confidence intervals

Since inaccurate estimates complicate downstream analyses, we used CIs to filter inaccurate estimates, which are expected to have large CIs. To validate the 95% CIs based on empirical FIM, we investigated the proportion of CIs that contained the true values, which should be 0.95 if the computed CIs are accurate, using simulated datasets. When selection coefficients are estimated both

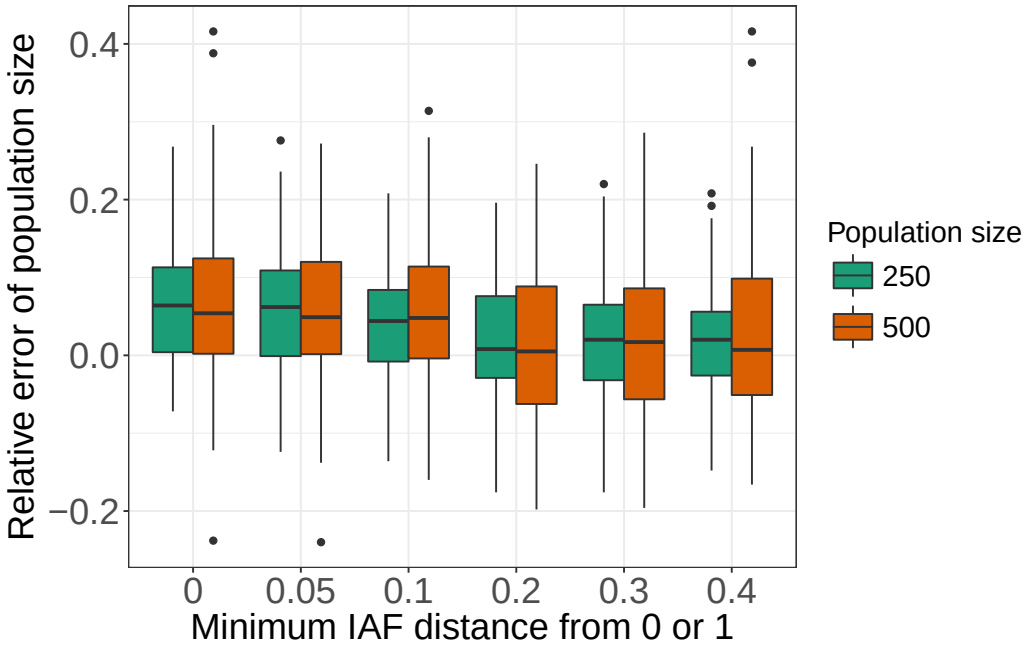


Figure 2.5: Distributions of relative errors of estimated effective population sizes for $N = 250, 500$. The y-axes represent relative errors from the true values. In this experiment, we exclude SNPs whose initial allele frequency (IAF) is near to the boundaries (0 or 1). We vary the minimum distance of IAF from the boundaries.

with and without dominance parameters, the proportions of the selection coefficients contained within the 95% CIs were 0.952 and 0.970, respectively (**Table. 2.1**). In contrast, the proportion of dominance parameter estimates within the 95% dominance parameter CIs was 0.801 when selection coefficients and dominance parameters were simultaneously estimated. This is caused by a flat ridge in the likelihood landscape around inaccurate estimates as described above, where the peak along a constant selection coefficient is narrow while the peak along a constant dominance parameter value is broad (**Fig. 2.8**).

However, we found that the proportion of dominance parameter CIs containing the true values was 0.902 if we limited the SNPs to those with small selection coefficient CIs (i.e. $CI_s < 0.1$). We

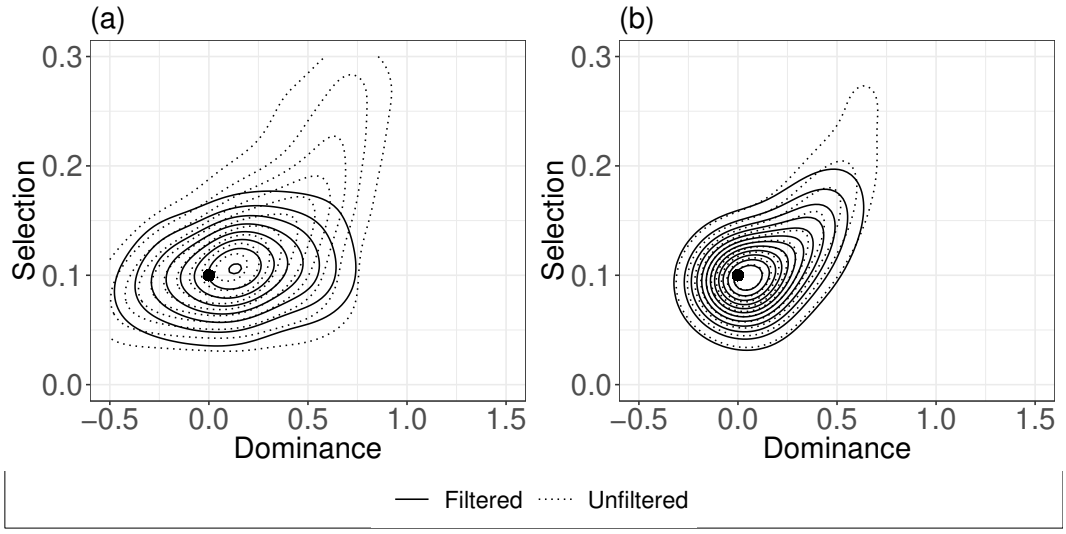


Figure 2.6: Two-dimensional density of selection coefficients and dominance parameter estimates based on simulated datasets before and after filtering. The dashed lines represent the contour for all estimates, while the solid lines represent the contour for estimate with small confidence intervals ($CI_s < 0.1, CI_h < 1.0$). The black point represents the true value used to generate the datasets ($s = 0.1, h = 0$). The number or replicates is 3 in (a) and 20 in (b).

confirmed that these observations are similar for various parameter values ($s = 0 \sim 0.2, h = 0 \sim 1.0$) (**Table. 2.1**). Therefore, our simulation study indicates that CIs computed from empirical FIMs are sufficiently accurate for realistic E&R data, even though it is theoretically guaranteed only at the limit of an infinite number of replicates.

To investigate whether inaccurate estimates can actually be removed by imposing a threshold on CIs, we calculated the fraction of inaccurate estimates (i.e. $|\hat{s} - s| > 0.1, |\hat{h} - h| > 0.5$) in all the estimates whose CIs are smaller than a threshold (**Fig. 2.9**). We confirmed that a lower threshold decrease the fraction of inaccurate estimates when only selection coefficients are estimated (**Fig. 2.9-a**). Furthermore, when selection coefficients and dominance parameters are simultaneously estimated, the fraction of inaccurate estimates is decreased with decreasing threshold of dominance CI if a threshold of selection CI ($CI_s < 0.1$) is applied at the same time (**Fig. 2.9-b**). The filtering based on CIs can efficiently exclude inaccurate estimates for stronger selection ($s > 0.06$), while almost all estimates for weaker selected loci ($s < 0.04$) are excluded after the filtering (**Fig. 2.10**). This is because it is intrinsically difficult to estimate the dominance parameters for weakly selected loci (see **Section 3.1.1**). Hence, these results show that the filtering is effective

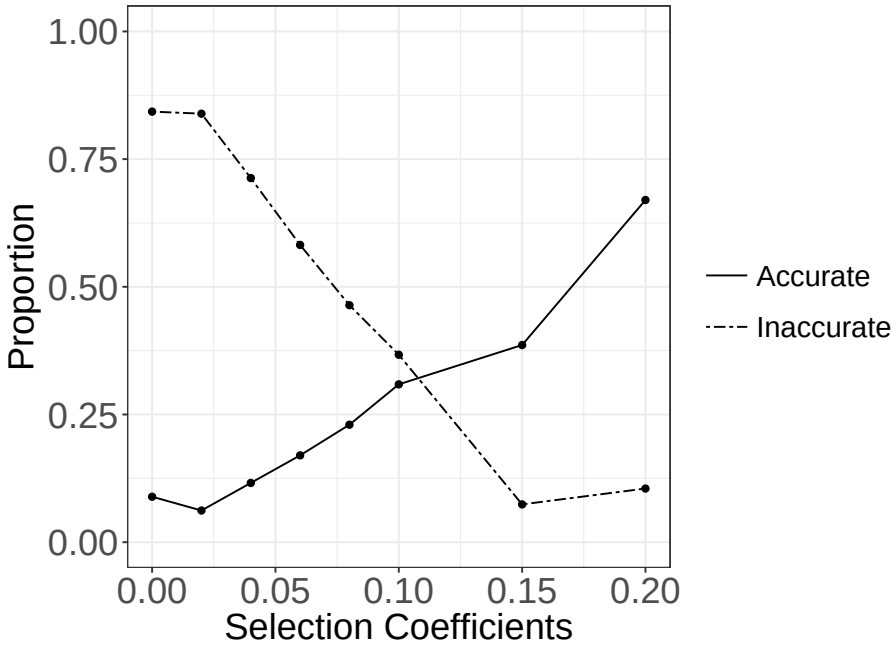


Figure 2.7: Proportion of accurate (i.e. $|\hat{s} - s| < 0.05 \wedge |\hat{h} - h| < 0.2$) and inaccurate (i.e. $|\hat{s} - s| > 0.1 \vee |\hat{h} - h| > 0.5$) estimates among all the estimates for varying true selection coefficients.

across varying selection coefficients ($s = 0 \sim 0.2$). In summary, our simulation shows that by imposing a CI threshold, we can selectively remove inaccurate estimates discussed above section (**Fig. 2.6**) and safely analyze the estimated values even when selection coefficients and dominance parameters are estimated simultaneously.

2.3.2.3 Comparison with other methods for estimating WF parameters

To make a comparison between EMWER and other methods of estimating WF parameters, we investigated the execution time and estimation accuracy using down-sampled real datasets and simulated datasets, respectively. When selection coefficients were estimated in this experiment, the execution time of EMWER was about 3-fold smaller than the most efficient existing method ([Iranmehr et al., 2017]) (**Table. 2.2**). While [Taus et al., 2017] achieved the efficiency closest to that of

Selection	Only.slection	With.dominance
0	0.961	0.991
0.02	0.982	0.993
0.04	0.964	0.968
0.06	0.957	0.963
0.08	0.947	0.973
0.1	0.952	0.970
0.15	0.937	0.980
0.2	0.930	0.963
Dominance	Before.exclusion	After.exclusion
0	0.780	0.912
0.2	0.786	0.900
0.5	0.801	0.902
0.8	0.835	0.933
1	0.861	0.943

Table 2.1: Proportion of confidence intervals (CIs) that included the true selection coefficients (above) and dominance parameters (below). For selection coefficients, the proportion is displayed when the selection coefficients are estimated both solely and simultaneously with dominance parameters. For dominance parameters, the proportion is displayed both after and before estimates with large CIs (i.e. $CI_s > 0.1$) are excluded.

EMWER and [Iranmehr et al., 2017], its calculations of P-values and confidence intervals, based on iterative simulation, are very slow. Additionally, we confirmed that the relative execution time of the existing EM algorithm [Mathieson and McVean, 2013] is improved when the number of generation is smaller, as it needs to calculate sufficient statistics for each generation under experimental evolution (**Fig. 2.11**).

For selection coefficients randomly sampled from the uniform distribution $s \sim [0, 0.3]$ with fixed dominance value $h = 0.5$, the MAE of EMWER after excluding the estimates with top 30% CI_s (0.0219) was almost identical to that of the most accurate method with the same filtering [Ferrer-Admetlla et al., 2016] (0.0214) (**Table. 2.2**). On the other hand, the accuracy of EMWER without the filtering (0.0354) was inferior to that of [Iranmehr et al., 2017] (0.0273), which has

Method	Time(s)	Time(s,h)	Error(s)	Error(s)+filter	Error(h)	Error(h)+filter
EMWER	**1.04**	43.52	0.0354	0.0219	0.339 (0.230†)	**0.140**
CLEAR	3.08	**7.49**	**0.0273**	NA	0.320 (0.231†)	NA
poolSeq	4.55 (1646*, 1875†)	20.51	0.0460	0.0322	0.415 (0.325†)	NA
WFABC	240.18	1886.47	0.0639	0.0556	**0.231** (0.225†)	0.208
Approx	998.63	1746.78	0.0297	**0.0214**	0.234 (0.230†)	0.245
Lattice	720.50	NA	0.0478	NA	NA	NA

Table 2.2: Accuracy and execution time of estimating WF parameters by our methods (EMWER) as well as those of [Taus et al., 2017] (poolSeq), [Iranmehr et al., 2017] (CLEAR), [Foll et al., 2015] (WFABC), [Mathieson and McVean, 2013] (Lattice) and [Ferrer-Admetlla et al., 2016] (Approx). We solely estimate selection coefficients parameters in "Time (s)", "Error (s)" and "Error (s) + filter", while we simultaneously estimate selection coefficients and dominance parameters in "Time (s,h)", "Error (h)" and "Error (h) + filter". "Time (s)" and "Time (s,h)" represents the time required to estimate WF parameters of one million SNPs with quad-core computing. Additionally, we represents the execution time of poolSeq with detection and confidence interval options in parentheses, which is marked by $*$ and † respectively in 'Time (s)". "Error (s)" and "Error (h)" represents MAEs of selection coefficients and dominance parameters. The true selection coefficients are randomly sampled from the uniform distribution $s \sim [0, 0.3]$ in both "Error (s)" and "Error (h)". The true dominance parameters fixed to 0.5 in "Error (s)", while they are randomly sampled from the uniform distribution $h \sim [0, 1.0]$ in "Error (h)". For EMWER, poolSeq, WFABC and Approx, we represents the selection errors after excluding the estimates with top 30% CIs in "Error (s) + filter". For EMWER, WFABC and Approx, we represent the dominance errors after excluding the estimates with top 70% CIs of either s or h "Error (h) + filter". For poolSeq, we represent the errors of only available estimates (14%) in "Error (h)" and do not represent the accuracy after filtering in "Error (h) + filter", because poolSeq failed to calculate CIs for over 70% SNPs (89%). Additionally, we represent the dominance errors after excluding the SNPs weakly selected ($s < 0.1$) in parentheses, which is marked by † in "Error (h)".

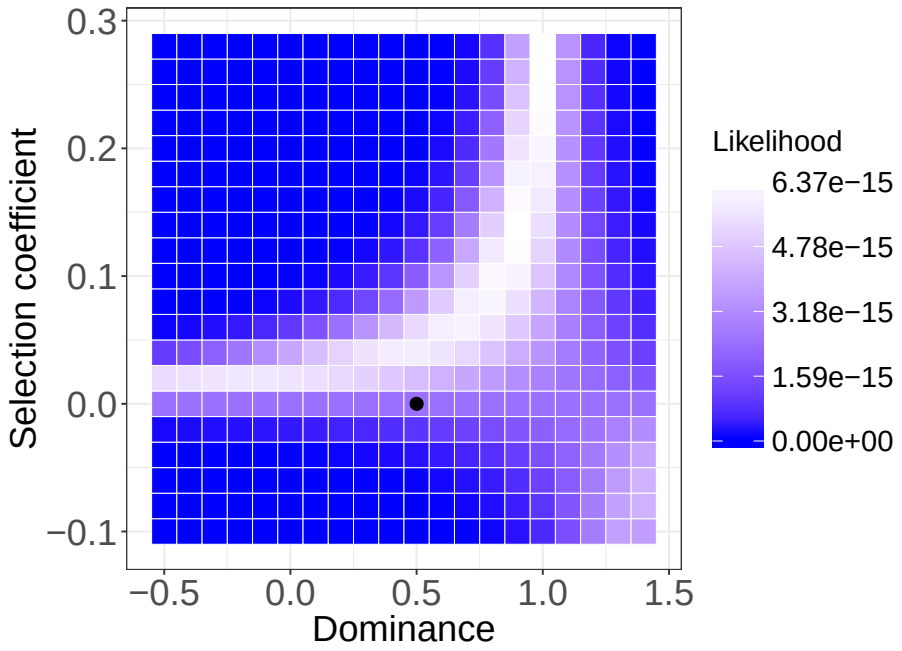


Figure 2.8: Likelihood landscape for simulated ER data ($s = 0, h = 0.5$). The black point indicates the true value.

no methods for excluding inaccurate estimates. We confirmed that the accuracy of EMWER and [Ferrer-Admetlla et al., 2016] were better than those of the other methods for varying filtering criteria when we exclude more than about 20% or more of all the estimates (**Fig. 2.12-a**). The superior accuracy of EMWER and [Ferrer-Admetlla et al., 2016] with CI filtering were confirmed under various experimental conditions (**Fig. 2.13**). These results show that, along with [Ferrer-Admetlla et al., 2016], EMWER can provide the most precise estimates by excluding inaccurate estimates in the range of our experiments, while the execution time of EMWER was about 20,000 times smaller than that of [Ferrer-Admetlla et al., 2016].

We compared the accuracy of dominance estimation for selection coefficients and dominance parameters randomly sampled from the uniform distribution $s \sim [0, 0.3]$ and $h \sim [0, 1.0]$. The MAE of EMWER after excluding the estimates with top 70% confidence intervals of either selection

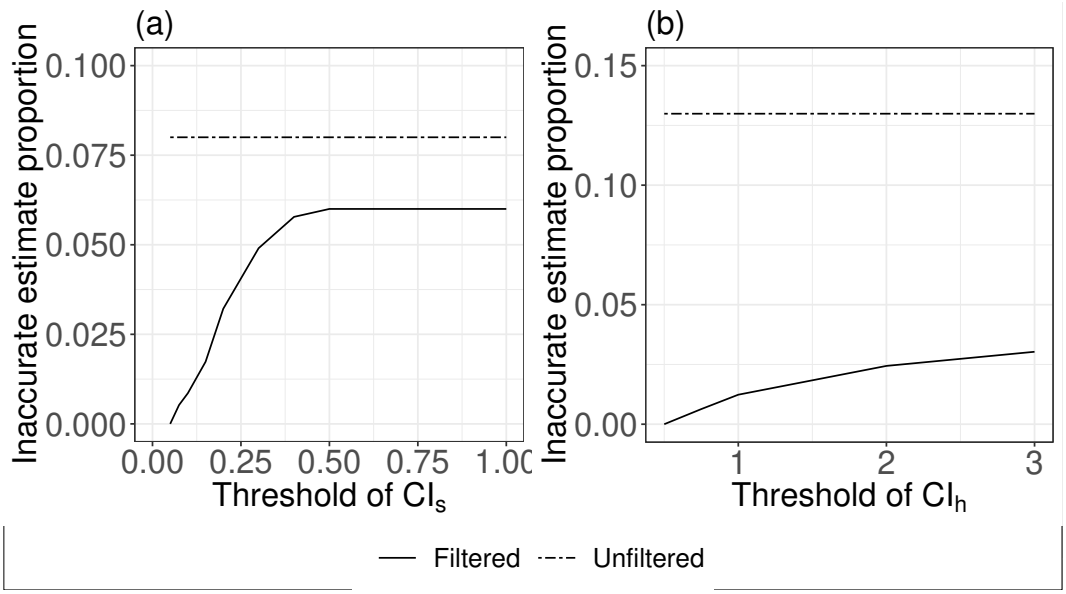


Figure 2.9: Proportion of the inaccurate estimates (i.e. $|\hat{s}-s| > 0.1 \vee |\hat{h}-h| > 0.5$) in the estimates whose confidence intervals (CIs) are under the varying threshold (solid lines). The threshold of selection CI (CI_s) and dominance CI (CI_h) are varied in (a) and (b) respectively. Only selection coefficients are estimated in (a). Selection coefficients and dominance parameters are simultaneously estimated, and the threshold of CI_s is fixed at 0.1 in (b). The true values were $(s, h) = (0.2, 0.5)$ in (a) and (b). Proportion of the inaccurate estimates in all estimates are plotted by dashed lines.

or dominance (0.140) was less than those of the other methods including the methods with the same filtering criteria (**Table. 2.2**). We confirmed that the accuracy of EMWER is better than those of the other methods for varying filtering criteria when we exclude more than about 75% or more of all the estimates (**Fig. 2.12-b**). On the other hand, the accuracy of [Taus et al., 2017] (0.415), [Iranmehr et al., 2017] (0.320) and EMWER before the filtering (0.339) was even inferior to [Foll et al., 2015] (0.231) and [Ferrer-Admetlla et al., 2016] (0.234), which estimate the dominance as 0.5 for the varying true dominance values (**Fig. 2.14**). This may be because the estimated dominance values are likely to be much larger than true values due to the flat surface of likelihoods, which is more distinct with weak selection. Indeed, under strong selection, the accuracy of EMWER and [Iranmehr et al., 2017] before the filtering are competitive to those of [Foll et al., 2015] and [Ferrer-Admetlla et al., 2016]. We note that [Taus et al., 2017] failed to estimate WF parameters and its CIs for many SNPs (86% and 89% respectively) in this simultaneous estimation experiment.

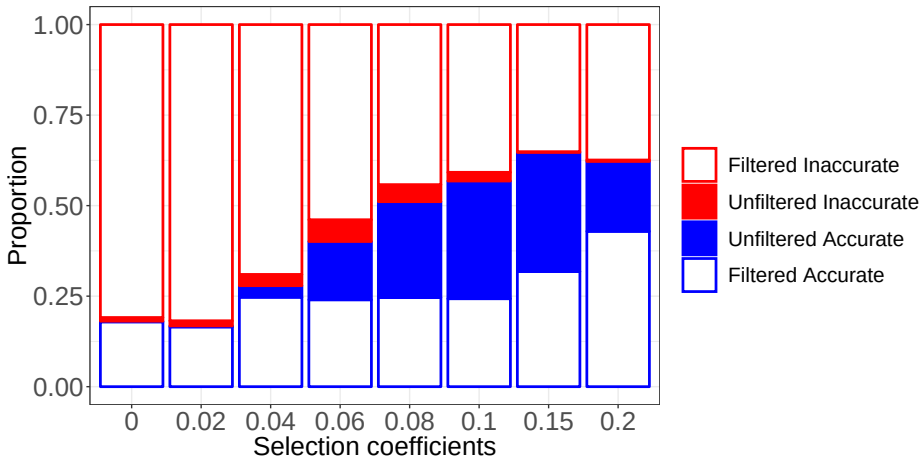


Figure 2.10: The performance of filtering for varying selection coefficients. We represents the proportion of estimates classified according to whether they are excluded by filtering ($CI_s > 0.1 \vee CIh > 1.0$) or not and whether they are inaccurate ($|\hat{s} - s| > 0.1 \vee |\hat{h} - h| > 0.5$) or not.

The execution time of EMWER was about 6 and 2 times larger than that of [Iranmehr et al., 2017] and [Taus et al., 2017] respectively, and 40 times less than [Ferrer-Admetlla et al., 2016] and [Foll et al., 2015], when we simultaneously estimated selection and dominance. One reason for the superiority of [Iranmehr et al., 2017] is that the grid points of selection and dominance in the estimation of [Iranmehr et al., 2017] are coarse grained in this experiment, while EMWER estimates continuous values for s and h. Hence, in some experimental conditions which require highly precise estimates, the execution time of [Iranmehr et al., 2017] can be larger than that of EMWER. These results show that EMEWR can provide the most accurate estimates of dominance in E&R data after filtering inaccurate estimates, while its computational efficiency is next to the top two efficient methods [Iranmehr et al., 2017, Taus et al., 2017] in our experimental conditions.

2.3.2.4 Performance of detecting selected SNPs

To compare our method with other methods used in E&R studies, we investigated the accuracies for detecting selected SNPs from simulated datasets. While the area under the curve (AUC) is very high for all four methods, our method and [Iranmehr et al., 2017] are distinctively better than the other

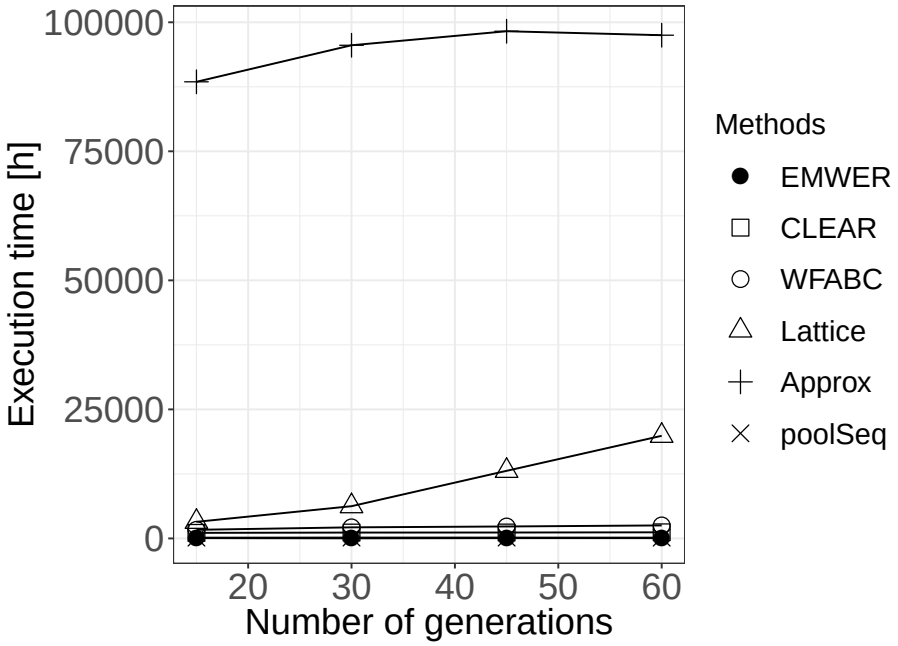


Figure 2.11: Comparison of performance in the selection estimation by EMWER with those estimated by poolSeq [Taus et al., 2017] , CLEAR [Iranmehr et al., 2017] , Lattice [Mathieson and McVean, 2013] and Approx [Ferrer-Admetlla et al., 2016]. The dependency of execution time on the interval between two sequencing time points are represented. The number of sequencing time points is two.

methods that are not based on the WF model in the parameter regions of low false positives, which are the regions of interest in most studies (i.e. the true positive rate of methods with WF model are about 10 % greater than those of the others when the significance level is set so that the false positive rate is 0.01) (**Table. 2.3**) (**Fig. 2.15**). Among the WF-based methods, the performance of [Iranmehr et al., 2017] is slightly better than our method (i.e. about 1% increases in the true positive rate). This is presumably because our method is based on the diffusion approximation of WF model, which is accurate for long period of generations, while [Iranmehr et al., 2017] is based on the discrete WF model, which is more close to the simulation model of MimicrEE.

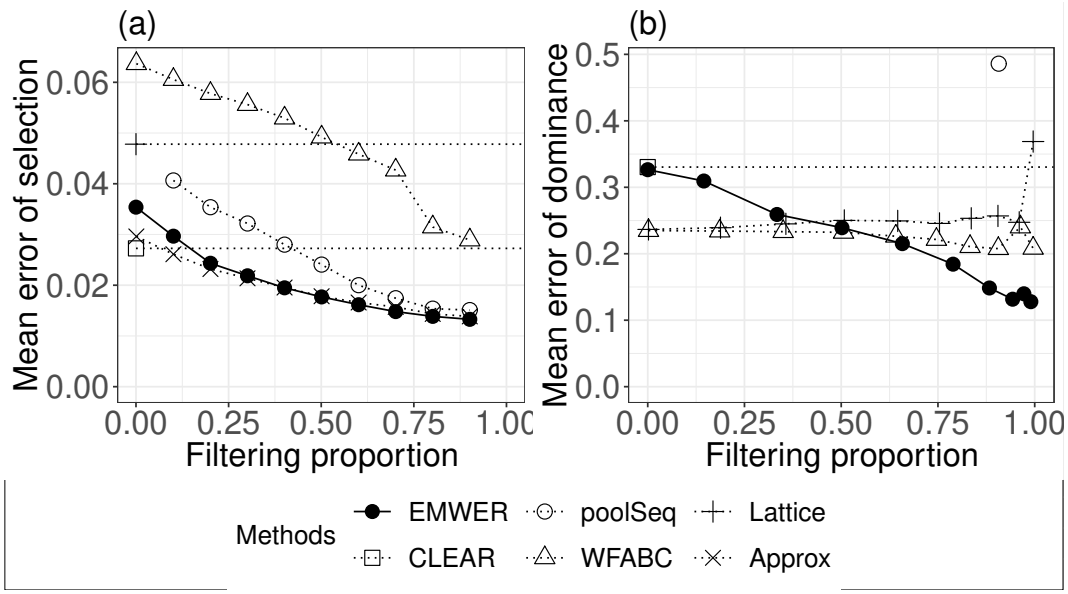


Figure 2.12: The estimation accuracy for varying filtering criteria. We represent the accuracy for EMWER with those estimated by CLEAR [Iranmehr et al., 2017], poolSeq [Taus et al., 2017], WFABC [Foll et al., 2015], Lattice [Mathieson and McVean, 2013] and Approx [Ferrer-Admetlla et al., 2016]. Although CLEAR and Lattice has no filtering methods, we represent the accuracy of these methods for the comparison. (a) We estimated selection coefficients, when the true selection coefficients are randomly sampled from the uniform distribution $s \sim [0, 0.3]$. We calculate the accuracy after excluding estimates with top $n\%$ of CI_s $(n = 0, 10, \ldots, 90)$. The x-axis represents the proportion of filtered estimates. The y-axis represents the MAE of selection coefficients, when the true selection coefficients and the true dominance parameters are randomly sampled from the uniform distribution $s \sim [0, 0.3]$ and $h \sim [0, 1.0]$. (b) We estimated simultaneously selection coefficients and dominance parameters. We calculate the accuracy after excluding estimates with top $n\%$ of either CI_s or CI_h $(n = 0, 10, \ldots, 90)$. For poolSeq, we represent the accuracy with only $n = 90$, because poolSeq failed to calculate CIs for 89% of SNPs. The x-axis represents the proportion of filtered estimates. The y-axis represents the MAE of dominance parameters.

We conducted the accuracy comparison of detecting selected SNPs under various conditions of selection coefficient, population size, number of generations, sequencing depth, number of sequencing, initial allele frequency, and number of replicates (**Fig. 2.16**) (**Fig. 2.17**). The performance of the BBGP-based test was relatively inferior when the number of individuals N was smaller (**Fig. 2.16-b**). This may be because the BBGP-based test assumes the common trajectory

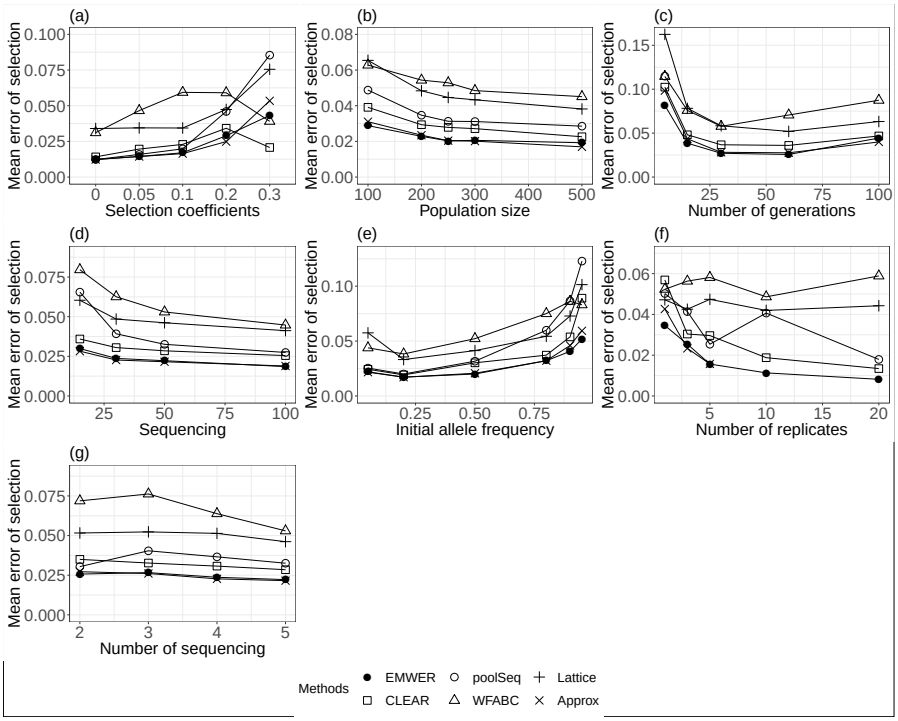


Figure 2.13: Comparison of accuracy in the selection estimation by EMWER with those estimated by CLEAR [Iranmehr et al., 2017], poolSeq [Taus et al., 2017], WFABC [Foll et al., 2015], Lattice [Mathieson and McVean, 2013] and Approx [Ferrer-Admetlla et al., 2016]. We evaluated the accuracy of estimation by the mean absolute errors (MAE). The accuracy of EMWER, poolSeq, WFABC and Approx are represented after filtering estimates with top 30% confidence intervals of selection. (a-d) The dependency of the accuracy on the selection coefficients , the population size, the interval between two sequencing time points, the sequence depth, the initial allele frequency, the number of replicates and the number of sequencing are represented. The sequence depth was set to 15 in (d). Except for (a), the true selection coefficients were sampled from the uniform distribution $s \sim [0, 0.3]$. In (b), we set the maximum number of allele frequency grids in EMWER to 200, because the larger number of individuals require the larger number of the grid points.

of allele frequencies for all replicates since the small number of individuals N are associated with strong genetic drift and violate the assumptions of the test. The performance of the CMH test was relatively inferior for large s values, which may be because it does not model the effects of fixation (**Fig. 2.16-a**). We have shown that the accuracy of the CMH test degrades for biased initial allele frequencies or large number of generations (**Fig. 2.16-c,e**). Furthermore, we investigated the influence of linkage on the accuracy of detecting selected SNPs. In stead of iterative simulations for single SNPs used in other sections, we conducted whole genome simulations for varying recombi-

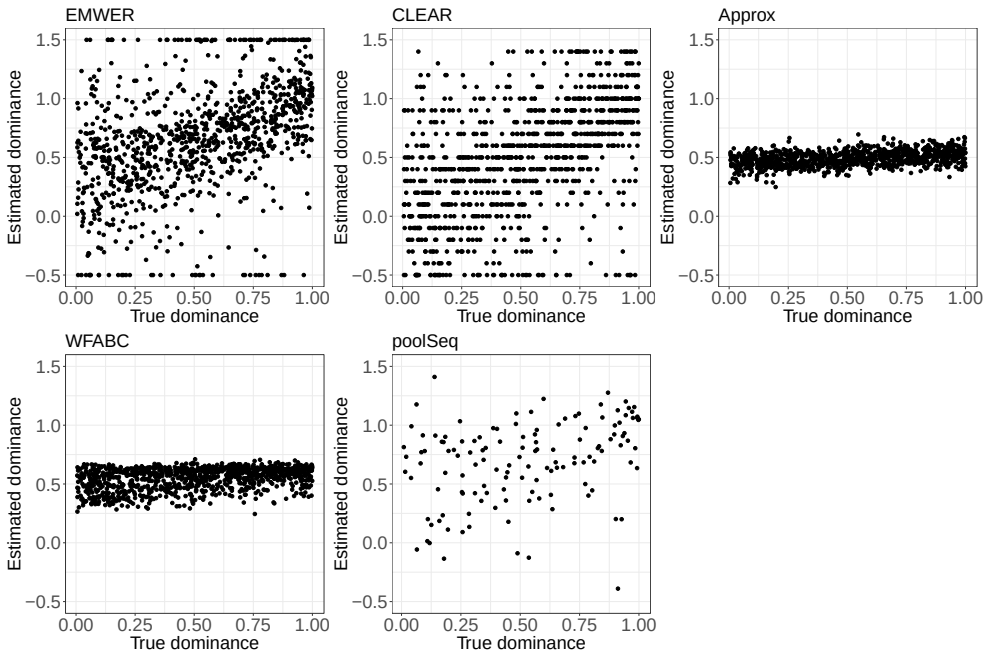


Figure 2.14: The correlation between estimated dominance (y-axe) and true dominance (x-axes). We represents the correlation for EMWER, [Taus et al., 2017] (poolSeq), [Iranmehr et al., 2017] (CLEAR), [Foll et al., 2015] (WFABC) and [Ferrer-Admetlla et al., 2016] (Approx).

nation rates. In this experiments, we simulate 100 selected loci and 9900 neutral loci, and measure the accuracy to detect the selected loci from all loci. We found that higher recombination rates increase the accuracy for all methods, while variability of the performance are relatively constant (**Fig. 2.18**). Additionally, we evaluated the detection accuracy when we excluded SNPs fixed or close to fixation at the end of E&R experiment. Specifically, we excluded SNPs whose mean allele frequencies x_{60} at 60th generation is close to the boundaries ($x_{60} < 0.05, x_{60} > 0.95$). We found that the accuracy of CMH test (0.954) is almost identical to the accuracy of EMWER (0.967) and CLEAR (0.973), but the methods based on the WF model (EMWER and CLEAR) are still superior to the other methods (CMH test and BBGP-based test) (**Table. 2.5**). This indicates that the largest advantage of using WF-based models in selection detection is the ability to capture the complicated dynamics of allele frequencies close to the boundaries, while their superiority is small in intermediate frequency ranges.

Relative to the non WF-based methods, our method and [Iranmehr et al., 2017] are consistently more accurate across the entire range of parameter values (**Fig. 2.16**) (**Fig. 2.17**) (**Fig. 2.18**) (**Table. 2.4**) (**Table. 2.5**), which is presumably because our method and [Iranmehr et al., 2017]

Method	AUC	TPR at FPR=0.01
EMWER	0.99	0.857
CLEAR	0.992	0.868
CMH	0.978	0.738
BBGP	0.985	0.734

Table 2.3: Accuracy of detecting selected SNPs for our method (EM algorithm of Wright–Fisher for E&R data; EMWER), the CLEAR [Iranmehr et al., 2017], the BBGP-based test [Topa et al., 2015] and the Cochran–Mantel–Haenszel (CMH) test [Agresti, 2002]. This table shows the area under the curve and the true positive rate (TPR) when the significance level is set so that the false positive rate (FPR) is 0.01.

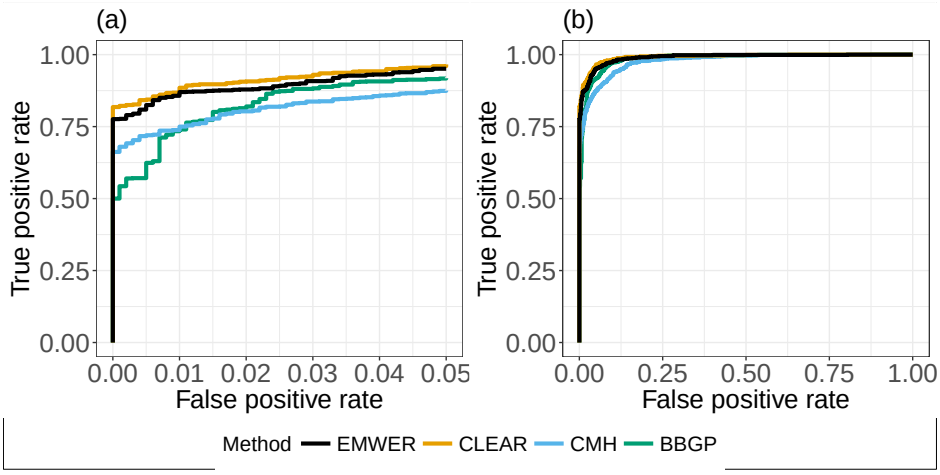


Figure 2.15: Receiver operator splitting (ROC) curve of detecting selected SNPs for our method (EM algorithm of Wright–Fisher for E&R data; EMWER), the CLEAR [Iranmehr et al., 2017], the BBGP-based test [Topa et al., 2015] and the Cochran–Mantel–Haenszel (CMH) test [Agresti, 2002].

consider the data generation processes more accurately. We concluded that, in terms of the accuracy of detecting selected SNPs, EMWER is similar to or slightly inferior to CLEAR, and both of them are consistently better than the CMH test and BBGP in the range of our experiments.

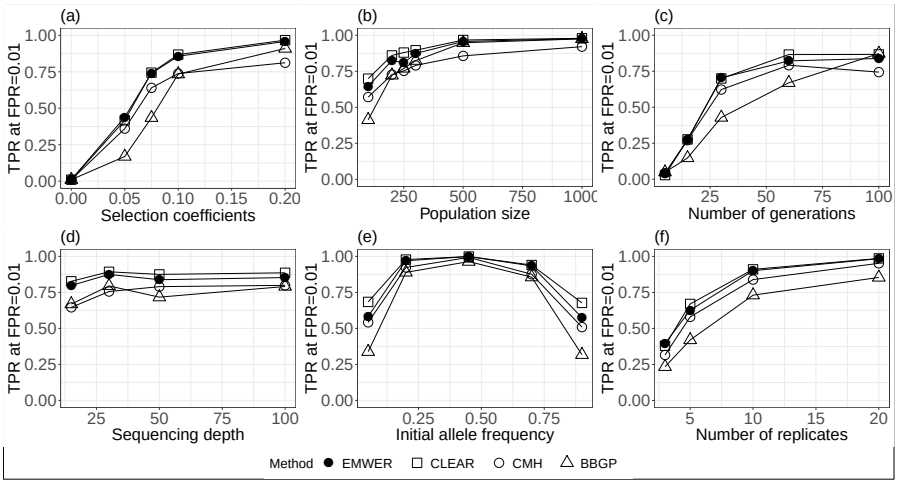


Figure 2.16: Comparison of the accuracies of detecting selected SNPs among three methods: our method (EMWER), the CLEAR, the BBGP-based test and the CMH test. Accuracies are measured as the true positive rate (TPR) when the significance level is set so that the false positive rate (FPR) is 0.01 for each method. (a-d) The dependency of the accuracy on the selection coefficients , the population size, the interval between two sequencing time points, the sequencing depth, the initial allele frequency and the number of replicates are represented. The sequence depth was set to 15 in (d).

2.3.3 Application to real E&R data

We applied our methods to data from a previous E&R study that investigated adaptation of *Drosophila* to a thermally fluctuating environment [OROZCO-terWENGEL et al., 2012]. We estimated the effective population size $\hat{N}_{\mathrm{e}}$ independently for each chromosome using randomly chosen SNPs with initial allele frequencies ($0.2 < x_{k0} < 0.8$), assuming most SNPs behave neutrally under artificial selection ($s = 0$). The estimated $\hat{N}_{\mathrm{e}}$ converges to around 200 as the number of SNPs is increased for all chromosomes except chromosome 3R (**Fig. 2.19**), for which the $\hat{N}_{\mathrm{e}}$ converges to about 115. The discordant estimate for chromosome 3R can be explained by the inferred large region under strong selection as shown below, which creates substantial allele frequency changes, resulting in an overestimate of genetic drift and an underestimate of the effective population size. On the other hand, the effective population sizes estimated by existing method [Jónás et al., 2016], which is included in the package of [Taus et al., 2017] , were larger (252(2L), 244(2R), 184(3L), 111(3R)

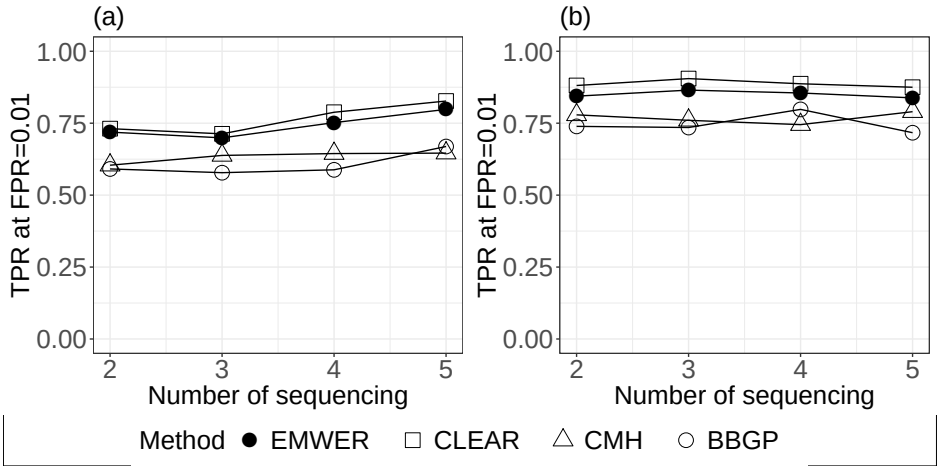


Figure 2.17: The accuracies of detecting selected SNPs among the methods (our method (EMWER), the CLEAR, the BBGP-based test and the CMH test) at read depths of 15 (left panel) and 50 (right panel). Accuracies are measured as the true positive rate (TPR) when the significance level is set so that the false positive rate (FPR) is 0.01 for each method. The x-axis is the number of sequencing time points.

and 270(X)) than our estimates (197(2L), 218(2R), 168(3L), 112(3R) and 208(X)) (**Fig. 2.19**). This is presumably due to the estimated variance of pooling process in E&R data, which is considered only in [Jónás et al., 2016]. Further, the estimation of [Jónás et al., 2016] is more stable to loci with biased initial allele frequencies (**Fig. 2.19**). This indicate that the restriction of the initial allele frequencies is important in our method (**Fig. 2.5**), but not in [Jónás et al., 2016]. Since the method by [Jónás et al., 2016] is more robust to biased initial allele frequencies, we used $\hat{N}_{\mathrm{e}} = 250$ in the following sections, because the estimated N_{e} by [Jónás et al., 2016] are distributed around 250 for chromosome 2L and 2R, which are expected to have relatively smaller background selection than other autosomes (3L and 3R) (**Fig. 2.20**).

Next, we estimated the selection coefficient s_k for each SNP k as well as the corresponding likelihood ratio statistic D_k as measures for the significance of selection acting on SNP k. In the following analysis, we use the estimates of SNPs with an initial allele frequency range $0.05 < x_{k0} < 0.95$, which corresponds to 1,361,739 SNPs out of 1,547,837 SNPs in total (88%). Although s_k is the

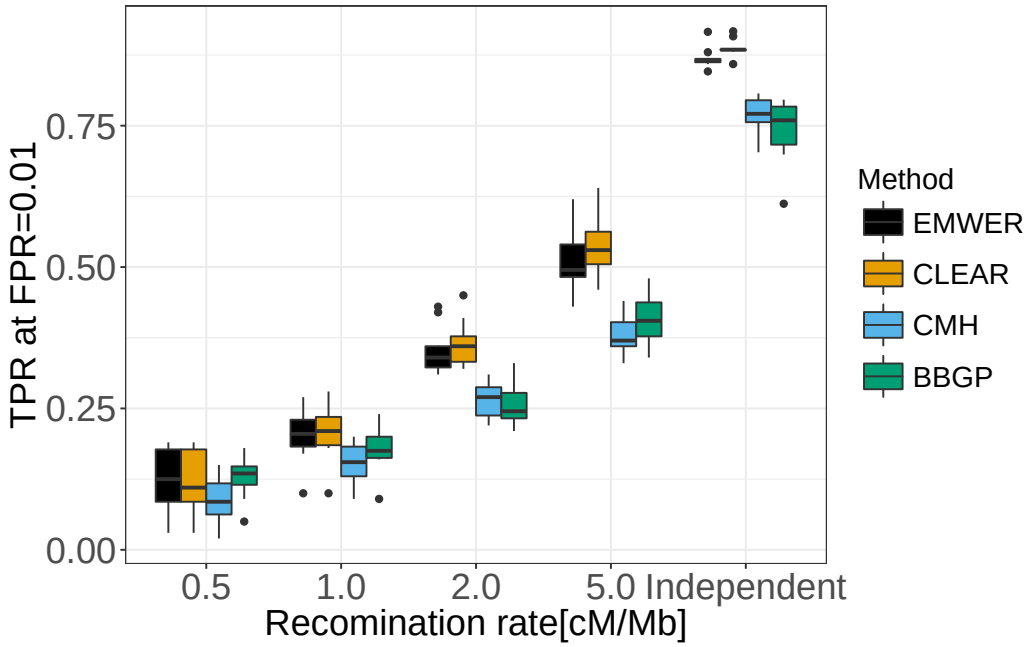


Figure 2.18: Comparison of the accuracies of detecting selected SNPs among three methods: our method (EMWER), the CLEAR, the BBGP-based test and the CMH test. Accuracies are measured as the true positive rate (TPR) when the significance level is set so that the false positive rate (FPR) is 0.01 for each method. The dependency of the accuracy on the recombination rate is displayed. In the condition "Independent", we conducted simulations for single SNP iteratively while simulations were conducted for whole genome in the other conditions.

selection coefficient for the variant allele of SNP k, in the following analysis, we do not distinguish between variant and reference alleles and convert negative s_k to selection coefficients corresponding to reference alleles as $s'_k = \frac{-s_k}{1+s_k}$, which must be positive. We conducted a GO enrichment analysis using Gowinda that accounts for gene length and linkage between multiple loci. We found the GO terms with the most significant P-values for EMWER (aminoacylase activity) were represented in the top 5 GO terms for the CMH test and the BBGP-based test, while the other GO-terms, all of which are related to proton-transporting ATPase complex, are not shared with the other

Dominance	Method	AUC	TPR at FPR 0.01
0	EMWER	0.989	0.888
0	CLEAR	0.989	0.906
0	CMH	0.975	0.714
0	BBGP	0.981	0.795
0.5	EMWER	0.99	0.841
0.5	CLEAR	0.993	0.884
0.5	CMH	0.978	0.762
0.5	BBGP	0.988	0.803
1.0	EMWER	0.877	0.531
1.0	CLEAR	0.876	0.53
1.0	CMH	0.854	0.52
1.0	BBGP	0.831	0.344

Table 2.4: Accuracy of detecting selected SNPs for EMWER, the CLEAR [Iranmehr et al., 2017], the BBGP-based test [Topa et al., 2015] and the Cochran–Mantel–Haenszel (CMH) test [Agresti, 2002] when the true dominance parameter ($h = 0, 1$) differs from the fixed value in our estimation ($h = 0.5$). This table shows the AUC and the true positive rate (TPR) when the significance level is set so that the false positive rate (FPR) is 0.01 for each method.

Method	Before exclusion	After exclusion
EMWER	0.844	0.967
CLEAR	0.884	0.973
CMH	0.742	0.954
BBGP	0.699	0.784

Table 2.5: Accuracy of detecting selected SNPs for EMWER, the CLEAR [Iranmehr et al., 2017], the BBGP-based test [Topa et al., 2015] and the Cochran–Mantel–Haenszel (CMH) test [Agresti, 2002] before and after we excluded SNPs whose mean allele frequencies x_{60} at 60th generation is close to boundary ($x_{60} < 0.05, x_{60} > 0.95$). This table shows the true positive rate (TPR) when the significance level is set so that the false positive rate (FPR) is 0.01 for each method.

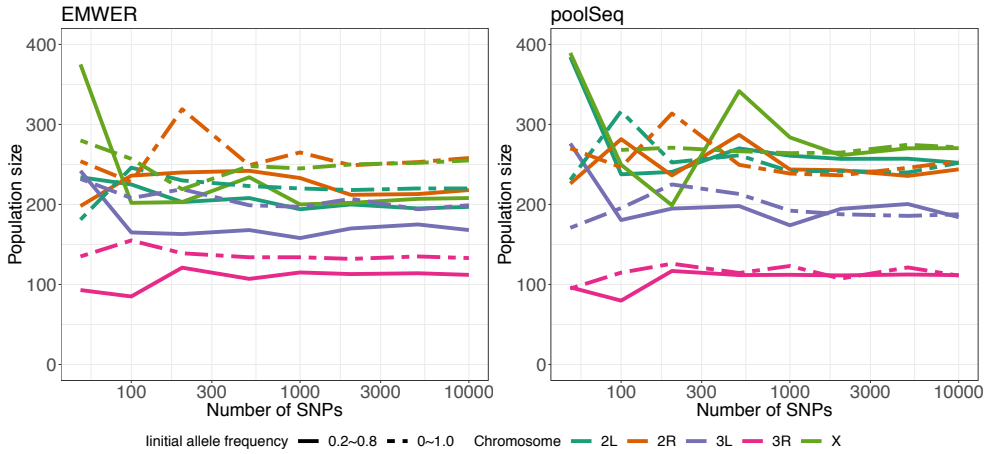


Figure 2.19: Estimated effective population size $\hat{N}_e$ computed for each chromosome by our method (EMWER, left) and ([Jónás et al., 2016], right). The x-axis represents the number of SNPs used. The solid lines represent the estimate for SNPs whose initial allele frequencies are between 0.2 and 0.8. The dashed lines represent the estimate for SNPs whose initial allele frequencies are between 0 and 1.0.

methods (**Table. 2.6**), although more than half of candidate SNPs (75.4%) are shared with either of other methods (**Fig. 2.21**). However, it is difficult to interpret the relationships between these gene sets and thermal adaptation, as it can be caused by hitchhiking effect of strong selection on cosmopolitan inversion described below. The composition of D_k across chromosomes (**Fig. 2.20**) shows that there exists a large proportion of SNPs with significant D_k (FDR < 0.05) on chromosome 3R (57.7%).

To study chromosome 3R further, we computed the mean estimated selection coefficients $\hat{s}_k$ with significant D_k (FDR < 0.05) averaged over 200-kbp windows for every 50-kbp. We can see an elevated level of selection strength in the *In(3R)P* inversion region (**Fig. 2.22**), which has been inferred to be related to climate change adaptation of natural populations [Rane et al., 2015].

Comparing the distributions of $\hat{s}_k$ for SNPs within *In(3R)P* and those within other regions of the whole genome, *In(3R)P* SNPs have a distinct peak around $s = 0.2$, which implies the allele frequencies of many SNPs within this region are driven by a common selective force, directly or

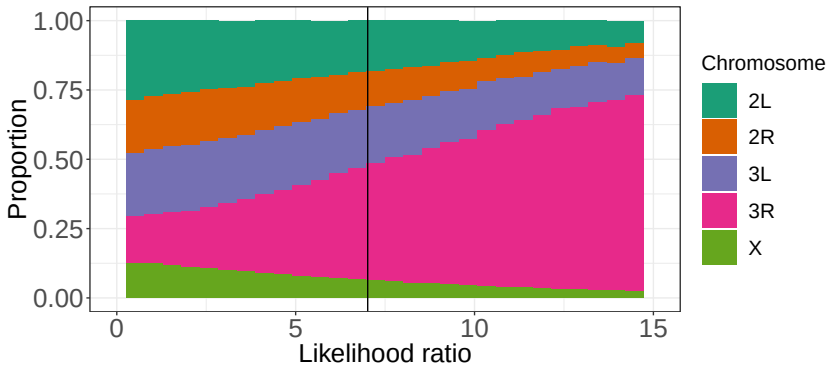


Figure 2.20: The proportion of each chromosome in bins with a specified likelihood ratio D_k. The x-axis represents the value of the specified likelihood ratio D_k. In this figure, we exclude chromosomes with fewer than 100,000 SNPs. The vertical line represents $D_k = 9.11$ (FDR < 0.05)

Method	GO.ID	Term	P.value	FDR
EMWER	GO:0004046	aminoacylase activity	1.3e-05	0.021825
EMWER	GO:0033179	proton-transporting V-type ATPase, V0 domain	3.5e-05	0.021825
EMWER	GO:0070070	proton-transporting V-type ATPase complex assembly	3.5e-05	0.021825
EMWER	GO:0070072	vacuolar proton-transporting V-type ATPase complex assembly	3.5e-05	0.021825
EMWER	GO:0033177	proton-transporting two-sector ATPase complex, proton-transporting domain	5.1e-05	0.025541
CMH	GO:0046914	transition metal ion binding	1e-05	0.02347
CMH	GO:0004046	aminoacylase activity	5e-05	0.05224667
CMH	GO:0045089	positive regulation of innate immune response	1e-04	0.05224667
CMH	GO:0005506	iron ion binding	0.00011	0.05224667
CMH	GO:1902600	hydrogen ion transmembrane transport	0.00013	0.05224667
BBGP	GO:0004046	aminoacylase activity	6e-05	0.13358
BBGP	GO:0044437	vacuolar part	3e-04	0.13358
BBGP	GO:0046914	transition metal ion binding	0.00037	0.13358
BBGP	GO:0065010	extracellular membrane-bounded organelle	0.00049	0.13358
BBGP	GO:1903561	extracellular vesicle	0.00049	0.13358

Table 2.6: Tables of Gene Ontology (GO) terms with the top five P-values calculated for significant SNPs identified by EMWER, the CMH test and the BBGP-based test.

indirectly due to hitchhiking effect driven by other selected loci (**Fig. 2.23**). On the other hand, such a difference cannot be seen in the distribution of significance measures of EMWER, the CMH test or the BBGP-based test (**Fig. 2.24**). In short, we detected a peak in the selection coefficient

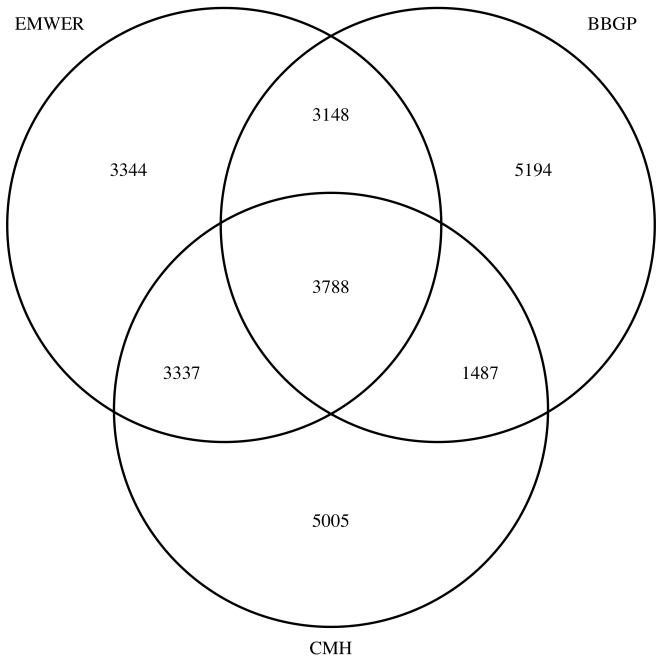


Figure 2.21: Overlaps of candidate SNPs (top 1% significant loci) among EMWER, CMH test, and BBGP-based test.

s_k within *In(3R)P* , which cannot be seen in the distribution of the significance statistic D_k.

To study the nature of the SNPs exhibiting *In(3R)P* -specific selection ($s_k \approx 0.2$), we compared these SNPs with annotated *In(3R)P* marker SNPs that were identified in a previous study and used for inferring the inversion frequency [Kapun et al., 2014]. We found that the estimated selection coefficients of the marker SNPs are also distributed around $s_k = 0.2$ with a median of 0.205 (**Fig. 2.23**). It indicates that the allele frequency paths of SNPs with *In(3R)P* -specific selection are very similar to those of the marker SNPs. While an excess of SNPs exhibiting *In(3R)P* -specific selection of $|s_k - 0.2| < 0.02$ is observed in the whole *In(3R)P* region, we also found three distinct peaks in the proportion of SNPs with *In(3R)P* -specific selection (**Fig. 2.25**); one is close to the proximal breakpoints of *In(3R)C* and *In(3R)Mo*, while two are close to the proximal and distal breakpoints of *In(3R)P* , where many *In(3R)P* -specific SNPs are expected to exist due to repressed

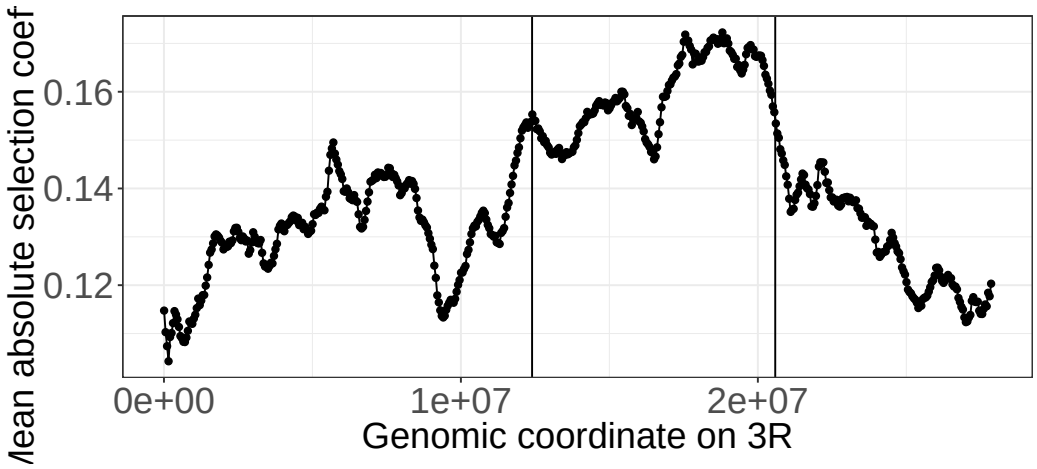


Figure 2.22: At every 50-kbp interval, we computed the mean estimated selection coefficients $\hat{s}_k$ with significant D_k (FDR = 0.05) for SNPs within each 1-Mbp window. The vertical lines represent the *In(3R)P* inversion breakpoints.

recombination. This result suggests that the recombination between chromosomes with and without *In(3R)P* is repressed not only around the *In(3R)P* breakpoints but also around those of *In(3R)C* or *In(3R)Mo*.

Finally, we simultaneously estimated selection coefficients and dominance parameters from the same data. Since the number of replicates in this experiment is rather small, we obtained only 1,181 estimates with relatively small CIs ($CI_s \leq 0.1, CI_h \leq 1.0$) within parameter ranges ($0 < s < 0.32, -0.5 < h < 1.5$), which corresponds to 0.08% of the total 1,547,837 SNPs. We found that the estimated dominance parameter values are distributed around $h = 1$ (**Fig. 2.26**), while those are distributed around $h = 0.5$ in simulation data mimicking the real data ($g = 0, 17, 23, 37, M = 3$) i.e) the selection coefficients were sampled from the empirical distribution estimated by EMWER assuming $h = 0.5$, the dominance parameters were sampled from uniform distribution $h \sim [0, 1.0]$, and the initial allele frequencies were sampled from the empirical distribution in the real data. As shown in **Fig. 2.27**, such beneficial dominant alleles show slowdown of allele frequency changes in later generations, which is very similar to the trend seen by [OROZCO-terWENGEL et al., 2012]. We confirmed that this slowdown can be reproduced by the simulation mimicking these SNPs (i.e. selection coefficients were set to the average values over the SNPs ($s = 0.13$), the dominance parameters were set to 1, and initial allele frequencies were sampled from uniform distribution

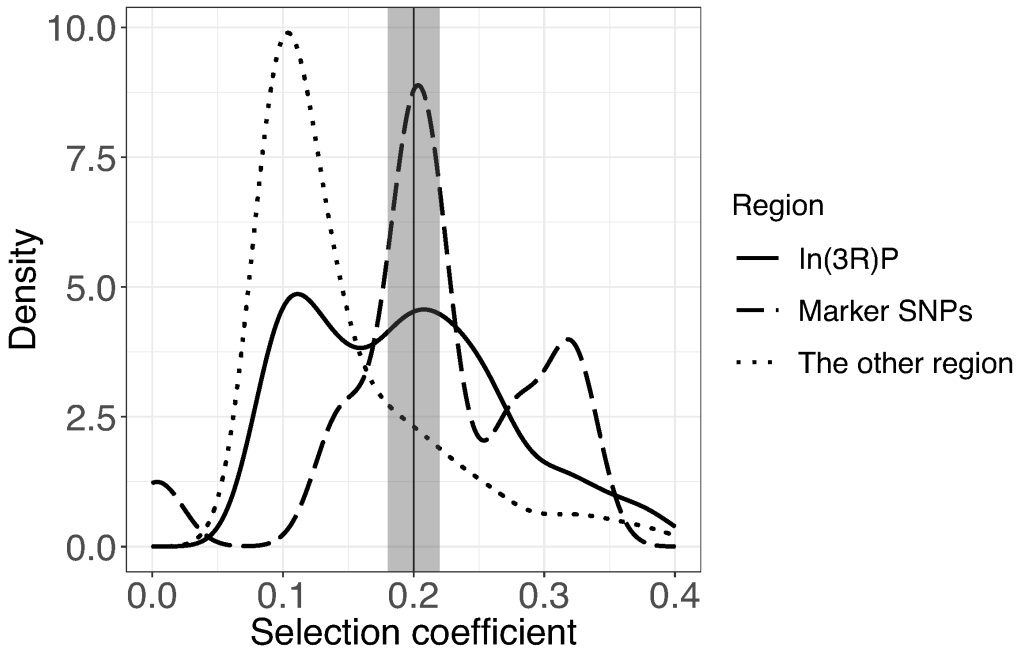


Figure 2.23: Distributions of selection coefficient s_k of SNPs within the *In(3R)P* inversion (solid line), *In(3R)P* inversion marker SNPs (dashed line), and SNPs within the remainder of the whole genome (dotted line). The vertical line represents $s = 0.2$, and the shading around the line represent the region $|s - 0.2| < 0.02$. The estimation which reached the boundary (0.41) were excluded in this figure.

between 0.2 and 0.5, which is similar to that of the SNPs). This slowdown of allele frequencies was described as 'plateauing' in original research [OROZCO-terWENGEL et al., 2012], and our observation indicate that the dominance of beneficial alleles might contribute to this trend.

2.4 Discussion and conclusion

We have developed a novel algorithm for estimating population genetic parameters (i.e. selection coefficient s, dominance parameter h and effective population size N_e) from time-course resequencing data. We further introduced a likelihood ratio test for detecting selected SNPs and CIs for each estimated value. We showed that our method could determine selection coefficients more efficiently than other WF optimization methods, while the accuracy was almost identical to the most accurate method after filtering inaccurate estimates based on inferred CIs. By the filtering, the accuracy of dominance estimation was also superior to those of the other methods. Furthermore, a comparison

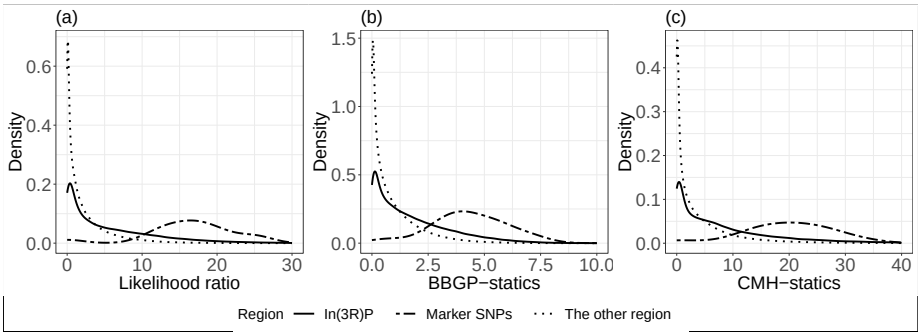


Figure 2.24: (a) Distribution of the likelihood ratio D_k (a), the significance statistics of the BBGP-based test (b) and that of the CMH test (c) among SNPs within the *In(3R)P* inversion (solid line), among *In(3R)P* marker SNPs (dashed line) and among SNPs within the remainder of the whole genome (dotted line).

with previous methods used in E&R studies showed WF-based methods including our algorithm was superior to non WF-based methods in detecting selected SNPs at thresholds with low false positive rates. However, we note that the performance difference between the methods may be dependent on the simulation settings, as they have so many parameters and it is difficult to test in all possible combinations.

The computational efficiency of our algorithm allowed us to investigate complex dependencies of estimation accuracy on internal parameters and experimental conditions. Our observations are not only consistent with traditional results of the WF model, but they also develop new insights from the WF model that may enhance data science and machine learning approaches. These useful insights include characterizations of the interactions between finite read depth and the number of sampling points, for example. We expect that our results can be useful for optimizing the experimental designs of E&R studies.

Our analyses of selection coefficients and likelihood ratios strongly suggested there is a common selection that affected the allele frequencies of many SNPs within the *In(3R)P* inversion. This observation is consistent with the significant reduction of *In(3R)P* frequency in population [Kapun et al., 2014], which was suggested using a small number of marker SNPs. On the other hand, there are no qualitative difference between the inside and outside of *In(3R)P* in terms of the distributions of statistical significance measures for detecting selection (**Fig. 2.24**). Indeed, in the original E&R

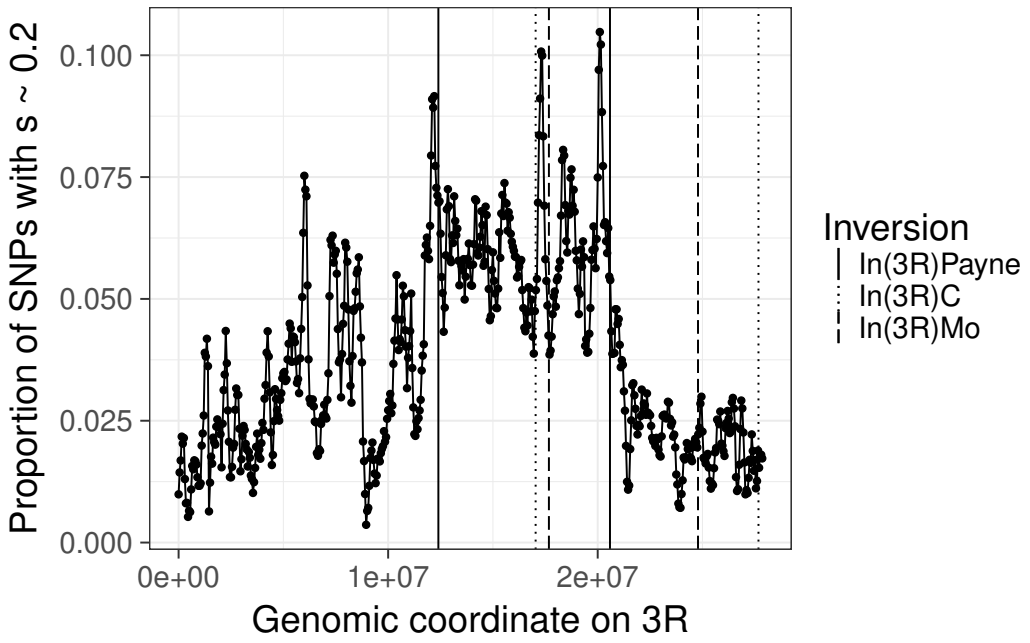


Figure 2.25: Local proportion of SNPs with a selection coefficient $|s_k - 0.2| < 0.02$. At every 50-kbp interval, the proportion were computed for SNPs within each 200-kbp window. The vertical lines represent the break points of the *In(3R)P* inversion (solid line), *In(3R)C* inversion (dotted line) and *In(3R)Mo* inversion (dashed line).

study, which used the CMH test, it was concluded that *In(3R)P* itself was unlikely to be the cause of the high number of significant SNPs on chromosome 3R [OROZCO-terWENGEL et al., 2012]. This demonstrates the value of estimating selection coefficients s_k in addition to significance measures such as the likelihood ratio D_k, as they contain different evolutionary information.

Our further analysis of the *In(3R)P* region suggests that, the SNPs specific to *In(3R)P* chromosomes are distributed not only around the breakpoints, but also throughout the *In(3R)P* region. This observation is consistent with a previous study in which the authors found that in a natural North American population SNPs associated with *In(3R)P* are distributed across the entire *In(3R)P* region, while many other cosmopolitan inversions have specific SNPs only around their breakpoints [Kapun et al., 2016]. The elevated local density of *In(3R)P* -specific SNP candidates not only around *In(3R)P* breakpoints but also around those of *In(3R)C* or *In(3R)Mo* is also consistent with the previous research [Kapun et al., 2016]. Furthermore, our analysis suggests that the local density of *In(3R)P* -specific SNPs at *In(3R)Mo* and *In(3R)C* breakpoints is approximately equivalent to the density of SNPs at *In(3R)P* breakpoints. This observation implies that there may

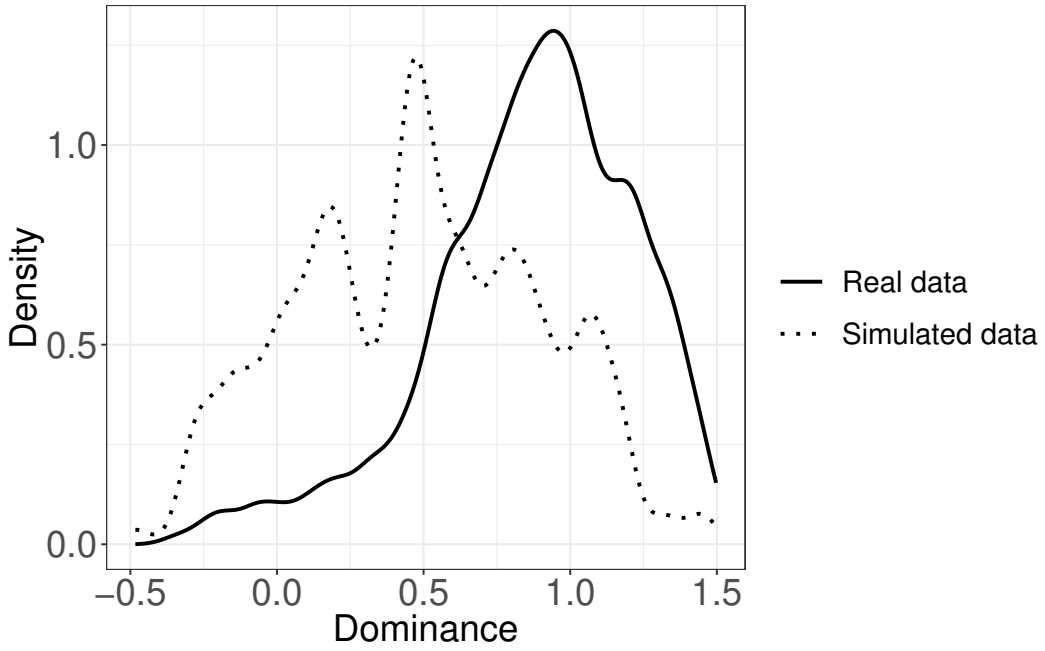


Figure 2.26: Distribution of estimated dominance parameters (h_k) with relatively high reliability, i.e. $CI_s < 0.1, CI_h < 1.0$. Solid and dotted lines represent the dominance parameters estimated from real E&R data [OROZCO-terWENGEL et al., 2012] and similar simulated E&R data ($g = 0, 17, 23, 37, M = 3$). The selection coefficients of the simulation data were sampled from the empirical distribution estimated by EMWER assuming $h = 0.5$, the dominance parameters were sampled from uniform distribution $h \sim [0, 1.0]$, and the initial allele frequencies of the simulation data were sampled from the empirical distribution in the real data.

exist some mechanical connection between the suppression of recombination around the *In(3R)P* breakpoints and the suppression of recombination around the *In(3R)Mo* and *In(3R)C* breakpoints.

We assumed a site independent model in this study like the other methods tested, and it does not model detailed linkage structure. This is because the accurate estimation of linkage parameters is quite difficult in general and especially for Pool-seq data which do not provide individual genotype or haplotype information. We independently estimate selection coefficient and dominance from allele frequency tragectories for each site. It indicates that we cannot distinguish physically selected sites and hitchhiking sites, but we should observe similar estimated selected coefficients and dominances among strongly linked sites as discussed above. Thus, our software can be used for the analysis of linkage structure in a data-driven manner even though it does not model linkage structure in itself.

Dominance is an important factor in population genetics, but there has been few reports about

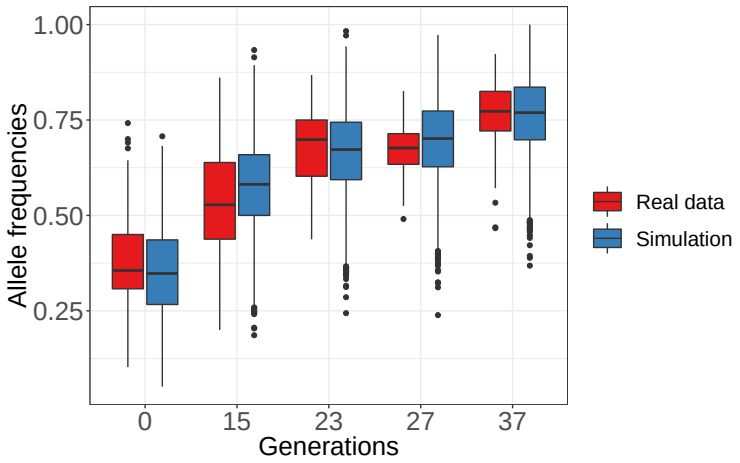


Figure 2.27: Allele frequency distribution of dominant alleles (i.e. $|h - 1.0| < 0.1, CI_s < 0.1, CI_h < 1.0$) at each sequencing time point (red). The allele frequency distribution of simulation data mimicking these SNPs are also represented (blue).

dominance distribution of standing variation at a genome-wide scale [Huber et al., 2018]. Previously, sevral studies have suggested that artificially introduced mutations are often deleterious recessive mutations [Agrawal and Whitlock, 2011, Manna et al., 2011]. A more recent theoretical analysis has suggested that a portion of deleterious recessive mutations have a longer waiting time prior to extinction relative to neutral mutations [Mafessoni and Lachmann, 2015]. Our analysis indicates that standing variation has similar properties; the estimated dominance parameter values were distributed around 1.0, suggesting many selected SNPs have either dominant beneficial alleles or recessive deleterious alleles. This observation is also consistent with the recent investigation of dominance within the natural population of *Arabidopsis* [Huber et al., 2018].

Our EM algorithm provides a general framework for inferring parameters of the stochastic differential equation $dX_t = \mu_\theta(X_t)dt + \phi^2_\theta(X_t)dW_t$, whose transition density can be expressed as a solution of the KE. However, this method can be computationally difficult to apply to multivariate cases since it requires the eigen decomposition of the rate matrix R, whose row and column dimensions correspond to the number of grid points in the multi-dimensional state space. Still, we expect this approach has various applications for fitting stochastic differential equations in theoretical and

mathematical biology to experimental data.

Chapter 3

Reconstructing spatiotemporal gene expression pattern during embryogenesis

3.1 Introduction

Embryogenesis produces a specific shape of each species from a single fertilized egg by repeated cell division and cell movement. In spite of its dynamic nature, this process is also accurate and reproducible. In the case of zebrafish, the shape of the embryo can almost be determined by the number of hours post fertilization (hpf) given a specified temperature [Kimmel et al., 1995]. To achieve this accuracy, the behavior of cells in each spatiotemporal context must be accurately specified.

One large element determining the cellular behavior is the expression of numerous genes. Recent technological advancement of single cell RNA-seq (scRNA-seq) provides us heterogeneous expression profile of each cell in various situations [Hwang et al., 2018]. Farrell *et al.* applied scRNA-seq to the whole embryos of zebrafish at 12 time points during early embryogenesis, and derived the temporal dynamics of gene expression during the differentiation into dozens of cell types [Farrell et al., 2018]. On the other hand, spatial contexts of each cell in scRNA-seq are not basically accessible due to tissue dissection into single cells before sequencing, although the region where each gene is expressed have important roles in embryogenesis. For example, specific proteins secreted from some tissues contribute to the differentiation of proximal tissues [Roelink et al., 1994].

Spatial patterns of gene expression during embryogenesis have been obtained by several techniques such as *in situ* hybridization (ISH) [Thisse et al., 2004], but it was difficult to observe the spatial patterns of numerous genes at one time. However, recent comprehensive observation technology also has brought about radical changes to the observation of the spatial gene expression patterns [Chen et al., 2015, Satija et al., 2015, Junker et al., 2014]. Satija *et al.* restored the positional information of each cell in scRNA-seq by using the spatial expression patterns of multiple genes obtained by ISH. In this approach, they evaluated the similarity of gene expression profiles between each scRNA-seq cells and each region within the integrated spatial patterns made from ISH observation, and projected the scRNA-seq cells onto the regions that resemble them. This approach succeeded in reconstructing the spatial patterns of tens of thousands of genes in shield stage of zebrafish embryogenesis. Junker *et al.* derived spatial transcriptome data by directly annotating the transcriptome data with spatial information, which they termed as tomo-seq. They sliced zebrafish embryos along multiple embryonic axes, and conducted a transcriptome after tagging RNA with unique barcode for each slice . Then, they derived the expression patterns of numerous genes along each axis, and reconstructed the original spatial expression patterns by combining them computationally.

These technologies now enable us to access spatial expression patterns of tens of thousands of genes. However, most of them have been performed in a few time points, and there are probably many developmental events within the observation intervals, which are not captured by the spatial transcriptome observation. Furthermore, the spatial patterns of each stage have been analyzed separately, presumably because extensive cell migration make it difficult to distinguish spatial pattern changes caused by transcriptional regulation from those caused by cell migration. Hence, it has veiled the generation process of the spatial expression patterns of numerous genes during embryogenesis. On the other hand, the cell migration during embryogenesis is now possible to be captured three-dimensionally by advanced optical technologies. Keller *et al.* tracked the cell movement of zebrafish embryo from 100 minutes to 24 hours after fertilization using digital scanned laser light sheet fluorescence microscopy [Keller et al., 2008]. Like this study, the movements of the cells in the early stage of embryogenesis were comprehensively obtained for several model organisms [Amat et al., 2014, McDole et al., 2018].

In this study, we integrated scRNA-seq data and spatial transcriptome data with cell movement data, and reconstructed the time-continuous spatiotemporal gene expression patterns during early stages of zebrafish embryogenesis. Our methodology for this estimation consists of two processes. In

the first process, a MAP-EM algorithm reconstructs positions of scRNA-seq cells and spatial gene expression patterns at the time points when the transcriptome observations are conducted. In the second process, Gaussian process regression predicts gene expression levels at any spatiotemporal points during embryogenesis described in the cell movement data. We confirmed that our method correctly reconstructed simulated spatiotemporal gene expression patterns, and recapitulated real anatomical structures of zebrafish embryos in real data application. Furthermore, we investigated the spatiotemporal dynamics of cell type differentiation by analyzing the estimated spatiotemporal gene expression patterns. The result of this analysis suggested that the boundary region of midbrain and hindbrain leads the differentiation process into both of them.

3.2 Methods

We modeled the spatiotemporal gene expression patterns during embryogenesis as Gaussian process with a expected correlation of the gene expression between cell movement cells. In particular, we defined the correlation based on the phylogenetic relationships of cells, time points and spatial coordinates. We also construct observation model of tomo-seq data and sc-RNA-seq, which have spatial information and more frequent temporal observation respectively. Since scRNA-seq has no spatial information, we assumed the original positions of scRNA-seq cells as hidden variables. Using MAP-EM algorithm, we estimated the maximum a posteriori (MAP) estimates of the spatial expression patterns at the observation time points and the expected positions of scRNA-seq cells given the MAP estimates of the spatial patterns. The MAP estimate of the spatial patterns at the observation time points enable us to construct the posterior distribution of the spatial gene expression patterns at any time points within cell movements data by Gaussian process regression.

In the following sections, we describe a probabilistic model of gene expression on cell movement data and an observation model of transcriptome data of single cells and sliced tissues with spatial annotation. Next, we introduce a MAP-EM algorithm for estimating the original location of the scRNA-seq cells and the spatial gene expression patterns at the observation time points, and describe Gaussian process regression for reconstructing spatial expression patterns at any time points. At the end of this section, we introduce generation procedure of simulation data set for validation of our method, and describe methods of acquisition and prepossessing of the real data used in this study.

3.2.1 Probabilistic model of spatiotemporal gene expression

During embryogenesis, there are many genes whose expression is localized to particular regions during particular duration of embryogenesis. Therefore, we modeled the spatiotemporal expression pattern of gene $g \in \{1, \ldots, G\}$ as a Gaussian process f_g indexed by cell movement cells where spatiotemporally more proximal cells have stronger correlation of gene expression levels.

$$\begin{aligned} f(i) &\sim GP(0, k(i,j)) \\ k(i,j) &= \sigma_f^2 \exp(-D_{st}(i,j)^2) \end{aligned}$$

where $D_{st}(i,j)$ represents a spatiotemporal distance between cell $i, j \in \{1, \cdots N\}$ in cell movement data, and N represents the total cell number of cell movement data. We quantified D_{st} as the square root of the square sum of spatial distance and temporal distance.

$$D_{st}(i,j) = \sqrt{\frac{D_t(i,j)^2}{\tau^2} + \frac{D_s(i,j)^2}{l^2}} \tag{3.1}$$

where τ and l are hyper parameters which correspond to the temporal and spatial correlation length of the gene expression. The temporal distance $D_t(i,j)$ is defined as $|t_i - t_j|$ where the time coordinates of cell i and j are t_i and t_j. When we evaluated the spatial distance between cells at different time coordinates, we compared the spatial location between the earlier stage cells and the ancestor cells of the later stage cells at the earlier time coordinates to consider the expectation that the gene expression also migrate along with the cell migration. Furthermore, after the gastrularation, it is necessary to pay attention to the fact that the deviation between the expression differences and the spatial distances tends to be significant due to tissue folding. Therefore, we considered the location of each cell at both the stage we are interested in and the earlier embryonic stage (50% epiboly, $t = 0$). Cosidering these two points, we defined spatial distance $D_s(i,j)$ as below:

$$D_s(i,j) = \sqrt{\frac{\|x_{I(i,0)} - x_{I(j,0)}\|_2^2 + \|x_{I(i,\min(t_i,t_j))} - x_{I(i,\min(t_i,t_j))}\|_2^2}{2}} \tag{3.2}$$

where x_i is a spatial coordinate of cell i in cell movement data, and $I(i,t)(t \leqq t_i)$ is a projection from cell i in cell movement data to a cell which is ancestor of cell i at time t. The projection I is recursively calculated as below:

$$I(i,t) = \begin{cases} i & (t = t_i) \\ \mathrm{argmin}_{j \in \{j' | t_{j'} = t\}} \|x_j - x_{I(x_i, t+1)}\|_2 & (t < t_i) \end{cases} \tag{3.3}$$

Here, we consider the expression patterns at observation time points $T^{(\mathrm{ref})}$ of spatial and single cell transcriptome data. To prevent huge calculation along with the increase of observation time points, we sampled a constant number $N^{(\mathrm{ref})}$ of cells $R_t \subset \{i|t_i = t\}$ at each observation time $t \in T^{(\mathrm{ref})}$, and constructed a subset of cell movement cells at the observation time points as $R = \bigcup_{t \in T^{(\mathrm{ref})}} R_t$. From the definition of Gaussian process, we formulated the prior distribution of gene expression of cells in R as below:

$$P(F_{R,*g}) = N(F_{R,*g}|0, K) \tag{3.4}$$

where $F_R \in \mathbb{R}^{|R| \times G}$ is the expression levels of cells in R, $F_{R,rg}$ represents the expression of gene g in cell $c(r)$ in cell movement data, $c(r)$ is a projection from an index in R to that in cell movement data, and K is a matrix with r, s elements as $k(c(r), c(s))$. In general, we denote the i-th row and the j-th column of a matrix X as X_{i*} and X_{*j} respectively.

3.2.2 Observation models of spatial and singel cell gene expression data

The expression level of each slice of tomo-seq is expected to be proportional to the sum of the expression level of cells contained in the slice. Hence, we assumed the expression level of gene g at slice k follows a Gaussian distribution whose variance and mean are a hyperparameter $\sigma^{(\mathrm{sl})2}$ and the sum of gene expression levels over cell set C_k which represents cells included in slice k.

$$P(Y^{(\mathrm{sl})}_{kg}) = N(Y^{(\mathrm{sl})}_{kg}| \sum_{i \in C_k} F_{R,c^{-1}(i)g}, \sigma^{(\mathrm{sl})2}) \tag{3.5}$$

Here, we formulated the predefined association between the slice indexes of tomo-seq and the cell indexes in cell movement data as a matrix A where

$$A_{kr} = \begin{cases} 1 & (c(r) \in C_k) \\ 0 & (\text{otherwise}). \end{cases}$$

Using this formulation, a linear transformation of expression levels at observation time $A_{k*}F_{R,*g}$ represents the mean of the Gaussian distribution (3.5).

The expression level of each scRNA-seq cell is assumed to be observed from any cell in the cell movement data at the observation time of the scRNA-seq cell. If the expression level of cell l in scRNA-seq is observed from cell $c(r) \in R$ in cell movement data, we assume that the expression level of gene g in cell l in scRNA-seq follows Gaussian distribution with mean $F_{R,rg}$ and variance $\sigma^{(\mathrm{sc})2}$ as below:

$$P(Y^{(\mathrm{sc})}_{lg}) = N(Y^{(\mathrm{sc})}_{lg}|F_{R,rg}, \sigma^{(\mathrm{sc})2}) \tag{3.6}$$

where $Y_{lg}^{(\mathrm{sc})}$ represents the expression level of gene g in scRNA-seq cell l. However, the actual correspondence between the scRNA-seq cells and the cell movement cells is not accessible, since the positional information of each scRNA-seq cell has been lost. Therefore, we formulated the correspondence as a hidden variable Z where

$$Z_{lr} = \begin{cases} 1 & (\text{scRNA-seq cell } l \text{ is observed from cell movement cell } c(r)) \\ 0 & (\text{otherwise}). \end{cases}$$

Prior distribution of Z is assumed to be categorical distribution so that each scRNA-seq cell at t is observed from one cell movement cell randomly selected out of R_t, whose size is $N^{(\mathrm{ref})}$. Hence, the formulation of the distribution is as below:

$$P(Z_{lr} = 1) = \begin{cases} \frac{1}{N^{(\mathrm{ref})}} & (t_{c(r)} = t_l^{(\mathrm{sc})}) \\ 0 & (t_{c(r)} \neq t_l^{(\mathrm{sc})}). \end{cases} \tag{3.7}$$

where $t_l^{(\mathrm{sc})}$ is the observation time of scRNA-seq cell l. Combining (3.6) and (3.7), the joint distribution of $Y^{(\mathrm{sc})}$ and Z given F_R is formulated as below:

$$\begin{aligned} P(Y^{(\mathrm{sc})}, Z|F_R) =& \prod_{l,r} \left(P(Z_{lr} = 1) \prod_g N(Y_{lg}^{(\mathrm{sc})}|F_{R,rg}, \sigma^{(\mathrm{sc})2}) \right)^{Z_{lr}} \\ =& \left(\frac{1}{N^{(\mathrm{ref})}} \right)^L \prod_{l, i \in R_{t_l^{(\mathrm{sc})}}} \left(\prod_g N(Y_{lg}^{(\mathrm{sc})}|F_{R,c^{-1}(i)g}, \sigma^{(\mathrm{sc})2}) \right)^{Z_{lr}}. \end{aligned} \tag{3.8}$$

where L represents the total number of scRNA-seq cells. We denote $c^{-1}(i)$ as the projection from cell movement cells i to cell indexes in R.

3.2.3 MAP-EM algorithm for cell position and gene expression patterns

Integrating the prior distribution of F_R (3.4) and the observation models of $Y^{(\mathrm{sl})}$ (3.5) and $Y^{(\mathrm{sc})}$ with Z (3.11), we formulated the joint distribution of them as below:

$$\begin{aligned} P(F_R, Z, Y^{(\mathrm{sl})}, Y^{(\mathrm{sc})}) =& P(F_R) P(Z, Y^{(\mathrm{sc})}|F_R) P(Y^{(\mathrm{sl})}|F_R) \\ =& (2\pi)^{-\frac{1}{2}|R|G} \det(K)^{-\frac{1}{2}} \exp\left(-\sum_g \frac{1}{2} F_{R,*g}^T K^{-1} F_{R,*g} \right) \\ & \left(\frac{1}{N^{(\mathrm{ref})}} \right)^L (2\pi\sigma^{(\mathrm{sc})2})^{-\frac{1}{2}LG} \prod_{g,l,i \in R_{t_l^{(\mathrm{sc})}}} \exp\left(-\frac{1}{2\sigma^{(\mathrm{sc})2}} (Y_{lg}^{(\mathrm{sc})} - F_{R,c^{-1}(i)g})^2 \right)^{Z_{lr}} \\ & (2\pi\sigma^{(\mathrm{sl})2})^{-\frac{1}{2}KG} \prod_{g,k} \exp\left(-\frac{1}{2\sigma^{(\mathrm{sl})2}} \left(Y_{lk}^{(\mathrm{sl})} - A_{k*} F_{R,*g} \right)^2 \right) \end{aligned}$$

It is difficult to directly analyze the posterior distribution of F_R and Z, because their inter-dependency prevent us from formulating the posterior distribution of them as a distribution with a known mean parameter. On the other hand, we can formulate the posterior distribution of Z as a categorical distribution given fixed value of another hidden variable $F_R^{(n)}$ as below:

$$
\begin{aligned}
P(Z|F_R^{(n)}, Y^{(\mathrm{sl})}, Y^{(\mathrm{sc})}) &\propto P(F_R^{(n)}, Z, Y^{(\mathrm{sl})}, Y^{(\mathrm{sc})}) \qquad (3.9)\\
&\propto P(Z, Y^{(\mathrm{sc})}|F_R^{(n)})\\
&\propto \prod_{g,l,i\in R_{t_l^{(\mathrm{sc})}}} \exp\left(-\frac{(Y_{lg}^{(\mathrm{sc})} - F_{Rc^{-1}(i)g}^{(n)})^2}{2\sigma^{(\mathrm{sc})2}}\right)^{Z_{lr}}\\
&= \exp\left(\sum_{g,l,i\in R_{t_l^{(\mathrm{sc})}}} Z_{lc^{-1}(i)}\left(-\frac{(Y_{lg}^{(\mathrm{sc})} - F_{Rc^{-1}(i)g}^{(n)})^2}{2\sigma^{(\mathrm{sc})2}}\right)\right)\\
&= \prod_{l,i\in R_{t_l^{(\mathrm{sc})}}} \exp\left(-\frac{\|Y_{l*}^{(\mathrm{sc})} - F_{Rc^{-1}(i)*}^{(n)}\|_2^2}{2\sigma^{(\mathrm{sc})2}}\right)^{Z_{lc^{-1}(i)}}\\
&\propto \prod_l \mathrm{Cat}(Z_{l*}|P_{l*}^{(n)}) \qquad (3.10)
\end{aligned}
$$

where

$$
P_{lr}^{(n)} = \begin{cases} \dfrac{\exp\left(-\frac{\|Y_{l*}^{(\mathrm{sc})} - F_{R,r*}^{(n)}\|_2^2}{2\sigma^{(\mathrm{sc})2}}\right)}{\sum_{i\in R_{t_l^{(\mathrm{sc})}}} \exp\left(-\frac{\|Y_{l*}^{(\mathrm{sc})} - F_{Rc^{-1}(i)*}^{(n)}\|_2^2}{2\sigma^{(\mathrm{sc})2}}\right)} & (t_{c(r)} = t_l^{(\mathrm{sc})}) \\ 0 & (t_{c(r)} \neq t_l^{(\mathrm{sc})}) \end{cases}.
$$

To set F_R, Z to reasonable values, we utilized a MAP-EM algorithm, and derived a MAP estimate of F_R and an expected value of Z given the MAP estimate of F_R. First, we marginalized the log joint posterior probability of F_R and Z over the posterior distribution of Z, and derived Q function of the MAP-EM algorithm as below:

$$
\begin{aligned}
Q_{EM}(F_R|F_R^{(n)}) &= \sum_Z P(Z|F_R^{(n)}, Y^{(\mathrm{sc})}, Y^{(\mathrm{sl})}) \log P(F_R, Z|Y^{(\mathrm{sc})}, Y^{(\mathrm{sl})}) \\
&\propto \sum_Z P(Z|F_R^{(n)}, Y^{(\mathrm{sc})}, Y^{(\mathrm{sl})}) \log P(F_R, Z, Y^{(\mathrm{sc})}, Y^{(\mathrm{sl})}) \\
&= \sum_Z P(Z|F_R^{(n)}, Y^{(\mathrm{sc})}, Y^{(\mathrm{sl})}) \log P(F_R) P(Y^{(\mathrm{sl})}|F_R) P(Z, Y^{(\mathrm{sc})}|F_R) \\
&\propto \sum_Z P(Z|F_R^{(n)}, Y^{(\mathrm{sc})}, Y^{(\mathrm{sl})}) \left(\sum_g -\frac{1}{2} F_{R,*g}^T K^{-1} F_{R,*g} \right. \\
&\quad \left. -\frac{1}{2\sigma^{(\mathrm{sl})2}} \|Y_{*g}^{(\mathrm{sl})} - AF_{R*g}\|_2^2 - \frac{1}{2\sigma^{(\mathrm{sc})2}} \sum_{lr} Z_{lr} \left(Y_{lg}^{(\mathrm{sc})} - F_{R,rg} \right)^2 \right) \\
&= \sum_g -\frac{1}{2} F_{R,*g}^T K^{-1} F_{R,*g} - \frac{1}{2\sigma^{(\mathrm{sl})2}} \|Y_{*g}^{(\mathrm{sl})} - AF_{R*g}\|_2^2 \\
&\quad -\frac{1}{2\sigma^{(\mathrm{sc})2}} \sum_{lr} P_{lr}^{(n)} \left(Y_{lg}^{(\mathrm{sc})} - F_{R,rg} \right)^2 \\
&\propto \sum_g -\frac{1}{2} F_{R,*g}^T K^{-1} F_{R,*g} - \frac{1}{2\sigma^{(\mathrm{sc})2}} \left(F_{R,*g}^T S^{(n)} F_{R,*g} - 2\overline{Y}_{*g}^{(\mathrm{sc})(n)} F_{R,*g}) \right)^2 \\
&\quad -\frac{1}{2\sigma^{(\mathrm{sl})2}} \left(F_{R,*g}^T A^T A F_{R,*g} - 2{Y^{(\mathrm{sl})}}^T A F_{R,*g} \right) \\
&= \exp\left(\sum_g -\frac{1}{2} (F_{R,*g}^T - \mu_{*g}^{(n)})^T \Sigma^{(n)-1} (F_{R,*g}^T - \mu_{*g}^{(n)})) \right) \qquad (3.11)
\end{aligned}
$$

where $S^{(n)}$ is a diagonal matrix withe elements $S_{rr}^{(n)} = \sum_l P_{lr}^{(n)}$,

$$
\begin{aligned}
\overline{Y}_{*g}^{(\mathrm{sc})(n)} &= P^{(n)T} Y_{*g}^{(\mathrm{sc})} \\
\Sigma^{(n)} &= \left(K^{-1} + \frac{S}{\sigma^{(\mathrm{sc})2}} + \frac{A^T A}{\sigma^{(\mathrm{sl})2}} \right)^{-1} \\
\mu_{*g}^{(n)} &= \Sigma \left(\frac{A^T Y_{*g}^{(\mathrm{sl})}}{\sigma^{(\mathrm{sl})2}} + \frac{\overline{Y}_{*g}^{(\mathrm{sc})(n)}}{\sigma^{(\mathrm{sc})2}} \right).
\end{aligned}
$$

Due to the formulation of Q function (3.11), we can easily maximize the Q function by $F_R^{(n+1)} = \mu^{(n)}$. $Q_{EM}(F_R|F_R^{(n)})$ is calcuated in E-step, and $F_R^{(n+1)}$ is set to $\mathrm{argmax}_{F_R} Q_{EM}(F_R|F_R^{(n)}) = \mu^{(n)}$ in M-step. When E-step and M-step are repeated, the values of $\mu^{(n)}$ and $P^{(n)}$ are converged to the MAP estimate $\hat{F}_R$ and the expected values $\hat{Z}$ given the MAP estimate.

3.2.4 Gaussian process regression for gene expression patterns at arbitrary time

Now we consider the gene expression of any cell i in cell movement data, which is notated as f_i, given a fixed value of F_R. The joint distribution of F_R and f_i is as below:

$$P\left(\begin{pmatrix} F_R \\ f_i \end{pmatrix}\right) = N\left(0, \begin{pmatrix} K & \boldsymbol{k} \\ \boldsymbol{k}^T & k(i,i) \end{pmatrix}\right) \tag{3.12}$$

where $\boldsymbol{k}$ represents a vector with elements $\boldsymbol{k}_r = k(c(r), i)$. This can be transformed into the posterior distribution of f_i given F_R as below:

$$P(f_i|F_R) = N(\boldsymbol{k}^T K^{-1} F_R, k(i,i) - \boldsymbol{k}^T K^{-1} \boldsymbol{k}) \tag{3.13}$$

Using this formulation, we can estimate the gene expression level of each gene at any spatiotemporal point in the cell movement data as the mean of this posterior distribution $\boldsymbol{k}^T K^{-1} F_R$.

3.2.5 Simulation

In order to evaluate the validity of our approach, we simulated spatiotemporal gene expression patterns, and generated scRNA-seq data and tomo-seq data from the simulated patterns. We evaluated the estimation performance of the scRNA-seq cells position and the spatiotemporal gene expression patterns using the simulated data. In this simulation, we generated the spatiotemporal patterns as the product of spatial patterns and temporal patterns as follows:

$$F_{ig} = S_{I(i,t,0)g} T_{gt_i} \tag{3.14}$$

where S_{ig} represents the spatial coefficient of gene g in cell $i \in R_{t=0}$, T_{gt} represents the time coefficient of the expression of gene g at time point t. When we generated spatial coefficient $S_{i,g}$, we picked one cell movement cell $o_g \in R_{t=0}$, and generated a radial spatial pattern around o_g as $S_{i,g} = a_g exp(-\|x_i - x_{o_g}\|_2^2)$ where a_g represents expression amplitude of gene g. We generated temporal coefficient T_{gt} as a sigmoid function $\frac{1}{1+e^{-(t-\tau_g)}}$ where τ_g represents the inflection point of T_{gt}, which is sampled from uniform distortion between 0 and $\max(T)$. For the generation of tomo-seq, we used previously calculated A in real data application described as 3.2.6, and sampled the gene expression of each gene g in each section k from a Poisson distribution with a mean parameter $\lambda_{kg} = A_{k*} F_{R,*g}$. For generation of simulated scRNA-seq, we randomly picked cell i out of cell movement data at a specified observation time, and sampled the gene expression level of gene g in cell l from a Poisson distribution therefrom with mean parameter $\mu_{lg} = F_{ig}$. We set the maximum

expression amplitude a_g to 100, the number of genes to 100 and the number of scRNA-seq cells at observation time to 1000 unless otherwise noted. We conducted scRNA-seq observation at 8, 10, 13, 16 and 19.5hpf, while we conducted tomo-seq observation at 8 and 19.5hpf.

3.2.6 Deriving and preprocessing of real data

3.2.6.1 Cell movement data

We downloaded the cell movement data of [Keller et al., 2008] from the cell movement data base `https://www.embl.de/digitalembryo/fish.html`. The cell movement data contains the 901 frame of three-dimensional cell coordinates with 90 second intervals, and covers the period from 100 minutes to 24 hours post fertilization (hpf)). Each frame contains $68 \sim 15582$ cells. We specified 8, 9.5, 10, 11, 12.5, 14.5, 16 and 19 hpf as the stage of 50% epiboly, shield, 75% epiboly, 90% epiboly, bud, 3 somites, 6 somites, 10 somites based on visual inspection.

3.2.6.2 Transcriptome data

Data of tomo-seq was downloaded from the supplementary data of [Junker et al., 2014]. In this study, we used the tomo-seq data at the shield and 10 somites stages. We specified the left-right axis as the axis which minimize the population variance of coordinates along itself, and specified the other axes, dorsal-ventral axis and animal-vegetable axis, based on visual inspection. The first slice position was determined so that the correlation between the total expression levels of tomo-seq slices and the cell numbers of cell movement slices was maximized. The width of tomo-seq slices was set to be $26\mu m$ so that the square sum of the difference between the total expression levels of tomo-seq slices and the cell numbers of cell movement slices was minimized after total quantity of both of them were normalized to 1. These information of the direction and width of tomo-seq slice was used to calculate the assignment matrix A. After the calculation of A, the total expression level of each tomo-seq slice was normalized to 1 per cell for each slice. We used scRNA-seq of [Farrell et al., 2018] downloaded from NCBI GEO (accession no. GSE106587). In this study, we used the data observed at the stages of the shield, 75% epiboly, 90% epiboly, 3 somites and 6 somites. The total amount of expression was normalized to 1 per cell.

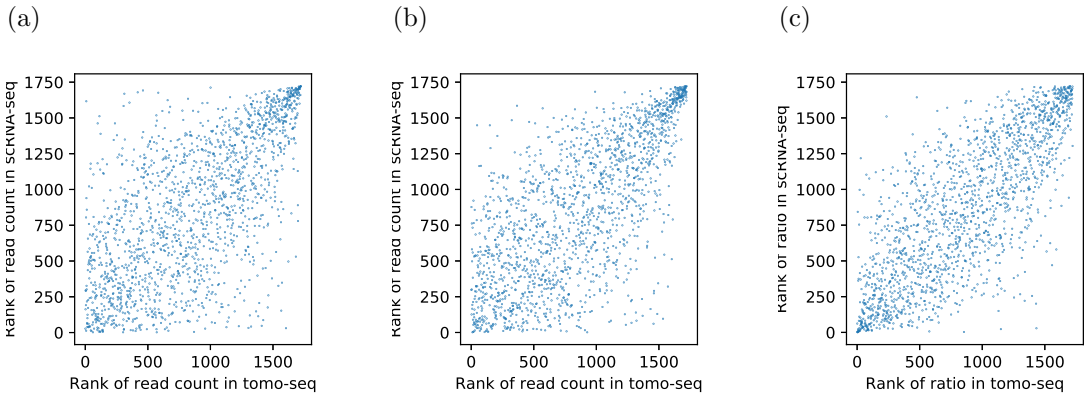


Figure 3.1: Correlation of total expression between tomo-seq and scRNA-seq. In (a) and (b), the rank of the total read count are compared at shield stage and somites stage (10 somites stage for tomo-seq data and 6 somites stage for scRNA-seq data). The rank of the ratio of the total read count between shield stage and somites stage are compared in (c). The Spearman's rank correlation coefficients are 0.66, 0.70 and 0.78 for (a), (b) and (c) respectively. The genes displayed here is genes whose total expression levels exceeds the 50% quantile at both shield and somites stages in both tomo-seq and scRNA-seq.

3.2.6.3 Normalization for integrating tomo-seq and scRNA-seq

The total expression levels of each gene was largely different between tomo-seq and scRNA seq, even if they are derived from the embryos at the same stage. On the other hand, the ratio of the total expression levels of each gene between shield stage and 10 somites stage in tomo-seq was consistent with those between shield stage and 6 somites stage in scRNA-seq (**Fig. 3.1**). Therefore, we selected a stage when each gene has more abundant expression between shield and 6 or 10 somites stages, and normalized the expression of the gene to 1 per cell in both scRNA-seq and tomo-seq at the selected stage.

After that, for each gene, we made the mean and standard deviation of the scRNA-seq expression levels 0 and 1 respectively so as to prevent a few genes with large expression levels from having too strong influence on positional estimation. For the tomo-seq, we subtracted the product of the cell number of each tomo-seq slice and the mean expression level of each gene in the pre-normalized scRNA-seq from the expression level of the gene in the tomo-seq slice, and divided them by the standard deviation of the gene expression level in the pre-normalized scRNA-seq data. This

complicated normalization of tomo-seq is consistent with the normalization of scRNA-seq described in the first half of this paragraph.

3.2.6.4 Feature selection for EM algorithm

It is unnecessary to consider the expression patterns of all the genes included in tomo-seq and scRNA-seq when we estimate the scRNA-seq assignment. In order to reduce the computational cost, we used a subset of the genes for conducting our MAP-EM algorithm. Since genes specifically expressed in cell l are crucial for calculate assignment of the cell Z_{l*}, we used the marker in each cell type of scRNA-seq data.

In particular, we used marker genes for cell types observed at the latest stage (6 somites stage), since the cell types are expected to be more divergent in later stage, and have more detailed information for each cell lineage. To derive the marker genes, we utilized "Seurat" pipeline in "R" language as described below. First, we filtered exclude genes expressed in low number of cells (< 3) and cells with low number of expressing genes (< 200). Next, we selected 2000 variable genes using "FindVariableFeatures" with "vst" method. Using the expression of these genes in each cell, we conducted principle component analysis (PCA) to reduce the expression profiles into 20 dimensions. We extracted cells observed at 6 somites stage, and find clusters by applying a shared nearest neighbor optimization method to PC coordinates of each cell with the resolution parameter as 0.5. Finally, we detected genes with excess expression at each cluster using Wilcoxon Rank Sum test ($P < 0.01$), and specified marker genes for each cluster as genes with top 10 log2 fold change among the excess genes.

On the other hand, large observation noise due to small number of reads is expected to obscure the expression patters needed for estimation of Z. Hence, we additionally filter the genes with low expression in either tomo-seq or scRNA-seq. In particular, we exclude the genes with total read numbers fewer than 1000 in tomo-seq data and the genes with total expression fewer than 1 in scRNA-seq data normalized as described in previous section. Finally, we used 271 gene for reconstructing cell assignment of scRNA-seq Z.

3.2.6.5 Practical estimation process

The values of these hyper parameters have a large effect on estimation performance of F_R and Z in our EM-algorithm. Although it is common to optimize hyper parameters $\sigma_f^2, \sigma^{(\mathrm{sl})2}, \sigma^{(\mathrm{sc})2}, l, \tau$ by maximizing the likelihood, we observed that the estimation with the parameters which locally

maximize the likelihood often posed the expression patterns inconsistent with those in the real embryos presumably due to some numerical errors. Hence, we manually specified these values based on insight to the nature of Gaussian process, and conducted our EM-algorithm with a few iteration as described below.

First, we aimed to reconstruct flat and coarse-grained spatiotemporal gene expression patterns F_R so that the assignment of cells in scRNA-seq Z are correct in wide perspective (e.g. not mapping the cells with expression profile of head to tail region of cell movement data). Hence, we set σ_f^2, smaller values of which make F_R more flat, to a very small value (0.01^2), and set spatial and temporal correlation length l, τ to large values (500, 10). After two times of E and M step calculation, we changed hyper parameter values to let F_R have more complicated spatiotemporal structure and make Z reflect the structure. In particular, we set σ_f^2 to 0.5^2, and conducted one time of E and M step calculation. After the reconstruction of Z, we set l, τ to smaller values (200, 5) and conduct the calculation of only F_R, so that it have more detailed spatiotemporal patterns. In this step, we set A to 0, and ignored the expression data of tomo-seq to reconstruct the spatiotemporal expression patterns of genes expressed only in scRNA-seq.

3.3 Results

3.3.1 Evaluating performance using simulation data

We describe the validation results of our algorithms using simulation data described in Section 3.2.5. We evaluated the estimation performance of scRNA-seq cell positions by the distance between expected position $\sum_r \hat{Z}_{lr} x_{c(r)}$ and true position in simulation for each cell, while we evaluated that of spatiotemporal gene expression patterns by the Pearson's correlation coefficient between estimated gene expression patterns and true simulated gene expression patterns for each gene. We evaluated these metrics for the stages of 8, 13 and 19.5 hpf. We note that both of scRNA-seq and tomo-seq were observed at 8 and 19.5hpf, while only scRNA-seq was observed at 13 hpf. First, we evaluated the performance across the stages. The reconstruction correlation of spatiotemporal gene expression exceeded 0.8 for 84 % of the total 100 genes (**Fig. 3.2-b**). On the other hand, the median of the distance between the expected cell positions and the true position was 11 % of the maximal cell distance in the cell movement data, which we term as an embryo diameter, and largely less than that between the randomly picked cells and the true positions (60 % of embryo diameter). (**Fig. 3.2-a**). When we compared these performance between the stages (**Fig. 3.3**),

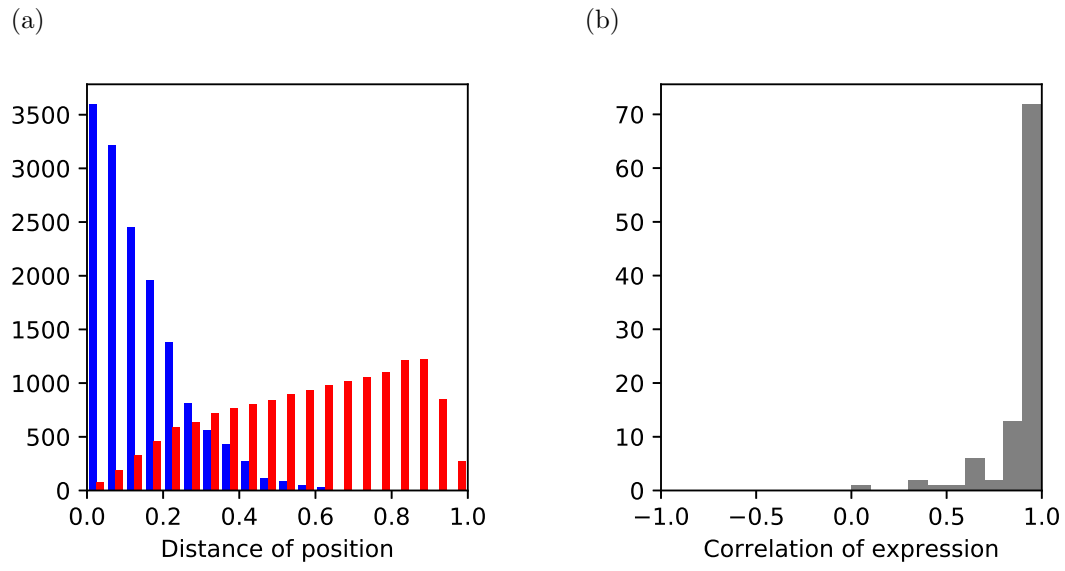


Figure 3.2: Distribution of the distance between the expected position and the true simulated position (blue) for each scRNA-seq cell in simulation (a) and Pearson's correlation coefficient between the estimated and true simulated expression patterns for each gene (b). We set the diameter of embryo in cell movement data as 1 in the x-axis of the left panel. We also displayed distribution of the distance of the randomly picked cells from each scRNA-seq cell in left panel (red). Both of the measure were calculated for shield, bud and 10 somites stage.

both of the performance metrics was the best at 8 hpf. On the other hand, the performance at 13 hpf was better than 19.5 hpf, although spatial expression patterns were available not at 13 hpf but at 19.5 hpf. This is presumably because the simple spatial expression patterns were generated at 8hpf and became more complicated in later stages along with cell migration.

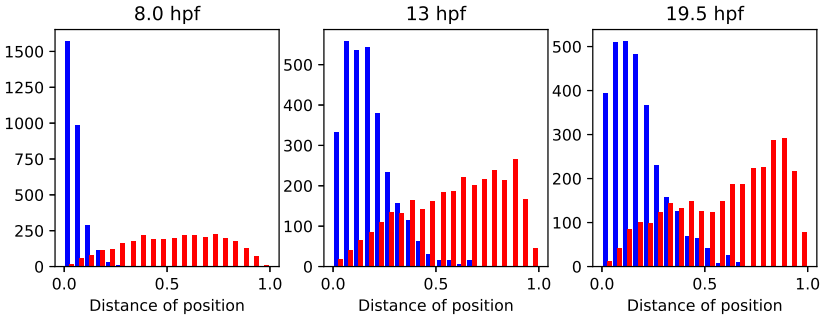
Next, we evaluated the performance for varying experimental conditions including the number of scRNA-seq cells, the number of genes and the maximum expression amplitude in original simulated spatiotemporal gene expression patterns (**Fig. 3.4**). The estimation performance of both cell position and gene expression was largely dropped when only 10 scRNA-seq cells at each stage are used in the estimation (**Fig. 3.4-a,d**). This is presumably because scRNA-seq data had almost no effect on the gene expression estimation. This might have resulted in loss of the detailed temporal dynamics included only in scRNA-seq, which had larger number of observation time points. If the scRNA-seq cell number was more than 100, median of the gene expression correlation between the

simulation and the estimates was larger than 0.85, and median of the distance between the simulated scRNA-seq cell position and the expected cell position was less than 12 % of the embryo diameter. This results shows our algorithm can work with relatively small number of scRNA-seq cells. When the number of the genes used in the estimation was set to 10, the estimation performance of the scRNA-seq cell position was largely dropped (**Fig. 3.4-b**). On the other hand, the gene expression reconstruction with 10 genes was comparable to the estimation conducted with more number of genes (**Fig. 3.4-e**). We hypothesized that this performance degradation of scRNA-seq cell position estimation was caused by cells with low expression of these small number of genes, and these cells had smaller effect on the expression reconstruction of these genes. Indeed, we found that the total expression levels were obviously lower in the scRNA-seq cells with larger distance between the expected and true position (**Fig. 3.5**). Compared to the two experimental conditions discussed above, the expression amplitude had smaller effect on the performance of both tasks (**Fig. 3.4-c, f**). These results showed our algorithm succeeded in the estimation of scRNA-seq cell positions and spatiotemporal gene expression patterns for the wide range of realistic experimental conditions.

3.3.2 Validating gene expression reconstruction on cell movement data

We applied our algorithm to the reconstruction of spatiotemporal gene expression patterns of real zebrafish embryo as described in Section 3.2.6, and validated the reconstructed spatial expression patterns of several genes at several stages by comparing them with those in ISH images of real embryos, which are searched from the Zebrafish Information Network (ZFIN), University of Oregon, Eugene, OR 97403-5274; URL: `http://zfin.org/` [Ruzicka et al., 2019]. First, we investigated the predicted expression patterns at shield stage (**Fig. 3.6**), at which both of tomo-seq and scRNA-seq was conducted. We found that the genes expressed in ventral, dorsal and animal side of the real embryo, *gsc* (Fig.7-A in [Junker et al., 2014]), *eve1* (Fig.4-E in [Joly et al., 1993]) and *sox2* (Fig.4-A in [Dee et al., 2008]), are expressed in the correct regions of cell movement data. Next, we validated the predicted expression patterns at 6 somites stage, at which only scRNA-seq observation was conducted and there was no information about the spatial gene expression patterns (**Fig. 3.7**). We found that *otx2*, which is specifically expressed in rostral side of zebrafish embryo (Fig.1-A in [Rhinn et al., 2005]), was predicted to have specific expression in rostral region of cell movement data, while *cdx4* (Fig.5-B in [Shimizu et al., 2005]), which is expressed in real caudal region, was predicted to have specific expression in the caudal region of cell movement data. These results indicated that our algorithm succeeded in the reconstruction of spatial gene expression patterns at multiple stages

(a)



(b)

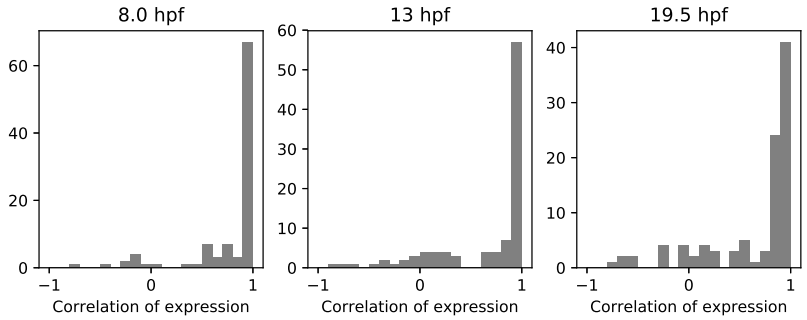


Figure 3.3: Distribution of the distance between the expected position and the true simulated position of each scRNA-seq cell in simulation (a, blue) and Pearson's correlation coefficients between the estimated and the true simulated expression level. We set the diameter of embryo in cell movement data as 1. We also displayed the distance of the randomly picked cells from each scRNA-seq cell in (a, red). Both of the measure are calculated for shield, bud and 10 somites stage separately.

regardless of the presence of the spatial expression data at the corresponding stages.

Next, we investigated the expression patterns of the marker genes for more detailed anatomical structures, since we confirmed the validity of our method from large perspective. The marker genes are derived from supplementary data of [Farrell et al., 2018]. First, we investigated the expression patterns within rostral region at 6 somites stage (**Fig. 3.8-a**). In actual zebrafish embryo, the forebrain exists in the most anterior region, and the midbrain and the optic primordium exists in posterior and lateral side of the forebrain region respectively. When we compared the predicted

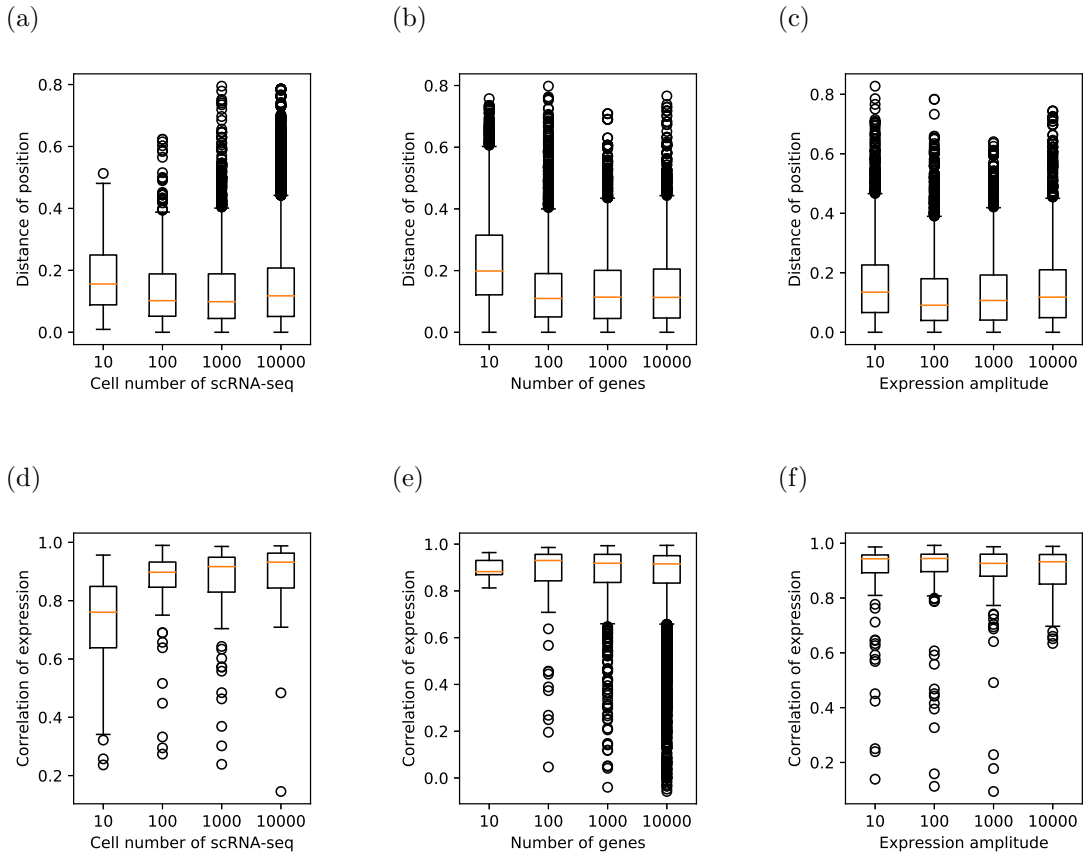


Figure 3.4: Distribution of the distance between the expected position and the true simulated position of each scRNA-seq cell (a-c) and Pearson's correlation coefficients between the estimated and true simulated expression levels of each gene (d-f) for varying experimental conditions. Both of the measure are calculated for shield, bud and 10 somites stage. The changed experimental conditions are the number of scRNA-seq cells, the number of genes and the maximum amplitude of expression in simulation.

expression patterns of maker genes for forebrain, midbrain and optic primordium, *foxg1a*, *eng2b* and *rx3* respectively, the expression of *foxg1a* was exhibited in the most anterior region, and *eng2b* and *rx3* was expressed in the posterior and lateral side of the expression region of *foxg1a* as expected. Next, we investigated expression patterns within caudal region (**Fig. 3.8-b**). In the caudal region of the real embryo, the tail bud, the presomatic mesoderm (PSM) and the somites are located in the line from posterior to anterior, while lateral plate located at the lateral sides of them. When we compared the predicted expression of *eve1*, *tbx6* and *myl13*, which are expressed in tail bud, PSM

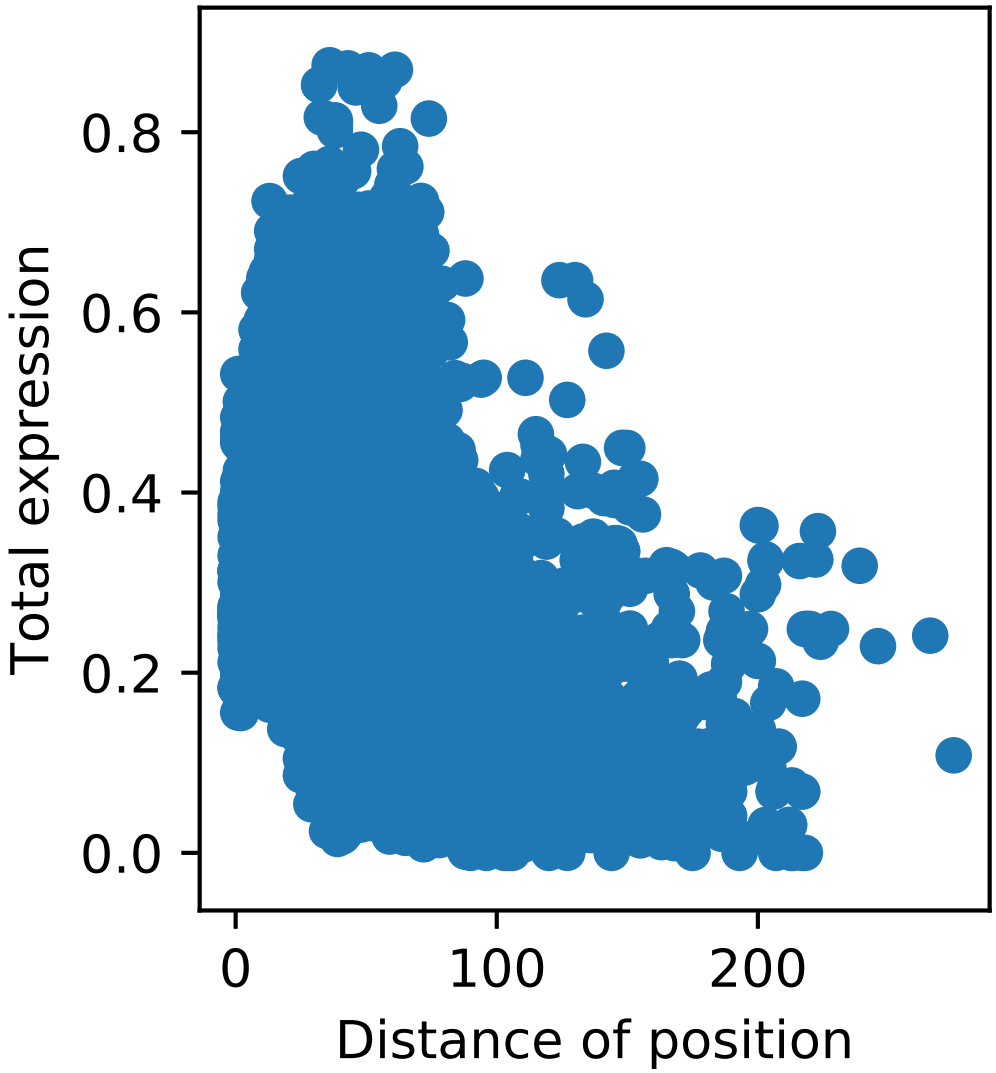


Figure 3.5: Correlation between total expression level and the distance between true and expected cell position for each scRNA-seq cells in simulation data.

and somites of the real embryo respectively, we found that *eve1* was expressed in the most posterior part, and *tbx6* and *myl13* were expressed in the second and third posterior regions as expected. On the other hand, the expression of *hand2*, which is expressed in real lateral plate, was restricted in the lateral side of the expression region of *myl13*, and absent in those of *eve1* and *tbx6*. These results showed that the spatiotemporal gene expression patterns predicted by our algorithm reflect some of the detailed anatomical structure, while it also suggested that our algorithm have some difficulty in reconstructing the expression patterns with thin and long shapes. This difficulty is expected from

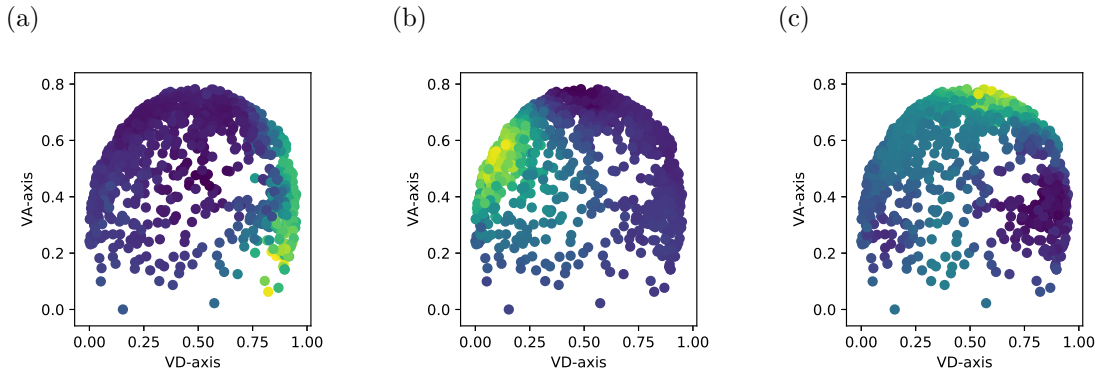


Figure 3.6: Reconstructed spatial gene expression patterns at shield stage for *gsc* (a), *eve*1 (b) and *sox*2 (c). Yellow regions represents regions with high expression level. "VD-axis" and "VA-axis" represents ventral-dorsal axis and vegetable-animal axis respectively.

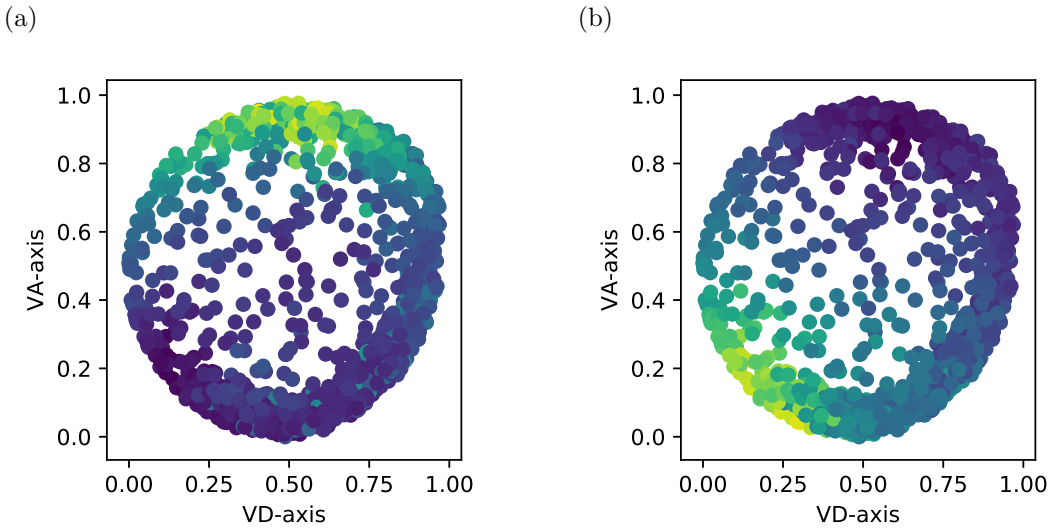


Figure 3.7: Reconstructed spatial gene expression patterns at 6 somites stage for *otx*2 (a)and *cdx*4 (b). Yellow regions represents regions with high expression level. "VD-axis" and "VA-axis" represents ventral-dorsal axis and vegetable-animal axis respectively.

our assumption that the expression between more proximal regions have stronger correlation.

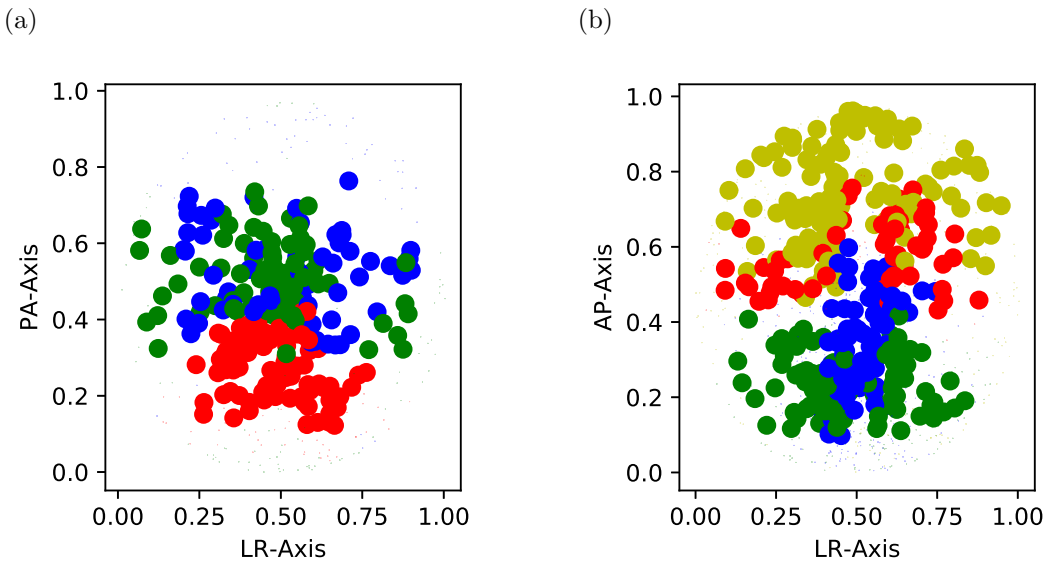


Figure 3.8: Combined expression of maker genes for substructures of head (a) and tail region (b). In (a), we displayed expression of the marker genes for forebrain (*foxg1a*, green), optic primordium (*rx3*, blue) and midbrain (*eng2b*, red). In (b), we displayed expression of the marker genes for tail bud (*eve1*, yellow), PSM (*tbx6*, red), somites (*myl13*, blue) and lateral lateral plate (*hand2*, green). The color of each cell is the color of genes with highest expression. We plotted the cells with expression higher than 0.5 when we normalize each predicted gene expression at 6 somites stage into the range from 0 to 1. "LR-axis", "PA-axis" and "AP-axis" represents left-right axis, posterior-anterior axis and anterior-posterior axis respectively.

3.3.3 Clustering analysis of cell movement cells based on predicted gene expression

Our method enable us to reconstruct the expression profile at each cell of cell movement data for all the genes observed in the scRNA-seq. Here, we investigated how accurately the predicted expression profiles at the cell movement cells recapitulated the cell types and the spatial configuration of them. In order to conduct this investigation, we do clustering of cell movement cells at 6 somites stages based on the predicted expression profiles of the cells, and confirmed the spatial location of cell movement cells belonging to each cluster. In particular, we selected 1843 genes by "FindVariableGenes" function in "Seurat" package, and reduced the dimensions of expression profiles of 1843 genes into 20 using principle component analysis (PCA). After the dimensional reduction, we

conducted the clustering of them by hierarchical clustering implemented as "scipy.cluster.hierarchy" in "scipy" package of "Python" language. We used Ward method and Euclidean distance, and set the root of maximum within-cluster variance to 30 based on the visual inspection for the deprogram of the hierarchical clustering (**Fig. 3.9**). Finally, we derived 18 clusters.

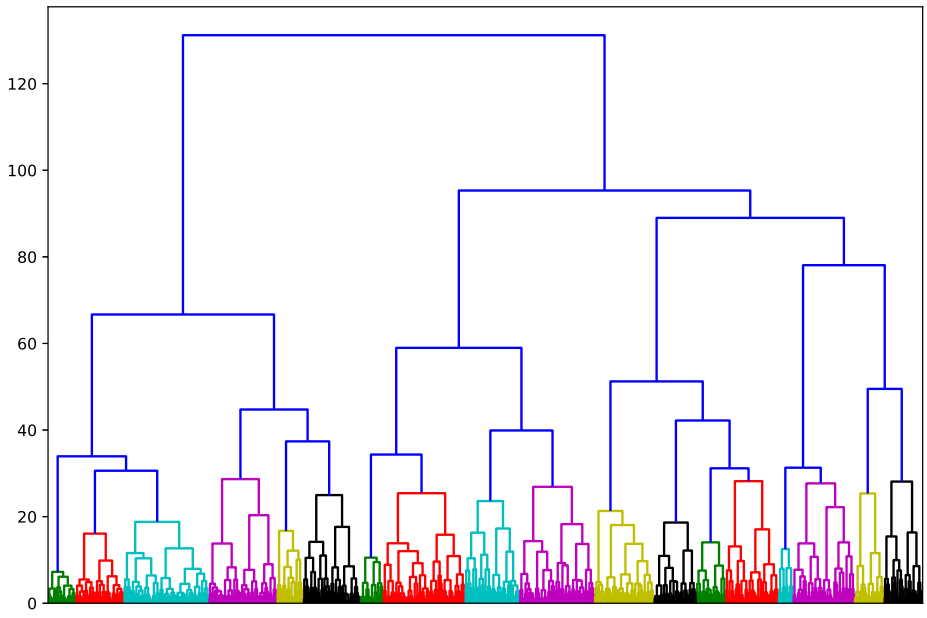


Figure 3.9: The dendrogram of hierarchical clustering. Each element of the x-axis represents each cell movement cell. The y-coordinates of the top of each U-link represents the root of the expression profile variance of all the cells below the U-link. The colors on x-axis represents clusters of each cell when we set the root of maximum within-cluster variance to 30.

In order to annotate the clusters by cell types expected to be observed in this stage, we investigated mean expression levels of each cluster for marker genes of the cell types (**Fig. 3.10**). As a result, we annotated the clusters as forebrain, midbrain, hindbrain, optic primordium, spinal cord, notochord, somites PSM and tail bud as shown in Table 3.1. Since cluster 3, 6, 14 and 18 didn't have specific expression of marker genes for any cell types investigated in this research, we didn't annotate them and excluded these cells from the following analysis.

Next, we investigated the spatial position of cell movement cells belonging to each cluster. In

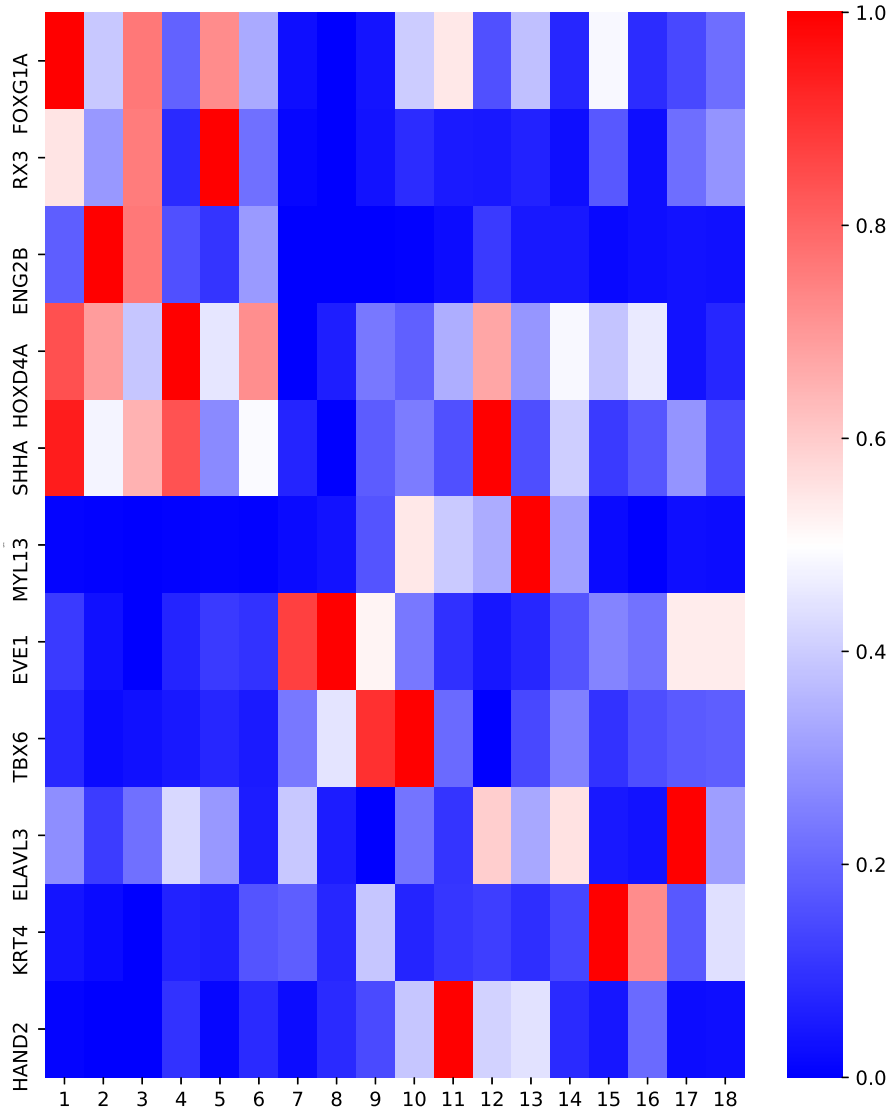


Figure 3.10: Marker gene expression of each cluster. The expression level of each gene is normalized into the range from 0 to 1. Red color indicate higher expression. Each row and column represent each gene and cluster respectively.

Cell type	Cluster	Marker gene
Lateral plate	11	hand2
Presomatic mesoderm	9,10	tbx6
Somites	13	myl13
Tail bud	7,8	eve1
Midbrain	2	eng2b
Optic primordium	5	rx3
Forebrain	1	foxg1a
Hindbrain	4	mafba
Spinal cord	17	elavl3
Epidermis	15, 16	krt4
Notochord	12	shha

Table 3.1: Annotation for each cluster of expression profiles of cell movement cells. "Cluster" column represents the cluster indexes which annotated as the "Cell type" of the same row. "Marker gene" column represents marker genes for the "Cell type" of the same row. The marker genes are derived from supplementary data of [Farrell et al., 2018].

the caudal region, the clusters of tail bud, PSM and somites were located in the line from posterior to anterior (**Fig. 3.12**), which is consistent with anatomical structure of the real zebrafish embryo. On the other hand, the lateral plate cluster was located in lateral side of only somites, and did not cover the lateral region of PSM and tail bud. This truncated shape of the lateral plate clusters were expected from the marker gene expression of lateral plate presented in the previous section. On the other hand, the notochord clusters, which is distributed at axial region from tail to head in the real embryo, also have truncated shape (**Fig. 3.13**). These truncated shape of two clusters also suggested that our method have some difficulty in estimating the expression patterns with slender shapes as described in the previous section. The spinal cord clusters were located at the ventral region, although it exists in the dorsal side of the real embryo. The epidermis cluster is located on the yolk regions, which is expected location from anatomical features of real zebrafish embryos. In the rostral region, forebrain, midbrain and hindbrain clusters were correctly located in the line from anterior to posterior. Further, the cluster of the optic primordium were located in the lateral side of forebrain as expected from real anatomical structure (**Fig. 3.12**). These results showed that

the gene expression profile of each cell movement cell was consistent with the known anatomical structures for many of cell types exhibited in this embryonic stage.

3.3.4 Spatiotemporal dynamics of molecular differentiation

One of the specific features of our method is the ability to investigate gene expression dynamics of each cell in cell movement data along with its spatial position trajectory. Utilizing this feature, we investigated the spatiotemporal differentiation dynamics of cells belonging to each cluster. In particular, we evaluated the predicted expression difference of each cell between 6 somites stage and earlier embryonic stages, and analyzed them with the spatial position of the cell. Using this method, we tried to reveal which region leads the differentiation process into each cell type. First, the application of this process to somites clusters suggested that the somites differentiation initiated at the anterior side of the population (**Fig. 3.15-b**), which is consistent with known dynamics of somitegenesis [Maroto et al., 2012].

The posterior part of PSM region is closer to the expression profiles at 6 somites stage (**Fig. 3.15-a**), while the anterior part is expected to lead the differentiation process into PSM in real embryo [Maroto et al., 2012]. This inconsistency presumably came from the non flat differentiation progress even at 6 somites stage. Since PSM is a transient cell stage before somites, the anterior part of them leads the differentiation process into somites even at 6 somites stage, and may have larger distance to expression profiles at earlier stages than the posterior part.

When we focused on the midbrain and hindbrain population, the expression profiles at the boundary region between the midbrain and the hindbrain were closer to gene expression profile at 6 somites stage than the other regions (**Fig. 3.15-c, d**). In real embryo, the Fgf singaling from this boundary is essential for the differentiation process of both hindbrain and midbrain [Harada et al., 2016]. Hence, the expected leading differentiation progress at the boundary may be due to the sooner arrival of Fgf signaling from the boundary region.

3.4 Discussion and conclusion

Recent technological developments for acquiring comprehensive gene expression profiles have enabled us to access the spatial expression patterns for numerous genes in actual 3D tissue structures. The methods to achieve this goal are ranging across multiplexed RNA molecule observation by imaging [Sheth et al., 2017, Eng et al., 2019], direct spatial annotation on transcriptome data [Junker

et al., 2014, Rodriques et al., 2019] the spatial reconstruction of scRNA-seq based on the spatial expression patterns of a few genes, which are derived by ISH [Satija et al., 2015, Karaiskos et al., 2017]. However, since all of them focused on the spatial expression patterns at fixed time points, it is difficult to analyze the spatiotemporal generation process of the spatial gene expression patterns derived by these method. To overcome this difficulty, we proposed a new method to estimate the time-continuous spatiotemporal gene expression patterns. For constructing our method, we integrated these spatially annotated transcriptome data and scRNA-seq data with cell movement data, utilizing the concept of reconstructing spatial position of scRNA-seq cells presented in a previous research [Satija et al., 2015]. Our newly developed method enables us to attribute the dynamics of spatial gene expression patterns to the cell movement and the transcription regulation separately as described in Section 3.3.4.

The major difficulty for developing our method was integrating the transcriptome data derived by different technologies with cell movement data. In order to overcome this challenge, we constructed a probabilistic model of gene expression on the cell movement data, and assumed that all the transcriptome data were observed from the probabilistic model. This approach succeeded in this challenging data integration and the estimation of spatial gene expression patterns at any time points covered by the cell movement data. Another point helped us to conduct the data integration was the reproducibility of the embryogenesis itself. Due to this reproducibility, we could assume the common spatiotemporal gene expression patterns behind all the data, and applied our new approach to the reconstruction of spatiotemporal gene expression patterns during zebrafish embryogenesis.

The nature of Gaussian process, which is utilized as prior distribution of spatiotemporal gene expression patterns, is highly dependent on the formulation of kernel function. Hence, we tried to make the kernel function of our method reflect the biological property of the spatiotemporal gene expression patterns in real embryos. In particular, we designed the kernel function so that cells with proximal spatiotemporal location have strong correlation of gene expression, based on the observation that the expression of many genes are restricted to specific spatial positions and specific embryonic stages. Another point we considered here was that some tissues with distinctly different gene expression profiles are adjacent each other at later embryonic stages due to the tissue folding processes, but they are expected to be distant each other in earlier embryonic stages. In order to reflect this biological insight, we made the spatiotemporal distance used in the kernel function reflect both position at the early embryonic stage (50 % epiboly) and the stages we are interested in. Due to this design of the kernel function, our method succeeded in recapitulating the complicated

anatomical structures as described in Section 3.3.2.

The embryonic stages we focused on this research is ranging from gastrulation to somitegenesis, during which three germ layer are differentiated into various cell types. The spatiotemporal initiation points of these differentiation processes presumably bring us new insights about the mechanism underlying the processes. We utilized the ability of our newly developed method to reconstruct gene expression profile of cells movement cells along with its spatial movement, and tried to uncover the spatiotemporal initiation points of cell type differentiation. We confirmed that the results of this analysis consistent with the known mechanisms of somites formation [Maroto et al., 2012]. Differentiation of hindbrain and midbrain was predicted to be led by the boundary region between them. This prediction indicate the hypothesis that the Fgf signaling at the boundary, which is said to required by the differentiation [Harada et al., 2016], makes the gradient of the differentiation progress due to the time lag of the signaling arrival. On the other hand, the initiation points of PSM was opposite side of the expected region [Maroto et al., 2012]. This is presumably because the metrics quantifying differentiation progress is the difference of the expression profiles to those of the same cells at the later stages, which is biased by the differentiation progress at the later stage. Recently, many methods were developed for reconstructing the differentiation progress of scRNA-seq cells. Hence, applying these methods to our estimated gene expression profiles could improve the performance of uncovering spatiotemporal initiation points of cell type differentiation.

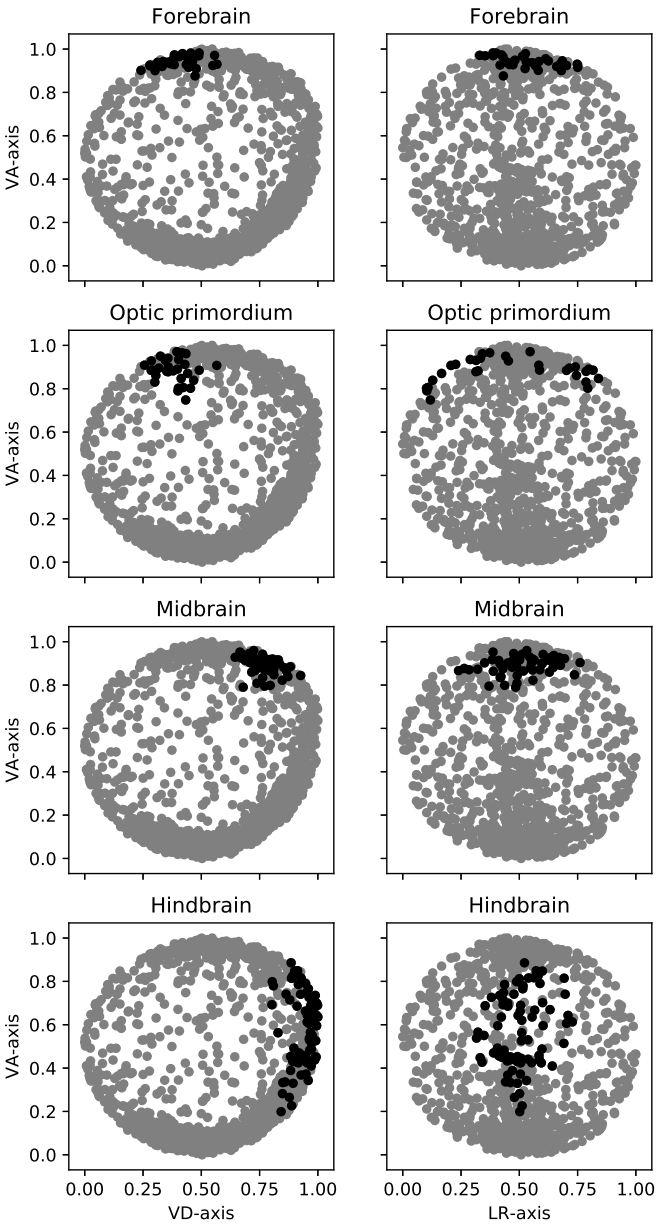


Figure 3.11: Location of cell type cluster for Forebrain, Optic primordium, Midbrain and Hindbrain. The black points represent cells belonging to a cluster annotated with each cell type, while the gray points represents the other cells.

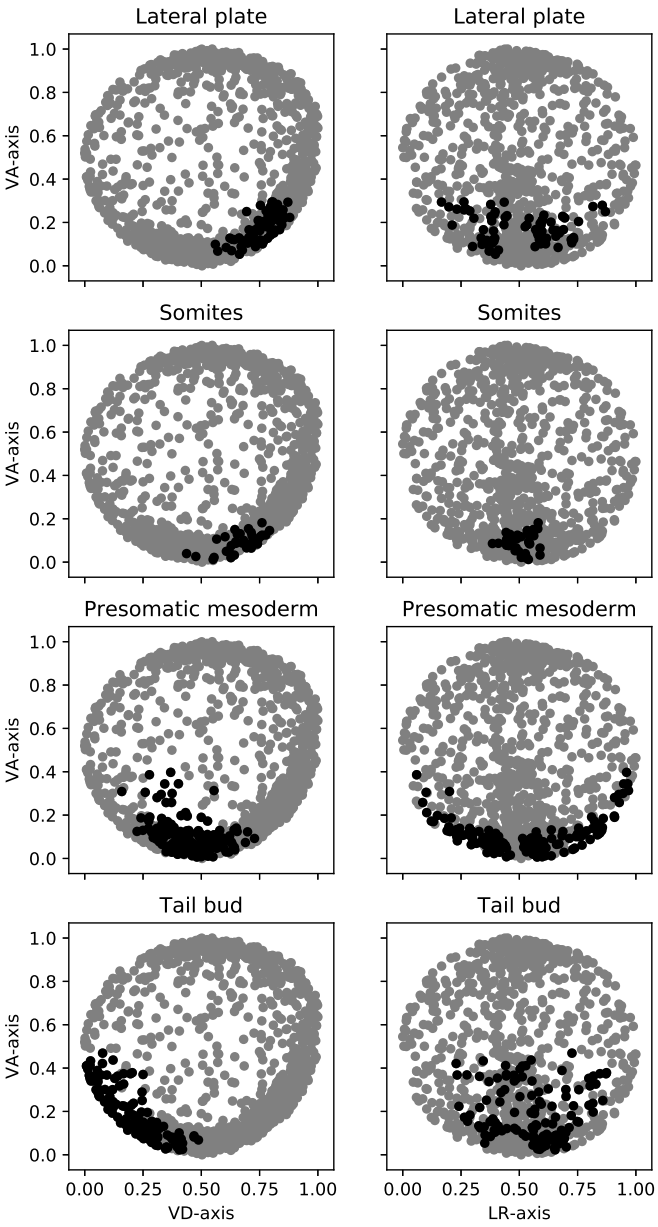


Figure 3.12: Location of cell type cluster for Lateral plate, Somites, Presomatic mesoderm and Tail bud. The black points represent cells belonging to a cluster annotated with each cell type, while the gray points represents the other cells.

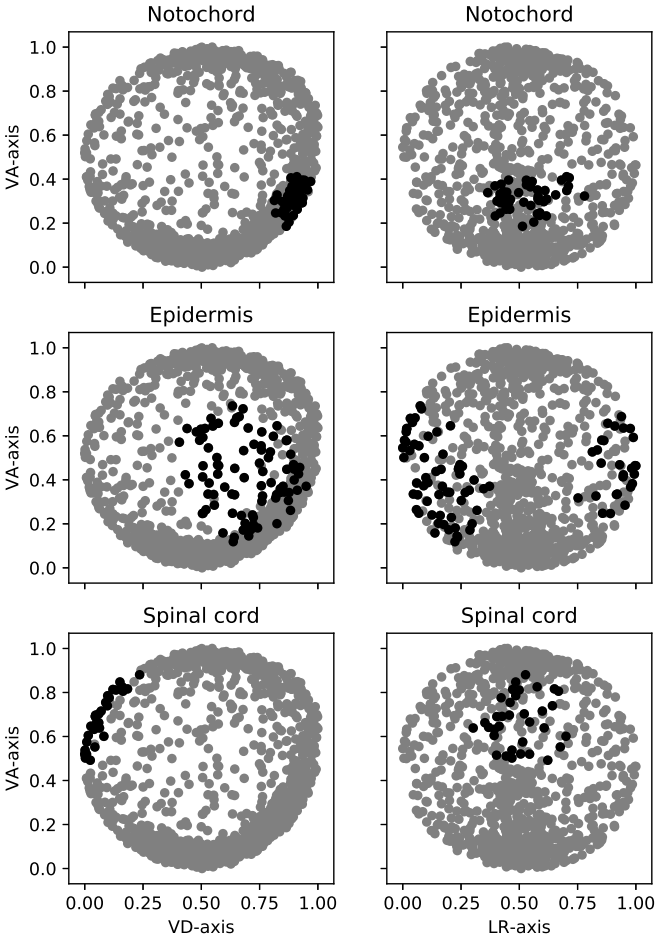


Figure 3.13: Location of cell type cluster for Notochord, Epidermis and Spinal cord. The black points represent cells belonging to a cluster annotated with each cell type, while the gray points represents the other cells.

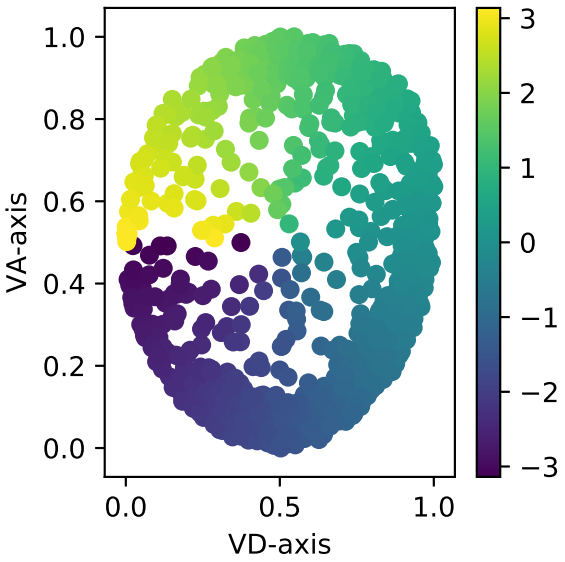


Figure 3.14: Angle coordinates of cell movement cells. Yellow indicate larger values (Rostral).

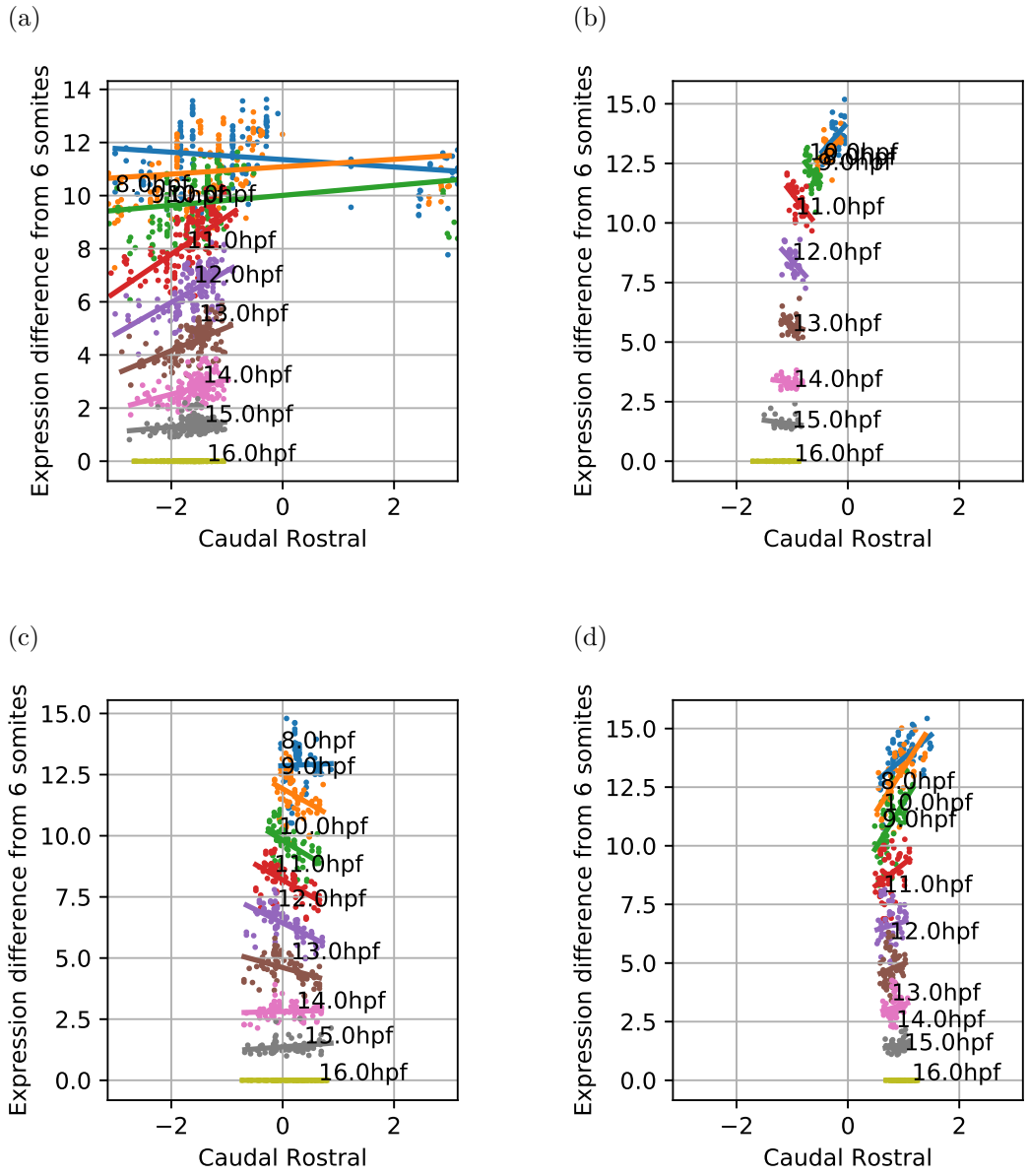


Figure 3.15: Relationships between spatial location and difference of gene expression profiles from 6 somites stages at 8, 9, 10, 11, 12, 13, 14, 15 and 16 hpf (blue, orange, green, red, purple, brown, pink, gray and gold). The displayed cells are belonging to clusters annotated with presomatic mesoderm (a), somites (b), hindbrain (c) and midbrain (d) at 6 somites stage. Spatial location is represented by the angle in the plane of VD-VA. Origin is set to the most dorsal side, and large angles represent rostral region of embryo as described in **Fig. 3.14**.

Chapter 4

General conclusions

In this thesis, we described two research with a approach where we assume probabilistic models with biologically interpretable metrics behind large scale data, and estimate these metrics using EM algorithms. As a result, we succeeded in calculating WF parameters and time-continuous spatiotemporal gene expression patters, both of which are difficult to calculate straightforward or directly observe. These metrics estimated for numerous biological units such as SNPs and genes enabled us to derive insights from a new perspective.

In the research described in Chapter 2, we developed a new method to estimate WF parameters from the E&R data, and showed the performance of our method was superior in some aspects specifically after filtering process of unreliable estimates. Analyzing selection coefficients estimated for about one million SNPs, we found a common allele frequency dynamics shared by SNPs included in *In(3R)P* region. Further, the distribution of SNPs with the *In(3R)P* specific selection coefficients suggested that other cosmopolitan inversions also shared the same specific allele frequency dynamics. On the other hand, filtering process based on the confidence intervals for estimated WF parameters enabled us to investigate the dominance distribution of existing SNPs, which was difficult to reveal due to the unidentifiability of simultaneous estimation of dominance parameters and selection coefficients. However, the small number of SNPs, which passed appropriate filtering process, indicated that the difficulty of estimating the dominance parameters based on the existing E&R data.

In the research described in Chapter 3, we integrated the spatial and single cell transcriptome data with the cell movement data to estimate the time-continuous spatiotemporal gene expression patterns. The reconstructed spatial gene expression patterns recapitulated the complicated anatomical structures even if there are no spatial transcriptome data at the corresponding embry-

onic stages. Further, comprehensive gene expression profiles on cell movement cells enabled us to investigate the initiation points of differentiation processes into each cell type. This investigation suggested that the differentiation of both hindbrain and midbrain at the boundary between them have larger progress than other regions.

These research suggested that the approach of estimating biological metrics within probabilistic models is an effective way to derive new biological statement as described above. However, the estimation is often difficult due to some unobserved variables included in the probabilistic model. The EM algorithms described in this thesis helped us to overcome this challenge and provide us the unique opportunity to investigate the biological processes from new perspectives.

Bibliography

[Agrawal and Whitlock, 2011] Agrawal, A. F. and Whitlock, M. C. (2011). Inferences about the distribution of dominance drawn from yeast gene knockout data. *Genetics*, 187(2):553–566.

[Agresti, 2002] Agresti, A. (2002). *Categorical Data Analysis*, volume 45.

[Amat et al., 2014] Amat, F., Lemon, W., Mossing, D. P., McDole, K., Wan, Y., Branson, K., Myers, E. W., and Keller, P. J. (2014). Fast, accurate reconstruction of cell lineages from large-scale fluorescence microscopy data. *Nature Methods*, 11(9):951–958.

[Bank et al., 2014] Bank, C., Ewing, G. B., Ferrer-Admettla, A., Foll, M., and Jensen, J. D. (2014). Thinking too positive? Revisiting current methods of population genetic selection inference.

[Burke et al., 2010] Burke, M. K., Dunham, J. P., Shahrestani, P., Thornton, K. R., Rose, M. R., and Long, A. D. (2010). Genome-wide analysis of a long-term evolution experiment with Drosophila. *Nature Materials*, 467(7315):1–6.

[Chen et al., 2015] Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S., and Zhuang, X. (2015). Spatially resolved, highly multiplexed RNA profiling in single cells. *Science (New York, N.Y.)*, 348(6233):aaa6090.

[Dee et al., 2008] Dee, C. T., Hirst, C. S., Shih, Y. H., Tripathi, V. B., Patient, R. K., and Scotting, P. J. (2008). Sox3 regulates both neural fate and differentiation in the zebrafish ectoderm. *Developmental Biology*, 320(1):289–301.

[Do and Batzoglou, 2008] Do, C. B. and Batzoglou, S. (2008). What is the expectation maximization algorithm? *Nature Biotechnology*, 26(8):897–899.

[Eng et al., 2019] Eng, C. H. L., Lawson, M., Zhu, Q., Dries, R., Koulena, N., Takei, Y., Yun, J., Cronin, C., Karp, C., Yuan, G. C., and Cai, L. (2019). Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature*.

[Ewens, 2004] Ewens, W. J. (2004). *Mathematical Population Genetics, I. Theoretical introduction*, volume 27.

[Farrell et al., 2018] Farrell, J. A., Wang, Y., Riesenfeld, S. J., Shekhar, K., Regev, A., and Schier, A. F. (2018). Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science*, 360(6392).

[Ferrer-Admetlla et al., 2016] Ferrer-Admetlla, A., Leuenberger, C., Jensen, J. D., and Wegmann, D. (2016). An Approximate Markov Model for the Wright–Fisher Diffusion and Its Application to Time Series Data. *Genetics*, 203(2):831–846.

[Fisher, 1930] Fisher, R. a. (1930). The Genetical Theory of Natural Selection. *Genetics*, 154:272.

[Foll et al., 2015] Foll, M., Shim, H., and Jensen, J. D. (2015). WFABC: A Wright-Fisher ABC-based approach for inferring effective population sizes and selection coefficients from time-sampled data. *Molecular Ecology Resources*, 15(1):87–98.

[Futschik and Schlötterer, 2010] Futschik, A. and Schlötterer, C. (2010). The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. *Genetics*, 186(1):207–218.

[Guennebaud et al., 2010] Guennebaud, G., Jacob, B., et al. (2010). Eigen v3. http://eigen.tuxfamily.org.

[Harada et al., 2016] Harada, H., Sato, T., and Nakamura, H. (2016). Fgf8 signaling for development of the midbrain and hindbrain. *Development Growth and Differentiation*, 58(5):437–445.

[Huber et al., 2018] Huber, C. D., Durvasula, A., Hancock, A. M., and Lohmueller, K. E. (2018). Gene expression drives the evolution of dominance. *Nature Communications*, 9(1).

[Hwang et al., 2018] Hwang, B., Lee, J. H., and Bang, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental and Molecular Medicine*, 50(8).

[Iranmehr et al., 2017] Iranmehr, A., Akbari, A., Schlötterer, C., and Bafna, V. (2017). CLEAR: Composition of likelihoods for evolve and resequence experiments. *Genetics*, 206(2):1011–1023.

[Joly et al., 1993] Joly, J. S., Maury, M., Joly, C., Boulekbache, H., and Condamine, H. (1993). Ventral and posterior expression of the homeo box gene eve1 in zebrafish (Brachydanio rerio) is

repressed in dorsalized embryos. *Comptes rendus des séances de la Société de biologie et de ses filiales*, 187(3):356–363.

[Jónás et al., 2016] Jónás, Á., Taus, T., Kosiol, C., Schlötterer, C., and Futschik, A. (2016). Estimating the effective population size from temporal allele frequency changes in experimental evolution. *Genetics*, 204(2):723–735.

[Junker et al., 2014] Junker, J. P., Noël, E. S., Guryev, V., Peterson, K. A., Shah, G., Huisken, J., McMahon, A. P., Berezikov, E., Bakkers, J., and Van Oudenaarden, A. (2014). Genome-wide RNA Tomography in the Zebrafish Embryo. *Cell*, 159(3):662–675.

[Kapun et al., 2016] Kapun, M., Fabian, D. K., Goudet, J., and Flatt, T. (2016). Genomic Evidence for Adaptive Inversion Clines in Drosophila melanogaster. *Molecular Biology and Evolution*, 33(5):1317–1336.

[Kapun et al., 2014] Kapun, M., Van Schalkwyk, H., McAllister, B., Flatt, T., and Schlötterer, C. (2014). Inference of chromosomal inversion dynamics from Pool-Seq data in natural and laboratory populations of Drosophila melanogaster. *Molecular Ecology*, 23(7):1813–1827.

[Karaiskos et al., 2017] Karaiskos, N., Wahle, P., Alles, J., Boltengagen, A., Ayoub, S., Kipar, C., Kocks, C., Rajewsky, N., and Zinzen, R. P. (2017). The Drosophila embryo at single-cell transcriptome resolution. *Science*, 358(6360):194–199.

[Keller et al., 2008] Keller, P. J., Schmidt, A. D., Wittbrodt, J., and Stelzer, E. H. (2008). Reconstruction of Zebrafish Early Embryonic Development by Scanned Light Sheet Microscopy. *Science*, 322(5904):1065–1069.

[Kimmel et al., 1995] Kimmel, C. B., Ballard, W. W., Kimmel, S. R., Ullmann, B., and Schilling, T. F. (1995). Stages of embryonic development of the zebrafish. *Developmental dynamics : an official public*, 203(3):253–310.

[Kiryu, 2011] Kiryu, H. (2011). Sufficient statistics and expectation maximization algorithms in phylogenetic tree models. *Bioinformatics*, 27(17):2346–2353.

[Kofler and Schlotterer, 2014] Kofler, R. and Schlotterer, C. (2014). A Guide for the Design of Evolve and Resequencing Studies. *Molecular Biology and Evolution*, 31(2):474–483.

[Mafessoni and Lachmann, 2015] Mafessoni, F. and Lachmann, M. (2015). Selective strolls: Fixation and extinction in diploids are slower for weakly selected mutations than for neutral ones. *Genetics*, 201(4):1581–1589.

[Manna et al., 2011] Manna, F., Martin, G., and Lenormand, T. (2011). Fitness landscapes: An alternative theory for the dominance of mutation. *Genetics*, 189(3):923–937.

[Maroto et al., 2012] Maroto, M., Bone, R. A., and Kim Dale, J. (2012). Somitogenesis. *Development (Cambridge)*, 139(14):2453–2456.

[Mathieson and McVean, 2013] Mathieson, I. and McVean, G. (2013). Estimating selection coefficients in spatially structured populations from time series data of allele frequencies. *Genetics*, 193(3):973–984.

[McDole et al., 2018] McDole, K., Guignard, L., Amat, F., Berger, A., Malandain, G., Royer, L. A., Turaga, S. C., Branson, K., and Keller, P. J. (2018). In Toto Imaging and Reconstruction of Post-Implantation Mouse Development at the Single-Cell Level. *Cell*, 175(3):859–876.e33.

[Oakes, 1999] Oakes, D. (1999). Direct Calculation of the Information Matrix via the {EM} Algorithm. *J. R. Statistical Society*, 61(2):479–482.

[OROZCO-terWENGEL et al., 2012] OROZCO-terWENGEL, P., Kapun, M., Nolte, V., Kofler, R., Flatt, T., and Schlötterer, C. (2012). Adaptation of Drosophila to a novel laboratory environment reveals temporally heterogeneous trajectories of selected alleles. *Molecular Ecology*, 21(20):4931–4941.

[Rane et al., 2015] Rane, R. V., Rako, L., Kapun, M., Lee, S. F., and Hoffmann, A. A. (2015). Genomic evidence for role of inversion 3RP of Drosophila melanogaster in facilitating climate change adaptation. *Molecular Ecology*, 24(10):2423–2432.

[Rhinn et al., 2005] Rhinn, M., Lun, K., Luz, M., Werner, M., and Brand, M. (2005). Positioning of the midbrain-hindbrain boundary organizer through global posteriorization of the neuroectoderm mediated by Wnt8 signaling. *Development*, 132(6):1261–1272.

[Rodriques et al., 2019] Rodriques, S. G., Stickels, R. R., Goeva, A., Martin, C. A., Murray, E., Vanderburg, C. R., Welch, J., Chen, L. M., Chen, F., and Macosko, E. Z. (2019). Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, 363(6434):1463–1467.

[Roelink et al., 1994] Roelink, H., Augsburger, A., Heemskerk, J., Korzh, V., Norlin, S., Ruiz i Altaba, A., Tanabe, Y., Placzek, M., Edlund, T., Jessell, T. M., and Dodd, J. (1994). Floor plate and motor neuron induction by vhh-1, a vertebrate homolog of hedgehog expressed by the notochord. *Cell*, 76(4):761–775.

[Ruzicka et al., 2019] Ruzicka, L., Howe, D. G., Ramachandran, S., Toro, S., Van Slyke, C. E., Bradford, Y. M., Eagle, A., Fashena, D., Frazer, K., Kalita, P., Mani, P., Martin, R., Moxon, S. T., Paddock, H., Pich, C., Schaper, K., Shao, X., Singer, A., and Westerfield, M. (2019). The Zebrafish Information Network: New support for non-coding genes, richer Gene Ontology annotations and the Alliance of Genome Resources. *Nucleic Acids Research*, 47(D1):D867–D873.

[Sandberg et al., 2018] Sandberg, R., Hagemann-Jensen, M., Segerstolpe, Å., Johnsson, P., Faridani, O. R., Larsson, A. J. M., Ren, B., Reinius, B., Hartmanis, L., and Rivera, C. M. (2018). Genomic encoding of transcriptional burst kinetics. *Nature*, 565(7738):251–254.

[Satija et al., 2015] Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, 33(5):495–502.

[Schlötterer et al., 2014] Schlötterer, C., Tobler, R., Kofler, R., and Nolte, V. (2014). Sequencing pools of individuals-mining genome-wide polymorphism data without big funding. *Nature Reviews Genetics*, 15(11):749–763.

[Sheth et al., 2017] Sheth, R. U., Yim, S. S., Wu, F. L., and Wang, H. H. (2017). Multiplex recording of cellular events over time on CRISPR biological tape. *Science*.

[Shimizu et al., 2005] Shimizu, T., Bae, Y. K., Muraoka, O., and Hibi, M. (2005). Interaction of Wnt and caudal-related genes in zebrafish posterior body formation. *Developmental Biology*, 279(1):125–141.

[Song and Steinrücken, 2012] Song, Y. S. and Steinrücken, M. (2012). A simple method for finding explicit analytic transition densities of diffusion processes with general diploid selection. *Genetics*, 190(3):1117–1129.

[Sorensen and Gianola, 2007] Sorensen, D. and Gianola, D. (2007). *Likelihood of Bayesian, and MCMC Methods in Quantitative Genetics.*

[Stark et al., 2019] Stark, R., Grzelak, M., and Hadfield, J. (2019). RNA sequencing: the teenage years. *Nature Reviews Genetics*, 20(11):631–656.

[Taus et al., 2017] Taus, T., Futschik, A., and Schlötterer, C. (2017). Quantifying Selection with Pool-Seq Time Series Data. *Molecular biology and evolution*, 34(11):3023–3034.

[Terhorst et al., 2015] Terhorst, J., Schlötterer, C., and Song, Y. S. (2015). Multi-locus Analysis of Genomic Time Series Data from Experimental Evolution. *PLOS Genetics*, 11(4):e1005069.

[Thisse et al., 2004] Thisse, B., Heyer, V., Lux, A., Alunni, V., Degrave, A., Seiliez, I., Kirchner, J., Parkhill, J. P., and Thisse, C. (2004). Spatial and temporal expression of the zebrafish genome by large-scale in situ hybridization screening. *Methods in Cell Biology*, 2004(77):505–519.

[Tobler et al., 2014] Tobler, R., Franssen, S. U., Kofler, R., Orozco-terWengel, P., Nolte, V., Hermisson, J., and Schlotterer, C. (2014). Massive Habitat-Specific Genomic Response in D. melanogaster Populations during Experimental Evolution in Hot and Cold Environments. *Molecular Biology and Evolution*, 31(2):364–375.

[Topa et al., 2015] Topa, H., Jonas, A., Kofler, R., Kosiol, C., and Honkela, A. (2015). Gaussian process test for high-throughput sequencing time series: application to experimental evolution. *Bioinformatics*, 31(11):1762–1770.

[Turner and Miller, 2012] Turner, T. L. and Miller, P. M. (2012). Investigating natural variation in drosophila courtship song by the evolve and resequence approach. *Genetics*, 191(2):633–642.

[Turner et al., 2011] Turner, T. L., Stewart, A. D., Fields, A. T., Rice, W. R., and Tarone, A. M. (2011). Population-based resequencing of experimentally evolved populations reveals the genetic basis of body size variation in Drosophila melanogaster. *PLoS Genetics*, 7(3).

[Wright, 1931] Wright, S. (1931). Evolution in mendelian populations. *Bulletin of Mathematical Biology*, 52(1-2):241–295.