UNIVERSITY OF CALIFORNIA
RIVERSIDE

Development and Application of Advanced Methodologies for Genome Dissection

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Plant Biology

by

Han Qu

December 2022

Dissertation Committee:
Dr. Zhenyu Jia, Chairperson
Dr. Shizhong Xu
Dr. Thomas Girke

The Dissertation of Han Qu is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

I want to thank the following people, without them, I would not have been able to complete my research, and without them, I would not have made it through my PhD degree.

Words cannot express my gratitude to my advisor Zhenyu Jia, for the enthusiasm, support, encouragement and patience. You have been an extraordinary director, tutor, educator, and guide. You have helped and guided me immensely in all directions; I am pleased to work under your remarkable management.

I could not have undertaken this journey without my dissertation committee, Dr. Shizhong Xu and Dr. Thomas Girke, who generously provided knowledge and expertise. The meetings and conversations were vital in inspiring me to think outside the box, from multiple perspectives to form comprehensive planning.

I am also thankful for other qualifying and guidance committee members, Dr. Bailian Li, Dr. Daniel Koenig, Dr. Dawn Nagel, and Dr. Min Xue. Thanks should also go to my collaborators, Dr. Steve Kay, Dr. Meng Qu, Dr. John Chater, Dr. Mikeal Roose, and Dr. Sergio Pietro Ferrante, for generously providing the data and valuable discussions.

The text of chapter 4 in this dissertation, in part, is a reprint of the material as it appears in “HNF4A defines tissue-specific circadian rhythms by beaconing BMAL1::CLOCK

chromatin binding and shaping the rhythmic chromatin landscape" (*Nature Communications*, 12(1), 6350, 2021). Thanks to Dr. Steve Kay and Dr. Meng Qu at the University of Southern California for your collaboration throughout the research. I appreciate your responsiveness and your candid communication.

I could not have undertaken the journey without the guidance of Dr. Ruidong Li, not only in research skills but also in my career development. I want to extend my sincere thanks to all former and present members of Dr. Jia and Dr. Xu's lab. Thank you for all of your help on my projects and lab activities.

The endeavor would not have been possible without the support and love of my best friends. Especially thank Fangjie Xie for standing by my side when times get hard. Thanks to Zenan Xing for making my life at UCR memorable. Thank you, Doudou and Wei Zhang, for always taking care of me. Many thanks to Congcong, Fanglei and Xinyang for always making me laugh.

I would be remiss in not mentioning my beloved parents, sister, grandpa and grandma. The strength of a family, like the strength of an army, is in its loyalty to each other.

This dissertation is dedicated to my beloved parents, sister, grandpa and grandma for their unconditional love.

ABSTRACT OF THE DISSERTATION

Development and Application of Advanced Methodologies
for Genome Dissection

by

Han Qu

Doctor of Philosophy, Graduate Program in Plant Biology
University of California, Riverside, December 2022
Dr. Zhenyu Jia, Chairperson

Next-generation sequencing (NGS) technologies have become an established and affordable framework for generating genomic information about living organisms. NGS data usually are bulky, complicated, and imperfect, which makes processing and analyzing NGS data challenging. Although many methods for NGS data analysis have emerged in the past years, the continuing development of advanced algorithms and tools is desperately wanted to tackle the dramatically growing data. Moreover, there are always misunderstandings between the end user and the developer. Proper, well-developed, explicit pipelines and many real data tests could bridge the gap between data generation and hypothesis testing.

The first chapter proposed an advanced algorithm *IIIandMe,* which infers chromosome-scale haplotypes using genomic data of single gametes. Theoretically, only three gametes

are sufficient in our hypothesis, and then the simulation of maize data and real data of citrus were tested as shreds of evidence. In the second chapter, the EM algorithm for probit and logistic regressions was introduced in a language style that is easy to understand by biologists to analyze binary traits in biology and agriculture and thus promotes wide applications of the generalized linear model (GLM) and generalized linear mixed model (GLMM) to biological problems. The third chapter performed whole genome phylogenomic analyses to decipher the phylogenetic relationships and diversification within the *Punica* genus. Our phylogenomic pipeline has empowered the use of low-coverage and fragmented whole genomes, providing productive perspectives for future research of other model groups. The fourth chapter provided a general mechanism for establishing circadian rhythm heterogeneity during development and disease progression governed by chromatin structure. We report that knockout of the lineage-specifying *Hnf4a* gene in mouse liver causes associated reductions in the genome-wide distribution of core clock component BMAL1 and accessible chromatin marks (H3K4me1 and H3K27ac), underlying circadian control of peripheral metabolism and its observed perturbation in human diseases.

Contents

List of Figures

List of Tables

Chapter 1 *IIIandMe*: An Algorithm for Chromosome-scale Haplotype Determination Using Genome-wide Variants of Three Haploid Reproductive Cells

Our recent algorithm, *Hapi*, infers chromosome-scale haplotypes using genomic data of a small number of single gametes. Its advanced version, *IIIandMe*, is proposed here to achieve comparable phasing accuracy with as few as three gametes, pushing the analysis to its limit. The new method is validated with simulation and a citrus gamete dataset. The rapid advances in genotyping technologies promise a broad application of *IIIandMe* in disclosing important genetic information.

1.1 Introduction

Mounting evidence has shown the benefit of using haplotype variants over single nucleotide polymorphisms (SNPs) in various genetic analyses, including genome-wide association studies (Yang et al. 2012; Howard et al. 2017; Lambert et al. 2013; Zhang et al. 2021), detection of the signatures of positive selection (Fariello et al. 2013), deducing genetic admixture, introgression, and demographic history (Lohmueller, Bustamante, and Clark 2009; Palamara et al. 2012). Accurate chromosomal haplotypes are needed to identify causal haplotype variants for further genetic dissection. We recently developed the *Hapi* algorithm to infer chromosome-length haplotypes using the genotypic data of several single gametes (Li et al. 2020). Multi-step preprocessing steps, including removal of erroneously genotyped markers and iterative imputation of missing markers, are

implemented by *Hapi* when the quality of gamete data is suboptimal. With the rapid advancement of biotechnologies, high-resolution genotyping with negligible errors will be achieved in the foreseeable future. Here we present a new and advanced chromosome-phasing algorithm, *IIIandMe*, which only requires three gametes when genotypic data are of high quality. Compared to *Hapi*, *IIIandMe* is logically more straightforward and computationally more efficient.

1.2 *IIIandMe* Algorithm

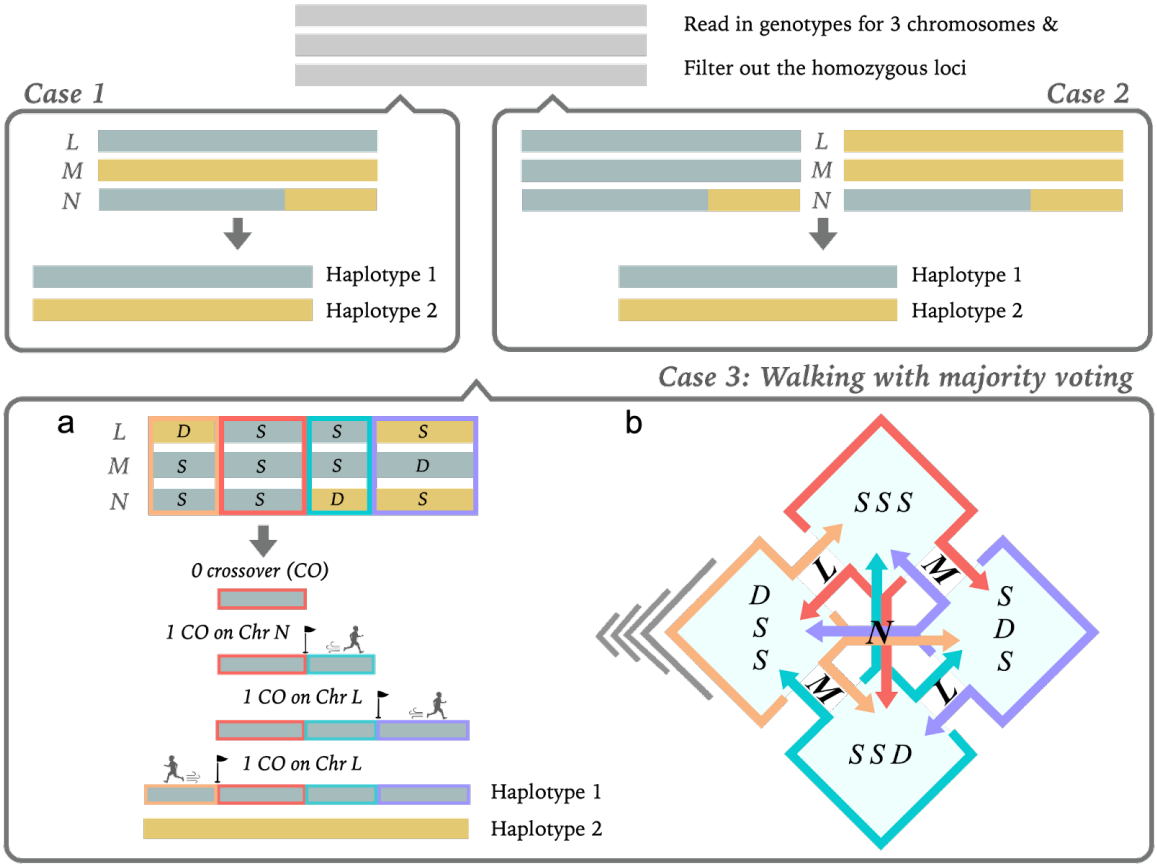


Figure 1.1 The rationale of the core 3-gamete algorithm implemented by *IIIandMe*. Cases 1 & 2: The genotypes of two out of three gamete chromosomes are complementary or identical, respectively. Therefore, the parental haplotypes can be immediately obtained. **Case 3:** Two or three gamete chromosomes have breakpoints. (a) The workflow of implementing one-locus-a-step walking with majority voting (*WMV*). A common region (*S*-*S*-*S*) is first identified, followed by the consequent detections of crossovers (labeled with flag) on both sides of this common region. Haplotype 1 is first inferred and haplotype 2 can be obtained by flipping haplotype 1. (b) Pattern-transition diagram facilitating the detection of the gamete chromosome with a crossover. The red, orange, blue and purple boxes represent *S*-*S*-*S, D*-*S*-*S, S*-*S*-*D* and *S*-*D*-*S,* respectively. The color arrows denote the pattern-to-pattern transition at a locus with the detected crossover and the associated letters (*L, M* and *N*) indicate the chromosome with that crossover.

We illustrate how *IIIandMe* works given the genotypic data of the ordered heterozygous SNPs (hetSNPs) along a chromosome, where two different colors (green and yellow in

Figure 1.1) represent two complementary (or reciprocal) paternal haplotypes for a diploid genome. The same principles apply to phasing other chromosomes independently. A few prerequisite rules (or assumptions) are needed for the subsequent reasoning and computation described in *IIIandMe*. Rule {1} – DNA breakpoints due to crossovers are randomly positioned (between two adjacent hetSNPs) along any haploid gamete chromosome. Rule {2} – Theoretically, there exists a complementary haploid chromosome of any observed haploid chromosome (recombinant or nonrecombinant) in a given gamete; these two reciprocal chromosomes may be thought of as the product from the same meiotic event. However, it is very unlikely (probability close to zero) to sample these two reciprocal gamete chromosomes since the sample space (the population of all gametes from the donor) is hypothetically large. Rule {3} – Based on rule {2}, the probability that two sampled gamete chromosomes having breakpoints at the same position is also close to zero. This is likely to be true if marker density is substantially higher relative to the frequency of crossovers. Rule {4} – If two gamete chromosomes in a sample have complementary or identical genotypes, then both must represent the intact parental haplotypes (without breakpoint).

We start from the haploid genotypes for three gamete chromosomes, labeled with *L*, *M* and *N*, respectively (**Figure 1.1**). In the simplest scenarios with two nonrecombinant gamete chromosomes (shown in Cases 1 or 2), the parental haplotypes for that chromosome can be immediately obtained according to rule {4}. Panel **a** of case 3 represents a common scenario where at least two gamete chromosomes have breakpoints. A *walking with majority voting* (*WMV*) strategy is implemented, described as follows, to

infer the parental phases of the chromosome. A common region, denoted by the pattern of *S*-*S*-*S* (red box), will be first identified across the three gamete chromosomes and used as an initial phased fragment of a parental haplotype. If no such a common region can be found, we simply flip the entire genotype of one or two original gamete chromosomes to yield a common region (**Appendix A**). The word 'flip' refers to the swap between the two reciprocal genotypes of the parents at heterozygous loci (rule{2}). This common region, representing a phased fragment of one parental haplotype (Haplotype 1 in **Figure 1.1**), is then used as the backbone of Haplotype 1 from which we can extend on both sides through *WMV*. There are four genotype patterns for gamete chromosomes *L*, *M* and *N*, *i.e.*, *S*-*S*-*S* (red box), *D*-*S*-*S* (orange box), *S*-*S*-*D* (blue box) and *S*-*D*-*S* (purple box), at each hetSNP locus, where *S* and *D* denote 'same' and 'different' genotypes, respectively. In the one-locus-a-step 'walk', a transition between any two genotype patterns indicates a crossover on only one of these three chromosomes based on the majority voting principle. For example, a transition from *S*-*S*-*S* to *D*-*S*-*S* implies a crossover on chromosome *L*. In *WMV*, we monitor transfers between these genotype patterns to detect crossover-bearing chromosomes and infer the genotypes along Haplotype 1 using a pattern-transition diagram (panel **b** of Case 3). Haplotype 2, which is complementary to haplotype 1, can be simply obtained by flipping the genotypes of the inferred Haplotype 1. See Online Methods for more complicated phasing scenarios with three gamete chromosomes.

1.3 Results

1.3.1 Simulation of maize data

Whole-genome sequencing data of 24 meiotic tetrads from a maize accession were initially published in (Li et al. 2015) and 24 'mutually independent' microspores were selected for developing *Hapi* (R. Li et al. 2020). A survey on the data of these 24 microspores showed that, on average, 1.015 microspores have a crossover at the same locus (between the same two adjacent hetSNPs), which is consistent with rule {3}. In the simulation, we selected chromosome 1 (82710 hetSNPs) of six microspores, in which two microspores have a crossover at the same locus. In each simulation scenario, we always included these two special microspores. Specifically, in a 3-gamete analysis with these two microspores, we randomly picked the third one from the rest of the four microspores, yielding four possible combinations (choose 1 out of 4) of a set of 3 microspores. The data of each combination were analyzed by *Hapi* and *IIIandMe*, respectively; the average performances for two chromosome-phasing methods, including time consumption, total RAM usage, and inference accuracy, are presented in **Figures 1.2 A-3**, **B-3** and **C-3**, accordingly. In a similar way, the 4-gamete, 5-gamete, and 6-gamete analyses were performed with 6 (choose 2 out of 4), 4 (choose 3 out of 4), and 1 (choose 4 out of 4) combinations, respectively, and the average performances of respective methods are also summarized in **Figure 1.2**. Compared to *Hapi*, *IIIandMe* had significantly reduced time consumption and RAM usage in the simulation (**Figures 1.2 A** and **B**). In the 3-gamete and 4-gamete analyses, neither method was able to correctly derive the parental haplotypes of chromosome 1, owing to the fact that two microspores have a common

crossover (**Figure 1.2 C-3** and **C-4**). Nevertheless, in the 5-gamete and 6-gamete analyses, greater than 50% of the microspores do not have crossover at this locus so the two rounds of majority votings were able to identify these two chromosomes with a common crossover and inferred the parental haplotypes of chromosome 1 accurately (**Figure 1.2 C-5** and **C-6**).

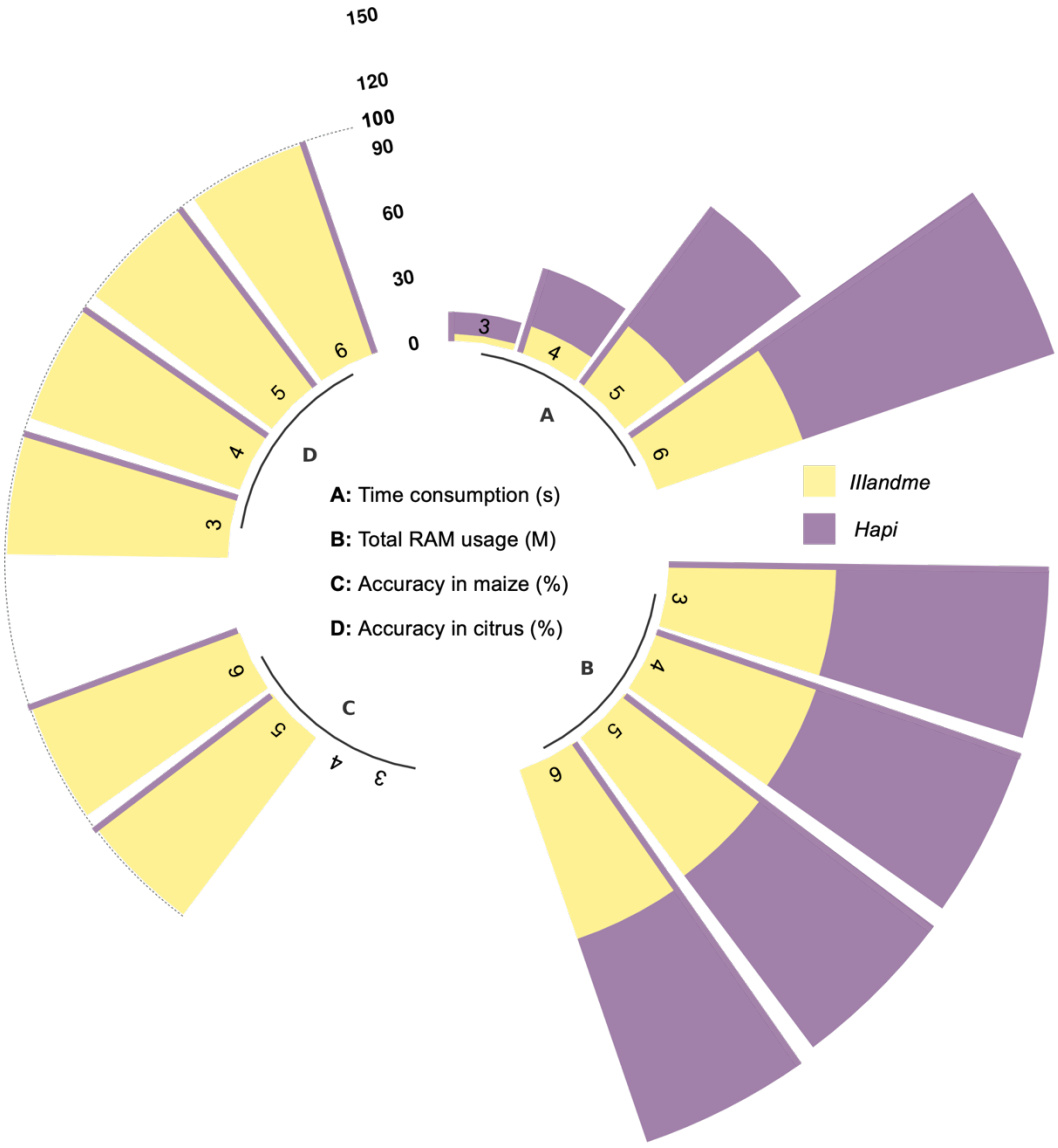


Figure 1.2 Performance comparison between *IIlandMe* and *Hapi* when analyzing 3, 4, 5 or 6 single gamete cells, respectively, in the simulation (**A, B** and **C**) and in the analysis of a citrus dataset (**D**).

1.3.2 Real citrus data analysis

We also demonstrated *IIIandMe* and *Hapi* using our dataset of a citrus accession, *Clementine de Nules*, including the diploid genotypic data for 540-2365 ordered hetSNPs on 9 chromosomes and haploid genotypic data for 6 single pollen grains. The cultivar has been assayed using a customized SNP array (AxiomTM Citrus56AX) that was designed at UCR by Dr. Mikeal Roose. This dataset has been deposited and is publicly available in the Citrus Genome Database (https://www.citrusgenomedb.org/). When analyzing any set of 3 pollen grains, 20 candidate haplotypes (choose 3 out of 6) were estimated for each of 9 chromosomes by each method. As shown in **Figure 1.2 D-3**, both methods were able to infer the parental haplotypes with 100% accuracy because no two or more pollen grains carry a common crossover at any loci. The inference accuracies for 4-gamete, 5-gamete, and 6-gamete analyses were also 100%, as displayed in **Figure 1.2 D**.

1.4 Discussion

We demonstrated that *IIIandMe*, which applies to genomic data of high quality, can accurately infer chromosomal haplotypes using three or a few more single gametes and is computationally much more efficient than *Hapi*. Our goal is to reduce the sample size substantially, and therefore operational cost, for phasing individual genomes, such that population-based genetics and clinical genetics studies become affordable and feasible. Theoretically, only three gametes are sufficient, likely pushing the boundary to its possible limit. However, errors may arise if two gametes have a common crossover at the same locus, which is a very rare occurrence in practice. To overcome this possible but

unlikely deficiency, we suggest using 5 or 6 gametes rather than 3 and propose a *k*-gamete strategy, where *k* is the number of gametes and $k > 3$. The same 3-gamete core algorithm is repeatedly applied to each combination set of 3 gametes from *k* gametes, resulting in multiple candidate chromosomal haplotypes. Consensus haplotypes are then derived with high-level of confidence, ruling out the potential adverse influences that stem from (1) a common crossover shared by two gametes and/or (2) occasional mistakes in genotypic data. In the near future, these two defect sources will likely be eliminated from the rapidly advancing technologies by further increasing the genetic marker density (DNA resolution) and decreasing the genotyping error rate, further perfecting the performance of *IIIandMe*. We foresee that *IIIandMe* will find its way to become impactful in many genetic research areas and applications.

Chapter 2 An Expectation and Maximization Algorithm for Binary Data Analyses

Probit and logistic regressions are routinely used to analyze binary traits in biology and agriculture. The expectation and maximization (EM) algorithm has been developed for probit regression analysis under the latent variable assumption. However, computing the variance matrix of the estimated parameters from the EM algorithm is not provided as a byproduct of the iteration process and thus requires additional efforts to derive the variance matrix. In this study, we presented Thomas Louis's (1982) observed information matrix for the EM algorithm. We also extended the methods from the probit regression analysis to the logistic regression analysis, including the EM algorithm and the variance matrix of the estimated parameters. Using a rice data set, we demonstrated that the EM estimated parameters and their estimated errors are identical to the results from PROC PROBIT and PROC LOGISTIC of the SAS software package, which use the Newton-Raphson algorithm to deliver the parameters and their standard errors.

2.1 Introduction

Binary traits are very common in farm animals and agricultural crops. Typical examples include disease resistance (resistant vs. susceptible), flower color (white vs. purple) and twinning in sheep (yes vs. no). Analyzing binary traits requires special statistical technologies beyond linear models. Although linear model analysis has been adopted by

geneticists to map quantitative trait loci (QTL) for binary traits (Visscher et al. 2009), it is not encouraged in general. One reason is that the predicted responses may fall outside of the natural range between 0 and 1. Generalized linear models (GLM) are particularly designed for analyzing discretely distributed traits, including binary traits as a special case (Nelder and Wedderburn 1972). The GLM is a hybrid technology between linear models and the maximum likelihood methods. The distribution of a discrete trait provides the likelihood function to be maximized. A link function of the expectation of the trait provides a linear model for parameter estimation. The GLM is sufficiently general to cover all traits with distributions from the exponential family (McCullagh and Nelder 1989).

A binary response variable is often assumed to be controlled by a hidden variable that is continuously distributed, called the liability or the latent variable. The distribution of the liability is also assumed to be normal for the probit regression analysis. If the liability is greater than zero (an assumed threshold), the individual will show one of the two phenotypes of the binary trait; otherwise, the individual will show the other phenotype. The assumed normal liability is then described by a linear model with interested factors as independent variables in the linear model. Under the liability model, all properties of the normal distribution can be used in parameter estimation and statistical tests.

The maximum likelihood (ML) method is a general method for parameter estimation. For most problems, the solution of the parameters is often implicit and thus iterations are

required. However, when applied to linear models with a normally distributed residual error, the ML solution of the parameters is explicit. Therefore, a variable associated with a normal error is better analyzed via the ML method. The assumed latent variable of the binary trait is normally distributed in the probit regression analysis. Therefore, the ML method is the ideal tool for binary data analysis under the liability model. The liability is a missing variable (missing value). When the missing value is not missing, the beautiful solution of the linear model applies. The expectation and maximization (EM) algorithm (Dempster et al. 1977) is a special algorithm for the ML method. There are two requirements in order to use the EM algorithm: (1) the problem must be formulated as a missing value problem; (2) if the missing value is not missing, the solution of the parameters is mathematically attractive (beautiful). Both requirements of the EM algorithm are satisfied for binary trait analysis. Therefore, the liability model provides the best example for the EM algorithm. McCulloch (McCulloch 2000) presented a mini review on generalized linear models, where he introduced the EM algorithm for binary data analysis. This mini review traced the original concept of "probit analysis" back to 1934 by Bliss (Bliss 1934; Bliss 1935) who plotted $\Phi^{-1}(\hat{p}_j)$ against the log dose ($\log d_j$) of nicotine to kill aphids at this dose and found that the relationship appeared to be described by a two-segment linear regression, where $\hat{p}_j = m_j / n_j$ is the binomial proportion with m_j killed aphids out of a total of n_j aphids at dose d_j. When $0 < \hat{p}_j < 1$, the probit of $\hat{p}_j$, $\Phi^{-1}(\hat{p}_j)$, can be treated as the response variable and the log dose ($\log d_j$) as an independent variable with a weight denoted by W_j, which is the inverse of

the variance for the quantile corresponding to $\hat{p}_j$. Bliss's (Bliss 1934; Bliss 1935) probit regression is limited to binomial data with large number of trials because $\hat{p}_j$ cannot take 0 or 1. Fisher (Fisher 1935), for the first time, modeled a binomial trait with the ML method that defines the probit link, which is drastically different from the probit transformation (Bliss 1935). With the probit link function, the expectation of the observed binomial data is defined as $p_j = \Phi(X_j\beta)$. This allowed Fisher (Fisher 1935) to explicitly write the binomial log likelihood function at each dose and then the overall log likelihood function of the entire sample. The Newton-Raphson algorithm is often used to search for the MLE of the parameters. When the link function is used, $\hat{p}_j$ never appears in the likelihood function; instead, m_j and n_j both enter the likelihood function separately. When $n_j = 1$ for all $j = 1, \cdots, n$, the binomial trait becomes a binary (Bernoulli) trait, where n is the sample size. So, Fisher's maximum likelihood can handle binary traits as a special case of binomial traits.

McCulloch (McCulloch 2000) introduced the EM algorithm for estimating β of a binary trait and found that the EM algorithm is remarkably similar to the "working probit" model developed by Finney (Fanney 1952), who proposed a pseudo response variable evaluated at the current parameter $\beta^{(t)}$ and used a weighted least squares method to update the parameters,

$$\beta^{(t+1)} = (X^T W X)^{-1} X^T W Z \quad (1)$$

where Z is a pseudo response variable (a function of the parameter at iteration *t*) and W is the weight, which is defined as

$$W_j = \frac{\phi^2(X_j\beta^{(t)})}{\Phi(X_j\beta^{(t)})\left[1-\Phi(X_j\beta^{(t)})\right]} \tag{2}$$

We prefer to call the pseudo response variable the "working response variable." In the EM algorithm of McCulloch (McCulloch 2000), the E-step involved the conditional expectation of the latent variable given the observed binary trait as shown below,

$$\mathrm{E}(\xi_j \mid y_j) = X_j\beta^{(t)} + \frac{y_j - \Phi(X_j\beta^{(t)})}{\phi(X_j\beta^{(t)})} W_j \tag{3}$$

where ξ_j is the normal liability. This is the expectation of a truncated standardized normal distribution. The maximization step is represented by

$$\beta^{(t+1)} = (X^T X)^{-1} X^T \mathrm{E}(\xi \mid y) \tag{4}$$

Like any other EM algorithms, there is no easy way to calculate the variance matrix of the estimated parameter, $\mathrm{var}(\hat{\beta})$, and a simple extension of the linear model variance matrix for β, $(X^T X)^{-1}$, does not apply here. This is one motivation of the current study.

The liability model makes it easier to analyze binary traits using the Bayesian approach (Albert and Chib 1993), where the missing liability for each individual can be simulated from a truncated normal distribution via the Markov chain Monte Carlo (MCMC) algorithm (Brooks 1998). Burton (Burton 1999) also developed Gibbs samplers under the Bayesian framework to study the generalized linear mixed model (GLMM) for binary data analysis. Although the Bayesian method can be computationally inferior relative to

the ML method, it does not depend on the large sample asymptotic theory for calculating the variance matrix of the estimate parameters and thus provides appropriate results of significance tests. Chakraborty and Khare (Chakraborty and Khare 2017) also studied the Bayesian method for binary data analysis and further investigated the convergence properties of the Gibb samplers. Yi and Xu (Yi and Xu 2000) adopted the Bayesian method for mapping quantitative trait loci (QTL) underlying binary disease traits. More Bayesian approaches to binary data analyses can be found in Sorensen et al (Sorensen et al. 1995), Czado (Czado 1994), Girolami and Rogers (Girolami and Rogers 2006) and McDermott et al (McDermott et al. 2016).

The GLMs reviewed so far only deal with fixed effects. When the model effects include both the fixed effects and the random effects, the model becomes GLMM (Breslow and Clayton 1993; Wolfinger and O'connell 1993). DeMaris (DeMaris 1995) first applied the GLMM to quantitative genetics to estimate the heritability for binary traits. McCulloch (McCulloch 1994) first investigated the EM algorithm applied to GLMM for estimation of variance components for binary data. The simplified version of the GLMM of McCulloch (McCulloch 1994), when the linear predictors only included fixed effects, is the EM algorithm for GLM. However, the information matrix of the estimated fixed effects was not given by McCulloch (McCulloch 1994). Beyond the EM algorithm for GLM of binary data analysis, measurement errors of independent variables of the fixed effects have been investigated (Schafer 1993; Xu et al. 2003). An interesting problem is the interval QTL mapping for binary traits, where the independent variable (genotype

indicator variable for a locus between two markers) is entirely missing (Xu et al. 2003). The liability model of Xu et al (Xu et al. 2003) involves missing values for the response variable and missing values for the predictors. Xu et al (Xu et al. 2003) were able to use a double layer EM algorithm to estimate genetic parameters (QTL effects). The approach is similar to what McCulloch (McCulloch 1994) did for the GLMM of binary data analysis where a similar double layer EM algorithm is involved.

While the GLM and GLMM for binary data analysis often take advantages of the probit link function because the latent variable is normal, which is consistent with the normal error distribution of the linear models and the linear mixed models, the logistic regression is more often used for classification involving two categorical groups. A similar EM algorithm does not exist for logistic regression. Therefore, the second motivation of this study is to propose an EM algorithm for logistic regression under the latent variable assumption. In addition to logistic regression, Liu (Liu 2004) and Azevedo & Andrade (Azevedo and Andrade 2013) suggested to use a *t* distribution with $df = 7$ to describe the latent variable. Such an analysis is called the robit analysis (Liu 2004). Link families beyond logit and probit were discussed by Czado (Czado 1994).

The overall motivations of this study are: (1) investigate the EM algorithm for probit analysis with an emphasis on developing the information matrix of the estimated parameters; (2) develop a similar EM algorithm for logistic regression; (3) illustrate the EM algorithms applied to agricultural data analysis and QTL mapping.

2.2 Theory and Methods

2.2.1 The liability model and the likelihood function

Let ξ_j be an underlying variable with a normal distribution of unit variance (called the liability). The variable controls the status of a binary trait, which is denoted by y_j, where $y_j = 1$ for the presence of a character and $y_j = 0$ for the absence of the character. The relationship between the binary phenotype and the liability is assumed to be

$$y_j = \begin{cases} 1 & \text{if} \quad \xi_j < 0 \\ 0 & \text{if} \quad \xi_j > 0 \end{cases} \tag{5}$$

All individuals with the character have a liability greater than 0 and all individuals without the character have a liability less than 0. So, we do not know the value of the liability for an individual but given the binary phenotype, we have partial information about the liability. The model for the liability is

$$\xi_j = X_j\beta + \varepsilon_j \tag{6}$$

where $\varepsilon_j \sim N(0,1)$ is the residual with a standardized normal distribution. The parameters are $\beta = \begin{bmatrix} \beta_0 & \beta_1 \end{bmatrix}^T$ and the independent variables are $X_j = \begin{bmatrix} X_{j0} & X_{j1} \end{bmatrix} = \begin{bmatrix} 1 & X_{j1} \end{bmatrix}$. The probability of $y_j = 1$ is

$$\Pr(y_j = 1) = \int_{-\infty}^{0} \phi(\xi_j)d\xi_j = \int_{-\infty}^{-X_j\beta} \phi(\xi_j - X_j\beta)d(\xi_j - X_j\beta) = \Phi(-X_j\beta) \tag{7}$$

and the probability of $y_j = 0$ is

$$\Pr(y_j = 0) = 1 - \Pr(y_j = 1) = 1 - \Phi(-X_j\beta) = \Phi(X_j\beta) \tag{8}$$

where $\Phi(X_j\beta)$ is the cumulative distribution function (CDF) of the standardized normal variable. Corresponding to $\Phi(X_j\beta)$, we call $\phi(X_j\beta)$ the probability density function (PDF) of the standardized normal distribution. Since the normal distribution is symmetrical around the mean (zero for the standardized normal distribution), $\Phi(-X_j\beta) = 1 - \Phi(X_j\beta)$ and $\phi(X_j\beta) = \phi(-X_j\beta)$. Define the log likelihood function for individual *j* by

$$L_j(\beta) = y_j \log\left[\Phi(-X_j\beta)\right] + (1 - y_j)\log\left[\Phi(X_j\beta)\right] \tag{9}$$

The population-wise log likelihood function takes the sum of all individual-wise log likelihood functions, which is

$$L(\beta) = \sum_{j=1}^{n} L_j(\beta) \tag{10}$$

We now review several existing algorithms to find the maximum likelihood estimates (MLE) of the parameters (β) and then introduce the EM algorithm under the liability model and develop the variance-covariance matrix of the estimated parameters.

2.2.2 The Newton-Raphson algorithm

The maximum likelihood estimates of the parameters are obtained via the Newton-Raphson algorithm described by

$$\beta^{(t+1)} = \beta^{(t)} - \left[\frac{\partial^2 L(\beta)}{\partial\beta\partial\beta^T}\right]^{-1}\left[\frac{\partial L(\beta)}{\partial\beta}\right] \tag{11}$$

where $\left[\partial L(\beta)/\partial\beta\right]$ is called the score vector (first order partial derivatives) and $\left[\partial^2 L(\beta)/\partial\beta\partial\beta^T\right]$ is called the Hessian matrix (second order partial derivatives).

Explicit expressions of the score and hessian matrices are available for the probit regression analysis, but numeric derivatives are often used in Once the iteration process converges, the observed information matrix is

$$I(\hat{\beta}) = -\left[\frac{\partial^2 L(\hat{\beta})}{\partial\hat{\beta}\partial\hat{\beta}^T}\right] \tag{12}$$

Therefore, the variance matrix of the estimated parameters is approximated by

$$\text{var}(\hat{\beta}) = \left[I(\hat{\beta})\right]^{-1} = -\left[\frac{\partial^2 L(\hat{\beta})}{\partial\hat{\beta}\partial\hat{\beta}^T}\right]^{-1} \tag{13}$$

Details of the variance matrix is

$$\text{var}\begin{bmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{bmatrix} = \begin{bmatrix}\text{var}(\hat{\beta}_0) & \text{cov}(\hat{\beta}_0,\hat{\beta}_1)\\ \text{cov}(\hat{\beta}_0,\hat{\beta}_1) & \text{var}(\hat{\beta}_1)\end{bmatrix} \tag{14}$$

The exact information matrix is

$$I\begin{bmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{bmatrix} = -\mathrm{E}\begin{bmatrix}\partial^2 L(\beta)/\partial\beta_0^2 & \partial^2 L(\beta)/\partial\beta_0\partial\beta_1\\ \partial^2 L(\beta)/\partial\beta_1\partial\beta_0 & \partial^2 L(\beta)/\partial\beta_1^2\end{bmatrix} \tag{15}$$

and the exact variance matrix should be

$$\text{var}\begin{bmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{bmatrix} = -\begin{bmatrix}\mathrm{E}\left(\partial^2 L(\beta)/\partial\beta_0^2\right) & \mathrm{E}\left(\partial^2 L(\beta)/\partial\beta_0\partial\beta_1\right)\\ \mathrm{E}\left(\partial^2 L(\beta)/\partial\beta_1\partial\beta_0\right) & \mathrm{E}\left(\partial^2 L(\beta)/\partial\beta_1^2\right)\end{bmatrix}^{-1} \tag{16}$$

We can test the null hypothesis $H_0 : \beta_1 = 0$ using the Wald test

$$W = \frac{\hat{\beta}_1^2}{\text{var}(\hat{\beta}_1)} \tag{17}$$

and the *p*-value is

$$p = 1 - \Pr(\chi_1^2 < W) \tag{18}$$

where χ_1^2 represents a variable from a Chi-square distribution with one degree of freedom.

The first and second numeric differentials of the likelihood function with respect to the parameters are available in R and other computer languages. Users do not have much control of the iteration process because the Newton-Raphson iteration equation is numerical, not explicit.

2.2.3 The generalized linear model approach

The generalized linear model takes advantage of the properties of the exponential family of distribution for the liability. With the normal distribution of the liability, the expectation and the variance of y_j are

$$\text{E}(y_j) = \mu_j = \Phi(-X_j\beta) \tag{19}$$

and

$$\text{var}(y_j) = V_j = \mu_j(1-\mu_j) \tag{20}$$

respectively. With the log likelihood function given in equation (10), the score matrix has an explicit form of

$$S(\beta) = \sum_{j=1}^{n} D_j^T W_j (Y_j - \mu_j) \tag{21}$$

and the negative Hessian matrix is

$$H(\beta) = \sum_{j=1}^{n} D_j^T W_j D_j \tag{22}$$

where D_j is the partial derivative of the expectation with respect to the parameters, i.e.,

$$D_j = \frac{\partial \mu_j}{\partial \beta} = X_j \phi(-X_j \beta) \tag{23}$$

The weight is

$$W_j = V_j^{-1} = \frac{1}{\mu_j (1 - \mu_j)} \tag{24}$$

The Newton-Raphson iterative equation is

$$\beta^{(t+1)} = \beta^{(t)} - \Delta\beta \tag{25}$$

where

$$\Delta\beta = H^{-1}(\beta) S(\beta) = \left[\sum_{j=1}^{n} D_j^T W_j D_j \right]^{-1} \left[\sum_{j=1}^{n} D_j^T W_j (y_j - \mu_j) \right] \tag{26}$$

The increment shown in equation (26) has a familiar weighted least square form. Therefore, the algorithm is also called the iteratively reweighted least squares (IRWLS) method.

2.2.4 The expectation and maximization algorithm

2.2.4.1 The EM algorithm

We have formulated the problem as a missing value problem. Is the solution mathematically beautiful if ξ_j is observed? If ξ_j is not missing, the solution for β is

$$\beta = (X^T X)^{-1} X^T \xi \tag{27}$$

This is the least squares solution and it is beautiful. So, the two conditions for application of the EM algorithm are met and thus we will go ahead to develop the EM algorithm for binary data analysis. Let $\phi(x)$ and $\Phi(x)$ be the probability density function (PDF) and the cumulative distribution function (CDF) for the standardized normal variable. Two properties of the standardized normal distribution are $\phi(x) = \phi(-x)$ and $\Phi(-x) = 1 - \Phi(x)$. The EM algorithm for estimating the parameters is

$$\beta = (X^T X)^{-1} X^T \mathrm{E}(\xi \mid y) \tag{28}$$

where

$$\mathrm{E}(\xi_j \mid y_j) = X_j\beta + \frac{(1-2y_j)\phi(X_j\beta)}{\Phi(X_j\beta)^{(1-y_j)}\Phi(-X_j\beta)^{y_j}} \tag{29}$$

An alternative expression of the above expectation is

$$\begin{aligned} \mathrm{E}(\xi_j \mid y_j) &= X_j\beta + \frac{(1-2y_j)\phi(X_j\beta)}{(1-y_j)\Phi(X_j\beta) + y_j\Phi(-X_j\beta)} \\ &= X_j\beta + (1-2y_j)\frac{\phi(X_j\beta)}{\Phi\left[(1-2y_j)X_j\beta\right]} \end{aligned} \tag{30}$$

Equation (30) is a smart way to write the conditional expectation. The conventional expression is in fact expressed as

$$\mathrm{E}(\xi_j \mid y_j) = \begin{cases} X_j\beta - \dfrac{\phi(X_j\beta)}{\Phi(-X_j\beta)} & \text{if} \quad y_j = 1 \\ X_j\beta + \dfrac{\phi(X_j\beta)}{\Phi(X_j\beta)} & \text{if} \quad y_j = 0 \end{cases} \tag{31}$$

The EM algorithm is represented by the two steps after the parameters are initialized.

Summary of the EM algorithm:

Step (0): Initialize the parameter values, $\beta^{(t)}$ for $t = 0$;

Step (1): Take the expectation step by calling equation (30);

Step (2): Take the maximization step by calling equation (28);

Step (3): Increment *t* by $t = t + 1$ and go back to step (1) if $\delta = (\beta^{(t+1)} - \beta^{(t)})^2$ is not sufficiently small, otherwise stop the loop and report the MLE of the parameters $\hat{\beta}$.

The Louis information matrix (Louis 1982) for the estimated parameter is

$$I(\hat{\beta}) = X^T X - X^T \operatorname{var}(\xi \mid y) X = X^T \left[I - \operatorname{var}(\xi \mid y) \right] X \tag{32}$$

The variance-covariance matrix of the estimated parameters is

$$\operatorname{var}(\hat{\beta}) = I^{-1}(\hat{\beta}) = \left\{ X^T \left[I - \operatorname{var}(\xi \mid y) \right] X \right\}^{-1} \tag{33}$$

Let us define

$$i_j = \frac{\phi(X_j\beta)}{(1 - y_j)\Phi(X_j\beta) + y_j\Phi(-X_j\beta)} \tag{34}$$

which is called selection intensity in quantitative genetics (Falconer and Mackay 1996). The conditional variance of the liability given y_j is

$$\text{var}(\xi_j \mid y_j) = 1 - i_j^2 \tag{35}$$

Let $W = \text{diag}(i_1^2, \cdots, i_n^2)$ be a diagonal matrix with elements holding the squared selection intensities. We can express the conditional variance matrix by $\text{var}(\xi \mid y) = I - W$ and thus $I - \text{var}(\xi \mid y) = I - (I - W) = W$. Therefore, the variance-covariance matrix of the estimated parameters is

$$\text{var}(\hat{\beta}) = (X^T W X)^{-1} \tag{36}$$

The Wald test for the regression coefficient is

$$W = \frac{\hat{\beta}_1^2}{\text{var}(\hat{\beta}_1)} \tag{37}$$

and the *p*-value is

$$p = 1 - \Pr(\chi_1^2 < W) \tag{38}$$

where χ_1^2 represents a variable from a Chi-square distribution with one degree of freedom.

2.2.4.2 Derivation of the EM algorithm

The expectation of the complete-data likelihood function is the target function for maximization. The complete-data log likelihood function is

$$L_C(\beta) = -\frac{1}{2}(\xi - X\beta)^T (\xi - X\beta) \tag{39}$$

The expectation of the complete log likelihood function is

$$\mathrm{E}\left[L_C(\beta)\right] = -\frac{1}{2}\mathrm{E}\left[(\xi - X\beta)^T(\xi - X\beta)\right] \tag{40}$$

The partial derivative of $\mathrm{E}[L_C(\beta)]$ with respect to β is

$$\frac{\partial}{\partial\beta}\mathrm{E}\left[L_C(\beta)\right] = X^T\mathrm{E}(\xi - X\beta) = X^T\mathrm{E}(\xi \mid y) - X^TX\beta \tag{41}$$

Setting $\frac{\partial}{\partial\beta}\mathrm{E}\left[L_C(\beta)\right] = 0$, yields

$$X^T\mathrm{E}(\xi \mid y) - X^TX\beta = 0 \tag{42}$$

Therefore,

$$\beta = (X^TX)^{-1}X^T\mathrm{E}(\xi \mid y) \tag{43}$$

which is the maximization step. The expectation step has been introduced before in equation (29) or equation (30).

The Thomas Louis (1982) information matrix for the estimate parameters is

$$I(\hat{\beta}) = -\mathrm{E}\left[\frac{\partial^2 L_C(\beta)}{\partial\beta\partial\beta^T}\right] - \mathrm{var}\left[\frac{\partial L_C(\beta)}{\partial\beta}\right] \tag{44}$$

where

$$\mathrm{E}\left[\frac{\partial^2 L_C(\beta)}{\partial\beta\partial\beta^T}\right] = -X^TX \tag{45}$$

and

$$\mathrm{var}\left[\frac{\partial L_C(\beta)}{\partial\beta}\right] = X^T\,\mathrm{var}(\xi - X\beta)X = X^T\,\mathrm{var}(\xi \mid y)X \tag{46}$$

Therefore,

$$I(\hat{\beta}) = X^T \left[I - \text{var}(\xi \mid y) \right] X \tag{47}$$

As a result,

$$\text{var}(\hat{\beta}) = I^{-1}(\hat{\beta}) = \left\{ X^T \left[I - \text{var}(\xi \mid y) \right] X \right\}^{-1} \tag{48}$$

2.2.5 EM Algorithm for logistic regression

The liability model for logistic regression is the same as the probit regression, which is

$$\xi_j = X_j \beta + \varepsilon_j \tag{49}$$

The residual of the model is assumed to follow a standardized logistic distribution, $\varepsilon_j \sim \text{Logis}(\mu = 0, s = 1)$, where $\mu = 0$ is the location parameter and $s = 1$ is the scale parameter. The scale parameter is not the standard deviation, which is $\pi / \sqrt{3}$, in the standardize logistic distribution. Unlike the normal liability model, the complete-data log-likelihood function for the logistic model does not have a clean form. This makes the EM algorithm for the logistic regression intractable. However, we simply copied the formula from the probit regression and adopted the maximization step by

$$\beta = (X^T X)^{-1} X^T \mathrm{E}(\xi \mid y) \tag{50}$$

where the conditional expectation is the expectation of a truncated logistic distribution. There is no explicit expression of the truncated expectation, but a numerical function is available in R (Nadarajah and Kotz 2006), which is

`extrunc(spec="logis",a=-Inf,b=0,location=-`$X_j\beta$`,scale=1)` (51)

if $y_j = 1$ and

$$\texttt{extrunc(spec="logis",a=0,b=Inf,location=-X}_j\beta\texttt{,scale=1)} \tag{52}$$

if $y_j = 0$. The Louis (Louis 1982) information matrix used in equation (32) does not apply here. We now propose an alternative method for the variance of $\hat{\beta}$ from the logistic regression model.

Recall that ξ_j defined in equation (49) is a logistic variable. If we know $\text{var}(\xi_j)$, we can take the inverse of $\text{var}(\xi_j)$ to obtain a weight, $W_j = \left[\text{var}(\xi_j)\right]^{-1}$, which will allow us to calculate

$$\text{var}(\hat{\beta}) = (X^T W X)^{-1} \tag{53}$$

We now use the delta method to approximate $\text{var}(\xi_j)$. We first use the Taylor series expansion to linearize the relationship between y_j and ξ_j at $\xi_j = X_j\beta$, which is

$$\begin{aligned} y_j &\approx \left[1-\Theta(\xi_j)\right]_{\xi_j = X_j\beta} + \left\{\partial\left[1-\Theta(\xi_j)\right]/\partial\xi_j\right\}_{\xi_j = X_j\beta}(\xi_j - X_j\beta) \\ &= \left[1-\Theta(\xi_j)\right]_{\xi_j = X_j\beta} - \left[\partial\Theta(\xi_j)/\partial\xi_j\right]_{\xi_j = X_j\beta}(\xi_j - X_j\beta) \\ &= \left[1-\Theta(\xi_j)\right]_{\xi_j = X_j\beta} - \left[\theta(\xi_j)\right]_{\xi_j = X_j\beta}(\xi_j - X_j\beta) \\ &= 1-\Theta(X_j\beta) - \theta(X_j\beta_j)(\xi_j - X_j\beta) \end{aligned} \tag{54}$$

The variance of y_j is

$$\text{var}(y_j) \approx \theta^2(X_j\beta_j)\,\text{var}(\xi_j - X_j\beta) = \theta^2(X_j\beta_j)\,\text{var}(\xi_j) \tag{55}$$

Since $\text{var}(y_j) = \Theta(X_j\beta)\left[1-\Theta(X_j\beta)\right]$, we have

$$\Theta(X_j\beta)\left[1-\Theta(X_j\beta)\right] \approx \theta^2(X_j\beta_j)\,\mathrm{var}(\xi_j) \tag{56}$$

Solving for the variance of the liability, we get

$$\mathrm{var}(\xi_j) \approx \frac{1}{\theta^2(X_j\beta_j)}\Theta(X_j\beta)\left[1-\Theta(X_j\beta)\right] \tag{57}$$

Therefore, the weight is approximated by

$$W_j \approx \left[\mathrm{var}(\xi_j)\right]^{-1} \approx \frac{\theta^2(X_j\beta_j)}{\Theta(X_j\beta)\left[1-\Theta(X_j\beta)\right]} \tag{58}$$

2.3 Results

2.3.1 Association between the "dark purple pericarp color" and agronomic traits in rice

This trait is controlled by a single gene called the OSC1 gene (Saitoh et al. 2004). Presence of this gene causes the entire rice plant to show the dark purple color. The phenotype was coded as 1 for the presence of the purple color and 0 for the absence of the purple color. The population consists of 210 recombinant inbred lines (RIL) from the cross between two elite rice cultivar. Among the 210 RILs, 91 were purple colored and 119 were regular green colored. The agronomic traits (variables) include yield (YD), tiller number per plant (TP), grain number (GN), 1000-grain-weight (KGW), grain length (GL), grain width (GW) and heading date (HD). The traits were evaluated with four replications (two years and two locations) (Yu et al. 2011). The data are provided in **Appendix B-S1**. The probit model is

$$\Phi^{-1}(\mu) = \beta_{INT} + X_{YD}\beta_{YD} + X_{TP}\beta_{TP} + X_{GN}\beta_{GN} + X_{KGW}\beta_{KGW} + X_{GL}\beta_{GL} + X_{GW}\beta_{GW} + X_{HD}\beta_{HD} \tag{59}$$

where β_{INT} is the intercept. The data were analyzed using three algorithms under both the probit model and the logistic model. The first algorithm is the Newton-Raphson (NR) algorithm using our own R code. The second algorithm is the EM algorithm developed in this project. The third algorithm is although the Newton-Raphson algorithm but implemented with PROC PROBIT and PROC LOGISTIC in SAS. Results from the probit model analysis are shown in **Table 2.1** while **Table 2.2** shows the results from the logistic model analysis.

Table 2.1 Estimated parameters and tests from the probit regression model analysis.

	EM		NR		SAS		Test	
Parameter	Estimate	StdErr	Estimate	StdErr	Estimate	StdErr	Wald	p-Value
Intercept	-9.5129	6.2314	-9.5129	6.2312	-9.5129	6.2314	2.33	0.1269
YD	0.0166	0.1017	0.0166	0.1017	0.0166	0.1017	0.03	0.8707
TP	0.1150	0.2384	0.1150	0.2384	0.1150	0.2384	0.23	0.6294
GN	0.0019	0.0253	0.0019	0.0253	0.0019	0.0253	0.01	0.9395
KGW	-0.2787	0.1239	-0.2787	0.1239	-0.2787	0.1239	5.06	0.0245
GL	0.9646	0.2945	0.9646	0.2945	0.9646	0.2945	10.72	0.0011
GW	3.0556	0.7804	3.0557	0.7804	3.0557	0.7804	15.33	0.0001
HD	-0.0408	0.0151	-0.0408	0.0151	-0.0408	0.0151	7.25	0.0071

Table 2.2 Estimated parameters and tests from the logistic regression model analysis.

	EM		NR		SAS		Test		
Parameter	Estimate	StdErr	Estimate	StdErr	Estimate	StdErr	Wald	p-Value	Odds Ratio
Intercept	-15.4488	10.2063	-15.9031	10.1854	-15.9031	10.2882	2.3894	0.1222	
YD	0.0291	0.1648	0.0224	0.1643	0.0224	0.1651	0.0184	0.892	1.023
TP	0.1814	0.3904	0.2022	0.3893	0.2022	0.3915	0.2669	0.6054	1.224
GN	0.0028	0.0411	0.0038	0.0410	0.0038	0.0412	0.0086	0.9262	1.004
KGW	-0.4595	0.2033	-0.4494	0.2028	-0.4494	0.2027	4.915	0.0266	0.638
GL	1.5842	0.4875	1.5701	0.4874	1.5700	0.4901	10.2631	0.0014	4.807
GW	4.9920	1.3152	5.0465	1.3172	5.0465	1.3258	14.4888	0.0001	155.478
HD	-0.0667	0.0250	-0.0666	0.0250	-0.0666	0.0250	7.108	0.0077	0.936

For the probit model analysis (**Table 2.1**), all three algorithms produced identical results for both the estimated parameters and the standard errors (StdErr) of the estimates. This observation validated the EM algorithm. Among the seven variables (agronomic traits), four variables appear to be related to the color trait. The p-value of $\hat{\beta}_{KGW}$ is significant at the 0.05 level. The p-values of the remaining three significant variables are all very small, especially trait GW with a p-value of 0.0001.

The conclusions from the probit regression analysis also apply to the logistic regression analysis (**Table 2.2**). The NR and the SAS results of the logistic analysis are identical, subject to small differences due to floating point errors of the computers. The EM algorithm, however, produced slightly different results from the NR and the SAS algorithms. The differences are negligibly small, which validated the EM algorithm. The

standard errors of the estimated parameters for the EM algorithm differ slightly from those of the NR and SAS analyses (**Table 2.2**). For the EM algorithm, we also drew 1000 samples with replacement to calculate the bootstrap variances and standard deviations among the 1000 samples (Efron 1979). **Appendix B-S2** provide the bootstrap variance-covariance matrix of the EM estimated parameters along with the NR and SAS variance-covariance matrices. From the covariance matrices, we extracted the diagonal elements and took square roots of the variances to get the standard errors, which are listed in **Table 2.3**. The bootstrap standard errors are consistently larger than the standard errors from the NR algorithm and the SAS procedure. This observation indicates that using the observed information of an estimated parameter to calculate the variance of the estimate is biased downward. The expected information may improve the estimation of the standard error of a parameter over the observed information.

The logistic regression analysis produced odds ratio statistics as by products. The odds ratio is a very common statistic in human genetics studies, although it is rarely reported in agricultural statistics. The odds ratios for the GL and GW are very high, much higher than unity, which is expected under the null model. The highest odds ratio occurs for trait GW, which is

$$\text{OR} = \frac{\pi(1)/(1-\pi(1))}{\pi(0)/(1-\pi(0))} = \frac{\exp(\beta_{GW} + \beta_{-GW})}{\exp(\beta_{-GW})} = \exp(\beta_{GW}) = \exp(5.0465) = 155.478 \qquad (60)$$

where β_{-GW} is the sum of all Betas except β_{GW}.

Table 2.3 Standard errors of the estimated parameters from the logistic regression analysis.

Parameter	Bootstrap	NR	SAS
Intercept	11.4194	10.2063	10.2882
YD	0.1765	0.1648	0.1651
TP	0.4364	0.3904	0.3915
GN	0.0449	0.0411	0.0412
KGW	0.2155	0.2033	0.2027
GL	0.5145	0.4875	0.4901
GW	1.4526	1.3152	1.3258
HD	0.0280	0.0250	0.0250

2.3.2 Association between the "dark purple pericarp color" and a molecular marker in rice

The purple color trait is controlled by a gene (OsC1) located on chromosome 6 at position 31.014 cM. This gene is within Bin868 (a marker) of the rice genome. We choose a neighboring marker (Bin871) that is about 3 cM away from Bin868 for this association study. The genotype data for Bin871 are also presented in **Appendix B-S1**. The probit model for this association study is

$$\Phi^{-1}(\mu) = \beta_{INT} + X_{BIN871}\beta_{BIN871} \tag{61}$$

Results from three algorithms (EM, NR and SAS) under two models (the probit and logistic models) are shown in **Table 2.4**. Again, all three algorithms produced almost

identical results under both the probit model and the logistic model. The test statistics, however, are different between the probit and the logistic models. This is because the Wald test statistics are extremely high, leading to extremely small p-values. The small model (a single independent variable) allows us to show the iteration processes of the EM algorithm and the NR algorithm under the two models (**Figure 2.1**). Clearly, the EM algorithm took more iterations (about 40) to converge at a predetermined criterion while the NR algorithm took just 6 iterations to converge at the same criterion.

Table 2.4 Estimated parameters for association between the color trait and Bin871.

		EM		NR		SAS	
Model	Parameter	Estimate	StdErr	Estimate	StdErr	Estimate	StdErr
Probit	Intercept	-1.6041	0.2145	-1.6041	0.2145	-1.6041	0.2145
	Effect	3.4300	0.3082	3.4304	0.3082	3.4304	0.3082
Logistic	Intercept	-2.8565	0.4599	-2.8565	0.4599	-2.8565	0.4599
	Effect	6.2059	0.6857	6.2064	0.6858	6.2063	0.6858

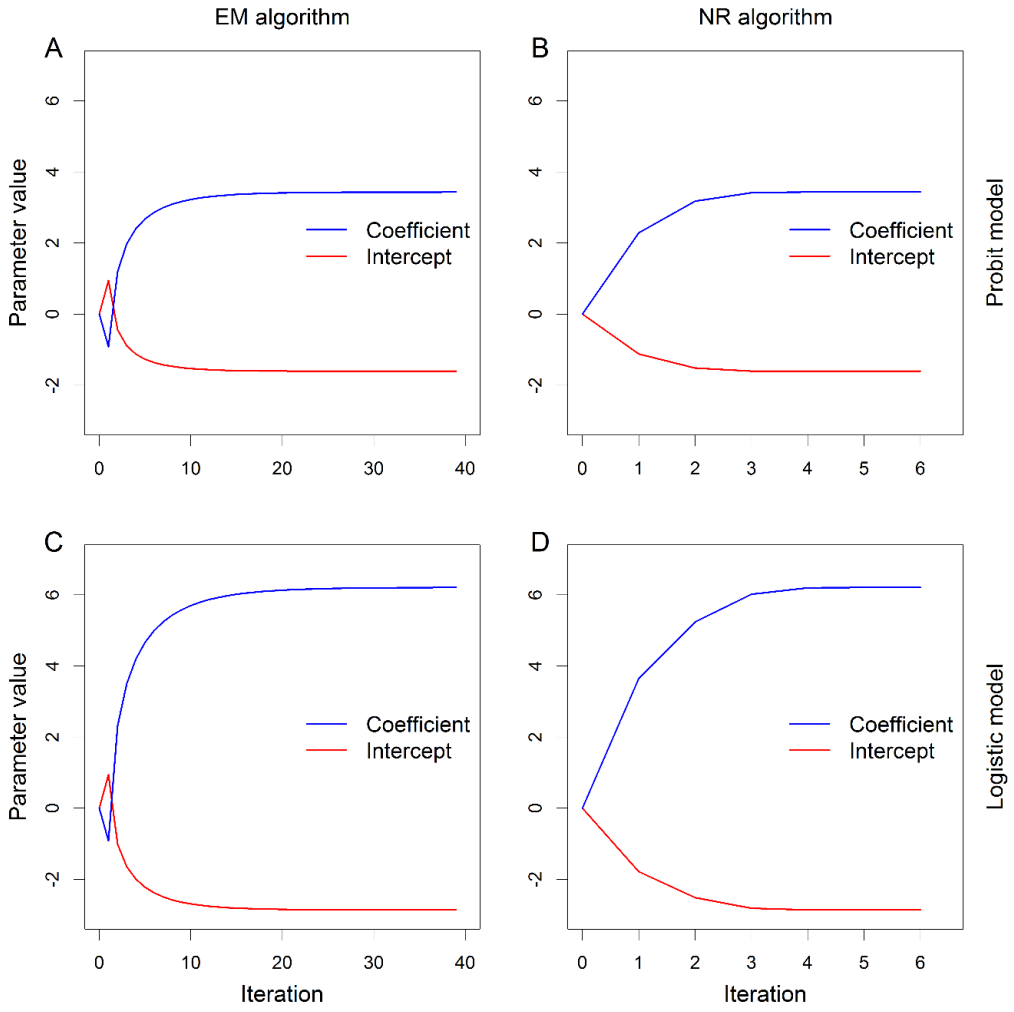


Figure 2.1 Convergence processes for parameters of the regression of color on Bin871. The first row (panels **A** and **B**) shows the result from the probit model analysis and the second row (panels **C** and **D**) shows the result from the logistic model analysis. The first column (panels **A** and **C**) shows the result from the EM algorithm and the second column (panels **B** and **D**) shows the result from the NR algorithm.

2.3.3 Genome scanning for the "dark purple pericarp color" trait in rice

The rice genome in this RIL population consists of 1619 bins (a bin is a block of SNP markers with identical segregation pattern). We scanned the entire genome to show the estimated marker effect for every bin of the genome. **Figure 2.2** illustrates the genome-

wide estimated parameters (intercepts and bin effects) for the color trait. First, there is no visual difference between the EM algorithm and the NR algorithm in both the estimated intercepts (upper panels) and the estimated marker effects (lower panels). Secondly, the estimated parameters (intercept and effect) from the logistic model analysis deviated more from zero than the effects from the probit model analysis. The estimates from the logistic model are approximately $\pi/\sqrt{3}=1.8138$ times the estimates from the probit model. Thirdly, the estimated parameters (intercept and effect) peak at Bin871 from chromosome 6, which is 3 cM away from Bin868 where the OsC1 gene is located. Three bins (Bin866, Bin867 and Bin868) have been excluded from the plot because the segregation patterns of the three bins match the color phenotypes and thus no legal estimates are available for the three bins.

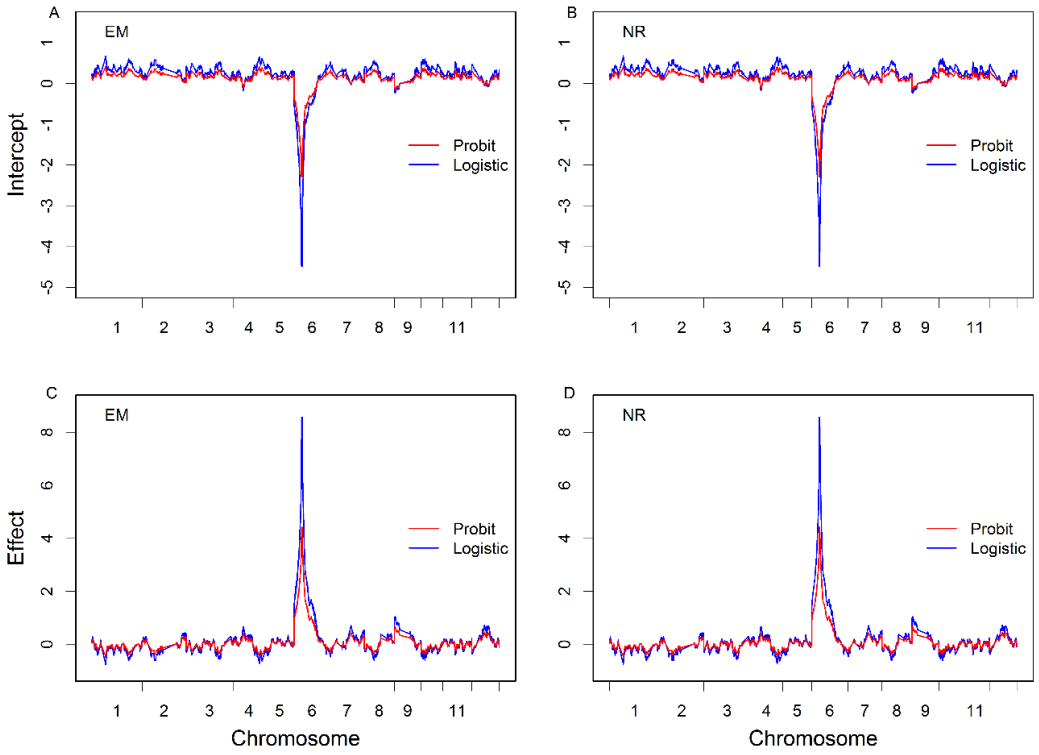


Figure 2.2 Genome-wide estimates of parameters for the color trait of the rice population. The first row (panels **A** and **B**) shows the intercept and the second row (panels **C** and **D**) shows the regression coefficient (effect). The first column (panels **A** and **C**) shows the result of the EM algorithm and the second column (panels **B** and D) shows the result of the NR algorithms.

We now compare the estimated parameters from the EM algorithm and those from the NR algorithm under the probit model and the logistic model. **Figure 2.3** shows the plots of the estimated parameters (intercept and regression coefficient) from the NR algorithm against the estimates from the EM algorithm for the 1619 bins. All points are on the diagonal lines, indicating that the EM algorithm and the NR algorithm produced identical results.

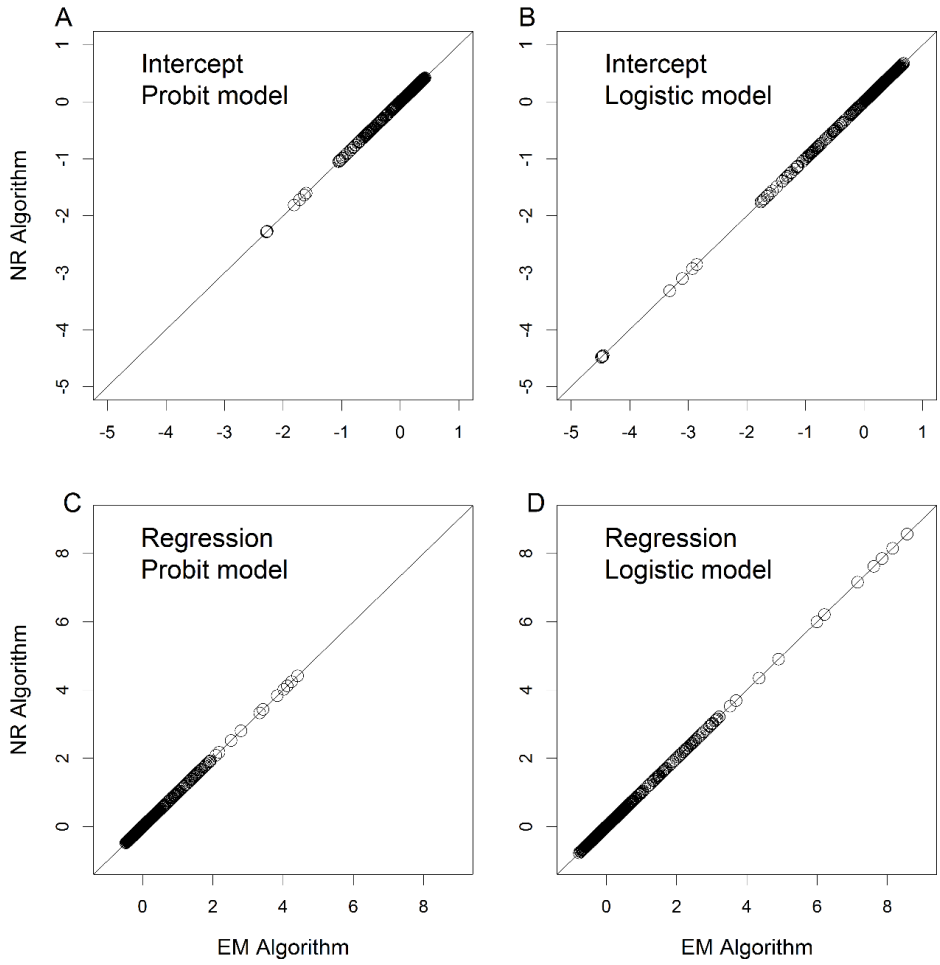


Figure 2.3 Comparison of estimated parameters from the EM algorithm with the estimates from the NR algorithm. The first row (panels **A** and **B**) shows the comparison of the intercept. The second row (panels **C** and **D**) shows the comparison of the regression coefficient (effect). The first column (panels **A** and **C**) shows the comparison from the probit model analysis and the second column (panels **B** and **D**) shows the comparison from the logistic model analysis.

Figure 2.4 compares the standard errors of the estimates from the EM algorithm and the standard errors from the NR algorithm. Again, all points are on the diagonals, indicating that the standard errors from Louis's (1982) method for the EM algorithm are the same as those from the NR algorithm.

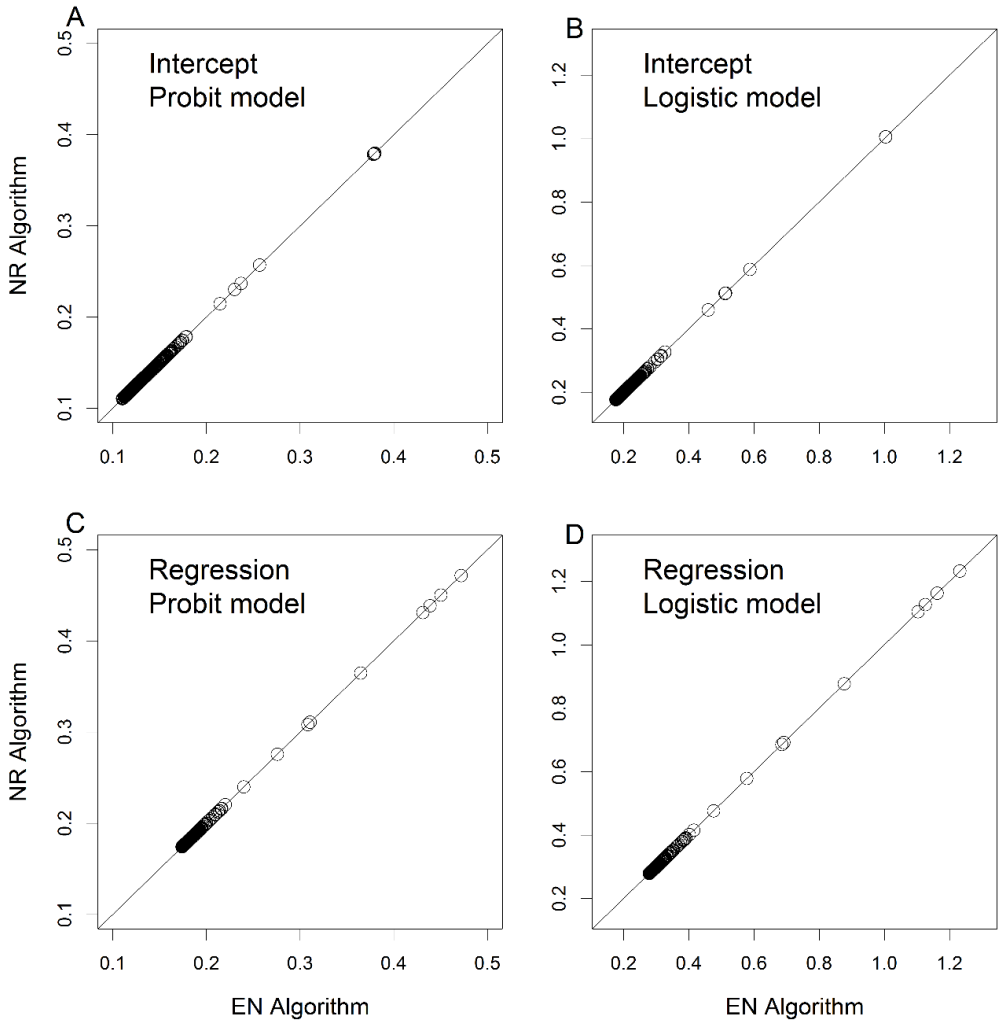


Figure 2.4 Comparison of standard errors of estimated parameters from the EM algorithm with the standard errors from the NR algorithm. The first row (panels **A** and **B**) shows the comparison of the intercept. The second row (panels **C** and **D**) shows the comparison of the regression coefficient (effect). The first column (panels **A** and **C**) shows the comparison from the probit model analysis and the second column (panels **B** and **D**) shows the comparison from the logistic model analysis.

The OsC1 gene is located within Bin868 on chromosome 6. This bin co-segregates exactly with the OsC1 gene. The NR algorithm cannot handle this degenerating issue properly, as shown in **Figure 2.5** where the estimated parameters at Bin868 are

drastically different from the nearby markers (Bin869, Bin870 and Bin871). This problem is called “complete separation” where the maximum likelihood estimates of the parameters do not exist (Hosmer and Lemeshow 1989; DeMaris 1995), . The segregations of Bin866 and Bin867 differ from the OsC1 gene (Bin868) by two and one individuals, respectively. These two bins are nearly identical to the OsC1 gene and thus the NR algorithm also failed to provide reasonable estimates of the parameters (see **Figure 2.5**). The EM algorithm, however, provides very good estimates of the parameters because the estimated parameters of the three bins are much the same as their neighboring bins (see **Figure 2.5**). This phenomenon demonstrates that the EM algorithm behaves better than the NR algorithm in handling special situations like this, although it took a very large number of iterations to converge.

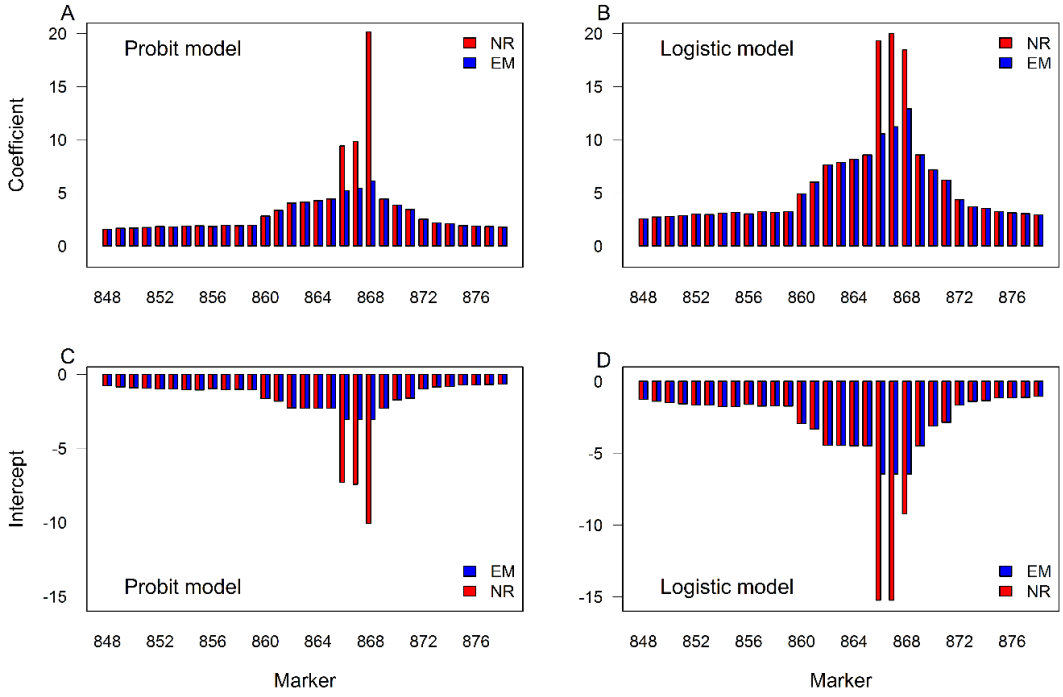


Figure 2.5 Estimated parameters (intercept and coefficient) for markers around the OSC1 gene (within Bin868) on chromosome 6 of the rice genome. Panel A shows the estimated marker effect (regression coefficient) from the probit model. Panel B shows the estimated marker effect (regression coefficient) from the logistic model. Panel C shows the estimated intercept from the probit model. Panel D shows the estimated intercept from the logistic model.

2.4 Discussion

Given the more general Newton-Raphson algorithm for the maximum likelihood estimation of parameters, why do we need the EM algorithm in the first place? Answers to this question depend on the problems to be addressed. In most problems, the first and the second order derivatives of the likelihood function do not have explicit forms. Numerical derivatives must be used to code the NR algorithm, which makes the iteration process much more like a black box. If the iteration process fails, it is hard to debug the

code. The EM algorithm, however, makes the coding much easier and users can easily code the EM algorithm themselves and thus have a better control for the iteration process. More importantly, the EM algorithm is often more stable than the NR algorithm (Camilleri 2009). This is because one of the properties of the EM algorithm: the objective function at the parameters of the next iteration always higher than the objective function at the parameters at the current iteration (Dempster et al. 1977). The NR algorithm, however, does not guarantee for the parameters to always move in the direction of increasing the objective function. Sometimes the parameters at the next iteration may cause the Hessian matrix not invertible and thus the iteration process gets crashed. This is why a Newton-Raphson ridge (NRR) algorithm may be adopted to improve the stability of the NR algorithm.

Unlike the NR algorithm that produces the variance matrix of the estimated parameters as a byproduct of the iteration process, the EM algorithm does not produce such a variance matrix in an automatic way. The Louis's (Louis 1982) information matrix must be derived if such a variance matrix is needed. In many problems, derivation of the Louis information matrix can be very complicated, which is often the criticism of the EM algorithm. However, situations where straightforward derivation of the Louis information matrix do exist. The EM algorithm for the probit regression analysis under the latent variable presented here is a typical example of the latter. In many cases, the EM algorithm is just an intermediate step of a larger problem and the variance of the estimated parameters is not needed.

The Louis information matrix for the EM algorithm (Louis 1982) and the information matrix for all the ML methods are only appropriate for large samples. For small samples where the asymptotic theory does not apply, Bayesian methods implemented via the MCMC algorithm are highly recommended because the empirical variance matrix of the parameters drawn from the posterior samples is less biased. If the MCMC algorithm is too costly in terms of computational time required, the bootstrap method (Efron 1979) should be adopted to obtain a less biased variance matrix, as demonstrated in the rice data analysis from this study.

The logistic regression is more often used to analyze binary traits than the probit regression. One of the reasons is that the CDF of the logistic distribution is explicit while the CDF of the normal distribution is not and it involves numerical integration. Prior to the advent of the popular computers, numerical integration was computationally costly. Being able to avoid numerical integration in the CDF of the logistic distribution is a big advantage over the normal distribution. The logistic regression under the latent variable assumption indicates that the error of the liability follows a standard logistic distribution. An EM algorithm similar to the probit regression has not been developed for the logistic regression analysis. We simply adopted the formulas of the EM algorithm from the probit regression to the logistic regression. To our surprise, the same EM formula works perfectly for both the probit regression and the logistic regression.

We tried to derive the EM algorithm for the logistic regression anew and realized that this was a bad example to use the EM algorithm. Given the PDF of the logistic distribution,

$$\theta(\xi_j - X_j\beta) = \frac{\exp\left[-(\xi_j - X_j\beta)\right]}{\left\{1+\exp\left[-(\xi_j - X_j\beta)\right]\right\}^2} \tag{62}$$

The complete-data log likelihood function is

$$L_C(\beta) = -\sum_{j=1}^{n}(\xi_j - X_j\beta) - 2\sum_{j=1}^{n}\ln\left\{1+\exp\left[-(\xi_j - X_j\beta)\right]\right\} \tag{63}$$

The expectation of the complete-data log likelihood function involves the expectation of a natural log function. Neither the first order derivative nor the second order derivative has an explicit form as what we see in the complete-data likelihood function for the normal distribution.

Generalized linear models, e.g., the probit regressions, and generalized linear mixed models are routinely used in statistics. However, many biologists may not be familiar with the technology and thus often try to avoid using GLM and GLMM. This study introduces the EM algorithm for probit and logistic regressions in a language style that is easy to understand by biologists and thus promotes wide applications of the GLM and GLMM to biological problems.

Chapter 3 Whole Genome-based Insights Into the Phylogeny of Punica

Pomegranate (*Punica granatum* L.) is a perennial fruit tree and has been widespread worldwide as a traditional medical product for over 4000 years. *Punica protopunica* Balf. is one of the only two species of the *Punica* genera, considered the "sister" of *P. granatum*. However, due to its unique independent evolutionary line, it was also hypothesized as the ancestor of *P. granatum*., beyond the taxonomic classification. Phylogenetic relationships and diversification within the *Punica* genus are classic and hot scientific topics that have been elucidated by fossil, morphological, molecular and environmental data. Further resolution of relationships within the genus is still needed and can be achieved by analysis of whole genomic data. In this study, important pomegranate germplasm from the United States was sequenced to resolve the complication, including 40 accessions of 2 species: *P. granatum*. and *P. protopunica*. We assembled the genomes, predicted and annotated genes, and identified orthologous coding sequences, which were then used to investigate the relevance and power of phylogenomic relationship inference. The phylogenetic tree of the whole genome data yielded highly node-supported, indicating *P. protopunica*. as both sister and ancestor groups of *P. granatum*., which was consistent with the traditional taxonomy. Our framework provides an efficient and inexpensive methodology for characterizing any unknown cultivars. These analyses also provided for a robust number of annotated genes among accessions to conduct studies on the genomic underpinnings of essential traits

such as pest and disease resistance, seed hardness, flavor, tree size, and reduced fruit cracking rates, among others. Moreover, the data here provide a valuable resource for analyzing pomegranates and facilitating future breeding and trait association studies.

3.1 Introduction

The rapid development of next-generation sequencing (NGS) has transformed the field of molecular phylogenetics into phylogenomic, where genome-scale data reconstructs the evolutionary biology of organisms (Kapli, Yang, and Telford 2020; Young and Gillung 2020). Traditional molecular phylogenetic studies include relatively few loci and are therefore limited by stochastic or sampling error (Young and Gillung 2020). This obstacle can be addressed successfully using much larger sequencing data, as modern phylogenomics analysis makes use of hundreds to thousands of loci across the whole genome (Zhang et al. 2019; Sims et al. 2009). Furthermore, whole genomes are particularly suited to resolving evolutionary relationships where sequence variation is limited by taxonomic level, early divergence, significant differences in morphology, rapid speciation, or slow genome evolution (Zhou et al. 2021).

Pomegranate (*P. granatum*) has been an ancient fruit tree since prehistoric times and becoming an arising profitable crop due to its attractive features, such as its bright appearance and abundant medicinally valuable compounds (Chandra et al. 2010). This fruit belongs to the family *Punicaceae* Horan. (*Lythraceae* Jaume St.-Hil.), which contains a single genus *Punica* L., with two species: *P. granatum* and *P. protopunica*. *P.*

granatum has its natural distribution in central Asia, from Iran to northern India, and spreads to the Mediterranean basin, East Asia, North and South America, and South Africa (Chen, Zhang, and Yuan 2019; Ja et al. 2020). The second species in *Punica*, *P. protopunica*, is only distributed on the Yemeni island of Socotra of the Arabian Peninsula and is considered an ancestral species (Zeynalova 2017) or an independent evolutionary branch (Chandra et al. 2010). This species exhibits several morphological differences compared with *P. granatum*, i.e., larger and coarser leaves, different foliage, smaller fruit size and pink flower, evergreen, continuous flowering, and white seeds (Ja et al. 2020). Even if it is the only congeneric species of *P. granatum* (2n = 16), the haploid number of *P. protopunica* (2n = 14) chromosomes is n = 7, unlike n = 8 in *P. granatum* (Teixeira da Silva et al. 2013). The difference was considered a primitive characteristic of *P. protopunica* as an ancestor of the domesticated species *P. granatum.* Recently, several researchers studied the genetic diversity and relationship between the two species based on morphological and biochemical characterization, molecular markers, and genotypes. Youssef et al. (Youssef et al. 2018) and Mohammad et al. (Shahsavari et al. 2022) supported the hypothesis that *P. protopunica* could be an ancestor of *P. granatum*. However, those genetic studies of pomegranate were based on analysis of selected loci, and the evolution of the *Punica* genera remains a difficult problem. Then whole genome-scale phylogenomic studies can be helpful in supplementing previous research.

The primary goal of this study was to increase the resolution of the molecular phylogeny of *Punica* genera by maximizing the number of taxa sampled and the number of genetic

markers used. We selected 40 pomegranate accessions, including 38 accessions of *P. granatum* and 2 accessions of *P. protopunica*, making up most of the available pomegranate germplasm from the United States. Our study presents a procedure for inferring complete genus-level phylogenies from averaged 16.3X Illumina genome data, making this the most comprehensive study to date. The pipeline builds on existing methods to (1) *de novo* assembly of 40 pomegranate accessions, (2) predict and annotate assembled genomes, (3) retrieve orthologous genes, and (4) reconstruct a phylogenomic tree (**Figure 3.1**). In addition to evaluating the hypothesis on relationships within the *Punica* genus, this study provides a valuable and complete analytical framework for phylogenomic analysis.

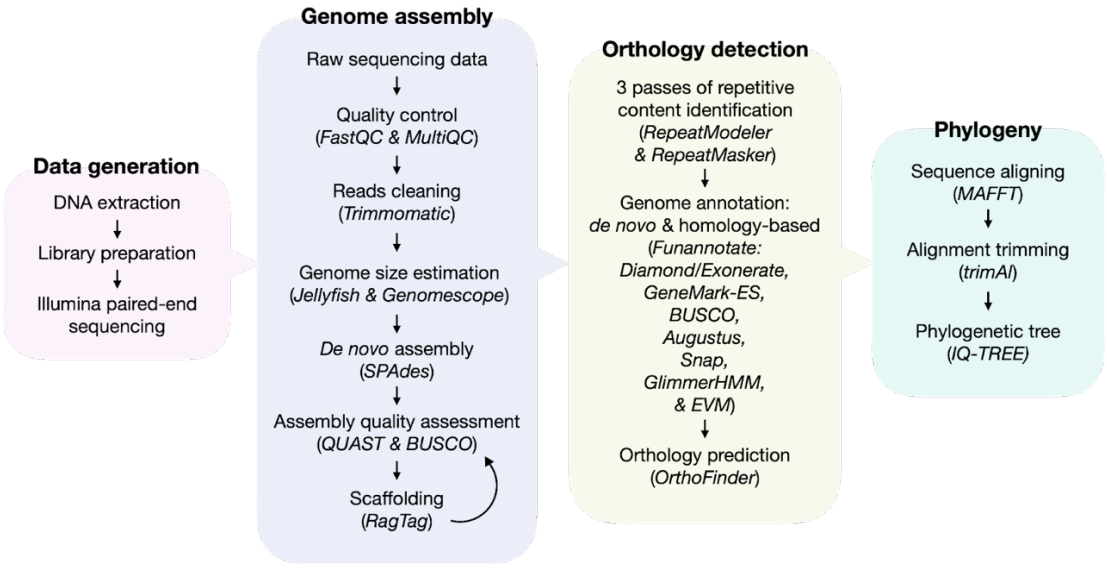


Figure 3.1 The flowchart illustrates creating and validating sequence data for 40 pomegranate accessions.

3.2 Materials and methods

3.2.1 Taxon sampling and sequencing

Forty samples were sent to the company for whole genome sequencing. Genomic DNA samples were individually processed in paired-end Illumina using TruSeq DNA libraries and Illumina-compatible barcoded DNA adaptors with an average insert size of 338 bp.

3.2.2 Quality check and pre-processing raw sequencing data

FastQC v.0.11.9 (Andrews 2010) and MultiQC v.1.10 (Ewels et al. 2016) were performed to access the quality of sequencing data as it calculates statistics about the composition and quality of raw sequences. Given the quality report, raw reads were cleaned using Trimmomatic v.0.36 (Bolger, Lohse, and Usadel 2014) by removing low-quality bases from their beginning (LEADING:5) and the end (TRAILING:5), by eliminating reads below 50 bp (MINLEN:50), by evaluating read quality with a sliding window strategy (SLIDINGWINDOW:5:15), and by trimming 'TruSeq3' adaptors.

3.2.3 Genome size estimation and heterozygosity

A k-mer count analysis was done using Jellyfish v.2.2.10 (Marçais and Kingsford 2011) on the trimmed sequencing data to avoid the lower quality part of the read. After converting the 17-mer counts into a histogram format, this file was analyzed using the Genomescope v.2.0 (Vurture et al. 2017) tool for genome size, heterozygosity ratio and repeat length.

3.2.4 *De novo* Assembly of whole genomes and evaluation

The trimmed sequencing data was preliminarily assembled using SPAdes genome assembler v.3.15.5 (Prjibelski et al. 2020) in '-carful' mode. The kmer size values 21, 33, 55, and 77 were tested for assemblies. RagTag v.2.1.0 (Alonge et al. 2021) was performed for scaffolding and improving genome assemblies with the draft reference genome. At last, the resulting assembly was analyzed for completeness and quality using QUAST v.5.0.2 (Gurevich et al. 2013) and BUSCO v.5.4.2 (Manni et al. 2021) with the eudicots_odb10 dataset.

3.2.5 Repetitive element identification

De novo and homology-based approaches were performed for repetitive content using three passes of the program RepeatMasker v.4.1.2-p1 (Smit et al. 2013) in soft-masking mode. An initial run was conducted using well-curated repeat libraries for the target organism. Simple, complex, and interspersed repeats are annotated using repeat consensus sequences from *Myrtales* included in the RepBase and DFam 5.0 (Bao, Kojima, and Kohany 2015; Hubley et al. 2016). The result was then passed into the second and third run of RepeatMasker with custom, species-specific known and unknown repeat libraries generated using RepeatModeler v.2.0.3 (Smit et al. 2008). Finally, three rounds of outputs were combined and summarized to create a GFF3 file, which is compatible with downstream annotation software that interprets masking.

3.2.6 Genome annotation

The soft-masked genome was annotated using funannotate v.1.8.13 (Palmer and Stajich 2019) with the options '--repeat2evm --organism other --busco_db eudicots_odb10', and the flag '--max_introlen' was set to 272000. Funannotate used Evidence Modeler (EVM) v.1.1.1 (Haas et al. 2008) to combine *ab initio* gene model predictions with protein evidence aligned to the draft reference genome (Luo et al. 2020) and the UniProtKB/SwissProt curated protein database. The protein evidence was mapped to the assembled genome using Diamond v.2.0.14 (Buchfink, Reuter, and Drost 2021) and Exonerate v.2.4.0 (Slater and Birney 2005). *Ab initio* gene predictions were synthesized using self-training GeneMark-ES v.4.69_lic (Brůna, Lomsadze, and Borodovsky 2020) and combined with identified BUSCO conserved orthologs as inputs to EVM. After double-checking that EVM BUSCO consensus models are correct, they were used to train Augustus v.3.3.3 (Stanke et al. 2006) to obtain high-quality Augustus predictions (HiQ). The BUSCO training set was also conveyed to train SNAP v.2006-07-28 (Korf 2004) and GlimmerHMM v.3.0.4 (Majoros, Pertea, and Salzberg 2004). Finally, the EVM combines all *ab initio* gene predictions and protein alignments into weighted consensus gene structures to generate a final annotation file.

3.2.7 Orthology detection and alignment cleaning

The newly annotated protein sequences from the forty pomegranate accessions were used to identify orthologous proteins with OrthoFinder v.2.5.4 (Emms and Kelly 2019). Multiple sequence alignments were achieved by MAFFT v.7.505 (Katoh et al. 2002) with

the ‘auto’ setting. The alignment trimming was conducted using TrimAl v.1.4.rev15 (Capella-Gutiérrez, Silla-Martínez, and Gabaldón 2009) with recommended ‘-automated1’ parameter.

3.2.8 Phylogenetic tree inference

Maximum likelihood models (LG+F+G & PMSF) were implemented for phylogenetic tree inference using IQ-TREE v.2.2.0.3 (Nguyen et al. 2015). Phylogenomic analyses were performed using ML methods on concatenated amino-acid datasets of selected 7120 orthologous proteins. First, an ML analysis was performed using a single LG model (Le and Gascuel 2008) for amino acids, one discrete gamma rate category (+G4 option) , and empirical among acid sequences estimated from data (+F option). In addition, node supports were calculated with 1000 ultrafast bootstrap replicates. These ML analyses assumed a single rate matrix for the whole data; however, rate heterogeneity is widespread in phylogenomic data sets and should be considered. The posterior mean site frequency (PMSF) (Wang et al. 2018) is the amino-acid profile for each alignment site computed from an input mixture model and a guide tree. A second ML analysis was performed in the ‘MFP’ mode, which made the IQ-TREE perform ModelFinder (Kalyaanamoorthy et al. 2017) to determine the best-fit model.

3.3 Results

3.3.1 Genome size estimation

Plant genomes span several orders of magnitude in size, gene duplications, alternative gene splicing, ploidy and gene retention following genome duplication, which make plant genomes large and complex (Claros et al. 2012). Therefore, essential characteristics of genomes must be deplored before selecting appropriate plant genome analysis methods, especially *de novo* assembler. For example, genome size appears to be related to the type of interspersion. Plant species with smaller genomes have longer interspersion and smaller lengths of repetitive sequences (Cvrčková 2016; McKain et al. 2018). Furthermore, high heterozygosity could introduce false segmental duplications in assemblies when heterozygous sequences from two haplotypes are assembled into separate contigs and scaffolded adjacent to each other rather than merged (Claros et al. 2012; Voshall and Moriyama 2018). These characteristics in advance can reveal if the following analysis could handle the full complexity of the genome. In our study, the Illumina data were analyzed for k-mer depth frequency distribution to estimate the genome size, heterozygosity and the number of repetitive sequences in the forty pomegranate accessions. The estimated genome size was calculated by the ratio of the total number and the average depth of the 17-mers. The estimated genome size ranges from 326 (‘Blaze’) to 371M (‘*Punica_protopunica*_S14258A’) (**Table 3.1**), which is approximately close to several published draft reference genomes size of 320-362 Mb (Luo et al. 2020; Qin et al. 2017; Yuan et al. 2018; Usha et al. 2022). The heterozygosity is from 0.23% (‘Toryu_Shibori’) to 0.57% (‘*Punica_protopunica*_S14258A’), and the

amounts of repetitive sequences were roughly between 51.5% to 54.1%. The reported characteristics showed the low-complicated personality of the forty pomegranate accessions.

Accessions	Species	Origin	Population	# Raw reads	Estimated genome size (M)	Mean %GC	Duplication ratio	SPAdes scaffolds N50 (K)	SPAdes scaffolds Genome fra	SPAdes scaffolds BUSCO	RagTag scaffolds N50 (M)	RagTag scaffolds Genome fra	RagTag scaffolds BUSCO	Masked repeats	Found genes	% genes in orthogroups	Protein length
Blaze	*P. granatum* L.	United States	Winters	37324300	326.48	39.18	1.033	19.056	75.165	88.80%	33.358853	79.471	92.00%	47.75%	27955	99.7	359.21
cv_857	*P. granatum* L.	United States	Somis	38296330	330.78	39.14	0.987	20.795	75.365	88.60%	33.600346	83.138	91.60%	48.83%	28042	99.7	365.85
Eversweet	*P. granatum* L.	United States	Riverside	43200256	328.11	39.25	0.998	19.181	78.207	89.10%	33.480797	81.873	92.10%	49.27%	28905	99.7	365.31
Haku_Botan	*P. granatum* L.	Japan	Riverside	33209640	337.24	39.05	1.047	16.385	74.589	86.80%	32.66677	77.08	92.00%	47.36%	29495	99.5	354.22
Kopetdag	*P. granatum* L.	Turkmenistan	Winters	45948412	332.97	39.07	1.009	26.336	75.738	89.50%	33.215102	80.28	92.10%	44.69%	29398	99.3	360.82
Myagkosemyannyi_Rosovyi	*P. granatum* L.	Turkmenistan	Winters	41366742	341.67	39.21	1.049	18.106	75.431	88.80%	33.486146	78.282	91.70%	48.74%	29222	99.6	357.45
Parfianka	*P. granatum* L.	Turkmenistan	Riverside	41380474	356.37	39.15	1.027	20.854	75.561	88.10%	33.591075	79.913	91.50%	48.43%	28409	99.7	359.7
Phoenicia	*P. granatum* L.	United States	Riverside	35908652	----	39.15	0.982	11.763	74.113	82.90%	33.167209	82.405	91.10%	49.28%	27564	99.5	356.43
Don_Somner_North	*P. granatum* L.	United States	Georgia	40002888	349.22	39.14	0.995	23.028	75.494	89.30%	33.475968	81.714	92.10%	49.25%	28803	99.8	371.11
Sakerdze	*P. granatum* L.	Russia	Florida	37436660	343.98	39.08	0.976	18.325	75.021	87.50%	33.376368	83.928	91.40%	48.43%	27803	99.7	360.59
Salavatski	*P. granatum* L.	Russia	Florida	38978958	355.28	39.11	1.007	17.166	75.059	88.00%	33.271788	81.397	91.70%	49.02%	29036	99.6	355.88
Surh_Anor	*P. granatum* L.	Russia	Florida	43996404	354.37	39.13	1.055	20.47	75.477	87.80%	33.821432	78.332	91.40%	49.64%	27662	99.7	365.89
Sweet_32	*P. granatum* L.	United States	Florida	43219202	339.94	39.15	1.023	21.31	78.642	88.40%	33.943949	81.078	91.30%	49.72%	27888	99.7	361.53
Vkusnyi	*P. granatum* L.	Turkmenistan	Florida	39708002	345.41	39.15	1.048	20.941	75.648	88.40%	33.553404	78.38	91.70%	49.56%	28086	99.8	361.17
Afganski	*P. granatum* L.	Russia	Florida	43674342	346.91	39.17	1.036	19.853	77.352	88.20%	33.899456	80.007	91.40%	49.43%	28919	99.6	355.17
Sirenevyi	*P. granatum* L.	Turkmenistan	Florida	45780618	354.45	39.21	1.05	19.8	75.756	87.90%	33.722447	78.691	91.60%	50.89%	28163	99.8	361.17
Cedar_Key_Sunset	*P. granatum* L.	United States	Florida	32240448	347.38	39.07	1.043	15.308	74.739	87.10%	32.472743	77.238	92.00%	47.43%	29535	99.5	357.19
Punica_protopunica_Hawaiian	*P. protopunica* Balf.	United States	Riverside	27888088	----	39.15	1.044	7.881	21.661	77.20%	22.859811	----	91.10%	41.48%	30182	95.7	341.38
Punica_protopunica_S14258A	*P. protopunica* Balf.	United States	Los Angeles	39645398	371.36	39.45	1.081	16.248	21.881	84.00%	24.446733	----	91.10%	52.97%	29092	96.3	342.8
Toryu_Shibori	*P. granatum* L.	Japan	Winters	45532494	363.02	39.21	0.983	20.276	76.322	88.50%	33.55831	83.844	91.80%	50.12%	29464	99.5	356.57
Al_Sirin_Nar	*P. granatum* L.	Russia	Florida	35450402	341.61	39.12	1.017	15.865	74.924	87.50%	33.231209	80.256	91.80%	48.19%	27853	99.7	359.96
Angel_Red	*P. granatum* L.	United States	Florida	42988416	352.56	39.18	0.957	19.07	75.558	87.70%	33.954543	86.564	91.50%	49.81%	27706	99.6	367.58
Arakta	*P. granatum* L.	India	Florida	47313818	338.12	39.2	1.039	25.38	76.316	89.30%	33.345423	78.86	91.90%	48.64%	28885	99.7	362.53
Azadi	*P. granatum* L.	Turkmenistan	Florida	47067084	352.33	39.21	1.008	21.957	75.926	88.90%	33.945513	82.104	92.00%	50.27%	28214	99.7	367.07
Christina	*P. granatum* L.	United States	North Florida	42008818	340.48	39.19	1.044	22.624	75.696	89.70%	33.760704	78.541	91.80%	49.97%	28338	99.7	360.55
Desertnyi	*P. granatum* L.	Turkmenistan	Florida	37925178	353.6	39.13	0.991	20.793	75.009	89.00%	33.042859	82.221	91.70%	48.23%	28333	99.7	362.44
Eversweet_Florida_budline	*P. granatum* L.	United States	Florida	36990116	346.56	39.08	1.051	18.287	74.978	88.30%	33.342917	77.693	91.50%	48.19%	27880	99.7	360.51
Fleischman	*P. granatum* L.	Unknown	Florida	45525178	339.29	39.16	1.06	20.18	78.12	87.90%	34.051444	78.554	91.50%	50.28%	28341	99.6	358.99
Gainey_Sweet	*P. granatum* L.	United States	Georgia	41081070	355.01	39.18	1.046	22.427	75.85	88.90%	33.603635	78.582	91.70%	49.83%	29413	99.7	357.64
Girkanets	*P. granatum* L.	Turkmenistan	Florida	44356004	349.49	39.13	1.058	20.415	75.381	88.00%	33.957273	78.359	91.60%	49.30%	28508	99.6	357.18
Gissarskii_Rozovyi	*P. granatum* L.	Turkmenistan	Florida	44243848	350.44	39.17	1.05	22.99	75.67	88.40%	33.695648	78.501	91.80%	49.68%	28214	99.8	365.6
Jimmy_Roppe	*P. granatum* L.	United States	Georgia	46452958	350.22	39.15	1.021	21.634	77.082	88.40%	33.868604	81.531	91.70%	50.07%	27592	99.7	362.1
Larkin	*P. granatum* L.	United States	Florida	40884830	342	39.1	1.059	18.844	75.235	87.60%	33.760954	78.057	91.20%	49.14%	27626	99.7	358.51
Mack_Glass	*P. granatum* L.	United States	Florida	39870808	344.72	39.08	1.057	18.409	75.021	88.20%	33.68646	77.862	91.40%	48.72%	28683	99.6	354.78
Nikitski_Ranni	*P. granatum* L.	Turkmenistan	Florida	43060376	357.87	39.11	1.057	19.233	75.146	87.50%	33.90975	77.98	91.50%	49.00%	27689	99.7	359.02
Rosavaya	*P. granatum* L.	India	Florida	39180354	348.77	39.1	1.055	17.746	75.146	87.70%	33.647707	77.944	91.50%	48.69%	28762	99.6	356.24
Bala_Miursal	*P. granatum* L.	Azerbaijan	Florida	40365516	357.45	39.16	1.006	17.322	77.562	87.50%	34.055343	82.324	91.50%	49.33%	28706	99.6	357.79
Bhagwa	*P. granatum* L.	India	Florida	40574406	343.18	39.17	1.004	18.748	79.53	88.40%	33.382968	81.68	91.90%	48.92%	29563	99.7	354.17
Boris_2	*P. granatum* L.	United States	North Carolina	36597783	343.43	39.25	1.005	13.925	75.513	87.30%	33.395903	81.518	92.10%	49.12%	28457	99.6	365.15

Table 3.1 Taxon sampling and genomic results of 40 pomegranate accessions. ---- means we have too low of coverage for the model to confidently identify the peaks corresponding to the homozygous kmers. This can be recognized by a lack of any peaks in the kmer plots or that the model fit doesn't match the observed kmer profile very well.

3.3.2 Genome assembly

We assembled the genomes using SPAdes, which contains reads error and mismatch correction tools in resulting contigs and scaffolds. To establish a near complete genome, RagTag was performed for automating assembly scaffolding and patching. To evaluate the accuracy and completeness of the SPAdes and RagTag genome assembly, we first compared the total length of scaffolds, the number of scaffolds, N50, and genome fraction percentage when scaffolds aligned to the draft reference genome. The total average length of SPAdes scaffolds was about 204.48 Mb and increased to 262.67 Mb after RagTag scaffolding, the average N50 was improved significantly from 19.22 Kb to 33.55 Mb (**Table 3.1**), and the average number of scaffolds (larger than 5 Kb) dropped from 11609 to 414. In addition, the average genome fraction was grown from 73.15% to 80.11% (**Table 3.1**), though having two exceptions, the 'Punica_protopunica_S14258A' and 'Punica_protopunica_Hawaiian' were roughly 21%. This result agrees with our preknowledge that the two accessions are relatively unrelated to the others.

Additionally, we have assessed the integrity of the genome assembly with single-copy orthologs with the eudicots_obd10 database. The RagTag assembly contained approximately 91.64% of the 2326 conserved eudicots genes, higher than 87.73% in the SPAdes assembly (**Table 3.1**), as more fragmented BUSCOs were present in the SPAdes assembly. All in all, the results of these assessments indicate that the forty pomegranate genome assembly is considered complete and high-quality given 16X sequencing depth.

3.3.3 Repetitive content identification

The initial run using RepBase and DFam canonical repeat library identified a relatively small portion of the repetitive sequences. On average, 4.04% of the assembly comprises simple repeats, retroelements, a few known interspersed repeats and DNA transposons. And then, two more rounds of masking were implemented using a library of known and unknown repeats generated by RepeatModeler. These rounds were split so known elements would be preferentially annotated over the unknown to the degree possible. The known elements are mostly LTR repeats, especially Gypsy/DIRS1 groups of retrotransposons, accounting for about 21.82% of the assembled genome. Ultimately, results from each round were analyzed together to produce the final repeat annotation. An average of 48.91% of sequences were masked as repeats, and '*Punica_protopunica*_S14258A' ranked first (**Table 3.1**).

3.3.4 Gene prediction and orthology detection

We detected protein-coding genes in the *P. granatum* and *P. protopunica* genome assembly by a combination of methods: *Ab initio* and homology-based prediction. Overall, an average of 28504 genes, 28098 mRNAs, and 405 tRNAs were predicted and annotated, with an average exon number per gene of 4.72 and an average CDS length of 200 bp (**Table 3.1**).

Phylogenetic relationships should always be estimated based on sequences that are related by orthology (Young and Gillung 2020; Zhang et al. 2019). Orthogroups are

sequence clusters containing genes that descended via speciation from a single gene in the last common ancestor of all the species. Typically, if more than 80% of the genes are found in the orthogroups, we assume most of the critical genes are involved in the analysis. In our case, 99.5% of genes were assigned to orthogroups (**Table 3.1**), suggesting that nearly all genes are considered for ortholog detection. To limit gene duplication problems, we selected 7120 orthogroups with only one gene per species. Multiple sequence alignments and the last trimming alignments steps were performed for large-scale phylogenomic analyses. Gap/ambiguity percents of the output are below 1.8% on average, expect the two accessions of *P. protopunica*, with 3.57% and 3.70%, separately.

3.3.5 Phylogenomic analyses

Despite the fragmented nature of the genomes, we obtained a resolved and relatively supported phylogeny displaying the relationships of the forty accessions of pomegranate. Two models were implemented here to compare: the PMSF model, and the VT+F+I+I+R10, which IQ-TREE determines as the 'best-fit' model. No matter which model was performed, when the branch length is shown, we can easily observe that the two accessions of *P. protopunica* are distinctly related to all other *P. granatum* accessions (**Figure 3.2A, 3.3A**). The dramatic genetic distance recognizes them as two species, which is in agreement with previous studies (Youssef et al. 2018; Shahsavari et al. 2022). To better understand the relationships of the other accessions of *P. granatum*, we ignored the branch length temporarily. The node support of the PMSF model is much

stronger than the ‘best-fit’ model, especially for the circled part in **Figures 3.2 B and 3.3 B**. The two models gave the same topology except the showed part, indicating that this topology part is robust. However, due to potentially high similarities among the circled accessions, the bootstrap support of the circled part is relatively low, making unstable topology. Another vital piece of information was gained from here, the ‘Eversweet’ grown at the USDA germplasm repository seems to be a different genotype than the ‘Eversweet’ grown in Florida. As debates have surrounded the phylogenetic positions of the two *Punica* species, our studies suggest the *P. protopunica* as both sister and ancestor groups of *P. granatum*, as eighteen *P. granatum* accessions viewed *P. protopunica* as an ancestor with strong bootstrap support.

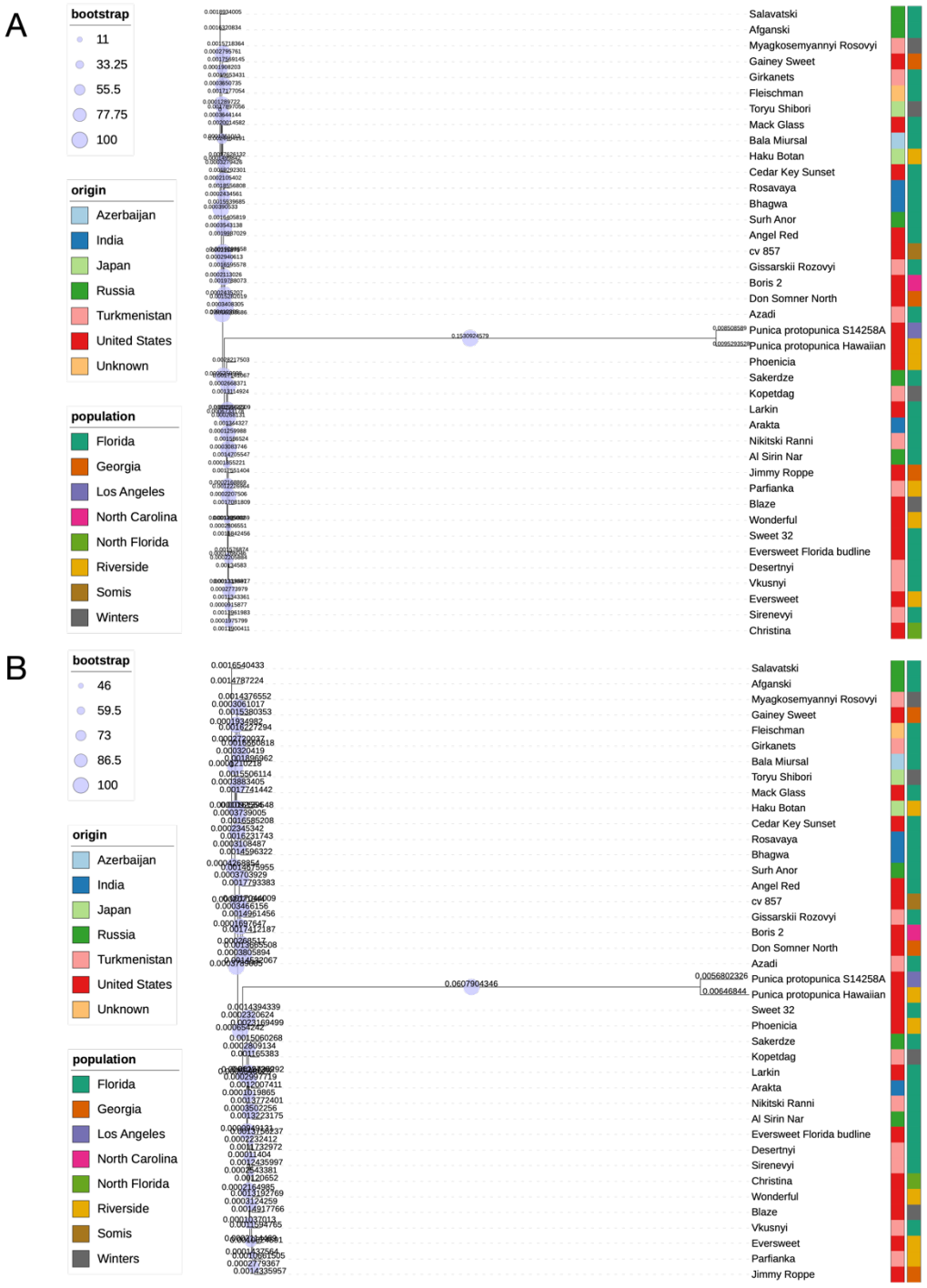


Figure 3.2 Phylogenomic relationships of pomegranate based on supermatrix analyses. The number on each line represents the branch length, and the purple circle represents support values. (**A**) 'best-fit' model detected in IQ-TREE (**B**) PMSF model.

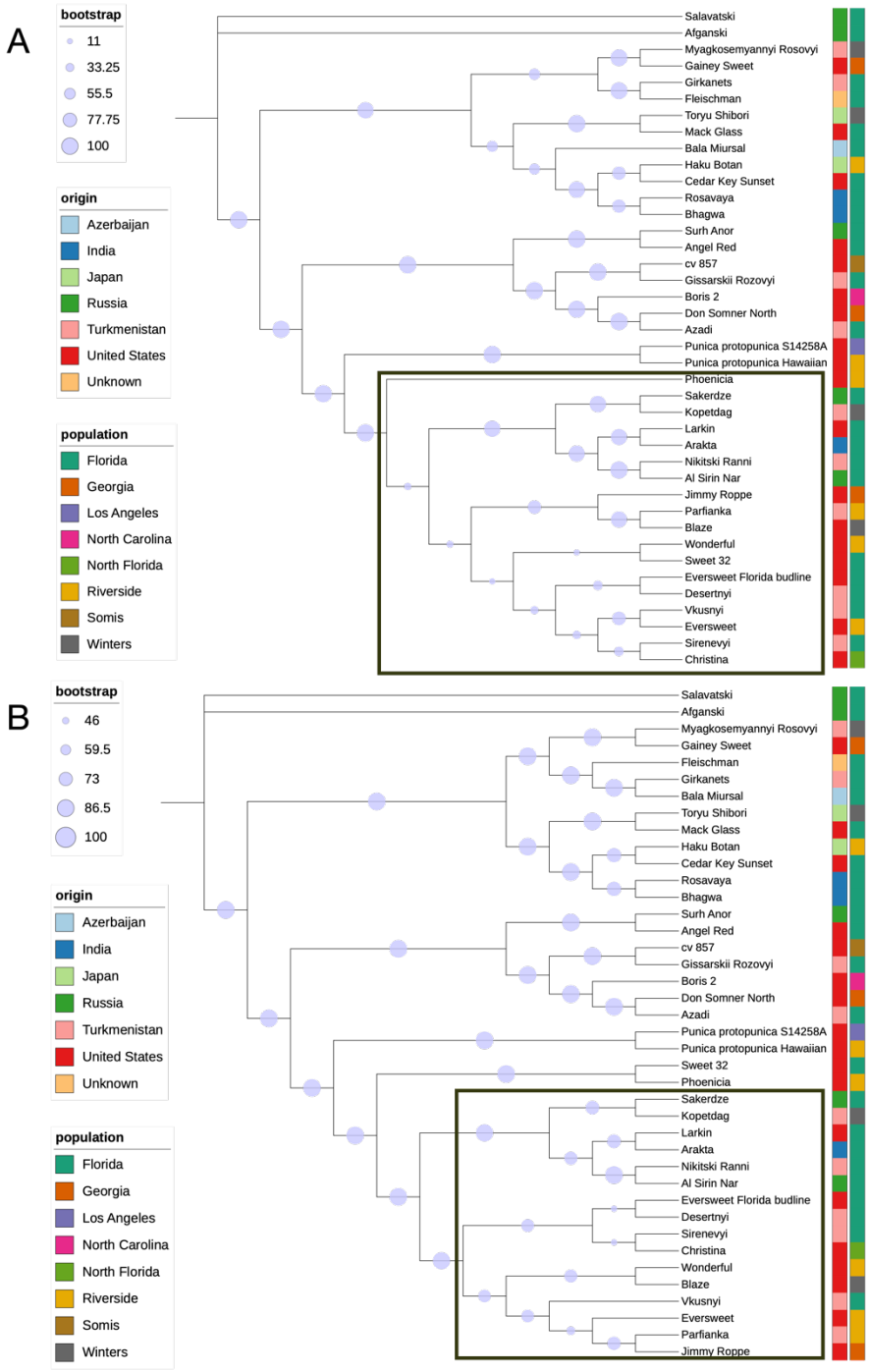


Figure 3.3 Phylogenomic relationships of pomegranate based on supermatrix analyses with branch length hiding. The purple circle represents support values. (**A**) 'best-fit' model detected in IQ-TREE (**B**) PMSF model.

3.4 Discussion

Pomegranate is one of the first cultivated fruits and is even considered sacred to many world religions and peoples. Evolutionary relationships have remained unsolved in most accessions of pomegranate, especially the precious germplasms in the United States. Whole genome sequencing allows us to resolve and understand evolutionary histories that are increasingly complex and complete. We still face substantial challenges in data accessibility and method shortcomings, such as few genomes available, method complexity and running time. Here, we meet the challenge of phylogenomic reconstruction by orthologous CDS identification from contigs obtained with whole-genome sequencing. Available annotation pipelines designed for highly fragmented and low-coverage genomes depend on a reference genome, selecting scaffolds similar to the reference protein (Allio et al. 2020). However, our phylogenomic pipeline has empowered the use of *ab initio* and similarity-based gene prediction in low-coverage genomic data, ensuring sufficient and novel transcripts are targeted even if the species is not closely related to the reference genome. Our study can provide productive perspectives for future research of other model groups and demonstrate the promising potential of low-coverage phylogenomic analyses. Moreover, based on the developed pipeline, we produced valued assembled genomes, well-annotated genes and new evidence of pomegranate classification, unveiling novel relationships and confirming previous hypotheses.

Chapter 4 HNF4A defines tissue-specific circadian rhythms by beaconing BMAL1::CLOCK chromatin binding and shaping rhythmic chromatin landscape

Transcription modulated by the circadian clock is diverse across cell types, underlying circadian control of peripheral metabolism and its observed perturbation in human diseases. We report that knockout of the lineage-specifying *Hnf4a* gene in mouse liver causes associated reductions in the genome-wide distribution of core clock component BMAL1 and accessible chromatin marks (H3K4me1 and H3K27ac). Ectopically expressing HNF4A remodels chromatin landscape and nucleates distinct tissue-specific BMAL1 chromatin binding events, predominantly in enhancer regions. Circadian rhythms are disturbed in *Hnf4a* knockout liver and HNF4A-MODY diabetic model cells. Additionally, the epigenetic state and accessibility of the liver genome dynamically change throughout the day, synchronized with chromatin occupancy of HNF4A and clustered expression of circadian outputs. Lastly, *Bmal1* knockout attenuates HNF4A genome-wide binding in the liver, likely due to downregulated *Hnf4a* transcription. Our results may provide a general mechanism for establishing circadian rhythm heterogeneity during development and disease progression, governed by chromatin structure.

4.1 Introduction

The circadian clock is a molecular oscillator that aligns behavior and physiology with daily light-dark cycles. The core of the mammalian circadian clock, composed of two interlocked transcriptional feedback loops, relies on chromatin occupancy of the master transcription factor heterodimer BMAL1::CLOCK at the E-box DNA element. BMAL1::CLOCK positively regulates expression of the *Period* (*Per1, Per2, Per3*), *Cryptochrome* (*Cry1, Cry2*), and *Rev-erb* (*Nr1d1, Nr1d2*) genes at the beginning of the feedback cycles. Protein dimer composed of PER and CRY suppresses the transcriptional activity of BMAL1::CLOCK, closing the first feedback loop. Formation of the second feedback loop is achieved by the nuclear receptor REV-ERBs to repress the transcription of *Arntl* (*Bmal1*) gene (and to a lesser extent on *Clock* gene) (Takahashi 2017).

While many peripheral organs have circadian clocks, the identities of rhythmic outputs are considerably divergent across tissues (Panda et al. 2002; Storch et al. 2002; R. Zhang et al. 2014; Ruben et al. 2018; Mure et al. 2018), contributing to organ-specific physiology and disorders associated with circadian misalignment (Bass and Lazar 2016). However, the molecular mechanisms involved in generating heterogeneous circadian rhythms remain unclear. Tissue-specific chromatin occupancy of the core clock transcription factors BMAL1::CLOCK and REV-ERB has been described, identifying co-occupancy of tissue-specific transcription factors (Perelis et al. 2015; Beytebiere et al. 2019; Y. Zhang et al. 2015). In the context of these prior studies, it will be intriguing to

apply genetic approaches to ascertain whether tissue-specific TFs influence clock TFs' loading onto chromatin, and the other way around.

In multicellular organisms, cells from different tissues exhibit specialized gene expression profiles in part achieved by physically sequestering unnecessary genes into heterochromatin. Genes that are required for particular tasks of a cell type display an accessible chromatin structure allowing for the binding of necessary machineries to facilitate gene expression (Clapier and Cairns 2009). Chromatin remodeling that opens condensed chromatin structures is initiated by the recruitment of lineage-specifying pioneer transcription factors to their target DNA sequences at enhancers. The pioneer TFs recruit histone methyltransferases MLL3/4 to deposit histone mark H3K4me1, whereby the condensed DNA wrapped around histones is loosened (Jozwik et al. 2016). Completely activated enhancers feature bimodal distribution of histone modifications H3K4me1 and H3K27ac, nucleosome depletion, and recruitment of other transcription factors and coactivators (Mayran and Drouin 2018). Instead of being simply correlated with chromatin accessibility, H3K4me1 has an active regulatory role by serving as docking sites for chromatin remodelers (Local et al. 2018). Due to the activity of ATP-dependent chromatin remodelers (Clapier et al. 2017) and three-dimensional chromatin folding (Yadon et al. 2013), chromatin remodeling commonly creates extended accessibility beyond the central nucleosomes that pioneer TFs bind.

With most (~16%) transcripts exhibiting circadian expression, the liver is the primary organ controlled by the circadian clock (R. Zhang et al. 2014). The hepatic circadian

transcripts are highly organ-specific and involved in most principal functions of the liver, including glucose homeostasis, lipogenesis, bile acid synthesis, mitochondrial biogenesis, oxidative metabolism, amino acid turnover, and xenobiotic detoxification. Indeed, environmental or genetic disruption of the circadian clock exacerbates the development of liver diseases such as non-alcoholic fatty liver disease (NAFLD), hepatitis, cirrhosis, and hepatocellular carcinoma (HCC).The hepatocyte nuclear factor 4A (HNF4A) is a nuclear receptor specifically expressed in the liver, kidney, pancreas, and intestinal tracts (Sladek et al. 1990). Mutation or dysregulation of the *Hnf4a* gene is associated with human diseases such as maturity-onset diabetes of the young (MODY) and HCC (Colclough et al. 2013; Hatziapostolou et al. 2011). Whole-body *Hnf4a* knockout resulted in embryonic lethality, and liver-specific knockout mice displayed severe hepatocyte differentiation defects and premature death by 8 weeks of age (W. S. Chen et al. 1994; Hayhurst et al. 2001a; Parviz et al. 2003). We previously demonstrated that HNF4A modulates peripheral circadian clocks in cell cultures (Qu et al. 2018). Here, we further interrogate the interface between HNF4A and the circadian clock in the liver tissue where they both play critical roles. We found that HNF4A supervises BMAL1 chromatin binding seemingly by remodeling chromatin accessibility. Synchronized with HNF4A recruitment (Qu et al. 2018), mouse liver displayed increased genome-wide chromatin accessibility during the night. Furthermore, the circadian clock contributes to chromatin remodeling likely through regulating HNF4A. Our results reveal a collaborative effort between HNF4A and the clock machinery in shaping tissue-specific chromatin landscape and circadian rhythms that are vital for liver biology.

4.2 Materials and methods

4.2.1 Raw Data

Raw data of ChIP-seq and ATAC-seq (fastq files) for NGS experiments are available on GEO under accession code GSE157452 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE157452]. GSE35262 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE35262] and E-MTAB-941 [https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-941/] were used to analyze PPARA, HNF1A, and LXR deposition at BMAL1 binding sites. GSE39860 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE39860] and SRA025656 [https://www.ncbi.nlm.nih.gov/sra/?term=SRA025656] were used for reanalysis of H3K4me1 circadian rhythms. CircaDB [http://circadb.hogeneschlab.org/] was used for identification of circadian transcripts.

4.2.2 ChIP-seq analysis

Single-end ChIP-seq reads were trimmed using Trimmomatic v.0.36 and then aligned to hg38 or mm10 genome with Bowtie2 v.2.3.4.1. BAM files were processed using SAMtools v.1.10 and PCR duplicates were removed with PicardTools v.2.18.3. Peaks were called in MACS2 v.2.1.2 using default settings and IgG mock ChIP files for normalization. BAM files of replicate samples were merged using SAMtools. BIGWIG track coverage files were generated from merged BAM files using the DeepTools v.3.3.0 bamCoverage command with RPGC normalization.

Heatmaps and metaplots were generated by the computeMatrix, plotHeatmap, and plotProfiles functions of DeepTools v.3.3.0 using BIGWIG files (replicates merged) and scaled regions. DiffBind v.3.2.7 was used to make PCA plots. Statistically significantly differential peaks were called and MA plots were generated by using the DESeq2 method within DiffBind, which selected differential regions based on ChIP signals in each replicate and FDR-corrected q-value of 0.05.

HOMER v.4.11.1 mergePeaks program was used to identify overlapping binding loci of two transcription factors. In order to define the sites as "overlapping," peak centers of the two binding sites must be at a distance less than or equal to 500 bp. Note that the peak numbers may not add up exactly since the function automatically resolves redundant overlaps by dropping one fragment during analysis. Motif enrichment analysis was performed using HOMER findMotifsGenome.pl command and scanned +/- 200 bp from the peak center for binding sites of transcription factors, and +/- 750 bp for histone modifications. HOMER annotatePeaks.pl command was used to make annotations of genomic features. The functional analyses of GO term ("Biological Process" sub-ontology) and KEGG pathway were performed using the clusterProfiler package in R or DAVID (https://david.ncifcrf.gov).

4.2.3 ATAC-seq analysis

Paired-end ATAC-seq reads were trimmed using Trimmomatic v.0.36 and then mapped to mm10 mouse genome using Bowtie2 v.2.3.4.1. SAMtools v.1.10 was used to generate BAM files, remove PCR duplicates, and remove mitochondrial DNA. MACS2 v.2.1.2

was used for peak calling with the following parameters: --nomodel --broad --shift -100 --extsize 200 --keep-dup all.

4.2.4 Quantification and statistical analysis

The significance of differences between peak distance, period length, and gene expression was evaluated by unpaired Student's t-test (two-tailed), with significant differences at $p < 0.05$. For motif analysis, HOMER findMotifsGenome.pl calculated P-values using cumulative binomial distribution. For GO term and KEGG pathway analyses, clusterProfiler calculated P-values using hypergeometric distribution which were then adjusted for multiple comparison.

4.3 Results

4.3.1 BMAL1 chromatin binding is attenuated in the *Hnf4a* knockout liver

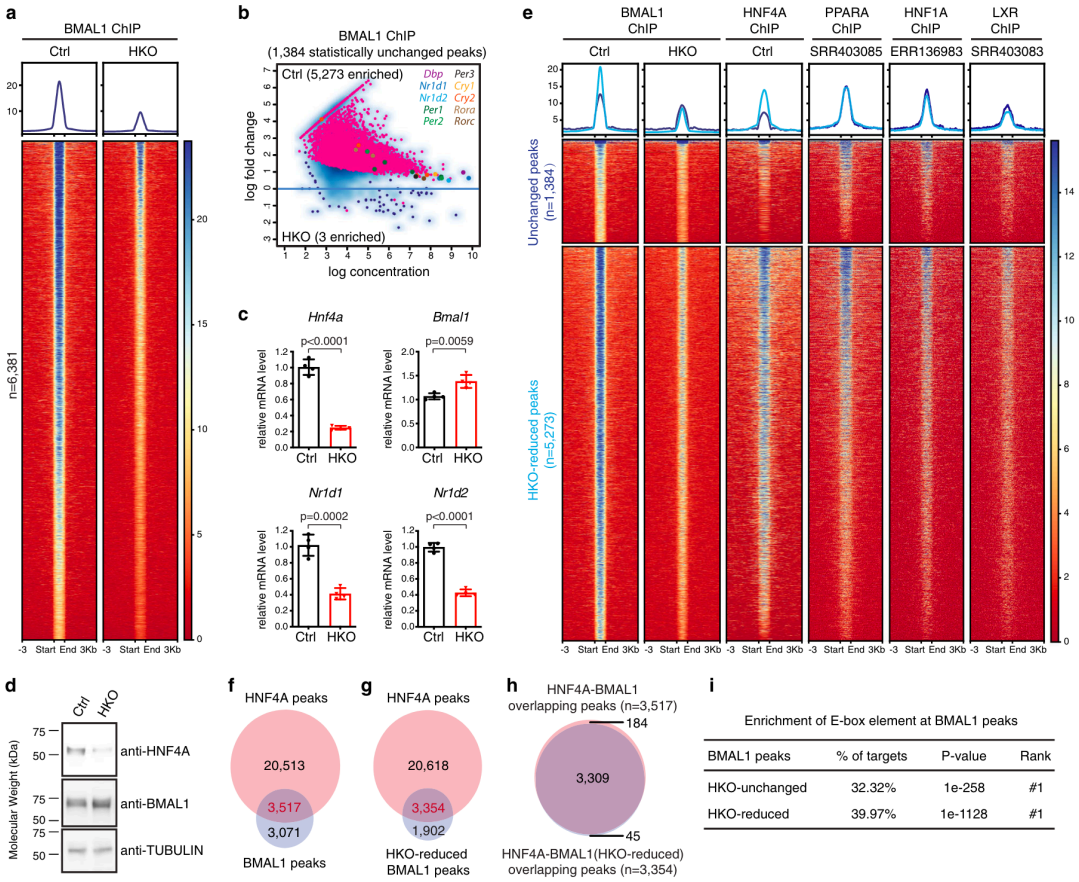
Previously we discovered an extensive genome-wide colocalization of HNF4A and BMAL1::CLOCK in the mouse liver (Qu et al. 2018). While physical interactions and genome-wide co-occupancy between the diurnal regulatory machinery and tissue-specific transcription factors have been reported (Kriebs et al. 2017; Menet, Pescatore, and Rosbash 2014; Trott and Menet 2018), to our knowledge, how the tissue-specific factors may affect BMAL1::CLOCK recruitment has not been studied. To investigate the influence of HNF4A on BMAL1::CLOCK chromatin occupancy and circadian rhythms, we crossed *Hnf4a* floxed mice (Hayhurst et al. 2001a) with Albumin-Cre mice and Per2-luciferase mice in the same C57BL/6J background to generate liver-specific *Hnf4a*

knockout (*Hnf4a*$^{fl/fl}$ *Alb-Cre*$^{+/-}$ *Per2-luc*$^{+/+}$; HKO) and control (*Hnf4a*$^{fl/fl}$ *Alb-Cre*$^{-/-}$ *Per2-luc*$^{+/+}$; Ctrl) mice (see Methods). In the HKO liver, RT-qPCR confirmed a ~75% decrease in *Hnf4a* transcript level accompanied by downregulation of the classic HNF4A target genes *ApoC3*, *Fabp1*, *Ppara*, and *Hnf1a* **(Appendix C-Figure 1a)**. The liver-to-body-weight ratio was significantly increased for the HKO mice **(Appendix C-Figure 1b)**. Histopathological analyses revealed extensive vacuolization in the HKO hepatocytes and marked lipid accumulation throughout the liver tissue **(Appendix C-Figure 1c)**. Remarkably, in contrast with the premature lethality of HKO mice constructed with Albumin-Cre mice in the FVB genetic background (Hayhurst et al. 2001a), the HKO mice we constructed here live to at least the age of 9 months. The *Hnf4a* knockout liver exhibited more severe pathological lesions and greater changes in gene expression in male mice than the female (Hayhurst et al. 2001b; Holloway et al. 2008), although the HCC development rate was sex-independent (Fekry et al. 2019). To eliminate sex as a confounder, we used male mice throughout the study. We mapped genome-wide BMAL1 binding profiles in liver samples collected from three HKO mice and three control mice at ZT6 when BMAL1 binding reaches maximum intensity (Koike et al. 2012). Principal component analyses (PCA) of the three ChIP-seq replicates revealed clustering of samples from the same genotype **(Appendix C-Figure 2a)**. Surprisingly, about 79% (5,273 out of 6,660) of the total BMAL1 peaks were prominently attenuated by *Hnf4a* removal, including ones located within the E-box-containing core clock genes **(Figure 4.1a, b)**. In addition to the clock genes, KEGG and gene ontology (GO) pathway analyses of the HKO-reduced BMAL1 binding genes identified enrichment of the metabolic

pathways, such as glucose and cholesterol metabolism, especially when compared with unchanged binding sites **(Appendix C-Figure 2b, c).** Therefore, circadian regulation of these key tissue-specific nodes (Tahara and Shibata 2016) is supervised by HNF4A. The strong impact HNF4A exerted on BMAL1::CLOCK cistrome seemed to occur post-translationally, because BMAL1 transcript and protein levels were not reduced but rather moderately increased in the HKO liver, potentially related to downregulated *Nr1d1* and *Nr1d2* encoding transcriptional repressors of *Bmal1* **(Figure 4.1c, d)**.

Motif analysis of the HKO-reduced BMAL1 peaks indicated an enrichment of the HNF4A-binding motif, apart from the E-box element **(Appendix C-Figure 2d)**. We parsed all BMAL1 peaks into three groups based on signal variation in response to *Hnf4a* knockout: ones that were reduced (5,273 peaks), enhanced (3 peaks), or not significantly changed (1,384 peaks). On average, BMAL1 peaks of higher intensity tended to be more responsive to *Hnf4a* ablation **(Figure 4.1e)**. We also plotted HNF4A ChIP-seq signals when they reach maximum at ZT16 (Qu et al. 2018) at each position of the BMAL1 peaks, finding HNF4A to display higher accumulation at the HKO-reduced BMAL1 peaks relative to the unchanged peaks **(Figure 4.1e)**. In contrast, for transcription factors PPARA, HNF1A, and LXR that were downregulated upon *Hnf4a* removal **(Appendix C-Figure 1a)**, by analyzing legacy ChIP-seq data (Boergesen et al. 2012; Faure et al. 2012), we did not observe their differential accumulation at the BMAL1 peaks **(Figure 4.1e)**. Consistently, their binding motifs ranked far behind the HNF4A-binding sequence at the HKO-reduced BMAL1 peaks **(Appendix C-Figure 2d)**. The distance from a BMAL1

peak to the nearest HNF4A peak was significantly smaller in general for the HKO-reduced BMAL1 binding sites than the unchanged ones **(Appendix C-Figure 2e)**. Out of the 3,517 BMAL1 peaks that colocalize with HNF4A occupancy, 3,309 (94%) were greatly reduced by *Hnf4a* removal **(Figure 4.1f-h and Appendix C-Figure 2f)**. These data collectively indicate that HNF4A directly regulates global BMAL1 chromatin binding in the mouse liver. The underlying mechanisms do not involve gene expression regulation but are likely achieved on chromatin in a spatially restricted manner.



i

Enrichment of E-box element at BMAL1 peaks

BMAL1 peaks	% of targets	P-value	Rank
HKO-unchanged	32.32%	1e-258	#1
HKO-reduced	39.97%	1e-1128	#1

Figure 4.1 *BMAL1* chromatin binding is attenuated in the Hnf4a knockout liver. a Heatmap of BMAL1 ChIP-seq signals at ZT6 in control (left) or HKO (right) liver centered at all BMAL1 peaks in control liver. Peaks are ordered vertically by signal strength. **b** MA plot showing differential BMAL1 occupancy in control and HKO livers, using a threshold of FDR < 0.05. The x-axis represents the mean number of reads (log scaled) within the peaks across all samples. The y-axis represents the log fold change between the two samples. BMAL1 bindings at the core clock genes are highlighted. **c** Transcript level of genes was determined by RT-qPCR using liver samples isolated from control or HKO mice at ZT6. Displayed are the means ± SD (n = 4) normalized to Rplp0 expression levels. Statistical significance was determined by two-tailed Student's t-test. **d** Protein levels were determined by western blot analysis using liver samples isolated from control or HKO mice at ZT6. Two independent experiments were repeated with similar results. **e** BMAL1 peaks in control and HKO livers were partitioned into three categories with DiffBind (the HKO-enriched group has only 3 peaks and couldn't be plotted), and then the corresponding TF occupancy at each BMAL1 binding site was plotted. Each horizontal line represents a single BMAL1 binding site. Peaks were ordered vertically by strength of BMAL1 ChIP signal in control liver. f Venn diagram showing overlap between all BMAL1 binding sites (at ZT6) and all HNF4A binding sites (at ZT16). Overlapping peaks were identified using the mergePeaks command in HOMER (see Methods). Note that the peak numbers may not add up exactly since the function automatically resolves redundant overlaps by dropping one fragment during analysis. **g** Venn diagram showing overlap between BMAL1-binding sites that were significantly reduced in HKO liver (at ZT6) and all HNF4A binding sites (at ZT16). h Venn diagram showing overlap between the HNF4A-BMAL1 co-occupancy sites identified in (**f**) and (**g**). i A summary of de novo motif analysis showing significance values of E-box enrichment at the HKO-unchanged or HKO-reduced BMAL1 peaks.

4.3.2 *Hnf4a* knockout alters genome-wide epigenetic landscape

The cooperative loading of transcription factors may involve two mechanisms: 1) a simultaneous loading mediated by protein-protein interactions; 2) a sequential loading that requires a pioneer TF to open up local chromatin for other factors to bind (Mayran and Drouin 2018). Notably, we detected physical interactions between BMAL1 and HNF4A in liver cells (Qu et al. 2018). To evaluate the possibility of HNF4A recruiting BMAL1 to the genome, we compared enrichments of the E-box element at HKO-unchanged and HKO-reduced BMAL1 binding sites. The "% of targets" and "p-value of enrichment" reported by HOMER analysis indicated that the E-box sequence was present at a similar frequency within the two categories of BMAL1 binding sites **(Figure 4.1i)**. Moreover, there were a considerable fraction (1,902/5,256=36%) of HKO-reduced BMAL1 binding sites indeed not displaying exactly overlapping HNF4A occupancy **(Figure 4.1g)**. Therefore, it is unlikely for the HNF4A-BMAL1 physical interactions to be generally responsible for the HNF4A-dependent BMAL1 occupancy. We were prompted to ask if HNF4A acts as a pioneer TF and facilitates the accessibility of a broad range of chromatin that is a prerequisite for BMAL1 binding to occur.

The chromatin loading of a pioneer TF initiates increases in accessible/primed enhancers marked by H3K4me1 and subsequent chromatin activation marked by H3K27ac (Mayran and Drouin 2018). Therefore, the intensity of H3K4me1 and H3K27ac defines chromatin landscape and is indicative of pioneer TFs' activity. In agreement with our prediction, we observed a clear reduction in genome-wide H3K4me1 and H3K27ac deposition upon

Hnf4a knockout **(Figure 4.2a, b and Appendix C-Figure 3a, b)**, with the HNF4A-binding motif overrepresented at the HKO-reduced sites for both histone marks **(Figure 4.2c and Appendix C-Figure 3c)**. To interrogate to what extent HNF4A is involved in early steps of chromatin remodeling, we looked into H3K4me1 and found it generally reduced at HNF4A binding sites upon *Hnf4a* knockout **(Appendix C-Figure 3d)**. In addition, HNF4A tended to accumulate more intensively at the H3K4me1 sites that would be significantly reduced by *Hnf4a* knockout (about 41.3% of total peaks), relative to the unchanged H3K4me1 sites **(Figure 4.2d)**. The distance from an H3K4me1 peak to the nearest HNF4A peak was noticeably smaller for the HKO-reduced H3K4me1 sites **(Appendix C-Figure 3e)**. Similarly, the extent of H3K27ac loss in the HKO liver was positively correlated with the intensity of local HNF4A binding **(Appendix C-Figure 3f)**. Motif analysis of all H3K4me1-marked regions in the control liver revealed maximal enrichment of the HNF4A-binding motif **(Figure 4.2e)**, in agreement with a global profiling finding HNF4A occupancy overrepresented in accessible regions of liver chromatin (C. Liu et al. 2019). Taken together, HNF4A potentially serves as a key pioneer factor remodeling the active chromatin landscape in the liver. Of note, we found the local deposition of H3K4me1 and H3K27ac marks was specifically reduced by *Hnf4a* knockout at the HKO-reduced BMAL1 sites **(Figure 4.2f and Appendix C-Figure 3g)**, supporting a working model that HNF4A supervises BMAL1 loading by helping establish a permissive chromatin landscape.

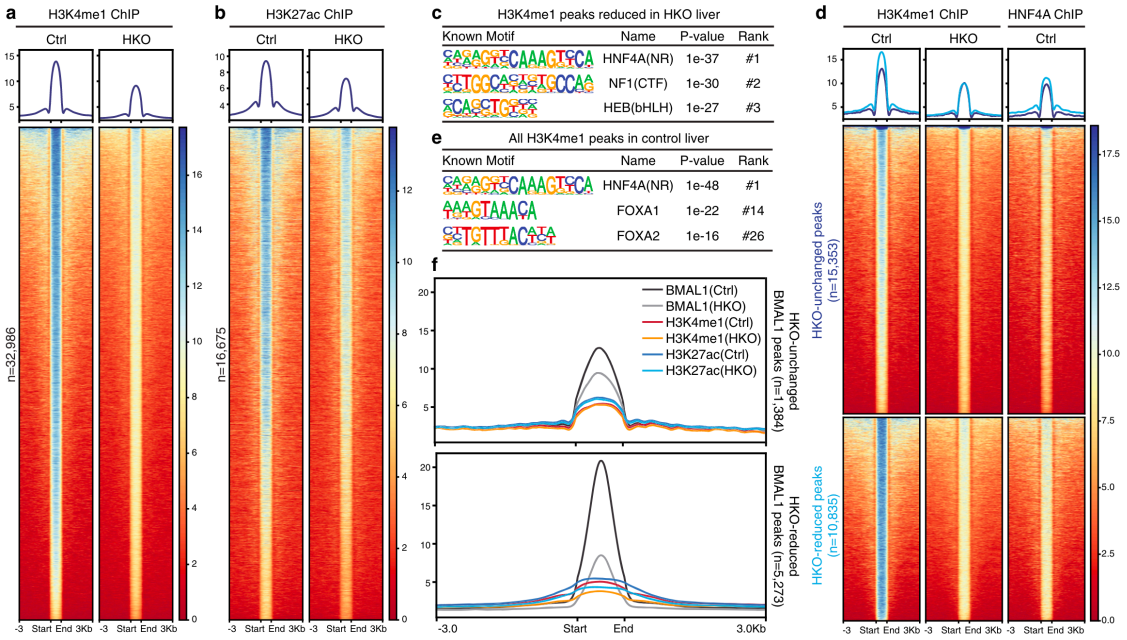


Figure 4.2 *Hnf4a* knockout alters the genome-wide epigenetic landscape. a, b Heatmap of H3K4me1 (a) or H3K27ac (b) ChIP-seq signals at ZT6 in control (left) or HKO (right) liver centered at all peaks in control liver. Peaks are ordered vertically by signal strength. **c** Motif analysis of HKO-deprived H3K4me1 sites. Known consensus motifs are shown with corresponding enrichment significance values. **d** H3K4me1 peaks in control and HKO livers were partitioned into three categories with DiffBind (the HKO-enriched group has only 23 peaks and couldn't be plotted), and then the corresponding HNF4A occupancy (at ZT16) at each H3K4me1 site was plotted. Each horizontal line represents a single H3K4me1 site. Peaks were ordered vertically by strength of H3K4me1 ChIP signal in control liver. **e** Motif analysis of all H3K4me1 marked sites in the control liver. Known consensus motifs are shown with corresponding enrichment significance values. **f** Metaplot showing average intensity of BMAL1, H3K4me1, and H3K27ac ChIP-seq signals (all at ZT6) in control or HKO livers surrounding HKO-unchanged (upper panel) or HKO-reduced (lower panel) BMAL1 peak centers.

4.3.3 Ectopic HNF4A expression reprograms epigenetic landscape and induces tissue-specific BMAL1 bindings

Next, we sought to assess BMAL1 cistromes before and after HNF4A action in a biological system that has never been exposed to HNF4A protein. We ectopically expressed the adult isoform HNF4A2 in human bone osteosarcoma epithelial U2OS cells

where the endogenous *Hnf4a* expression is negligible ("The Human Protein Atlas" n.d.). 1,742 BMAL1 peaks were moderately reduced by HNF4A2 expression **(Figure 4.3a and Appendix C-Figure 4a)**, seemingly resulting from the downregulation of *Bmal1* transcription **(Figure 4.3b)**. In the meanwhile, we identified 311 BMAL1 binding events that were significantly enhanced or gained *de novo* in response to HNF4A2 expression, compared with the GFP expression group **(Figure 4.3a and Appendix C-Figure 4a)**. These HNF4A2-induced BMAL1 peaks were more frequently located at distal or intronic enhancer regions **(Figure 4.3c)** and enriched with the HNF4A-binding motif ranking second only to the E-box element **(Figure 4.3d)**. To interrogate the biological relevance of HNF4A2-induced BMAL1 bindings, we examined whether they occur in cells where HNF4A is naturally expressed. BMAL1 and HNF4A ChIP-seq signals from human liver cancer Hep3B or HepG2 cells were plotted correspondingly at each position of the BMAL1 binding sites we just profiled in U2OS-GFP and U2OS-HNF4A2 cells. Interestingly, BMAL1 ChIP signals in Hep3B cells displayed an analogous pattern to the U2OS-HNF4A2 dataset, *i.e.* signals at the U2OS-HNF4A2-enriched peak sites were stronger than those at the U2OS-GFP-enriched ones **(Figure 4.3e)**, indicating the U2OS-HNF4A2-induced BMAL1 peaks to be specifically expressed in liver cell cultures. Furthermore, endogenously expressed HNF4A in Hep3B or HepG2 cells was found to accumulate more abundantly at the U2OS-HNF4A2-induced BMAL1 peaks than the other sites **(Figure 4.3e)**. Therefore, the BMAL1 binding events we have induced in U2OS cells by introducing genome-wide occupancy of HNF4A2 may represent a true aspect of tissue-specific BMAL1 cistromes.

Chromatin landscape was confirmed to be remodeled by HNF4A2 expression, according to ChIP-seq profiling of H3K4me1 and H3K27ac **(Figure 4.3f and Appendix C-Figure 4b-d)**. In line with that observed at the gained BMAL1 peaks, the HNF4A2-enhanced H3K4me1 and H3K27ac sites were more likely located in distal or intronic enhancer regions **(Appendix C-Figure 4e, f)**, concordant with a general recognition that lineage-specifying transcription factors exert physiologic effects through interactions with tissue-specific enhancers (Mayran and Drouin 2018). The U2OS-HNF4A2-enhanced H3K4me1 sites were confirmed to enrich more HNF4A occupation than the other sites in liver cells **(Figure 4.3g)**, suggesting that HNF4A2 binding is directly responsible for the induced H3K4me1 deposition. The subset of H3K4me1 sites that were mildly reduced by HNF4A2 expression, considering the minimal on-site HNF4A localization in liver cells **(Figure 4.3g)**, likely resulted from indirect effects of HNF4A2 ectopic expression. Lastly, distinct from the other BMAL1 peaks, the HNF4A2-induced BMAL1 peaks were marked by locally enhanced deposition of H3K4me1 and H3K27ac upon HNF4A expression **(Figure 4.3h and Appendix C-Figure 4g)**. To exhibit the HNF4A2-reprogrammed BMAL1, H3K4me1, and H3K27ac peaks in higher resolution, we present genome tracks of representative genes (*SLC25A42*, *DOK4*, *CDHR2*, and *PLPP3*) in **Figure 4.3i** and **Appendix C-Figure 5**. The fetal HNF4A isoforms lacking the N-terminal activation domain AF-1 relative to the adult isoforms are specifically expressed in the embryonic liver and diseased liver. They occupy much the same set of genome loci as the adult isoforms do yet exhibit a lower transcriptional activity (Deans et al. 2021; Lambert et al. 2020). We found that ectopically expressing the fetal isoform HNF4A8

induced tissue-specific BMAL1 binding likewise **(Appendix C-Figure 6)**, arguing that HNF4A-regulated BMAL1 recruitment is invariable during liver development and disease transition. Taken together, we programmed tissue-specific BMAL1 bindings by remodeling E-box-containing enhancers which are otherwise actively masked by nucleosomes. Existing literature has demonstrated that functional BMAL1::CLOCK occupancy at circadian enhancers closely correlates with the oscillation of the target genes (Fang et al. 2014; Vollmers et al. 2012). We speculate that in some cases HNF4A expression alone is not enough for achieving efficient chromatin opening and the presence of additional chromatin remodeling factors is necessary.

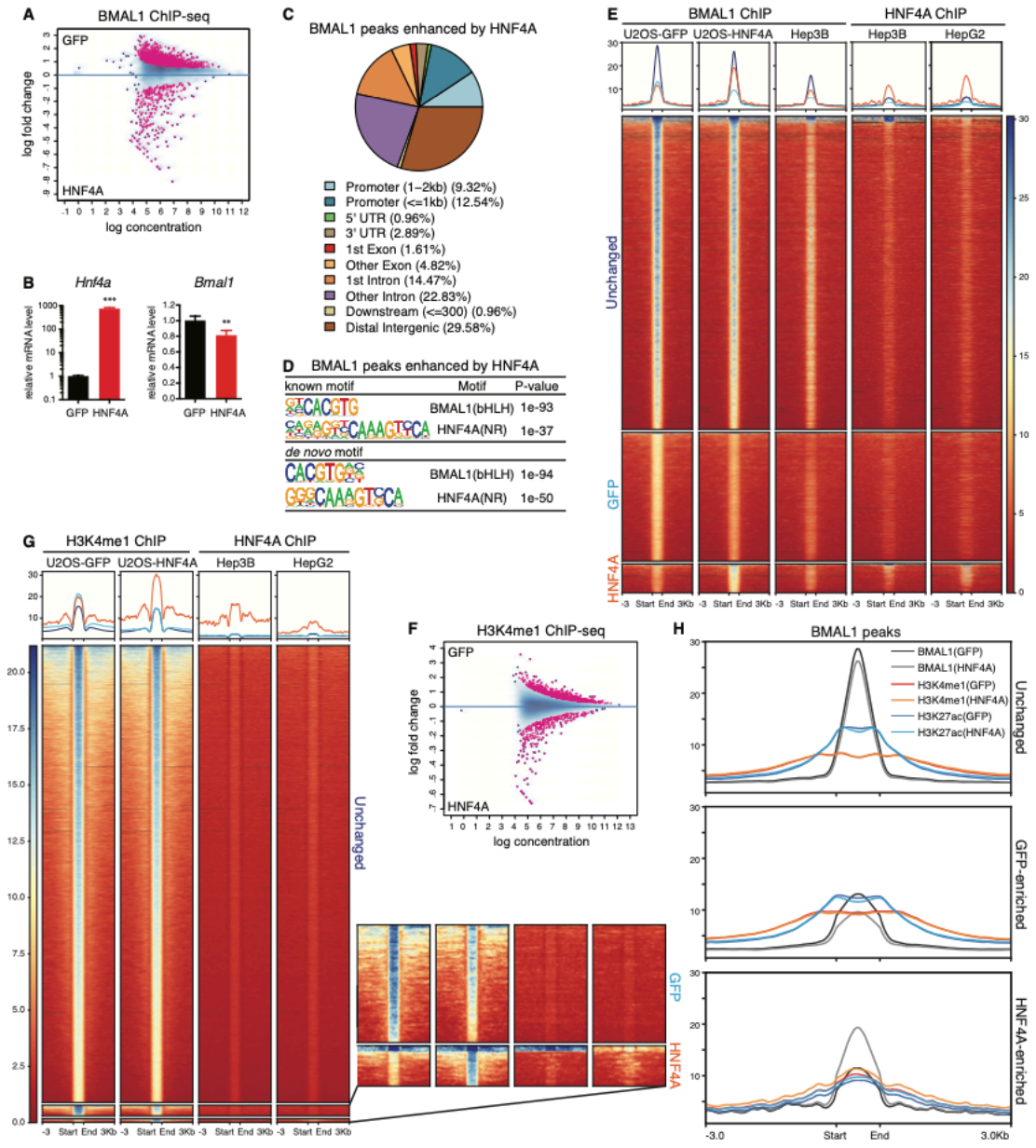

A
BMAL1 ChIP-seq
GFP
HNF4A
log fold change
log concentration
B
Hnf4a
Bmal1
relative mRNA level
GFP HNF4A
C
BMAL1 peaks enhanced by HNF4A
Promoter (1–2kb) (9.32%)
Promoter (<=1kb) (12.54%)
5' UTR (0.96%)
3' UTR (2.89%)
1st Exon (1.61%)
Other Exon (4.82%)
1st Intron (14.47%)
Other Intron (22.83%)
Downstream (<=300) (0.96%)
Distal Intergenic (29.58%)
D
BMAL1 peaks enhanced by HNF4A
known motif
Motif
P-value
BMAL1(bHLH) 1e-93
HNF4A(NR) 1e-37
de novo motif
BMAL1(bHLH) 1e-94
HNF4A(NR) 1e-50
E
BMAL1 ChIP
HNF4A ChIP
U2OS-GFP
U2OS-HNF4A
Hep3B
Hep3B
HepG2
Unchanged
GFP
HNF4A
-3 Start End 3Kb
F
H3K4me1 ChIP-seq
GFP
HNF4A
log fold change
log concentration
G
H3K4me1 ChIP
HNF4A ChIP
U2OS-GFP
U2OS-HNF4A
Hep3B
HepG2
Unchanged
GFP
HNF4A
H
BMAL1 peaks
BMAL1(GFP)
BMAL1(HNF4A)
H3K4me1(GFP)
H3K4me1(HNF4A)
H3K27ac(GFP)
H3K27ac(HNF4A)
Unchanged
GFP-enriched
HNF4A-enriched
-3.0
Start
End
3.0Kb

Figure 4.3 Ectopic HNF4A2 expression reprograms epigenetic landscape and induces tissue-specific BMAL1 bindings. a MA plot showing differential BMAL1 occupancy in U2OS-GFP and U2OS-HNF4A2 cells, using threshold of FDR < 0.05. The x-axis represents the mean number of reads (log scaled) within the peaks across all samples. The y-axis represents the log fold change between the two samples. **b** Transcript level of genes in U2OS-GFP or U2OS-HNF4A2 cells was determined by RT-qPCR. Displayed are the means ± SD (n = 3 cell culture wells) normalized to Rplp0 expression levels. Statistical significance was determined by two-tailed Student's t-test. **c** Distribution of genomic annotations of HNF4A2-enhanced BMAL1 peaks. **d** Motif analysis of HNF4A2-enhanced BMAL1 binding sites. de novo consensus motifs are shown with corresponding enrichment significance values. **e** BMAL1 peaks in U2OS-GFP and U2OS-HNF4A2 cells were partitioned into three categories with DiffBind. Then the corresponding BMAL1 and HNF4A occupancy in Hep3B or HepG2 cells were plotted by centering at each BMAL1 binding site in U2OS cells. Each horizontal line represents a single BMAL1 binding site in U2OS. Peaks were ordered vertically by strength of BMAL1 ChIP signal in U2OS. **f** MA plot showing differential H3K4me1 occupancy in U2OS-GFP and U2OS-HNF4A2 cells, using threshold of FDR < 0.05. The x-axis represents the mean number of reads (log scaled) within the peaks across all samples. The y-axis represents the log fold change between the two samples. **g** H3K4me1 peaks in U2OS-GFP and U2OS-HNF4A2 were partitioned into three categories with DiffBind. Then the corresponding HNF4A occupancy in Hep3B or HepG2 cells was plotted by centering at each H3K4me1 site. Each horizontal line represents a single H3K4me1 site. Peaks were ordered vertically by strength of H3K4me1 ChIP signal in U2OS. Heatmaps of GFP- and HNF4A2-enriched peaks are highlighted in the inset. **h** Metaplot showing average intensity of BMAL1, H3K4me1, or H3K27ac ChIP-seq signals in U2OS-GFP or U2OS-HNF4A2 cells surrounding centers of BMAL1 peaks of indicated groups. **i** IGV genome tracks showing BMAL1, HNF4A, H3K4me1, and H3K27ac enrichment at the SLC25A42 gene in indicated cells, based on normalized ChIP-seq read coverage. Track heights are indicated.

4.3.4 Circadian rhythms are disturbed by *Hnf4a* knockout and HNF4A-MODY mutation

We previously showed that *Hnf4a* knockdown caused varying degrees of circadian rhythm disruption in cell cultures including period shortening and complete arrhythmicity (Qu et al. 2018). Consistently, tissue explants of HKO liver exhibited a shorter period of *Per2-Luc* oscillation *ex vivo* **(Figure 4.4a, b)**. Control and HKO liver tissues were collected every four hours from mice housed under a 12-h light:12-h dark cycle (LD 12:12). RT-qPCR quantification of the core clock transcripts revealed robust circadian oscillations in the control liver, while a dampening was observed after *Hnf4a* ablation **(Figure 4.4c)**. This phenotype was especially clear for *Dbp, Nr1d1,* and *Nr1d2* **(Figure 4.4c)**, genes that are distinct from the other E-box-containing clock genes and lost expression in the *Bmal1* knockout mice (Fang et al. 2014; A. C. Liu et al. 2008). Downregulation of the three BMAL1::CLOCK-dependent genes was confirmed by dimmed local H3K4me1 and H3K27ac signals and associated with dysregulated BMAL1 recruitment **(Figure 4.4d and Appendix C-Figure 7)**. Since *Hnf4a* is minimally expressed outside the liver, kidney, pancreas, and intestinal tracts, we do not expect it to act on the master circadian clock in the SCN or animal behaviors.

Hnf4a mutations were frequently identified in patients with MODY, a rare form of diabetes (Colclough et al. 2013). In agreement, disrupting *Hnf4a* expression in mouse islets or insulinoma cells resulted in impaired glucose-stimulated insulin secretion (Gupta et al. 2005). Interestingly, insulin secretion by pancreatic □ cells is rhythmic, and

perturbation of the circadian cycles contributes to diabetes (Perelis et al. 2015; Marcheva et al. 2010). Our results provide an excellent opportunity for investigating whether HNF4A-MODY mutations connect clock dysregulation to the development of diabetes. R85W is a mutation within the DNA-binding domain of HNF4A that was repeatedly identified in MODY patients (Flanagan et al. 2010; Improda et al. 2016). To investigate this connection, we generated homozygous R85W mutation using CRISPR-Cas9 and surprisedly found the mutant cells to exhibit fundamentally disrupted circadian rhythms **(Figure 4.4e and Appendix C-Figure 8a)**, resembling cells carrying *Hnf4a* homozygous knockout **(Figure 4.4f and Appendix C-Figure 8b)**. Therefore, HNF4A-MODY patients may express dysregulated circadian rhythms which potentially contribute to the disease's pathogenesis and progression.

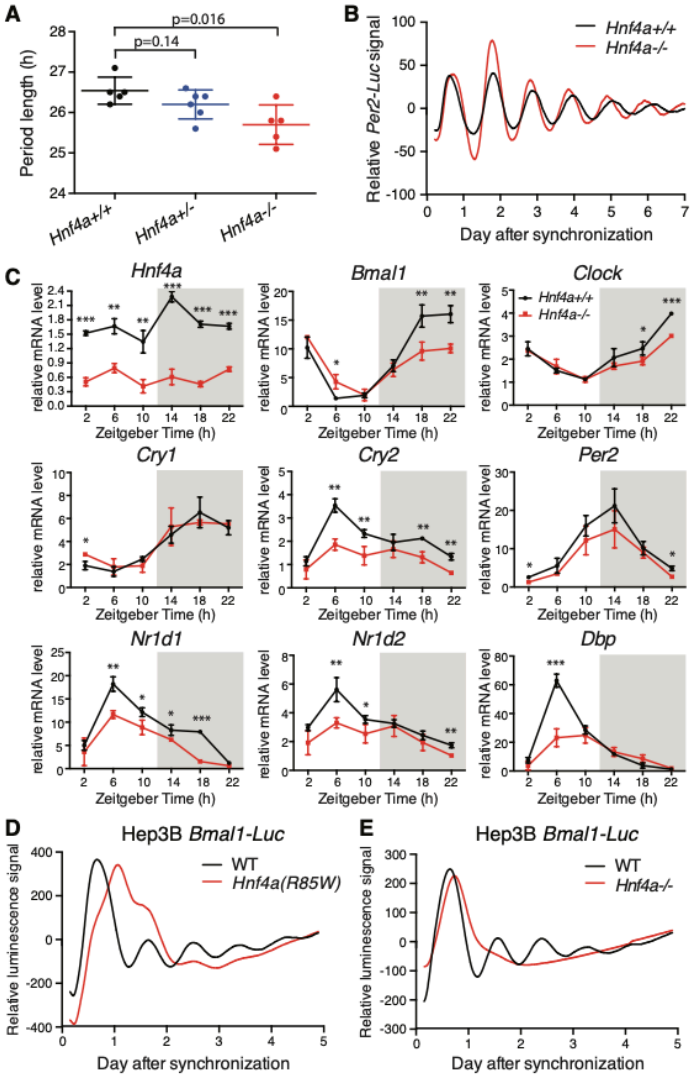


Figure 4.4 Circadian rhythms are disturbed by Hnf4a knockout and HNF4A-MODY mutation. a, b Liver tissue explants were isolated from mice of indicated genotypes and recorded for Per2-Luc bioluminescence. Period lengths of Per2-Luc oscillation were plotted (means ± SD, n = 5 or 6) and statistical significance was determined by two-tailed Student's t-test (a). Representative bioluminescence records show Per2-Luc circadian profiles in control or HKO liver (b). **c** Control and HKO mouse livers were harvested at 4-h intervals over the course of 24 h. Transcript level of genes was analyzed by using RT-qPCR. Displayed are the means ± SD (n = 3 or 4) normalized to non-oscillating Rplp0 expression levels. P-values determined by two-tailed Student's t-test were displayed. **d** IGV genome tracks showing HNF4A (at ZT16), BMAL1 (at ZT6), H3K4me1 (at ZT6), and H3K27ac (at ZT6) enrichment at the Dbp gene in liver tissues, based on normalized ChIP-seq read coverage. Track heights are indicated. **e, f** Representative effect of Hnf4a(R85W) point mutation (e) or Hnf4a knockout (f) on Bmal1-Luc oscillation in human Hep3B cells (n = 3).

4.3.5 HNF4A governs liver-specific circadian transcription

The chromatin remodeling activity and rhythmic recruitment (Qu et al. 2018) of HNF4A prompted us to test whether the hepatic chromatin landscape is dynamically shaped throughout the day. We performed ChIP-seq analyses of H3K4me1 and H3K27ac with wild-type mouse livers collected at ZT16, the peak time of HNF4A binding (Qu et al. 2018), or the antiphase ZT6. Interestingly, the genome-wide deposition of H3K4me1 or H3K27ac was overall higher at ZT16 **(Figure 4.5a, b and Appendix C-Figure 9a, b)**. ATAC-seq that assesses genome-wide chromatin accessibility by probing open chromatin showed an analogous pattern **(Figure 4.5c)**. Indeed, our results indicating that chromatin accessibility in the liver is greater at night are in agreement with observations that the phases of cycling transcripts remarkably clustered between midnight and dawn in the developmentally related liver and kidney where HNF4A is tissue-specifically expressed (R. Zhang et al. 2014; Koike et al. 2012). Therefore, the genome-wide HNF4A occupancy, chromatin opening, and circadian output gene expression are in phase and potentially causally linked. Indeed, we identified the HNF4A-binding motif most enriched at the ZT16-enhanced H3K4me1 or H3K27ac sites **(Appendix C-Figure 9c-f)**. The night-time enhanced HNF4A recruitment potentially induces bursts of genome-wide gene expression by facilitating DNA accessibility by transcriptional machinery.

We identified 6,995 H3K4me1 peaks that were prominently stronger at ZT16 than ZT6 and 44,765 statistically unchanged peaks **(Appendix C-Figure 9c)**. Relative to the unchanged peaks, genes with ZT16-enhanced H3K4me1 peaks were more likely involved

in the circadian rhythm pathway, along with critical aspects of hepatic functions, in particular cholesterol metabolism, gluconeogenesis, insulin resistance, drug metabolism, and autophagy **(Figure 4.5d)**. Indeed, all of these cellular processes were characterized to operate under circadian control (Tahara and Shibata 2016) and feature rhythmic expression of key regulatory genes (Panda et al. 2002). Other than glucose metabolism, HNF4A is well characterized in the regulation of lipid and xenobiotic metabolisms (Hwang-Verslues and Sladek 2010; Yin Liya et al. 2011). HNF4A-MODY patients also exhibit liver disorders such as increased LDL cholesterol levels owing to altered expression of apolipoprotein genes (Ng et al. 2019; Pearson et al. 2005). Therefore, the central mechanisms underlying HNF4A-regulated hepatic metabolisms may involve circadian regulation whereby HNF4A synchronizes the metabolic activities with active food intake after dark.

To further interrogate HNF4A roles in tissue-specific circadian transcription, we assessed circadian rhythms of genes that were significantly downregulated by *Hnf4a* knockout (Walesky et al. 2013) or most differentially expressed at all time points by *Bmal1* knockout (Yang et al. 2016). CircaDB (Pizarro et al. 2013) identified 38% of HNF4A-downregulated and 37% of BMAL1-regulated genes robustly rhythmic, higher than the ratio of 16% for general hepatic transcripts (R. Zhang et al. 2014). The HNF4A-regulated circadian transcripts tend to peak at the pre-dawn "rush hours" **(Appendix C-Figure 10a, b)** and are highly enriched in the pathways of circadian rhythm, lipid and cholesterol metabolism, amino acid metabolism, redox reactions, and liver development **(Appendix**

C-Figure 10c-f), strongly arguing that HNF4A regulates tissue-specific circadian rhythms.

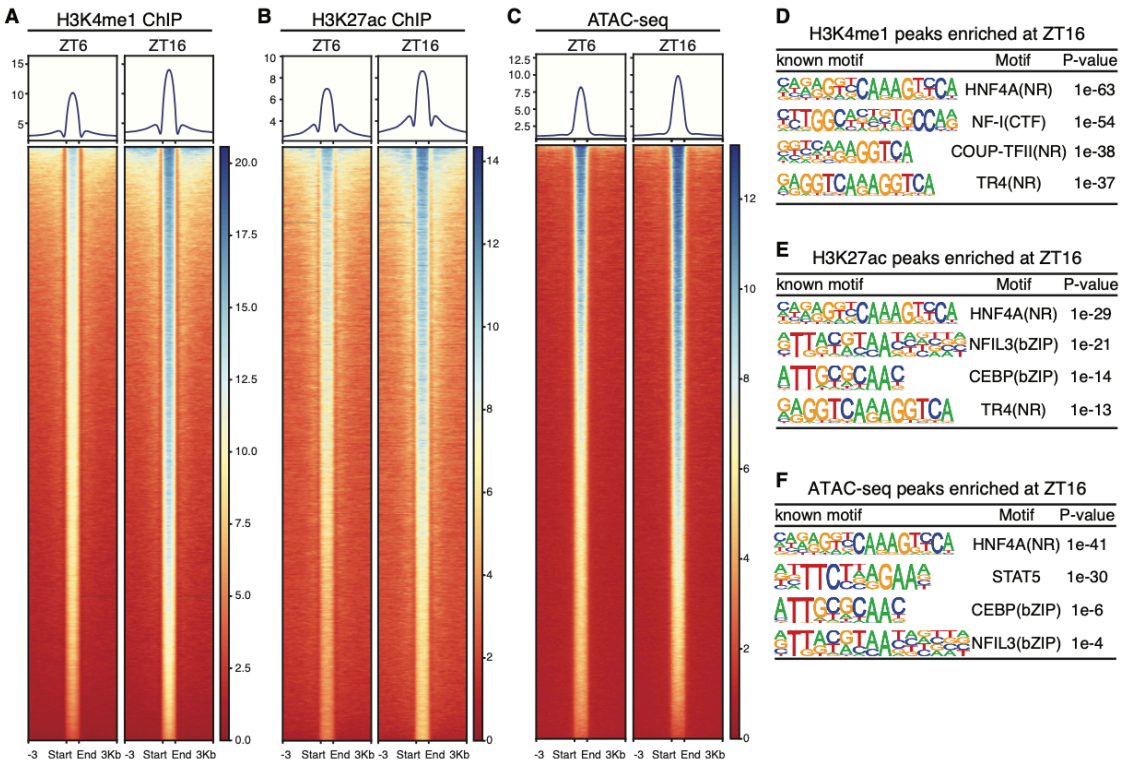


Figure 4.5 Mouse liver chromatin is more accessible at night, synchronized with HNF4A recruitment. a-c Heatmap of H3K4me1 ChIP-seq (a), H3K27ac ChIP-seq (b), or ATAC-seq (c) signals within liver tissues sampled at ZT16 (left) or ZT6 (right) and centered at all peaks of ZT16. Peaks are ordered vertically by signal strength. **d** KEGG pathway analysis was performed for genes having ZT16-ZT6-common or ZT16-enriched H3K4me1 peaks as defined in Appendix C-Figure 9c.

4.3.6 The circadian clock modulates genome-wide DNA binding of HNF4A

To evaluate how HNF4A activity is supervised by the circadian clock, we first induced chronic circadian disruption in mice by performing a jet lag protocol for four weeks. At the end of the treatment, remarkably, the night-time enhanced HNF4A recruitment was

reversed, *i.e.* HNF4A ChIP-seq signals at ZT16 were no longer greater than those at ZT4 (Qu et al. 2018) **(Figure 4.6a and Appendix C-Figure 11a)**. Therefore, the daily cycle of HNF4A chromatin loading is generated by the circadian clock. We then assessed whether BMAL1 in turn regulates HNF4A chromatin binding by using the liver-specific *Bmal1* knockout mouse model (Storch et al. 2007) (see Methods). We mapped genome-wide DNA binding of HNF4A at ZT16 in liver samples collected from liver-specific *Bmal1* knockout (*Bmal1*$^{fl/fl}$*Alb-Cre*$^{+/-}$; BKO) or control mice (*Bmal1*$^{fl/fl}$*Alb-Cre*$^{-/-}$; Ctrl), identifying about 14% (4,576 out of 32,201) of total HNF4A ChIP-seq peaks reduced and about 1% (321 out of 32,201) enhanced in BKO liver **(Figure 4.6b, c and Appendix C-Figure 11b)**. KEGG and GO term analyses of the BKO-reduced HNF4A binding sites revealed genes involved in cancer pathogenesis among most enriched. Other overrepresented categories included Wnt/β-catenin signaling and cell cycle pathways **(Appendix C-Figure 11c, d)**. HNF4A inhibits Wnt/β-catenin signaling and cell cycle progression, potentially underlying its tumor-suppressive roles (Lv, Zhou, and Tang 2021). In comparison, genome-wide binding of the well-characterized hepatic pioneer factor FOXA2 (Mayran and Drouin 2018) was barely affected by *Bmal1* knockout **(Appendix C-Figure 11e, f)**. BMAL1 co-occupancy was only slightly more enriched at the BKO-reduced HNF4A binding sites (19.2% colocalized with BMAL1 binding) than the control sites (14.0% for total HNF4A peaks; 10.3% for BKO-unchanged peaks) **(Appendix C-Figure 11g, h)**, therefore, it is unlikely for chromatin recruitment mediated by protein-protein interactions to play a dominant role in the regulation. Instead, we found *Hnf4a* transcription steadily downregulated by 20-30% upon *Bmal1* removal at all

sampling times **(Figure 4.6d)**. Since BMAL1 directly binds to the *Hnf4a* gene body **(Appendix C-Figure 11i)**, BMAL1::CLOCK likely modulates HNF4A chromatin binding through transcriptional regulation. Analogous to *Per2* transcripts, although dampened, *Hnf4a* oscillation was not abolished by *Bmal1* removal **(Figure 4.6d)**. Since the night-enhanced *Hnf4a* expression was not altered by fasting (Qu et al. 2018), mechanisms rather than feeding behavior may be involved in clock-independent *Hnf4a* oscillation.

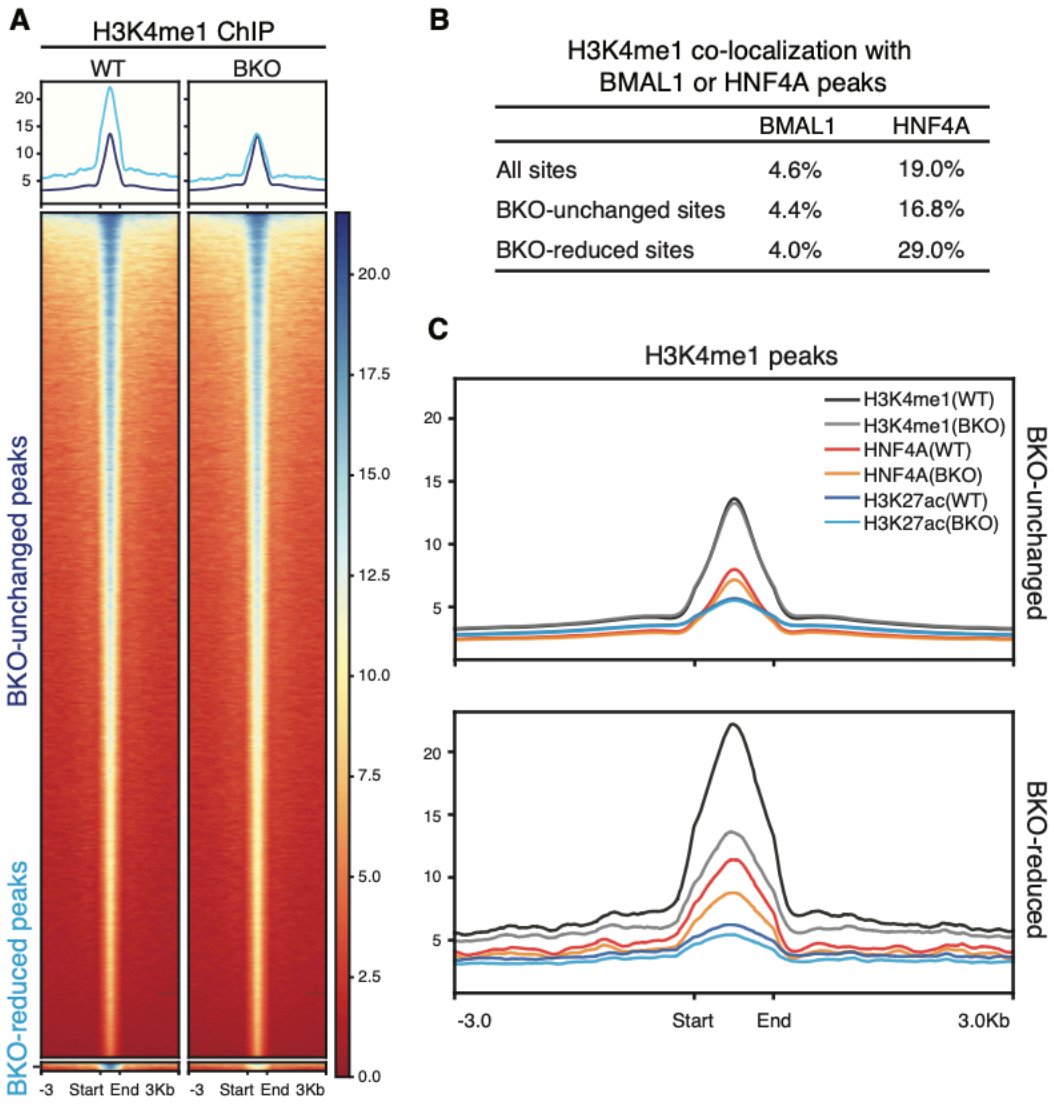


Figure 4.6 The circadian clock modulates genome-wide DNA binding of HNF4A. a Heatmap of HNF4A ChIP-seq signals within liver tissues sampled at ZT4 (left) or ZT16 (right) after jet lag treatment and centered at all HNF4A peaks of ZT4. Peaks are ordered vertically by signal strength. **b** Heatmap of HNF4A ChIP-seq signals at ZT16 in control (left) or BKO (right) liver centered at all HNF4A peaks in control liver. Peaks are ordered vertically by signal strength. **c** MA plot showing differential HNF4A peaks in control and BKO livers, using threshold of FDR < 0.05. The x-axis represents the mean number of reads (log scaled) within the peaks across all samples. The y-axis represents the log fold change between the two samples. **d** Control and BKO mouse livers were harvested at 4-h intervals over the course of 24 h. Transcript level of genes was analyzed by using RT-qPCR. Displayed are the means ± SD (n = 3) normalized to non-oscillating Rplp0 expression levels. P-values determined by two-tailed Student's t-test were displayed.

4.3.7 *Bmal1* knockout alters epigenetic landscape seemingly due to attenuated HNF4A activity

To assess whether chromatin remodeling is responsible for BMAL1-regulated HNF4A genome binding, we profiled genome-wide locations of H3K4me1 and H3K27ac at ZT16 in control or BKO liver tissues. Overall, the histone marks were not greatly changed by BKO **(Figure 4.7a, b and Appendix C-Figure 12a, b)** especially when compared with HKO **(Figure 4.2a, b)**. Statistical analysis identified small subgroups that were significantly reduced or enhanced by *Bmal1* knockout **(Figure 4.7c, d)**. For instance, genes exhibiting significantly reduced histone modifications included *Nr1d2*; genes exhibiting significantly enhanced histone modifications included *Npas2* **(Figure 4.7e)**.

Motif analysis of the BKO-enhanced H3K4me1 (n=264) or H3K27ac (n=134) peaks identified the ROR response element (RORE), binding motif of transcriptional repressors REV-ERBs **(Appendix C-Figure 12c, d)**. Interestingly, at the BKO-reduced H3K4me1 (n=987) or H3K27ac (n=101) peaks, we did not identify the E-box element but instead found an enrichment of nuclear receptor binding sites, with the HNF4A-binding motif top-ranked **(Appendix C-Figure 12e, f)**. About 4.6% of total H3K4me1 peaks display colocalization with BMAL1 binding within a distance of 500 bp. This degree of BMAL1 colocalization remained similar for the BKO-unchanged (4.5%) and BKO-reduced (4.0%) subgroups of H3K4me1 sites **(Figure 4.7f and Appendix C-Figure 12g)**, indicating BMAL1 occupancy not enriched at the BKO-reduced H3K4me1 sites. In contrast, H3K4me1 peaks having HNF4A colocalization increased from a background of

19.0% to 27.5% at the BKO-reduced sites and decreased to 16.6% for the BKO-unchanged sites **(Figure 4.7f and Appendix C-Figure 12h)**. We consider HNF4A co-occupancy enriched at the BKO-reduced H3K4me1 sites, given that only 42% of the HKO-reduced H3K4me1 peaks exhibited HNF4A colocalization within the same distance of 500 bp **(Appendix C-Figure 12i)**. We noticed that HNF4A occupancy was selectively reduced at the BKO-reduced H3K4me1 sites compared with the unchanged sites **(Appendix C-Figure 12j)**. Importantly, we plotted H3K4me1 and H3K27ac ChIP-seq signals at each position of the BMAL1 or HNF4A peaks to find both histone marks specifically attenuated by BKO at HNF4A peaks **(Figure 4.7g)** rather than at BMAL1 peaks **(Figure 4.7h)**. Therefore, BMAL1::CLOCK occupancy does not directly regulate active epigenetic modifications at ZT16 but through positively modulating HNF4A. Taken together, it is unlikely for BMAL1 to regulate HNF4A cistrome through chromatin remodeling.

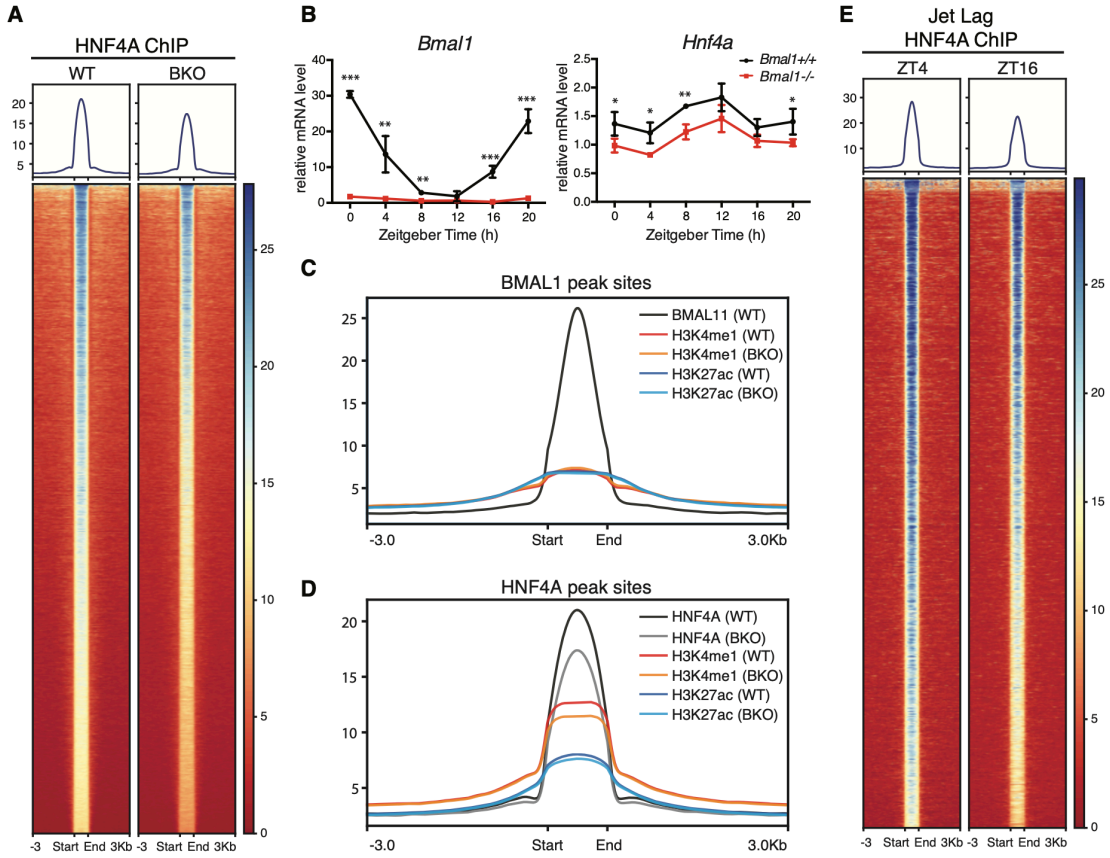


Figure 4.7 *Bmal1* knockout alters epigenetic landscape in the liver, seemingly due to attenuated HNF4A activity. a, b Heatmap of H3K4me1 (a) or H3K27ac (b) ChIP-seq signals at ZT16 in control (left) or BKO (right) liver and centered at all peaks in control liver. Peaks are ordered vertically by signal strength. **c, d** MA plot showing differential H3K4me1 (c) or H3K27ac (d) peaks in control and BKO livers, using threshold of FDR < 0.05. The x-axis represents the mean number of reads (log scaled) within the peaks across all samples. The y-axis represents the log fold change between the two samples. **e** IGV genome tracks showing BMAL1 (at ZT6), HNF4A (at ZT16), H3K4me1 (at ZT16), and H3K27ac (at ZT16) enrichment at Nr1d2 and Npas2 genes in control or BKO liver, based on normalized ChIP-seq read coverage. Track heights are indicated. f Percentages of three groups of H3K4me1 sites colocalizing with BMAL1 or HNF4A peaks. Peak numbers for percentage calculation are in **Appendix C-Figure 12g, h. g** Metaplot showing average intensity of HNF4A, H3K4me1, and H3K27ac ChIP-seq signals (all at ZT16) in control or BKO livers surrounding HNF4A peak centers in control liver. h Metaplot showing average intensity of BMAL1 (at ZT6), H3K4me1 (at ZT16), and H3K27ac (at ZT16) ChIP-seq signals in control or BKO livers surrounding BMAL1 peak centers in control liver.

4.4 Discussion

Our findings demonstrate that HNF4A may act as a pioneer TF creating tissue-specific repertoires of accessible *cis*-regulatory elements. Consistently, the HNF4-binding element was top-scoring in accessible chromatin regions in the intestinal duodenal epithelium (L. Chen, Toke, Luo, Vasoya, Fullem, et al. 2019; L. Chen, Toke, Luo, Vasoya, Aita, et al. 2019). HNF4A was essential for maintaining active histone signature H3K27ac in the intestine and liver (L. Chen, Toke, Luo, Vasoya, Fullem, et al. 2019; Thakur et al. 2019). While HNF4A was an established fundamental liver development regulator, it was not as well characterized in the process of chromatin remodeling as another hepatic TF FOXA/HNF3 (Mayran and Drouin 2018; Nagy, Bisgaard, and Thorgeirsson 1994; Li, Ning, and Duncan 2000). Nevertheless, we found the HNF4A-binding motif more enriched than the FOXA motifs in H3K4me1-positive liver genome regions **(Figure 4.2e)**. HNF4A was essential and to some extent sufficient for establishing liver-specific chromatin landscape **(Figure 4.2, 4.3)**. Notably, among all hepatic TFs, HNF4A was the most important in converting human fibroblasts to hepatocyte-like cells (hiHeps) (Sekiya and Suzuki 2011; Nakamori et al. 2017). These observations collectively suggest that HNF4A remodels the chromatin landscape for active gene expression changes during development. To gain mechanistic insights, direct nucleosome binding studies will be needed in future to clarify whether HNF4A can independently displace histones like FOXA/HNF3 does or engages ATP-dependent enzymes to expand the ‘‘openness’’ of local chromatin (Mayran and Drouin 2018).

The REV-ERB regulation of metabolic genes was reported to require chromatin recruitment by hepatic transcription factors (Y. Zhang et al. 2015; 2016). The activity of pancreatic cycling gene expression displayed a correlation with the binding of the pancreas-specific transcription factor PDX1 (Perelis et al. 2015). Despite these insights into a role of the lineage-specifying TFs, there has been a gap in understanding the molecular basis of tissue-specific rhythmicity whose misalignment is closely associated with organ-specific disorders (Bass and Lazar 2016). By using loss-of-function and gain-of-function genetic models, we demonstrate that the lineage-specifying HNF4A is critical and in some cases sufficient for establishing liver-specific BMAL1 cistrome, seemingly independent of direct recruitment but by means of providing permissive chromatin structure. Our results may provide a molecular basis for tissue-specific BMAL1::CLOCK cistromes depending on the chromatin structures likely arising from early events in tissue development. Systematic profiling of 20 diverse human cell types identified ~25% of genes displaying cell-type-specific expression that is explained by alterations in chromatin structures (Marstrand and Storey 2014). Our discoveries suggest that tissue-specific chromatin landscape profoundly shapes the circadian rhythms, providing a unifying mechanism for circadian rhythm heterogeneity across tissue types. We recently reported that the genome-wide BMAL1::CLOCK occupancy in glioblastoma stem cells was more expanded as compared with normal neural stem cells (Dong et al. 2019). Given that chromatin structure alterations are prevalent in tumor tissues (Corces et al. 2018), our findings may provide additional insights into reprogrammed circadian clocks now found in cancer and many other disease states.

Liver-specific *Hnf4a* removal undermined BMAL1 occupancy at most of its target genes including the E-box-containing core clock genes **(Figure 4.1b)**, downregulated transcription of the BMAL1::CLOCK-dependent core clock genes *Dbp, Nr1d1*, and *Nr1d2* **(Figure 4.4c),** and shortened the period of *Per2-Luc* oscillation **(Figure 4.4a, b)**. Potentially resulting from altered *Nr1d1* and *Nr1d2* expression, *Bmal1* transcription was upregulated in HKO liver at ZT6 **(Figure 4.1c)** and downregulated by HNF4A ectopic expression **(Figure 4.3b and Appendix C-Figure 6a)**. The adult HNF4A was reported to repress *Bmal1* expression less than the fetal form (Fekry et al. 2018), potentially underlying the mild *Bmal1* upregulation in HKO liver where the adult HNF4A is specifically targeted due to spatiotemporal expression of Albumin-Cre. It seems that the HNF4A actions on BMAL1::CLOCK activity are multilayered, including promoting the chromatin binding, transrepressing the transcriptional activity (Qu et al. 2018), and negatively regulating *Bmal1* transcription. These seemingly contradictory modes of action are indeed prevalently employed by circadian clock regulators so as to maintain circadian homeostasis, by virtue of the nature of the interlocking negative feedback loops (Mohawk, Green, and Takahashi 2012). For instance, CRY stabilization lead to suppression of BMAL1::CLOCK transcriptional activity and a simultaneous increase in *Bmal1* transcription which was largely attributable to downregulated *Rev-erb* genes (Hirota et al. 2012). Given that *Dbp, Nr1d1,* and *Nr1d2* were downregulated in the HKO liver **(Figure 4.4c)**, the chromatin remodeling activity of HNF4A seems to play a dominant role here by positively impacting BMAL1::CLOCK activities. The robust BMAL1::CLOCK transrepression activity we have characterized (Qu et al. 2018) can

serve as a second mechanism for HNF4A to fine-tune circadian rhythms only after BMAL1::CLOCK is efficiently recruited to the target genes. In aggregate, our results indicate that HNF4A is a key modulator of the core circadian clock machinery.

Largely in agreement with prior studies (Koike et al. 2012; Vollmers et al. 2012) **(Appendix C-Figure 9g-h)**, genome-wide H3K4me1 and H3K27ac deposition as well as chromatin accessibility assessed by ATAC-seq were increased during the night **(Figure 4.5a-c)**. The rhythmic recruitment of HNF4A may stimulate a synchronized day-night transition of chromatin accessibility, which intriguingly coincides with the predawn "rush hours" of circadian gene transcription in the liver (Koike et al. 2012; R. Zhang et al. 2014). Zhang *et al.* (R. Zhang et al. 2014) profiled circadian transcriptomes of 12 different mouse organs and found the phase distribution of circadian transcripts in the liver and kidney to exhibit patterns distinct from the other ten organs, *i.e.* being clustered between midnight and dawn. Given that out of the 12 organs investigated, only the liver and kidney indeed express the HNF4A protein, HNF4A is likely responsible for the unique repertoire and phase distribution of circadian outputs in the two organs. Even though the pioneer activity of HNF4A is higher at night, BMAL1, H3K4me1, and H3K27ac ChIP-seq signals were all considerably reduced at noon (ZT6) by *Hnf4a* knockout **(Figure 4.1, 4.2)**. Therefore, HNF4A seems to determine the hepatic chromatin landscape from morning till night. Collectively, the chromatin remodeling activities of HNF4A may control tissue-specific circadian rhythms through two mechanisms: 1) to facilitate BMAL1::CLOCK recruitment during the day and secure the operation of the

core clock; 2) to open chromatin maximally during the night and promote predawn-clustered expression of tissue-specific circadian outputs.

Finally, the rhythmic HNF4A genome binding was disrupted by chronic jet lag **(Figure 4.6a)**. BMAL1 promoted efficient genome binding of HNF4A, likely independent of protein-protein interactions or chromatin remodeling but through activating *Hnf4a* transcription **(Figure 4.6, 4.7)**. *Bmal1* knockout only slightly altered H3K4me1 and H3K27ac at ZT16 **(Figure 4.7a-d)**. The RORE element was enriched at BKO-enhanced modification sites; the BKO-reduced modification sites did not enrich E-box element or BMAL1 colocalization but were associated with HNF4A binding **(Figure 4.7f-h)**. Therefore, BMAL1::CLOCK modulates hepatic epigenetic landscape potentially by activating target genes, namely *Rev-erbs* and *Hnf4a*. These results incidentally support our main finding that HNF4A shapes hepatic chromatin landscape. The circadian clock regulates HNF4A transcription and rhythmic DNA binding whereby it contributes to hepatic epigenetic landscape.

Appendix A

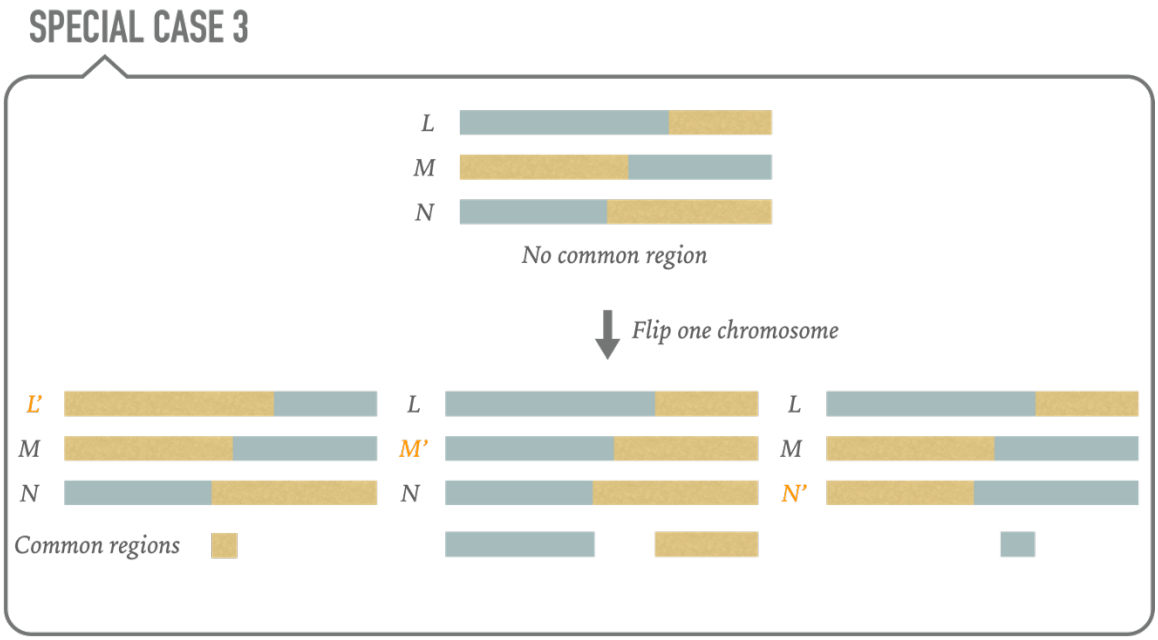


Figure S1. The strategy of flipping gametes when no common region is found.

Appendix B

Table S1. The population consists of 210 recombinant inbred lines (RIL) from the cross between two elite rice cultivars.

RIL	yd	tp	gn	kgw	gl	gw	hd	OsC1	Bin871
R001	26.55	11.575	98.6	24.2325	9.74	3	80.1	0	1
R002	18.65	12.975	61.8	23.395	9.595	2.78	74.33333	0	1
R003	21.3625	12.25	74.1	24.02	9.61	2.905	78.56666667	0	1
R004	26.94	9.9	126.65	22.2575	8.36	2.94	86.66666667	1	0
R005	26.0325	11.4	88.55	26.1275	8.88	3.25	93.66666667	1	0
R006	26.0225	11.425	106.1	21.62	8.815	2.94	85.9	0	1
R007	25.885	11.15	91.675	25.44	8.925	3.045	82.56666667	0	1
R008	31.955	12.025	106.75	25.1875	9.555	2.91	86	0	1
R009	31.1625	9.825	123.175	25.6825	9.375	2.94	91.33333	1	0

R010	28.0925	11.45	87.85	28.745	9.42	3.45	79.76666667	0	1
R012	31.2275	11.525	89.3	29.915	9.595	3.43	94.23333333	0	1
R013	24.81	9.875	102.7	24.1075	8.23	3.415	97.23333333	1	0
R014	21.7725	10.2	79.125	26.5675	9.185	3.07	84.9	1	0
R015	28.8675	8.25	149.325	23.3925	8.475	3.155	89.76666667	1	0
R016	33.7575	12.95	99.325	25.525	9.26	2.98	94.56666667	1	0
R017	30.7225	12.175	98.5	25.35	9.27	3.105	91.56666667	1	0
R018	27.69	10.8	97.25	25.6275	9.17	3.37	85	0	1
R019	27.46	10.35	98.575	25.9475	8.805	3.43	83	1	0
R020	29.4025	11.1	102.225	26.0525	8.975	3.445	83.76666667	1	1
R021	34.32	10.625	125.6	26.0975	8.165	3.655	99.33333333	0	1

R022	13.835	7.775	74.75	23.785	8.48	3.14	66.333 33333	1	0
R023	23.83	12.45	94.225	20.38	8.205	2.86	70.666 66667	1	0
R025	26.38	11.25	122.5	19.022 5	8.125	2.77	90.666 66667	1	0
R027	27.08	8.55	134.92 5	23.512 5	8.285	3.445	81.433 33333	0	0
R028	21.627 5	11.1	80.875	24.665	8.75	3.065	73.766 66667	1	0
R029	28.132 5	11.9	93.25	25.405	8.6	3.24	80.9	0	0
R030	27.66	10.95	91.45	27.772 5	9.77	3.145	81.233 33333	0	1
R031	24.22	12.375	77.125	26.245	9.405	3.305	76.233 33333	0	1
R032	20.63	11.1	82.325	22.57	9.115	2.76	79.766 66667	0	1
R033	31.262 5	10.875	114.67 5	25.065	8.85	3.065	81.766 66667	0	1
R034	30.085	10.375	117.15	24.852 5	8.85	3.095	80.566 66667	0	1

R036	29.3475	12.8	109.025	21.44	8.075	2.985	87.23333333	1	0
R039	20.9275	11.675	69.05	26.32	10.33	2.82	72.56666667	0	1
R040	33.4025	11.8	128.95	22.235	8.725	2.96	80.33333333	0	1
R041	33.85	11.95	126.475	22.665	8.57	2.815	86	1	0
R042	28.3975	10.7	101	26.295	9.525	3.235	92.1	1	0
R043	25.58	11.075	83	27.7525	9.885	3.305	80.9	0	1
R044	20.7275	10.975	75.775	25.345	9.61	2.97	91.43333333	0	1
R045	25.49	10.225	92.9	26.3175	8.67	3.395	80.9	0	1
R046	30.44	10.2	119.875	24.79	8.625	3.325	96.33333333	0	1
R048	29.9325	12.1	115.775	21.395	8.755	3.015	81.1	0	1
R049	19.015	11.45	73.75	20.3375	7.97	2.9	81.1	1	0

R050	23.885	8.65	111.15	24.2975	8.875	3.025	84.1	1	0
R051	29.41	11.025	97.3	27.7925	9.765	3.26	79.9	0	1
R053	14.47	9.75	68.15	22.3525	8.91	2.9	66.1	1	0
R054	24.77	10.825	96.125	24.22	8.1	3.275	72.43333333	1	0
R055	21.1475	12.375	87.275	20.005	8.155	2.825	77.33333333	0	1
R056	27.3575	12.625	85.85	25.28	9.74	2.8	86.25426731	1	0
R057	27.93	11.1	119.175	21.395	8.38	3.14	81	0	1
R059	25.5275	12.275	82.8	25.56	9.455	3.255	82.9	1	0
R060	16.4325	10.35	73.875	21.0975	8.685	2.785	70.12093398	1	0
R061	26.37	10.575	98.7	25.2925	8.595	3.3	82.9	1	0
R062	27.1725	9.875	110.4	24.9325	8.415	3.335	80.56666667	1	0

R063	17.27	9.65	71.625	25.35	8.875	3.385	68.566 66667	1	0
R064	27.652 5	10.975	95.95	26.512 5	8.665	3.35	78.333 33333	0	1
R065	26.742 5	9.425	102.65	27.965	8.775	3.355	78	1	0
R066	21.997 5	11.575	96.4	20.032 5	8.265	2.965	76.9	0	1
R067	21.695	10.425	101.9	20.397 5	9.035	2.84	82.9	0	1
R068	27.9	12.225	86.275	26.532 5	8.64	3.47	84	1	1
R069	28.047 5	11.875	91.375	26.662 5	8.18	3.12	86.766 66667	1	0
R070	25.532 5	11.1	108.02 5	21.367 5	8.55	2.99	84.333 33333	0	1
R071	31.562 5	11.025	124.7	22.87	8.255	3.195	84.666 66667	0	1
R072	26.277 5	12.8	89.1	23.832 5	8.35	3.14	87.1	0	1
R073	16.867 5	10.475	66.25	25.685	8.97	3.215	78.333 33333	0	1

R074	20.965	10.825	74.175	26.9525	9.025	3.25	79.9	0	1
R075	23.7325	11.575	76.425	26.5875	9.24	2.98	91.02093398	0	1
R076	30.5075	9.2	142.525	22.975	8.185	3.02	87.9	0	1
R077	25.165	11.375	80.8	27.9225	10.22	3.065	78.43333333	0	1
R078	26.5175	10.475	93.25	26.5675	9.9	3.14	96.9	0	1
R079	31.1875	11.15	118.475	23.4825	8.5	3.055	89	0	1
R080	32.7	11.2	130.925	21.8825	7.805	3.47	99.23333333	0	1
R081	31.1475	10.5	122.875	24.0175	8.525	3.15	81.1	0	1
R082	30.455	11.05	111.7	24.56	8.465	3.115	79.76666667	0	1
R083	32.985	10	112.65	28.5775	8.655	3.405	97.23333333	1	0
R084	23.98	11.8	114.425	18.2025	7.795	2.795	91.76666667	1	0

R086	26.985	11.925	90.625	25.0125	9.655	2.885	88	1	0
R087	26.3375	11.15	89.175	25.84	9.425	2.92	90.76666667	1	0
R089	11.87	11.275	58.8	18.025	8.51	2.59	62.9	0	1
R090	26.145	9.9	117.475	22.695	8.105	3.1	82.1	0	1
R092	34.3225	10.25	139.2	24.8675	8.45	3.235	80.9	1	0
R093	26.005	9.475	102.775	27.61	9.31	3.275	81.76666667	0	1
R094	29.5425	9.4	138.35	23.005	7.65	3.525	95.1	1	0
R097	19.5425	13.125	61.125	24.6225	8.545	3.105	75.9	0	1
R098	24.7	11.9	86.55	23.855	9.045	2.995	85.76666667	0	1
R099	26.8625	11.625	111	21.02	8.41	2.93	80.1	0	1
R100	30.8275	11	135.175	20.9825	8.27	2.99	81.76666667	0	1
R101	27.797	9.325	131.05	22.905	8.125	3.34	95	0	1

	5								
R103	24.2575	10.8	102.625	21.8575	8.955	2.78	95	0	1
R105	36.99	11.825	109.55	28.0925	9.63	3.2	99.66666667	1	0
R106	25.3275	10.475	88.3	28.1925	9.295	3.255	73.9	0	1
R107	34.425	11.075	133.15	23.3675	8.23	3.4	100.3333333	1	0
R108	21.7975	10.675	76.25	27.4925	9.13	3.235	91.1	1	0
R109	26.0075	13.4	90.175	22.38	8.625	2.81	83.9	1	0
R111	18.4125	11.8	77.25	21.9	8.655	2.66	66.33333333	1	0
R112	20.755	11.15	84.85	23.505	9.29	2.905	72.43333333	0	1
R113	26.0625	10.55	105	23.76	8.58	3.085	76.33333333	0	1
R115	26.445	9.325	115.925	25.0275	8.855	3.025	81.43333333	1	0
R116	24.49	11.6	103.9	20.625	8.08	2.805	78.766	1	0

							66667		
R118	9.96	11.075	50.175	18.945	8.945	2.635	58	0	1
R119	28.775	11.25	94.9	27.367 5	8.535	3.435	71.9	0	1
R120	26.557 5	12.325	91.55	23.78	8.725	2.99	82.566 66667	0	1
R121	27.565	11.05	123.55	20.95	8.375	2.81	84.333 33333	1	0
R122	27.365	10.95	93.4	27.205	8.61	3.325	82.1	1	0
R123	24.945	10.4	110.2	22.107 5	8.595	2.82	85.766 66667	0	1
R125	20.962 5	11.625	79.225	22.492 5	8.05	3.12	84.566 66667	0	1
R126	23.675	11.125	97.8	21.895	8.045	2.98	78.566 66667	0	1
R127	21.177 5	10.15	85.125	24.89	8.995	3.24	82.9	0	1
R128	28.268 87805	11.545 97561	110.94 69512	22.247 37805	8.765	2.825	94.233 33333	0	1
R129	17.772 5	12.4	61.5	23.587 5	9.53	2.66	66	1	0
R130	24.988	8.9209	111.19	25.029	8.255	3.305	99.233	1	0

	87805	7561	69512	87805			33333		
R131	24.452	9.45	112.32	23.845	8.495	3.295	79.766	0	1
	5		5				66667		
R132	29.24	11.4	87.075	29.057	9.005	3.38	86.9	0	1
				5					
R133	23.442	11.825	94.25	21.607	8.78	3.05	85.433	1	0
	5			5			33333		
R135	28.082	9.325	138.7	22.297	8.48	2.875	87.433	1	0
	5			5			33333		
R137	27.767	11.925	84.075	28.56	9.515	3.115	88.333	1	0
	5						33333		
R138	28.437	12.525	104.65	22.042	8.92	2.825	90.566	1	0
	5			5			66667		
R139	22.6	10.675	83.05	26.167	8.63	3.36	79.233	0	1
				5			33333		
R140	24.472	13.5	78.225	23.645	9.575	2.85	77.333	0	1
	5						33333		
R141	22.812	13	78.7	22.845	8.15	3.04	69.433	1	1
	5						33333		
R142	28.317	10.575	104.3	25.257	9.975	2.955	87.1	1	0
	5			5					
R143	30.28	10.55	122.82	23.637	9.01	2.99	83.9	0	1

			5	5					
R144	21.8025	10.225	102.4	21.89	8.805	2.69	71.76666667	0	1
R145	18.0625	13.725	63.725	20.6175	8.635	2.735	66.1	1	0
R146	24.5775	7.1	134.65	26.1875	8.515	3.45	102	1	0
R147	30.175	10.925	92.9	29.635	9.775	3.325	91.33333333	1	0
R148	29.4925	10.575	97.175	28.6625	9.445	3.29	93	1	0
R149	21.0625	10.425	83.1	24.9275	8.655	3.31	75.9	0	1
R150	20.98	9.85	93.525	22.6375	9.45	2.86	79.9	0	1
R151	23.36	10.575	107.975	20.6725	8.355	3.055	88.43333333	0	1
R152	26.4925	10.075	110.075	24.1025	8.755	3.49	82.76666667	0	1
R153	26.315	11.55	97.15	24.2325	8.435	3.385	79.9	1	0
R154	20.127	10.75	88.3	21.902	7.77	3.26	75.9	0	1

	5			5					
R155	29.0575	14.55	99.8	20.045	8.04	2.76	80	0	1
R156	27.5525	12.575	86.275	25.645	9.2	3.06	79.43333333	0	1
R157	26.0725	12.6	89.175	23.8975	8.6	2.855	77.43333333	0	1
R158	23.1475	10.35	89.35	24.7925	9.62	3.17	94.43333333	0	1
R159	24.635	9.6	105.475	24.505	8.605	3.345	82.23333333	0	1
R160	29.695	10.025	99.05	29.505	9.5	3.275	91.33333333	1	0
R161	24.7425	10.35	89.925	26.455	9.595	3.17	83.1	0	1
R163	28.0125	11.975	83.65	28.0775	9.97	3.07	78.43333333	1	0
R164	29.4475	12.2	90	26.9525	9.255	3.215	86.66666667	1	0
R166	25.7	9.075	115.6	24.6675	9.075	3.33	93.33333333	0	1
R167	17.385	10.475	68.625	24.787	8.06	3.31	68.433	1	0

				5			33333		
R168	30.4075	11.125	112.125	24.5125	8.58	3.07	77.566 66667	0	1
R169	33.0675	10.8	118.475	25.4875	8.69	3.125	83.1	1	0
R170	26.6325	10.4	91.025	28.4625	9.49	3.405	78.333 33333	0	1
R171	25.7775	9.3	118.875	23.545	9.425	2.885	78.333 33333	1	0
R172	25.1325	11.325	111.175	20.1275	8.035	3.165	86.9	0	1
R173	25.8525	10.6	110.375	22.4575	7.97	3.125	78.766 66667	1	0
R174	24.2825	11	101.225	22.2425	7.895	3.15	73.333 33333	0	1
R175	22.805	9.075	88.95	28.5875	8.935	3.41	78.566 66667	0	1
R176	22.66	11.725	77.15	24.9925	9.32	3.02	80.333 33333	0	1
R177	24.55	10.075	84.85	29.0675	9.42	3.315	80.666 66667	0	1
R178	27.882	12.625	105.2	21.207	7.845	2.98	86.566	1	0

	5			5			66667		
R179	27.485	11	93.5	27.007	9.62	3.185	88.433	0	1
				5			33333		
R180	31.265	9.2	119.42	28.015	8.88	3.345	82	1	0
			5						
R181	24.747	11.2	87.775	25.485	9.24	3.155	81	1	0
	5								
R182	22.635	12.05	75.925	25.47	9.165	3.16	78.233	1	0
							33333		
R183	25.492	10.775	81.35	29.312	9.805	3.09	87.233	1	0
	5			5			33333		
R184	26.443	9.7209	104.99	26.172	8.89	3.24	84.233	1	0
	87805	7561	69512	37805			33333		
R185	25.135	9.75	110.57	23.042	8.565	2.88	93.433	1	0
			5	5			33333		
R186	30.122	10.45	115.65	25.28	9.52	2.855	88.766	1	0
	5						66667		
R187	29.07	9.55	111.72	27.24	9.105	3.18	85.1	0	1
			5						
R188	27.162	9.8	103.42	26.96	9.47	3.18	89.333	0	1
	5		5				33333		
R189	26.88	12.425	81.925	27.075	8.88	3.075	76.433	0	0

							33333		
R190	25.925	12.875	80.75	25.825	9.925	2.84	78.1	0	1
R191	29.552 5	10.65	138.27 5	20.407 5	8.475	2.78	88.666 66667	1	0
R193	29.225	14.375	88.15	23.835	8.365	2.96	79.433 33333	0	1
R194	31.162 5	11.1	118.5	23.205	8.395	3.24	93.666 66667	0	1
R195	29.852 5	10.275	130.32 5	22.707 5	8.705	2.995	85.766 66667	1	0
R196	25.352 5	8.2	122.52 5	25.777 5	8.33	3.4	92.333 33333	0	0
R197	23.775	11.25	90.725	23.425	8.095	3.33	90.566 66667	1	0
R198	27.44	11.35	106.92 5	22.567 5	8.79	2.9	87.666 66667	0	1
R199	24.032 5	10.975	84.175	26.77	8.54	3.34	75.666 66667	0	1
R200	18.862 5	9.525	83.1	23.692 5	8.265	3.295	81.766 66667	1	0
R202	22.245	9.725	109.97 5	21.317 5	8.47	2.96	82.9	0	1

R203	24.8525	11.1	84.8	26.575	9.08	3.33	81.56666667	0	1
R204	17.975	12.55	67.875	21.1625	8.755	2.68	67.23333333	1	0
R205	27.175	12	80.65	28.3525	9.52	2.97	93.9	1	0
R206	27.5825	10.275	111.975	24.6325	8.02	3.165	82.33333333	1	0
R208	29.83	9	113.95	29.43	10.08	3.33	84.1	1	1
R209	30.525	10.4	139.2	21.4725	7.945	3.02	87.1	1	0
R210	29.545	10.7	104.7	26.6125	9.32	3.265	97.56666667	0	1
R211	24.22887805	10.64597561	97.34695123	24.09237805	8.2	3.215	73.23333333	1	0
R213	18.9425	8.9	83.2	25.02	8.42	3.245	80.56666667	0	1
R214	26.6975	10.475	118.75	21.66	9.055	2.93	94.76666667	1	0
R216	24.2775	9.4	97.825	26.5175	8.99	3.145	88.76666667	1	0
R218	29.765	10.85	127.6	22.092	8.315	2.99	80.1	0	1

				5					
R219	16.675	10.6	66.125	24.107 5	8.14	3.29	79.766 66667	0	1
R220	32.547 5	11.575	126.6	22.29	8.785	2.94	86.9	0	1
R221	20.545	10.525	77.9	25.37	9.56	2.975	74.1	0	1
R222	25.292 5	8.65	100.67 5	29.09	9.325	3.415	89.433 33333	1	0
R223	28.195	10.1	103.8	26.532 5	8.565	3.55	77.9	0	1
R224	23.77	10.675	108.72 5	21.142 5	8.34	3.005	76.9	0	1
R225	31.052 5	10.325	111.62 5	27.355	8.745	3.22	75.566 66667	0	1
R226	20.997 5	11.625	74.55	24.962 5	8.7	2.95	72.766 66667	0	1
R227	26.715	10.675	111.82 5	23.177 5	8.32	3.385	79.233 33333	0	1
R228	25.545	9.775	121.25	22.127 5	8.145	3.305	77.233 33333	0	1
R229	29.032 5	9.95	120.1	24.485	8.29	3.34	86.433 33333	0	1

R230	28.21	12.85	101.575	21.6075	8.455	2.95	86.43333	0	0
R231	15.40637805	8.12097561	64.74695123	21.9637805	8.822559809	3.032464115	57.56666667	0	1
R232	26.71	10.6	112.725	22.4125	8.705	3.09	80.76666667	0	1
R233	29.305	11.525	105.95	24.0475	9.25	2.885	88.56666667	1	0
R234	25.59	10.85	104.7	23.5825	8.71	2.915	85.43333	1	0
R235	29.7575	10.825	122.2	22.6475	8.21	3.16	82	1	0
R236	22.5125	11.225	91.125	22.325	8.5	2.805	74.56666667	0	1
R237	19.7425	12.975	71.5	21.4125	9.175	2.66	79.1	1	0
R238	28.1975	8.525	121.525	27.1525	8.84	3.18	86.9	1	0
R239	30.7625	10.275	112.975	26.6425	8.905	3.285	83.23333	0	1
R241	27.57558613	11.6076555	94.84342107	26.2685407	8.72	3.455	86.9	0	1

Table S2. The bootstrap variance-covariance matrix of the EM estimated parameters along with the NR and SAS variance-covariance matrices.

Bootstrap	Intercept	YD	TP	GN	KGW	GL	GW	HD		Variance	StdErr
Intercept	130.4037	1.670197	-4.44165	-0.43559	-1.06067	-3.00958	-9.24343	-0.00866		130.4037	11.41944
YD	1.670197	0.031164	-0.07144	-0.00761	-0.0288	-0.00967	-0.03671	-0.00046		0.031164	0.176533
TP	-4.44165	-0.07144	0.190418	0.018249	0.06652	0.030598	0.156139	0.000251		0.190418	0.436369
GN	-0.43559	-0.00761	0.018249	0.002013	0.007343	0.003705	0.009123	-8.40E-05		0.002013	0.044869
KGW	-1.06067	-0.0288	0.06652	0.007343	0.046435	-0.04716	-0.1238	0.000204		0.046435	0.215488
GL	-3.00958	-0.00967	0.030598	0.003705	-0.04716	0.264735	0.518094	-0.00281		0.264735	0.514524
GW	-9.24343	-0.03671	0.156139	0.009123	-0.1238	0.518094	2.109994	-0.00614		2.109994	1.452582
HD	-0.00866	-0.00046	0.000251	-8.40E-05	0.000204	-0.00281	-0.00614	0.000784		0.000784	0.027994
Hessian	Intercept	YD	TP	GN	KGW	GL	GW	HD		Variance	StdErr
Intercept	104.1677	1.380834	-3.52248	-0.35571	-0.8744	-2.56441	-7.07747	-0.00248		104.1677	10.20626
YD	1.380834	0.027144	-0.05917	-0.00652	-0.02549	-0.0072	-0.0217	-0.00045		0.027144	0.164754
TP	-3.52248	-0.05917	0.152383	0.014944	0.055211	0.02515	0.106685	6.90E-05		0.152383	0.390362
GN	-0.35571	-0.00652	0.014944	0.001691	0.00626	0.002941	0.006197	-5.57E-05		0.001691	0.041116
KGW	-0.8744	-0.02549	0.055211	0.00626	0.041346	-0.04314	-0.11573	0.000474		0.041346	0.203337
GL	-2.56441	-0.0072	0.02515	0.002941	-0.04314	0.237687	0.440239	-0.00268		0.237687	0.487532
GW	-7.07747	-0.0217	0.106685	0.006197	-0.11573	0.440239	1.729701	-0.00674		1.729701	1.315181
HD	-0.00248	-0.00045	6.90E-05	-5.57E-05	0.000474	-0.00268	-0.00674	0.000623		0.000623	0.024968
SAS	Intercept	YD	TP	GN	KGW	GL	GW	HD		Variance	StdErr
Intercept	105.8479	1.3951	-3.56353	-0.3594	-0.86896	-2.64673	-7.32216	-0.00112		105.8479	10.28824
YD	1.3951	0.027242	-0.05945	-0.00654	-0.02536	-0.00812	-0.02426	-0.00043		0.027242	0.165052
TP	-3.56353	-0.05945	0.153237	0.015019	0.05487	0.02766	0.113951	0.000016		0.153237	0.391455
GN	-0.3594	-0.00654	0.015019	0.001697	0.006229	0.003175	0.006843	-0.00006		0.001697	0.041195
KGW	-0.86896	-0.02536	0.05487	0.006229	0.041095	-0.04273	-0.11496	0.000451		0.041095	0.202719
GL	-2.64673	-0.00812	0.02766	0.003175	-0.04273	0.240186	0.447966	-0.00268		0.240186	0.490088
GW	-7.32216	-0.02426	0.113951	0.006843	-0.11496	0.447966	1.757715	-0.00681		1.757715	1.325788
HD	-0.00112	-0.00043	0.000016	-0.00006	0.000451	-0.00268	-0.00681	0.000623		0.000623	0.02496

Parameter	Bootstrap	NR	SAS
Intercept	11.4194	10.2063	10.2882
YD	0.1765	0.1648	0.1651
TP	0.4364	0.3904	0.3915
GN	0.0449	0.0411	0.0412
KGW	0.2155	0.2033	0.2027
GL	0.5145	0.4875	0.4901
GW	1.4526	1.3152	1.3258
HD	0.028	0.025	0.025

Appendix C

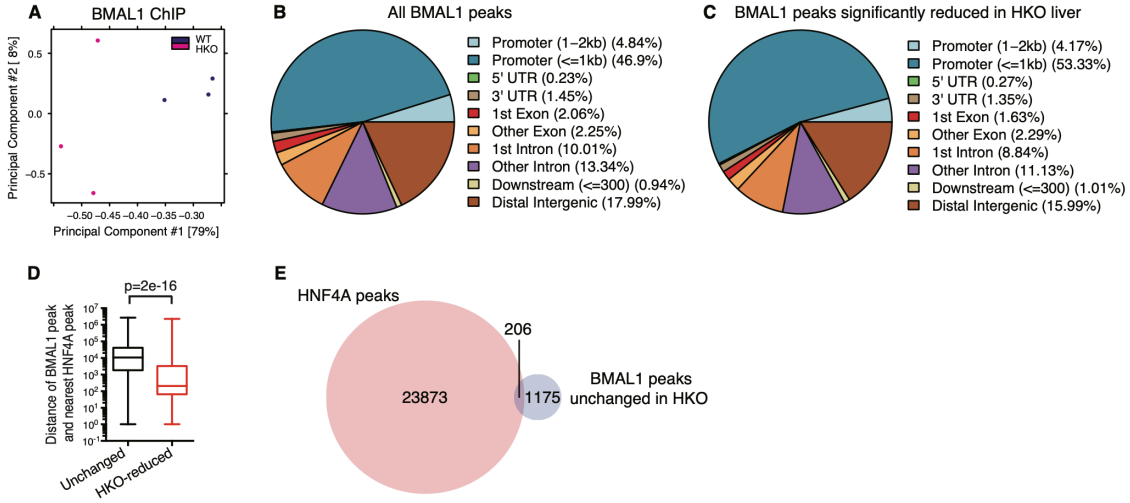


Figure S1. BMAL1 chromatin binding was substantially attenuated in *Hnf4a* knockout liver. (A) PCA plot of BMAL1 ChIP-seq counts across consensus BMAL1 peaks in WT and HKO (at ZT6). **(B)** Distribution of genomic annotations of all the BMAL1 peaks in WT. **(C)** Distribution of genomic annotations of BMAL1 peaks significanlty reduced by *Hnf4a* knockout. **(D)** Base pair unit distance from each BMAL1 peak to the closest HNF4A peak was calculated and box-plotted. Statistical significance was determined by Student's t-test. **(E)** Venn diagram showing overlap between BMAL1-binding sites that were not significantly changed in HKO (at ZT6) and all HNF4A binding sites (at ZT16).

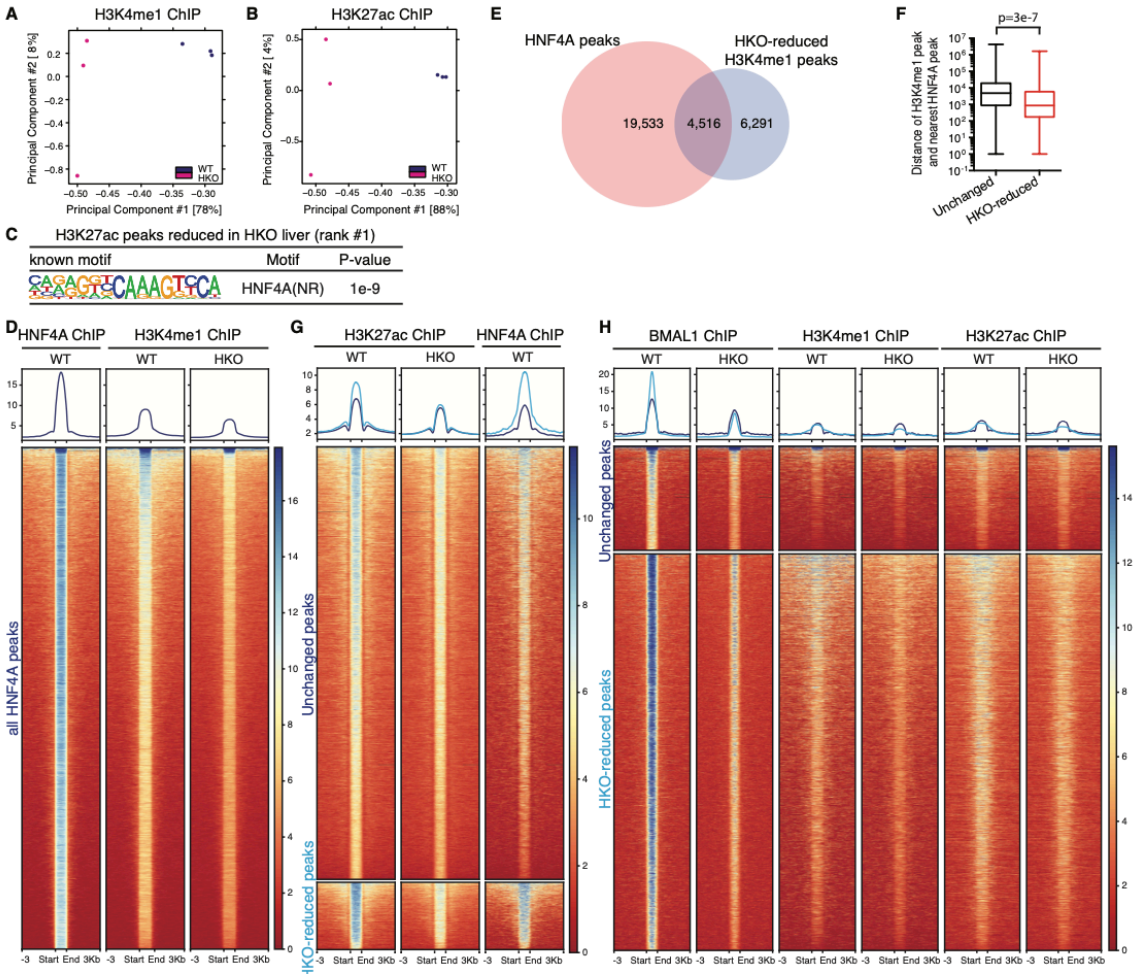


Figure S2. *Hnf4a* knockout alters genome-wide epigenetic landscape. (A-B) PCA plot of H3K4me1 **(A)** or H3K27ac **(B)** ChIP-seq counts at ZT6 across consensus peaks in control and HKO liver. **(C)** Motif analysis of HKO-deprived H3K27ac sites. Known consensus motifs are shown with corresponding enrichment significance values. **(D)** H3K4me1 occupancy in control or HKO liver was plotted at each HNF4A binding site (at ZT16). Each horizontal line represents a single HNF4A binding site. **(E)** Base pair unit distance from each H3K4me1 peak to the closest HNF4A peak was calculated and box-plotted. Statistical significance was determined by Student's t-test. **(F)** H3K27ac peaks in control and HKO livers were partitioned into three categories with DiffBind (the HKO-enriched group has only 3 peaks and couldn't be plotted), and then the corresponding HNF4A occupancy (at ZT16) at each H3K27ac site was plotted. Each horizontal line represents a single H3K27ac site. **(G)** BMAL1 peaks in control and HKO livers were partitioned into three categories with DiffBind (the HKO-enriched group has only 3 peaks and couldn't be plotted), and then the corresponding H3K4me1 or H3K27ac occupancy (at ZT6) at each BMAL1 binding site was plotted. Each horizontal line represents a single BMAL1 binding site.

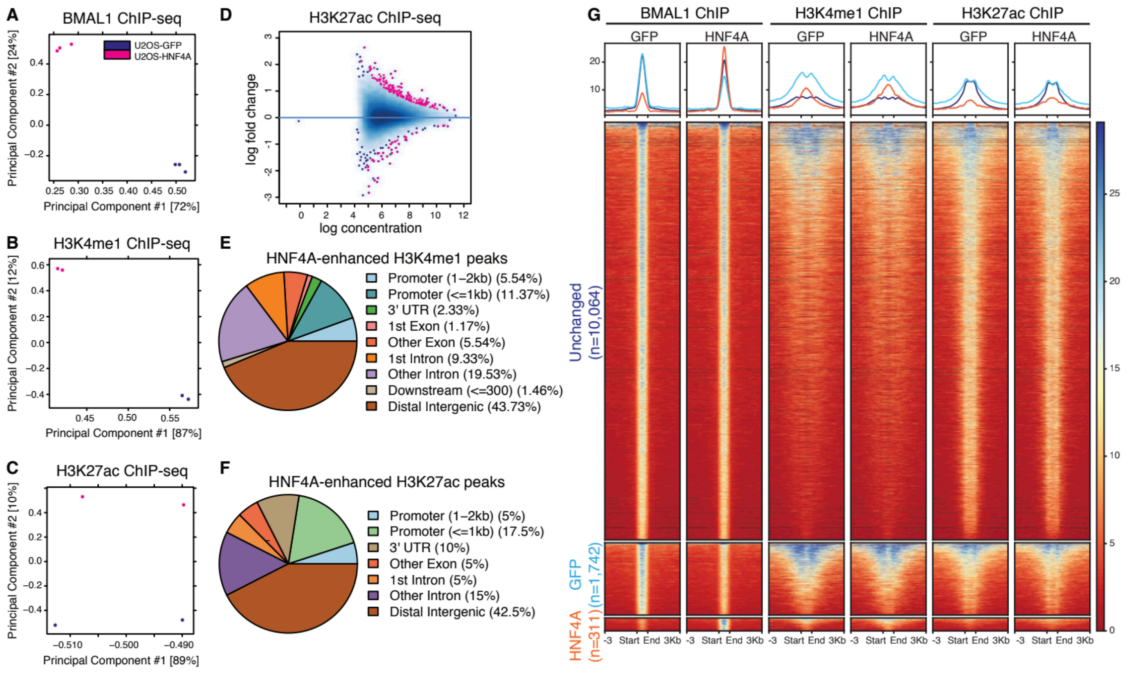


Figure S3. Ectopic HNF4A expression created tissue-specific BMAL1 binding events by stimulation of chromatin accessibility. (A) PCA plot of BMAL1 ChIP-seq counts across consensus peaks in U2OS-GFP and U2OS-HNF4A. **(B)** PCA plot of H3K4me1 ChIP-seq counts across consensus peaks in U2OS-GFP and U2OS-HNF4A. **(C)** PCA plot of H3K27ac ChIP-seq counts across consensus peaks in U2OS-GFP and U2OS-HNF4A. **(D)** MA plot showing differential H3K27ac occupancy in U2OS-GFP or U2OS-HNF4A cells, using threshold of FDR < 0.05. The x-axis represents the mean number of reads (log scaled) within the peaks across all samples. The y-axis represents the log fold change between the two samples. **(E)** Motif analysis of HNF4A-enhanced H3K4me1 sites. Known consensus motif was shown with corresponding enrichment significance values. **(F)** Distribution of genomic annotations of HNF4A-induced H3K4me1 sites. **(G)** Distribution of genomic annotations of HNF4A-induced H3K27ac sites. **(H)** BMAL1 peaks in U2OS-GFP and U2OS-HNF4A cells were partitioned into three categories with DiffBind, and then the corresponding H3K4me1 and H3K27ac occupancy at each BMAL1 binding site was plotted. Each horizontal line represents a single BMAL1 binding site.

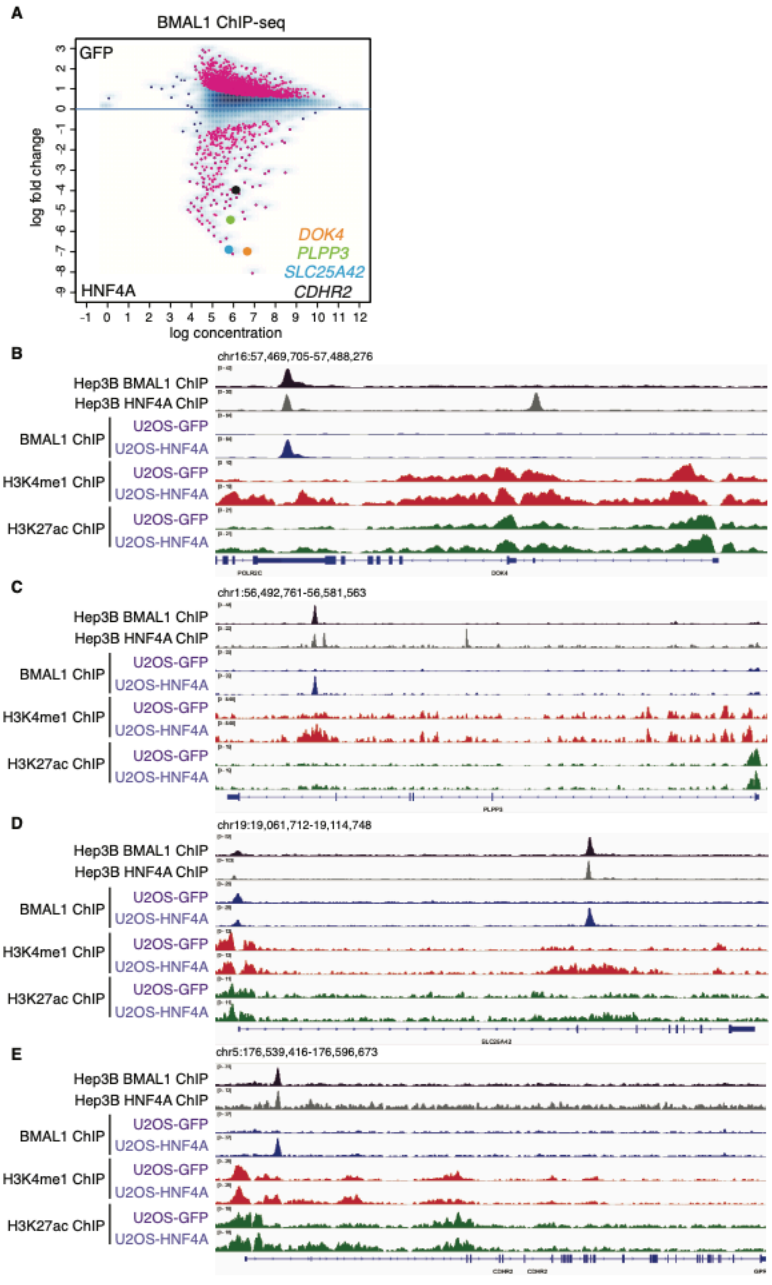


Figure S4. Representative genome tracks showing HNF4A-induced BMAL1 peaks and locally enhanced H3K4me1 and H3K27ac marks. (A) Four representative BMAL1 binding events significantly induced by HNF4A were highlighted in Fig. 3A MA plot. **(B-E)** IGV genome tracks of BMAL1, HNF4A, H3K4me1, and H3K27ac enrichment at *DOK4* **(B)**, *PLPP3* **(C)**, *SLC25A42* **(D)**, *CDHR2* **(E)** gene loci in indicated cells based on normalized ChIP-seq read coverage. Track heights are indicated.

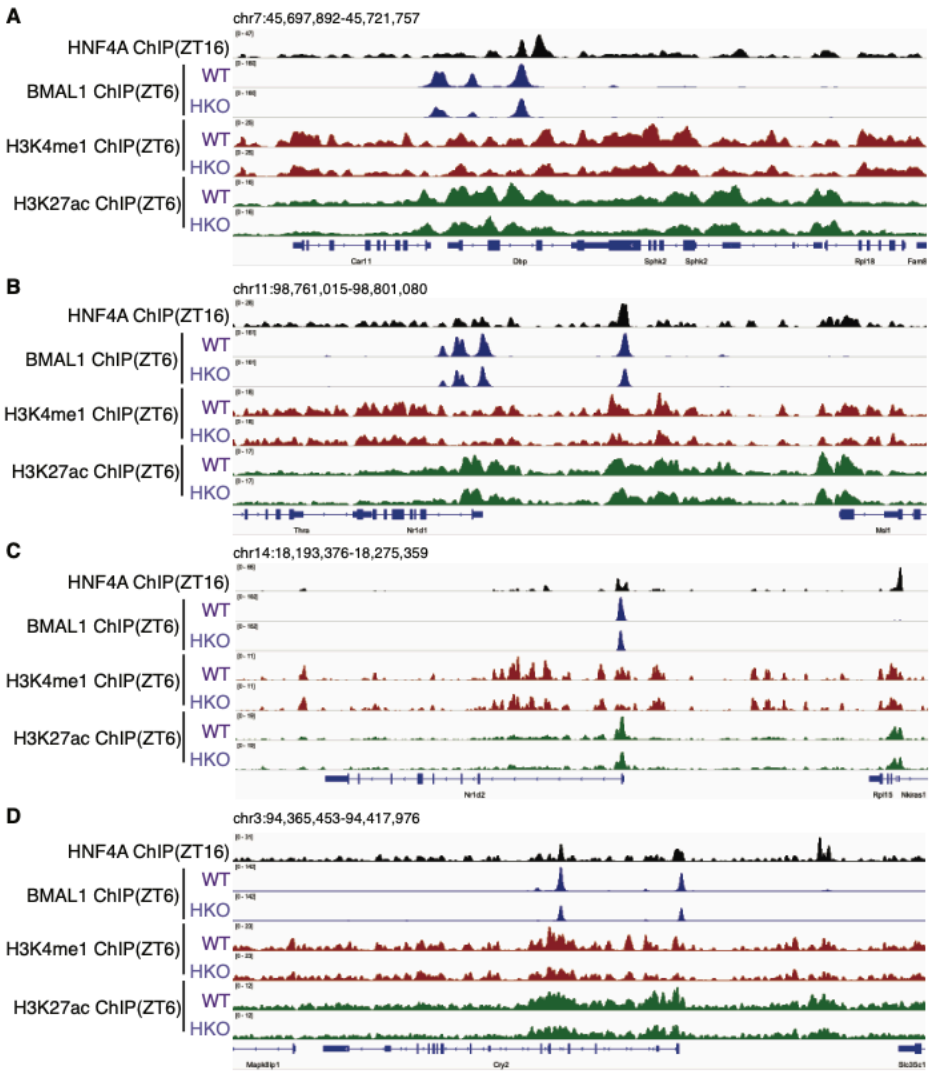


Figure S5. Genome tracks showing HKO-reduced BMAL1 peaks and locally decreased H3K4me1 and H3K27ac marks at core clock genes in the mouse liver. IGV genome tracks of BMAL1, HNF4A, H3K4me1, and H3K27ac enrichment at *Dbp* **(A)**, *Nr1d1* **(B)**, *Nr1d2* **(C)**, and *Cry2* **(D)** gene loci in indicated genotypes based on normalized ChIP-seq read coverage. Track heights are indicated.

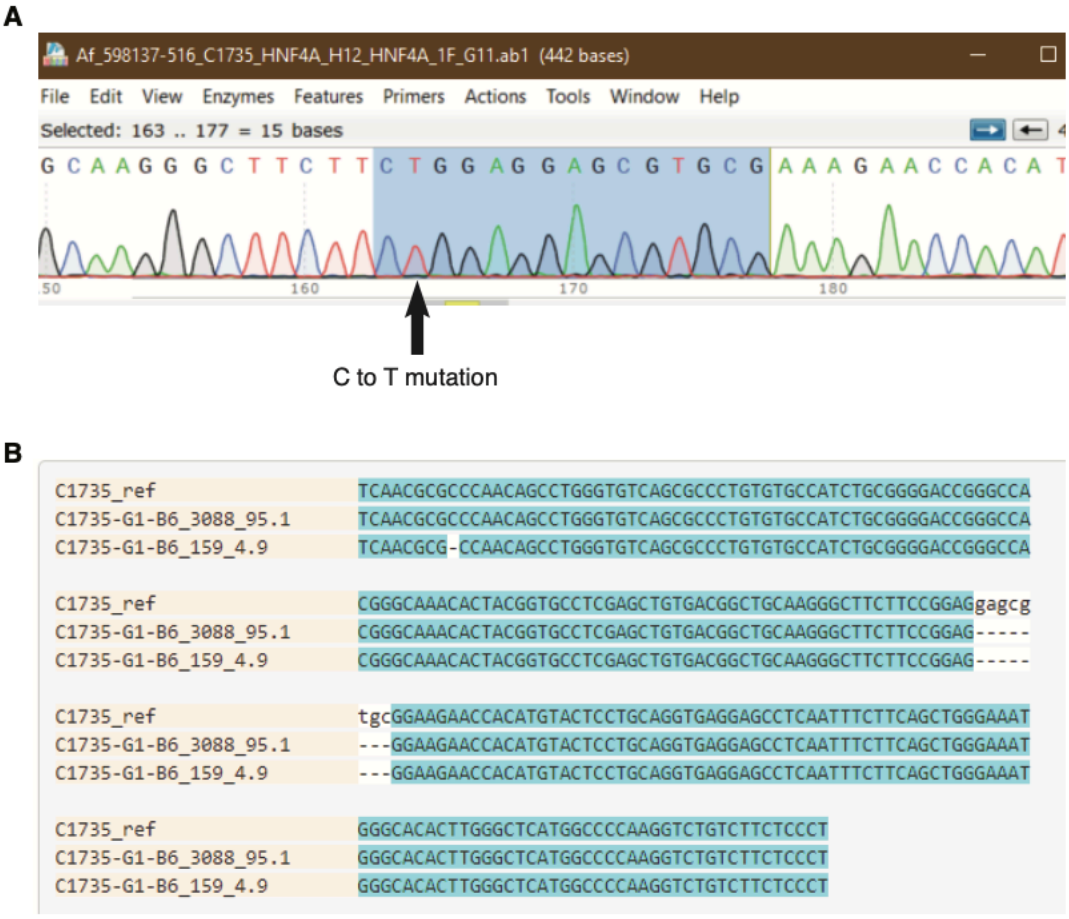


Figure S6. *Hnf4a* knockout and R85W mutation were generated in Hep3B cells using CRISPR-CAS9. (A) The homozygous *Hnf4a-R85W* mutant line was validated by Sanger sequencing. **(B)** The *Hnf4a* knockout line was validated by NGS.

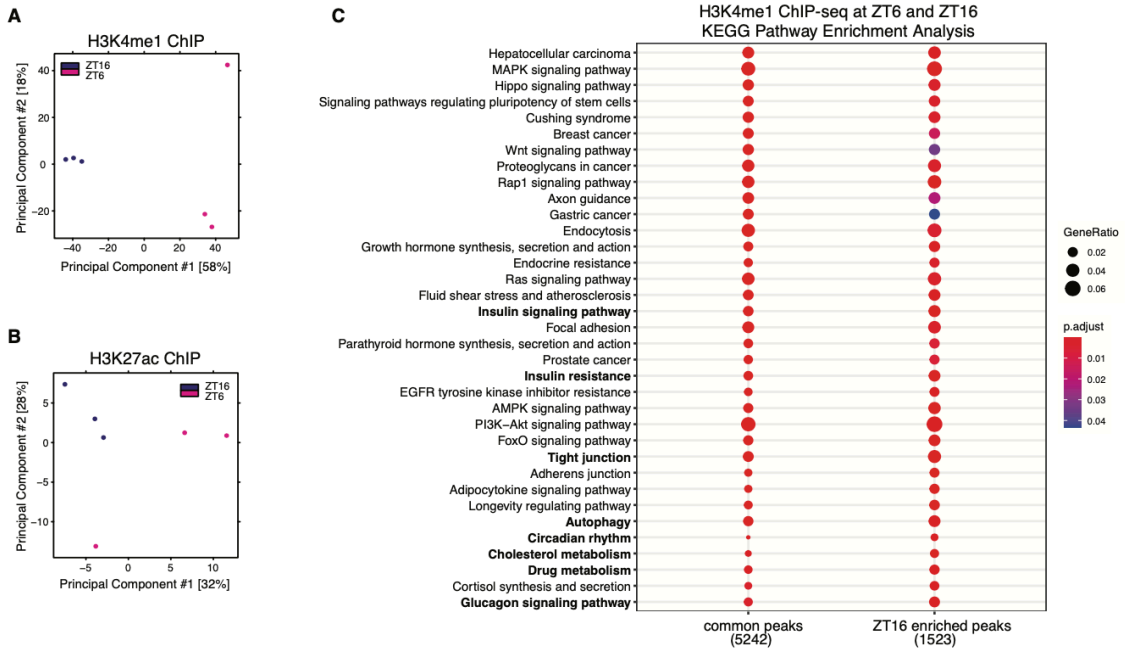


Figure S7. Liver chromatin is more accessible in the evening. (A) PCA plot of H3K4me1 ChIP-seq counts across consensus H3K4me1 peaks at ZT16 and ZT6. **(B)** PCA plot of H3K27ac ChIP-seq counts across consensus H3K27ac peaks at ZT16 and ZT6. **(C)** KEGG pathway analysis of genes at the common or ZT16-enriched peaks identified from a comparison between H3K4me1 ChIP-seq signals at ZT6 and ZT16.

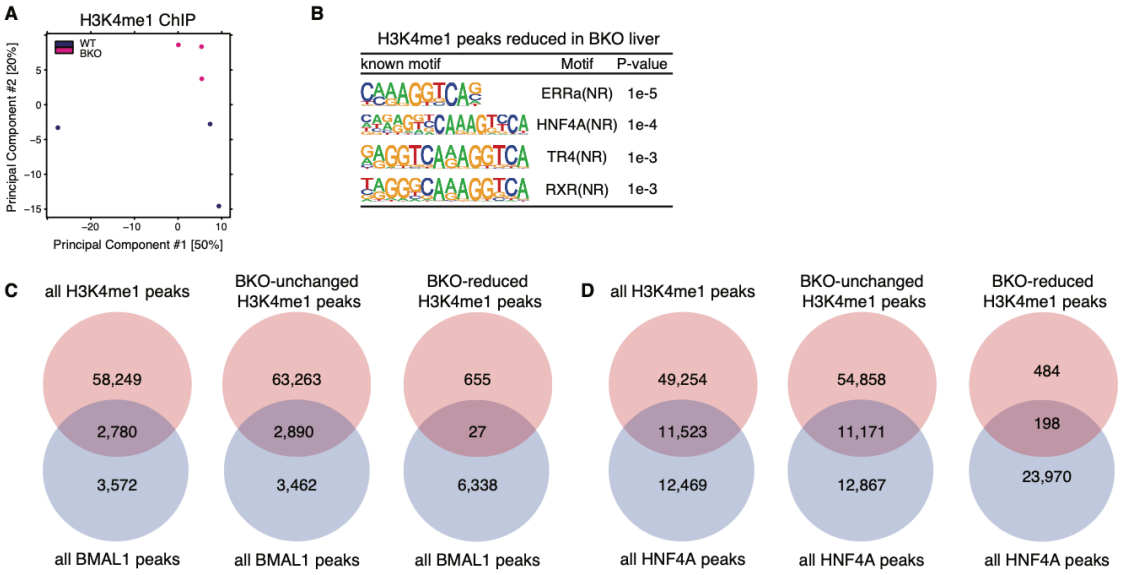


Figure S8. BMAL1 is involved in chromatin remodeling. (A) PCA plot of H3K4me1 ChIP-seq counts across consensus H3K4me1 peaks in WT and BKO (at ZT16). **(B)** Motif analysis of BKO-reduced H3K4me1 sites defined in (Fig. 6A). Known consensus motifs are shown with corresponding enrichment significance values. **(C)** Venn diagram showing overlap between H3K4me1 (at ZT16) and BMAL1 peaks (at ZT6). **(D)** Venn diagram showing overlap between H3K4me1 and HNF4A peaks (both at ZT16).

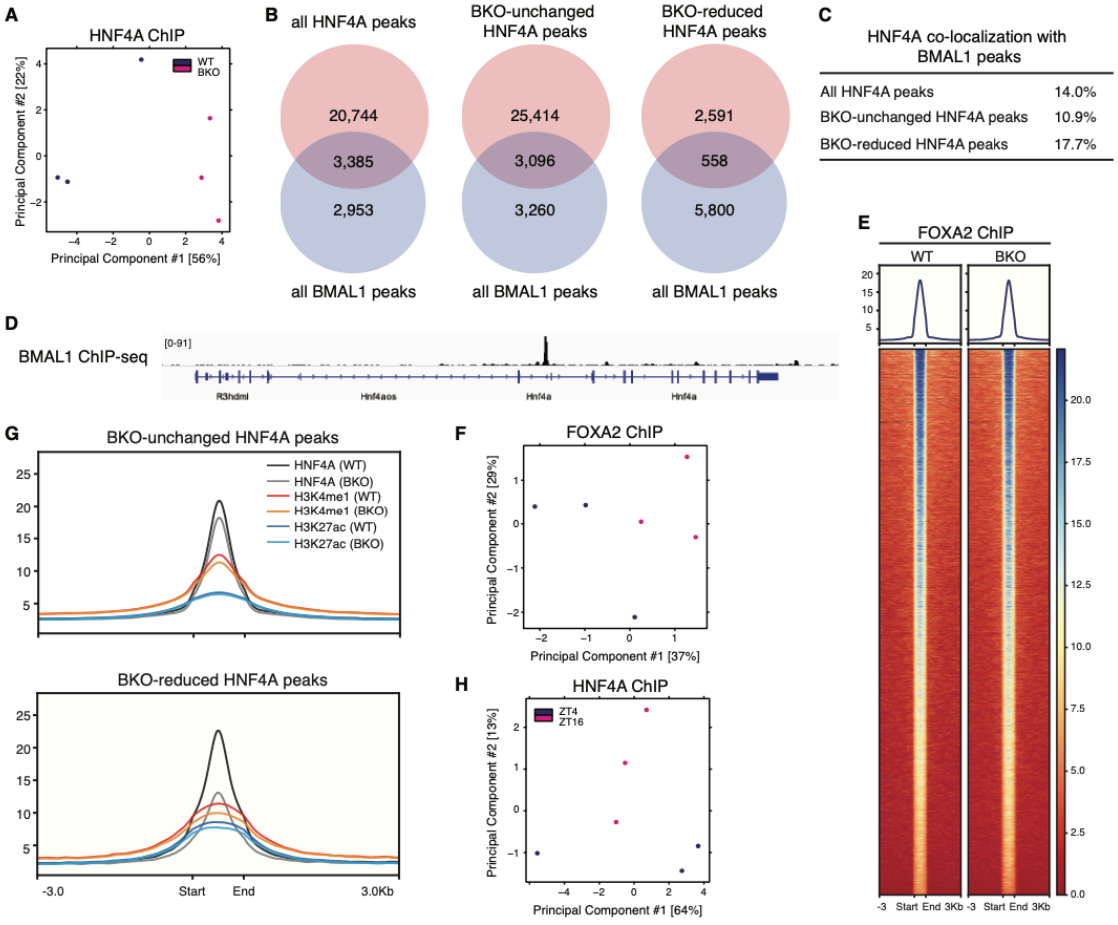


Figure S9. BMAL1 controls chromatin remodeling through regulation of HNF4A. (A) PCA plot of HNF4A ChIP-seq counts across consensus HNF4A peaks in WT and BKO (at ZT16). **(B-C)** Venn diagram showing overlap between BMAL1 (at ZT6) and HNF4A (at ZT16) peaks. **(D)** IGV genome tracks showing BMAL1 enrichment at *Hnf4a* gene. **(E)** Heatmap of FOXA2 ChIP-seq signals (at ZT16) in WT (left) or BKO (right) liver centered at all FOXA2 peaks of WT. Peaks are ordered vertically by signal strength. **(F)** PCA plot of FOXA2 ChIP-seq counts across consensus FOXA2 peaks in WT and BKO. **(G)** Metaplot showing average intensity of HNF4A, H3K4me1, and H3K27ac signals surrounding HNF4A peak centers in WT or BKO liver. **(H)** PCA plot of HNF4A ChIP-seq counts across consensus HNF4A peaks at ZT4 and ZT16 after chronic jet lag.

Bibliography

Fariello, María Inés, Simon Boitard, Hugo Naya, Magali SanCristobal, and Bertrand Servin. 2013. “Detecting Signatures of Selection Through Haplotype Differentiation Among Hierarchically Structured Populations.” Genetics 193 (3): 929–41. https://doi.org/10.1534/genetics.112.147231.

Howard, David M., Lynsey S. Hall, Jonathan D. Hafferty, Yanni Zeng, Mark J. Adams, Toni-Kim Clarke, David J. Porteous, et al. 2017. “Genome-Wide Haplotype-Based Association Analysis of Major Depressive Disorder in Generation Scotland and UK Biobank.” Translational Psychiatry 7 (11): 1–9. https://doi.org/10.1038/s41398-017-0010-9.

Lambert, J.-C., B. Grenier-Boley, D. Harold, D. Zelenika, V. Chouraki, Y. Kamatani, K. Sleegers, et al. 2013. “Genome-Wide Haplotype Association Study Identifies the FRMD4A Gene as a Risk Locus for Alzheimer’s Disease.” Molecular Psychiatry 18 (4): 461–70. https://doi.org/10.1038/mp.2012.14.

Li, Ruidong, Han Qu, Jinfeng Chen, Shibo Wang, John M Chater, Le Zhang, Julong Wei, et al. 2020. “Inference of Chromosome-Length Haplotypes Using Genomic Data of Three or a Few More Single Gametes.” Molecular Biology and Evolution 37 (12): 3684–98. https://doi.org/10.1093/molbev/msaa176.

Li, Xiang, Lin Li, and Jianbing Yan. 2015. “Dissecting Meiotic Recombination Based on Tetrad Analysis by Single-Microspore Sequencing in Maize.” Nature Communications 6 (March): 6648. https://doi.org/10.1038/ncomms7648.

Lohmueller, Kirk E., Carlos D. Bustamante, and Andrew G. Clark. 2009. “Methods for Human Demographic Inference Using Haplotype Patterns From Genomewide Single-Nucleotide Polymorphism Data.” Genetics 182 (1): 217–31. https://doi.org/10.1534/genetics.108.099275.

Palamara, Pier Francesco, Todd Lencz, Ariel Darvasi, and Itsik Pe’er. 2012. “Length Distributions of Identity by Descent Reveal Fine-Scale Demographic History.” American Journal of Human Genetics 91 (5): 809–22. https://doi.org/10.1016/j.ajhg.2012.08.030.

Yang, Jian, Teresa Ferreira, Andrew P. Morris, Sarah E. Medland, Genetic Investigation of ANthropometric Traits (GIANT) Consortium, DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, Pamela A. F. Madden, et al. 2012. “Conditional and Joint Multiple-SNP Analysis of GWAS Summary Statistics Identifies Additional Variants Influencing Complex Traits.” Nature Genetics 44 (4): 369–75. https://doi.org/10.1038/ng.2213.

Zhang, Fan, Chunchao Wang, Min Li, Yanru Cui, Yingyao Shi, Zhichao Wu, Zhiqiang Hu, Wensheng Wang, Jianlong Xu, and Zhikang Li. 2021. “The Landscape of Gene-CDS-Haplotype Diversity in Rice: Properties, Population Organization, Footprints of Domestication and Breeding, and Implications for Genetic Improvement.” Molecular Plant 14 (5): 787–804. https://doi.org/10.1016/j.molp.2021.02.003.

Albert JH, Chib S. 1993. Bayesian Analysis of Binary and Polychotomous Response Data. Journal of the American Statistical Association 88: 669-679.

Azevedo C, Andrade D. 2013. CADEM: A conditional augmented data EM algorithm for fitting one parameter probit models. Brazilian Journal of Probability and Statistics 27: 245-262.

Bliss CI. 1934. The method of probits. Science 79: 38-39.

Bliss CI. 1935. The calculation of the dosage-mortality curve. Annals of Applied Biology 22: 134-167.

Breslow NE, Clayton DG. 1993. Approximate inference in generalized linear mixed models. J Am Stat Assoc 88: 9-25.

Brooks S. 1998. Markov chain Monte Carlo method and its application. Journal of the Royal Statistical Society: Series D (The Statistician) 47: 69-100.

Burton PR, et. al. 1999. Genetic variance components analysis for binary phenotypes using generalized linear mixed models (GLMMs) and Gibbs sampling. Genetic Epidemiology 17: 118-140.

Camilleri L. 2009. Bias of Standard Errors in Latent Class Model Applications Using Newton-Raphson and EM Algorithms. JACIII 13: 537-541.

Chakraborty S, Khare K. 2017. Convergence properties of Gibbs samplers for Bayesian probit regression with proper priors. Electronic Journal of Statistics 11: 177-210, 134.

Czado C. 1994. Bayesian inference of binary regression models with parametric link. Journal of Statistical Planning and Inference 41: 121-140.

DeMaris A. 1995. A Tutorial in Logistic Regression. Journal of Marriage and Family 57: 956-968.

Dempster AP, Laird NM, Rubin DB. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society Series B (Methodological) 39: 1-38.

Efron B. 1979. Bootstrap methods: Another look at the Jackknife. The Annals of Statistics 7: 1-26.

Fanney DJ. 1952. Probit Analysis. Cambridge University Press, U. K.

Fisher RA. 1935. Appendix to “The Calculation of the Dose-Mortality Curve” by C. Bliss. Annals of Applied Biology 22: 164-165.

Girolami M, Rogers S. 2006. Variational Bayesian Multinomial Probit Regression with Gaussian Process Priors. Neural Computation 18: 1790-1817.

Hosmer DW, Lemeshow S. 1989. Applied Logistic Regression. New York, Wiley.

Liu C. 2004. Robit Regression: A Simple Robust Alternative to Logistic and Probit Regression. In Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives, doi:https://doi.org/10.1002/0470090456.ch21, pp. 227-238.

Louis TA. 1982. Finding the observed information matrix when using the EM algorithm. Journal of the Royal Statistical Society Series B (Methodological) 44: 226-233.

McCullagh P, Nelder JA. 1989. Generalized Linear Mixed Models. Chapman & Hall/CRC, New York.

McCulloch CE. 1994. Maximum likelihood variance components estimation for binary data. Journal of the American Statistical Association 89: 330-335.

McCulloch CE. 2000. Generalized linear models. Journal of the American Statistical Association 95: 1320-1324.

McDermott P, Snyder J, Willison R. 2016. Methods for Bayesian variable selection with binary response data using the EM algorithm. arXiv preprint arXiv:160505429.

Nadarajah S, Kotz S. 2006. R Programs for Truncated Distributions. Journal of Statistical Software, Code Snippets 16: 1 - 8.

Nelder JA, Wedderburn RWM. 1972. Generalized Linear Models. Journal of the Royal Statistical Society, Ser A 135: 370-384.

Saitoh K, Onishi K, Mikami I, Thidar K, Sano Y. 2004. Allelic diversification at the C (OsC1) locus of wild and cultivated rice: nucleotide changes associated with phenotypes. Genetics 168: 997-1007.

Schafer DW. 1993. Likelihood analysis for probit regression with measurement errors. Biometrika 80: 899-904.

Sorensen DA, Andersen S, Gianola D, Korsgaard I. 1995. Bayesian inference in threshold models using Gibbs sampling. Genet Sel Evol 27: 229-249.

Visscher PM, Haley CS, Knott SA. 2009. Mapping QTLs for binary traits in backcross and F2 populations. Genetical Research 68: 55-63.

Wolfinger c, O'connell M. 1993. Generalized linear mixed models a pseudolikelihood approach. Journal of Statistical Computation and Simulation 48: 233-243.

Xu S, Yi N, Burke D, Galecki A, Miller RA. 2003. An EM algorithm for mapping binary disease loci: application to fibrosarcoma in a four-way cross mouse family. Genetical Research 82: 127-138.

Yi N, Xu S. 2000. Bayesian mapping of quantitative trait loci for complex binary traits. Genetics 155: 1391-1403.

Yu H, Xie W, Wang J, Xing Y, Xu C, Li X, Xiao J, Zhang Q. 2011. Gains in QTL detection using an ultra-high density SNP map based on population sequencing relative to traditional RFLP/SSR markers. PLoS One 6: e17595. doi:17510.11371/journal.pone.0017595.

Alonge, Michael, Ludivine Lebeigle, Melanie Kirsche, Sergey Aganezov, Xingang Wang, Zachary B. Lippman, Michael C. Schatz, and Sebastian Soyk. 2021. "Automated Assembly Scaffolding Elevates a New Tomato System for High-Throughput Genome Editing." bioRxiv. https://doi.org/10.1101/2021.11.18.469135.

Andrews, S. 2010. "FastQC A Quality Control Tool for High Throughput Sequence Data." 2010. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Bao, Weidong, Kenji K. Kojima, and Oleksiy Kohany. 2015. "Repbase Update, a Database of Repetitive Elements in Eukaryotic Genomes." Mobile DNA 6 (1): 11. https://doi.org/10.1186/s13100-015-0041-9.

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. "Trimmomatic: A Flexible Trimmer for Illumina Sequence Data." Bioinformatics 30 (15): 2114–20. https://doi.org/10.1093/bioinformatics/btu170.

Brůna, Tomáš, Alexandre Lomsadze, and Mark Borodovsky. 2020. "GeneMark-EP+: Eukaryotic Gene Prediction with Self-Training in the Space of Genes and Proteins." NAR Genomics and Bioinformatics 2 (2): lqaa026. https://doi.org/10.1093/nargab/lqaa026.

Buchfink, Benjamin, Klaus Reuter, and Hajk-Georg Drost. 2021. "Sensitive Protein Alignments at Tree-of-Life Scale Using DIAMOND." Nature Methods 18 (4): 366–68. https://doi.org/10.1038/s41592-021-01101-x.

Capella-Gutiérrez, Salvador, José M. Silla-Martínez, and Toni Gabaldón. 2009. "TrimAl: A Tool for Automated Alignment Trimming in Large-Scale Phylogenetic Analyses." Bioinformatics 25 (15): 1972–73. https://doi.org/10.1093/bioinformatics/btp348.

Chandra, Ram, Dhinesh Babu, Vilas Jadhav, and Jaime Teixeira da Silva. 2010. "Origin, History and Domestication of Pomegranate." Fruit, Vegetable and Cereal Science and Biotechnology 4 (December): 1–6.

Chen, M., T.K. Zhang, and Z.H. Yuan. 2019. "Evolution and Classification of Pomegranate." Acta Horticulturae, no. 1254 (October): 41–48. https://doi.org/10.17660/ActaHortic.2019.1254.7.

Claros, Manuel Gonzalo, Rocío Bautista, Darío Guerrero-Fernández, Hicham Benzerki, Pedro Seoane, and Noé Fernández-Pozo. 2012. "Why Assembling Plant Genome Sequences Is So Challenging." Biology 1 (2): 439–59. https://doi.org/10.3390/biology1020439.

Cvrčková, F. 2016. "A Plant Biologists' Guide to Phylogenetic Analysis of Biological Macromolecule Sequences." Biologia Plantarum 60 (4): 619–27. https://doi.org/10.1007/s10535-016-0649-8.

Emms, David M., and Steven Kelly. 2019. “OrthoFinder: Phylogenetic Orthology Inference for Comparative Genomics.” Genome Biology 20 (1): 238. https://doi.org/10.1186/s13059-019-1832-y.

Ewels, Philip, Måns Magnusson, Sverker Lundin, and Max Käller. 2016. “MultiQC: Summarize Analysis Results for Multiple Tools and Samples in a Single Report.” Bioinformatics 32 (19): 3047–48. https://doi.org/10.1093/bioinformatics/btw354.

Gurevich, Alexey, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. 2013. “QUAST: Quality Assessment Tool for Genome Assemblies.” Bioinformatics (Oxford, England) 29 (8): 1072–75. https://doi.org/10.1093/bioinformatics/btt086.

Haas, Brian J, Steven L Salzberg, Wei Zhu, Mihaela Pertea, Jonathan E Allen, Joshua Orvis, Owen White, C Robin Buell, and Jennifer R Wortman. 2008. “Automated Eukaryotic Gene Structure Annotation Using EVidenceModeler and the Program to Assemble Spliced Alignments.” Genome Biology 9 (1): R7. https://doi.org/10.1186/gb-2008-9-1-r7.

Hubley, Robert, Robert D. Finn, Jody Clements, Sean R. Eddy, Thomas A. Jones, Weidong Bao, Arian F.A. Smit, and Travis J. Wheeler. 2016. “The Dfam Database of Repetitive DNA Families.” Nucleic Acids Research 44 (Database issue): D81–89. https://doi.org/10.1093/nar/gkv1272.

Ja, Guerrero-Solano, Jaramillo-Morales Oa, Jiménez-Cabrera T, Urrutia-Hernández Ta, Chehue-Romero A, Olvera-Hernández Eg, and Bautista M. 2020. “Punica Protopunica Balf., the Forgotten Sister of the Common Pomegranate (Punica Granatum L.): Features and Medicinal Properties-A Review.” Plants (Basel, Switzerland) 9 (9). https://doi.org/10.3390/plants9091214.

Kalyaanamoorthy, Subha, Bui Quang Minh, Thomas K. F. Wong, Arndt von Haeseler, and Lars S. Jermiin. 2017. “ModelFinder: Fast Model Selection for Accurate Phylogenetic Estimates.” Nature Methods 14 (6): 587–89. https://doi.org/10.1038/nmeth.4285.

Kapli, Paschalia, Ziheng Yang, and Maximilian J. Telford. 2020. “Phylogenetic Tree Building in the Genomic Age.” Nature Reviews Genetics 21 (7): 428–44. https://doi.org/10.1038/s41576-020-0233-0.

Katoh, Kazutaka, Kazuharu Misawa, Kei-ichi Kuma, and Takashi Miyata. 2002. “MAFFT: A Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform.” Nucleic Acids Research 30 (14): 3059–66. https://doi.org/10.1093/nar/gkf436.

Korf, Ian. 2004. “Gene Finding in Novel Genomes.” BMC Bioinformatics 5 (1): 59. https://doi.org/10.1186/1471-2105-5-59.

Le, Si Quang, and Olivier Gascuel. 2008. “An Improved General Amino Acid Replacement Matrix.” Molecular Biology and Evolution 25 (7): 1307–20. https://doi.org/10.1093/molbev/msn067.

Luo, Xiang, Haoxian Li, Zhikun Wu, Wen Yao, Peng Zhao, Da Cao, Haiyan Yu, et al. 2020. "The Pomegranate (Punica Granatum L.) Draft Genome Dissects Genetic Divergence between Soft- and Hard-Seeded Cultivars." Plant Biotechnology Journal 18 (4): 955–68. https://doi.org/10.1111/pbi.13260.

Majoros, W. H., M. Pertea, and S. L. Salzberg. 2004. "TigrScan and GlimmerHMM: Two Open Source Ab Initio Eukaryotic Gene-Finders." Bioinformatics (Oxford, England) 20 (16): 2878–79. https://doi.org/10.1093/bioinformatics/bth315.

Manni, Mosè, Matthew R Berkeley, Mathieu Seppey, Felipe A Simão, and Evgeny M Zdobnov. 2021. "BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes." Molecular Biology and Evolution 38 (10): 4647–54. https://doi.org/10.1093/molbev/msab199.

Marçais, Guillaume, and Carl Kingsford. 2011. "A Fast, Lock-Free Approach for Efficient Parallel Counting of Occurrences of k-Mers." Bioinformatics 27 (6): 764–70. https://doi.org/10.1093/bioinformatics/btr011.

McKain, Michael R., Matthew G. Johnson, Simon Uribe-Convers, Deren Eaton, and Ya Yang. 2018. "Practical Considerations for Plant Phylogenomics." Applications in Plant Sciences 6 (3). https://doi.org/10.1002/aps3.1038.

Nguyen, Lam-Tung, Heiko A. Schmidt, Arndt von Haeseler, and Bui Quang Minh. 2015. "IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies." Molecular Biology and Evolution 32 (1): 268–74. https://doi.org/10.1093/molbev/msu300.

Palmer, Jon, and Jason Stajich. 2019. "Nextgenusfs/Funannotate: Funannotate v1.5.3." Zenodo. https://doi.org/10.5281/zenodo.2604804.

Prjibelski, Andrey, Dmitry Antipov, Dmitry Meleshko, Alla Lapidus, and Anton Korobeynikov. 2020. "Using SPAdes De Novo Assembler." Current Protocols in Bioinformatics 70 (1): e102. https://doi.org/10.1002/cpbi.102.

Qin, Gaihua, Chunyan Xu, Ray Ming, Haibao Tang, Romain Guyot, Elena M. Kramer, Yudong Hu, et al. 2017. "The Pomegranate (Punica Granatum L.) Genome and the Genomics of Punicalagin Biosynthesis." The Plant Journal 91 (6): 1108–28. https://doi.org/10.1111/tpj.13625.

Shahsavari, Shiva, Zahra Noormohammadi, Masoud Sheidai, Farah Farahani, and Mohammad Reza Vazifeshenas. 2022. "A Bioinformatic Insight into the Genetic Diversity within Pomegranate Cultivars: From Nuclear to Chloroplast Genes." Genetic Resources and Crop Evolution 69 (3): 1207–17. https://doi.org/10.1007/s10722-021-01297-z.

Sims, Gregory E., Se-Ran Jun, Guohong Albert Wu, and Sung-Hou Kim. 2009. “Whole-Genome Phylogeny of Mammals: Evolutionary Information in Genic and Nongenic Regions.” Proceedings of the National Academy of Sciences 106 (40): 17077–82. https://doi.org/10.1073/pnas.0909377106.

Slater, Guy St C, and Ewan Birney. 2005. “Automated Generation of Heuristics for Biological Sequence Comparison.” BMC Bioinformatics 6 (February): 31. https://doi.org/10.1186/1471-2105-6-31.

Smit, Arian F., Robert Hubley, Jullien M. Flynn, Clément Goubert, Jeb Rosen, Andrew G. Clark, and Cédric Feschotte. 2008. “RepeatModeler Open-1.0.” Preprint. Genomics. https://doi.org/10.1101/856591.

Smit, Arian F., Robert Hubley, Jullien M. Flynn. 2013. “RepeatMasker Open-4.0.” Preprint. Genomics. https://doi.org/10.1101/856591.

Stanke, Mario, Oliver Keller, Irfan Gunduz, Alec Hayes, Stephan Waack, and Burkhard Morgenstern. 2006. “AUGUSTUS: Ab Initio Prediction of Alternative Transcripts.” Nucleic Acids Research 34 (suppl_2): W435–39. https://doi.org/10.1093/nar/gkl200.

Teixeira da Silva, Jaime, Tikam Rana, Diganta Narzary, Nidhi Verma, Deodas Meshram, and Shirish Ranade. 2013. “Pomegranate Biology and Biotechnology: A Review.” Scientia Horticulturae 160 (August): 85–107.

Usha, Talambedu, Sushil Kumar Middha, Dinesh Babu, Arvind Kumar Goyal, Anupam J. Das, Deepti Saini, Aditya Sarangi, et al. 2022. “Hybrid Assembly and Annotation of the Genome of the Indian Punica Granatum, a Superfood.” Frontiers in Genetics 13: 786825. https://doi.org/10.3389/fgene.2022.786825.

Voshall, Adam, and Etsuko N. Moriyama. 2018. “Next-Generation Transcriptome Assembly: Strategies and Performance Analysis.” Bioinformatics in the Era of Post Genomics and Big Data, June. https://doi.org/10.5772/intechopen.73497.

Vurture, Gregory W, Fritz J Sedlazeck, Maria Nattestad, Charles J Underwood, Han Fang, James Gurtowski, and Michael C Schatz. 2017. “GenomeScope: Fast Reference-Free Genome Profiling from Short Reads.” Bioinformatics 33 (14): 2202–4. https://doi.org/10.1093/bioinformatics/btx153.

Wang, Huai-Chun, Bui Quang Minh, Edward Susko, and Andrew J. Roger. 2018. “Modeling Site Heterogeneity with Posterior Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation.” Systematic Biology 67 (2): 216–35. https://doi.org/10.1093/sysbio/syx068.

Young, Andrew D., and Jessica P. Gillung. 2020. “Phylogenomics — Principles, Opportunities and Pitfalls of Big-Data Phylogenetics.” Systematic Entomology 45 (2): 225–47. https://doi.org/10.1111/syen.12406.

Youssef, Muhammad, Arif Saeed Alhammadi, Jorge Humberto Ramírez-Prado, Lorenzo Felipe Sánchez-Teyer, and Rosa María Escobedo-GraciaMedrano. 2018. “Remarks on Genetic Diversity and Relationship of Punica Protopunica and P. Granatum Assessed by Molecular Analyses.” Genetic Resources and Crop Evolution 65 (2): 577–90. https://doi.org/10.1007/s10722-017-0556-7.

Yuan, Zhaohe, Yanming Fang, Taikui Zhang, Zhangjun Fei, Fengming Han, Cuiyu Liu, Min Liu, et al. 2018. “The Pomegranate (Punica Granatum L.) Genome Provides Insights into Fruit Quality and Ovule Developmental Biology.” Plant Biotechnology Journal 16 (7): 1363–74. https://doi.org/10.1111/pbi.12875.

Zeynalova, Aydan. 2017. “ORIGIN, TAXONOMY AND SYSTEMATICS OF POMEGRANATE.” Journal of Botany of ANAS, January. https://www.academia.edu/40364877/ORIGIN_TAXONOMY_AND_SYSTEMATICS_OF_POMEGRANATE.

Zhang, Feng, Yinhuan Ding, Chao-Dong Zhu, Xin Zhou, Michael C. Orr, Stefan Scheu, and Yun-Xia Luan. 2019. “Phylogenomics from Low-Coverage Whole-Genome Sequencing.” Methods in Ecology and Evolution 10 (4): 507–17. https://doi.org/10.1111/2041-210X.13145.

Zhou, Huijuan, Yiheng Hu, Aziz Ebrahimi, Peiliang Liu, Keith Woeste, Peng Zhao, and Shuoxin Zhang. 2021. “Whole Genome Based Insights into the Phylogeny and Evolution of the Juglandaceae.” BMC Ecology and Evolution 21 (1): 191. https://doi.org/10.1186/s12862-021-01917-3.

Bass, Joseph, and Mitchell A. Lazar. 2016. “Circadian Time Signatures of Fitness and Disease.” Science 354 (6315): 994–99. https://doi.org/10.1126/science.aah4965.

Beytebiere, Joshua R., Alexandra J. Trott, Ben J. Greenwell, Collin A. Osborne, Helene Vitet, Jessica Spence, Seung-Hee Yoo, et al. 2019. “Tissue-Specific BMAL1 Cistromes Reveal That Rhythmic Transcription Is Associated with Rhythmic Enhancer–Enhancer Interactions.” Genes & Development 33 (5–6): 294–309. https://doi.org/10.1101/gad.322198.118.

Boergesen, Michael, Thomas Åskov Pedersen, Barbara Gross, Simon J. van Heeringen, Dik Hagenbeek, Christian Bindesbøll, Sandrine Caron, et al. 2012. “Genome-Wide Profiling of Liver X Receptor, Retinoid X Receptor, and Peroxisome Proliferator-Activated Receptor α in Mouse Liver Reveals Extensive Sharing of Binding Sites.” Molecular and Cellular Biology 32 (4): 852–67. https://doi.org/10.1128/MCB.06175-11.

Chen, Lei, Natalie H. Toke, Shirley Luo, Roshan P. Vasoya, Rohit Aita, Aditya Parthasarathy, Yu-Hwai Tsai, Jason R. Spence, and Michael P. Verzi. 2019. “HNF4 Factors Control Chromatin Accessibility and Are Redundantly Required for Maturation of the Fetal Intestine.” Development 146 (19). https://doi.org/10.1242/dev.179432.

Chen, Lei, Natalie H. Toke, Shirley Luo, Roshan P. Vasoya, Robert L. Fullem, Aditya Parthasarathy, Ansu O. Perekatt, and Michael P. Verzi. 2019. “A Reinforcing HNF4–SMAD4 Feed-Forward Module Stabilizes Enterocyte Identity.” Nature Genetics 51 (5): 777–85. https://doi.org/10.1038/s41588-019-0384-0.

Chen, W. S., K. Manova, D. C. Weinstein, S. A. Duncan, A. S. Plump, V. R. Prezioso, R. F. Bachvarova, and J. E. Darnell. 1994. “Disruption of the HNF-4 Gene, Expressed in Visceral Endoderm, Leads to Cell Death in Embryonic Ectoderm and Impaired Gastrulation of Mouse Embryos.” Genes & Development 8 (20): 2466–77. https://doi.org/10.1101/gad.8.20.2466.

Clapier, Cedric R., and Bradley R. Cairns. 2009. “The Biology of Chromatin Remodeling Complexes.” Annual Review of Biochemistry 78 (1): 273–304. https://doi.org/10.1146/annurev.biochem.77.062706.153223.

Clapier, Cedric R., Janet Iwasa, Bradley R. Cairns, and Craig L. Peterson. 2017. “Mechanisms of Action and Regulation of ATP-Dependent Chromatin-Remodelling Complexes.” Nature Reviews Molecular Cell Biology 18 (7): 407–22. https://doi.org/10.1038/nrm.2017.26.

Colclough, Kevin, Christine Bellanne-Chantelot, Cecile Saint-Martin, Sarah E. Flanagan, and Sian Ellard. 2013. “Mutations in the Genes Encoding the Transcription Factors Hepatocyte Nuclear Factor 1 Alpha and 4 Alpha in Maturity-Onset Diabetes of the Young and Hyperinsulinemic Hypoglycemia.” Human Mutation 34 (5): 669–85. https://doi.org/10.1002/humu.22279.

Corces, M. Ryan, Jeffrey M. Granja, Shadi Shams, Bryan H. Louie, Jose A. Seoane, Wanding Zhou, Tiago C. Silva, et al. 2018. “The Chromatin Accessibility Landscape of Primary Human Cancers.” Science 362 (6413). https://doi.org/10.1126/science.aav1898.

Deans, Jonathan R., Poonamjot Deol, Nina Titova, Sarah H. Radi, Linh M. Vuong, Jane R. Evans, Songqin Pan, et al. 2021. “HNF4α Isoforms Regulate the Circadian Balance between Carbohydrate and Lipid Metabolism in the Liver.” BioRxiv, February, 2021.02.28.433261. https://doi.org/10.1101/2021.02.28.433261.

Dong, Zhen, Guoxin Zhang, Meng Qu, Ryan C. Gimple, Qiulian Wu, Zhixin Qiu, Briana C. Prager, et al. 2019. “Targeting Glioblastoma Stem Cells through Disruption of the Circadian Clock.” Cancer Discovery 9 (11): 1556–73. https://doi.org/10.1158/2159-8290.CD-19-0215.

Fang, Bin, Logan J. Everett, Jennifer Jager, Erika Briggs, Sean M. Armour, Dan Feng, Ankur Roy, Zachary Gerhart-Hines, Zheng Sun, and Mitchell A. Lazar. 2014. “Circadian Enhancers Coordinate Multiple Phases of Rhythmic Gene Transcription In Vivo.” Cell 159 (5): 1140–52. https://doi.org/10.1016/j.cell.2014.10.022.

Faure, Andre J., Dominic Schmidt, Stephen Watt, Petra C. Schwalie, Michael D. Wilson, Huiling Xu, Robert G. Ramsay, Duncan T. Odom, and Paul Flicek. 2012. “Cohesin Regulates Tissue-Specific Expression by Stabilizing Highly Occupied Cis-Regulatory Modules.” Genome Research 22 (11): 2163–75. https://doi.org/10.1101/gr.136507.111.

Fekry, Baharan, Aleix Ribas-Latre, Corrine Baumgartner, Jonathan R. Deans, Christopher Kwok, Pooja Patel, Loning Fu, et al. 2018. “Incompatibility of the Circadian Protein BMAL1 and HNF4α in Hepatocellular Carcinoma.” Nature Communications 9 (1). https://doi.org/10.1038/s41467-018-06648-6.

Fekry, Baharan, Aleix Ribas-Latre, Corrine Baumgartner, Alaa M. T. Mohamed, Mikhail G. Kolonin, Frances M. Sladek, Mamoun Younes, and Kristin L. Eckel-Mahan. 2019. “HNF4α-Deficient Fatty Liver Provides a Permissive Environment for Sex-Independent Hepatocellular Carcinoma.” Cancer Research 79 (22): 5860–73.

Flanagan, S. E., R. R. Kapoor, G. Mali, D. Cody, N. Murphy, B. Schwahn, T. Siahanidou, et al. 2010. “Diazoxide-Responsive Hyperinsulinemic Hypoglycemia Caused by HNF4A Gene Mutations.” European Journal of Endocrinology 162 (5): 987–92. https://doi.org/10.1530/EJE-09-0861.

Gupta, Rana K., Marko Z. Vatamaniuk, Catherine S. Lee, Reed C. Flaschen, James T. Fulmer, Franz M. Matschinsky, Stephen A. Duncan, and Klaus H. Kaestner. 2005. “The MODY1 Gene HNF-4α Regulates Selected Genes Involved in Insulin Secretion.” Journal of Clinical Investigation 115 (4): 1006–15. https://doi.org/10.1172/JCI200522365.

Hatziapostolou, Maria, Christos Polytarchou, Eleni Aggelidou, Alexandra Drakaki, George A. Poultsides, Savina A. Jaeger, Hisanobu Ogata, et al. 2011. “An HNF4α-MiRNA Inflammatory Feedback Circuit Regulates Hepatocellular Oncogenesis.” Cell 147 (6): 1233–47. https://doi.org/10.1016/j.cell.2011.10.043.

Hayhurst, Graham P., Ying-Hue Lee, Gilles Lambert, Jerrold M. Ward, and Frank J. Gonzalez. 2001a. “Hepatocyte Nuclear Factor 4α (Nuclear Receptor 2A1) Is Essential for Maintenance of Hepatic Gene Expression and Lipid Homeostasis.” Molecular and Cellular Biology 21 (4): 1393–1403. https://doi.org/10.1128/MCB.21.4.1393-1403.2001.

Hayhurst, G. P., Lee, Y.-H., Lambert, G., Ward, J. M. & Gonzalez, F. J. 2001b. “Hepatocyte Nuclear Factor 4α (Nuclear Receptor 2A1) Is Essential for Maintenance of Hepatic Gene Expression and Lipid Homeostasis.” Molecular and Cellular Biology 21 (4): 1393–1403. https://doi.org/10.1128/MCB.21.4.1393-1403.2001.

Hirota, Tsuyoshi, Jae Wook Lee, Peter C. St John, Mariko Sawa, Keiko Iwaisako, Takako Noguchi, Pagkapol Y. Pongsawakul, et al. 2012. “Identification of Small Molecule Activators of Cryptochrome.” Science 337 (6098): 1094–97. https://doi.org/10.1126/science.1223710.

Holloway, Minita G., Gregory D. Miles, Alan A. Dombkowski, and David J. Waxman. 2008. “Liver-Specific Hepatocyte Nuclear Factor-4α Deficiency: Greater Impact on Gene Expression in Male than in Female Mouse Liver.” Molecular Endocrinology 22 (5): 1274–86. https://doi.org/10.1210/me.2007-0564.

Hwang-Verslues, Wendy W, and Frances M Sladek. 2010. “HNF4α—Role in Drug Metabolism and Potential Drug Target?” Current Opinion in Pharmacology, Endocrine and metabolic diseases/New technologies - the importance of protein dynamics, 10 (6): 698–705. https://doi.org/10.1016/j.coph.2010.08.010.

Improda, Nicola, Pratik Shah, Maria Güemes, Clare Gilbert, Kate Morgan, Neil Sebire, Detlef Bockenhauer, and Khalid Hussain. 2016. “Hepatocyte Nuclear Factor-4 Alfa Mutation Associated with Hyperinsulinaemic Hypoglycaemia and Atypical Renal Fanconi Syndrome: Expanding the Clinical Phenotype.” Hormone Research in Paediatrics 86 (5): 337–41. https://doi.org/10.1159/000446396.

Jozwik, Kamila M., Igor Chernukhin, Aurelien A. Serandour, Sankari Nagarajan, and Jason S. Carroll. 2016. “FOXA1 Directs H3K4 Monomethylation at Enhancers via Recruitment of the Methyltransferase MLL3.” Cell Reports 17 (10): 2715–23. https://doi.org/10.1016/j.celrep.2016.11.028.

Koike, Nobuya, Seung-Hee Yoo, Hung-Chung Huang, Vivek Kumar, Choogon Lee, Tae-Kyung Kim, and Joseph S. Takahashi. 2012. “Transcriptional Architecture and Chromatin Landscape of the Core Circadian Clock in Mammals.” Science 338 (6105): 349–54. https://doi.org/10.1126/science.1226339.

Kriebs, Anna, Sabine D. Jordan, Erin Soto, Emma Henriksson, Colby R. Sandate, Megan E. Vaughan, Alanna B. Chan, et al. 2017. “Circadian Repressors CRY1 and CRY2 Broadly Interact with Nuclear Receptors and Modulate Transcriptional Activity.” Proceedings of the National Academy of Sciences, July, 201704955. https://doi.org/10.1073/pnas.1704955114.

Lambert, Élie, Jean-Philippe Babeu, Joël Simoneau, Jennifer Raisch, Laurie Lavergne, Dominique Lévesque, Émilie Jolibois, et al. 2020. “Human Hepatocyte Nuclear Factor 4-α Encodes Isoforms with Distinct Transcriptional Functions*.” Molecular & Cellular Proteomics 19 (5): 808–27. https://doi.org/10.1074/mcp.RA119.001909.

Li, Jixuan, Gang Ning, and Stephen A. Duncan. 2000. “Mammalian Hepatocyte Differentiation Requires the Transcription Factor HNF-4α.” Genes & Development 14 (4): 464–74. https://doi.org/10.1101/gad.14.4.464.

Liu, Andrew C., Hien G. Tran, Eric E. Zhang, Aaron A. Priest, David K. Welsh, and Steve A. Kay. 2008. “Redundant Function of REV-ERBα and β and Non-Essential Role for Bmal1 Cycling in Transcriptional Regulation of Intracellular Circadian Rhythms.” PLoS Genet 4 (2): e1000023. https://doi.org/10.1371/journal.pgen.1000023.

Liu, Chuanyu, Mingyue Wang, Xiaoyu Wei, Liang Wu, Jiangshan Xu, Xi Dai, Jun Xia, et al. 2019. “An ATAC-Seq Atlas of Chromatin Accessibility in Mouse Tissues.” Scientific Data 6 (1): 65. https://doi.org/10.1038/s41597-019-0071-0.

“The Human Protein Atlas.” n.d. Accessed March 28, 2021. https://www.proteinatlas.org/.

Local, Andrea, Hui Huang, Claudio P. Albuquerque, Namit Singh, Ah Young Lee, Wei Wang, Chaochen Wang, et al. 2018. “Identification of H3K4me1-Associated Proteins at Mammalian Enhancers.” Nature Genetics 50 (1): 73–82. https://doi.org/10.1038/s41588-017-0015-6.

Lv, Duo-Duo, Ling-Yun Zhou, and Hong Tang. 2021. “Hepatocyte Nuclear Factor 4α and Cancer-Related Cell Signaling Pathways: A Promising Insight into Cancer Treatment.” Experimental & Molecular Medicine 53 (1): 8–18. https://doi.org/10.1038/s12276-020-00551-1.

Marcheva, Biliana, Kathryn Moynihan Ramsey, Ethan D. Buhr, Yumiko Kobayashi, Hong Su, Caroline H. Ko, Ganka Ivanova, et al. 2010. “Disruption of the Clock Components CLOCK and BMAL1 Leads to Hypoinsulinaemia and Diabetes.” Nature 466 (7306): 627–31. https://doi.org/10.1038/nature09253.

Marstrand, Troels T., and John D. Storey. 2014. “Identifying and Mapping Cell-Type-Specific Chromatin Programming of Gene Expression.” Proceedings of the National Academy of Sciences 111 (6): E645–54. https://doi.org/10.1073/pnas.1312523111.

Mayran, Alexandre, and Jacques Drouin. 2018. “Pioneer Transcription Factors Shape the Epigenetic Landscape.” Journal of Biological Chemistry 293 (36): 13795–804. https://doi.org/10.1074/jbc.R117.001232.

Menet, Jerome S., Stefan Pescatore, and Michael Rosbash. 2014. “CLOCK:BMAL1 Is a Pioneer-like Transcription Factor.” Genes & Development 28 (1): 8–13. https://doi.org/10.1101/gad.228536.113.

Mohawk, Jennifer A., Carla B. Green, and Joseph S. Takahashi. 2012. “Central and Peripheral Circadian Clocks in Mammals.” Annual Review of Neuroscience 35 (1): 445–62. https://doi.org/10.1146/annurev-neuro-060909-153128.

Mure, Ludovic S., Hiep D. Le, Giorgia Benegiamo, Max W. Chang, Luis Rios, Ngalla Jillani, Maina Ngotho, et al. 2018. “Diurnal Transcriptome Atlas of a Primate across Major Neural and Peripheral Tissues.” Science 359 (6381). https://doi.org/10.1126/science.aao0318.

Nagy, P, H C Bisgaard, and S S Thorgeirsson. 1994. “Expression of Hepatic Transcription Factors during Liver Development and Oval Cell Differentiation.” The Journal of Cell Biology 126 (1): 223–33. https://doi.org/10.1083/jcb.126.1.223.

Nakamori, Daiki, Hiroki Akamine, Kazuo Takayama, Fuminori Sakurai, and Hiroyuki Mizuguchi. 2017. “Direct Conversion of Human Fibroblasts into Hepatocyte-like Cells by ATF5, PROX1, FOXA2, FOXA3, and HNF4A Transduction.” Scientific Reports 7 (1): 1–9. https://doi.org/10.1038/s41598-017-16856-7.

Ng, Natasha Hui Jin, Joanita Binte Jasmen, Chang Siang Lim, Hwee Hui Lau, Vidhya Gomathi Krishnan, Juned Kadiwala, Rohit N. Kulkarni, et al. 2019. “HNF4A Haploinsufficiency in MODY1 Abrogates Liver and Pancreas Differentiation from Patient-Derived Induced Pluripotent Stem Cells.” IScience 16 (June): 192–205. https://doi.org/10.1016/j.isci.2019.05.032.

Panda, Satchidananda, Marina P. Antoch, Brooke H. Miller, Andrew I. Su, Andrew B. Schook, Marty Straume, Peter G. Schultz, Steve A. Kay, Joseph S. Takahashi, and John B. Hogenesch. 2002. “Coordinated Transcription of Key Pathways in the Mouse by the Circadian Clock.” Cell 109 (3): 307–20. https://doi.org/10.1016/S0092-8674(02)00722-5.

Parviz, Fereshteh, Christine Matullo, Wendy D. Garrison, Laura Savatski, John W. Adamson, Gang Ning, Klaus H. Kaestner, Jennifer M. Rossi, Kenneth S. Zaret, and Stephen A. Duncan. 2003. “Hepatocyte Nuclear Factor 4α Controls the Development of a Hepatic Epithelium and Liver Morphogenesis.” Nature Genetics 34 (3): 292–96. https://doi.org/10.1038/ng1175.

Pearson, E. R., S. Pruhova, C. J. Tack, A. Johansen, H. A. J. Castleden, P. J. Lumb, A. S. Wierzbicki, et al. 2005. “Molecular Genetics and Phenotypic Characteristics of MODY Caused by Hepatocyte Nuclear Factor 4α Mutations in a Large European Collection.” Diabetologia 48 (5): 878–85. https://doi.org/10.1007/s00125-005-1738-y.

Perelis, Mark, Biliana Marcheva, Kathryn Moynihan Ramsey, Matthew J. Schipma, Alan L. Hutchison, Akihiko Taguchi, Clara Bien Peek, et al. 2015. “Pancreatic β Cell Enhancers Regulate Rhythmic Transcription of Genes Controlling Insulin Secretion.” Science 350 (6261). https://doi.org/10.1126/science.aac4250.

Pizarro, Angel, Katharina Hayer, Nicholas F. Lahens, and John B. Hogenesch. 2013. “CircaDB: A Database of Mammalian Circadian Gene Expression Profiles.” Nucleic Acids Research 41 (D1): D1009–13. https://doi.org/10.1093/nar/gks1161.

Qu, Meng, Tomas Duffy, Tsuyoshi Hirota, and Steve A. Kay. 2018. “Nuclear Receptor HNF4A Transrepresses CLOCK:BMAL1 and Modulates Tissue-Specific Circadian Networks.” Proceedings of the National Academy of Sciences 115 (52): E12305–12. https://doi.org/10.1073/pnas.1816411115.

Ruben, Marc D, Gang Wu, David F Smith, Robert E Schmidt, Lauren J Francey, Yin Yeng Lee, Ron C Anafi, and John B Hogenesch. 2018. “A Database of Tissue-Specific Rhythmically Expressed Human Genes Has Potential Applications in Circadian Medicine.” SCIENCE TRANSLATIONAL MEDICINE, 8.

Sekiya, Sayaka, and Atsushi Suzuki. 2011. “Direct Conversion of Mouse Fibroblasts to Hepatocyte-like Cells by Defined Factors.” Nature 475 (7356): 390–93. https://doi.org/10.1038/nature10263.

Sladek, F. M., W. M. Zhong, E. Lai, and J. E. Darnell. 1990. “Liver-Enriched Transcription Factor HNF-4 Is a Novel Member of the Steroid Hormone Receptor Superfamily.” Genes & Development 4 (12b): 2353–65. https://doi.org/10.1101/gad.4.12b.2353.

Storch, Kai-Florian, Ovidiu Lipan, Igor Leykin, N. Viswanathan, Fred C. Davis, Wing H. Wong, and Charles J. Weitz. 2002. “Extensive and Divergent Circadian Gene Expression in Liver and Heart.” Nature 417 (6884): 78–83. https://doi.org/10.1038/nature744.

Storch, Kai-Florian, Carlos Paz, James Signorovitch, Elio Raviola, Basil Pawlyk, Tiansen Li, and Charles J. Weitz. 2007. “Intrinsic Circadian Clock of the Mammalian Retina: Importance for Retinal Processing of Visual Information.” Cell 130 (4): 730–41. https://doi.org/10.1016/j.cell.2007.06.045.

Tahara, Yu, and Shigenobu Shibata. 2016. “Circadian Rhythms of Liver Physiology and Disease: Experimental and Clinical Evidence.” Nature Reviews Gastroenterology & Hepatology 13 (4): 217–26. https://doi.org/10.1038/nrgastro.2016.8.

Takahashi, Joseph S. 2017. “Transcriptional Architecture of the Mammalian Circadian Clock.” Nature Reviews Genetics 18 (3): 164–79. https://doi.org/10.1038/nrg.2016.150.

Thakur, Avinash, Jasper C. H. Wong, Evan Y. Wang, Jeremy Lotto, Donghwan Kim, Jung-Chien Cheng, Matthew Mingay, et al. 2019. “Hepatocyte Nuclear Factor 4-Alpha Is Essential for the Active Epigenetic State at Enhancers in Mouse Liver.” Hepatology 70 (4): 1360–76. https://doi.org/10.1002/hep.30631.

Trott, Alexandra J., and Jerome S. Menet. 2018. “Regulation of Circadian Clock Transcriptional Output by CLOCK:BMAL1.” PLOS Genetics 14 (1): e1007156. https://doi.org/10.1371/journal.pgen.1007156.

Vollmers, Christopher, Robert J. Schmitz, Jason Nathanson, Gene Yeo, Joseph R. Ecker, and Satchidananda Panda. 2012. “Circadian Oscillations of Protein-Coding and Regulatory RNAs in a Highly Dynamic Mammalian Liver Epigenome.” Cell Metabolism 16 (6): 833–45. https://doi.org/10.1016/j.cmet.2012.11.004.

Walesky, Chad, Genea Edwards, Prachi Borude, Sumedha Gunewardena, Maura O’Neil, Byunggil Yoo, and Udayan Apte. 2013. “Hepatocyte Nuclear Factor 4 Alpha Deletion Promotes Diethylnitrosamine-Induced Hepatocellular Carcinoma in Rodents.” Hepatology 57 (6): 2480–90. https://doi.org/10.1002/hep.26251.

Yadon, Adam N, Badri Nath Singh, Michael Hampsey, and Toshio Tsukiyama. 2013. “DNA Looping Facilitates Targeting of a Chromatin Remodeling Enzyme.” Molecular Cell 50 (1): 93–103. https://doi.org/10.1016/j.molcel.2013.02.005.

Yang, Guangrui, Lihong Chen, Gregory R. Grant, Georgios Paschos, Wen-Liang Song, Erik S. Musiek, Vivian Lee, et al. 2016. “Timing of Expression of the Core Clock Gene Bmal1 Influences Its Effects on Aging and Survival.” Science Translational Medicine 8 (324): 324ra16-324ra16. https://doi.org/10.1126/scitranslmed.aad3305.

Yin Liya, Ma Huiyan, Ge Xuemei, Edwards Peter A., and Zhang Yanqiao. 2011. “Hepatic Hepatocyte Nuclear Factor 4α Is Essential for Maintaining Triglyceride and Cholesterol Homeostasis.” Arteriosclerosis, Thrombosis, and Vascular Biology 31 (2): 328–36. https://doi.org/10.1161/ATVBAHA.110.217828.

Zhang, Ray, Nicholas F. Lahens, Heather I. Ballance, Michael E. Hughes, and John B. Hogenesch. 2014. “A Circadian Gene Expression Atlas in Mammals: Implications for Biology and Medicine.” Proceedings of the National Academy of Sciences 111 (45): 16219–24. https://doi.org/10.1073/pnas.1408886111.

Zhang, Yuxiang, Bin Fang, Manashree Damle, Dongyin Guan, Zhenghui Li, Yong Hoon Kim, Maureen Gannon, and Mitchell A. Lazar. 2016. “HNF6 and Rev-Erbα Integrate Hepatic Lipid Metabolism by Overlapping and Distinct Transcriptional Mechanisms.” Genes & Development 30 (14): 1636–44. https://doi.org/10.1101/gad.281972.116.

Zhang, Yuxiang, Bin Fang, Matthew J. Emmett, Manashree Damle, Zheng Sun, Dan Feng, Sean M. Armour, et al. 2015. “Discrete Functions of Nuclear Receptor Rev-Erbα Couple Metabolism to the Clock.” Science 348 (6242): 1488–92. https://doi.org/10.1126/science.aab3021.