UNIVERSITY OF CALIFORNIA
RIVERSIDE

Advanced Bioinformatics Tools and Quantitative Methods for Understanding Complex Traits Using Multi-Omic Data

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

Doctor of Philosophy

in

Genetics, Genomics and Bioinformatics

by

Ruidong Li

June 2018

Dissertation Committee:
Dr. Zhenyu Jia, Chairperson
Dr. Shizhong Xu
Dr. Shouwei Ding

The Dissertation of Ruidong Li is approved:

_____

_____

_____

Committee Chairperson

University of California, Riverside

Acknowledgements

I would like to take this chance to acknowledge all the helps I received during my PhD study.

First of all, I sincerely thank my advisor, Dr. Zhenyu Jia for your invaluable support, encouragement, and guidance on both research as well as on my career. You have been a tremendous mentor for me and I am very lucky to have worked with you in the most critical period of my academic life.

I would like to express my gratitude to Dr. Shizhong Xu and Dr. Shouwei Ding, who have been in my guidance, qualifying, and dissertation committees since I joined UCR. Both of you are great mentors and provide a lot of insightful suggestions on my projects.

I am also grateful to other members in my guidance and qualifying committees, Dr. Xuemei Chen, Dr. Thomas Girke, Dr. Ted Karginov, Dr. Yinsheng Wang, and especially thank Dr. Weifeng Gu for your guidance and training on RNA biology research.

I would like to thank the former and present directors of Genetics, Genomics, and Bioinformatics program, Dr. Shizhong Xu, Dr. Hailing Jin and Dr. Xuemei Chen for providing an ideal research environment to us. I thank Deidra Kornfeld and Julio Sosa for your constant help on all student affairs.

I greatly appreciate the collaboration and discussion with Dr. Mikeal Roose, Dr. Sergio Pietro Ferrante and Dr. Jinfeng Chen at UCR. Jinfeng is a great brother who helps me a lot not only on data analysis skills but also on my career development. Thanks for Dr. Jianbing Yan and Dr. Xiang Li in Huazhong Agricultural University for kindly providing the maize microspore dataset.

I thank all the members, former and present, in Dr. Jia's lab: Han Qu, Dr. Shibo Wang, Dr. John Chater, Dr. Julong Wei, Le Zhang, Dr. Jianming Lu and thank all the members in Dr. Xu's lab: Dr. Yanru Cui, Dr. Weibo Xie, Dr. Xuehai Hu, Dr. Tiantian Zhu, Meiyue Wang, Fangjie Xie, Chen Lin, and Sakar Sigdel. Thanks for your helps on the projects and I really enjoy the great collaboration, discussion and lab activities with you.

I would also like to thank my friends and fellow graduate students, especially thank Yike Ding, Duluo Nie, Yahui Li, Ruoying Lu, and Lichao Li for making my life in UCR memorable. With great thanks to my friends Yaocai Bai and Xue Di, whom I have known since high school. I am very lucky to meet you again at UCR and thanks a lot for your helps since always.

The text of chapter 2 in this dissertation, in part or in full, is a reprint of the material as it appears in "*GDCRNATools*: an R/Bioconductor package for integrative analysis of lncRNA, miRNA and mRNA data in GDC" (*Bioinformatics*, bty124-bty124, 2018). The major corresponding author, Dr. Zhenyu Jia directed and supervised the research.

Last, but not least, I would like to express my special thanks to my parents Yuyuan Li and Fuqin Liu, my grandma Xiuying Wang, my younger sister Na Li, my cute niece Xinyue Liu and all the family members for everything you have done for me. And thanks to Hongmei Ke for your love and support. We have known each other for 12 years since college and I want to experience the most beautiful journey with you for the rest of my life.

ABSTRACT OF THE DISSERTATION

Advanced Bioinformatics Tools and Quantitative Methods for Understanding Complex Traits Using Multi-Omic Data

by

Ruidong Li

Doctor of Philosophy, Graduate Program in Genetics, Genomics, and Bioinformatics
University of California, Riverside, June 2018
Dr. Zhenyu Jia, Chairperson

Most common diseases in humans are complex traits that are controlled by genetic variants in multiple genes and their interaction with environmental factors. The rapid evolution of high-throughput sequencing technology has led to tremendous increase in the volume of multi-dimensional omics data with dramatically reduced cost. Advanced bioinformatics tools and quantitative methods are urgently required to understand the molecular and genetic basis of human complex diseases including cancer to advance precision medicine.

In this dissertation, the first chapter introduces the most comprehensive cancer genomic data repository Genomic Data Commons (GDC), the genomic prediction models especially the Best Linear Unbiased Prediction (BLUP) method for common disease risk prediction, and the haplotype phasing methodologies. In the second chapter, a novel R

package is developed to download, organize and analyze RNA-seq and miRNA-seq data in GDC to decipher the lncRNA-mRNA related competing endogenous RNAs (ceRNAs) regulatory networks in cancer. In the third chapter, a BLUP-HAT method is proposed to test the hypotheses that the inclusion of a large number of genes selected from transcriptome and integration of other omic data will greatly improve the predictive power for cancer prognosis. In the fourth chapter, a haplotype phasing method is developed to infer high-resolution chromosome-scale haplotypes using genotype data of a few single gamete cells to facilitate genetic studies of complex traits.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The rapid advancement of high-throughput sequencing technology including the next-generation sequencing (NGS) and the third-generation sequencing (also known as long-read sequencing) has dramatically reduced the cost per genome from $100M to $1K (https://www.genome.gov/27541954/dna-sequencing-costs-data/). Taking advantage of the cost-effective sequencing technology, massive datasets have been generated including those from large collaboration projects such as The Cancer Genome Atlas (TCGA), the Genomics of Drug Sensitivity in Cancer (GDSC), the UK Biobank, and 1000 Genomes Project, etc. Novel bioinformatic and statistical tools are required to make effective use of the large-scale data to decipher the genetic and molecular basis of human common diseases and ultimately improve the prevention, diagnosis, and treatment in the era of precision medicine.

1.1 The Genomic Data Commons (GDC)

Cancer is a collection of related diseases including more than 100 types and many more subtypes caused by abnormal cell growth. It remains the second leading cause of death in US. Understanding the genomic and molecular changes that drive cancer will lay the foundation for precision medicine in oncology. TCGA is a pilot project supported by the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) that has generated 2.5 petabytes of comprehensive large-scale multi-dimensional omics data of more than 11,000 patients from 33 types of cancer. Most of the TCGA data is freely available to the public while data with unique information to an individual may be accessed with certification. TCGA data was initially collected, stored, and distributed in the Data Coordinating Center (DCC) and has moved to a newly launched data portal, the Genomic Data Commons (GDC) since July 15th, 2016.

GDC is a unified data sharing system that contains comprehensive cancer genomic datasets generated from NCI-supported programs such as TCGA and Therapeutically Applicable Research To Generate Effective Treatments (TARGET) (Grossman et al., 2016; Jensen et al., 2017). It also accepts high-quality genomic and clinical data from researchers who wish to share their data to the cancer research community, such as the Foundation Medicine Adult Cancer Clinical Dataset (FM-AD). All the data in GDC is harmonized using standardized pipelines so that they are accessible and reproducible to any researchers around the world. To date, multi-omics data including genomic, transcriptomic, epigenomic and proteomic data from 40 projects with 32,555 cancer cases has been generated and maintained in GDC. Moreover, many bioinformatic pipelines

have been developed by the GDC team for data processing and analysis, which have become invaluable resources for the research community.

A few web services including UCSC Xena Browser (https://xenabrowser.net/) (Goldman et al., 2017), Broad Institute GDAC Firehose (https://gdac.broadinstitute.org/), Oncomine (https://www.oncomine.org/) (Rhodes et al., 2004), and cBioPortal (http://www.cbioportal.org/) (Cerami et al., 2012; Gao et al., 2013), etc. as well as bioinformatics tools such as *TCGA-Assembler* (Zhu et al., 2014), *TCGA2STAT* (Wan et al., 2015), *TCGAbiolinks* (Colaprico et al., 2015), and *RTCGAToolbox* (Samur, 2014), etc. have also been developed to access, organize and analyze GDC/TCGA data. Many more bioinformatic and statistical tools are needed to decipher the genetic and molecular basis of cancer to advance precision medicine.

1.2 Genomic prediction

1.2.1 Genomic prediction for human complex traits

Many important human complex traits and common diseases have a polygenic nature that are controlled by many genetic variants with minor effects. Although an unprecedented number of genetic loci associated with human traits and disorders have been identified through genome-wide association studies (GWAS), only a small proportion of phenotypic variation can be explained, thus resulting in limited application in clinical practice. The great success of genomic prediction in plant and animal breeding stimulated the interest of predicting human traits and disease risk using genome-wide markers. Numerous studies have shown that inclusion of a large amount of markers with small

effect has the potential to capture all the genetic variance to improve the prediction accuracy. In the study of genetic architecture of human height, Allen et al. (2010) found that the explained proportion of phenotypic variation can be increased from 10.5% to 13.3% when associated SNPs at low-significance levels were used. Yang et al. (2010) estimated the genetic variance of human height explained by genome-wide SNPs using the restricted maximum likelihood (REML) in a linear model and found that common SNPs in total can explain 45% of genetic variance, which is much higher than the ~5% explained by a small number of significant and validated SNPs. Because the large amount of variance accounted for may not necessarily lead to improved predictive accuracy in predicting complex traits, Makowsky et al. (2011) then systematically assessed the relationship between proportion of explained genetic variance and predictive ability using Whole Genome Prediction (WGP) methods. The results indicated that much higher prediction accuracy can be achieved using WGP models than that only based on a small number of pre-selected significant SNPs.

Realizing the potential of genomic prediction in precision medicine, a few studies have been conducted to predict the risk of disease using WGP models. For example, Abraham et al. (2014) used L1-penalized support vector machine (SVM) methods to simultaneously model genome-wide SNPs to develop genomic risk scores (GRS) for celiac disease (CD) risk stratification and diagnosis. Validation studies showed that the GRS can generate robust and high predictive accuracy both within each of the six cohorts *via* cross-validation (AUC of 0.87-0.89) and in external validation datasets (AUC of 0.86-0.9). Abraham et al. (2016) have recently generated another GRS to predict the

lifetime coronary heart disease (CHD) risk using 49310 SNPs. The GRS was validated in five large cohorts and results indicated that it outperformed other models based on a small number of SNPs as well as the traditional clinical risk scores. Chen et al. (2014) performed a comprehensive evaluation of genomic risk prediction for Inflammatory bowel disease (IBD) in very large cohorts using two high density immunochip (each with 909,763 and 123,437 SNPs, respectively). They compared the prediction performance with different number of SNPs and varying sample sizes. The authors concluded in the study that the predictive power for IBD mainly benefits from those strongly associated SNPs with considerable effect sizes as well as increased sample size of training datasets.

1.2.2 Genomic prediction models

Conventional statistical methods usually cannot efficiently handle highly saturated models with $p \gg n$, where p is the number of parameters (selected markers) of the models and n is the sample size. To overcome this limitation, many whole-genome regression (WGR) models have been adopted to estimate genetic effects of thousands of genome-wide markers simultaneously for genomic prediction.

Linear models such as BLUP (Henderson, 1975), LASSO (Tibshirani, 1996), and BayesB (George and McCulloch, 1993; Hayes and Goddard, 2001) treat the effects of markers as random effects with different assumptions of effect size distribution. BLUP assumes a normal distribution with common variance across all the genomic variants. Details of BLUP are discussed in the next section. The LASSO regression method performs both variable section and regularization. It puts a constraint on the sum of the absolute value of the regression coefficients to force it to be less than a fixed value by

shrinking some coefficients to 0, thus only a subset of explanatory variables is included in the final model. The LASSO method can be reformulated as a Bayesian hierarchical model with a mixture of normal distributions for genetic effects and exponential distributions for the variances. The elastic net (ENET) is an extension of LASSO which overcomes its limitation that at most n variables can be selected where n is the number of observations (Zou and Hastie, 2005). Elastic net can also outperform LASSO on data with highly correlated variables because LASSO tends to select one variable and ignore the others. BayesB assumes that a large proportion of markers have no effect and the *a priori* distribution of effect size is a mixture of normal distributions of zero variance for some markers with probability π and a scaled inverse chi-square distribution of variances for the rest of markers with probability $1-\pi$. PLS is a dimension reduction methodology that combines features from principal component analysis (PCA) and multiple regression analysis. The original variables are transformed into latent components and the first few components which have the best predictive power are used as new predictors. Unlike principal component regression (PCR), the latent components in PLS are constructed by taking response variable into account. Support vector machines (SVMs) are supervised learning algorithms developed for classification and regression analysis, which can also be used for genomic prediction. Kernel functions such the (Gaussian) radial basis function kernel (SVM-RBF) and the polynomial kernel (SVM-POLY) are commonly used in SVM to make it computation-efficient in high-dimensional data.

Besides the general polygenic models mentioned above, a few extended and advanced algorithms have also been proposed. Zhou et al. (2013) developed a Bayesian sparse linear mixed model (BSLMM), which combines the advantages of both LMM and sparse regression models. Comparison of prediction performance for seven human diseases in the WTCCC datasets indicated that BSLMM always outperformed either of the two models. MultiBLUP is an extension of the BLUP model which accommodates multiple random effects with distinct effect-size variances for different SNP classes (*eg.* SNPs that are classified by functional annotations) (Speed and Balding, 2014). In a simulated human dataset of unrelated individuals, SNPs were divided into 5 distinct regions first with each contributing a predetermined heritability. The prediction accuracy for MultiBLUP was very similar to that for BLUP if the 5 regions contribute equally and was dramatically improved if the contribution to heritability is unequal in each region. The simulation study in a related mice dataset indicated that MultiBLUP always outperformed BLUP in all scenarios. In the human disease datasets with real phenotype, performance of the Adaptive MultiBLUP which can automatically detect genomic regions with different effect sizes was compared with BLUP, genetic risk scores, stepwise regression, and BSLMM. The results indicated that Adaptive MultiBLUP consistently achieved better prediction than other methods. A Bayesian non-parametric model named latent Dirichlet process regression (DPR) allowing for greater flexibility on the *a priori* effect size distribution was recently developed to adapt to a broad spectrum polygenic architecture of different complex traits (Zeng and Zhou, 2017). Compared with other commonly used models in the simulation datasets suggested that the Markov chain Monte Carlo (MCMC)

version of DPR performed robustly well in all scenarios although other methods may work better if the assumption of the model was satisfied. The comparison of DPR with other methods in predicting gene expression and phenotype in multiple real datasets generated very consistent results with that in the simulation dataset.

1.2.3 Mixed model

Mixed model is the most commonly used method in genomic prediction, which incorporates both fixed and random effects in a single regression model. It is generally described as:

$$y = X\beta + Z\gamma + \varepsilon$$

where y is a vector of the observed phenotypic values of n individuals; β is a $p \times 1$ vector of the non-genetic fixed-effects regression coefficients; X is an $n \times p$ design matrix for the fixed effects β; γ is a $q \times 1$ column vector of marker effects. Z is an $n \times q$ design matrix for the random effects γ; ε is an $n \times 1$ vector of residuals.

1.2.4 Best Linear Unbiased Prediction (BLUP)

BLUP is the most robust and well recognized method for estimating random genetic effects of a mixed model. The random effects are assumed to follow a normal distribution with $\gamma \sim N\left(0, G\sigma_\gamma^2\right)$ and residual errors $\varepsilon \sim N\left(0, R\sigma^2\right)$ distribution.

The expectation of the model is

$$E(y) = X\beta$$

and the variance-covariance matrix of model is

$$Var(y) = ZGZ^T\sigma_\gamma^2 + R\sigma^2$$

The variance components, $\theta = \{\sigma_\gamma^2, \sigma^2\}$, can be estimated using the restricted maximum likelihood (REML) method to maximize the log likelihood function

$$L(\theta) = \frac{1}{2}\ln|V| - \frac{1}{2}\ln|X^T V^{-1} X| - \frac{1}{2}\left(y - X\hat{\beta}\right)^T V^{-1}\left(y - X\hat{\beta}\right)$$

where $\hat{\beta} = \left(X^T V^{-1} X\right)^{-1}\left(X^T V^{-1} y\right)$.

The random and fixed effects are estimated from Henderson's mixed model equation,

$$\begin{bmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + G^{-1}/\lambda \end{bmatrix} \begin{bmatrix} \beta \\ \gamma \end{bmatrix} = \begin{bmatrix} X^T R^{-1} y \\ Z^T R^{-1} y \end{bmatrix},$$

where $\lambda = \sigma_\gamma^2 / \sigma^2$.

Both the BLUE (best linear unbiased estimation) for the fixed effects and the BLUP for the random effects can be computed via

$$\begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + G^{-1}/\lambda \end{bmatrix}^{-1} \begin{bmatrix} X^T R^{-1} y \\ Z^T R^{-1} y \end{bmatrix}$$

In genomic prediction, it is assumed that $\gamma \sim N\left(0, I\sigma_\gamma^2\right)$ and $\varepsilon \sim N\left(0, I\sigma^2\right)$ so that

$$Var(y) = ZZ^T\sigma_\gamma^2 + I\sigma^2$$

By defining

$$K = \frac{1}{a} ZZ^T$$

as the marker-inferred kinship matrix (Yu et al., 2006), where $a = \mathrm{tr}(ZZ^T)/n$. The variance can be rewritten as

$$Var(y) = K\sigma_A^2 + I\sigma^2$$

where $\sigma_A^2 = a\sigma_\gamma^2$ is called the polygenic variance.

In BLUP prediction, it's not necessary to estimate random effect of each individual marker, rather, the information from observed phenotypic values of individuals in the training set (y_1) can be used directly to predict the phenotypic values for individuals in the test set (y_2). Let ξ be the polygene, the model can be rewritten as,

$$y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1\beta \\ X_2\beta \end{bmatrix} + \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}$$

The variance-covariance matrix is partitioned as

$$Var(y) = Var\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix}\sigma_A^2 + \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}\sigma^2$$

The predicted phenotypic values can be calculated as the conditional expectation of y_2 given y_1, so

$$\hat{y}_2 = \mathrm{E}(y_2 \mid y_1) = X_2\hat{\beta} + \mathrm{cov}(y_2, y_1)\left[\mathrm{var}(y_1)\right]^{-1}(y_1 - X_1\hat{\beta}_1)$$

where $\mathrm{var}(y_1) = K_{11}\hat{\sigma}_A^2 + I\hat{\sigma}^2$ and $\mathrm{cov}(y_2, y_1) = K_{21}\hat{\sigma}_A^2$

1.2.5 Evaluation of prediction models

(1) Cross validation (CV)

CV is often adopted to assess the performance of a model by randomly partitioning the data into two parts: the training set which is used to estimate model parameters and the test set which is used for model evaluation. The advantage of CV is that the model development solely relies on the training dataset while the test dataset doesn't contribute to parameter estimation.

***k*-fold CV**. In a *k*-fold CV, the population is arbitrarily partitioned into *k* portions with approximately equal size. In each iteration, *k-1* portions are used as the training data to develop the model and the remaining one portion is used as the test data for model evaluation. This process is repeated for *k* times with each portion has been used for test exactly once. The predictability is calculated as the squared Pearson correlation coefficient between the observed and predicted values of all the individuals. Because the dataset can be split in many different ways, usually the entire *k*-fold CV is replicated for a few times to reduce the variation caused by random partitioning.

Leave-one-out cross-validation (LOOCV). Leave-one-out cross-validation is a special case of cross validation which uses an observation as the test set and the remaining observations as the training set. For a dataset with *n* individuals, LOOCV is exactly the *k*-fold CV when k equals to n. Although LOOCV can eliminate problems associated with random partitioning, it is more computationally expensive, which makes it infeasible for large samples.

(2) HAT method

To overcome the heavy computational burden of CV, the HAT algorithm is developed by Xu (2017) to evaluate the approximate predictability of a linear mixed model. With the HAT method, the predictability is calculated by

$$R^2_{HAT} = 1 - PRESS / SS$$

where $PRESS$ is the predicted residual error sum of squares, and SS is the total sum of squares for observed phenotypes adjusted by the fixed effects. Performance of the HAT method was evaluated using 7 agronomic and 1000 metabolomic traits in a Recombinant

Inbred Line (RIL) population of rice and results indicated that HAT method can generate very similar predictabilities as CV methods. The HAT method makes it possible to efficiently evaluate a lot of models to find the best ones.

1.3 Haplotype phasing

A haplotype is a set of DNA variants (or alleles) that tend to inherit together from a single parent. Knowing the haplotype information is very critical to the accurate interpretation of personal genomes for precision medicine such as assessment of the phase of potentially disease-causing mutations and prediction of the clinical outcomes or drug response, etc. (Crawford and Nickerson, 2005; De Bakker et al., 2006; Drysdale et al., 2000; Petersdorf et al., 2007). Haplotypes also have many applications in genetic studies including imputation of low-frequent variants (Huang et al., 2015b; McCarthy et al., 2016), characterization of genotype-phenotype relationship and genomic prediction of common disease risk (Musone et al., 2008), etc.

1.3.1 Applications of haplotype in genetic studies

(1) Genotype imputation

Phased haplotypes provide very good resources for genotype imputation in quantitative genetics and population genetics studies. For example, Huang et al. (2015b) reported that a haplotype reference panel consisting of 3,781 samples from the UK10K Cohorts project can significantly improve the imputation accuracy of low-frequency and rare variants in a UK and an Italian populations. A much larger reference panel of 64,976 human haplotypes at 39,235,157 SNPs from 20 studies were constructed and further increased

the genotype imputation accuracy even for variants at minor allele frequencies of 0.1% (McCarthy et al., 2016). The imputed genotypes can provide immediate benefits to improve the statistical power of genome-wide association studies.

(2) Genome-wide haplotype association studies (GWHAS)

During the past decade, GWAS have been ubiquitous and become the standard tool for gene discovery, by which statistical associations between genotypes and phenotypes are simultaneously tested on a large number of single nucleotide polymorphisms (SNPs) to find genes or genomic loci contributing to traits of interest, including important agronomic traits in crops and common diseases in human. Mounting evidences indicated the analysis of haplotypes that incorporates the grouping and interaction of several variants is superior to any individual SNP analysis technique. Compared with individual SNP-based association studies, the use of multi-allelic haplotypes has resulted in an increase of heritability, yielding improved power and robustness of association studies. Many specific haplotypes have been identified to be associated with a particular disease susceptibility or drug response. Trégouët et al. (2009) performed a GWHAS study using a sliding windows approach and identified the SLC22A3-LPAL2-LPA gene cluster as a strong susceptibility locus for coronary artery disease (CAD), which were not reported in previous conventional single marker-based GWAS analyses. Using the same strategy, Lambert et al. (2013) identified 91 regions with suggestive haplotype effects for Alzheimer's disease (AD). One of the haplotype associations can be replicated in all three datasets and the gene FRMD4A located in the locus were identified as a new genetic risk factor of AD.

(3) Genomic prediction

Strong linkage disequilibrium (LD) between genotyped markers and the causal variants is an essential determinant of the predictive power in genomic prediction. The utilization of haplotype information can capture more LD than individual markers to improve the predictive accuracy. A simulation study was performed to compare the accuracy of predicted breeding values with different methods of defining haplotypes (Calus et al., 2008). The results suggested that all the haplotype-based models performed better than the single-marker model and inclusion of linkage analysis information in the haplotype model would considerably increase the predictive accuracy for high-heritability trait. Edriss et al. (2013) genotyped 4429 Danish Holstein bulls with a 50K SNP chip to construct genealogy-based haplotypes for genomic prediction. Compared with single marker-based approach, slight increasing of predictive accuracy of direct genomic values (DGV) were observed for the haplotype-based prediction. Other traits such as protein yield and mastitis can achieve more gains in accuracy. The authors also pointed that improved methods for haplotype construction could further increase the predictive accuracy. Another genomic prediction study using LD-based haplotype in a 5214 Nordic Holstein bull population genotyped with a 54K bovine SNP chip showed that the haploblock approach can significantly improve prediction accuracy for all three traits (milk protein, fertility and mastitis) compared with SNP approach (Cuyabano et al., 2014). The most significant improvement in accuracy (up to 3.1%) was achieved for milk protein. Hess et al (2017) used a very large New Zealand dairy cattle population consisting of 58, 000 individuals genotyped with 37,740 SNPs to evaluate the effect of

fix-length haplotypes on genomic prediction accuracy. The results indicated that using short haplotypes (125 to 250 kb) can increase prediction accuracy compared with that using individual SNPs, whereas fitting covariates for long haplotypes (>500 kb) generated lower accuracy than single marker-based prediction.

1.3.2 Haplotype phasing methods

Although deep-sequencing of individual genome has made it easy to determine the genotypes for millions of single nucleotide polymorphisms (SNPs), the genotype data usually takes unphased format. It remains challenging to phase these molecular variants into specific haplotype for species with heterozygous genomes like human, thus limited the applications of haplotype in personalized medicine. Many strategies have been proposed for haplotype phasing, including population-based inference, whole-genome experimental phasing, molecular haplotype assembly, single somatic cell sequencing-based haplotyping, and haplotype phasing by gamete cell sequencing.

(1) Population-based inference

Most popular statistical approaches for phasing genetic variants from population data are based on coalescent methods and Hidden Markov Models (HMMs), including *PHASE* (Stephens and Scheet, 2005; Stephens et al., 2001), *fastPHASE* (Scheet and Stephens, 2006), *Beagle* (Browning and Browning, 2007), *MaCH* (Li et al., 2010), *IMPUTE2* (Howie et al., 2009), *SHAPEIT3* (O'Connell et al., 2016) and *Eagle2* (Loh et al., 2016). *PHASE* is a Bayesian method that utilizes a coalescent model to capture the fact that haplotypes tend to group into clusters of similar haplotypes over short genomic regions. Despite that it was the most accurate algorithm and was considered as a gold standard for

haplotype inference of unrelated individuals in the early years, *PHASE* is very time-cost and is not suitable for large scale genomic data and large populations. *fastPHASE* is more practicable to phase genome-wide variants with increased speed in the sacrifice of a little accuracy compared with *PHASE*. Browning et al. (2007) later proposed a new HMM-based haplotype phasing algorithm, *Beagle*, which uses localized haplotype-cluster model to resolve the issues of either slow computational speed or low accuracy in previous methods. The comparison in both simulated and real datasets indicated that *Beagle* was one or two orders of magnitude faster and was more accurate than *fastPHASE* and other previously existing methods. *IMPUTE2* and *MaCH* were originally developed for genotype imputation but also used similar approximate coalescent models as *PHASE* for haplotype phasing. A comprehensive evaluation of *Beagle*, *IMPUTE2*, and *MaCH* with default parameters indicated that *IMPUTE2* and *MaCH* had higher accuracy for small sample sizes and lower accuracy for large population compared with *Beagle*. The computing time of *MaCH* is an order of magnitude slower than *Beagle*, while *IMPUTE2* is moderate in speed.

SHAPEIT enhances the HMMs that are used in *IMPUTE2* and *MaCH* to speed up the phasing process with higher accuracy (Delaneau et al., 2012). *SHAPEIT* can handle trios, duos, and unrelated samples simultaneously. To further increase the phasing accuracy, *SHAPEIT2* was developed to incorporate information of genotype calls and base-quality scores in a probabilistic model (Delaneau et al., 2013a; Delaneau et al., 2013b). *SHAPEIT3* is an extension to SHAPEIT2, which significantly reduced the computational complexity from quadratic *O(N2)* to *O(NlogN)*, where *N* is the sample size. The

algorithm enhancements in *SHAPEIT3* makes it possible to deal with biobank-scale data consisting of hundreds of thousands of individuals such as UK Biobank. Generally, large sample sizes lead to higher accuracy for population-based phasing algorithms. However, the accuracy is limited in small cohorts for most of the statistical algorithms. *Eagle2* is a reference-based phasing algorithm which aims to improve phasing accuracy in small populations based on the positional Burrows-Wheeler transform (BWT) and a new search-based algorithm. The Haplotype Reference Consortium (HRC) panel comprising of 32, 470 samples is used as a reference in *Eagle2*. It achieves a ~20x speedup and ~10% improvement in accuracy compared with *SHAPEIT2*. The limitations of population-based phasing methods are that *de novo* mutations, rare variants or structural variants are not capable of being phased and only short haplotype blocks can be inferred.

(2) Whole-genome experimental phasing

Whole-genome experimental phasing methods usually depend on the physical separation of homologous chromosomes in a diploid cell followed by genotyping. Ma et al. (2010) adopted a strategy to microdissect metaphase chromosomes into several pieces and collectively phased 24,245 hetSNPs on 15 chromosomes from a HapMap individual. Yang et al. (2011) used the fluorescence-activated cell sorting (FACS) instrument to separate single chromosomes into 96-well plate based on the fluorescence patterns of Hoechst 33258 and Chromomycin A3 staining. The single chromosomes were then amplified and tagged for multiplex sequencing to construct chromosome-specific haplotypes. Fan et al. (2011) developed a microfluidic device to separate and amplify isolated chromosomes in a single metaphase cell. Three to four single-cell experiments

for each of the four individuals and 2~3 biological replicates for each homologous chromosome were performed. The haplotype phasing can achieve an accuracy of ~98% in the study. Although capable of accurately generate long-range haplotypes, experimental phasing is very expensive and labor-intensive, which is economically unfeasible in most studies.

(3) Molecular haplotype assembly

Peters et al. (2012) developed the long fragment read (LFR) technology, which is similar to single-molecule sequencing of long fragment to generate long-range haplotypes at low cost. This approach doesn't require the direct isolation of metaphase chromosomes and only 10-20 cells are enough to phase up to 97% of detected hetSNPs. HaploSeq is a proximity ligation-based approach (eg. Hi-C) that captures distal DNA fragments on a homolog for chromosome-spanning haplotype (Selvaraj et al., 2013). Short fragments in the Hi-C library can generate small local haplotype blocks and large fragments can be used to assemble the small blocks to a consensus chromosome-length haplotype. This strategy successfully phased ~95% of hetSNPs in mouse cells (30x sequencing coverage) with an accuracy of ~99.5% and ~81% of hetSNPs in human cells (17x sequencing coverage) with ~98% accuracy. A robust bioinformatics tool, *HAPCUT2*, was also developed to assemble haplotypes using data generated from diverse technologies, including fosmid-based dilution pool sequencing, PacBio single molecule real-time (SMRT) sequencing, 10X Genomics linked-read sequencing, and proximity ligation (Hi-C) sequencing (Edge et al., 2017).

(4) Single somatic cell sequencing-based haplotyping

Porubský et al (2016) used the single-cell DNA template strand sequencing (Strand-seq) technique to sequence either Watson or Crick strand of a chromosome in a somatic cell and pooled multiple Strand-seq libraries to phase chromosome-length haplotypes of a diploid individual. 183 libraries were used in their study and ~80% of the genotyped hetSNPs were phased with a concordance of 99.3% compared with the HapMap reference. Porubský et al (2017) further evaluated the combination of Strand-seq with various sequencing technologies for reconstructing dense haplotypes. The results suggested that using 10 Strand-seq libraries and 10x coverage PacBio long-read or 10X Genomics linked-read sequencing data can successfully phase more than 95% of the total number of hetSNPs.

(5) Haplotype phasing by gamete cell sequencing

Genotyping of single haploid cells from the heterozygous individuals for whole-genome haplotype phasing has many advantages over other phasing strategies because it exploits the haploid nature of gamete cells directly. Lu et al. (2012) performed whole genome amplification (WGA) using their newly developed multiple annealing and looping-based amplification cycles (MALBAC) method followed by Illumina HiSeq 2000 sequencing on 99 sperm cells from an Asian male donor to infer whole-genome haplotypes. 93 sperms were sequenced at ~1x genome depth and 6 sperms at ~5x depth. A two-stage algorithm was developed to phased hetSNPs into chromosome-length haplotypes. hetSNPs that were genotyped in > 40 sperms were used to generate draft haplotypes via counting the number of links between two adjacent markers first and other hetSNPs were

filled in to make high-resolution haplotypes in the second step. ~82% of the identified hetSNPs were successfully phased with high confidence. Kirkness et al. (2013) used the multiple displacement amplification (MDA) method to amplify genomic DNA of isolated sperm cells to infer haplotypes of the HuRef donor. 16 sperms were genotyped on an Illumina HumanOmni-Quad v1.0 BeadChip and 11 sperms were sequenced at 1.5x~3.7x depth on the Illumina GAIIx and HiSeq2000 platforms. They adopted a pairwise comparison strategy to infer crossovers in each gamete cell to assemble draft haplotypes using hetSNPs on the BeadChip and to phase the other hetSNPs genotyped by low-coverage sequencing to construct consensus haplotypes. 94% of hetSNPs identified in the HuRef genome was phased by the combination of BeadChip genotyping and low-coverage sequencing in the study. Hou et al. (2013) also performed MALBAC method for WGA on single human oocytes and inferred haplotypes using the haploid second polar body (PB2) cells. 4 to 14 cells were isolated from each of the 8 donors and were sequenced at ~1x genome depth. The authors used the two strategies proposed by Lu et al. and Kirkness et al. with some modifications (eg. introduce an HMM) to infer haplotypes of each individual. Both algorithms could phase ~95% of the hetSNPs with >95% of consistency. Despite that some efforts have been made to infer chromosomal haplotypes using gamete cells in the past few years, no user-friendly program is available.

1.4 Objectives of the dissertation

The objective of this dissertation is to develop advanced bioinformatics tools and quantitative methods for understanding complex traits using multi-omic data. In the

following chapters, I will introduce the tools and methods that we have developed in facilitating the study of complex traits. In chapter 2, an R package is developed to download, organize, and analyze RNA data in GDC with an emphasis on deciphering the lncRNA-mRNA related competing endogenous RNAs (ceRNAs) regulatory network in cancers. In chapter 3, by using transcriptomic, epigenomic, and miRNA data of prostate cancer in GDC, we developed a BLUP-HAT method to prove that the inclusion of a large number of genes selected from transcriptome and integration of other omic data will greatly improve the predictive power for cancer prognosis. In chapter 4, an algorithm implemented in our newly developed R package is proposed to infer high-resolution chromosome-length haplotypes using imperfect genomic data of single gametes to facilitate genetic studies of complex traits and advance precision medicine.

Chapter 2

GDCRNATools: an R/Bioconductor package for downloading, organizing, and integrative analyzing of data in GDC

The large-scale multi-dimensional omics data in the GDC provides opportunities to investigate the crosstalk among different classes of RNAs and their regulatory mechanisms in cancers. Easy-to-use bioinformatics pipelines are needed to facilitate such studies. In this study, we have developed a user-friendly R/Bioconductor package, named *GDCRNATools*, for downloading, organizing, and analyzing RNA data in GDC with an emphasis on deciphering the lncRNA-mRNA related competing endogenous RNAs (ceRNAs) regulatory network in cancers. Many widely used bioinformatics tools and databases are utilized in our package. Users can easily pack preferred downstream analysis pipelines or integrate their own pipelines into the workflow. Interactive *shiny*

web apps built in *GDCRNATools* greatly improve visualization of results from the analysis.

2.1 Introduction

Competing endogenous RNAs (ceRNAs) are RNA molecules that indirectly regulate other transcripts by competing for shared miRNAs. Although only a fraction of long non-coding RNAs has been functionally characterized, increasing evidences show that lncRNAs harboring multiple miRNA response elements (MREs) can act as ceRNAs to sequester miRNAs activity and thus reduce the inhibition of miRNAs on its targets. Deregulation of ceRNA networks may lead to human diseases. For example, long non-coding RNA HOTAIR was reported to play a critical role in cancer progression and metastasis. Liu et al. (2014) found that HOTAIR showed an oncogenic role in gastric pathogenesis by functioning as a ceRNA to regulate human epithelial growth factor receptor 2 (HER2) expression through competitively binding miR-331-3p. Another example is the imprinted oncofetal long non-coding RNA H19, which is actively involved in the tumorigenesis process and is expressed in many kinds of human cancers. Wang et al. (2016) found that H19 can regulate FOXM1 expression by sponging miR-342-3p to promote tumor development in gallbladder cancer (GBC). Although some lncRNA-associated ceRNAs have been identified to play critical roles in cancer, the regulatory significance of a large portion of ceRNAs remains to be unraveled.

The Genomic Data Commons (GDC) provides the cancer research community with a repository of standardized genomic and clinical data from National Cancer Institute (NCI)

programs including The Cancer Genome Atlas (TCGA) and Therapeutically Applicable Research To Generate Effective Treatments (TARGET). It also supports the import and standardization of high-quality genomic and clinical data from non-NCI supported cancer research programs, such as the Foundation Medicine Adult Cancer Clinical Dataset (FM-AD). The mission of GDC is to establish a cancer genomic data sharing platform that promotes precision medicine in oncology. To date, multi-omics data including genomic, transcriptomic, epigenomic, and proteomic data from 40 projects with 32,555 cancer cases has been generated and maintained in GDC.

Besides the raw sequencing and microarray data, many bioinformatics pipelines are developed by the GDC to generate derived data products such as somatic mutation, copy number segment, gene expression quantification, and methylation beta values, etc. All the data are harmonized against the latest human reference genome by the genomic harmonization pipelines. Data can be accessed via tools developed by the GDC, including the web-based GDC Data Portal, the command-line driven application GDC Data Transfer Tool, the UI version of the Data Transfer Tool for users prefer the graphical interface, and the GDC Application Programming Interface (API) which enables programmatic access to GDC functionality.

A few web services are available to access, organize and analyze GDC/TCGA data, such as UCSC Xena Browser, the Broad Institute GDAC Firehose, Oncomine, and cBioPortal. Some bioinformatic tools including *TCGA-Assembler*, *TCGA2STAT*, *TCGAbiolinks*, and *RTCGAToolbox*, etc. have also been developed. However, all the tools were initially developed for retrieving and analyzing data hosted in the TCGA Data

Coordinating Center (DCC) only, which uses the old human genome version GRCh37 (hg19) or GRCh36 (hg18) as reference. Tools such as *TCGA-Assembler* and *TCGAbiolinks* have been updated, thus are able to access the dynamic GDC data now. However, none of them provides a comprehensive workflow for systematic analysis of RNA-seq and miRNA-seq data in GDC.

In this study, a new R/Bioconductor package, named *GDCRNATools*, is developed for downloading, organizing, and integrative analyzing RNA data in GDC. By using *GDCRNATools*, data can be easily accessed and prepared for downstream analyses, including differential gene expression analysis, functional enrichment analysis, univariate survival analysis, and the ceRNAs network analysis. A newly developed algorithm *spongeScan* is used to predict miRNA response elements (MREs) in lncRNAs acting as ceRNAs (Furió-Tarí et al., 2016). In addition, databases including *starBase v2.0* (a collection of lncRNA-miRNA and mRNA-miRNA interactions that are predicted by 5 bioinformatics tools and experimentally validated in 108 CLIP-Seq datasets) (Li et al., 2013) , *miRcode* (a whole transcriptome human miRNA target predictions including > 10,000 lncRNAs) (Jeggari et al., 2012) and *miRTarBase* (accumulated 360,000 experimentally validated miRNA-target interactions by manually surverying pertinent literatures) (Chou et al., 2017) are also used as evidence basis for miRNA-mRNA and miRNA-lncRNA interactions in the package to identify ceRNAs in cancers. *GDCRNATools* allows users to easily perform the comprehensive analysis or integrate their own pipelines such as molecular subtype classification, weighted correlation

network analysis (WGCNA) (Langfelder and Horvath, 2008), and TF-miRNA co-regulatory network analysis, etc. into the workflow.

2.2 The *GDCRNATools* R package

The R package *GDCRNATools* consists of 3 modules: data download, data organization, and data analysis. Many easy-to-use functions are developed and many commonly used bioinformatics tools are integrated in the package.

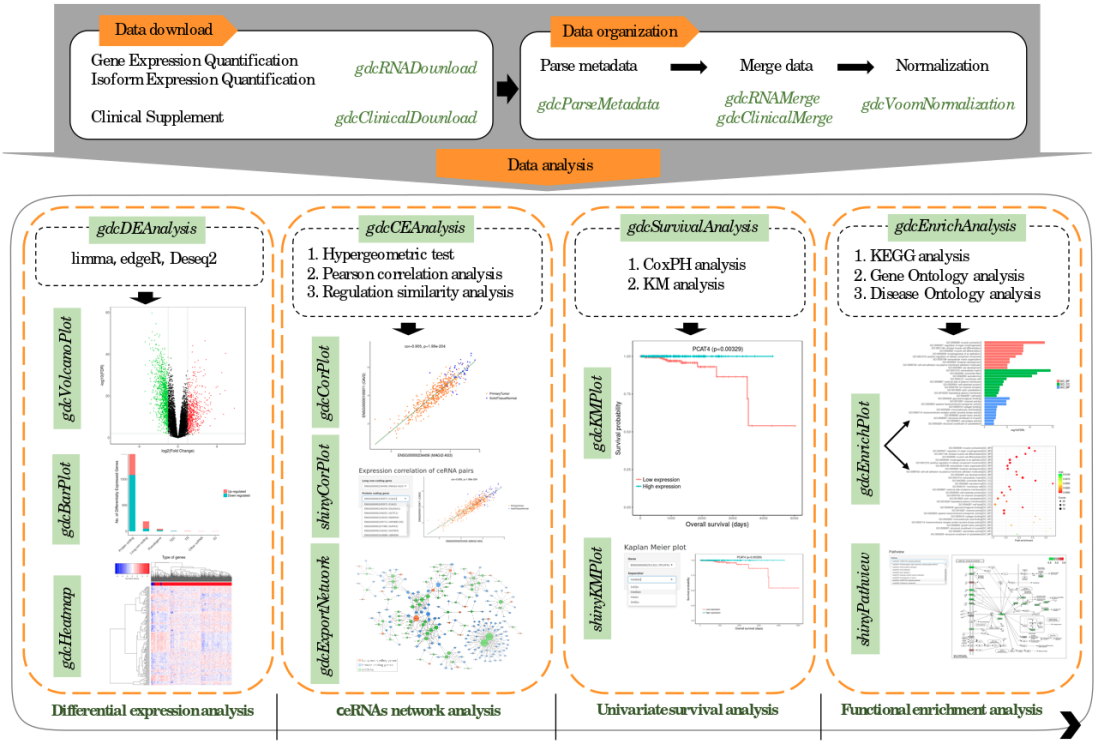


Figure 2.1: Workflow of *GDCRNATools*. The R package consists of three modules: data download, data organization, and data analysis

2.2.1 Data download module

Gene Expression Quantification (HTSeq-Counts), Isoform Expression Quantification (BCGSC miRNA Profiling), and Clinical (Clinical Supplement) data in GDC can be

easily downloaded by *gdcRNADownload* and *gdcClinicalDownload* functions, respectively. Two methods are provided in the functions. By default, data can be downloaded automatically through the GDC Application Programming Interface (API) by specifying data type and project id. An alternative method is also developed to download data using the GDC Data Transfer Tool *gdc-client* by manually providing a manifest file that is downloaded from the GDC cart.

Main functions:

- ***gdcRNADownload*** can download HTSeq-Counts data of RNA-seq and isoform quantification data of miRNA-seq into a separate folder for each sample
- ***gdcClinicalDownload*** can download clinical data in XML format into a separate folder for each patient

2.2.2 Data organization module

Data downloaded from GDC are in separate folders named by the corresponding Universal Unique Identifiers (UUIDs). To merge the data for downstream analysis, metadata associated with the downloaded files including file id, file name, sample id, sample type, as well as basic information of the patient (gender, age, tumor stage, tumor grade, days to death, days to last follow up, and vital status, etc.) are required to be retrieved. A series of functions are developed to parse the metadata, filter out samples, merge the clinical and RNA counts data, and normalize the RNA counts data in the data organization module.

Main functions:

- ***gdcParseMetadata*** parses metadata associated with downloaded files to facilitate downstream analysis.
- ***gdcFilterDuplicate*** and ***gdcFilterSampleType*** functions filter out duplicated samples from the same patient and samples that are neither Primary Tumor (code: TP) nor Solid Tissue Normal (code: NT) type.
- ***gdcRNAMerge*** merges total read counts for 5p and 3p strands of miRNAs in isoform quantification data and HTSeq read counts of gene quantification data to single expression matrices, respectively. Gene IDs are updated to the latest Ensembl 90 annotation of human genome, and unified mature miRNA IDs are updated based on the new release miRBase 21 (Kozomara and Griffiths-Jones, 2013).
- ***gdcClinicalMerge*** parses and merges clinical information of each patient stored in each single XML file. Either the complete table with all the information in the XML files or a well-organized table with the most important clinical information can be retrieved.
- ***gdcVoomNormalization*** is a function that normalizes gene/miRNA counts data based on the Trimmed Mean of M-values (TMM) method implemented in *edgeR* (Robinson et al., 2010) and further transforms normalized data by the voom method provided in *limma* (Law et al., 2014). Low expression genes (counts per million reads (CPM) < 1 in more than half of the total number of samples) are filtered out by default.

2.2.3 Data analysis module

The data analysis module in *GDCRNATools* provides many routine analysis methods for RNA-seq studies including differential gene expression analysis, survival analysis, and functional enrichment analysis. Most importantly, the availability of both RNA-seq and

miRNA-seq data from the same cohort makes it possible to study the lncRNA and mRNA associated ceRNA networks in cancer. The *gdcDEAnalysis*, *gdcSurvivalAnalysis*, *gdcEnrichAnalysis*, and *gdcCEAnalysis* functions can be easily implemented to perform the comprehensive analysis.

(1) Differential gene expression analysis

- ***gdcDEAnalysis*** is a convenient wrapper that can implement three most widely used methods: *limma*, *edgeR*, and *DESeq2* (Love et al., 2014) to identify differentially expressed genes (DEGs) between any two groups defined by users (eg., Primary Tumor vs. Solid Tissue Normal).
- ***gdcDEReport*** can report DEGs that are determined by the given threshold of fold change in log scale (logFC) and the False Discovery Rate (FDR) adjusted with Benjamini & Hochberg (BH) method. Gene id, official gene symbol and biotype of each DEG based on the Ensembl 90 genome annotation are also reported in the output.

(2) Survival analysis

- ***gdcSurvivalAnalysis*** can perform both Cox Proportional-Hazards (CoxPH) regression and Kaplan-Meier (KM) survival analyses to detect genes that are associated with overall survival (OS) or relapse free survival (RFS) of the patients. The hazard ratio, 95% confidence intervals, and p value for the tested genes are reported.

(3) Functional enrichment analysis

- ***gdcEnrichAnalysis*** performs Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) functional enrichment analyses using the latest databases through the R/Bioconductor package *clusterProfiler* (Yu et al., 2012). Disease Ontology analysis

using *DOSE* package (Yu et al., 2014) is also included in the *gdcEnrichAnalysis* function to detect gene-disease associations.

(4) ceRNAs network analysis

- ***gdcCEAnalysis*** uses three criteria to identify competing lncRNA-mRNA pairs: (1) the number and hypergeometric probability of shared miRNAs between a lncRNA and mRNA (2) the strength of positive expression correlation between the lncRNA and mRNA, and (3) the overall regulation similarity of all shared miRNAs on the lncRNA-mRNA pair.

To identify common miRNAs targeting both lncRNA and mRNA, three miRNA-mRNA interaction databases including *StarBase v2.0, miRcode,* and *mirTarBase 7.0,* as well as three miRNA-lncRNA interaction databases, including *StarBase v2.0*, *miRcode*, and *spongeScan* are incorporated and used in the *gdcCEAnalysis* function internally. Gene IDs in these databases are updated to the latest Ensembl 90 annotation of human genome, and unified mature miRNA IDs are updated based on the new release miRBase 21. *gdcCEAnalysis* also provides a portal *via* which the user-provided datasets of miRNA-mRNA and miRNA-lncRNA interactions (either predicted using other algorithms or validated through experiments) can be included and utilized for the ceRNAs regulatory network analysis.

Hypergeometric test is performed to test whether a lncRNA and mRNA share many miRNAs significantly.

$$p = 1 - \sum_{k=0}^{m} \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

where m is the number of shared miRNAs, N is the total number of miRNAs in the database, n is the number of miRNAs targeting the lncRNA, K is the number of miRNAs targeting the mRNA.

Pearson's correlation is calculated to measure the expression correlation between lncRNA and mRNA. Because miRNAs are negative regulators of gene expression, if more common miRNAs are occupied by a lncRNA, less of them will bind to the target mRNA, thus increasing the expression level of mRNA. As a result, expression of the lncRNA and mRNA in a ceRNA pair should be positively correlated.

The overall regulation similarity of all shared miRNAs on the lncRNA-mRNA pair is defined as:

$$Regulation\,similarity = 1 - \frac{1}{M}\sum_{k=0}^{M}\left[\frac{|corr(m_k, l) - corr(m_k, g)|}{|corr(m_k, l)| + |corr(m_k, g)|}\right]^M$$

where M is the total number of shared miRNAs, k is the k th shared miRNAs with $k = 1, \ldots, M$, and $corr(m_k, l)$ and $corr(m_k, g)$ represents the Pearson's correlation between the k th miRNA with lncRNA, and with mRNA, respectively.

gdcCEAnalysis can also compute sensitivity correlation (the difference between the Pearson's correlation and partial correlation coefficients) for each lncRNA-miRNA-mRNA triplet, defined as:

$$Sensitivity\,correlation = corr(l, g) - \frac{corr(l, g) - corr(m_k, l)corr(m_k, g)}{\sqrt{1 - corr(m_k, l)^2}\sqrt{1 - corr(m_k, g)^2}}$$

to measure the contribution of a miRNA in mediating the expression correlation between a lncRNA and mRNA (Paci et al., 2014), where $corr(l, g)$ is the Pearson's correlation between lncRNA and mRNA.

2.3 Case Study: TCGA-PRAD

Data from TCGA-PRAD project was used to demonstrate the usage of *GDCRNATools*. HTSeq-Counts data of RNA-seq and isoform quantification data of miRNA-seq were downloaded by *gdcRNADownload*. The associated metadata were parsed by *gdcParseMeta* function. A total of 52 normal (NT) and 495 prostate cancer (TP) samples were kept after filtering out duplicated samples as well as non-NT and non-TP samples. Raw counts of gene expression and mature miRNA expression were merged and normalized by *gdcRNAMerge* and *gdcVoomNormalization* functions, respectively. Finally, 15,524 high-expression genes were kept for downstream analysis after filtering out low expression ones.

2.3.1 DE analysis

DEGs between the 52 normal and 495 tumor samples were identified using the *limma* method in *gdcDEAnalysis* function. A total of 3,946 DEGs are determined with the absolute FC > 1.5 and FDR < 0.05. Among them, 3,391 are protein coding genes (1,218 up-regulated and 2,173 down-regulated in tumor) and 427 are lncRNAs (338 up-regulated and 338 down-regulated in tumor).

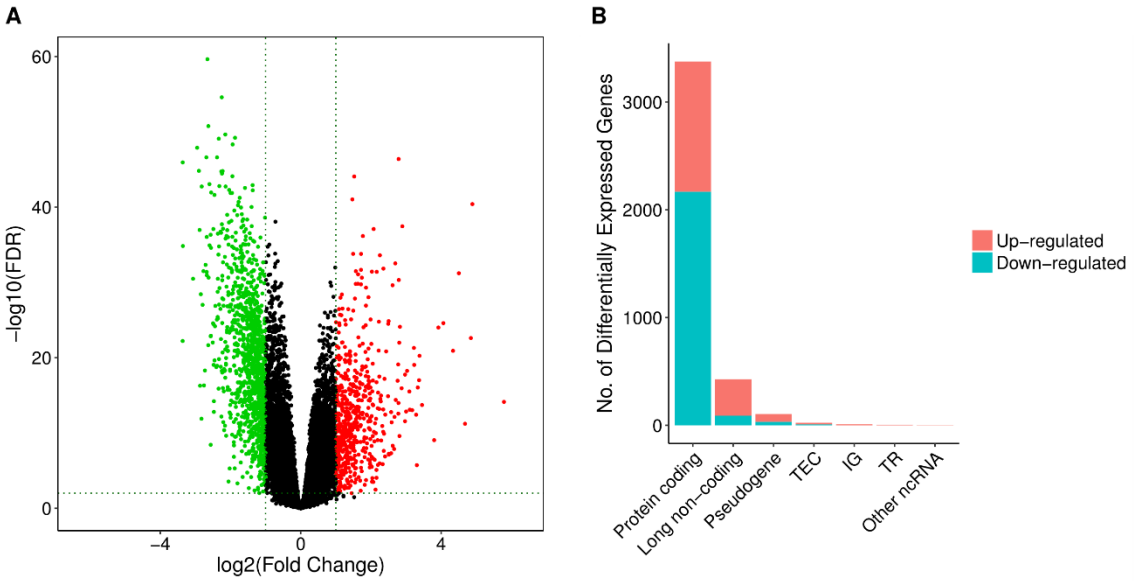


Figure 2.2: Differentially expressed genes between TP and NT samples. (A) Volcano plot (red: up-regulated in TP; green: down-regulated in TP) and (B) Bar plot of biotype for DEGs that are up- or down-regulated in TP

2.3.2 Functional enrichment analysis

Functional enrichment analysis including GO, DO, and KEGG analyses were performed on the list of DEGs using the *gdcEnrichAnalysi* function. A cutoff of FDR < 0.01 was used to determine the significantly enriched terms. Top 10 terms in each of the three ontology domains (BP: biological process; CC: cellular component; MF: molecular function) are shown in Figure 2.3.

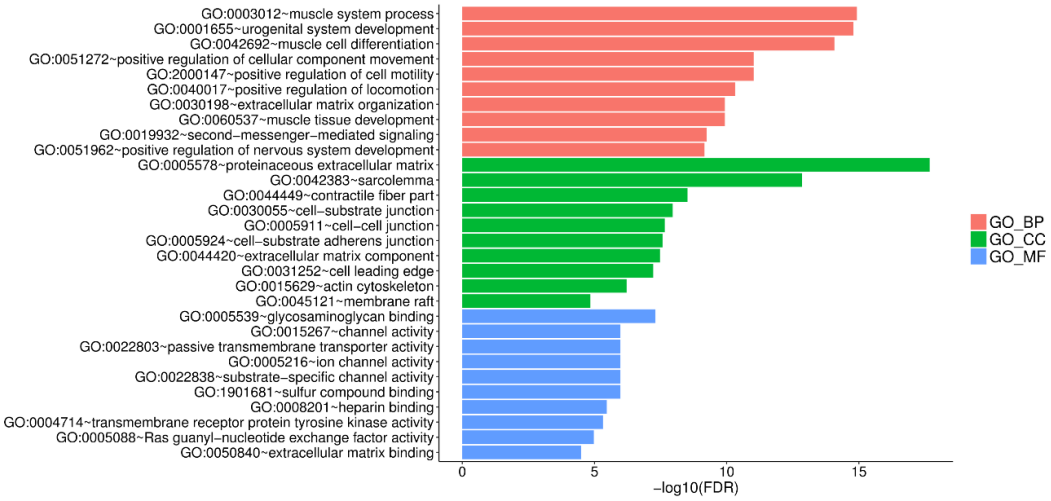


Figure 2.3: GO enrichment analysis. (Only the top 10 terms for each domain are shown)

DO enrichment analysis indicated that 65 disease ontology terms were enriched and the majority of them were related to human cancers. The DO term prostate cancer (DOID:10283) ranked the second in the enriched DO list with 145 DEGs were involved. Top 30 enriched DO terms are shown in Figure 2.4.

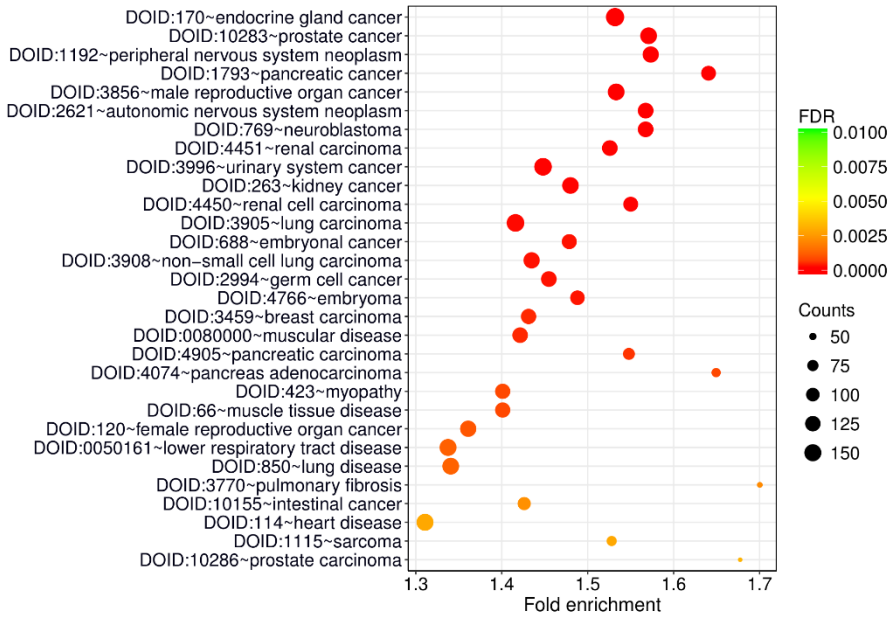


Figure 2.4: DO enrichment analysis. (Only the top 30 terms are shown)

The DEGs were significantly enriched in 30 KEGG pathways with many of which were involved in tumor initiation and progression (Figure 2.5).

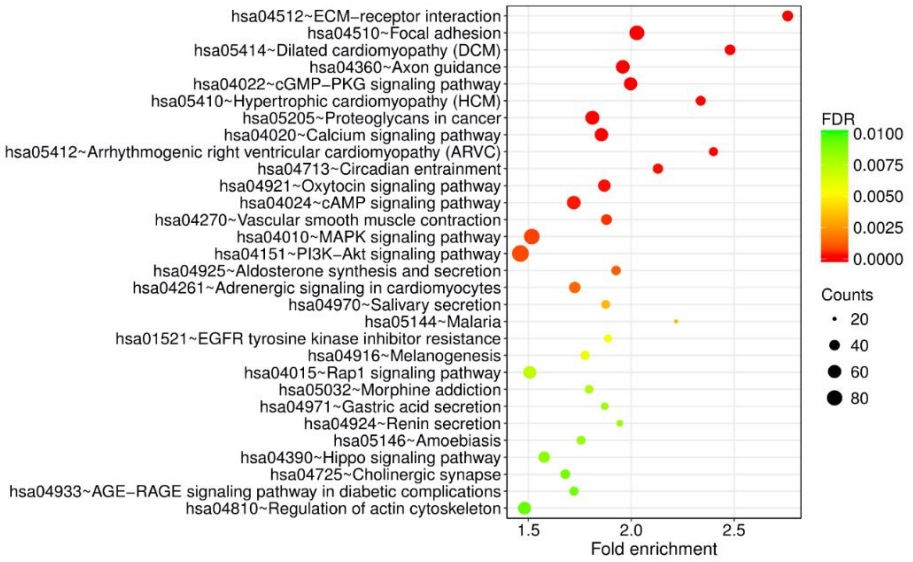


Figure 2.5: KEGG pathway enrichment analysis

The shiny app *shinyPathview* in *GDCRNATools* provides a very convenient tool for users to visualize the KEGG pathway maps on a local webpage (Figure 2.6).

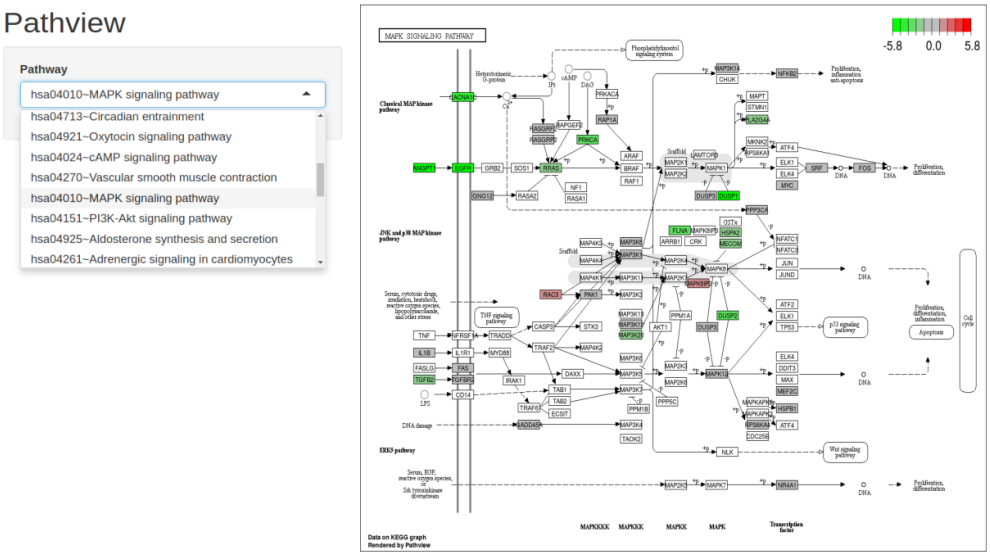


Figure 2.6: Visualization of enriched KEGG pathway maps on a local webpage

2.3.3 ceRNAs network analysis

gdcCEAnalysis was performed to identify lncRNAs that potentially function as ceRNAs in prostate cancer. DE lncRNAs and DE protein coding genes were used as input and *starBase V2.0* database was selected to provide the evidences of lncRNA-miRNA and mRNA-miRNA interactions. lncRNA-mRNA pairs that shared significant number of miRNAs (hypergeometric test $p < 0.05$), show significant positive correlations (Pearson's correlation test $p < 0.05$) and had non-zero regulation similarity scores mediated by shared miRNAs were determined for the ceRNAs regulatory network construction. A total of 838 lncRNA-miRNA-mRNA triplets were identified, which involved in 26 lncRNAs, 85 miRNAs, and 289 protein coding genes. Many of the 26 lncRNAs were reported to act as ceRNAs in cancers in previous studies. For example, Huang et al. (2017) found that NEAT1 can promote pancreatic cancer progression through sponging hsa-miR-506-3p. Sun et al. (2016) reported that NEAT1 can function as a ceRNA sponging hsa-miR-377-3p to activate the E2F3 pathway in non-small cell lung cancer (NSCLC) tumorigenesis and progression. Another NEAT1-associated ceRNA triplet NEAT1-hsa-miR-98-5p-CTR1 was also identified in NSCLC (Jiang et al., 2016). The interactions of NEAT1 with hsa-miR-506-3p, hsa-miR-377-3p, and hsa-miR-98-5p were all detected in our study.

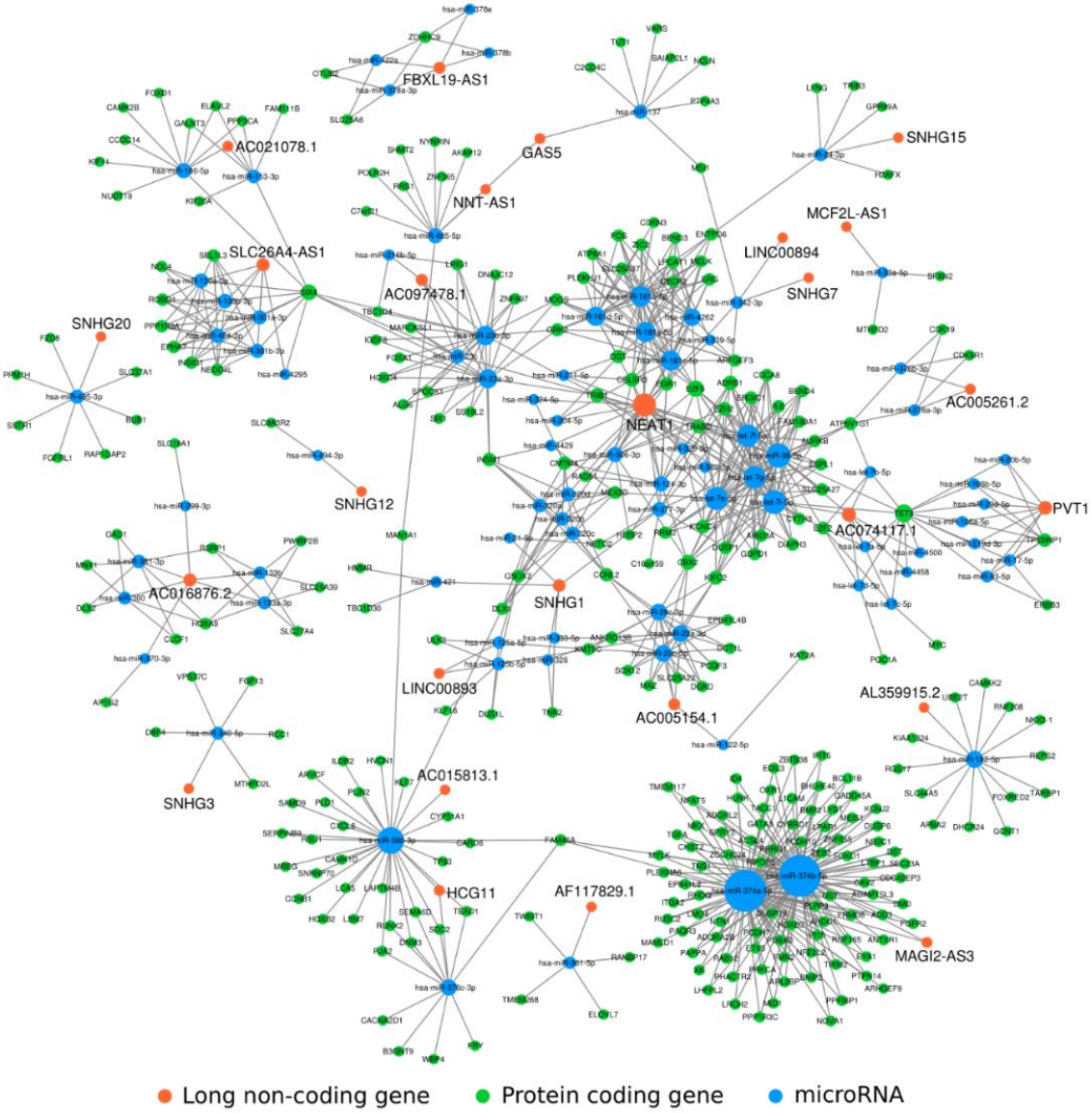


Figure 2.7: ceRNAs regulatory network. The size of each circle indicates the number of interactions

2.3.4 Survival analysis

A univariate CoxPH analysis using the *gdcSurvAnalysis* function showed that 15 out of the 26 lncRNAs in the ceRNAs network exhibited significant prognostic values ($p < 0.01$) on the RFS of prostate cancer. The *gdcSurvAnalysis* function was also performed to

compute the KM survival curves. Significant survival differences ($p < 0.01$) were detected between low-expression and high-expression groups of 10 lncRNAs that were divided by the median expression values, respectively. The prognosis roles of many of the lncRNAs were very consistent with previous studies. For example, it is reported that high expression of SNHG1 predicted poor prognosis in hepatocellular carcinoma (HCC) patients and was associated with a short biochemical recurrence-free survival time in prostate cancer (Zhang et al., 2016). Highly expressed lncRNA PVT1 showed poor prognosis in many type of cancers, such as cervical cancer, pancreatic cancer, nasopharyngeal cancer and gastric cancer, etc (He et al., 2018; Huang et al., 2015a; Iden et al., 2016; Kong et al., 2015). A meta-analysis of 11 cancer studies including 1,354 patients showed that elevated NEAT1 expression was significantly correlated with poor prognosis (Yang et al., 2017). In our study, the high expression levels of all the 3 lncRNAs were related to poor RFS.

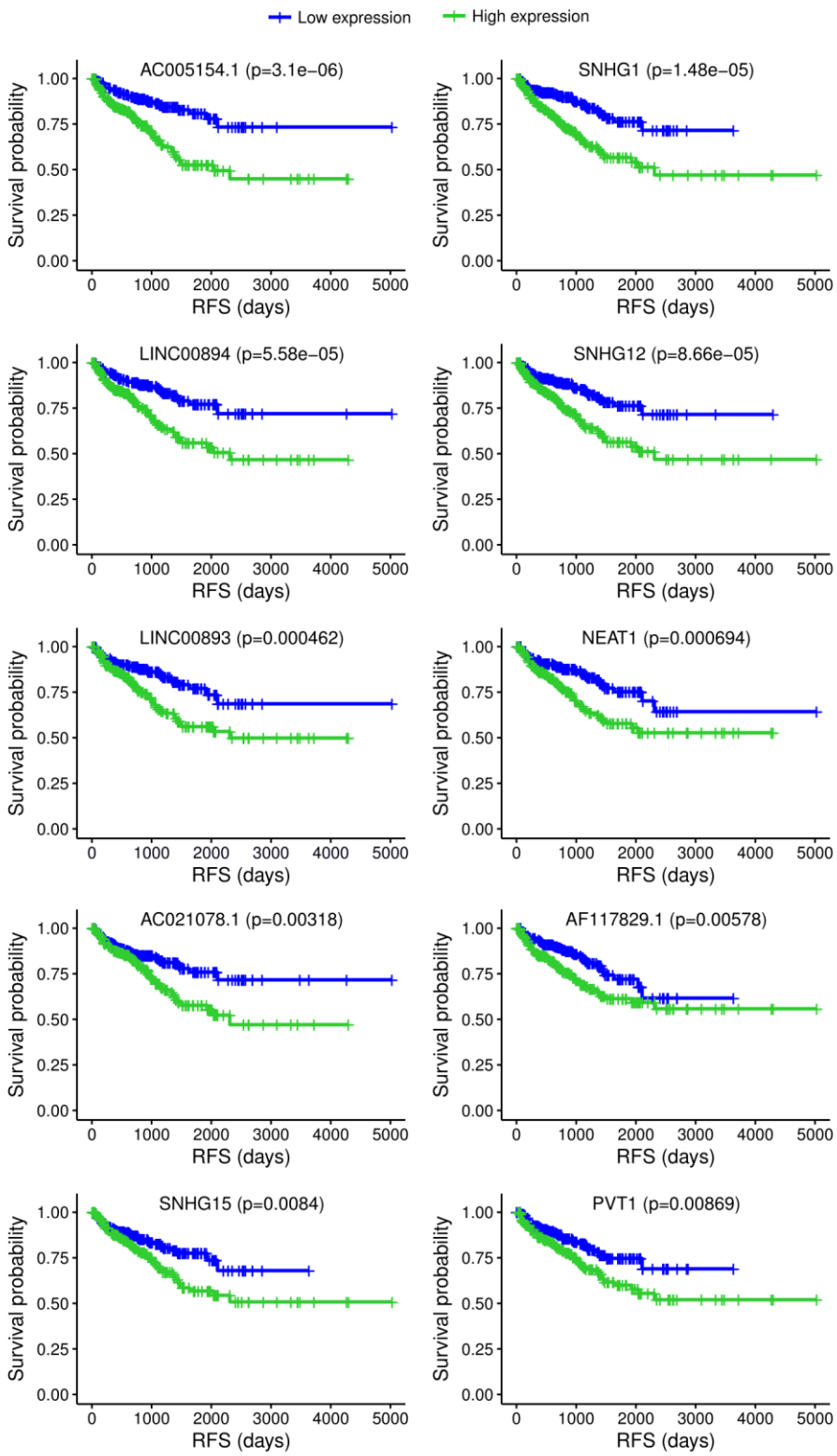


Figure 2.8: RFS survival analysis of lncRNAs in the ceRNAs regulatory network

2.4 Discussion

With the rapid growing of multi-dimensional omics data from cancer research programs, it is critical to provide the community a platform for storing, sharing, and standardizing such data to promote precision medicine in cancer. GDC is the largest and most comprehensive data repository, which contains standardized genomic and clinical data from both NCI-supported programs including TCGA and TARGET, and non-NCI generated data such as FM-AD. Comprehensive workflows are urgently required to exploit the data by integrating advanced bioinformatic or statistical methods. Although a few tools initially developed for retrieving and analyzing TCGA data hosted in DCC have been updated to access the up-to-date GDC data, none of them provides a comprehensive workflow for systematic analysis of RNA-seq and miRNA-seq data.

In this study, we have developed a novel R/Bioconductor package, named *GDCRNATools*, to access, organize, and integrative analyze RNA-seq and miRNA-seq data in GDC. Data can be easily downloaded and organized by a few functions in the package. Besides the routine analyses such as differential gene expression analysis, functional enrichment analysis, and univariate survival analysis, a function for ceRNAs network analysis is designed to identify lncRNA-miRNA-mRNA triplets that may play important roles in cancer. Many databases with predicted and/or experimentally validated miRNA-mRNA and miRNA-lncRNA interactions are incorporated and used as evidence basis. The ceRNA analysis function also provides a portal *via* which the user-provided datasets of miRNA-mRNA and miRNA-lncRNA interactions can be utilized for the ceRNAs regulatory network analysis. A case study of prostate cancer identified 26

lncRNA-related candidate ceRNAs with many of which have been reported in previous studies.

This easy-to-use package allows users with little coding experience to perform the entire analysis smoothly. As standardized data from other programs would be submitted to GDC, we believe that *GDCRNATools* will gain ground in cancer research for deciphering the crosstalk among multiple RNA species and their regulatory mechanisms, and will also greatly facilitate the exploit of multi-omics data using other advanced tools.

Chapter 3

BLUP-HAT method for accurately predicting outcomes of prostate cancer

Many molecular markers or gene expression signatures have been used for the prognosis of prostate cancer (PCa). The mediocre predictive ability for the current prognostic tests is mainly because of the limited number of genes used in the simple linear model, rather than incorporating a large number of genes into the whole-genome regression models. In this chapter, transcriptomic, miRNA, and methylomic data are used to evaluate the potential of improving prognosis accuracy by including a large number of genes selected from transcriptome and integrating multi-omic data in the BLUP-HAT model.

3.1 Introduction

PCa is the second most common cancer in men worldwide. An estimated 164,690 men will be newly diagnosed with PCa in United States in 2018 and 29,430 are predicted to die of PCa (Siegel et al., 2018). One major challenge in PCa is the accurate prediction of

tumor progression and clinical outcomes of diagnosed patients. PCa usually develops slowly that would probably never cause problems in most cases. Only a small portion of the patients harbor aggressive cancers that cause significant morbidity and mortality and need immediate treatments. The misclassification might lead to the overtreatment on patients with indolent cancers which cause potential side effects that impact their lives. Effective tests are urgently needed to advice regarding which patients harbor aggressive disease requiring radical treatment possibly followed by adjuvant therapy and which patients may be suitable for a more conservative active surveillance program.

Biochemical recurrence (BCR), which is defined as the prostate-specific antigen (PSA) value of at least 4.0 ng/mL followed by another increase after radical prostatectomy (RP), has been used as principal measure of clinical outcomes for PCa (Thompson et al., 2004). Although serum PSA have been utilized for over 20 years, it has serious limitations for the prognosis of PCa due to its lack of sensitivity and specificity (Filella and Giménez, 2013; Kretschmer and Tilki, 2017). Numerous nomograms have been created based on clinical variables, such as PSA level, biopsy Gleason score, tumor stage at the time of diagnosis, and percentage of positive biopsy cores to predict indolent PCa and clinical outcomes after the surgery (Chun et al., 2008; Kattan et al., 2003; Nakanishi et al., 2007; Steyerberg et al., 2007). But such prediction tools, to date, have provided limited power to distinguish aggressive prostate tumors from the indolent forms (Wang et al., 2014). Individual genetic biomarkers, including overexpression of prostate cancer antigen 3 (PCA3) or alpha-methylacyl-CoA racemase (AMACR), ETS gene fusions, and Glutathione S-transferase Pi 1 (GSTP1) hypermethylation, have been

employed for risk stratification of aggressive PCa (Ananthanarayanan et al., 2005; Cairns et al., 2001; Hessels and Schalken, 2009; Tomlins et al., 2005). To increase the predictive ability, several clinically applicable prognostic RNA expression signatures have been developed to calculate risk scores. For example, the OncotypeDX Genomic Prostate Score (GPS) from Genomic Health, Inc., which consists of 17 genes (12 genes in 4 biological pathways and 5 reference genes) can be calculated to predict adverse pathology at the time of radical prostatectomy (Klein et al., 2014). The 22-marker panel, Decipher, from GenomeDx biosciences Inc. was developed to predict systemic progression after PSA recurrence (Nakagawa et al., 2008). The Prolaris panel developed by Myriad Genetics Inc. is another gene expression signature assay that is based on 31 genes involved in cell cycle progression for cancer risk stratification (Cuzick et al., 2011). These multiple-gene tests have only provided a moderate improvement to classify tumor aggressiveness compared to nomograms. The gap between clinical practice and its objective needs to be filled.

During the past decade, the clinical tests for PCa prognosis have been constrained to contain a small number of genes mainly because of (1) the cost of tests and (2) convenience of the statistical algorithms for the development of prognostic models. Numerous studies indicated that using genome-wide markers as predictors yielded much higher predictability of complex traits than using major QTL only (Allen et al., 2010; Makowsky et al., 2011; Yang et al., 2010). The emergence of many cost-effective methods such as microarray and high-throughput sequencing make it feasible to apply transcriptome profiling to clinical use. A number of advanced methods, including BLUP

(Henderson, 1975; VanRaden, 2008), LASSO (Tibshirani, 1996), and BayesB (Meuwissen et al., 2001) have been proposed and applied to handle saturated linear regression models with $p \gg n$, where p is the number of parameters (selected genes) of the models and n is the sample size.

In this study, we compared 6 statistical models (BLUP, LASSO, PLS, BayesB, SVM-RBF, SVM-POLY) using 3 types of omic data including transcriptome (Tr), miRNAs (Mi), and methylome (Me) as well as their combinations (Tr+Mi, Tr+Me, Mi+Me, Tr+Mi+Me) for predicting 8 nomogram probabilities and 5-year RFS (RFS5YR). We have also developed BLUP-HAT method, an optimized version of BLUP, to substantially increase computational efficiency by avoiding arduous CV which used to be mandatory for model development. The BLUP-HAT method was then adopted to test our two hypotheses that (1) using a large number of genes selected from transcriptome to predict outcomes of PCa patients will outperform the clinically employed prognostic tests which only rely on several tens of major gene expression, and (2) the predictive power will be further increased if other omic predictors are also factored into the prognostic models.

3.2 Materials and Methods

3.2.1 TCGA data

HTSeq-Counts of RNA-seq, BCGSC miRNA Profiling of miRNA-seq, Beta value of Illumina Human Methylation 450 array, and clinical data of PCa patients from the TCGA-PRAD project were downloaded and processed by a series of functions in *GDCRNATools* package (Li et al., 2018). Low-expression genes/miRNAs with CPM < 1

in more than half of the total number of patients, and probes of methylation array with *NA* values in the cohort were filtered out. Some clinical traits that are not available in GDC such as pre-operative PSA values were retrieved from Broad GDAC Firehose (https://gdac.broadinstitute.org/). Clinical characteristics are summarized in Table 3.1.

Table 3.1: Clinical characteristics of the patients in TCGA-PRAD project

		Patients (N = 495)
Age at diagnosis (years)	$\leq$ 65	353
	> 65	142
Clinical tumor stage	T1a	1
	T1b	2
	T1c	172
	T2a	54
	T2b	54
	T2c	50
	T3a	36
	T3b	17
	T4	2
Gleason score	$\leq$ 6	45
	7 (3+4)	149
	7 (4+3)	98
	$\geq$ 8	203
Pre-operative PSA (ng/mL)	0-3.9	52
	4-9.9	273
	10-19.9	99
	$\geq$ 20	55

3.2.2 Pre-radical prostatectomy nomogram for PCa

Pre-radical prostatectomy nomogram that contains eight predictors have been developed by the Memorial Sloan-Kettering Cancer Center (https://www.mskcc.org/nomograms/) to predict the extent of the cancer and long-term results following radical prostatectomy. The following information was required for a patient: age (20 to 99), pre-operative PSA value (0.1 to 100), Gleason pattern and score (primary Gleason: pattern 1 to 5 and secondary Gleason: pattern 1 to 5), clinical tumor stage using the AJCC version 1/2010 staging system (T1a, T1b, T1c, T2a, T2b, T2c, T3a, T3b, T3c) and information on biopsy cores (optional). By entering the information, 4 primary treatment outcomes and 4 extent of disease probabilities were reported including:

4 primary treatment outcomes:

Probability of cancer-specific survival after radical prostatectomy in 10 years (OS10YR)

Probability of cancer-specific survival after radical prostatectomy in 15 years (OS15YR)

Progression-free probability after radical prostatectomy in 5 years (PFR5YR)

Progression-free probability after radical prostatectomy in 10 years (PFR10YR)

4 extent of disease probabilities:

Organ-confined disease (OCD)

Extracapsular extension (ECE)

Lymph node involvement (LNI)

Seminal vesicle invasion (SVI)

3.2.3 Methods of genomic prediction

Six prediction methods, including BLUP, LASSO, PLS, BayesB, SVM-RBF, and SVM-POLY were used in this study for comparison. The BLUP method was implemented using lab custom R script (Xu, 2013). LASSO, PLS, and BayesB were performed in the R packages *glmnet* (Friedman et al., 2010), *pls* (Wehrens and Mevik, 2007), and *BGLR* (Pérez and de Los Campos, 2014), respectively. The two SVM methods, SVM-RBF and SVM-POLY were implemented in the R *kernlab* package (Karatzoglou et al., 2004) using the radial basis function and the polynomial kernel function, respectively.

3.2.4 Evaluation of prediction methods

(1) 10-fold CV

To assess performances of the 6 predictive methods using 7 omic data combinations, 10-fold CVs were conducted. In a 10-fold CV, the population was randomly partitioned into 10 portions with equal size. In each iteration, 9 portions were used as training data to develop the model and the remaining 1 portion was used as test data for model evaluation. This process was repeated for 10 times with each portion had been used for test exactly once. After the CV, the predicted phenotype of each individual can be obtained and the predictability (squared Pearson's correlation coefficient between predicted and observed phenotypes) was calculated. We replicated the 10-fold CV analysis for 10 times.

(2) LOOCV

LOOCV is another strategy to measure the predictive ability of a model which consider one individual as the test set and the remaining individuals as the training set. Although

it's very time consuming, this method allows an efficient utilization of available data because only one sample is excluded for parameter estimation in each step. LOOCV was performed to obtain predicted RFS5YR values for models using different gene or miRNA expression datasets that were generated from BLUP-HAT analysis.

3.2.5 Commercial panels for PCa prognosis

(1) OncotypeDX GPS® (Genomic health Inc., Redwood City, USA)

OncotypeDX GPS consists of 17 genes (12 genes in 4 biological pathways and 5 reference genes). Expression of the 12 genes were all quantified in the TCGA dataset and were used for prediction (Table 3.2).

(2) Decipher® (GenomeDX Inc., Vancouver, Canada)

The Decipher is a 22-marker panel involving 19 genes because 2 markers may be derived from the same gene (eg., one in the coding region, and the other one in the intronic region). One of the 19 genes, Prostate Cancer Associated Transcript 32 (PCAT-32) doesn't have a stable id in the Ensembl genome annotation, so expression of 18 genes with stable Ensembl ids were used to represent this panel (Table 3.3).

(3) Prolaris® (Myriad Genetics Inc., Salt Lake City, US)

The Prolaris gene signature consists of 31 cell cycle genes and 15 house-keeping genes. All of the 31 genes can map to Ensembl gene ids in the TCGA gene expression dataset (Table 3.4). The 15 house-keeping genes were not included in the panel for prediction.

Table 3.2: Genes in the OncotypeDX panel

Gene symbol	Ensembl gene id	Biological pathways
AZGP1	ENSG00000160862	Androgen signaling
FAM13C	ENSG00000148541	Androgen signaling
KLK2	ENSG00000167751	Androgen signaling
SRD5A2	ENSG00000277893	Androgen signaling
FLNC	ENSG00000128591	Cellular organization
GSN	ENSG00000148180	Cellular organization
GSTM2	ENSG00000213366	Cellular organization
TPM2	ENSG00000198467	Cellular organization
BGN	ENSG00000182492	Stromal response
COL1A1	ENSG00000108821	Stromal response
SFRP4	ENSG00000106483	Stromal response
TPX2	ENSG00000088325	Cellular proliferation

Table 3.3: Genes in the Decipher panel

Gene symbol	Ensembl gene id	Marker type
LASP1	ENSG00000002834	Coding
IQGAP3	ENSG00000183856	3’ UTR
NFIB	ENSG00000147862	Intronic
S1PR4	ENSG00000125910	3’ UTR
THBS2	ENSG00000186340	3’ UTR
ANO7	ENSG00000146205	3’ UTR, Non-coding transcript
PCDH7	ENSG00000169851	Intronic
MYBPC1	ENSG00000196091	Coding, Intronic
EPPK1	ENSG00000261150	3’ UTR
TSBP	ENSG00000204296	Intronic
PBX1	ENSG00000185630	Coding
NUSAP1	ENSG00000137804	3’ UTR
ZWILCH	ENSG00000174442	3’ UTR
UBE2C	ENSG00000175063	3’ UTR, Coding antisense
CAMK2N1	ENSG00000162545	Coding antisense
RABGAP1	ENSG00000011454	Exon/intron junction antisense
PCAT-32	NA	Non-coding transcript
GLYATL1P4/PCAT-80	ENSG00000254399	Non-coding transcript
TNFRSF19	ENSG00000127863	Intronic

Table 3.4: Genes in the Prolaris panel

Gene symbol	Ensembl gene id	Gene symbol	Ensembl gene id
FOXM1	ENSG00000111206	TK1	ENSG00000167900
CDC20	ENSG00000117399	PBK	ENSG00000168078
CDKN3	ENSG00000100526	ASF1B	ENSG00000105011
CDC2	ENSG00000170312	C18orf24	ENSG00000154839
KIF11	ENSG00000138160	RAD54L	ENSG00000085999
KIAA0101	ENSG00000166803	PTTG1	ENSG00000164611
NUSAP1	ENSG00000137804	CDCA3	ENSG00000111665
CENPF	ENSG00000117724	MCM10	ENSG00000065328
ASPM	ENSG00000066279	PRC1	ENSG00000198901
BUB1B	ENSG00000156970	DTL	ENSG00000143476
RRM2	ENSG00000171848	CEP55	ENSG00000138180
DLGAP5	ENSG00000126787	RAD51	ENSG00000051180
BIRC5	ENSG00000089685	CENPM	ENSG00000100162
KIF20A	ENSG00000112984	CDCA8	ENSG00000134690
PLK1	ENSG00000166851	ORC6L	ENSG00000091651
TOP2A	ENSG00000131747		

3.3 Results

3.3.1 Prediction of 8 nomogram probabilities

The predictabilities of 8 PCa nomogram probabilities using 6 statistical models and 7 omic data combinations were evaluated via 10-fold CV. In the cohort, only 285 patients with all the clinical parameters for calculating nomogram and with all the omic data were used. The predictability of a trait was averaged across all methods and omic data combinations for comparison. The results indicated that predictabilities of different traits vary a lot with PFR5YR and PFR10YR having the highest predictabilities, whereas OS10YR and OS15YR having the lowest predictabilities. The 4 extent of disease probabilities (OCD, CEC, LNI, and SVI) had moderate predictabilities (Figure 3.1 upper panel). Comparing the three single omic data, transcriptomic prediction performed the best followed by the miRNA prediction, whereas the methylomic prediction was the worst. The combinations of multi-omic data did not seem to improve the predictability in general. The Tr+Mi model performed the best among the 4 multi-omic models, which produced very similar result as the Tr model. The Mi+Me model performed the worst, which was similar to the model using methylomic data only. Among the 6 statistical methods, overall, BLUP performed the best followed by BayesB and LASSO, whereas the SVM-RBF and SVM-POLY performed the worst. PLS was slightly better than SVM-RBF and SVM-POLY.

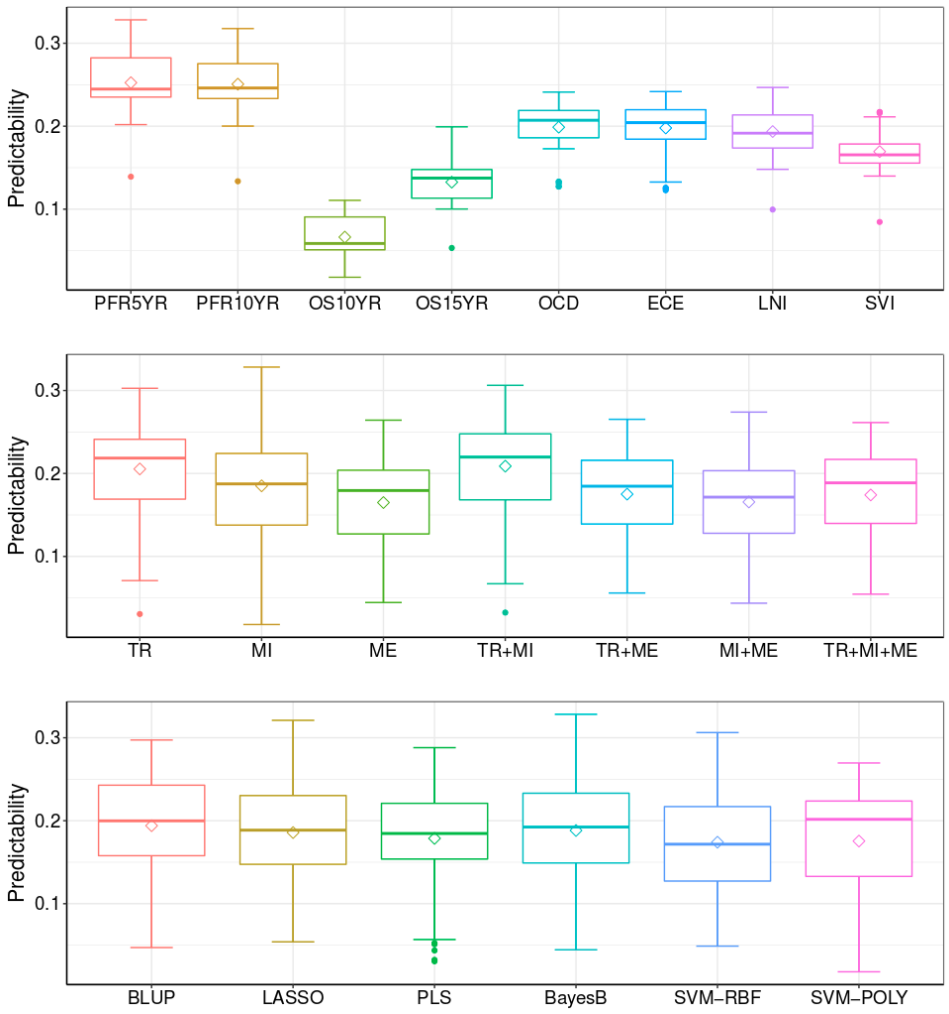


Figure 3.1: Comparison of predictabilities for different nomogram probabilities (upper panel), omic data (middle panel), and predictive methods (lower panel)

For a given trait, performances of the six statistical methods showed different pattern with different omic data (Figure 3.2). For example, in the prediction of PRF5YR, miRNA data was always the best predictor using LASSO, PLS, and BayesB, while Tr and Tr+Mi models showed similarly highest predictive power with the BLUP, SVM-RBF, and SVM-POLY methods. Methylome was the worst predictor in almost all the six methods except shat SVM-POLY with miRNA data worked much worse than using other omic data. The other 3 primary treatment outcomes had similar pattern as PFR5YR. In the

prediction of the 4 extent of disease probabilities, Tr and Tr+Mi were almost always the best omic models no matter which method was used. The only exception is that LASSO preferred miRNA data in predicting LNI.

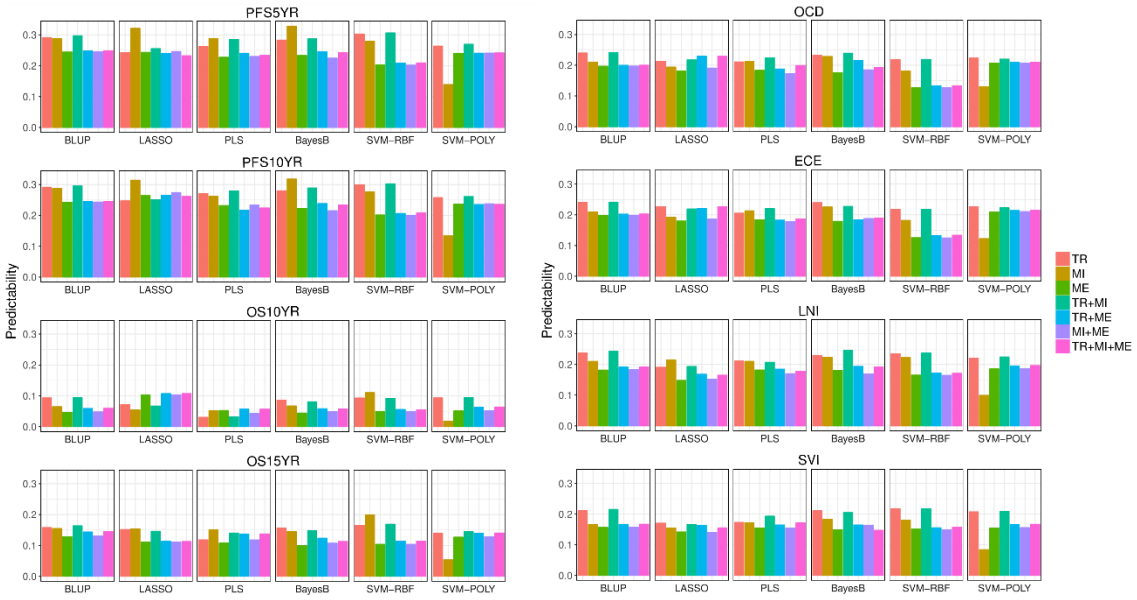


Figure 3.2: Barplot of predictabilities for different omic data and statistical methods in predicting each nomogram probability

3.3.2 Prediction of RFS5YR

To evaluate the predictability of RFS5YR rate directly, days to first BCR of each patient was transformed as follows: for a patient, if days to first BCR or days to last follow up without BCR of a patient is greater than 5*365 (1825), the phenotype value is determined as 1; otherwise, the phenotype value is calculated by dividing days to first BCR by 1825. A total of 153 patients were included with 93 of them underwent disease relapse in 5 years. Overall, the highest predictability that can be achieved using each method are as follows: BayesB with Tr+Mi (0.147), PLS with Tr (0.142), BLUP with Tr+Mi (0.141), LASSO with Mi+Me (0.139), SVM-RBF with Mi (0.133) and SVM-POLY with Tr+Mi

(0.133). For BLUP, BayesB, and SVM-POLY, transcriptome was the best predictor among the 3 single omic data and the combination of transcriptomic and miRNA data can future improve the predictive power for these three methods. For PLS, transcriptome was the best among all the 7 omic data combinations while for SVM-RBF, Tr, Mi, and Tr+Mi models performed similar and were much better than the other 4 omic data combinations. LASSO was very different from other methods with methylome itself and the combination of methylome with miRNAs were the two best predictors.

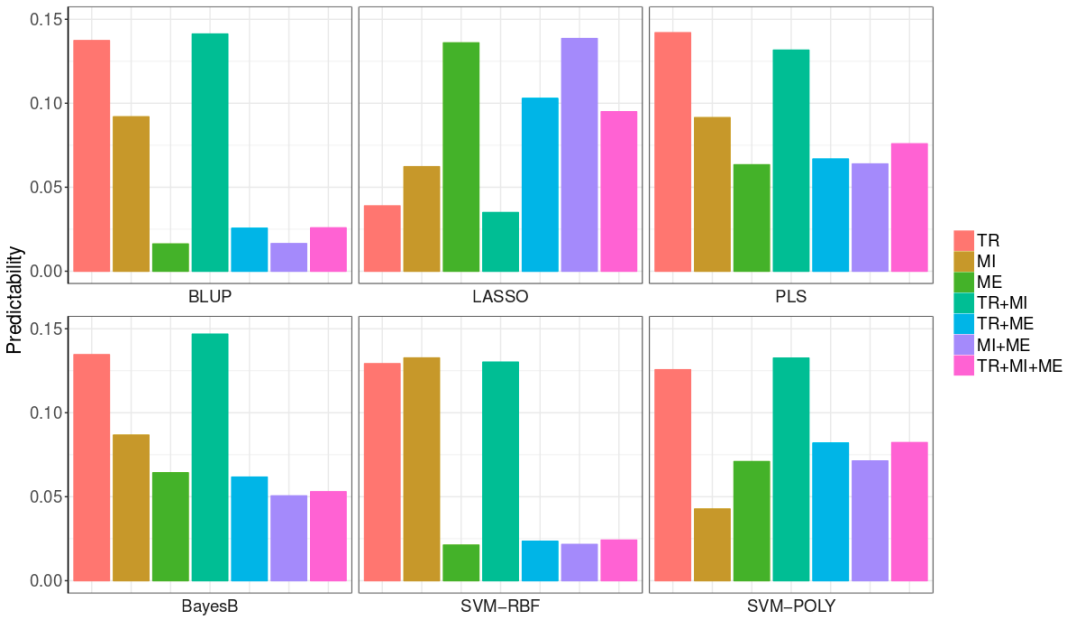


Figure 3.3: Barplot of predictabilities for different omic data and statistical methods in predicting RFS5YR

3.3.3 BLUP-HAT method for prediction of nomogram probabilities

Enlightened by the study that HAT method can be applied to mixed models as a very good approximation of the CV method and significantly improves the computational efficiency (Xu, 2017), a BLUP-HAT method was developed to test the two proposed

hypotheses that (1) using a large number of genes selected from transcriptome to predict the outcomes of PCa patients will outperform the clinically employed prognostic tests which only rely on several tens of major gene expression, and (2) the predictive power will be further increased if other omic predictors are also factored into the prognostic models.

First, transcriptomic data was used to evaluate if a prediction model using a large number of genes selected from transcriptome will outcompete the models with only a few tens of highly ranked genes. For each trait, genes were sorted in descending order according to the absolute Pearson's correlation coefficient between the gene expression values and phenotypes. Top N correlated genes with high-expression (N ranges from 5 to 15536) were sequentially added to the model one by one and HAT value for each scenario was calculated. Three commercially available prognosis panels Oncotype, Decipher, and Prolaris consisting of 12, 18, and 31 genes with stable Ensembl gene ids, respectively, were also compared.

The results indicated that predictabilities of all the 8 nomogram probabilities can be significantly improved by including hundreds of correlated genes in the BLUP model, rather than using only a few significant genes (Figure 3.4). It was interesting to notice that by adding more genes to some extent, the predictability will be reduced, resulting an obvious peak of predictability for each trait. The minimum and maximum number of genes for the corresponding peaks were 143 for OS15YR and 1248 for OCD, respectively.

For the 6 traits PFR5YR, PFR10YR, OCD, ECE, LNI, and SVI, all the models were better than the three commercial panels no matter how many genes were included. For

the other two traits OS10YR and OS15YR, models with too few or too many genes may be a little worse than one of the three commercial panels. For example, Prolaris performed better than models with less than 47 or more than 10,976 genes for OS10YR.

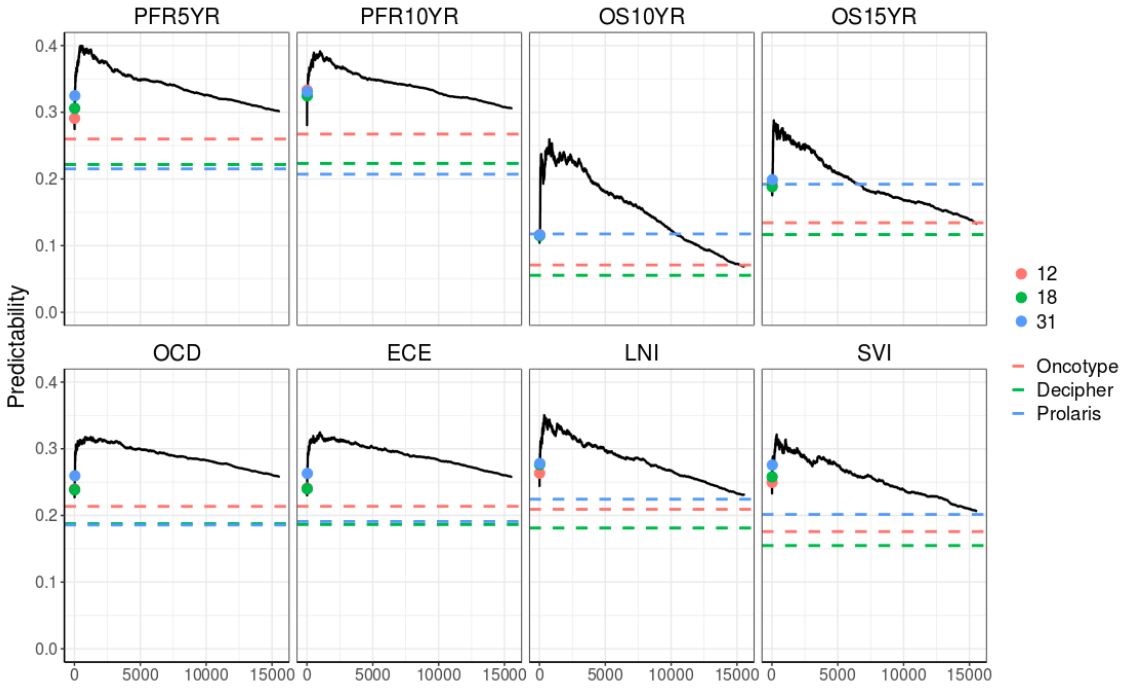


Figure 3.4: Evaluation of models using different number of genes selected from the transcriptome in predicting 8 nomogram probabilities by the BLUP-HAT method

To evaluate if the predictability can be further improved by integrating panels from other omic data, BLUP-HAT was performed to detect the top N correlated miRNAs that corresponded to the highest HAT values. We compared averaged predictabilities for gene expression panels, miRNA expression panels, and the combination of gene with miRNA expression panels, respectively, across the 8 nomogram probabilities. The result indicated that gene expression panel still performed better than miRNA expression panel, and the integration of gene and miRNA expression panels can significantly improve the predictability (Figure 3.5).

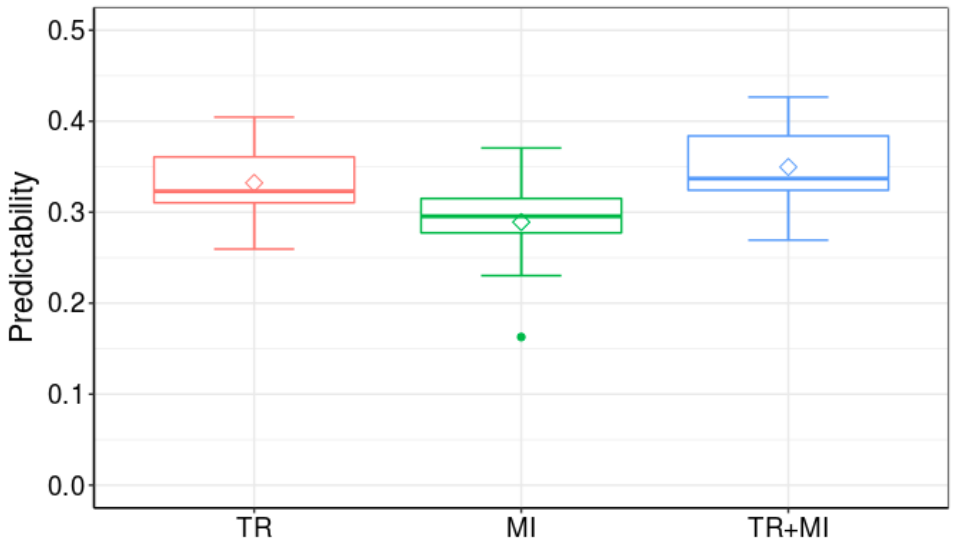


Figure 3.5: The performances of different expression panels in predicting the 8 nomogram probabilities

3.3.4 Development of a multi-omic signature for RFS5YR prediction

We further adopted the BLUP-HAT method to generate a multi-omic signature for RFS5YR prediction. It is indicated that by using the BLUP method, the top 359 genes (GENE359) and top 61 miRNAs (MIR61) can achieve the highest HAT values of 0.347 and 0.270, respectively. The predictabilities of the three commercial panels are all around 0.1 (Figure 3.6).

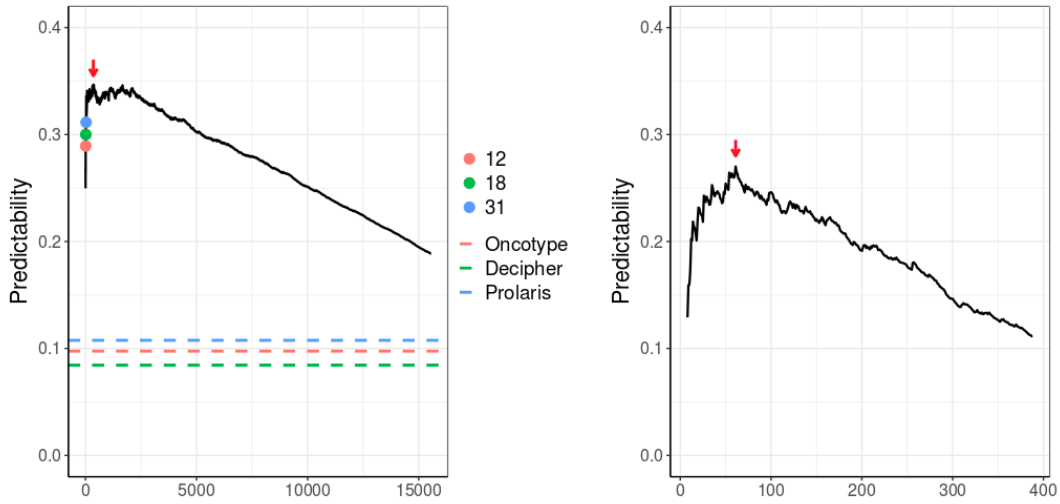


Figure 3.6: Identification of the gene and miRNA expression signatures that have the highest HAT values in predicting RFS5YR

We used LOOCV to get the predicted values of RFS5YR by using different classifiers including the three commercial panels Oncotype, Decipher, and Prolaris, the three omic data Tr, Mi, and Tr+Mi, as well as GENE359, MIR61, and GENE359+MIR61. The Receiver operating characteristic (ROC) curve of each model is shown in Figure 3.7. The area under the curve (AUC) for the three commercial panels were all below 0.7, with the highest was 0.685 for Oncotype and the lowest was 0.614 for Prolaris. The use of whole transcriptomic data and the transcriptomic plus miRNA data performed better than the three commercial panels, which had similar AUC of 0.732, while using the entire miRNA dataset did not perform good (AUC=0.646). The GENE359 panel can significantly improve the predictive power with an AUC of 0.809 and the integration of the GENE359 and MIR61 panels further improved the prediction accuracy (AUC=0.821).

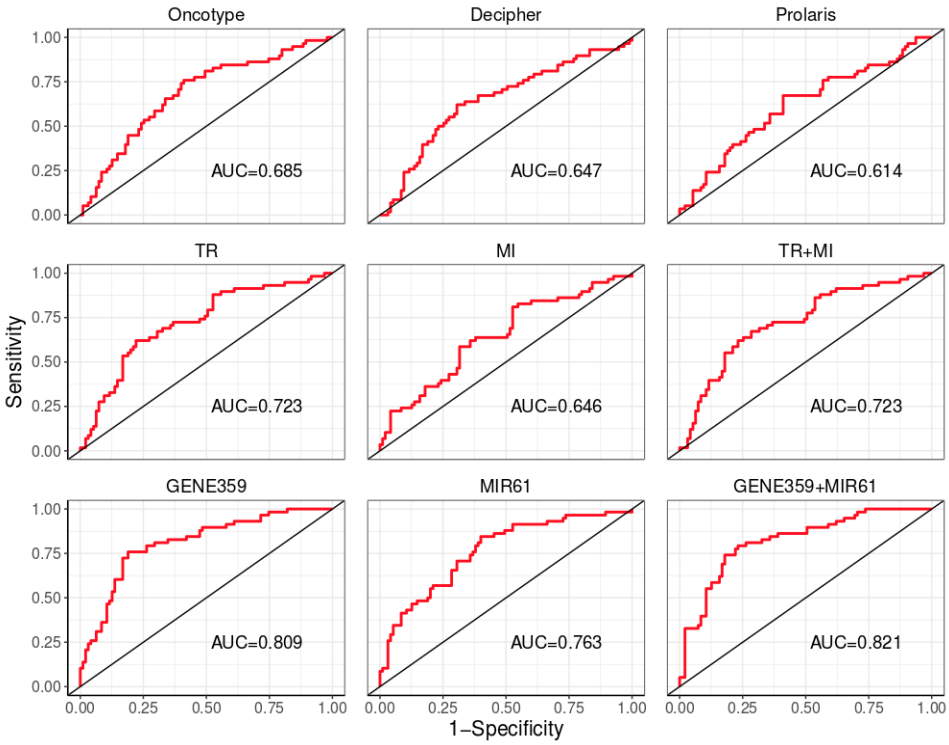


Figure 3.7: ROC curves for models using different gene/miRNA expression dataset

3.4 Discussion

Despite a few gene signatures have been developed to distinguish aggressive prostate tumors from the indolent forms in the past decade, they only achieve mediocre predictive ability. We proposed a hypothesis that the power for the current prognostic tests is constrained by the limited number of genes included in the models. To test the hypothesis, we first compared performances of six statistical methods using 3 omic data as well as their combinations. We found that BLUP was overall the most robust method and transcriptomic prediction was more accurate than miRNA and methylomic predictions. The combination of multi-omic data did not improve the predictive ability in general. The efficient BLUP-HAT method was then developed to evaluate the predictive power of

models using expression data of the top N genes that were correlated with the clinical traits. The results proved our first hypothesis that models using a large number of genes selected from transcriptome outperform current commercial panels with several tens of genes in predicting outcomes of PCa patients. Integration of the top correlated genes and miRNAs corresponding to the peak predictabilities, respectively, can further improve the predictive accuracy, just as anticipated in the second hypothesis.

In this study, we also adopted the BLUP-HAT method to develop gene and miRNA signatures to predict the risk of RFS5YR. ROC analysis indicated that the identified GENE359 signature outperformed all the three commercial panels and the predictive accuracy can be further improved by the combination of GENE359 with MIR61 panels.

The positive results of the study not only proved the concept that inclusion of transcriptomic data of a large number of small effect genes and integration of other omic data can significantly improve the predictability of PCa outcomes, but also provided a promising strategy to guide the development of new practice standard using more accurate biomarker-based diagnosis and prognosis of PCa.

Chapter 4

Inference of chromosome-length haplotypes using genomic data of three to five single gametes

Knowledge of chromosome-length haplotypes will not only advance our understanding of the relationship between DNA and phenotypes, but also promote a variety of important genetic applications. The current diploid-based phasing methods are either costly or only produce haplotype fragments, whereas, the alternatives based on analysis of haploid gametes, which are still in their early development stage, are computationally challenging and error-prone. In the study, we developed an innovative method, named *Hapi*, for a fully-automatic inference of chromosomal haplotypes for individual diploid genome using only 3 to 5 gametes. Analyses of simulated data and real gamete datasets showed that *Hapi* outperformed the other two haploid-based methods in terms of accuracy, reliability, and cost efficiency. This highly cost-effective phasing method will increase power of widely employed genome-wide association studies (GWASs) by revealing

novel haplotype variants that are entirely undetectable by conventional GWAS, and facilitate human disease studies as well as animal and plant breeding. Moreover, *Hapi* can detect meiotic recombination events in gametes, which has promise for adoption in the public health sector including the diagnosis of abnormal recombination activity in human reproductive cells to aid in reducing infant mortality, birth defects, and miscarriages.

4.1 Introduction

A haplotype in a diploid individual is a set of DNA variants (or alleles) on a chromosome that are co-inherited from a parent. Knowledge of parental haplotypes is critical to advance our understanding of the relationship between DNA and phenotypes, and promote a variety of genetic applications. For example, in precision medicine, haplotype data has an essential role in interpreting personal genomes and guiding individualized treatment plans. Haplotype data have been utilized in many areas of genetic studies, including imputation of low-frequent variants (Huang et al., 2015b; McCarthy et al., 2016) and characterization of DNA-phenotype associations (Lambert et al., 2013; Trégouët et al., 2009). Numerous GWAS studies have indicated that while single-SNP analysis is not optimal, joint analysis of multiple SNPs along chromosomes, i.e., haplotypes, showed significantly increased power for detection of genetic determinants for complex traits. The cartoon in Figure 4.1 illustrates how the lack of haplotype information limits the interpretation of existing genomes.

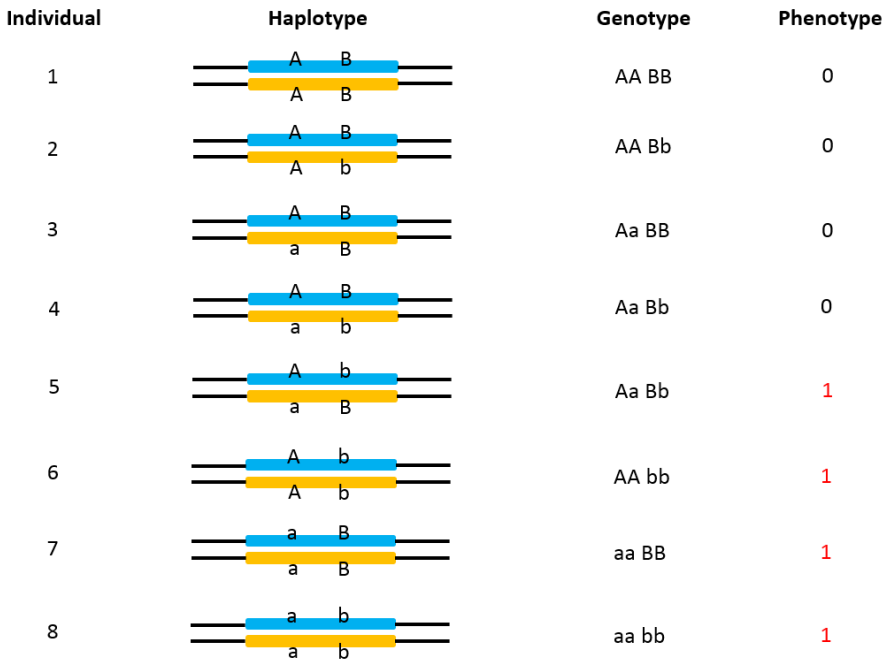


Figure 4.1: A hypothetical example showing the advantage of using haplotype data over individual SNPs data

Determination of haplotypes, termed phasing or haplotyping, is the process of inferring haplotype architecture based on genotypic data using statistical or bioinformatic approaches. The most widely used haplotyping strategy is to phase common genetic variants using population data (Browning and Browning, 2007; Delaneau et al., 2013a; Delaneau et al., 2012; Delaneau et al., 2013b; Howie et al., 2009; Li et al., 2010; Loh et al., 2016; O'Connell et al., 2016; Scheet and Stephens, 2006; Stephens and Scheet, 2005; Stephens et al., 2001), however, this approach is incapable to phase *de novo* mutations, rare variants, or structural variants, and is limited to infer short-range haplotype fragments, which constrains its use in genetic studies as well as precision medicine. Experimental approaches targeting whole-chromosome phasing involve the physical separation of homologous chromosomes in diploid cells using chromosome

microdissection, FACS-mediated chromosome sorting, or microfluidics, followed by single-chromosome sequencing (Fan et al., 2011; Ma et al., 2010; Yang et al., 2011). Nevertheless, these approaches usually require specialized equipment which are considerably expensive and are typically time-consuming. Numerous sequencing technologies including fosmid-based dilution pool sequencing, long fragment read (LFR) technology, PacBio single molecule real-time (SMRT) long-read sequencing, 10X Genomics linked-read sequencing, and proximity ligation (Hi-C) sequencing can also be employed to generate long-range haplotype fragments (Peters et al., 2012; Selvaraj et al., 2013). Bioinformatics tools, such as HAPCUT2, have been developed to assemble haplotypes using data generated from diverse sequencing technologies (Edge et al., 2017). A novel single-cell DNA template strand sequencing (Strand-seq) technique has been invented to sequence either Watson strand or Crick strand of a chromosome in a diploid somatic cell and phase chromosomal haplotypes using pooled Strand-seq libraries (Porubský et al., 2016). With 183 libraries sequenced by the Illumina HiSeq 2500 sequencing platform, approximate 80% of the genotyped hetSNPs were phased with a concordance of 99.3% compared with the HapMap reference. A comparison among various sequencing technologies coupled with Strand-seq method suggested that using 10 Strand-seq libraries and 10x coverage PacBio long-read or 10X Genomics linked-read sequencing data can successfully phase more than 95% of the total number of hetSNPs (Porubsky et al., 2017). However, the cost associated with these methods to phase individual genome using diploid genotype data are still high, making large-scale research infeasible.

Gametes produced by an individual, such as pollen grains in plants or sperms and eggs in animals, are the natural packaging of haploid complements that are formed by meiotic recombination. Using haploid data of single gamete cells may substantially reduce the complexity in inferring the donor's chromosomal haplotypes, compared to the phasing approaches using diploid data. However, the development of gamete-based phasing methodologies has been premature and inadequate. Some recent efforts have been made to reconstruct chromosome-scale haplotypes with gamete cells, but these methods required either a large number of gametes for the analysis or requires manual inspection for assembly to ensure phasing accuracy (Hou et al., 2013; Kirkness et al., 2013; Lu et al., 2012). No cost-efficient and user-friendly software has been made available for phasing chromosome-length haplotypes with gamete data. To fill this void, we developed an innovative method, named *Hapi* (short for Haplotyping with imperfect genotype data), for a fully-automatic inference of an individual's chromosomal haplotypes using 3 to 5 gametes. A comprehensive comparison, involving the use of a simulated dataset, a maize microspore dataset, and a human sperm sequencing dataset, demonstrated that the new *Hapi* method outperformed two existing approaches in terms of phasing accuracy and cost efficiency. The results also suggested that chromosomal haplotypes may be inferred by using only 3 gamete cells if the genotype data are of high quality. The simple, inexpensive and reliable methods for isolation, lysis, and whole-genome amplification (WGA) of single gamete cells together with the dramatically reduced number of gametes required in *Hapi* for phasing an individual genome, will make the genome-wide haplotype association study (GWHAS) – the next generation

GWAS – affordable and feasible. In addition, the crossover analysis module in the *Hapi* R package can be used to detect the crossovers on gamete chromosomes, which will facilitate the recombination-relevant researches and also hold promise of adoption by the public health sector including in the diagnosis of abnormal recombination activity in human sperms and eggs.

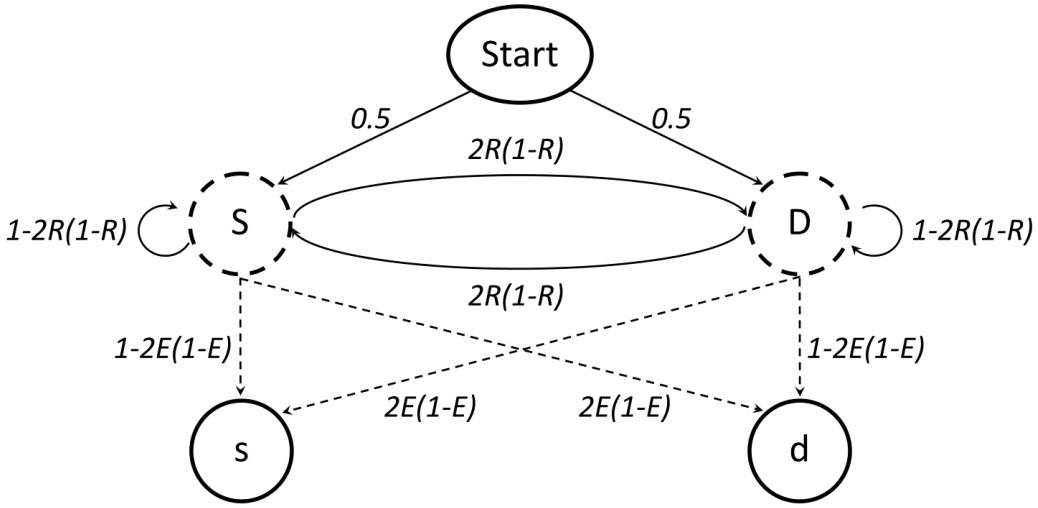
4.2 Materials and Methods

4.2.1 Key Component Algorithms Employed in Hapi

(1) Hidden Markov Model (HMM)

Enlightened by a previous study (Hou et al., 2013), an HMM is adopted to linearly scrutinize hetSNP markers along the chromosome in two gametes to identify markers bearing genotyping errors. In the HMM, there are two observations ‘s’ and ‘d’ indicating the two possible outcomes, either same or different, in terms of the relationship of observed genotype calls at a hetSNP locus between two gametes. Two hidden states, ‘S’ and ‘D’, represent the invisible relationship between the true genotypes of this marker in these two gametes, with ‘S’ and ‘D’ denoting the same and different genotypes, respectively. The initial probabilities of the two states are 0.5. Because the observed genotype outcomes may be different from the hidden states due to the genotyping errors at rate *E*, the emission probabilities to observe the same genotype calls, i.e., s, given the S hidden state is *1-2E(1-E)* and to observe the different genotype calls, i.e., d, is *2E(1-E)*. The emission probabilities given the D state are defined in the same way. A transition is defined as a change in state when scanning two adjacent markers, indicating that a

meiotic recombination likely occurs between these two markers on either gamete chromosome. Suppose the recombination frequency is *R*, the transition probabilities from one state to itself is *1-2R(1-R)*, and to the other state is *2R(1-R)*. After defining the HMM, Viterbi's algorithm can be used to determine the most likely hidden state for each marker. Markers with genotyping errors are determined where there are conflicts between the observed outcomes and the inferred states. The HMM is iteratively applied to all gamete pairs for the detection of disputable SNP loci with potential genotyping errors.



gamete 1	1	0	1	1	0	0	0	1	0	0
gamete 2	1	0	0	1	0	0	1	1	0	0
Observations	s	s	d	s	s	s	d	s	s	s
Inferred states	S	S	S	S	S	S	S	S	S	S
	√	√	×	√	√	√	×	√	√	√

Figure 4.2: HMM for detection of hetSNPs with potential genotyping errors

(2) Imputation of missing genotypes

We define a framework as a set of selected hetSNP markers for constructing draft haplotypes for the chromosome. Missing data for the framework markers in the gametes are imputed in an iterative manner. When a missing region (either a single marker or consecutive markers) of a 'target' gamete is to be imputed, the two markers immediately around this region, called comparator markers, are first compared with those in other 'support' gametes. The missing region can be imputed with the information from a support gamete cell only if the genotype calls for these two comparator markers in the target gamete are either both identical or both complementary to those in the support gamete. For example, if genotype calls of the two comparator markers in the target gamete are both identical to those in the support gamete, the missing region on the target gamete is simply imputed with genotype calls of markers in the same region in the support gamete. Otherwise, the missing region in the target gamete is imputed with the reciprocal genotypes in the support gamete. Missing genotypes in one gamete can be eventually resolved only if the imputations are supported by more than 2 support gametes and no imputation conflict is incurred. Once all the gametes are imputed in one iteration, genotypes in the missing regions are updated and the entire process described above will be repeated until no more missing data can be further imputed.

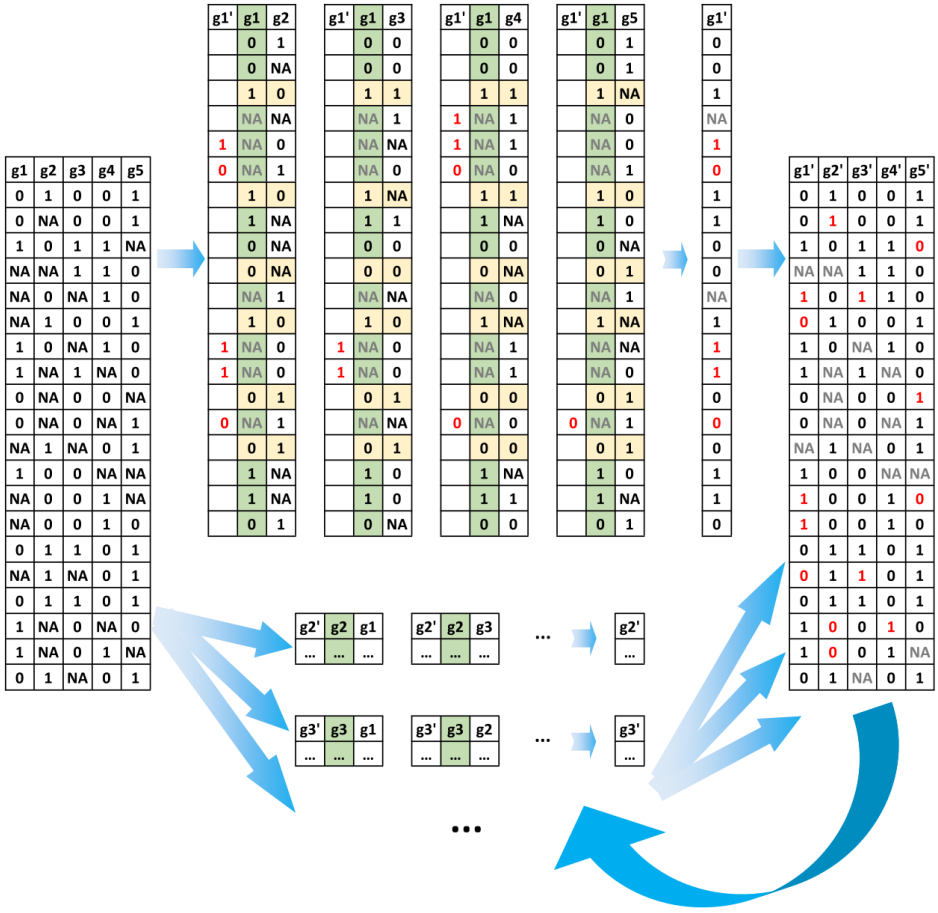


Figure 4.3: Imputation of missing genotypes. (g1, …, g5 are 5 gamete cells. g1', …, g5' represent imputed gametes)

(3) Majority voting

With the assumption that recombination is generally rare on the chromosome and even rarer to occur between two neighboring framework markers (a small region) in multiple gametes, the haplotypes of these two adjacent framework markers are deduced by analyzing genotype links (genotype patterns for these two markers) across all gametes based on the majority voting principle. There are two types of links between these two neighboring framework markers, i.e., type I links include genotype patterns 0-0 and 1-1

and type II links include genotype patterns 0-1 and 1-0, where 1 and 0 represent two complementary genotype calls that are arbitrarily and independently assigned at either locus (Figure 4.4). The most frequent link type is determined as hap-link which represents the likely haplotypes for the two framework markers, whereas the minority link type is considered as cv-link arising from a crossover. The final draft haplotypes can be deduced through walking and voting along the framework of the chromosome.

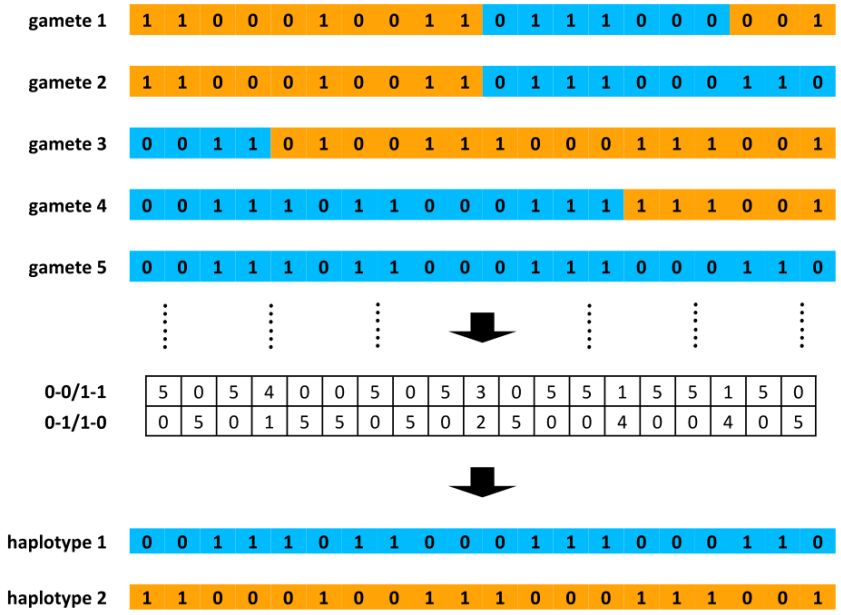


Figure 4.4: Majority voting strategy for draft haplotype inference. Different colors indicate different haplotypes

(4) Maximum parsimony of recombination (MPR)

MPR, an optimality criterion to search for the haplotype arrangement with minimum number of crossovers in a chromosomal region across all gametes, is adopted by *Hapi* to proofread the equivocal regions (two adjacent framework markers) of draft haplotypes where disputable cv-links have been observed (Figure 4.5). When five or more gametes are analyzed, we treat any two adjacent markers with 2 or more cv-links as candidate

regions for proofreading. If very few (e.g., 3 or 4) gametes are in use, every two adjacent markers with cv-link(s) are subject to proofreading. The draft haplotypes are first segmented into blocks by the equivocal regions. Small blocks with little genotypic data are excluded from the construction of the draft haplotypes. To phase two neighboring blocks, raw genotype calls (with possible missing data) of joining hetSNPs markers, i.e., the last 100 consecutive hetSNPs in the first block and the first 100 consecutive hetSNPs in the second block, are retrieved. Since haplotypes within each block are unambiguous, there are only two possible combining haplotypes for these two blocks. The total number of crossovers in all gametes are counted given the two combining haplotypes, and the one generating less crossovers is preferred by the MPR algorithm.

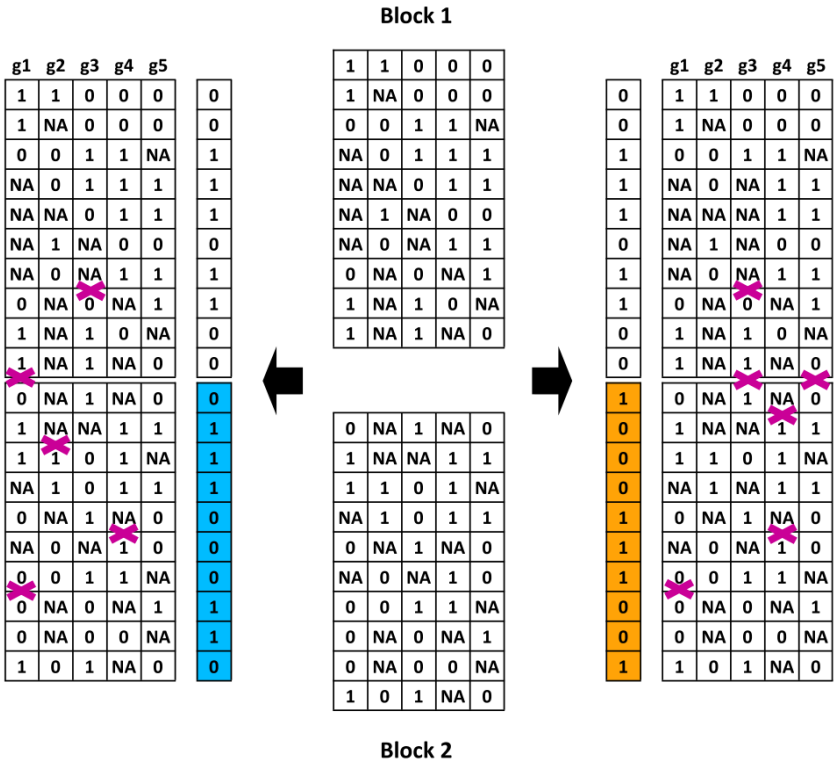


Figure 4.5: MPR for draft haplotype proofreading. The combining haplotype that generates less crossovers is determined as the true haplotype

(5) Assembly of chromosome-length haplotypes

We arbitrarily select one of the inferred draft haplotypes and use it as a blueprint to deduce gamete-specific haplotypes and eventually assemble the chromosome-length consensus haplotypes through 3 steps (Figure 4.6). In step 1, genotype calls of framework markers in each gamete chromosome are compared to the blueprint to identify haplotype-converting points (HCPs) which are caused by potential recombination. These HCPs partition the gamete chromosome into k haplotype segments, where k-1 is the number of HCPs identified for this gamete chromosome. For the segments 1 through k, genotype calls of hetSNPs in every second segment are flipped to form a gamete-specific haplotype, where 'flip' refers to switching the current genotype call to its reciprocal genotype. In step 2, each gamete-specific haplotype is synchronized with the blueprint by either remaining the same or flipping over the genotypes of entire chromosomal hetSNPs. In step 3, the first consensus chromosome-length haplotype is reconstructed via voting for the most frequent allele at each hetSNP locus across all the gamete-specific haplotypes. The second consensus haplotype is obtained by simply flipping genotypes of hetSNPs on the first chromosome-length haplotype.

If a crossover occurs at the end of a gamete chromosome where hetSNPs are not enclosed in the framework, it becomes very challenging to correctly infer the haplotypes for this chromosome-tip region. *Hapi* employs an additional capping strategy to polish two ends of chromosomal haplotypes. First, hetSNPs in such region are combined with the immediately adjacent 200 consecutive hetSNPs at the joining end of the framework to form a capping block, which is treated as a small chromosome. Then the majority of

gametes with high-consistent genotype calls in the capping block are used to build small draft haplotypes. The same strategy is adopted to generate gamete-specific haplotypes to deduce consensus haplotypes in this small region. The inferred haplotypes for the capping block are integrated into the chromosome-length haplotypes to accomplish the assembly.

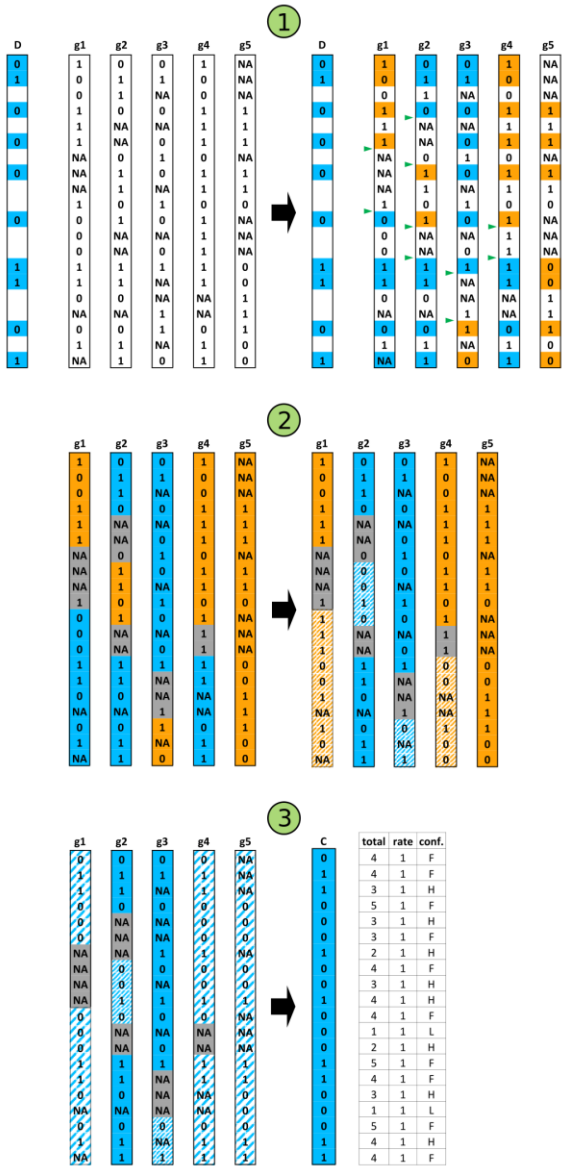


Figure 4.6: High-resolution consensus haplotype assembly. (g1, …, g5 are 5 gametes; D: draft haplotype; C: high-resolution consensus haplotype)

4.2.2 Rival Phasing Methods

(1) One-versus-All (*OVA*) pipeline

Kirkness et al. (2013) proposed a two-stage strategy to infer chromosome-scale haplotypes by combining the use of genotyping array data and next-generation sequencing data of sperm cells. In the first stage, array data with relatively high call rate (50.9% on average in their study) were analyzed in a one-versus-all fashion to identify crossovers in the gametes, which were then used to construct the draft haplotypes. When phasing a chromosome, a gamete is set as a reference, and the other gametes are considered as offspring. HCPs are identified for all reference-offspring pairs, where a HCP indicates the position with a potential crossover either on the reference or on offspring chromosome. A crossover is assigned to the reference chromosome if the HCP is identified in the majority of pairwise comparisons, for example, 13 out of 15 as indicated in the original paper. Otherwise, multiple crossovers must have taken place on the offspring chromosomes. A manual inspection step is required to confirm the crossover locations on each reference chromosome. As a result, a gamete-specific chromosome-scale haplotype can be inferred by the crossovers assigned to the reference chromosome. The entire process described above is repeated until each gamete has been set as a reference for one time. Draft haplotypes can be constructed using these gamete-specific haplotypes by voting for the major allele at each locus. In the second stage, the inferred crossover positions are employed again to assist the analysis of the additional sequencing data in the gametes to infer the high-resolution consensus chromosome-scale haplotypes.

To perform the comparison analysis, this algorithm was written in R language by us with a few optimizations. (1) Rather than using two sets of gamete genotype data, i.e., SNP array data and sequencing data in the original study, only one dataset is used for the modified two-stage *OVA* pipeline. (2) The gamete genotype data are preprocessed to remove markers with potential genotyping errors and a subset of high-quality hetSNPs is selected to infer crossovers in the gametes. (3) A HMM is used to detect HCPs with higher level of accuracy. (4) A well-written function is developed to automatically determine crossovers to replace the manual inspection required in the original pipeline.

(2) Pairwise HMM (*PHMM*)

The *PHMM* pipeline developed by Hou et al. (2013) evolved from the *OVA* pipeline by introducing a HMM-based HCP detection approach to the reference-offspring pairwise-comparison scheme. For each reference chromosome, a crossover can be directly inferred if, within a 1Mb sliding window, HCPs can be identified in over 60% of the reference-offspring pairs. Detailed description of the pipeline can be found in the original paper (Hou et al., 2013). Source code of a series of C++ programs and perl scripts for implementing the PHMM pipeline are publicly available. To facilitate the comparison analysis in this study, we directly applied the C++ programs for crossover identification but rewrote the perl scripts in R language (without changing the original algorithm) for the inference of consensus haplotypes.

4.2.3 Maize microspore dataset

The maize microspore sequencing dataset was generated by Li et al (2015). A total of 96 (24 × 4) microspores from 24 tetrads were isolated from F1 hybrid individuals of a cross

between two inbred lines (SK and ZHENG58) and were sequenced at ~1.4× depth coverage. Parents of the F1 hybrid were also sequenced at up to 8× (SK) and 15.7× (ZHENG58) genome coverage depth, respectively. With a stringent filtering process, a total of 599,154 high-quality SNPs was obtained for both parents and the microspores. Note that, for the F1 hybrid, two parental haplotypes are known.

4.2.4 Human sperm dataset

Single sperm cell sequencing data were downloaded from the NCBI Sequencing Read Archive (SRA) (https://www.ncbi.nlm.nih.gov/sra) with the accession number SRP017516 (Kirkness et al., 2013). Sequences were aligned to the human GRCh37 reference genome using BWA-MEM (Li and Durbin, 2009) implemented in the SpeedSeq software (Chiang et al., 2015). Duplicate-marked, sorted, and indexed BAM files were produced by the SpeedSeq align module, which utilizes SAMBLASTER (Faust and Hall, 2014) to mark duplicates and uses Sambamba (Tarasov et al., 2015) to sort and index BAM files. For each sperm, the genotypes involving 1.95 million heterozygous SNP loci in the HuRef genome were determined using Genome Analysis Toolkit (GATK) following the recommended best practices (DePristo et al., 2011).

4.3 Results

4.3.1 Implementation of *Hapi*

Phasing two chromosomes are completely independent of one another. We first demonstrate the strategy for inferring haplotypes for a chromosome using gamete data, and the strategy can be simply applied to haplotyping other chromosomes in the same way. Implementing the *Hapi* method to phase an entire chromosome consists of three

steps: (1) data preprocessing, (2) construction of draft haplotypes and (3) inference of high-resolution chromosomal haplotypes (Figure 4.7). In step (1), markers with potential genotyping errors in any gamete cells are filtered out via an iterative HMM analysis of gamete pairs. A subset of markers, which have been successfully genotyped in at least 3 gametes, are selected to form a "precursor" framework. In the framework, missing data in each gamete are iteratively imputed using the supportive data in other gametes. The markers, usually of a small number, with missing data that cannot be fully resolved by imputation are eliminated, resulting in the final framework for building draft haplotypes. In step (2), the draft haplotypes are derived by sequentially analyzing two adjacent framework markers using the majority voting method, through which the haplotypes for these two markers are determined by the link type represented in the majority of the gametes. The MPR principle is then adopted to proofread the draft haplotypes at the positions where disputable cv-links appear. In step (3), each gamete chromosome is compared to the draft haplotypes to deduce gamete-specific haplotypes, with the non-framework markers being phased. Consensus high-resolution haplotypes are eventually determined by these gamete-specific haplotypes.

A user-friendly R package has been developed for implementing the *Hapi* method to infer chromosome-length haplotypes using genotype data of single gamete cells. The gamete genotype data may be generated from various platforms including genotyping arrays and sequencing. The *Hapi* method uses genotype data of hetSNPs in individual gametes and outputs the high-resolution chromosomal haplotypes as well as confidence

level of each phased hetSNPs. The package also includes a module allowing downstream analyses and visualization of crossovers identified in the gametes.

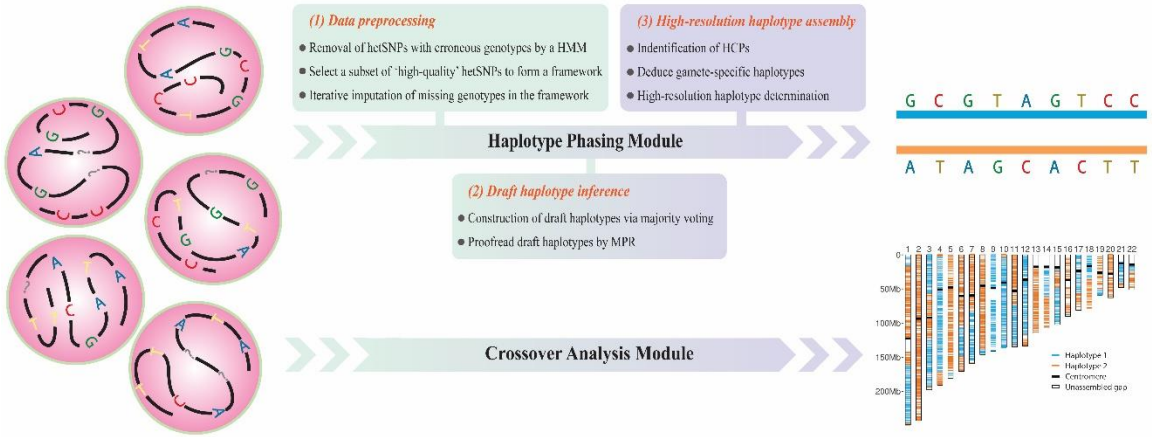


Figure 4.7: Overview of the *Hapi* pipeline

4.3.2 Analysis of simulated data

To evaluate the performance of *Hapi* compared with the other two competitive methods, *OVA* and *PHMM*, we carried out a comprehensive simulation study in which a single chromosome of 100Mb is considered. Three factors that may affect phasing accuracy and cost efficiency were considered in each scenario, *i.e.*, (1) the number of hetSNP markers on the chromosome, (2) the number of gametes, and (3) the rate of missing genotype data. In the simulation, a pool of 100 haploid gametes is generated from a single diploid donor. The number of hetSNPs on the chromosome ranged from 5,000 to 1,000,000. 3 to 15 gametes, each with 0 to 3 crossovers on the chromosome, were arbitrarily selected from the 100 haploid gametes. 10% to 70% of missing genotype data were randomly introduced to each simulated gamete chromosome. Moreover, 1% genotyping errors were randomly placed on the simulated gamete chromosomes. The simulated data in each

scenario was analyzed using the three methods, respectively, and replicated for 100 times. A successful inference was defined if more than 99% of hetSNPs were correctly phased in each replicate run.

The results indicated that *Hapi* outperformed the other two methods in both phasing accuracy and cost efficiency (Figure 4.8). When 5000 hetSNPs on the chromosome (low marker density) were considered, *Hapi* only needed 6 gametes to correctly infer haplotypes even with 60% of missing genotype data. For *OVA*, at the missing rate of 50%, the first 100% correct inference of haplotypes occurred when 7 gametes were used. However, when more gametes were included in the analysis, the performance of *OVA* was not monotonically increased, indicating a lack of reliability and robustness of the method. If 70% of the marker data were missing (extreme situation), *Hapi* was able to reconstruct haplotypes correctly with 11 or more gametes; whereas, *OVA* failed to do so even all 15 gametes have been used. With increased density of hetSNPs, fewer gametes were needed and higher rate of missing genotypes can be tolerated for both methods to correctly phase the chromosome, however, *Hapi* always outcompeted *OVA* by requiring even less gametes and allowing more missing data. The results also indicated that only 3 gametes may be enough for successful inference of chromosomal haplotypes when gamete data are of high quality. *PHMM* behaved quite differently from the other two methods. The performance of *PHMM* did not change with the rate of missing data, while the performance was barely improved with the increase in number of hetSNPs. Rather, the phasing accuracy of *PHMM* depended on the number of gametes used in analysis. In

general, much more gametes are required for *PHMM* to infer correct haplotypes than the other two methods.

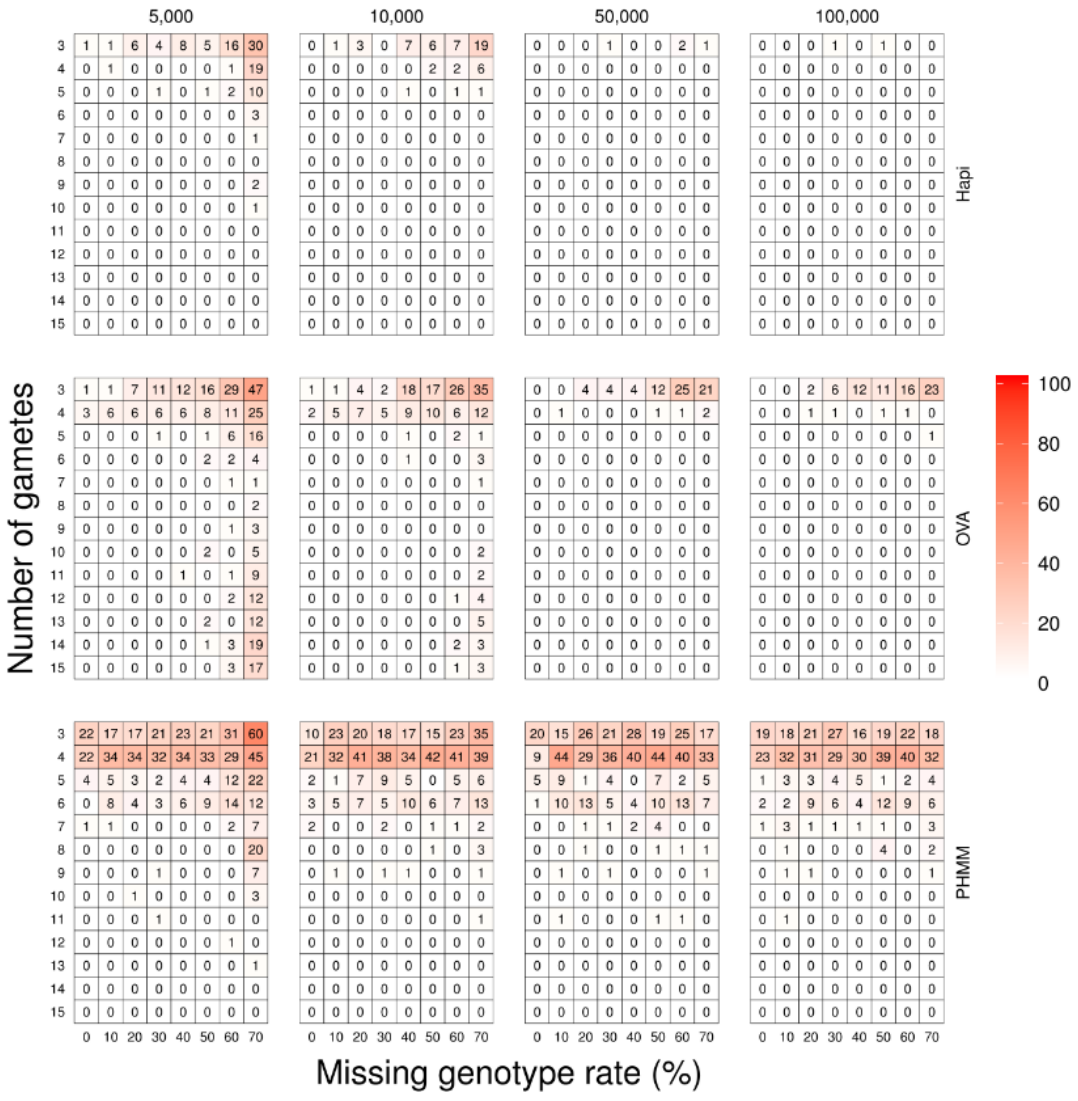


Figure 4.8: Performances of three methods (*Hapi*, *OVA*, and *PHMM*) in the simulated dataset. The number in each heatmap grid denotes for how many times out of the 100 replicates the haplotypes are incorrectly inferred in that scenario

4.3.3 Analysis of maize microspore dataset

A maize microspore sequencing dataset from F1 hybrid individuals of a cross between two inbred lines is used to further evaluate the performances of three methods. This is an ideal validation dataset since the parental haplotypes are known. To avoid using microspores from the same meiosis event, one microspore from each of the 24 tetrads was randomly selected to form a 24-gamete pool. The number of hetSNPs on maize chromosomes ranges from 42691 (Chr10) to 82689 (Chr1). The average rate of missing genotype data for 10 chromosomes across the 24 selected gametes is about 50%, with the maximum missing rate equal to 72.46% (Table 4.1). When phasing a chromosome, 24 selected gametes were sorted in descending order of missing rates on that chromosome, *i.e.*, the first gamete in the sorted list has the most missing data for the chromosome. 3 to 15 gametes were sequentially selected from the sorted list and analyzed with the three methods, respectively, to infer haplotypes for that chromosome. This process is repeated to phase all 10 chromosomes, yielding a total of 390 scenarios (13 numbers of gametes × 10 chromosomes × 3 methods). In each scenario, the phased chromosome was compared with the known parental haplotypes to calculate phasing accuracy. A successful inference of chromosomal haplotypes is defined if > 99% of the markers can be correctly phased.

Table 4.1: Missing genotype rate of the 24 maize microspores on each chromosome

	Chr1	Chr2	Chr3	Chr4	Chr5	Chr6	Chr7	Chr8	Chr9	Chr10
T12	0.56	0.55	0.56	0.53	0.50	0.55	0.55	0.62	0.58	0.61
T24	0.49	0.45	0.46	0.50	0.42	0.46	0.46	0.55	0.49	0.52
T33	0.54	0.51	0.51	0.50	0.47	0.53	0.50	0.59	0.54	0.47
T42	0.56	0.54	0.58	0.53	0.51	0.57	0.55	0.60	0.56	0.60
T54	0.60	0.60	0.57	0.59	0.54	0.59	0.58	0.67	0.60	0.64
T61	0.64	0.61	0.64	0.62	0.59	0.63	0.61	0.57	0.63	0.66
T73	0.56	0.56	0.53	0.57	0.59	0.54	0.52	0.60	0.54	0.46
T83	0.56	0.55	0.55	0.54	0.50	0.56	0.54	0.48	0.55	0.49
T93	0.51	0.48	0.52	0.50	0.47	0.50	0.49	0.57	0.47	0.54
T102	0.54	0.54	0.56	0.53	0.57	0.55	0.50	0.60	0.55	0.60
T113	0.69	0.67	0.67	0.68	0.65	0.71	0.66	0.64	0.67	0.72
T123	0.49	0.50	0.47	0.49	0.47	0.51	0.48	0.49	0.49	0.43
T133	0.57	0.50	0.54	0.50	0.55	0.51	0.52	0.58	0.49	0.47
T143	0.48	0.46	0.50	0.46	0.50	0.47	0.45	0.46	0.48	0.41
T151	0.58	0.54	0.58	0.57	0.60	0.57	0.56	0.61	0.56	0.48
T163	0.47	0.46	0.48	0.49	0.49	0.49	0.48	0.41	0.46	0.52
T174	0.63	0.60	0.58	0.58	0.65	0.60	0.60	0.70	0.61	0.50
T181	0.47	0.45	0.43	0.45	0.52	0.45	0.46	0.51	0.46	0.40
T191	0.55	0.53	0.51	0.52	0.56	0.51	0.51	0.62	0.56	0.48
T202	0.51	0.53	0.57	0.54	0.58	0.54	0.53	0.50	0.55	0.47
T214	0.54	0.64	0.57	0.54	0.58	0.60	0.53	0.47	0.55	0.59
T224	0.53	0.52	0.53	0.51	0.46	0.52	0.50	0.44	0.50	0.56
T232	0.55	0.53	0.53	0.56	0.50	0.56	0.54	0.47	0.57	0.48
T241	0.57	0.56	0.60	0.58	0.53	0.54	0.54	0.51	0.55	0.49

The results indicated that the *Hapi* method can achieve phasing accuracies of greater than 99.9% for most of the time, with two exceptions at 98.46% for chr2, and 99.89% for chr6, respectively (Figure 4.9). A close look at chr2 of 3 gametes disclosed two crossovers on two gamete chromosomes in a small region (39 hetSNPs in between) which approaches one end of the chromosome. Because, by default, a small block (< 100 hetSNPs) delimited by two close-in crossovers will be excluded from the draft haplotypes by the MPR, haplotypes of the two merging framework markers were incorrectly inferred by misinterpreting the link types in between due to the removed crossovers. To achieve a

phasing accuracy of > 99% for all the 10 chromosomes, at least 6 and 7 gametes are required for *OVA* and *PHMM*, respectively.

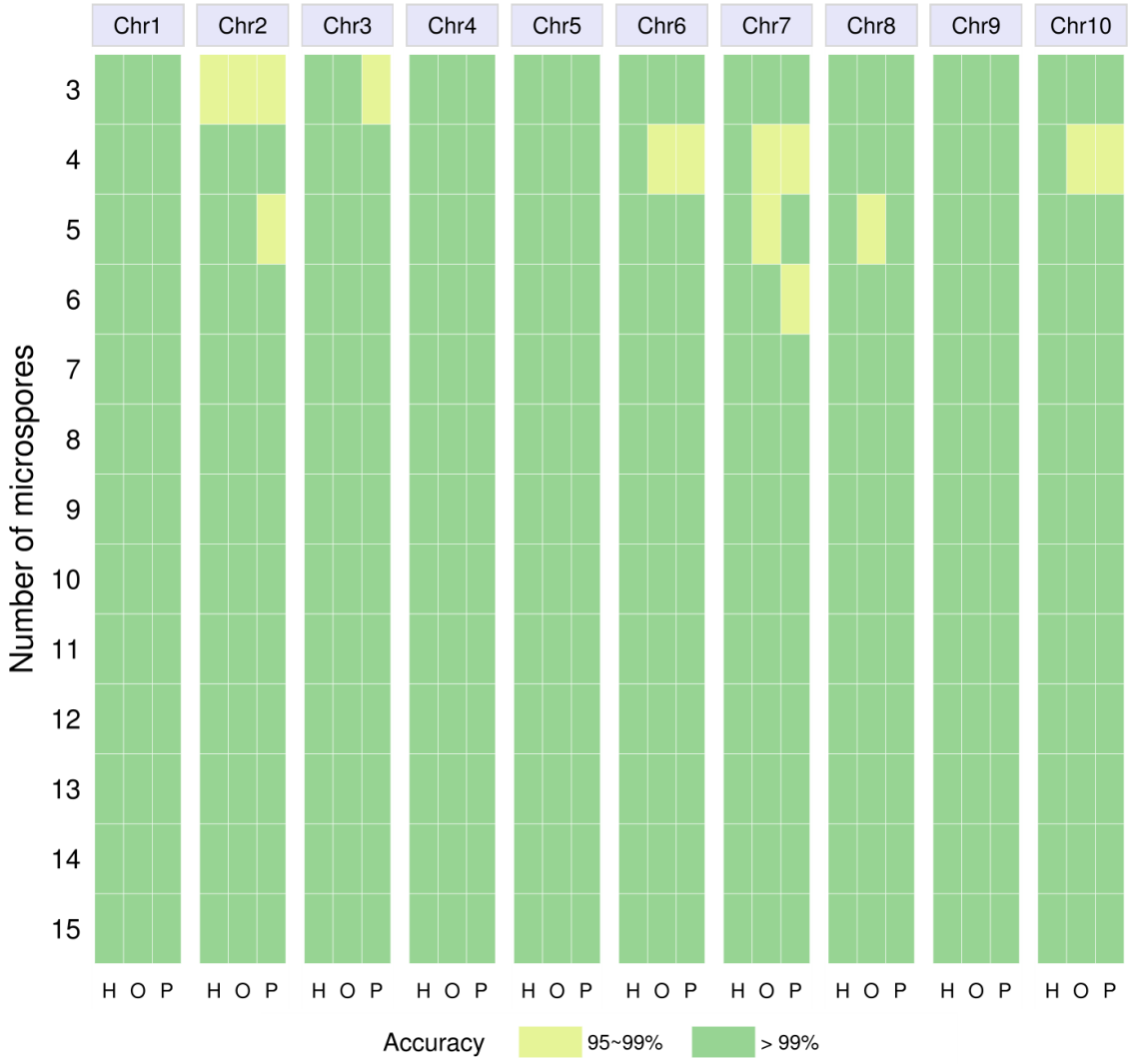


Figure 4.9: Performances of three methods (H: *Hapi*, O: *OVA*, and P: *PHMM*) in the maize microspore dataset

4.3.4 Analysis of human sperm dataset

To further benchmark the three phasing methods, a human sperm sequencing dataset consisting of 11 independent sperm cells from the donor of the HuRef diploid genome sequence was used (Kirkness et al., 2013). Although the true haplotypes for this donor are unknown, a 'phased' genome consisting of 1.82 million hetSNPs has been suggested based on a joint analysis of these 11 sperms sequenced at 1.5~3.7× coverage and 16 additional sperms genotyped using the Illumina HumanOmni-Quad v1.0 BeadChip (array data not publicly available). The raw sequencing data of the 11 sperm cells were downloaded and 1.66 million out of the 1.82 million hetSNPs were called in at least one sperm by GATK. The number of hetSNPs on 22 autosomes ranges from 15340 (Chr22) to 141669 (Chr2), and the rate of missing genotype data ranges from 70.95% to 86.49% (Table zx). When phasing a chromosome, the 11 sperms were sorted in a similar manner as for the maize data based on the missing genotype rate. 3 to 11 sperms were sequentially selected from the sorted list and analyzed using three methods, respectively, to infer chromosomal haplotypes which were then compared with the hetSNPs 'phased' in the original study to calculate the concordance rate. Since the chromosomal haplotypes recommended in the original study may be subject to errors, we relaxed the criterion in the sperm analysis by defining a successful inference of haplotypes if > 95% of phased markers are in agreement with the haplotypes suggested by Kirkness et al. (2013).

Table 4.2: Missing genotype rate of the 11 human sperms on each autosome

	X06	X34	X37	X44	X82	X91	X01	X45	Y47	X69	X22
Chr1	0.68	0.79	0.72	0.74	0.73	0.70	0.75	0.80	0.80	0.76	0.75
Chr2	0.67	0.78	0.72	0.74	0.71	0.68	0.74	0.81	0.80	0.75	0.76
Chr3	0.66	0.76	0.70	0.72	0.70	0.69	0.75	0.80	0.79	0.74	0.73
Chr4	0.63	0.74	0.70	0.71	0.70	0.66	0.74	0.79	0.78	0.72	0.74
Chr5	0.67	0.76	0.72	0.73	0.71	0.68	0.75	0.79	0.79	0.72	0.73
Chr6	0.66	0.75	0.69	0.72	0.71	0.66	0.74	0.79	0.79	0.74	0.74
Chr7	0.66	0.79	0.73	0.74	0.73	0.68	0.76	0.81	0.80	0.76	0.76
Chr8	0.65	0.77	0.72	0.74	0.71	0.70	0.75	0.81	0.78	0.76	0.76
Chr9	0.70	0.78	0.73	0.74	0.72	0.67	0.75	0.80	0.80	0.74	0.77
Chr10	0.67	0.78	0.72	0.74	0.73	0.72	0.77	0.81	0.80	0.78	0.79
Chr11	0.63	0.78	0.70	0.73	0.72	0.69	0.74	0.80	0.80	0.76	0.74
Chr12	0.68	0.76	0.71	0.72	0.72	0.68	0.77	0.81	0.81	0.74	0.79
Chr13	0.65	0.77	0.70	0.72	0.71	0.66	0.76	0.79	0.80	0.74	0.75
Chr14	0.67	0.79	0.71	0.73	0.72	0.69	0.77	0.81	0.80	0.76	0.78
Chr15	0.69	0.78	0.75	0.77	0.74	0.69	0.79	0.74	0.83	0.77	0.81
Chr16	0.75	0.83	0.78	0.80	0.77	0.74	0.81	0.84	0.84	0.81	0.79
Chr17	0.76	0.84	0.78	0.81	0.80	0.76	0.83	0.87	0.87	0.83	0.82
Chr18	0.69	0.79	0.72	0.73	0.74	0.68	0.76	0.82	0.80	0.75	0.78
Chr19	0.84	0.86	0.82	0.84	0.84	0.81	0.87	0.90	0.89	0.86	0.85
Chr20	0.71	0.80	0.78	0.77	0.76	0.73	0.82	0.83	0.86	0.81	0.78
Chr21	0.70	0.79	0.77	0.77	0.75	0.72	0.81	0.83	0.79	0.78	0.72
Chr22	0.82	0.87	0.82	0.83	0.83	0.82	0.88	0.91	0.87	0.85	0.84

The results showed that *Hapi* can correctly phase all 22 autosomes with 3 sperms; whereas, *OVA* and *PHMM* required at least 7 and 8 sperms, respectively, to achieve the same level of accuracy (Figure 4.10). When 7 or less sperms were used, *Hapi* performed consistently well but the performances of *OVA* and *PHMM* fluctuated wildly, indicating *Hapi* provides more reliable phasing results with small sample. Interestingly, *PHMM* can correctly infer the haplotypes of chromosome 1 with 6 to 10 gametes but failed when all 11 sperms had been used. Although a consistency of 95% was used to determine the success of haplotype inference, *Hapi* achieved > 99% of consistency for 82% of the scenarios (164 out of 198). For *Hapi*, the majority of scenarios with consistencies of

95%~99% were for the analyses of Chr15, Chr16, and Chr21, which also appeared to be challenging to the other two approaches, suggesting a complication in the genotype data for these chromosomes. Overall, among the 1.66 million hetSNPs phased by *Hapi* using all the 11 sperms, 99.73% (1,658,197/1,662,611) of them are concordant with the haplotypes suggested by Kirkness et al. (2013). An inspection of the non-concordant hetSNPs showed that 49.1% of them are only supported by 1 sperm and 33.4% of them have discordancy among 2 or more supporting sperms. The disputably phased hetSNPs tend to cluster around the centromere or at either end of the chromosomes (Figure 4.11). The hetSNPs that are not in agreement between *Hapi* and the suggested haplotypes on Chr15 are evenly distributed along the chromosome, which might be ascribed to the complication in data of sperm Y47 contaminated by DNA from other lysed cells (discussed in the original paper). Phasing Chr15 is equally challenging for *OVA* and *PHMM* as well. Compared with *Hapi*, the major deficiency in haplotype phasing with *OVA* and *PHMM* are due to their core strategy of a direct inference of crossover positions, which is sensitive to the regions with ambiguous genotypes or complication caused by multiple crossovers in more than one gamete. For example, when phasing Chr1 by *PHMM*, if 10 sperms were analyzed, a crossover on the Chr1 in the sperm X69 (reference chromosome) was not claimed because it was only supported by 5 out of 9 other sperms and missed the cutoff of ≥0.6 for determining a crossover. However, when including the 11th sperm, the crossover became supported by 6 out of 10 sperms, which claimed a false crossover and yielded an incorrect gamete-specific haplotype. In *Hapi*, genomic regions harboring complicated multiple cv-links will be excluded from the draft haplotypes to

reduce the chance of phasing errors. In addition, A special capping function has been designed in *Hapi* to phase either end of the chromosome, which are usually excluded from the framework but may also involve crossovers. The *OVA* method, which also leverages draft haplotypes for phasing a chromosome, cannot handle recombination beyond the framework.

Figure 4.10: Performances of three methods (H: *Hapi*, O: *OVA* and P: *PHMM*) in the human sperm dataset. The 3 outer circles show the phasing concordance with the suggested haplotypes (Kirkness et al., 2013) for each method using 3 to 11 sperms. The 6 inner circles are the haplotypes inferred by the 3 methods using 3 sperms with the most missing genotypes

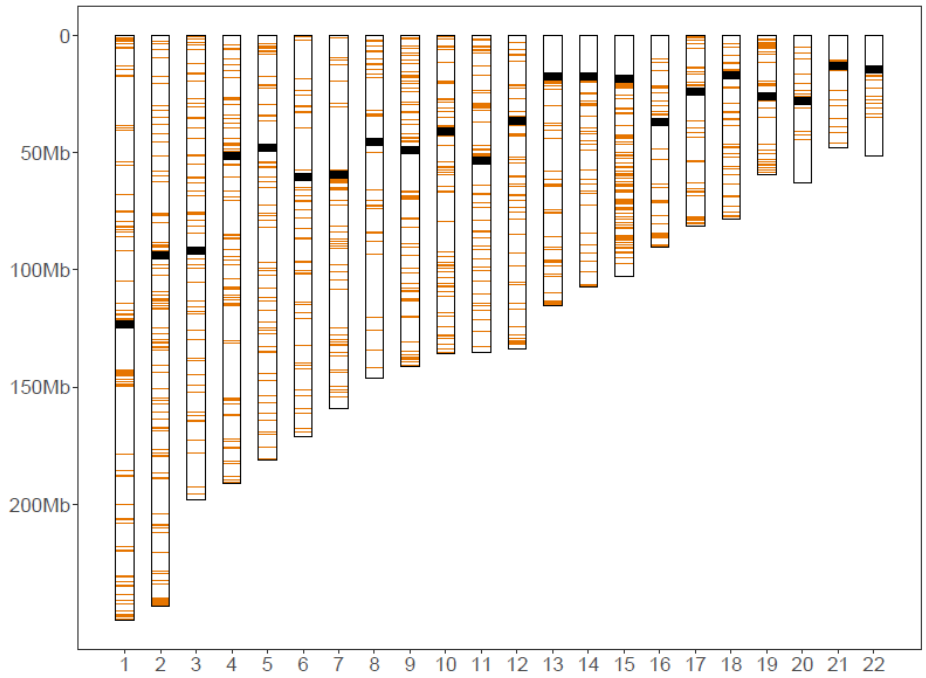


Figure 4.11: Distribution of hetSNPs that are not agreeable between *Hapi* and the suggested haplotypes (Kirkness et al., 2013)

4.3.5 Recombination analysis in sperm sequencing dataset

With the phased chromosome-length haplotypes, a HMM (a slightly different HMM from the one for detection of genotyping errors) was used to infer crossover positions in the sperm genomes by successively contrasting hetSNPs in each sperm with the chromosomal haplotypes. This HMM also consists of two observations (f and m) and two hidden states (F and M), representing the paternal and maternal haplotypes, respectively. The same initial probabilities of 0.5 are assigned to the two states. Given the F (or M) state, the emission probability of observing the f (or m) haplotype is 1-*E* and observing the complementary haplotype m (or f) is E, respectively. The transition probabilities from one state to itself is 1-R, and to the other state is R. A sequence of hidden states for the

‘chained’ markers are also inferred by Viterbi's algorithm and crossover positions are determined where there exist state swaps.

A total of 254 crossovers along the 22 autosomes were identified in the 11 sperms with an average of 1.05 per chromosome. Compared with the 260 crossovers identified in the original paper, 251 were also identified by the *Hapi* method (Table 4.3). The 12 inconsistent crossovers are all located at the ends of chromosomes, and such inconsistency may be ascribed to either of two following reasons. (1) The *OVA* method in the original paper ignores the crossovers at the chromosome ends which are not included in the draft haplotypes; thus, incorrect inference of haplotypes was occasionally made by the *OVA* method at the chromosome ends. (2) The observed double crossovers in a very small region are considered to be either caused by a gene conversion event or consecutive genotyping errors thus are filtered out by *Hapi*. The number of crossovers was counted in each bin (5Mb in length) along 22 autosomes and distributions of the 254 crossovers are depicted in Figure 4.12A. The resolution of crossover locations ranges from 79bp~788kb with a median of 89.3kb, which is roughly the same with the 82.5kb resolution reported in the original paper. Over 75% of the 255 crossovers were located within an interval of < 200kb (Figure 4.12B). Distribution of distances between any two chromosomally adjacent crossovers was provided, which can be used for recombination-relevant research including coexistent crossovers and interference in the formation of chromosomal crossovers during meiosis (Figure 12C). Functions for downstream analysis and visualization are included in the ‘crossover analysis’ module of the *Hapi* package.

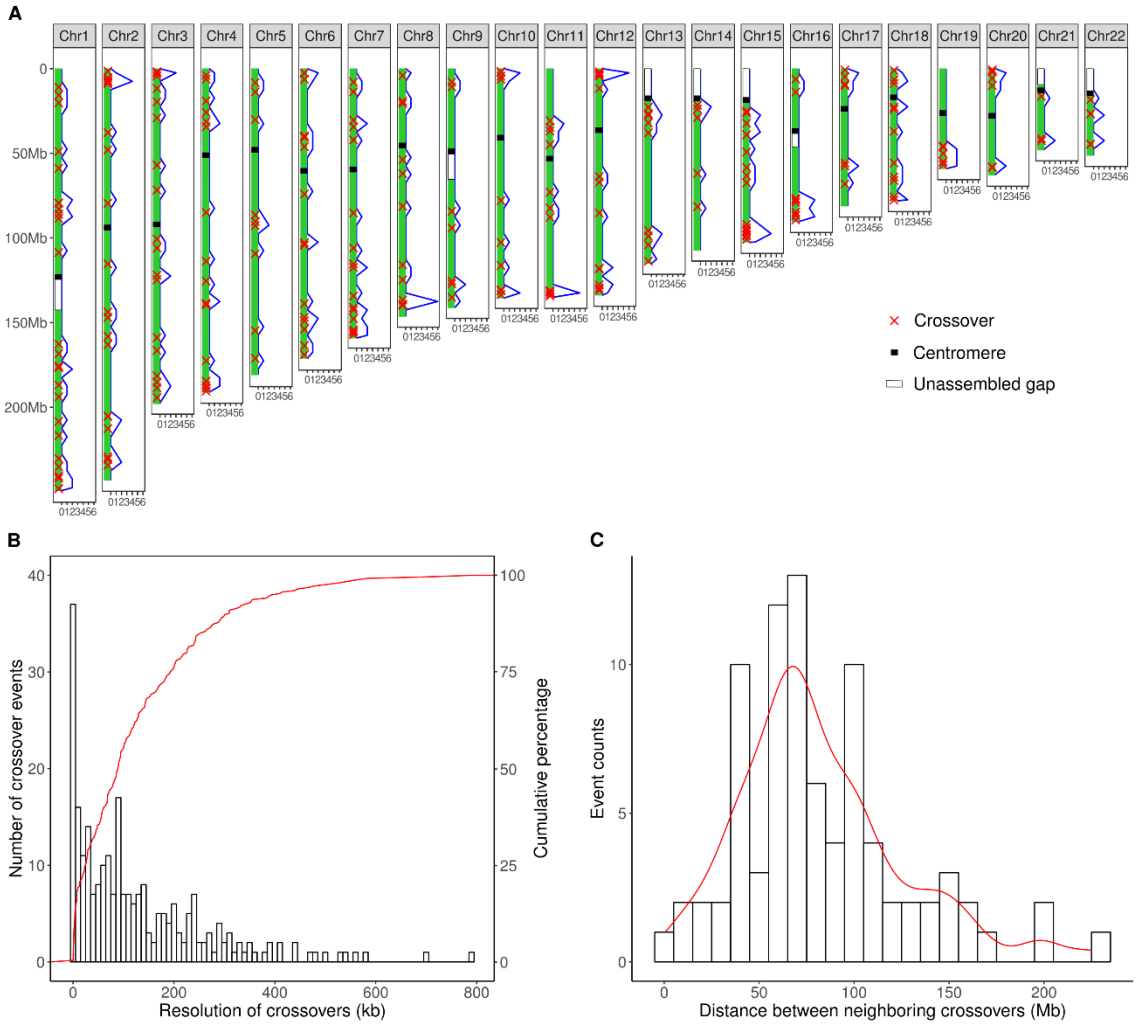


Figure 4.12: Crossover analysis in the human sperm sequencing dataset. (A) Distribution of 254 identified crossovers on the 22 autosomes. (B) Resolution of crossovers (distance between two adjacent markers that involve a crossover). (C) The distribution of distances between two neighboring crossovers on the same chromosome

Table 4. 3: Comparison of crossovers identified by *Hapi* with those reported in the original paper (Kirkness et al., 2013)

		Hapi			*OVA*		
sperm	chr	start	end	resolution	start	end	resolution
X01	1	82963262	82990167	26905	82956657	82987838	31181
X01	2	79566482	79568318	1836	79562010	79573515	11505
X01	2	158143819	158177489	33670	158080876	158177489	96613
X01	2				240479364	240490213	10849
X01	3	105979470	106007364	27894	105988605	105992865	4260
X01	3	185102541	185202707	100166	185099352	185199555	100203
X01	4	138521200	138593037	71837	138533327	138564014	30687
X01	6				465707	642667	176960
X01	6	39682646	39825364	142718	39769455	39851679	82224
X01	6	102884588	102892160	7572	102884588	102896943	12355
X01	7	13678506	13678959	453	13678247	13679931	1684
X01	7	115412350	115435818	23468	115412350	115439353	27003
X01	7	157048918	157141277	92359	157048403	157151234	102831
X01	8	81045489	81571548	526059	81045600	81710837	665237
X01	8	139578620	139826496	247876	139774409	139782538	8129
X01	11	131215076	131222575	7499	131217695	131224448	6753
X01	12	67038722	67236182	197460	67043027	67295238	252211
X01	15	63208349	63224483	16134	63207340	63215651	8311
X01	15	98586859	98719873	133014	98583617	98704009	120392
X01	16	78953605	78982911	29306	78950001	78979560	29559
X01	18	24024531	24026191	1660	24024531	24042875	18344
X01	19	54699386	55203458	504072	54705641	54851327	145686
X01	20	720916	722411	1495	719854	731744	11890
X01	22	18214192	18395952	181760	18234535	18403629	169094
Y47	1	88316577	88559595	243018	88297927	88569442	271515
Y47	1	186759597	186889473	129876	186599524	186894402	294878
Y47	2	6835876	7061691	225815	6835876	7056183	220307
Y47	2	205054041	205221016	166975	205057668	205114083	56415
Y47	3	2815838	2863911	48073	2815263	2867637	52374
Y47	3	100533853	100641397	107544	100550650	100561439	10789
Y47	3	166400514	166493066	92552	166403296	166503411	100115
Y47	3	181346937	181641361	294424	181426554	181583317	156763
Y47	5	109331459	109383586	52127	109371759	109381583	9824
Y47	6	138803449	138832027	28578	138821708	138823548	1840
Y47	7	34266979	34355100	88121	34276627	34293822	17195
Y47	7	117429299	117452215	22916	117429299	117451124	21825
Y47	9	84262895	84320069	57174	84275014	84281312	6298
Y47	10	1981253	2186314	205061	2104415	2148432	44017
Y47	11				2309898	2492629	182731
Y47	12	3855463	3900287	44824	3861721	4003134	141413
Y47	13	113575792	113906097	330305	112667537	113310547	643010
Y47	15	26189666	26190826	1160	26096382	26189969	93587
Y47	16	6143701	6285410	141709	6241183	6348492	107309

Y47	16	76511928	76800221	288293	76576729	76767431	190702
Y47	18	63941026	64135538	194512	63435260	64253268	818008
Y47	20	9975874	10081023	105149	9935686	10068488	132802
X45	1	79240690	79382934	142244	79106967	79330525	223558
X45	1	176496663	176610018	113355	176497685	176607426	109741
X45	1	235167914	235168970	1056	235152757	235539020	386263
X45	2	212434442	212675467	241025	212529506	212581902	52396
X45	3	1859218	1860457	1239	1856166	1865505	9339
X45	3	71652888	71756375	103487	71652888	71872014	219126
X45	5	154720821	154774815	53994	30089045	30195390	106345
X45	5	30091322	30136472	45150	154761863	154774815	12952
X45	6	163494295	163620207	125912	163494295	163620616	126321
X45	7	42071183	42160492	89309	42070252	42159463	89211
X45	8	124671283	124862560	191277	124804669	124947910	143241
X45	9	7939601	8022935	83334	7939672	7997399	57727
X45	11	44650927	45129274	478347	44698338	45044671	346333
X45	12	4298099	4330215	32116	4324976	4329977	5001
X45	13	37940514	37942995	2481	37430041	37943033	512992
X45	15				42437431	42922656	485225
X45	18	65930657	66132972	202315	65995883	66130015	134132
X06	1	20135822	20363654	227832	20138436	20141920	3484
X06	1	175700236	175727238	27002	175650399	175732289	81890
X06	1	240645826	240781448	135622	240646538	240763283	116745
X06	2	1321663	1675155	353492	1426346	1664654	238308
X06	2	146993207	146999990	6783	146971415	147004201	32786
X06	2				241541640	241595246	53606
X06	3	11625202	11744899	119697	11625202	11744584	119382
X06	4	19070123	19073685	3562	19057430	19101375	43945
X06	4	125419126	125513521	94395	125475882	125513521	37639
X06	4	184664543	184807842	143299	184668956	184790301	121345
X06	5	92685426	92687568	2142	92680028	92718948	38920
X06	6	6242593	6355602	113009	6229348	6310536	81188
X06	7	7939688	7939907	219	7925560	8025311	99751
X06	7	105910061	106075312	165251	105884848	105896527	11679
X06	8	20513428	20571619	58191	20559099	20590319	31220
X06	8	61981870	62108746	126876	61998100	62346192	348092
X06	8	139660695	139970419	309724	139764396	139821059	56663
X06	9	126860507	127300052	439545	126850913	126949213	98300
X06	12	4340608	4557597	216989	4416085	4500711	84626
X06	12	63785678	63805028	19350	63777954	63892595	114641
X06	12	130782820	131001280	218460	131145655	131288444	142789
X06	13	31994364	32016690	22326	31990598	32066605	76007
X06	15	58345001	58430763	85762	58338651	58461814	123163
X06	15	97875586	98101565	225979	97869259	97888848	19589
X06	16	77521178	77526335	5157	77522911	77527427	4516
X06	17	4660036	4874171	214135	4658971	4874215	215244
X06	18	6617744	6635823	18079	6603954	6649965	46011

X06	18	77459418	77869232	409814	76601547	76763842	162295
X06	21	41246661	41307457	60796	41260260	41305324	45064
X06	22	26791352	26791431	79	26789849	26792236	2387
X22	1	193827754	193829436	1682	193826947	193833027	6080
X22	1	230277115	230345325	68210	229798874	229869669	70795
X22	2	8551119	9094634	543515	8545555	9017034	471479
X22	2	37766497	37771234	4737	37718181	37777752	59571
X22	2	162975397	163041404	66007	162969662	163007420	37758
X22	3	29001627	29123737	122110	28993693	29086494	92801
X22	3	124661292	124756589	95297	124656695	124768366	111671
X22	4	25794869	25924291	129422	25721489	25745088	23599
X22	4	139433846	139617036	183190	139434076	139462440	28364
X22	4	187201211	187246449	45238	187213027	187315811	102784
X22	7	30454165	30559253	105088	30147299	30211788	64489
X22	8	136478555	136722584	244029	136570502	136743544	173042
X22	10	6286117	6298724	12607	6288755	6311749	22994
X22	10	77960360	77998443	38083	77962354	78063891	101537
X22	11	30762964	30831065	68101	30579485	30626660	47175
X22	12	85210329	85270228	59899	85195303	85550132	354829
X22	12	129633387	129675385	41998	129637348	129655509	18161
X22	13	95621197	95751017	129820	95616611	95717498	100887
X22	14	21619002	21631923	12921	21620695	21631651	10956
X22	15	38873115	39224334	351219	38868788	39222614	353826
X22	15	95933037	96721093	788056	96257703	96325643	67940
X22	17	716824	818466	101642	840805	1123682	282877
X22	17				76832133	77295774	463641
X22	19				5422553	5892486	469933
X22	21	16346881	16355020	8139	16346881	16361057	14176
X34	1	15610004	15811692	201688	15613478	15753398	139920
X34	1	162552738	162668336	115598	162551486	162724636	173150
X34	1	248108306	248152010	43704	248843314	248907678	64364
X34	3	56953174	57031977	78803	56942973	57032690	89717
X34	3	121939668	122022417	82749	121916325	122020194	103869
X34	3	194199814	194464045	264231	194260332	194459164	198832
X34	4	3919530	4034168	114638	3691191	4183931	492740
X34	4	172405998	172648301	242303	171963260	172031172	67912
X34	5	171139868	171174801	34933	171172825	171174351	1526
X34	6	104198257	104232520	34263	104198552	104222812	24260
X34	6	148917941	148921469	3528	148901519	148921825	20306
X34	7	85078623	85379359	300736	85111529	85365197	253668
X34	7	155531080	155840773	309693	155650354	155836562	186208
X34	8	139198704	139375834	177130	139202536	139286075	83539
X34	10	4431716	4517381	85665	4435485	4517381	81896
X34	11	72821069	73287121	466052	72957320	73337929	380609
X34	11	132931152	133318489	387337	132828835	132919385	90550
X34	12	117803515	118158758	355243	118010988	118047175	36187
X34	14	61817119	61917178	100059	61665229	61930678	265449
X34	15	25106101	25112623	6522	25105069	25114867	9798

X34	15	49191253	49220457	29204	49199392	49209632	10240
X34	16	86768472	86768959	487	86768472	86769827	1355
X34	17	57375766	57466504	90738	57371432	57421666	50234
X34	18	36946786	37146739	199953	37012708	37412228	399520
X34	19	49837439	50420112	582673	49867913	50140068	272155
X34	21	42507126	42651447	144321	42524741	42657187	132446
X34	22	26517370	26594197	76827	26520564	26589752	69188
X37	1	85939476	85943180	3704	85940105	85955772	15667
X37	2	115237760	115305050	67290	115298931	115306432	7501
X37	3	188298989	188690718	391729	188302851	188479543	176692
X37	5	90249105	90318391	69286	90263916	90494114	230198
X37	6	46034367	46114531	80164	46034491	46043561	9070
X37	7	142470574	142633833	163259	142474939	142509338	34399
X37	7	154413406	154414010	604	154413406	154414010	604
X37	9	10759908	10763165	3257	10752338	10837199	84861
X37	10	131053604	131279791	226187	131043467	131250665	207198
X37	11	34659833	34666863	7030	34660618	34680053	19435
X37	11	134172670	134434549	261879	134330418	134342005	11587
X37	12	11698217	11707406	9189	11687997	11708953	20956
X37	12	127661932	127729994	68062	127665366	127726384	61018
X37	13	22710477	22902249	191772	22782909	22827752	44843
X37	15	95476246	95629510	153264	95473492	95674002	200510
X37	16	84754964	84780613	25649	85955304	86009740	54436
X37	17	55593013	55777985	184972	55496093	55612236	116143
X37	18	55659197	55757544	98347	55695737	55699955	4218
X37	19	55173814	55176702	2888	55173844	55176262	2418
X37	20	5582598	5822368	239770	5633512	5843479	209967
X37	20	58406205	58406750	545	58243535	58476841	233306
X44	1	79483492	79540020	56528	79484287	79542356	58069
X44	1	216667294	216684321	17027	216236311	216701242	464931
X44	2	5469833	5507276	37443	5488174	5498227	10053
X44	2	230779625	230834675	55050	230822296	230829594	7298
X44	3	4417689	4485568	67879	4403614	4486303	82689
X44	4	34345958	34348295	2337	34323177	34466283	143106
X44	4	184584338	184587701	3363	184579914	184590150	10236
X44	6	146822249	146848013	25764	146944056	146944800	744
X44	9	94165651	94166861	1210	94049879	94188141	138262
X44	9	135133039	135133735	696	135132930	135343839	210909
X44	10	116152112	116364424	212312	116257523	116362835	105312
X44	11	87664122	87938024	273902	87606160	87937521	331361
X44	12	1878392	1968790	90398	1923040	1962759	39719
X44	14	81481849	81510740	28891	81480822	81520260	39438
X44	16	13873088	13901337	28249	13697975	13901337	203362
X44	16	86472884	86558013	85129	86474899	86511915	37016
X44	17	68103004	68103347	343	68103004	68103347	343
X44	18	1039426	1047261	7835	968021	1051047	83026
X44	19				5422553	5591735	169182

X44	19	46247641	46526950	279309	46361607	46472915	111308
X44	22	44643624	44912511	268887	44705361	44759155	53794
X69	1	10565488	11266251	700763	10814191	11514206	700015
X69	1	208170487	208584867	414380	208164620	208403258	238638
X69	2	234070789	234369487	298698	233929545	234107070	177525
X69	3	158722633	159055355	332722	158714381	159056759	342378
X69	4	6035861	6376471	340610	6235553	6441967	206414
X69	4	84809785	85048963	239178	84809102	85054547	245445
X69	4	188705868	188784610	78742	188680866	188730709	49843
X69	5	13741223	13748662	7439	13399475	13750898	351423
X69	6	2598145	2602357	4212	2599592	2602357	2765
X69	6	73923937	74216472	292535	74019879	74246392	226513
X69	7	141091985	141116780	24795	141098424	141224952	126528
X69	8	4070767	4071104	337	4067273	4096546	29273
X69	9				136569061	136658987	89926
X69	10	130552522	130843008	290486	130750248	130840798	90550
X69	11	36867813	36868287	474	36863551	36884595	21044
X69	11	131217695	131222575	4880	131204953	131224448	19495
X69	12	2861331	2925386	64055	2852249	2869552	17303
X69	14	23804719	24004429	199710	23964419	24002887	38468
X69	15	38974456	39015860	41404	38960882	39033746	72864
X69	15	100708857	100728663	19806	100691456	100728663	37207
X69	16	84800419	84970412	169993	86325241	86348568	23327
X69	17	9019269	9341204	321935	9145968	9336370	190402
X69	18	22718292	22775930	57638	22721515	22772210	50695
X69	19	56814440	56815923	1483	56806348	56815923	9575
X69	20	57838786	58113377	274591	57164080	58118584	954504
X82	1	48789172	48935977	146805	48558786	49033734	474948
X82	1	108621812	108644247	22435	108616600	108646557	29957
X82	1	248058130	248152010	93880			
X82	2	7957773	7967358	9585	7949693	7968275	18582
X82	2	47750740	48195264	444524	47703984	47814721	110737
X82	2	143722839	143726055	3216	143722839	143729878	7039
X82	2	229364186	229400777	36591	229300887	229631893	331006
X82	4	31279108	31389374	110266	31067347	31319406	252059
X82	4	190465333	190470163	4830			
X82	5	86403941	86541971	138030	86095152	86605943	510791
X82	6	2441986	2825591	383605	2460168	2608995	148827
X82	6	40546190	40599176	52986	40517063	40601537	84474
X82	6	153867035	153874482	7447	153775310	153955693	180383
X82	7	154531649	154617207	85558	154512110	154599918	87808
X82	8	19293565	19382890	89325	19326083	19392548	66465
X82	8	115866116	116023442	157326	115866116	116161903	295787
X82	9	125372279	125552744	180465	125391241	125606617	215376
X82	10	1665821	1720840	55019	1745254	1747614	2360
X82	10	133343436	133595963	252527	134295644	134309194	13550
X82	13	27410834	27614334	203500	27403167	27545885	142718

X82	13	98937421	98938919	1498	98922498	98952600	30102
X82	15	32300235	32303657	3422	32299953	32303657	3704
X82	15	92079944	92174033	94089	92087341	92090795	3454
X82	20	1814266	1931582	117316	1886493	1917290	30797
X82	22	44523593	44523900	307	44514267	44527714	13447
X91	1	58859948	58947574	87626	58868360	59114438	246078
X91	1	168312082	168432357	120275	168311842	168407955	96113
X91	1	241774286	242017826	243540	241706533	242025098	318565
X91	2	205362217	205375579	13362	205358028	205803707	445679
X91	3	19568009	19624459	56450	19568009	19653387	85378
X91	4	113747792	113796033	48241	113740351	113825869	85518
X91	5	8188763	8282810	94047	8179705	8411931	232226
X91	6	168896772	169126013	229241	168895058	168956803	61745
X91	7	134225827	134273949	48122	134225827	134288628	62801
X91	7	147753561	147754816	1255	147753884	147754816	932
X91	8	53830883	53843748	12865	53829203	53843748	14545
X91	8	139235991	139304863	68872	139232371	139255690	23319
X91	10	102609006	102753788	144782	102707526	102744376	36850
X91	11	82340435	82441634	101199	82352407	82411471	59064
X91	11	131779557	131901805	122248	131779557	131904225	124668
X91	12	118194643	118367536	172893	118365682	118375486	9804
X91	13	26595081	26764008	168927	26600933	26603361	2428
X91	13	104036436	104117720	81284	104036311	104067224	30913
X91	14	28974396	28995903	21507	28913965	29030750	116785
X91	15	67370389	67523043	152654	67407899	67451954	44055
X91	15	94134022	94205530	71508	94106508	94205530	99022
X91	16	89307255	89493405	186150			
X91	17	10006770	10096877	90107	10006770	10021341	14571
X91	18	8763794	8838397	74603	8784612	8841460	56848
X91	18	75697212	75825410	128198	75780436	75831848	51412
X91	19	45410444	45972408	561964	45413576	45962799	549223

4.4 Discussion

The knowledge of complete rather than fragmental haplotypes is critical in precision medicine and will also advance our understanding in many areas of genetic research. Current diploid-based haplotyping methods are costly or only produce small haplotype fragments, whereas, haploid-based alternatives using gamete data may break through such boundary and infer chromosome-scale haplotypes for individual genomes. Two existing haploid-based phasing methods rely on the accurate detection of crossover positions in gamete genomes. However, complex chromosomal regions with many repetitive DNA elements, such as large segmental duplications, make it challenging to infer the positions of recombination. The existence of missing and ambiguous genotype calls makes the task even harder, leading to inaccurately phased chromosomes. In this study, we developed a highly efficient and fully automatic method, *Hapi*, that only requires 3 to 5 gametes to correctly reconstruct high-resolution chromosomal haplotypes. The *Hapi* method circumvents the direct inference of crossovers, and infers chromosome-scale haplotypes through a 3-step strategy. In the first step, hetSNPs with erroneous genotypes are removed by an HMM. High-quality hetSNPs are then selected to form a framework and missing genotypes in the framework are imputed iteratively. In the second step, draft haplotypes with high level of confidence are construct by the majority voting and proofread by the MPR. In the third step, the draft haplotypes are used as a blueprint to derive gamete-specific chromosomal haplotypes, which are eventually used to assemble the consensus high-resolution chromosome-length haplotypes. The phasing

pipeline can be easily implemented by the functions built in the *Hapi* package written in R language.

Using simulated and real datasets (maize microspores and human sperms), we demonstrated that *Hapi* outperforms the other two methods in phasing accuracy, reliability, and cost efficiency. To achieve the same level of phasing accuracy, *Hapi* required less gametes and can tolerate more missing genotypes than the other two methods, not only because of the ameliorated phasing strategy, but also due to the equipped algorithms handling imperfect data (missing or erroneous data). When different numbers of gametes are used for phasing, *Hapi* performed consistently well but the performances of *OVA* and *PHMM* fluctuated wildly, indicating the new *Hapi* method handles ambiguous data from a small number of gametes very well and produces reliable phasing results. In addition, since ambiguous genotype data sometimes occur at the chromosome ends, a special capping algorithm has been designed in the *Hapi* method to polish the process of constructing draft haplotypes.

Our study also indicated that 3 gametes may be enough to reconstruct chromosome-length haplotypes by *Hapi* if the genotype data are of high quality, i.e., with few missing or erroneous data. In the simulation study, the *Hapi* method used 3 gametes to correctly infer the chromosomal haplotypes 99 out of 100 replicates with low marker density (5000 hetSNPs) and low missing genotype rate (0~10%). When the number of makers is increased to 50,000 hetSNPs, by using 3 gametes, the *Hapi* method can correctly infer the chromosomal haplotypes for all 100 replicates with the missing rate of no more than 20% and only returned incorrectly phased chromosomes in 1 or 2 out of 100 replicates even

the missing genotype rate is increased to 70%. It should be noted that using 3 gametes may fail in a special scenario when two crossovers within a very small region occur in two gametes, respectively. The reason is that, in the step of proofreading draft haplotypes, small blocks with little genotype information are excluded from the draft haplotypes by default, assuming the probability of having multiple crossovers within these blocks on more than one gamete is low. In this specific but rare scenario, removal of such blocks may lead to the wrong determination of the major link type and thereafter the haplotypes. If only 3 gametes are available, it is recommended to implement the *Hapi* method with and without removing blocks in constructing draft haplotypes, and check the consistency in results from two different settings.

Unlike other haplotype phasing algorithms that demand sequencing long-reads or linked-reads in diploid cells, the *Hapi* method can analyze hetSNPs data of single gamete cells generated using any genotyping platforms. Either nucleobases (A/T/C/G) or binary code (0/1) can be used as the input genotypic data for hetSNPs in gamete genomes. Advanced technologies, such as 10X Genomics linked-read sequencing, are not necessary for the *Hapi* method, but may be used as ancillary approaches to generate designated long-range haplotype fragments for the complex and challenging genomic regions, further perfecting the chromosomal haplotypes inferred by the *Hapi* method.

Abnormality in meiotic recombination is the leading cause of miscarriage and birth defects. Studies have shown that reduced recombination activity could be associated with male infertility and sperm aneuploidy. On the other hand, abnormally increased recombination activity may indicate the existence of deleterious stresses for which

activated recombination is an evolved response. An important application of the *Hapi* package is to implement the crossover analysis module to derive the map of recombination in gametes based on the inferred chromosome-length haplotypes. This function has the potential to be applied in clinical labs to manage human diseases that are associated with abnormal recombination, and can also be used to monitor the crossovers on plant genomes to facilitate more rapid introgression of target genes or to break up undesirable linkages for crop improvement.

Conclusion

In this dissertation, three bioinformatic and statistical methods were developed to advance the study of human complex traits. In the second chapter, an easy-to-use R package, named *GDCRNATools* was developed to download, organize, and analyze RNA-seq and miRNA-seq data in GDC to construct the lncRNA-mRNA related ceRNAs regulatory networks. We believe that the *GDCRNATools* will gain ground in cancer research for deciphering the crosstalk among multiple RNA species and their regulatory mechanisms in cancer. In the third chapter, we developed a BLUP-HAT method and used transcriptomic, miRNA, and epigenomic data from the TCGA-PRAD project in GDC to demonstrate that the predictive power for prognosis of prostate cancer can be greatly improved by using a large number of genes selected from transcriptome and can be further improved by the integration of other omic data into the BLUP model. This study provided a promising strategy to guide the development of multi-omic signatures for the accurate diagnosis and prognosis of cancer. In the fourth chapter, an R package, named *Hapi* was developed to implement our new algorithm for high-resolution chromosome-length haplotype phasing by using genotype data of a few single gamete cells. The gamete-based haplotyping strategy is probably the most feasible method for large-scale

applications in many areas of genetic studies and precision medicine. Recombination events in each gamete cell can also be detected by the crossover analysis module in *Hapi* to study human diseases that are associated with abnormal recombination.

We anticipated that the advanced tools developed in the dissertation will greatly facilitate the understanding of the genetic and molecular basis of human complex traits.

Bibliography

Abraham, G., Havulinna, A.S., Bhalala, O.G., Byars, S.G., De Livera, A.M., Yetukuri, L., Tikkanen, E., Perola, M., Schunkert, H., and Sijbrands, E.J. (2016). Genomic prediction of coronary heart disease. European heart journal *37*, 3267-3278.

Abraham, G., Tye-Din, J.A., Bhalala, O.G., Kowalczyk, A., Zobel, J., and Inouye, M. (2014). Accurate and robust genomic prediction of celiac disease using statistical learning. PLoS genetics *10*, e1004137.

Allen, H.L., Estrada, K., Lettre, G., Berndt, S.I., Weedon, M.N., Rivadeneira, F., Willer, C.J., Jackson, A.U., Vedantam, S., and Raychaudhuri, S. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature *467*, 832.

Ananthanarayanan, V., Deaton, R.J., Yang, X.J., Pins, M.R., and Gann, P.H. (2005). Alpha-methylacyl-CoA racemase (AMACR) expression in normal prostatic glands and high-grade prostatic intraepithelial neoplasia (HGPIN): Association with diagnosis of prostate cancer. The Prostate *63*, 341-346.

Browning, S.R., and Browning, B.L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. The American Journal of Human Genetics *81*, 1084-1097.

Cairns, P., Esteller, M., Herman, J.G., Schoenberg, M., Jeronimo, C., Sanchez-Cespedes, M., Chow, N.-H., Grasso, M., Wu, L., and Westra, W.B. (2001). Molecular detection of prostate cancer in urine by GSTP1 hypermethylation. Clinical Cancer Research *7*, 2727-2730.

Calus, M., De Roos, A., and Veerkamp, R. (2008). Accuracy of genomic selection using different methods to define haplotypes. Genetics *178*, 553-561.

Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B.A., Jacobsen, A., Byrne, C.J., Heuer, M.L., and Larsson, E. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data (AACR).

Chen, G.-B., Lee, S.H., Brion, M.-J.A., Montgomery, G.W., Wray, N.R., Radford-Smith, G.L., Visscher, P.M., and Consortium, I.I.G. (2014). Estimation and partitioning of (co) heritability of inflammatory bowel disease from GWAS and immunochip data. Human molecular genetics *23*, 4710-4720.

Chiang, C., Layer, R.M., Faust, G.G., Lindberg, M.R., Rose, D.B., Garrison, E.P., Marth, G.T., Quinlan, A.R., and Hall, I.M. (2015). SpeedSeq: ultra-fast personal genome analysis and interpretation. Nature methods *12*, 966.

Chou, C.-H., Shrestha, S., Yang, C.-D., Chang, N.-W., Lin, Y.-L., Liao, K.-W., Huang, W.-C., Sun, T.-H., Tu, S.-J., and Lee, W.-H. (2017). miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. Nucleic acids research *46*, D296-D302.

Chun, F.K.H., Haese, A., Ahyai, S.A., Walz, J., Suardi, N., Capitanio, U., Graefen, M., Erbersdobler, A., Huland, H., and Karakiewicz, P.I. (2008). Critical assessment of tools to predict clinically insignificant prostate cancer at radical prostatectomy in contemporary men. Cancer *113*, 701-709.

Colaprico, A., Silva, T.C., Olsen, C., Garofano, L., Cava, C., Garolini, D., Sabedot, T.S., Malta, T.M., Pagnotta, S.M., and Castiglioni, I. (2015). TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic acids research *44*, e71-e71.

Crawford, D.C., and Nickerson, D.A. (2005). Definition and clinical importance of haplotypes. Annu Rev Med *56*, 303-320.

Cuyabano, B.C., Su, G., and Lund, M.S. (2014). Genomic prediction of genetic merit using LD-based haplotypes in the Nordic Holstein population. BMC genomics *15*, 1171.

Cuzick, J., Swanson, G.P., Fisher, G., Brothman, A.R., Berney, D.M., Reid, J.E., Mesher, D., Speights, V., Stankiewicz, E., and Foster, C.S. (2011). Prognostic value of an RNA expression signature derived from cell cycle proliferation genes in patients with prostate cancer: a retrospective study. The lancet oncology *12*, 245-255.

De Bakker, P.I., McVean, G., Sabeti, P.C., Miretti, M.M., Green, T., Marchini, J., Ke, X., Monsuur, A.J., Whittaker, P., and Delgado, M. (2006). A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. Nature genetics *38*, 1166.

Delaneau, O., Howie, B., Cox, A.J., Zagury, J.-F., and Marchini, J. (2013a). Haplotype estimation using sequencing reads. The American Journal of Human Genetics *93*, 687-696.

Delaneau, O., Marchini, J., and Zagury, J.-F. (2012). A linear complexity phasing method for thousands of genomes. Nature methods *9*, 179.

Delaneau, O., Zagury, J.-F., and Marchini, J. (2013b). Improved whole-chromosome phasing for disease and population genetic studies. Nature methods *10*, 5.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., Del Angel, G., Rivas, M.A., and Hanna, M. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature genetics *43*, 491.

Drysdale, C.M., McGraw, D.W., Stack, C.B., Stephens, J.C., Judson, R.S., Nandabalan, K., Arnold, K., Ruano, G., and Liggett, S.B. (2000). Complex promoter and coding region β2-adrenergic receptor haplotypes alter receptor expression and predict in vivo responsiveness. Proceedings of the National Academy of Sciences *97*, 10483-10488.

Edge, P., Bafna, V., and Bansal, V. (2017). HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. Genome research *27*, 801-812.

Edriss, V., Fernando, R.L., Su, G., Lund, M.S., and Guldbrandtsen, B. (2013). The effect of using genealogy-based haplotypes for genomic prediction. Genetics Selection Evolution *45*, 5.

Fan, H.C., Wang, J., Potanina, A., and Quake, S.R. (2011). Whole-genome molecular haplotyping of single cells. Nature biotechnology *29*, 51.

Faust, G.G., and Hall, I.M. (2014). SAMBLASTER: fast duplicate marking and structural variant read extraction. Bioinformatics *30*, 2503-2505.

Filella, X., and Giménez, N. (2013). Evaluation of [−2] proPSA and Prostate Health Index (phi) for the detection of prostate cancer: a systematic review and meta-analysis. Clinical chemistry and laboratory medicine *51*, 729-739.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. Journal of statistical software *33*, 1.

Furió-Tarí, P., Tarazona, S., Gabaldón, T., Enright, A.J., and Conesa, A. (2016). spongeScan: A web for detecting microRNA binding elements in lncRNA sequences. Nucleic acids research *44*, W176-W180.

Gao, J., Aksoy, B.A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S.O., Sun, Y., Jacobsen, A., Sinha, R., and Larsson, E. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal *6*, pl1-pl1.

George, E.I., and McCulloch, R.E. (1993). Variable selection via Gibbs sampling. Journal of the American Statistical Association *88*, 881-889.

Goldman, M., Craft, B., Zhu, J., and Haussler, D. (2017). The UCSC Xena system for cancer genomics data visualization and interpretation (AACR).

Grossman, R.L., Heath, A.P., Ferretti, V., Varmus, H.E., Lowy, D.R., Kibbe, W.A., and Staudt, L.M. (2016). Toward a shared vision for cancer genomic data. New England Journal of Medicine *375*, 1109-1112.

Hayes, B., and Goddard, M. (2001). Prediction of total genetic value using genome-wide dense marker maps. Genetics *157*, 1819-1829.

He, Y., Jing, Y., Wei, F., Tang, Y., Yang, L., Luo, J., Yang, P., Ni, Q., Pang, J., and Liao, Q. (2018). Long non-coding RNA PVT1 predicts poor prognosis and induces radioresistance by regulating DNA repair and cell apoptosis in nasopharyngeal carcinoma. Cell death & disease *9*, 235.

Henderson, C.R. (1975). Best linear unbiased estimation and prediction under a selection model. Biometrics, 423-447.

Hess, M., Druet, T., Hess, A., and Garrick, D. (2017). Fixed-length haplotypes can improve genomic prediction accuracy in an admixed dairy cattle population. Genetics Selection Evolution *49*, 54.

Hessels, D., and Schalken, J.A. (2009). The use of PCA3 in the diagnosis of prostate cancer. Nature Reviews Urology *6*, 255.

Hou, Y., Fan, W., Yan, L., Li, R., Lian, Y., Huang, J., Li, J., Xu, L., Tang, F., and Xie, X.S. (2013). Genome analyses of single human oocytes. Cell *155*, 1492-1506.

Howie, B.N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS genetics *5*, e1000529.

Huang, B., Liu, C., Wu, Q., Zhang, J., Min, Q., Sheng, T., Wang, X., and Zou, Y. (2017). Long non-coding RNA NEAT1 facilitates pancreatic cancer progression through negative modulation of miR-506-3p. Biochemical and biophysical research communications *482*, 828-834.

Huang, C., Yu, W., Wang, Q., Cui, H., Wang, Y., Zhang, L., Han, F., and Huang, T. (2015a). Increased expression of the lncRNA PVT1 is associated with poor prognosis in pancreatic cancer patients. Minerva medica *106*, 143-149.

Huang, J., Howie, B., McCarthy, S., Memari, Y., Walter, K., Min, J.L., Danecek, P., Malerba, G., Trabetti, E., and Zheng, H.-F. (2015b). Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. Nature communications *6*, 8111.

Iden, M., Fye, S., Li, K., Chowdhury, T., Ramchandran, R., and Rader, J.S. (2016). The lncRNA PVT1 contributes to the cervical cancer phenotype and associates with poor patient prognosis. PLoS One *11*, e0156274.

Jeggari, A., Marks, D.S., and Larsson, E. (2012). miRcode: a map of putative microRNA target sites in the long non-coding transcriptome. Bioinformatics *28*, 2062-2063.

Jensen, M.A., Ferretti, V., Grossman, R.L., and Staudt, L.M. (2017). The NCI Genomic Data Commons as an engine for precision medicine. Blood *130*, 453-459.

Jiang, P., Wu, X., Wang, X., Huang, W., and Feng, Q. (2016). NEAT1 upregulates EGCG-induced CTR1 to enhance cisplatin sensitivity in lung cancer cells. Oncotarget *7*, 43337.

Karatzoglou, A., Smola, A., Hornik, K., and Zeileis, A. (2004). kernlab-an S4 package for kernel methods in R. Journal of statistical software *11*, 1-20.

Kattan, M.W., Eastham, J.A., Wheeler, T.M., Maru, N., Scardino, P.T., Erbersdobler, A., Graefen, M., Huland, H., Koh, H., and Shariat, S.F. (2003). Counseling men with prostate cancer: a nomogram for predicting the presence of small, moderately differentiated, confined tumors. The Journal of urology *170*, 1792-1797.

Kirkness, E.F., Grindberg, R.V., Yee-Greenbaum, J., Marshall, C.R., Scherer, S.W., Lasken, R.S., and Venter, J.C. (2013). Sequencing of isolated sperm cells for direct haplotyping of a human genome. Genome research *23*, 826-832.

Klein, E.A., Cooperberg, M.R., Magi-Galluzzi, C., Simko, J.P., Falzarano, S.M., Maddala, T., Chan, J.M., Li, J., Cowan, J.E., and Tsiatis, A.C. (2014). A 17-gene assay to predict prostate cancer aggressiveness in the context of Gleason grade heterogeneity, tumor multifocality, and biopsy undersampling. European urology *66*, 550-560.

Kong, R., Zhang, E.-b., Yin, D.-d., You, L.-h., Xu, T.-p., Chen, W.-m., Xia, R., Wan, L., Sun, M., and Wang, Z.-x. (2015). Long noncoding RNA PVT1 indicates a poor prognosis of gastric cancer and promotes cell proliferation through epigenetically regulating p15 and p16. Molecular cancer *14*, 82.

Kozomara, A., and Griffiths-Jones, S. (2013). miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic acids research *42*, D68-D73.

Kretschmer, A., and Tilki, D. (2017). Biomarkers in prostate cancer–Current clinical utility and future perspectives. Critical reviews in oncology/hematology *120*, 180-193.

Lambert, J.-C., Grenier-Boley, B., Harold, D., Zelenika, D., Chouraki, V., Kamatani, Y., Sleegers, K., Ikram, M., Hiltunen, M., and Reitz, C. (2013). Genome-wide haplotype association study identifies the FRMD4A gene as a risk locus for Alzheimer's disease. Molecular psychiatry *18*, 461.

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC bioinformatics *9*, 559.

Law, C.W., Chen, Y., Shi, W., and Smyth, G.K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome biology *15*, R29.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics *25*, 1754-1760.

Li, J.-H., Liu, S., Zhou, H., Qu, L.-H., and Yang, J.-H. (2013). starBase v2. 0: decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data. Nucleic acids research *42*, D92-D97.

Li, R., Qu, H., Wang, S., Wei, J., Zhang, L., Ma, R., Lu, J., Zhu, J., Zhong, W.-D., and Jia, Z. (2018). GDCRNATools: an R/Bioconductor package for integrative analysis of lncRNA, miRNA and mRNA data in GDC. Bioinformatics, bty124-bty124.

Li, X., Li, L., and Yan, J. (2015). Dissecting meiotic recombination based on tetrad analysis by single-microspore sequencing in maize. Nature communications *6*, 6648.

Li, Y., Willer, C.J., Ding, J., Scheet, P., and Abecasis, G.R. (2010). MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genetic epidemiology *34*, 816-834.

Liu, X.-h., Sun, M., Nie, F.-q., Ge, Y.-b., Zhang, E.-b., Yin, D.-d., Kong, R., Xia, R., Lu, K.-h., and Li, J.-h. (2014). Lnc RNA HOTAIR functions as a competing endogenous RNA to regulate HER2 expression by sponging miR-331-3p in gastric cancer. Molecular cancer *13*, 92.

Loh, P.-R., Danecek, P., Palamara, P.F., Fuchsberger, C., Reshef, Y.A., Finucane, H.K., Schoenherr, S., Forer, L., McCarthy, S., and Abecasis, G.R. (2016). Reference-based phasing using the Haplotype Reference Consortium panel. Nature genetics *48*, 1443.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology *15*, 550.

Lu, S., Zong, C., Fan, W., Yang, M., Li, J., Chapman, A.R., Zhu, P., Hu, X., Xu, L., and Yan, L. (2012). Probing meiotic recombination and aneuploidy of single sperm cells by whole-genome sequencing. Science *338*, 1627-1630.

Ma, L., Xiao, Y., Huang, H., Wang, Q., Rao, W., Feng, Y., Zhang, K., and Song, Q. (2010). Direct determination of molecular haplotypes by chromosome microdissection. Nature methods *7*, 299.

Makowsky, R., Pajewski, N.M., Klimentidis, Y.C., Vazquez, A.I., Duarte, C.W., Allison, D.B., and de Los Campos, G. (2011). Beyond missing heritability: prediction of complex traits. PLoS genetics *7*, e1002051.

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M., Fuchsberger, C., Danecek, P., and Sharp, K. (2016). A reference panel of 64,976 haplotypes for genotype imputation. Nature genetics *48*, 1279.

Meuwissen, T., Hayes, B., and Goddard, M. (2001). Prediction of total genetic value using genome-wide dense marker maps. Genetics *157*, 1819-1829.

Musone, S.L., Taylor, K.E., Lu, T.T., Nititham, J., Ferreira, R.C., Ortmann, W., Shifrin, N., Petri, M.A., Kamboh, M.I., and Manzi, S. (2008). Multiple polymorphisms in the TNFAIP3 region are independently associated with systemic lupus erythematosus. Nature genetics *40*, 1062.

Nakagawa, T., Kollmeyer, T.M., Morlan, B.W., Anderson, S.K., Bergstralh, E.J., Davis, B.J., Asmann, Y.W., Klee, G.G., Ballman, K.V., and Jenkins, R.B. (2008). A tissue biomarker panel predicting systemic progression after PSA recurrence post-definitive prostate cancer therapy. PloS one *3*, e2318.

Nakanishi, H., Wang, X., Ochiai, A., Trpkov, K., Yilmaz, A., Donnelly, J.B., Davis, J.W., Troncoso, P., and Babaian, R.J. (2007). A nomogram for predicting low-volume/low-grade prostate cancer. Cancer *110*, 2441-2447.

O'Connell, J., Sharp, K., Shrine, N., Wain, L., Hall, I., Tobin, M., Zagury, J.-F., Delaneau, O., and Marchini, J. (2016). Haplotype estimation for biobank-scale data sets. Nature genetics *48*, 817.

Paci, P., Colombo, T., and Farina, L. (2014). Computational analysis identifies a sponge interaction network between long non-coding RNAs and messenger RNAs in human breast cancer. BMC systems biology *8*, 83.

Pérez, P., and de Los Campos, G. (2014). Genome-wide regression & prediction with the BGLR statistical package. Genetics, genetics. 114.164442.

Peters, B.A., Kermani, B.G., Sparks, A.B., Alferov, O., Hong, P., Alexeev, A., Jiang, Y., Dahl, F., Tang, Y.T., and Haas, J. (2012). Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. Nature *487*, 190.

Petersdorf, E.W., Malkki, M., Gooley, T.A., Martin, P.J., and Guo, Z. (2007). MHC haplotype matching for unrelated hematopoietic cell transplantation. PLoS medicine *4*, e8.

Porubsky, D., Garg, S., Sanders, A.D., Korbel, J.O., Guryev, V., Lansdorp, P.M., and Marschall, T. (2017). Dense and accurate whole-chromosome haplotyping of individual genomes. Nature Communications *8*, 1293.

Porubský, D., Sanders, A.D., van Wietmarschen, N., Falconer, E., Hills, M., Spierings, D.C., Bevova, M.R., Guryev, V., and Lansdorp, P.M. (2016). Direct chromosome-length haplotyping by single-cell sequencing. Genome research *26*, 1565-1574.

Rhodes, D.R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., Barrette, T., Pander, A., and Chinnaiyan, A.M. (2004). ONCOMINE: a cancer microarray database and integrated data-mining platform. Neoplasia *6*, 1-6.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139-140.

Samur, M.K. (2014). RTCGAToolbox: a new tool for exporting TCGA Firehose data. PloS one *9*, e106397.

Scheet, P., and Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. The American Journal of Human Genetics *78*, 629-644.

Selvaraj, S., Dixon, J.R., Bansal, V., and Ren, B. (2013). Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. Nature biotechnology *31*, 1111.

Siegel, R.L., Miller, K.D., and Jemal, A. (2018). Cancer statistics, 2018. CA: a cancer journal for clinicians *68*, 7-30.

Speed, D., and Balding, D.J. (2014). MultiBLUP: improved SNP-based prediction for complex traits. Genome research *24*, 1550-1557.

Stephens, M., and Scheet, P. (2005). Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. The American Journal of Human Genetics *76*, 449-462.

Stephens, M., Smith, N.J., and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. The American Journal of Human Genetics *68*, 978-989.

Steyerberg, E., Roobol, M., Kattan, M., Van der Kwast, T., De Koning, H., and Schröder, F. (2007). Prediction of indolent prostate cancer: validation and updating of a prognostic nomogram. The Journal of urology *177*, 107-112.

Sun, C., Li, S., Zhang, F., Xi, Y., Wang, L., Bi, Y., and Li, D. (2016). Long non-coding RNA NEAT1 promotes non-small cell lung cancer progression through regulation of miR-377-3p-E2F3 pathway. Oncotarget *7*, 51784.

Tarasov, A., Vilella, A.J., Cuppen, E., Nijman, I.J., and Prins, P. (2015). Sambamba: fast processing of NGS alignment formats. Bioinformatics *31*, 2032-2034.

Thompson, I.M., Pauler, D.K., Goodman, P.J., Tangen, C.M., Lucia, M.S., Parnes, H.L., Minasian, L.M., Ford, L.G., Lippman, S.M., and Crawford, E.D. (2004). Prevalence of prostate cancer among men with a prostate-specific antigen level≤ 4.0 ng per milliliter. New England Journal of Medicine *350*, 2239-2246.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B (Methodological), 267-288.

Tomlins, S.A., Rhodes, D.R., Perner, S., Dhanasekaran, S.M., Mehra, R., Sun, X.-W., Varambally, S., Cao, X., Tchinda, J., and Kuefer, R. (2005). Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. science *310*, 644-648.

Trégouët, D.-A., König, I.R., Erdmann, J., Munteanu, A., Braund, P.S., Hall, A.S., Großhennig, A., Linsel-Nitschke, P., Perret, C., and DeSuremain, M. (2009). Genome-wide haplotype association study identifies the SLC22A3-LPAL2-LPA gene cluster as a risk locus for coronary artery disease. Nature genetics *41*, 283.

VanRaden, P.M. (2008). Efficient methods to compute genomic predictions. Journal of dairy science *91*, 4414-4423.

Wan, Y.-W., Allen, G.I., and Liu, Z. (2015). TCGA2STAT: simple TCGA data access for integrated statistical analysis in R. Bioinformatics *32*, 952-954.

Wang, S.-H., Ma, F., Tang, Z.-h., Wu, X.-C., Cai, Q., Zhang, M.-D., Weng, M.-Z., Zhou, D., Wang, J.-D., and Quan, Z.-W. (2016). Long non-coding RNA H19 regulates FOXM1 expression by competitively binding endogenous miR-342-3p in gallbladder cancer. Journal of Experimental & Clinical Cancer Research *35*, 160.

Wang, S.Y., Cowan, J.E., Cary, K.C., Chan, J.M., Carroll, P.R., and Cooperberg, M.R. (2014). Limited ability of existing nomograms to predict outcomes in men undergoing active surveillance for prostate cancer. BJU international *114*.

Wehrens, R., and Mevik, B.-H. (2007). The pls package: principal component and partial least squares regression in R.

Xu, S. (2013). Mapping quantitative trait loci by controlling polygenic background effects. Genetics *195*, 1209-1222.

Xu, S. (2017). Predicted Residual Error Sum of Squares of Mixed Models: An Application for Genomic Prediction. G3: Genes, Genomes, Genetics *7*, 895-909.

Yang, C., Li, Z., Li, Y., Xu, R., Wang, Y., Tian, Y., and Chen, W. (2017). Long non-coding RNA NEAT1 overexpression is associated with poor prognosis in cancer patients: a systematic review and meta-analysis. Oncotarget *8*, 2672.

Yang, H., Chen, X., and Wong, W.H. (2011). Completely phased genome sequencing through chromosome sorting. Proceedings of the National Academy of Sciences *108*, 12-17.

Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., and Montgomery, G.W. (2010). Common SNPs explain a large proportion of the heritability for human height. Nature genetics *42*, 565.

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. Omics: a journal of integrative biology *16*, 284-287.

Yu, G., Wang, L.-G., Yan, G.-R., and He, Q.-Y. (2014). DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. Bioinformatics *31*, 608-609.

Yu, J., Pressoir, G., Briggs, W.H., Bi, I.V., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., and Holland, J.B. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature genetics *38*, 203.

Zeng, P., and Zhou, X. (2017). Non-parametric genetic prediction of complex traits with latent Dirichlet process regression models. Nature communications *8*, 456.

Zhang, M., Wang, W., Li, T., Yu, X., Zhu, Y., Ding, F., Li, D., and Yang, T. (2016). Long noncoding RNA SNHG1 predicts a poor prognosis and promotes hepatocellular carcinoma tumorigenesis. Biomedicine & pharmacotherapy *80*, 73-79.

Zhou, X., Carbonetto, P., and Stephens, M. (2013). Polygenic modeling with Bayesian sparse linear mixed models. PLoS genetics *9*, e1003264.

Zhu, Y., Qiu, P., and Ji, Y. (2014). TCGA-assembler: open-source software for retrieving and processing TCGA data. Nature methods *11*, 599.

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) *67*, 301-320.