

5-2017

## Statistical Methods for Two Problems in Cancer Research: Analysis of RNA-seq Data from Archival Samples and Characterization of Onset of Multiple Primary Cancers

Jialu Li

Follow this and additional works at: [https://digitalcommons.library.tmc.edu/utgsbs\\_dissertations](https://digitalcommons.library.tmc.edu/utgsbs_dissertations)



Part of the [Bioinformatics Commons](#), [Biostatistics Commons](#), [Computational Biology Commons](#), [Genomics Commons](#), [Medicine and Health Sciences Commons](#), [Statistical Models Commons](#), and the [Survival Analysis Commons](#)

---

### Recommended Citation

Li, Jialu, "Statistical Methods for Two Problems in Cancer Research: Analysis of RNA-seq Data from Archival Samples and Characterization of Onset of Multiple Primary Cancers" (2017). *The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences Dissertations and Theses (Open Access)*. 740.

[https://digitalcommons.library.tmc.edu/utgsbs\\_dissertations/740](https://digitalcommons.library.tmc.edu/utgsbs_dissertations/740)

This Dissertation (PhD) is brought to you for free and open access by the The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences at DigitalCommons@TMC. It has been accepted for inclusion in The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences Dissertations and Theses (Open Access) by an authorized administrator of DigitalCommons@TMC. For more information, please contact [digitalcommons@library.tmc.edu](mailto:digitalcommons@library.tmc.edu).

**STATISTICAL METHODS FOR TWO PROBLEMS IN CANCER RESEARCH:  
ANALYSIS OF RNA-SEQ DATA FROM ARCHIVAL SAMPLES AND  
CHARACTERIZATION OF ONSET OF MULTIPLE PRIMARY CANCERS**

by

Jialu Li, M.S.

APPROVED:

---

Wenyi Wang, Ph.D., Supervisory Professor

---

Guillermina Lozano, Ph.D.

---

Jeffrey S. Morris, Ph.D.

---

Jing Ning, Ph.D.

---

Han Liang, Ph.D.

---

Paul A. Scheet, Ph.D.

APPROVED:

---

Dean, The University of Texas  
MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences

**STATISTICAL METHODS FOR TWO PROBLEMS IN CANCER  
RESEARCH: ANALYSIS OF RNA-SEQ DATA FROM ARCHIVAL  
SAMPLES AND CHARACTERIZATION OF ONSET OF MULTIPLE  
PRIMARY CANCERS**

A

DISSERTATION

Presented to the Faculty of

The University of Texas

MD Anderson Cancer Center UTHealth

Graduate School of Biomedical Sciences

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

by

Jialu Li, M.S.

Houston, Texas

May, 2017

## ACKNOWLEDGEMENTS

First and foremost, I want to give my utmost gratitude to my advisor, Dr. Wenyi Wang, for her excellent mentoring, generous support and kind encouragement for my quantitative training and dissertation research. Her ingenuity, optimistic and enthusiastic personality, and her dedication to scientific research and education have set a real role model for me. It is my great honor and pleasure to study with her in the past four years, and the experience will be extremely valuable for my future career.

I would also like to convey my sincere gratitude to my advisory committee members Drs. Guillermina Lozano, Jeffrey S. Morris, Han Liang, Jing Ning, and Paul A. Scheet. Their incisive advising and patient guidance greatly help improve my dissertation work.

I want to thank our collaborators Drs. Louise. C Strong, W. Fraser Symmans, Chunxiao Fu for nicely providing Li-Fraumeni syndrome data and archival tumor RNA sequencing data that make my PhD research a more complete story. I also want to sincerely thank Drs. Terence P. Speed, Jing Ning, Seung Jun Shin for their insightful comments and critical instructions on my data analysis and methodology development.

I thank Dr. William W. Mattox, Dr. Andrew Bean, Ms. Lourdes Perez, Ms. Brenda Gaughan and other GSBS staffs for their help and support for my Ph.D study in the past five years.

I also want to thank all members in Dr. Wenyi Wang's group: Gang Peng, Elissa B. Dodd, Xuedong Pan, Shaolong Cao, Zeya Wang, Rongjie Liu and Ruoxuan Tian. I learned a lot from discussion with them.

Lastly, I would like to express my gratitude to my parents. Without their unconditional love and support, I cannot be here to pursue what I want.



**STATISTICAL METHODS FOR TWO PROBLEMS IN CANCER RESEARCH:  
ANALYSIS OF RNA-SEQ DATA FROM ARCHIVAL SAMPLES AND  
CHARACTERIZATION OF ONSET OF MULTIPLE PRIMARY CANCERS**

Jialu Li, M.S.

Advisory Professor: Wenyi Wang, Ph.D.

My dissertation is focused on quantitative methodology development and application for two important topics in translational and clinical cancer research.

The first topic was motivated by the challenge of applying transcriptome sequencing (RNA-seq) to formalin-fixation and paraffin-embedding (FFPE) tumor samples for reliable diagnostic development. We designed a biospecimen study to directly compare gene expression results from different protocols to prepare libraries for RNA-seq from human breast cancer tissues, with randomization to fresh-frozen (FF) or FFPE conditions. To comprehensively evaluate the FFPE RNA-seq data quality for expression profiling, we developed multiple computational methods for assessment, such as the uniformity and continuity of coverage, the variance and correlation of overall gene expression, patterns of measuring coding sequence expression, phenotypic patterns of gene expression, and measurements from representative multi-gene signatures. Our results showed that the principle determinant of variance from these protocols was use of exon capture probes, followed by the conditions of preservation (FF versus FFPE), then phenotypic differences between breast cancers. We also successfully identified one protocol, with RNase H-based ribosomal RNA (rRNA) depletion, exhibited least variability of gene expression

measurements, strongest correlation between FF and FFPE samples, and was generally representative of the transcriptome.

In the second topic, we focused on *TP53* penetrance estimation for multiple primary cancers (MPC). The study was motivated by the high proportion of MPC patients observed in Li-Fraumeni syndrome (LFS) families, but no MPC risk estimates so far have been provided for a better clinical management of LFS. To this end, we proposed a Bayesian recurrent event model based on a non-homogeneous Poisson process in order to estimate a set of penetrance for MPC related to LFS. Toward the associated inference, we employed the familywise likelihood that allows for utilizing genetic information inherited through the family. The ascertainment bias, which is inevitable in rare disease studies, was also properly adjusted by inverse probability weighting scheme. We applied the proposed method to the LFS data, a family cohort collected through pediatric sarcoma patients at MD Anderson Cancer Center from 1944 to 1982. Both internal and external validation studies show that the proposed model provides reliable penetrance estimates for MPC in LFS, which, to the best of our knowledge, have never been reported in the LFS literatures yet.

The research I conducted during my PhD study will be useful to translational scientists who want to obtain accurate gene expression by applying RNA-seq technology to FFPE tumor tissue samples. This research will also be helpful to genetic counselors or genetic epidemiologists who need high-resolution penetrance estimates for primary cancer risk assessment.

## Table of Contents

APPROVAL SHEET.....	i
TITLE PAGE.....	ii
ACKNOWLEDGEMENTS.....	iii
ABSTRACT.....	iv
LIST OF FIGURES.....	vi
LIST OF TABLES.....	viii
ABBREVIATIONS.....	ix
1. Introduction.....	1
1.1 Quality evaluation of formalin-fixed and paraffin-embedding tumor biopsies	
RNA sequencing data .....	1
1.1.1 Using RNA sequencing for expression profiling .....	1
1.1.2 RNA-seq data analysis.....	1
1.1.3 Effects of tumor sample preservation in the clinic on RNA.....	2
1.1.4 Limitations of previous studies on FFPE samples RNA-seq expression	
profiling quality evaluation .....	4
1.2 Characterization of the onset of multiple primary cancers .....	5
1.2.1 The Li-Fraumeni syndrome data .....	5
1.2.2 Multiple primary cancers and the penetrance .....	6
1.2.3 Challenges of estimating MPC penetrance using LFS data .....	7
1.3 Dissertation organization .....	9
2. Protocols for transcriptome sequencing of formalin-fixed tumor biopsies that best	
represent high quality frozen tissue .....	11

2.1 Introduction .....	11
2.2 Methods .....	13
2.2.1 Tumor tissue samples .....	13
2.2.2 RNA-seq library construction and sequencing .....	15
2.2.3 Sequence alignment, post-alignment statistics and expression quantification .....	17
2.2.4 Data analysis .....	20
2.2.5 Quantification of CDS-expression pattern dissimilarity .....	22
2.3 Results .....	23
2.3.1 Post-alignment statistics .....	23
2.3.2 Uniformity and continuity of read coverage of transcripts .....	26
2.3.3 Pre-analytical sources of variance .....	29
2.3.4 Protocols that target mRNA or deplete rRNA .....	30
2.3.5 Protocol with subsequent exon capture .....	40
2.3.6 Pattern dissimilarity in measurement of coding sequence .....	46
2.3.7 Gene expression patterns associated with tumor phenotype .....	48
2.3.8 Representative gene signatures of prognosis .....	55
2.4 Discussion .....	56
3. A Bayesian estimation of semiparametric recurrent event model with applications to the penetrance estimation of multiple primary cancers in Li-Fraumeni Syndrome .....	60
3.1 Introduction .....	60
3.2 The motivating data .....	60

3.3 Preliminary Analysis of the LFS Data .....	62
3.4 The model .....	64
3.4.1 MPC-specific age-at-onset penetrance .....	64
3.4.2 Semiparametric Recurrent Event Model for MPC .....	65
3.5 Computing likelihood .....	68
3.5.1 Individual likelihood .....	68
3.5.2 Familywise likelihood .....	69
3.5.3 Ascertainment bias correction .....	70
3.6 Posterior Sampling through MCMC .....	72
3.7 Case study.....	72
3.7.1 Model fitting.....	72
3.7.2 Cancer risk prediction .....	73
3.7.3 The MPC penetrance estimates .....	77
3.7.4 Comparison with penetrance estimates from literature.....	81
3.8 Discussion.....	82
4. Conclusions and Future Research	
4.1 Conclusions .....	85
4.2 Future Research .....	87
Appendix .....	89
Bibliography .....	93
VITA .....	105

## LIST OF FIGURES

Figure 1.1: Overview of pre-analytical and analytical factors. ....	4
Figure 2.1: Workflows of RNA-seq library preparation. ....	13
Figure 2.2: Overview of main RNA-seq data analysis. ....	18
Figure 2.3: Summary of concordant pair alignment rate. ....	24
Figure 2.4: Summary of exonic, intronic and intergenic region alignment rate of all mapped reads. ....	25
Figure 2.5: Summary of rRNA alignment rate. ....	25
Figure 2.6: Summary of number of genes with TPM values greater than 0.1. ....	26
Figure 2.7: An illustration of mean read coverage.....	27
Figure 2.8: Summary of mean coefficient of variation. ....	28
Figure 2.9: Summary of the percentage of gaps. ....	29
Figure 2.10: Illustration of adjustment of mean-variance dependence. ....	30
Figure 2.11: Scatter plot of the first three principal components for CPM-normalized and variance stabilizing transformed counts. ....	30
Figure 2.12: Hierarchical clustering results.....	31
Figure 2.13: Illustration of technical reproducibility. ....	32
Figure 2.14-2.17: Illustration of MA plot for overall gene expression. ....	33
Figure 2.18-2.20: Summary of between-protocol correlation coefficients. ....	37
Figure 2.21-2.23: MA-plot for FF.CR protocol as compared to other FF references. ....	41
Figure 2.24-2.25: Number of false positives identified by FFPE RNA-seq data. ...	44
Figure 2.26-2.27: Pattern similarity of coding DNA sequencing. ....	46

Figure 2.28-2.32: Phenotypic differential expression analysis. ....	50
Figure 3.1: Kaplan-Meier estimates of survival distribution for the first or the second gap times after removing data from probands.....	64
Figure 3.2. ROC of 5-year risk of developing next primary cancer assessed by 10-fold cross-validation. ....	74
Figure 3.3: Comparison of validation performance between our multiple primary cancer-specific penetrance and those estimated from Kaplan-Meier (KM) method in predicting the first or the second primary cancer occurrence using the MD Anderson prospective data. ....	77
Figure 3.4: Age-at-onset penetrance for females or males without a history of cancer. The shaded area is the 95% credible bands. ....	79
Figure 3.5: Penetrance estimates of the second primary cancer since the first primary cancer diagnosis time, stratified by the first primary cancer diagnosis time, genotype and gender. ....	80
Figure 3.6: Age-at-onset penetrance when with or without a history of cancer. ....	82
Figure Appendix B: A hypothetical pedigree for illustrating likelihood calculation using the Elston- Stewart algorithm. ....	92

## LIST OF TABLES

Table 2.1: Summary of starting RNA materials and the related cost for the RNA-seq data generated in this study. ....	14
Table 2.2: Histopathology annotation, extracted RNA integrity and sample fixation and storage time for FFPE condition of the nine breast tumors. ....	15
Table 2.3: Summary of the median correlation coefficients. ....	40
Table 2.4: Summary of the median of mean dissimilarity scores across nine tumor samples. ....	48
Table 2.5: Summary of median AUC values of between tumor phenotype differential expression. ....	49
Table 2.6: Summary of the median spearman correlation coefficients across nine tumor samples for five signature gene sets. ....	56
Table 3.1: Summary of number of families of LFS data. ....	61
Table 3.2: Number of primary cancer patients in LFS data. ....	62
Table 3.3: Number of primary cancer patients by the <i>TP53</i> mutation status and gender in MD Anderson prospective data. ....	75
Table 3.4 Summary of BIC for model selection. ....	78
Table 3.5 Summary of posterior estimates. ....	78
Table 3.6: Median second primary cancer-free times since the first primary cancer diagnosis age and their 95% confidence intervals. ....	80



## ABBREVIATIONS

AUC	Area under the curve
BIC	Bayesian information criteria
CDS	Coding region sequences
CI	Confidence interval
CPM	Count per million
CR	Coding region targeted protocol
CV	Coefficient of variation
deM	De-methylation/ de-modification
DE	Differential expression
DV200	Percentage of RNA fragments longer than 200bp
ER	Estrogen receptors
FF	fresh frozen
FFPE	Formalin-fixation and paraffin-embedding
FP	False positive
FPR	False positive rate
FPKM	Fragments per kilobase of exon per million reads mapped
GLM	Generalized linear model
HR	Hormone receptors
I.TotalRNA	Total RNA library protocol with bead-based ribosomal
RNA depletion method	
IPCW	Inverse probability of censoring weight
KM	Kaplan-Meier method

K.TotalRNA	Total RNA library protocol with enzyme-based ribosomal
RNA depletion method	
LFS	Li-Fraumeni syndrome
lowess	Locally weighted scatterplot smoothing
MCMC	Markov chain Monte Carlo
mRNA	Messenger RNA
MPC	Multiple primary cancers
NHPP	Non-homogeneous Poisson process
PCA	Principal component analysis
PCR	Polymerase chain reaction
PR	Progesterone receptors
qRT-PCR	Quantitative reverse transcription PCR
RIN	RNA integrity number
RLE	relative log expression
RNA-seq	RNA sequencing
rRNA	ribosomal RNA
ROC	Receiver operating characteristics
SPC	Single primary cancer
sRNA	Sense RNA
TP	True positive
TPR	True positive rate
TPM	Transcript per million
TN	Triple receptor-negative

TMM	Trimmed mean of M values
UQ	Upper-quartile
vst	Variance stabilizing transformation

## **1. Introduction**

### **1.1 Quality evaluation of formalin-fixed and paraffin-embedding tumor biopsies**

#### **RNA sequencing data**

##### **1.1.1 Using RNA sequencing for expression profiling**

The development of gene expression biomarkers for cancer tissues typically relies on high-dimensional technologies to discover transcripts from fresh frozen (FF) samples with high quality nucleic acids. Biomarkers that measure strong signals from a few transcripts generally translate with customized PCR or hybridization assays, but other biomarker indications may require more complicated algorithms involving many transcripts from a large set of samples with mature clinical annotations(1). RNA sequencing (RNA-seq) is a powerful tool that has been successfully implemented for the quantification of whole transcriptome abundance using FF samples(2-4). Compared to traditional RNA measurement methods, such as quantitative reverse transcription PCR (qRT-PCR) and microarray, RNA-seq can interrogate both pre-defined and novel RNA species at a greater dynamic range, allowing a more comprehensive exploration for non-coding RNA biomarkers. Many previous studies have shown that RNA-seq can generate accurate expression profiling comparable to that of microarray, preserve biological variability(5), and have performance that is reproducible across laboratories and robust to the variation of pre-analytical factors(2, 6-9). These also make RNA-seq a promising platform for multigene mRNA signature-based assays with clinical validity(10).

##### **1.1.2 RNA-seq data analysis**

There are many variations of pipelines for RNA-seq data analysis for expression profiling, but no optimal pipeline exists for all RNA experiments(11). The best practices of RNA-seq data analysis depend on the scientific questions of interest, as well as pre-analytical and analytical factors involved in the study. For the comparative analysis of feature expression between formalin-fixation and paraffin embedding (FFPE) and FF samples, the factors that could influence the choice of RNA-seq data analysis are summarized in Figure 1.1. In general, the major steps of RNA-seq data analysis involve read alignment, quality control, quantification of feature counts and expression data normalization(11-13). For read alignment, different algorithms have been developed so that raw reads can be either mapped to the genome reference(14) or transcriptome reference(15). Multiple quality checks have been proposed to evaluate the quality of raw reads or of the after-alignment read coverage(16, 17). The feature counts quantified from read alignment file are a biased measurement of the true abundance because of differences in library size and feature length. Hence, proper count data normalization is required before performing expression-based statistical analysis. Common within-samples normalization methods for RNA-seq data include count per million (CPM), fragments per kilobase of exon per million reads mapped (FPKM) and transcript per million (TPM) (2, 18, 19). TPM is reported as a preferred method as it can adjust for both library size and feature length effects, and it is more invariant to the change of mean expressed transcript length(15, 19).

### **1.1.3 Effects of tumor sample preservation in the clinic on RNA**

In diagnostic pathology, FFPE is the standard method for preserving and storing tissue samples. FFPE samples are commonly used for analyzing protein, cell

morphology(20), and even DNA(21), but are incompatible for the analysis of RNA, as RNA is fragmented and chemically modified caused by FFPE. Multiple factors in FFPE procedure can influence the RNA integrity(22). For example, the formalin fixation process cross-links nucleic acids and proteins, and modifies the RNA by adding the mono-methylol to all four bases(23). The high temperature required for paraffin to penetrate the tissue during embedding step, as well as the storage at room temperature, facilitate this chemical modification, which leads to RNA degradation over time(22). As a result, the RNA extracted from FFPE samples has much lower yield and shorter fragment length compared to the high quality RNA extracted from FF samples.

Increasing number of studies support that RNA-seq can be used to reliably profile FFPE specimens, despite that the RNA derived from FFPE samples is fragmented and variably chemically modified. For example, Sinicropi *et.al* used 5-12 year old FFPE tumor sample RNA-seq data to successfully re-discover breast cancer recurrence risk RNA biomarkers that were developed based on RT-PCR(24). Adiconis *et.al*, Li *et.al*, Liu *et.al* and Graw *et.al* showed that overall gene expression is highly correlated ( $r>0.8$ ) between FFPE and matched FF RNA-seq data using different types of tumor samples(25-27). Lin *et.al* applied RNA-seq to FFPE bladder tumor samples to identify a gene signature that can predict the risk of developing non-muscle invasive versus muscle invasive tumors for patients with high-grade T1 bladder cancer(28).

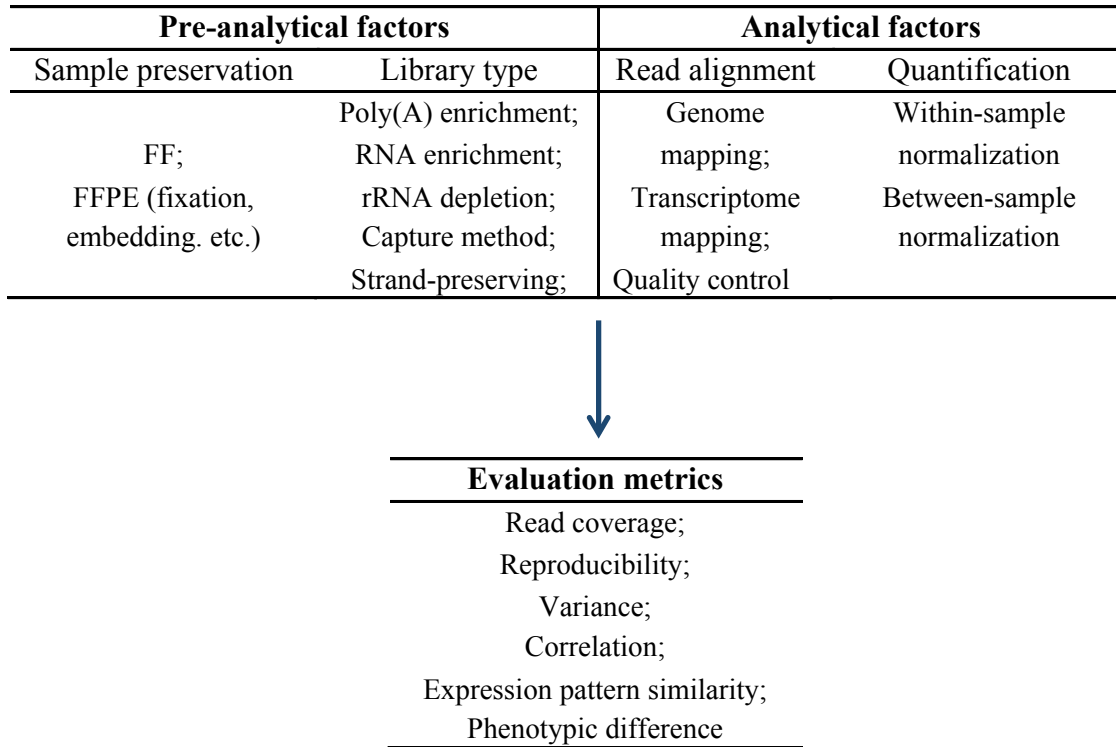


Figure 1.1: Overview of pre-analytical and analytical factors relevant to the evaluation of FFPE sample RNA-seq data quality for expression profiling.

#### 1.1.4 Limitations of previous studies on FFPE samples RNA-seq expression profiling quality evaluation

One major limitation of previous FFPE RNA-seq expression data quality evaluation studies is on the experimental design. For example, the reference standards or the “gold-standards” used for the comparative analysis are usually those generated by mRNA sequencing or total RNA sequencing from matched FF samples(24-27). The variation and concordance within those FF RNA-seq data generated by different library preparation protocols have not been thoroughly evaluated. On the other hand, several library preparation protocols designed for FFPE samples, such as rRNA depletion, template

RNA random amplification, capture sequencing and chemical de-modification, have already been tested(10). But none of previous studies has evaluated all of these protocols together in one study.

Another limitation of previous studies is the scope of evaluation metrics used for comparative analysis. To be analytically valid, FFPE RNA-seq data must demonstrate both reproducibility and accuracy in uncovering intelligent expression signals. For example, we expect valid FFPE RNA-seq data not only have high expression concordance with match FF references for both genes and coding sequences, but also allow us to re-discover true differential expressed genes that we have identified from the FF “gold-standards”. We’re also interested in whether we can use FFPE RNA-seq data to correctly cluster the tumor sample phenotypes or what are the dominant factors causing differences between FFPE and FF RNA-seq data. However, previous studies have no comprehensive evaluations based on these criteria.

## **1.2 Characterization of the onset of multiple primary cancers**

### **1.2.1 The Li-Fraumeni syndrome data**

Li-Fraumeni syndrome (LFS) is a hereditary cancer syndrome first recognized by two physicians, Frederick Pei Li and Joseph F. Fraumeni Jr., after evaluating the medical records and death certificates from pediatric sarcoma patients(29). Individuals with LFS are characterized with high risk of developing early-onset, multiple and multiple types of primary cancers throughout his/her lifetime(30). The syndrome is associated with germline mutation in *TP53* tumor suppressor gene, and follows an autosomal dominant inheritance rule(31, 32). Recent studies have shown that cancer risk in LFS patients is



also associated with gender(33) and the interaction effect between *TP53* genotype and gender(34).

The data that motivated our study is a family cohort of LFS collected through probands with pediatric sarcoma treated at MD Anderson Cancer Center from January 1944 to December 1982 and their extended relatives. The data was collected based on probands with sarcoma diagnosed before age 16 and with at least 3 years after-diagnosis survival. The data collection was extended to the probands' blood relatives, which includes the probands' grandparents, parents, parental siblings, siblings and offspring. For each individual, the gender and the diagnoses of any malignant cancer except the non-melanoma skin cancer were recorded from the date of birth until the data of death or the study termination date, whichever came first. All cancer diagnoses were confirmed by medical records and death certificates. The primary cancer diagnoses were determined based on the histology and site information recorded for each cancer event. More details on inclusion criteria and cancer diagnosis confirmation can be found elsewhere (31, 35). We define mutation carrier in this study as someone with missense or truncation mutations in exons 2-11 of the *TP53* gene tested from peripheral-blood samples. All probands were tested for the *TP53* mutation status, and once positive, all of their first-degree relatives and any other family members with a high risk of being mutation carrier were also tested. More information about mutation testing can be found elsewhere(33).

### **1.2.2 Multiple primary cancers and the penetrance**

A primary cancer develops independently at different sites and histology from original cancer, which is not caused by extension, recurrence or metastasis(36). Multiple

primary cancers (MPC) refer to the case when primary cancers occur more than once per subject over the follow-up time. The MPC cases are getting increasingly common due to advances of the cancer treatment and related medical technologies. The National Cancer Institute estimates that there are around eleven million cancer survivors in the US as of 2005, which is more than three times than that in 1970(37). Furthermore, surviving a cancer does not necessarily suggest a decreasing risk of developing another cancer. For example, Eggermond *et.al* reported that the risk for a second primary among Hodgkin lymphoma survivors is 4.7-fold increased compared with that in the general population(38). The risk of developing MPC varies by genetic susceptibility factors as well. For example, LFS is associated with germline mutation in *TP53*(39, 40).

Penetrance is defined as the proportion of individuals with the genetic variants (genotype) that cause a particular trait and who have clinical symptoms of the trait (phenotype). It plays a crucial role in many genetic epidemiology studies to characterize the association of germline mutation with disease outcomes(41). For example, penetrance is an essential quantity for disease risk assessment, which is clinically important to identify at-risk individuals and to provide prompt disease prevention strategies. To be more specific, popular risk assessment models often require the penetrance estimates as inputs(42, 43).

### **1.2.3 Challenges of estimating MPC penetrance using LFS data**

The goal of this study is to estimate MPC-specific penetrance in LFS, which is defined as  $\Pr(\text{developing the next primary cancer by age } t \mid \text{Genotype} \ \& \ (\text{Cancer history}, \text{Gender}))$ . It shall therefore lead to more accurate cancer risk assessment in LFS for both

cancer survivors and no-cancer-history individuals by utilizing more detailed individual cancer histories with MPC.

Few attempts have been made for taking into account MPC in the penetrance estimation. Wang *et.al* used Bayes' rule to calculate multiple primary Melanoma (MPM)-specific penetrance, based on penetrance estimates for carriers, the ratio of MPM patients for carriers and non-carriers, and the ratio of MPM and single primary Melanoma (SPM) patients for carriers(44). However, they do not account for age and other factors that may contribute to variations observed in SPM and MPM patients, and rely on previous population estimates of penetrance and relative risk. To the best of our knowledge, this is the only work that has tried to estimate MPC-specific penetrance.

We remark that MPC can naturally be regarded as recurrent events which have been extensively studied in statistics(45). However, the MPC-specific penetrance estimation from the LFS data is more challenging than the conventional recurrent event model due to the following reasons.

First, the majority of individuals (74%) has unknown *TP53* genotype in the LFS family data. Since the genotypes within a family are highly correlated through the rule of Mendelian inheritance, we cannot simply ignore the missing information. Instead, we need to consider all possible genotypes for untested individuals, with the probability of each inferred genotype calculated based on family structure.

Second, the rate of cancer occurrence is time-varying and we need to take into account for time-dependent covariates like cancer status. To tackle this issue, we exploit the non-homogeneous Poisson process (NHPP) with time-varying occurrence rate(45-47). One may suggest Andersen-Gill model that extends Cox regression to the recurrent event

data context(48, 49). However, it cannot be directly applied to our case since there is no clear way to extend the partial likelihood to the family data in which the complicated pedigree structure should be taken into account for the estimation.

Finally, the LFS data are collected through high-risk probands with pediatric sarcoma and hence are not random samples. This is often referred to as the ascertainment bias, which commonly occurs in rare diseases studies and should be properly corrected for an unbiased estimation.

### **1.3 Dissertation organization**

This dissertation focuses on addressing the two challenges described in above sections. In chapter 2, we developed multiple evaluation criteria, accounting for different read alignment algorithms and count data normalization methods, to assess the expression profiling quality using FFPE tissue samples RNA-seq data, as compared to high quality FF references. In this study, we applied RNA-seq, following 6 different RNA-seq library preparation protocols, to identical pairs of breast cancer tissue that were randomized to FF or FFPE conditions. The parameters we used for evaluation covers post-alignment quality checks, read coverage quality, overall data variation, correlation, differential test and expression pattern similarity in coding sequences. We identified one RNA-seq library preparation protocol with consistent good transcript coverage uniformity and continuity, most concordant expression for overall and specific signature genes, and least differential expression when compared to the different non-capture sequenced FF samples.

In chapter 3, we developed a novel statistical model that can estimate the MPC penetrance using genotype-incomplete LFS family data. In brief, we consider the MPC

occurrence in a randomly selected individual as a Poisson process and build the model with the following two major components: 1) Recurrent events modeling, which is devised to estimate the time varying hazard that fully characterizes the primary cancer occurrence process. We used a proportional hazard function where the baseline is a function of current age and the exponential component can incorporate covariates of interest. The model can thus consider effects from current age, cancer history or genetic factors when estimating the risk for next primary cancer development. 2) Missing genotype imputation via the Elston-Stewart algorithm(50), which significantly increases the statistical power for parameter estimation using incomplete real data. This approach improves computational efficiency by exploiting the Mendelian inheritance property when inferring the missing genotype and recursively partitioning the original family into nuclear ones. We also correct the ascertainment bias in the model and finally make inference on model parameters via the Markov Chain Monte Carlo method. Our method shows reasonable cancer risk prediction performance in both internal and external validations.

In chapter 4, we conclude the dissertation with discussion and future research.

## **2. Protocols for transcriptome sequencing of formalin-fixed tumor biopsies that best represent high quality frozen tissue**

### **2.1 Introduction**

While it is generally best to identify gene expression biomarkers from cancer tissues using the highest quality of ribonucleic acids (RNA) purified from fresh frozen (FF) samples, any subsequent development toward diagnostic testing will require its translation for use with formalin-fixed paraffin-embedded (FFPE) tissue samples. However, the variably fragmented and chemically modified RNA derived from FFPE samples presents a challenge for accurate measurement of gene-expression(23, 51).

In a different context, there is great interest to perform transcriptome sequencing (RNA-seq) for biomarker discovery research using large cohorts of precious archival FFPE samples from completed clinical trials. However, an unfavorable signal-to-noise ratio from FFPE samples could reduce the accuracy of biomarker discovery. Therefore, it is essential to select a protocol for FFPE RNA-seq libraries that yields data that is comparable with a “gold standard” result from FF samples. But there is more than one standard protocol for RNA-seq of high-quality RNA from FF tumor samples.

We have summarized different approaches for RNA-seq library preparation in Figure 2.1. Those include: 1) selection of messenger RNA by targeting the poly(A) 3' tail (mRNA protocol), 2) depletion of more abundant ribosomal RNA (rRNA depletion) using bead-based method (I.TotalRNA protocol) or enzymic method (K.TotalRNA protocol), and 3) exon capture probes for known coding region sequence (CDS) from an RNA-seq library prepared (CR protocol). Data generated from the popular mRNA protocol using FF tissue samples (FF.mRNA library) are highly concordant with

microarray data in tumor gene expression signature study (8). But this protocol is not appropriate for degraded mRNAs from FFPE samples (52). On the other hand, total RNA library protocols do not restrict enrichment to poly(A)<sup>+</sup> tailed mRNA, allowing less biased quantification of isoform abundance (52, 53).

Corresponding protocols for RNA-seq from FFPE tumor samples include an adaptation of the mRNA protocol that combines random and poly(A) primers (sRNA protocol) was optimized for gene expression microarrays (SensationPlus kit, Affymetrix, CA); or are unchanged for the I.TotalRNA, K.TotalRNA and CR protocols (Figure 2.1). Total RNA protocols have achieved Pearson correlations with FF counterparts of >0.9 (26, 52, 54). Exon capture using the CR protocol has potential for stronger correlation, but involves selected coverage(55). Finally, since pre-treatment heat and methyl saturation have been claimed to reduce methylol adducts on FFPE RNA, we evaluated pre-analytical demethylation (deM) of total RNA prior to library preparation using the CR protocol and the sRNA protocols (Figure 2.1).

Consequently, this study was designed to directly compare the results from RNA-seq library protocols between optimally matched sample pairs (FF and FFPE) from representative breast cancers, in order to address three scenarios in translational research: 1) biomarker discovery from FF samples phase with intention to translate for FFPE samples in future studies for validation and diagnostic development, 2) biomarker discovery from FFPE samples that is intended to be representative had high quality FF samples been available, and 3) translation of existing biomarkers, developed using a different method (such as microarrays or RNA-seq using mRNA protocol), for use with RNA-seq data from FFPE samples.

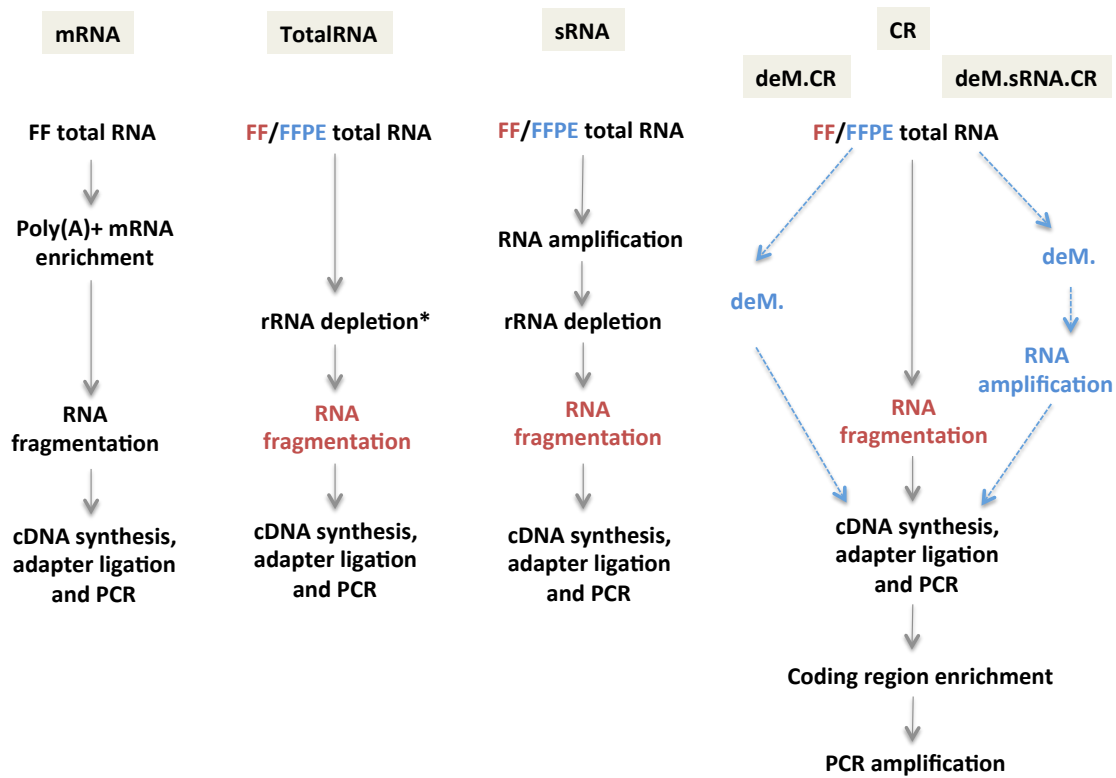


Figure 2.1: Workflows of RNA-seq library preparation. The red color indicates steps only applied to FF samples, while the blue indicates steps only applied to FFPE samples. The grey shaded boxes contain the names for each protocol. The \* indicates different rRNA depletion methods that result in two different TotalRNA protocols, that is, RiboZero for I.TotalRNA and Rnase H for K.TotalRNA protocol.

## 2.2 Methods

### 2.2.1 Tumor tissue samples

In order to minimize any impact from intratumoral heterogeneity, we collected fresh tissue, diced it into pieces of 1-2mm diameter, stirred, and then randomly assigned half to RNAlater solution or 10% neutral buffered formalin. The tissue in RNAlater was frozen



and stored at -80°C freezer (FF). The tissue in formalin was processed as a FFPE tissue block within the Histology and Tissue Processing Facility in MD Anderson Cancer Center. The phenotypes of the nine breast cancers, defined by pathologic status of hormone receptors (HR) and HER2 receptor were: HR+/HER2- in five, HR+/HER2+ in one, and triple receptor-negative (TN) in three (Table 2.1, 2.2).

Chemistry Procedure	Sample Size	Starting RNA( <i>ng</i> )	Cost (\$)	Time (days)
FF.mRNA	18	500	75	2
FF.CR	9	10	160	3
FF.I.TotalRNA	9	100	120	2
FF.K.TotalRNA	9	100	105	2
FFPE.K.TotalRNA	9	100	105	2
FFPE.I.TotalRNA	9	100	120	2
FFPE.sRNA	18	100	170	3
FFPE.CR	9	20	160	3
FFPE.deM.CR	9	20	160	3
FFPE.deM.sRNA.CR	9	100	170	3

Table 2.1: Summary of starting RNA materials and the related cost for the RNA-seq data generated in this study.

Tumor ID	Histopathology annotation			RIN		DV200	Duration of fixation (days)	Cut slides Storage time (days)
	ER	PR	HER2	FF	FFPE	FFPE	FFPE	FFPE
C	+	+	-	6.7	1.7	65	2	8
E	+	+	-	6.7	1.7	79	1	264
F	+	+	-	6.9	1.6	80	3	242
N	+	+	-	9.3	1.3	77	3	239
T	+	+	-	8.2	1.2	80	1	197
R	+	+	+	6.8	1.9	78	1	155
G	-	-	-	9.3	2.2	85	1	123
L	-	-	-	6.9	2.2	76	1	81
S	-	-	-	7.6	2.0	79	1	66

Table 2.2: Histopathology annotation, extracted RNA integrity and sample fixation and storage time for FFPE condition of the nine breast tumors. All slides were cut right after the block was prepared. All the cut slides were stored at 4°C in sealed cases until they were used for RNA extraction.

### **2.2.2 RNA-seq library construction and sequencing**

The FF RNA was purified from the sample in RNAlater using the RNeasy Mini Kit (Qiagen, Valencia, CA), while the FFPE RNA was purified from 10 $\mu$ m freshly-cut sections using High Pure FFPE RNA Isolation Kit (Roche, Indianapolis, IN). A DNase-I treatment step was included in both the FF and FFPE RNA isolation protocols. RNA concentration was quantified using Nanodrop (Nanodrop Technologies, Wilmington, DE), and its integrity was assessed using a Bioanalyzer 2100 and an RNA Chip assay (Agilent Technologies, Wilmington, DE).

The mRNA protocol began with poly(A)<sup>+</sup> mRNA enrichment on 500ng of total RNA using oligo-dT beads followed by standard procedures of TruSeq RNA Sample Prep Kit v2 (Illumina, San Diego, CA). Briefly, the poly(A)<sup>+</sup> mRNA was fragmented, then double-stranded cDNA was synthesized using random primers. After end repair and ligation of dsDNA adapters, the library was amplified with 10 cycles of PCR.

The I.TotalRNA protocol used Ribo-Zero<sup>TM</sup> Magnetic Gold Kit to deplete ribosomal RNA (rRNA) from 100ng of total RNA, followed by library preparation using the Truseq Stranded Total RNA Sample Prep Kit (Illumina, San Diego, CA).

The K.TotalRNA protocol used an RNase H-based method to deplete rRNA from 100*ng* of total RNA, followed by library preparation using KAPA Stranded RNA-Seq Kit with RiboErase (Kapa Biosystems, Wilmington, MA)

The sRNA protocol began with whole-transcriptome amplification on 100*ng* of total RNA using SensationPlus™ Amplification Kit (Affymetrix, Santa Clara, CA). The protocol used the same methods to amplify the RNA template as for gene expression microarrays. In brief, the template RNA was reverse-transcribed into the first-strand cDNA using random and oligo-dT primers, then the sense RNA (sRNA) was synthesized by in vitro transcription. Next, 4.5*μg* of sRNA was subjected to rRNA depletion using the Ribo-Zero™ Magnetic Gold Kit and then 50*ng* of rRNA-depleted sRNA was used as input for library construction using Truseq RNA Sample Prep Kit v2 as described in mRNA protocol, bypassing the poly(A)+ mRNA purification step.

The Coding-Region (CR) protocol was performed using Truseq Access RNAseq kit (Illumina, San Diego, CA) following manufacturer's instruction. In brief, cDNA was generated using random primers from the 10*ng* of RNA from FF, or 20*ng* of RNA from FFPE samples. Next, sequencing adapters were ligated to the resulting cDNA followed by the 1st round PCR amplification (15 cycles). After validation, a 4-plex pool of libraries was made and the coding regions of the transcriptome were enriched by two cycles of hybridization and capture to ensure high specificity. Finally, the 2nd round of PCR (10cycles) was performed to further amplify the enriched library for sequencing.

We also developed a de-modification (deM) protocol to leach methyl adducts from FFPE-derived RNA by heating it at 70°C for 30 min in a de-modification solution (1x TE buffer containing 20*μM* NH<sub>4</sub>Cl, pH7.0) (56)(57). This deM proved effective in restoring

the template activity of RNA in RT-PCR (unpublished data). Starting with de-modified RNA, we tested three additional FFPE library preparation methods: FFPE.deM.CR, FFPE.deM.sRNA.CR, FFPE.deM.sRNA. These methods followed the same main protocols mentioned above, with same amount of de-modified FFPE RNA as input.

In each protocol, the FF RNA was subjected to fragmentation prior to reverse transcription and cDNA generation, but no fragmentation was performed on FFPE RNA, except in the K.TotalRNA protocol where the FFPE RNA was fragmented at 85°C for 3 min according to manufacturer's instructions. For the mRNA and sRNA protocols, the libraries were prepared with two technical replicates to test reproducibility.

The size distribution of RNA-seq libraries was measured to be in the range of 200–600 bp and peaked around 270 bp using Agilent High Sensitivity DNA kit on a Bioanalyzer. Libraries were quantified using KAPA Library Quantification Kits (Kapa Biosystems, Wilmington, MA) and then paired-end sequenced on Illumina Hi-Seq 2000 Sequencing System with two or four libraries pooled in one lane. All libraries were randomly assigned to a lane (4 per lane) of the Hi-Seq 2000 following a rule that no technical replicates could share the same lane. We generated 100 base-paired reads for sample C and 50 base-paired reads for the other eight samples for the FF.mRNA and FFPE.sRNA protocols. All remaining libraries had 75 base-paired reads.

### **2.2.3 Sequence alignment, post-alignment statistics and expression quantification**

The computational analysis of RNA-seq data performed in this study can be summarized in Figure 2.2. We mapped reads to the human reference genome hg19 using Tophat2(14) (v.2.0.4, default parameters and supplying the -G option with GTF

annotation file downloaded from UCSC genome browser). The concordant pair alignment rate was obtained from the Tophat2 output. For rRNA alignment, we mapped reads to manually merged human rRNA references using BWA(58) in paired-end mode as previously described(26). Gene-level expression was quantified by htseq-count(59) in the "union" mode and using same GTF annotation file for mapping. To quantitate CDS-level expression, we first modified the GTF annotation file by adding a new feature ID "exon\_id" into the attribute. The exon\_id concatenates the feature type, start and end position and gene id for each row. The number of reads mapped to coding sequence was counted by htseq-count ("intersection-nonempty" mode, supplying the -t option with "CDS" and the -i option with "exon\_id").

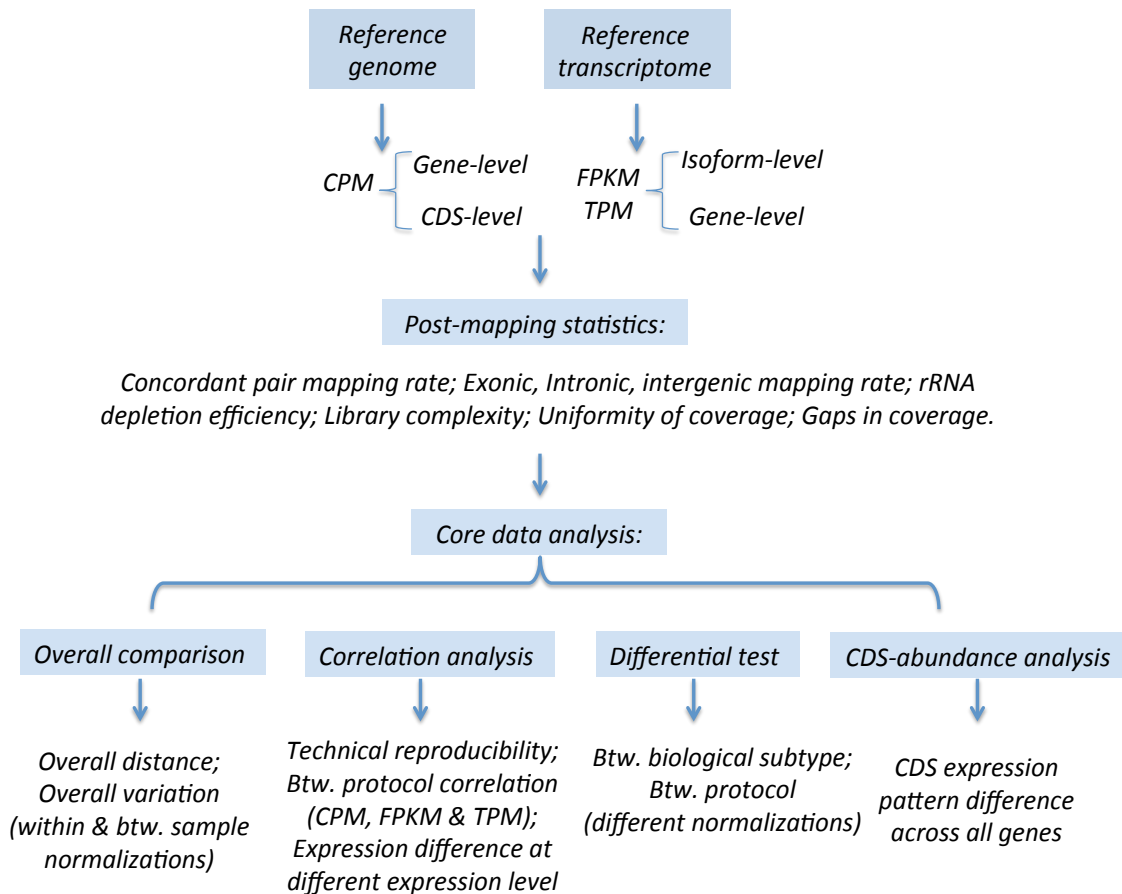


Figure 2.2: Overview of main RNA-seq data analysis performed in this study.

We used RNA-SeQC (16) (v.1.1.8 and same genome reference and GTF annotation file as that used for Tophat2 alignment), with genomic coordinate-sorted alignment file as the input, to calculate the mapping rate for exonic, intronic and intergenic regions, the coefficient of variation and the number of gaps in reads coverage. To calculate the coefficient of variation, the transcript length was normalized to 100 quantiles and the mean coverage signal for each quantile was calculated. The coefficient of variation of a transcript was calculated by dividing the standard deviation with the mean read coverage for that transcript. A smaller value of coefficient of variation indicates a greater uniformity of read coverage. The percentage of gaps (defined as >5 consecutive bases without coverage) was calculated by dividing the cumulative gap length by the cumulative transcript length. A smaller value of gap percentage indicates a greater continuity of read coverage.

We also mapped reads to the human reference transcriptome using RSEM(15) (v.1.2.11, Bowtie v.1.0.0 with default setting, and supplying the rsem-prepare-reference with UCSC knownGene transcriptome fasta file, and the rsem-calculate-expression with paired end mode). In contrast to Tophat2, RSEM avoids dealing directly with the splicing junction problem, by aligning the reads to the transcript reference and making inference on the relative abundance of each isoform from a mixture model built based on the RNA-seq data generative process(15, 18). The relative abundance was further adjusted by the effective length of isoform as an expression measure named fragments per kilobase of exon per million reads mapped (FPKM). An alternative expression measure, named

transcripts per million (TPM), was calculated by normalizing FPKM with the sum of per-nucleotide relative abundance over all isoforms. TPM is preferred to FPKM in some previous studies as it is more invariant to the change of mean expressed transcript length(18, 19).

#### **2.2.4 Data analysis**

Genes or CDS targeted by the CR protocol were identified using the manifest file for the Nextera Rapid Capture Exome preparation kit (Illumina, San Diego, CA), which cover same genes as the Truseq Access RNAseq kit. The poly(A)<sup>+</sup> genes were kept by filtering out poly(A)<sup>-</sup> genes as previously reported(60). The resultant 20,381 coding-region targeted and ploy(A)<sup>+</sup> genes were then included for further analysis whenever the mRNA and CR protocols are involved. For reproducibility and correlation analysis, the gene expression data were normalized to  $X$  by either CPM or FPKM or TPM and log transformed using the formula  $X' = \log_2(X + 1)$ .

A variance stabilizing transformation was applied to the CPM-normalized count data based on the empirically estimated functional relation between variance and mean as previously described(61). Principal component analysis (PCA) was performed on the transformed data using the 'prcomp' function in R after the gene variables were centered to zero and scaled to unit variance. A total of 17,395 Poly(A)<sup>+</sup> genes targeted by CR and with at least 1 normalized counts in five or more samples were included for the analysis. A total of 3543 genes with variance greater than 1 across all libraries were included for the hierarchical clustering analysis, where Euclidean distance and average linkage criteria were used. The 'pvclust' R package(62) was used to perform 1000 bootstrap resamples

on the clustering, and the bootstrap probability (bp) or the frequency that a cluster appears in bootstrap replicates was calculated as a measure of cluster uncertainty.

For differential gene expression analysis on distinct biological groups, the raw gene-expression data were normalized by two representative methods, (i) upper-quartile (UQ)(63): a global scaling method by the top quantile of the per-sample count distribution; and (ii) the trimmed mean of M values (TMM)(64): a global scaling method using an empirical estimate of relative RNA production of two samples. The TMM is based on the assumption that the majority of genes are not differentially expressed between groups. It doubly trims the noisy genes whose expression contributes to the bias of log-fold-changes (M values), and normalizes the raw gene count data with the weighted mean of adjusted M values, where the weight is the inverse of variance of the M values. Only genes with at least 5 reads in two or more samples prepared by one library construction method were retained for normalization. This resulted in an average of 16,810 (sd = 265) genes for further analysis. The relative log expression (RLE) is defined, for each gene, as the log ratio of read counts to the median count across all samples. The normalized counts were fit into negative binomial GLM for differential expression analysis using edgeR(65), with tag-wise dispersion. For receiver operating characteristic curves (ROC), either one of FF measures (FF.mRNA, FF.I.TotalRNA, FF.K.TotalRNA or FF.CR) was used as the gold standard to define truly differentially expressed genes. True positives are defined as genes measured as differentially expressed in both the gold standard and any one of other protocols, and the true positive rate (TPR) is defined as the number of true positives divided by the number of differentially expressed genes identified by the gold standard at a specific threshold. The false positive rate (FPR) is



analogously defined as the number of false positives divided by the number of non-differentially expressed genes according to the gold standard. The genes identifiable in every library preparation group were included for ROC. The most strongly differentially expressed genes were removed by filtering out genes with adjusted p-values smaller than 0.01. This resulted in a total ~15,000 genes for the ROC.

For library preparation method-based differential expression analysis, we used the paired design in edgeR to identify genes differentially expressed in response to library preparation method compared to the reference group for all nine tumors, adjusting for baseline difference between tumors. Only genes with at least 5 reads in five or more samples out of all 90 libraries were retained for normalization. This resulted in a total of 18,177 CR-targeted and poly(A)<sup>+</sup> genes for further analysis. All analysis and data visualization are performed using R (<http://www.r-project.org>).

### 2.2.5 Quantification of CDS-expression pattern dissimilarity

Let  $X_{ijk}$  be the CPM-normalized counts for  $i$ th CDS of  $j$ th gene in  $k$ th sample. We define the within-gene relative expression of  $i$ th CDS as

$$Y_{ijk} = \frac{X_{ijk}}{\sum_i X_{ijk}}$$

where  $i = 1, \dots, N_j$ ;  $j = 1, \dots, J$ ;  $k = 1, \dots, K$ . The pattern dissimilarity score used to measure pattern dissimilarity of  $j$ th gene between any two samples (i.e.  $k = 1$  or  $2$ ) is defined as

$$d_j = \sum_i \frac{|Y_{ij1} - Y_{ij2}|}{N_j}$$

The mean dissimilarity score between the two samples is then

$$\bar{d} = \frac{\sum_j \sum_i \frac{|Y_{ij1} - Y_{ij2}|}{N_j}}{J}$$

In our study, we only consider CR-targeted poly(A)<sup>+</sup> genes with two or more non-zero expressed CDS. This results in an average of 15,670 (*sd* = 134) genes.

## 2.3 Results

RNA extracted from FFPE samples was severely degraded, with RNA integrity number (RIN) of 1.2-2.2, versus 6.7-9.3 from FF samples (Table 2.2). All libraries generated >49 million raw reads (mean= 113 million, *sd*= 27 million).

### 2.3.1 Post-alignment statistics

We calculated the alignment rate for exonic, intronic, intergenic and all genomic regions for all libraries (Figure 2.3 and 2.4). Libraries from protocols that did not include exon capture probes (I.TotalRNA, K.TotalRNA, sRNA) had different mapping rates from FFPE samples than from FF samples, with the following mean differences: lower for exonic (overall mean difference= 0.335,  $P < 10^{-14}$ ), higher for intronic (overall mean difference= 0.309,  $P < 10^{-15}$ ), and comparable for intergenic sequence reads. The CR protocol (that used exon capture probes) had highly concordant mapping between FF and FFPE. Efficiency of rRNA depletion was highest for the CR protocol, followed by FFPE.K.TotalR (Figure 2.5). Additionally, the FFPE.sRNA protocol had lowest mapping rate for concordant pairs of reads, and least efficient rRNA depletion. Overall, The number of genes with read coverage (TPM > 0.1) was slightly higher in FFPE samples

than in FF samples for both non-CR and CR protocols (Figure 2.6), consistent with another report (27).

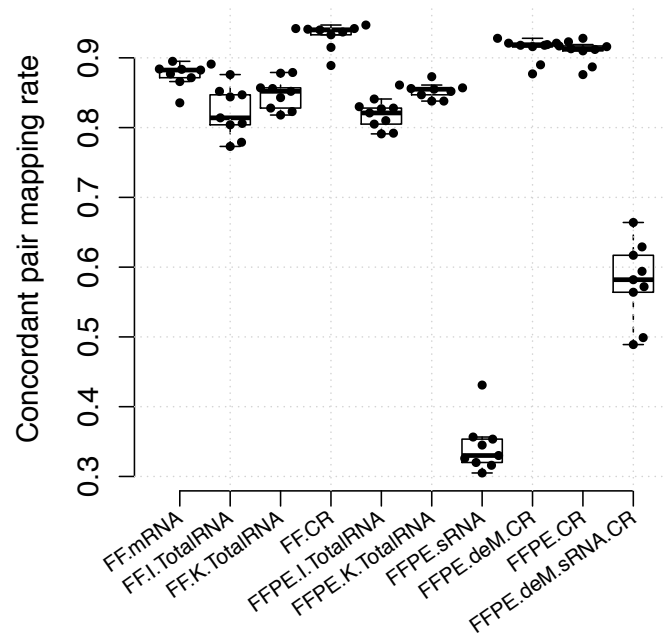


Figure 2.3: Summary of concordant pair alignment rate. Each box contains the mapping rate from nine tumor samples. The concordant pairs are those aligned with proper orientation and distance between the pair.

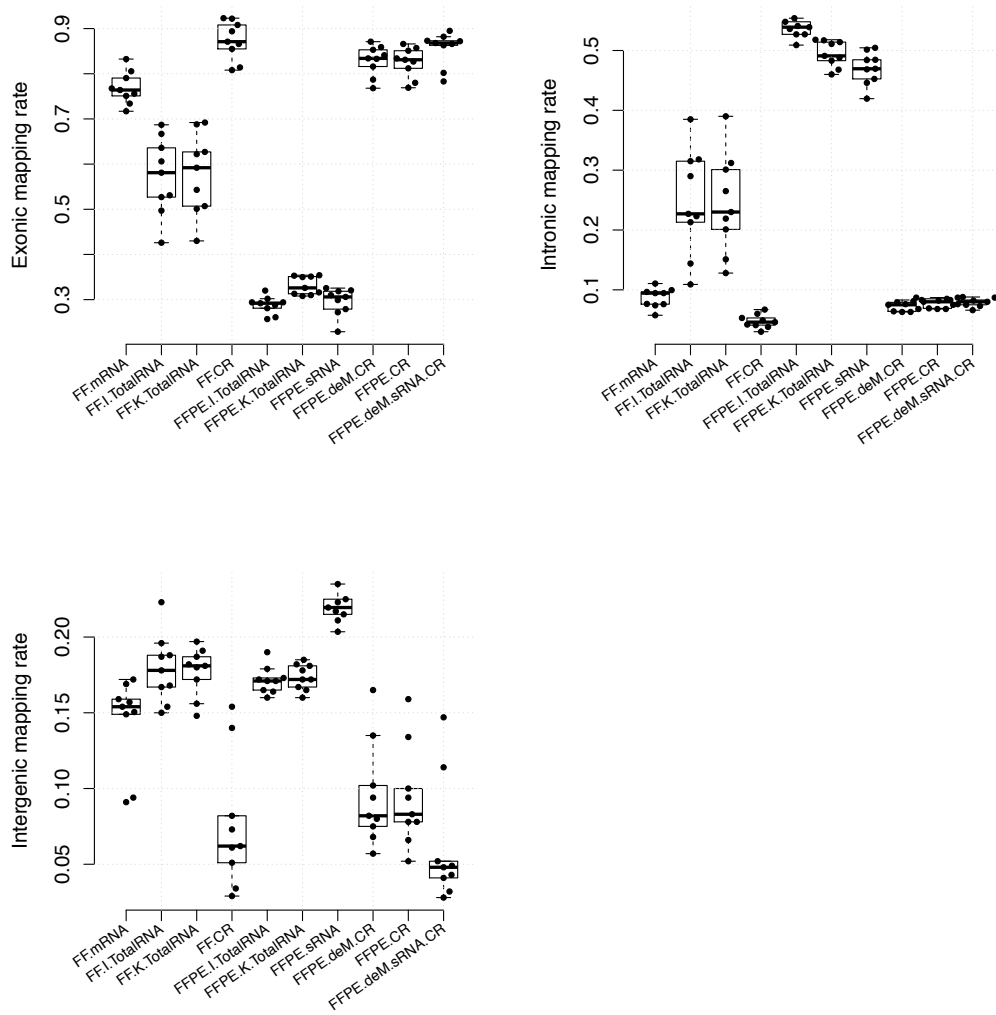


Figure 2.4: Summary of exonic, intronic and intergenic region alignment rate of all mapped reads. Each box contains the mapping rate from nine tumor samples.

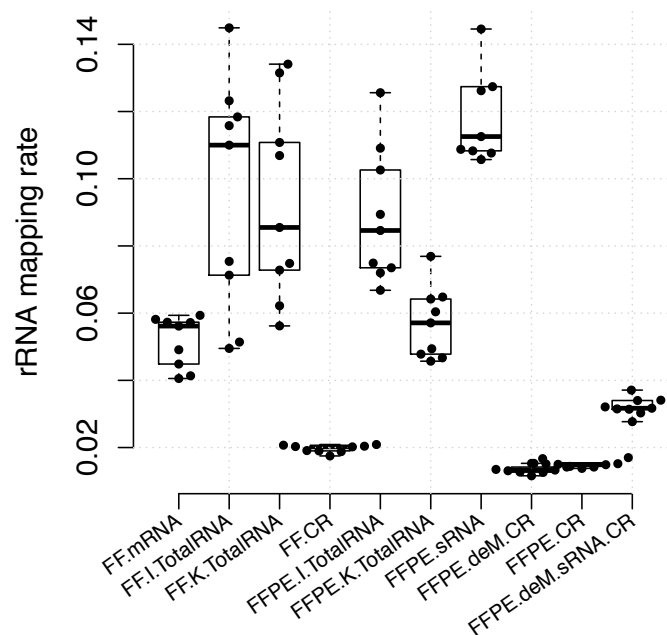


Figure 2.5: Summary of rRNA alignment rate. Each box contains the mapping rate from nine tumor samples.

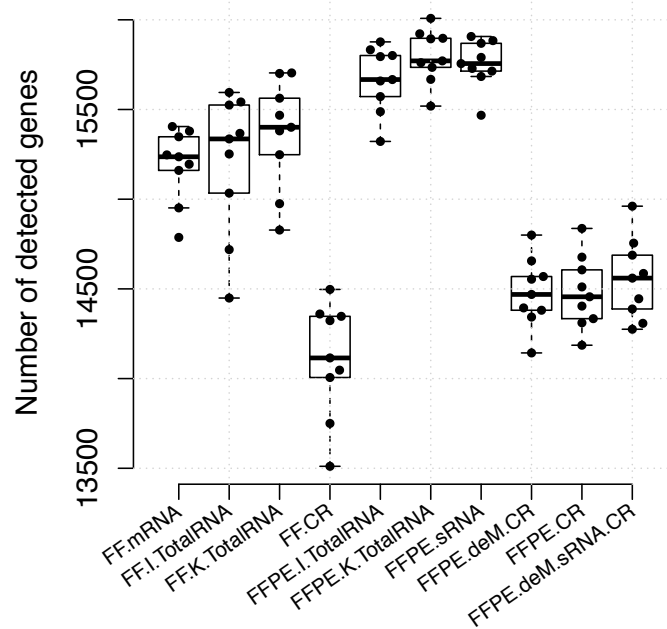


Figure 2.6: Summary of number of genes with TPM values greater than 0.1.

### 2.3.2 Uniformity and continuity of read coverage of transcripts

The uniformity of read coverage, as illustrated in Figure 2.7, was measured by the mean coefficient of variation (CV) across the top 1000 highly expressed transcripts, and coverage continuity was evaluated through the percentage of gaps without read coverage (Figure 2.8 and 2.9). FFPE.I.TotalRNA and FFPE.K.TotalRNA libraries demonstrated the most uniform and continuous coverage among protocols for FFPE samples, and were equivalent to protocols for FF samples. In contrast, the CR protocol produced non-uniform coverage, with high percentage of gaps, in both FF and FFPE libraries. The FFPE.sRNA protocol also introduced non-uniformity.

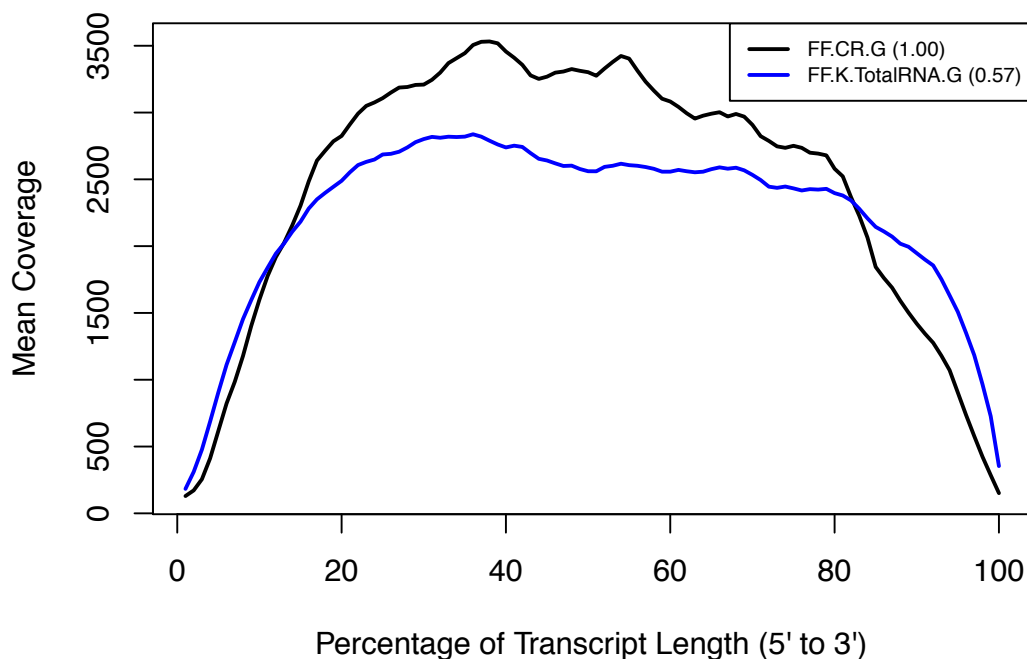


Figure 2.7: An illustration of mean read coverage for each normalized base position from top 1000 highly expressed transcripts for two libraries from FF sample G. The transcript

length is normalized to 100 quantiles and the mean coverage signal for each quantile is calculated. The coefficient of variation (in parenthesis) of a sample is the standard deviation divided by the mean of mean read coverage for that sample.

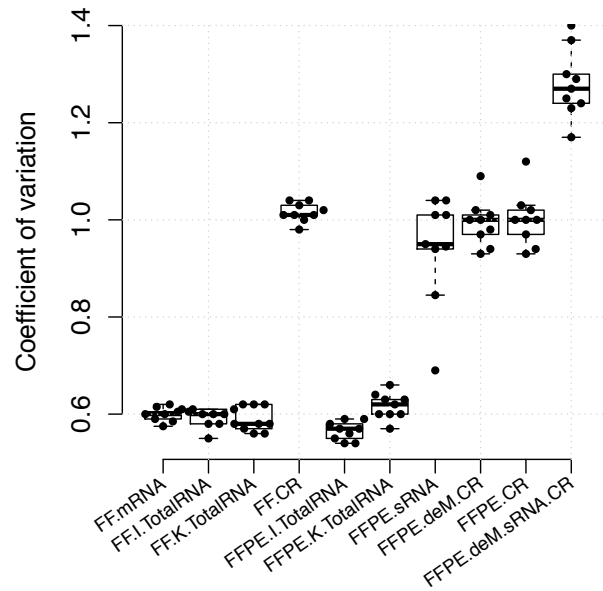


Figure 2.8: Summary of mean coefficient of variation (cv) of top 1000 highly expressed transcripts for all samples. Each box summarizes the mean cv from nine samples for one library preparation protocol. A lower cv value indicates better uniformity of read coverage.

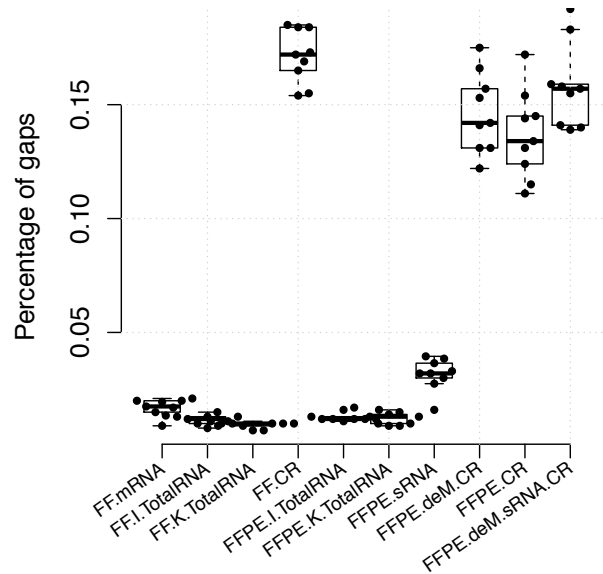


Figure 2.9: Summary of the percentage of gaps of top 1000 highly expressed transcripts for all samples. Each box summarizes the percentage from nine samples for one library preparation protocol.

### 2.3.3 Pre-analytical sources of variance

In RNA-seq studies, the variance across samples usually grows with the mean of gene expression (also known as heteroscedasticity), and this can be problematic for correctly uncovering the underlying pattern in data using techniques such as distance-based clustering (66). We therefore applied the variance-stabilizing transformation method to approximate the independence between variance and mean (Figure 2.10 and Methods). Principal component analysis (PCA) of expression of a total of 20,381 CR protocol targeted poly(A)<sup>+</sup> genes for all libraries showed that the 38.8% of total variation captured by the first two principal components was due to use of exon capture probes (CR protocol), 20.6% from the second and third components (effects of FFPE and biological differences) (Figure 2.11). Hierarchical clustering results, with high confidence



(average bootstrap probability= 0.93), showed that the major tumor phenotypes (HR+ vs. HR-) and the source tumor, clustered together with FFPE samples (Figure 2.12).

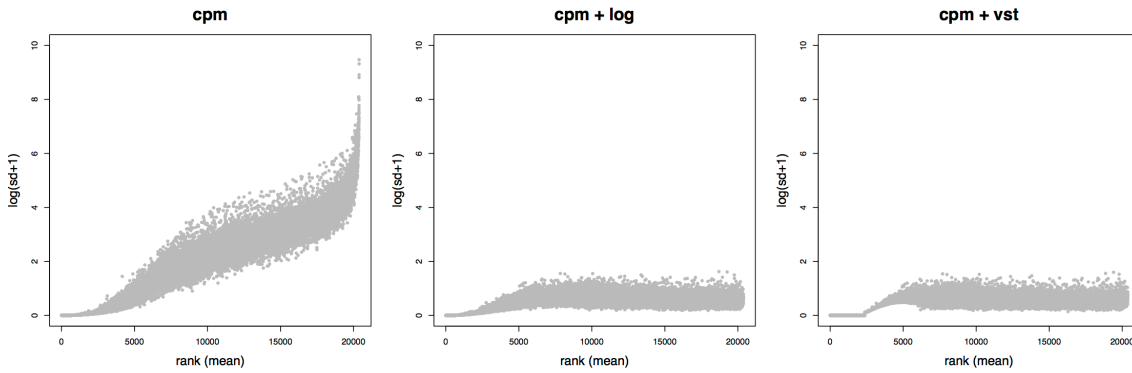
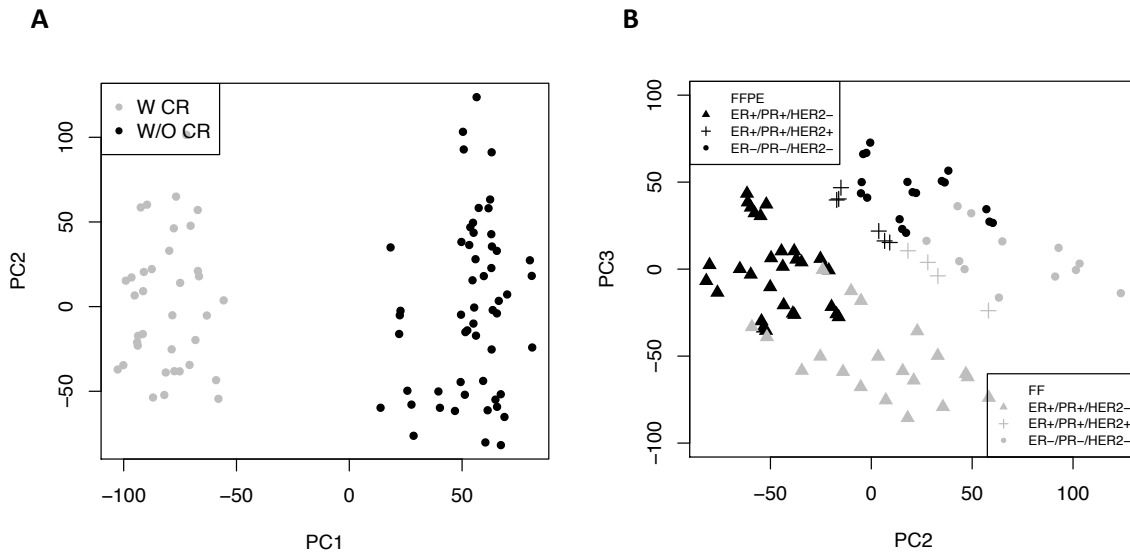


Figure 2.10: Scatter plot of per-gene standard deviation across all 90 libraries, against the rank of the mean expression level when with or without between sample normalization methods. Both log transformation and the variance stabilizing transformation (vst) can approximate variance-mean independence or homoscedasticity. Note that in the figure the standard deviation was added with one and then log transformed.



**Figure 2.11** Scatter plot of the first three principal components for CPM-normalized and variance stabilizing transformed counts of 20,381 CR-targeted poly(A)<sup>+</sup> genes. Each

point corresponds to one of 90 libraries. **A)** the gray color indicates samples prepared with CR and the black for those without CR treatment. A 38.8% of total variation comes from CR treatment. **B)** the gray color indicates FF samples and the black for FFPE samples. The symbol shape indicates the different biological group. The biological differences and FFPE effects are captured, which accounts for 20.6% of total variation.

#### **2.3.4 Protocols that target mRNA or deplete rRNA**

We performed technical replicates from source RNA for the FF.mRNA and FFPE.sRNA protocols in all 9 tumors, with replicate library preparation occurring on different days. The raw expression values were scale-normalized by total count and transformed to  $\log_2$  count per million (CPM). Technical replicates were highly correlated (Spearman  $\rho \geq 0.992$ ) for all samples (Figure 2.13).

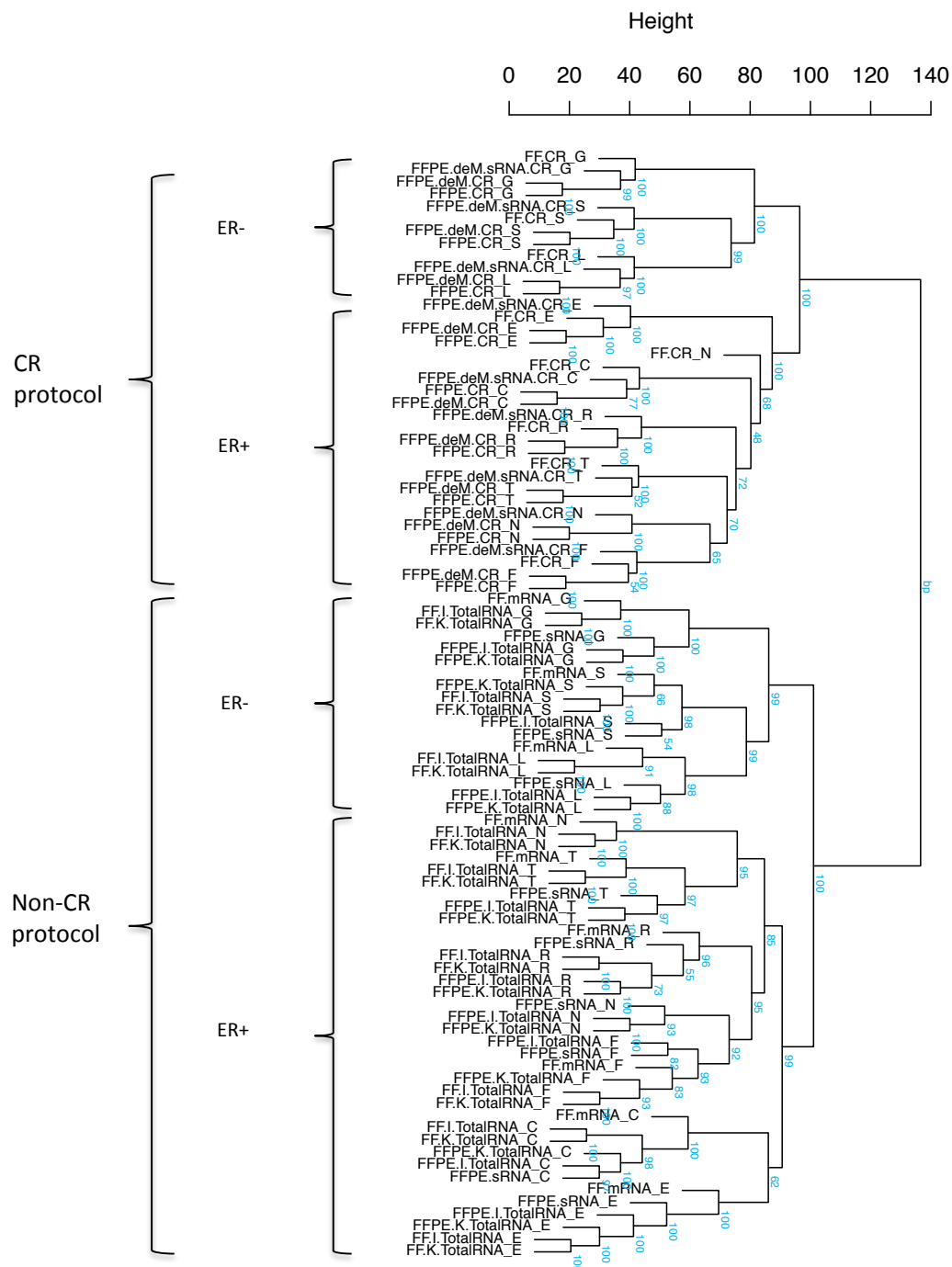


Figure 2.12: Hierarchical clustering of all 90 samples. The bootstrap probability (bp) or the frequency that a cluster appears in bootstrap replicates is annotated in blue.

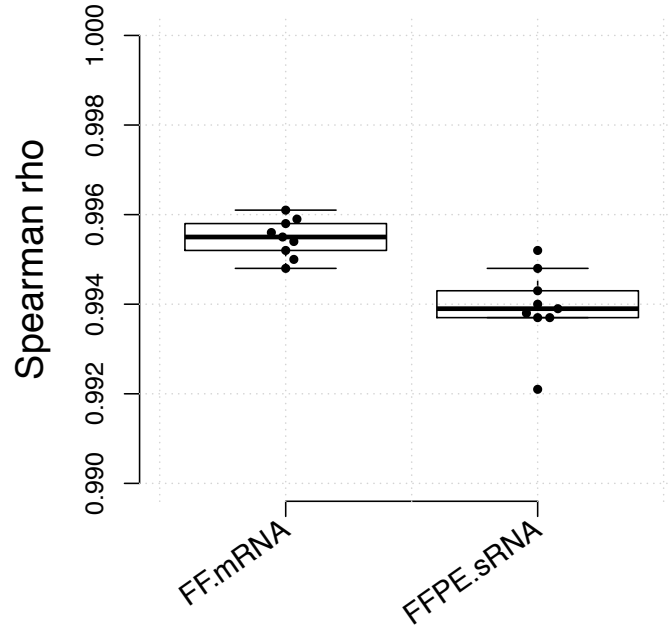


Figure 2.13: Summary of spearman's rho between two replicates using FF.mRNA or FFPE.sRNA protocols as a measure of technical reproducibility.

We next evaluated which library preparation protocol best represented the gene expression measurements that would be obtained from FF libraries. We used MA plots to compare, for each gene, the differences in expression between two different FF libraries against the average expression in the two libraries from tumor C. The log ratios between the two FF libraries for CR targeted genes were centered around zero, with small variation across different mean expression levels (Figure 2.14). We then compared two types of FFPE library protocols to FF.K.TotalRNA using libraries from tumor C (Figure 2.14). Although more variable than within-FF libraries, the log ratio values of

FFPE.K.TotalRNA were still centered around zero at different mean expression levels. However, the log ratio values of FFPE.CR to the FF reference deviated from zero at both low and high expression levels. The same patterns were observed for all other tumor samples (Figure 2.15-2.17). These observations suggest that the TotalRNA protocols produced high-quality FFPE RNA-seq data that was comparable to the FF RNA-seq data.

The FFPE.K.TotalRNA and FF.K.TotalRNA libraries were highly correlated (median rank correlation 0.973 using the TPM measure), significantly higher than FF.K.TotalRNA with FF.CR (mean difference = 0.066,  $P < 10^{-6}$ ), or any other FFPE protocol (lowest mean difference = 0.019,  $P = 0.031$ ) (Figure 2.18). Results were similar using CPM and FPKM measures (Figure 2.19-2.20). The FFPE.K.TotalRNA also had the highest median rank correlation with FF.mRNA and FF.I.TotalRNA, in spite of normalization methods used (Figure 2.18-2.20). We did note consistently low correlation between FF and FFPE for sample N across all the protocols.

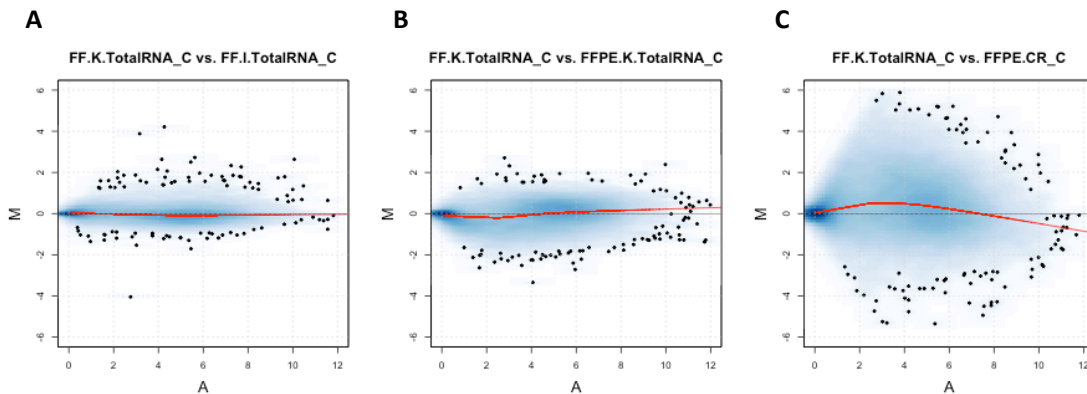


Figure 2.14: MA-plot of 20,381 CR targeted poly(A)<sup>+</sup> genes for tumor sample C when using FF.K.TotalRNA sample C library as the reference. A) MA plot for tumor C between FF.K.TotalRNA and FF.I.TotalRNA; B) MA plot for tumor C between FF.K.TotalRNA and FFPE.K.TotalRNA; C) MA plot for tumor C between

FF.K.TotalRNA and FFPE.CR. M is the  $\log_2$ -transformed expression of a gene from first library divided by that from the second library, while the A is the mean  $\log_2$ -transformed expression of the gene. The red curve indicates the locally weighted scatterplot smoother (lowess) fitted to the data.

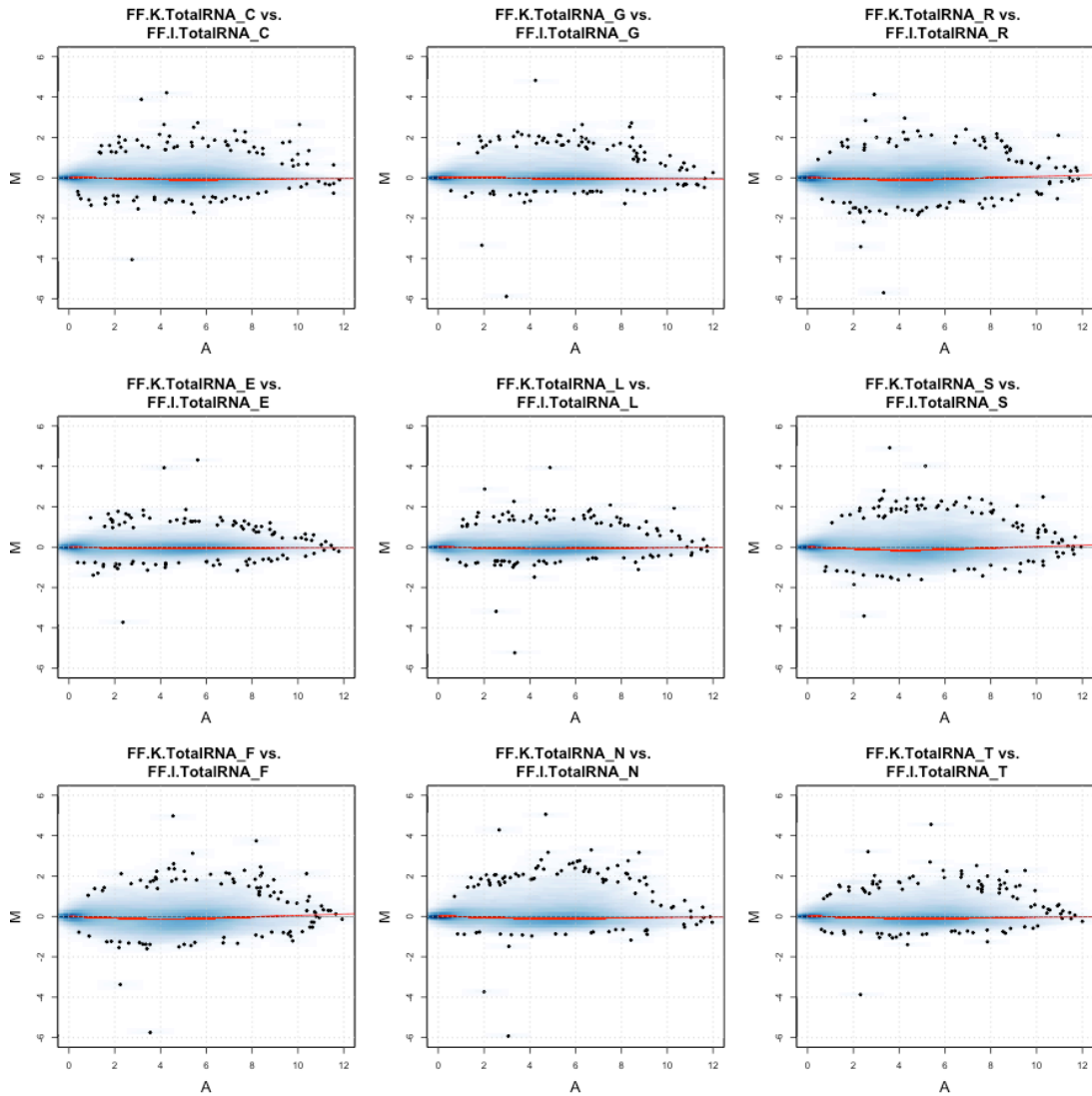


Figure 2.15: MA-plot for FF.I.TotalRNA protocol as compared to FF.K.TotalRNA for nine breast tumors. The red curve indicates the lowess smoother fitted to the data.

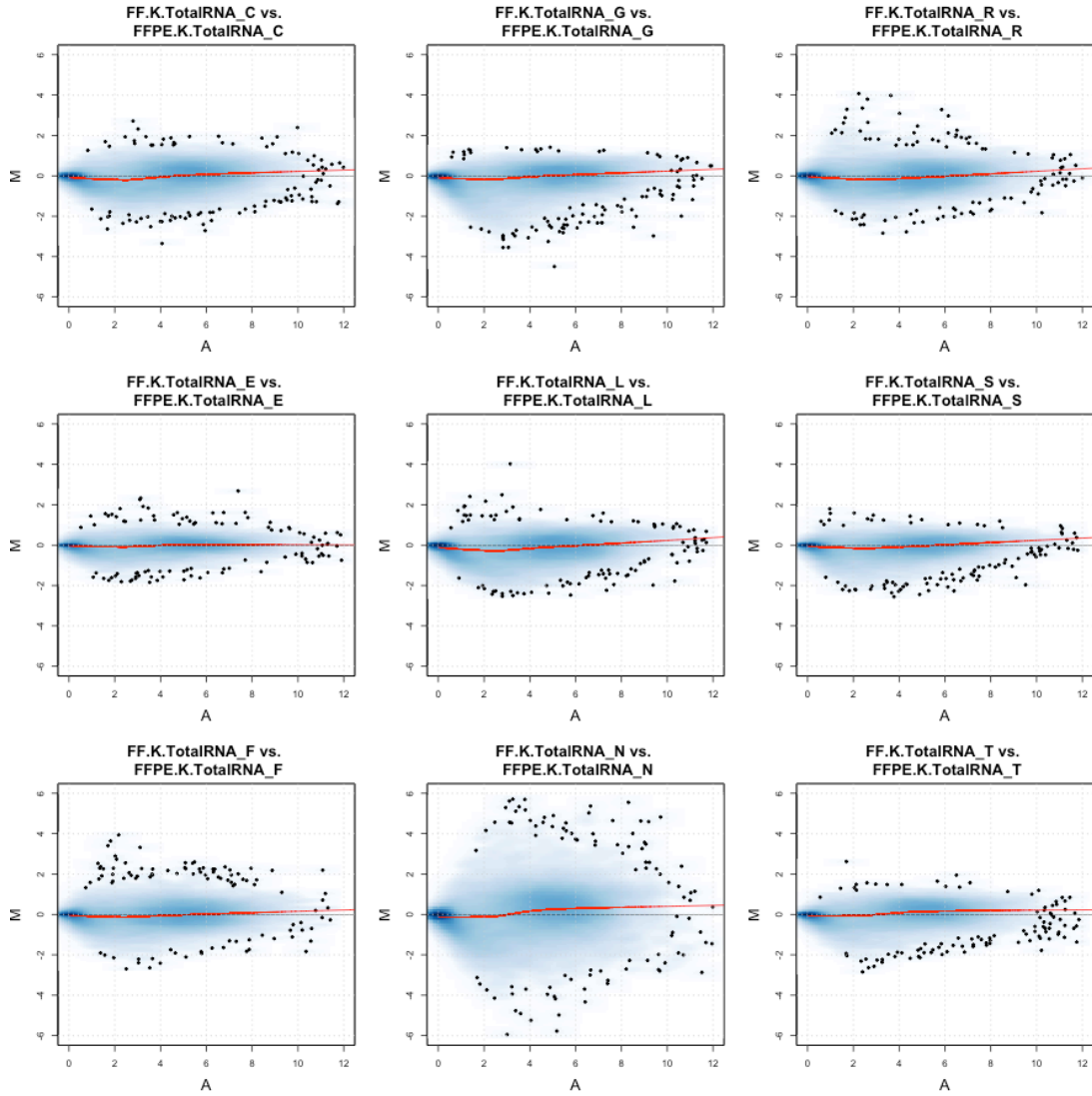


Figure 2.16: MA-plot for FFPE.K.TotalRNA protocol as compared to FF.K.TotalRNA for nine breast tumors. The red curve indicates the lowess smoother fitted to the data.

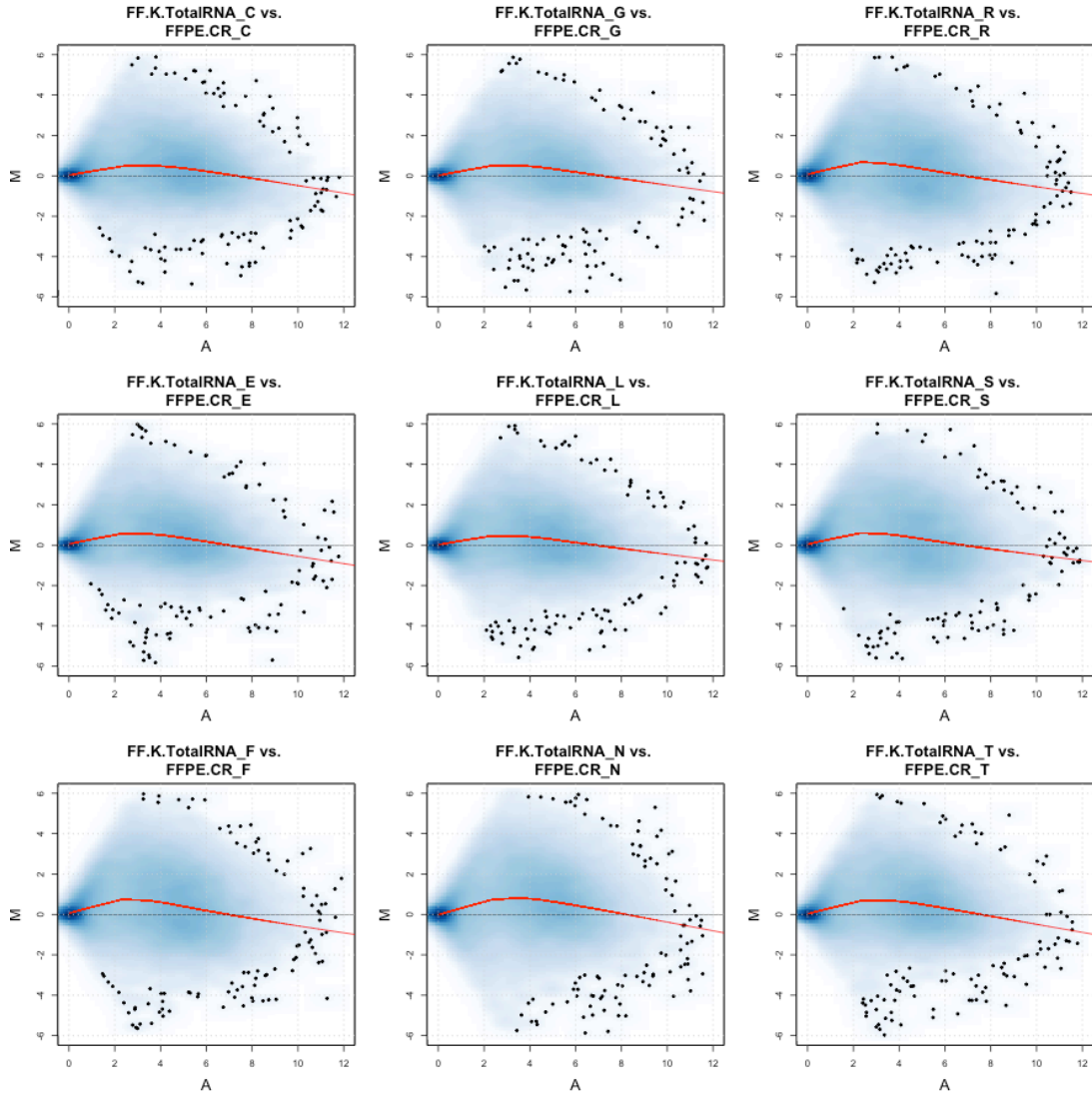


Figure 2.17: MA-plot for FFPE.CR protocol as compared to FF.K.TotalRNA for nine breast tumors. The red curve indicates the lowest smoother fitted to the data.



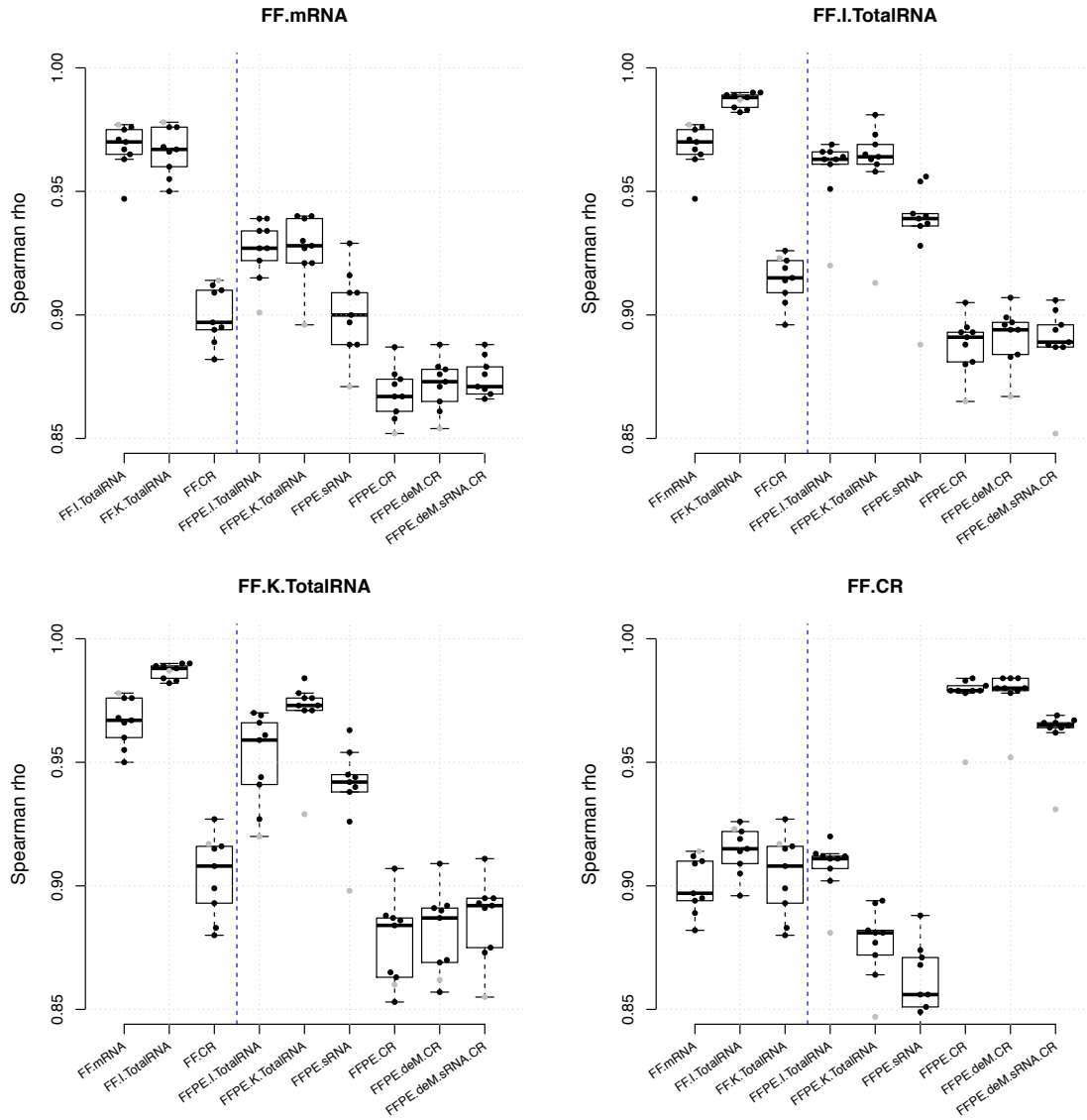


Figure 2.18. Summary of between-protocol correlation coefficients based on TPM. The main title of each figure is the reference protocol used for comparison. Each dot is the Spearman rho estimate calculated between the reference library and the library showing on the x axis. Each box summarizes the Spearman rho estimates from nine breast tumor samples. The gray dot indicates the tumor sample N.

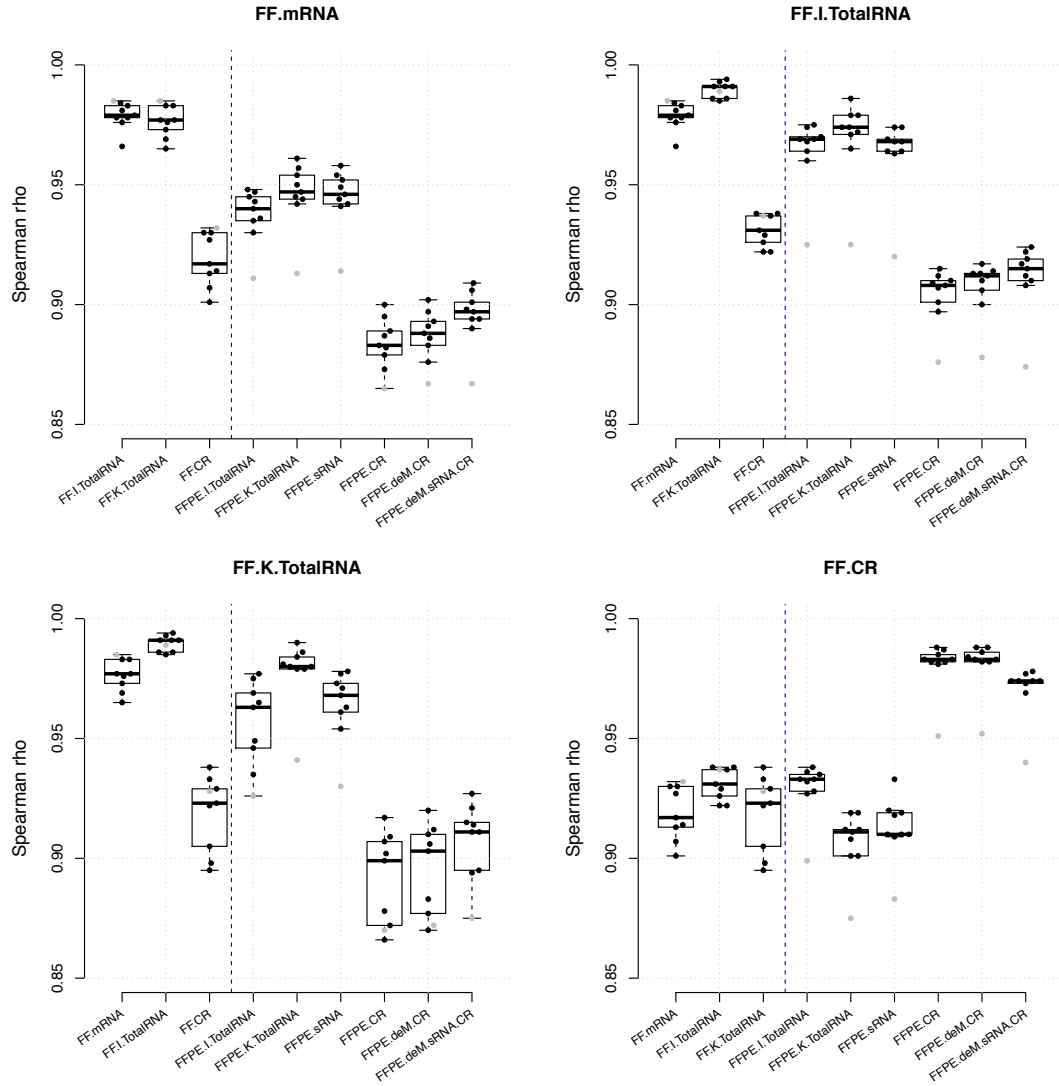


Figure 2.19: Summary of between-protocol correlation coefficients based on CPM. The main title of each figure is the reference protocol used for comparison. Each dot is the Spearman's rho estimate calculated between the reference library and the library showing on the x axis. Each box summarizes the Spearman's rho estimates from nine breast tumor samples. The gray dot indicates the tumor sample N.

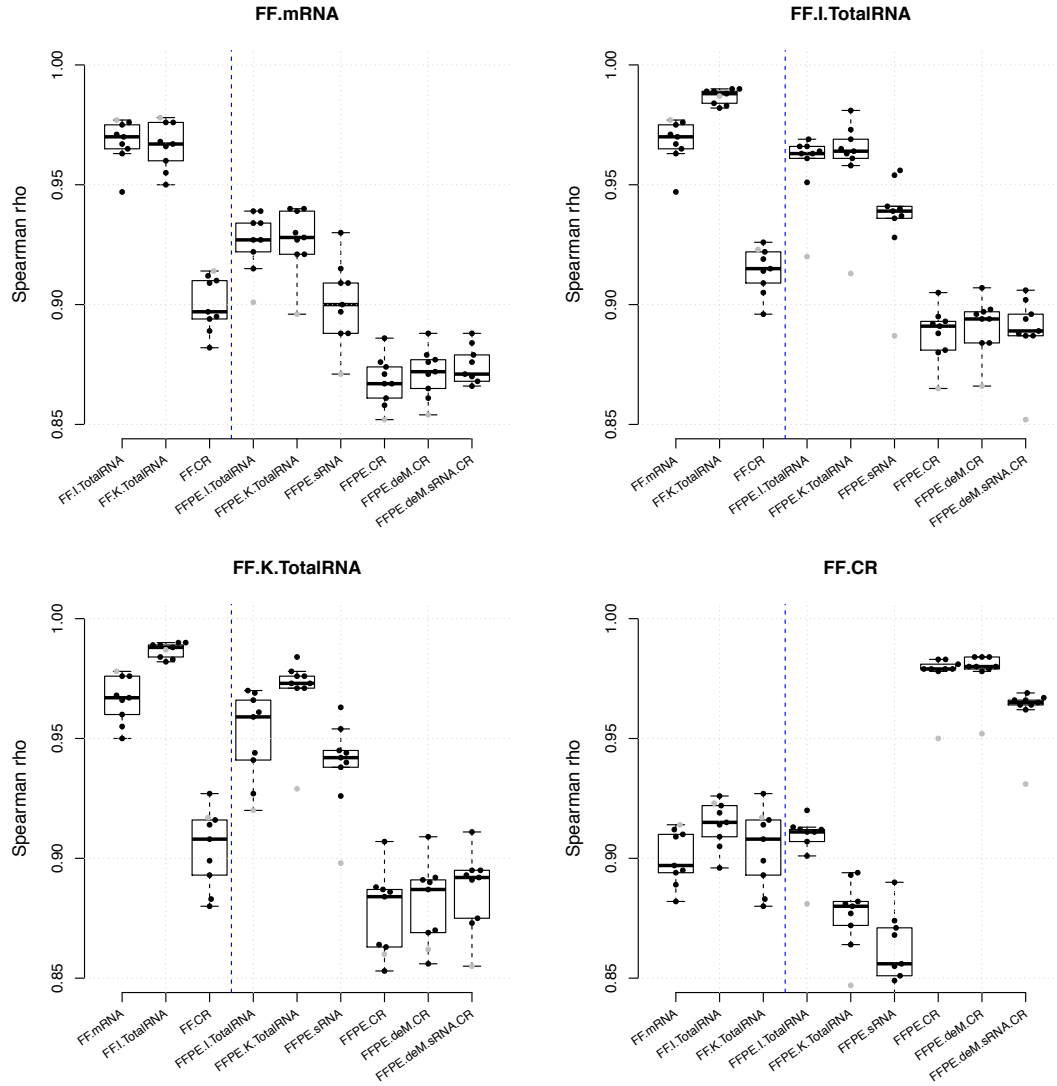


Figure 2.20: Summary of between-protocol correlation coefficients based on FPKM. The main title of each figure is the reference protocol used for comparison. Each dot is the Spearman's rho estimate calculated between the reference library and the library showing on the x axis. Each box summarizes the Spearman's rho estimates from nine breast tumor samples. The gray dot indicates the tumor sample N.

### 2.3.5 Protocol with subsequent exon capture

Subsequent use of exon capture probes after TotalRNA library preparation (CR protocol) resulted in a median rank correlation of 0.980 between FF and FFPE, but the FF.CR had much lower correlation with non-CR libraries (lowest mean difference = 0.063,  $P < 10^{-9}$  using TPM) (Figure 2.18 and Table 2.3). Comparing the log ratio values across all protocols for FF samples, the CR protocol tended to overly enrich the highly expressed genes, and was more likely to not capture low expressed genes (Figure 2.14 and Figure 2.21-2.23). Pre-analytical approaches to de-modification (deM) of methylol adducts from FFPE tissue-derived RNA using heat and amines, or random and dT primers for mRNA (sRNA protocol) had little effect on the FFPE.CR protocol (Figure 2.18 and Table 2.3). Addition of the deM method (FFPE.deM.CR) slightly increased concordance of expression but was not statistically significant. Similarly, the sRNA method (FFPE.deM.sRNA.CR) slightly increased the concordance of expression but was not statistically significant.

Case \ Reference	FF.mRNA	FF.I.TotalRNA	FF.K.TotalRNA	FF.CR
FF.mRNA	-	-	-	-
FF.I.TotalRNA	0.979 / 0.970	-	-	-
FF.K.TotalRNA	0.977 / 0.967	0.991 / 0.988	-	-
FF.CR	0.917 / 0.897	0.931 / 0.915	0.923 / 0.908	-
FFPE.I.TotalRNA	0.940 / 0.927	0.969 / 0.963	0.963 / 0.959	0.933 / 0.911
FFPE.K.TotalRNA	<b>0.947 / 0.928</b>	<b>0.974 / 0.964</b>	<b>0.980 / 0.973</b>	0.911 / 0.880
FFPE.sRNA	0.946 / 0.900	0.968 / 0.939	0.968 / 0.942	0.910 / 0.856
FFPE.deM.CR	0.883 / 0.867	0.908 / 0.891	0.899 / 0.884	<b>0.983 / 0.979</b>
FFPE.CR	0.888 / 0.872	0.912 / 0.894	0.903 / 0.887	<b>0.983 / 0.980</b>
FFPE.deM.sRNA.CR	0.897 / 0.871	0.915 / 0.889	0.911 / 0.892	0.974 / 0.965

Table 2.3: Summary of the median correlation coefficients using either CPM (left) or TPM (right). The highest median for FFPE protocols are highlighted in bold.

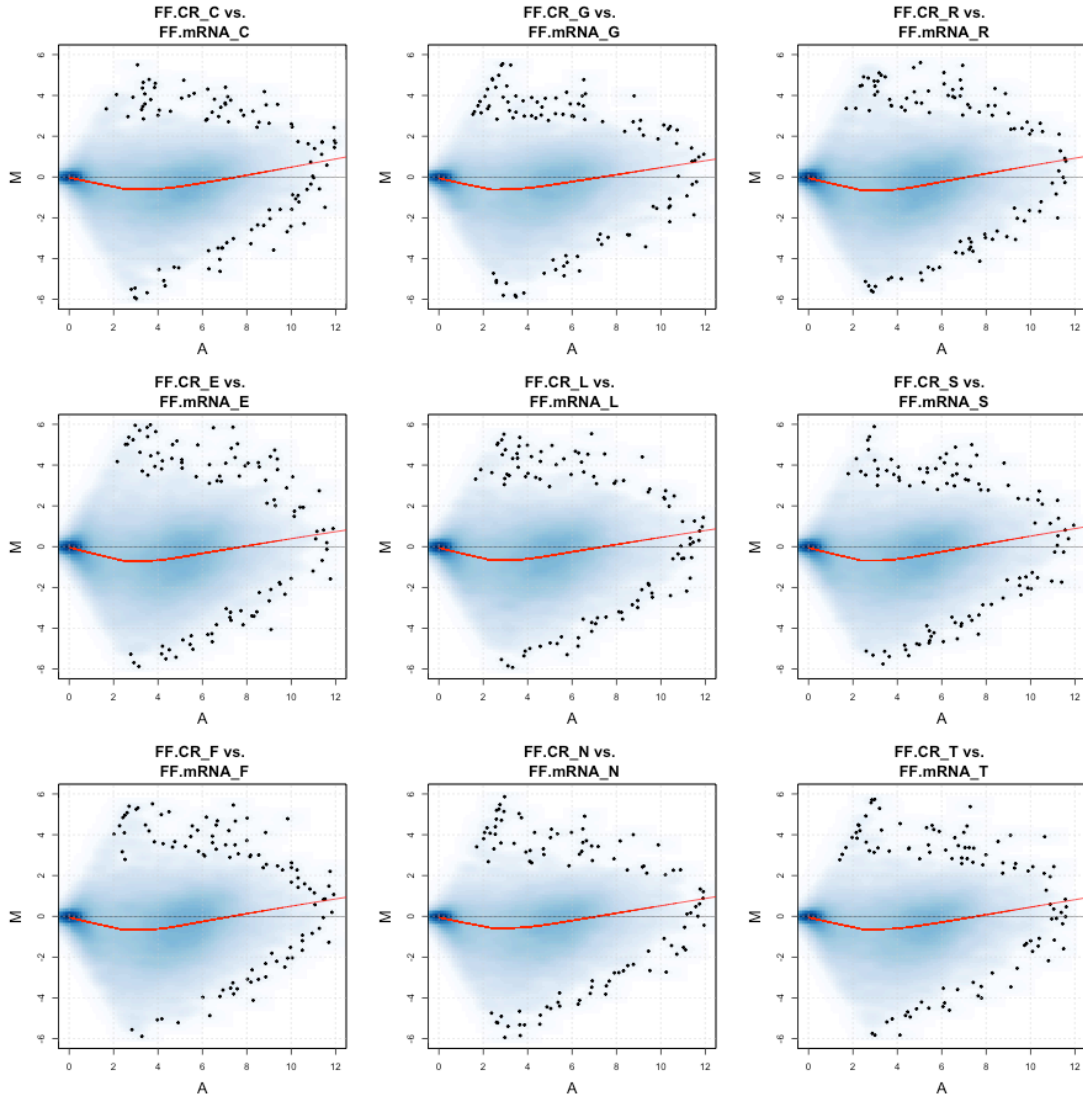


Figure 2.21: MA-plot for FF.CR protocol as compared to FF.mRNA for nine breast tumors. The red curve indicates the lowess smoother fitted to the data.

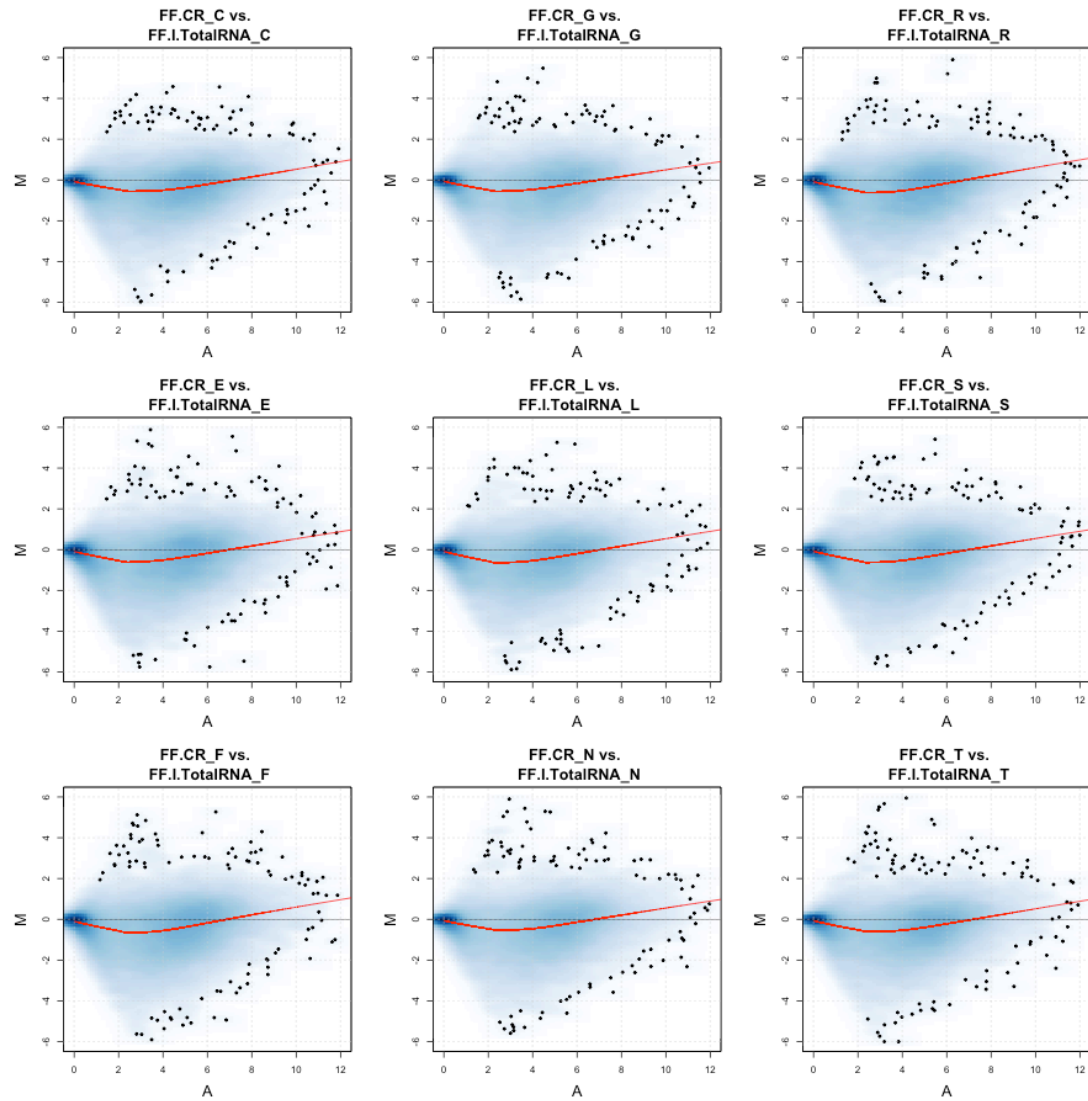


Figure 2.22: MA-plot for FF.CR protocol as compared to FF.I.TotalRNA for nine breast tumors. The red curve indicates the lowess smoother fitted to the data.

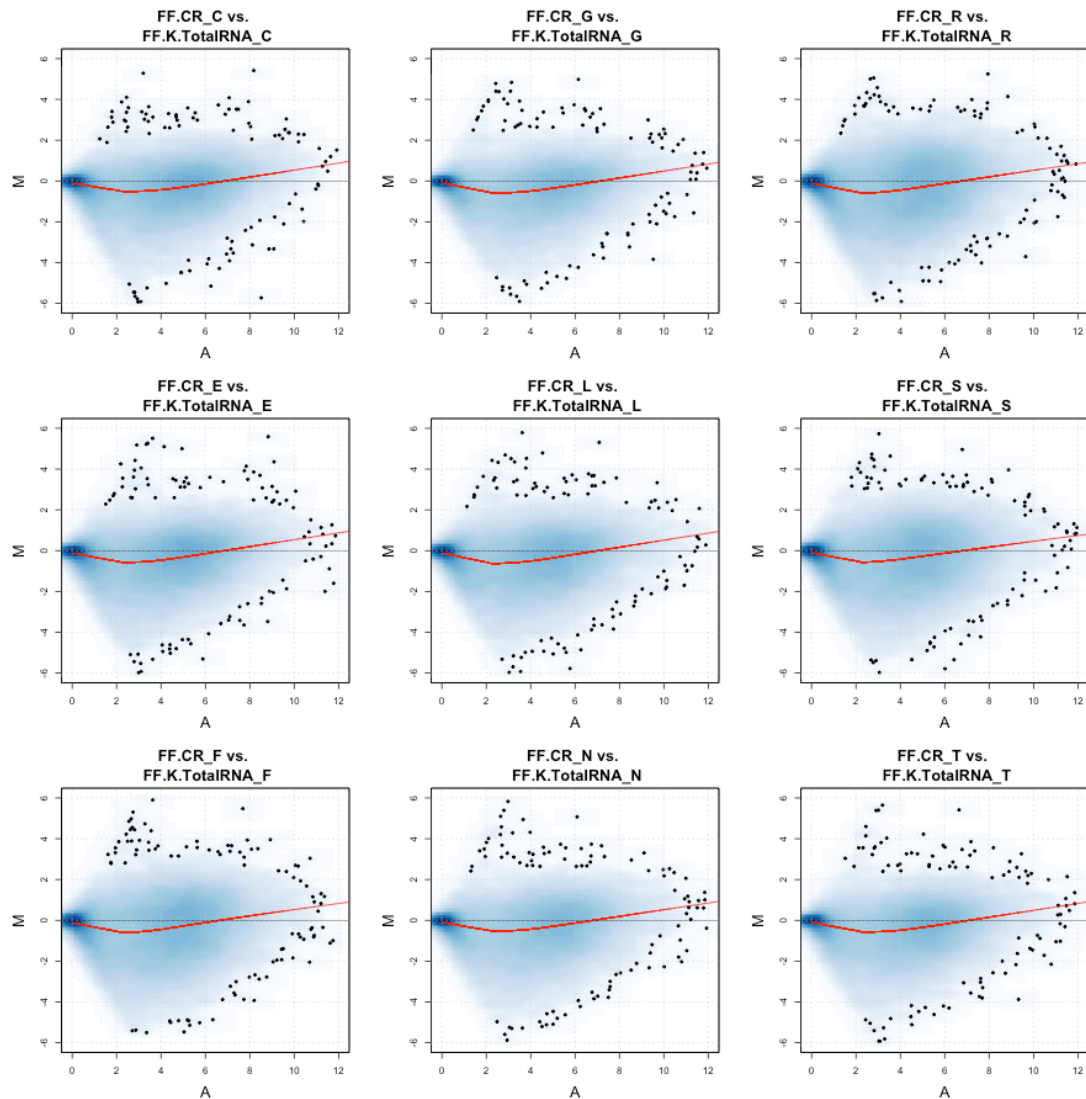


Figure 2.23: MA-plot for FF.CR protocol as compared to FF.K.TotalRNA for nine breast tumors. The red curve indicates the lowest smoother fitted to the data.

Further investigating these protocol-induced biases, we looked at the number of genes that would be considered as “differentially expressed” or “false positive” (FP), when we compared data from different library protocols with those from the FF reference standards (Figure 2.24 and 2.25). Fewer FP genes would suggest fewer artifacts

introduced by a protocol. FFPE.K.TotalRNA RNA-seq data, among all FFPE data, gave the fewest genes with significant expression differences at various p-value thresholds using different data normalization methods. In contrast, FF.CR, rather than a FFPE protocol, was the most biased method, with 84.2% of all genes identified as significantly different in expression from FF.mRNA at an adjusted p-value cutoff of 0.01. Also, there were fewer FP genes from FFPE.CR data when compared to FF.CR data, but not when either library used a non-CR protocol.

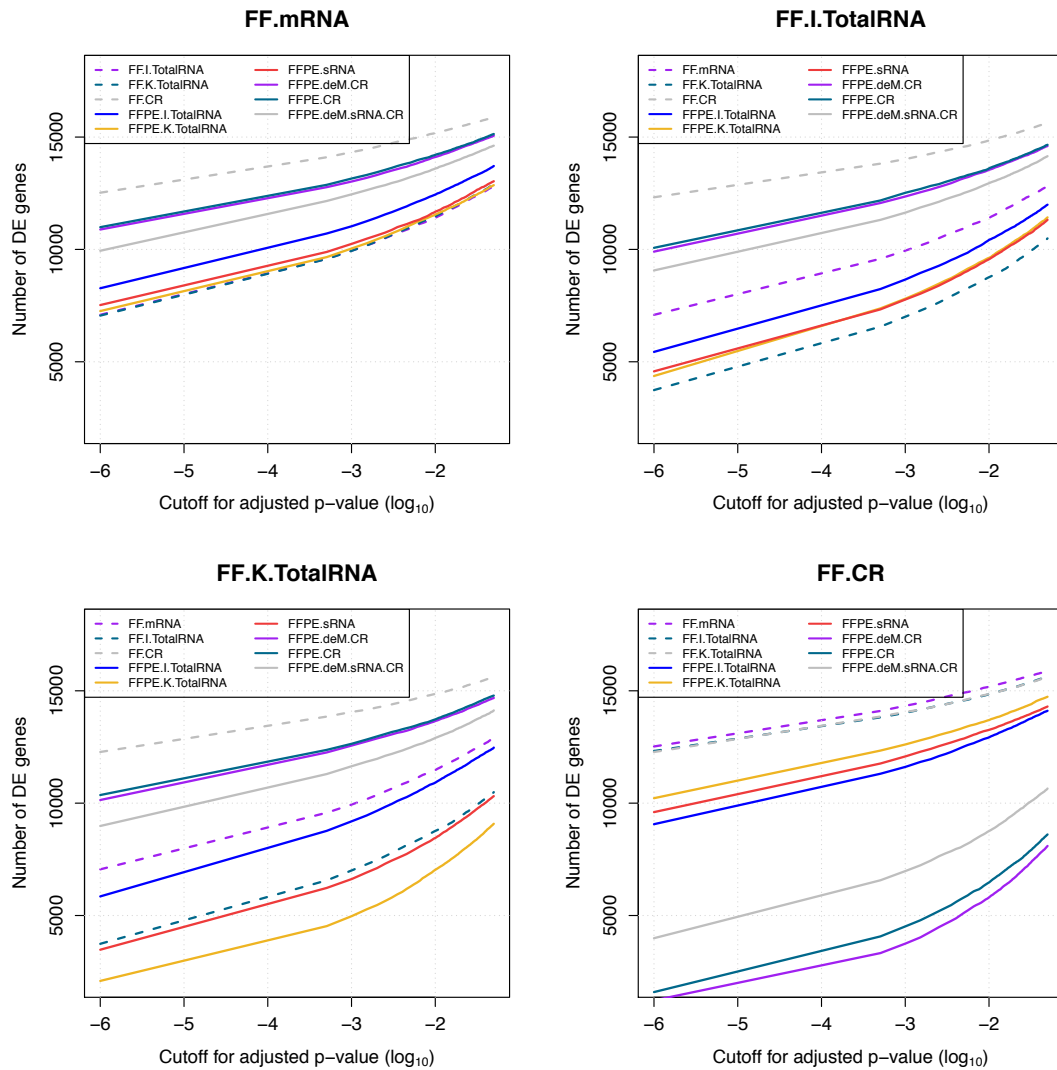




Figure 2.24: Number of genes identified to be differentially expressed between a reference FF protocol, as shown in the main title for each plot, and one of the other library preparation methods. A gene is considered as differentially expressed if its adjusted p-value from a test of differential expression is lower than the selected cutoff. The data is normalized by UQ method.

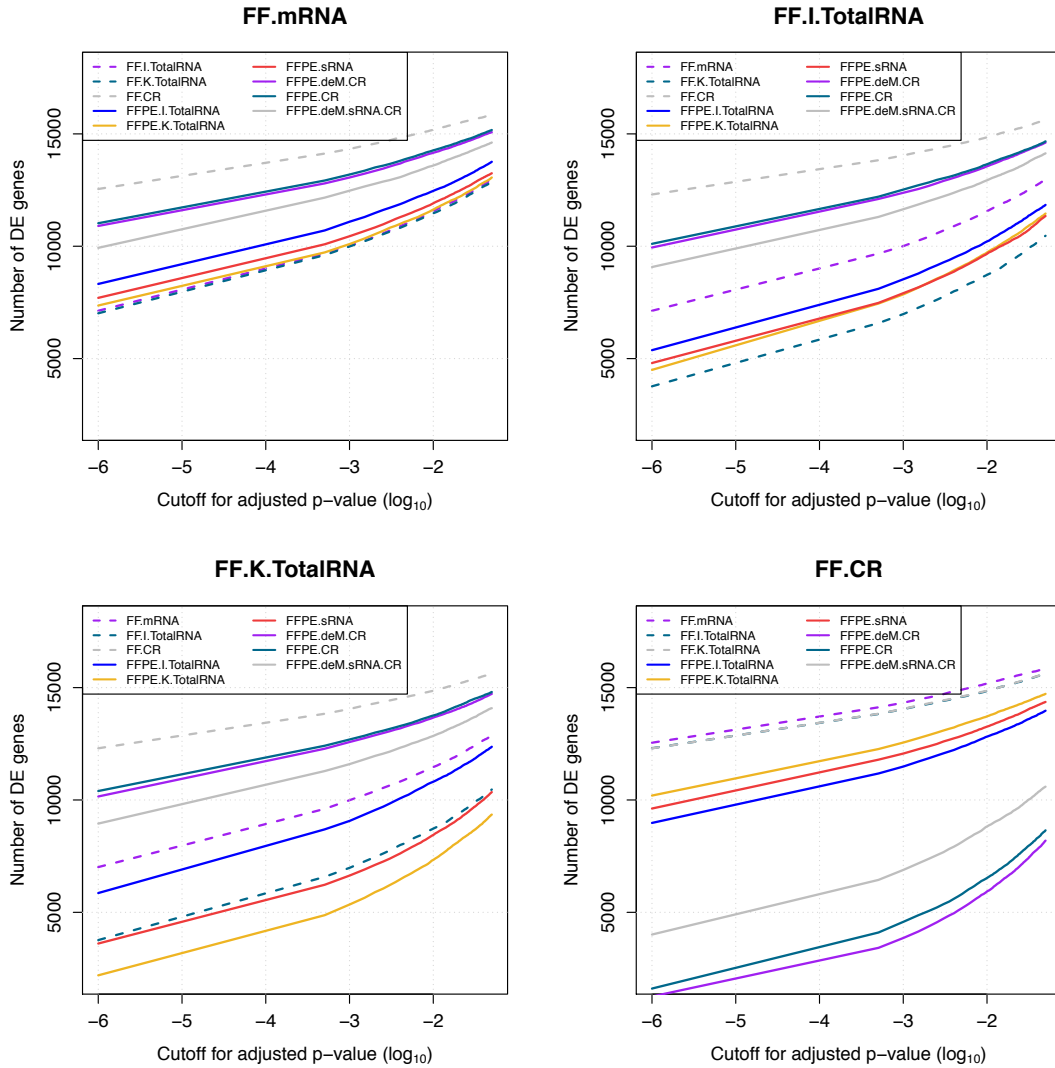


Figure 2.25: Number of genes identified to be differentially expressed between a reference FF protocol, as shown in the main title for each plot, and one of the other library preparation methods. A gene is considered as differentially expressed if its

adjusted p-value from a test of differential expression is lower than the selected cutoff.

The data is normalized by TMM method.

### 2.3.6 Pattern dissimilarity in measurement of coding sequence

We used a pattern dissimilarity score to measure the differences in expression patterns of CDS between library protocols, allowing direct comparison of non-CR and CR protocols. A smaller value of the score indicates higher similarity between a protocol and a FF reference. The distributions of dissimilarity scores across all genes were similar within each protocol, but varied across protocols (Figure 2.26). FFPE.K.TotalRNA had the lowest mean dissimilarity score when using FF non-CR libraries as the reference (Figure 2.27 and Table 2.4).

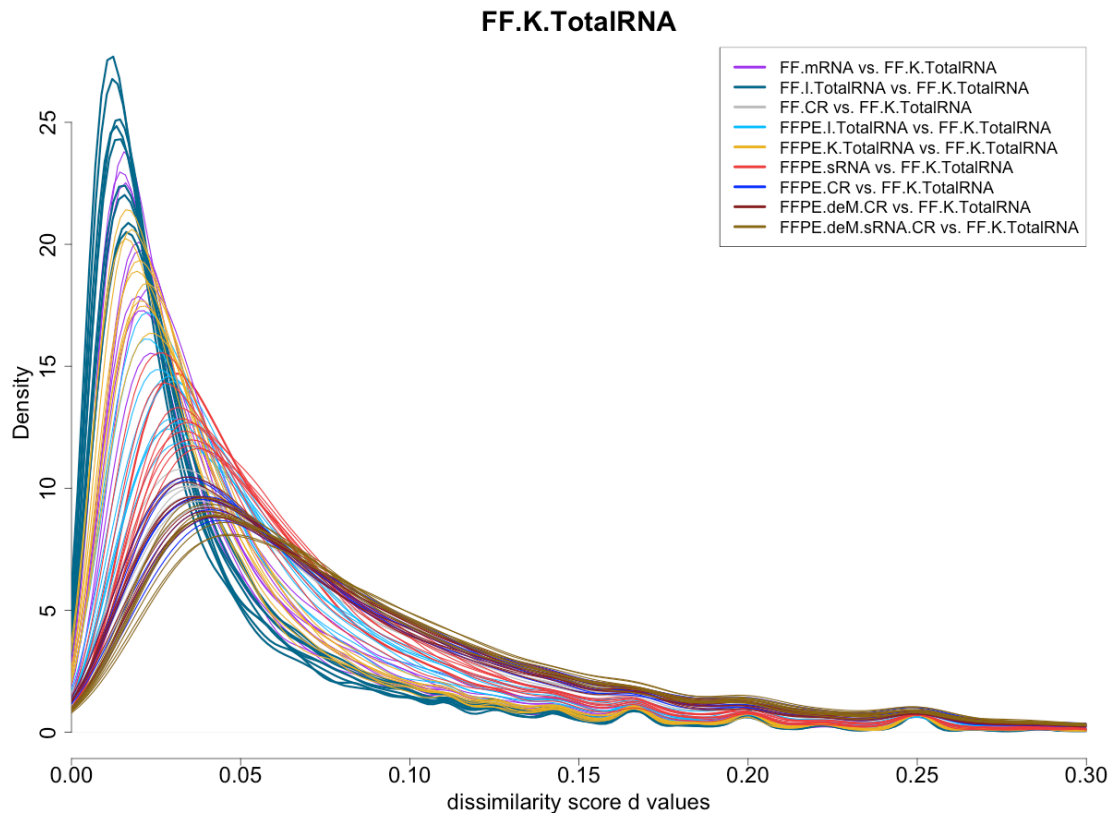


Figure 2.26: Distribution of dissimilarity score  $d$  values for all genes in each sample. Tumor samples processed by the same library preparation method are shown in the same color.

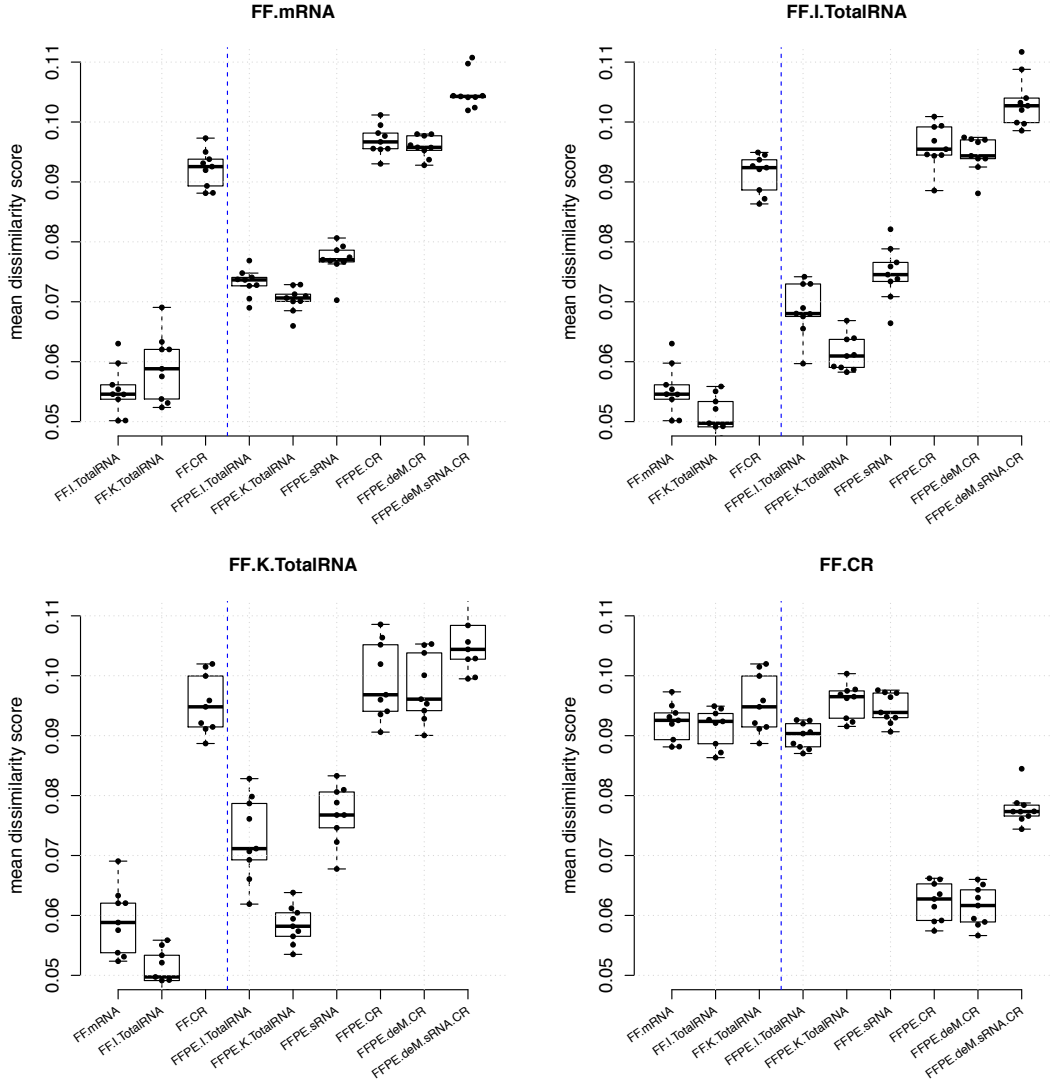


Figure 2.27: Boxplot of mean dissimilarity scores for CR-targeted poly(A)<sup>+</sup> genes with two or more expressed cds. Each point is the mean dissimilarity score calculated between a reference FF protocol, as shown in the main plot title, and one of the other library preparation methods.

Reference Case	FF.mRNA	FF.I.TotalRNA	FF.K.TotalRNA	FF.CR
FF.mRNA	-	-	-	-
FF.I.TotalRNA	0.055	-	-	-
FF.K.TotalRNA	0.059	0.05	-	-
FF.CR	0.093	0.092	0.095	-
FFPE.I.TotalRNA	0.074	0.068	0.071	0.09
FFPE.K.TotalRNA	<b>0.071</b>	<b>0.061</b>	<b>0.058</b>	0.097
FFPE.sRNA	0.077	0.075	0.077	0.094
FFPE.deM.CR	0.097	0.095	0.097	0.063
FFPE.CR	0.096	0.094	0.096	<b>0.062</b>
FFPE.deM.sRNA.CR	0.104	0.103	0.104	0.077

Table 2.4: Summary of the median of mean dissimilarity scores across nine tumor samples. The lowest median scores for FFPE protocols are highlighted in bold.

### 2.3.7 Gene expression patterns associated with tumor phenotype

We analyzed differential expression (DE) of genes comparing HR+/HER2- and TN breast cancers within each protocol. Overall, the normalized data were distributed around zero relative log expression, and were clustered by tumor phenotypes in the first two principal components. The p-value from DE analysis followed the ideal uniform distribution for non-DE genes, with a spike close to zero for the DE genes (Figure 2.28). ROC curves represented the sensitivity and specificity of the DE analyses using each FF reference as the gold standard. FFPE.K.TotalRNA achieved high and stable area under the curve (AUC) (0.921 - 0.933) at different cutoffs set for each FF gold standard, even after the strongest DE genes in the gold standards had been filtered out (Figure 2.29-2.32 and Table 2.5). The best agreement between FFPE protocols and each FF standards was as follows: FFPE.sRNA with FF.mRNA, FFPE.K.TotalRNA with both FF.I.TotalRNA and FF.K.TotalRNA, and FFPE.CR with FF.CR (Table 2.5).

Reference Case	FF.mRNA	FF.I.TotalRNA	FF.K.TotalRNA	FF.CR
FF.mRNA	-	-	-	-
FF.I.TotalRNA	0.977 / 0.975	-	-	-
FF.K.TotalRNA	0.977 / 0.976	0.987 / 0.987	-	-
FF.CR	0.963 / 0.962	0.967 / 0.966	0.968 / 0.966	-
FFPE.I.TotalRNA	0.919 / 0.917	0.929 / 0.923	0.931 / 0.929	0.917 / 0.913
FFPE.K.TotalRNA	0.921 / 0.918	<b>0.932 / 0.925</b>	<b>0.933 / 0.930</b>	0.910 / 0.907
FFPE.sRNA	<b>0.926 / 0.923</b>	0.928 / 0.922	0.933 / 0.929	0.904 / 0.901
FFPE.deM.CR	0.920 / 0.919	0.924 / 0.918	0.929 / 0.925	0.923 / 0.921
FFPE.CR	0.921 / 0.919	0.924 / 0.918	0.928 / 0.924	<b>0.928 / 0.925</b>
FFPE.deM.sRNA.CR	0.911 / 0.910	0.912 / 0.907	0.921 / 0.918	0.913 / 0.911

Table 2.5: Summary of median AUC values of between tumor phenotype differential expression using either UQ (left) or TMM (right) normalization. The highest median AUC values for FFPE protocols are highlighted in bold.

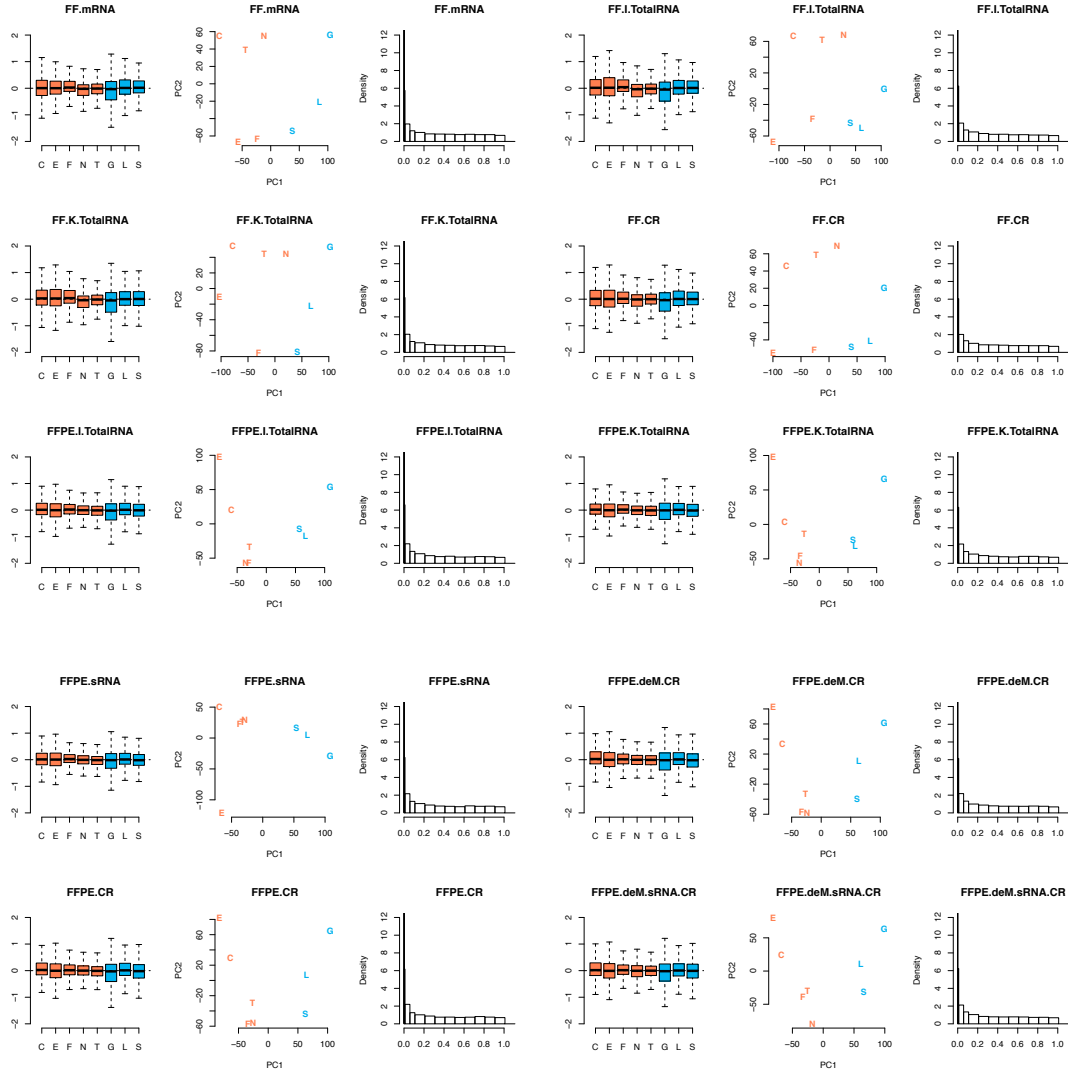


Figure 2.28: Relative log expression boxplot, principal component analysis, and p-value distribution for ER&PR positive (tumor ID: C, E, F, N, T) and triple negative (tumor ID: G, L, S) tumors for each library preparation group. The data were normalized by UQ method.

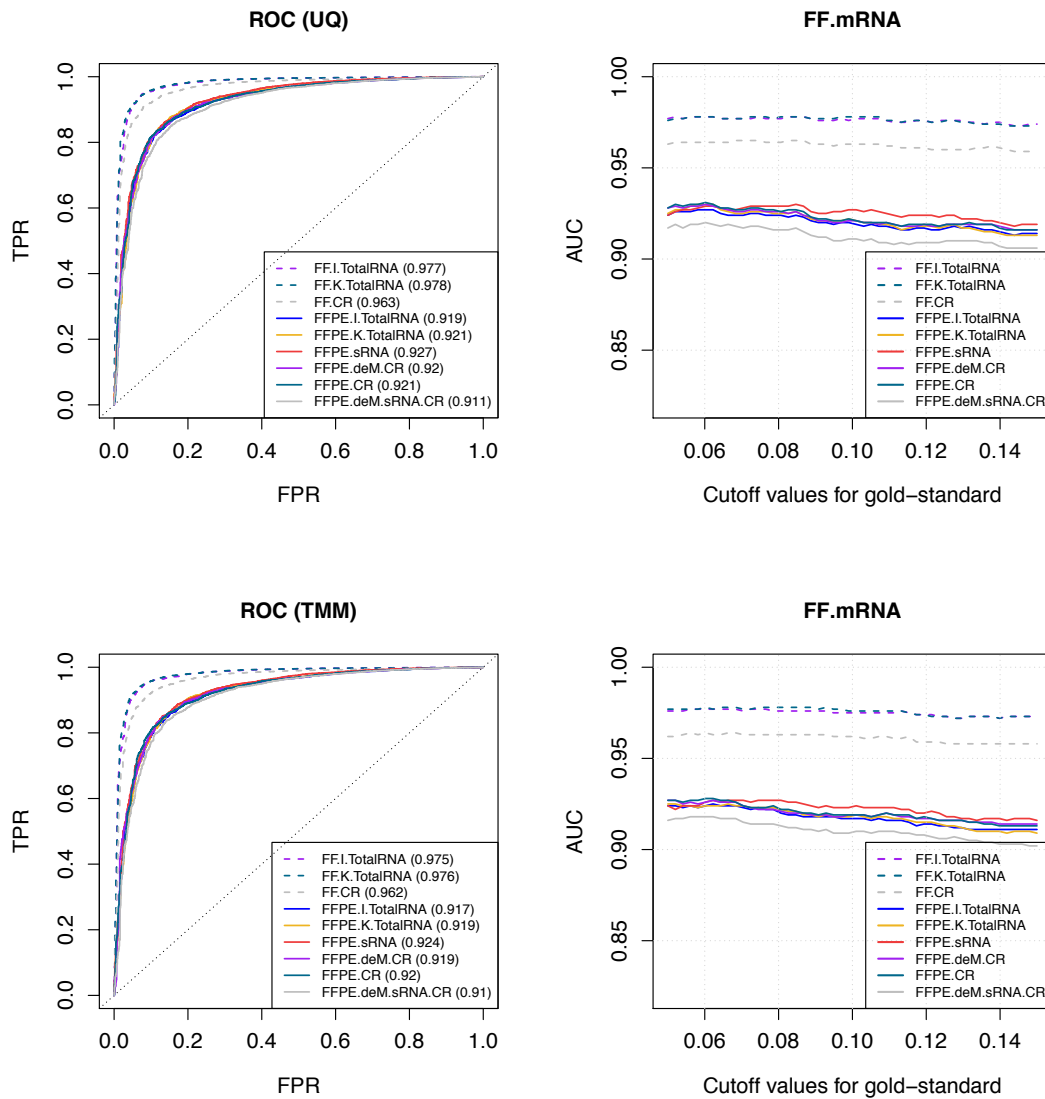


Figure 2.29: Between-tumor phenotype differential expression analysis results with FF.mRNA as the reference. **First row:** differential expression analysis results based on data normalized by UQ; **Second row:** differential expression analysis results based on data normalized by TMM; **Left column:** ROC curve for the differential expression analysis between ER+/PR+/HER2- and ER-/PR-/HER2- tumor samples. The adjusted p-value cutoff is 0.10 for gold standard, which is the FF.mRNA measures. AUC for each curve is included in the parenthesis in the figure legend. Abbreviations: TPR, true positive rate; FPR, false positive rate; **Right column:** Plot of AUC as a function of cutoff values for gold standard. Genes with adjusted p-value smaller than 0.01 in FF.mRNA group were removed.

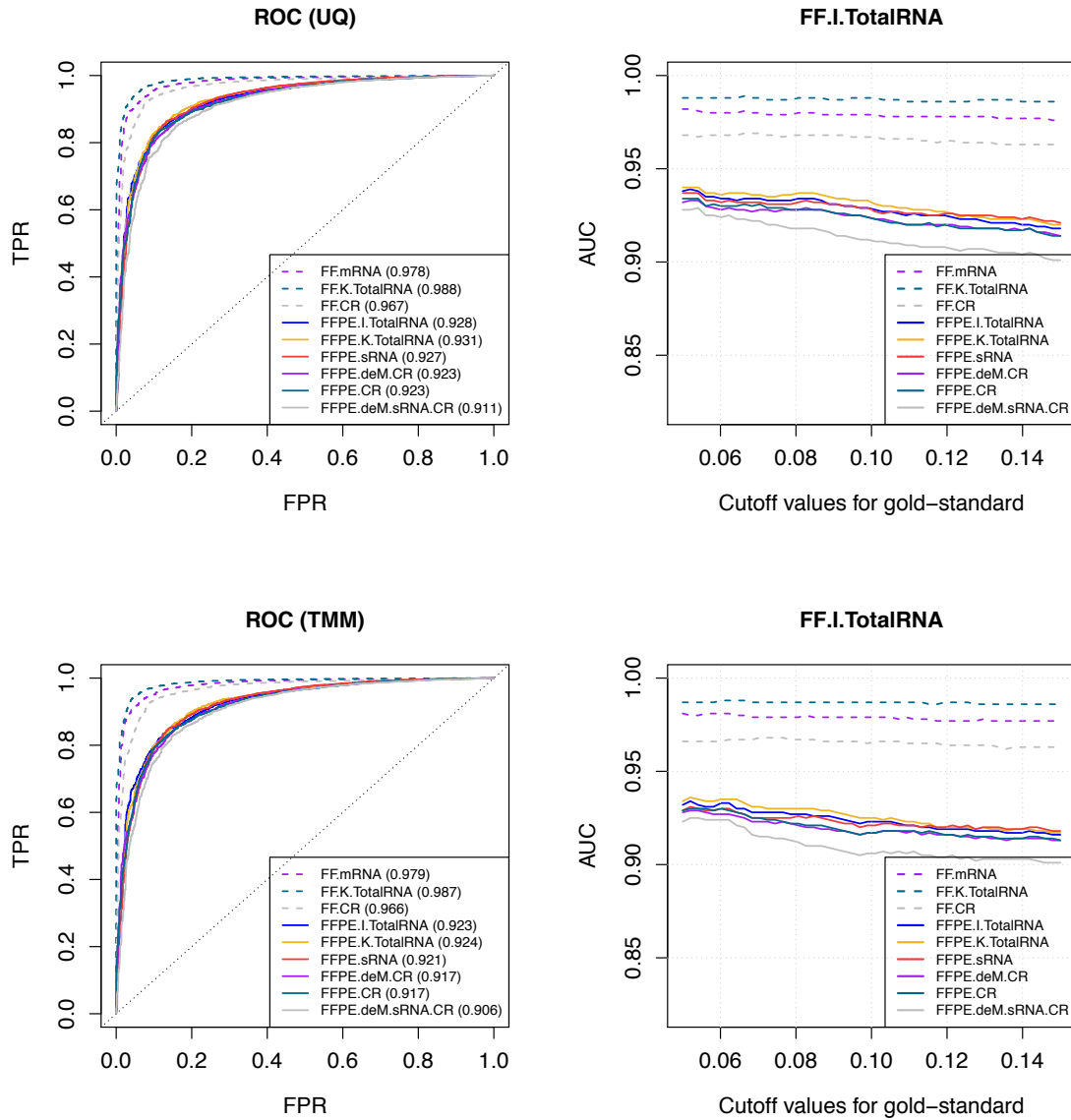


Figure 2.30: Between-tumor phenotype differential expression analysis results with FF.I.TotalRNA as the reference. **First row:** differential expression analysis results based on data normalized by UQ; **Second row:** differential expression analysis results based on data normalized by TMM; **Left column:** ROC curve for the differential expression analysis between ER+/PR+/HER2- and ER-/PR-/HER2- tumor samples. The adjusted p-value cutoff is 0.10 for gold standard, which is the FF.I.TotalRNA measures. AUC for each curve is included in the parenthesis in the figure legend. Abbreviations: TPR, true positive rate; FPR, false positive rate; **Right column:** Plot of AUC as a function of cutoff



values for gold standard. Genes with adjusted p-value smaller than 0.01 in FF.I.TotalRNA group were removed.

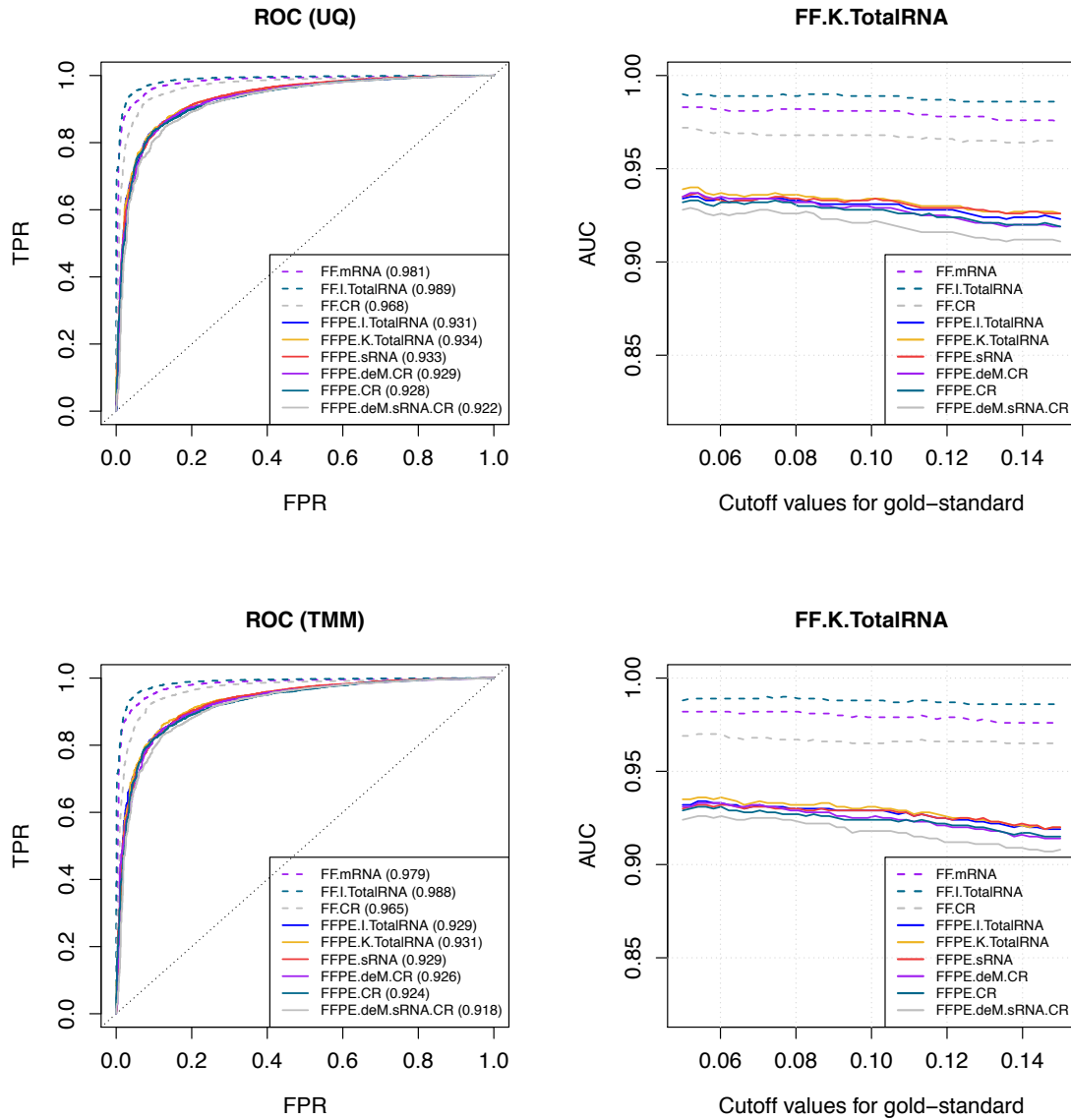


Figure 2.31: Between-tumor phenotype differential expression analysis results with FF.K.TotalRNA as the reference. **First row:** differential expression analysis results based on data normalized by UQ; **Second row:** differential expression analysis results based on data normalized by TMM; **Left column:** ROC curve for the differential expression

analysis between ER+/PR+/HER2- and ER-/PR-/HER2- tumor samples. The adjusted p-value cutoff is 0.10 for gold standard, which is the FF.K.TotalRNA measures. AUC for each curve is included in the parenthesis in the figure legend. Abbreviations: TPR, true positive rate; FPR, false positive rate; **Right column**: Plot of AUC as a function of cutoff values for gold standard. Genes with adjusted p-value smaller than 0.01 in FF.K.TotalRNA group were removed.

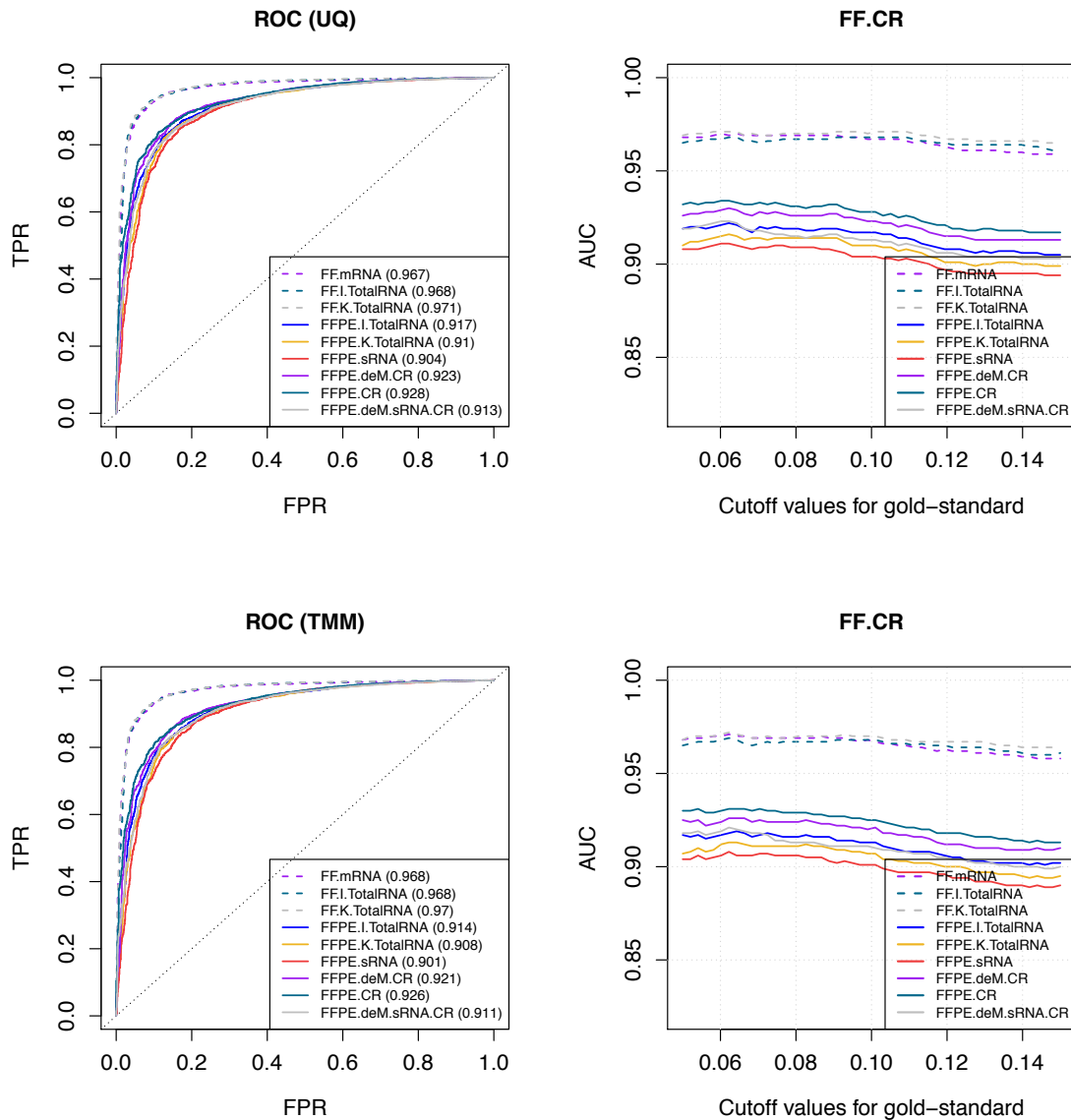


Figure 2.32: Between-tumor phenotype differential expression analysis results with FF.CR as the reference. **First row:** differential expression analysis results based on data normalized by UQ; **Second row:** differential expression analysis results based on data normalized by TMM; **Left column:** ROC curve for the differential expression analysis between ER+/PR+/HER2- and ER-/PR-/HER2- tumor samples. The adjusted p-value cutoff is 0.10 for gold standard, which is the FF.CR measures. AUC for each curve is included in the parenthesis in the figure legend. Abbreviations: TPR, true positive rate; FPR, false positive rate; **Right column:** Plot of AUC as a function of cutoff values for gold standard. Genes with adjusted p-value smaller than 0.01 in FF.CR group were removed.

### 2.3.8 Representative gene signatures of prognosis

We compared 5 published breast cancer gene expression signatures (recurrence score (Oncotype DX), PAM50, sensitivity to endocrine therapy (SET) index, mammaprint and PI3-kinase index (PI3K)) across the 3 FFPE protocols (I.TotalRNA, K.TotalRNA and sRNA) and 3 FF protocols as standards (mRNA, I.totalRNA and K.TotalRNA) (10, 67-71). Best correlations using FFPE protocols with FF.mRNA (range 0.911 - 0.934) were not as strong as with FF.I.TotalRNA (range 0.952 - 0.975) or FF.K.TotalRNA (range 0.956 - 0.986) protocols (Table 2.6). The FFPE.K.TotalRNA protocol had the highest observed Spearman correlation coefficient in 13 of these 15 comparisons.

FF reference	mRNA			I.TotalRNA			K.TotalRNA		
FFPE protocol	I.Total RNA	K.Total RNA	sRNA	I.Total RNA	K.Total RNA	sRNA	I.Total RNA	K.Total RNA	sRNA
Oncotype DX	0.909	0.93	<b>0.934</b>	0.97	<b>0.975</b>	0.96	0.969	<b>0.986</b>	0.974
PAM50	0.901	<b>0.911</b>	0.901	<b>0.953</b>	<b>0.953</b>	0.937	0.957	<b>0.972</b>	0.954
SET	0.898	<b>0.911</b>	0.898	0.94	<b>0.96</b>	0.942	0.936	<b>0.968</b>	0.95
Mammaprint	0.905	<b>0.924</b>	0.921	0.947	<b>0.952</b>	0.95	0.932	<b>0.956</b>	0.936
PI3K	<b>0.926</b>	0.909	0.912	0.956	<b>0.961</b>	0.953	0.955	<b>0.971</b>	0.954

Table 2.6: Summary of the median spearman correlation coefficients across nine tumor samples for five signature gene sets. The highest median scores for FFPE protocols are highlighted in bold.

## 2.4 Discussion

Overall, FFPE RNA-seq data reliably captured transcriptional profiles and differences in tumor phenotype-based expression in breast cancer samples, just not quite as well as FF RNA-seq data. Principle component analyses demonstrated the following order of variables influencing gene expression measurements from RNA-sequencing: i) whether the library preparation protocol used exon capture for coding region (CR); ii) whether the samples was from FF tissue or FFPE tissue; and iii) the biological phenotype of the breast cancer based on hormone receptors and HER2 receptor status (Figure 2.2). Generally, we observed small differences in performance between non-CR protocols. However, even small differences can have important effects on large-scale genomic data for biomarker discovery, validation or subsequent diagnostic development. Nevertheless, we identified one protocol, FFPE.K.TotalRNA, with consistently good transcript coverage uniformity and continuity; most concordant expression; and least differential expression when compared to the different non-CR protocols with fresh tissue. This

protocol utilized RNase H-based rRNA depletion method and outperformed another similar TotalRNA-seq method, which used RiboZero to remove rRNA. It had a reasonable requirement of total RNA input (100ng) for FFPE samples, which is crucial for studies using tumor biopsy samples.

The first translational research scenario that we posed, in the Background section, considered the best pairing of protocols that would enable discovery using FF samples with intention to later translate for use with FFPE samples. Overall, we favor the K.TotalRNA as consistently best, or close to best performance with FFPE protocols, when compared to FF.mRNA, FF.I.TotalRNA or FF.K.TotalRNA as reference FF protocols. This interpretation was supported by most parameters that we studied – including the quality of read coverage, pattern of coding sequence expression, translation of overall or phenotype-related gene expression profiles and prognostic signatures.

The CR protocols yielded concordant results, but very different from all other (non-CR) protocols. So a CR protocol used for discovery (FF) would preclude other protocols for later translation to FFPE samples (Figures 2, 4). Hence, future application of customized assays might also be biased. Also, changes to the population of exon capture probes within a commercial kit over time could be a potential risk to this approach.

The most generalizable results from FFPE samples were obtained using the Total.RNA protocols without exon capture. Although similar, the FFPE.K.TotalRNA protocol produced slightly stronger results than the FFPE.I.TotalRNA protocol. So for our second scenario, we prefer the K.TotalRNA protocol for best representation of the transcriptome in FFPE samples utilized for discovery research – aiming to represent the transcriptional information that FF samples would have provided.

Our third translational research scenario involves the translation of an existing gene expression signature that was previously developed using a different method (e.g. microarray) or a particular RNA-seq protocol. Again, the FFPE.K.TotalRNA protocol had the best performance for total transcriptional profile, coding sequence, phenotypic discrimination, and for specific gene expression signatures.

The formalin fixation process is known to cause cross-linkage between nucleic acids and proteins, and mono-methyl addition to the RNA bases(23). Although we tested a method of chemical de-modification of total RNA, our results showed negligible effect, and argue against the incorporation of this method for RNA-seq of FFPE samples (Figure 4). However, due to limited tumor sample total RNAs, we did not test the performance of potential protocols combining de-modification with sRNA alone or TotalRNA methods.

The inclusion of random and dT primers (sRNA protocol) to simulate the FF.mRNA protocol produced good concordance overall, but introduced non-uniformity and discontinuity of read coverage across the transcriptome. So there seems to be no advantage to incorporating these innovations for RNA-seq of FFPE samples.

Limitations to our study include small sample size (although cancers were selected to represent biologic diversity), optimally short time to fixation of tissues, and lack of generalizability (single institution conditions of tissue processing). Also, the effects of long-term storage of FFPE samples could not be tested – but would be expected from a completed clinical trial. Also, several of the cases had prolonged storage of cut FFPE sections (at 4°C) until RNA purification. This could have compromised the FFPE library protocols for this comparison, but can also be viewed as stress-testing the FFPE-derived RNA. Notwithstanding these limitations, we believe that the results from this study will

be helpful to translational researchers as they consider how to obtain accurate gene expression by applying RNA-seq methods to FFPE tumor samples.

### **3. A Bayesian estimation of semiparametric recurrent event model with applications to the penetrance estimation of multiple primary cancers in Li-Fraumeni Syndrome**

#### **3.1 Introduction**

In this chapter, we propose a Bayesian semiparametric recurrent event model based on NHPP. We define and exploit what we call a familywise likelihood in order to maximally utilize the genetic information shared by individuals within the same family. In particular, we apply the peeling algorithm(50) to evaluate the familywise likelihood with large amount of missing genotype information. The ascertainment-corrected joint model(72, 73) is used to correct the ascertainment bias.

The rest of this paper is organized as follows. In Section 3.2, we introduce the LFS family data that motivate this study. In Section 3.3, we provide an explorative analysis for the data to give a justification of our approach. In Section 3.4, we propose a semiparametric recurrent event model for MPC based on NHPP. In Section 3.5, we describe how to construct the familywise likelihood including the ascertainment bias correction in a great detail, and the posterior updating scheme via MCMC is given in Section 3.6. In Section 3.7, we apply the proposed method to the LFS data and the estimated age-at-onset MPC-specific penetrances then follow. We also carry out both internal and external validation analysis. Final discussions follow in Section 3.8.

#### **3.2 The Motivating Data**

The pediatric sarcoma cohort data consists of 189 unrelated families, with 17 of them being *TP53* positive families in which there is at least one *TP53* mutation carrier within the family (Table 3.1). The *TP53* status was determined by PCR of *TP53* exonic regions,



and once a mutation was identified from the proband, all of his/her first-degree relatives and any family members at risk of carrying the mutation were also tested. Among a total of 3,706 individuals, 964 (26.0%) of them had *TP53* mutation status testing results. The age at the diagnosis of each invasive primary tumor for each individual was recorded. The follow-up periods for each family ranges from 22-62 years starting from the ascertainment date of probands. Among 570 (15.4%) individuals with a history of cancer, a total of 52 (1.4%) had been diagnosed with more than one primary cancer (Table 3.2). In the data, we have approximately equal number of cancer patients or healthy individuals for the two genders. Further details on data collection and germline mutation testing can be found elsewhere(33, 74).

	With carriers	W/O carriers	Total
Number of families	17	172	189
Number of individuals	2,409	1,297	3,706
Average family size	141.71	7.54	19.61

Table 3.1: Summary of number of families of LFS data

Number of primaries	Gender	Wildtype	Mutation	Unknown
0	Male	295	9	1276
	Female	341	8	1207
1	Male	105	25	139
	Female	121	23	105
2	Male	3	9	8
	Female	3	12	5
3	Male	0	3	0
	Female	0	2	2
4	Male	0	2	0
	Female	0	1	0
5	Male	0	0	0
	Female	0	1	0
7	Male	0	0	0
	Female	0	1	0
Total number of individuals		868	96	2742
Total number of cancer patients		232	79	259
Total number of MPC patients		6	31	15

Table 3.2: Number of primary cancer patients in LFS data

### 3.3 Preliminary Analysis of the LFS Data

Let  $T_k$ ,  $k = 1, 2, \dots, K$  denotes the age of the  $k$ th primary cancer-onset (i.e., the age of diagnosis of  $k$ th primary cancer), and  $W_k = T_k - T_{k-1}$  denote the  $k$ th gap time between two adjacent primary cancers with  $T_0 = 0$ . A common issue in serial gap time analysis is that the censoring time, although independent of  $T_k$ , is dependent of  $W_k$  if gap times are associated with each other(75). The dependent censoring makes it inappropriate to fit marginal models for  $k$ th gap times  $W_k$  ( $k \geq 2$ ). For example, Cook *et.al* shows that ignoring dependent censoring can lead to underestimation of the survival functions of second and subsequent gap times(45). We therefore use the inverse probability of censoring weighted (IPCW) estimates of Kendall's  $\tau$  to assess the association between  $W_1$  and  $W_2$  in the LFS data after adjusting the induced dependent censoring issue(76).

We compute the Kendall's  $\tau$  using data without those from probands as these individuals are not randomly selected for genotype testing (detailed computation can be found in Appendix A). The estimated IPCW Kendall's  $\tau = -0.017$  (Jackknife se= 0.005), which indicates a very weak negative but statistically significant correlation between the two gap times within subjects.

Figure 3.1 shows Kaplan-Meier estimates of survival functions  $S_1(t) = \Pr(W_1 > t)$  and  $S_2(t) = \Pr(W_2 > t)$ , stratified by genotype. The risk set used for calculating  $S_2(t)$  considers only single and multiple primary patients starting from the first cancer, while  $S_1(t)$  includes all individuals. For both of the *TP53* mutation carriers and non-carriers or untested individuals, the lengths of the first and the second gap time are not identically distributed, with the first gap time significantly longer than the second one. This suggests a time trend in the process where the rate of event occurrence would increase with aging. Moreover, the mutation carriers appear to have different length distribution from wildtype and untested individuals. This empirical difference in successive survival again suggests the importance of providing subgroup-specific and MPC-specific penetrance.

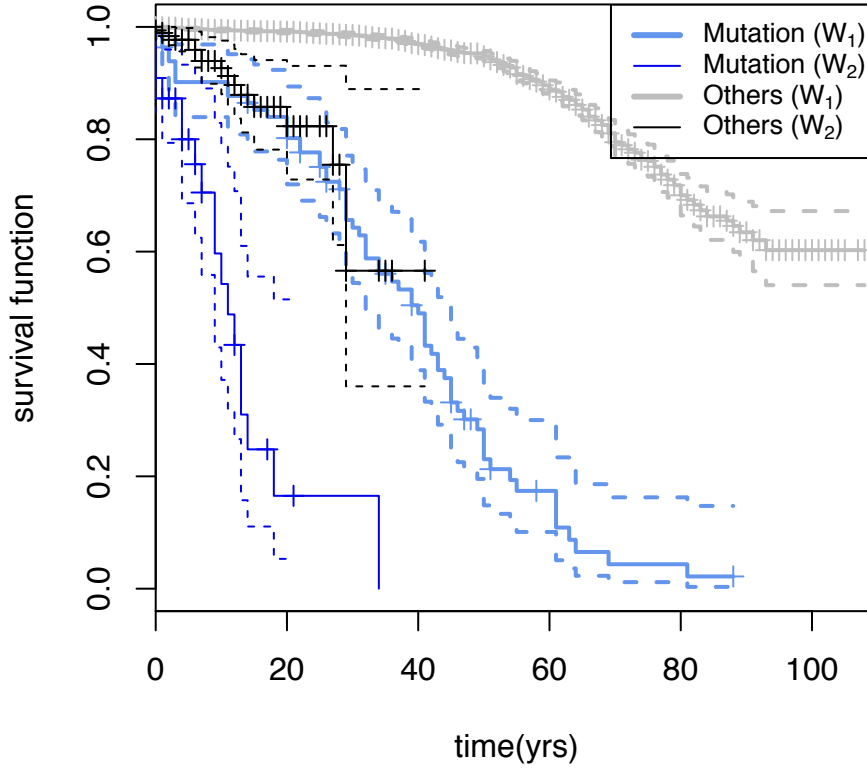


Figure 3.1: Kaplan-Meier estimates of survival distribution for the first ( $W_1$ ) or the second ( $W_2$ ) gap times after removing data from probands. The dashed lines are the 95% confidence bounds of the estimated survival function. Log rank test gave a p-value of  $< 10^{-7}$  either comparing the first and second gap time distribution for individuals with no mutations or unknown genotype of *TP53* (Others group), or comparing the first and second gap time distribution for individuals with a mutation in *TP53* (Mutation group).

### 3.4 The Model

#### 3.4.1 MPC-specific age-at-onset penetrance

The MPC-specific age-at-onset penetrance can be written as

(3.4.1)

$$Pr(W_k \leq w | T_{k-1}, \mathbf{X})$$

where  $\mathbf{X}$  denotes a vector of covariates. In particular, we set  $\mathbf{X}^T = (G, S, G \times S, C, G \times C)$ , where  $G$  and  $S$  denotes individual's genotype (0 for wildtype, 1 for *TP53* mutation), gender (0 for female, 1 for male), respectively, and  $C$  is one's cancer status at a specific age. Notice that  $C$  is a periodically fixed covariate during follow-up of an individual as its value will change at the age of cancer diagnosis. For example, let  $t_1$  and  $\tau$  denote the observed age of the first cancer onset and censoring time, respectively, then  $C$  is given by

$$C = \begin{cases} 0, & t \in [0, t_1) \\ 1, & t \in [t_1, \tau) \end{cases}$$

### 3.4.2 Semiparametric Recurrent Event Model for MPC

There are two canonical approaches in modeling recurrent events: one approach models the event counts via counting process and another approach models gap times via renewal process. We will use NHPP for our modeling because of its flexibility in dealing with our primary cancer data.

The NHPP-based approach directly models the number of primary cancers occurred by age  $t$  denoted by  $N(t)$ . The rate function of  $N(t)$  that characterizes the counting process  $\{N(t), t \geq 0\}$  is the probability of events occurring at time  $t$ , and is defined as

(4.2)

$$\lambda(t) = \lim_{\Delta t \rightarrow 0} \frac{Pr\{N(t + \Delta t) - N(t) > 0\}}{\Delta t}$$

Note that under Poisson process, the rate function is equivalent to the intensity function defined as

(4.3)

$$\eta(t) = \lim_{\Delta t \rightarrow 0} \frac{\Pr\{N(t + \Delta t) - N(t) > 0 | H(t)\}}{\Delta t}$$

where  $H(t)$  is the event history up to time  $t$  (45, 77). In particular, the NHPP assumes that  $N(t)$  for a given  $t$  follows a Poisson distribution:

(4.4)

$$\Pr(N(t) = k) = \frac{\Lambda(t)^k}{k!} e^{-\Lambda(t)}, \quad k = 0, 1, 2, \dots$$

where  $\Lambda(t)$  is a cumulative rate function defined as

$$\Lambda(t) = \int_0^t \lambda(u) du$$

Notice that if  $\lambda(t) = \lambda$ , NHPP becomes the homogeneous Poisson process.

Incorporating the covariates  $\mathbf{X}$ , we consider the following multiplicative model for the conditional rate function given  $\mathbf{X}$  denoted by  $\lambda(t|\mathbf{X})$ :

(4.5)

$$\lambda(t|\mathbf{X}) = \lambda_0(t) \exp(\boldsymbol{\beta}^T \mathbf{X})$$

where  $\lambda_0(t)$  is a baseline rate function,  $\boldsymbol{\beta}$  denotes the regression coefficient parameter vector associated with the covariate  $\mathbf{X}$ .

We assume that  $t \in [0, 1]$  without loss of generality. Toward modeling the baseline rate function  $\lambda_0(t)$ , we propose a nonparametric model for the cumulative baseline rate function  $\Lambda_0(t) = \int_0^t \lambda_0(s) ds$  via Bernstein polynomials. To be more precise, we approximate  $\Lambda_0(t)$  by Bernstein polynomials of degree  $M$  (78-80) as follows:

(4.6)

$$\Lambda_0(t) \approx \sum_{l=1}^M \omega_l \binom{M}{l} t^l (1-t)^{M-l}$$

where  $\omega_l = \Lambda_0\left(\frac{l}{M}\right)$  and  $\omega_l \leq \dots \leq \omega_M$  to ensure  $\Lambda_0(t)$  monotone increasing.

Introducing the following transformation of  $\gamma_1 = \omega_1$  and  $\gamma_m = \omega_m - \omega_{m-1}$  for  $m = 2, \dots, M$ , (4.6) can be equivalently rewritten as a linear function of  $\boldsymbol{\gamma}^T = (\gamma_1, \dots, \gamma_M)$

(4.7)

$$\Lambda_0(t) \approx \boldsymbol{\gamma}^T \mathbf{B}_M(t)$$

subject to  $\gamma_m \geq 0$ ,  $m = 1, \dots, M$ . Here  $\mathbf{B}_M(t) = (B_M(t, 1), \dots, B_M(t, M))^T$  denotes the beta distribution function with parameters  $m$  and  $M - m + 1$  evaluated at  $t$  (81). The baseline rate function  $\lambda_0(t)$  is then given by

(4.8)

$$\lambda(t) \approx \boldsymbol{\gamma}^T \mathbf{b}_M(t)$$

where  $\mathbf{b}_M(t)^T = (b_M(t, 1), \dots, b_M(t, M))$  denotes the beta density with parameters  $m$  and  $M - m + 1$  evaluated at  $t$ . Finally, we have

(4.9)

$$\lambda(t|\mathbf{X}) = \boldsymbol{\gamma}^T \mathbf{b}_M(t) \exp(\boldsymbol{\beta}^T \mathbf{X})$$

For the value of  $M$ , a large  $M$  provides more flexibility to model the shape of baseline rate function, but at the cost of increasing computations. We set  $M = 5$  as this works well in most studies(82).

Finally, the MPC-specific age-at-onset penetrance (4.1) is obtained by

(4.10)

$$\Pr(W_k \leq w \mid T_{k-1} = t_{k-1}, \mathbf{X}) = 1 - \exp \left( - \int_{t_{k-1}}^{t_{k-1}+w} \boldsymbol{\gamma}^T \mathbf{b}_M(u) \exp(\boldsymbol{\beta}^T \mathbf{X}) du \right)$$

since  $\Pr(W_k > w \mid T_{k-1} = t_{k-1}, \mathbf{X}) = \exp \left( - \int_{t_{k-1}}^{t_{k-1}+w} \lambda(u|\mathbf{X}) du \right)$

### 3.5 Computing Likelihood

In this work, computing likelihood is not trivial due to a large number of missing genotypes and the ascertainment bias. In this section we define a familywise likelihood and propose a way to correct the ascertainment bias.

Let  $\tau_{ij}$  and  $K_{ij}$  denote the censoring time and the total number of primary cancers developed for individual  $j = 1, \dots, n_i$  from family  $i = 1, \dots, I$ , respectively. Suppose we are given a set of data  $(\mathbf{t}_{ij}, \tau_{ij}, \mathbf{x}_{ij})$  where  $\mathbf{t}_{ij}^T = (t_{ij,k} : k = 1, \dots, K_{ij})$ , and  $\mathbf{x}_{ij}^T = (g_{ij}, s_{ij}, g_{ij} \times s_{ij}, c_{ij}, g_{ij} \times c_{ij})$  are observed covariates for  $j$ th individual. Note that  $t_{ij,k} = \tau_{ij}$  if the individual has not developed a primary cancer when being censored.

#### 3.5.1 Individual likelihood

Let  $t_{ij,k} = 0$  and  $\tau_{ij} \geq t_{ij,K_{ij}}$ , the likelihood contribution of the  $k$ th event since  $(k - 1)$ th event is

(5.1)

$$\lambda(t_{ij,k}) \exp \left( - \int_{t_{ij,k-1}}^{t_{ij,k}} \lambda(s) ds \right)$$

where  $\lambda(\cdot)$  in the integrand denotes the rate function with fixed covariates  $\mathbf{x}_{ij,k}$  for any time points in  $[t_{ij,k-1}, t_{ij,k})$ , during which the covariate  $c_{ij}$  is time-invariant. Note the  $\lambda(\cdot)$  is still time-varying within this time interval. See Cook *et.al* for more details on the derivation(45). We show that the likelihood of the  $j$ th individual of the  $i$ th family with primary cancer events at  $t_{ij}$ , denoted by  $L_{ij}(\boldsymbol{\theta})$ , is given by

(5.2)



$$L_{ij}(\boldsymbol{\theta}) \propto \left\{ \prod_{k=1}^{K_{ij}} \lambda(t_{ij,k}) \right\} \exp \left\{ - \sum_{k=1}^{K_{ij}} \int_{t_{ij,k-1}}^{t_{ij,k}} \lambda(s) ds \right\} \exp \left\{ - \int_{t_{ij,\square_{ij}}}^{\tau_{ij}} \lambda_{\tau_{ij}}(s) ds \right\}$$

where the covariate  $c_{ij}$  in  $\lambda_{\tau_{ij}}(\cdot)$  is the cancer status within the time interval  $[t_{ij,K_{ij}}, \tau_{ij})$ .

In our model, the full likelihood is extended by considering each event for each individual as one term of the likelihood in order to incorporate the periodically fixed covariates.

### 3.5.2 Familywise Likelihood

Assume data from different processes are independent given covariates, the likelihood for the  $i$ th family is given by

(5.3)

$$L_i(\boldsymbol{\theta}) = \prod_{j=1}^{n_i} L_{ij}(\boldsymbol{\theta})$$

This likelihood construction assumes that the covariates  $\mathbf{x}_{ij}$  are observed for every individual. However, in LFS data, most individuals have their *TP53* mutation status untested. Let  $\mathbf{g}_i = \{\mathbf{g}_{i,obs}, \mathbf{g}_{i,mis}\}$  and  $\mathbf{h}_i$  denotes the *TP53* genotype vector and cancer phenotype data (eg: cancer status and age of cancer diagnosis), respectively, for the  $i$ th family. For simplicity, we denote the  $i$ th family likelihood  $L_i(\boldsymbol{\theta})$  as  $L_i(\mathbf{h}_i|\mathbf{g}_i)$ . By the law of total probability, the likelihood for the observed data is

(5.4)

$$L_i(\mathbf{h}_i|\mathbf{g}_{i,obs}) = \sum_{\mathbf{g}_{i,mis} \in S} \Pr(\mathbf{h}_i, \mathbf{g}_{i,mis}|\mathbf{g}_{i,obs})$$

where  $S$  is a set of all possible values of genotypes  $\mathbf{g}_{i,mis}$  conditional on  $\mathbf{g}_{i,obs}$ . Because the set  $S$  increases exponentially with the number of individuals with missing genotype, we use Elston-Stewart's peeling algorithm to recursively calculate  $L_i(\mathbf{h}_i|\mathbf{g}_{i,obs})$  (50, 83, 84). The algorithm proceeds by “peeling” out nuclear families from the whole family and its computational complexity is approximately linear in the number of individuals with unknown genotype. A simple example of how the algorithm can improve the efficiency of likelihood calculation is given in the Appendix B. The likelihood for  $I$  independent families is then

(5.5)

$$L(\boldsymbol{\theta}) = \prod_{i=1}^I L_i\{\mathbf{h}_i|\mathbf{g}_{i,obs}\}$$

### 3.5.3 Ascertainment bias correction

The ascertainment bias exists in rare disease studies like LFS study because the data were collected from a high-risk population. The familywise likelihood (5.5) we construct is then a biased one for the LFS data. To estimate penetrances for a general population, we will need to correct for the ascertainment bias.

Let indicator variable  $A_i = 1$  denotes that the  $i$ th family is ascertained, and  $A_i = 0$  otherwise.  $A_i$  is a subset variable of  $\mathbf{h}_i$ . When  $Pr(A_i = 1)$  is independent of family history, we can assume no ascertainment bias. However, in the dataset with ascertainment bias, we estimate  $Pr(A_i = 1|\mathbf{x}_i, \boldsymbol{\theta})$  from the data. We use ascertainment-corrected joint

model to correct the bias(72, 73). Ascertainment bias of  $i$ th familywise likelihood is corrected by inverse weighting of the probability that  $i$ th family is ascertained

(5.6)

$$Pr(\mathbf{h}_i | A_i = 1, \mathbf{g}_i, \mathbf{x}_i, \boldsymbol{\theta}) \propto \frac{Pr(\mathbf{h}_i | \mathbf{g}_i, \mathbf{x}_i, \boldsymbol{\theta})}{Pr(A_i = 1 | \mathbf{x}_i, \boldsymbol{\theta})}$$

Assuming  $j = 1$  in each family is the proband, the weight can be calculated as,

(5.7)

$$Pr(A_i = 1 | \mathbf{x}_i, \boldsymbol{\theta}) = \sum_{h_{i1}} Pr(A_i = 1 | \mathbf{h}_{i1}) Pr(\mathbf{h}_{i1} | \mathbf{x}_i, \boldsymbol{\theta})$$

Because in the LFS data, we ascertained a family by the fact the proband was diagnosed with a primary cancer. The weight can be rewritten as,

(5.8)

$$Pr(A_i = 1 | \mathbf{x}_i, \boldsymbol{\theta}) = \sum_{g_{i1} \in \{0,1\}} Pr(h_{i1} | g_{i1}, \mathbf{x}_i, \boldsymbol{\theta}) Pr(g_{i1})$$

where  $Pr(h_{i1} | g_{i1}, \mathbf{x}_i, \boldsymbol{\theta})$  is the data likelihood for the proband. The probability of genotype  $Pr(g_{i1})$  can be calculated based on the mutated allele frequency  $\phi_A$ . In the case of autosomal dominant inheritance disease,  $Pr(g_{i1} = 0) = (1 - \phi_A)^2$  and  $Pr(g_{i1} = 1) = (1 - \phi_A)^2$ . The ascertainment bias-corrected familywise likelihood in our study is then given by

(5.9)

$$L(\boldsymbol{\theta}) = \prod_{i=1}^I \frac{L_i(\mathbf{h}_i | \mathbf{g}_{i,obs}, \mathbf{x}_i, \boldsymbol{\theta})}{Pr(A_i = 1 | \mathbf{x}_i, \boldsymbol{\theta})}$$

### 3.6 Posterior Sampling through MCMC

Let  $Pr(\boldsymbol{\theta})$  denotes the prior distribution of  $\boldsymbol{\theta}$ , our goal is to estimate  $\boldsymbol{\theta}$  from the posterior distribution, which is given by

(6.1)

$$Pr(\boldsymbol{\theta} | \mathbf{H}) \propto Pr(\boldsymbol{\theta}) Pr(\mathbf{H} | \boldsymbol{\theta})$$

We set an independent normal prior for  $\boldsymbol{\beta}$  where  $\beta \sim N(0, \sigma^2)$ , and  $\sigma = 100$  for vague priors. We assign noninformative flat priors for  $\gamma$ . We use a random-walk Metropolis-Hastings algorithm within Gibbs to generate 50,000 posterior estimates in total with first 5,000 as burn-in.

### 3.7 Case Study

We applied our method to the LFS data (Section 3.2, 3.3) and estimated the parameters using the MCMC algorithm as described in Section 3.6. We performed a cross-validation, in which we compared our prediction of a 5-year risk for developing the next cancer given cancer history and genotype information for an individual with the observed outcome, based on our penetrance estimates. We also compared our penetrance results with population estimates and previous studies on *TP53* penetrance.

#### 3.7.1 Model fitting

We fit our model to the LFS data up to the second cancer event due to limited number of individuals with third or more cancers in this dataset (Table 3.2). Our model contains

three relevant covariates including genotype ( $G$ ), gender ( $S$ ) and cancer status ( $C$ ). We also include two interaction effects on genotype. The mutated allele frequency  $\phi_A$  is set as  $\phi_A = 0.0001$  in this study. Sensitivity prior analysis of the Bayesian estimation shows that the posterior parameter estimates are insensitive to the setting of  $\gamma$  prior distributions or hyper-parameters.

### **3.7.2 Cancer risk prediction**

#### **3.7.2.1 Internal validation**

We assessed our model in cancer risk prediction using a 10-fold cross-validation. We randomly split the 189 families into 10 portions. Our model was repeatedly fit to the 9 portions of all families to estimate the penetrance, based on which we made prediction using remaining 1 portion of the data. The individuals used for prediction are those who have known genotype information. We removed the probands because they were not randomly selected for genotype testing. We rolled back five years from the age of diagnosis of cancer or the censoring age. Based on the rolled-back time, we then calculated a 5-year cumulative cancer risk. We made two types of risk prediction that are of clinical interests. In the first scenario, we predicted the 5-year risk of developing a cancer given that the individual has no history of cancer (affected versus unaffected). In the second scenario, we predicted the risk of developing next cancer when the individual already developed a cancer before (MPC versus SPC). We combined the results from these 10-fold of cross-validation together and evaluated them using the receiver operating characteristic (ROC) curves. To assess the variation of prediction caused by data partition, we performed 25 times of the random splits for cross-validation. Figure 3.2 shows the

results on risk prediction from each random split. The median area under the curve (AUC) is 0.810 for predicting the status of being affected by cancer over healthy status, given that the individual has no cancer before. The median AUC is 0.706 for predicting the status of next cancer when the subject has a history of cancer. The validation performance is robust to random splits.

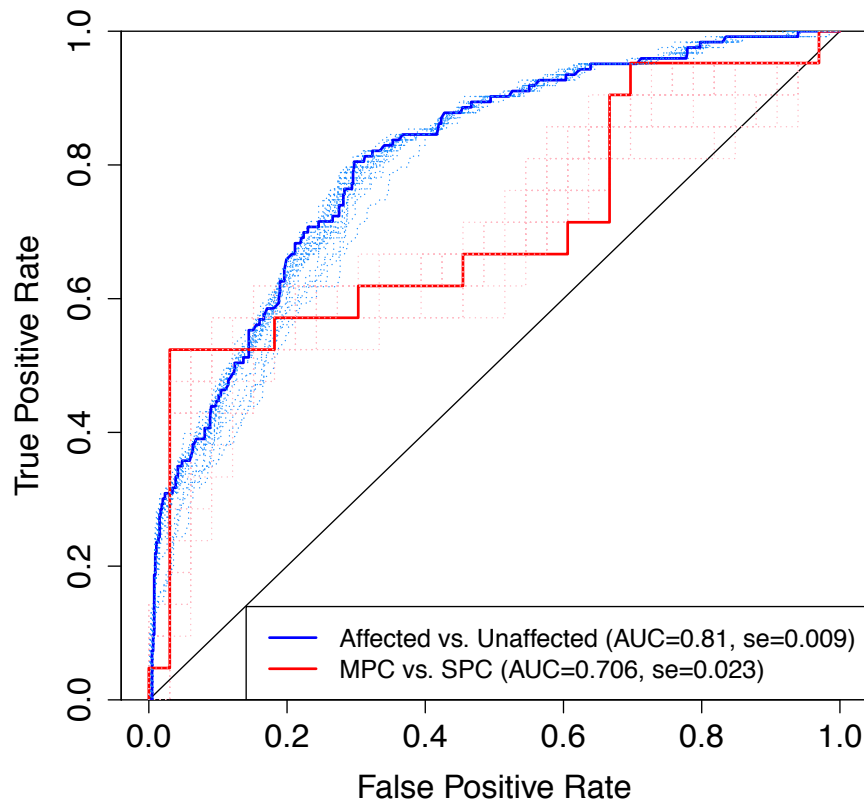


Figure 3.2. ROC of 5-year risk of developing next primary cancer assessed by 10-fold cross-validation. The dotted lines are the ROC curves generated from 25 times of random splits of the data for cross-validation, and the solid line is the one with median AUC value. Sample size:  $n(\text{Affected})=123$ ,  $n(\text{Unaffected})=643$ ,  $n(\text{MPC})=21$ ,  $n(\text{SPC})=33$ . Abbreviation: se, standard error.

### 3.7.2.2 External validation

We used the MD Anderson prospective data, collected independently from the model training data, for model prediction performance validation. These data have the same inclusion criteria, cancer diagnosis confirmation, mutation testing method as that for MD Anderson pediatric data. The number of primary cancers in this data is summarized in Table 3.3. We only used the individuals with available genotype information for validation purpose.

		Wildtype	Mutation
Healthy individuals	Male	95	27
	Female	115	21
SPC	Male	56	30
	Female	116	62
MPC	Male	35	35
	Female	112	70
Total number of individuals		529	245

Table 3.3: Number of primary cancer patients by the *TP53* mutation status and gender in MD Anderson prospective data. Abbreviations: SPC, single primary cancer patients; MPC, multiple primary cancer patients.

We evaluated the model prediction performance on primary cancer risk using the average annual risk computed using our penetrance estimates. The risk was calculated as the cumulative probability of developing next primary cancer divided by the follow-up time. The receiver operating characteristic (ROC) curve was used to evaluate the sensitivity and specificity of predicting a primary cancer incidence using the estimated risk probability at various cutoffs. Such model discrimination evaluation method has also been used for pancreatic cancer risk prediction in a previous study(85). For Kaplan-Meier

(KM) method-based risk prediction, we obtained KM survival functions for the time from date of birth to first primary cancer, and the time from first primary cancer diagnosis age to second primary cancer diagnosis age, respectively. These survival probabilities were then converted to penetrance estimate to compute the average annual risk. We used Jackknife to compute the standard error of prediction performance(86, 87). In brief, each subsample was generated by omitting the  $i$ th family and the AUC was calculated for this subsample as previously described. The standard error (se) was calculated using the Jackknife technique

$$se_{Jackknife} = \sqrt{\frac{n-1}{n} \sum_{i=1}^n (AUC_i - \overline{AUC})^2}$$

where  $n$  is the number of families, and  $\overline{AUC}$  is the mean estimate of AUC values among all Jackknife subsamples. As shown in Figure 3.3, our model achieves better performance compared to that of KM method for predicting either the first primary cancer (AUC: 0.754 vs. 0.698) or the second primary cancer (AUC: 0.731 vs. 0.658).



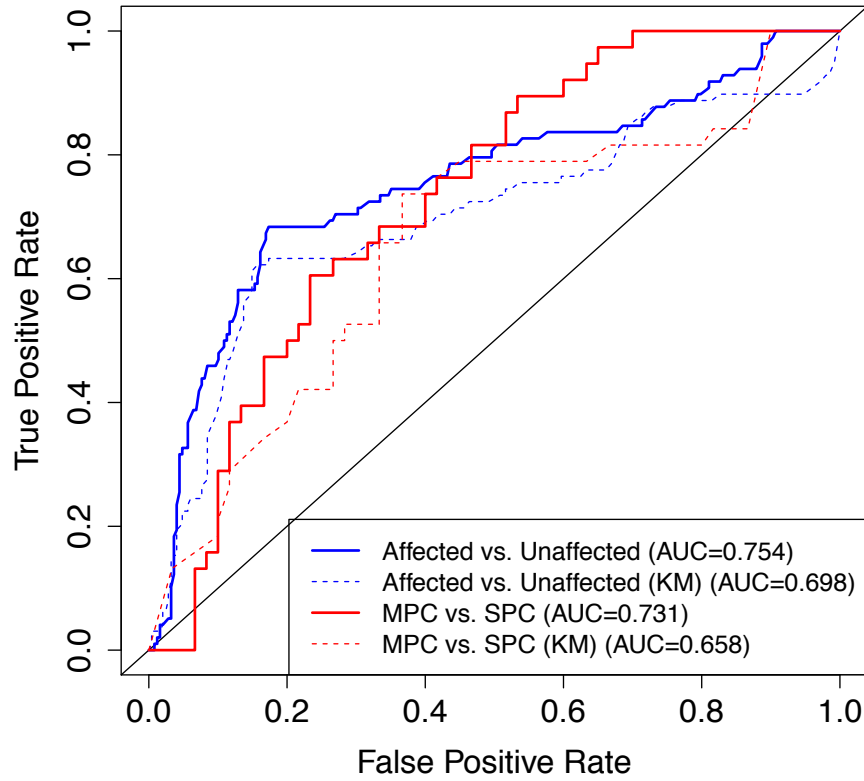


Figure 3.3: Comparison of validation performance between our multiple primary cancer-specific penetrance and those estimated from Kaplan-Meier (KM) method in predicting the first or the second primary cancer occurrence using the MD Anderson prospective data. Sample size:  $n(\text{Affected})=98$ ,  $n(\text{Unaffected})=248$ ,  $n(\text{MPC})=38$ ,  $n(\text{SPC})=60$ . Standard error:  $se(\text{Affected vs. Unaffected})=0.028$ ,  $se(\text{Affected vs. Unaffected (KM)})=0.032$ ,  $se(\text{MPC vs. SPC})=0.046$ ,  $se(\text{MPC vs. SPC(KM)})=0.055$ .

### 3.7.3 The MPC penetrance estimates

We applied the proposed method to the entire pediatric sarcoma cohort dataset to obtain penetrance estimates for single and multiple primary cancers given mutation status in *TP53*. We used Bayesian information criterion (BIC) for model selection. Table 3.3

shows that two models have best goodness-of-fit to the data. We decided to use the model with the interaction effect on gender  $G \times S$  as it has been reported that gender has different effects on cancer risk for mutation carriers and non-carriers(33). All posterior estimates of the model generated from MCMC converged well and had reasonable acceptance ratios. The summary of posterior estimates is shown in Table 3.4. The genotype has dominant effects on increasing cancer risk, both through main effect and interaction with the cancer history, as expected from the exploratory analysis (Section 3.3). Figure 3.4 illustrates the age-at-onset penetrance for a female and male individual over all ages with specified cancer history and mutation status.

Model	BIC
$\{G, S, C\}$	2807
$\{G, S, C, G \times S\}$	2805
$\{G, S, C, G \times C\}$	2800
$\{G, S, C, G \times S, G \times C\}$	2800
$\{G, S, C, G \times S, G \times C, S \times C\}$	2807

Table 3.4 Summary of BIC for model selection

Coefficient	Median	sd	95% CI	AR
$\beta_G$	3.016	3.016	(2.618, 3.391)	3.016
$\beta_S$	0.298	0.298	(0.024, 0.573)	0.298
$\beta_{G \times S}$	-0.721	-0.721	(-1.220, -0.183)	-0.721
$\beta_C$	-1.765	-1.765	(-2.949, -0.713)	-1.765
$\beta_{G \times C}$	2.117	2.117	(1.006, 3.340)	2.117

Table 3.5 Summary of posterior estimates. Abbreviations: sd, standard deviation; AR, acceptance ratio.

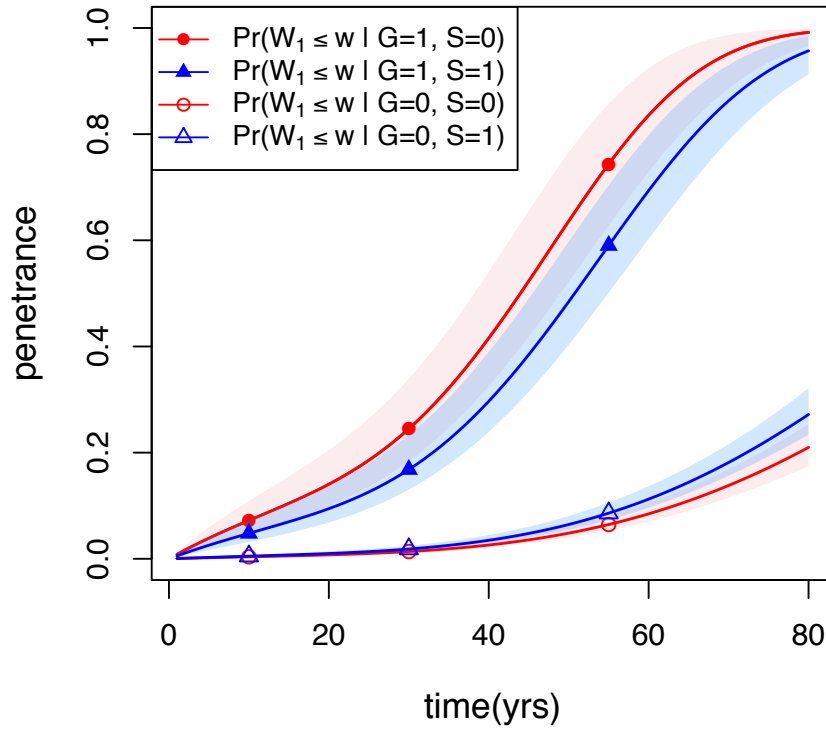


Figure 3.4: Age-at-onset penetrance for females or males without a history of cancer. The shaded area is the 95% credible bands.

For the second primary cancer risk, our penetrance estimates show that having a primary cancer developed before could have a positive effect on increasing cancer occurrence rate for mutation carriers but not for non-carriers (Table 3.5), with a hazard ratio of  $\exp^{2.117} = 8.3$ . The second primary cancer risk is also associated with the age of first primary cancer diagnosis, with a higher cancer risk for older first primary cancer diagnosis age (Figure 3.5 and Table 3.6).

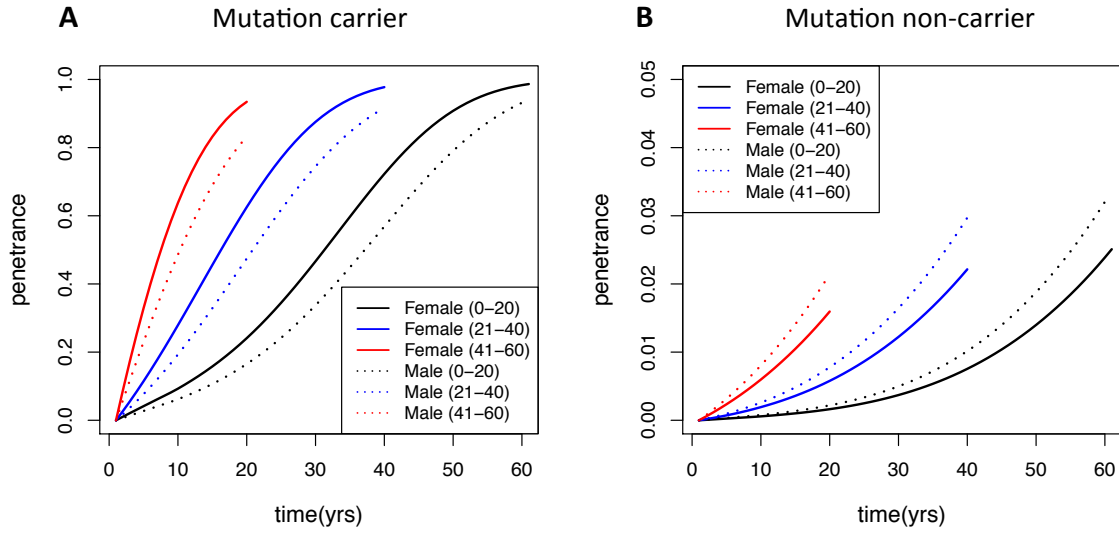


Figure 3.5: Penetrance estimates of the second primary cancer since the first primary cancer diagnosis time, stratified by the first primary cancer diagnosis time and gender, for A) *TP53* mutation carriers and B) non-carriers. Each curve represents the median penetrance estimates among all penetrance estimates in the first primary cancer diagnosis time group. The figures only show penetrance estimates up to age 80. Note that the two figures have different y-axis scales.

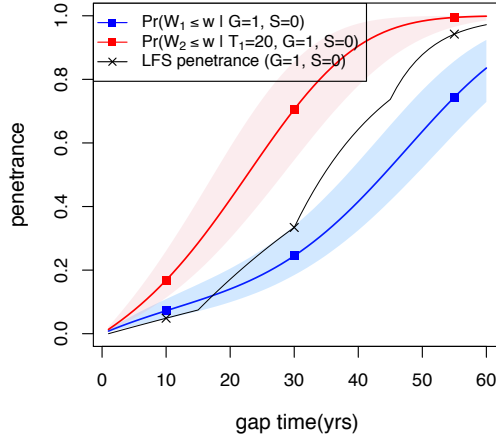
Age of diagnosis of the first primary cancer	Female	Male
0 - 20	29 (25-37)	35 (30-42)
21 - 40	14 (11-19)	19 (14-25)
41 - 60	6 (4-9)	8 (6-13)

Table 3.6: Median second primary cancer-free times since the first primary cancer diagnosis age and their 95% confidence intervals (in parenthesis) estimated for *TP53* mutation carriers, stratified by gender and first primary cancer diagnosis age.

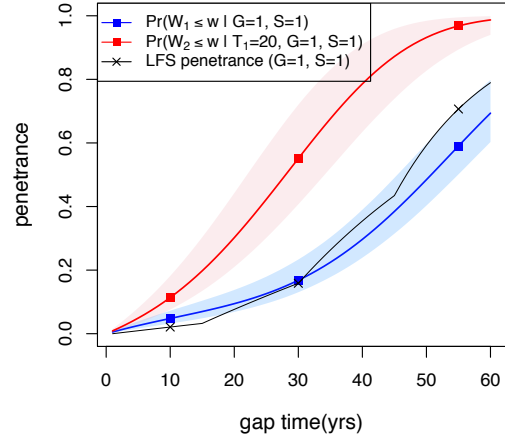
#### 3.7.4 Comparison with penetrance estimates from literature

Figure 3.4 compares penetrance estimates at different ages for females and males, stratified by genotype, respectively. As expected, *TP53* mutation has a clear effect on the increase of cancer risk, especially when the individual has a recent history of cancer. For a wildtype subject, a history of cancer does not have positive effect on increasing the risk of developing a subsequent cancer.

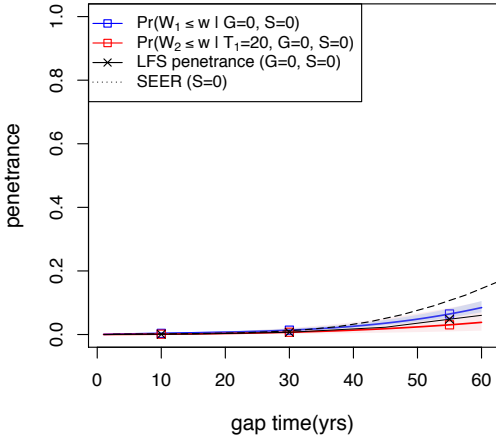
Wu *et.al* estimated *TP53* penetrance from six pediatric sarcoma families for both mutation carriers and non-carriers(34). This estimate can be considered as a weighted average of probability for SPC and MPC patients. Figure 3.4 shows that, for mutation carriers, this age-at-onset *TP53* penetrance estimate lies between those from cancer survivors and non- cancer survivors, as it should be. For non-carriers, the previous estimates are very slightly lower than our estimates for individuals without cancer history, but higher than those with early age at first diagnosis. When comparing with population estimates from Surveillance, Epidemiology, and End (SEER) Results program(88), our estimates for non-carriers overlap with the SEER estimate.



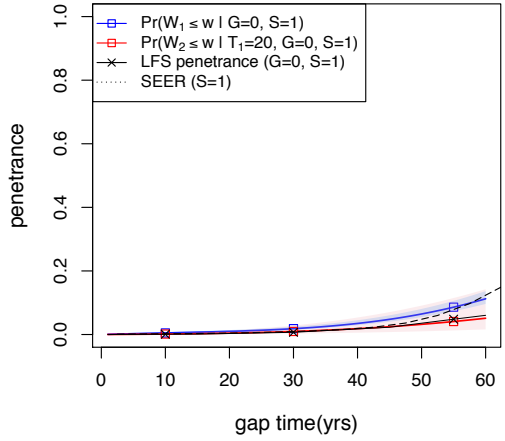
(a)



(b)



(c)



(d)

Figure 3.6: Age-at-onset penetrance when with or without a history of cancer for (a) female mutation carriers, (b) male mutation carriers, (c) female mutation non-carriers and (d) male mutation non-carriers. The shaded area is the 95% credible bands.

### 3.8 Discussions

To our knowledge, this is the first attempt to estimate MPC-specific penetrance for germline mutation in *TP53* with a large amount of missing genotype information in individuals that are genetically related. We developed a novel NHPP incorporated with

familywise likelihood so that it can model MPC events in the family context, while properly accounting for age effect and time-varying cancer status. A Bayesian framework was applied to estimate unknown parameters in the model. We also adjusted for ascertainment bias in the likelihood calculation so our penetrance estimates can be compared to those generated from the general population. Our new method provides a flexible framework for the penetrance estimation of MPC data, and shows reasonable predictive performance of cancer risk. As the number of multiple primary cancer patients becomes increasing in the general population, our method will be useful for prediction and clinical management of such diseases.

We are still left with a few possible extensions. First, we restricted our analysis up to the second primary cancer because of limited power in LFS data for the wildtype and untested groups. This makes our penetrance estimation unsuitable for individuals with a history of two or more cancers. It is straightforward to extend our model to account for three or more cancers if we have such cases for each subpopulation.

Second, the occurrence of primary cancers may be dependent on other factors such as cancer treatment. For example, radiotherapy can damage normal cells in tumor adjacent area and is associated with excess incidence of solid cancers(89). Our model can include additional covariates, as we set for cancer history, to adjust for such dependency between successive events.

Third, because the correlation between first two gap times in the real data is very small, the recurrent events model we used in this study does not explicitly consider such association. For future datasets that do exhibit a stronger level of correlation between gap times, it would be expected the prediction performance of second or subsequent primary

cancers can be improved by properly utilizing such correlation information. We note that Bayesian parametric copula models have been developed for sequential gap time analyses(90). It will be interesting to incorporate such methods into our Bayesian framework to deal with missing genotype and ascertainment bias for a more flexible and accurate penetrance estimation. However, this is beyond the scope of this study.

Finally, in MPC studies, there usually exist multiple types of cancers. For example, the LFS is characterized by several cancer types such as sarcoma, breast cancer and lung cancer. MPC patients are then under the competing risk of multiple type of cancers. In our current model, we assume all cancers are of the same type and do not take into account of this nontrivial competing risk. Future work may focus on extending our methodology to provide a MPC-specific and cancer-specific penetrance estimation.



## 4. Conclusions and Future Research

### 4.1 Conclusions

In this dissertation, we developed statistics and bioinformatics methods to specifically solve two important problems in cancer research. The first problem is on assessing the accuracy of using FFPE RNA-seq data for gene expression profiling. To this end, we designed a FFPE breast tumor biopsies study, with matched high-quality FF samples as the reference standards for comparison. We devised multiple computational evaluation criteria, which cover almost all major parameters relevant to the discovery and translational application of mRNA expression biomarkers and take into account of the variation of analytical factors, to extensively investigate the concordance between FFPE and FF RNA-seq data, as well as the effects of pre-analytical factors of RNA-seq on such concordance (Figure 1.1). We found in this study that capture sequencing, rather than FFPE conditioning, is the dominant determinant for the variation of RNA-seq data. We also successfully identified one FFPE library preparation protocol that can generate RNA-seq data consistently highly concordant with and being least deviated from any types of non-capture sequenced FF references. The computational methods we applied in this study will be useful for other comparative analysis aiming to study the influences of pre-analytical factors of RNA-seq on mRNA expression data quality.

In the second project, we were challenged by estimating second primary cancer-specific penetrance of germline *TP53* mutation from individuals with missing genotype information. Justified by careful preliminary analysis of the real data, we proposed a Bayesian semiparametric recurrent events model based on NHPP in order to reflect the age-dependent and time-varying nature of the cancer occurrence rate in LFS study.

Following the idea of Shin *et. al*(80), we defined the familywise likelihood by averaging individual likelihoods within the family over the missing genotypes. This is possible since the exact distribution of missing genotypes is available according to the Mendelian law of inheritance. The familywise likelihood can minimize the efficiency loss due to the missing genetic information by utilizing family structure. We therefore developed the ascertainment corrected familywise likelihood for the proposed NHPP model and estimated the penetrance parameters via the MCMC algorithm. The MPC-specific penetrance we provided here for LFS study is stratified not only by genotype and gender, but also by the interaction of previous primary cancer diagnosis and genotype, as well as the age of first primary cancer diagnosis. Our penetrance estimates have a reliable cancer risk prediction performance on an independent dataset when comparing to that of the penetrance estimated by KM method. To the best of our knowledge, this is the first time that a high-resolution MPC-specific penetrance is reported and its cancer risk prediction performance is thoroughly evaluated.

## 4.2 Future Research

My dissertation is still left with a few possible extensions. For the FFPE RNA-seq project, we focused our comparative analysis on mRNA expression profiling. In future research, we may consider extending the analysis to other RNA species, like long non-coding RNA and microRNA, as they play important regulatory role in human cancer and have potential utility as cancer prognosis biomarkers(91-96). Also, in future the analysis should cover other aspects of detecting aberrant transcription in human cancer, such as gene fusion and alternative splicing analysis, as increasing evidences suggest their utility in cancer diagnostics and prognosis(10). For example, the detection of *RUNXI-RUNXITI* fusion has been suggested by World Health Organization as an alternative diagnostic method acute myeloid leukemia(97), and the *TMPRSS2-ERG* fusion has been shown associated with prostate cancer prognosis(98). Future work should be focused on incorporating these aspects of analysis to achieve a more comprehensive assessment of FFPE RNA-seq data quality.

For the multiple primary cancers penetrance estimation, we limited our analysis up to the second cancer diagnosis, as there are no mutation non-carriers with more than two primary cancers during follow-up in the training data. Our model is flexible to generate penetrance for third or fourth primary cancer if in future we have sufficient number of individuals with a history of two or more cancers. Also, we can easily modify our model to incorporate the treatment information to account the effects of radiotherapy or tissue-resection on future cancer occurrence rate. Finally, our model does not consider competing risk from different types of primary cancers, but in LFS MPC patients do

usually exhibit multiple types of cancers, such as sarcoma, leukemia and brain tumor.

Future research should extend the current model to take into account this competing risk.

## Appendix

### Appendix A: Computation of IPCW Kendall's $\tau$

Let  $(X_1, Y_1)$  and  $(X_2, Y_2)$  be two independent realizations of  $(X, Y)$ , the first and the second gap time, and let  $\psi_{12} = I\{(X_1 - X_2)(Y_1 - Y_2) > 0\} - I\{(X_1 - X_2)(Y_1 - Y_2) < 0\}$  indicate the concordance/discordant status of the pair, the Kendall's  $\tau$  (99) can be estimated from uncensored bivariate data  $\{(X_i, Y_i), i = 1, \dots, n\}$  by

$$\binom{n}{2}^{-1} \sum_{i < j} \psi_{ij}$$

In the presence of censoring events  $(D_X, D_Y)$  related to the two gap times, respectively, the estimation of  $\tau$  can only be based on orderable pairs. Let one observation be denoted as  $(\tilde{X}, \tilde{Y}, \delta_X, \delta_Y)$ , where  $\tilde{X} = \min(X, D_X)$ ,  $\tilde{Y} = \min(Y, D_Y)$ ,  $\delta_X = I(X < D_X)$  and  $\delta_Y = I(Y < D_Y)$ . Oakes *et.al* showed that the pair  $(i, j)$  is orderable if  $\{\tilde{X}_{ij} < \tilde{D}_{X_{ij}}, \tilde{Y}_{ij} < \tilde{D}_{Y_{ij}}\}$ , where  $\tilde{X}_{ij} = \min(X_i, X_j)$ ,  $\tilde{Y}_{ij} = \min(Y_i, Y_j)$ ,  $\tilde{D}_{X_{ij}} = \min(D_{X_i}, D_{X_j})$ ,  $\tilde{D}_{Y_{ij}} = \min(D_{Y_i}, D_{Y_j})$  (100). Let  $L_{ij}$  be the indicator of this event, and  $\hat{p}_{ij}$  be an estimator of the probability of being orderable  $p_{ij} = \Pr(D_X > \tilde{X}_{ij}; D_Y > \tilde{Y}_{ij} | \tilde{X}_{ij}, \tilde{Y}_{ij})$ , Lakhal-Chaieb *et.al* proposed the weighted estimate as

$$\hat{\tau}_m = \left( \sum_{i < j} \frac{L_{ij}}{\hat{p}_{ij}} \right)^{-1} \sum_{i < j} \frac{L_{ij} \psi_{ij}}{\hat{p}_{ij}}$$

To identify orderable pairs and estimate the corresponding  $p_{ij}$ , Lakhal-Chaieb *et.al* showed that  $L_{ij}$  can be reduced to that  $X_i$  and  $X_j$  are uncensored,  $\tilde{Y}_{ij}$  is observed, and that  $\{D_{X_i} > X_i + \tilde{Y}_{ij}; D_{X_j} > X_j + \tilde{Y}_{ij}\}$ . The conditional probability of a pair being orderable is then

$$\begin{aligned}
p_{ij} &= \Pr \left\{ D_{X_i} > X_i + \tilde{Y}_{ij}, D_{X_j} > X_j + \tilde{Y}_{ij} \mid X_i, X_j, \tilde{Y}_{ij} \right\} \\
&= G(X_i + \tilde{Y}_{ij}) \times G(X_j + \tilde{Y}_{ij})
\end{aligned}$$

The probability is estimated by

$$\hat{p}_{ij} = \hat{G}(X_i + \tilde{Y}_{ij}) \times \hat{G}(X_j + \tilde{Y}_{ij})$$

where  $\hat{G}(\cdot)$  is the Kaplan-Meier estimator of  $G(\cdot)$  based on  $\{(\tilde{X}_k + \tilde{Y}_k, 1 - \delta_{Y_k}), k = 1, \dots, n\}$ . The standard error of the kendall's  $\tau$  is estimated by the Jackknife.

## Appendix B: An example of using peeling algorithm to calculate familywise likelihood

Figure 3.7 shows an example of a hypothetical family with three generations. Assume that the genotype is known for the 4th individual and unknown for all other members, or  $\mathbf{g}_{mis}^T = (g_1, g_2, g_3, g_5, g_6, g_7)$ . Let  $\mathbf{h}^T = (h_1, \dots, h_7)$  denotes the cancer phenotype for the family, we want to calculate the familywise likelihood by marginalizing out  $\mathbf{g}_{mis}$ ,

$$L(\mathbf{h}|\mathbf{g}_4) = \sum_{\mathbf{g}_{mis}} L(\mathbf{h}, \mathbf{g}_{mis}|\mathbf{g}_4) = \sum_{\mathbf{g}_{mis}} p(\mathbf{h}|\mathbf{g}_{mis}, \mathbf{g}_4)p(\mathbf{g}_{mis}|\mathbf{g}_4)$$

The Elston-Stewart algorithm exploits the family structure by Mendelian inheritance property and introduces a “peeling” method, which rewrites the likelihood as,

$$\begin{aligned} &= p(h_4|\mathbf{g}_4) \\ &\left[ \sum_{g_1} p(h_1|g_1)p(g_1|\mathbf{g}_4) \left\{ \sum_{g_2} p(h_2|g_2)p(g_2|g_1, \mathbf{g}_4) \left( \sum_{g_3} p(h_3|g_3)p(g_3|g_1, g_2) \right) \right\} \right] \\ &\left[ \sum_{g_5} p(h_5|g_5)p(g_5|\mathbf{g}_4) \left\{ \sum_{g_6} p(h_6|g_6)p(g_6|g_4, g_5) \left( \sum_{g_7} p(h_7|g_7)p(g_7|g_4, g_5) \right) \right\} \right] \end{aligned}$$

Note that in our example, the partial likelihood of upper part of the family (anterior) can be evaluated separately from that of the lower part of the family (posterior) given the genotype of the 4th individual (pivot element), or Anterior  $\perp$  Posterior | Pivot element.

Also, within a nuclear family the likelihood for some members can be evaluated separately (e.g.: the 6th and 7th individual). This is based on the Mendelian inheritance property that a child’s genotype only depends on his parents’ genotypes. The computation complexity is then reduced by the algorithm from  $O(3^n)$  to  $O(n \log(n))$  if *TP53* has three genotypes.

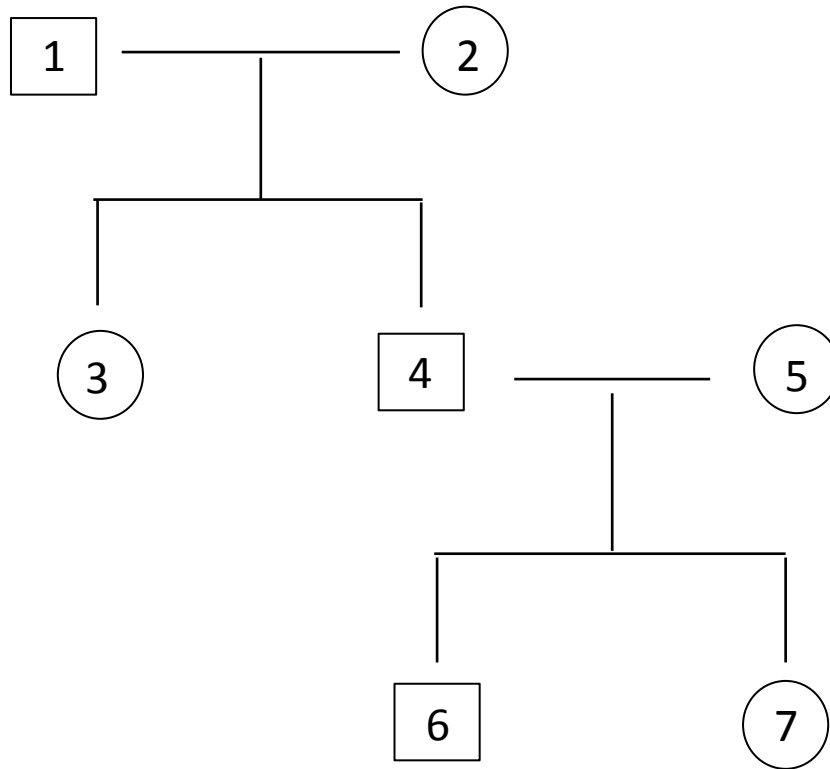


Figure Appendix B: A hypothetical pedigree for illustrating likelihood calculation using the Elston- Stewart algorithm. The family consists of three generations. The circle indicates the female member while the square indicates the male. In this example, the genotype is assumed unknown for every members except the 4th individual.



## Bibliography

1. Hatzis, C., L. Pusztai, V. Valero, D. J. Booser, L. Esserman, A. Lluch, T. Vidaurre, F. Holmes, E. Souchon, H. Wang, M. Martin, J. Cotrina, H. Gomez, R. Hubbard, J. I. Chacon, J. Ferrer-Lozano, R. Dyer, M. Buxton, Y. Gong, Y. Wu, N. Ibrahim, E. Andreopoulou, N. T. Ueno, K. Hunt, W. Yang, A. Nazario, A. DeMichele, J. O'Shaughnessy, G. N. Hortobagyi, and W. F. Symmans. 2011. A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. *Jama* 305: 1873-1881.
2. Mortazavi, A., B. A. Williams, K. McCue, L. Schaeffer, and B. Wold. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods* 5: 621-628.
3. Cancer Genome Atlas Network. Electronic address, i. m. o., and N. Cancer Genome Atlas. 2015. Genomic Classification of Cutaneous Melanoma. *Cell* 161: 1681-1696.
4. Cancer Genome Atlas, N. 2012. Comprehensive molecular portraits of human breast tumours. *Nature* 490: 61-70.
5. Hansen, K. D., Z. J. Wu, R. A. Irizarry, and J. T. Leek. 2011. Sequencing technology does not eliminate biological variability. *Nature biotechnology* 29: 572-573.
6. t Hoen, P. A., M. R. Friedlander, J. Almlof, M. Sammeth, I. Pulyakhina, S. Y. Anvar, J. F. Laros, H. P. Buermans, O. Karlberg, M. Brannvall, G. Consortium, J. T. den Dunnen, G. J. van Ommen, I. G. Gut, R. Guigo, X. Estivill, A. C. Syvanen, E. T. Dermitzakis, and T. Lappalainen. 2013. Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. *Nature biotechnology* 31: 1015-1022.
7. Marioni, J. C., C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad. 2008. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research* 18: 1509-1517.
8. Fumagalli, D., A. Blanchet-Cohen, D. Brown, C. Desmedt, D. Gacquer, S. Michiels, F. Rothe, S. Majjaj, R. Salgado, D. Larsimont, M. Ignatiadis, M. Maetens, M. Piccart, V. Detours, C. Sotiriou, and B. Haibe-Kains. 2014. Transfer

- of clinically relevant gene expression signatures in breast cancer: from Affymetrix microarray to Illumina RNA-Sequencing technology. *BMC genomics* 15: 1008.
9. Zhang, W., Y. Yu, F. Hertwig, J. Thierry-Mieg, W. Zhang, D. Thierry-Mieg, J. Wang, C. Furlanello, V. Devanarayan, J. Cheng, Y. Deng, B. Hero, H. Hong, M. Jia, L. Li, S. M. Lin, Y. Nikolsky, A. Oberthuer, T. Qing, Z. Su, R. Volland, C. Wang, M. D. Wang, J. Ai, D. Albanese, S. Asgharzadeh, S. Avigad, W. Bao, M. Bessarabova, M. H. Brilliant, B. Brors, M. Chierici, T. M. Chu, J. Zhang, R. G. Grundy, M. M. He, S. Hebring, H. L. Kaufman, S. Lababidi, L. J. Lancashire, Y. Li, X. X. Lu, H. Luo, X. Ma, B. Ning, R. Noguera, M. Peifer, J. H. Phan, F. Roels, C. Rosswog, S. Shao, J. Shen, J. Theissen, G. P. Tonini, J. Vandesompele, P. Y. Wu, W. Xiao, J. Xu, W. Xu, J. Xuan, Y. Yang, Z. Ye, Z. Dong, K. K. Zhang, Y. Yin, C. Zhao, Y. Zheng, R. D. Wolfinger, T. Shi, L. H. Malkas, F. Berthold, J. Wang, W. Tong, L. Shi, Z. Peng, and M. Fischer. 2015. Comparison of RNA-seq and microarray-based models for clinical endpoint prediction. *Genome biology* 16: 133.
  10. Byron, S. A., K. R. Van Keuren-Jensen, D. M. Engelthaler, J. D. Carpten, and D. W. Craig. 2016. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. *Nat Rev Genet* 17: 257-271.
  11. Conesa, A., P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szczesniak, D. J. Gaffney, L. L. Elo, X. Zhang, and A. Mortazavi. 2016. A survey of best practices for RNA-seq data analysis. *Genome biology* 17: 13.
  12. Trapnell, C., A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn, and L. Pachter. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols* 7: 562-578.
  13. Anders, S., D. J. McCarthy, Y. Chen, M. Okoniewski, G. K. Smyth, W. Huber, and M. D. Robinson. 2013. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nature protocols* 8: 1765-1786.

14. Kim, D., G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S. L. Salzberg. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology* 14: R36.
15. Li, B., and C. N. Dewey. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics* 12: 323.
16. DeLuca, D. S., J. Z. Levin, A. Sivachenko, T. Fennell, M. D. Nazaire, C. Williams, M. Reich, W. Winckler, and G. Getz. 2012. RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics* 28: 1530-1532.
17. Wang, L., S. Wang, and W. Li. 2012. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* 28: 2184-2185.
18. Li, B., V. Ruotti, R. M. Stewart, J. A. Thomson, and C. N. Dewey. 2010. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics* 26: 493-500.
19. Wagner, G. P., K. Kin, and V. J. Lynch. 2012. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory in biosciences = Theorie in den Biowissenschaften* 131: 281-285.
20. Scicchitano, M. S., D. A. Dalmas, R. W. Boyce, H. C. Thomas, and K. S. Frazier. 2009. Protein extraction of formalin-fixed, paraffin-embedded tissue enables robust proteomic profiles by mass spectrometry. *The journal of histochemistry and cytochemistry : official journal of the Histochemistry Society* 57: 849-860.
21. Dedhia, P., S. Tarale, G. Dhongde, R. Khadapkar, and B. Das. 2007. Evaluation of DNA extraction methods and real time PCR optimization on formalin-fixed paraffin-embedded tissues. *Asian Pacific journal of cancer prevention : APJCP* 8: 55-59.
22. von Ahlfen, S., A. Missel, K. Bendrat, and M. Schlumpberger. 2007. Determinants of RNA quality from FFPE samples. *PloS one* 2: e1261.
23. Masuda, N., T. Ohnishi, S. Kawamoto, M. Monden, and K. Okubo. 1999. Analysis of chemical modification of RNA from formalin-fixed samples and optimization of molecular biology applications for such samples. *Nucleic acids research* 27: 4436-4443.

24. Sinicropi, D., K. Qu, F. Collin, M. Crager, M. L. Liu, R. J. Pelham, M. Pho, A. Dei Rossi, J. Jeong, A. Scott, R. Ambannavar, C. Zheng, R. Mena, J. Esteban, J. Stephans, J. Morlan, and J. Baker. 2012. Whole transcriptome RNA-Seq analysis of breast cancer recurrence risk using formalin-fixed paraffin-embedded tumor tissue. *PloS one* 7: e40092.
25. Graw, S., R. Meier, K. Minn, C. Bloomer, A. K. Godwin, B. Fridley, A. Vlad, P. Beyerlein, and J. Chien. 2015. Robust gene expression and mutation analyses of RNA-sequencing of formalin-fixed diagnostic tumor samples. *Scientific reports* 5: 12335.
26. Adiconis, X., D. Borges-Rivera, R. Satija, D. S. DeLuca, M. A. Busby, A. M. Berlin, A. Sivachenko, D. A. Thompson, A. Wysoker, T. Fennell, A. Gnirke, N. Pochet, A. Regev, and J. Z. Levin. 2013. Comparative analysis of RNA sequencing methods for degraded or low-input samples. *Nature methods* 10: 623-629.
27. Li, P., A. Conley, H. Zhang, and H. L. Kim. 2014. Whole-Transcriptome profiling of formalin-fixed, paraffin-embedded renal cell carcinoma by RNA-seq. *BMC genomics* 15: 1087.
28. Sharron Lin, X., L. Hu, K. Sandy, M. Correll, J. Quackenbush, C. L. Wu, and W. Scott McDougal. 2014. Differentiating progressive from nonprogressive T1 bladder cancer by gene expression profiling: applying RNA-sequencing analysis on archived specimens. *Urologic oncology* 32: 327-336.
29. Li, F. P., and J. F. Fraumeni, Jr. 1969. Soft-tissue sarcomas, breast cancer, and other neoplasms. A familial syndrome? *Ann Intern Med* 71: 747-752.
30. Hisada, M., J. E. Garber, C. Y. Fung, J. F. Fraumeni, Jr., and F. P. Li. 1998. Multiple primary cancers in families with Li-Fraumeni syndrome. *J Natl Cancer Inst* 90: 606-611.
31. Lustbader, E. D., W. R. Williams, M. L. Bondy, S. Strom, and L. C. Strong. 1992. Segregation analysis of cancer in families of childhood soft-tissue-sarcoma patients. *American journal of human genetics* 51: 344-356.
32. Varley, J. M. 2003. Germline TP53 mutations and Li-Fraumeni syndrome. *Human mutation* 21: 313-320.

33. Hwang, S. J., G. Lozano, C. I. Amos, and L. C. Strong. 2003. Germline p53 mutations in a cohort with childhood sarcoma: sex differences in cancer risk. *American journal of human genetics* 72: 975-983.
34. Wu, C. C., L. C. Strong, and S. Shete. 2010. Effects of measured susceptibility genes on cancer risk in family studies. *Hum Genet* 127: 349-357.
35. Strong, L. C., M. Stine, and T. L. Norsted. 1987. Cancer in Survivors of Childhood Soft-Tissue Sarcoma and Their Relatives. *J Natl Cancer I* 79: 1213-1220.
36. Hayat, M. J., N. Howlader, M. E. Reichman, and B. K. Edwards. 2007. Cancer statistics, trends, and multiple primary cancer analyses from the Surveillance, Epidemiology, and End Results (SEER) Program. *The oncologist* 12: 20-37.
37. Curtis, R. E. 2006. *New malignancies among cancer survivors : SEER cancer registries, 1973-2000*. U.S. Dept. of Health and Human Services, National Institutes of Health, National Cancer Institute, Washington, D.C.
38. van Eggermond, A. M., M. Schaapveld, P. J. Lugtenburg, A. D. Krol, J. P. de Boer, J. M. Zijlstra, J. M. Raemaekers, L. C. Kremer, J. M. Roesink, M. W. Louwman, B. M. Aleman, and F. E. van Leeuwen. 2014. Risk of multiple primary malignancies following treatment of Hodgkin lymphoma. *Blood* 124: 319-327; quiz 466.
39. Malkin, D., F. P. Li, L. C. Strong, J. F. Fraumeni, Jr., C. E. Nelson, D. H. Kim, J. Kassel, M. A. Gryka, F. Z. Bischoff, M. A. Tainsky, and et al. 1990. Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. *Science* 250: 1233-1238.
40. Eeles, R. A. 1995. Germline mutations in the TP53 gene. *Cancer surveys* 25: 101-124.
41. Khoury, M. J., W. D. Flanders, and T. H. Beaty. 1988. Penetrance in the presence of genetic susceptibility to environmental factors. *American journal of medical genetics* 29: 397-403.
42. Domchek, S. M., A. Eisen, K. Calzone, J. Stopfer, A. Blackwood, and B. L. Weber. 2003. Application of breast cancer risk prediction models in clinical

- practice. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 21: 593-601.
43. Chen, S., and G. Parmigiani. 2007. Meta-analysis of BRCA1 and BRCA2 penetrance. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 25: 1329-1333.
  44. Wang, W. Y., K. B. Niendorf, D. Patel, A. Blackford, F. Marroni, A. J. Sober, G. Parmigiani, and H. Tsao. 2010. Estimating CDKN2A Carrier Probability and Personalizing Cancer Risk Assessments in Hereditary Melanoma Using MelaPRO. *Cancer Res* 70: 552-559.
  45. Cook, R. J., and J. F. Lawless. 2007. *The statistical analysis of recurrent events*. Springer, New York.
  46. Brown, L., N. Gans, A. Mandelbaum, A. Sakov, H. P. Shen, S. Zeltyn, and L. Zhao. 2005. Statistical analysis of a telephone call center: A queueing-science perspective. *J Am Stat Assoc* 100: 36-50.
  47. Weinberg, J., L. D. Brown, and J. R. Stroud. 2007. Bayesian forecasting of an inhomogeneous Poisson process with applications to call center data. *J Am Stat Assoc* 102: 1185-1198.
  48. Andersen, P. K., and R. D. Gill. 1982. Cox Regression-Model for Counting-Processes - a Large Sample Study. *Ann Stat* 10: 1100-1120.
  49. Courgeau, D. 1994. Statistical-Models Based on Counting-Processes - Andersen,Pk, Borgan,O, Gill,R, Keiding,N. *Eur J Popul* 10: 199-202.
  50. Elston, R. C., and J. Stewart. 1971. A general model for the genetic analysis of pedigree data. *Human heredity* 21: 523-542.
  51. Penland, S. K., T. O. Keku, C. Torrice, X. He, J. Krishnamurthy, K. A. Hoadley, J. T. Woosley, N. E. Thomas, C. M. Perou, R. S. Sandler, and N. E. Sharpless. 2007. RNA expression analysis of formalin-fixed paraffin-embedded tumors. *Laboratory investigation; a journal of technical methods and pathology* 87: 383-391.
  52. Zhao, W., X. He, K. A. Hoadley, J. S. Parker, D. N. Hayes, and C. M. Perou. 2014. Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. *BMC genomics* 15: 419.

53. Matranga, C. B., K. G. Andersen, S. Winnicki, M. Busby, A. D. Gladden, R. Tewhey, M. Stremlau, A. Berlin, S. K. Gire, E. England, L. M. Moses, T. S. Mikkelsen, I. Odia, P. E. Ehiane, O. Folarin, A. Goba, S. H. Kahn, D. S. Grant, A. Honko, L. Hensley, C. Happi, R. F. Garry, C. M. Malboeuf, B. W. Birren, A. Gnirke, J. Z. Levin, and P. C. Sabeti. 2014. Enhanced methods for unbiased deep sequencing of Lassa and Ebola RNA viruses from clinical and biological samples. *Genome biology* 15: 519.
54. Norton, N., Z. Sun, Y. W. Asmann, D. J. Serie, B. M. Necela, A. Bhagwate, J. Jen, B. W. Eckloff, K. R. Kalari, K. J. Thompson, J. M. Carr, J. M. Kachergus, X. J. Geiger, E. A. Perez, and E. A. Thompson. 2013. Gene expression, single nucleotide variant and fusion transcript discovery in archival material from breast tumors. *PloS one* 8: e81925.
55. <http://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/datasheet-truseq-rna-access.pdf>.
56. Method for optimized isolation of RNA from fixed tissue. WO 2009127350 A1. <http://www.google.com/patents/WO2009127350A1?cl=en>
57. Serena Bonin, Giorgio Stanta. RNA Temperature Demodification. Chapter of Guidelines for Molecular Analysis in Archive Tissues, pp 67-69 .
58. Li, H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25: 1754-1760.
59. Anders, S., P. T. Pyl, and W. Huber. 2015. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31: 166-169.
60. Yang, L., M. O. Duff, B. R. Graveley, G. G. Carmichael, and L. L. Chen. 2011. Genomewide characterization of non-polyadenylated RNAs. *Genome biology* 12: R16.
61. Anders, S., and W. Huber. 2010. Differential expression analysis for sequence count data. *Genome biology* 11: R106.
62. Suzuki, R., and H. Shimodaira. 2006. Pvcust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* 22: 1540-1542.

63. Bullard, J. H., E. Purdom, K. D. Hansen, and S. Dudoit. 2010. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC bioinformatics* 11: 94.
64. Robinson, M. D., and A. Oshlack. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome biology* 11: R25.
65. Robinson, M. D., D. J. McCarthy, and G. K. Smyth. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26: 139-140.
66. Zwiener, I., B. Frisch, and H. Binder. 2014. Transforming RNA-Seq data to improve the performance of prognostic gene signatures. *PloS one* 9: e85150.
67. Paik, S., S. Shak, G. Tang, C. Kim, J. Baker, M. Cronin, F. L. Baehner, M. G. Walker, D. Watson, T. Park, W. Hiller, E. R. Fisher, D. L. Wickerham, J. Bryant, and N. Wolmark. 2004. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *New Engl J Med* 351: 2817-2826.
68. Parker, J. S., M. Mullins, M. C. Cheang, S. Leung, D. Voduc, T. Vickery, S. Davies, C. Fauron, X. He, Z. Hu, J. F. Quackenbush, I. J. Stijleman, J. Palazzo, J. S. Marron, A. B. Nobel, E. Mardis, T. O. Nielsen, M. J. Ellis, C. M. Perou, and P. S. Bernard. 2009. Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 27: 1160-1167.
69. Loi, S., B. Haibe-Kains, S. Majjaj, F. Lallemand, V. Durbecq, D. Larsimont, A. M. Gonzalez-Angulo, L. Pusztai, W. F. Symmans, A. Bardelli, P. Ellis, A. N. J. Tutt, C. E. Gillett, B. T. Hennessy, G. B. Mills, W. A. Phillips, M. J. Piccart, T. P. Speed, G. A. McArthur, and C. Sotiriou. 2010. PIK3CA mutations associated with gene signature of low mTORC1 signaling and better outcomes in estrogen receptor-positive breast cancer. *P Natl Acad Sci USA* 107: 10208-10213.
70. Symmans, W. F., C. Hatzis, C. Sotiriou, F. Andre, F. Peintinger, P. Regitnig, G. Daxenbichler, C. Desmedt, J. Domont, C. Marth, S. Delaloge, T. Bauernhofer, V. Valero, D. J. Booser, G. N. Hortobagyi, and L. Pusztai. 2010. Genomic Index of Sensitivity to Endocrine Therapy for Breast Cancer. *Journal of Clinical Oncology* 28: 4111-4119.



71. van't Veer, L. J., H. Y. Dai, M. J. van de Vijver, Y. D. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend. 2002. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415: 530-536.
72. Kraft, P., and D. C. Thomas. 2000. Bias and efficiency in family-based gene-characterization studies: conditional, prospective, retrospective, and joint likelihoods. *American journal of human genetics* 66: 1119-1131.
73. Iversen, E. S., Jr., and S. Chen. 2005. Population-Calibrated Gene Characterization: Estimating Age at Onset Distributions Associated With Cancer Genes. *J Am Stat Assoc* 100: 399-409.
74. Peng, G., J. Bojadzieva, M. L. Ballinger, J. Li, A. L. Blackford, P. L. Mai, S. A. Savage, D. M. Thomas, L. C. Strong, and W. Wang. 2017. Estimating TP53 Mutation Carrier Probability in Families with Li-Fraumeni Syndrome Using LFSPRO. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*.
75. Lin, D. Y., W. Sun, and Z. L. Ying. 1999. Nonparametric estimation of the gap time distributions for serial events with censored data. *Biometrika* 86: 59-70.
76. Lakhal-Chaieb, L., R. J. Cook, and X. Lin. 2010. Inverse probability of censoring weighted estimates of Kendall's tau for gap time analyses. *Biometrics* 66: 1145-1152.
77. Ning, J., Y. Chen, C. Cai, X. Huang, and W. M. 2015. On the dependence structure of bivariate recurrent event processes: inference and estimation. *Biometrika* 102 (2): 345-358.
78. Chang, I., C. A. Hsiung, Y. J. Wu, and C. C. Yang. 2005. Bayesian survival analysis using Bernstein polynomials. *Scandinavian Journal of Statistics* Vol. 32, No. 3 (Sep., 2005), pp. 447-466.
79. Lorentz, G. G. 2012. *Bernstein polynomials*. American Mathematical Soc.

80. Shin, S. J., L. C. Strong, J. Bojadzieva, W. Wang, and Y. Yuan. 2017. Bayesian Semiparametric Estimation of Cancer-specific Age-at-onset Penetrance with Application to Li-Fraumeni Syndrome. *arXiv:1701.01558*.
81. Curtis, S. M., and S. K. Ghosh. 2011. A variable selection approach to monotonic regression with Bernstein polynomials. *Journal of Applied Statistics* vol. 38: pages 961-976.
82. Gelfand, A. E., and B. K. Mallick. 1995. Bayesian analysis of proportional hazards models built from monotone functions. *Biometrics* 51: 843-852.
83. Lange, K., and R. C. Elston. 1975. Extensions to pedigree analysis I. Likelihood calculations for simple and complex pedigrees. *Human heredity* 25: 95-105.
84. Fernando, R. L., C. Stricker, and R. C. Elston. 1993. An efficient algorithm to compute the posterior genotypic distribution for every member of a pedigree without loops. *Theor Appl Genet* 87: 89-93.
85. Wang, W., S. Chen, K. A. Brune, R. H. Hruban, G. Parmigiani, and A. P. Klein. 2007. PancPRO: risk assessment for individuals with a family history of pancreatic cancer. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 25: 1417-1422.
86. Efron, B., and C. Stein. 1981. The Jackknife Estimate of Variance. *Ann Stat* 9: 586-596.
87. Low, L. Y. 1983. The Jackknife, the Bootstrap and Other Resampling Plans - Efron, B. *Journal of the American Statistical Association* 78: 987-987.
88. SEER Program (National Cancer Institute (U.S.)), National Center for Health Statistics (U.S.), National Cancer Institute (U.S.). Surveillance Program., National Cancer Institute (U.S.). Cancer Statistics Branch., and National Cancer Institute (U.S.). Cancer Control Research Program. 2007. SEER cancer statistics review. In *NIH publication*. U.S. Dept. of Health and Human Services, Public Health Service, National Institutes of Health, National Cancer Institute, Bethesda, Md. volumes.
89. Inskip, P. D., and R. E. Curtis. 2007. New malignancies following childhood cancer in the United States, 1973-2002. *Int J Cancer* 121: 2233-2240.
90. Meyer, R., and J. S. Romeo. 2015. Bayesian semiparametric analysis of recurrent failure time data using copulas. *Biom J* 57: 982-1001.

91. Meiri, E., W. C. Mueller, S. Rosenwald, M. Zepeniuk, E. Klinke, T. B. Edmonston, M. Werner, U. Lass, I. Barshack, M. Feinmesser, M. Huszar, F. Fogt, K. Ashkenazi, M. Sanden, E. Goren, N. Dromi, O. Zion, I. Burnstein, A. Chajut, Y. Spector, and R. Aharonov. 2012. A second-generation microRNA-based assay for diagnosing tumor tissue origin. *The oncologist* 17: 801-812.
92. Taubert, H., T. Greither, D. Kaushal, P. Wurl, M. Bache, F. Bartel, A. Kehlen, C. Lautenschlager, L. Harris, K. Kraemer, A. Meye, M. Kappler, H. Schmidt, H. J. Holzhausen, and S. Hauptmann. 2007. Expression of the stem cell self-renewal gene Hiwi and risk of tumour-related death in patients with soft-tissue sarcoma. *Oncogene* 26: 1098-1100.
93. Baraniskin, A., S. Nopel-Dunnebacke, M. Ahrens, S. G. Jensen, H. Zollner, A. Maghnouj, A. Wos, J. Mayerle, J. Munding, D. Kost, A. Reinacher-Schick, S. Liffers, R. Schroers, A. M. Chromik, H. E. Meyer, W. Uhl, S. Klein-Scory, F. U. Weiss, C. Stephan, I. Schwarte-Waldhoff, M. M. Lerch, A. Tannapfel, W. Schmiegel, C. L. Andersen, and S. A. Hahn. 2013. Circulating U2 small nuclear RNA fragments as a novel diagnostic biomarker for pancreatic and colorectal adenocarcinoma. *International Journal of Cancer* 132: E48-E57.
94. Su, J., J. Liao, L. Gao, J. Shen, M. A. Guarnera, M. Zhan, H. Fang, S. A. Stass, and F. Jiang. 2016. Analysis of small nucleolar RNAs in sputum for lung cancer diagnosis. *Oncotarget* 7: 5131-5142.
95. Du, Z., T. Fei, R. G. Verhaak, Z. Su, Y. Zhang, M. Brown, Y. Chen, and X. S. Liu. 2013. Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer. *Nat Struct Mol Biol* 20: 908-913.
96. Guo, Y., A. Bosompem, S. Mohan, B. Erdogan, F. Ye, K. C. Vickers, Q. Sheng, S. Zhao, C. I. Li, P. F. Su, M. Jagasia, S. A. Strickland, E. A. Griffiths, and A. S. Kim. 2015. Transfer RNA detection by small RNA deep sequencing and disease association with myelodysplastic syndromes. *BMC genomics* 16: 727.
97. Vardiman, J. W., J. Thiele, D. A. Arber, R. D. Brunning, M. J. Borowitz, A. Porwit, N. L. Harris, M. M. Le Beau, E. Hellstrom-Lindberg, A. Tefferi, and C. D. Bloomfield. 2009. The 2008 revision of the World Health Organization (WHO)

- classification of myeloid neoplasms and acute leukemia: rationale and important changes. *Blood* 114: 937-951.
98. Font-Tello, A., N. Juanpere, S. de Muga, M. Lorenzo, J. A. Lorente, L. Fumado, L. Serrano, S. Serrano, J. Lloreta, and S. Hernandez. 2015. Association of ERG and TMPRSS2-ERG with grade, stage, and prognosis of prostate cancer is dependent on their expression levels. *Prostate* 75: 1216-1226.
99. Kendall, M. G., and J. D. Gibbons. 1990. *Rank correlation methods*. E. Arnold ; Oxford University Press, London  
New York, NY.
100. Oakes, D. 1982. A concordance test for independence in the presence of censoring. *Biometrics* 38: 451-455.

## **VITA**

Jialu Li, the son of Bangming Li and Youling Huang, was born in 1988 in LuAn, Anhui, China. In 2005, he entered Anhui Normal University, China and obtained his Bachelor of Science degree in biological sciences in 2009. In the same year, he was enrolled at Chinese Academy of Sciences and received his Master of Sciences in biochemistry and molecular biology in 2012. Three months after graduation, he left for USA to pursue his Ph.D degree in biostatistics and bioinformatics at the University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences. In May 2013, he joined Dr. Wenyi Wang's group to prepare for his Ph.D dissertation in the University of Texas MD Anderson Cancer Center. He is expected to obtain his Doctor of Philosophy degree in biostatistics and bioinformatics in May 2017.