META-analysis of microarray data to assess gender biased differential gene expression in hepatic tissue

Amira Salah eldin Mahmoud ElBakry

Follow this and additional works at: https://fount.aucegypt.edu/etds

Recommended Citation

APA Citation

MLA Citation

This Dissertation is brought to you for free and open access by AUC Knowledge Fountain. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of AUC Knowledge Fountain. For more information, please contact mark.muehlhaeusler@aucegypt.edu.
META-ANALYSIS OF MICROARRAY DATA TO ASSESS GENDER BIASED DIFFERENTIAL GENE EXPRESSION IN HEPATIC TISSUE

A Thesis Submitted to
The Biotechnology Graduate Program

In partial fulfillment for the pre-requisite requirement for full admission to the Ph.D Program in Applied Sciences

By: Amira ElBakry

Under the Supervision of: Dr. Rania Siam
Chair of the Biology Department

May/2015
ACKNOWLEDGMENTS

My unreserved gratitude goes to Dr. Rania Siam, for her unconditional support, invaluable help and constant encouragement. Thank you for making this thesis possible, and for pushing everything forward for me and with me. Working with you is a pleasure and a privilege.

Additionally, Mustafa Adel has provided much-needed help in many aspects of this thesis. Thank you for taking the time, putting the effort, and always replying to my emails and answering my many questions.

Appreciation is directed to Dr. Ahmed Moustafa, for supporting my idea and his invaluable coaching in bioinformatics. I learned an incredible amount in a very short time.

Special thanks to Dr. Sherif El-Khamisy for inspiring this thesis, it was his idea I built most of the foundation for the work on.

Also, acknowledgment is due for Ali El Behery, for helping me multiple times with my hopeless programming and resolving every single error. Also, my appreciation for Yasmeen El Howeedy, for taking time to share her experience and providing valuable help.

Thank you for everyone in the biology department that made my day easier, happier or just listened to my complaining.

Finally, my eternal gratitude goes for my family, for their endless support every step of the way.
ABSTRACT

Hepatocellular carcinoma (HCC) is the second deadliest cancer globally, and with an estimated 782,000 new cases in 2012, it is the fifth most common cancer in men and ninth in women. HCC is of particular concern in Egypt because of the high prevalence of Hepatitis C Virus (HCV). Due to its poor prognosis, HCC is the leading cause of cancer-related deaths in Egypt. A gender disparity is observed in liver cancer cases, with higher prevalence in men by three to five fold. This sex bias is even more pronounced in mouse models of HCC, which was found to be sex hormone-dependent. Some studies have attempted to elucidate the molecular mechanisms of this disparity; but with inconclusive and sometimes contradicting outcomes, they remain largely unresolved. Understanding the natural protective mechanisms in females would allow for the development of preventative and therapeutic strategies for patients at risk for HCC or already inflicted with the disease. In this study, we applied a meta-analysis approach on already available microarray data from human normal liver tissues to identify differentially expressed genes between males and females. Microarray datasets were downloaded from the Gene Expression Omnibus database, Robust Multiarray Average pre-processed and analyzed for differential expression. The combination of 2 distinct datasets and analysis using a p-value cut-off of 0.05 and fold change cut-off of 2 revealed male up-regulated genes including RPS4Y1, EIF1AY, CYorf15B, UTY, DDX3Y and USP9Y. Female up-regulated genes included XIST, PNPLA4 and PZP. Our results confirm gender-specific differential expression patterns found in other tissues and call for further investigation using a larger sample size and more sensitive approaches such as RNA-Sequencing and, targeted protein-level studies.
TABLE OF CONTENTS

LIST OF TABLES............................................................................................................................................. v
LIST OF FIGURES................................................................................................................................................ vi
ABBREVIATIONS .................................................................................................................................................... vii
1. Introduction ...................................................................................................................................................... 1
   1.1 Gender Bias in Hepatocellular Carcinoma ................................................................................................. 1
   1.2 Molecular Mechanisms of Gender Bias in HCC......................................................................................... 2
   1.3 Studying Gene Expression using Microarrays ......................................................................................... 4
   1.4 Microarray Data pre-processing .............................................................................................................. 6
   1.5 Differential Expression Analysis ............................................................................................................ 8
   1.6 Accounting for Batch Effects ................................................................................................................ 10
2. Materials and Methods .................................................................................................................................. 12
   2.1 Data Collection and Processing ............................................................................................................. 12
   2.2 Data Exploration and Differential Expression Analysis ...................................................................... 12
   2.3 Dataset Merging and Batch Effect Removal ...................................................................................... 13
   2.4 Gene ID conversion and functional annotation .................................................................................. 14
3. Results ............................................................................................................................................................ 14
   3.1 Dataset Collection and Processing ....................................................................................................... 14
   3.2 Individual Dataset Analysis Using the T-test Method ........................................................................ 15
   3.3 Individual Dataset Analysis Using the Limma Package and Bayesian statistics .................................. 16
   3.4 Analysis of Merged Datasets Using the Student’s T-test ................................................................. 16
   3.5 Batch Effects Removal ........................................................................................................................ 17
   3.6 Gene Signature Validation .................................................................................................................... 18
4. Discussion ....................................................................................................................................................... 18
TABLES............................................................................................................................................................... 26
FIGURES............................................................................................................................................................. 35
REFERENCES..................................................................................................................................................... 43
LIST OF TABLES

Table 1. Summary of the datasets of microarray studies using human normal liver tissue .......... 26
Table 2. Sample information of three microarray datasets with gender information .................. 27
Table 3. Summary of differentially expressed probes found in all three data set and their corresponding gene names .......................................................................................................................... 28
Table 4. Summary of differentially expressed probes identified in GSE14323 using the Limma Package ............................................................................................................................................. 29
Table 5. Differentially expressed probes identified in GSE 23343 using the Limma Package .... 30
Table 6. Differentially expressed probes and their corresponding gene names, identified in merged datasets using the t-test method ............................................................................................................ 31
Table 7. Differentially expressed probes and their corresponding genes in dataset GSE14343 after batch effect removal using fRMA ................................................................. 32
Table 8. Differentially expressed probes and their corresponding genes in dataset GSE23343 after batch effect removal using fRMA ................................................................. 33
Table 9. Differentially expressed probes and their corresponding genes in merged datasets after batch effect removal using fRMA or using ComBat ...................................................... 34
LIST OF FIGURES

Figure 1: Hierarchical Cluster Analysis of 3 microarray datasets………………………………35
Figure 2: Differentially Expressed probes in individual datasets (t-test)………………………….36
Figure 3: Differentially Expressed probes in individual datasets (Limma)…………………………37
Figure 4: Differentially Expressed probes in merged datasets……………………………………38
Figure 5: Hierarchical Cluster Analysis of merged microarray datasets…………………………39
Figure 6: Differentially Expressed probes in individual datasets after batch effect removal……40
Figure 7: Differentially Expressed probes in merged datasets after batch effect removal………41
Figure 8: Gene Signature Validation………………………………………………………………..42
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANOVA</td>
<td>Analysis of variance</td>
</tr>
<tr>
<td>ComBat</td>
<td>Combining Batches of Gene Expression Microarray Data</td>
</tr>
<tr>
<td>DAVID</td>
<td>Database for Annotation, Visualization and Integrated Discovery</td>
</tr>
<tr>
<td>DDX3Y</td>
<td>DEAD (Asp-Glu-Ala-Asp) box polypeptide 3Y</td>
</tr>
<tr>
<td>DEN</td>
<td>Diethylnitrosamine</td>
</tr>
<tr>
<td>DWD</td>
<td>Distance-weighted discrimination</td>
</tr>
<tr>
<td>EIF1AY</td>
<td>Eukaryotic translation initiation factor 1A-Y</td>
</tr>
<tr>
<td>FC</td>
<td>Fold Change</td>
</tr>
<tr>
<td>fRMA</td>
<td>Frozen robust multichip average</td>
</tr>
<tr>
<td>GEO</td>
<td>Gene Expression Omnibus</td>
</tr>
<tr>
<td>HBV</td>
<td>Hepatitis B Virus</td>
</tr>
<tr>
<td>HCC</td>
<td>Hepatocellular carcinoma</td>
</tr>
<tr>
<td>HCV</td>
<td>Hepatitis C Virus</td>
</tr>
<tr>
<td>IL-6</td>
<td>Interleukin 6</td>
</tr>
<tr>
<td>KDM5D</td>
<td>(K)-specific demethylase 6A</td>
</tr>
<tr>
<td>LARP4B</td>
<td>La ribonucleoprotein domain family, member 4B</td>
</tr>
<tr>
<td>MM</td>
<td>Mismatch</td>
</tr>
<tr>
<td>PM</td>
<td>Perfect Match</td>
</tr>
<tr>
<td>PNPLA4</td>
<td>Patatin-like phospholipase domain containing 4</td>
</tr>
<tr>
<td>PRL</td>
<td>Prolactin</td>
</tr>
<tr>
<td>PRLR</td>
<td>Prolactin Receptor</td>
</tr>
<tr>
<td>PZP</td>
<td>Pregnancy Zone Protein</td>
</tr>
<tr>
<td>RMA</td>
<td>Robust Multichip Average</td>
</tr>
<tr>
<td>RNA-Seq</td>
<td>RNA Sequencing</td>
</tr>
<tr>
<td>RPS4Y1</td>
<td>Ribosomal protein S4</td>
</tr>
<tr>
<td>RVM</td>
<td>random variance model</td>
</tr>
<tr>
<td>SAM</td>
<td>significance analysis of microarrays</td>
</tr>
<tr>
<td>SVA</td>
<td>Surrogate variable analysis</td>
</tr>
<tr>
<td>TLR</td>
<td>Toll-like Receptor</td>
</tr>
<tr>
<td>TNF-α</td>
<td>Tumor Necrosis Factor-α</td>
</tr>
<tr>
<td>USP94</td>
<td>ubiquitin specific peptidase 9</td>
</tr>
<tr>
<td>UTY</td>
<td>ubiquitously transcribed tetratricopeptide repeat gene</td>
</tr>
<tr>
<td>WT</td>
<td>Wild Type</td>
</tr>
<tr>
<td>XIST</td>
<td>X (inactive)-specific transcript</td>
</tr>
</tbody>
</table>
1. Introduction

1.1 Gender Bias in Hepatocellular Carcinoma

Hepatocellular carcinoma (HCC) is one of the most common cancers in the world, with an estimated 782,000 new cases in 2012\(^1\). HCC is the fifth most common cancer in men and ninth in women and with about 746,000 deaths in 2012, it is the second deadliest cancer globally\(^1\). Risk factors for HCC include alcoholic liver disease, infection with Hepatitis C virus (HCV) or Hepatitis B Virus (HBV), and aflatoxin exposure\(^2\). HCC is of particular concern in Egypt because of the high prevalence of HCV. A recent study showed that more than 91% of HCC patients surveyed tested positive for HCV\(^3\), which is in agreement with a previous report showing that almost 89% of surveyed HCC patients were HCV positive\(^4\). Owing to its poor prognosis, HCC is the leading cause of cancer-related death in Egypt, and second in incidence to breast cancer\(^1\).

One important aspect of HCC is a global gender disparity, with a higher prevalence in men by two to four fold, despite equal exposure to risk factors such as HCV and HBV\(^5\). This female protection seems to be hormone-dependent, as indicated by the rise in liver cancer incidence amongst women who undergo menopause\(^6\). Furthermore, hormone replacement therapy in women is associated with lower HCC incidence\(^7\). This sex bias is even more pronounced in mouse models of HCC, which was also found to be sex hormone-dependent\(^8,9\). HCC is chemically induced in mice using a post natal injection of diethylnitrosamine (DEN), which causes DNA damage and hepatocyte death. This triggers inflammatory responses from Kupffer cells, which further promotes compensatory hepatocyte proliferation and leads to tumor formation\(^10,11\). In experiments with DEN-induced HCC, almost all male mice develop the disease, while only 30% of female mice progress to HCC\(^9\). This effect is significantly reduced
by estrogen administration or castration. On the other hand, ovariectomy and / or testosterone supplement significantly reduced female protection\(^9\). This disparity has also been shown to be androgen-receptor-dependent in DEN-induced HCC mouse models\(^{12}\). Mice lacking functional androgen receptors showed resistance to DEN-induced carcinogenesis, and females treated with testosterone showed a higher incidence of liver tumors than untreated females\(^{12}\). In another study, transgenic mice expressing Hepatitis B Surface Antigen and/or p53 were exposed to aflatoxin in different combinations\(^8\). All male mice with these three risk factors developed tumors, compared to only 62% of females\(^8\). Furthermore, males in all other groups developed HCC more frequently than their female counterparts. Some studies have addressed the molecular mechanisms of this disparity in an attempt to elucidate the underlying causes to this phenomenon, but the picture is far from complete.

1.2 Molecular Mechanisms of Gender Bias in HCC

Liver cancer is strongly associated with inflammation\(^{13}\), and thus studies investigating the molecular mechanisms underlying the observed gender bias in HCC found links between sex hormones and inflammatory responses in the liver.

In a key article published in 2007, Naugler \textit{et al.} found that interleukin-6 (IL-6) plays an important role in promoting DEN-induced liver carcinogenesis in mice through signaling via a Toll-like receptor (TLR) adaptor protein\(^{14}\). Males had a higher level of serum IL-6 after DEN administration and a higher level of liver injury than females. IL6\(^{-/-}\) males had lower HCC incidence than wild type (WT), while no difference was observed between WT and IL6\(^{-/-}\). Estrogen administration to males and ovariectomized females reduced liver injury through suppression of IL-6 production. While this study identified the estrogen receptor ER\(\alpha\) as the receptor responsible for this effect, this was later questioned by another group. Bigsby \textit{et al.}
found that ERα status did not affect tumorigenesis in females, and its absence in male mice resulted in less tumors. Interestingly, estradiol treatment reduced the number of tumors in WT male mice, but not in ERα−/− mice, suggesting a protective effect of ERα only for exogenously administered estradiol. Another hormone, prolactin (PRL), has also been indicated as an important factor in HCC resistance in women. Hartwell et. al. found that the pituitary hormone PRL (expressed more in females) restricts innate immune responses in the liver by inhibiting c-Myc activation. PRL was found to signal through a hepatocyte-predominant short-form prolactin receptors (PRLR-S) to inhibit IL-1β, TNF-α, and TLR-4 induced innate responses. PRL acts by ubiquitinating a group of proteins in a “Trafasome” thereby inhibiting their downstream interactions with c-Myc activating pathways. This is in contrast to previous studies which show no role for PRL in HCC gender bias. Tumor incidence in PRLR knockout was comparable to WT, and in females, ovariectomy induced tumorigenesis regardless of PRLR status. This difference has been attributed mainly to the biological differences between PRLR and PRL knockout mice.

In an attempt to identify potential oncogenes and their role in sex-bias occurrence on a larger scale, one group used the sleeping beauty transposon system to screen for potential oncogenes in mice. They found that transposon insertions in the epidermal growth factor receptor (EGFR) gene were more common in tumors from male mice than female mice. Additionally, gene expression analysis in human liver samples revealed differences in male and female liver tissues predominantly pertaining to inflammation, carcinogenesis and reproduction. Together, these results indicate a role for sex hormones, pituitary hormones and the innate immune system in the gender disparity observed in HCC, the details of which remain largely uncharacterized.
The paradoxical results obtained by different groups highlight important points of consideration while studying this aspect of HCC. First, the use of DEN-induced HCC model represents a point of variation as the treatment regimen varies in dose, age of mice at administration, and time of sample collection post DEN administration. Another factor to be considered is the mice genetic background, which as demonstrated by Bigsby et al., affects the number of tumors obtained by a certain treatment and the sex-dependent response\textsuperscript{15}. For example, ovariecctomy did not affect the tumorigenesis in C57Bl/6J females, but showed an effect in the mixed background strain 129Ola-X-C57Bl/6J. Additionally, HCC in humans is largely associated with a other external factors\textsuperscript{2} (such as alcohol, viral infection, chemical exposure), which are not represented in animal models. Therefore, it is evident that while mice provide an invaluable experimental tool, they do not accurately reflect the pathogenesis of HCC in humans, nor the sex bias observed. A valuable source of information would be clinical samples from HCC patients and human liver tissues to provide a more accurate and reliable

1.3 Studying Gene Expression using Microarrays

One way to extract valuable information from clinical data is to utilize already available genomewide microarray data from human liver tissues to identify genes with potential roles in the sex bias observed in HCC. High density oligonucleotide arrays represent a very useful tool for studying gene expression as they allow the quantitative detection of mRNA. Microarray chips, such as the Affymetrix GeneChip, use tens of thousands of synthetic oligonucleotides (25 bases long) that hybridize to specific sequences of target genes\textsuperscript{18,19}. Commercially available platforms allowed the widespread use of microarrays for high throughput transcriptome studies, and the accumulation of large amounts of gene expression data. Microarray studies are often used to study disease pathogenesis or treatment effectiveness by comparing gene expression in normal
and diseased tissues, treated and untreated samples or multiple stages of one disease, among many other alternatives. These approaches allow the large-scale identification of genes that are up-regulated and down-regulated in experiment versus control, allowing subsequent downstream functional analyses. For example, microarray studies have been used to identify gene expression signatures in cancers of the colon\textsuperscript{20}, prostate\textsuperscript{21}, breast\textsuperscript{22}, liver\textsuperscript{23,24} and lung\textsuperscript{25} to identify diagnostic and prognostic biomarkers and therapeutic targets. Public databases, such as the NCBI Gene Expression Omnibus (GEO)\textsuperscript{26,27} and ArrayExpress\textsuperscript{28} allow access to a huge number of microarray data.

Although various gene expression studies addressed different aspect of HCC pathogenesis, only one study attempted to dissect the molecular basis of the observed gender bias. Hartwell \textit{et al.} carried out gene expression analysis on 7 male and 7 female human liver samples to identify a gene signature that could be protective of females as compared to males\textsuperscript{16}. Their findings distinguished about 500 genes that are differentially expressed between the 2 groups, many of which have functions in cell cycle, inflammation and cancer\textsuperscript{16}.

Our aim in the current study is to build on these results by using a larger sample size to define a more specific signature, which can be followed in patients that develop HCC. One approach to achieve this is to use data from multiple studies that are already published. However, lack of correlation between platforms and experimental variations among labs do not allow a direct comparison between heterogeneous studies\textsuperscript{29,30}. One way to overcome this is to carry out a meta-analysis of the data. Meta-analysis is a statistical approach, consisting of a set of statistical techniques that allow the combination of data from independent, but relevant, studies\textsuperscript{31}. Combining data from various studies using a meta-analysis approach increases the statistical power to allow the detection of small, but consistent changes that are otherwise missed in single-
study analyses. Furthermore, statistical approaches directed to overcome heterogeneity among the different datasets increases the sensitivity and reproducibility of the result, when compared to individual studies. Another advantage of meta-analysis is its relative low-cost, as the data are already available and many analysis tools are open-source.

The success of this approach is evident by the increasing number of researchers employing it to generate novel information from already existing data, such as common transcriptional profiles or gene signature of certain cancers, and even new prognostic biomarkers. Rhodes et al. used meta-analysis on 40 independent microarray cancer studies and identified a “meta-signature” of cancer transformation and progression. In a more clinically-driven approach, Mehra et al. combined 417 samples and identified GATA3 as a prognostic marker for breast cancer. Also, Ewald et al. used patient-derived microarray data to identify pathways involved in the stage-wise progression of bladder cancer from papillary to muscle-invasive tumors. In addition, efforts to compare and unify meta-analysis approaches and workflows have been emerging to enhance the applicability of such studies and enhance their outcome. Overall, meta-analysis is a powerful approach to use already-available data to answer novel questions, which is what we aim to accomplish in this study. However, as explained below, careful consideration must be given to using public microarray datasets for differential expression analysis, owing to the heterogeneity of their experimental design, the platforms used, and the data processing methods.

1.4 Microarray Data Pre-processing

The output of microarray experiments, i.e. raw data, has to be pre-processed in order to reach meaningful, informative values. Affymetrix microarray chips contain around 11-20 probes for
each gene, arranged in pairs known as “probe pairs”. Probe pairs consist of perfect match (PM), which
match the target gene completely, and mismatch (MM) probes, which are the same sequences as
PM probes, but with a mismatched base in the middle (13th base) to account for non-specific
binding. Therefore, in order to have a single expression value for each probe, the data from all
probes in a probe set needs to be summarized in an expression value. In addition, normalization
of the data is crucial to account for obscuring variations that occur during sample preparations,
handling, production and processing of the arrays.

One of the most widely-used methods for pre-processing of Affymetrix microarray data
is the Robust Multichip Average (RMA) method. The RMA expression measure is generated by
background correction of probe-level PM values, followed by quantile normalization of the data
and then linear fitting using the median polish method, to generate log expression values. The
quantile normalization approach utilizes information from all arrays with the goal of making the
distribution of probe intensities the same for all arrays in a particular set analyzed. Therefore,
it is important to process all arrays from an experiment together as one batch when using RMA.
When compared to other normalization methods, such as non-linear and scaling methods (used in
the Affymetrix algorithms), quantile normalization was most successful at reducing variance and
produced the smallest distances between arrays in pairwise comparisons which remained
constant across intensities. It also performed better in terms of bias and speed. This
improvement in performance could be attributed to using all arrays in the normalization process,
rather than a single baseline, which is less representative of the complete data.
Additionally, algorithms such as the Affymetrix MAS 5.0 algorithm and dChip software developed by Li and Wong rely on subtracting the MM signal to correct for non-specific binding \(^{43}\). However, RMA does not use the MM probes for subtraction, and relies only on PM values, as studies have shown that MM probes detect signal in addition to non-specific binding, and that methods utilizing PM-MM or PM/MM values add noise and result in a biased signal, respectively \(^{41,44}\). In a series of spike-in experiments, RMA was compared to other expression measures, such as Li and Wong’s model-based expression indexes (MBEI) and the MAS 5.0 signal, in which it performed better in 3 criteria: (1) precision, as estimated by the gene-specific standard deviations (SD) across replicates; (2) consistency of estimates for fold change; and (3) the specificity and sensitivity when using fold change to detect differential expression \(^{41,45}\).

1.5 Differential Expression Analysis

The simplest way to identify differentially expressed gene is using the Fold Change (FC) indicator, which evaluates the average log ratio between 2 conditions or groups and considers gene with FC above a certain (arbitrary) threshold differentially expressed. For example, if the FC threshold is set at 2, genes that are found to be more abundant under one condition or less abundant in the other by more than 2 fold are considered to be differentially expressed. However, FC is not a statistical test; it does not provide any level of confidence and does not account for variance across samples. Therefore, it is important to employ a statistical method to account for this variance and to standardize differential expression.

A simple and popular method to employ is a standard t-test, which has been widely used to detect differentially expressed gene for microarray studies \(^{46}\). The t-test is conducted for each gene and the error variance is estimated based on the log ratios. For each gene, a t-statistic is computed and then converted to a p-value. Typically, genes with p-values falling below a
specific threshold are considered significant. Despite the popular application of the t-test to microarray data analysis, this approach has its drawbacks, particularly with low-variance genes and small sample size, where some bias is introduced and the statistical power is compromised, respectively \(^{47,48}\). Consequent to these criticisms, numerous statistical approaches have been developed to address different areas of concern, in attempts to improve variance estimates, accuracy and statistical power. The abundance and variety of these methodologies make it difficult to decide on which method would be most effective in analyzing gene expression data and which method would be most appropriate to specific experimental settings.

Fortunately, studies were designed to specifically address this issue, and some have utilized data in which the differentially expressed genes are already known (for example, using spike-in experiments) to evaluate the performance of statistical tests \(^{47-50}\). For example, one study has compared the performance of 8 statistical methods of variance modeling in gene expression data analysis: Welch's t-test, analysis of variance (ANOVA), Wilcoxon's test, significance analysis of microarrays (SAM), random variance model (RVM), Limma, VarMixt and SMVar \(^{47}\). Using Limma resulted in the best overall performance compared to the other 7 methods, where the most improvement was achieved compared to the t-test, especially with small sample size, and in terms of ease of use and speed of execution.

Limma is a popular method used for microarray data analysis, which has recently been adapted for RNA-sequencing (RNA-Seq) data as well. The principal idea in Limma is the use of gene-wise linear model fitting to analyze an entire experiment as a whole \(^{51}\). The linear models are fitted to each row, and regression coefficients and standard error generated for the compared group. The parallel nature of gene expression experiments motivates the use of empirical Bayes statistics which allows sharing of information between genes to account for variances across
genes and samples to obtain Bayes posterior variance estimators\textsuperscript{51}. The estimated sample variances are squeezed towards a common variance, leading to more stable inferences from small sample size\textsuperscript{52}. In addition, the use of gene-wise linear modeling allows flexibility in handling different experimental designs. In testing for differential expression, empirical Bayes statistics, such as moderated t-statistic and p-values are generated for each coefficient of the linear model in order to assess the significance of these changes. Limma is able to reduce the false positive rate in genes with low variance and improve the power for detecting genes with large variances. The use of empirical Bayes statistics in Limma has been found favorable to other significance testing methods in gene expression analysis, and its performance favored especially with small samples\textsuperscript{49,50}.

1.6 Accounting for Batch Effects

Another important source of variation in microarray data that needs to be addressed is the variation introduced by non-biological factor that causes differences between samples, known as the “batch effect”. Batch effects are introduced when samples are run in different “batches”, which could be different runs, different days, using different reagents or different technicians \textsuperscript{53–55}. Batch effects can affect the downstream processing of the data, resulting in lower power to detect real changes, or more the serious consequences of false or misleading biological conclusions. Although batch effects have been demonstrated in early microarray experiments \textsuperscript{56}, surprisingly, they have been reported only in small percentage of studies. For example, one study reported that out of 219 papers published using microarray data, less than 10% addressed batch effects \textsuperscript{53}. Moreover, upon examining data from 9 high throughput studies, Leek et al. found all of them had considerable batch effects, with substantial percentages (32.1-99.5%). While batch effects can be minimized through careful experimental design, the only way to avoid them
completely is by running samples as one batch. Therefore, several methods for batch effect removal have been developed, including Distance-weighted discrimination (DWD)\textsuperscript{57}, mean-centering (PAMR)\textsuperscript{58}, Surrogate variable analysis (SVA)\textsuperscript{59}, Geometric ratio-based method (Ratio\_G)\textsuperscript{60} and Combating Batch Effects When Combining Batches of Gene Expression Microarray Data (ComBat)\textsuperscript{54}. When all these 5 methods were compared and performance assessed based on accuracy, precision and variation reduction (batch effect removal), using ComBat resulted in the best overall performance. In addition, ComBat was robust for handling small batches that other methods did not perform well with\textsuperscript{53}. This is attributable to the empirical Bayes framework employed in the ComBat approach, which estimates location and scale adjustment parameters for each gene independently by borrowing information across genes to “shrink” the batch effect parameter estimates towards an overall mean of the estimates\textsuperscript{54}. These estimates are then used to adjust the data for batch effects, providing robust adjustment for small batches (10\textless)\textsuperscript{54}.

Although normalization of microarray data does not account for batch effects, some modifications of existing methods attempt to address these effects in batches of datasets or single arrays that can be analyzed individually or after before combining them with others. The frozen robust multichip average (fRMA) is used to achieve the advantages of multichip processing to single-array analysis by using a large dataset of representative samples to create a reference distribution for the subsequent quantile normalization. Pre-computed estimates of probe effects are used in concert with data from the set being analyzed to generate the summary expression values. fRMA was found to perform similarly to RMA when the data was preprocessed as one batch, but outperformed RMA in terms of precision when analyzing multiple batches.
Using the above tools for analyzing gene expression data from human liver tissues, we aim to identify differentially expressed genes between males and females that could be responsible for the inherent HCC resistance in females, which would be a starting point for subsequent *in silico* analysis and *in vitro* and *in vivo* experiments.

2. Materials and Methods

2.1 Data Collection and Processing

The NCBI GEO database ([http://www.ncbi.nlm.nih.gov/geo/](http://www.ncbi.nlm.nih.gov/geo/)) was searched for “human liver”, with the search filters set as follows: “Organism” set to “Homo sapiens”, “Data type” set to “expression profiling by array”, “Attribute name” set to “tissue to satisfy our inclusion criteria for whole-transcriptome studies using human tissues from normal livers. In order to unify gene IDs and normalization, we only selected data produced using Affymetrix chips. Studies using cell lines, animals, non-coding RNA profiling, or different platforms were excluded. For the selected datasets, raw (.cel) files were downloaded and pre-processed using the RMA method to generate expression value output\(^{41,45}\), which is summary measure of background-corrected, normalized log-transformed probe intensities. RMA was applied by implementing the “just RMA()” function in the Bioconductor Affy Package \(^{61}\). As with all other statistical analyses used in this study, all methods were implemented in the R statistical computing environment ([http://www.r-project.org/](http://www.r-project.org/)), and all packages are available from the open source Bioconductor project\(^{62,63}\) ([http://www.bioconductor.org/about/](http://www.bioconductor.org/about/)).

2.2 Data Exploration and Differential Expression Analysis

As an initial exploratory step, hierarchical cluster analysis was carried out using the “hclust” function in R, with the default method of complete linkage, which defines the distance between 2 clusters as the maximum distances between its components. Distance matrices were constructed
by using the Pearson distances between columns (1- Pearson’s correlation coefficient), and the analysis displayed as a dendrogram.

To identify differentially expressed genes between male and female groups, we used 2 methods: a standard t-test between the 2 groups and the Limma Bioconductor package which implements linear modeling and Bayesian statistics. Using the t-test, p-values were calculated for each probe, and a cut-off of 0.05 was set so that only probes with p-values below this limit are selected. Absolute fold change cut-off was set as 2, so genes which are up- or down-regulated by at least 2 fold are selected. Probes that met both of these criteria were filtered for further downstream analysis.

For the Limma Package, the “lmFit” function was used for the estimation of fold changes and standard errors through fitting a linear model, which is specific by the design matrix, for each gene. The fitted model was then processed using the “eBayes” function to apply empirical Bayes statistical methods and generate statistics such as the moderated t-statistic and its associated p-value, adjusted p-values for multiple testing and average log expression values. The selection cut-off for differentially expressed genes was set to a p-adjusted value of 0.05 (using the Holm’s sequential Bonferoni multiple testing correction method). This ensures that errors introduced due to testing multiple hypotheses are corrected for and so minimizes false positives.

2.3 Dataset Merging and Batch Effect Removal

The datasets were combined together using the inSilicoMerging Bioconductor package, which allows the merging of different datasets and the use of different methods for batch effect removal. Using this package, we used the ComBat approach, which utilizes empirical Bayes statistics
to correct for variations between arrays due to use of different methodology, or data generated in different laboratories, such as the case in this study\textsuperscript{54}.

Another method for batch effect adjustment that was employed was the fRMA method, which allows pre-processing of individual datasets by using pre-computed probe effects and variances during the normalization process\textsuperscript{66}. This way, individually fRMA-processed datasets can be computed without the need for further batch effect removal. Datasets were fRMA pre-processed using the “fRMA” function in the fRMA Bioconductor package, then combined using the inSiliocMerging package and then analyzed for differentially expressed probes using the Limma package.

2.4 Gene ID conversion and functional annotation

To convert probe IDs to gene names and identify the location and function of these genes, The Database for Annotation, Visualization and Integrated Discovery version 6.7 (DAVID, http://david.abcc.ncifcrf.gov/) tool was used\textsuperscript{67,68}. Each list of probes was copied and pasted into the Gene List Manager and analyzed either by the Gene ID Conversion tool to identify the gene corresponding to the selected probes or the Gene Functional Classification tool to group genes into functional groups.

3. Results

3.1 Dataset Collection and Processing

The GEO search yielded 739 studies on January 2015, of which 7 studies were selected (summarized in Table 1) that provided expression data for normal human liver. All selected datasets were generated using the Affymetrix Human Genome U133 Plus 2.0 Array, except GSE14323, which was generated using the Affymetrix GeneChip Human Genome U133A 2.0
Array. Out of those, only 2 (GSE23343, GSE14951) had gender information readily available in the database. We contacted the principal author for each study, and only the author for GSE143323 provided the missing gender data. Consequently we had a total of 3 studies qualifying for the subsequent analysis. Collectively, the datasets included 27 male and 19 female normal liver samples, adding up to 46 samples in total (summarized in Table 2). As the data provided in the expression matrices was processed differently by each group, raw .cel files were downloaded and pre-processed using the same RMA method.

3.2 Individual Dataset Analysis Using the T-test Method

For each dataset, hierarchical clustering analysis was performed, using the complete linkage method, and plotted as dendrograms (Figure 1). Hierarchical clustering for all three datasets did not generate a distinct gender cluster (data not shown). Each dataset was then divided into 2 groups of male and female samples and compared using a t-test statistic and FC. A cut-off of 0.05 for p-value and 2 for FC was used to filter the probes that were differentially expressed between the 2 groups. The data are represented in a form of a heat maps depicting up-regulated and down-regulated probes in each dataset (Figure 2). For dataset GSE14323, a total of 6 probes were differentially expressed and in GSE23343 a total of 19 differentially expressed probes were identified. Finally, dataset GSE14951 showed only 2 differentially expressed probes, that did not produce complete segregation of the 2 groups. The DAVID tool was used to map the probe ID to genes names. Table 3 summarizes the differentially expressed probes in all datasets, highlighting their overlapping probes and their corresponding genes. Since multiple probes are used to monitor the same gene, the total number of genes is sometimes smaller than the number of probes. Therefore, for GSE 14323, 6 probes corresponded to 5 genes including genes involved in spermatogenesis, X-inactivation and protein biosynthesis. For GSE23343, 19 probes...
corresponded to only 11 genes, including genes involved in hexose metabolism, histone demethylatoin and protein deubiquitination. Only one gene was found to be differentially expressed in all three datasets, which is the Ribosomal protein S4 (RPS4Y1), a Y-linked gene. Most of the genes found were either Y-linked or X-linked, with only 3 out of 20 found on autosomes. Additionally, none of the genes in all datasets clustered into functional groups when using the DAVID gene functional classification tool. Finally, for the apparent heterogeneity of the GSE14951 dataset sample pool, it was excluded from further analysis.

3.3 Individual Dataset Analysis Using the Limma Package and Bayesian statistics

The Limma Bioconductor package\textsuperscript{51} was used to further analyze the data as an alternative method to determine differentially expressed genes between the 2 groups in 2 datasets: GSE14323 and GSE23343 (figure 3). For GSE14323, 9 probes were found to be differentially expressed which included 5 of the 6 probes found by the t-test method. The 4 “new” probes were already identified for the other dataset (GSE23343) using the t-test. On the other hand, only 8 probes were indentified for GSE23343, and again, considerable overlap was observed, as 7 out of the 8 probes were also differentially expressed among the 19 probes identified in the t-test method (data summarized in tables 4 and 5). Using the Limma package added only one new gene, which is La ribonucleoprotein domain family, member 4B (LARP4B), found on chromosome 10.

3.4 Analysis of Merged Datasets Using the t-test

To test if combining these 2 datasets would provide additional information, or allow for the detection of changes that were not previously detected, the 2 datasets were merged together and a t-test was used to identify significantly up- or down- regulated probes (figure 4). Ten probes were identified to be differentially expressed, corresponding to 7 genes. All of these probes have
been identified in individual dataset analysis, and the genes they correspond to are Y- or X-linked genes, with the exception of just one (summarized in table 6). A hierarchical cluster analysis of the merged datasets resulted in the clustering of the 2 datasets into separate groups (figure 5), regardless of the gender of the sample.

3.5 Batch Effects Removal

To correct for differences between data generated in different laboratories, or at distinct time periods, various batch effect removal tools were used. Two methods were employed to correct for batch effect correction: either the fRMA normalization\(^6\) prior to merging the datasets or using the ComBat\(^5\) tool after merging the datasets. In both cases, the Limma package\(^5\) was employed to identify differentially expressed genes. For GSE14323, 12 probes were identified, corresponding to 9 genes (figure 6A). Three new genes for this dataset were identified after batch effect removal: patatin-like phospholipase domain containing 4 (PNPLA4), ubiquitously transcribed tetratricopeptide repeat gene (UTY), and ubiquitin specific peptidase 9 (USP94) (table 7). PNPLA4 is X-linked while UTY and USP94 are Y-linked. Also, USP94 was identified before for the other dataset using Limma. For GSE23433, 9 probes were found to be differentially expressed (figure 6B). As shown in table 8, these probes correspond to only 4 genes, none of which are new compared to previously identified probes for this dataset.

After merging the 2 fRMA-normalized datasets, we obtained 10 differentially expressed probes (figure 7). These corresponded to 8 genes, 7 of which were sex-chromosome linked, and all of which have been already identified in earlier individual dataset analyses. Using the ComBat method for batch effect removal, 13 probes were identified as differentially expressed—corresponding to 10 genes, with considerable overlap with those identified using fRMA (table 9).
Out of these ten, only one was not previously identified for either dataset, which is lysine (K)-specific demethylase 6A (KDM5D), an X-linked gene.

3.6 Gene Signature Validation

Recently published microarray data showed around 500 genes to be significantly differentially expressed between normal male and female human liver tissue\textsuperscript{16}. Since our analysis did not reveal gender bias in human liver tissues gene expression, we tested if this gene set would produce a similar signature using our two datasets. The gene set was successful in distinguishing between their male and female samples, as they clustered into two distinct classes, but there was no clear signature of up- and down-regulated genes that could be distinguished.

4. Discussion

We used data from microarray studies of human liver tissue to compare female and male gene expression and identify differential expression patterns. We identified a subset of X- and Y-linked genes that are up-regulated in females and males, respectively.

As microarray data available from public data bases are found in a host of different formats, it is usually recommended to start any analysis with raw files to be pre-processed in a unified manner, to make the output comparable. We chose the RMA pre-processing approach due to its effectiveness and precision, as it is becoming the method of choice for Affymetrix microarray data processing.

The initial clustering of the individual datasets did not show any gender-based segregation, but as the analysis was done for all probes for each dataset, subsets of differentially expressed genes could not be determined or visualized using this exploratory step. Therefore, the t-test statistical analysis was carried out to identify differentially expressed genes between males and females. Although using the t-test for this application has been criticized, it was used in the
initial exploration of the data because the experimental design is simple (only 2 groups compared, representing 2 conditions) and also for comparison to other methods. Furthermore, the fold change was used as an additional filtering criterion in the selection for biologically meaningful differences. Although the number of identified probes was small for each dataset (19 for GSE23343, 6 for GSE14323, and 2 for GSE14951), there was considerable overlap between them, indicating consistency for those genes. The most notable is RPS4Y1, that is up-regulated in males in all datasets, which is not surprising given it is a Y-linked gene (functional analysis for relevant probes will be discussed in later sections).

To verify these results and refine our analysis by using improved statistical approaches that are more appropriate for small sample size, we used the Limma package for differential expression. The Limma package has been shown to be superior in performance and accuracy than various other statistical and increased power compared to the t-test. When we used Limma for individual dataset analysis, we identified more probes for GSE14323 and less for GSE23343, and overall identified just one new gene: LARP4B, an autosomal gene found on chromosome 10. It is worth noting that the cut-off used for Limma was a p-adjusted value that corrects for multiple testing, which is a stricter cut-off than using standard p-values in a t-test. This indicates that Limma is indeed more powerful than the t-test, and was able to detect changes that are more significant than the t-test could account for, and therefore Limma was used for all subsequent differential expression analysis. This would also explain the lower number of probes obtained for GSE23343. Similar to the individual analysis using the t-test, considerable overlap was found between the differentially expressed genes identified using Limma for the 2 datasets.

Merging the two datasets together in a “meta-analysis” approach is expected to increase the statistical power and allow for the detection of other small changes in gene expression.
However, using a t-test and fold change to filter out differentially expressed genes did not add any new probes compared to those identified by individual analysis. To further explore the data, a hierarchical analysis revealed that the datasets clustered into 2 different groups, regardless of the gender. This indicates that non-biological differences arising due to the different origin of the samples (i.e. batch effects) are interfering with the analysis and could be obscuring gender-specific changes. Therefore, before carrying out any further analysis on the merged datasets, we used batch effect removal tools to overcome this variation.

For this purpose, 2 methods were used: fRMA and ComBat. fRMA was used to preprocess the raw data for both sets before merging. The advantage of this approach is that each set is analyzed individually and more datasets can be added to the analysis without the need to repeat pre-processing or batch effect removal steps that were already done. For GSE14323, three new genes were identified using Limma, while none were newly found for GSE23343. This could mean that the inter-sample non-biological variations are marginal, or that these batch effects were not obscuring considerable biological changes.

Finally, for the purpose of the meta-analysis approach we are employing in this study, the 2 datasets were merged and processed using ComBat for batch effect removal and limma for the identification of differentially expressed genes. This was compared to merging the fRMA-processed datasets. Again, similar previous results, merging the fRMA datasets did not result in detecting new or different genes, and using ComBat only added one new gene: KDM5D.

Throughout our analysis, the differences in output we observed were obtained when we used different statistical methods, e.g. t-test versus Limma, or RMA versus fRMA or ComBat. Merging the datasets using the same pre-processing and downstream analysis algorithms resulted in almost the same outcome, which indicated that merging the 2 datasets together does not
provide additional power to the analysis and does not add information that would have been otherwise undetected. However, we only used 2 (relatively small) datasets, with sample sizes below 25 for each group, which may have reduced the power of the study.

Compared to the ~500 genes identified by Hatrwell et al. to be differentially expressed between males and females in human normal liver tissues, our identified genes are very few. Although using these genes in hierarchical analysis allowed us to cluster male and females in each dataset, they did not exhibit any consistent pattern for up-and down-regulated genes. However, we examine below the genes that were found in common between this study and ours, and additionally, genes that were found to be consistently deferentially expressed in our analysis.

For the male-dominated genes, we identified 6 that are consistently found to be up-regulated in males: RPS4Y1, EIF1AY, CYorf15B, UTY, DDX3Y and USP9Y, all of which are Y-linked. The first three were also found to be up-regulated in males by Hartwell et al. The most prominent gene that is identified by all approaches employed in this study is RPS4Y1, which is part of the 40S ribosomal small subunit\(^73\). The sex-bias in RPS4Y1 expression was observed in human brain\(^74\), heart\(^75\) and was found to be pronounced in human prostate cancer tissues and cell lines\(^76\). RPS4Y1 has a homologue on the X chromosome, RPS4X, which is known to escape X-inactivation, but was not found to be differentially expressed in our sample pool\(^73\). Therefore, up-regulation of RPS4Y1 seems to be tissue non-specific, and its role in any potential tumor transformation is unknown.

DEAD (Asp-Glu-Ala-Asp) box polypeptide 3Y (DDX3Y) is an RNA helicase that is essential for spermatogenesis. DDX3Y was found to be widely transcribed but the protein was only found in the testis\(^77\). The X-linked homologue, DDX3X is also ubiquitously expressed and both function as nucleo-cytoplasmic shuttles for RNA\(^77,78\). DDX3Y was also up-regulated in
human heart, while DBY (another member of the DEAD box protein family) was also up-regulated in male brain tissue. Another protein that is important for spermatogenesis is USP9Y, is ubiquitously expressed in embryonic and adult tissues and is found up-regulated also in brain, heart and prostate tissues of males\textsuperscript{74–76}. It is a ubiquitin-specific protease that functions in the de-ubiquitination of target proteins, and thus plays a role in protein turnover and regulation\textsuperscript{79,80}.

Eukaryotic translation initiation factor 1A-Y (EIF1AY) is a translation initiation complex that interacts with the ribosome, and is required for achieving the maximum rate in protein biosynthesis\textsuperscript{81}. It also may function in stabilizing the binding of the initiator Met-tRNA to 40S ribosomal subunits\textsuperscript{82}. UTY is also found to be up-regulated in male prostate, heart and brain tissues, and it encodes a protein with tetratricopeptide repeats that has been found to have multiple splice variants, the functions of which are unknown\textsuperscript{83}. It has been found to code for a male-specific minor histocompatibility antigen involved in stem cell graft rejection\textsuperscript{84}. Finally, CYrof15B has also been found to be up-regulated in liver and heart tissue, but so far is found to be a pseudogene of the human taxilin gamma (TXLNG).

For the genes found to be up-regulated in females, 2 are X-linked: X (inactive)-specific transcript (XIST), PNPLA4, and the third, PZP, is autosomal, found on chromosome 12. XIST is found on the X inactivation center (XIC)\textsuperscript{85} and which functions in silencing of X-linked genes through X-chromosome inactivation. XIST was found to be non-protein coding and instead functions as structural RNA in the nucleus\textsuperscript{86}. XIST is found to be up-regulated in human brain (specifically neurons), heart and liver\textsuperscript{16,74,83}. PNPLA4 belongs to a family of lipid hyrolases and is a potent retinylester hydrolase in keratinocytes, affecting their morphology\textsuperscript{87}. It is expressed in a variety of tissues, albeit in different transcript lengths, suggesting differential processing across tissues\textsuperscript{88}. Finally, the only autosomal gene to be up-regulated in liver of females in our study that has been
previously reported is PZP. PZP is a plasma protein that functions as a peptidase inhibitor\textsuperscript{89}. It has been found to associate with TGF-beta-1 and TGF-beta-2 and regulates their plasma clearance\textsuperscript{90}. Since TGF-beta signaling is highly implicated in hepatic carcinogenesis, PZP may play a role in this pathway. This is further corroborated by data from Genome-wide associated study where single nucleotide polymorphisms in PZP associated with high serum AST\textsuperscript{91}.

It is not unexpected to find X- and Y-linked genes to be differentially expressed among females and males, but to explain why only a few of the >1300 gene on the X-chromosomes have been found to be up-regulated in females and only a few Y-linked genes have been up-regulated in males, some considerations have to be taken. X-inactivation of the X-chromosome in somatic cells of females results in similar expression patterns for X-linked genes in males and females. Dosage compensation for the active X-chromosome results in hypertranscription of the X-linked genes to reach the level of autosomal genes (present in 2 copies). This could explain the lack of differential expression of most X-linked genes observed in our study\textsuperscript{93}. To add another level of complication, some genes are found to escape this X-inactivation, which results in higher expression of these genes in females compared to males\textsuperscript{92}. However, this inactivation is found not be consistent between individuals, or within cells and tissues of an individual, leading to variations in expression that are yet to be further analyzed\textsuperscript{94}. Additionally, due to the presence of various X-Y homologues (homologues of the same gene serving the same function), cross-hybridization of the Y-specific probes to X-homologues has been reported and could compromises the ability of the Y-chromosome probes to differentiate male and female samples, thereby resulting in similar expression levels\textsuperscript{74,95}. Additionally, a lack of correlation between mRNA levels and protein levels has been observed for some genes, as the case for DDX3Y,
necessitating downstream functional analysis on the protein level before solid conclusions can be made\textsuperscript{77}.

Overall, while the differentially expressed genes identified in this study did not produce any functional clusters using the gene functional classification tool in DAVID, they are consistent with other studies comparing male and female human tissues. The expression of these genes is not specific to the liver, as they are found in other tissues as well including brain, heart and prostate. However, several studies have linked sex chromosomal aberrations and Y-linked gene expression to cancer. For example, up-regulation of Y-linked gene such as RPS4Y1, UTY, EIF1AY and USP9Y has been found in prostate cancer tissue and cell lines, compared to benign prostatic hyperplasia and normal testis\textsuperscript{76,96}. In contrast, Y-chromosomal deletions have been also associated with prostate cancer\textsuperscript{97–99}, male breast carcinomas\textsuperscript{100} and pancreatic adenocarcinomas\textsuperscript{101}. Loss of Y-Chromosome in peripheral blood cells was found to be associated with higher risk of cancer in men\textsuperscript{102}. Additionally, rearrangements of the X-chromosomes, including deletions and gains, have been associated with breast\textsuperscript{103}, ovarian\textsuperscript{104} and uterine cervix cancer\textsuperscript{105}. Loss of heterozygosity (LOH) has been found in endocrine carcinomas of the gastroenteropancreatic tract, lung and colorectal cancer\textsuperscript{106–108}. Therefore there is a link between sex-chromosome genes and both gender-specific and non-specific cancer, which remains to be fully elucidated.

Although this study provides a preliminary meta-analysis of the gender-bias in normal liver tissue utilizing two microarray datasets, further studies should analyze more samples, with gender information for the utilized datasets. A larger sample pool would allow the stratification of patients into age groups. As cancer is largely a disease of ageing, analyzing data from more uniform age groups could be more informative. In addition, more advanced techniques, such as
RNA-Seq could provide more information about mutations in X- and Y- chromosome genes, as well as a more accurate and sensitive measurement of X-Y homologues that could have interfered with the microarray analysis. Finally, specific studies of already identified genes in the liver and their corresponding proteins would provide more information about their role in liver biology and regulation during carcinogenesis.
**TABLES**

**Table 1.** Summary of the datasets of microarray studies using human normal liver tissue. Datasets with available gender information are highlighted.

<table>
<thead>
<tr>
<th>Accession</th>
<th>Sample size</th>
<th>Sample types</th>
<th>Gender</th>
<th>Platform</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSE14323</td>
<td>19</td>
<td>Normal Liver</td>
<td>Yes</td>
<td>Affymetrix (GPL571)</td>
</tr>
<tr>
<td>GSE14951</td>
<td>4</td>
<td>Normal liver (prior to transplantation)</td>
<td>yes</td>
<td>Affymetrix (GPL570)</td>
</tr>
<tr>
<td>GSE23343</td>
<td>7</td>
<td>Normal liver</td>
<td>yes</td>
<td>Affymetrix (GPL570)</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>Normal liver (with diabetes)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSE13471</td>
<td>4</td>
<td>Normal liver</td>
<td>none</td>
<td>Affymetrix (GPL570)</td>
</tr>
<tr>
<td>GSE6222</td>
<td>2</td>
<td>Normal liver</td>
<td>none</td>
<td>Affymetrix (GPL570)</td>
</tr>
<tr>
<td>GSE45436</td>
<td>39</td>
<td>Normal liver</td>
<td>None</td>
<td>Affymetrix (GPL570)</td>
</tr>
<tr>
<td>GSE38941</td>
<td>10</td>
<td>Normal liver</td>
<td>None</td>
<td>Affymetrix (GPL570)</td>
</tr>
</tbody>
</table>
Table 2. Sample information of three microarray datasets with gender information

<table>
<thead>
<tr>
<th>Accession</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSE14323</td>
<td>12</td>
<td>7</td>
</tr>
<tr>
<td>GSE14951</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>GSE23343</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>27</strong></td>
<td><strong>19</strong></td>
</tr>
</tbody>
</table>
Table 3. Summary of differentially expressed probes found in all three data set and their corresponding gene names. Common probes found between datasets are highlighted.

<table>
<thead>
<tr>
<th>GSE14951</th>
<th>GSE143233</th>
<th>GSE23343</th>
<th>Gene Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>201909_at</td>
<td>201909_at</td>
<td>201909_at</td>
<td>Ribosomal protein S4, Y-linked 1</td>
</tr>
<tr>
<td>224590_at</td>
<td>224590_at</td>
<td></td>
<td>X (inactive)-specific transcript (non-protein coding)</td>
</tr>
<tr>
<td></td>
<td>203649_s_at</td>
<td></td>
<td>Phospholipase A2, group IIA (platelets, synovial fluid)</td>
</tr>
<tr>
<td>204409_s_at</td>
<td>204409_s_at</td>
<td></td>
<td>Eukaryotic translation initiation factor 1A, Y-linked</td>
</tr>
<tr>
<td>205000_at</td>
<td>205000_at</td>
<td></td>
<td>DEAD (Asp-Glu-Ala-Asp) box polypeptide 3, Y-linked</td>
</tr>
<tr>
<td>214218_s_at</td>
<td>214218_s_at</td>
<td></td>
<td>X (inactive)-specific transcript (non-protein coding)</td>
</tr>
<tr>
<td>221728_x_at</td>
<td>221728_x_at</td>
<td></td>
<td>X (inactive)-specific transcript (non-protein coding)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>206700_s_at</td>
<td>Lysine (K)-specific demethylase 5D</td>
</tr>
<tr>
<td></td>
<td></td>
<td>207063_at</td>
<td>Chromosome Y open reading frame 14</td>
</tr>
<tr>
<td></td>
<td></td>
<td>207330_at</td>
<td>Pregnancy-zone protein</td>
</tr>
<tr>
<td></td>
<td></td>
<td>214131_at</td>
<td>Chromosome Y open reading frame 15B</td>
</tr>
<tr>
<td></td>
<td></td>
<td>204410_at</td>
<td>Eukaryotic translation initiation factor 1A, Y-linked</td>
</tr>
<tr>
<td></td>
<td></td>
<td>223645_s_at</td>
<td>Chromosome Y open reading frame 15B</td>
</tr>
<tr>
<td></td>
<td></td>
<td>223646_s_at</td>
<td>Chromosome Y open reading frame 15B</td>
</tr>
<tr>
<td></td>
<td></td>
<td>224588_at</td>
<td>X (inactive)-specific transcript (non-protein coding)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>205001_s_at</td>
<td>DEAD (Asp-Glu-Ala-Asp) box polypeptide 3, Y-linked</td>
</tr>
<tr>
<td></td>
<td></td>
<td>227614_at</td>
<td>Hexokinase domain containing 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>227671_at</td>
<td>X (inactive)-specific transcript (non-protein coding)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>228492_at</td>
<td>Ubiquitin specific peptidase 9, Y-linked</td>
</tr>
<tr>
<td></td>
<td></td>
<td>235942_at</td>
<td>Hypothetical LOC401629</td>
</tr>
</tbody>
</table>
Table 4 presents a summary of differentially expressed probes identified in GSE14323 using the Limma Package compared to those found using the T-test method. Common probes are highlighted.

<table>
<thead>
<tr>
<th>Limma</th>
<th>T-test</th>
<th>Gene Name</th>
<th>Y or X-linked</th>
</tr>
</thead>
<tbody>
<tr>
<td>201909_at</td>
<td>201909_at</td>
<td>Ribosomal protein S4, Y-linked 1</td>
<td>Y</td>
</tr>
<tr>
<td>204409_s_at</td>
<td>204409_s_at</td>
<td>Eukaryotic translation initiation factor 1A, Y-linked</td>
<td>Y</td>
</tr>
<tr>
<td>203649_s_at</td>
<td></td>
<td>Phospholipase A2, group IIA (platelets, synovial fluid)</td>
<td></td>
</tr>
<tr>
<td>204410_at</td>
<td></td>
<td>Eukaryotic translation initiation factor 1A, Y-linked</td>
<td>Y</td>
</tr>
<tr>
<td>205000_at</td>
<td>205000_at</td>
<td>DEAD (Asp-Glu-Ala-Asp) box polypeptide 3, Y-linked</td>
<td>Y</td>
</tr>
<tr>
<td>205001_s_at</td>
<td></td>
<td>DEAD (Asp-Glu-Ala-Asp) box polypeptide 3, Y-linked</td>
<td>Y</td>
</tr>
<tr>
<td>206700_s_at</td>
<td></td>
<td>Lysine (K)-specific demethylase 5D</td>
<td>Y</td>
</tr>
<tr>
<td>214131_at</td>
<td></td>
<td>Chromosome Y open reading frame 15B</td>
<td>Y</td>
</tr>
<tr>
<td>214218_s_at</td>
<td>214218_s_at</td>
<td>X (inactive)-specific transcript (non-protein coding)</td>
<td>X</td>
</tr>
<tr>
<td>221728_x_at</td>
<td>221728_x_at</td>
<td>X (inactive)-specific transcript (non-protein coding)</td>
<td>X</td>
</tr>
</tbody>
</table>
Table 5. Differentially expressed probes identified in GSE 23343 using the Limma Package compared to those found using the T-test method, common probes are highlighted.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>Gene Name</th>
<th>Y or X-linked</th>
</tr>
</thead>
<tbody>
<tr>
<td>201909_at</td>
<td>201909_at</td>
<td>Ribosomal protein S4, Y-linked 1</td>
<td>Y</td>
</tr>
<tr>
<td>204409_s_at</td>
<td>204409_s_at</td>
<td>Eukaryotic translation initiation factor 1A, Y-linked</td>
<td>Y</td>
</tr>
<tr>
<td>204410_at</td>
<td>204410_at</td>
<td>Eukaryotic translation initiation factor 1A, Y-linked</td>
<td>Y</td>
</tr>
<tr>
<td>205000_at</td>
<td>205000_at</td>
<td>DEAD (Asp-Glu-Ala-Asp) box polypeptide 3, Y-linked</td>
<td>Y</td>
</tr>
<tr>
<td>205001_s_at</td>
<td>205001_s_at</td>
<td>DEAD (Asp-Glu-Ala-Asp) box polypeptide 3, Y-linked</td>
<td>Y</td>
</tr>
<tr>
<td>206700_s_at</td>
<td>206700_s_at</td>
<td>Lysine (K)-specific demethylase 5D</td>
<td>Y</td>
</tr>
<tr>
<td>207063_at</td>
<td>207063_at</td>
<td>Chromosome Y open reading frame 14</td>
<td>Y</td>
</tr>
<tr>
<td>207330_at</td>
<td>207330_at</td>
<td>Pregnancy-zone protein</td>
<td>Chr. 12</td>
</tr>
<tr>
<td>214131_at</td>
<td>214131_at</td>
<td>Chromosome Y open reading frame 15B</td>
<td>Y</td>
</tr>
<tr>
<td>214218_s_at</td>
<td>214218_s_at</td>
<td>X (inactive)-specific transcript (non-protein coding)</td>
<td>X</td>
</tr>
<tr>
<td>221728_x_at</td>
<td>221728_x_at</td>
<td>X (inactive)-specific transcript (non-protein coding)</td>
<td>X</td>
</tr>
<tr>
<td>223645_s_at</td>
<td>223645_s_at</td>
<td>Chromosome Y open reading frame 15B</td>
<td>Y</td>
</tr>
<tr>
<td>223646_s_at</td>
<td>223646_s_at</td>
<td>Chromosome Y open reading frame 15B</td>
<td>Y</td>
</tr>
<tr>
<td>224588_at</td>
<td>224588_at</td>
<td>X (inactive)-specific transcript (non-protein coding)</td>
<td>X</td>
</tr>
<tr>
<td>224590_at</td>
<td>224590_at</td>
<td>X (inactive)-specific transcript (non-protein coding)</td>
<td>X</td>
</tr>
<tr>
<td>227614_at</td>
<td>227614_at</td>
<td>Hexokinase domain containing 1</td>
<td>Chr. 10</td>
</tr>
<tr>
<td>227671_at</td>
<td>227671_at</td>
<td>X (inactive)-specific transcript (non-protein coding)</td>
<td>X</td>
</tr>
<tr>
<td>228492_at</td>
<td>228492_at</td>
<td>Ubiquitin specific peptidase 9, Y-linked</td>
<td>Y</td>
</tr>
<tr>
<td>235942_at</td>
<td>235942_at</td>
<td>Hypothetical LOC401629</td>
<td>Y</td>
</tr>
<tr>
<td>214216_s_at</td>
<td>214216_s_at</td>
<td>La ribonucleoprotein domain family, member 4B</td>
<td>Chr. 10</td>
</tr>
</tbody>
</table>
Table 6. Differentially expressed probes and their corresponding gene names, identified in merged datasets using the t-test method.

<table>
<thead>
<tr>
<th>Probe ID</th>
<th>Gene Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>201909_at</td>
<td>Ribosomal protein S4, Y-linked 1</td>
</tr>
<tr>
<td>204409_s_at</td>
<td>Eukaryotic translation initiation factor 1A, Y-linked</td>
</tr>
<tr>
<td>204410_at</td>
<td>Eukaryotic translation initiation factor 1A, Y-linked</td>
</tr>
<tr>
<td>205000_at</td>
<td>DEAD (Asp-Glu-Ala-Asp) box polypeptide 3, Y-linked</td>
</tr>
<tr>
<td>205001_s_at</td>
<td>DEAD (Asp-Glu-Ala-Asp) box polypeptide 3, Y-linked</td>
</tr>
<tr>
<td>206700_s_at</td>
<td>Lysine (K)-specific demethylase 5D</td>
</tr>
<tr>
<td>207330_at</td>
<td>Pregnancy-zone protein</td>
</tr>
<tr>
<td>214131_at</td>
<td>Chromosome Y open reading frame 15B</td>
</tr>
<tr>
<td>214218_s_at</td>
<td>X (inactive)-specific transcript (non-protein coding)</td>
</tr>
<tr>
<td>221728_x_at</td>
<td>X (inactive)-specific transcript (non-protein coding)</td>
</tr>
</tbody>
</table>
Table 7. Differentially expressed probes and their corresponding genes in dataset GSE14343 after batch effect removal using fRMA. Newly identified genes are highlighted.

<table>
<thead>
<tr>
<th>Probe ID</th>
<th>Gene Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>214131_at</td>
<td>Chromosome Y open reading frame 15B</td>
</tr>
<tr>
<td>204410_at</td>
<td>Eukaryotic translation initiation factor 1A, Y-linked</td>
</tr>
<tr>
<td>205001_s_at</td>
<td>DEAD (Asp-Glu-Ala-Asp) box polypeptide 3, Y-linked</td>
</tr>
<tr>
<td>206700_s_at</td>
<td>Lysine (K)-specific demethylase 5D</td>
</tr>
<tr>
<td>204409_s_at</td>
<td>Eukaryotic translation initiation factor 1A, Y-linked</td>
</tr>
<tr>
<td>205000_at</td>
<td>DEAD (Asp-Glu-Ala-Asp) box polypeptide 3, Y-linked</td>
</tr>
<tr>
<td>201909_at</td>
<td>Ribosomal protein S4, Y-linked 1</td>
</tr>
<tr>
<td>209739_s_at</td>
<td>patatin-like phospholipase domain containing 4</td>
</tr>
<tr>
<td>221728_x_at</td>
<td>X (inactive)-specific transcript (non-protein coding)</td>
</tr>
<tr>
<td>214218_s_at</td>
<td>X (inactive)-specific transcript (non-protein coding)</td>
</tr>
<tr>
<td>206624_at</td>
<td>ubiquitin specific peptidase 9, Y-linked</td>
</tr>
<tr>
<td>211149_at</td>
<td>ubiquitously transcribed tetratricopeptide repeat gene, Y-linked</td>
</tr>
</tbody>
</table>
Table 8. Differentially expressed probes and their corresponding genes in dataset GSE23343 after batch effect removal using fRMA.

<table>
<thead>
<tr>
<th>Probe ID</th>
<th>Gene Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>221728_x_at</td>
<td>X (inactive)-specific transcript (non-protein coding)</td>
</tr>
<tr>
<td>227671_at</td>
<td>X (inactive)-specific transcript (non-protein coding)</td>
</tr>
<tr>
<td>214218_s_at</td>
<td>X (inactive)-specific transcript (non-protein coding)</td>
</tr>
<tr>
<td>224588_at</td>
<td>X (inactive)-specific transcript (non-protein coding)</td>
</tr>
<tr>
<td>224590_at</td>
<td>X (inactive)-specific transcript (non-protein coding)</td>
</tr>
<tr>
<td>205001_s_at</td>
<td>DEAD (Asp-Glu-Ala-Asp) box polypeptide 3, Y-linked</td>
</tr>
<tr>
<td>204409_s_at</td>
<td>Eukaryotic translation initiation factor 1A, Y-linked</td>
</tr>
<tr>
<td>205000_at</td>
<td>DEAD (Asp-Glu-Ala-Asp) box polypeptide 3, Y-linked</td>
</tr>
<tr>
<td>201909_at</td>
<td>Ribosomal protein S4, Y-linked 1</td>
</tr>
</tbody>
</table>
Table 9. Differentially expressed probes and their corresponding genes in merged datasets after batch effect removal using fRMA or using ComBat, common probes are highlighted.

<table>
<thead>
<tr>
<th>Probe ID</th>
<th>Gene Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Merged fRMA</td>
<td>Merged ComBat</td>
</tr>
<tr>
<td>201909_at</td>
<td>201909_at</td>
</tr>
<tr>
<td>204409_s_at</td>
<td>204409_s_at</td>
</tr>
<tr>
<td>204410_at</td>
<td>204410_at</td>
</tr>
<tr>
<td>205000_at</td>
<td>205000_at</td>
</tr>
<tr>
<td>205001_s_at</td>
<td>205001_s_at</td>
</tr>
<tr>
<td>206700_s_at</td>
<td>206700_s_at</td>
</tr>
<tr>
<td>211149_at</td>
<td>211149_at</td>
</tr>
<tr>
<td>214131_at</td>
<td>214131_at</td>
</tr>
<tr>
<td>214216_s_at</td>
<td>214216_s_at</td>
</tr>
<tr>
<td>221728_x_at</td>
<td>221728_x_at</td>
</tr>
<tr>
<td>214218_s_at</td>
<td>X (inactive)-specific transcript (non-protein coding)</td>
</tr>
<tr>
<td>203992_s_at</td>
<td>lysine (K)-specific demethylase 6A</td>
</tr>
<tr>
<td>206624_at</td>
<td>ubiquitin specific peptidase 9, Y-linked</td>
</tr>
<tr>
<td>209739_s_at</td>
<td></td>
</tr>
</tbody>
</table>

Ribosomal protein S4, Y-linked 1
Eukaryotic translation initiation factor 1A, Y-linked
Eukaryotic translation initiation factor 1A, Y-linked
DEAD (Asp-Glu-Ala-Asp) box polypeptide 3, Y-linked
DEAD (Asp-Glu-Ala-Asp) box polypeptide 3, Y-linked
Lysine (K)-specific demethylase 5D
ubiquitously transcribed tetratricopeptide repeat gene, Y-linked
Chromosome Y open reading frame 15B
La ribonucleoprotein domain family, member 4B
X (inactive)-specific transcript (non-protein coding)
X (inactive)-specific transcript (non-protein coding)
X (inactive)-specific transcript (non-protein coding)
lysine (K)-specific demethylase 6A
ubiquitin specific peptidase 9, Y-linked
Figure 1: Hierarchical Cluster Analysis of 3 microarray datasets of human normal liver tissues: A. GSE14951 B. GSE14323 C. GSE23343
Figure 2: Differentially Expressed probes in individual datasets (t-test). Heat Maps of differentially expressed probes (t-test, P<0.05, fold change >2) between male and female samples of human normal liver: Datasets: A. GSE14951 B. GSE14323 C. GSE23343
Figure 3: Differentially Expressed probes in individual datasets (Limma). Heat maps of differentially expressed probes between male and female samples of human normal liver identified using the Limma Package (P<0.05). A. Dataset GSE14323 B. Dataset GSE23343.
Figure 4: Differentially Expressed probes in merged datasets. Heat map of differentially expressed probes between male and female samples of human normal liver tissues, identified in merged datasets (GSE14323, GSE23343) using the t-test method ((P<0.05, Fold Change>2).
Figure 5: Hierarchical Cluster Analysis of merged microarray datasets (GSE14323, GSE23343) of human normal liver tissues.
Figure 6. Differentially Expressed probes in individual datasets after batch effect removal. Heat maps of differentially expressed probes (Limma, P<0.05) between male and female human normal liver samples in microarray datasets A. GSE14323 and B. GSE23343 after removal of batch effects using the fRMA method.
Figure 7: Differentially Expressed probes in merged datasets after batch effect removal.
Heat maps of differentially expressed probes (Limma, P<0.05) between male and female human normal liver samples in merged microarray data sets (GSE14323 and GSE23343) after removal of batch effects using A. fRMA and B. ComBat methods.
Figure 8: Gene Signature Validation. Hierarchical cluster analysis for microarray datasets A. GSE14323 and B. GSE23343, using a subset of 500 genes identified previously by Hartwell et al\textsuperscript{16}.
REFERENCES


