DEVELOPMENT OF NOVEL TOOLS FOR CANCER DIAGNOSIS, PROGNOSIS AND TREATMENT USING INTRA- OR INTER-SPECIES TRANSCRIPTOME META-ANALYSIS

A DISSERTATION SUBMITTED TO
THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN
MOLECULAR BIOLOGY AND GENETICS

By
Huma Shehwana
August 2017

DEVELOPMENT OF NOVEL TOOLS FOR CANCER DIAGNOSIS, PROGNOSIS AND TREATMENT USING INTRA- OR INTER-SPECIES TRANSCRIPTOME META-ANALYSIS

By Huma Shehwana

August 2017

We certify that we have read this dissertation and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Doctor of Philosophy.

_____
Özlen Konu Karakayalı (Advisor)

_____
Murat Alper Cevher

_____
Hilal Özdağ

_____
Ali Osmay Güre

_____
Nurcan Tunçbağ

Approved for the Graduate School of Engineering and Science

_____
Ezhan Karaşan
Director of the Graduate School of Engineering and Science

Abstract

DEVELOPMENT OF NOVEL TOOLS FOR CANCER DIAGNOSIS, PROGNOSIS AND TREATMENT USING INTRA- OR INTER-SPECIES TRANSCRIPTOME META-ANALYSIS

Huma Shehwana

Ph.D. in Molecular Biology and Genetics

Advisor: Özlen Konu Karakayalı

August 2017

In the past decades, a considerable number of studies have performed meta-analysis on large data collections to prioritize sets of genes, pathways or types/categories of disease focusing on either differential expression, survival analysis, or co-expression networks. However, not many web applications or databases have been developed from these studies thus findings largely remained restricted to the addressed questions and it was not possible for other researchers to use the collected data for the evaluation of novel hypotheses. In this thesis, transcriptomic meta-analysis strategies have been applied to untangle complexities in multiple aspects of cancer research including treatment, diagnosis, and prognosis. Furthermore, three different web-tools have been developed which are not limited to a single type of meta-analysis. In this context, in addition to interesting cancer related findings, novel methodologies have been proposed and tested in the field of meta-analysis and cancer research.

First chapter of the thesis presented a general introduction on the concepts of the thesis. Second chapter focused on a pathway comparison strategy based on meta-analysis that was used to reveal concordant/discordant aspects of rapamycin-mediation on transcriptomes of zebrafish and mouse. Analysis has shown that ribosomal terms were significantly upregulated while proteasome was downregulated in both species.

Zebrafish has undergone a whole-genome duplication event; I also found out that rapamycin treatment resulted in largely concordant behavior of duplicated gene pairs. In addition, an online database, CompariZome, was developed to evaluate the duplicate zebrafish gene pairs in multiple GEO datasets in zebrafish in comparison to respective human expression datasets. In the third chapter of this thesis, I focused on identification of correlation between a trio of genes, CDH1, HNF4A, and GRHL3, using Cancer Cell Line Encyclopedia (CCLE) dataset to reveal the significance of association between these genes in different cancers, including breast and other epithelial cancers. The findings indicated correlation within the module and has demonstrated the power of meta-analysis using CCLE dataset. In the fourth chapter of this thesis, I focused on understanding the association of CHRNA5, a subunit of cholinergic receptors, with epithelial-to-mesenchymal transition (EMT) as well as epithelial differentiation, TP53 induction, and estrogen (E2) signaling with respect to breast cancer. Meta-analysis of *in-vitro* and *in-vivo* microarray expression datasets showed that CHRNA5, itself, and its positively co-expressed neighbors, were likely secondary targets of E2-signaling; overexpressed in ER- breast cancer patients; and indicators of worse prognosis. Functional annotation revealed that CHRNA5 and its co-expression network was indeed associated with proliferation related pathways. Based on meta-analysis of different cohorts processed in the study, an online database E2S (Estrogen (E2) to Survival) was developed that can facilitate user to query any gene for evaluation of E2-mediated effects, regulation by estrogen receptor (ER), prognostic importance and co-expression network along with functional annotations.

In the fifth chapter of this thesis I focused on deciphering the correlation and deregulation between a human parolog pair of genes, i.e., mineralocorticoid and glucocorticoid receptors (MR and GR, respectively) in breast cancer. Meta-analysis of a separate normal/tumor cohort revealed that both genes were downregulated in breast cancer and their expressions were highly positively correlated. However, deregulation analysis predicted that expression of MR and GR was more tightly regulated in normal breast hence its regulation might be lost with the onset of tumorigenesis. Another Shiny database, DualExpBC, was developed to evaluate differential expression of a gene in breast cancer as well as correlation and deregulation of expression between any two

input genes in the breast normal/cancer expression cohort. With this thesis, I have developed novel tools and approaches for intra- and inter-species comparative transcriptomics and meta-analysis providing potential diagnostic, prognostic and therapeutic biomarkers.

Keywords: Meta-analysis, Co-expression network, Breast cancer, E2 signaling, CHRNA5, Inter-species and Intra-species transcriptome analysis, Mineralocorticoid and Glucocorticoid receptors.

Özet

TÜR İÇİ VE TÜRLER ARASı TRANSKRİPTOM META-ANALİZİ KULLANıLARAK KANSER TEŞHİSİ, PROGNOZU VE TEDAVİSİ İÇİN YENİ ARAÇLARıN GELİŞTİRİLMESİ

Huma Shehwana
Moleküler Biyoloji ve Genetik, Doktora
Tez Danışmanı: Özlen Konu Karakayalı
Ağustos 2017

Geçtiğimiz yıllarda, ifade farklılığının analizi, hayatta kalma ya da ortak ifade olma açısından önemli gen setlerine, yolaklara ya da hastalık türü/kategorilerine ulaşabilmek için, dikkate değer miktarda çalışmada ve geniş data koleksiyonlarının meta-analizi ile gerçekleştirilmişitr. Ancak, bu çalışmalardan yeterince web temelli uygulama ya da databaz geliştirilmemiştir. Bu nedenle bu çalışmaların çoğu, belirlenmiş hipotezlerle sınırlı kaldığından toplanmış olan bu data, diğer araştırmacılar tarafından yeni hipotezlerin denenmesine imkân sağlamamaktadır. Bu tez çalışmasında transcriptomik meta-analiz stratejileri, kanser tedavisi ve teşhisi ve prognozundaki komplikasyonların çözülebilmesi için kullanılmıştır. Ayrıca, tek tip meta-analizle sınırlı olmayan üç farklı web-uygulaması da geliştirilmiştir. Bu bağlamda, kanserle ilgili ilgi çekici buluşların yanı sıra, yeni metodlar geliştirilmiş ve meta-analiz ve kanser alanlarında test edilmiştir.

Tezin ilk bölümünde tezdeki kavramlar ile ilgili genel bir giriş bulunmaktadır. Tezin ikinci bölümünde, rapamycin ile değişen fare ve zebrabalığı transkriptomlarının arasındaki türler arası benzerlik ve farklılıkları ortaya çıkartmak için meta-analiz temelli yolak karşılaştırması yapılmıştır. Analiz her iki tür için de ribozomla ilişkili terimlerin anlamlı miktarda arttığını ancak proteazome'un anlamlı miktarda azaldığını göstermiştir. Zebrabalığı tüm genom duplikasyonuna uğramış bir organizmadır; bu tezde, rapamycin uygulaması sonucunda, genel olarak paralog gen çiftlerinin benzer hareket ettikleri

sonucuna ulaşılmıştır. Buna ek olarak, pek çok GEO zebrabalığı datasetinde paralog çiftlerin davranışlarını incelemeye imkan sağlayan CompariZome isimli bir online veritabanı geliştirilmiştir. CompariZome ile, zebrabalığı datasetleri, benzeri insan datasetleri ile karşılaştırılabilmektedir. Tezin üçüncü bölümünde, Cancer Cell Line Encyclopedia (CCLE) dataseti kullanılarak CDH1, HNF4A ve GRHL3 genleri arasındaki korelasyon ve anlamlı ilişki, meme kanseri ve epitelyal kanserler için ortaya çıkartılmıştır. Bu sonuçlar, modül içinde korelasyon olduğunu göstermiş olup CCLE dataseti kullanılarak gerçekleştirilen meta-analizlerin gücünü ortaya koymuştur. Tezin dördüncü bölümünde, bir kolinerjik reseptör altbirimi olan CHRNA5'nın, epitelyal-mezenkimal geçiş (EMT) ve epitelyal değişim, TP53 induksiyonu ve östrojen (E2) sinyal yolları ile ilişkisini, meme kanserinde ortaya çıkarmak amaçlanmıştır. *İn-vivo* ve *in-vitro* datasetlerinin meta-analizi, CHRNA5 ve onunla pozitif korelasyon gösteren genlerin E2 yolağının ikincil hedefleri olduğunu, ER- meme kanseri hastalarında bu genlerin ifade miktarlarının arttığını ve bu artışın kötü prognoz belirtisi olduğunu göstermiştir. Fonksiyonel anotasyon çalışması, CHRNA5 ve onunla ilişkili protein ağının proliferasyon yolakları ile de ilişkili olduğunu ortaya çıkartmıştır. Bu çalışmada kullanılan farklı kohortların meta-analizi sayesinde E2S (Östrojenden (E2) Hayatta Kalıma) isimli bir veritabanı geliştirilmiştir. Bu veritabanı, kullanıcının herhangi bir genin E2 temelli değişimini, Östrojen Reseptörü (ER) ile regülasyonunu, prognostik önemini ve işlevsel anotasyon ile birlikte beraber ifade olduğu genler ağını analiz etmesine imkân sağlamaktadır. Beşinci bölümde, insandaki bir paralog çiftinin, yani mineralkortikoid ve glukokortikoid reseptörlerinin (MR ve GR), meme kanserindeki korelasyonu ve deregülasyonu incelenmiştir. Farklı normal/tümör datasetlerinin meta analizi ile iki genin de meme kanserinde ifade seviyesinin azaldığı ve birbirleri ile yüksek pozitif korelasyon gösterdikleri ortaya çıkarılmıştır. Ancak, deregulasyon analizi sonuçları, MR ve GR genlerinin ifadesinin normal meme dokusunda daha sıkı kontrol edildiğini göstermiştir. Bunun nedeni, regülasyonun kanserleşme başlangıcı ile kaybolması olabileceğidir. İstenilen genin meme kanseri dokusundaki ifadesi ve ilgili iki genin normal meme ve meme kanseri dokusundaki korelasyonu ve deregulasyonunu analiz etmeye imkân sağlayan, DualExpBC isimli bir diğer Shiny veritabanı geliştirilmiştir. Bu tezle, potansiyel teşhis, prognoz ve biyogösterge belirlemek için

türler arası ve tür içi transkriptom karşılaştırması ve meta-analizi yapan yeni araçlar ve metodlar geliştirilmiştir.

Anahtar sözcükler: Meta-analiz, Ortak ifade ağı, Meme Kanseri, Österojen Sinyal yolağı, CHRNA5, Tür-içi ve türler-arası transcriptom analizi, mineralokortikoid ve glukokortikoid reseptörleri

To my parents…

Acknowledgement

I would like to thank my advisor Assoc. Prof. Dr. Ozlen Konu for her assistance and guidance. Her amazing research ideas and trusting me with them have always been a motivation to work harder. I am very obliged to her for providing me the chance to work and advance on several of her own and collaborative projects.

I am deeply grateful to MBG faculty members for the research perspective they provided during my course work. Especially, I really appreciate Dr. Ozgur Sahin for enlightening me with cancer research during his lectures. Working in collaboration with Dr. Ceren Sucularli, Dr. Gulcin Cakan, Dr. Hani Alotaibi, Dr. Ozgur Sahin, Dr. Ender Avci has provided me a pleasant exposure to gain knowledge on different aspects of cancer research. Besides, I am also very grateful to Dr. Gokhan Karakulah who guided me during the learning of RNAseq data analysis. I am thankful to Merzie Razekh and Dr. Can Alkan for letting me perform heavy computational executions on their server. Dr. Hilal Ozdag and Dr. Oznur Tastan, being my thesis progress committee members, always encouraged me and provided valuable suggestions. I would also like to thank Dr. Murat Cevher, Dr. Ali Gure, and Dr. Nurcan Tuncbag for being in my thesis committee and for their helpful comments. I pay my thanks to Dr. Isik Yulug, the former head of the MBG Department, for her kind assistance and help during my PhD studies.

I am deeply thankful to Ermira Jahja for her care, sincerity, love and support in academic and personal problems during the entire PhD period. Her optimistic attitude and calm nature was always a relief during stressful times. I also owe special thanks to Ermira Jahja, Seniye Targen and Sahika Cingir for guiding me during wet-lab experiments. PhD studies at MBG gave me a chance to meet many amazing people and make new friends. It is not possible to mention all the names but I am very grateful to Zeynep Ozge Ayyildiz, Secil Demirkol, Pinar Demirayak, Sohail Ansari, Rashmi Mishra, Ayse Gokce Keskus, Basak Ozgursoy, Fatma Betul Dincaslan, Ilgin Cagnan,

Alperen Taciroglu, Said Tiryaki, Murat Yaman, Bilge abla and Yavuz abi for making my stay at MBG an interesting and memorable experience.

PhD is a long journey with many ups and downs. Friends, who supported me, will always have a very special place in my heart. I would like to express my deepest gratitude to Alia Khan, Amna Adnan, Rida Arif, Ayman Rehan for always motivating me to focus and complete my PhD on time. I am also very thankful to Mahzeb Fiaz and Muhaddisa Barat Ali for their caring attitude, sincerity and help. I deeply respect aunt Naeem Fatima Mehmud for providing a family like environment in a foreign land and I will always remember her valuable advices.

I am highly obliged to my parents who taught me to work hard for my goal and to not settle without achieving them. The trust of my father and the passion of my mother to see me as a confidant, successful and an independent person was always a major push to pursue my dreams and never quit. In spite of many problems, my parents always kept my studies as a priority and let nothing come in the way. My elder brother Khurram Yaqub and his successful PhD career was a great motivation for me to come abroad, pursue and complete my studies. I am also thankful to my brothers Shariq Yaqub, Haris Yaqub and my sister Fatima Seemab for always cheering me up. Besides, I am very grateful to my uncles Muhammad Ejaz and Muhammad Nawaz for their immense support during my entire educational career. Without having a loving, understanding and cooperative family, it was not possible for me to spend even one day of my PhD.

I am thankful to funding bodies Higher Education Commission (HEC), Pakistan, and The Scientific and Technological Research Council of Turkey (TUBITAK). TUBITAK Grant No. 103T038 allowed extraction of the microarray expression profile of rapamycin exposure in zebrafish (Chapter 2); TUBITAK Grant No. 111T316 supported the CHRNA5 related studies (Chapter 4); TUBITAK Grant No. 114S226 and COST Action BM1301 supported the MR/GR studies (Chapter 5). I have been supported in my graduate studies by Bilkent University, Department of Molecular Biology and Genetics, HEC, as well as TUBITAK grants 111S316, 114S226 and COST Action BM1301.

Contents

Abstract ..... **iii**

Özet ..... **vi**

Acknowledgement ..... **x**

Abbreviations ..... **xxv**

Chapter 1

Introduction **1**

1.1. Microarray Technology **1**

1.1.1. Types of microarray 2

1.1.2. Platforms of microarrays 2

1.2. Use of microarray studies in cancer research **5**

1.3. Analysis steps in expression analysis **7**

1.3.1. Data preprocessing 8

1.3.1.1. Quality control 8

1.3.1.2. Methods of normalization 9

1.3.2. Statistical analysis for identification of important genes 10

1.3.2.1. Differential expression analysis 10

1.3.2.2. Co-expression analysis 10

1.3.2.3. Supervised and unsupervised analysis 11

1.3.2.4. Parametric and non-parametric statistical analysis 12

1.3.2.5. Survival analysis 12

1.3.3. Annotation 15

1.3.4. Enrichment of biological knowledge 15

1.3.5. Regulatory analysis 16

1.3.5.1. Transcription factor enrichment 16

1.3.5.2. MicroRNA enrichment 17

1.3.5.3. Chromosomal enrichment 18

1.3.6. Protein-protein interactions 18

1.3.6.1. Network visualization tools 18

1.4. Meta-analysis **20**

1.4.1. Meta-analysis algorithms 22
1.4.1.1. Combining P-values 22
1.4.1.2. Effect-size based meta-analysis 23
1.4.1.3. Ranking based meta-analysis 24
1.4.1.4. Merging data 24
1.4.2. R packages for meta-analysis 24
1.4.3. Pathway enrichment using meta-analysis 25
1.4.4. Co-expression using meta-analysis 25
1.4.5. Heterogeneity analysis and publication bias 26
1.5. Shiny package for web-application development 26
1.5.1. Shiny Architecture 27
1.5.2. Sharing of web applications 27
1.5.3. Bioinformatics applications developed in Shiny 28
1.6. Sources of expression data 29
1.6.1. Gene Expression Omnibus (GEO) 29
1.6.2. COSMIC 29
1.6.3. ARRAYEXPRESS 30
1.6.4. CCLE 30
1.6.5. TCGA and Firebrowse 30
1.7. Outline of work in thesis chapters 32

Chapter 2

Comparative transcriptomics in zebrafish, mouse, and humans 34
2.1. Introduction 34
2.1.1. Model Organisms 34
2.1.2. Mouse as a model organism 34
2.1.3. Zebrafish as cancer model 35
2.1.4. Cross-species oncogenomics in zebrafish 35
2.1.5. Comparative toxicological studies 36
2.1.6. Duplicated genome in zebrafish 37
2.1.7. Computational databases/tools for inter-species comparison 38
2.1.8. Bioinformatics approaches to identify duplicated zebrafish genomes 40
2.1.9. mTOR pathway 40
2.2. Aims and rationale 42
2.3. Methodology 43

2.3.1. Rapamycin treated expression dataset collection and processing 43
2.3.2. Identification of differentially expressed probesets in zebrafish and mouse 43
2.3.3. Comparison of functional enrichment between zebrafish and mouse transcriptome 44
2.3.4. Duplicated gene pair retrieval and processing 45
2.3.5. Duplicated gene pairs analysis 45
2.3.6. Mapping of rapamycin treated zebrafish profile with duplicated pairs 46
2.3.7. Collection and processing of zebrafish datasets for CompariZome 47
2.3.8. Design of web application, CompariZome 48
2.3.9. Instruction messages 50
2.3.10. Making rds object of large files 50
2.4. Results 51
2.4.1. Rapamycin treatment in zebrafish cell line 51
2.4.2. Case Study I: Comparison of rapamycin treated expression profile among model organisms (zebrafish and mouse) 51
2.4.2.1. Rapamycin treatment expression profile in mouse 51
2.4.2.2. Meta-analysis of rapamycin treated mouse expression data 53
2.4.2.3. Rapamycin regulated KEGG pathways in zebrafish and mouse 55
2.4.2.4. Functional comparisons between zebrafish and mouse rapamycin treatment expression profiles 55
2.4.2.5. Robustness of comparative concordant signature 56
2.4.3. Case Study II: Comparison of zebrafish duplicated genes in their response to rapamycin 60
2.4.3.1. Identification of duplicated gene pairs 60
2.4.3.2. Conserved/diverged expression patterns among paralog pairs in response to rapamycin 60
2.4.4. Case Study III: Development of a database for human-zebrafish expression comparison: CompariZome 64
2.4.4.1. Data collection 64
2.4.4.2. CompariZome input interface 67
2.4.4.3. Modules in Comparizome 68
Module I: Orthologs 68
Module II: Zebrafish expression module 69
Module III: Human expression 70
Module IV: Gene ontology enrichment 72

Module V: MicroRNA regulation 72
2.5. Discussion 73
2.6. Future Perspectives 77

Chapter 3

Correlation analysis of GRHL3, HNF4A and CDH1 in CCLE datasets 79
3.1. Introduction 79
3.1.1. Use of CCLE in cancer research 79
3.1.2. Genes evaluated in the study 81
3.1.2.1. Grainy head-like 3 (GRHL3) 81
3.1.2.2. Hepatocyte nuclear factor 4A (HNF4A) 82
3.1.2.3. Cadherin 1 (CDH1) 82
3.1.3. EMT and MET in cancer 83
3.2. Aims and rationale 85
3.3. Materials and methods 86
3.4. Results 87
3.4.1. Correlation between GRHL3, HNF4A and CDH1 87
3.4.2. Meta-correlation analysis in cancers originating form different tissues 87
3.5. Discussion 94
3.6. Future perspective 95

Chapter 4

Identification of CHRNA5 as a breast cancer biomarker using meta-analysis and development of E2S (Estrogen (E2) to Survival) database 96
4.1. Introduction 96
4.1.1. Hierarchy and development of breast tissue 96
4.1.2. Differentiation in cancer 97
4.1.3. Involvement of different receptors in breast cancer 97
4.1.4. Estrogen signaling in breast cancer 99
4.1.5. ER-antagonists in breast cancer treatment 100
4.1.6. Estrogens, ER, and meta-analysis studies in the literature 101
4.1.7. Cholinergic receptors 102
4.1.7.1. Cholinergic signaling in cancer 103
4.1.7.2. CHRNA5 in cancer 104
4.1.8. Genes important in mammary differentiation and cancer cell morphology 104

4.2. Aims and rationale 107
4.3. Materials and methods 109
4.3.1. CHRNA5 depletion expression data analysis (GSE89333) 109
4.3.2. Cellular differentiation expression data collection and processing 109
4.3.3. CEMA (Core Expression Module Analysis) data analysis 110
4.3.4. Mapping of TP53 and EMT markers onto CHRNA5 depletion profile 110
4.3.5. CCLE (Cancer Cell Line Encyclopedia) data analysis for CHRNA5 expression ... 111
4.3.6. Chromosome enrichment 111
4.3.7. Primer design and primer efficiency 111
4.3.8. RT-qPCR and expression analysis 112
4.3.9. Collection and preprocessing of E2 treated MCF7 datasets 113
4.3.10. Collection and pre-processing of *ER cohort* 114
4.3.11. Collection and pre-processing of *survival cohort* 115
4.3.12. Meta-analyses of *E2* and *ER cohorts* 116
4.3.13. Meta-analysis of *survival cohort* 116
4.3.14. CHRNA5 correlation analysis 121
4.3.15. CHRNA5 co-expression network 121
4.3.16. Secondary E2 target prediction 122
4.3.17. ER-target mapping to meta-analysis scores 123
4.3.18. COSMIC cell lines and TCGA data analysis 123
4.3.19. Shiny web-application: E2S 123
4.3.19.1. Data collection 123
4.3.19.2. Ready to use results and run-time processing 125
4.3.19.3. Graphical presentation 125
4.4. Results 127
4.4.1. Evaluation of epithelial to mesenchymal transition (EMT) in CHRNA5 depleted MCF7 cells 127
4.4.2. The association between CHRNA5 RNAi and differentiation expression profiles. 129
4.4.2.1. Epithelial differentiation datasets and association with CHRNA5 RNAi 131
4.4.2.2. Comparison of non-epithelial differentiation with CHRNA5 RNAi signature 138
4.4.2.3. Identification of well-known differentiation markers in CHRNA5 depleted transcriptome 142
4.4.2.4. Meta-analysis of epithelial differentiation and comparison with CHRNA5 depletion profile 144

4.4.2.5. A shared interaction network of epithelial differentiation and CHRNA5 depletion 146
4.4.2.6. Validation of differentiation signature using CEMA cell states expression profile 152
4.4.2.7. Evaluation of luminal and myoepithelial markers in CHRNA5 depleted MCF7 cells 153
4.4.3. Evaluation of TP53 signaling in CHRNA5 RNAi mediated effect 157
4.4.4. Comparison of CHRNA5 co-expression profile from CCLE breast cell lines with CHRNA5 depletion profile 159
4.4.4.1. Functional analysis of CCLE vs CHRNA5 RNA expression signature 161
4.4.5. CHRNA5 and its neighboring genes exhibit high correlation with each other 165
4.4.6. CHRNA5 and E2-mediated signaling 168
4.4.7. CHRNA5 depletion profile in comparison with E2 treatment and starvation profiles 168
4.4.8. Effects of time-dependent E2 treatment on CHRNA5 expression 172
4.4.9. Prediction of secondary E2 targets 175
4.4.10. Correlation between CHRNA5 and Estrogen Receptor (ER) expressions 179
4.4.11. Meta-analysis of CHRNA5 expression in ER dependent breast cancer subtypes . 180
4.4.12. Prognostic importance of CHRNA5 187
4.4.13. Mapping of genome-wide CHRNA5 correlation profile with effect size of *ER cohort* and hazard ratio of *untreated survival cohort* 189
4.4.14. Evaluation of secondary E2-induced targets in meta-analysis results. 192
4.4.15. CHRNA5 co-expression network 193
4.4.16. Development of an online tool for E2-mediated signaling in breast cancer 202
4.4.16.1. Modules in E2S 202
Module I: E2 Regulation 202
Module II: ER regulation 203
Module III: Prognosis 204
Module IV: Signaling and network analysis 204
4.4.16.2. Features of E2S 205
1) Multiple probesets for one gene 205
2) Download data and plots 206
3) Visual graphics 206
4) Heterogeneity analysis 206

5) Error messages 207
6) Wait messages 207
7) Forest plot visualization 207
4.4.16.3. E2S Home page 208
4.5. Discussion 210
4.6. Future perspectives 216

Chapter 5

Concordant/discordant transcriptomics of mineralocorticoid and glucocorticoid receptors in breast cancer and development of DualExpBC 218
5.1. Introduction 218
5.1.1. Nuclear Receptors 218
5.1.2. Steroid receptors 219
5.1.2.1. Steroid receptors in breast cancer 219
5.1.2.2. Mineralocorticoid (MR) and Glucocorticoid (GR) receptors 221
5.1.2.3. Regulation of MR in cancer 221
5.1.2.4. Glucocorticoid receptor in cancer 223
5.1.3. Paired and unpaired expression data analysis in cancer 224
5.1.4. Deregulation of gene pairs in cancer 225
5.2. Aims and rationale 226
5.3. Materials and methods 227
5.3.1. Breast normal/tumor expression cohort collection 227
5.3.2. Splitting in paired and unpaired cohorts 227
5.3.3. Meta-differential expression analysis 230
5.3.4. Correlation analysis of normal/tumor cohort 230
5.3.5. MR and GR deregulation 230
5.3.6. DualExpBC 230
5.3.6.1. Search from stored data 231
5.3.6.2. Run-time executions 231
5.4. Results 233
5.4.1. Meta-analysis of MR and GR differential expression in breast normal and cancer tissues 233
5.4.2. Correlation between paired and unpaired breast cancer subcohorts 236
5.4.3. Correlation between MR and GR expression in breast normal/tumor cohort 241
5.4.4. Correlation of MR and GR co-expression profiles 241

5.4.5. Prognositic importance of MR and GR in breast cancer 242
5.4.6. Differential co-expression of MR and GR in breast cancer..... 243
5.4.7. MR and GR deregulation in breast cancer..... 248
5.4.8. Development of a web tool for normal/tumor expression comparison 249
5.5. Discussion..... 251
5.6. Future perspectives..... 254
Appendix I..... 256
Appendix II 262
Appendix III..... 270
Appendix IV..... 281
REFERENCES 286
Appendix V..... 320

List of Figures

Figure 1.1: Sample preparation for microarray and processing of Affymetrix gene-chip [4]. 4
Figure 1.2: Distribution of pubmed results returned by query term 'cancer gene expression signature' in range of 1982-2012 [27]. 6
Figure 1.3: Guideline for microarray analysis [37]. 7
Figure 1.4: Workflow of WGCNA [58]. 11
Figure 1.5: Statistics of different meta-analysis used in different research papers [121]. 22
Figure 2.1: Comparison of paralog pairs in rapamycin treated zebrafish profile. 61
Figure 2.2: Home page/Input interface of CompariZome database. 68
Figure 2.3: Orthologs module showing input human gene and corresponding zebrafish ortholog genes. 69
Figure 2.4: IRS2 expression analysis in human rapamycin treated breast cancer dataset. 71
Figure 3.1: Meta-correlation between GRHL3 and CDH1 in CCLE data (R.mean=0.31, p=5.71E=08). 90
Figure 3.2: Meta-correlation between HNF4A and CDH1 in CCLE data (R.mean=0.17, p=4.03E=06). 91
Figure 3.3: Meta-correlation between HNF4A and GRHL3 in CCLE data (R.mean=0.13, p=3.56E=05). 92
Figure 3.4: GRHL3, CDH1 and HNF4A scatter-plot in CCLE cell lines 93
Figure 4.1: Cellular hierarchy in breast development and tumorigenesis [305]. 98
Figure 4.2: Regulatory mechanisms of different nuclear receptors in breast cancer [311]. 99
Figure 4.3: PRISMA guideline for E2 cohort meta-analysis. 118
Figure 4.4: PRISMA guideline for ER cohort meta-analysis. 119
Figure 4.5: PRISMA guideline for *untreated survival cohort* meta-analysis. 120
Figure 4.6: PRISMA guidelines for tamoxifen survival cohort meta-analysis. 124
Figure 4.7: Analysis of epithelial to mesenchymal transition in CHRNA5 depleted MCF7 cells. 128
Figure 4.8: Comparison of CHRNA5 RNAi expression profile with epithelial differentiation expression cohort. 136
Figure 4.9: Association of CRHNA5 depletion expression profile with lung differentiation study. 137

Figure 4.10: Comparison of CHRNA5 RNAi expression profile with non-epithelial differentiation expression cohort. 141
Figure 4.11: Meta-differential expression of epithelial datasets and its comparison with CHRNA5 depletion profile. 146
Figure 4.12: Interaction network of shared epithelial differentiation and siRNA-1 induced profile. 148
Figure 4.13: Regulation of autophagy machinery genes in CHRNA5 depletion profile. 149
Figure 4.14: Evaluation of differentiation signature in CEMA data. 153
Figure 4.15: Evaluation of luminal and myoepithelial marker genes in CHRNA5 depleted MCF7 cells. 155
Figure 4.16: TP53 genes regulation in CHRNA5 depleted cells. 158
Figure 4.17: EMT and differentiation biomarkers eexpression in three cell lines. 159
Figure 4.18: CCLE breast cell lines association with CRHNA5 depletion profile. 161
Figure 4.19: Graphical visualization of chromosome 15 highlighting CHRNA5 location. 165
Figure 4.20: CHRNA5 probeset (206533_AT) in E2 treatment and starvation profile (GSE4668). 170
Figure 4.21*:* Comparison of CHRNA5 depletion datasets with E2 treatment and hormone starvation profile. 171
Figure 4.22: Meta-analysis of CHRNA5 expression in E2 treated MCF7 expression cohort. . 175
Figure 4.23: CHRNA5 expression in E2 and cycloheximide treated MCF7 cells (GSE8597). 176
Figure 4.24: Mapping of PS-score on ER-bound gene list (retrieved from Jia et. al. [389]). ... 178
Figure 4.25: In-vitro and in-vivo correlation between expression of CHRNA5 and ESR1 179
Figure 4.26: Meta-analysis of ESR1 expression in E2 treated MCF7 expression datasets. 180
Figure 4.27: Funnel plot of CRHNA5 effect-size in ER-cohort (31 datasets). 184
Figure 4.28: Forest plot of meta-differential expression analysis in ER-/ER+ groups of ER cohort. 185
Figure 4.29: Meta-correlation analysis of CHRNA5 and ESR1 probesets in ER cohort. 186
Figure 4.30: Meta-survival analysis results of CHRNA5 probeset in survival cohort. 188
Figure 4.31: Separate meta-survival analysis of ER- and ER+ breast cancer patients from untreated survival cohort. 189
Figure 4.32: Comparison of CHRNA5 meta-correlation profile with ER-/ER+ effect-size and meta-hazard ratio. 191
Figure 4.33: Comparison of secondary E2-induced targets with different meta-analysis score. 193

Figure 4.34: CHRNA5 co-expression module retrieved from WGCNA. 196
Figure 4.35: Interaction network of yellow module. 198
Figure 4.36: Interaction network of brown module..... 199
Figure 4.37: Distribution of effect-size (ER-cohort) and CHRNA5 meta-correlation score in ER bound target genes retrieved from [389]..... 200
Figure 4.38: CHRNA5 co-expressed modules enrichment with calculated statistical scores... 201
Figure 4.39: Home page of E2S. 209
Figure 5.1: PRISMA guideline for paired/unpaired breast normal/tumor cohort meta-analysis. 229
Figure 5.2: Differential expression of MR in breast normal/tumor meta-analysis..... 234
Figure 5.3: Differential expression of GR in breast normal tumor meta-analysis. 234
Figure 5.4: MR /GR meta-differential expression analysis in paired and unpaired sub-cohort.236
Figure 5.5: Correlation between paired and unpaired effect-size obtained from respective meta-analysis..... 237
Figure 5.6: Distribution of shared and paired/unpaired specific significantly regulated genes (FDR<0.1). 238
Figure 5.7: Correlation between MR and GR expression and co-expression profiles. 242
Figure 5.8: Survival analysis of MR and GR expression in breast cancer patients (KM plotter database). 243
Figure 5.9: Distribution of shared MR/GR and MR/GR specific correlation profile. 244
Figure 5.10: MR/GR deregulation in normal/tumor samples of breast cancer. 248

List of Tables

Table 1.1: List of paramteric and nonparametric tests [72]. 13
Table 1.2: List of protein-protein interaction databases [114]. 19
Table 1.3: Amount of data available in GEO database (according to July, 2017). 29
Table 1.4: Summary Statistics of COSMIC data resource. 31
Table 2.1: Mouse datasets used in meta-analysis of rapamycin treated expression profile. 52
Table 2.2: Correlation among logFC of rapamycin treated mouse expression profiles. 54
Table 2.3: Number of differentially expressed genes according to different meta-analysis threshold 54
Table 2.4: Contingency table and association statistics of significantly regulated ontology terms ($p<0.05$) 57
Table 2.5: Validation of robust rapamycin regulated profile in zebrafish and mouse using different meta-analysis thresholds. 59
Table 2.6: Association statistics of significantly regulated paralog pairs (paralog $p<0.1$). 62
Table 2.7: Duplicated gene pairs regulated in same direction in rapamycin treated zebrafish expression profile. 62
Table 2.8: Duplicated gene pairs regulated in opposite direction in rapamycin treated zebrafish expression profile. 64
Table 2.9: Datasets included in CompariZome_V1 (version 1). 65
Table 3.1: Correlation between GRHL3, HNF4A and CDH1 in CCLE tissues. 88
Table 4.1: Primer sequences used for RT-qPCR in CHRNA5 depleted MCF7, BT20 and MDA-MB-231 samples. 112
Table 4.2: Differentiation datasets used for the comparative analysis. 130
Table 4.3: Differentiation regulators along with their expression in CHRNA5 depleted MCF7 cells and normal breast differentiation (limma results). 143
Table 4.4: Selected differentiation related genes (Top 10 with known literature) induced by CHRNA5 depletion (Figure 4.12). 151
Table 4.5: Regulation of luminal and myoepithelial genes in MCF7 CHRNA5 RNAi microarray analysis (limma results). 156
Table 4.6: Two-way ANOVA results of RT-qPCR evaluated genes using cell line and siRNA-1 treatment (Figure 4.16). 159

Table 4.7: Functional annotation of shared significant CHRNA5 siRNA-1 and CCLE expression profile (using KEGG). 163
Table 4.8: Chromosome enrichment (Top 5) of CHRNA5 modulated genes (depleted by the siRNA-1 in MCF7 and positively correlated with CHRNA5 in CCLE). 166
Table 4.9: CRHNA5 co-expressed genes enriched on chromosome 15 (DAVID version 6.8). 167
Table 4.10: Datasets used in meta-analysis of E2 cohort (MCF7 cell line data)..... 174
Table 4.11: Description of datasets used in ER cohort..... 182
Table 4.12: Datasets of untreated breast cancer patient expression data used in survival cohort analysis (KM plotter, 2010). 187
Table 4.13: Contingency table and association statistics of CHRNA5 meta-correlation profile. 192
Table 4.14: KEGG pathways of the CHRNA5 co-expression modules (Figure 4.33)..... 195
Table 5.1 : Gene expression datasets included for meta-analysis. 228
Table 5.2: Upregulated Gene Ontology (biological processes) terms specific to unpaired and paired sub-cohort. 239
Table 5.3: Downregulated Gene Ontology (biological processes) terms specific to unpaired and paired sub-cohort. 240
Table 5.4: Top 10 significantly regulated gene ontology terms (biological processes) enriched in MR/GR shared co-expression profile. 245
Table 5.5: Top 10 significantly regulated gene ontology terms (biological processes) enriched in MR specific co-expression profile. 246
Table 5.6: Top 10 singificantly regulated gene ontology terms (biological processes) enriched in GR specific co-expression profile..... 247

Abbreviations

ANOVA	Analysis of Variance
BP	Biological Processes
CC	Cellular Component
CCLE	Cancer Cell Line Encyclopedia
ccRCC	clear cell Renal Cell Carcinoma
CEMA	Core Expression Module Analysis
CINSARC	Complexity Index in Sarcomas
CompariZome	Comparison of Zebrafish transcriptome
COSMIC	Catalogue Of Somatic Mutations In Cancer
DPES	Duplicated Pair Effect-Size
DPP	Duplicated Pair Pvalue
DSL	DerSimonian-Laird
DualExpBC	Dual gene Expression in Breast Cancer
E2	Estrogen/Estradiol
E2S	Estrogen (E2) to Survival
EMT	Epithelial to Mesenchymal Transition
ER	Estrogen Receptor
ERE	Estrogen Response Element
ERE	Estrogen Response Element
FDR	False Discovery Rate
FEM	Fixed Effect Model
GEMS	Gene Expression Meta-Signature
GEO	Gene Expression Omnibus
GIST	Gastrointestinal Stromal Tumors
GO	Gene Ontology
GR	Glucocorticoid Receptor
GSEA	Gene set enrichment analysis
hESC	Human Embryonic Stem Cells
HNSCCs	Head andNeck SquamousCcell Carcinomas
MEA	Modular enrichment analysis
MET	Mesenchymal to epithelial transition
MF	Molecular Function

mHR	Meta-hazard ratio
MIAME	Minimum Information About a Microarray Experiment
MM	MisMatch
MR	Mineralocorticoid Receptor
N	Number of samples
nACHR	Nicotinic Acetyl Choline Receptors
NCBI	National Center for Biotechnology Information
NR	Nuclear Receptor
NUSE	Normalized Unscaled Standard Error
OR	Odds Ratio
PM	Perfect Match
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PS score	Predicted Secondary Score
PSC	Pluripotent stem cells
Qpvalue	Pvalue of Cochran's Q statistics
R	Correlation Coefficient
R	Correlation Coefficient
R.mean	Meta-correlation score
REM	Random Effect Model
RLE	Relative Log Expression
RMA	Robust Multichip Average
SAM	Significance analysis of microarray data
SEA	Singular enrichment analysis
SMD	Standardized Mean Difference
TCGA	The Cancer Genome Atlas
TF	Transcription Factor
UI	User Interface

Chapter 1

Introduction

In this main introduction chapter of my thesis I will give general introductory concepts which were used in the subsequent chapters dealing with inter- and intra-species meta-analysis and comparative transcriptomics. The aim of this introduction chapter is to provide the background in expression data, mainly microarray expression datasets and repositories, as well as statistical methods of co-expression, network analyses and meta-analysis methodologies. At the end of the chapter, brief introduction of subsequent chapters will also be summarized.

1.1. Microarray Technology

Advent of microarray technology is an important step in transition of molecular biology into the post-genomic era. Microarray technology superseded the trend of observing single gene expression and allowed evaluation of thousands of genes at one time [1]. Popularity of microarrays gave rise to huge amount of high throughput transcriptomics data on different platforms [2].

Microarray technology works on the principle of hybridization based on the use of complementary genetic code between a probe and a target DNA. Similar to a high throughput dot-blot system, known DNA fragments of particular length are attached on the chip and fluorescently labeled product (amplified RNA/cDNA) is used as a target for hybridization [3, 4]. Scanners with high resolution are used to access the signal produced by binding of probes to target RNA/cDNA. These signals are recorded in pixels and ultimately provide the expression estimate of a particular gene/probe in a sample [5].

1.1.1. Types of microarray

The fame of microarray technology sequentially gave rise to many variants and platforms for microarray studies. Two main types of microarrays include cDNA [6] and oligonucleotide microarrays [7], which deviate on the basis of the number and types of probes present on a chip. Immobilization of longer DNA fragments (500-5000bp) or shorter probes (20-25mer) distinguishes cDNA and oligonucleotide microarrays, respectively [3, 8]. cDNA microarrays can be used for custom (small size) or large-scale screening of selected libraries [9] while oligonucleotide microarrays allow for the evaluation of gene mutation, gene expression, as well as alternative transcripts [3, 10-12]. Oligonucleotide probesets are either synthesized off-chip by using traditional procedures and are later immobilized on the chip; or another advanced procedure is the light directed on-chip synthesis, as in the case of Affymetrix arrays [13].

Second way of microarray categorization is based on the number of samples observed on a chip. Single channel or dual channel microarrays depend on one-color or two-color detection of fluorescent signal, respectively. As the name implies, dual channel microarrays utilize two samples (one sample form each group) and label each of them with either Cy3 (570nm; green dye) or Cy5 (670nm; red dye) [14, 15]. Both samples after fluorescent labeling are hybridized on to a single array and the intensity of Cy3/Cy5 ratio provides an estimate of relative abundance of hybridized material (RNA/cDNA) onto the same spot [16]. On the other hand, single channel microarrays are only labeled with Cy3 or in the case of Affymetrix biotin [17]. Studies have revealed that dual channel microarrays may not indicate the absolute gene expression but rather provide a relative abundance between two samples and is comparative [15].

1.1.2. Platforms of microarrays

Onset of high-throughput gene expression technologies in mid 1990s have given rise to a wide variety of commercial and customized arrays [18]. Three different kinds of experimental platforms exist, including spotted cDNA arrays, spotted oligonucleotide arrays and Affymetrix gene chips. Longer cDNA fragments are generally less used due

to the lack of specificity and no control over clone nucleotide sequence [4, 19]. Spotted oligonucleotides (60-70-mers) are synthetic and the process is largely automated however higher costs and exhaustible resources have been the main disadvantages [4]. Yet, a study has shown good correlation between expression values of spotted oligonucleotides and 25-mer probesets [20]. Affymetrix gene-chip involves light directed synthesis of 25-mers series (11 probes) [21]. These 25-mers are designed through software programs using 3' end of transcripts/predicted transcripts. In addition, another set of 11 probes (25-mers) with a mutation/mismatch in the center are also generated to control hybridization sensitivity, i.e., mismatch probes [22]. Signal from all 22 probes (11 perfect match and 11 mismatch) is recorded and summarized in a single value using different probeset algorithms [4]. Sample preparation followed by processing of Affymetrix gene chip are shown in Figure 1.1.

Platforms can be selected on the basis of study focus and generally vary depending on number of genes on the chip, repertoire of genes of interest, costs spared for microarray experiment and accessibility [5]. Extensive use of microarray has also given rise to customized platforms [23]. Comparisons between customized microarray with commercial platforms (i.e. Affymetrix and Agilent platforms) revealed that commercial platforms showed higher correlation among each other than microarray data obtained from customized microarray [24]. Affymetrix [7] is one of the pioneer commercial microarray companies holding attention of large number of microarray research while Agilent [25] joined the market later and both platforms are well-known in the area [26].

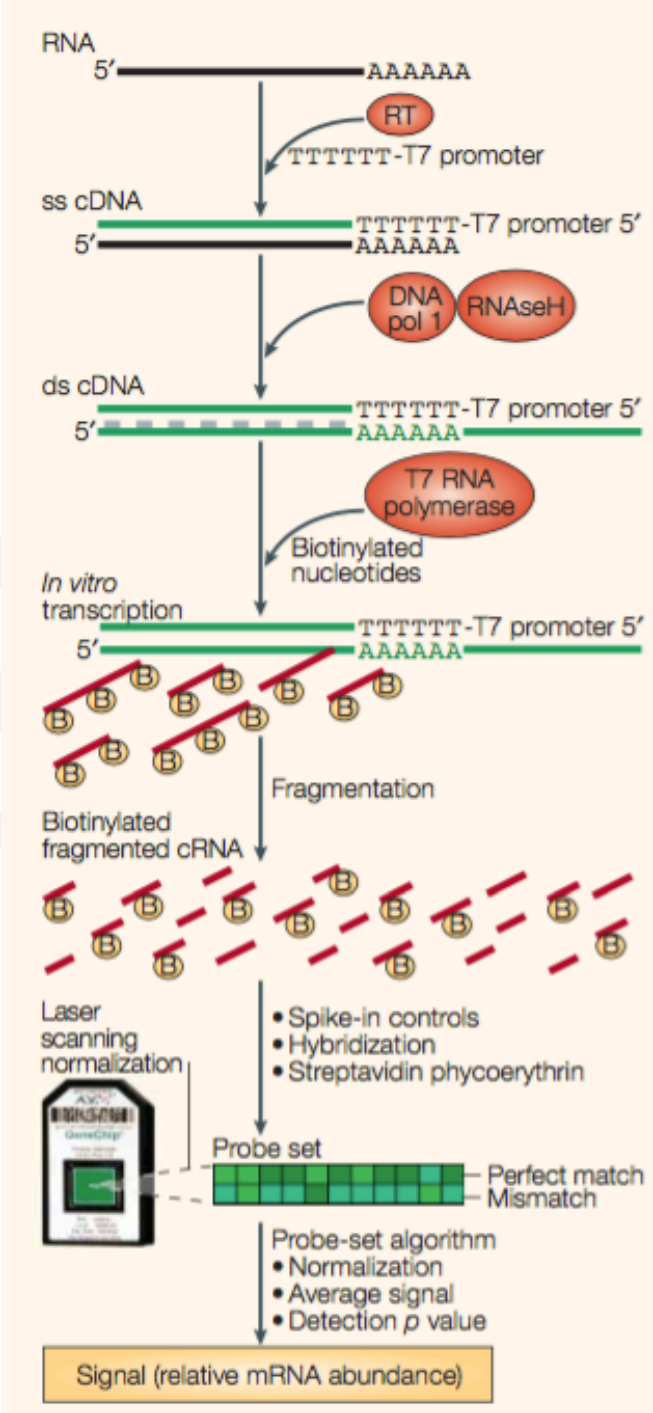


Figure 1.1: Sample preparation for microarray and processing of Affymetrix gene-chip [4].

The figure is reprinted from *Tumor analysis best practices working group, 2004* [4] and copyright permission can be found in Appendix V.

1.2. Use of microarray studies in cancer research

A review evaluated the statistics of cancer microarray studies between 1998-2012 and showed that over the years, cancer research papers based on microarray data have remarkably increased in the Pubmed database (Figure 1.2) [27]. Analysis of cancer microarray expression datasets thus facilitated gene prioritization and categorization of cancer types. For instance, Dawany et al. have collected and integrated microarray expression data of 13 cancer types to identify a list of shared genes and pathways in multiple cancer types [28]. Another microarray-based study has presented an expression signature of 70 genes, which can classify young breast cancer patients in prognostically different classes, i.e., high risk and low risk groups [29]. Furthermore, using primary head and neck squamous cell carcinomas (HNSCCs) expression profile, 102 selected genes signature was presented for prediction of local lymph node metastasis [30]. This signature predicted HNSCCs with 86% efficacy outperforming the existing diagnostic efficacy of 68% true prediction [30].

Microarray data were not only used to identify important biomarkers but also to predict primary source of tumors with unknown origins [31]. Multiple microarray datasets have helped differentiation between different groups/classes of tumors and also for the identification of novel subclasses. For example, gene expression analysis has added cutaneous melanoma to the subtypes of melanoma and also revealed that cutaneous melanoma subtype was comparatively the less invasive subtype of melanoma [32]. Similarly, a signature of 496 genes has classified breast cancer into four major breast cancer subtypes including basal-like, Erbb2, luminal like and normal-like [33]. In 2009, a 67 gene-signature called complexity index in sarcomas (CINSARC) was used for prediction of metastasis in sarcoma. CINSARC signature was also successfully validated in gastrointestinal stromal tumors (GISTs), breast carcinomas and lymphomas [34]. Biological significance of these 67 signature genes revealed involvement of mitosis and chromosomal abnormalities related genes [35].

Another interesting study focused on hepatocellular carcinoma, showing that not only tumor cells, but also other factors like normal hepatic tissue around tumor stroma can

play a significant role in cancer development [36]. Gene signature associated with tumor recurrence was obtained using machine learning approaches and results showed that good prognostic markers were indeed involved in regular tissue-specific functions of hepatocytes while bad prognostic marker genes were mostly related to inflammation [36]. Above-mentioned studies are just a few examples indicating widespread use of microarray technology in cancer research.

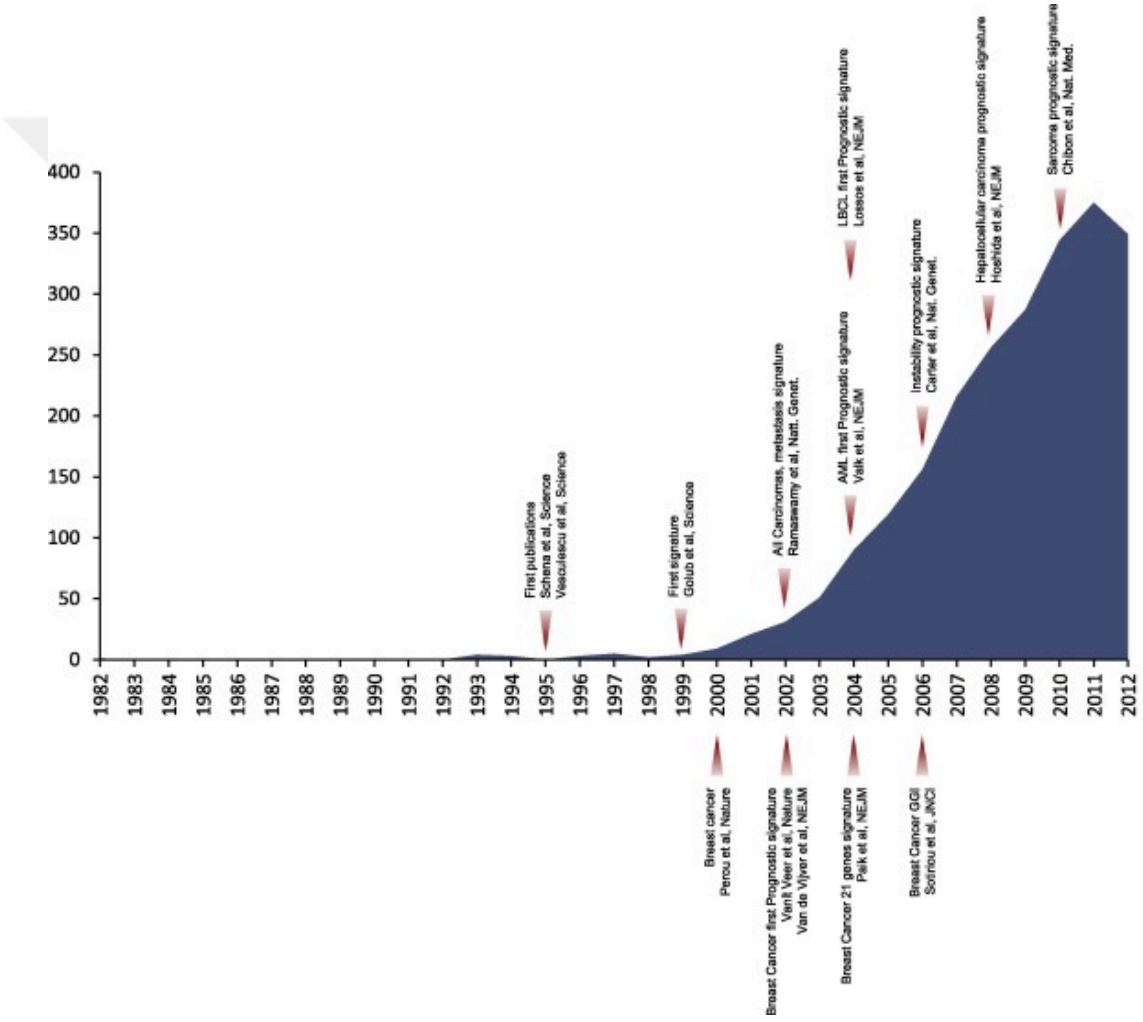


Figure 1.2: Distribution of pubmed results returned by query term 'cancer gene expression signature' in range of 1982-2012 [27].

Major research in some of the years are highlighted by arrows [27]. This figure is reprinted from *Chibon et al., 2013* [27] and copyright permission can be found in Appendix V.

1.3. Analysis steps in expression analysis

Depending on the biological question several kinds of analysis can be performed. First, data should be carefully pre-processed before statistical analysis of differentially expressed genes. An outline of workflow along with considerations is shown in Figure 1.3.

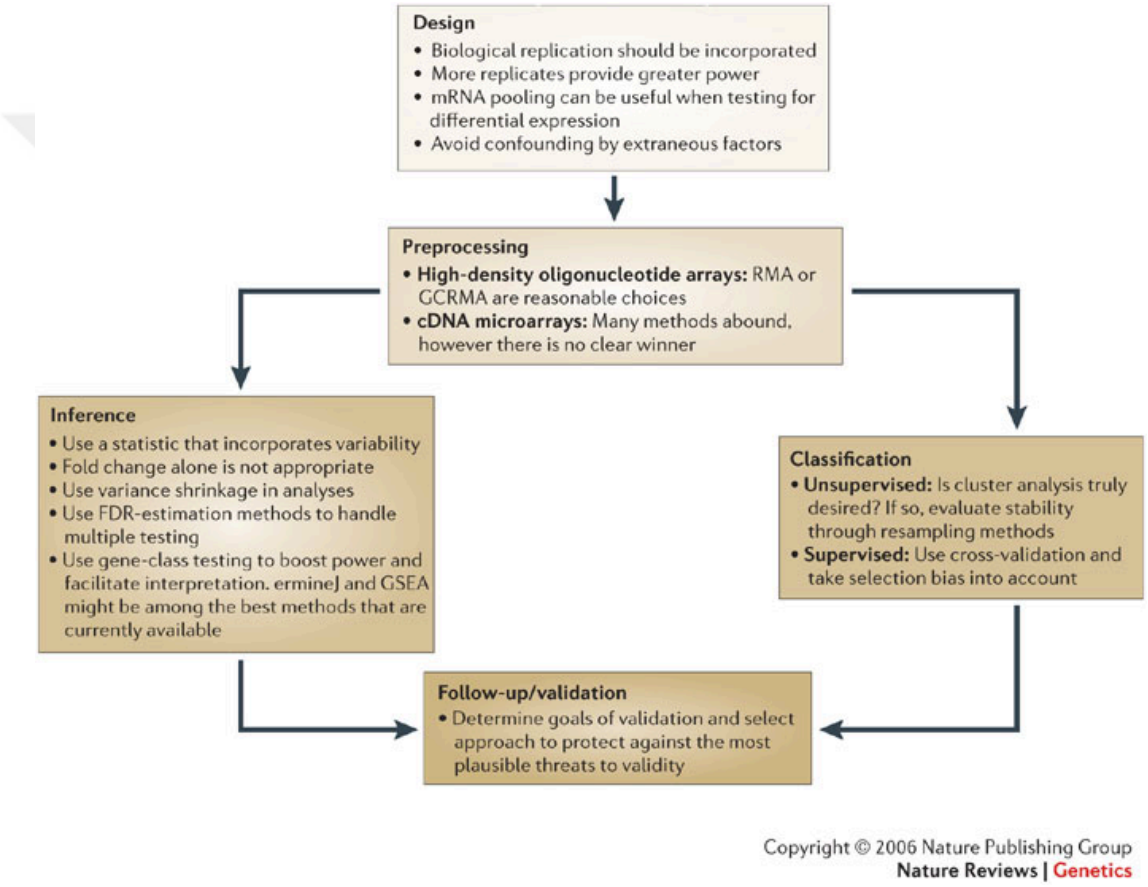


Figure 1.3: Guideline for microarray analysis [37].

This figure is reprinted from *Allison et al., 2006* [37] and copyright permission can be found in Appendix V.

1.3.1. Data preprocessing

1.3.1.1. Quality control

Before evaluation of data, processing and expression analysis, it is important to detect the quality of microarray samples [38]. Quality control is not always performed in the original publication and hence when data are reanalyzed after retrieval form public resources it is important to evaluate its quality using predefined quality scores and plots [39]. Several criteria have been suggested for separation of good and bad samples and are centered around two main points, i.e., homogeneity between samples and identification of any artifact in individual samples [39, 40]. Some of these measures, are given below:

Scaling factors are calculated by analyzing summary of MAS5.0 analysis. MAS5.0 uses a common central value to scale all analyzed arrays. Larger difference of scaling factor between arrays indicates a lack of homogeneity [41].

Average background represents pixel variation between probes on a chip and suggested acceptable range is 20-100 [42]. Average background measures hybridization quality by detecting number of spots which are expressed more than the background noise [43].

Percent present, as the name indicates, is the ratio of probesets detected as present by MAS5.0 to the number of total probesets on array. Replicates are expected to have similar percent present values [42]. If average background, scaling factor and percent present are considered together, large differences of quality measures between samples can be considered as an underlying problem [41].

Control probes are evaluated for RNA quality analysis. Affymetrix chips have special probesets, which are designed to bind either end of certain known transcripts including β-actin and GAPDH [42]. RNA degradation generally starts form 5’ end so the ratio between 3’ and 5’ probesets [42] can be a good measure to evaluate quality of RNA [41].

Several R packages are available for quality control which include *affyplm* [44], *simpleAffy* [41], *affyQCReport* [45], *arrayQualityMetrics* [46] and *yaqcaffy* [47]. *Yaqcaffy* package, in addition to output of quality scores, facilitates outlier prediction by using mean+/-standard deviation [47]. *AffyPLM* package fits a probe level model to probe intensity values [44]. Two important quality control plots include RLE (relative log expression) and NUSE (normalized unscaled standard error), which are generated by using PLM fitted data. RLE value of a probeset is calculated by comparison between the expression value on an individual array and median expression value across all arrays [44]. NUSE, on the other hand considers standard error of each probeset and normalizes each array across others in such a way that median standard error becomes 1 [42].

1.3.1.2. Methods of normalization

After evaluating the quality of a sample, intensity values are normalized using three main steps, i.e., background signal estimation and correction, normalization and summarizing of probe values to a probeset expression value [43]. Three widely used normalization methods include RMA [48], MAS5 [49], and GCRMA [50] for Affymetrix arrays and are briefly described below:

Unlike RMA and GCRMA, MAS5 normalization method evaluates mismatch (MM) probes for background correction; and uses one-step tukey-biweight method, as the summarization method [49]. RMA, instead of using MM probeset, uses noise distribution from perfect matched (PM) probes for background correction; and performs a quantile normalization followed by summarization by median-polish [48]. GCRMA algorithm works on the same principle as RMA but also takes the GC content of probesets into account [50]. *Affy* package in R is used for normalization [51].

1.3.2. Statistical analysis for identification of important genes

1.3.2.1. Differential expression analysis

Differential expression analysis is used to find genes significantly regulated between at least two experimental conditions. Some widely used statistical methods include:

- Fold change [52]
- T-test and ANOVA (Analysis of Variance) [53, 54]
- SAM: Significance analysis of microarray data [55]
- Linear model and empirical Bayes methods [56]

Linear model and empirical Bayes methods can be implemented by using *limma* package [56]. *Limma* uses a moderated t.test by fitting a linear model on the data with an empirical Bayes test. The strategy is suitable for small sample size and is very widely used [43]. SAM is another iteration-based analysis that is applicable to small sample sizes and can also be used for RNAseq datasets [55].

1.3.2.2. Co-expression analysis

Co-expression analysis has also been used in many studies and helps identification of gene modules behaving in similar manners however it does not provide information about causal relationships at the transcript level [57]. Weighted Gene Co-expression Network Analysis (WGCNA) is widely accepted strategy and is distributed as an R package [58]. Aims of WGCN analysis are: 1) identification of highly connected hub nodes in modules which can be considered as representatives of an expression module; 2) extracting modules which are significantly associated with any observed trait; 3) comparison of modules from two or more separate networks [58]. The following figure summarizes the flow of WGCNA pipeline (Figure 1.4).

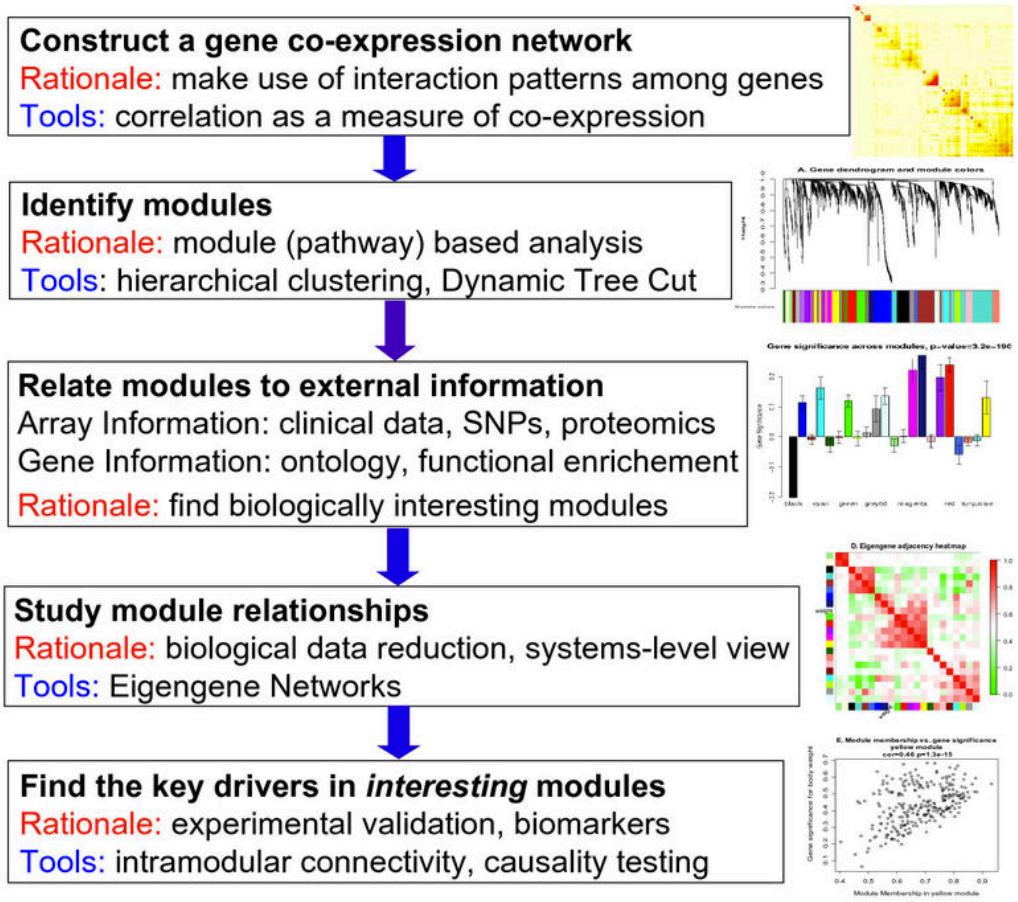


Figure 1.4: Workflow of WGCNA [58].

This figure is reprinted from *Langfelder et al, 2008* [58] and copyright permission can be found in Appendix V.

1.3.2.3. Supervised and unsupervised analysis

In case of prior hypothesis and knowledge about two or more groups, a supervised gene expression data analysis is performed to identify a gene signature, which is characteristic of a specific condition [59]. Significance of differentially expressed genes is evaluated by using parametric or non-parametric tests [60]. Analysis can be based on identification of a single gene with the nearest neighbor approach [61] or by using decision tree for multiple groups of genes [62]. In case of unknown class labels of few samples, supervised analysis can still be used by dividing data in training and test set [63]. On the other hand, if no prior information is known about class labels, unsupervised methods can facilitate the analysis [64]. Unsupervised analyses are generally used to formulate

novel hypotheses, to identify experimental artifacts or estimation of samples' relationships within data [64].

1.3.2.4. Parametric and non-parametric statistical analysis

Choice of a parametric or non-parametric test depends on normality of data. A parametric statistical test is based on the assumption that data is normal distributed while the non-parametric expression analysis is used when this assumption does not hold or when distribution of data is not known [65]. A parametric test of two groups generally relies on a t- or t-like statistics. Bayes method has been also suggested for two group comparisons [56]. On the other hand, Wilcoxon rank-sum test is an example of non-parametric test [66]. A list of commonly used parametric and non-parametric tests is given in Table 1.1. Pan et al. has shown that using Wilcoxon test in leukemia data could not result in any significant differentially expressed genes revealing a loss of power in non-parametric tests under certain conditions [67]. This might also be connected to the fact that Wilcoxon function assumes both samples in data to have the same shape of distribution function while their location parameter may change [67].

1.3.2.5. Survival analysis

Survival analysis is a well-established way to associate expression of a gene to its prognostic value in terms of a disease, e.g., cancer. Time of survival indicates the follow-up time from starting point of study to the occurrence of an observed survival related event. For instance, this time can be duration between the disease occurrence and observed event, e.g., recurrence. Handling survival data can be tricky because of the probable presence of censored data. Censored data represent ambiguous cases in which during the follow-up time, the event may not have occurred yet, or the patient might have dropped out [68]. As survival information is usually not available for all patients because of the censored data, routinely used regression analysis cannot be conducted [69]. Two notable methods to predict survival time are log-rank test [70] and Cox-proportional hazard model [71] .

Table 1.1: List of paramteric and nonparametric tests [72].

This table is reprinted from *Rosenthol et al., 2014* [72] and copyright permission can be found in Appendix V.

Outcome	Parametric	Non-Parametric	Multivariate Analysis (Example)
Continuous			
Unpaired measurement (comparison of 2 groups)	Unpaired t-test	Wilcoxon's rank-sum test (Mann-Whitney U-test)	Continuous: Linear Regression Discrete: Poisson Regression
Paired measurements (comparison of 2 groups)	Paired t-test	Wilcoxon's signed-rank test, sign test	
Comparison of >2 groups	Analysis of variance	Kruskal-Wallis test	
Categorical			
Unpaired measurements	Chi-square test	Fisher's exact test*	Binary: logistic regression
Paired measurements	Mc-Nemar's Chi-square test	Methods based on exact probabilities	
Time to event	Log-rank test		Regression analysis of survival/time-to-event data (e.g., Cox or Poisson regression or parametric models)

**Applicable if the expected value (under the null) of any of the cells in the table is <5. Adapted from Kirkwood and Sterne.*

Log rank test

Log rank test is based on the null hypothesis that there is an equal probability of an event to occur in each population [70]. The following calculations are used to test the hypothesis:

$$\frac{(O_{g1}-E_{g1})^2}{E_{g1}} + \frac{(O_{g2}-E_{g2})^2}{E_{g2}} \qquad \textit{Eq 1.1}$$

O_{g1}= Number of observed event in patients group 1;

O_{g2}= Number of observed event in patients group 2;

E_{g1}= Number of expected event in patients group 1;

E_{g2}= Number of expected event in patients group 2;

Number of total expected events is calculated by adding all expected events at a particular time.

Expected events at time of event T1=Death risk at time T1*number of alive patients in the group;

Death Risk =count of dead patients/count of alive samples;

Combined data from both groups is used for risk of death calculation.

Cox's proportional hazard model

Log rank test simply gives the information about difference in survival times between groups. Cox's proportional hazard model also works on similar principle but also allows inclusion of other factors during the hypothesis testing [70]. However, model uses 'hazard' as a dependent variable. Hazard provides death risk at the moment knowing that patient survived until the observed point. It is assumed that hazard ratio is independent of time and remains the same over time [68, 70]. In Cox's model, no assumption is made about the hazard probability distribution. Hazard score calculation is given in Eq 1.2 and 1.3. Hazard value over time are integrated in cumulative hazard function for interpretation [68].

$$ln\, h(t) = ln\, h_0(t) + b_1 x_1 + \cdots + b_p x_p \qquad \textit{Eq 1.2}$$

$$\frac{ln\, h(t)}{ln\, h_0(t)} = b_1 x_1 + \cdots + b_p x_p \qquad \textit{Eq 1.3}$$

h(t) = hazard at any time t; x_1 ... x_p are the explanatory variables; $h_0(t)$ indicates the baseline hazard without any explanatory variable. b_1 ... b_p are coefficients which can be estimated by using available statistical packages.

Survcomp is an R package which can be used to apply different risk models on expression data [73].

1.3.3. Annotation

By default, normalized microarray data are represented by unique labels, e.g., probeset ID. In order to make it more applicable to further analyses and cross-platform comparisons, probesets are annotated to a more routinely used ID, e.g., gene symbol, ensemble ID or entrez ID. Annotation is also important for functional enrichment analysis [43]. Multiple R packages like *BioMart* and Affymetrix specific annotation packages are available for this purpose [74].

1.3.4. Enrichment of biological knowledge

Differential expression profile is not an end-product of microarray analysis. Linking biological activities to these differentially expressed profiles can give more insight about underlying mechanisms [42]. Gene-ontology (GO) is a widely used enrichment method, which consists of three main terms, i.e., molecular function (MF), biological processes (BP) and cellular component (CC) [75]. KEGG [76] and Reactome [77] are also widely used databases for mapping differentially expressed gene list to pathways. Functional enrichment analysis of over- or under-represented terms is based on the comparison between distribution of terms in list of significantly regulated genes and the background, i.e., distribution of terms in list of all genes. Statistics methods (chi-square [78], binomial probability or hypergeometric distribution [79], Fisher's exact test [80]) are used for calculation of enrichment scores [81, 82]. A review in 2009 revealed that from 2005-2009, 68 different tools were presented for functional enrichment analysis tool [82, 83]. Accordingly, enrichment analysis can be performed using three different approaches, i.e., singular enrichment analysis (SEA), gene set enrichment analysis (GSEA), and modular enrichment analysis (MEA) [83]. SEA maps each annotation term independently with significantly regulated gene list. GOStat [84] and Bingo [85] are common examples based on SEA algorithm. GSEA, instead of using significantly regulated profile, takes all genes into consideration and a quantitative value such that logFC is used to rank these genes [83]. MSigDBis a very well-known example of GSEA [86]. Modular enrichment analysis (MEA) works on the same principle as SEA but

additionally it incorporates network identification algorithms after incorporation of term-to-term relationships [83]. DAVID is an online bioinformatics tool which utilizes MEA algorithm [87].

Different R packages have been suggested for performing functional annotation of microarray data. Some of these packages include *clusterprofiler* for gene-ontology and KEGG annotation [88] and *reactomePA* to retrieve reactome pathways enrichment [89].

1.3.5. Regulatory analysis

Besides functional enrichment analysis, several efforts have been conducted to identify shared regulatory networks in differentially expressed gene lists. Two common regulatory analyses corresponding to 5' and 3' UTR include a) transcription factor and b) microRNA binding site enrichments and another regulatory analysis pertains to chromosomal localization of co-expression genes.

1.3.5.1. Transcription factor enrichment

Almost 1800 sequence-specific TFs are present in human body and hundreds of genes are regulated by each transcription factor [90]. As transcription factors regulate diverse range of cellular processes and almost ~40% of cancer driver mutations reside in transcription factors, hence it is very important to identify TF regulators of differentially expressed networks [91].

Experimental procedures such as ChIP-ChIP is a renowned strategy to identify transcription factors binding sites on DNA-sequences [92]. After retrieval of these motifs, next challenge is to predict the genome-wide targets. Position weight matrix (PWM) has been introduced to indicate probability of any nucleotide at each position of TFBS (transcription factor binding site) [93]. TRANSFAC is one of the widely used database [94] which stores a large collection of experimentally identified eukaryotic TF binding sites and also provides the PWM for these TF binding sites. Number of other tools including MatInd and MatInspector [95], MATRIX SEARCH [96], rVISTA [97],

and SIGNAL SCAN [98] are used for PWM enrichment analysis that require input of the query sequence [99]. iRegulon plugin is available for Cytoscape which help user identify the shared transcription factors in a network [100]. iRegulon has combined predictions by using a large range of PWMs from multiple species and validation by ChIPseq data [91]. iRegulon enrichment analysis provides Normalized Enrichment Score (NES) which is established on normalized ratio of area under curve. Default NES threshold value is 3. Enrichment >3 are considered as predicted transcription factor [97]. Finding common TF-binding sequences in shared transcripts have been a useful strategy in the literature to understand the regulatory component of co-expression [58].

1.3.5.2. MicroRNA enrichment

MicroRNAs are small noncoding RNAs with lengths of ~22 nucleotides and are involved in expression regulation of thousands of genes [101]. miRbase is among the initially developed, the most comprehensive database and provides standardized nomenclature [102]. TargetScan [103] uses computational algorithms to predict miRNA-targets relationship while TarBase [104] consists of collection of experimentally collected microRNAs [105].

Targetscan is a widely used microRNA enrichment database and is available for five species, i.e., human, fish, mouse, fly and worm. It is known that 5' end of a microRNA is likely to bind to a target gene [103]. MicroRNAs contain a seed region of 2-8 nucleotides in their sequence, playing an important role in targets identification [101]. Targetscan evaluates the complementary base pairing of miRNA seed value in 3' end of target genes. 6, 7 and 8-mer seed sequences are generally conserved and used for matching. Targetscan provides conserved and non-conserved miRNA sites depending on their homology, i.e., if corresponding genomic sequence is conserved in other orthologous species, it is added into the conserved miRNA database otherwise it is considered as non-conserved miRNA [106]. Use of microRNA target prediction algorithms to compare 3'UTR binding capacity of paralogous genes is an interesting idea, which is not explored in the literature widely.

1.3.5.3. Chromosomal enrichment

Chromosomal enrichment of a significantly regulated profile is comparatively less studied field but can provide meaningful information in disease association with genomic alterations. For instance, a study about Down syndrome has indicated concordant upregulation of multiple genes on chromosome 21 [107]. Computational tools which can facilitate chromosomal enrichment analysis include online tool DAVID [87] as well as the *category* package developed in R [108]. In *category* package, over-representation of any chromosomal band is evaluated by hypergeometric test and Odds ratio [108]. Associated p-value can be used to measure significance of over-representation [108].

1.3.6. Protein-protein interactions

Experimentally, protein interactions are predicted by affinity purification by mass spectrometry [109] and yeast two-hybrid system [110]. Protein-protein interaction database development initiated by retrieving experimental interactions from scientific papers, and later on, the present databases, also started incorporating the predicted interactions using different computational algorithms to the gene lists [111]. STRING database contains interaction data for a large number of organisms from multiple resources, i.e., experimentally verified, text mining and computationally predicted interactions. Reliability of each interaction can be estimated from the confidence score calculated based on the source evidence [112]. Different protein interaction databases are listed in Table 1.2.

1.3.6.1. Network visualization tools

Cytoscape is a widely used and user-friendly network visualization tool [100]. Cytoscape facilitates the retrieval of interactions directly from multiple online protein-protein interaction databases. Different plugins designed for Cytoscape can also enable the functional enrichment of network, e.g., CLUEGO [113], iRegulon [91]. Interactive

setup of visual properties helps mapping of different statistical scores as features and annotation of the network, e.g., logFC, to node color, size, shape, helps visualize degree of change in a gene onto its protein-protein interaction network [100].

Table 1.2: List of protein-protein interaction databases [114].

This table is reprinted form *Sans-Pamplona et al., 2007* [114] and copyright permission can be found in Appendix V.

	Kind of interactions	**Included databases**	**Input**	**Distinctive features**
BIANA	Experimentally determined and predictions	String, Intact, DPI, Degg, IPI, SCOP, Uniprot, Reactome, MINT, cog and psi_msi	Accepts a variety of identifiers (Uniprot, Ensembl, GeneSymbol)	Online-interface Flexibility: It is possible to choose the network level, the relation types, restrict interactions by method and add interlogs. The output could be downloaded and visualized in Cytoscape
Poinet	Experimentally determined and predictions	DIP, MINT, BIND, HPRD, MIPS, CYGD, BioGRID and NCBI	NCBI or UniProt identifiers	By default, only experimentally determined interactions were retrieved. It is possible to filter interactions based on the number of shared GO terms between the two interacting proteins. PPI could be filtered with tissue-specific expression data from public resources. The output will be directly visualized with POINET or be downloaded and visualised in Cytoscape

	Kind of interactions	Included databases	Input	Distinctive features
SNOW	Only experimentally determined	HPRD, IntAct, BIND, DIP and MINT	Gene, transcript or protein	Construct minimal connected network (MCN); a graph containing only seed and linker proteins Maps seeds proteins onto an interactome of reference calculating the network parameters degree, clustering coefficient and betweenness
UniHi	Experimentally determined and predictions	MDC_Y2H, CCSB, HPRD, DIP, BIND, IntAct, BioGRID, COCIT, REACTOME, ORTHO, HOMOMINT and OPHID	Entrez Gene, Gene Symbol, UniProt, NCBI, Ensembl, RefSeq, BioGRID, HPRD, OMIM	Uses Human Gene Atlas data to construct tissue-specific interaction networks Annotate networks with pathway information from KEGG database Only accept a maximum of 15 proteins as an input

1.4. Meta-analysis

As microarray or RNAseq expression experiments have become a common practice, the number of expression datasets proportionally increased in public databases. Meta-analysis approach of combining several studies exhibiting a commonality was proposed to minimize the artifacts arising from an individual dataset [115]. After identification of differentially expressed genes using meta-analysis methodologies, biological functions and regulatory analysis can be conducted in the same way as mentioned in Section 1.3 [115]. It has been shown that detection of small but homogeneous effects in multiple studies obtained by meta-analysis is indeed powerful [116]. Secondly, by including studies from different labs performed by different people helps transition of case-specific results to generalized terms [117].

Seven major steps in conducting gene expression meta-analysis are mentioned below [118]:

1) Selection of suitable studies by review of the literature and databases;

2) Downloading the expression data;

3) Preparation of data to remove bad quality samples, make data from different platforms compatible with each other, and handling of multiple technical replicates from single biological replicates;

4) Annotate the probesets into gene symbols;

5) Removing probesets having many-to-many relationship to gene;

6) Selection of a proper meta-analysis methodology (described below);

7) Interpretation and presentation of the obtained results.

Different statistical ways to use the meta-analysis profiles include differential expression [119], co-expression analysis [57], pathway analysis [119] and survival analysis [120]. A review conducted in 2012 showed that the largest number of meta-analysis studies belonged to genome-wide expression profiling; and the differential expression analysis has been the most prominent [121] (Figure 1.5).

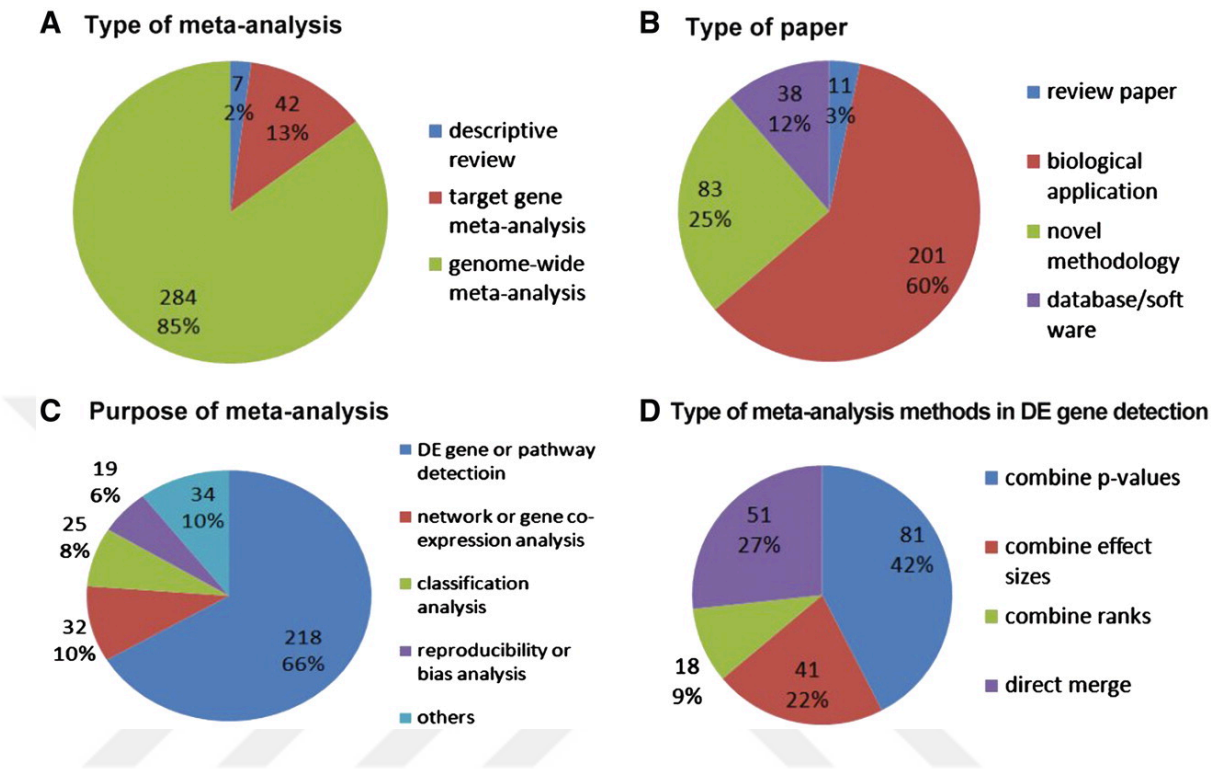


Figure 1.5: Statistics of different meta-analysis used in different research papers [121].

This figure is reprinted from *Tseng et al., 2012* [121] and copyright permission can be found in Appendix V.

1.4.1. Meta-analysis algorithms

Four different methods of meta-analysis have been presented and include combining 1) p-values [122]; 2) effect-sizes [123]; 3) ranks [124]; and data directly from included studies [125].

1.4.1.1. Combining P-values

In 1932, Fisher proposed a method to combine independent p-values [126]. Meta-analysis on the basis of p-value is simpler and can be extendible to multiple types of outcome [122]. It can also be used in case of multiclass comparisons. Fisher's method transforms individual p-values by taking log of them and summarizes the results by adding all p-values (Eq 1.4)[126] .

$$S_i = -2\Sigma_{k=1}^{K} log\ (p_{ik}) \qquad \textit{Eq 1.4}$$

K = Total number of datasets, I = gene, p_{ik} = p-value of *ith* gene in *kth* dataset

Stouffer presented another method by taking inverse cumulative distribution function [127]. Two other methods depending on the minimum and maximum p-values (more stringent) from all individual studies are used to represent the meta-analysis p-values [121, 128]. However, using p-value meta-analysis can vanish the effect of direction so it is advised to calculate one-sided p-value for each study [129, 130].

1.4.1.2. Effect-size based meta-analysis

Choi et al., in 2003, have presented a strategy of combining studies by using the differential effect (between groups) from each expression study, i.e., effect-size [123]. For two groups comparisons, a t-like statistics can be presented [123]. Unlike p-values or rank based methods, using the effect-size has an advantage for identification of up- and down-regulated genes. For a single gene of each dataset, calculation of effect-size are given in Eq 1.5 [116].

$$d_j = \frac{\bar{T}_j - \overline{C_j}}{s_{pj}} \qquad \textit{Eq 1.5}$$

$\bar{T}_k$ and $\overline{C_k}$ represent the mean value of both group respectively in *jth* study;

S_{pj} represent pooled standard deviation for the *jth* study.

Based on the above model, two methods, i.e., fixed and random effect models, FEM and REM, respectively, have been developed and are widely used in expression data meta-analysis [131]. Basic difference in both methods lies in their way of handling heterogeneity. FEM can be applied when data are more homogeneous and it incorporates true effect-size along with the measurement error (experimental or sampling error). REM, on the other hand, also incorporates between study heterogeneity [121, 131]. Applicability of fixed or random effect models is evaluated by Q-statistics, i.e.,

heterogeneity score [121]. Permutation test is conducted to calculate the p-value and associated false discovery rate (FDR) [128].

1.4.1.3. Ranking based meta-analysis

As the name implies, this methodology ranks differential expression scores form all genes. It is a non-parametric statistic measure [132]. Ranks of individual studies are obtained by logFC and are summarized by the product, mean or sum [116, 121].

1.4.1.4. Merging data

Instead of using any meta-analysis method, several studies have selected direct merging of expression from different resources after applying inter-study normalization methods [133]. Although the method requires less effort and statistical knowledge it does not take between study heterogeneity into consideration and may need further improvements by use of preprocessing and batch correction methods [121].

1.4.2. R packages for meta-analysis

In the recent years, with increasing trend of combining expression data, multiple R packages have been presented for meta-analysis differential expression which can use either of the rank based, p-value or effect-size based approach, e.g., *MAMA*, *geneMeta* [134], *metaMA* [129], *metaArray* [135], *RankProd* [121, 124, 136].

METADE belongs to a suite of R packages focusing on different aspects of meta-analysis, i.e., quality control (*MetaQC*), meta-differential expression analysis (*MetaDE*) and meta-pathway analysis (*MetaPath*) [137]. *MetaDE* package facilitates meta-analysis by using variety of approaches including p-value based meta-analysis, i.e., Fisher, maxP or minP approach), rank based strategies (rank product or rank sum) or effect-size based strategy (fixed and random effect models). *MetaDE* also allows matching gene names before meta-analysis [137]. One important aspect is that both paired and unpaired data can be analyzed after labeling each study [137].

1.4.3. Pathway enrichment using meta-analysis

Pathway analysis is usually conducted after meta-analysis driven differential expression profile is obtained. List of significantly modulated genes is enriched with biological pathways using pathway information from KEGG [76] or GO [75]. However, few efforts have also been made to perform meta-analysis on the basis of pathway enrichment. Shen and Tseng [138] have provided a strategy along with an R package to perform meta-analysis at the pathway level, i.e., Meta-Analysis for Pathway Enrichment (*MAPE*).

MAPE includes three main strategies, i.e., meta-analysis at gene level (MAPE_G), pathway level (MAPE_P) or using a hybrid approach (MAPE_I). MAPE_G allows meta-analysis at the gene level, which is followed by a pathway enrichment analysis. Nevertheless, MAPE_P is designed to perform pathway analysis at individual study level and this information is then combined as meta-analysis [121, 137].

1.4.4. Co-expression using meta-analysis

Meta-co-expression analysis is a less studied field with fewer available approaches. Multiple studies have used direct merging of data for building a co-expression network [139, 140]. Few studies have retrieved gene interactions from each study and afterwards, the confidence on each study has been determined by vote counting or Fisher's method [141, 142]. In 2005, Zhou et al. has used 2nd order correlation analysis to correlate expression pattern in 39 microarray studies [143]. In 2006, Hutten et al. has calculated the correlation between each gene-gene pair from individual study and combined these correlation scores into a z-score (meta-correlation score), which is ultimately used for co-expression network [121, 144].

Metacor is an R package that requires input of individual correlation coefficients from each study along with a number of samples [145]. *Metacor* combines these correlation coefficients using classical random effect model, i.e., Der Simonian and Laird (DSL) method [145].

1.4.5. Heterogeneity analysis and publication bias

Selection of fixed and random effect models is dependent on the heterogeneity between studies. Two mainly used statistics for calculation of heterogeneity include Cochran Q-statistics and I^2 [146]. Cochran statistics is described in Eq 1.6 [146] and is generally followed by a p-value to identify significance in heterogeneity and can be calculated as:

$$Q = \sum_{i=1}^{k} w_i \ (y_i - \mu)^2 \qquad Eq1.6$$

$$wi = 1/v_i$$

Vi is variance of study *i*; μ is the pooled estimate and *y* is the effect-size; and *k* is number of total studies [147].

I^2 is a another quantitative measure ranging from 0-100% and a cut-off is selected to estimate heterogeneity [146]. I^2 calculation is presented in Eq 1.7 [147]

$$I^2 = \frac{H^2 - 1}{H^2} \qquad Eq1.7$$

While H^2 is defined as *Q/K-1*. *Q* represents the Cochran Q-statistics [147].

Publication bias is generally investigated by visual observation of a funnel plot. Funnel plot is comparison of study size against the effect-size obtained from each study [148]. In case of no bias, a funnel appears symmetric and is more variable for smaller studies as compared to larger studies. Symmetry of funnel plot can also be evaluated using regression based test, e.g., Egger's test finds association between effect-size and standard error of each study and $p<0.05$ indicate significant presence of publication bias [147, 148].

1.5. Shiny package for web-application development

In past several years, variety of HTML and JavaScript based web tools have been excessively developed and are made available over World Wide Web. R is one of the open source programming language and has increasingly become popular among

bioinformaticians. Recently R has introduced *shiny* package to facilitate online web tools development [149].

1.5.1. Shiny Architecture

Shiny web apps consist of two parts designated for web interface designing (UI.R) and server activities (server.R). Both scripts should be placed in the same folder. Shiny package works on reactive programming [149, 150]. The main concept behind Shiny is to take input and produce output; thus, due to underlying reactive programming, Shiny responds to changed input and reflects changes immediately in the output plots/text. The procedure allows Shiny to immediately respond to changing output without refreshing [149, 150]. In addition, multiple packages have been designed to facilitate Shiny-based web development, e.g., to decrease processing time of heavy programs *doparallel* and *parallel* packages are developed [151, 152]. *Compiler* package is also designed to increase the speed in R [150, 152].

Web development in Shiny is easier as large collection of widgets are available which facilitate decorating interfaces for input/output [149]. Shiny mediates fast communication between web page and R via *websockets* package. This fast communication is especially useful in cases when there is a continuous dialogue between the user and web-apps. *Websockets* strategy is well established, supported by multiple web browsers and is widely used in online applications, e.g., online gaming [149, 150].

1.5.2. Sharing of web applications

Shiny web application can be shared by several options. Both UI.R and Server.R can be passed to a colleague and can be run on any computer having R and Rstudio installed [153]. Application can be opened in R and using *RunApp* button or command, app can be visualized on web page in an interactive way. Second option is to deploy application to public repositories, e.g., GitHub [154]. Another alternative can be use of the servers

hosted by shiny, i.e., ShinyApps.io. ShinyApps.io doesn't need R on local computer and can be used over web [153].

1.5.3. Bioinformatics applications developed in Shiny

Although Shiny-based web application development is a relatively new concept several web applications already have been designed in the past couple of years. Recently, in 2016, shinyGEO web-app is developed which facilitates the user to download any data from GEO database and perform differential expression along with survival analysis [155]. Digital droplet PCR (ddPCR) enables exact quantification of DNA and a Shiny web-app has been designed based on analysis of ddPCR data [156]. VennDiagramWeb, as the name indicates, is designed to summarize and visualize data in form of Venn or Euler's diagram [157]. START app is developed to upload RNAseq data and visualize it with different kinds of plot graphics which ease interpretation of RNAseq data [158]. rTRM-web is designed to predict transcriptional regulatory module by using ChIPseq, gene expression and protein-protein interaction data [159]. PAEA (Principle Angle Enrichment Analysis) is developed for gene set enrichment analysis based on dimension reduction [160]. For calculation of sample and power size of RNAseq data, RNAseqPS has been developed [161]. Seed is developed to visualize and explore microbial community data by performing principal component analysis (PCA), hierarchical clustering, scatter plot, and heatmap [162]. CoRegNet is intended to construct co-regulatory network [163]. Some other web applications developed using *shiny* include visPIG [164], pophelper [165], ClustVis [166], Methylation Plotter [167], Structure Plot [168], Methyl Aid [169], shiny-phyloseq [170]. However, there is not many meta-analysis applications generated by using Shiny in the literature.

1.6. Sources of expression data

1.6.1. Gene Expression Omnibus (GEO)

National Center for Biotechnology Information (NCBI) has designed Gene Expression Omnibus (GEO) database, which facilitates researchers to upload their data, and make it available to research community for free. GEO requires the user to upload raw as well as processed data. GEO also requires short description of experimental groups, characteristics of samples and strategy used for expression analysis [171]. Initiated in 2002, journals started requests of depositing microarray data in public databases, which has led to a huge increase in the number and variety of expression profiles in GEO [171]. Current statistics of GEO data is presented in Table 1.3. To standardize information from different experiments, GEO upload works on the Minimum Information About a Microarray Experiment (MIAME) guidelines [171].

GEO accepts high-throughput data from multiple technologies. Database has its specific nomenclature to represent platform (GPLxxx), samples (GSMxxx) and series (GSExxx). In addition to data, GEO has also provided a statistical tool GEO2R to perform statistical analysis and visualize expression of specific gene [171].

Table 1.3: Amount of data available in GEO database (according to July, 2017). (*https://www.ncbi.nlm.nih.gov/geo/summary/?type=platforms*).

	Public	**Unreleased**	**Total**
Series	86,911	10,721	97,632
Platforms	17,493	300	17,793
Samples	2,138,213	328,398	2,466,611

1.6.2. COSMIC

In 2004, COSMIC has been developed to collect data about somatic mutations in human cancer [172]. With time, COSMIC has expanded to different sources of cancer data including coding and non-coding mutations, fusion genes, and re-arrangements of

genome, copy number variations and gene expression data. Statistics of COSMIC database (v70) released in 2014 is shown in Table 1.4 [172].

1.6.3. ARRAYEXPRESS

ARRAYEXPRESS has been developed in 2003 and like GEO, database enables user to upload data and retrieve publicly available data. However, in addition, this database also imports expression data form GEO database to collect large repertoire at one place [173].

1.6.4. CCLE

Cancer cell line encyclopedia (CCLE) is a platform hosting cancer genomic data of 947 cell lines on 36 tumor types. CCLE data includes mutation, DNA copy number and mRNA expression data. CCLE facilitate its user to download raw as well as processed data [174].

1.6.5. TCGA and Firebrowse

The Cancer Genome Atlas (TCGA) is a large repertoire of RNAseq data encompassing gene, protein expression and epigenomic data for 20 different cancer types with thousands of samples in each type. TCGA has categorized data in three different levels. Level 1 and level 2 data contain raw information about data such as DNA sequence and germline variants. However, level 3 data is preprocessed and publicly available for use [175, 176]. Since, mining of TCGA data can rather be difficult, Firebrowse is developed to facilitate easy download and analysis of TCGA data (http://firebrowse.org/).

Table 1.4: Summary Statistics of COSMIC data resource.

Source	Number of samples
Genes (transcripts)	28735
Tumor samples	1029547
Coding mutations	2002811
Curated publications	19703
Fusion mutations	10435
Genomic rearrangements	61299
Whole genomes	12542
Copy number aberrations	695504
Gene expression variants	60119787

1.7. Outline of work in thesis chapters

In the present thesis, I have performed multiple meta-analyses focusing on different contexts of inter- or intra-species comparative transcriptomics.

In chapter 2, I have compared a zebrafish microarray dataset (GSE5308) which was obtained in Konu Lab, with respective and meta-analyzed mouse expression datasets to identify commonly shared KEGG pathways and GO terms in response to rapamycin exposure. In addition, paralogous gene expression in zebrafish has been studied for the first time regarding rapamycin treatment. I also presented a novel database, called CompariZome, for paralogous gene expression analysis in zebrafish and its comparative transcriptomics with similar human datasets.

In chapter 3, I have used CCLE dataset and obtained correlations between CDH1, HNF4A, and GRHL3, three genes that have been shown to be important in mammary differentiation in mouse, for different cancer types and performed a meta-analysis for CCLE data.

In chapter 4, I have presented multiple meta-analysis strategies for comparison of cellular differentiation and breast cancer E2 mediated signaling using independent cohorts and applied these meta-analyses to a case study exemplified by CHRNA5, cholinergic receptor nicotinic alpha 5. In addition, an online Shinny app, called E2S (Estrogen (E2) to Survival), allowing for meta-analysis of any gene in the context of E2-signaling based on *in vitro* MCF7 cell line and *in vivo* breast cancer patient cohorts, has been presented.

In Chapter 5, I have collected normal tumor datasets in paired and unpaired fashions for breast cancer and performed meta-analysis of correlation between mineralocorticoid and glucocorticoid receptors (MR and GR, respectively) and generalized the conclusion that MR and GR were downregulated in tumor samples and unpaired and paired datasets were comparable. In addition to the correlations I have obtained a measure of deregulation of MR and GR in tumor tissue with respect to its adjacent normal tissue. All of these has been also made into an online Shiny tool called DualExpBC for testing

correlation and deregulation of expression between any two genes of interest using meta-analysis techniques and visualizations. Hence, in the present thesis, novel methods, approaches and findings are presented with respect to different aspects of cancer treatment, diagnosis and prognosis.

Chapter 2

Comparative transcriptomics in zebrafish, mouse, and humans

2.1. Introduction

2.1.1. Model Organisms

Model organisms are non-human species, which can be used for understanding and analyzing of experimental results so that the hypothesis can be generalized to humans [177]. Since model organisms are easy to breed and can be maintained in large numbers in the lab environment, it is possible to observe long-term effects of a treatment in next generations. 13 model organisms are recognized by U.S. National Institutes of Health and include rat, mouse, zebrafish, *Drosophila*, and nematode worm. Choice of an organism depends on the biological question and availability of suitable techniques to answer that question [178].

2.1.2. Mouse as a model organism

Mouse is a well known, widely used cancer model and a lot of techniques/tools have been developed and applied to facilitate mouse experiments [179]. Several factors including its smaller size, easy breeding, shorter lifespan, molecular similarities and sequenced genome have made many cancer studies possible in mouse. Advancement and research in the field have given rise to successful xenograft models based on tumor cell lines, carcinogen treatment and genetically engineered mice. However, challenges of modeling cancer in human include disparities between two species because of differences in orthologs and telomere length [180-182]. Moreover, GEO and ArrayExpress house a high number of mouse expression data from drug treatment and

knockout models enabling comparison with human or other model organism datasets [171, 173].

2.1.3. Zebrafish as cancer model

Zebrafish has emerged as a useful model organism because of several characteristics including its smaller size (1-2 inch), transparent embryos to facilitate observation of morphological abnormalities, large number of eggs produced per setup and consequently large progeny, easy fertilization of eggs *in-vitro*, rapid development outside and giving rise to free-swimming larvae in a short period, e.g., 4-7 days. Organogenesis can be easily studied as all organs become recognizable on 5^{th} day [183]. Besides, 70 percent of genomic similarity (coding genes) between human and zebrafish makes inter-species comparisons possible. 84 percent of human disease related genes are also known to have a zebrafish counterpart [184]. These characteristics have increased the worth of zebrafish as model organism in transcriptomics and genomics [184].

2.1.4. Cross-species oncogenomics in zebrafish

Comparison of human, mouse and/or zebrafish genomic abnormalities can provide insight of shared or deregulated mechanisms across species. Human cancer has heterogeneous genetic makeup, which makes it difficult to identify common driver genes. Comparative genomics can help identify the most conserved pathways and genes showing alterations. Comparative genomic hybridization of T-cell acute lymphoblastic leukemia in zebrafish has resulted in identification of 893 genes with differential copy number alterations [185]. However, transcriptome comparison with public human tumors datasets resulted in only 9 shared genes between both species, which is quite a small number [185]. Literature review of the shared profile indicated that these genes indeed possessed functional importance highlighting probability of cancer driver genes [185]. Transcriptome comparison between zebrafish liver cancer and four different human cancer types (liver, gastric, prostate and lung) was performed using binomial test and GSEA (Gene set enrichment analysis) [186]. Individual comparison of each human

dataset with that of zebrafish liver cancer showed that the shared number of genes between zebrafish and human liver cancer were much more than the other evaluated cancer types [186]. This shared liver cancer signature was comprised of 76 genes and functional enrichment indicated its association with pathways involved in pathogenesis of hepatocellular carcinoma, e.g., WNT–β-catenin and RAS–MAPK pathways [186]. Medaka, another fish species, has similar characteristics to zebrafish and is used as a complementary model organism [187]. Cross-species pigment tumor expression profile of medaka, zebrafish and humans showed a highly conserved expression pattern. Comparison with human dataset was based on 1) an independent human melanoma microarray dataset and also on 2) previously published meta-analysis signature [187]. Cross-species comparisons of microarray or expression data have not been limited to zebrafish or cancers. Using microarray profiles, common regulators of hematopoietic stem cells across organisms including zebrafish, mouse and humans were identified [188]. Hence, the overlapping genes and pathways in up- and down- regulated categories of human, mouse and zebrafish were used to represent the shared signature of hematopoietic stem cells [188]. Transcriptome conservation of distal renal convoluted tube among zebrafish and mouse was also analyzed by considering overlapping significantly modulated genes which resulted in 13 shared genes between both species [189].

2.1.5. Comparative toxicological studies

A toxicology-oriented study involves the comparison of mercury exposure in zebrafish liver and with mercury treated public human HCC HEPG2 cell dataset [190]. In this study, significantly affected pathways with a given FDR and enrichment scores were identified yet no significant association statistics were used. Comparative studies revealed a similar hepatotoxic effect in both species of significantly upregulated genes that involved proteasome and DNA damage pathways while downregulated-shared pathways included electron transport chain and nuclear receptor signaling pathways. Moreover proliferative pathways were specific to HepG2 cell line [190]. Another toxicology based study evaluated transcriptomics effects of multiple xenobiotics

(cyclosporine A, amiodarone, and acetaminophen) in zebrafish embryo, *in-vitro* and *in-vivo* mouse, rat and human primary hepatocytes [191]. Comparative toxicogenomics revealed that gene-overlap could not result in a concordant effect and model-specific changes were more prevalent. On the other hand, the pathway overlap could identify conserved hepatotoxic effect of different drugs in different model organisms. Pathway "regulation of metabolism – bile acids regulation of glucose and lipid metabolism via FXR" was shared in analyzed toxicogenomics comparison indicating that zebrafish embryos can be used for prescreening of drugs [191].

Similarly, another study compared the microbiota suppression of HNF4A activity in human Inflammatory Bowel Disease patients and zebrafish using comparative transcriptomics [192]. Yet another study compared estrogen (E2) treatment profile in *in-vivo* zebrafish and human breast cancer cell lines putting forward a signature of 139 genes, which was shared between both species [193]. In this study, differentially expressed signature of multiple human breast cancer cell lines was individually compared with zebrafish E2-mediated profile and shared regulation between two species was identified by evaluating overlap of the significantly modulated genes [193]. Many others exist thus showing the importance and value of such comparative transcriptomics analysis between species. Although a significant amount of work has been performed for anticancer drug expression profiles in humans and rodents their corresponding zebrafish datasets were not analyzed in detail. Comparison of H2O2 signaling between larval zebrafish and human keratinocytes [194] is another example yet this subject remains an open area of research in comparative onco-transcriptomics.

2.1.6. Duplicated genome in zebrafish

Despite of having high molecular similarity with human, zebrafish genome is interesting because of the presence of an additional whole genome duplication (WGD) event [195]. Duplicated gene pairs (i.e., paralogous) are referred as 'ohnologues'. Number of total protein coding genes in zebrafish, i.e., 26206, is higher than known vertebrates, which might also be due to WGD. Based on orthology information of one-to-one relationships retrieved from Ensemble, 47% of human genes with human-zebrafish orthologous pair

genes have been identified [184]. One-to-many relationship (One human genes with multiple zebrafish genes) resulted in average of 2.28 zebrafish genes corresponding to each human gene. This finding of 1-2.28 ratio might be explained by WGD [184]. Interestingly, few important genes, e.g., leukemia inhibitory factor (*LIF*), oncostatin M (*OSM*) or interleukin-6 (*IL6*), did not have any orthologues in zebrafish but their receptors were found in zebrafish [184]. Gene duplication thus can lead to redundancy in the genome as well as pseudogenization, conserved functionality, sub-functionalization or neo-functionalization.

Pseudogenization: As a result of evolution, most likely fate of duplicated pair is that one of the duplicated genes accumulates mutations and eventually loses its function, i.e., pseudogene. According to evolutionary point of view this pseudogene is meant to be deleted from the genome or will become distinct from the duplicated ancestor [196, 197].

Conserved gene function: Concerted evolution is referred as two paralogous genes having similar sequence and function [198].

Sub-functionalization: This arises when both genes in the duplicate pair show mixture of similar and different characteristics, e.g., engrailed-1a and engrailed-1b have arisen after genome duplication [199] and both are transcription factors differing in their location such that *engrailed-1a* is expressed in pectoral appendage bud while *engrailed-1b* is specific to neurons [200].

Neo-functionalization: This refers to to gaining a completely new function. Mostly related functionality instead of completely novel functions arises as the result of WGD. For the phenomenon of neo-functionalization to occur, multiple amino acid changes are required [198].

2.1.7. Computational databases/tools for inter-species comparison

Zebrafish information network (ZFIN) was developed in 2003 to facilitate the curation of zebrafish genetic, genomic, expression and functional data [201]. OGEE database was

developed to determine essential and non-essential genes in multiple species [202]. Genes required for viability and fertility were classified as essential genes. Essentiality of a gene was determined by considering several factors including gene duplication, interaction between genes, evolutionary relationship, earliest expression during development [202]. Inparanoid database, encompassing 273 species data, has been developed to retrieve orthologs between two species; orthologs were identified by alignment using BLAST followed by clustering [203]. Several other recent comparative genomics approaches between zebrafish and human include CoexpressDB, a database that integrates multiple species expression data for comparison among them [204]; this database is comprised of gene expression data on 11 species. CoexpressDB provides confidence on resulting co-expression network by calculating reliability score [193]. Protein classification from different organisms based on their evolutionary and functional relationship has been processed and made available on PANTHER database [205]. Yet another database ECRbase has been designed for identification of conserved regulatory regions in different vertebrate organisms [206]. Another computational strategy, ping-pong algorithm was put forward to facilitate inter-species comparisons of co-modules on the basis of gene and organ homology and was evaluated on human and mouse expression profile however no online tool was presented[207].

Gene duplication is a well-known phenomenon, which can also give rise to pseudogenes. To store repository of pseudogenes in multiple species, pseudogenes.org was presented [208]. PhyloFish database was developed for duplicated genome analysis and processed the duplicated pairs and expression data of 23 different fish in 10 different tissues [197]. Duplicated genes database is another attempt to combined duplicated gene pair information with expression profiles for nine species [209]. WANDA database was developed to identify duplicated genes which have one copy in mammals and two or more copies in ray-finned fish but did not link the orthology information to expression changes [210]. Even though, multiple efforts have been made, there is no platform, which can enable analysis of duplicated gene pairs under different biological conditions (e.g., knockout studies, drug exposure, disease condition) which can help evaluate condition based duplicated gene pair converged/diverged expression. Besides, a platform

which can evaluate duplicated gene pairs in context of expression data and different functional enrichments, including microRNA binding, is still missing.

2.1.8. Bioinformatics approaches to identify duplicated zebrafish genomes

Taylor et al. in 2003 has used blastp and CLUSTALW alignment along with clustering tools to identify zebrafish-duplicated genes. Duplicated genes were selected if they exhibit two copies in zebrafish but a single corresponding copy in humans, mouse, chicken, frogs (one or two copies). Furthermore, comparison of identified orthologs with another ray-finned fish (also have duplicated genome), i.e., pufferfish, revealed that 22 of these genes also had ohnologues in pufferfish [211]. These findings indicate that these 22 gene pairs could have originated as a result of WGD event specific to ray-finned fish [211]. However, this study focused on earlier versions of the genomes and did not provide a database/web-tool. Another approach of alignment BLASTN/BLASTX has also been used to identify human/zebrafish orthologs [212]. In 2012, a duplicated genome database has been established, which combined alignment information with the gene expression and gene ontology data to predict duplicated genes in nine different species [209]. A more recent strategy in 2016 presented has also used blastp, an analysis comparing each zebrafish gene with all other zebrafish genome [213]. This analysis helped to identify duplicated genes based on sequence similarity. Evolutionary time was then calculated by the ratio of non-synonymous to synonymous substitution (ka/ks) and genes with the lowest ks score were selected [213]. Ensembl Biomart is another reliable source to retrieve orthology information providing one-to-one or one-to-many status as well as ka/ks rates between species thus can be implemented easily in comparative transcriptomics studies [214]. As Ensembl Biomart is regularly updated, it might be advantageous for the retrieval of orthology/paralogy information.

2.1.9. mTOR pathway

Increased metabolic activity is one of the major hallmarks in cancer. Deregulation of glucose pathway in cancer is regulated by mutations in oncogenes or downstream

effector of PI3K/AKT mediated signaling, an important survival, metabolism and cell growth-controlling pathway. Because of its vast applications, PI3K/AKT pathway has been a major focus in cancer therapeutic attempts [215, 216].

PI3K enzymes are involved in variety of cellular processes including proliferation, survival and angiogenesis. In mid 1990s, another family of enzymes with resemblance to PI3K has been identified and named as PI3K and related kinases (PIKKs). Like PI3K, this family regulates proliferation and survival related processes [217]. One of PIKK subfamily includes the TOR proteins. Multiple of mTOR pathway genes including PI3K, PTEN, AKT, 4EBP1 have been found to be deregulated in variety of cancer types [217]; besides TOR pathway has been highly conserved in eukaryotes which can facilitate its manipulation in a variety of different model organisms [218]. Targeting of mTOR pathway is a promising strategy for cancer therapy. Rapamycin and its analogous compounds specifically bind to members of Target of Rapamycin (TOR) subfamily [219]. TOR mediated anti-cancer activities involve G1 arrest in cancer cells and immune-suppression [220, 221].

mTOR proteins are comprised of two complexes, i.e., mTORC1 and mTORC2. mTORC1, targeted by rapamycin, is comparatively well studied and is known to regulate growth and proliferation [219]. TSC1 and TSC2 complex are inhibitors of mTORC1 and are also known as tumor suppressor genes [222]; hence deregulation of TSC1 and TSC2 leads to continuous cell proliferation [219]. Zebrafish has two homologs of TSC1 and are called as TSC1a and TSC1b [223]. Rapamycin treatment in zebrafish has also been shown to result in developmental delay [224].

2.2. Aims and rationale

mTOR pathway has been studied in several species from yeast to humans and was found to be highly conserved. Previously, Ceren Sucularli (PhD Thesis, 2011) has performed rapamycin treatment (100nM and 48 hours) in zebrafish embryonic fibroblast cell line ZF4 as well as *in-vivo* dose- and time- dependent (2,10 and 20uM; 1-5dpf) analysis in zebrafish embryos. Dose and time- dependent (1-5dpf) rapamycin exposure in zebrafish embryos resulted in delayed embryo development with larger yolks and smaller body size. Rapamycin treatment in ZF4 cell line did not show significant effects on cytotoxicity and DNA synthesis but resulted in reduced levels of apoptosis. qPCR and microarray experiment performed on ZF4 cell line showed concordance in gene regulation of 9 evaluated genes. Microarray data has been submitted to GEO (GSE53086) (Ceren Sucularli, PhD Thesis, 2011; [225]).

Meta-analysis has been widely applied in expression studies, both microarray and RNAseq. However, its application in inter-species transcriptomics comparisons is relatively lacking. In order to address this deficiency in the field, I aimed to address the following points in this thesis:

- Determination of concordant/discordant rapamycin-mediated changes between two widely used model organisms, i.e., zebrafish and mouse. Zebrafish dataset obtained in our lab (GSE53086) and a meta-analysis of four mouse public datasets were used for comparative transcriptomics between zebrafish and mouse rapamycin exposure.
- Identification of shared/diverged duplicated gene pairs in rapamycin treated zebrafish expression profile (GSE53086).
- Development of a web-app, called CompariZome, that can generalize comparative transcriptomics in zebrafish and human with respect to duplicated gene pairs.

2.3. Methodology

2.3.1. Rapamycin treated expression dataset collection and processing

Rapamycin treatment microarray studies performed on the mouse Affymetrix platform GPL1261 were searched form GEO database [226]. GPL1261 (Affymetrix Mouse Genome 430 2.0 Array) is a widely used platform for mouse expression analysis. Mouse cell lines expression data with at least two samples in each of the control and rapamycin treatment groups were selected. Raw data of three datasets were downloaded using GEOquery package (Table 2.1) [227]. GSE21755 expression data were based on two different cell lines with respective control and rapamycin treated samples [228]. Both cell lines data from GSE21755 were separately analyzed. Quality control of raw data was performed using *affyPLM* [229] and *affyQCReport* [45] packages in R. Rapamycin treated samples were RMA normalized with respective control samples using *affy* package [51]. Normalized data were further used for analysis and comparison as described below:

1) For each probeset in the transcriptome, a log fold change (logFC) was calculated for all four datasets as described in Eq 2.1

   LogFC=Average (rapamycin treated samples)-average (control samples) *Eq 2.1*

   Pearson correlation between the logFC of datasets was calculated in a pair-wise fashion using *stat* package in R [152].
2) Meta-analysis of mouse expression data was conducted using *metaDE* package [230] in R. Parameters used for meta-analysis score calculation included random effect model (REM) with seed value of 123 and 1000 permutations.

2.3.2. Identification of differentially expressed probesets in zebrafish and mouse

Rapamycin treated zebrafish dataset had been obtained and analyzed by Ceren Sucularli using BRB-ArrayTools (Sucularli, PhD Thesis, 2011). A P-value threshold of 0.05 was used to extract 2186 significantly rapamycin regulated zebrafish probesets [225].

List of significant probesets were retrieved from mouse meta-analysis results using FDR<0.005 for functional enrichment analysis and comparison studies (with the zebrafish expression profile). Three other thresholds FDR<0.05, FDR<0.01 and top 2186 probesets were also used for validation. Top 2186 probesets from mouse profile were retrieved by ranking the probesets with respect to their meta-analysis p-values. Upregulated (logFC>0 in zebrafish/mu.hat>0 in mouse meta-analysis) and downregulated (logFC<0 in zebrafish/mu.hat<0 in mouse meta-analysis) probesets were extracted from significantly regulated gene list to represent the interspecies rapamycin meta-analysis signature.

2.3.3. Comparison of functional enrichment between zebrafish and mouse transcriptome

Significantly modulated transcriptome (p<0.05 from zebrafish and FDR<0.005 from mouse) was used for KEGG and Gene ontology enrichment using the online tool DAVID [87]. Functional annotation terms having p<0.05 were considered as significantly enriched in the respective profile. Following steps were performed for comparison between concordant/divergent functional regulations of significantly modulated profiles.

Step 1: calculation of count data

Contingency matrix of count data was generated as below:

UPUP=Count of shared terms in the upregulated profiles of both species

DOWNDOWN=Count of shared terms in the downregulated profiles of both species

UPDOWN=Count of terms enriched in the upregulated zebrafish profile and downregulated mouse profile

DOWNUP=Count of terms in the downregulated zebrafish profile but upregulated in mouse profile

Step 2: Association statistics

Using count data, Fisher's exact test along with Odd ratio (OR) was used to test the association between the zebrafish and mouse profiles. *Exact2x2* package [231] in R was used for calculation of OR and fisher's exact p-value.

Step-1 and step-2 were repeated for different gene ontologies (biological processes, BP, cellular component, CC, and molecular function, MF) and differentially expressed gene list obtained from three different mouse meta-analysis thresholds: FDR<0.05, FDR<0.01 and top 2186 probesets.

2.3.4. Duplicated gene pair retrieval and processing

Ensembl biomart database [214] was used to retrieve zebrafish orthologs of human genes. Retrieved human-zebrafish ortholog databases were processed as below:

1) Human-zebrafish orthologs having one-to-one and many-to-many relationships were ignored.
2) One-to-many orthologs (i.e., one human gene mapping to multiple zebrafish paralogs) were processed for retrieval of duplicated zebrafish gene pairs. More than two zebrafish genes mapping to single human ensemble ID were not used in further analysis and only those paralogs having two different zebrafish ensemble IDs mapping to single human ID were considered as duplicated pairs. In addition, if both genes of identified duplicated pairs mapped to the same Affymetrix probeset; these pairs were not used.

2.3.5. Duplicated gene pairs analysis

Effect-size and p-value were used to indicate the converged/diverged expression of genes in duplicated pairs and were calculated as given in Eq 2.2 and 2.3. For each duplicated pair:

logFC1=log fold change of duplicated pair gene 1

logFC2=log fold change of duplicated pair gene 2

Duplicated pairs effect-size (DPES) was calculated as

$$^{sign(logFC1*logFC2)}\sqrt{abs(logFC1 * logFC2)} \qquad Eq\ 2.2$$

In addition, p-value of a duplicated pair was calculated as:

P1=p-value of *limma* analysis from duplicated pair gene 1

P2=p-value of *limma* analysis from duplicated pair gene 2

Duplicated pair p-value (DPP-value) was identified as

$$P1+P2 \qquad Eq\ 2.3$$

Significant p-value indicates that both genes of paralog pairs were regulated during any treatment. DPES>0 was used to identify duplicated pairs in which both genes are regulated in same way, i.e., either both genes of pair were upregulated or both genes were downregulated while on the other hand DPES<0 indicated that the direction of modulation was opposite in both genes of a paralog pair.

2.3.6. Mapping of rapamycin treated zebrafish profile with duplicated pairs

Rapamycin treated and DMSO control .cel files were normalized using RMA form *affy* package [51]. Differential expression analysis was conducted using *limma* package in R [56] to obtain log fold change and p-value for each probesets. Mapping of differential expression results with duplicated pairs resulted in 606 pairs. These pairs were processed to calculate DPES and DPP-value as described in previous section. Duplicated pairs having DPP-value<0.1 and DPES>0 were identified as similarly behaving paralog pairs while paralog pairs having DPP-value<0.1 and DPES<0 were considered as oppositely behaving paralog pairs.

2.3.7. Collection and processing of zebrafish datasets for CompariZome

GEO database [226] was searched for zebrafish expression datasets based on Affymetrix platform GPL1319, that contained 15,600 probesets. Datasets based on only two experimental groups with no requirement of the experimental type (i.e., development, drug treatment, morpholino etc) were selected for differential expression analysis to be incorporated into CompariZome. Search resulted in a local database for CompariZome version 1.1 consisting of 28 zebrafish and 3 corresponding human datasets (Table 2.11).

In order to obtain a differential expression signature for each dataset, data was processed as mentioned below:

- *GEOQuery* package [227] was used to download the raw data (i.e., CEL files) and series matrix files providing normalized expression data along with phenotypic descriptions.
- Phenotypic data were extracted from the series matrix files. Phenotypic information in the series matrix file contained multiple sources of information including sample description, experimental details, data submission and processing however for each dataset, I extracted the data which gave the information about experiment type in the dataset along with the identity of the control/treatment groups. Columns describing samples and experimental condition were retrieved by using the keywords ‘‘title’’, ‘‘description’’ and ‘‘characteristics’’.
- Raw data were evaluated for the quality of samples in a dataset using *affyplm* and *affyQCReport* [45] package. All datasets were RMA normalized using *affy* package [51].
- RMA normalized expression data were used to perform differential expression using *limma* package [56].
- Normalized expression data were processed to obtain one gene-one probeset relationships. In case of multiple probesets hitting a single gene, the mean expression of probesets was considered as representative expression of the corresponding gene.
- Examplary human expression data corresponding to the biological background of collected zebrafish datasets were searched from GEO database (GPL570 or GPL96

platforms only). Manually, retrieved datasets were assembled in a file having similar human-zebrafish expression GEO IDs pairs in each row. This remains an on-going effort.

- Human expression datasets were also downloaded and processed as in the case of zebrafish datasets following the above-mentioned pipeline comprising quality control assessment, RMA normalization and *limma* differential expression analysis.

In addition to expression data, annotation data were obtained and stored using the following steps:

- Zebrafish and human microarray probesets to gene annotation files were downloaded from Ensemble Biomart database [232].
- Human-zebrafish orthologs information was also retrieved from the homology module of Ensemble Biomart [214] by filtering human genes, having associated zebrafish orthologs.
- Gene ontology terms (biological processes, cellular component, molecular function) mapping to each zebrafish gene were also downloaded from Ensemble Biomart [214].
- Conserved microRNAs database and microRNAs mapping to zebrafish paralogs were downloaded from Targetscan [103].

2.3.8. Design of web application, CompariZome

Shiny R package was used to develop a web-application, called CompariZome (Comparative Zebrafish transcriptome analysis), for analysis of paralog expression in zebrafish in connection with corresponding human datasets [153]. *Ggplot2* package was used for plotting of bar and scatter plots [233] while *igraph* package [234] was used for interaction graphs (having nodes and edges). Using an automated pipeline mentioned in the previous section, datasets were downloaded, checked for quality, RMA normalized and *limma* analysis was performed; analyzed data were locally stored and only searched/visualized upon query entered by the user.

While running CompariZome with a user entered human gene symbol, following run-

time processes take place:

- Using human-zebrafish orthology file, the local database is searched for zebrafish orthologs and findings are displayed in ***Orthologs*** tab.
- In ***zebrafish expression module***, list of GEO IDs representing all zebrafish expression datasets are displayed and upon selection of aGEO ID, already prepared quality control files and description of samples/dataset are displayed.
- Affymetrix IDs of zebrafish paralogs are retrieved from zebrafish annotation file and *limma* analysis file of the respective dataset is then searched for differential expression of these matched Affymetrix IDs.
- Normalized expression of paralog pairs is retrieved and in case of multiple probesets for a gene, the mean expression is calculated. Mean expression along with standard deviation of each group (control and treatment) are used to draw the bar plots of paralog genes under different experimental conditions.
- Normalized expression value (i.e., mean expression for multiple probesets hitting a single gene) is used to perform *t.test* between paralog genes by taking all samples in the dataset without segregating on the basis of experimental condition. t.test statistics are described in a table and a scatter plot is used to visualize the expressions of paralog pairs on the x and y-axes, respectively.
- Human-zebrafish GEO ID file is searched for a corresponding human dataset (with a matching biological background) and the quality control, phenotypic data along with differential expression (for user entered gene) is retrieved from already stored human expression dataset files. Results are visualized in ***human expression module.***
- Local downloaded database of gene-ontology mapping (Ensemble Biomart) is searched for functional annotation of both genes in a paralog pair. Annotation is visually presented in ***functional annotation module*** such that circular nodes represent gene names of paralog pairs, square nodes represent gene-ontology term while edge represents mapping between ontology term and gene. *Igraph* package [234] is used for graphical representation of gene-ontology enrichment.
- For ***miRNA enrichment module***, local downloaded database of conserved

microRNA targets (retrieved from targetscan) is searched for both genes of paralog pair and are visually represented in such a way that nodes represent microRNAs (circular) and genes (squares) while the edges represent predicted hits of microRNA to the gene. MicroRNA conservation score is calculated by using the Jaccard index, provided in Eq 2.4.

$$(number\ of\ microRNAs\ hitting\ gene\ 1\ \cap number\ of\ microRNAs\ hitting\ gene\ 2\ ) \div \cup (total\ number\ of\ microRNAs\ hitting\ both\ genes) \qquad Eq\ 2.4$$

2.3.9. Instruction messages

HTML tags were used to provide small instructions about each module.

2.3.10. Making rds object of large files

Zebrafish and human gene annotation, orthology information and *limma* result files were all converted to RDS objects in R prior to loading in CompariZome. RDS objects are easier and faster to load as compared to csv/txt files which is time-consuming.

2.4. Results

Deviant pool of probesets in distinct species makes it difficult to perform interspecies transcriptome analysis. Using rapamycin treated zebrafish microarray expression profile (GSE53086) I have performed two different case studies by combining a) zebrafish-mouse drug (rapamycin) treatment profile and b) zebrafish-human ortholog-paralog transcriptome for rapamycin treatment.

2.4.1. Rapamycin treatment in zebrafish cell line

ZF4 cell line was exposed to 100nM rapamycin or DMSO for 48 hours and transcriptomic changes were analyzed by Affymetrix Zebrafish Genome Array (GPL1319) (Table 3.1; GSE53086). Differentially regulated transcriptome upon rapamycin exposure was analyzed using the BRB-Array tool (Ceren Sucularli, PhD thesis; [225]). Significance threshold of $p<0.05$ resulted in 2186 differentially expressed genes. Considering direction of logFC, 906 of these probesets were upregulated while expression of 1282 genes was downregulated upon rapamycin exposure. Microarray experiment and differential expression analysis was performed by Ceren Sucularli [225].

2.4.2. Case Study I: Comparison of rapamycin treated expression profile among model organisms (zebrafish and mouse)

Zebrafish and mouse are widely used model organisms and large numbers of transcriptomics studies have been performed in these species. Using zebrafish expression data generated in our lab (GSE53086) and mouse data obtained form public database GEO [226], I have compared the shared and divergent changes in transcriptomes in response to rapamycin between both of the model organisms.

2.4.2.1. Rapamycin treatment expression profile in mouse

Rapamycin treated mouse cell line expression data were collected from GEO [226].

Three datasets based on Affymetrix Mouse Genome 430 2.0 Array (GPL1261) were retrieved (Table 2.1) and were specified in detail below.

Table 2.1: Mouse datasets used in meta-analysis of rapamycin treated expression profile.

GEO ID	Sample type		Control (n) / treated (n)	Treatment time	Treatment dose
GSE5332	Mouse embryo fibroblast (tsc2-/-)		3/3	14 hours	20nM
GSE19885	Mouse brain tumor cell line BC3H1		3/3	24 hours	100nM
GSE21755	Mouse embryonic fibroblast	TSC2-/-, p53-/-	3/2	24 hours	20nM
		TSC1-/-	3/2	24 hours	20nM

1) TSC2 deficient Mouse Embryonic Fibroblasts (GSE5332)

In this study, Cunningham et al., have used a mouse embryonic fibroblast cell line, which is TSC2 mutant (TSC2-/-). TSC2 null mutation results in continuous activation of mTOR pathway in this cell line. The authors carried out 20nM rapamycin treatment for 14 hours in triplicates while using DMSO treatment as control. Transcriptomic changes were quantified by microarray analysis (Table 3.1). Raw data was downloaded and after performing RMA normalization using *affy* package [51], rapamycin control along with rapamycin treated samples were used in mouse data meta-analysis.

2) Mouse brain tumor cell line (GSE19885)

BC3H1 cell line, i.e., a brain cell line, has been used to identify the effect of rapamycin effect on transcriptome. Experimental setup has consisted of three groups 1) a resistant BC3H1 cell line treated with 1μM rapamycin for more than 6 months; 2) BC3H1 cell

line treated with 100nM rapamycin for 24 hours (rapamycin sensitive group); 3) a control group comprising of 24 hours DMSO treated cells. I have used two groups (control and rapamycin sensitive) with triplicate samples for meta-analysis. Results of quality control showed that one sample, i.e., GSM497116 from rapamycin sensitive group, was not of good quality and was not used in further analysis [235].

3) Mouse embryonic fibroblast (GSE21774)

GSE21774 is a public time-course rapamycin treatment study with samples collected after exposure of 2, 6, 12 and 24-hours; this study has been conducted in mouse embryonic fibroblast cell lines, which are null for either TSC1 or TSC2 to enable the constitutive activation of mTORC1 complex [228]. TSC2 null cell lines are also TP53 null. It is already known that TSC2 is a downstream gene of TP53 [236]. I have used the longest time-point, i.e., 24 hours rapamycin treatment, for meta-analysis considering the expression of both cell lines as separate datasets.

Number of samples along with dose of rapamycin treatment in different datasets have been described in Table 2.1.

2.4.2.2. Meta-analysis of rapamycin treated mouse expression data

Collected datasets were normalized with their respective control samples. Datasets originated from different sources (i.e. TSC1 mutant, TSC2 mutant, rapamycin sensitive, embryonic fibroblast or brain cells) so coherence in rapamycin-induced effect was evaluated by correlation coefficient analysis among the logFC values. All datasets showed significantly positive correlation with each other suggesting commonality in the effects of rapamycin regardless of genetic and tissue background differences (Table 2.2).

Table 2.2: Correlation among logFC of rapamycin treated mouse expression profiles.

	GSE21755 TSC1-/-	GSE21755 TSC2-/-	GSE19885	GSE5332
GSE21755 TSC1-/-		**0.49**	**0.21**	**0.39**
GSE21755 TSC2-/-			**0.5**	**0.59**
GSE19885				**0.59**

*Significant correlation coefficient ($p<0.05$) is indicated by bold font face.

The significant positive correlation among logFC values indicated that datasets were comparable and could be combined in meta-analysis. Random effect model was used for calculating meta-analysis effect-size because of different origin of data and possible heterogeneity in data. FDR<0.05 resulted in a large number of differentially expressed probesets (n=9659). Different thresholds were tried to extract the rapamycin meta-signature, i.e., FDR<0.05, FDR<0.01, FDR<0.005 and top 2186 probesets to evaluate robustness of signature at different threshold. Number of differentially expressed probesets in selected threshold was given in Table 2.3.

Table 2.3: Number of differentially expressed genes according to different meta-analysis threshold.

	Differentially expressed probesets	**Upregulated probesets**	**Downregulated probesets**
FDR<0.05	9659	4834	4825
FDR<0.01	6135	2846	3289
FDR<0.005	5019	2243	2776
Top 2186	2186	871	1315

Upregulated probesets show effect-size, i.e., mu.hat>0 and downregulated probesets have effect-size<0

2.4.2.3. Rapamycin regulated KEGG pathways in zebrafish and mouse

KEGG pathway enrichment was retrieved for the significantly modulated profile in zebrafish ($p<0.05$) and mouse meta-analysis (FDR<0.005). Mapping of significantly regulated zebrafish pathways with enrichment scores of respective pathways in mouse meta-signature revealed that multiple similar pathways were also significantly regulated in mouse. Concordantly upregulated profile was enriched with ribosomal components while concordantly downregulated pathways contained the proteasome, cell cycle and oxidative phosphorylation [225]. Complete list of KEGG pathways has been shown in Table 2 of Sucularli et al., 2016 [225].

2.4.2.4. Functional comparisons between zebrafish and mouse rapamycin treatment expression profiles

Interspecies transcriptome expression comparison based on microarray probesets has challenges to overcome. Instead of individual genes, I have used functional annotation comparison to identify concordant/discordant rapamycin mediated processes in the zebrafish and mouse.

Gene-ontology annotations (biological processes, cellular component and molecular functions) were retrieved for the significantly modulated zebrafish profile and mouse meta-signature using online tool DAVID [87]. Significantly modulated gene-ontology terms ($p<0.05$) were extracted from rapamycin treatment signature of both species. After counting shared and diverged ontologies ($p<0.05$) in profiles, Odds ratio and Fisher's exact test statistics was used to find association between processes regulated in both species.

a) Biological processes

Biological process enrichment analysis showed that the shared downregulated terms included cell cycle, DNA replication or glucose metabolic processes. Like KEGG pathways, enrichment of proliferative processes in the downregulated pathways indicated a strong shared anti-cancer potential of rapamycin in both species.

Comparison of biological processes revealed that the number of significant processes regulated in same direction was much higher than the ones regulated in the opposite direction (Same direction=72; opposite direction: 16; Table 2.4). These shared terms resulted in a highly significant positive association between biological processes of both species (OR=23.68, p=6.3E-08; Table 2.4). Shared upregulated terms included the development related processes, i.e., skeletal muscle development or embryonic development.

b) Molecular Function

Like biological processes, molecular function enrichment between both species resulted in a large number of concordantly regulated molecular functions in comparison to the ones regulated in the opposite direction (Shared functions: 24, opposite regulated functions: 7; Table 2.6). Cross tabulation statistics resulted in very high Odds ratio along with a positive association (OR=9.79, p=1.2E-02; Table 2.4). Common upregulated molecular functions included DNA and chromatin binding while downregulated ones included threonine type peptidase activity [225].

c) Cellular component

Comparison of cellular component ontologies did not result in significant association between mouse and zebrafish expression profiles (Table 2.4). Interestingly, ribosome and ribonucleoprotein complex came up as significantly upregulated terms in upregulated profiles of both species. Shared downregulated terms included proteasome, endoplasmic reticulum and mitochondria.

2.4.2.5. Robustness of comparative concordant signature

Ontology enrichment revealed that some of generalized terms like translation and cell cycle appeared in list of both significantly upregulated and downregulated terms. These

processes encompass wide-transcriptomic regulation and positive or negative regulators might justify enrichment in both upregulated and downregulated profiles. However, to avoid any kind of bias, the test strategy has been improved by removing duplicated terms.

Comparison of shared terms was repeated by removing these duplicated ambiguous terms, i.e., if any ontology term appeared in both the upregulated and downregulated category of given species, it was not used in the association analysis. Again, findings resulted in significant and improved positive association in biological processes (OR=253.65, p=9.1E-12; Table 2.4) and molecular function (OR=18.22, p=7.9E-03; Table 2.4) ontologies and with no significant association in cellular component ontologies. These findings highlighted robustness of the analysis in identification of concordant rapamycin treated expression profiles in zebrafish and mice.

Table 2.4: Contingency table and association statistics of significantly regulated ontology terms ($p<0.05$) [225].

Rapamycin regulated zebrafish ($p<0.05$) and mouse profile (FDR<0.005).

		Mouse			
		All terms		**Duplicated terms Removed**	
Zebrafish	**BP**	Down	Up	Down	Up
	Downregulated	54	13	43	2
	Upregulated	3	18	1	16
	P-value	**6.3E-08**		**9.1E-12**	
	OR	23.68		253.65	
	CC	Down	Up	Down	Up
	Downregulated	23	8	17	3
	Upregulated	12	8	3	0
	P-value	0.36		1.00	
	OR	1.89		0.00	
	MF	Down	Up	Down	Up
	Downregulated	18	2	17	1
	Upregulated	5	6	4	5
	P-value	**1.2E-02**		**7.9E-03**	
	OR	9.79		18.22	

Another important aspect in comparison between different data sources has been the selection of a significance threshold. Setting the hard-threshold in comparison studies can give rise to different number of probesets passing the threshold in independently performed analysis. P-value<0.05 is widely accepted significance threshold. Zebrafish expression profile resulted in 2186 probesets using p<0.05, however, meta-analysis (FDR<0.05) resulted in a larger number of probesets. Unbalanced number of significantly regulated probesets can lead to results that are biased toward the study having more number of probesets passing the threshold. To address the issue, analysis was repeated for three additional meta-analysis thresholds. Two of them were hard thresholds using FDR, i.e., FDR<0.05, FDR<0.01, while the third one was rank-based to equalize the number of probesets. Expression profile having lesser number of significant probesets (zebrafish: 2186 probesets) was used as reference and mouse meta-analysis having large number of significantly regulated probesets was ranked to obtain top 2186 probesets. Analysis was performed for all shared terms in BP, CC and MF and was repeated by removing the duplicated terms. Results again confirmed the significant positive association in BP and MF at all three thresholds however CC remained non-significant (Table 2.5). These findings reveal robustness of concordantly regulated processes in rapamycin treated zebrafish and mouse profiles.

Table 2.5: Validation of robust rapamycin regulated profile in zebrafish and mouse using different meta-analysis thresholds.

A) All terms

		Mouse					
		FDR<0.05		**FDR<0.01**		**Top 2186**	
Zebrafish	**BP**	Down	Up	Down	Up	Down	Up
	Down	56	7	53	11	42	9
	Up	4	30	4	23	2	10
	P-value	**3.1E-14**		**1.2E-09**		**3.4E-05**	
	Odd ratio	55.40		26.25		21.73	
	CC	Down	Up	Down	Up	Down	Up
	Down	29	8	25	8	17	9
	Up	13	9	13	5	10	4
	P-value	0.14		1.00		1.00	
	Odd ratio	2.47		1.20		0.76	
	MF	Down	Up	Down	Up	Down	Up
	Down	24	11	25	3	17	2
	Up	6	8	5	6	5	5
	P-value	0.12		**7.8E-03**		**3.0E-02**	
	Odd ratio	2.84		9.22		7.75	

B) Duplicate Terms removed

		Mouse					
		FDR<0.05		**FDR<0.01**		**Top 2186**	
Zebrafish	**BP**	Down	Up	Down	Up	Down	Up
	Down	49	0	43	1	35	2
	Up	1	27	1	20	1	9
	P-value	**6.4E-20**		**1.5E-14**		**3.8E-07**	
	Odd ratio	Inf		531.93		116.67	
	CC	Down	Up	Down	Up	Down	Up
	Down	22	2	18	2	12	4
	Up	3	0	7	0	6	0
	P-value	1.00		1.00		0.54	
	Odd ratio	0.00		0.00		0.00	
	MF	Down	Up	Down	Up	Down	Up
	Down	14	1	23	1	16	1
	Up	2	4	4	5	4	4
	P-value	**1.1E-02**		**2.8E-03**		**2.3E-02**	
	Odd ratio	21.55		24.56		13.81	

2.4.3. Case Study II: Comparison of zebrafish duplicated genes in their response to rapamycin

As described in the introduction, zebrafish genome has undergone an additional whole genome duplication event and as a result, multiple copies of genes are present and can exhibit divergent regulation [237]. Zebrafish is widely used as a model organism and multiple cancer related research trials are heavily conducted on zebrafish [238]. While analyzing expression modulation in anti-cancer drugs, e.g., rapamycin, on zebrafish, it is important to observe whether duplicated copies of genes are affected in the same or the opposite manners.

2.4.3.1. Identification of duplicated gene pairs

Biomart homology tool was used to retrieve human-zebrafish orthologs and was processed in R as described in materials and methods section 2.3.4. *Limma* analysis was performed in R to identify differentially regulated expression profile. Mapping the orthology information with rapamycin experiment probesets from zebrafish Affymetrix arrays resulted in 606 duplicated pairs.

2.4.3.2. Conserved/diverged expression patterns among paralog pairs in response to rapamycin

When 606 duplicated pairs were plotted against each other, a mild but significant positive correlation was found (R=0.25, p=9.5e-10; Figure 2.1) revealing that rapamycin mediated the changes in these duplicated pairs, similarly. However, there were considerable number of paralogous pairs either behaving in the opposite direction or only one gene from the pair was significantly modulated by rapamycin exposure. Considering duplicated pair p-value (*limma* p-value of gene1 + p-value of gene2; described in materials and methods) threshold of 0.1, we found that 37 paralog pairs showed differential expression in rapamycin treatment. These 37 paralog pairs belonged to group where both genes of paralog pairs were significantly modulated. Considering

paralog pair effect-size (discussed in materials and methods), 32 of 37 behaved concordantly, i.e., both genes behaved in the similar direction while the remaining 5 paralog pairs were regulated in the opposite direction, i.e., logFC direction of both genes in these paralog pairs were different (Table 2.6). Association statistics showed that number of paralog pairs concordantly modulated were much more than the ones acting in opposition (OR=33.84, p=1.76e-05, Table 2.6). The descriptions of these 37 pairs were provided in Table 2.7. Besides 160 out of 606 paralog pairs showed significant regulation in one gene ($p<0.05$) with non-significant modulation in the other gene of the paralog pair.

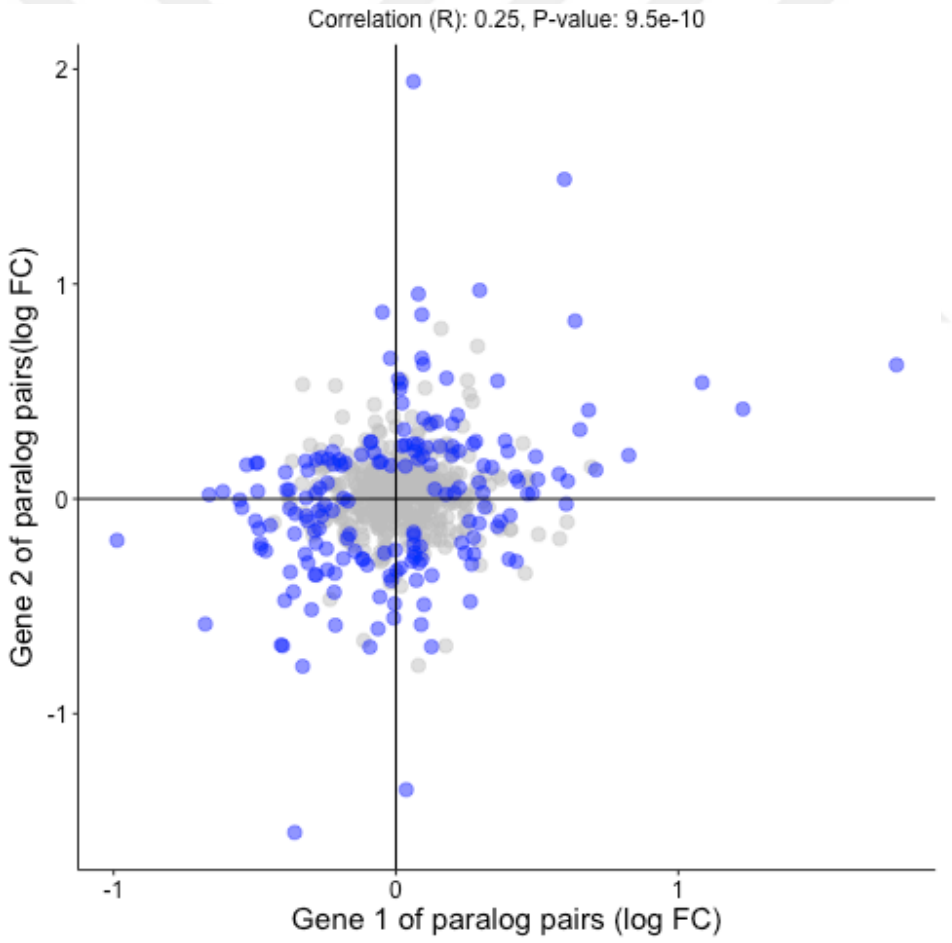


Figure 2.1: Comparison of paralog pairs in rapamycin treated zebrafish profile.

Each dot represents one paralog pair. Blue color indicate that either one or both gene of paralog pair is differentially expressed, i.e., limma $p<0.05$. Grey color indicates both genes of pair are non- non-significantly regulated.

Table 2.6: Association statistics of significantly regulated paralog pairs (paralog p<0.1).

	Up	Down
Up	12	2
Down	3	20
Odd ratio	33.84	
P-value	1.76E-05	

Table 2.7: Duplicated gene pairs regulated in same direction in rapamycin treated zebrafish expression profile.

Gene name	LogFC	P-value	Gene name	LogFC	P-value	Paralog P-value	Paralog Effect-size
nr2f6b	0.39	7.17E-02	nr2f6a	0.27	5.93E-03	7.76E-02	0.32
kdelr2b	-0.29	1.11E-02	kdelr2a	-0.35	4.07E-03	1.51E-02	0.32
got2a	-0.48	5.33E-03	got2b	-0.21	2.46E-02	3.00E-02	0.32
prc1b	-0.33	5.90E-02	prc1a	-0.78	4.57E-04	5.94E-02	0.51
skia	0.30	8.42E-02	skib	0.97	1.28E-03	8.55E-02	0.54
hsp90aa1.2	-0.27	9.42E-03	hsp90aa1.1	-0.14	7.65E-02	8.59E-02	0.19
thbs1b	0.36	5.82E-02	THBS1 (1 of many)	0.55	7.19E-03	6.54E-02	0.44
cry1aa	-0.41	3.35E-02	cry1ab	-0.68	5.94E-02	9.29E-02	0.53
FOXK2 (1 of many)	0.58	1.79E-02	foxk2	0.12	8.12E-02	9.90E-02	0.26
anxa13	0.68	5.17E-03	anxa13l	0.41	9.10E-02	9.62E-02	0.53
eno1a	-0.28	1.73E-02	eno1b	-0.35	4.87E-03	2.22E-02	0.32
rrbp1a	-0.36	3.15E-03	rrbp1b	-0.43	2.60E-02	2.91E-02	0.40
atp5j	-0.19	2.95E-02	si:ch211-140m22.7	-0.28	3.52E-02	6.47E-02	0.23
ndufab1b	-0.39	4.15E-02	ndufab1a	-0.47	1.14E-03	4.26E-02	0.43
hsd17b12b	-0.36	4.10E-03	hsd17b12a	-1.55	5.63E-06	4.10E-03	0.75

Gene name	LogFC	P-value	Gene name	LogFC	P-value	Paralog P-value	Paralog Effect-size
rpl5b	0.28	3.57E-03	rpl5a	0.27	1.68E-02	2.04E-02	0.28
mdh1aa	-0.40	9.64E-04	mdh1ab	-0.68	3.73E-03	4.70E-03	0.52
eif4g2b	-0.37	1.64E-02	eif4g2a	-0.34	3.09E-03	1.95E-02	0.36
tnc	0.34	2.27E-02	TNC (1 of many)	0.15	6.40E-02	8.67E-02	0.22
msrb1a	-0.21	1.97E-02	msrb1b	-0.59	1.53E-03	2.12E-02	0.36
eef1db	0.63	9.48E-04	eef1da	0.83	5.57E-05	1.00E-03	0.72
psmc1a	-0.32	4.20E-03	psmc1b	-0.26	3.74E-02	4.17E-02	0.29
btbd6b	0.60	8.86E-02	btbd6a	1.49	1.52E-03	9.01E-02	0.94
arpp19b	-0.29	3.23E-03	arpp19a	-0.15	5.15E-02	5.47E-02	0.21
tagln	1.77	1.24E-04	tagln2	0.62	7.24E-04	8.48E-04	1.05
cited4b	1.23	7.40E-04	cited4a	0.42	6.06E-04	1.35E-03	0.72
irs2b	1.08	2.45E-04	irs2a	0.54	3.75E-02	3.77E-02	0.77
ckbb	-0.99	1.81E-02	ckba	-0.19	7.07E-02	8.88E-02	0.44
porb	-0.45	1.42E-02	pora	-0.12	7.26E-02	8.68E-02	0.23
CPEB4 (1 of many)	-0.30	8.11E-03	cpeb4	-0.52	1.00E-03	9.11E-03	0.39
sox11a	-0.68	3.09E-02	sox11b	-0.58	1.03E-02	4.12E-02	0.63
cox5aa	-0.24	2.26E-02	cox5ab	-0.33	2.20E-03	2.48E-02	0.28

Table 2.8: Duplicated gene pairs regulated in opposite direction in rapamycin treated zebrafish expression profile.

Gene name	LogFC	P-value	Gene name	LogFC	P-value	Paralog P-value	Paralog Effect-size
cbx1b	0.27	2.11E-02	cbx1a	-0.30	1.24E-02	3.35E-02	-0.29
fn1a	-0.49	2.39E-02	fn1b	0.17	3.39E-02	5.78E-02	-0.29
pak2b	-0.22	1.47E-02	pak2a	0.22	6.43E-02	7.89E-02	-0.22
tefa	-0.53	2.97E-02	tefb	0.16	6.70E-02	9.66E-02	-0.29
appa	0.28	1.68E-02	appb	-0.26	4.24E-02	5.92E-02	-0.27

2.4.4. Case Study III: Development of a database for human-zebrafish expression comparison: CompariZome

Up to date there is no online tool, which addresses similar/deviant gene expression of duplicated zebrafish genes pairs and their conserved/diverged expression association with respective human orthologs. Using transcriptomics data from Zebrafish Affymetrix arrays, we have developed a tool called 'CompariZome: COMPARIson of Zebrafish transcriptOME' which can facilitate scientists to query any human gene to find the orthologous gene expression in zebrafish and functional annotation of the selected zebrafish paralog pair. CompariZome also facilitates the comparison of expression changes of the queried gene in the respective human data obtained under similar conditions.

2.4.4.1. Data collection

Human and zebrafish gene expression data collection for CompariZome_V1 (version 1) have been described in material and methods section. Accordingly, 28 zebrafish datasets have been obtained as a result of two-group expression search of GEO database (as of June, 2016). From the collected cohort of datasets, 7 datasets belonged to drug treatment

while 10 belonged to knockdown or overexpression of studies of a gene. Zebrafish and human datasets included in CompariZome_V1 were described in Table 2.9.

Table 2.9: Datasets included in CompariZome_V1 (version 1).

	Dataset	GEO Title	Number of samples in experimental groups
Zebrafish	GSE55339	Uhrf1 mutant expression profile at 120hpf	2 (UHRF1 Mutant) / 3 (wild type)
	GSE61062	Morphine treatment in zebrafish embryos	6 (Control Embryo) / 6 (Embryo exposed to Morphine)
	GSE55540	Expression profile of zebrafish heart after reatment with FGFR inhibitor (SU5402)	2 (wild type) / 2 (SU5402 treated)
	GSE56365	Exposure to live Salmonella enterica serovar Typhimurium bacteria at 2dpf	3 (Control Embryo) /3 (infected Embryo)
	GSE46844	Gene expression of multiciliated cells after miR-34b knockdown	2 (Control) /2 (miR-34B morphants)
	GSE41059	Gene expression of zebrafish liver in wild type and Def (Digestive-organ expansion factor) overexpressing transgenic cell line	3 (Control) /3 (Tg(fabp10a:def))
	GSE43186	Effect of glucose metabolism by glucose treatment at 36hpf	2 (Control) / 2 (glucose treated)
	GSE39707	Treatment of epoxyeicosatrienoic acid in zebrafish at 36hpf	3 (DMSO) / 3 (hpf EET)
	GSE35889	Comparison between wild type and zCdx1b transgenic zebrafish (zCdx1b is otholog of mammalian cdx2)	2 (wild type) / 3 (transgenic)
	GSE31186	Comparison of small round blue cell tumors with malignant peripheral nerve sheath tumors in zebrafish	3 (SRBCT) /3 (MPNST)
	GSE31185	Comparison of zebrafish wild type strain with transgenic zebrafish expressing Ewing's Sarcoma onvoprotein EWS-FLI1	3 (Control) / 3 (Tg(HSP:EWS-FLI1))
	GSE32092	Comparison of zebrafish brain expression in	2 (Control) / 2 (Experiment)

Dataset	GEO Title	Number of samples in experimental groups
	normal and cold stress	
GSE17711	hi559 mutant zebrafish larvae comparison with wildtype	3 (Wild Type) /3 (Mutant)
GSE24527	Leflunomide treatment in 24 hpf zebrafish embryo	3(DMSO)/3(LEFLUNOMIDE)
GSE28199	Comparison of prdm1a mutant zebrafish with wild type	3 (Wild Type) / 3 (Mutant)
GSE21539	Expression analysis of ova1 mutant in zebrafish	3 (Control) / 3 (Ovo1 morphant)
GSE16857	Zebrafish larvae at 6 days post-fertilization in presence/absence of microbiota	3 (Germ free) /3 (Conventionalized)
GSE14718	Exposure of aromatase inhibitor Fadrozole in zebrafish telencephalon	3 (Control) / 3 (Treated)
GSE10766	Expression of caudal fin after Beclomethasone treatment (2 day post fertilization)	3 (Control) /3 (Beclo)
GSE13999	Light exposure experiment for evaluation of heparin-binding growth proteins (midkine-a and midkine-b) during retinal development	4 (Control) / 4 (Light lesion)
GSE12955	Lymphoild cells obtained from kidney are divided in two populations, i.e. SP and MP	2 (zSP1) /2 (zMP1)
GSE12826	Gene expression of SPT mutant gene (sk8) in 24hpf zebrafish	3 (wild type) / 3 (Mutant sk8)
GSE7658	Gene Expression Profile of Hematopoietic Stem Cells during Zebrafish Development	4 (GFP positive) / 4 (GFP negative)
GSE5048	Gene Expression Profiling of Zebrafish Embryonic Retinal Pigment Epithelium in vivo.	3 (Wild type) / 3 (Wild type retina epithelium attached)
GSE4989	Effect of hypoxia on zebrafish heart	5 (hypoxia) / 5 (normoxia)
GSE1995	Gene expression in foggy mutant	1 (WT) / 3 (foggy mutant)
GSE32914	Study of Mid-Blastula Transition of zebrafish embryos	3 (4.3 hours post-fertilization) / 2 (2.5 hours post fertilization)

	Dataset	GEO Title	Number of samples in experimental groups
	GSE53086	Effects of rapamycin exposure on zebrafish ZF4 cells (Performed by Ceren Sucularli)	3 (DMSO) / 2 (Rapamycin treatment)
Human	GSE67338	Expression data of human neuroblastoma IMR32 cells treated with the DHODH inhibitor Teriflunomide [118 μM]	3 (DMSO) / 3 (Terfilunomide)
	GSE30478	PHD finger recognition of unmodified histone H3R2 links UHRF1 to regulation of euchromatic gene expression	1 (Scr_shRNA) /2 (UHRF1 shRNA)
	GSE18571	Rapamycin treatment of MDA-MB-468 breast cancer cell line and MDA-MB-468 xenografts	3 rapamycin treated/3 DMSO
	GSE1825	SRBCT-Comparison between neuroblastoma samples and Ewing family tumor samples	5 (Ewing family Tumor Samples) / 5 (Neuroblastoma)

2.4.4.2. CompariZome input interface

Home page of CompariZome has an input panel on the sidebar where the user can enter any human gene symbol. All of the 5 modules are shown as tabs on the home page along with the 6th tab representing meta-analysis, which is not incorporated yet and remains a future perspective (Figure 2.2).



Figure 2.2*:* Home page/Input interface of CompariZome database.

2.4.4.3. Modules in Comparizome

CompariZome is comprised of 5 modules. First module describes the human-zebrafish orthology association. Next two modules are based on gene expression analysis of zebrafish and human data while last two modules are based on functional annotation and regulatory associations between paralog pairs, i.e., gene-ontology and miRNA enrichment, respectively.

Module I: Orthologs

After entering a human gene symbol and pressing search, orthology information (retrieved from Biomart, described in materials and methods) is displayed under 'Orthologs' tab. Visualized orthology information includes description of human queried gene (Human Ensembl ID, human gene symbol, description of gene symbol) and respective zebrafish orthology information including gene symbol, confidence on orthology (0 indicate low while 1 indicates high confidence), type of relationship, i.e., one-to-one, one-to-many along with identity between human and zebrafish gene

sequence. Fibronectin (FN1) is an important extracellular matrix protein and has been shown to be associated with angiogenesis [239]. Zebrafish duplicated pair consist of FN1a and FN1b (Figure 2.3). Rapamycin treatment showed mild but opposite regulation of both genes in paralog pair (Table 2.8). Complete case study of FN1 analysis showing functionality of different CompariZome modules is shown in Appendix-I.

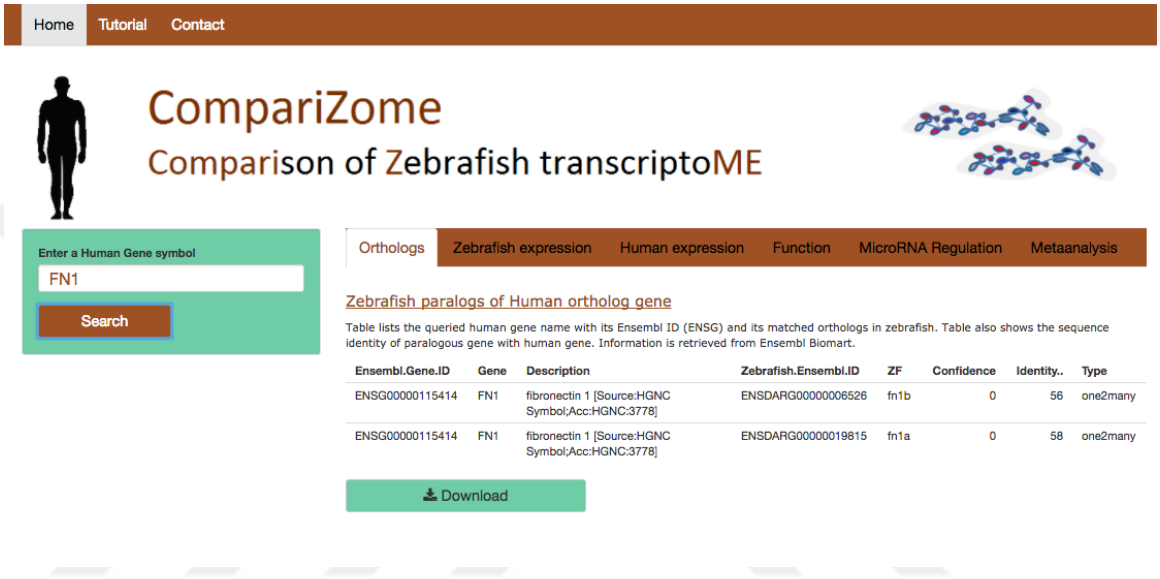


Figure 2.3: Orthologs module showing input human gene and corresponding zebrafish ortholog genes.

Module II: Zebrafish expression module

Switching to 'zebrafish expression' tab adds the list of zebrafish expression IDs to the side bar of the input panel. By default, GSE55339 is selected. 'Zebrafish expression' module has three further sub-modules, which includes *phenotypic data*, *quality control* and *expression analysis.*

Phenotypic data: User can select any dataset from the list in the input panel and by clicking the search button; description of the dataset can help user select data of his/her own interest. Phenotypic data can give the user insights about experimental setup of the selected dataset and experimental groups, number of samples in each group and origin of data.

Quality control: Quality control is an important preliminary evaluation criterion. Several of quality control measures are visualized that includes RNA degradation plot,

Relative Log Expression (RLE), Normalized Unscaled Standard Errors (NUSE), log intensity vs density plots, actin3/actin5, gapdh3/gapdh5 levels, percent present, positive and negative border elements, and the correlation coefficients between arrays.

Expression analysis: Expression module is further categorized in two types of output:

i) Experimental group comparison is aimed to visualize the differential expression of paralogous genes between the control and experimental condition provided in the dataset. This analysis helps user observe concordant/discordant modulation of expression of different genes making up the paralog pairs under the specific condition(s)/treatment(s). *Limma* differential expression analysis of paralogous genes was shown in the form of a table and includes log fold change, p-value and adjusted p-value (FDR) of the associated probesets for.gene modulated by rapamycin treatment. Besides, bar plot of expression values in given experimental conditions are also presented. (Case study in Appendix-I).

ii) Paralog comparison is destined for differential expression between paralog genes. This module considers all samples without defining a control or an experimental group and performs differential expression analysis between paralog genes (t.test). Scatter plot with expressions of both paralogous genes on the x and y-axis, respectively can also be visualized (Case study in Appendix-I).

Module III: Human expression

Based on selected zebrafish dataset, human expression module lists analysis of matched human expression data (Table 2.9). This matching is done manually based on underlying biological problem (described in materials and methods) and is linked with selection of zebrafish dataset. If there were multiple human GEO datasets corresponding to the zebrafish dataset then the one with most relevant biological background was retained.

For example, upon manual searching of rapamycin cancer treated datasets, I retrieved three rapamycin treatment datasets on breast cancer cell lines. 1) GSE36847 was based on rapamycin treatment in MCF7 cell line (12 hours; 100nM) in MCF7; 2) GSE18571 included rapamycin treated MDA-MB-468 cells (24h; 100nM) and 3) rapamycin treatment on three cell lines, i.e., SKBR3, C5.2, HB4a with treatment profile of 20nM

and 24 hours were analyzed in GSE26599. Upon manual evaluation, GSE18571 was considered comparatively similar to zebrafish rapamycin treatment study (GSE53086) as both datasets were treated with 100nM and time-point of 24 hours treatment in human dataset was closer to 48 hour exposure in zebrafish expression profile.

Like the zebrafish expression module, this module also has three sub-modules allowing for visualization of quality control results; phenotypic data associated with dataset; and *limma* results of the selected gene. However, in this case, results are generally displayed for single gene, i.e., human ortholog of the gene of interest. The example for human and zebrafish rapamycin modulated gene was *IRS2 gene,* which has duplicated pair in zebrafish. Based on previous analysis (Case study II), both genes of *IRS2* duplicated pair were upregulated upon rapamycin treatment (Table 2.7). Analysis of human rapamycin treatment expression profile revealed milder regulation, which could not reach significant expression difference between treated/untreated groups (Figure 2.4).

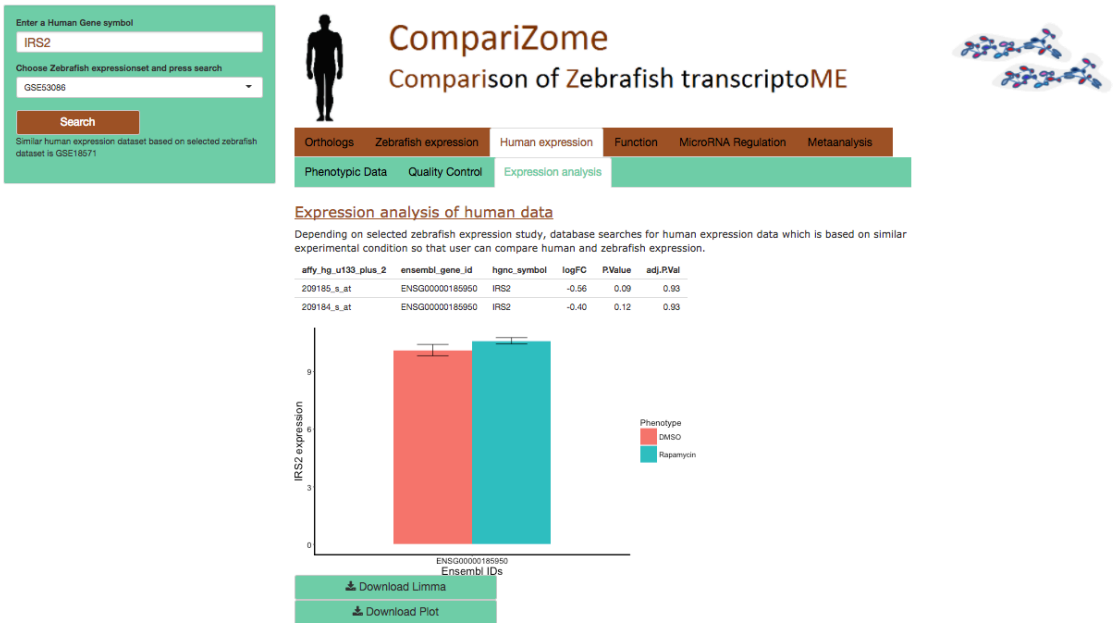


Figure 2.4: IRS2 expression analysis in human rapamycin treated breast cancer dataset.

Module IV: Gene ontology enrichment

Gene ontologies for each paralogous zebrafish gene retrieved from Biomart have been used in the module helping user identify common/distinct gene ontologies associated with the duplicate gene pairs. Three types of ontologies, i.e., biological processes, cellular component and molecular functions are visualized individually. Both genes and ontology terms are represented by nodes but are differentiated such that genes are represented with green circular nodes and ontologies with purple squares. Edges between both gene and ontology term indicates the link between term and gene.

Module V: MicroRNA regulation

MicroRNA enrichment data are collected form Targetscan database can be visualized using this module thus helping identify differential microRNA regulation between the duplicated pairs. Visual presentation includes nodes representing genes (orange squares) and microRNAs (yellow circles) and edges indicating the predicted targets of microRNAs. Moreover, a shared microRNA regulation score is calculated using the Jaccard index.

2.5. Discussion

Rapamycin treatment expression profile obtained in this study was based on zebrafish. Comparing it with another model organism, i.e., mouse, revealed that drug treatment on both model organisms was significantly shared. Gene expression comparison generally is based on significant number of shared genes but in this study, I have first performed pathway enrichment analysis on significantly regulated genes in zebrafish and mouse, individually. *MetaPath* package is also based on same underlying concept, i.e., comparison of different datasets based on their functional enrichment supporting robustness of our approach [137].

As a previous study has also reported that gene based analysis might not be a good criterion for transcriptome between species [191], my analysis have verified that shared inter-species transcriptomic changes can indeed be explored via pathway comparisons. The comparison of rapamycin treatment profiles between zebrafish and mouse has resulted in identification of proteasome being highly downregulated in both species. Proteasome inhibition has already been used as a therapeutic target in cancer [240]. The other pathways significantly downregulated in both zebrafish and mice included cell cycle and oxidative phosphorylation. The findings of this study suggested that anti-cancer activity of rapamycin has been conserved in zebrafish and mouse.

Due to easy breeding and fast generation time, zebrafish has been widely used as a model organism for the identification of drug mediated effects. However, duplicated genome of zebrafish can complicate the situation as WGD (whole genome duplication) can lead to sub- or neo- functionalization hence it is important to observe expression of both genes in duplicated pairs. This study signifies the importance of duplicated gene pair analysis while using zebrafish as a model organism. Comparison of duplicated gene pair profile in rapamycin treated zebrafish transcriptome revealed that duplicated genes largely behaved in same manner upon rapamycin treatment. However, the strategy presented for duplicated gene pair analysis can be used for any kind of treatment/group comparison.

Recent studies show that toxicogenomics comparisons between zebrafish embryos and

mammalian models are appropriate [241]. Previously multiple studies have been performed that compare zebrafish gene expression profiles with those of human and mouse. Nevertheless, comparative transcriptomics of drug exposure and/or toxicology has not been very prevalent when zebrafish is concerned. A recent study comparing human keratinocytes and larval zebrafish expression when exposed to H2O2 revealed small but a significant number of genes as similarly regulated (n=23) between the species [194]. However, this study did not involve meta-analysis of H2O2 exposure neither in humans nor zebrafish. Similarly, in another study comparing the microbiota suppression of HNF4A activity in human Inflammatory Bowel Disease patients and zebrafish using comparative transcriptomics [192], only human orthologs (one-to-one) to zebrafish orthologs (only one-to-one to prevent problems that might originate from duplicate genes) were compared. This study that was published later than ours also used an associative analysis based on hypergeometric tests and compared zebrafish and human genes (but not pathways) across the individual datasets, exhibiting the importance of application of proper meta-analysis methodologies in comparative genomics. Another meta-analysis study that involved zebrafish along with other organisms focusing on neuronal connectivity and social behavior [242]. In this study, IFNgamma expression signatures from multiple organisms including zebrafish, in social behavior and neural concepts, have been compared using GSEA (gene-set-enrichment-analysis) as a meta-analysis approach [242]. The rapamycin study presented in this thesis differs from the perspective that first a mouse rapamycin signature based on effect size is obtained by meta-analysis and then compared with the logFC differences of zebrafish not on the gene level but via KEGG and GO terms. Proper statistical measure of Fisher's exact tests has been used to identify positive-/negative association between species transcriptomes. This approach allows, for the first time, identification of common and divergent pathways between fish and mouse in response to rapamycin with high confidence. Another interesting study has focused on meta-analysis of ovarian transcriptome from different fish species to identify heterogeneity associated with expression [243]. This study was interesting at it showed that there was batch variability increasing heterogeneity even within the same tissue and within fish species supporting the idea that pathway comparisons might produce a more reliable outcome when

comparing across species [243].

In this thesis, I have shown that comparative analysis might not be enough in case of transcriptomic studies on zebrafish model organism. WGD event in zebrafish should also be considered and probable heterogeneity in duplicated gene pairs should also be analyzed. CompariZome has been designed for this purpose and can help evaluate expression of both genes in same duplicated pair. In the literature, one of the first attempts that allows for mapping across transcriptomes including human, mouse, and zebrafish in a database format, is TRAM [244]; this software has been published in 2011. In the last decade, the genomics information accumulated at an incredibly high rate thus it is important to have the most recent updates on chromosomal assemblies as well as annotations, which include Gene Ontology (GO), KEGG, mircoRNA targets, transcription factor binding sites. Recently, a comprehensive meta-analysis of toxicogenomics studies obtained on zebrafish embryos (based on concentration, time of exposure, mode of action) has been performed [245]. This is an excellent attempt in understanding the shared and divergent pathways/genes across a wide range of drug exposures in zebrafish. However, it still lacks an online platform and no further analysis can be performed. Since above study is not an interspecies comparison the study presented in this thesis remains a pioneer in the analysis of a user selected zebrafish dataset in comparison with a corresponding human dataset as well as addressing expression differences along with pathway and miRNA target differences among the paralogous pairs. Multiple studies have performed identification of duplicated gene pairs and presented potential databases (mentioned in introduction) of duplicated gene pairs. However, what makes CompariZome unique is the link between expression data and duplicated genes, not only in zebrafish but also in matching human dataset. In addition, CompariZome can enable analysis of duplicated gene pairs in biologically diverse conditions, e.g., drug treatment and knockdown studies.

Zebrafish studies that question the importance of paralogs at the transcriptomics level are not many. Among these Kozol et. al. (2016) evaluated the problem from a neurological point of view and tested the prevalence of duplication with respect to different neurodevelopmental and neurodegenerative diseases [246]. Furthermore, there

are many proteomics studies becoming available in zebrafish, which provides new avenues to expand on CompariZOME by incorporating proteomics [247].

However, duplicated genes divergence might not only be associated to gene expression but several other factors, e.g., tissue specificity and microRNA binding can contribute to diverged behavior. One recent study has shown that evolution of 3' UTR in zebrafish has led to varying poly-adenylation lengths and genes having smaller lengths were more expressed [248]. Similarly, CompariZome has microarray data stored in its first version. Paralog pairs have high sequence similarity among themselves. Microarray probe design limits the identification of duplicated gene pairs. Rapamycin data resulted in 606 duplicated pairs. Secondly, probeset design of both genes in paralog pairs might be hitting the homologous region and observed similar expression might not reveal the true underlying picture. For this reason, CompariZome will be updated in the future to incorporate RNAseq data.

2.6. Future Perspectives

Above described CompariZome database represents a first version and can be extended to several aspects of duplicated gene pair. Wide range of datasets having multiple groups are planned to be added. Two-group *limma* has been used as the statistical analysis package in this version 1. However, *limma* also enables multi-comparison analysis by modification of contrast matrix to incorporate several combinations. Similarly, the addition of RNAseq data can help evaluation of a larger number of duplicate pairs. Since infrastructure of CompariZome has an automated pipeline for addition of new datasets with minimal effort; addition of new datasets would not be difficult and are planned in the next version.

Another module is under development by which the use of database will not be restricted to available datasets in CompariZome. This module is based on the concept that user can enter GEO ID and select groups to be compared for evaluation of duplicated gene pairs in an interactive way. GEO ID input of a particular study will start the automated pipeline of downloading of phenotypic and expression data, RMA normalization and presentation of phenotypic properties to 'Phenotypic data' tab so that the user can assign samples to experimental groups that they want to compare. Afterwards, *limma* analysis could be automatically performed and results displayed in the respective module.

At the moment CompariZome can only display analysis for one gene at a time. CompariZome update will also enable visualization/download of study-based results in addition to already present gene-based analysis.

Duplicated gene score and p-value presented in Eq 2.2 and 2.3 will be reported in the results of analysis to conclude convergent/divergent expression pattern of duplicated gene pair expression.

Graphics used for gene ontology and microRNA enrichment can be improved to present an interactive network graphs. It will also be important in the future to incorporate non-

coding RNA information and RNAseq datasets into CompariZOME to fully understand the extent of transcriptomic/pathway conservation between species.

Meta-analysis module of CompariZome will also be designed such that it can enable selection of multiple studies by user and combine *duplicated gene score* (Eq 2.2) and *duplicated gene p-value* (Eq 2.3) of each pair from all selected datasets. This module can enable homogeneous/heterogeneous effect of multiple drugs, different cancer types, and developmental stages on duplicated gene pairs.

Chapter 3

Correlation analysis of GRHL3, HNF4A and CDH1 in CCLE datasets

3.1. Introduction

3.1.1. Use of CCLE in cancer research

CCLE (Cancer Cell Line Encyclopedia) represents a large cohort of cancer cell line data belonging to different tissues and has been described in section 1.6.4. CCLE data have been extensively used in previous years to reveal hidden cancer patterns using gene expression, copy number, mutation or pharmacological data. Few of these studies are exemplified below:

CCLE data have been used for prediction of novel subtypes and identification of new and interesting concepts on the basis of genome-wide changes. Epstein-Barr virus (EBV), Kaposi's Sarcoma Herpes viruses (KSHV) and human T-lymphotropic virus type 1 (HTLV-1) are generally considered as important contributors in T and B-cell lymphoma [249]. A recent study has used RNAseq of only lymphoma cell lines from CCLE and performed customized alignment with 740 mammalian virus genomes to identify virus-host interactions taking place during lymphoma progression [249]. Pharmacological data of GDSC (Genomics of drug sensitivity in cancer) and CCLE also showed good correlation indicating reliability of CCLE data [250]. Using 799 samples and various available data sources of CCLE including mutation, expression, copy number LOH (loss of heterozygosity), compound sensitivity and gene set activity data, functional status of 69 tumor suppressor genes was classified [251]. Four subtypes of gastric cancer were classified on the basis of their prognostic importance. For this, gastro-esophageal cell line data from CCLE were examined for validation of gastric

cancer subtypes. Interestingly, CCLE classification exhibited under-representation of TP53 active subtype in these cancer cell line models [252].

CCLE data has been used also as a translation model for confirmation of experimental findings based on single or cluster of genes in a tissue of interest. For example, inhibition of bromo-domain was found to suppress IL7R mediated proliferation in B-cell acute lymphoblastic leukemia (B-ALL) and thus, was considered as a potential therapeutic target. B-ALL cell lines (CCLE) were used to evaluate the expression of IL7R in various tissues and the findings revealed that B-ALL indeed exhibited highest expression of IL7R in comparison to other cell types [253]. PVT1, a long non-coding RNA gene, has been known as an oncogene. Knockdown of PVT-1 in colorectal cancer cell lines resulted in reduced proliferative potential [254]. Using colorectal cancer cell line copy number and gene expression data from CCLE, a positive correlation between 8q24 copy-number gain and *PVT-1* expression was shown which highlighted the role of 8q24 in colon cancer [254]. In another recent study, a subset of melanoma tumors exhibited overexpression of *PGC1α*. It was shown that oncogenic melanocyte lineage-specification transcription factor MITF has regulated the expression of PGC1*α*. This finding was verified by using gene expression data of CCLE; correlation of PGC1*α* with MITF and downstream targets of MITF exhibited positive correlations [255].

In addition, CCLE data has also been used to detect different types of phenomena in a variety of tissue types. Copy number data from CCLE confirmed that SIRT6 locus, involved in genome stability, aging as well as tumor suppression, was lost in 35% of cancer cell lines [256]. Recently, a novel TERT promoter mutation also has been identified which eventually led to 2-4 folds increases in activity of telomerase reverse transcriptase. Using mutation data of 150 cancer cell lines form CCLE, TERT mutation was verified in 24 cases (16%) belonging to different cancer types [257]. Wnt/β-catenin signaling has been largely of interest because of its involvement in cancer progression. Recent research showed that complex of oncogenic β-catenin with YAP1 and TBX5 was essential for cancer pathogenesis. β-catenin active cancers were dependent on YAP1 mediated signaling to regulate its oncogenic effects. However, CCLE analysis of YAP1 showed comparable levels of YAP1 in all cancer types including β-catenin-inactive

cancer types suggesting β-catenin might not be the primary regulator of YAP1 in these independent cell lines [258].

All these studies demonstrate the use of CCLE data in understanding of the mechanisms underlying different cancers as well as its importance in discovery and validation of gene(s) in cancer. There is still a need for testing CCLE for meta-analysis of gene-expression modules. In the present study, I have evaluated and confirmed the correlations between three genes for which experimental studies in mouse indicated strong interactions in breast tissue. This correlation study was extended to other cell lines using meta-analysis,

3.1.2. Genes evaluated in the study

In this study I have analyzed expression changes of GRHL3, HNF4A and CDH1. Genes and their functional background has been described below:

3.1.2.1. Grainy head-like 3 (GRHL3)

Grainy head is a highly conserved gene and was first identified in *Drosophila melanogaster*. GRHL1-3 are mammalian ortholog members of the same family. GRHL3 belongs to a family of transcription factors that is an important regulator of development [259]. GRHL3 regulates embryonic development and later on, is confined to cuticle producing tissues including skin and also to central nervous system [260, 261]. Knockdown of GRHL3 gene in mice results in deregulation of several epidermal differentiation programs along with dehydration in mice and ultimately death of pups [262]. GRHL3 along with GRHL1 are also required for transition of epidermal cells from undifferentiated to a differentiate state [262]. Another interesting study has shown that GRHL2 regulates apical junction complex and resembles to HNF4A [263]. However, GRHL3 connection with HNF4A is not well studied.

GRHL3 has also been shown as a tumor suppressor in squamous cell carcinoma via regulation of its downstream target, PTEN [261]. GRHL3 is also associated with anti-

apoptotic actions and consequently plays a role in migration of endothelial cells [264]. Involvement of GRHL3 in migration has been determined in breast cancer cell line MCF7 where it is found to be angiogenic and promoted tumorigenicity [265].

3.1.2.2. Hepatocyte nuclear factor 4A (HNF4A)

HNF4A (Hepatocyte nuclear factor 4A) belongs to the nuclear receptor family and is expressed in epithelial cells of stomach, gut, liver and pancreas. HNF4A controls the fate of hepatic progenitors by regulating expression of various hepatic lineage specific transcription factors [266]. Knockdown of Hnf4a in colon cells of mouse embryo results in deregulated differentiation of epithelial cells [267]. Hnf4a regulates differentiation of hepatic cells and inhibits epithelial to mesenchymal transition (EMT) [268]. It is known to maintain epithelial integrity by controlling three major cell-cell junction programs including adherens junctions, tight junctions and desmosomes. Hnf4a null mice liver also loses the expression of epithelial marker E-cadherin [269]. This is an important lead suggesting that HNF4A and CDH1 may be co-expressed.

TGF-β is a well-known factor treatment to induce mesenchymal phenotype. TGF-β treatment in human fetal hepatocytes resulted in downregulation of HNF4A, hence indicating its specificity to epithelial cells. However, differentiation of cells, which have undergone mesenchymal to epithelial phenotype results in increased expression of HNF4A and well known epithelial marker CDH1 along with other genes [270]. Moreover, it has been shown that balance between proliferation and differentiation along with EMT and mesenchymal to epithelial transition (MET) is maintained by CDH1 and HNF4A via miR122 such that expressions of both genes are upregulated in differentiated and epithelial hepatocytes [271]. Hnf4a is also a discriminative marker between the primary and metastatic tumors in breast and gastric tissue [272].

3.1.2.3. Cadherin 1 (CDH1)

CDH1 as a cell membrane receptor is a member of adherens junctions thus playing an important role in cell-cell adhesion. Cadherins are mainly expressed in continuous cell

layers of endothelial (VE-cadherins) and epithelial tissues (E-cadherins) [273]. Different members of cadherin superfamily have differing roles. Vascular endothelial cadherin (VE-cadherin) in blood vessels is involved in maintenance of the mechanical barrier, while neural cadherin (N-cadherin) is involved in cell motility. Due to their divergent roles in signaling pathways via interaction with growth factor receptors and actin cytoskeleton, cadherins are shown to be involved in the regulation of cell proliferation, motility and tissue homeostasis [274]. Cadherins are disrupted during cancer progression and metastasis, and are highly involved in EMT process [275]. EMT is accompanied by decrease in expression of E-cadherin (which is accomplished by the activity of transcription factors such as SNAI) [276] and increase in N-cadherin, which induces invasiveness and a more aggressive phenotype [277]. In breast cancer progression, downregulation of E-cadherin is thought to result in disruption of contacts between epithelial cells, while an increase in N-cadherin induces the formation of new contracts between cancer cells and stromal cells [278]. This changes in expression of cadherins were also detected in mice models where they initially induced formation of cysts and fibrosis and resulted in malignancy of breast cancer [279].

3.1.3. EMT and MET in cancer

Both processes epithelial to mesenchymal and mesenchymal to epithelial transition, EMT and MET, respectively, have initially been identified as main regulators of early embryogenesis; and later differentiated epithelial cells are shown to have capability to transform in the mesenchymal cell type [280]. In cancer cells, EMT process is generally associated with aggressiveness of tumors and is an important step in spreading locally confined tumors to secondary organs via blood circulation. Reverse process, i.e., MET, takes place at the metastatic site to develop secondary solid tumors [281]. EMT remains the focus of interest for cancer treatment and can be used for the targeting of tumors. EMT begins with loss of tight junctions, epithelial marker CDH1, and integrins along with the gain of mesenchymal markers, N-cadherin and vimentin. These changes are also associated with changes in shape from cuboidal to spindle cells [282]. Acquired features facilitate the cell to invade without its contact to neighboring cells. The reverse

process takes place when these migrating cells reach at their destination [282]. EMT and MET represent dynamic processes [283]. Interestingly, in a recent study it has been shown that circulating tumor cells in patients with metastasis exhibit expression of markers specific for epithelial, mesenchymal and stem cells [284].

3.2. Aims and rationale

Rationale, which has given rise to the following study, is based on identification of enhancers and corresponding transcription factors regulating cdh1 expression in mice. Non-tumorigenic mouse mammary gland epithelial cell line (NMuMG) has been used as a model and reversible EMT could be induced by TGFβ treatment. Two enhancer regions in intron 2 of cdh1 have been identified whose actions mediated by GRHL3 and HNF4A [285].

With the above provided background, aims of this study include the following questions to answer whether

1) GRHL3, HNF4A, and CHD1 are correlated with each other in cancer cell lines expression data (CCLE).
2) There is any consistency across tissues/cancer cell lines based on meta-analysis of GRHL3, CDH1 and HNF4A correlation coefficients.

3.3. Materials and methods

GCRMA normalized CCLE expression data were downloaded from Broad Institute [174]. Correlation analysis was conducted in R using *stat* package [152]. Cell lines belonging to individual tissues were separately analyzed for their correlation coefficients. Tissues having <=4 cell lines were not used in analysis. GRHL3-CDH1 correlation coefficients were used for ranking of each tissue and top 9 tissues were used for meta-correlation analysis as well as scatter plots. Meta-correlation analysis was conducted using *metacor* package in R [145]. Correlation coefficients along with number of cell lines in each tissue were provided for meta-correlation analysis. *Ggplot2* package was used to draw plots [233].

3.4. Results

3.4.1. Correlation between GRHL3, HNF4A and CDH1

CCLE data consisted of cancer cell line expression data on 24 tissues. If number of cell lines were less than 4 for any tissue, it was excluded from correlation analysis; thus, pairwise correlation coefficients among three genes were calculated for each of the 22 tissues and were presented in Table 3.1. Generally, a significant positive correlation was observed between CDH1 and GRHL3/HNF4A in several tissues including the tumors of central nervous system, lung, breast, skin, stomach, large intestine and lung (Table 3.1).

3.4.2. Meta-correlation analysis in cancers originating form different tissues

Meta-correlation analysis for 22 tissues (>=4 cell lines) was conducted to provide an overall effect of the trio in cancer. Meta-correlation analysis results showed a significant positive association between GRHL3 and CDH1 (r.mean =0.31, p=5.7E-08; Figure 3.1). Milder but a significant association was also found between HNF4A and CDH1 (r.mean =0.17, p=4.03E-06; Figure 3.2). GRHL3 and HNF4A meta-correlation also resulted in a weak but significant association (r.mean=0.13, p=3.56E-05; Figure 3.3). Findings revealed that there was a positive correlation of both genes with CDH1 and this correlation was higher for GRHL3. Top 9 tissues showing highest correlation between GRHL3 and CDH1 were plotted and linear increasing trend was observed (Figure 3.4). Findings revealed that expressions of three genes were highly correlated in various cancer types suggesting that findings about the association of CDH1 with GRHL3 and HNF4A in mouse can be extended in humans.

Table 3.1: Correlation between GRHL3, HNF4A and CDH1 in CCLE tissues.

	Correlation coefficient (R)			P-value			
Tissues	GRHL3_CDH1	HNF4A_CDH1	GRHL3_HNF4A	GRHL3_CDH1	HNF4A_CDH1	GRHL3_HNF4A	N
OVARY	0.68	0.16	0.18	3.51E-08	2.66E-01	2.13E-01	52
URINARY_TRACT	0.61	0.16	0.20	7.40E-04	4.13E-01	3.09E-01	27
BILIARY_TRACT	0.61	0.53	0.05	1.11E-01	1.79E-01	9.08E-01	8
LIVER	0.57	0.58	0.30	1.39E-03	1.30E-03	1.27E-01	28
CENTRAL_NERVOUS_SYSTEM	0.47	0.28	0.08	4.68E-05	1.87E-02	5.12E-01	69
BREAST	0.44	0.01	0.01	4.19E-04	9.55E-01	9.43E-01	59
LUNG	0.44	0.21	0.16	2.52E-10	4.61E-03	2.98E-02	187
LARGE_INTESTINE	0.43	0.29	-0.10	5.97E-04	2.50E-02	4.40E-01	61
UPPER_AERODIGESTIVE_TRACT	0.42	0.24	0.16	1.58E-02	1.79E-01	3.73E-01	32
STOMACH	0.42	0.36	0.09	8.34E-03	2.58E-02	5.95E-01	38
SKIN	0.37	0.13	-0.04	3.41E-03	3.23E-01	7.61E-01	62
SOFT_TISSUE	0.32	0.37	0.01	1.61E-01	9.53E-02	9.74E-01	21
PANCREAS	0.30	-0.07	0.15	4.85E-02	6.62E-01	3.42E-01	44
HAEMATOPOIETIC_AND_LYMPHOID_TISSUE	0.11	0.04	0.15	1.49E-01	6.22E-01	4.44E-02	180
PLEURA	0.09	0.01	0.70	7.88E-01	9.85E-01	1.67E-02	11

	Correlation coefficient (R)			P-value			
Tissues	GRHL3_CDH1	HNF4A_CDH1	GRHL3_HNF4A	GRHL3_CDH1	HNF4A_CDH1	GRHL3_HNF4A	N
ENDOMETRIUM	0.05	0.12	0.04	8.22E-01	5.51E-01	8.38E-01	27
AUTONOMIC_GANGLIA	0.03	0.38	0.14	9.02E-01	1.32E-01	5.86E-01	17
THYROID	-0.01	0.19	-0.09	9.72E-01	5.57E-01	7.82E-01	12
KIDNEY	-0.03	-0.20	0.35	8.62E-01	2.43E-01	3.37E-02	36
OESOPHAGUS	-0.10	0.25	0.47	6.36E-01	2.25E-01	1.54E-02	26
BONE	-0.17	0.24	-0.02	3.66E-01	2.13E-01	9.35E-01	29
PROSTATE	-0.52	-0.36	0.27	1.85E-01	3.79E-01	5.11E-01	8

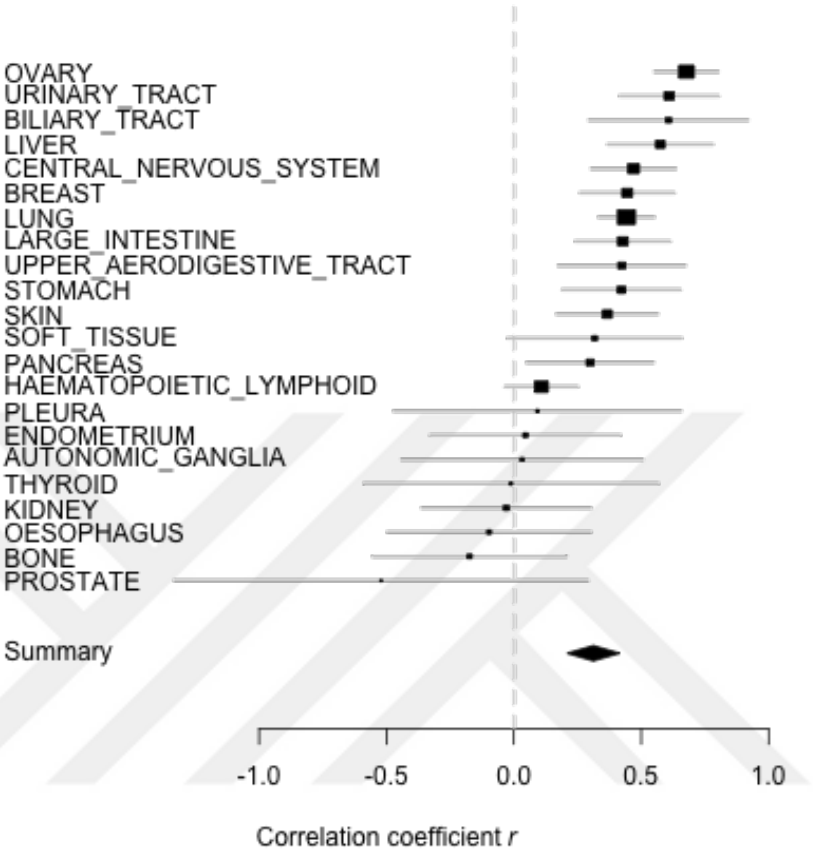


Figure 3.1: Meta-correlation between GRHL3 and CDH1 in CCLE data (R.mean=0.31, p=5.71E=08).

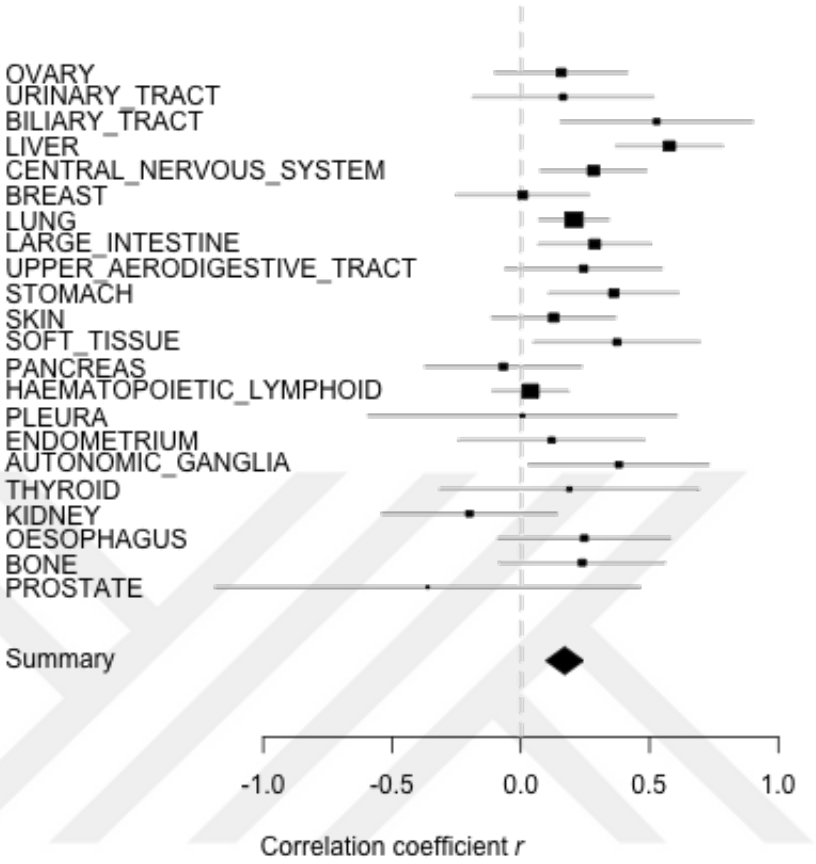


Figure 3.2: Meta-correlation between HNF4A and CDH1 in CCLE data (R.mean=0.17, p=4.03E=06).

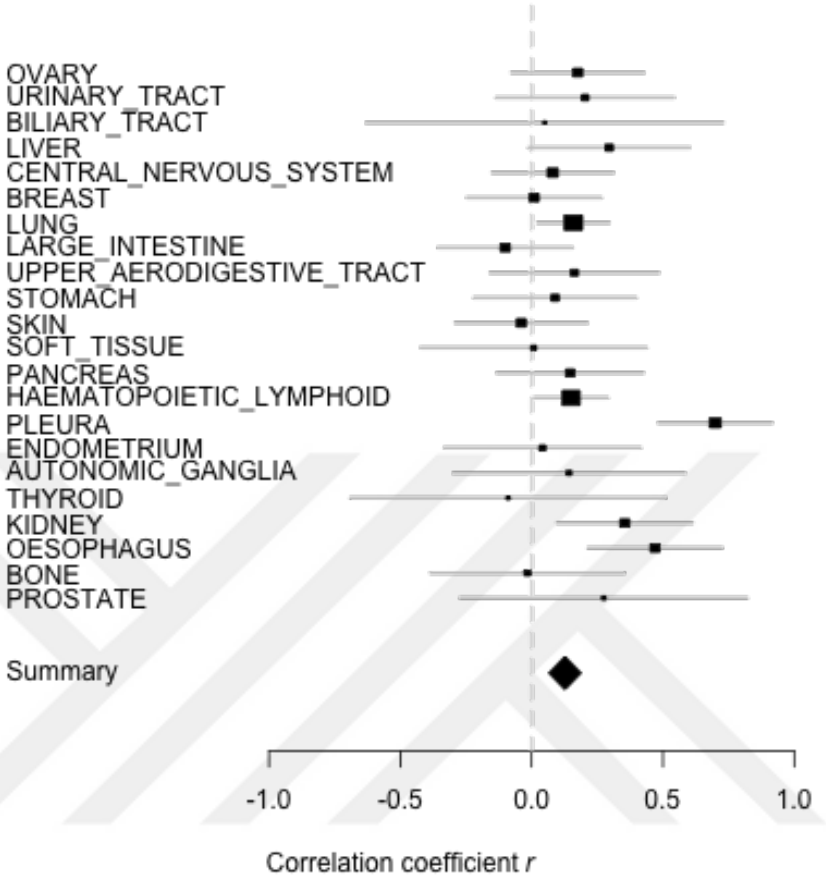


Figure 3.3: Meta-correlation between HNF4A and GRHL3 in CCLE data (R.mean=0.13, p=3.56E=05).

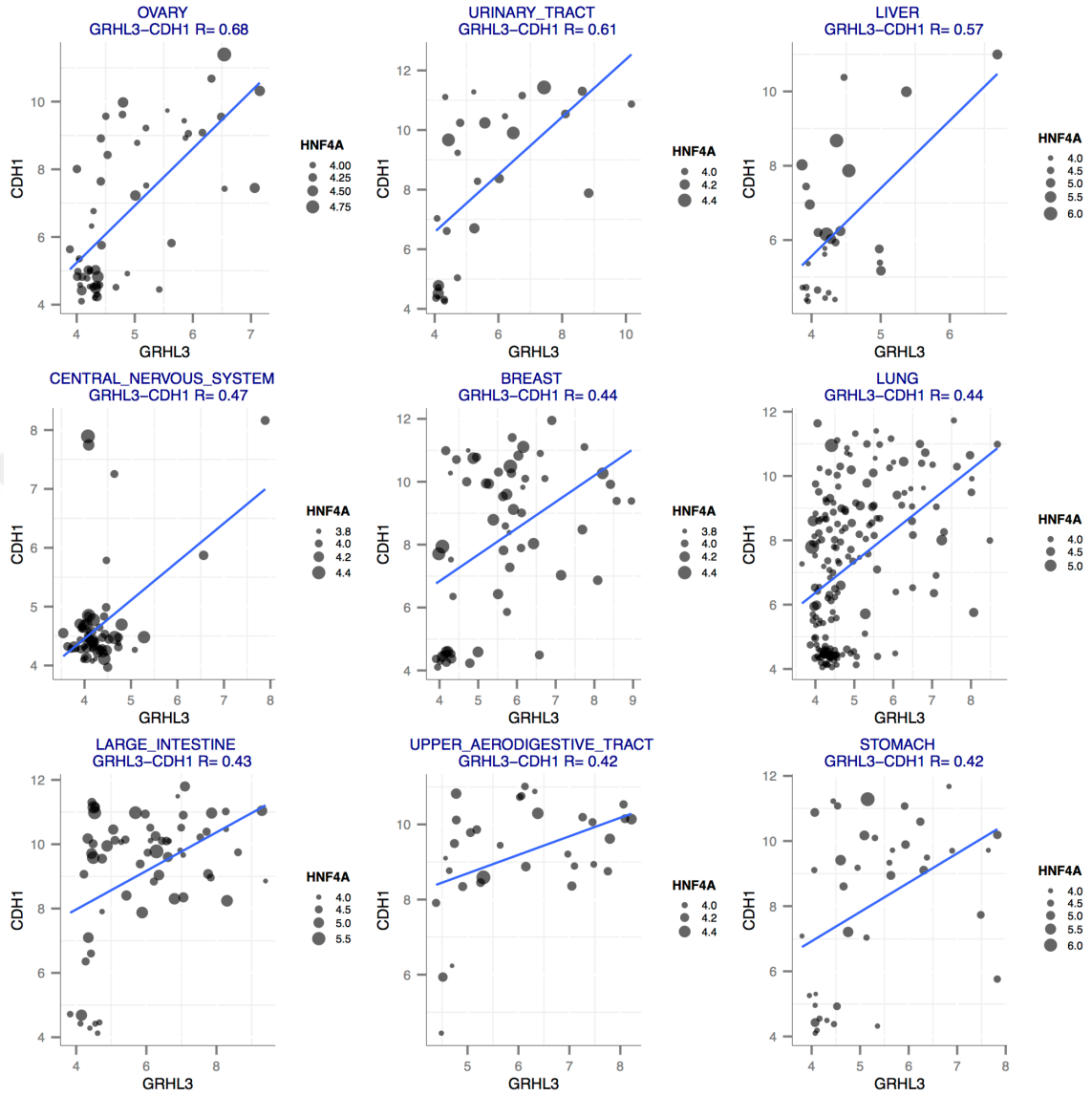


Figure 3.4: GRHL3, CDH1 and HNF4A scatter-plot in CCLE cell lines

9 cell lines showing highest GRHL3-CDH1 correlation coefficient are plotted.

3.5. Discussion

CDH1 is an important epithelial marker and its upregulation during transition to epithelial phenotype is a well-known fact. A previous study by Dr. Hani Alotaibi showed regulation of CDH1 by GRHL3 and HNF4A during MET process [285]. However, I have used a wide range of cancer cell lines to show that this correlation between these three genes might not be limited to only MET process but was observed in a wide range of human cancer cell line repertoire with different lineages. Besides, this correlation was solely specific to breast but multiple tissues of epithelial origin exhibited the phenomenon. GRHL3 is generally confined to epithelial tissues and is mostly found in skin [286] and mammary tissue [285]. Our results also verified positive significant correlation of GRHL3 and CDH1 in skin and breast cell lines.

Generally correlation coefficients were the highest between GRHL3 and CDH1 indicating co-expression of both genes in epithelial tissues. HNF4A and CDH1 correlation coefficient being intermediate was followed by least values between GRHL3 and HNF4A. Highest correlation between GRHL3 and CDH1 were also justified by the fact that CDH1 expression has been majorly regulated by GRHL3 however; HNF4A might be involved in the process in a tissue dependent manner [285].

Our finding also exhibits the potential of CCLE data to translate an experimental finding in mouse into a vast range of human cancer cell lines. Correlation analysis can provide a measure of co-expression between different genes and can help prioritize tissues on the basis of co-expression profile. Meta-correlation analysis, on the other hand, provided collective co-expression measure in different epithelial tissue.

3.6. Future perspective

The study highlighted the fact that several of regulatory processes during MET are not yet fully explored. To cover the gap, ongoing work on the project includes induction of EMT by TGFβ followed by time-series MET experiment (personal communication with Dr. Alotaibi). Gene expression changes are to be evaluated by RNAseq data analysis. Differential expression analysis along with co-expression studies will be used to identify network of transcription factor regulating different stages of MET.

One of the following studies has shown that GRHL3 is involved in repression of CDH1 expression in some of the epithelial tissues, which makes cancer cells more aggressive [287]. However, in line with our findings, GRHL3 expression regulation in mammary epithelial cells has recently been verified [288]. Future studies are needed to study the regulation between GRHL3 and CDH1 in a cell/tissue specific manner to verify the positive/negative, or linear/non-linear associations.

Besides, GRHL3 also regulates the expression of HNF4A [285]. However, from CCLE correlation analysis, a strong positive correlation between these genes was not observed. Future studies are needed to identify other factors, which might be co-regulating the expression of HNF4A.

Moreover, future studies can lead to development of a Shiny tool for meta-analysis of CCLE datasets.

Chapter 4

Identification of CHRNA5 as a breast cancer biomarker using meta-analysis and development of E2S (Estrogen (E2) to Survival) database

4.1. Introduction

Breast cancer is one of the most prevalent cancers among women and is also a major reason of death [289, 290]. Prognosis of breast cancer differs with age of a patient and type of breast cancer. Cancer, when localized to breast, is less aggressive. However, regional breast cancer spreads out to lymph nodes and afterwards to any distant secondary site in body, and thus becomes a challenge for treatment therapies [291]. It is thus important to identify novel biomarkers of breast cancer that help diagnosis, prognosis and/or treatment of breast cancer.

4.1.1. Hierarchy and development of breast tissue

Breast tissue has a well-organized cellular hierarchy. Development of mammary gland starts at 4^{th} week of embryogenesis in human [292]. Mammary gland consists of branching epithelial tree of ductal structure, which is embedded in extracellular stroma and fat-pad. Epithelium of mammary gland is a bi-layered structure consisting of luminal and myoepithelial cells [293]. Luminal cells give rise to lumen of ducts and alveoli; which are covered by myoepithelial cells, which separate mammary gland form stroma and basement membrane [294, 295]. Myoepithelial cells are spindle shaped cells with contractility to squeeze milk in the ducts during pregnancy [296, 297]. Mammary epithelium is a highly dynamic structure and morphologically changes with onset of pregnancy [298]. Luminal and myoepithelial cells can be differentiated from each other based on lineage specific markers, e.g., cytokeratin (CKs) 8/18 [299]

and 19 [300], MUC1 [301], GATA3, and CD24 for luminal cells [301] and CKs 5 [300], 14 [302], and 17 [299], smooth muscle actin (SMA/ACTA2) [303] and vimentin [304] for myoepithelial cells [295, 297]. All cell lineages are highly organized but with the onset of tumorigenesis, cell organization is lost (Figure 4.1) [305].

4.1.2. Differentiation in cancer

Although differentiation is a common process in embryonic and organ development, de-differentiation and trans-differentiation can be as important in cancer. Traditional cancer treatment strategies focus on killing of the cells by chemo- or radio- therapy [306]. However, another approach of treatment is based on reverting cancer cells toward a more differentiated normal cell [307]. All-trans-retinoic acid is one such therapy for acute promyelocytic leukemia, which mediate its action by inducing terminal differentiation [307]. Another study has focused on comparative expression analysis of stem cells differentiating to adipogenesis as well as to osteoblastogenesis. They found that gene repression has taken place before cell commitment and a signature of 65 genes has been presented controlling osteoblast/adipocyte differentiation [308]. However, meta-analysis or comparative transcriptomics of differentiation process has received little attention.

4.1.3. Involvement of different receptors in breast cancer

Breast cancer has multiple subtypes, which differ in their prognostic importance. These subtypes can be differentiated on the basis of important receptors found in breast tissue (ER, PR, HER2) [309]. Several of steroid nuclear receptors have been found to be extremely important in this classification [310] (Figure 4.2). In this thesis, I have studied three nuclear receptors (Estrogen Receptor, ER (Chapter 4); Mineralocorticoid Receptor, MR and Glucocorticoid Receptor, GR (Chapter 5)) and one membrane receptor subunit, i.e., Cholinergic Receptor alpha 5, nicotinic (CHRNA5; Chapter 4). Nuclear receptors and their role in cancer have been dicussed in Section 5.1. However, this Chapter 4 has focused on CHRNA5 and its role in differentiation as well as regulation by ER.

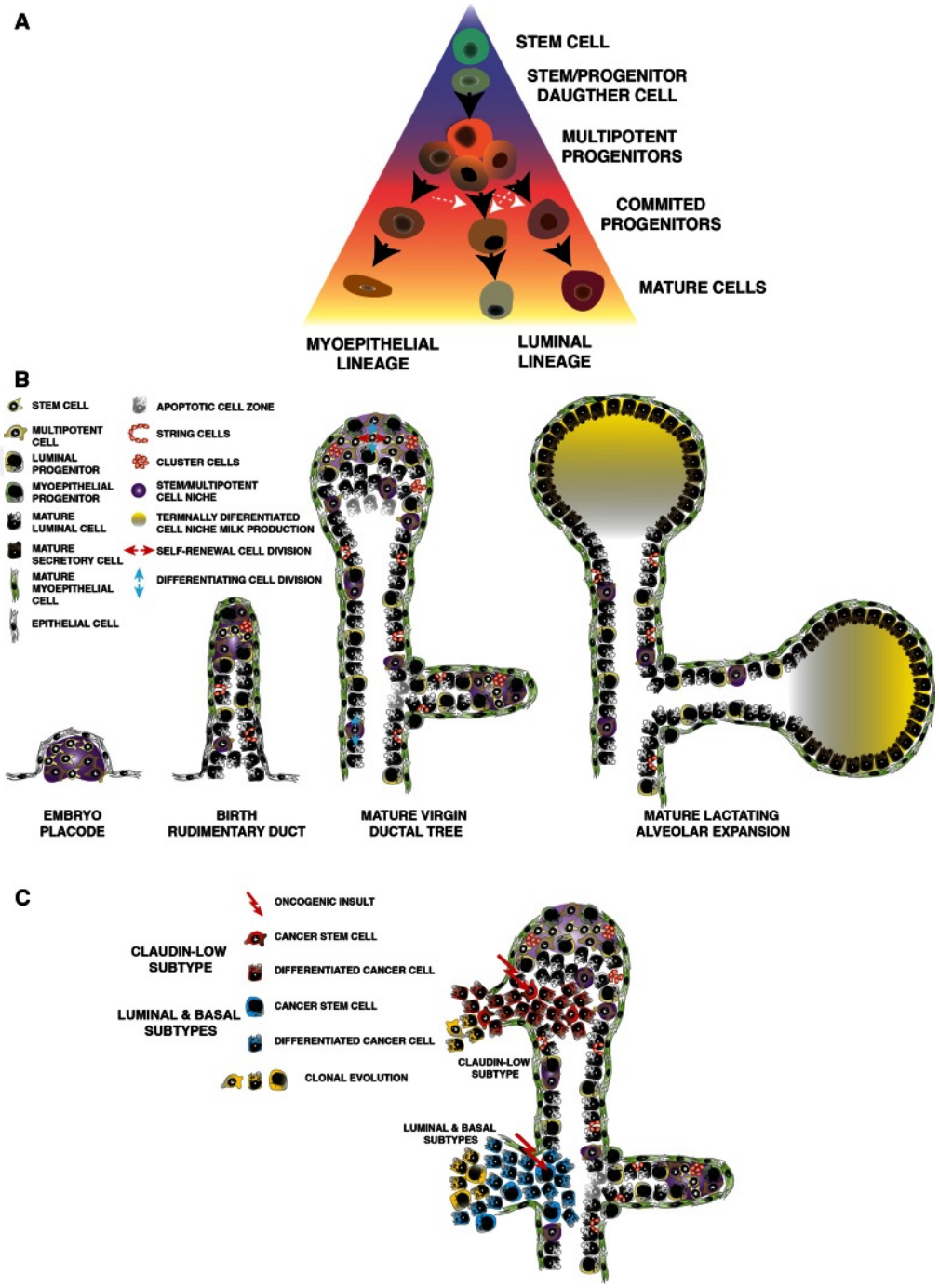


Figure 4.1: Cellular hierarchy in breast development and tumorigenesis [305].

A) Hierarchy of cells lineage. B) Normal organization of different cell types. C) Disrupted cell organization in carcinogenesis. This figure is reprinted from *Oakes et al., 2014* [305] and the copyright permission can be found in Appendix V.

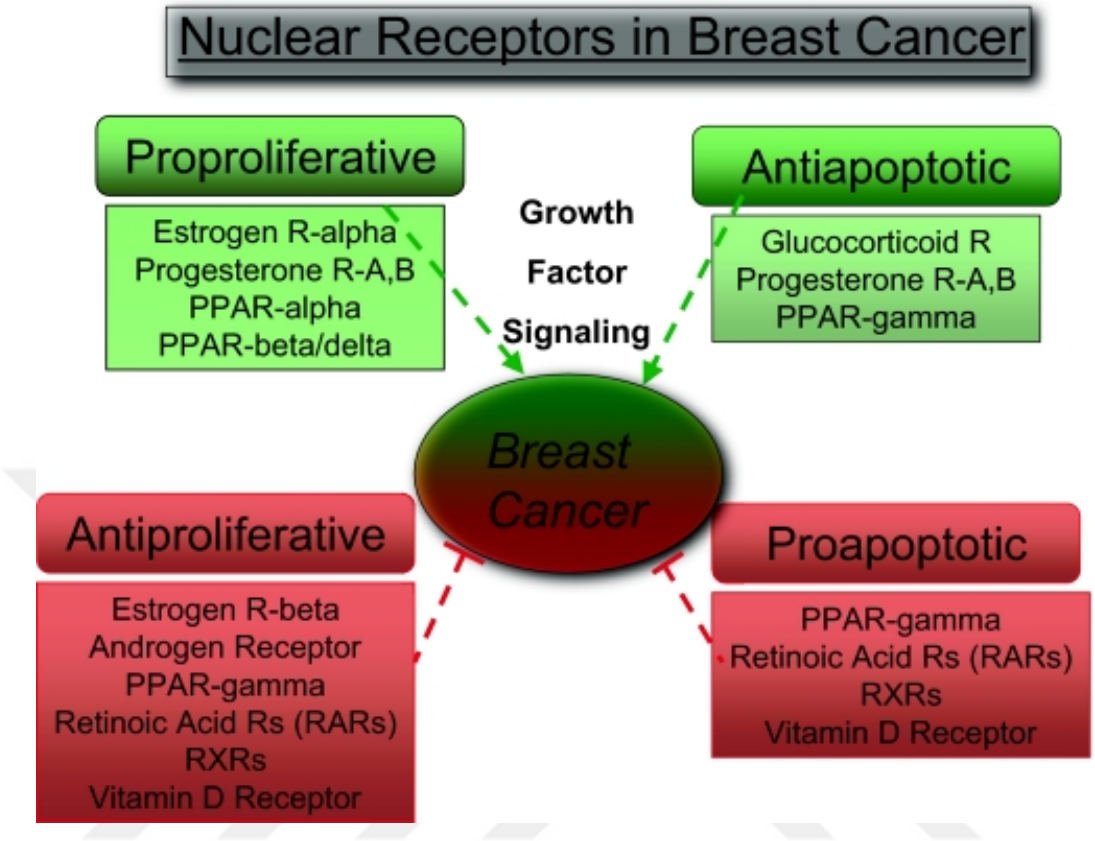


Figure 4.2: Regulatory mechanisms of different nuclear receptors in breast cancer [311].

This figure is reprinted from *Conzen et al., 2008* [311] and copyright permission can be found in Appendix V

4.1.4. Estrogen signaling in breast cancer

Estrogen (E2) is a steroid hormone, which shows its action through estrogen receptors (ERs). Breast classification based on ER status gives rise to ER+ (estrogen receptor positive) breast cancer subtypes encompassing luminal A and luminal B while ER- (estrogen receptor negative) subtypes can be classified as basal and HER2+/ER- subtype; all of these subtypes are associated with different prognoses [312]. Therefore E2 and E2-mediated signaling remains a big player in breast cancer research. Upon activation by ligand binding, ERα interacts with >10,000 genome sites, facilitates co-regulators' binding and recruits RNA polymerase II for active transcriptional regulation of target genes [313].

Genomic action of E2, which involves binding of E2-ER, transport to nucleus, interaction with estrogen response element (EREs) on gene promoters and ultimately transcriptional regulation of target genes [314], has remained a main focus for a long time. Another nuclear but non-classical type of E2 signaling includes ER binding to nuclear proteins and mediates downstream signaling, e.g., interaction of ER with receptors NFYB and SP1 also leads to transactivation of several genes [315]. Both of these nuclear events take hours or sometime days [316]. In recent findings, it has been observed that upon E2 treatment, instant ESR1 mediated actions are also observed on plasma membrane. These findings provide new nomenclature to E2 signaling (E2-ER mediated transcriptional regulation) as "nuclear-initiated steroid signaling" (NISS), and the immediate membrane associated activity is known as "membrane-initiated steroid signaling" (MISS) [316]. Primary targets of E2 are regulated upon E2-mediated transcriptional regulation however secondary E2 targets are induced later and need primary targets for their transcriptional regulation [317, 318].

In the healthy mammary gland, E2 facilitates growth of ER+ cells and in addition blocks the cell death. E2 signaling in the cell produces growth factors and by paracrine signaling, these growth factors are distributed to ER- cells in the surroundings [319]. Cancer cells also take advantage of paracrine signaling to regulate proliferation. ERα has also been associated with metastasis [319]. Research has shown that there is a correlation between ERα+ tumor and metastasis to lung and bone tissue [320]. In case of metastasis of ERα positive breast cancer, 80% of lymph nodes and 65-70% of distant metastatic cells are also found to be ERα positive [321].

4.1.5. ER-antagonists in breast cancer treatment

With the increasing significance of E2/ER signaling, breast cancer drugs targeting E2/ER signaling started invading the treatment scenario [322]. Tamoxifen is one of the well-studied and widely used drugs in ER+ breast tumors. Tamoxifen shows its action by competing for E2 binding to ER and suppressing E2-mediated autocrine or paracrine growth regulatory actions [323]. Ultimately G1-arrest and reduced proliferation is observed in treated patients [324, 325]. Affinity of tamoxifen for estrogen receptor is

very high and is comparable to affinity of its primary ligand 17β-estradiol [324]. It has been shown that tamoxifen exhibits its action against invasive as well as non-invasive breast cancer [326]. Tamoxifen associated benefits are maximum for node negative patients having positive estrogen receptor status [327]. These findings show that identification of breast cancer subtype along with specific genes contributing to this subtype can have implications for complementing existing therapies.

4.1.6. Estrogens, ER, and meta-analysis studies in the literature

Meta-analysis methods as described in Chapter 1 are divergent. The commonality among them is use of stringent parametric or non-parametric statistics to obtain a highly consistent meta-statistic and associated p-values. Accordingly, very few studies have attempted the meta-analysis of E2 signaling. One of the E2 treatment meta-analysis focused on treatment of MCF7 cells, probably because a large number of studies are available. After categorization of early and long E2 treatment profiles, separate meta-analyses were performed using weighted t.tests; the authors have also developed a web-tool Gene Expression Meta-Signature (GEMS) to facilitate evaluation of any gene [328]. A following study used ranking based meta-analysis approach and declared that early E2 exposure network identified in GEMS interestingly overlapped with their late E2 treatment network [329]. Therefore, there is still need of well-controlled time-dependent E2 meta-analysis in E2-treated MCF7 cells. In another online tool NetworkAnalyst, which can facilitate individual dataset analysis as well as meta-analysis, an E2 dataset has been provided as a default example and thus can facilitate users to evaluate the signature associated with E2 signaling, but in the individual dataset analysis module [330]. A fourth example is a zebrafish human comparative transcriptomics study that analyzes the conserved pathways driven by E2 [193].

Similarly, few meta-analysis or expression data collection studies have also been conducted on ER-/ER+ breast cancer patients. Focusing on eight different datasets, an ER-/ER+ subtype meta-analysis was performed using random effects model (REM) and showed that ER- patients were enriched with proliferation-inducing and MYC/E2F target genes [331]. Another interesting study has focused on E2 treatment in MCF7 and

T47D along with ER-/ER+ breast cancer expression profiling in six datasets. In this study, instead of using any meta-analysis strategy, individual datasets were analyzed [332]. Association between physical activity and breast cancer survival were also analyzed using meta-analysis of six-studies and it was found that physical activity was associated with improved survival in ER+ patient but no such effect was observed in ER- patients [333]. Yet another study used a large cohort of breast cancer patient data with eighteen datasets and has performed co-expression analysis along with meta-survival analysis (using FEM and REM). Interestingly, they identified modules based on expression and also associated these modules with good or bad prognosis. Results concluded that ER- and ERBB+ subtypes resulted in bad outcome and this differential prognosis was associated with increased expression of proliferation genes [334]. However, there is still a need to improve meta-analysis strategies by making meta-analysis based comparisons between independent cohorts, i.e., a comparison of meta-analysis of a molecular type (ER-/+) with that of the survival outcome at the genome-wide context has not yet been performed. In addition, mapping multiple sources of information, e.g., secondary E2 target scores, thus providing more than one-layer of transcriptomics information in meta-analysis studies will improve the predictive potential of meta-analyses. Finally, there is no comprehensive online tool for meta-analysis of E2-ER signaling and its association with prognosis.

4.1.7. Cholinergic receptors

Cell membrane acts as a barrier between cell content and extra-cellular environment but has also been adapted to regulate ion movement in and out of the cell via ion-channels [335]. Opening/closing of these channels is facilitated by electrical, chemical or mechanical signals. The family of ligand-gated ion-channels monitors and guides the transport of chemical messengers into the cell [335]. Nicotinic acetylcholine receptors (nACHRs) are prototype receptors of this family and respond to binding of acetylcholine or the drug, nicotine. nACHRs are categorized in two groups: 1) the muscle type: generally localized in skeletal muscles; 2) the neuronal type: present in different parts of nervous system [335]. Neuronal nAChRs regulate cerebral blood flow and metabolism

along with cognition, learning and memory, arousal and pathophysiological conditions [336]. Neuronal nACHRs have also been found and studied in several non-neuronal/non-muscle tissues including keratinocytes, blood cells, epithelia and macrophages [335, 337].

In vertebrates, nACHR subunits, coding seventeen cDNAs, have been cloned (α1-α10, β1-4, γ, δ, ε). α-subunits were distinguished from 'complementary' non-α-subunits by the presence of specific sequence of amino acid required for ligand binding at the N-terminal [338]. Multiple subunits of nACHRs combine to form homo- or hetero-pentamers receptor complex having the same underlying structure (large extracellular and intracellular along with four trans-membrane domains) [339]. Homo-pentamers are generally composed of one type of α subunit, e.g., α7, while heteropentameric receptors are constituted by the combination of α and β subunits, e.g., $\alpha_2\beta_3$ [338].

4.1.7.1. Cholinergic signaling in cancer

nACHRs bind to nicotine with high affinities and thus are involved in nicotine addiction in tobacco users [340]. nACHR/nicotine and lung cancer/nicotine association generate a strong motivation to study nACHR regulation in lung cancer. Multiple studies have shown role of nACHRs in a multiple cancer related processes including cell growth [341], resistance to apoptosis [342], proliferation, tumor invasion and angiogenesis [343, 344]. Involvement of nACHRs in lung cancer has been highlighted by significant upregulation of β4, downregulation of α4 in lung cancer and upregulated expression of α6β3 in non-smokers as compared to smokers [345]. Proliferative effects of nicotine are arbitrated by α7 subunit of nACHR and knockdown of α7 abrogates nicotine mediated proliferative effects [343]. α7 activates ERK and PI3K/mTOR signaling pathways ultimately leading to upregulation of EMT marker fibronectin and its receptor α5β1 integrin [346]. α7 also facilitates calcium signaling and ultimately activation of MEKK-1, ERK1/2 and $p90^{RSK}$ and NF-KB, inducing cell cycle progression through entry in to S-phase [341]. Proliferative effects of nACHRs are mediated by activation of NFKB, Src, the 5-lipooxygenase pathways, AKT, HIF-1α and COX-2. nACHRs mediate anti-apoptotic effects on cancel cells via overexpression of XIAP and survivin via regulation

of α3-nAChR [343]. Nicotine can induce angiogenesis and atherogenesis by binding to nAChRs on endothelial and smooth muscle cells α7 nAChR has been implicated in mediating the proliferative effects of nicotine [347].

4.1.7.2. CHRNA5 in cancer

CHRNA5 is located on chromosome 15q and is organized in a gene cluster where two other cholinergic receptor genes (*CHRNA3* and *CHRNB4*) are also located [348]. Multiple studies have shown involvement of gene cluster on nicotine dependent actions and lung cancer [349]. Nicotine treatment in lung cancer cells results in increased CHRNA5 expression [350]. Another study has shown that expression of multiple CHRNA5 isoforms increases during lung cancer [351]. There is no comprehensive study in the literature focusing on its effects on differentiation and regulation by E2- or ER-mediated signaling in breast cancer cells.

4.1.8. Genes important in mammary differentiation and cancer cell morphology

Cadherin 1 (CDH1) is an epithelial marker and has been discussed in section 3.1.2.3.

Vimentin (VIM) is a member of intermediate filament family and provides stability for cytoskeletal proteins, integrity of cytoplasm and cell shape. It is expressed in different cell types but has gained important attention as a marker of EMT due to its overexpression in different cancer types [352]. Many *in-vitro* experiments showed that vimentin might have a tumor promoter role in cancer [353] and in several aggressive breast cancer cell lines vimentin was shown to be overexpressed [354]. Overexpressing of vimentin in MCF7 has led to increases in migration while its downregulation in MDA-MB-231 resulted in less motility and increased migrative capability of the cell line [355]. Histological studies of breast cancer samples show that the more aggressive tumors (high-grade, low ER) get, they express higher levels of vimentin [356].

Actin alpha2, smooth muscle, aorta (ACTA2) is a member of actin proteins. Being an alpha subunit, ACTA2 protein is mainly involved in contractility of muscle cells,

healing of wounds and lesions in non-muscle cells [357]. Knockout experiments in mice have shown impairment of blood pressure and vascular contractility [358] while mutations of this gene have been linked to pathologies like coronary artery disease and premature ischemic stroke [359]. Its expression is reported to be regulated in de-differentiation of parotid gland [360] and also in differentiation of myoepithelial cells [361]. ACTA2 has also been shown to regulate contraction of myoepithelial cells [362].

Mucin1, cell surface associated (MUC1) is expressed in glandular or luminal epithelial cells of lung, stomach, mammary gland, prostate, uterus, and pancreas, where it is involved in formation of mucus barriers by providing lubrification [363]. It regulates cell adhesion as well as several cell signaling pathways [363]. In addition, this protein plays an important role in transcriptional regulation of several important cancer markers like TP53, STAT3 and β-catenin [364, 365]. MUC1 overexpression in pancreatic cancer is accompanied by its binding to β-catenin, translocation to nucleus, upregulation of the several EMT markers (VIM, SNAI1 and TWIST) and downregulation of E-cadherin expression [366]. Interaction between MUC1 and β-catenin is also observed in breast cancer carcinomas [367]. MUC1 in breast carcinomas is characterized by overexpression and change in its glycosylation pattern [368]. *In-vivo* studies also associated its expression with the transformation of mammary glands [367] while pancreatic and lung cancer studies implicate its role with differentiation [369, 370].

Snail family transcriptional repressor 2 (SNAI2) is a zip finger transcriptional factor/repressor having important roles in development, especially expressed in the migratory cells undergoing EMT [371]. Due to its role in motility, SNAI2 is involved in melanocyte migration, spermatogenesis and hematopoiesis [372]. SNAI2 expression is associated with different pathologies including several cancer types [372]. It is implicated with drug resistance due to its role in protecting the cancer cells from apoptosis [373]. It is described as an anti-apoptotic factor and EMT regulator in cancer [374]. Its involvement in EMT is usually associated with the E-cadherin repression during embryogenesis as well as tumor progression [375]. SNAI2 is also involved in maintenance of stem-ness in specific cancer cells which are capable of regenerating new

tumors [376]. In addition, SNAI2 shows implications in keeping the non-epithelial phenotype during tumorigenesis [377].

4.2. Aims and rationale

1) Previous work from our lab has shown functional pathways associated with CHRNA5 downregulation in breast cancer (Ermira Jahja, PhD thesis, 2017). Ermira Jahja has shown that downregulation of CHRNA5 resulted in changed cell morphology with filopodia-like projections in MCF7 cells. Based on these findings, the work of my thesis is focused on the following points:

- Evaluation of EMT markers by qPCR in CHRNA5 depleted cells in addition to comparative transcriptomics study with an EMT gene list.
- Comprehensive meta-analysis of differentiation datasets in the epithelial or non-epithelial tissues. I aimed to obtain a differentiation meta-signature against which any expression logFC (associated p-value) dataset can be compared to address the degree of association with differentiation status.
- Use of CHRNA5 RNAi data to test whether CHRNA5 downregulation is associated with any kind of differentiation. Furthermore, I evaluated luminal and myoepithelial markers to validate lineage pattern of CHRNA5 RNAi induced differentiation.

2) Previous individual microarray expression dataset analyses (Rumeysa Biyik, MS thesis, 2009) suggested that CHRNA5 is an E2 secondary target and exhibits expression differences between ER+ and ER- breast cancer subtypes. However, REM/FEM methodology of meta-analysis approach was not applied and the scale of the study needed expansion. In my thesis, I aimed to verify E2-mediated CHRNA5 regulation by collecting larger microarray cohorts and using a statistical meta-analysis strategy. Besides, using two independent cohorts of patient data I aimed to analyze prognostic importance of CHRNA5 using survival meta-analysis and CHRNA5 expression in ER-/ER+ subtypes using differential expression meta-analysis. Identification of co-expression network to evaluate CHRNA5 mediated signaling and its prognostic importance in breast cancer also became a focus of the study.

3) In addition to CHRNA5 mediated signaling, whole transcriptome-wide meta-analysis was performed for collected cohorts (*E2 cohort*, *ER patient cohort*, *survival cohort*). A highly comprehensive web-tool collating *in-vitro* and *in-vivo* expression

datasets for performing a FEM or REM methodology is lacking in breast cancer. I aimed to extend and generalize this methodology using a Shiny app so that any user queried gene can be searched for its association with E2 mediated signaling as primary or secondary target of E2 as well as its association with survival/prognosis.

4.3. Materials and methods

4.3.1. CHRNA5 depletion expression data analysis (GSE89333)

CHRNA5 depletion and microarray experiments were performed in MCF7 cells (Ermira Jahja, PhD thesis, 2017). Cel files were evaluated for samples quality using *affyQCreport* package in R and were normalized by RMA. Jetset (version No 3.1.2) was used to remove gene-probeset redundancy by considering only best probesets for each gene. Differential expression analysis was performed using *limma* [56].

4.3.2. Cellular differentiation expression data collection and processing

Cellular differentiation microarray studies of normal tissues/cells based on Affymetrix HG-U133 Plus 2 platform (GPL570) were collected from GEO database [171]. The retrieved 11 datasets were described in Table 4.2. Raw data (cel files) of each dataset were downloaded using *GEOquery* package [227] and normalized by RMA using *affy* package [51]. *limma* analysis was used to test for differential expression [56]. Paired differential expression analysis was conducted only for the datasets obtained from the human donors with known matched origin in both the control and differentiated samples otherwise unpaired differential expression analysis was performed. For time-series datasets, all time points were compared with the earliest differentiation time-point in the dataset. Expression signatures of these differentiation datasets were compared with the CHRNA5 depletion profile. Association between logFC of significantly regulated probesets in each differentiation profile and those from CHRNA5 depletion profile ($p<0.05$) was evaluated using Odds ratio (OR) and Fisher's exact test from *exact2x2* package [231]. Meta-analysis was conducted by considering a single time-point from each of the epithelial differentiation datasets. Selection of a single time-point was based on the most significant Fisher's exact p-value and significant CHRNA5 downregulation in dataset. Paired and unpaired datasets were specified for meta-analysis before *MetaDE* package was used for meta-differential expression analysis [230]. Parameters for *MetaDE* package included 1000 permutation and a seed value of 123.

4.3.3. CEMA (Core Expression Module Analysis) data analysis

Raw Expression data (GSE47796; [378]) comprising 17 tissues was downloaded and normalized using RMA. *Limma* differential expression was conducted to compare each of 16 tissues with the undifferentiated pluripotent stem cell group. CHRNA5 associated differentiation signature was extracted by filtering genes significantly modulated by CHRNA5 depletion ($p<0.05$) and meta-analysis of epithelial differentiation (FDR<0.05). This filtered profile was split in two separate signatures: 1) upregulated profile, i.e., logFC of CHRNA5 depletion data>0 and mu.hat of meta-analysis>0 and 2) downregulated profile, i.e., logFC of CHRNA5 depletion data<0 and mu.hat of meta-analysis<0. *PVclust* package was used for clustering of upregulated and downregulated logFC of these 16 tissues with respect to pluripotent stem cells [379]. Hierarchical clustering was performed using uncentered distance, pairwise complete correlation, and 1000 permutations.

4.3.4. Mapping of TP53 and EMT markers onto CHRNA5 depletion profile

List of epithelial and mesenchymal marker genes were retrieved from a dataset by Chang et al. [380]. R package *biomart* was used to map genes to Affymetrix probesets [74]. Respective log fold change and p-values of epithelial/mesenchymal genes were extracted from *limma* analysis and the distribution plot of significant genes ($p<0.05$) was generated using *ggplot2* package [233].

List of TP53 targets, predicted by RNAseq and also validated by ChIPseq, were retrieved from Janky et al. [91]. Study has also provided logFC values of these targets obtained from RNAseq data analysis. Again, *biomart* package of R [74] was used to extract Affymetrix probesets of the given gene list. RNAseq log fold changes were compared with those of CHRNA5 depletion expression profile and the correlation analysis between these two datasets was performed in R.

4.3.5. CCLE (Cancer Cell Line Encyclopedia) data analysis for CHRNA5 expression

GCRMA normalized expression data of cancer cell lines from all tissues were downloaded from Broad Institute website [174]. CCLE expression profiling has been performed on a customized version of Affy HG-U133 Plus 2.0 microarray platform. For comparison with CHRNA5 depletion profile (Affy HG-U133 Plus 2.0), probeset mapping of this customized version were also downloaded from Broad Institute website. Only jetset best probesets [381] were selected for mapping with the customized CCLE platform. Expression data of 59 breast cancer cell lines were extracted and expression correlation analysis was performed between CHRNA5 and each gene to retrieve correlation coefficients and p-values. *Exact2x2* package was used to perform Fisher's exact test and to obtain an Odds Ratio [231].

4.3.6. Chromosome enrichment

R package *category* was used for chromosomal enrichment analysis [108]. CHRNA5 correlated genes ($R>0.3$ and $p<0.05$) across the CCLE breast cancer cell lines, which were also downregulated in the CHRNA5 depletion profile ($logFC<-0.5$ and $p<0.05$), were extracted for the enrichment analysis. All Ensemble IDs corresponding to Affymetrix HG-U133 plus 2.0 was used as the background list.

4.3.7. Primer design and primer efficiency

Primer-BLAST [382], the primer designing tool provided by NCBI (National center for biotechnology information), and *in-silico* PCR, which is developed by UCSC [383] were used to design primers and for evaluation of the specificity and reliability of designed primers. Forward and reverse sequences of primers were described in Table 4.1.

Table 4.1: Primer sequences used for RT-qPCR in CHRNA5 depleted MCF7, BT20 and MDA-MB-231 samples.

Name	Primer sequence (5'-3')	Size of amplicon (bp)	Tm (°C)	Amplification efficiency
SDHA	F: TGGGAACAAGAGGGCATCTG R: CCACCACTGCATCAAATTCATG	86	78.24	2.01
VIM	F: CGTCACCTTCGTGAATACCA R: CCAGAGGGAGTGAATCCAGA	169	82.94	1.95
CDH1	F: AGCGTGTGTGACTGTGAAGG R: CTCTTCTCCGCCTCCTTCTT	240	85.28	1.96
ACTA2**	F: GGGAATGGGACAAAAAGACA R: CTTCAGGGGCAACACGAA	164	82.77	1.89
MUC1	F: TACCGATCGTAGCCCCTATG R: ACCTGAGTGGAGTGGAATGG	157	87.79	1.89
SNAI2***	F: CGAACTGGACACACATACAGTG R: CTGAGGATCTCTGGTTGTGGT	87	78.54	1.73

** (Zhang et al. 2013)

*** (Zhao et al. 2016)

Before amplification of the specified gene, primer efficiencies were evaluated by performing RT-qPCR of cDNA which was serially diluted as 1/5, 1/25, 1/125, 1/625; 1/10, 1/100, 1/1000, 1/10000 or 1/2, 1/4, 1/8, 1/16. Ct values were plotted against logged dilution values and a linear model was fit on the plot to retrieve the slope values. Efficiency was calculated as below (Eq 4.1):

$$Efficiency = 10^{-1/slope} \qquad Eq\ 4.1$$

4.3.8. RT-qPCR and expression analysis

Sybr Green mix (04707516001, Roche) was used to perform RT-qPCR in LightCycler® 480 Instrument. Experiments were conducted in duplicates. 10µl was the total volume of each tube and it consisted of the following components:

- Nuclease free H2O: 1µl
- Sybr Green Mix (2X): 5µl
- Forward primer (10 pmol): 1µl
- Reverse primer (10 pmol): 1µl
- 1/10 diluted cDNA: 2µl

Ct (cycle threshold) values of technical replicates were averaged and used for calculation of the relative expression. Expression of a housekeeping gene, i.e., SDHA (RT-qPCR), was used as a reference (Table 4.1). Ct values obtained from RT-qPCR were normalized with untreated samples, and also with the housekeeping gene to observe the treatment effects. Difference between Ct values of untreated and treated samples was calculated for both the target and house-keeping gene and data were normalized with each other as described in Eq 4.2 by utilizing specific primer efficiencies (Table 4.1).

$$Relative\ expression\ (\Delta\Delta Ct) = \frac{Eff_{target}{}^{\Delta CP_{target}(control-treatment)}}{Eff_{ref}{}^{\Delta CP_{ref}(control-treatment)}} \qquad Eq\ 4.2$$

4.3.9. Collection and preprocessing of E2 treated MCF7 datasets

E2-treated microarray expression data were collected form GEO database [227]. ‘estrogen /estradiol and/or MCF7’ were used as query terms and datasets based on platforms GPL570/GPL96/GPL571 were collected. Datasets having a single replicate in any group, i.e., control or treated, were discarded from analysis. Collected cohort comprised of 13 datasets. One of these 13 datasets, GSE4668 included both E2 treatment and hormone starvation experiment in MCF7 but raw data was not available; for this dataset processed data (MAS5 normalized) was log transformed and used separately for preliminary analysis by performing *limma* differential expression between E2-treated/hormone starved samples with respective controls. Only datasets having raw data available were used in meta-analysis. Although the raw data was available for GSE8597, this dataset was separately analyzed and not used in meta-analysis since it included cycloheximide treated profiles in addition to E2 treated samples and was found to be

useful for prediction of secondary E2-induced targets by the use of a PS-score (Predicted Secondary score) I have developed (described in section 4.3.16). Remaining 11 datasets were processed for meta-analysis. In case of a time series experiment with respective control samples, each time-point along with its control was split and used independently. Furthermore, E2 treated cohort was split in two parts depending on the duration of E2 exposure. *Early E2 treatment cohort* comprised of 2-4 hours E2 treated dataset while 6-48 hours treatment datasets were categorized in *late E2 treatment group*. Quality control was performed using *affyQCreport* [45] and *affyplm* [51] package in R. All E2 treated samples in both cohorts were normalized with their respective controls by RMA using *affy* package [51].

4.3.10. Collection and pre-processing of *ER cohort*

Breast cancer patient microarray datasets were searched in GEO database [226]. Breast cancer microarray expression studies based on GPL570/GPL571/GPL96 and known ER status were then collected. The collected cohort was evaluated for availability of raw data and only those studies for which raw data available were considered for further analysis. CEL files and series matrix files (for phenotypic data) were downloaded using *GEOquery* [227] package. All datasets in *ER cohort* were processed as below:

1) Based on phenotypic data similarity, it was noticed that multiple samples in collected *ER cohort* were duplicated without any indication in GEO profiles. *Dupchecker* package [384] was used to identify duplicated expression profiles in *ER cohort*. Only a single sample from the set of the duplicated profiles was selected on the basis of data submission date to GEO. Dataset submitted on the earlier date was identified as the original sample and was kept while remaining duplicates submitted on a later date were considered a secondary source and were discarded from further analysis.
2) ER status in the cohort was generally evaluated by immunohistochemistry (IHC). In case of duplicate samples possessing conflicting ER status, none of the samples was used in the analysis.

3) Datasets having < 5 samples in any of ER-/ER+ groups were not used in further analysis.
4) Replicate samples from same patients were averaged before analysis.
5) In some of the datasets, microarray expression analyses of same patients were conducted before and after particular treatment. To avoid redundancy, only pre-treatment samples were used.
6) If the same sample was analyzed more than once, e.g., using FNA biopsy or core needle biopsy, to be able to represent a patient only once and to minimize loss of data, the strategy used for evaluation of maximum samples was selected, e.g., GSE11001 consisted of 30 samples from which RNA of 24 samples were extracted by kit A while 6 samples were extracted by kit B. 24 samples obtained from kit A were retrieved form GSE11001 and were added to *ER cohort* while 6 replicates form kit B were discarded.

Due to large number of samples in each dataset, visual analysis of quality control graphs was not possible. Quality of *ER cohort* samples was determined by outlier detection on the basis of five quality control scores. Samples considered as outliers in all of the five attributes were removed from further analysis. These five scores were scaling factor, *GAPDH*, *ACTB*, percent present and average background values. Quality evaluation was performed by using *yaqcaffy* package in R [47]. Quality controlled datasets were separately normalized using RMA.

4.3.11. Collection and pre-processing of *survival cohort*

KM plotter database, version 2010, was used to retrieve survival data of untreated and tamoxifen treated breast cancer patients [385]. Expression data of corresponding survival datasets were downloaded from GEO database. *ER cohort* and *survival cohort* were made independent by removing common breast cancer studies in both cohorts: common datasets were deleted from *ER cohort* as *survival cohort* was comparatively smaller in size. Due to large sample numbers in each dataset, quality control was performed in the same way using outlier detection method as *ER cohort*. KM plotter database has already processed samples by removing duplicated profiles. By

downloading raw data of survival datasets and running *dupchecker* script, I confirmed the KM plotter processing of duplicate samples and also received the same results [384]. RMA normalization was performed for each dataset in the cohort.

4.3.12. Meta-analyses of *E2* and *ER cohorts*

Meta-analyses of early and late *E2 treatment cohort* and *ER cohort* were separately conducted using *metaDE* package [230]. PRISMA guidelines (Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) were followed to perform meta-analysis of *E2 cohort* or *ER cohort* (Figure 4.3 and 4.4). Class labels, i.e., control and treated samples, in *E2 cohort* as well as ER- and ER+ samples in *ER cohort* along with normalized expression of all datasets were given as inputs to *metaDE* package for the calculation of effect-size and associated significance. Parameters used for *metaDE* analysis included seed value of 123 and 1000 permutations. Effect-size was determined using random effect model (REM) in case of significant heterogeneity (Cochran's Q-statistics p-value: qp-value<0.05) while fixed effect model (FEM) was used for genes having less heterogeneity in their expression (qp-value>0.05). Forest plot of specific gene was visualized using *meta* package in R [386].

4.3.13. Meta-analysis of *survival cohort*

Survival analysis of untreated patients was performed using Cox-proportional hazard model [387]. An auto best cut-off criterion to split low and high expression data had been introduced by KM plotter [385]. Using this auto-best cut-off criterion, hazard ratio (HR) was predicted for each gene in *survival cohort* datasets. Meta-analysis was used to combine the HRs obtained from multiple datasets to obtain meta-hazard ratio (mHR) and its significance. PRISMA guidelines for untreated and tamoxifen treated *survival cohorts* were shown in Figure 4.5 and 4.6. Algorithm for survival meta-analysis was described below:

For each gene in each dataset:

- Expression values were ranked in the ascending order.
- For all expression values between the first and third quartile, Cox-proportional hazard model was executed to obtain p-values.
- Expression value, which resulted in the best p-value, was selected as a cut-off.
- *Meta* package was used to combine these individual hazard ratios of each dataset into mHR [386].
- For genes showing considerable heterogeneity in expression (I^2>50%), REM was used for meta-analysis otherwise calculations were based on FEM.

Process was repeated for all genes and all datasets.

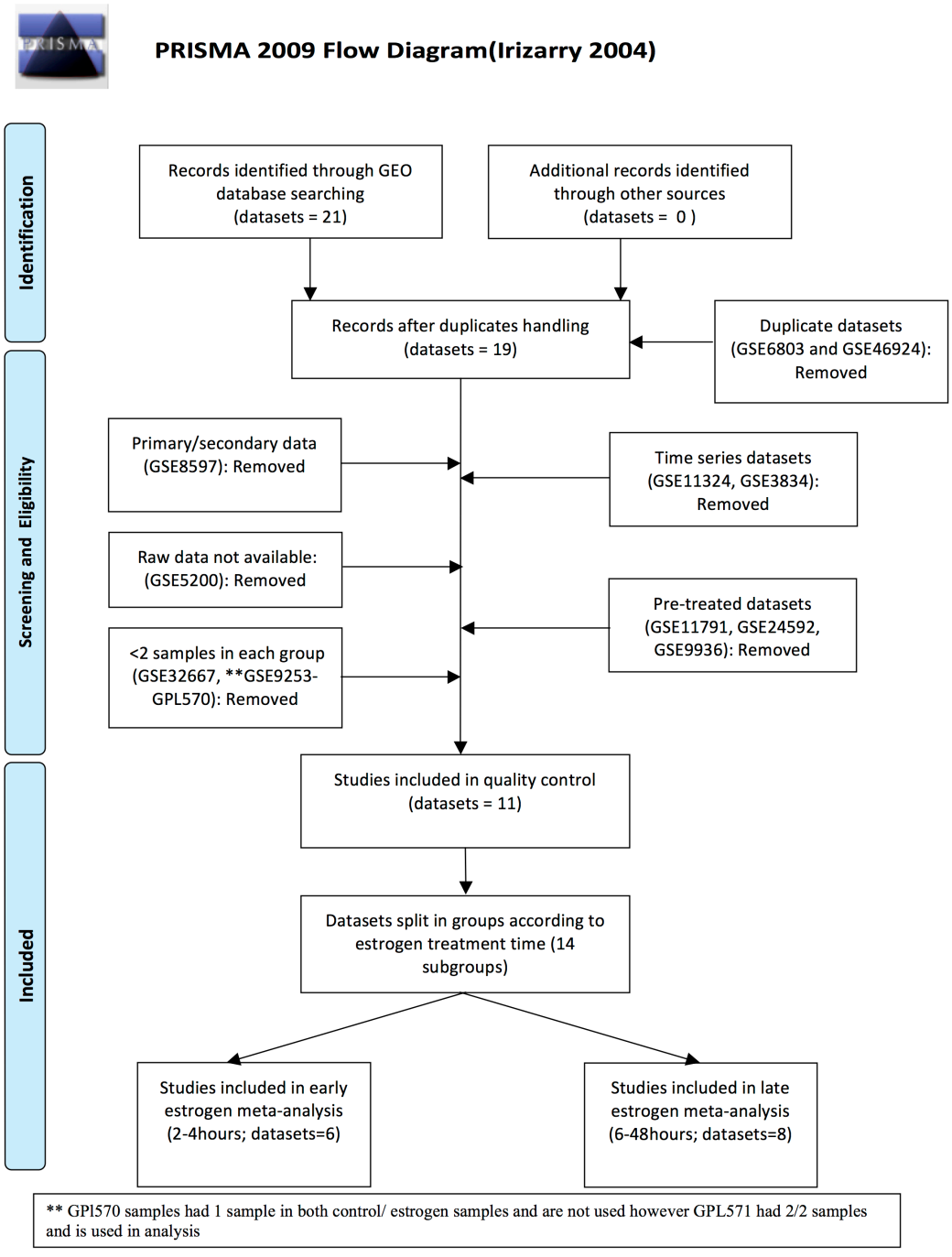


Figure 4.3: PRISMA guideline for *E2 cohort* meta-analysis.

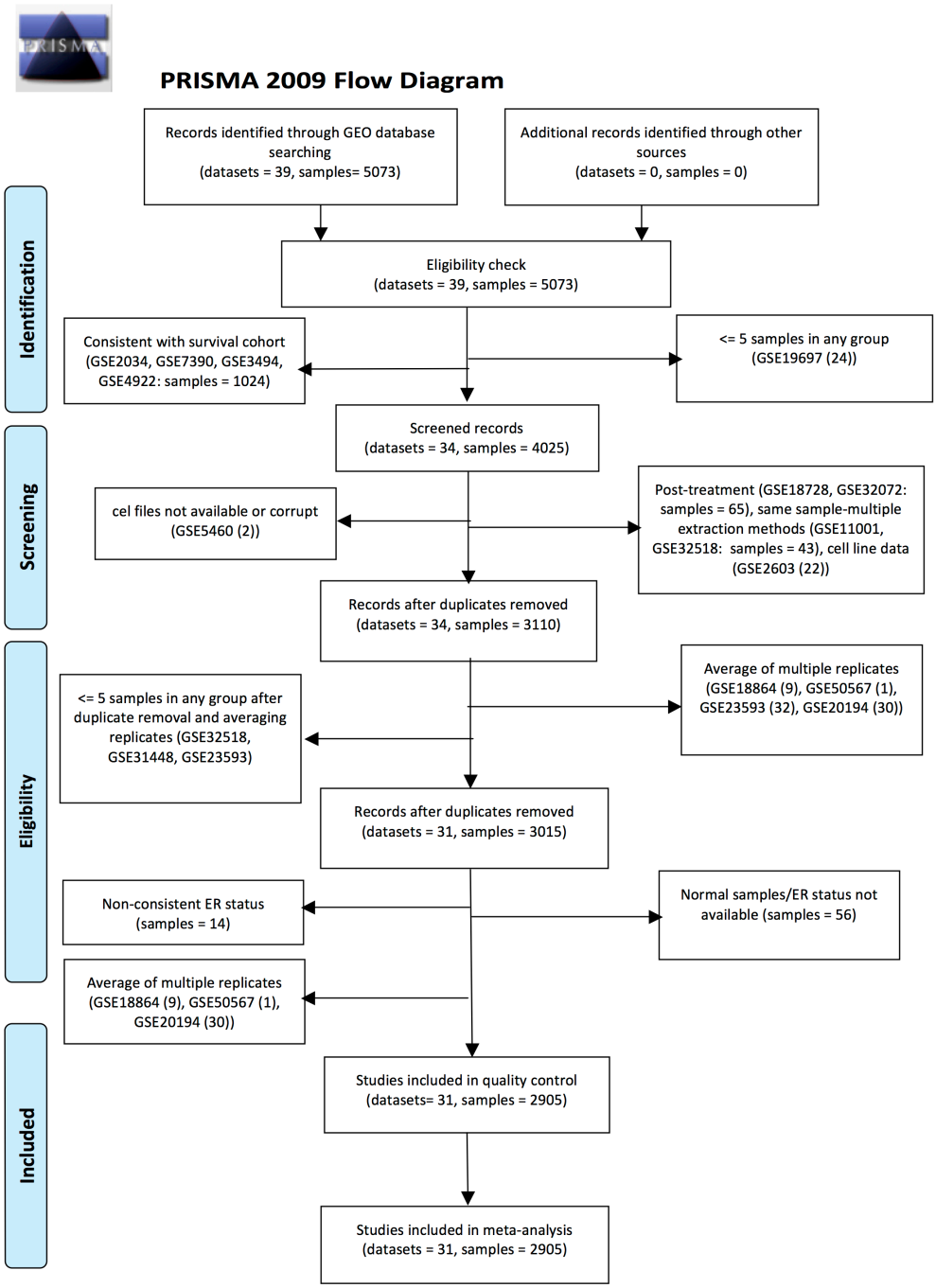


Figure 4.4: PRISMA guideline for *ER cohort* meta-analysis.

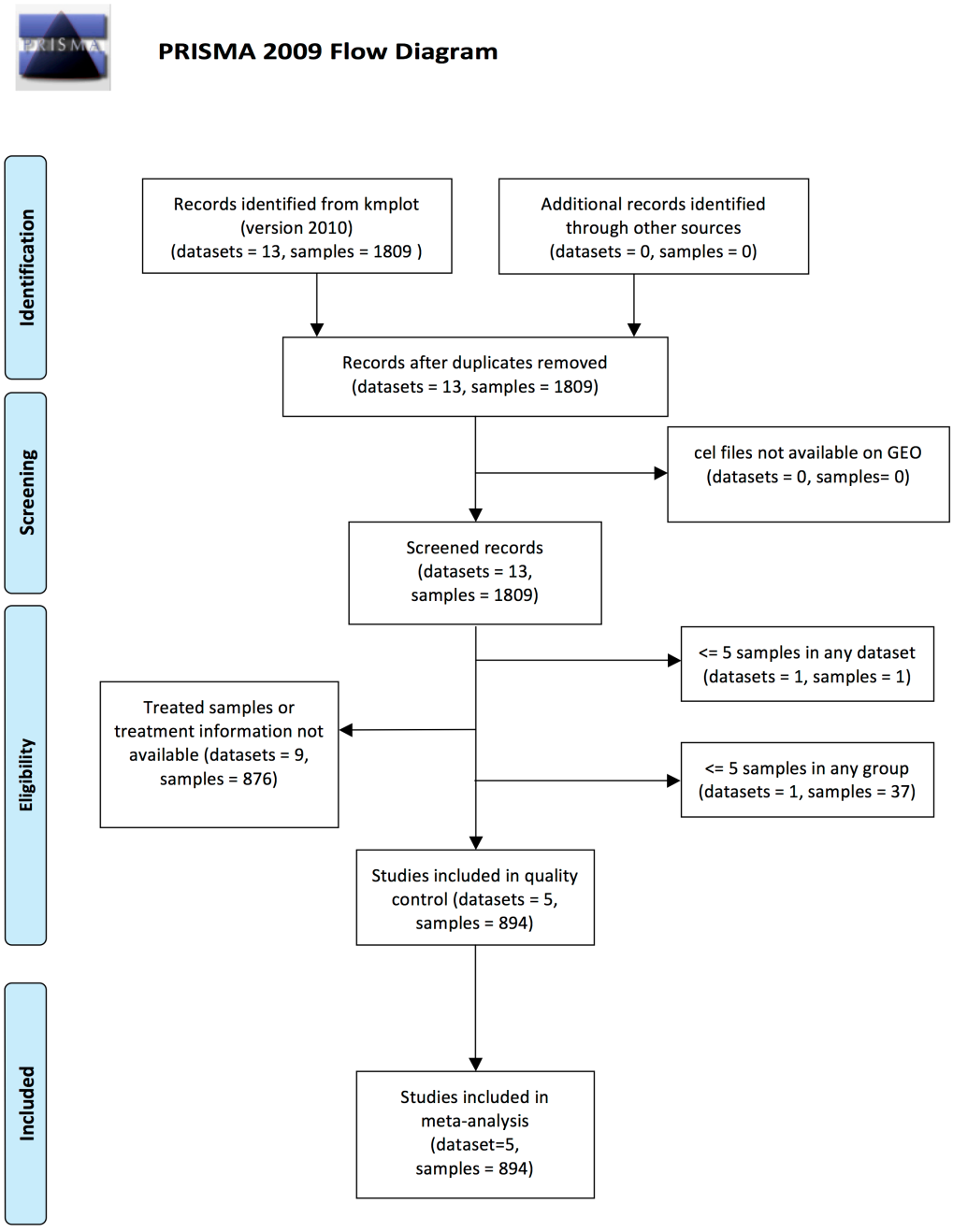


Figure 4.5: PRISMA guideline for *untreated survival cohort* meta-analysis.

4.3.14. CHRNA5 correlation analysis

Both *ER* and *survival cohorts* were used to identify the CHRNA5 correlation profile. Correlation coefficient for expression between each probeset and that of CHRNA5 was calculated for all the datasets in both cohorts. *Metacor* package was used for meta-correlation analysis by DerSimonian-Laird (DSL) random-effect meta-analytical approach [388].

4.3.15. CHRNA5 co-expression network

Weighted gene co-expression network analysis (WGCNA) pipeline using meta-analysis correlation matrix was followed to obtain and rank CHRNA5 co-expressed modules [58]. WGCNA analysis was performed on a selected gene list showing significant correlation with CHRNA5. Datasets with > 12 samples were used for co-expression analysis. Hence, 29 out of 31 datasets in *ER* cohort passed the filter. Top CHRNA5 correlated genes from *ER cohort* were selected by using meta-correlation scores (absolute r.mean > 0.2 and p<0.05, jetset best probesets). 3070 selected probesets were categorized into modules by modifying WGCNA pipeline. Strategy to obtain co-expression network has been described below:

- Pairwise correlations among 3070 genes were performed to obtain a matrix with a dimension of 3070x3070. 29 matrices of the same dimension were obtained after repeating analysis for all datasets in *ER cohort.*
- Meta-correlation analysis was performed using *metacor* package to combine these matrices into single matrix where each element (rows and columns referring to probesets with intersecting cells indicating meta-correlation score) was represented by r.mean. The resulting matrix had the same 3070x3070 dimensions.
- Adjacency matrix for the signed co-expression module was derived from resulting meta-correlation matrix and recommended soft-power threshold of 13 (Eq 4.3).

$$adjacency = 0.5\times(1 + r.mean)^{softpowerthreshold} \qquad Eq\ 4.3$$

- Adjacency matrix was converted to TOM-similarity matrix. Distance matrix was derived from TOM matrix as follows:

  Dist= *1*-TOM *Eq 4.4*

- Hierarchical clustering was performed by setting the parameters: deep-split=3; and the module size threshold =30 probesets.

4.3.16. Secondary E2 target prediction

GSE8597, that had not been included in the MCF7 E2 treatment cohort, consisted of cycloheximide treated samples in addition to the E2 treatment profile. Experimental setup of GSE8597 was explained as shown below:

Experimental groups

E2 = E2 treatment

EtOH = Control

Cyc+E2 = cycloheximide treatment 1 hour prior to E2 treatment

Cyc+EtOH = Control group of cycloheximide and E2 treatment

limma analysis was performed between two contrasts:

Group 1: E2 – EtOH; and

Group 2: Cyc + E2 – Cyc + EtOH.

PS-score was calculated as the difference between the log fold change of group 1 and group 2.

Finally, secondary E2-induced targets were selected by using the following filters:

Group 1: logFC>0.5, $p<0.05$; and PS-score>0.5

4.3.17. ER-target mapping to meta-analysis scores

Direct ER targets were obtained from a recent study, which has used ChIPseq to identify ER interacting genes [389]. Target list was retrieved and *biomart* package was used to obtain corresponding Affymetrix probesets [74]. These probesets were mapped to different scores (PS-score, depicting the degree of secondary E2 induction; CHRNA5 correlation profile obtained from survival analysis; and meta-differential expression results of *ER cohort*) to observe the behavior of E2-induced targets across the meta-analysis profiles.

4.3.18. COSMIC cell lines and TCGA data analysis

COSMIC cell line data were obtained from COSMIC website and was based on Affymetrix Human Genome U219 Array [172]. TCGA RNAseq level-3 RSEM normalized count data of breast cancer patients were downloaded from Firebrowse (http://firebrowse.org). Scatter plot and correlation analysis was performed in R.

4.3.19. Shiny web-application: E2S

4.3.19.1. Data collection

Gathered results of *E2 treatment cohorts*, *ER cohort* and *survival cohort* were used to develop a web-tool named 'E2S: Estrogen (E2) to Survival'. In addition to the untreated breast cancer patient data, tamoxifen treated breast cancer patient data from KM plotter database (2013) were also added into E2S and was processed similar to untreated patient data (section 4.3.13). Similar to the meta-analysis of untreated breast cancer patients, meta-survival analysis of tamoxifen treated patients was performed as described above. PRISMA guidelines of the tamoxifen treated cohort were shown in Figure 4.6.

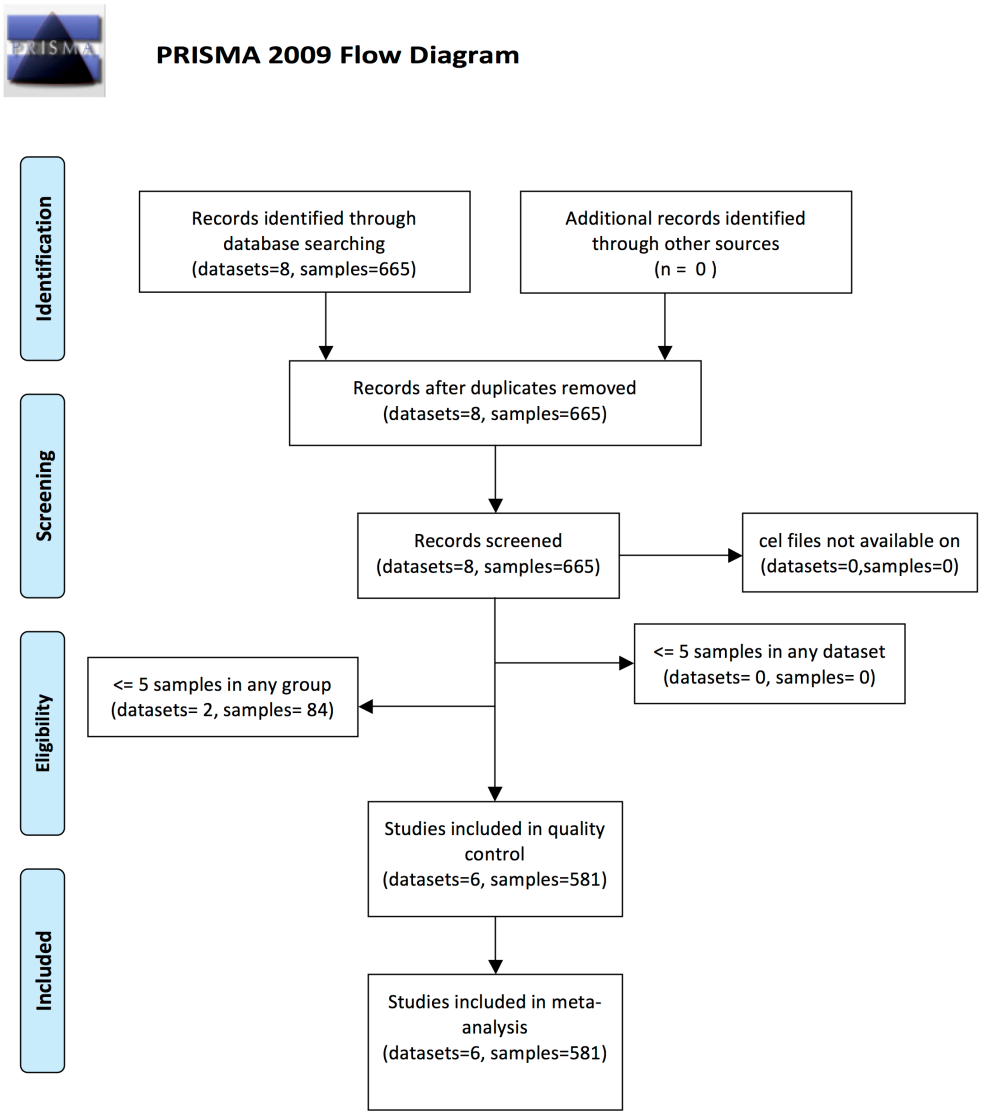


Figure 4.6: PRISMA guidelines for *tamoxifen survival cohort* meta-analysis.

4.3.19.2. Ready to use results and run-time processing

Normalized expression data from *E2 cohort*, *ER cohort* and *survival cohort* were already stored; upon user request, a queried gene could be searched from the addressed cohort; and using *meta* package meta-analysis results can then be calculated and visualized in the form of a forest plot [386].

Genome-wide meta-differential expression results of *late E2 cohort* and *ER cohort* were already prepared as mentioned in previous section (section 4.3.12). Similarly, genome-wide meta-survival analysis results of untreated and tamoxifen treated patients were also prepared and placed in E2S to extract data of user requested gene. The run-time calls during the E2S execution are described below:

Above mentioned four genome-wide meta-analysis scores are called and used in '***module IV: signaling and network***'; and are also visualized either as a color code or on the x-axis. However, meta-correlation analysis of user entered query gene with whole transcriptome (presented on y-axis) is calculated at run-time using *survival cohort*. After calculation of meta-correlation scores between user entered gene and whole transcriptome, scores are ranked to identify the most and least correlated genes. Functional analysis of top correlated genes is also conducted at the run-time using *clusterprofiler* [89, 390] and *ReactomePA* packages [89] in R. Interaction network of correlated gene is obtained from the well-known interaction database STRING at run-time using *stringdb* package [391]. GSE8597 genome-wide differential expression analysis results and PS-scores are ready to use and only searched and plotted upon entry of the query gene. Jetset annotation system is used for mapping between probesets and genes [381].

4.3.19.3. Graphical presentation

RcytoscapeJS package was used to draw and present customizable interaction network [392]. Functional analysis results were arranged in panels using *ShinyBS* package [393]. *Ggvis* package was used to visualize meta-correlation profile graph. Graph was

customized to visualize tooltip on each dot representing gene name [394]. Javascript was used to visualize wait messages during processing time. HTML tags were used to provide small notes about each module.

4.4. Results

4.4.1. Evaluation of epithelial to mesenchymal transition (EMT) in CHRNA5 depleted MCF7 cells

MCF7 is an epithelial cell line; and observed lamellipodia and filopodia-like projections are generally not associated with epithelial cells but might indicate underlying transition of epithelial cells towards mesenchymal cell type [395]. I evaluated the effects of CHRNA5 RNAi on EMT by analyzing mRNA expression (by RT-qPCR) of two well-reputed epithelial and mesenchymal markers (CDH1 and VIM, respectively). Increased expression of *VIM* (vimentin) and downregulation of *CDH1* (e-cadherin) are expected to be primary indicators during EMT. *limma* results of CHRNA5 depleted MCF7 microarray analysis (Ermira Jahja, PhD thesis, 2017) revealed a mild but significant upregulation of *VIM* in CHRNA5 depleted cells (probeset: 201426_S_AT, logFC: 0.25, p=0.02; Figure 4.7A). RT-qPCR evaluation showed no significant changes of this gene at 72h (logFC=0.15, p=0.12; Figure 4.7A) but relatively upregulated expression of VIM at 120h was approaching significance (logFC=1, p=0.08; Figure 4.7B). An independent 120h siRNA-1 treatment validated the significant but mild increase in VIM expression (logFC=0.84, p=0.048; Figure 4.7C). On the other hand, a significant change in the epithelial marker *CDH1* was observed neither in microarray nor in the RT-qPCR data at the considered time points or with tested siRNAs (Figure 4.7A-C).

Recently, Chang et al. have induced EMT by using a time-course TGF-β treatment in human small cell lung cancer cell line (A549) and analyzed the gene expression changes by RNAseq [380]. 1633 differentially expressed genes have been categorized in three groups, i.e., epithelial (930 genes), partial EMT (200), and mesenchymal (503) associated genes [380]. After retrieving the list of epithelial and mesenchymal associated genes and matching them with Affymetrix HG-U133 Plus 2 probesets, the distribution of this gene list with respect to CHRNA5 RNAi fold change data (1325 jetset best probesets) was evaluated. Considering p-values obtained from the *limma* analysis, I found that expressions of 756 out of 1325 probesets were significantly modulated upon CHRNA5 depletion. Distribution of these epithelial and mesenchymal marker genes

showed that neither epithelial nor mesenchymal genes were confined to one-side of the graph (i.e., entirely to positive or negative logFC). However, epithelial genes were largely clustered in downregulated category upon CHRNA5 depletion (445 out of 547 probesets) yet 102 epithelial marker probesets were still significantly upregulated. On the other hand, 209 out of 454 mesenchymal marker probesets were significantly affected by the siRNA-1 treatment; 125 of these probesets exhibited significant upregulation but a considerable number (84 marker probesets) were also downregulated (Figure 4.7D). Mixed expression of mesenchymal and epithelial genes and no significant downregulation of CDH1 revealed that downregulation of CHRNA5 did not initiate a complete EMT transcriptomic signature in response to CHRNA5 RNAi.

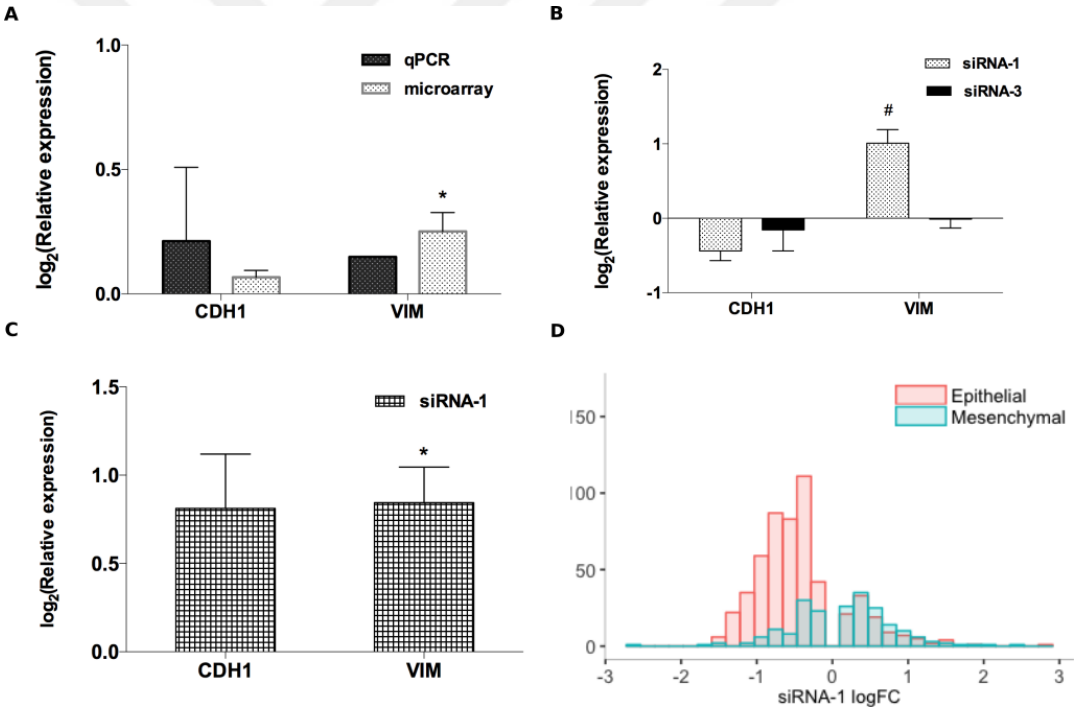


Figure 4.7: Analysis of epithelial to mesenchymal transition in CHRNA5 depleted MCF7 cells.

Epithelial and mesenchymal marker genes expression (CDH1 and VIM respectively) (A-C) A) Comparison of RT-qPCR and microarray (siRNA-1; 72h; n=2 per group). B) Effects of multiple siRNAs and time points i.e., siRNA-1: 120h (2 siRNA treated compared with 1 scrambled) and siRNA-3: 120h (2 siRNA treated samples compared with 1 scrambled). C) Validation by an independent siRNA-1 (120h) experiment (n = 2 per group). D) Distribution of significantly modulated epithelial and mesenchymal marker genes (retrieved from [380]) in siRNA-1 microarray data using histograms. # $p<0.1$, * $P<0.05$.

4.4.2. The association between CHRNA5 RNAi and differentiation expression profiles

A repository of the expression data from different differentiation experiments was collected by searching and categorizing the GEO for normal cell differentiation (except in the case of neuroblastoma) microarray studies based on GPL570 [226]. No gene silencing or overexpression study leading to or preventing from differentiation was collected. In addition, no cancer cell line studies were included. Collected datasets (n=11) were processed and categorized into epithelial datasets (i.e., mammary, keratinocytes, bronchial, urothelial, and hepatocytes) or non-epithelial categories (i.e., erythrocytes, adipocytes, chondrocytes, and neuronal cells) (Table 4.2). Splitting of datasets and respective *limma* comparison groups were also described in Table 4.2. *Limma* results (jetset best probesets) from each dataset were compared with the expression profile of the CHRNA5 RNAi microarray analysis (GSE89333; Ermira Jahja, PhD Thesis, 2017). Association statistics (Odds ratio along with Fisher's exact tests) were used to measure the overlap between each differentiation dataset and the CHRNA5 depletion expression signature (Figures 4.8, 4.9, 4.10).

Percentage of significantly shared probesets was identified as the ratio between the number of shared significant probes (in each differentiation category and CHRNA5 depletion profile) and total significant probesets in respective differentiation dataset/time-point (Table 4.2). Interestingly, CHRNA5 depletion signature shared a large number of significantly modulated probesets with several datasets, most of which can be classified as epithelial tissue differentiation profiles (Table 4.2). These percentages ranged between 40-48% for epithelial and 35-41% for non-epithelial datasets except neuronal cell differentiation that is non-epithelial but still percentage of shared significant probesets reached up to 48%. The details of data and analyses for each tissue have been provided in the following paragraphs:

Table 4.2: Differentiation datasets used for the comparative analysis.

Dataset	Type	Differentiation time point (N) versus control time point (N)	Analysis	Shared significant probesets (Percent)	Meta-analysis (differentiation / control)
GSE4655 [396]	Human adult-derived peripheral blood CD34+ cells	Day 3 (3), 5 (3), 7 (3), 9 (3), 11 (3) versus day 1 (3)	U	35%, 39%, 41%, 42%, 41%	--
GSE10070 [397]	Breast differentiation on Transwells (MCF10A)	Base (3), Midpoint (3), Plateau (4) versus Monolayer (3)	U	46%, 45%, 45%	Plateau/ Monolayer
GSE11839 [398]	Human urothelial cells; control subjects	Day 2 (3) versus day 0 (3)	P	47%	2 days/0 day
GSE16464 [399]	Human chondrocytes; normal donors	Day 14 (3) versus day 0 (3)	P	42%	--
GSE18590 [400]	Primary human keratinocytes	Day 3 (2) versus day 0 (2)	U	42%	3 days / 0 day
GSE52651 [401]	Primary human keratinocytes	Day 1 (2), 2 (2), 3 (2), 4 (2), 5 (2), 6 (2), 7 (2) versus day 0 (2)	U	43%, 45%, 45%, 47%, 47%, 47% 48%	7 days / 0 day
GSE39059 [402]	Human bronchial epithelial cells	Day 28 (3) versus confluent (3)	U	43%	28 days / confluent
GSE61302 [403]	Adipose derived Stem Cells	Day 7 (4), 21 (6) versus undifferentiated (5)	U	41%, 40%	--
GSE9169 [404]	Neuronal cells differentiation; SH-SY5Y (ATCC)	Hour 6 (2); day 1 (2), 2 (2), 3 (2), 5 (2) versus hour 0 (2) Retinoic Acid	U	38%, 44%, 47%, 48%, 48%	--

Dataset	Type	Differentiation time point (N) versus control time point (N)	Analysis	Shared significant probesets (Percent)	Meta-analysis (differentiation / control)
		treatment			
GSE14897 [405]	Hepatic cells differentiation; WAO9 cells	Day 20 (3) versus undifferentiated (3)	U	40%	20 days / undifferentiated

Columns indicate GEO identification number; origin of sample; time points used for comparison along with number of samples used are indicated in parentheses; kind of *limma* analysis (paired (P) and unpaired (U)); percent of shared significant probesets between each comparison and CRHNA5 depletion profile; comparison group used in epithelial meta-analysis (Single time point was chosen from each epithelial differentiation repository; described in materials and methods section). Shared probesets were calculated as percentage of ratio among the number of shared significant probesets (in each differentiation comparison and CHRNA5 depletion profile) and total number of significant probesets in respective differentiation combinations.

4.4.2.1. Epithelial differentiation datasets and association with CHRNA5 RNAi

Seven out of 11 collected differentiation profiles were based on epithelial tissues. Analysis of each dataset and its comparison with CHRNA5 depletion profile was described below:

Mammary cells

Breast differentiation has been studied in the literature using transwell differentiation of MCF10A [397]. MCF10A transwell model mimics multiple features of *in-vivo* differentiation model. Transwell cultures offer permeable support to both anchorage dependent and independent cells representing an efficient way of differentiation compared to pellet culture. Transwell culture provides cell-cell contact, and diffusion from both sides (above and below) enabling easy access of nutrients to all the cells [406]. Tight junctions of epithelial cells in transwell culture model are generally evaluated by quantitative measurement of trans-epithelial/trans-endothelial electrical resistance (TEER) [407]. MCF10A differentiation expression data [397] that I

reanalyzed, four experimental conditions were used for gene-expression analysis. The authors obtained the monolayer cell from conventional culture at 5th day after cells reached confluence (control group). Differentiated samples obtained at three different states according to TEER measurement were reported: transwell samples obtained at 200–300 Ω*cm^2, midpoint at 1400–1600 Ω*cm^2 and Plateau at 3000–3200 Ω*cm^2 [397].

I compared expression data from all three-differentiation states against the monolayer using *limma* analysis. CHRNA5 expression in each of the three-differentiation states was highly downregulated (Base: logFC=-1.74, p=1.84E-05; Mid-point: logFC=-2.33, p=5.4E-07; Plateau: logFC=-2.45, p=5.3E-09; Figure 4.8A). When expression profiles of all three-differentiation states were compared with CHRNA5 RNAi expression profile using shared significant probesets, the number of probesets modulated in same direction (i.e. upregulated or downregulated in both datasets; Base: 2826, Mid-point: 3126, Plateau: 3109; Figure 4.8A) were higher in number than the ones modulated in the opposite direction (Base: 1337, Mid-point: 1536, Plateau: 1525). Comparison of the shared significant profiles using Odds ratio and Fisher's exact test resulted in similar and highly significant positive associations in all three-differentiation states (Base: OR=4.49, p=2e-120; Mid-point: OR=4.19, p=1.6E-123; Plateau: OR=4.22, p=9.1E-124; Figure 4.8B). These findings indicated that CHRNA5 was indeed downregulated during transwell mammary epithelial cell differentiation and CHRNA5 mediated expression changes in breast tumorigenic cell line MCF7 could also be highly associated with normal mammary differentiation program.

Keratinocytes

Two keratinocyte differentiation datasets studied in epithelial differentiation profiles were GSE18590 [400] and GSE52651 [401]. They represented calcium-induced differentiation and time-course differentiation, respectively, in primary keratinocytes isolated from foreskin. Interestingly, calcium induced keratinocyte differentiation (GSE18590; day 3 [400]), like MCF10A differentiation, also exhibited a strongly significant *CHRNA5* downregulation (logFC=-2.51; p=2.22E-06; Figure 4.8A).

Comparison of shared, significant keratinocyte differentiation transcriptome with CHRNA5 depletion profile again showed a significant positive association ([400]; OR=3.04, p=8E-77; Figure. 4.8B). The keratinocyte dataset based on the differentiation time-series (GSE52651; 0-7 days [401]) was evaluated by comparing the day 1-7 expression profiles with the day 0 expression profile. Although expression profile at day 1 did not result in significant CHRNA5 downregulation and the number of shared significant probesets was comparatively less (43%), a significant positive association between day 1 and CHRNA5 RNAi expression signatures was still present. However, starting from day 2, the number of shared significant probesets started increasing with time (day 2: 45%, day 3: 45%, day 4: 47%, day 5: 47%, day 6: 47%, day 7: 48%), along with significant downregulation of CHRNA5 probeset (day 2: logFC=-0.89, p= 0.02; day 3: logFC=-0.88, p=0.01; day 4: logFC= -1.26, p=5.5E-03; day 5: logFC= -1.27, p=2.5E-03; day 6: logFC= -0.97, p=0.02; day 7: logFC= -1.07, p = 0.02; Figure 4.8A). Positive association between the keratinocyte differentiation and CHRNA5 RNAi also became more prominent with time (day 2: OR=3.12, p=5.49E-32; day 3: OR=2.57, p=1.21E-22; day 4: OR=3.73, p=4.88E-53, day 5: OR=4.01, p=7.93E-62; day 6: 3.45, p=4.23E-50; day 7: 4.04, p=4.26E-64; Figure 4.8B) yet when compared with calcium induced keratinocyte differentiation, these statistics were lesser (Figure. 4.8; Table 4.2). Results indicated that CHRNA5 depletion and keratinocyte differentiation programs indeed shared an underlying transcriptomic regulation; and this overlap was more enhanced in the calcium-induced differentiation. This observation can most probably relate to a previously attributed role of CHRNA5 to regulate calcium signaling by facilitating calcium influx into the cell [408].

Lung

Two datasets, GSE39059 [402] and GSE5264 [409] were used to obtain a differentiation signature in lung tissue. GSE39059 samples were obtained from primary normal human bronchial epithelial cells and differentiated on air-liquid interface (ALI) culture [402]. ALI culture is an efficient medium of lung cells differentiation and polarization because of its close resemblance to *in-vivo* airway epithelial niche [410]. The authors performed

gene expression analysis at three experimental conditions with 2 stages depending on cell confluency 1) sub-confluent cells and 2) confluent cells, and the differentiated samples were obtained at 28 days after ALI culture. In *limma* re-analysis I performed, the confluent cells (control) were compared with 28 days differentiated samples. CHRNA5 downregulation in GSE39059 (logFC=-2.2, p=4.63E-07; Figure 4.8A) and a significant positive association of bronchial differentiation with CHRNA5 depletion signature (OR: 2.17, p=9.41E-41; Figure 4.8B) validated that CHRNA5 mediated changes might be involved in differentiation of lung tissue as well. However, the second dataset, GSE5264 did not show any significant modulations in CHRNA5 expression levels upon differentiation. Samples were based on bronchial epithelial cells form three different donors and gene expression was performed in time-series manner (day 1, 4, 8, 10, 14, 21, 28) upon culturing in ALI system. The CHRNA5 expression modulation between undifferentiated and differentiated cells was in general insignificant (Day 1: logFC=0.49, p=0.03; day 4: logFC=0.11, p=0.29, day 8: logFC=-0.05, p=0.8, day 10: logFC=-0.18, p=0.53, day 14: logFC=-0.21, p=0.35, day 21: logFC=-0.2, p=0.4, day 28: logFC=-0.01, p=0.98; Figure 4.9A). Although shared expression profile of time-series and CHRNA5 RNAi expression signature ranged from 44-47%, association with CHRNA5 depletion profile was highly variable and overall insignificant. Even in the cases where it was significant, Odds-ratio was comparatively less than other evaluated epithelial datasets and ranged between 0.31-1.53 (Figure 4.9B).

Hepatocytes

In the hepatocyte differentiation dataset (GSE14897), the authors presented an efficient way of reprogramming human embryonic stem cells (hESC) to hepatic cells so that 80% of resulting differentiated cells expressed albumin protein, a characteristic feature of differentiated hepatic cells [405]. WAO9 human embryonic stem cells harvested by following hepatic differentiation protocol at day 20 and hESC along with differentiated samples were compared using microarray analysis [405]. I have found upon *limma* tests that consistent with the other epithelial dataset findings, the reprogramming of ESC to hepatocytes resulted in significant CHRNA5 depletion (logFC: -2.41, p=1.94E-06;

Figure 4.8A) and significant positive association with CHRNA5 depletion profile (OR=2.16, p=5.9E-36; Figure 4.8B).

Bladder

Bladder biopsy of interstitial cystitis and control (normal women having stress inconsistency) was differentiated by culturing in high-calcium and serum containing media for 2 days and gene expression was evaluated by microarrays (GSE11839) [398]. I compared expression profile of only normal samples (between the undifferentiated and differentiated samples) using *limma* analysis. 47% of significant probesets from bladder differentiation dataset were shared with those from CHRNA5 depletion profile. Similarly, as expected, a significant CHRNA5 depletion (logFC=-0.58, p=5.68E-03; Figure 4.8A) along with a significant positive association with CHRNA5 siRNA-1 profile (OR: 2.19, p=3.07E-20; Figure 4.8B) was observed validating that association of CHRNA5 with differentiation could be relatively widespread across different epithelial tissues.

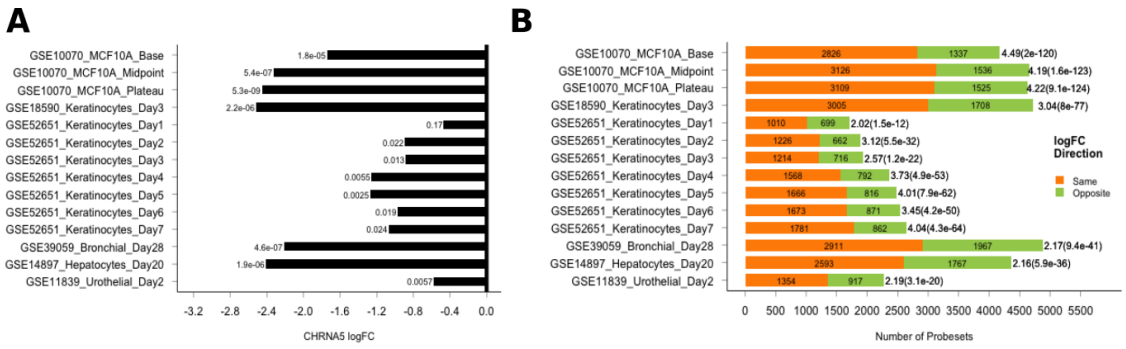


Figure 4.8: Comparison of CHRNA5 RNAi expression profile with epithelial differentiation expression cohort.

A) Expression of CHRNA5 probeset in epithelial differentiation datasets. P-value obtained form limma are shown in front of each bar. B) Association of CRHNA5 depletion expression profile with epithelial differentiation transcriptome studies. Orange bar indicate significantly shared probesets in same direction, i.e., (concordent up or downregulation based on limma fold change) while green bar indicates opposite direction of significantly shared probeset (upregulation in CHRNA5 depletion and downregulation in differentiation profile and vice-versa) in both differentiation and CHRNA5 depletion profile.

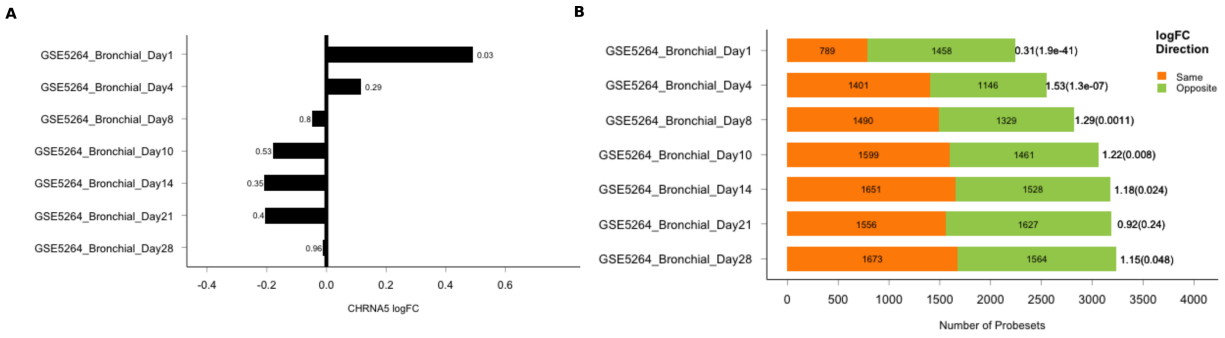


Figure 4.9: Association of CRHNA5 depletion expression profile with lung differentiation study.

This study is an outlier from epithelial differentiation pattern. A) Expression of CHRNA5 probeset in lung differentiation datasets. P-values obtained form limma are shown in front of each bar, B) Association of CRHNA5 depletion expression profile with epithelial differentiation transcriptome studies. Orange bar indicate significantly shared probesets in same direction i.e. (concordent up or downregulation based on limma fold change) while green bar indicates opposite direction of significantly shared probeset (upregulation in CHRNA5 depletion and downregulation in differentiation profile and vice-versa) in both differentiation and CHRNA5 depletion profile. Number written on bar indicate number of probeset in each category while number written in front of bar indicate association statistics as Odds ratio score (Fisher's exact p-value).

4.4.2.2. Comparison of non-epithelial differentiation with CHRNA5 RNAi signature

Collected differentiation profiles consisted of 4 non-epithelial tissues including erythrocytes, chondrocytes, adipocytes and neuronal cells and details of analyses were given below:

Erythroid cells

In GSE4655, human adult-derived peripheral blood CD34+ hematopoietic progenitor cells were differentiated and the expression profiling conducted at 6 different time-points (day 1, 3, 5, 7, 9, 11) was used for understanding of the transcriptomic changes associated with erythroid differentiation [396]. As day 1 was earliest time-point available, I based the *limma* analysis by comparing it with all the other time points. The other time point profiles, i.e., day 3 and day 5, when compared with CHRNA5 RNAi expression profile did not reveal any significant expression change for CHRNA5 probeset (day 3: logFC=-0.59, p=0.44; day 5: logFC=0.17, p= 0.71; Figure 4.10A); and contrary to the epithelial differentiation profile, a significant negative association was observed (day 3: OR=0.31, p =3.38E-14, day 5: OR=0.57, p=8.25E-08; Figure 4.10B) yet the number of shared significant probesets was comparatively less (35%, and 39%). Day 7 comparison showed a considerable number of shared probesets (41%) yet again no significant CHRNA5 expression change (logFC=-1.33, p=0.22; Figure 4.10A) and lesser but a significant negative association with CHRNA5 depletion profile (OR=0.7, p=7.55E-05; Figure 4.10B) were observed. On the other hand, at day 9, percentage of shared probesets (42% and 41%) increased and a significant reduction of CHRNA5 expression was observed (day 9: logFC=-1.86, p=0.04, day 11: logFC=-1.52, p=0.03; Figure 4.10A) and the association with CHRNA5 depletion profile started showing a smaller but positive trend (day 9: OR=1.19, p=0.099, day 11: OR=1.28, p=0.0054; Figure 4.10B). These findings revealed inconsistent association patterns between

CHRNA5 and the erythroid differentiation profile yet with increasing time even the erythrocyte differentiation exhibited similarity with CHRNA5 RNAi profile.

Chondrocytes

In GSE16464, the samples were retrieved from human donors (normal and osteoarthritis patients) and chondrogenic differentiation was induced using 3D culture for 14 days [411]. Gene expression analysis of monolayer versus 3D culture samples obtained from normal donors was used to obtain the chondrocyte differentiation expression signature. No significant CHRNA5 depletion was observed during chondrocyte differentiation (logFC=0.02, p=0.89; Figure 4.10A) yet a significant positive association with CHRNA5 RNAi profile was observed. However, relatively smaller number of genes, in comparison with epithelial differentiation profiles, were positively (656 probesets) or negatively modulated (491 probesets; OR=1.79, p=1.2E-06; Figure 4.10B).

Adipocytes

In GSE61302 [403], the authors obtained adipocyte tissue derived stromal stem cells from 6 females and in differentiation media these cells were converted to mature fat cells. 7- and 21-day differentiation profiles were separately compared with undifferentiated samples herein. CHRNA5 expression was not modulated during adipogenesis (day 7: logFC=0.21, p=0.54; day 21: logFC=-0.35, p=0.14; Figure 4.10A). In accord with other non-epithelial datasets, no association with CHRNA5 depletion profile was observed at an earlier time of differentiation (day 7) (OR=1.03, p=0.75) while day 21-differentiation profile resulted in a significant positive association (OR=1.37, p=4e-05; Figure 4.10B).

Neuronal cells

Human neuroblastoma cell lines (SK-N-SH and SH-SY5Y) were treated with retinoic acid to induce differentiation (GSE9169; [404]). I reanalyzed samples collected at different time-points (0h, 6h, 1, 2, 3, 5, day) using *limma*. In my analysis, I used SH-

SY5Y-A samples (obtained from ATCC) for comparative analysis that resulted in efficient differentiation. Although 6h samples had very less number of shared probesets with CHRNA5 RNAi profile (UPUP+DOWNDOWN: 119, UPDOWN+DOWNUP: 170), a significant negative association with CHRNA5 depletion profile was observed (OR=0.45, p=0.0012; Figure 4.10B). Starting form day 1, number of shared probesets comparatively increased and the trend of association with CHRNA5 depletion profile changed to positive (day 1: OR=1.68, p=4.5E-05; day 2: OR=4.82, p=1.82E-62; day 3: OR=5.61, p= 8.4E-87; day 5: OR=5.8, p =1.1E-93; Figure 4.10B). However, interestingly at all tested time-points, the expression of CHRNA5 probeset was not modulated (6h: logFC=0.05, p=0.74; day1: logFC=-0.11, p=0.45; day2: -0.14, p=0.37; day3: logFC=-0.05, p=0.73; day5: logFC=0.01, p=0.96; Figure 4.10A). Results indicated a probable CHRNA5-independent mechanism of differentiation; similar to the one induced by CHRNA5 RNAi, in neuronal cells upon retinoic acid treatment.

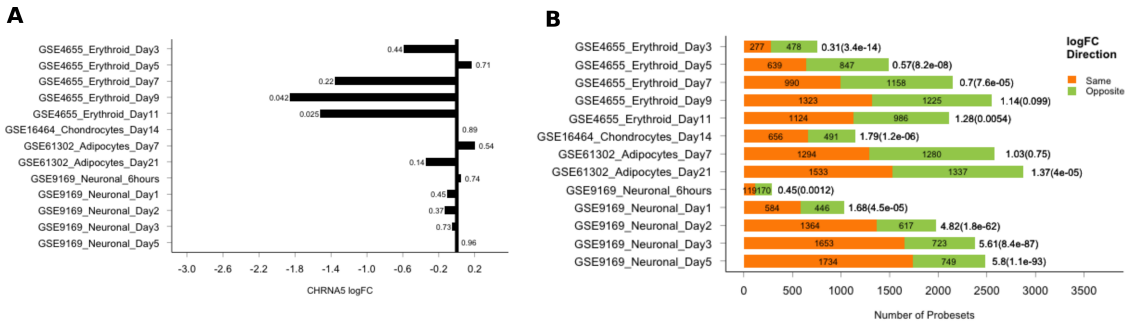


Figure 4.10: Comparison of CHRNA5 RNAi expression profile with non-epithelial differentiation expression cohort.

A) Expression of CHRNA5 probeset in non-epithelial differentiation datasets. P-value obtained form limma are shown in front of each bar. B) Association of CRHNA5 depletion expression profile in non-epithelial differentiation cohort. Orange bar indicate significantly shared probesets in same direction i.e. (concordent up or downregulation based on *limma* fold change) while green bar indicates opposite direction of significantly shared probesets (upregulation in CHRNA5 depletion and downregulation in differentiation profile and vice-versa) in both differentiation and CHRNA5 depletion profile. Number written on bar indicate number of probeset in each category while number written on top of bar indicate association statistics as Odds ratio score (Fisher's exact p-value).

4.4.2.3. Identification of well-known differentiation markers in CHRNA5 depleted transcriptome

Several well-known regulators of differentiation were searched form literature to compile their role along with *limma* results of CHRNA5 RNAi and MCF10a transwell differentiation and were presented in Table 4.3. DNMT1 and UHRF1, known to be negative regulators of differentiation, and expectedly were found to be downregulated in CHRNA5 depletion and normal breast differentiation profile. Positive regulators of differentiation included GADD45A, GADD45B, FDXR, MPZL3, ZNF750, KLF4, GRHL3 and TP63; four of these regulators (MPZL3, GADD45A, ZNF750 and GRHL3) were significantly induced in both CHRNA5 depleted MCF7 and normal breast differentiation signature. However, TP63 and GADD45B showed no significant changes in CHRNA5 depletion profile; FDXR (downregulated) and KLF4 (non-significant) did not follow the trend in normal breast differentiation profile.

Table 4.3: Differentiation regulators along with their expression in CHRNA5 depleted MCF7 cells and normal breast differentiation (limma results).

	Gene description	**SiRNA-1 (MCF-7; *limma*)**		**Breast differentiation (GSE10070: Plateau; *limma*)**		**Differentiation related role**
Symbol	**Probesets**	**logFC**	**P.Value**	**logFC**	**P.Value**	
DNMT1	201697_S_AT	-0.59	**1.6E-04**	-2.06	**2.1E-08**	Repress differentiation by methylating differentiation genes [401].
UHRF1	225655_AT	-1.25	**3.0E-06**	-2.94	**2.1E-09**	Interacts with DNMT1 to suppress differentiation and maintaining progenitor state [412, 413].
FDXR	207813_S_AT	1.42	**4.8E-06**	-0.59	**4.9E-06**	Interacts with MPZL3 to facilitate ROS during epidermal differentiation [414, 415].
GADD45A	203725_AT	2.31	**9.6E-08**	0.80	**4.5E-05**	Regulate calcium induced differentiation and demethylation of differentiation genes [400, 416].
MPZL3	227747_AT	1.39	**5.1E-06**	1.18	**9.6E-08**	Together with FDXR, regulates epidermal differentiation by ROS production [414, 415].
ZNF750	219995_S_AT	0.53	**4.3E-04**	4.17	**3.3E-11**	Transcription factor: regulates expression of early and late differentiation gene [401, 415].
KLF4	221841_S_AT	1.9	**1.3E-05**	0.11	3.3E-01	Transcription factor: regulates expression of early and late differentiation gene [401, 415].

	Gene description	SiRNA-1 (MCF-7; *limma*)		Breast differentiation (GSE10070: Plateau; *limma*)		Differentiation related role
Symbol	**Probesets**	**logFC**	**P.Value**	**logFC**	**P.Value**	
GRHL3	232116_AT	0.72	**1.0E-02**	2.58	**2.3E-10**	Transcription factor: regulates expression of early and late differentiation gene [401, 415].
TP63	209863_S_AT	-0.01	0.88	0.75	5.8E-07	Transcription factor: regulates expression of early and late differentiation gene [401, 415].
GADD45B	207574_S_AT	0.10	0.42	0.94	4.5E-07	Regulate calcium induced differentiation and demethylation of differentiation genes [400, 416].

4.4.2.4. Meta-analysis of epithelial differentiation and comparison with CHRNA5 depletion profile

CHRNA5 depletion exhibited a highly positive association with a variety of epithelial tissue differentiation profiles. However, a meta-analysis of various epithelial differentiation profiles has been lacking in the literature. This prompted me to obtain a generalized epithelial meta-differentiation signature using the datasets included in Figure 4.8. If datasets consisted of a single differentiation time-point and one control group, it was directly used in the meta-analysis. To avoid redundancy, only one time-point was chosen from each dataset having multiple time-points. A time-point showing the most significant Fisher's exact p-value indicating highest association with CHRNA5 depletion profile was selected. In addition, it was mandatory for the chosen time-point expression

profile to show significant downregulation of CHRNA5 probeset ($p<0.05$). These two filtration criteria helped us establish a link of epithelial differentiation meta-signature with the CHRNA5 depletion signature (Figure 4.8; Table 4.2). Six datasets were used in the meta-analysis (GSE10070: mammary cell differentiation (Plateau), GSE18590: keratinocyte differentiation, GSE52651: keratinocyte differentiation (7 day), GSE39059: lung differentiation, GSE14897: Hepatocytes, GSE11839: Urothelial differentiation; Figure. 4.11). Included datasets were considered highly heterogeneous because of different tissue of origin so the random REM was used as the meta-analysis method. Expectedly, the meta-analysis of CHRNA5 probeset expression resulted in enhanced downregulation by differentiation. Comparison of effect-size for genes significantly regulated in epithelial differentiation and logFC of CHRNA5 depletion profile resulted in a significantly positive correlation coefficient ($R= 0.49$, $p=7.9E-142$; Figure 4.11B). There were more genes commonly upregulated/downregulated between the meta-differentiation and CHRNA5 RNAi expression profiles (i.e., UPUP: 687 and DOWNDOWN: 1028) compared to the number of genes showing opposite effect, i.e., significantly downregulated in one dataset but upregulated in other (UPDOWN: 261 and DOWNUP: 316). This analysis allowed demonstration of the similarity between expressions of differentiated cells with CHRNA5 siRNA-1 treated MCF7 cells. The association statistics OR and Fisher's exact tests confirmed the high positive association between both profiles ($OR=8.55$, $p=1.72E-123$; Figure. 4.11C).

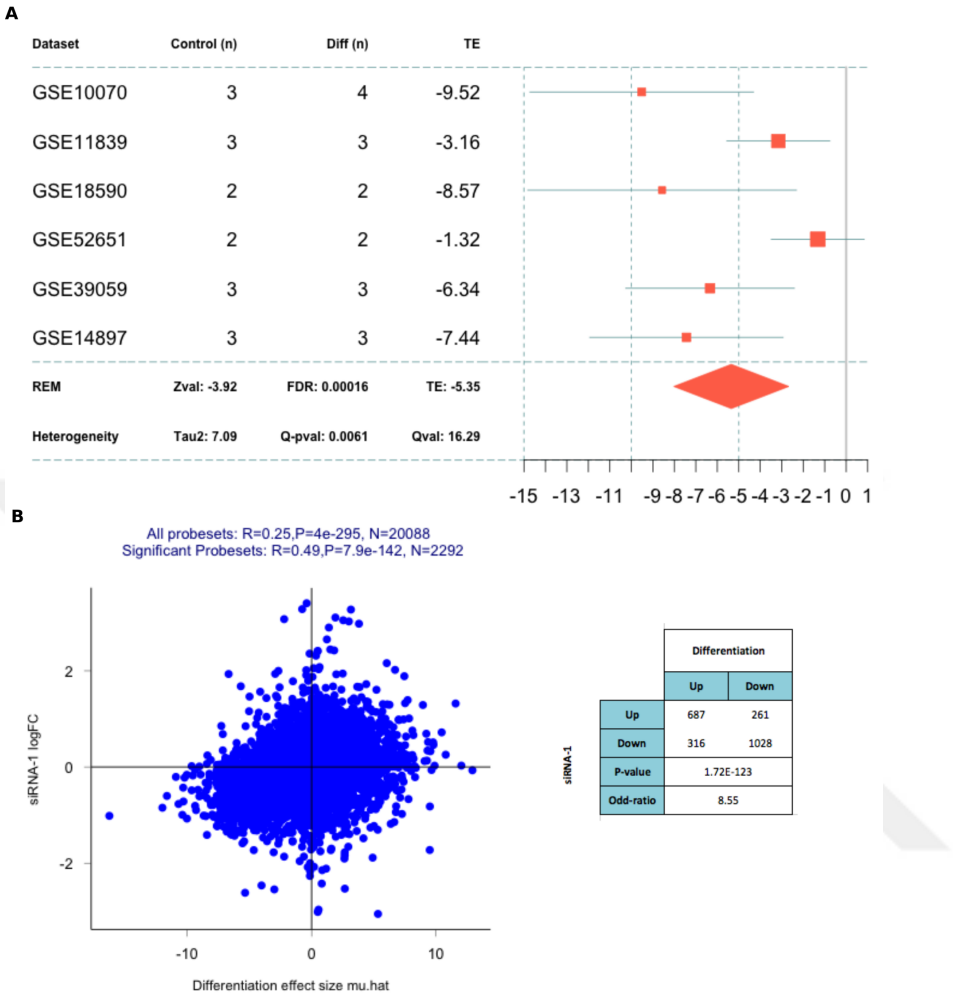


Figure 4.11: Meta-differential expression of epithelial datasets and its comparison with CHRNA5 depletion profile.

A) CHRNA5 meta-analysis in epithelial differentiation datasets. Control (n) and diff (n) indicates number of samples in control and differentiation group. TE indicates treatment effect calculated by standardized mean-difference. Mean effect-size along with 95% confidence interval is shown respectively by squares and lines. REM represents Random effect model; Tau2 indicates Squared-Tau value. B) Correlation between effect-size of epithelial differentiation meta-analysis and CHRNA5 depletion profile. C) Contingency table along with association statistics between epithelial differentiation and CHRNA5 depletion profile.

4.4.2.5. A shared interaction network of epithelial differentiation and CHRNA5 depletion

Next, differentiation-associated CHRNA5 expression signature was obtained by taking the genes highly induced by CHRNA5 depletion in MCF7 that were also significantly

upregulated in meta-differentiation profile (CHRNA5 depletion profile: $p<0.05$ and logFC>0.3; epithelial differentiation meta-analysis: FDR<0.05 and meta-score>0). Interactions of the extracted gene list was obtained using STRING database and the largest network was visualized in Figure 4.12. LogFC of CHRNA5 depletion profile and meta-score of epithelial differentiation meta-analysis results were mapped to show the significance of nodes using Cytoscape visual properties [417]. Connection degree of each node was calculated in Cytoscape and highly connected nodes were obtained from the network. Associated literature studies showing their role in differentiation was described in Table 4.4. The specific clusters of genes in the network were explained in more detail below:

Histone variants and epigenetic regulators

One of the modules in the network was enriched with histone variants (i.e. HIST1H1C, HIST1H2AE, HIST1H2BD, HIST1H2BG, HIST1H2BJ, HIST1H2BK, HIST4H4) and histone modifying machinery (histone demethylases (KDM6B and KDM4B), histone acetyl-transferase (KAT2B) and histone deacetylase (HDAC11)). Expression of Hist1H1C is observed in differentiating photo-receptor cells [418]. Similarly, a recent study has shown that it is terminal differentiation, and not cell cycle arrest, which causes an increase in poly-adenylated histone expression at mRNA level. This signifies the link between CHRNA5 RNAi induced differentiation and histone sub-network [419]. Histone deacetylase (HDAC) and acetyl transferase both are known to aid differentiation program by monitoring epigenetic modifications in embryonic stem cells [420]. KAT2B has been shown as a critical regulator of muscle differentiation [421]. One recent study has shown that EZH2 maintains stem cells state in mammary gland by suppressing differentiation while KDM6B opposes the activity of EZH2 and thus is a positive regulator [422]. KDM6B and KAT2B were among the highly connected nodes in the network (Figure 4.12; Table 4.4).

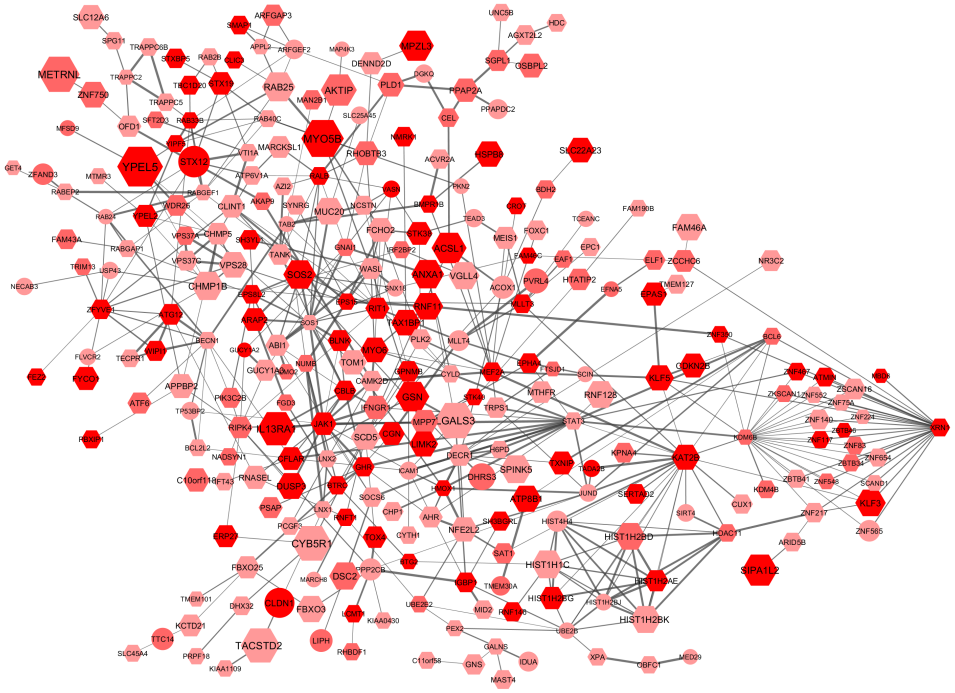


Figure 4.12: Interaction network of shared epithelial differentiation and siRNA-1 induced profile.

Largest connected sub-network of common epithelial differentiation and siRNA-1 induced genes (CHRNA5 siRNA-1: $p<0.05$ and $logFC>0.3$, meta-analysis: $FDR<0.05$ and meta-score>0). Each node represents a gene. Node color indicates logFC of siRNA-1 and three discrete shades of color intensity from pink to red illustrate increased fold change) [423]. Node size represents meta-score (mu.hat) (continuous increase) of epithelial differentiation meta-analysis, while node shape indicates heterogeneity in meta-analysis results (hexagon: qp-value<0.05; elliptical: qp-value>0.05). Edge width represents strength of interaction between two genes (based on STRING score) [424].

Autophagosome processing and trafficking and protein degradation

Another module in the network included the autophagy/autophagosome processing and trafficking genes such as BECN1, ATG12, and ZFYVE1. Autophagy is a requirement for differentiation of myoblast [425] and glioma initiating cells [426]. BECN1 is also known with its tumor suppressive role and has been suggested to play a role in mammary development [427]. ATG12 has been shown to make a conjugate with ATG5. This ATG5-ATG12 conjugate is important for chondrocyte differentiation and viability [428].

Regulation of autophagy in CHRNA5 was further evaluated by collecting and analyzing autophagy machinery genes from the list given by the Qiagen autophagy PCR assay kit (http://www.sabiosciences.com/rt_pcr_product/HTML/PAHS-084A.html#function). Matching of these retrieved genes (n=45) with Affymetrix probesets resulted in 25 probesets and 11 of them were significantly regulated by CHRNA5 RNAi. Interaction network showed small but a well-connected network indicating potential regulation of autophagy (poster presentation in FEBS Advance lecture course: 360^{o}, 23-28th October, 2014).

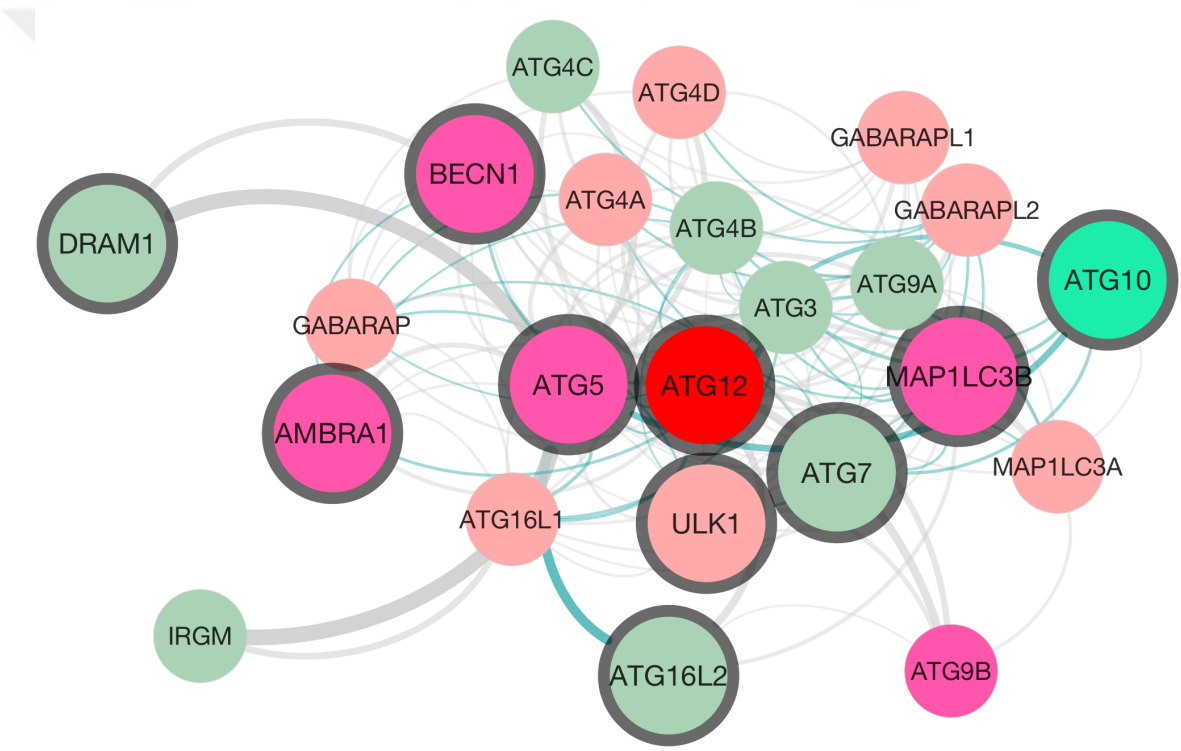


Figure 4.13: Regulation of autophagy machinery genes in CHRNA5 depletion profile.

Upregulated genes are indicated by red color (three different shades representing intensity) and downregulated genes are represented by different shades of green color. Size of the node and grey border represent significant regulation of gene in CHRNA5 depletion profile ($p<0.05$).

Endosomal sorting complex, which is required for transport (ESCRT) machinery also came up in the network (Figure 4. 12). These genes monitor protein turnover in cell by regulating recycling and degradation of ubiquitin tagged proteins [429, 430]. ESCRT machinery is comprised of five protein complexes (ESCRT-0, ESCRT-I, ESCRT-II,

ESCRT-III and AAA-ATPase complex Vps4) [429]; and few proteins belonging to two of these complexes showed up in the network. ESCRT-1 genes (VPS28, 37A and VPS37C) and ESCRT-III genes (CHMP1B and CHMP5) [431] were found in network. Previous studies have stated that knockdown of ESCRT protein disrupts polarity of epithelial cells revealing a tumor suppressive role of ESCRT proteins. ESCRT system regulates localization as well as recycling of claudin-1 (CLDN1), a tight junction protein [432]. *CLDN1* was also found to be significantly upregulated by RT-qPCR in the CHRNA5 RNAi experiment performed in MCF7 (Ermira Jahja, PhD Thesis, 2017). Rabs and SNARE proteins are shown to be involved in membrane trafficking or autophagosome maturation. Several of rab proteins (RAB24, RAB25, RAB2B, RAB33B, RAB40C) and VTI1A (which is a SNARE protein) were also identified in the epithelial differentiation network (Figure 4.12). Accordingly, these findings show a coordinated modulation of ESCRT modules in differentiation induced by several factors including calcium, transwell differentiation, as well as CHRNA5 depletion.

In addition, several other genes present in the epithelial differentiation-CHRNA5 RNAi network are known to be important regulators of differentiation. Muscle enhancer factor 2-A (MEF2A) is involved in myocyte differentiation and calcium dependent gene expression changes. Mice lacking MEF2A shows impaired myoblast differentiation [433]. BCL6, a known oncogene, is induced during keratinocyte terminal differentiation [434]. STAT3, which has been also identified in the meta-analysis as a highly connected hub node of epithelial differentiation, regulates the expression of BCL6 [435]. Another member of BCL family found in the network, BCL2, is a critical factor of epithelial differentiation and morphogenesis [436] as well as a regulator of neural differentiation [437]. BCL2 is also a known to regulator of intracellular calcium levels [436]. 5'-3' exoribonuclease-1 (Xrn1), a part of XRN family enzymes has a role in RNA-decay and during developmental processes. Loss of XRN1 prevents epithelial sheet sealing in *C.elegans* and *D.melanogaster*, while reduced expression of XRN1 is seen in human osteosarcoma cells, not able to differentiate properly. These findings illustrates that XRN1 might be an important factor that responds to developmental changes as well as targets cell shape, movement and proliferation genes. During development, cells might need to switch-off genes which interfere with developmental programs and XRN1 might

be involved in this monitoring [438]. Kruppel-like factors (KLFs) belong to family of C2H2 zinc-finger proteins, which are known to be involved in differentiation [439]. Two of these factors (KLF3 and KLF5) are also included in the epithelial differentiation-CHRNA5 RNAi network; klf5 is especially involved in lung morphogenesis [440] and adipocyte differentiation [441].

Table 4.4: Selected differentiation related genes (Top 10 with known literature) induced by CHRNA5 depletion (Figure 4.12).

Genes	**Degree of connection in network**	**Pubmed Ids (PMIDs) for known association with differentiation**
Histone demethylases (KDM6B)	25	PMID: 25998860, PMID: 27215382
5'-3' exoribonuclease-1 (XRN1)	24	PMID: 18652574, PMID: 23517755
Histone acetyl-transferase (KAT2B)	22	PMID: 23468178, PMID: 23570531, PMID: 24265859, PMID: 20940255, PMID: 9659901, PMID: 15831463
JAK1/STAT3	18	PMID: 24481815, PMID: 19491198, PMID: 10521404, PMID: 12193580, PMID: 8906870, PMID: 23541951, PMID: 11467447
Son of sevenless homolog 1 (SOS1)	12	PMID: 21746917, PMID: 23452850, PMID: 10880441, PMID: 9865697
Beclin-1 (BECN1)	11	PMID: 19261983, PMID: 19261986, PMID: 25483966, PMID: 22273805, PMID: 21861179
Muscle enhancer factor 2-A (MEF2A)	11	PMID: 27040525 PMID: 21868385, PMID: 12379849, PMID: 9271430, PMID: 9162005
Receptor interacting protein kinase 4 (RIPK4)	10	PMID: 25246526, PMID: 27667567, PMID: 23371553, PMID: 24519546
Ras like proto-oncogene B (RALB)	10	PMID: 10022860, PMID: 16648489

4.4.2.6. Validation of differentiation signature using CEMA cell states expression profile

Recently, Germanguz et al. has presented an algorithm CEMA (Core Expression Module Analysis) to define an expression signature which can differentiate between 17 different cell developmental states and to obtain cell specific expression modules [378]. We first tested whether CHRNA5 expression was significantly different between the undifferentiated states versus each of the differentiated states in the CEMA dataset. I showed that CHRNA5 expression was highly significantly reduced in some of the tissue states including pluripotent stem cells (PSC)-derived and tissue fibroblasts, PSC-derived and tissue hepatocytes, human vasculture and tissue keratinocytes (Figure 4.14A). This was in accord with the epithelial meta-analysis results. In addition, CEMA data showed that fibroblasts, hepatocytes and human vasculture had highly reduced CHRNA5 expression when differentiated but neuronal (although similar to epithelial differentiation meta-analysis) as well as PSC endoderm did not exhibit significant CHRNA5 downregulation. I used the differentiation meta-analysis signature after being filtered by the significant genes from CHRNA5 RNAi study (i.e., UPUP and DOWNDOWN categories) and performed hierarchical clustering (Figure 4.14B-C). LogFC of 16 differentiated CEMA states compared with the undifferentiated PSC and CHRNA5 siRNA-1 (logFC) clustering showed that CHRNA5 RNAi signature were grouped together with those where CHRNA5 expressed was significantly lowered. Upregulated signature showed that CHRNA5 siRNA-1 clustered with PSC derived hepatocytes (highly downregulated CHRNA5 expression) (Figure 4.14B) and downregulated signature was successfully able to separate the neuronal differentiation signature from that of the endothelial (fibroblasts and vasculature), epithelial (hepatocytes and keratinocytes) and myoepithelial clusters. Tumor tissue along with PSC-derived endoderm came as an outgroup in downregulated signature (Figure 4.14C).

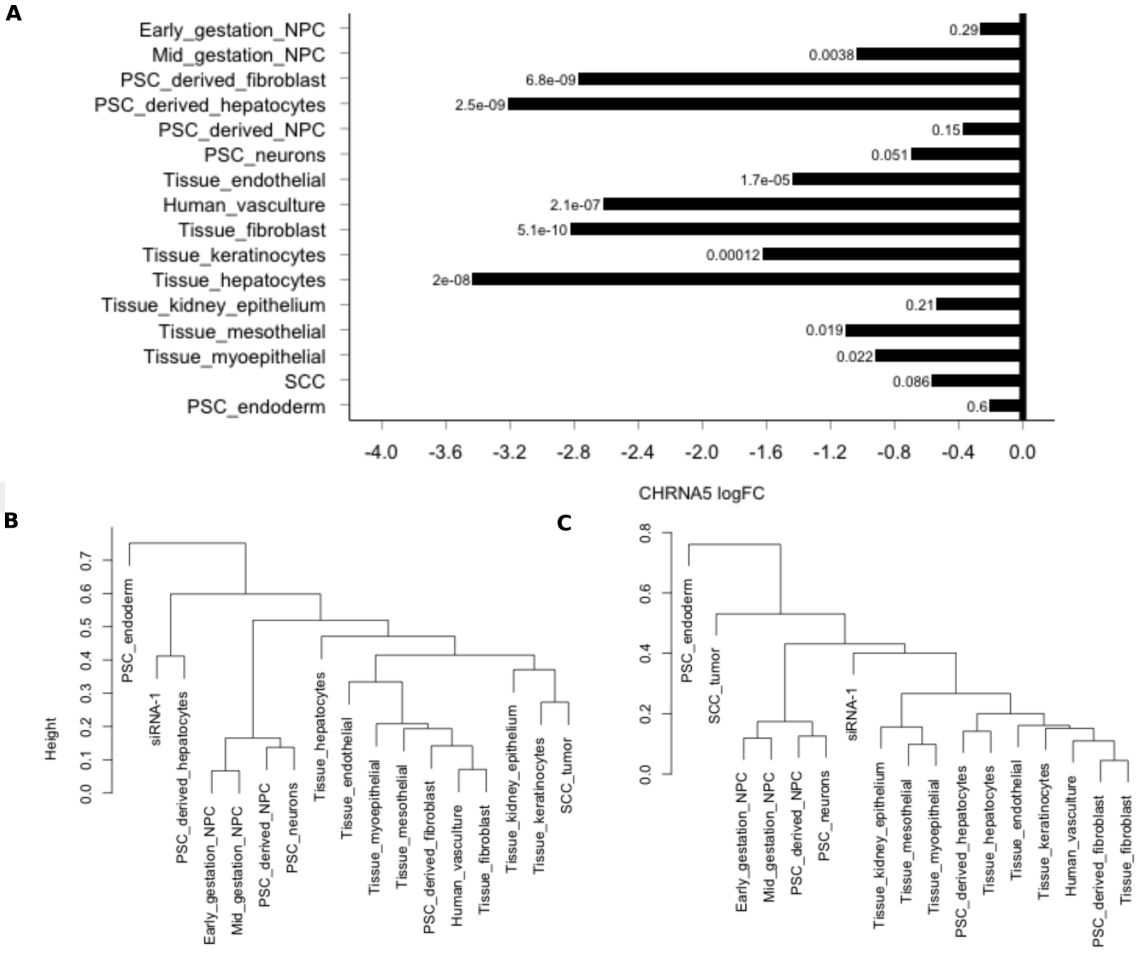


Figure 4.14: Evaluation of differentiation signature in CEMA data.

A) CHRNA5 expression in CEMA data comparison. Bar length indicates log fold change while p-value obtained from limma are described in front of each bar. Clustering of CEMA data and siRNA-1 profile based on shared siRNA-1 and epithelial differentiation meta-analysis. B) induced, and C) repressed genes.

4.4.2.7. Evaluation of luminal and myoepithelial markers in CHRNA5 depleted MCF7 cells

Bipotent/multipotent stem cells are the precursors, giving rise to lineage specific precursors and ultimately differentiated epithelial cells [297]. Two types of epithelial cells, i.e., luminal and myoepithelial, are present in the mammary gland [442, 443]. In the above section, the association between the EMT and CHRNA5 RNAi signature implicated that CHRNA5 depletion, although increasing VIM mildly, did not result in EMT. Differentiation meta-analysis on the other hand suggested a strong association

with epithelial differentiation profiles. Accordingly, the microarray dataset analysis of CHRNA5 RNAi application identified ACTA2, a known myoepithelial marker, as a significantly upregulated gene in CHRNA5 depleted MCF7 cells (logFC=3.41, p=3.48E-08; Figure 4.15A). On the other hand, although MCF7 is a known luminal cell line, a mild downregulation of a well-reputed luminal marker, MUC1 was still observed in CHRNA5 siRNA-1 treated cells (logFC=-0.45, p=0.0015; Figure 4.15A). RT-qPCR was used to validate these findings. ACTA2, in accord with the microarray findings, was significantly upregulated by treatment with siRNA-1 for both 72h (logFC=2.5, p=0.002; Figure 4.15A) and 120h (logFC=2.77, p=0.009; Figure 4.15B). This phenomenon was also observed in siRNA-3 (logFC=1.27, p= 0.048; Figure 4.15B). In case of MUC1, downregulation at 72h siRNA treatment was on the borderline of significance (logFC=-0.84, p=0.056; Figure 4.15A). Although a downregulation pattern was also seen in 120h experiment (logFC=-1.2, p=0.13; Figure 4.15B), it could not reach significance perhaps due to variability between biological replicates and less number of samples (i.e., 2 siRNA samples compared to one scrambled). Validation by siRNA-3 did not show any significant change either. Another independent 120h siRNA experiment however confirmed increased expression of *ACTA2* (logFC=3.8, p=0.007; Figure 4.15C) and downregulation of *MUC*1 expression (logFC=-0.37, p=0.035; Figure 4.15C) upon CHRNA5 depletion. ACTA2 and VIM are interesting genes because both of them are markers of EMT as well as myoepithelial differentiation but as CDH1 expression did not change significantly, no EMT was assumed and the mild upregulation observed in VIM expression can be associated with a probable initiation of myoepithelial differentiation [444]. SNAI2, another known myoepithelial marker, was found to be upregulated in the microarray study (logFC=0.48, p=8.37E-04; Figure 4.15A). RT-qPCR did not result in a significant change at 72h (logFC=-0.3, p=0.59; Figure 4.15A). While120h siRNA treatment showed an increasing trend with a p-value approaching significance (logFC=0.99, p=0.08; Figure 4.15B), this change could not be confirmed in the independent 120h siRNA treatment (Figure 4.15C). Evaluation of multiple luminal/myoepithelial markers based on microarray results showed that there was a mixed trend of luminal and myoepithelial markers upon CHRNA5 depletion in the previous section. Several myoepithelial markers were upregulated but downregulation of

well-known luminal markers was not remarkable (Table 4.5). With upregulation of multiple myoepithelial markers and significant positive association with differentiation, these results highlighted the role of CHRNA5 in switching MCF7 towards a more myoepithelial type but this differentiation might be in its early stages or incomplete.

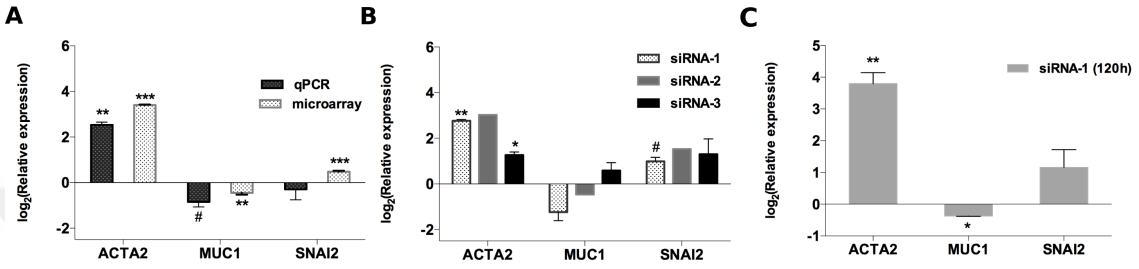


Figure 4.15: Evaluation of luminal and myoepithelial marker genes in CHRNA5 depleted MCF7 cells.

Breast lineage specific marker genes evaluation in A) microarray probesets and RT-qPCR (72h siRNA-1 transfection) (microarray calculations: subtraction of average siRNA-CN (controls) logFC from each of siRNA-1 treated samples). B) siRNA-3 as well as siRNA-1 treatment (120h exposure) **C)** an independent siRNA-1 (120h) transfection. SDHA was used as a reference gene in RT-qPCR calculations. *p<0.05, **p<0.01, ***p<0.001, # p<0.1. Statistical tests were performed using One-Way ANOVA or single sample t.test.

Table 4.5: Regulation of luminal and myoepithelial genes in MCF7 CHRNA5 RNAi microarray analysis (*limma* results).

	Gene	Probeset	logFC	Pvalue	Pubmed ID
Luminal Markers	KRT18	201596_X_AT	-0.16	7.58E-02	PMCID: PMC3325967, PMID: 26861772, PMID: 25482617, PMID: 27558284, PMID: 18469806, PMID: 15886886, PMCID: PMC4594335
	MUC1	213693_S_AT	-0.45	1.51E-03	PMID: 26861772, PMID: 15886886, PMCID: PMC4594335
	KRT8	209008_X_AT	-0.18	5.62E-02	PMCID: PMC3325967, PMID: 26861772, PMID: 25482617, PMID: 27558284, PMCID: PMC4594335
	KRT19	201650_AT	0.19	4.31E-02	PMID: 15886886, PMID: 27558284, PMID: 25482617, PMCID: PMC4594335
	GATA3	209602_S_AT	0.22	8.83E-02	PMID: 25482617, PMCID: PMC4594335
	ITGA6/CD49f	201656_AT	0.50	1.69E-03	PMID: 26861772, PMCID: PMC4594335, PMID: 15886886, PMID: 25482617
Myoepithelial markers	VIM	201426_S_AT	0.25	1.88E-02	PMID: 26861772, PMID: 25482617, PMID: 15886886, PMCID: PMC4594335
	KRT14	209351_AT	0.16	4.09E-01	PMID: 15886886, PMID: 25482617, PMCID: PMC4594335, PMID: 18469806
	ACTA2	200974_AT	3.41	3.48E-08	PMCID: PMC3325967, PMID: 25482617, PMID: 15886886, PMCID: PMC4594335
	SNAI2	213139_AT	0.48	8.37E-04	PMCID: PMC4594335
	ITGB1	211945_S_AT	0.89	1.73E-05	PMID: 18469806, PMCID: PMC4594335, PMID: 25482617

4.4.3. Evaluation of TP53 signaling in CHRNA5 RNAi mediated effect

CHRNA5 depletion experiments via siRNA application were performed in three different cell lines (cell culture, RNA isolation, and cDNA extraction were performed by Ermira Jahja, PhD thesis, 2017). MCF7 was primarily used to understand CHRNA5 associated transcriptional (microarray, qPCR) or functional (multiple functional assays related to cell viability and cellular morphology) effects. In these studies, the effects of CHRNA5 RNAi was found to be TP53 dependent, i.e., induced only in the presence of a functional TP53 protein, by using qPCR in TP53 positive and negative cell lines (Ermira Jahja, PhD thesis, 2017).

I used comparative transcriptomics by comparing CHRNA5 siRNA-1 expression profile with a public dataset in which MCF7 cells were treated with Nutlin-3a. Induction of TP53 targets was quantified using RNAseq; and ChipSeq was used to filter out the genes bound by TP53 [445]. In these datasets, 212 TP53 target genes were predicted and validated by using both RNAseq and ChIPseq, respectively. After mapping to Affymetrix HGU-133 Plus 2 arrays using jetset best probesets, 187 probesets were obtained. I retrieved CHRNA5 siRNA-1 *limma* results to compare with RNAseq logFC provided by Janky et al. [445]. The comparative transcriptomics identified that 106 out of 187 TP53 target probesets were significantly modulated also in CHRNA5 siRNA-1 treated MCF7 cells ($p<0.05$). Of these 106 targets, most (90 probesets) were indeed upregulated revealing a significant induction of TP53 signaling by targeting CHRNA5 using RNAi and TP53 signaling might be an important factor in transducing the CHRNA5 mediated effect. Considering 106 significantly regulated probesets, a significant positive correlation was identified between logFC of siRNA-1 and RNAseq of nutlin-3 treatment ($R=0.47$, $p=3.9E-07$) (Figure 4.16).

To evaluate CHRNA5 associated changes in other cell lines CHRNA5 was downregulated in MDA-MB-231 and BT-20 (siRNA-1 transfection for 72h; Ermira Jahja, PhD thesis, 2017). I then tested expression changes of two EMT and three differentiation markers (*VIM, CDH1, ACTA2, MUC1* and *SNAI2* via RT-qPCR (Figure 4.17A). Univariate analysis revealed that only *SNAI2* changed significantly in BT20.

Baseline expressions of analyzed genes were also variable among the cell lines (Figure 4.17B) and Two-way ANOVA results indicated a strong impact of cell line (Table 4.6). As both of the cell lines tested were TP53 mutant, I hypothesized that CHRNA5 depletion induced changes were carried by TP53 signaling. TP53 also came as one of the significant pathway upon functional annotation of CHRNA5 siRNA-1 significant genes (Ermira Jahja, PhD Thesis, 2017).

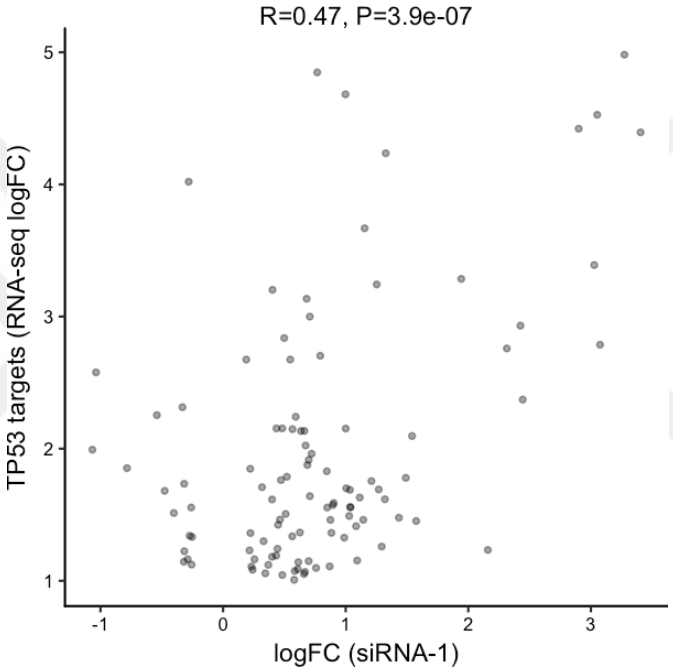


Figure 4.16: TP53 genes regulation in CHRNA5 depleted cells.

Correlation between RNAseq logFC of TP53 targets (confirmed by chip-seq, i.e., chipseq peak=yes) (retrieved from Table S4 of [91]) and siRNA-1 logFC (only significantly moduled genes).

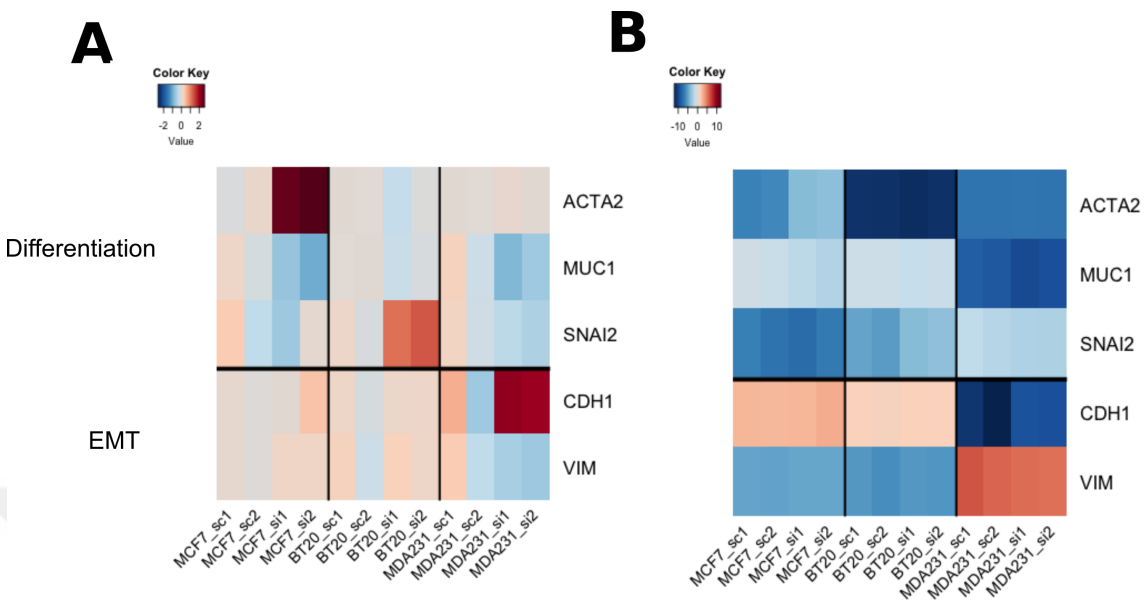


Figure 4.17: EMT and differentiation biomarkers eexpression in three cell lines.

Heatmap of EMT and differentiation genes in MCF7, MDA-MB-231 and BT-20 after CHRNA5 depletion analyzed via qPCR (A) ΔΔCT. B)-ΔCT.

Table 4.6: Two-way ANOVA results of RT-qPCR evaluated genes using cell line and siRNA-1 treatment (Figure 4.16).

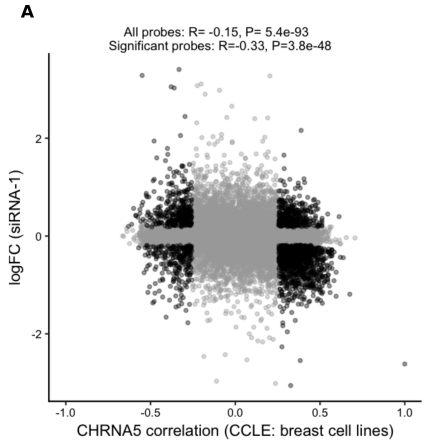
Category	Genes	Cell line	siRNA-1 treatment	Interaction
Differentiation	**ACTA2**	0.31	0.62	0.53
	MUC1	**2.03E-03**	0.66	0.97
	SNAI2	**4.16E-06**	0.56	0.83
EMT	**CDH1**	**3.09E-04**	0.64	0.66
	VIM	**2.68E-03**	0.97	0.88

4.4.4. Comparison of CHRNA5 co-expression profile from CCLE breast cell lines with CHRNA5 depletion profile

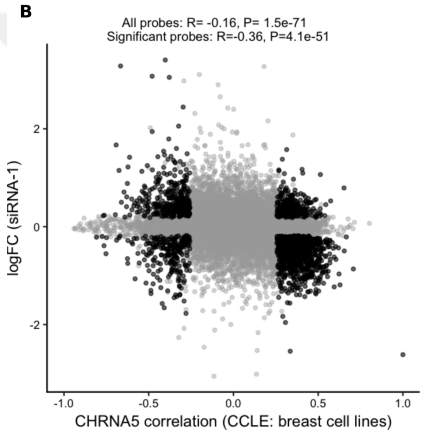
CCLE (Cancer Cell Line Encyclopedia) encompasses gene expression data of 1036 cancer cell lines belonging to different tissues. I extracted breast data (59 cell lines) and identified the correlation coefficient between expression profile of each gene with that of

CHRNA5. The relationship between these correlation values across the CCLE breast cancer cell lines with the logFC of CHRNA5 RNAi expression profile exhibited a great potential to validate the effects of CHRNA5 siRNA-1 on MCF7 for generalization to multiple cell lines belonging to different cancer subtypes. Expectedly, a significant negative correlation between the CCLE correlation profile and CHRNA5 siRNA-1 expression profile was found (R=-0.15; p=5.4E-93; Figure 4.18A). By using only significant probesets (CCLE correlation: $p<0.05$ and siRNA-1: $p<0.05$), this correlation value increased to (R=-0.33, p=3.8E-48; Figure 4.18A). Considering these shared significant probesets, Fisher's exact test and Odds ratio also showed a significant negative association between both profiles (OR=0.27; p=1.94E-38; Figure 4.18A).

In addition, we also used a filtering criterion to exclude cell lines with low expression of the gene of interest (other than CHRNA5) before finding this correlation. A normalized expression value less than 5 was set as a cut-off and the correlation analysis was performed between each gene and CHRNA5 as mentioned above but this time only for the number of cell lines which passed the cut-off. The filtering of non-expressing cell lines slightly increased the negative correlation (R=-0.16, p=1.5E-71; Figure 4.18B). The correlation value increased further by considering only significant probesets (R=-0.36, p=4.1E-51; Figure 4.18B). Similarly, Fisher's exact test p-value decreased further with a strongly negative association represented by OR equaling to 0.22 (p=1.71E-44) (Figure 4.18B). Our results validated that transcriptomic response associated with CHRNA5 depletion were not limited only to MCF7 and could be extended to other cell lines with public expression profiles.



		CCLE breast cell lines	
		Up	Down
siRNA-1	Up	353	351
	Down	942	249
pvalue		1.94E-38	
Odd-ratio estimate		0.27	



		CCLE breast cell lines	
		Up	Down
siRNA-1	Up	252	347
	Down	793	240
P-value		1.71E-44	
Odd-ratio estimate		0.22	

Figure 4.18: CCLE breast cell lines association with CRHNA5 depletion profile.

Comparison of CHRNA5 correlation profile in A) unfiltered and B) filtered CCLE breast cancer cell lines and siRNA-1 logFC. Significantly modulated genes in both categories are indicated by black color and non-significant (or significant in only one category) are shown in grey. Association statistics result, i.e., contingency table and Fisher's exact test result of shared significant genes are shown below scatter plot.

4.4.4.1. Functional analysis of CCLE vs CHRNA5 RNA expression signature

Online tool DAVID was used to perform KEGG analysis on the shared CCLE and CHRNA5 depletion profile separately for both unfiltered and filtered CCLE co-expression profiles. Functional annotation of CHRNA5 co-expressed genes (i.e., downregulated in CHRNA5 RNAi and positively correlated with CHRNA5 in CCLE) irrespective of unfiltered/filtered CCLE profile resulted in identification of cellular growth-related processes. Proliferative pathways enriched in the category included cell cycle, DNA replication and repair (Table 4.7). Findings revealed that CHRNA5 expression not only in MCF7 but in a wide-range of breast cell lines might be associated

with aggressiveness; and targeting of CHRNA5 in breast cancer can help control cell growth. On the other hand, CHRNA5 negatively co-expressed genes (upregulated in CHRNA5 depletion profile and negatively correlated in CCLE profile) resulted in cancer related pathways, wnt-signaling pathway and focal adhesion (Table 4.7). Deregulated genes, i.e., upregulated by siRNA-1 along with CCLE positively correlated profile or downregulated by siRNA-1 along with negatively correlated in CCLE profile, included tight junction and endocytosis revealing that CHRNA5 mediated effects on these pathways might be specific to MCF7 and similar cell lines, which might differ across other breast cell line types (Table 4.7).

Table 4.7: Functional annotation of shared significant CHRNA5 siRNA-1 and CCLE expression profile (using KEGG).

Analysis was performed for unfiltered/filtered (normalized expression value >5) CCLE correlation analysis ($p<0.05$). NA represents term was not found significant.

CCLE correlations based on unfiltered data / filtered data (expression>5)			
Term	**Count**	**PValue**	**Benjamini**
Upregulated by siRNA-1 and positively correlated with CHRNA5 in CCLE breast data			
hsa04530: Tight junction	11/6	1.59E-04 / 3.09E-02	1.84E-02 / 7.85E-01
hsa04012: ErbB signaling pathway	6 / NA	2.18E-02 / NA	7.24E-01/ NA
hsa04514: Cell adhesion molecules (CAMs)	7 / NA	3.46E-02 / NA	7.47 E-01 / NA
hsa04660: T cell receptor signaling pathway	6 / NA	4.87E-02/ NA	7.68E-01 / NA
hsa03010: Ribosome	NA / 5	NA / 2.71E-02	0 / 9.32E-01
Downregulated by siRNA-1 and negative correlated with CHRNA5 in CCLE breast data			
hsa04144: Endocytosis	7 / 7	2.73E-02 / 2.37E-02	8.94E-01 / 8.27E-01
hsa04810: Regulation of actin cytoskeleton	NA / 7	NA / 4.57E-02	NA / 8.19E-01
Upregulated by siRNA-1 and negatively correlated with CHRNA5 in CCLE breast data			
hsa05219: Bladder cancer	5 / 5	4.61E-03 / 5.02E-03	4.09E-01 / 2.34E-01

CCLE correlations based on unfiltered data / filtered data (expression>5)			
Term	**Count**	**PValue**	**Benjamini**
hsa00520: Amino sugar and nucleotide sugar metabolism	5 / 4	5.45E-03 / 3.65E-02	2.68E-01 / 3.89E-01
hsa04310: Wnt signaling pathway	8 / 9	1.11E-02 / 3.49E-03	3.47E-01 / 3.10E-01
hsa04510: Focal adhesion	9 / 9	1.61E-02 / 1.84E-02	3.70E-01 / 3.88E-01
hsa00500: Starch and sucrose metabolism	4 / NA	3.05E-02 / NA	5.06E-01 / NA
hsa04142: Lysosome	6 / NA	4.15E-02 / NA	5.53E-01 / NA
hsa04350: TGF-beta signaling pathway	NA / 6	NA / 1.47E-02	NA / 4.07E-01
hsa05200: Pathways in cancer	NA / 12	NA / 1.97E-02	NA / 3.45E-01
hsa04115: p53 signaling pathway	NA / 5	NA / 2.63E-02	NA / 3.76E-01
hsa04144: Endocytosis	NA / 8	NA / 3.33E-02	NA / 4.01E-01
hsa00520: Amino sugar and nucleotide sugar metabolism	NA / 4	NA / 3.65E-02	NA / 3.89E-01
Downregulated by siRNA-1 and positively correlated with CHRNA5 in CCLE breast data			
hsa03030: DNA replication	24 / 24	1.33E-22 / 2.48E-24	2.0E-20 / 3.25E-22
hsa04110: Cell cycle	34 / 31	6.04E-17 / 2.63E-16	8.44E-15 / 1.45E-14
hsa03430: Mismatch repair	13 / 13	7.56E-11 / 1.03E-11	3.81E-09 / 3.37E-10
hsa03040: Spliceosome	27 / 27	7.59E-11 / 1.52E-12	2.87E-09 / 6.65E-11
hsa03410: Base excision repair	14 / 14	2.32E-09 / 2.86E-10	7.01E-08 / 7.49E-09
hsa00240: Pyrimidine metabolism	21 / 20	9.18E-09 / 3.27E-09	2.31E-07 / 7.14E-08
hsa03420: Nucleotide excision repair	13 / 13	4.87E-07 / 7.68E-08	1.05E-05 / 1.44E-06

CCLE correlations based on unfiltered data / filtered data (expression>5)			
Term	**Count**	**PValue**	**Benjamini**
hsa03440: Homologous recombination	10 / 10	3.04E-06 / 7.39E-07	5.73E-05 / 1.21E-05
hsa00230: Purine metabolism	22 / 19	7.35E-06 / 2.89E-05	1.23E-04 / 3.79E-04
hsa03018: RNA degradation	12 / 12	5.31E-05 / 1.10E-05	8.01E-04 / 1.60E-04
hsa04114: Oocyte meiosis	16 / 15	1.67E-04/9.79E-05	2.29E-03/ 1.16E-03
hsa04914: Progesterone-mediated oocyte maturation	12 / 10	2.11E-03/ 6.69E-03	2.62E-02/ 7.06E-02
hsa04115: p53 signaling pathway	9 / 8	1.36E-02/ 1.75E-02	1.47E-01 / 1.63E-01
hsa00900: Terpenoid backbone biosynthesis	4 / 4	3.07E-02 / 1.97E-02	2.85E-01 / 1.70E-01
hsa00670: One carbon pool by folate	4 / 4	3.65E-02 / 2.36E-02	3.12E-01 / 1.88E-01

4.4.5. CHRNA5 and its neighboring genes exhibit high correlation with each other

CHRNA5 itself is located on the q-arm of chromosome 15 i.e. 15q25.1 along with multiple other genes (Figure 4.19).

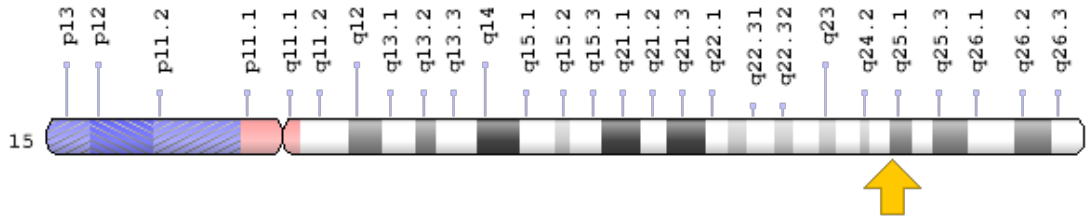


Figure 4.19: Graphical visualization of chromosome 15 highlighting CHRNA5 location.

Figure is retrieved from Genes database: https://ghr.nlm.nih.gov/gene.

Chromosomal enrichment was performed to evaluate the existence of clusters, whose expression might be regulated similarly. Highly co-expressed genes were selected by

two criteria: 1) extracting positively correlated genes with CHRNA5 in CCLE cell lines (correlation analysis: $R>0.3$ and $p<0.05$); and 2) extracting genes that were also downregulated upon siRNA-1 treatment in MCF7 (*limma* analysis: $logFC<-0.5$ and $p<0.05$). Chromosome enrichment using Bioconductor package *category* [108] showed that highly CHRNA5 co-expressed genes were also located on chromosome 15 and in particular, different regions of q-arm of chromosome 15 (Table 4.8). Genes present on chromosome 15 were retrieved from DAVID bioinformatics tool and are listed in Table 4.9. Interestingly many of these genes belong to centromeric region; thus modulation of CHRNA5 along with kinetochore related activities make CHRNA5 an important indicator of cell division.

Table 4.8: Chromosome enrichment (Top 5) of CHRNA5 modulated genes (depleted by the siRNA-1 in MCF7 and positively correlated with CHRNA5 in CCLE).

ChrMapID	PValue	OddsRatio	ExpCount	Count	Size
15	6.06E-04	2.35	9.49	21	474
15q22.31	2.21E-03	8.24	0.56	4	28
15q26.1	2.53E-03	7.91	0.58	4	29
1p34.1	5.07E-03	6.38	0.7	4	35
15q22	7.45E-03	4.42	1.22	5	61

Table 4.9: CRHNA5 co-expressed genes enriched on chromosome 15 (DAVID version 6.8).

ID	Gene Name
203755_AT	BUB1 mitotic checkpoint serine/threonine kinase B (BUB1B)
205733_AT	Bloom syndrome RecQ like helicase (BLM)
213007_AT	Fanconi anemia complementation group I (FANCI)
1558750_A_AT	OTU deubiquitinase 7A pseudogene (LOC100288637)
213599_AT	Opa interacting protein 5 (OIP5)
228252_AT	PIF1 5'-to-3' DNA helicase (PIF1)
219258_AT	TIMELESS interacting protein (TIPIN)
230021_AT	TOPBP1 interacting checkpoint and replication regulator (TICRR)
239680_AT	WD repeat domain 76 (WDR76)
239413_AT	centrosomal protein 152 (CEP152)
206533_AT	Cholinergic receptor nicotinic alpha 5 subunit (CHRNA5)
202705_AT	cyclin B2 (CCNB2)
208955_AT	deoxyuridine triphosphatase (DUT)
224779_S_AT	family with sequence similarity 96 member A(FAM96A)
202070_S_AT	isocitrate dehydrogenase 3 (NAD(+)) alpha (IDH3A)
204709_S_AT	kinesin family member 23 (KIF23)
225300_AT	kinetochore localized astrin/SPAG5 binding protein (KNSTRN)
218039_AT	nucleolar and spindle associated protein 1 (NUSAP1)
218009_S_AT	protein regulator of cytokinesis 1 (PRC1)
205768_S_AT	solute carrier family 27 member 2 (SLC27A2)
222606_AT	zwilch kinetochore protein (ZWILCH)

4.4.6. CHRNA5 and E2-mediated signaling

Functional analysis of CHRNA5 depletion revealed its significance in inducing G1 arrest (Ermira Jahja, PhD thesis, 2017) as well as differentiation in breast cancer cells (as shown above). These findings highlight the proliferative potential of CHRNA5 in breast cancer.

Estrogen/17b-estradiol (E2) is the main steroid hormone; apart from regulating mammary development it is also the main contributor to tumorigenicity in breast cancer [446]. E2 helps cancer cell proliferation by facilitating G1/S transition during cell cycle and in addition, by blocking apoptosis, while on the other hand, hormone starvation in cancer cells can revert these proliferative effects and drives breast cancer cells toward apoptosis [447]. Understanding the effect of E2 signaling on CHRNA5 expression or vice-versa can provide more insight about its regulation in breast cancer.

Previously Rumeyse Biyik in her Master's thesis (2009) has shown by using bioinformatics analysis that CHRNA5 expression was responsive to a) E2 in multiple E2-treated breast cancer cell line datasets; b) the expression of CHRNA5 was increased in ER- breast cancer patients as compared to ER+ patients; and c) E2 was a likely secondary target of E2-mediated signaling. She has used regression, correlation analysis and class comparison analyses on public datasets (*in-vitro* and *in-vivo*) to present these findings. However a proper meta-analysis using more comprehensive cohorts was missing. Following analyses were conducted to understand the impact of E2-mediated regulation on CHRNA5 transcript levels.

4.4.7. CHRNA5 depletion profile in comparison with E2 treatment and starvation profiles

Previously Rumeysa Biyik has reanalyzed 1) the expression dataset GSE4668 performed by Coser et al. using MCF7/BUS cells treated with different doses (0, 10, 30, 60, 100pM) of E2 for 48 hours and 2) a hormone starvation experiment (GSE4668) performed for 48 hours (2 days) in a time-series manner by collecting samples at 0, 1

and 2 days [448]. Rumeysa Biyik has shown that CHRNA5 expression increased in response to E2 while it decreased in E2 starvation, expectedly (Figure 4.20).

This reanalysis study is of significance as both the E2 treatment and starvation experiments were conducted in MCF7/BUS stock and processed gene expression data was publicly available. One recent study has evaluated the effect of E2 treatment/starvation on four routinely used different MCF7 stocks and has found that MCF7/BUS has been the best performing stock with most proliferative capability in the presence of E2 and strong opposite effect in hormone starved cells [449]. Therefore, GSE4668 provided an excellent model to study effect of E2 treatment and starvation on CHRNA5 mediated transcriptomic changes. Both sets of experiments were evaluated for their gene expression studies using Affymetrix Human Genome U133A Array (HGU133A). Processed normalized data (MAS 5.0) was submitted to GEO database [226] with accession number GSE4668 [448]. For this, I have performed the *limma* analysis of the normalized expression values obtained from the E2 treatment/starvation datasets and compared them with that of the CHRNA5 RNAi expression profile (Figure 4.21). Selection of the time point for comparison was determined as below: Highest concentration of E2 treatment (100pM) and the maximum time point of hormone starvation (2 days) were most effective on CHRNA5 expression and were chosen to compare with their respective controls (0pM or 0 day) using *limma* analysis. Sequential comparison with CHRNA5 depletion profile showed that starvation expression profile was highly positively correlated with CHRNA5 depletion profile (significant genes: R=0.53, p= 3.2E-94; Figure 4.21A). Number of genes modulated in the same direction (i.e., upregulated in both profiles=468, downregulated in both profiles=523, Total=991) were much higher than the number of genes regulated in opposite direction (upregulated in starvation and downregulated in siRNA-1=222, downregulated in starvation and upregulated in siRNA=50; Total=272). Fisher's exact test validated the high positive association between the log fold changes of these significantly regulated genes (OR=21.98, p=7.03E-111; Figure 4.21C).

In accord with upregulation of CHRNA5 probesets in the E2 treatment profile, a significant negative correlation was found between CHRNA5 depletion and E2

treatment (Figure 4.21B). Contrary to starvation comparative analysis, significant genes regulated in the opposite direction (upregulated in siRNA-1 and downregulated in E2 treatment=357, downregulated in siRNA-1 and upregulated in E2 treatment=518; Total=875) were much higher than the ones regulated in the same direction (upregulated in both profiles=108, downregulated in both profiles=178, Total=286) resulting in significant negative association (OR=0.10, p=2.18E-68; Figure 4.21C). These findings showed that genes upregulated in the presence of E2 were largely downregulated in both CRHNA5 depletion profile and starvation and vice-versa (Figure 4.21D-E). KEGG annotation of these common genes (downregulated in siRNA-1 and starvation profile but upregulated in E2 treatment) showed proliferation related mechanisms while in the opposite category (upregulated in siRNA-1 and starvation profile but downregulated by E2 treatment), only viral carcinogenesis was found to be significant (BH<0.05). Our findings indicated that CHRNA5 could be induced by E2 and may facilitate E2-induced proliferation processes in breast tumors while targeting of CHRNA5 can induce anti-proliferative actions mimicing E2 starvation.

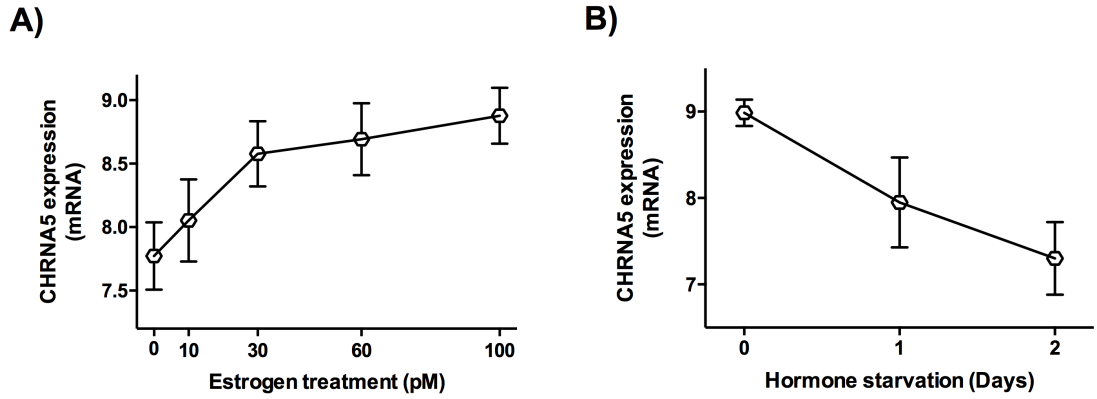


Figure 4.20: CHRNA5 probeset (206533_AT) in E2 treatment and starvation profile (GSE4668).

Circles represents mean of replicate samples along with standard deviation indicated by error bars A) in increasing dose of E2 or B) hormone starvation profile (GSE4668). The figure is re-drawn from Rumeysa Biyik MS thesis (2009) based on GSE4668 dataset.

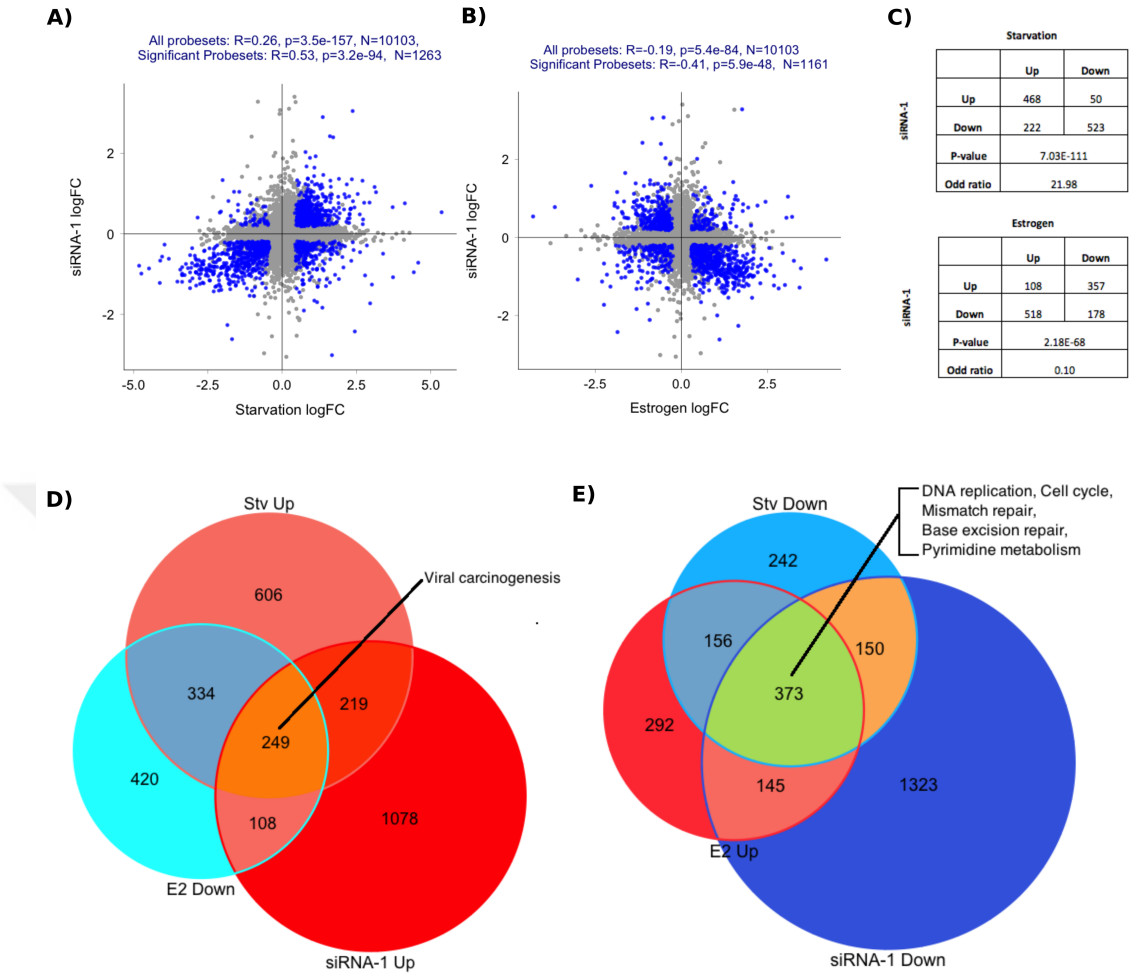


Figure 4.21*:* Comparison of CHRNA5 depletion datasets with E2 treatment and hormone starvation profile.

Scatter plot between logFC of CHRNA5 depletion profile and A) starvation (2 day; GSE4668), B) E2 treatment profile (48 hours; GSE4668). Each dot represents one probeset. Grey dots represent non-significant change in gene expression while blue dots indicate significant regulation of gene expression according to both *limma* analyses (x or y-axis; $p<0.05$). Correlation statistics indicated in plot title is based on all probesets (including grey and blue) as well as only significant genes (blue dots). C) Contingency table along with association statistics of significantly regulated genes ($p<0.05$; blue dots in A and B) using Fisher's exact p-value and OR. Up and down genes were identified on basis of logFC direction. Overlap of significantly regulated genes ($p<0.05$) which are D) upregulated in siRNA-1 and starvation but downregulated in E2 treatment E) downregulated in siRNA-1 and starvation while upregulated in E2 treatment. Stv=starvation, siRNA-1=CHRNA5 depletion profile, R=Correlation coefficient. Significant KEGG pathways (BH<0.05) for genes common in all three categories are listed. In case of multiple pathways, only top 5 are written.

4.4.8. Effects of time-dependent E2 treatment on CHRNA5 expression

CHRNA5 expression induction in E2-treated MCF7 cells was observed by analysis of GSE4668 but it was important to identify whether the observed upregulation was limited to one study/one particular MCF7 stock or was widespread. GEO expression database was searched for E2-treated MCF7 cells based on GPL570 or GPL96. As of 2016, 21 E2 treated datasets were collected, filtered and processed as mentioned in Materials and Methods section above; and ultimately 11 datasets passed the selection criteria (Figure 4.3; PRISMA guidelines). Acquired *E2 cohort* consisted of different treatment time-points ranging from 2-48 hours. Few of these datasets were also previously analyzed individually to show CHRNA5 expression was E2 responsive (Rumeyse Biyik; MS thesis, 2009), however statistical meta-analysis of a larger cohort can provide a more comprehensive interpretation.

In case of one GEO dataset consisting of several treatment time points with respective control samples, each time-point with its control was considered as a separate dataset. Splitting of 11 gene expression studies resulted in 14 datasets. Based on E2 exposure time points, datasets were further split in two categories as described previously: *early E2 treatment cohort* (2-4 hours E2 treated samples) and *late E2 treatment cohort* (6-48 hours of E2 treated samples). Categorization of E2 treated datasets was described in Table 4.10. Datasets having 16 control and 17 E2 treated samples constituted early E2 treatment group (Table 4.10). Evaluation of CHRNA5 probeset expression revealed that heterogeneity among the datasets was less than 50% (I^2 = 45.2%). FEM statistics were more suitable for less heterogeneous studies. Findings revealed highly variable expression with no conclusive effect-size (z-value=0.45, SMD=0.21, p=0.65; Figure 4.22A). *Late E2 treatment cohort* was comprised of 8 datasets (29 controls and 29 E2-treated samples) (Table 4.10). CHRNA5 probeset expression in this cohort was highly heterogeneous (I^2=65.9%) making REM statistics more applicable. Meta-analysis results revealed that CHRNA5 expression was significantly induced upon longer E2 treatment (REM z-value=3.28, SMD=3.62, p=0.0011; Figure 4.22B). These findings confirmed CHRNA5 regulation by E2 treatment and also indicated that CHRNA5 might not be a

direct E2 target since its transcription might be indirectly regulated by E2 in a time-dependent manner.

Table 4.10: Datasets used in meta-analysis of *E2 cohor*t (MCF7 cell line data).

'Meta-analysis group' was decided on the basis of E2 treatment duration, i.e., *early treatment* (2-4 hours) and *late treatment* (6-48 hours). GSE8597 was used for evaluation of primary/secondary E2 effect only.

*Time (h) indicate hours

Dataset	Platform	PMID	Time (h)	E2 Dose	Control/ E2 (n)	Meta-analysis group
GSE11506	GPL570	PMID: 17510434	3	1e-9 M	3/3	Early treatment
GSE11467	GPL96	PMID: 19920189	2	10 nM	3/3	Early treatment
GSE9253	GPL571	PMID: 17515612	3	100nM	2/2	Early treatment
GSE79761	GPl570	PMID: 27141101	4	10nM	3/3	Early treatment
GSE26459	GPl570	PMID: 21233418	4	10 to 9M	3/3	Early treatment
GSE26834	GPL571	PMID: 21541704	3	10nM	2/3	Early treatment
			24	10nM	3/3	Late treatment
GSE23610	GPl570	PMID: 21464939	6	0.1 μM	3/3	Late treatment
GSE6800	GPl570	PMID: 17880733	24	1 nM	2/2	Late treatment
GSE11352	GPl570	PMID: 17542648	12	10 nM	3/3	Late treatment
			24	10 nM	3/3	Late treatment
			48	10 nM	3/3	Late treatment
GSE35428	GPL570	PMID: 22570330	24	1 nM	10/10	Late treatment
GSE3529	GPl96	PMID: 15986123	24	10-8 M	2/2	Late treatment
GSE8597	GPL570	PMID: 17986456	24	E2=25nM	4/4	Primary/secondary Estrogen targets
				Cyc=10 ug/ml	4/4	

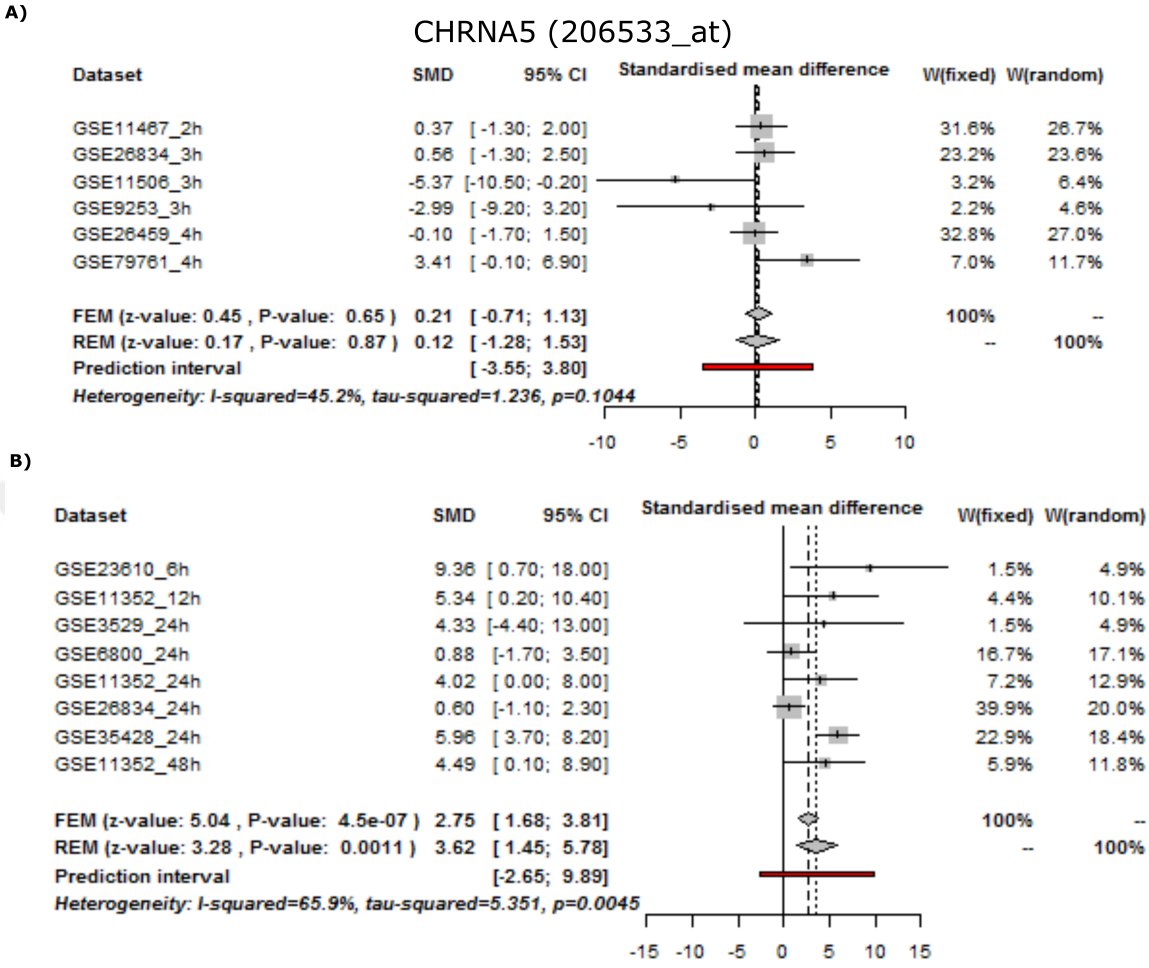


Figure 4.22: Meta-analysis of CHRNA5 expression in E2 treated MCF7 expression cohort.

Forest plot of CHRNA5 probeset expression in A) early E2 treatment cohort (2-4h) and B) Late E2 treatment cohort (6-48h). Study name represents GEO accession numbers along with time-point used in analysis. SMD= Standardized mean difference; W(fixed)= weight assigned to dataset according to fixed effect model; W(random)=Weight assigned to dataset according to random effect model; FEM=Fixed effect model; REM=Random Effect Model; CI=Confidence Interval.

4.4.9. Prediction of secondary E2 targets

Cycloheximide is a well-known inhibitor of protein translation [318, 450] and ultimately blocks the transcription of downstream targets. In light of this fact, transcriptomic changes observed after cycloheximide treatment merely represent primary targets of E2 signaling. Previously in our lab, GSE8597 public dataset was reanalyzed to show that CHRNA5 was induced by E2 but not by E2 in the presence of cycloheximide using a One-way ANOVA (Rumeysa Biyik, MS Thesis, 2009; Table 4.10) thus was suspected

of being a secondary E2 target. In my thesis, GSE8597 that was excluded from late E2 meta-analysis group was separately analyzed for calculating a PS-score (as described in Materials and Methods) for CHRNA5 and for the whole transcriptome. Expectedly, CHRNA5 probeset was significantly induced in E2 treated samples but its expression was highly reduced in cycloheximide+estrogen treatment group, with a large PS-score. *Limma* findings confirmed that CHRNA5 induction in response to E2 treatment might be secondarily regulated (Figure 4.23).

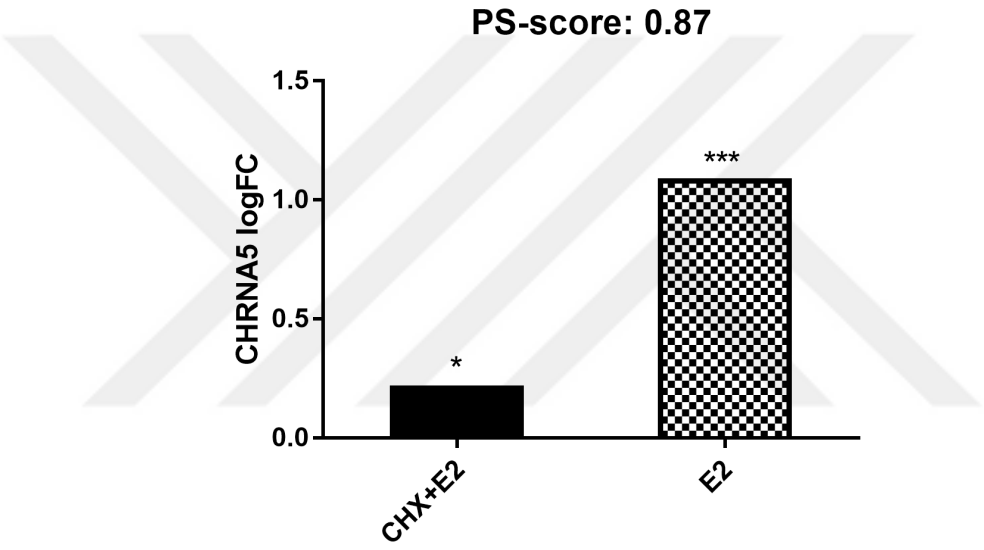


Figure 4.23: CHRNA5 expression in E2 and cycloheximide treated MCF7 cells (GSE8597). Significance is shown by p-values obtained from limma analysis. *p<0.05, **p<0.01, ***p<0.001.

In addition to the CHRNA5 probeset expression, list of all E2-induced secondary targets was predicted by setting a PS-score threshold of 0.5. CHRNA5 inclusive, 367 secondarily induced E2 targets were predicted. These probesets and PS-scores were presented in Appendix II.

List of independently predicted E2-induced secondary targets was previously not available so this study was the first to quantitatively determine these targets through reanalysis of GSE8597. For this reason, evaluation of the true-positive prediction on the basis of PS-score was not possible. In order to take into account the false positive predictions of E2- induced secondary targets the list of ER-bound genes was retrieved from a recently performed study in breast cancer cell lines. Jia et al. performed ChIPseq

in MCF7 and T47D and combined these predictions with E2-treatment gene expression profiles (in MCF7) to present the list of direct ER bound targets (207 genes) [389]. Hypothetically, these direct ER-bound targets cannot be only secondarily induced and their overlap with the predicted list can help evaluate reliability. After retrieving Affymetrix probesets (147 jetset best probesets), PS-score (logFC of E2 treatment – logFC of E2+cycloheximide) along with PS-score based secondary E2-induced target list were mapped onto ER-bound direct targets. From the list of these 147 ER interacting genes, only 9 genes were shared with the predicted secondary E2 targets. Findings revealed a low false positivity of predicted secondarily E2-induced targets (Figure 4.24).

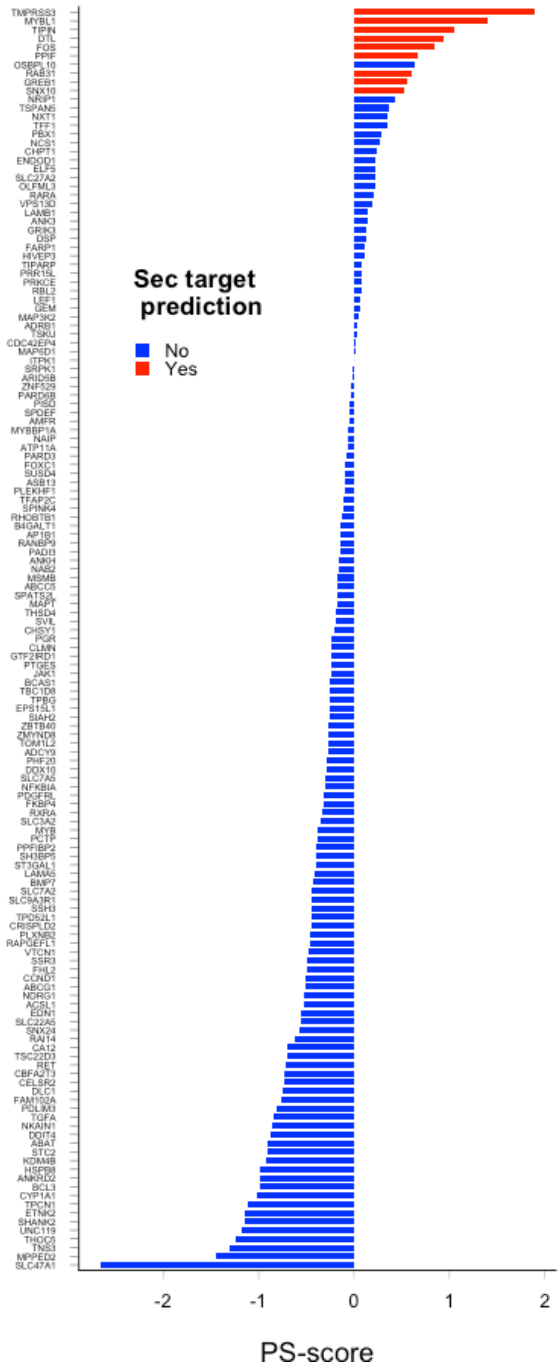


Figure 4.24: Mapping of PS-score on ER-bound gene list (retrieved from Jia et. al. [389]).

PS-score calculations were described in materials and methods. Prediction of secondary E2 target is based on two filters 1) E2 induced genes (E2: logFC>0.5; p<0.05) and 2) PS-score>0.5.

4.4.10. Correlation between CHRNA5 and Estrogen Receptor (ER) expressions

Previous finding in our lab have shown that expression of CHRNA5 in breast cancer cell line MCF7 can be regulated by E2 exposure. It's a known fact that most of the E2-mediated effects are regulated through ER-α [446]. In order to understand E2-induced CHRNA5 regulation, it is important to consider impact of ER on CHRNA5 expression. At the mRNA level, correlation between both genes was addressed by *in-vitro* analysis of COSMIC breast cell lines (n=50) and *in-vivo* investigation of TCGA breast cancer patient RNAseq data (obtained from Firebrowse). Interestingly we found that CHRNA5 and ESR1 probesets were significantly negatively correlated (R=-0.42, p=0.0027; Figure 4.25) in COSMIC breast cancer cell lines. Correlation analysis in TCGA data also confirmed that a negative correlation existed *in-vivo* as well (R=-0.48, p=2.2e-64; Figure 4.25).

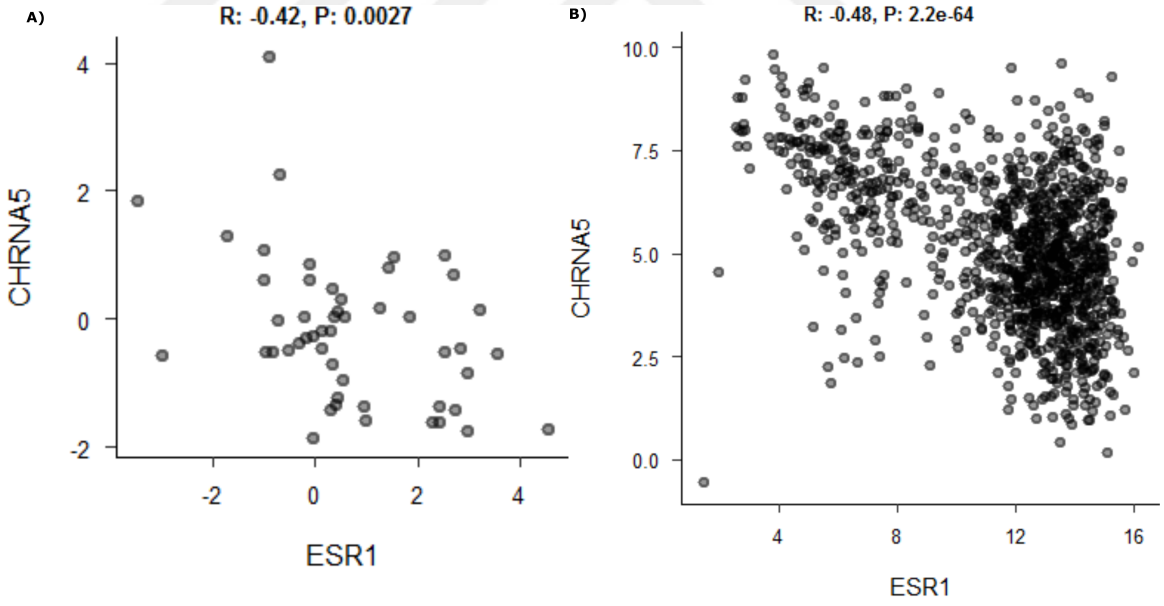


Figure 4.25: In-vitro and in-vivo correlation between expression of CHRNA5 and ESR1.

CHRNA5 and ESR1 mRNA expression in A) Cosmic breast cell lines (n=50) B) TCGA breast cancer patient data (1092 patients; level 3 RNAseq normalized data retrieved from firebrowse).

E2 target genes exhibiting negative correlations with ESR1 is not a novel concept as several studies have shown downregulation of ESR1 expression upon E2 treatment [451]. I also confirmed downregulation of ESR1 in the E2-treated MCF7 cohort (Figure 4.26). Although *early E2 treatment cohort* did not result in significant changes late *E2*

treatment cohort revealed significantly reduced ESR1 expression upon E2 treatment (high heterogeneity: I^2=68%; REM: z-value=-3.09, SMD=-5.31, p=0.002; Figure 4.26).

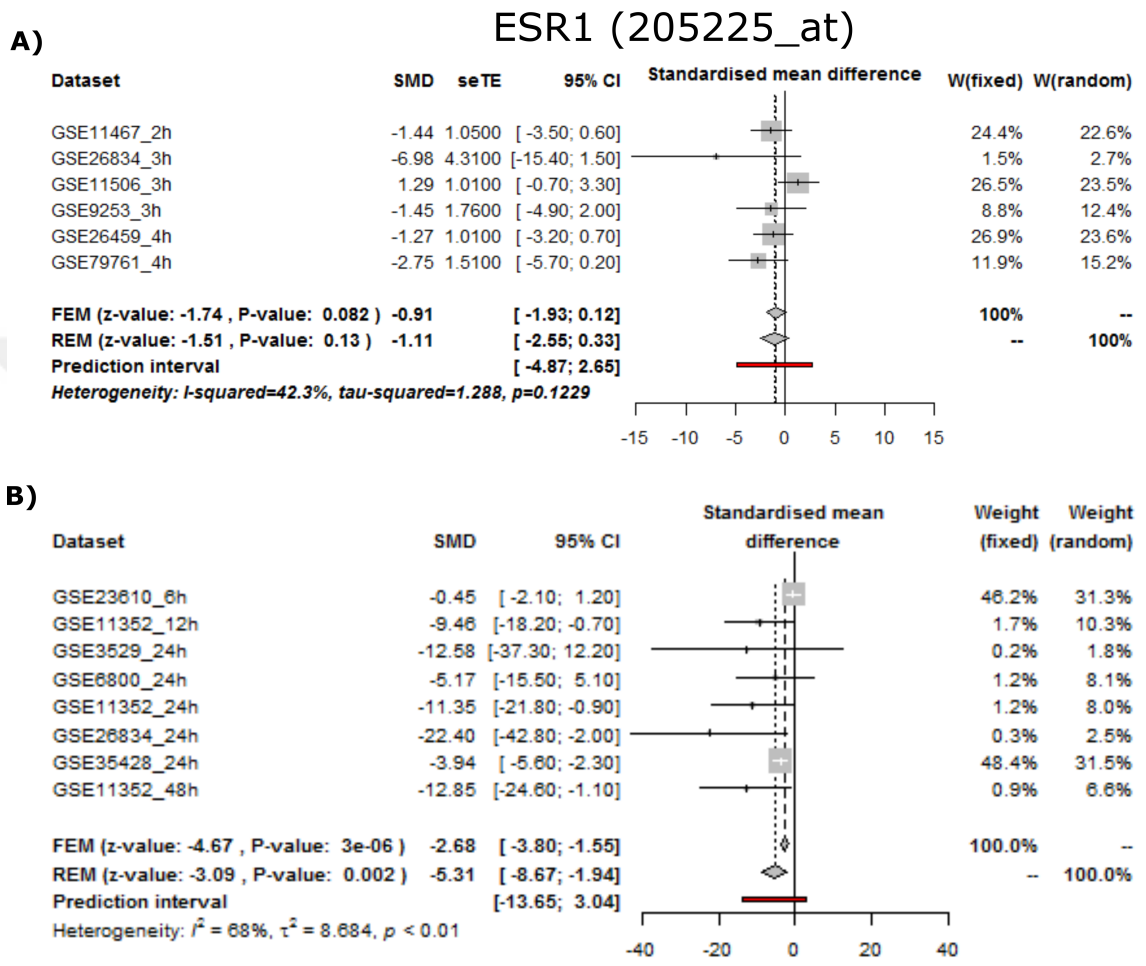


Figure 4.26: Meta-analysis of ESR1 expression in E2 treated MCF7 expression datasets.

A) Early E2 treatment cohort (2-4 hours) B) Late E2 treatment cohort (6-48 hours). Study name represents GEO accession numbers along with time-point used in analysis. SMD= Standardized mean difference; W(fixed)= weight assigned to dataset according to fixed effect model; W(random)=Weight assigned to dataset according to random effect model; FEM=Fixed effect model; REM=Random Effect Model; CI=Confidence Interval.

4.4.11. Meta-analysis of CHRNA5 expression in ER dependent breast cancer subtypes

An independent large breast cancer patient microarray expression cohort based on GPL570/GPL96 was collected form GEO database (as of 2016) and was filtered/

processed (as mentioned in Materials and Methods; Figure 2.4; PRISMA guidelines). *ER cohort* consisted of 31 datasets with 1650 ER+ and 1255 ER- breast cancer patients (Table 4.11). Visual inspection of the funnel plot and heterogeneity analysis revealed that CHRNA5 expression was highly variable in *ER cohort* (I-squared=72.6%). Egger's test showed that this heterogeneity was not associated with publication bias (p=0.99; Figure 4.27). High heterogeneity in the data classified REM scores more appropriate for analysis. Consistent with COSMIC cell lines and TCGA data, meta-differential expression analysis between ER+ and ER- samples of *ER cohort* revealed that CHRNA5 was over-expressed in ER- patients (Z-score=7.4, SMD=0.60, p=1.4e-13; Figure 4.28). Evaluation of ESR1 jetset best probeset with CHRNA5 expression also signified a negative correlation between both genes (R=-0.37, p=3.78E-18; Figure 4.29) highlighting that the negative association between both genes existed at mRNA level as well.

Table 4.11: Description of datasets used in *ER cohort.*

Dataset	Sample numbers	GPL	PMID	Duplicate samples found (removed)	ER-/ER+/NA/not consistent
GSE11001	30	GPL570	PMID: 21339180	0	10/14
GSE17907	55	GPL570	PMID: 20932292	47(0)	25/22/8
GSE18728	61	GPL570	PMID: 20012355	0	10/10/1
GSE18864	84	GPL570	PMID: 20189874, PMID: 20098429, PMID: 20100965	0	47/28
GSE20194	278	GPL96	PMID: 20064235, PMID: 20676074	124(4)	99/145
GSE20271	178	GPL96	PMID: 20829329	69 (34)	67/76/0/1
GSE20711	90	GPL570	PMID: 21910250	0	45/42/3
GSE21653	266	GPL570	PMID: 2049065, PMID: 22110708	266(27)	98/139/2
GSE22093	103	GPL96	PMID: 21191116	41(35)	39/23/5/1
GSE22597	82	GPL96	PMID: 21153052	45(38)	27/16/0/1
GSE23720	197	GPL570	PMID: 22110708, PMID: 21339811	195(148)	16/33
GSE23988	61	GPL96	PMID: 21191116	60(0)	27/27/0/7
GSE24185	103	GPL96	PMID: 21750966	0	42/58/3
GSE25055	310	GPL96	PMID: 21558518	64(64)	112/129/5
GSE25065	198	GPL96	PMID: 21558518	0	74/123/1
GSE2603	121	GPL96	PMID: 16049480	0	42/57

Dataset	Sample numbers	GPL	PMID	Duplicate samples found (removed)	ER-/ER+/NA/not consistent
GSE32072	50	GPL96	PMID: 22235097	13	6/6
GSE32646	115	GPL570	PMID: 22320227	0	44/71
GSE36771	107	GPL570	PMID: 22564725	0	28/79
GSE37946	50	GPL96	PMID: 22460789	17(17)	19/16
GSE42568	121	GPL570	PMID: 23740839	0	34/67/20
GSE42822	91	GPL96	PMID: 23158478	60(60)	22/9
GSE43358	57	GPL570	PMID: 25412710	0	19/38
GSE43365	111	GPL570	PMID: 23990869	0	18/93
GSE45255	139	GPL96	PMID: 23618380	0	48/89/2
GSE48390	81	GPL570	PMID: 24098497	0	28/53
GSE50567	41	GPL570	PMID: 21196292	0	28/6/6
GSE50948	156	GPL570	PMID: 24443618	0	104/52
GSE51450	12	GPL570	PMID: 25844937	0	6/6
GSE5460	129	GPL570	PMID: 18297396	0	53/74
GSE76274	67	GPL570	PMID: 25208879, PMID: 26921331	0	18/49

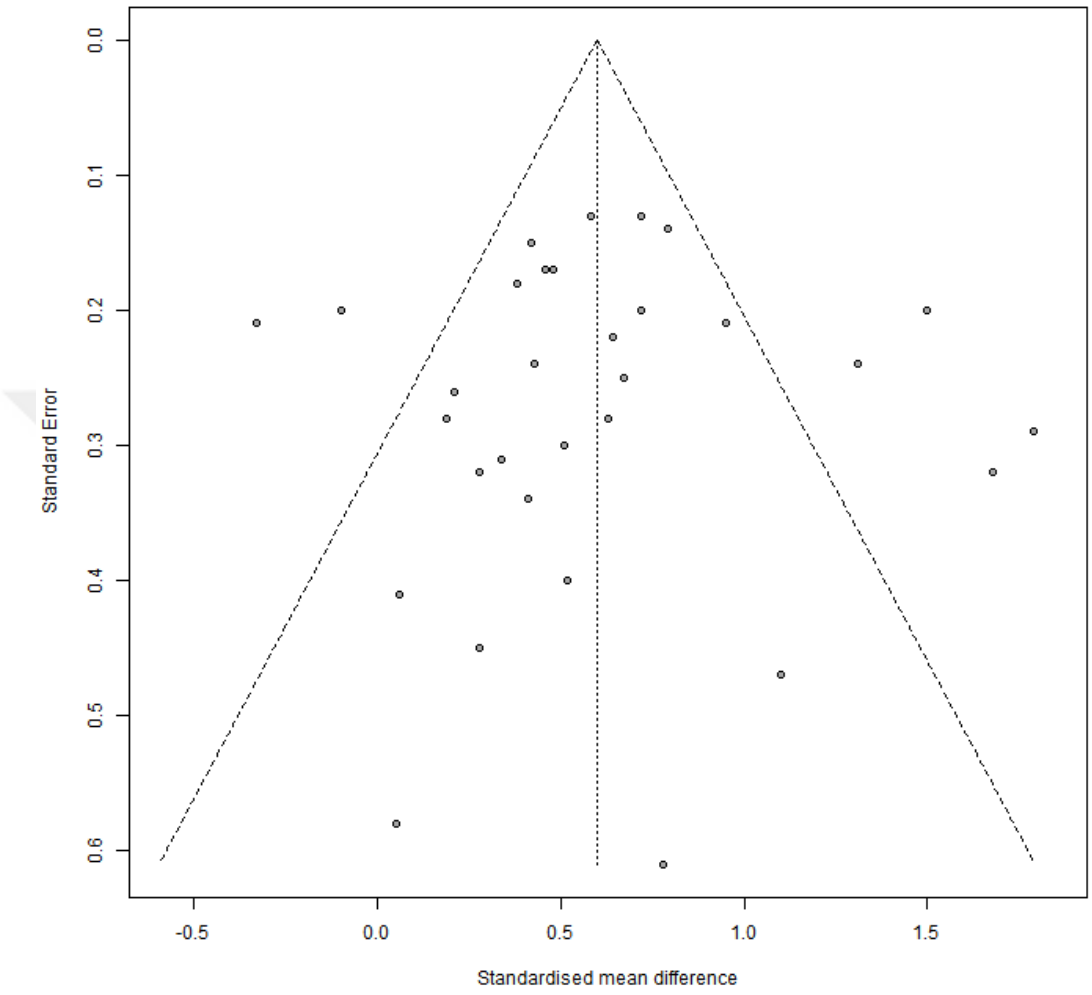


Figure 4.27: Funnel plot of CRHNA5 effect-size in ER-cohort (31 datasets).

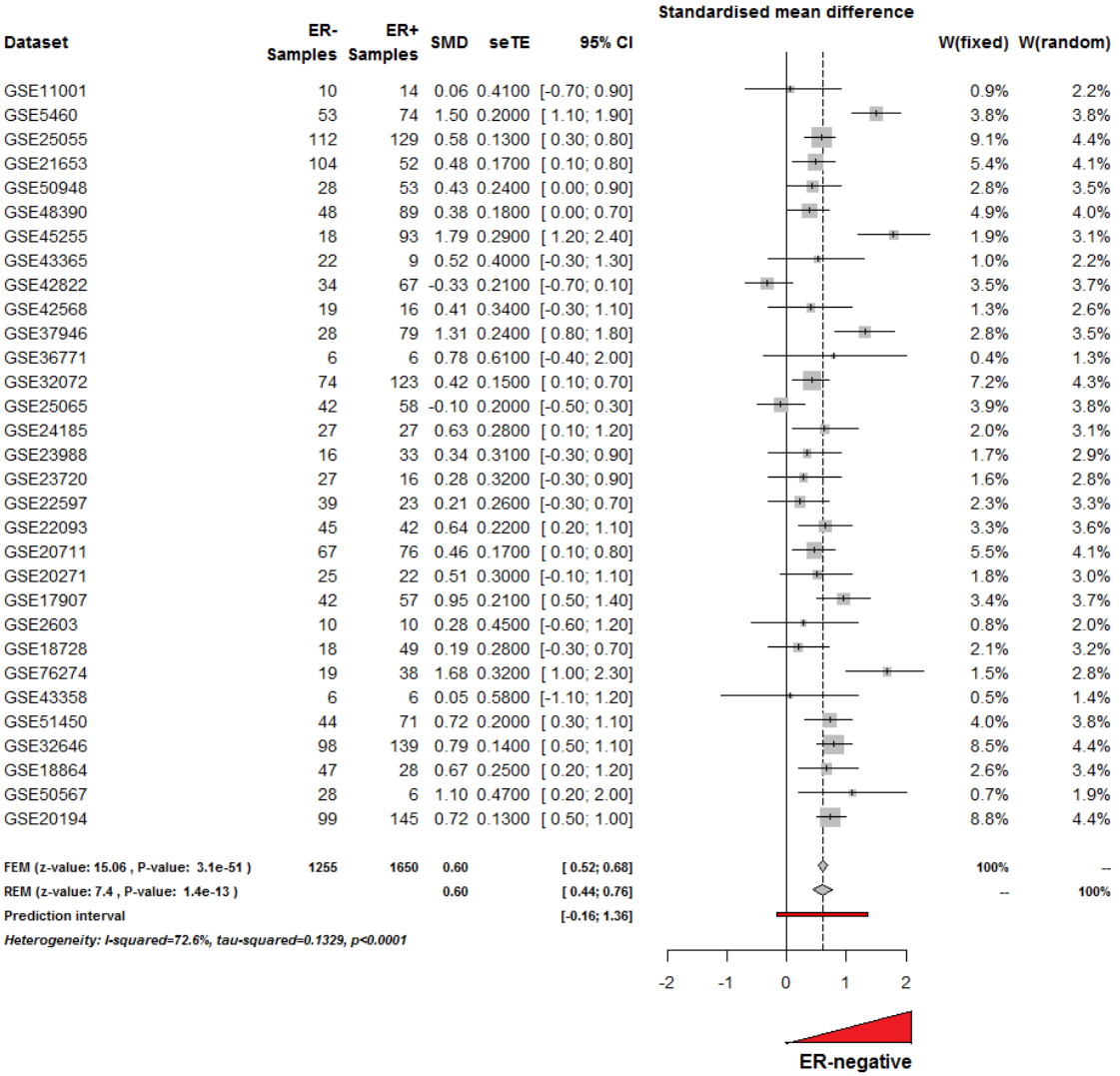


Figure 4.28: Forest plot of meta-differential expression analysis in ER-/ER+ groups of ER cohort.

Study name represents GEO accession. SMD=Standardized mean difference; seTE=standard error of treatment effect; W (fixed)= weight assigned to dataset according to fixed effect model; W (random) = Weight assigned to dataset according to random effect model; FEM=Fixed effect model; REM=Random Effect Model; CI=Confidence Interval.

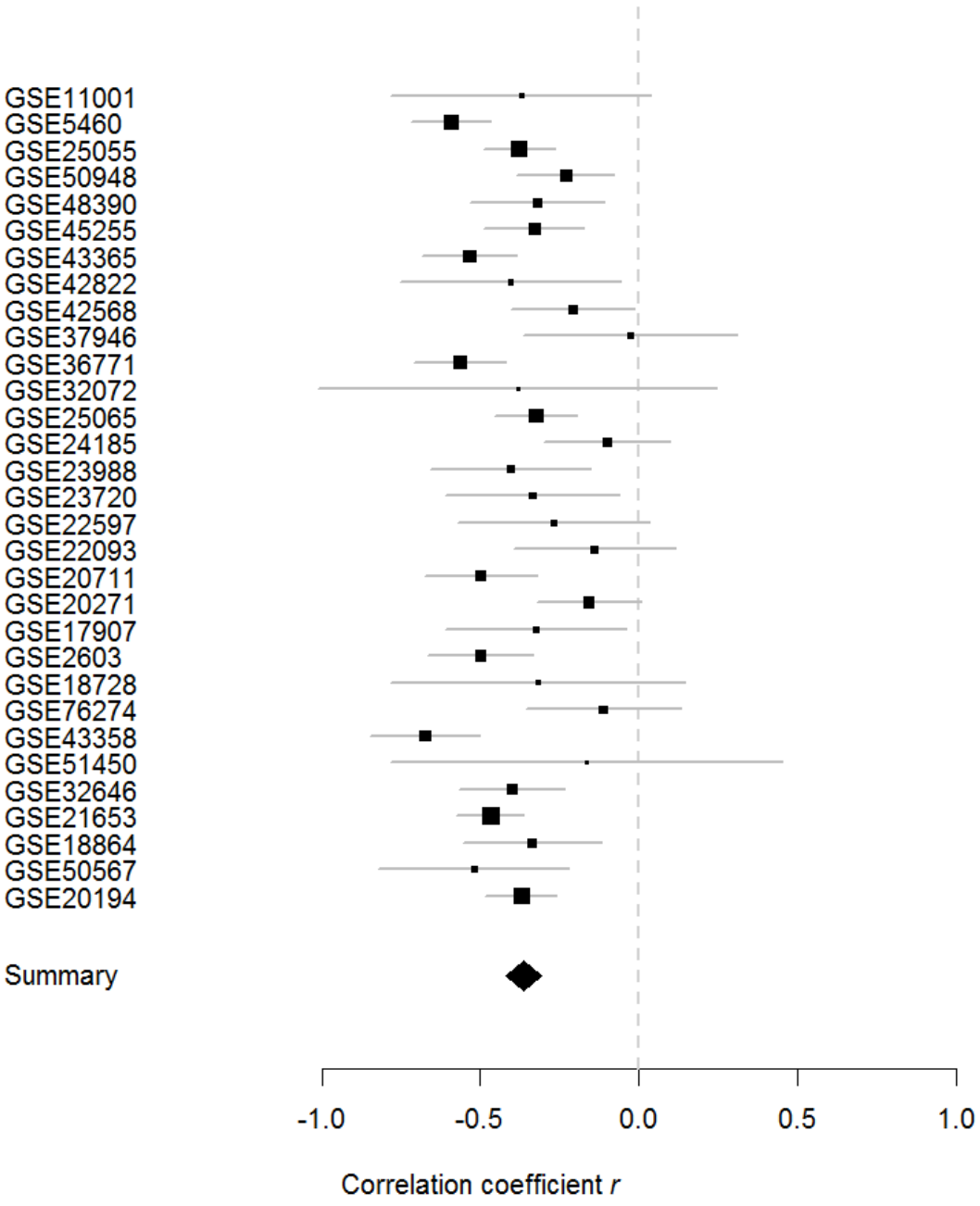


Figure 4.29: Meta-correlation analysis of CHRNA5 and ESR1 probesets in ER cohort.

Correlation score of each dataset is shown with square while lines indicate 95% confidence interval. Summary shows mean meta-correlation score.

4.4.12. Prognostic importance of CHRNA5

Prognostic importance of CHRNA5 was estimated by survival analysis of an independent untreated breast cancer patient cohort, *survival cohort*. This independent cohort was also based on gene expression data but was collected from untreated patient database of KM plotter online tool, 2010 version (5 datasets; 894 patients; Table 4.12) (PRISMA guidelines: Figure 4.5) [385]. *Survival cohort* and *ER cohort* were made independent by removing the common studies from *ER cohort*, as *ER cohort* was comparatively larger.

Table 4.12: Datasets of untreated breast cancer patient expression data used in survival cohort analysis (KM plotter, 2010).

Datasets	Platform	Pubmed ID	RFS (1/0) 1= Relapse
GSE2034	GPL96	PMID: 15721472	107/179
GSE2990	GPL96	PMID: 16478745	29/33
GSE3494	GPL96	PMID: 16141321	45/103
GSE7390	GPL96	PMID: 25788628, PMID: 17545524	91/107
GSE11121	GPL96	PMID: 18593943	46/154

CHRNA5 meta-analysis results of *survival cohort* revealed less heterogeneity (I^2=47.4%; p-value=0.1). FEM showed that higher expression of CHRNA5 in breast cancer patient was associated with significantly high mHR indicating worse prognosis (FEM z-value: 2.73, SMD=0.32, p=0.0063; Figure 4.30).

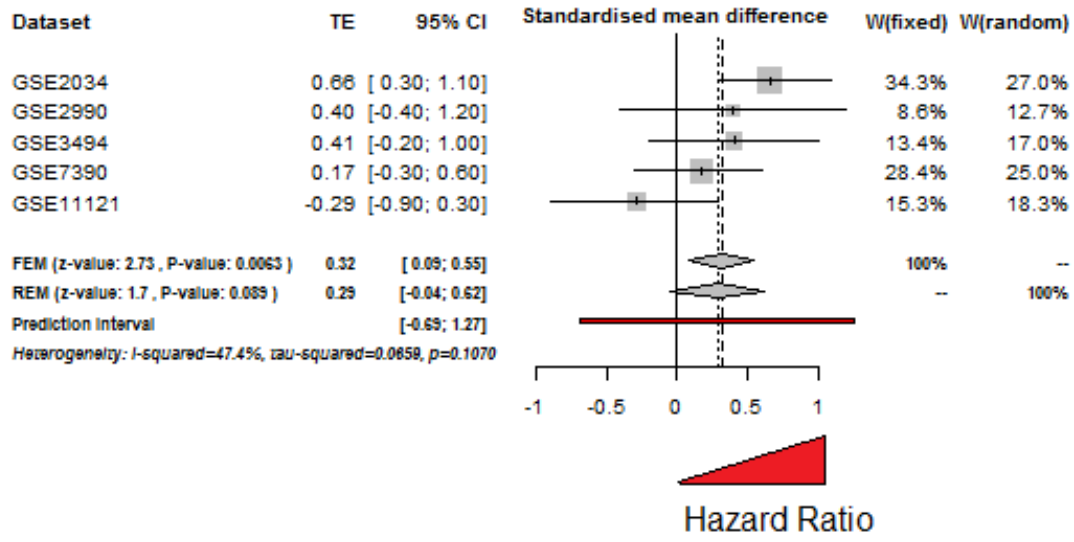


Figure 4.30: Meta-survival analysis results of CHRNA5 probeset in survival cohort.

Hazard ratios were combined using meta-analysis. TE indicates standardized mean difference; W (fixed)= weight assigned to dataset according to fixed effect model; W (random)=Weight assigned to dataset according to random effect model; FEM=Fixed effect model; REM=Random Effect Model; CI=Confidence Interval.

Furthermore, splitting each dataset of *survival cohort* for known ER status resulted in 4 datasets with 199 ER- patients and 629 ER+ patients. Separate meta-analysis again showed that higher expression of CHRNA5 was indeed hazardous in untreated ER+ breast cancer patients (I^2=0%, p=0.81; FEM: z-value=3.29, SMD=0.52, p=9.9E-4; Figure 4.31A) but non-significant changes were observed in untreated ER- breast cancer patients (Figure 4.31B). Observed effects might be related to less number of samples in ER- breast cancer patients.

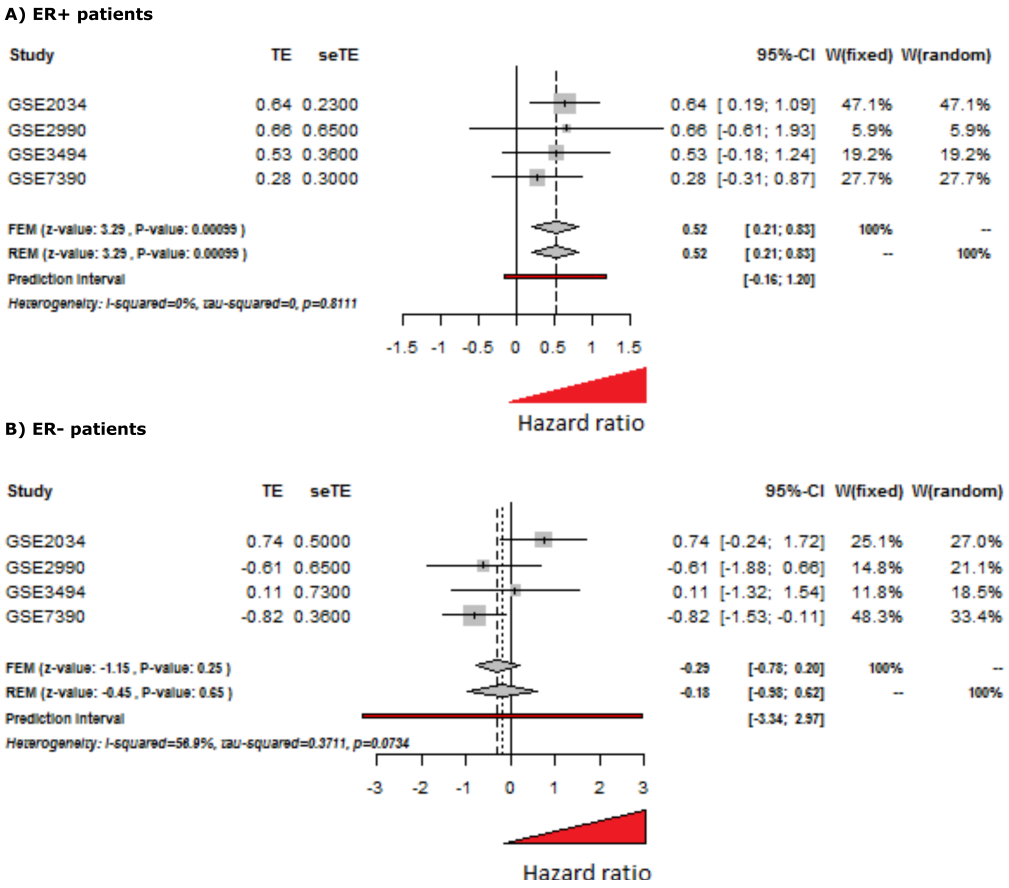


Figure 4.31: Separate meta-survival analysis of ER- and ER+ breast cancer patients from untreated *survival cohort.*

Hazard ratios were combined using meta-analysis. TE indicates standardized mean difference; W (fixed)= weight assigned to dataset according to fixed effect model; W (random)=Weight assigned to dataset according to random effect model; FEM=Fixed effect model; REM=Random Effect Model; CI=Confidence Interval.

4.4.13. Mapping of genome-wide CHRNA5 correlation profile with effect size of *ER cohort* and hazard ratio of *untreated survival cohort*

CHRNA5 co-expression profile was evaluated by performing expression meta-correlation analysis of each gene with the CHRNA5 probeset values separately in *ER* and *survival cohort*s. Meta-correlation analysis helped identify genes facilitating (positively correlated) or inhibiting (negatively correlated) CHRNA5 mediated actions. Besides, genome-wide meta-differential expression analysis of *ER cohort* and meta-

survival analysis of *survival cohort* were performed to obtain transcriptome-wide ER-/ER+ effect-sizes and mHRs, respectively.

CHRNA5 meta-correlation profile was evaluated for differential ER-/ER+ expression by plotting CHRNA5 meta-correlation scores against ER-/ER+ effect sizes of all genes. Like CHRNA5 itself, positively correlated genes were generally over-expressed in ER- breast cancer and negatively correlated profile had increased expression in ER+ patients. The phenomenon was consistently observed for meta-correlation profile obtained from both *ER* and *survival cohorts*. A significant positive correlation was found between transcriptome-wide CHRNA5 correlation (of both *ER* and *survival cohort*) and the effect-size in ER-/ER+ patients (CHRNA5 meta-correlation based on *ER cohort*: All probesets: R=0.66, p=0; significant probeset: R=0.79, p=0; Figure 4.32A) (CHRNA5 meta-correlation based on *survival cohort*: All probeset: R=0.78, p=0; significant probeset: R=0.88, p=0; Figure 4.32B). Considering highly co-expressed genes in both directions (meta-correlation score>0.3 or <-0.3 and p<0.05), which were significantly differentially expressed in *ER cohort* (FDR<0.05), association statistics was calculated using Fisher's exact test and Odds ratio. Findings revealed a highly significant positive association (CHRNA5 correlation based on *survival cohort*: Fisher's exact p=1.82E-228; CHRNA5 correlation based on *ER cohort*: Fisher's exact p=2.68E-138). Odds ratio resulted in infinity since no probesets existed in the oppositely regulated direction, i.e., positively correlated with CHRNA5 and over-expressed in ER+ patients or vice-versa in both cohorts. Strong statistics and validation from these two cohorts emphasized that CHRNA5 along with its positively co-expressed genes were part of a network that was over-expressed in ER- patients.

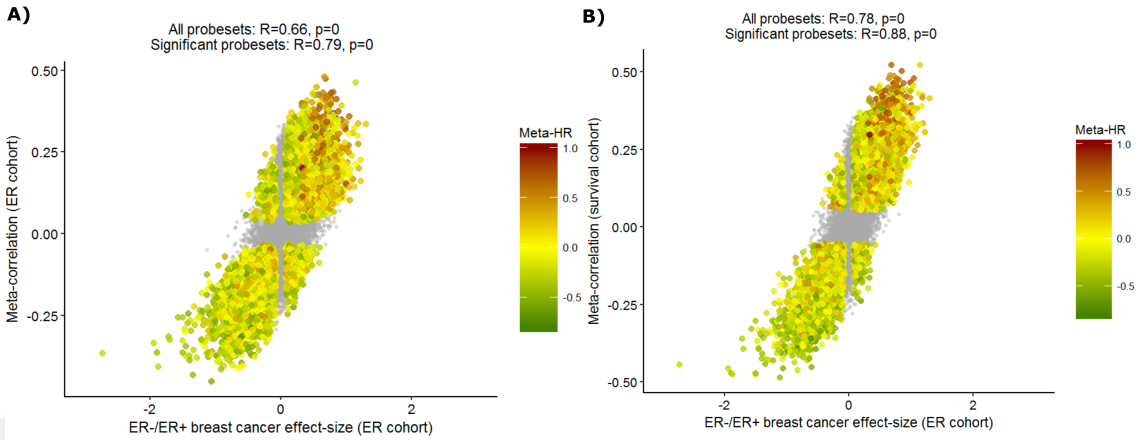


Figure 4.32: Comparison of CHRNA5 meta-correlation profile with ER-/ER+ effect-size and meta-hazard ratio.

Meta-correlation expression profile calculated from A) ER cohort and B) survival cohort is shown on y-axis with Genome wide ER-/ER+ effect-size on x-axis. Grey color indicates that gene showed non-significant ER-/ER+ meta-differential expression and/or non-significant correlation with CHRNA5. As a third layer, meta-hazard ratio is mapped with color code. Correlation analysis results written on top of plot title are either based on all probesets or genes significantly regulated, i.e., colored dots on plot.

As shown above, higher expression of CHRNA5 was associated with worse prognosis (FEM z-value: 2.73, SMD=0.32, p=0.0063; Figure 4.30). Interestingly, mapping survival information on top of correlation analysis and differential expression showed that genes positively correlated with CHRNA5 also exhibited high HRs. Comparative statistics again resulted in highly positive associations between significant CHRNA5 correlation profile and survival meta-analysis (*survival cohort* derived CHRNA5 co-expression profile: p=4.36E-51, OR=66.53; *ER cohort* based CHRNA5 correlation profile: p=7.38E-15, OR=20.21; Table 4.13). These findings suggested that CHRNA5 and its co-expressed network might be related to worse prognosis of ER- patients (Figure 4.32).

Table 4.13: Contingency table and association statistics of CHRNA5 meta-correlation profile.

CHRNA5 meta-correlation analysis was separately conducted for *survival* and *ER cohort* and was then compared with hazard ration and differential effect in ER-/ER+ breast cacner.

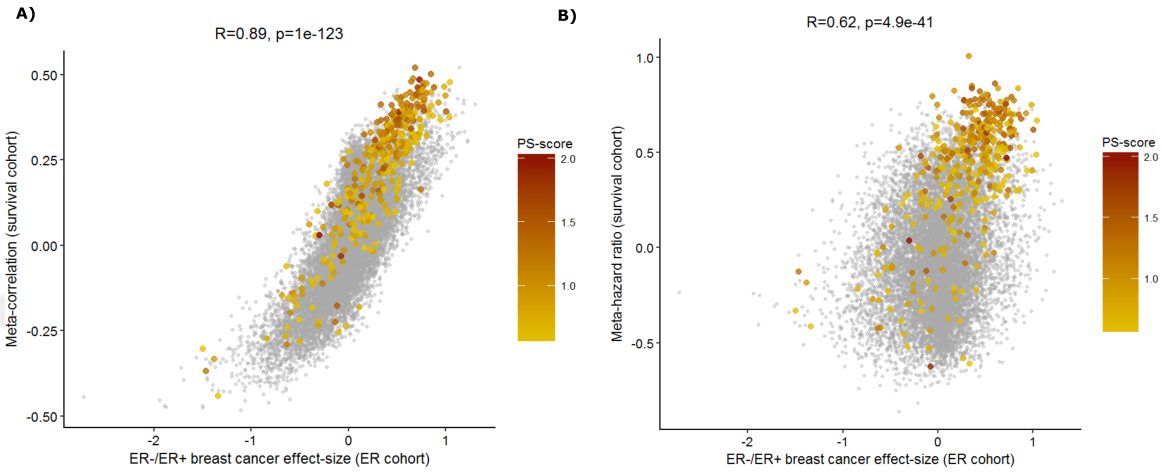
		Survival Cohort				*ER cohort*			
		ER-/ER+ effect-size		Meta-hazard ratio		ER-/ER+ effect-size		Meta-Hazard ratio	
		Up	Down	Up	Down	Up	Down	Up	Down
CHRNA5 R.mean	Up	450	0	209	49	251	0	82	52
	Down	0	326	8	127	0	213	5	65
	Odds ratio	Inf		66.53		Inf		20.21	
	Pvalue	1.82E-228		4.36E-51		2.68E-138		7.38E-15	

4.4.14. Evaluation of secondary E2-induced targets in meta-analysis results.

As mentioned before, PS-score helped us predict 367 secondarily E2-induced targets, including CHRNA5. Elevated expression of CHRNA5 in ER- patients along with a high HR prompted me to investigate behavior of the other 366 genes in different meta-analysis signatures (i.e., CHRNA5 meta-correlation profile, ER-/ER+ meta-differential expression analysis and survival mHR analysis). Mapping these secondary E2-induced targets for their CHRNA5 correlations and respective ER-/ER+ effect-sizes indicated that secondarily E2-induced target genes were highly enriched in the cluster, which was over-expressed in ER- patients and was positively correlated with CHRNA5 (in *survival cohort*). Fisher's exact test showed significant positive association (p=9.06E-40; Figure 4.33A, C).

Our findings suggested that secondary E2 target genes were enriched in ER- patients and were positively correlated with CHRNA5. Prognostic analysis of these target genes showed that besides having higher expression in ER- patients, these targets also possessed high HRs with significant positive correlations between the effect-size and mHRs (R=0.62, p=4.9E-41; Figure 4.33B). Large number of these secondary E2-

induced targets were clustered in UPUP category, i.e., high ER-/ER+ effect-size and high HR (202 out of 366 probesets; Figure 4.33D).



C)

CHRNA5 R.mean	ER-/ER+ effect-size: Up	ER-/ER+ effect-size: Down
Up	259	5
Down	0	34
P-value	9.06E-40	

D)

Meta-Hazard ratio	ER-/ER+ effect-size: Up	ER-/ER+ effect-size: Down
Up	202	10
Down	4	9
Odd ratio	43.15	
Pvalue	1.61E-08	

Figure 4.33: Comparison of secondary E2-induced targets with different meta-analysis score.

A) CHRNA5 meta-correlation profile from survival cohort vs ER-/ER+ effect-size B) meta-HR vs ER-/ER+ effect-size and association statistics of A and B are summarized in C and D. Background genes (all affymetrix probesets) are grey in color while colored dots represent secondary E2-induced targets and their PS-score.

4.4.15. CHRNA5 co-expression network

Although CHRNA5 meta-correlation analysis helped ranking the whole transcriptome for its correlation with CHRNA5, clustering of this profile can help identifying closely linked modules. Setting a meta-correlation score threshold of 0.2 (r.mean>0.2 and r.mean<-0.2) resulted in larger list of genes (3070 jetset best probesets). WGCNA pipeline was followed to cluster these genes in modules and then ranking of those modules based on mean meta-correlation values of participating genes provided tight

CHRNA5 co-expression clusters. WGCNA pipeline is designed for a single dataset. However, to analyze multiple datasets, WGCNA allows a comparison strategy after obtaining modules from each study. As *ER cohort* based on patient data was extensive with 31 datasets, comparison of each dataset can be difficult and time-consuming. WGCNA analysis pipeline therefore was slightly modified to allow meta-correlation score incorporation instead of regular correlation analysis for calculation of adjacency matrix. Strategy has been described in Materials and Methods. Six modules retrieved from WGCNA were shown in Figure 4.34. KEGG annotation of these modules illustrated the proliferative potential of CHRNA5 positively co-expressed modules. Enriched pathways included cell cycle, DNA replication and repair pathways (Table 4.14) however negatively co-expressed modules were enriched with cell junction related pathways, e.g., focal adhesion along with extracellular matrix adhesion pathway.

Table 4.14: KEGG pathways of the CHRNA5 co-expression modules (Figure 4.33).

	Term	Count	PValue	Benjamini
Yellow (positive)	hsa04110: Cell cycle	35	1.12E-29	1.49E-27
	hsa03030: DNA replication	17	3.05E-18	2.03E-16
	hsa04115: p53 signaling pathway	12	1.12E-07	4.94E-06
	hsa04114: Oocyte meiosis	13	2.53E-06	8.40E-05
	hsa03430: Mismatch repair	7	5.79E-06	1.54E-04
	hsa00240: Pyrimidine metabolism	12	9.96E-06	2.21E-04
	hsa04914: Progesterone-mediated oocyte maturation	11	1.22E-05	2.32E-04
	hsa03440: Homologous recombination	6	3.00E-04	4.98E-03
	hsa00230: Purine metabolism	12	1.12E-03	1.65E-02
	hsa03420: Nucleotide excision repair	6	2.87E-03	3.75E-02
Turquoise (positive)	hsa04080: Neuroactive ligand-receptor interaction	83	1.29E-23	3.51E-21
	hsa05033: Nicotine addiction	15	5.82E-06	7.88E-04
	hsa04060: Cytokine-cytokine receptor interaction	41	3.72E-05	3.35E-03
	hsa05032: Morphine addiction	22	4.95E-05	3.35E-03
	hsa04723: Retrograde endocannabinoid signaling	20	1.02E-03	5.38E-02
	hsa04020: Calcium signaling pathway	30	1.04E-03	4.60E-02
Green (negative)	hsa04510: Focal adhesion	16	5.44E-07	9.80E-05
	hsa04512: ECM-receptor interaction	11	7.87E-07	7.08E-05
	hsa04151: PI3K-Akt signaling pathway	17	7.37E-05	4.41E-03
	hsa05200: Pathways in cancer	18	1.02E-04	4.58E-03
	hsa04142: Lysosome	9	5.48E-04	1.95E-02
	hsa05146: Amoebiasis	8	1.23E-03	3.64E-02

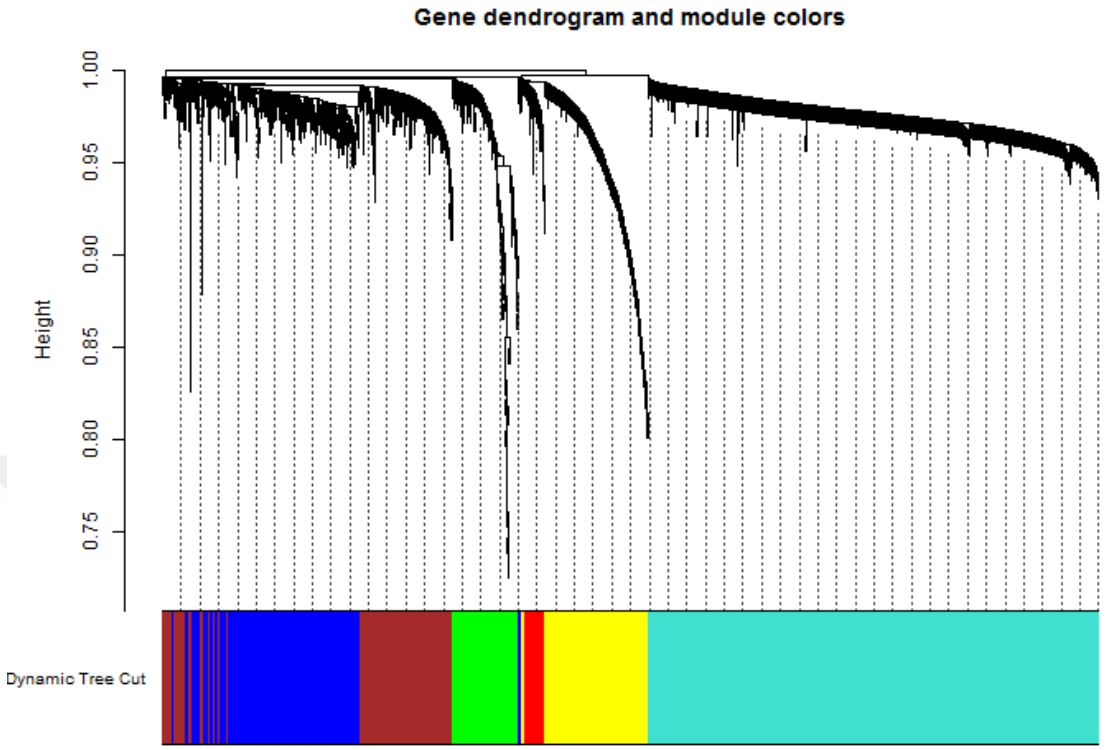


Figure 4.34: CHRNA5 co-expression module retrieved from WGCNA.

Yellow and brown, being top positively and negatively co-expressed modules respectively were further accessed. STRING interaction of both modules showed highly connected modules (Figure 4.35, 4.36). Transcription factor enrichment using *iRegulon* plugin were shown by octagons in figures. Transcription factor enrichment revealed that genes present in the positively co-expressed module (yellow) were regulated by multiple members of E2F transcription factor (TF) family. E2Fs are known primary E2-targets and enrichment of E2Fs highlight the presence of secondary E2-targets in the network of positively co-expressed module [318] and their elevated expression has already been shown in basal breast cancer [331]. Predicted secondary E2-induced targets (by PS-score) were also shown by yellow border color of nodes in both networks. Mapping these targets validated that predicted secondary E2 targets were clustered in CHRNA5 positively co-expressed module.

Interestingly, estrogen receptor (ESR1) was one of the transcription factors that were enriched in the negatively co-expressed brown module. E2 shows most of its actions by

binding to ER, which highlights presence of primary E2 targets in the negatively co-expressed module. This finding was confirmed using the list of direct ER targets predicted by ChIPseq [389]. Results showed that a large number of ER-bound targets had a negative effect-size indicating over-expression in ER+ breast cancer patients (Figure 4.37) and also had negative correlation with CHRNA5 indicating antagonistic effect between CHRNA5 and direct targets of E2/ESR1 (Figure 4.37).

Distribution of different scores for genes present in each WGCNA module was shown in Figure 4.38. From 366 secondary E2-induced targets, 180 were clustered in highly ranked positively co-expressed module (yellow). Statistics validated that all positively co-expressed modules seemed to possess large number of genes enriched in ER- breast cancer patients and high mHRs while opposite phenomenon was observed in the negatively co-expressed modules (Figure 4.38).

Genes of each module were evaluated for distribution of different calculated scores (meta differential expression analysis form *ER cohort*, meta-hazard ratio analysis from *survival cohort* and predicted secondary E2-induced targets) in CHRNA5 co-expressed modules. Enrichment of hazardous, ER- breast cancer specific genes and secondary E2-induced targets in all three CRHNA5 positively co-expressed modules helped us adding CHRNA5 to the list of proliferative genes in breast cancer (Figure 4.38).

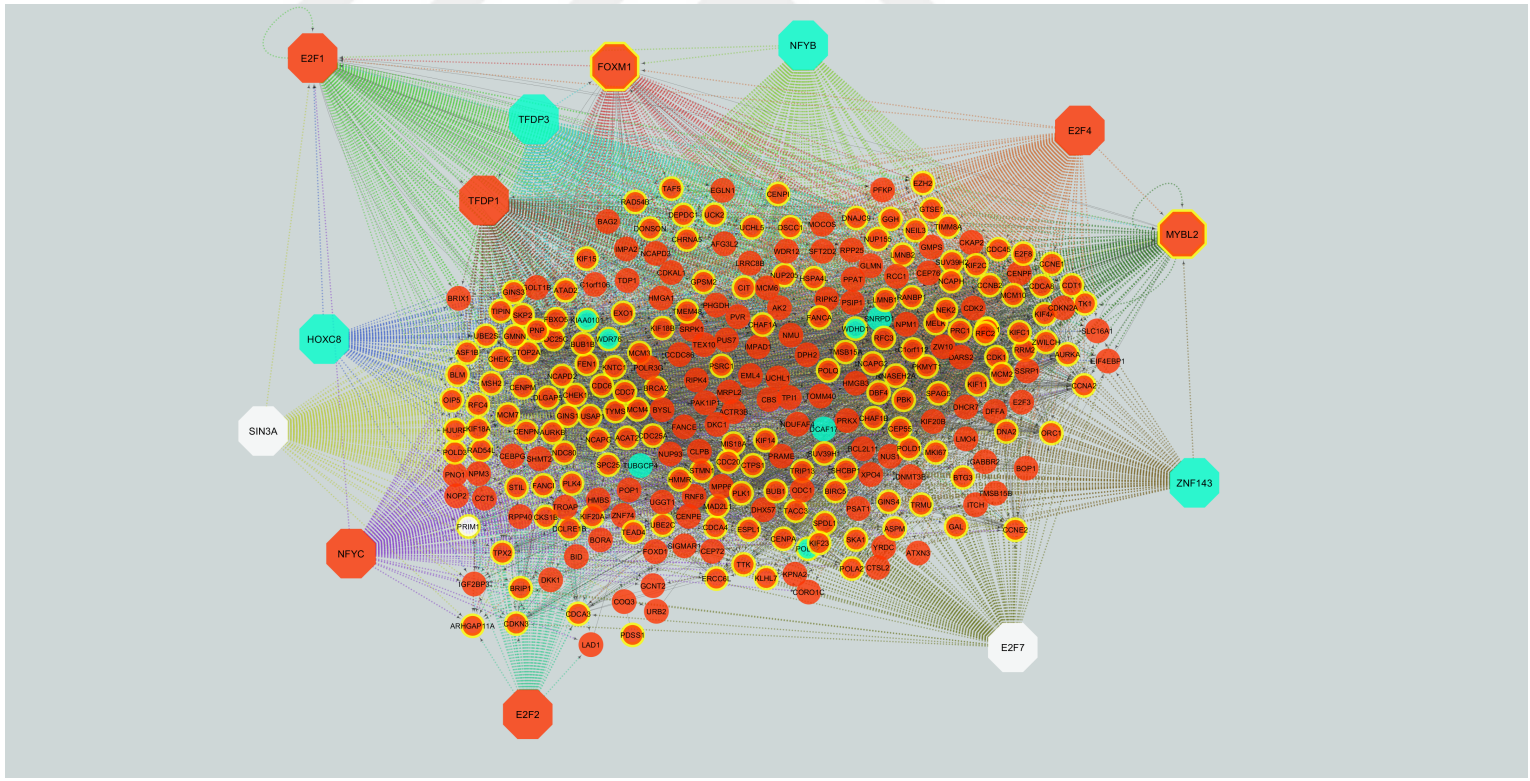


Figure 4.35: Interaction network of yellow module.

Octagonal nodes represent enriched TFs from IRegulon. Brown nodes show gene upregulation in ER- patients and turqoise nodes indicate upregulation in ER+ breast cancer patients. White nodes represent either no-significant changes or unavailable data for gene. Yellow node borders indicate genes predicted to be secondary E2-induced targets according to PS score.

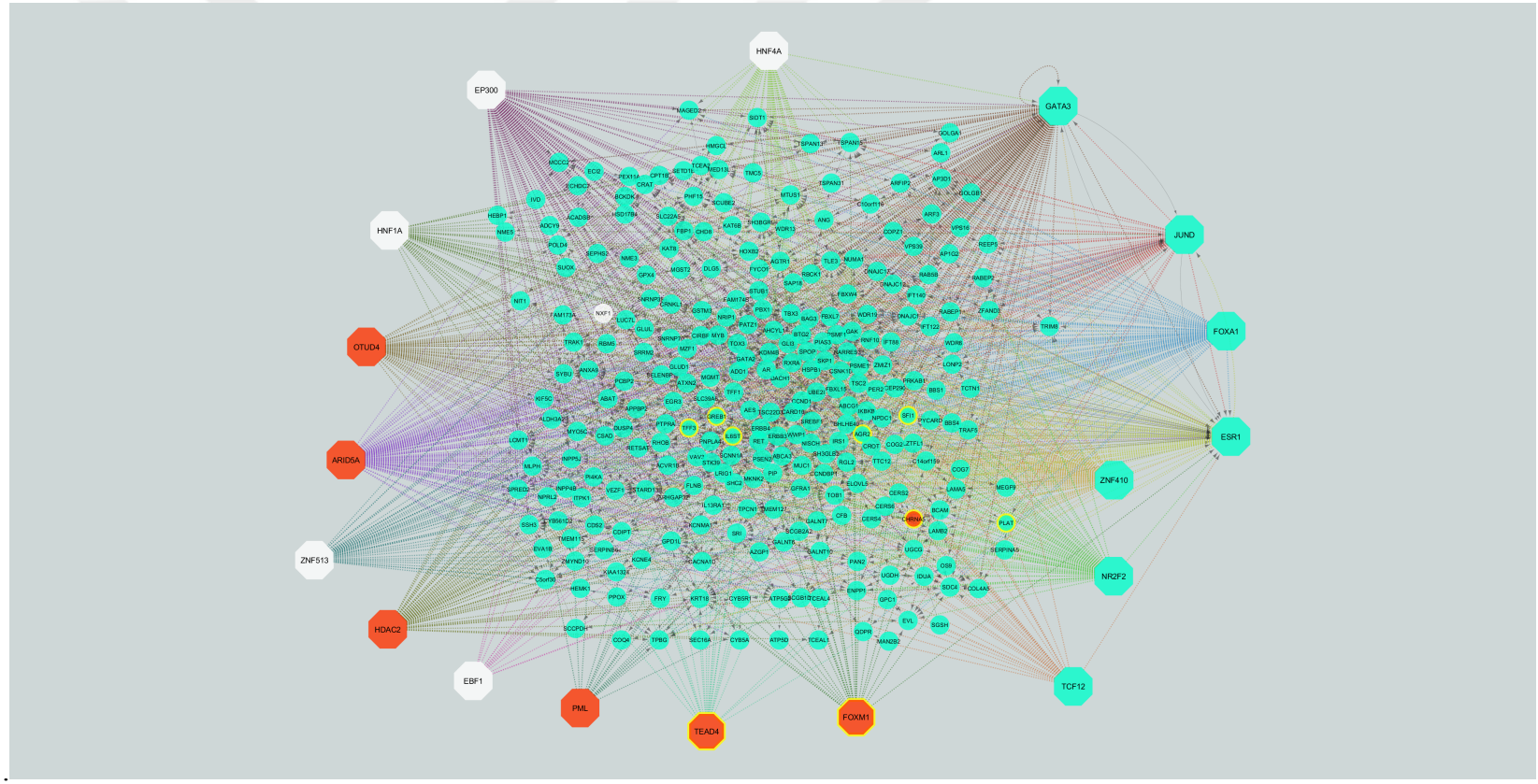


Figure 4.36: Interaction network of brown module.

Octagonal nodes represent enriched TFs from IRegulon. Brown nodes show gene upregulation in ER- patients and turqoise nodes indicate upregulation in ER+ breast cancer patients. White nodes represent either no-significant changes or unavailable data for gene. Yellow node borders indicate genes predicted to be secondary E2-induced targets according to PS score.

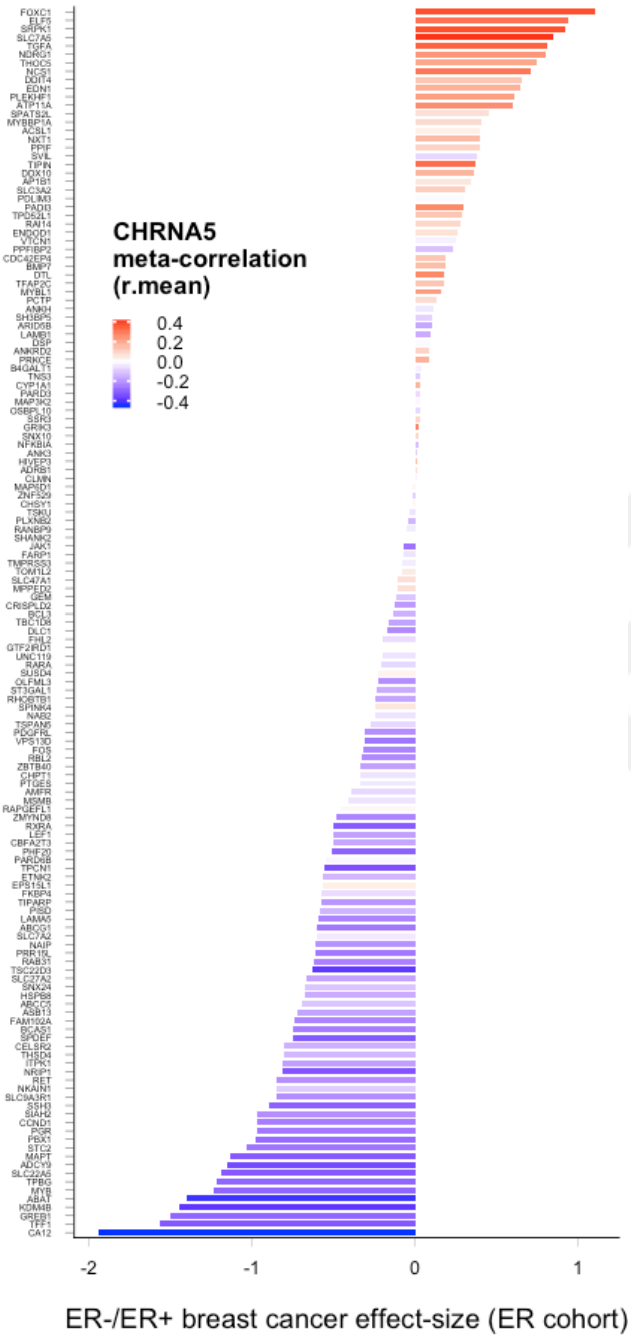


Figure 4.37: Distribution of effect-size (ER-cohort) and CHRNA5 meta-correlation score in ER bound target genes retrieved from [389].

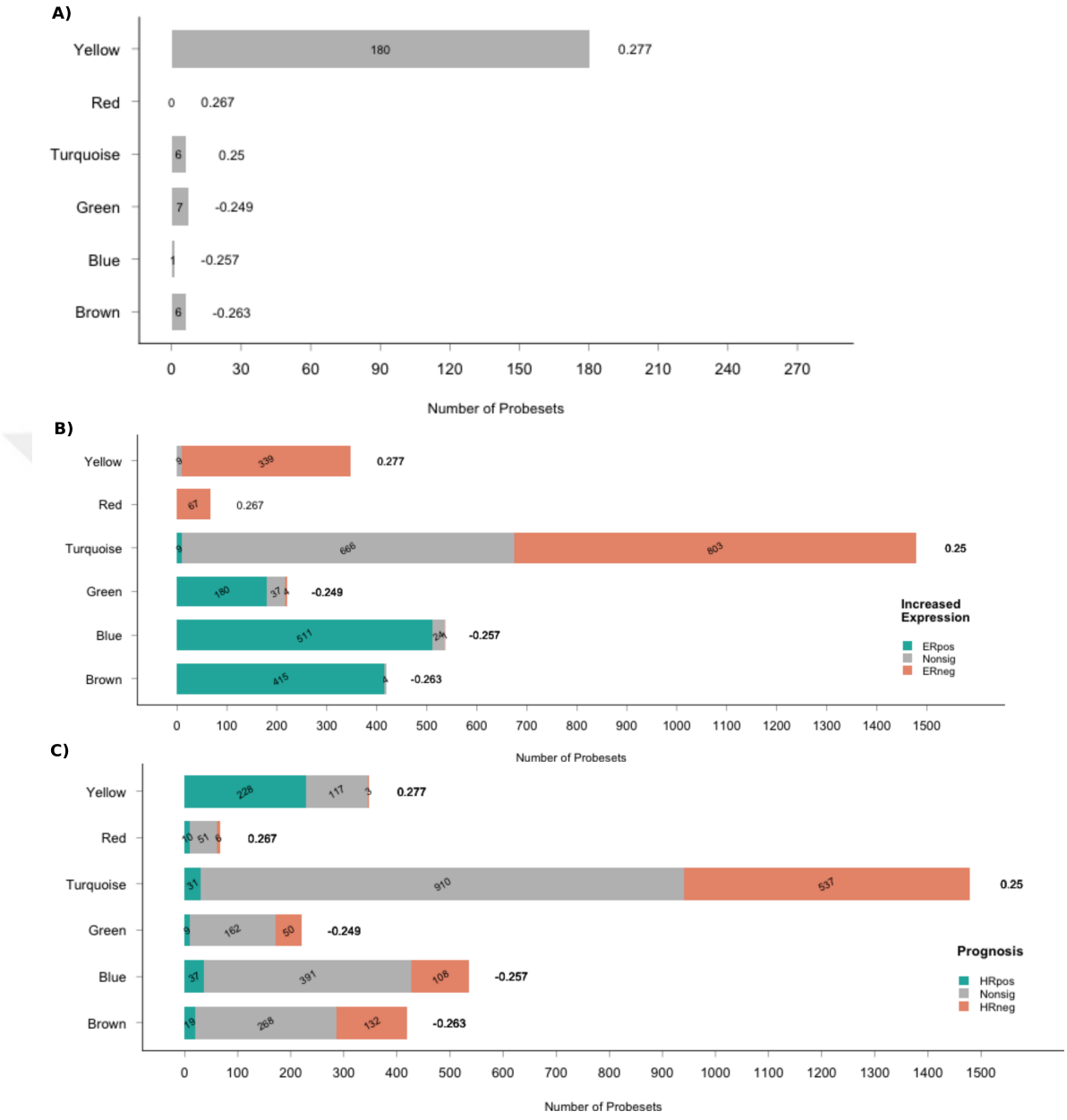


Figure 4.38: CHRNA5 co-expressed modules enrichment with calculated statistical scores.

Enrichment of WGCNA module genes with A) number of secondary E2-induced targets; B) number of genes significantly differentially expressed (FDR<0.05) in ER cohort (genes having mu.hat>0 were counted as ER- specific and mu.hat<0 with significance (FDR<0.05) were counted as ER+ specific genes); C) number of good and worse prognostic markers on the basis of significantly regulated meta-hazard ratio (p<0.05) in survival analysis. Grey area in the bar indicates non-significant genes regulation.

4.4.16. Development of an online tool for E2-mediated signaling in breast cancer

E2-mediated signaling is the main regulatory mechanism of breast cancer and for the identification of any breast cancer specific biomarkers. In the previously mentioned CHRNA5 regulation analyses, large amounts of transcriptome data were collected, pre-processed and combined with meta-analysis strategies; hence making these cohorts a valuable resource to study any gene and its co-expression profile with respect to E2-mediated signaling. Hence, I have also developed a web tool and named it E2S (Estrogen (E2) to Survival), which can facilitate users to query any gene and evaluate different aspect of regulation in E2-mediated signaling.

4.4.16.1. Modules in E2S

E2S is organized in different modules and can be extended for the incorporation of new features. E2S has four modules, each, addressing a different biological problem. First three modules are designed for individual gene regulation while the fourth module links the queried gene to the transcriptome and different type of expression datasets and functional analysis. Details are described below:

Module I: E2 Regulation

The module I is based on *in-vitro E2 cohort* described in Table 4.10 and processed using PRISMA guidelines (Figure 4.3). *E2 cohort* consists E2-treated MCF7 gene-expression profiles regulated in a time-dependent manner with another dataset focusing on primary/secondary E2 target prediction. E2 regulation module is divided in two further sub-modules:

1) For any user queried gene, early and late E2 treatment meta-analysis results, based on both FEM and REM are visualized in the form of a forest plot. For each expression profile included in early and late E2 treatment profile, dose (in description) and time (in forest plot) of E2 exposure are provided. Moreover, the hyperlink of the study can lead the user to GEO page of the dataset.

2) GSE8597 differential expression analysis results of E2 treated samples in the presence or absence of cycloheximide are visualized by showing the mean expressions of all four groups (E2 and its control in presence or absence of cycloheximide) along with a bar plot of logFC in both groups, i.e., E2 with control and cycloheximide+E2 with its control. Significance in these logFC values is indicated by the color of bars in barplot, i.e., maroon indicating significant change ($p<0.05$) in logFC and grey indicating non-significant ($p>0.05$) logFC regulation. This analysis signifies that observed E2-mediated effect on any gene could be primary or secondary. Difference in the fold-change of both groups is shown on top of the plot. Greater the difference, more chances of a gene to be secondary target and vice-versa.

Module II: ER regulation

Module II is based on *ER cohort* expression profiles obtained on 31 breast cancer datasets. ER status in a cell is generally measured by two scales: a) ESR1 expression at the transcriptome level; and b) the presence of ER protein in a cell measured by immunohistochemistry staining [452]. ER module addresses both scales to analyze ER mediated gene regulation. Module II is further divided in three sub-modules:

1) **Single dataset analysis** allows visualization of the queried gene expression (y-axis) and its correlation with expression of ESR1 (jetset best probeset) on x-axis along with color-coding of ER status (IHC) in any dataset. User can select any individual dataset from the list of 31 studies.
2) **Meta-differential expression** is intended to find differential expression between ER+ and ER- patients. Heterogeneity analysis help visualization of the effect size and standard error on a funnel plot and the heterogeneity for publication bias is identified by Egger's test. $P<0.05$ indicates potential publication bias in studies for the selected gene while $p>0.05$ illustrates heterogeneity in the dataset might be due to any other reason. After analyzing publication bias, differential expression of gene between ER-/ER+ can be visualized by the forest plot. ER-, due to being a more aggressive subtype, was used as the treatment group for comparisons with

ER+ (control). Effect-size > 0 indicates over-expression in ER- while effect-size < 0 shows an increased level of gene in ER+ patients.

3) **Meta-correlation analysis** uses ESR1 expression (jetset best probeset) for meta-analysis instead of the IHC measured ER status. Module identifies the association between gene and ER specifically at the transcript level. Correlation between the selected gene and respective ESR1 expression from each study are combined to identify meta-correlation score using DerSimonian-Laird (DSL) random-effect meta-analysis approach.

Module III: Prognosis

Prognosis module is divided in two separate survival meta-analyses, i.e., based on untreated or Tamoxifen treated patients. PRISMA guidelines for both untreated and tamoxifen treatment cohort have been provided in Material and Methods (Figures 4.5, 4.6). Survival analysis for each dataset is based on auto-best cut-off. Individual HRs from each dataset was combined using meta-analysis. Effect-size > 0 indicates that higher expression of the gene is more hazardous thus is associated with worse prognosis of breast cancer patients, and vice-versa.

Module IV: Signaling and network analysis

The significance of this module lies in connecting the queried gene with the whole transcriptome and overlaying of different kinds of meta-analysis onto each other. E2S takes the queried gene and finds its correlation with each gene in the transcriptome for all datasets found in untreated *survival cohort* and then combines these correlations obtained from each dataset to generate a meta-correlation statistics for gene of interest. Ranking of meta-correlation scores helps find the closely associated or distantly co-expressed genes. By default top 200 correlated (positive or negative) genes are visualized but this can be extended to whole transcriptome. In addition, this module enables comparison of co-expression profiles with different meta-analysis scores in the database, i.e., late E2 treatment meta-differential expression profile, ER-/ER+ meta-

differential expression profile, meta-survival analysis of untreated patients, meta-survival analysis of tamoxifen treated patients and the PS-scores, (i.e., difference in logFC of cycloheximide+estrogen and E2 treated profiles). These layers of different effect sizes can be chosen for visualization on the x-axis or as a color code on the top of x or y-axes. By default, ER-/ER+ effect-size is shown on the x-axis with a mHRs of untreated patient indicated as color codes. Meta-correlation score with queried gene is fixed on the y-axis.

Second sub-module is based on functional enrichment of the co-expression network. Functional enrichment can be performed using multiple options, e.g., KEGG, Reactome pathways and Gene Ontology annotations (biological processes, cellular component or molecular function). Functional annotation is further divided in two categories: positively co-expressed genes (meta-correlation score>0) or negatively co-expressed genes (meta-correlation score<0). E2S offers the user to choose a number of top correlated genes in the range of 50-500 for positive or negative co-expressed genes before performing the functional annotation. By default, enrichment is obtained for the top 100 co-expressed genes.

Third sub-module combines co-expression and interaction network by retrieving protein-protein STRING interactions of top co-expressed genes (irrespective of direction). This module can also be customized for the top 50-500 genes. Retrieved interactions are shown by edges while nodes represent genes. Color of node provides the direction of correlation direction with queried gene (green: negative meta-correlation score; maroon: positive meta-correlation score).

Step by step tutorial of E2S is provided in Appendix III by a case study of a well-known ER+ breast cancer biomarker (GATA3).

4.4.16.2. Features of E2S

1) Multiple probesets for one gene

A well-known problem in dealing with microarray data is many-to-many relationship in gene-probesets relationship. Jetset has provided a way to score each probeset and

identify single best performing probeset for each gene [381]. While entering the gene, E2S displays all possible probesets for the queried gene and user can select any particular probeset he/she is interested in or the jetset best probeset, if preferred.

2) Download data and plots

For all the modules in E2S, user can download plots for each analysis along with numerical data/statistical analysis results as well. This feature is especially useful for co-expression profile, which enables the user to download different meta-analysis scores including prognostic importance of gene, its differential expression in ER-/ER+ subtypes, and the E2-induced effect on co-expression profile.

3) Visual graphics

Graphics of E2S are customized to provide the best presentation. Forest plots, which are widely used plots for presentation of meta-analysis statistics, show most of the meta-analysis results in E2S. Co-expression profile of any gene is provided on y-axis and can be over-layered with two different layers (on x-axis or color-code) with user selected analysis profiles. Grey color of gene indicates a non-significant regulation (in either of the analysis on x or y axis). Respective scores can be visualized on x and y-axes while gene-names are provided as a tool-tip on each dot. Visualization of interaction network allows the user to zoom in or out and move any particular node in network.

4) Heterogeneity analysis

Combining data from different studies conducted in different labs might be challenging due to heterogeneity in the data. Hence, heterogeneity analysis is important in analyzing large number of studies. Funnel plot facilitates the visualization of heterogeneity in effect-size along with a standard error of the gene in all datasets used in analysis. One source of this heterogeneity can be publication bias, i.e., only selecting datasets with favorable outcome. Publication bias is evaluated by Egger's test. In case of significant

publication bias ($p<0.05$), meta-analysis results can be misleading. On the other hand, non-significant publication bias ($p>0.05$) indicates that heterogeneity in data might not be related to publication bias. However, funnel plot and Eggers's test can only be applied in case of more than 10 studies and thus can be provided for *ER cohort* having 31 studies. For this reason, no heterogeneity source is provided for the *E2 cohort* or *survival cohort*.

5) Error messages

E2S displays meaningful error messages to help its user with source/possibility of problems in retrieving results.

6) Wait messages

In case of long processing time, E2S displays waiting sign to indicate that command has been taken and E2S is preparing results.

7) Forest plot visualization

Forest plot summarizes meta-analysis results and following information:

1) GEO accession number and in case of E2 treatment profile, treatment time is also shown.
2) Numbers of samples in control or treatment group of any particular study are shown. As analysis in *survival cohort* is based on auto-best cut-off generation, the number of samples in control and treatment is not provided for the survival related forest plots.
3) Treatment effect is represented by standardized mean difference in numerical form and also shown on the forest plot with squares in front of each study.
4) Standard error of treatment effect for each study.
5) 95% confidence intervals in numerical form along with visual representation by a horizontal line around square (SMD) are shown.
6) Weights assigned to each study on the basis of FEM or REM are provided.

7) After individual statistics of each study, meta-analysis results based on FEM and REM are provided in two separate lines: z-value along with p-value, total number of samples in the treatment or control groups written beneath the number of samples of each study, SMD along with 95% confidence interval beneath the respective results of individual study, visual presentation of effect-size by diamond shape with area of the diamond representing the 95% confidence interval.
8) Limits of prediction interval are numerically written and also shown by a red line.
9) Measures of heterogeneity analysis are provided on the last line with I-squared, Tau-squared and p-values. P-value<0.05 makes REM statistics more applicable while p-value>0.05 leads user towards FEM score evaluation.

4.4.16.3. E2S Home page

Home page of E2S provides summary of the data in different cohorts of E2S including number of samples, platform and source of data (Figure 4.39). No input is taken on the page but 'Go to analysis' button or analysis tab on the top helps navigation to input page. Input page notifies the user to enter a query gene. Only valid human gene symbols are accepted in E2S at the moment. After entering the query gene, user can navigate to any of the four modules described above by clicking the respective button or selecting the tabs.

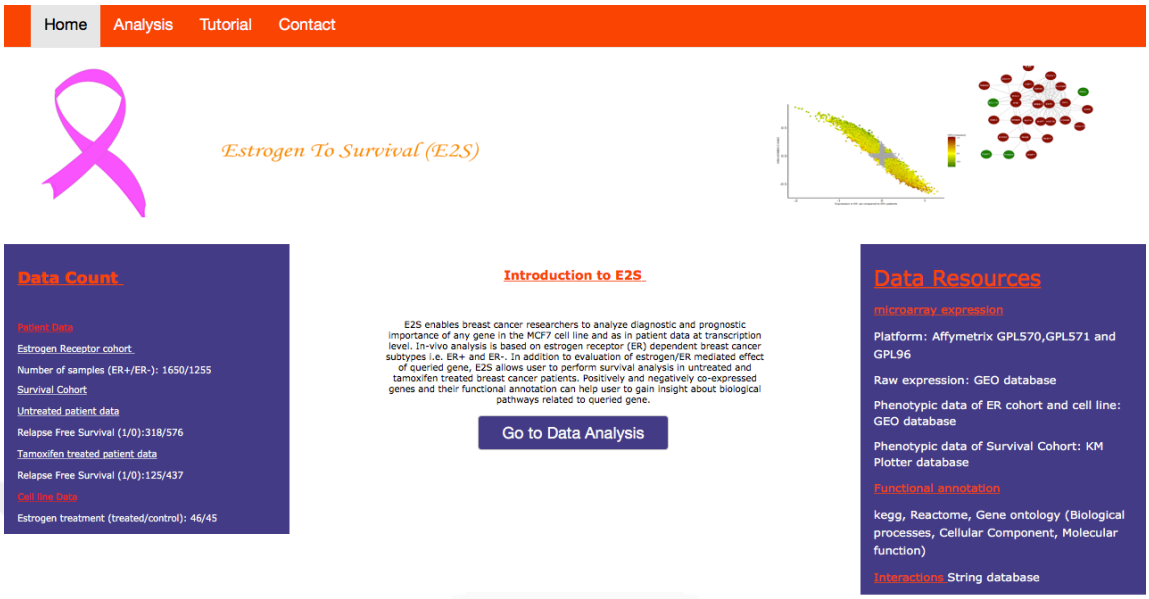


Figure 4.39: Home page of E2S.

Step by step tutorial of E2S is provided in Appendix III by a case study of a well-known ER+ breast cancer biomarker (GATA3).

4.5. Discussion

The strength of this chapter lies in two major points 1) identification of CHRNA5 as a breast cancer biomarker with respect to differentiation and E2 signaling; and 2) development of a web-tool for study of any gene in the context of E2 signaling and its prognostic importance in breast cancer. Regulation of CHRNA5 expression in breast cancer is not well studied before. However, two previous studies from our lab have focused on 1) functional pathways related to CHRNA5 depletion in breast cancer; and 2) effect of E2 on CHRNA5 isoform expression in dose and time- dependent manners. The main subject of this chapter thus has been also functional involvement and CHRNA5 regulation in E2-mediated signaling but from a comparative transcriptomics perspective and meta-analyses. 1) Two aspects of functional involvement, i.e., EMT and differentiation, have not been covered before and are thus explored in this study. 2) Besides, induction by E2, association of CHRNA5 with multiple aspects of E2 signaling (time-dependent meta-analysis, primary/secondary E2 target modulation), differential regulation by ER status, prognostic significance and co-expression networks also have been elaborated.

The study presented in this thesis successfully associates CHRNA5 with ER status and implicates its importance in prognosis of ER+ tumors. Moreover, CHRNA5 is one of the secondary signaling molecules downstream of E2-ESR1 induction and is part of a larger co-expression module with significant prognostic effects based on meta-analysis. Each question regarding E2 signaling and its prognostic value has been answered using different and large microarray cohorts**.** In addition to understanding of CHRNA5, an online user-friendly tool has been developed helping users study any gene of interest in the collected cohorts. One potential application of E2S lies in the identification of secondary E2-induced target enrichment in ER- breast cancer patients and their association with the worse prognosis. This finding validates the already described study that indirect targets of E2 are upregulated in ER- breast cancer [331], which also provides confidence on our predicted secondary E2 induced targets. In a previously conducted meta-analysis of breast cancer, it has been shown that MYC is highly upregulated in ER- basal breast cancer and might be responsible for mimicking ER

mediated signaling in its absence. Furthermore, this study has also shown upregulation of E2F family members in basal breast cancer, which again confirms findings in this thesis. In addition, our study for the first time provides the list of secondary E2-induced targets; and also indicates the prognostic value associated with each of these secondary E2-induced targets to conclude that these secondary targets of E2 signaling might be associated with worse prognosis of ER- patients as a whole.

In the literature, there are examples in which meta-analysis of a large number breast cancer datasets have been used to address different aspects of a gene's biomarker capacity in breast cancer. Two of these genes are DACH1 and CD44; Xu et al. (2017) have performed a meta-analysis based on 19 different breast cancer datasets using REM in case of significant heterogeneity otherwise using FEM and calculated the HRs [453]. Our collection of datasets and partitioning differs from Xu's in terms of treatment specificity and survival cut-off strategy. Dataset cohort they have used was larger while no treatment information was provided and the cohort was based on multiple Affymetrix as well as Illumina platforms. In my collection, I have conducted survival analysis by categorizing and specifying untreated (CHRNA5 case study and E2S) and tamoxifen (E2S) treated patient, separately, from KM plotter, which allows comparisons in between. To reduce platform-based heterogeneity, I have limited the dataset search to two majorly used Affymetrix platforms. In addition, the meta-analyses of different cohorts were also compiled in a database (E2S) extending the meta-analysis to all transcriptome other than the case study of CHRNA5. Another gene based meta-analysis study on CD44 expression used 23 different GEO datasets to address the importance of CD44 in breast cancer clinical parameters and survival [454]. The analysis used in Xu et al. (2016) study differed from the present one as it was specific to few genes defining breast cancer subtypes, EMT, cancer stem cells or survival [454]. Meta-analysis for the identification of CD44 was conducted by pooling of Odds ratio however I have used HR, effect-size, or meta-correlation score depending on the different underlying question.

Meta-analysis of primary epithelial and non-epithelial differentiation related datasets also is lacking in the literature. The present study thus allows the collection and

processing of epithelial differentiation datasets from normal tissue to which any transcriptional signature can be mapped onto and tested for its association with differentiation processes. In the literature, a transcriptomics signature for mesenchymal stem cells (MSC) was obtained through meta-analysis [455]. MSC signature was obtained using multiple arrays from GEO after careful selection and batch correction using *frozenRMA* and hence the data were combined and analyzed using Pearson correlation coefficients. Epithelial differentiation signature obtained in this thesis is different from the MSC signature study in terms of a) datasets used; and b) the use of different meta-analysis techniques (e.g., combining datasets vs. analyzing each dataset separately and also by combining effect sizes, respectively). A meta-analysis of cancer stem cells (CSC) in hepatocellular carcinoma (HCC) and differentiation based on fixed- and random-effect models also existed [456] yet it was not comprehensive enough extending to multiple tissues of origin. In addition, Odds ratio was used to find the association between few selected genes and differentiation [176]. One of the most relevant meta-analysis studies, which has been done in breast tissue [457], was based on interspecies analysis of mouse and human datasets deriving signatures for mammary cell populations (e.g., luminal progenitor, mammary stem cell, stromal cell populations). This was rather a signature-based analysis and did not employ a classical meta-analysis approach or multiple tissues to address their commonalities and divergences. However, genes signature available from this study can help compare the meta-differentiation signatures obtained together with CHRNA5 RNAi signature. Meta-analysis of EGF (a regulator of differentiation) in keratinocytes has also been performed in connection with other tissues; this study was rather EGF specific and included effect of external chemical stimulus [458]. However, in this chapter, I have focused on differentiation datasets without any external stimulus. Instead of using statistical meta-analysis strategies, authors also have focused on evaluation of shared and diverged functionalities. Meta-analyses also helped compare EMT in normal breast and identified that mesenchymal cells partially resembled myoepithelial profile and claudin-low breast cancer cells [459]. This study also could be relevant for understanding the mammary differentiation effects of CHRNA5 RNAi and can be used for future comparisons. Meta-analysis also was used for addressing the relationship between claudin low tumors, EMT and cyclin D1 using

more than thousand expression arrays from 6 datasets in breast cancer [460]. However, to the best of my knowledge, the differentiation signature meta-analysis performed in my dissertation study, is one of the most highly comprehensive, encompassing multiple epithelial tissues, thus a pioneering study. Furthermore, it is the first time that transcriptomics signature of CHRNA5 RNAi in MCF7 cells has been associated with a large collection of epithelial differentiation processes and demonstrated a strong correlation with differentiation meta-signature. This study complements previous findings on the TP63 connection with CHRNA5 RNAi [408] and leads to multiple biomarker genes to test in connection with epithelial differentiation. Interesting finding is that CHRNA5 has not emerged as a regulator for cartilage, adipose or neural differentiation suggesting that there could be another molecule mimicking CHRNA5 since the CHRNA5 signature is highly correlated with neural differentiation in the absence of a change in CHRNA5. Future studies can focus on this hypothesis.

One of the first attempts in meta-analysis of E2-induced effects in which early or late effects of E2 on breast cancer cell lines *in-vitro* has also been investigated in my thesis [329, 460]. A study performed in 2009 used E2 response as a case study while introducing a software called GEMS (Gene Expression Meta-Signatures) [328]. This study has been rather limited in the number of datasets included however they have used a proper meta-analysis approach, i.e., weighted z-score. They also provided an online database to evaluate the effect-size of any given gene in E2 treated samples. GEMS have now become part of a broader database 'transcriptomine' including wide range of nuclear receptors [461]. The database currently covers a range of nuclear receptors with respect to pathways, species, tissue and receptor. Comparatively, E2S is specific for E2 signaling and prognosis. However, what make E2S different is the visualization as well as overlaying of meta-analysis results. Upon entering a query gene, Transcriptomine also visualizes the logFC in individual dataset yet with limited information about dataset. However, E2S shows a forest plot summarizing both individual and meta-analysis effects along with heterogeneity scores reporting both fixed/random effect model scores. Studies are labeled with GSE numbers, which makes it easier to track dataset from GEO database. Another following study has also used early and late E2 treatment datasets (9 datasets in total and shared with GEMS) and RankProd for meta-analysis; however,

there is no online platform to evaluate any human gene in their meta-analysis profile. Both previously reported studies have categorized 3-4 hour exposure in early E2 treatment group while 24-hour exposure to the late E2 treatment. My study has used a wider range of datasets and categorized 2-4 hours exposure in early E2 treatment group and 6-48 hours in late E2 treatment group. Although E2S shares datasets with both of these previous meta-analysis studies, it includes additional studies, as well. Previously described meta-analyses have used normalized data (by RMA or GCRMA) from GEO. However, I have normalized each dataset by RMA to avoid any bias.

Development of a Shiny tool for analysis of candidate genes in response to E2 as well as the genes' prognostic values in breast cancer where additional datasets, e.g., secondary signaling scores (PS-score) can be overlaid, has been performed for the first time in the literature to my knowledge. Future studies will also add the functionality of overlaying any user introduced transcriptomics signature onto the meta-effect size, meta-correlation, as well as mHRs that are already present in the E2S tool.

There are also other more comprehensive and general meta-analysis approaches for gene expression analysis. One of these is NetworkAnalyst [330]. E2S is different from the NetworkAnalyst meta-analysis tool in a way that E2S is specific and contains already processed breast cancer data. NetworkAnalyst allows you to upload the data and perform meta-analysis using different strategies but it does not remove duplicate samples found in different studies as well as quality control has not been the focus of database. Strength of E2S lies in manual processing of the samples to remove duplicates, averaging of replicates, removing samples with conflicting phenotypic data. Its aim is allowing the user to test a gene in specifically breast cancer using four different independent cohorts which are independently meta-analyzed: one for evaluation of E2 treatment effect; second for estimating gene's effect size in ER+ and ER- tumor populations; and another one for estimating gene's meta-hazard ratio in untreated or tamoxifen treated patients; and the forth one is for PS-scores to overlay. E2S is unique in that it allows comparison of two separate data cohorts thus while discovering, it validates the finding. The underlying structure of E2S can be applied to other cancers where two independent cohorts need to be compared and, as a third input, an additional dataset can be associated

with these effect sizes and HRs. Furthermore, as in the case of NetworkAnalyst, E2S allows for high quality visualization using Cytoscape to color and represent the correlations between genes.

In both of the meta-analysis tools (NetwrokAnalyst and Transcriptomine) reviewed, it is possible to evaluate differential expression but the survival outcome associated with a gene and genome-wide co-expression profile cannot be retrieved. Online web-tools to evaluate survival information are also available. One of the most commonly used online tool is KM plotter. KM plotter allows user to query a gene for its prognostic importance in breaat, lung, gastric and ovarian cancer [385]. E2S has also obtained phenotypic data of untreated and tamoxifen treated samples from KM plotter database. E2S consists of rather smaller cohorts as compared to KM plotter database but its uniqueness lies in proper meta-survival analysis strategy rather than direct merging of data. Besides, instead of providing HR for one queried gene, it is possible to retrieve HRs of co-expressed transcriptome for scatterplots. Hence, E2S is specific to E2-mediated signaling in breast cancer and is a platform to evaluate differential expression, prognostic importance, co-expression profile, functional and protein interaction information for any gene and its co-expressed neighbors in breast cancer patient/MCF7 cell line data. Future studies will focus on identifying novel candidate genes that exhibit poor or good prognosis as well as significant effect sizes between ER+ and ER- tumors, as in the case of CHRNA5. E2S is a powerful tool to identify candidate biomarkers for future functional validation as well as reliable identification of candidates.

4.6. Future perspectives

- CRHNA5 downregulation study in ER+ breast cancer cell line has shown regulation of incomplete differentiation processes toward myoepithelial cell type. To clearly understand the process, fluorescent labeling of luminal (ACTA2) and myoepithelial (MUC1) antibodies can provide the extent of regulation at protein level.

- This thesis has validated that CHRNA5 is an E2-mediated target and is secondary in nature. Using the P2S scores, I also predicted list of secondary E2-ESR1 targets. Wet-lab validation of few secondary E2 induced targets along with CHRNA5 can validate this finding. Besides, in this study, it is shown that secondary E2 targets are enriched in ER- patients and are associated with worse prognosis. Comparative experiments on ER+ and ER- cell lines by knockdown of top ranked secondary E2 targets can be verified by evaluation of proliferation. Overlap of these secondary E2 targets with known signatures of MAMMAPRINT and PAM50 can be interesting to identify known/novel biomarkers from the list.

- Because of limited number of samples in ER- patient collection of *survival cohort*, no significant prognostic importance could be observed for CHRNA5. Repetition of analysis by collecting RNAseq data might provide more insight into subtype-dependent prognostic value of CHRNA5.

- An online tool E2S has also been presented in this study; E2S is designed for E2-mediated signaling in breast cancer and along with patient data, it also includes E2 treatment expression profiles in MCF7 (ER+ cell line). However, the addition of data on multiple breast cancer ER+ and ER- cell lines can be helpful, as E2 has been shown to exhibit signaling mechanisms dependent or independent of ER. Moreover, data on tamoxifen resistant MCF7 can help identify novel targets for tamoxifen resistance.

- PS-score calculation is based on cycloheximide treatment in the presence or absence of E2. However there are several other protein synthesis inhibitors, e.g.,

anisomycin, pactamycin, tylocrebrine [462]. With respect to E2 treatment, expression data of these inhibitors should be searched and incorporated in E2S. Addition of ChIPseq and RNAseq datasets in different contexts of E2-mediated signaling and/or ESR1 regulation could also broaden the perspective of database.

- Execution time of String network and pathway annotation module is 2-5 minutes depending on the speed of Internet. An alternative and faster strategy to make this execution time shorter can be identified.

- Co-expression analysis allows user to analyze different scores on one plot (using y-axis or color coding). Module will be generalized by allowing user to upload any of their own score (logFC, effect-size, mHR) and analyze its correlation with meta-scores present in the database (ER-/ER+ differential expression meta-analysis, E2 meta-analysis, PS-score, meta-hazard ratio of untreated or tamoxifen treated patient data).

- Upon completion, E2S will be made online to facilitate breast cancer research.

Chapter 5

Concordant/discordant transcriptomics of mineralocorticoid and glucocorticoid receptors in breast cancer and development of DualExpBC

5.1. Introduction

5.1.1. Nuclear Receptors

Nuclear receptors (NRs) represent a transcription factors superfamily, which have evolutionary similar DNA binding sites [463]. Binding of receptor to its specific ligand initiates a downstream signaling cascade. These ligands include estrogen, progesterone, retinoic acid, oxysterols, and thyroid hormone [464, 465]. Receptor-ligand binding does not take place at the cell surface, instead after directly crossing of the ligand through membrane inside the cell, it interacts with nuclear receptors to regulate a transcriptional program involved in diverse biological processes, e.g., cell development, proliferation, and metabolism [466].

Upon ligand binding, a nuclear receptor can mediate non-genomic actions via interaction with other intracellular proteins (e.g., kinases present in cytoplasm). Genomic actions are mediated by modification of chromatin packaging [467]. In the absence of a proper ligand, nuclear receptor co-repressor (N-CoR) or silencing mediators may interact with nuclear receptors to repress downstream signaling [463]. One particular repressor event includes complex of nuclear receptor with HDAC (histone deacetylase complex) to suppress transcriptional process however, upon receptor activation it is replaced by activator complex of HAT (histone acetyl transferase) and chromatin remodeling complex which makes chromatin more accessible for transcriptional machinery to exhibit its action [467].

In total 48 nuclear receptors have been studied in human cells [468] and homology comparison of different nuclear receptors have categorized them in six subfamilies with varying number of receptors in each class [469]. Interestingly, few nuclear receptors have been identified with no known ligand mediating their activation and they are known as 'orphan nuclear receptors' [468].

5.1.2. Steroid receptors

Steroid receptor represents one important family of nuclear receptors, which mediates its biological actions via different kind of steroid hormones. Like all classes of nuclear receptors, steroid molecule diffuses into cells to interact with cognate receptor. Inactive receptor remains repressed by its interaction with heat shock proteins. Once the ligand binds, receptor dissociates from heat shock protein complex and is translocated to nucleus [470]. Features which make steroid receptors different from other nuclear receptor includes: 1) use of steroids, which are different cholesterol derived molecules, as ligands [471]; 2) unlike other nuclear receptors, steroid receptors, in their inactive form sequester in cytoplasm instead of nucleus [472, 473]; 3) non-steroid hormone receptors generally attaches to DNA response elements in the form of a heterodimer however steroid receptors assemble in form of homo-dimers to recognize hormone response element [473, 474]; 4) steroid and non-steroid receptors recognize differently structured hormone response elements. Steroid receptor specific hormone response elements consist of 2 palindromic sequences separated by 3 nucleotides however non-steroid receptors recognize direct repeat of two half sites [470].

5.1.2.1. Steroid receptors in breast cancer

Steroid receptors include estrogen, progesterone, glucocorticoid, mineralocorticoid, retinoid, vitamin D, and estrogen-related receptors. Both progesterone and estrogen receptors are important regulators of the breast cancer and are frequently used to classify breast cancer in subtypes [475]. Both receptors are expressed in luminal cells of the breast which are widely distributed among the peripheral alveoli and intra-lobular ducts

[476]. Breast cancer subtypes, negative for both estrogen and progesterone receptor are generally associated with worse prognosis [477].

ER gene exists in two forms (ER-alpha and ER-beta); it has been shown that ER-alpha contributes to tumorigenicity in breast. On the other hand, ER-beta is expressed in normal breast tissue and its loss in tumor provides proliferative potential to cells [478]. Progesterone receptor is also expressed in two forms, i.e., PRA and PRB. Both forms are equally expressed in normal breast however PRA dominates during carcinogenesis [479].

Retinoid X receptors and retinoic acid receptors also belong to steroid-thyroid family. Each of these receptors have three subtypes known as alpha, beta and gamma [480, 481]; and it has been shown that loss of RAR-beta increases tumorigenicity of breast cancer [480]. RAR-alpha gene is also an estrogen responsive gene making a transcriptional complex with ER to mediate proliferation [482].

Polymorphisms in vitamin D receptor has also been linked with risk of developing breast cancer [483, 484]. Breast tumor patients with vitamin D receptor show better survival and anti-proliferative potential [485]. Androgen receptor (NR3C4) is also used for classification of breast cancer and is widely expressed, even more than estrogen and progesterone receptor [486]. ER-patients expressing androgen receptors exhibit ER+ like phenotype while overexpression of AR enhances resistance to tamoxifen treatment [487, 488].

Steroids mediate their action by interaction with steroid receptors, which is facilitated by co-activator proteins. AIB1, steroid receptor co-activator protein has been upregulated in a proportion of breast cancer besides, its expression correlates negatively with ER and PR [489]. Mineralocorticoid and glucocorticoid receptors (MR and GR, respectively) are comparatively less studies and are described in next section.

5.1.2.2. Mineralocorticoid (MR) and Glucocorticoid (GR) receptors

MR (NR3C2) and GR (NR3C1) belong to same steroid class of nuclear receptors and have highly homologous structures [490]. MR and GR interact within themselves or with each other to function in the form of homo- or hetero-dimers. Both act as transcription factors upon ligand binding and may compete for the same ligands although with different affinities [491]. In addition, studies have shown that both transcription factors recognize the same DNA response element to initiate transcriptional program [492]. MR and GR are also involved in several rapid non-genomic/extra-nuclear pathways. Although both genes share several functionalities, separate MR and GR mediated effects also exist. Primary ligand (i.e., aldosterone or cortisol, respectively) is involved in regulation of diverse MR/GR specific actions [491]. MR regulates homeostasis by controlling ion and fluid transport, injury response, excitation in neurons and muscles; GR, on the other hand, regulates energy homeostasis, e.g., gluconeogenesis, and mediates inflammation and stress response [491].

DNA binding domain of steroid receptors including androgen receptor (AR), progesterone receptor (PR), MR and GR is highly conserved. For this reason, DNA response element of not only MR and GR, but of multiple steroid receptors is mostly overlapping. Recent studies have revealed an interesting finding that transcriptional regulation mediated by MR-GR heterodimer is different from homo-dimers of MR/GR [493]. Ratio of MR to GR is important to mediate cell adaptation in changing extra-cellular environment, e.g., response to stress. Both genes are expressed in multiple cell types and as they interact to undergo biological activities; both are usually found in same cells. One study has also shown that MR compensates for the loss of GR during mammary development [490].

5.1.2.3. Regulation of MR in cancer

Using microarray analysis, NR3C2/MR has also been found to be downregulated in gastric cancer [494]. Several recent studies have shown tumor suppressor role of MR in different cancer types. Expression of MR is downregulated in 81% of hepatocellular

carcinoma and associated with poor prognosis. Overexpression of MR results in growth suppression of hepatocellular carcinoma. MR exhibits its anti-carcinogenic effect by suppressing Warburg effect, which is main source of energy for cancer cells [495].

In line with previous findings, studies in lung cancer also put forwarded MR as a biomarker and showed that in lung cancer also, a reduced MR expression is associated with poor prognosis [496]. Tumor suppressor role of MR in colorectal cancer is mediated by inhibition of VEGFA, which is an important angiogenesis marker [497]. Classification of clear cell renal cell carcinoma (ccRCC) in early as well as late stages can be performed using microarray data and NR3C2 performs best for classification with 71.12% accuracy and is downregulated in late stage of ccRCC [498]. Besides, research has also focused on NR3C2 regulation in non-epithelial cells. Array based comparative genomic hybridization results have identified gain/loss of several chromosomal regions in acute lymphoblastic leukemia and loss of 4q31.23, where NR3C2 is located, in 17% of cases validating MR downregulation in non-epithelial cells [37]. Reduced expression of NR3C2 indicates poor prognosis also in pancreatic ductal adenocarcinoma. Overexpression of NR3C2 suppresses EMT and also resulting in enhanced sensitivity to chemotherapy [38].

A recent study has shown that both MR and GR cross talk with progesterone receptor to induce focal adhesion and suppress tumor growth in breast cancer cell lines [28]. Another study based prognostic significance of nuclear receptor has shown that MR expression is downregulated in breast cancer in comparison to normal samples and was able to predict metastasis-free survival in tamoxifen treated patients [39]. Downregulation of NR3C2 in breast cancer was also identified using gene expression data [40]. Although there are several studies showing expression of MR in breast normal and tumors there is no meta-analysis of expression datasets to evaluate the effect size between tumor and normal breast tissue.

5.1.2.4. Glucocorticoid receptor in cancer

Consistent with MR's tumor suppressor role, GR receptor also exhibits anti-cancer potential in multiple cancers, including skin cancer [497]. Anti-proliferative mechanism mediated by GR is largely similar to anti-inflammatory processes in skin cancer [497]. A recent study also has covered multiple cancer types and shown the tumor suppressor role of GR in liver, esophageal, breast, lung, prostate, ovarian and colon cancers [499]. GR mediated effect, however, is tissue dependent with no significant expression difference in the bone and gastric cancer. In contrary, upregulation of GR has been observed in renal and pancreatic cancer [499]. Furthermore, the same study using mouse experiments has identified that haplo-insufficiency of GR is associated with increasing chromosomal aberrations and DNA damage in cancer cells [499]. In an independent study, tumor suppressive role of GR by reducing tumor proliferation has been shown in prostate cancer [500].

GR, as compared to MR, is more studied in breast cancer because of its close association with important breast cancer regulator, i.e., progesterone receptor (PR). PR is well known to play a role in mammary development and pregnancy. Both PR and GR are also known to regulate/inhibit breast cancer growth [490] and share comparable number of targets genes [501]. Expression analysis reveals that expression of NR3C1/GR declines upon carcinogenesis in breast and changes its location from nucleus to cytoplasm [502]. Another study focusing on breast cancer subtypes has shown that GR mediated actions are subtype specific. Higher expression of GR indicates good prognosis in ER+ but worse prognosis in ER- patients [503]. In a subsequent study, multiple ER- breast cancer cell lines have been studied and protective effects of GR in ER- cancer could be inhibited by Mifepristone (MR and GR antagonist) treatment, resulting in caspase-3 and PARP mediated apoptosis. Anti-apoptotic effects of GR are mediated by its well-known direct target genes, i.e., SGK1 and MKP1/DUSP1 [504]. In a recent attempt to identify important players in cancer associated fatigue, the increased NF-KB and decreased expression of GR responsive genes are found to be more prevalent in fatigued breast cancer survivors. As GR is involved in inhibition of NF-KB, fatigue related to breast cancer can largely be mediated by glucocorticoid receptor

signaling [505]. As in the case of MR/NR3C2, there is no meta-analysis of tumor suppressor role of GR/NR3C1 in breast cancer normal and tumor tissues. Such an analysis will reveal heterogeneity as well as effect size of GR in breast cancer pathogenesis.

5.1.3. Paired and unpaired expression data analysis in cancer

Paired and unpaired data differ in their origin of the control samples: If both the control and treatment samples are taken from same individual then paired, otherwise unpaired data assumptions are applicable [506]. Few interesting comparative paired and unpaired studies have been reported in literature using expression data from cancers and are described in the following lines. For example, microarray data helped clustering of prostate cancer subgroups using paired and unpaired datasets upon filtering of low expressed genes or flooring low expressed genes to a certain value. Filtered (by removing low expressed genes) analysis resulted in the same prediction for either paired or unpaired data. Flooring approach slightly increased prediction accuracy as compared to filtered analysis; and this accuracy was even more in paired data cohort [507]. Another study was used for normal/renal cancer differential expression analysis using the paired as well as unpaired t-tests and showed that the unpaired t-test was more powerful in identification of differentially expressed genes [508]. Unpaired tests identified 622 more significant genes than the paired data analysis. Hence the authors concluded that tumor/normal heterogeneity outstands inter-patient heterogeneity making unpaired tests more accurate [508]. miRNA expression analysis of breast cancer also revealed varying capability of paired and unpaired data for extraction of significantly modulated expression profiles. 10 microRNAs were found common between the paired (total 35 differentially expressed) and the unpaired (total 98 differentially expressed miRNAs) significantly regulated profile [509]. These findings suggest that both paired and unpaired analyses can provide common as well as unique genes/gene sets. The reasons behind this could be that paired data allows for incorporation of microenvironment thus takes into account the individual differences in a gene's expression thus could separate cancer specific effects from the common variability. Both

types of analyses however can be useful since unpaired tests will have higher statistical power at the expense of sample heterogeneity. However, there is no tool for comprehensively analyzing breast cancer tumor/normal datasets in a paired and unpaired fashion. Furthermore, meta-analysis of normal-tumor datasets in breast cancer can benefit from incorporation of paired and unpaired options and comparisons in between to identify genes with potential inter-individual variability and micro-environmental influences.

5.1.4. Deregulation of gene pairs in cancer

Gene pair correlation in normal samples and deregulation in cancer is comparatively less studied field. A very interesting study has focused on use of gene pairs instead of raw values by taking sum or difference of original expression values. They have revealed that using gene pair in place of single gene analysis improves the classification of cancer [510]. Similarly another study has used gene-pairs to classify normal-tumor samples and they stated that there are some cancer related signals, which can only be explored by gene-pair analysis [511]. Anglani et. al. has presented a methodology in which they compared co-expression network in normal and tumor samples. They overlaid differential co-expression with differential connectivity to identify gene pairs which are correlated in normal samples but have lost connectivity in cancer [512]. However, to the best of my knowledge, there is no online tool/database, which can help identification of deregulated gene pairs by using the difference in expression between the gene pairs as a measure of expression variability between groups tested (e.g., cancer vs. normal) across multiple studies of breast cancer.

5.2. Aims and rationale

From the literature review, I have found that MR is comparatively less studied in breast cancer and MR/GR co-expression has not been discussed in detail. Previous MS thesis study by Bircan Coban from Konu lab has shown the correlation of MR and GR in breast cancer at the protein level (2016). She has shown that both MR and GR are expressed in several breast cancer cell lines, possibly in a subtype dependent manner and could be expressed dissimilarly in certain breast cancer cell lines, e.g., MCF7, CAL51. However, there is lack of extensive expression analysis in larger cohorts focusing MR/GR differential and co-regulation in breast cancer.

Accordingly, the aims of this study included:

1) Collection and processing of large normal/ tumor breast expression cohorts with categorization of paired and unpaired samples/datasets.
2) Determination of differential MR and GR expression in breast cancer as compared to normal tissue in paired and unpaired breast cancer cohorts.
3) Identification of transcriptome wide differences in normal/tumor meta-analysis signature in paired and unpaired cohorts, separately and together.
4) Investigation of shared and MR/GR specific co-expression profiles in breast cancer and identification of underlying pathways.
5) Testing deregulation of MR/GR expression in breast cancer as compared to normal tissues in the paired data.
6) Development of a web-tool from the collected paired and unpaired breast cancer cohort, which can facilitate meta-differential and co-expression analysis in normal/breast tumor, i.e., Dual gene Expression in Breast Cancer (DualExpBC).

5.3. Materials and methods

5.3.1. Breast normal/tumor expression cohort collection

Microarray expression data of breast cancer patients were obtained from GEO database [171]. Datasets having both normal and breast tumor samples with raw data available were downloaded. Quality control using outlier detection of five-quality measures and duplicate sample removal-using *dupchecker* were performed as mentioned in the *ER* and *survival cohort* pre-processing of Chapter 4 (section 4.3.10 and 4.3.11). Replicate samples obtained from the same patient were averaged prior to analysis. Datasets having less than three samples in normal or tumor groups were discarded. Studies based on stromal origin were not included and datasets included were based on tumor tissues, biopsy or epithelial samples.

5.3.2. Splitting in paired and unpaired cohorts

The collected normal/tumor cohorts consisted of several studies with known paired and unpaired data origins. Cohorts were split into *paired* and *unpaired subcohorts*. In case of datasets having both paired and unpaired samples (more than 3 samples in each group of both datasets), the data were split in two datasets, i.e., paired sub-dataset and unpaired sub-dataset. In case of GSE50567, which had 41 samples in total with 6 normal and 35 breast tumor profiles and in which 5 of these tumor samples were paired with normal samples, the following datasets were obtained upon splitting: unpaired data with 5 tumor/5 normal samples in the paired sub-dataset; and 30 tumors/1 normal in the unpaired sub-dataset. Unpaired sub-dataset fell below the defined threshold (<3 normal samples). In such cases, where splitting of dataset in paired and unpaired categories led to number of samples less then threshold value (<3 samples) in any of normal/tumor group, dataset was placed in sub-cohort where loss of samples was minimum. In the light of mentioned rule, 5 tumor samples of paired data were removed and GSE50567 data was added to the unpaired cohort resulting in 30 tumor and 6 normal samples. Datasets in breast normal/tumor cohort were described in Table 5.1.

Table 5.1 : Gene expression datasets included for meta-analysis.

Dataset	Platform	Meta-differential expression analysis		Meta-correlation analysis
		Normal/Tumors	Type of analysis	Number of samples
GSE3744	GPL570	7/40	unpaired	47
GSE5764	GPL570	10/10	paired	30
GSE10810	GPL570	26/26	paired	58
GSE45827	GPL570	11/130	unpaired	141
GSE42568	GPL570	17/104	unpaired	121
GSE61304	GPL570	4/58	unpaired	62
GSE50567	GPL570	6/29	unpaired	40
GSE31192	GPL570	6/8	unpaired	16
GSE29431	GPL570	12/54	unpaired	66
GSE29044	GPL570	18/18 18/24	paired unpaired	78
GSE22544	GPL570	4/11	unpaired	17
GSE21422	GPL570	5/14	unpaired	19
GSE17907	GPL570	4/51	unpaired	55
GSE16873	GPL96	12/12	paired	24
GSE15852	GPL96	43/43	paired	86

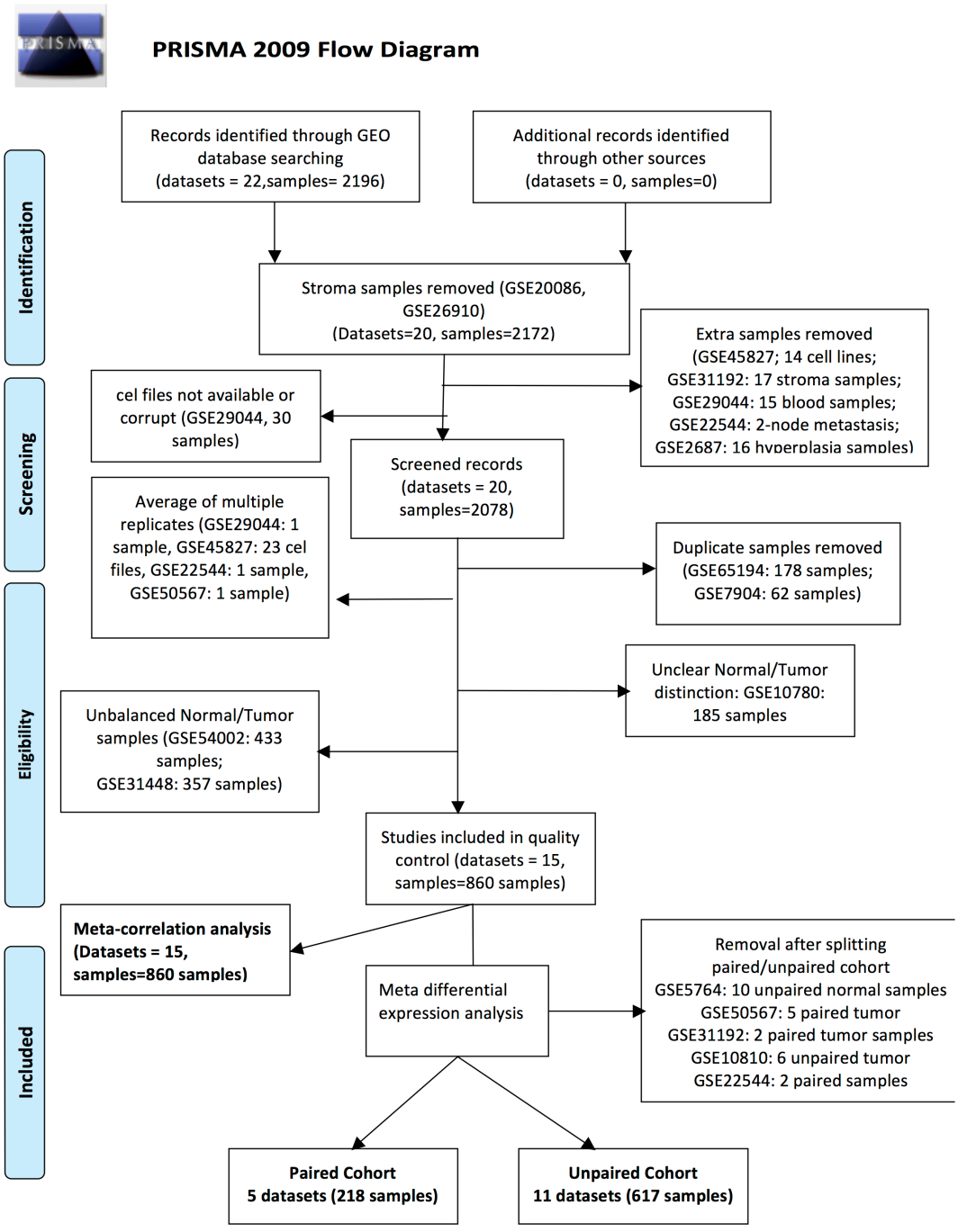


Figure 5.1: PRISMA guideline for paired/unpaired breast normal/tumor cohort meta-analysis.

5.3.3. Meta-differential expression analysis

PRISMA meta-analysis guideline of datasets (Table 5.1) has been shown in Figure 5.1. Separate meta-differential expression analysis was performed for 1) paired, 2) unpaired, and 3) combined meta-analysis of both sub-cohorts upon defining the paired and unpaired data. *MetaDE* package with seed value of 123 was used for meta-scores calculation [230]. *Forestplot* package was used to visualize meta-analysis results by forest plots [513].

5.3.4. Correlation analysis of normal/tumor cohort

Correlation analysis was performed without categorizing the data into paired and unpaired sub-cohorts. Meta-correlation analysis was performed separately for NR3C2 and NR3C1 using *metacor* package [145] as described in section 4.3.14.

5.3.5. MR and GR deregulation

MR and GR deregulation was individually studied in paired datasets. Deregulation in normal and tumor samples was calculated as a difference in MR and GR expression in normal samples as well as tumor samples and is described below. F-statistics [514] was used to evaluate difference in variance of both Dereg_N and Dereg_T

Dereg_N = MR-GR in normal samples

Dereg_T = MR-GR in tumor samples

5.3.6. DualExpBC

Similar to previously mentioned two databases (CompariZome (Chapter 2) and E2S (Chapter 4)), I have developed DualExpBC using shiny web-app [515] for visualization of normal and breast tumor meta-analysis results.

5.3.6.1. Search from stored data

Normalized gene expression data of all 16 datasets in *breast normal/tumor cohort* were stored in the database. Label of each sample, i.e., normal/tumor are also appended to normalized data before saving. Upon user-query, gene symbol was matched with the jetset file to retrieve best probeset(s) for the selected gene(s). The respective probeset was searched in normalized data of normal/tumor cohort.

5.3.6.2. Run-time executions

Run-time executions included meta-differential 1) expression analysis of queried gene and drawing of forest plot in all three categories (paired, unpaired and combined); 2) meta-correlation analysis between user entered two genes and visual representation using forest plot; 3) scatter plot of an individual dataset and scatter plot after retrieving normal/tumor labels of samples; and 4) deregulation graphics with F-statistics reported on the output page.

Deregulation of two input genes was calculated given the user selected input option, specific for deregulation module, i.e., 'without direction', 'gene1-gene2', 'gene2-gene1' as described below:

Gene1_N=expression of gene1 in normal samples

Gene1_T=expression of gene1 in tumor samples

Gene2_N=expression of gene2 in normal samples

Gene2_T=expression of gene2 in tumor samples

Without direction

Dereg_N=abs (Gene1_N-Gene2_N)

Dereg_T=abs (Gene1_T-Gene2_T)

Gene1-Gene2

Dereg_N=abs (Gene1_N-Gene2_N)

Dereg_T=abs (Gene1_T-Gene2_T)

Gene2-Gene1

Dereg_N=abs (Gene2_N-Gene1_N)

Dereg_T=abs (Gene2_T-Gene1_T)

Significance of variance difference between two groups, i.e., Dereg_N and Dereg_T were calculated using F-test in R.

5.4. Results

In this study, MR (NR3C2) and GR (NR3C1) co-expression and deregulation in breast cancer were identified by using meta-analysis of microarray expression data. A large transcriptome cohort of breast cancer/normal tissues was collected from GEO and processed as described in the materials and methods sections.

5.4.1. Meta-analysis of MR and GR differential expression in breast normal and cancer tissues

Initially I asked that whether both genes were differentially expressed between the normal and tumor tissues originating from breast and if there was any difference of this differential expression in between the paired and unpaired datasets. Collected and processed *breast cancer/normal cohort* was comprised of 16 distinct processed datasets with a total of 203 normal samples and 632 breast tumor patients (Table 5.1; PRISMA guideline: Figure 5.1). Meta-differential expression results of MR probeset (205259_AT) and GR probeset (216321_S_AT) were visualized by a forest plot of combined paired and unpaired data. Results showed that both MR and GR were highly downregulated in breast cancer as compared to normal breast tissue and their meta-analysis scores were comparable (MR: mu.hat=-1.97, FDR=9.7E-20; GR: mu.hat=-1.82, FDR=9.7E-20; Figure 5.2 and 5.3, respectively).

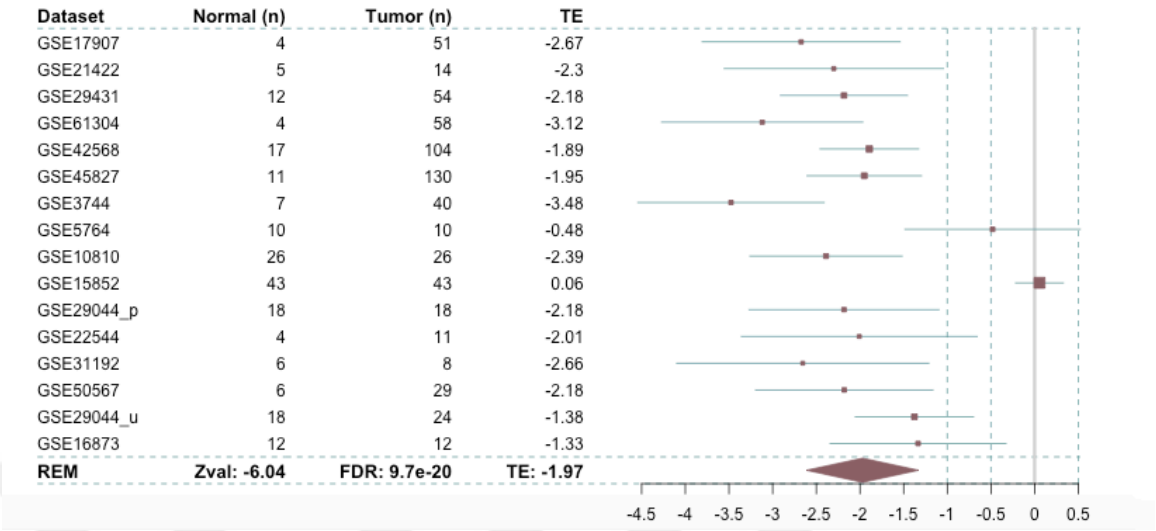


Figure 5.2: Differential expression of MR in breast normal/tumor meta-analysis.

Analysis is performed in combined paired and unpaired sub-cohort. GSE29044_p indicates paired samples of GSE29044 and GSE29044_u indicates unpaired samples analyzed separately. Number of samples (n) in normal and tumor are shown for each study . Effect-size>0 indicates upregulation in breast tumor while effect-size<0 indicate upregulation in normal. TE (Treatment effect) is calculated using SMD (Standardized Mean Difference). REM=Random Effect Model.

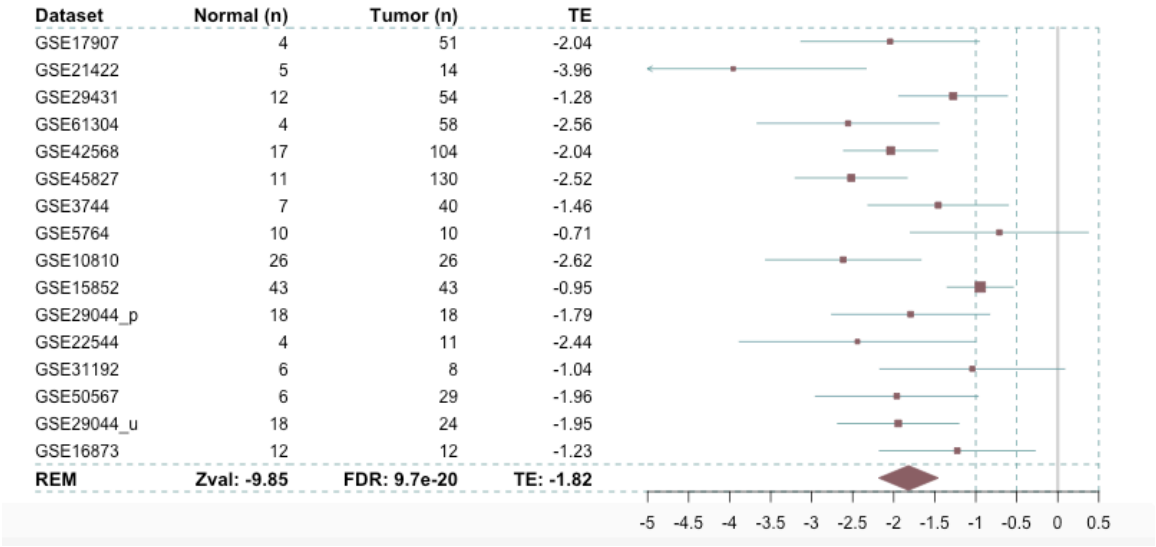


Figure 5.3: Differential expression of GR in breast normal tumor meta-analysis.

Analysis is performed in combined paired and unpaired sub-cohort. GSE29044_p indicates paired samples of GSE29044 and GSE29044_u indicates unpaired samples analyzed separately. Number of samples in normal and tumor are shown for each study. Effect-size>0 indicates upregulation in breast tumor while effect-size<0 indicate upregulation in normal; TE (Treatment effect) is calculated using SMD (Standardized Mean Difference); REM=Random Effect Model.

The meta-analysis was repeated by splitting the paired and unpaired datasets. Categorization of data into *paired and unpaired sub-cohort* resulted in 5 and 11 datasets, respectively (Table 5.1). Consistent with combined meta-analysis, GR expression was highly downregulated in tumors in both the paired and unpaired sub-cohorts while MR expression, although significantly downregulated in the unpaired breast tumor, it only approached significance in the paired data (Figure 5.4). Downregulation of MR and GR in unpaired cohort were even more pronounced then combined meta-analysis (MR: mu.hat=-2.22, FDR=9.4e-20; GR: mu.hat=-1.99; FDR=9.4e-20; Figure 5.4). In the case of paired data analysis, GR expression was again significantly downregulated but the effect was milder compared to previous unpaired and combined analysis (GR: mu.hat=-1.43, FDR=1.4e-4; Figure 5.4). Interestingly, MR downregulation trend in paired data was even less prominent and almost approached significance (MR: mu.hat=-1.23, FDR=0.083; Figure 5.4). These results showed that MR expression might have more microenvironment issues in comparison to GR. Paired cohort was comparatively smaller with 10-18 samples in each normal/tumor group except one dataset (GSE15852), which consisted of 43 samples in each group. Forest plot reveals that although MR expression showed a downregulation in other 4 datasets, that trend was not observed in GSE15852 (Figure 5.4). As GSE15852 consisted of large number of samples and did not show MR downregulation, the observed milder change in MR expression might be related to heterogeneity in GSE15852.

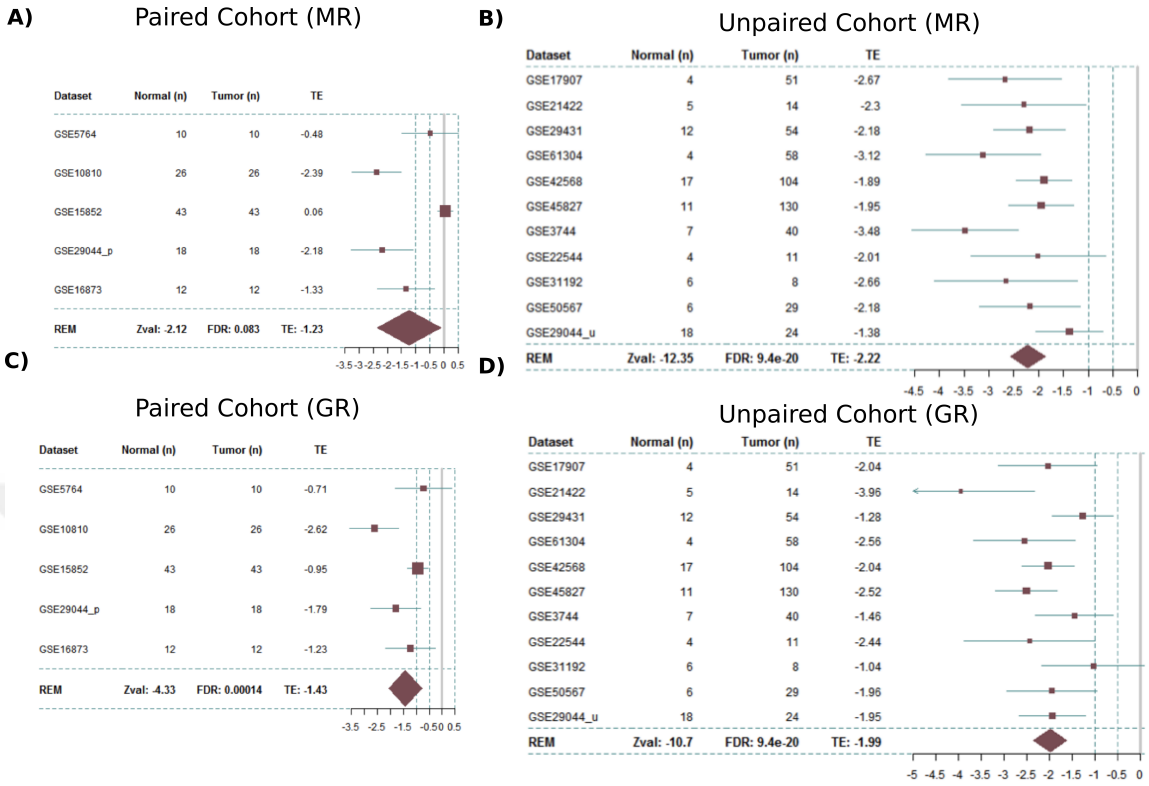


Figure 5.4: MR /GR meta-differential expression analysis in paired and unpaired sub-cohort.

Meta-differential expression of MR in A) paired B) unpaired and GR C) paired and D) unpaired were separately performed. Number of samples in normal and tumor are shown for each study. Effect-size>0 indicates upregulation in breast tumor while effect-size<0 indicate upregulation in normal. TE (Treatment effect) is calculated using SMD (Standardized Mean Difference). REM=Random Effect Model.

5.4.2. Correlation between paired and unpaired breast cancer subcohorts

Whole transcriptome paired and unpaired meta-analysis results were obtained, i.e., an effect size (mu.hat) for each gene in the a) paired and b) unpaired cohorts. Next, these effect-sizes were plotted against each other to identify the correlation in between as well as concordant/discordant gene sets. Effect-size indicated good correlation between both cohorts (Figure 5.5) suggesting that paired and unpaired cohorts' contained similar information about genes in their differential expression of tumor and normal tissues in breast.

Since transcriptome wide paired and unpaired analysis comparison in breast cancer normal/tumor cohorts has been conducted first time in literature, I have thought to

identify concordant/discordant gene sets. Significant probesets with 10% false discovery rate (FDR<0.1) were obtained from both sub-cohorts. Figure 5.6 summarized that 5035 significant probesets were shared between both meta-analysis. Unpaired cohort also had a larger number of significant probesets, which did not appear in the paired cohort. Functional analysis of these cohort specific genes revealed that genes specifically upregulated in the unpaired cohort included DNA replication and repair related processes. However, upregulated genes specific to the paired cohort were enriched with pathways involved in vesicular transport (Table 5.2). Interestingly, downregulated genes in the unpaired cohort were related to post-transcriptional events while NF-KB signaling and metabolic processes were pronounced among the paired cohort (Table 5.3). Genes significantly regulated in the paired data indicate that these genes might have high inter-patient heterogeneity and possible effect of microenvironment which cannot be caught using the unpaired data alone.

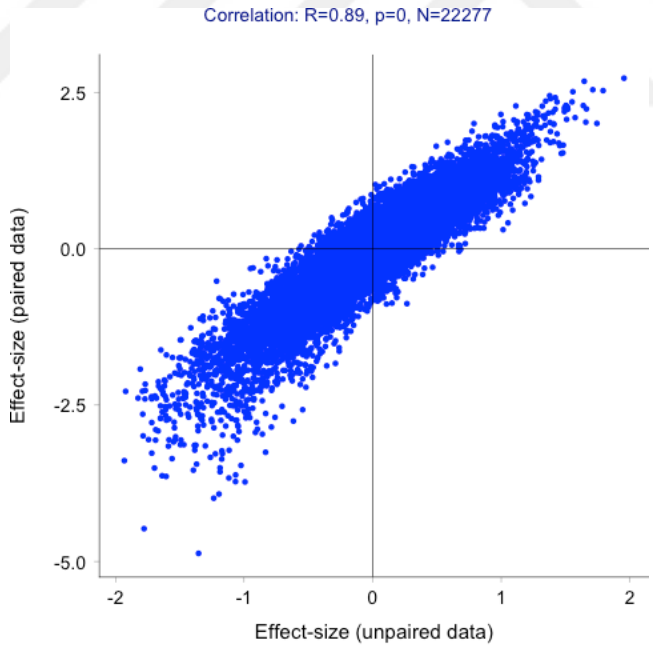


Figure 5.5: Correlation between paired and unpaired effect-size obtained from respective meta-analysis.

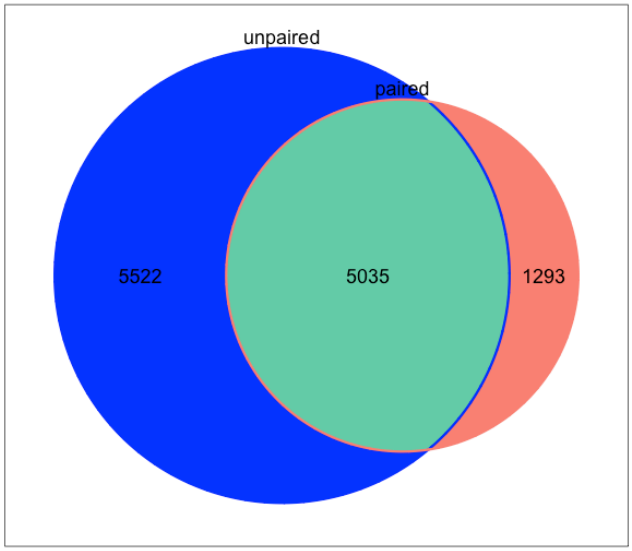


Figure 5.6: Distribution of shared and paired/unpaired specific significantly regulated genes (FDR<0.1).

Table 5.2: Upregulated Gene Ontology (biological processes) terms specific to unpaired and paired sub-cohort.

If more than 10 pathways came as significant (BH<0.05), then 10 pathways otherwise all significant pathways are described.

Term	Count	PValue	Benjamini
Upregulated genes in unpaired cohort			
GO:0016032~viral process	76	1.07E-08	6.02E-05
GO:0006260~DNA replication	48	1.30E-08	3.65E-05
GO:0006281~DNA repair	62	6.65E-08	1.25E-04
GO:0098609~cell-cell adhesion	66	5.76E-07	8.11E-04
GO:0050852~T cell receptor signaling pathway	41	4.00E-06	4.50E-03
GO:0006468~protein phosphorylation	94	5.98E-06	5.60E-03
GO:0019886~antigen processing and presentation of exogenous peptide antigen via MHC class II	29	9.95E-06	7.97E-03
GO:0006886~intracellular protein transport	56	1.03E-05	7.21E-03
GO:0016925~protein sumoylation	33	2.75E-05	1.71E-02
GO:0006303~double-strand break repair via non-homologous end joining	22	2.77E-05	1.55E-02
Upregulated genes in paired cohort			
GO:0006120~mitochondrial electron transport, NADH to ubiquinone	13	1.63E-07	4.13E-04
GO:0030433~ER-associated ubiquitin-dependent protein catabolic process	13	1.70E-06	2.15E-03
GO:0048208~COPII vesicle coating	12	1.32E-05	1.10E-02
GO:0006521~regulation of cellular amino acid metabolic process	11	1.51E-05	9.47E-03
GO:0002479~antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP-dependent	12	1.81E-05	9.13E-03
GO:0032981~mitochondrial respiratory chain complex I assembly	12	1.81E-05	9.13E-03
GO:0006888~ER to Golgi vesicle-mediated transport	19	2.86E-05	1.20E-02
GO:0031145~anaphase-promoting complex-dependent catabolic process	13	3.27E-05	1.18E-02
GO:0051436~negative regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle	12	5.76E-05	1.80E-02

Table 5.3: Downregulated Gene Ontology (biological processes) terms specific to unpaired and paired sub-cohort.

Term	Count	PValue	Benjamini
Downregulated genes specific to unpaired cohort			
GO:0006413~translational initiation	64	1.70E-26	8.68E-23
GO:0006614~SRP-dependent cotranslational protein targeting to membrane	52	5.77E-26	1.47E-22
GO:0000184~nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	58	3.18E-25	5.39E-22
GO:0019083~viral transcription	54	3.08E-23	3.92E-20
GO:0006364~rRNA processing	66	7.27E-16	7.91E-13
GO:0006412~translation	68	3.82E-13	3.24E-10
GO:0001525~angiogenesis	48	2.08E-06	1.51E-03
GO:0043547~positive regulation of GTPase activity	95	3.27E-06	2.08E-03
GO:0007507~heart development	40	1.13E-05	6.40E-03
GO:0002181~cytoplasmic translation	12	1.76E-05	8.90E-03
Downregulated genes specific to paired cohort			
GO:0043123~positive regulation of I-kappaB kinase/NF-kappaB signaling	13	1.36E-04	2.30E-01
GO:0002933~lipid hydroxylation	4	1.68E-04	1.49E-01
GO:0006805~xenobiotic metabolic process	8	1.16E-03	5.23E-01
GO:0051028~mRNA transport	6	2.79E-03	7.38E-01
GO:0071456~cellular response to hypoxia	8	3.81E-03	7.68E-01
GO:0006970~response to osmotic stress	4	5.71E-03	8.40E-01
GO:0045944~positive regulation of transcription from RNA polymerase II promoter	33	7.52E-03	8.73E-01
GO:0044255~cellular lipid metabolic process	5	1.00E-02	9.11E-01
GO:0002576~platelet degranulation	7	2.03E-02	9.87E-01
GO:0070989~oxidative demethylation	3	2.46E-02	9.92E-01

5.4.3. Correlation between MR and GR expression in breast normal/tumor cohort

Both MR and GR have originated from the same ancestral gene. Downregulation of both genes in breast cancer and their comparable effect-sizes thus can implicate a significant association between the expression of both genes. Thus, I performed a correlation analysis between the MR and GR probeset expression values. Meta-correlation analysis was performed without segregating the paired and unpaired data to increase the sample size. Pipeline of meta-correlation analysis was shown by a PRISMA diagram (Figure 5.1). Results illustrated a significant positive association existing between MR and GR genes at the transcript level (R.mean=0.56, p=9.1e-29; Figure 5.7A). TCGA RNAseq normalized data was used to validate this correlation observed between both genes based on microarray data. Results confirmed that a positive correlation existed also in RNAseq breast cancer samples (R=0.48, p=7.9E-65; Figure 5.7B).

5.4.4. Correlation of MR and GR co-expression profiles

A significant positive correlation between expression of both MR and GR in two independent cohorts (microarray and RNAseq) highlighted that the downstream transcriptomic regulation by these genes might also be shared. Transcriptome-wide meta-correlation was performed between the expression of every gene and that of MR or GR probesets, separately giving rise to MR- and GR-meta-correlation profiles. Plotting genome-wide meta-correlation scores across each other revealed that indeed the co-expression profile of MR was very highly correlated with that of GR in a positive manner (R=0.93, p=0; Figure 5.7C).

Consistent with MR and GR downregulation in breast cancer, MR and GR co-expression profiles also seemed to be downregulated in breast cancer revealing a significant negative correlation between breast cancer effect-size and MR/GR meta-correlation profile (MR co-expression signature: R=-0.84, p=0; GR co-expression profile: R=-0.77, p = 0; Figure 5.7D-E). Although correlation was strong for both genes it was comparatively less for the GR correlated profile. These findings revealed that MR and GR might work together in breast tissue and their combined downregulation in breast

cancer when compared with normal tissue suggested that this was not a random process but both receptors may share an underlying regulatory mechanism.

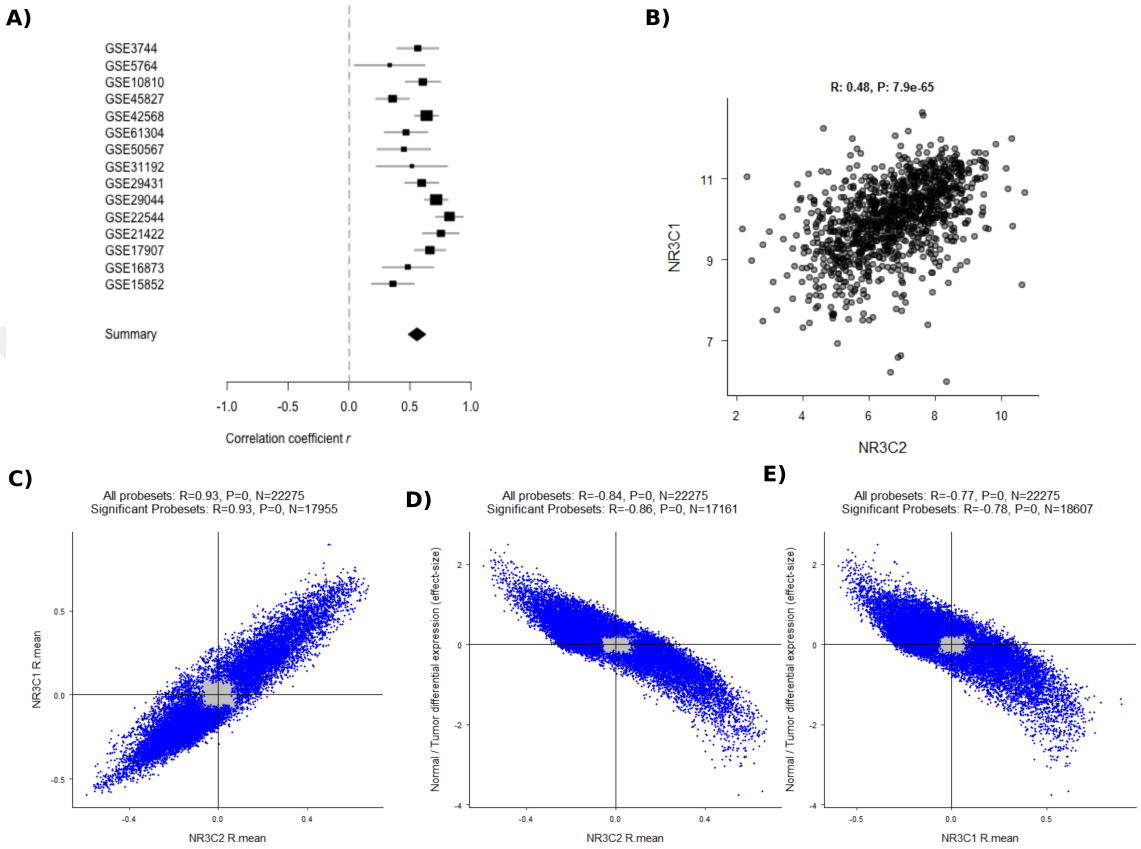


Figure 5.7: Correlation between MR and GR expression and co-expression profiles.

Correlation between both genes in A) breast normal tumor cohort meta-analysis and B) TCGA breast tumor expression data. Transcriptome-wide meta-correlation profile of C) MR and GR, D) MR and transcriptome-wide effect-size in breast cancer, E) GR and transcriptome-wide effect-size in breast cancer. Effect-size of breast cancer is based on combined paired and unpaired meta-analysis.

5.4.5. Prognositic importance of MR and GR in breast cancer

Strong downregulation of MR and GR expression along with shared co-expression profiles prompted us to investigate the prognostic importance of MR and GR. Online database KM plotter was used to analyze survival of patients. Results revealed that there was a separation of survival probability between the low and high MR expression. Low hazard ratio (HR<1) indicated that indeed higher MR expression was associated with a

better survival. Although MR expression showed differential prognostic importance, association of GR with survival could not be identified in significant terms (Figure 5.8).

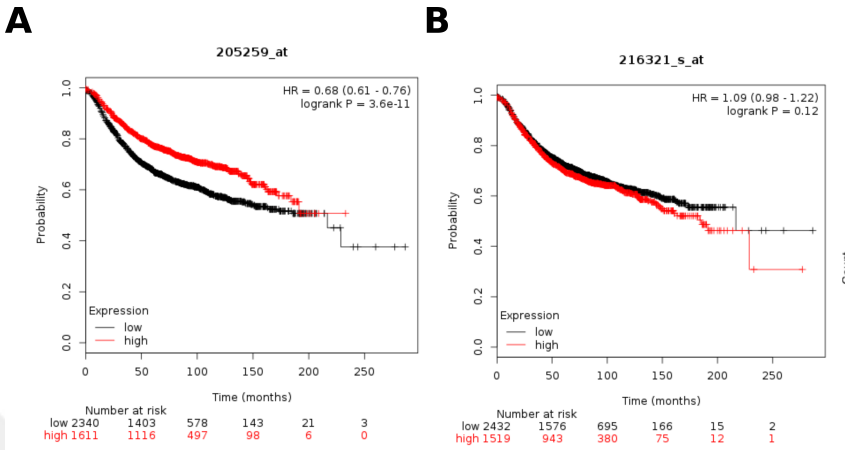


Figure 5.8: Survival analysis of MR and GR expression in breast cancer patients (KM plotter database).

5.4.6. Differential co-expression of MR and GR in breast cancer

Difference in prognostic importance of MR and GR provoked the thinking that although the expression profiles of MR and GR shared a high degree of commonality, a smaller network of genes specific to MR/GR might be present leading to differential prognostic implications. MR-specific co-expression module was extracted by filtering genes which were significantly correlated with MR but not with GR (MR meta-correlation $p<0.05$ and GR meta-correlation $p>0.05$); and GR-specific co-expression profile by extracting GR correlated genes with no significant correlation with GR (GR meta-correlation $p<0.05$ and MR meta-correlation $p>0.05$). Shared MR and GR correlated profile consisted of genes, each showing a significant correlation with both MR and GR in the same direction (MR meta-correlation $p<0.05$ and GR meta-correlation $p<0.05$). Distribution plot of these genes showed that although MR/GR correlated genes were higher in number and possessed a higher correlation scores, a smaller module of genes showing less but significant correlation specifically with MR and GR also existed (Figure 5.9).

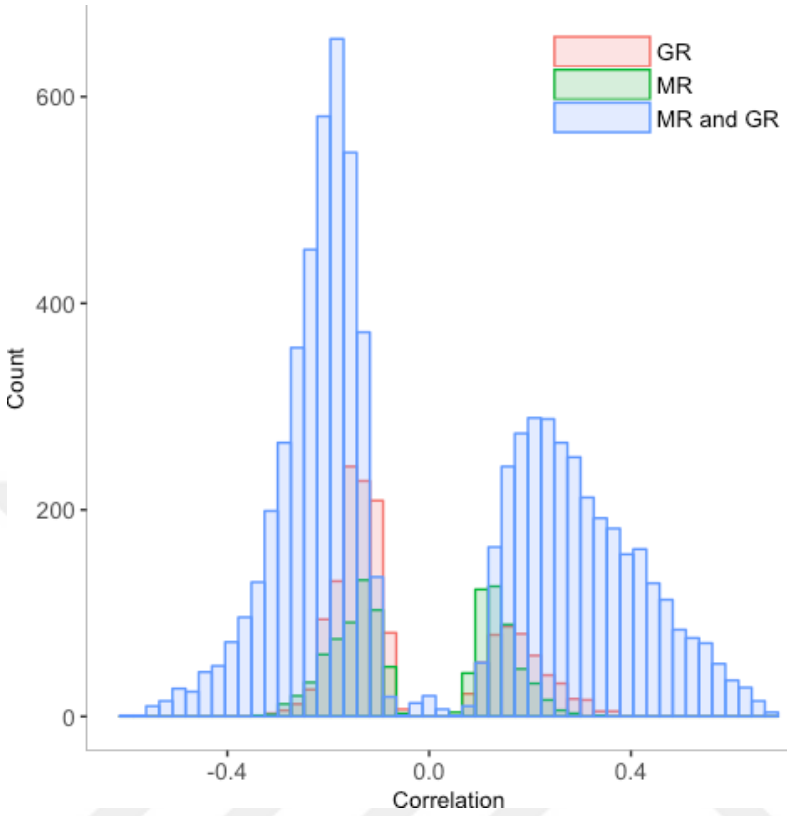


Figure 5.9: Distribution of shared MR/GR and MR/GR specific correlation profile.

Biological enrichment of shared and MR/GR specific coexpression profile was conducted using gene ontology enrichment (DAVID). Shared MR/GR coexpression profile (MR meta-correlation: absolute r.mean>0.4, p<0.05 and GR meta-correlation score: absolute r.mean>0.4, p<0.05) revealed that both genes cooperate to regulate PI3K signaling, cell adhesion, angiogenesis (Table 5.4) while negatively regulated pathways including cell division, proliferation and DNA replication (Table 5.4). Results highlight that anti-proliferative capability is shared by MR and GR however, future studies are needed to verify the findings.

MR specific coexpression profile (MR meta-correlation score: absolute r.mean>0.1 and p<0.05; GR meta-correlation score: p>0.05) revealed that upregulated MR-specific profile was involved in differentiation and morphogenesis while negatively regulated profile was enriched with immune response (Table 5.5).

Table 5.4: Top 10 significantly regulated gene ontology terms (biological processes) enriched in MR/GR shared co-expression profile.

Term	Count	PValue	Benjamini
Upregulated			
GO:0014068~positive regulation of phosphatidylinositol 3-kinase signaling	15	8.33E-09	2.42E-05
GO:0001525~angiogenesis	26	2.57E-08	3.72E-05
GO:0045944~positive regulation of transcription from RNA polymerase II promoter	61	2.74E-07	2.65E-04
GO:0008285~negative regulation of cell proliferation	33	7.00E-07	5.08E-04
GO:0007155~cell adhesion	35	2.25E-06	1.31E-03
GO:0046326~positive regulation of glucose import	9	2.46E-06	1.19E-03
GO:0032869~cellular response to insulin stimulus	13	3.80E-06	1.57E-03
GO:0030335~positive regulation of cell migration	20	4.18E-06	1.52E-03
GO:0032870~cellular response to hormone stimulus	10	7.65E-06	2.46E-03
GO:0001938~positive regulation of endothelial cell proliferation	12	7.68E-06	2.23E-03
Downregulated			
GO:0051301~cell division	35	6.29E-29	4.07E-26
GO:0007067~mitotic nuclear division	26	1.11E-21	3.60E-19
GO:0007062~sister chromatid cohesion	16	8.95E-16	1.92E-13
GO:0000070~mitotic sister chromatid segregation	9	6.41E-12	1.04E-09
GO:0000086~G2/M transition of mitotic cell cycle	14	2.28E-11	2.95E-09
GO:0007018~microtubule-based movement	11	3.72E-10	4.01E-08
GO:0000281~mitotic cytokinesis	8	1.30E-09	1.20E-07
GO:0006260~DNA replication	13	1.49E-09	1.20E-07
GO:0051726~regulation of cell cycle	12	1.74E-09	1.25E-07
GO:0008283~cell proliferation	17	1.06E-08	6.89E-07

On the other hand, GR specific profile in breast cancer (GR meta-correlation score: absolute r.mean>0.1 and $p<0.05$; MR meta-correlation score: $p>0.05$) was upregulated in immune related processes while steroid metabolism and chemical synaptic transmission were repressed (Table 5.6).

Table 5.5: Top 10 significantly regulated gene ontology terms (biological processes) enriched in MR specific co-expression profile.

Term	Count	PValue	Benjamini
Upregulated			
GO:0035176~social behavior	6	3.71E-03	9.99E-01
GO:0045787~positive regulation of cell cycle	5	6.03E-03	9.96E-01
GO:0086091~regulation of heart rate by cardiac conduction	5	6.69E-03	9.85E-01
GO:0007010~cytoskeleton organization	10	8.57E-03	9.82E-01
GO:0007416~synapse assembly	6	1.03E-02	9.79E-01
GO:0006355~regulation of transcription, DNA-templated	47	1.08E-02	9.66E-01
GO:0060042~retina morphogenesis in camera-type eye	3	1.20E-02	9.60E-01
GO:0022604~regulation of cell morphogenesis	4	1.63E-02	9.78E-01
GO:0045893~positive regulation of transcription, DNA-templated	20	1.75E-02	9.74E-01
GO:0045944~positive regulation of transcription from RNA polymerase II promoter	32	2.30E-02	9.87E-01
Downregulated			
GO:0006955~immune response	38	3.60E-10	8.37E-07
GO:0060337~type I interferon signaling pathway	15	1.38E-09	1.61E-06
GO:0002479~antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP-dependent	14	1.15E-08	8.91E-06
GO:0050852~T cell receptor signaling pathway	20	2.09E-08	1.22E-05
GO:0033209~tumor necrosis factor-mediated signaling pathway	17	1.27E-07	5.92E-05
GO:0045087~innate immune response	33	2.84E-07	1.10E-04
GO:0034341~response to interferon-gamma	8	2.45E-06	8.13E-04
GO:0002250~adaptive immune response	17	2.83E-06	8.21E-04
GO:0006928~movement of cell or subcellular component	13	3.53E-06	9.11E-04
GO:0042110~T cell activation	10	4.12E-06	9.57E-04

Table 5.6: Top 10 singificantly regulated gene ontology terms (biological processes) enriched in GR specific co-expression profile.

Term	Count	PValue	Benjamini
Upregulated pathways			
GO:0006954~inflammatory response	35	1.48E-10	3.14E-07
GO:0006955~immune response	36	6.22E-10	6.60E-07
GO:0032496~response to lipopolysaccharide	17	4.04E-06	2.85E-03
GO:0060333~interferon-gamma-mediated signaling pathway	11	1.08E-05	5.73E-03
GO:0007165~signal transduction	53	3.73E-05	1.57E-02
GO:0050729~positive regulation of inflammatory response	10	8.75E-05	3.05E-02
GO:0042535~positive regulation of tumor necrosis factor biosynthetic process	5	1.14E-04	3.39E-02
GO:0007259~JAK-STAT cascade	7	1.28E-04	3.34E-02
GO:0098869~cellular oxidant detoxification	9	3.66E-04	8.26E-02
GO:0097190~apoptotic signaling pathway	9	4.03E-04	8.20E-02
Downregulated pathways			
GO:0007268~chemical synaptic transmission	28	1.30E-05	3.52E-02
GO:0019373~epoxygenase P450 pathway	7	9.87E-05	1.27E-01
GO:0007339~binding of sperm to zona pellucida	9	1.36E-04	1.17E-01
GO:0008202~steroid metabolic process	9	6.07E-04	3.41E-01
GO:0006805~xenobiotic metabolic process	12	7.49E-04	3.38E-01
GO:0006810~transport	30	1.24E-03	4.33E-01
GO:0007338~single fertilization	10	1.26E-03	3.92E-01
GO:0042738~exogenous drug catabolic process	5	1.55E-03	4.13E-01
GO:0055114~oxidation-reduction process	42	4.73E-03	7.65E-01
GO:0042493~response to drug	25	6.15E-03	8.17E-01

Hence our findings indicated that MR and GR were highly co-expressed and shared profile indicated their anti-proliferative roles in breast cancer; however MR and GR

specific co-expression networks were involved in regulation of specific immune responses which needs to be verified by wet-lab experiments.

5.4.7. MR and GR deregulation in breast cancer

Since MR and GR largely share their co-expression profiles and are highly correlated, I asked whether coregulation between both genes remained the same in normal and tumor samples. Using the paired cohort, which consisted of equal normal and tumor samples in each dataset, a deregulation analysis was performed as mentioned in Materials and Methods section. The results revealed that 3 out of 5 paired datasets showed significant difference in variance of MR/GR expression measurement between normal and tumor sample groups. Normal samples showed tighter regulation of both genes while in tumor samples, expressions of both genes were dispersed within high range of variability (Figure 5.10).

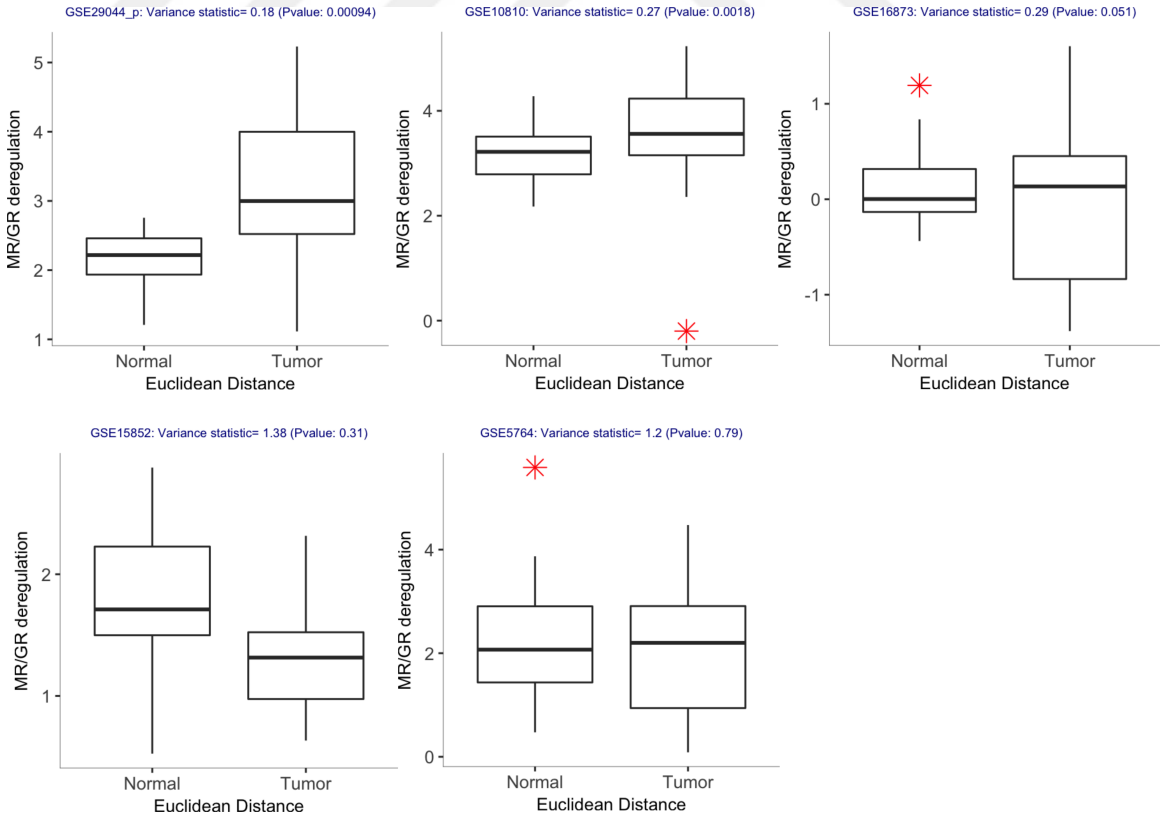


Figure 5.10: MR/GR deregulation in normal/tumor samples of breast cancer.

5.4.8. Development of a web tool for normal/tumor expression comparison

Based on normal tumor breast cancer expression data, an online database has been created which facilitates the user to query any gene and observe its differential regulation in breast cancer and normal tissues. DualExpBC has been divided into several modules as listed below:

1) Differential expression

Upon user entering the query gene, DualExpBC consults the stored meta-differential expression analysis of paired, unpaired and combined profiles and visualizes the result in form of three separate forest plots under the same tab. Forest plot shows effect-size in individual dataset with 95% confidence interval along with meta-analysis results at the bottom.

2) Meta-correlation analysis

This module asks user to enter another gene as it is designed to find association between the two input genes. After user enters a second gene, DualExpBC connects with all normalized data files to retrieve expression of both genes. Expression of both genes from each dataset is summarized in the form of correlation coefficients. These correlation coefficients are combined in the form of meta-correlation analysis and then results are visualized as forest plots.

3) Individual dataset expression

Like the meta-correlation analysis, this module also accepts two query genes and in addition, facilitates the user to select any of the individual datasets from cohort of 16 studies. Similar to the meta-correlation profile, this module obtains expression of both genes from a user-selected dataset and also obtains normal/tumor sample labels. Results are visualized in the form of scatter plot with x and y- axes showing normalized expression of both genes and color labels highlight normal/tumor label of samples in the selected dataset.

4) Gene deregulation

Based on user entered genes and a selected dataset, the gene deregulation module is designed to visualize the degree of dissimilarity between the two selected genes in

normal/tumor samples. F-test is used to find significance of variance of deregulation (i.e., based on the difference of expression between genes) in between the normal and tumor groups. Gene deregulation analysis can be performed with or without considering direction of expression deregulation. Three input options are provided for this purpose, i.e., A) without direction (absolute difference in both genes), B) Gene1-Gene2 C) Gene2-Gene1. Results show box plot of normal and tumor on x- axis and deregulation value on y-axis.

Step-by-step tour of DualExpBC using cell cycle gene CCNE2 is provided in Appendix-IV.

5.5. Discussion

Meta-analysis of MR and GR expression with respect to each other and across a large set of breast cancer/normal datasets has not been previously performed. Nuclear Receptors MR and GR originate from the same ancestor so their functions are also related. At the same time, they share regulatory components as well, e.g., transcription factors. Although they compete with each other they also have relatively distinct functions due to differential enzymatic, synthetic machineries and ligand affinities.

In this thesis, several questions have been answered regarding the dual action of these two receptors. One of them was whether MR and GR were meta-correlated across a large cohort of breast cancer datasets to facilitate comparisons between the breast tumor and normal tissues. The hypothesis was based on the assumption that there was a correlation between MR and GR expressions. Ongoing work by Seniye Targen and Bircan Coban indicated that qPCR and western blotting support co-expression of MR and GR, with a degree of deregulation, in breast cancer cell lines (Bircan Coban, MS Thesis, 2016; Seniye Targen, PhD Candidate).

Furthermore, ongoing studies and previous published research suggested that normal breast tissue had higher expression of both genes when compared with tumor [516]. However, this is a first meta-analysis attempt in the literature in regard to MR and GR correlation as well as their association with tumor or normal tissue. In addition, this analysis separates datasets in paired and unpaired fashions for the sake of accuracy thus the findings can also indicate tumor and adjacent unaffected tissue effects. Most meta-analysis studies focused on unpaired expression datasets yet there are several meta-analysis studies in the literature that deal with paired datasets (e.g., papillary thyroid cancer [517]; thyroid carcinoma [518]; B-ALL [519]; gastric cancer primary/metastasis [520]). These studies show that the information obtained from paired meta-analysis studies can be interesting. However very few studies (mentioned in introduction) have focused on comparative paired and unpaired analysis. Specifically, meta-analysis of paired and unpaired data in breast cancer could not be identified, hence ours is the first. In GR and MR, we have found that both the paired and unpaired data support each other

and the direction of change is the same although MR exhibited more variability among the paired datasets thus leading to a smaller and less significant effect size. This methodology was also extended to transcriptome-wide consistency/discrepancy in paired/unpaired samples. Our results found that although effect-size of the paired/unpaired meta-analysis showed high correlation, the number of significantly regulated genes showed deviation. Diverged significance pattern could also be related to varying number of studies in the paired and unpaired cohorts, as the paired cohort was comparatively smaller. Analysis could be useful to identify discrepant meta-analysis scores thus can help identify effects of microenvironment on gene expression. Genome-wide comparative paired and unpaired meta-analysis has put forwarded several important pathways, which need to be verified and indicated that vesicular transport related pathways were probably affected by microenvironment and were specific to paired data. It might also indicate that paired data specific pathways might show high heterogeneity between patients. However, results showed that replication and repair related processes did not differ due to inter-patient heterogeneity and were enriched in unpaired data.

Gene deregulation is an important concept for the pair of genes, which work in close relationship. MR/GR is one such pair with highly conserved and shared profile. The concept has already been highlighted in literature and in previous study, it has been shown that cytoplasmic and mitochondrial nucleotide metabolism coordinated regulation present in normal tissue was lost in transformed tissues [521]. Routinely used expression analysis does not encounter coordinated expression changes. Our study has presented a simple and interesting strategy to evaluate deregulation between normal and tumor samples. Besides, incorporating this module in DualExpBC can help user to identify deregulation of any two genes in breast cancer.

No online tool exists to meta-analyze any two genes in breast cancer with respect to tumor normal for the identification of meta-correlation profiles as well as the extent of deregulation between genes of interest. MR and GR, for example, are highly correlated but are associated with varying hazard ratios based on the patient survival data. This suggests that their correlated partners may differ thus make them diverge in prognosis.

This kind of tool would be very useful for comparing other paralogs pairs (like MR and GR), or members of a receptor groups to answer questions, such as “Does PR network correlate with MR network?” In addition, shared and divergent networks of co-expressed genes are of interest in the literature. Therefore, the DualExpBC is a first version of a Shiny Application that allows any two human genes to be compared for their divergence in breast tumors and normal tissue in a paired and unpaired fashion. In the literature, there is no Shiny tool that does two gene comparisons in the context of breast cancer meta-analysis. Not specific to breast, but there are several Shiny packages in the literature that focus on cancer expression. Among these, hemaclass.org [522] focuses on hematological cancers and their classification based on multiple GEO datasets. GAM, on the other hand, focuses more on integration of metabolomics and transcriptomics datasets with the aim of creating graphs from the given data and testing of the hypothesis [523]. Another integrative tool named PACMEN uses TCGA data along with annotations to represent mutation-expression networks [524]. There are several other R shiny apps for integrative genomics but none of them incorporated meta-analysis of breast cancer tumor/normal datasets with meta-analytic statistics, automatic drawing of forest plots, and visualization capacities. There are online databases that are applicable to multiple cancers, such as oncomine, one can test tumor and normal expression differences [525]. However, DualExpBC is a customized breast cancer oriented meta-analysis tool using REM methodology and forest plots. It is unique and modular, allowing incorporation of new features for future releases. For example, a potential application already included but requiring further studies is assessment of gene deregulation module. In addition, other platforms and data from RNAseq can be incorporated into DualExpBC.

5.6. Future perspectives

This study has shown MR/GR co-expression and deregulation in breast cancer using transcriptome meta-analysis. In future, the study can be extended for co-expression network analysis using WGCNA, which can help prioritize important shared and specific MR/GR modules. MR/GR shared co-expression profile highlighted its anti-proliferative role in cancer but experimental verification of highly deregulated gene via qPCR can validate its significance.

Deregulation analysis has shown that MR/GR work in the cell in close relationship especially in normal samples however; this tight co-expression seems to be relatively lost in tumor samples. The phenomenon was significantly observed in 3 out of 5 observed datasets indicating 60% cases. Further verification using a larger cohort of paired data (RNAseq, microarray, protein) can be helpful to generalize the MR/GR deregulation in breast cancer. Experimental verification using either normal/cancer patients or normal/tumor breast cell lines can also help verification. The number of normal/tumor samples in paired cohort was smaller and correlation analysis does not efficiently work with less number of samples. Hence, larger cohorts with more number of samples in both groups of all datasets can help generating separate meta-correlation profile with respect to MR/GR in normal and tumor condition. Separate co-expression network of normal and tumor samples can be compared using network comparison packages (e.g. WGCNA) to identify MR/GR associated network, which is also deregulated with the onset of tumor. Comparison of normal/tumor network has already been practiced with different underlying hypotheses. Loss of transcriptional interaction network in colon tumor samples was observed in comparison to large transcriptional network in normal tissues [526]. Since MR/GR expression shows variance in normal/tumor samples, it would be interesting to analyze differential co-expression network in normal/tumor samples.

Paired/unpaired meta-analysis of breast cancer/ normal patient data has revealed that unpaired data analysis resulted in larger number of significant genes compared to paired data meta-analysis. As a result of meta-analysis few interesting pathways like DNA

replication and cell-cell adhesion has been found prominent in unpaired meta-analysis. However, experimental verification of these pathways using biopsy or tissues can verify these findings.

In addition, I have developed a database DualExpBC that can provide a base to transcriptome-wide co-expression network using paired/unpaired meta-analysis and normal/tumor deregulation approaches described in the thesis. Besides DualExpBC can be extended in several ways

- DualExpBC evaluates gene expression of paired and unpaired breast cancer patients. However, database is designed to analyze single gene for differential expression analysis and two genes for meta-correlation analysis. Co-expression analysis could be extended to whole genome to help gene prioritization. Similarly, the database should be updated to allow parallel analysis of genome-wise co-expression profile with effect-size of paired and unpaired breast cancer patients.
- Deregulation among genes can be combined using meta-analysis instead of its representation in individual datasets.
- Addition of RNAseq data can make DualExpBC more extensive.
- Functional analysis using KEGG, GO, Reactome databases can be included in DualExpBC for co-expression profile.
- Differential functional analysis and co-expression of paired and unpaired meta-analysis can be an interesting step to understand biological differences in paired/unpaired data.

Appendix I

Working Tutorial of CompariZome

FN1 is an important angiogenesis gene having two orthologs in zebrafish. The duplicated gene pair (FN1a and FN1b) was identified by processing of Biomart orthology information. Literature search also revealed that in 2005, FN1b was discovered as duplicated gene pair of FN1a [527]. FN plays important role in embryogenesis [527] and is associated with angiogenesis [528]. FN1a and FN1b showed opposite regulation in rapamycin treatment analysis (Table 3.8). CompariZome has collection of 7 drug treatment datasets and one of these datasets (GSE24527, Leflunomide treatment; Table 3.9) also has corresponding human datasets (GSE67338, Teriflunomide; Table 3.9).

Leflunomide and Teriflunomide

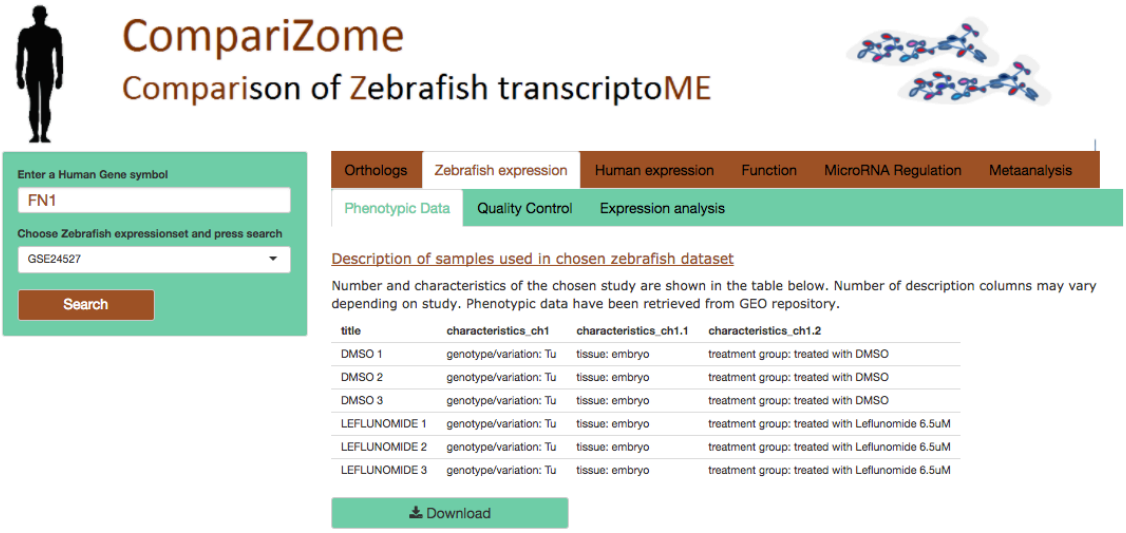
Leflunomide was initially suggested for treatment of rheumatoid arthritis. Leflunomide exhibit its action by rargeting mitochondrial enzyme dihydro-orotic acid dehydrogenase (DHODH). DHODH is involved in regulation of pyrimidine synthesis. Leflunomide treated cells become depleted of ribonucleotide, DNA and RNA synthesis mechanisms are slowed down and cell proliferation is arrested in the G1 phase [529]. Leflunomide has also been tested in multiple cancer types including melanoma, gliomastoma and breast cancer [530]. Teriflunomide is active metabolite of leflunomide and studies have revealed that leflunomide completely convert to Teriflunomide for its downstream actions [531].

Comparison of zebrafish dataset (GSE24527) with human dataset (GSE67338)

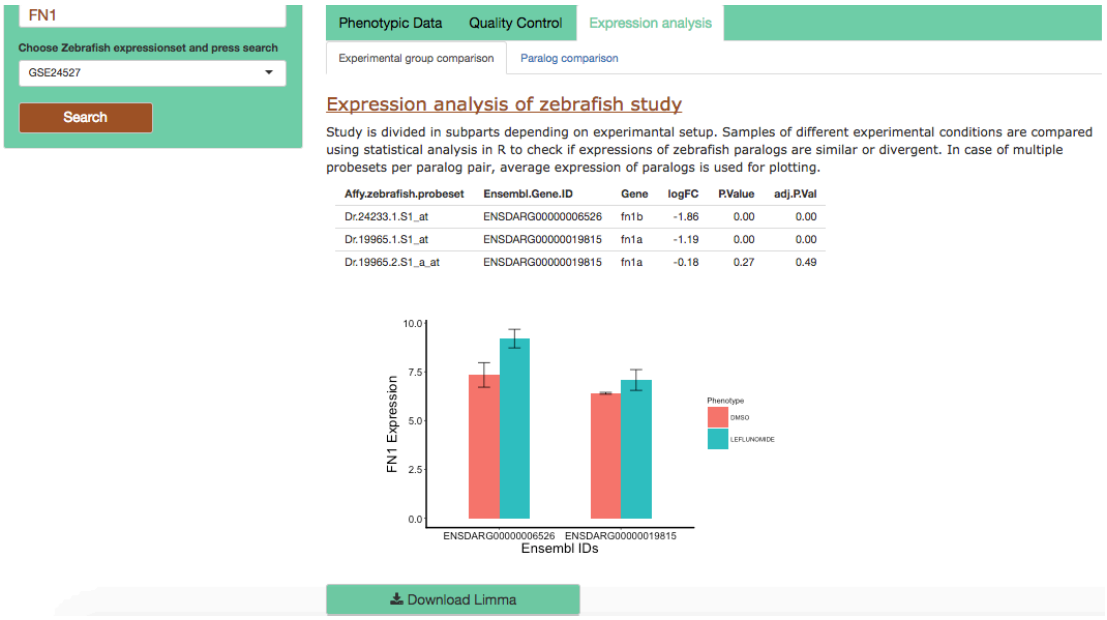
Human experiment of teriflunomide is conducted on neuroblastoma cell line (IMR32) the concentration of 118 μM for 12 hours while leflunomide treatment in zebrafish is conducted on zebrafish embryo until 24hpf (hours post fertilization) with 6.5uM drug. Control group in both treatments was DMSO. FN1 was analyzed using CompariZome to

evaluate if other drugs used in cancer treatment also result in opposite regulation of FN1a and FN1b expression.

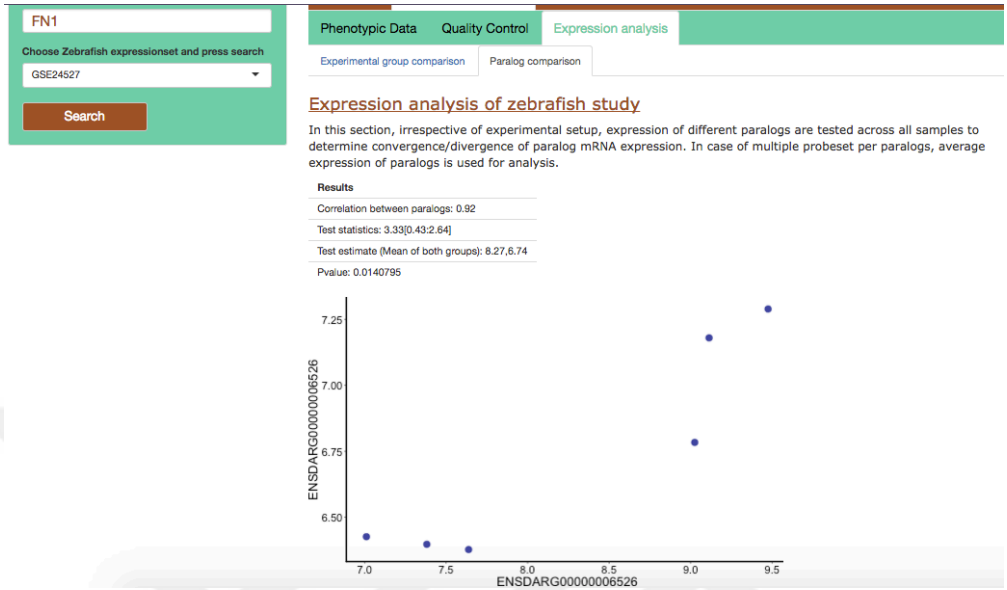
Phenotypic data of zebrafish experiment is shown in the figure.



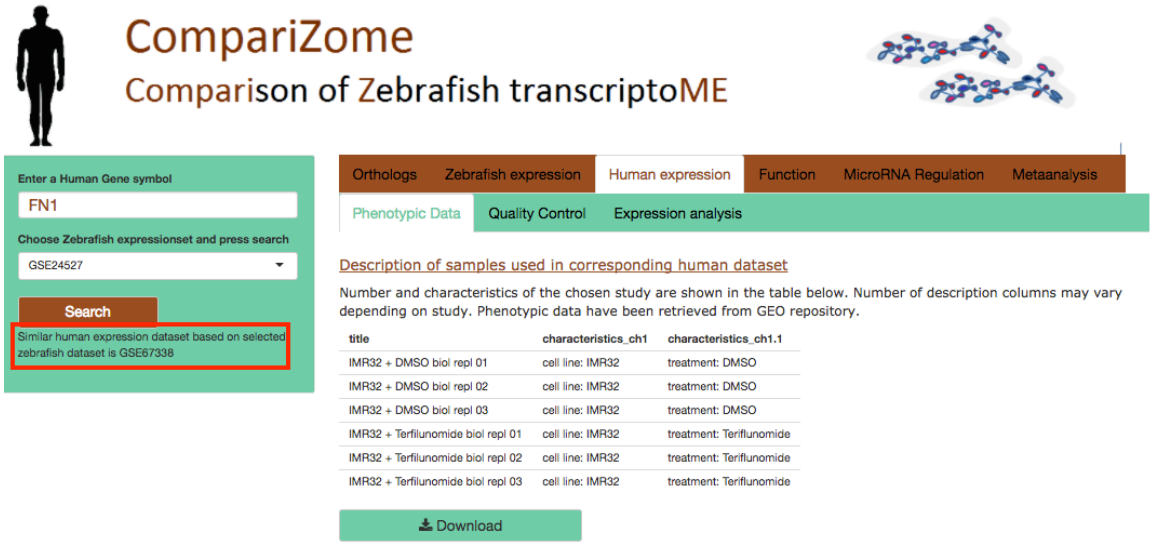
Dwtails of Zebrafish expression regulattion can be viewed under **Zebrafish expression analysis,** which has two further modules. **Experimental group comparison** allows limma analysis and along with logFC and p-value, a bar plot of mean expression is also shown. Bar plot are labeled with experimental groups and can help identifying direction of *limma* logFC.



Paralog comparison module visualizes the scatter plot between both paralog pairs.



By clicking on '**Human expression**' module, sidebar shows a text line indicating respective human dataset.

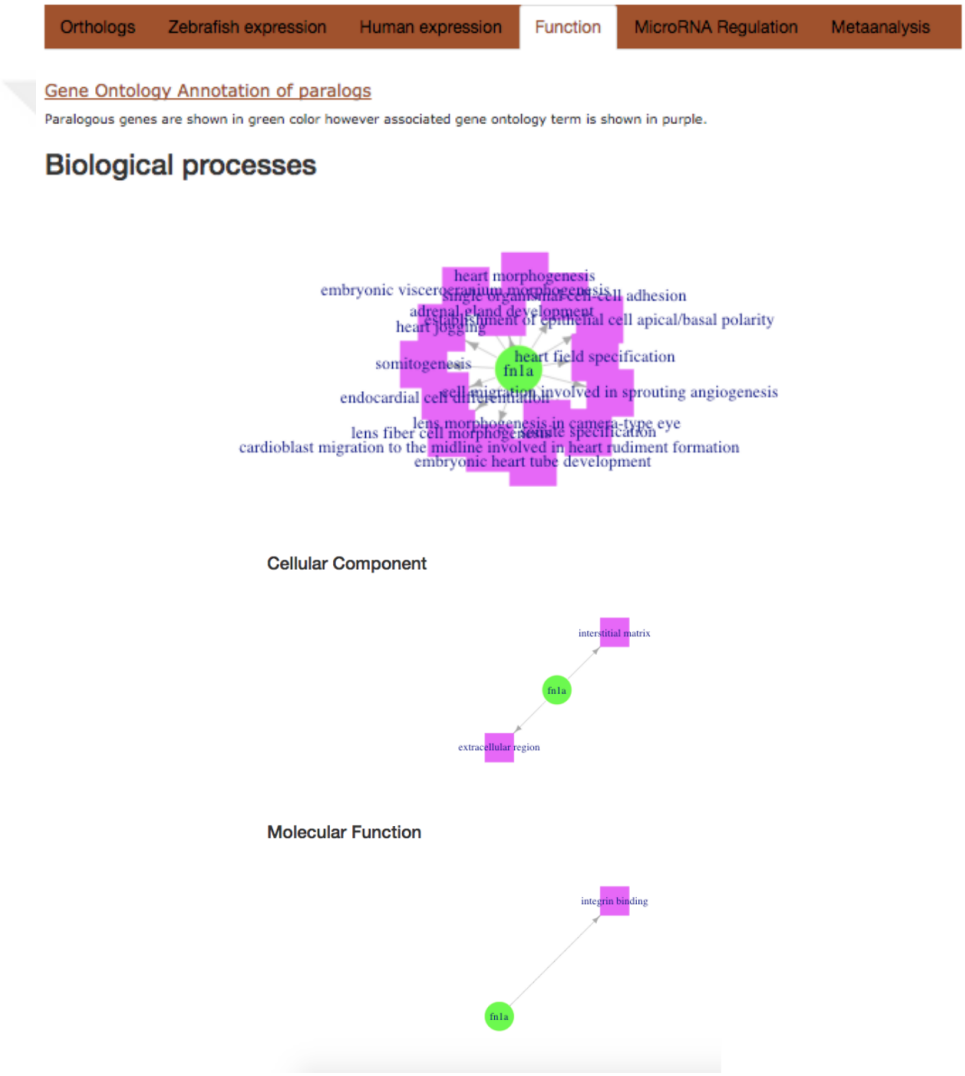


Expression module of human data visualized *limma* results along with graphical representation in form of bar plot.

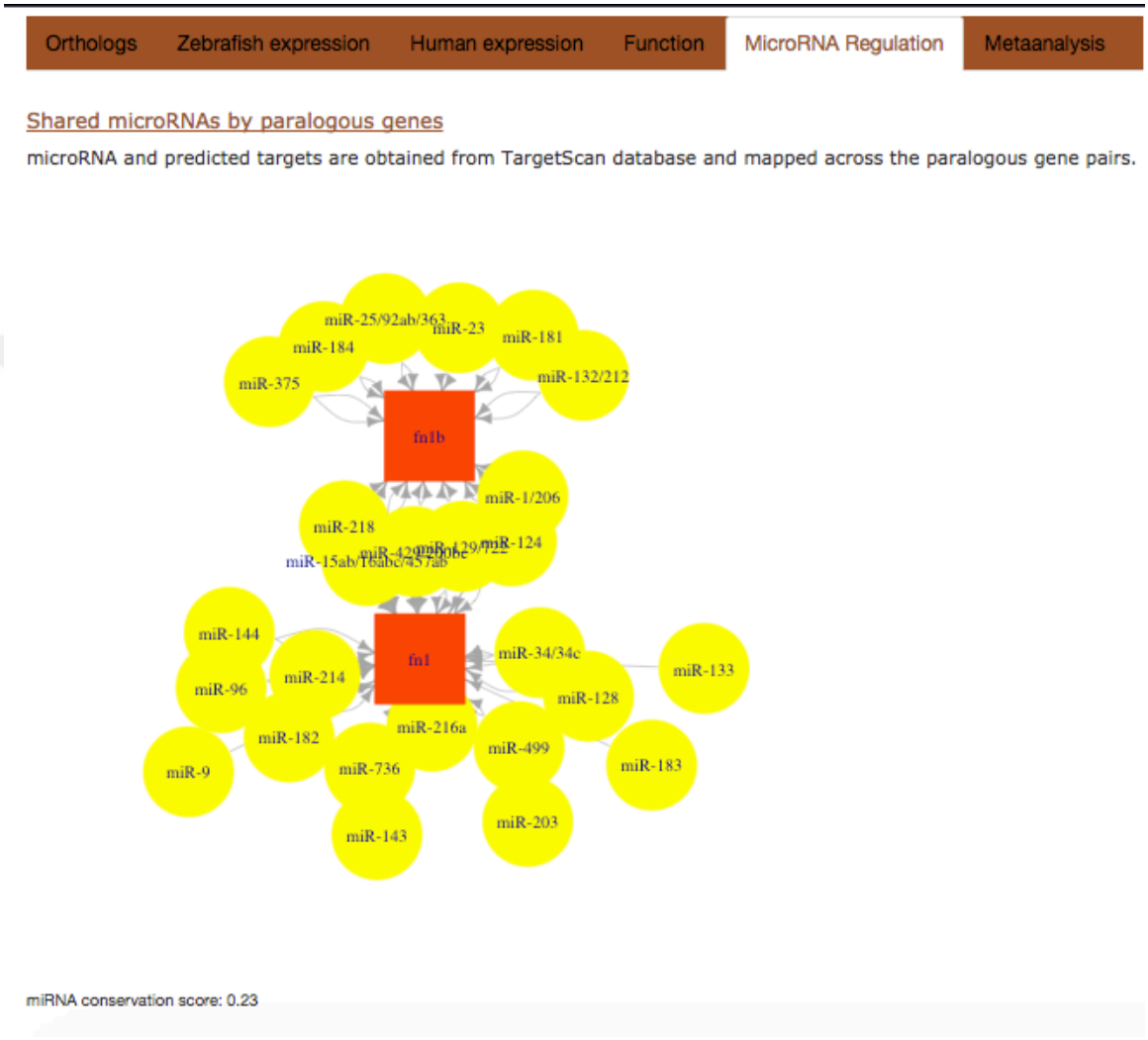


Function module can be used to find gene ontologies term associated with FN1a and FN1b duplicated gene pair.

Results reveal that significantly modulated probesets of FN1a and FN1b were upregulated as result of leflunomide treatment in zebrafish embryo and teriflunomide treatment in human neuroblastoma cell line resulted in similar expression in both genes of paralog pair.



MicroRNA regulation module visualizes shared or gene specific microRNAs hitting duplicated pair.



Appendix II

Predicted Secondary E2-induced targets with PS-score

Probeset	Symbol	PS-score	Probeset	Symbol	PS-score
204603_AT	EXO1	2.08	205733_AT	BLM	1.37
219702_AT	PLAC1	2.04	201292_AT	TOP2A	1.37
220177_S_AT	TMPRSS3	1.89	220651_S_AT	MCM10	1.36
219000_S_AT	DSCC1	1.87	207590_S_AT	CENPI	1.36
205034_AT	CCNE2	1.70	209714_S_AT	CDKN3	1.36
209891_AT	SPC25	1.67	202589_AT	TYMS	1.36
204162_AT	NDC80	1.67	202532_S_AT	DHFR	1.35
219985_AT	HS3ST3A1	1.65	218875_S_AT	FBXO5	1.35
208955_AT	DUT	1.63	203967_AT	CDC6	1.34
204126_S_AT	CDC45	1.61	202726_AT	LIG1	1.34
204444_AT	KIF11	1.61	204825_AT	MELK	1.34
218782_S_AT	ATAD2	1.60	221703_AT	BRIP1	1.33
218883_S_AT	CENPU	1.60	219502_AT	NEIL3	1.33
218662_S_AT	NCAPG	1.60	201540_AT	FHL1	1.32
219148_AT	PBK	1.60	213226_AT	CCNA2	1.32
202338_AT	TK1	1.59	213599_AT	OIP5	1.31
212023_S_AT	MKI67	1.58	203560_AT	GGH	1.30
205394_AT	CHEK1	1.57	218355_AT	KIF4A	1.29
203213_AT	CDK1	1.56	205519_AT	WDR76	1.29
219306_AT	KIF15	1.56	203022_AT	RNASEH2A	1.28
201710_AT	MYBL2	1.55	204558_AT	RAD54L	1.28
201890_AT	RRM2	1.54	203362_S_AT	MAD2L1	1.27

Probeset	Symbol	PS-score	Probeset	Symbol	PS-score
209464_AT	AURKB	1.54	219555_S_AT	CENPN	1.27
204146_AT	RAD51AP1	1.54	205085_AT	ORC1	1.27
202580_X_AT	FOXM1	1.52	209173_AT	AGR2	1.26
218542_AT	CEP55	1.52	203764_AT	DLGAP5	1.25
203625_X_AT	SKP2	1.51	219588_S_AT	NCAPG2	1.25
213253_AT	SMC2	1.50	202345_S_AT	FABP5	1.25
204531_S_AT	BRCA1	1.49	211767_AT	GINS4	1.24
218741_AT	CENPM	1.48	202954_AT	UBE2C	1.23
219990_AT	E2F8	1.46	219320_AT	MYO19	1.22
202291_S_AT	MGP	1.45	218308_AT	TACC3	1.22
203755_AT	BUB1B	1.44	204159_AT	CDKN2C	1.21
204695_AT	CDC25A	1.44	202094_AT	BIRC5	1.21
212949_AT	NCAPH	1.44	218039_AT	NUSAP1	1.21
204033_AT	TRIP13	1.44	207871_S_AT	ST7	1.20
219493_AT	SHCBP1	1.43	205967_AT	HIST1H4C	1.19
204822_AT	TTK	1.43	205053_AT	PRIM1	1.19
213007_AT	FANCI	1.41	49452_AT	ACACB	1.19
213906_AT	MYBL1	1.39	219510_AT	POLQ	1.19
218115_AT	ASF1B	1.38	205229_S_AT	COCH	1.19
204709_S_AT	KIF23	1.37	206102_AT	GINS1	1.18
204727_AT	WDHD1	1.18	221521_S_AT	GINS2	1.02
211713_X_AT	KIAA0101	1.17	206205_AT	MPHOSPH9	1.01
207165_AT	HMMR	1.17	204766_S_AT	NUDT1	1.01
218009_S_AT	PRC1	1.16	219294_AT	CENPQ	1.01
209408_AT	KIF2C	1.16	221436_S_AT	CDCA3	1.01
204886_AT	PLK4	1.15	201202_AT	PCNA	1.01

Probeset	Symbol	PS-score	Probeset	Symbol	PS-score
205339_AT	STIL	1.15	219787_S_AT	ECT2	1.01
204641_AT	NEK2	1.15	218726_AT	HJURP	1.01
38158_AT	ESPL1	1.15	205176_S_AT	ITGB3BP	0.99
219650_AT	ERCC6L	1.14	221520_S_AT	CDCA8	0.99
204441_S_AT	POLA2	1.14	212801_AT	CIT	0.99
218350_S_AT	GMNN	1.13	204775_AT	CHAF1B	0.99
213647_AT	DNA2	1.13	202735_AT	EBP	0.98
220295_X_AT	DEPDC1	1.13	39966_AT	CSPG5	0.98
208002_S_AT	ACOT7	1.13	202854_AT	HPRT1	0.98
203145_AT	SPAG5	1.13	206632_S_AT	APOBEC3B	0.98
205543_AT	HSPA4L	1.13	204026_S_AT	ZWINT	0.97
203276_AT	LMNB1	1.12	210115_AT	RPL39L	0.97
202705_AT	CCNB2	1.12	204317_AT	GTSE1	0.97
219494_AT	RAD54B	1.12	202870_S_AT	CDC20	0.96
214727_AT	BRCA2	1.11	209832_S_AT	CDT1	0.96
217640_X_AT	SKA1	1.11	220840_S_AT	C1orf112	0.96
222039_AT	KIF18B	1.10	204510_AT	CDC7	0.96
200650_S_AT	LDHA	1.10	206140_AT	LHX2	0.95
201897_S_AT	CKS1B	1.10	222077_S_AT	RACGAP1	0.95
210416_S_AT	CHEK2	1.09	218585_S_AT	DTL	0.94
220060_S_AT	PARPBP	1.09	206316_S_AT	KNTC1	0.94
203696_S_AT	RFC2	1.09	201664_AT	SMC4	0.94
201563_AT	SORD	1.09	218755_AT	KIF20A	0.93
210052_S_AT	TPX2	1.07	203564_AT	FANCG	0.93
206364_AT	KIF14	1.07	218248_AT	FAM111A	0.93
205909_AT	POLE2	1.07	203270_AT	DTYMK	0.92

Probeset	Symbol	PS-score	Probeset	Symbol	PS-score
201477_S_AT	RRM1	1.07	219918_S_AT	ASPM	0.92
45633_AT	GINS3	1.05	205167_S_AT	CDC25C	0.92
209642_AT	BUB1	1.05	212621_AT	TMEM194A	0.91
219258_AT	TIPIN	1.04	201764_AT	TMEM106C	0.91
221677_S_AT	DONSON	1.04	201036_S_AT	HADH	0.90
213951_S_AT	PSMC3IP	1.03	204521_AT	FAM216A	0.90
211538_S_AT	HSPA2	1.02	220936_S_AT	H2AFJ	0.90
202330_S_AT	UNG	1.02	201969_AT	NASP	0.90
218979_AT	RMI1	1.02	207695_S_AT	IGSF1	0.90
218384_AT	CARHSP1	1.02	204623_AT	TFF3	0.90
214633_AT	SOX3	0.89	204835_AT	POLA1	0.82
203432_AT	TMPO	0.89	204767_S_AT	FEN1	0.82
218399_S_AT	CDCA4	0.89	221258_S_AT	KIF18A	0.81
218493_AT	SNRNP25	0.89	213454_AT	APITD1	0.81
215380_S_AT	GGCT	0.88	213520_AT	RECQL4	0.81
203422_AT	POLD1	0.88	218073_S_AT	NDC1	0.81
202107_S_AT	MCM2	0.88	205436_S_AT	H2AFX	0.80
202613_AT	CTPS1	0.88	203805_S_AT	FANCA	0.79
205470_S_AT	KLK11	0.87	221058_S_AT	CKLF	0.79
206280_AT	CDH18	0.87	205239_AT	AREG	0.79
204407_AT	TTF2	0.87	209608_S_AT	ACAT2	0.79
206533_AT	CHRNA5	0.87	218349_S_AT	ZWILCH	0.78
210792_X_AT	SIVA1	0.87	205395_S_AT	MRE11A	0.78
201774_S_AT	NCAPD2	0.87	221505_AT	ANP32E	0.78
213088_S_AT	DNAJC9	0.86	202483_S_AT	RANBP1	0.78
204244_S_AT	DBF4	0.86	209421_AT	MSH2	0.78

Probeset	Symbol	PS-score	Probeset	Symbol	PS-score
205628_AT	PRIM2	0.86	200762_AT	DPYSL2	0.78
212858_AT	PAQR4	0.86	217949_S_AT	VKORC1	0.77
221685_S_AT	SPDL1	0.86	201275_AT	FDPS	0.77
204023_AT	RFC4	0.85	209980_S_AT	SHMT1	0.77
205549_AT	PCP4	0.85	202468_S_AT	CTNNAL1	0.77
205347_S_AT	TMSB15A	0.85	204347_AT	AK4	0.77
208079_S_AT	AURKA	0.85	213761_AT	MDM1	0.77
201849_AT	BNIP3	0.85	203856_AT	VRK1	0.76
213534_S_AT	PASK	0.85	200934_AT	DEK	0.76
202779_S_AT	UBE2S	0.85	201923_AT	PRDX4	0.75
202240_AT	PLK1	0.85	204120_S_AT	ADK	0.75
203358_S_AT	EZH2	0.84	212097_AT	CAV1	0.74
212836_AT	POLD3	0.84	220239_AT	KLHL7	0.73
209189_AT	FOS	0.84	201860_S_AT	PLAT	0.73
209507_AT	RPA3	0.84	33322_I_AT	SFN	0.73
208096_S_AT	COL21A1	0.84	41037_AT	TEAD4	0.73
207392_X_AT	UGT2B15	0.84	202183_S_AT	KIF22	0.72
204127_AT	RFC3	0.83	204504_S_AT	HIRIP3	0.72
204900_X_AT	SAP30	0.83	204267_X_AT	PKMYT1	0.72
202309_AT	MTHFD1	0.83	221591_S_AT	FAM64A	0.72
208107_S_AT	LOC81691	0.83	205449_AT	SAC3D1	0.72
208808_S_AT	HMGB2	0.82	203302_AT	DCK	0.72
203209_AT	RFC5	0.82	205345_AT	BARD1	0.71
209267_S_AT	SLC39A8	0.82	201714_AT	TUBG1	0.71
213346_AT	TEX30	0.82	220466_AT	CCDC15	0.71
205677_S_AT	DLEU1	0.82	212192_AT	KCTD12	0.71

Probeset	Symbol	PS-score	Probeset	Symbol	PS-score
209680_S_AT	KIFC1	0.71	208795_S_AT	MCM7	0.63
204962_S_AT	CENPA	0.70	218619_S_AT	SUV39H1	0.63
202427_S_AT	MPC2	0.70	219698_S_AT	METTL4	0.62
218866_S_AT	POLR3K	0.70	209849_S_AT	RAD51C	0.62
200783_S_AT	STMN1	0.70	219960_S_AT	UCHL5	0.62
219763_AT	DENND1A	0.70	207267_S_AT	RIPPLY3	0.62
204492_AT	ARHGAP11A	0.69	218738_S_AT	RNF138	0.62
219403_S_AT	HPSE	0.69	206500_S_AT	MIS18BP1	0.62
214710_S_AT	CCNB1	0.69	203132_AT	RB1	0.61
201695_S_AT	PNP	0.68	216299_S_AT	XRCC3	0.61
201896_S_AT	PSRC1	0.68	217764_S_AT	RAB31	0.61
201489_AT	PPIF	0.67	219703_AT	MNS1	0.61
205431_S_AT	BMP5	0.67	203690_AT	TUBGCP3	0.61
201136_AT	PLP2	0.67	213334_X_AT	HAUS7	0.61
218383_AT	HAUS4	0.67	36545_S_AT	SFI1	0.60
201555_AT	MCM3	0.67	218447_AT	CMC2	0.60
214061_AT	TBC1D31	0.67	208644_AT	PARP1	0.60
204560_AT	FKBP5	0.67	204662_AT	CCP110	0.60
201013_S_AT	PAICS	0.67	203405_AT	PSMG1	0.60
218951_S_AT	PLCXD1	0.66	222216_S_AT	MRPL17	0.60
218108_AT	UBR7	0.66	218461_AT	GPN3	0.60
219105_X_AT	ORC6	0.66	205841_AT	JAK2	0.60
219419_AT	RBFA	0.66	205217_AT	TIMM8A	0.59
204776_AT	THBS4	0.66	206698_AT	XK	0.59
37512_AT	HSD17B6	0.66	219205_AT	SRR	0.59
213262_AT	SACS	0.65	222201_S_AT	CASP8AP2	0.59

Probeset	Symbol	PS-score	Probeset	Symbol	PS-score
206550_S_AT	NUP155	0.65	209259_S_AT	SMC3	0.57
219215_S_AT	SLC39A4	0.65	216952_S_AT	LMNB2	0.57
216026_S_AT	POLE	0.65	219038_AT	MORC4	0.57
205282_AT	LRP8	0.65	201020_AT	YWHAH	0.57
209572_S_AT	EED	0.65	219262_AT	SUV39H2	0.57
201697_S_AT	DNMT1	0.65	212195_AT	IL6ST	0.57
219490_S_AT	DCLRE1B	0.64	209263_X_AT	TSPAN4	0.57
205296_AT	RBL1	0.64	222036_S_AT	MCM4	0.57
216237_S_AT	MCM5	0.64	209272_AT	NAB1	0.57
206052_S_AT	SLBP	0.64	201577_AT	NME1	0.56
201111_AT	CSE1L	0.64	204256_AT	ELOVL6	0.56
202413_S_AT	USP1	0.64	218894_S_AT	MAGOHB	0.56
215136_S_AT	EXOSC8	0.64	205061_S_AT	EXOSC9	0.56
205598_AT	TRAIP	0.63	218768_AT	NUP107	0.56
219003_S_AT	MANEA	0.63	210053_AT	TAF5	0.56
205500_AT	C5	0.63	212247_AT	NUP205	0.56
209054_S_AT	WHSC1	0.56	214426_X_AT	CHAF1A	0.53
205862_AT	GREB1	0.55	206526_AT	RIBC2	0.53
214240_AT	GAL	0.55	204082_AT	PBX3	0.52
214430_AT	GLA	0.55	218911_AT	YEATS4	0.52
206176_AT	BMP6	0.55	220865_S_AT	PDSS1	0.52
219596_AT	THAP10	0.55	202911_AT	MSH6	0.52
213911_S_AT	H2AFZ	0.54	218404_AT	SNX10	0.52
213523_AT	CCNE1	0.54	221922_AT	GPSM2	0.52
218904_S_AT	C9orf40	0.54	218102_AT	DERA	0.52
218209_S_AT	RPRD1A	0.54	213664_AT	SLC1A1	0.52

Probeset	Symbol	PS-score	Probeset	Symbol	PS-score
205548_S_AT	BTG3	0.54	213379_AT	COQ2	0.52
214086_S_AT	PARP2	0.54	219004_S_AT	MIS18A	0.51
209825_S_AT	UCK2	0.54	218564_AT	RFWD3	0.51
218494_S_AT	SLC2A4RG	0.54	201115_AT	POLD2	0.50
214723_X_AT	ANKRD36	0.54	213677_S_AT	PMS1	0.50
205260_S_AT	ACYP1	0.53			

Appendix III

Case study of GATA3 using E2S database

Homepage of E2S

By clicking of 'Go to Analysis' or 'analysis' tab, E2S redirects to input page



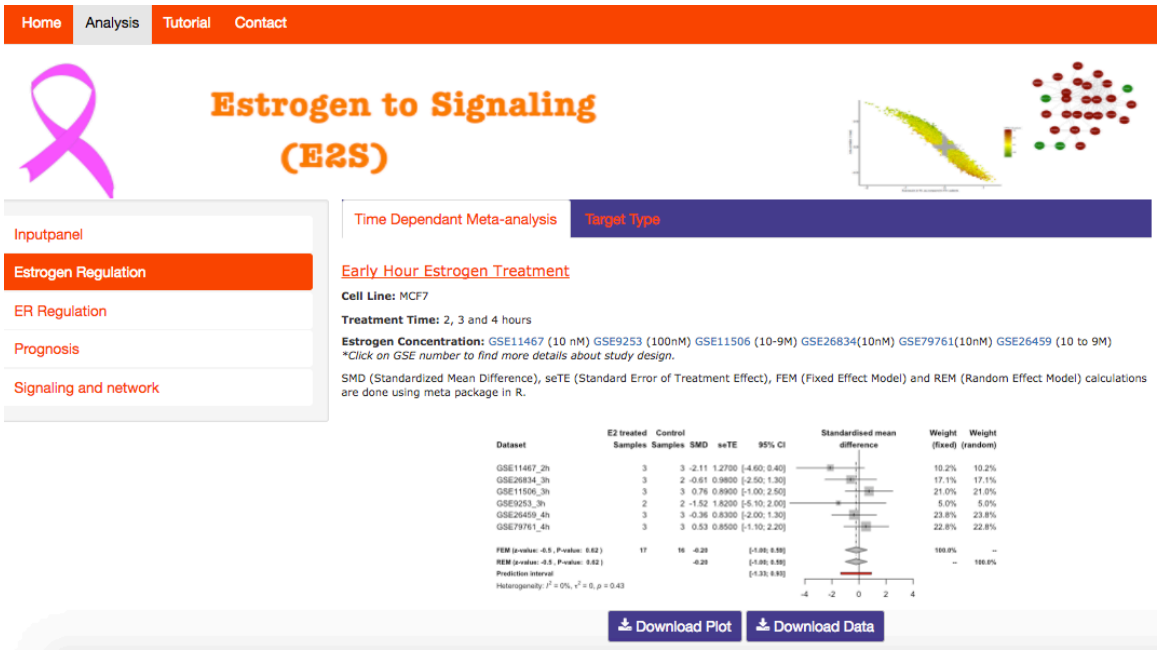
User can enter a valid gene symbol (e.g., GATA3) for querying in E2S database.



If entered gene symbol is associated with multiple probesets, E2S visualizes all probesets corresponding to the gene in drop- down menu and allows user to choose any specific probeset or jetset best probeset.

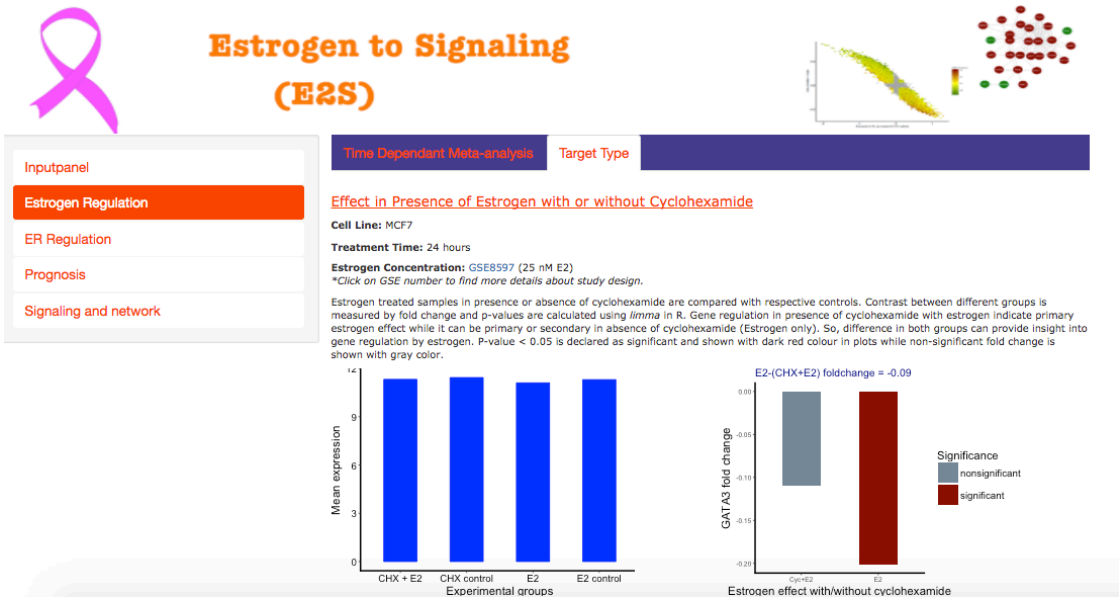


Estrogen Regulation tab shows the meta-analysis result for early and late hour estrogen exposure under the same tab with two separate forest plots.





Under the estrogen regulation module, 'Target Type' tab is designed to evaluate effect of protein synthesis inhibitor 'cycloheximide' on estrogen mediated effects.



Module II: ER regulation

Sub-module: Single dataset analysis

Correlation of queried gene with ESR1 mRNA expression and ER-status in any individual data can be studied by selecting dataset of interest.

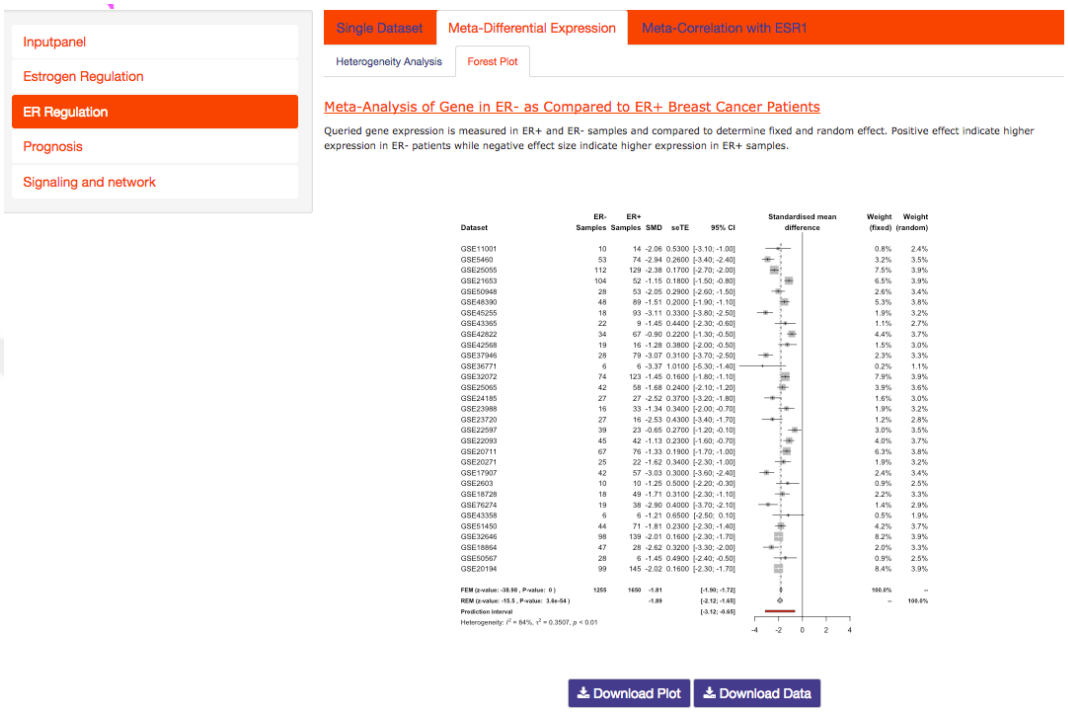


Sub-module: Meta-differential expression

ER-/ER+ meta-analysis returns heterogeneity measures for entered gene by visualization of funnel plot. Significance of publication bias is given by Egger's test. P<0.05 indicate potential publication bias while p>0.05 rejects the probability of publication bias and indicate that observed heterogeneity might be due to any other reason.

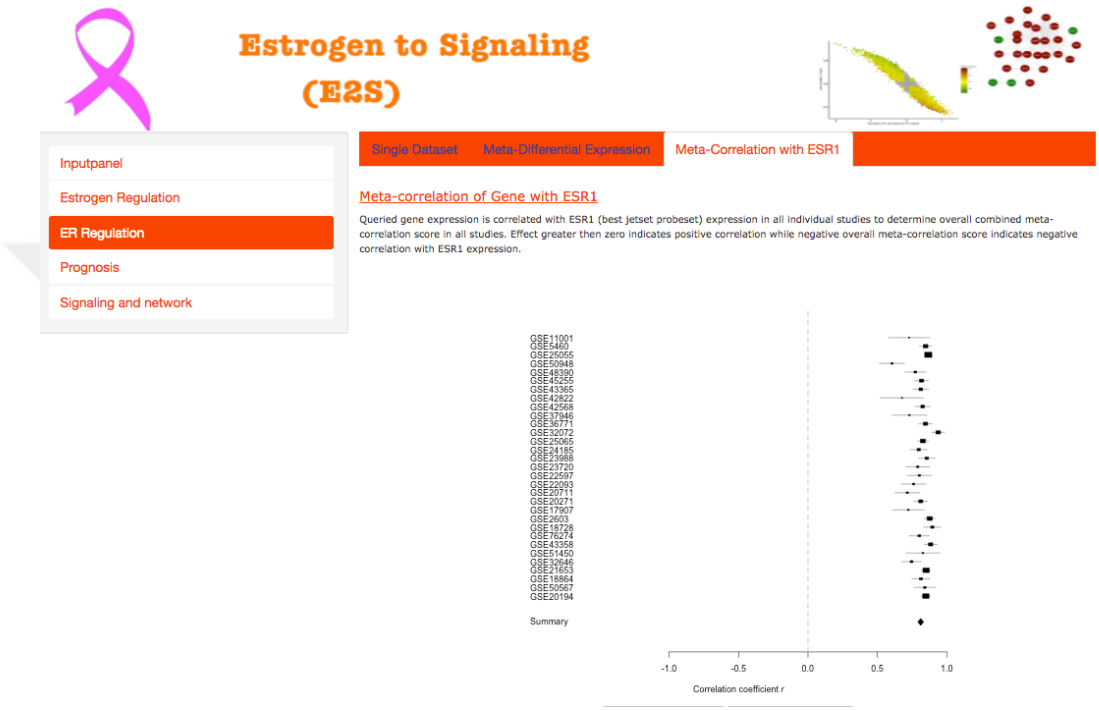


ER-/ER+ meta-differential expression analysis results are visualized by a forest plot with positive scores indicating upregulation in ER- breast cancer patients and vice-versa.



Sub-module: Meta-Correlation with ESR1

'Meta-Correlation with ESR1' tab performs meta-correlation analysis between queried gene and ESR1 mRNA expression.



Module III: Prognosis

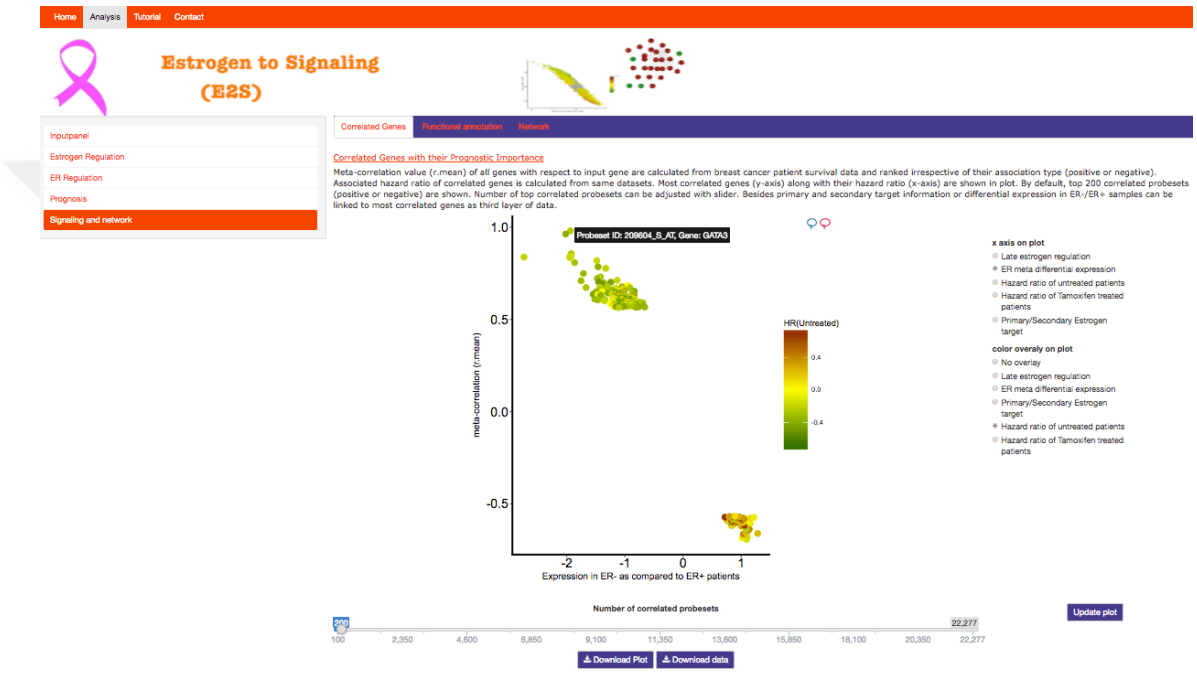
Meta-survival analysis of untreated and tamoxifen treated patients is visualized by forest plot. Meta-analysis scores>0 lying under the significance threshold ($p<0.05$) indicate potential bad prognosis associated with patient and vice-versa.





Module IV: Signaling and network
Sub-module: Correlated genes

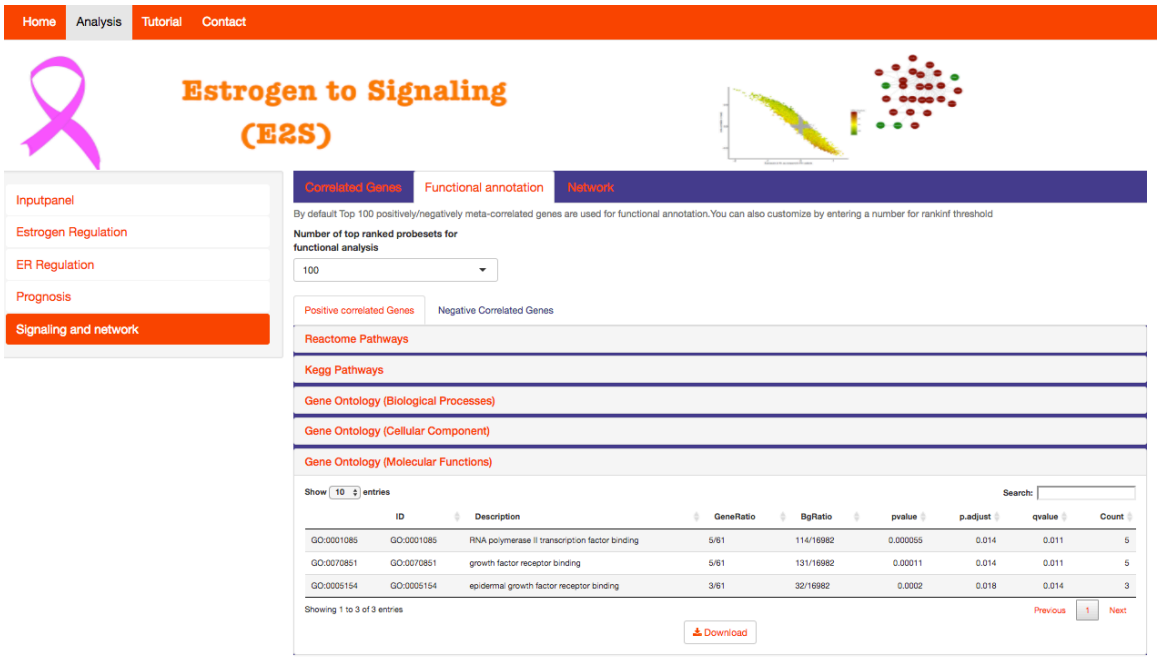
Module performs met-correlation analysis and visualized most to least correlated genes on y-axis. Number of displayed gene is adjustable by slider at the bottom. Different layer of already stored meta-analysis output can be overlaid on data.



Sub-module: Functional annotation

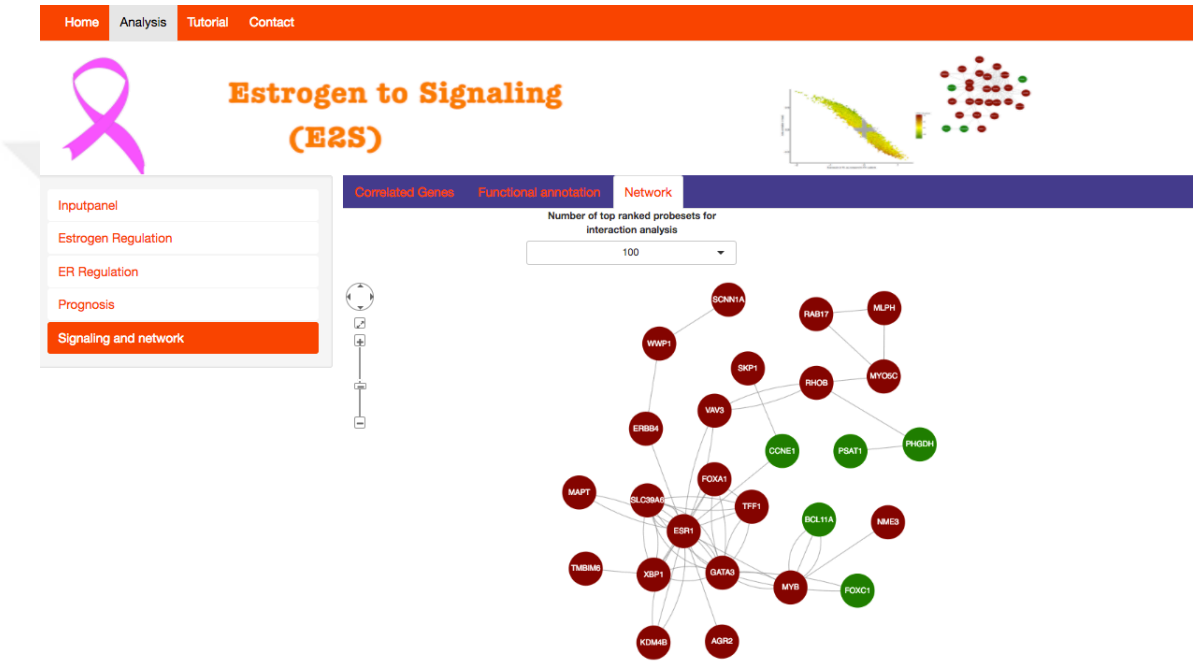
Functional annotation offers the user to choose any database (KEGG, Gene Ontology and Reactome) for the enrichment analysis of top correlated genes. Number of top correlated genes can be adjusted between 50-500.





Sub-module: Signal and Network Module

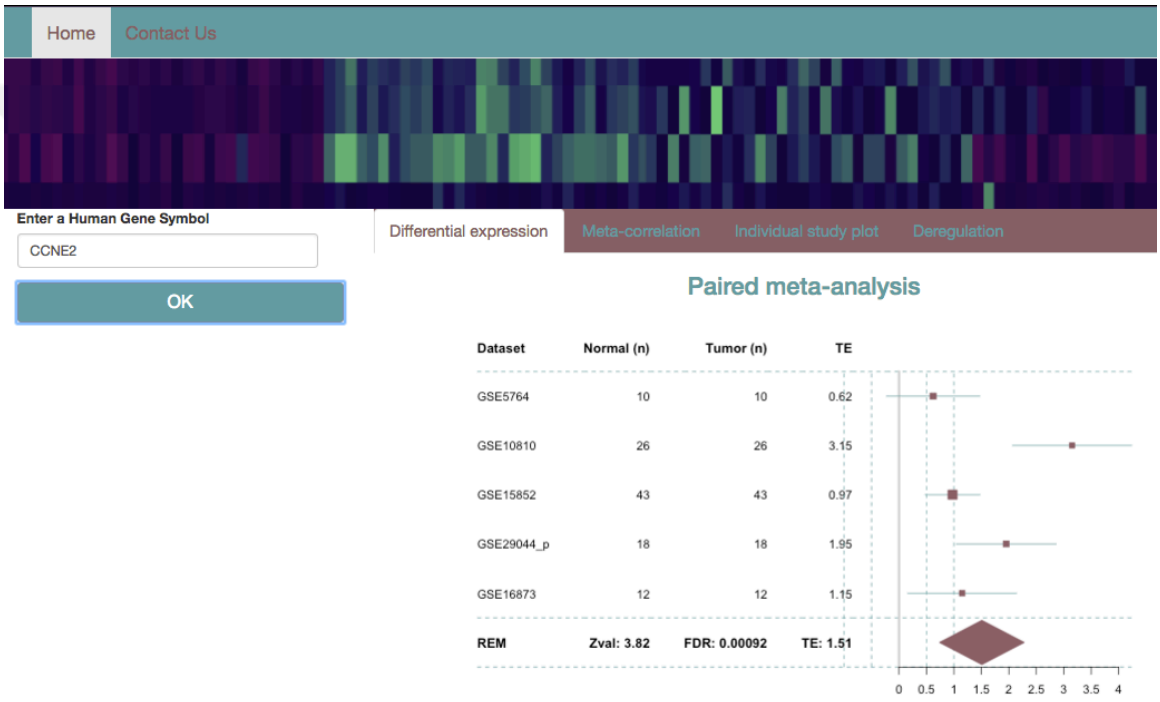
This module enables visualization of protein-protein interaction for top correlated profile and is also adjustable between 50-500 probesets. Maroon color indicates positive correlation with queried gene and green indicate negative correlation. Edges represent gene-gene interaction retrieved from STRING database.



Appendix IV

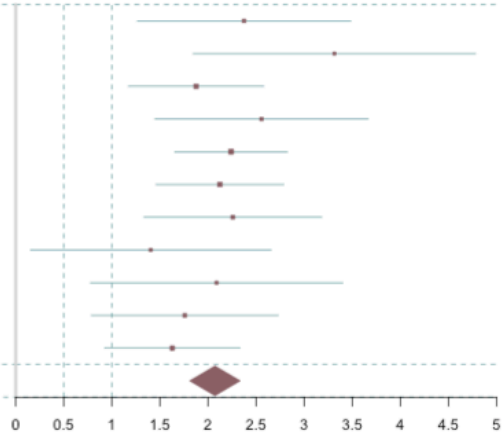
Case Study of CCNE2 gene using DualExpBC

Module I: Differential expression



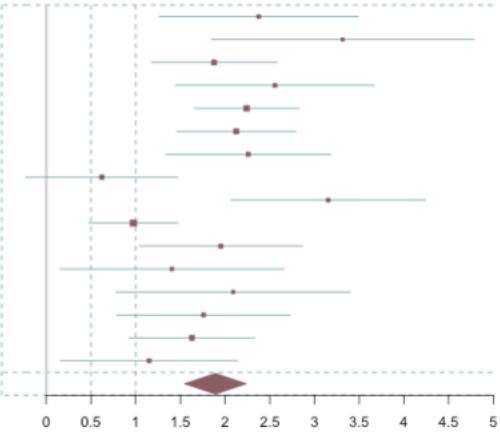
Unpaired meta-analysis

Dataset	Normal (n)	Tumor (n)	TE
GSE17907	4	51	2.38
GSE21422	5	14	3.31
GSE29431	12	54	1.88
GSE61304	4	58	2.56
GSE42568	17	104	2.24
GSE45827	11	130	2.12
GSE3744	7	40	2.26
GSE22544	4	11	1.41
GSE31192	6	8	2.09
GSE50567	6	29	1.76
GSE29044_u	18	24	1.63
REM	**Zval: 15.62**	**FDR: 9.4e-20**	**TE: 2.07**

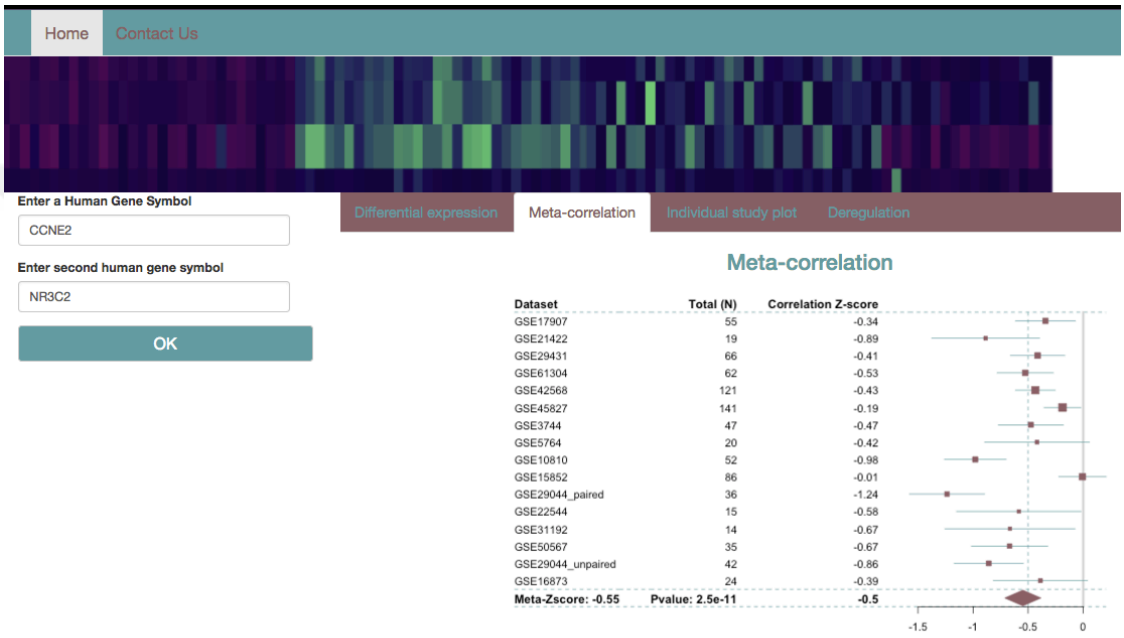


Combined paired and unpaired meta-analysis

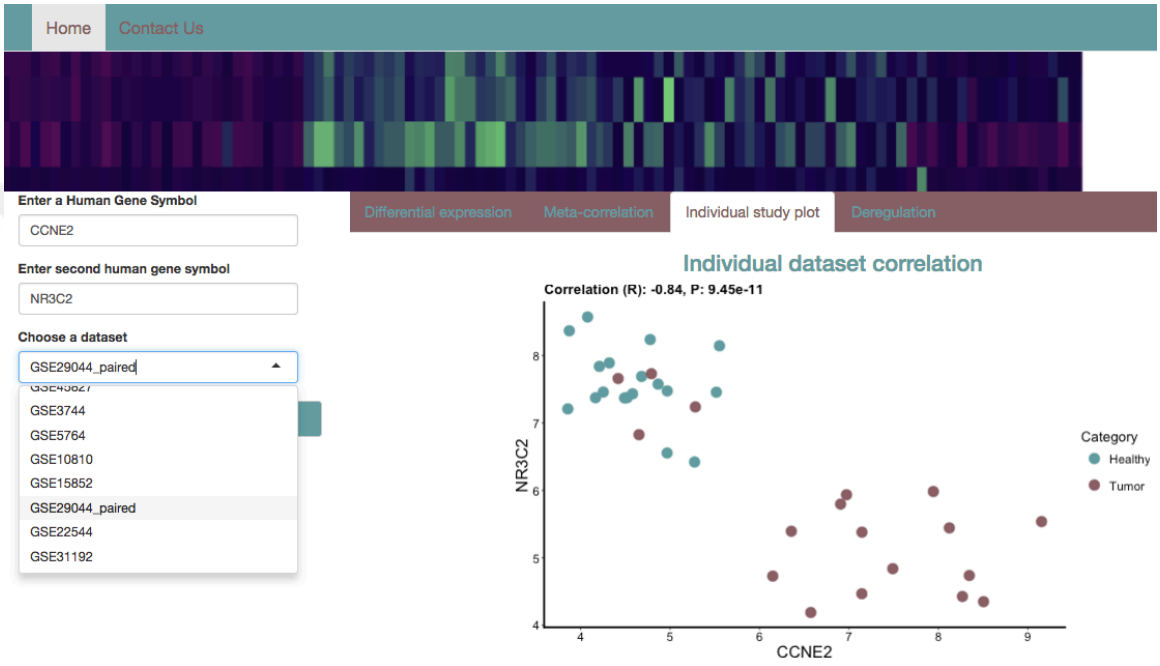
Dataset	Normal (n)	Tumor (n)	TE
GSE17907	4	51	2.38
GSE21422	5	14	3.31
GSE29431	12	54	1.88
GSE61304	4	58	2.56
GSE42568	17	104	2.24
GSE45827	11	130	2.12
GSE3744	7	40	2.26
GSE5764	10	10	0.62
GSE10810	26	26	3.15
GSE15852	43	43	0.97
GSE29044_p	18	18	1.95
GSE22544	4	11	1.41
GSE31192	6	8	2.09
GSE50567	6	29	1.76
GSE29044_u	18	24	1.63
GSE16873	12	12	1.15
REM	**Zval: 10.76**	**FDR: 9.7e-20**	**TE: 1.89**



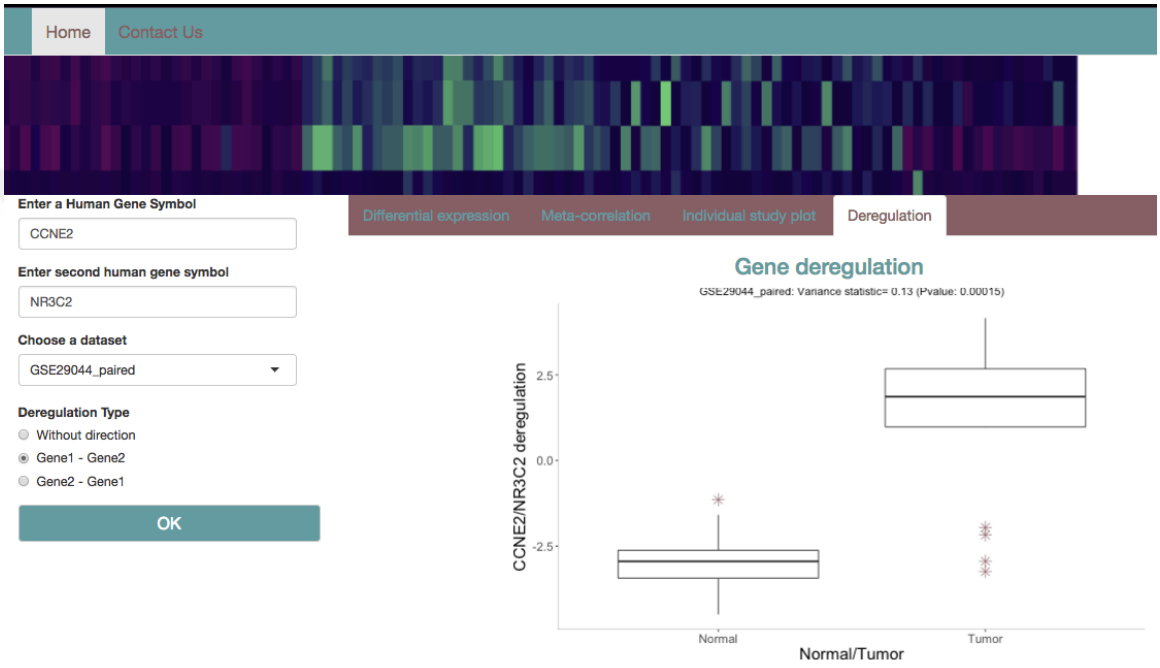
Module II: Meta-correlation: This module requires two input genes to perform meta-correlation analysis in breast cancer cohort.



Module III: Individual Study Plot: In addition to meta-analysis, user can also look at correlation between both genes in a selected dataset where each dot represents one sample and is labeled by its phenotype, i.e., normal/tumor.



Module IV: Deregulation: Like individual study plot, deregulation tab acceptstwo input gene and selection of a dataset. While third input is direction of deregulation (mentioned in Chapter 5).



REFERENCES

1. Schena, M., et al., *Quantitative monitoring of gene expression patterns with a complementary DNA microarray.* Science, 1995. **270**(5235): p. 467-70.
2. Irizarry, R.A., et al., *Multiple-laboratory comparison of microarray platforms.* Nat Methods, 2005. **2**(5): p. 345-50.
3. Russo, G., C. Zegar, and A. Giordano, *Advantages and limitations of microarray technology in human cancer.* Oncogene, 2003. **22**(42): p. 6497-507.
4. Tumor Analysis Best Practices Working, G., *Expression profiling--best practices for data generation and interpretation in clinical trials.* Nat Rev Genet, 2004. **5**(3): p. 229-37.
5. Mirnics, K. and J. Pevsner, *Progress in the use of microarray technology to study the neurobiology of disease.* Nature Neuroscience, 2004. **7**(5): p. 434-439.
6. Schena, M., et al., *Parallel human genome analysis: microarray-based expression monitoring of 1000 genes.* Proc Natl Acad Sci U S A, 1996. **93**(20): p. 10614-9.
7. Lockhart, D.J., et al., *Expression monitoring by hybridization to high-density oligonucleotide arrays.* Nat Biotechnol, 1996. **14**(13): p. 1675-80.
8. Xiang, C.C. and Y. Chen, *cDNA microarray technology and its applications.* Biotechnol Adv, 2000. **18**(1): p. 35-46.
9. Kuhn, E., *From library screening to microarray technology: Strategies to determine gene expression profiles and to identify differentially regulated genes in plants.* Annals of Botany, 2001. **87**(2): p. 139-155.
10. Kim, I.J., et al., *Oligonucleotide microarray analysis of distinct gene expression patterns in colorectal cancer tissues harboring BRAF and K-ras mutations.* Carcinogenesis, 2006. **27**(3): p. 392-404.
11. Nuwaysir, E.F., et al., *Gene expression analysis using oligonucleotide arrays produced by maskless photolithography.* Genome Research, 2002. **12**(11): p. 1749-1755.
12. Lee, C. and M. Roy, *Analysis of alternative splicing with microarrays: successes and challenges.* Genome Biol, 2004. **5**(7): p. 231.
13. Eroshenko, N., et al., *Gene Assembly from Chip-Synthesized Oligonucleotides.* Curr Protoc Chem Biol, 2012. **2012**.
14. Altman, N.S. and J. Hua, *Extending the loop design for two-channel microarray experiments.* Genet Res, 2006. **88**(3): p. 153-63.

15. Guo, Y., et al., *Large scale comparison of gene expression levels by microarrays and RNAseq using TCGA data.* PLoS One, 2013. **8**(8): p. e71462.

16. Lee, M.L., et al., *Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations.* Proc Natl Acad Sci U S A, 2000. **97**(18): p. 9834-9.

17. Yang, Y.H. and T. Speed, *Design issues for cDNA microarray experiments.* Nat Rev Genet, 2002. **3**(8): p. 579-88.

18. Woo, Y., et al., *A comparison of cDNA, oligonucleotide, and Affymetrix GeneChip gene expression microarray platforms.* J Biomol Tech, 2004. **15**(4): p. 276-84.

19. Chou, C.C., et al., *Optimization of probe length and the number of probes per gene for optimal microarray analysis of gene expression.* Nucleic Acids Res, 2004. **32**(12): p. e99.

20. Barczak, A., et al., *Spotted long oligonucleotide arrays for human gene expression analysis.* Genome Res, 2003. **13**(7): p. 1775-85.

21. Lipshutz, R.J., et al., *High density synthetic oligonucleotide arrays.* Nat Genet, 1999. **21**(1 Suppl): p. 20-4.

22. Suzuki, S., et al., *Experimental optimization of probe length to increase the sequence specificity of high-density oligonucleotide microarrays.* BMC Genomics, 2007. **8**: p. 373.

23. Tan, J.C., et al., *An optimized microarray platform for assaying genomic variation in Plasmodium falciparum field populations.* Genome Biol, 2011. **12**(4): p. R35.

24. Jarvinen, A.K., et al., *Are data from different gene expression microarray platforms comparable?* Genomics, 2004. **83**(6): p. 1164-8.

25. Wolber, P.K., et al., *The Agilent in situ-synthesized microarray platform.* Methods Enzymol, 2006. **410**: p. 28-57.

26. Gershon, D., *Microarray technology: an array of opportunities.* Nature, 2002. **416**(6883): p. 885-91.

27. Chibon, F., *Cancer gene expression signatures - the rise and fall?* Eur J Cancer, 2013. **49**(8): p. 2000-9.

28. Dawany, N.B., W.N. Dampier, and A. Tozeren, *Large-scale integration of microarray data reveals genes and pathways common to multiple cancer types.* Int J Cancer, 2011. **128**(12): p. 2881-91.

29. Wittner, B.S., et al., *Analysis of the MammaPrint breast cancer assay in a predominantly postmenopausal cohort.* Clin Cancer Res, 2008. **14**(10): p. 2988-93.

30. Roepman, P., et al., *An expression profile for diagnosis of lymph node metastases from primary head and neck squamous cell carcinomas.* Nat Genet, 2005. **37**(2): p. 182-6.

31. van Laar, R.K., et al., *Implementation of a novel microarray-based diagnostic test for cancer of unknown primary.* Int J Cancer, 2009. **125**(6): p. 1390-7.

32. Bittner, M., et al., *Molecular classification of cutaneous malignant melanoma by gene expression profiling.* Nature, 2000. **406**(6795): p. 536-40.

33. Perou, C.M., et al., *Molecular portraits of human breast tumours.* Nature, 2000. **406**(6797): p. 747-52.

34. Chibon, F., et al., *Validated prediction of clinical outcome in sarcomas and multiple types of cancer on the basis of a gene expression signature related to genome complexity.* Nat Med, 2010. **16**(7): p. 781-7.

35. Brodin, B.A., *CINSARC: a new look into an old concept gives hope for new treatments for synovial sarcomas.* Transl Pediatr, 2013. **2**(2): p. 70-2.

36. Ye, Q.H., et al., *Predicting hepatitis B virus-positive metastatic hepatocellular carcinomas using gene expression profiling and supervised machine learning.* Nature Medicine, 2003. **9**(4): p. 416-423.

37. Allison, D.B., et al., *Microarray data analysis: from disarray to consolidation and consensus.* Nat Rev Genet, 2006. **7**(1): p. 55-65.

38. Wang, X., S. Ghosh, and S.W. Guo, *Quantitative quality control in microarray image processing and data acquisition.* Nucleic Acids Res, 2001. **29**(15): p. E75-5.

39. Eijssen, L.M., et al., *User-friendly solutions for microarray quality control and pre-processing on ArrayAnalysis.org.* Nucleic Acids Res, 2013. **41**(Web Server issue): p. W71-6.

40. Reimers, M. and J.N. Weinstein, *Quality assessment of microarrays: visualization of spatial artifacts and quantitation of regional biases.* BMC Bioinformatics, 2005. **6**: p. 166.

41. Wilson, C.L. and C.J. Miller, *Simpleaffy: a BioConductor package for Affymetrix Quality Control and data analysis.* Bioinformatics, 2005. **21**(18): p. 3683-5.

42. Cordero, F., M. Botta, and R.A. Calogero, *Microarray data analysis and mining approaches.* Brief Funct Genomic Proteomic, 2007. **6**(4): p. 265-81.

43. Durinck, S., *Pre-processing of microarray data and analysis of differential expression.* Methods Mol Biol, 2008. **452**: p. 89-110.

44. Brettschneider J, C.F., Bolstad BM, and Speed TP, *Quality assessment for short oligonucleotide arrays.* Technometrics (in press), 2007.

45. Parman C, H.C.a.G.R., *affyQCReport: QC Report Generation for affyBatch objects*. 2017.

46. Kauffmann, A., R. Gentleman, and W. Huber, *arrayQualityMetrics--a bioconductor package for quality assessment of microarray data.* Bioinformatics, 2009. **25**(3): p. 415-6.

47. Gatto, L., *yaqcaffy: Affymetrix expression data quality control and reproducibility analysis*. 2017.

48. Irizarry, R.A., et al., *Summaries of Affymetrix GeneChip probe level data.* Nucleic Acids Res, 2003. **31**(4): p. e15.

49. Hubbell, E., W.M. Liu, and R. Mei, *Robust estimators for expression analysis.* Bioinformatics, 2002. **18**(12): p. 1585-92.

50. Wu, Z.J., et al., *A model-based background adjustment for oligonucleotide expression arrays.* Journal of the American Statistical Association, 2004. **99**(468): p. 909-917.

51. Gautier, L., et al., *affy - analysis of Affymetrix GeneChip data at the probe level.* Bioinformatics, 2004. **20**(3): p. 307-315.

52. DeRisi, J., et al., *Use of a cDNA microarray to analyse gene expression patterns in human cancer.* Nature Genetics, 1996. **14**(4): p. 457-460.

53. Witt, P.L. and P. McGrain, *Comparing two sample means t tests.* Phys Ther, 1985. **65**(11): p. 1730-3.

54. Huang, J.Z., *Functional ANOVA models for generalized regression.* Journal of Multivariate Analysis, 1998. **67**(1): p. 49-71.

55. Tusher, V.G., R. Tibshirani, and G. Chu, *Significance analysis of microarrays applied to the ionizing radiation response.* Proc Natl Acad Sci U S A, 2001. **98**(9): p. 5116-21.

56. Ritchie, M.E., et al., *limma powers differential expression analyses for RNA-sequencing and microarray studies.* Nucleic Acids Res, 2015. **43**(7): p. e47.

57. Lee, H.K., et al., *Coexpression analysis of human genes across many microarray data sets.* Genome Res, 2004. **14**(6): p. 1085-94.

58. Langfelder, P. and S. Horvath, *WGCNA: an R package for weighted correlation network analysis.* BMC Bioinformatics, 2008. **9**: p. 559.

59. Kim, S.Y. and D.J. Volsky, *PAGE: parametric analysis of gene set enrichment.* BMC Bioinformatics, 2005. **6**: p. 144.

60. Butte, A., *The use and analysis of microarray data.* Nature Reviews Drug Discovery, 2002. **1**(12): p. 951-960.

61. Armstrong, S.A., et al., *MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia.* Nature Genetics, 2002. **30**(1): p. 41-47.

62. Czajkowski, M., M. Grzes, and M. Kretowski, *Multi-test decision tree and its application to microarray data classification.* Artif Intell Med, 2014. **61**(1): p. 35-44.

63. Dobbin, K.K. and R.M. Simon, *Optimally splitting cases for training and testing high dimensional classifiers.* BMC Med Genomics, 2011. **4**: p. 31.

64. Ringner, M., C. Peterson, and J. Khan, *Analyzing array data using supervised methods.* Pharmacogenomics, 2002. **3**(3): p. 403-415.

65. Gao, X. and P.X. Song, *Nonparametric tests for differential gene expression and interaction effects in multi-factorial microarray experiments.* BMC Bioinformatics, 2005. **6**: p. 186.

66. Tanizaki, H., *Power comparison of non-parametric tests: Small-sample properties from Monte Carlo experiments.* Journal of Applied Statistics, 1997. **24**(5): p. 603-632.

67. Pan, W., *A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments.* Bioinformatics, 2002. **18**(4): p. 546-54.

68. Bewick, V., L. Cheek, and J. Ball, *Statistics review 12: survival analysis.* Crit Care, 2004. **8**(5): p. 389-94.

69. van Wieringen, W.N., et al., *Survival prediction using gene expression data: A review and comparison.* Computational Statistics & Data Analysis, 2009. **53**(5): p. 1590-1603.

70. Schoenfeld, D.A. and A.A. Tsiatis, *A Modified Log Rank Test for Highly Stratified Data.* Biometrika, 1987. **74**(1): p. 167-175.

71. Cox, D., *Citation-Classic - Regression-Models and Life-Tables.* Current Contents/Agriculture Biology & Environmental Sciences, 1986(42): p. 16-16.

72. Rosenthal, R., et al., *How to write a surgical clinical research protocol: literature review and practical guide.* American Journal of Surgery, 2014. **207**(2): p. 299-312.

73. Schroder, M.S., et al., *survcomp: an R/Bioconductor package for performance assessment and comparison of survival models.* Bioinformatics, 2011. **27**(22): p. 3206-8.

74. Durinck, S., et al., *Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt.* Nat Protoc, 2009. **4**(8): p. 1184-91.

75. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.

76. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes.* Nucleic Acids Res, 2000. **28**(1): p. 27-30.

77. Croft, D., et al., *Reactome: a database of reactions, pathways and biological processes.* Nucleic Acids Res, 2011. **39**(Database issue): p. D691-7.

78. Wilson, E.B. and M.M. Hilferty, *The Distribution of Chi-Square.* Proc Natl Acad Sci U S A, 1931. **17**(12): p. 684-8.

79. Wu, T., *An Accurate Computation of the Hypergeometric Distribution Function.* Acm Transactions on Mathematical Software, 1993. **19**(1): p. 33-43.

80. Sloane, D. and S.P. Morgan, *An introduction to categorical data analysis.* Annual Review of Sociology, 1996. **22**: p. 351-375.

81. Tipney, H. and L. Hunter, *An introduction to effective use of enrichment analysis software.* Hum Genomics, 2010. **4**(3): p. 202-6.

82. Rivals, I., et al., *Enrichment or depletion of a GO category within a class of genes: which test?* Bioinformatics, 2007. **23**(4): p. 401-7.

83. Huang da, W., B.T. Sherman, and R.A. Lempicki, *Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.* Nucleic Acids Res, 2009. **37**(1): p. 1-13.

84. Beissbarth, T. and T.P. Speed, *GOstat: find statistically overrepresented Gene Ontologies within a group of genes.* Bioinformatics, 2004. **20**(9): p. 1464-1465.

85. Maere, S., K. Heymans, and M. Kuiper, *BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks.* Bioinformatics, 2005. **21**(16): p. 3448-9.

86. Liberzon, A., et al., *The Molecular Signatures Database (MSigDB) hallmark gene set collection.* Cell Syst, 2015. **1**(6): p. 417-425.

87. Huang, D.W., B.T. Sherman, and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.* Nature Protocols, 2009. **4**(1): p. 44-57.

88. Yu, G., et al., *clusterProfiler: an R package for comparing biological themes among gene clusters.* OMICS, 2012. **16**(5): p. 284-7.

89. Yu, G. and Q.Y. He, *ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization.* Mol Biosyst, 2016. **12**(2): p. 477-9.

90. Sheffield, N.C., et al., *Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions.* Genome Res, 2013. **23**(5): p. 777-88.

91. Janky, R., et al., *iRegulon: from a gene list to a gene regulatory network using large motif and track collections.* PLoS Comput Biol, 2014. **10**(7): p. e1003731.

92. Geertz, M. and S.J. Maerkl, *Experimental strategies for studying transcription factor-DNA binding specificities.* Brief Funct Genomics, 2010. **9**(5-6): p. 362-73.

93. Siddharthan, R., *Dinucleotide weight matrices for predicting transcription factor binding sites: generalizing the position weight matrix.* PLoS One, 2010. **5**(3): p. e9722.

94. Wingender, E., et al., *TRANSFAC: a database on transcription factors and their DNA binding sites.* Nucleic Acids Res, 1996. **24**(1): p. 238-41.

95. Quandt, K., et al., *MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data.* Nucleic Acids Res, 1995. **23**(23): p. 4878-84.

96. Chen, Q.K., G.Z. Hertz, and G.D. Stormo, *MATRIX SEARCH 1.0: a computer program that scans DNA sequences for transcriptional elements using a database of weight matrices.* Comput Appl Biosci, 1995. **11**(5): p. 563-6.

97. Loots, G.G. and I. Ovcharenko, *rVISTA 2.0: evolutionary analysis of transcription factor binding sites.* Nucleic Acids Res, 2004. **32**(Web Server issue): p. W217-21.

98. Prestridge, D.S., *SIGNAL SCAN: a computer program that scans DNA sequences for eukaryotic transcriptional elements.* Comput Appl Biosci, 1991. **7**(2): p. 203-6.

99. Bulyk, M.L., *Computational prediction of transcription-factor binding site locations.* Genome Biol, 2003. **5**(1): p. 201.

100. Shannon, P., et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks.* Genome Res, 2003. **13**(11): p. 2498-504.

101. Macfarlane, L.A. and P.R. Murphy, *MicroRNA: Biogenesis, Function and Role in Cancer.* Curr Genomics, 2010. **11**(7): p. 537-61.

102. Griffiths-Jones, S., et al., *miRBase: microRNA sequences, targets and gene nomenclature.* Nucleic Acids Res, 2006. **34**(Database issue): p. D140-4.

103. Lewis, B.P., et al., *Prediction of mammalian microRNA targets.* Cell, 2003. **115**(7): p. 787-98.

104. Sethupathy, P., B. Corda, and A.G. Hatzigeorgiou, *TarBase: A comprehensive database of experimentally supported animal microRNA targets.* RNA, 2006. **12**(2): p. 192-7.

105. Wang, X., *miRDB: a microRNA target prediction and functional annotation database with a wiki interface.* RNA, 2008. **14**(6): p. 1012-7.

106. Friedman, R.C., et al., *Most mammalian mRNAs are conserved targets of microRNAs.* Genome Research, 2009. **19**(1): p. 92-105.

107. De Preter, K., et al., *Positional gene enrichment analysis of gene sets for high-resolution identification of overrepresented chromosomal regions.* Nucleic Acids Res, 2008. **36**(7): p. e43.

108. D.Sarkar, F.R.a., *Category: Category Analysis*. 2017.

109. Rigaut, G., et al., *A generic protein purification method for protein complex characterization and proteome exploration.* Nat Biotechnol, 1999. **17**(10): p. 1030-2.

110. Fields, S. and O. Song, *A novel genetic system to detect protein-protein interactions.* Nature, 1989. **340**(6230): p. 245-6.

111. Lehne, B. and T. Schlitt, *Protein-protein interaction databases: keeping up with growing interactomes.* Hum Genomics, 2009. **3**(3): p. 291-7.

112. Szklarczyk, D., et al., *The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible.* Nucleic Acids Res, 2017. **45**(D1): p. D362-D368.

113. Bindea, G., et al., *ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks.* Bioinformatics, 2009. **25**(8): p. 1091-3.

114. Sanz-Pamplona, R., et al., *Tools for protein-protein interaction network analysis in cancer research.* Clin Transl Oncol, 2012. **14**(1): p. 3-14.

115. Rhodes, D.R., et al., *Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer.* Cancer Res, 2002. **62**(15): p. 4427-33.

116. Hong, F. and R. Breitling, *A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments.* Bioinformatics, 2008. **24**(3): p. 374-82.

117. Grutzmann, R., et al., *Meta-analysis of microarray data on pancreatic cancer defines a set of commonly dysregulated genes.* Oncogene, 2005. **24**(32): p. 5079-88.

118. Ramasamy, A., et al., *Key issues in conducting a meta-analysis of gene expression microarray datasets.* PLoS Med, 2008. **5**(9): p. e184.

119. Sweeney, T.E., et al., *Methods to increase reproducibility in differential gene expression via meta-analysis.* Nucleic Acids Res, 2017. **45**(1): p. e1.

120. Barrett, J.K., et al., *Two-stage meta-analysis of survival data from individual participants using percentile ratios.* Statistics in Medicine, 2012. **31**(30): p. 4296-4308.

121. Tseng, G.C., D. Ghosh, and E. Feingold, *Comprehensive literature review and statistical considerations for microarray meta-analysis.* Nucleic Acids Res, 2012. **40**(9): p. 3785-99.

122. Won, S.H., et al., *Choosing an optimal method to combine P-values.* Statistics in Medicine, 2009. **28**(11): p. 1537-1553.

123. Choi, J.K., et al., *Combining multiple microarray studies and modeling interstudy variation.* Bioinformatics, 2003. **19 Suppl 1**: p. i84-90.

124. Carratore FD, A., Wittner B, Breitling R and Battke F, *RankProd: Rank Product method for identifying differentially expressed genes with application in meta-analysis*. 2016. p. R package.

125. Taminau, J., et al., *Comparison of merging and meta-analysis as alternative approaches for integrative gene expression analysis.* ISRN Bioinform, 2014. **2014**: p. 345106.

126. RA, F., *Statistical Methods for Research Workers*. 1950.

127. Stouffer SA, S.E., Devinney LC, Star SA, Williams RM, *The American soldier: adjustment during army life*. 1949, Princeton University Press: Princeton, NJ.

128. Campain, A. and Y.H. Yang, *Comparison study of microarray meta-analysis methods.* BMC Bioinformatics, 2010. **11**: p. 408.

129. Marot, G., et al., *Moderated effect size and P-value combinations for microarray meta-analyses.* Bioinformatics, 2009. **25**(20): p. 2692-9.

130. Moreau, Y., et al., *Comparison and meta-analysis of microarray data: from the bench to the computer desk.* Trends Genet, 2003. **19**(10): p. 570-7.

131. Hedges, L.V. and J.L. Vevea, *Fixed- and random-effects models in meta-analysis.* Psychological Methods, 1998. **3**(4): p. 486-504.

132. Breitling, R. and P. Herzyk, *Rank-based methods as a non-parametric alternative of the T-statistic for the analysis of biological microarray data.* J Bioinform Comput Biol, 2005. **3**(5): p. 1171-89.

133. Lagani, V., et al., *A comparative evaluation of data-merging and meta-analysis methods for reconstructing gene-gene interactions.* BMC Bioinformatics, 2016. **17 Suppl 5**: p. 194.

134. Lusa L, G.R.a.R.M., *GeneMeta: MetaAnalysis for High Throughput Experiments*. I2017: R package.

135. Choi, H., *metaArray: Integration of Microarray Data for Meta-analysis*. 2017, R package Version 1.54.0.

136. Hong, F., et al., *RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis.* Bioinformatics, 2006. **22**(22): p. 2825-7.

137. Wang, X., et al., *An R package suite for microarray meta-analysis in quality control, differentially expressed gene analysis and pathway enrichment detection.* Bioinformatics, 2012. **28**(19): p. 2534-6.

138. Shen, K. and G.C. Tseng, *Meta-analysis for pathway enrichment analysis when combining multiple genomic studies.* Bioinformatics, 2010. **26**(10): p. 1316-1323.

139. Carrera, J., et al., *Reverse-engineering the Arabidopsis thaliana transcriptional network under changing environmental conditions.* Genome Biol, 2009. **10**(9): p. R96.

140. Ucar, D., et al., *Construction of a reference gene association network from multiple profiling data: application to data analysis.* Bioinformatics, 2007. **23**(20): p. 2716-24.

141. Niida, A., et al., *A novel meta-analysis approach of cancer transcriptomes reveals prevailing transcriptional networks in cancer cells.* Genome Inform, 2010. **22**: p. 121-31.

142. Varrault, A., et al., *Zac1 regulates an imprinted gene network critically involved in the control of embryonic growth.* Dev Cell, 2006. **11**(5): p. 711-22.

143. Zhou, X.H.J., et al., *Functional annotation and network reconstruction through cross-platform integration of microarray data.* Nature Biotechnology, 2005. **23**(2): p. 238-243.

144. Huttenhower, C., et al., *A scalable method for integration and functional analysis of multiple microarray datasets.* Bioinformatics, 2006. **22**(23): p. 2890-7.

145. Laliberté, E., *metacor: Meta-analysis of correlation coefficients*. 2011.

146. Ioannidis, J.P., *Interpretation of tests of heterogeneity and bias in meta-analysis.* J Eval Clin Pract, 2008. **14**(5): p. 951-7.

147. Peters, J.L., et al., *Assessing publication bias in meta-analyses in the presence of between-study heterogeneity.* Journal of the Royal Statistical Society Series a-Statistics in Society, 2010. **173**: p. 575-591.

148. Egger, M., et al., *Bias in meta-analysis detected by a simple, graphical test.* BMJ, 1997. **315**(7109): p. 629-34.

149. Jahanshiri, E. and A.R.M. Shariff, *Developing web-based data analysis tools for precision farming using R and Shiny.* 7th Igrsm International Remote Sensing & Gis Conference and Exhibition, 2014. **20**.

150. Wojciechowski, J., A.M. Hopkins, and R.N. Upton, *Interactive Pharmacometric Applications Using R and the Shiny Package.* CPT Pharmacometrics Syst Pharmacol, 2015. **4**(3): p. e00021.

151. Weston, R.A.a.S., *doParallel: Foreach parallel adaptor for the parallel package*. 2014.

152. Team, R.C., *R: A language and environment for statistical computing*. 2017, R Foundation for Statistical Computing: Vienna, Austria.

153. McPherson, W.C.a.J.C.a.J.A.a.Y.X.a.J., *shiny: Web Application Framework for R.* 2017.

154. Blischak, J.D., E.R. Davenport, and G. Wilson, *A Quick Introduction to Version Control with Git and GitHub.* PLoS Comput Biol, 2016. **12**(1): p. e1004668.

155. Dumas, J., M.A. Gargano, and G.M. Dancik, *shinyGEO: a web-based application for analyzing gene expression omnibus datasets.* Bioinformatics, 2016. **32**(23): p. 3679-3681.

156. Attali, D., et al., *ddpcr: an R package and web application for analysis of droplet digital PCR data.* F1000Res, 2016. **5**: p. 1411.

157. Lam, F., et al., *VennDiagramWeb: a web application for the generation of highly customizable Venn and Euler diagrams.* BMC Bioinformatics, 2016. **17**(1): p. 401.

158. Nelson, J.W., et al., *The START App: a web-based RNAseq analysis and visualization resource.* Bioinformatics, 2017. **33**(3): p. 447-449.

159. Liu, F. and D. Miranda-Saavedra, *rTRM-web: a web tool for predicting transcriptional regulatory modules for ChIP-seq-ed transcription factors.* Gene, 2014. **546**(2): p. 417-20.

160. Clark, N.R., et al., *Principle Angle Enrichment Analysis (PAEA): Dimensionally Reduced Multivariate Gene Set Enrichment Analysis Tool.* Proceedings 2015 Ieee International Conference on Bioinformatics and Biomedicine, 2015: p. 256-262.

161. Guo, Y., et al., *RNAseqPS: A Web Tool for Estimating Sample Size and Power for RNAseq Experiment.* Cancer Inform, 2014. **13**(Suppl 6): p. 1-5.

162. Beck, D., C. Dennis, and J.A. Foster, *Seed: a user-friendly tool for exploring and visualizing microbial community data.* Bioinformatics, 2015. **31**(4): p. 602-3.

163. Nicolle, R., F. Radvanyi, and M. Elati, *CoRegNet: reconstruction and integrated analysis of co-regulatory networks.* Bioinformatics, 2015. **31**(18): p. 3066-8.

164. Scales, M., et al., *visPIG--a web tool for producing multi-region, multi-track, multi-scale plots of genetic data.* PLoS One, 2014. **9**(9): p. e107497.

165. Francis, R.M., *pophelper: an R package and web app to analyse and visualize population structure.* Mol Ecol Resour, 2017. **17**(1): p. 27-32.

166. Metsalu, T. and J. Vilo, *ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap.* Nucleic Acids Res, 2015. **43**(W1): p. W566-70.

167. Mallona, I., A. Diez-Villanueva, and M.A. Peinado, *Methylation plotter: a web tool for dynamic visualization of DNA methylation data.* Source Code Biol Med, 2014. **9**: p. 11.

168. Ramasamy, R.K., et al., *STRUCTURE PLOT: a program for drawing elegant STRUCTURE bar plots in user friendly interface.* Springerplus, 2014. **3**: p. 431.

169. van Iterson, M., et al., *MethylAid: visual and interactive quality control of large Illumina 450k datasets.* Bioinformatics, 2014. **30**(23): p. 3435-7.

170. McMurdie, P.J. and S. Holmes, *Shiny-phyloseq: Web application for interactive microbiome analysis with provenance tracking.* Bioinformatics, 2015. **31**(2): p. 282-3.

171. Clough, E. and T. Barrett, *The Gene Expression Omnibus Database.* Methods Mol Biol, 2016. **1418**: p. 93-110.

172. Forbes, S.A., et al., *COSMIC: exploring the world's knowledge of somatic mutations in human cancer.* Nucleic Acids Res, 2015. **43**(Database issue): p. D805-11.

173. Kolesnikov, N., et al., *ArrayExpress update--simplifying data submissions.* Nucleic Acids Res, 2015. **43**(Database issue): p. D1113-6.

174. Barretina, J., et al., *The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity (vol 483, pg 603, 2012).* Nature, 2012. **492**(7428): p. 290-290.

175. Zhu, Y., P. Qiu, and Y. Ji, *TCGA-assembler: open-source software for retrieving and processing TCGA data.* Nat Methods, 2014. **11**(6): p. 599-600.

176. Deng, M., et al., *Web-TCGA: an online platform for integrated analysis of molecular cancer data sets.* BMC Bioinformatics, 2016. **17**: p. 72.

177. Reed, L.K., C.F. Baer, and A.S. Edison, *Considerations when choosing a genetic model organism for metabolomics studies.* Current Opinion in Chemical Biology, 2017. **36**: p. 7-14.

178. Ankeny, R.A. and S. Leonelli, *What's so special about model organisms?* Studies in History and Philosophy of Science, 2011. **42**(2): p. 313-323.

179. Amatruda, J.F., et al., *Zebrafish as a cancer model system.* Cancer Cell, 2002. **1**(3): p. 229-231.

180. Kipling, D. and H.J. Cooke, *Hypervariable ultra-long telomeres in mice.* Nature, 1990. **347**(6291): p. 400-2.

181. Rosenthal, N. and S. Brown, *The mouse ascending: perspectives for human-disease models.* Nat Cell Biol, 2007. **9**(9): p. 993-9.

182. Liu, K., et al., *Recombinase-mediated cassette exchange to rapidly and efficiently generate mice with human cardiac sodium channels.* Genesis, 2006. **44**(11): p. 556-64.

183. Gerhard, G.S., *Comparative aspects of zebrafish (Danio rerio) as a model for aging research.* Experimental Gerontology, 2003. **38**(11-12): p. 1333-1341.

184. Howe, K., et al., *The zebrafish reference genome sequence and its relationship to the human genome.* Nature, 2013. **496**(7446): p. 498-503.

185. Rudner, L.A., et al., *Shared acquired genomic changes in zebrafish and human T-ALL.* Oncogene, 2011. **30**(41): p. 4289-96.

186. Lam, S.H., et al., *Conservation of gene expression signatures between zebrafish and human liver tumors and tumor progression.* Nature Biotechnology, 2006. **24**(1): p. 73-75.

187. Schartl, M., et al., *Conserved Expression Signatures between Medaka and Human Pigment Cell Tumors.* Plos One, 2012. **7**(5).

188. Kobayashi, I., et al., *Comparative gene expression analysis of zebrafish and mammals identifies common regulators in hematopoietic stem cells.* Blood, 2010. **115**(2): p. e1-9.

189. Sugano, Y., et al., *Comparative transcriptomic analysis identifies evolutionarily conserved gene products in the vertebrate renal distal convoluted tubule.* Pflugers Arch, 2017. **469**(7-8): p. 859-867.

190. Ung, C.Y., et al., *Mercury-induced hepatotoxicity in zebrafish: in vivo mechanistic insights from transcriptome analysis, phenotype anchoring and targeted gene expression validation.* BMC Genomics, 2010. **11**: p. 212.

191. Driessen, M., et al., *A transcriptomics-based hepatotoxicity comparison between the zebrafish embryo and established human and rodent in vitro and in vivo models using cyclosporine A, amiodarone and acetaminophen.* Toxicology Letters, 2015. **232**(2): p. 403-412.

192. Davison, J.M., et al., *Microbiota regulate intestinal epithelial gene expression by suppressing the transcription factor Hepatocyte nuclear factor 4 alpha.* Genome Res, 2017. **27**(7): p. 1195-1206.

193. Lam, S.H., et al., *Molecular conservation of estrogen-response associated with cell cycle regulation, hormonal carcinogenesis and cancer in zebrafish and human cancer cell lines.* BMC Med Genomics, 2011. **4**: p. 41.

194. Lisse, T.S., B.L. King, and S. Rieger, *Comparative transcriptomic profiling of hydrogen peroxide signaling networks in zebrafish and human keratinocytes: Implications toward conservation, migration and wound healing.* Sci Rep, 2016. **6**: p. 20328.

195. Postlethwait, J.H., et al., *Zebrafish comparative genomics and the origins of vertebrate chromosomes.* Genome Research, 2000. **10**(12): p. 1890-1902.

196. Zhao, H., et al., *Pseudogenization of the umami taste receptor gene Tas1r1 in the giant panda coincided with its dietary switch to bamboo.* Mol Biol Evol, 2010. **27**(12): p. 2669-73.

197. Pasquier, J., et al., *Gene evolution and gene expression after whole genome duplication in fish: the PhyloFish database.* BMC Genomics, 2016. **17**: p. 368.

198. Zhang, J.Z., *Evolution by gene duplication: an update.* Trends in Ecology & Evolution, 2003. **18**(6): p. 292-298.

199. Prince, V.E. and F.B. Pickett, *Splitting pairs: the diverging fates of duplicated genes.* Nat Rev Genet, 2002. **3**(11): p. 827-37.

200. Gibert, J.M., *The evolution of engrailed genes after duplication and speciation events.* Dev Genes Evol, 2002. **212**(7): p. 307-18.

201. Sprague, J., et al., *The Zebrafish Information Network (ZFIN): the zebrafish model organism database.* Nucleic Acids Res, 2003. **31**(1): p. 241-3.

202. Chen, W.H., et al., *OGEE: an online gene essentiality database.* Nucleic Acids Res, 2012. **40**(Database issue): p. D901-6.

203. Sonnhammer, E.L. and G. Ostlund, *InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic.* Nucleic Acids Res, 2015. **43**(Database issue): p. D234-9.

204. Obayashi, T., et al., *COXPRESdb: a database of comparative gene coexpression networks of eleven species for mammals.* Nucleic Acids Res, 2013. **41**(Database issue): p. D1014-20.

205. Mi, H., et al., *PANTHER version 10: expanded protein families and functions, and analysis tools.* Nucleic Acids Res, 2016. **44**(D1): p. D336-42.

206. Loots, G. and I. Ovcharenko, *ECRbase: database of evolutionary conserved regions, promoters, and transcription factor binding sites in vertebrate genomes.* Bioinformatics, 2007. **23**(1): p. 122-4.

207. Piasecka, B., et al., *Comparative modular analysis of gene expression in vertebrate organs.* BMC Genomics, 2012. **13**: p. 124.

208. Karro, J.E., et al., *Pseudogene.org: a comprehensive database and comparison platform for pseudogene annotation.* Nucleic Acids Res, 2007. **35**(Database issue): p. D55-60.

209. Ouedraogo, M., et al., *The duplicated genes database: identification and functional annotation of co-localised duplicated genes across genomes.* PLoS One, 2012. **7**(11): p. e50653.

210. Van de Peer, Y., et al., *Wanda: a database of duplicated fish genes.* Nucleic Acids Res, 2002. **30**(1): p. 109-12.

211. Taylor, J.S., et al., *Genome duplication, a trait shared by 22000 species of ray-finned fish.* Genome Res, 2003. **13**(3): p. 382-90.

212. Barbazuk, W.B., et al., *The syntenic relationship of the zebrafish and human genomes.* Genome Res, 2000. **10**(9): p. 1351-8.

213. Zhong, Z.X., et al., *Divergent DNA Methylation Provides Insights into the Evolution of Duplicate Genes in Zebrafish.* G3-Genes Genomes Genetics, 2016. **6**(11): p. 3581-3591.

214. Kinsella, R.J., et al., *Ensembl BioMarts: a hub for data retrieval across taxonomic space.* Database (Oxford), 2011. **2011**: p. bar030.

215. Boroughs, L.K. and R.J. DeBerardinis, *Metabolic pathways promoting cancer cell survival and growth.* Nat Cell Biol, 2015. **17**(4): p. 351-9.

216. Sasore, T. and B. Kennedy, *Deciphering combinations of PI3K/AKT/mTOR pathway drugs augmenting anti-angiogenic efficacy in vivo.* PLoS One, 2014. **9**(8): p. e105280.

217. LoPiccolo, J., et al., *Targeting the PI3K/Akt/mTOR pathway: effective combinations and clinical considerations.* Drug Resist Updat, 2008. **11**(1-2): p. 32-50.

218. Shertz, C.A., et al., *Conservation, duplication, and loss of the Tor signaling pathway in the fungal kingdom.* BMC Genomics, 2010. **11**: p. 510.

219. Huang, J.X., et al., *The TSC1-TSC2 complex is required for proper activation of mTOR complex 2.* Molecular and Cellular Biology, 2008. **28**(12): p. 4104-4115.

220. Law, B.K., *Rapamycin: an anti-cancer immunosuppressant?* Crit Rev Oncol Hematol, 2005. **56**(1): p. 47-60.

221. Chan, S., *Targeting the mammalian target of rapamycin (mTOR): a new approach to treating cancer.* Br J Cancer, 2004. **91**(8): p. 1420-4.

222. Huang, J. and B.D. Manning, *The TSC1-TSC2 complex: a molecular switchboard controlling cell growth.* Biochem J, 2008. **412**(2): p. 179-90.

223. DiBella, L.M., A. Park, and Z. Sun, *Zebrafish Tsc1 reveals functional interactions between the cilium and the TOR pathway.* Hum Mol Genet, 2009. **18**(4): p. 595-606.

224. Makky, K., J. Tekiela, and A.N. Mayer, *Target of rapamycin (TOR) signaling controls epithelial morphogenesis in the vertebrate intestine.* Dev Biol, 2007. **303**(2): p. 501-13.

225. Sucularli, C., et al., *Functionally conserved effects of rapamycin exposure on zebrafish.* Mol Med Rep, 2016. **13**(5): p. 4421-30.

226. Edgar, R., M. Domrachev, and A.E. Lash, *Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.* Nucleic Acids Res, 2002. **30**(1): p. 207-10.

227. Sean, D. and P.S. Meltzer, *GEOquery: a bridge between the gene expression omnibus (GEO) and BioConductor.* Bioinformatics, 2007. **23**(14): p. 1846-1847.

228. Duvel, K., et al., *Activation of a metabolic gene regulatory network downstream of mTOR complex 1.* Mol Cell, 2010. **39**(2): p. 171-83.

229. Gentleman, R., *Bioinformatics and computational biology solutions using R and Bioconductor*, in *Statistics for biology and health*. 2005, Springer: New York ; London. p. xix, 473 p.

230. Tseng, X.W.a.J.L.a.G.C., *MetaDE: Microarray meta-analysis for differentially expressed*

gene detection. 2012. **R package version 1.0.5**.

231. Fay, M.P., *Confidence intervals that match Fisher's exact or Blaker's exact tests.* Biostatistics, 2010. **11**(2): p. 373-374.

232. Aken, B.L., et al., *The Ensembl gene annotation system.* Database (Oxford), 2016. **2016**.

233. Wickham, H., *ggplot2: Elegant Graphics for Data Analysis*. 2009, Springer-Verlag New York.

234. Nepusz, G.C.a.T., *The igraph software package for complex network research.* InterJournal, 2006: p. 1695.

235. Totary-Jain, H., et al., *Reprogramming of the microRNA transcriptome mediates resistance to rapamycin.* J Biol Chem, 2013. **288**(9): p. 6034-44.

236. Feng, Z., et al., *The regulation of AMPK beta1, TSC2, and PTEN expression by p53: stress, cell and tissue specificity, and the role of these gene products in modulating the IGF-1-AKT-mTOR pathways.* Cancer Res, 2007. **67**(7): p. 3043-53.

237. Postlethwait, J.H., et al., *Zebrafish comparative genomics and the origins of vertebrate chromosomes.* Genome Res, 2000. **10**(12): p. 1890-902.

238. Zhao, S., J. Huang, and J. Ye, *A fresh look at zebrafish from the perspective of cancer research.* Journal of Experimental & Clinical Cancer Research, 2015. **34**.

239. Kumra, H. and D.P. Reinhardt, *Fibronectin-targeted drug delivery in cancer.* Advanced Drug Delivery Reviews, 2016. **97**: p. 101-110.

240. Osmulski, P.A. and M. Gaczynska, *Rapamycin allosterically inhibits the proteasome.* Mol Pharmacol, 2013. **84**(1): p. 104-13.

241. Ducharme, N.A., et al., *Comparison of toxicity values across zebrafish early life stages and mammalian studies: Implications for chemical testing.* Reprod Toxicol, 2015. **55**: p. 3-10.

242. Filiano, A.J., et al., *Unexpected role of interferon-gamma in regulating neuronal connectivity and social behaviour.* Nature, 2016. **535**(7612): p. 425-9.

243. Wang, R.L., et al., *Natural Variation in Fish Transcriptomes: Comparative Analysis of the Fathead Minnow (Pimephales promelas) and Zebrafish (Danio rerio).* PLoS One, 2014. **9**(12): p. e114178.

244. Lenzi, L., et al., *TRAM (Transcriptome Mapper): database-driven creation and analysis of transcriptome maps from multiple sources.* BMC Genomics, 2011. **12**: p. 121.

245. Schuttler, A., et al., *The Transcriptome of the Zebrafish Embryo After Chemical Exposure: A Meta-Analysis.* Toxicol Sci, 2017. **157**(2): p. 291-304.

246. Kozol, R.A., et al., *Function Over Form: Modeling Groups of Inherited Neurological Conditions in Zebrafish.* Front Mol Neurosci, 2016. **9**: p. 55.

247. Groh, K.J. and M.J. Suter, *Stressor-induced proteome alterations in zebrafish: a meta-analysis of response patterns.* Aquat Toxicol, 2015. **159**: p. 1-12.

248. Ulitsky, I., et al., *Extensive alternative polyadenylation during zebrafish development.* Genome Res, 2012. **22**(10): p. 2054-66.

249. Cao, S., et al., *High-throughput RNA sequencing-based virome analysis of 50 lymphoma cell lines from the Cancer Cell Line Encyclopedia project.* J Virol, 2015. **89**(1): p. 713-29.

250. Cancer Cell Line Encyclopedia, C. and C. Genomics of Drug Sensitivity in Cancer, *Pharmacogenomic agreement between two cancer cell line data sets.* Nature, 2015. **528**(7580): p. 84-7.

251. Sonkin, D., et al., *Tumor suppressors status in cancer cell line Encyclopedia.* Mol Oncol, 2013. **7**(4): p. 791-8.

252. Cristescu, R., et al., *Molecular analysis of gastric cancer identifies subtypes associated with distinct clinical outcomes.* Nat Med, 2015. **21**(5): p. 449-56.

253. Ott, C., et al., *BET Bromodomain Inhibition Targets Both c-Myc and IL7R in Acute Lymphoblastic Leukemia.* European Journal of Cancer, 2012. **48**: p. 166-166.

254. Takahashi, Y., et al., *Amplification of PVT-1 is involved in poor prognosis via apoptosis inhibition in colorectal cancers.* British Journal of Cancer, 2014. **110**(1): p. 164-171.

255. Vazquez, F., et al., *PGC1alpha expression defines a subset of human melanoma tumors with increased mitochondrial capacity and resistance to oxidative stress.* Cancer Cell, 2013. **23**(3): p. 287-301.

256. Sebastian, C., et al., *The Histone Deacetylase SIRT6 Is a Tumor Suppressor that Controls Cancer Metabolism.* Cell, 2012. **151**(6): p. 1185-1199.

257. Huang, F.W., et al., *Highly recurrent TERT promoter mutations in human melanoma.* Science, 2013. **339**(6122): p. 957-9.

258. Rosenbluh, J., et al., *beta-Catenin-driven cancers require a YAP1 transcriptional complex for survival and tumorigenesis.* Cell, 2012. **151**(7): p. 1457-73.

259. Bray, S.J., et al., *Embryonic Expression Pattern of a Family of Drosophila Proteins That Interact with a Central Nervous-System Regulatory Element.* Genes & Development, 1989. **3**(8): p. 1130-1145.

260. Wilanowski, T., et al., *A highly conserved novel family of mammalian developmental transcription factors related to Drosophila grainyhead.* Mech Dev, 2002. **114**(1-2): p. 37-50.

261. Darido, C., et al., *Targeting of the tumor suppressor GRHL3 by a miR-21-dependent proto-oncogenic network results in PTEN loss and tumorigenesis.* Cancer Cell, 2011. **20**(5): p. 635-48.

262. Yu, Z., et al., *The Grainyhead-like epithelial transactivator Get-1/Grhl3 regulates epidermal terminal differentiation and interacts functionally with LMO4.* Dev Biol, 2006. **299**(1): p. 122-36.

263. Werth, M., et al., *The transcription factor grainyhead-like 2 regulates the molecular composition of the epithelial apical junctional complex.* Development, 2010. **137**(22): p. 3835-45.

264. Lukosz, M., et al., *The transcription factor Grainyhead like 3 (GRHL3) affects endothelial cell apoptosis and migration in a NO-dependent manner.* Biochem Biophys Res Commun, 2011. **412**(4): p. 648-53.

265. Guardiola-Serrano, F., et al., *Gene trapping identifies a putative tumor suppressor and a new inducer of cell migration.* Biochem Biophys Res Commun, 2008. **376**(4): p. 748-52.

266. DeLaForest, A., et al., *HNF4A is essential for specification of hepatic progenitors from human pluripotent stem cells.* Development, 2011. **138**(19): p. 4143-4153.

267. Garrison, W.D., et al., *Hepatocyte nuclear factor 4 alpha is essential for embryonic development of the mouse colon.* Gastroenterology, 2006. **130**(4): p. 1207-1220.

268. Yang, M., et al., *A double-negative feedback loop between Wnt-beta-catenin signaling and HNF4alpha regulates epithelial-mesenchymal transition in hepatocellular carcinoma.* J Cell Sci, 2013. **126**(Pt 24): p. 5692-703.

269. Battle, M.A., et al., *Hepatocyte nuclear factor 4alpha orchestrates expression of cell adhesion proteins during the epithelial transformation of the developing liver.* Proc Natl Acad Sci U S A, 2006. **103**(22): p. 8419-24.

270. Caja, L., et al., *The transforming growth factor-beta (TGF-beta) mediates acquisition of a mesenchymal stem cell-like phenotype in human liver cells.* J Cell Physiol, 2011. **226**(5): p. 1214-23.

271. Deng, X.G., et al., *Overexpression of miR-122 promotes the hepatic differentiation and maturation of mouse ESCs through a miR-122/FoxA1/HNF4a-positive feedback loop.* Liver Int, 2014. **34**(2): p. 281-95.

272. van der Post, R.S., et al., *HNF4A immunohistochemistry facilitates distinction between primary and metastatic breast and gastric carcinoma.* Virchows Archiv, 2014. **464**(6): p. 673-679.

273. Breier, G., M. Grosser, and M. Rezaei, *Endothelial cadherins in cancer.* Cell Tissue Res, 2014. **355**(3): p. 523-7.

274. Dejana, E., *Endothelial cell-cell junctions: happy together.* Nat Rev Mol Cell Biol, 2004. **5**(4): p. 261-70.

275. Yilmaz, M. and G. Christofori, *EMT, the cytoskeleton, and cancer cell invasion.* Cancer Metastasis Rev, 2009. **28**(1-2): p. 15-33.

276. Peinado, H., D. Olmeda, and A. Cano, *Snail, Zeb and bHLH factors in tumour progression: an alliance against the epithelial phenotype?* Nat Rev Cancer, 2007. **7**(6): p. 415-28.

277. Hazan, R.B., et al., *Exogenous expression of N-cadherin in breast cancer cells induces cell migration, invasion, and metastasis.* J Cell Biol, 2000. **148**(4): p. 779-90.

278. Cavallaro, U. and G. Christofori, *Cell adhesion and signalling by cadherins and Ig-CAMs in cancer.* Nat Rev Cancer, 2004. **4**(2): p. 118-32.

279. Scully, O.J., et al., *Breast cancer metastasis.* Cancer Genomics Proteomics, 2012. **9**(5): p. 311-20.

280. Hay, E.D., *An overview of epithelio-mesenchymal transformation.* Acta Anat (Basel), 1995. **154**(1): p. 8-20.

281. Yoshida, T., et al., *Eribulin mesilate suppresses experimental metastasis of breast cancer cells by reversing phenotype from epithelial-mesenchymal transition (EMT) to mesenchymal-epithelial transition (MET) states.* Br J Cancer, 2014. **110**(6): p. 1497-505.

282. Micalizzi, D.S., S.M. Farabaugh, and H.L. Ford, *Epithelial-Mesenchymal Transition in Cancer: Parallels Between Normal Development and Tumor Progression.* Journal of Mammary Gland Biology and Neoplasia, 2010. **15**(2): p. 117-134.

283. Thompson, E.W. and E.D. Williams, *EMT and MET in carcinoma - clinical observations, regulatory pathways and new models.* Clinical & Experimental Metastasis, 2008. **25**(6): p. 591-592.

284. Armstrong, A.J., et al., *Circulating Tumor Cells from Patients with Advanced Prostate and Breast Cancer Display Both Epithelial and Mesenchymal Markers.* Molecular Cancer Research, 2011. **9**(8): p. 997-1007.

285. Alotaibi, H., et al., *Enhancer cooperativity as a novel mechanism underlying the transcriptional regulation of E-cadherin during mesenchymal to epithelial transition.* Biochimica Et Biophysica Acta-Gene Regulatory Mechanisms, 2015. **1849**(6): p. 731-742.

286. Auden, A., et al., *Spatial and temporal expression of the Grainyhead-like transcription factor family during murine development.* Gene Expr Patterns, 2006. **6**(8): p. 964-70.

287. Zhao, P., et al., *Grhl3 induces human epithelial tumor cell migration and invasion via downregulation of E-cadherin.* Acta Biochim Biophys Sin (Shanghai), 2016. **48**(3): p. 266-74.

288. Nagarajan, S., et al., *BRD4 promotes p63 and GRHL3 expression downstream of FOXO in mammary epithelial cells.* Nucleic Acids Res, 2017. **45**(6): p. 3130-3145.

289. Benson, J.R. and I. Jatoi, *The global breast cancer burden.* Future Oncol, 2012. **8**(6): p. 697-702.

290. Siegel, R.L., K.D. Miller, and A. Jemal, *Cancer statistics, 2016.* CA Cancer J Clin, 2016. **66**(1): p. 7-30.

291. Blackburn, H.L., et al., *Breast Cancer Metastasis to the Axillary Lymph Nodes: Are Changes to the Lymph Node "Soil" Localized or Systemic?* Breast Cancer (Auckl), 2017. **11**: p. 1178223417691246.

292. Medina, D., *The mammary gland: a unique organ for the study of development and tumorigenesis.* J Mammary Gland Biol Neoplasia, 1996. **1**(1): p. 5-19.

293. Visvader, J.E., *Keeping abreast of the mammary epithelial hierarchy and breast tumorigenesis.* Genes Dev, 2009. **23**(22): p. 2563-77.

294. Glukhova, M., et al., *Adhesion systems in normal breast and in invasive breast carcinoma.* Am J Pathol, 1995. **146**(3): p. 706-16.

295. Visvader, J.E. and J. Stingl, *Mammary stem cells and the differentiation hierarchy: current status and perspectives.* Genes Dev, 2014. **28**(11): p. 1143-58.

296. Warburton, M.J., et al., *Distribution of myoepithelial cells and basement membrane proteins in the resting, pregnant, lactating, and involuting rat mammary gland.* J Histochem Cytochem, 1982. **30**(7): p. 667-76.

297. Phillips, S. and C. Kuperwasser, *SLUG: Critical regulator of epithelial cell identity in breast development and cancer.* Cell Adh Migr, 2014. **8**(6): p. 578-87.

298. Sreekumar, A., K. Roarty, and J.M. Rosen, *The mammary stem cell hierarchy: a looking glass into heterogeneous breast cancer landscapes.* Endocrine-Related Cancer, 2015. **22**(6): p. T161-T176.

299. Guelstein, V.I., et al., *Monoclonal antibody mapping of keratins 8 and 17 and of vimentin in normal human mammary gland, benign tumors, dysplasias and breast cancer.* Int J Cancer, 1988. **42**(2): p. 147-53.

300. Nagle, R.B., et al., *Characterization of breast carcinomas by two monoclonal antibodies distinguishing myoepithelial from luminal epithelial cells.* J Histochem Cytochem, 1986. **34**(7): p. 869-81.

301. Shehata, M., et al., *Phenotypic and functional characterisation of the luminal cell hierarchy of the mammary gland.* Breast Cancer Res, 2012. **14**(5): p. R134.

302. Dairkee, S.H., et al., *Monoclonal-Antibody That Defines Human Myoepithelium.* Proceedings of the National Academy of Sciences of the United States of America, 1985. **82**(21): p. 7409-7413.

303. Skalli, O., et al., *A monoclonal antibody against alpha-smooth muscle actin: a new probe for smooth muscle differentiation.* J Cell Biol, 1986. **103**(6 Pt 2): p. 2787-96.

304. Mork, C., B. van Deurs, and O.W. Petersen, *Regulation of vimentin expression in cultured human mammary epithelial cells.* Differentiation, 1990. **43**(2): p. 146-56.

305. Oakes, S.R., D. Gallego-Ortega, and C.J. Ormandy, *The mammary cellular hierarchy and breast cancer.* Cell Mol Life Sci, 2014. **71**(22): p. 4301-24.

306. Dhermain, F., *Radiotherapy of high-grade gliomas: current standards and new concepts, innovations in imaging and radiotherapy, and new therapeutic approaches.* Chin J Cancer, 2014. **33**(1): p. 16-24.

307. Warrell, R.P., Jr., et al., *Differentiation therapy of acute promyelocytic leukemia with tretinoin (all-trans-retinoic acid).* N Engl J Med, 1991. **324**(20): p. 1385-93.

308. Scheideler, M., et al., *Comparative transcriptomics of human multipotent stem cells during adipogenesis and osteoblastogenesis.* BMC Genomics, 2008. **9**: p. 340.

309. Onitilo, A.A., et al., *Breast cancer subtypes based on ER/PR and Her2 expression: comparison of clinicopathologic features and survival.* Clin Med Res, 2009. **7**(1-2): p. 4-13.

310. Lin, M.L., et al., *Expression profiling of nuclear receptors in breast cancer identifies TLX as a mediator of growth and invasion in triple-negative breast cancer.* Oncotarget, 2015. **6**(25): p. 21685-21703.

311. Conzen, S.D., *Minireview: nuclear receptors and breast cancer.* Mol Endocrinol, 2008. **22**(10): p. 2215-28.

312. Carey, L.A., et al., *Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study.* JAMA, 2006. **295**(21): p. 2492-502.

313. Hah, N., et al., *A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells.* Cell, 2011. **145**(4): p. 622-34.

314. McDevitt, M.A., et al., *New insights into the classical and non-classical actions of estrogen: evidence from estrogen receptor knock-out and knock-in mice.* Mol Cell Endocrinol, 2008. **290**(1-2): p. 24-30.

315. Ngwenya, S. and S. Safe, *Cell context-dependent differences in the induction of E2F-1 gene expression by 17 beta-estradiol in MCF-7 and ZR-75 cells.* Endocrinology, 2003. **144**(5): p. 1675-85.

316. Song, R.X. and R.J. Santen, *Membrane initiated estrogen signaling in breast cancer.* Biol Reprod, 2006. **75**(1): p. 9-16.

317. Drabovich, A.P., et al., *Dynamics of Protein Expression Reveals Primary Targets and Secondary Messengers of Estrogen Receptor Alpha Signaling in MCF-7 Breast Cancer Cells.* Mol Cell Proteomics, 2016. **15**(6): p. 2093-107.

318. Bourdeau, V., et al., *Mechanisms of primary and secondary estrogen target gene regulation in breast cancer cells.* Nucleic Acids Res, 2008. **36**(1): p. 76-93.

319. Fillmore, C.M., et al., *Estrogen expands breast cancer stem-like cells through paracrine FGF/Tbx3 signaling.* Proc Natl Acad Sci U S A, 2010. **107**(50): p. 21737-42.

320. Chiu, J.H., et al., *Role of estrogen receptors and Src signaling in mechanisms of bone metastasis by estrogen receptor positive breast cancers.* J Transl Med, 2017. **15**(1): p. 97.

321. Saha Roy, S. and R.K. Vadlamudi, *Role of estrogen receptor signaling in breast cancer metastasis.* Int J Breast Cancer, 2012. **2012**: p. 654698.

322. Ariazi, E.A., et al., *Estrogen receptors as therapeutic targets in breast cancer.* Current Topics in Medicinal Chemistry, 2006. **6**(3): p. 181-202.

323. Sawka, C.A., et al., *Role and mechanism of action of tamoxifen in premenopausal women with metastatic breast carcinoma.* Cancer Res, 1986. **46**(6): p. 3152-6.

324. Jordan, V.C., *The role of tamoxifen in the treatment and prevention of breast cancer.* Curr Probl Cancer, 1992. **16**(3): p. 129-76.

325. Osborne, C.K., *Tamoxifen in the treatment of breast cancer.* N Engl J Med, 1998. **339**(22): p. 1609-18.

326. Fisher, B., et al., *Tamoxifen for prevention of breast cancer: report of the National Surgical Adjuvant Breast and Bowel Project P-1 Study.* J Natl Cancer Inst, 1998. **90**(18): p. 1371-88.

327. Fisher, B., et al., *Treatment of lymph-node-negative, oestrogen-receptor-positive breast cancer: long-term findings from National Surgical Adjuvant Breast and Bowel Project randomised clinical trials.* Lancet, 2004. **364**(9437): p. 858-68.

328. Ochsner, S.A., et al., *GEMS (Gene Expression MetaSignatures), a Web resource for querying meta-analysis of expression microarray datasets: 17beta-estradiol in MCF-7 cells.* Cancer Res, 2009. **69**(1): p. 23-6.

329. Jagannathan, V. and M. Robinson-Rechavi, *Meta-analysis of estrogen response in MCF-7 distinguishes early target genes involved in signaling and cell proliferation from later target genes involved in cell cycle and DNA repair.* BMC Syst Biol, 2011. **5**: p. 138.

330. Xia, J., E.E. Gill, and R.E. Hancock, *NetworkAnalyst for statistical, visual and network-based meta-analysis of gene expression data.* Nat Protoc, 2015. **10**(6): p. 823-44.

331. Alles, M.C., et al., *Meta-analysis and gene set enrichment relative to er status reveal elevated activity of MYC and E2F in the "basal" breast cancer subgroup.* PLoS One, 2009. **4**(3): p. e4710.

332. Lin, C.Y., et al., *Discovery of estrogen receptor alpha target genes and response elements in breast tumor cells.* Genome Biol, 2004. **5**(9): p. R66.

333. Ibrahim, E.M. and A. Al-Homaidh, *Physical activity and survival after breast cancer diagnosis: meta-analysis of published studies.* Med Oncol, 2011. **28**(3): p. 753-65.

334. Wirapati, P., et al., *Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures.* Breast Cancer Res, 2008. **10**(4): p. R65.

335. Kalamida, D., et al., *Muscle and neuronal nicotinic acetylcholine receptors. Structure, function and pathogenicity.* FEBS J, 2007. **274**(15): p. 3799-845.

336. Posadas, I., B. Lopez-Hernandez, and V. Cena, *Nicotinic receptors in neurodegeneration.* Curr Neuropharmacol, 2013. **11**(3): p. 298-314.

337. Lindstrom, J., *Nicotinic acetylcholine receptors in health and disease.* Mol Neurobiol, 1997. **15**(2): p. 193-222.

338. Weiland, S., D. Bertrand, and S. Leonard, *Neuronal nicotinic acetylcholine receptors: from the gene to the disease.* Behav Brain Res, 2000. **113**(1-2): p. 43-56.

339. Colombo, S.F., et al., *Biogenesis, trafficking and up-regulation of nicotinic ACh receptors.* Biochem Pharmacol, 2013. **86**(8): p. 1063-73.

340. Paterson, D. and A. Nordberg, *Neuronal nicotinic receptors in the human brain.* Prog Neurobiol, 2000. **61**(1): p. 75-111.

341. Trombino, S., et al., *Alpha7-nicotinic acetylcholine receptors affect growth regulation of human mesothelioma cells: role of mitogen-activated protein kinase pathway.* Cancer Res, 2004. **64**(1): p. 135-45.

342. Zeidler, R., K. Albermann, and S. Lang, *Nicotine and apoptosis.* Apoptosis, 2007. **12**(11): p. 1927-43.

343. Egleton, R.D., K.C. Brown, and P. Dasgupta, *Nicotinic acetylcholine receptors in cancer: multiple roles in proliferation and inhibition of apoptosis.* Trends Pharmacol Sci, 2008. **29**(3): p. 151-8.

344. Schuller, H.M., *Cell type specific, receptor-mediated modulation of growth kinetics in human lung cancer cell lines by nicotine and tobacco-related nitrosamines.* Biochem Pharmacol, 1989. **38**(20): p. 3439-42.

345. Lam, D.C., et al., *Expression of nicotinic acetylcholine receptor subunit genes in non-small-cell lung cancer reveals differences between smokers and nonsmokers.* Cancer Res, 2007. **67**(10): p. 4638-47.

346. Zheng, Y., et al., *Nicotine stimulates human lung cancer cell growth by inducing fibronectin expression.* Am J Respir Cell Mol Biol, 2007. **37**(6): p. 681-90.

347. Dom, A.M., et al., *The alpha7-nicotinic acetylcholine receptor and MMP-2/-9 pathway mediate the proangiogenic effect of nicotine in human retinal endothelial cells.* Invest Ophthalmol Vis Sci, 2011. **52**(7): p. 4428-38.

348. Wang, J.C., et al., *Genetic variation in the CHRNA5 gene affects mRNA levels and is associated with risk for alcohol dependence.* Mol Psychiatry, 2009. **14**(5): p. 501-10.

349. Spitz, M.R., et al., *The CHRNA5-A3 Region on Chromosome 15q24-25.1 Is a Risk Factor Both for Nicotine Dependence and for Lung Cancer.* Journal of the National Cancer Institute, 2008. **100**(21).

350. Sun, H.J., et al., *CHRNA5 Variants Related to Nicotine Sensitivity in Lung Cancer in Vitro.* 2010 3rd International Conference on Biomedical Engineering and Informatics (Bmei 2010), Vols 1-7, 2010: p. 2001-2003.

351. Falvella, F.S., et al., *Multiple isoforms and differential allelic expression of CHRNA5 in lung tissue and lung adenocarcinoma.* Carcinogenesis, 2013. **34**(6): p. 1281-5.

352. Satelli, A. and S. Li, *Vimentin in cancer and its potential as a molecular target for cancer therapy.* Cell Mol Life Sci, 2011. **68**(18): p. 3033-46.

353. Ivaska, J., et al., *Novel functions of vimentin in cell adhesion, migration, and signaling.* Exp Cell Res, 2007. **313**(10): p. 2050-62.

354. Dave, J.M. and K.J. Bayless, *Vimentin as an integral regulator of cell adhesion and endothelial sprouting.* Microcirculation, 2014. **21**(4): p. 333-44.

355. Korsching, E., et al., *The origin of vimentin expression in invasive breast cancer: epithelial-mesenchymal transition, myoepithelial histogenesis or histogenesis from progenitor cells with bilinear differentiation potential?* J Pathol, 2005. **206**(4): p. 451-7.

356. Domagala, W., et al., *Vimentin is preferentially expressed in high-grade ductal and medullary, but not in lobular breast carcinomas.* Am J Pathol, 1990. **137**(5): p. 1059-64.

357. Yuan, S.M., *alpha-Smooth Muscle Actin and ACTA2 Gene Expressions in Vasculopathies.* Braz J Cardiovasc Surg, 2015. **30**(6): p. 644-9.

358. Schildmeyer, L.A., et al., *Impaired vascular contractility and blood pressure homeostasis in the smooth muscle alpha-actin null mouse.* FASEB J, 2000. **14**(14): p. 2213-20.

359. Guo, D.C., et al., *Mutations in smooth muscle alpha-actin (ACTA2) cause coronary artery disease, stroke, and Moyamoya disease, along with thoracic aortic disease.* Am J Hum Genet, 2009. **84**(5): p. 617-27.

360. Kusafuka, K., et al., *Dedifferentiated epithelial-myoepithelial carcinoma of the parotid gland: a rare case report of immunohistochemical analysis and review of the literature.* Oral Surg Oral Med Oral Pathol Oral Radiol Endod, 2008. **106**(1): p. 85-91.

361. Ogawa, Y., *Immunocytochemistry of myoepithelial cells in the salivary glands.* Prog Histochem Cytochem, 2003. **38**(4): p. 343-426.

362. Haaksma, C.J., R.J. Schwartz, and J.J. Tomasek, *Myoepithelial cell contraction and milk ejection are impaired in mammary glands of mice lacking smooth muscle alpha-actin.* Biol Reprod, 2011. **85**(1): p. 13-21.

363. Pillai, K., et al., *MUC1 as a potential target in anticancer therapies.* Am J Clin Oncol, 2015. **38**(1): p. 108-18.

364. Ahmad, R., et al., *MUC1-C oncoprotein promotes STAT3 activation in an autoinductive regulatory loop.* Sci Signal, 2011. **4**(160): p. ra9.

365. Li, Y., et al., *Interaction of glycogen synthase kinase 3beta with the DF3/MUC1 carcinoma-associated antigen and beta-catenin.* Mol Cell Biol, 1998. **18**(12): p. 7216-24.

366. Roy, L.D., et al., *MUC1 enhances invasiveness of pancreatic cancer cells by inducing epithelial to mesenchymal transition.* Oncogene, 2011. **30**(12): p. 1449-59.

367. Schroeder, J.A., et al., *MUC1 overexpression results in mammary gland tumorigenesis and prolonged alveolar differentiation.* Oncogene, 2004. **23**(34): p. 5739-47.

368. Whitehouse, C., et al., *A transfected sialyltransferase that is elevated in breast cancer and localizes to the medial/trans-Golgi apparatus inhibits the development of core-2-based O-glycans.* J Cell Biol, 1997. **137**(6): p. 1229-41.

369. Nagai, S., et al., *A novel classification of MUC1 expression is correlated with tumor differentiation and postoperative prognosis in non-small cell lung cancer.* J Thorac Oncol, 2006. **1**(1): p. 46-51.

370. Westgaard, A., et al., *Differentiation markers in pancreatic head adenocarcinomas: MUC1 and MUC4 expression indicates poor prognosis in pancreatobiliary differentiated tumours.* Histopathology, 2009. **54**(3): p. 337-47.

371. Nieto, M.A., et al., *Control of cell behavior during vertebrate development by Slug, a zinc finger gene.* Science, 1994. **264**(5160): p. 835-9.

372. Cobaleda, C., et al., *Function of the zinc-finger transcription factor SNAI2 in cancer and development.* Annu Rev Genet, 2007. **41**: p. 41-61.

373. Inoue, A., et al., *Slug, a highly conserved zinc finger transcriptional repressor, protects hematopoietic progenitor cells from radiation-induced apoptosis in vivo.* Cancer Cell, 2002. **2**(4): p. 279-88.

374. Perez-Losada, J., et al., *Zinc-finger transcription factor Slug contributes to the function of the stem cell factor c-kit signaling pathway.* Blood, 2002. **100**(4): p. 1274-86.

375. Nieto, M.A., *The snail superfamily of zinc-finger transcription factors.* Nat Rev Mol Cell Biol, 2002. **3**(3): p. 155-66.

376. Becker, M.W. and M.F. Clarke, *SLUGging away at cell death.* Cancer Cell, 2002. **2**(4): p. 249-51.

377. Perez-Mancera, P.A., et al., *SLUG in cancer development.* Oncogene, 2005. **24**(19): p. 3073-82.

378. Germanguz, I., et al., *Identifying gene expression modules that define human cell fates.* Stem Cell Research, 2016. **16**(3): p. 712-724.

379. Suzuki, R. and H. Shimodaira, *Pvclust: an R package for assessing the uncertainty in hierarchical clustering.* Bioinformatics, 2006. **22**(12): p. 1540-2.

380. Chang, H., et al., *Synergistic action of master transcription factors controls epithelial-to-mesenchymal transition.* Nucleic Acids Res, 2016. **44**(6): p. 2514-27.

381. Li, Q., et al., *Jetset: selecting the optimal microarray probe set to represent a gene.* BMC Bioinformatics, 2011. **12**: p. 474.

382. Ye, J., et al., *Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction.* BMC Bioinformatics, 2012. **13**: p. 134.

383. Kent, W.J., et al., *The human genome browser at UCSC.* Genome Research, 2002. **12**(6): p. 996-1006.

384. Sheng, Q., Y. Shyr, and X. Chen, *DupChecker: a bioconductor package for checking high-throughput genomic data redundancy in meta-analysis.* BMC Bioinformatics, 2014. **15**: p. 323.

385. Gyorffy, B., et al., *An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients.* Breast Cancer Res Treat, 2010. **123**(3): p. 725-31.

386. Schwarzer, G., *meta: An R package for meta-analysis.* R News, 2007. **7**: p. 40--45.

387. Lin, D.Y. and L.J. Wei, *The Robust Inference for the Cox Proportional Hazards Model.* Journal of the American Statistical Association, 1989. **84**(408): p. 1074-1078.

388. Laliberté, E., *metacor: Meta-analysis of correlation coefficients.* 2011.

389. Jia, M., et al., *Estrogen Receptor alpha Promotes Breast Cancer by Reprogramming Choline Metabolism.* Cancer Res, 2016. **76**(19): p. 5634-5646.

390. Yu, G.C., et al., *clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters.* Omics-a Journal of Integrative Biology, 2012. **16**(5): p. 284-287.

391. Franceschini, A., et al., *STRING v9.1: protein-protein interaction networks, with increased coverage and integration.* Nucleic Acids Res, 2013. **41**(Database issue): p. D808-15.

392. Franz, M., et al., *Cytoscape.js: a graph theory library for visualisation and analysis.* Bioinformatics, 2016. **32**(2): p. 309-11.

393. Bailey, E., *shinyBS: Twitter Bootstrap Components for Shiny*. 2015.

394. Wickham, W.C.a.H., *ggvis: Interactive Grammar of Graphics*. 2016.

395. Mendez, M.G., S. Kojima, and R.D. Goldman, *Vimentin induces changes in cell shape, motility, and adhesion during the epithelial to mesenchymal transition.* FASEB J, 2010. **24**(6): p. 1838-51.

396. Keller, M.A., et al., *Transcriptional regulatory network analysis of developing human erythroid progenitors reveals patterns of coregulation and potential transcriptional regulators.* Physiol Genomics, 2006. **28**(1): p. 114-28.

397. Marshall, A.M., et al., *In vitro multipotent differentiation and barrier function of a human mammary epithelium.* Cell Tissue Res, 2009. **335**(2): p. 383-95.

398. Erickson, D.R., et al., *Differentiation associated changes in gene expression profiles of interstitial cystitis and control urothelial cells.* J Urol, 2008. **180**(6): p. 2681-7.

399. Dehne, T., et al., *Chondrogenic differentiation potential of osteoarthritic chondrocytes and their possible use in matrix-associated autologous chondrocyte transplantation.* Arthritis Res Ther, 2009. **11**(5): p. R133.

400. Sen, G.L., et al., *DNMT1 maintains progenitor function in self-renewing somatic tissue.* Nature, 2010. **463**(7280): p. 563-7.

401. Lopez-Pajares, V., et al., *A LncRNA-MAF:MAFB transcription factor network regulates epidermal differentiation.* Dev Cell, 2015. **32**(6): p. 693-706.

402. Martinez-Anton, A., et al., *Changes in microRNA and mRNA expression with differentiation of human bronchial epithelial cells.* Am J Respir Cell Mol Biol, 2013. **49**(3): p. 384-95.

403. Satish, L., et al., *Expression analysis of human adipose-derived stem cells during in vitro differentiation to an adipocyte lineage.* BMC Med Genomics, 2015. **8**: p. 41.

404. Nishida, Y., et al., *Identification and classification of genes regulated by phosphatidylinositol 3-kinase- and TRKB-mediated signalling pathways during neuronal differentiation in two subtypes of the human neuroblastoma cell line SH-SY5Y.* BMC Res Notes, 2008. **1**: p. 95.

405. Si-Tayeb, K., et al., *Highly efficient generation of human hepatocyte-like cells from induced pluripotent stem cells.* Hepatology, 2010. **51**(1): p. 297-305.

406. Murdoch, A.D., et al., *Chondrogenic differentiation of human bone marrow stem cells in transwell cultures: generation of scaffold-free cartilage.* Stem Cells, 2007. **25**(11): p. 2786-96.

407. Srinivasan, B., et al., *TEER measurement techniques for in vitro barrier model systems.* J Lab Autom, 2015. **20**(2): p. 107-26.

408. Krais, A.M., et al., *CHRNA5 as negative regulator of nicotine signaling in normal and cancer bronchial cells: effects on motility, migration and p63 expression.* Carcinogenesis, 2011. **32**(9): p. 1388-1395.

409. Ross, A.J., et al., *Transcriptional profiling of mucociliary differentiation in human airway epithelial cells.* Am J Respir Cell Mol Biol, 2007. **37**(2): p. 169-85.

410. Wong, A.P., et al., *Directed differentiation of human pluripotent stem cells into mature airway epithelia expressing functional CFTR protein.* Nature Biotechnology, 2012. **30**(9): p. 876-U108.

411. Dehne, T., et al., *Chondrogenic differentiation potential of osteoarthritic chondrocytes and their possible use in matrix-associated autologous chondrocyte transplantation.* Arthritis Research & Therapy, 2009. **11**(5).

412. Wang, B.C., et al., *UHRF1 suppression promotes cell differentiation and reduces inflammatory reaction in anaplastic thyroid cancer.* Oncotarget, 2016.

413. Bostick, M., et al., *UHRF1 plays a role in maintaining DNA methylation in mammalian cells.* Science, 2007. **317**(5845): p. 1760-4.

414. Pichiorri, F., et al., *Correlation of fragile histidine triad (Fhit) protein structural features with effector interactions and biological functions.* J Biol Chem, 2009. **284**(2): p. 1040-9.

415. Bhaduri, A., et al., *Network Analysis Identifies Mitochondrial Regulation of Epidermal Differentiation by MPZL3 and FDXR.* Dev Cell, 2015. **35**(4): p. 444-57.

416. Khavari, D.A., G.L. Sen, and J.L. Rinn, *DNA methylation and epigenetic control of cellular differentiation.* Cell Cycle, 2010. **9**(19): p. 3880-3.

417. Cline, M.S., et al., *Integration of biological networks and gene expression data using Cytoscape.* Nature Protocols, 2007. **2**(10): p. 2366-2382.

418. Banday, A.R., et al., *Replication-dependent histone genes are actively transcribed in differentiating and aging retinal neurons.* Cell Cycle, 2014. **13**(16): p. 2526-41.

419. Lyons, S.M., et al., *A subset of replication-dependent histone mRNAs are expressed as polyadenylated RNAs in terminally differentiated tissues.* Nucleic Acids Res, 2016. **44**(19): p. 9190-9205.

420. Saraiva, N.Z., C.S. Oliveira, and J.M. Garcia, *Histone acetylation and its role in embryonic stem cell differentiation.* World J Stem Cells, 2010. **2**(6): p. 121-6.

421. Modak, R., et al., *Probing p300/CBP associated factor (PCAF)-dependent pathways with a small molecule inhibitor.* ACS Chem Biol, 2013. **8**(6): p. 1311-23.

422. Wu, J. and D.L. Crowe, *The histone methyltransferase EZH2 promotes mammary stem and luminal progenitor cell expansion, metastasis and inhibits estrogen receptor-positive cellular differentiation in a model of basal breast cancer.* Oncol Rep, 2015. **34**(1): p. 455-60.

423. Tuncbag, N., et al., *Network-Based Interpretation of Diverse High-Throughput Datasets through the Omics Integrator Software Package.* Plos Computational Biology, 2016. **12**(4).

424. Bassel, G.W., et al., *Genome-wide network model capturing seed germination reveals coordinated regulation of plant cellular phase transitions.* Proc Natl Acad Sci U S A, 2011. **108**(23): p. 9709-14.

425. McMillan, E.M. and J. Quadrilatero, *Autophagy is required and protects against apoptosis during myoblast differentiation.* Biochemical Journal, 2014. **462**: p. 267-277.

426. Zhuang, W.Z., et al., *Induction of autophagy promotes differentiation of glioma-initiating cells and their radiosensitivity.* International Journal of Cancer, 2011. **129**(11): p. 2720-2731.

427. Cicchini, M., et al., *Autophagy regulator BECN1 suppresses mammary tumorigenesis driven by WNT1 activation and following parity.* Autophagy, 2014. **10**(11): p. 2036-52.

428. Wang, X.F., et al., *FGFR3/fibroblast growth factor receptor 3 inhibits autophagy through decreasing the ATG12-ATG5 conjugate, leading to the delay of cartilage development in achondroplasia.* Autophagy, 2015. **11**(11): p. 1998-2013.

429. Mattissek, C. and D. Teis, *The role of the endosomal sorting complexes required for transport (ESCRT) in tumorigenesis.* Mol Membr Biol, 2014. **31**(4): p. 111-9.

430. Tanaka, N., M. Kyuuma, and K. Sugamura, *Endosomal sorting complex required for transport proteins in cancer pathogenesis, vesicular transport, and non-endosomal functions.* Cancer Sci, 2008. **99**(7): p. 1293-303.

431. Hurley, J.H., *The ESCRT complexes.* Crit Rev Biochem Mol Biol, 2010. **45**(6): p. 463-87.

432. Dukes, J.D., et al., *Functional ESCRT machinery is required for constitutive recycling of claudin-1 and maintenance of polarity in vertebrate epithelial cells.* Mol Biol Cell, 2011. **22**(17): p. 3192-205.

433. Seok, H.Y., et al., *miR-155 inhibits expression of the MEF2A protein to repress skeletal muscle differentiation.* J Biol Chem, 2011. **286**(41): p. 35339-46.

434. Yoshida, T., et al., *The BCL6 gene is predominantly expressed in keratinocytes at their terminal differentiation stage.* Biochemical and Biophysical Research Communications, 1996. **228**(1): p. 216-220.

435. Walker, S.R., et al., *STAT5 outcompetes STAT3 to regulate the expression of the oncogenic transcriptional modulator BCL6.* Mol Cell Biol, 2013. **33**(15): p. 2879-90.

436. Lu, Q.L., et al., *bcl-2: role in epithelial differentiation and oncogenesis.* Hum Pathol, 1996. **27**(2): p. 102-10.

437. Zhang, K.Z., et al., *BCL2 regulates neural differentiation.* Proceedings of the National Academy of Sciences of the United States of America, 1996. **93**(9): p. 4504-4508.

438. Nagarajan, V.K., et al., *XRN 5'-->3' exoribonucleases: structure, mechanisms and functions.* Biochim Biophys Acta, 2013. **1829**(6-7): p. 590-603.

439. Nemer, M. and M.E. Horb, *The KLF family of transcriptional regulators in cardiomyocyte proliferation and differentiation.* Cell Cycle, 2007. **6**(2): p. 117-21.

440. Wan, H.J., et al., *Kruppel-like factor 5 is required for perinatal lung morphogenesis and function.* Development, 2008. **135**(15): p. 2563-2572.

441. Rosen, E.D. and O.A. MacDougald, *Adipocyte differentiation from the inside out.* Nat Rev Mol Cell Biol, 2006. **7**(12): p. 885-96.

442. Gordon, L.A., et al., *Breast cell invasive potential relates to the myoepithelial phenotype.* Int J Cancer, 2003. **106**(1): p. 8-16.

443. Deugnier, M.A., et al., *The importance of being a myoepithelial cell.* Breast Cancer Research, 2002. **4**(6): p. 224-230.

444. Tomaskovic-Crook, E., E.W. Thompson, and J.P. Thiery, *Epithelial to mesenchymal transition and breast cancer.* Breast Cancer Res, 2009. **11**(6): p. 213.

445. Janky, R., et al., *iRegulon: From a Gene List to a Gene Regulatory Network Using Large Motif and Track Collections.* Plos Computational Biology, 2014. **10**(7).

446. Manavathi, B., et al., *Derailed estrogen signaling and breast cancer: an authentic couple.* Endocr Rev, 2013. **34**(1): p. 1-32.

447. Fernando, R.I. and J. Wimalasena, *Estradiol abrogates apoptosis in breast cancer cells through inactivation of BAD: Ras-dependent nongenomic pathways requiring signaling through ERK and Akt.* Molecular Biology of the Cell, 2004. **15**(7): p. 3266-3284.

448. Coser, K.R., et al., *Global analysis of ligand sensitivity of estrogen inducible and suppressible genes in MCF7/BUS breast cancer cells by DNA microarray.* Proc Natl Acad Sci U S A, 2003. **100**(24): p. 13994-9.

449. Villalobos, M., et al., *The E-screen assay: a comparison of different MCF7 cell stocks.* Environ Health Perspect, 1995. **103**(9): p. 844-50.

450. Schneider-Poetsch, T., et al., *Inhibition of eukaryotic translation elongation by cycloheximide and lactimidomycin.* Nat Chem Biol, 2010. **6**(3): p. 209-217.

451. Ellison-Zelski, S.J., N.M. Solodin, and E.T. Alarid, *Repression of ESR1 through Actions of Estrogen Receptor Alpha and Sin3A at the Proximal Promoter.* Molecular and Cellular Biology, 2009. **29**(18): p. 4949-4958.

452. Iwamoto, T., et al., *Estrogen Receptor (ER) mRNA and ER-Related Gene Expression in Breast Cancers That Are 1% to 10% ER-Positive by Immunohistochemistry.* Journal of Clinical Oncology, 2012. **30**(7): p. 729-734.

453. Xu, H., et al., *DACH1 suppresses breast cancer as a negative regulator of CD44.* Sci Rep, 2017. **7**(1): p. 4361.

454. Xu, H., et al., *Enrichment of CD44 in basal-type breast cancer correlates with EMT, cancer stem cell gene profile, and prognosis.* Onco Targets Ther, 2016. **9**: p. 431-44.

455. Roson-Burgo, B., et al., *Insights into the human mesenchymal stromal/stem cell identity through integrative transcriptomic profiling.* BMC Genomics, 2016. **17**(1): p. 944.

456. Liu, R., et al., *Association Between Expression of Cancer Stem Cell Markers and Poor Differentiation of Hepatocellular Carcinoma: A Meta-Analysis (PRISMA).* Medicine (Baltimore), 2015. **94**(31): p. e1306.

457. Pfefferle, A.D., et al., *Luminal progenitor and fetal mammary stem cell expression features predict breast tumor response to neoadjuvant chemotherapy.* Breast Cancer Res Treat, 2015. **149**(2): p. 425-37.

458. Blumenberg, M., *Profiling and metaanalysis of epidermal keratinocytes responses to epidermal growth factor.* BMC Genomics, 2013. **14**: p. 85.

459. Sarrio, D., et al., *Epithelial and mesenchymal subpopulations within normal basal breast cell lines exhibit distinct stem cell/progenitor properties.* Stem Cells, 2012. **30**(2): p. 292-303.

460. Tobin, N.P., et al., *Cyclin D1, Id1 and EMT in breast cancer.* BMC Cancer, 2011. **11**: p. 417.

461. Ochsner, S.A., et al., *Transcriptomine, a web resource for nuclear receptor signaling transcriptomes.* Physiol Genomics, 2012. **44**(17): p. 853-63.

462. Chan, J., et al., *Eukaryotic protein synthesis inhibitors identified by comparison of cytotoxicity profiles.* RNA, 2004. **10**(3): p. 528-43.

463. Lavinsky, R.M., et al., *Diverse signaling pathways modulate nuclear receptor recruitment of N-CoR and SMRT complexes.* Proc Natl Acad Sci U S A, 1998. **95**(6): p. 2920-5.

464. Mangelsdorf, D.J., et al., *The nuclear receptor superfamily: the second decade.* Cell, 1995. **83**(6): p. 835-9.

465. Sever, R. and C.K. Glass, *Signaling by nuclear receptors.* Cold Spring Harb Perspect Biol, 2013. **5**(3): p. a016709.

466. Wu, Q., et al., *Non-nuclear estrogen receptor signaling in the endothelium.* J Biol Chem, 2011. **286**(17): p. 14737-43.

467. Gronemeyer, H., J.A. Gustafsson, and V. Laudet, *Principles for modulation of the nuclear receptor superfamily.* Nat Rev Drug Discov, 2004. **3**(11): p. 950-64.

468. Burris, T.P., S.A. Busby, and P.R. Griffin, *Targeting orphan nuclear receptors for treatment of metabolic diseases and autoimmunity.* Chem Biol, 2012. **19**(1): p. 51-9.

469. Germain, P., et al., *Overview of nomenclature of nuclear receptors.* Pharmacol Rev, 2006. **58**(4): p. 685-704.

470. Umesono, K., et al., *Direct Repeats as Selective Response Elements for the Thyroid-Hormone, Retinoic Acid, and Vitamin-D3 Receptors.* Cell, 1991. **65**(7): p. 1255-1266.

471. Yamamoto, K.R., *Steroid receptor regulated transcription of specific genes and gene networks.* Annu Rev Genet, 1985. **19**: p. 209-52.

472. Kliewer, S.A., et al., *An orphan nuclear receptor activated by pregnanes defines a novel steroid signaling pathway.* Cell, 1998. **92**(1): p. 73-82.

473. Glass, C.K., *Differential recognition of target genes by nuclear receptor monomers, dimers, and heterodimers.* Endocr Rev, 1994. **15**(3): p. 391-407.

474. Mangelsdorf, D.J. and R.M. Evans, *The RXR heterodimers and orphan receptors.* Cell, 1995. **83**(6): p. 841-50.

475. Badowska-Kozakiewicz, A.M., et al., *The role of oestrogen and progesterone receptors in breast cancer - immunohistochemical evaluation of oestrogen and progesterone receptor expression in invasive breast cancer in women.* Contemp Oncol (Pozn), 2015. **19**(3): p. 220-5.

476. Petersen, O.W., P.E. Hoyer, and B. van Deurs, *Frequency and distribution of estrogen receptor-positive cells in normal, nonlactating human breast tissue.* Cancer Res, 1987. **47**(21): p. 5748-51.

477. Bauer, K.R., et al., *Descriptive analysis of estrogen receptor (ER)-negative, progesterone receptor (PR)-negative, and HER2-negative invasive breast cancer, the so-called triple-negative phenotype: a population-based study from the California cancer Registry.* Cancer, 2007. **109**(9): p. 1721-8.

478. Roger, P., et al., *Decreased expression of estrogen receptor beta protein in proliferative preinvasive mammary tumors.* Cancer Res, 2001. **61**(6): p. 2537-41.

479. Mote, P.A., et al., *Loss of co-ordinate expression of progesterone receptors A and B is an early event in breast carcinogenesis.* Breast Cancer Res Treat, 2002. **72**(2): p. 163-72.

480. Liu, Y., et al., *Retinoic acid receptor beta mediates the growth-inhibitory effect of retinoic acid by promoting apoptosis in human breast cancer cells.* Mol Cell Biol, 1996. **16**(3): p. 1138-49.

481. Huang, P.X., V. Chandra, and F. Rastinejad, *Retinoic Acid Actions through Mammalian Nuclear Receptors.* Chemical Reviews, 2014. **114**(1): p. 233-254.

482. Ross-Innes, C.S., et al., *Cooperative interaction between retinoic acid receptor-alpha and estrogen receptor in breast cancer.* Genes & Development, 2010. **24**(2): p. 171-182.

483. Ingles, S.A., et al., *Vitamin D receptor genotype and breast cancer in Latinas (United States).* Cancer Causes Control, 2000. **11**(1): p. 25-30.

484. Bretherton-Watt, D., et al., *Vitamin D receptor gene polymorphisms are associated with breast cancer risk in a UK Caucasian population.* Br J Cancer, 2001. **85**(2): p. 171-5.

485. Colston, K.W., U. Berger, and R.C. Coombes, *Possible role for vitamin D in controlling breast cancer cell proliferation.* Lancet, 1989. **1**(8631): p. 188-91.

486. Guedj, M., et al., *A refined molecular taxonomy of breast cancer.* Oncogene, 2012. **31**(9): p. 1196-206.

487. Cochrane, D.R., et al., *Role of the androgen receptor in breast cancer and preclinical analysis of enzalutamide.* Breast Cancer Res, 2014. **16**(1): p. R7.

488. De Amicis, F., et al., *Androgen receptor overexpression induces tamoxifen resistance in human breast cancer cells.* Breast Cancer Res Treat, 2010. **121**(1): p. 1-11.

489. Bouras, T., M.C. Southey, and D.J. Venter, *Overexpression of the steroid receptor coactivator AIB1 in breast cancer correlates with the absence of estrogen and progesterone receptors and positivity for p53 and HER2/neu.* Cancer Res, 2001. **61**(3): p. 903-7.

490. Leo, J.C., et al., *Glucocorticoid and mineralocorticoid cross-talk with progesterone receptor to induce focal adhesion and growth inhibition in breast cancer cells.* Endocrinology, 2004. **145**(3): p. 1314-21.

491. Gomez-Sanchez, E. and C.E. Gomez-Sanchez, *The multifaceted mineralocorticoid receptor.* Compr Physiol, 2014. **4**(3): p. 965-94.

492. Massaad, C., et al., *A natural sequence consisting of overlapping glucocorticoid-responsive elements mediates glucocorticoid, but not androgen, regulation of gene expression.* Biochem J, 2000. **350 Pt 1**: p. 123-9.

493. Savory, J.G., et al., *Glucocorticoid receptor homodimers and glucocorticoid-mineralocorticoid receptor heterodimers form in the cytoplasm through alternative dimerization interfaces.* Mol Cell Biol, 2001. **21**(3): p. 781-93.

494. Kim, H.K., et al., *Distinctions in gastric cancer gene expression signatures derived from laser capture microdissection versus histologic macrodissection.* BMC Med Genomics, 2011. **4**: p. 48.

495. Nie, H., et al., *Mineralocorticoid receptor suppresses cancer progression and the Warburg effect by modulating the miR-338-3p-PKLR axis in hepatocellular carcinoma.* Hepatology, 2015. **62**(4): p. 1145-59.

496. Jeong, Y., et al., *Nuclear receptor expression defines a set of prognostic biomarkers for lung cancer.* PLoS Med, 2010. **7**(12): p. e1000378.

497. Chebotaev, D., A. Yemelyanov, and I. Budunova, *The mechanisms of tumor suppressor effect of glucocorticoid receptor in skin.* Mol Carcinog, 2007. **46**(8): p. 732-40.

498. Bhalla, S., et al., *Gene expression-based biomarkers for discriminating early and late stage of clear cell renal cancer.* Sci Rep, 2017. **7**: p. 44997.

499. Matthews, L.C., et al., *Glucocorticoid receptor regulates accurate chromosome segregation and is associated with malignancy.* Proc Natl Acad Sci U S A, 2015. **112**(17): p. 5479-84.

500. Yemelyanov, A., et al., *Tumor suppressor activity of glucocorticoid receptor in the prostate.* Oncogene, 2007. **26**(13): p. 1885-96.

501. Hu, X. and J.W. Funder, *The evolution of mineralocorticoid receptors.* Mol Endocrinol, 2006. **20**(7): p. 1471-8.

502. Conde, I., et al., *Glucocorticoid receptor changes its cellular location with breast cancer development.* Histol Histopathol, 2008. **23**(1): p. 77-85.

503. Pan, D., M. Kocherginsky, and S.D. Conzen, *Activation of the Glucocorticoid Receptor Is Associated with Poor Prognosis in Estrogen Receptor-Negative Breast Cancer.* Cancer Research, 2011. **71**(20): p. 6360-6370.

504. Skor, M.N., et al., *Glucocorticoid Receptor Antagonism as a Novel Therapy for Triple-Negative Breast Cancer.* Clinical Cancer Research, 2013. **19**(22): p. 6163-6172.

505. Bower, J.E., et al., *Fatigue and gene expression in human leukocytes: increased NF-kappaB and decreased glucocorticoid signaling in breast cancer survivors with persistent fatigue.* Brain Behav Immun, 2011. **25**(1): p. 147-50.

506. Stedman, M.R., et al., *Meta-analyses involving cross-over trials: methodological issues.* Int J Epidemiol, 2011. **40**(6): p. 1732-4.

507. Halvorsen, O.J., et al., *Gene expression profiles in prostate cancer: association with patient subgroups and tumour differentiation.* Int J Oncol, 2005. **26**(2): p. 329-36.

508. Marshall, F.F., *Previously unidentified changes in renal cell carcinoma gene expression identified by parametric analysis of microarray data.* J Urol, 2005. **173**(4): p. 1112.

509. Chen, L., et al., *Role of Deregulated microRNAs in Breast Cancer Progression Using FFPE Tissue.* Plos One, 2013. **8**(1).

510. Chopra, P., et al., *Improving Cancer Classification Accuracy Using Gene Pairs.* Plos One, 2010. **5**(12).

511. Yap, Y.L., et al., *Classification between normal and tumor tissues based on the pair-wise gene expression ratio.* Bmc Cancer, 2004. **4**.

512. Anglani, R., et al., *Loss of Connectivity in Cancer Co-Expression Networks.* Plos One, 2014. **9**(1).

513. Lumley, M.G.a.T., *forestplot: Advanced Forest Plot Using 'grid' Graphics*. 2017.

514. Weir, B.S. and C.C. Cockerham, *Estimating F-Statistics for the Analysis of Population Structure.* Evolution, 1984. **38**(6): p. 1358-1370.

515. RStudio, I., *Easy web applications in R.* 2014.

516. Doan, T.B., et al., *Breast cancer prognosis predicted by nuclear receptor-coregulator networks.* Mol Oncol, 2014. **8**(5): p. 998-1013.

517. Zhang, P., et al., *MiR-646 inhibited cell proliferation and EMT-induced metastasis by targeting FOXK1 in gastric cancer.* Br J Cancer, 2017.

518. Hu, Y., et al., *Candidate microRNAs as biomarkers of thyroid carcinoma: a systematic review, meta-analysis, and experimental validation.* Cancer Med, 2016. **5**(9): p. 2602-14.

519. Chow, Y.P., H. Alias, and R. Jamal, *Meta-analysis of gene expression in relapsed childhood B-acute lymphoblastic leukemia.* BMC Cancer, 2017. **17**(1): p. 120.

520. Peng, Z., et al., *HER2 discordance between paired primary gastric cancer and metastasis: a meta-analysis.* Chin J Cancer Res, 2015. **27**(2): p. 163-71.

521. Gandhi, V.V. and D.C. Samuels, *Correlated tissue expression of genes of cytoplasmic and mitochondrial nucleotide metabolisms in normal tissues is disrupted in transformed tissues.* Nucleosides Nucleotides Nucleic Acids, 2012. **31**(2): p. 112-29.

522. Falgreen, S., et al., *hemaClass.org: Online One-By-One Microarray Normalization and Classification of Hematological Cancers for Precision Medicine.* PLoS One, 2016. **11**(10): p. e0163711.

523. Sergushichev, A.A., et al., *GAM: a web-service for integrated transcriptional and metabolic network analysis.* Nucleic Acids Res, 2016. **44**(W1): p. W194-200.

524. Ghazanfar, S. and J.Y. Yang, *Characterizing mutation-expression network relationships in multiple cancers.* Comput Biol Chem, 2016. **63**: p. 73-82.

525. Rhodes, D.R., et al., *Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles.* Neoplasia, 2007. **9**(2): p. 166-80.

526. Cordero, D., et al., *Large differences in global transcriptional regulatory programs of normal and tumor colon cells.* BMC Cancer, 2014. **14**: p. 708.

527. Sun, L., et al., *Identification and characterization of a second fibronectin gene in zebrafish.* Matrix Biol, 2005. **24**(1): p. 69-77.

528. Yao, Y., S.C. Mack, and M.D. Taylor, *Molecular genetics of ependymoma.* Chin J Cancer, 2011. **30**(10): p. 669-81.

529. Jones, P.B. and D.H. White, *Reappraisal of the clinical use of leflunomide in rheumatoid arthritis and psoriatic arthritis.* Open Access Rheumatol, 2010. **2**: p. 53-71.

530. Zhu, S., et al., *Leflunomide reduces proliferation and induces apoptosis in neuroblastoma cells in vitro and in vivo.* PLoS One, 2013. **8**(8): p. e71555.

531. Alhefdhi, A., et al., *Leflunomide suppresses growth in human medullary thyroid cancer cells.* J Surg Res, 2013. **185**(1): p. 212-6.

Appendix V

Copyright permission for Figure 1.1.

NATURE PUBLISHING GROUP LICENSE
TERMS AND CONDITIONS

Aug 25, 2017

This Agreement between Ms. Huma Shehwana ("You") and Nature Publishing Group ("Nature Publishing Group") consists of your license details and the terms and conditions provided by Nature Publishing Group and Copyright Clearance Center.

License Number	4174810546463
License date	Aug 23, 2017
Licensed Content Publisher	Nature Publishing Group
Licensed Content Publication	Nature Reviews Genetics
Licensed Content Title	Expression profiling [mdash] best practices for data generation and interpretation in clinical trials
Licensed Content Author	
Licensed Content Date	Mar 1, 2004
Licensed Content Volume	5
Licensed Content Issue	3
Type of Use	reuse in a dissertation / thesis
Requestor type	academic/educational
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
High-res required	no
Figures	Figure 1
Author of this NPG article	no
Your reference number	Figure 1.1
Title of your thesis / dissertation	DEVELOPMENT OF NOVEL TOOLS FOR CANCER DIAGNOSIS, PROGNOSIS AND TREATMENT USING INTRA- OR INTER-SPECIES TRANSCRIPTOME META-ANALYSIS
Expected completion date	Sep 2017
Estimated size (number of pages)	280
Requestor Location	Ms. Huma Shehwana Bilkent University Ankara, Ankara 06800 Turkey Attn: Ms. Huma Shehwana

Copyright permission for Figure 1.2.

ELSEVIER LICENSE
TERMS AND CONDITIONS

Aug 25, 2017

This Agreement between Ms. Huma Shehwana ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

License Number	4174960258182
License date	Aug 23, 2017
Licensed Content Publisher	Elsevier
Licensed Content Publication	European Journal of Cancer
Licensed Content Title	Cancer gene expression signatures – The rise and fall?
Licensed Content Author	Frederic Chibon
Licensed Content Date	May 1, 2013
Licensed Content Volume	49
Licensed Content Issue	8
Licensed Content Pages	10
Start Page	2000
End Page	2009
Type of Use	reuse in a thesis/dissertation
Intended publisher of new work	other
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
Format	both print and electronic
Are you the author of this Elsevier article?	No
Will you be translating?	No
Order reference number	Figure 1.2
Original figure numbers	Figure 1
Title of your thesis/dissertation	DEVELOPMENT OF NOVEL TOOLS FOR CANCER DIAGNOSIS, PROGNOSIS AND TREATMENT USING INTRA- OR INTER-SPECIES TRANSCRIPTOME META-ANALYSIS
Expected completion date	Sep 2017
Estimated size (number of pages)	280
Requestor Location	Ms. Huma Shehwana Bilkent University

Copyright permission for Figure 1.3.

NATURE PUBLISHING GROUP LICENSE
TERMS AND CONDITIONS

Aug 25, 2017

This Agreement between Ms. Huma Shehwana ("You") and Nature Publishing Group ("Nature Publishing Group") consists of your license details and the terms and conditions provided by Nature Publishing Group and Copyright Clearance Center.

License Number	4174961186092
License date	Aug 23, 2017
Licensed Content Publisher	Nature Publishing Group
Licensed Content Publication	Nature Reviews Genetics
Licensed Content Title	Microarray data analysis: from disarray to consolidation and consensus
Licensed Content Author	David B. Allison, Xiangqin Cui, Grier P. Page and Mahyar Sabripour
Licensed Content Date	Jan 1, 2006
Licensed Content Volume	7
Licensed Content Issue	1
Type of Use	reuse in a dissertation / thesis
Requestor type	academic/educational
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
High-res required	no
Figures	Figure 2 of paper is used as Figure 1.3
Author of this NPG article	no
Your reference number	Figure 2
Title of your thesis / dissertation	DEVELOPMENT OF NOVEL TOOLS FOR CANCER DIAGNOSIS, PROGNOSIS AND TREATMENT USING INTRA- OR INTER-SPECIES TRANSCRIPTOME META-ANALYSIS
Expected completion date	Sep 2017
Estimated size (number of pages)	280
Requestor Location	Ms. Huma Shehwana Bilkent University Ankara, Ankara 06800 Turkey Attn: Ms. Huma Shehwana
Billing Type	Invoice
Billing Address	Ms. Huma Shehwana Bilkent University

Copyright permission for Figure 1.4.



BMC Bioinformatics. 2008; 9: 559.
Published online 2008 Dec 29. doi: 10.1186/1471-2105-9-559
PMCID: PMC2631488

WGCNA: an R package for weighted correlation network analysis

Reviewed by Peter Langfelder[1] and Steve Horvath[2]

Author information ▶ Article notes ▶ Copyright and License information ▼

Copyright © 2008 Langfelder and Horvath; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Copyright permission for Figure 1.5.

OXFORD UNIVERSITY PRESS LICENSE
TERMS AND CONDITIONS

Aug 25, 2017

This Agreement between Ms. Huma Shehwana ("You") and Oxford University Press ("Oxford University Press") consists of your license details and the terms and conditions provided by Oxford University Press and Copyright Clearance Center.

License Number	4174870084779
License date	Aug 23, 2017
Licensed content publisher	Oxford University Press
Licensed content publication	Nucleic Acids Research
Licensed content title	Comprehensive literature review and statistical considerations for microarray meta-analysis
Licensed content author	Tseng, George C.; Ghosh, Debashis
Licensed content date	Jan 19, 2012
Type of Use	Thesis/Dissertation
Institution name	
Title of your work	DEVELOPMENT OF NOVEL TOOLS FOR CANCER DIAGNOSIS, PROGNOSIS AND TREATMENT USING INTRA- OR INTER-SPECIES TRANSCRIPTOME META-ANALYSIS
Publisher of your work	n/a
Expected publication date	Sep 2017
Permissions cost	0.00 USD
Value added tax	0.00 USD
Total	0.00 USD
Requestor Location	Ms. Huma Shehwana Bilkent University Ankara, Ankara 06800 Turkey Attn: Ms. Huma Shehwana
Publisher Tax ID	GB125506730
Billing Type	Invoice
Billing Address	Ms. Huma Shehwana Bilkent University Ankara, Turkey 06800 Attn: Ms. Huma Shehwana
Total	0.00 USD
Terms and Conditions	

Copyright permission for Table 1.1.

ELSEVIER LICENSE
TERMS AND CONDITIONS

Aug 25, 2017

This Agreement between Ms. Huma Shehwana ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

License Number	4174850327818
License date	Aug 23, 2017
Licensed Content Publisher	Elsevier
Licensed Content Publication	The American Journal of Surgery
Licensed Content Title	How to write a surgical clinical research protocol: literature review and practical guide
Licensed Content Author	Rachel Rosenthal,Juliane Schäfer,Matthias Briel,Heiner C. Bucher,Daniel Oertli,Salome Dell-Kuster
Licensed Content Date	Feb 1, 2014
Licensed Content Volume	207
Licensed Content Issue	2
Licensed Content Pages	14
Start Page	299
End Page	312
Type of Use	reuse in a thesis/dissertation
Intended publisher of new work	other
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
Format	both print and electronic
Are you the author of this Elsevier article?	No
Will you be translating?	No
Order reference number	Table 1.2
Original figure numbers	Table 6
Title of your thesis/dissertation	DEVELOPMENT OF NOVEL TOOLS FOR CANCER DIAGNOSIS, PROGNOSIS AND TREATMENT USING INTRA- OR INTER-SPECIES TRANSCRIPTOME META-ANALYSIS
Expected completion date	Sep 2017
Estimated size (number of pages)	280
Requestor Location	Ms. Huma Shehwana Bilkent University

Copyright permission for Table 1.2.

SPRINGER LICENSE
TERMS AND CONDITIONS

Aug 26, 2017

This Agreement between Ms. Huma Shehwana ("You") and Springer ("Springer") consists of your license details and the terms and conditions provided by Springer and Copyright Clearance Center.

License Number	4176540595940
License date	Aug 26, 2017
Licensed Content Publisher	Springer
Licensed Content Publication	Clinical and Translational Oncology
Licensed Content Title	Tools for protein-protein interaction network analysis in cancer research
Licensed Content Author	Rebeca Sanz-Pamplona
Licensed Content Date	Jan 1, 2012
Licensed Content Volume	14
Licensed Content Issue	1
Type of Use	Thesis/Dissertation
Portion	Figures/tables/illustrations
Number of figures/tables/illustrations	1
Author of this Springer article	No
Country of republication	other
Order reference number	Table 1 of paper as Table 1.3 in thesis
Original figure numbers	Table 1
Title of your thesis / dissertation	DEVELOPMENT OF NOVEL TOOLS FOR CANCER DIAGNOSIS, PROGNOSIS AND TREATMENT USING INTRA- OR INTER-SPECIES TRANSCRIPTOME META-ANALYSIS
Expected completion date	Sep 2017
Estimated size(pages)	280
Requestor Location	Ms. Huma Shehwana Bilkent University Ankara, Ankara 06800 Turkey Attn: Ms. Huma Shehwana
Billing Type	Invoice
Billing Address	Ms. Huma Shehwana Bilkent University Ankara, Turkey 06800 Attn: Ms. Huma Shehwana

Copyright permission for Table 1.4.





Title:	COSMIC: exploring the world's knowledge of somatic mutations in human cancer
Author:	Forbes, Simon A.; Beare, David
Publication:	Nucleic Acids Research
Publisher:	Oxford University Press
Date:	2014-10-29

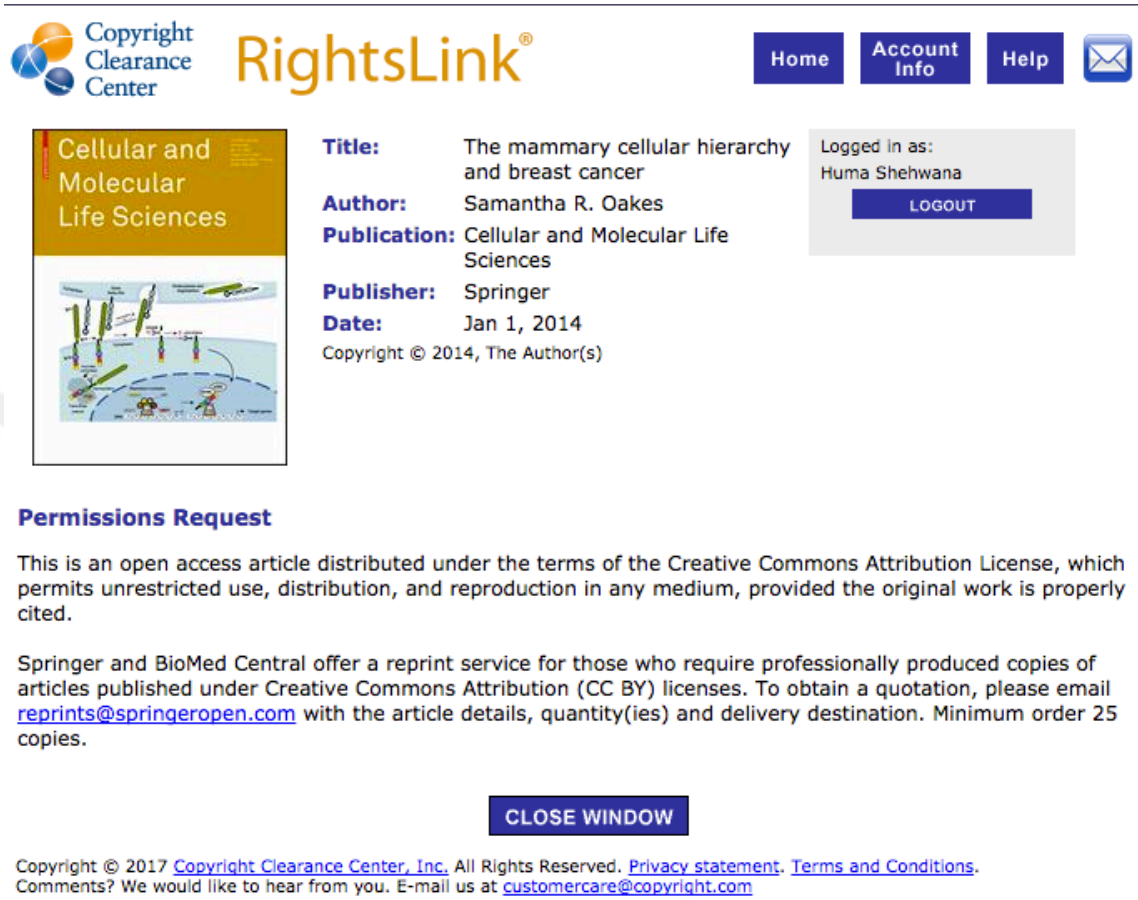
Copyright © 2014, Oxford University Press

Creative Commons

This is an open access article distributed under the terms of the Creative Commons CC BY license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

You are not required to obtain permission to reuse this article.

Copyright permission for Fig 4.1.



Copyright Clearance Center RightsLink®

Home | Account Info | Help

Title: The mammary cellular hierarchy and breast cancer
Author: Samantha R. Oakes
Publication: Cellular and Molecular Life Sciences
Publisher: Springer
Date: Jan 1, 2014
Copyright © 2014, The Author(s)

Logged in as:
Huma Shehwana
LOGOUT

Permissions Request

This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Springer and BioMed Central offer a reprint service for those who require professionally produced copies of articles published under Creative Commons Attribution (CC BY) licenses. To obtain a quotation, please email reprints@springeropen.com with the article details, quantity(ies) and delivery destination. Minimum order 25 copies.

CLOSE WINDOW

Copyright © 2017 Copyright Clearance Center, Inc. All Rights Reserved. Privacy statement. Terms and Conditions.
Comments? We would like to hear from you. E-mail us at customercare@copyright.com

Copyright permission for Figure 4.2.

OXFORD UNIVERSITY PRESS LICENSE
TERMS AND CONDITIONS

Aug 22, 2017

This Agreement between Ms. Huma Shehwana ("You") and Oxford University Press ("Oxford University Press") consists of your license details and the terms and conditions provided by Oxford University Press and Copyright Clearance Center.

License Number	4174440392544
License date	Aug 22, 2017
Licensed content publisher	Oxford University Press
Licensed content publication	Molecular Endocrinology
Licensed content title	Minireview: Nuclear Receptors and Breast Cancer
Licensed content author	Conzen, Suzanne D.
Licensed content date	Oct 1, 2008
Type of Use	Thesis/Dissertation
Institution name	
Title of your work	DEVELOPMENT OF NOVEL TOOLS FOR CANCER DIAGNOSIS, PROGNOSIS AND TREATMENT USING INTRA- OR INTER-SPECIES TRANSCRIPTOME META-ANALYSIS
Publisher of your work	n/a
Expected publication date	Sep 2017
Permissions cost	0.00 USD
Value added tax	0.00 USD
Total	0.00 USD
Requestor Location	Ms. Huma Shehwana Bilkent University Ankara, Ankara 06800 Turkey Attn: Ms. Huma Shehwana
Publisher Tax ID	GB125506730
Billing Type	Invoice
Billing Address	Ms. Huma Shehwana Bilkent University