BRCA:Cohort-vs-TCGA: A WEBTOOL FOR COMPREHENSIVE EXPLORATION AND COMPARISON OF MUTATIONAL LANDSCAPE OF BREAST CANCER COHORTS WITH TCGA-BRCA

A THESIS SUBMITTED TO

THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE

OF BILKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENETS FOR

THE DEGREE OF

MASTER OF SCIENCE

IN

MOLECULAR BIOLOGY AND GENETICS

By

Farid Ahadli

August 2022

BRCA:Cohort-vs-TCGA: A WEBTOOL FOR COMPREHENSIVE EXPLORATION AND COMPARISON OF MUTATIONAL LANDSCAPE OF BREAST CANCER COHORTS WITH TCGA-BRCA

By Farid Ahadli,

July, 2022

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

Özlen Konu Karakayalı
(advisor)

Serkan Belkaya

Nurcan Tunçbağ

Approved for the Graduate School of Engineering and Science

Orhan Arıkan
Director of the Graduate School of Engineering and Science

ABSTRACT

BRCA:Cohort-vs-TCGA: A WEBTOOL FOR COMPREHENSIVE EXPLORATION AND COMPARISON OF MUTATIONAL LANDSCAPE OF BREAST CANCER COHORTS WITH TCGA-BRCA

Farid Ahadli

M.S in Molecular Biology and Genetics

Advisor: Özlen Konu Karakayalı

August 2022

Comparisons of mutational landscapes between independent breast cancer (BRCA) cohorts in a comprehensive and statistical manner using online tools are likely to help advance our understanding of the diversity of somatic interactions and mutational signatures. In this thesis I have developed a webtool called BRCA:Cohort-vs-TCGA, which makes the mutational landscape comparison between a breast cancer cohort and TCGA-BRCA cohort accessible to the general public. BRCA:Cohort-vs-TCGA app contains modules to conduct mutational signature identification and comparison, differentially mutated gene and pathway identification and comparison, driver gene identification, and somatic interaction identification and comparison. Among the applications generated for similar purposes in literature, BRCA:Cohort-vs-TCGA

provides advantages because of its comparative, statistical, and signature related features.

When BRCA:Cohort-vs-TCGA was applied to compare the Gustave Roussy Institute (IGR) metastatic breast cancer cohort with the TCGA-BRCA cohort, a unique to IGR mutational signature related to SBS17b, a reactive oxygen species associated signature, which was missed by the original analysis of the authors, was identified. Interestingly, contribution of SBS17b was significantly higher in TNBC samples when compared to HER+ and HR+ samples. Differentially mutated gene analysis identified four genes, namely, *ESR1*, *FSIP2*, *ARMCX4* and *MALRD4*, of which the latter two were unique to re-analysis by BRCA:Cohort-vs-TCGA. Differentially mutated pathway analysis on the other hand pointed to the differential mutation of DNA damage related pathways, NRF2 pathway and WNT pathway. The influence of the hypermutator samples in Fisher's exact test based analyses is well documented, as a result of many spurious mutations these samples contain. Re-analysis of the IGR cohort after removal of the hypermutator samples showed *ARMCX4* was no longer differentially mutated anymore.

Re-analysis of the differentially mutated pathways resulted in identification of three pathways, namely Mismatch Repair - Core, LIN37 independent P53 Targets, and U5, all of which had an odds ratio of less than one. These results, however, could be biased by the presence of hypermutators in the TCGA-BRCA cohort.

In the second case study, I have demonstrated how the extracted somatic interactions for a user selected gene, i.e., *CHRNA5* and *TP53*, can be used to prioritize and filter genes. This was done by merging the somatic interactions with the differential expression profiles of MCF7 cells treated with siRNA against *CHRNA5* and *TP53*. Accordingly,

BRCA:Cohort-vs-TCGA can help annotate/enrich other high throughput data with somatic interactions of custom genes.

Key words: TCGA-BRCA; BRCA; TNBC; somatic mutation; MAF; VCF; hypermutation; Shiny; cancer driver genes

ÖZET

BRCA:Cohort-vs-TCGA: MEME KANSERİ KOHORTLARININ MUTASYON PROFİLLERİNİN KAPSAMLI İNCELENMESİ VE TCGA-BRCA İLE KIYASLANMASI İÇİN BİR WEB ARACI

Farid Ahadli

Moleküler Biyoloji ve Genetik, Yüksek Lisans

Tez Danışmanı: Özlen Konu Karakayalı

Ağustos 2022

Çevrimiçi araçları kullanarak bağımsız meme kanseri (BRCA) kohortları arasındaki mutasyonel farklılıkların kapsamlı ve istatistiksel bir şekilde karşılaştırılması, somatik etkileşimlerin ve mutasyon imzalarının çeşitliliği konusundaki anlayışımızı geliştirmemize yardımcı olabilir. Bu tezde, bir meme kanseri kohortu ile TCGA-BRCA kohortu arasındaki mutasyonel karşılaştırma analizlerini araştırmacılar için erişilebilir kılan BRCA:Cohort-vs-TCGA adlı bir web aracı geliştirdim. BRCA:Cohort-vs-TCGA uygulaması, mutasyonel imza belirleme ve karşılaştırma, diferansiyel mutasyona uğramış gen ve yolak belirleme ve karşılaştırma, kansere neden olan gen tanımlama, ve somatik etkileşim tanımlama ve karşılaştırma yapmak için modüller içerir. Literatürde benzer amaçlarla oluşturulan uygulamalardan BRCA:Cohort-vs-TCGA karşılaştırmalı, istatistiksel ve imza ile ilgili özellikleri nedeniyle üstündür.

Gustave Roussy Enstitüsü (IGR) metastatik meme kanseri kohortunu TCGA-BRCA kohortuyla karşılaştırmak için BRCA:Cohort-vs-TCGA uygulandığında, reaktif oksijen türleri kaynaklı olan, SBS17b ile benzer, ve yazarların orijinal analizinde gözden kaçan yeni bir mutasyonel imza tespit edildi. İlginç bir şekilde, SBS17b'nin katkısı, HER+ ve HR+ numunelerine kıyasla TNBC numunelerinde önemli ölçüde daha yüksekti. Diferansiyel olarak mutasyona uğramış gen analizi sonucunda, *ESR1, FSIP2, ARMCX4* ve *MALRD4* olmak üzere dört genin metastatik örneklere özgü olduğu bulgulandı. Bunlardan son ikisi BRCA:Cohort-vs-TCGA tarafından yapılan analizlere özeldi. Diferansiyel olarak mutasyona uğramış yolak analizi ise, DNA hasarı ile ilgili yolakların, NRF2 ve WNT yolağının diferansiyel mutasyonuna işaret etti. Yolak temelli analiz bu teze özgü olmakla beraber ne IGR kohortunun orijinal analizi ne de başkaları tarafından denenmiştir. Hipermutatör numunelerin analizlerindeki etkisi iyi belgelenmiş olup bu numunelerin içerdiği birçok passenger mutasyonların sonucu olduğu düşünülmektedir. Hipermutatör numunelerin çıkarılmasından sonra IGR kohortunun yeniden analizi farklı sonuçlar verdi. Örneğin, hipermutatör numunelerde çok sayıda bulunması nedeniyle *ARMCX4*'ün artık farklı şekilde mutasyona uğramadığı bulgulandı..

Diferansiyel olarak mutasyona uğramış yolakların yeniden analizi, tümü birden küçük olasılık oranına sahip olan Mismatch Repair - Core, LIN37 bağımsız P53 Hedefleri ve U5 olmak üzere üç yeni yolağın bulunmasıyla sonuçlandı. Bununla birlikte, bu sonuçlar, TCGA-BRCA kohortunda hipermutatörlerin varlığından etkilenmiş olabilir.

İkinci vaka çalışmasında, kullanıcı tarafından seçilen bir gen, yani *CHRNA5* ve *TP53,* için çıkarılan somatik etkileşimlerin, genleri önceliklendirmek ve filtrelemek için nasıl kullanılabileceğini gösterdim. Bu, *CHRNA5* ve *TP53*'e karşı siRNA'a maruz bırakılan

MCF7 hücrelerinin diferansiyel ekspresyon profilleri ile somatik etkileşimleri birleştirerek yapıldı. Buna göre, BRCA:Cohort-vs-TCGA, özel genlerin somatik etkileşimleriyle diğer yüksek verimli verisetlerini anlamdırmaya yardımcı olabilir.

Anahtar sözcükler: TCGA-BRCA; BRCA; TNBC; somatik mutasyon; MAF; VCF; hipermutasyon; Shiny; kansere neden olan genler.

Aldada bilməmiş dünyanın varı
Bir məslək eşqilə yaşayanları!

-Səməd Vurğun, "Vaqif" dramı

ACKNOWLEDGEMENTS

I would like to start by extending gratitude and love to my family. They allowed me to embark on a journey unknown to them and fully supported me financially and emotionally. I will forever be indebted to them.

Next, I would like to thank my advisor, Associate Professor Özlen Konu Karakayalı. She accepted me to her lab by compromising her preferences for the duration of a master's degree. Participating in various projects in her lab motivated me to learn a wide array of techniques and helped me to build a competitive curriculum vitae. She was also tolerant throughout the duration of our work and treated my mental well-being with utmost care.

I am grateful to Professor Ali Osmay Güre, my previous advisor, for supervising me from the senior year of my undergraduate studies to the second year of my master's studies. It was under his supervision that I discovered my propensity for computational biology and ability to self-learn. I learned a lot from him in matters outside of academia as well, such as persuasion.

If it was not for our collaboration with Professor Cengiz Yakıcıer, we would not have been introduced to the world of cancer cohort studies, and this thesis would not exist. He was extremely polite during our interactions and always used plural subject pronouns in Turkish when addressing me. I am grateful to him.

Past and previous members of the Konu and Güre labs were always there to support and guide me in bad times. They were faithful friends and brilliant co-workers, and I would like to thank them for making graduate life tolerable.

I am grateful to the dear jury members, Assistant Professor Serkan Belkaya and Associate Professor Nurcan Tunçbağ, for dedicating their time to read this thesis and suggesting invaluable additions without which this thesis would be incomplete.

I would like to acknowledge that during my thesis research I was partially supported by the graduate stipend of Bilkent University and the TÜBİTAK 1001 research grant № 219Z029 awarded to Associate Professor Özlen Konu Karakayalı.

TABLE OF CONTENTS

ABSTRACT ..... **I**

ÖZET ..... **IV**

1 CHAPTER 1 INTRODUCTION **1**

1.1 Cancer and Its Hallmarks **1**

1.2 Breast Cancer **2**

1.2.1 Epidemiology of Breast Cancer 2

1.2.2 Intrinsic Subtypes of Breast Cancer 2

1.2.3 Molecular Subtypes of Breast Cancer 3

1.3 Mutation Detection and Filtration **3**

1.3.1 Mutation Calling 3

1.3.2 Detection of Somatic Variants 5

1.3.3 Filtration of the Variants 6

1.3.4 Mutation Annotation Format (MAF). 7

1.3.5 Multi-Center Mutation Calling in Multiple Cancers (MC3) pipeline 7

1.4 Mutational Signatures **8**

1.4.1 Introduction to The Concept of Mutational Signatures 8

1.4.2 Non-Negative Matrix Factorization 9

1.4.3 Selection of Optimal Number of Signatures 10

1.4.4 Comparing The Similarity of Signatures 11

1.5 Cancer Driver Genes **11**

1.5.1 Introduction 11

1.5.2 Mutational Rate Based Methods 12

1.5.3 Positional Clustering Based Methods 13

1.5.4 Latent Driver Gene Approach 14

1.5.5 Tumor Suppressor Gene and Oncogene Identification 14

1.6 Methods for Identification of Somatic Interactions and Their Visualization. 15

1.7 Oncogenic Pathways in Cancer 16

1.7.1 RTK/RAS signaling pathway 17

1.7.2 PI3K signaling pathway 18

1.7.3 MYC signaling pathway 20

1.7.4 TP53 Pathway 22

1.7.5 Cell Cycle Pathway 23

1.7.6 Applications For Analysis of Cancer Cohorts. 25

2 CHAPTER 2 AIMS AND RATIONALE 27

3 CHAPTER 3 MATERIALS AND METHODS 28

3.1 Materials 28

3.2 Methods 32

3.2.1 Data Access and Storage 33

3.2.2 Pathway annotations 33

3.2.3 Mutational Signature Analysis 33

3.2.4 Identification of Differentially Mutated Genes and Pathways 34

3.2.5 Identification of Somatic Interactions. 34

3.2.6 Identification of Driver Genes 34

3.2.7 Harmonization of Gene Names 35

3.2.8 Identification of Optimal Variant Allele Frequency Cutoff 35

3.2.9 Reactivity 35

3.2.10 Hypermutator Detection 36

4 CHAPTER 4 RESULTS 38

4.1 BRCA:Cohort-vs-TCGA App 38

4.1.1 "Data Upload" Tab 38

4.1.2 "Mutation and Clinical Data" Tab 39

4.1.3 "Hypermutator Detection and Filtration" 41

4.1.4 "Oncoplot" Tab 42

4.1.5 "Comparison of TMB" Tab 43

4.1.6 "Mutational Signatures" Tab 45

4.2 "Clinical Enrichment" Tab 54

4.3 "Comparison with TCGA" Tab 56

4.3.1 "Differentially Mutated Genes" Subtab 56

4.3.2 "Cancer Driver Genes Based on Positional Clustering" Subtab 58

4.3.3 "Somatic Interactions" Subtab 61

4.3.4 "Comparison of Mutational Signatures" Subtab 64

4.3.5 "Comparison of Pathways" Subtab. 68

4.4 Case Study of The Gustave Roussy Institute Cohort 71

4.4.1 Features of Cohort Used for The Case Study 71

4.4.2 Data Processing Pipeline Used in The Original Study 71

4.4.3 Overview of The Mutation Spectrum of The IGR Cohort 72

4.4.4 Determination of The Optimal VAF cutoff for The Cohort Comparison 74

4.4.5 Mutational Signature Analysis 76

4.4.5.1 Detection of Optimal Number of Mutational Signatures 76

4.4.5.2 Fitting of Signatures 78

4.4.5.3 Comparison of the Signatures 80

4.4.5.4 Contribution of Signatures to Each Patient in The IGR Cohort. 83

4.4.5.5 Correlation Between Signatures and TMB in the IGR cohort. 85

4.4.5.6 Relationship Between Signatures and the HR and HER Status of Samples in the IGR Cohort 89

4.4.5.7 Relationship Between Signatures and the Metastasis Location of Samples in the IGR Cohort 92

4.4.6 Differentially Mutated Gene Analysis between IGR and TCGA Cohorts 98

4.4.7 Identification of the Driver Genes in IGR and TCGA cohorts 99

4.4.8 Clinical Enrichment Analysis of Mutations in The IGR Cohort 104

4.4.8.1 Clinical Enrichment Analysis by HR/HER subtypes in The IGR Cohort. 104

4.4.8.2 Clinical Enrichment Analysis by Metastasis Site in The IGR Cohort 106

4.4.9 Somatic Interaction Analysis in TCGA and IGR cohorts. 108

4.4.9.1 Somatic Interaction Analysis in the IGR cohort. 108

4.4.9.2 Somatic Interaction Analysis in the TCGA cohort. 108

4.4.9.3 Comparison of somatic interactions between TCGA and IGR cohorts. 108

4.4.10 Differentially Mutated Pathways between IGR and TCGA cohorts. 112

4.4.10.1 Testing the Validity of the Differential Mutation of the DNA Damage Repair related Pathways with Mutational Signatures 114

4.4.10.2 Testing the Validity of the Differential Mutation of the NRF2 Pathways with Mutational Signatures 114

4.5 Case Study of the Network of CHRNA5 Somatic Interactions in TCGA-BRCA Dataset 118

4.6 Comparison of BRCA:Cohort-vs-TCGA App to Similar Apps 125

5 CHAPTER 5 DISCUSSION 127

5.1 Case Studies 127

5.1.1 Mutational Signature Analysis 127

5.1.2 Differentially Mutated Genes and Cancer Driver Genes. 129

5.1.3 Determination of Optimal VAF cutoff 132

5.1.4 Clinical Enrichment Analyses 133

5.1.5 Somatic Interactions 135

5.1.6 Differentially Mutated Pathway Analysis 135

5.1.7 *CHRNA5* Case Study in the TCGA-BRCA dataset 138

6 FUTURE PERSPECTIVES 140

6.1 Development of BRCA:Cohort-vs-TCGA **140**

6.2 Potential Case Studies with BRCA:Cohort-vs-TCGA **140**

7 REFERENCES **142**

8 APPENDIX **165**

LIST OF FIGURES

Figure 1. Components of the RTK/RAS pathway and interaction between them. Pointed arrows denote the activatory interactions, while blunted arrows denote the inhibitory interactions. The image was generated via the PathwayMapper software [75]. 18

Figure 2. Components of the PI3K pathway and interactions between them. Pointed arrows denote the activatory interactions, while blunted arrows denote the inhibitory interactions. The image was generated via the PathwayMapper software [75]. 19

Figure 3. Components of the MYC pathway and interactions between them. Pointed arrows denote the activatory interactions, while blunted arrows denote the inhibitory interactions. The image was generated via the PathwayMapper software [75]. 21

Figure 4. Components of the TP53 pathway and interactions between them. Pointed arrows denote the activatory interactions, while blunted arrows denote the inhibitory interactions. The image was generated via the PathwayMapper software [75]. 23

Figure 5. Components of the cell cycle pathway and interactions between them. Pointed arrows denote the activatory interactions, while blunted arrows denote the inhibitory interactions. The image was generated via the PathwayMapper software [75]. 25

Figure 6. A flowchart that describes the workflow of the application. 37

Figure 7. A screenshot from the “Data Upload” tab of the “BRCA:Cohort-vs-TCGA” app. 39

Figure 8. A screenshot from the “Mutation and Clinical Data” tab of the “BRCA:Cohort-vs-TCGA” app. 39

Figure 9. A screenshot displaying the filter options for the numeric columns, from the “Mutation and Clinical Data” tab of the “BRCA:Cohort-vs-TCGA” app. 41

Figure 10. A screenshot from the “Hypermutator Detection and Filtration” tab of the “BRCA:Cohort-vs-TCGA” app. 42

Figure 11. A screenshot from the “Oncoplot” tab of the “BRCA:Cohort-vs-TCGA” app. 43

Figure 12. A screenshot from the “Comparison of TMB” tab of the “BRCA:Cohort-vs-TCGA” app. 45

Figure 13. A screenshot from the “Mutational Signatures” tab of the “BRCA:Cohort-vs-TCGA” app. 46

Figure 14. A screenshot from the “Identified signatures and their cosine similarity to SBS signatures” subtab of the “Mutational Signatures” tab. 47

Figure 15. A screenshot from the “Signature Contribution” subtab of the “Mutational Signatures” tab. 48

Figure 16. A screenshot from the “TMB vs Contribution” subtab of the “Mutational Signatures” tab. 49

Figure 17. A screenshot from the “ANOVA of contributions” subtab of the “Mutational Signatures” tab. 50

Figure 18. A screenshot from the “Correlation with a Continuous Variable” subtab of the “Mutational Signatures” tab. 51

Figure 19. A screenshot from the “ANOVA by a character column” subtab of the “Mutational Signatures” tab. 52

Figure 20. A screenshot from the “Pairwise Comparison” subtab of the “Mutational Signatures” tab. 53

Figure 21. A screenshot from the “Pairwise Comparison within Signatures” subtab of the “Mutational Signatures” tab. 54

Figure 22. A screenshot from the “Pairwise Enrichment Results Table” subtab of the “Clinical enrichment” tab. 55

Figure 23. A screenshot from the “Groupwise Enrichment results” subtab of the “Clinical enrichment” tab. 55

Figure 24. A screenshot from the “Clinical Enrichment plot” subtab of the “Clinical enrichment” tab. The height of the barplots shows the fraction of samples with the specified gene mutations, and the colors represent each level of the chosen clinical column. The numerator in each representation indicates the number of samples with mutation meanwhile, the denominator indicates the number of samples in each level. 56

Figure 25. A screenshot from the “Fisher’s exact test results” subtab of the “Differentially Mutated Genes” tab. 57

Figure 26. A screenshot from the “Forest Plot” subtab of the “Differentially Mutated Genes” tab. 58

Figure 27. A screenshot from the “Uploaded Cohort OncoCLUST results” subtab of the “Cancer Driver Genes Based on Positional Clustering” tab. 59

Figure 28. A screenshot from the “TCGA Cohort OncoCLUST results” subtab of the “Cancer Driver Genes Based on Positional Clustering” tab. 59

Figure 29. A screenshot from the “OncoCLUST plot” subtab of the “Cancer Driver Genes Based on Positional Clustering” tab. 60

Figure 30. A screenshot from the “Co-lollipop plot” subtab of the “Cancer Driver Genes Based on Positional Clustering” tab. 60

Figure 31. A screenshot from the “Uploaded cohort somatic interaction table” subtab of the “Somatic Interactions” tab. 61

Figure 32. A screenshot from the “Uploaded cohort somatic interaction plot” subtab of the “Somatic Interactions” tab. “*” denotes p < 0.01 and “.” denotes p < 0.05. Color intensities of cells increase according to the log10(p-value), with co-occurring genes colored with shades of blue and mutually exclusive genes colored with shades of brown. 62

Figure 33. A screenshot from the “Somatic Interaction Network” subtab of the “Somatic Interactions” tab. 63

Figure 34. A screenshot from the “Somatic Interactions with Custom Gene” subtab of the “Somatic Interactions” tab. 64

Figure 35. A screenshot from the “Cophenetic Correlation vs Signature Number, TCGA cohort” subtab of the “Comparison of Mutational Signatures” tab. The vertical line denotes the optimal number of signatures. 66

Figure 36. A screenshot from the “Identified signature and their cosine similarity to SBS signatures, TCGA cohort” subtab of the “Comparison of Mutational Signatures” tab. 67

Figure 37. A screenshot from the “Cosine similarity between TCGA and Uploaded cohort signatures” subtab of the “Comparison of Mutational Signatures” tab. 68

Figure 38. A screenshot from the “Summary Table” subtab of the “Comparison of Pathways” tab. 68

Figure 39. A screenshot from the “Summary Plot” subtab of the “Comparison of Pathways” tab. 69

Figure 40. A screenshot from the “Fisher’s exact test table” subtab of the “Comparison of Pathways” tab. 69

Figure 41. A screenshot from the “Fisher’s exact test table” subtab of the “Comparison of Pathways” tab. 70

Figure 42. A screenshot from the “Pathways” subtab of the “Comparison of Pathways” tab. 70

Figure 43. Oncoplot of the IGR cohort, summarizing the mutational profile. Samples are grouped by the HR and HER+ status. Top 25 highly mutated genes are shown. 73

Figure 44. Oncoplot of the TCGA cohort, summarizing the mutational profile. Samples are grouped by the PAM50 subtype. Top 25 highly mutated genes are shown. 74

Figure 45. Comparison of TMB between the IGR cohort and the TCGA cohort. P-values are calculated by the Student's t-test. The upper and lower edges of the boxes represent 75th and 25th quantiles, respectively, meanwhile black rod represents the median. Whiskers extend until 1.5* interquartile range (IQR) if there are observations extremum to those points, and till the minimum and maximum values otherwise. 76

Figure 46. Cophenetic correlation analysis for the IGR dataset. The x-axis represents the number of signatures, y-axis represents the cophenetic correlation coefficient. 77

Figure 47. Cophenetic correlation analysis for the IGR dataset. the x-axis represents the number of signatures, y-axis represents the cophenetic correlation coefficient. The vertical line represents the chosen number of signatures to fit. 78

Figure 48. Signatures identified in the IGR cohort. The contribution of each SBS context to each signature is represented by the height of the barplots, which add up to 1. The title of each mutational signature includes the code of the best-matching COSMIC SBS signature, its name, and its cosine similarity to the SBS signature. 79

Figure 49. Signatures identified in the TCGA cohort. Contribution of each SBS context to each signature is represented by the height of the barplots, which add up to 1. Title of each mutational signature includes the code of the best-matching COSMIC SBS signature, its name, and its similarity to the signature in cosine similarity. 81

Figure 50. Cosine similarity matrix of the signatures identified in IGR and TCGA datasets. Matrix is clustered using Euclidian distance and average linkage. 82

Figure 51. Signature contribution to each patient in the IGR dataset, represented as stacked barplots. 83

Figure 52. Distribution of contributions of each signature in the IGR dataset. Each dot represents a patient. One-way ANOVA results are also included in the graph. The upper and lower edges of the boxes represent 75th and 25th quantiles, respectively, while black rod represents the median. Whiskers extend until 1.5* IQR if there are observations extremum to those points, and till the minimum and maximum values otherwise. 84

Figure 53. Linear Regression analysis of the signature contributions with TMB. The blue line represents the best fit line, and the grey area represents the confidence interval while each dot represents a patient. 86

Figure 54. Linear Regression analysis of the signature contributions with TMB, at TMB cutoff of 250. Blue line represents the best fit line, and the grey area represents the confidence interval meanwhile each dot represents a patient. 87

Figure 55. Linear Regression analysis of the signature contributions with TMB, at TMB cutoff of 150. Blue line represents the best fit line, and the grey area represents the confidence interval meanwhile each dot represents a patient. 88

Figure 56. Distribution of contributions of each signature in the IGR dataset, within each HR/HER subtype. Each dot represents a patient. One-way ANOVA results are also included in the graph. The upper and lower edges of the boxes represent 75th and 25th quantiles, respectively, while the black rod represents the median. Whiskers extend until 1.5* IQR if there are observations extremum to those points, and till the minimum and maximum values otherwise. 90

Figure 57. Distribution of contributions of each signature in the IGR dataset compared between HR+ and TNBC samples. Each dot represents a patient. Student's t-test results are also included in the graph. The upper and lower edges of the boxes represent 75th and 25th quantiles, respectively, while a black rod represents the median. Whiskers extend until 1.5* IQR if there

are observations extremum to those points, and till the minimum and maximum values otherwise. 91

Figure 58. Distribution of contributions of each signature in the IGR dataset compared between HER2+ and TNBC samples. Each dot represents a patient. Student's t-test results are also included in the graph. The upper and lower edges of the boxes represent 75th and 25th quantiles, respectively, while a black rod represents the median. Whiskers extend until 1.5* IQR if there are observations extremum to those points, and till the minimum and maximum values otherwise. 93

Figure 59. Distribution of contributions of each signature in the IGR dataset compared between HER2+ and HR+ samples. Each dot represents a patient. Student's t-test results are also included in the graph. The upper and lower edges of the boxes represent 75th and 25th quantiles, respectively, while a black rod represents the median. Whiskers extend until 1.5* IQR if there are observations extremum to those points, and till the minimum and maximum values otherwise. 94

Figure 60. Distribution of contributions of each signature in the IGR dataset compared between Liver and Lymph node biopsies. Each dot represents a patient. Student's t-test results are also included in the graph. The upper and lower edges of the boxes represent 75th and 25th quantiles, respectively, while a black rod represents the median. Whiskers extend until 1.5* IQR if there are observations extremum to those points, and till the minimum and maximum values otherwise. 95

Figure 61. Distribution of contributions of each signature in the IGR dataset compared between Liver and Skin biopsies. Each dot represents a patient. Student's t-test results are also included in the graph. The upper and lower edges of the boxes represent 75th and 25th quantiles, respectively, while a black rod represents the median. Whiskers extend until 1.5* IQR if there are observations extremum to those points, and till the minimum and maximum values otherwise. 96

Figure 62. Distribution of contributions of each signature in the IGR dataset compared between Lymph node and Skin biopsies. Each dot represents a patient. Student's t-test results are also included in the graph. The upper and lower edges of the boxes represent 75th and 25th quantiles, respectively, while a black rod represents the median. Whiskers extends until 1.5* IQR if there are observations extremum to those points, and till the minimum and maximum values otherwise. 97

Figure 63. Forest plot of the genes that are differentially mutated between IGR and TCGA cohorts. The dot represents the OR, and if the confidence interval is out of the interval shown an arrow is drawn. The number of the mutated samples in each dataset along with the respective OR is shown. * p< 0.05, **p<0.01, *** p<0.001 99

Figure 64. Co-lollipop plot of the ESR1 gene. The upper part of the plot represents the mutated residues in the IGR cohort, meanwhile the lower part represents the mutated residues in the TCGA cohort. Height of the lollipop corresponds to the number of events observed in each cohort 102

Figure 65. Co-lollipop plot of the GJA8 gene. The upper part of the plot represents the mutated residues in the IGR cohort, meanwhile the lower part represents the mutated residues in the TCGA cohort. Height of the lollipop corresponds to the number of events observed in each cohort 103

Figure 66. Clinical enrichment results for HR/HER subtypes. The upper part represents the fraction of samples with the mutation in the indicated subtype for the shown gene, whereas the lower part represents the fraction of samples with the mutation in the rest of the samples for the shown gene. Genes with FDR <0.06 were selected. 105

Figure 67. Clinical enrichment results for metastasis sites. The upper part represents the fraction of samples with the mutation in the indicated metastasis site for the shown gene, whereas the lower part represents the fraction of samples with the mutation in the rest of the samples for the

shown gene. Genes with positive enrichment with lower OR > 1 or negative enrichment with higher OR < 1 are shown. 107

Figure 68. Somatic interactions in the IGR dataset. Cells are colored according to the -log10(p-value) value of the interaction. "*" indicates significance at p-value < 0.01, "." indicates significance at p-value < 0.05. 110

Figure 69. Somatic interactions in the TCGA dataset. Cells are colored according to the -log10(p-value) value of the interaction. "*" indicates significance at p-value < 0.01, "." indicates significance at p-value < 0.05. 111

Figure 70. Forest plot of the genes differentially mutated pathway analysis between IGR and TCGA cohorts. The dot represents the OR, and if the confidence interval is out of the interval shown, an arrow is drawn. Number of the mutated samples in each dataset, along with the respective OR, is shown. * p < 0.05, ** p < 0.01, *** p < 0.001, NS p ≥ 0.05. 113

Figure 71. Comparison of the DNA damage related signatures between cohorts. Each dot represents a patient. Student's t-test results are also included in the graph. The upper and lower edges of the boxes represent 75th and 25th quantiles, respectively, while a black rod represents the median. Whiskers extend until 1.5* IQR if there are observations extremum to those points, and till the minimum and maximum values otherwise.....115

Figure 72. Oncoplot of the NRF2 mutated samples of the IGR dataset.....117

Figure 73. Comparison of the contribution of Signature 3 to NRF2 pathway wild-type and mutant samples. Each dot represents a patient. Student's t-test results are also included in the graph. The upper and lower edges of the boxes represent 75th and 25th quantiles, respectively, while a black rod represents the median. Whiskers extend until 1.5* IQR if there are observations extremum to those points, and till the minimum and maximum values otherwise. 117

Figure 74. PPI network of significantly co-occurring and positively correlated genes. Each node represents a gene, and each edge represents PPI. Red nodes denote regulation of protein-

containing complex assembly, blue nodes denote organelle localization, and green nodes denote posttranscriptional regulation of gene expression gene ontology (GO) terms [146]. Yellow nodes represent the cell cycle Reactome pathway [147]. 121

Figure 75. PPI network of significantly co-occurring and positively correlated genes. Each node represents a gene, and each edge represents PPI. Red nodes denote regulation of signaling, blue nodes denote regulation of cell communication, and green nodes denote regulation of signal transduction gene ontology (GO) terms [146]. 123

Figure 76. Hierarchical clustering of the differentially expressed genes ($p < 0.05$) in at at least one of the treatment groups, namely siA5, siP53 and siA5+siP53. Somaticaly interacting genes of the *CHRNA5* and *TP53*, as well as the components of the DREAM pathway and the LIN37 independent targets of TP53 were added as annotation. 125

LIST OF TABLES

Table 1. List of software packages used.....28

Table 2. List of reference files used.....31

Table 3. Summary of the hormone receptor and HER2 status of the IGR cohort.....71

Table 4. Summary of the biopsy locations of the IGR cohort.....71

Table 5. Driver genes identified in each cohort by the OncodriveCLUST algorithm [50].....100

Table 6. Summary of the VAF, altered residues, reference and variant allele read numbers of GJA8 mutations in normal and tumor samples of the IGR cohort.....101

Table 7. Detailed Fisher's exact test results for the HR/HER subtype clinical enrichment analysis.....106

Table 8. Detailed Fisher's exact test results for metastasis site clinical enrichment analysis.....108

Table 9. List of the somatic interactions unique to the IGR cohort. Pairs that are composed of the highly mutated genes common to both IGR and TCGA cohorts are marked with an asterisk.....111

Table 10. Detailed Fisher's exact test results of the pathway comparison.....114

Table 11. Comparison of features of BRCA:Cohort-vs-TCGA app with other apps.....125

LIST OF EQUATIONS

Equation 1 4

Equation 2 4

Equation 3 4

Equation 4 5

Equation 5 5

Equation 6 5

Equation 7 6

Equation 8 6

Equation 9 9

Equation 10 10

Equation 11 10

Equation 12 11

Equation 13 13

ABBREVIATIONS

ANOVA	: analysis of variance
BRCA	: breast invasive carcinoma
CDK	: cyclin dependent kinase
CNA	: copy number alteration
COSMIC	: catalogue of somatic mutations in cancer
DDR	: DNA damage repair
DSB	: double-strand break
EMT	: epithelial mesenchymal transition
ER	: estrogen receptor
FDR	: false discovery rate
GO	: gene ontology
GSEA	: geneset enrichment analysis
HER2	: human epidermal growth factor receptor 2
HR	: hormone receptor
IGR	: Gustave Roussy Institute
IHC	: immunohistochemistry
IQR	: interquartile range
MAF	: mutation annotation format
MC3	: multi-center mutation calling in multiple cancers
MET	: mesenchymal epithelial transition
mTOR	: mammalian target of rapamycin
mTORC1	: mammalian target of rapamycin complex 1
mTORC1	: mammalian target of rapamycin complex 2
NER	: nucleotide excision repair
NGS	: next-generation sequencing
NMF	: non-negative matrix factorization
OR	: odds ratio
PI3K	: PI-3-Kinase
PIP_2	: phosphatidylinositol (3,4)-bisphosphate
PIP_3	: phosphatidylinositol (3,4,5)-trisphosphate
PON	: panel of normals
PPI	: protein-protein interaction
PR	: progesterone receptor
ROS	: reactive oxygen species
RTK	: receptor tyrosine kinase
SBS	: single base substitution
TCGA	: the cancer genome atlas
TMB	: tumor mutation border
TNBC	: triple-negative breast carcinoma

TSG	: tumor suppressor gene
VAF	: variant allele frequencies
VCF	: variant call format
VEGF	: Vascular endothelial growth factor
VEP	: variant effect predictor
VUS	: variants of unknown significance

1 Chapter 1 INTRODUCTION

1.1 Cancer and Its Hallmarks

Prominent cancer researchers Douglas Hanahan and Robert A. Weinberg introduced the concept of cancer hallmarks to describe and delineate the necessary steps in cancer progression [1]. The first step is to acquire the ability to proliferate without the dependence on growth factors. Activating mutations in the gene coding the catalytic subunit of the phosphatidylinositol 3-kinase (PI3K), for example, constitutes this type of an event [2]. In order to survive intrinsic and extrinsic stressors such as oncogene induced senescence/apoptosis and genomic alterations, cancer cells should also acquire mutations that deactivate cell death pathways [1]. Perhaps the most widespread and well-known example of this type of mutation is the missense and the frameshift mutations in the *TP53* [3]. A cell that can divide without stimulation and can evade cell death and division suppression mechanisms still faces the existential threat of shortening telomeres, a process inherent in the lagging strand replication [1], [4]–[6]. Tumor cells solve this problem either by re-expressing the telomerase protein or utilizing the Alternative Lengthening of Telomeres Pathway [1], [7]–[9]. As the tumor grows, so does the need for vasculature to sustain the exchange of metabolites and fuel. Thus, tuning the vascular endothelial growth factor (VEGF) pathway components to allow neovascularization is necessary and another hallmark of cancer progression. Yet another property that defines invasive/metastatic tumors is their ability to disseminate and colonize nearby or distant tissues. This process which involves transitions back and

forth from adhesive, or epithelial phenotype to mesenchymal phenotype is called epithelial to mesenchymal transition (EMT) and MET, respectively [1].

1.2 Breast Cancer

Breast cancer is the most common malignancy among women, which is most frequently of either ductal or lobular origin [10].

1.2.1 Epidemiology of Breast Cancer

GLOBOCAN estimates that in 2020, the most recent published figures, around 9.2 million women were diagnosed with cancer. Breast cancer accounts for 24.5 % of the newly diagnosed cases, which makes this malignancy the most widespread one among women. Breast cancer is also the lead cause of mortality among women, with 14.5 % of 4.4 million female cancer related deaths attributed to it. In Turkey, the incidence of breast cancer among women is 24175, and the mortality is 7161 [11].

1.2.2 Intrinsic Subtypes of Breast Cancer

The idea of intrinsic molecular subtyping was first put forward in the paper of Perou *et al.*, in which gene expression profiles of breast cancer cell lines, and patients, some of which were sampled and profiled several times, including from metastatic sites, were quantified. Interestingly, the final list of 496 genes chosen can recapitulate the patient of origin and cluster samples together regardless of them being a primary or metastatic sample or before and after chemotherapy treatment. More interestingly, however, this classification using the “intrinsic gene list” mainly was faithful, and concordant with the immunohistochemical (IHC) classifications used at the time. The identified subtypes were Luminal A, Luminal B, Basal, Normal-like, and Her2 over-expressed [12].

Later, a paper further narrowed the intrinsic gene list down to 50 genes based on the Prediction Analysis of Microarray (PAM50) [13], [14]. This paper further showcased the power of PAM50 subtyping by showing that it increased both the prognostic and predictive value of models that incorporate other clinicopathological variables as well [14].

1.2.3 Molecular Subtypes of Breast Cancer

Breast cancers can also be classified according to the hormone receptors they express, namely estrogen receptor (ER) and progesterone receptor (PR), Human Epidermal Growth Factor Receptor 2 (HER2) expression, and Ki67 staining via IHC [15]. Molecular subtypes are essential since they can be used to decide the treatment strategy by clinicians.

1.3 Mutation Detection and Filtration

1.3.1 Mutation Calling

Facilitated by the ever-decreasing price of the next-generation sequencing (NGS) the number of large scale genome and transcriptome projects is now increasing [16]. NGS platforms have inherent error rates for calling bases at the specific mutation hence require statistical methods to call the variants confidently [17]. Additionally, since most cancer cases are sporadic, and not caused by inherited traits, detecting the somatic mutations in the tumor tissue is another challenge that needs to be addressed [18]. A statistical software used in this thesis, Mutect2, addresses both of these problems by using Bayesian classifiers [19], [20].

Let's suppose that the reference allele is $g \in \{A, T, G, C\}$ and the true variant allele is $v \in \{A, T, G, C\}$. Let's suppose that the machine used to sequence the library has a base

calling error rate of e_i which is given by the Phred scaled error rate. This error rate is invariant to the base being called. Finally, the true variant allele fraction can be represented by f. If a specific locus is sequenced at the depth of d the likelihood of the observed base b_i can be written as

$$L(M) = \prod_{i=1}^{d} P(b_i|g, v, e_i, f)$$

Equation 1

$$P(b_i|g, v, e, f) = \begin{cases} f\frac{e_i}{3} + (1-f)(1-e_i) \text{ if } b_i = g \\ f(1-e_i) + (1-f)\frac{e_i}{3} \text{ if } b_i = v \\ \frac{e_i}{3} \text{ otherwise} \end{cases}$$

Equation 2

Given that now we have the equations for the likelihood and the probability of observing a base given the variant allele fraction f , one can calculate the log odds (or the log posterior odds in the proper Bayesian terms) of observing the base in $f = 0$ and in $f = \hat{f}$, the observed variant allele fraction. Then one can define this log odds as

$$LOD(v, f) = log\ (\frac{L(M_{f=\hat{f}})P(v, f)}{L(M_{f=0})(1 - P(v, f))})$$

Equation 3

In the **Equation 3** the $P(v, f)$ serves as the prior and is the joint distribution of the variant frequency and the chance of getting mutated. The tool assumes that these events are independent and hence can be re-written $P(v, f) = P(v)P(f)$. Furthermore, $P(f)$ is assumed to be uniformly distributed, hence $P(f) = 1$ and $P(v)$ is set to the one-third of the mutation rate. Mutect2 wants this posterior log odds to be at least two.

$$LOD(v, f) \geq log\ (\varphi)$$

Equation 4

$$LOD(v, f) = log\ (\frac{L(M_{f=\hat{f}})P(v, f)}{L(M_{f=0})(1 - P(v, f))}) \geq log\ (\varphi)$$

Equation 5

$$LOD(v, f) = \log\left(\frac{L(M_{f=\hat{f}})}{L(M_{f=0})}\right) \geq \log(\varphi) - \log\left(\frac{P(v, f)}{1 - P(v, f)}\right) = \alpha$$

Equation 6

Given that we want $\varphi \geq 2$, by plugging in the value for $P(v, f)$ we can estimate that the $\alpha = 6.3$. Simply put, one needs to ensure that the log-likelihood ratio is at least 6.3 to ensure that the called variant is at least twice as likely to originate from a real mutation compared to the case where the given site is wild-type [19], [20].

1.3.2 Detection of Somatic Variants

To classify the variants as somatic Mutect2 uses a similar log odds introduced in the **Equation 3-6**. However, in this case, one desires to make sure that the mutation

observed is not likely to originate from a germline mutation. Therefore, in the normal sample, the log odds is defined as

$$LOD = log\left(\frac{L(M_{f=0})P(v,f)}{L(M_{f=0.5})P(I)}\right) \geq log\ (\omega)$$

Equation 7

With ω desired to be at least 10. $L(M_{f=0.5})$ represents the likelihood under the assumption that the observed variant is heterozygous. $P(I)$ represents the prior for the probability of germline mutation. After modifying equation **Equation 7** we get

$$LOD = \log\left(\frac{L\left(M_{f=0}\right)}{L\left(M_{f=0.5}\right)}\right) \geq \log(\omega) - \log\left(\frac{P(v,f)}{P(I)}\right) = \beta$$

Equation 8

Since authors calculate different $P(I)$ depending on locus being present in dbSNP or not using the average variant number in the human genome and the fraction of these variants belonging to the dbSNP sites. In the dbSNP sites, β is estimated to be 2.2 for the non-dbSNP sites and 5.5 for the dbSNP sites [21]. In the end, researcher is sure that the detected variant is at least 10 times more likely to be a somatic variant [19], [20].

1.3.3 Filtration of the Variants

Mutect2 uses the concept of the panel of normal (PON) to detect and filter out the possible germline or the recurrent sequencing artefacts. Briefly, Mutect2 tumor variant detection is run on PON composed of the normal samples from unrelated patients to the

patients under study. Variants that occur in at least two of the normal are then flagged and excluded from the downstream analyses [19], [20].

1.3.4 Mutation Annotation Format (MAF).

The mutation annotation format is developed to streamline and harmonize the annotation of somatic mutations. MAF file is generated via vcf2maf [22] which runs the Ensembl's variant effect predictor (VEP) to convert a variant call format (VCF) file to a MAF file with preset options [23], [24]. The required fields of a MAF file are Hugo_Symbol, Chromosome, Start_Position, End_Position, Reference_Allele, Tumor_Seq_Allele2, Variant_Classification, Variant_Type, and Tumor_Sample_Barcode [25]. Additional annotations can be added to the file with the proper VEP options.

1.3.5 Multi-Center Mutation Calling in Multiple Cancers (MC3) pipeline

Multi-Center Mutation Calling in Multiple Cancers (MC3) pipeline is the pipeline used to detect the somatic mutations in the PanCancer Atlas and this list was used in all the analyses presented in the series [26]. MC3 pipeline is characterized by the custom software used for the variant filtration and the multiple software used for calling the somatic variants. It can be argued that the MC3 pipeline is the most robust pipeline generated for the detection of the somatic variants so far, yet has a high computational burden and long runtime owning to the use of multiple software [26].

Overall, seven tools were used to detect the variants, namely Mutect [19], Varscan2 [27], Indelocator [28], Pindel [29], SomaticSniper [30], RADIA [31] and MuSE [32]. Out of these tools, Indelocator and Pindel were used to detect insertions and deletions

(indels), Varscan2 was used to detect both indels and single nucleotide variations (SNVs), and the rest was used to detect SNVs. Variants called at least by the two of the callers were allowed in the public access files. Furthermore, specific filters for the alternate allele count at the ExAC [33] datasets, different read depth requirements for the dbSNP and non-dbSNP sites [21], PONs filter, capture kit coverage filter, and strand bias filters were also applied.

In addition to the variant level filters, several filters were also applied at the sample level. These filters include contamination filters [34] and 8-Oxoguanine lesion filter [35], which may happen during the library preparation step. It should be noted that the highest number of variants are filtered by the non-exonic and PON filters [26].

1.4 Mutational Signatures

1.4.1 Introduction to The Concept of Mutational Signatures

The idea of the mutational signature was developed with the hope that one can identify or decompose the somatic mutation profiles into certain fingerprints which in turn are representative of the underlying mutational processes. The early studies on the carcinogens implicated that these carcinogens might leave a unique mutational fingerprint. For example, a research team led by Mehmet Ozturk found in 1991 that patients with a high risk of exposure to aflatoxin were more likely to have a recurrent mutation at the 294th codon of the *TP53* [36]. Around the same time, research conducted on a leukemia cell line that can be induced to produce oxygen radicals identified that these radicals caused C-C dimers, and mutations also occurred at the recurrent sites known at the time [37]. Moreover, the first mutational profiling of a lung cancer cell line revealed features related to the expression dependent DNA damage repair, and a

novel repair mechanism that protected both transcribed and untranscribed strands of highly expressed genes[38].

The modern approaches for the detection of the mutational signatures rely on the mutational context. Single-base substitution based signature analysis considers bases upstream and downstream of 6 substitution mutations which in total makes up the 96 different types of mutations with their respective context. The first ever methods developed for this purpose utilized non-negative matrix factorization (NMF) to extract the signatures [39]. The most recent set of signatures identified and presented in the catalogue of somatic mutations in cancer (COSMIC) identifies contexts relevant to the deficiencies in the DNA damage repair pathways, exposure to carcinogens, chemotherapy drugs as well as the signatures which are likely to represent sequencing errors [40], [41].

1.4.2 Non-Negative Matrix Factorization

NMF technique is widely applied in image processing. A paper that that describes its application and algorithm was published in 1999 [42]. The goal of the NMF technique is to estimate a non-negative matrix by the matrix multiplication of two matrices that themselves are non-negative.

$$M \sim VS$$

Equation 9

In the case of the mutational signature extraction, columns of the matrix M represent each sample while the rows are 96 mutational contexts. The matrix V is composed of n columns each representing a mutational signature and 96 rows, overall representing the

contribution of each mutational context to the derived signature. Columns of the matrix S on the other hand represent each sample and the rows correspond to the signatures, representing the contribution of each signature in each patient [42].

The matrices V and S are initialized randomly, and their values are updated as follows

$$V_{nc} := V_{nc} \frac{\sum_p \frac{S_{cp} M_{np}}{(VS)_{np}}}{\sum_v S_{cv}}$$

Equation 10

$$S_{cp} := S_{cp} \frac{\sum_n \frac{V_{nc} M_{np}}{(VS)_{np}}}{\sum_k V_{kc}}$$

Equation 11

This algorithm is first implemented in the reference [43] and is the one used in this thesis.

1.4.3 Selection of Optimal Number of Signatures

To determine the optimal number of signatures as well as decrease the dependence on the initialization values of the matrices, NMF can be run several times. The dominant signature can represent the cluster each sample belongs to and these values can be used to determine the consensus adjacency matrix. The Pearson correlation between the consensus adjacency matrix and the cluster assignments when a clustering algorithm is run can give the stability of the signature. This correlation analysis is called the cophenetic correlation, and was used by the reference [43] to assess the signature stability when they tried to decompose a high-dimensional microarray dataset to a few “metagenes” via NMF.

1.4.4 Comparing The Similarity of Signatures

Similarity of the signatures can be compared by the cosine similarity. Cosine similarity simply calculates the cosine of the angle made between the two vectors

$$\cos\theta = \frac{A \cdot B}{||A||\,||B||} = \frac{\sum_i A_i B_i}{\sqrt{\sum_i A_i^2}\sqrt{\sum_i B_i^2}}$$

Equation 12

Intuitively, this similarity metric only measures the direction of vectors, rather than the magnitude.

1.5 Cancer Driver Genes

1.5.1 Introduction

Even though large scale studies have identified thousands of somatic mutations; most of these are variants of unknown significance (VUS) and passenger mutations, which by definition do not confer any selective advantage to the cancer cells and simply happen randomly because of the mutagenic processes [44]. Driver mutations on the other hand provide the mutant cells with a significant evolutionary advantage and hence are under positive selection [44]. American College of Medical Genetics and Genomics published guidelines in the annotation of the variants with regard to their possible effect [45]. This guideline however relies strictly on the annotations with known databases and does not confer the researchers with the ability to identify new driver genes/variants which might be present in their data. The first report regarding the presence of transforming genes or oncogenes within the genome of the animals came with the

observation that cancer causing genes of the viruses can hybridize with some genes from the animals [46]. Later on, it was identified that deactivating mutations in some genes can cause cancer and these genes were “named tumor suppressor” genes [47]. However, it is not feasible to test all of the mutations for their transformation activity. Consequently, the development of computational methods for the identification of the candidate driver genes was necessary. Several approaches have been developed for the detection of the cancer driver genes and hotspot mutations. Broadly, these approaches can be classified into mutational rate based and proximity based methods [48]–[53]. A rather new method of detecting driver mutations is to consider their combination. Some mutations on their own can be passengers, however, their co-occurrence can provide a selective advantage. This type of driver mutations is called latent driver mutations [54], [55].

1.5.2 Mutational Rate Based Methods

The most commonly used mutational rate based method for identifying the driver mutations is MutSigCV [56]. MutSigCV builds up on the previous versions of the tool in which the mutation category specific mutation rate and patient specific mutation rate are used to give each gene a score that is compared to the background distribution to calculate a p-value for that gene [57], [58]. MutSigCV extends this paradigm by calculating a gene specific mutation rate as well as accounting for covariates such as the chromatin state of the gene, its duplication time and expression. Even though calculating the background mutation rate for long genes has enough power, doing so for the shorter genes is less robust. MutSigCV increases this power by averaging the mutation rate for the genes which share the same neighborhood in the covariate space. As can be seen from the supplementary figure 1 of the reference, [56] this approach can

both increase the power of detection for the genes with low background mutation rate as well as decrease the false positive rate for the genes with intrinsically high mutation rate.

1.5.3 Positional Clustering Based Methods

Philosophy behind this approach is that some genes tend to have clusters of mutations in neighboring regions. The MutSig suit has a module dedicated to this purpose, MutSigCL [49], [59]. MutSigCL defines a hotspot region as a 3 base-pair region that contains two or more mutations and at least 2% of the mutations in that gene. Then a score, S_{CL}, is calculated as the proportion of the mutations in that gene that fall onto the hotspots. To calculate a p-value for the gene under investigation, mutations on the gene are permuted and the S_{CL} score is calculated for each permutation. Afterward, an empirical p-value is calculated for the gene.

OncodriveCLUST is another method based on the positional clustering of the mutations [50]. This approach takes into account the length and mutation rate of the gene to identify seeds of the clusters, which are defined as positions likely to be observed in less than 1 % of the observations given the binomial null distribution. These cluster seeds are then extended to define the cluster regions. Afterward, a cluster score for each cluster is calculated as

$$S_c = \sum_i \frac{f_i}{(\sqrt{2})^{d_i}}$$

Equation 13

Where i represents each position within the cluster, f_i represents the proportion of mutation falling to the position i in that protein and d_i represents the distance from that mutation to the seed of the cluster. Afterward, a score for each gene is computed by summing up the scores of the clusters within that gene. This score is compared to the background generated by the clustering scores of the non-coding mutations. A simple Z-score based approach is then used to compute the p-value [50].

The methods discussed so far defines the proximity in the linear amino acid sequence. A far more accurate approach would be to quantify the clustering of the mutations in the 3 dimensional (3D) conformation of the folded protein [51]. Again, hypothesis testing in this method is carried out empirically by permuting the mutations across residues.

1.5.4 Latent Driver Gene Approach

This approach states that two or more mutations which are passengers or VUS on their own can act as driver mutations when they are present in the same patient. Indeed, different co-occurrent mutations of the *PIK3CA* gene are shown to manifest different phenotypes, including increased sensitivity to inhibitors compared to their single mutant counterparts [60]. A recent study tried to identify these latent driver mutations on pan-cancer datasets. Researchers identified 113 novel double mutations classified as latent drivers, and through analysis of the drug response data from cell lines and patient derived xenografts they show that the presence of these latent drivers can influence drug response [55].

1.5.5 Tumor Suppressor Gene and Oncogene Identification

One of the problems encountered in cohort studies is to assign genes as tumor suppressor genes (TSGs) or oncogenes without conducting functional experiments. A simple yet effective method is called “20/20” [53], [61]. In this method authors suggest that if at least 20% or more of the mutations with the coding consequence happen at the hotspot regions are missense mutations, then the gene is classified as oncogenes, likewise, if at least 20% of mutations with the coding consequence are loss-of-function mutations, then the gene is said to be TSG. The authors support their method with the example of *IDH1*. The “20/20” method identifies this gene as an oncogene, whereas multiple conflicting reports exist in the literature, which at the end, concluded the *IDH1* to be a TSG [53], [61].

1.6 Methods for Identification of Somatic Interactions and Their Visualization.

Somatic interaction analyses can reveal genes that act synergistically or alternatively, they can also reveal the genes with redundant functions or synthetic lethal interactions. The simples approach for the identification of such interactions is Fisher’s exact test [62]. This test however, can have a high number of false positives if the cohort contains a considerable number of hypermutated samples, because they can lead to spurious co-occurrences [62]. To that end, WeSME tool suggested a weighted sampling scheme to overcome this problem, and identified novel mutually exclusive pairs of mutations [63]. A tool called Dendrix takes an algorithmic approach to the problem. This tool tries to maximize the number of genes, *u*, for which most of the patients have at least one mutation and for which most of the patients harbor mutations in only one of the genes in *u* [64]. CoMEt technique can also identify mutual exclusivity relationships among multiple genes, but unlike Dendrix it calculates a p-value [65]. Some methods are able

to incorporate *a priori* pathway knowledge. For example, a study identified the candidate mutually exclusive genes via expanding the gene list such that all members of the list share a common downstream effector without crossing a gene which is not part of the gene list. Finally, a p-value is calculated via a permutation based methodology [66].

Networks made from mutually exclusive or co-occurring genes can be informative as well. For example, a network of the somatic interactions identified by using COSMIC shows that genes cluster into separate modules [67]. Interestingly, closer genes in the network have similar expression values. Co-occurring genes have similar mutation frequencies, and they influence the functionally related pathways, the authors identify [67].

1.7 Oncogenic Pathways in Cancer

It is usually observed that mutations in the components of the same pathways are mutually exclusive because they have redundant functions, and a new mutation in the same pathway does not provide an evolutionary advantage [68]–[71]. This approach can lead to the identification of the differentially mutated pathways, with the increased power of detection owing to considering multiple genes in the same pathway. For example, in the pan-cancer analysis of the pathway interactions, it was discovered that alterations in Hippo and cell TP53 pathways, TP53 and cell cycle pathways, NRF2 and PI-3-Kinase (PI3K) pathways were co-occurring [68]. Alterations in the RAS and NRF2, NRF2 and cell cycle, and receptor tyrosine kinase pathway (RTK) and PI3K pathways, on the other hand, were mutually exclusive.

In this thesis, I define the oncogenic pathways in the same way that is defined in the PanCancer Atlas [68]. The selected pathways are the PI3K pathway, RTK/ RAS

pathway, MYC pathway, P53 pathway, and the cell cycle pathway. Below the components of each pathway are described in detail.

1.7.1 RTK/RAS signaling pathway

RTKs are transmembrane proteins that can initiate a signaling cascade dependent on the tyrosine phosphorylation upon the binding of their corresponding ligands [72]. Furthermore, RTKs can be classified into families. EGFR, a member of the ErbB family of RTKs, was shown to be amplified in around 6 % of breast cancers in a study [73]. In this subset of the patients, amplification also correlated with the expression of the same gene. Additionally, three of the EGFR amplified patients were also positive for *HER2*, yet another member of the ErbB family. *HER2* amplifications are one of the defining features of Triple-Negative Breast Carcinoma (TNBC), a subtype of breast carcinoma with a high recurrence rate and worse prognosis. VEGF family of receptors is also implicated in breast carcinogenesis. For example, *VEGFR2* expression mediated by the mutated *TP53* with the help of the SWI/SNF chromatin remodeler is shown to increase cell growth and migration [74]. Furthermore, a pathological study conducted on 98 invasive breast carcinomas identified that *VEGFR2* expression is related to lymph node metastasis, associated with poor survival outcome, and *VEGFR2* high samples express the epithelial-mesenchymal transition related markers. Some of the developed VEGFR inhibitors have also undergone clinical trials. A phase II trial conducted on metastatic breast cancer patients identified that sunitinib, a multi-targeting TK inhibitor that also targets VEGFRs, was effective on 11% of the patients. Responding patients were either TNBC or HER2 positive subtype. A detailed description of the RTK/RAS pathway components is shown in **Figure 1**, which was downloaded from PathwayMapper [75].

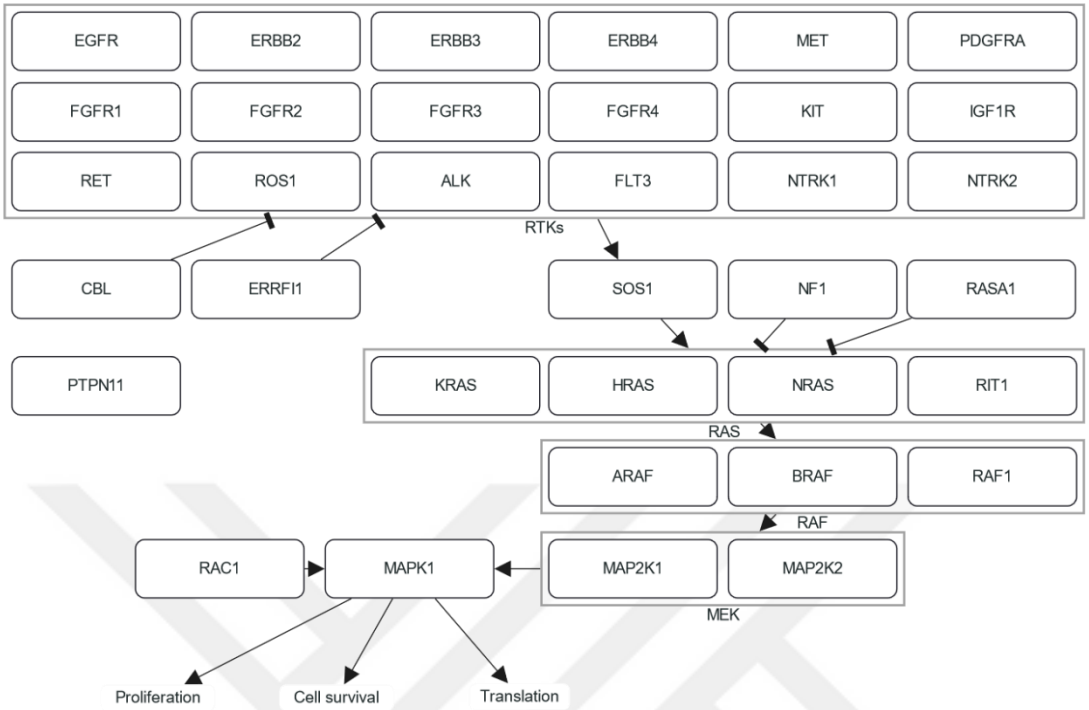


Figure 1. Components of the RTK/RAS pathway and interaction between them. Pointed arrows denote the activatory interactions, while blunted arrows denote the inhibitory interactions. The image was generated via the PathwayMapper software [75].

1.7.2 PI3K signaling pathway

PI3K catalyzes the conversion of phosphatidylinositol (3,4)-bisphosphate (PIP_2) to phosphatidylinositol (3,4,5)-trisphosphate (PIP_3) upon activation by a receptor or adaptor protein, and is first isolated from a rat liver sample [76]. The class IA of the PI3Ks is composed of the one p110 and one p85 subunits, catalytic and the regulatory subunits, respectively [77]. The signal from the PI3K is transmitted downstream via

PIP_3 to AKT [78]. AKT, in turn, phosphorylates and inhibits a complex made up of TSC1 and TSC2, a suppressor of the

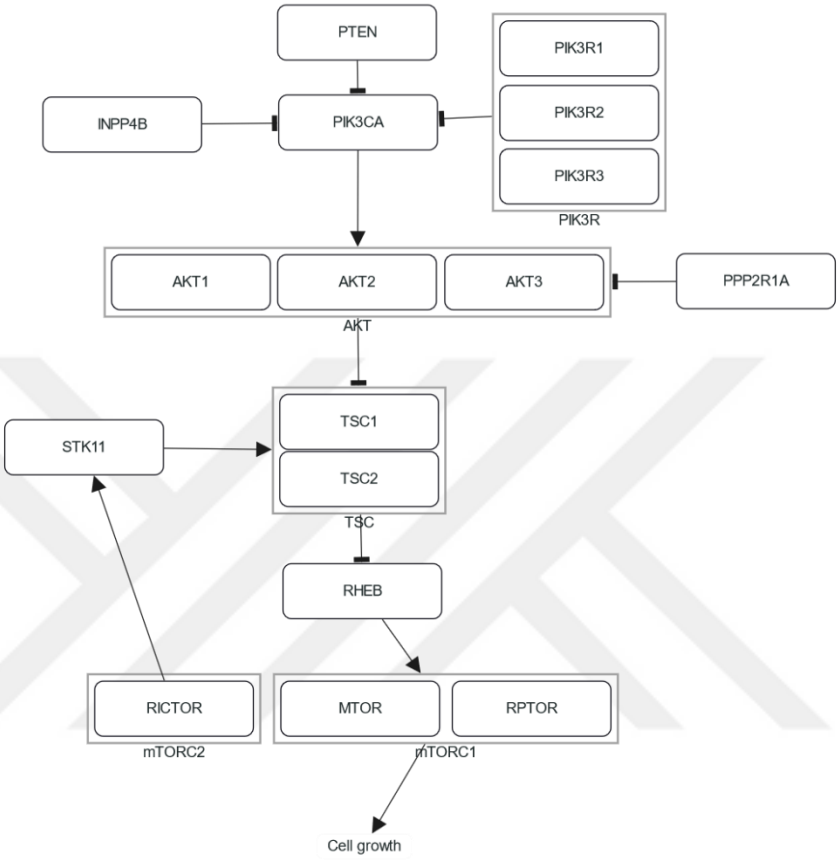


Figure 2. Components of the PI3K pathway and interactions between them. Pointed arrows denote the activatory interactions, while blunted arrows denote the inhibitory interactions. The image was generated via the PathwayMapper software [75].

mammalian target of rapamycin (mTOR) complex 1 (mTORC1). The other complex which contains the mTOR, mTORC2, also controls the activity of all 3 AKTs [79]. Widespread *PIK3CA* mutations were first identified in numerous cancer in a study published in Science [2]. Authors reported that these mutations were mainly clustered in the kinase and the helical domains. Moreover, overexpression of one of the mutants allowed them to demonstrate that these mutations are activating mutations [2]. The first

TCGA report on the breast cancer cohort also identified this gene as the second highest mutated gene, TP53 being the first. This study also reports that for Luminal A, *PIK3CA* is the gene with the highest mutation frequency [80]. A negative regulator of the PI3K pathway, *PTEN*, is also observed to be mutated in high frequencies in the families with inherited Cowden disease [81]. Patients suffering from this disease are susceptible to developing multiple malignancies, including breast cancer. Yet another gene of this pathway, *AKT1*, was found to harbor an activating mutation in the pleckstrin homology domain across multiple malignancies [81]. A detailed network of the PI3K pathway components is represented in **Figure 2** [75].

1.7.3 MYC signaling pathway

MYC family members are basic helix-loop-helix leucine zipper (bHLHZip) transcription factors and require interaction with MAX to be able to promote transcription [82]. MAX, however, can also dimerize with the other binding partners, which creates competition for MAX between MYCs and them. One of such transcription factors is MGA [83]. MAD family of proteins also competes for the binding MAX [84], [85]. The complex made of the MLX, MONOA, and MONOB controls the metabolic side of the pathway, and acts as the sensor for glucose and other nutrients [86], [87]. The most frequently altered member of the MYC pathway is the *MYC* itself, presenting itself as the copy number alteration (CNA) in many cancer types.

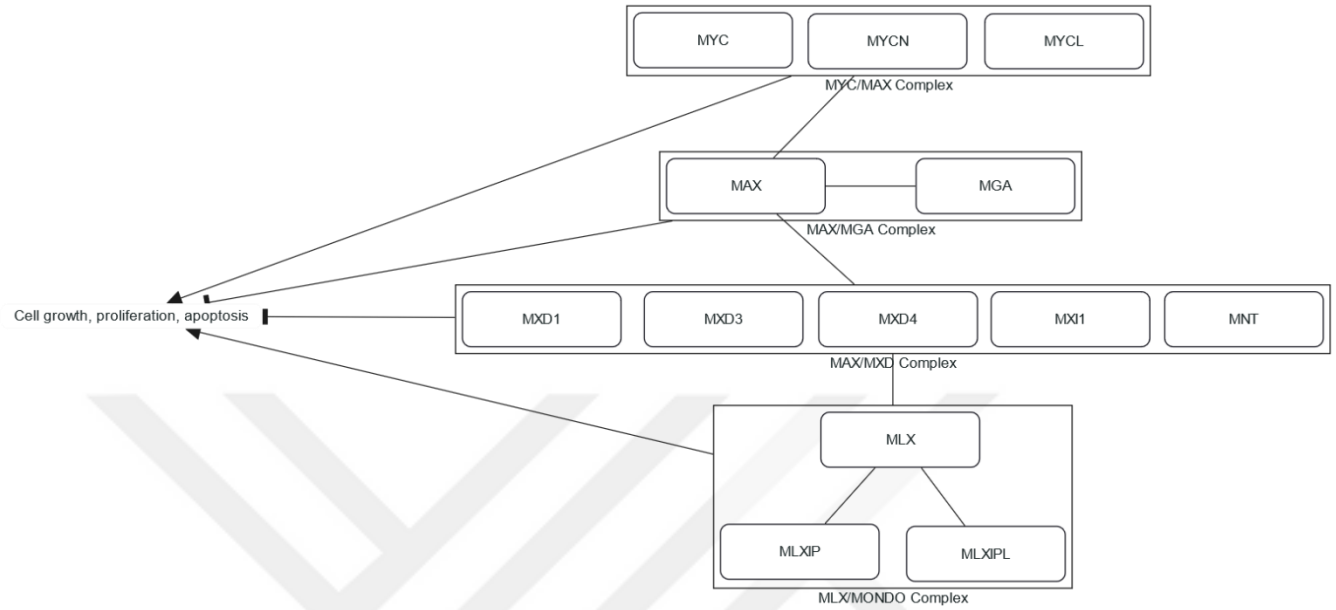


Figure 3. Components of the MYC pathway and interactions between them. Pointed arrows denote the activatory interactions, while blunted arrows denote the inhibitory interactions. The image was generated via the PathwayMapper software [75].

Amplification of *MYC* is observed in almost 14 % of the TCGA samples [88]. The MYC binding partner *MAX* is also mutated in cancer as well. Especially, small-cell lung carcinoma displays *MAX* deficiency. Given the role of *MAX*, this result is unexpected. However, it was identified that *MAX* deficiency resulted in the malfunctioning of the non-canonical polycomb repressor complex 1.6, which was essential to repress the cell-cycle genes [89]. A detailed network of the MYC pathway components is represented in **Figure 3** [75]**.**

1.7.4 TP53 Pathway

TP53 is the gene that is mutated the most in sporadic breast cancer [68]. TP53 is often regarded as the protector of the genome, responding to mitogenic stress and DNA damage. One of the regulators of the *TP53* gene is the *MDM2* gene which promotes ubiquitination and subsequent degradation of the *TP53* gene product and suppresses the TP53 pathway [90]. ATM and ATR kinases are activated in response to various DNA damage types. They phosphorylate and activate CHEK1 and CHEK2 proteins, which orchestrate the DNA-damage response pathways. ATM can directly phosphorylate TP53 at the 46th serine residue to help with the preferential activation of the proapoptotic genes [91]. Moreover, DNA-PK can

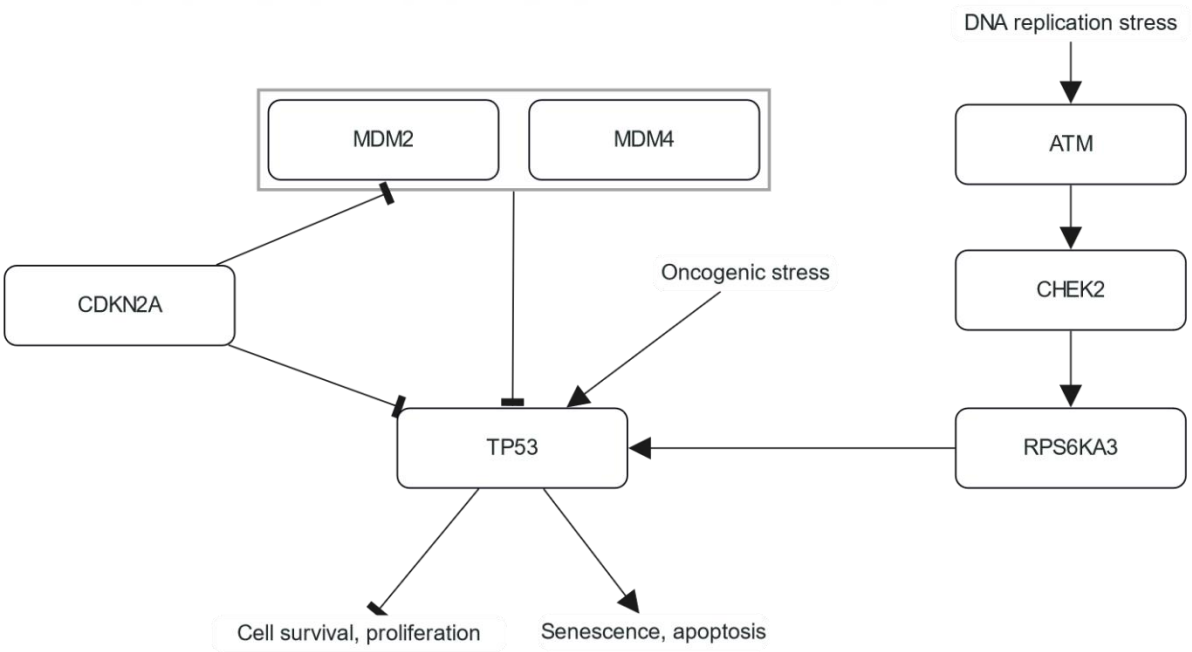


Figure 4. Components of the TP53 pathway and interactions between them. Pointed arrows denote the activatory interactions, while blunted arrows denote the inhibitory interactions. The image was generated via the PathwayMapper software [75].

can phosphorylate the TP53 protein at 15th and 37th serine residues, preventing the interaction of MDM2 and TP53, thus TP53 stabilization [92]. DNA damage can also activate the RPS6KA3 protein, which phosphorylates the TP53 at serines 18 and 16 [93]. *TP53* gene harbors both damaging mutations, such as nonsense mutations and frameshifts, as well as hotspot mutations. For example, point mutations of codons 248, 273, 175 , 245, 220, 179, 213, 282, 163, and 237 together account for 30.2 % of the *TP53* mutations in the IARC TP53 database [94]. In addition to ablating the associated TP53 pathway, the gain of function mutations of *TP53* can confer new oncogenic abilities. For example, mutant *TP53* was found to enhance the expression *MYC* and *BCL2L1* in glioblastoma cell lines and patients [95]. Additionally, mutant (R175H) TP53 and SMAD complex was shown to enhance metastasis in TGFβ dependent manner [96]. Other components of the TP53 pathway are also mutated in cancer. For instance, Germline mutations in the *ATM* gene can dispose patients towards developing early onset breast cancer [97]. Overall, *ATM* and *ATR* are mutated in the 5.5 and 3 % of the cBioPortal datasets, respectively [98], [99]. A study conducted on a cohort composed of German familial breast cancer patients suggested a modest effect in the breast cancer susceptibility for the most common CHEK2 mutations [100]. A detailed view of the TP53 pathway can be found in **Figure 4** [75].

1.7.5 Cell Cycle Pathway

Cell cycle pathway controls the progression of the cell through the cell cycle stages and ultimately decides whether the cell will divide or not. MDM2, ATM, and TP53

components of this pathway have already been discussed within **1.6.4**. The progression through the cell cycle is controlled by the cyclins and their interacting kinases, cyclin dependent kinases (CDKs). The cyclin D family of proteins is upregulated in the G1 phase of the cell cycle and interacts with CDK4 or CDK6 [101]–[103]. The activated CDKs hyperphosphorylated the RB family of proteins hence preventing them from interacting and repressing the E2F family of proteins [104], [105]. The CDK2-cyclin E complex, activated in late G1, also phosphorylates the RB family of proteins [106]. There exist a set of proteins called cyclin dependent kinase inhibitors which inhibit the function of CDKs. Two of these proteins are encoded by the *CDKN2A* gene and inhibit CDK4 and CDK6 interaction with their respective cyclins [107]. Another cyclin inhibitor, CDKN1A, prevents the interaction between CDK2 and Cyclin E [108].

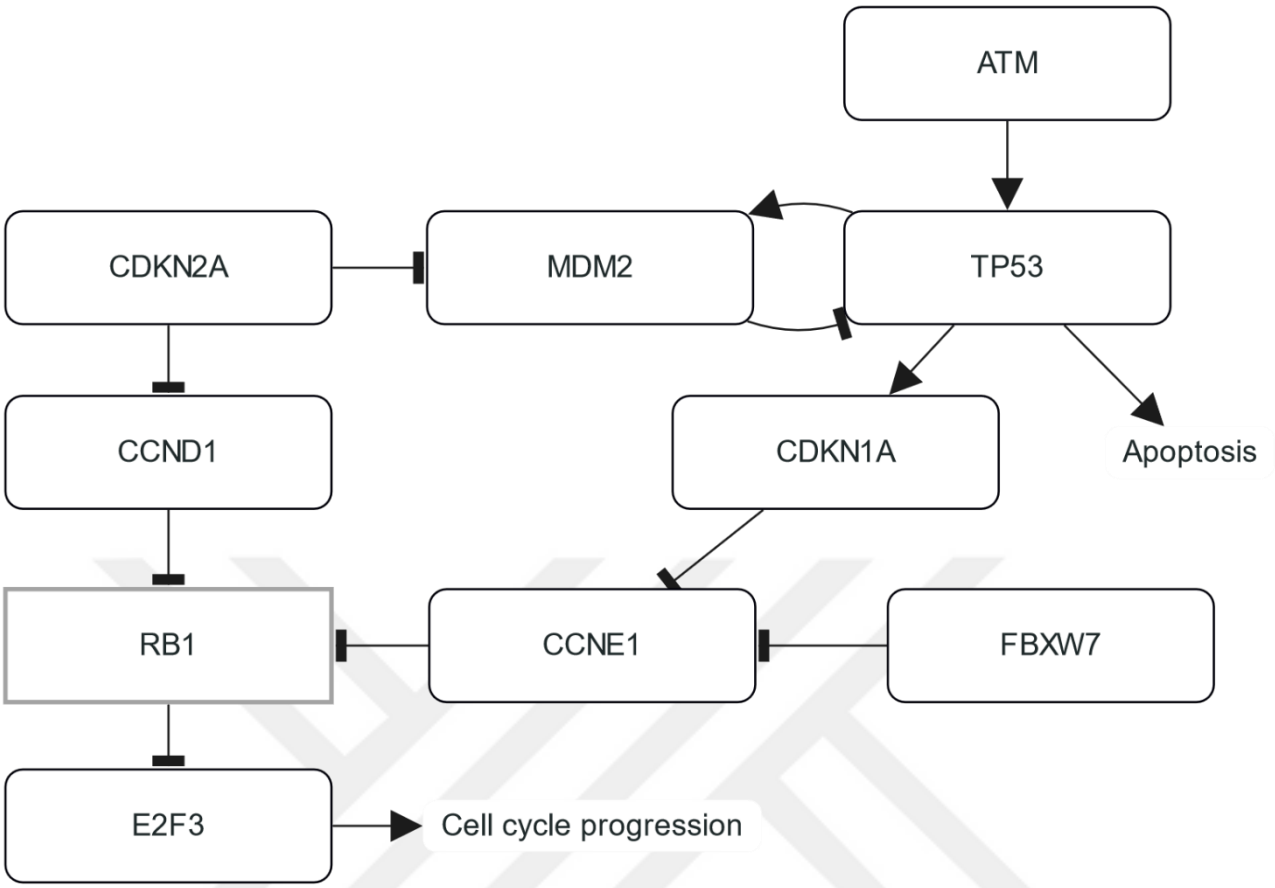


Figure 5. Components of the cell cycle pathway and interactions between them. Pointed arrows denote the activatory interactions, while blunted arrows denote the inhibitory interactions. The image was generated via the PathwayMapper software [75].

1.7.6 Applications For Analysis of Cancer Cohorts.

Over the years, various apps have been developed for the interactive analysis of mutation data. Some of these apps, however, usually serve a very specific purpose, such as mutational signature extraction, some offer a subset of analysis, and most do not offer a cross-cohort comparison. Perhaps the app with the highest number of functionalities is MutScape [109]. This app can accept multiple input types, help the user with variant filtering, offers co-mutation plot and various visualizations, signature identification, and *de novo* driver gene identification. Along the same line, cBioPortal hosts various

big datasets which include expression, mutation, and copy number alteration data. cBioPortal does not allow for the upload of new datasets because of the very specific file and annotation formats its functionalities depend on. This application allows various visualizations such as oncoplots and co-mutations plots. Additionally, comparisons can be made on all data types available with the groups defined based on clinical information or alteration presence. Additionally, this tool provides a pathway visualization scheme but does not offer any enrichment or comparison statistics [99].

Other apps are specialized in specific analyses. vcfView, for instance, works on VCF data and aims to aid the user in visualizing the effects of filters based on the VCF field specifications. This app also provides lollipop plot visualization and signature fitting [110]. CoMut package only performs co-mutation analysis [111]. MuSiCa specializes specifically in the signature fitting of the supplied cancer cohort [112]. PathwayMapper, on the other hand, offers a framework to both explore the pre-defined pathways and also create new ones [75].

2 Chapter 2 AIMS AND RATIONALE

In this thesis I aimed to develop an interactive web application that can:

- Simplify and make complex analyses accessible to wet lab scientists;
- Provide them with means to explore and filter their data
- Aid them on *de novo* mutational signature extraction and clinical correlate identification
- Guide data harmonization for TCGA-BRCA cohort comparison
- Provide utility for signature comparison with TCGA-BRCA cohort
- Allow identification of *de novo* driver genes and differentially mutated genes
- Help with extensive pathway comparison between TCGA-BRCA and uploaded cohorts

I aimed to demonstrate the utility of this app via case studies, and feature comparison with previously published apps. Finally, I aimed to use the custom gene somatic interactions module to annotate the somatic interactions of *CHRNA5* and *TP53* genes with a list of *TP53* targets by expression profiles of *CHRNA5* and *TP53* depleted breast cancer cells. Such a novel approach can help understand the significant mutation-expression associations.

3 Chapter 3 MATERIALS AND METHODS

3.1 Materials

Table 1. List of software packages used

Name	Reference	Package Github Repository
data.table	NA	https://github.com/Rdatatable/data.table
ggpmisc	NA	https://github.com/aphalo/ggpmisc
ggpubR	NA	https://github.com/kassambara/ggpubr
MutationalPatterns	[113]	https://github.com/UMCUGenetics/MutationalPatterns
ggfortify	[114]	https://github.com/sinhrks/ggfortify
ggrepel	NA	https://github.com/slowkow/ggrepel
NMF	[115]	https://github.com/renozao/NMF
ggsignif	[116]	https://github.com/const-ae/ggsignif
ggbreak	[117]	https://github.com/YuLab-SMU/ggbreak
ggbeeswarm	NA	https://github.com/eclarke/ggbeeswarm
BSgenome.Hsapiens.UCSC.hg19	NA	https://github.com/Przemol/BSgenome.Hsapiens.UCSC.hg19

EnsDb.Hsapiens.v75	[118]	https://github.com/jorainer/EnsDb.Hsapiens.v75
AnnotationDbi	NA	https://github.com/Bioconductor/AnnotationDbi
clusterProfiler	[119]	https://github.com/YuLab-SMU/clusterProfiler
shiny	NA	https://github.com/rstudio/shiny
shinydashboard	NA	https://github.com/rstudio/shinydashboard
DT	NA	https://github.com/rstudio/DT
rmarkdown	NA	https://github.com/rstudio/rmarkdown
RColorBrewer	NA	https://github.com/cran/RColorBrewer
dplyr	NA	https://github.com/tidyverse/dplyr
shinycssloaders	NA	https://github.com/daattali/shinycssloaders
gplots	NA	https://github.com/cran/gplots
genefu	[120]	https://github.com/bhklab/genefu
ggplot2	[121]	https://github.com/tidyverse/ggplot2
plotly	NA	https://github.com/plotly
shinythemes	NA	https://github.com/rstudio/shinythemes

heatmaply	[122]	https://github.com/talgalili/heatmaply
maftools	[123]	https://github.com/PoisonAlien/maftools
shinyjs	NA	https://github.com/daattali/shinyjs
Hmisc	NA	https://github.com/harrelfe/Hmisc
shinyWidgets	NA	https://github.com/dreamRs/shinyWidgets
visNetwork	NA	https://github.com/datastorm-open/visNetwork
segmented	[124]	https://github.com/cran/segmented

Table 2. List of reference files used

Name	Reference
Oncogenic Pathways	[68], [123]
Gene mapping to MutSig Gene names	[123]
DNA Damage Repair pathways genes	[125]
Splicing Factor genes	[126]
Metastatic Breast Cancer MAF file of Gustave Roussy Institute (IGR) dataset	[99], [127]
Metastatic Breast Cancer clinical information file of IGR dataset	[127]

Metastatic Breast Cancer MAF file of alpelisib trial dataset	[99], [128]
Metastatic Breast Cancer clinical information file of alpelisib trial dataset	[99], [128]
DREAM complex Targets	[129]
RB-E2F targets	[129]
MMB-FOXM1 Targets	[129]
Cell Cycle Genes	[129]
LIN37 Dependent TP53 Targets	[130]
LIN37 Independent TP53 Targets	[130]

3.2 Methods

3.2.1 Data Access and Storage

MAF files of the IGR and alpelisib trial datasets wereaccessed and downloaded from the cBioPortal [99], [127], [128]. The clinical information of the IGR dataset was obtained directly from the original publication instead of cBioPortal because it contained more information regarding the site of biopsy of the sample [127]. The clinical and mutational information was merged via the MAF object of the maftools [123]. Mutational and clinical information of the TCGA samples were accessed via the tcgaLoad() function of the maftools [123].

3.2.2 Pathway annotations

Oncogenic pathways come pre-packaged as "oncogenic_pathways.tsv" with the maftools package, which gathered them from the reference [68]. Splicing factors, DNA damage pathway genes, DREAM complex targets, RB-E2F targets, MMM-FOXM1 targets, cell cycle genes, LIN37 dependent TP53 targets, and LIN37 independent TP53 targets annotations were taken from their respective papers [125], [126], [129], [130].

3.2.3 Mutational Signature Analysis

Mutational signatures were done using the maftools package. First, the trinucleotide matrix of the genome was extracted by the trinucleotideMatrix() function, followed by the estimation of the signatures via the estimateSignatures() function with the user defined number of maximum signatures to try, and the user defined bootstrap values. Afterward, the desired number of signatures was extracted by the extractSignatures() function. Finally, similarities of the signatures were compared to the signatures defined in the COSMIC database via the compareSignatures() function [41].

The same steps were carried out for the TCGA BRCA dataset, except the number of signatures were predefined by me [26].

3.2.4 Identification of Differentially Mutated Genes and Pathways

Differentially mutated genes were detected by the mafCompare() function meanwhile differentially mutated pathways were detected by the mafCompare_pathway() function. Forest plots were drawn by the forestPlot() function [123].

3.2.5 Identification of Somatic Interactions.

Co-occurring and mutually exclusive sets of genes were identified between top 25 mutated genes within each dataset and plotted by the somaticInteractions() function of the maftools. Function calls Fisher's exact test internally and calculates the respective odds ratio (OR) and p-value. The network displaying both of the datasets was generated with the help of the visNetwork package.

For the somatic interaction analysis of the user selected gene, I conducted a Fisher's exact test based analysis by comparing the mutational profile of the user selected gene to other genes in the dataset via a custom function. Finally, p-values and FDRs (false discovery rate) werereported for each comparison.

3.2.6 Identification of Driver Genes

Driver genes were identified by the oncodrive() function of the maftools, which invokes an R implementation of the oncodriveCLUST internally. Briefly, this tool divides each gene into the clusters of positional neighborhoods and estimates a score for each neighborhood based on its width in terms of amino acids and the number of mutations falling on that neighbor. Afterward, a single score is calculated for each gene by

summing up the scores of positional neighborhoods within that gene. Finally, a p-value for each gene is calculated by comparing its score with the background distribution of the scores [50].

3.2.7 Harmonization of Gene Names

As datasets can use different nomenclature for the gene names; hence, a reference file within the maftools that provides common aliases for the HUGO symbols was used to convert the TCGA dataset, uploaded dataset, and the target file to the standard HUGO symbols [131].

3.2.8 Identification of Optimal Variant Allele Frequency Cutoff

I computed the tumor mutation border (TMB), defined as the number of non-synonymous alterations per patient, for each patient in TCGA, and the uploaded dataset for each variant allele frequencies (VAF) cutoff between 0 and 0.5 with an increment of 0.05. The optimal cutoff is defined as the lowest cutoff at which the two cohorts have the same TMB distribution.

3.2.9 Reactivity

It is important for a user interface to be interactive and to respond to a variety of inputs, called reactivity. Shiny is a package/framework of RStudio that allows for putting together a UI and server function that implements the user interface and the functions to run within this interface, respectively (https://github.com/rstudio/shiny). In the present thesis, I have used Shiny and R packages stated above to generate a reactive and user-interactive analysis and visualization tool for comparing a MAF file from a user supplied cohort of BRCA with that of TCGA-BRCA with respect to mutation signatures, driver mutations, oncogenic pathways all in one webtool. The details of the

app have been provided in the results section of the thesis. **Figure 6** summarizes the overall workflow.

3.2.10 Hypermutator Detection

To detect the hypermutators, i.e., the samples with extremely high TMB, I utilized the segmented R package [124] Briefly, this package identifies the optimal points at which the slope of the fitted regression line changes significantly. Here, the TMB of the samples and their rank were used to perform the segmented correlation and the subsequent identification of the TMB cutoff. Every sample above this cutoff is regarded as a hypermutator. These samples can be excluded from the downstream analyses.

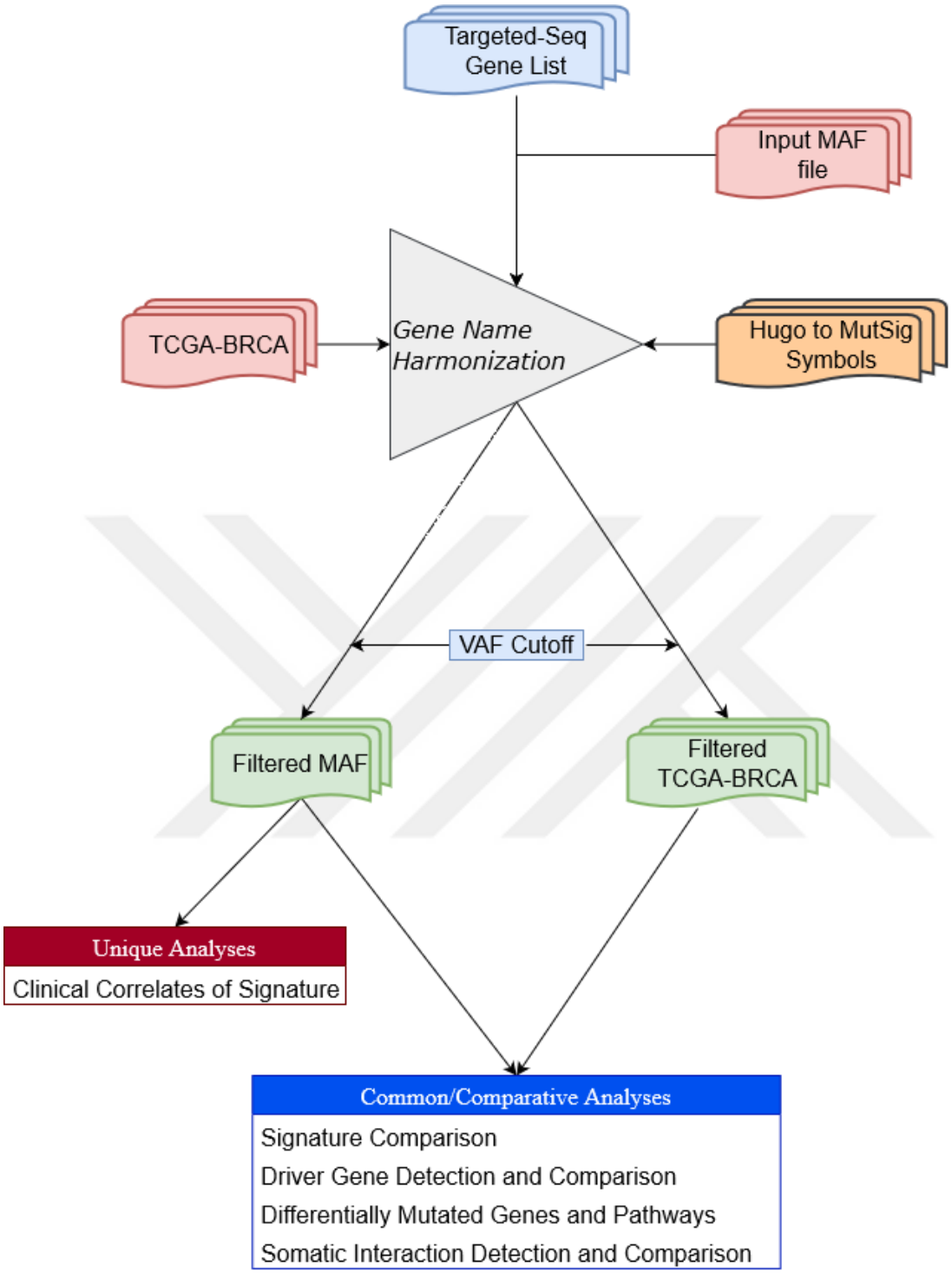


Figure 6. A flowchart that describes the workflow of the application.

4 Chapter 4 RESULTS

4.1 BRCA:Cohort-vs-TCGA App

This app provides an interactive user interface to explore and compare the MAF file of an uploaded breast cancer cohort with that of the TCGA-BRCA cohort. App also contains download buttons to extract both the source data and the figures.

4.1.1 "Data Upload" Tab

The "Data Upload" tab requires the user to upload a mutation file in the MAF format and the associated clinical information. The user may choose to upload a "target" file containing gene names targeted by their protocol if they generated their data via a targeted sequencing protocol. After uploading the required files, the user has to click the "Click to upload" button to start their analysis. An information button is also available with the links to the MAF format specification and vcf2maf utility which they can use to convert their VCF file to the MAF file [22] (**Figure 7**).

Figure 7. A screenshot from the “Data Upload” tab of the “BRCA:Cohort-vs-TCGA” app.

4.1.2 “Mutation and Clinical Data” Tab

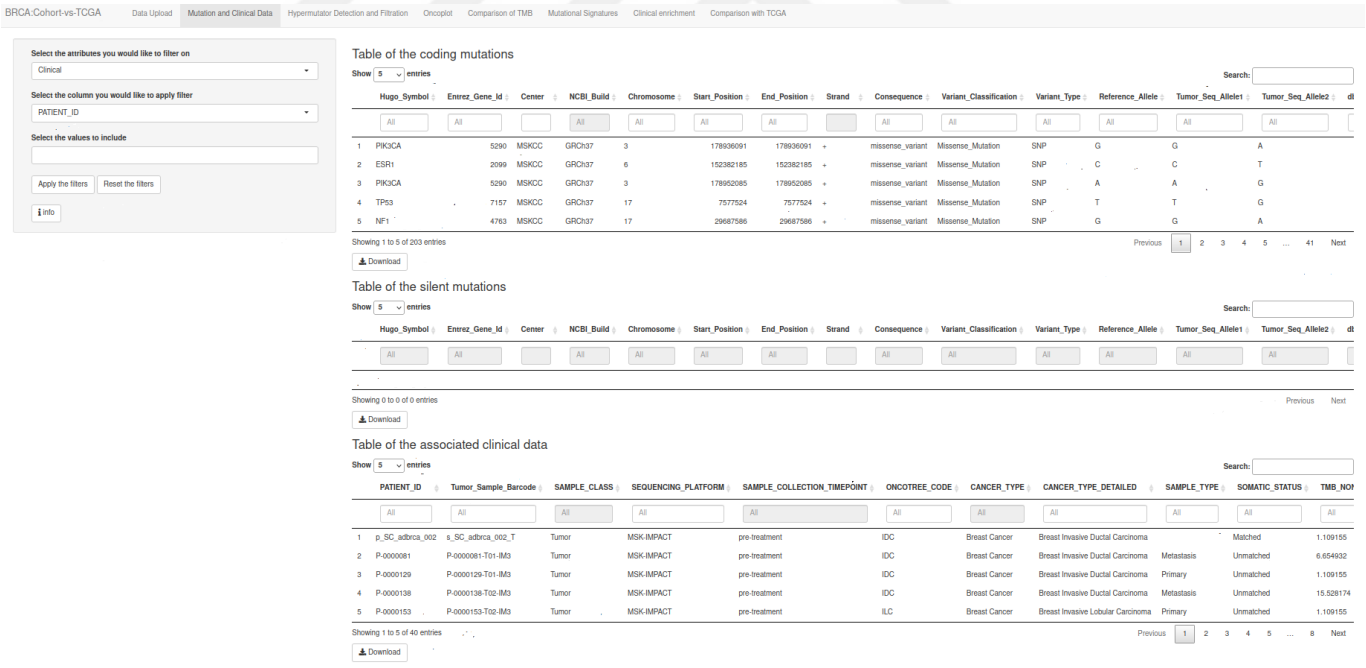


Figure 8. A screenshot from the “Mutation and Clinical Data” tab of the “BRCA:Cohort-vs-TCGA” app.

The “Mutation and Clinical Data” tab provides the overall view of the coding mutations, silent mutations, and the clinical data for the selected cohort. Each of the displayed

tables can be filtered by the values of the column with the help of the search bars, for display and exploration purposes. The sidebar allows the user to filter their dataset according to a column from either the mutation or the clinical file. The selected filter is applied via the “Apply the filters” button. Once the user clicks the “Apply the filters” button, the updated dataset will be displayed. Each consecutive click to this button will apply the chosen filters on the new, filtered dataset. As a result, “Apply the filters” button works as an AND gate. The user can reset the applied filters via the “Reset the filters” button if they wish. This page contains download buttons for each of the tables displayed as well as the information button which explains the functions of the buttons (**Figure 8**). The sidebar is able to understand the difference between the numeric and character filters and hence adapts the filters for each accordingly (**Figure 9**). The resultant dataset is used for the subsequent analyses.

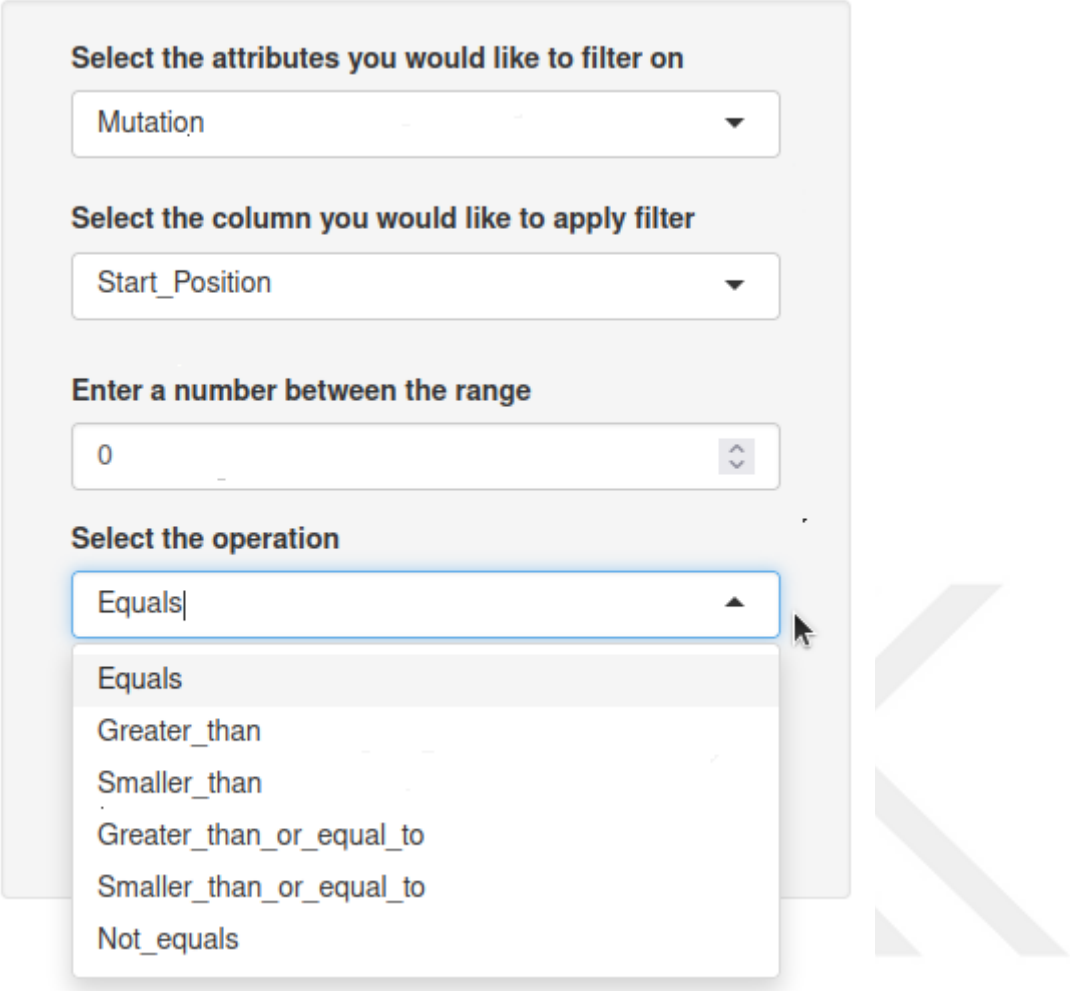


Figure 9. A screenshot displaying the filter options for the numeric columns, from the "Mutation and Clinical Data" tab of the "BRCA:Cohort-vs-TCGA" app.

4.1.3 "Hypermutator Detection and Filtration"

Presence of hypermutators can influence the Fisher's exact test based analyses and increase the false positive rate [62]. Hence, identification of hypermutator samples and the option of excluding them from the downstream analyses is desirable. The "Hypermutator Detection and Filtration" tab first conducts the segmented regression analysis to identify the optimal TMB cutoff, as used by many recent publications [132]–[137]. Upon the completion of segmented regression analysis, a plot showing the regression result and the table with the source data are outputted. The user has the option to download both

the plot and the table. After detection of the optimal TMB cutoff, the user can choose to filter the dataset with the TMB cutoff. Figure 10 displays the result of an example segmented regression analysis. Since the dataset used in this demonstration comes from a targeted sequencing analysis, the results are not informative in this case.

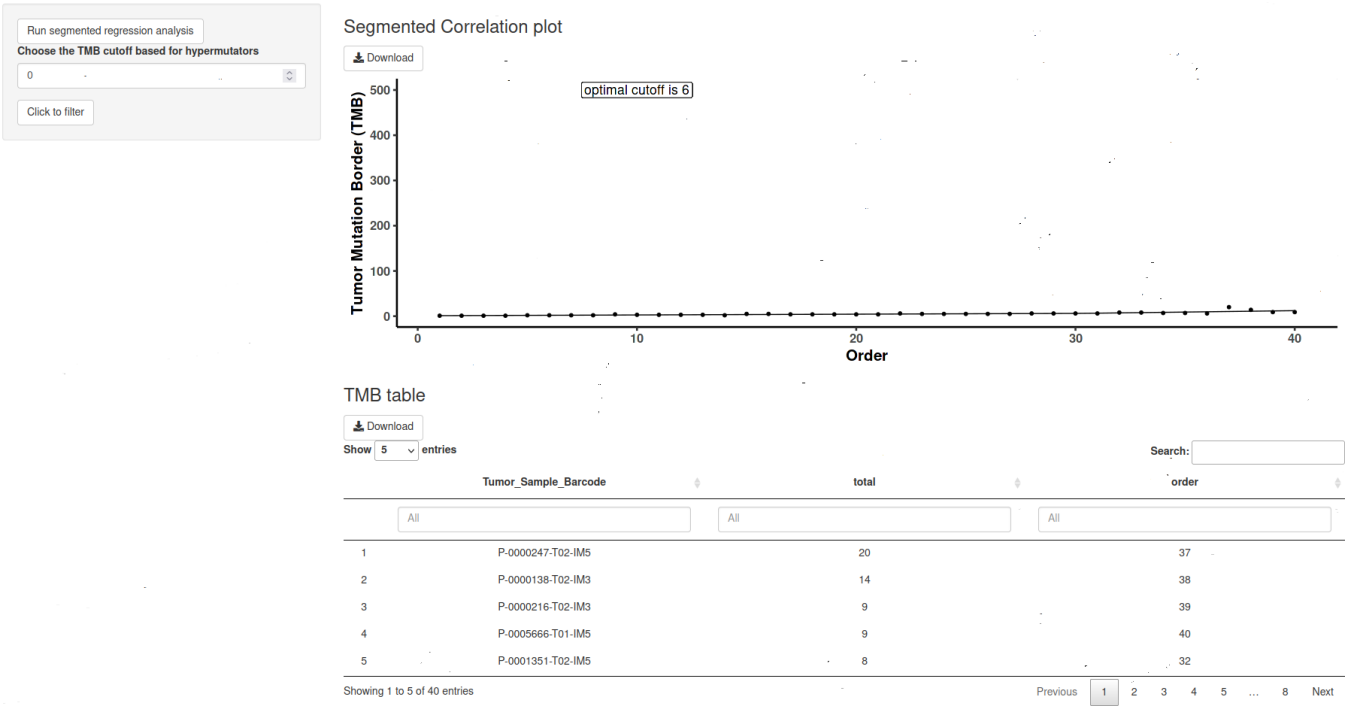


Figure 10. A screenshot from the “Hypermutator Detection and Filtration” tab of the “BRCA:Cohort-vs-TCGA” app.

4.1.4 “Oncoplot” Tab

Oncoplot is a convenient way to summarize the mutational profile of a cohort. The upper panel of the oncoplot includes the TMB barplot for each patient, colored by the number of each mutation consequence in the corresponding patient. The right panel summarizes the number of mutations for that gene within the cohort by a stacked barplot, again colored by the consequence of the mutations with the black color

representing the patients with multiple mutations for that gene. The plot also gives the percent of the patients mutant for each gene. Genes are ordered according to the number of mutant patients. The sidebar allows for multiple annotations from the clinical data, and can order the patients according to the first chosen column. Additionally, the user can choose to summarize transitions and transversions for each patient. The number of genes displayed is chosen by the user. Moreover, the user can control several aesthetic parameters such as font size of the genes, title, legend, annotation and samples. A "Download" button is also made available (**Figure 11**). In the example dataset, it can be seen that the genes with the highest mutation rate are *PIK3CA*, *CDH1*, and *ESR1* (**Figure 11**).

Figure 11. A screenshot from the "Oncoplot" tab of the "BRCA:Cohort-vs-TCGA" app.

4.1.5 "Comparison of TMB" Tab

The data generated from the different studies can come from the different protocols. Differences can exist in the regions captured, the depth, and the mutation calling algorithms and these differences can hamper the subsequent comparison with a published study. Published studies try to overcome these differences via different approaches. For example, a cohort study of clear cell renal cell carcinoma claimed that their cohort is comparable to the TCGA cohort by comparing the TMB of each cohort and showing that they do not differ [138]. Yet another study on metastatic breast cancer reanalyzed a subset of TCGA samples with their pipeline and compared the number of mutations identified from the TCGA pipeline and their pipeline. They identified that the number of mutations correlated with R^2 of 0.98 along the x=y line and concluded that these two cohorts are comparable [127]. Based on these findings, I reasoned that a VAF cutoff at which the difference between the TMBs of two cohorts disappears can be used in order to make them comparable. I implemented this technique in the "Comparison of TMB" tab. The TMB difference is assessed at different cutoffs ranging from 0 to 0.5 with an increment of 0.05. The user can choose to implement either a Student's t-test or a Wilcoxon signed-rank test. The selected cutoff can then be used to filter the dataset using the "Data Upload" tab. Several aesthetics can be modified by the user and both the plot and source data can be downloaded afterward (**Figure 12**). As seen in **Figure 12**, the TMB difference was significant at all of the VAF cutoffs. This is expected since TCGA-BRCA was profiled by WES and the example dataset was profiled by targeted sequencing.

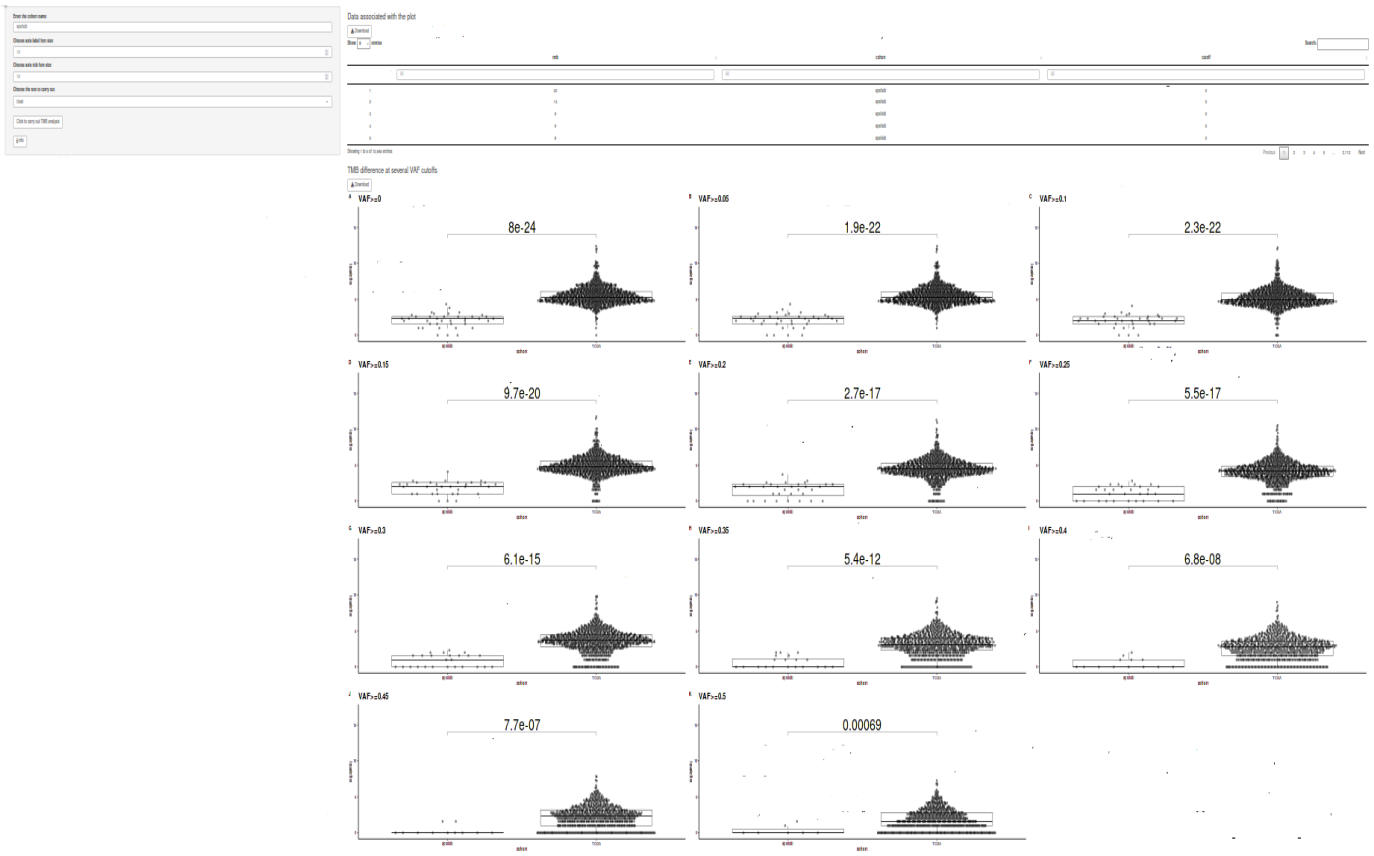


Figure 12. A screenshot from the “Comparison of TMB” tab of the “BRCA:Cohort-vs-TCGA” app.

4.1.6 “Mutational Signatures” Tab

This tab is designed to help the user to detect the optimal number of signatures and extract them

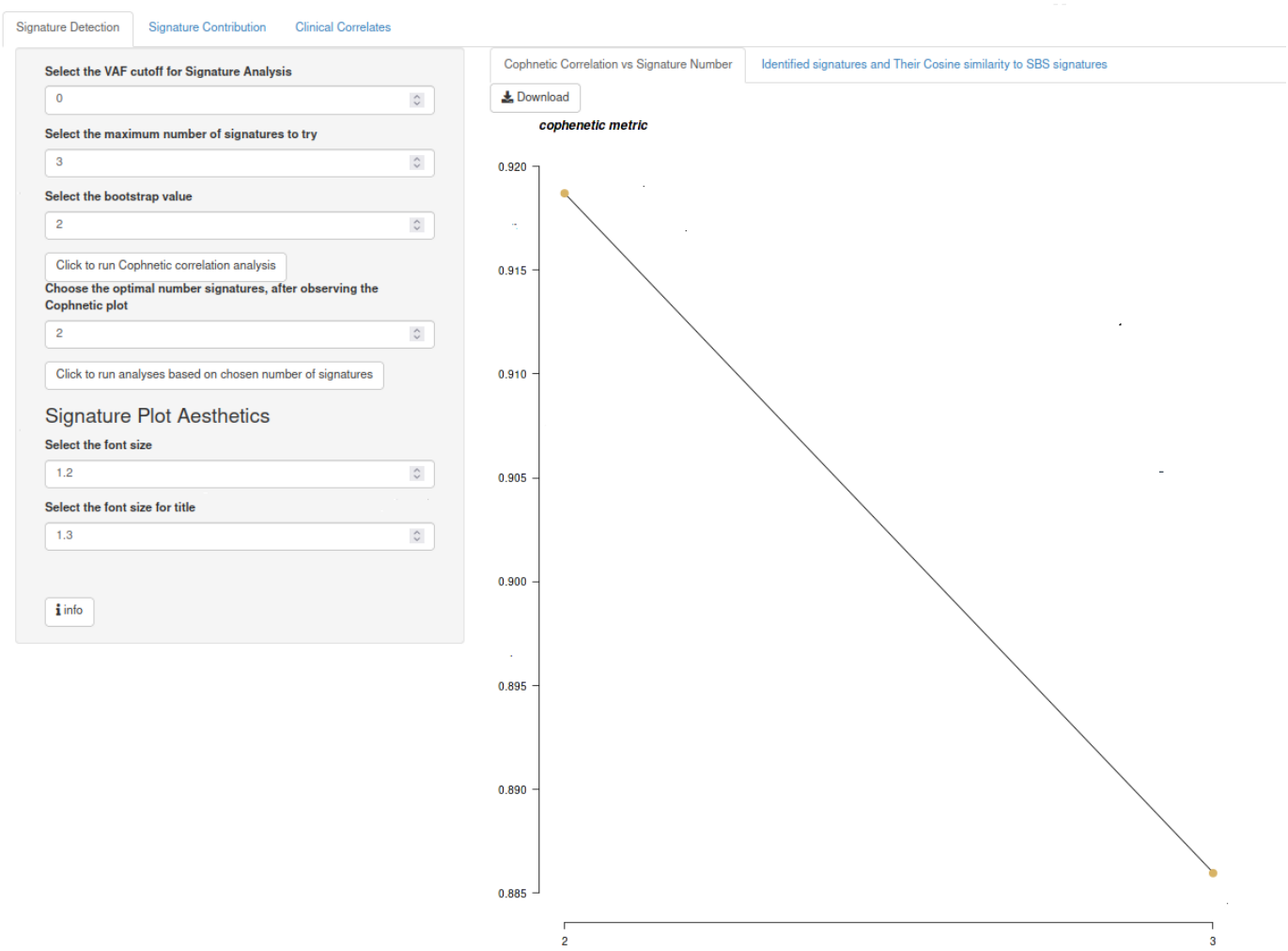


Figure 13. A screenshot from the "Mutational Signatures" tab of the "BRCA:Cohort-vs-TCGA" app.

The tab has three sub-tabs. The first step is to run the cophenetic correlation analysis to determine the optimal number of signatures. The user can select a VAF cutoff, determine the maximal number of signatures to carry out the cophenetic correlation analysis and select the number of bootstraps. When the analysis is finished, the user can download the plot (**Figure 13**).

Once the user has decided on the optimal number of signatures, usually determined with a "knee" on the cophenetic correlation plot, they can click the "Click to run analyses based on the chosen number of signatures" to extract the signatures. Fitted signatures,

their values, and the COSMIC single base substitution (SBS) signature they are most similar to can be viewed in the “Identified signatures and their cosine similarity to SBS signatures” tab. The aesthetics of the plot can be modified by the user via the sidebar. Finally, the user can download the signatures plot (**Figure 14**). Signature analysis conducted on the example dataset identifies three signatures that are similar to SBS2, SBS5 and SBS15 of the COSMIC database (**Figure 14**).

Figure 14. A screenshot from the “Identified signatures and their cosine similarity to SBS signatures” subtab of the “Mutational Signatures” tab.

The next subtab, “Signature Contribution”, reports each signature's contribution to each patient's mutational profile. Results are made available as a table and as a plot with their respective download buttons (**Figure 15**).

The “Clinical Correlates” subtab tries to correlate the contribution from the signatures with clinical data available. The first subtab of the “Clinical Correlates”, “TMB vs Contributions” fits a linear line between the TMB of each patient and the contribution of each signature.

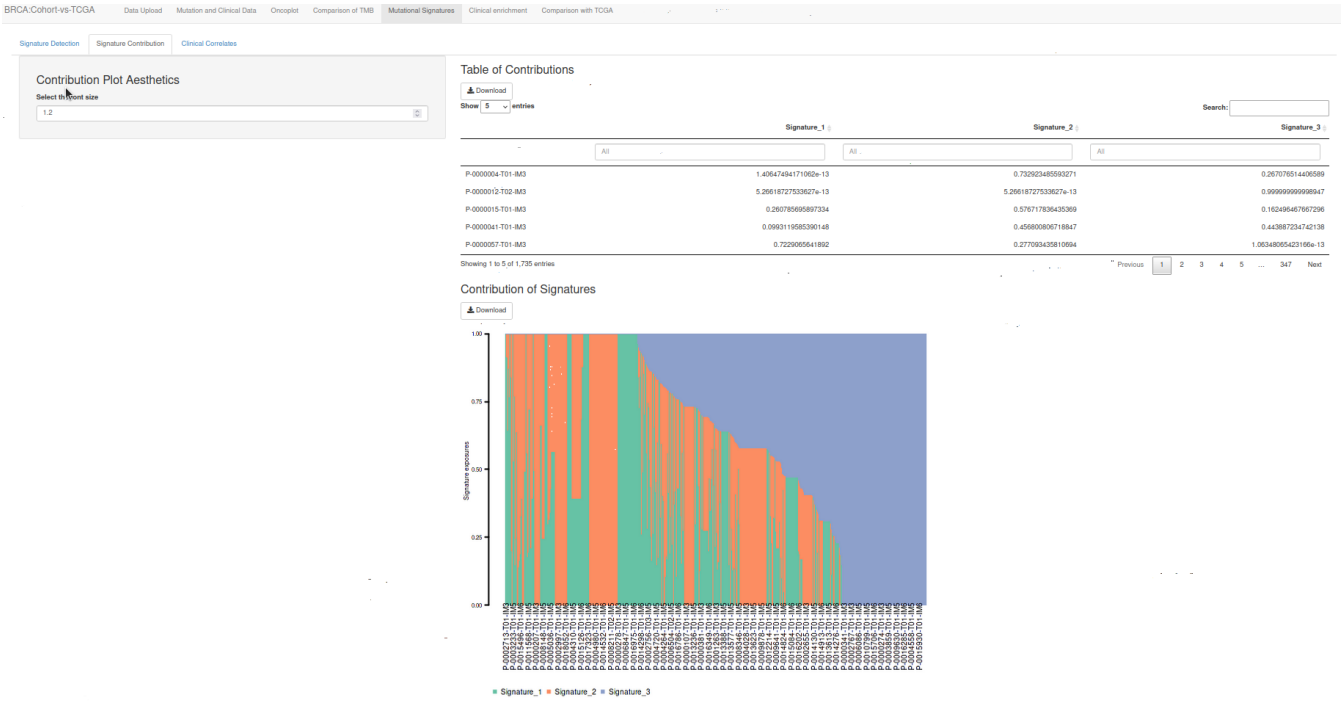


Figure 15. A screenshot from the “Signature Contribution” subtab of the “Mutational Signatures” tab.

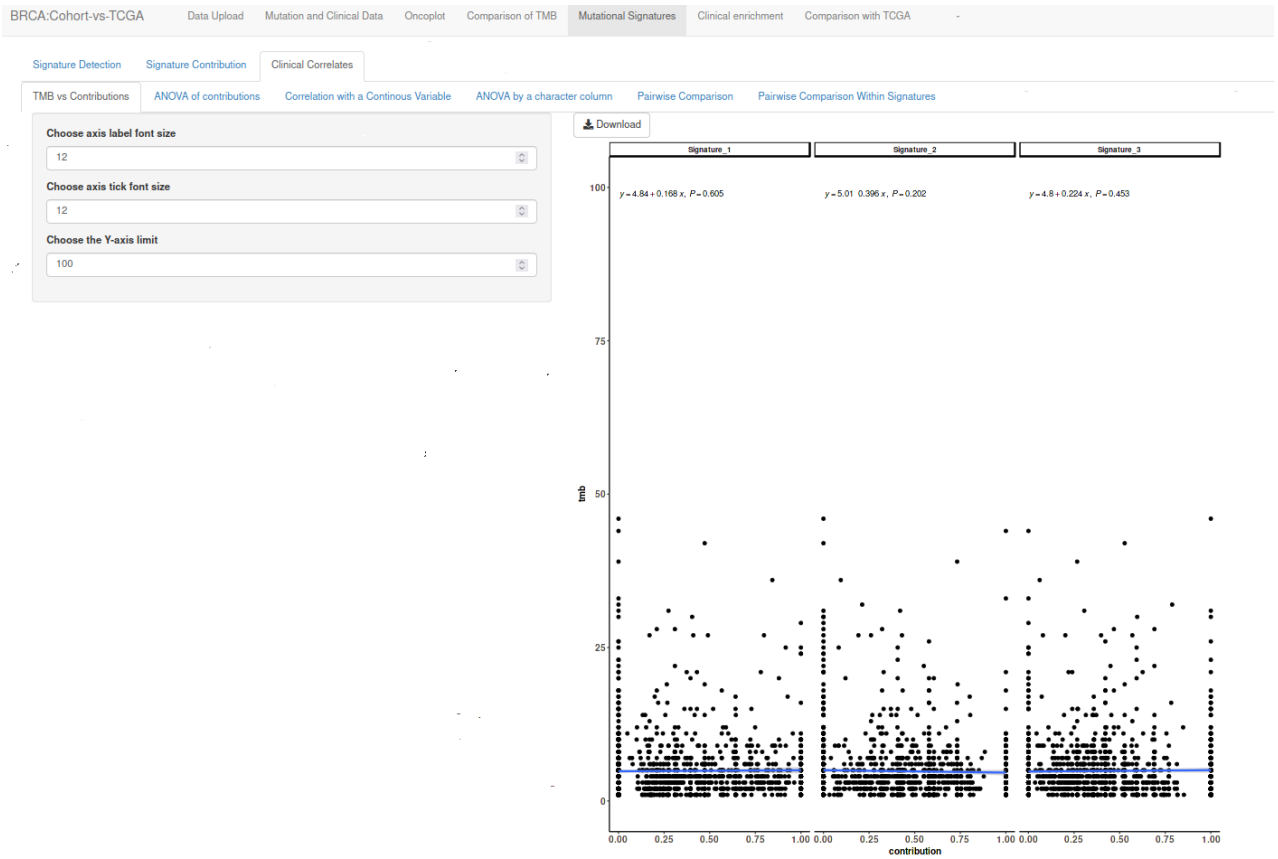


Figure 16. A screenshot from the "TMB vs Contribution" subtab of the "Mutational Signatures" tab.

The aesthetics of the plot can be adjusted, and it can be downloaded. The user can choose to select a TMB cutoff to eliminate the effect of any outliers (**Figure 16**).

The next subtab of the "Clinical Correlates" tab, "ANOVA of contributions" carries out an analysis of variance (ANOVA) analysis to test whether the contributions of the signatures differ within the cohort. There are controls to modify the appearance of the plot and to download it (**Figure 17**). **Figure 17** showed that the contributions of signatures were not the same in the example cohort.

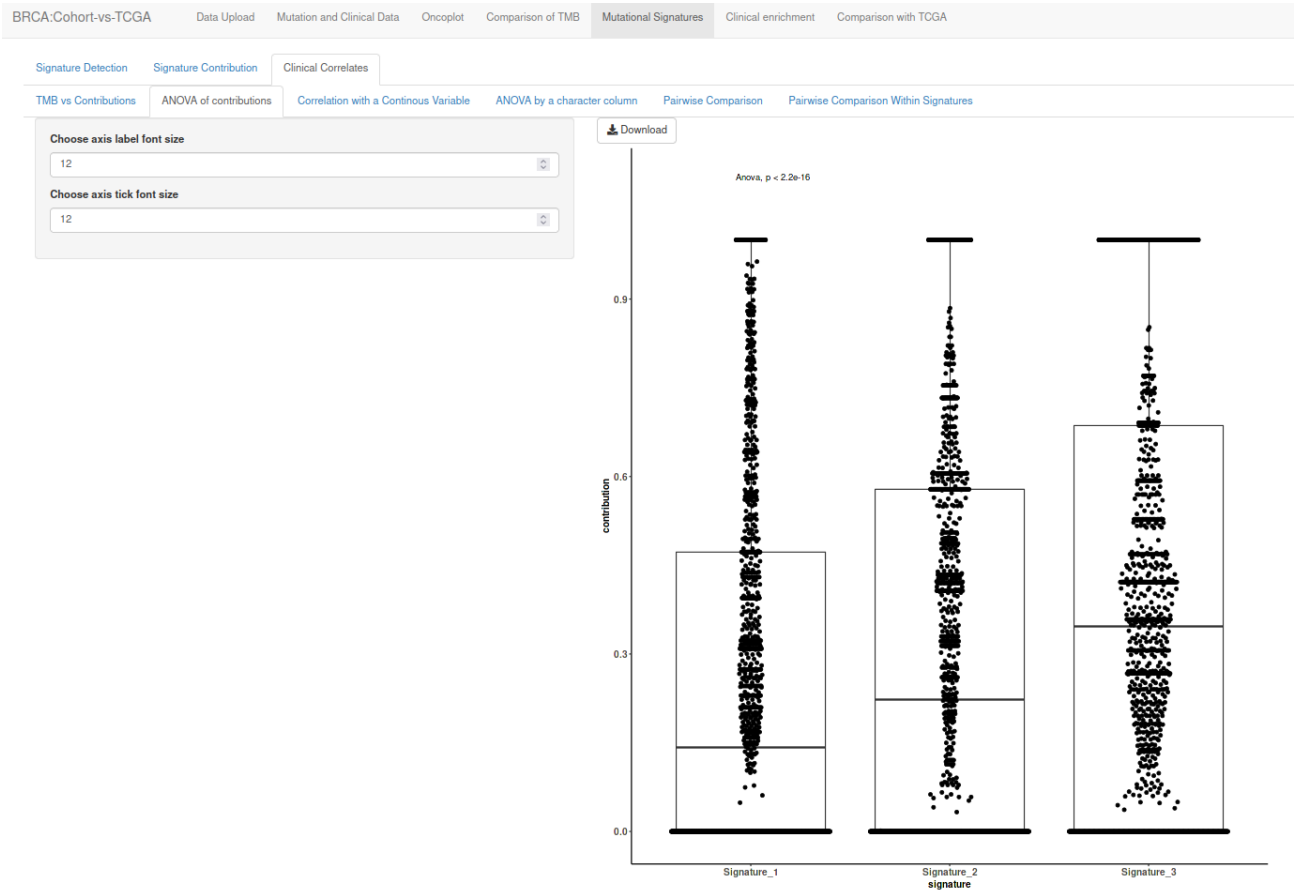


Figure 17. A screenshot from the "ANOVA of contributions" subtab of the "Mutational Signatures" tab.

"Correlation with a Continuous Variable" subtab lets the user carry out a regression analysis with a continuous variable from the clinical data. The user has the option to modify the plot and download it (**Figure 18**). In **Figure 18,** time to death was correlated with the signature contributions, and none of the signatures were found to contribute significantly to this variable.

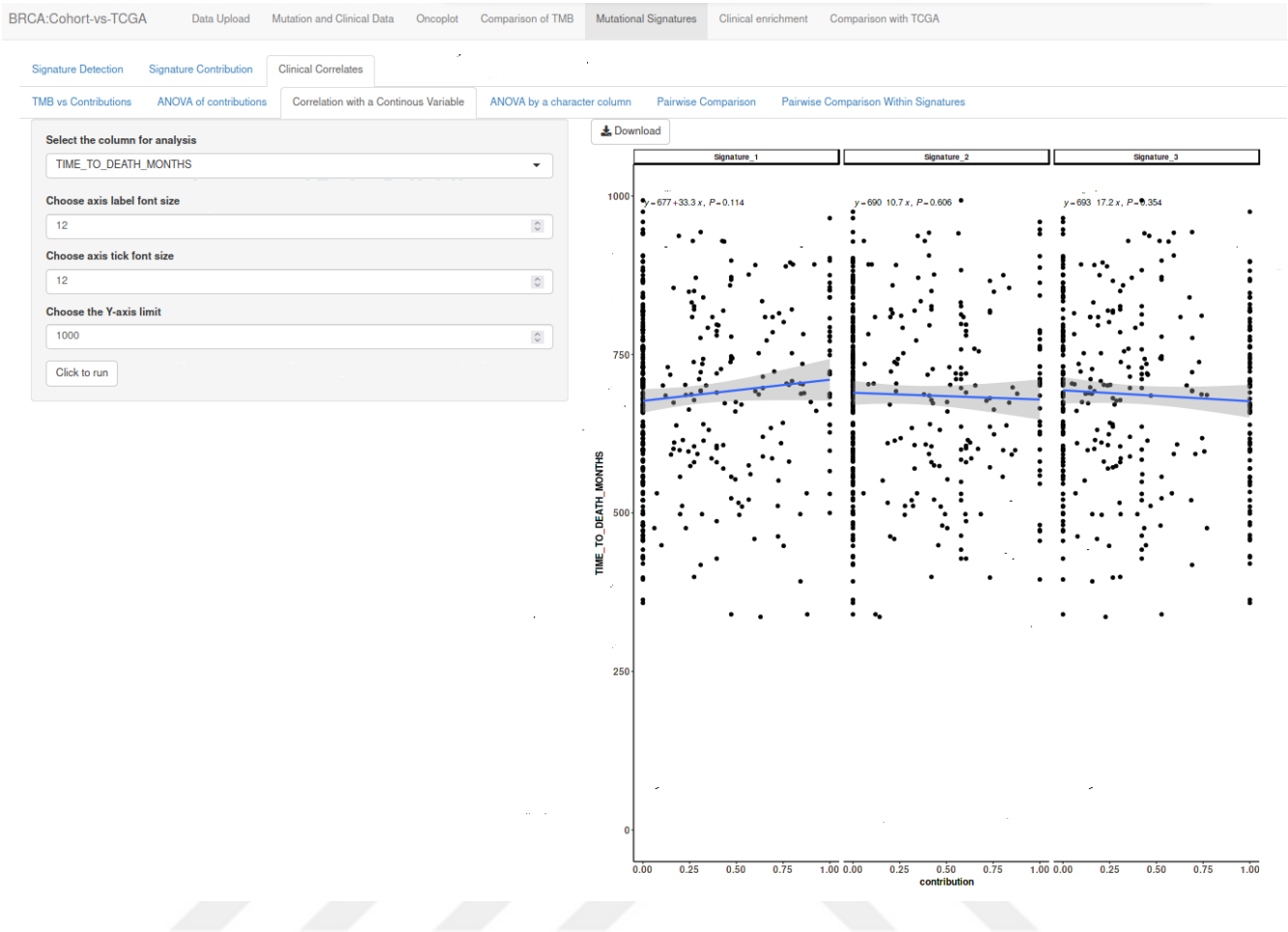


Figure 18. A screenshot from the "Correlation with a Continuous Variable" subtab of the "Mutational Signatures" tab.

The following subtab of the "Clinical Correlates" tab, "ANOVA by a character column" subtab, allows the user to carry out an ANOVA analysis of the contribution of signatures within the levels of a clinical column. User has the choice to adjust the aesthetic parameters and download the plot (**Figure 19**). In the example dataset, signatures exhibited significant differences in their contributions to the patients with

no previous breast cancer history (**Figure 19**).

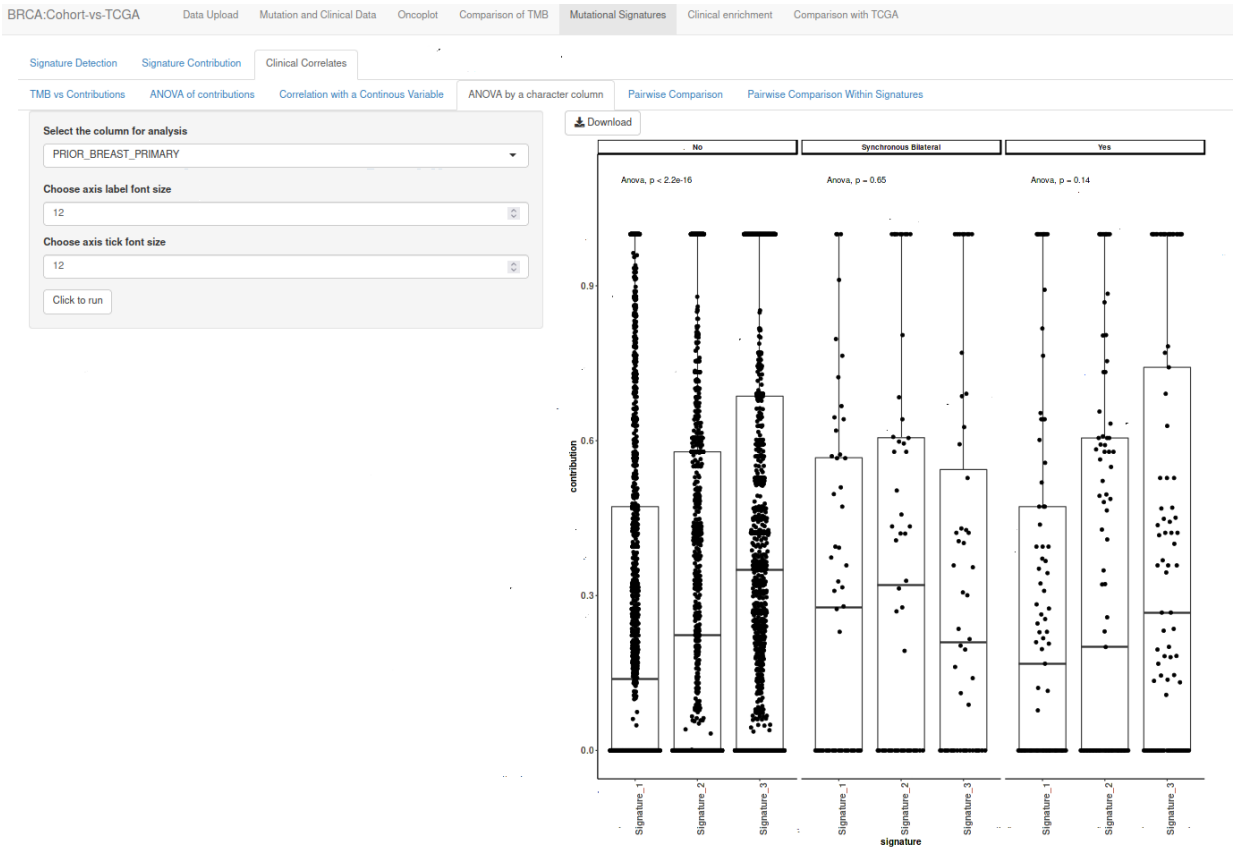


Figure 19. A screenshot from the "ANOVA by a character column" subtab of the "Mutational Signatures" tab.

The "Pairwise Comparison" subtab of the "Clinical Correlates" tab makes it possible to compare if two signatures contribute differentially to the two selected levels of a chosen clinical column via Student's t-test. User can alter the aesthetics of the plot and subsequently download it (**Figure 20**). In **Figure 20**, it is shown that none of the signatures exhibited a differential contribution between patients with and without previous history of breast cancer.

Finally, the last subtab of the "Clinical Correlates" tab, "Pairwise Comparison within Signatures" lets the user compare the contribution of two selected signatures within the levels of a column of the clinical data with the help of the via Student's t-test. The user

is presented with the option to modify the aesthetics of the plot and subsequently download it (**Figure 21)**. In the example dataset, Signature 2 was found to contribute significantly higher to the mutational profile of the patients without previous breast cancer history than Signature 1 (**Figure 21**).

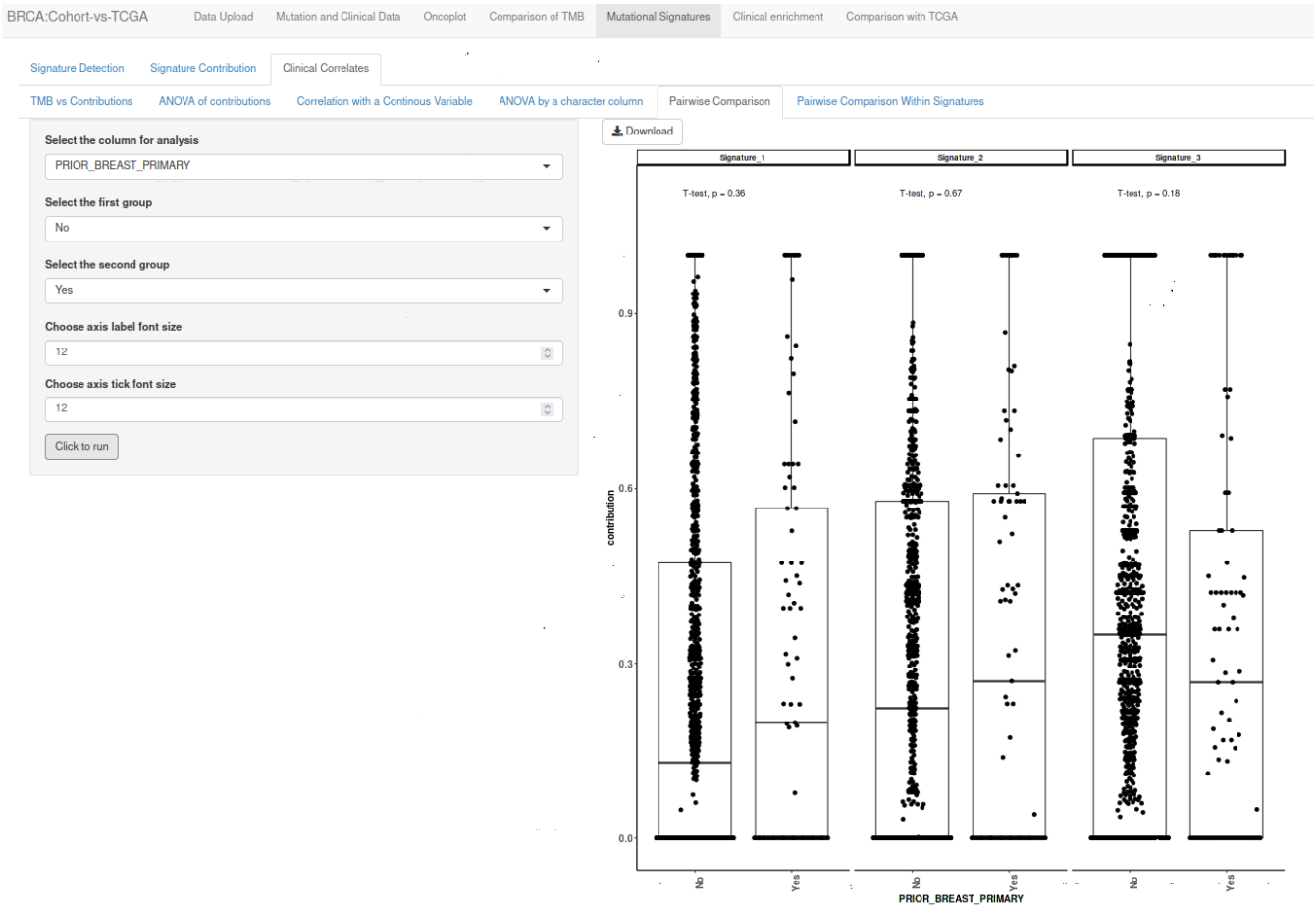


Figure 20. A screenshot from the “Pairwise Comparison” subtab of the “Mutational Signatures” tab.

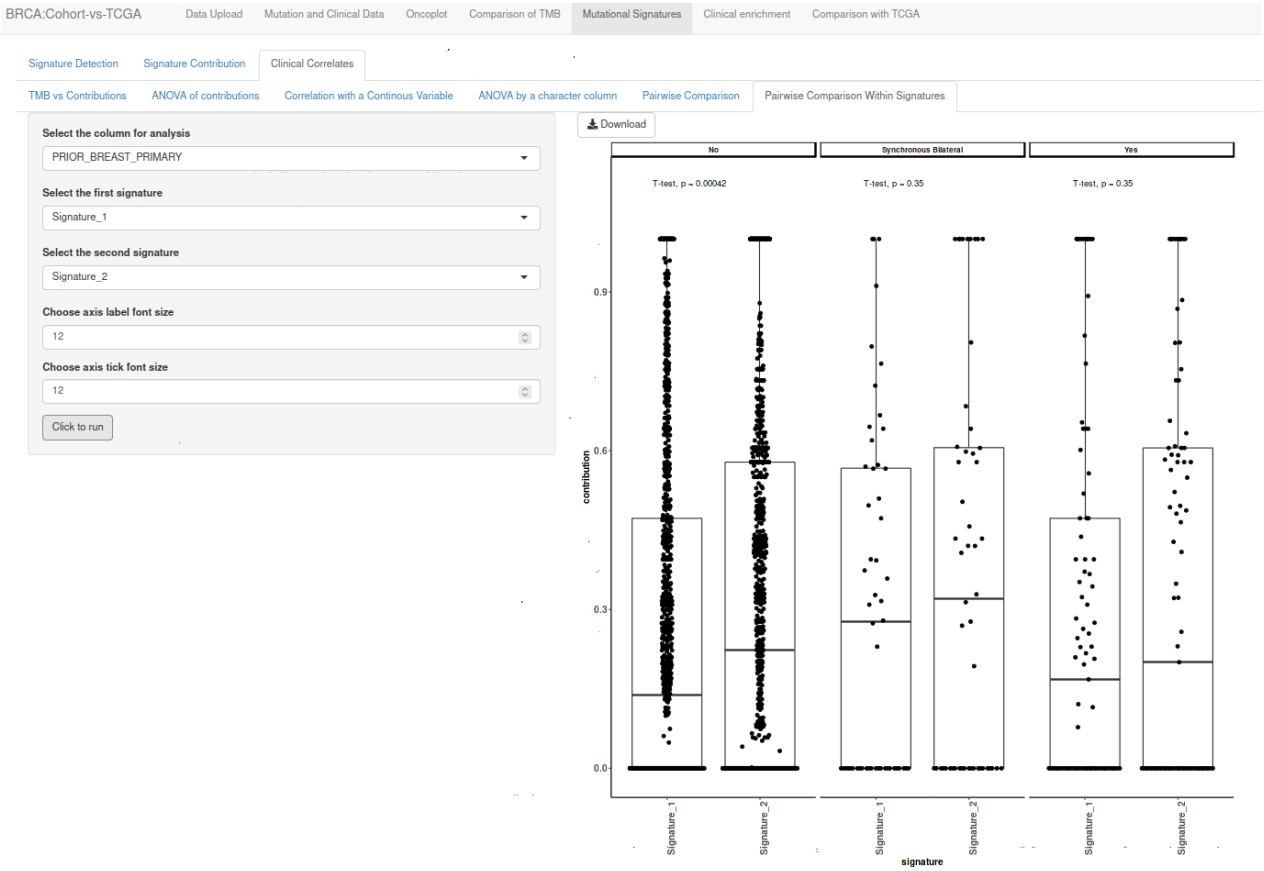


Figure 21. A screenshot from the “Pairwise Comparison within Signatures” subtab of the “Mutational Signatures” tab.

4.2 “Clinical Enrichment” Tab

The “Clinical enrichment” tab presents the user with the option to conduct clinical enrichment analysis for each gene within levels of a selected column. The analysis is done with the help of Fisher’s exact test, first via comparing each group to another and second via comparing each group to the rest of the sample. The results for the pairwise enrichment analysis are reported on the “Pairwise Enrichment Results Table” tab, meanwhile, results for the groupwise enrichment are reported on the “Groupwise Enrichment Results” tab. Results of the groupwise enrichment analysis can also be plotted on the “Clinical Enrichment plot” tab with the control over the p-value threshold, OR threshold, and aesthetics. Users are given the option to download the

results of each analysis from each tab (**Figures 22-24**). *NCOR1* mutations were found to be enriched in patients with a previous history of breast cancer when compared to the patients with no history of breast cancer (**Figure 22**). Identical results were observed when these patients were compared to the rest of the patients (**Figure 23**). **Figure 24** represents the tabular results of **Figure 23** in a barplot format.

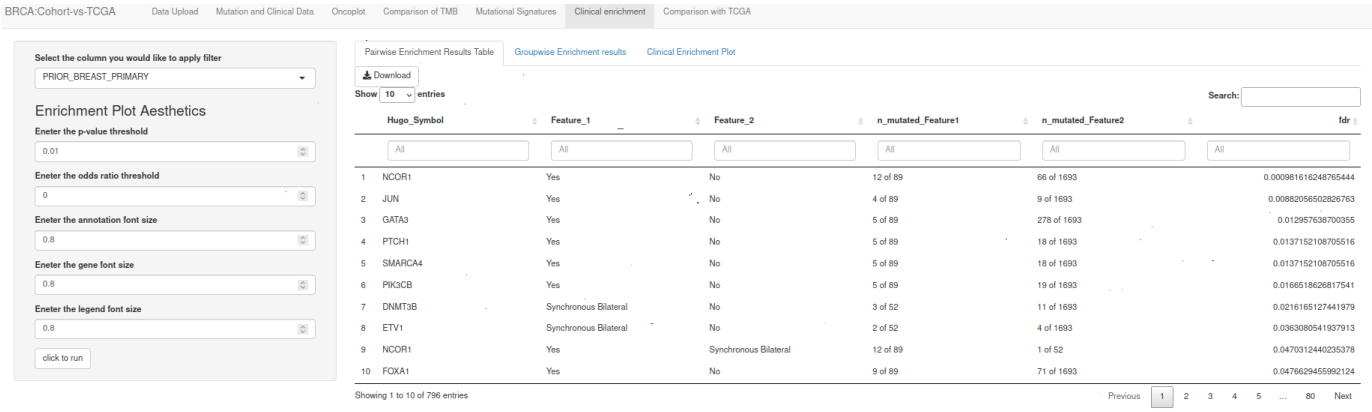


Figure 22. A screenshot from the "Pairwise Enrichment Results Table" subtab of the "Clinical enrichment" tab.

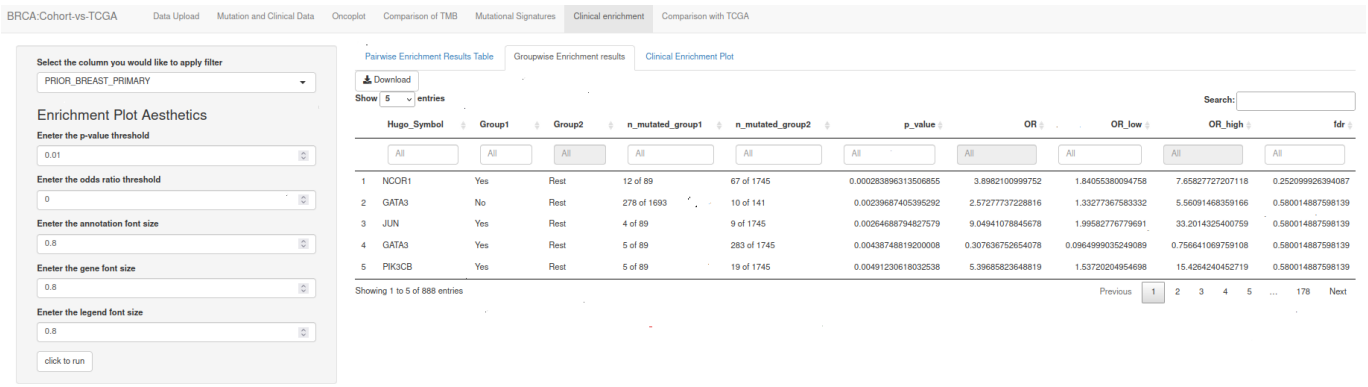


Figure 23. A screenshot from the "Groupwise Enrichment results" subtab of the "Clinical enrichment" tab.

Figure 24. A screenshot from the "Clinical Enrichment plot" subtab of the "Clinical enrichment" tab. The height of the barplots shows the fraction of samples with the specified gene mutations, and the colors represent each level of the chosen clinical column. The numerator in each representation indicates the number of samples with mutation meanwhile, the denominator indicates the number of samples in each level.

4.3 "Comparison with TCGA" Tab

This tab provides the users with a chance to compare their dataset with the TCGA-BRCA dataset on the level of differentially mutated genes and pathways, driver genes, mutational signatures, and somatic interactions.

4.3.1 "Differentially Mutated Genes" Subtab

In this subtab, users are first prompted to provide a VAF cutoff to apply to the dataset. Afterward, a Fisher's exact test based analysis is run and the results are provided first in the tabular format, under the "Fisher's exact test results" subtab, and second as a forest plot under the "Forest Plot" subtab. The user is free to choose an arbitrary p-value cutoff and the aesthetics for the forest plot. Both table and the forest plot include download buttons (**Figures 25, 26**). **Figure 25** shows the result of the Fisher's exact test in tabular format, from which it can be seen that the example dataset has lower mutation rates for *TTN*, *MUC16*, *FLG*, *RYR2*, and *HMCN1* genes. **Figure 26** presents the same results but in a forest plot format.

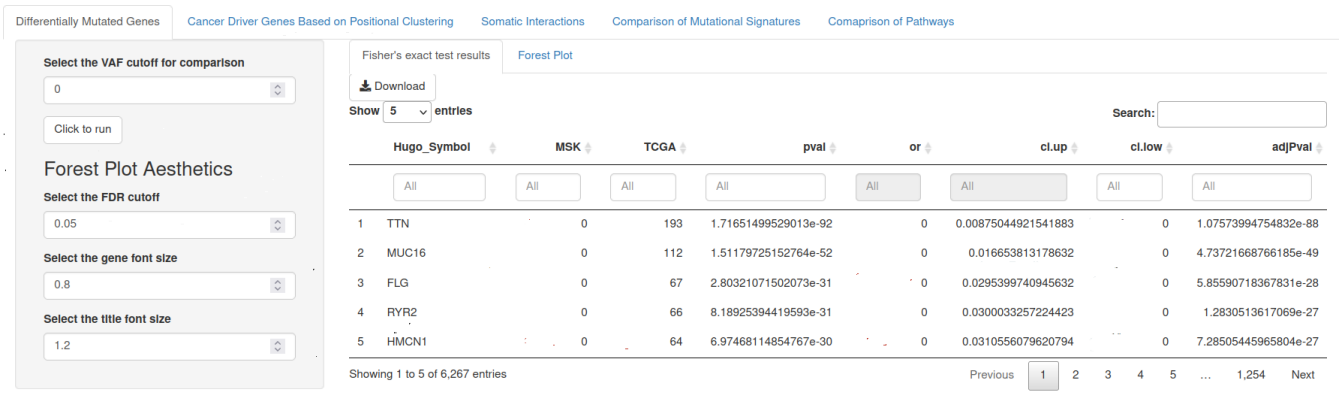


Figure 25. A screenshot from the "Fisher's exact test results" subtab of the "Differentially Mutated Genes" tab.

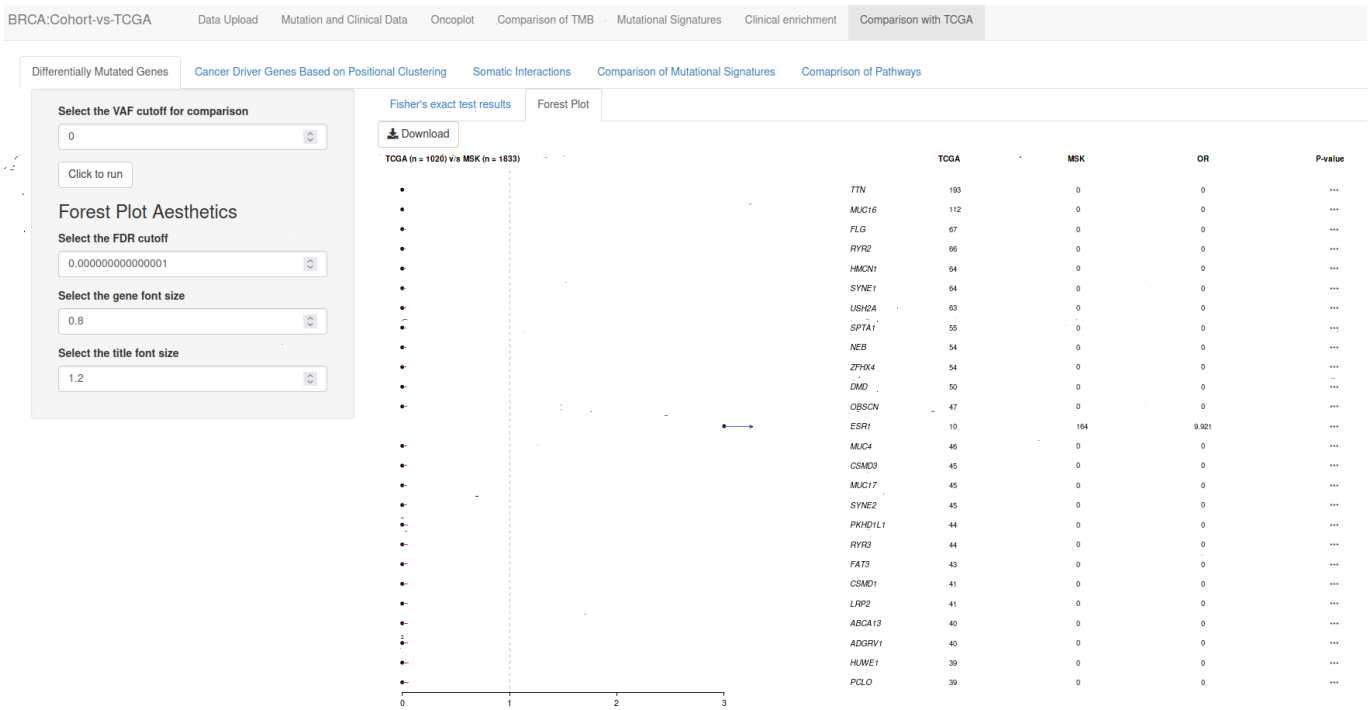


Figure 26. A screenshot from the “Forest Plot” subtab of the “Differentially Mutated Genes” tab.

4.3.2 “Cancer Driver Genes Based on Positional Clustering” Subtab

In this subtab, OncodriveCLUST analysis is carried out for each dataset and presented in tabular formats in the “Uploaded Cohort OncoCLUST results” and “TCGA Cohort OncoCLUST results” subtabs. A comparison of the FDRs for each gene for each cohort is presented as a scatter in the “OncoCLUST plot” tab. Finally, the “Co-lollipop” subtab allows users to get the mutation data of each data overlayed on the linear protein domain plot. Aesthetics can be modified for both the OncoCLUST plot and co-lollipop plot. Tables and the plots are downloadable (**Figure 27-30**). Analysis of the example dataset revealed that *KRAS* and *AKT1* were common driver genes between the example dataset (**Figure 27**) and TCGA-BRCA (**Figure 28**) meanwhile, *ESR1* was a unique driver gene in the example dataset. **Figure 30** overlays the *ESR1* mutations of the two datasets over a protein structure.

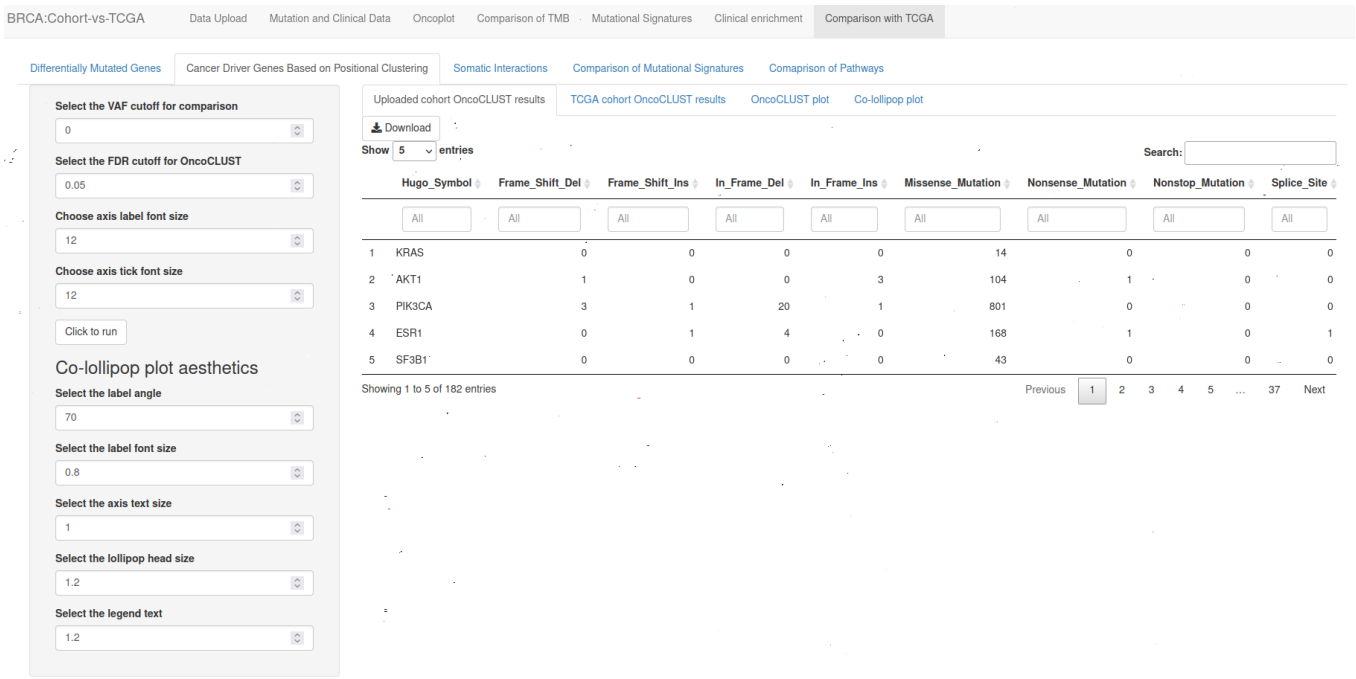


Figure 27. A screenshot from the “Uploaded Cohort OncoCLUST results” subtab of the “Cancer Driver Genes Based on Positional Clustering” tab.

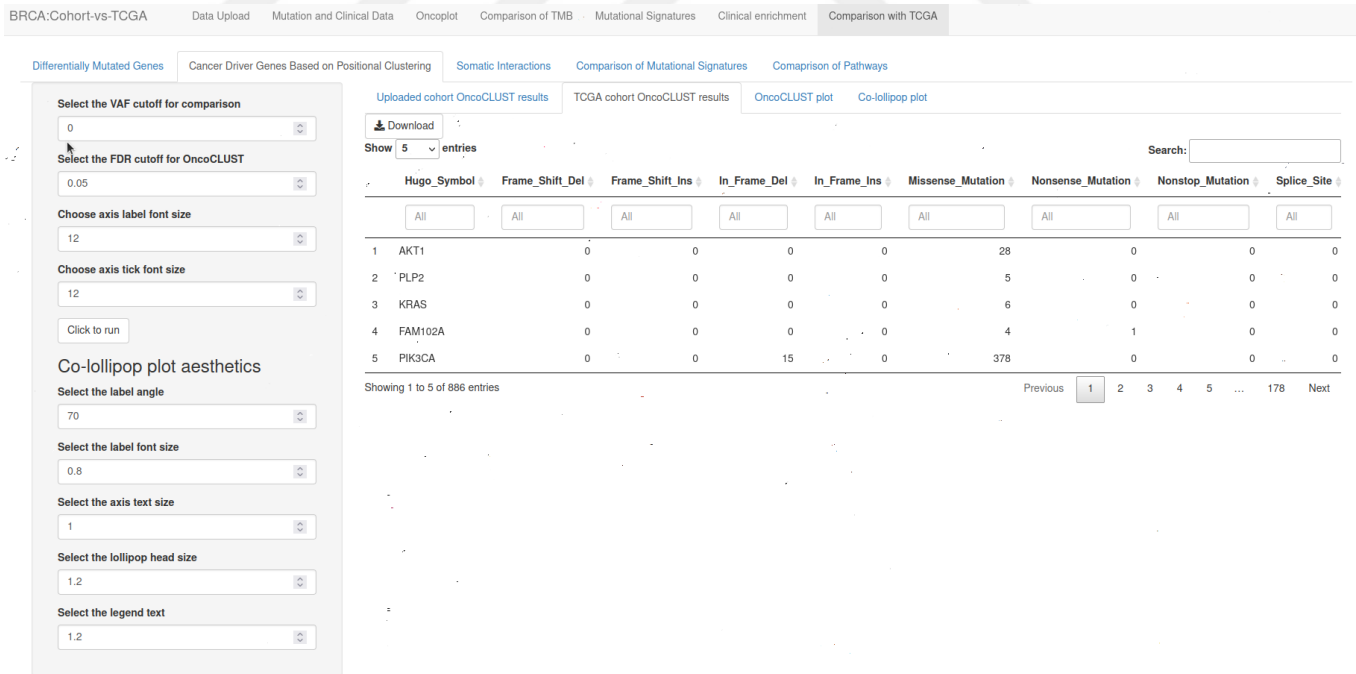


Figure 28. A screenshot from the “TCGA Cohort OncoCLUST results” subtab of the “Cancer Driver Genes Based on Positional Clustering” tab.

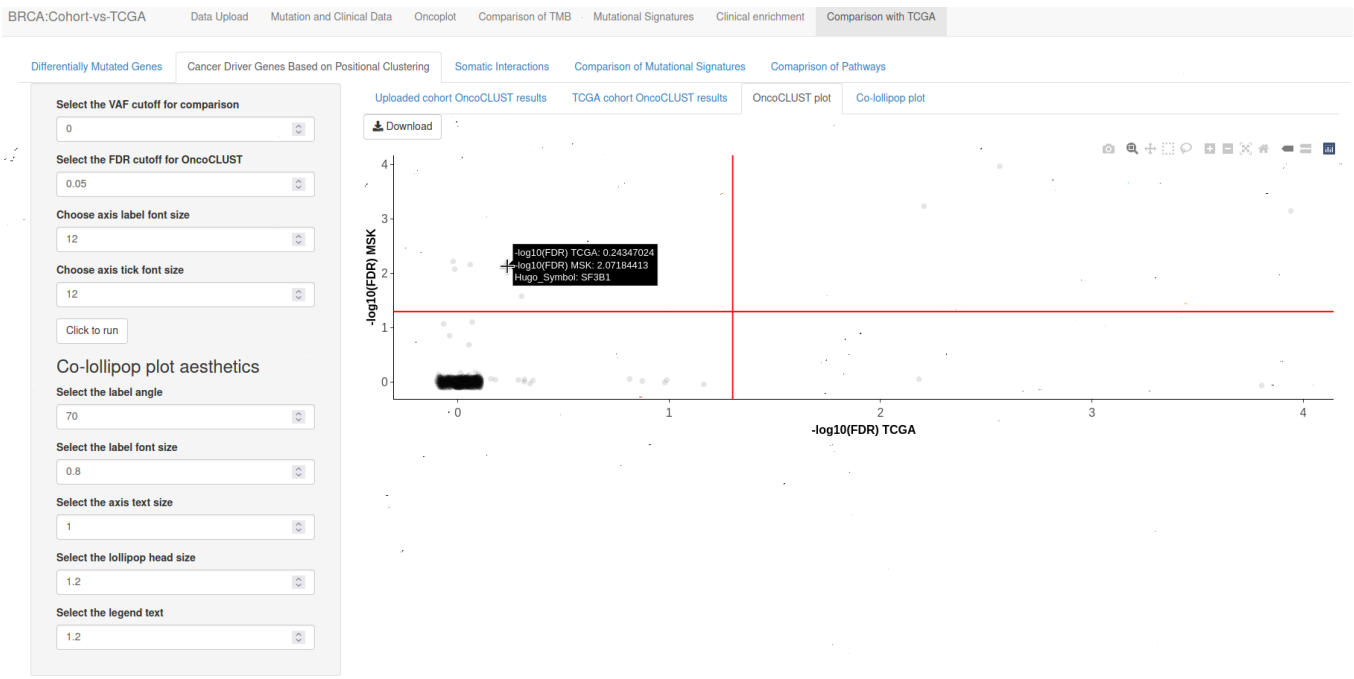


Figure 29. A screenshot from the “OncoCLUST plot” subtab of the “Cancer Driver Genes Based on Positional Clustering” tab.

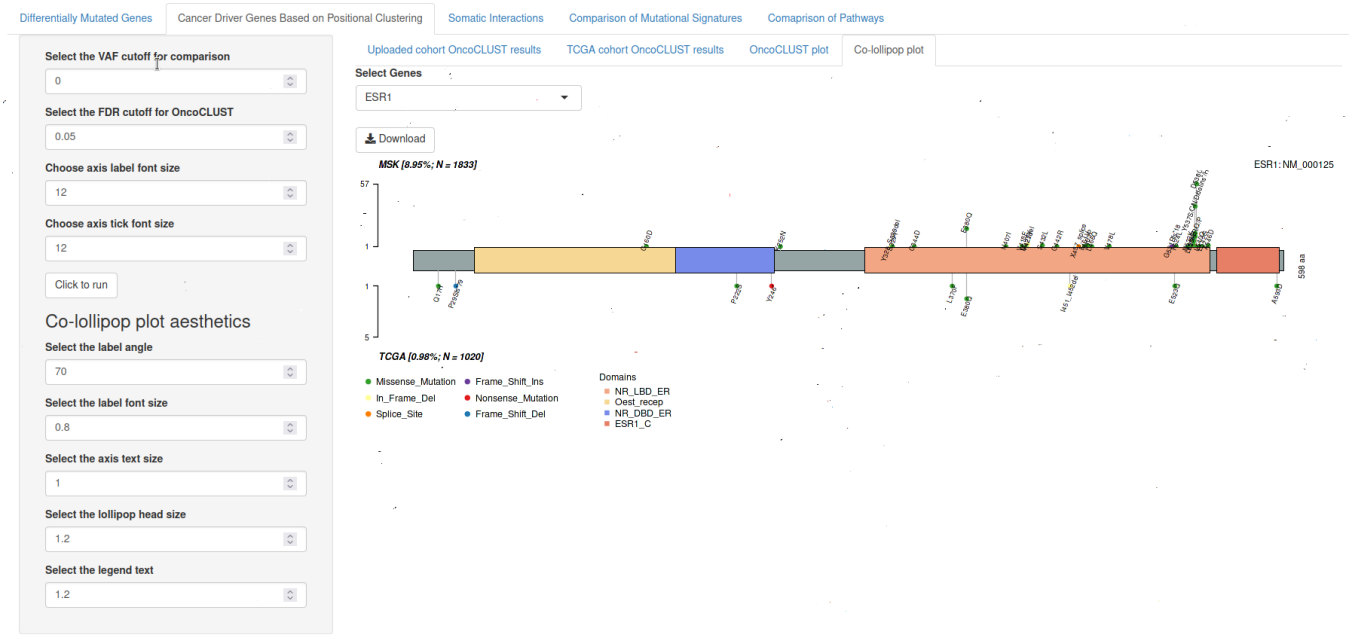


Figure 30. A screenshot from the “Co-lollipop plot” subtab of the “Cancer Driver Genes Based on Positional Clustering” tab.

4.3.3 “Somatic Interactions” Subtab

“Somatic Interactions” subtab conducts mutual exclusivity and co-occurrence analysis on the dataset for which the user decides on the VAF cutoff. Fisher’s exact test results are presented in a

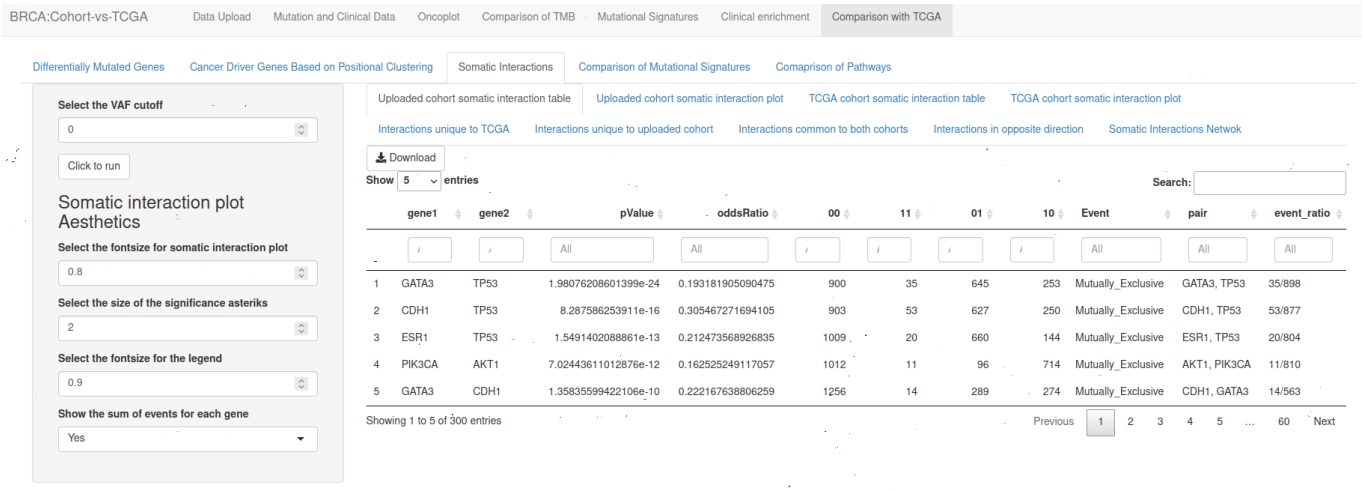


Figure 31. A screenshot from the “Uploaded cohort somatic interaction table” subtab of the “Somatic Interactions” tab.

tabular format in the “Uploaded cohort somatic interaction table” and “TCGA cohort somatic interaction table” for the uploaded and TCGA cohorts, respectively. The user has the option to download both tables (**Figure 31**). The example cohort was identified to have an interesting, mutually exclusive pair of genes, namely, *ESR1* and *TP53* (**Figure 31**).

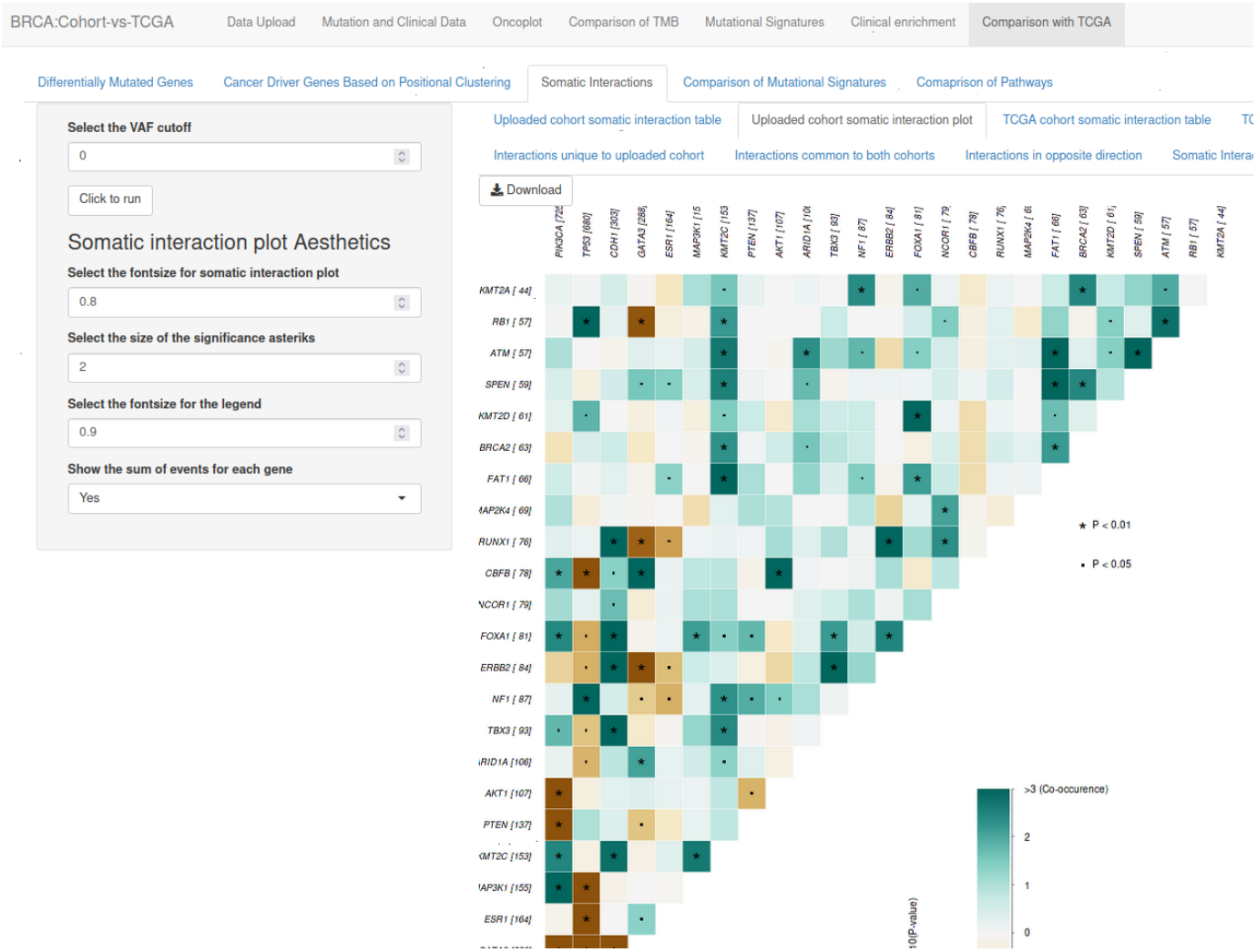


Figure 32. A screenshot from the "Uploaded cohort somatic interaction plot" subtab of the "Somatic Interactions" tab. "*" denotes $p < 0.01$ and "." denotes $p < 0.05$. Color intensities of cells increase according to the log10(p-value), with co-occurring genes colored with shades of blue and mutually exclusive genes colored with shades of brown.

The results are also visualized as an upper triangular heatmap in which the color represents the OR, a dot represents a p-value less than 0.05, and an asterisk represents a p-value less than 0.01. User has the option to modify different aesthetic parameters and to download the resultant plots (**Figure 32**). In the case of the example dataset one can easily observe highly mutually exclusive genes of *GATA3*, *PIK3CA*, *TP53*, and *CDH1* (**Figure 32**)

In addition to the tabular representation of the common and unique interactions in each dataset, the app also provides a network view, in which the nodes represent the genes and edges represent somatic interactions. Blue edges represent co-occurrence meanwhile, red edges denote mutual exclusivity. Datasets are coded by the shape of the edges, the TCGA cohort is represented via the solid edges, and the uploaded cohort is represented via the dashed edges (**Figure 33**).

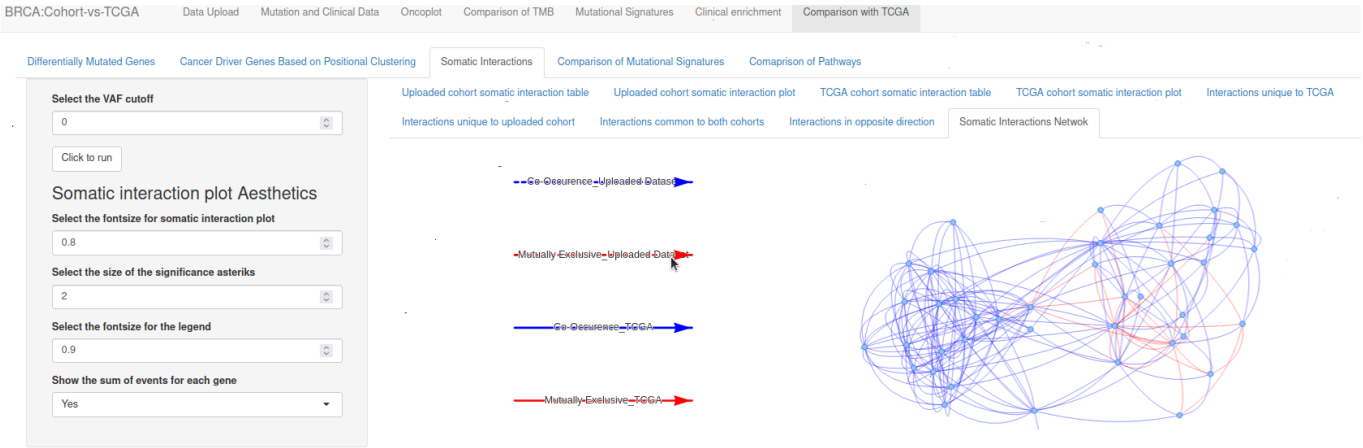


Figure 33. A screenshot from the "Somatic Interaction Network" subtab of the "Somatic Interactions" tab.

Additionally, there is a tab called "Somatic Interactions with Custom Gene" which performs the fisher's exact test analysis for a gene selected by the user against all other available genes in both uploaded and TCGA datasets. This tab also presents the unique and common interactions identified in each dataset. For example, the *CHRNA5* was found to be co-occurring with the *UTP20* gene with an OR of 40.69 in the TCGA-BRCA cohort (**Figure 34**.)

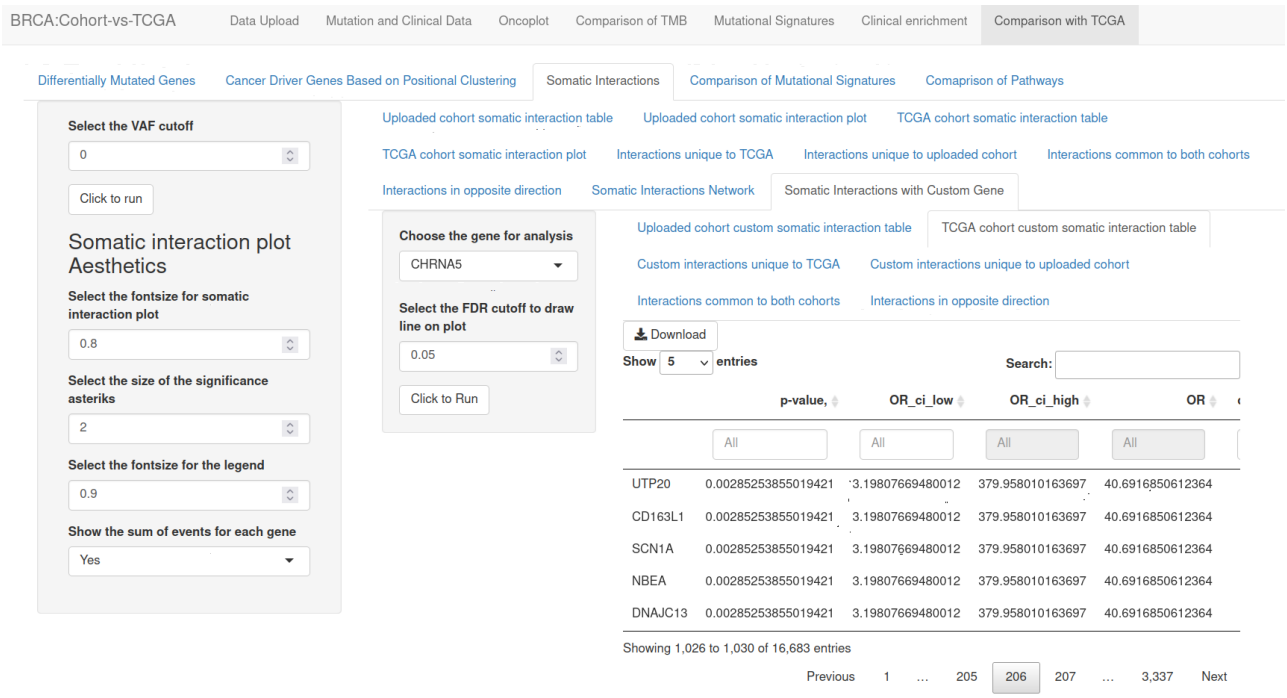


Figure 34. A screenshot from the "Somatic Interactions with Custom Gene" subtab of the "Somatic Interactions" tab.

4.3.4 "Comparison of Mutational Signatures" Subtab

This subtab first represents the fitted signatures and cophenetic plot of the uploaded cohort, followed by the precomputed results for the TCGA cohort. Afterward, the similarity of each signature in one dataset to every other signature in the other dataset is calculated via cosine similarity metric and presented as a heatmap. The user has the option to download every plot generated (**Figure 35-37**). **Figure 35** shows the cophenetic plot of the TCGA dataset. The best number of signatures was determined to be seven, the point at which a "knee" occurs. Fitted signatures of the TCGA-BRCA cohort and their cosine similarities to COSMIC signatures are depicted in **Figure 36**. **Figure 37** shows the hierarchical clustering of the matrix of cosine similarities between TCGA-BRCA and the example cohort. It was observed that signatures 2 and 3 of the example cohort have low similarity to the TCGA-BRCA signatures (**Figure 37**).

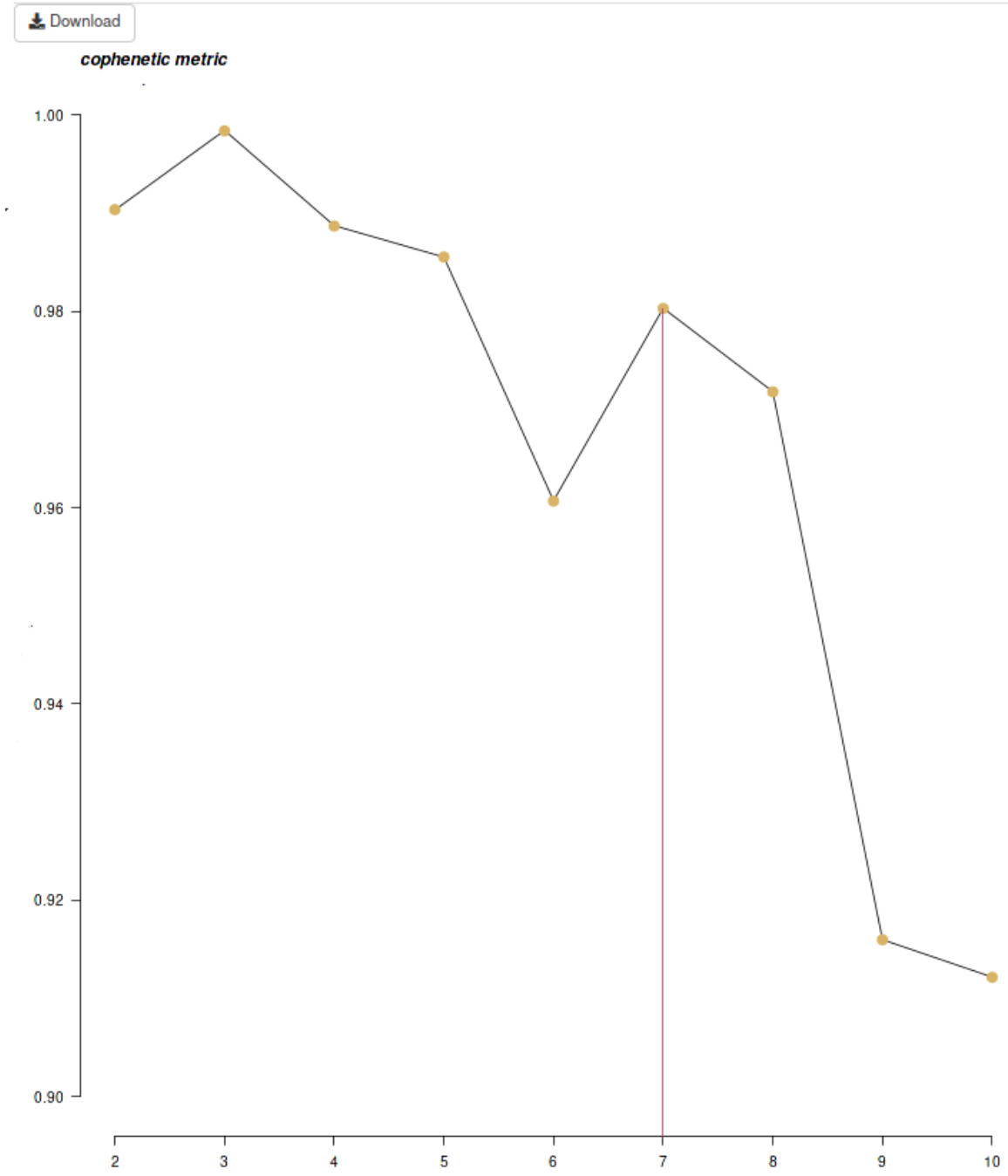

Download
cophenetic metric
1.00
0.98
0.96
0.94
0.92
0.90
2
3
4
5
6
7
8
9
10

Figure 35. A screenshot from the "Cophenetic Correlation vs Signature Number, TCGA cohort" subtab of the "Comparison of Mutational Signatures" tab. The vertical line denotes the optimal number of signatures.

Figure 36. A screenshot from the “Identified signature and their cosine similarity to SBS signatures, TCGA cohort” subtab of the “Comparison of Mutational Signatures” tab.

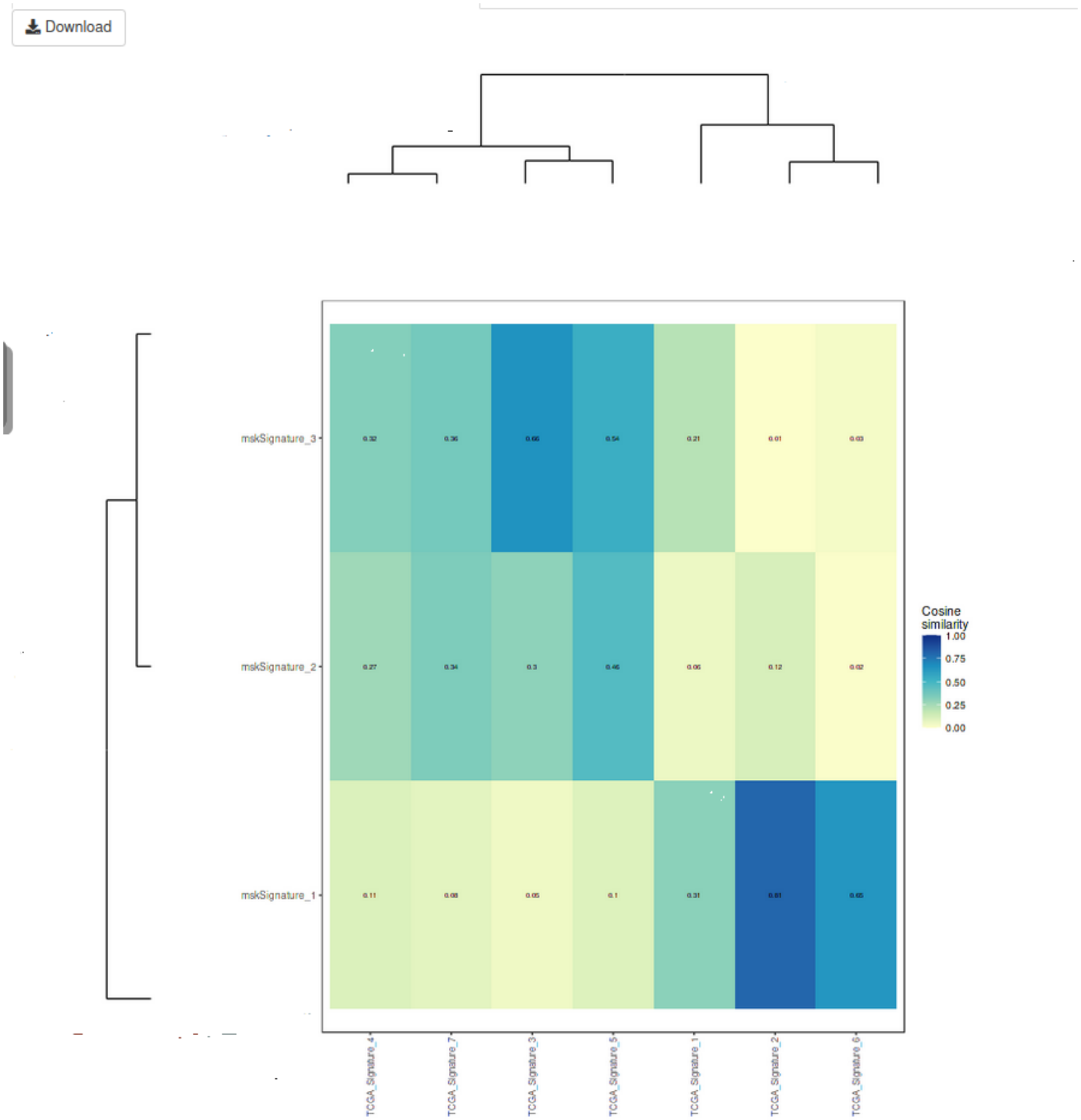


Figure 37. A screenshot from the “Cosine similarity between TCGA and Uploaded cohort signatures” subtab of the “Comparison of Mutational Signatures” tab.

4.3.5 “Comparison of Pathways” Subtab.

This subtab is composed of three subtabs. The first subtab, “Summary” subtab, allows the user first to select a VAF cutoff. Afterward, the user is prompted to select a dataset, namely TCGA or the uploaded dataset. This subtab then summarizes the number of affected genes and affected individuals for each pathway in a tabular form or as a barplot (**Figure 38, 39**).

The next subtab, “Comparison” subtab, performs a Fisher’s exact test based differentially mutated pathway analysis, and presents the results in tabular as well as the forest plot form (**Figure 40, 41**)

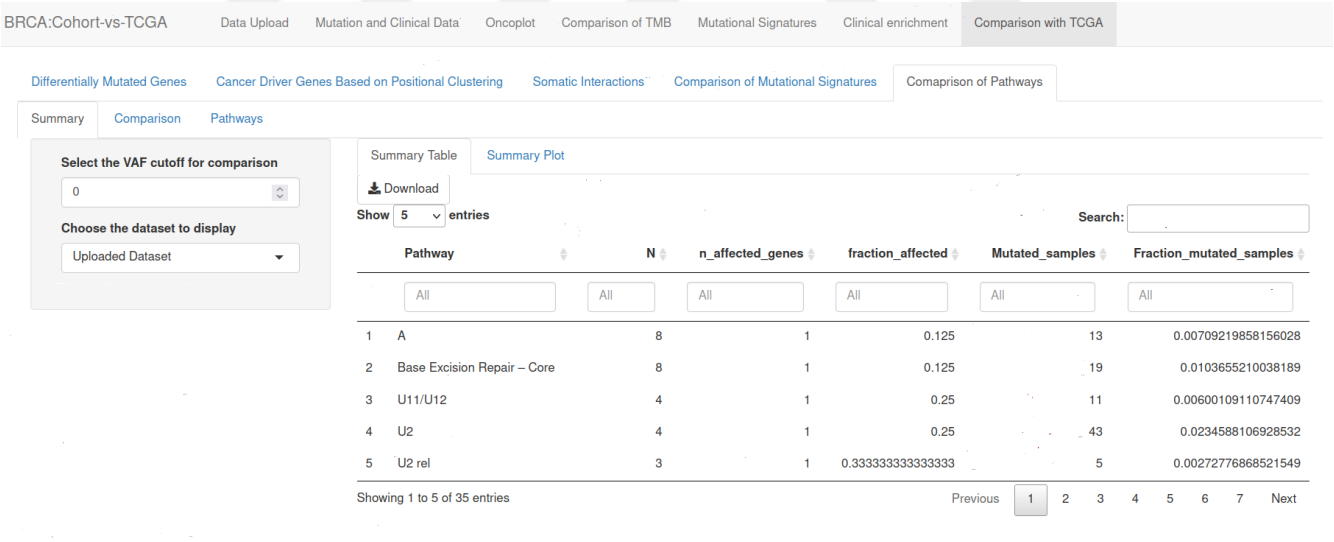


Figure 38. A screenshot from the “Summary Table” subtab of the “Comparison of Pathways” tab.

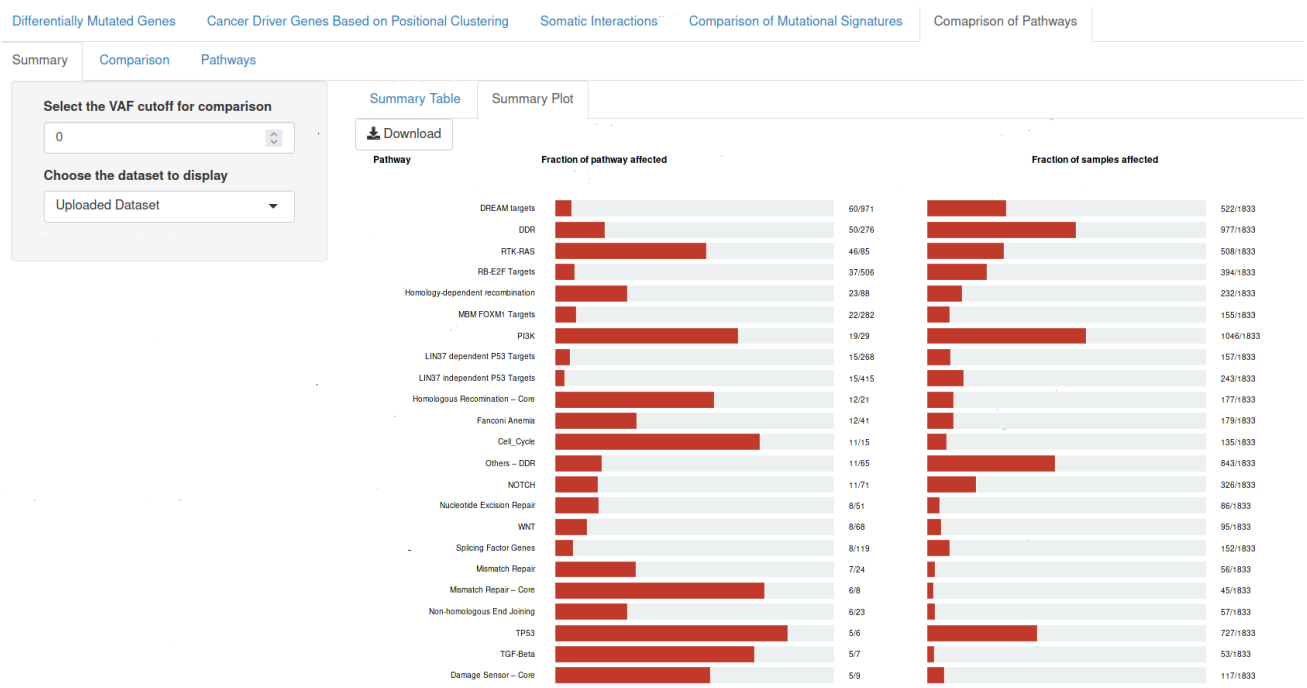
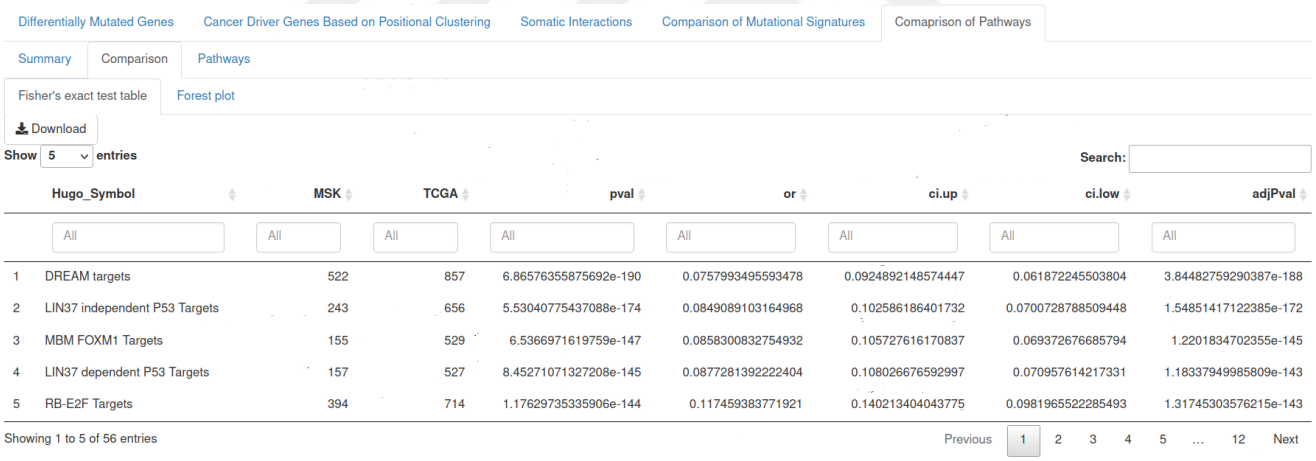


Figure 39. A screenshot from the “Summary Plot” subtab of the “Comparison of Pathways” tab.



	Hugo_Symbol	MSK	TCGA	pval	or	ci.up	ci.low	adjPval
	All	All	All	All	All	All	All	All
1	DREAM targets	522	857	6.86576355875692e-190	0.0757993495593478	0.0924892148574447	0.061872245503804	3.84482759290387e-188
2	LIN37 independent P53 Targets	243	656	5.53040775437088e-174	0.0849089103164968	0.102586186401732	0.0700728788509448	1.54851417122385e-172
3	MBM FOXM1 Targets	155	529	6.5366971619759e-147	0.0858300832754932	0.105727616170837	0.069372676685794	1.2201834702355e-145
4	LIN37 dependent P53 Targets	157	527	8.45271071327208e-145	0.0877281392222404	0.108026676592997	0.070957614217331	1.18337949985809e-143
5	RB-E2F Targets	394	714	1.17629735335906e-144	0.117459383771921	0.140213404043775	0.0981965522285493	1.31745303576215e-143

Figure 40. A screenshot from the “Fisher’s exact test table” subtab of the “Comparison of Pathways” tab.

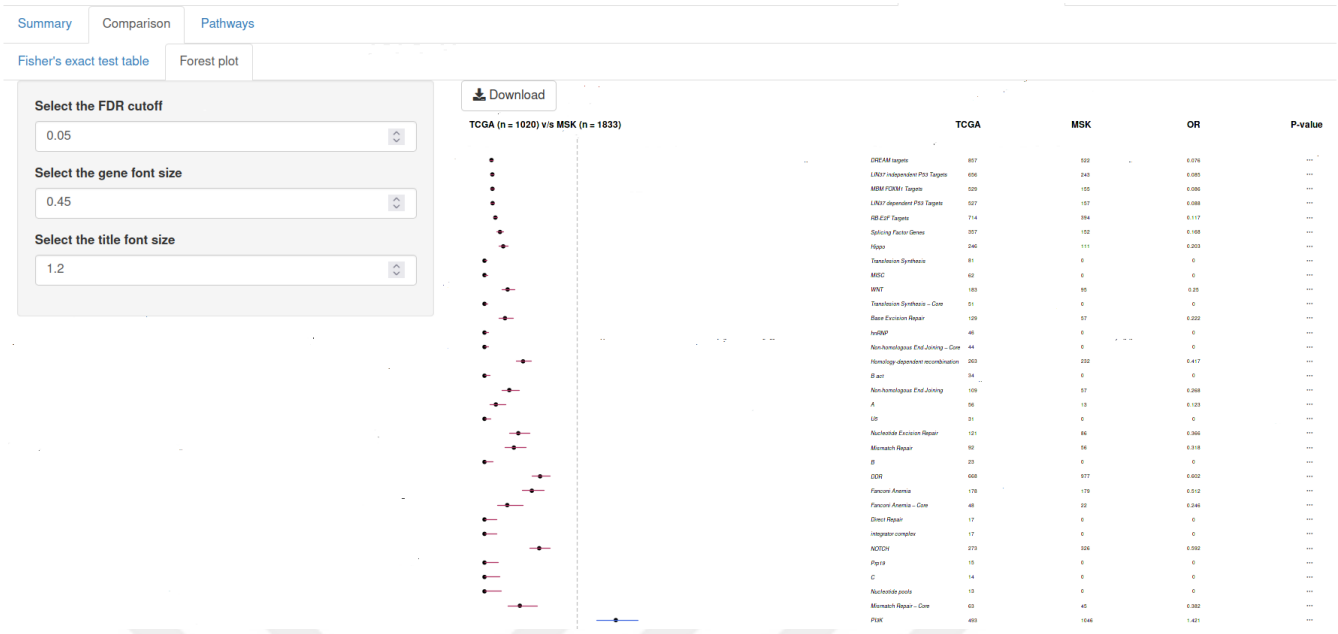


Figure 41. A screenshot from the "Fisher's exact test table" subtab of the "Comparison of Pathways" tab.

Additionally, the user is presented with the functionality to create an oncoplot from the components of the pathway of interest (**Figure 42**).

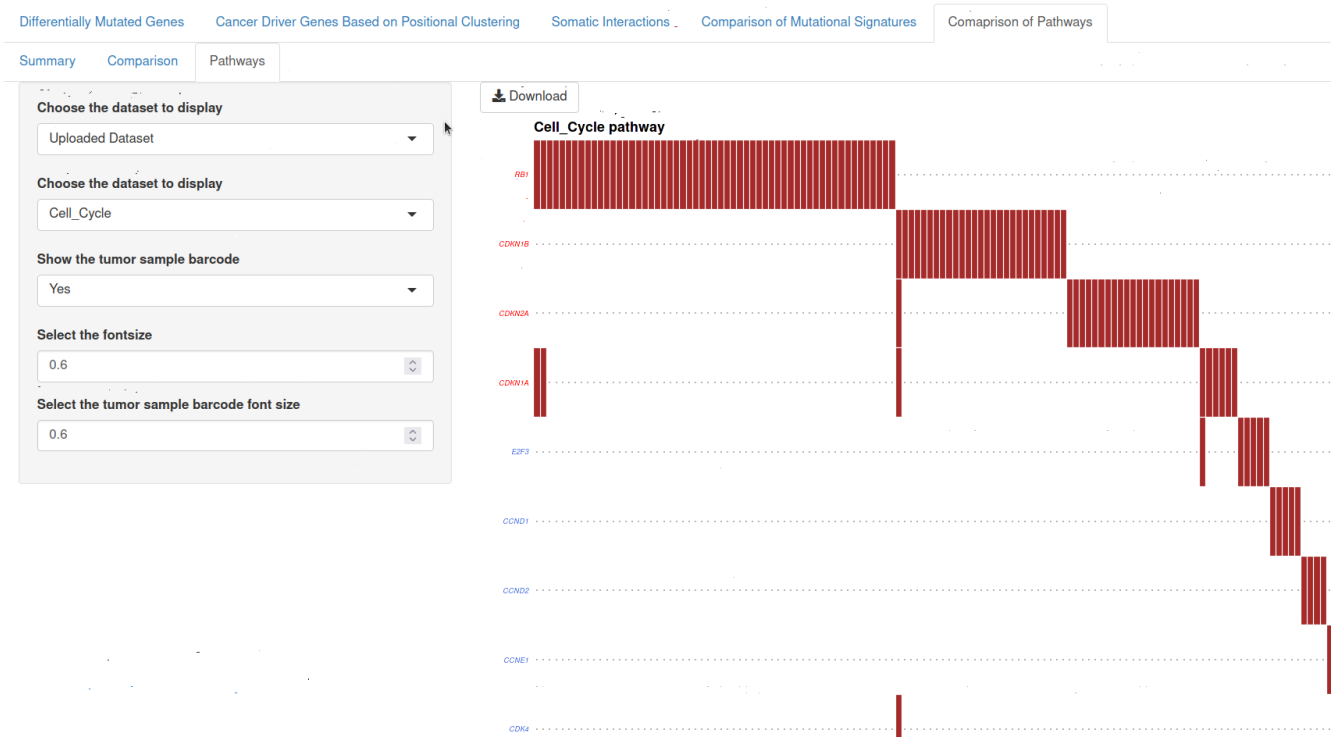


Figure 42. A screenshot from the "Pathways" subtab of the "Comparison of Pathways" tab.

4.4 Case Study of The Gustave Roussy Institute Cohort

4.4.1 Features of Cohort Used for The Case Study

This study has originally been conducted on samples from metastatic breast cancer patients who were enrolled in three different clinical trials. The study presents the whole-exome sequencing results for each patient, accompanied with the clinical information regarding hormone receptor (HR) status, HER2 status and location of the sample collected. The study was conducted by the Gustave Roussy Institute (IGR) thus, the dataset will be referred as IGR henceforth [127]. **Table 3 and 4** summarizes the clinical features of the dataset.

Table 3. Summary of the hormone receptor and HER2 status of the IGR cohort

	HER2+	HR-/HER-	HR+/HER2-	Not Determined
Count	14	51	140	8

Table 4. Summary of the biopsy locations of the IGR cohort

	Liver	Lung	Lymph Node	Other	Skin
Count	53	1	25	27	28

4.4.2 Data Processing Pipeline Used in The Original Study

Whole exome sequencing data was processed as follows and is the input for the case study conducted in this thesis. Mutect was used to identify substitutions and the Scalpel was used to identify small insertions and deletions (indels). Variants that were detected in at least one of the normal samples in the cohort or annotated as known

polymorphisms by 1000 Genomes or the ESP databases were removed unless they were also reported in COSMIC or ClinVar databases. Variants passing the following filters were considered for the subsequent analysis: VAF in the tumor > 0.1, number of the variant reads in the tumor ≥ 5, VAF in the normal < 0.02, number of variant reads in the normal ≤ 3; and total depth in normal and tumor samples ≥ 10. Please note that the same filters were not applied to the TCGA dataset in the original study. Additionally, the TCGA cohort used by the authors was small compared to the TCGA cohort used in this study because they chose to consider the samples generated by the Washington University School of Medicine only [127].

4.4.3 Overview of The Mutation Spectrum of The IGR Cohort

Figures 43 and **44** showed the oncoplots for the IGR and the TCGA datasets, respectively. Overall, the top 25 genes between the datasets were found to be fairly similar. 12 genes in the top 25 highly mutated genes in both datasets matched, namely, *TP53, PIK3CA, TTN, GATA3, RYR2, MAP3K1, CDH1, MUC16, HMCN1, OBSCN, FLG, NEB, RYR3, SYNE1, USH2A* and *KMT2C*. *ESR1, FSIP2, RYR1, DNAH6, XIRP2, NF1, PCLO, FRAS1,* and *TENM3* genes, on the other hand, were unique to the IGR dataset, whereas, *PTEN, SPTA1, ZFHX4, NCOR1, DMD, MUC4, SYNE2, CSMD3* and *MUC17* were unique to the TCGA dataset.

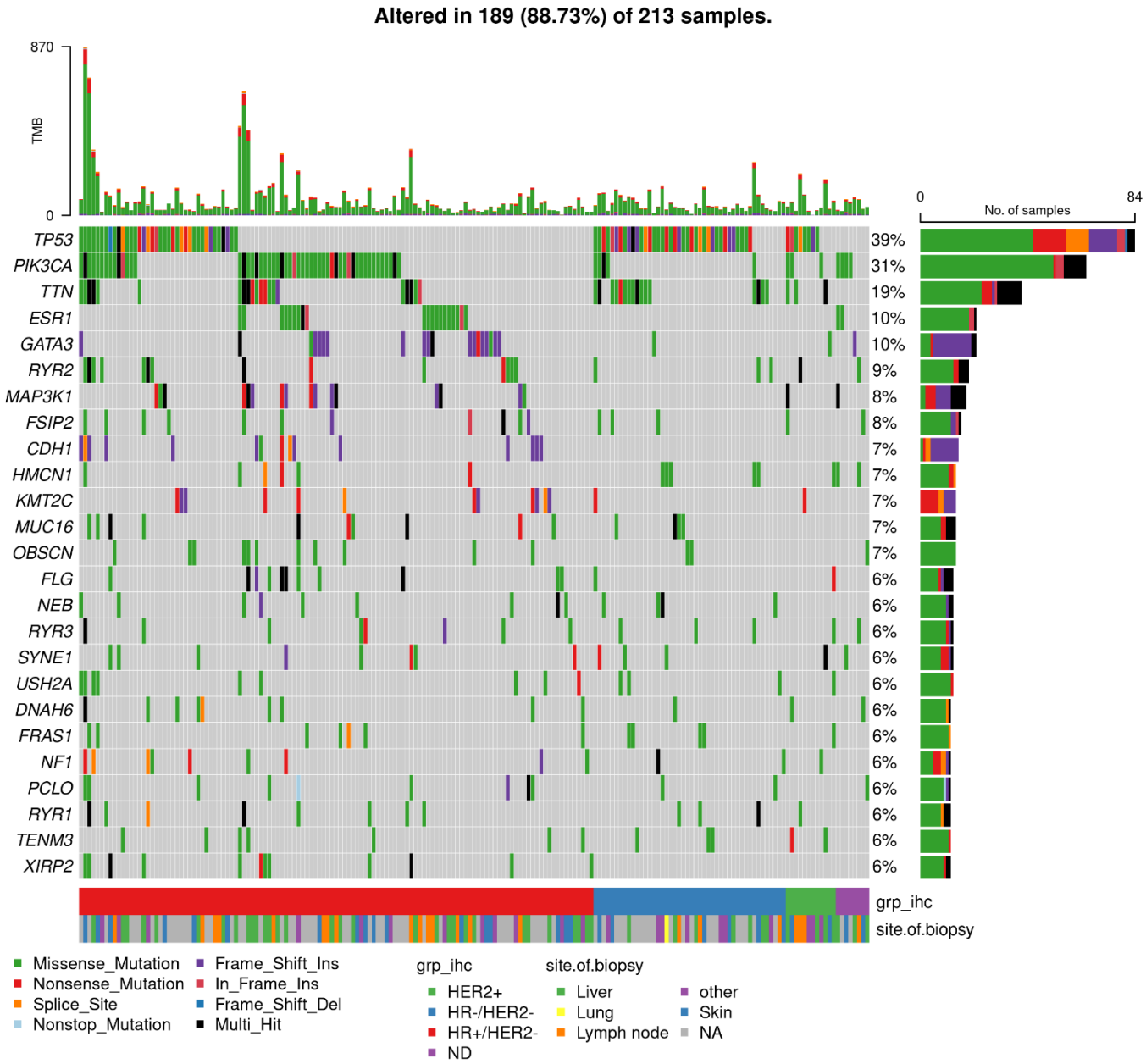


Figure 43. Oncoplot of the IGR cohort, summarizing the mutational profile. Samples are grouped by the HR and HER+ status. Top 25 highly mutated genes are shown.

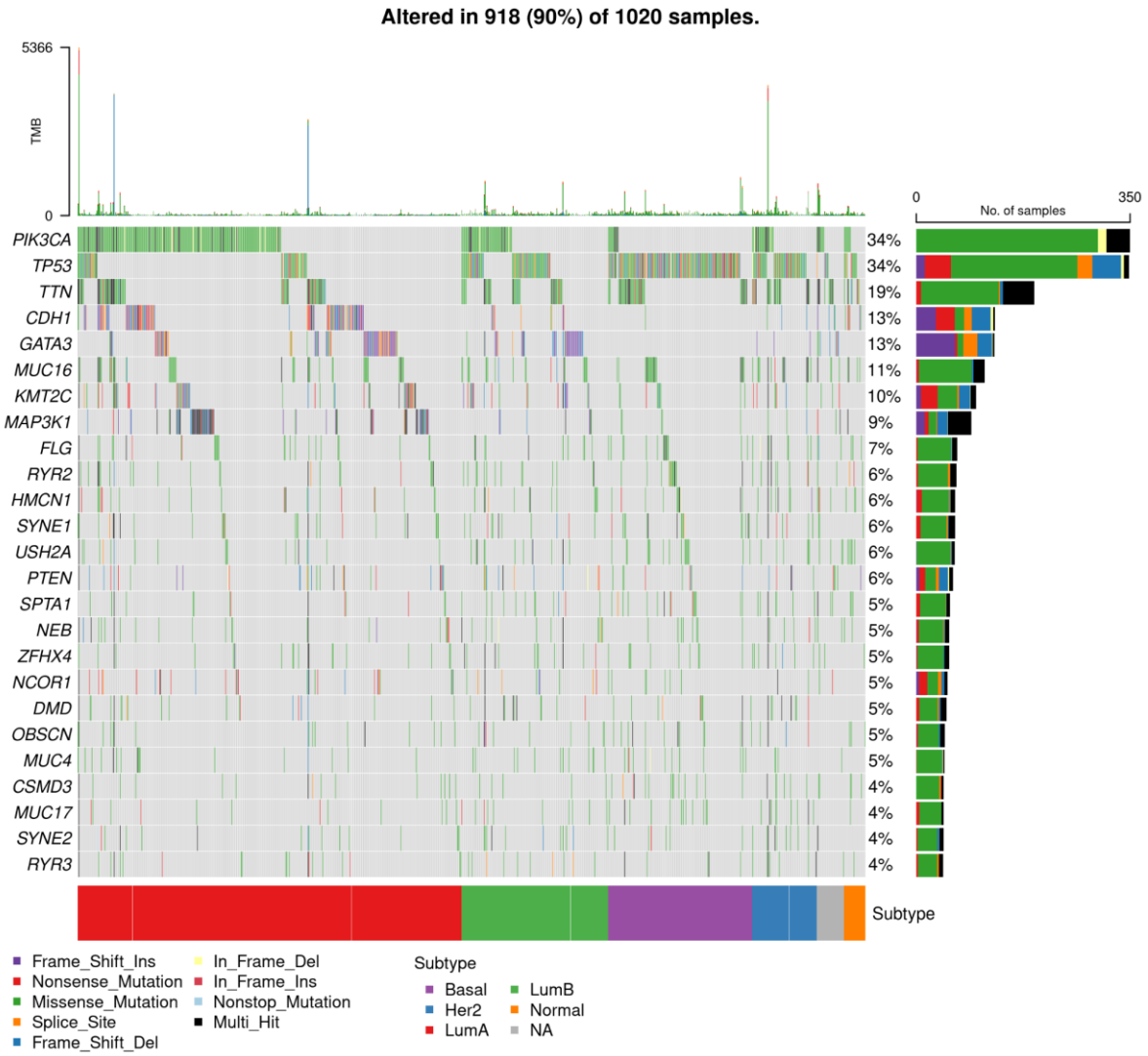


Figure 44. Oncoplot of the TCGA cohort, summarizing the mutational profile. Samples are grouped by the PAM50 subtype. Top 25 highly mutated genes are shown.

4.4.4 Determination of The Optimal VAF cutoff for The Cohort Comparison

The optimal VAF cutoff was determined by subsetting each dataset by the corresponding VAF cutoff and comparing their mean TMBs. Interestingly, no difference in TMB was observed until VAF 0.1, after which the difference becomes

significant until and including VAF 0.15. Setting this cutoff at VAF 0.2 will probably result in the loss of valuable sub-clonal mutations, hence VAF 0 was decided to be the optimal VAF cutoff. The significance of the VAF difference between 0.1 and 0.15 represents an interesting case which is discussed in detail in the discussion section (**Figure 45**). Hereon, VAF cutoff 0 was used for every analysis.

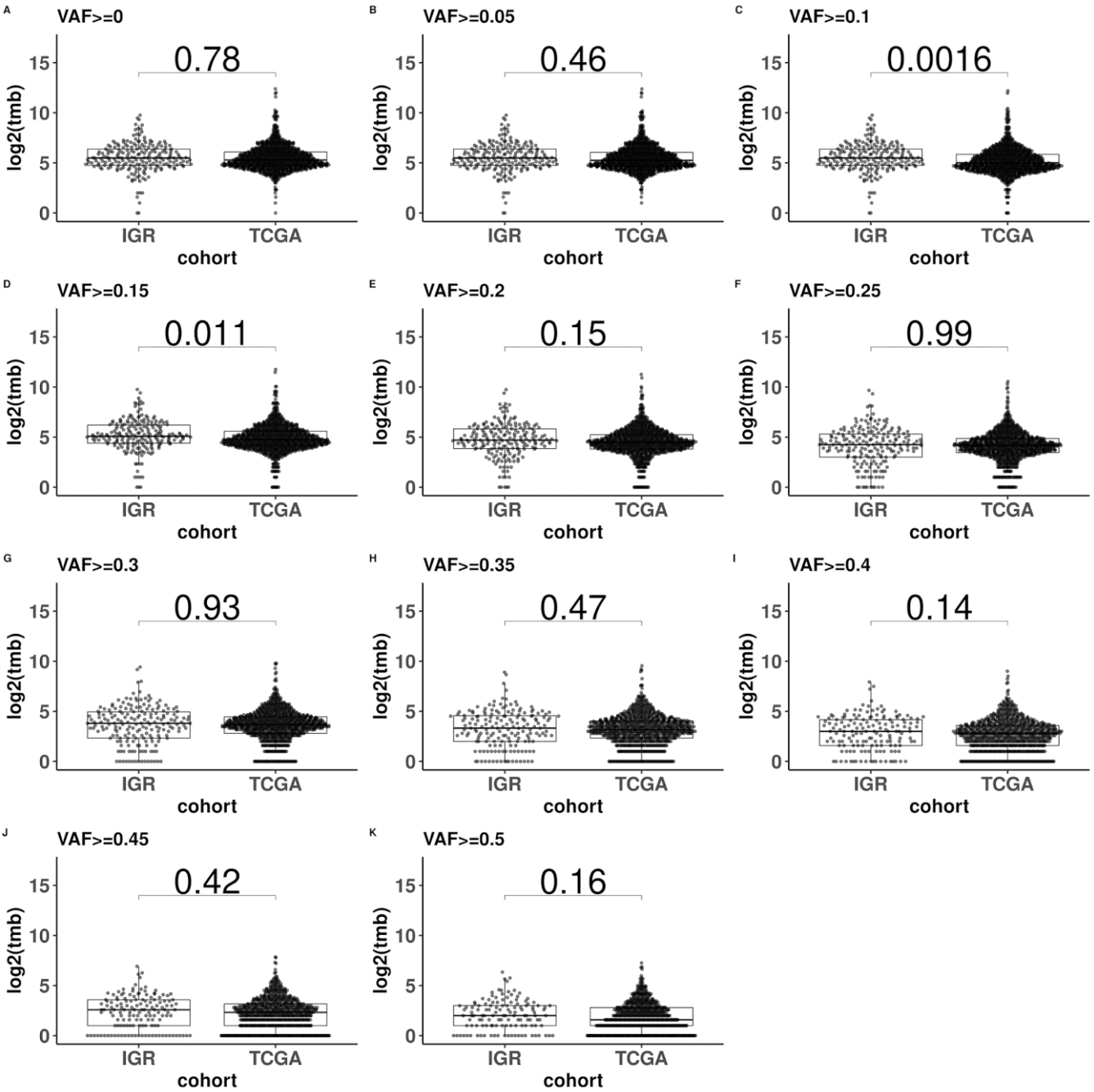


Figure 45. Comparison of TMB between the IGR cohort and the TCGA cohort. P-values are calculated by the Student's t-test. The upper and lower edges of the boxes represent 75th and 25th quantiles, respectively, meanwhile black rod represents the median. Whiskers extend until 1.5* interquartile range (IQR) if there are observations extremum to those points, and till the minimum and maximum values otherwise.

4.4.5 Mutational Signature Analysis

4.4.5.1 Detection of Optimal Number of Mutational Signatures

As mentioned before, I used the cophenetic correlation analysis, with 10 signatures to try and 5 bootstraps, to determine the optimal number of signatures for both TCGA and IGR datasets.

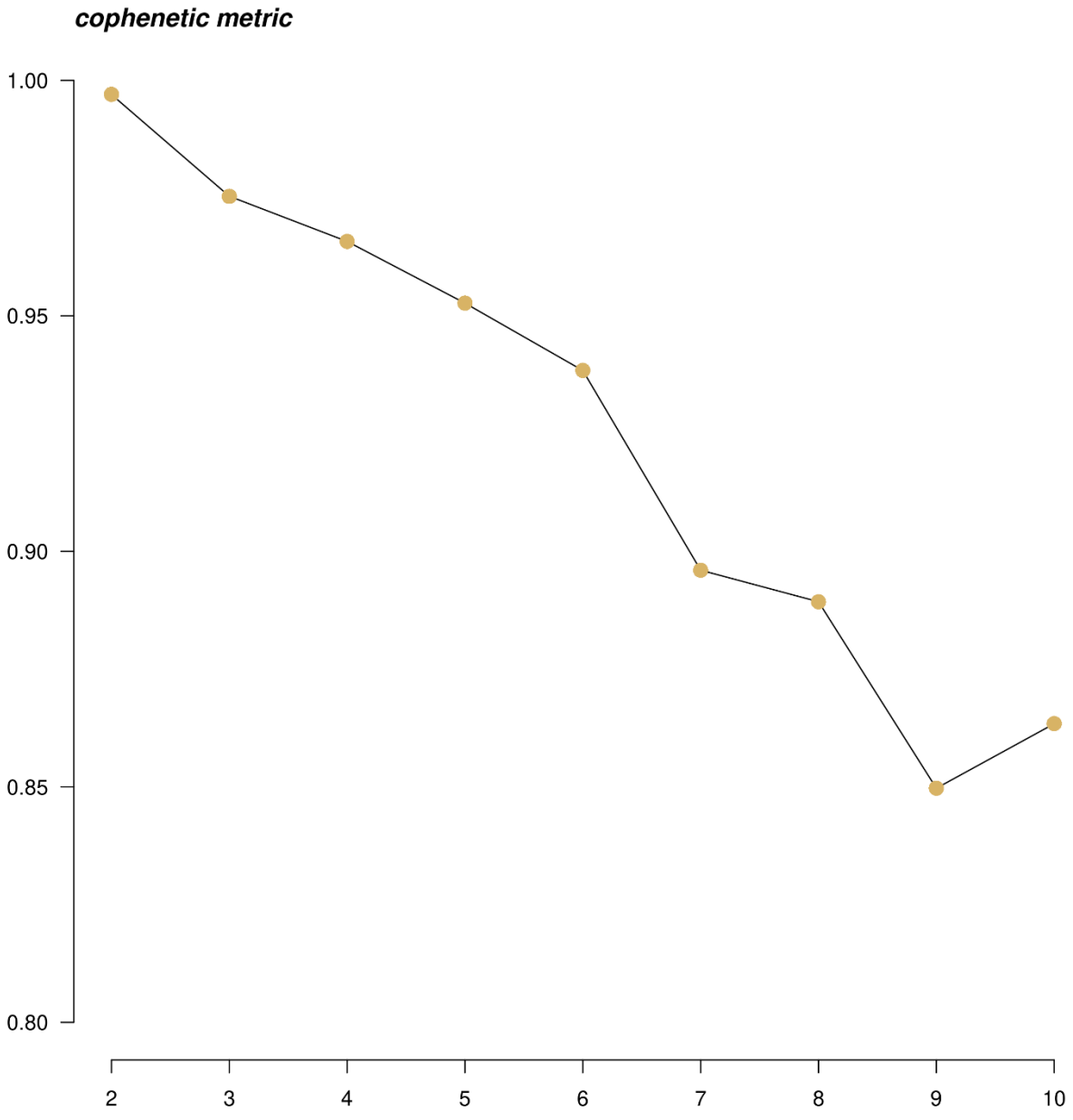


Figure 46. Cophenetic correlation analysis for the IGR dataset. The x-axis represents the number of signatures, y-axis represents the cophenetic correlation coefficient.

For the IGR dataset, cophenetic correlation displays a knee at the 6th signature, hence should be chosen as the optimal number of signatures to carry the analysis out (**Figure 46**). However, the original study chose to fit 5 signatures instead, thus I also used the 5 signatures to make the results more comparable. The same analysis was also carried out

on the TCGA cohort and the results are presented in **Figure 47**. I observed a “knee” at 7, as result, I fitted 7 signatures to the data.

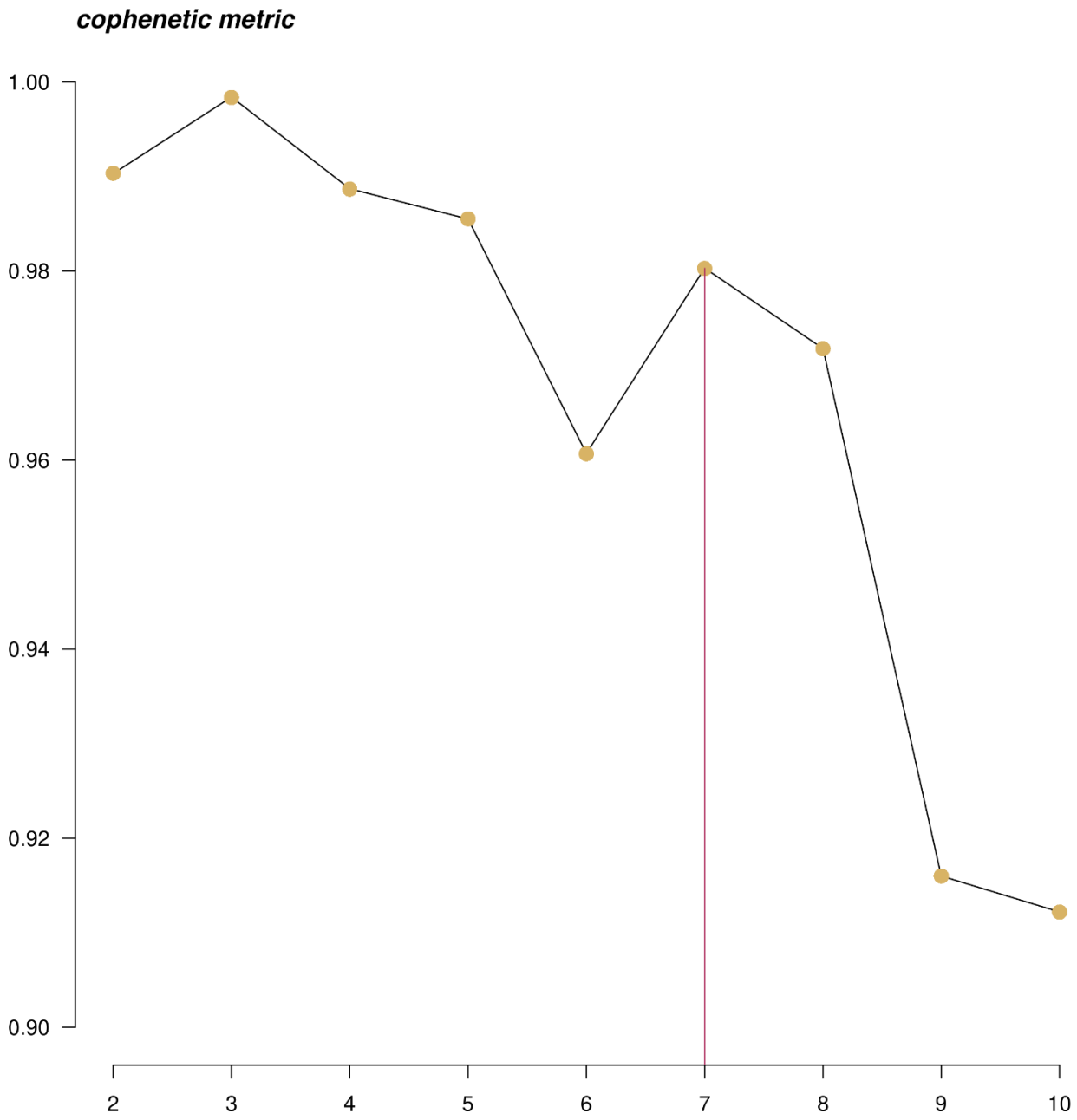


Figure 47. Cophenetic correlation analysis for the IGR dataset. the x-axis represents the number of signatures, y-axis represents the cophenetic correlation coefficient. The vertical line represents the chosen number of signatures to fit.

4.4.5.2 Fitting of Signatures

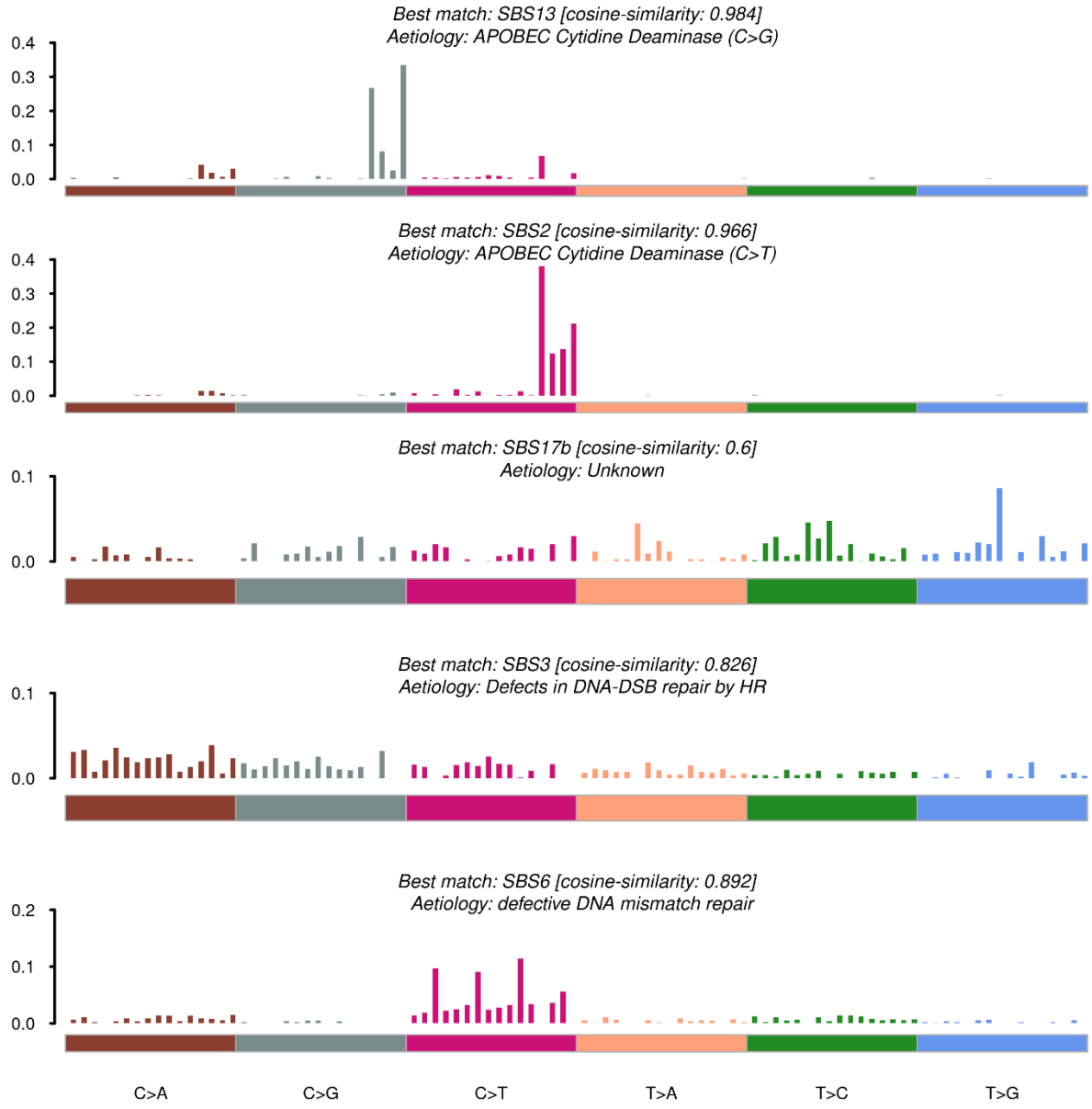


Figure 48. Signatures identified in the IGR cohort. The contribution of each SBS context to each signature is represented by the height of the barplots, which add up to 1. The title of each mutational signature includes the code of the best-matching COSMIC SBS signature, its name, and its cosine similarity to the SBS signature.

Figure 48 shows the signatures identified in the IGR dataset. Overall, the quality of the high analysis was supported by the high cosine similarity of our *de novo* identified signatures to the SBS signatures present in the COSMIC database. Signatures 1 and 2

correspond to a APOBEC related etiology, Signature 4 is matched best to a defective double-strand break (DSB) repair related signature, and Signature 5 represents a defective DNA mismatch repair related etiology. Signature 3 is reminiscent of SBS17b, which has unknown etiology. A more thorough examination of this signature in the COSMIC website however indicated that signature might be related to reactive oxygen species exposure or 5FU treatment. The quality of this annotation is low hence it is annotated to be caused by an unknown etiology.

The same analysis in the TCGA cohort revealed similar signatures, though not exactly the same. Two signatures caused by APOBEC deaminase activities, one caused by defective DNA mismatch repair, and another caused by defective DNA DSB repair were identified and these had their counterparts in the IGR dataset as well. On the other hand, defective polymerase epsilon (POLE) related signature, spontaneous 5-methylcytosine deamination related signature and a signature of unknown etiology were unique to the TCGA dataset. Signature 3 of IGR dataset did not have any similar signature in the TCGA dataset (**Figure 49**).

4.4.5.3 Comparison of the Signatures

I have given a qualitative comparison of the signatures in the previous section. A quantitative approach was taken in this section by using the cosine similarity metric. Overall, the results were in agreement with the previous section. APOBEC related signatures were perfectly similar to each other with a cosine similarity of 1. The cosine similarity of the DNA damage related signatures was around 0.9, implying an excellent agreement. Signature 3 of the IGR dataset on the other hand

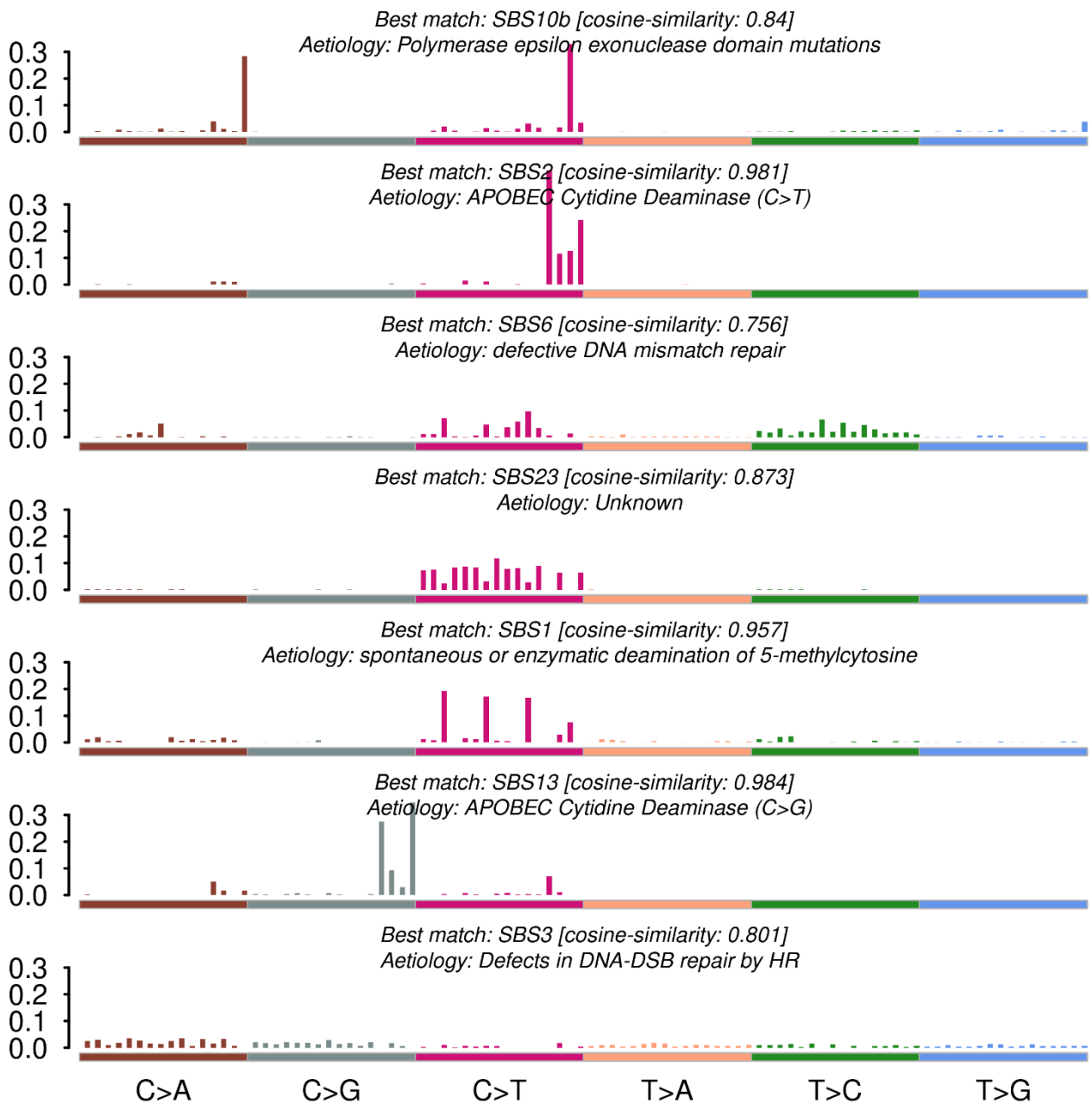


Figure 49. Signatures identified in the TCGA cohort. Contribution of each SBS context to each signature is represented by the height of the barplots, which add up to 1. Title of each mutational signature includes the code of the best-matching COSMIC SBS signature, its name, and its similarity to the signature in cosine similarity.

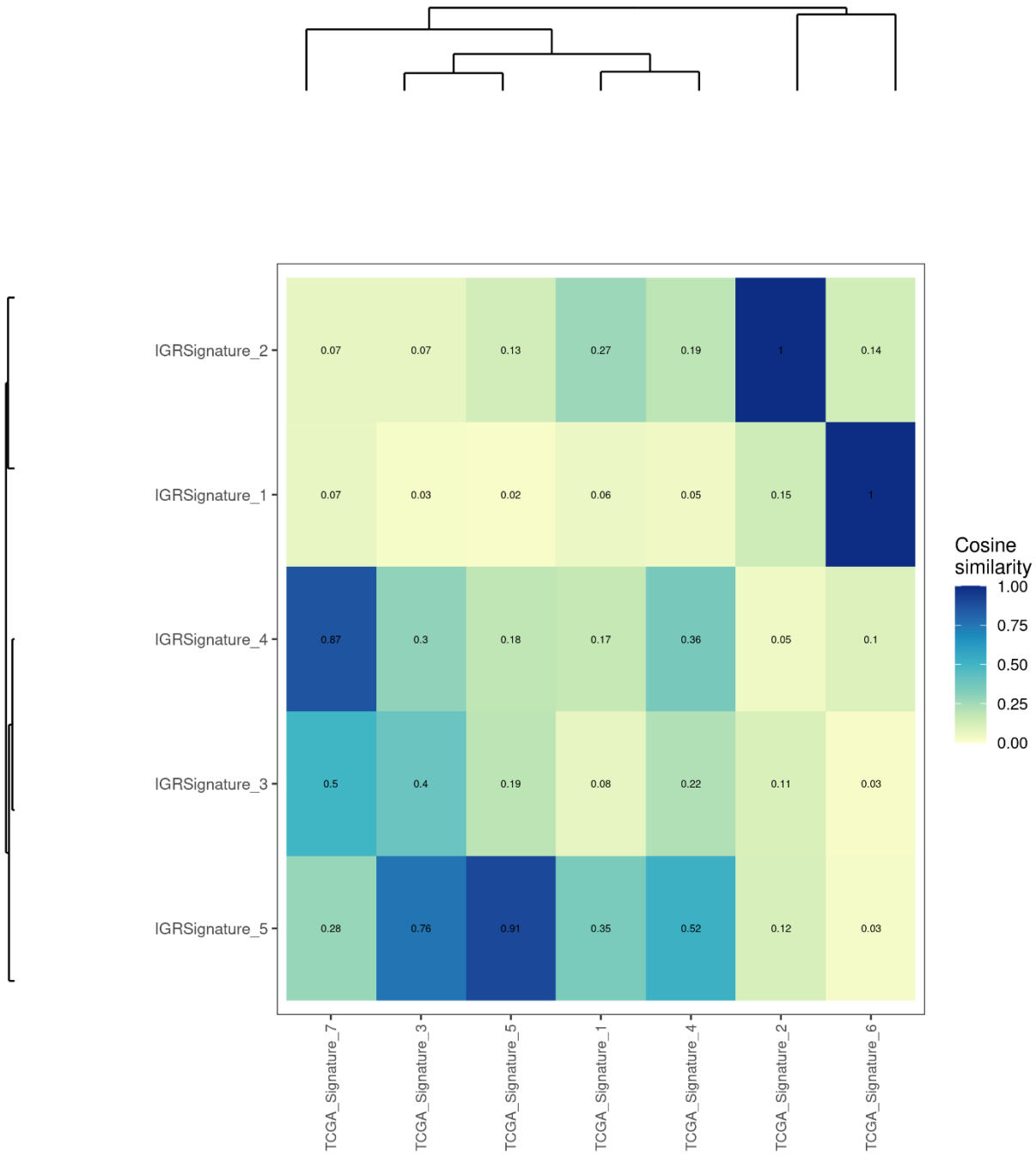


Figure 50. Cosine similarity matrix of the signatures identified in IGR and TCGA datasets. Matrix is clustered using Euclidian distance and average linkage.

had a maximum cosine similarity of 0.5, with the signature 7 of the TCGA dataset. POLE related signature and the signature of unknown etiology (SBS23) of the TCGA

dataset were also poorly related to the signature of the IGR dataset, with a cosine similarity of 0.35 and 0.52, respectively (**Figure 50**).

4.4.5.4 Contribution of Signatures to Each Patient in The IGR Cohort.

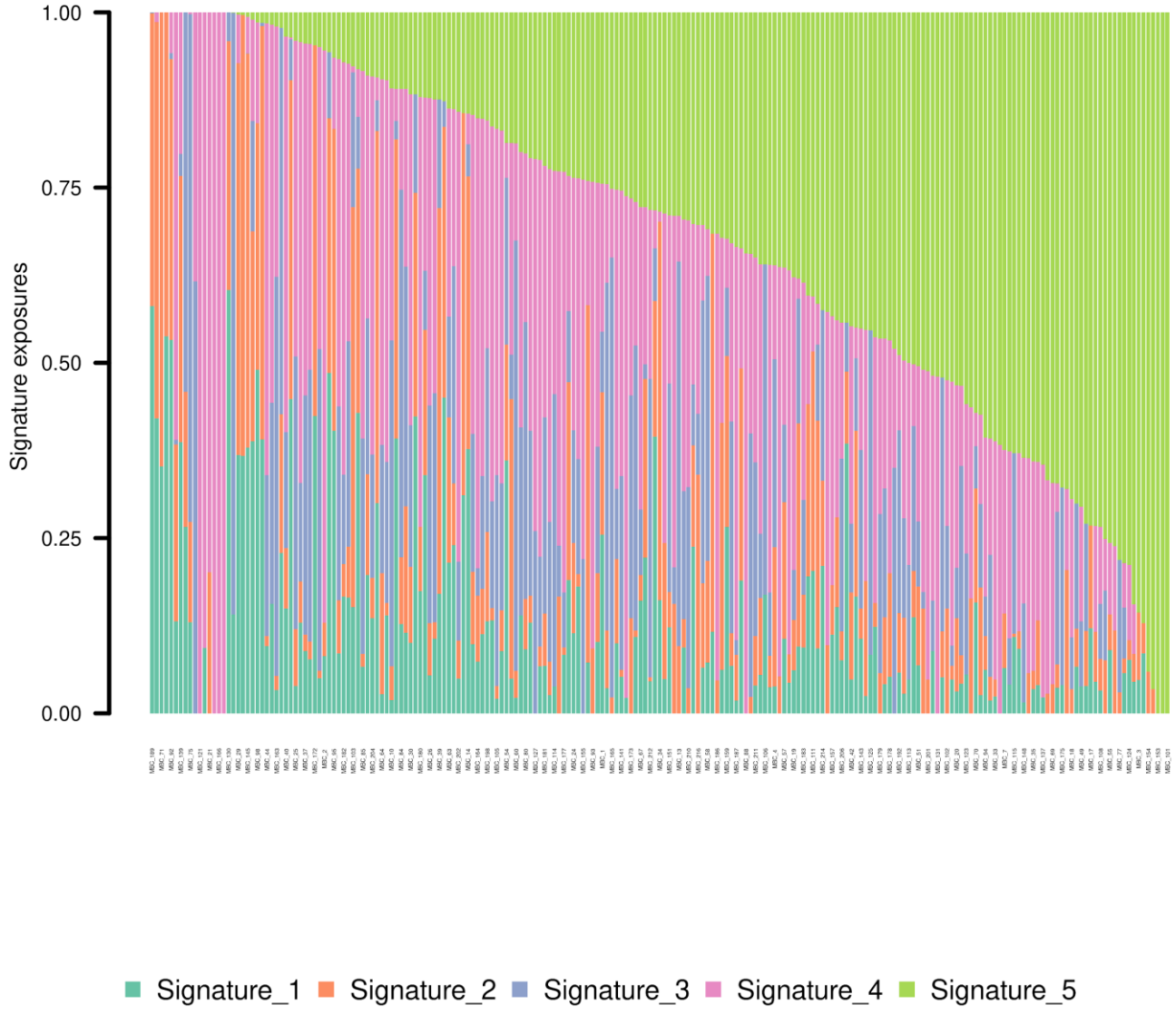


Figure 51. Signature contribution to each patient in the IGR dataset, represented as stacked barplots.

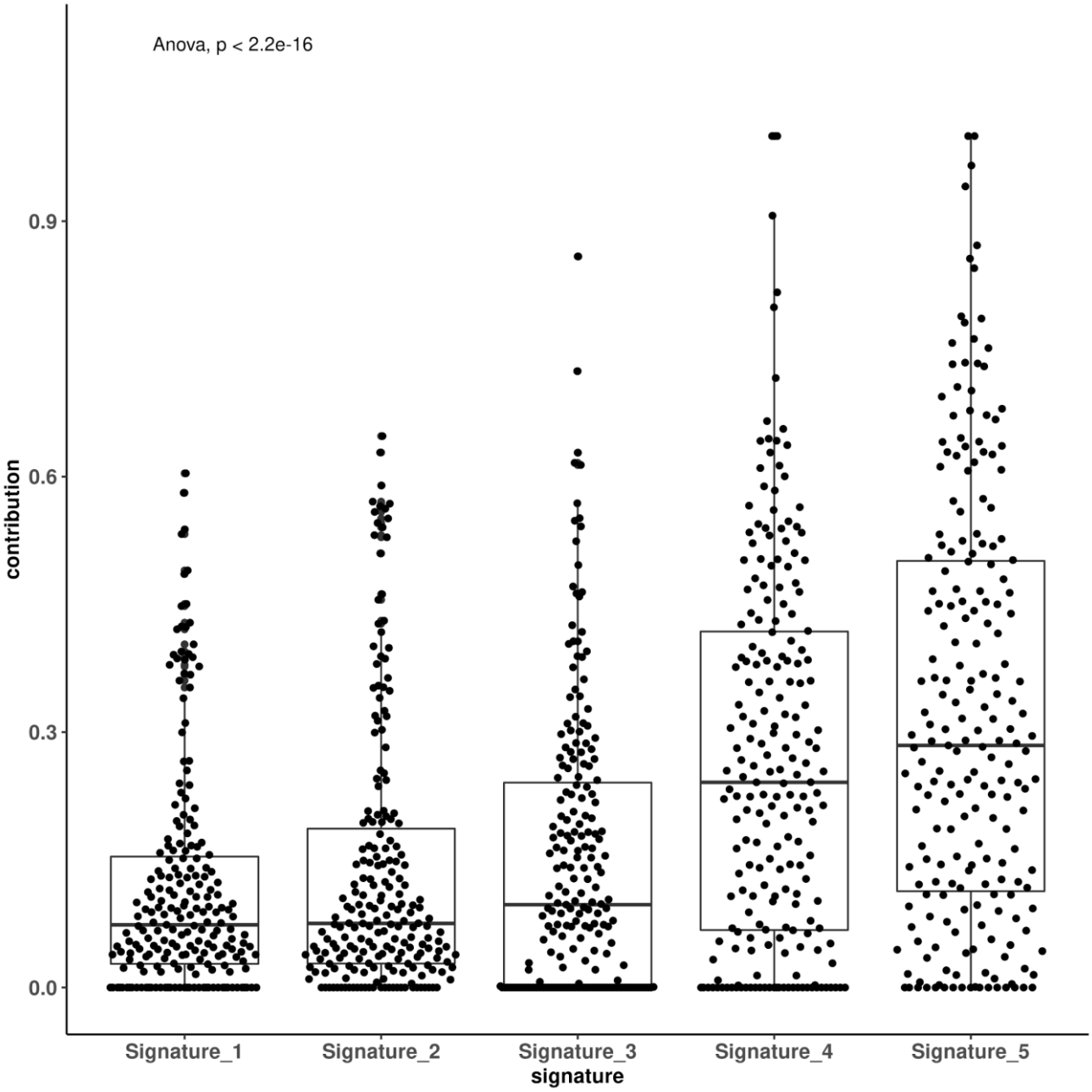


Figure 52. Distribution of contributions of each signature in the IGR dataset. Each dot represents a patient. One-way ANOVA results are also included in the graph. The upper and lower edges of the boxes represent 75th and 25th quantiles, respectively, while black rod

represents the median. Whiskers extend until 1.5* IQR if there are observations extremum to those points, and till the minimum and maximum values otherwise.

Figure 51 represents the contribution of each signature to the mutational profile of each patient as stacked barplots. From this plot it can be observed that the most dominant signatures were Signature 4 and 5, which were related to the defects in the DNA repair pathways. Indeed, the ANOVA analysis and boxplot of contributions agreed with signature 4 and 5 being the dominant signatures, followed by signatures 3, 2, and 1, in descending order (**Figure 52**).

4.4.5.5 Correlation Between Signatures and TMB in the IGR cohort.

Next, the natural course of action in the mutational signature analysis is to try and identify the clinical correlates for each signature in the dataset being analyzed. I started this analysis by first considering TMB. When no filter was applied to the TMB values, it could be observed that the APOBEC related signatures, Signature 1 and Signature 2, displayed a highly significant relationship with the TMB. Signatures 3, 4 and 5, on the other hand, displayed a significant negative relationship with TMB (**Figure 53**). From inspection of **Figure 53,** I reasoned that outliers, i.e., hypermutators, could significantly alter the nature of the relationship. Consequently, I decided to repeat the same correlation analysis at the TMB cutoffs of 250 (**Figure 54**) and 150 (**Figure 55**). The relationship between the TMB and Signatures 1 and 2 were unaffected by the cutoff changes, indicating that these results were robust. On the contrary, Signatures 3 and 4 lost their significance at these new cutoffs. Lastly, signature 4 was also unaffected by the different TMB cutoffs (**Figure 54, 55**).

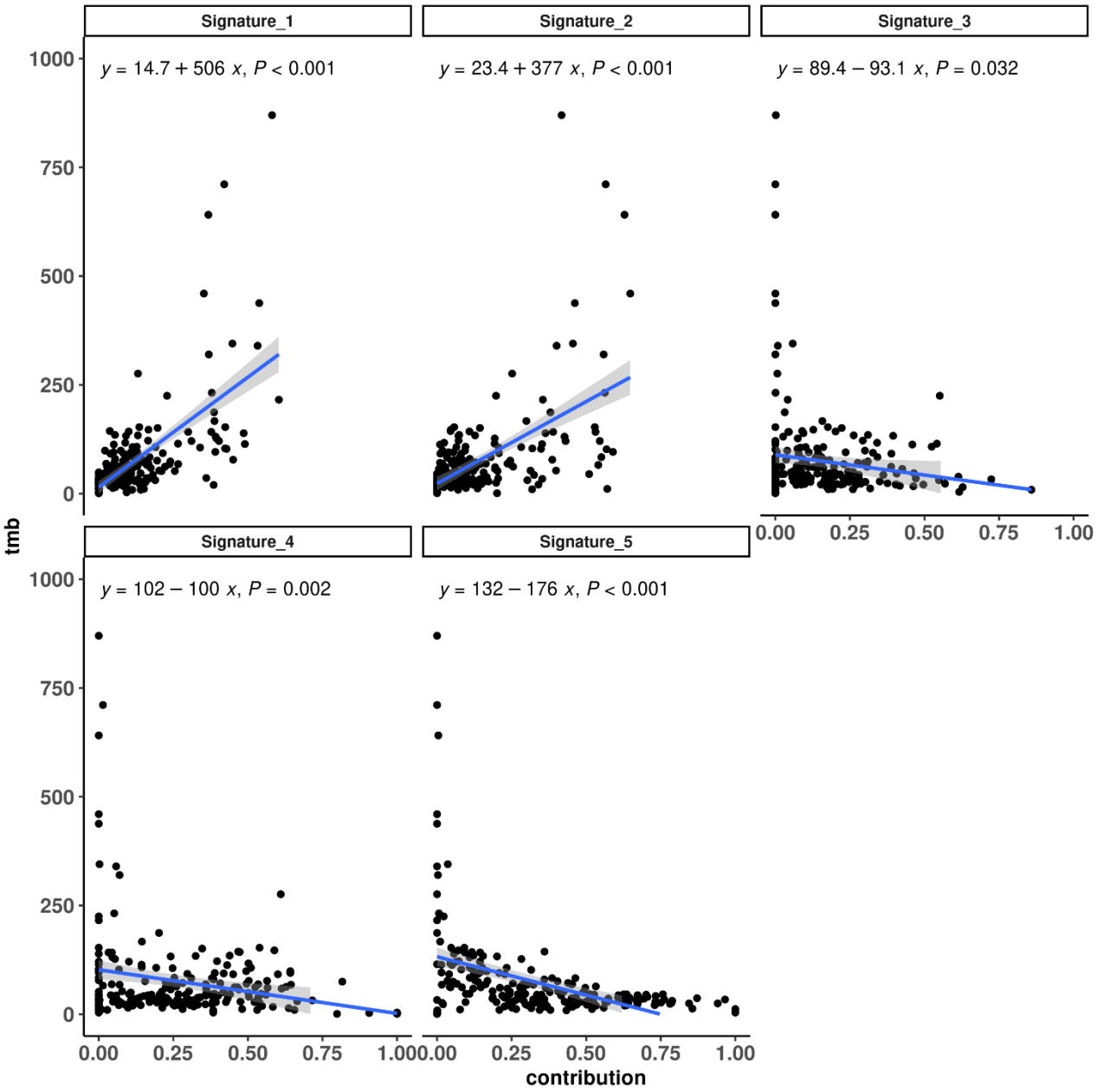


Figure 53. Linear Regression analysis of the signature contributions with TMB. The blue line represents the best fit line, and the grey area represents the confidence interval while each dot represents a patient.

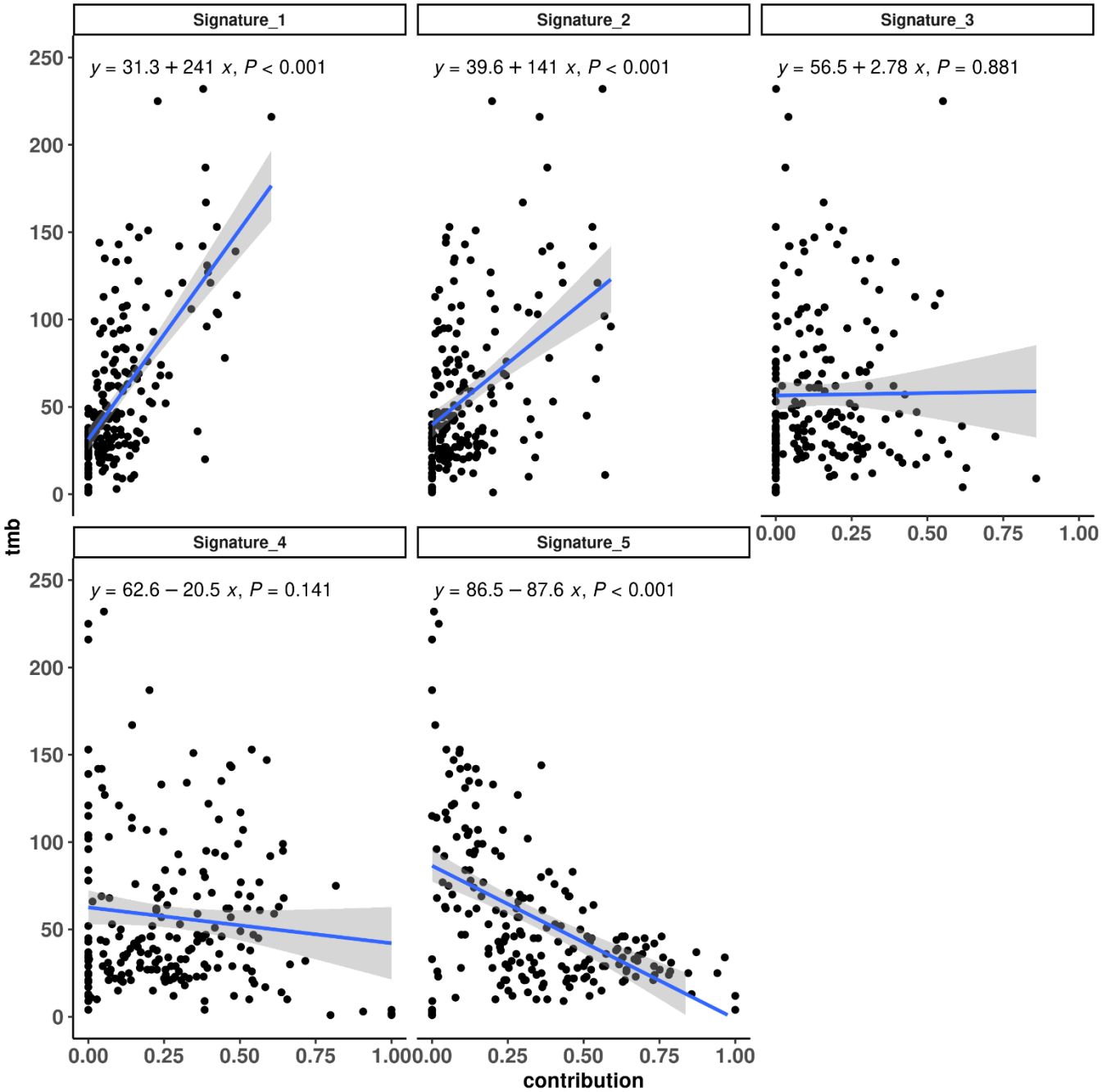


Figure 54. Linear Regression analysis of the signature contributions with TMB, at TMB cutoff of 250. Blue line represents the best fit line, and the grey area represents the confidence interval meanwhile each dot represents a patient.

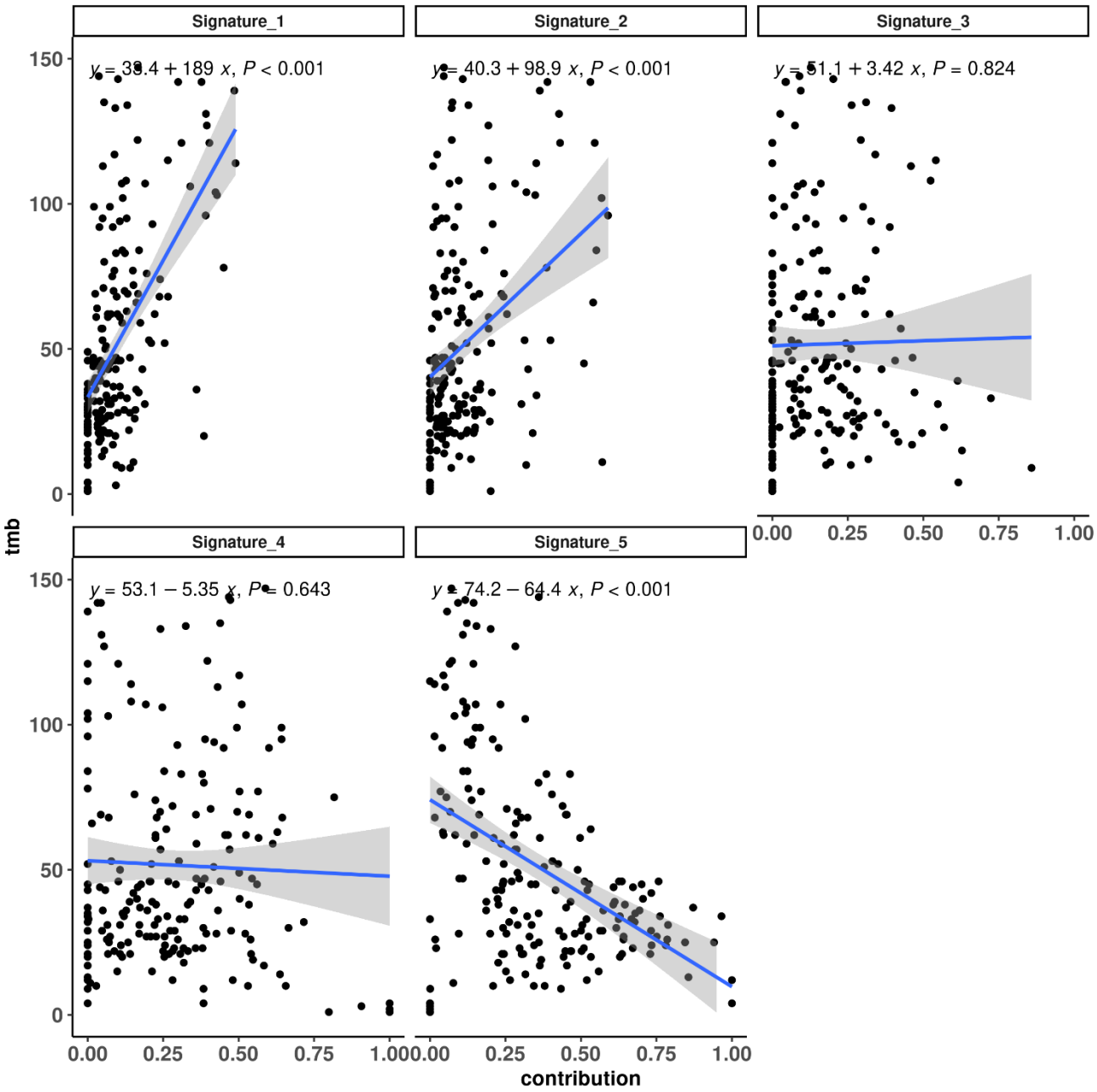


Figure 55. Linear Regression analysis of the signature contributions with TMB, at TMB cutoff of 150. Blue line represents the best fit line, and the grey area represents the confidence interval meanwhile each dot represents a patient.

4.4.5.6 Relationship Between Signatures and the HR and HER Status of Samples in the IGR Cohort

Next, I have tried to elucidate any relationship that might exist between the signatures and the HR and HER status of the patients. To this end, I first conducted an ANOVA analysis of the contribution of each signature within the HR/HER subtypes (**Figure 56**). From this analysis, it was evident that the overall contribution of the signatures within each subtype was reminiscent of the overall population (**Figure 52**), with Signature 4 and 5 being the dominant signatures,

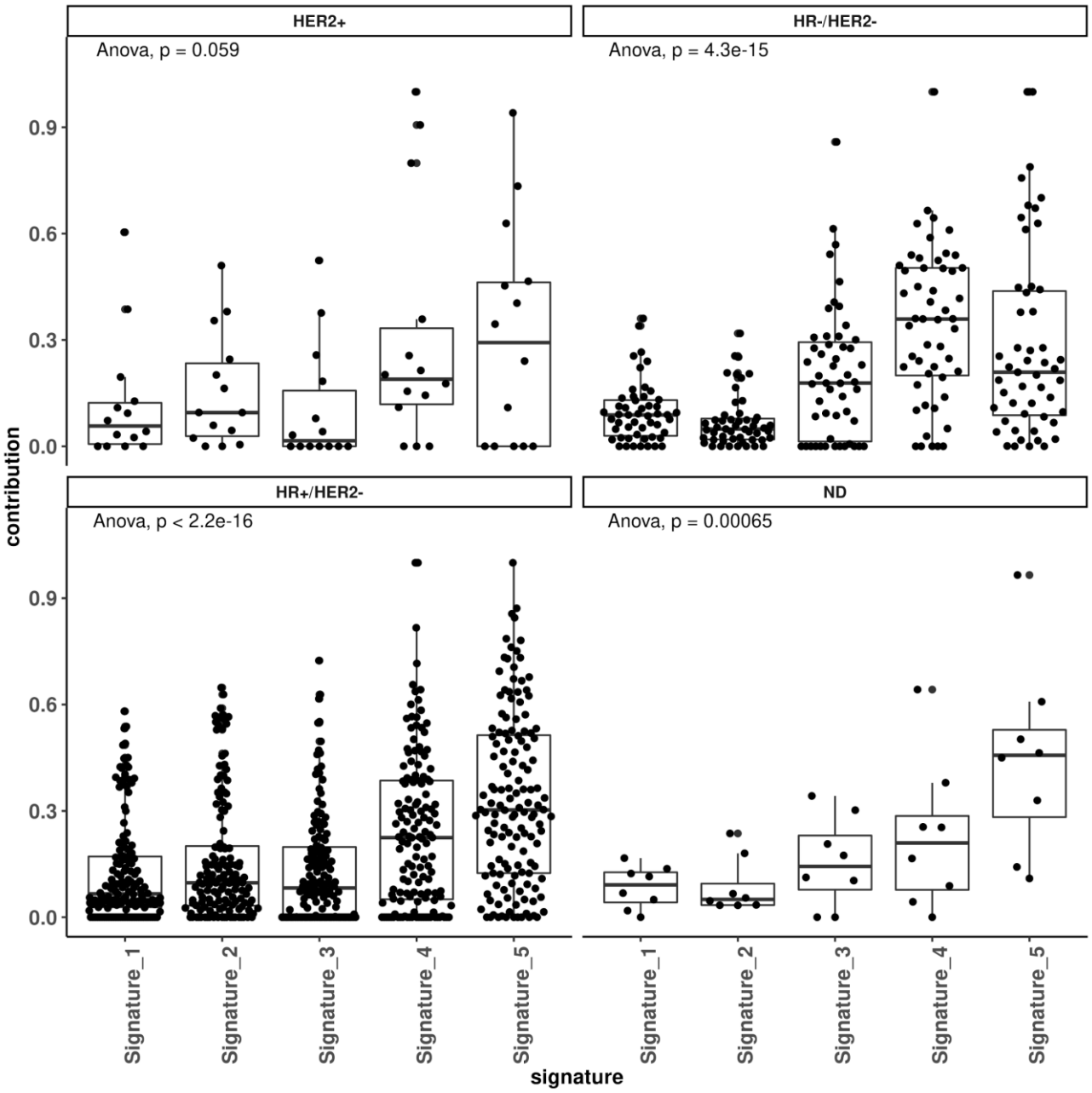


Figure 56. Distribution of contributions of each signature in the IGR dataset, within each HR/HER subtype. Each dot represents a patient. One-way ANOVA results are also included in the graph. The upper and lower edges of the boxes represent 75th and 25th quantiles, respectively, while the black rod represents the median. Whiskers extend until 1.5* IQR if there are observations extremum to those points, and till the minimum and maximum values otherwise.

followed by signatures 4, 3, 1. To better compare, I compared the contribution of each signature between the HR/HER subtypes. First, I compared the contribution differences

between HR+ and TNBC samples (**Figure 57**). Signatures 3 and 4 were contributing more

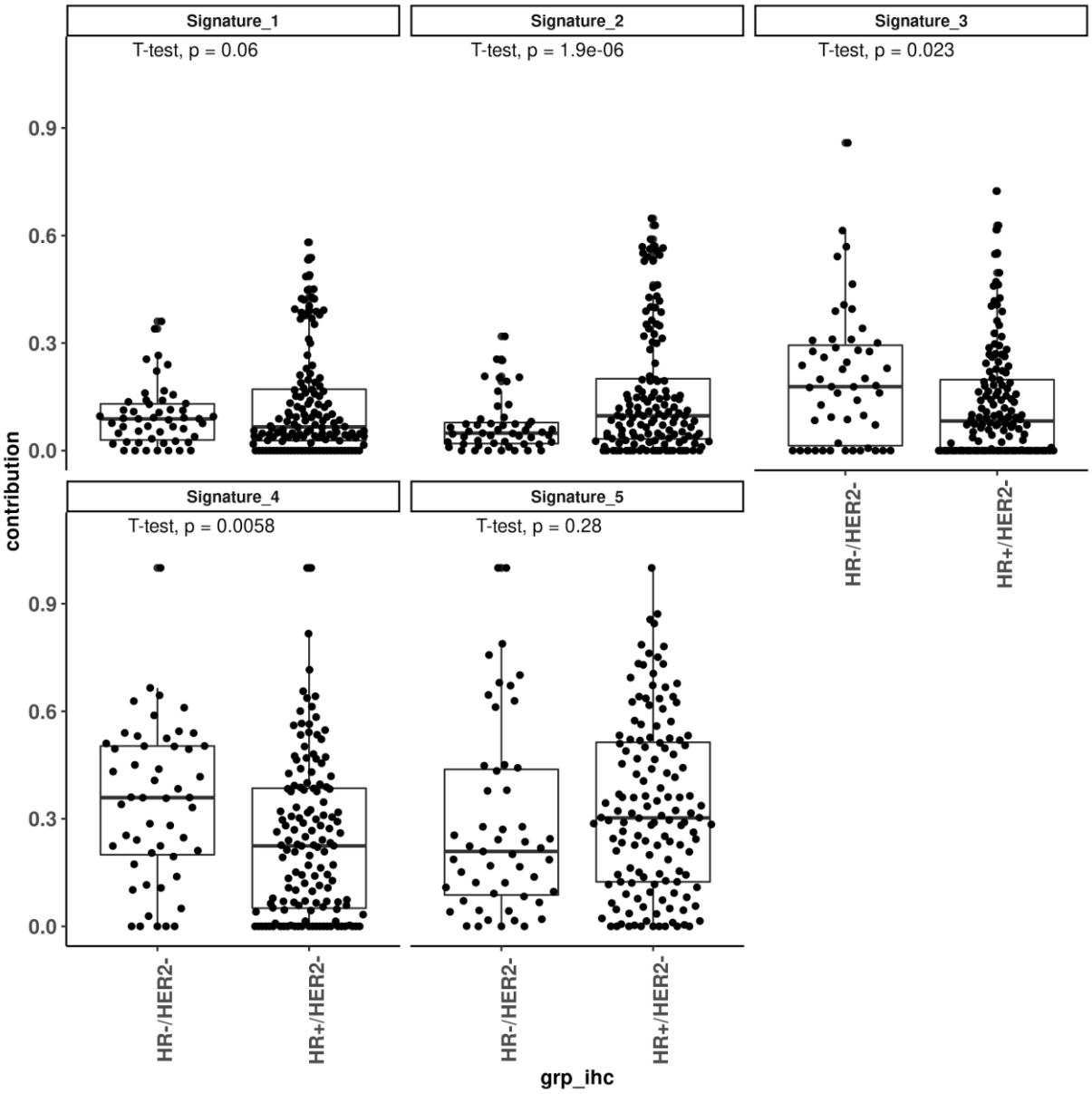


Figure 57. Distribution of contributions of each signature in the IGR dataset compared between HR+ and TNBC samples. Each dot represents a patient. Student's t-test results are also included in the graph. The upper and lower edges of the boxes represent 75th and 25th quantiles, respectively, while a black rod represents the median. Whiskers extend until 1.5* IQR if there are observations extremum to those points, and till the minimum and maximum values otherwise.

to the mutational profile of TNBC samples compared to HR+ samples. I would like to mention that signature 3 was unique to the IGR dataset. Additionally, HR+ samples seemed to have a higher contribution of the APOBEC related signatures, with Signature 1 approaching significance and Signature 2 being highly significant.

A similar analysis between the HER+ and TNBC samples revealed that the difference in the contribution of Signature 3 reached near significance, with TNBC samples having a higher mean. On the contrary, signature 2 had a higher mean contribution in the HER+ samples, which approached near significance (**Figure 58**).

Next, I conducted the same analysis between HR+ and HER2+ samples. In this analysis, however, none of the comparisons yielded a significant result (**Figure 59**).

4.4.5.7 Relationship Between Signatures and the Metastasis Location of Samples in the IGR Cohort

The last type of clinical information available in the IGR dataset was the information regarding the biopsy site of the samples. The same pairwise analyses were also conducted on the biopsy site information, however, none of the comparisons yielded a significant result (**Figure 60-62**). Indeed, when the metastatic samples from liver and lymph nodes were compared, medians of contributions were all similar for every signature (**Figure 60**). The same pattern was also observed when the liver and skin biopsies were compared (**Figure 61**). However, in the lymph node versus skin biopsy comparison, it was possible to observe visual differences. Signature 2 seemed to have a higher median contribution in lymph node metastasis, and vice versa was true for signature 5 (Figure 62).

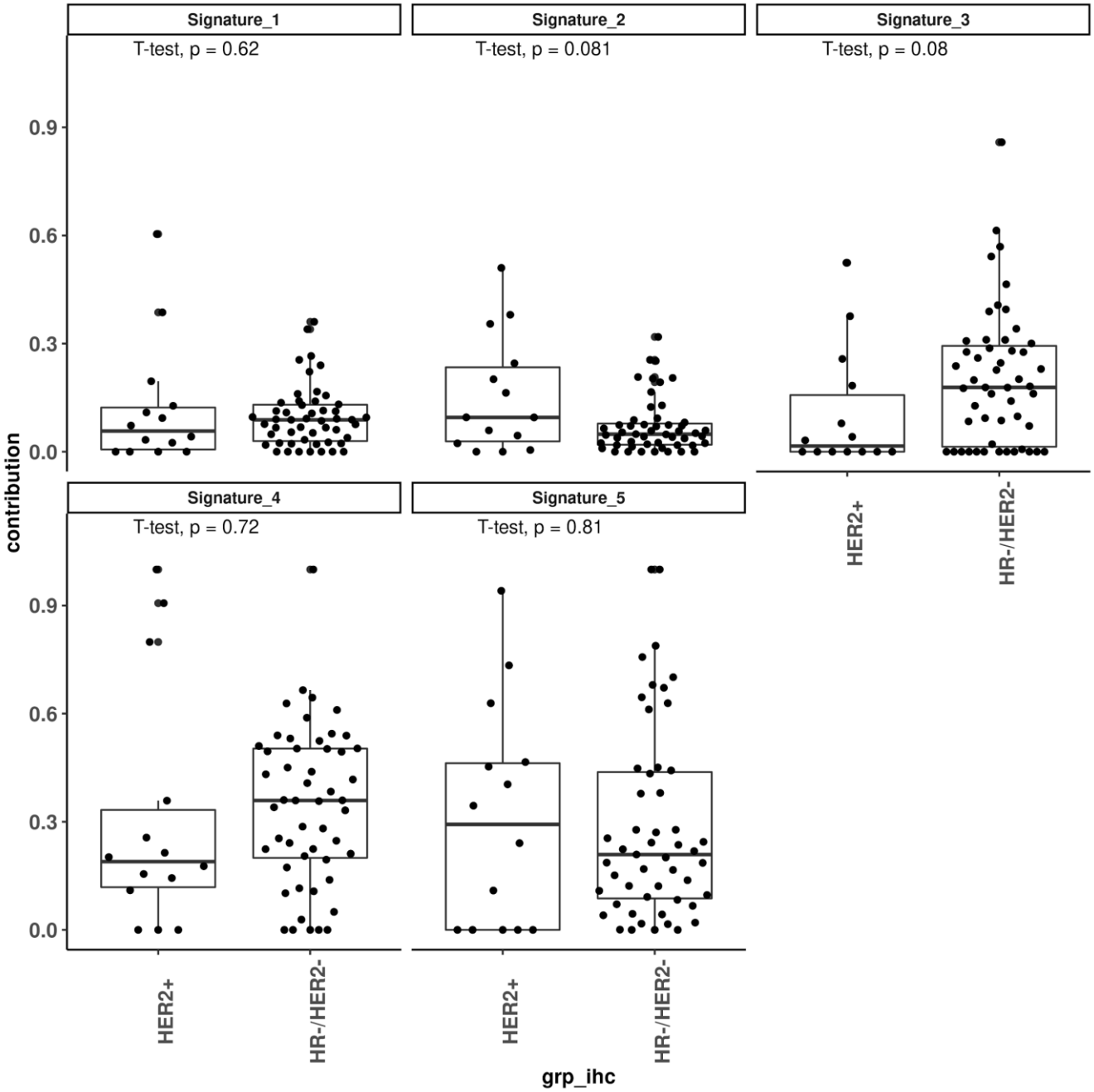


Figure 58. Distribution of contributions of each signature in the IGR dataset compared between HER2+ and TNBC samples. Each dot represents a patient. Student's t-test results are also included in the graph. The upper and lower edges of the boxes represent 75th and 25th quantiles, respectively, while a black rod represents the median. Whiskers extend until 1.5* IQR if there are observations extremum to those points, and till the minimum and maximum values otherwise.

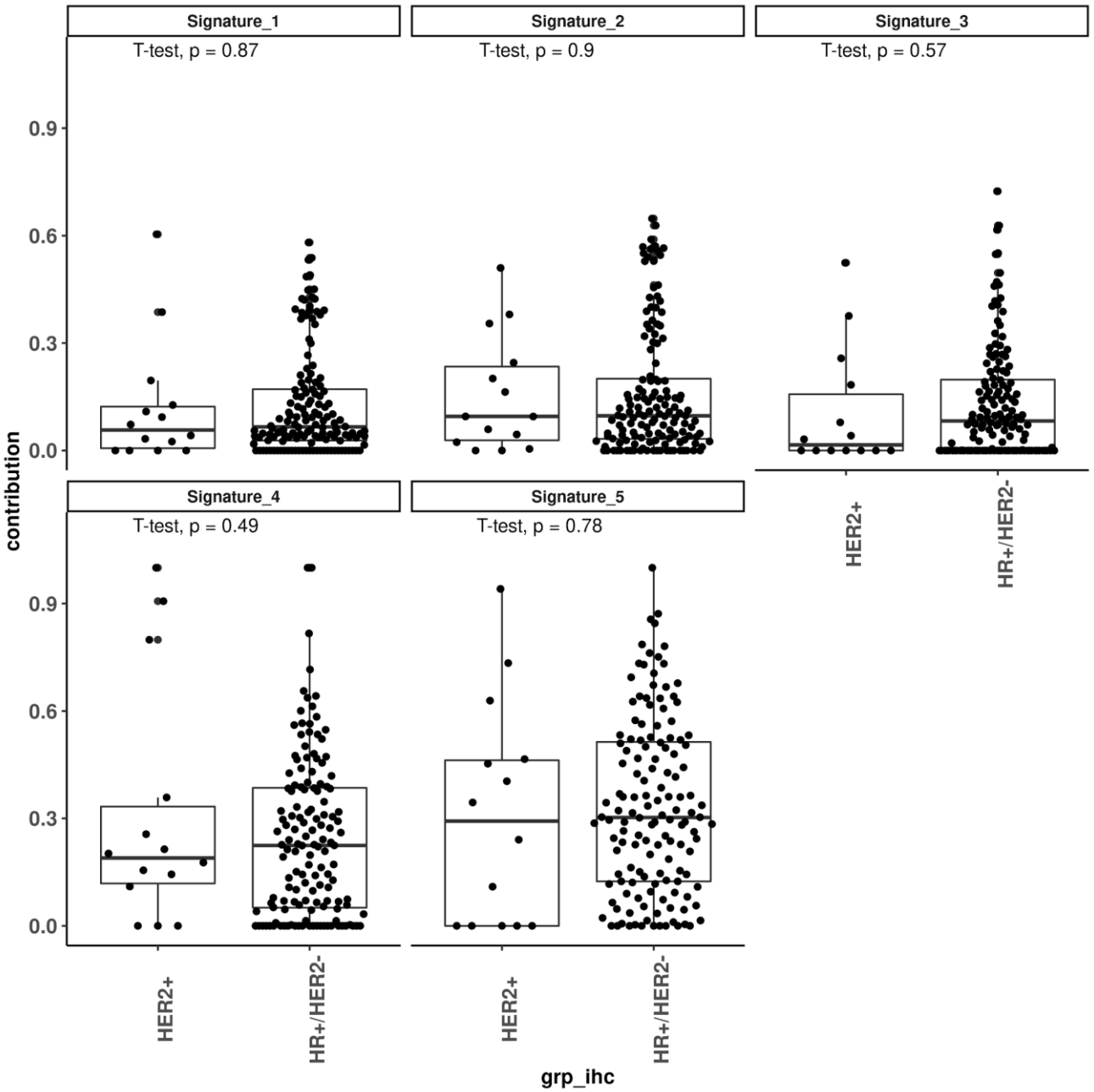


Figure 59. Distribution of contributions of each signature in the IGR dataset compared between HER2+ and HR+ samples. Each dot represents a patient. Student's t-test results are also included in the graph. The upper and lower edges of the boxes represent 75th and 25th quantiles, respectively, while a black rod represents the median. Whiskers extend until 1.5* IQR if there are observations extremum to those points, and till the minimum and maximum values otherwise.

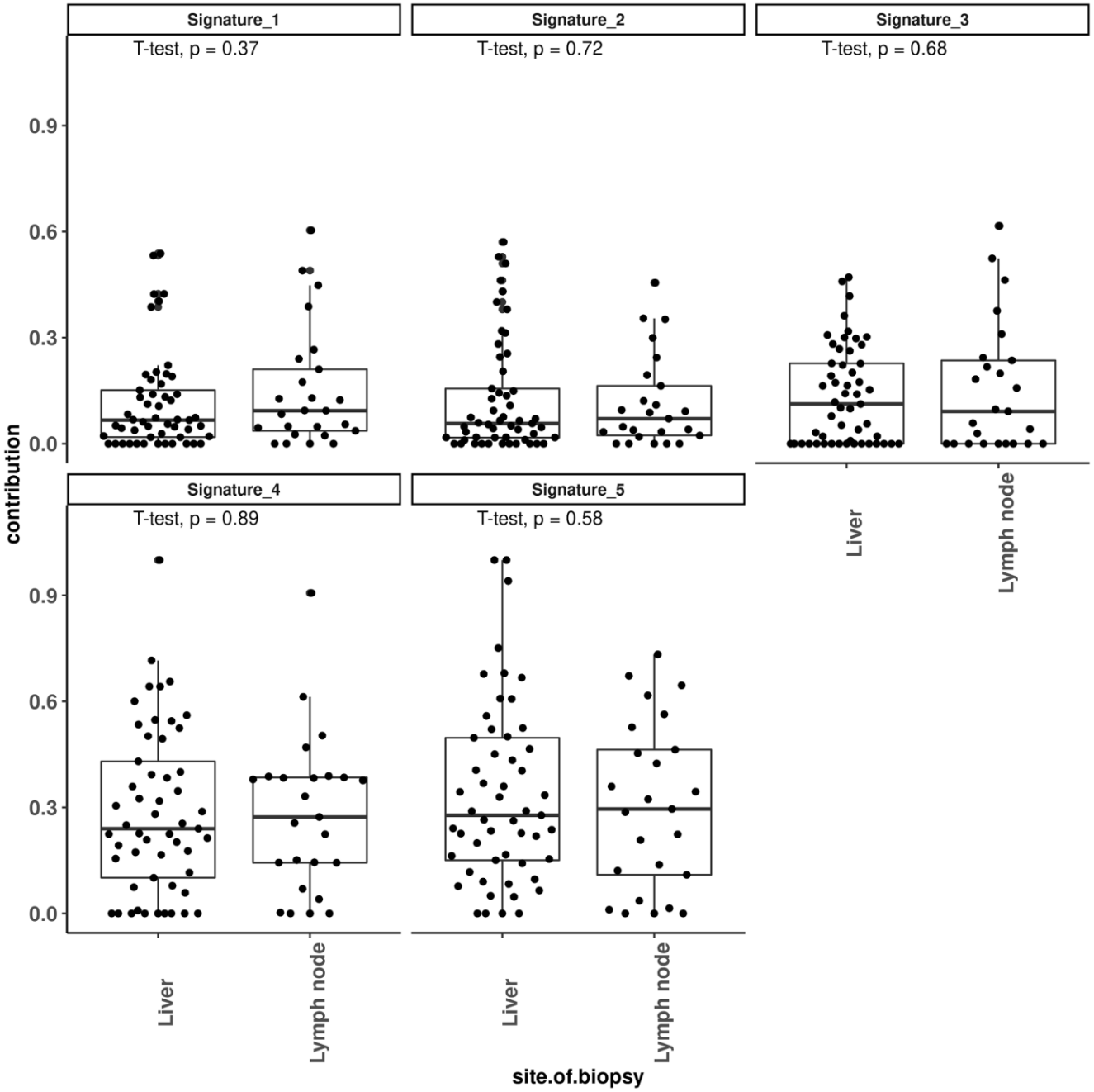


Figure 60. Distribution of contributions of each signature in the IGR dataset compared between Liver and Lymph node biopsies. Each dot represents a patient. Student's t-test results are also included in the graph. The upper and lower edges of the boxes represent 75th and 25th quantiles, respectively, while a black rod represents the median. Whiskers extend until 1.5* IQR if there are observations extremum to those points, and till the minimum and maximum values otherwise.

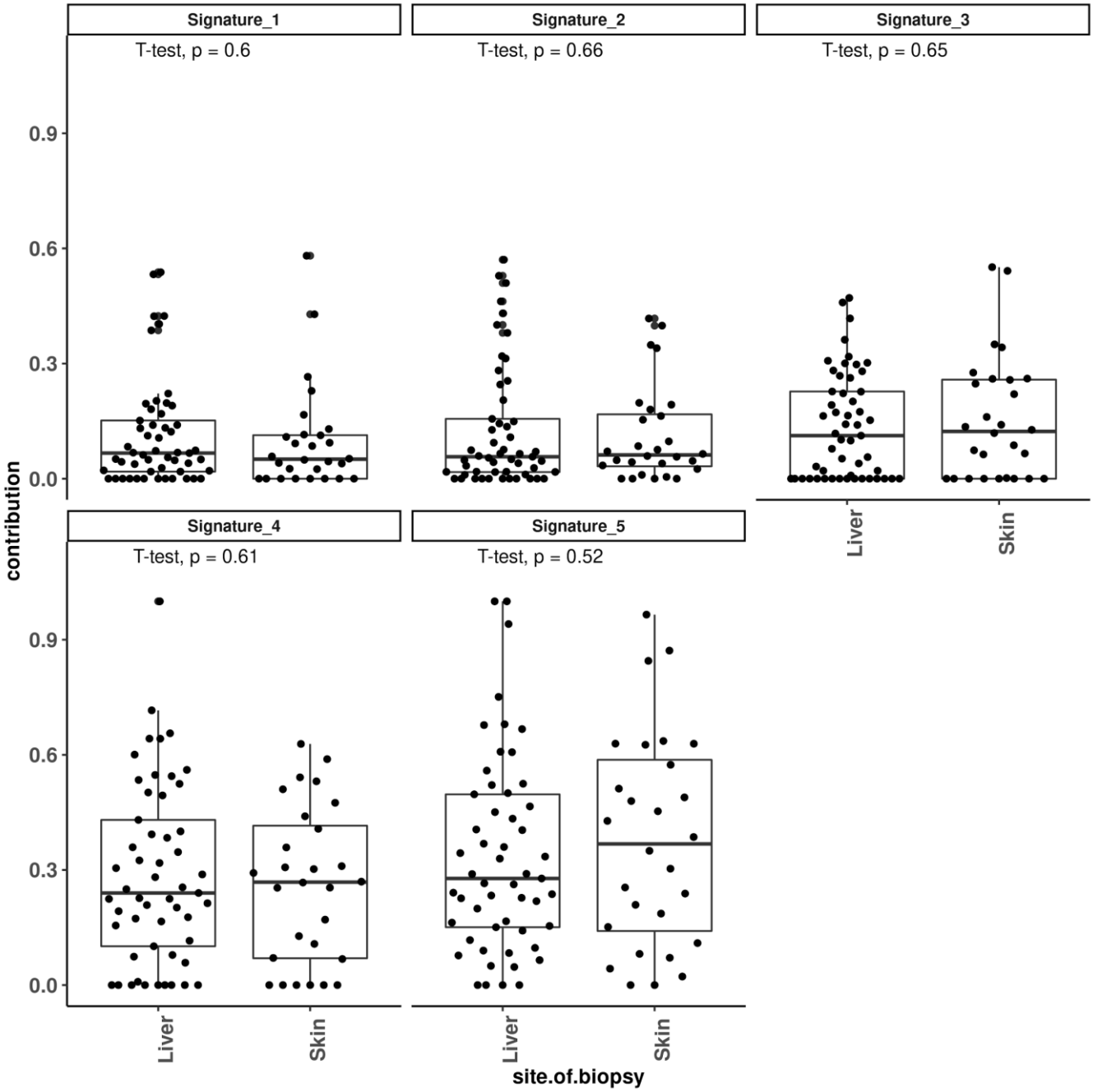


Figure 61. Distribution of contributions of each signature in the IGR dataset compared between Liver and Skin biopsies. Each dot represents a patient. Student's t-test results are also included in the graph. The upper and lower edges of the boxes represent 75th and 25th quantiles, respectively, while a black rod represents the median. Whiskers extend until 1.5* IQR if there are observations extremum to those points, and till the minimum and maximum values otherwise.

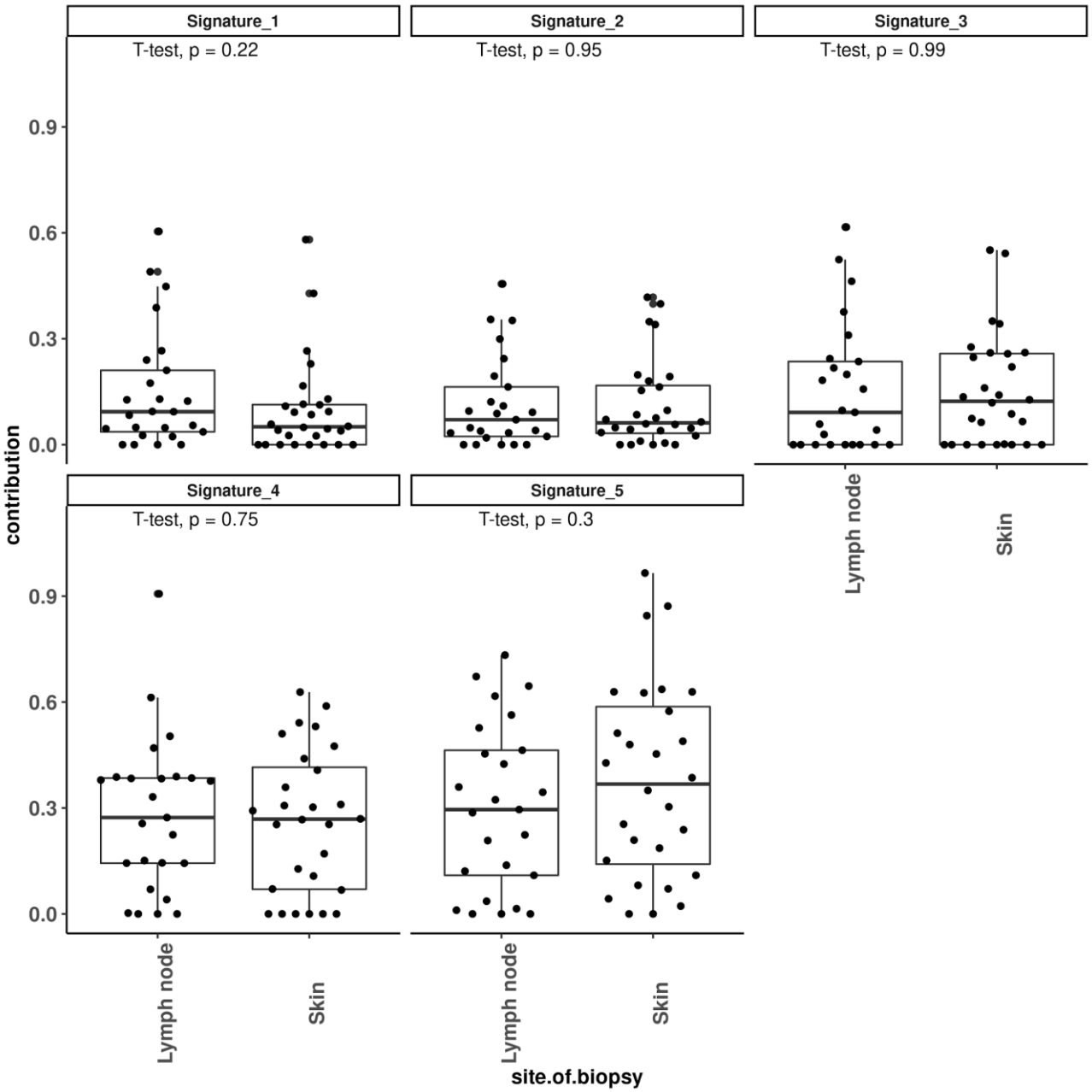


Figure 62. Distribution of contributions of each signature in the IGR dataset compared between Lymph node and Skin biopsies. Each dot represents a patient. Student's t-test results are also included in the graph. The upper and lower edges of the boxes represent 75th and 25th quantiles, respectively, while a black rod represents the median. Whiskers extends until 1.5* IQR if there are observations extremum to those points, and till the minimum and maximum values otherwise.

4.4.6 Differentially Mutated Gene Analysis between IGR and TCGA Cohorts

Next, to determine the genes which were differentially mutated between these two datasets, I performed Fisher's exact test. Overall, 4 genes passed the FDR threshold of 0.05. Among those, ESR1 and FSIP2 genes were also part of the top 25 highly mutated genes and were unique to the IGR cohort (**Figure 43, 44, 63**). In order to control for the possible influence of hypermutators I repeated the same analysis after excluding the hypermutators (**Appendix Figure 1**). This analysis revealed similar results, except for *ARMCX4*, which was not significantly differentially mutated anymore (**Appendix Table 1**).

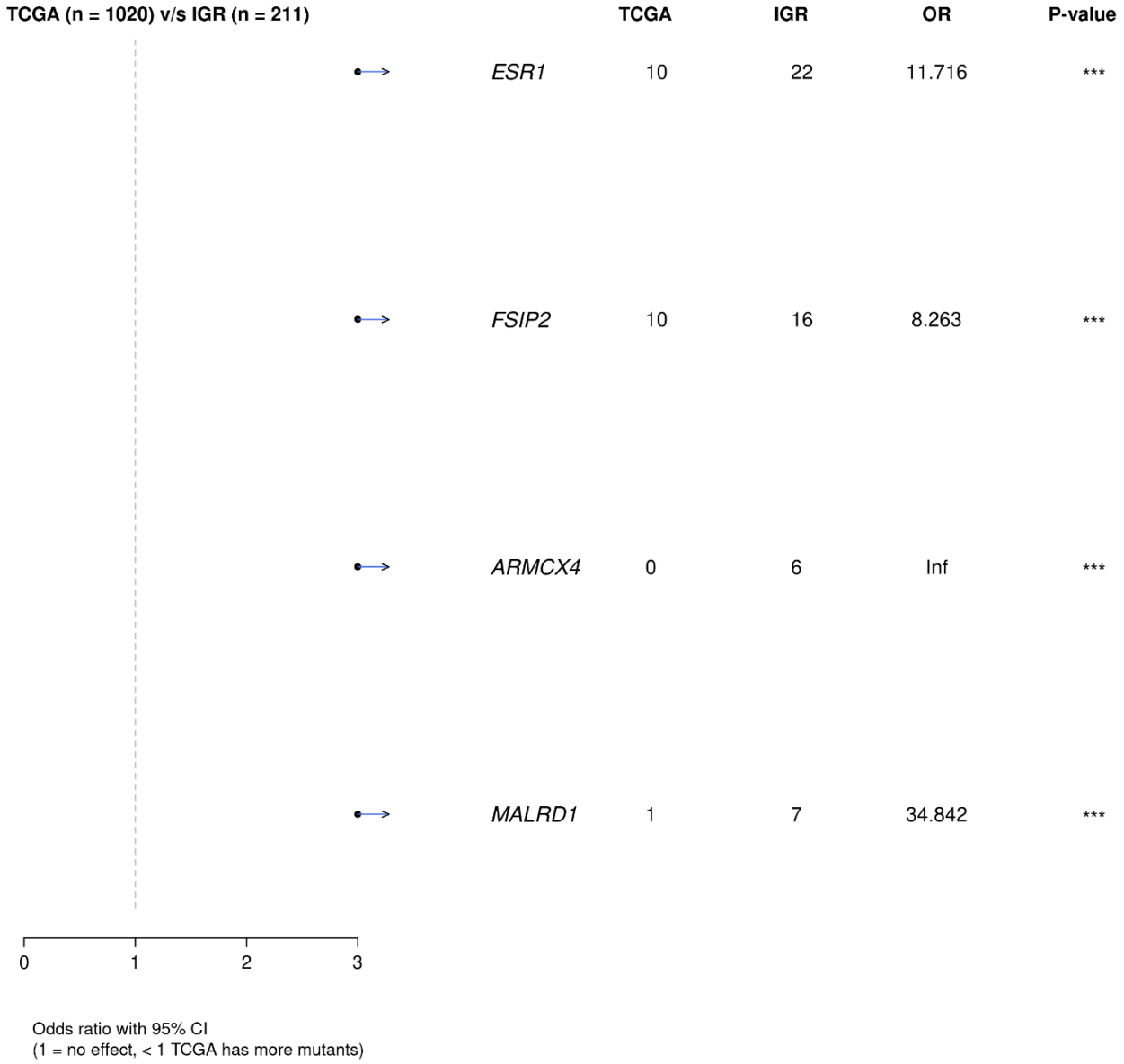


Figure 63. Forest plot of the genes that are differentially mutated between IGR and TCGA cohorts. The dot represents the OR, and if the confidence interval is out of the interval shown an arrow is drawn. The number of the mutated samples in each dataset along with the respective OR is shown. * p< 0.05, **p<0.01, *** p<0.001.

4.4.7 Identification of the Driver Genes in IGR and TCGA cohorts

BRCA:Cohort-vs-TCGA uses a positional proximity based driver mutation identification algorithm, whereby genes that appear to have mutations in the same or nearby residues are defined as cancer driver genes [50]. Consequently, I identified four and six near significant driver genes in IGR and TCGA cohorts, respectively. Out of four driver genes identified in the IGR cohort two of them, *ESR1* and *GJA8*, were unique to the IGR cohort. *KRAS, FAM102A, PLP2, PIK3R3,* on the other hand, were only identified in the TCGA cohort (**Table 5**).

Table 5. Driver genes identified in each cohort by the OncodriveCLUST algorithm [50].

Gene Symbol	**FDR in IGR**	**FDR in TCGA-BRCA**
AKT1	0.00657199732734	0.000129175900095
ESR1	0.00657199732734	0.97225028
PIK3CA	0.00657199732734	0.005432502829452
GJA8	0.050776449903783	NA
PLP2	NA	0.000129175900095
KRAS	NA	0.002963776943342
FAM102A	NA	0.005432502829452
PIK3R3	NA	0.060021915413002

Furthermore, I visually inspected some of those driver genes with the help of the co-lollipop plots. In the case of the *ESR1* all of the mutations observed fall within the nuclear receptor ligand binding domain. These mutations were significantly clustered around residues 537, 538, and 380. Moreover, such enrichment hasn't been seen in the TCGA cohort. It is worth noting that ESR1 was also among the differentially mutated genes (**Figure 63, 64**).

Next, I explored the mutation distribution of another gene that was unique to the IGR dataset, *GJA8*. This gene had only five mutations in the IGR dataset, however, three of those happened on the same residue and resulted in the conversion of glutamate, a negatively charged amino acid, to lysine, a positively charged amino acid [139]. Moreover, an additional mutation existed within 30 amino acids of this mutation, p.Ile82Phe. All four of these mutations were either within the connexin domain or near it. Moreover, VAF of these four mutations ranged from 0.17 to 0.34, suggesting subclonal selection (**Table 6, Figure 65**).

I repeated the same oncodriveCLUST analysis in IGR after exclusion of the hypermutators. Results indicated that *GJA8* gene lost its significance meaning most of its mutations were in the hypermutator samples (**Appendix Table 2**.).

Table 6. Summary of the VAF, altered residues, reference and variant allele read numbers of GJA8 mutations in normal and tumor samples of the IGR cohort.

Hugo_Symbol	t_ref_count	t_alt_count	n_ref_count	n_alt_count	HGVSp	VAF
GJA8	135	69	184	0	p.Glu110Lys	0.338235
GJA8	175	36	170	0	p.Glu110Lys	0.170616
GJA8	307	90	295	1	p.Ile82Phe	0.2267
GJA8	77	18	88	0	p.Gly341Glu	0.189474
GJA8	179	67	169	0	p.Glu110Lys	0.272358

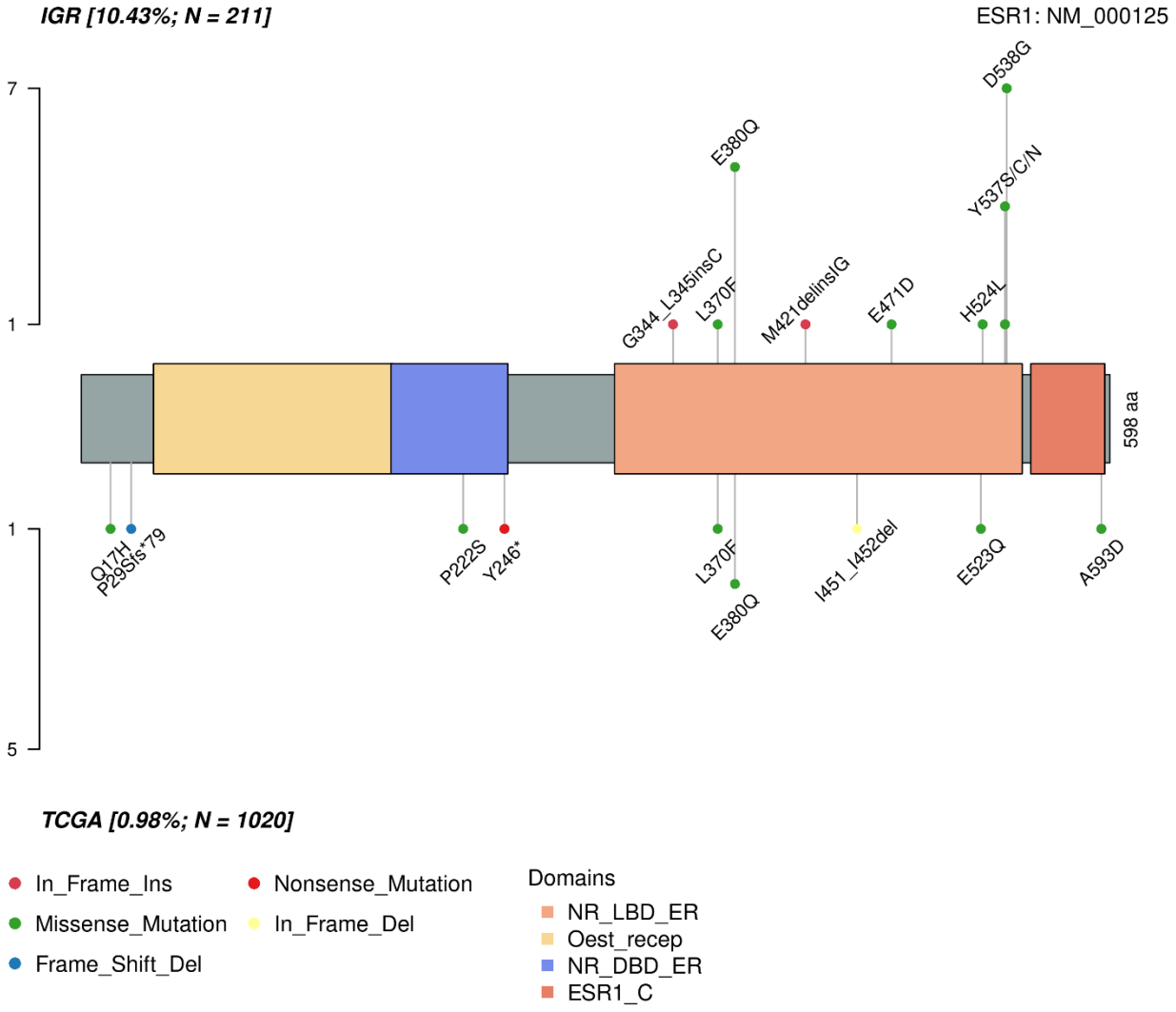


Figure 64. Co-lollipop plot of the ESR1 gene. The upper part of the plot represents the mutated residues in the IGR cohort, meanwhile the lower part represents the mutated residues in the TCGA cohort. Height of the lollipop corresponds to the number of events observed in each cohort.

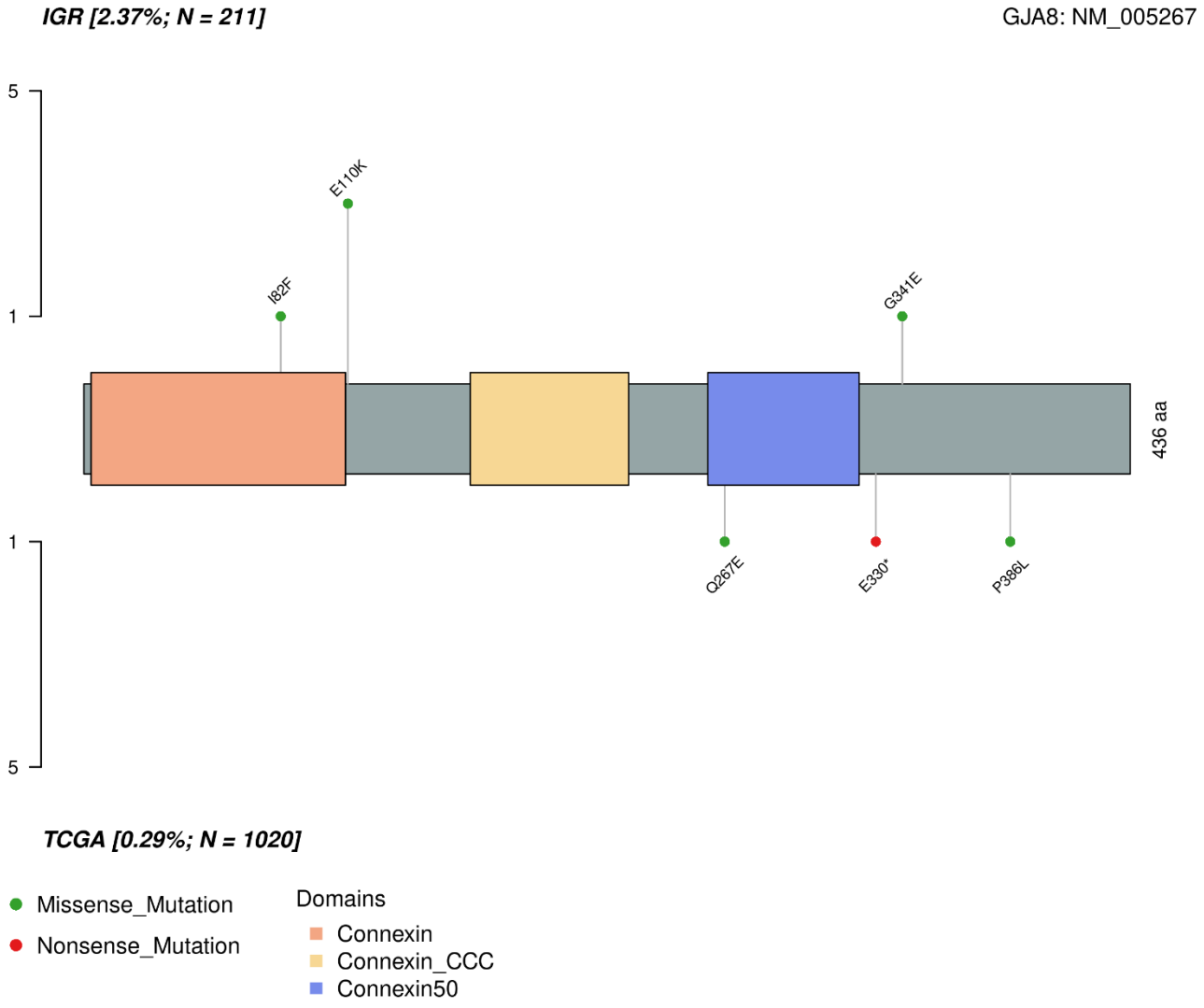


Figure 65. Co-lollipop plot of the GJA8 gene. The upper part of the plot represents the mutated residues in the IGR cohort, meanwhile the lower part represents the mutated residues in the TCGA cohort. Height of the lollipop corresponds to the number of events observed in each cohort.

4.4.8 Clinical Enrichment Analysis of Mutations in The IGR Cohort

4.4.8.1 Clinical Enrichment Analysis by HR/HER subtypes in The IGR Cohort.

I carried out the clinical enrichment analysis for all genes for each HR/HER subtype vs the rest. Fisher's exact test was used to compute the OR and the respective p-values and FDRs. At the end, I treated the results with FDR < 0.06 as significant. Overall, three genes across five comparisons were identified as significant. *ESR1* and *PIK3CA* mutations were significantly

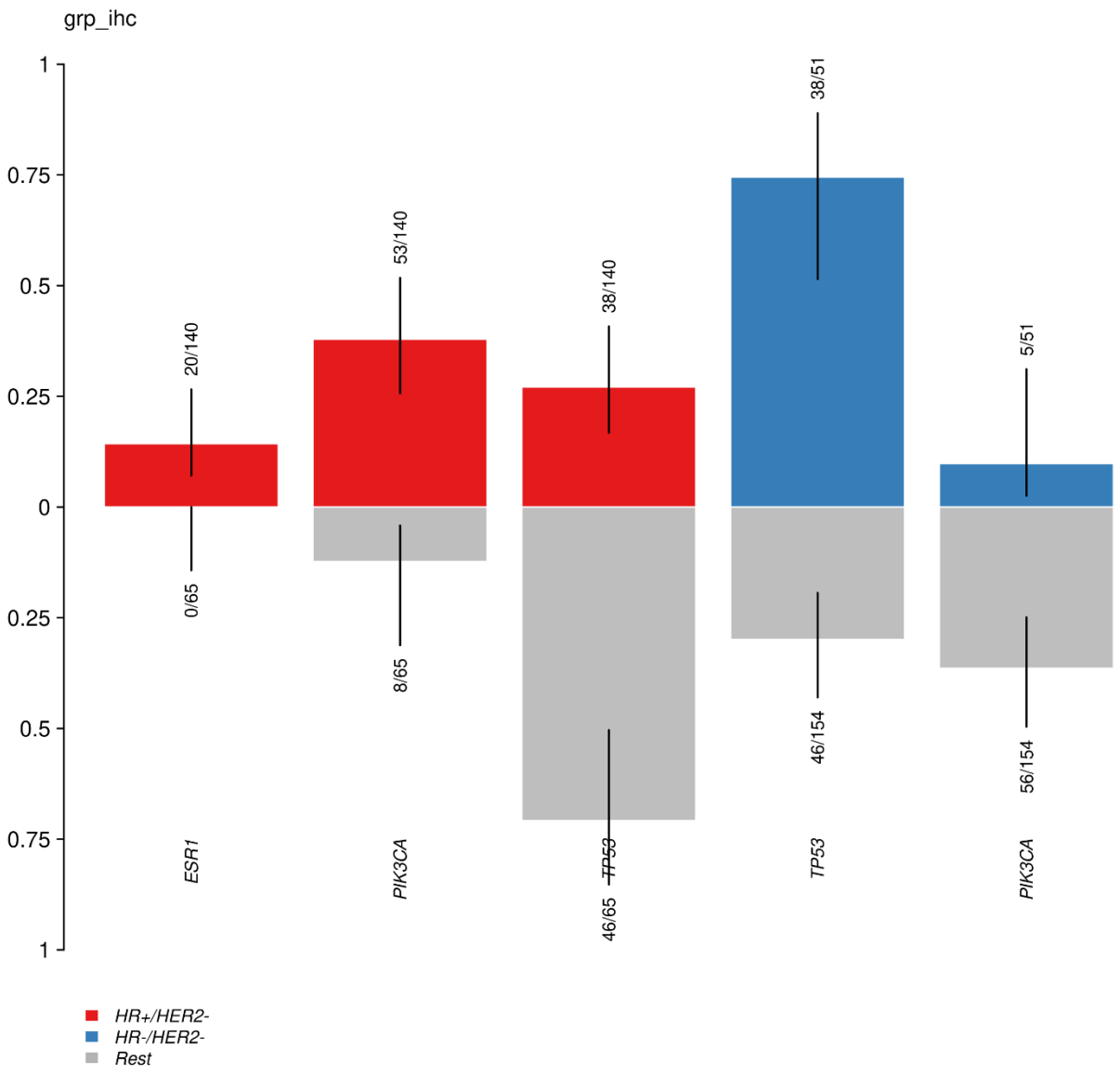


Figure 66. Clinical enrichment results for HR/HER subtypes. The upper part represents the fraction of samples with the mutation in the indicated subtype for the shown gene, whereas the lower part represents the fraction of samples with the mutation in the rest of the samples for the shown gene. Genes with FDR <0.06 were selected.

enriched in the HR+ samples, while TP53 mutations were enriched in the TNBC samples (**Figure 66**). **Table 7** displays the detailed Fisher's exact test for the significant genes. In order to eliminate the influence of hypermutators, the same analysis was

carried out after the exclusion of the hypermutator samples (**Appendix Table 3**). Interestingly, the *ESR1* enrichment in the HR+ group lost its significance. Additionally, TTN and MYH4 gene mutations appeared to be enriched in the HR+ samples (Appendix Table 3).

Table 7. Detailed Fisher's exact test results for the HR/HER subtype clinical enrichment analysis.

Hugo_Symbol	Group1	Group2	p_value	OR	OR_low	OR_high	fdr
TP53	HR+/HER2-	Rest	6.97E-09	0.155524	0.075734	0.308999	3.68E-06
TP53	HR-/HER2-	Rest	2.66E-08	6.790291	3.19085	15.26198	7.02E-06
PIK3CA	HR+/HER2-	Rest	0.000142	4.312883	1.855453	11.29831	0.02435
PIK3CA	HR-/HER2-	Rest	0.000184	0.191489	0.056096	0.5199	0.02435
ESR1	HR+/HER2-	Rest	0.000552	Inf	2.555006	Inf	0.058312

4.4.8.2 Clinical Enrichment Analysis by Metastasis Site in The IGR Cohort

I carried out the clinical enrichment analysis for all genes for each metastasis location vs the rest. Fisher's exact test was used to compute the OR and the respective p-values and FDRs. Since the analysis did not yield any significant results at FDR<0.05, I treated any positive enrichment with lower OR > 1 and any negative enrichment with higher OR < 1 as significant. In the end, the analysis yielded four genes significant across three comparisons. *MAP3K1* gene was highly mutated in the liver metastasis samples compared to the rest. For skin metastasis, two genes were positively enriched, *IGSF10* and *DYNC2H1*. *TTN* gene, on the other hand, was depleted from the "other" category, which was composed of the samples collected from tissues other than liver, lung, skin, or lymph node. Results are shown in **Figure 67** and **Table 8.** When the same analysis was repeated after the exclusion of the hypermutator samples, the previously identified associations remained the same (**Appendix Table 4**).

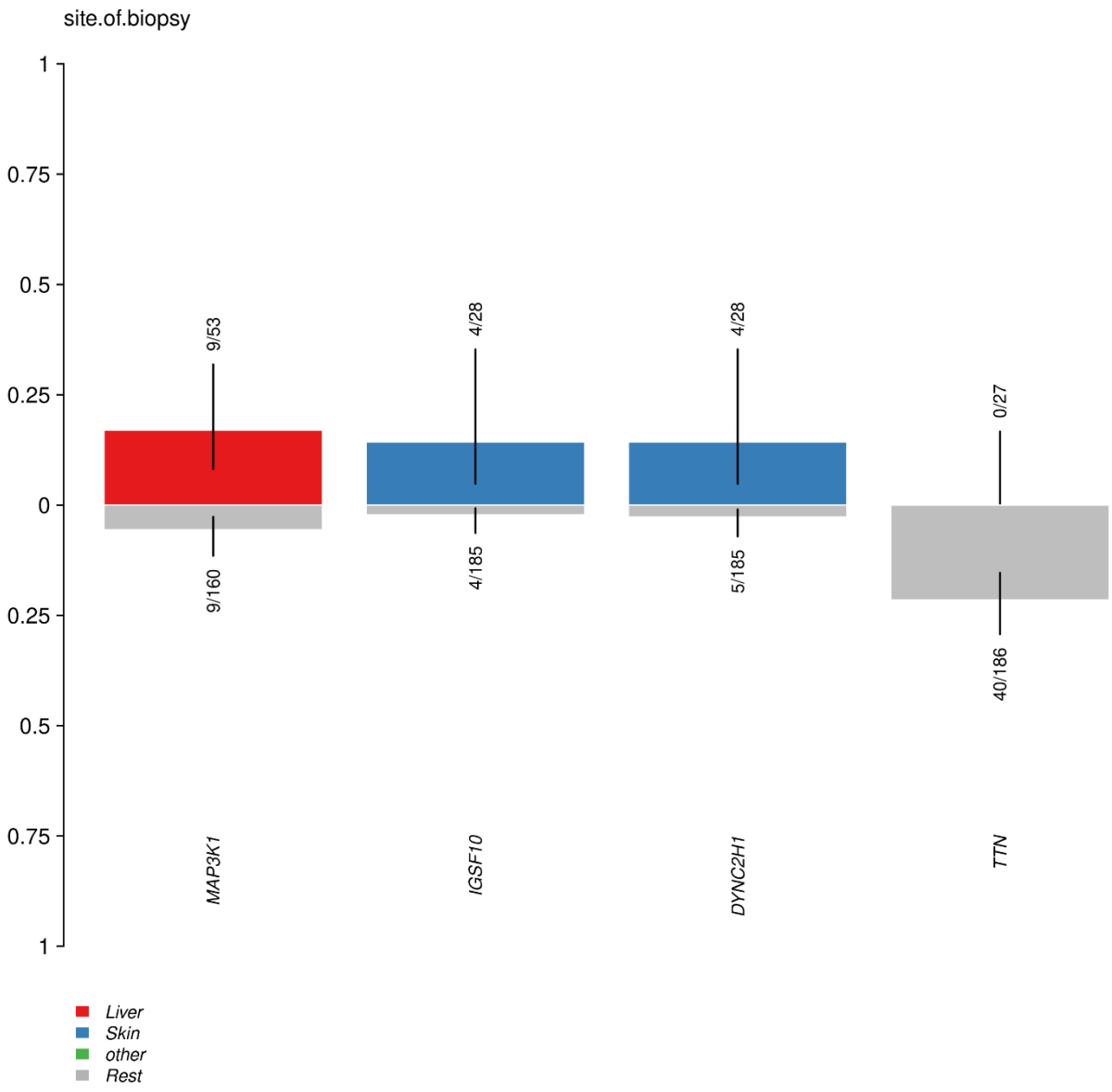


Figure 67. Clinical enrichment results for metastasis sites. The upper part represents the fraction of samples with the mutation in the indicated metastasis site for the shown gene, whereas the lower part represents the fraction of samples with the mutation in the rest of the samples for the shown gene. Genes with positive enrichment with lower OR > 1 or negative enrichment with higher OR < 1 are shown.

Table 8. Detailed Fisher's exact test results for metastasis site clinical enrichment analysis.

Hugo_Symbol	Group1	Group2	p_value	OR	OR_low	OR_high	fdr
TTN	other	Rest	0.003266	0	0	0.567008	1
IGSF10	Skin	Rest	0.011696	7.418062	1.293646	42.65705	1
MAP3K1	Liver	Rest	0.019033	3.407306	1.123428	10.35334	1
DYNC2H1	Skin	Rest	0.019122	5.916704	1.096356	29.65542	1

4.4.9 Somatic Interaction Analysis in TCGA and IGR cohorts.

4.4.9.1 Somatic Interaction Analysis in the IGR cohort.

Co-occurrence and mutual exclusivity status of each gene were determined via the Fisher's exact test, among the top 25 mutated genes. Overall, there were 37 significant somatic interactions at p-value <0.05. Out of those 37, only two were mutually exclusive interactions. Results are represented in **Figure 68.**

4.4.9.2 Somatic Interaction Analysis in the TCGA cohort.

Co-occurrence and mutual exclusivity status of each gene were determined via Fisher's exact test, among the top 25 mutated genes. Overall, there were 160 significant somatic interactions at p-value <0.05. Out of those 160, only six were mutually exclusive interactions. Results are represented in **Figure 69.**

4.4.9.3 Comparison of somatic interactions between TCGA and IGR cohorts.

Using the set operations, I identified common and unique somatic interactions for each dataset. Of special interest were the interactions unique to the IGR cohort, because TCGA had many unique interactions, owing to its very large sample size. Interactions unique to the IGR dataset are given in **Table 9**. *ESR1, FSIP2, RYR1, DNAH6, XIRP2, NF1, PCLO, FRAS1,* and *TENM3* genes were highly mutated genes unique to the IGR dataset. Except two unique pairs, all of the unique somatic interactions were composed

of the aforementioned genes. Additionally, *TP53* and *ERS1*, and *RYR1* and *RYR2* pairs were significant at the FDR cutoff <0.05. When this analysis was repeated after the exclusion of hypermutators from the IGR dataset, it was observed that the *RYR1* and *RYR2* pair lost its significance (**Appendix Table 5**).

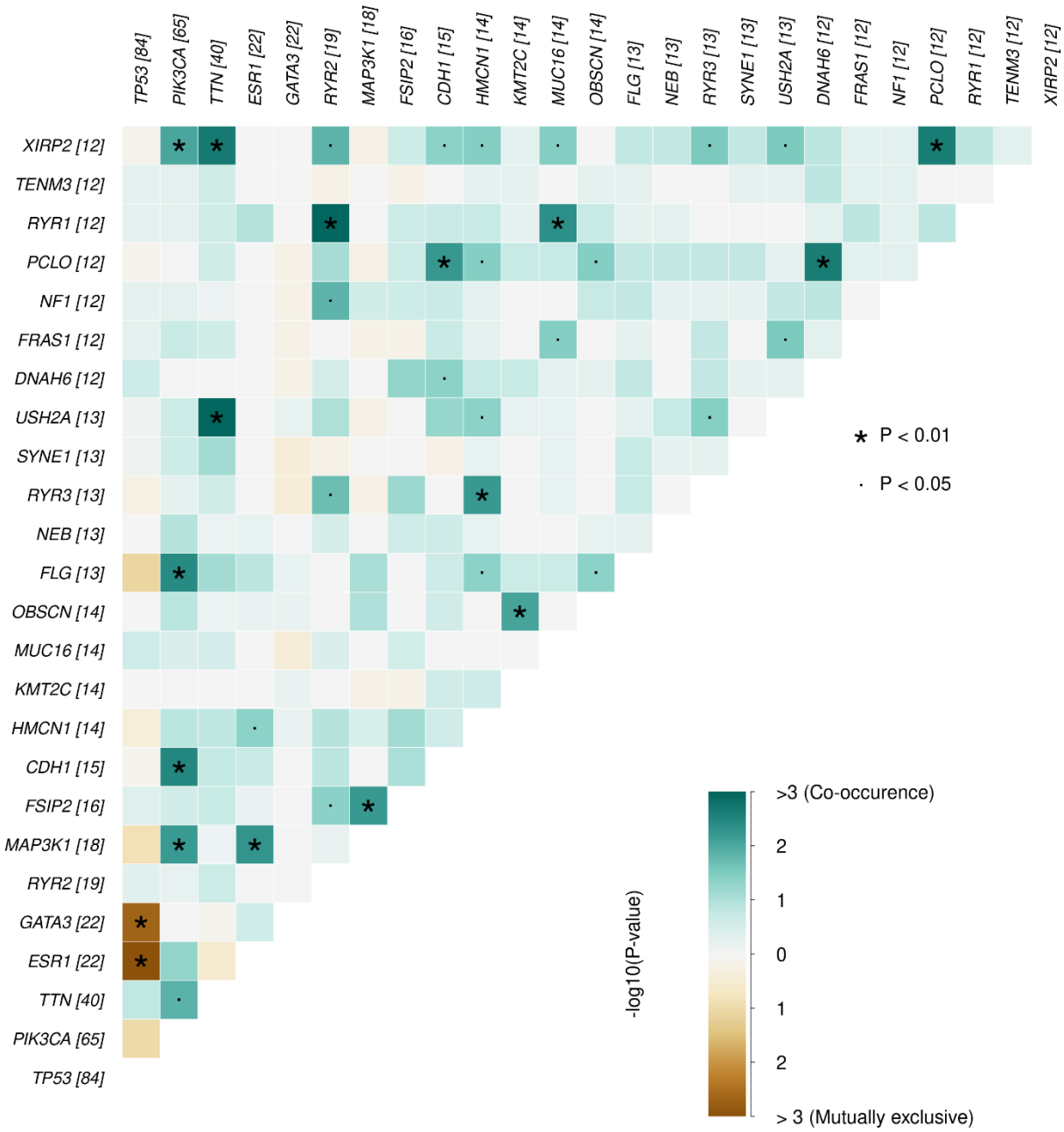


Figure 68. Somatic interactions in the IGR dataset. Cells are colored according to the -log10(p-value) value of the interaction. "*" indicates significance at p-value < 0.01, "." indicates significance at p-value < 0.05.

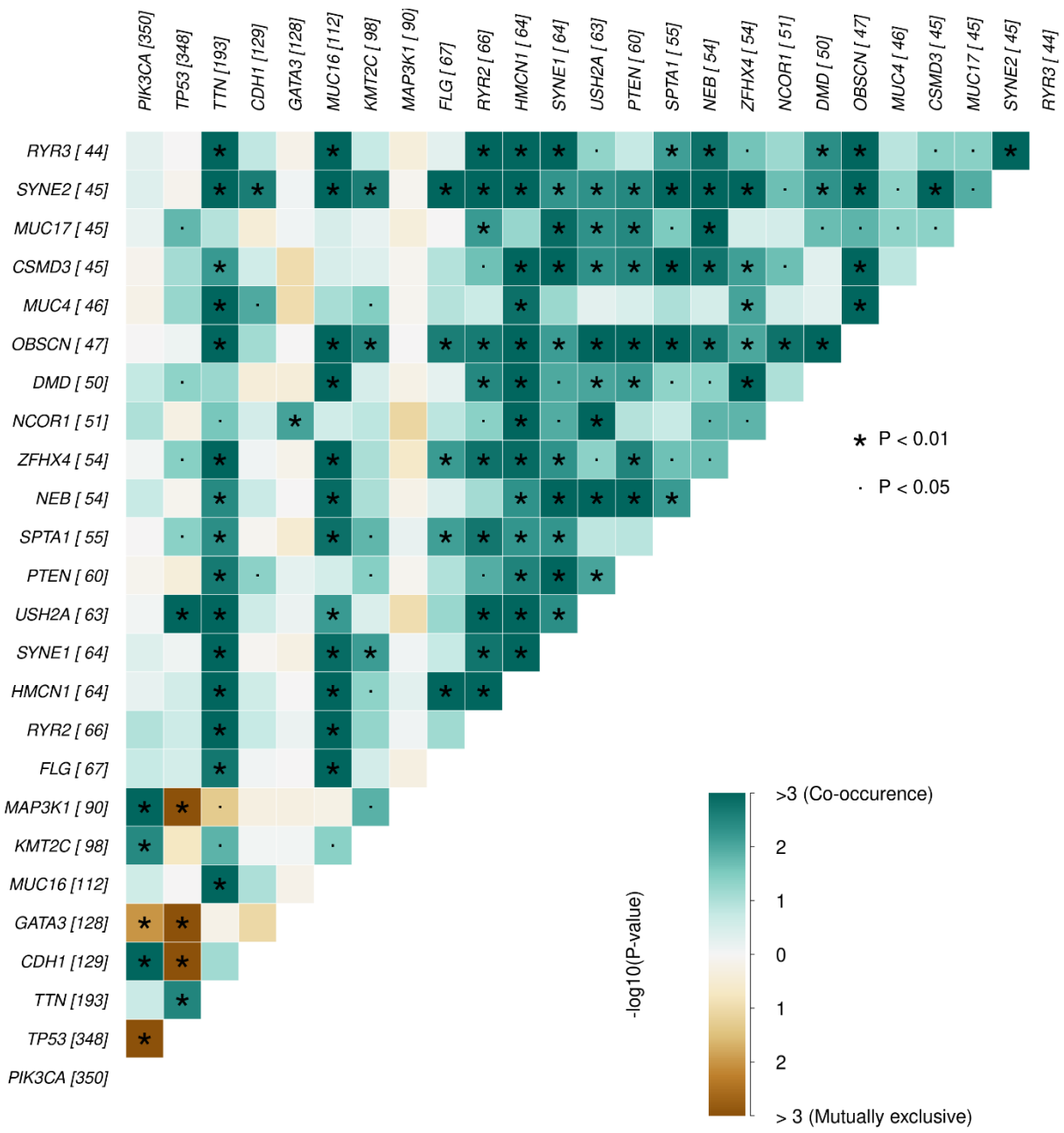


Figure 69. Somatic interactions in the TCGA dataset. Cells are colored according to the -log10(p-value) value of the interaction. "*" indicates significance at p-value < 0.01, "." indicates significance at p-value < 0.05.

Table 9. List of the somatic interactions unique to the IGR cohort. Pairs that are composed of the highly mutated genes common to both IGR and TCGA cohorts are marked with an asterisk.

Uploaded_gene1	Uploaded_gene2	Uploaded_pValue	Uploaded_oddsRatio
TP53	ESR1	9.41E-06	0
RYR1	RYR2	0.000157	13.89660277
TTN	XIRP2	0.002055	6.946560784
PCLO	DNAH6	0.002369	11.60126004
XIRP2	PCLO	0.002369	11.60126004
FLG	PIK3CA*	0.003767	5.651291323
MUC16	RYR1	0.004492	9.232182672
MAP3K1	ESR1	0.005223	5.457141315
CDH1	PCLO	0.005932	8.365449198
MAP3K1	FSIP2	0.00617	6.263740377
XIRP2	PIK3CA	0.009301	4.93900714
TTN	PIK3CA*	0.013836	2.447002779
NF1	RYR2	0.014792	6.037577563
XIRP2	RYR2	0.014792	6.037577563
FRAS1	USH2A	0.029244	6.193484783
XIRP2	RYR3	0.029244	6.193484783
XIRP2	USH2A	0.029244	6.193484783
FRAS1	MUC16	0.035994	5.609753505
PCLO	HMCN1	0.035994	5.609753505
XIRP2	HMCN1	0.035994	5.609753505
XIRP2	MUC16	0.035994	5.609753505
PCLO	OBSCN	0.035994	5.609753505
RYR2	FSIP2	0.042715	3.959250284
CDH1	DNAH6	0.04351	5.121885812
CDH1	XIRP2	0.04351	5.121885812
HMCN1	ESR1	0.04427	3.937973309

4.4.10 Differentially Mutated Pathways between IGR and TCGA cohorts.

To determine the pathways which are differentially mutated between these two datasets, hence might contribute to metastasis, I performed Fisher's exact test. None of the pathways tested passed the FDR threshold of 0.05. Hence, I treated any positive enrichment with lower OR > 1 and any negative enrichment with higher OR < 1 as significant. As a result, nucleotide excision repair (NER), DNA damage repair (DDR), NRF2, and WNT pathways were identified to be differentially mutated (**Figure 70, Table 10, Appendix Table 6**). Re-analysis after the exclusion of hypermutators from the IGR dataset, however, revealed completely different results (**Appendix Table 7**). First, none of the previous significantly mutated pathways were significant. The new significantly differentially mutated pathways were Mismatch Repair - Core, LIN37 independent P53 Targets, and U5 (**Appendix Table 7)**.

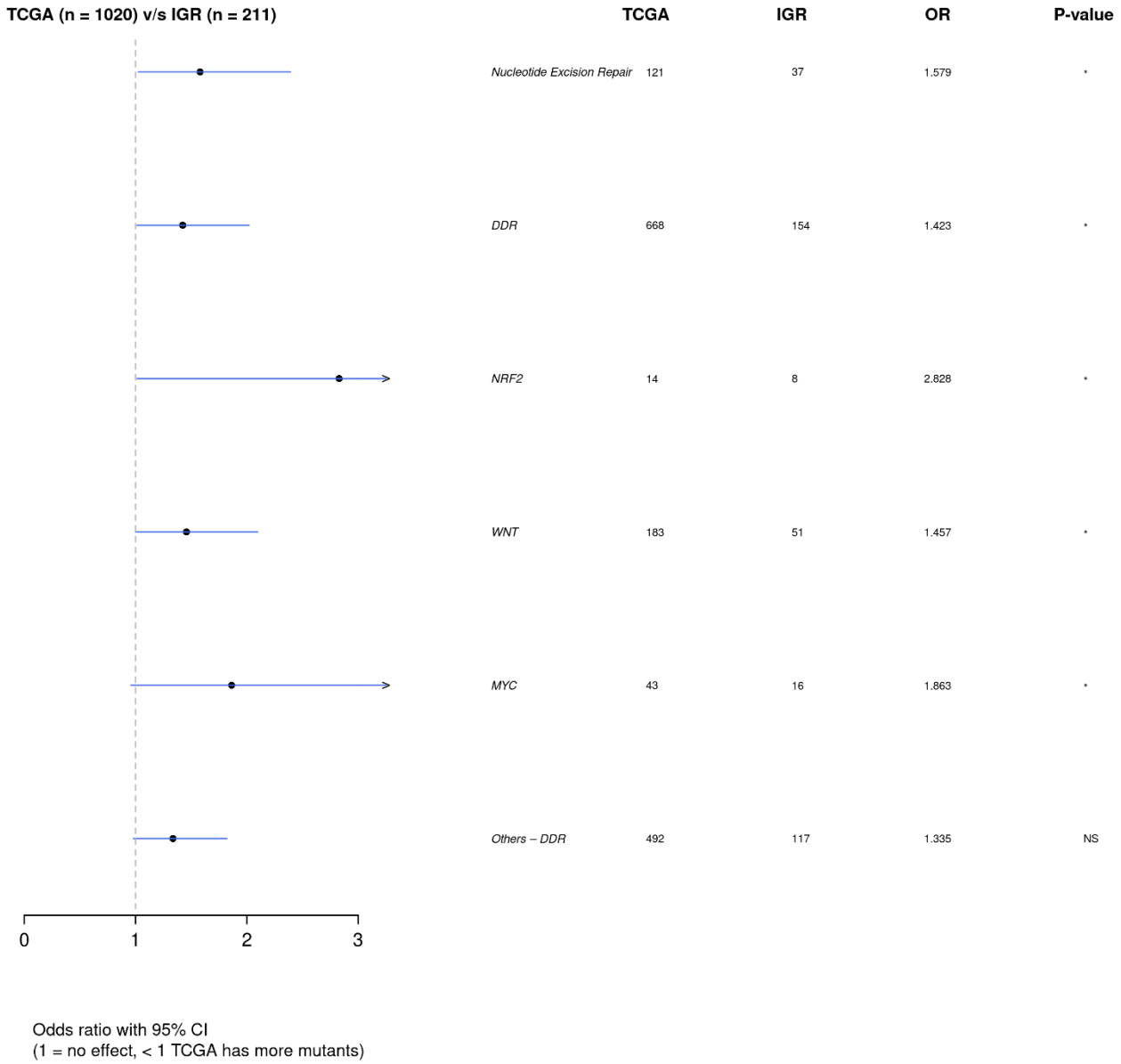


Figure 70. Forest plot of the genes differentially mutated pathway analysis between IGR and TCGA cohorts. The dot represents the OR, and if the confidence interval is out of the interval shown, an arrow is drawn. Number of the mutated samples in each dataset, along with the respective OR, is shown. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, NS $p \geq 0.05$.

Table 10. Detailed Fisher's exact test results of the pathway comparison

Pathways	IGR	TCGA	pval	or	ci.up	ci.low	adjPval
Nucleotide Excision Repair	37	121	0.031222	1.579234	2.391035	1.025241	0.549749
DDR	154	668	0.036919	1.423285	2.018543	1.013922	0.549749
NRF2	8	14	0.038797	2.828469	7.336885	1.014171	0.549749
WNT	51	183	0.042646	1.457408	2.097546	1.001611	0.549749
MYC	16	43	0.049599	1.863079	3.457933	0.959234	0.549749

4.4.10.1 Testing the Validity of the Differential Mutation of the DNA Damage Repair related Pathways with Mutational Signatures

Previously, in this thesis, it was shown that Signature 4 of the IGR dataset and Signature 7 of the TCGA dataset are caused by the deficiencies in the DSB repair. It was also shown that Signature 5 of the IGR dataset and the Signature 3 or the TCGA dataset had the etiology of deficient mismatch repair. Consequently, one can reason that the mean contribution of these signatures in the IGR dataset should be higher compared to the TCGA dataset. Indeed, when Student's t-test was carried out, I observed that for both repair pathways, IGR had a higher mean (**Figure 71**).

4.4.10.2 Testing the Validity of the Differential Mutation of the NRF2 Pathways with Mutational Signatures

Figure 72 displays the oncoplot of the samples which harbored a mutation in the NRF2 pathway. All of the samples, except one, have mutations either in KEAP1 or the CUL3, which are required to suppress the activity of the NFE2L2 (NRF2). Hence, one can expect that in the absence of these regulators NRF2 pathway will be constitutively active, resulting in lower than expected levels of the reactive oxygen species. Consequently, the contribution of the ROS related signature (Signature 3) to the mutational profiles of these samples was expected to be less. Indeed, when the contribution of signature 3 was tested between the NRF2 pathway mutant samples and

the rest, NRF2 pathway mutant samples displayed a highly significantly less contribution of Signature 3 to their mutational profiles (**Figure 73**). To further demonstrate the robustness of this result, I repeated the same procedure after subsampling the NRF wildtype samples to the sample sizes of 5, 10, 15, and 20, with 50 repetitions for each sample size. Starting from the sample size of 15, more than 70% of the tests became significant (**Appendix Figure 2**).

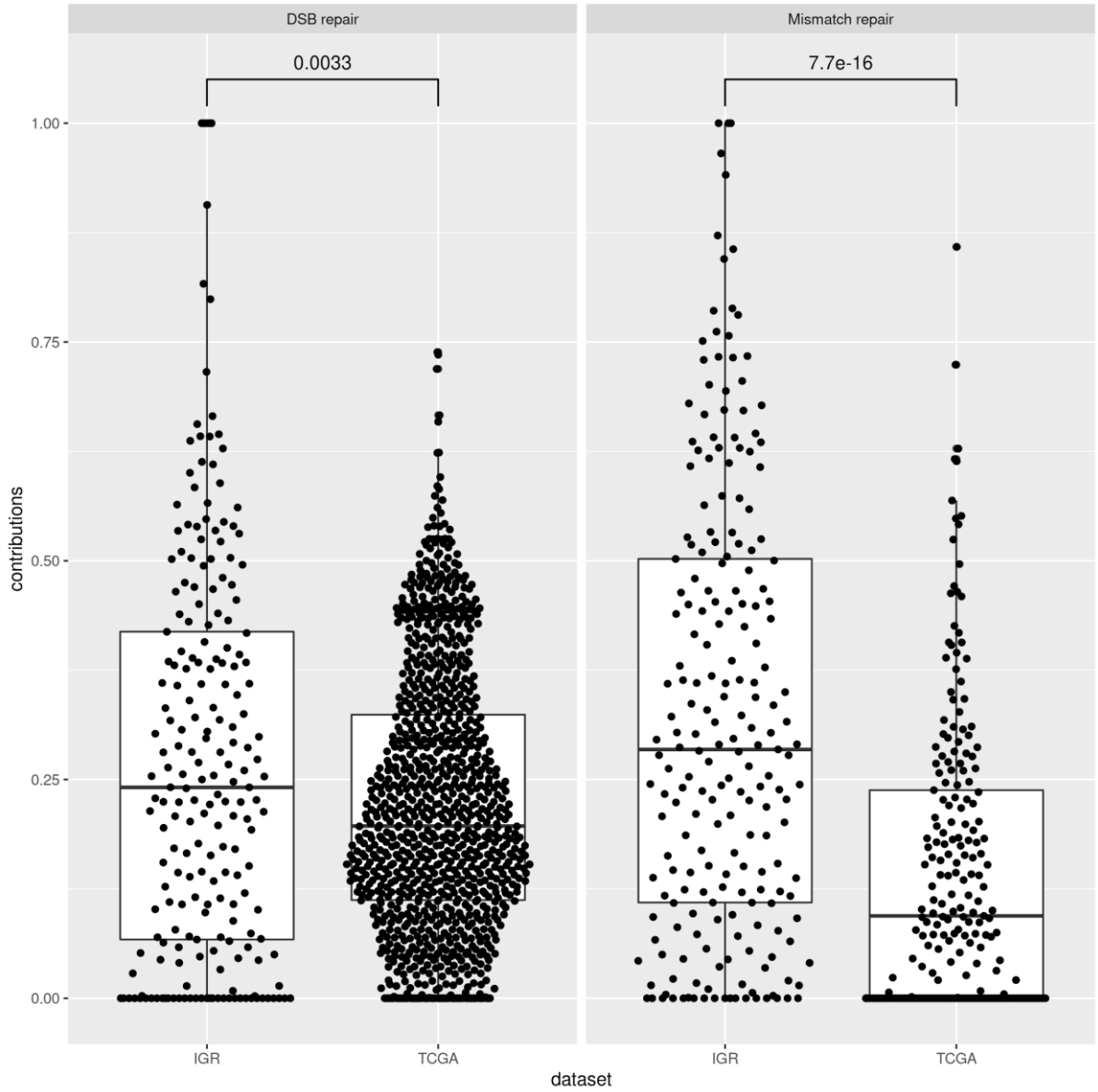


Figure 71. Comparison of the DNA damage related signatures between cohorts. Each dot represents a patient. Student's t-test results are also included in the graph. The upper and lower

edges of the boxes represent 75th and 25th quantiles, respectively, while a black rod represents the median. Whiskers extend until 1.5* IQR if there are observations extremum to those points, and till the minimum and maximum values otherwise.

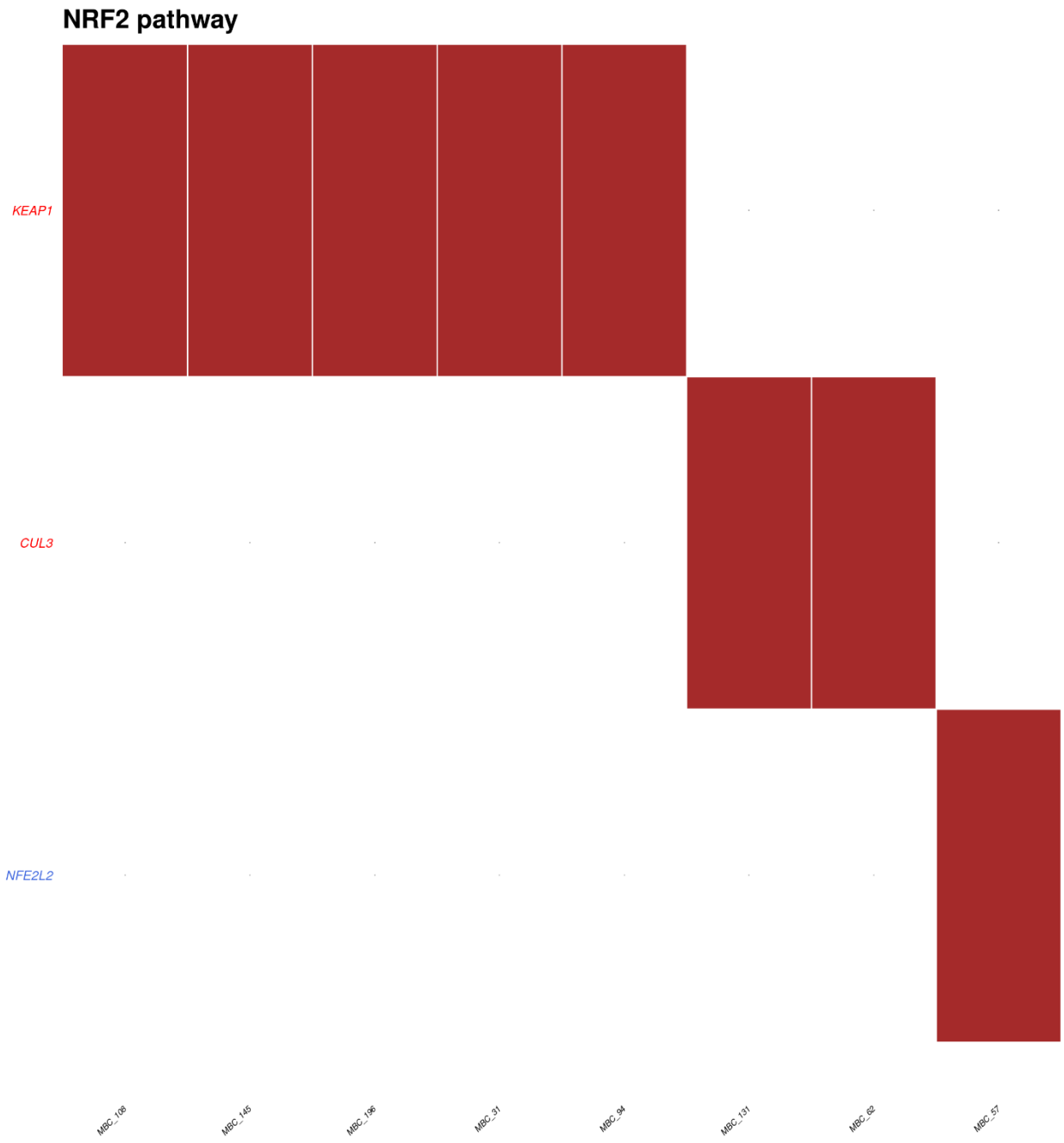


Figure 72. Oncoplot of the NRF2 mutated samples of the IGR dataset.

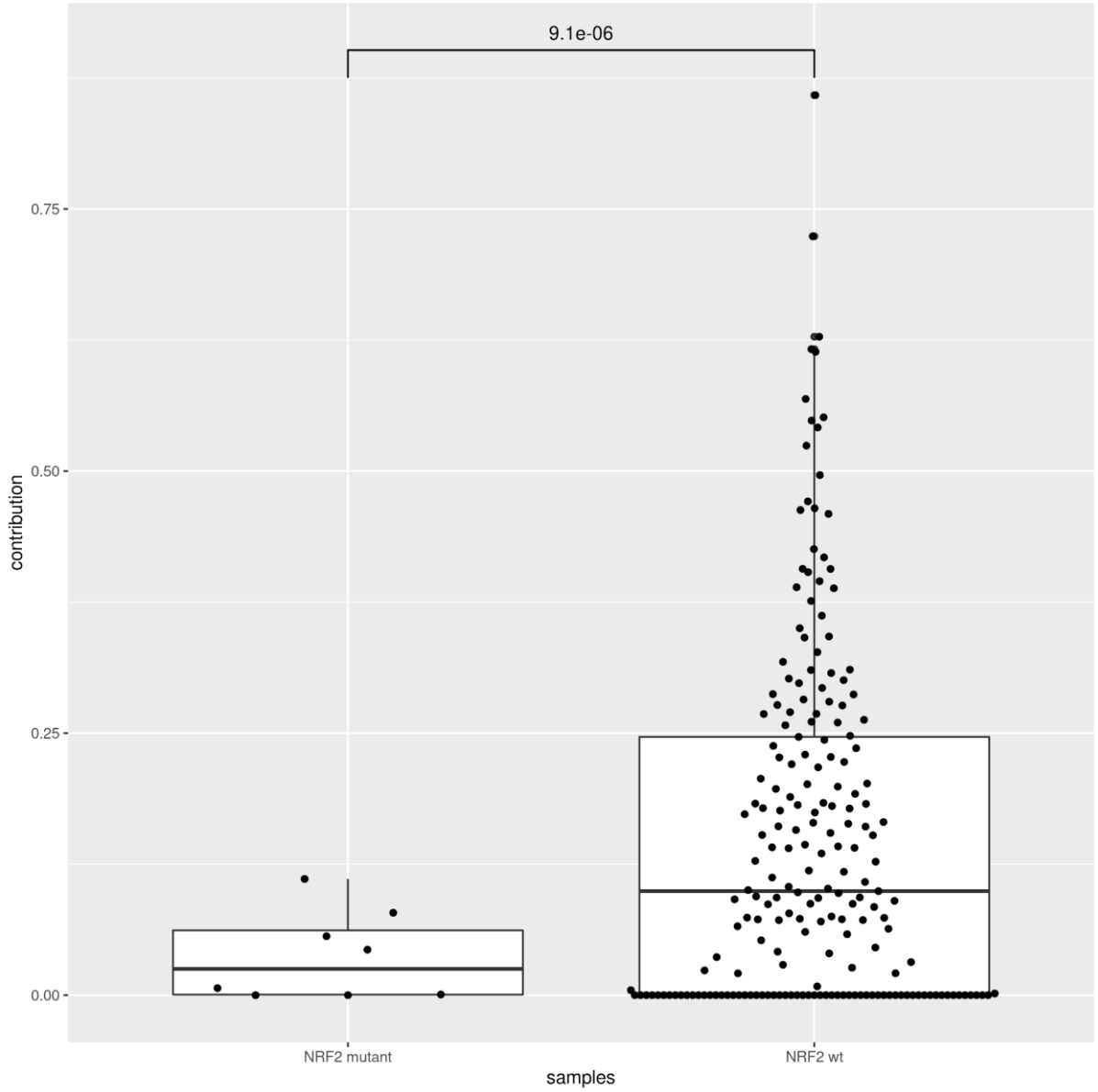


Figure 73. Comparison of the contribution of Signature 3 to NRF2 pathway wild-type and mutant samples. Each dot represents a patient. Student's t-test results are also included in the

graph. The upper and lower edges of the boxes represent 75th and 25th quantiles, respectively, while a black rod represents the median. Whiskers extend until 1.5* IQR if there are observations extremum to those points, and till the minimum and maximum values otherwise.

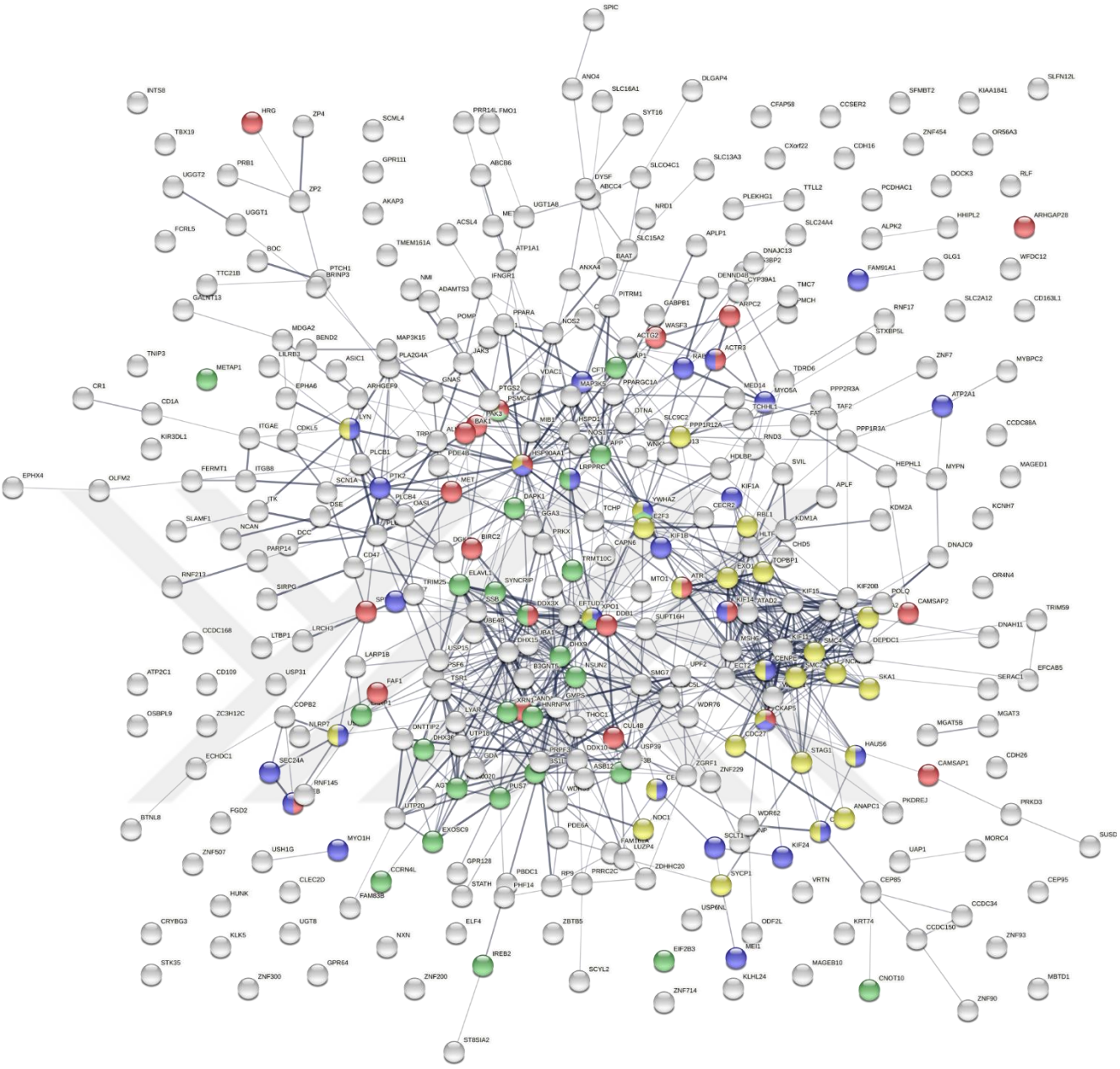
4.5 Case Study of the Network of CHRNA5 Somatic Interactions in TCGA-BRCA Dataset

CHRNA5 is one of the genes coding for the nicotinic acetylcholine receptor (nAChR) subunits in the human genome. nAChRs are formed from homopentamers or hetero heteropentamers nAChR subunits. CHRNA5 cannot form homopentamers and need to interact with other subunits to form a functioning receptor [140]. CHRNA5 gene variants have been implicated in susceptibility to nicotine addiction and lung cancer [141], [142]. A previous study on lung cancer cell lines demonstrated that CHRNA5 knockdown increased the motility of cells, calcium influx, and EMT. Investigation of this gene in our lab demonstrated that CHRNA5 knockdown could cause cell cycle arrest, DNA damage response, and increased sensitivity to chemotherapeutics [143]. Additionally, the latest work from our lab identified that the effects of CHRNA5 knockdown are TP53 dependent. The same work identified 14q32.31 miRNA cluster to be downregulated by siCHRNA5 treatment, adding another level of complexity to the pathway [144].

Considering the previous body of work, I identified co-occuring and mutually exclusive mutant genes of *CHRNA5* in the TCGA-BRCA cohort using BRCA:Cohort-vs-TCGA webtool. Additionally, I identified the genes that have a correlation with RNA expression of CHRNA5 with cBioPortal [99]. Finally, the significant somatic interactions (FDR<0.05) were split according to the sign of their correlation coefficient

(q-value<0.05) and fed into STRING webpage to generate protein-protein interaction (PPI) networks [145].

PPI networks for both positively and negatively correlated genes were significant (**Figure 74, 75**). The network of the positively correlated genes was enriched for regulation of protein-containing complex assembly, organelle localization, and posttranscriptional regulation of gene expression gene ontology (GO) terms. A reactome pathway for cell cycle was also enriched



Network Stats

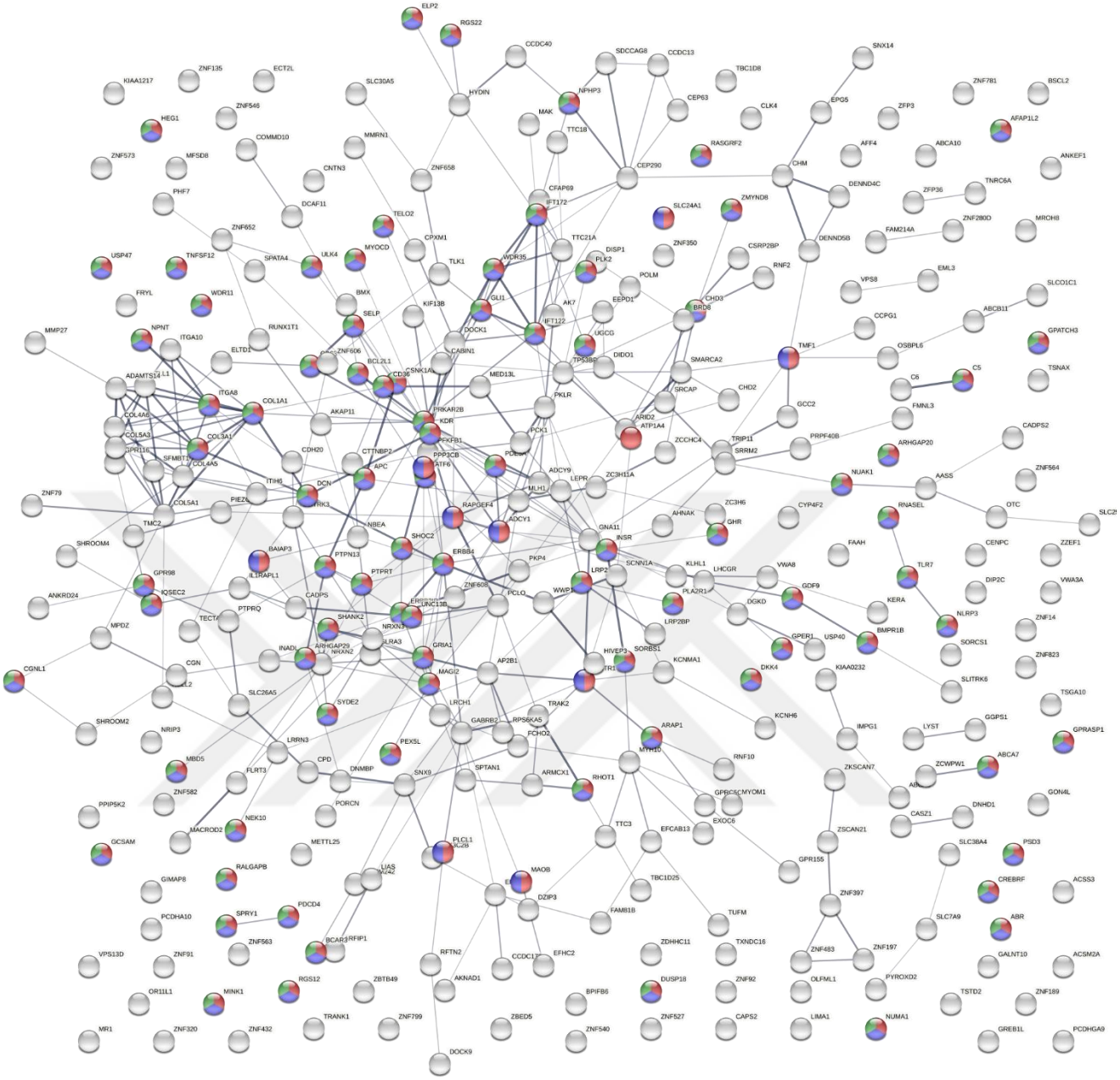
number of nodes:	351	expected number of edges:	631
number of edges:	824	PPI enrichment p-value:	1.36e-13
average node degree:	4.7	*your network has significantly more interactions than expected (what does that mean?)*	
avg. local clustering coefficient:	0.397		

Functional enrichments in your network

explain columns

	Biological Process (Gene Ontology)			
GO-term	*description*	*count in network*	*strength*	*false discovery rate*
GO:0043254	Regulation of protein-containing complex assembly	23 of 451	0.45	0.0442
GO:0051640	Organelle localization	28 of 598	0.42	0.0442
GO:0010608	Posttranscriptional regulation of gene expression	27 of 574	0.42	0.0442
GO:0009987	Cellular process	304 of 15024	0.05	0.0413

Figure 74. PPI network of significantly co-occurring and positively correlated genes. Each node represents a gene**,** and each edge represents PPI. Red nodes denote regulation of protein-containing complex assembly, blue nodes denote organelle localization, and green nodes denote posttranscriptional regulation of gene expression gene ontology (GO) terms [146]. Yellow nodes represent the cell cycle Reactome pathway [147].



Network Stats

number of nodes:	338	expected number of edges:	256
number of edges:	396	PPI enrichment p-value:	3.33e-16
average node degree:	2.34	*your network has significantly more interactions than expected (what does that mean?)*	
avg. local clustering coefficient:	0.314		

Functional enrichments in your network

explain columns

Biological Process (Gene Ontology)

GO-term	*description*	*count in network*	*strength*	*false discovery rate*
GO:0023051	Regulation of signaling	94 of 3553	0.19	0.0337
GO:0010646	Regulation of cell communication	93 of 3514	0.19	0.0337
GO:0009966	Regulation of signal transduction	84 of 3107	0.19	0.0366
GO:0050794	Regulation of cellular process	231 of 10932	0.09	0.0148
GO:0050789	Regulation of biological process	241 of 11475	0.08	0.0148
GO:0065007	Biological regulation	250 of 12171	0.08	0.0162

(less ...)

Figure 75. PPI network of significantly co-occurring and positively correlated genes. Each node represents a gene, and each edge represents PPI. Red nodes denote regulation of signaling, blue nodes denote regulation of cell communication, and green nodes denote regulation of signal transduction gene ontology (GO) terms [146].

(**Figure 74**). **Figure 74** indicated that there are three modules in the network. One of those modules was enriched for cell cycle genes, another one was enriched for the posttranscriptional regulation of gene expression genes, and the last one did not display any marked enrichment. The network of the downregulated genes had only a single module (**Figure 75**). This PPI network was enriched for the regulation of signaling, regulation of cell communication, and regulation of signal transduction GO terms. The module did not contain any enrichment for a particular GO term (**Figure 75**).

Considering the previous body of work regarding the functional interaction of *CHRNA5* and *TP53*, and the importance of DREAM complex in mediating the repressive functions of TP53, I identified the co-occurring genes of *TP53* in the TCGA cohort [129], [130], [144]. Afterwards, using the RNA-seq samples generated by the Dr. Ayşe Gökçe Keşküş , I performed both geneset enrichment analysis (GSEA) and clustering analysis using the components of the DREAM pathway and the LIN37 independent targets of TP53 (**Figure 76**) [148]. These RNA-seq samples were generated by the treatment of MCF7 cell line with siRNAs against TP53 (siP53) , CHRNA5 (siA5), or cotreatment with siP53 and siA5 [144].

The GSEA analysis of the *TP53* and *CHRNA5* co-occuring genes using the differentially expressed genes in each of the treatment groups did not yield any significant results (**Appendix Table 8, 9**). However, when the logFCs were clustered with their annotations, some patterns emerged. Firstly, the clustering of *CHRNA5* co-

occurring genes were indiscernible owing to size of this genelist, however, *CHRNA5* gene depleted clusters appeared which were mostly dominated by the components of the DREAM pathway and the LIN37 independent targets of TP53. A subset of the *TP53* co-occurring genes were densely clustered. This subset had negative logFCs in both siP53, and siA5 and siP53 treatments. As expected, the DREAM pathway and the LIN37 independent targets of TP53 formed clusters as well. Two of the clusters of DREAM pathway had high logFC values in siA5+siP53 treatment and no discernible change in siA5 only treatment. This result is in line with the fact that the DREAM complex silences its targets upon TP53 induction hence, its targets should be upregulated in the absence of TP53 (**Figure 76**).

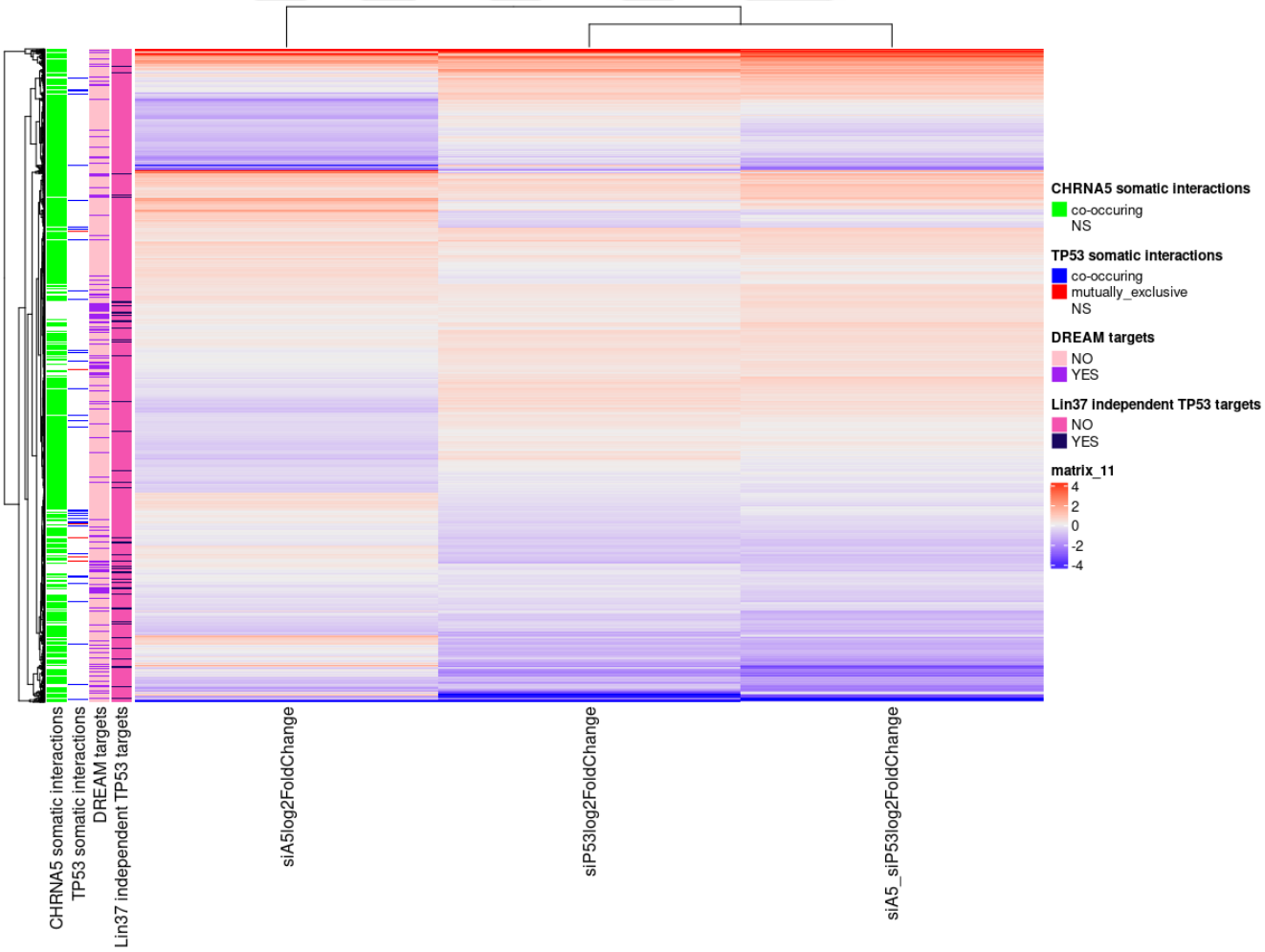


Figure 76. Hierarchical clustering of the differentially expressed genes ($p < 0.05$) in at at least one of the treatment groups, namely siA5, siP53 and siA5+siP53. Somaticaly interacting genes of the *CHRNA5* and *TP53*, as well as the components of the DREAM pathway and the LIN37 independent targets of TP53 were added as annotation.

4.6 Comparison of BRCA:Cohort-vs-TCGA App to Similar Apps

I have also compared, feature by feature, the existing apps with similar functionalities with the BRCA:Cohort-vs-TCGA app, presented in this thesis (**Table 11**). One of the most outstanding attributes of BRCA:Cohort-vs-TCGA was its comparative ability. For example, BRCA:Cohort-vs-TCGA is the only app that offers signature comparison, clinical correlate identification, and differentially mutated gene identification functionalities. BRCA:Cohort-vs-TCGA also performed well in *de novo* signature and driver gene identification, matched only by the MutScape app. BRCA:Cohort-vs-TCGA also has the widest functionality spectrum among the compared apps. However, BRCA:Cohort-vs-TCGA does not currently offer functionalities to accept multiple input formats as well as for the fitting of known signatures (**Table 11**).

Table 11. Comparison of features of BRCA:Cohort-vs-TCGA app with other apps.

	Multiple Input Format	De Novo Signature Analysis	Signature Fitting	Signature Comparison of Uploaded dataset with TCGA	Signature vs Clinical Correlates	Pathway Based Comparison	Differential Mutation Analysis	CoMut Plots	De Novo Driver Gene identification	Lollipop plots	Data Filtering
BRCA:Cohort-vs-TCGA	No	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
MutScape [109]	Yes	Yes	No	No	No	No	No	Yes	Yes	No	Yes

CoMut [111]	No	No	No	No	No	No	No	Yes	No	No	No
vcfView [110]	No	No	Yes	No	No	No	No	No	No	Yes	Yes
MuSica [112]	Yes	No	Yes	No	No	No	No	No	No	No	No
cBioPortal [99]	No	No	No	No	No	No	No	Yes	No	Yes	Yes
PathwayMapper [75]	No	No	No	No	No	Yes	No	No	No	No	No

5 Chapter 5 DISCUSSION

5.1 Case Studies

In this thesis, I developed an app for the interactive exploration of a breast cancer dataset and comparison with the TCGA-BRCA cohort. I have used a previously published metastatic breast cancer dataset to benchmark the app [127]. Additionally, a network based study of the CHRNA5 somatic interactions was also carried out

5.1.1 Mutational Signature Analysis

Authors of the original article have taken several approaches for the signature detection. First, using the Welcome Trust Sanger Institute's method for signature detection which automatically decides on the number of signatures, they identified only two signatures [149]. Even though this frame works well for the cohorts with big sample sizes, it rids one from the freedom of choosing the number of signatures [150]. Unsatisfied with the result, they tried to merge their cohort with the TCGA and redid the signature extraction process, which did not yield any unique signature for the IGR dataset [127]. In another attempt, authors fit the already known signatures to the mutational profiles. In none of these attempts they identify a unique signature for the IGR cohort.

Our pipeline, on the other hand, identified a signature specific to the IGR cohort, which was similar to the signature SBS17b of the COSMIC database. Another study conducted on the metastatic breast cancer cohort identified that signature 17 was enriched in the metastatic breast cancer samples who had received treatment, however do not describe

this signature as novel [151]. Along the same lines, another larger cohort identified that the signature 17 was enriched in the HR+ metastatic breast cancer compared to their primary counterparts [152]. In this regard, BRCA:Cohort-vs-TCGA outperformed the IGR study by detecting the signature SBS17b, using a comparative strategy.

Another finding of the original IGR study was that the deficient DNA damage repair related signatures were contributing more to the metastatic samples. Our study could recapture this relationship, and showed that both SBS3 and SBS6 related signatures contributed differentially to metastatic and primary breast cancers. Literature on the other hand is conflicting both with our results and with each other. SBS3 has been found to contribute differentially in both [151] and [152]. Angus et al., [151] reported this difference to be significant only in the TNBC samples, meanwhile another study by Bertucci et al., [152] reported it to be true only for the HR+ samples. Additionally, the latter [152] reported that presence of signature 17's contribution was a bad prognostic marker for HR+ patients.

One type of analysis that BRCA:Cohort-vs-TCGA attempted but the original IGR study did not was to find the clinical correlates of the mutational signatures. Of these analyses, I first measured the relationship between signature contributions and TMB. As expected, the app identified a strong correlation between Signature 1, 2 and TMB. These signatures were APOBEC related signatures, which have been associated with the hypermutation and kataegis, a phenomenon that results in clusters of point mutations in close proximity [153]–[156]. More surprising observation however was the negative correlation of Signature 5, defective mismatch repair associated signature, with TMB. This might be explained by stating that defective DNA repair associated signatures

forms the baseline mutagenetic process, hence whenever they are absent, they are replaced by a stronger mutagenetic process, such as APOBEC related ones.

Interestingly, I identified some relationships between subtypes and signatures which were overlooked both by the original IGR study and the two other big cohort studies [127], [151], [152]. First, when compared with HR+ and HER+ metastatic breast cancers, TNBC samples displayed a higher contribution of signature 3, SBS17b, which was the novel signature identified in our analysis. Signature 2, an APOBEC related signature, on the other hand was more enriched in HER+ compared with the TNBC samples. A similar result was the identified when TNBC and HR+ samples were compared. HR+ samples were enriched for both of the APOBEC related signatures and Signature 4. There were no differences in signature profile of HR+ and HER+ samples, possible due to the underrepresentation of HER+ in the IGR cohort which was the subtype with the least number of samples (14 samples).

5.1.2 Differentially Mutated Genes and Cancer Driver Genes.

The original IGR study only considered 1569 genes for the differential mutations analysis, meanwhile BRCA:Cohort-vs-TCGA considered all of the 6184 genes. The only two common differentially mutated genes between the analyses were *ESR1* and *FSIP2*. Most of the differentially mutated genes identified as significant in the original IGR study disappeared from our analysis simply because the additional ~ 300 new TCGA samples contained mutations for these. This shows the power of a comparative approach of BRCA:Cohort-vs-TCGA. One of such genes was AGRN. Genes unique to our analysis were missed by the original study because they were not part of the 1569 genes considered (**S6 Table**, [127]). When compared with the two other studies [152] and [151], the only common differentially mutated gene was *ESR1*. These larger studies,

[152] and [151] on the other hand, had ESR1, *AKT1*, *NF1*, *KMT2C* and *TP53* in common between them. The small size of the IGR cohort compared to the other datasets might hinder the identification of new differentially mutated genes.

FSIP2 has been associated with other types of cancer. For example, a study conducted on the Paget's disease patients identified that this gene was mutated in 29 % of the cases, making it the fourth most frequently mutated gene in that cohort [157]. Another study cohort of the esophageal squamous cell carcinoma patients annotated this gene as a driver gene [158]. Interestingly, *FSIP2* mutant skin cutaneous melanoma patients had higher TMB and a better response to the immune checkpoint inhibition therapy [159]. The most interesting study however was the one conducted on the primary and relapse samples from the patients with non-small cell lung cancer. *FSIP2* was observed to be differentially mutated but the difference was not significant, owing to the small sample size [160].

ARMCX4 was yet another differentially mutated gene unique to my re-analysis. Interestingly, a cohort study on the Chinese cervical cancer patient primary samples and metastasis samples also identified this gene as the highly mutated gene in both the primary and metastatic samples although this gene was not differentially mutated between them [161]. Additionally, in a cohort study of germline lung adenocarcinoma, *ARMCX4* gene was found to be a driver mutation in when the somatic mutation profile of these patients were analyzed [162].

The last differentially mutated gene identified in this thesis was *MALRD1*. This gene was previously identified as one of the top ten driver genes in relapsed multiple myeloma [163]. A copy number loss of this gene has also been associated with

endometriosis, a disease which can develop into cancer [164]. Additionally, copy number loss of this gene has been implicated in the TCGA glioblastoma cohort [165].

With my approach, I identified *ESR1* and *GJA8* as cancer driver genes. Both of these genes were also identified as cancer driver genes by the original authors [127]. Overall, all of the ESR1 mutations fell within the ligand binding domain, with residues 538-537 and 380 standing out. First of such mutations were first described in a metastatic breast cancer sample. In this sample the ER could be active without the need for stimulation by the ligand [166]. Afterwards, these mutations were identified in many cohort studies conducted on metastatic and endocrine therapy resistant breast cancers [151], [152], [167, p. 1], [168, p. 1].

None of the large cohort studies annotated the *GJA8* gene as cancer driver gene. This gene is part of the connexin family of proteins which are constituents of gap junctions [169]. As part of the gap junctions, this protein might help to preserve the tumor integrity. The only study which investigated mutant connexin cx43 used mouse as their model. By crossing cx43 mutant and *HER2* overexpressing mice they were able to show that the progeny displayed a very high propensity towards lung metastasis [170]. Other studies were mostly concerned with the expression profiling, and they agree that overexpression of these proteins can lead to metastasis [169]. These show that our app may help for identification of driver genes that might be associated with a clinical parameter.

Considering the influence hypermutators can have on aforementioned analyses, because of the presence of vast numbers of variants of unknown significance in their genome, All of these analyses were repeated after exclusion of these samples. A segmented linear

regression based approach was used to determine the optimal TMB, identified to be 167 (**Appendix Figure 1**), cutoff for hypermutator detection.

Re-analysis of the differentially mutated genes identified exactly the same genes with the previous analysis, except for *ARMCX4*, because most of its mutations were in the hypermutant samples. *ARMCX4* can as well be a driver gene specific to hypermutators or a mere passenger gene. A recent publication devoted to the hypermutator samples tried to identify driver mutations specific to the hypermutators in *POLE* and *POLD1* genes [133]. One of their filters was the absence of the candidate mutation in non-hypermutated samples, for which *ARMCX4* mutations qualify, with the exception of K160N.

In the reanalysis of the driver genes, *GJA8* lost its significance. This was due to the loss of one of the three samples bearing E110K mutation to the hypermutation filtration.

5.1.3 Determination of Optimal VAF cutoff

In our app I offer a tab which makes the dataset comparison easy and statistically sound. This tab is especially helpful when one wants to compare two datasets generated at different sequencing depths. This is motivated by the fact that a protocol with higher depth can capture rare variants and also might give rise to more artifacts, clustered in the rare VAF cutoffs. As a result, one can reason that at a sufficiently large VAF cutoff, both high and moderate depth protocols will have a comparable power to detect the mutations. In literature, the mainstream approach is to compare the TMB of the datasets. If the TMB profile of the two datasets are comparable authors decide that these two cohorts are comparable [127], [138]. An alternative approach however is to re-process both datasets with the same pipeline. The two big metastatic breast cancer cohorts for

example, did not employ any cutoff as they sequenced their samples at depth similar to TCGA and called variants with an analogous pipeline [151], [152].

The preprocessing of the IGR mutation files included removal of any variants with VAF less than 0.1. Hence our VAF comparisons till this point affected the TCGA dataset only. Interestingly, between the VAF cutoffs 0.1 and 0.15 we see a significant difference in the TMB of the two datasets which disappear at 0.2. However, for comparison purposes this threshold might be considered high as many subclonal mutations would be lost. When I checked the two large cohort studies of the metastatic breast cancers, both of these studies concluded that the metastatic samples exhibit a higher TMB compared to their primary counterparts [151], [152]. This explains why a VAF cutoff of 0.2, a high one, would be needed to equalize TMBs of these two datasets. If differential VAF cutoffs were applied to each dataset, 0 to TCGA and 0.1 to IGR as the original analysis, this difference also disappears. Hence, these results show an inherent limitation of this approach, namely if two datasets have markedly different TMBs, a high VAF cutoff will be required to normalize them.

5.1.4 Clinical Enrichment Analyses

In the clinical enrichment analyses, I identified four genes related to three metastasis sites. Out of those, TTN will not be described here because it was found to be depleted in the “other” category. MAP3K1 gene was associated with liver metastasis, meanwhile *IGSF10* and *DYNC2H1* were associated with skin metastasis at p-value < 0.05. None of these genes were identified as differentially mutated by our analysis, by the analysis of the original authors, or by the two large cohort studies [127], [151], [152]. These studies focused their attention to subtype specific comparisons, hence metastatic site related comparisons are not present. However, a meta-analysis study considered all of

the aforementioned studies and tried to address this question [171]. In this study as well, none of the above mentioned genes were associated with the organotropic tendencies. This might be a direct consequence of using p-value threshold instead of FDR threshold. However, the small sample size of this study was limiting discovery of new associations.

The second kind of enrichment analysis I performed was concentrated on the HR and HER subtypes. In this analysis the results yielded significant FDR values hence, FDR cutoff was used. As a result, I identified *ESR1* and *PIK3CA* mutations to be enriched in HR+ samples, and *TP53* to be enriched in the TNBC samples. These results are not surprising and align well with the literature. Enrichment of ESR1 mutations in HR+ samples indicate that these are not caused by mere genetic drift, as would be expected if the *ESR1* was enriched in TNBC samples. Previously, *ESR1* mutations have been identified in endocrine resistant breast cancers [167, p. 1], [168], [172]. Association of the *TP53* mutations with TNBC and *PIK3CA* mutations with HR+ samples have also been demonstrated before [80].

Subtype specific clinical enrichment analysis conducted after the removal of the hypermutator samples yielded two additional enrichments for *TTN* and *MYH4* genes in HR+ samples. These results are interesting because one expects that the mutation number of the long genes will drop after removal of the hypermutators [173]. However, some studies highlight the potential importance the long genes might have. For example, Hofree et al., using their network based approach showed that removal of the long genes with mutations resulted in the worse performance of their algorithm in cancer subtype prediction [174]. Yet another study reported that in the cancer related pathways are enriched with long genes [175]. Clinical enrichment of *TTN* and *MYH4*,

especially after the exclusion of hypermutators, warrants further study. In the reanalysis of the clinical enrichment by biopsy site I observed no difference with the previous analysis.

5.1.5 Somatic Interactions

This kind of analysis was not performed by the original IGR study. When BRCA:Cohort-vs-TCGA used an FDR based cutoff the app identified two somatic interactions, one between *TP53* and *ESR1* which was mutually exclusive, and another between *RYR1* and *RYR2* which was co-occuring. *TP53* and *ESR1* mutual exclusivity has also been identified by large cohort studies, but *RYR1* and *RYR2* interaction is unique to this [127], [151], [152]. Interestingly, I have not found any report regarding the co-occurrence of the RYR genes, and this novel finding warrants further study. However, a re-analysis after exclusion of hypermutator samples, resulted in the loss of interaction between the RYR genes, supporting the notion that it was the length of these genes that drove their interaction and mutation.

5.1.6 Differentially Mutated Pathway Analysis

In our analyses it was found that DNA damage related pathways, NRF2 pathway and WNT pathways were significantly highly mutated in metastatic breast samples. Using a p-value threshold instead of an FDR threshold undermines the robustness of this analysis. In order to justify these pathways, I conducted additional analyses.

Both my analyses and the results from the large cohort studies as well as the original analyses of the IGR cohort showed that the contribution of DNA damage related signatures to the mutational profiles of the metastatic breast cancer samples were markedly higher than their primary counterparts [127], [151], [152]. None of the large

cohort studies attempted to conduct a differentially mutated pathway analysis. One of the studies reported that the mutations in the DDR genes are associated with worse survival outcomes [176]. These results justify the use of p-value threshold in the case of the DDR related pathways and highlights the importance of considering pathways as a whole.

In the case of the NRF2 pathway, all of the mutations identified except one were in CUL3 and KEAP1 genes, all of which are the inhibitors of NFE2L2 and the NRF2 pathway [177]. First, to show that these mutations had functional consequences, I assessed the contribution of SBS17b signature in NRF2 pathway mutant and wild-type samples. SBS17b can be caused by the reactive oxygen species (ROS) **SBS**

present in cells [178]. Hence, it can be stated that the samples with the NRF2 pathway mutations will constitutively activate NRF2 activity and thus a reduced ROS signature [177, p. 2]. Indeed, NRF2 mutant samples show a markedly reduced contribution of SBS17b to their mutational profile. NRF2 activation has been associated with breast cancer proliferation and migration before. For example, a study identified that overexpression of *NFE2L2* or knockdown of *KEAP1* increases the proliferation and migration of the breast cancer cell lines [179]. Yet another study showed that NRF2 activation in breast cancer resulted in lower levels of ROS and NRF2 knockdown markedly reduced lung metastasis [180]. Overall, it can be concluded that the NRF2 pathway mutations in the IGR cohort may well be functional, and various *in vitro* studies indicate that this pathway might indeed be differentially mutated in metastatic breast cancer.

Unfortunately, activity of the WNT pathway could not be assessed due to lack of reports associating it with mutational signatures. Nonetheless, the pathway itself has been

associated with breast cancer pathogenesis before. For instance, a recent paper associated WNT pathway activity in breast tumors with inflammation and subsequent metastasis [181]. A similar study conducted on HER2+ breast cancer cell lines concluded that the WNT activation led to EMT and resulted in spreading of the cells [182]. As a result, it can be concluded that the increased mutation of WNT pathway components in metastatic breast cancer samples does not contradict the findings of literature.

In contrast, none of the above mentioned pathways were found to be significantly differentially mutated. Mismatch Repair - Core, LIN37 independent P53 Targets, and the U5 pathway components on the other hand, were found to be significantly less mutated in the IGR cohort. First, it should be mentioned that the hypermutator filtration was not applied to the TCGA dataset. As a result, these new findings might simply be an artifact of having more hypermutators in the TCGA-BRCA cohort (Please also note that these pathways have an OR < 0). Nonetheless, it was previously shown that the mutations in the mismatch repair pathway genes can lead to the hypermutator phenotype both in sporadic and inherited cancer types. Hence, filtration of hypermutators from the IGR cohort can lead to the decreased mutation of the components of this pathway. LIN37 independent P53 targets pathway was curated by the analysis of gene expression datasets [143]. As a result, this pathway has hundreds of components. Even though the gene expression based target curation can enhance the power of activation detection, let's say for a transcription factor, use of such gene sets in the mutation analyses might not be appropriate. Because multiple mutations of this RNA expression based pathway in the same patient would be required to reach the same effect as a single mutation in a gene set composed of the actual regulators. In the end, it is possible that the observed results, possibly with the exception of the Mismatch

Repair - Core, for the reanalysis is the mere artifact of not having hypermutators in the IGR dataset while having them them in the TCGA-BRCA dataset

5.1.7 *CHRNA5* Case Study in the TCGA-BRCA dataset

I observed that both negatively and positively correlated genes formed significant PPI networks. It is worth noting that all of the genes in these networks are co-occurently mutated with CHRNA5. Positively correlated gene network exhibited three modules, one enriched for cell cycle genes, such as *CENPA*, *KIF11*, *KIF15*, and *KIF14*, another one enriched for posttranscriptional gene regulation genes, such as *DHX9* and *ELAVL1*, and the last with no marked enrichment. Most of the previous studies are in agreement that the co-occurently mutated genes should belong to separate pathways [68]–[71]. Given the previous observations that CHRNA5 knockdown arrests the cell cycle, additional mutations in cell cycle genes might relieve this arrest and be of functional importance if one assumes CHRNA5 mutations are loss-of-function mutations. Other than cell cycle pathway and posttranscriptional gene regulation pathway, pathways enriched in these networks are all too general to suggest a mechanism of action.

Observing that the somaticly interacting genes of *CHRNA5* and the genes that positively correlate with *CHRNA5* can form a meaningful network of PPI, I also analyzed their enrichment in previously generated RNA-seq dataset of siA5, siP53, and siA5 and siP53 co-treated MCF7 cells [144]. In the light of the findings regarding the functional interaction between *CHRNA5* and *TP53* genes, and the importance of the DREAM complex in repression of the various TP53 targets, we also included the components of the DREAM pathway, the LIN37 independent targets of TP53 as well as the somatically interacting genes of *CHRNA5* and *TP53* in our analyses [129], [143], [144].

The GSEA analysis of the *CHRNA5* and *TP53* somatically interacting genes in the knockdown experiments did not yield any significant results, which prompted me to perform a clustering analysis. Clustering results however displayed interesting patterns. Firstly, the clustering of *CHRNA5* co-occurring genes were indiscernible owing to the sheer size of this genelist, however, *CHRNA5* gene depleted clusters appeared which were mostly dominated by the components of the DREAM pathway. A subset of the *TP53* co-occurring genes were densely clustered. This subset had negative logFCs in both siP53, and siA5 and siP53 treatments, indicating their functional relevance with TP53. This observation indicated that abolishment of TP53 induces the expression of these genes. As expected, there was no upregulation of these genes in the siA5 treatment, since siA5 is known to induce TP53 rather than suppressing it.

Overall, it can be concluded that concurrent analysis of the somatically interacting genes and co-expressing genes of a target gene might reveal new insight as was the case in CHRNA5 PPI network, and *TP53* expression in siP53 and siP53+siA5 treatments

However, the existence of only 5 patients with *CHRNA5* mutations makes the inference hard. Additionally, three of these mutations displayed very low VAF, all below 0.05, and the remaining 2 mutations were in the samples with high TMB (**Appendix Table 10**).

6 FUTURE PERSPECTIVES

6.1 Development of BRCA:Cohort-vs-TCGA

BRCA:Cohort-vs-TCGA webtool allows for the exploration of the input breast cancer cohort and its comparison with the BRCA-TCGA cohort. The most important missing feature is the option to allow the user to perform filteration on TCGA dataset. For example, the user might be interested in HR+ breast cancers only and the proper filteration of the TCGA-BRCA dataset would be necessary for appropriate comparison. The better visualization of the pathways, reminiscent of PathwayMapper, can be added to provide the user with mechanistic insight, such as activation or repression dependencies between the pair of nodes. [75]. In the future, additional algorithms for somatic interaction detection [63]–[66], and mutational signature detection [41], [150], [183] can be implemented to offer complementary approaches and increase the power of detection. The same framework used in TCGA-BRCA analysis can, in principle, also be extended to other TCGA cohorts. Moreover, annotations for the known driver genes such as the ones from OncoKB [184], and variant effect prediction results such as the ones from dbNSFP, VEP, and PolyPhen [23], [185], [186] can be added to further help the users with variant/gene prioritization.

More fundamental changes however would be support for multiple input types, such as VCF and ANNOVAR outputs, functionalities to perform the quality control of the variants, and the option to perform the comparison of the two uploaded datasets [24], [187].

6.2 Potential Case Studies with BRCA:Cohort-vs-TCGA

The possibilities are virtually endless. For example, one can use the IGR cohort as a discovery cohort followed by validation analysis in Angus et al., and Bertucci et al., cohorts [127], [151], [152]. Another possibility is combination of the different cohorts such the ones from TCGA and ICGC, to increase the power for the discovery of cancer driver genes, somatic interactions, and clinical enrichments [80], [188]. This combination, however, does not need to be restricted to the same cancer type or subtype, as exemplified by the TCGA PanCancer studies [26], [126], [189].

I restricted the custom gene analysis to *CHRNA5* and *TP53* genes. However, special genes of interest, such as *FSIP2, ARMCX4, ESR1* and *MALRD1* can further be explored in the IGR dataset as well as in other metastatic breast cancer cohorts in which an expression profiling was conducted, to identify and prioritize their somatically interacting genes with meaningful expression patterns. A similar analysis in myelodysplastic syndrome patients, for example, identified that gene regulatory networks of the mutually exclusive driver genes in the spliceosome pathway did not overlap [190]. Additionally, both univariate and multivariate survival analysis with the differentially mutated and/or driver genes will provide an orthogonal approach to assess their significance.

7 REFERENCES

[1] D. Hanahan and R. A. Weinberg, “Hallmarks of Cancer: The Next Generation,” *Cell*, vol. 144, no. 5, pp. 646–674, Mar. 2011, doi: 10.1016/j.cell.2011.02.013.

[2] Y. Samuels *et al.*, “High Frequency of Mutations of the PIK3CA Gene in Human Cancers,” *Science*, vol. 304, no. 5670, pp. 554–554, Apr. 2004, doi: 10.1126/science.1096502.

[3] M. Hollstein, D. Sidransky, B. Vogelstein, and C. C. Harris, “p53 Mutations in Human Cancers,” *Science*, vol. 253, no. 5015, pp. 49–53, Jul. 1991, doi: 10.1126/science.1905840.

[4] R. J. Wellinger, “In the End, What’s the Problem?,” *Mol. Cell*, vol. 53, no. 6, pp. 855–856, Mar. 2014, doi: 10.1016/j.molcel.2014.03.008.

[5] A. M. Olovnikov, “A theory of marginotomy: The incomplete copying of template margin in enzymic synthesis of polynucleotides and biological significance of the phenomenon,” *J. Theor. Biol.*, vol. 41, no. 1, pp. 181–190, Sep. 1973, doi: 10.1016/0022-5193(73)90198-7.

[6] J. Soudet, P. Jolivet, and M. T. Teixeira, “Elucidation of the DNA end-replication problem in Saccharomyces cerevisiae,” *Mol. Cell*, vol. 53, no. 6, pp. 954–964, Mar. 2014, doi: 10.1016/j.molcel.2014.02.030.

[7] N. W. Kim *et al.*, “Specific Association of Human Telomerase Activity with Immortal Cells and Cancer,” *Science*, vol. 266, no. 5193, pp. 2011–2015, Dec. 1994, doi: 10.1126/science.7605428.

[8] M. A. Dunham, A. A. Neumann, C. L. Fasching, and R. R. Reddel, “Telomere maintenance by recombination in human cells,” *Nat. Genet.*, vol. 26, no. 4, Art. no. 4, Dec. 2000, doi: 10.1038/82586.

[9] A. Muntoni, A. A. Neumann, M. Hills, and R. R. Reddel, “Telomere elongation involves intra-molecular DNA replication in cells utilizing alternative lengthening of

telomeres," *Hum. Mol. Genet.*, vol. 18, no. 6, pp. 1017–1027, Mar. 2009, doi: 10.1093/hmg/ddn436.

[10] N. Harbeck *et al.*, "Breast cancer," *Nat. Rev. Dis. Primer*, vol. 5, no. 1, Art. no. 1, Sep. 2019, doi: 10.1038/s41572-019-0111-2.

[11] H. Sung *et al.*, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," *CA. Cancer J. Clin.*, vol. 71, no. 3, pp. 209–249, 2021, doi: 10.3322/caac.21660.

[12] C. M. Perou *et al.*, "Molecular portraits of human breast tumours," *Nature*, vol. 406, no. 6797, Art. no. 6797, Aug. 2000, doi: 10.1038/35021093.

[13] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu, "Diagnosis of multiple cancer types by shrunken centroids of gene expression," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 99, no. 10, pp. 6567–6572, May 2002, doi: 10.1073/pnas.082099299.

[14] J. S. Parker *et al.*, "Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes," *J. Clin. Oncol.*, vol. 27, no. 8, pp. 1160–1167, Mar. 2009, doi: 10.1200/JCO.2008.18.1370.

[15] D. C. Zaha, "Significance of immunohistochemistry in breast cancer," *World J. Clin. Oncol.*, vol. 5, no. 3, pp. 382–392, Aug. 2014, doi: 10.5306/wjco.v5.i3.382.

[16] *Summary of findings – cost-effectiveness of next generation sequencing*. Canadian Agency for Drugs and Technologies in Health, 2014. Accessed: May 06, 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK274079/

[17] P. J. A. Cock, C. J. Fields, N. Goto, M. L. Heuer, and P. M. Rice, "The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants," *Nucleic Acids Res.*, vol. 38, no. 6, pp. 1767–1771, Apr. 2010, doi: 10.1093/nar/gkp1137.

[18] C. E. DeSantis *et al.*, “Breast cancer statistics, 2019,” *CA. Cancer J. Clin.*, vol. 69, no. 6, pp. 438–451, 2019, doi: 10.3322/caac.21583.

[19] K. Cibulskis *et al.*, “Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples,” *Nat. Biotechnol.*, vol. 31, no. 3, Art. no. 3, Mar. 2013, doi: 10.1038/nbt.2514.

[20] D. Benjamin, T. Sato, K. Cibulskis, G. Getz, C. Stewart, and L. Lichtenstein, “Calling Somatic SNVs and Indels with Mutect2.” bioRxiv, p. 861054, Dec. 02, 2019. doi: 10.1101/861054.

[21] S. T. Sherry *et al.*, “dbSNP: the NCBI database of genetic variation,” *Nucleic Acids Res.*, vol. 29, no. 1, pp. 308–311, Jan. 2001, doi: 10.1093/nar/29.1.308.

[22] C. Kandoth *et al.*, “mskcc/vcf2maf: vcf2maf v1.6.16.” Zenodo, Feb. 27, 2018. doi: 10.5281/zenodo.1185418.

[23] W. McLaren *et al.*, “The Ensembl Variant Effect Predictor,” *Genome Biol.*, vol. 17, no. 1, p. 122, Jun. 2016, doi: 10.1186/s13059-016-0974-4.

[24] “VCF - Variant Call Format,” *GATK*. https://gatk.broadinstitute.org/hc/en-us/articles/360035531692-VCF-Variant-Call-Format (accessed Jul. 22, 2022).

[25] “File Format: MAF - GDC Docs.” https://docs.gdc.cancer.gov/Data/File_Formats/MAF_Format/ (accessed Jul. 22, 2022).

[26] K. Ellrott *et al.*, “Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines,” *Cell Syst.*, vol. 6, no. 3, pp. 271-281.e7, Mar. 2018, doi: 10.1016/j.cels.2018.03.002.

[27] D. C. Koboldt *et al.*, “VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing,” *Genome Res.*, vol. 22, no. 3, pp. 568–576, Mar. 2012, doi: 10.1101/gr.129684.111.

[28] A. Ratan, T. L. Olson, T. P. Loughran, and W. Miller, “Identification of indels in next-generation sequencing data,” *BMC Bioinformatics*, vol. 16, no. 1, p. 42, Feb. 2015, doi: 10.1186/s12859-015-0483-6.

[29] K. Ye, M. H. Schulz, Q. Long, R. Apweiler, and Z. Ning, “Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads,” *Bioinforma. Oxf. Engl.*, vol. 25, no. 21, pp. 2865–2871, Nov. 2009, doi: 10.1093/bioinformatics/btp394.

[30] D. E. Larson *et al.*, “SomaticSniper: identification of somatic point mutations in whole genome sequencing data,” *Bioinformatics*, vol. 28, no. 3, pp. 311–317, Feb. 2012, doi: 10.1093/bioinformatics/btr665.

[31] A. J. Radenbaugh *et al.*, “RADIA: RNA and DNA Integrated Analysis for Somatic Mutation Detection,” *PLOS ONE*, vol. 9, no. 11, p. e111516, Nov. 2014, doi: 10.1371/journal.pone.0111516.

[32] Y. Fan *et al.*, “MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data,” *Genome Biol.*, vol. 17, no. 1, p. 178, Aug. 2016, doi: 10.1186/s13059-016-1029-6.

[33] M. Lek *et al.*, “Analysis of protein-coding genetic variation in 60,706 humans,” *Nature*, vol. 536, no. 7616, Art. no. 7616, Aug. 2016, doi: 10.1038/nature19057.

[34] K. Cibulskis, A. McKenna, T. Fennell, E. Banks, M. DePristo, and G. Getz, “ContEst: estimating cross-contamination of human samples in next-generation sequencing data,” *Bioinformatics*, vol. 27, no. 18, pp. 2601–2602, Sep. 2011, doi: 10.1093/bioinformatics/btr446.

[35] M. Costello *et al.*, “Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation,” *Nucleic Acids Res.*, vol. 41, no. 6, p. e67, Apr. 2013, doi: 10.1093/nar/gks1443.

[36] B. Bressac *et al.*, “p53 mutation in hepatocellular carcinoma after aflatoxin exposure,” *The Lancet*, vol. 338, no. 8779, pp. 1356–1359, Nov. 1991, doi: 10.1016/0140-6736(91)92236-U.

[37] T. M. Reid and L. A. Loeb, “Mutagenic specificity of oxygen radicals produced by human leukemia cells,” *Cancer Res.*, vol. 52, no. 5, pp. 1082–1086, Mar. 1992.

[38] E. D. Pleasance *et al.*, “A small-cell lung cancer genome with complex signatures of tobacco exposure,” *Nature*, vol. 463, no. 7278, Art. no. 7278, Jan. 2010, doi: 10.1038/nature08629.

[39] L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, P. J. Campbell, and M. R. Stratton, “Deciphering Signatures of Mutational Processes Operative in Human Cancer,” *Cell Rep.*, vol. 3, no. 1, pp. 246–259, Jan. 2013, doi: 10.1016/j.celrep.2012.12.008.

[40] J. G. Tate *et al.*, “COSMIC: the Catalogue Of Somatic Mutations In Cancer,” *Nucleic Acids Res.*, vol. 47, no. D1, pp. D941–D947, Jan. 2019, doi: 10.1093/nar/gky1015.

[41] L. B. Alexandrov *et al.*, “The repertoire of mutational signatures in human cancer,” *Nature*, vol. 578, no. 7793, Art. no. 7793, Feb. 2020, doi: 10.1038/s41586-020-1943-3.

[42] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” *Nature*, vol. 401, no. 6755, Art. no. 6755, Oct. 1999, doi: 10.1038/44565.

[43] J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, “Metagenes and molecular pattern discovery using matrix factorization,” *Proc. Natl. Acad. Sci.*, vol. 101, no. 12, pp. 4164–4169, Mar. 2004, doi: 10.1073/pnas.0308531101.

[44] M. R. Stratton, P. J. Campbell, and P. A. Futreal, “The cancer genome,” *Nature*, vol. 458, no. 7239, Art. no. 7239, Apr. 2009, doi: 10.1038/nature07943.

[45] S. Richards *et al.*, “Standards and Guidelines for the Interpretation of Sequence Variants: A Joint Consensus Recommendation of the American College of Medical Genetics

and Genomics and the Association for Molecular Pathology," *Genet. Med. Off. J. Am. Coll. Med. Genet.*, vol. 17, no. 5, pp. 405–424, May 2015, doi: 10.1038/gim.2015.30.

[46] D. Stehelin, H. E. Varmus, J. M. Bishop, and P. K. Vogt, "DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA," *Nature*, vol. 260, no. 5547, Art. no. 5547, Mar. 1976, doi: 10.1038/260170a0.

[47] A. G. Knudson, "Mutation and Cancer: Statistical Study of Retinoblastoma," *Proc. Natl. Acad. Sci.*, vol. 68, no. 4, pp. 820–823, Apr. 1971, doi: 10.1073/pnas.68.4.820.

[48] F. Martínez-Jiménez *et al.*, "A compendium of mutational cancer driver genes," *Nat. Rev. Cancer*, vol. 20, no. 10, Art. no. 10, Oct. 2020, doi: 10.1038/s41568-020-0290-x.

[49] M. S. Lawrence *et al.*, "Discovery and saturation analysis of cancer genes across 21 tumour types," *Nature*, vol. 505, no. 7484, Art. no. 7484, Jan. 2014, doi: 10.1038/nature12912.

[50] D. Tamborero, A. Gonzalez-Perez, and N. Lopez-Bigas, "OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes," *Bioinformatics*, vol. 29, no. 18, pp. 2238–2244, Sep. 2013, doi: 10.1093/bioinformatics/btt395.

[51] C. Tokheim *et al.*, "Exome-Scale Discovery of Hotspot Mutation Regions in Human Cancer Using 3D Protein Structure," *Cancer Res.*, vol. 76, no. 13, pp. 3719–3731, Jun. 2016, doi: 10.1158/0008-5472.CAN-15-3190.

[52] N. D. Dees *et al.*, "MuSiC: Identifying mutational significance in cancer genomes," *Genome Res.*, vol. 22, no. 8, pp. 1589–1598, Aug. 2012, doi: 10.1101/gr.134635.111.

[53] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz, and K. W. Kinzler, "Cancer Genome Landscapes," *Science*, vol. 339, no. 6127, pp. 1546–1558, Mar. 2013, doi: 10.1126/science.1235122.

[54] R. Nussinov and C.-J. Tsai, "'Latent drivers' expand the cancer mutational landscape," *Curr. Opin. Struct. Biol.*, vol. 32, pp. 25–32, Jun. 2015, doi: 10.1016/j.sbi.2015.01.004.

[55] B. R. Yavuz, C.-J. Tsai, R. Nussinov, and N. Tuncbag, “Discovery of Latent Drivers from Double Mutations in Pan-Cancer Data Reveal their Clinical Impact.” bioRxiv, p. 2021.04.02.438239, Apr. 04, 2021. doi: 10.1101/2021.04.02.438239.

[56] M. S. Lawrence *et al.*, “Mutational heterogeneity in cancer and the search for new cancer-associated genes,” *Nature*, vol. 499, no. 7457, Art. no. 7457, Jul. 2013, doi: 10.1038/nature12213.

[57] M. A. Chapman *et al.*, “Initial genome sequencing and analysis of multiple myeloma,” *Nature*, vol. 471, no. 7339, Art. no. 7339, Mar. 2011, doi: 10.1038/nature09837.

[58] S. Banerji *et al.*, “Sequence analysis of mutations and translocations across breast cancer subtypes,” *Nature*, vol. 486, no. 7403, Art. no. 7403, Jun. 2012, doi: 10.1038/nature11154.

[59] J. G. Lohr *et al.*, “Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing,” *Proc. Natl. Acad. Sci.*, vol. 109, no. 10, pp. 3879–3884, Mar. 2012, doi: 10.1073/pnas.1121343109.

[60] N. Vasan *et al.*, “Double PIK3CA mutations in cis increase oncogenicity and sensitivity to PI3Kα inhibitors,” *Science*, vol. 366, no. 6466, pp. 714–723, Nov. 2019, doi: 10.1126/science.aaw9032.

[61] C. J. Tokheim, N. Papadopoulos, K. W. Kinzler, B. Vogelstein, and R. Karchin, “Evaluating the evaluation of cancer driver genes,” *Proc. Natl. Acad. Sci.*, vol. 113, no. 50, pp. 14330–14335, Dec. 2016, doi: 10.1073/pnas.1616440113.

[62] Y. Deng *et al.*, “Identifying mutual exclusivity across cancer genomes: computational approaches to discover genetic interaction and reveal tumor vulnerability,” *Brief. Bioinform.*, vol. 20, no. 1, pp. 254–266, Jan. 2019, doi: 10.1093/bib/bbx109.

[63] Y.-A. Kim, S. Madan, and T. M. Przytycka, “WeSME: uncovering mutual exclusivity of cancer drivers and beyond,” *Bioinformatics*, vol. 33, no. 6, pp. 814–821, Mar. 2017, doi: 10.1093/bioinformatics/btw242.

[64] F. Vandin, E. Upfal, and B. J. Raphael, “De novo discovery of mutated driver pathways in cancer,” *Genome Res.*, vol. 22, no. 2, pp. 375–385, Feb. 2012, doi: 10.1101/gr.120477.111.

[65] M. D. Leiserson, H.-T. Wu, F. Vandin, and B. J. Raphael, “CoMEt: a statistical approach to identify combinations of mutually exclusive alterations in cancer,” *Genome Biol.*, vol. 16, no. 1, p. 160, Aug. 2015, doi: 10.1186/s13059-015-0700-7.

[66] Ö. Babur *et al.*, “Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations,” *Genome Biol.*, vol. 16, no. 1, p. 45, Feb. 2015, doi: 10.1186/s13059-015-0612-6.

[67] Q. Cui, “A Network of Cancer Genes with Co-Occurring and Anti-Co-Occurring Mutations,” *PLOS ONE*, vol. 5, no. 10, p. e13180, Oct. 2010, doi: 10.1371/journal.pone.0013180.

[68] F. Sanchez-Vega *et al.*, “Oncogenic Signaling Pathways in The Cancer Genome Atlas,” *Cell*, vol. 173, no. 2, pp. 321-337.e10, Apr. 2018, doi: 10.1016/j.cell.2018.03.035.

[69] M. Mina *et al.*, “Conditional Selection of Genomic Alterations Dictates Cancer Evolution and Oncogenic Dependencies,” *Cancer Cell*, vol. 32, no. 2, pp. 155-168.e6, Aug. 2017, doi: 10.1016/j.ccell.2017.06.010.

[70] G. Ciriello, E. Cerami, C. Sander, and N. Schultz, “Mutual exclusivity analysis identifies oncogenic network modules,” *Genome Res.*, vol. 22, no. 2, pp. 398–406, Feb. 2012, doi: 10.1101/gr.125567.111.

[71] T. I. Zack *et al.*, “Pan-cancer patterns of somatic copy number alteration,” *Nat. Genet.*, vol. 45, no. 10, Art. no. 10, Oct. 2013, doi: 10.1038/ng.2760.

[72] H. Yamaguchi and W. A. Hendrickson, “Structural basis for activation of human lymphocyte kinase Lck upon tyrosine phosphorylation,” *Nature*, vol. 384, no. 6608, Art. no. 6608, Dec. 1996, doi: 10.1038/384484a0.

[73] R. Bhargava *et al.*, “EGFR gene amplification in breast cancer: correlation with epidermal growth factor receptor mRNA and protein expression and HER-2 status and absence of EGFR-activating mutations,” *Mod. Pathol.*, vol. 18, no. 8, Art. no. 8, Aug. 2005, doi: 10.1038/modpathol.3800438.

[74] N. T. Pfister *et al.*, “Mutant p53 cooperates with the SWI/SNF chromatin remodeling complex to regulate VEGFR2 in breast cancer cells,” *Genes Dev.*, vol. 29, no. 12, pp. 1298–1315, Jun. 2015, doi: 10.1101/gad.263202.115.

[75] I. Bahceci, U. Dogrusoz, K. C. La, Ö. Babur, J. Gao, and N. Schultz, “PathwayMapper: a collaborative visual web editor for cancer pathways and genomic data,” *Bioinformatics*, vol. 33, no. 14, pp. 2238–2240, Jul. 2017, doi: 10.1093/bioinformatics/btx149.

[76] C. L. Carpenter, B. C. Duckworth, K. R. Auger, B. Cohen, B. S. Schaffhausen, and L. C. Cantley, “Purification and characterization of phosphoinositide 3-kinase from rat liver.,” *J. Biol. Chem.*, vol. 265, no. 32, pp. 19704–19711, Nov. 1990, doi: 10.1016/S0021-9258(17)45429-9.

[77] J. Yu, Y. Zhang, J. McIlroy, T. Rordorf-Nikolic, G. A. Orr, and J. M. Backer, “Regulation of the p85/p110 Phosphatidylinositol 3′-Kinase: Stabilization and Inhibition of the p110α Catalytic Subunit by the p85 Regulatory Subunit,” *Mol. Cell. Biol.*, vol. 18, no. 3, pp. 1379–1387, Mar. 1998, doi: 10.1128/MCB.18.3.1379.

[78] T. F. Franke *et al.*, “The protein kinase encoded by the Akt proto-oncogene is a target of the PDGF-activated phosphatidylinositol 3-kinase,” *Cell*, vol. 81, no. 5, pp. 727–736, Jun. 1995, doi: 10.1016/0092-8674(95)90534-0.

[79] D. D. Sarbassov, D. A. Guertin, S. M. Ali, and D. M. Sabatini, “Phosphorylation and Regulation of Akt/PKB by the Rictor-mTOR Complex,” *Science*, vol. 307, no. 5712, pp. 1098–1101, Feb. 2005, doi: 10.1126/science.1106148.

[80] D. C. Koboldt *et al.*, “Comprehensive molecular portraits of human breast tumours,” *Nature*, vol. 490, no. 7418, Art. no. 7418, Oct. 2012, doi: 10.1038/nature11412.

[81] J. D. Carpten *et al.*, “A transforming mutation in the pleckstrin homology domain of AKT1 in cancer,” *Nature*, vol. 448, no. 7152, Art. no. 7152, Jul. 2007, doi: 10.1038/nature05933.

[82] B. Amati, M. W. Brooks, N. Levy, T. D. Littlewood, G. I. Evan, and H. Land, “Oncogenic activity of the c-Myc protein requires dimerization with Max,” *Cell*, vol. 72, no. 2, pp. 233–245, Jan. 1993, doi: 10.1016/0092-8674(93)90663-b.

[83] “Mga, a dual-specificity transcription factor that interacts with Max and contains a T-domain DNA-binding motif,” *EMBO J.*, vol. 18, no. 24, pp. 7019–7028, Dec. 1999, doi: 10.1093/emboj/18.24.7019.

[84] D. E. Ayer, L. Kretzner, and R. N. Eisenman, “Mad: a heterodimeric partner for Max that antagonizes Myc transcriptional activity,” *Cell*, vol. 72, no. 2, pp. 211–222, Jan. 1993, doi: 10.1016/0092-8674(93)90661-9.

[85] A. V. Grinberg, C.-D. Hu, and T. K. Kerppola, “Visualization of Myc/Max/Mad Family Dimers and the Competition for Dimerization in Living Cells,” *Mol. Cell. Biol.*, vol. 24, no. 10, pp. 4294–4308, May 2004, doi: 10.1128/MCB.24.10.4294-4308.2004.

[86] C. W. Peterson, C. A. Stoltzman, M. P. Sighinolfi, K.-S. Han, and D. E. Ayer, “Glucose Controls Nuclear Accumulation, Promoter Binding, and Transcriptional Activity of the MondoA-Mlx Heterodimer,” *Mol. Cell. Biol.*, vol. 30, no. 12, pp. 2887–2895, Jun. 2010, doi: 10.1128/MCB.01613-09.

[87] C. L. Sans, D. J. Satterwhite, C. A. Stoltzman, K. T. Breen, and D. E. Ayer, “MondoA-Mlx Heterodimers Are Candidate Sensors of Cellular Energy Status: Mitochondrial Localization and Direct Regulation of Glycolysis,” *Mol. Cell. Biol.*, vol. 26, no. 13, pp. 4863–4871, Jul. 2006, doi: 10.1128/MCB.00657-05.

[88] R. Beroukhim *et al.*, “The landscape of somatic copy-number alteration across human cancers,” *Nature*, vol. 463, no. 7283, Art. no. 7283, Feb. 2010, doi: 10.1038/nature08822.

[89] P. Llabata *et al.*, “MAX mutant small-cell lung cancers exhibit impaired activities of MGA-dependent noncanonical polycomb repressive complex,” *Proc. Natl. Acad. Sci.*, vol. 118, no. 37, p. e2024824118, Sep. 2021, doi: 10.1073/pnas.2024824118.

[90] R. Honda, H. Tanaka, and H. Yasuda, “Oncoprotein MDM2 is a ubiquitin ligase E3 for tumor suppressor p53,” *FEBS Lett.*, vol. 420, no. 1, pp. 25–27, 1997, doi: 10.1016/S0014-5793(97)01480-4.

[91] M. Kodama, C. Otsubo, T. Hirota, J. Yokota, M. Enari, and Y. Taya, “Requirement of ATM for rapid p53 phosphorylation at Ser46 without Ser/Thr-Gln sequences,” *Mol. Cell. Biol.*, vol. 30, no. 7, pp. 1620–1633, Apr. 2010, doi: 10.1128/MCB.00810-09.

[92] S.-Y. Shieh, M. Ikeda, Y. Taya, and C. Prives, “DNA Damage-Induced Phosphorylation of p53 Alleviates Inhibition by MDM2,” *Cell*, vol. 91, no. 3, pp. 325–334, Oct. 1997, doi: 10.1016/S0092-8674(00)80416-X.

[93] H. C. Lim *et al.*, “Ribosomal S6 Kinase 2 (RSK2) Maintains Genomic Stability by Activating the Atm/p53-Dependent DNA Damage Pathway,” *PLOS ONE*, vol. 8, no. 9, p. e74334, Sep. 2013, doi: 10.1371/journal.pone.0074334.

[94] K. C. de Andrade *et al.*, “The TP53 Database: transition from the International Agency for Research on Cancer to the US National Cancer Institute,” *Cell Death Differ.*, vol. 29, no. 5, pp. 1071–1073, May 2022, doi: 10.1038/s41418-022-00976-3.

[95] X. Huang *et al.*, “A Novel PTEN/Mutant p53/c-Myc/Bcl-XL Axis Mediates Context-Dependent Oncogenic Effects of PTEN with Implications for Cancer Prognosis and Therapy,” *Neoplasia*, vol. 15, no. 8, pp. 952–965, Aug. 2013, doi: 10.1593/neo.13376.

[96] M. Adorno *et al.*, “A Mutant-p53/Smad Complex Opposes p63 to Empower TGFβ-Induced Metastasis,” *Cell*, vol. 137, no. 1, pp. 87–98, Apr. 2009, doi: 10.1016/j.cell.2009.01.039.

[97] A. Renwick *et al.*, “ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles,” *Nat. Genet.*, vol. 38, no. 8, Art. no. 8, Aug. 2006, doi: 10.1038/ng1837.

[98] E. Waskiewicz, M. Vasiliou, I. Corcoles-Saez, and R. S. Cha, “Cancer genome datamining and functional genetic analysis implicate mechanisms of ATM/ATR dysfunction underpinning carcinogenesis,” *Commun. Biol.*, vol. 4, no. 1, Art. no. 1, Mar. 2021, doi: 10.1038/s42003-021-01884-x.

[99] J. Gao *et al.*, “Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal,” *Sci. Signal.*, vol. 6, no. 269, pp. pl1–pl1, Apr. 2013, doi: 10.1126/scisignal.2004088.

[100] M. R. Dufault *et al.*, “Limited relevance of the CHEK2 gene in hereditary breast cancer,” *Int. J. Cancer*, vol. 110, no. 3, pp. 320–325, Jun. 2004, doi: 10.1002/ijc.20073.

[101] H. Matsushime, M. F. Roussel, R. A. Ashmun, and C. J. Sherr, “Colony-stimulating factor 1 regulates novel cyclins during the G1 phase of the cell cycle,” *Cell*, vol. 65, no. 4, pp. 701–713, May 1991, doi: 10.1016/0092-8674(91)90101-4.

[102] H. Matsushime *et al.*, “Identification and properties of an atypical catalytic subunit (p34PSK-J3/cdk4) for mammalian D type G1 cyclins,” *Cell*, vol. 71, no. 2, pp. 323–334, Oct. 1992, doi: 10.1016/0092-8674(92)90360-O.

[103] M. Meyerson and E. Harlow, “Identification of G1 kinase activity for cdk6, a novel cyclin D partner,” *Mol. Cell. Biol.*, vol. 14, no. 3, pp. 2077–2086, Mar. 1994, doi: 10.1128/mcb.14.3.2077-2086.1994.

[104] J. Kato, H. Matsushime, S. W. Hiebert, M. E. Ewen, and C. J. Sherr, “Direct binding of cyclin D to the retinoblastoma gene product (pRb) and pRb phosphorylation by the cyclin D-dependent kinase CDK4,” *Genes Dev.*, vol. 7, no. 3, pp. 331–342, Mar. 1993, doi: 10.1101/gad.7.3.331.

[105] M. E. Ewen, H. K. Sluss, C. J. Sherr, H. Matsushime, J. Kato, and D. M. Livingston, “Functional interactions of the retinoblastoma protein with mammalian D-type cyclins,” *Cell*, vol. 73, no. 3, pp. 487–497, May 1993, doi: 10.1016/0092-8674(93)90136-E.

[106] P. W. Hinds, S. Mittnacht, V. Dulic, A. Arnold, S. I. Reed, and R. A. Weinberg, “Regulation of retinoblastoma protein functions by ectopic expression of human cyclins,” *Cell*, vol. 70, no. 6, pp. 993–1006, Sep. 1992, doi: 10.1016/0092-8674(92)90249-c.

[107] L. Bockstaele *et al.*, “Regulated Activating Thr172 Phosphorylation of Cyclin-Dependent Kinase 4(CDK4): Its Relationship with Cyclins and CDK ‘Inhibitors,’” *Mol. Cell. Biol.*, vol. 26, no. 13, pp. 5070–5085, Jul. 2006, doi: 10.1128/MCB.02006-05.

[108] C. J. Sherr and J. M. Roberts, “Living with or without cyclins and cyclin-dependent kinases,” *Genes Dev.*, vol. 18, no. 22, pp. 2699–2711, Nov. 2004, doi: 10.1101/gad.1256504.

[109] C.-H. Lu, C.-H. Wu, M.-H. Tsai, L.-C. Lai, and E. Y. Chuang, “MutScape: an analytical toolkit for probing the mutational landscape in cancer genomics,” *NAR Genomics Bioinforma.*, vol. 3, no. 4, p. lqab099, Nov. 2021, doi: 10.1093/nargab/lqab099.

[110] B. O’Sullivan and C. Seoighe, “vcfView: An Extensible Data Visualization and Quality Assurance Platform for Integrated Somatic Variant Analysis,” *Cancer Inform.*, vol. 19, p. 1176935120972377, Jan. 2020, doi: 10.1177/1176935120972377.

[111] J. Crowdis, M. X. He, B. Reardon, and E. M. Van Allen, “CoMut: visualizing integrated molecular information with comutation plots,” *Bioinformatics*, vol. 36, no. 15, pp. 4348–4349, Jun. 2020, doi: 10.1093/bioinformatics/btaa554.

[112] M. Díaz-Gay, M. Vila-Casadesús, S. Franch-Expósito, E. Hernández-Illán, J. J. Lozano, and S. Castellví-Bel, “Mutational Signatures in Cancer (MuSiCa): a web application to implement mutational signatures analysis in cancer samples,” *BMC Bioinformatics*, vol. 19, no. 1, p. 224, Jun. 2018, doi: 10.1186/s12859-018-2234-y.

[113] F. Blokzijl, R. Janssen, R. van Boxtel, and E. Cuppen, “MutationalPatterns: comprehensive genome-wide analysis of mutational processes,” *Genome Med.*, vol. 10, no. 1, p. 33, Apr. 2018, doi: 10.1186/s13073-018-0539-0.

[114] Y. Tang, M. Horikoshi, and W. Li, “ggfortify: Unified Interface to Visualize Statistical Results of Popular R Packages,” *R J.*, vol. 8, no. 2, pp. 474–485, 2016.

[115] R. Gaujoux and C. Seoighe, “A flexible R package for nonnegative matrix factorization,” *BMC Bioinformatics*, vol. 11, no. 1, p. 367, Jul. 2010, doi: 10.1186/1471-2105-11-367.

[116] C. Ahlmann-Eltze and I. Patil, “ggsignif: R Package for Displaying Significance Brackets for ‘ggplot2.’” PsyArXiv, Mar. 31, 2021. doi: 10.31234/osf.io/7awm6.

[117] S. Xu, M. Chen, T. Feng, L. Zhan, L. Zhou, and G. Yu, “Use ggbreak to Effectively Utilize Plotting Space to Deal With Large Datasets and Outliers,” *Front. Genet.*, vol. 12, 2021, Accessed: Jul. 14, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fgene.2021.774846

[118] F. Cunningham *et al.*, “Ensembl 2022,” *Nucleic Acids Res.*, vol. 50, no. D1, pp. D988–D995, Jan. 2022, doi: 10.1093/nar/gkab1049.

[119] T. Wu *et al.*, “clusterProfiler 4.0: A universal enrichment tool for interpreting omics data,” *The Innovation*, vol. 2, no. 3, p. 100141, Aug. 2021, doi: 10.1016/j.xinn.2021.100141.

[120] B. Haibe-Kains *et al.*, “A Three-Gene Model to Robustly Identify Breast Cancer Molecular Subtypes,” *JNCI J. Natl. Cancer Inst.*, vol. 104, no. 4, pp. 311–325, Feb. 2012, doi: 10.1093/jnci/djr545.

[121] H. Wickham, *ggplot2*. New York, NY: Springer New York, 2009. doi: 10.1007/978-0-387-98141-3.

[122] T. Galili, A. O’Callaghan, J. Sidi, and C. Sievert, “heatmaply: an R package for creating interactive cluster heatmaps for online publishing,” *Bioinformatics*, vol. 34, no. 9, pp. 1600–1602, May 2018, doi: 10.1093/bioinformatics/btx657.

[123] A. Mayakonda, D.-C. Lin, Y. Assenov, C. Plass, and H. P. Koeffler, “Maftools: efficient and comprehensive analysis of somatic variants in cancer,” *Genome Res.*, vol. 28, no. 11, pp. 1747–1756, Nov. 2018, doi: 10.1101/gr.239244.118.

[124] V. M. R. Muggeo, “Interval estimation for the breakpoint in segmented regression: a smoothed score-based approach,” *Aust. N. Z. J. Stat.*, vol. 59, no. 3, pp. 311–322, 2017, doi: 10.1111/anzs.12200.

[125] T. A. Knijnenburg *et al.*, “Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas,” *Cell Rep.*, vol. 23, no. 1, pp. 239-254.e6, Apr. 2018, doi: 10.1016/j.celrep.2018.03.076.

[126] M. Seiler *et al.*, “Somatic Mutational Landscape of Splicing Factor Genes and Their Functional Consequences across 33 Cancer Types,” *Cell Rep.*, vol. 23, no. 1, pp. 282-296.e4, Apr. 2018, doi: 10.1016/j.celrep.2018.01.088.

[127] C. Lefebvre *et al.*, “Mutational Profile of Metastatic Breast Cancers: A Retrospective Analysis,” *PLOS Med.*, vol. 13, no. 12, p. e1002201, Dec. 2016, doi: 10.1371/journal.pmed.1002201.

[128] M. J. Nixon *et al.*, “PIK3CA and MAP3K1 alterations imply luminal A status and are associated with clinical benefit from pan-PI3K inhibitor buparlisib and letrozole in ER+

metastatic breast cancer," *NPJ Breast Cancer*, vol. 5, p. 31, 2019, doi: 10.1038/s41523-019-0126-6.

[129] M. Fischer, P. Grossmann, M. Padi, and J. A. DeCaprio, "Integration of TP53, DREAM, MMB-FOXM1 and RB-E2F target gene analyses identifies cell cycle gene regulatory networks," *Nucleic Acids Res.*, vol. 44, no. 13, pp. 6070–6086, Jul. 2016, doi: 10.1093/nar/gkw523.

[130] S. Uxa *et al.*, "DREAM and RB cooperate to induce gene repression and cell-cycle arrest in response to p53 activation," *Nucleic Acids Res.*, vol. 47, no. 17, pp. 9087–9103, Sep. 2019, doi: 10.1093/nar/gkz635.

[131] J. M. Mudge *et al.*, "Standardized annotation of translated open reading frames," *Nat. Biotechnol.*, Jul. 2022, doi: 10.1038/s41587-022-01369-0.

[132] M. Touat *et al.*, "Mechanisms and therapeutic implications of hypermutation in gliomas," *Nature*, vol. 580, no. 7804, Art. no. 7804, Apr. 2020, doi: 10.1038/s41586-020-2209-9.

[133] B. B. Campbell *et al.*, "Comprehensive Analysis of Hypermutation in Human Cancer," *Cell*, vol. 171, no. 5, pp. 1042-1056.e10, Nov. 2017, doi: 10.1016/j.cell.2017.09.048.

[134] H. Yuan *et al.*, "Characteristics of Pan-Cancer Patients With Ultrahigh Tumor Mutation Burden," *Front. Oncol.*, vol. 11, 2021, Accessed: Aug. 05, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fonc.2021.682017

[135] K. Hatakeyama *et al.*, "Molecular profiling and sequential somatic mutation shift in hypermutator tumours harbouring POLE mutations," *Sci. Rep.*, vol. 8, no. 1, Art. no. 1, Jun. 2018, doi: 10.1038/s41598-018-26967-4.

[136] H. Zhang *et al.*, "Predicting Tumor Mutational Burden from Liver Cancer Pathological Images Using Convolutional Neural Network," in *2019 IEEE International Conference on*

Bioinformatics and Biomedicine (BIBM), Nov. 2019, pp. 920–925. doi: 10.1109/BIBM47256.2019.8983139.

[137] J. Ying *et al.*, "Additive effects of variants of unknown significance in replication repair-associated DNA polymerase genes on mutational burden and prognosis across diverse cancers," *J. Immunother. Cancer*, vol. 9, no. 9, p. e002336, Sep. 2021, doi: 10.1136/jitc-2021-002336.

[138] X.-M. Wang *et al.*, "Integrative genomic study of Chinese clear cell renal cell carcinoma reveals features associated with thrombus," *Nat. Commun.*, vol. 11, no. 1, Art. no. 1, Feb. 2020, doi: 10.1038/s41467-020-14601-9.

[139] H. F. Lodish, *Molecular cell biology*, Eighth edition. New York: W.H. Freeman-Macmillan Learning, 2016.

[140] A. Taly, P.-J. Corringer, D. Guedin, P. Lestage, and J.-P. Changeux, "Nicotinic receptors: allosteric transitions and therapeutic targets in the nervous system," *Nat. Rev. Drug Discov.*, vol. 8, no. 9, Art. no. 9, Sep. 2009, doi: 10.1038/nrd2927.

[141] W. H. Berrettini and G. A. Doyle, "The CHRNA5–A3–B4 gene cluster in nicotine addiction," *Mol. Psychiatry*, vol. 17, no. 9, Art. no. 9, Sep. 2012, doi: 10.1038/mp.2011.122.

[142] R. J. Hung *et al.*, "A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25," *Nature*, vol. 452, no. 7187, Art. no. 7187, Apr. 2008, doi: 10.1038/nature06885.

[143] S. Cingir Koker, E. Jahja, H. Shehwana, A. G. Keskus, and O. Konu, "Cholinergic Receptor Nicotinic Alpha 5 (CHRNA5) RNAi is associated with cell cycle inhibition, apoptosis, DNA damage response and drug sensitivity in breast cancer," *PloS One*, vol. 13, no. 12, p. e0208982, 2018, doi: 10.1371/journal.pone.0208982.

[144] A. G. Keşküş, "Identification of modulatory functions of TP53, estrogen signaling, and 14q32.31 miRNA cluster on CHRNA5 knock-down expression profile and development of

syneRgy APP," Thesis, Bilkent University, 2021. Accessed: Jul. 23, 2022. [Online]. Available: http://repository.bilkent.edu.tr/handle/11693/76589

[145] D. Szklarczyk *et al.*, "The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D605–D612, Jan. 2021, doi: 10.1093/nar/gkaa1074.

[146] Gene Ontology Consortium, "The Gene Ontology resource: enriching a GOld mine," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D325–D334, Jan. 2021, doi: 10.1093/nar/gkaa1113.

[147] M. Gillespie *et al.*, "The reactome pathway knowledgebase 2022," *Nucleic Acids Res.*, vol. 50, no. D1, pp. D687–D692, Jan. 2022, doi: 10.1093/nar/gkab1028.

[148] G. Korotkevich, V. Sukhov, N. Budin, B. Shpak, M. N. Artyomov, and A. Sergushichev, "Fast gene set enrichment analysis." bioRxiv, p. 060012, Feb. 01, 2021. doi: 10.1101/060012.

[149] L. B. Alexandrov *et al.*, "Signatures of mutational processes in human cancer," *Nature*, vol. 500, no. 7463, Art. no. 7463, Aug. 2013, doi: 10.1038/nature12477.

[150] S. M. A. Islam *et al.*, "Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor." bioRxiv, p. 2020.12.13.422570, Apr. 10, 2022. doi: 10.1101/2020.12.13.422570.

[151] L. Angus *et al.*, "The genomic landscape of metastatic breast cancer highlights changes in mutation and signature frequencies," *Nat. Genet.*, vol. 51, no. 10, Art. no. 10, Oct. 2019, doi: 10.1038/s41588-019-0507-7.

[152] F. Bertucci *et al.*, "Genomic characterization of metastatic breast cancers," *Nature*, vol. 569, no. 7757, Art. no. 7757, May 2019, doi: 10.1038/s41586-019-1056-z.

[153] A. G. Lada *et al.*, "AID/APOBEC cytosine deaminase induces genome-wide kataegis," *Biol. Direct*, vol. 7, no. 1, p. 47, Dec. 2012, doi: 10.1186/1745-6150-7-47.

[154] C. J. Sakofsky *et al.*, “Break-Induced Replication Is a Source of Mutation Clusters Underlying Kataegis,” *Cell Rep.*, vol. 7, no. 5, pp. 1640–1648, Jun. 2014, doi: 10.1016/j.celrep.2014.04.053.

[155] B. A. Walker *et al.*, “APOBEC family mutational signatures are associated with poor prognosis translocations in multiple myeloma,” *Nat. Commun.*, vol. 6, no. 1, Art. no. 1, Apr. 2015, doi: 10.1038/ncomms7997.

[156] B. J. Taylor *et al.*, “DNA deaminases induce break-associated mutation showers with implication of APOBEC3B and 3A in breast cancer kataegis,” *eLife*, vol. 2, p. e00534, Apr. 2013, doi: 10.7554/eLife.00534.

[157] G. Zhang *et al.*, “Whole-Exome Sequencing Reveals Frequent Mutations in Chromatin Remodeling Genes in Mammary and Extramammary Paget’s Diseases,” *J. Invest. Dermatol.*, vol. 139, no. 4, pp. 789–795, Apr. 2019, doi: 10.1016/j.jid.2018.08.030.

[158] P. Yuan *et al.*, “Genomic analyses reveal SCN7A is associated with the prognosis of esophageal squamous cell carcinoma,” *Esophagus*, vol. 19, no. 2, pp. 303–315, Apr. 2022, doi: 10.1007/s10388-021-00898-y.

[159] H. Ying, A. Lin, J. Liang, J. Zhang, and P. Luo, “Association Between FSIP2 Mutation and an Improved Efficacy of Immune Checkpoint Inhibitors in Patients With Skin Cutaneous Melanoma,” *Front. Mol. Biosci.*, vol. 8, 2021, Accessed: Jul. 17, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fmolb.2021.629330

[160] S. Zhang *et al.*, “Dysregulated Immune and Metabolic Microenvironment Is Associated with the Post-Operative Relapse in Stage I Non-Small Cell Lung Cancer,” *Cancers*, vol. 14, no. 13, Art. no. 13, Jan. 2022, doi: 10.3390/cancers14133061.

[161] H. Tang *et al.*, “Genomic characterization of primary cervical cancer and lung metastasis.,” *J. Clin. Oncol.*, vol. 40, no. 16_suppl, pp. e17527–e17527, Jun. 2022, doi: 10.1200/JCO.2022.40.16_suppl.e17527.

[162] G. Bao *et al.*, “A Germline Mutation in ATR Is Associated With Lung Adenocarcinoma in Asian Patients,” *Front. Oncol.*, vol. 12, p. 855305, Jan. 2022, doi: 10.3389/fonc.2022.855305.

[163] F. Towfic *et al.*, “Large-Scale Analysis of Relapsed/Refractory Multiple Myeloma Genome Reveals Increased Prevalence of High-Risk Molecular Features and Oncogenic Drivers,” *Blood*, vol. 134, p. 365, Nov. 2019, doi: 10.1182/blood-2019-124939.

[164] R. Chettier, K. Ward, and H. M. Albertsen, “Endometriosis Is Associated with Rare Copy Number Variants,” *PLOS ONE*, vol. 9, no. 8, p. e103968, Aug. 2014, doi: 10.1371/journal.pone.0103968.

[165] Q. Huang and M. Baudis, “Candidate targets of copy number deletion events across 17 cancer types.” bioRxiv, p. 2022.06.29.498080, Jun. 30, 2022. doi: 10.1101/2022.06.29.498080.

[166] Q.-X. Zhang, Å. Borg, D. M. Wolf, S. Oesterreich, and S. A. W. Fuqua, “An Estrogen Receptor Mutant with Strong Hormone-independent Activity from a Metastatic Breast Cancer1,” *Cancer Res.*, vol. 57, no. 7, pp. 1244–1249, Apr. 1997.

[167] W. Toy *et al.*, “ESR1 ligand-binding domain mutations in hormone-resistant breast cancer,” *Nat. Genet.*, vol. 45, no. 12, Art. no. 12, Dec. 2013, doi: 10.1038/ng.2822.

[168] S. Li *et al.*, “Endocrine-Therapy-Resistant ESR1 Variants Revealed by Genomic Characterization of Breast-Cancer-Derived Xenografts,” *Cell Rep.*, vol. 4, no. 6, pp. 1116–1130, Sep. 2013, doi: 10.1016/j.celrep.2013.08.022.

[169] D. Banerjee, “Connexin’s Connection in Breast Cancer Growth and Progression,” *Int. J. Cell Biol.*, vol. 2016, p. e9025905, Aug. 2016, doi: 10.1155/2016/9025905.

[170] I. Plante, M. K. G. Stewart, K. Barr, A. L. Allan, and D. W. Laird, “Cx43 suppresses mammary tumor metastasis to the lung in a Cx43 mutant mouse model of human disease,” *Oncogene*, vol. 30, no. 14, Art. no. 14, Apr. 2011, doi: 10.1038/onc.2010.551.

[171] S. Cha, E. Lee, and H.-H. Won, “Comprehensive characterization of distinct genetic alterations in metastatic breast cancer across various metastatic sites,” *Npj Breast Cancer*, vol. 7, no. 1, Art. no. 1, Jul. 2021, doi: 10.1038/s41523-021-00303-y.

[172] P. Razavi *et al.*, “The Genomic Landscape of Endocrine-Resistant Advanced Breast Cancers,” *Cancer Cell*, vol. 34, no. 3, pp. 427-438.e6, Sep. 2018, doi: 10.1016/j.ccell.2018.08.008.

[173] J.-H. Oh *et al.*, “Spontaneous mutations in the single TTN gene represent high tumor mutation burden,” *Npj Genomic Med.*, vol. 5, no. 1, Art. no. 1, Jan. 2020, doi: 10.1038/s41525-019-0107-6.

[174] M. Hofree, J. P. Shen, H. Carter, A. Gross, and T. Ideker, “Network-based stratification of tumor mutations,” *Nat. Methods*, vol. 10, no. 11, Art. no. 11, Nov. 2013, doi: 10.1038/nmeth.2651.

[175] A. B. Sahakyan and S. Balasubramanian, “Long genes and genes with multiple splice variants are enriched in pathways linked to cancer and other multigenic diseases,” *BMC Genomics*, vol. 17, no. 1, p. 225, Mar. 2016, doi: 10.1186/s12864-016-2582-9.

[176] G. Rong *et al.*, “DNA damage response as a prognostic indicator in metastatic breast cancer via mutational analysis,” *Ann. Transl. Med.*, vol. 9, no. 3, p. 220, Feb. 2021, doi: 10.21037/atm-20-2137.

[177] S. Saha, B. Buttari, E. Panieri, E. Profumo, and L. Saso, “An Overview of Nrf2 Signaling Pathway and Its Role in Inflammation,” *Molecules*, vol. 25, no. 22, Art. no. 22, Jan. 2020, doi: 10.3390/molecules25225474.

[178] S. Christensen *et al.*, “5-Fluorouracil treatment induces characteristic T>G mutations in human cancer,” *Nat. Commun.*, vol. 10, no. 1, Art. no. 1, Oct. 2019, doi: 10.1038/s41467-019-12594-8.

[179] H. Zhang *et al.*, “Nrf2 promotes breast cancer cell migration via up-regulation of G6PD/HIF-1α/Notch1 axis,” *J. Cell. Mol. Med.*, vol. 23, no. 5, pp. 3451–3463, May 2019, doi: 10.1111/jcmm.14241.

[180] C. Wolowczyk *et al.*, “NRF2 drives an oxidative stress response predictive of breast cancer,” *Free Radic. Biol. Med.*, vol. 184, pp. 170–184, May 2022, doi: 10.1016/j.freeradbiomed.2022.03.029.

[181] M. D. Wellenstein *et al.*, “Loss of p53 triggers WNT-dependent systemic inflammation to drive breast cancer metastasis,” *Nature*, vol. 572, no. 7770, Art. no. 7770, Aug. 2019, doi: 10.1038/s41586-019-1450-6.

[182] K. L. Harper *et al.*, “Mechanism of early dissemination and metastasis in Her2+ mammary cancer,” *Nature*, vol. 540, no. 7634, Art. no. 7634, Dec. 2016, doi: 10.1038/nature20609.

[183] R. Rosenthal, N. McGranahan, J. Herrero, B. S. Taylor, and C. Swanton, “deconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution,” *Genome Biol.*, vol. 17, no. 1, p. 31, Feb. 2016, doi: 10.1186/s13059-016-0893-4.

[184] D. Chakravarty *et al.*, “OncoKB: A Precision Oncology Knowledge Base,” *JCO Precis. Oncol.*, no. 1, pp. 1–16, Nov. 2017, doi: 10.1200/PO.17.00011.

[185] X. Liu, C. Li, C. Mou, Y. Dong, and Y. Tu, “dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs,” *Genome Med.*, vol. 12, no. 1, p. 103, Dec. 2020, doi: 10.1186/s13073-020-00803-9.

[186] I. Adzhubei, D. M. Jordan, and S. R. Sunyaev, “Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2,” *Curr. Protoc. Hum. Genet. Editor. Board Jonathan Haines Al*, vol. 0 7, p. Unit7.20, Jan. 2013, doi: 10.1002/0471142905.hg0720s76.

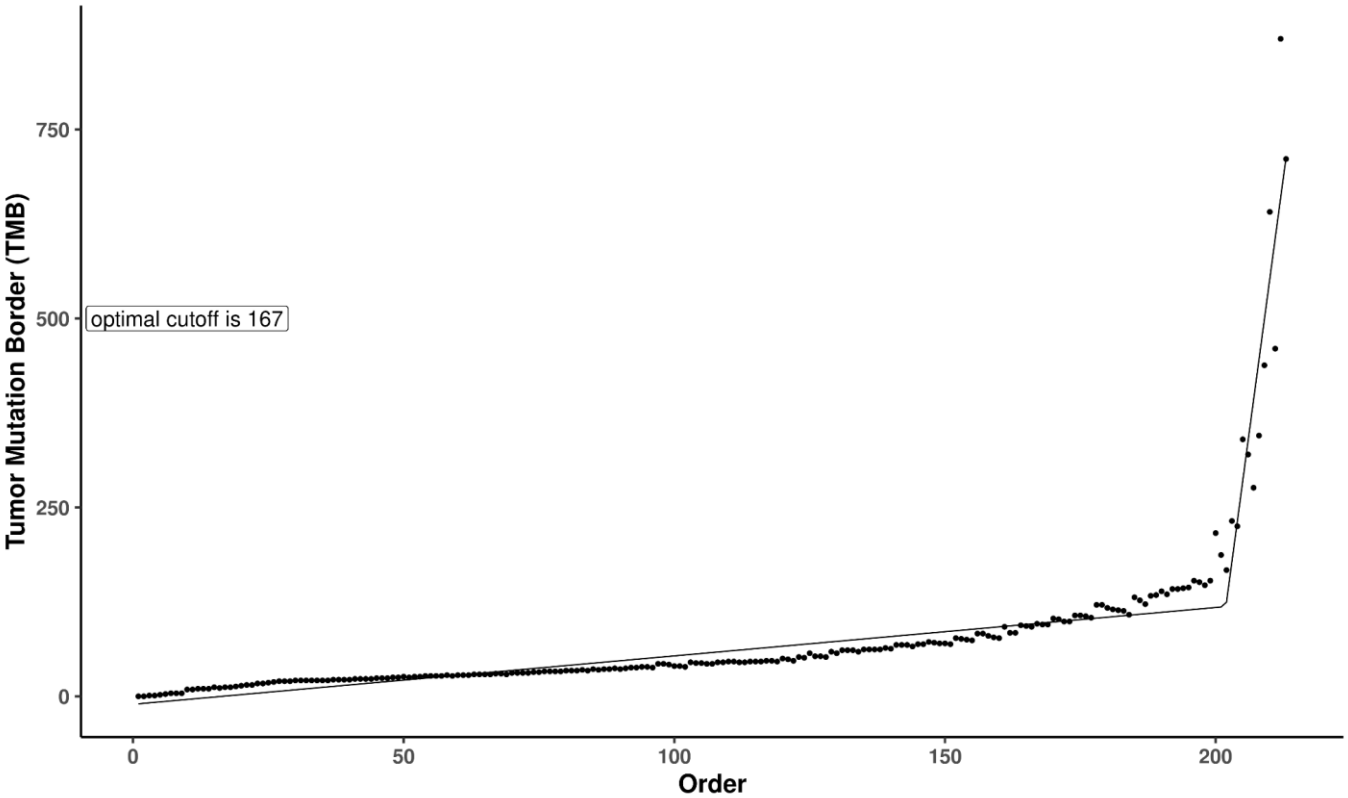
[187] K. Wang, M. Li, and H. Hakonarson, “ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data,” *Nucleic Acids Res.*, vol. 38, no. 16, p. e164, Sep. 2010, doi: 10.1093/nar/gkq603.

[188] “International network of cancer genome projects,” *Nature*, vol. 464, no. 7291, pp. 993–998, Apr. 2010, doi: 10.1038/nature08987.

[189] Z. Ge *et al.*, “Integrated Genomic Analysis of the Ubiquitin Pathway across Cancer Types,” *Cell Rep.*, vol. 23, no. 1, pp. 213-226.e3, Apr. 2018, doi: 10.1016/j.celrep.2018.03.047.

[190] M. Gerstung *et al.*, “Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes,” *Nat. Commun.*, vol. 6, no. 1, Art. no. 1, Jan. 2015, doi: 10.1038/ncomms6901.

8 APPENDIX



Appendix Figure 1**.** Segmented regression analysis conducted on the IGR dataset. The line represents the fitted segmented regression line. Each dot is patient. The y-axis represents TMB meanwhile, the x-axis represents the rank of the respective TMB in the IGR cohort. Optimal cutoff suggested by the algorithm

Appendix Table 1. Fihser's exact test results for the differentially mutated gene analysis after removal of the hypermutated samples identified by the segmented regression analysis. The table displays the top 25 genes with the lowest p-value.

Hugo_Symbol	IGR	TCGA	pval	or	ci.up	ci.low	adjPval
ESR1	18	10	8.45E-09	10.12489	24.99324	4.349141	5.22E-05
FSIP2	13	10	9.24E-06	7.11771	18.42092	2.835336	0.028509
MALRD1	7	1	1.84E-05	37.40171	1677.506	4.757801	0.037756

EDC4	7	6	0.001889	6.211844	22.63762	1.766043	1
AGMO	6	4	0.001986	7.958924	38.68435	1.867044	1
SREBF1	6	4	0.001986	7.958924	38.68435	1.867044	1
CACNA1I	7	7	0.003266	5.320537	17.99814	1.572785	1
IGFN1	8	10	0.004138	4.267748	12.18625	1.443532	1
MYH9	0	31	0.005789	0	0.64158	0	1
CDH1	12	129	0.007206	0.448229	0.832258	0.221047	1
ANKRD31	5	4	0.007641	6.598595	33.55508	1.406398	1
RUNX1	1	40	0.009039	0.125108	0.747749	0.003077	1
LRP2	1	41	0.009111	0.121933	0.72774	0.003001	1
AGRN	6	7	0.010565	4.537771	15.9621	1.245241	1
OSBPL3	6	7	0.010565	4.537771	15.9621	1.245241	1
PALB2	6	7	0.010565	4.537771	15.9621	1.245241	1
TENM3	9	16	0.012198	3.000288	7.337822	1.1506	1
P2RX2	5	5	0.013296	5.275588	23.15896	1.201778	1
MYO18B	0	28	0.015561	0	0.717029	0	1
ZFHX4	3	54	0.016637	0.276823	0.867279	0.054812	1
CNKSR2	7	11	0.016959	3.374688	9.674692	1.094606	1
MUC16	11	112	0.019884	0.479714	0.914865	0.228175	1

ATP1A2	5	6	0.021221	4.393247	17.47162	1.049513	1
NPAP1	5	6	0.021221	4.393247	17.47162	1.049513	1
PSD	5	6	0.021221	4.393247	17.47162	1.049513	1

Appendix Table 2. OncodriveCLUST analysis results of the IGR cohort conducted after the removal of the hypermutated samples identified by the segmented regression analysis.

Hugo_Symbol	clusters	muts_in_clusters	pval	fdr	fract_muts_in_clusters
AKT1	1	7	6.23E-05	0.000436	0.777778
PIK3CA	4	50	4.34E-05	0.000436	0.833333
ESR1	2	15	0.000428	0.001999	0.789474
ERBB2	1	4	0.026953	0.094336	0.571429
P2RX2	1	2	0.044565	0.124783	0.4
GATA3	3	11	0.078951	0.18422	0.5
MUC4	1	2	0.175986	0.351973	0.4
TP53	5	34	0.250426	0.438245	0.414634
NCOR1	1	2	0.337992	0.525766	0.333333
FASN	1	2	0.479404	0.671166	0.285714
NF1	1	2	0.668854	0.851269	0.2
ANK3	1	2	0.772635	0.901407	0.181818
MAP3K1	1	2	0.915732	0.915732	0.1

NEB	1	2	0.852508	0.915732	0.142857

Appendix Table 3. Results of clinical enrichment analysis of the IGR cohort conducted on HR/HER groups after the removal of the hypermutated samples identified by the segmented regression analysis. The table displays the top 25 genes with the lowest p-value.

Hugo_Symbol	Group1	Group2	p_value	OR	OR_low	OR_high	fdr
TP53	HR+/HER2-	Rest	5.11E-10	0.120908	0.055016	0.25343	1.03E-07
TP53	HR-/HER2-	Rest	1.86E-09	9.143833	4.016464	22.60313	1.87E-07
PIK3CA	HR-/HER2-	Rest	0.000609	0.18907	0.046534	0.566386	0.040782
TTN	HR-/HER2-	Rest	0.000914	4.085193	1.657595	10.17884	0.045928
PIK3CA	HR+/HER2-	Rest	0.001401	3.762383	1.52915	10.64341	0.050494
MYH4	HR+/HER2-	Rest	0.001507	0.06004	0.001308	0.485533	0.050494
TTN	HR+/HER2-	Rest	0.001857	0.261585	0.104481	0.634905	0.053336
ESR1	HR+/HER2-	Rest	0.00331	Inf	1.952446	Inf	0.083155
MYH4	HR-/HER2-	Rest	0.003795	9.780703	1.669947	102.5812	0.084755
ANK3	HER2+	Rest	0.009242	11.11318	1.496145	68.61355	0.185758
CDH1	HR+/HER2-	Rest	0.010537	Inf	1.362525	Inf	0.192532
ESR1	HR-/HER2-	Rest	0.013252	0	0	0.711346	0.221968
MAP3K1	HR-/HER2-	Rest	0.022634	0	0	0.839123	0.349961
GATA3	HR-/HER2-	Rest	0.028349	0.137504	0.003225	0.914709	0.407005
HMCN1	HR+/HER2-	Rest	0.034	0.175043	0.01619	1.108494	0.429632
RELN	HR-/HER2-	Rest	0.037211	6.241315	0.861638	71.32585	0.429632

CDH1	HR-/HER2-	Rest	0.038952	0	0	1.015854	0.429632
GATA3	HR+/HER2-	Rest	0.039292	4.674167	1.059548	42.8942	0.429632
MAP3K1	HR+/HER2-	Rest	0.040612	6.569355	0.943268	285.3341	0.429632
XIRP2	HR+/HER2-	Rest	0.057263	Inf	0.813048	Inf	0.575492
FAT1	HER2+	Rest	0.065083	6.735147	0.574475	48.13467	0.616098
HMCN1	HR-/HER2-	Rest	0.070499	4.142383	0.673041	29.3826	0.616098
KMT2D	HR-/HER2-	Rest	0.070499	4.142383	0.673041	29.3826	0.616098
NLRP3	HR+/HER2-	Rest	0.081817	0.222456	0.019594	1.603276	0.657808
RELN	HR+/HER2-	Rest	0.081817	0.222456	0.019594	1.603276	0.657808

Appendix Table 4. Results of clinical enrichment analysis of the IGR cohort conducted on biopsy groups after the removal of the hypermutated samples identified by the segmented regression analysis. The table displays the top 25 genes with the lowest p-value.

Hugo_Symbol	Group1	Group2	p_value	OR	OR_low	OR_high	fdr
IGSF10	Skin	Rest	0.006453	9.965137	1.57541	72.57813	1
MAP3K1	Liver	Rest	0.013206	3.896573	1.158244	13.45542	1
TTN	other	Rest	0.016861	0	0	0.804939	1
SREBF1	Lung	Rest	0.030457	Inf	0.816239	Inf	1
DYNC2H1	Skin	Rest	0.031362	7.176761	0.907553	56.8727	1
NEB	Liver	Rest	0.040087	0	0	1.047483	1

HMCN1	Lung	Rest	0.045685	Inf	0.535606	Inf	1
CACNA1I	Skin	Rest	0.050043	5.369767	0.74001	33.97332	1
FASN	Skin	Rest	0.050043	5.369767	0.74001	33.97332	1
SYNE1	Lung	Rest	0.055838	Inf	0.433566	Inf	1
TP53	Liver	Rest	0.065442	0.512463	0.234097	1.074345	1
GATA3	Lymph node	Rest	0.06574	2.901287	0.738937	9.738927	1
RYR2	other	Rest	0.094664	2.905518	0.612526	11.21834	1
NF1	other	Rest	0.099951	3.552648	0.538537	18.05557	1
PCLO	Liver	Rest	0.107576	3.176973	0.567693	17.78937	1
XIRP2	Liver	Rest	0.107576	3.176973	0.567693	17.78937	1
GATA3	Liver	Rest	0.110065	0.290288	0.031612	1.279339	1
FLG	Skin	Rest	0.130329	3.031684	0.473072	14.48624	1
SETD2	Lymph node	Rest	0.135534	4.224698	0.360819	31.66236	1
SYNE1	Skin	Rest	0.16363	2.639909	0.421301	12.05289	1
CACNA1E	Liver	Rest	0.16434	3.129818	0.405193	24.18465	1
COL12A1	Liver	Rest	0.16434	3.129818	0.405193	24.18465	1
DMXL1	Liver	Rest	0.16434	3.129818	0.405193	24.18465	1
OSBPL3	Liver	Rest	0.16434	3.129818	0.405193	24.18465	1

MYO3A	other	Rest	0.180033	3.447002	0.296825	25.57767	1

Appendix Table 5. Results of the somatic interaction analysis of the IGR cohort conducted after the removal of the hypermutated samples identified by the segmented regression analysis. The table displays the top 25 genes with the lowest p-value.

gene1	gene2	pValue	oddsRatio
ESR1	TP53	6.73E-05	0
GATA3	TP53	0.001908	0.136367
ANK3	PIK3CA	0.013955	5.770214
PIK3CA	CDH1	0.020205	4.074975
OBSCN	KMT2C	0.027392	6.476698
PIK3CA	FLG	0.027411	4.304566
MAP3K1	PIK3CA	0.031439	3.354074
RYR2	RYR3	0.033764	5.864784
MAP3K1	ESR1	0.035372	4.311693
FLG	TTN	0.042573	4.272645
MAP3K1	FSIP2	0.064396	4.24762
DNAH6	TENM3	0.069098	6.303627
SYNE1	PIK3CA	0.073825	3.424226
PIK3CA	TP53	0.074246	0.536336

TTN	ESR1	0.080601	0
RYR3	AKT1	0.082422	5.585164
RYR3	HMCN1	0.082422	5.585164
RYR3	MAP2K4	0.082422	5.585164
NEB	PIK3CA	0.093368	2.836548
ANK3	NEB	0.096527	5.008023
TTN	TP53	0.099807	2.04354
RYR3	FLG	0.099866	4.870857
SYNE1	FLG	0.099866	4.870857
RYR3	FRAS1	0.099866	4.870857
ANK3	FSIP2	0.111332	4.534462

Appendix Table 6. Results of the differentially mutated pathway analysis conducted on the IGR cohort

Hugo_Symbol	IGR	TCGA	pval	or	ci.up	ci.low	adjPval
Nucleotide Excision Repair	37	121	0.031222	1.579234	2.391035	1.025241	0.549749
DDR	154	668	0.036919	1.423285	2.018543	1.013922	0.549749
NRF2	8	14	0.038797	2.828469	7.336885	1.014171	0.549749
WNT	51	183	0.042646	1.457408	2.097546	1.001611	0.549749

MYC	16	43	0.049599	1.863079	3.457933	0.959234	0.549749
Others - DDR	117	492	0.058902	1.335431	1.821356	0.981035	0.549749
LIN37 dependent P53 Targets	123	527	0.0819	1.307263	1.787632	0.958763	0.577805
hnRNP	4	46	0.085891	0.409375	1.139615	0.105855	0.577805
LIN37 independent P53 Targets	123	656	0.100184	0.775748	1.063196	0.567536	0.577805
B	1	23	0.103179	0.206574	1.28667	0.004991	0.577805
Cell_Cycle	19	60	0.120841	1.582661	2.762768	0.871182	0.615191
NOTCH	66	273	0.203963	1.245199	1.73602	0.887015	0.846126
1st step	2	3	0.205187	3.239756	28.46962	0.269045	0.846126
TP53	89	381	0.212903	1.223289	1.67059	0.893494	0.846126
U2 rel	0	10	0.226641	0	2.154746	0	0.846126
SR	2	4	0.27454	2.428359	17.06843	0.218327	0.960889
Homologous Recomination - Core	31	124	0.306432	1.244211	1.925676	0.785323	0.964403
Mismatch Repair - Core	9	63	0.335339	0.677001	1.397969	0.291128	0.964403
Hippo	57	246	0.380624	1.164385	1.645497	0.816141	0.964403
U4/U6	3	7	0.388974	2.085716	9.230418	0.345223	0.964403

Base Excision Repair - Core	9	31	0.391452	1.420978	3.116791	0.585771	0.964403
A	8	56	0.394653	0.678591	1.460388	0.275058	0.964403
RB-E2F Targets	154	714	0.407858	1.157792	1.645115	0.822961	0.964403
U11/U12	3	8	0.413316	1.82345	7.677954	0.309033	0.964403
U5	4	31	0.495721	0.6167	1.771622	0.15647	1
Damage Sensor - Core	10	62	0.521895	0.768892	1.544164	0.34544	1
Sm	1	3	0.529099	1.613552	20.20591	0.030606	1
MISC	15	62	0.535289	1.182311	2.156453	0.611588	1
Homology-dependent recombination	59	263	0.546989	1.117135	1.571811	0.786866	1
Base Excision Repair	30	129	0.572689	1.144638	1.77679	0.719275	1
RTK-RAS	73	331	0.573083	1.101031	1.519673	0.793274	1
Splicing Factor Genes	70	357	0.634336	0.922031	1.274396	0.662835	1
Fanconi Anemia	39	178	0.692745	1.072522	1.589821	0.710824	1
Nucleotide pools	3	13	0.745579	1.11716	4.115423	0.20242	1
Prp19	2	15	0.752	0.641352	2.790431	0.070648	1

Direct Repair	2	17	0.757367	0.564815	2.408266	0.062828	1
Translesion Synthesis	15	81	0.778524	0.887272	1.592161	0.464739	1
Mismatch Repair	20	92	0.793693	1.056185	1.778255	0.601352	1
Non-homologous End Joining	24	109	0.807412	1.072607	1.735493	0.641008	1
PI3K	100	493	0.820714	0.96305	1.310018	0.707356	1
B act	6	34	0.833562	0.848889	2.081493	0.287535	1
C	3	14	1	1.036375	3.759936	0.189293	1
C2	2	11	1	0.877862	4.066691	0.093882	1
DREAM targets	178	857	1	1.02591	1.593709	0.675557	1
Direct Repair - Core	1	6	1	0.804895	6.687209	0.017414	1
EJC/TREX	1	9	1	0.535145	3.897751	0.01215	1
Fanconi Anemia - Core	10	48	1	1.007457	2.060046	0.446762	1
MBM FOXM1 Targets	110	529	1	1.010866	1.375965	0.743167	1
Non-homologous End Joining - Core	9	44	1	0.988307	2.093364	0.417414	1

Nucleotide Excision Repair - Core	9	47	1	0.922452	1.943381	0.391029	1
Prp19 rel	1	4	1	1.209298	12.30428	0.024445	1
RES	0	2	1	0	25.79499	0	1
TGF-Beta	8	42	1	0.917744	2.017658	0.366503	1
Translesion Synthesis - Core	10	51	1	0.945295	1.923915	0.420614	1
U2	5	28	1	0.860018	2.292543	0.256266	1
integrator complex	3	17	1	0.851068	2.981071	0.158357	1

Appendix Table 7. Results of the differentially mutated pathway analysis conducted on the IGR cohort after the removal of the hypermutated samples identified by the segmented regression analysis.

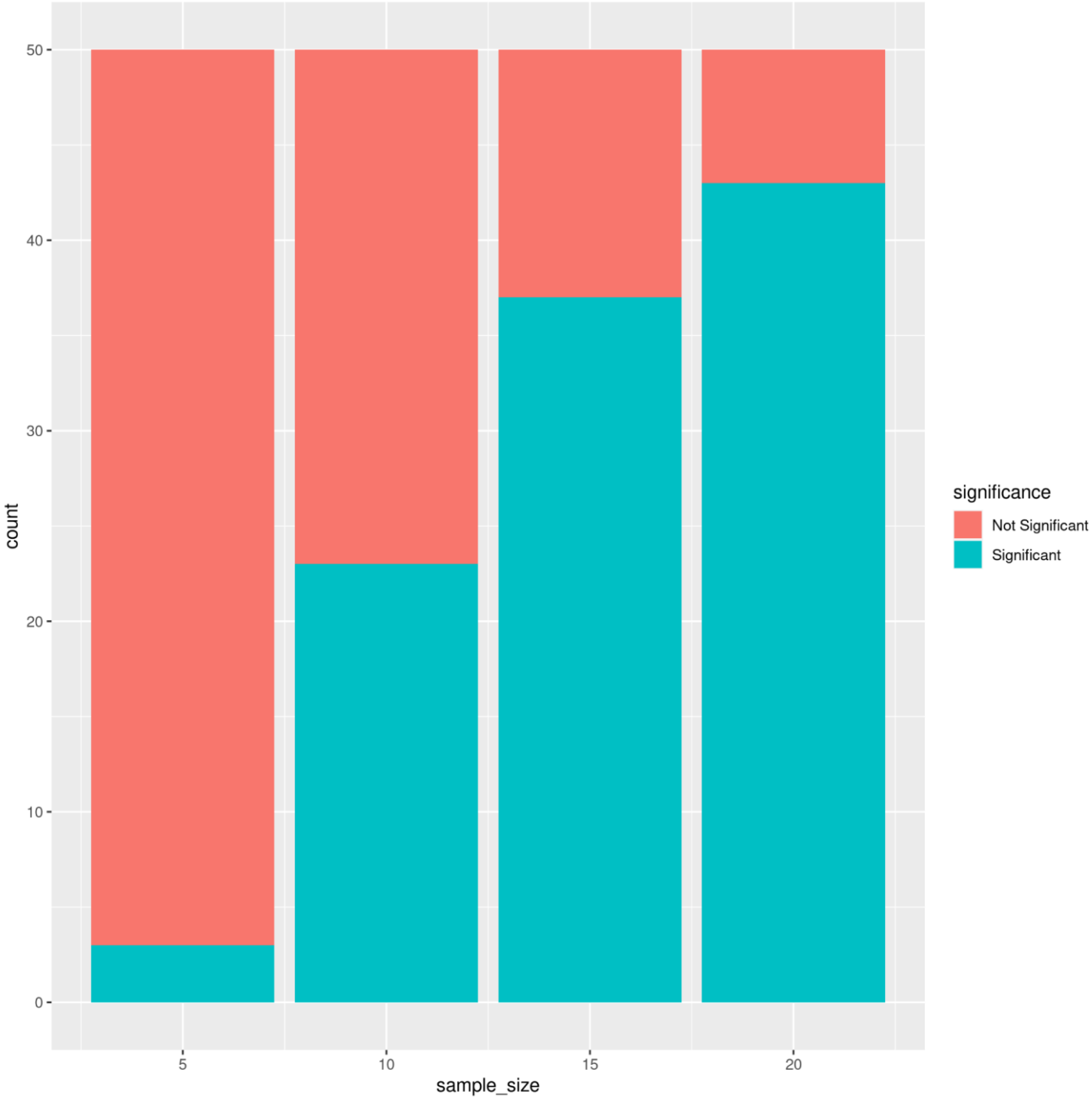
Hugo_Symbol	IGR	TCGA	pval	or	ci.up	ci.low	adjPval
Mismatch Repair - Core	4	63	0.016304	0.315032	0.862576	0.08231	0.543783
LIN37 independent P53 Targets	109	656	0.019421	0.687538	0.948546	0.499212	0.543783
U5	1	31	0.047959	0.162904	0.989773	0.003978	0.819321
NRF2	6	14	0.117665	2.255372	6.343598	0.701329	0.819321

hnRNP	4	46	0.119489	0.439059	1.22343	0.113462	0.819321
Damage Sensor - Core	6	62	0.12485	0.48563	1.140684	0.169216	0.819321
DDR	140	668	0.138477	1.293987	1.842049	0.917799	0.819321
Splicing Factor Genes	58	357	0.139988	0.775081	1.09104	0.545502	0.819321
B	1	23	0.157617	0.221326	1.379488	0.005346	0.819321
WNT	43	183	0.194963	1.276822	1.875946	0.856351	0.819321
A	6	56	0.213222	0.540996	1.278038	0.187764	0.819321
Others - DDR	104	492	0.244192	1.199922	1.649212	0.873965	0.819321
Nucleotide Excision Repair - Core	5	47	0.247702	0.539349	1.375135	0.165241	0.819321
SR	2	4	0.251513	2.602396	18.29989	0.233893	0.819321
B act	3	34	0.254761	0.448686	1.449052	0.0873	0.819321
MYC	12	43	0.259888	1.473261	2.911338	0.693494	0.819321
TP53	82	381	0.262739	1.195703	1.648752	0.864534	0.819321
PI3K	87	493	0.311233	0.845572	1.161986	0.613942	0.819321
Mismatch Repair	13	92	0.331627	0.712852	1.31484	0.357955	0.819321
Cell_Cycle	15	60	0.334709	1.318356	2.413941	0.679866	0.819321

Direct Repair	1	17	0.33651	0.301216	1.944463	0.00717	0.819321
Nucleotide Excision Repair	28	121	0.344244	1.230743	1.939434	0.760224	0.819321
Fanconi Anemia - Core	6	48	0.350294	0.636335	1.518224	0.219409	0.819321
LIN37 dependent P53 Targets	109	527	0.351137	1.158583	1.594998	0.843285	0.819321
U2 rel	0	10	0.380714	0	2.308886	0	0.844281
Fanconi Anemia	29	178	0.407343	0.816684	1.262865	0.513708	0.844281
TGF-Beta	5	42	0.417599	0.606616	1.559076	0.184923	0.844281
MBM FOXM1 Targets	96	529	0.43689	0.882309	1.211307	0.642368	0.844281
Non-homologous End Joining	17	109	0.444375	0.789496	1.363286	0.432867	0.844281
Translesion Synthesis	12	81	0.463683	0.752118	1.422654	0.365683	0.844281
Translesion Synthesis - Core	7	51	0.46737	0.700187	1.581518	0.264025	0.844281
Prp19	1	15	0.492616	0.342044	2.248206	0.008083	0.850413
1st step	1	3	0.507026	1.728662	21.65481	0.032781	0.850413

Hippo	43	246	0.523016	0.878605	1.280824	0.593351	0.850413
Homology-dependent recombination	46	263	0.531508	0.876922	1.267439	0.598293	0.850413
Non-homologous End Joining - Core	6	44	0.555946	0.696999	1.673439	0.239333	0.864805
Prp19 rel	1	4	0.587089	1.295617	13.18782	0.026182	0.880245
Direct Repair - Core	0	6	0.597309	0	4.406567	0	0.880245
RTK-RAS	60	331	0.617631	0.911691	1.281371	0.643199	0.883879
Homologous Recomination - Core	21	124	0.631342	0.862273	1.422797	0.501311	0.883879
U11/U12	2	8	0.669104	1.297133	6.568014	0.133249	0.913898
MISC	10	62	0.741235	0.826413	1.661972	0.37086	0.970995
integrator complex	2	17	0.754449	0.605342	2.582932	0.067307	0.970995
RB-E2F Targets	140	714	0.799027	1.052592	1.50134	0.74499	0.970995
U2	4	28	0.807328	0.734441	2.133483	0.185097	0.970995
Base Excision Repair	23	129	0.813903	0.913058	1.480626	0.542768	0.970995
Base Excision Repair - Core	5	31	0.82196	0.830933	2.193184	0.249084	0.970995

DREAM targets	164	857	0.832281	0.945267	1.471829	0.620785	0.970995
EJC/TREX	1	9	1	0.573349	4.178306	0.013014	1
C	2	14	1	0.737164	3.249187	0.080691	1
C2	2	11	1	0.940824	4.36115	0.100575	1
NOTCH	53	273	1	1.007091	1.434804	0.699262	1
Nucleotide pools	2	13	1	0.794617	3.552554	0.086395	1
RES	0	2	1	0	27.61366	0	1
Sm	0	3	1	0	12.55994	0	1
U4/U6	1	7	1	0.738505	5.800112	0.0163	1



Appendix Figure 2. Contribution of the ROS related signature was tested for difference between NRF2 mutant and wildtype samples of the IGR dataset. Comparison was conducted with the subsampling of the NRF2 wildtype samples into the sample sizes of 5, 10, 15, and 20 with 50 repetitions for each sample size. Red color denotes the non-significant results, while blue color denotes the significant results. Comparisons were made with the Student's t-test and $p < 0.05$ was treated as significant.

Appendix Table 8. GSEA results of CHRNA5 co-occurring genes in logFCs of the siA5, siTP53, and siA5+siTP53 treated MCF7 cell lines.

pathway	pval	padj	log2err	ES	NES	size	Logfcs_Used
CHRNA5 Co-occuring genes	0.651706	0.651706	0.060509	-0.16266	-0.91975	147	siA5 LogFCs FDR<0.05
CHRNA5 Co-occuring genes	0.274212	0.274212	0.082434	-0.23379	-1.12462	136	siP53 LogFCs FDR<0.05
CHRNA5 Co-occuring genes	0.256243	0.256243	0.082205	-0.24309	-1.13054	x	siA5+P53 LogFCs FDR<0.05
CHRNA5 Co-occuring genes	0.829365	0.829365	0.053898	-0.16572	-0.8549	225	siA5 LogFCs p.value<0.05
CHRNA5 Co-occuring genes	0.271131	0.271131	0.079771	-0.26371	-1.09755	221	siP53 LogFCs p.value<0.05
CHRNA5 Co-occuring genes	0.54717	0.54717	0.043713	-0.22784	-0.96986	260	siA5+P53 LogFCs p.value<0.05

Appendix Table 9. GSEA results of CHRNA5 co-occuring genes in logFCs of the siA5, siTP53, and siA5+siTP53 treated MCF7 cell lines.

pathway	pval	padj	log2err	ES	NES	size	Logfcs_Used
P53 Co-occuring	0.631808	0.631808	0.071479	0.178296	0.896175	73	siA5 LogFCs FDR<0.05

P53 Co-occuring	0.207792	0.207792	0.104733	-0.28704	-1.21718	61	siP53 LogFCs FDR<0.05
P53 Co-occuring	0.282325	0.282325	0.081752	-0.26523	-1.1291	97	siA5+P53 LogFCs FDR<0.05
P53 Co-occuring	0.864811	0.864811	0.052057	0.173261	0.790968	112	siA5 LogFCs p.value<0.05
P53 Co-occuring	0.442893	0.442893	0.062386	-0.26529	-1.00398	92	siP53 LogFCs p.value<0.05
P53 Co-occuring	0.64342	0.64342	0.039517	-0.22774	-0.91384	138	siA5+P53 LogFCs p.value<0.05

Appendix Table 10. Overview of the CHRNA5 mutations in the TCGA-BRCA dataset

Tumor_Sample_Barcode	HGVSp_Short	t_depth	t_ref_count	t_alt_count	VAF	TMB
TCGA-A2-A0T1	p.N155S	217	199	18	0.082949	23
TCGA-A8-A09G	p.E88D	892	867	25	0.028027	234
TCGA-AC-A23H	p.I152M	174	90	84	0.482759	4161

TCGA-AN-A046	p.W110C	404	308	96	0.237624	5366
TCGA-BH-A1F0	p.T254N	269	257	12	0.04461	59