ANALYZING GENE REPLICATION TIME USING DNA REPLICATION TIME

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS OF
THE MIDDLE EAST TECHNICAL UNIVERSITY
BY

ABDULLAH TERCAN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
BIOINFORMATICS

JUNE 2026

ANALYZING GENE REPLICATION TIME USING DNA REPLICATION TIME

submitted by **ABDULLAH TERCAN** in partial fulfillment of the requirements for the degree of **Master of Science in Health Informatics Department, Middle East Technical University** by,

Prof. Dr. Banu Günel Kılıç _____
Dean, **Graduate School of Informatics**

Prof. Dr. Yeşim Aydın Son _____
Head of Department, **Health Informatics**

Assist. Prof. Dr. Burçak Otlu Sarıtaş _____
Supervisor, **Health Informatics, METU**

Examining Committee Members:

Prof. Dr. Yeşim Aydın Son
Health Informatics, METU _____

Assist. Prof. Dr. Burçak Otlu Sarıtaş
Health Informatics, METU _____

Prof. Dr. Tunca Doğan
Computer Engineering, Hacettepe University _____

Date: 23.06.2026

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Abdullah Tercan

Signature :

ABSTRACT

ANALYZING GENE REPLICATION TIME USING DNA REPLICATION TIME

Tercan, Abdullah

M.S., Department of Health Informatics

Supervisor: Assist. Prof. Dr. Burçak Otlu Sarıtaş

June 2026, 105 pages

DNA replication, while essential for genetic inheritance, is also a significant source of mutations. This process is highly orchestrated, with different genomic regions replicating at specific times during the S phase of the cell cycle. The replication timing is intimately linked to mutation rates, both in germline and somatic cells. Certain mutational signatures, such as those associated with UV exposure and tobacco smoking, exhibit a preference for late-replicating regions, while others, like those linked to DNA repair deficiency and alcohol consumption, tend to occur in early-replicating regions. This specificity highlights the complex relationship between replication timing and the mutation rate. Replication timing of genomic regions can be used to explain the accumulation of mutations. In this study, we quantitatively analyzed the replication timing of protein coding genes across multiple cell lines using SigProfilerTopography. We assigned higher scores to genes that replicate earlier. We developed a model that can predict the replication time of protein coding genes. However, since the expected results could not be obtained, the approach of analyzing the replication time differences between protein coding genes approach. This research has the potential to provide an understanding of whether protein coding genes replicate early or late in different cell lines. It can serve as a resource for future studies focusing on mutation rate, and genome evolution.

Keywords: DNA Replication time, bioinformatics, gene ontology

ÖZ

DNA REPLİKASYON VERİSİ KULLANILARAK GEN REPLİKASYON SÜRESİ ANALİZİ

Tercan, Abdullah

Yüksek Lisans, Sağlık Bilişimi Bölümü

Tez Yöneticisi: Dr. Öğr. Üyesi. Burçak Otlu Sarıtaş

Haziran 2026, 105 sayfa

DNA replikasyonu, genetik mirasın devamı için gerekli olmasının yanı sıra mutasyonların önemli bir kaynağıdır. Bu süreç oldukça düzenli şekilde gerçekleşir ve genomun farklı bölgeleri hücre döngüsünün S fazında belirli zamanlarda replike olur. Replikasyon zamanlaması, hem germline hem de somatik hücrelerdeki mutasyon oranlarıyla yakından ilişkilidir. UV maruziyeti ve tütün kullanımıyla ilişkili bazı mutasyon imzaları geç replike olan bölgelerde daha sık görülürken, DNA onarım eksikliği ve alkol tüketimiyle ilişkili imzalar erken replike olan bölgelerde daha yaygındır. Bu durum, replikasyon zamanlaması ile mutasyon oranı arasındaki karmaşık ilişkiyi ortaya koymaktadır. Bu çalışmada, SigProfilerTopography kullanılarak farklı hücre hatlarında protein kodlayan genlerin replikasyon zamanlaması nicel olarak analiz edilmiştir. Erken replike olan genlere daha yüksek skorlar atanmıştır. Protein kodlayan genlerin replikasyon zamanını tahmin edebilen bir model geliştirilmiştir. Ancak beklenen sonuçlar elde edilemediği için çalışma, protein kodlayan genler arasındaki replikasyon zamanı farklılıklarının incelenmesine yönelmiştir. Bu araştırma, farklı hücre hatlarında protein kodlayan genlerin erken mi yoksa geç mi replike olduğunu anlamaya katkı sağlayabilir. Ayrıca mutasyon oranı ve genom evrimi üzerine yapılacak gelecekteki çalışmalar için bir kaynak niteliği taşıyabilir.

Anahtar Kelimeler: DNA replikasyon süresi, biyoenformatik, gen ontolojisi

To my family...

ACKNOWLEDGMENTS

I would like to express my sincere appreciation to my supervisor, Dr. Burçak Otlu Sarıtaş, for their invaluable guidance, support, and encouragement throughout this thesis. The numerical calculations reported in this paper were fully/partially performed at TUBITAK ULAKBIM, High Performance and Grid Computing Center (TRUBA resources).

TABLE OF CONTENTS

ABSTRACT iv

ÖZ v

DEDICATION vi

ACKNOWLEDGMENTS vii

TABLE OF CONTENTS viii

LIST OF TABLES xiii

LIST OF FIGURES xiv

LIST OF ABBREVIATIONS xix

CHAPTERS

1 INTRODUCTION 1

1.1 Research Questions 2

1.2 Target of Research 3

1.3 Importance of Research 3

2 LITERATURE REVIEW 5

3 METHODOLOGY 11

3.1 Research Design 11

3.2 Methodology for Model 11

3.2.1 Data Collection 11
3.2.2 VCF Files 12
3.2.3 SigProfilerTopography 12
3.2.4 Methods 13
3.2.4.1 Ridge Regression 15
3.2.4.2 Random Forest Regressor 16
3.2.4.3 XGBRegressor 16
3.2.4.4 Artifical Neural Network 17
3.2.5 Parameters 18
3.2.6 Metrics 18
3.2.7 Correlation Matrix 19
3.2.8 Feature Importance 19
3.3 Methodology for GO Terms 20
3.3.1 Data Collection For GO Term Analysis 20
3.3.2 Creating DataFrame 21
3.3.3 Score 21
3.3.4 Early and Late Genes 21
3.4 Workflow 23
4 RESULTS 25
4.1 Results For Model 25
4.1.1 Model Predictions 26
4.1.2 Random Forest Feature Importance 27

4.1.3 Correlation Matrix 28

4.2 Results for GO Term Analysis for Common Pathways 29

4.2.1 Pathways in Detail 33

4.2.1.1 Biological Process - Response to Stimulus 33

4.2.1.2 Biological Process - Multicellular Organismal Process 34

4.2.1.3 Biological Process - Localization 35

4.2.1.4 Biological Process - Immune System Process 35

4.2.1.5 Biological Process - Homeostatic Process 35

4.2.1.6 Biological Process - Developmental Process 36

4.2.1.7 Biological Process - Cellular Process 36

4.2.1.8 Biological Process - Biological Regulation 36

4.2.1.9 Biological Process - Biological Process Involved In Interspecies Interaction Between 37

4.2.1.10 Cellular Component - Protein Containing Complex 37

4.2.1.11 Cellular Component - Cellular Anatomical Structure 38

4.2.1.12 Molecular Function - Transporter Activity 39

4.2.1.13 Molecular Function - Molecular Transducer Activity 39

4.2.1.14 Molecular Function - Molecular Function Regulator Activity . . . 39

4.2.1.15 Molecular Function - Catalytic Activity 40

4.2.1.16 Molecular Function - Binding 40

4.2.2 DNA Replication Time Distribution 40

4.2.3 Results for GO Term Analysis for Individual Cell Lines 43

4.2.3.1 BG02ES 43

4.2.3.2 BJ 44
4.2.3.3 GM12878 44
4.2.3.4 HELAS3 44
4.2.3.5 HEPG2 44
4.2.3.6 IMR90 44
4.2.3.7 K562 45
4.2.3.8 MCF7 45
4.2.3.9 NHEK 46
4.2.3.10 SKNSH 46
5 DISCUSSION 47
5.1 Model 47
5.2 Pathways Affected by Early Replicated Genes 48
5.3 Pathways Affected by Late Replicated Genes 48
5.4 Cell Line Individual Pathway Analysis 49
5.4.1 BG02ES 49
5.4.2 BJ 49
5.4.3 GM12878 50
5.4.4 HELAS3 50
5.4.5 HEPG2 50
5.4.6 IMR90 50
5.4.7 K562 50
5.4.8 MCF7 50

5.4.9 NHEK 51

5.4.10 SKNSH 51

5.5 Contribution of This Study 51

REFERENCES 53

APPENDICES

A ANCESTRAL PATHWAYS 57

B CELL-LINE SPECIFIC PATHWAYS 67

C ANCESTOR WISE COMMON THEMES AND PATHWAYS 85

D CELL LINE SPECIFIC GO TERMS COMMON THEMES 99

LIST OF TABLES

Table 1 Cell Lines With Tissues 20

Table 2 Early and Late Replicating Gene Counts 22

Table 3 R^2 and MSE Scores For Algorithms 25

LIST OF FIGURES

Figure 1 Replication Time Figure. Source of Figure [1] 2

Figure 2 Schematic for Replication Timing Prediction Model 14

Figure 3 Workflow of Study 24

Figure 4 ANN Loss Graph 26

Figure 5 PTPRD with SBS15 Prediction 27

Figure 6 Correlation Matrix 28

Figure 7 Ancestor-Based GO Term Distribution for Biological Process 29

Figure 8 Ancestor-Based GO Term Distribution for Cellular Component 30

Figure 9 Ancestor-Based GO Term Distribution for Molecular Function 30

Figure 10 Ancestor-Based GO Terms for Biological Process 31

Figure 11 Ancestor-Based GO Terms for Cellular Component 32

Figure 12 Ancestor-Based GO Terms for Molecular Function 33

Figure 13 GO Terms for Response to Stimulus and Cell Lines 34

Figure 14 Common Themes for Early Replicating GO Terms Ancestors for Response to Stimulus 34

Figure 15 Common Themes for Late Replicating GO Terms Ancestors for Response to Stimulus 34

Figure 16 Cell Line Distribution 41

Figure 17 KDE Comparison for Cell Lines 41

Figure 18 Mean and Standard Deviation for Genes 42

Figure 19 Gene Replication Timing Data Distribution 42

Figure 20 BG02ES Specific Early Replicating GO Terms Common Themes for Biological Process 43

Figure 21 BG02ES Specific Early Replicating GO Terms Common Themes for Cellular Component 43

Figure 22 GO Terms for Multicellular Organismal Process and Cell Lines 85

Figure 23 Common Themes for Early Replicating GO Terms Ancestors for Multicellular Organismal Process 85

Figure 24 Common Themes for Late Replicating GO Terms Ancestors for Multicellular Organismal Process 86

Figure 25 GO Terms for Localization and Cell Lines 86

Figure 26 Common Themes for Early Replicating GO Terms Ancestors for Localization 86

Figure 27 Common Themes for Late Replicating GO Terms Ancestors for Localization 87

Figure 28 GO Terms for Immune System Process and Cell Lines 87

Figure 29 Common Themes for Late Replicating GO Terms Ancestors for Immune System Process 87

Figure 30 GO Terms for Homeostatic Process and Cell Lines 87

Figure 31 Common Themes for Early Replicating GO Terms Ancestors for Homeostatic Process 87

Figure 32 GO Terms for Developmental Process and Cell Lines 88

Figure 33 Common Themes for Early Replicating GO Terms Ancestors for Developmental Process 88

Figure 34 Common Themes for Late Replicating GO Terms Ancestors for Developmental Process 88

Figure 35 Common Themes for Early Replicating GO Terms Ancestors for Cellular Process 89

Figure 36 Common Themes for Late Replicating GO Terms Ancestors for Cellular Process 89

Figure 37 GO Terms for Cellular Process and Cell Lines 90

Figure 38 Common Themes for Early Replicating GO Terms Ancestors for Biological Regulation 91

Figure 39 Common Themes for Late Replicating GO Terms Ancestors for Biological Regulation 91

Figure 40 GO Terms for Biological Regulation and Cell Lines 92

Figure 41 GO Terms for Biological Process Involved in Interspecies Interaction Between and Cell Lines 92

Figure 42 Common Themes for Late Replicating GO Terms Ancestors for Biological Process Involved In Interspecies Interaction Between 92

Figure 43 GO Terms for Protein Containing Complex and Cell Lines 93

Figure 44 Common Themes for Early Replicating GO Terms Ancestors for Protein-Containing Complex 93

Figure 45 Common Themes for Late Replicating GO Terms Ancestors for Protein-Containing Complex 93

Figure 46 Common Themes for Early Replicating GO Terms Ancestors for Cellular Anatomical Structure 94

Figure 47 Common Themes for Late Replicating GO Terms Ancestors for Cellular Anatomical Structure 94

Figure 48 GO Terms for Cellular Anatomical Structure and Cell Lines 95

Figure 49 GO Terms for Transporter Activity and Cell Lines 95

Figure 50 Common Themes for Late Replicating GO Terms Ancestors for Transporter Activity 96

Figure 51 GO Terms for Molecular Transducer Activity and Cell Lines 96

Figure 52 Common Themes for Late Replicating GO Terms Ancestors for Molecular Transducer Activity 96

Figure 53 GO Terms for Molecular Function Regulatory Activity and Cell Lines . . . 96

Figure 54 Common Themes for Late Replicating GO Terms Ancestors for Molecular Function Regulator Activity 97

Figure 55 GO Terms for Catalytic Activity and Cell Lines 97

Figure 56 Common Themes for Late Replicating GO Terms Ancestors for Catalytic Activity 97

Figure 57 GO Terms for Binding and Cell Lines 97

Figure 58 Common Themes for Early Replicating GO Terms Ancestors for Binding 98

Figure 59 Common Themes for Late Replicating GO Terms Ancestors for Binding . 98

Figure 60 BJ Specific Early Replicating GO Terms Common Themes for Biological Process 99

Figure 61 BJ Specific Late Replicating GO Terms Common Themes for Biological Process 99

Figure 62 BJ Specific Late Replicating GO Terms Common Themes for Cellular Component 99

Figure 63 GM12878 Specific Early Replicating GO Terms Common Themes for Biological Process 99

Figure 64 GM12878 Specific Late Replicating GO Terms Common Themes for Biological Process 100

Figure 65 HELAS3 Specific Early Replicating GO Terms Common Themes for Biological Process 100

Figure 66 HELAS3 Specific Early Replicating GO Terms Common Themes for Cellular Component 100

Figure 67 HEPG2 Specific Early Replicating GO Terms Common Themes for Cellular Component 100

Figure 68 IMR90 Specific Early Replicating GO Terms Common Themes for Biological Process 101

Figure 69 IMR90 Specific Early Replicating GO Terms Common Themes for Cellular Component 101

Figure 70 K562 Specific Early Replicating GO Terms Common Themes for Biological Process 101

Figure 71 K562 Specific Early Replicating GO Terms Common Themes for Cellular Component 101

Figure 72 K562 Specific Late Replicating GO Terms Common Themes for Biological Process 102

Figure 73 K562 Specific Late Replicating GO Terms Common Themes for Cellular Component 102

Figure 74 K562 Specific Late Replicating GO Terms Common Themes for Molecular Function 102

Figure 75 MCF7 Specific Early Replicating GO Terms Common Themes for Biological Process 102

Figure 76 MCF7 Specific Late Replicating GO Terms Common Themes for Biological Process 103

Figure 77 MCF7 Specific Late Replicating GO Terms Common Themes for Molecular Function 103

Figure 78 NHEK Specific Early Replicating GO Terms Common Themes for Biological Process 103

Figure 79 NHEK Specific Late Replicating GO Terms Common Themes for Biological Process 103

Figure 80 NHEK Specific Late Replicating GO Terms Common Themes for Cellular Component 104

Figure 81 NHEK Specific Late Replicating GO Terms Common Themes for Molecular Function 104

Figure 82 SKNSH Specific Early Replicating GO Terms Common Themes for Cellular Component 104

Figure 83 SKNSH Specific Late Replicating GO Terms Common Themes for Biological Process 104

Figure 84 SKNSH Specific Late Replicating GO Terms Common Themes for Molecular Function 105

LIST OF ABBREVIATIONS

GO	Gene Ontology
SBS	Single Base Substitution
NGS	Next-Generation Sequencing
SNP	Single Nucleotide Polymorphism
RT	Replication Timing
VCF	Variant Call Format
MSE	Mean Squared Error
KDE	Kernel Density Estimation

CHAPTER 1

INTRODUCTION

Mutations may emerge as a result of external influences as well as it may occur during the cell division. For investigating mutations which occur during the cell division, replication time can be used to obtain general information. In cell division, which consists of 6 stages, replication time is calculated according to these phases.

Replication timing data shows exactly when a gene replicates during the cell cycle. This metric is derived by applying decreasing coefficients across six consecutive cell cycle fractions, assigning the highest weights to the earliest fractions and the lowest to the latest. Consequently, regions characterized by early replication yield relatively high values, whereas late-replicating regions exhibit lower scores.

By partitioning these calculated scores into equal bins, the replication timing status of the genomic features is classified as either early or late. Consequently, the presence of data within the initial bins, as opposed to the final bins, indicates that the feature resides in an early-replicating region. This system enables the characterization of given genomic features as either early- or late-replicating during DNA replication.

Figure 1 provides a schematic overview illustrating replication timing. In this representation, valleys correspond to late-replicating regions, while peaks represent early-replicating regions.

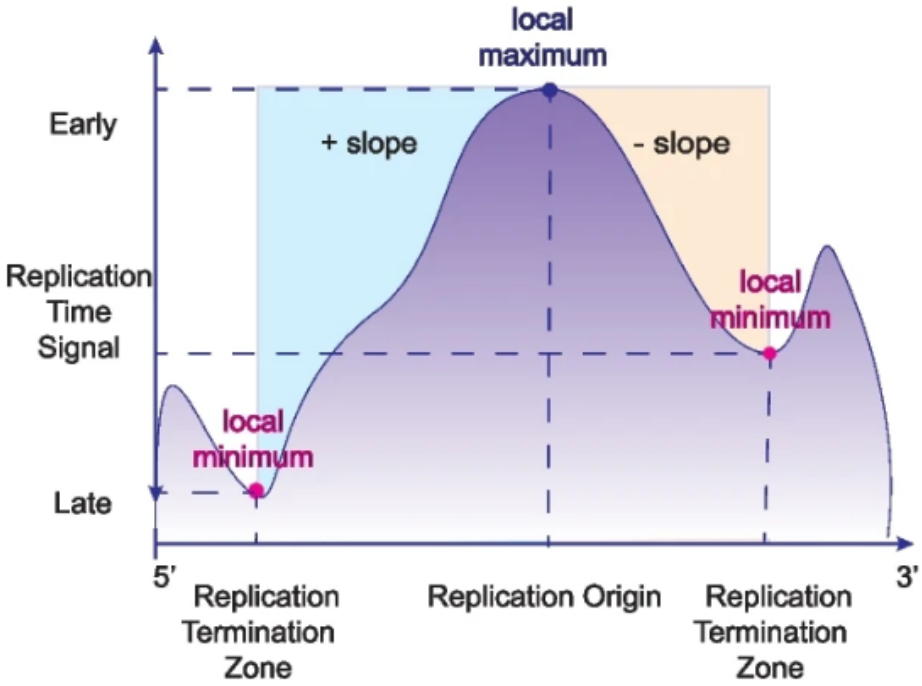

Figure 1: Replication Time Figure. Source of Figure [1]

Being able to estimate replication time seems to be a time-saving task for further studies. Also, replication time can show which gene replicated early or late during cell division. With this information, we can interpret the genes' behaviour in the course of cell division.

Despite existing literature on this topic, there is a clear gap regarding how gene behavior is influenced by replication timing. Further investigation into this gap is necessary as it may help us understand the replication times' effect on genes.

Since the calculation of replication time values takes a long time, this thesis' initial aim was to create a model that can predict the replication time. However, due to lack of data and resemblance between gene replication time and signature mutations of genes' replication time, the aim has changed to finding gene behaviors in terms of replication time for different cell lines. In this study whether protein coding genes have behaviors that depend on replication time is focused.

1.1 Research Questions

In this study, we first investigated the predictability of replication timing. Furthermore, we directed our study to see if we could explain the behavior of genes with replication times. The reason we conducted these studies was to contribute to the in-depth understanding of replication times and to make replication times more accessible through estimation. This thesis aims to address the following research questions: Can we develop a model that predicts replication

time? Is it possible to find a link between Single Base Substition (SBS) mutational signature replication time and gene replication time? Which protein coding genes replicate early and which replicate late in the specified cell lines? Do genes replicating at the same time give consistent results in terms of Gene Ontology (GO) terms?

1.2 Target of Research

Calculation of replication timing is a job that requires a lot of time, especially if it is calculated individually for different cell lines. To solve this problem, we tried to create a model which can predict the replication time. For this purpose, we used the replication timings of protein coding genes and the replication timings of SBS signatures. Using these values as features, we estimated the replication timing of overlapping nucleotides in the relevant signature in protein coding genes.

Replication timing is usually calculated for signatures to interpret reasonable information about the signature. In this study, we calculated replication timing for protein coding genes due to understanding the genes' behaviour in terms of mutation.

Another objective of this study was to determine the replication timing (early versus late) of specific protein coding genes within the designated cell lines. Consequently, this allows for an investigation into whether early- or late-replicating genes yield significant outcomes. Furthermore, the study aimed to identify the GO terms influenced by the retrieved genes, evaluating the results across three primary domains: molecular function, biological process, and cellular component.

These two aspects enable the model to predict mutations of protein coding genes within their corresponding SBS signatures, while simultaneously revealing the specific biological processes, macromolecular complexes, and functions these genes influence. By stratifying the predicted mutations into early- and late-replicating categories, subsequent GO Term analysis can pinpoint the exact GO terms affected. Additionally, this model-driven approach yields relatively shorter computational runtimes.

1.3 Importance of Research

This study shows us whether the replication time can be predicted for the relevant SBS signatures and what problems are encountered when it is. It also reveals the effects of early and late-replicating genes in certain cell lines. It can be seen which biological process, cellular component or molecular function is affected by early replicated or late replicated genes. Thus, the relationships between genes that replicate early or late during cell replication and the cell can be seen more clearly. This perspective is also useful in terms of creating an idea for further studies about replication timing.

CHAPTER 2

LITERATURE REVIEW

Researchers have conducted studies to calculate DNA replication time, its effects on systems, and its relationship with diseases. At the same time, DNA replication time has been a subject of study not only for human cells but also for many animal cells. Whether there is a similarity between the DNA replication times of human and animal cells is also one of the subjects of research. The effect of DNA replication time on genes and systems in the human body is another area of research. Another area of research has been whether the DNA replication time can be predicted with the machine learning models. This literature review discusses some studies on DNA replication time and the research gaps.

The organization and biological significance of DNA replication timing (RT) programs have been reviewed, discussing RT as a regulated genomic feature associated with chromatin organization and nuclear structure [2]. The study summarized evidence showing that some replication domains maintain relatively stable timing patterns, while others undergo changes during developmental processes and cellular differentiation. By examining the relationship between RT, chromatin states, and gene regulation, this review highlighted that the timing of genome duplication is coordinated with broader aspects of genome organization and cell-type-specific regulatory programs.

An important connection between genome architecture and RT organization was described by [3], which investigated the relationship between replication timing domains and topologically associating domains (TADs) identified through high-resolution chromosome conformation capture (Hi-C) analysis. Their study showed that replication domain boundaries frequently overlap with TAD boundaries, suggesting a close relationship between three-dimensional genome organization and the temporal regulation of DNA replication. These findings supported the idea that TADs represent stable structural units associated with the organization of replication timing programs during the S phase, providing insights into how spatial genome architecture relates to replication dynamics.

The dynamic changes in replication timing programs during mammalian development were investigated through genome-wide replication timing profiling of mouse embryonic stem cells (ESCs) and differentiated cell populations [4]. Their study showed that cellular differentiation is accompanied by substantial changes in the organization of replication domains, with a fraction of the genome undergoing shifts between early- and late-replication patterns. These changes were associated with alterations in the structure and distribution of replication domains, reflecting a reorganization of the temporal genome duplication program during development.

The study further observed that pluripotent ESCs display distinct replication timing patterns compared with differentiated cells, suggesting that RT organization is linked to changes in cellular state and lineage progression.

The relationship between replication timing and patterns of human genetic variation was investigated by a former study [5], which examined how different types of genomic variation are distributed across the replication timing landscape. Their analysis revealed that single-nucleotide polymorphisms (SNPs) show a gradual increase in density toward later-replicating regions, whereas structural variations display more diverse relationships with replication timing depending on their underlying formation mechanisms. For example, certain classes of structural variants, including those associated with non-allelic homologous recombination (NAHR), showed preferential localization in early-replicating regions. The study suggested that differences in replication-associated processes, including DNA repair dynamics and chromatin context, may contribute to the observed variation patterns. These findings highlight replication timing as an important genomic feature associated with the distribution of human genetic variation.

A standardized approach for genome-wide replication timing profiling was established through the development of the early/late Repli-seq (e/l Repli-seq) protocol [6]. This method is based on labeling newly synthesized DNA in proliferating cells with the nucleotide analog bromodeoxyuridine (BrdU), followed by fluorescence-activated cell sorting (FACS) to separate cell populations representing early and late stages of S phase. After genomic DNA extraction and BrdU-based enrichment of nascent DNA, next-generation sequencing is used to generate genome-wide replication profiles. The resulting RT values are calculated by comparing normalized sequencing signals from early- and late-S-phase fractions, commonly expressed as a $\log_2(\text{ratio})$ across genomic windows. This experimental and computational framework enables genome-wide characterization and comparison of replication timing programs across different mammalian cell types.

Emerging computational approaches for studying replication timing at the single-cell level, such as single-cell Repli-seq, have been reviewed to address cell-to-cell heterogeneity in DNA replication programs [7]. Unlike population-based replication timing measurements that provide averaged profiles across many cells, single-cell approaches capture replication states within individual cells by analyzing changes in DNA copy number during S phase progression. These methods often use computational models, including probabilistic approaches such as hidden Markov models (HMMs), to infer whether genomic regions have been replicated or remain unreplicated in individual cells. By integrating data from many single cells at different stages of the cell cycle, these approaches enable the estimation of replication probability patterns and provide insights into the variability of replication timing programs across individual cells.

Advancing the quantitative analysis of DNA replication dynamics, single-cell 5-ethynyl-2-deoxyuridine sequencing (scEdU-seq) was developed to estimate replication fork speeds and replication progression in individual human cells [8]. The method is based on short EdU labeling of newly synthesized DNA fragments, followed by their enrichment and sequencing, with unique molecular identifiers (UMIs) used to reduce potential amplification biases. By analyzing the genomic distribution of EdU-labeled DNA segments through a dedicated computational framework, the study was able to estimate replication fork velocity at the single-cell level. Their results showed that replication speed varies during the cell cycle and increases during

progression through S phase, providing additional insights into the dynamic properties of DNA replication beyond conventional population-based timing measurements.

Beyond experimental measurements, computational modeling of RT provides a way to infer potential replication dynamics from genome-wide timing profiles. To investigate replication-related features and their relationship with genomic properties such as transcriptional activity and genome fragility, a modeling framework was applied to Repli-seq data [9]. By analyzing replication timing patterns across the genome, this approach enabled the estimation of replication parameters and the identification of regions with distinct replication behaviors. The study further suggested that local variations in replication dynamics may be associated with transcription-related processes and genomic instability. These findings highlight the potential of mathematical models for interpreting replication timing data and exploring the mechanisms underlying variation in genome organization and fragility.

To apply deep learning approaches to the analysis of genome-wide replication timing organization, a hybrid computational framework combining a deep neural network with a Hidden Markov Model (DNN-HMM) was developed for the de novo identification of replication timing domains [10]. The model was trained on human Repli-seq data and was designed to learn informative patterns directly from RT profiles, reducing the need for manually defined features. The deep learning component was used to extract relevant features from replication timing signals, while the HMM framework enabled the segmentation of the genome into distinct RT domains and transition regions. Compared with conventional approaches such as Gaussian mixture models, this method showed improved performance in identifying replication timing domain boundaries. Overall, the study provided a computational framework for characterizing the large-scale organization of replication timing patterns across the human genome.

A notable advance in sequence-based predictive modeling was the development of CONCERT [11], an interpretable deep learning framework designed to predict genome-wide replication timing profiles and identify sequence features associated with replication timing variation. Rather than relying solely on post-hoc interpretation methods, CONCERT incorporates interpretability into its architecture through two coordinated modules: a selector and a predictor. The selector identifies informative genomic regions, while the predictor uses bidirectional recurrent neural networks (RNNs) with self-attention mechanisms to learn sequence patterns related to replication timing from DNA sequence information. Applied to mouse embryonic stem cells and human cell lineages, CONCERT showed strong predictive performance and enabled the identification of candidate sequence elements associated with replication timing regulation. Further analyses suggested that these elements exhibit relationships with repetitive sequences, chromatin organization features, and replication-associated genomic regions, providing insights into the sequence determinants underlying replication timing variation.

To investigate the genomic features associated with replication timing organization, the relationship between differentiation-induced changes in replication timing and underlying sequence composition was examined [12]. Their study showed that early-replicating regions are generally associated with gene-dense, GC-rich isochores, whereas late-replicating regions are more frequently found in AT-rich regions containing higher densities of long interspersed nuclear elements (LINEs). The study further demonstrated that replication timing changes occurring during cellular differentiation are largely concentrated within AT-rich and LINE-rich genomic regions, while many GC-rich regions maintain relatively stable early-replication

patterns. These findings highlight the relationship between genome composition and the dynamic regulation of replication timing during cellular differentiation.

To characterize the large-scale organization of replication timing in mammalian genomes, cross-species comparative analysis was employed to investigate the conservation of early- and late-replicating domains [13]. Their study showed that megabase-scale replication timing domains are broadly conserved between mammalian species and are associated with genomic features such as gene density and transcriptional activity. Although evolutionary rearrangements have altered chromosome organization, the overall spatial distribution of early- and late-replicating regions remains substantially preserved. These findings suggest that replication timing represents a conserved aspect of genome organization and provide insights into the relationship between temporal replication programs and large-scale chromosomal architecture.

To investigate the evolutionary conservation of genome replication programs, replication timing profiles were compared between human and mouse cells [14]. Their study demonstrated that large-scale replication timing patterns are substantially conserved across mammalian species, suggesting that temporal regulation of DNA replication represents a stable feature of genome organization. In addition, conserved RT domains were associated with similar patterns of chromatin organization and long-range genomic interactions, highlighting a relationship between replication timing and higher-order genome architecture. These findings indicate that replication timing programs are not only preserved during evolution but also provide valuable information for understanding the relationship between genome replication and chromatin organization.

To investigate the relationship between replication timing and human cell differentiation, dynamic changes in replication timing profiles and gene expression patterns were analyzed during the lineage specification of human pluripotent stem cells (hPSCs) [15]. By examining multiple differentiation pathways, the study showed that replication timing programs are progressively remodeled during cellular differentiation and that these changes are associated with alterations in gene expression. Functional analyses indicated that regions with conserved early replication timing are often linked to broadly expressed cellular functions, whereas regions displaying lineage-specific timing changes are associated with genes involved in specialized cellular processes. These findings highlight the dynamic nature of replication timing during development and provide insights into how temporal genome organization relates to cell-type-specific regulatory programs.

The relationship between replication timing and somatic mutation patterns in cancer genomes was highlighted through an analysis of mutational heterogeneity across human cancers [16]. Their study showed that replication timing is a strong predictor of background somatic mutation frequency, with late-replicating genomic regions generally exhibiting higher mutation rates than early-replicating regions. These findings indicate that the temporal organization of DNA replication contributes to the uneven distribution of mutations across the genome. By incorporating replication timing and other genomic features into models of background mutation rates, this work provided an important framework for distinguishing potential cancer driver mutations from the broader landscape of passenger mutations.

To address the resolution limitations and experimental costs associated with traditional RT profiling techniques, such as Repli-seq, recent computational approaches have increasingly focused on predictive modeling. A notable example is Soffritto, a deep-learning framework developed to predict high-resolution RT profiles from widely available two-fraction RT data by integrating epigenomic features, including histone ChIP-seq signals, together with genomic characteristics such as GC content and gene density [17]. By reconstructing high-resolution replication timing profiles, Soffritto provides a computationally efficient and scalable alternative for studying genome-wide replication dynamics. This approach complements experimental RT profiling and enables more detailed characterization of replication timing patterns, thereby supporting subsequent analyses of genome organization and replication-associated biological processes.

Beyond experimental characterization of genome organization, DNA RT has emerged as a valuable predictive feature for studying chromatin interactions in cancer. A notable advancement in this area is RepliChrom, an interpretable machine learning framework developed to predict cancer-associated enhancer-promoter interactions (EPIs) using DNA replication timing profiles [18]. By integrating replication timing-derived features into its predictive framework, RepliChrom demonstrates that temporal replication information can substantially improve the identification of enhancer-promoter interactions. Moreover, the model's interpretable design enables the assessment of the relative contributions of individual replication timing-derived features to prediction performance, providing insights into the characteristics most closely associated with EPIs. As a computational approach, RepliChrom complements experimental methods and supports further investigation of the relationship between DNA replication timing and gene regulatory interactions in cancer.

Although the association between active transcription and early DNA replication timing (RT) has long been recognized, determining whether replication timing contributes to gene expression or simply reflects the underlying chromatin state has remained an important question. Addressing this issue, the study demonstrated that DNA replication timing can influence steady-state gene expression levels through a gene dosage effect associated with DNA replication [19]. Their findings show that genes replicated early during S phase generally exhibit higher mRNA abundance because they remain in a duplicated state for a longer portion of the cell cycle, whereas genes replicated later experience a shorter period with increased copy number and therefore tend to display lower expression levels. Rather than suggesting that replication timing is the sole determinant of transcription, this study identifies replication-dependent gene dosage as an important contributor to differences in gene expression. These findings provide a useful framework for understanding how replication timing can shape transcriptional output alongside other regulatory mechanisms.

Although much of the variability in DNA RT is shaped by tissue-specific epigenetic regulation, inherited genetic variation also contributes to differences in replication timing among individuals. Addressing this aspect, identified genetic variants associated with human DNA replication timing, demonstrating that replication timing is a heritable trait influenced by specific genomic loci [20]. Their study mapped replication timing quantitative trait loci (rtQTLs), showing that most of these variants act in cis and are associated with differences in the replication timing of nearby genomic regions. Many rtQTLs were also found to coincide with structural variants and other genomic features, highlighting the contribution of local genetic variation to the regulation of replication timing. By establishing a link between inherited genetic diversity and individual

replication timing profiles, this work provides a valuable framework for investigating how genetic variation influences the temporal organization of genome replication.

Moving beyond the analysis of individual genomic regions, the characterization of RT programs across different cell types has enabled a broader understanding of genome organization. A systematic framework based on cell-type-specific replication timing profiles, termed RT States, was developed to annotate the human genome according to its temporal replication patterns [21]. By integrating RT profiles from multiple cell types, this approach defines reproducible genomic categories that distinguish relatively conserved early- and late-replicating regions from cell-type-dependent changes in replication timing. These RT States show associations with genomic and epigenomic features, including chromatin characteristics, gene density, and functional annotations. Overall, this framework provides a resource for exploring how cell-type-specific regulatory programs relate to the temporal organization of genome replication.

To better understand the relationship between DNA RT and transcriptional regulation, approaches that enable the simultaneous measurement of both processes can provide valuable insights. A notable methodology addressing this need is PARTAGE [22], a parallel analysis framework developed to profile replication timing and gene expression within the same biological context. By integrating measurements of DNA replication dynamics and transcriptional output, this approach reduces potential limitations associated with comparing independently generated datasets. The resulting combined profiles enable the investigation of how replication timing patterns relate to changes in steady-state mRNA levels across different cellular conditions. Overall, PARTAGE provides a computational and experimental framework for examining the relationship between genome replication programs and transcriptional regulation, supporting further studies on how these processes are coordinated across diverse biological contexts.

The spatial distribution and strand-specific patterns of somatic mutations provide important insights into the molecular processes contributing to genome instability. By analyzing mutational strand asymmetries across cancer genomes, the study revealed associations between distinct mutation patterns and underlying mechanisms of DNA damage, replication, and repair [23]. Their study demonstrated that replication fork orientation, including differences between leading and lagging strand synthesis, contributes to the distribution of mutations associated with processes such as APOBEC activity and polymerase epsilon proofreading defects. In addition, strand-specific mutation patterns provided evidence for the influence of transcription-coupled nucleotide excision repair (TC-NER) on the distribution of DNA damage signatures between transcribed and untranscribed regions. Overall, these findings highlight how replication- and repair-related processes can shape mutational patterns in cancer genomes and provide a framework for investigating the mechanisms underlying genomic instability.

CHAPTER 3

METHODOLOGY

3.1 Research Design

This study was prepared to solve the mentioned problems using computational biology and bioinformatics methods. Specifically, this thesis aims to gather and analyze the necessary data, create a model, and evaluate the results obtained. Accordingly, the study described in this thesis has a quantitative design that utilizes computational biology, bioinformatics, and machine learning tools.

3.2 Methodology for Model

3.2.1 Data Collection

The COSMIC database [24] hosts some of the data used in the model building phase. This data is replication data belonging to SBS signatures. The signatures used for model training in this study are SBS1, SBS2, SBS3, SBS5, SBS13, SBS15, SBS18, and SBS40.

In determining the signatures used in model generation, care was taken to ensure that the relevant protein coding genes' had more than 100 nucleotides in those signatures. Therefore, replication timing was not calculated for each signature; only the signatures mentioned were included in the model except SBS15. Replication timing data for SBS signatures were obtained from the COSMIC database.

Other data was obtained using the SigProfilerTopography tool [1]. In this process, artificial VCF files were prepared for 20,117 protein coding genes individually, and the replication timing of the prepared VCF files were calculated using the SigProfilerTopography tool. In addition, the replication timing of overlapping nucleotides in the relevant SBS signatures of 20,117 protein coding genes were also used as data during model development. For this process as well, the SigProfilerTopography tool was used to calculate the replication timing of the created artificial VCF files.

To identify chromosomal loci where SBS signatures intersect with protein coding genes, SigProfilerTopography tool was executed using 552 VCF files. Based on this analysis, the tool generated chromosome-based annotations indicating the genomic positions of detected

mutations. These mutation coordinates were then compared with the start and end positions of protein coding genes to determine regions of overlap. Subsequently, artificial VCF files were constructed using these intersecting loci, and the SigProfilerTopography tool was applied again to calculate the replication timing values for regions where protein coding genes and SBS signatures overlap.

3.2.2 VCF Files

VCF files corresponding to each protein coding gene and SBS-overlapping loci were generated using the hg19 human reference genome. The nucleotide identities within the regions defined by the start and end coordinates of protein coding genes were determined by consulting the hg19 fasta reference sequences.

Each VCF file consists of four columns. The first column provides the genomic position as a numerical value. The second column contains the name of the associated gene. The third column specifies the reference nucleotide present at the given position (A, C, G, or T). The fourth column lists an alternative nucleotide, selected at random based on the reference base in the third column. For instance, if the reference nucleotide is A, the alternative nucleotide is randomly chosen from C, G, or T. This random selection process was applied consistently across all entries.

Artificial VCF files were also generated for nucleotides within protein coding genes that overlap with SBS signatures, again using the hg19 human reference genome. During this process, care was taken to ensure that each protein coding gene included at least 100 nucleotides overlapping with a given SBS signature. This threshold was imposed to obtain statistically meaningful results in downstream replication timing analyses.

For example, if only 10 overlapping nucleotides were identified and these were evenly distributed across replication timing bins, it would not be possible to draw reliable conclusions about the relationship between the data and replication timing. Therefore, a minimum threshold of 100 overlapping nucleotides was established to enhance the robustness and interpretability of the analysis. These VCF files were created using the Python programming language.[25]

DNA replication timing is calculated based on the duration of the stages that occur during DNA division. The equation below provides a clearer representation of how this time is determined.

3.2.3 SigProfilerTopography

DNA replication timing is calculated based on the duration of the stages that occur during DNA division. The equation (1) provides a clearer representation of how this time is determined.

$$(0.917\,G_{1b}) + (0.750\,S_1) + (0.583\,S_2) + (0.417\,S_3) + (0.250\,S_4) + (0 \cdot G_2) \qquad (1)$$

According to the equation [26], the stages that occur earlier during DNA division are assigned higher coefficients, whereas the stages that take place later are associated with lower coefficients.

This weighting scheme is, in fact, used to infer whether the corresponding DNA replication timing reflects an early or a late replication phase.

The SigProfilerTopography tool partitions the calculated DNA replication timing into 10 equal bins. Each gene is then assigned to a corresponding bin, enabling the classification of genes as either early- or late-replicating.

3.2.4 Methods

Using artificially generated VCF files, the replication timing of nucleotides overlapping between protein coding gene regions and SBS signature loci was determined.

The replication timing values calculated for the protein coding genes were used as input features to the model, whereas those derived for the overlapped mutation set were treated as the model outputs. In addition, the replication timing metrics obtained for protein coding genes did not constitute the sole input provided to the model.

An additional input to the model was the replication timing associated with the relevant SBS signatures [27]. These data were obtained from the COSMIC database. Using the aggregated information, a structured dataframe was constructed for model training. The first column contained gene identifiers, the next ten columns corresponded to the replication timing values for the protein coding genes, columns 12 through 21 captured the replication timing of the SBS signatures, and columns 22 through 31 represented the replication timing values for the overlapped mutations.

The constructed dataset consists of input and output variables. The model input is formed by concatenating the replication timing data of a gene and its corresponding SBS signature. Both components are 1-dimensional vectors with 10 elements, combined end-to-end to create the final input. The output represents the replication timing of SBS mutations overlapping with the relevant gene, which is likewise a 1-dimensional vectors with 10 elements. The resulting dataframe consists of 3584 observations. Notably, some genes appear multiple times within the dataset. This occurs because a single gene can overlap with more than one SBS signature, with overlaps exceeding 100 nucleotides. For rows containing the same gene, the replication timing values of the protein coding gene variables remain consistent, whereas the replication timing associated with SBS signatures and the overlapped mutation variables varies. Figure 2 provides a general overview of the inputs and outputs.

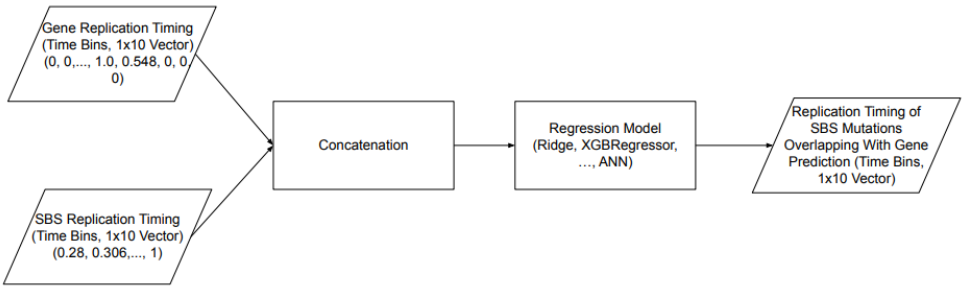

Figure 2: Schematic for Replication Timing Prediction Model

As illustrated in Figure 2, the gene replication timing data and the SBS signature replication timing data are concatenated into a single vector, which serves as the input to the model. The model is subsequently trained to predict the replication timing of SBS mutations overlapping with the gene.

To enhance the reliability of the results and to evaluate model performance, model training was conducted using 10-fold cross-validation. Subsequently, the models were evaluated on an independent test set derived from SBS15 data, and the reported R^2 and MSE values were obtained through this testing procedure. Training was performed using all available SBS signatures except for SBS15.

Since the model to be produced within the scope of this study is the replication timing of SBS mutations overlapping with the gene, the problem is seen as a regression model. The models evaluated within the scope of this study are Ridge Regression, XGBRegressor, Random Forest Regression, Bayesian Ridge Regression, and Artificial Neural Network (ANN).

In this study, a diverse set of regression algorithms was employed to ensure a comprehensive evaluation of model performance under different assumptions and learning paradigms. Linear models such as Ridge Regression and Bayesian Ridge Regression were selected as foundational benchmarks due to their robustness against multicollinearity and their ability to prevent overfitting through regularization [28] [29]. While Ridge Regression applies an L_2 penalty to stabilize coefficient estimates, Bayesian Ridge Regression incorporates probabilistic inference, providing uncertainty estimates alongside point predictions, which aids in assessing the reliability of the replication timing predictions.

To capture the non-linear relationships and complex interactions inherent in genomic datasets, ensemble learning techniques were utilized. Random Forest Regression [30] was selected for its capacity to handle high-dimensional data through bagging (bootstrap aggregating). By constructing an ensemble of decision trees, it reduces variance and improves generalizability, making it particularly effective at identifying stable patterns without overfitting to noisy features. Complementing this, XGBRegressor [31] was implemented to leverage gradient-based optimization. Unlike the bagging approach, this boosting methodology sequentially minimizes residuals, iteratively refining the model's predictive accuracy by focusing on patterns that were poorly captured by previous iterations.

Finally, ANN [32] were incorporated to capture high-dimensional, non-linear mappings between input and output features. By utilizing deep architectures with multiple hidden layers and non-linear activation functions, the ANN is uniquely suited to identify hierarchical features and latent dependencies within biological data structures that traditional linear or ensemble models might overlook. Collectively, this comparative framework allows for a multi-faceted analysis, balancing the interpretability of linear models with the predictive power of advanced machine learning architectures.

Since the precise linear or non-linear nature of the constructed dataset remained ambiguous, all four previously mentioned algorithms were evaluated during model development. While the upward trend inherent to all training SBS signatures suggests a potential linear relationship, the presence of multiple SBS signatures for a single protein coding gene within the dataset could disrupt this linearity. Consequently, both linear and non-linear models were assessed. Additionally, an ANN was trained to account for potential complex interactions. Compared to standard linear and non-linear models, ANNs excel at identifying intricate relationships and generally perform optimally with larger datasets. Although our dataset size is somewhat constrained for such architectures, the ANN was implemented specifically to ensure any complex underlying patterns were captured.

By integrating these diverse approaches, the study aims to compare linear, ensemble, and deep learning-based methods, thereby providing a more reliable and well-rounded assessment of predictive performance.

3.2.4.1 Ridge Regression

To model the linear relationship between the predictors and the target variable while controlling for model complexity, we employed Ridge Regression. Standard ordinary least squares estimation can become unstable and exhibit high variance, particularly in the presence of multicollinearity or when the feature space is relatively large. Ridge regression addresses this limitation by adding a continuous shrinkage penalty proportional to the square of the L_2-norm of the coefficient vector. Equation (2) presents the mathematical formulation of Ridge Regression.

$$\hat{\beta} = \arg\min_{\beta} \left(\|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 \right) \tag{2}$$

Specifically, the regression coefficients $\hat{\beta}$ are estimated by minimizing the penalized residual sum of squares. Where y represents the response vector, X denotes the design matrix of input predictors, β is the vector of coefficients, and $\lambda \geq 0$ is the tuning parameter that governs the strength of the regularization.

The inclusion of the L_2 penalty continuously shrinks the coefficient estimates toward zero, achieving a systematic trade-off that substantially reduces model variance at the cost of introducing a slight bias. Because the regularization term penalizes the magnitude of the coefficients uniformly, predictors with larger numerical scales would be penalized disproportionately if left

unscaled. Therefore, all continuous input variables were standardized to have a mean of zero and unit variance prior to fitting the model.

3.2.4.2 Random Forest Regressor

To capture non-linear patterns and complex interactions among the predictors without imposing strict parametric assumptions, we implemented a Random Forest Regressor. While individual decision trees are inherently prone to high variance and overfitting, the random forest algorithm mitigates this limitation by constructing an ensemble of uncorrelated trees through a combination of bootstrap aggregating (bagging) and random feature selection. Equation (3) presents the mathematical formulation of Random Forest Regressor.

$$\hat{f}_{\text{rf}}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_b(x) \tag{3}$$

During training, the algorithm draws B independent bootstrap samples from the original dataset and fits a decision tree to each sample. To ensure that the individual trees remain decorrelated, at each candidate node split, the algorithm evaluates only a random subset of the available predictors rather than the entire feature space. For a given input vector x, the final ensemble prediction is obtained by averaging the outputs of all individual trees.

Where $\hat{f}_b(x)$ denotes the prediction of the b-th regression tree, and B represents the total number of trees in the ensemble. By averaging across multiple diverse trees, the overall prediction variance is significantly reduced while maintaining a low bias. Furthermore, because decision tree splits are based on relative feature rankings and are invariant to monotonic transformations, continuous predictor variables do not require standardization or scaling prior to model fitting.

3.2.4.3 XGBRegressor

As an alternative tree-based methodology, we utilized eXtreme Gradient Boosting (XGBoost), which shifts the ensemble strategy from parallel bagging to sequential boosting. Rather than averaging independently grown trees, the XGBoost algorithm builds decision trees iteratively; each new tree is specifically trained to predict and correct the residual errors of the existing ensemble, progressively refining the model's predictions.

To guide this sequential learning process without fitting to noise, the algorithm optimizes a regularized objective function at each iteration t. This formulation explicitly balances predictive accuracy against structural complexity by integrating a regularization term directly into the loss calculation. Equation (4) presents the mathematical formulation of XGBRegressor.

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{4}$$

Where l represents a differentiable convex loss function measuring the discrepancy between the true target y_i and the prediction $\hat{y}_i^{(t-1)}$ updated by the newly added tree $f_t(x_i)$. The second term acts as a complexity penalty, where T denotes the total number of leaves in tree f_t, γ is the minimum loss reduction required to make a further partition, and λ applies an L_2-norm regularization on the leaf weights w.

By employing a second-order Taylor expansion to approximate this objective function, the algorithm determines optimal split points and leaf weights with high efficiency. Because the underlying architecture relies on decision tree node splitting rather than distance metrics, continuous input features remain invariant to monotonic transformations, negating the need for prior standardization.

3.2.4.4 Artifical Neural Network

To account for complex and non-linear interactions between the variables, we also implemented an ANN. This model processes input data through layers of interconnected nodes, where each node applies a weighted linear combination followed by a non-linear activation function to map inputs to the target variable. Equation (5) presents the mathematical formulation of Artificial Neural Network.

For a network configuration with a single hidden layer, the forward propagation that computes the predicted output $\hat{y}$ from an input vector x is expressed as:

$$\hat{y} = W^{(2)}\sigma\left(W^{(1)}x + b^{(1)}\right) + b^{(2)} \tag{5}$$

where $W^{(1)}$ and $b^{(1)}$ represent the weight matrix and bias vector for the hidden layer, $\sigma(\cdot)$ is the activation function (such as the Rectified Linear Unit), and $W^{(2)}$ and $b^{(2)}$ denote the weights and biases for the final output layer.

The network learns by iteratively adjusting these weights via backpropagation, using a gradient-based optimization algorithm (such as the Adam optimizer) to minimize the prediction error. Because gradient descent algorithms are highly sensitive to the scale of the input values, all continuous features were standardized to have a mean of zero and unit variance before the training process began.

All of the models were given gene replication timing of gene and the corresponding SBS signature as input. The output was the replication timing of SBS mutations overlapping with the relevant gene.

Each model was trained using the remaining SBS signatures and their associated protein coding genes, with the exception of SBS15. Genes containing more than 100 mutations within SBS15 were reserved to evaluate these trained models, resulting in a final test set of 10 genes. This 10-sample test set was kept entirely independent and was strictly excluded from the training phase. SBS15 was selected as the test set primarily due to its limited number of genes. Since the other SBS signatures contain substantially more data, isolating SBS15 for testing mitigates

potential class imbalance during model training. Finally, cross-validation was performed exclusively on the training set composed of the remaining signatures.

Since a vector needed to be obtained as a result of the problem, a Multi Output Regressor was used and the replication timing in each bin was estimated. For each algorithm, R^2 score and mean square error (MSE) were used as metrics [33] [34]. All tests were conducted using the Python programming language.

3.2.5 Parameters

Within the scope of this study, five different models were trained and evaluated. The Ridge Regression model was trained using an alpha value of 1. The R^2 and MSE scores presented in this study were obtained with this alpha configuration. Initially, the model was trained with an alpha value of 0.1; however, due to inferior performance, the alpha parameter was subsequently adjusted to 1.

For the XGBoost Regressor model, the number of estimators was set to 100 during training, while all remaining hyperparameters were retained at their default values. The Random Forest algorithm was trained using the default parameter settings. Under the default configuration, the number of estimators is defined as 100.

Similarly, the Bayesian Ridge algorithm was trained using the default hyperparameter configuration. The ANN architecture consists of two hidden layers. The first hidden layer contains 20 neurons, while the second hidden layer consists of 32 neurons. Since the target data to be estimated is represented as a 10-element matrix, the network generates 10 output values. A ReLU activation function [35] was applied after the first hidden layer, whereas a sigmoid activation function [32] was employed after the second hidden layer. The sigmoid function was selected because the target values are constrained within the range of 0 to 1.

During the training process, the Adam optimizer [36] was utilized with a learning rate of 0.001. As the task addressed in this study is a regression problem, MSELoss [32] was adopted as the loss criterion. The model was trained for 500 epochs.

3.2.6 Metrics

In regression analysis, evaluation metrics play a critical role in assessing how well a model captures the underlying relationship between input variables and the target outcome. Among the most widely used metrics are MSE and the coefficient of determination (R^2), each offering distinct yet complementary insights into model performance.

MSE measures the average of the squared differences between the observed values and the predicted values. This metric is particularly useful because it penalizes larger errors more heavily than smaller ones due to the squaring operation. As a result, MSE is sensitive to outliers and provides a clear indication of how far predictions deviate from actual values in terms of squared magnitude. Its mathematical properties also make it convenient for optimization, as

it is differentiable and thus well-suited for many machine learning algorithms that rely on gradient-based methods.

On the other hand, R^2 evaluates how well the model explains the variability of the dependent variable. It represents the proportion of variance in the observed data that is accounted for by the model. An R^2 value closer to 1 indicates that the model explains a large portion of the variance, while a value near 0 suggests limited explanatory power. Unlike MSE, R^2 is a relative measure, making it especially useful for comparing different models or understanding their explanatory strength in a standardized way.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{6}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{7}$$

The combined use of MSE and R^2 provides a more comprehensive evaluation of a regression model. While MSE focuses on the magnitude of prediction errors, R^2 offers insight into the model's ability to capture data variability. Using both metrics together allows researchers to balance error minimization with explanatory capability, leading to more robust and interpretable model assessment.

3.2.7 Correlation Matrix

A correlation matrix [37] is employed in research to systematically examine the strength and direction of linear relationships among multiple variables within a dataset. By presenting pairwise correlation coefficients in a structured format, it enables researchers to quickly identify patterns of association, such as positive, negative, or negligible linear relationships. This is particularly useful in the early stages of data analysis, as it helps detect potential pairwise multicollinearity issues, guides preliminary variable selection, and informs the initial development of statistical models. Moreover, while it primarily captures linear interdependencies, a correlation matrix provides a concise summary of the data structure, allowing for clearer interpretation of complex datasets without requiring extensive individual comparisons.

3.2.8 Feature Importance

Feature importance in a Random Forest regressor [30] is used to quantify the relative contribution of each input variable in predicting the target outcome. This is achieved by evaluating how much each feature reduces prediction error (measured by mean squared error) across all decision trees in the ensemble. Features that consistently lead to larger reductions in error when used for splitting are considered more important. This approach allows researchers to identify the most influential predictors, improve model interpretability, and potentially reduce dimensionality by removing less relevant variables. Additionally, feature importance derived from Random Forest

models is particularly valuable because it captures non-linear relationships and interactions between variables, providing a more robust understanding of the underlying data structure compared to simpler linear methods.

3.3 Methodology for GO Terms

3.3.1 Data Collection For GO Term Analysis

In the analysis performed from the perspective of GO Terms, the replication timing calculated for 20,117 protein coding genes were used as data. The replication timing of the mentioned protein coding genes was calculated for 10 different cell lines. The list of relevant cell lines is shown in Table 1.

Table 1: Cell Lines With Tissues

Cell Lines	**Tissues**	**Organsim/Cell Type**
MCF7	Breast Tissue	Human, Mammary
HEPG2	Liver Tissue	Human, Liver Cells (Hepatocytes)
HELAS3	Cervix Tissue	Human, Epithelial-Like Cervical Cells
SKNSH	Brain Tissue	Human, Neuronal-Like Cells
K562	Bone Marrow Tissue	Human, Lymphoblast Cells
IMR90	Lung Tissue	Human, Fibroblast
NHEK	Skin Tissue	Human, Keratinocyte
BJ	Skin Tissue	Human, Fibroblast
BG02ES	Early Development Stage of an Embryo Tissue	Human, Embyronic Stem Cell
GM12878	Blood Tissue	Human, B-Lymphocyte

3.3.2 Creating DataFrame

The replication timing profiles of the artificially generated VCF files were computed using the SigProfilerTopography tool, after which the resulting data were transformed into a structured dataframe. During this transformation process, each replication timing value was incorporated into the dataframe as part of a ten-dimensional matrix representation. The final dataframe comprises gene identifiers, replication timing bins, and scores.

3.3.3 Score

A unique score was created for the calculated replication timing data. This score allows us to see whether the gene replicates early or late in the relevant cell line.

Appropriate scores were generated from the obtained replication timing data, based on whether the genes replicated early or late.

The replication timing is obtained as a vector with 10 elements. The obtained values are normalized between 0 and 1. These values are multiplied by their current position and then divided by the sum of vectors. With this calculation, a unique score is obtained between 0.1 and 1. Accordingly, scores close to the 0.1 value can be interpreted as late replicate, while scores close to the 1 value can be interpreted as early replicate. To assign higher scores to early-replicating genes and lower scores to late-replicating genes, the input vector was first reversed. The values were then weighted using coefficients ranging from 1 to 10 according to their positions, and the weighted sum was normalized by dividing it by ten times the sum of all values in the vector. This procedure yielded a normalized replication-timing score ranging from 0.1 to 1 for each gene.

3.3.4 Early and Late Genes

The scores derived from replication timing data provide insight into whether genes replicate during early or late phases of the cell cycle. These scores range from 0.1 to 1.0, where a value of 0.1 (the lowest value) indicates late replication, while a value of 1.0 (the highest value) corresponds to early replication timing.

A predefined threshold was applied to the genes selected for analysis. Specifically, genes were classified as late-replicating if their scores were 0.2 or lower, and as early-replicating if their scores were 0.9 or higher. Throughout the study, GO terms associated with late-replicating genes were identified based on genes with scores at or below 0.2. Similarly, GO terms linked to early-replicating genes were determined using genes with scores equal to or greater than 0.9.

The quantitative distribution of early and late replicating genes varies across the examined cell lines. In the BG02ES cell line, 5,416 genes were identified as early-replicating, while 518 were classified as late-replicating. The BJ cell line exhibited 4,737 early and 436 late replicating genes, whereas GM12878 showed 5,781 early and 336 late replicating genes. The HELAS3 cell line comprised 5,965 early and 542 late replicating genes, and HEPG2 contained 6,096 early

and 309 late replicating genes. Analysis of IMR90 revealed 5,088 early and 444 late replicating genes. Furthermore, the K562 cell line displayed 6,521 early and 373 late replicating genes, compared to MCF7 with 6,522 early and 620 late replicating genes. Finally, the NHEK cell line featured 5,946 early and 351 late replicating genes, while SKNSH presented 6,340 early and 656 late replicating genes. The relevant data, along with the number of genes discarded during the study, are presented in Table 2.

Table 2: Early and Late Replicating Gene Counts

Cell Lines	**Early Replicating Genes**	**Late Replicating Genes**	**Discarded Genes**
BG02ES	5,416	518	14,183
BJ	4,737	436	14,944
GM12878	5,781	336	14,000
HELAS3	5,965	542	13,610
HEPG2	6,096	309	13,712
IMR90	5,088	444	14,585
K562	6,521	373	13,223
MCF7	6,522	620	12,975
NHEK	5,946	351	13,820
SKNSH	6,340	656	13,121

For each cell line, the pathways enriched by the protein coding genes were calculated. Subsequently, these pathways were analyzed to identify both those shared across the cell lines and those that are cell-line-specific. This approach was adopted because only 19 genes were identified as late-replicating in each cell line, compared to 5,057 early-replicating genes. Consequently, the limited number of late-replicating genes would have significantly constrained the pathway enrichment analysis.

The selection of relatively stringent threshold values was motivated by the disproportionately high number of early-replicating genes. Furthermore, restricting the thresholds to these ranges enabled a more accurate and robust analysis of the GO terms associated with both early- and late-replicating genes.

After identifying the early- and late-replicating genes, the GSEApy library [38] implemented in the Python programming language was utilized to investigate the pathways associated with these genes. Using this library, GO terms [39] related to the identified genes were analyzed

across the categories of biological process, cellular component, and molecular function. For all three categories, the 2025 versions of the corresponding databases were employed.

To classify the pathways identified using the GSEApy library and determine statistically significant pathways, the adjusted p-value threshold was set to below 0.05. The library provides both raw p-values and adjusted p-values for pathway enrichment analyses. In this study, pathway classification was performed based on the adjusted p-value in order to more clearly identify the common biological themes associated with the pathways influenced by the genes.

A new dataframe was generated by categorizing the pathways according to their adjusted p-values and annotating them with their corresponding pathway categories. This dataframe includes the names of pathways with adjusted p-values below 0.05, the cell lines in which these pathways were observed, and the categories to which they belong. If a pathway was associated with multiple categories, all ancestor categories related to that pathway were included in the annotation. The purpose of this approach was to narrow the scope of the analysis and to investigate, from a categorical perspective, whether the pathways influenced by early- and late-replicating genes exhibit common biological themes.

3.4 Workflow

Figure 3 illustrates the comprehensive workflow diagram of this study. Because the research encompasses two distinct components, it yields two separate outputs. Accordingly, the chart delineates both the predictive modeling pipeline and the subsequent Gene Ontology term analysis.

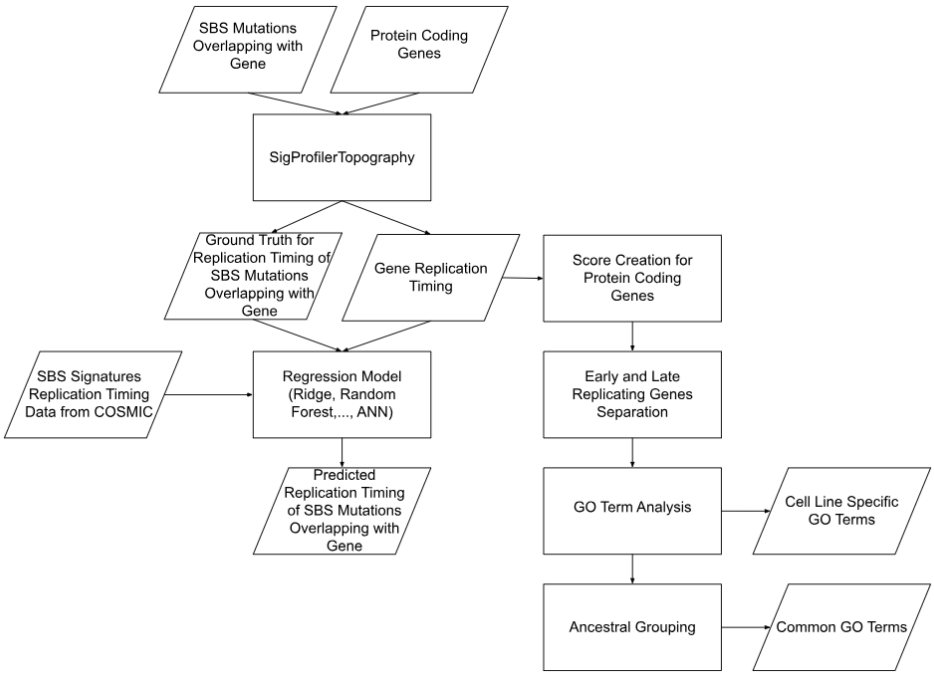

Figure 3: Workflow of Study

As depicted in Figure 3, we utilized SigProfilerTopography to calculate replication timing for protein coding genes, and SBS mutations overlapping with the genes. Following integration with the SBS signature replication timing profiles, this dataset was used to train the models, which aim to predict the replication timing of SBS mutations overlapping the with genes.

In parallel, a replication timing score was calculated for protein coding genes using score creation method. Based on this score, genes were classified into early- and late-replicating groups, followed by GO enrichment analysis. The enriched GO terms were subsequently organized according to their ancestral relationships, allowing the identification of shared functional categories. Furthermore, following the GO enrichment analysis, these gene sets were also examined in a cell line–specific manner to evaluate lineage-dependent functional patterns.

CHAPTER 4

RESULTS

4.1 Results For Model

This study aimed to train a regression-based model. Accordingly, the models used were evaluated based on the MSE and R^2 metrics. Each of the results was obtained after performing k=10 cross-validation.

According to the R^2 metric, the model with the highest score is the Random Forest model. Similarly, based on the Mean Squared Error (MSE) metric, the Random Forest model achieved the lowest MSE value. However, the R^2 and MSE scores obtained by the other models are not substantially different from those of the Random Forest model. Overall, the Random Forest model demonstrates slightly better performance compared to the other models. As observed, the models exhibit very high R^2 scores and very low MSE values. Based on these results, it can be inferred that the models have either achieved excellent learning performance or may be overfitting the training data. Therefore, the trained models were further evaluated using a separate test set. The R^2 and MSE metrics were again employed for performance assessment. The R^2 and MSE scores of all models are presented in Table 3.

Table 3: R^2 and MSE Scores For Algorithms

Algorithms	**Test R^2 Score**	**Test MSE**	**CV R^2 Score**	**CV MSE**
Ridge Regression	0.843	0.004	0.968	0.003
XGBRegressor	0.809	0.006	0.973	0.002
Random Forest Regressor	0.935	0.004	0.976	0.002
Bayesian Ridge	0.833	0.004	0.968	0.003

Accordingly, when the test and cross-validation scores are compared, it can be concluded that the Random Forest model does not exhibit overfitting. This is because the cross-validation and test scores are quite similar, indicating that the model does not merely memorize the training data but instead demonstrates genuine predictive capability.

In addition to these algorithms, a model was also developed using an Artificial Neural Network (ANN). Initially, a relatively more complex ANN architecture consisting of three layers was implemented; however, the model was observed to overfit the training data. Consequently, the ANN architecture was restructured into a two-layer configuration. The loss graph of the revised ANN model is presented in Figure 4.

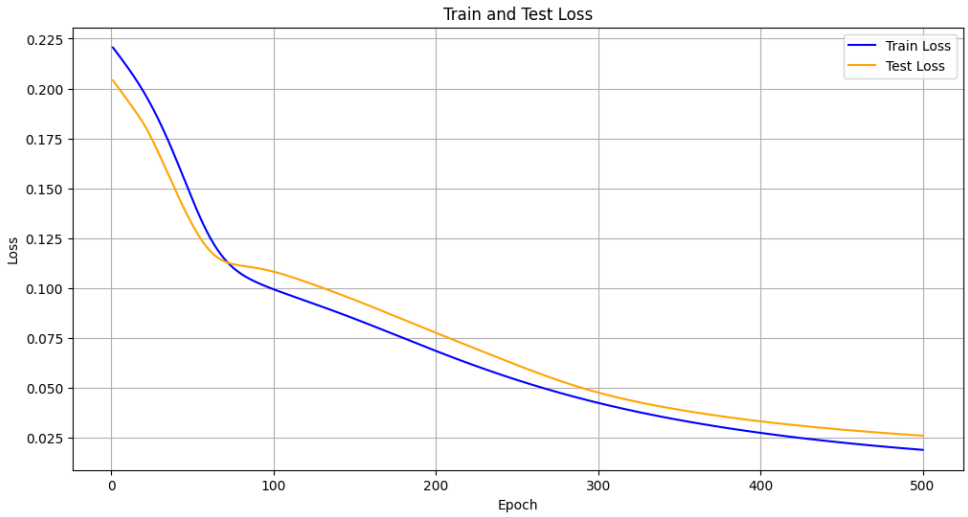

Figure 4: ANN Loss Graph

The model does not appear to be overfitted after 500 epochs training and actually has low loss values.

In light of the obtained results, and considering that the models were trained on a relatively limited dataset, the performance metrics appear almost too good to be true. Therefore, the predictions of the best-performing models—namely Random Forest, ANN, and Ridge Regression—were visualized in order to further analyze how the models perform in practice and to better understand their actual predictive behavior.

4.1.1 Model Predictions

Among the evaluated models, the three with the highest overall performance were selected to further assess their predictive capability on the dataset. For this purpose, a representative input feature protein coding gene and SBS signature were chosen. Specifically, the selected gene was PTPRD, and the signature was SBS15.

As illustrated in Figure 5, the model received the first two graphs as input. The third graph represents the ground truth values for the target variable. The final three graphs depict the predictions generated by the three best-performing models. These visual comparisons enable a direct evaluation of how closely each model's predictions align with the ground truth, thereby providing a qualitative assessment of their predictive accuracy.

Figure 5 illustrates the model's prediction for the PTPRD gene and the SBS15 signature.

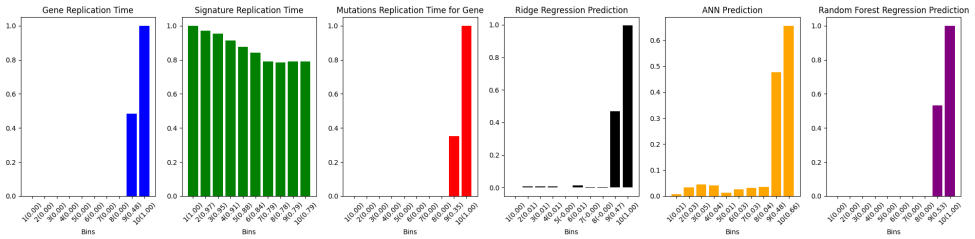

Figure 5: PTPRD with SBS15 Prediction

The ground truth and the gene replication time provided to the model as input exhibit a highly similar pattern. Due to this strong resemblance, it remains unclear whether the model is genuinely learning and predicting the underlying relationship or merely replicating the input signal. Consequently, the observed predictive performance may partly reflect the structural similarity between the input feature and the target variable rather than the model's true generalization capability.

Considering the predictions, the Random Forest model demonstrates the most accurate performance. This is evident because, similar to the ground truth, nonzero values appear only in the last two bins. In contrast, the Artificial Neural Network (ANN) produced values across every bin, many of which deviate substantially from the ground truth. The Ridge Regression model also predicted values in other bins, though to a lesser extent than the ANN. Based on these observations, the Random Forest model appears to be the most suitable choice for practical application.

4.1.2 Random Forest Feature Importance

Among the evaluated models, the Random Forest model produced comparatively the closest predictions to the actual values. Therefore, in order to determine whether the model was genuinely learning underlying patterns rather than producing coincidental results, the feature importance scores of the model were examined.

Accordingly, while the highest feature importance value associated with the gene replication bins reached 0.10, all SBS replication values remained below 0.005. Based on these results, it was concluded that the model did not exhibit overfitting. Rather, the findings indicate that the SBS replication variable did not have a meaningful influence on the prediction process.

The model assigned minimal weight to the SBS replication time during the prediction process and instead appeared to replicate the gene replication time pattern. Consequently, the model was not performing a genuine predictive task in a meaningful sense. This outcome indicates that the model is not suitable for practical application.

4.1.3 Correlation Matrix

A correlation matrix was constructed to analyze the relationships among the variables used in the model training process. The primary objective of this analysis was to determine whether the model's performance limitations stem from the model itself or whether the underlying data genuinely exhibit trends consistent with the model's predictions. By examining the correlation structure of the dataset, it becomes possible to assess the strength and direction of linear relationships between variables and to evaluate whether the observed predictive behavior is supported by the intrinsic characteristics of the data. The correlation matrix of the dataset is presented in Figure 6.

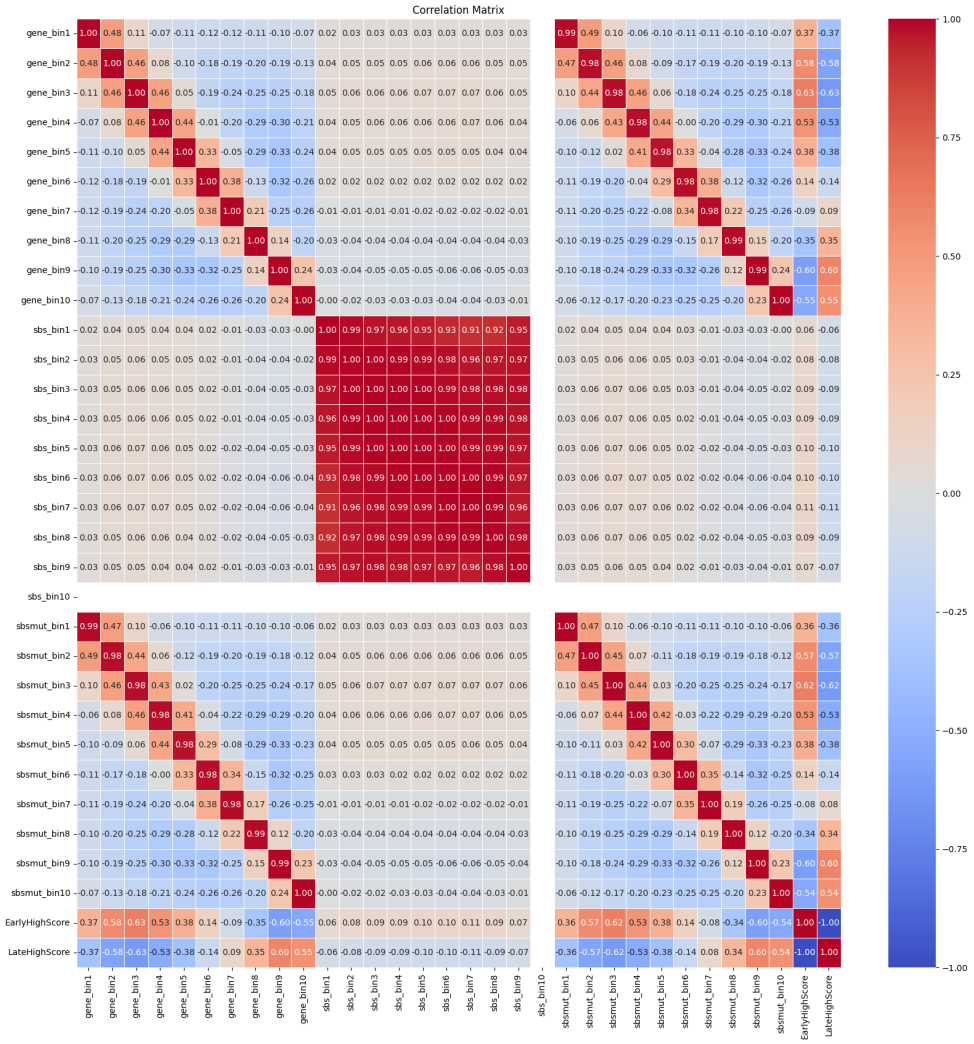

Figure 6: Correlation Matrix

According to the correlation matrix, the SBS bins exhibit a high degree of correlation among themselves, indicating strong internal linear relationships. In contrast, both the replication time of genes and the replication times of mutations overlapping within SBS show very low

correlation with the replication times of SBS signatures. However, the replication times of genes display a relatively strong correlation with the replication times of the overlapping mutations within SBS.

Based on these findings, it can be inferred that the model's limited predictive performance does not arise from structural inadequacy of the model itself. Rather, the results suggest that the replication times of SBS do not exert a sufficiently strong influence on the replication times of the overlapping mutations. Therefore, the weak predictive behavior appears to be primarily attributable to the intrinsic characteristics and relational structure of the data, rather than to deficiencies in the modeling approach.

4.2 Results for GO Term Analysis for Common Pathways

The Gene Ontology term analysis results were examined under three main categories: biological process, cellular component, and molecular function. The analysis was initially conducted to identify commonly enriched pathways shared across all experimental groups. Subsequently, a more detailed evaluation was performed to determine cell line–specific pathways, allowing for the identification of distinct functional patterns associated with each cell line. The number displayed in the legend of the figures represents the total number of pathways, whereas the number shown in the figures indicates how many pathways are categorized under the number of enriched ancestor pathway.

Figure 7, 8, and 9 present the number of pathways identified for biological process, cellular component, and molecular function, respectively. The y-axis in each figure represents the ancestral ontology terms to which the respective enriched pathways are mapped. Since a given pathway may be linked to multiple ancestors, the total number of unique enriched pathways may be lower than the sum of all enriched ancestors

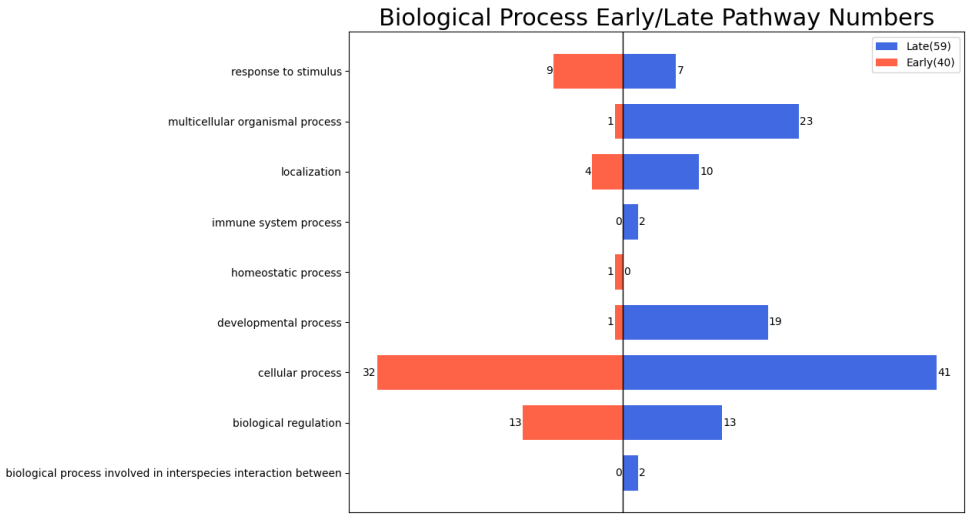

Figure 7: Ancestor-Based GO Term Distribution for Biological Process

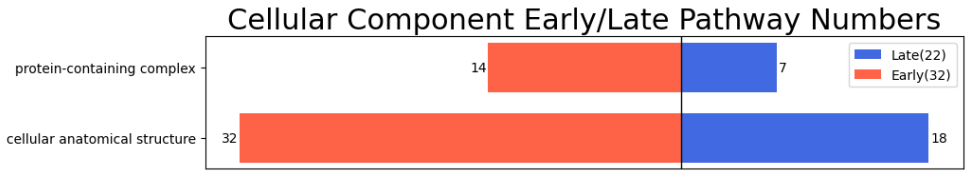

Figure 8: Ancestor-Based GO Term Distribution for Cellular Component

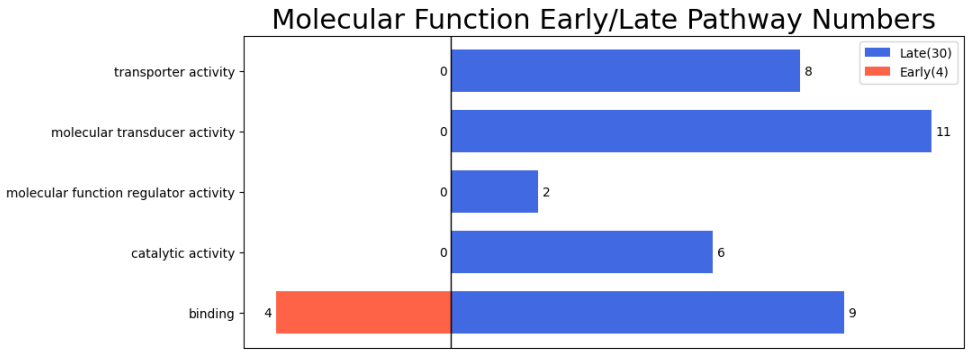

Figure 9: Ancestor-Based GO Term Distribution for Molecular Function

Figures 10, 11, and 12 present all pathways identified across the biological process, cellular component, and molecular function categories.

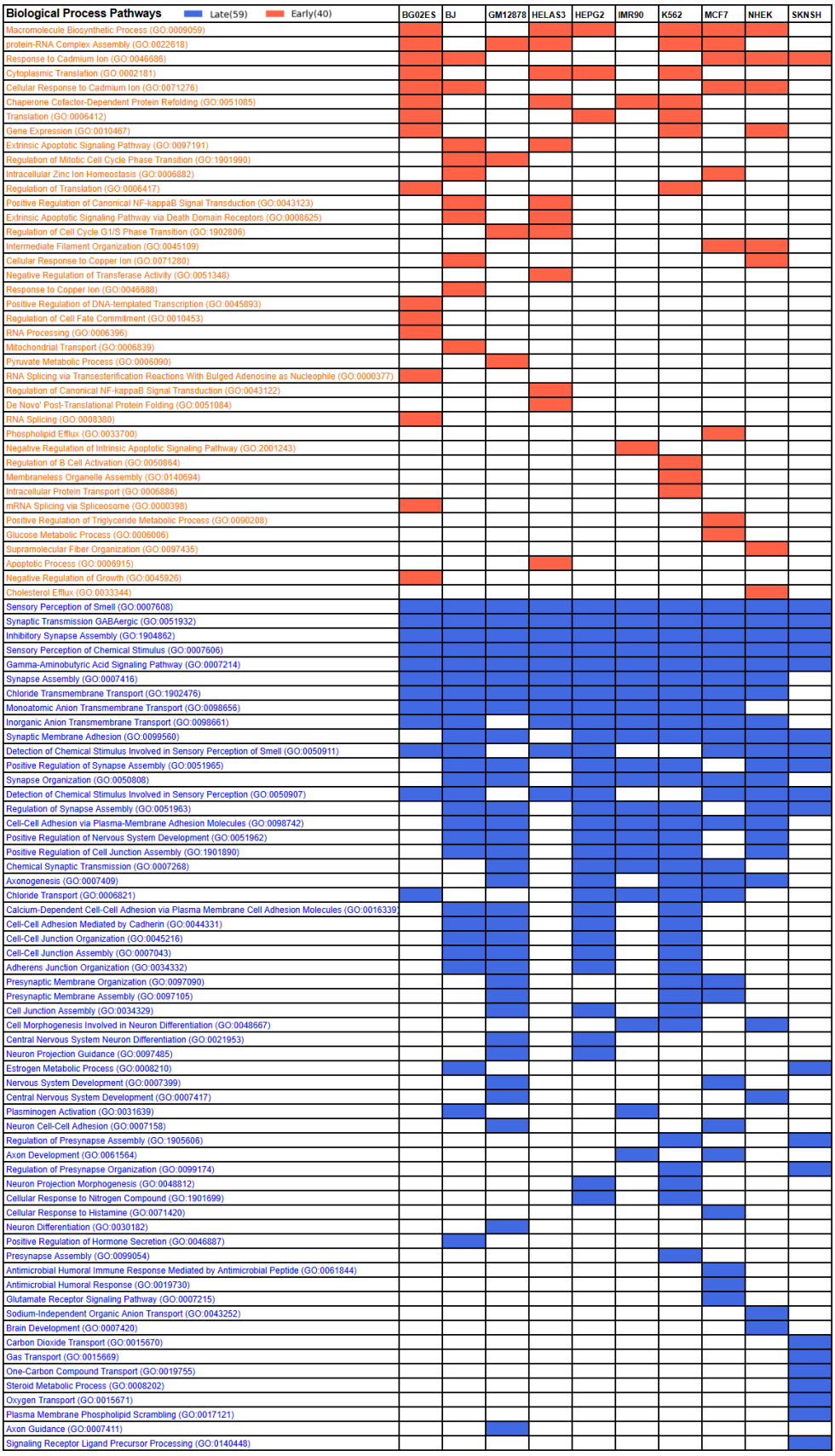

Figure 10: Ancestor-Based GO Terms for Biological Process

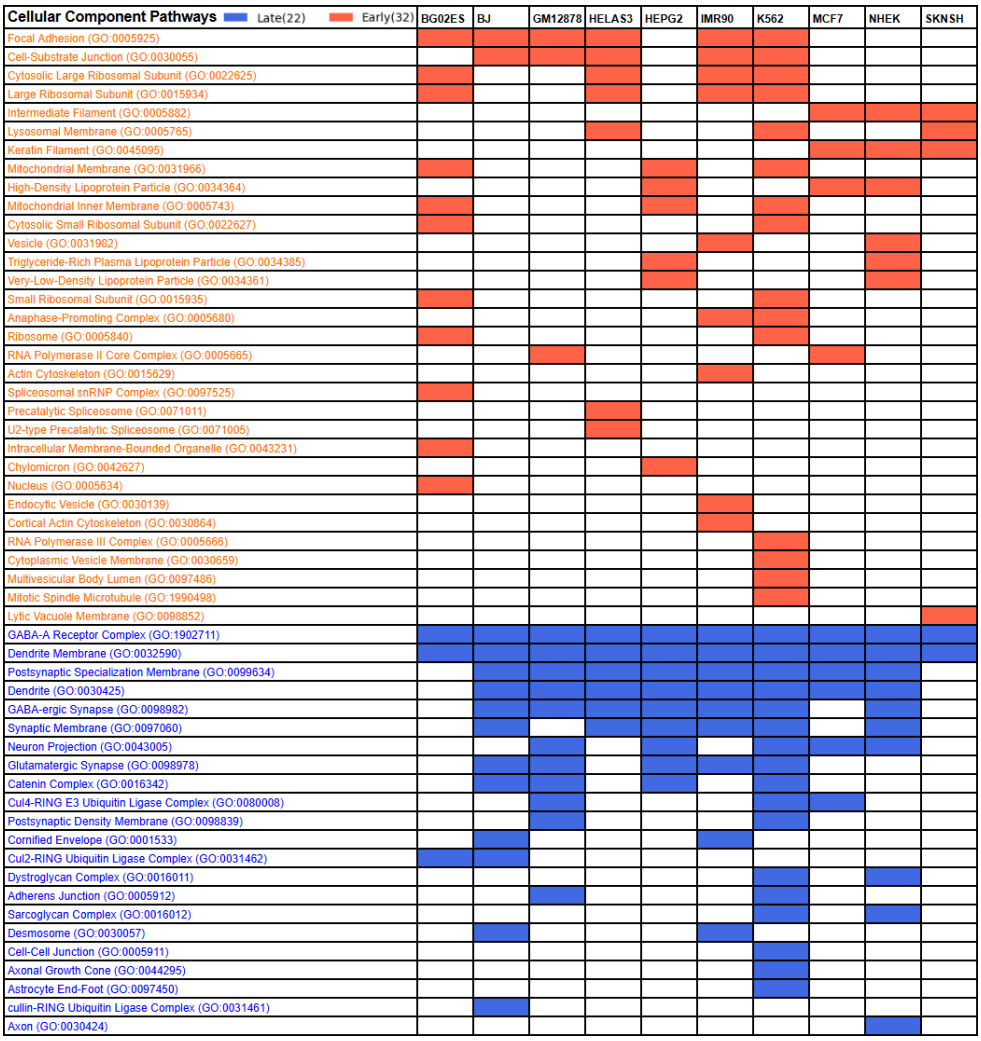

Figure 11: Ancestor-Based GO Terms for Cellular Component

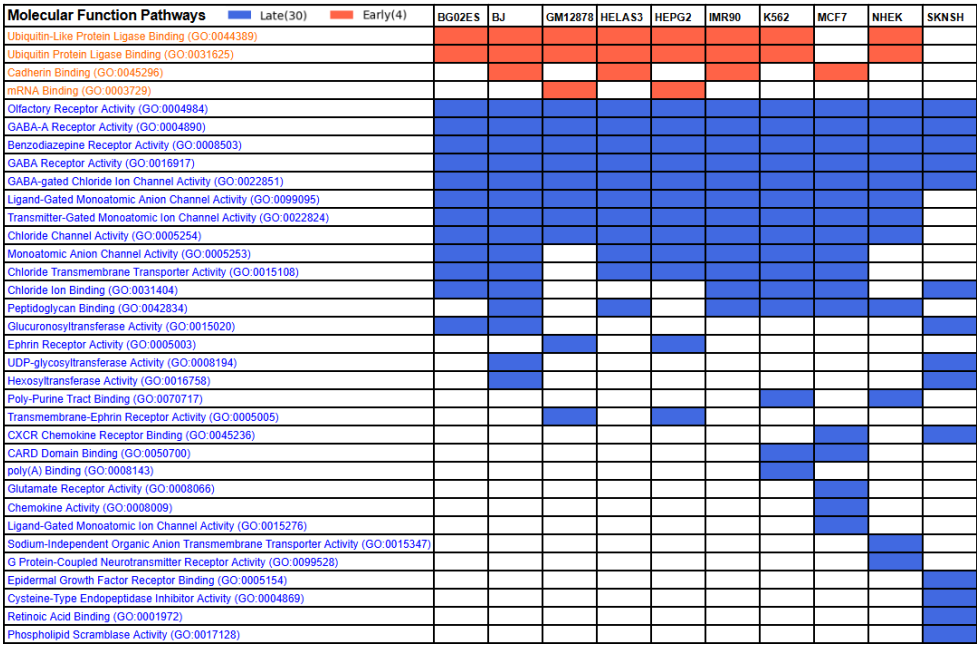

Figure 12: Ancestor-Based GO Terms for Molecular Function

4.2.1 Pathways in Detail

As presented in the figures, pathways classified under biological processes, cellular components, and molecular functions are systematically grouped based on their respective ancestors. The number of distinct cell lines in which each pathway overlaps is indicated alongside the pathway names, while pie charts illustrate the total number of pathways associated with each ancestor.

Appendices A and B present all the ancestral pathways and cell line specific pathways respectively.

4.2.1.1 Biological Process - Response to Stimulus

Figure 13 presents the enriched GO terms mapped to the 'Response to Stimulus' ancestor category across various cell lines. Figures 14 and 15 illustrate the common themes among the early- and late-enriched pathways under this section.

Replication timing reveals a temporal shift from internal homeostasis to environmental interaction. Early-replicating genes were significantly enriched in GO terms related to heavy metal ion response, extrinsic apoptotic signaling, and NF-B pathway regulation. This enrichment suggests that genes governing cellular stress responses and apoptotic regulation are maintained in a transcriptionally permissive chromatin state, consistent with their role as rapidly inducible defense mechanisms. In contrast, late-replicating genes were enriched in GO terms associated with sensory perception, glutamate receptor signaling, histamine

response, and antimicrobial humoral immunity, reflecting their localization within condensed, transcriptionally inactive chromatin.

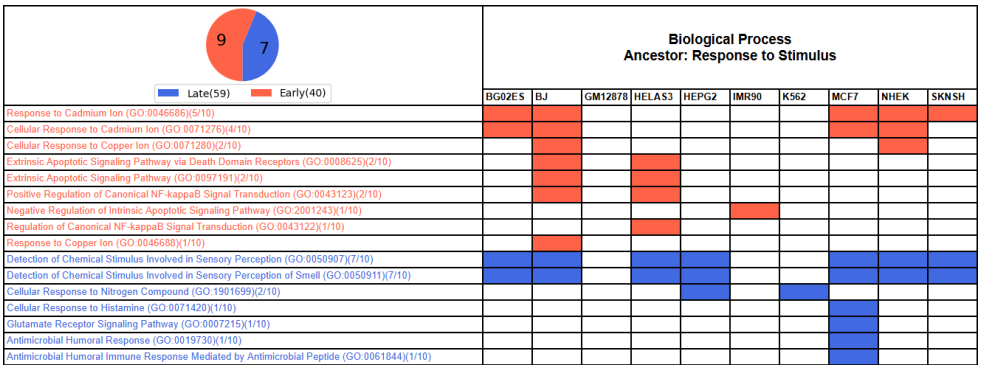

Figure 13: GO Terms for Response to Stimulus and Cell Lines

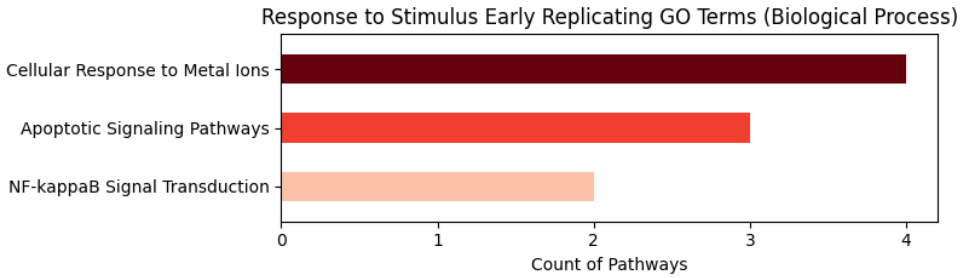

Figure 14: Common Themes for Early Replicating GO Terms Ancestors for Response to Stimulus

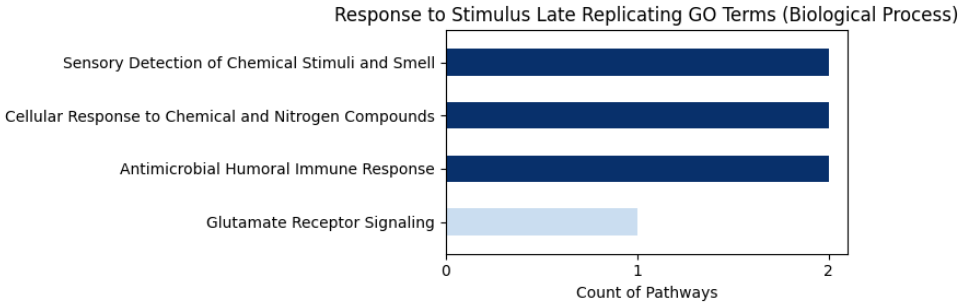

Figure 15: Common Themes for Late Replicating GO Terms Ancestors for Response to Stimulus

4.2.1.2 Biological Process - Multicellular Organismal Process

Pathways and common themes for the remaining ancestors are available in Appendix C.

Early-replicating genes are enriched in Regulation of B Cell Activation. B cell activation regulatory genes replicate early, suggesting they are constitutively ready for expression when immune challenge occurs. Late-replicating genes are significantly enriched in GO terms spanning two closely related biological domains: neuronal development and wiring, and chemosensory and olfactory perception. Genes underpinning nervous system development, particularly those governing neuronal morphogenesis, synapse formation, and olfactory chemosensation, are consistently replicated late during S-phase.

4.2.1.3 Biological Process - Localization

This is consistent with the well-established principle that replication timing is a proxy for chromatin accessibility and gene regulation. The early-replicating genes are enriched in vesicle-based and intracellular transport processes, moving lipids (phospholipids, cholesterol) and proteins within the cell. The late-replicating genes tell a strikingly different story. They are enriched in plasma membrane-level transport, moving small molecules (chloride, anions, CO, O, monocarbon compounds) across cellular boundaries.

4.2.1.4 Biological Process - Immune System Process

Only two pathways are enriched among late-replicating genes, both related to defense against microbial agents. The first pathway, Antimicrobial Humoral Response (GO:0019730), encompasses a broad range of secreted immune defense functions, whereas the second, Antimicrobial Humoral Immune Response Mediated by Antimicrobial Peptide (GO:0061844), specifically represents the host response mediated by antimicrobial peptides. The enrichment of these genes in late-replicating domains of MCF7 reflects the epigenetic silencing of lineage-inappropriate innate immune effector genes in a breast cancer cell line. This is consistent with the established chromatin organization principle that cell-type-specific genes adopt late replication timing in cell types where they are not expressed, and in the cancer context, may additionally reflect the suppression of genes with potential anti-tumor or immune-recruiting functions.

4.2.1.5 Biological Process - Homeostatic Process

Early-replicating genes are enriched for intracellular zinc ion homeostasis (GO:0006882) in both a normal fibroblast line (BJ) and a breast cancer line (MCF7), under the ancestor term Homeostatic Process. This shared enrichment suggests that zinc homeostasis is a genomically privileged, early-replicating function conserved from normal to cancerous cells. In the context of MCF7, this may not be coincidental, given zinc's established role in fueling estrogen receptor-driven breast cancer proliferation. [40]

4.2.1.6 Biological Process - Developmental Process

Early-replicating genes are enriched in Cell Fate Commitment, the molecular machinery that defines and maintains cellular identity, in the BG02ES human embryonic stem cell (hESC) line. Because these genes govern the fundamental decision of what a cell will become, their faithful replication before cell division is essential; errors here would be catastrophic. Late-replicating genes, by contrast, are enriched in a coherent cluster of neuronal GO terms spanning synaptic biology, axon development, neuron differentiation, and broader nervous system and brain development.

4.2.1.7 Biological Process - Cellular Process

The early-replicating GO terms form a coherent biological story centred on housekeeping and broadly expressed genes. These terms span RNA processing, splicing, translation, and protein folding, alongside cell cycle regulation, apoptosis, core carbon metabolism (glucose, pyruvate), mitochondrial transport, and stress signalling (including NF-B). These are the genes a cell must have on at all times in virtually every proliferating cell type.

Late-replicating genes are enriched for the following GO terms that can be summarized into four main themes. The largest theme is synaptic biology, dominated by inhibitory GABAergic signaling, from GABA receptor signaling and chloride transport to synapse assembly and regulation. Glutamate receptor signaling (GO:0007215) also appears, likely reflecting co-regulation of excitatory/inhibitory balance rather than independent enrichment for excitatory function. The second cluster covers cell adhesion and junction assembly, including cadherin-mediated cell-cell adhesion, adherens junction organization, cell-cell junction assembly, and synaptic membrane adhesion. The third cluster is neuron development: axonogenesis, axon guidance, axon development, neuron projection morphogenesis, neuron differentiation, and CNS neuron differentiation. These are the programs neurons use to grow, navigate, and wire up during development. The fourth cluster is smaller and more heterogeneous: steroid and estrogen metabolism, plasminogen activation, phospholipid scrambling, histamine response, and hormone secretion regulation. These likely reflect genomic co-localization with synaptic gene clusters rather than an independent functional enrichment.

4.2.1.8 Biological Process - Biological Regulation

Early-replicating genes are enriched for GO terms that fall into four themes. The first is cell cycle and proliferation control, covering the G1/S and mitotic phase transitions where cells commit to dividing. The second is apoptosis regulation, with both pro-apoptotic (extrinsic signaling via death domain receptors) and anti-apoptotic (negative regulation of intrinsic apoptosis) terms present, reflecting that cells keep both arms of the death machinery accessible. The third is NF-B signaling and immune regulation, including canonical NF-B transduction and B cell activation, consistent with broadly expressed survival and immune response genes in constitutively open chromatin. The fourth is a heterogeneous housekeeping cluster including translation regulation, transcription, growth control, cell fate commitment, and triglyceride metabolism.

Late-replicating genes are enriched in biological processes centered on synaptic signaling and neural circuit organization. The most prominent theme is GABAergic neurotransmission (GO:0007214, GO:0051932), alongside glutamate receptor signaling (GO:0007215) and broader chemical synaptic transmission (GO:0007268), indicating a strong association with synaptic communication. A second major theme is the regulation of synapse and cell junction assembly, including regulation and positive regulation of synapse assembly (GO:0051963, GO:0051965), regulation of presynapse organization and assembly (GO:0099174, GO:1905606), and positive regulation of cell junction assembly (GO:1901890). Nervous system development is also represented through positive regulation of nervous system development (GO:0051962). Beyond synaptic function, late-replicating genes are additionally enriched for estrogen metabolic process (GO:0008210), plasma membrane phospholipid scrambling (GO:0017121), and positive regulation of hormone secretion (GO:0046887), suggesting secondary roles in lipid membrane dynamics and neuroendocrine regulation.

4.2.1.9 Biological Process - Biological Process Involved In Interspecies Interaction Between

The pathways influenced by early-replicated genes are not observed in this context. Late-replicating genes are enriched in two GO terms related to antimicrobial immunity: Antimicrobial Humoral Response (GO:0019730) and Antimicrobial Humoral Immune Response Mediated by Antimicrobial Peptide (GO:0061844). Both terms describe the body's defense against microbial pathogens through secreted antimicrobial factors. This enrichment was exclusive to MCF7, a well-differentiated breast epithelial cancer cell line, and was not observed in any other cell line tested. This suggests that genes governing antimicrobial humoral defense occupy late-replicating genomic regions specifically in MCF7 cells.

4.2.1.10 Cellular Component - Protein Containing Complex

An assessment of the pathways influenced by early-replicating genes reveals multiple overlapping functional themes. Within the ribosomal subunit category, for example, several pathways gain prominence: Large Ribosomal Subunit (GO:0015934), Cytosolic Large Ribosomal Subunit (GO:0022625), Small Ribosomal Subunit (GO:0015935), and Cytosolic Small Ribosomal Subunit (GO:0022627). Similarly, lipoprotein-associated pathways specifically High-Density Lipoprotein Particle (GO:0034364), Very-Low-Density Lipoprotein Particle (GO:0034361), Triglyceride-Rich Plasma Lipoprotein Particle (GO:0034385), and Chylomicron (GO:0042627) constitute another notable shared theme falling under the "protein-containing complex" ancestor term. RNA polymerase complexes represent another major structural group affected by early-replicating genes, supported by pathways such as RNA Polymerase II, Core Complex (GO:0005665) and RNA Polymerase III Complex (GO:0005666). Additionally, the identified pathways cluster heavily around the spliceosome, featuring terms like the Spliceosomal snRNP Complex (GO:0097525), Precatalytic Spliceosome (GO:0071011), and U2-type Precatalytic Spliceosome (GO:0071005). Apart from these broader thematic clusters, the sole remaining pathway is the Anaphase-Promoting Complex (GO:0005680), which is directly linked to the cell cycle.

Pathways affected by late-replicating genes also share common themes. The most notable ones are related to ubiquitin complexes, such as the cullin-RING Ubiquitin Ligase Complex (GO:0031461), Cul2-RING Ubiquitin Ligase Complex (GO:0031462), and Cul4-RING E3 Ubiquitin Ligase Complex (GO:0080008). Another theme includes the Dystroglycan Complex (GO:0016011) and Sarcoglycan Complex (GO:0016012), which help the cell connect with internal and external substances. The list also contains components related to how cells connect with each other, such as the Catenin Complex (GO:0016342). Finally, the GABA-A Receptor Complex (GO:1902711) stands out for its role in synaptic transmission.

4.2.1.11 Cellular Component - Cellular Anatomical Structure

One notable theme among the pathways affected by early-replicating genes is the ribosomal component. This theme includes pathways like Ribosome (GO:0005840), Large Ribosomal Subunit (GO:0015934), Cytosolic Large Ribosomal Subunit (GO:0022625), Small Ribosomal Subunit (GO:0015935), and Cytosolic Small Ribosomal Subunit (GO:0022627), which are all directly related to protein synthesis. Another shared theme consists of pathways that maintain the cell's structural integrity and cytoskeleton. These include the Actin Cytoskeleton (GO:0015629), Cortical Actin Cytoskeleton (GO:0030864), Intermediate Filament (GO:0005882), Keratin Filament (GO:0045095), Cell-Substrate Junction (GO:0030055), and Focal Adhesion (GO:0005925). In addition to structural integrity, these pathways also highlight cell adhesion to external matrices.

Pathways such as RNA Polymerase II, Core Complex (GO:0005665), RNA Polymerase III Complex (GO:0005666), Spliceosomal snRNP Complex (GO:0097525), Precatalytic Spliceosome (GO:0071011), and U2-type Precatalytic Spliceosome (GO:0071005) represent components involved in transcription and splicing processes. Processes like intracellular transport, waste excretion, and cellular uptake form another shared theme in this list. This theme features pathways such as Vesicle (GO:0031982), Cytoplasmic Vesicle Membrane (GO:0030659), Endocytic Vesicle (GO:0030139), Multivesicular Body Lumen (GO:0097486), Lysosomal Membrane (GO:0005765), and Lytic Vacuole Membrane (GO:0098852).

Alongside the broad Intracellular Membrane-Bounded Organelle (GO:0043231) pathway, terms like Nucleus (GO:0005634), Mitochondrial Membrane (GO:0031966), and Mitochondrial Inner Membrane (GO:0005743) point specifically to the nucleus and mitochondria. For lipids and lipoproteins, the prominent pathways are High-Density Lipoprotein Particle (GO:0034364), Very-Low-Density Lipoprotein Particle (GO:0034361), Triglyceride-Rich Plasma Lipoprotein Particle (GO:0034385), and Chylomicron (GO:0042627). Finally, pathways active during cell division, such as the Mitotic Spindle Microtubule (GO:1990498) and Anaphase-Promoting Complex (GO:0005680), are also included in this group.

Pathways affected by late-replicating genes show different themes compared to early-replicating ones. The first theme involves cellular component pathways related to nerve cell communication. These include Neuron Projection (GO:0043005), Axon (GO:0030424), Axonal Growth Cone (GO:0044295), Dendrite (GO:0030425), and Dendrite Membrane (GO:0032590). Another shared theme is formed by synaptic components, which generally include synaptic membranes and different types of synapses. This group consists of Synaptic Membrane (GO:0097060), Postsynaptic Specialization Membrane (GO:0099634), Postsynaptic Density

Membrane (GO:0098839), Glutamatergic Synapse (GO:0098978), and GABAergic Synapse (GO:0098982).

Under the theme of intercellular connectivity, we find the Cell-Cell Junction (GO:0005911), Adherens Junction (GO:0005912), Catenin Complex (GO:0016342), and Desmosome (GO:0030057) pathways. These terms stand out as components involved in how cells connect to one another. Additionally, two pathways are related to the connection of the cell membrane with internal and external substances: the Dystroglycan Complex (GO:0016011) and Sarcoglycan Complex (GO:0016012). Finally, pathways related to structures that act as barriers are also notable among late-replicating genes, as seen in Astrocyte End-Foot (GO:0097450) and Cornified Envelope (GO:0001533).

4.2.1.12 Molecular Function - Transporter Activity

Under the "transporter activity" ancestor term, only pathways affected by late-replicating genes are found. Within this category, Ligand-Gated Monoatomic Ion Channel Activity (GO:0015276), Transmitter-Gated Monoatomic Ion Channel Activity (GO:0022824), Ligand-Gated Monoatomic Anion Channel Activity (GO:0099095), and GABA-gated Chloride Ion Channel Activity (GO:0022851) show functions related to ion channels. Other pathways, such as Monoatomic Anion Channel Activity (GO:0005253), Chloride Channel Activity (GO:0005254), and Chloride Transmembrane Transporter Activity (GO:0015108), specialize in the transport of negatively charged ions, particularly chloride. Finally, the Phospholipid Scramblase Activity (GO:0017128) pathway plays a role in phospholipid transport.

4.2.1.13 Molecular Function - Molecular Transducer Activity

Similar to the "transporter activity" heading, there are no pathways affected by early-replicating genes under "molecular transducer activity." Instead, this group contains pathways affected by late-replicating genes. These include GABA-related pathways such as GABA Receptor Activity (GO:0016917), GABA-A Receptor Activity (GO:0004890), GABA-gated Chloride Ion Channel Activity (GO:0022851), and Benzodiazepine Receptor Activity (GO:0008503). For neurotransmitter functions, pathways like Transmitter-Gated Monoatomic Ion Channel Activity (GO:0022824), Glutamate Receptor Activity (GO:0008066), and G Protein-Coupled Neurotransmitter Receptor Activity (GO:0099528) are present.

Additionally, the list features pathways related to ephrin receptors: Ephrin Receptor Activity (GO:0005003) and Transmembrane-Ephrin Receptor Activity (GO:0005005). Finally, the Chemokine Activity (GO:0008009) pathway, which is related to immunity, and the Olfactory Receptor Activity (GO:0004984) pathway also stand out in this set.

4.2.1.14 Molecular Function - Molecular Function Regulator Activity

Under the "molecular function regulator activity" heading, the Chemokine Activity (GO:0008009) pathway is present in relation to immunity, alongside the Cysteine-Type Endopeptidase Inhibitor

Activity (GO:0004869) pathway for endopeptidase inhibitor activity. No pathways affected by early-replicating genes were found under this heading either.

4.2.1.15 Molecular Function - Catalytic Activity

Some pathways affected by late-replicating genes can be grouped under sugar group functions, which include Hexosyltransferase Activity (GO:0016758), UDP-glycosyltransferase Activity (GO:0008194), and Glucuronosyltransferase Activity (GO:0015020) targeting the UDP acid molecule. Another cluster consists of ephrin receptor activities: Ephrin Receptor Activity (GO:0005003) and Transmembrane-Ephrin Receptor Activity (GO:0005005). Finally, the Cysteine-Type Endopeptidase Inhibitor Activity (GO:0004869) pathway is present as an endopeptidase inhibitor activity. Similar to the previous groups, no pathways affected by early-replicating genes were found here.

4.2.1.16 Molecular Function - Binding

Pathways affected by early-replicating genes can first be analyzed under the "ubiquitin protein ligase binding" heading. This category can be summarized as the binding of proteins to relevant units. Under this heading, we find the Ubiquitin Protein Ligase Binding (GO:0031625) and Ubiquitin-Like Protein Ligase Binding (GO:0044389) pathways. Other pathways in this group include Cadherin Binding (GO:0045296), which shows binding to cadherins, and mRNA Binding (GO:0003729), which is related to transcriptional binding.

Regarding pathways affected by late-replicating genes, Chemokine Activity (GO:0008009) and CXCR Chemokine Receptor Binding (GO:0045236) are notable for their effects on immune cells. This group also includes Peptidoglycan Binding (GO:0042834) for binding to bacterial cell walls, and CARD Domain Binding (GO:0050700), which is also related to immune signaling. For binding to purine regions, the Poly-Purine Tract Binding (GO:0070717) pathway is present, along with its specialized form, poly(A) Binding (GO:0008143). Additionally, Epidermal Growth Factor Receptor Binding (GO:0005154) is related to cell growth, while Retinoic Acid Binding (GO:0001972) is involved in cell differentiation. Finally, the Chloride Ion Binding (GO:0031404) pathway is included in this group for chloride ion binding.

4.2.2 DNA Replication Time Distribution

The results were obtained for 10 different cell lines. Accordingly, DNA replication timing values were calculated for each of these cell lines in this study. Figure 16 presents the box plots of scores assigned to protein-coding genes across the cell lines. As shown in the plot, the score distributions are largely similar across all cell lines. Figure 17 illustrates the Kernel Density Estimation (KDE) distributions of the cell lines. This plot illustrates the density distribution of scores assigned to protein coding genes for each cell line.

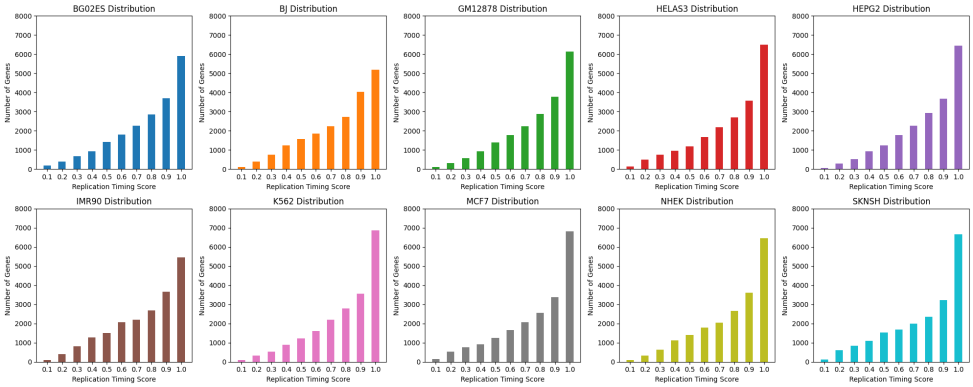

Figure 16: Cell Line Distribution

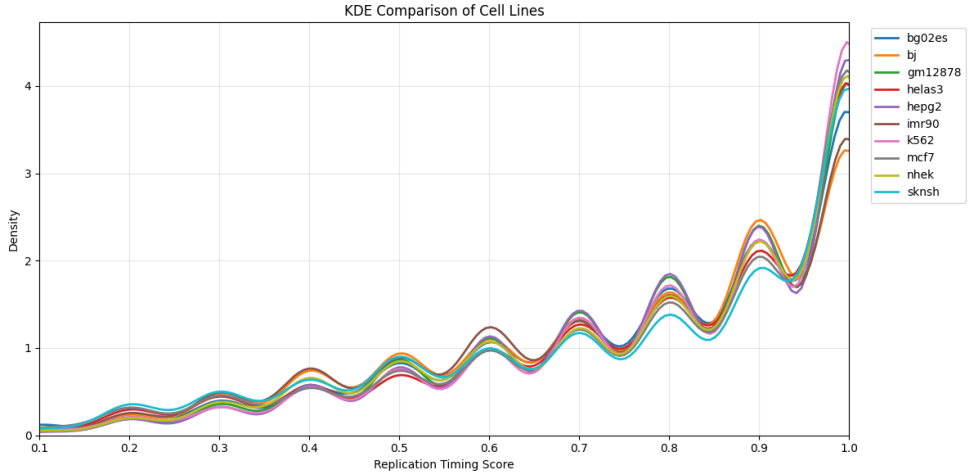

Figure 17: KDE Comparison for Cell Lines

Examination of the distributions reveals that genes predominantly exhibit high scores across all cell lines, indicating that they reside in early replicating regions. This observation is biologically plausible, as early replicating domains are known to be evolutionarily conserved. Furthermore, given that this study focuses specifically on protein coding genes, the obtained distributions appear highly consistent with established genomic patterns.

The replication timing profiles generated by SigProfilerTopography exhibit a non-uniform distribution for gene replication timings. Specifically, a gene-level assessment indicated that values are limited to only two or three bins, while the majority remain unoccupied. We hypothesize that this inherent data sparsity is the primary driver of the value concentration at integers observed during the scoring process.

Figures 18 and 19 illustrate the gene replication timing data, presenting both the mean and standard deviation values alongside the corresponding violin plots. Despite similar means and standard deviations, the enrichment of distinct pathways in the GO analysis indicates that different genes populate these distributions. As illustrated by the violin plot, the similar

distribution of protein coding genes across cell lines is consistent with our earlier observation that these genes predominantly localize to early replicating regions. Additionally, our scoring methodology may also contribute to this pattern.

The score creation method was utilized to generate replication timing scores, as it provides a more comprehensive representation of the temporal data compared to other statistical heuristics. Unlike the 'argmax' method, which identifies only the peak position and neglects the surrounding data, or the median-based approach, which tends to diminish the significance of the non-zero time bins, our score creation approach preserves the integrity of the overall replication timing distribution. Because this scoring approach bounds all protein coding genes between 0.1 and 1.0, the resulting density distributions appear remarkably similar.

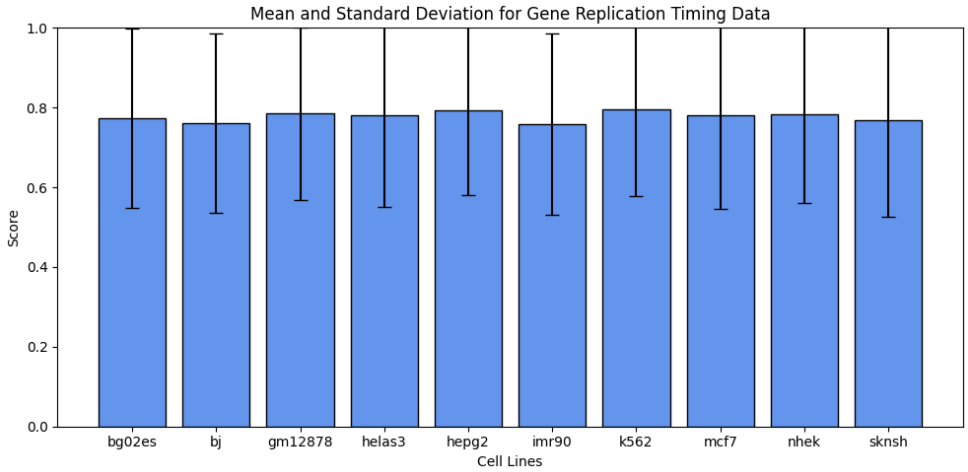

Figure 18: Mean and Standard Deviation for Genes

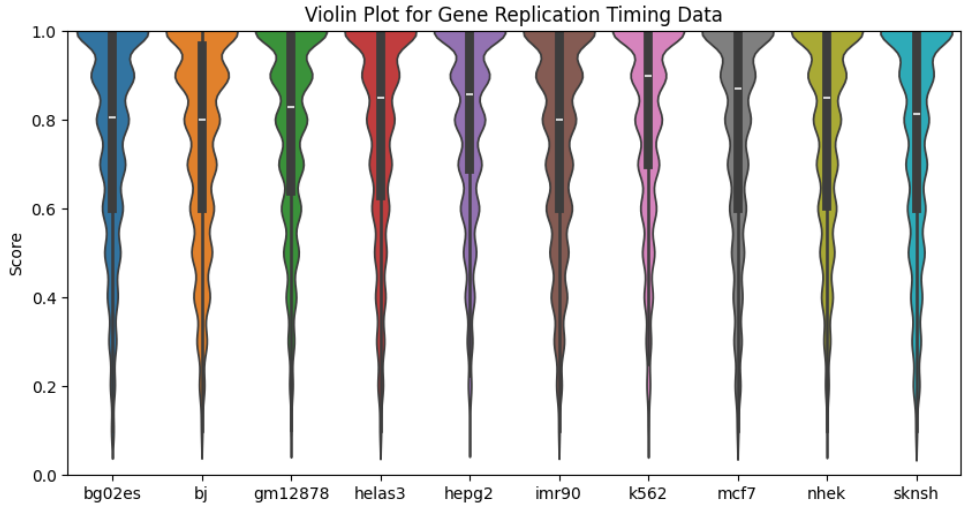

Figure 19: Gene Replication Timing Data Distribution

4.2.3 Results for GO Term Analysis for Individual Cell Lines

4.2.3.1 BG02ES

The common themes among the pathways uniquely enriched in the BG02ES cell line is presented in Figures 20 and 21.

Among the pathways unique to this cell line, none are affected by late-replicating genes. For early-replicating genes, pathways are only present under the biological process and cellular component categories.

The pathways specific to this cell line that are related to RNA processing and the spliceosome include RNA Processing (GO:0006396), RNA Splicing (GO:0008380), mRNA Splicing, via Spliceosome (GO:0000398), and RNA Splicing, via Transesterification Reactions With Bulged Adenosine as Nucleophile (GO:0000377). Additionally, the transcription-related Positive Regulation of DNA-templated Transcription (GO:0045893) pathway is uniquely found in this cell line. For cell growth and fate, the Negative Regulation of Growth (GO:0045926) and Regulation of Cell Fate Commitment (GO:0010453) pathways are notable.

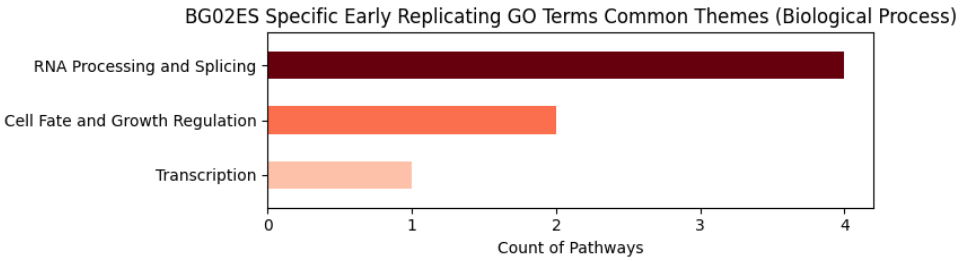

Figure 20: BG02ES Specific Early Replicating GO Terms Common Themes for Biological Process

Finally, under cellular components, the Spliceosomal snRNP Complex (GO:0097525), Nucleus (GO:0005634), and Intracellular Membrane-Bounded Organelle (GO:0043231) pathways stand out in relation to nuclear structure.

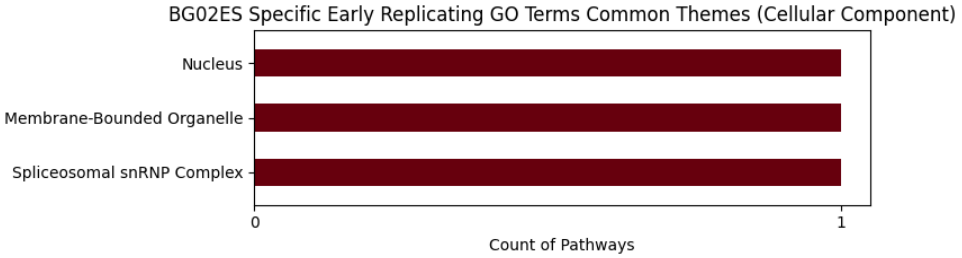

Figure 21: BG02ES Specific Early Replicating GO Terms Common Themes for Cellular Component

4.2.3.2 BJ

Cell line specific common themes for other cell lines can be found in Appendix D.

In this cell line, pathways affected by early-replicating genes are found only under the Biological Process category. These include the Mitochondrial Transport (GO:0006839) pathway, which is related to mitochondrial transport, and the Response to Copper Ion (GO:0046688) pathway, reflecting a cellular response. For late-replicating genes, the affected pathways include Positive Regulation of Hormone Secretion (GO:0046887), which regulates hormone secretion, and the cullin-RING Ubiquitin Ligase Complex (GO:0031461) pathway, related to protein degradation.

4.2.3.3 GM12878

Among the pathways affected by early-replicating genes, only the Pyruvate Metabolic Process (GO:0006090) is present. This pathway covers processes such as the synthesis and breakdown of pyruvate. On the other hand, looking at the pathways affected by late-replicating genes, both pathways specific to this cell line are related to nervous system processes: Neuron Differentiation (GO:0030182) and Axon Guidance (GO:0007411).

4.2.3.4 HELAS3

No pathways affected by late-replicating genes are found specific to this cell line. Instead, the Negative Regulation of Transferase Activity (GO:0051348) pathway is present, which reduces transferase activity. The 'De Novo' Post-Translational Protein Folding (GO:0051084) pathway is also included in the list as a type of protein control mechanism. While the Apoptotic Process (GO:0006915) highlights processes related to cell death, the Regulation of Canonical NF-kB Signal Transduction (GO:0043122) pathway is associated with NF-kB signaling activity. Finally, the Precatalytic Spliceosome (GO:0071011) and U2-type Precatalytic Spliceosome (GO:0071005) pathways, which represent spliceosome complexes, are also among the enriched terms for this cell line.

4.2.3.5 HEPG2

Only a single cell line-specific pathway is identified for the HEPG2 cell line. This is the Chylomicron (GO:0042627) pathway, which is associated with lipoprotein complexes and transport.

4.2.3.6 IMR90

The Negative Regulation of Intrinsic Apoptotic Signaling Pathway (GO:2001243), which slows down or potentially halts the cell death process, appears to be enriched in this cell line. Additionally, components related to intracellular stability and cell structure, such as the

Actin Cytoskeleton (GO:0015629), Cortical Actin Cytoskeleton (GO:0030864), and Endocytic Vesicle (GO:0030139), are also enriched in this cell line.

4.2.3.7 K562

In this cell line, the Intracellular Protein Transport (GO:0006886) pathway and the Membraneless Organelle Assembly (GO:0140694) pathway, which reflects the assembly process of certain intracellular organelles, are enriched. Additionally, the Regulation of B Cell Activation (GO:0050864) pathway is enriched in relation to the activity of B cells, which play an important role in the immune system. Regarding intracellular transport, the Multivesicular Body Lumen (GO:0097486) and Cytoplasmic Vesicle Membrane (GO:0030659) pathways stand out under the cellular component category. Furthermore, the Mitotic Spindle Microtubule (GO:1990498) structure, which functions during cell division, and the RNA Polymerase III Complex (GO:0005666), involved in protein synthesis, are also observed among the pathways affected by early-replicating genes.

Regarding the pathways affected by late-replicating genes, Presynapse Assembly (GO:0099054), which plays a role in connecting neuronal structures, and the transcriptional function poly(A) Binding (GO:0008143) are prominent. Additionally, components involved in synapse formation, such as Astrocyte End-Foot (GO:0097450) and Axonal Growth Cone (GO:0044295), along with the Cell-Cell Junction (GO:0005911) component related to intercellular connectivity, stand out.

4.2.3.8 MCF7

As a result of the analyses, the Glucose Metabolic Process (GO:0006006) pathway, which involves the metabolic activities of glucose—regarded as the primary cellular fuel and the Positive Regulation of Triglyceride Metabolic Process (GO:0090208) pathway, which covers the metabolism of lipids as an alternative energy source, emerge as pathways affected by early-replicating genes. Similarly, the Phospholipid Efflux (GO:0033700) pathway, associated with phospholipid transport, is also included in this list.

Regarding the pathways affected by late-replicating genes, the Antimicrobial Humoral Response (GO:0019730) and Antimicrobial Humoral Immune Response Mediated by Antimicrobial Peptide (GO:0061844) pathways, both involved in responses against pathogens, are observed. The Glutamate Receptor Signaling Pathway (GO:0007215), associated with synaptic transmission, and the Cellular Response to Histamine (GO:0071420) pathway, which covers the cellular response to histamine, are also present in this list. Additionally, the Ligand-Gated Monoatomic Ion Channel Activity (GO:0015276) related to ion transport, and the Chemokine Activity (GO:0008009) pathway involved in cell movement, stand out. Finally, the Glutamate Receptor Activity (GO:0008066) pathway, which functions in synaptic transmission, is also found among the terms affected by late-replicating genes.

4.2.3.9 NHEK

The Supramolecular Fiber Organization (GO:0097435) pathway, which is involved in the organization of intracellular and extracellular fiber structures, and the Cholesterol Efflux (GO:0033344) pathway, related to the transport of cholesterol out of the cell, stand out as pathways affected by early-replicating genes.

On the other hand, under the biological process category, Brain Development (GO:0007420), which describes neural development, and the Sodium-Independent Organic Anion Transport (GO:0043252) pathway, representing the sodium-independent export of organic anions from the cell, appear to be enriched by late-replicating genes. Similarly, the Axon (GO:0030424) component, which functions in electrical signal transmission in neurons, stands out. Furthermore, mirroring these patterns in the cellular component category, the function of sodium-independent organic anion transport is also present under the molecular function category via the Sodium-Independent Organic Anion Transmembrane Transporter Activity (GO:0015347) pathway. Finally, the G Protein-Coupled Neurotransmitter Receptor Activity (GO:0099528) pathway, associated with signal transduction, is also included in this list.

4.2.3.10 SKNSH

Regarding the lysosome organelle responsible for cellular waste management, the Lytic Vacuole Membrane (GO:0098852) pathway appears to be the only pathway enriched by early-replicating genes.

On the other hand, several pathways enriched by late-replicating genes are present. Under the biological process category, the terms Gas Transport (GO:0015669), Oxygen Transport (GO:0015671), and Carbon Dioxide Transport (GO:0015670), which are directly associated with gas transport, are observed. The Steroid Metabolic Process (GO:0008202) pathway, representing the metabolic processes of steroid hormones, and the Plasma Membrane Phospholipid Scrambling (GO:0017121) pathway, which reflects the translocation of phospholipid structures, are also included in this list. Furthermore, the One-Carbon Compound Transport (GO:0019755) and Signaling Receptor Ligand Precursor Processing (GO:0140448) pathways, which function in DNA synthesis and the activation of signaling proteins, are identified.

Regarding cell growth and differentiation, Epidermal Growth Factor Receptor Binding (GO:0005154) and Retinoic Acid Binding (GO:0001972) are present, alongside Cysteine-Type Endopeptidase Inhibitor Activity (GO:0004869), which is related to cell death. Finally, corresponding to the phospholipid transport process mentioned in the biological process category, the Phospholipid Scramblase Activity (GO:0017128) pathway is found under the molecular function heading.

CHAPTER 5

DISCUSSION

5.1 Model

Initial R^2 and MSE scores looked promising, but a closer check showed these metrics were misleading. Our initial suspicion was that the models were overfitting the training data. However, an assessment of the learning curves contradicted this hypothesis; the loss trajectory of the ANN indicated a state closer to underfitting, suggesting that extending the training phase could theoretically push the model toward a perfect fit. Yet, an analysis of the Random Forest Regressor which outwardly outperformed the ANN demonstrated that the core issue was entirely independent of training duration or algorithmic tuning.

The Random Forest model produced exceptionally high R^2 and MSE scores, creating an illusion of highly consistent predictive capability. The reality behind these metrics surfaced during the feature importance analysis, which clearly highlighted the structural limitations of the dataset: a lack of sufficient sample size, low diversity, and most critically, an absence of meaningful underlying relationships. Specifically, the algorithm almost entirely bypassed the SBS data. Instead of extracting functional patterns, the model minimized its error metrics by merely mimicking the inherent gene replication timings. Because the prediction targets closely align with these replication timings, the model found a mathematical shortcut it ignored the actual input and replicated the baseline timing data. While this could loosely be interpreted as a form of feature-specific overfitting, it is more accurately a complete failure to learn the intended mapping.

To definitively diagnose the dataset's viability for regression tasks, we conducted a detailed correlation analysis. The resulting matrix provided a clear explanation for the predictive failure: while the SBS data points exhibit consistent intra-feature correlations, they show absolutely no correlation with gene replication timings. More importantly, the input SBS values are entirely uncorrelated with the replication timings of the specific nucleotides expected as the output. This fundamental absence of signal clearly explains why the trained models failed to perform beyond superficial metric optimization.

With this weak relationship, conventional strategies such as expanding the dataset or adding new features are unlikely to improve performance, as the existing variables appear to have only a limited relationship with the target. This realization also justifies our decision not to pursue deeper or more complex ANN architectures. While neural networks excel at capturing

highly complex, non-linear dependencies, no architecture can extract meaningful predictive rules from uncorrelated data.

Therefore, adjusting hyperparameters like the number of epochs did not improve the loss. The data structure causes the network to memorize the dominant feature instead of learning to generalize. We hesitate to classify this strictly as classical overfitting, primarily because the model does not fail catastrophically on unseen data; when presented with novel test instances, it simply repeats the same baseline prediction strategy. Ultimately, the limited scope and diversity of the dataset prevent a definitive classification of the error topology, but the central conclusion remains clear: the lack of intrinsic feature-target correlation renders the current regression approach unviable.

5.2 Pathways Affected by Early Replicated Genes

Functional pathways linked to early-replicating genes converge around a few core themes. Across all evaluated cell lines, these gene sets are consistently tied to gene expression, post-translational quality control, and structural scaffolding. The prominence of structural scaffolding highlights a strict prioritization of processes designed to preserve overall cellular integrity. Coupled with the observed integration of metal ion responses and metabolic reprogramming, these molecular signatures clearly constitute the foundational infrastructure necessary for the cell to sustain its essential functions and manage internal stress.

These observations reinforce the understanding that the cellular replication program initially prioritizes core survival mechanisms. Strikingly, our analysis revealed an absolute functional divergence between early and late replication domains. There is zero intersection between the pathways influenced by these two groups; they orchestrate entirely distinct biological operations. This clear separation is further supported by the distribution of the genes. Early-replicating genes significantly outnumber their late-replicating counterparts, yet this large group of early genes influences only a relatively narrow, highly focused set of pathways.

While certain pathways appeared exclusive to individual cell lines, mapping these annotations to their ancestral nodes allowed us to isolate and clarify the broad biological logic. When viewed through this ancestral hierarchy, it became evident that pathways sharing multiple upstream ancestors were heavily concentrated within core 'biological process' categories. To maintain analytical rigor and ensure the relevance of this functional network, all obsolete pathway annotations were excluded from the evaluation.

5.3 Pathways Affected by Late Replicated Genes

Although late-replicating genes are less abundant than the early-replicating genes, they still share a defined set of functional themes. Generally, the pathways driven by late-replicating genes are heavily implicated in environmental interactions and neurodevelopmental processes. They consistently aggregate around mechanisms of intercellular communication, cell adhesion, and molecular transport. Aligning closely with existing literature, our data indicate that

this temporal replication domain is fundamentally dedicated to mediating the cell's extrinsic relationships and specialized developmental trajectories.

This functional dichotomy clearly explains the complete absence of overlap between the early and late replicating networks. Early-replicating domains are enriched for essential cellular processes and core expression, whereas late-replicating domains govern specialized activities that proceed once homeostasis is secured.

Mapping these late-replicating pathways to their ancestral nodes revealed topological patterns similar to the early cohort. Although certain pathways remain unique to individual cell lines, the proportion of universally conserved pathways across all analyzed cell lines is substantially higher within the late-replicating group. This consistency across cell lines supports the reliability of our functional mapping.

The underlying distribution of the replication timing scores provides a plausible statistical explanation for these observed differences in pathway diversity. In our methodology, genes with a replication score above or equal 0.9 were classified as early, whereas those scoring below or equal 0.2 were designated as late. The early-replicating category encompasses a massive volume of genes, saturated particularly with scores peaking exactly at 1.0. This sheer numerical advantage likely drives the broader diversity and variance of functional pathways observed in the early cohort. The late-replicating group, conversely, lacks this density. Genes scoring near the extreme low end (e.g., 0.1) are rare. In fact, even with the relatively permissive 0.2 threshold, the absolute count of late-replicating genes remains lower than just the subset of early genes scoring a perfect 1.0. Finally, consistent with the methodology applied to the early cohort, all obsolete pathway annotations were excluded from this evaluation to maintain analytical precision.

5.4 Cell Line Individual Pathway Analysis

5.4.1 BG02ES

An analysis of the pathways specific to embryonic stem cells shows that no pathways are affected by late-replicating genes. Instead, pathways related to nuclear structure, RNA splicing, and RNA processing stand out. These pathways provide a consistent outcome for embryonic stem cells. The identified GO terms are fully consistent with the ability of embryonic stem cells to maintain their cellular identity, as well as to develop and differentiate.

5.4.2 BJ

When examining these pathways specific to skin tissue, the involvement of mitochondrial transport is logical given the energy required for cell proliferation and protein synthesis. For pathways affected by late-replicating genes, the results are consistent when considering protein degradation and signaling in response to external stimuli at the stage where cells form a protective barrier.

5.4.3 GM12878

Since axon and neuron processes are not active in blood tissue, it is completely natural that these pathways are enriched by late-replicating genes. On the other hand, the pathway affected by early-replicating genes is consistent with generating the energy required by blood cells.

5.4.4 HELAS3

When examining these pathways specifically in cervix tissue, they appear logical due to factors such as immunity, the controlled death of damaged cells, cancer prevention, and the avoidance of errors during intensive protein synthesis. Additionally, due to other reasons like the rapid renewal of epithelial cells, the resulting pathway list seems to align well with the cervix list.

5.4.5 HEPG2

In liver tissue, the enrichment of pathway by early-replicating genes is logical given the liver's active role in lipid and cholesterol trafficking, which is one of its primary functions.

5.4.6 IMR90

The pathways identified from the analysis appear to present a consistent pattern for lung tissue. These pathways seem specialized to protect cells against mechanical stress and to delay or prevent cell death in the presence of excess oxygen or toxic substances. This provides a logical explanation for the enrichment of these pathways by early-replicating genes in lung tissue.

5.4.7 K562

Regarding bone marrow tissue, the high rate of cell division and the direct involvement of this tissue with B lymphocytes clearly explain why these pathways are enriched by early-replicating genes. Furthermore, the pathways affected by late-replicating genes appear highly plausible, considering the absence of neuronal structures in bone marrow tissue and the fact that the produced blood cells do not need to form junctions with other cells.

5.4.8 MCF7

When examined specifically in breast tissue, the early-enriched pathways demonstrate a consistent pattern regarding fat cell storage and energy requirements. Similarly, the presence of immune and neural processes among the late-enriched pathways for this tissue is logical. This is because breast tissue has no functional requirement to actively participate in the immune system like bone tissue, nor does it need to maintain neural processes like brain cells.

5.4.9 NHEK

In this cell line, which represents another skin tissue of a different cell type, the enrichment of brain tissue-related pathways by late-replicating genes can be considered entirely natural. Similarly, the inclusion of anion transport systems that actively function in organs like the liver and kidney presents a consistent pattern. Furthermore, considering the mechanical flexibility, structural integrity, and protective function of skin tissue, the pathways enriched by early-replicating genes also appear highly consistent.

5.4.10 SKNSH

Considering that cells in brain tissue do not divide, the activity of lysosomal systems to prevent protein accumulation appears entirely natural. On the other hand, since gas transport is a function of red blood cells rather than brain cells, the late enrichment of these processes is also consistent. Furthermore, the suppression of growth hormones to prevent uncontrolled cell growth and thus tumor formation presents a biologically logical and coherent framework.

5.5 Contribution of This Study

In this study, we aimed to develop a model to mitigate the significant computational cost and time associated with calculating replication timing data. Due to data limitations and incompatibility, we refined the scope of this study. We analyzed both common and cell-line-specific GO terms for protein-coding genes.

Following this, we investigated the protein-coding genes across ten distinct cell lines, distinguishing between common and cell-line-specific GO terms. The analysis of these common GO terms indicated that early-replicating genes are mainly involved in core processes like transcription, translation, and gene expression. In contrast, late-replicating genes are clustered around transport, cell signaling, and neuronal development.

In the analysis of cell line specific GO terms, we observed that the identified terms consistently aligned with the characteristics of each respective cell line. For example, transport processes are linked to late-replicating genes in brain tissue, whereas pathways handling mechanical stress are managed by early-replicating genes in lung tissue. The presence of the 'Regulation of Cell Fate Commitment' (GO:0010453) term in the BG02ES cell line serves as a compelling example of our findings' validity. As an embryonic stem cell line, the association of this GO term with early-replicating genes aligns well with the cell line's developmental status. As another example, GO terms related to B-cell activation in bone marrow tissue were found to be enriched within the early-replicating gene set. These demonstrate a clear, consistent link between tissue types and functional pathways at the cell-line level.

The contribution of this study lies in identifying the biological processes, cellular components, and molecular functions associated with early and late-replicating protein-coding genes, both as shared GO Terms across cell lines and as cell-line-specific.

REFERENCES

[1] B. Otlu and L. B. Alexandrov, "Evaluating topography of mutational signatures with sigprofilertopography," *Genome Biology*, vol. 26, no. 1, p. 134, 2025.

[2] N. Rhind and D. M. Gilbert, "Dna replication timing," *Cold Spring Harbor perspectives in biology*, vol. 5, no. 8, p. a010132, 2013.

[3] B. D. Pope, T. Ryba, V. Dileep, F. Yue, W. Wu, O. Denas, D. L. Vera, Y. Wang, R. S. Hansen, T. K. Canfield, *et al.*, "Topologically associating domains are stable units of replication-timing regulation," *Nature*, vol. 515, no. 7527, pp. 402–405, 2014.

[4] I. Hiratani, T. Ryba, M. Itoh, T. Yokochi, M. Schwaiger, C.-W. Chang, Y. Lyou, T. M. Townes, D. Schübeler, and D. M. Gilbert, "Global reorganization of replication domains during embryonic stem cell differentiation," *PLoS biology*, vol. 6, no. 10, p. e245, 2008.

[5] A. Koren, P. Polak, J. Nemesh, J. J. Michaelson, J. Sebat, S. R. Sunyaev, and S. A. McCarroll, "Differential relationship of dna replication timing to different forms of human mutation and variation," *The American Journal of Human Genetics*, vol. 91, no. 6, pp. 1033–1040, 2012.

[6] C. Marchal, T. Sasaki, D. Vera, K. Wilson, J. Sima, J. C. Rivera-Mulia, C. Trevilla-García, C. Nogues, E. Nafie, and D. M. Gilbert, "Genome-wide analysis of replication timing by next-generation sequencing with e/l repli-seq," *Nature protocols*, vol. 13, no. 5, pp. 819–839, 2018.

[7] I. Hiratani and S. Takahashi, "Dna replication timing enters the single-cell era," *Genes*, vol. 10, no. 3, p. 221, 2019.

[8] J. van den Berg, V. van Batenburg, C. Geisenberger, R. B. Tjeerdsma, A. de Jaime-Soguero, S. P. Acebrón, M. A. van Vugt, and A. van Oudenaarden, "Quantifying dna replication speeds in single cells by scedu-seq," *Nature methods*, vol. 21, no. 7, pp. 1175–1184, 2024.

[9] F. Berkemeier, P. R. Cook, and M. A. Boemo, "Dna replication timing reveals genome-wide features of transcription and fragility," *Nature Communications*, vol. 16, no. 1, p. 4658, 2025.

[10] F. Liu, C. Ren, H. Li, P. Zhou, X. Bo, and W. Shu, "De novo identification of replication-timing domains in the human genome by deep learning," *Bioinformatics*, vol. 32, no. 5, pp. 641–649, 2016.

[11] Y. Yang, Y. Wang, Y. Zhang, and J. Ma, "Concert: genome-wide prediction of sequence elements that modulate dna replication timing," *bioRxiv*, pp. 2022–04, 2022.

[12] I. Hiratani, A. Leskovar, and D. M. Gilbert, "Differentiation-induced replication-timing changes are restricted to at-rich/long interspersed nuclear element (line)-rich isochores,"

Proceedings of the National Academy of Sciences, vol. 101, no. 48, pp. 16861–16866, 2004.

[13] E. Yaffe, S. Farkash-Amar, A. Polten, Z. Yakhini, A. Tanay, and I. Simon, "Comparative analysis of dna replication timing reveals conserved large-scale chromosomal architecture," *PLoS genetics*, vol. 6, no. 7, p. e1001011, 2010.

[14] T. Ryba, I. Hiratani, J. Lu, M. Itoh, M. Kulik, J. Zhang, T. C. Schulz, A. J. Robins, S. Dalton, and D. M. Gilbert, "Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types," *Genome research*, vol. 20, no. 6, p. 761, 2010.

[15] J. C. Rivera-Mulia, Q. Buckley, T. Sasaki, J. Zimmerman, R. A. Didier, K. Nazor, J. F. Loring, Z. Lian, S. Weissman, A. J. Robins, *et al.*, "Dynamic changes in replication timing and gene expression during lineage specification of human pluripotent stem cells," *Genome research*, vol. 25, no. 8, p. 1091, 2015.

[16] M. S. Lawrence, P. Stojanov, P. Polak, G. V. Kryukov, K. Cibulskis, A. Sivachenko, S. L. Carter, C. Stewart, C. H. Mermel, S. A. Roberts, *et al.*, "Mutational heterogeneity in cancer and the search for new cancer-associated genes," *Nature*, vol. 499, no. 7457, pp. 214–218, 2013.

[17] D. Bolzan and F. Ay, "Soffritto: a deep learning model for predicting high-resolution replication timing," *Bioinformatics*, vol. 41, no. Supplement_1, pp. i580–i589, 2025.

[18] F. Dao, B. Lebeau, C. C. Y. Ling, M. Yang, X. Xie, M. J. Fullwood, H. Lin, and H. Lyu, "Replichrom: Interpretable machine learning predicts cancer-associated enhancer-promoter interactions using dna replication timing," *Imeta*, vol. 4, no. 4, p. e70052, 2025.

[19] C. A. Müller and C. A. Nieduszynski, "Dna replication timing influences gene expression level," *Journal of Cell Biology*, vol. 216, no. 7, pp. 1907–1914, 2017.

[20] A. Koren, R. E. Handsaker, N. Kamitaki, R. Karlić, S. Ghosh, P. Polak, K. Eggan, and S. A. McCarroll, "Genetic variation in human dna replication timing," *Cell*, vol. 159, no. 5, pp. 1015–1026, 2014.

[21] A. Poulet, B. Li, T. Dubos, J. C. Rivera-Mulia, D. M. Gilbert, and Z. S. Qin, "Rt states: systematic annotation of the human genome using cell type-specific replication timing programs," *Bioinformatics*, vol. 35, no. 13, pp. 2167–2176, 2019.

[22] L. S. S. Murari, Q. Dickinson, S. Meyer-Nava, and J. C. Rivera-Mulia, "Partage: Parallel analysis of replication timing and gene expression," *bioRxiv*, 2025.

[23] N. J. Haradhvala, P. Polak, P. Stojanov, K. R. Covington, E. Shinbrot, J. M. Hess, E. Rheinbay, J. Kim, Y. E. Maruvka, L. Z. Braunstein, *et al.*, "Mutational strand asymmetries in cancer genomes reveal mechanisms of dna damage and repair," *Cell*, vol. 164, no. 3, pp. 538–549, 2016.

[24] S. A. Forbes, D. Beare, P. Gunasekaran, K. Leung, N. Bindal, H. Boutselakis, M. Ding, S. Bamford, C. Cole, S. Ward, *et al.*, "Cosmic: exploring the world's knowledge of somatic

mutations in human cancer," *Nucleic acids research*, vol. 43, no. D1, pp. D805–D811, 2015.

[25] G. Van Rossum *et al.*, "Python programming language.," in *USENIX ATC*, 2007.

[26] G. Perez, G. P. Barber, A. Benet-Pages, J. Casper, H. Clawson, M. Diekhans, C. Fischer, J. N. Gonzalez, A. S. Hinrichs, C. M. Lee, *et al.*, "The ucsc genome browser database: 2025 update," *Nucleic acids research*, vol. 53, no. D1, pp. D1243–D1249, 2025.

[27] L. B. Alexandrov, J. Kim, N. J. Haradhvala, M. N. Huang, A. W. Tian Ng, Y. Wu, A. Boot, K. R. Covington, D. A. Gordenin, E. N. Bergstrom, *et al.*, "The repertoire of mutational signatures in human cancer," *Nature*, vol. 578, no. 7793, pp. 94–101, 2020.

[28] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[29] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*, vol. 4. Springer, 2006.

[30] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[31] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

[32] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[33] T. Hastie, R. Tibshirani, J. Friedman, *et al.*, "The elements of statistical learning," 2009.

[34] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to linear regression analysis*. John Wiley & Sons, 2021.

[35] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 807–814, 2010.

[36] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[37] K. Pearson, "Vii. mathematical contributions to the theory of evolution.—iii. regression, heredity, and panmixia," *Philosophical Transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, no. 187, pp. 253–318, 1896.

[38] Z. Fang, "Gseapy: gene set enrichment analysis in python," *Zenodo*, 2020.

[39] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, *et al.*, "Gene ontology: tool for the unification of biology," *Nature genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[40] L. H. Wang, X. Y. Yang, X. Zhang, K. Mihalic, Y.-X. Fan, W. Xiao, O. Z. Howard, E. Appella, A. T. Maynard, and W. L. Farrar, "Suppression of breast cancer by chemical modulation of vulnerable zinc fingers in estrogen receptor," *Nature medicine*, vol. 10, no. 1, pp. 40–47, 2004.

APPENDIX A

ANCESTRAL PATHWAYS

Biological Process Ancestor Pathways

Cellular Process Early
1. Macromolecule Biosynthetic Process (GO:0009059)
2. protein-RNA Complex Assembly (GO:0022618)
3. Cytoplasmic Translation (GO:0002181)
4. Cellular Response to Cadmium Ion (GO:0071276)
5. Translation (GO:0006412)
6. Gene Expression (GO:0010467)
7. Extrinsic Apoptotic Signaling Pathway (GO:0097191)
8. Regulation of Mitotic Cell Cycle Phase Transition (GO:1901990)
9. Positive Regulation of Canonical NF-kB Signal Transduction (GO:0043123)
10. Regulation of Translation (GO:0006417)
11. Extrinsic Apoptotic Signaling Pathway via Death Domain Receptors (GO:0008625)
12. Regulation of Cell Cycle G1/S Phase Transition (GO:1902806)
13. Intermediate Filament Organization (GO:0045109)
14. Cellular Response to Copper Ion (GO:0071280)
15. RNA Splicing (GO:0008380)
16. Regulation of Cell Fate Commitment (GO:0010453)
17. RNA Processing (GO:0006396)
18. Mitochondrial Transport (GO:0006839)
19. Pyruvate Metabolic Process (GO:0006090)
20. "RNA Splicing, via Transesterification Reactions With Bulged Adenosine as Nucleophile (GO:0000377)"
21. Regulation of Canonical NF-kB Signal Transduction (GO:0043122)
22. 'De Novo' Post-Translational Protein Folding (GO:0051084)
23. Glucose Metabolic Process (GO:0006006)
24. Negative Regulation of Intrinsic Apoptotic Signaling Pathway (GO:2001243)
25. Regulation of B Cell Activation (GO:0050864)
26. Membraneless Organelle Assembly (GO:0140694)
27. Intracellular Protein Transport (GO:0006886)
28. "mRNA Splicing, via Spliceosome (GO:0000398)"
29. Positive Regulation of Triglyceride Metabolic Process (GO:0090208)

30. Supramolecular Fiber Organization (GO:0097435)
31. Apoptotic Process (GO:0006915)
32. Positive Regulation of DNA-templated Transcription (GO:0045893)

Cellular Process Late
1. Gamma-Aminobutyric Acid Signaling Pathway (GO:0007214)
2. Inhibitory Synapse Assembly (GO:1904862)
3. "Synaptic Transmission, GABAergic (GO:0051932)"
4. Chloride Transmembrane Transport (GO:1902476)
5. Synapse Assembly (GO:0007416)
6. Monoatomic Anion Transmembrane Transport (GO:0098656)
7. Synaptic Membrane Adhesion (GO:0099560)
8. Positive Regulation of Synapse Assembly (GO:0051965)
9. Synapse Organization (GO:0050808)
10. Regulation of Synapse Assembly (GO:0051963)
11. Positive Regulation of Cell Junction Assembly (GO:1901890)
12. Chemical Synaptic Transmission (GO:0007268)
13. Axonogenesis (GO:0007409)
14. Calcium-Dependent Cell-Cell Adhesion via Plasma Membrane Cell Adhesion Molecules (GO:0016339)
15. Cell-Cell Adhesion Mediated by Cadherin (GO:0044331)
16. Cell-Cell Junction Organization (GO:0045216)
17. Cell-Cell Junction Assembly (GO:0007043)
18. Adherens Junction Organization (GO:0034332)
19. Presynaptic Membrane Assembly (GO:0097105)
20. Presynaptic Membrane Organization (GO:0097090)
21. Cell Junction Assembly (GO:0034329)
22. Cell Morphogenesis Involved in Neuron Differentiation (GO:0048667)
23. Neuron Cell-Cell Adhesion (GO:0007158)
24. Neuron Projection Guidance (GO:0097485)
25. Plasminogen Activation (GO:0031639)
26. Central Nervous System Neuron Differentiation (GO:0021953)
27. Estrogen Metabolic Process (GO:0008210)
28. Regulation of Presynapse Assembly (GO:1905606)
29. Axon Development (GO:0061564)
30. Regulation of Presynapse Organization (GO:0099174)
31. Neuron Projection Morphogenesis (GO:0048812)
32. Cellular Response to Nitrogen Compound (GO:1901699)
33. Axon Guidance (GO:0007411)
34. Plasma Membrane Phospholipid Scrambling (GO:0017121)
35. Steroid Metabolic Process (GO:0008202)
36. Cellular Response to Histamine (GO:0071420)
37. Glutamate Receptor Signaling Pathway (GO:0007215)
38. Presynapse Assembly (GO:0099054)
39. Positive Regulation of Hormone Secretion (GO:0046887)
40. Neuron Differentiation (GO:0030182)

41. Signaling Receptor Ligand Precursor Processing (GO:0140448)

Response to Stimulus Early
1. Response to Cadmium Ion (GO:0046686)
2. Cellular Response to Cadmium Ion (GO:0071276)
3. Cellular Response to Copper Ion (GO:0071280)
4. Extrinsic Apoptotic Signaling Pathway via Death Domain Receptors (GO:0008625)
5. Extrinsic Apoptotic Signaling Pathway (GO:0097191)
6. Positive Regulation of Canonical NF-kB Signal Transduction (GO:0043123)
7. Negative Regulation of Intrinsic Apoptotic Signaling Pathway (GO:2001243)
8. Regulation of Canonical NF-kappaB Signal Transduction (GO:0043122)
9. Response to Copper Ion (GO:0046688)

Response to Stimulus Late
1. Detection of Chemical Stimulus Involved in Sensory Perception (GO:0050907)
2. Detection of Chemical Stimulus Involved in Sensory Perception of Smell (GO:0050911)
3. Cellular Response to Nitrogen Compound (GO:1901699)
4. Cellular Response to Histamine (GO:0071420)
5. Glutamate Receptor Signaling Pathway (GO:0007215)
6. Antimicrobial Humoral Response (GO:0019730)
7. Antimicrobial Humoral Immune Response Mediated by Antimicrobial Peptide (GO:0061844)

Biological Regulation Early
1. Regulation of Cell Cycle G1/S Phase Transition (GO:1902806)
2. Extrinsic Apoptotic Signaling Pathway via Death Domain Receptors (GO:0008625)
3. Extrinsic Apoptotic Signaling Pathway (GO:0097191)
4. Regulation of Translation (GO:0006417)
5. Positive Regulation of Canonical NF-kB Signal Transduction (GO:0043123)
6. Regulation of Mitotic Cell Cycle Phase Transition (GO:1901990)
7. Negative Regulation of Growth (GO:0045926)
8. Positive Regulation of Triglyceride Metabolic Process (GO:0090208)
9. Regulation of B Cell Activation (GO:0050864)
10. Negative Regulation of Intrinsic Apoptotic Signaling Pathway (GO:2001243)
11. Regulation of Canonical NF-kB Signal Transduction (GO:0043122)
12. Regulation of Cell Fate Commitment (GO:0010453)
13. Positive Regulation of DNA-templated Transcription (GO:0045893)

Biological Regulation Late
1. Gamma-Aminobutyric Acid Signaling Pathway (GO:0007214)
2. "Synaptic Transmission, GABAergic (GO:0051932)"
3. Regulation of Synapse Assembly (GO:0051963)
4. Positive Regulation of Synapse Assembly (GO:0051965)
5. Positive Regulation of Nervous System Development (GO:0051962)
6. Positive Regulation of Cell Junction Assembly (GO:1901890)

7. Chemical Synaptic Transmission (GO:0007268)
8. Regulation of Presynapse Organization (GO:0099174)
9. Regulation of Presynapse Assembly (GO:1905606)
10. Estrogen Metabolic Process (GO:0008210)
11. Plasma Membrane Phospholipid Scrambling (GO:0017121)
12. Glutamate Receptor Signaling Pathway (GO:0007215)
13. Positive Regulation of Hormone Secretion (GO:0046887)

Homeostatic Process Early
1. Intracellular Zinc Ion Homeostasis (GO:0006882)

Localization Early
1. Phospholipid Efflux (GO:0033700)
2. Intracellular Protein Transport (GO:0006886)
3. Mitochondrial Transport (GO:0006839)
4. Cholesterol Efflux (GO:0033344)

Localization Late
1. Chloride Transmembrane Transport (GO:1902476)
2. Monoatomic Anion Transmembrane Transport (GO:0098656)
3. Chloride Transport (GO:0006821)
4. Plasma Membrane Phospholipid Scrambling (GO:0017121)
5. Oxygen Transport (GO:0015671)
6. One-Carbon Compound Transport (GO:0019755)
7. Gas Transport (GO:0015669)
8. Carbon Dioxide Transport (GO:0015670)
9. Sodium-Independent Organic Anion Transport (GO:0043252)
10. Positive Regulation of Hormone Secretion (GO:0046887)

Multicellular Organismal Process Early
1. Regulation of B Cell Activation (GO:0050864)

Multicellular Organismal Process Late
1. Sensory Perception of Smell (GO:0007608)
2. Sensory Perception of Chemical Stimulus (GO:0007606)
3. Inhibitory Synapse Assembly (GO:1904862)
4. Synapse Assembly (GO:0007416)
5. Regulation of Synapse Assembly (GO:0051963)
6. Detection of Chemical Stimulus Involved in Sensory Perception (GO:0050907)
7. Detection of Chemical Stimulus Involved in Sensory Perception of Smell (GO:0050911)
8. Positive Regulation of Synapse Assembly (GO:0051965)
9. Positive Regulation of Nervous System Development (GO:0051962)
10. Axonogenesis (GO:0007409)

11. Presynaptic Membrane Assembly (GO:0097105)
12. Cell Morphogenesis Involved in Neuron Differentiation (GO:0048667)
13. Neuron Projection Morphogenesis (GO:0048812)
14. Axon Development (GO:0061564)
15. Regulation of Presynapse Assembly (GO:1905606)
16. Central Nervous System Neuron Differentiation (GO:0021953)
17. Central Nervous System Development (GO:0007417)
18. Nervous System Development (GO:0007399)
19. Neuron Projection Guidance (GO:0097485)
20. Axon Guidance (GO:0007411)
21. Brain Development (GO:0007420)
22. Presynapse Assembly (GO:0099054)
23. Neuron Differentiation (GO:0030182)

Developmental Process Early
1. Regulation of Cell Fate Commitment (GO:0010453)

Developmental Process Late
1. Inhibitory Synapse Assembly (GO:1904862)
2. Synapse Assembly (GO:0007416)
3. Regulation of Synapse Assembly (GO:0051963)
4. Positive Regulation of Synapse Assembly (GO:0051965)
5. Positive Regulation of Nervous System Development (GO:0051962)
6. Axonogenesis (GO:0007409)
7. Cell Morphogenesis Involved in Neuron Differentiation (GO:0048667)
8. Presynaptic Membrane Assembly (GO:0097105)
9. Central Nervous System Development (GO:0007417)
10. Neuron Projection Guidance (GO:0097485)
11. Nervous System Development (GO:0007399)
12. Axon Development (GO:0061564)
13. Central Nervous System Neuron Differentiation (GO:0021953)
14. Regulation of Presynapse Assembly (GO:1905606)
15. Neuron Projection Morphogenesis (GO:0048812)
16. Axon Guidance (GO:0007411)
17. Brain Development (GO:0007420)
18. Presynapse Assembly (GO:0099054)
19. Neuron Differentiation (GO:0030182)

Biological Process Involved in Interspecies Interaction Between Late
1. Antimicrobial Humoral Response (GO:0019730)
2. Antimicrobial Humoral Immune Response Mediated by Antimicrobial Peptide (GO:0061844)

Immune System Process Late
1. Antimicrobial Humoral Response (GO:0019730)

2. Antimicrobial Humoral Immune Response Mediated by Antimicrobial Peptide (GO:0061844)

Cellular Component

Protein-Containing Complex Early
1. Cytosolic Large Ribosomal Subunit (GO:0022625)
2. Large Ribosomal Subunit (GO:0015934)
3. High-Density Lipoprotein Particle (GO:0034364)
4. Anaphase-Promoting Complex (GO:0005680)
5. "RNA Polymerase II, Core Complex (GO:0005665)"
6. Small Ribosomal Subunit (GO:0015935)
7. Cytosolic Small Ribosomal Subunit (GO:0022627)
8. Triglyceride-Rich Plasma Lipoprotein Particle (GO:0034385)
9. Very-Low-Density Lipoprotein Particle (GO:0034361)
10. RNA Polymerase III Complex (GO:0005666)
11. Chylomicron (GO:0042627)
12. U2-type Precatalytic Spliceosome (GO:0071005)
13. Precatalytic Spliceosome (GO:0071011)
14. Spliceosomal snRNP Complex (GO:0097525)

Protein-Containing Complex Late
1. GABA-A Receptor Complex (GO:1902711)
2. Catenin Complex (GO:0016342)
3. Cul4-RING E3 Ubiquitin Ligase Complex (GO:0080008)
4. Dystroglycan Complex (GO:0016011)
5. Sarcoglycan Complex (GO:0016012)
6. Cul2-RING Ubiquitin Ligase Complex (GO:0031462)
7. cullin-RING Ubiquitin Ligase Complex (GO:0031461)

Cellular Anatomical Structure Early
1. Focal Adhesion (GO:0005925)
2. Cell-Substrate Junction (GO:0030055)
3. Cytosolic Large Ribosomal Subunit (GO:0022625)
4. Large Ribosomal Subunit (GO:0015934)
5. Intermediate Filament (GO:0005882)
6. Lysosomal Membrane (GO:0005765)
7. Keratin Filament (GO:0045095)
8. Mitochondrial Membrane (GO:0031966)
9. High-Density Lipoprotein Particle (GO:0034364)
10. Mitochondrial Inner Membrane (GO:0005743)
11. Cytosolic Small Ribosomal Subunit (GO:0022627)
12. Vesicle (GO:0031982)
13. Triglyceride-Rich Plasma Lipoprotein Particle (GO:0034385)
14. Very-Low-Density Lipoprotein Particle (GO:0034361)

15. Small Ribosomal Subunit (GO:0015935)
16. Anaphase-Promoting Complex (GO:0005680)
17. Ribosome (GO:0005840)
18. "RNA Polymerase II, Core Complex (GO:0005665)"
19. Actin Cytoskeleton (GO:0015629)
20. Spliceosomal snRNP Complex (GO:0097525)
21. Precatalytic Spliceosome (GO:0071011)
22. U2-type Precatalytic Spliceosome (GO:0071005)
23. Intracellular Membrane-Bounded Organelle (GO:0043231)
24. Chylomicron (GO:0042627)
25. Nucleus (GO:0005634)
26. Endocytic Vesicle (GO:0030139)
27. Cortical Actin Cytoskeleton (GO:0030864)
28. RNA Polymerase III Complex (GO:0005666)
29. Cytoplasmic Vesicle Membrane (GO:0030659)
30. Multivesicular Body Lumen (GO:0097486)
31. Mitotic Spindle Microtubule (GO:1990498)
32. Lytic Vacuole Membrane (GO:0098852)

Cellular Anatomical Structure Late
1. Dendrite Membrane (GO:0032590)
2. Dendrite (GO:0030425)
3. Postsynaptic Specialization Membrane (GO:0099634)
4. GABAergic Synapse (GO:0098982)
5. Synaptic Membrane (GO:0097060)
6. Neuron Projection (GO:0043005)
7. Glutamatergic Synapse (GO:0098978)
8. Catenin Complex (GO:0016342)
9. Adherens Junction (GO:0005912)
10. Cornified Envelope (GO:0001533)
11. Postsynaptic Density Membrane (GO:0098839)
12. Dystroglycan Complex (GO:0016011)
13. Sarcoglycan Complex (GO:0016012)
14. Desmosome (GO:0030057)
15. Cell-Cell Junction (GO:0005911)
16. Axonal Growth Cone (GO:0044295)
17. Astrocyte End-Foot (GO:0097450)
18. Axon (GO:0030424)

Molecular Function

Binding Early
1. Ubiquitin-Like Protein Ligase Binding (GO:0044389)
2. Ubiquitin Protein Ligase Binding (GO:0031625)
3. Cadherin Binding (GO:0045296)

4. mRNA Binding (GO:0003729)

Binding Late
1. Chloride Ion Binding (GO:0031404)
2. Peptidoglycan Binding (GO:0042834)
3. CARD Domain Binding (GO:0050700)
4. CXCR Chemokine Receptor Binding (GO:0045236)
5. Poly-Purine Tract Binding (GO:0070717)
6. poly(A) Binding (GO:0008143)
7. Chemokine Activity (GO:0008009)
8. Epidermal Growth Factor Receptor Binding (GO:0005154)
9. Retinoic Acid Binding (GO:0001972)

Transporter Activity Late
1. GABA-gated Chloride Ion Channel Activity (GO:0022851)
2. Ligand-Gated Monoatomic Anion Channel Activity (GO:0099095)
3. Transmitter-Gated Monoatomic Ion Channel Activity (GO:0022824)
4. Chloride Channel Activity (GO:0005254)
5. Monoatomic Anion Channel Activity (GO:0005253)
6. Chloride Transmembrane Transporter Activity (GO:0015108)
7. Ligand-Gated Monoatomic Ion Channel Activity (GO:0015276)
8. Phospholipid Scramblase Activity (GO:0017128)

Molecular Transducer Activity Late
1. Olfactory Receptor Activity (GO:0004984)
2. GABA-gated Chloride Ion Channel Activity (GO:0022851)
3. GABA-A Receptor Activity (GO:0004890)
4. Benzodiazepine Receptor Activity (GO:0008503)
5. GABA Receptor Activity (GO:0016917)
6. Transmitter-Gated Monoatomic Ion Channel Activity (GO:0022824)
7. Transmembrane-Ephrin Receptor Activity (GO:0005005)
8. Ephrin Receptor Activity (GO:0005003)
9. Glutamate Receptor Activity (GO:0008066)
10. Chemokine Activity (GO:0008009)
11. G Protein-Coupled Neurotransmitter Receptor Activity (GO:0099528)

Molecular Function Regulator Activity Late
1. Chemokine Activity (GO:0008009)
2. Cysteine-Type Endopeptidase Inhibitor Activity (GO:0004869)

Catalytic Activity Late
1. Glucuronosyltransferase Activity (GO:0015020)
2. Transmembrane-Ephrin Receptor Activity (GO:0005005)

3. Ephrin Receptor Activity (GO:0005003)
4. Hexosyltransferase Activity (GO:0016758)
5. UDP-glycosyltransferase Activity (GO:0008194)
6. Cysteine-Type Endopeptidase Inhibitor Activity (GO:0004869)

APPENDIX B

CELL-LINE SPECIFIC PATHWAYS

BG02ES

Biological Process Early
1. Macromolecule Biosynthetic Process (GO:0009059)
2. Response to Cadmium Ion (GO:0046686)
3. protein-RNA Complex Assembly (GO:0022618)
4. Cytoplasmic Translation (GO:0002181)
5. Cellular Response to Cadmium Ion (GO:0071276)
6. Chaperone Cofactor-Dependent Protein Refolding (GO:0051085)
7. Translation (GO:0006412)
8. Gene Expression (GO:0010467)
9. Regulation of Translation (GO:0006417)
10. Negative Regulation of Growth (GO:0045926)
11. mRNA Splicing, via Spliceosome (GO:0000398)
12. RNA Splicing (GO:0008380)
13. RNA Splicing, via Transesterification Reactions With Bulged Adenosine as Nucleophile (GO:0000377)
14. RNA Processing (GO:0006396)
15. Regulation of Cell Fate Commitment (GO:0010453)
16. Positive Regulation of DNA-templated Transcription (GO:0045893)

Biological Process Late
1. Sensory Perception of Smell (GO:0007608)
2. Sensory Perception of Chemical Stimulus (GO:0007606)
3. Gamma-Aminobutyric Acid Signaling Pathway (GO:0007214)
4. Synaptic Transmission, GABAergic (GO:0051932)
5. Inhibitory Synapse Assembly (GO:1904862)
6. Chloride Transmembrane Transport (GO:1902476)
7. Synapse Assembly (GO:0007416)
8. Monoatomic Anion Transmembrane Transport (GO:0098656)
9. Inorganic Anion Transmembrane Transport (GO:0098661)
10. Detection of Chemical Stimulus Involved in Sensory Perception (GO:0050907)
11. Detection of Chemical Stimulus Involved in Sensory Perception of Smell (GO:0050911)

12. Chloride Transport (GO:0006821)

Cellular Component Early
1. Focal Adhesion (GO:0005925)
2. Cytosolic Large Ribosomal Subunit (GO:0022625)
3. Large Ribosomal Subunit (GO:0015934)
4. Mitochondrial Membrane (GO:0031966)
5. Mitochondrial Inner Membrane (GO:0005743)
6. Ribosome (GO:0005840)
7. Small Ribosomal Subunit (GO:0015935)
8. Cytosolic Small Ribosomal Subunit (GO:0022627)
9. Nucleus (GO:0005634)
10. Intracellular Membrane-Bounded Organelle (GO:0043231)
11. Spliceosomal snRNP Complex (GO:0097525)

Cellular Component Late
1. GABA-A Receptor Complex (GO:1902711)
2. Dendrite Membrane (GO:0032590)
3. Cul2-RING Ubiquitin Ligase Complex (GO:0031462)

Molecular Function Early
1. Ubiquitin-Like Protein Ligase Binding (GO:0044389)
2. Ubiquitin Protein Ligase Binding (GO:0031625)

Molecular Function Late
1. Olfactory Receptor Activity (GO:0004984)
2. GABA-gated Chloride Ion Channel Activity (GO:0022851)
3. GABA-A Receptor Activity (GO:0004890)
4. Benzodiazepine Receptor Activity (GO:0008503)
5. GABA Receptor Activity (GO:0016917)
6. Ligand-Gated Monoatomic Anion Channel Activity (GO:0099095)
7. Transmitter-Gated Monoatomic Ion Channel Activity (GO:0022824)
8. Chloride Channel Activity (GO:0005254)
9. Monoatomic Anion Channel Activity (GO:0005253)
10. Chloride Transmembrane Transporter Activity (GO:0015108)
11. Chloride Ion Binding (GO:0031404)
12. Glucuronosyltransferase Activity (GO:0015020)

BJ

Biological Process Early
1. Response to Cadmium Ion (GO:0046686)

2. Cellular Response to Cadmium Ion (GO:0071276)
3. Cellular Response to Copper Ion (GO:0071280)
4. Extrinsic Apoptotic Signaling Pathway via Death Domain Receptors (GO:0008625)
5. Extrinsic Apoptotic Signaling Pathway (GO:0097191)
6. Positive Regulation of Canonical NF-kappaB Signal Transduction (GO:0043123)
7. Intracellular Zinc Ion Homeostasis (GO:0006882)
8. Regulation of Mitotic Cell Cycle Phase Transition (GO:1901990)
9. Mitochondrial Transport (GO:0006839)
10. Response to Copper Ion (GO:0046688)

Biological Process Late
1. Sensory Perception of Smell (GO:0007608)
2. Sensory Perception of Chemical Stimulus (GO:0007606)
3. Gamma-Aminobutyric Acid Signaling Pathway (GO:0007214)
4. Synaptic Transmission, GABAergic (GO:0051932)
5. Inhibitory Synapse Assembly (GO:1904862)
6. Chloride Transmembrane Transport (GO:1902476)
7. Synapse Assembly (GO:0007416)
8. Monoatomic Anion Transmembrane Transport (GO:0098656)
9. Inorganic Anion Transmembrane Transport (GO:0098661)
10. Synaptic Membrane Adhesion (GO:0099560)
11. Regulation of Synapse Assembly (GO:0051963)
12. Detection of Chemical Stimulus Involved in Sensory Perception (GO:0050907)
13. Detection of Chemical Stimulus Involved in Sensory Perception of Smell (GO:0050911)
14. Cell-Cell Adhesion via Plasma-Membrane Adhesion Molecules (GO:0098742)
15. Positive Regulation of Synapse Assembly (GO:0051965)
16. Synapse Organization (GO:0050808)
17. Positive Regulation of Nervous System Development (GO:0051962)
18. Positive Regulation of Cell Junction Assembly (GO:1901890)
19. Calcium-Dependent Cell-Cell Adhesion via Plasma Membrane Cell Adhesion Molecules (GO:0016339)
20. Cell-Cell Adhesion Mediated by Cadherin (GO:0044331)
21. Cell-Cell Junction Organization (GO:0045216)
22. Cell-Cell Junction Assembly (GO:0007043)
23. Adherens Junction Organization (GO:0034332)
24. Plasminogen Activation (GO:0031639)
25. Estrogen Metabolic Process (GO:0008210)
26. Positive Regulation of Hormone Secretion (GO:0046887)

Cellular Component Early
1. Focal Adhesion (GO:0005925)
2. Cell-Substrate Junction (GO:0030055)

Cellular Component Late
1. GABA-A Receptor Complex (GO:1902711)

2. Dendrite Membrane (GO:0032590)
3. Postsynaptic Specialization Membrane (GO:0099634)
4. Dendrite (GO:0030425)
5. GABA-ergic Synapse (GO:0098982)
6. Synaptic Membrane (GO:0097060)
7. Glutamatergic Synapse (GO:0098978)
8. Catenin Complex (GO:0016342)
9. Desmosome (GO:0030057)
10. Cul2-RING Ubiquitin Ligase Complex (GO:0031462)
11. Cornified Envelope (GO:0001533)
12. cullin-RING Ubiquitin Ligase Complex (GO:0031461)

Molecular Function Early
1. Ubiquitin-Like Protein Ligase Binding (GO:0044389)
2. Ubiquitin Protein Ligase Binding (GO:0031625)
3. Cadherin Binding (GO:0045296)

Molecular Function Late
1. Olfactory Receptor Activity (GO:0004984)
2. GABA-gated Chloride Ion Channel Activity (GO:0022851)
3. GABA-A Receptor Activity (GO:0004890)
4. Benzodiazepine Receptor Activity (GO:0008503)
5. GABA Receptor Activity (GO:0016917)
6. Ligand-Gated Monoatomic Anion Channel Activity (GO:0099095)
7. Transmitter-Gated Monoatomic Ion Channel Activity (GO:0022824)
8. Chloride Channel Activity (GO:0005254)
9. Monoatomic Anion Channel Activity (GO:0005253)
10. Chloride Transmembrane Transporter Activity (GO:0015108)
11. Chloride Ion Binding (GO:0031404)
12. Peptidoglycan Binding (GO:0042834)
13. Glucuronosyltransferase Activity (GO:0015020)
14. Hexosyltransferase Activity (GO:0016758)
15. UDP-glycosyltransferase Activity (GO:0008194)

GM12878

Biological Process Early
1. protein-RNA Complex Assembly (GO:0022618)
2. Regulation of Cell Cycle G1/S Phase Transition (GO:1902806)
3. Regulation of Mitotic Cell Cycle Phase Transition (GO:1901990)
4. Pyruvate Metabolic Process (GO:0006090)

Biological Process Late
1. Sensory Perception of Smell (GO:0007608)
2. Sensory Perception of Chemical Stimulus (GO:0007606)
3. Gamma-Aminobutyric Acid Signaling Pathway (GO:0007214)
4. Synaptic Transmission, GABAergic (GO:0051932)
5. Inhibitory Synapse Assembly (GO:1904862)
6. Chloride Transmembrane Transport (GO:1902476)
7. Synapse Assembly (GO:0007416)
8. Monoatomic Anion Transmembrane Transport (GO:0098656)
9. Synaptic Membrane Adhesion (GO:0099560)
10. Regulation of Synapse Assembly (GO:0051963)
11. Cell-Cell Adhesion via Plasma-Membrane Adhesion Molecules (GO:0098742)
12. Positive Regulation of Synapse Assembly (GO:0051965)
13. Synapse Organization (GO:0050808)
14. Positive Regulation of Nervous System Development (GO:0051962)
15. Positive Regulation of Cell Junction Assembly (GO:1901890)
16. Chemical Synaptic Transmission (GO:0007268)
17. Axonogenesis (GO:0007409)
18. Calcium-Dependent Cell-Cell Adhesion via Plasma Membrane Cell Adhesion Molecules (GO:0016339)
19. Cell-Cell Adhesion Mediated by Cadherin (GO:0044331)
20. Cell-Cell Junction Organization (GO:0045216)
21. Cell-Cell Junction Assembly (GO:0007043)
22. Adherens Junction Organization (GO:0034332)
23. Cell Junction Assembly (GO:0034329)
24. Presynaptic Membrane Organization (GO:0097090)
25. Presynaptic Membrane Assembly (GO:0097105)
26. Neuron Cell-Cell Adhesion (GO:0007158)
27. Central Nervous System Neuron Differentiation (GO:0021953)
28. Central Nervous System Development (GO:0007417)
29. Nervous System Development (GO:0007399)
30. Neuron Projection Guidance (GO:0097485)
31. Axon Guidance (GO:0007411)
32. Neuron Differentiation (GO:0030182)

Cellular Component Early
1. Focal Adhesion (GO:0005925)
2. Cell-Substrate Junction (GO:0030055)
3. RNA Polymerase II, Core Complex (GO:0005665)

Cellular Component Late
1. GABA-A Receptor Complex (GO:1902711)
2. Dendrite Membrane (GO:0032590)
3. Postsynaptic Specialization Membrane (GO:0099634)
4. Dendrite (GO:0030425)

5. GABA-ergic Synapse (GO:0098982)
6. Neuron Projection (GO:0043005)
7. Glutamatergic Synapse (GO:0098978)
8. Catenin Complex (GO:0016342)
9. Cul4-RING E3 Ubiquitin Ligase Complex (GO:0080008)
10. Adherens Junction (GO:0005912)
11. Postsynaptic Density Membrane (GO:0098839)

Molecular Function Early
1. Ubiquitin-Like Protein Ligase Binding (GO:0044389)
2. Ubiquitin Protein Ligase Binding (GO:0031625)
3. mRNA Binding (GO:0003729)

Molecular Function Late
1. Olfactory Receptor Activity (GO:0004984)
2. GABA-gated Chloride Ion Channel Activity (GO:0022851)
3. GABA-A Receptor Activity (GO:0004890)
4. Benzodiazepine Receptor Activity (GO:0008503)
5. GABA Receptor Activity (GO:0016917)
6. Ligand-Gated Monoatomic Anion Channel Activity (GO:0099095)
7. Transmitter-Gated Monoatomic Ion Channel Activity (GO:0022824)
8. Chloride Channel Activity (GO:0005254)
9. Transmembrane-Ephrin Receptor Activity (GO:0005005)
10. Ephrin Receptor Activity (GO:0005003)

HELAS3

Biological Process Early
1. Macromolecule Biosynthetic Process (GO:0009059)
2. protein-RNA Complex Assembly (GO:0022618)
3. Cytoplasmic Translation (GO:0002181)
4. Chaperone Cofactor-Dependent Protein Refolding (GO:0051085)
5. Regulation of Cell Cycle G1/S Phase Transition (GO:1902806)
6. Extrinsic Apoptotic Signaling Pathway via Death Domain Receptors (GO:0008625)
7. Extrinsic Apoptotic Signaling Pathway (GO:0097191)
8. Positive Regulation of Canonical NF-kappaB Signal Transduction (GO:0043123)
9. Apoptotic Process (GO:0006915)
10. Negative Regulation of Transferase Activity (GO:0051348)
11. 'De Novo' Post-Translational Protein Folding (GO:0051084)
12. Regulation of Canonical NF-kappaB Signal Transduction (GO:0043122)

Biological Process Late
1. Sensory Perception of Smell (GO:0007608)

2. Sensory Perception of Chemical Stimulus (GO:0007606)
3. Gamma-Aminobutyric Acid Signaling Pathway (GO:0007214)
4. Synaptic Transmission, GABAergic (GO:0051932)
5. Inhibitory Synapse Assembly (GO:1904862)
6. Chloride Transmembrane Transport (GO:1902476)
7. Synapse Assembly (GO:0007416)
8. Monoatomic Anion Transmembrane Transport (GO:0098656)
9. Inorganic Anion Transmembrane Transport (GO:0098661)
10. Detection of Chemical Stimulus Involved in Sensory Perception (GO:0050907)
11. Detection of Chemical Stimulus Involved in Sensory Perception of Smell (GO:0050911)

Cellular Component Early
1. Focal Adhesion (GO:0005925)
2. Cell-Substrate Junction (GO:0030055)
3. Cytosolic Large Ribosomal Subunit (GO:0022625)
4. Large Ribosomal Subunit (GO:0015934)
5. Lysosomal Membrane (GO:0005765)
6. U2-type Precatalytic Spliceosome (GO:0071005)
7. Precatalytic Spliceosome (GO:0071011)

Cellular Component Late
1. GABA-A Receptor Complex (GO:1902711)
2. Dendrite Membrane (GO:0032590)
3. Postsynaptic Specialization Membrane (GO:0099634)
4. Dendrite (GO:0030425)
5. GABA-ergic Synapse (GO:0098982)
6. Synaptic Membrane (GO:0097060)

Molecular Function Early
1. Ubiquitin-Like Protein Ligase Binding (GO:0044389)
2. Ubiquitin Protein Ligase Binding (GO:0031625)
3. Cadherin Binding (GO:0045296)

Molecular Function Late
1. Olfactory Receptor Activity (GO:0004984)
2. GABA-gated Chloride Ion Channel Activity (GO:0022851)
3. GABA-A Receptor Activity (GO:0004890)
4. Benzodiazepine Receptor Activity (GO:0008503)
5. GABA Receptor Activity (GO:0016917)
6. Ligand-Gated Monoatomic Anion Channel Activity (GO:0099095)
7. Transmitter-Gated Monoatomic Ion Channel Activity (GO:0022824)
8. Chloride Channel Activity (GO:0005254)
9. Monoatomic Anion Channel Activity (GO:0005253)
10. Chloride Transmembrane Transporter Activity (GO:0015108)

11. Peptidoglycan Binding (GO:0042834)

HEPG2

Biological Process Early
1. Macromolecule Biosynthetic Process (GO:0009059)
2. Cytoplasmic Translation (GO:0002181)
3. Translation (GO:0006412)

Biological Process Late
1. Sensory Perception of Smell (GO:0007608)
2. Sensory Perception of Chemical Stimulus (GO:0007606)
3. Gamma-Aminobutyric Acid Signaling Pathway (GO:0007214)
4. Synaptic Transmission, GABAergic (GO:0051932)
5. Inhibitory Synapse Assembly (GO:1904862)
6. Chloride Transmembrane Transport (GO:1902476)
7. Synapse Assembly (GO:0007416)
8. Monoatomic Anion Transmembrane Transport (GO:0098656)
9. Inorganic Anion Transmembrane Transport (GO:0098661)
10. Synaptic Membrane Adhesion (GO:0099560)
11. Regulation of Synapse Assembly (GO:0051963)
12. Detection of Chemical Stimulus Involved in Sensory Perception (GO:0050907)
13. Detection of Chemical Stimulus Involved in Sensory Perception of Smell (GO:0050911)
14. Cell-Cell Adhesion via Plasma-Membrane Adhesion Molecules (GO:0098742)
15. Positive Regulation of Synapse Assembly (GO:0051965)
16. Synapse Organization (GO:0050808)
17. Positive Regulation of Nervous System Development (GO:0051962)
18. Positive Regulation of Cell Junction Assembly (GO:1901890)
19. Chemical Synaptic Transmission (GO:0007268)
20. Axonogenesis (GO:0007409)
21. Chloride Transport (GO:0006821)
22. Calcium-Dependent Cell-Cell Adhesion via Plasma Membrane Cell Adhesion Molecules (GO:0016339)
23. Cell-Cell Adhesion Mediated by Cadherin (GO:0044331)
24. Cell-Cell Junction Organization (GO:0045216)
25. Cell-Cell Junction Assembly (GO:0007043)
26. Adherens Junction Organization (GO:0034332)
27. Cell Junction Assembly (GO:0034329)
28. Cellular Response to Nitrogen Compound (GO:1901699)
29. Neuron Projection Morphogenesis (GO:0048812)
30. Central Nervous System Neuron Differentiation (GO:0021953)
31. Neuron Projection Guidance (GO:0097485)

Cellular Component Early
1. Mitochondrial Membrane (GO:0031966)
2. High-Density Lipoprotein Particle (GO:0034364)
3. Mitochondrial Inner Membrane (GO:0005743)
4. Triglyceride-Rich Plasma Lipoprotein Particle (GO:0034385)
5. Very-Low-Density Lipoprotein Particle (GO:0034361)
6. Chylomicron (GO:0042627)

Cellular Component Late
1. GABA-A Receptor Complex (GO:1902711)
2. Dendrite Membrane (GO:0032590)
3. Postsynaptic Specialization Membrane (GO:0099634)
4. Dendrite (GO:0030425)
5. GABA-ergic Synapse (GO:0098982)
6. Synaptic Membrane (GO:0097060)
7. Neuron Projection (GO:0043005)
8. Glutamatergic Synapse (GO:0098978)
9. Catenin Complex (GO:0016342)

Molecular Function Early
1. Ubiquitin-Like Protein Ligase Binding (GO:0044389)
2. Ubiquitin Protein Ligase Binding (GO:0031625)
3. mRNA Binding (GO:0003729)

Molecular Function Late
1. Olfactory Receptor Activity (GO:0004984)
2. GABA-gated Chloride Ion Channel Activity (GO:0022851)
3. GABA-A Receptor Activity (GO:0004890)
4. Benzodiazepine Receptor Activity (GO:0008503)
5. GABA Receptor Activity (GO:0016917)
6. Ligand-Gated Monoatomic Anion Channel Activity (GO:0099095)
7. Transmitter-Gated Monoatomic Ion Channel Activity (GO:0022824)
8. Chloride Channel Activity (GO:0005254)
9. Monoatomic Anion Channel Activity (GO:0005253)
10. Chloride Transmembrane Transporter Activity (GO:0015108)
11. Transmembrane-Ephrin Receptor Activity (GO:0005005)
12. Ephrin Receptor Activity (GO:0005003)

IMR90

Biological Process Early
1. Chaperone Cofactor-Dependent Protein Refolding (GO:0051085)

2. Negative Regulation of Intrinsic Apoptotic Signaling Pathway (GO:2001243)

Biological Process Late
1. Sensory Perception of Smell (GO:0007608)
2. Sensory Perception of Chemical Stimulus (GO:0007606)
3. Gamma-Aminobutyric Acid Signaling Pathway (GO:0007214)
4. Synaptic Transmission, GABAergic (GO:0051932)
5. Inhibitory Synapse Assembly (GO:1904862)
6. Chloride Transmembrane Transport (GO:1902476)
7. Synapse Assembly (GO:0007416)
8. Monoatomic Anion Transmembrane Transport (GO:0098656)
9. Inorganic Anion Transmembrane Transport (GO:0098661)
10. Synaptic Membrane Adhesion (GO:0099560)
11. Regulation of Synapse Assembly (GO:0051963)
12. Cell-Cell Adhesion via Plasma-Membrane Adhesion Molecules (GO:0098742)
13. Positive Regulation of Synapse Assembly (GO:0051965)
14. Synapse Organization (GO:0050808)
15. Positive Regulation of Nervous System Development (GO:0051962)
16. Positive Regulation of Cell Junction Assembly (GO:1901890)
17. Chemical Synaptic Transmission (GO:0007268)
18. Chloride Transport (GO:0006821)
19. Cell Morphogenesis Involved in Neuron Differentiation (GO:0048667)
20. Axon Development (GO:0061564)
21. Plasminogen Activation (GO:0031639)

Cellular Component Early
1. Focal Adhesion (GO:0005925)
2. Cell-Substrate Junction (GO:0030055)
3. Cytosolic Large Ribosomal Subunit (GO:0022625)
4. Large Ribosomal Subunit (GO:0015934)
5. Anaphase-Promoting Complex (GO:0005680)
6. Vesicle (GO:0031982)
7. Cortical Actin Cytoskeleton (GO:0030864)
8. Endocytic Vesicle (GO:0030139)
9. Actin Cytoskeleton (GO:0015629)

Cellular Component Late
1. GABA-A Receptor Complex (GO:1902711)
2. Dendrite Membrane (GO:0032590)
3. Postsynaptic Specialization Membrane (GO:0099634)
4. Dendrite (GO:0030425)
5. GABA-ergic Synapse (GO:0098982)
6. Synaptic Membrane (GO:0097060)
7. Glutamatergic Synapse (GO:0098978)
8. Desmosome (GO:0030057)

9. Cornified Envelope (GO:0001533)

Molecular Function Early
1. Ubiquitin-Like Protein Ligase Binding (GO:0044389)
2. Ubiquitin Protein Ligase Binding (GO:0031625)
3. Cadherin Binding (GO:0045296)

Molecular Function Late
1. Olfactory Receptor Activity (GO:0004984)
2. GABA-gated Chloride Ion Channel Activity (GO:0022851)
3. GABA-A Receptor Activity (GO:0004890)
4. Benzodiazepine Receptor Activity (GO:0008503)
5. GABA Receptor Activity (GO:0016917)
6. Ligand-Gated Monoatomic Anion Channel Activity (GO:0099095)
7. Transmitter-Gated Monoatomic Ion Channel Activity (GO:0022824)
8. Chloride Channel Activity (GO:0005254)
9. Monoatomic Anion Channel Activity (GO:0005253)
10. Chloride Transmembrane Transporter Activity (GO:0015108)
11. Chloride Ion Binding (GO:0031404)
12. Peptidoglycan Binding (GO:0042834)

K562

Biological Process Early
1. Macromolecule Biosynthetic Process (GO:0009059)
2. protein-RNA Complex Assembly (GO:0022618)
3. Cytoplasmic Translation (GO:0002181)
4. Chaperone Cofactor-Dependent Protein Refolding (GO:0051085)
5. Translation (GO:0006412)
6. Gene Expression (GO:0010467)
7. Regulation of Translation (GO:0006417)
8. Intracellular Protein Transport (GO:0006886)
9. Membraneless Organelle Assembly (GO:0140694)
10. Regulation of B Cell Activation (GO:0050864)

Biological Process Late
1. Sensory Perception of Smell (GO:0007608)
2. Sensory Perception of Chemical Stimulus (GO:0007606)
3. Gamma-Aminobutyric Acid Signaling Pathway (GO:0007214)
4. Synaptic Transmission, GABAergic (GO:0051932)
5. Inhibitory Synapse Assembly (GO:1904862)
6. Chloride Transmembrane Transport (GO:1902476)
7. Synapse Assembly (GO:0007416)

8. Monoatomic Anion Transmembrane Transport (GO:0098656)
9. Inorganic Anion Transmembrane Transport (GO:0098661)
10. Synaptic Membrane Adhesion (GO:0099560)
11. Regulation of Synapse Assembly (GO:0051963)
12. Cell-Cell Adhesion via Plasma-Membrane Adhesion Molecules (GO:0098742)
13. Positive Regulation of Synapse Assembly (GO:0051965)
14. Synapse Organization (GO:0050808)
15. Positive Regulation of Nervous System Development (GO:0051962)
16. Positive Regulation of Cell Junction Assembly (GO:1901890)
17. Chemical Synaptic Transmission (GO:0007268)
18. Axonogenesis (GO:0007409)
19. Chloride Transport (GO:0006821)
20. Calcium-Dependent Cell-Cell Adhesion via Plasma Membrane Cell Adhesion Molecules (GO:0016339)
21. Cell-Cell Adhesion Mediated by Cadherin (GO:0044331)
22. Cell-Cell Junction Organization (GO:0045216)
23. Cell-Cell Junction Assembly (GO:0007043)
24. Adherens Junction Organization (GO:0034332)
25. Cell Morphogenesis Involved in Neuron Differentiation (GO:0048667)
26. Cell Junction Assembly (GO:0034329)
27. Presynaptic Membrane Organization (GO:0097090)
28. Presynaptic Membrane Assembly (GO:0097105)
29. Cellular Response to Nitrogen Compound (GO:1901699)
30. Neuron Projection Morphogenesis (GO:0048812)
31. Regulation of Presynapse Organization (GO:0099174)
32. Regulation of Presynapse Assembly (GO:1905606)
33. Presynapse Assembly (GO:0099054)

Cellular Component Early
1. Focal Adhesion (GO:0005925)
2. Cell-Substrate Junction (GO:0030055)
3. Cytosolic Large Ribosomal Subunit (GO:0022625)
4. Large Ribosomal Subunit (GO:0015934)
5. Lysosomal Membrane (GO:0005765)
6. Mitochondrial Membrane (GO:0031966)
7. Mitochondrial Inner Membrane (GO:0005743)
8. Ribosome (GO:0005840)
9. Anaphase-Promoting Complex (GO:0005680)
10. Small Ribosomal Subunit (GO:0015935)
11. Cytosolic Small Ribosomal Subunit (GO:0022627)
12. Mitotic Spindle Microtubule (GO:1990498)
13. Multivesicular Body Lumen (GO:0097486)
14. Cytoplasmic Vesicle Membrane (GO:0030659)
15. RNA Polymerase III Complex (GO:0005666)

Cellular Component Late
1. GABA-A Receptor Complex (GO:1902711)
2. Dendrite Membrane (GO:0032590)
3. Postsynaptic Specialization Membrane (GO:0099634)
4. Dendrite (GO:0030425)
5. GABA-ergic Synapse (GO:0098982)
6. Synaptic Membrane (GO:0097060)
7. Neuron Projection (GO:0043005)
8. Glutamatergic Synapse (GO:0098978)
9. Catenin Complex (GO:0016342)
10. Cul4-RING E3 Ubiquitin Ligase Complex (GO:0080008)
11. Dystroglycan Complex (GO:0016011)
12. Sarcoglycan Complex (GO:0016012)
13. Adherens Junction (GO:0005912)
14. Postsynaptic Density Membrane (GO:0098839)
15. Cell-Cell Junction (GO:0005911)
16. Axonal Growth Cone (GO:0044295)
17. Astrocyte End-Foot (GO:0097450)

Molecular Function Early
1. Ubiquitin-Like Protein Ligase Binding (GO:0044389)
2. Ubiquitin Protein Ligase Binding (GO:0031625)

Molecular Function Late
1. Olfactory Receptor Activity (GO:0004984)
2. GABA-gated Chloride Ion Channel Activity (GO:0022851)
3. GABA-A Receptor Activity (GO:0004890)
4. Benzodiazepine Receptor Activity (GO:0008503)
5. GABA Receptor Activity (GO:0016917)
6. Ligand-Gated Monoatomic Anion Channel Activity (GO:0099095)
7. Transmitter-Gated Monoatomic Ion Channel Activity (GO:0022824)
8. Chloride Channel Activity (GO:0005254)
9. Monoatomic Anion Channel Activity (GO:0005253)
10. Chloride Transmembrane Transporter Activity (GO:0015108)
11. Chloride Ion Binding (GO:0031404)
12. Peptidoglycan Binding (GO:0042834)
13. CARD Domain Binding (GO:0050700)
14. Poly-Purine Tract Binding (GO:0070717)
15. poly(A) Binding (GO:0008143)

MCF7

Biological Process Early
1. Macromolecule Biosynthetic Process (GO:0009059)

2. Response to Cadmium Ion (GO:0046686)
3. protein-RNA Complex Assembly (GO:0022618)
4. Cellular Response to Cadmium Ion (GO:0071276)
5. Intermediate Filament Organization (GO:0045109)
6. Intracellular Zinc Ion Homeostasis (GO:0006882)
7. Phospholipid Efflux (GO:0033700)
8. Glucose Metabolic Process (GO:0006006)
9. Positive Regulation of Triglyceride Metabolic Process (GO:0090208)

Biological Process Late
1. Sensory Perception of Smell (GO:0007608)
2. Sensory Perception of Chemical Stimulus (GO:0007606)
3. Gamma-Aminobutyric Acid Signaling Pathway (GO:0007214)
4. Synaptic Transmission, GABAergic (GO:0051932)
5. Inhibitory Synapse Assembly (GO:1904862)
6. Chloride Transmembrane Transport (GO:1902476)
7. Synapse Assembly (GO:0007416)
8. Monoatomic Anion Transmembrane Transport (GO:0098656)
9. Inorganic Anion Transmembrane Transport (GO:0098661)
10. Synaptic Membrane Adhesion (GO:0099560)
11. Detection of Chemical Stimulus Involved in Sensory Perception (GO:0050907)
12. Detection of Chemical Stimulus Involved in Sensory Perception of Smell (GO:0050911)
13. Cell-Cell Adhesion via Plasma-Membrane Adhesion Molecules (GO:0098742)
14. Synapse Organization (GO:0050808)
15. Chemical Synaptic Transmission (GO:0007268)
16. Axonogenesis (GO:0007409)
17. Chloride Transport (GO:0006821)
18. Presynaptic Membrane Organization (GO:0097090)
19. Presynaptic Membrane Assembly (GO:0097105)
20. Axon Development (GO:0061564)
21. Neuron Cell-Cell Adhesion (GO:0007158)
22. Nervous System Development (GO:0007399)
23. Cellular Response to Histamine (GO:0071420)
24. Glutamate Receptor Signaling Pathway (GO:0007215)
25. Antimicrobial Humoral Response (GO:0019730)
26. Antimicrobial Humoral Immune Response Mediated by Antimicrobial Peptide (GO:0061844)

Cellular Component Early
1. Intermediate Filament (GO:0005882)
2. Keratin Filament (GO:0045095)
3. High-Density Lipoprotein Particle (GO:0034364)
4. RNA Polymerase II, Core Complex (GO:0005665)

Cellular Component Late
1. GABA-A Receptor Complex (GO:1902711)

2. Dendrite Membrane (GO:0032590)
3. Postsynaptic Specialization Membrane (GO:0099634)
4. Dendrite (GO:0030425)
5. Neuron Projection (GO:0043005)
6. Cul4-RING E3 Ubiquitin Ligase Complex (GO:0080008)

Molecular Function Early
1. Cadherin Binding (GO:0045296)

Molecular Function Late
1. Olfactory Receptor Activity (GO:0004984)
2. GABA-gated Chloride Ion Channel Activity (GO:0022851)
3. GABA-A Receptor Activity (GO:0004890)
4. Benzodiazepine Receptor Activity (GO:0008503)
5. GABA Receptor Activity (GO:0016917)
6. Ligand-Gated Monoatomic Anion Channel Activity (GO:0099095)
7. Transmitter-Gated Monoatomic Ion Channel Activity (GO:0022824)
8. Chloride Channel Activity (GO:0005254)
9. Monoatomic Anion Channel Activity (GO:0005253)
10. Chloride Transmembrane Transporter Activity (GO:0015108)
11. Chloride Ion Binding (GO:0031404)
12. Peptidoglycan Binding (GO:0042834)
13. CARD Domain Binding (GO:0050700)
14. CXCR Chemokine Receptor Binding (GO:0045236)
15. Glutamate Receptor Activity (GO:0008066)
16. Chemokine Activity (GO:0008009)
17. Ligand-Gated Monoatomic Ion Channel Activity (GO:0015276)

NHEK

Biological Process Early
1. Macromolecule Biosynthetic Process (GO:0009059)
2. Response to Cadmium Ion (GO:0046686)
3. Cellular Response to Cadmium Ion (GO:0071276)
4. Gene Expression (GO:0010467)
5. Cellular Response to Copper Ion (GO:0071280)
6. Intermediate Filament Organization (GO:0045109)
7. Supramolecular Fiber Organization (GO:0097435)
8. Cholesterol Efflux (GO:0033344)

Biological Process Late
1. Sensory Perception of Smell (GO:0007608)
2. Sensory Perception of Chemical Stimulus (GO:0007606)

3. Gamma-Aminobutyric Acid Signaling Pathway (GO:0007214)
4. Synaptic Transmission, GABAergic (GO:0051932)
5. Inhibitory Synapse Assembly (GO:1904862)
6. Chloride Transmembrane Transport (GO:1902476)
7. Synapse Assembly (GO:0007416)
8. Inorganic Anion Transmembrane Transport (GO:0098661)
9. Synaptic Membrane Adhesion (GO:0099560)
10. Regulation of Synapse Assembly (GO:0051963)
11. Detection of Chemical Stimulus Involved in Sensory Perception (GO:0050907)
12. Detection of Chemical Stimulus Involved in Sensory Perception of Smell (GO:0050911)
13. Cell-Cell Adhesion via Plasma-Membrane Adhesion Molecules (GO:0098742)
14. Positive Regulation of Synapse Assembly (GO:0051965)
15. Synapse Organization (GO:0050808)
16. Positive Regulation of Nervous System Development (GO:0051962)
17. Positive Regulation of Cell Junction Assembly (GO:1901890)
18. Axonogenesis (GO:0007409)
19. Cell Morphogenesis Involved in Neuron Differentiation (GO:0048667)
20. Central Nervous System Development (GO:0007417)
21. Brain Development (GO:0007420)
22. Sodium-Independent Organic Anion Transport (GO:0043252)

Cellular Component Early
1. Intermediate Filament (GO:0005882)
2. Keratin Filament (GO:0045095)
3. High-Density Lipoprotein Particle (GO:0034364)
4. Triglyceride-Rich Plasma Lipoprotein Particle (GO:0034385)
5. Very-Low-Density Lipoprotein Particle (GO:0034361)
6. Vesicle (GO:0031982)

Cellular Component Late
1. GABA-A Receptor Complex (GO:1902711)
2. Dendrite Membrane (GO:0032590)
3. Postsynaptic Specialization Membrane (GO:0099634)
4. Dendrite (GO:0030425)
5. GABA-ergic Synapse (GO:0098982)
6. Synaptic Membrane (GO:0097060)
7. Neuron Projection (GO:0043005)
8. Dystroglycan Complex (GO:0016011)
9. Sarcoglycan Complex (GO:0016012)
10. Axon (GO:0030424)

Molecular Function Early
1. Ubiquitin-Like Protein Ligase Binding (GO:0044389)
2. Ubiquitin Protein Ligase Binding (GO:0031625)

Molecular Function Late
1. Olfactory Receptor Activity (GO:0004984)
2. GABA-gated Chloride Ion Channel Activity (GO:0022851)
3. GABA-A Receptor Activity (GO:0004890)
4. Benzodiazepine Receptor Activity (GO:0008503)
5. GABA Receptor Activity (GO:0016917)
6. Ligand-Gated Monoatomic Anion Channel Activity (GO:0099095)
7. Transmitter-Gated Monoatomic Ion Channel Activity (GO:0022824)
8. Chloride Channel Activity (GO:0005254)
9. Peptidoglycan Binding (GO:0042834)
10. Poly-Purine Tract Binding (GO:0070717)
11. Sodium-Independent Organic Anion Transmembrane Transporter Activity (GO:0015347)
12. G Protein-Coupled Neurotransmitter Receptor Activity (GO:0099528)

SKNSH

Biological Process Early
1. Response to Cadmium Ion (GO:0046686)

Biological Process Late
1. Sensory Perception of Smell (GO:0007608)
2. Sensory Perception of Chemical Stimulus (GO:0007606)
3. Gamma-Aminobutyric Acid Signaling Pathway (GO:0007214)
4. Synaptic Transmission, GABAergic (GO:0051932)
5. Inhibitory Synapse Assembly (GO:1904862)
6. Synaptic Membrane Adhesion (GO:0099560)
7. Regulation of Synapse Assembly (GO:0051963)
8. Detection of Chemical Stimulus Involved in Sensory Perception (GO:0050907)
9. Detection of Chemical Stimulus Involved in Sensory Perception of Smell (GO:0050911)
10. Positive Regulation of Synapse Assembly (GO:0051965)
11. Regulation of Presynapse Organization (GO:0099174)
12. Regulation of Presynapse Assembly (GO:1905606)
13. Estrogen Metabolic Process (GO:0008210)
14. Plasma Membrane Phospholipid Scrambling (GO:0017121)
15. Oxygen Transport (GO:0015671)
16. Steroid Metabolic Process (GO:0008202)
17. One-Carbon Compound Transport (GO:0019755)
18. Gas Transport (GO:0015669)
19. Carbon Dioxide Transport (GO:0015670)
20. Signaling Receptor Ligand Precursor Processing (GO:0140448)

Cellular Component Early
1. Intermediate Filament (GO:0005882)
2. Lysosomal Membrane (GO:0005765)

3. Keratin Filament (GO:0045095)
4. Lytic Vacuole Membrane (GO:0098852)

Cellular Component Late
1. GABA-A Receptor Complex (GO:1902711)
2. Dendrite Membrane (GO:0032590)

Molecular Function Late
1. Olfactory Receptor Activity (GO:0004984)
2. GABA-gated Chloride Ion Channel Activity (GO:0022851)
3. GABA-A Receptor Activity (GO:0004890)
4. Benzodiazepine Receptor Activity (GO:0008503)
5. GABA Receptor Activity (GO:0016917)
6. Chloride Ion Binding (GO:0031404)
7. Glucuronosyltransferase Activity (GO:0015020)
8. CXCR Chemokine Receptor Binding (GO:0045236)
9. Hexosyltransferase Activity (GO:0016758)
10. UDP-glycosyltransferase Activity (GO:0008194)
11. Epidermal Growth Factor Receptor Binding (GO:0005154)
12. Cysteine-Type Endopeptidase Inhibitor Activity (GO:0004869)
13. Retinoic Acid Binding (GO:0001972)
14. Phospholipid Scramblase Activity (GO:0017128)

APPENDIX C

ANCESTOR WISE COMMON THEMES AND PATHWAYS

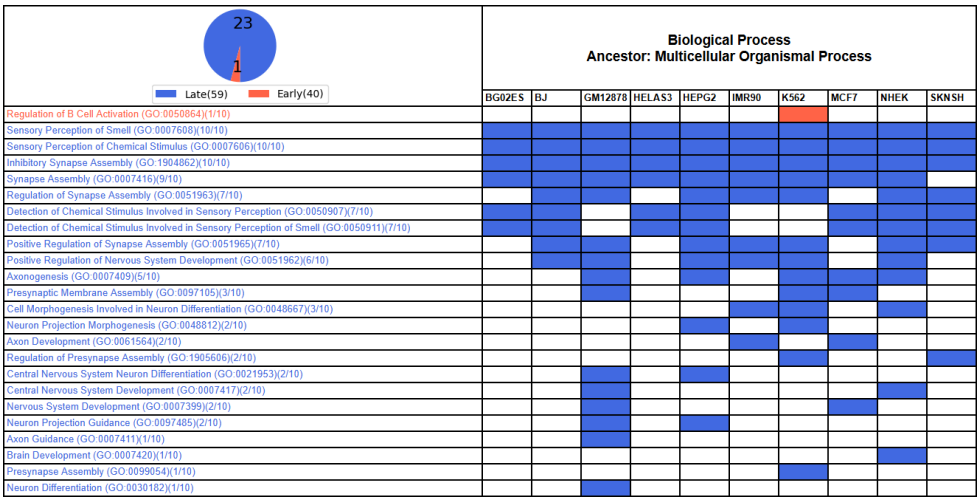

Figure 22: GO Terms for Multicellular Organismal Process and Cell Lines

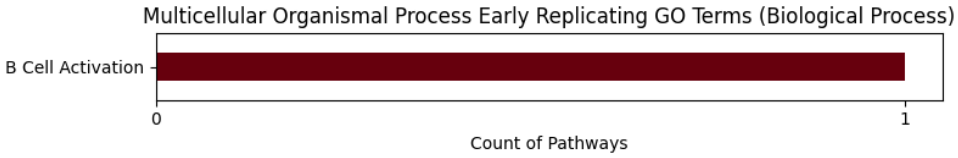

Figure 23: Common Themes for Early Replicating GO Terms Ancestors for Multicellular Organismal Process

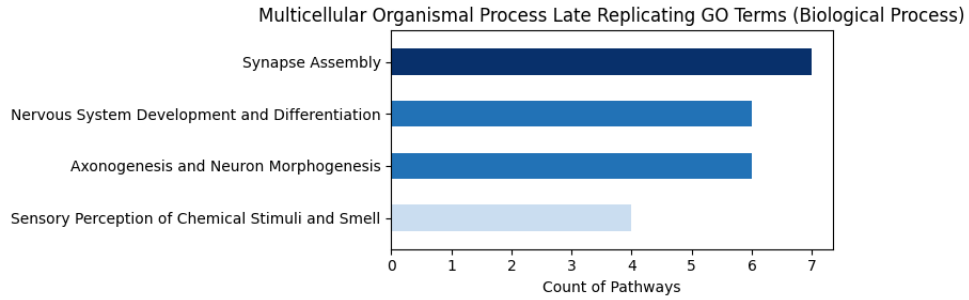

Figure 24: Common Themes for Late Replicating GO Terms Ancestors for Multicellular Organismal Process

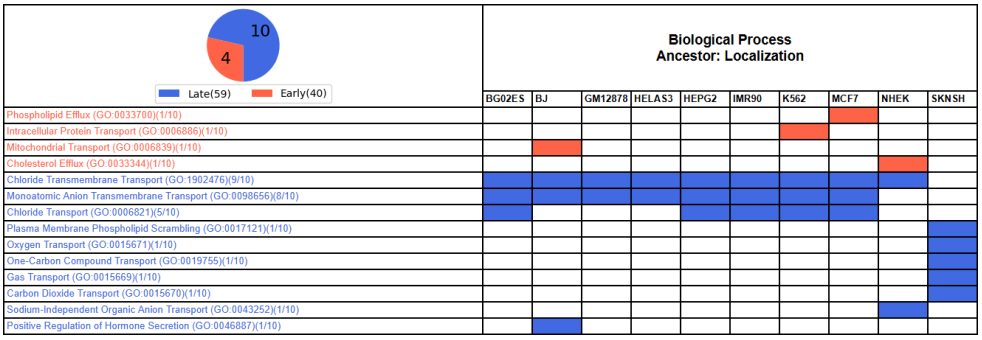

Figure 25: GO Terms for Localization and Cell Lines

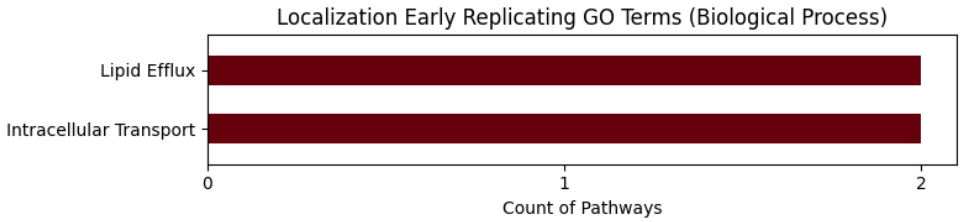

Figure 26: Common Themes for Early Replicating GO Terms Ancestors for Localization

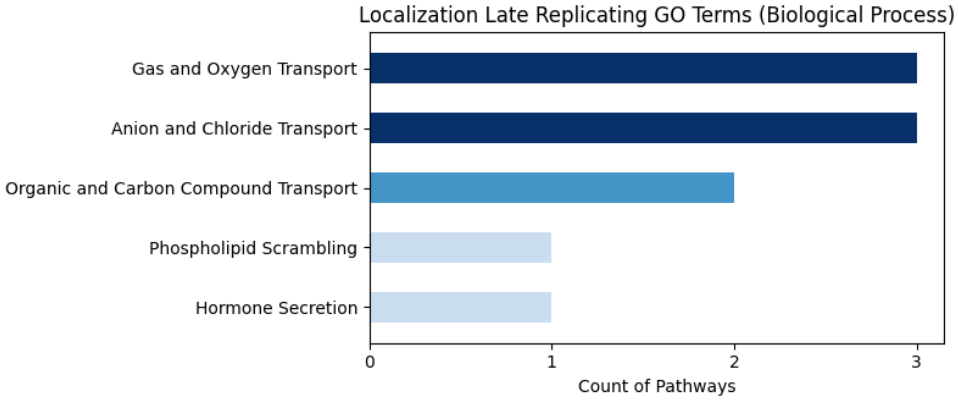

Figure 27: Common Themes for Late Replicating GO Terms Ancestors for Localization

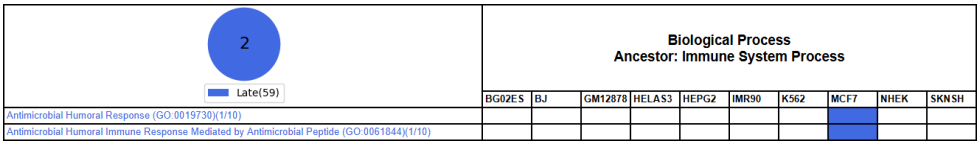

Figure 28: GO Terms for Immune System Process and Cell Lines

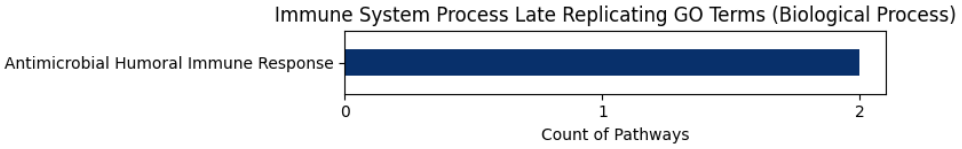

Figure 29: Common Themes for Late Replicating GO Terms Ancestors for Immune System Process

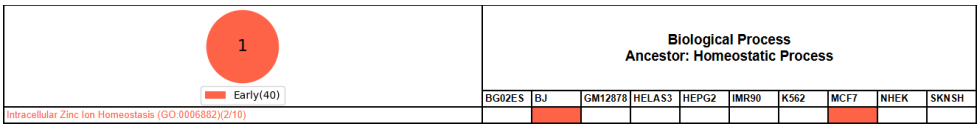

Figure 30: GO Terms for Homeostatic Process and Cell Lines

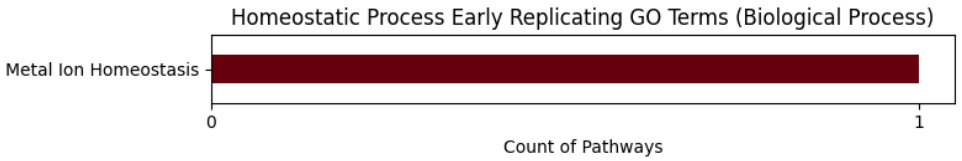

Figure 31: Common Themes for Early Replicating GO Terms Ancestors for Homeostatic Process

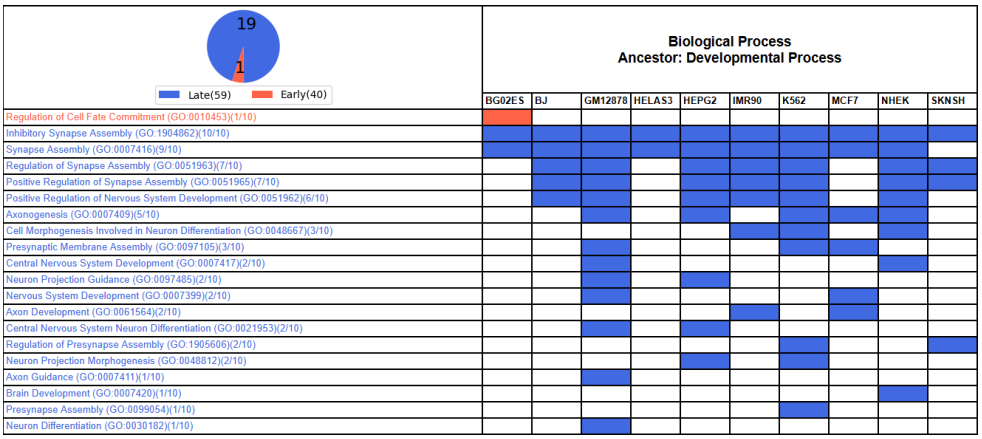

Figure 32: GO Terms for Developmental Process and Cell Lines

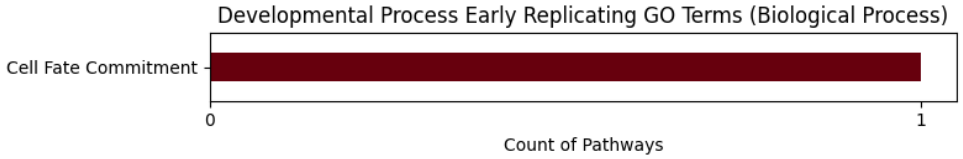

Figure 33: Common Themes for Early Replicating GO Terms Ancestors for Developmental Process

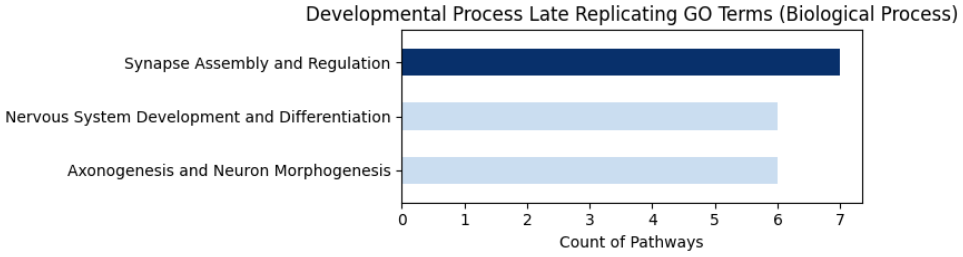

Figure 34: Common Themes for Late Replicating GO Terms Ancestors for Developmental Process

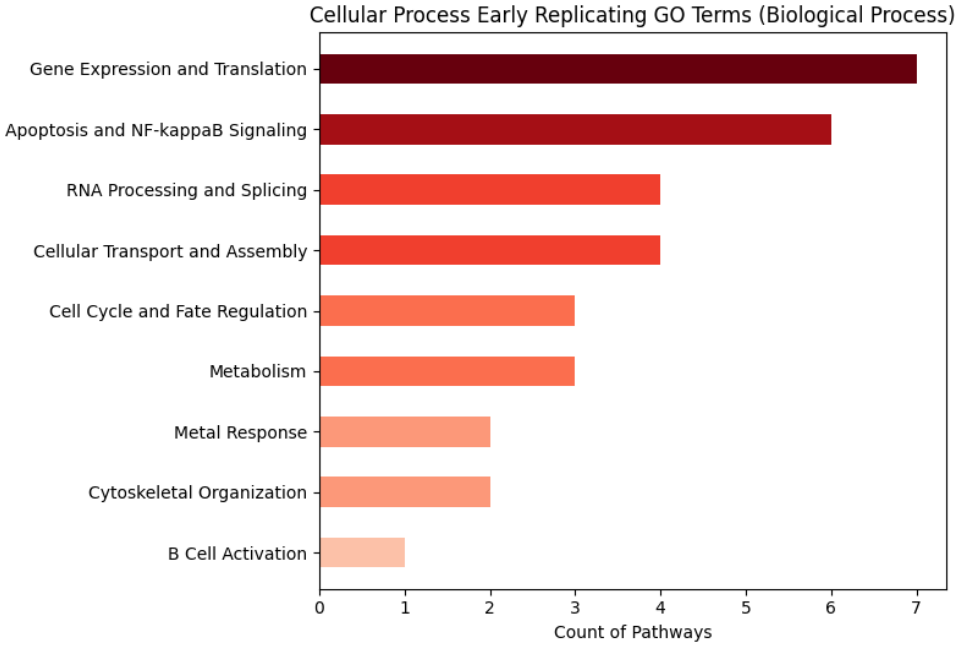

Figure 35: Common Themes for Early Replicating GO Terms Ancestors for Cellular Process

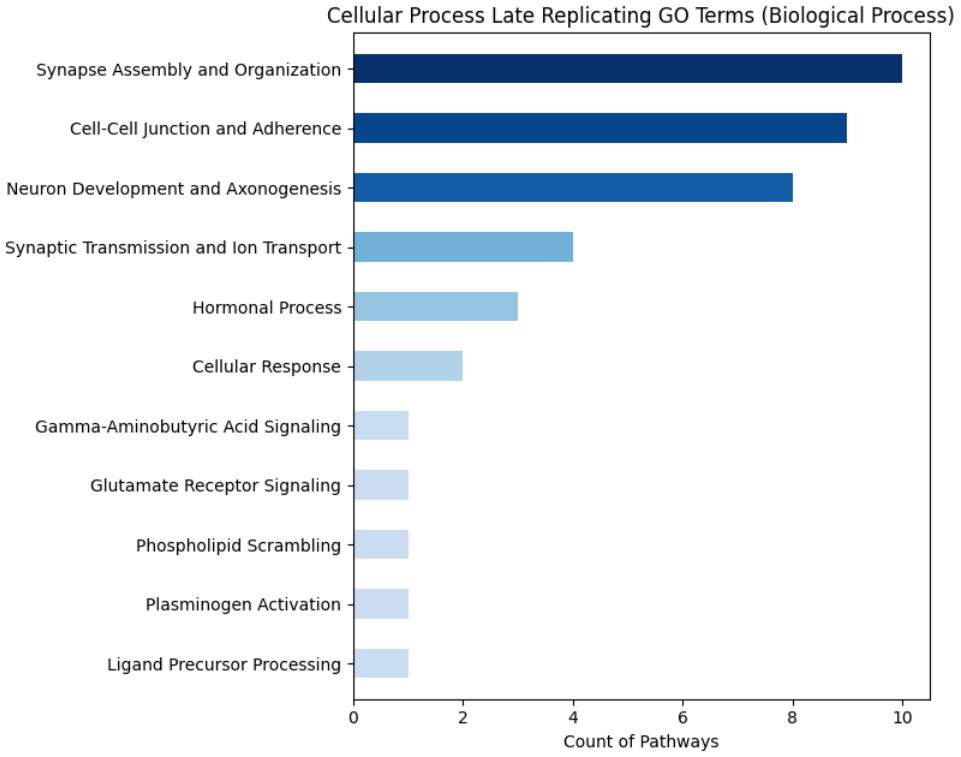

Figure 36: Common Themes for Late Replicating GO Terms Ancestors for Cellular Process

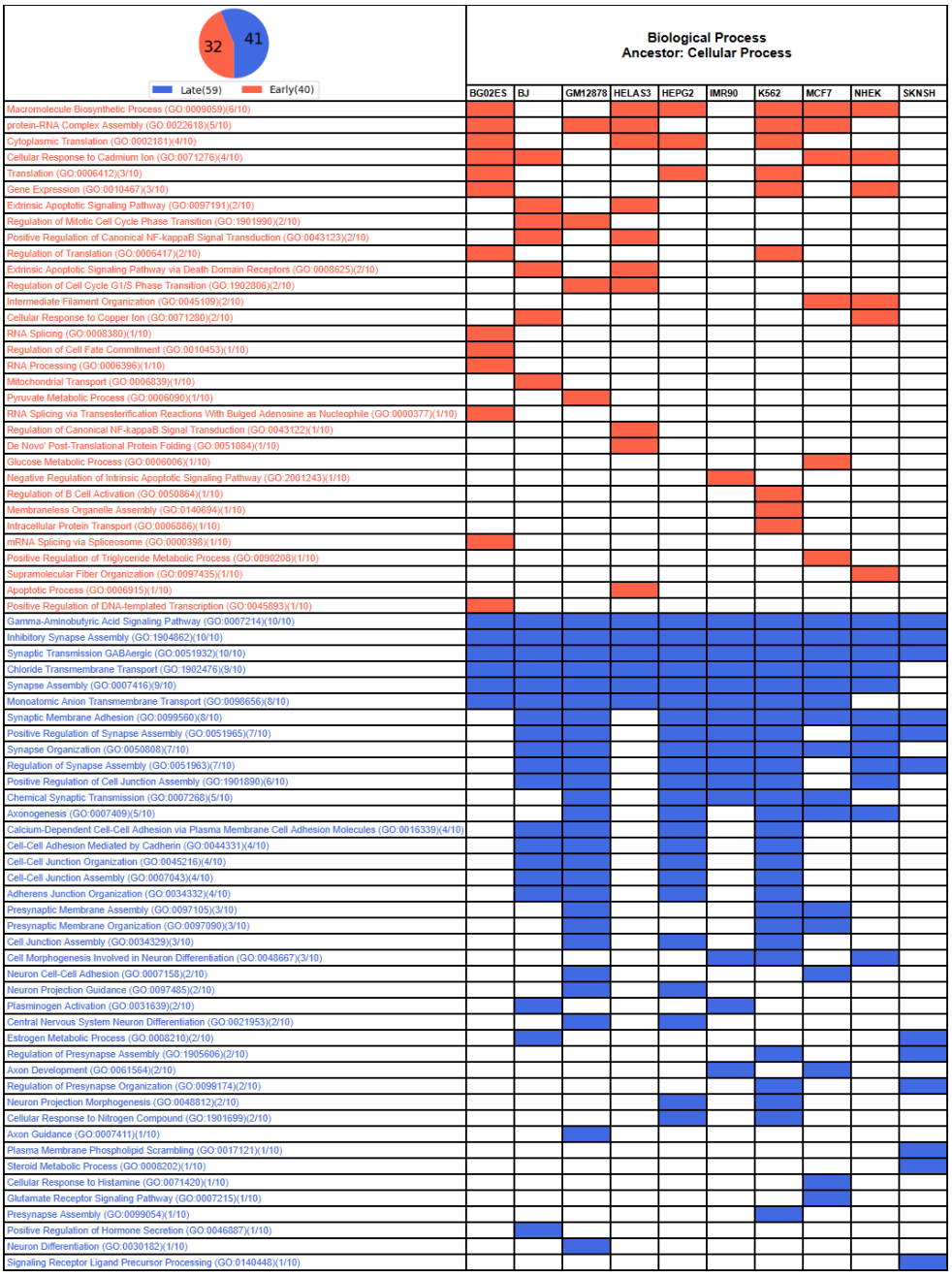

Figure 37: GO Terms for Cellular Process and Cell Lines

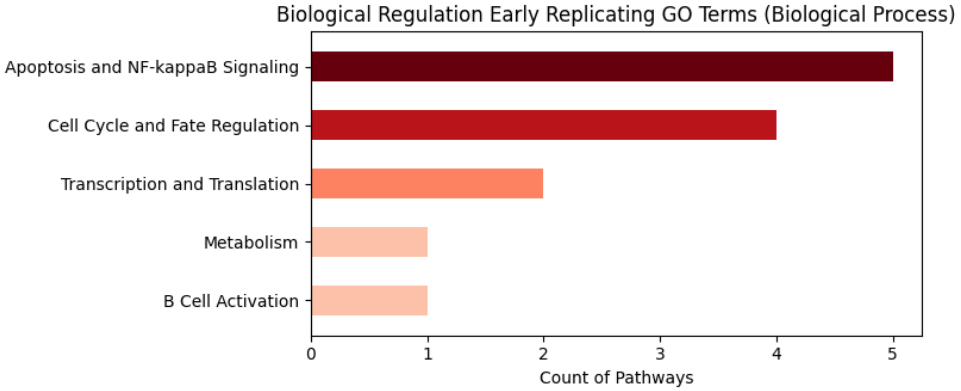

Figure 38: Common Themes for Early Replicating GO Terms Ancestors for Biological Regulation

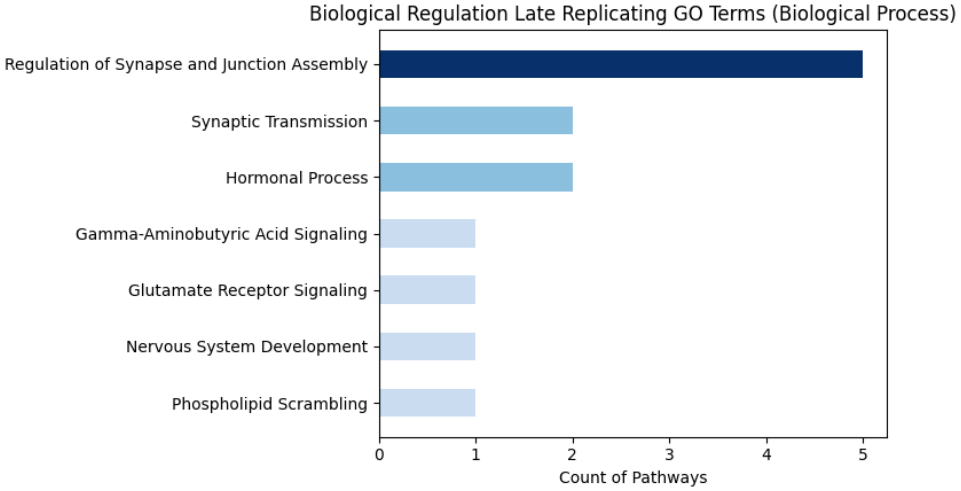

Figure 39: Common Themes for Late Replicating GO Terms Ancestors for Biological Regulation

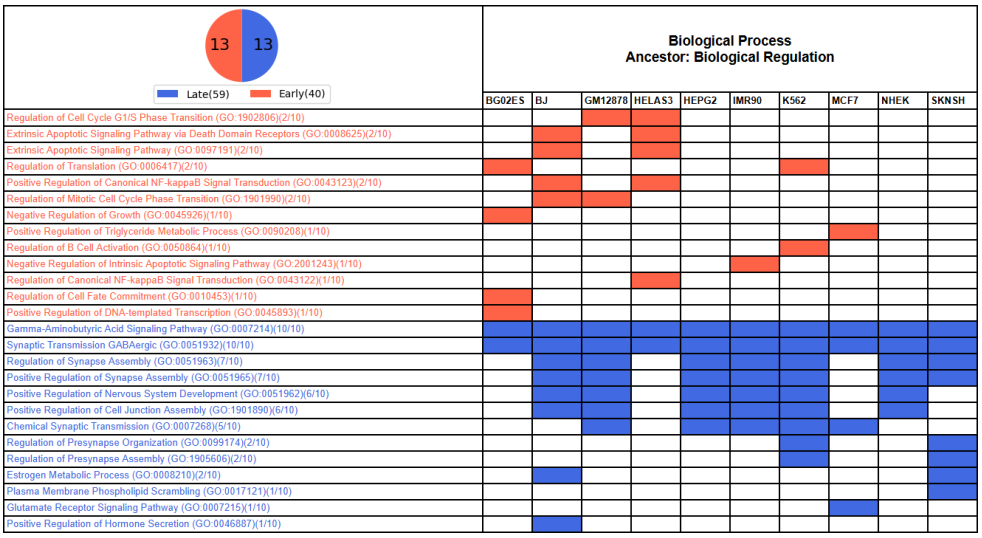

Figure 40: GO Terms for Biological Regulation and Cell Lines

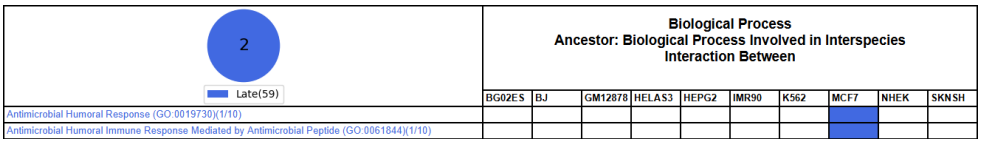

Figure 41: GO Terms for Biological Process Involved in Interspecies Interaction Between and Cell Lines

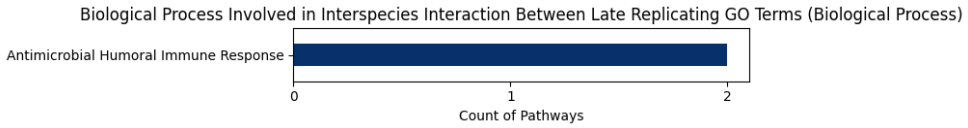

Figure 42: Common Themes for Late Replicating GO Terms Ancestors for Biological Process Involved In Interspecies Interaction Between

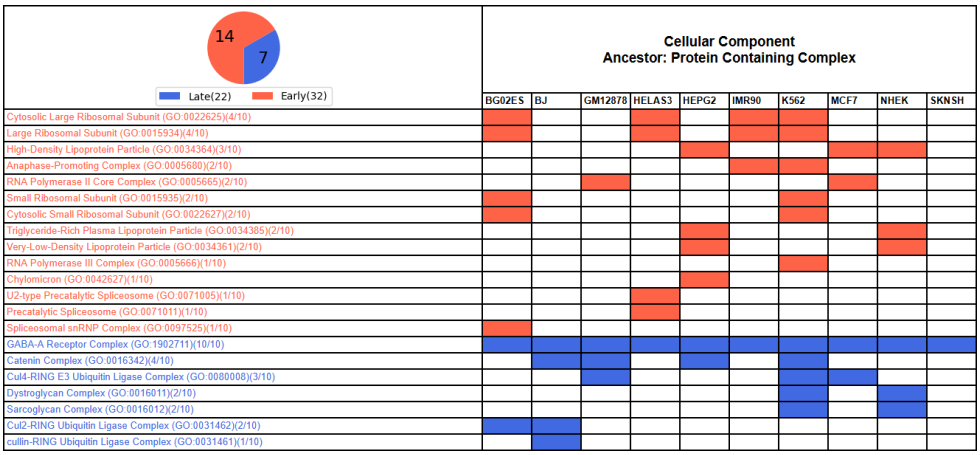

Figure 43: GO Terms for Protein Containing Complex and Cell Lines

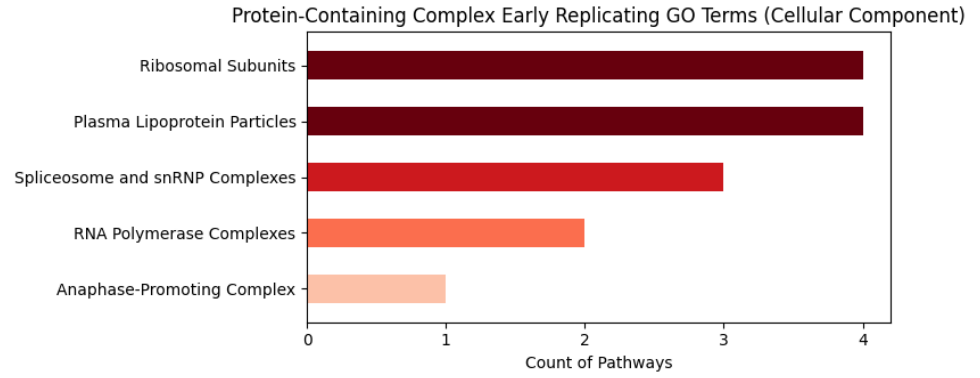

Figure 44: Common Themes for Early Replicating GO Terms Ancestors for Protein-Containing Complex

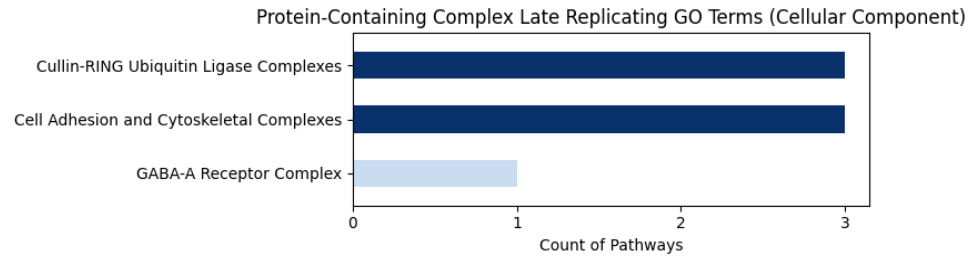

Figure 45: Common Themes for Late Replicating GO Terms Ancestors for Protein-Containing Complex

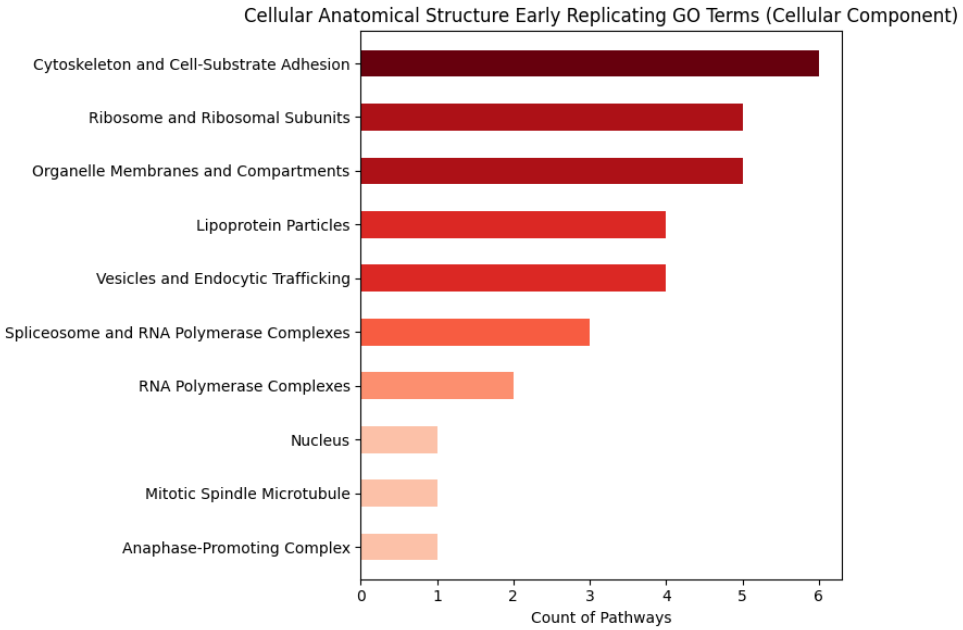

Figure 46: Common Themes for Early Replicating GO Terms Ancestors for Cellular Anatomical Structure

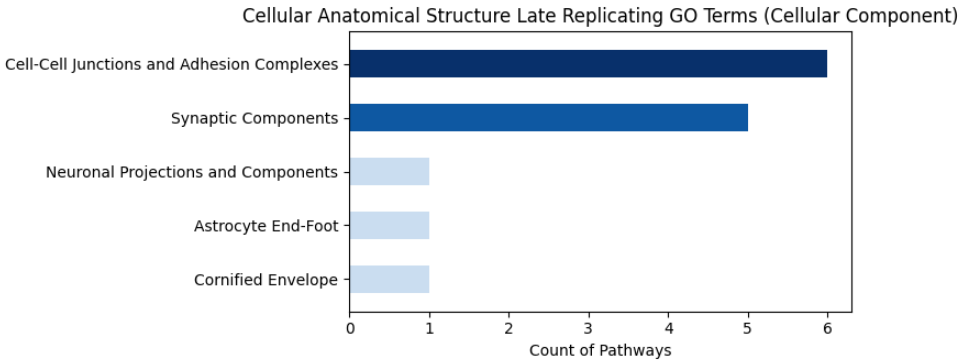

Figure 47: Common Themes for Late Replicating GO Terms Ancestors for Cellular Anatomical Structure

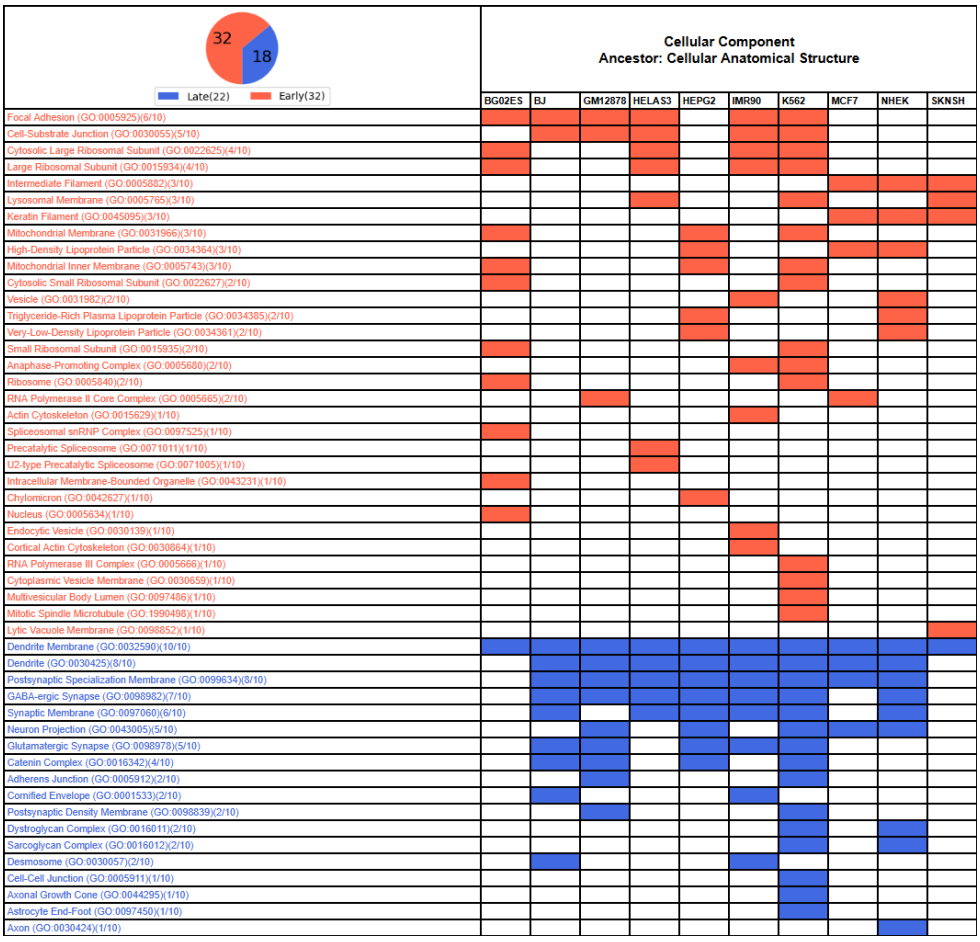

Figure 48: GO Terms for Cellular Anatomical Structure and Cell Lines

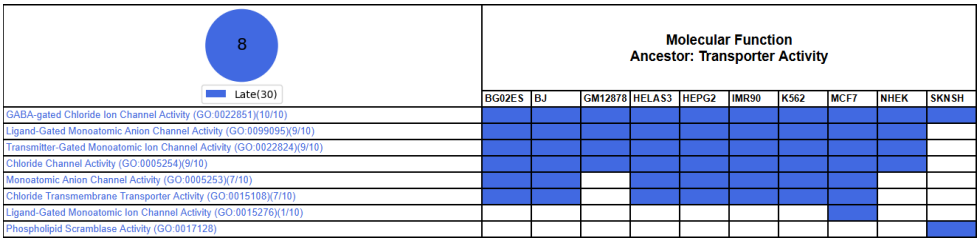

Figure 49: GO Terms for Transporter Activity and Cell Lines

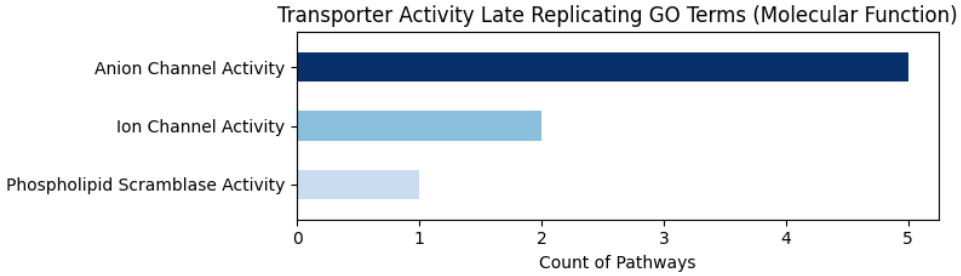

Figure 50: Common Themes for Late Replicating GO Terms Ancestors for Transporter Activity

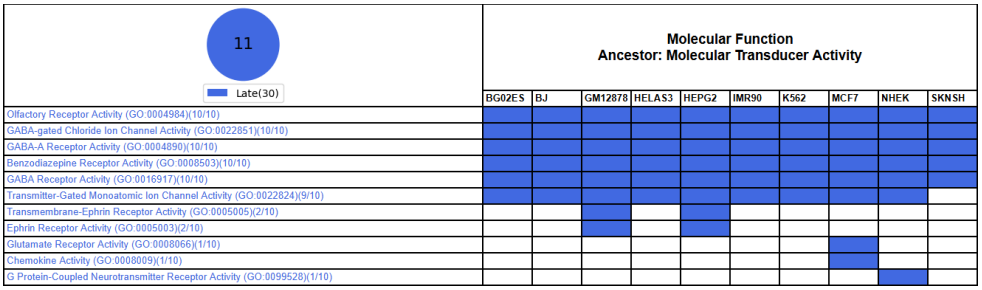

Figure 51: GO Terms for Molecular Transducer Activity and Cell Lines

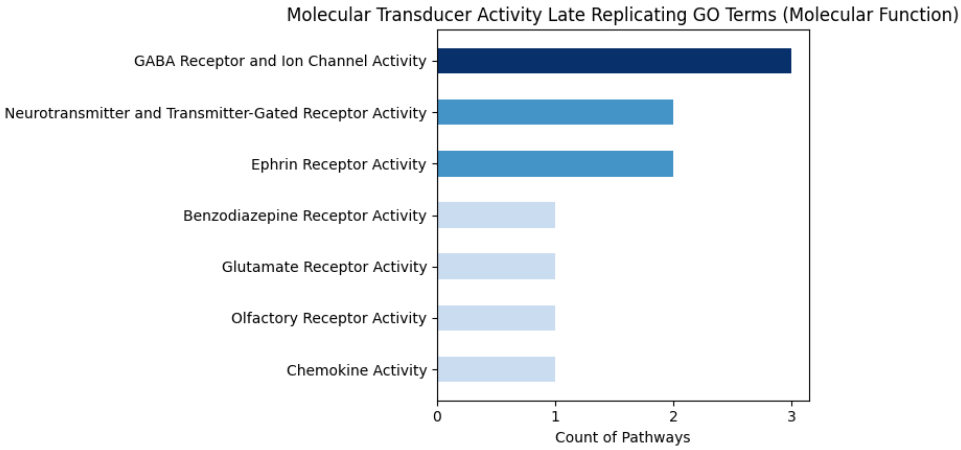

Figure 52: Common Themes for Late Replicating GO Terms Ancestors for Molecular Transducer Activity

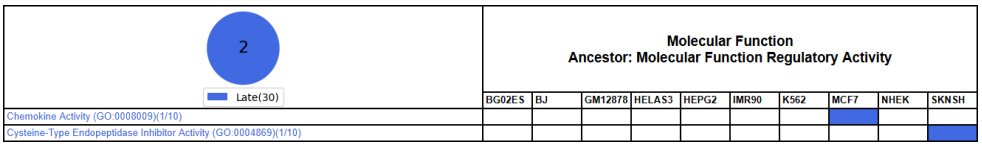

Figure 53: GO Terms for Molecular Function Regulatory Activity and Cell Lines

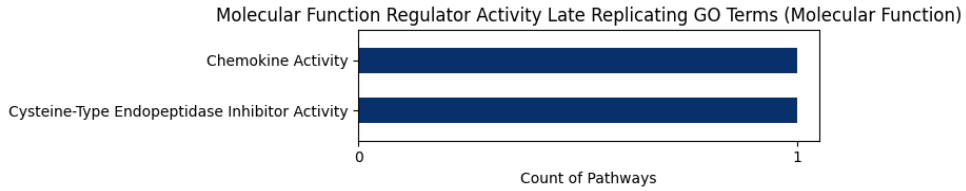

Figure 54: Common Themes for Late Replicating GO Terms Ancestors for Molecular Function Regulator Activity

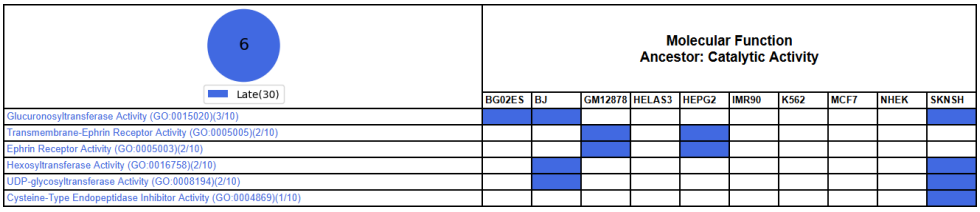

Figure 55: GO Terms for Catalytic Activity and Cell Lines

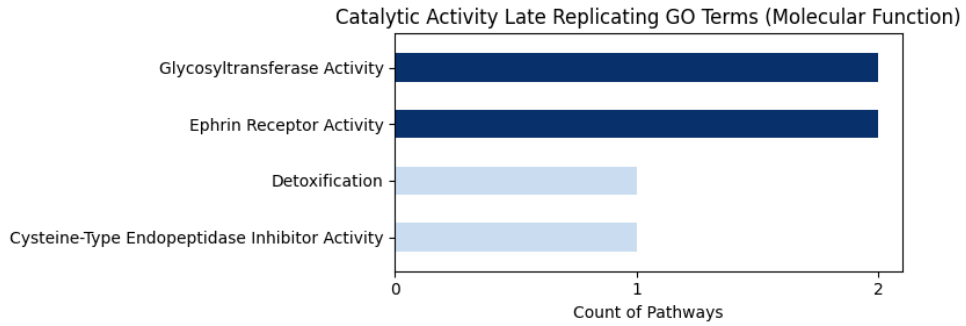

Figure 56: Common Themes for Late Replicating GO Terms Ancestors for Catalytic Activity

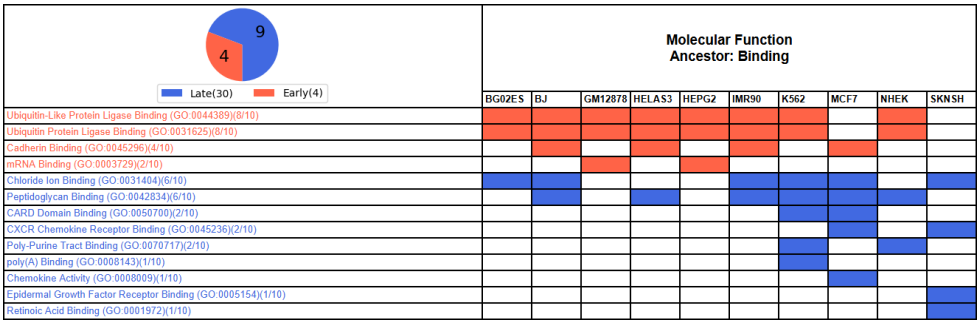

Figure 57: GO Terms for Binding and Cell Lines

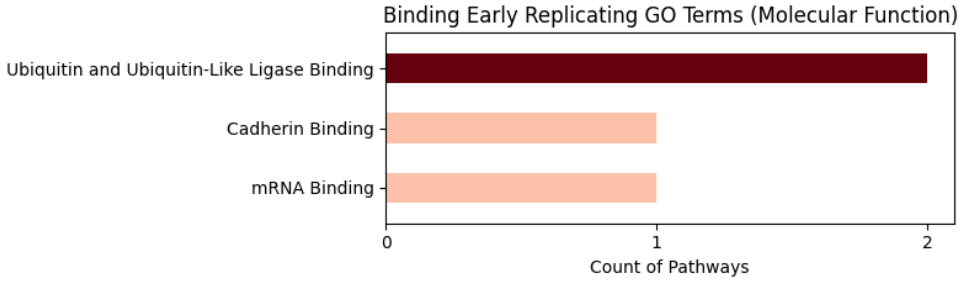

Figure 58: Common Themes for Early Replicating GO Terms Ancestors for Binding

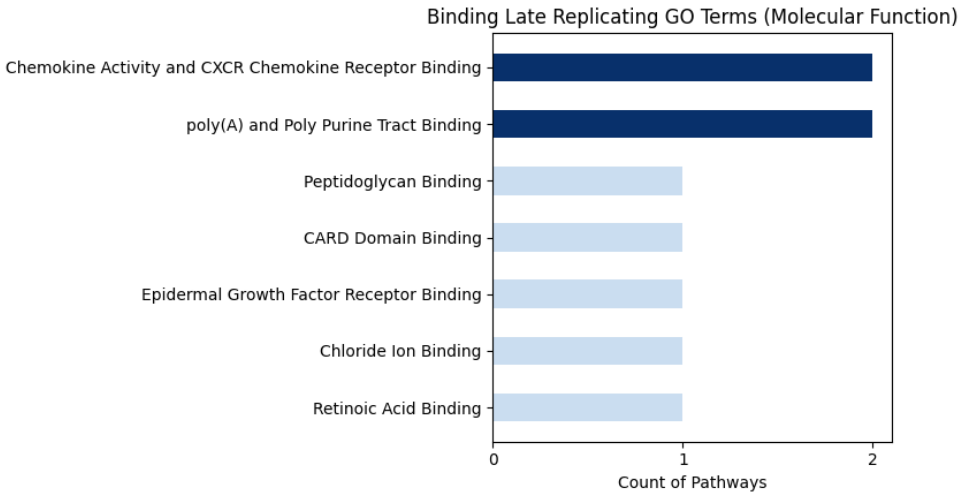

Figure 59: Common Themes for Late Replicating GO Terms Ancestors for Binding

APPENDIX D

CELL LINE SPECIFIC GO TERMS COMMON THEMES

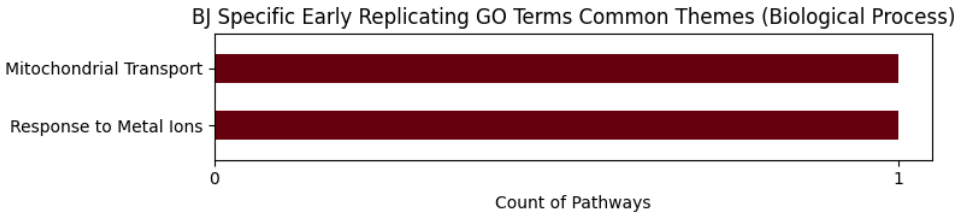

Figure 60: BJ Specific Early Replicating GO Terms Common Themes for Biological Process

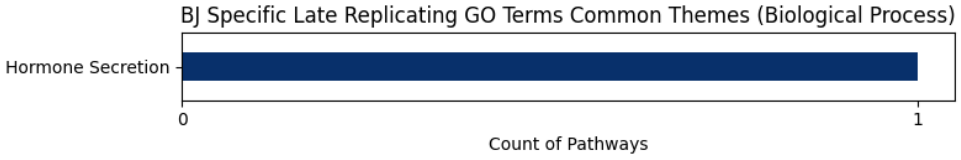

Figure 61: BJ Specific Late Replicating GO Terms Common Themes for Biological Process

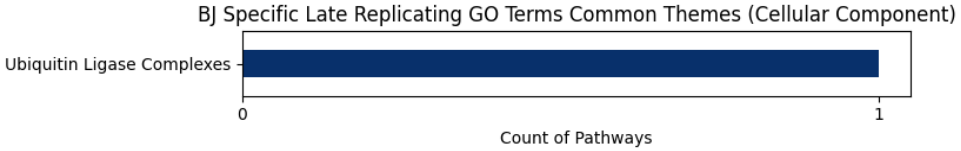

Figure 62: BJ Specific Late Replicating GO Terms Common Themes for Cellular Component

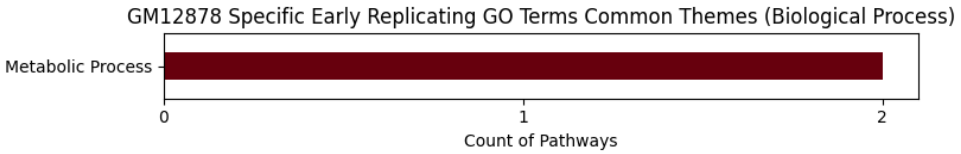

Figure 63: GM12878 Specific Early Replicating GO Terms Common Themes for Biological Process

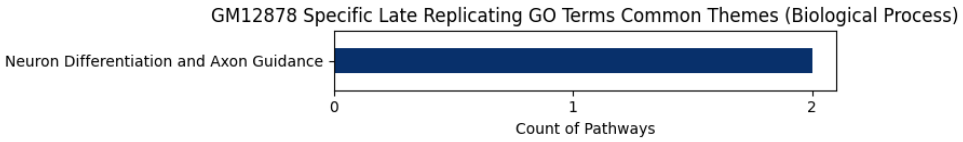

Figure 64: GM12878 Specific Late Replicating GO Terms Common Themes for Biological Process

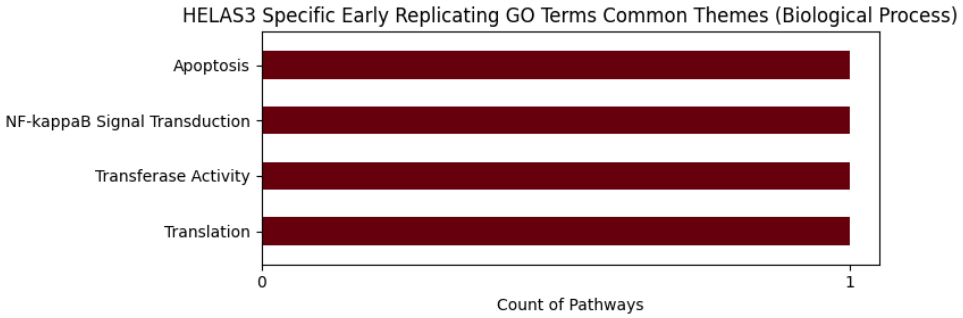

Figure 65: HELAS3 Specific Early Replicating GO Terms Common Themes for Biological Process

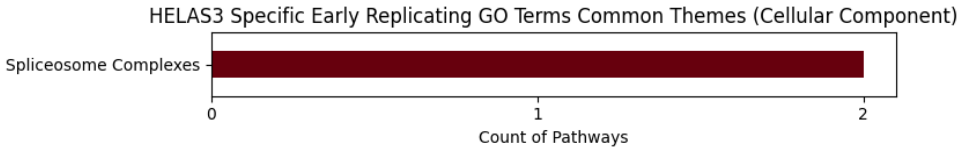

Figure 66: HELAS3 Specific Early Replicating GO Terms Common Themes for Cellular Component

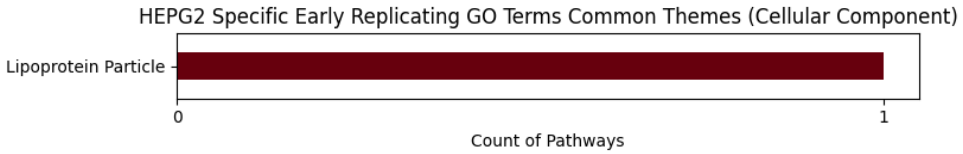

Figure 67: HEPG2 Specific Early Replicating GO Terms Common Themes for Cellular Component

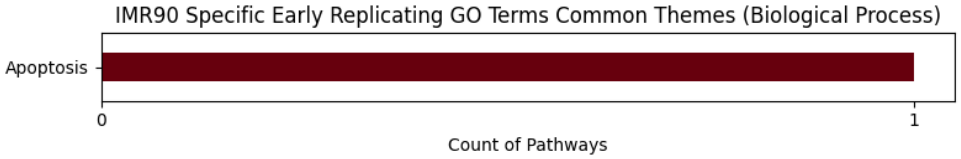

Figure 68: IMR90 Specific Early Replicating GO Terms Common Themes for Biological Process

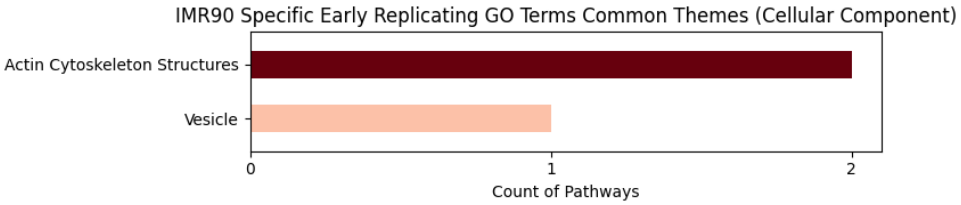

Figure 69: IMR90 Specific Early Replicating GO Terms Common Themes for Cellular Component

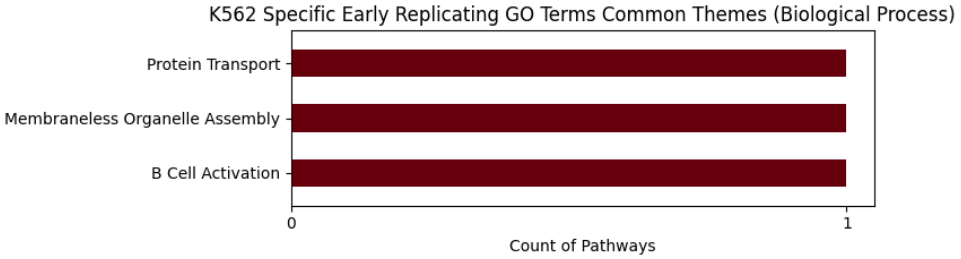

Figure 70: K562 Specific Early Replicating GO Terms Common Themes for Biological Process

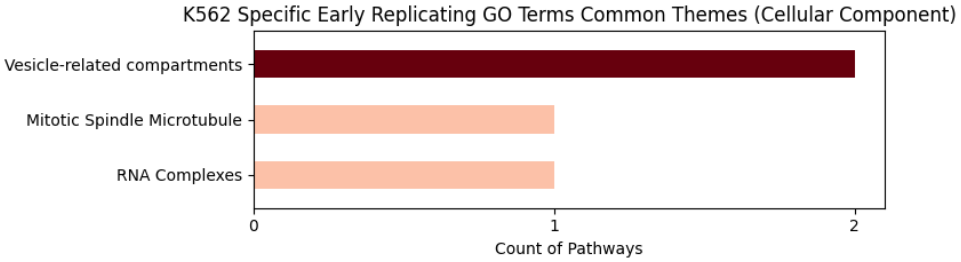

Figure 71: K562 Specific Early Replicating GO Terms Common Themes for Cellular Component

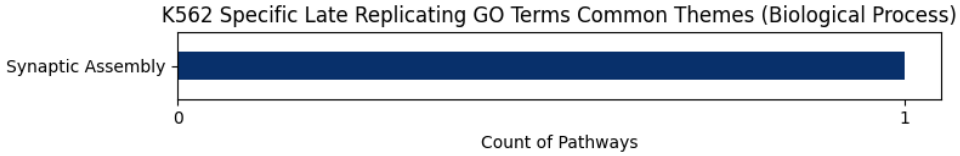

Figure 72: K562 Specific Late Replicating GO Terms Common Themes for Biological Process

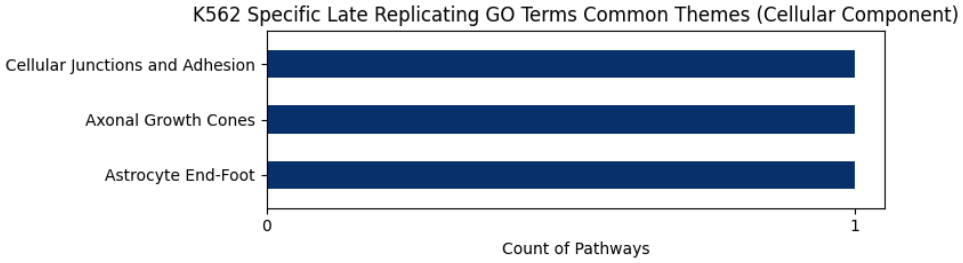

Figure 73: K562 Specific Late Replicating GO Terms Common Themes for Cellular Component

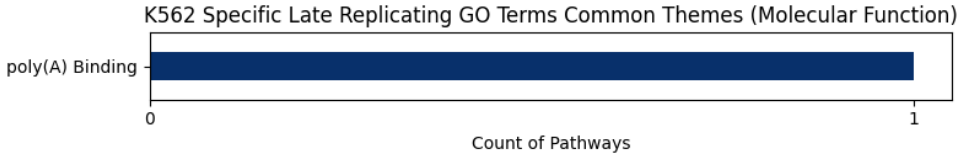

Figure 74: K562 Specific Late Replicating GO Terms Common Themes for Molecular Function

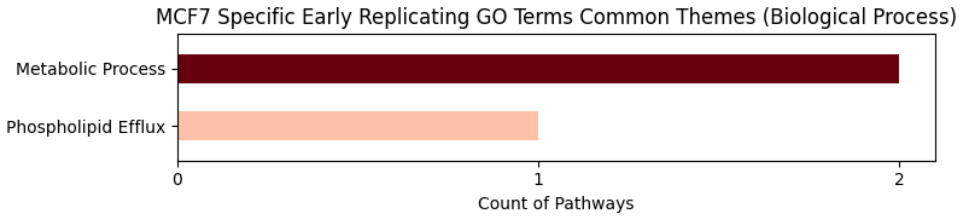

Figure 75: MCF7 Specific Early Replicating GO Terms Common Themes for Biological Process

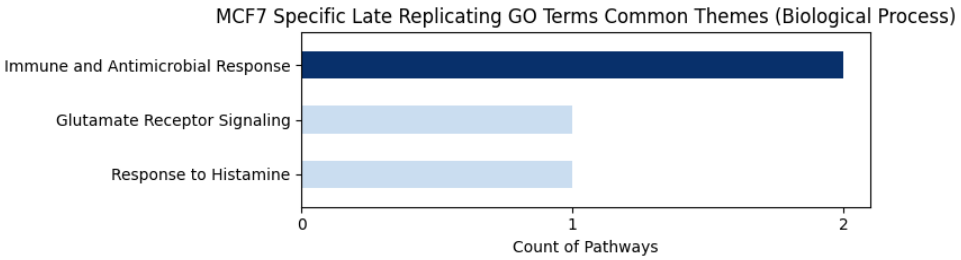

Figure 76: MCF7 Specific Late Replicating GO Terms Common Themes for Biological Process

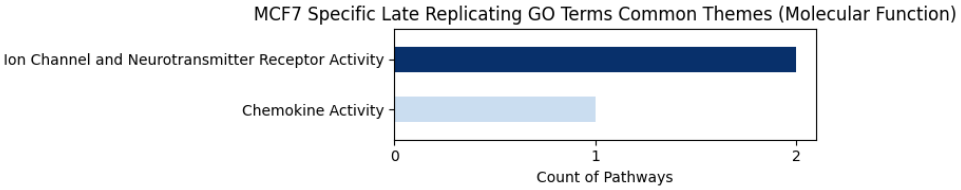

Figure 77: MCF7 Specific Late Replicating GO Terms Common Themes for Molecular Function

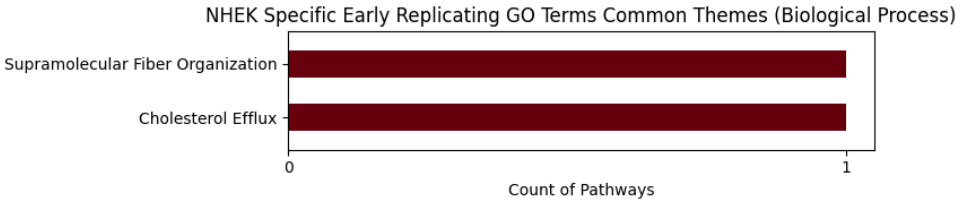

Figure 78: NHEK Specific Early Replicating GO Terms Common Themes for Biological Process

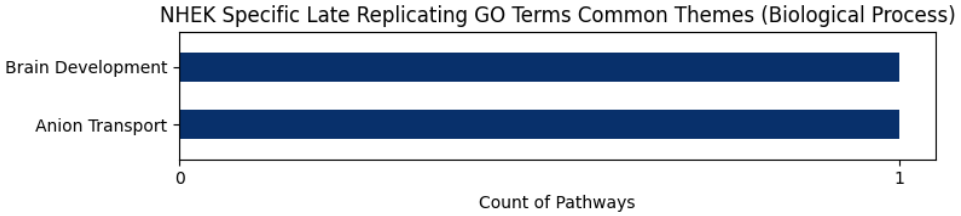

Figure 79: NHEK Specific Late Replicating GO Terms Common Themes for Biological Process

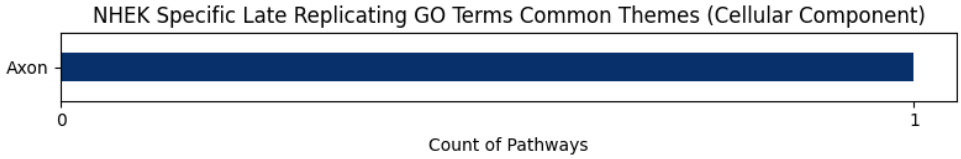

Figure 80: NHEK Specific Late Replicating GO Terms Common Themes for Cellular Component

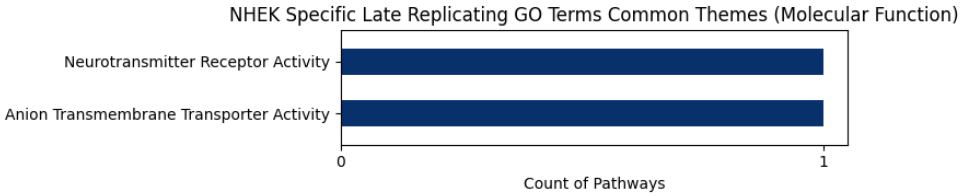

Figure 81: NHEK Specific Late Replicating GO Terms Common Themes for Molecular Function

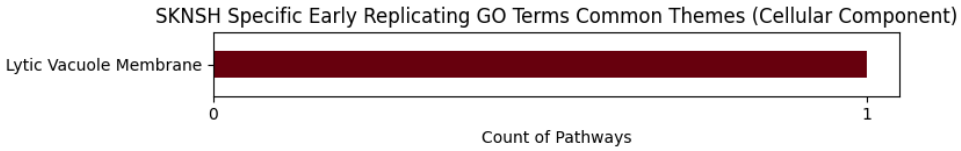

Figure 82: SKNSH Specific Early Replicating GO Terms Common Themes for Cellular Component

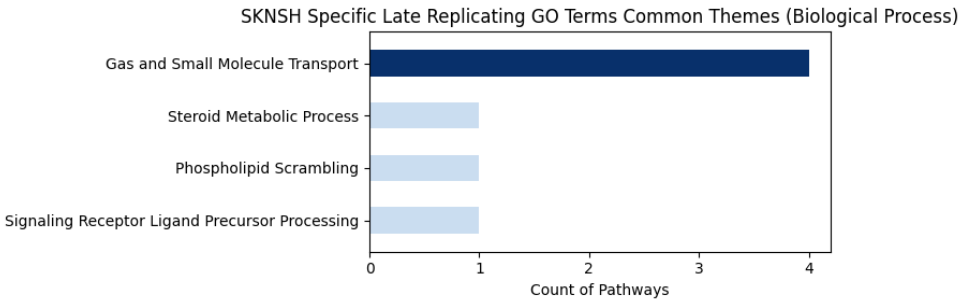

Figure 83: SKNSH Specific Late Replicating GO Terms Common Themes for Biological Process

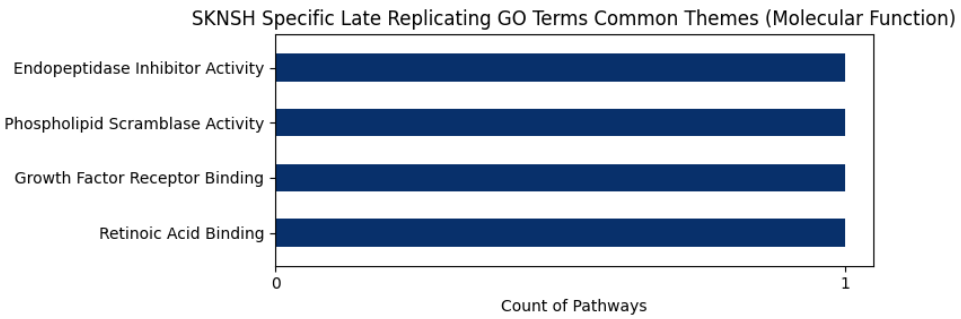

Figure 84: SKNSH Specific Late Replicating GO Terms Common Themes for Molecular Function