Nick Santoso (49135866) , Elena Kovacevic (91499111), Amine Benhammou (89250187)

BMEG 310 project report: BRCA analysis

Abstract

This paper explores the genetic characteristics and clinical aspects of breast cancer using data from The Cancer Genome Atlas (TCGA). The study focuses on the genetic alterations in the BRCA genes, which play a significant role in breast cancer development. This study utilizes unsupervised learning techniques to analyze TCGA data, with the aim of identifying highly transcribed genes and their impact on biochemical pathways, as well as factors influencing overall patient survival. Three distinct clusters are derived from mutation analysis, revealing patterns in patient characteristics. Differential expression analysis highlights the significance of certain pathways, such as cell cycle and NK cell-mediated cytotoxicity, in breast cancer progression. The study discusses the clinical implications of demographic characteristics, mutation types, and pathway dysregulation. Despite limitations in the clustering method, the findings contribute valuable insights for potential targeted therapies and further investigations into breast cancer treatment strategies.

Introduction

One of the most common types of cancer amongst women is breast cancer. It is the second cause of cancer related deaths in the US, after lung cancer. Ductal Carcinoma and Breast Invasive Carcinoma (BRCA) for example claimed the life of around 43'000 patients in 2022 alone. Though this number may seem high, the death rate of breast cancer has been steadily declining since 1989, dropping with an estimated percentage of 43% over that time period [1]. This is due to the improvement of the early detection tools, as well as the ever evolving research in the field.

At a molecular level, one of the main drivers for breast cancer is the occurrence of one or more genetic alterations in the BRCA genes. These alterations may occur in one of the two types of epithelial cells inside a mammary gland, basal and luminal, and can yield a neoplasm that can be split into five distinct subtypes: Basal-like, Luminal A, Luminal B, human epidermal growth factor receptor 2 (HER2), and normal-like. These subtypes were discovered during the early 2000's, using unsupervised learning models, and have since greatly improved the quality of prognosis, since it offered a measure for tumor classification [2].

According to data from the National Cancer Institute and over a duration of 5 years since diagnosis, LumA and LumB have the highest survival rate, both above 90%, HER2 was around 85%, and basal had the worst with 77.1%. But the subtype of the tumor does not determine the survival rate accurately, stage of the disease at diagnosis is the crucial factor when determining this metric [3]. These metrics are great in comprehending the disease's progression, but leaves a lot of gaps in our understanding of the underlying genetic causes of this progression. of the intricate molecular mechanisms and genetic factors contributing to cancer development and progression is essential for refining treatment strategies and developing targeted therapies.

In our analysis, we aim to look at the raw TCGA data, pre-process it, eliminating irrelevant noise, then use unsupervised learning techniques to collect different insights into the data. Mainly, we aim to determine the most transcribed genes and the implications of those on the upregulation/downregulation of different biochemical pathways within the cells, and how that influences the overall survival. Furthermore, we aim to determine the aspect that is most influential in analyzing the survival of patients, whether that would be subtype of BRCA, stage of tumor, age, or any other clinical as well. In doing so, we hope to be able to shed light on valuable pathways or parameters that can be tackled in biomedical research to help mitigate some of the symptoms of BRCA.

Methods

In this project, three datasets from The Cancer Genome Atlas (TCGA) were utilized, encompassing mutation, clinical, and RNA sequencing data for patients with BRCA. Using R subsetting methods, common patients were identified, and data corresponding to those patients across all three datasets were extracted.

This study began by analyzing patient demographics in clinical data and their survival outcomes based on their disease-specific survival status. This was done by creating Kaplan-Meier plots using R's 'survival' and 'survminer' packages to visualize how long a patient remains in the study before dying with a tumor. This was to better understand how BRCA affects each group and whether or not the study should focus on a specific demographic.

Next, to enhance clarity in subsequent analyses, various preprocessing steps were explored. Specifically, mutation data underwent multiple filtering approaches, such as the inclusion of HIGH and MODERATE impact mutations while excluding wildtype patients or including HIGH impact mutations while excluding wildtype patients, among others. Through trial and error it was found that preprocessing which excluded LOW impact mutations as well as wildtype patients yielded the best results. To implement this filter, wildtype patients were identified by analyzing the Mutation Annotation Format (MAF) and examining patient samples with an insertion mutation, specifically those with a dash in one of the two 'tumorseqAllele' columns.

Mutation analysis first involved creating bar charts to visualize the frequency of various mutation data aspects. Subsequently, an oncoplot was generated using R's pheatmap function, incorporating patient IDs and gene names (HUGO ID). This binary plot highlighted the top three mutated genes, representing the majority of mutations in the dataset. Hierarchical clustering was then applied to the oncoplot data, resulting in the selection of four clusters for further investigation. These clusters were analyzed for trends in subtype, age, race, radiation therapy attendance, sex, and ethnicity.

The selection of four clusters was based on the observation that, once clustered, the first three were consistently similar in size, as well as the second cluster predominantly comprising the basal subtype. This choice aimed to highlight clearer distinctions in the underlying cellular pathways driving each cluster. Subsequent analysis of patient characteristics within each cluster served to validate the decision to choose four clusters over alternative numbers, demonstrating the most distinct separation between each group compared to other options when comparing cancer subtypes. Upon observation, it was noted that cluster 4 had a significantly different number of participants compared to clusters 1-3. Consequently, cluster 4 and its participants were excluded from further analysis.

Survival analysis was then conducted using the clusters derived from mutation analysis alongside the clinical data. A Kaplan-Meier plot was generated based on progression-free survival status and the duration of progression-free survival obtained from clinical data.

Differential analysis was conducted on patients within each cluster with the RNA sequencing data to uncover significant variations among distinct clusters using R's DESeq2 library. The dataset was refined by excluding genes with fewer than 100 counts across all samples, ensuring a more robust analysis of differential gene expression. The results were further filtered to include only significant findings with an adjusted p-value of 0.05. To analyze the quality of the differential analysis mapping, a heatmap was made using the top 10 most down and upregulated genes.

As there were 3 clusters, pairwise pathway analysis was subsequently conducted based on the results from the earlier differential analysis. Gene annotation was performed using R's AnnotationDbi and org.Hs.eg.db library, translating gene IDs from ENSEMBL to ENTREZ format. Utilizing the pathview and gage libraries, along with kegg.sets.hs and segment.idx.hs datasets, gene names and their corresponding log2fold change data from the differential analysis were mapped to their respective signaling and metabolic pathways using the gage() function. Finally, the top 5 upregulated and downregulated pathways were extracted for further examination.

Results

Clinical Summary

From the clinical data, a few demographic characteristics of patients were determined. Most patients were female, though there were 10 samples from men. 697/1006 patients were White, 162/1006 were Black or African American, 59/1006 were Asian, and 1/1006 were American Indian or Alaska Native. There were 5 subtypes designated in the clinical data, LumA, LumB, Basal, HER+, and Normal. Black/African American participants were most likely to have Basal mutations in comparison to other ethnicities (appendix A, fig 1a). Similarly, Asian participants were most likely to have Her2 mutations (appendix A, fig 1b). Something to note is that because there is only 1 participant with subtype data that falls within the "American Indian or Alaska Native", so no meaningful conclusions can be observed in this demographic. Survival analysis between subtypes with similar samples sizes for each race were performed (Appendix A, fig 2). No significant conclusions could be drawn from the Kaplan-Meier plots of White and Asian participants due to high P values, however between Basal and LumA subtypes in African populations, those with Basal subtypes would have worse survival outcomes. Higher stages of tumors were observed to have worse survival outcomes (appendix A, fig 3)

Mutation data oncoplot

The oncoplot (Fig 1.)shows 4 distinct clusters (from left to right):

1. The large blue section, indicating no mutation on any of the 3 genes.
2. The section in the middle, representing the intersection between TP53 and TTN.
3. The section after section 2 represents the cluster containing only PIK3CA.
4. The rightmost section represents the intersection between all 3 genes.

Mutation summaries

A closer inspection of our datasets revealed a predominance of missense mutations, composing 65% (66620/103314) of the total variations in the pre-processed sample. The second closest are frameshift at 9% (9245/103314) of the total sample (Appendix A, Fig 4a.). The most common mutation type, and by a large margin, are SNVs accounting for 89% of total mutation types (Appendix A, Fig 4b.). Similarly, 89% of the variant types are SNPs, followed by deletion at 11%, and with the insertion and oligo-nucleotide polymorphism (ONP) numbers being almost negligible (Appendix A, Fig 4c.). The most commonly mutated genes within our sample were PIK3CA, TP53 and TTN. (Appendix A, Fig 4d.)

By looking at the plots, we can determine that SNV missense mutations on either PIK3CA, TP53, or TTN are the main variations that are common to all the patients within our sample. To visualize the prevalence of each mutation type in each gene, we generated a bar chart breaking down the involvement of each variant class in the mutation of each gene (Appendix A, Fig 5.). As concluded, missense mutations dominate the total mutations occurring in the 3 most mutated genes. Using this data, we proceeded to generate a heatmap of the top 3 genes, which will help us further cluster our data.

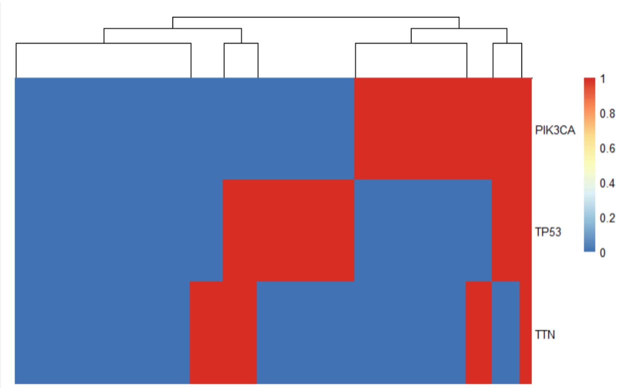


Fig 1. Heat Map of the top 3 genes with dendrogram above.

Survival analysis

Survival analysis did not yield any meaningful results as the survival trends between each cluster are too similar to one-another. The Kaplan-Meier plot (Fig. 2) displaying lines very close to one-another implies that the survival of each cluster is not dependent on the clusters formed/information used to form the clusters.

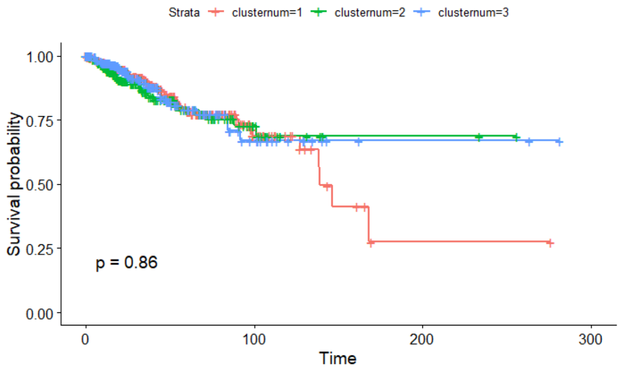


Fig 2. Kaplan-Meier plot of 4 clusters from mutation data (progression free survival)

Cluster composition analysis

We can observe that a majority of the basal subgroups are in cluster 2, that a majority of cluster 3 is composed of luminal A and that cluster 1 is mainly a split of all subtypes (Appendix A, fig 6). No other significant differences were observed in any of the other clinical features.

Differential analysis

Differential analysis yielded results that can be displayed using an MA plot (Appendix A, fig 7), in which a majority of the samples in the dataset have a log fold change of ~ 1 to -1, which can be observed especially as the mean of normalized counts increases.

Pathway analysis

A heatmap was formed to examine the quality of mapping from mutation analysis to differential analysis by plotting each sample vs the top 10 up and downregulated genes (Appendix A, figure 8). From the top to the bottom are the top 10 downregulated genes and top 10 upregulated genes respectively. The clusters in the top row are not clear sections like in mutation analysis (Fig 1), implying that the mapping from mutation data to RNA sequence data is not very good.

Pairwise analysis

As mentioned in the methods section, pairwise analysis was performed on the results of differential analysis due to the nature of log2 fold change as a method to measure changes in expression level. The log2 fold changes of 3 pairs were examined, and will be denoted by P# (numerator, denominator), the pairs being: P1 (cluster 1, cluster 2); P2 (cluster 1, cluster 3); P3 (cluster 2, cluster 3). The up and down regulated pathways are pathways that are either up or downregulated in the numerator cluster relative to the denominator cluster.

$$log_2(fold\ change) \ = \ log_2(\frac{condition\ 1}{condition\ 2})$$

- If log2(fold change) > 0, it indicates upregulation in Condition 1 compared to Condition 2.
- If log2(fold change) < 0, it indicates downregulation in Condition 1 compared to Condition 2.
- If log2(fold change) = 0, it indicates no change in expression.

The top 5 most up and down regulated pathways can be found in appendix A, figure 9. Only pairs 1 and 3 were chosen for analysis based on the cluster distribution found on the pathway analysis heatmap. This is because there is little to no separation between cluster pair 1 and 3 in comparison to cluster pair 1 and 2, and cluster pair 2 and 3.

Discussion

Clinical Analysis

From the clinical data, a few demographic characteristics of patients were determined, all of which have been observed in literature. Majority of patients were female, which is the case with breast cancer, however men make up 10% of breast cancer cases, which means they were underrepresented in the dataset [4]. Most breast cancer patients are between the age of 40 and 70, with the mean age being 58.5, this is consistent with literature reports of median age of 62 [1]. Within different races, different distributions of breast cancer subtypes were observed. LumA was the predominant type across White, Asian and Black or African American patients. Black or African American patients had the highest likelihood of having basal type (compared to other races), and kaplan-meier survival analysis found lower survival outcomes in those with basal type compared to lumA; both these findings are consistent with literature [3]. Kaplan-meyer analysis found Black or African Americans with the basal subtype have lower survival outcomes than those with the LumA subtype. Asian patients have the highest likelihood of having the HER2+ subtype, which has also been reported in literature [5, 6]. Though the analysis did not show correlation between being HER2+ and having lower survival outcome for Asian patients (or in general), literature does suggest HER2+ patients have a worse prognosis compared to LumA [3]. It was found that higher stages have worse disease specific survival outcomes, which is also observed in literature [3]. There were no findings during analysis for the DSS outcomes of all patients with different subtypes, however literature reports heavily that there are differences [3].

Mutation Analysis - Clustering

It was observed during initial data exploration, and difficulty clustering, that many subtypes of breast cancer have the same driver genes, TP53, TTN, PIK3CA. This matches the findings of research, which finds TP53 and PIK3CA as the most dominant driver genes across all subtypes of BRCA. In fact, the activation of PIK3CA is an early sign of malignant transformation, meaning the cells started to acquire cancer characteristics, resulting in a metabolic reprogramming. TP53 on the other hand, responsible for regulating cellular division, is characterized by its inactivation, which leads to the frantic reproduction of tumor cells [7].

Due to the prevalence of activity around these genes in all cases of BRCA, determining clear clusters containing a somewhat even distribution of patients across each cluster proved difficult. Upon a thorough investigation of our data, we determined that certain characteristics were skewing our results. For example, many patients had low impact mutations, such as silent mutations, which had no contribution to our analysis. Furthermore, patients with wild type characteristics, mainly having an insertion or ONP variant types, were discarded, to only keep deletion and SNP variants.

Survival analysis on the generated clusters was done using Kaplan-Meier curves; looking at disease specific survival and then progression free survival turned up no significant differences between the three clusters. However, it is known through literature that there are differences in survival outcomes in different subtypes of breast cancer [3]. This would indicate the clusters had overlap between the known subtypes of cancer, and were not very well differentiated. However, the clustering algorithm did separate Luminal subtype types (making up the large majority cluster 3), and Basal and Her2 subtypes (making up the majority of cluster 2). So although the clustering was not optimal, it could still yield some results on different clusters of breast cancer and the different pathways involved.

Differential Expression Analysis

By looking at the heat map generated using the genes, we can see the clusters formed by the dendrogram. We can notice that the separation between the clusters is not optimal. Likely due to the complications during mutation analysis (many types of breast cancer having the same driver genes which meant the hierarchical clustering algorithm could not differentiate well between them). Cluster 1 (C1) and cluster 3 (C3) for example form the wide majority of the patients in terms of variation, meanwhile cluster 2 (C2) is not as much scattered, with most observations being concentrated on one side of the heat map. Comparing C1 and C3 is therefore pointless due to the widespread of the observations. We therefore focused on comparing C1 and C2, and C2 and C3.

The differentially expressed pathways between C1 and C2 seem to have importance in how breast cancer progresses, and these pathways have been cited in literature as being important in current immunotherapies, or as having potential to be such. The top three downregulated (in C1 with respect to (w.r.t.) C2) pathways observed during the first piece-wise round of DEA were cell cycle, natural killer cell mediated cytotoxicity, and antigen processing and presentation. All three of these pathways are currently being targeted in immunotherapies to treat breast cancer.

First, the cell cycle being differentially expressed is expected in cancer, as the cell begins to shift its metabolism to become a tumor, so does its cycle, inactivating secondary pathways, and diverting all energy pathways towards the one goal of growing the tumor. Certain genes, such as GADD45, involved in tumor suppression [8], are differentially expressed in different subtypes of cancer. The cell cycle pathway map shows GADD45 being overexpressed for C1. GADD45A is typically overexpressed in LumA and LumB subtypes, which C1is largely composed of, and underexpressed in the triple-negative type (which largely falls into the basal subtype). Studies have found that within the triple-negative breast cancer type, expression of GADD45A is associated with worse outcomes and categorizing cancer with both of those features as high risk may help with finding appropriate treatment plans [9]. Other genes such CDK1, which is involved in tumor growth, are downregulated in C1 w.r.t C2, suggesting inhibition of this pathway may be more effective for treating C2 than C1 (10).

Second, the natural killer (NK) cell mediated cytotoxicity is the second most downregulated pathway in our analysis. This finding matches the research, NK cells are major contributors to the anti-tumor response, and downregulating them would offer the best conditions for the tumor to grow [11]. In treating HER2+ breast cancer specifically, better outcomes have been achieved when using targeted immunotherapy that features inducing the cytotoxic factors that are part of the antibody dependent cell-mediated cytotoxicity (ADCC) [12]. This therapy involves the patient receiving enriched natural killer (NK) cells, which causes upregulated ADCC specifically to target and destroy breast cancer cells [12]. In some breast cancer patients, it has been found that NK cells have lower cytotoxicity, this is characterized by lower levels of p-30 related proteins; this includes NKp46 [12] which appears as downregulated in C1 (relative to C2) in the natural killer cell mediated cytotoxicity pathway. Low concentrations of NK cells are typically seen in HER2 and Luminal cancer subtypes, as these subtypes have a higher tendency to lose the ability to synthesize NK cells in comparison to basal-type [12]. C1 is largely composed of luminal subtype cancer, and C2 is majoritarily formed of basal subtype, so this finding is consistent with literature.

Other genes in this pathway such as KIR, which codes for inhibitory/activation receptors necessary for natural killer function [13], and RAE1 which is overexpressed in certain lines of breast cancer and contributes to induction of EMT features in cells [14], were found to be differentially expressed in our analysis, matching literature results.

Third, the antigen processing and presentation pathway is involved in the activation of NK mediated cytotoxicity, and the CD8 and CD4C T-cell receptor signaling pathways. It is reasonable that the pathway contributing to inhibition of growth/reproduction of tumour cells would be less expressed in patients with cancer. The NK cytotoxicity pathway is downregulated in C1, so antigen processing and presentation pathway being downregulated in C1 as well is consistent with this result. All of these pathways are involved in the body's immune response, and as discussed previously can potentially be exploited to target cancer cells. Current explorations of activating this pathway to induce ADCC have shown promising results, suggesting inhibition of this pathway is important to cancer progression [15]

DNA replication is the fourth most down regulated pathway in C1 w.r.t. C2, indicating DNA replication is occurring at higher rates in C2. High levels of DNA replication are associated with poorer outcomes in breast cancer; clinical trials are being undergone (as of 2022) looking at targeting this pathway for the treatment of HER2 and triple negative type breast cancer [16]. This implies Her2+ and triple negative (which mostly make up C2) are better candidates for this type of therapy, perhaps due to more up regulation as this analysis would suggest.

The fifth most down regulated pathway (C1 w.r.t. C2) is the cell adhesion molecules (CAM) pathway. CAMs are important to maintaining cell connectivity, and in cancer, the deregulation in expression of certain proteins is a large contributor to epithelial mesenchymal transition (EMT), and subsequently cancer metastasis. In the pathway, CLDN was highlighted as one of the differentially expressed genes. CLDN codes for a family of transmembrane proteins that play a major role in the integrity of tight junction TJs, which keep cells connected [17]. TJ dysregulation is frequently seen during EMT, and is therefore in highly metastatic cancers, low expression of proteins involved in TJ, such as CLDN proteins, is typically seen [18]. Changes in the expression of CLDN proteins are frequently seen in breast cancer [17,18], and studies have found that their inhibition or over-expression can play a role in a tumor's progression [17].

Other differentially expressed proteins that were highlighted in this pathway, including MAG and CDH2, have also been cited in literature as important in determining progression or presence of cancer, and potentially important in creating more targeted therapies or earlier detection strategies [19, 20]. Based on literature, the CAM pathway seems to be very significant in breast cancer progression, and potentially important in developing better classifications and therapies of cancer.

Upregulated pathways of C1 with respect to C2 include MAPK signaling pathway, endocytosis, fatty acid metabolism, vascular smooth cell contraction, circadian rhythm - mammal.

The MAPK signaling pathway links together extracellular signals and intracellular responses, which can change the cells ability to proliferate, differentiate, and undergo apoptosis (programmed cell death). In pathway analysis of how the genes are differentially expressed, there is a mix of genes that are either up or down regulated in C1 with respect to C2. AP1 is an example of the gene that is upregulated in cluster 1; it is related to a cell's ability to adapt to different environments, which is a trait commonly exhibited by cancer cells especially those that metastasize. AP1 has been described in breast cancer, and its overexpression is typically indicative of an invasive cancer as the cancer cell gains better ability to travel, proliferate, and survive in a variety of microenvironments [21]. This may seem contradictory to earlier described cell cycle pathway downregulation (of C1 with respect to C2), however a study has found that different types of breast cancer have different methods of metastasis that can change the rate of progression or the site of metastasis [22]. The signaling pathway did show cluster 1 had downregulation in certain genes, including NFkB, which is responsible for the distribution in normal cell proliferation and cell apoptosis balance in breast cancer cell lines [23]. It is therefore possible the clusters are representing two different mechanisms that occur for different groupings of breast cancer.

In breast cancer, overexpression disruptions in the endocytosis pathway are responsible for helping the cancer cell monitor and evade tumor immune responses, as well as helping in the nutrient scavenging processes [24]. In treatment, these disruptions make it difficult for antibodies to bind to the cell surface, which reduces (or eliminates) the efficiency of immunotherapies previously described involving activation of ADCC [24]. This means C1 would likely have a poorer response to traditional immunotherapies, which is in line with findings from the downregulated pathways which found C2 to potentially be a better candidate for those types of therapies.

Fatty acid metabolism is commonly altered in cancer, to induce creation of more fatty acid building blocks, which can then be used to construct membranes. This plays a role in the cancer cells ability to signal, and grow or metastasize [25]. It has been found that different subtypes of breast cancer have different changes in the fatty acid metabolism, and these could suggest that they need different approaches to treatment [26]. The differential expression in genes in the fatty acid metabolism pathway supports that the two clusters need to be approached differently due to the cancers involving different pathways for growth, proliferation, and signaling.

The circadian rhythm is a potential indicator for patient prognosis and responsiveness to therapy/drugs [27]. Although as a whole the pathway is more upregulated in C1 w.r.t. C2, there are certain genes, such as CLOCK, that are upregulated in C2 (w.r.t. C1). Upregulation of CLOCK is associated with high grade glioma (brain/spine cancer), and is one of the genes that could be used as an indicator of a patient's prognosis and/or response to certain kinds of treatment [27]. On the other hand, PER is upregulated in C1 w.r.t. C2, and can also be used as an indicator to assess the same outcomes to treatment as CLOCK [27]. This could be due to the duality of CLOCK; it act as either an oncogene or a tumor suppressor in different cancers [27], so it may be the case that it is already acting as a tumor suppressor in C2, meaning only C1 would respond to “clock” treatments.

Now looking at the comparison of C2 w.r.t. C3, upregulated pathways include cell cycle, NK cell mediated cytotoxicity, antigen processing and presentation, DNA replication, and CAM. All of these were also observed to be downregulated in C1 w.r.t. C2, which is not only suggestive of C1 and C3 being very similar (as assumed after generating heatmap), but also confirm the opposing behaviors between basal and Her2 type, vs luminal types. This also confirms that our findings match the literature.

As for the downregulated pathways, the 5 most downregulated are Circadian rhythm - mammal, Valine, leucine and degradation, focal adhesion, fatty acid metabolism, and endocytosis. Circadian rhythm, fatty acid metabolism, and endocytosis being upregulated in C2 with respect to C1 and C3, suggests these pathways are indeed differences between C2 and the other clusters. For the circadian rhythm, certain papers associate variations in regulations of certain genes such as PER and BMAL1 to the breast cancer subtypes. Additionally variations occur the most within the luminal A group when compared to basal types [28].

The focal adhesion pathway covers the much broader signal transduction pathways that are characteristic of breast cancer. This is because genes such as FAK and PAK are heavily expressed in cancer patients, especially at the stage of metastasis [29][30]. In this analysis case, the downregulation of this pathway in C2 when compared to C3 suggests that basal and HER2 types require these pathways to a lesser extent than luminal A and B. Additionally, Bcl2, a protein used as prognostic for breast cancer, was found by research to be highly upregulated in luminal A [31], a fact that our findings support, since C2 is downregulated compared to C3.

The valine, leucine, and isoleucine, also called branch chained amino acids (BCAA), are part of a nine amino-acid group who's dysregulation is directly associated with cancer progression [32]. There is no clear cut decision as to the exact utility of BCAAs, as they contribute to the inhibition of cancer, by stopping tumor migration, but other research also suggests that with the right dosage, these amino-acids can lead to the opposite effect, contributing to the size growth, and progression of the tumor [32]. In the case of our clusters, C2 was found to be downregulated compared to C3.

Limitations

Most of the limitations in this study come from limitations in the heatmap hierarchical clustering method of mutation data. Despite careful pre-processing, the heat map method of clustering is limited in that it can only see whether a sample has a mutation on a gene or doesn't, but it is unable to differentiate what kind of mutation. Scaling the onco-matrix before generating the heat map led to genes with much lower significance in literature being prioritized, so this method was disregarded. The mutation clusters having overlap with one another led to poorer mapping in the differential expression phase of the analysis, which meant all three clusters could not be compared to each other piece-wise as originally planned. This limitation meant it could only be suggested how expression of cluster 2 pathways were relative to clusters 1 and 3, and that results were perhaps not as meaningful as they would have been with better mapping and subsequently more piece-wise analysis run.

Conclusion

In our analysis, we aimed to uncover the different factors responsible for the growth and propagation of breast cancer using TCGA data. The dataset we investigated gave us a lot of insight into the different pathways involved with the proliferation of the tumors, with specific data about factors such as age, race, stage of disease, tumor subtypes, and variant types. We further examined the dataset to determine the pathways that are upregulated/downregulated the most within our genes. From that data we were able to cross-examine our findings with the literature, which matched our conclusions in most of the cases. The limitations related to heat map clustering were thoroughly investigated during the pre-processing stage to minimize. Using a different way of clustering could potentially yield better defined clusters which could help enhance the quality of our findings. Nonetheless, our conclusions could be used for a deeper investigation of treatment plans that could target gene expressions directly, effectively hindering a tumor's growth.

Contribution

Clinical analysis

- Elena did most of the clinical analysis

Mutation analysis

- All team members worked equally on mutation analysis

Gene expression analysis

- Nick did a majority of the expression analysis work

Final report

- Amine worked on the introduction
- Nick worked on the methods and results sections
- Amine and Elena worked on the Discussion
- Amine worked on the conclusion

Presentation

- All team members worked equally on the presentation

References

1. Giaquinto, A. N., Sung, H., Miller, K. D., Kramer, J. L., Newman, L. A., Minihan, A., et al. (2022). Breast cancer statistics, 2022. CA: a cancer journal for clinicians, 72(6), 524-541. Doi: 10.3322/caac.21754
2. Hu, Z., Fan, C., Oh, D. S., Marron, J. S., He, X., Qaqish, B. F., ... & Perou, C. M. (2006). The molecular portraits of breast tumors are conserved across microarray platforms. BMC genomics, 7(1), 1-12. Doi: 10.1186/1471-2164-7-96
3. Orrantia-Borunda, E., Anchondo-Nuñez, P., Acuña-Aguilar, et al. (2022). Subtypes of breast cancer. *Breast Cancer [Internet]*. Link: https://www.ncbi.nlm.nih.gov/books/NBK583808/
4. Konduri, S., Singh, M., Bobustuc, G., Rovin, R., & Kassam, A. (2020). Epidemiology of male breast cancer. The Breast, 54, 8-14. Link: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7476060/
5. Yu, A. Y. L., Thomas, S. M., DiLalla, G. D., Greenup, R. A., Hwang, E. S., Hyslop, T., ... & Fayanju, O. M. (2022). Disease characteristics and mortality among Asian women with breast cancer. Cancer, 128(5), 1024-1037. Doi: 10.1002/cncr.34015
6. Telli, M. L., Chang, E. T., Kurian, A. W., Keegan, T. H., McClure, L. A., Lichtensztajn, D., ... & Gomez, S. L. (2011). Asian ethnicity and breast cancer subtypes: a study from the California Cancer Registry. Breast cancer research and treatment, 127, 471-478. Doi: 10.1007/s10549-010-1173-8
7. Kostecka, A., Nowikiewicz, T., Olszewski, et al(2022). High prevalence of somatic PIK3CA and TP53 pathogenic variants in the normal mammary gland tissue of sporadic breast cancer patients revealed by duplex sequencing. NPJ Breast Cancer, 8(1), 76. Doi: 10.1038/s41523-022-00443-9
8. E Tamura, R., F de Vasconcellos, J., Sarkar, D., A Libermann, T., B Fisher, P., & F Zerbini, L. (2012). GADD45 proteins: central players in tumorigenesis. Current molecular medicine, 12(5), 634-651. Doi: 10.2174/156652412800619978
9. Wang, J., Wang, Y., Long, F., Yan, F., Wang, N., & Wang, Y. (2018). The expression and clinical significance of GADD45A in breast cancer patients. PeerJ, 6, e5344. Doi: 10.7717/peerj.5344
10. Sofi, S., Mehraj, U., Qayoom, H., Aisha, S., et al (2022). Targeting cyclin-dependent kinase 1 (CDK1) in cancer: molecular docking and dynamic simulations of potential CDK1 inhibitors. Medical Oncology, 39(9), 133. Doi: 10.1007/s12032-022-01748-2
11. Wolf, N. K., Kissiov, D. U., & Raulet, D. H. (2023). Roles of natural killer cells in immunity to cancer, and applications to immunotherapy. Nature Reviews Immunology, 23(2), 90-105. Doi: 10.1038/s41577-022-00732-1
12. Li, F., & Liu, S. (2022). Focusing on NK cells and ADCC: a promising immunotherapy approach in targeted therapy for HER2-positive breast cancer. Frontiers in Immunology, 13, 1083462. Doi: 10.3389/fimmu.2022.1083462
13. Hematian Larki, M., Barani, S., Talei, A. R., & Ghaderi, A. (2020). Diversity of KIRs in invasive breast cancer patients and healthy controls along with the clinical significance in ER/PR/HER2+ patients. Genes & Immunity, 21(6-8), 380-389. Doi: 10.1038/s41435-020-00117-1
14. Oh, J. H., Lee, J. Y., Yu, S., Cho, Y., Hur, S., Nam, K. T., & Kim, M. H. (2019). RAE1 mediated ZEB1 expression promotes epithelial–mesenchymal transition in breast cancer. Scientific reports, 9(1), 2977. Doi: 10.1038/s41598-019-39574-8
15. Yang, K., Halima, A., & Chan, T. A. (2023). Antigen presentation in cancer—mechanisms and clinical implications for immunotherapy. Nature Reviews Clinical Oncology, 1-20. Doi: 10.1038/s41571-023-00789-4
16. Zhang, J., Chan, D. W., & Lin, S. Y. (2022). Exploiting DNA Replication Stress as a Therapeutic Strategy for Breast Cancer. Biomedicines, 10(11), 2775. Doi: 10.3390/biomedicines10112775

17. Du, H., Yang, X., Fan, J., & Du, X. (2021). Claudin 6: Therapeutic prospects for tumours, and mechanisms of expression and regulation. Molecular Medicine Reports, 24(3), 1-9. Doi: 10.3892/mmr.2021.12316
18. Salvador, E., Burek, M., & Förster, C. Y. (2016). Tight junctions and the tumor microenvironment. Current pathobiology reports, 4, 135-145. Doi: 10.1007/s40139-016-0106-6
19. Milosevic, B., Stojanovic, B., Cvetkovic, A., et al. (2023). The Enigma of Mammaglobin: Redefining the Biomarker Paradigm in Breast Carcinoma. International Journal of Molecular Sciences, 24(17), 13407. Doi: 10.3390/ijms241713407
20. Guvakova, M. A., Prabakaran, I., Wu, Z., Hoffman, D. et al (2020). CDH2/N-cadherin and early diagnosis of invasion in patients with ductal carcinoma in situ. Breast Cancer Research and Treatment, 183, 333-346. Doi: 10.1007/s10549-020-05797-x
21. Song, D., Lian, Y., & Zhang, L. (2023). The potential of activator protein 1 (AP-1) in cancer targeted therapy. Frontiers in Immunology, 14. Doi: 10.3389/fimmu.2023.1224892
22. Guo, Y., Arciero, C. A., Jiang, R.,et al (2020). Different breast cancer subtypes show different metastatic patterns: A study from a large public database. Asian Pacific Journal of Cancer Prevention: APJCP, 21(12), 3587. Doi: 10.31557/APJCP.2020.21.12.3587
23. Shostak, K., & Chariot, A. (2011). NF-κB, stem cells and breast cancer: the links get stronger. Breast Cancer Research, 13(4), 1-7. Doi: 10.1186/bcr2886
24. Banushi, B., Joseph, S. R., Lum, B., Lee, J. J., & Simpson, F. (2023). Endocytosis in cancer and cancer therapy. Nature Reviews Cancer, 1-24. Doi: 10.1038/s41568-023-00574-6
25. Pham, D. V., & Park, P. H. (2022). Adiponectin triggers breast cancer cell death via fatty acid metabolic reprogramming. Journal of Experimental & Clinical Cancer Research, 41(1), 1-20. Doi: 10.1186/s13046-021-02223-y
26. Monaco, M. E. (2017). Fatty acid metabolism in breast cancer subtypes. Oncotarget, 8(17), 29487. Doi: 10.18632/oncotarget.15494
27. Liu, Y., Guo, S., Sun, Y., Zhang, C., Gan, J., Ning, S., & Wang, J. (2023). CRS: a circadian rhythm score model for predicting prognosis and treatment response in cancer patients. Journal of Translational Medicine, 21(1), 1-16. Doi: 10.1186/s12967-023-04013-w
28. Don, Lin, H.-H., Furtado, J. J., Maan Qraitem, Taylor, S. S., & Farkas, M. E. (2019). Circadian oscillations persist in low malignancy breast cancer cells. Cell Cycle, 18(19), 2447–2453. Doi: 10.1080/15384101.2019.1648957
29. Luo, M., & Guan, J. L. (2010). Focal adhesion kinase: a prominent determinant in breast cancer initiation, progression and metastasis. Cancer letters, 289(2), 127-139. Doi: 10.1016/j.canlet.2009.07.005
30. Ye, D. Z., & Field, J. (2012). PAK signaling in cancer. Cellular logistics, 2(2), 105-116. Doi: 10.4161/cl.21882
31. Eom, Y. H., Kim, H. S., Lee, A., Song, B. J., & Chae, B. J. (2016). BCL2 as a Subtype-Specific Prognostic Marker for Breast Cancer. Journal of Breast Cancer, 19(3), 252. Doi: 10.4048/jbc.2016.19.3.252
32. Er, X., Ji, B., Jin, K., & Chen, Y. (2023). Branched-chain amino acids catabolism and cancer progression: focus on therapeutic interventions. Frontiers in Oncology, 13. Doi: 10.3389/fonc.2023.1220638

Appendix A: Tables and figures

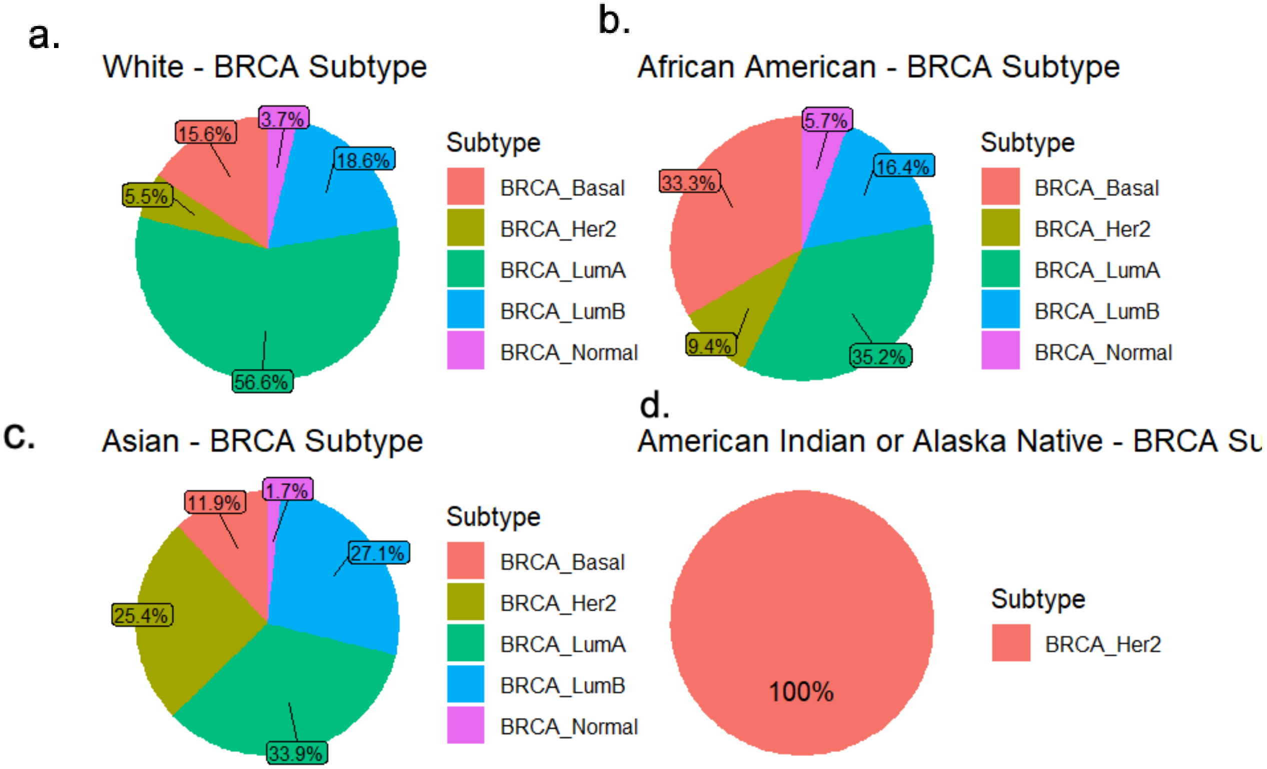


Figure 1: Pie charts of subtype distributions within: *a.* White participants. *b.*Black/African American participants, *c.* Asian participants, *d.* American Indian or Alaskan Native participants

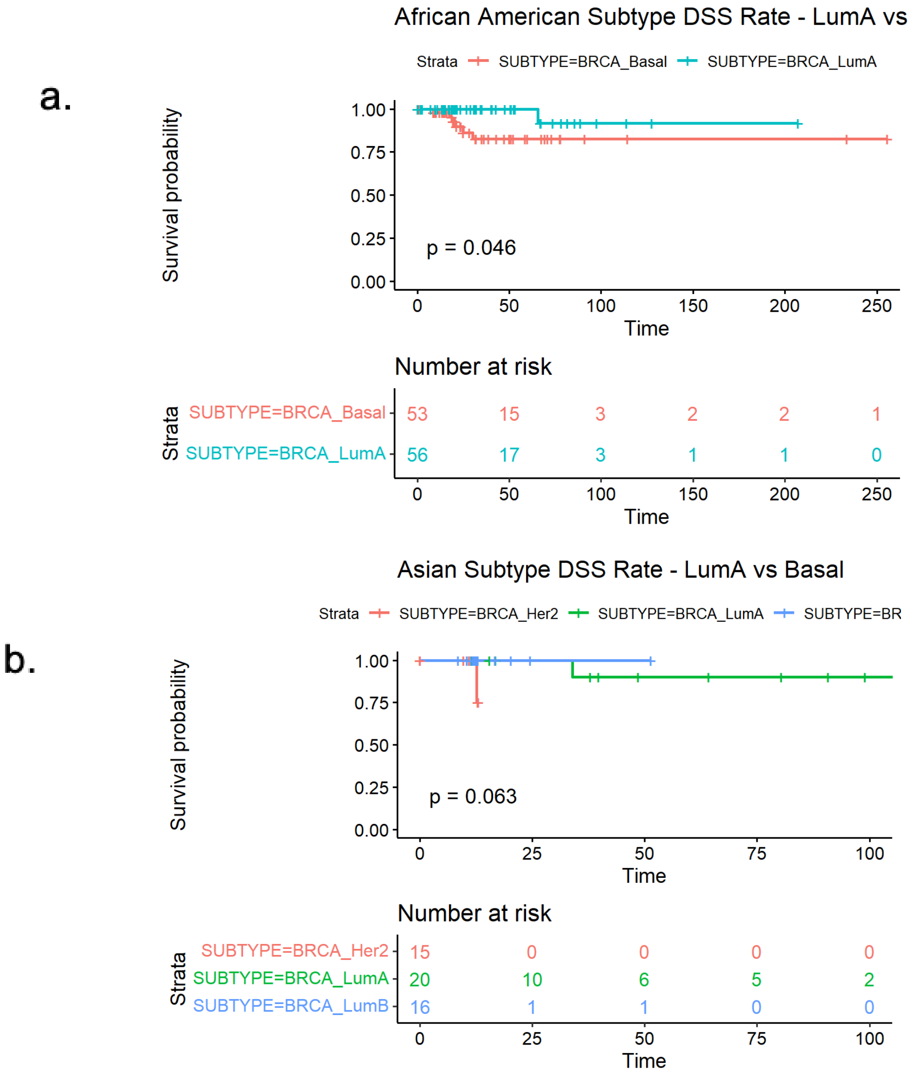


Figure 2: Survival analysis based on subtypes for: *a.* White, *b.*Black/African American, *c.*Asian participants

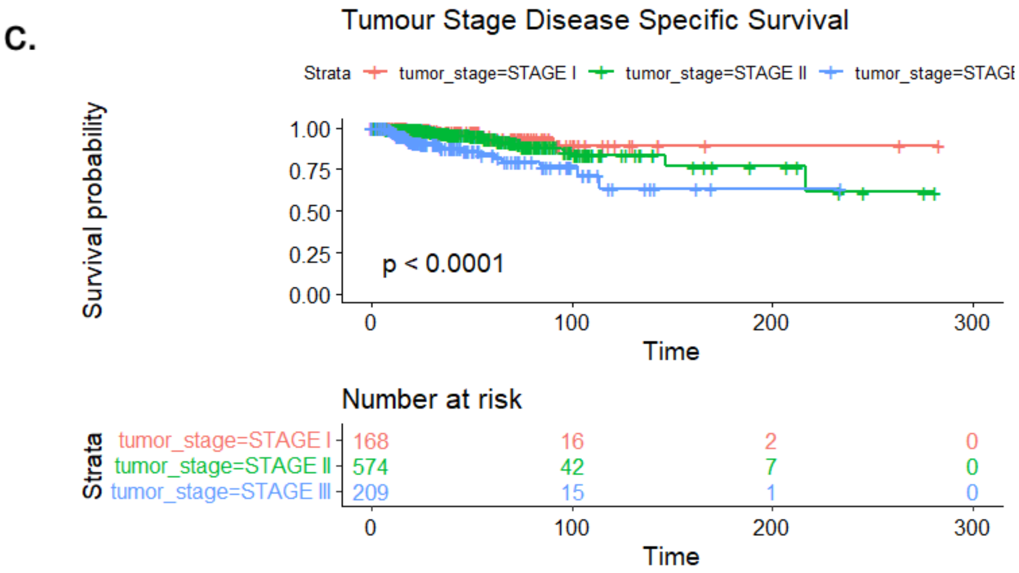


Figure 3: Survival analysis based on tumor state using disease specific survival status

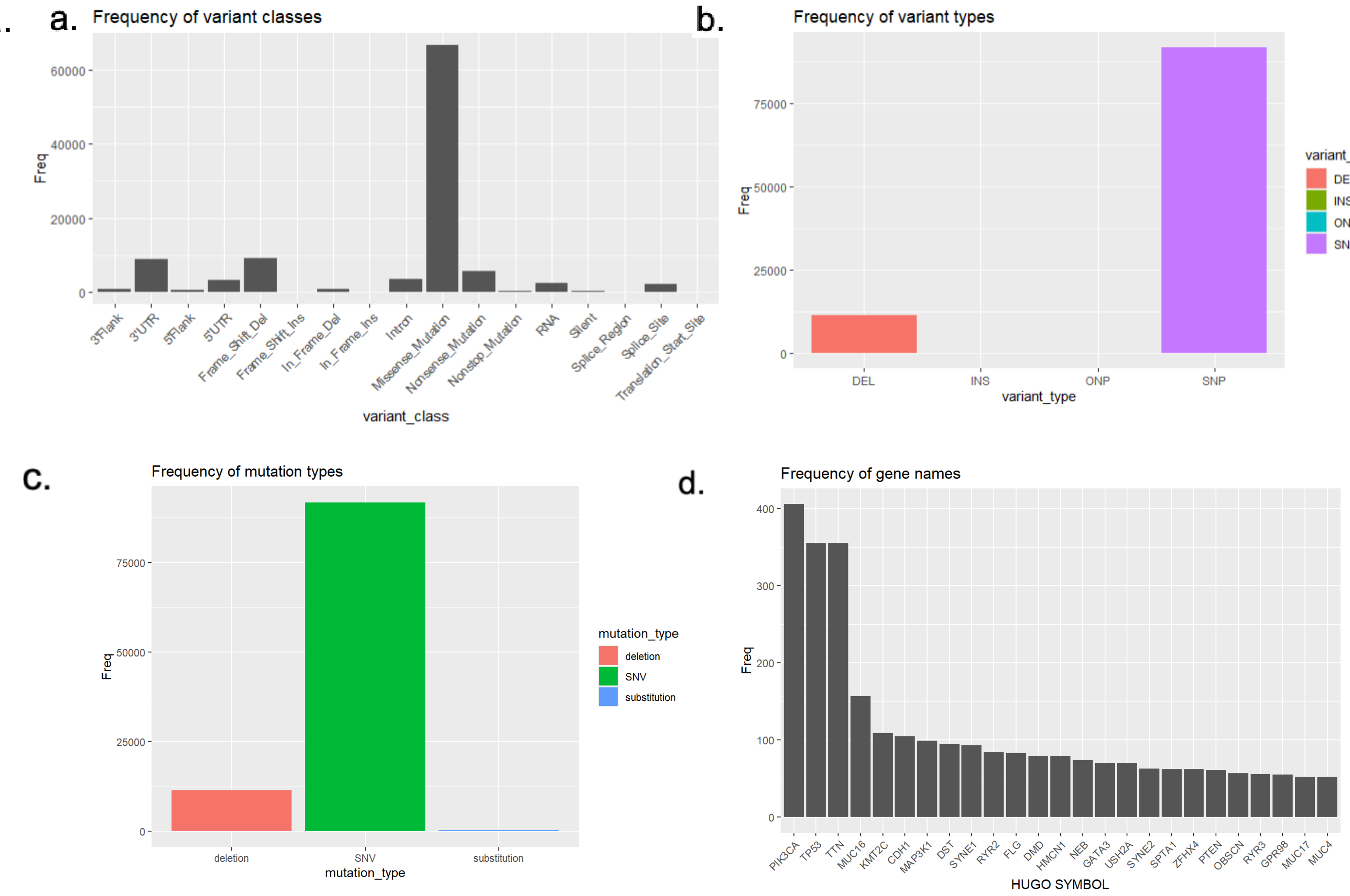


Fig 4. Plots representing the frequency of: *a.* Classes of Variants, *b.* Types of Variants, *c.* Types of Mutations, *d.* Mutated Genes

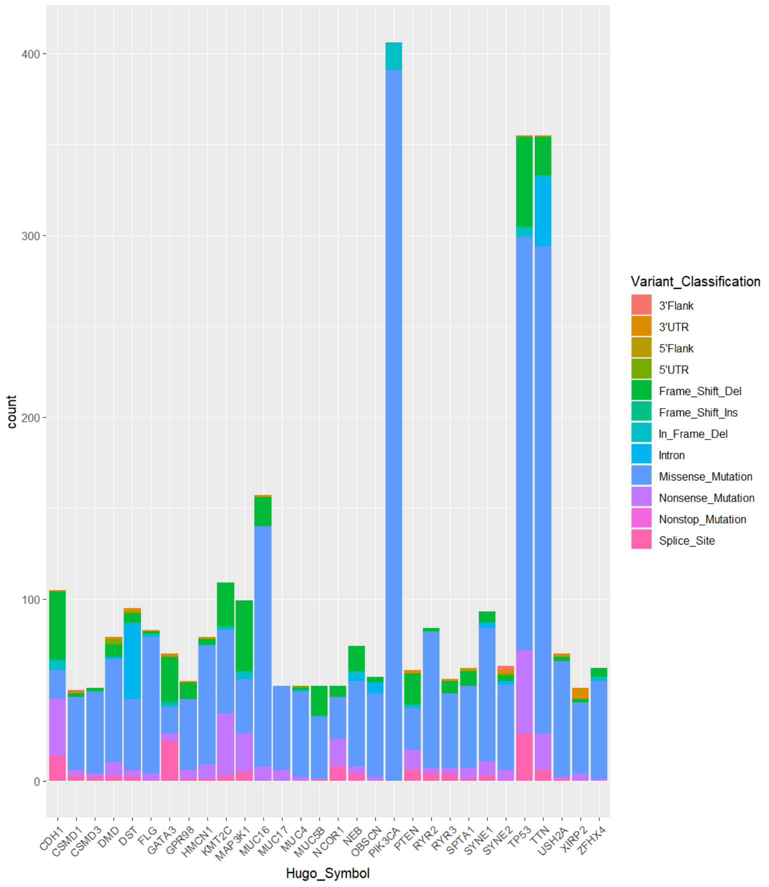


Fig 5. Bar Chart of the mutations occurring in each gene

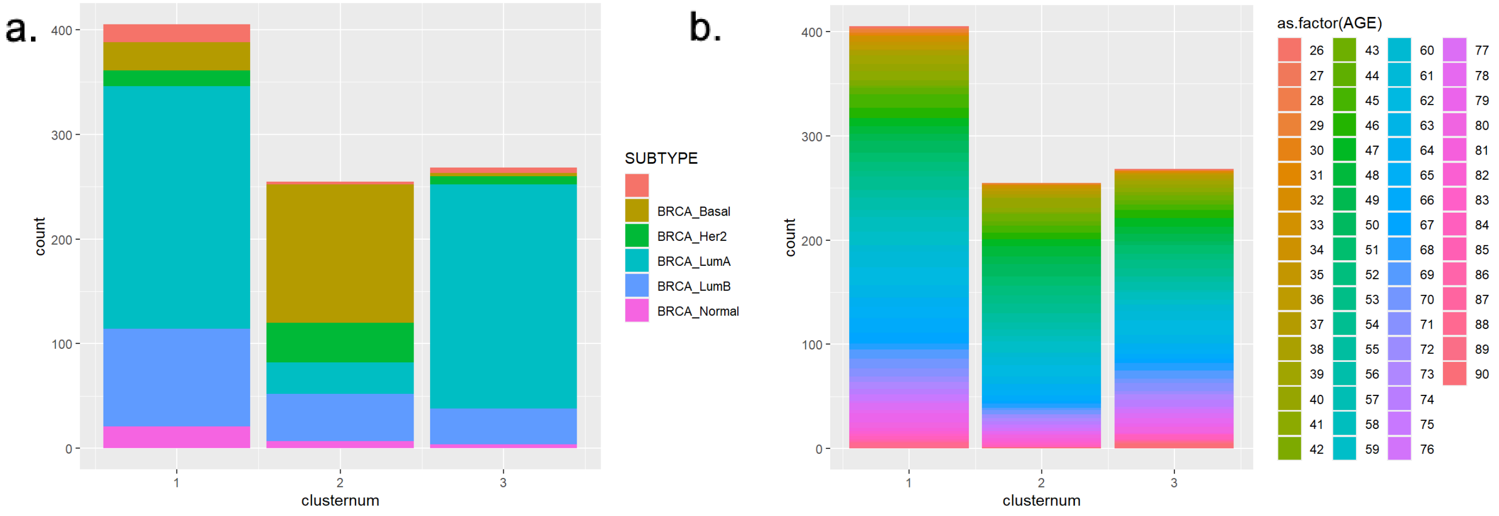


Fig 6. Cluster composition barchart: *a.* Subtypes in each cluster, *b:* Age distribution in each cluster

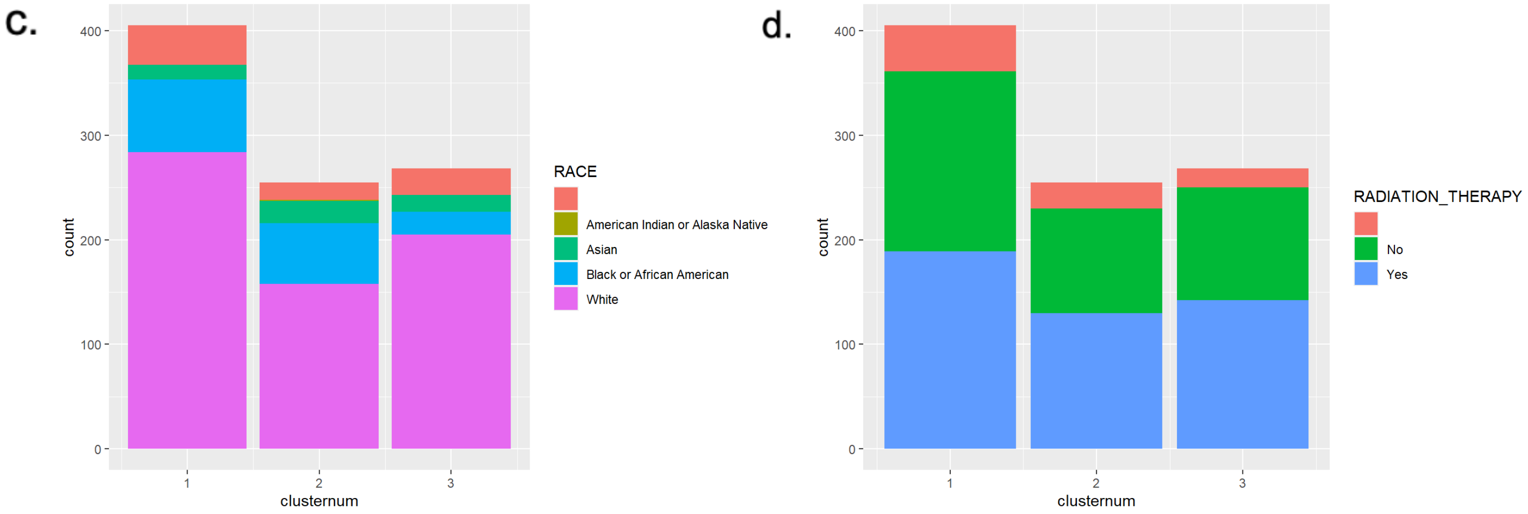


Fig 6. Cluster composition barchart: *c.* Races in each cluster ,*d:* Presence of radiation therapy in each cluster

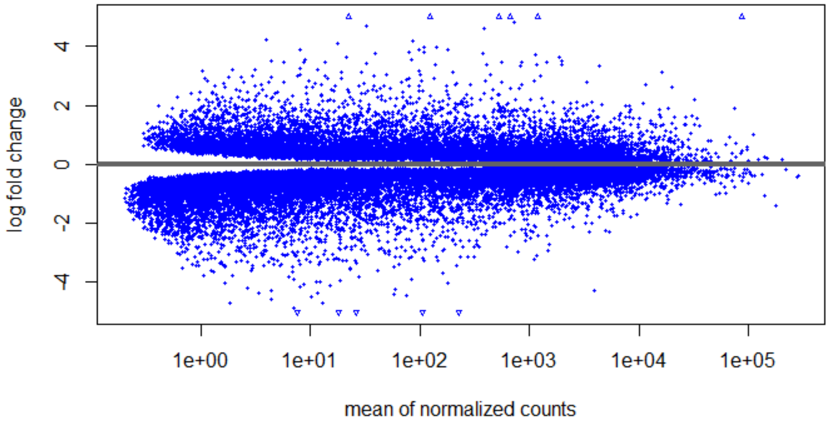


Fig 7. MA plot based on significant results (padj < 0.05) of differential analysis

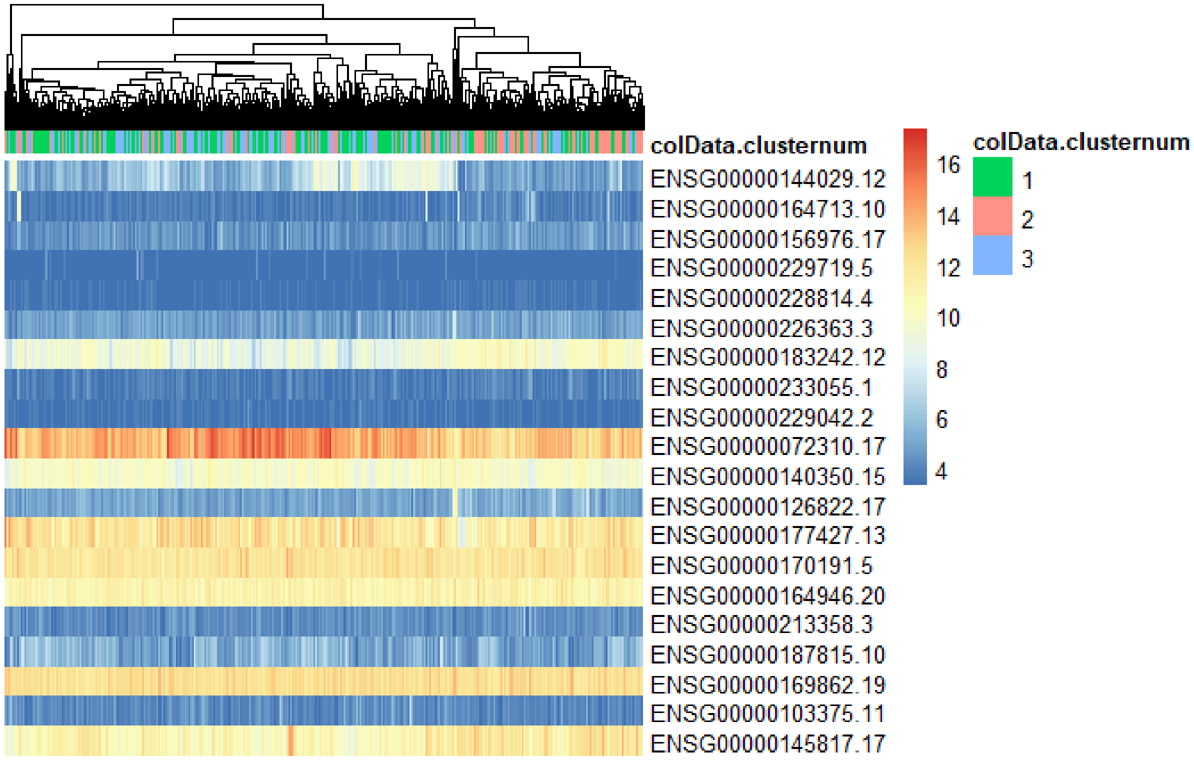
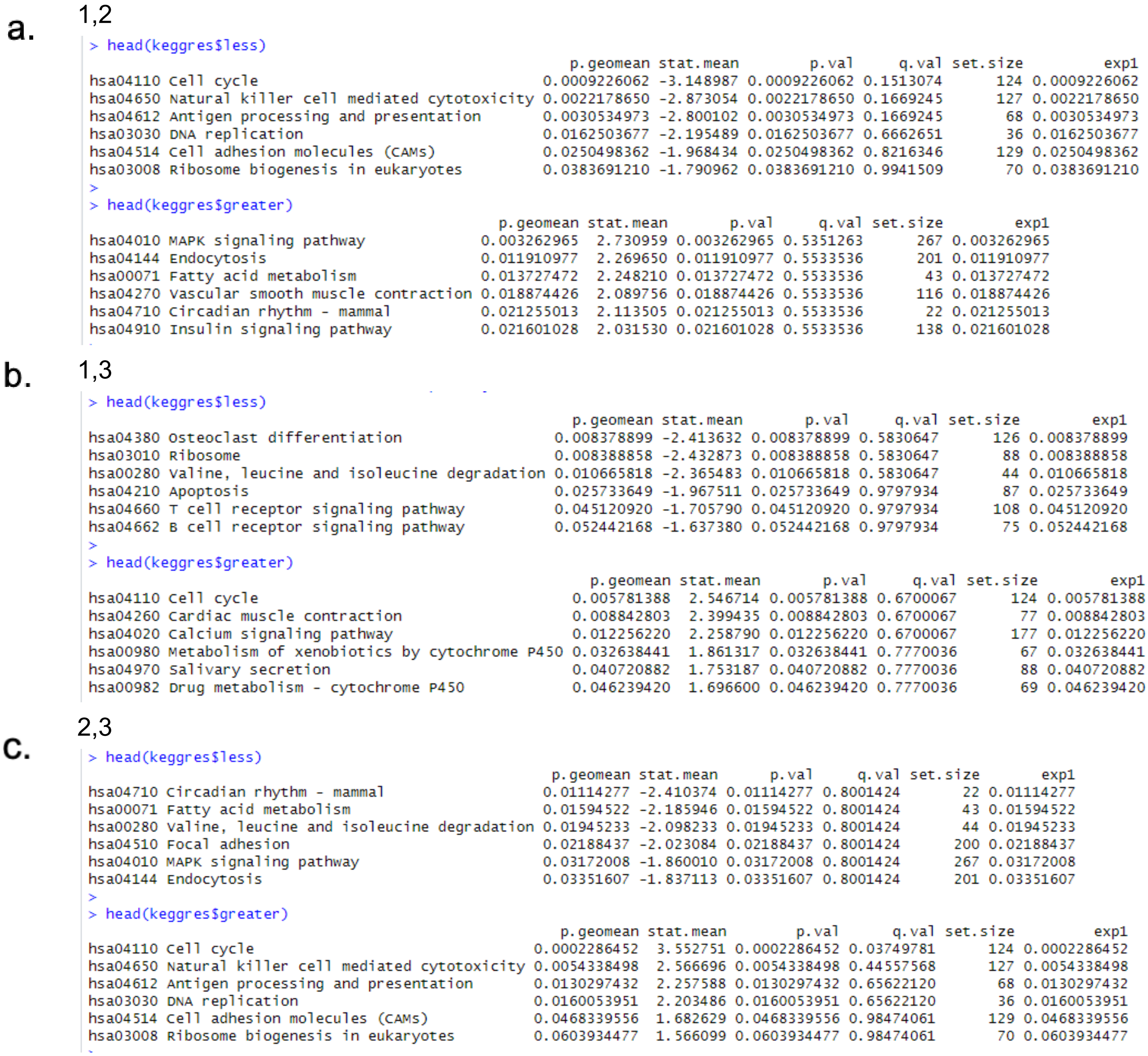


Fig 8: Heatmap of top 10 downregulated genes and top 10 upregulated genes



a. 1,2

```
> head(keggres$less)
                                             p.geomean stat.mean        p.val     q.val set.size         exp1
hsa04110 Cell cycle                       0.0009226062 -3.148987 0.0009226062 0.1513074      124 0.0009226062
hsa04650 Natural killer cell mediated cytotoxicity 0.0022178650 -2.873054 0.0022178650 0.1669245      127 0.0022178650
hsa04612 Antigen processing and presentation      0.0030534973 -2.800102 0.0030534973 0.1669245       68 0.0030534973
hsa03030 DNA replication                          0.0162503677 -2.195489 0.0162503677 0.6662651       36 0.0162503677
hsa04514 Cell adhesion molecules (CAMs)           0.0250498362 -1.968434 0.0250498362 0.8216346      129 0.0250498362
hsa03008 Ribosome biogenesis in eukaryotes        0.0383691210 -1.790962 0.0383691210 0.9941509       70 0.0383691210
>
> head(keggres$greater)
                                          p.geomean stat.mean       p.val     q.val set.size        exp1
hsa04010 MAPK signaling pathway           0.003262965  2.730959 0.003262965 0.5351263      267 0.003262965
hsa04144 Endocytosis                      0.011910977  2.269650 0.011910977 0.5533536      201 0.011910977
hsa00071 Fatty acid metabolism            0.013727472  2.248210 0.013727472 0.5533536       43 0.013727472
hsa04270 Vascular smooth muscle contraction 0.018874426  2.089756 0.018874426 0.5533536      116 0.018874426
hsa04710 Circadian rhythm - mammal        0.021255013  2.113505 0.021255013 0.5533536       22 0.021255013
hsa04910 Insulin signaling pathway        0.021601028  2.031530 0.021601028 0.5533536      138 0.021601028
```

b. 1,3

```
> head(keggres$less)
                                                p.geomean stat.mean       p.val     q.val set.size        exp1
hsa04380 Osteoclast differentiation             0.008378899 -2.413632 0.008378899 0.5830647      126 0.008378899
hsa03010 Ribosome                               0.008388858 -2.432873 0.008388858 0.5830647       88 0.008388858
hsa00280 Valine, leucine and isoleucine degradation 0.010665818 -2.365483 0.010665818 0.5830647       44 0.010665818
hsa04210 Apoptosis                              0.025733649 -1.967511 0.025733649 0.9797934       87 0.025733649
hsa04660 T cell receptor signaling pathway      0.045120920 -1.705790 0.045120920 0.9797934      108 0.045120920
hsa04662 B cell receptor signaling pathway      0.052442168 -1.637380 0.052442168 0.9797934       75 0.052442168
>
> head(keggres$greater)
                                                   p.geomean stat.mean       p.val     q.val set.size        exp1
hsa04110 Cell cycle                                0.005781388  2.546714 0.005781388 0.6700067      124 0.005781388
hsa04260 Cardiac muscle contraction                0.008842803  2.399435 0.008842803 0.6700067       77 0.008842803
hsa04020 Calcium signaling pathway                 0.012256220  2.258790 0.012256220 0.6700067      177 0.012256220
hsa00980 Metabolism of xenobiotics by cytochrome P450 0.032638441  1.861317 0.032638441 0.7770036       67 0.032638441
hsa04970 Salivary secretion                        0.040720882  1.753187 0.040720882 0.7770036       88 0.040720882
hsa00982 Drug metabolism - cytochrome P450         0.046239420  1.696600 0.046239420 0.7770036       69 0.046239420
```

c. 2,3

```
> head(keggres$less)
                                               p.geomean stat.mean      p.val     q.val set.size       exp1
hsa04710 Circadian rhythm - mammal             0.01114277 -2.410374 0.01114277 0.8001424       22 0.01114277
hsa00071 Fatty acid metabolism                 0.01594522 -2.185946 0.01594522 0.8001424       43 0.01594522
hsa00280 Valine, leucine and isoleucine degradation 0.01945233 -2.098233 0.01945233 0.8001424       44 0.01945233
hsa04510 Focal adhesion                        0.02188437 -2.023084 0.02188437 0.8001424      200 0.02188437
hsa04010 MAPK signaling pathway                0.03172008 -1.860010 0.03172008 0.8001424      267 0.03172008
hsa04144 Endocytosis                           0.03351607 -1.837113 0.03351607 0.8001424      201 0.03351607
>
> head(keggres$greater)
                                                p.geomean stat.mean        p.val      q.val set.size         exp1
hsa04110 Cell cycle                          0.0002286452  3.552751 0.0002286452 0.03749781      124 0.0002286452
hsa04650 Natural killer cell mediated cytotoxicity 0.0054338498  2.566696 0.0054338498 0.44557568      127 0.0054338498
hsa04612 Antigen processing and presentation 0.0130297432  2.257588 0.0130297432 0.65622120       68 0.0130297432
hsa03030 DNA replication                     0.0160053951  2.203486 0.0160053951 0.65622120       36 0.0160053951
hsa04514 Cell adhesion molecules (CAMs)      0.0468339556  1.682629 0.0468339556 0.98474061      129 0.0468339556
hsa03008 Ribosome biogenesis in eukaryotes   0.0603934477  1.566099 0.0603934477 0.98474061       70 0.0603934477
```

Figure 9: top 5 upregulated and down regulated pathways for pairs *a.*1,2; *b.*1,3; *c.*2,3

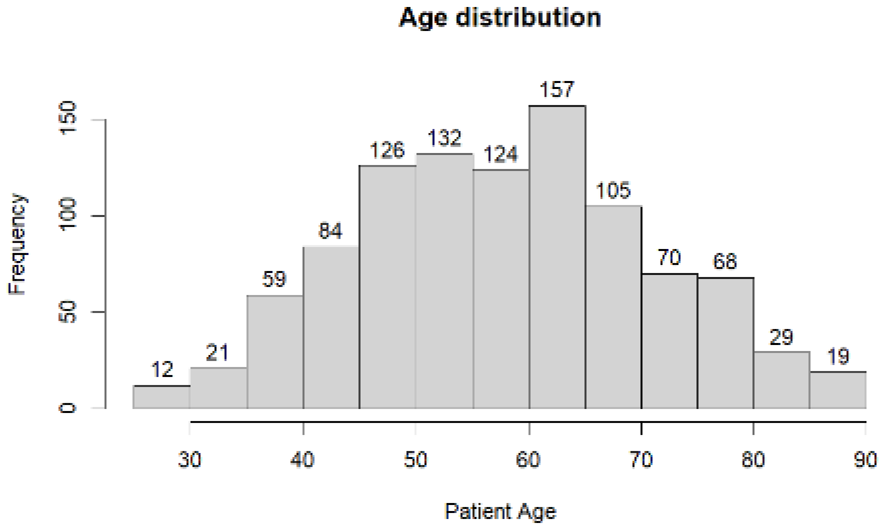


Figure 10: Age distribution of patients

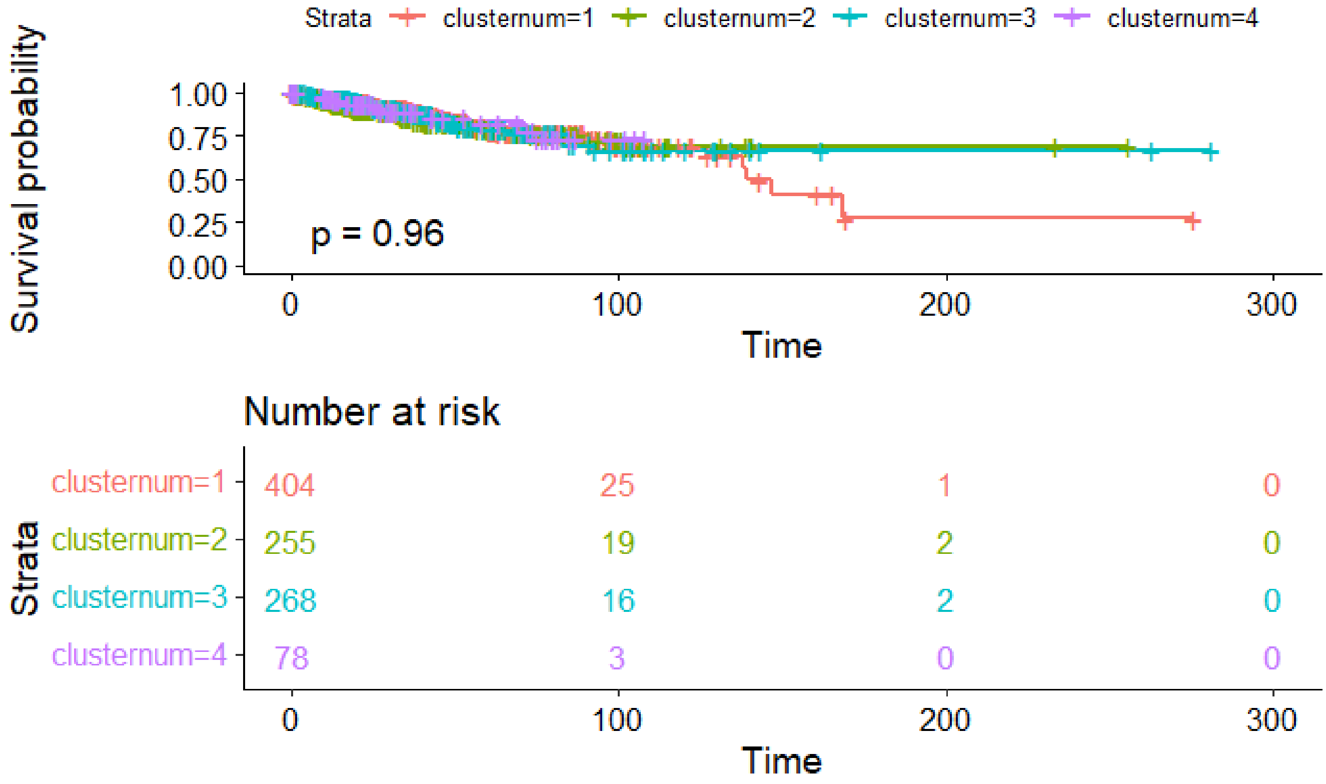


Figure 11: Kaplan-Meier plot with 4 clusters

1,2 down regulated pathways

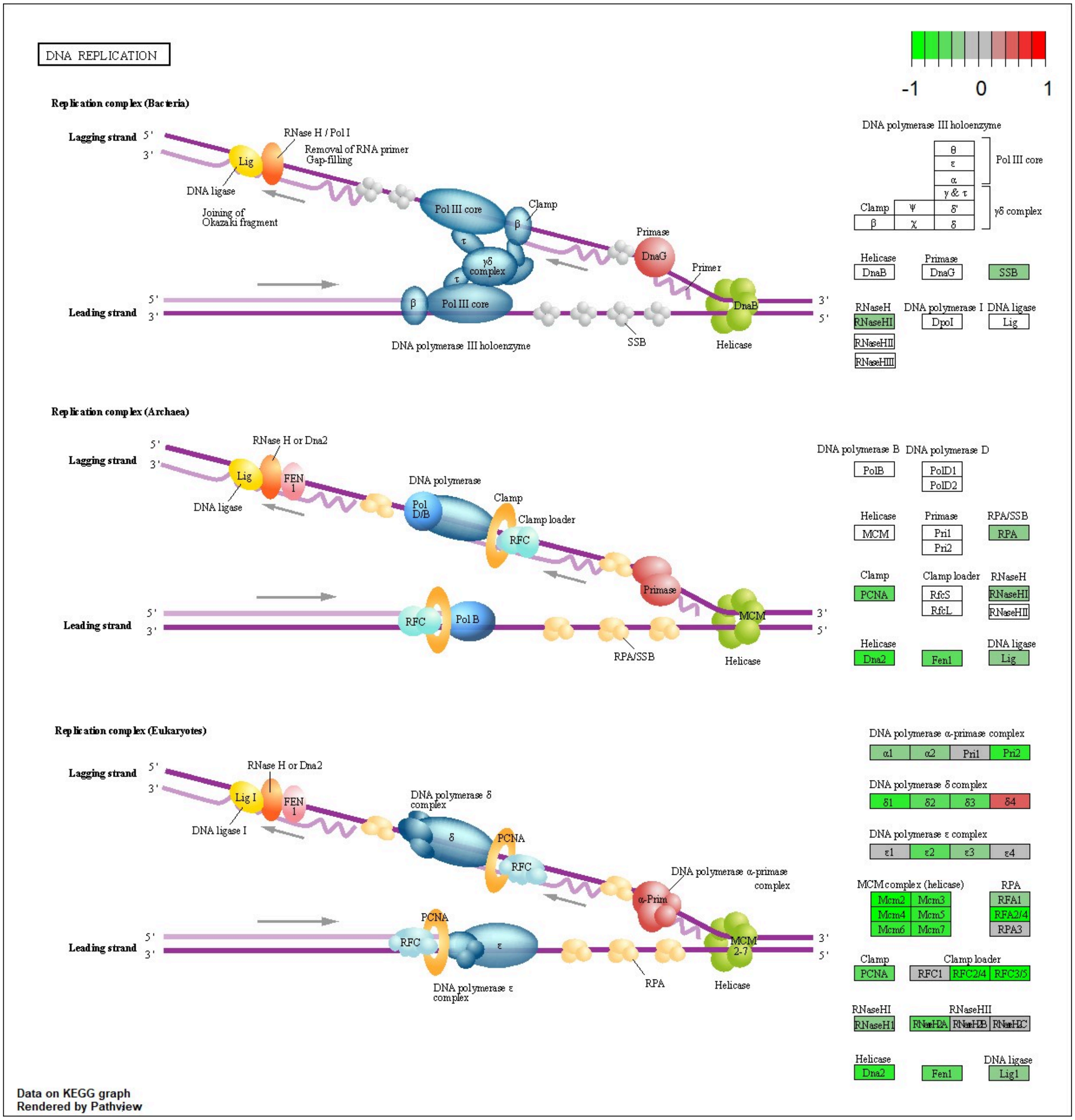


Figure 12 DNA replication pathway

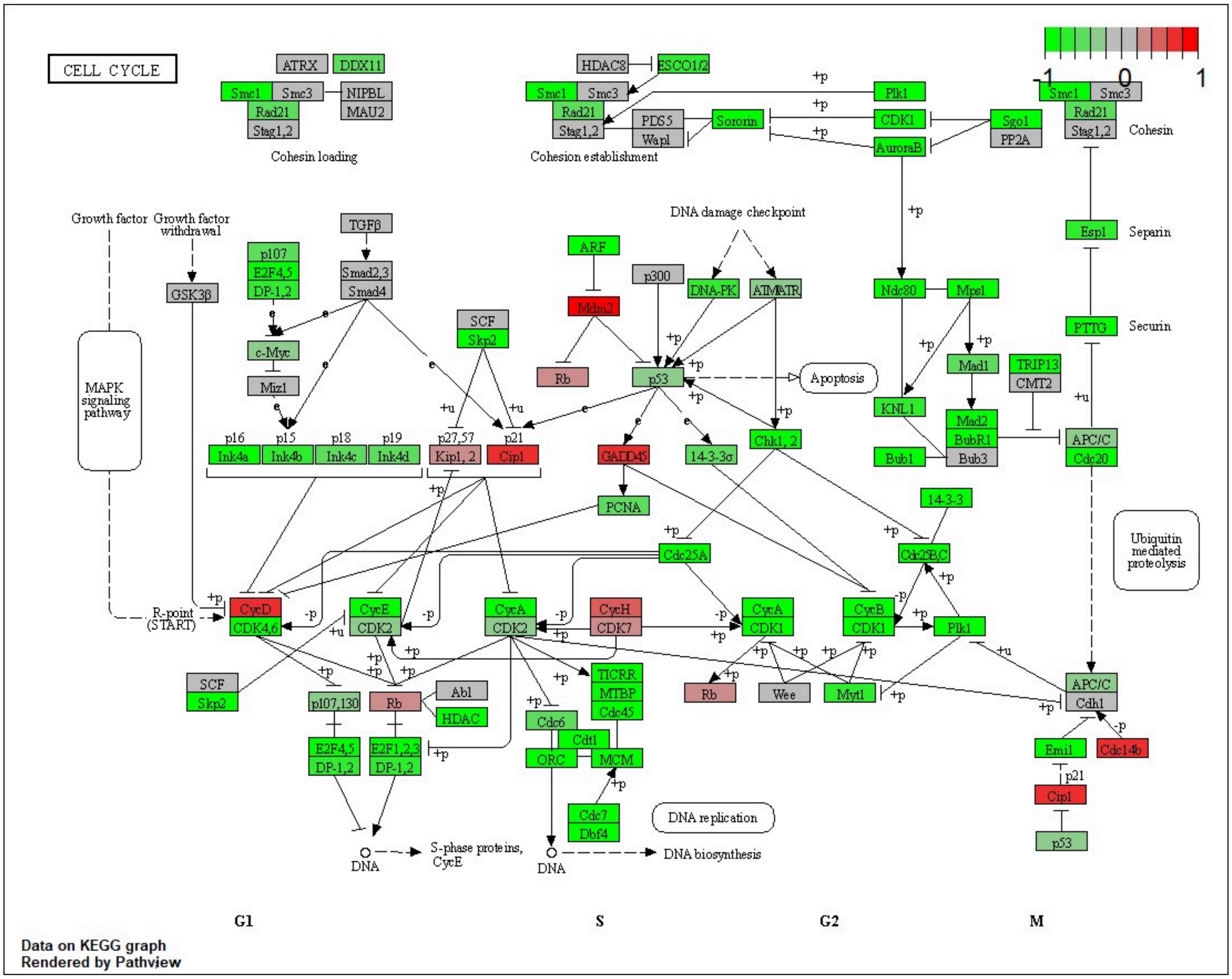


Figure 13 Cell cycle pathway

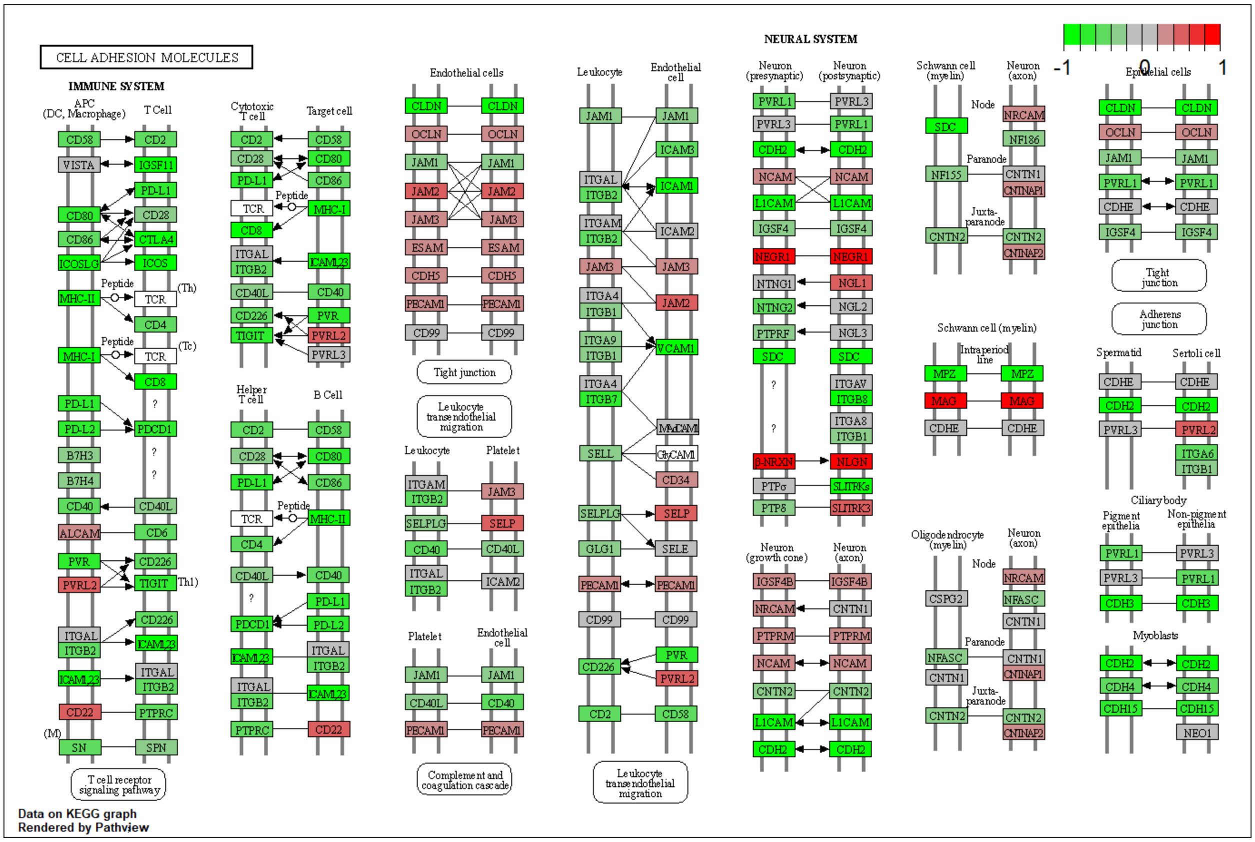


Figure 14 Cell adhesion molecules pathway

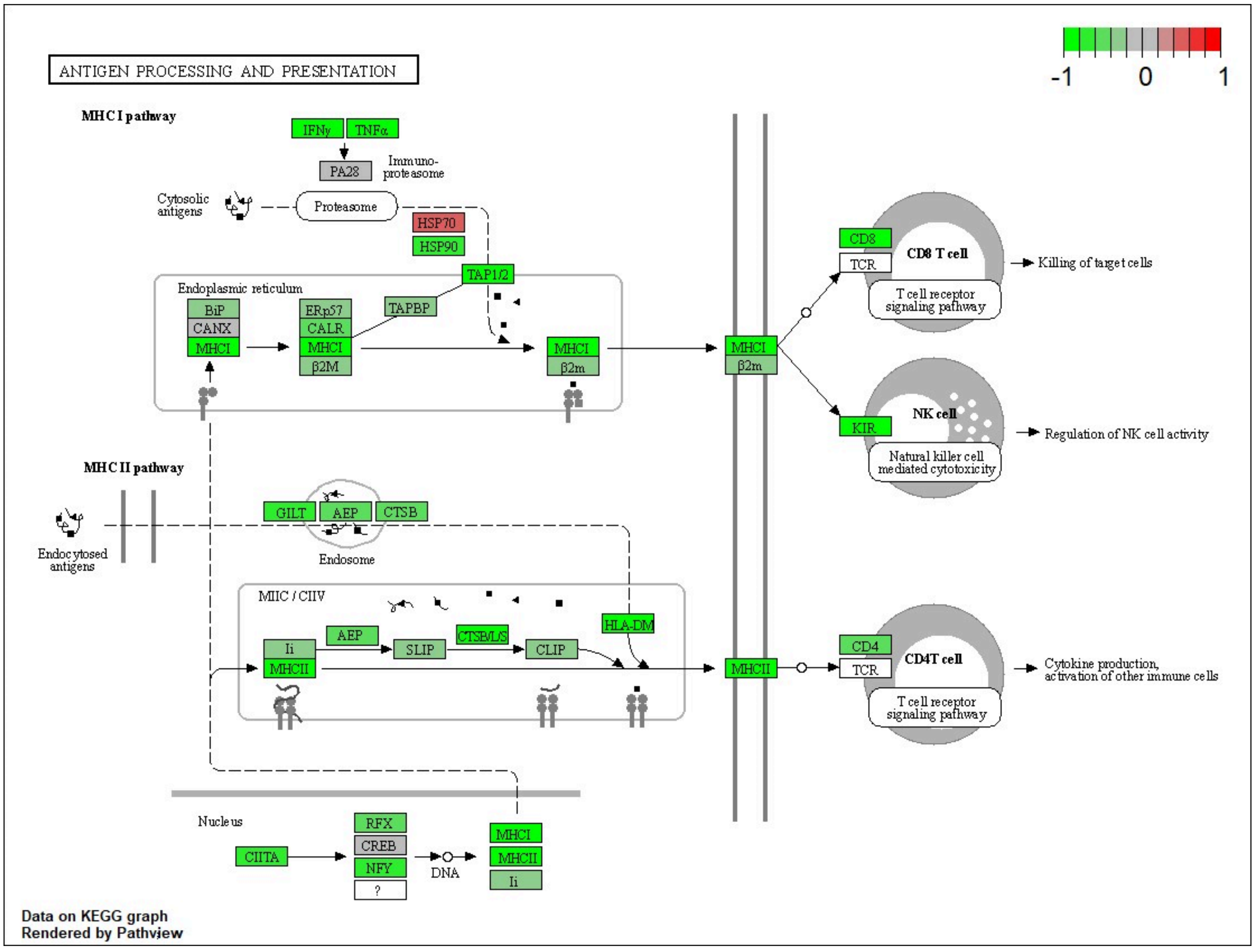


Figure 15 Antigen processing and presentation pathway

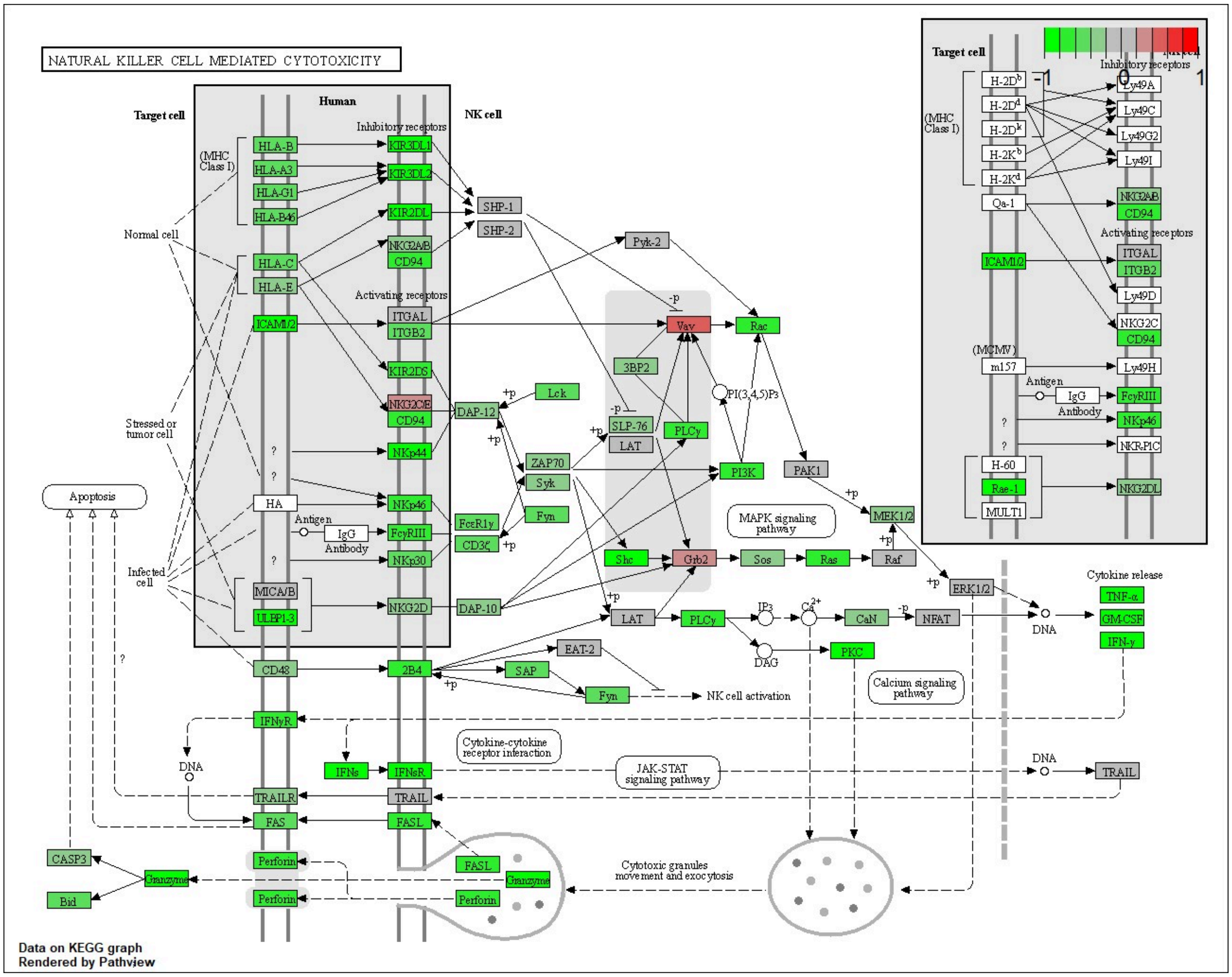


Figure 16 Natural killer cell mediated cytotoxicity pathway

1,2 upregulated pathways

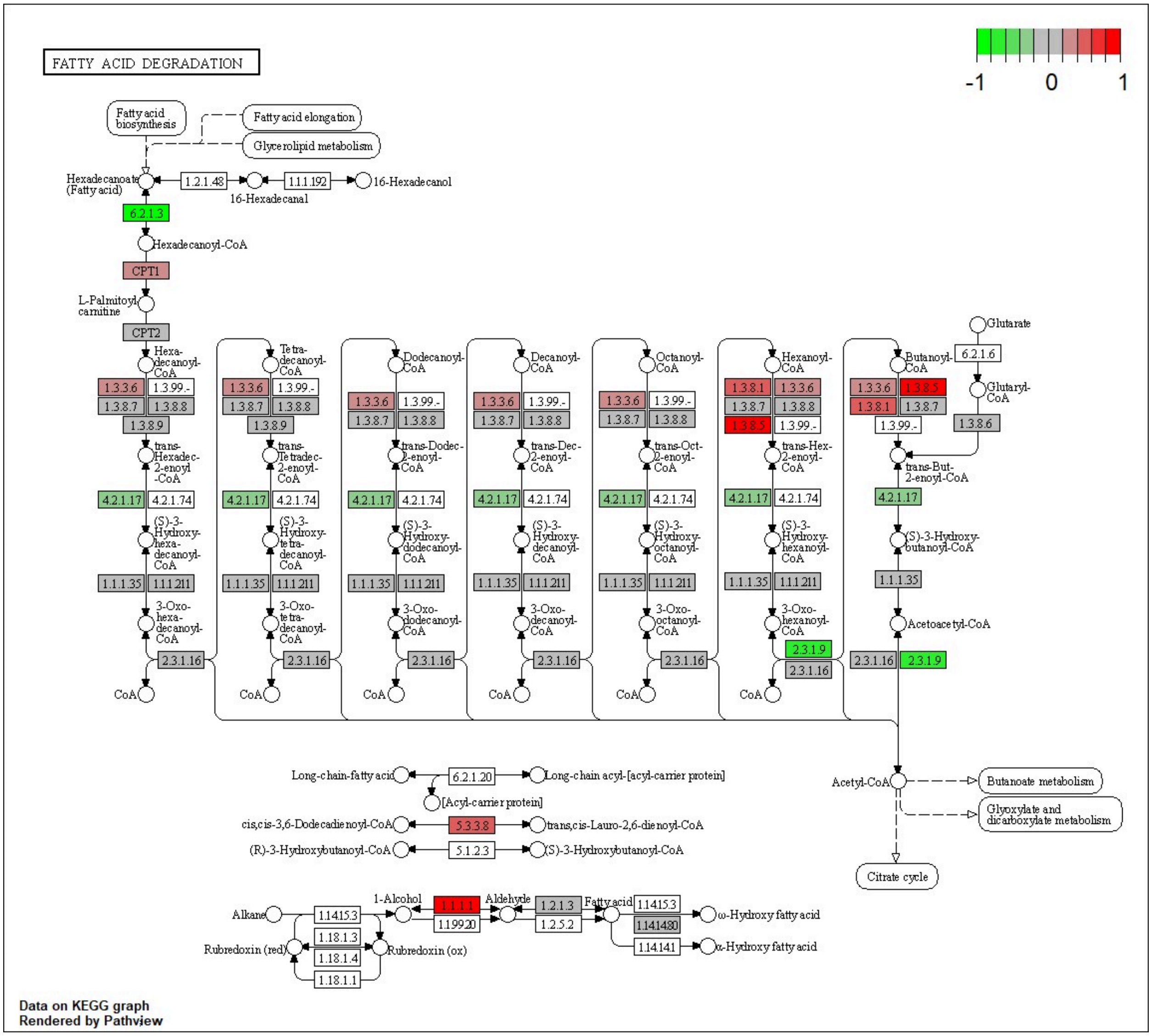


Figure 17 fatty acid degradation pathway

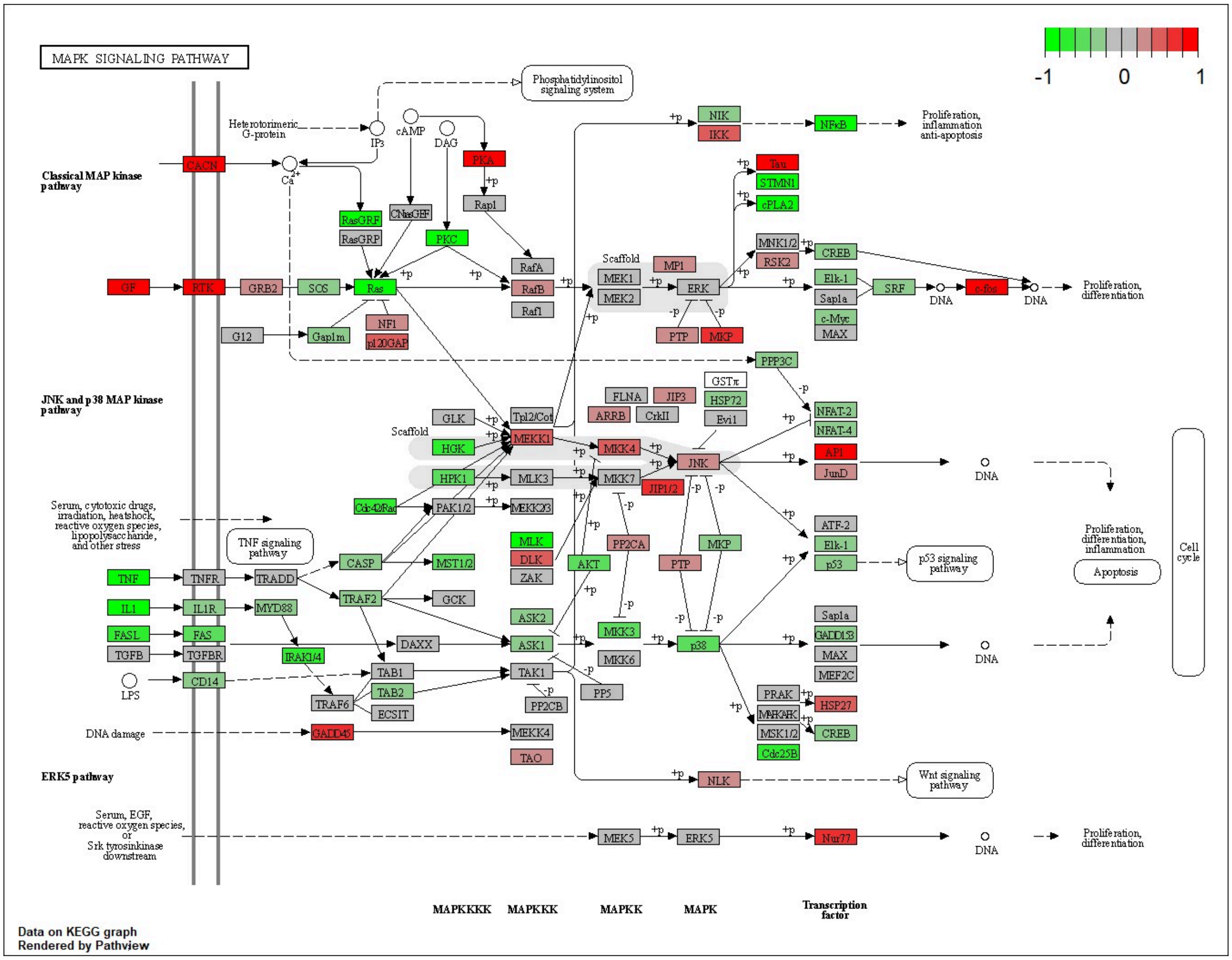


Figure 18 MAPK signaling pathway

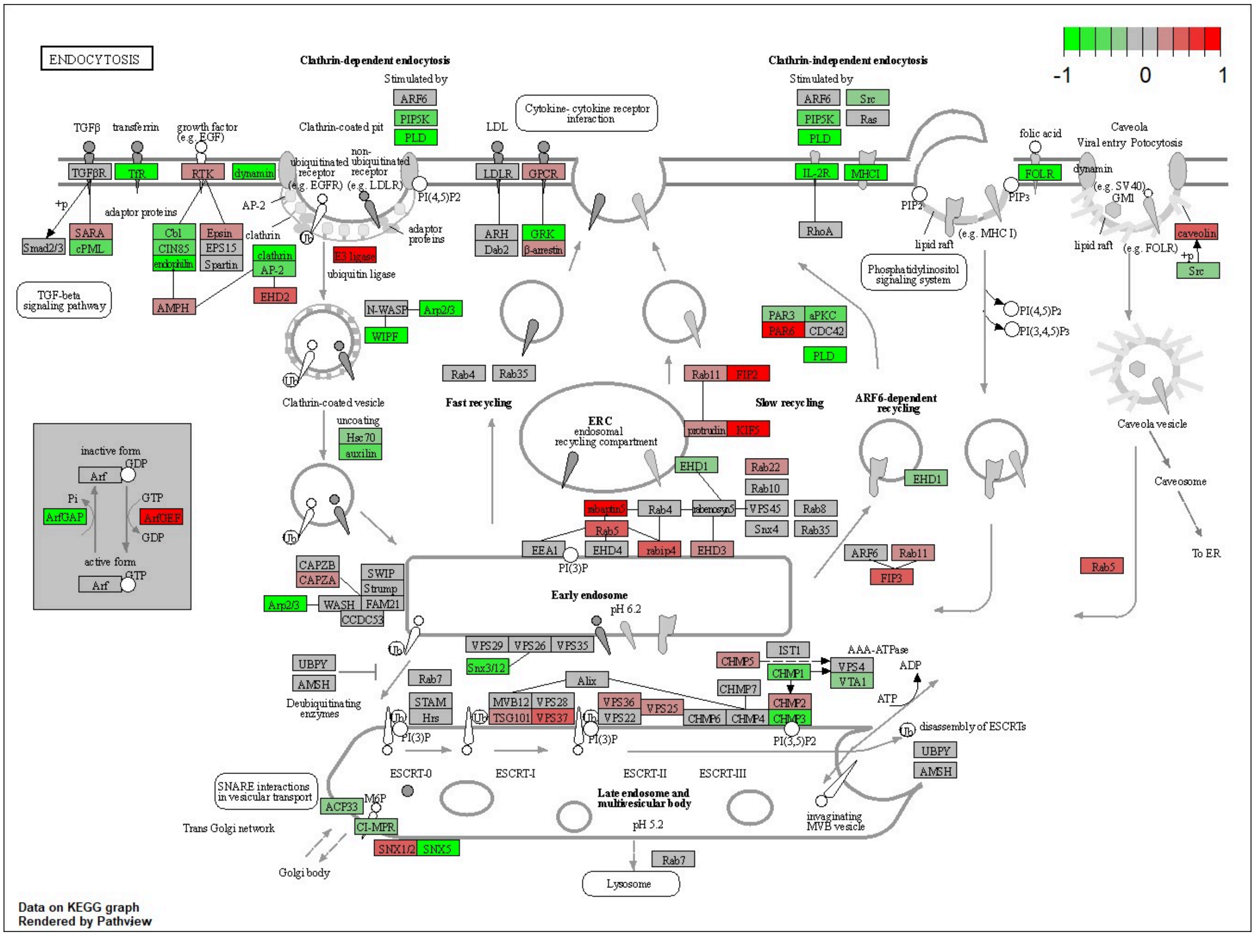


Figure 19 endocytosis pathway

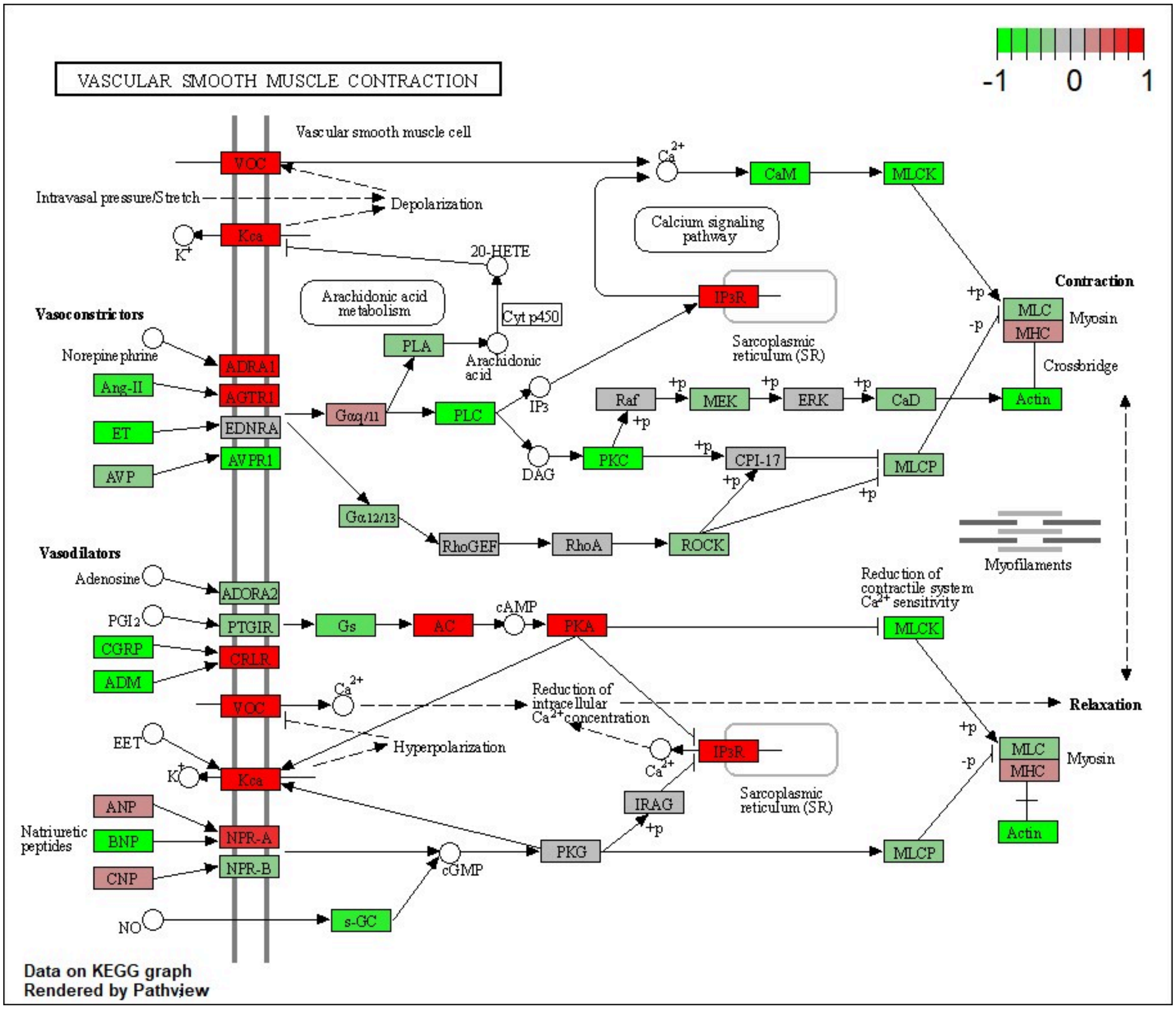


Figure 20 vascular smooth muscle contraction pathway

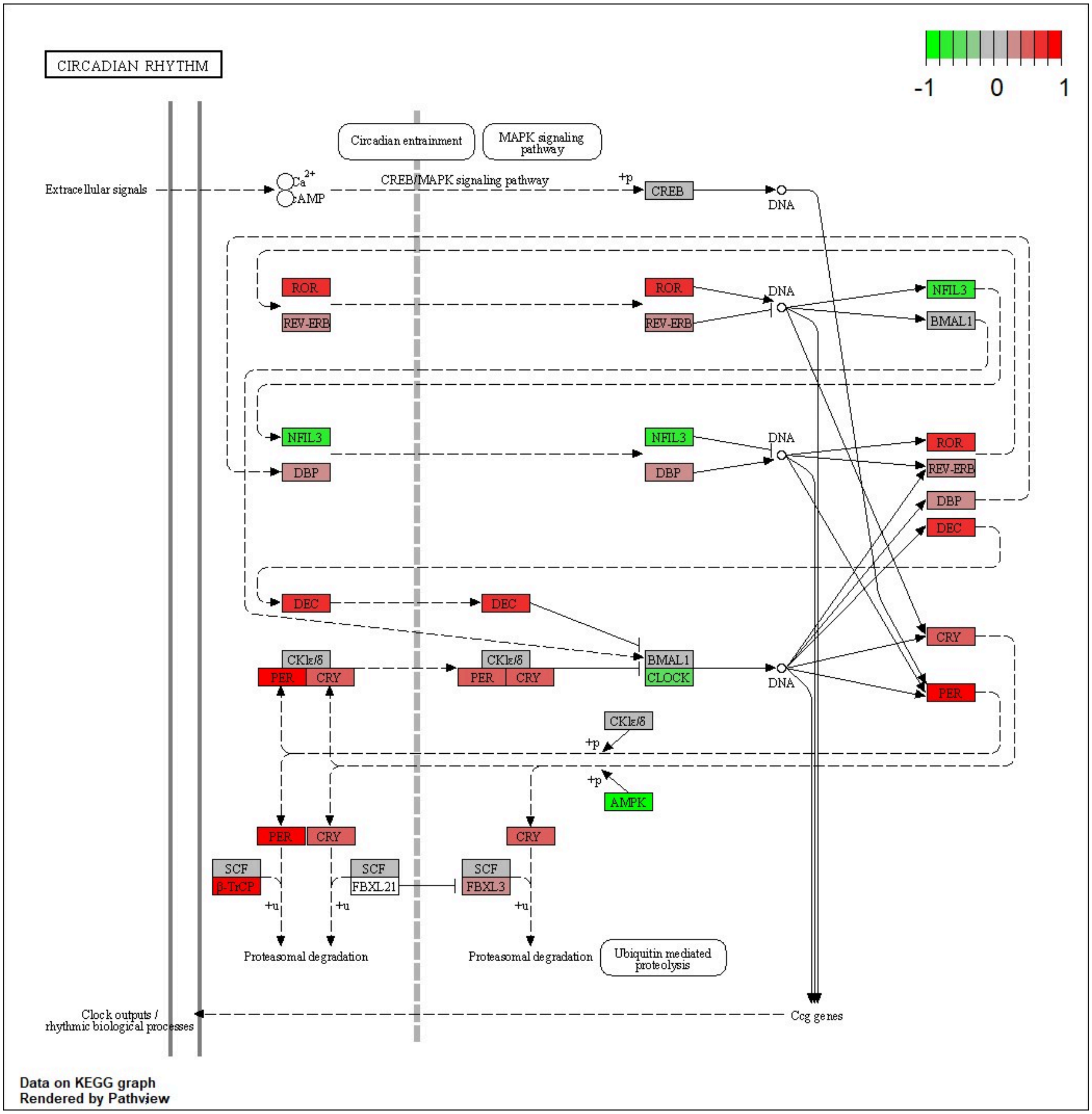


Figure 21 circadian rhythm pathway

2,3 downregulated pathways

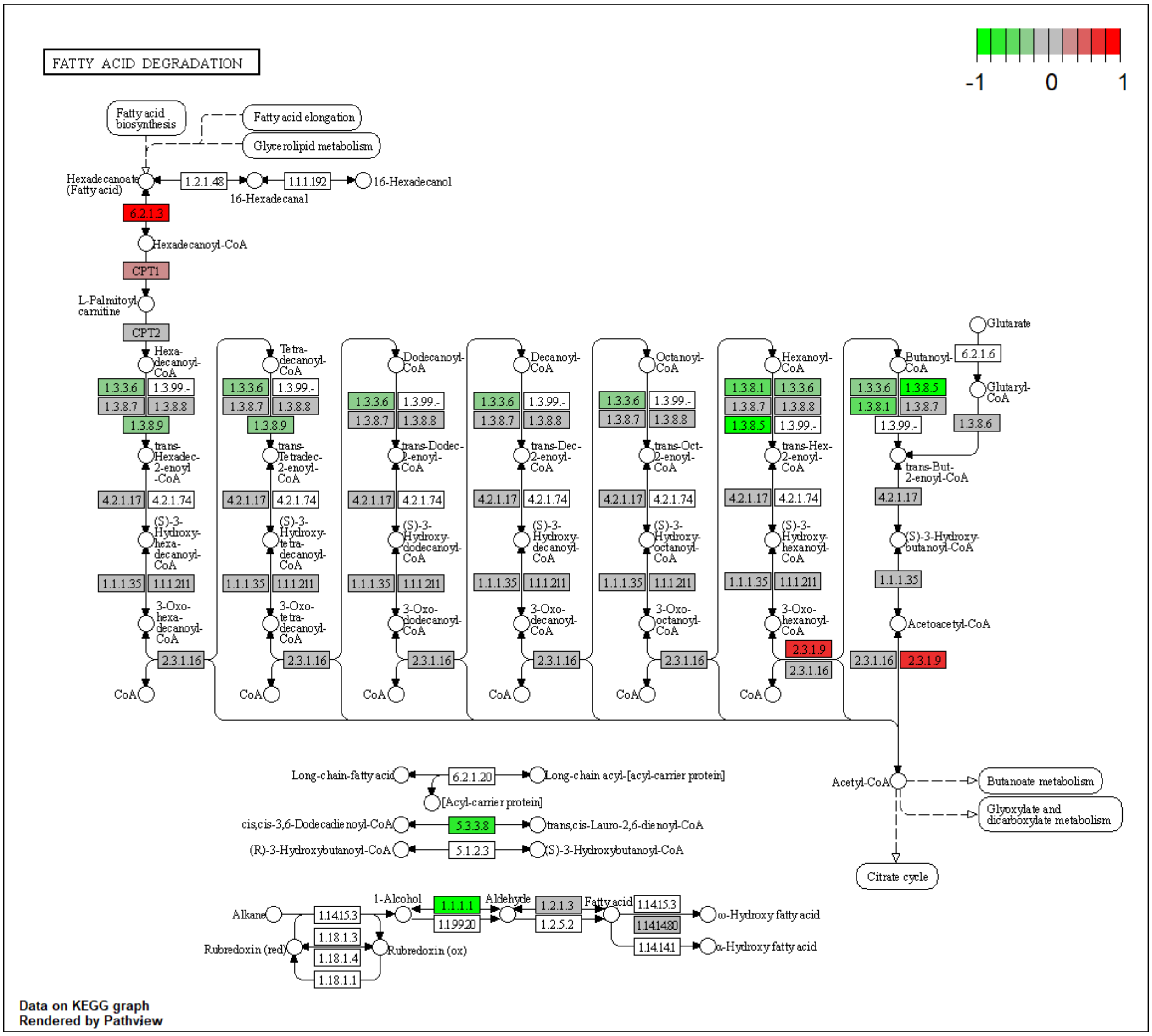


Figure 22 fatty acid degradation pathway

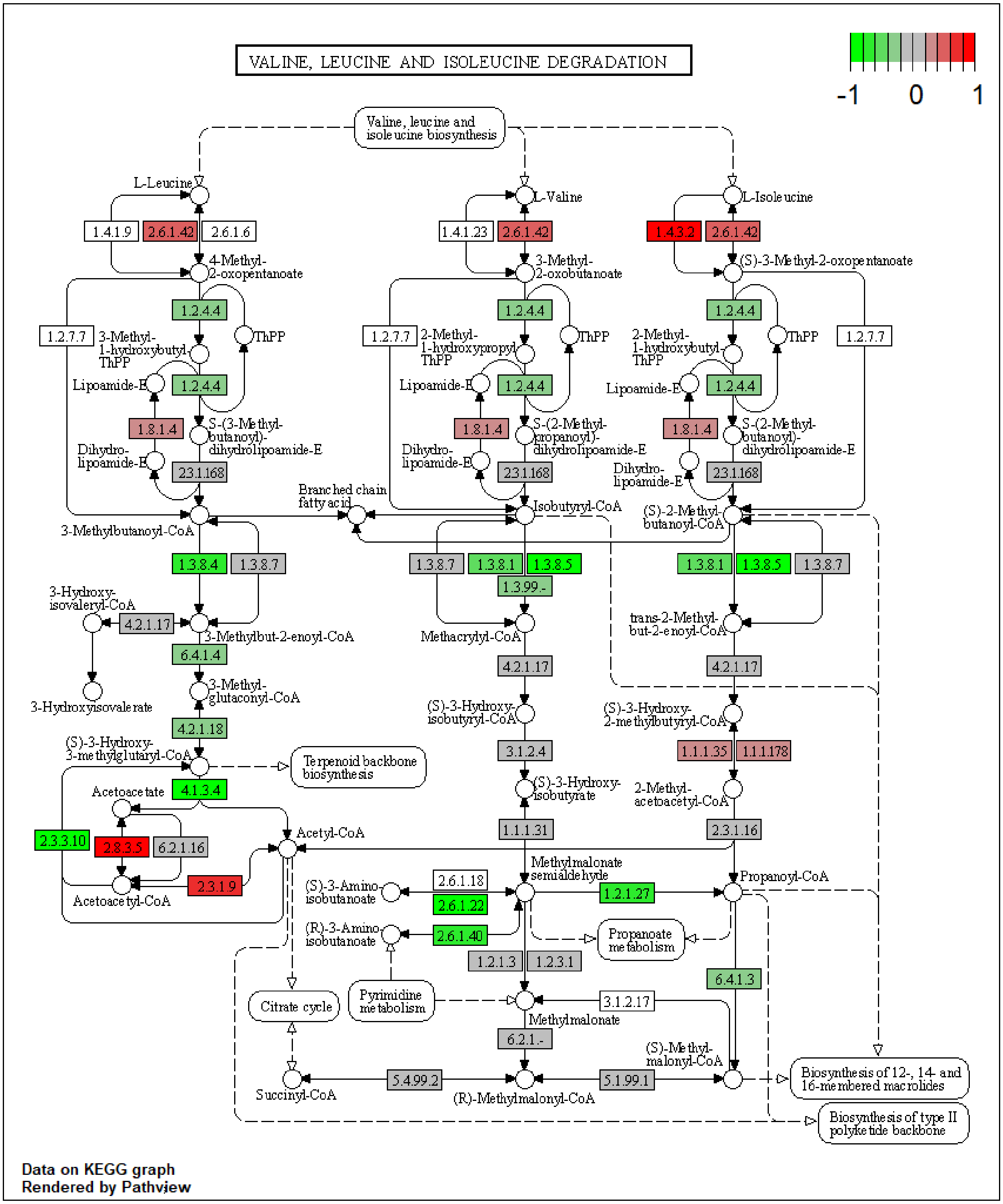


Figure 23 valine, leucine and isoleucine degradation pathway

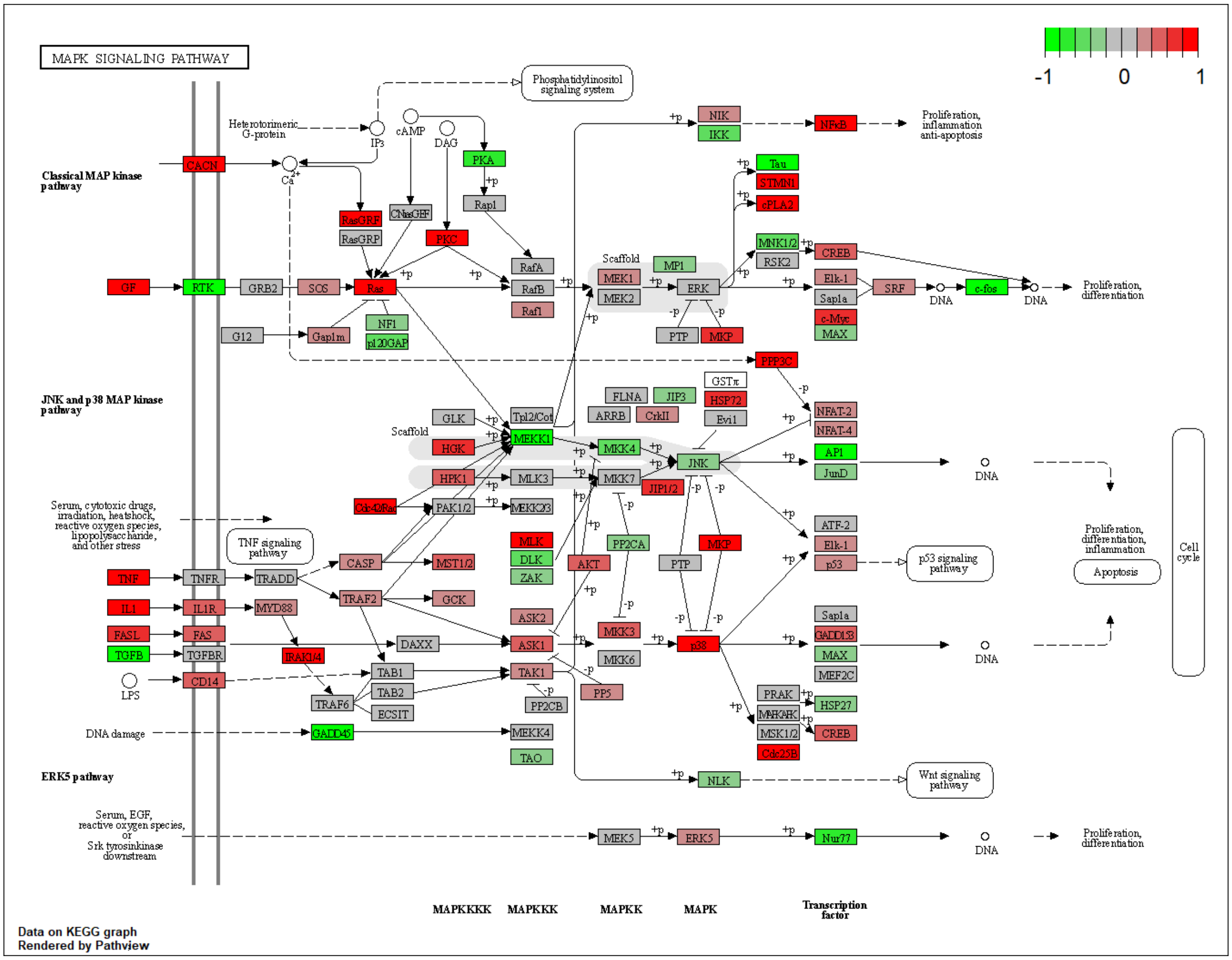
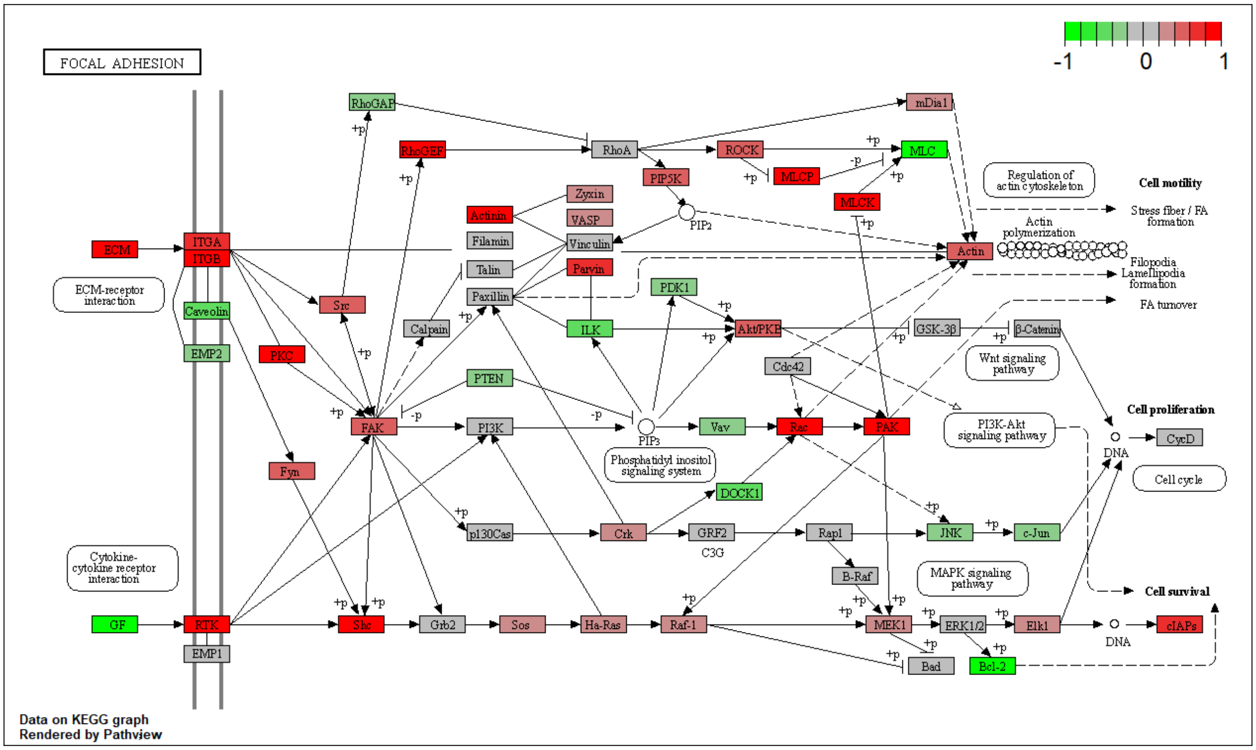


Figure 24 MAPK signaling pathway



Data on KEGG graph
Rendered by Pathview

Figure 25 focal adhesion pathway

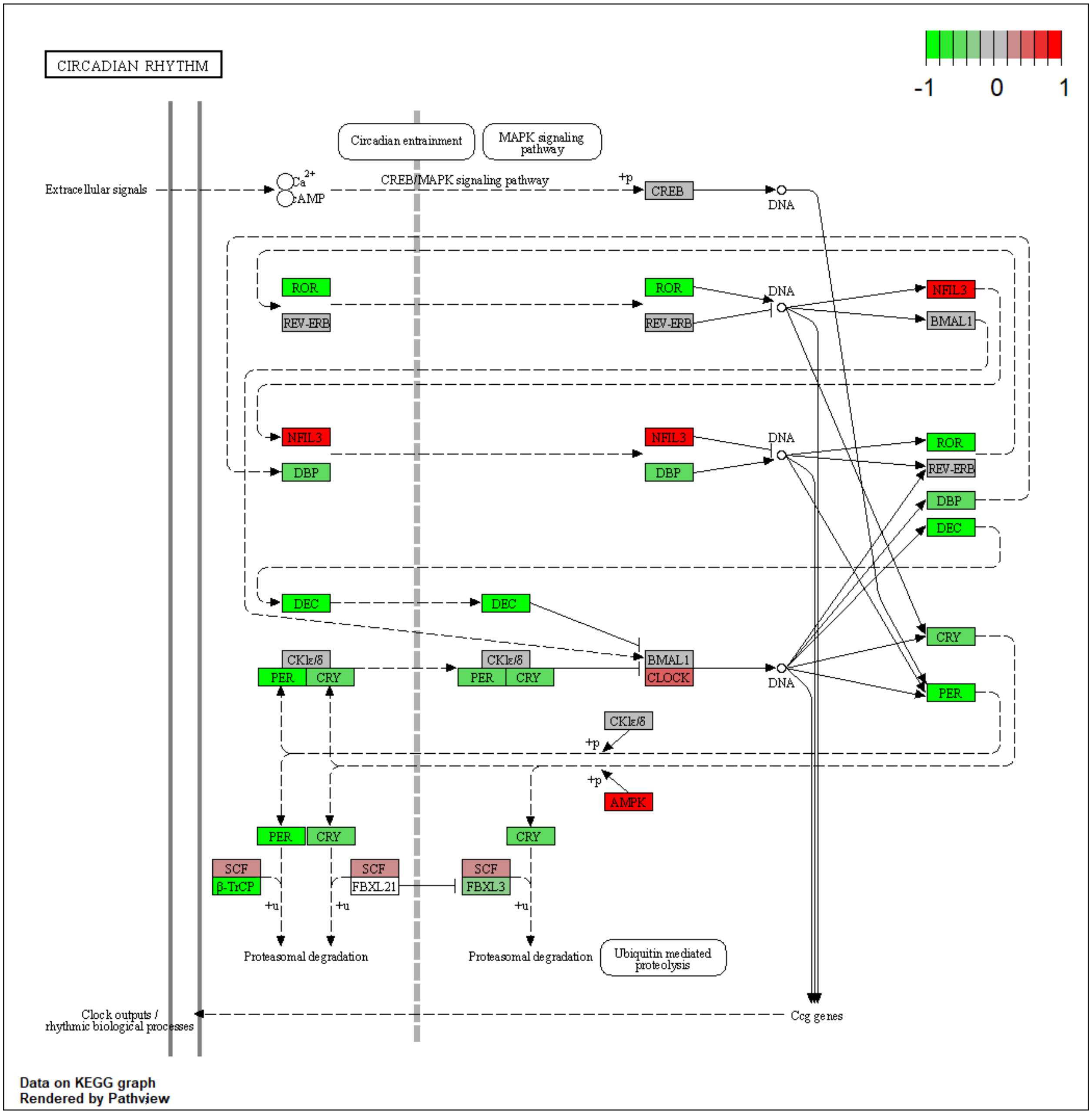


Figure 26 circadian rhythm pathway

2,3 up regulated pathways

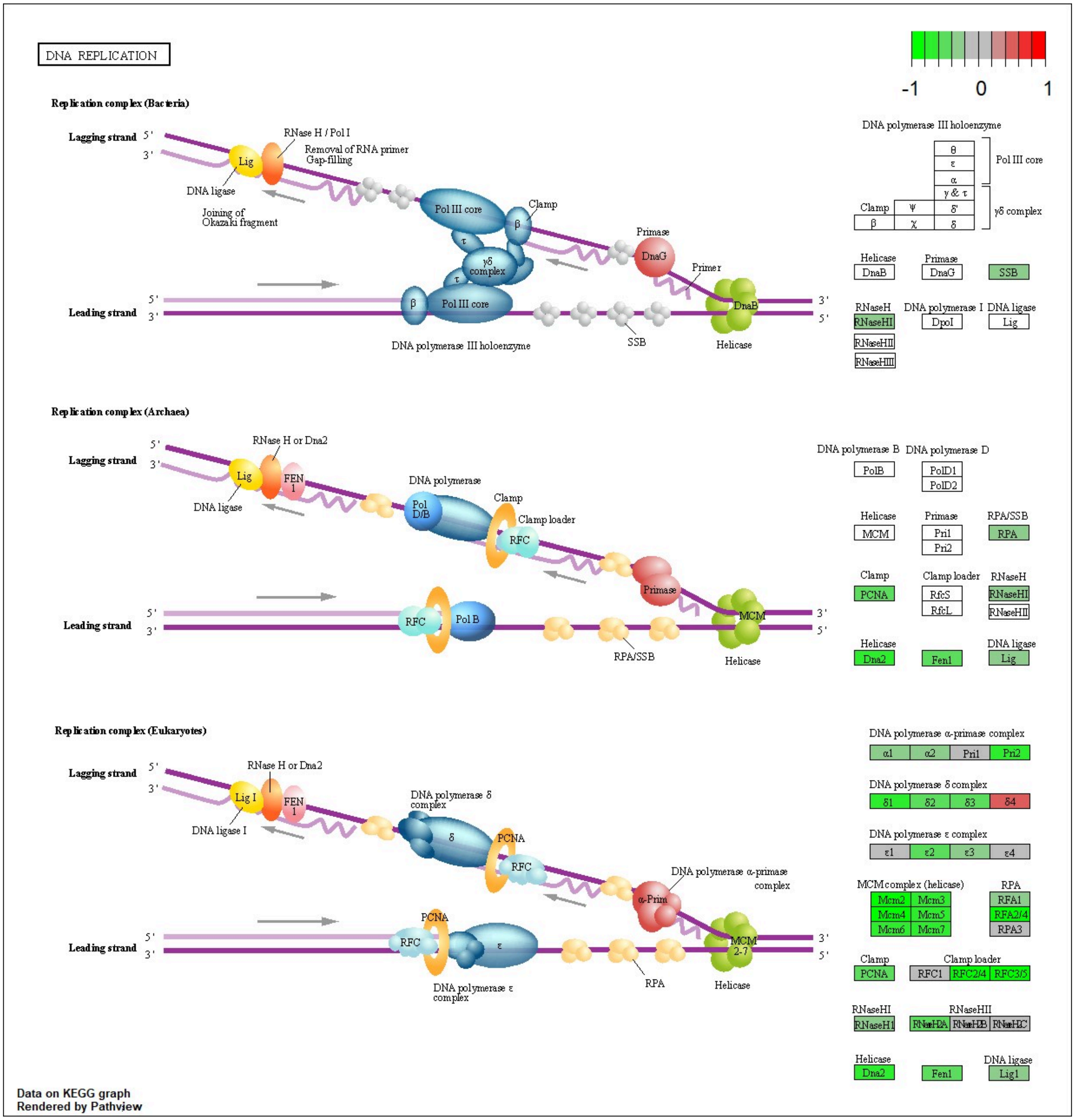


Figure 27 DNA replication pathway

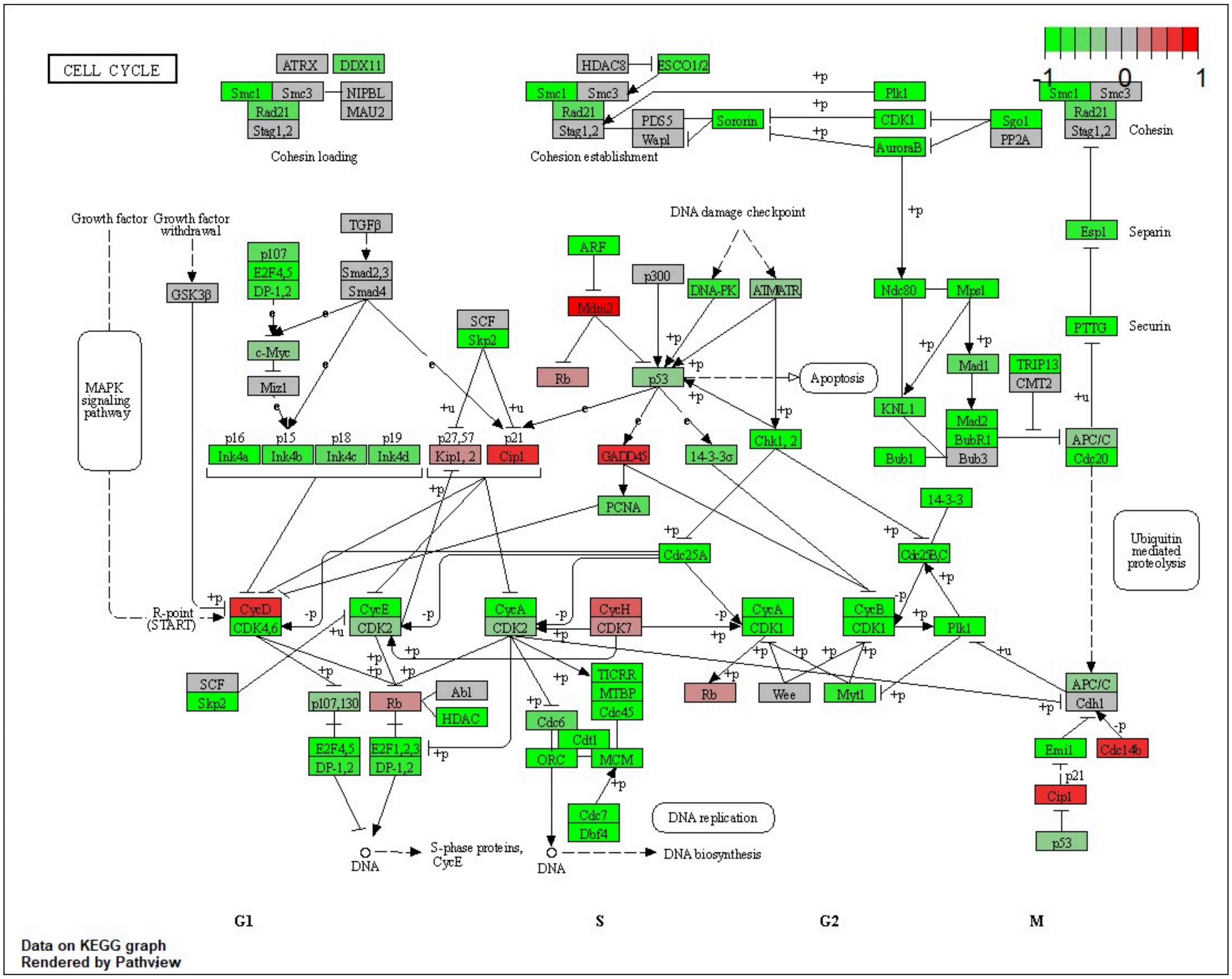


Figure 28 Cell cycle pathway

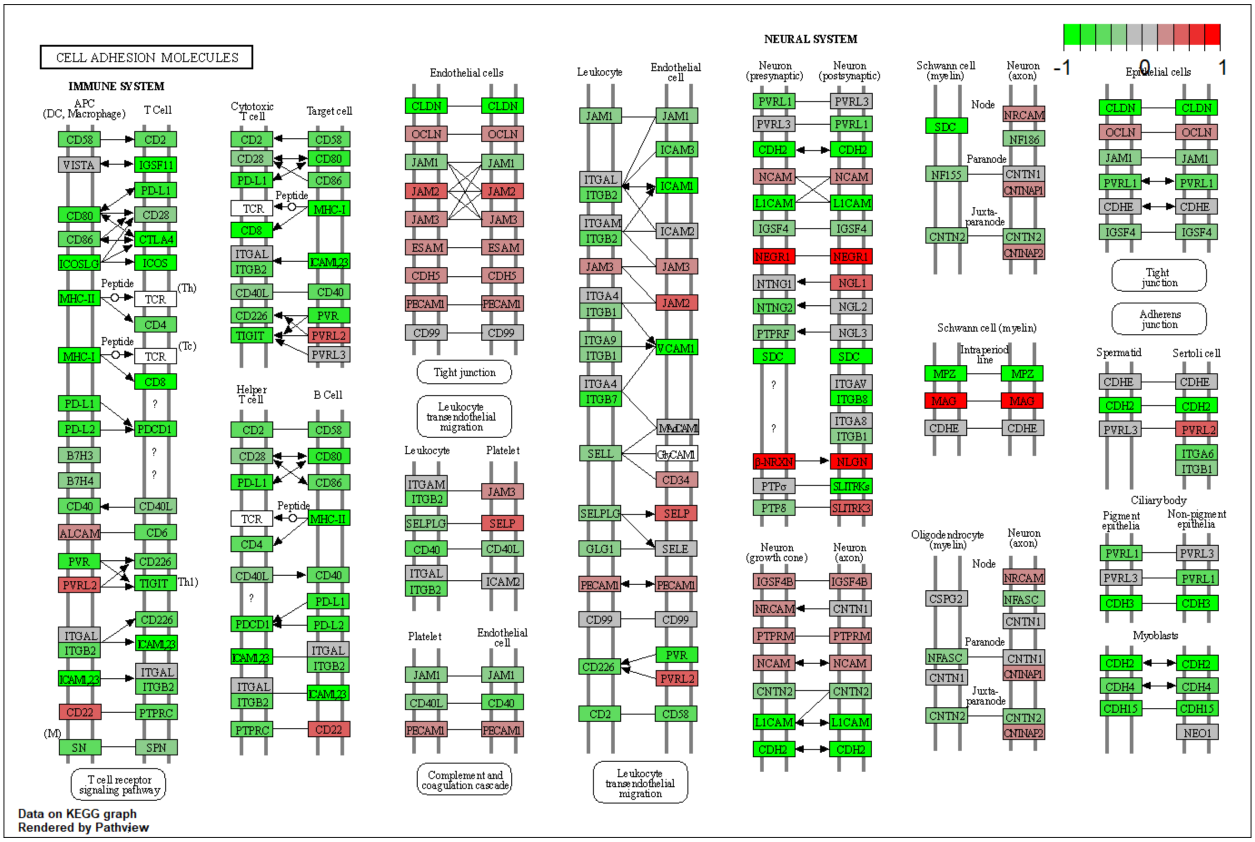


Figure 29 Cell adhesion molecules pathway

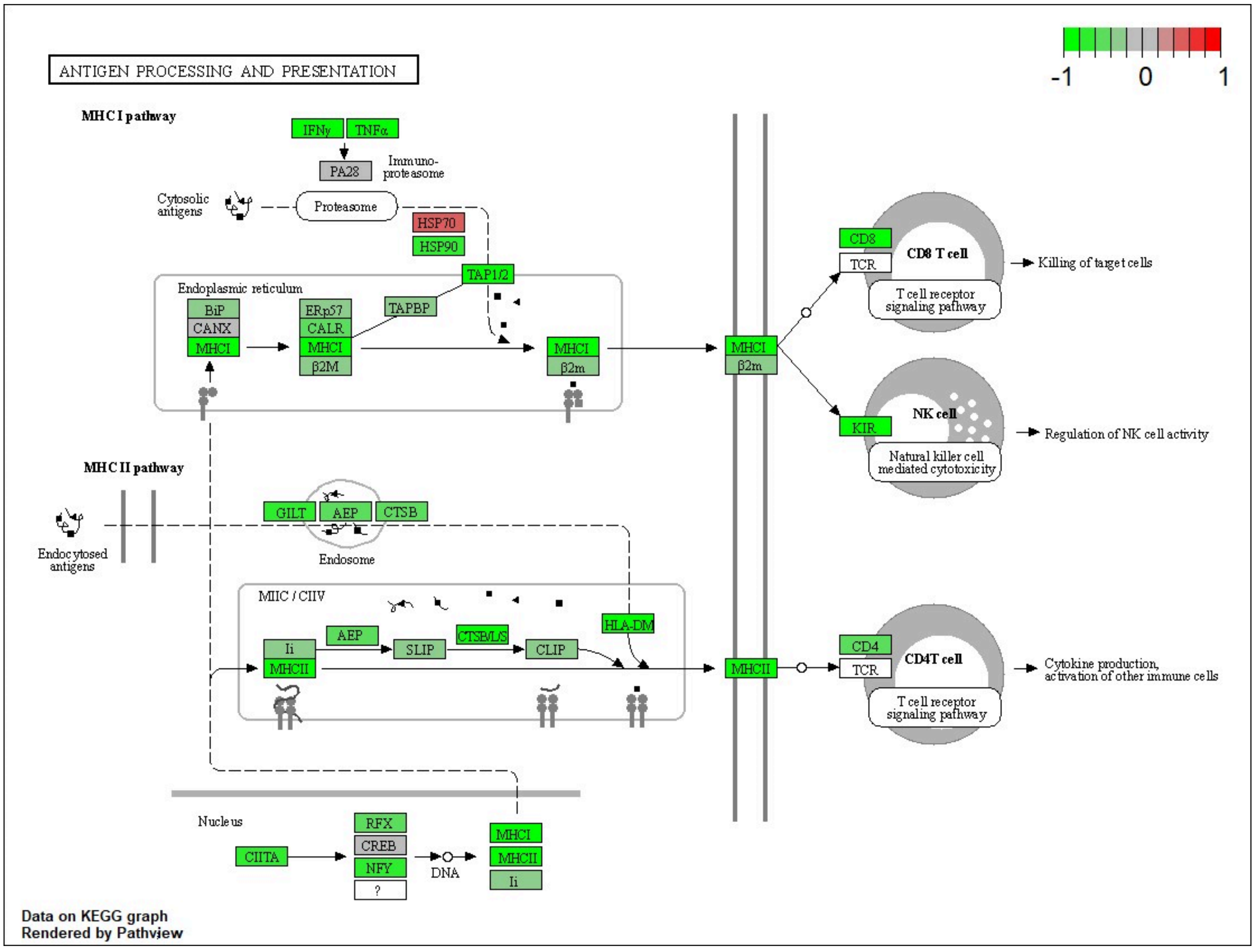


Figure 30 Antigen processing and presentation pathway

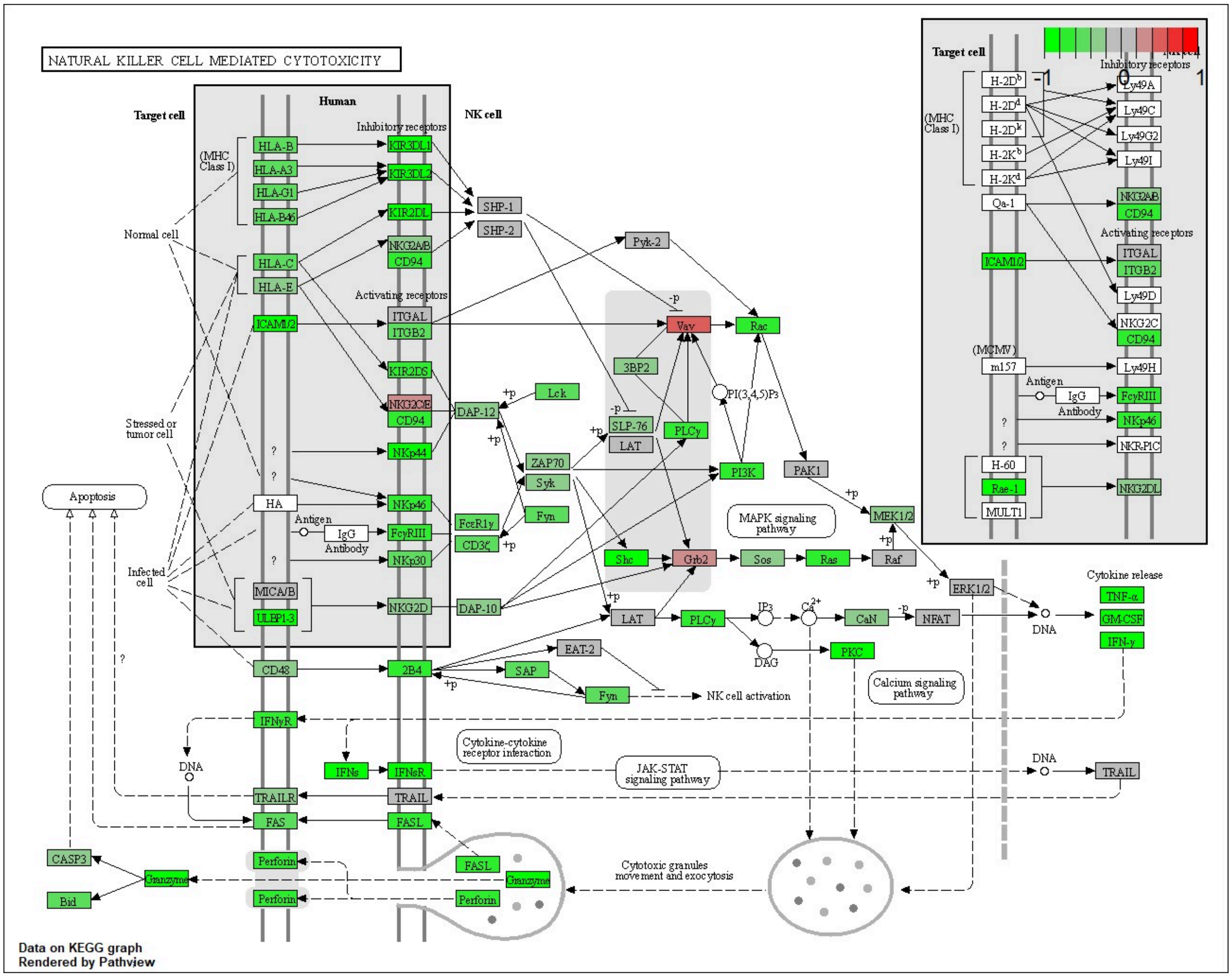


Figure 31 Natural killer cell mediated cytotoxicity pathway