Genome-wide differential expression in multiple myeloma patients with autologous stem cell transplant

Erin Pennington

Introduction

Summary

Multiple myeloma is a cancer of plasma cells, a type of white blood cell. Typically found in bone marrow, the function of plasma cells is to make antibodies (immunoglobulin) to protect the body from infection. However, cancerous plasma cells (myeloma cells) crowd out functional plasma cells and instead produce identical non-functional antibodies. These monoclonal antibodies fail to protect the body from infections, compromising the immune system. The overgrowth of myeloma cells also interferes with both blood formation and bone maintenance, causing anemia and osteoporosis (American Cancer Society, 2018).

There are multiple treatment options for different stages of multiple myeloma, one of them being stem cell transplant. This treatment involves killing the existing myeloma cells using drug treatments and chemotherapy, then introducing new healthy stem cells from blood which will replace them with functional plasma cells. There are two types of stem cell transplants: autologous, where the new stem cells originate from the patient's own blood, and allogeneic, where they originate from a donor. However, while stem cell transplant is a common treatment, it has severe side effects and puts the patient at a higher risk for serious infection and bleeding (American Cancer Society, 2018).

Previous research has shown that the outcome of these stem cell transplants depends on genetic factors, and gene expression profiling has helped to identify specific genes and mutations which affect outcomes (Gam, Rihab et al., 2017). In order to identify genes that might have an impact on outcomes due to changed expression after transplant, I will compare gene expression between multiple myeloma tumor samples after stem cell transplant, and samples without stem cell transplant.

Data

The samples used were collected by the Multiple Myeloma Research Foundation (MMRF) CoMMpass (Relating Clinical Outcomes in MM to Personal Assessment of Genetic Profile) study, made available through The Cancer Genome Atlas Program (TCGA) project MMRF-COMMPASS. The data originates from an ongoing longitudinal observation study of myeloma patients undergoing approved treatments and includes genetic information from tumor tissue samples and clinical outcomes.

To compare gene expression, I will use differential expression analysis using the R Bioconductor package DESeq (Anders & Huber, 2010) using RNA-seq data in order to determine quantitative expression level differences for each gene between treatment levels. For this analysis, I will be excluding samples from patients which have undergone allogeneic stem cell transplant, as it could confound our results due to differing risks versus the autologous method. 426 samples taken after autologous stem cell transplant will make up one treatment level for comparison, while the other comprises 417 samples with no stem cell transplant.

The result of the differential expression analysis will be used to determine how many genes had significantly changed expression, and which were up or downregulated. Investigating the function of these genes could give indications of whether they might be relevant to patient recovery after stem cell transplant and help inform which genes are of interest for future studies on the mechanism of that recovery.

Methods

R was used to perform this analysis (R Core Team, 2017). Using the TCGABiolinks library (Colaprico, Antonio, et al, 2015), I queried for and downloaded "HTSeq – Counts" from the "Gene Expression Quantification" data type in the "Transcriptome Profiling" category of the project "MMRF-COMMPASS" hosted by TCGA. I also used the biomaRt library (Durinck, S., Spellman, P., Birney, E. et al., 2009) in order to annotate the genes with names to facilitate interpretation of the results.

I then collected the presence and types of stem cell transplant treatment performed for each sample, and

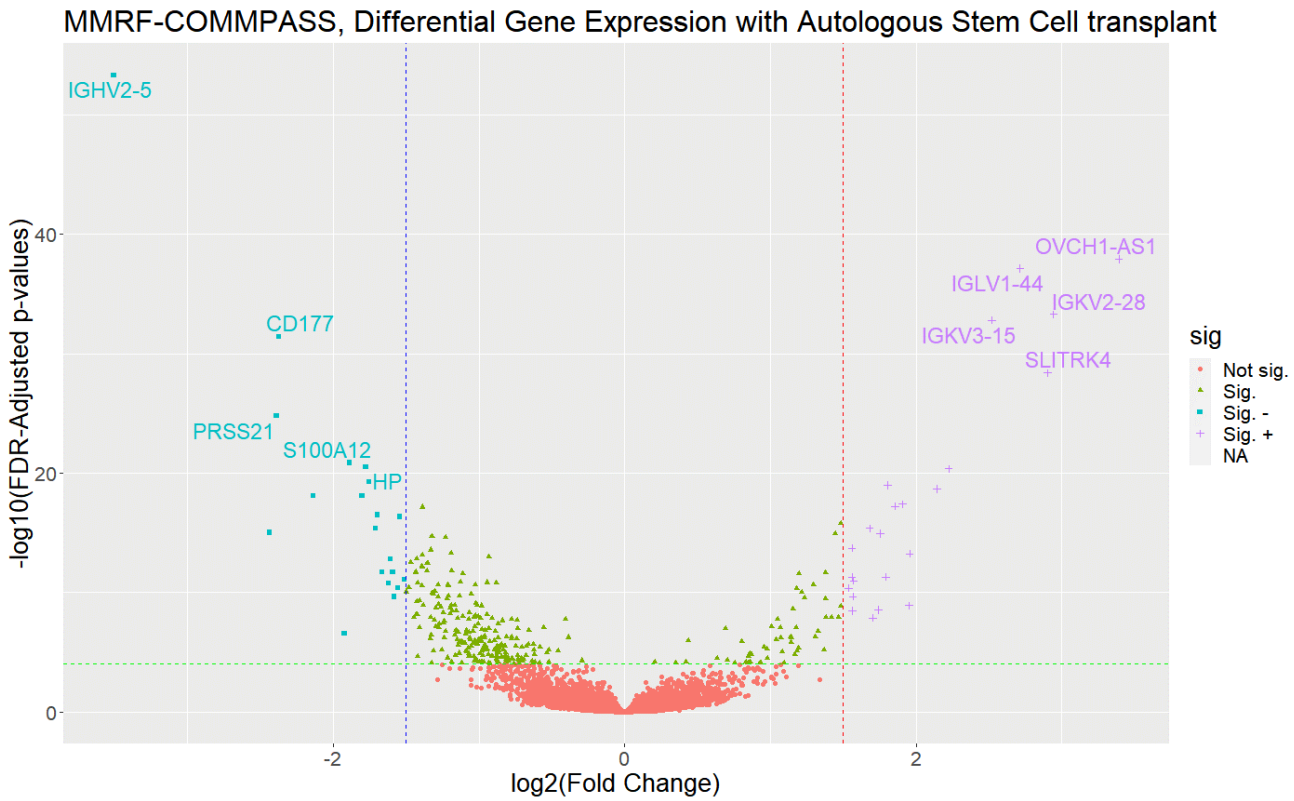

Figure 1. Differentially expressed genes in samples with stem cell transplant. *Volcano plot of genes significantly up (Sig +) or downregulated (Sig -) with autologous stem cell transplant compared to without transplant. X-axis represents log_2 transformation of fold change in gene expression in the stem cell transplant group compared to the control. Y-axis represents $-log_{10}$ adjusted p-values for gene expression change. Genes below the green line are not significantly differently expressed, and those above are significant (False discovery rate adjusted p-value cutoff = 0.0001). Genes to the left of the blue line are downregulated, and those to the right of the red line are upregulated (Log fold difference cutoff = 1.50).*

as previously mentioned, I excluded any samples with allogeneic stem cell transplant. Next, a visual analysis of the samples revealed an outlier with a significantly higher mean count standard deviation, so any samples with mean count SD over 100,000 was removed. In order to only focus on expressed genes, I removed genes with low baseline expression, only keeping genes where over 50% of samples had more than 0 reads.

Finally, I used the DESeq2 library (Anders & Huber, 2010) to perform a DESeq differential expression analysis. The package defaults were used except for the design parameter, which was provided as a factor vector containing FALSE values as the control group, and TRUE values for the stem cell transplant group. The results were visualized in a volcano plot using the ggplot library (Wickham, 2009) and ggrepel (Slowikowski, Kamil et al., 2021). An adjusted p-value cut-off of 0.0001 was used in addition to a minimum log fold difference cut-off of 1.50 to determine significant up or downregulation (Brandon Burgman, personal communication). A p-value cut-off of 0.01 was considered originally which provided numerous significant results. In order to target a smaller portion of genes with more power, I updated this to a p-value cut-off of 0.0001.

Results

Filtered genes and samples

	Genes	*Samples Untreated*	*Samples Treated*
Before Filtering	56602	418	441
After Filtering	29955	417	426

Table 1. Gene counts before and after filtering. *26647 genes were removed due to low baseline expression (< 50% of samples had > 0 reads). In the untreated samples (autologous stem cell transplant), 1 sample with allogeneic stem cell transplant was removed. In the treated samples (with autologous stem cell transplant), 15 samples which also had allogeneic stem cell transplant were removed.*

Differentially expressed genes

	Normalized Counts Mean ± SD Untreated	*Normalized Counts Mean ± SD Treated*	*# Genes*
Significant	6,133 ± 155,365	7,352 ± 175,060	318
Sig. Upregulated	16,808 ± 254,201	30,693 ± 390,457	23
Sig. Downregulated	4980 ± 62,844	1,547 ± 28,476	20

***Table 2. Counts and significance of differentially expressed genes**. Significant (p-adj < 0.0001), upregulated (log_2 fold change > 1.5), downregulated (log_2 fold change < -1.5). There were no genes meeting the up or downregulation fold change which were not significant, due to the large sample sizes.*

Differentially expressed gene biotypes

	IGJ Gene	*IGV Gene*	*IGV Pseudogene*	*lncRNA*	*Protein Coding*
Upregulated	1	14	2	1	5
Downregulated	0	3	1	1	15

***Table 3. Number of genes up and downregulated from different biotypes.** Immunoglobulin Junction (IGJ) genes are linkers for antigen binding. Immunoglobulin Variable (IGV) genes are part of the immunoglobulin chain. lncRNA are long non-coding RNA with various regulation functions. Protein coding genes are transcribed into proteins with various functions.*

After filtering, there were 29955 genes remaining, 417 control samples, and 426 samples with autologous stem cell transplant (Table 1). The differential expression analysis identified 318 genes which had a significant change in expression (p-adj < 0.0001), with 23 genes that were significantly upregulated and 20 which were significantly downregulated ($|\log_2$ fold change$| > 1.5$) (Table 2). Interestingly, all of these upregulated or downregulated genes also had a statistically significant adjusted p-value.

On inspection of the volcano plot (Figure 1), there was a clearly identifiable group of genes which were the most significantly upregulated across the autologous stem cell transplant samples compared to the control samples. This highly upregulated group included OVCH1-AS1, IGKV2-28, SLITRK4, IGLV1-44, and IGKV3-15. Of the downregulated genes, a few of the most significant include IGHV2-5, CD177, PRSS21, S100A12 and HP. IGHV2-5 stood out as the most significantly differentially expressed gene, with a more intense downregulation compared to any of the other genes investigated.

The biotypes of the significantly up or downregulated genes were also identified (Table 3). Many immunoglobulin variable genes were found to be upregulated, while most downregulated genes were protein coding.

Analysis

Immunoglobulin variable (IGV) genes make up most of those upregulated with stem cell transplant. IGKV2-28, IGLV1-44, and IGKV3-15 stand out as part of the immunoglobulin light chain, which has a function in antigen recognition (Stelzer, G. et al., 2016). Since multiple myeloma results in the creation of dysfunctional monoclonal antibodies, this upregulation of antibody genes might imply a return in immunoglobulin function.

Two other notable upregulated genes currently have no explicitly known relation to myeloma. SLITRK4 is a transmembrane protein which suppresses neurite outgrowth, but it has been correlated with myeloma treatments in the past despite its seemingly unrelated function (Manasanch, E. et al., 2019). On the other hand, OVCH1-AS1 is a long non-coding RNA which has a role in preventing osteolysis, which can be a complication of multiple myeloma (Yang, G. et al., 2021). This would be an interesting interaction to investigate in further research, as its upregulation would indicate better maintenance of bone health, but its interaction with myeloma has not yet been explored.

Although most of the downregulated genes code for various proteins, there are a few exceptions which are IGV genes. This was an unexpected result, since as previously mentioned, IGV gene upregulation would seemingly indicate proper antibody function. To determine the reason for this conflict, I looked into the

most significantly downregulated gene, IGHV2-5, which is an immunoglobulin heavy chain gene also involved in antigen recognition. Previous research has actually found that this gene is overrepresented in multiple myeloma tumors (Ferrero, Simone et al, 2012), which indicates that abundance of this allele could be indicative of the cancer, or part of the antibody dysfunction itself. However, more research would be necessary to determine if downregulation of IGHV2-5 in stem cell transplant patients is truly related to recovery of immunoglobulin function.

S100A12 was one of the protein coding genes downregulated in the stem cell transplant group. It plays a proinflammatory part in the immune response and is considered a biomarker of inflammation (Meijer, B. et al., 2012). Although S100A12 itself has not been directly researched with this cancer, high expression of the S100 family of genes is correlated with negative multiple myeloma outcomes. Higher expression of S100A12 results in more drug resistance and its copy number is increased in some patients, which is problematic for treating relapse. Since S100A12 expression was downregulated after stem cell transplant, this could help prevent such drug resistance issues for future drug treatments.

Another downregulated protein coding gene, CD177, codes for a cell surface glycoprotein for neutrophil (white blood cell) activation. Mutations in this gene are associated with myeloproliferative diseases, which results in the overproduction of blood cells (Stelzer, G. et al., 2016). Myeloproliferative diseases are not the same cancer as multiple myeloma, although they can occur simultaneously. The downregulation of CD177 does not mean that it is mutated though, so its downregulation is not necessarily an indicator of recovery. On the contrary, CD177 has been shown to be positively correlated with breast cancer survival (Kluz, P. et al., 2020), so it is possible that its downregulation is an undesired result of stem cell transplant. Altogether, it is unclear what relation CD177 expression has with multiple myeloma incidence or recovery.

There was a very high standard deviation for each significant group of differentially expressed genes (Table 2). This is potentially due to a minority of genes having very high count numbers and count deviations (Figure S1, Table S1, S2).

Conclusion

This differential expression analysis was performed to determine the genetic expression context for multiple myeloma recovery after autologous stem cell transplant, with a goal to identify key genes and their functions that relate to better outcomes after treatment. The results showed that in samples from patients who had the transplant, many immunoglobulin variable chain genes had higher expression, possibly indicating a recovery of antibody function. In addition, several genes that are often associated with inflammation and immunoglobulin dysfunction had lower expression, which also points to improved outcomes.

Since stem cell transplant is a high risk treatment, the genes identified in this analysis could be compared to those differentially expressed by alternative lower risk drug treatments, in order to identify which genes could serve as additional drug targets not already regulated by existing drugs.

Some of these differentially expressed genes do not have a known mechanistic relationship with multiple myeloma, such as CD177, SLITRK4 and OVCH1-AS1. With further investigation, these genes could be found to have some mechanism that ties into myeloma, serve as markers for certain outcomes, or may be considered as targets for future therapies. In particular, future research could investigate if CD177 is associated with multiple myeloma survival similarly to breast cancer, and whether its downregulation by stem cell transplant represents a negative result of the treatment.

Works Cited

American Cancer Society. (2018). *Stem Cell Transplant for Multiple Myeloma*. Retrieved from American Cancer Society: https://www.cancer.org/cancer/multiple-myeloma/treating/stem-cell-transplant.html

American Cancer Society. (2018). *What is multiple myeloma?* Retrieved from American Cancer Society: https://www.cancer.org/cancer/multiple-myeloma/about/what-is-multiple-myeloma.html

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology, 11*, R106. doi:10.1186/gb-2010-11-10-r106

Colaprico, Antonio, et al. (2015). TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic acids research, 44*(8), e71. doi:10.1093/nar/gkv1507

Durinck, S., Spellman, P., Birney, E. et al. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols, 4*, 1184-1191.

Ferrero, Simone et al. (2012, June). Multiple myeloma shows no intra-disease clustering of immunoglobulin heavy chain genes. *Haematologica, 97*(6), 849-853. doi:10.3324/haematol.2011.052852

Gam, Rihab et al. (2017). Genetic Association of Hematopoietic Stem Cell Transplantation Outcome beyond Histocompatibility Genes. *Frontiers in immunology, 8*, 380. doi:10.3389/fimmu.2017.00380

Liu, M. et al. (2021). S100 Calcium Binding Protein Family Members Associate With Poor Patient Outcome and Response to Proteasome Inhibition in Multiple Myeloma. *Frontiers in Cell and Developmental Biology, 9*, 2261. doi:10.3389/fcell.2021.723016

Love, M. I., & Anderson, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology, 15*(12), 550. doi:10.1186/s13059-014-0550-8

Manasanch, E. et al. (2019). A pilot study of pembrolizumab in smoldering myeloma: report of the clinical, immune, and genomic analysis. *Blood Advances, 3*(15), 2400–2408. doi:10.1182/bloodadvances.2019000300

Meijer, B. et al. (2012). The Role of S100A12 as a Systemic Marker of Inflammation. *International Journal of Inflammation, 907078*. doi:10.1155/2012/907078

Mounir, Mohamed, et al. (2019). New functionalities in the TCGAbiolinks package for the study and integration of cancer data from GDC and GTEx. *PLoS computational biology, 15*(3), e1006701. doi:10.1371/journal.pcbi.1006701

R Core Team. (2017). R: A Language and Environment for Statistical Computing. Retrieved from https://www.R-project.org/

Silva, Tiago C., et al. (2016). TCGA Workflow: Analyze cancer genomics and epigenomics data using Bioconductor packages. *F1000Research, 5*. Retrieved from https://f1000research.com/articles/5-1542/v2

Slowikowski, Kamil et al. (2021). *ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'*. Retrieved from https://github.com/slowkow/ggrepel

Springer. (2021, November 9). *Submission guidelines*. Retrieved from Genes & Genomics: https://www.springer.com/journal/13258/submission-guidelines

Stelzer, G. et al. (2016). The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Current Protocols in Bioinformatics, 54*, 1.30.1 - 1.30.33. doi:10.1002/cpbi.5

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York.

The results shown here are based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.