Review of:

“*Genome-wide transcriptional analysis of T cell activation reveals differential gene expression associated with psoriasis”* [1]

Sam Brown, Wayne Kunze, Caleb Perry

EECS E6690, Spring 2018

# Introduction

The paper selected for review investigates the genetic mechanisms that cause activation of T cells, responsible for the inflammation of skin is sufferers of Psoriasis, a common autoimmune disease.

**Short description of method**

**Short details on data set background. Large p**

**Our methods**

**Supervised**

**Feature Reduction**

**Unsupervised**

# Data Set and Paper

The paper examines Gene expression profiles from in vitro activated T cells from 17 psoriasis patients and 7 control subjects. The data set developed for this study contains 47,222 transcripts for each sample cataloging the level of gene expression in the activated T-cells for each gene in each individual.

**More details on data set**

**WCGNA**

**Focus on ID of genes, not pathways or biological background**

# Results

The paper, “Genome-wide transcriptional analysis of T cell activation reveals differential gene expression associated with psoriasis”, asserts several genes are significantly upregulated or down regulated in patients with psoriasis compared to those who don’t . these results were found by first pruning the data by removing all genes that were not expressed in at least 3 patients. Then the average expression rate of those who had the disease was computed and compared to the average expression rate of those who didn’t have the gene. This was used to produce tables of relative fold changes. Attempts to reproduce these results were made however results differ from the paper. Table 1 shows the 13 most upregulated gene indicators as a result of this analysis.

Table : List of most upgregulated genes

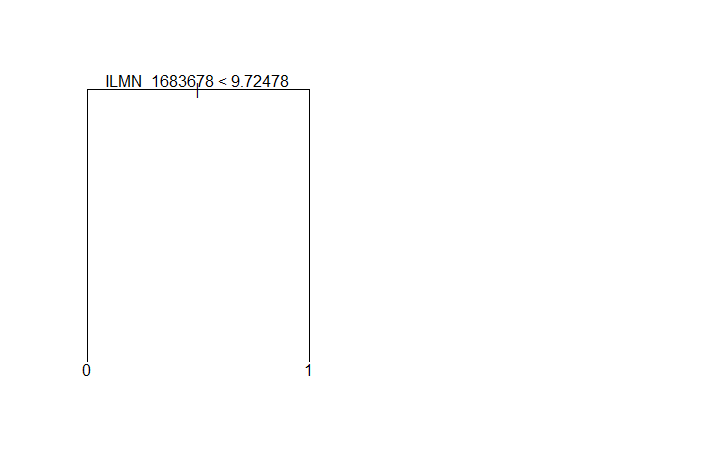
|  |
| --- |
| "ILMN\_2058782" "ILMN\_2305112" "ILMN\_1701789" "ILMN\_2410826" "ILMN\_1721113"  “ILMN\_1658247" "ILMN\_2054297" "ILMN\_2184373" "ILMN\_1739428" "ILMN\_1729749"  “ILMN\_1700967" "ILMN\_2347798" "ILMN\_1670134" |

The results from Table 1 differ from the results displayed in the paper. In the paper the principle gene discussed is SPATS2L, which the paper presented as being upregulated 1.37 fold in patients with psoriasis, however in our analysis, the gene was only upregulated 1.003101 fold. The cause of this discrepancy is unknown.

# New methods used

**Trees and Random Forests**

We decided to utilize Random Forests to analyze the data. This is a process that can be computationally intensive. We focused on large forests (each forest with ten thousand trees) but we only considered the default number of factors for each tree.

Working with more than forty-eight thousand factors creates a unique challenge. This is more than RStudio can handle with the randomForest function without crashing. To address this problem we had to choose a form of factor reduction. We chose to use importance from randomForest to take a subset of the “more important” factors from subsets of one thousand (or less) factors.

Once the “more important” factors have been identified, we used randomForest to create trees amongst these factors to identify the “most important” factors across the entire dataset.

We were surprised to note very limited overlap from Random Forests to other methods used by ourselves and the original paper. It’s not surprising that ILMN\_1683678 (SPATS2L) always makes the top slot (it has 100 percent accuracy in predicting the end result by itself, see Figure 1), but other factors not predicted elsewhere are also ranked very highly in terms of importance for Random Forests (see Figure 2).

Figure SPATS2L Tree

Creating individual trees from these “most important” factors has fairly good accuracy. We only see 100% accuracy in the case of ILMN\_1683678, but other trees use two factors for 96% accuracy (for example, see Figures 3 and 4).

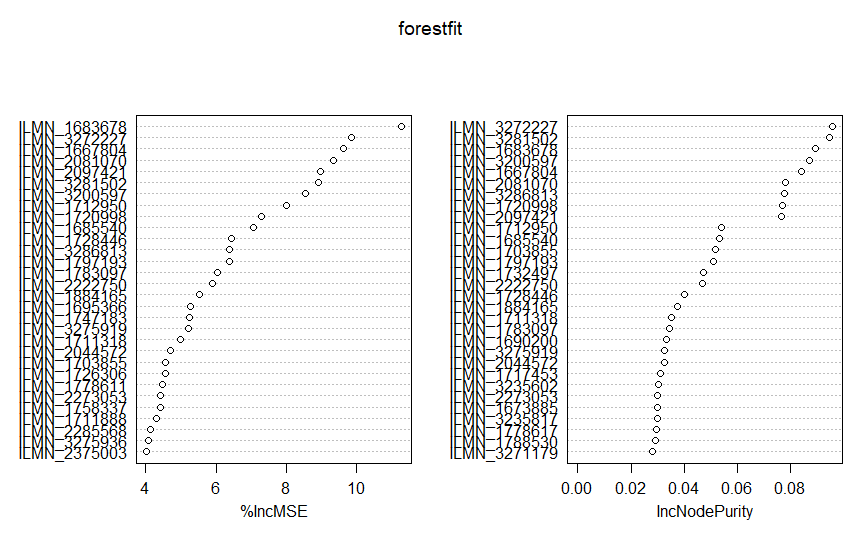


Figure Relative Importance of "Most Important" Factors

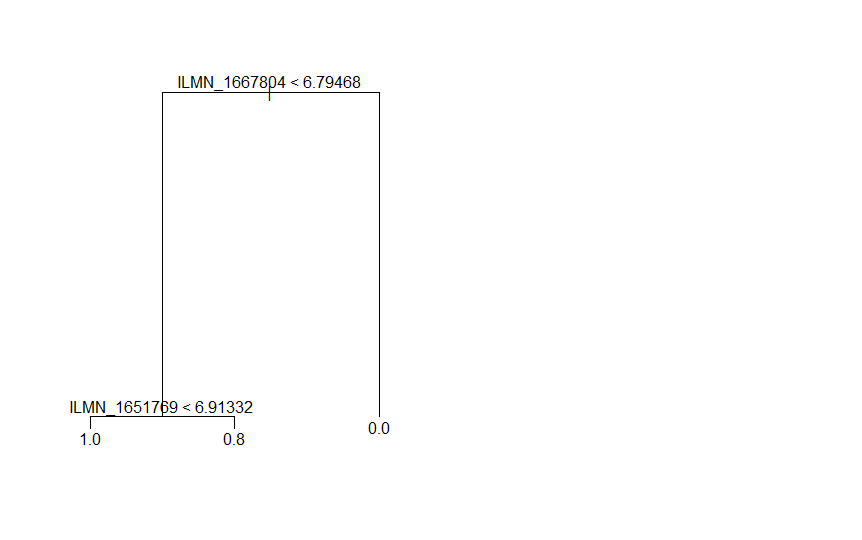
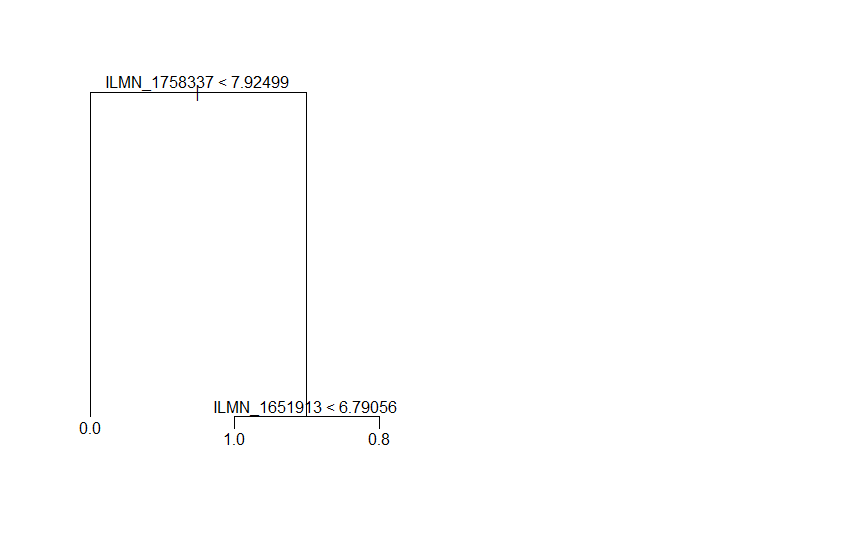


Figure Sample Tree 1

Figure Sample Tree 2

**PCA**

Another statistical learning method that was used was Support Vector Machines (SVM). Here many SVMs were trained on the data using a grid search with epsilon ranging from 0 to 1 in increments of 0.01, and the cost ranging from 22 to 29 in powers of 2. Each of these 700 different SVMs were tested using 10 fold cross validation to determine optimal choices for cost and epsilon. The Process was highly computationally expensive and took over 8 hours of computation time, however it yielded strong results. The best SVM utilized an epsilon value of 0 and a cost of 4. This SVM had a Root Mean Squared error of 0.0001073526, which is highly accurate. Figure 1 shows the error in each of the 24 patients. The data points were found to be completely separable. Figure – shows the results of the training. As can be seen larger values of epsilon produced higher error, with cost being somewhat independent of error. These results are to be expected, with data where the number of components from each sample far exceeds the number of samples. Since the data is perfectly separable, cost therefore has nearly no impact on the SVM.

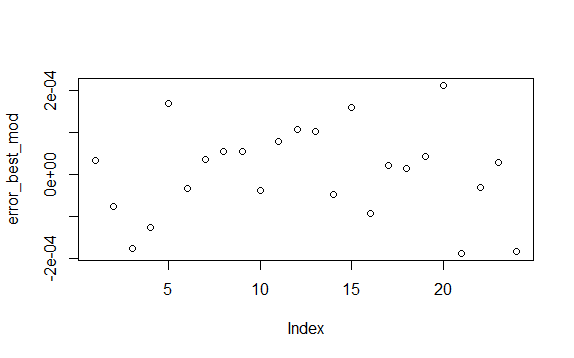


Figure : SVM error for each Patient

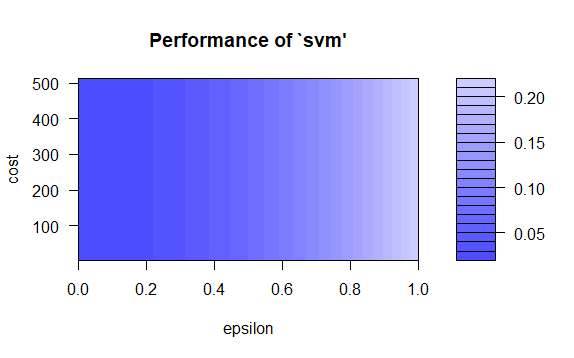


Figure : Performance of SVM using Grid search on Parameters

Unsupervised learning methods were also tried on the data. In particular K-means was tried, to see if the clusters would form around the disease without knowing who actually had the disease. An important metric for K-means clustering is the distance. Figure 3 shows Euclidean distance between all points while Figure 4 shows the Manhattan distance. From these distance plots, one can see that some patients had more similar gene expression rates than others. Kmeans clustering was done using the Eucliean distance, with 50 random starts, taking the best of the 50. Attempting to split the data into 2 clusters produced clusters of size 9 and 15, however the patients in each clusters had a similar rate of the disease compared to the original cluster. PCA was done to make the data view able and Figure 5 shows the 2 clusters. Expanding the search to look from 2 to 8 clusters, the presence of the disease did not seem to be hugely impactful. Using 6 or more clusters, the patents who did not have the disease seemed to be clustered more; however, the disease still did not seem to be a significant factor. Table 2 shows the results of clusters with various sizes. Here patients who did not have the disease are denoted as ‘1’, while those who did are ‘0’.

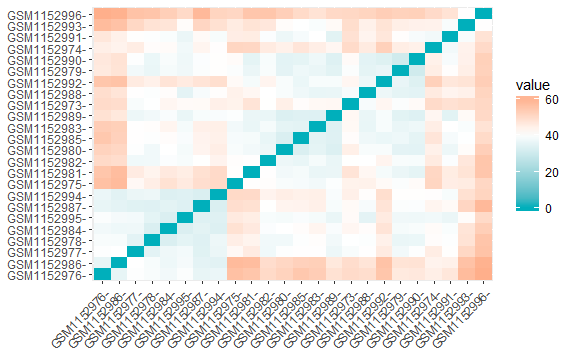


Figure : Euclidean Distance

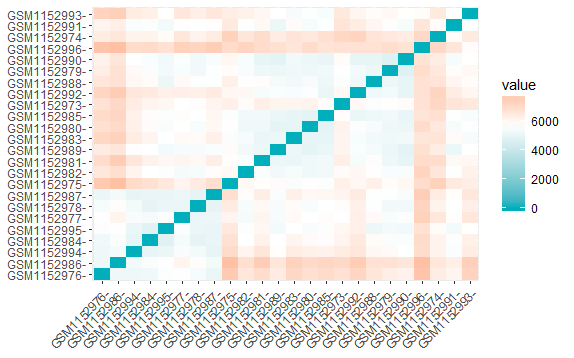


Figure : Manhattan Distance

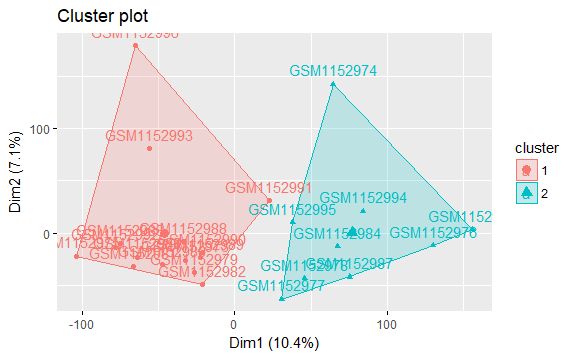


Figure : Clusters found using Kmeans

Table : results of Clusters with many different cluster sizes

|  |
| --- |
| [1] "Number of Cluster : 2"  Actual: 0 0 0 0 0 1 0 1 1[1] "Cluster: 1 mean: 0.333333333333333 "  Actual: 0 0 0 0 0 0 0 0 1 0 1 0 1 0 1[1] "Cluster: 2 mean: 0.266666666666667 "  [1] "==========================="  [1] "Number of Cluster : 3"  Actual: 0 0 0 1[1] "Cluster: 1 mean: 0.25 "  Actual: 0 0 0 0 1 0 1 1[1] "Cluster: 2 mean: 0.375 "  Actual: 0 0 0 0 0 0 0 0 1 0 1 1[1] "Cluster: 3 mean: 0.25 "  [1] "==========================="  [1] "Number of Cluster : 4"  Actual: 0 0[1] "Cluster: 1 mean: 0 "  Actual: 0 1[1] "Cluster: 2 mean: 0.5 "  Actual: 0 0 0 0 0 0 0 0 1 0 1 1[1] "Cluster: 3 mean: 0.25 "  Actual: 0 0 0 0 1 0 1 1[1] "Cluster: 4 mean: 0.375 "  [1] "==========================="  [1] "Number of Cluster : 5"  Actual: 0 0 0 0 0 0 0[1] "Cluster: 1 mean: 0 "  Actual: 0 1[1] "Cluster: 2 mean: 0.5 "  Actual: 0 0 0 0 1 0 1 1[1] "Cluster: 3 mean: 0.375 "  Actual: 0 0[1] "Cluster: 4 mean: 0 "  Actual: 0 0 1 1 1[1] "Cluster: 5 mean: 0.6 "  [1] "==========================="  [1] "Number of Cluster : 6"  Actual: 0 1[1] "Cluster: 1 mean: 0.5 "  Actual: 0[1] "Cluster: 2 mean: 0 "  Actual: 0 0 0 0 1 0 1 1[1] "Cluster: 3 mean: 0.375 "  Actual: 0 1 1 1[1] "Cluster: 4 mean: 0.75 "  Actual: 0 0 0 0 0 0 0[1] "Cluster: 5 mean: 0 "  Actual: 0 0[1] "Cluster: 6 mean: 0 "  [1] "==========================="  [1] "Number of Cluster : 7"  Actual: 0 1[1] "Cluster: 1 mean: 0.5 "  Actual: 0 0 0 0 1 1[1] "Cluster: 2 mean: 0.333333333333333 "  Actual: 0 0[1] "Cluster: 3 mean: 0 "  Actual: 0 0 0 0 0 0 0[1] "Cluster: 4 mean: 0 "  Actual: 0 1 1 1[1] "Cluster: 5 mean: 0.75 "  Actual: 0 1[1] "Cluster: 6 mean: 0.5 "  Actual: 0[1] "Cluster: 7 mean: 0 "  [1] "==========================="  [1] "Number of Cluster : 8"  Actual: 0 0 0 0 1 1[1] "Cluster: 1 mean: 0.333333333333333 "  Actual: 0 1[1] "Cluster: 2 mean: 0.5 "  Actual: 0 1 1 1[1] "Cluster: 3 mean: 0.75 "  Actual: 0 1[1] "Cluster: 4 mean: 0.5 "  Actual: 0[1] "Cluster: 5 mean: 0 "  Actual: 0[1] "Cluster: 6 mean: 0 "  Actual: 0 0 0 0 0 0[1] "Cluster: 7 mean: 0 "  Actual: 0 0[1] "Cluster: 8 mean: 0 "  [1] "===========================" |

# Conclusions

**Comparisons to original paper findings**

**Difficulty with large feature/small sample data set**

**Cross-Validation**

**Future: SVM?**

# References

**There are no sources in the current document.**14:825

**There are no sources in the current document.There are no sources in the current document.**

# Source Code

Table 1: Significantly Upregulated genes in Psoriasis patients compared to healthy controls

|  |  |  |  |
| --- | --- | --- | --- |
| Probe | Gene | Fold Change | P value |
| |  | | --- | | ILMN\_1683678 | | ILMN\_1735014 | | ILMN\_1703263 | | ILMN\_2322498 | | ILMN\_2246956 | | ILMN\_2397721 | | ILMN\_1781285 | | ILMN\_2321064 | | ILMN\_1731107 | | ILMN\_1729749 | | ILMN\_2095660 | | ILMN\_1681301 | | ILMN\_1767470 | | ILMN\_1735979 | | ILMN\_1719543 | | ILMN\_2188333 | | ILMN\_1776723 | | ILMN\_1721626 | | ILMN\_1773742 | | ILMN\_1741003 | | ILMN\_1729374 | | ILMN\_1660368 | | ILMN\_2406410 | | ILMN\_2305112 | | ILMN\_2284998 | | |  | | --- | | *SPATS2L* | | *KLF6* | | *SP140* | | *RORA* | | *BCL2* | | *GLB1* | | *DUSP1* | | *BAX* | | *CCDC92* | | *HERC5* | | *TMEM156* | | *AIM2* | | *SCPEP1* | | *BCKDHA* | | *MAF* | | *CD69* | | *PHF11* | | *ARID5B* | | *DNAJB9* | | *ANXA5* | | *ETFB* | | *TRRAP* | | *RHBDD2* | | *CTH* | | *SP100* | | |  | | --- | | 1.37 | | 1.32 | | 1.38 | | 1.31 | | 1.23 | | 1.23 | | 1.21 | | 1.24 | | 1.33 | | 1.72 | | 1.29 | | 1.42 | | 1.26 | | 1.21 | | 1.36 | | 1.33 | | 1.24 | | 1.27 | | 1.23 | | 1.27 | | 1.23 | | 1.20 | | 1.39 | | 1.57 | | 1.27 | | |  | | --- | | 0.0009 | | 0.0012 | | 0.0025 | | 0.0041 | | 0.0062 | | 0.0062 | | 0.0071 | | 0.0096 | | 0.0136 | | 0.0267 | | 0.0453 | | 0.0464 | | 0.0469 | | 0.0639 | | 0.0658 | | 0.0677 | | 0.0683 | | 0.0788 | | 0.0805 | | 0.0896 | | 0.0921 | | 0.0922 | | 0.0925 | | 0.0944 | | 0.0944 | |

Table 2: Significantly Upregulated genes in Psoriasis patients compared to healthy controls

|  |  |  |  |
| --- | --- | --- | --- |
| Probe | Gene | Fold Change | P value |
| |  | | --- | | ILMN\_3286813 | | ILMN\_3281502 | | ILMN\_1778617 | | ILMN\_1689294 | | ILMN\_2143250 | | ILMN\_1720114 | | ILMN\_2135175 | | ILMN\_1764163 | | ILMN\_1746148 | | ILMN\_1715401 | | ILMN\_2299072 | | ILMN\_2124802 | | ILMN\_1655827 | | ILMN\_2402936 | | ILMN\_3230435 | | ILMN\_1803799 | | ILMN\_1704873 | | |  | | --- | | *LOC391019* | | *LOC653375* | | *TAF9* | | *LOC85390* | | *FAR1* | | *GMNN* | | *SNORD36A* | | *LOC644330* | | *LRRC33* | | *MT1G* | | *CROP* | | *MT1H* | | *COPS2* | | *LOC440926* | | *LOC729086* | | *LOC649555* | | *TCEB1* | | |  | | --- | | -1.38 | | -1.31 | | -1.25 | | -1.20 | | -1.20 | | -1.20 | | -1.27 | | -1.20 | | -1.22 | | -1.82 | | -1.20 | | -1.64 | | -1.20 | | -1.22 | | -1.24 | | -1.21 | | -1.21 | | |  | | --- | | 0.0001 | | 0.0009 | | 0.0009 | | 0.0120 | | 0.0207 | | 0.0269 | | 0.0303 | | 0.0577 | | 0.0656 | | 0.0677 | | 0.0683 | | 0.0778 | | 0.0800 | | 0.0827 | | 0.0886 | | 0.0921 | | 0.0970 | |

[Research papers that use MLA format do not include a cover page unless requested by your instructor. Instead, start with the information shown at the top of this page. Do not bold the title or use all capital letters. Capitalize the first and last words of the title, and all principal words. If your paper includes a subtitle, separate it from the title by a colon and space, as shown. For more specific guidance on capitalization, see the MLA Handbook for Writers of Research Papers, 7th Edition (MLA 7th Edition).]

[All text—including titles, quotations, notes, and list of works cited—uses double line spacing. Body text and note text use a half-inch first-line indent. The list of works cited uses a half-inch hanging indent. Table titles and source text use a quarter-inch indent. Access all of these text formats on the Home tab, in the Styles gallery.]

[MLA format discourages extensive use of content notes. But, when you need to add notes, you can use either endnotes or footnotes. MLA 7th Edition indicates that you should use a superscript, Arabic numeral at an appropriate place in the text for the note reference. To begin the note text, use the same numeral, not superscript, and followed by a period.]

[If you use endnotes, they should be on a separate page, at the end of your text and preceding the list of works cited. If you use footnotes, consult your professor for preferred format.]

[For quotations of more than four lines, indent the quote one inch from the left margin and do not use quotation marks. To apply this formatting, on the Home tab, in the Styles gallery, click Quote. For shorter quotations, you can put them in quotation marks, and incorporate them directly into text.]

Table 1

[This Table Title Uses a Style Named “Table Title”, Available on the Home Tab, in the Styles Gallery]

| Column Heading | Column Heading | Column Heading | Column Heading |
| --- | --- | --- | --- |
| Row Heading | Table data | Table data | Table data |
| Row Heading | Table data | Table data | Table data |
| Row Heading | Table data | Table data | Table data |

Source: [This source text uses a style named “Table Source”, available on the Home tab, in the Styles gallery.]

1. [This note text uses a style named “Table Note”, available on the Home tab, in the Styles gallery. Table notes use a lowercase letter instead of Arabic numerals to differentiate them from the notes to body content.]

Fig. 1. [This figure caption uses the No Indent style, available on the Home tab, in the Styles gallery. Label figures with the abbreviated “Fig.” and a figure number.]

[The sample Works Cited list that follows was created using the Bibliography feature available on the References tab. This feature offers the option to specify MLA style, so that your references are automatically formatted correctly. You can also use this feature to add in-text citations, such as the one shown at the end of this paragraph. To add page numbers to a citation after you insert it, right-click the citation and then click Edit Citation. Note also that MLA rules for citations and references are extensive. So it’s a good idea to refer to MLA 7th Edition for further information.] (AuthorLastName Pages)

[To use this template when creating the outline for your paper, on the Home tab, in the Styles gallery, click No Indent. Then, on the same tab, in the Paragraph group, click the Multilevel List icon and then click the MLA Outline style that appears under List Styles. The first six levels of this list style correspond to the outline levels defined in MLA 7th Edition.]

For additional guidance on formatting your research paper, consult MLA 7th Edition as well as your instructor.

Works Cited

**There are no sources in the current document.**