CHARACTERIZATION OF PSORIASIS THROUGH CLUSTERING ANALYSIS OF MICROARRAY DATA

Bachelor Endproject (BEP) of	:	Manouk Groels (0904057)
Daily supervisor	:	Ir. Z.C. Félix Garza
Supervising staff member	:	Prof.dr. P.A.J. Hilbers
Head of CBio	:	Prof.dr. P.A.J. Hilbers
Period BEP	:	05/17 – 07/17
Report number	:	B17/4

Abstract

Psoriasis is an autoimmune disease expressed with thick and inflamed skin plaques. The effectiveness of treatments for this disease fluctuates between patients. To design a personalized treatment it is necessary to define patient specific factors. This study analyses the gene expression in patients with psoriasis. Two public microarray datasets are used, GSE13355 and GSE41662 with gene expression of tissue of lesional skin from psoriasis cases, tissue of non-lesional skin from cases and skin tissue from healthy controls. At first filtering is applied to the datasets to remove the genes that are not differentially expressed in all the tissue groups. Secondly, k-means clustering is applied to select genes that are able to classify the samples according to their tissue type. The selected genes of both datasets are compared and 63 core genes are found. A number of these genes are involved in the immune system and some are associated with cancer. It could be concluded that this analysis method may be useful to detect genes that increase the risk of psoriasis. But to make sure, more biomedical research should be done to identify the influence of the genes in the human body.

Contents

1. Introduction

Worldwide 2-3% of the population is affected by psoriasis (Parisi, Symmons, Griffiths, & Ashcroft, 2013). In the Netherlands, this percentage is 2,4% (Centraal bureau voor statistieken, 2014). Psoriasis is an autoimmune condition characterized by lesional areas of thick, red, and flaky skin (Nestle, Kaplan, & Barker, 2009). The quality of life of patients with this condition is affected due to the itching, pain, stress, and embarrassment they experience (Gelfand, et al., 2004). This makes it important to seek an effective treatment that can reduce the symptoms and improve their quality of life. Common therapeutic approaches for mild to severe psoriasis are topical corticosteroids (Smith & Barker, 2006) (Samarasekera, Sawyer, Wonderling, Tucker, & Smith, 2013), phototherapy (Hönigsmann, 2001), systemic therapies (Lebwohl & Ali, 2011), and biological immunomodulatory agents (Kormeili, Lowe, & Yamauchi, 2004). Recent studies of blue light-based phototherapy show a high variability in the responsiveness of individual subjects to this treatment (Pfaff, Liebman, & Born, 2015) (Weinstabl, Hoff-Lesch, Merk, & von Felbert, 2011). To explain this variability it is important to define patient specific factors. Ainali et al. (Ainali, Valeyev, Perera, Williams, & Gudjonsson, 2012) concluded that gene expression patterns revealed differences in the phenotype of psoriatic plaques. One subgroup of psoriatic plaques can be treated by targeting the TGFb and ErbB signalling pathways while other subgroups won't react to that. Thus, the focus of this study is on genetic factors.

In this study, a clustering method is implemented to analyse gene expression data from psoriasis patients to identify characteristic genes of psoriasis lesions and the skin on their surroundings, which may contribute to the response to a given treatment. The physiological background of psoriasis is described in section 1.1 and 1.2. The methodology used in this study is described in chapter 2. Chapter 3 presents the results obtained with this approach. Finally, the discussion and conclusion of this work are presented in chapter 4.

1.1 Skin structure

The skin is formed by three layers, i.e. epidermis, the dermis and subcutaneous fat (Shimizu, 2007) (figure 1). The fat layer is used to protect the body from external pressure and it preservers natural fat. Blood capillaries, lymphatic vessels, hair follicles and sweat glands are located in the dermis along with a dense structure of fibres produced by fibroblasts. These fibres maintain the dynamic strength of the skin and its elasticity. The epidermis is the outermost layer of the skin. The function of this layer is to protect the body from water loss, pathogens and physical injuries (Shwayder & Akland, 2005). The epidermis is composed of 4 layers: the basal, spinous, granular and cornified cell layers formed by keratinocytes at different stages of differentiation (Montagna, Kligman, & Carlisle, 1992). 95% of the cells in the epidermis are keratinocytes, some of them move from the basal to the cornified layer (Burns, Breathnach, Cox, & Griffiths, 2010). This process takes about 28 days in normal skin. When these cells are in the cornified layer they are enucleated and called corneocytes (Shimizu, 2007).

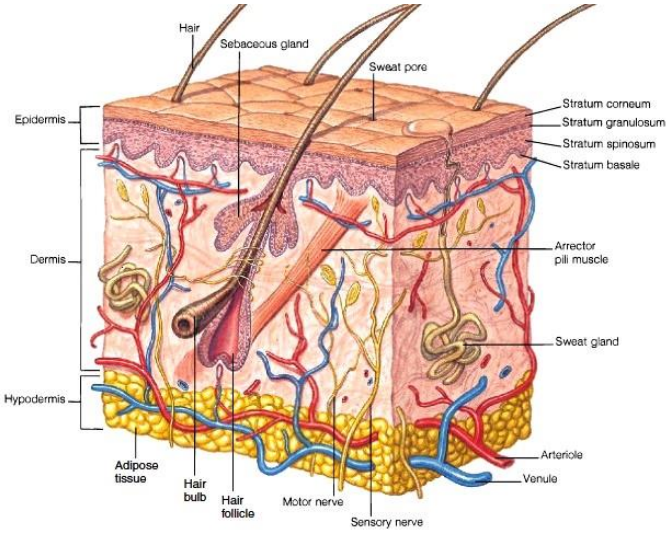

Figure 1. Structure of the skin. Figure taken from *(Shier, Lewis, & Butler)*.

1.2 Psoriatic skin

In psoriasis, a disturbance in the autoimmune regulation causes an overproduction and hyper-proliferation of keratinocytes (Dainichi, Hanakawa, & Kabashima, 2014). This causes a thickened and inflamed epidermis with well-defined borders (figure 2). The affected skin is called lesional skin and other parts of the skin of psoriasis patients are called non-lesional skin. Currently there is no cure for this disease, only symptom management treatments are available (Nestle, Kaplan, & Barker, 2009).

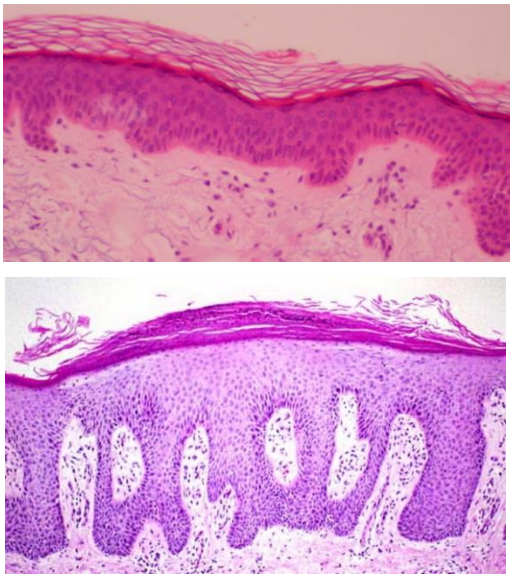

Figure 2. Hisology of a) healthy skin and b) psoriatic lesional skin. Figure taken from *(Swanson & Melton)* & *(University of Iowa, 1995)*.

It is known that this disease has genetic factors, but not everyone who carries the genes suffer from it (Chandra, Ray, Senapati, & Chatterjee, 2015). Nevertheless, twelve chromosomal loci have been linked to psoriasis, i.e. PSORS1 through PSORS12 (Alwan & Nestle, 2015). These loci are identified with genome wide linkage analysis in family-based studies. Additionally, the p40 subunit of IL-12 and IL23, the p19 subunit of IL-23 and IL-23 receptor are elements of genetic risk (Lowes, Suárez-Fariñas, & Krueger, 2014). A mutation in the CARD14 gene region is also found to be a psoriatic indicator (Jordan, Cao, Roberson, Duan, & Helms, 2012). The expression of these genes depends on environmental factors like cold and dry weather, stress, microorganisms, skin injuries, trauma and excessive alcohol intake. (Dika, Bardazzi, Balestri, & Maibach, 2007)

Environmental triggers induce stress on the cells and lead to the hyper activation of the immune system. First, dendritic cells are activated and produce cytokines like interleukins IL-12 and IL-23 in the lymph nodes. Those cytokines activate T-helper cells who migrate into skin tissue to present autoantigens and produce their own keratinocyte proliferating cytokines like interferon-γ (INF-γ) and tumor necrosis factor-

α (TNF-α). This will start and maintain a positive feedback cycle of inflammatory reactions which will lead to hyper-proliferation and overproduction of keratinocytes (figure 3) (Nestle, Kaplan, & Barker, 2009).

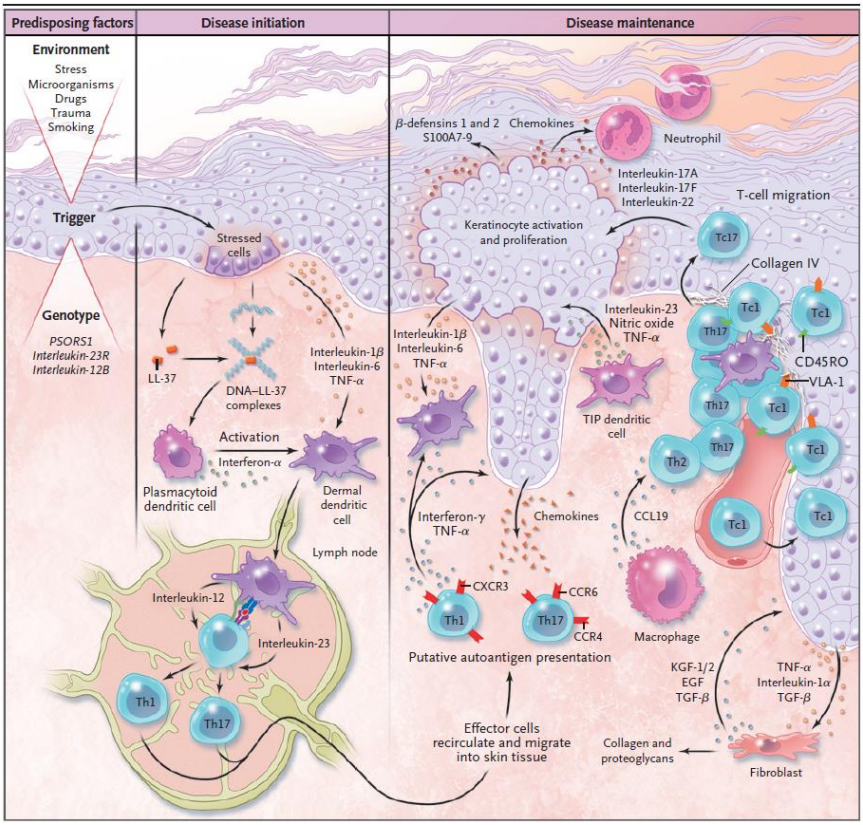

Figure 3. Proposed scheme of the evolution of a psoriatic lesion from initiation to maintenance of the disease. Figure taken from (Nestle, Kaplan, & Barker, 2009).

2. Materials and methods

2.1 Materials

To know which genes play an important role, it is necessary to analyse genetic data obtained from healthy and psoriatic people. In this study, datasets GSE13355 (Gudjonsson, Ding, Johnston, Tejasvi, & Guzman, 2010) and GSE41662 (Bigler, Rand, Kerkof, Timour, & Russell, 2013) are used, details on these datasets are shown in table 1. To create these datasets, 6mm punch biopsies were taken from lesional skin of psoriasis patients (PP), non-lesional skin (at least 10cm away from any active plaque) of psoriasis patients (PN) and normal skin from healthy controls (NN). Gene expression data was then acquired from these samples using Affymetrix platform HU133 Plus 2.0, which contain over 54.000 gene probes. These probes have a base sequence which is complementary to a target sequence of one specific gene. The data was normalized with Raw Multi-array Average (RMA) (Abecasis, 2017).

Table 1. Details of the datasets.

Dataset	Samples description	Number of samples	Reference
GSE13355	Lesional and non-lesional skin from psoriasis patients and normal healthy controls	58 PP 58 PN 64 NN	(Gudjonsson, Ding, Johnston, Tejasvi, & Guzman, 2010)
GSE41662	Lesional and non-lesional skin from psoriasis patients	24 PP 24 PN	(Bigler, Rand, Kerkof, Timour, & Russell, 2013)

2.2 Method

To bring down the number of genes from over 54000 to an accessible amount of 60, the microarray data is analysed. The approach comprised three steps. First, the microarray data is filtered. Second, the genes are clustered using k-means clustering. Finally, each gene-cluster is clustered based on the individual samples of lesional, non-lesional and healthy skin. The genes in the gene-clusters that provide the right sample cluster are selected and analysed. This is summarized in figure 4. The details of the steps are described in the following sections of this chapter.

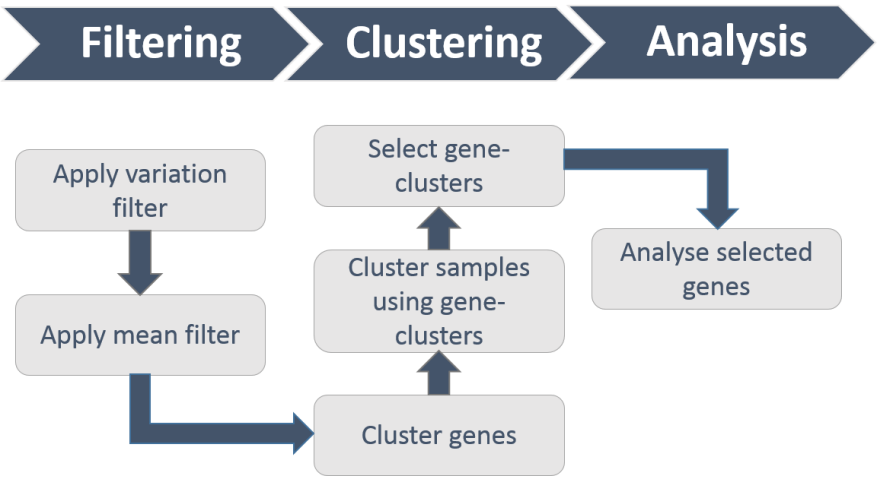

Figure 4. Scheme of the clustering analysis.

2.2.1 Filtering

Two filters are applied to the raw data. First, a low variance filter is used to delete 10% of the genes with the lowest variation in gene expression. Those genes are not relevant for this study. Then, the mean gene expression for healthy, non-lesional, and lesional samples is derived. The difference of these means is displayed in three boxplots. An example of a boxplot can be seen in figure 5. The outliers are of importance for this study due to the high variability of gene expression of those genes among tissue groups of interest. A gene is an outlier when the difference in mean gene expression value is greater than 75% + 1.5 * inner quartile range (IQR) of the total differences. When a gene is an outlier in all three of the boxplots, and so is differentially expressed in all three of the groups, the gene is selected. These genes are used to perform clustering, as explained in the next paragraph.

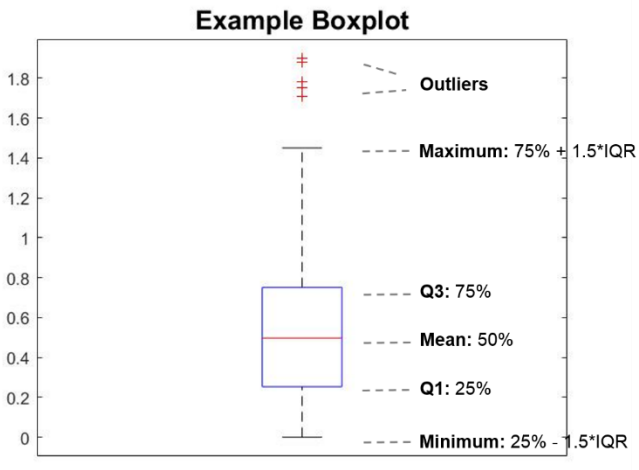

Figure 5. Example of a boxplot.

2.2.2 Clustering

The remaining genes are clustered with k-means, these clusters are called gene-clusters. K-means is an algorithm that assigns M data points in N dimensions into K clusters (Hartigan & Wong, 1979). The data points are assigned to the nearest centroid point in Euclidean space. This algorithm is iterated until the within-cluster sum of squares (WCSS) is minimal. The criterion used to determine the distance between the data points and the centroids is squared Euclidean distance, which expresses both homogeneity and separation of the clusters, see formula 1 (Aloise, Deshpande, & Hansen, 2009).

$$d(x, c) = (x - c)^2 \quad 1)$$

The maximum of iterations is set to 100 because it was observed that the cluster centroid varied less than 0.001 before even reaching the 100th iteration. The number of repetitions of clustering with another random initial cluster centroid position is called replicate and is set to 500.

To determine the minimum number of clusters (K), the total WCSS was analysed in relation to the number of clusters (figure 6). When adding another cluster does not lower the total WCSS significantly the optimal number of clusters is reached. In the graph this so called “elbow” is visible, marked with a red arrow.

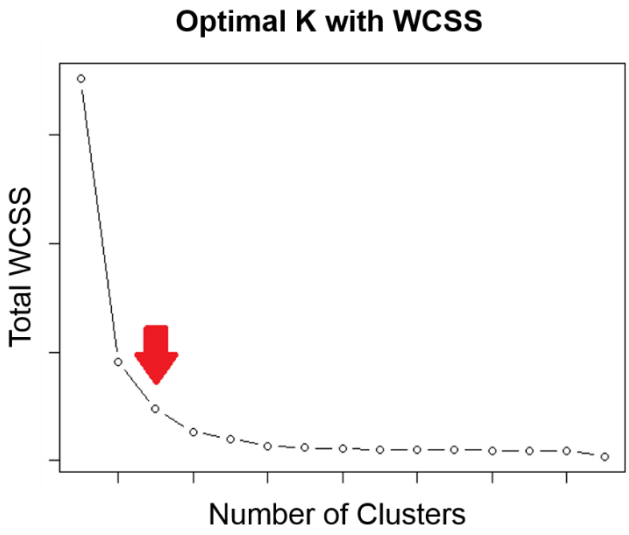

Figure 6. Example of optimal K with WCSS.

After the clustering, a silhouette value is calculated for all the genes in a gene-cluster. This silhouette value ranges from -1 to 1. The higher the value the better a data point, in this case a gene, belongs to its classified cluster (Rousseeuw, 1987). Negative or low positive silhouette ciphers indicate that bad clustering has taken place. Thus, gene-clusters with negative values are removed. The remaining gene-clusters are used to classify the samples into tissue groups (PP, PN or NN). This is also done with k-means clustering. If more than 95% of the samples are classified in the right group, the genes in the particular gene-cluster are selected. An example of a bad and good clustering for the samples has been given in figure 7. On the x-axis are the groups to which the samples belong. On the y-axis the sample-clusters are visible to which the samples are classified.

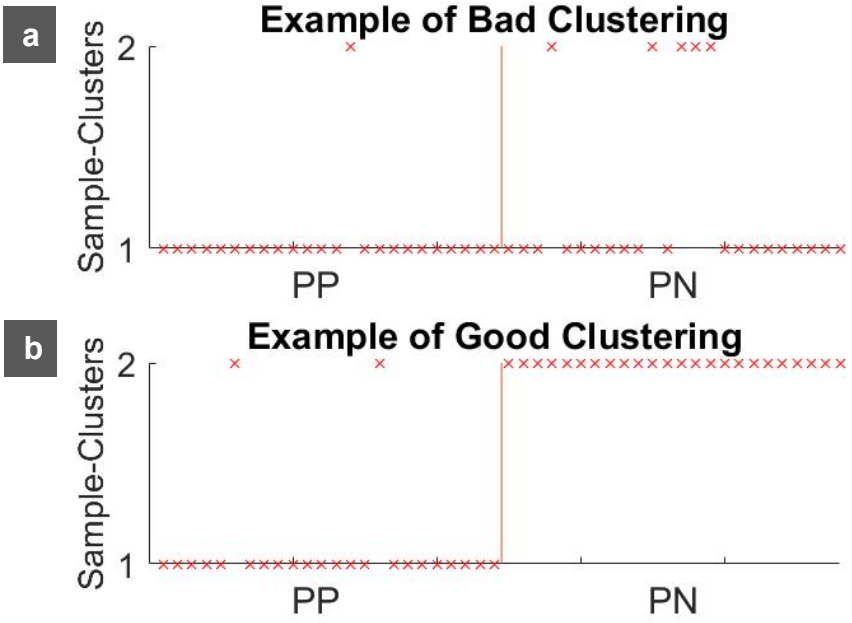

Figure 7. Example of **a.** samples that are badly clustered and **b.** samples that are correctly clustered, except for two.

2.2.3 Analysis

During the last step, the selected genes of both datasets are compared and the genes occurring in both are analysed. The function of the genes in the human body, the pathways they are involved in and the diseases they are associated with are looked up in The Human Gene Database (Weizmann Institute of Science, 2017).

3. Results

3.1 Filtering

The variation filter brought the number of genes in both datasets down from 54675 to 49207.

In figure 8 the boxplots of difference in mean gene expression between groups NN, PN and PP are shown of dataset GSE13355.

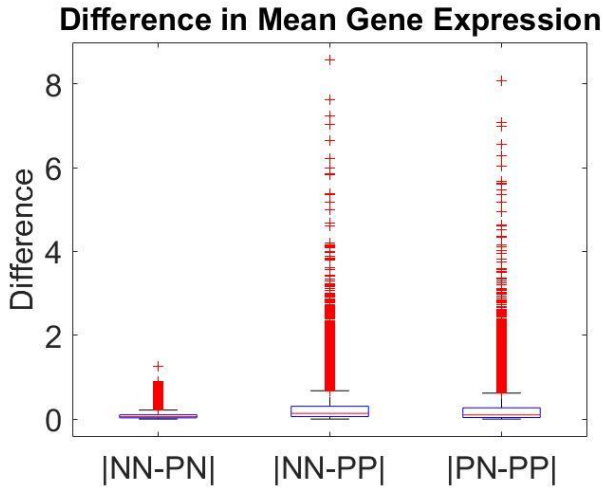

Figure 8. Boxplots of the difference in mean gene expression between the groups NN, PN and PP of dataset GSE13355.

The boxplot of the difference between NN and PN had a much smaller amount of outliers then the other two boxplots, 1917 compared to 3451 and 3357 for |NN-PP| and |PN-PP| respectively. The mean of the outliers of |NN-PN| is also lower than the maximum value of |NN-PP| and |PN-PP| (table 2 appendix). Thus, for the rest of the analysis in this work, the difference between the non-lesional skin from psoriasis patients and the healthy skin from controls is not included. The healthy and non-lesional groups are considered as one and are referred to as NNPN. In figure 9 the boxplots of GSE13355 and GSE41662 are shown.

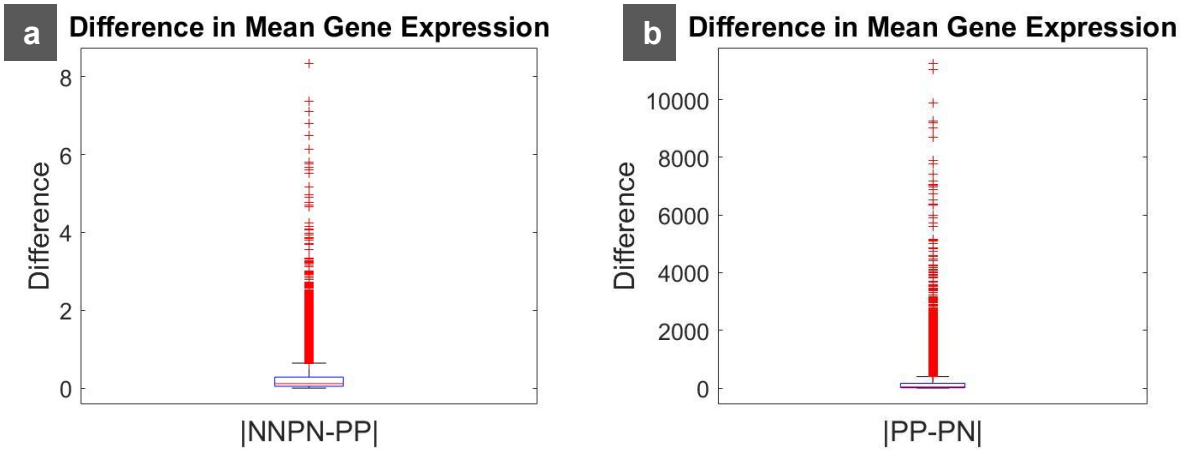

Figure 9. Boxplot of the differences in mean gene expression between the groups **a.** NNPN and PP from dataset GSE13355 and **b.** PP and PN from dataset GSE41662.

The 3442 outliers of GSE13355 and the 5561 outliers of GSE41662 are selected and used for clustering.

3.2 Clustering

At first a range of numbers of clusters from 1 to 500 is tested to find out which K is the correct choice to cluster the data (figure 10).

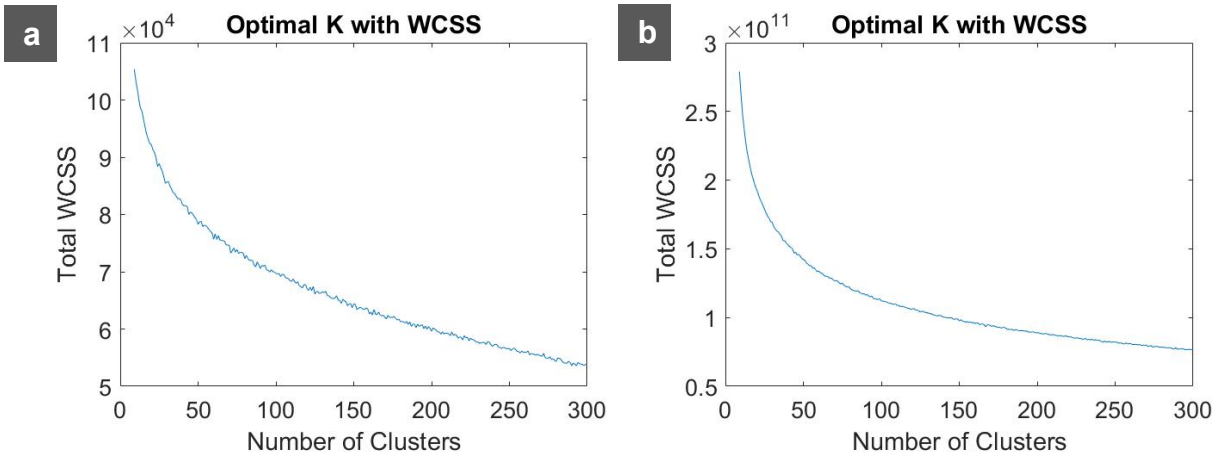

Figure 10. Optimal K with WCSS for **a.** GSE13355 and **b.** GSE41662

In both of these graphs an elbow shape is visible around K=50. So it is concluded that 50 is an optimal number of gene-clusters.

The silhouette value of all the genes in the gene-clusters are calculated and displayed in figure 11.

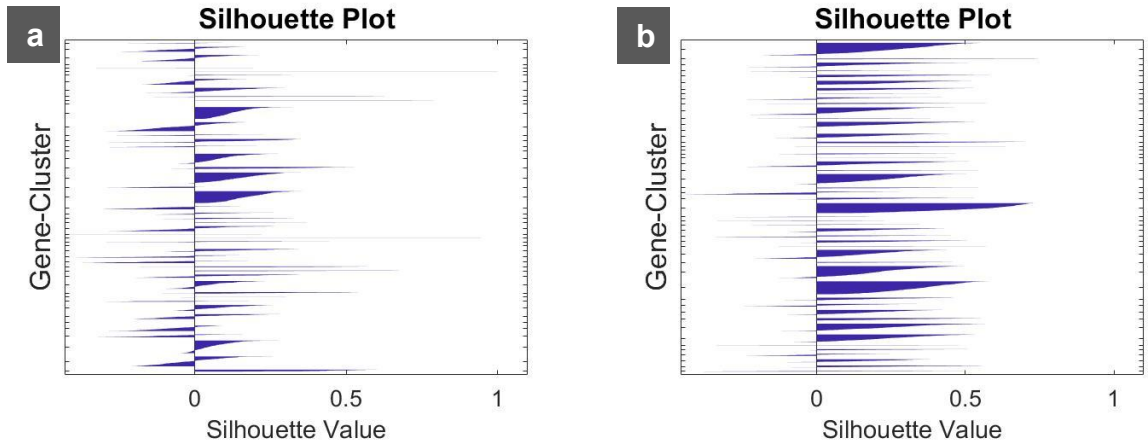

Figure 11. Silhouette plot of **a.** GSE13355 and **b.** GSE41662.

More than half of the gene-clusters contain negative silhouette values. The gene-clusters with genes with a silhouette value below zero are removed. The other gene-clusters are used to cluster the samples. For GSE13355, eleven gene-clusters remain, for GSE41662, this number is 10. The results of this clustering can be seen in figure 12. Above each graph, the number of well-clustered samples is displayed.

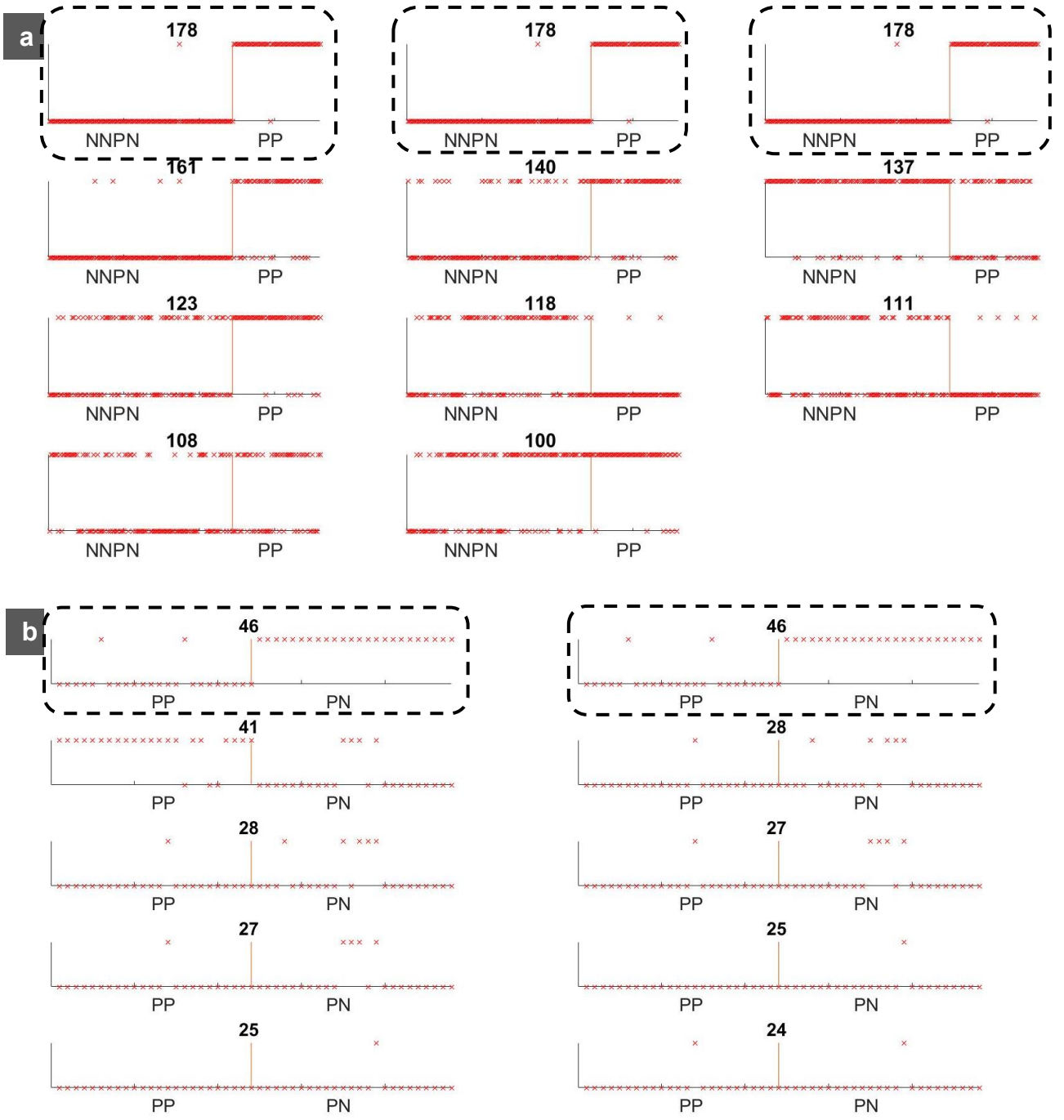

Figure 12. Samples clustered with remaining gene-cluster of datasets **a.** GSE13355 and **b.** GSE41662. The number of well clustered samples is displayed above each plot. When this number is greater than 95% of the total samples, the graph is delineated.

For GSE13355 the first three gene-clusters are selected, because 178 of the 180 samples are correctly classified. These gene-clusters contain a total of 563 genes. 349 genes are selected from GSE41662 divided over 2 gene-clusters. Both of these gene-clusters classified 46 of the 48 samples correct. The selected genes of both datasets are compared to each other and 63 genes occur in both. So, with the

filtering and clustering method, the number of genes is brought down from 54675 to an accessible amount of 63. These 63 genes are referred to as core genes and are analysed in the next paragraph.

3.3 Analysis

Of the 63 core genes, 18 are associated with a form of cancer, i.e. prostate cancer (e.g. WDR77), liver cancer (e.g. CDKN3) and breast cancer (e.g. CDK1). There are 2 genes found with an association to another dermal disease, Dermatophytosis, these are CCL22 and SLC7A1. Also, 2 genes are found with association to Cystic Firbrosis, APOL6 and S100A9. The last one, S100A9, has already been known to be a mediator of psoriasis (Schonthaler, et al., 2013).

Furthermore, 14 genes are involved in the immune system pathway including SLAMF7 and UBE2C and 11 genes in the metabolic pathway, among which HMOX2 and TIMM10. A function contributing to the cell cycle is found in 21 genes. This involves microtubule formation (e.g. AURKA), stabilizing chromosomes during mitose/meisose (e.g. NCAPG) or metaphase/anaphase promoting factors (e.g. NDC80).

Other interesting genes are IL1B and PTAFR as those play a role in pro-inflammatory processes. Interleukine 1 bèta (IL1B) mediates chronic inflammation. Binding of platelet activating factor (PAF) to its receptor (PTAFR) activates leukocytes.

For an overview of the 63 core genes and its functions, involved pathways, and associated diseases, see table 3 in the appendix.

4. Discussion and conclusion

A method has been made to analyse gene expression in Psoriasis. Two datasets are compared to each other and 63 genes are found to have an interesting gene expression. These genes could be seen as indicators of the disease. A number of these genes are involved in the immune system pathway, as expected, given that psoriasis is an autoimmune disease. Also the number of core genes involved in cancer is high. This agrees with a study of Pouplard et al. (Pouplard, et al., 2013) which found a small increased risk of cancer in psoriasis patients . Another explanation of the high number of genes involved in cancer is that both cancer and psoriasis involve malfunctioning in cell division, cell proliferation and apoptosis. Immune-suppressive treatments and UV-treatments used to treat psoriasis could also increase the change of cancer (Chiesa Fuxench, Shin, Ogdie Beatty, & Gelfand, 2016).

A study from Nair et al. (Nair, Duffin, Helms, & Ding, 2009) involving dataset GSE13355 found ten susceptible genetic loci with strong evidence of association with psoriasis. The eleven genes that are positioned near these loci where neither found in the selected genes of GSE13355 nor GSE41662 in this study. Another study involving dataset GSE13355 form Ainali et al. (Ainali, Valeyev, Perera, Williams, & Gudjonsson, 2012) found 24 important genes for psoriasis. He analysed gene expression through differential expression, a random forest classifier and comparison with biological literature. One of those genes, vanin 3 (VNN3), has also been found among the core genes of this study. Vanin 3 is expressed in the spinous cell layer of the epidermis and is induced by proinflammatory cytokines associated with psoriasis (Jansen, et al., 2009). A whole transcriptome analysis study found 29 differentially expressed genes in psoriasis patients compared to healthy controls, two of them are also found in this study, namely phospholipase A2, group IVD (PLA2G4D) and apolipoprotein L6 (APOL6) (Keermann, et al., 2015).
All of the three studies mentioned above indicated interleukin genes of importance, especially IL10, IL17, IL20, IL23, IL36. It is remarkable that only one gene coding for interleukin was found in the set of core genes in this study, IL1B. IL17A was selected in GSE13355. Further research concluded that the other interleukins of importance did not pass the mean filter. This means, those genes are not differentially expressed in lesional-skin compared to non-lesional and healty skin. Chances are, those interleukins were not removed during filtering if only difference in mean gene expression of lesional skin compared to healthy skin was looked into as was done in the study of Nair et al and Keermann et al.

Only 11% of the genes selected in GSE13355 are also selected in GSE41662. This implies inconsistencies between the two datasets. While there is only one difference in protocol. GSE13355 used cRNA as labelling method while GSE41662 used cDNA. Research of Bigler et al. (Bigler, Rand, Kerkof, Timour, & Russell, 2013) stated that gene expression of psoriatic lesions were consistent in different studies and consequently in different datasets. His research included datasets GSE13355 and GSE41662. But, note that the current study only refers to gene expression of PN and PP. Thus, the assumption that NN and PN of GSE13355 are considered as one group may explain the low similarity between the selected genes of these datasets. Therefore, a recommendation for next research with

these two datasets would be: either don't compare them with each other, or exclude the NN tissue group of dataset GSE13355 in your analysis.

In addition to the merging of these tissue groups, the clustering method has been done including the outliers of the difference between NN and PN as was initially stated in the method. After clustering the samples using the gene-clusters, there was no classification of NN or PN separately. This is not astonishing, considering figure 13 where principal component analysis visualised the coherence of NN and PN with respect to PP.

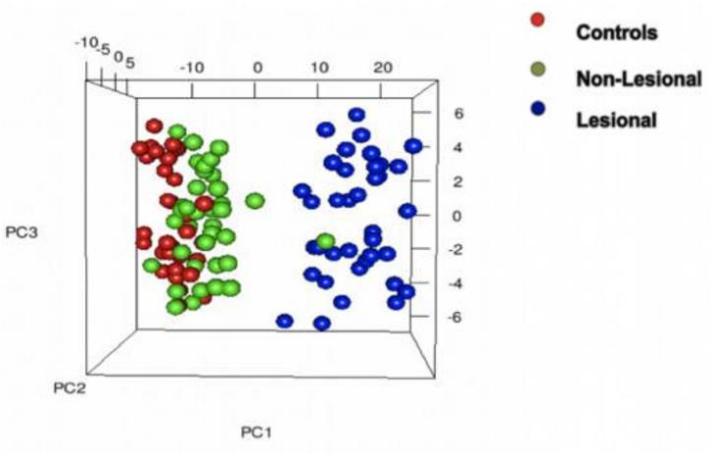

Figure 13. Principal Component Analysis to suggest sample clustering across skin types according to gene expression patterns. Good separation of inflamed (PP) and non-inflamed (PN, NN) tissues was observed, indicating a progression from normal (red) to lesional skin (blue) through the non-lesional cases (green). Figure taken from *(Ainali, Valeyev, Perera, Williams, & Gudjonsson, 2012)*

Keerman et al. did take a closer look at the difference between PN and NN to find genes involved in background inflammation. IL1F6 and IL6 were found to be of particular interest, as those genes are known to be involved in the pathogenesis of psoriasis.

The silhouette plot (especially for GSE13355) showed a lot of negative and low positive values. This could be an indication of a wrong number of clusters. To exclude this factor, variating the number of gene-clusters from 30 to 100 has been tried and a silhouette plot has been made of these different numbers of clusters. These plots showed the same pattern as K=50 and so this factor is excluded. Other factors that may have badly affected the silhouette values could be, number of replicates or filtering. Excluding all the gene-clusters that contain a negative silhouette value was necessary to reduce the number of genes but now, also genes are excluded with a positive silhouette value that might have been of interest for this study.

The initial goal of this study was to specify patient specific genes to be able to personalize treatment. To do this a closer look should be taken at the tissue samples that where not classified in the right sample group. As the expression of the genes in this gene-cluster is different compared to the expression of those genes in other patients. This would be interesting for a follow-up study.

The method here implemented to analyse microarray data is valuable to identify genes that increase the chance of psoriasis. However, here only gene expression at specific time points was analysed. The following questions are also important to explore: How does the protein encoded by the gene interact with other proteins and what is their influence on important pathway, such as the immune system? Are the genes up- or downregulated? And in what way do they affect keratinocytes? Thus, to complete this approach it is necessary to take a detailed look at the influence of the genes in the human body.

5. References

Abecasis, G. (2017). *Gene Expression Omnibus*. Retrieved from NCBI:

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE13355

Ainali, C., Valeyev, N., Perera, G., Williams, A., & Gudjonsson, J. E. (2012). Transcriptome classification reveals molecular. *BMC Genomics*, 13: 472.

Aloise, D., Deshpande, A., & Hansen, P. (2009). NP-hardness of Euclidean sum-of-squares clustering. *Machine Learning*, 75: 245–248.

Alwan, W., & Nestle, F. O. (2015). Pathogenesis and treatment of psoriasis: exploiting pathophysiological pathways for precision medicine. *Clinical and Experimental Rheumatology*, 33(93): S2-S6.

Bigler, J., Rand, H., Kerkof, K., Timour, M., & Russell, C. (2013). Cross-study homogeneity of psoriasis gene expression in skin across a large expression range. *PLoS One*, 8(1).

Burns, T., Breathnach, S., Cox, N., & Griffiths, C. (2010). *Rook's textbook of dermatology.* Oxford: UK: Wiley-Blackwell.

Centraal bureau voor statistieken. (2014). Retrieved from https://www.cbs.nl/nl-nl/nieuws/2015/46/ruim-1-2-miljoen-nederlanders-hebben-eczeem-of-psoriasis

Chandra, A., Ray, A., Senapati, S., & Chatterjee, R. (2015). Genetic and epigenetic basis of psoriasis pathogenesis. *Molecular Immunology*, 64(2):313-23.

Chiesa Fuxench, Z. C., Shin, D. B., Ogdie Beatty, A., & Gelfand, J. M. (2016). The Risk of Cancer in Patients With Psoriasis: A Population-Based Cohort Study in the Health Improvement Network. *JAMA Dermatology*, 152(3):282-90.

Dainichi, T., Hanakawa, S., & Kabashima, K. (2014). Classification of inflammatory skin diseases: A proposal based on the. *Journal of Dermatological Science*, vol 76: 81–89.

Dika, E., Bardazzi, F., Balestri, R., & Maibach, H. (2007). Environmental factors and psoriasis. *Current Problems in Dermatology*, 35:118-35.

Farber, E. M., & Nall, M. L. (1974). The natural history of psoriasis in 5,600 patients. *Dermatological*, 148: 1-18.

Gelfand, J., Feldman, S., Stern, R., Thomas, J., Rolstad, T., & Margolis, D. (2004). Determinants of quality of life in patients with psoriasis: a study from the US population. *Journal of the American Academy of Dermatology*, 51(5):704-8.

Gerrity, P. (n.d.). *Psoriasis vs Healthy Skin.* PegGerrity certified medical illustrator.

Gudjonsson, J. E., Ding, J., Johnston, A., Tejasvi, T., & Guzman, A. M. (2010). Assessment of the psoriatic transcriptome in a large sample: additional regulated. *Journal of Investigative Dermatology*, 130: 1829–1840.

Hartigan, J. A., & Wong, M. A. (1979). A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society.* , 28(1): 100-108.

Hönigsmann, H. (2001). Phototherapy for psoriasis. *Clinical and Experimental Dermatology*, 26(4): 343-350.

Jansen, P. A., Kamsteeg, M., Rodijk-Olthuis, D., van Vlijmen-Willems, I. M., de Jongh, G., Bergers, M., . . . Schalkwijk, J. (2009). Expression of the vanin gene family in normal and inflamed human skin: induction by proinflammatory cytokines. *Journal of Investigative Dermatology*, 129(9):2167-74.

Jordan, C., Cao, L., Roberson, E., Duan, S., & Helms, C. (2012). Rare and common variants in CARD14, encoding an epidermal regulator of NF-κB, in psoriasis. *The American Journal of Human Genetics*, 90: 796–808.

Keermann, M., Kõks, S., Reimann, E., Prans, E., Abram, K., & Kingo, K. (2015). Transcriptional landscape of psoriasis identifies the involvement of IL36 and IL36RN. *BMC Genomics*, 16:322 .

Kormeili, T., Lowe, N. J., & Yamauchi, P. S. (2004). Psoriasis: immunopathogenesis and evolving immunomodulators and systemic therapies; U.S. experiences. *British Journal of Dermatology*, 151(1): 3-15.

Lebwohl, M., & Ali, S. (2011). Treatment of psoriasis. Part 2. Systemic therapie. *Journal of the American Academy of Dermatology*, 45(5): 649-664.

Lowes, M. A., Suárez-Fariñas, M., & Krueger, J. G. (2014). Immunology of Psoriasis. *Annual Review of Immunology*, 32: 227–255.

Montagna, W., Kligman, A. M., & Carlisle, K. S. (1992). *Atlas of normal human skin.* New York: NY: Springer.

Nair, R., Duffin, K., Helms, C., & Ding, J. (2009). Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways. *Nature Genetics*, 41(2):199-204.

Nestle, F. O., Kaplan, D. H., & Barker, J. (2009). Psoriasis. *New England Journal of Medicine*, 361:496-509.

Parisi, R., Symmons, D. P., Griffiths, C. E., & Ashcroft, D. M. (2013). Global epidemiology of psoriasis: a systematic review of incidence. *Journal of Investigative Dermatology*, 133(2):377-85.

Pfaff, S., Liebman, J., & Born, M. (2015). Prospective Randomized Long-Term Study on the Efficacy and Safety of UV-Free Blue Light for. *Dermatology*, Vol. 231, pp. 24-34.

Pouplard, C., Brenaut, E., Horreau, C., Barnetche, T., Misery, L., Richard, M., . . . Paul, C. (2013). Risk of cancer in psoriasis: a systematic review and meta-analysis of epidemiological studies. *Journal of the European Academy of Dermatology and Venereology*, 3:36-46.

Rousseeuw, P. J. (1987). Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Computational and Applied Mathematics.*, 20: 53–65.

Samarasekera, E., Sawyer, L., Wonderling, D., Tucker, R., & Smith, C. (2013). Topical therapies for the treatment of plaque psoriasis: systematic review and network meta-analyses. *British Journal of Dermatology*, 168: 954–967.

Schonthaler, H., Guinea-Viniegra, J., Wculek, S., Ruppen, I., Ximénez-Embún, P., Guío-Carrión, A., . . . Wagner, E. (2013). S100A8-S100A9 protein complex mediates psoriasis by regulating the expression of complement factor C3. *Immunity*, 39(6):1171-81.

Shier, D., Lewis, R., & Butler, J. (n.d.). A section of skin an d the subcutaneous layer. *Hole's Human Anatomy and Physiology.* McGraw-Hill, New York.

Shimizu, H. (2007). Structure and Function of the Skin. In H. Shimizu, *Shimizu's textbook of dermatology.* Japan: Hokkaido University Press/Nakayama Shoten.

Shwayder, T., & Akland, T. (2005). Neonatal skin barrier: structure, function, and disorders. *Dermatology and Therapy*, 18(2):87-103.

Smith, C. H., & Barker, J. N. (2006). Psoriasis and its management. *British Medical Journal*, 333(7564): 380–384.

Swanson, J. R., & Melton, J. L. (n.d.). Haematoxylin and eosin stained section of normal skin ×20 magnification. *Jason R. Swanson and Jeffrey L. Melton, M.D.*

University of Iowa. (1995). *Dermatology.* University of Iowa, Iowa.

Weinstabl, A., Hoff-Lesch, S., Merk, H. F., & von Felbert, V. (2011). Prospective randomized study on the efficacy of blue light in the treatment of psoriasis vulgaris. *Dermatology*, 223(3):251-9.

Weizmann Institute of Science. (2017, 7 4). *The Human Gene Database*. Retrieved from GeneCards: http://www.genecards.org/

6. Appendices

Table 2. Values obtained from the boxplots of figure 8a.

	|NN-PN|	**|NN-PP|**	**|PP-PN|**
N outliers	1917	3451	3357
Mean outliers	0.2914	1.0769	1.0148
Maximum value (Q3+1.5*IQR)	0.2141	0.6750	0.6182

Table 3. Analysis of the 63 core genes.

Gene	**Function**	**Involved pathway**	**Associated diseases**
ADAP2	binds beta-tubulin and increases the stability of microtubules	B Cell Receptor Signaling Pathway	Chromosome 17Q11.2 Deletion Syndrome, 1.4-Mb
ALDH4A1	catalyzes the second step of the proline degradation pathway, converting pyrroline-5-carboxylate to glutamate	Metabolism and Arginine and proline metabolism.	Hyperprolinemia
APOL6	found in the cytoplasm, where it may affect the movement of lipids	Cholesterol and Sphingolipids transport / Recycling to plasma membrane in lung (normal and CF	Cystic Fibrosis
ATP6V1D	mediates acidification of eukaryotic intracellular organelles	Immune System and Ion channel transport	
AURKA	involved in microtubule formation and/or stabilization at the spindle pole during chromosome segregation	Integrated Breast Cancer Pathway and Regulation of PLK1 Activity at G2/M Transition	Colorectal Cancer and Colorectal Adenocarcinoma.
BAK1	act as anti- or pro-apoptotic regulators	Integrated Breast Cancer Pathway and Apoptosis and survival Regulation of Apoptosis by Mitochondrial Proteins	Benign Mammary Dysplasia and Chronic Gonococcal Salpingitis
BUB1B	plays a role in the inhibition of the anaphase-promoting complex/cyclosome (APC/C), delaying the onset of anaphase and ensuring proper chromosome segregation	HTLV-I infection and Mitotic Metaphase and Anaphase.	Mosaic Variegated Aneuploidy Syndrome 1 and Mosaic Variegated Aneuploidy Syndrome Many forms of cancer
CARHSP 1	Binds mRNA and regulates the stability of target mRNA		
CASP1	plays a central role in the execution-phase of cell apoptosis	Immune System and G-protein signaling_Regulation of CDC42 activity.	Cowpox and Shigellosis
CCL22	displays chemotactic activity for monocytes, dendritic cells, natural killer cells and for chronically activated T lymphocytes.	Immune System and PEDF Induced Signaling.	Atopic Dermatitis and Eosinophilic Pneumonia.

CCNB1	regulatory protein involved in mitosis	Mitotic G1-G1/S phases and Arrhythmogenic right ventricular cardiomyopathy (ARVC)	Thyroid Lymphoma and Adrenal Carcinoma.
CCNB2	essential components of the cell cycle regulatory machinery	HTLV-I infection and Arrhythmogenic right ventricular cardiomyopathy (ARVC)	Anauxetic Dysplasia and Breast Cancer
CCNE2	plays a role in cell cycle G1/S transition	Mitotic G1-G1/S phases and GPCR Pathway	A significantly increased expression level of this gene was observed in tumor-derived cells
CD274	immune inhibitory receptor ligand that is expressed by hematopoietic and non-hematopoietic cells, such as T cells and B cells and various types of tumor cells.	Immune System and IgA-Producing B Cells in the Intestine.	Scrotum Pagets Disease and Lymphoepithelioma-Like Carcinoma.
CDC20	regulatory protein interacting with several other proteins at multiple points in the cell cycle	HTLV-I infection and Regulation of activated PAK-2p34 by proteasome mediated degradation	
CDK1	catalytic subunit of the M-phase promoting factor (MPF), which is essential for G1/S and G2/M phase transitions of eukaryotic cell cycle	mediated regulation of DNA replication and Mitotic G1-G1/S phases.	Breast Cancer and Hepatocellular Carcinoma
CDKN3	cyclin-dependent kinase inhibitor overexpressed in several kinds of cancers		Hepatocellular Carcinoma and Bannayan-Riley-Ruvalcaba Syndrome
CKAP4	Mediates the anchoring of the ER to microtubules	Immune System and Surfactant metabolism	Interstitial Cystitis and Exstrophy-Epispadias Complex.
DLGAP5	Cell cycle regulator that may play a role in carcinogenesis of cancer cells.	Aurora A signaling.	Hepatocellular Carcinoma.
ECT2	important role in the regulation of cytokinesis	Signaling by GPCR and G-protein signaling_Regulation of CDC42 activity.	
EIF4E2	Recognizes and binds the 7-methylguanosine-containing mRNA cap during an early step in the initiation of protein synthesis	Interferon gamma signaling and Immune System.	
ENTPD7	Preferentially hydrolyzes nucleoside 5-triphosphates	Metabolism and Purine metabolism	
EPHB2	involved in diverse cellular processes including motility, division, and differentiation	Activation of cAMP-Dependent PKA and GPCR Pathway.	Prostate Cancer/Brain Cancer Susceptibility, Somatic and Prostate Cancer.
FAM83D	proto-oncogene that regulates cell, growth,		

	migration and epithelial to mesenchymal transition, may also play a role in cell proliferation		
HMMR	involved in cell motility	Metabolism and Regulation of PLK1 Activity at G2/M Transition.	Breast Cancer and Fibrosarco ma.
HMOX2	essential enzyme in heme catabolism	Metabolism and Immune System.	Neonatal Jaundice and Glass Syndrome
IL1B	important mediator of the inflammatory response, and is involved in a variety of cellular activities, including cell proliferation, differentiation, and apoptosis	Interleukin-1 signaling and AGE-RAGE signaling pathway in diabetic complications	Gastric Cancer Risk After H. Pylori Infection and Periodont al Disease
KIF20A	May act as a motor required for the retrograde RAB6 regulated transport of Golgi membranes and associated vesicles along microtubule	Immune System and Vesicle-mediated transport	Charcot-Marie-Tooth Disease, Type 4C.
LRP8	play roles in both signal transduction and receptor-mediated endocytosis of specific ligands for lysosomal degradation	Metabolism of fat-soluble vitamins and Metabolism.	Myocardial Infarction and Myocardi al Infarction, Susceptibility To, 1
MELK	involved in various processes such as cell cycle regulation, self-renewal of stem cells, apoptosis and splicing regulation	Neuroscience.	
MKI67	s associated with and may be necessary for cellular proliferation	Primary Focal Segmental Glomerulosclerosis FSGS and Neuroscience	Cervical Intraepithelial Neoplasia and Pancreat ic Neuroendocrine Tumor
MRPS17	help in protein synthesis within the mitochondrion	Mitochondrial translation and Viral mRNA Translation.	
NCAPG	responsible for the condensation and stabilization of chromosomes during mitosis and meiosis	Cell cycle_Chromosome condensation in prometaphase and Aurora B signaling	
NDC80	functions to organize and stabilize microtubule-kinetochore interactions and is required for proper chromosome segregation	Mitotic Metaphase and Anaphase and Aurora B signaling.	Female Reproductive Organ Cancer.
NDRG4	required for cell cycle progression and survival in primary astrocytes and may be involved in the regulation of mitogenic signalling in vascular smooth muscles cells	Wnt / Hedgehog / Notch and Apoptosis and Autophagy	Infantile Myofibromatosis and R eflex Epilepsy.
NPM3	related to the nuclear chaperone		Chondrosarcoma, Extraskeletal Myxoid

	phosphoproteins, nucleoplasmin and nucleophosmin		
NUSAP1	plays a role in spindle microtubule organization		
PBK	involved in the activation of lymphoid cells and support testicular functions, overexpression of this gene has been implicated in tumorigenesis	DNA Damage.	
PLA2G4D	catalyze the hydrolysis of glycerophospholipids at the sn-2 position and then liberate free fatty acids and lysophospholipids	Metabolism and Acyl chain remodelling of PE	
PRKCQ-AS1			
PRR11	Plays a critical role in cell cycle progression.		
PTAFR	plays a significant role in oncogenic transformation, tumor growth, angiogenesis, metastasis, and pro-inflammatory processes.	Interferon gamma signaling and Signaling by GPCR	Melanoma Metastasis
RCC1	Plays a key role in nucleo-cytoplasmic transport, mitosis and nuclear-envelope assembly	Transport of the SLBP independent Mature mRNA and HIV Life Cycle.	Renal-Hepatic-Pancreatic Dysplasia and Retinitis Pigmentosa.
RGS1	attenuates the signalling activity of G-proteins	Signaling by GPCR and Peptide ligand-binding receptors	
S100A9	involved in the regulation of a number of cellular processes such as cell cycle progression and differentiation	Immune System and Activated TLR4 signalling.	Cystic Fibrosis and Psoriasis.
SERPINB3	may act as a papain-like cysteine protease inhibitor to modulate the host immune response against tumor cells	Immune System and Amoebiasis	Cervical Squamous Cell Carcinoma and Anus Cancer
SERTAD1	Renders activity of CDK4, which involves in tumorigenesis of a variety of cancers		
SHMT2	catalyzes the reversible reaction of serine and tetrahydrofolate to glycine and 5,10-methylene tetrahydrofolate	Glycine, serine and threonine metabolism and Metabolism	Pyridoxine Deficiency
SLAMF7	involved in the regulation and interconnection of both innate and adaptive immune response	Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell and Immune System	

SLC7A1	involved in the transport of the cationic amino acids (arginine, lysine and ornithine) in non-hepatic tissues	Transport of glucose and other sugars, bile salts and organic acids, metal ions and amine compounds and MicroRNAs in cancer	Tinea Capitis and Dermatophytosis
SLC7A11	predominant mediator of Kaposi sarcoma-associated herpesvirus fusion and entry permissiveness into cells	Transport of glucose and other sugars, bile salts and organic acids, metal ions and amine compounds and Cell surface interactions at the vascular wall	Kaposi Sarcoma and Dyscalculia
STAT3	mediates the expression of a variety of genes in response to cell stimuli, and thus plays a key role in many cellular processes such as cell growth and apoptosis	IL27-mediated signaling events and CXCR4-mediated signaling events.	Hyper-Ige Recurrent Infection Syndrome and Autoimmune Disease, Multisystem, Infantile-Onset.
SYNCRIP	plays a role in multiple aspects of mRNA maturation	Translational Control	
TIMM10	mediate the import and insertion of hydrophobic membrane proteins into the mitochondrial inner membrane	Mitochondrial protein import and Metabolism of proteins.	
TPX2	Required for normal assembly of microtubules during apoptosis	Regulation of PLK1 Activity at G2/M Transition and Gene Expression.	Hepatocellular Carcinoma.
TRIM14		Interferon gamma signaling and Immune System	
UBE2C	important cellular mechanism for targeting abnormal or short-lived proteins for degradation	Immune System and Regulation of activated PAK-2p34 by proteasome mediated degradation.	
UBE2N	suggest that this protein plays a role in DNA postreplication repair	Interferon gamma signaling and Activated TLR4 signalling	
UBE2T	catalyzes the covalent attachment of ubiquitin to protein substrates	Fanconi anemia pathway and Metabolism of proteins	Fanconi Anemia, Complementation Group T and Ube2t-Related Fanconi Anemia
VNN3		Metabolism of proteins and Post-translational modification-synthesis of GPI-anchored proteins	
WDR77	involved in the early stages of prostate cancer, with most of the protein being nuclear-localized in benign cells but cytoplasmic in cancer cells	Transport of the SLBP independent Mature mRNA and Activated PKN1 stimulates transcription of AR (androgen receptor) regulated genes KLK2 and KLK3	Prostate Cancer.

ZNF593	transcriptional regulatory activity		

MATLAB script

```
%% load data
load('matlabGSE13355.mat')
load('GPLData.mat')
%% Values
k=50;
K=2;
indexNN=1:64;
indexPN=65:122;
indexNNPN=1:122;
indexPP=123:180;
nsample=180
%% filtering
Mask=genevarfilter(gseData.Data, 'Percentile',10); %apply
variation filter of 10%
V10Data=gseData.Data(Mask,:);

nGenesAfterV=numel(V10Data.RowNames)

nNN=numel(indexNN) %count the number of samples per group
nPN=numel(indexPN)
nPP=numel(indexPP)
nNNPN=numel(indexNNPN)

for i=1:nGenesAfterV
    expressionNNPN=sum(V10Data(i,indexNNPN));
    meanNNPN=sum(expressionNNPN)/nNNPN; %calculate the
mean gene expression per group
    expressionPP=sum(V10Data(i,indexPP));
    meanPP=sum(expressionPP)/nPP;
    diff=abs(meanNNPN-meanPP); %calculate the difference
in mean gene expression
    Difference(i)=diff;
end

% figure, %visualize difference
% boxplot(Difference, 'labels', {'|NNPN-PP|'})
% title('Difference in mean gene expression')
% ylabel('Difference')

ThresholdOutlier=quantile(Difference,0.75)+1.5*iqr(Differ
ence)

for i=1:nGenesAfterV %this for loop gives the indeces of
the outliers.
    if ge(Difference(i),ThresholdOutlier)
        index(i)=1;
```

```
    else
        index(i)=0;
    end
end

sum(index)
Index=logical(index);
V10MData=V10Data(Index,:); %creat new dataset with only
filtered data

%% clusteren
ngene=numel(V10MData.RowNames)
% create gene-clusters
[idxk50r20g3442, ctrs] = kmeans(V10MData, k,
'dist','sqeuclidean', 'rep',500, 'display', 'final');

for i=1:k;  %calculates number of genes per cluster
    t(i)=sum(idxk50r20g3442==i);
end
maxn=max(t);

for i=1:k
    X{i}=find(idxk50r20g3442==i,maxn); %gives geneindeces
per cluster
end

DV10MData=double(V10MData);
s=silhouette(DV10MData,idxk50r20g3442); %calculate
silhouette value for every gene in every gene-cluster

for i=1:k
    indexlist=X{1,i};
    length=numel(indexlist);
    for j=1:length
        index=indexlist(j);
        if ge(s(index),0)
            Genegezero{i,j}=1;
        else
            Genegezero{i,j}=0;
        end
    end

end

for i=1:k %create list (xt) with value 1 for gene-
clusters without silhouette values beneath zero
    vector=Genegezero(i,:);
    Vector=cell2mat(vector);
```

```
    NOnes=sum(Vector);
    Ntot=t(i);
    NZeros=Ntot-NOnes;
    percOnes=NOnes/Ntot;
    percZeros=NZeros/Ntot;
    if ge(percOnes,0.99)
        xt(i)=1;
    else
        xt(i)=0;
    end
end

for i=1:k  %create cells with gene-clusters
    clustindex=X{1,i};
    maxs=size(X{1,i});
    g=1:maxs;
    w=[];
    w([clustindex(g)])=1;
    W=logical(w);
    WW=W';
    Cluster{i}=V10MData(WW,:);
    clusterT{1,i}=Cluster{1,i}';
end

for i=1:k %cluster samples

     [idxs,cent]=kmeans(clusterT{1,i},K,'dist','sqeuclidea
     n', 'rep',500);
    Y{i,1}=idxs;
end

for i=find(xt==1) %Gives for the left gene-clusters the
number of right clustered samples
    clust=Y{i,1};
    clustNNPN=clust(indexNNPN);
    clustPP=clust(indexPP);

    for j=1:2
        num=sum(clust(:)==j);
        numNNPN=sum(clustNNPN(:)==j);
        numPP=sum(clustPP(:)==j);

        G(1,j)=numNNPN;
        G(2,j)=numPP;

    end
    G1=G(1,1)+G(2,2);
```

```
    G2=G(1,2)+G(2,1);
    if ge(G1,G2)
        NumGood(i)=G1;
    else
        NumGood(i)=G2;
    end

end

nWC=sort(NumGood,'descend');
nClustersLeft=sum(xt);
sortnWC=nWC(1:nClustersLeft);

figure,
for i=1:k %display the the left gene-clusters and create
ImportantGenes which contains cell with selected genes
    nWellClustered=sortnWC(i);
    cluster=find(NumGood==nWellClustered);
    Sample=1:nsample;
    if ge(nWellClustered,171)
        ImportantGenes{i}=Cluster{1,cluster(1)}.RowNames;
    end
    idxs=Y{cluster,1};
    subplot(4,4,i)
    scatter(Sample,idxs,'r','x')

    hold on
    y(1:K)=nNNPN;
    z=1:K;
    plot(y,z)
    hold off


    formatspec='%d %';
    str=sprintf(formatspec,nWellClustered);
    title(str);
    xlabel 'NNPN PP'
    set(gca, 'XTickLabel','')
    ylabel ''
    set(gca, 'YTick', [1,2])
end

GPLData=gplData.Data;
rijmetID=GPLData(:,1);
rijmetgenen=GPLData(:,10);
rijmetgenesymb=GPLData(:,11);
```

```
Genes13355=[ImportantGenes{1,1};ImportantGenes{1,2};Impor
tantGenes{1,3};ImportantGenes{1,4}];
nimpgene=numel(ImportantGenes{1,1})+numel(ImportantGenes{
1,2})+numel(ImportantGenes{1,3})+numel(ImportantGenes{1,4
});

for i=1:nimpgene %gives the names of the selected genes
    grod=strcmp(rijmetID, Genes13355{i});
    sds=find(grod==1);
    names13355(i)=rijmetgenesymb(sds);
end
```