Stockholms universitet

Do the Chemical Properties of Amino Acids Have an Impact on the Risk of Rheumatoid Arthritis?

Hannah Rossland

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2018:19
Matematisk statistik
Juni 2018

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Mathematical Statistics
Stockholm University
Bachelor Thesis **2018:19**
http://www.math.su.se

Do the Chemical Properties of Amino Acids Have an Impact on the Risk of Rheumatoid Arthritis?

Hannah Rossland*

June 2018

Abstract

Rheumatoid arthritis is a chronic and autoimmune joint disease. Despite that it is the most common inflammatory joint disease that affects around 0.5-1% of the population it is unknown why the disease occurs. The disease develops when the immune system fails to distinguish between self and non-self antigens and mistakenly attacks its own tissue. The main unit responsible for this function in the human body is the human leucocyte antigen system consisting of amino acids which determine their characteristics. There are theories about that the chemical properties of the amino acids can have an impact on the development of the disease and this is what is going to be investigated in this report. Accordingly, the question is: do the chemical properties of amino acids have an impact on the risk of rheumatoid arthritis?

The main approach was logistic regression with data from a case-control study. For each amino acid three properties (hydrophilicity, bulk and electronic properties) were used as explanatory variables, one at a time, in simple logistic regression models, and also multiple logistic regression models, with all three of them included. Because two different approaches were used initially to reduce the original data in consequence of missing values, two subsets have been treated parallel, with slightly different results.

The result in this study concluded that the properties of some amino acids in a couple of already known positions associated with risk of rheumatoid arthritis do have an impact on the presence of the disease. Because most of the results in this study correspond to findings of previous studies, it indicates that an extension of this work would seemingly be an appropriate future work to continue investigating the connection between amino acids in the human leucocyte antigen region and risk of rheumatoid arthritis.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: h.rossland@live.se. Supervisor: Disa Hansson, Ola Hössjer.

Contents

1 Introduction

Rheumatoid arthritis (RA) is an inflammatory, chronic and autoimmune disease with main symptoms such as swelling, pain and stiffness in the joints. It is the most common inflammatory joint disease occurring when the immune system fails to distinguish between self and non-self antigens and mistakenly attacks its own tissue. The main unit responsible for this commission in the human body is the Human Leucocyte Antigen System, a system with a wide-ranging genetic variation and therefore a diversity in the response of different peoples' immune system. The antigens consist of amino acids which determines their characteristics. Even though there are only 20 possible amino acids, the variation among the sequences is huge. There are some similar properties for some amino acids, but they are not entirely interchangeable and a replacement can work out or be critical depending on the circumstances. [1]

Why rheumatoid arthritis occurs is unknown but there are theories about that the chemical properties of amino acids can have an impact on the development of the disease. Between different ethnic groups the prevalence of human leucocyte antigens varies, and this may also be related to their role in the prevalence of the disease in different parts of the world. [4]

Currently the aim is to give an early diagnose in order to as soon as possible start an individualized treatment aimed at remission or at least enable a low disease activity. With this more customized approach the progression of joint damage is prevented, and the objective is that life in general, with physical functioning, social participation and work will not be affected. Rheumatoid arthritis can still not be cured, but as mentioned, remission is an achievable goal. The improved understanding of the disease and the development of it during the past two decades have enlarged both the diagnosing process and the development of new drugs and accordingly the lives of patients with rheumatoid arthritis. [3, 4, 5]

With the objective to increase the clarity about the theories that the chemical properties of amino acids can have an impact on the development of the disease, this will be investigated and the question is thus: Do the chemical properties of amino acids have an impact on the risk of rheumatoid arthritis?

1.1 Aim and Limitations

To address the problem there are several different scales of measurements for the variables of the amino acids that exist and can be used. One constraint is that the so called Hellberg z-scales are chosen for this purpose.

1.2 Outline

In section 2 the biological background about the disease, the human leucocyte antigen system and amino acids will briefly be described. Starting with the human leucocyte antigen system in subsection 2.1 we then introduce amino acids in subsection 2.2 and rheumatoid arthritis in subsection 2.3. In subsection 2.4 we explain what a genome-wide association study is and in subsection 2.5 we end with a description of the data.

Section 3 covers the mathematical theory behind the methods used. First in subsection 3.1 we explain principal component analysis, what that is and how to derive the principal components. Subsection 3.2 deals with logistic regression, how to interpret the results and how to perform it with nominal variables. Section 3 ends with subsection 3.3 that describes how to perform a test of independency between two variables.

The methods used in this study will be explained in section 4. First how the data was reduced in subsection 4.1 and then the principal component analysis and logistic regression in subsections 4.2 and 4.3 respectively.

Section 5 contains the analysis starting with subsection 5.1, where it will be demonstrated more precisely how the data was reduced and in subsection 5.2 the results are presented.

The final section 6 includes a discussion, where at first the results will be analyzed in a bigger context in subsection 6.1. Some comments about the methods are found in subsection 6.2, with different ways of handling missing values and the risk of losing information in subsection 6.2.1 and 6.2.2 respectively. Subsection 6.2.3 deals with how the variables were treated and could have been treated before the principal component analysis, and as an expansion of different ways to treat the variables, subsection 6.2.4 covers different approaches for categorization before the logistic regression. The whole section ends with some suggestions for future work in subsection 6.3.

2 Background

2.1 Human Leukocyte Antigen System

MHC (Major Histocompatibility Complex) is the genetic region consisting of more than 200 genes located close together on chromosome six. All genes in this complex can be divided into three classes, I, II and III. The human major MHC is called the Human Leucocyte Antigen System (HLA) and is known for being the most polymorphic genetic system in the human body, meaning this is the system with the most genetic variation. The main function for the human leucocyte antigen system is the control of self-recognition, to distinguish between the body's own proteins and intruders, as a defense mechanism against microorganisms. [3]

Because of the polymorphism of the human leucocyte antigen system, different peoples' immune response will react in different ways. Some of the human leucocyte antigens have hundreds of identified versions, so called alleles. All alleles get a name, for example HLA-B27 and if they are closely located they are categorized together. For HLA-B27 there are 43 subgroups, and these are designated HLA-B2701 up to HLA-B*2743. [4]

The human leucocyte antigens of both class I and class II consist among other things of amino acids which determines their characteristics. Some diseases and especially autoimmune diseases are associated with some alleles. How strong the association is varies between different diseases. It is usually unclear what role the human leucocyte antigen genes play in the risk of developing these diseases. Usually other genetic and environmental factors may also be of importance, and the function of some human leucocyte antigens is still unknown. [3]

The distribution and prevalence of human leucocyte antigens vary considerably between different ethnic groups, and this may also be related to the role of the human leucocyte antigen molecules in the prevalence of various diseases in different parts of the world. [4]

2.2 Amino acids

Many biological processes can partly be explained by the amino acids positions on the human leucocyte antigen and their properties. Even if there are only 20 different amino acids that can be incorporated into a protein; in an amino acid sequence, called a peptide, the variation can still be enormous. Nowadays when millions of these sequences are known it means we have knowledge of a lot of mutations that may occur. Some mutations can be crucial and cause diseases while some are more subtle. Whether or not the mutation has a drastic effect on the protein function is often unknown. Some amino acids have similar properties, but they are not completely commutable and can not always replace each other, a substitution in one context can be critical in another. [1]

The amino acids can be described with some simple descriptors such as molecular weight, volumes and polar/non-polar surfaces and these can be used in chemobioinformatics investigations. A problem that arises is that only a few of these would not be sufficient to describe all important physiochemical properties and by instead using a large set of descriptors, a hopeless amount of descriptors would have to be used. Fortunately, many of the properties are to a certain extent correlated and by using a principal component analysis (as will be explained in section 3) the correlated descriptors can be reduced to a lower number of uncorrelated descriptors. Several so called principal amino acid property scales have been developed and the so called Hellberg z-scales will be used here, where 29 physiochemical variables for the 20 amino acids have been reduced to only three components. The three main components z_1, z_2 and z_3 describe most of the variation in the sets of peptides and can tentatively be interpreted as the properties hydrophilicity, bulk and electronic properties. With these three scales about 70 % of the variation in the properties of the amino acids are captured. [2, 9]

Amino acid	z_1	z_2	z_3
Alanine	0.07	-1.73	0.09
Valine	-2.69	-2.53	-1.29
Leucine	-4.19	-1.03	-0.98
Isoleucine	-4.44	-1.68	-1.03
Proline	-1.22	0.88	2.23
Phenylalanine	-4.92	1.30	0.45
Tryptophan	-4.75	3.65	0.85
Methionine	-2.49	-0.27	-0.41
Lysine	2.84	1.41	-3.14
Arginine	2.88	2.52	-3.44
Histidine	2.41	1.74	1.11
Glycine	2.23	-5.36	0.30
Serine	1.96	-1.63	0.57
Threonine	0.92	-2.09	-1.4
Cysteine	0.71	-0.97	4.13
Tyrosine	-1.39	2.32	0.01
Asparginie	3.22	1.45	0.84
Glutamine	2.18	0.53	-1.14
Aspartic acid	3.64	1.13	2.36
Glutamic acid	3.08	0.39	-0.07
Min value	-4.92	-5.36	-3.44
Max value	3.64	3.65	4.13

Table 1: The Hellberg z-scales.

2.3 Rheumatoid arthritis

Rheumatoid arthritis (RA) is an inflammatory, chronic and autoimmune disease that mainly affects the joints. The main symptoms are swelling in the joints, pain and stiffness, together with decreased appetite and fatigue. [4] For the majority of cases the disease development begins years before clinical disease is evident. When insufficiently treated rheumatoid arthritis occurs, it leads to irreversible joint damage and disability. [4, 5] Most studies in rheumatoid arthritis have been done in Western countries where it shows a prevalence of 0.5-1%. There is a lack of epidemiological studies in some regions, but the prevalence seems to be rather homogeneous around the world. However, some ethnicities stand out, in the native American population a high prevalence of 5-6% has been reported and in rural Africa a notably low prevalence. [4, 5] The disease may occur at any age but it is most common in the age span 40-70, with mean of 66 years, thus the occurrence increases with age. [4] Some of the risk factors of the disease are genetics, sex (where females possess an enlarged risk with a ratio of 2-3 compared to men) and some environmental factors such as smoking, silica exposure, vitamin D deficiency and obesity. For some of these factors the studies are not that rigorous and it is incompletely understood how they contribute to the disease. [5]

No diagnostic criteria for rheumatoid arthritis exist, only classification criteria as for most of the rheumatological conditions and hence the diagnosing of rheumatoid arthritis is a highly individualized process. The reason for the non-occurrence of diagnostic criteria is partly a potential consequence of misdiagnosis and partly the diversity of symptoms of the disease between individuals. The classification criteria include clinical manifestations and serological measurements. Currently the ones that are in use are those by American College of Rheumatology (ACR) and the European League Against Rheumatism (EULAR). [4, 5]

Some specific class II human leukocyte antigen regions containing a certain amino acid sequence show a very strong association with rheumatoid arthritis. Some amino acid positions that are significantly associated with the risk of developing rheumatoid arthritis are 11, 13, 71 and 74 in

HLA-DRB1 and position 9 in both of HLA-B and HLA-DPB1. Other risk regions have also been identified. They have weaker associations but are still related with immune and inflammatory pathways. When several of these risk alleles are present, even though they might only have a weak association, modest cumulative effects have been observed. [5]

There is also a specific antibody called Anti-Citrullinated Protein Antibody (ACPA), that has arisen as a suspect in the development and/or progression of rheumatoid arthritis. Up to 10 years before the disease arrives the presence of circulating ACPAs can be detected and therefore it cannot only be used as a diagnostic marker, but even as a predictive marker as well. [8]

2.4 Genome-wide association studies

An organism's complete set of DNA including all of its genes is called a genome and each genome contains all of the information needed to build and maintain that organism. To find genetic variations associated with a particular disease one approach is a genome-wide association study where markers across the genomes are scanned. When new genetic associations are identified they can be used to detect inheritable diseases in an early stage and contribute to the development of treating and preventing the occurrence of such diseases. Genome-wide association studies are a useful tool in finding genetic variations that contribute to some common and complex diseases such as autoimmune diseases, cancer and mental illnesses. [10]

In genome-wide association studies $5 \cdot 10^{-8}$ has become an established and standard threshold for the p-value, a lot smaller than the more standard value of 0.05. The reason is the large number of tests performed. By having a small type I, positive error rate for each single test, the overall type I error for all tests combined, can still be kept at a reasonably low level. [12]

2.5 Description of the data

The data used in the study is from a case-control study with 2762 cases, patients with rheumatoid arthritis, and 1940 controls. Originally the data was collected during the period May 1996 - June 2000 with the objective to identify risk factors for rheumatoid arthritis. The population investigated was from a defined area in Sweden with ages in the range of 18-70 years. [11]

The variables consist of 399 different alleles in the HLA region and because of two copies of chromosome 6 in the DNA of all humans, they are all found in pairs. This leaves us with 798 unique measurement points. For each point the three properties hydrophilicity (z_1), bulk (z_2) and electronic properties (z_3) are measured, as mentioned above. In total this becomes 2394 explanatory variables related to the amino acids and these are supplemented by the sex of the patients, thus the total number of explanatory variables is 2395.

For all of the explanatory variables the minimum value, maximum value, the median, the number of positive observations and the number of negative observations, with respect to all individuals, were observed in order to get an idea of the distribution and behavior of the variables. In Table 1 there is a summary of this information, with the median of all of the above mentioned quantities over all individuals, for the three properties z_1, z_2 and z_3, complemented with the overall minimum- and maximum values.

	Min	Max	Median	Number positive	Number negative	MIN	MAX
z_1	-1.22	2.23	0.92	4329	87	-4.92	3.64
z_2	-1.68	1.13	-0.97	1650.5	2965	-5.36	3.65
z_3	-1.29	0.3	-0.07	1170	2512	-3.44	4.13

Table 2: A summary of the ranges and distributions for the three properties of amino acids.

3 Theory

3.1 Principal Component Analysis

The purpose with a Principal Component Analysis is to examine the relationship between a set of p correlated variables in order to perform a variable reduction. This is done by a transformation of the original set of variables into a new set of variables, the so called principal components, that are uncorrelated with each other. The transformation is an (orthogonal) rotation in the p-space and creates linear combinations of the original variables. The new set of variables are ordered in decreasing order of importance, which means that the first principal component explain as much as possible of the variation in the original data. Because of this, it is easy to find a certain number of principal components that accounts for a desired level of explanation of the variation from the original data. If the original variables are highly correlated it might be that just a few principal components are needed, and the dimensionality of the problem can be reduced from p. This also implies that if the original variables are nearly uncorrelated there is no point in carrying out a principal component analysis because then the analysis will only find almost the same components, based on the variances arranged in decreasing order. This is a so called variable-directed technique which is appropriate when there is no dependent variable. [6]

What can be seen as a problem or disadvantage is that it might be hard to find an innate interpretation of the new variables and label them.

3.1.1 Derivation of the components

Suppose we have a p-dimensional random variable $\mathbf{X}^T = (X_1, \ldots, X_p)$ with mean value $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The objective is to find linear combinations

$$Y_j = a_{1j}X_1 + a_{2j}X_2 + \ldots + a_{pj}X_p = \mathbf{a}_j^T\mathbf{X}, \tag{3.1}$$

where $\mathbf{a}_j^T = (a_{1j}, \ldots, a_{pj})$ is a vector of constants. Because equation (3.1) contains an arbitrary scale factor we introduce the constraint $a_j^T a_j = \sum_{k=1}^{p} a_{kj}^2 = 1$. This constraint is required so that a unique answer may be obtained and will ensure the overall transformation to be orthonormal.

The first principal component is found by choosing $\mathbf{a}_1$ so Y_1 has the largest possible variance, that is maximizing the variance of $\mathbf{a}_1^T\mathbf{X}$ subject to the constraint $a_j^T a_j = 1$. The second principal component is then found by choosing $\mathbf{a}_2$ so that Y_2 is uncorrelated with Y_1 and has the largest possible variance of all combinations of the form (3.1). This process continues for $Y_3, \ldots, Y_p$ such that these random variables are all uncorrelated and have non-increasing variance.

The task is to choose $\mathbf{a}_1$ to maximize $\mathrm{Var}(Y_1) = \mathbf{a}_1^T \sum \mathbf{a}_1$. The method of Lagrange multipliers is the standard procedure for maximizing a function, $f(x_1, \ldots x_p)$, of several variables subject to one or more constraints. Now with only one constraint, $g(x_1, \ldots, x_p)$, this method uses the fact that for the stationary points of the function f there exists a number λ called the Lagrange multiplier such that

$$\frac{\partial f}{\partial x_i} - \lambda \frac{\partial g}{\partial x_i} = 0, \quad i = 1, \ldots, p, \tag{3.2}$$

at the stationary points. Together with the constraint these p equations can determine the coordinates of the stationary points and the corresponding values of λ as well. The next step is to examine the character of the stationary points, which can be maximum, minimum or saddle points. To simplify this, a function $L(\mathbf{x}) = f(\mathbf{x}) - \lambda(g(\mathbf{x}) - c)$ is formed and for this particular case we have $L(\mathbf{a}_1) = \mathbf{a}_1^T \Sigma \mathbf{a}_1 - \lambda(\mathbf{a}_1^T \mathbf{a}_1 - 1)$.

Now the equations from (3.2) can be written as $\partial L/\partial \mathbf{x} = \mathbf{0}$ in the general case and $\partial L/\partial \mathbf{a}_1 = 2\boldsymbol{\Sigma}\boldsymbol{a}_1 - 2\lambda\mathbf{a}_1$ for this case. And by setting this equation equal to zero we have

$$(\boldsymbol{\Sigma} - \lambda\mathbf{I})\mathbf{a}_1 = \mathbf{0}, \tag{3.3}$$

where $\mathbf{I}$ is the $p \times p$ identity matrix. Assuming that (3.3) has a solution except for the null vector then $(\mathbf{\Sigma} - \lambda \mathbf{I})$ must be a singular matrix and λ must be chosen so that $|\mathbf{\Sigma} - \lambda \mathbf{I}| = 0$, which means λ is an eigenvalue of Σ. In general Σ will have p nonnegative eigenvalues, since Σ is positive semidefinite, and if they are denoted $\lambda_1, \ldots, \lambda_p$ they can be ordered as $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$. Because the aim is to maximize the variance and $\mathrm{Var}(Y_1) = \mathbf{a}_1^T \mathbf{\Sigma} \mathbf{a}_1 = \mathbf{a}_1^T \lambda \mathbf{I} \mathbf{a}_1 = \lambda$, we choose the largest eigenvalue λ_1 as λ and the corresponding eigenvector is $\mathbf{a}_1$.

For the second principal component there are now two constraints. First $\mathbf{a}_2^T \mathbf{a}_2 = 1$ and the second is that Y_2 should be uncorrelated with Y_1. That means $\mathrm{Cov}(Y_2, Y_1) = \mathbf{a}_2^T \mathbf{\Sigma} \mathbf{a}_1 = 0$. This can be rewritten as $\mathbf{a}_2^T \lambda_1 \mathbf{a}_1 = 0$ so an equivalent condition is $\mathbf{a}_2^T \mathbf{a}_1 = 0$.

Subject to the two constraints we need to Lagrange multipliers in order to maximize the variance of Y_2 that is $\mathbf{a}_2^T \mathbf{\Sigma} \mathbf{a}_2$. Calling the Lagrange multipliers λ and δ the function to be considered is

$$L(\mathbf{a}_2) = \mathbf{a}_2^T \mathbf{\Sigma} \mathbf{a}_2 - \lambda(\mathbf{a}_2^T \mathbf{a}_2 - 1) - \delta \mathbf{a}_2^T \mathbf{a}_1.$$

At the stationary point(s) we have

$$\frac{\partial L}{\partial \mathbf{a}_2} = 2(\mathbf{\Sigma} - \lambda \mathbf{I})\mathbf{a}_2 - \delta \mathbf{a}_1 = \mathbf{0}. \tag{3.4}$$

Multiplying this equation by $\mathbf{a}_1^T$ gives together with the two constraints $-\delta = 0$ and therefore δ has to be zero at the stationary point(s). Equation (3.4) then becomes $(\mathbf{\Sigma} - \lambda \mathbf{I})\mathbf{a}_2 = \mathbf{0}$ and now we chose the second largest eigenvalue of $\mathbf{\Sigma}$ with $\mathbf{a}_2$ as the corresponding eigenvector.

Denoting the $(p \times p)$ matrix of eigenvectors by $\mathbf{A}$ then $\boldsymbol{Y} = \mathbf{A}^T \mathbf{X}$ and the covariance matrix of $\mathbf{Y}$ denoted by Λ is given by

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_p \end{bmatrix}.$$

If some of the eigenvalues of $\mathbf{\Sigma}$ are equal there is no unique way of choosing the corresponding eigenvectors, but that is not a problem as long as the eigenvectors associated with multiple roots are chosen to be orthogonal. [6]

3.1.2 Component loadings

When tabulating the principal components it is common to present the scaled vectors $\mathbf{a}_j^* = \sqrt{\lambda_j}\mathbf{a}_j$, for $j = 1, 2, \ldots, p$ rather than just the eigenvectors $\{\mathbf{a}_j\}$. The sum of squares of these scaled vectors equals the corresponding eigenvalue λ_j because $\mathbf{a}_j^{*T} \mathbf{a}_j^* = \lambda_j \mathbf{a}_j^T \mathbf{a}_j = \lambda_j$. Setting $\mathbf{C} = [\mathbf{a}_1^*, \mathbf{a}_2^*, \ldots, \mathbf{a}_p^*]$ then $\mathbf{C} = \mathbf{A}\Lambda^{1/2}$. In $\mathbf{C}$ the elements are such that the coefficients of the more important components are scaled to be generally larger than those of the less important components.

To get an interpretation of the scaled vectors $\{\mathbf{a}_j\}$ we scale the components by $\mathbf{Y}^* = \Lambda^{-1/2}\mathbf{Y}$ so that they all have unit variance. Now the inverse transformation becomes $\mathbf{X} = \mathbf{A}\mathbf{Y} = \mathbf{A}\Lambda^{1/2}\mathbf{Y}^* = \mathbf{C}\mathbf{Y}^*$. The elements of $\mathbf{C}$ work as weights to scale more important components to generally be larger than those of the less important components as already mentioned and they are usually called the component loadings.

3.2 Logistic regression

The logistic regression model is a type of generalized linear model. It is the most important model for categorical response data and has for a long time been used in biomedical studies. The model is used to estimate the probability of a binary outcome Y, for example pass/fail or healthy/sick.

In case of more potential outcome categories there is also a multinomial logistic regression model, but it will not be discussed here. The predictors however can be either continuous or categorical.

If $P(Y = 1|\mathbf{X} = \mathbf{x}) = p(\mathbf{x})$ where $\mathbf{X}$ is a vector of n independent variables and $\mathbf{x} = (x_1, \ldots, x_n)$ the observed values of $\mathbf{X}$, the model can be expressed either as the logodds

$$\log\left(\frac{p(\mathbf{x})}{1 - p(\mathbf{x})}\right) = \alpha + \sum_i^n \beta_i x_i, \tag{3.5}$$

or it can be written as the probability

$$p(\mathbf{x}) = \frac{e^{\alpha + \sum_i \beta_i x_i}}{1 + e^{\alpha + \sum_i \beta_i x_i}}.$$

The sign of each effect parameter β_i determines whether $p(\mathbf{x})$ is increasing or decreasing as x_i is increasing and if $\beta_i = 0$, Y is independent of X_i. Another interpretation is by exponentiating both sides of (3.5), then the odds increases multiplicatively by the odds ratio e^{β_i} for every 1-unit increase in x_i. [7]

3.2.1 Hypothesis testing for the effect parameters

To test if X_i has any effect on Y a hypothesis test can be performed with the hypothesis

$$H_0 : \beta_i = 0,$$
$$H_a : \beta_i \neq 0.$$

A Wald test statistic is $z^2 = \hat{\beta}_i^{\,2} / \widehat{\mathrm{Var}}(\hat{\beta}_i) \sim \chi_1^2$ under H_0, and the p-value is $\mathrm{P}(\chi_1^2 \geq z^2) = p$, where a low p-value indicates that H_0 can be rejected and hence that X_i has an effect on Y.

3.2.2 Logistic regression with dummy variables

Just like ordinary regression, logistic regression extends to include qualitative explanatory variables and one way of handling them is to use dummy variables. For one single explanatory variable with I categories the model is

$$\log\left(\frac{p(\mathbf{x})}{1 - p(\mathbf{x})}\right) = \alpha + \beta_1 x_1 + \ldots + \beta_{I-1} x_{I-1}$$

where $x_i = 1$ for observations in row i ($x_i = 0, i = 1, \ldots, I - 1$) and otherwise.

By not creating a dummy variable for category I this avoids a parameter redundancy and the choice of category to exclude from having a dummy variable is arbitrary. The values for the β_i for a single category is then irrelevant, it is only meaningful to compare them to each other and accordingly comparing the effect of the categories.

For more than one explanatory variable, say n variables, we now write the model formula as

$$\log\left(\frac{P(Y = 1)}{1 - P(Y = 1)}\right) = \alpha + \beta_{j_1}^{X_1} + \beta_{j_2}^{X_2} + \ldots + \beta_{j_n}^{X_n}.$$

The model have parameters $\{\beta_{j_k}^{X_i}\}$ which represent the effects of X_i. Important to stress is that the X_i superscripts do not represent powers, they are only labels. The covariates X_i can have any number of categories that possibly varies between them; $j_1 = 1, 2, \ldots, n_1$ up to $j_n = 1, 2, \ldots, n_n$. A parameter $\beta_{j_k}^{X_i}$ quantifies the effect on the log odds of the probability in category j_k of X_i. One parameter for each factor is redundant and therefore needs to be fixed at 0, the category without a dummy variable.

3.2.3 Logistic regression for retrospective studies

In retrospective sampling design such as case-control studies it is the explanatory variable X rather than the response variable Y that is random and even for these cases logistic regression can be extended. The sample sizes of each group in retrospective studies are fixed before the analysis and cases are typically oversampled compared to the whole population. The advantage is that the odds ratios still can be estimated and the effect parameters will be the same as for a prospective study, but it is important to know and remember that the intercept needs to be adjusted for the oversampling of cases.

3.3 Test of independency

In order to test for independency between the variables for a multinomial sampling in a I x J contingency table, suppose the I rows are independent with each of the J columns. Let μ_{ij} be the expected number of observations in cell (i, j), n_{ij} is the observed number of observations in each each cell, $\hat{\mu}_{ij} = n_{i+}n_{+j}/n$ is the estimated expected number of observations in each cell and $\pi_{ij} = P(X = i, Y = j)$ is the probability for an observation to be in cell (i, j).

To test if X (row affiliation) has any effect on Y (column affiliation) the probability that a randomly chosen subject has outcome $Y = j$ is $\pi_{+j} = \sum_{I}^{i=1} \pi_{i+}\pi_{j|i}$. The null and alternative hypotheses can be expressed as:

$H_0 : \pi_{j|i} = \pi_{+j}$ for all i, j and

$H_a : \pi_{j|i} \neq \pi_{+j}$, for at least some i, j.

They can also be written with words as:

H_0 : There is no association between rows and columns; they are independent.

H_a :The row and column variables are not independent.

The chisquare test statistic is

$$X^2 = \sum_{i,j} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} \sim \chi^2_{df}$$

under the null hypothesis, where $df = (I - 1)(J - 1)$ and the p-value is $P(\chi^2_{df} \geq X^2) = p$. A small p-value indicates strong evidence of an association between the variables and then H_0 can be rejected. If H_0 is not rejected the variables are independent and that corresponds to homogeneity of each outcome probability among the rows. [7]

4 Method

4.1 Reducing data

Initially the data is reduced in two different ways. The first reduction is straightforward, removing all observations with missing values followed by removing the variables without any variation at all. From now on this will be called subset 1.

The second reduction starts with removing all the variables where more than 5% of the observations have missing values and then do the same as for subset 1, eliminate the observations with missing values to complete cases together with removing the variables without variation. This one is called subset 2.

4.2 Performing the principal component analysis

The first approach applied is the principal component analysis. Because this is a variable-directed technique, without a dependent variable, the status of the patients is ignored but all the explanatory variables are used. The analysis is performed twice, both on subsets 1 and 2.

4.3 Performing the logistic regression

4.3.1 Coding dummy variables

Because of the composition of the data with three properties in pairs of two for each amino acid, some variables are now further removed in order to get complete sets, with both variables in a pair for all three properties for all amino acids.

A limit between positive and negative values is determined, and for all three properties the two variables are divided into three classes: both values over the limit, both values under the limit or one value of both. We refer to these categories as "both positive", "one of each" and "both negative". Setting the category "both negative" as baseline the other two categories then generates dummy variables that are to be used in the logistic regression model.

4.3.2 The models

The next approach is the logistic regression analysis, and it is going to be done separately for all amino acids. The explanatory variables are z_1, z_2 and z_3 dummy-coded as explained above, and the sex of the patient. The dependent variable is the status of the patient, concretely rheumatoid arthritis or not.

For each amino acid four different models are to be examined, three univariable models and one multivariable model. With $y = 1$ when the patient has rheumatoid arthritis and $y = 0$ otherwise, $P(Y = 1)$ is the probability of having the disease. The models are then written as

$$P(Y=1) = \frac{e^{\beta_j^{z1}+\text{sex}}}{1+e^{\beta_j^{z1}+\text{sex}}},$$

$$P(Y=1) = \frac{e^{\beta_j^{z2}+\text{sex}}}{1+e^{\beta_j^{z2}+\text{sex}}},$$

$$P(Y=1) = \frac{e^{\beta_j^{z3}+\text{sex}}}{1+e^{\beta_j^{z3}+\text{sex}}},$$

$$P(Y=1) = \frac{e^{\beta_{j1}^{z1}+\beta_{j2}^{z2}+\beta_{j3}^{z3}+\text{sex}}}{1+e^{\beta_{j1}^{z1}+\beta_{j2}^{z2}+\beta_{j3}^{z3}+\text{sex}}}$$

where $j, j_1, j_2, j_3 \in \{(++), (+-)\}$ are the categories for the dummy variables when $(--)$ is used as baseline.

5 Analysis

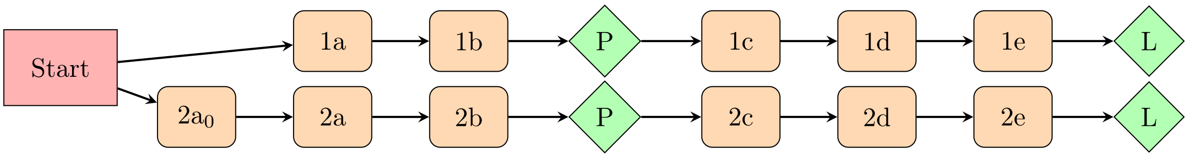
5.1 Handling the data

As described in section 4 some reduction of data is needed because of missing values and lack of variation in the data. The exact steps are explained in Table 3 along with the dimensions of the subsets with the variable sex excluded.

		Dimensions	
		Subset 1	Subset 2
	Original data set	4702 x 2394	4702 x 2394
Step a_0).	Removing all variables with more than 5% missing values in the observations.	Not done for this subset	4702 x 2256
Step a).	Removing all observations with missing values.	1877 x 2395	4633 x 2256
Step b).	Removing all variables without any variation.	1877 x 1974	4633 x 1989
Step c).	Removing all variables whose "twin" already been removed.	1877 x 1824	4633 x 1896
Step d).	Combining and categorizing the pairs of each property for all variables.	1877 x 912	4633 x 948
Step e).	Removing all variables with no variation after the categorization in at least one of the properties $z_1 - z_3$.	1877 x 72	4633 x 75

Table 3: All steps of data reduction. The full data set consists of 4702 individuals, and 2394 genetic explanatory variables related to amino acids.

In the flow chart below, the different steps of the analysis are illustrated. Here P stands for the principal component analysis and L symbolizes the logistic regression.



5.2 Biplots of the principal components

All principal components were plotted against each other, and for future references a representative sample is shown in Figure 1 below.

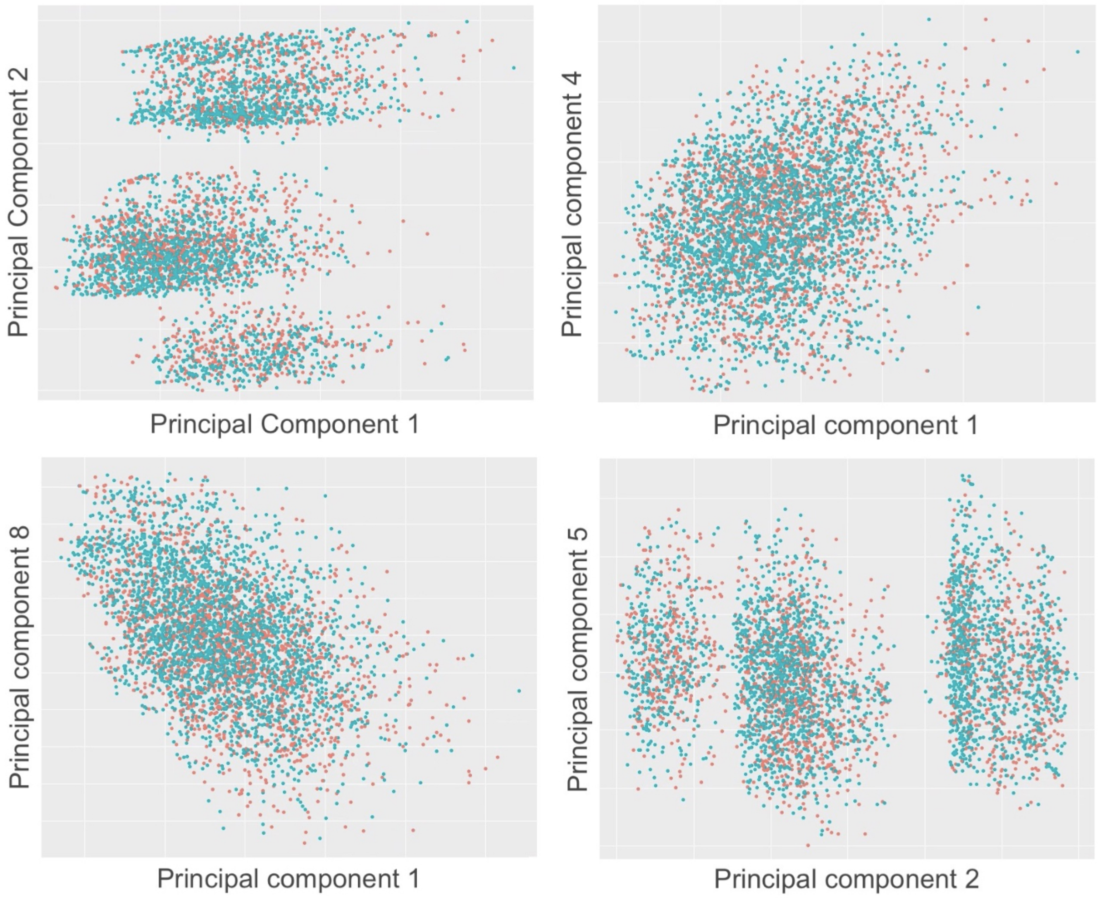


Figure 1: A sample of the biplots for the principal components. The green dots correspond to cases and the red dots correspond to controls.

5.3 Results

Seven variables where standing out with highly significant results for at least one of the properties. These were amino acid positions 11, 13, 37 and 74 in HLA-DRB1 and amino acid positions 5, 55 and 57 in HLA-DQB1. In the tables below the properties $z_1 - z_3$ with a genome-wide significance in the univariable models are presented and if they still had a genome-wide significance in the multivariable models that p-value is shown as well. Many of the amino acids occur twice because both dummy variables were significant. What is not presented is what category the variables represent, since, as mentioned above, no biological interpretations are made in this report.

Table 4: Significant variables for z_1.

Position	P-value	P-value
	Univariable	Multivariable
HLA-DRB1-37	5.531e-19	2.434e-15
HLA-DQB1-55	7.858e-18	7.799e-18
HLA-DRB1-11	1.913e-09	
HLA-DRB1-37	1.869e-08	2.608e-09

(a) Subset 1.

Position	P-value	P-value
	Univariable	Multivariable
HLA-DRB1-11	1.505e-34	
HLA-DRB1-11	2.031e-18	
HLA-DRB1-37	1.336e-17	5.510e-22
HLA-DQB1-55	4.458e-9	1.251e-17
HLA-DRB1-37	1.066e-08	7.834e-13

(b) Subset 2.

Table 5: Significant variables for z_2.

Position	P-value	P-value
	Univariable	Multivariable
HLA-DQB1-5	3.949e-18	2.044e-19
HLA-DRB1-11	3.024e-10	
HLA-DQB1-57	5.154e-9	1.362e-9
HLA-DRB1-37	2.916e-8	
HLA-DQB1-5	3.234e-8	5.700e-10

(a) Subset 1.

Position	P-value	P-value
	Univariable	Multivariable
HLA-DRB1-13	1.716e-18	5.843e-39
HLA-DRB1-74	3.043e-9	1.322e-9

(b) Subset 2.

Table 6: Significant variables for z_3.

Position	P-value	P-value
	Univariable	Multivariable
HLA-DRB1-11	1.539e-25	
HLA-DQB1-55	9.205e-18	
HLA-DQB1-5	1.035e-17	
HLA-DRB1-11	3.651e-12	
HLA-DRB1-13	2.516e-9	1.630e-14
HLA-DQB1-55	3.855e-8	

(a) Subset 1.

Position	P-value	P-value
	Univariable	Multivariable
HLA-DRB1-11	4.244e-53	
HLA-DQB1-55	2.716e-24	
HLA-DRB1-11	3.024e-20	
HLA-DRB1-74	3.585e-11	
HLA-DQB1-55	3.483e-9	

(b) Subset 2.

5.3.1 Test of independence between the amino acid properties

A χ^2 test was performed to test if there are any significant difference in occurrence, if there are any indications that one property seems to be more important and have more impact than the others. The p-values for subset 1 and subset 2 were 0.77 and 0.42 respectively. Therefore the null hypothesis that there is no association between the properties for the amino acids and a significant outcome in the logistic regression cannot be rejected, no property seems to have more impact than the others.

6 Discussion

In this project, we started with an exploratory data analysis. For example, the ranges and distributions of the variables, their variance and missing values were found. This was a necessary preparation of the subsequent steps. As a consequence of the lack of variation in some of the variables or principally the occurrence of missing values, a lot of data reduction has been made along the way throughout the analysis. As a first result of this, two different subsets have been studied parallel to each other.

The first actual analysis step was a principal component analysis. The principal components were not used further, and this step was mainly included to see if it was possible to detect a mismatch between cases and controls. However, when examining the graphs and result, some of the derived principal components were remarkable and accordingly, some of the variables required a more careful inspection. Yet, no results from this analysis were further used.

After this the pairs of amino acids and their properties were combined and categorized in order to use dummy variables in a logistic regression. Some further reduction of the variables was performed due to lack of variation in some of the variables after the categorization. We analyzed many of the variables separately, and we also compared the results from the two subsets and likewise the properties for the same amino acid.

6.1 Results

6.1.1 Interpretation of the PCA

When examining the biplots of the principal components from the principal components analysis, from which a sample can be seen in Figure 1, two things are noticed. The first thing is that for some of the graphs the observations have clustered into normally three groups and the second thing is that cases and controls overlaps, there is no evident difference between the groups observable.

The overlapping of cases and controls is for the rest of the analysis a good aspect and can be seen as a data check for testing if there are any hidden population structures. A hidden structure can be revealed by a separation of the groups in a biplot of the principal components. Hidden population structures involve confounders, other variables that may cause a difference in the results such as age or ethnicity. For example, if all of the cases have one ethnicity and the control group have another, then the variation along the genome may not be due to the disease but caused by ethnicity. The aim is to match confounders as well as possible in order to avoid that they will have an impact on the results. Fortunately, in this case no hidden population structures where found, and therefore the logistic regression could be performed with more confidence in having reliable results.

Biplots actually contain a lot of information and can be helpful in interpreting relationships between experimental groups and compounds or to identify outliers. Sometimes the observations from a certain group are closer together in clusters than other groups and there can be patterns of the observations in the biplots. The biplots are able to highlight groups of homogeneous observations. Because most of the variables only admit a few values, between 2-5 categories, it is not surprising that they more or less coincide and create the observed clusters. The principal component analysis captured some differences between certain groups. These groups are not cases and controls, but instead observations with different sets of amino acids.

6.1.2 Expectations from previous studies

Of the amino acid positions mentioned in subsection 2.3, not all of them had significant results. For example, HLA-DRB1-74 remained in the analysis for subset 2 and had some significant variables there, but it was removed for subset 1. The elimination occurred in step e, where all amino acids without any variation in at least one of the three properties $z_1 - z_3$ were removed. In this particular case the reason was no variation in z_1. Comments about the effects of this method of eliminating variables will be done in subsection 6.2.1 . Even for HLA-DRB-71, a position that was known as a risk factor from previous studies, elimination occurred for both subsets in the last step as a result of no variation in z_1. For HLA-B-9, which was also known from before but removed in these analyzes, the reduction took place in the same step. For this position the lack of variation occurred in both z_2 and z_3.

6.1.3 Odds ratio interpretation

In the result section the odds ratios (OR) are not presented. The reason is that at first when they were inspected, for some of the amino acids one property could have an impact on rheumatoid arthritis for positive values of the property, but for another amino acid the same property could have an impact on the disease for negative values of the property. This may seem contradictory, but it implies that for each amino acid the properties need to be interpreted separately. For some amino acids a negative value is critical but for others a positive value is associated with risk of the disease. However, after some consultation it was found that these observations would require too advanced biology and chemistry in order to be interpreted, even for a biologist these finding would be hard to explain and understand.

Therefore it was only established *if* the amino acid properties have a significant impact on rheumatoid arthritis and not *how*.

6.1.4 Difference between amino acid properties

Not all properties were significant for the same amino acid, for the majority of the amino acids only one or two properties got significant result. And there were not even any clear recurrent patterns or analogues between the properties, for example, when z_1 had significant impact even z_2 had it or when z_2 had significant results, z_3 had not. As pointed out above, no biological interpretations of these findings will be made.

6.2 Methods

6.2.1 Reducing the data set

Depending on where the missing values were found in the data set, the reduction was done in two different ways. At the end, about twice as many observations were left when using the second approach, about the same number of variables and almost the same variables as well. The first method started with removing all observations with at least one missing value and the second approach started with removing all variables with more than 5% missing values and then removing the observations with at least one missing value. The 5% limit was used because it is a default threshold in the field of genetics, but additionally it was a natural limit for this data set since in practice all variables removed had more than 7% missing values and the variables kept had less than 0.5% missing values.

In the principal component analysis some of the variables with the highest load on the ten first principal components differed but in general the principal component analysis came out with the same results with both subsets. It was primarily in the logistic regression the results came out with a more noticeable difference and this will be discussed more extensively in subsection 6.2.2.

Except for the initial steps of data reduction the subsets were treated equally, as seen in subsection 5.1. As a consequence of how it was decided to categorize the properties of the amino acids prior to the logistic regression, steps *c* and *d* were essential. The other steps where all variables without variation are removed, namely step *b* and *e*, were not as essential. But a variable without any variation do not contribute to the result and therefore these steps are not questionable. However, if step *b* would not have been performed then step *c* would be unnecessary. Even if one of the properties in a pair did not have any variation, the combined variable would still have variation if the other property in that pair had variation and therefore, it could have been used in the regression. Regarding step *e* it is definitely a questionable step to remove an amino acid with all three properties even if only one of the properties lacked variation. Together with the differences between the results from subset 1 and subset 2, this will be discussed in subsection 6.2.2.

6.2.2 Risk of losing information

When reducing the data some information loss is inescapable. For example, a variable that indeed has an important role can be eliminated as a result of a lot of missing values, caused by previous steps in the process. Another example is when observations are removed because of a lack of variation. For all other variables the results would probably turn out with stronger associations or at least with more solid results with more observations in the analysis. Because there are many more observations in subset 2 and about the same number of variables left in each step, the results using that subset can in one way be considered more reliable. But the results from subset 1 can still give a good indication as to which amino acids are associated with the disease.

One example is HLA-DQB1-5 that was eliminated already in the beginning of step a_0 for subset 2, because of around 8% missing values. However, in subset 1 it was still included and got significant results for both of z_2 and z_3. Even if these results might need to be reconsidered a bit more carefully because of all the missing values that possibly could change the result if they would not have been missing, this information would have been lost if the data were reduced only with the second approach. Even if HLA-DQB1-5, after more investigation, does not turns out to be a risk amino acid position, it might still be good to investigate this more closely.

Another example that is a possible case of information loss, is HLA-DRB1-74. This position was removed for subset 1 in step e because of no variation in z_1. In subset 2 however it remained and got significance for some variables. This shows the strength in having more observations, as in subset 2. Then it is more likely that the variation in the variable is more similar to the true proportions in the whole population, and the variable does not need to be eliminated, even though it contains useful information.

For the previously known amino acid positions HLA-DRB1-71 and HLA-B-9, they were both also removed prior to the logistic regression in step e on the grounds of lack in variation. As seen in section 5.1 this was a huge reduction, where a lot of amino acids were removed. For HLA-DRB1-71 it was in z_1 the lacking variation was found, and when it comes to HLA-B-9, both z_2 and z_3 lacked variation in data. This reduction could have been left out and the regressions could have been performed anyway. What would have happened then is that the variables without variation would not have had any influence on the dependent variable at all. The coefficients of these variables would not be estimable and as outcome instead we would get Not Available (NA), that can be ignored. By still keeping all of the amino acids, some information that now may have been lost, would still be kept in the analysis. Since not all properties are significant for the same amino acid and sometimes only one of them had significant result, it would not automatically be a disadvantage not to have them all, for some amino acids. In addition, even if there is no variation in one of the variables the others could still have been investigated. For example, we could have analyzed univariable models with z_2 and z_3 as explanatory variables, for HLA-DRB1-71. Indeed, when looking at the results there are sometimes strong associations for some of the properties even if there is almost no variation for some other property in the same position. This could have provided more information, since it might have changed the proportion of significant variables for each property, leading to a conclusion that one of them is more crucial than the others. But it could also have given nothing except for a lot of more work with a substantially larger number of variables. Now the aim was to do an initial analysis, but a future goal would be to conduct a more extensive analysis for all amino acids, or at least those previously known.

6.2.3 Treating the variables before the principal component analysis

Before the principal component analysis, there are some possible actions, with the aim to make the involved variables equally important. These are standardization and centering, and neither of them were done. Below we will explain why.

6.2.3.1 Standardization of variables The objective of standardizing the variables is that they should contribute equally to the analysis so that no variable with a wider range should outweigh a variable with a more narrow range. Without the standardization the variable with the wider range will have a larger effect in the analysis, and a transformation to equal scales can prevent this problem. Especially in principal component analysis the consequences of standardizing the variables or not can be essential, because the aim is to capture the total variance in the set of variables and the analysis will give more load to those variables that have higher variances.

If the variables measure different aspects and therefore do not have the same units of measurement, for example meters and kilograms, a standardization transforms these to comparable scales, hence it can be a useful tool to make it possible to compare variables representing various properties.

There may be reasons for not scaling as well, for instance, if the specific scale of the variables matters.

In this case the explanatory variables are themselves derived from a principal component analysis and the properties they represent, $z_1 - z_3$, do not correspond directly to the bulk, the electronic properties and the hydrophilicity, and therefore they do not have a specific unit and are already unitless. They are consequently measured on the same scale, but they still have some diversity in the variance.

No standardization was done with the justification that the variables already have similar variance. At a first look, the variances do not seem to be that similar with ranges from $8.5 \cdot 10^{-7}$ to 18.4 in subset 1 and from $3.5 \cdot 10^{-7}$ to 19.5 in subset 2. But when looking more closely we have for example the 50%-quantiles 0.14 respectively 0.18 and the 75%-quantiles 0.84 and 0.97. It is only about 10% of the variables that have a considerably larger variance.

When examining the ten variables with the highest load on the first 10 principal components, it is not surprising that almost all of them have high variance, most of them are found over the 95th percentile. Only one principal component stands out and attracts attention - the first one. The first principal component has the maximal variance value of one of the variables way below the minimum for any of the others. Despite the low variances in the involved variables, the first principal component by itself explains in subset 1 80% and in subset 2 79% of the variance in the whole data set.

Even when looking at more variables, the first 20 or the first 50 variables with the highest load on the principal component we observe a similar pattern. The variances of the variables with the highest load on the first principal component are still surprisingly low. In order to explain this, the ranges of the variables with the highest load were examined and it was concluded that the variables with the highest load on the first principal component have the same ranges, but on the other hand many observations in one of the end points. Therefore, they could still explain the variation in the data without having a large variance themselves.

6.2.3.2 Centering of variables Centering is done by subtracting the mean from the observed value in order to centralize all observations around zero. By doing this it is easy to examine if an observation is above or below the mean. Between different variables it erases the variations between the samples and leaves us only with different ranges of variations within the samples.

No centering was done in this case since we reasoned that the original variation in data could be of importance. Even if some scales are shifted, that variation was to be kept in the analysis.

6.2.4 Dummy coding of the variables in the logistic regression

Prior to the logistic regression a dummy coding of the variables was performed. In this case the coding was done with three categories (giving two dummies) depending on whether the properties of the amino acids had positive or negative values, or one of both. Because the majority of all amino acids only admitted positive or negative values a lot of them fell into the same category and after the dummy-coding there was no variation left for a considerable number of the explanatory variables. As a result, they were removed from further analysis. In exact numbers, 304 amino acids were available for the analysis before this step, but only 75 remained afterwards.

An alternative approach of coding the variables would be based on their values being below or above the median. For all of the 304 amino acids, three different properties had at least two different values, so all observations can obviously not be on one side of the median. Thus, by using this method, no amino acids would have been removed. For this approach some different options are available. The coding can still be based on both the amino acids jointly, like the previous method, and the coding would correspond to both variables being below, both above or one of each, creating three possible categories and two dummies. Alterntively, one can take the sum of the variables and compare the sums with the median sum, thereby generating two categories and just one dummy.

A third method could be to only choose for example the highest value among the two possible in each pair, or their sum, and then deal with the variables as continuous. For all the three methods mentioned (categories defined as positive or negative values, categories defined as above or below the average and a continuous variable) it would have been possible not to combine the pairs of amino acids but to treat them separately. The latter approach would have doubled the number of explanatory variables.

These are some possibilities that can be considered regarding the coding of the variables for the logistic regression. However, the important thing is to focus on what is relevant for the analysis. Is the intention to create a predictive model or is the aim to investigate some specific properties of the explanatory variables? What aspects of the explanatory variables are of interest? Here the aim was to investigate the properties of the amino acids and consequently we wanted to keep the distinction between positive and negative values. Since the amino acids always occur in sets of two it is of no use to separate them.

The problem has already been mentioned that the data set was reduced by three quarters of the variables, because of the chosen coding approach. Some alternative coding schemes would possibly reduce data less, depending on how the range of the variables is divided into categories.

6.3 Extensions and future work

A collection of suggestions of future work has already been mentioned. Starting with the categorization, some ideas of how that could have been done were discussed in subsection 6.2.4. Because of the extensive reduction of variables due to the lack of variation after the categorization, other categorization methods could have been investigated as well. On the other hand, another categorization method could possibly disguise some of the associations this method could identify. This argument can also be used in reverse, therefore, a comparison between different categorization approaches is one way to go.

As mentioned in subsection 6.2.2, step *e* in the reduction of variables could have been done differently. In addition, some of the variables known from before could be worth a closer inspection, even if they do not have variation in one of the three properties.

The most interesting associations are between those with the antibody ACPA and rheumatoid arthritis. Because of this a reasonable task would be to separate the cases into two groups, ACPA positives and ACPA negatives. When separated the whole analysis could be performed once again, not only to compare cases and controls, but also to compare the two subgroups of cases and controls.

When analyzing the results, all variables with a genome-wide significance in the univariable models were considered, and then we investigated whether they were still significant in the multivariable model. For a few of the amino acids all variables were significant in the multivariable model, but not in the univariable models, and these amino acids could be of interest to analyze further.

References

[1] Betts, M.J; B. Russell, R.B. *Bioinformatics for Geneticists, Second Edition.* John Wiley & Sons, New York, 2007. 311-339

[2] Hellberg, S; Sjöström, M; Skagerberg, B; Wold, S. Peptide Quantitative Structure-Activity Relationships, a Multivariate Approach. *Journal of Medicinal Chemistry*, 30 (1987), 1126-1135.

[3] Shankarkumar, U. The Human Leukocyte Antigen (HLA) System. *International Journal of Human Genetics*, 4 (2004), 91-103.

[4] Jiang, X. *Gene-environment interactions in rheumatoid arthritis: quantification and characterization of contributing factors.* PhD thesis, Karolinska Institutet, Stockholm, 2015.

[5] Smolen, J.S; Aletaha, D. Rheumatoid arthritis. *Nature reviews*, 4 (2018), 1-19.

[6] Chatfield, C; Collins, A.J, *Introduction to Multivariate Analysis.* Springer, New York, 1980.

[7] Agresti, A. *Categorical data analysis, Second Edition.* John Wiley and Sons, Hoboken New Jersey, 2002.

[8] Willemze, A; Trouw,L.A; Toes, R.E.M; Huizinga, T.W.J. The influence of ACPA status and characteristics on the course of RA. *Nature reviews*, 8 (2012), 114-150.

[9] Wikberg, J;Eklund, M; Willighagen, E; Spjuth, O; Lapins, M; Engkvist, O; Alvarsson, J. *Introduction to Pharmaceutical Bioinformatics.* Oakleaf Academic, Stockholm, 2010.

[10] National Human Genome Research Institute. *Genome-Wide Association Studies*, `https://www.genome.gov/20019523/genomewide-association-studies-fact-sheet/#gwas-1`(2015), Acessed 2018-04-20.

[11] P Stolt, P; Bengtsson, C; Nordmark, B; Lindblad, S; Lundberg, I; Klareskog, L; Alfredsson, L. Quantification of the influence of cigarette smoking on rheumatoid arthritis: results from a population based case-control study, using incident cases, *Annals of the Rheumatic Diseases*, 62 (2003), 835-840.

[12] Fadista, J; Manning, A.K; Florez, J.C. The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants, *European Journal of Human genetics*, 24 (2016), 1202-1205.

7 Appendix

In Tables 7 to 30 below, the estimates for the coefficients, 95% confidence intervals and p-values for all models and amino acids using subset 1 are presented. Each table includes all four (three univariable and the multivariable) models connected to an amino acid, only separated by an empty row. In Tables 31 to 55 the models using subset 2 are shown.

For some amino acids all four columns show “NA” for one or two of the properties in the multivariable model. This is because they are linearly dependent of the other properties and can therefore not be estimated.

Subset 1

	Estimate	2.5%	97.5%	p-value
HLA-A-9 z_1+-	1.0802284462	0.8581797167	1.3597309204	0.5109844439
HLA-A-9 z_1++	0.9272055539	0.4771153279	1.8018916785	0.8235717359
HLA-A-9 z_2+-	1.1650366434	0.5869221229	2.3125902527	0.6623489125
HLA-A-9 z_2++	1.0785095017	0.5549723170	2.0959293096	0.8235717359
HLA-A-9 z_3+-	1.3307846275	0.2161532859	8.1932028792	0.7579587298
HLA-A-9 z_3++	1.3259965168	0.2202848693	7.9817863467	0.7580092419
HLA-A-9 z_1+-	1.1121325276	0.8419696734	1.4689825515	0.4541517311
HLA-A-9 z_1++	0.9969287133	0.4705591268	2.1120977213	0.9935927473
HLA-A-9 z_3+-	1.2071727284	0.1748277719	8.3354376738	0.8485389048
HLA-A-9 z_3++	1.3006053645	0.1869189840	9.0497726772	0.7905882070
HLA-A-9 z_2+-	NA	NA	NA	NA
HLA-A-9 z_2++	NA	NA	NA	NA

Table 7: Estimate, 95% confidence interval and p-value for all three univariable models and the multivariable model for the amino acid on position HLA-A-9, using subset 1.

	Estimate	2.5%	97.5%	p-value
HLA-A-62 z_1+-	9.5196758e-06	3.598465e-282	2.518413e+271	0.9715981971
HLA-A-62 z_1++	7.9204212e-06	2.995078e-282	2.094538e+271	0.9711465945
HLA-A-62 z_2+-	0.7739362715	0.5858541693	1.0224000847	0.0712288196
HLA-A-62 z_2++	0.6247452853	0.4668936653	0.8359648042	0.0015470520
HLA-A-62 z_3+-	1.2370876743	1.0025238541	1.5265331669	0.0473157280
HLA-A-62 z_3++	1.5713330881	1.1700164698	2.1103016389	0.0026688256
HLA-A-62 z_1+-	1.1281482e-05	4.264199e-282	2.984660e+271	0.9720151364
HLA-A-62 z_1++	1.0631429e-05	4.020001e-282	2.811623e+271	0.9718693940
HLA-A-62 z_2+-	0.7212912301	0.3441435028	1.5117560968	0.3868474823
HLA-A-62 z_2++	0.6313717367	0.4694883777	0.8490737765	0.0023471270
HLA-A-62 z_3+-	1.0847339852	0.5175483695	2.2735030925	0.8294343701
HLA-A-62 z_3++	NA	NA	NA	NA

Table 8: Estimate, 95% confidence interval and p-value for all three univariable models and the multivariable model for the amino acid on position HLA-A-62, using subset 1.

	Estimate	2.5%	97.5%	p-value
HLA-A-76 z_1+-	0.9162226783	0.7498452247	1.1195163595	0.3921285545
HLA-A-76 z_1++	1.4288651438	0.9113245039	2.2403168032	0.1198760207
HLA-A-76 z_2+-	0.9572773956	0.7646853052	1.1983753394	0.7032280172
HLA-A-76 z_2++	1.6169356713	0.7232531698	3.6148904345	0.2417367940
HLA-A-76 z_3+-	0.9520060813	0.7479211920	1.2117795143	0.6894952054
HLA-A-76 z_3++	2.5849202871	0.7478682915	8.9344781244	0.1333988354
HLA-A-76 z_1+-	1.3656122944	0.6712175453	2.7783793071	0.3898645954
HLA-A-76 z_1++	2.5529817632	0.7372247275	8.8408807242	0.1391623473
HLA-A-76 z_2+-	0.6752684840	0.3357605746	1.3580734606	0.2707173428
HLA-A-76 z_2++	0.6262898486	0.1439451986	2.7249187757	0.5327919809
HLA-A-76 z_3+-	0.6656342487	0.3309375720	1.3388294063	0.2536367223
HLA-A-76 z_3++	NA	NA	NA	NA

Table 9: Estimate, 95% confidence interval and p-value for all three univariable models and the multivariable model for the amino acid on position HLA-A-76, using subset 1.

	Estimate	2.5%	97.5%	p-value
HLA-A-152 z_1+-	0.8067048510	0.6535484327	0.9957528532	0.0455450308
HLA-A-152 z_1++	0.7218101891	0.5396427474	0.9654719750	0.0280384902
HLA-A-152 z_2+-	0.7844086403	0.6417346816	0.9588026527	0.0177548350
HLA-A-152 z_2++	0.6603061365	0.4540191700	0.9603211112	0.0298691437
HLA-A-152 z_3+-	1.0184460455	0.8092362740	1.2817422810	0.8761935177
HLA-A-152 z_3++	1.1372792939	0.4637109314	2.7892467156	0.7786820451
HLA-A-152 z_1+-	0.8725713595	0.3670382632	2.0743907482	0.7576930954
HLA-A-152 z_1++	0.8499906035	0.1494672375	4.8337283675	0.8545850355
HLA-A-152 z_2+-	0.8709723834	0.3663604141	2.0706191593	0.7545407688
HLA-A-152 z_2++	0.7562330419	0.1275061709	4.4851822437	0.7583695156
HLA-A-152 z_3+-	1.0464913717	0.4401948685	2.4878622388	0.9180815127
HLA-A-152 z_3++	1.1408839547	0.1609230438	8.0884388412	0.8950689496

Table 10: Estimate, 95% confidence interval and p-value for all three univariable models and the multivariable model for the amino acid on position HLA-A-152, using subset 1.

	Estimate	2.5%	97.5%	p-value
HLA-A-156 z_1+-	1.0520183892	0.8554221351	1.2937971159	0.6309079212
HLA-A-156 z_1++	1.0170896551	0.6127276166	1.6883054369	0.9477475890
HLA-A-156 z_2+-	0.9150210152	0.7491307892	1.1176465717	0.3842133504
HLA-A-156 z_2++	0.9182751928	0.6401919540	1.3171507771	0.6431947499
HLA-A-156 z_3+-	0.8152571125	0.6333130994	1.0494716755	0.1129192390
HLA-A-156 z_3++	0.5957355153	0.2145570838	1.6541090041	0.3201817222
HLA-A-156 z_1+-	1.4556289377	0.8066225265	2.6268242390	0.2125875077
HLA-A-156 z_1++	1.6154700098	0.5222385303	4.9972248329	0.4051551076
HLA-A-156 z_2+-	0.6807457832	0.3714764046	1.2474946338	0.2133532127
HLA-A-156 z_2++	0.5923885005	0.2126031870	1.6506061849	0.3166085585
HLA-A-156 z_3+-	1.1346561591	0.6289438856	2.0469943803	0.6747534352
HLA-A-156 z_3++	NA	NA	NA	NA

Table 11: Estimate, 95% confidence interval and p-value for all three univariable models and the multivariable model for the amino acid on position HLA-A-156, using subset 1.

	Estimate	2.5%	97.5%	p-value
HLA-C-156 z_1+-	1.2135357844	0.9899986088	1.4875466358	0.0624347360
HLA-C-156 z_1++	1.1536246090	0.8424056402	1.5798205462	0.3729913464
HLA-C-156 z_2+-	1.0923856464	0.8732376119	1.3665311528	0.4392407564
HLA-C-156 z_2++	1.2482841574	0.9530063623	1.6350502990	0.1073038738
HLA-C-156 z_3+-	0.9707039647	0.7796733004	1.2085397649	0.7902952180
HLA-C-156 z_3++	0.9191744313	0.5112739104	1.6525029304	0.7782420122
HLA-C-156 z_1+-	0.7748109409	0.4546294372	1.3204864116	0.3482676584
HLA-C-156 z_1++	0.3912282521	0.1383947748	1.1059633238	0.0767262186
HLA-C-156 z_2+-	1.5279512749	0.8616289097	2.7095598492	0.1469424352
HLA-C-156 z_2++	3.0678689024	1.0598892588	8.8800028153	0.0387128536
HLA-C-156 z_3+-	0.6306886657	0.3804135842	1.0456203709	0.0739349938
HLA-C-156 z_3++	0.3617370168	0.1240772319	1.0546146732	0.0625239904

Table 12: Estimate, 95% confidence interval and p-value for all three univariable models and the multivariable model for the amino acid on position HLA-C-156, using subset 1.

	Estimate	2.5%	97.5%	p-value
HLA-C-116 z_1+-	0.8944281979	0.7225563368	1.1071825967	0.3054726832
HLA-C-116 z_1++	0.6724580279	0.5066770129	0.8924813791	0.0060038219
HLA-C-116 z_2+-	1.2328075992	0.9523341424	1.5958837440	0.1120298491
HLA-C-116 z_2++	1.4636927633	1.1081618872	1.9332883850	0.0072885310
HLA-C-116 z_3+-	5.6717549e-06	2.144468e-282	1.500083e+271	0.9703266588
HLA-C-116 z_3++	6.3587249e-06	2.404537e-282	1.681544e+271	0.9706073633
HLA-C-116 z_1+-	0.0031729782	1.950550e-141	5.161512e+135	0.9717356642
HLA-C-116 z_1++	4.7886809e-06	1.810743e-282	1.266411e+271	0.9699111204
HLA-C-116 z_2+-	0.0019936603	1.225576e-141	3.243110e+135	0.9694537214
HLA-C-116 z_2++	7.3807570e-06	2.790938e-282	1.951872e+271	0.9709733260
HLA-C-116 z_3+-	0.0023745901	1.459750e-141	3.862767e+135	0.9703123221
HLA-C-116 z_3++	NA	NA	NA	NA

Table 13: Estimate, 95% confidence interval and p-value for all three univariable models and the multivariable model for the amino acid on position HLA-C-116, using subset 1.

	Estimate	2.5%	97.5%	p-value
HLA-B-156 z_1+-	0.8069337197	0.6538060300	0.9959253940	0.0457174609
HLA-B-156 z_1++	0.9450206346	0.7019842384	1.2721995038	0.7092943571
HLA-B-156 z_2+-	0.9075851904	0.6948158963	1.1855095461	0.4768138739
HLA-B-156 z_2++	0.9942451252	0.7457295622	1.3255788948	0.9686278338
HLA-B-156 z_3+-	1.3995564312	1.1394238460	1.7190777699	0.0013551992
HLA-B-156 z_3++	1.9331223308	1.4053375540	2.6591205331	5.0872112e-05
HLA-B-156 z_1+-	1.0514287934	0.7839153953	1.4102319132	0.7377926451
HLA-B-156 z_1++	1.5248919888	0.9518941934	2.4428088684	0.0792755724
HLA-B-156 z_2+-	0.5587596267	0.3678810192	0.8486774369	0.0063454333
HLA-B-156 z_2++	0.3342326535	0.1842503400	0.6063026352	0.0003100785
HLA-B-156 z_3+-	1.9839637940	1.5016710677	2.6211548059	1.4276682e-06
HLA-B-156 z_3++	3.7339734030	2.3335802399	5.9747495013	3.9464366e-08

Table 14: Estimate, 95% confidence interval and p-value for all three univariable models and the multivariable model for the amino acid on position HLA-B-156, using subset 1.

	Estimate	2.5%	97.5%	p-value
HLA-B-116 z_1+-	1.3929638924	1.0944271211	1.7729352352	0.0070777650
HLA-B-116 z_1++	1.7075983027	1.2929983109	2.2551398087	0.0001627188
HLA-B-116 z_2+-	0.8264288825	0.6050022733	1.1288960852	0.2308969421
HLA-B-116 z_2++	0.7981095720	0.5839208336	1.0908651521	0.1572277045
HLA-B-116 z_3+-	1.1269766e-05	4.261163e-282	2.980585e+271	0.9720125707
HLA-B-116 z_3++	7.7221002e-06	2.920034e-282	2.042127e+271	0.9710843323
HLA-B-116 z_1+-	1.7087052774	1.2969154379	2.2512444833	0.0001401411
HLA-B-116 z_1++	2.4542998710	1.7043136933	3.5343187585	1.3969396e-06
HLA-B-116 z_2+-	1.1434321329	0.8020754510	1.6300673970	0.4587727682
HLA-B-116 z_2++	1.5231820097	1.0127675567	2.2908350680	0.0432909655
HLA-B-116 z_3+-	7.6569079e-06	2.894934e-282	2.025200e+271	0.9710635225
HLA-B-116 z_3++	3.5056614e-06	1.325409e-282	9.272349e+270	0.9691454176

Table 15: Estimate, 95% confidence interval and p-value for all three univariable models and the multivariable model for the amino acid on position HLA-B-116, using subset 1.

	Estimate	2.5%	97.5%	p-value
HLA-B-97 z_1++	1.5947734920	0.9413024533	2.7018972298	0.0827252404
HLA-B-97 z_2+-	1.3789179695	0.9996437159	1.9020924520	0.0502543507
HLA-B-97 z_2++	1.8735633269	1.3480605191	2.6039183627	0.0001852862
HLA-B-97 z_3+-	0.8385516890	0.6841844105	1.0277476721	0.0898304631
HLA-B-97 z_3++	0.5712169061	0.4134521821	0.7891813561	0.0006847615
HLA-B-97 z_1++	1.5074713902	0.8830348060	2.5734772593	0.1325527633
HLA-B-97 z_2+-	1.2546197229	0.8868437829	1.7749131012	0.2000110855
HLA-B-97 z_2++	1.6593076606	1.1315129566	2.4332924306	0.0095280117
HLA-B-97 z_3+-	0.9833754021	0.7788676979	1.2415808025	0.8879262456
HLA-B-97 z_3++	0.7687598920	0.5268784757	1.1216851681	0.1724896900

Table 16: Estimate, 95% confidence interval and p-value for all three univariable models and the multivariable model for the amino acid on position HLA-B-97, using subset 1.

	Estimate	2.5%	97.5%	p-value
HLA-B-80 z_1+-	0.8039276029	0.2963232119	2.1810629903	0.6682241361
HLA-B-80 z_1++	0.9302002731	0.3510635786	2.4647175063	0.8842885869
HLA-B-80 z_2+-	0.8339067047	0.6039033175	1.1515094751	0.2699607962
HLA-B-80 z_2++	0.6207780799	0.4514561482	0.8536054411	0.0033459288
HLA-B-80 z_3+-	0.8339067047	0.6039033175	1.1515094751	0.2699607962
HLA-B-80 z_3++	0.6207780799	0.4514561482	0.8536054411	0.0033459288
HLA-B-80 z_1+-	0.9752084145	0.3492797297	2.7228360844	0.9617800996
HLA-B-80 z_1++	1.5063570002	0.5413021117	4.1919500459	0.4327054647
HLA-B-80 z_2+-	0.7579121402	0.5390517078	1.0656321166	0.1108617087
HLA-B-80 z_2++	0.5038943692	0.3521207772	0.7210864901	0.0001780873
HLA-B-80 z_3+-	NA	NA	NA	NA
HLA-B-80 z_3++	NA	NA	NA	NA

Table 17: Estimate, 95% confidence interval and p-value for all three univariable models and the multivariable model for the amino acid on position HLA-B-80, using subset 1.

	Estimate	2.5%	97.5%	p-value
HLA-B-67 z_1+-	1.2584814765	0.9580907432	1.6530538866	0.0984777600
HLA-B-67 z_1++	1.9474563984	1.4547703048	2.6070001642	7.5058532e-06
HLA-B-67 z_2+-	0.6550392905	0.5285032474	0.8118710229	0.0001119945
HLA-B-67 z_2++	0.4935995045	0.3679449371	0.6621655750	2.4755734e-06
HLA-B-67 z_3++	0.6790499763	0.3017063221	1.5283367851	0.3497154049
HLA-B-67 z_1+-	0.3473679682	0.0755607197	1.5969210690	0.1742869679
HLA-B-67 z_1++	0.4369801738	0.0711482774	2.6838551695	0.3713597455
HLA-B-67 z_2+-	0.8095040452	0.3043391189	2.1531796553	0.6720014568
HLA-B-67 z_2++	0.2142946851	0.0346112817	1.3267989450	0.0977261244
HLA-B-67 z_3++	NA	NA	NA	NA

Table 18: Estimate, 95% confidence interval and p-value for all three univariable models and the multivariable model for the amino acid on position HLA-B-67, using subset 1.

	Estimate	2.5%	97.5%	p-value
HLA-DRB1-57 z_1+-	600847.81448	2.271731e-271	1.589176e+282	0.9673165122
HLA-DRB1-57 z_1++	662890.82992	2.506683e-271	1.753010e+282	0.9670752659
HLA-DRB1-57 z_2+-	4.7452643598	0.9307714350	24.192334439	0.0609799479
HLA-DRB1-57 z_2++	5.7731768680	1.1601771931	28.728000642	0.0322389845
HLA-DRB1-57 z_3+-	600847.81448	2.271731e-271	1.589176e+282	0.9673165122
HLA-DRB1-57 z_3++	662890.82992	2.506683e-271	1.753010e+282	0.9670752659
HLA-DRB1-57 z_1+-	165812.91378	6.255531e-272	4.395137e+281	0.9704775064
HLA-DRB1-57 z_1++	140487.16581	5.301098e-272	3.723123e+281	0.9708844487
HLA-DRB1-57 z_2+-	3.6551893738	0.6834856058	19.547462660	0.1297412379
HLA-DRB1-57 z_2++	4.7832672035	0.9213114680	24.833778731	0.0625413602
HLA-DRB1-57 z_3+-	NA	NA	NA	NA
HLA-DRB1-57 z_3++	NA	NA	NA	NA

Table 19: Estimate, 95% confidence interval and p-value for all three univariable models and the multivariable model for the amino acid on position HLA-DRB1-57, using subset 1.

	Estimate	2.5%	97.5%	p-value
HLA-DRB1-37 z_1+-	0.4581249294	0.3490043367	0.6013634471	1.8689456e-08
HLA-DRB1-37 z_1++	0.2692043764	0.2016459113	0.3593973010	5.5313798e-19
HLA-DRB1-37 z_2+-	1.4508644803	1.1230105660	1.8744327115	0.0044041273
HLA-DRB1-37 z_2++	2.1845190650	1.6574324122	2.8792266340	2.9159919e-08
HLA-DRB1-37 z_3++	521260.11881	1.971115e-271	1.378468e+282	0.9676653313
HLA-DRB1-37 z_1+-	0.3327063024	0.2308467025	0.4795107855	3.6083531e-09
HLA-DRB1-37 z_1++	0.1596993401	0.1014070802	0.2514999858	2.4341952e-15
HLA-DRB1-37 z_2+-	0.8087546991	0.5680204000	1.1515152681	0.2390351273
HLA-DRB1-37 z_2++	0.5367544533	0.3438545871	0.8378697099	0.0061718416
HLA-DRB1-37 z_3++	708814.73984	2.680065e-271	1.874649e+282	0.9669108349

Table 20: Estimate, 95% confidence interval and p-value for all three univariable models and the multivariable model for the amino acid on position HLA-DRB1-37, using subset 1.

	Estimate	2.5%	97.5%	p-value
HLA-DRB1-30 z_1+-	1.2543194974	1.0230917713	1.5378067203	0.0292905916
HLA-DRB1-30 z_1++	0.9403882669	0.6254881571	1.4138238788	0.7676663568
HLA-DRB1-30 z_2+-	1.5811734620	1.0032806409	2.4919343752	0.0483731090
HLA-DRB1-30 z_2++	1.3280585884	0.8544248321	2.0642419883	0.2073744083
HLA-DRB1-30 z_3+-	1.5445975e-05	5.839465e-282	4.085616e+271	0.9727866096
HLA-DRB1-30 z_3++	7.7814364e-06	2.942478e-282	2.057814e+271	0.9711031269
HLA-DRB1-30 z_1+-	0.5311358916	0.0329615402	8.5586211518	0.6554915460
HLA-DRB1-30 z_1++	0.2699373106	0.0009055952	80.462159615	0.6523443923
HLA-DRB1-30 z_2+-	0.7893829313	0.0363915464	17.122806631	0.8802514173
HLA-DRB1-30 z_2++	0.3469094743	0.0011471284	104.91081619	0.7163939535
HLA-DRB1-30 z_3+-	7.5529494e-06	2.836113e-282	2.011451e+271	0.9710302696
HLA-DRB1-30 z_3++	1.9889368e-06	7.331857e-283	5.395454e+270	0.9677552308

Table 21: Estimate, 95% confidence interval and p-value for all three univariable models and the multivariable model for the amino acid on position HLA-DRB1-30, using subset 1.

	Estimate	2.5%	97.5%	p-value
HLA-DRB1-13 z_1+-	1.3374560453	0.8785070245	2.0361688901	0.1751221355
HLA-DRB1-13 z_1++	1.0620608404	0.7064443544	1.5966908385	0.7722433154
HLA-DRB1-13 z_2+-	1.7047183725	0.8437190886	3.4443510511	0.1371708667
HLA-DRB1-13 z_2++	3.5259163589	1.7864404169	6.9591384369	0.0002806288
HLA-DRB1-13 z_3+-	2.1931085237	1.3980344708	3.4403479293	0.0006296211
HLA-DRB1-13 z_3++	3.8061431632	2.4524692333	5.9069959295	2.5162506e-09
HLA-DRB1-13 z_1+-	2.0551136307	1.3332953653	3.1677092302	0.0011024287
HLA-DRB1-13 z_1++	2.4752108656	1.5930696893	3.8458259990	5.5511242e-05
HLA-DRB1-13 z_2+-	2.2464405518	1.1035345082	4.5730288587	0.0256409031
HLA-DRB1-13 z_2++	6.3984184689	3.1691295757	12.918297571	2.2468520e-07
HLA-DRB1-13 z_3+-	2.7344329057	1.7265405847	4.3306965283	1.8040559e-05
HLA-DRB1-13 z_3++	6.2909669139	3.9336764391	10.060884601	1.6302597e-14

Table 22: Estimate, 95% confidence interval and p-value for all three univariable models and the multivariable model for the amino acid on position HLA-DRB1-13, using subset 1.

	Estimate	2.5%	97.5%	p-value
HLA-DRB1-11 z_1+-	0.5030981788	0.4020464291	0.6295486271	1.9128618e-09
HLA-DRB1-11 z_1++	0.3451021787	0.1868989603	0.6372187067	0.0006734229
HLA-DRB1-11 z_2+-	0.5953705654	0.4858574226	0.7295681690	5.7276194e-07
HLA-DRB1-11 z_2++	0.2711398243	0.1806294388	0.4070034475	3.0235599e-10
HLA-DRB1-11 z_3+-	0.4633160453	0.3729518479	0.5755749946	3.6507836e-12
HLA-DRB1-11 z_3++	0.1957094102	0.1441084930	0.2657870639	1.5391899e-25
HLA-DRB1-11 z_1+-	1.2615390093	0.7792548602	2.0423108706	0.3445424769
HLA-DRB1-11 z_1++	2.4628632832	0.8005541273	7.5768712502	0.1159529642
HLA-DRB1-11 z_2+-	1.6500124927	0.9866759704	2.7593063051	0.0562833332
HLA-DRB1-11 z_2++	2.2107621170	0.7864245471	6.2147972817	0.1324856187
HLA-DRB1-11 z_3+-	0.3001791654	0.1719380527	0.5240697444	2.3104780e-05
HLA-DRB1-11 z_3++	0.0859885241	0.0302795394	0.2441921646	4.0788984e-06

Table 23: Estimate, 95% confidence interval and p-value for all three univariable models and the multivariable model for the amino acid on position HLA-DRB1-11, using subset 1.

	Estimate	2.5%	97.5%	p-value
HLA-DQB1-57 z_1+-	1.2811935751	0.8169982725	2.0091315148	0.2803794142
HLA-DQB1-57 z_1++	1.0574014784	0.6834182733	1.6360374463	0.8020936306
HLA-DQB1-57 z_2+-	0.7241882148	0.5728795124	0.9154605098	0.0069630641
HLA-DQB1-57 z_2++	0.4443460880	0.3384824959	0.5833195168	5.1536368e-09
HLA-DQB1-57 z_3+-	1.2811935751	0.8169982725	2.0091315148	0.2803794142
HLA-DQB1-57 z_3++	1.0574014784	0.6834182733	1.6360374463	0.8020936306
HLA-DQB1-57 z_1+-	1.7030831766	1.0543217112	2.7510505338	0.0295429475
HLA-DQB1-57 z_1++	1.9375109423	1.1837526559	3.1712272264	0.0085135881
HLA-DQB1-57 z_2+-	0.6329794884	0.4888066238	0.8196759480	0.0005247420
HLA-DQB1-57 z_2++	0.3670309695	0.2654078026	0.5075650800	1.3620861e-09
HLA-DQB1-57 z_3+-	NA	NA	NA	NA
HLA-DQB1-57 z_3++	NA	NA	NA	NA

Table 24: Estimate, 95% confidence interval and p-value for all three univariable models and the multivariable model for the amino acid on position HLA-DQB1-57, using subset 1.

	Estimate	2.5%	97.5%	p-value
HLA-DQB1-55 z_1+-	0.5030844401	0.3790477903	0.6677098782	1.9716781e-06
HLA-DQB1-55 z_1++	0.2737559870	0.2037801057	0.3677608282	7.8579991e-18
HLA-DQB1-55 z_2++	7.9579470e-06	3.009199e-282	2.104510e+271	0.9711582012
HLA-DQB1-55 z_3+-	1.8293963241	1.4749864582	2.2689638214	3.8547224e-08
HLA-DQB1-55 z_3++	3.6417034144	2.7110365694	4.8918571990	9.2051529e-18
HLA-DQB1-55 z_1+-	0.5022758874	0.3784313505	0.6666494907	1.8687827e-06
HLA-DQB1-55 z_1++	0.2737244204	0.2037573060	0.3677171621	7.7989488e-18
HLA-DQB1-55 z_2++	8.3164943e-06	3.144719e-282	2.199371e+271	0.9712664095
HLA-DQB1-55 z_3+-	NA	NA	NA	NA
HLA-DQB1-55 z_3++	NA	NA	NA	NA

Table 25: Estimate, 95% confidence interval and p-value for all three univariable models and the multivariable model for the amino acid on position HLA-DQB1-55, using subset 1.

	Estimate	2.5%	97.5%	p-value
HLA-DQB1-37 z_1+-	1.2126510512	0.3776236493	3.8941485115	0.7460037706
HLA-DQB1-37 z_2++	7.9579470e-06	3.009199e-282	2.104510e+271	0.9711582012
HLA-DQB1-37 z_3++	7.9579470e-06	3.009199e-282	2.104510e+271	0.9711582012
HLA-DQB1-37 z_1+-	1.2144908333	0.3782075686	3.8999430637	0.7440701737
HLA-DQB1-37 z_2++	7.9353296e-06	3.000646e-282	2.098529e+271	0.9711512129
HLA-DQB1-37 z_3++	NA	NA	NA	NA

Table 26: Estimate, 95% confidence interval and p-value for all three univariable models and the multivariable model for the amino acid on position HLA-DQB1-37, using subset 1.

	Estimate	2.5%	97.5%	p-value
HLA-DQB1-26 z_1+-	1.2342125516	1.0077089298	1.5116275917	0.0419316080
HLA-DQB1-26 z_1++	0.6485999694	0.4407793633	0.9544047552	0.0280377997
HLA-DQB1-26 z_2+-	1.8478259547	1.4480310668	2.3580024193	7.9715306e-07
HLA-DQB1-26 z_2++	4.5412824872	1.7753553625	11.616404841	0.0015896322
HLA-DQB1-26 z_3+-	1.5757076229	1.2790282792	1.9412037662	1.9352428e-05
HLA-DQB1-26 z_3++	1.6360758833	1.2171950183	2.1991088163	0.0011043157
HLA-DQB1-26 z_1+-	1.3773060969	1.0482728835	1.8096166698	0.0215364764
HLA-DQB1-26 z_1++	0.8221050723	0.5536564840	1.2207149548	0.3314573353
HLA-DQB1-26 z_2+-	1.8813820188	1.4319112463	2.4719397308	5.6936457e-06
HLA-DQB1-26 z_2++	5.1903015888	2.0164576166	13.359681036	0.0006404439
HLA-DQB1-26 z_3+-	1.0354298272	0.7880401341	1.3604826463	0.8026337369
HLA-DQB1-26 z_3++	NA	NA	NA	NA

Table 27: Estimate, 95% confidence interval and p-value for all three univariable models and the multivariable model for the amino acid on position HLA-DQB1-26, using subset 1.

	Estimate	2.5%	97.5%	p-value
HLA-DQB1-5 z_1+-	1.2155768368	0.9926284474	1.4886003418	0.0589731790
HLA-DQB1-5 z_1++	0.6575932907	0.4461402631	0.9692667794	0.0342033571
HLA-DQB1-5 z_2+-	1.8405689793	1.4825762964	2.2850049443	3.2343792e-08
HLA-DQB1-5 z_2++	3.6680079928	2.7351690228	4.9189949590	3.9492811e-18
HLA-DQB1-5 z_3+-	2.2085925286	1.5591690149	3.1285132725	8.1953932e-06
HLA-DQB1-5 z_3++	4.5685088379	3.2276249975	6.4664491749	1.0351536e-17
HLA-DQB1-5 z_1+-	2.1670409862	1.6986702235	2.7645546329	4.8319675e-10
HLA-DQB1-5 z_1++	1.8260524794	1.1269671035	2.9587976857	0.0144699623
HLA-DQB1-5 z_2+-	2.1586170171	1.6924248916	2.7532255343	5.7002980e-10
HLA-DQB1-5 z_2++	6.2427538046	4.1915326996	9.2977862414	2.0444945e-19
HLA-DQB1-5 z_3+-	1.0246340479	0.8031597133	1.3071807695	0.8447276228
HLA-DQB1-5 z_3++	NA	NA	NA	NA

Table 28: Estimate, 95% confidence interval and p-value for all three univariable models and the multivariable model for the amino acid on position HLA-DQB1-5, using subset 1.

	Estimate	2.5%	97.5%	p-value
HLA-DPA1-228 z_1+-	0.2525429030	0.0290369612	2.1964391277	0.2124015208
HLA-DPA1-228 z_1++	0.3998572701	0.0465439192	3.4351605788	0.4035207662
HLA-DPA1-228 z_2+-	0.6315826219	0.4807416856	0.8297524850	0.0009657155
HLA-DPA1-228 z_2++	2.5008923802	0.2911072065	21.485083701	0.4035207662
HLA-DPA1-228 z_3+-	0.6315826219	0.4807416856	0.8297524850	0.0009657155
HLA-DPA1-228 z_3++	2.5008923802	0.2911072065	21.485083701	0.4035207662
HLA-DPA1-228 z_1+-	0.2525429030	0.0290369612	2.1964391277	0.2124015208
HLA-DPA1-228 z_1++	0.3998572701	0.0465439192	3.4351605788	0.4035207662
HLA-DPA1-228 z_2+-	NA	NA	NA	NA
HLA-DPA1-228 z_2++	NA	NA	NA	NA
HLA-DPA1-228 z_3+-	NA	NA	NA	NA
HLA-DPA1-228 z_3++	NA	NA	NA	NA

Table 29: Estimate, 95% confidence interval and p-value for all three univariable models and the multivariable model for the amino acid on position HLA-DPA1-228, using subset 1.

	Estimate	2.5%	97.5%	p-value
HLA-DPB1-84 z_1++	1.4998681386	0.5553204927	4.0510020112	0.4239119707
HLA-DPB1-84 z_2+-	0.7438100792	0.5985684331	0.9242943719	0.0075803592
HLA-DPB1-84 z_2++	0.3708960324	0.1947740761	0.7062740056	0.0025429251
HLA-DPB1-84 z_3++	1.4998681386	0.5553204927	4.0510020112	0.4239119707
HLA-DPB1-84 z_1++	1.5687942507	0.5799482748	4.2436808728	0.3751268176
HLA-DPB1-84 z_2+-	0.7425562020	0.5975142876	0.9228059054	0.0072638455
HLA-DPB1-84 z_2++	0.3691649484	0.1938468639	0.7030434044	0.0024295387
HLA-DPB1-84 z_3++	NA	NA	NA	NA

Table 30: Estimate, 95% confidence interval and p-value for all three univariable models and the multivariable model for the amino acid on position HLA-DPB1-84, using subset 1.

Subset 2

	Estimate	2.5%	97.5%	p-value
HLA-A-9 z_1+-	0.9909488401	0.8630101864	1.1378540130	0.8974250722
HLA-A-9 z_1++	0.7475585782	0.5110233439	1.0935778854	0.1338590767
HLA-A-9 z_2+-	1.3255801873	0.8947166359	1.9639322245	0.1599358268
HLA-A-9 z_2++	1.3376878135	0.9144296106	1.9568577676	0.1338590767
HLA-A-9 z_3+-	0.4017567993	0.0842694386	1.9153862717	0.2524705419
HLA-A-9 z_3++	0.3539473968	0.0750297560	1.6697210061	0.1894386876
HLA-A-9 z_1+-	0.9044879406	0.7698146366	1.0627213304	0.2223094881
HLA-A-9 z_1++	0.5766665890	0.3782927612	0.8790661333	0.0104917360
HLA-A-9 z_3+-	0.2717934060	0.0547302689	1.3497404092	0.1111198130
HLA-A-9 z_3++	0.2089446992	0.0419801233	1.0399656766	0.0558626389
HLA-A-9 z_2+-	NA	NA	NA	NA
HLA-A-9 z_2++	NA	NA	NA	NA

Table 31: Estimate, 95% confidence interval and p-value for all three univariable models and the multivariable model for the amino acid on position HLA-A-9, using subset 2.

	Estimate	2.5%	97.5%	p-value
HLA-A-62 z_1+-	1.3685761e-05	2.999220e-173	6.244958e+162	0.9546584661
HLA-A-62 z_1++	1.4769857e-05	3.237280e-173	6.738639e+162	0.9549667588
HLA-A-62 z_2+-	0.8752663642	0.7355964878	1.0414557723	0.1330931484
HLA-A-62 z_2++	0.7712186504	0.6439214846	0.9236812577	0.0047650187
HLA-A-62 z_3+-	1.1317645340	0.9969717465	1.2847816049	0.0557366783
HLA-A-62 z_3++	1.3155722053	1.0942647941	1.5816374947	0.0035167639
HLA-A-62 z_1+-	1.4678812e-05	3.216744e-173	6.698309e+162	0.9549417721
HLA-A-62 z_1++	1.7643797e-05	3.867050e-173	8.050156e+162	0.9556858565
HLA-A-62 z_2+-	0.9518320562	0.6580699953	1.3767293290	0.7931989443
HLA-A-62 z_2++	0.7578121083	0.6295620937	0.9121883245	0.0033730293
HLA-A-62 z_3+-	0.9064823352	0.6267015794	1.3111666717	0.6021121342
HLA-A-62 z_3++	NA	NA	NA	NA

Table 32: Estimate, 95% confidence interval and p-value for all three univariable models and the multivariable model for the amino acid on position HLA-A-62, using subset 2.

	Estimate	2.5%	97.5%	p-value
HLA-A-76 z_1+-	0.8573369138	0.7574701289	0.9703703891	0.0148523021
HLA-A-76 z_1++	0.8282004879	0.6782749080	1.0112655504	0.0643110490
HLA-A-76 z_2+-	0.9281054718	0.8087350596	1.0650951218	0.2881601320
HLA-A-76 z_2++	0.8362333724	0.5550491510	1.2598636568	0.3923997400
HLA-A-76 z_3+-	0.8994905060	0.7906050434	1.0233721340	0.1076104866
HLA-A-76 z_3++	0.8079768249	0.5819124713	1.1218638226	0.2029189405
HLA-A-76 z_1+-	0.7994839831	0.6285171525	1.0169565567	0.0683044273
HLA-A-76 z_1++	0.7702597783	0.5527969263	1.0732695820	0.1230254338
HLA-A-76 z_2+-	1.0822384495	0.8645316436	1.3547682960	0.4904022002
HLA-A-76 z_2++	1.0176742062	0.6076248776	1.7044410591	0.9469133221
HLA-A-76 z_3+-	1.0648606916	0.8506375354	1.3330334547	0.5834353297
HLA-A-76 z_3++	NA	NA	NA	NA

Table 33: Estimate, 95% confidence interval and p-value for all three univariable models and the multivariable model for the amino acid on position HLA-A-76, using subset 2.

	Estimate	2.5%	97.5%	p-value
HLA-A-152 z_1+-	0.9232245273	0.8112092259	1.0507073891	0.2261061974
HLA-A-152 z_1++	0.8304666934	0.6968033839	0.9897697755	0.0380039158
HLA-A-152 z_2+-	0.9380624960	0.8279611408	1.0628050075	0.3155029188
HLA-A-152 z_2++	0.9022802598	0.6846396390	1.1891068247	0.4653026692
HLA-A-152 z_3+-	0.9589552953	0.8442155953	1.0892895885	0.5191961041
HLA-A-152 z_3++	0.7428640437	0.5401864955	1.0215860485	0.0674620619
HLA-A-152 z_1+-	0.9677440480	0.6864333968	1.3643400025	0.8515787589
HLA-A-152 z_1++	0.9454041149	0.4907243883	1.8213664568	0.8667329239
HLA-A-152 z_2+-	0.9428011232	0.6752531736	1.3163565795	0.7294370577
HLA-A-152 z_2++	0.9144915408	0.4548241589	1.8387211006	0.8019440331
HLA-A-152 z_3+-	0.9670765875	0.6908327522	1.3537822623	0.8453446000
HLA-A-152 z_3++	0.7518996026	0.3769065324	1.4999819951	0.4183580272

Table 34: Estimate, 95% confidence interval and p-value for all three univariable models and the multivariable model for the amino acid on position HLA-A-152, using subset 2.

	Estimate	2.5%	97.5%	p-value
HLA-A-156 z_1+-	0.9338421096	0.8257193914	1.0561228122	0.2756117451
HLA-A-156 z_1++	0.7184663686	0.5687032705	0.9076682861	0.0055674662
HLA-A-156 z_2+-	0.8680785332	0.7654321035	0.9844901152	0.0275650976
HLA-A-156 z_2++	0.7630332455	0.6329300653	0.9198800398	0.0045749136
HLA-A-156 z_3+-	0.9036705766	0.7743563084	1.0545797873	0.1986126661
HLA-A-156 z_3++	0.8908447500	0.4790135697	1.6567471542	0.7150132372
HLA-A-156 z_1+-	1.0829272233	0.7636648918	1.5356622826	0.6548511930
HLA-A-156 z_1++	0.8320696754	0.4321983511	1.6019032535	0.5822663333
HLA-A-156 z_2+-	0.8206068929	0.5718108045	1.1776546846	0.2833948419
HLA-A-156 z_2++	0.8209222047	0.4398906434	1.5320018196	0.5353270977
HLA-A-156 z_3+-	0.9920159720	0.6995697866	1.4067155381	0.9641212389
HLA-A-156 z_3++	NA	NA	NA	NA

Table 35: Estimate, 95% confidence interval and p-value for all three univariable models and the multivariable model for the amino acid on position HLA-A-156, using subset 2.

	Estimate	2.5%	97.5%	p-value
HLA-C-156 z_1+-	1.0983890793	0.9713230548	1.2420775597	0.1346268005
HLA-C-156 z_1++	0.9970000002	0.8018504108	1.2396439374	0.9784328066
HLA-C-156 z_2+-	0.9130509342	0.7981980568	1.0444300151	0.1847788100
HLA-C-156 z_2++	0.9916776734	0.8371105704	1.1747846015	0.9229895926
HLA-C-156 z_3+-	0.8293889949	0.7288862836	0.9437495537	0.0045338015
HLA-C-156 z_3++	1.0145624894	0.7166506186	1.4363164116	0.9350334628
HLA-C-156 z_1+-	0.8851435365	0.6430804153	1.2183220970	0.4541671273
HLA-C-156 z_1++	0.5446542710	0.2915796702	1.0173832583	0.0566624729
HLA-C-156 z_2+-	1.1570082470	0.8201916728	1.6321405447	0.4060937281
HLA-C-156 z_2++	1.7830288288	0.9376975877	3.3904233581	0.0777701291
HLA-C-156 z_3+-	0.6716016593	0.4938202612	0.9133865582	0.0111671908
HLA-C-156 z_3++	0.5981623394	0.3145049057	1.1376553365	0.1171705532

Table 36: Estimate, 95% confidence interval and p-value for all three univariable models and the multivariable model for the amino acid on position HLA-C-156, using subset 2.

	Estimate	2.5%	97.5%	p-value
HLA-C-116 z_1+-	0.9511118540	0.8267565498	1.0941718684	0.4832345982
HLA-C-116 z_1++	0.7858388509	0.6654277430	0.9280387031	0.0045109587
HLA-C-116 z_2+-	1.1762208461	1.0191239021	1.3575341289	0.0264905177
HLA-C-116 z_2++	1.2920511261	1.0937497620	1.5263053493	0.0025775786
HLA-C-116 z_3+-	1.2108869032	0.0745569866	19.666125990	0.8929730660
HLA-C-116 z_3++	1.3708655395	0.0856700372	21.936167983	0.8235510224
HLA-C-116 z_1+-	1.3875555303	0.3360496972	5.7292429225	0.6507510602
HLA-C-116 z_1++	1.1595783972	0.0723170171	18.593439180	0.9167085735
HLA-C-116 z_2+-	1.0105297394	0.2447269008	4.1726935242	0.9884492236
HLA-C-116 z_2++	1.5212606401	0.0948949428	24.387326300	0.7669485541
HLA-C-116 z_3+-	1.0372314553	0.2512182422	4.2825277437	0.9597024418
HLA-C-116 z_3++	NA	NA	NA	NA

Table 37: Estimate, 95% confidence interval and p-value for all three univariable models and the multivariable model for the amino acid on position HLA-C-116, using subset 2.

	Estimate	2.5%	97.5%	p-value
HLA-B-156 z_1+-	0.9475859710	0.8325220996	1.0785529572	0.4150210604
HLA-B-156 z_1++	0.9783724976	0.8212892697	1.1655001220	0.8065646807
HLA-B-156 z_2+-	1.1271846531	0.9683563585	1.3120637160	0.1223445120
HLA-B-156 z_2++	1.2415619195	1.0497628113	1.4684040847	0.0114978758
HLA-B-156 z_3+-	1.2975836083	1.1411241638	1.4754951949	7.0769895e-05
HLA-B-156 z_3++	1.4768441981	1.2332491030	1.7685549335	2.2394940e-05
HLA-B-156 z_1+-	0.8151549910	0.6757926318	0.9832567388	0.0326408830
HLA-B-156 z_1++	0.7408117358	0.5555100187	0.9879246267	0.0410838743
HLA-B-156 z_2+-	1.0569182775	0.8082759907	1.3820480356	0.6858233985
HLA-B-156 z_2++	1.1212263086	0.7675037880	1.6379703330	0.5540691020
HLA-B-156 z_3+-	1.3410943134	1.1177395303	1.6090814621	0.0015913961
HLA-B-156 z_3++	1.5431966559	1.1631937489	2.0473424321	0.0026290890

Table 38: Estimate, 95% confidence interval and p-value for all three univariable models and the multivariable model for the amino acid on position HLA-B-156, using subset 2.

	Estimate	2.5%	97.5%	p-value
HLA-B-116 z_1+-	1.1195850414	0.9776717412	1.2820976736	0.1023790907
HLA-B-116 z_1++	1.2969753645	1.0953163948	1.5357618164	0.0025623194
HLA-B-116 z_2+-	0.8427798894	0.6821818634	1.0411856137	0.1127873294
HLA-B-116 z_2++	0.7720883428	0.6270362933	0.9506952236	0.0148430507
HLA-B-116 z_3+-	2.1841881172	0.4825308461	9.8867829265	0.3105459415
HLA-B-116 z_3++	1.9013881849	0.4246223685	8.5140993455	0.4008466904
HLA-B-116 z_1+-	1.1361480554	0.9729339573	1.3267420611	0.1067045600
HLA-B-116 z_1++	1.3173878681	1.0558191502	1.6437576405	0.0146485650
HLA-B-116 z_2+-	0.9291499251	0.7345075010	1.1753720447	0.5400720278
HLA-B-116 z_2++	0.9418553882	0.7184215265	1.2347786634	0.6646005588
HLA-B-116 z_3+-	2.1646686642	0.4727174455	9.9124550397	0.3198324927
HLA-B-116 z_3++	1.7792121516	0.3861578989	8.1976722200	0.4597804005

Table 39: Estimate, 95% confidence interval and p-value for all three univariable models and the multivariable model for the amino acid on position HLA-B-116, using subset 2.

	Estimate	2.5%	97.5%	p-value
HLA-B-97 z_1+-	1.7633207335	0.3185116958	9.7619649449	0.5159385213
HLA-B-97 z_1++	2.8973010652	0.5297992427	15.844404417	0.2197669246
HLA-B-97 z_2+-	1.1461067237	0.9674567342	1.3577461147	0.1147266857
HLA-B-97 z_2++	1.3066237534	1.0925903469	1.5625853164	0.0033882378
HLA-B-97 z_3+-	0.9125951284	0.8038394630	1.0360649194	0.1577377122
HLA-B-97 z_3++	0.8494365682	0.7086499772	1.0181930523	0.0775707416
HLA-B-97 z_1+-	1.6982024627	0.3063734003	9.4129960397	0.5444561889
HLA-B-97 z_1++	2.7425225180	0.5006454793	15.023464853	0.2449703345
HLA-B-97 z_2+-	1.1241785459	0.9330988745	1.3543874476	0.2181470035
HLA-B-97 z_2++	1.2281408898	0.9899230420	1.5236841464	0.0617753242
HLA-B-97 z_3+-	0.9536464499	0.8247830598	1.1026433443	0.5216627581
HLA-B-97 z_3++	0.9508544082	0.7655779882	1.1809693062	0.6485816505

Table 40: Estimate, 95% confidence interval and p-value for all three univariable models and the multivariable model for the amino acid on position HLA-B-97, using subset 2.

	Estimate	2.5%	97.5%	p-value
HLA-B-80 z_1+-	0.7733767965	0.4190611568	1.4272658289	0.4110686334
HLA-B-80 z_1++	0.9950832959	0.5459735148	1.8136241758	0.9871596513
HLA-B-80 z_2+-	0.9855937616	0.8090452405	1.2006684104	0.8854324985
HLA-B-80 z_2++	0.9955909537	0.8178654699	1.2119369060	0.9648696268
HLA-B-80 z_3+-	0.9855937616	0.8090452405	1.2006684104	0.8854324985
HLA-B-80 z_3++	0.9955909537	0.8178654699	1.2119369060	0.9648696268
HLA-B-80 z_1+-	0.8126928485	0.4323506005	1.5276251847	0.5195121778
HLA-B-80 z_1++	1.1097088743	0.5908245113	2.0842970494	0.7461787981
HLA-B-80 z_2+-	0.9338593523	0.7585417074	1.1496972169	0.5189096133
HLA-B-80 z_2++	0.8612780659	0.6905624764	1.0741966615	0.1851864303
HLA-B-80 z_3+-	NA	NA	NA	NA
HLA-B-80 z_3++	NA	NA	NA	NA

Table 41: Estimate, 95% confidence interval and p-value for all three univariable models and the multivariable model for the amino acid on position HLA-B-80, using subset 2.

	Estimate	2.5%	97.5%	p-value
HLA-B-67 z_1+-	1.1392642088	0.9704657362	1.3374227333	0.1110358956
HLA-B-67 z_1++	1.3568453773	1.1410464548	1.6134569895	0.0005545041
HLA-B-67 z_2+-	0.8874515266	0.7795384765	1.0103031932	0.0710737142
HLA-B-67 z_2++	0.7792196012	0.6527995542	0.9301219386	0.0057448154
HLA-B-67 z_3+-	2.0930961524	0.1873030693	23.390174644	0.5486406041
HLA-B-67 z_3++	2.7695723170	0.2508770233	30.574863805	0.4057461652
HLA-B-67 z_1+-	1.5829658065	0.4633759285	5.4076627428	0.4637029123
HLA-B-67 z_1++	3.1427718799	0.2842471228	34.747986149	0.3503151535
HLA-B-67 z_2+-	1.6988065833	0.4973123177	5.8030812922	0.3978451962
HLA-B-67 z_2++	2.3411853204	0.2112817326	25.942369155	0.4881947046
HLA-B-67 z_3+-	1.2873027540	0.3768077661	4.3978615341	0.6870237894
HLA-B-67 z_3++	NA	NA	NA	NA

Table 42: Estimate, 95% confidence interval and p-value for all three univariable models and the multivariable model for the amino acid on position HLA-B-67, using subset 2.

	Estimate	2.5%	97.5%	p-value
HLA-B-66 z_1+-	0.7557470659	0.5850429967	0.9762592336	0.0320406712
HLA-B-66 z_1++	0.3610665783	0.0327066052	3.9860166818	0.4057461652
HLA-B-66 z_2+-	0.7557470659	0.5850429967	0.9762592336	0.0320406712
HLA-B-66 z_2++	0.3610665783	0.0327066052	3.9860166818	0.4057461652
HLA-B-66 z_3+-	0.7557470659	0.5850429967	0.9762592336	0.0320406712
HLA-B-66 z_3++	0.3610665783	0.0327066052	3.9860166818	0.4057461652
HLA-B-66 z_1+-	0.7557470659	0.5850429967	0.9762592336	0.0320406712
HLA-B-66 z_1++	0.3610665783	0.0327066052	3.9860166818	0.4057461652
HLA-B-66 z_2+-	NA	NA	NA	NA
HLA-B-66 z_2++	NA	NA	NA	NA
HLA-B-66 z_3+-	NA	NA	NA	NA
HLA-B-66 z_3++	NA	NA	NA	NA

Table 43: Estimate, 95% confidence interval and p-value for all three univariable models and the multivariable model for the amino acid on position HLA-B-66, using subset 2.

	Estimate	2.5%	97.5%	p-value
HLA-DRB1-74 z_1+-	0.7912002399	0.2192053346	2.8557599690	0.7206210444
HLA-DRB1-74 z_1++	0.9725354848	0.2738570187	3.4537192928	0.9656458273
HLA-DRB1-74 z_2+-	0.7466682399	0.6587839727	0.8462765999	4.8230527e-06
HLA-DRB1-74 z_2++	0.4558405218	0.3515860839	0.5910091179	3.0422070e-09
HLA-DRB1-74 z_3+-	1.6009478759	1.2738823736	2.0119864708	5.4352581e-05
HLA-DRB1-74 z_3++	2.1229170825	1.6988034068	2.6529125861	3.5850227e-11
HLA-DRB1-74 z_1+-	0.8302160214	0.2286481697	3.0144944654	0.7773193377
HLA-DRB1-74 z_1++	1.1311265076	0.3181805181	4.0211361264	0.8489949756
HLA-DRB1-74 z_2+-	0.6938789916	0.5427247318	0.8871312228	0.0035530817
HLA-DRB1-74 z_2++	0.4460848731	0.3436504670	0.5790526513	1.3223512e-09
HLA-DRB1-74 z_3+-	1.0781283623	0.8432731765	1.3783917216	0.5484335614
HLA-DRB1-74 z_3++	NA	NA	NA	NA

Table 44: Estimate, 95% confidence interval and p-value for all three univariable models and the multivariable model for the amino acid on position HLA-DRB1-74, using subset 2.

	Estimate	2.5%	97.5%	p-value
HLA-DRB1-57 z_1+-	1.2655667297	0.7311318078	2.1906571844	0.4001748024
HLA-DRB1-57 z_1++	1.8524749236	1.0847647910	3.1635091504	0.0239489051
HLA-DRB1-57 z_2+-	1.0134187273	0.7174287275	1.4315254985	0.9397091356
HLA-DRB1-57 z_2++	1.4542856767	1.0406844199	2.0322652948	0.0282690108
HLA-DRB1-57 z_3+-	1.2655667297	0.7311318078	2.1906571844	0.4001748024
HLA-DRB1-57 z_3++	1.8524749236	1.0847647910	3.1635091504	0.0239489051
HLA-DRB1-57 z_1+-	1.4800870133	0.7493166993	2.9235402988	0.2588980909
HLA-DRB1-57 z_1++	1.7554325024	0.8781428915	3.5091592730	0.1113247064
HLA-DRB1-57 z_2+-	0.8444003814	0.5471446114	1.3031509208	0.4449005081
HLA-DRB1-57 z_2++	1.0929798035	0.6969484233	1.7140505825	0.6985510769
HLA-DRB1-57 z_3+-	NA	NA	NA	NA
HLA-DRB1-57 z_3++	NA	NA	NA	NA

Table 45: Estimate, 95% confidence interval and p-value for all three univariable models and the multivariable model for the amino acid on position HLA-DRB1-57, using subset 2.

	Estimate	2.5%	97.5%	p-value
HLA-DRB1-37 z_1+-	0.6484787397	0.5590358659	0.7522320149	1.0663259e-08
HLA-DRB1-37 z_1++	0.4824237274	0.4081116759	0.5702670776	1.3357583e-17
HLA-DRB1-37 z_2+-	1.1738157001	0.9474119478	1.4543233290	0.1426898153
HLA-DRB1-37 z_2++	1.2317344025	0.9984333940	1.5195501749	0.0517352121
HLA-DRB1-37 z_3+-	0.8798279929	0.1432421491	5.4041167483	0.8900507027
HLA-DRB1-37 z_3++	0.9881818176	0.1647277524	5.9279829323	0.9896230147
HLA-DRB1-37 z_1+-	0.5469923134	0.4637639472	0.6451570735	7.8344663e-13
HLA-DRB1-37 z_1++	0.3562031765	0.2887584603	0.4394008155	5.5097667e-22
HLA-DRB1-37 z_2+-	0.8706681444	0.6880179141	1.1018070927	0.2489544371
HLA-DRB1-37 z_2++	0.6358801612	0.4911226524	0.8233046824	0.0005921477
HLA-DRB1-37 z_3+-	1.3549444123	0.2185040061	8.4020169382	0.7442155538
HLA-DRB1-37 z_3++	2.3620608961	0.3850764170	14.488894750	0.3530057013

Table 46: Estimate, 95% confidence interval and p-value for all three univariable models and the multivariable model for the amino acid on position HLA-DRB1-37, using subset 2.

	Estimate	2.5%	97.5%	p-value
HLA-DRB1-30 z_1+-	1.2808512319	1.1246756755	1.4587137554	0.0001907371
HLA-DRB1-30 z_1++	1.2052673839	0.8717572467	1.6663692471	0.2586460988
HLA-DRB1-30 z_2+-	1.1972088907	0.9032783733	1.5867856138	0.2104800649
HLA-DRB1-30 z_2++	1.2540074858	0.9543811937	1.6477009236	0.1042074063
HLA-DRB1-30 z_3+-	1.7431219097	0.8709527518	3.4886783306	0.1164918339
HLA-DRB1-30 z_3++	2.2116175657	1.1209848566	4.3633526607	0.0220573498
HLA-DRB1-30 z_1+-	1.7392816316	1.4264933451	2.1206552449	4.4567701e-08
HLA-DRB1-30 z_1++	2.0557594780	1.3322845027	3.1721055247	0.0011285781
HLA-DRB1-30 z_2+-	1.3766424262	0.9709088358	1.9519282345	0.0727737524
HLA-DRB1-30 z_2++	2.0760562891	1.3651729832	3.1571161813	0.0006368998
HLA-DRB1-30 z_3+-	1.3947966528	0.6699673984	2.9038095098	0.3737876471
HLA-DRB1-30 z_3++	1.3613689150	0.6447782481	2.8743608028	0.4184897139

Table 47: Estimate, 95% confidence interval and p-value for all three univariable models and the multivariable model for the amino acid on position HLA-DRB1-30, using subset 2.

	Estimate	2.5%	97.5%	p-value
HLA-DRB1-13 z_1+-	1.1267454118	0.8653619904	1.4670799470	0.3755408802
HLA-DRB1-13 z_1++	1.0947653320	0.8478712376	1.4135532365	0.4874564077
HLA-DRB1-13 z_2+-	1.3465536560	1.1243415735	1.6126831838	0.0012223653
HLA-DRB1-13 z_2++	2.2574031317	1.8820187390	2.7076610841	1.7155511e-18
HLA-DRB1-13 z_3+-	1.8752599475	1.2127674422	2.8996489749	0.0046920526
HLA-DRB1-13 z_3++	2.3488903674	1.5366115286	3.5905535362	8.0126921e-05
HLA-DRB1-13 z_1+-	1.9126740694	1.4491462465	2.5244671507	4.6526254e-06
HLA-DRB1-13 z_1++	3.0676485089	2.2885150833	4.1120407913	6.4848207e-14
HLA-DRB1-13 z_2+-	1.8539803512	1.5264884814	2.2517321187	4.8103918e-10
HLA-DRB1-13 z_2++	4.4926749316	3.5855359712	5.6293196341	5.8426002e-39
HLA-DRB1-13 z_3+-	3.0420024731	1.9514739641	4.7419433803	9.0253599e-07
HLA-DRB1-13 z_3++	6.1071081600	3.8976341990	9.5690791319	2.8544063e-15

Table 48: Estimate, 95% confidence interval and p-value for all three univariable models and the multivariable model for the amino acid on position HLA-DRB1-13, using subset 2.

	Estimate	2.5%	97.5%	p-value
HLA-DRB1-11 z_1+-	0.5414097937	0.4719273129	0.6211222719	2.0305122e-18
HLA-DRB1-11 z_1++	0.3454630308	0.2914746949	0.4094513444	1.5051416e-34
HLA-DRB1-11 z_2+-	0.8071024048	0.7057841128	0.9229653658	0.0017406369
HLA-DRB1-11 z_2++	0.4957547890	0.3399135560	0.7230450403	0.0002683062
HLA-DRB1-11 z_3+-	0.4367045140	0.3661709105	0.5208246398	3.0244317e-20
HLA-DRB1-11 z_3++	0.2358849384	0.1961291671	0.2836992833	4.2439253e-53
HLA-DRB1-11 z_1+-	1.2229309575	0.8491797187	1.7611821075	0.2794957067
HLA-DRB1-11 z_1++	1.6552204722	0.8346009641	3.2827122533	0.1491771964
HLA-DRB1-11 z_2+-	1.4231690806	1.0152505897	1.9949855263	0.0405800935
HLA-DRB1-11 z_2++	1.5070054101	0.7318282358	3.1032764179	0.2657847931
HLA-DRB1-11 z_3+-	0.3401437190	0.2284608463	0.5064226603	1.0932067e-07
HLA-DRB1-11 z_3++	0.1387240247	0.0680950115	0.2826103499	5.3093254e-08

Table 49: Estimate, 95% confidence interval and p-value for all three univariable models and the multivariable model for the amino acid on position HLA-DRB1-11, using subset 2.

	Estimate	2.5%	97.5%	p-value
HLA-DQB1-57 z_1+-	1.1722959198	0.8459498210	1.6245381102	0.3395989364
HLA-DQB1-57 z_1++	1.0748612895	0.7829540139	1.4755998068	0.6552151222
HLA-DQB1-57 z_2+-	0.8570269671	0.7487122101	0.9810114120	0.0252171556
HLA-DQB1-57 z_2++	0.6624394872	0.5601872404	0.7833560685	1.4768266e-06
HLA-DQB1-57 z_3+-	1.1722959198	0.8459498210	1.6245381102	0.3395989364
HLA-DQB1-57 z_3++	1.0748612895	0.7829540139	1.4755998068	0.6552151222
HLA-DQB1-57 z_1+-	1.3036288384	0.9320986908	1.8232491528	0.1213477176
HLA-DQB1-57 z_1++	1.3666769429	0.9762352505	1.9132743520	0.0687824603
HLA-DQB1-57 z_2+-	0.8220543666	0.7126401136	0.9482673916	0.0071692507
HLA-DQB1-57 z_2++	0.6252764656	0.5188203822	0.7535761352	8.1763736e-07
HLA-DQB1-57 z_3+-	NA	NA	NA	NA
HLA-DQB1-57 z_3++	NA	NA	NA	NA

Table 50: Estimate, 95% confidence interval and p-value for all three univariable models and the multivariable model for the amino acid on position HLA-DQB1-57, using subset 2.

	Estimate	2.5%	97.5%	p-value
HLA-DQB1-55 z_1+-	0.7624258966	0.6662198032	0.8725247208	8.1002947e-05
HLA-DQB1-55 z_1++	0.6035458800	0.5098513541	0.7144584913	4.4579520e-09
HLA-DQB1-55 z_2+-	1.8517191210	1.2630737179	2.7146980055	0.0015968665
HLA-DQB1-55 z_2++	2.4745696410	1.7051298433	3.5912191276	1.8573141e-06
HLA-DQB1-55 z_3+-	1.4744887301	1.2962417117	1.6772466090	3.4830871e-09
HLA-DQB1-55 z_3++	2.5844291609	2.1522309222	3.1034188845	2.7161230e-24
HLA-DQB1-55 z_1+-	0.6567985954	0.5676030102	0.7600107597	1.6512518e-08
HLA-DQB1-55 z_1++	0.4112008108	0.3353990108	0.5041341845	1.2512775e-17
HLA-DQB1-55 z_2+-	2.4402302145	1.6459880244	3.6177198201	8.9734057e-06
HLA-DQB1-55 z_2++	4.1313361363	2.7763670680	6.1475798600	2.6435299e-12
HLA-DQB1-55 z_3+-	0.8937660560	0.7723924564	1.0342122794	0.1314994301
HLA-DQB1-55 z_3++	NA	NA	NA	NA

Table 51: Estimate, 95% confidence interval and p-value for all three univariable models and the multivariable model for the amino acid on position HLA-DQB1-55, using subset 2.

	Estimate	2.5%	97.5%	p-value
HLA-DQB1-37 z_1+-	2.4586085228	0.9901124672	6.1051204473	0.0525560149
HLA-DQB1-37 z_2+-	1.8517191210	1.2630737179	2.7146980055	0.0015968665
HLA-DQB1-37 z_2++	2.4745696410	1.7051298433	3.5912191276	1.8573141e-06
HLA-DQB1-37 z_3+-	1.8517191210	1.2630737179	2.7146980055	0.0015968665
HLA-DQB1-37 z_3++	2.4745696410	1.7051298433	3.5912191276	1.8573141e-06
HLA-DQB1-37 z_1+-	2.3244618017	0.9349487600	5.7790575258	0.0694914999
HLA-DQB1-37 z_2+-	1.8469964071	1.2598384964	2.7078040062	0.0016705153
HLA-DQB1-37 z_2++	2.4618592091	1.6963208951	3.5728798620	2.1279318e-06
HLA-DQB1-37 z_3+-	NA	NA	NA	NA
HLA-DQB1-37 z_3++	NA	NA	NA	NA

Table 52: Estimate, 95% confidence interval and p-value for all three univariable models and the multivariable model for the amino acid on position HLA-DQB1-37, using subset 2.

	Estimate	2.5%	97.5%	p-value
HLA-DQB1-26 z_1+-	1.1915707260	1.0505808036	1.3514817616	0.0063730902
HLA-DQB1-26 z_1++	0.9624312527	0.7198235343	1.2868069353	0.7961029306
HLA-DQB1-26 z_2+-	1.2994433260	1.1371493529	1.4848998975	0.0001190205
HLA-DQB1-26 z_2++	1.5284749854	1.0571665821	2.2099031699	0.0241017572
HLA-DQB1-26 z_3+-	1.3836955787	1.2191444393	1.5704566193	4.9713456e-07
HLA-DQB1-26 z_3++	1.4176175387	1.1780435409	1.7059127411	0.0002200946
HLA-DQB1-26 z_1+-	1.1952836804	1.0265320884	1.3917763437	0.0216082543
HLA-DQB1-26 z_1++	1.1117697418	0.8269702179	1.4946511156	0.4828591393
HLA-DQB1-26 z_2+-	1.2848301275	1.1034165939	1.4960699935	0.0012503618
HLA-DQB1-26 z_2++	1.7218142455	1.1847359138	2.5023672041	0.0043901482
HLA-DQB1-26 z_3+-	1.1220447133	0.9636356410	1.3064941615	0.1380912531
HLA-DQB1-26 z_3++	NA	NA	NA	NA

Table 53: Estimate, 95% confidence interval and p-value for all three univariable models and the multivariable model for the amino acid on position HLA-DQB1-26, using subset 2.

	Estimate	2.5%	97.5%	p-value
HLA-DPA1-228 z_1+-	1.4318070817	0.8909893371	2.3008934380	0.1380586483
HLA-DPA1-228 z_1++	1.7563865025	1.1062807604	2.7885267976	0.0169301799
HLA-DPA1-228 z_2+-	0.8152004582	0.7069809571	0.9399854131	0.0049290177
HLA-DPA1-228 z_2++	0.5693507656	0.3586122969	0.9039296675	0.0169301799
HLA-DPA1-228 z_3+-	0.8152004582	0.7069809571	0.9399854131	0.0049290177
HLA-DPA1-228 z_3++	0.5693507656	0.3586122969	0.9039296675	0.0169301799
HLA-DPA1-228 z_1+-	1.4318070817	0.8909893371	2.3008934380	0.1380586483
HLA-DPA1-228 z_1++	1.7563865025	1.1062807604	2.7885267976	0.0169301799
HLA-DPA1-228 z_2+-	NA	NA	NA	NA
HLA-DPA1-228 z_2++	NA	NA	NA	NA
HLA-DPA1-228 z_3+-	NA	NA	NA	NA
HLA-DPA1-228 z_3++	NA	NA	NA	NA

Table 54: Estimate, 95% confidence interval and p-value for all three univariable models and the multivariable model for the amino acid on position HLA-DPA1-228, using subset 2.

	Estimate	2.5%	97.5%	p-value
HLA-DPB1-84 z_1++	1.9674177441	1.1588021708	3.3402876498	0.0122212038
HLA-DPB1-84 z_2+-	0.8339300062	0.7367937864	0.9438723128	0.0040509562
HLA-DPB1-84 z_2++	0.5177670063	0.4030085007	0.6652035188	2.6224728e-07
HLA-DPB1-84 z_3++	1.9674177441	1.1588021708	3.3402876498	0.0122212038
HLA-DPB1-84 z_1++	2.1063936675	1.2394877763	3.5796192324	0.0058960008
HLA-DPB1-84 z_2+-	0.8281118791	0.7315082527	0.9374730657	0.0028806286
HLA-DPB1-84 z_2++	0.5111982762	0.3978192142	0.6568905378	1.5663449e-07
HLA-DPB1-84 z_3++	NA	NA	NA	NA

Table 55: Estimate, 95% confidence interval and p-value for all three univariable models and the multivariable model for the amino acid on position HLA-DPB1-84, using subset 2.