Vriezinga, S. L. et al. (2015) Nat. Rev. Gastroenterol. Hepatol. doi: 10. 1038/nrgastro.2015.98.

In some centres, genetic typing of subjects at the HLA locus is also used in the diagnosis. This is because >99% of patients with coeliac disease are positive for HLA-DQ2 and/ or HLA-DQ8. Therefore, patients with other HLA types and some features suspicious for coeliac disease are generally assumed either to have non-coeliac gluten-sensitivity or not have coeliac disease at all.

Unfortunately current diagnostic tests are far from perfect, and it is not uncommon for subjects with severe symptoms to be negative for autoantibodies in the blood and to have only minor intestinal changes with normal villi. Furthermore, for the tests to be most efficacious the biopsy should be taken where the subject is on a gluten containing diet, which may mean the subject has severe symptoms, so it can be an unpleasant test. The results may still be inconclusive as the morphological changes may be mild, or the subject may not have been exposed to sufficient gluten. People may have severe symptoms and be investigated for years before a diagnosis is achieved.

Summary

In one aspect, provided herein is a computer implemented method of determining the disease status in a subject, the method comprising the steps of:

ai) obtaining nucleotide or amino acid sequence data for the T-cell and/or B-cell receptor repertoire in a plurality of samples, wherein the samples comprise at least one reference sample obtained from a subject having a known disease status and at least one control sample obtained from a subject having a known negative disease status;

aii) identifying a data set of overlapping short sequences of nucleotides or amino acids of length k, known as k-mers, present within the nucleotide or amino acid sequence data, wherein the data set further provides information on a relative position of k-mers of amino acids or bases within the T-cell receptor or B-cell receptor amino acid or base sequence;

aiii) determining k-mer frequencies of each identified k-mer in the data set from each sample, each k-mer frequency indicating the number of times each individual k-mer sequence is present within the data set of each sample;

aiv) normalizing the k-mer frequencies to give a relative k-mer frequency, by scaling the k-mer frequencies by the total number of k-mers identified in the nucleotide or amino acid sequence data from that sample;

av) generating a matrix containing the relative k-mer frequencies of each k-mer observed in each of the samples;

avi) reducing dimensionality of the matrix to generate a reduced data set of k-mer frequencies;

avii) using a selected subset of the reduced data set of k-mer frequencies to classify the samples;

aviii) optimising one or more parameters of the data set to optimise classification of each reference sample into its known disease status, said optimised parameters comprising one or more of:

1. the set of principal components, or similarly reduced data sets of k-mer frequencies, used in the classification;
2. the length, k, of the k-mers;
3. the degree of scaling of the k-mer frequencies;
4. the relative positions of the k-mers of amino acids or bases within the T-cell receptor or B-cell receptor amino acid or base sequence, and/or whether or not they are within the CDR3 sequence; and
5. annotation of the k-mers, such that the frequencies of differentially annotated k-mers, even if they have the same amino acid or base sequence, are considered separately, such k-mer annotation comprising one or more of:

positional annotation of the k-mers indicating whether they were derived from the beginning, middle or end of the CDR3 region of the neucleotide or amino acid sequence data or at any specifically defined position within the CDR3 region; and/or

relative associations of k-mers with a particular CDR1, CDR2 or TRG CDR4/HV4 sequence in the nucleotide or amino acid sequence data; and

bi) repeating steps aii) to avi) for the plurality of samples of ai) and at least one unknown sample obtained from a subject having an unknown disease status;

bii) determining the one or more optimised parameters for the data set of the at least one unknown sample; and

c) classifying the unknown sample based on the optimised parameters to determine which reference sample or plurality of reference samples of known disease status the unknown sample is most similar to.

In another aspect, provided herein is a system for classifying a sample, said system comprising a microprocessor and memory, wherein the sample comprises a T-cell receptor repertoire and/ or a B-cell receptor repertoire, wherein the processor is configured to undertake the method as described herein.

In another aspect, provided herein is a k-mer dataset produced using the method of claim 1, derived from k-mer analysis of sequence data of TRA, TRB, TRG, TRD, IGH, IGK and/ or IGL from samples of known disease status, known physiological status or known pathophysiological status, for use as a training or reference data set which can be used to classify new or test samples.

In another aspect, provided herein is a computer readable medium storing instructions executable by one or more processors to perform operations according to the method as described herein.

Detailed Description

The present invention has a number of advantages over the current diagnostic protocol, these advantages may include one or more of the following:

- increase the predictive value of a biopsy when used in addition to a biopsy;
- it may reduce or avoid the subj ectivity of the analysis of a biopsy;
- it may reduce or avoid the need for the subject to be exposed to gluten prior to the test; and/or

- it may avoid the need for an endoscopy as the test may be carried out on a blood sample, and, if duodenal lymphocytes are needed, they are very superficial and may be obtained by less invasive means than endoscopy.

According to a first aspect the present invention provides a method of determining the disease status in a subject, the method comprising:

CLAIMS

1. A computer implemented method of determining the disease status in a subject, the method comprising the steps of:

ai) obtaining nucleotide or amino acid sequence data for the T-cell and/or B-cell receptor repertoire in a plurality of samples, wherein the samples comprise at least one reference sample obtained from a subject having a known disease status and at least one control sample obtained from a subject having a known negative disease status;

aii) identifying a data set of overlapping short sequences of nucleotides or amino acids of length k, known as k-mers, present within the nucleotide or amino acid sequence data, wherein the data set further provides information on a relative position of k-mers of amino acids or bases within the T-cell receptor or B-cell receptor amino acid or base sequence;

aiii) determining k-mer frequencies of each identified k-mer in the data set from each sample, each k-mer frequency indicating the number of times each individual k-mer sequence is present within the data set of each sample;

aiv) normalizing the k-mer frequencies to give a relative k-mer frequency, by scaling the k-mer frequencies by the total number of k-mers identified in the nucleotide or amino acid sequence data from that sample;

av) generating a matrix containing the relative k-mer frequencies of each k-mer observed in each of the samples;

avi) reducing dimensionality of the matrix to generate a reduced data set of k-mer frequencies;

avii) using a selected subset of the reduced data set of k-mer frequencies to classify the samples;

aviii) optimising one or more parameters of the data set to optimise classification of each reference sample into its known disease status, said optimised parameters comprising one or more of:

1. the set of principal components, or similarly reduced data sets of k-mer frequencies, used in the classification;

2. the length, k, of the k-mers;

3. the degree of scaling of the k-mer frequencies;

4. the relative positions of the k-mers of amino acids or bases within the T-cell receptor or B-cell receptor amino acid or base sequence, and/or whether or not they are within the CDR3 sequence; and

5. annotation of the k-mers, such that the frequencies of differentially annotated k-mers, even if they have the same amino acid or base sequence, are considered separately, such k-mer annotation comprising one or more of:

positional annotation of the k-mers indicating whether they were derived from the beginning, middle or end of the CDR3 region of the neucleotide or amino acid sequence data or at any specifically defined position within the CDR3 region; and/or

relative associations of k-mers with a particular CDR1, CDR2 or TRG CDR4/HV4 sequence in the nucleotide or amino acid sequence data; and

bi) repeating steps aii) to avi) for the plurality of samples of ai) and at least one unknown sample obtained from a subject having an unknown disease status;

bii) determining the one or more optimised parameters for the data set of the at least one unknown sample; and

c) classifying the unknown sample based on the optimised parameters to determine which reference sample or plurality of reference samples of known disease status the unknown sample is most similar to.

2. The method of claim 1 wherein the reduced data set of k-mer frequencies comprises principal components of the data set.

3. The method of claim 1 wherein reducing dimensionality of the matrix to generate a reduced data set of k-mer frequencies comprises the step of one of:

i) performing principal component analysis on the data set such that the reduced data set comprises the principal components of the data set;

ii) performing contrastive principal component analysis (CPCA) on the data set such that the principal components or the reduced data set comprises contrastive principal components of the data set;

iii) variable selection or feature selection on the data set such that the principal components are a dimensionality reduced data set;

iv) performing any suitable analysis on the data set that mediates dimensionality reduction; and

v) classification of the data in such way that a specific dimensionality reduction approach is not required, such as linear discriminant analysis, quadratic discriminant analysis, generalised discriminant analysis and canonical correlation analysis

4. The method of claim 1 wherein the step of classifying the unknown sample comprises one or more of: hierarchical cluster analysis; non-hierarchical cluster analysis; supervised cluster analysis; unsupervised cluster analysis; any other form of clustering; quadratic discriminant analysis; linear discriminant analysis; generalised discriminant analysis; nearest neighbour analysis; decision trees; support vector machines; logistic regression; neural networks; and any suitable technique for statistically mediated classification.

5. The method claim 1 wherein the step of classifying the unknown sample further includes the step of:

performing cluster analysis of the unknown sample with reference samples having a known disease status of coeliac disease and/or a known negative disease status of coeliac disease; or

performing cluster analysis of the unknown sample with reference samples having a first and second known disease status; or

performing cluster analysis of the unknown sample with reference samples having three or more known different known disease statuses; or

performing cluster analysis of the unknown sample with reference samples having two or more known physiological and/ or or pathophysiological statuses.

6. The method of claim 1 wherein the disease status is the status of any condition mediated or modulated by the immune system.

7. The method of claim 1 wherein the disease status is a condition is selected from the group comprising an autoimmune condition, hypersensitivity, allergy, transplantation, transplant rejection, cancer, all forms of neoplasia, infectious diseases, and vaccination.

8. The method of claim 1 wherein the disease status is coeliac disease status; or gluten sensitivity status.

9. The method of claim 1 wherein the sample is a bodily fluid, such as one or more of: blood or a product derived from blood; lymph or a product derived from lymph; pericardial, pleural or ascitic (peritoneal) fluid(s); joint aspirate fluid; or urine; or wherein the sample is a biopsy, a fine needle aspirate sample or a buccal scrape.

10. The method of claim 1 wherein the sequence data of step a) is one or more of

i) a library of sequences representing the T-cell and/or B-cell receptor repertoire in the sample provided by the subject; or

ii) a library of DNA sequences representing the T-cell receptor repertoire in the sample provided by the subject; or

iii) a library of RNA sequences representing the T-cell receptor repertoire in the sample provided by the subject; or

iv) a library of protein sequences representing the T-cell receptor repertoire in the sample provided by the subject; or

v) a library of DNA sequences representing the B-cell receptor repertoire in the sample provided by the subject; or

vi) a library of RNA sequences representing the B-cell receptor repertoire in the sample provided by the subject; or

vii) a library of protein sequences representing the B-cell receptor repertoire in the sample provided by the subject.

11. The method of claim 1 wherein k-mers length ks of between 3 and 10 amino acids are used.

12. The method of claim 1 wherein the output of step d) is either a numerical assessment based on statistical classification, optionally including a ratio, percentage, proportion or relative likelihood; or a yes/no based on the outcome of statistical classification.

13. The method of claim 1 for use in the diagnosis of coeliac disease and/or gluten sensitivity.

14. The method of claim 1 for use in the diagnosis of Crohn's disease, by comparing data from a test sample of unknown disease status with data from known samples from Crohn's disease and normal; or for use in the diagnosis of ulcerative colitis, by comparing data from a test sample of unknown disease status with data from known samples from ulcerative colitis and normal; or for distinguishing between Crohn's disease and ulcerative colitis, optionally in order to avoid a diagnosis such as indeterminate colitis, by comparing data from a test sample of unknown disease status with data from known samples from Crohn's disease and ulcerative colitis; or for use in the determination of prognosis in melanoma patients, by comparing data from a test sample from a melanoma patient of unknown prognosis or outcome with data from melanoma patient samples with known prognosis or outcome; or for use in the diagnosis of autoimmune conditions, including but not limited to such as multiple sclerosis, pre- or early type I insulin-dependent diabetes mellitus, polymyositis, dermatomyositis, systemic lupus erythemtososus (SLE), rheumatoid arthritis, HLA-B27-associated arthritides (e.g., ankylosing spondylitis), autoimmune hepatitis, primary biliary cirrhosis and primary sclerosing cholangitis, by comparing data from a test sample of unknown disease status with data from

known samples from one or more known autoimmune condition(s) with or without normal samples; or for use in the prediction of prognosis of autoimmune conditions (including but not limited to such as multiple sclerosis, pre- or early type I insulin-dependent diabetes mellitus, polymyositis, dermatomyositis, systemic lupus erythemtososus (SLE), rheumatoid arthritis, HLA-B27-associated arthritides (e.g., ankylosing spondylitis), autoimmune hepatitis, primary biliary cirrhosis and primary sclerosing cholangitis), by comparing data from a test sample of unknown disease status with data from samples with known severity or outcome in autoimmune conditionsor for use in the diagnosis of hypersensitivity conditions, by comparing data from a test sample of unknown disease status with data from known samples from a known hypersensitivity condition and normal or any other suitable comparator condition or physiological/ pathophysiological status; or for use in the prediction of prognosis of hypersensitivity conditions, by comparing data from a test sample of unknown prognosis in a hypersensitivity condition with data from samples in a known hypersensitivity condition with unknown severity, outcome status or precipitating antigen with data from hypersensitivity condition samples with known severity, outcome status or precipitating antigen; or for use in the diagnosis of allergic conditions, by comparing data from a test sample of unknown disease status with data from known samples from a known allergic condition and normal or any other suitable comparator condition and/or physiological/ pathophysiological status; or for use in the prediction of prognosis of allergic conditions, by comparing data from a test sample of unknown prognosis in a hypersensitivity condition with data from samples in a known allergic condition with unknown severity, outcome status or precipitating antigen with data from allergic condition samples with known severity, outcome status or precipitating antigen; or for use in the diagnosis of transplant rejection of any organ or tissue, by comparing data from a test sample of unknown disease status with data from known samples with a known rejection status and normal or any other suitable comparator condition or physiological/ pathophysiological status; or for use in the prediction of prognosis or outcome in transplant rejection of any organ or tissue, by

comparing data from a test sample of unknown disease status with data from samples with a known transplant rejection status, prognosis or outcome; or for use in the prediction of prognosis or outcome in cancer or any form of neoplasia, by comparing data from a test sample of unknown cancer or neoplasia outcome status with data from known cancer or neoplasia samples with a known prognosis or outcome status; or for use in the diagnosis of an infectious disease, by comparing data from a test sample of unknown infectious disease status with data from samples from individuals with a particular infectious disease status and normal and/ or other infectious diseases; or for use in the prediction of prognosis or outcome of an infectious disease (of any organ or tissue), by comparing data from a test sample of unknown infectious disease prognosis or outcome status with data from samples from individuals with that particular infectious disease with known prognosis or outcome status; or for use in the prediction of prognosis or outcome of a vaccination, by comparing data from a test sample of unknown vaccination prognosis or outcome status with data from samples from individuals with particular prognosis or outcome statuses following vaccination; or for use in the determination of the specific response to one or more antigens following vaccination, by comparing data from a test sample of unknown vaccination response with data from samples from individuals with a known specific immune response to one or more specific antigens; or for use in the determination of the specific response to one or more antigens in the setting of infection, by comparing data from a sample of unknown infection status with data from samples from individuals with a known specific immune response to one or more specific antigens; or for use in the diagnosis or prediction of prognosis or outcome status of any particular physiological or pathophysiological state that is, at least in part, determined, mediated by or modulated by the immune system, by comparing data from a test sample of unknown diagnosis, prognosis or outcome status with data from samples with a known diagnosis, prognosis or outcome status; or for use in the determination of biological similarity with respect to the immune system of any group(s) of samples for the purpose of diagnosis; or for use in the determination of biological similarity with respect to the immune system of any group(s) of

samples for the purpose of prediction of prognosis or outcome; or for use in the determination of physiological or pathophysiological state with respect to the immune system of any group(s) of samples for any purpose that involves comparison of the samples with samples from subjects with known physiological or pathophysiological states; or for use in the determination of biological similarity with respect to the immune system of any group(s) of samples for any purpose that involves comparison of the similarity or differences between the samples; or for use in the determination of whether there is a specific response to one or more antigens in a sample, by comparing data from that sample with data from samples from subjects with a known specific immune response to that or those specific antigen(s).

15. The method of claim 1, wherein the sequence data used in the k-mer analysis comprises sequences of any one or two or more of the following genes: TRA, TRB, TRG and TRD (considered components of the T-cell receptor repertoire) IGH, IGK and IGL (considered components of the B-cell receptor repertoire).

16. The method of claim 1 wherein the sequence data used in the k-mer analysis comprises sequences of TRA

wherein the sequence data used in the k-mer analysis comprises sequences of TRB; and/or

wherein the sequence data used in the k-mer analysis comprises sequences of TRG; and/or

wherein the sequence data used in the k-mer analysis comprises sequences of TRD; and/or

wherein the sequence data used in the k-mer analysis comprises sequences of IGH; and/or

wherein the sequence data used in the k-mer analysis comprises sequences of IGK; and/or

wherein the sequence data used in the k-mer analysis comprises sequences of IGL; and/or

wherein the sequence data used in the k-mer analysis comprises sequences of TRA and TRB, TRG and/ or TRD; and/or

wherein the sequence data used in the k-mer analysis comprises sequences of TRA and TRB, combined with a 1:1 weighting or any other weighting between 10,000:1 and 1:10,000; and/or

wherein the sequence data used in the k-mer analysis comprises sequences of TRG and TRD, combined with a 1:1 weighting or any other weighting between 10,000:1 and 1:10,000; and/or

wherein the sequence data used in the k-mer analysis comprises sequences of TRA and TRG, combined with a 1:1 weighting or any other weighting between 10,000:1 and 1:10,000; and/or

wherein the sequence data used in the k-mer analysis comprises sequences of TRA and TRD, combined with a 1:1 weighting or any other weighting between 10,000:1 and 1:10,000; and/or

wherein the sequence data used in the k-mer analysis comprises sequences of TRB and TRG, combined with a 1:1 weighting or any other weighting between 10,000:1 and 1:10,000; and/or

wherein the sequence data used in the k-mer analysis comprises sequences of TRB and TRD, combined with a 1:1 weighting or any other weighting between 10,000:1 and 1:10,000; and/or

wherein the sequence data used in the k-mer analysis comprises sequences of any combination of two, three or four of TRA, TRB, TRG and/ or TRD, combined using a 1:1 weighting or any other relative weightings; and/or

wherein the sequence data used in the k-mer analysis comprises sequences of any two of IGH, IGK and/ or IGL, combined with a 1:1 weighting or any other weighting between 10,000:1 and 1:10,000; and/or

wherein the sequence data used in the k-mer analysis comprises sequences of all three of IGH, IGK and/ or IGL, combined with a 1:1 weighting or any other weighting between 10,000:1 and 1:10,000; and/or

wherein the sequence data used in the k-mer analysis comprises sequences of any combination of two or more of TRA, TRB, TRG, TRD, IGH, IGK and/ or IGL, combined with a 1:1 weighting or any other relative weightings.

17. A system for classifying a sample, said system comprising a microprocessor and memory, wherein the sample comprises a T-cell receptor repertoire and/ or a B-cell receptor repertoire, wherein the processor is configured to undertake the method of claim 1.

18. A k-mer dataset produced using the method of claim 1, derived from k-mer analysis of sequence data of TRA, TRB, TRG, TRD, IGH, IGK and/ or IGL from samples of known disease status, known physiological status or known pathophysiological status, for use as a training or reference data set which can be used to classify new or test samples.

19. A computer readable medium storing instructions executable by one or more processors to perform operations according to the method of claim 1.