



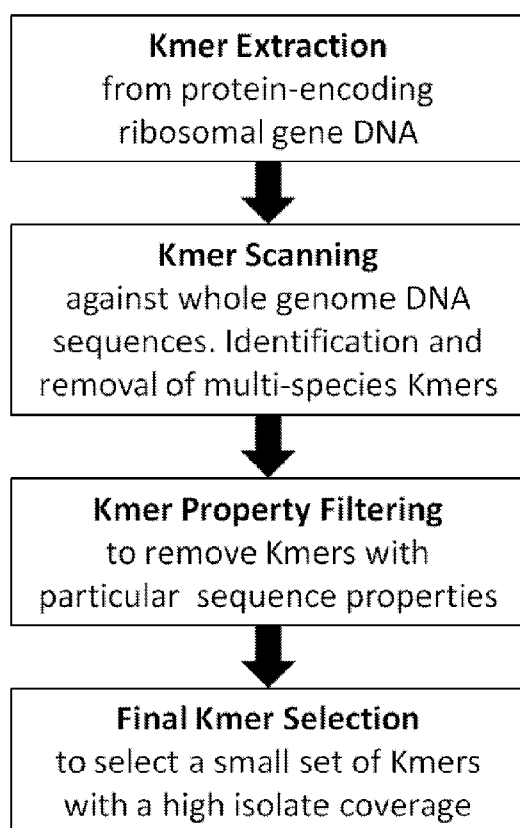
- (51) International Patent Classification:
G06F 19/18 (2011.01) *G06F 19/22* (2011.01)
- (21) International Application Number:
PCT/GB2016/051802
- (22) International Filing Date:
16 June 2016 (16.06.2016)
- (25) Filing Language: English
- (26) Publication Language: English
- (30) Priority Data:
1510649.5 17 June 2015 (17.06.2015) GB
- (71) Applicant: **ISIS INNOVATION LIMITED** [GB/GB];
Buxton Court, 3 West Way, Oxford Oxfordshire OX2 0SZ
(GB).

- (72) Inventors: **BRAY, James**; Department of Zoology, University of Oxford, The Tinbergen Building, South Parks Road, Oxford Oxfordshire OX1 3PS (GB). **JOLLEY, Keith**; Department of Zoology, University of Oxford, The Tinbergen Building, South Parks Road, Oxford Oxfordshire OX1 3PS (GB). **MAIDEN, Martin**; Department of Zoology, University of Oxford, The Tinbergen Building, South Parks Road, Oxford Oxfordshire OX1 3PS (GB).
- (74) Agent: **JA KEMP**; NICHOLLS, James Ronald, 14 South Square, Gray's Inn, London Greater London WC1R 5JJ (GB).
- (81) Designated States (*unless otherwise indicated, for every kind of national protection available*): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM,

[Continued on next page]

(54) Title: METHOD

Figure 1



(57) Abstract: The invention relates to identifying genomic sequences which are characteristic of an organism. The invention also relates to using the genomic sequences to distinguish the organism from other organisms.



WO 2016/203246 A1



DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE,

DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

- with international search report (Art. 21(3))
- before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))
- with sequence listing part of description (Rule 5.2(a))

METHOD

Field of the Invention

The invention relates to identifying genomic sequences which are characteristic of an organism. The invention also relates to using the genomic sequences to distinguish the organism from other organisms.

Background of the Invention

It is important to be able to rapidly identify organisms and rapidly differentiate them from other organisms. For instance, infections with microorganisms, occurring primarily in hospitalised patients, can delay recovery and in some instances can be life threatening. It has been established that a decrease in the time before diagnosis of microorganism infection improves patient outcome. The ability to rapidly identify a microorganism is therefore important for improved patient treatment and survival.

Summary of the Invention

The inventors have surprisingly shown that it is possible to identify in the genome of an organism one or more polynucleotide sequences (one or more k-mers) which are capable of distinguishing the organism from one or more different organisms. In particular, the inventors have surprisingly shown that it is possible to identify one or more k-mers which are present in all known versions of the genome of an organism, but which are not present in the genomes of one or more different organisms or all different organisms. Such k-mers allow the organism to be distinguished from the one or more different organisms. Most surprisingly of all, the inventors have shown that it is possible to identify one k-mer in all known versions of the genome of an organism which is capable of distinguishing the organism from one or more different organisms. This makes it possible to rapidly and efficiently identify the organism.

The invention provides a method for identifying one or more nucleotide k-mers in the genome of an organism which are capable of distinguishing the organism from one or more different organisms, comprising:

- (a) extracting all nucleotide k-mers from the genome of the organism; and
- (b) comparing the nucleotide k-mers extracted in step (a) with the genome(s) of the one or more different organisms and identifying those nucleotide k-mers which do not appear in the genome(s) of the one or more different organisms.

The invention also provides:

- a computer program configured to carry out a method of the invention;

- a computer program medium comprising a computer program of the invention;
- a computer for carrying out a method of the invention or programmed with a computer program of the invention;
- one or more nucleotide k-mers identified using a method of the invention or a computer program of the invention;
- and oligonucleotide probe which comprises a sequence which is complementary to a nucleotide k-mer of the invention;
- a plurality of oligonucleotide probes each of which comprises a sequence which is complementary to one of the nucleotide k-mers in a population of two or more nucleotide k-mers of the invention;
- a support having attached thereto an oligonucleotide probe of the invention or a plurality of oligonucleotide probes of the invention;
- a method for detecting the presence or absence of an organism in a sample, comprising detecting the presence or absence in the sample of one or more nucleotide k-mers of the invention, wherein the presence of the one or more nucleotide k-mers is indicative of the presence of the organism in the sample and the wherein the absence of the one or more nucleotide k-mers is indicative of the absence of the organism from the sample;
- a nucleotide k-mer capable of distinguishing an organism from one or more different organisms, wherein the k-mer comprises about 10 or more consecutive nucleotides from any one of the sequences shown in SEQ ID NOs: 1 to 213;
- an oligonucleotide probe which comprises a sequence which is complementary to a nucleotide k-mer of the invention and derived from any one of SEQ ID NOs: 1 to 213;
- a support having attached thereto a oligonucleotide probe of the invention and derived from any one of SEQ ID NOs: 1 to 213; and
- a method for detecting the presence or absence of an organism in a sample, comprising detecting the presence or absence in the sample of a nucleotide k-mer of the invention and derived from any one of SEQ ID NOs: 1 to 213, wherein the presence of the nucleotide k-mer is indicative of the presence of the organism in the sample and the wherein the absence of the nucleotide k-mer is indicative of the absence of the organism from the sample.

Description of the Figures

Figure 1 shows a flowchart describing the four computational stages required to generate a small single species k-mer library from protein-encoding ribosomal gene DNA.

Description of the Sequences

SEQ ID NOS: 1 to 5 show k-mer cluster sequences from *Aggregatibacter actinomycetemcomitans*.

SEQ ID NOS: 6 to 10 show k-mer cluster sequences *Bacillus anthracis*.

5 SEQ ID NOS: 11 to 15 show k-mer cluster sequences from *Bacillus licheniformis*.

SEQ ID NOS: 16 to 20 show k-mer cluster sequences from *Bacteroides fragilis*.

SEQ ID NOS: 21 to 25 to 5 show k-mer cluster sequences from *Bartonella henselae*.

SEQ ID NOS: 26 to 30 show k-mer cluster sequences from *Bordetella pertussis*.

SEQ ID NOS: 31 to 35 show k-mer cluster sequences from *Borrelia burgdorferi*.

10 SEQ ID NOS: 36 to 40 show k-mer cluster sequences from *Brucella abortus*.

SEQ ID NOS: 41 to 44 show k-mer cluster sequences from *Campylobacter jejuni*.

SEQ ID NOS: 45 to 49 show k-mer cluster sequences from *Chlamydia trachomatis*.

SEQ ID NOS: 50 to 54 show k-mer cluster sequences from *Chlamydophila pneumonia*.

SEQ ID NOS: 55 to 59 show k-mer cluster sequences from *Clostridium difficile*.

15 SEQ ID NOS: 60 to 64 show k-mer cluster sequences from *Clostridium perfringens*.

SEQ ID NOS: 65 to 69 show k-mer cluster sequences from from *Enterobacter aerogenes*.

SEQ ID NOS: 70 to 74 show k-mer cluster sequences from from *Enterococcus faecalis*.

SEQ ID NOS: 75 to 79 show k-mer cluster sequences from from *Enterococcus faecium*.

SEQ ID NOS: 80 to 84 show k-mer cluster sequences from *Francisella tularensis*.

20 SEQ ID NO: 85 shows a k-mer cluster sequence from *Haemophilus influenza*.

SEQ ID NOS: 86 and 87 show k-mer cluster sequences from *Helicobacter pylori*.

SEQ ID NOS: 88 to 92 show k-mer cluster sequences from *Klebsiella oxytoca*

SEQ ID NOS: 93 to 97 show k-mer cluster sequences from *Legionella pneumophila*.

SEQ ID NOS: 98 to 102 show k-mer cluster sequences from *Listeria monocytogenes*.

25 SEQ ID NOS: 103 to 107 show k-mer cluster sequences from *Moraxella catarrhalis*.

SEQ ID NOS: 108 to 112 show k-mer cluster sequences from *Mycobacterium avium*.

SEQ ID NOS: 113 to 116 show k-mer cluster sequences from *Mycobacterium bovis*.

SEQ ID NOS: 117 to 121 show k-mer cluster sequences from *Mycoplasma genitalium*.

SEQ ID NOS: 122 to 126 show k-mer cluster sequences from *Mycoplasma pneumonia*.

30 SEQ ID NOS: 127 to 131 show k-mer cluster sequences from *Neisseria gonorrhoeae*.

SEQ ID NOS: 132 to 136 show k-mer cluster sequences from *Neisseria meningitides*.

SEQ ID NOS: 137 to 141 show k-mer cluster sequences from *Porphyromonas gingivalis*.

SEQ ID NOS: 142 to 146 show k-mer cluster sequences from *Proteus mirabilis*.

SEQ ID NOS: 147 to 151 show k-mer cluster sequences from *Pseudomonas aeruginosa*.

35 SEQ ID NOS: 152 to 156 show k-mer cluster sequences from *Salmonella enterica*.

SEQ ID NOs: 157 to 161 show k-mer cluster sequences from *Serratia marcescens*.

SEQ ID NOs: 162 to 166 show k-mer cluster sequences from *Staphylococcus aureus*.

SEQ ID NOs: 167 to 171 show k-mer cluster sequences from *Staphylococcus epidermidis*.

5 SEQ ID NOs: 172 to 176 show k-mer cluster sequences from *Staphylococcus haemolyticus*.

SEQ ID NO: 177 shows a k-mer cluster sequence from *Stenotrophomonas maltophilia*.

SEQ ID NOs: 178 to 182 show k-mer cluster sequences from *Streptococcus mutans*.

SEQ ID NOs: 183 to 187 show k-mer cluster sequences from *Streptococcus pyogenes*.

10 SEQ ID NOs: 188 to 192 show k-mer cluster sequences from *Streptococcus salivarius*.

SEQ ID NOs: 193 to 197 show k-mer cluster sequences from *Streptococcus sanguinis*.

SEQ ID NOs: 198 to 202 show k-mer cluster sequences from *Treponema pallidum*.

SEQ ID NO: 203 shows a k-mer cluster sequence from *Vibrio cholera*.

SEQ ID NOs: 204 to 208 show k-mer cluster sequences from *Vibrio parahaemolyticus*.

15 SEQ ID NOs: 209 to 213 show k-mer cluster sequences from *Yersinia enterocolitica*.

Detailed Description of the Invention

It is to be understood that different applications of the disclosed methods may be tailored to the specific needs in the art. It is also to be understood that the terminology used herein is for the purpose of describing particular embodiments of the invention only, and is not intended to be limiting.

In addition, as used in this specification and the appended claims, the singular forms "a", "an", and "the" include plural referents unless the content clearly dictates otherwise. Thus, for example, reference to "an organism" includes two or more organisms, reference to "a k-mer" includes two or more k-mers, reference to "a probe" includes two or more such probes and the like.

All publications, patents and patent applications cited herein are hereby incorporated by reference in their entirety.

Methods of identifying k-mers

The invention concerns identifying one or more nucleotide k-mers in the genome of an organism. A nucleotide k-mer is a nucleotide sequence containing a whole number value, k, of nucleotides. Suitable lengths of the k-mers are discussed in more detail below.

The one or more k-mers in the genome of the organism are capable of distinguishing the organism from one or more different organisms. The one or more k-mers in the genome of the

organism are characteristic of the organism. The one or more k-mers in the genome of the organism are specific for the organism.

The one or more k-mers in the genome of the organism are capable of distinguishing the organism from any number of different organisms. The one or more k-mers in the genome of the organism are preferably capable of distinguishing the organism from at least about 2, at least about 3, at least about 4, at least about 5, at least about 10, at least about 20, at least about 30, at least about 50, at least about 100, at least about 150, at least about 200, at least about 250, at least about 500, at least about 1000, at least about 5000 or at least about 10,000 different organisms or more. The one or more k-mers in the genome of the organism are preferably capable of distinguishing the organism from all different organisms.

An organism is typically different from the organism in which the one or more k-mers are identified (*i.e.* the organism of interest) if it belongs to a different species.

Organisms may be distinguished at any level of taxonomy. The one or more k-mers in the genome of an organism in a particular kingdom are preferably capable of distinguishing the organism from one or more different organisms in a different kingdom or different kingdoms, such as any number of organisms as described above. The one or more k-mers in the genome of the organism are more preferably capable of distinguishing the organism from all organisms in different kingdoms. For instance, the one or more k-mers in the genome of a bacterium are preferably capable of distinguishing the bacterium from one or more different fungi, such as any number of fungi as described above, and *vice versa*. The one or more k-mers in the genome of a bacterium are more preferably capable of distinguishing the bacterium from all different fungi and *vice versa*.

The one or more k-mers in the genome of an organism in a particular kingdom are preferably capable of distinguishing the organism from one or more different organisms in the same kingdom, such as any number of organisms as described above. The one or more k-mers in the genome of the organism are more preferably capable of distinguishing the organism from all organisms in the same kingdom. For instance, the one or more k-mers in the genome of a bacterium are preferably capable of distinguishing the bacterium from one or more different bacteria, such as any number of bacteria as described above. The one or more k-mers in the genome of a bacterium are more preferably capable of distinguishing the bacterium from all different bacteria. The one or more k-mers in the genome of a fungus are preferably capable of distinguishing the fungus from one or more different fungi, such as any number of fungi as described above. The one or more k-mers in the genome of a fungus are more preferably capable of distinguishing the fungus from all different fungi.

The one or more k-mers in the genome of an organism in a particular phylum are preferably capable of distinguishing the organism from one or more organisms in a different phylum or different phyla, such as any number of organisms as described above. The one or more k-mers in the genome of the organism are more preferably capable of distinguishing the organism from all organisms in different phyla.

Organisms within the same phylum may be distinguished in accordance with the invention. The one or more k-mers in the genome of an organism in a particular phylum are preferably capable of distinguishing the organism from one or more different organisms in the same phylum, such as any number of organisms as described above. The one or more k-mers in the genome of the organism are more preferably capable of distinguishing the organism from all different organisms in the same phylum.

The one or more k-mers in the genome of an organism in a particular family are preferably capable of distinguishing the organism from one or more organisms in a different family or different families, such as any number of organisms as described above. The one or more k-mers in the genome of the organism are more preferably capable of distinguishing the organism from all organisms in different families.

Organisms within the same family may be distinguished in accordance with the invention. The one or more k-mers in the genome of an organism in a particular family are preferably capable of distinguishing the organism from one or more different organisms in the same family, such as any number of organisms as described above. The one or more k-mers in the genome of the organism are more preferably capable of distinguishing the organism from all different organisms in the same family.

The one or more k-mers in the genome of an organism in a particular genus are preferably capable of distinguishing the organism from one or more organisms in a different genus or different genera, such as any number of organisms as described above. The one or more k-mers in the genome of the organism are more preferably capable of distinguishing the organism from all organisms in different genera.

Organisms within the same genus may be distinguished in accordance with the invention. The one or more k-mers in the genome of the organism are preferably capable of distinguishing the organism from one or more different species in the same genus, such as any number of different species as discussed above. The one or more k-mers in the genome of the organism are more preferably capable of distinguishing the organism from all of the different species in the same genus.

The one or more k-mers may be identified in the genome of any organism.

The organism may be eukaryotic. The organism may be an animal or a plant. The organism may be human or another mammalian animal, such as a commercially farmed animal, such as a horse, a cow, a sheep, a fish, a chicken or a pig, a laboratory animal, such as a mouse or a rat, or a pet, such as a guinea pig, a hamster, a rabbit, a cat or a dog. The organism may be a plant, such a cereal, legume, fruit or vegetable, for example wheat, barley, oats, canola, maize, soya, rice, rhubarb, bananas, apples, tomatoes, potatoes, grapes, tobacco, beans, lentils, sugar cane, cocoa or cotton. The organism may be a fungus.

The organism may be prokaryotic, such a bacterium or archaeon.

Preferably, the organism is a microorganism, a fungus or a virus and the one or more different organisms are one or more different microorganisms, fungi or viruses.

The organism is preferably a bacterium. The bacterium may be Gram negative or Gram positive. The Gram positive bacterium is preferably from the genus *Bacillus*, *Clostridium*, *Enterococcus*, *Mycobacterium*, *Staphylococcus* or *Streptococcus*. The Gram positive bacterium may be from the genus *Pasteurella* or *Nocardia*.

The Gram negative bacterium is preferably from the genus *Aggregatibacter*, *Bacteroides*, *Bartonella*, *Brucella*, *Campylobacter*, *Chlamydia*, *Enterobacter*, *Francisella*, *Haemophilus*, *Heliobacter*, *Klebsiella*, *Legionella*, *Moraxella*, *Neisseria*, *Porphyromonas*, *Pseudomonas*, *Salmonella*, *Serratia*, *Stenotrophomonas*, *Vibrio* or *Yersinia*. The Gram negative bacterium may be from the genus *Escherichia* or *Pseudomonas*.

The bacterium may be from the genus *Borrelia*, *Chlamydia*, *Listeria*, *Mycoplasma*, *Proteus* or *Treponema*,

The bacterium is preferably *Aggregatibacter actinomycetemcomitans*, *Bacillus anthracis*, *Bacillus licheniformis*, *Bacteroides fragilis*, *Bartonella henselae*, *Bordetella pertussis*, *Borrelia burgdorferi*, *Brucella abortus*, *Campylobacter jejuni*, *Chlamydia trachomatis*, *Chlamydia pneumoniae*, *Clostridium difficile*, *Clostridium perfringens*, *Enterobacter aerogenes*, *Enterococcus faecalis*, *Enterococcus faecium*, *Francisella tularensis*, *Haemophilus influenzae*, *Helicobacter pylori*, *Klebsiella oxytoca*, *Legionella pneumophila*, *Listeria monocytogenes*, *Moraxella catarrhalis*, *Mycobacterium avium*, *Mycobacterium bovis*, *Mycoplasma genitalium*, *Mycoplasma pneumoniae*, *Neisseria gonorrhoeae*, *Neisseria meningitidis*, *Porphyromonas gingivalis*, *Proteus mirabilis*, *Pseudomonas aeruginosa*, *Salmonella enterica*, *Serratia marcescens*, *Staphylococcus aureus*, *Staphylococcus epidermidis*, *Staphylococcus haemolyticus*, *Stenotrophomonas maltophilia*, *Streptococcus mutans*, *Streptococcus pyogenes*, *Streptococcus salivarius*, *Streptococcus sanguinis*, *Treponema pallidum*, *Vibrio cholera*, *Vibrio parahaemolyticus* or *Yersinia enterocolitica*. Specific and preferred k-mers present in these

bacteria are discussed below.

Other specific examples of bacteria include, but are not limited, to *Mycobacterium tuberculosis*, *Mycobacterium intracellulare*, *Mycobacterium kansasii*, *Mycobacterium gordonae*, *Streptococcus agalactiae*, *Streptococcus viridans* group, *Streptococcus faecalis*, *Streptococcus bovis*, *Streptococcus pneumoniae*, *Corynebacterium diphtheriae*, *Erysipelothrix rhusiopathie*,
 5 *Clostridium tetani*, *Klebsiella pneumoniae*, *Pasteurella multocida*, *Fusobacterium nucleatum*, *Streptobacillus moniliformis*, *Treponema pertenue* and *Actinomyces israelii*.

Preferably, the organism is a bacterium and the one or more different organisms are one or more different bacteria. There may be any number of different bacteria as discussed above. Preferably, the organism is a bacterium and the one or more different organisms are all different
 10 bacteria.

Preferably, the organism is a bacterium and the one or more different organisms are one or more bacteria from one or more different genera of bacteria, such as all different genera of bacterium. For instance, (a) the organism may be a bacterium from the genus *Bacillus* and (b) the one or more different organisms may be one or more bacteria from one or more of, or all of,
 15 *Aggregatibacter*, *Bacteroides*, *Bartonella*, *Bordetella*, *Borrelia*, *Brucella*, *Campylobacter*, *Chlamydia*, *Chlamydophila*, *Clostridium*, *Enterobacter*, *Enterococcus*, *Francisella*, *Haemophilus*, *Helicobacter*, *Klebsiella*, *Legionella*, *Listeria*, *Moraxella*, *Mycobacterium*, *Mycoplasma*, *Neisseria*, *Porphyromonas*, *Proteus*, *Pseudomonas*, *Salmonella*, *Serratia*, *Staphylococcus*, *Stenotrophomonas*, *Streptococcus*, *Treponema*, *Vibrio* and *Yersinia*.

Alternatively, the organism in (a) may be a bacterium from one of the genera in list (b) and the one or more different organisms may be one or more bacteria from one or more of, or all of, *Bacillus* and the remaining genera in list (b).

More preferably, (i) the organism is a bacterium and (ii) the one or more different organisms are one or more different species of bacteria. (i) is preferably selected from the list
 25 *Aggregatibacter actinomycetemcomitans*, *Bacillus anthracis*, *Bacillus licheniformis*, *Bacteroides fragilis*, *Bartonella henselae*, *Bordetella pertussis*, *Borrelia burgdorferi*, *Brucella abortus*, *Campylobacter jejuni*, *Chlamydia trachomatis*, *Chlamydophila pneumoniae*, *Clostridium difficile*, *Clostridium perfringens*, *Enterobacter aerogenes*, *Enterococcus faecalis*, *Enterococcus faecium*, *Francisella tularensis*, *Haemophilus influenzae*, *Helicobacter pylori*, *Klebsiella oxytoca*, *Legionella pneumophila*, *Listeria monocytogenes*, *Moraxella catarrhalis*,
 30 *Mycobacterium avium*, *Mycobacterium bovis*, *Mycoplasma genitalium*, *Mycoplasma pneumoniae*, *Neisseria gonorrhoeae*, *Neisseria meningitidis*, *Porphyromonas gingivalis*, *Proteus mirabilis*, *Pseudomonas aeruginosa*, *Salmonella enterica*, *Serratia marcescens*, *Staphylococcus aureus*, *Staphylococcus epidermidis*, *Staphylococcus haemolyticus*, *Stenotrophomonas maltophilia*, *Streptococcus mutans*, *Streptococcus pyogenes*, *Streptococcus salivarius*,
 35

Streptococcus sanguinis, *Treponema pallidum*, *Vibrio cholera*, *Vibrio parahaemolyticus* or *Yersinia enterocolitica* and the one or more species in (ii) are preferably the species remaining in the list.

The organism is preferably a fungus. Preferably, the organism is a fungus and the one or more different organisms are one or more different fungi. The fungus is preferably from the genus *Absidia*, *Acremonium*, *Aspergillus*, *Aureobasidium*, *Basidiobolus*, *Blastomyces*, *Blastoschizomyces*, *Candida*, *Cladosporium*, *Coccidioides*, *Cryptococcus*, *Cunninghamella*, *Curvularia*, *Debaryomyces*, *Exophiala*, *Exserohilum*, *Fonsecea*, *Fusarium*, *Geotrichum*, *Histoplasma*, *Issatchenkia*, *Khuyveromyces*, *Malezzesia*, *Mucor*, *Paracoccidioides*, *Paecilomyces*, *Penicillium*, *Pichia*, *Pneumocystis*, *Rhizomucor*, *Rhizopus*, *Rhodotorula*, *Saccharomyces*, *Scedosporium*, *Schizophyllum*, *Scopulariopsis*, *Sporothrix*, *Trichoderma*, *Trichophyton* or *Trichosporon*.

The organism is preferably *Aspergillus fumigatus*, *Aspergillus flavus*, *Aspergillus lentulus*, *Aspergillus terreus*, *Aspergillus nidulans*, *Aspergillus oryzae*, *Aspergillus niger*, *Candida albicans*, *Candida caribbica* (*Candida fermentati*), *Candida dubliniensis*, *Candida famata* (*Debaryomyces hansenii*), *Candida fukuyamaensis* (*Candida xestobii* or *Candida carpophila*), *Candida guilliermondii*, *Candida kefyr* (*Khuyveromyces marxianus*), *Candida krusei* (*Issatchenkia orientalis*), *Candida metapsilosis*, *Candida orthopsilosis*, *Candida parapsilosis*, *Candida parapsilosis*, *Candida pelliculosa*, *Candida psychrophila*, *Candida rugosa*, *Candida smithsonii*, *Candida tropicalis*, *Candida utilis*, *Coccidioides immitis*, *Cryptococcus bacillisporus*, *Cryptococcus gattii*, *Cryptococcus grubii*, *Cryptococcus neoformans*, *Debaryomyces coudertii*, *Debaryomyces maramus*, *Debaryomyces nepalensis*, *Debaryomyces prosopidis*, *Debaryomyces robertsiae*, *Debaryomyces udonii*, *Histoplasma capsulatum*, *Khuyveromyces lactis*, *Pichia cecembensis*, *Rhodotorula araucariae*, *Rhodotorula babjevae*, *Rhodotorula dairensis*, *Rhodotorula diobovatum*, *Rhodotorula glutinis*, *Rhodotorula kratochvilovae*, *Rhodotorula paludigenum*, *Rhodotorula sphaerocarpum*, *Rhodotorula toruloides*, *Rhodotorula mueliaginata*, *Saccharomyces 'sensu stricto'*, *Saccharomyces bayanus*, *Saccharomyces boulardii*, *Saccharomyces cariocanus*, *Saccharomyces kudavzevii*, *Saccharomyces mikatae*, *Saccharomyces paradixus*, *Saccharomyces pastorianus*, *Saccharomyces uvarum*, *Saccharomyces cerevisiae* and *Tsuchiyaea wingfieldii*.

Preferably, the organism is a fungus and the one or more different organisms are one or more different fungi. There may be any number of different fungi as discussed above. Preferably, the organism is a fungus and the one or more different organisms are all different fungi.

Preferably, the organism is a fungi and the one or more different organisms are one or more fungi from one or more different genera of fungi, such as all different genera of fungi. For instance, (a) the organism may be a fungus from the genus *Candida* and (b) the one or more different organisms may be one or more fungi from one or more of, or all of, *Absidia*,

5 *Acremonium*, *Aspergillus*, *Aureobasidium*, *Basidiobolus*, *Blastomyces*, *Blastoschizomyces*, *Candida*, *Cladosporium*, *Coccidioides*, *Cryptococcus*, *Cunninghamella*, *Curvularia*, *Debaryomyces*, *Exophiala*, *Exserohilum*, *Fonsecea*, *Fusarium*, *Geotrichum*, *Histoplasma*, *Issatchenkia*, *Kluyveromyces*, *Malezzesia*, *Mucor*, *Paracoccidioides*, *Paecilomyces*, *Penicillium*, *Pichia*, *Pneumocystis*, *Rhizomucor*, *Rhizopus*, *Rhodotorula*, *Saccharomyces*, *Scedosporium*,
10 *Schizophyllum*, *Scopulariopsis*, *Sporothrix*, *Trichoderma*, *Trichophyton* and *Trichosporon*.

Alternatively, the organism in (a) may be a fungus from one of the genera in list (b) (such as *Aspergillus*) and the one or more different organisms may be one or more bacteria from one or more of, or all of, *Candida* and the remaining genera in list (b) (such as everything except *Aspergillus*).

15 More preferably, (i) the organism is a fungus and (ii) the one or more different organisms are one or more different species of fungus. (i) is preferably selected from the list of fungal species above and the one or more species in (ii) are preferably the species remaining in the list.

The organism is preferably a virus. The virus may belong to the family Retroviridae, such as human deficiency viruses, such as HIV-I (also referred to as HTLV- III), HIV-II, LAC,
20 IDLV-III/LAV, HIV-III or other isolates such as HIV-LP, the family Picornaviridae, such as poliovirus, hepatitis A, enteroviruses, human Coxsackie viruses, rhinoviruses, echoviruses, the family Calciviridae, such as viruses that cause gastroenteritis, the family Togaviridae, such as equine encephalitis viruses and rubella viruses, the family Flaviviridae, such as dengue viruses, encephalitis viruses and yellow fever viruses, the family Coronaviridae, such as coronaviruses,
25 the family Rhabdoviridae, such as vesicular stomata viruses and rabies viruses, the family Filoviridae, such as Ebola viruses, the family Paramyxoviridae, such as parainfluenza viruses, mumps viruses, measles virus and respiratory syncytial virus, the family Orthomyxoviridae, such as influenza viruses, the family Bungaviridae, such as Hantaan viruses, bunya viruses, phleboviruses and Nairo viruses, the family Arenaviridae, such as hemorrhagic fever viruses,
30 the family Reoviridae, such as reoviruses, orbiviruses and rotaviruses, the family Birnaviridae, the family Hepadnaviridae, such as hepatitis B virus, the family Parvoviridae, such as parvoviruses, the Papovaviridae, such as papilloma viruses and polyoma viruses, the family Adenoviridae, such as adenoviruses, the family Herpesviridae, such as herpes simplex virus (HSV) I and II, varicella zoster virus and pox viruses, or the family Iridoviridae, such as African
35 swine fever virus). The virus may be an unclassified virus, such as the etiologic agents of

Spongiform encephalopathies, the agent of delta hepatitis, the agents of non-A, non-B hepatitis (class 1 enterally transmitted; class 2 parenterally transmitted such as Hepatitis C); Norwalk and related viruses and astroviruses.

Preferably, the organism is a virus and the one or more different organisms are one or more different virus. There may be any number of different viruses as discussed above. Preferably, the organism is a virus and the one or more different organisms are all different viruses.

Step (a)

The method comprises extracting all nucleotide k-mers from the genome of the organism. The genome of the organism is typically publically available, for instance on GenBank. A computer is capable of extracting all nucleotide k-mers from the genome.

A nucleotide typically contains a nucleobase, a sugar and at least one linking group, such as a phosphate, 2'-O-methyl, 2' methoxy-ethyl, phosphoramidate, methylphosphonate or phosphorothioate group. The nucleobase is typically heterocyclic. Nucleobases include, but are not limited to, purines and pyrimidines and more specifically adenine (A), guanine (G), thymine (T), uracil (U) and cytosine (C). The sugar is typically a pentose sugar. Nucleotide sugars include, but are not limited to, ribose and deoxyribose.

The nucleotides in the genome of the organism are typically ribonucleotides or deoxyribonucleotides. The nucleotides typically contain a monophosphate, diphosphate or triphosphate. Phosphates may be attached on the 5' or 3' side of a nucleotide.

Nucleotides include, but are not limited to, adenosine monophosphate (AMP), adenosine diphosphate (ADP), adenosine triphosphate (ATP), guanosine monophosphate (GMP), guanosine diphosphate (GDP), guanosine triphosphate (GTP), thymidine monophosphate (TMP), thymidine diphosphate (TDP), thymidine triphosphate (TTP), uridine monophosphate (UMP), uridine diphosphate (UDP), uridine triphosphate (UTP), cytidine monophosphate (CMP), cytidine diphosphate (CDP), cytidine triphosphate (CTP), 5-methylcytidine monophosphate, 5-methylcytidine diphosphate, 5-methylcytidine triphosphate, 5-hydroxymethylcytidine monophosphate, 5-hydroxymethylcytidine diphosphate, 5-hydroxymethylcytidine triphosphate, cyclic adenosine monophosphate (cAMP), cyclic guanosine monophosphate (cGMP), deoxyadenosine monophosphate (dAMP), deoxyadenosine diphosphate (dADP), deoxyadenosine triphosphate (dATP), deoxyguanosine monophosphate (dGMP), deoxyguanosine diphosphate (dGDP), deoxyguanosine triphosphate (dGTP), deoxythymidine monophosphate (dTMP), deoxythymidine diphosphate (dTDP), deoxythymidine triphosphate (dTTP), deoxyuridine monophosphate (dUMP), deoxyuridine diphosphate (dUDP), deoxyuridine

triphosphate (dUTP), deoxycytidine monophosphate (dCMP), deoxycytidine diphosphate (dCDP) and deoxycytidine triphosphate (dCTP), 5-methyl-2'-deoxycytidine monophosphate, 5-methyl-2'-deoxycytidine diphosphate, 5-methyl-2'-deoxycytidine triphosphate, 5-hydroxymethyl-2'-deoxycytidine monophosphate, 5-hydroxymethyl-2'-deoxycytidine diphosphate and 5-hydroxymethyl-2'-deoxycytidine triphosphate. The nucleotides are preferably selected from AMP, TMP, GMP, UMP, dAMP, dTMP, dGMP or dCMP.

The organism preferably comprises a genome which is based on deoxyribonucleic acid (DNA) and which comprises dAMP, dTMP, dGMP and dCMP. If the organism's genome is based on DNA, k-mers which comprise nucleotides other than dAMP, dTMP, dGMP and dCMP are preferably be excluded from the comparison in step (b). The organism, such as a virus, may comprise a genome which is based on ribonucleic acid (RNA) and comprises AMP, TMP, GMP and UMP. If the organism's genome is based on RNA, k-mers which comprise one or more nucleotides other than AMP, TMP, GMP and UMP are preferably excluded from the comparison in step (b).

The nucleotide k-mers may be any length. Step (a) preferably comprises defining a length for the one or more nucleotide k-mers and extracting all nucleotide k-mers having that length from the genome of the organism. Step (a) more preferably comprises defining a length of from about 10 to about 40 nucleotides for the one or more nucleotide k-mers and extracting all nucleotide k-mers having that length from the genome of the organism. The method may comprise defining a length of from about 11, about 12, about 13, about 14, about 15, about 16, about 17, about 18, or about 19 nucleotides to about 20, about 21, about 22, about 23, about 24, about 25, about 26, about 27, about 28, about 29, about 30, about 31, about 32, about 33, about 34, about 35, about 36, about 37, about 38 or about 39 nucleotides.

Preferred k-mer lengths include, but are not limited to, about 17, about 20, about 25 and about 30 nucleotides.

Certain nucleotide k-mers may appear more than once in the genome of the organism. All instances of theses nucleotide k-mers may be extracted and compared in step (b). A nucleotide k-mer which appears more than once in the organism's genome is preferably extracted only once in step (a) and only compared once with the genome(s) of the one or more different organisms in step (b).

Several different versions of the genome of the organism may be publically available. The ability to distinguish the organism from one or more different organisms may be improved if two or more different versions of the genome are analysed using the method of the invention. Step (a) preferably comprises extracting all nucleotide k-mers from two or more different versions of the genome of the organism. Step (a) more preferably comprises extracting all

nucleotide k-mers from (i) two hundred or more or (ii) two thousand or more different versions of the genome of the organism. Any number of different versions of the genome may be analysed such as about five or more, about ten or more, about fifty or more, about one hundred or more, about five hundred or more, about one thousand or more, about five thousand or more or even more.

Step (a) most preferably comprises extracting all nucleotide k-mers from all known versions of the genome of the organism.

In some instances, the nucleotide k-mers may be extracted from certain parts or fragments of the genome. For instance, the method may be focussed on those parts or fragments of the genome which are expected to be conserved amongst different versions of the genome of the organism and/or not conserved in the one or more different organisms. Thus, the method may comprise using only a subset of sequences from the genome of the organism. This may advantageously reduce complexity and input data for genome comparisons. The person skilled in the art is capable of identifying such parts or fragments of the genomes. Step (a) preferably comprises extracting all nucleotide k-mers from a predefined or specific part or fragment of the genome of the organism. The part or fragment may be any part or fragment. The nucleotide k-mers are preferably extracted from the parts or fragments of the genome which encode the ribosomal protein subunits. This is disclosed in the Example. The nucleotide k-mers are preferably extracted from the parts or fragments of the genome which are involved in DNA replication, such as a part or fragment of the genome which encode a polymerase, a helicase and/or a topoisomerases, such as a gyrase.

Step (b)

The method also comprises comparing the nucleotide k-mers extracted in step (a) with the genome(s) of the one or more different organisms, such as all different organisms, and identifying those nucleotide k-mers which do not appear in the genome(s) of the one or more different organisms. This is straightforward using a computer. Once the k-mers have been extracted, the computer can investigate the genome(s) of the one or more different organisms to determine whether or not the k-mers are present. Those k-mers which do not appear in (i.e. are not present in) the genome(s) of the one or more different organisms may be used to distinguish the organism from the one or more different organisms. These k-mers may be used in the subsequent optional steps of the method of the invention. Those k-mers which appear in (i.e. are present in) the genome(s) of the one or more different organisms may not be used to distinguish the organism from the one or more different organisms. These k-mers are typically discarded and are not used in the subsequent optional steps of the method of the invention.

Step (b) preferably comprises comparing the nucleotide k-mers extracted in step (a) with two or more versions of the genome of each the one or more different organisms, such as all different organisms, and identifying those nucleotide k-mers which do not appear in the genomes of the one or more different organisms. Step (b) more preferably comprises comparing the nucleotide k-mers extracted in step (a) with about two hundred or more, about two thousand or more or all known versions of the genome of each the one or more different organisms, such as all different organisms, and identifying those nucleotide k-mers which do not appear in the genomes of the one or more different organisms. Any number of versions of the one or more different organisms may be investigated as discussed above for the organism of interest.

Optional step (c)

The method of the invention may also involve removing those k-mers which may not be helpful in distinguishing the organism from the one or more different organisms. In particular, the method preferably comprises (c) discounting any nucleotide k-mers identified in step (b) which have an undesirable property. The undesirable property is preferably (i) its presence in the human genome, (ii) a homopolymer or repetitive sequence, (iii) the ability to form secondary structure or (iv) its presence in a contaminating organism.

Optional step (d)

At the end of step (b) or step (c), there may be hundreds, thousands or millions of unique k-mers identified in the organism's genome. This may be acceptable for computational applications, but a smaller collection of k-mers (e.g. tens or hundreds) may be desirable in some experimental contexts, such as microarrays or genotyping by species-specific primers.

The method preferably further comprises (d) analysing the nucleotide k-mers identified in step (b) or step (c) and reducing their number by removing at least some sequence redundancy. This can be done using know method, such as set covering. A preferred method is disclosed in the example. Step (d) preferably comprises minimising the number of nucleotide k-mers by removing all sequence redundancy.

Numbers of identified k-mers

The method of the invention may comprise identifying any number of k-mers. The method preferably comprises identifying about 25 or fewer k-mers which are capable of distinguishing the organism from the one or more different organisms such as all different organisms, such as about 24 or fewer k-mers, about 23 or fewer k-mers, about 22 or fewer k-mers, about 21 or fewer k-mers, about 20 or fewer k-mers, about 19 or fewer k-mers, about 18 or

fewer k-mers, about 17 or fewer k-mers, about 16 or fewer k-mers, about 15 or fewer k-mers, about 14 or fewer k-mers, about 13 or fewer k-mers, about 12 or fewer k-mers, about 11 or fewer k-mers, about 10 or fewer k-mers, about 9 or fewer k-mers, about 8 or fewer k-mers, about 7 or fewer k-mers, about 6 or fewer k-mers, about 5 or fewer k-mers, about 4 or fewer k-mers, about 3 or fewer k-mers, or about 2 or fewer k-mers. The method may identify two or more nucleotide k-mers each having the same length. Alternatively, the method may identify two or more nucleotide k-mers having different lengths. The method may identify a population of k-mers in which two or more k-mers have the same length and two or more k-mers have different lengths.

The method preferably identifies only one k-mer which is capable of distinguishing the organism from the one or more different organisms, such as all different organisms.

As explained above, step (a) may involve extracting one or nucleotide k-mers from all of the different versions in the genome of the organism. The method is preferably for identifying about 25 or fewer nucleotide k-mers (or any of the numbers listed above) which are present in all of the different versions of the genome of the organism and which are capable of distinguishing the organism from the one or more different organisms, such as all different organisms.

The method is most preferably for identifying only one nucleotide k-mer which is present in all of the different versions of the genome of the organism and which is capable of distinguishing the organism from the one or more different organisms, such as all different organisms.

The method of the invention may comprise identifying a cluster of k-mer sequences in the genome of the organism of interest, wherein more than one k-mer in the cluster is capable of distinguishing the organism from one or more different organisms. The k-mer sequences in the cluster which are capable of distinguishing the organism may be overlapping. Each nucleotide in the cluster is typically found in at least one k-mer which is capable of distinguishing the organism. The method may comprise identification of a cluster of k-mer sequences in the genome of the organism that has the highest number of overlapping k-mer sequences which are capable of distinguishing the organism. The cluster of k-mer sequences may comprise a k-mer that is not capable of distinguishing the organism from one or more different organisms, in addition to the k-mer sequences which are capable of distinguishing the organism. Examples of k-mer clusters identified according to the invention include SEQ ID NOs: 1 to 213, as further described below.

Phenotype

The invention may also be used to identifying one or more nucleotide k-mers in the genome of an organism which can be associated with a phenotype of the organism. Any of the embodiments apply to this method.

5 Step (a) comprises extracting all nucleotide k-mers from the genome of the organism as described above. K-mers may be extracted from any number of different versions of the genome of the organism as discussed above.

10 Step (b) comprises comparing the nucleotide k-mers extracted in step (a) with the genome(s) of one or more different organisms, such as all different organisms, which do not display the phenotype and identifying those nucleotide k-mers which do not appear in the genome(s) of the one or more different organisms which do not display the phenotype. Any number of different versions of the genomes of the one or more different organisms, such as all different organisms, which do not display the phenotype may be investigated as discussed above.

15 Those k-mers which do not appear in (i.e. are not present in) the genome(s) of the one or more different organisms may be associated with a phenotype of the organism. These k-mers may be used in the subsequent optional steps of the method of the invention. Those k-mers which appear in (i.e. are present in) the genome(s) of the one or more different organisms may not be associated with a phenotype of the organism. These k-mers are typically discarded and are not used in the subsequent optional steps of the method of the invention.

20 The phenotype may be any phenotype of the organism. Preferred phenotypes include, but are not limited to, pathogenicity, resistance to antibiotics, host specificity, tissue specificity, transmissibility, virulence, antigenicity and biochemical properties. The one or more different organisms are typically from the same phylum, family or genus as the organism of interest.

Computer program

25 The invention also provides a computer program configured to carry out the method of the invention. The program can be written using routine methods.

The invention also provides a computer program medium, such as hard drive, USD or disk, comprising the computer program of the invention.

30 The invention further provides a computer for carrying out the method of the invention programmed with the computer program of the invention.

K-mers and probes

35 The invention also provides one or more nucleotide k-mers identified using the method of the invention or the computer program of the invention. The one or more nucleotide k-mers may have any length as discussed above. There may be any number of nucleotide k-mers as

discussed above. The one or more k-mers can preferably be associated with a phenotype of the organism. The invention further provides one or more nucleotide k-mers capable of distinguishing an organism from one or more different organism, wherein the nucleotide k-mer(s) are from the parts of the organism's genome which encode the ribosomal protein subunits, and wherein the nucleotide k-mer(s) are from about 10 to about 40 nucleotides in length. The nucleotide k-mer(s) may be of a preferred number, length or other characteristic as described above.

The invention also provides an oligonucleotide probe which comprises a sequence which is complementary to a nucleotide k-mer of the invention. The invention also provides an oligonucleotide probe which comprises a sequence which is the reverse complement of a nucleotide k-mer of the invention. The invention also provides a plurality of oligonucleotide probes each of which comprises a sequence which is complementary to one of the nucleotide k-mers in a population of two or more nucleotide k-mers of the invention. The invention also provides a plurality of oligonucleotide probes each of which comprises a sequence which is the reverse complement of one of the nucleotide k-mers in a population of two or more nucleotide k-mers of the invention.

Oligonucleotides are short nucleotide polymers which typically have about 50 or fewer nucleotides, such about 40 or fewer, about 30 or fewer or about 20 or fewer nucleotides. The sequence in the oligonucleotide probe/probes which is complementary to or the reverse complement of a nucleotide k-mer is typically the same length as the k-mer, such as from about 10 to about 40 nucleotides. The sequence may have a length of from about 11, about 12, about 13, about 14, about 15, about 16, about 17, about 18, or about 19 nucleotides to about 20, about 21, about 22, about 23, about 24, about 25, about 26, about 27, about 28, about 29, about 30, about 31, about 32, about 33, about 34, about 35, about 36, about 37, about 38 or about 39 nucleotides.

Preferred sequence lengths include, but are not limited to, about 17, about 20, about 25 and about 30 nucleotides.

The oligonucleotide probe/probes may comprise any of the nucleotide discussed above. The nucleotides are preferably selected from AMP, TMP, GMP, UMP, dAMP, dTMP, dGMP or dCMP.

The nucleotides may contain additional modifications. In particular, suitable modified nucleotides include, but are not limited to, 2'-amino pyrimidines (such as 2'-amino cytidine and 2'-amino uridine), 2'-hydroxyl purines (such as , 2'-fluoro pyrimidines (such as 2'-fluorocytidine and 2'fluoro uridine), hydroxyl pyrimidines (such as 5'- α -P-borano uridine), 2'-O-methyl nucleotides (such as 2'-O-methyl adenosine, 2'-O-methyl guanosine, 2'-O-methyl

cytidine and 2'-O-methyl uridine), 4'-thio pyrimidines (such as 4'-thio uridine and 4'-thio cytidine) and nucleotides have modifications of the nucleobase (such as 5-pentynyl-2'-deoxy uridine, 5-(3-aminopropyl)-uridine and 1,6-diaminohexyl-N-5-carbamoylmethyl uridine).

One or more nucleotides in the oligonucleotide probe/probes can be oxidized or
5 methylated.

The nucleotides in the oligonucleotide probe/probes may be attached to each other in any manner. The nucleotides may be linked by phosphate, 2'-O-methyl, 2' methoxy-ethyl, phosphoramidate, methylphosphonate or phosphorothioate linkages. The nucleotides are typically attached by their sugar and phosphate groups as in nucleic acids. The nucleotides may
10 be connected via their nucleobases as in pyrimidine dimers.

The oligonucleotide probe/probes can be a nucleic acid, such as deoxyribonucleic acid (DNA) or a ribonucleic acid (RNA). The polynucleotide may be any synthetic nucleic acid known in the art, such as peptide nucleic acid (PNA), glycerol nucleic acid (GNA), threose nucleic acid (TNA), locked nucleic acid (LNA), morpholino nucleic acid or other synthetic
15 polymers with nucleotide side chains.

The oligonucleotide probe/probes may be single stranded. The oligonucleotide probe/probes may be double stranded. The oligonucleotide probe/probes may comprise a hairpin. Oligonucleotides may be synthesised using standard techniques known in the art.

The oligonucleotide probe/probes preferably comprise a sequence which complementary
20 to or the reverse complement of the k-mer/k-mers through Watson and Crick base pairing. If the nucleotide k-mer/k-mers is/are derived from an organism whose genome is based on deoxyribonucleic acid (DNA), the probe/probes preferably comprise(s) dAMP, dTMP, dGMP and dCMP. If the nucleotide k-mer/k-mers are derived from an organism whose genome is based on ribonucleic acid (RNA), the probe/probes preferably comprise(s) AMP, TMP, GMP
25 and UMP.

Preferably, the probe or probes of the invention is/are detectably-labelled. The detectable label preferably allows the presence or absence of a hybridization product formed by specific hybridization between the probe and the complementary nucleotide k-mer (and thereby the presence or absence of the nucleotide k-mer) to be determined. Any label can be used. Suitable
30 labels include, but are not limited to, fluorescent molecules, radioisotopes, e.g. ¹²⁵I, ³⁵S, enzymes, antibodies and linkers such as biotin.

The probe can be a molecular beacon probe comprising a fluorescent label at one end and a quenching molecule at the other. In the absence of the complementary nucleotide k-mer, the probe forms a hairpin loop and the quenching molecule is brought into close proximity with the
35 fluorescent label so that no signal can be detected. Upon hybridization of the probe to the

nucleotide k-mer, the loop unzips and the fluorescent molecule is separated from the quencher such that a signal can be detected. Suitable fluorescent molecules and quencher for use in molecular beacons are known in the art. These include, but are not limited to, the fluorophores carboxyfluorescein (FAM) and HEX and the quenchers dabcy1, DDQ1 and DDQ2.

5 The probe can be a scorpion probe, which is a probe linked to primer. The primer part of the probe can be designed to amplify the nucleotide k-mer to be detected and the probe part can be designed to detect the amplified nucleotide k-mer. Scorpion probes are well-known in the art. They are described in, for example, Whitcombe et al. (Nat. Biotechnol., 1994; 17: 804- 807).

10 The probe/probes may be attached to or immobilised on a support using any technology which is known in the art. Suitable solid supports are well-known in the art and include plates, such as multi-well plates, filters, membranes, beads, chips, pins, array, dipsticks, nanoparticles and porous carriers. The probe may be used as part of an array-based detection method.

The invention also provides a support having attached thereto or immobilised thereon a oligonucleotide probe of the invention or a plurality of oligonucleotide probes of the invention.

15 The support preferably comprises a chip, pin, array or dipstick.

Detection method

The invention also provides a method for detecting the presence or absence of an organism in a sample. The method comprises detecting the presence or absence in the sample of
20 one or more nucleotide k-mers which are capable of distinguishing the organism from one or more different organisms, such as all different organisms. The one or more k-mers are characteristic of the organism. The one or more k-mers are specific for the organism.

The presence of the one or more nucleotide k-mers is indicative of the presence of the organism in the sample. The absence of the one or more nucleotide k-mers is indicative of the
25 absence of the organism from the sample. The method is preferably for detecting the presence or absence of an organism with a phenotype in a sample and the method comprises detecting the presence or absence in the sample of one or more nucleotide k-mers which are associated with the phenotype. The method preferably comprises the rapid detection of the presence or absence in the sample of one or more nucleotide k-mers.

30 The method may be carried out on any suitable sample. The method is typically carried out on a sample that is known to contain or suspected to contain the organism. Alternatively, the invention may be carried out on a sample to confirm the presence of an organism whose presence in the sample is known or expected.

The sample may be a biological sample. The method may be carried out *in vitro* using a
35 sample obtained from or extracted from any organism or microorganism. The organism or

microorganism is typically archaeal, prokaryotic or eukaryotic and typically belongs to one of the five kingdoms: plantae, animalia, fungi, monera and protista. The invention may be carried out *in vitro* on a sample obtained from or extracted from any virus. The sample is preferably a fluid sample. The sample typically comprises a body fluid of the patient. The sample may be urine, lymph, saliva, mucus or amniotic fluid but is preferably blood, plasma or serum.

Typically, the sample is human in origin, but alternatively it may be from another mammal animal such as from commercially farmed animals such as horses, cattle, sheep, fish, chickens or pigs or may alternatively be pets such as cats or dogs. Alternatively, the sample may be of plant origin, such as a sample obtained from a commercial crop, such as a cereal, legume, fruit or vegetable, for example wheat, barley, oats, canola, maize, soya, rice, rhubarb, bananas, apples, tomatoes, potatoes, grapes, tobacco, beans, lentils, sugar cane, cocoa, cotton.

The sample may be a non-biological sample. The non-biological sample is preferably a fluid sample. Examples of non-biological samples include surgical fluids, water such as drinking water, sea water or river water, and reagents for laboratory tests.

The sample is typically processed prior to being used in the invention, for example by centrifugation or by passage through a membrane that filters out unwanted molecules or cells, such as red blood cells. The sample may be measured immediately upon being taken. The sample may also be typically stored prior to assay, preferably below -70°C.

The one or more nucleotide k-mers are present in the genome of the organism. The method typically comprises extracting the genomic material, such as DNA or RNA, from the sample before the presence or absence of the one or more nucleotide k-mers is detected. The genomic material can be extracted using routine methods known in the art. For instance, commercially available extraction kits may be used such as MycXtra (Myconostica, UK), QIAamp Blood mini kit (Qiagen, Germany), QIAamp DNA mini kit (Qiagen, Germany) and BiOstic Bacteremia DNA Isolation kit (MoBio, USA). Suitable methods of extracting RNA are disclosed in the art such as the commercially available RNeasy mini kit (Qiagen, Germany).

The method preferably comprises the step of amplifying the one or more nucleotide k-mers. In one embodiment, the one or more nucleotide k-mers are amplified before their presence or absence is detected. In another embodiment, the one or more nucleotide k-mers are amplified in real time as their presence is determined. Real-time methods are described in the art.

In one embodiment, only the one or more nucleotide k-mers to be detected are amplified. In other embodiments, the one or more nucleotide k-mers to be detected is amplified as part of a much larger length of genomic material, such as DNA or RNA. Sequences of genomic material, such as DNA or RNA, having at least about 50, at least about 70, at least about 90, at least about 110, at least about 120, at least about 130, at least about 140, at least about 150, at least about

160, at least about 170, at least about 180, at least about 190, at least about 200, at least about 250, at least about 300, at least about 400 or at least about 500 nucleotides and comprising the region to be detected can be amplified. For example, sequences having from 10 to 2000, from 20 to 1500, from 50 to 1000 or from 100 to 500 nucleotides can be amplified.

5 The one or more nucleotide k-mers can be amplified using routine methods that are known in the art. The amplification of DNA is preferably carried out using polymerase chain reaction (PCR), nucleic acid sequence based analysis (NASBA) or Loop-mediated isothermal amplification (LAMP). RNA can be amplified using routine methods in the art, such as reverse transcription-PCR.

10 The one or more nucleotide k-mers can be detected using any method known in the art. The one or more k-mers may be detected using next generation sequencing such as nanopore sequencing disclosed in WO 2012/107778, WO 2013/014451, WO 2013/121224 and WO 2013/153359.

15 The one or more nucleotide k-mers can be detected using TaqMan PCR or TaqMan realtime PCR. These techniques are well-known in the art.

20 The one or more nucleotide k-mers are preferably detected using the probe or probes of the invention. Typically, the detecting comprises contacting the probe or probes with the sample under conditions in which the probe hybridizes to the one or more nucleotide k-mers, if present, and determining the presence or absence of the hybridization product. The presence of the hybridization product indicates the presence of the one or more nucleotide k-mers. Conversely, the absence of the hybridization product indicates the absence of the one or more nucleotide k-mers. Conditions that permit the hybridization are well-known in the art (for example, Sambrook *et al.*, 2001, Molecular Cloning: a laboratory manual, 3rd edition, Cold Spring Harbour Laboratory Press; and Current Protocols in Molecular Biology, Chapter 2, Ausubel *et al.*, Eds., Greene Publishing and Wiley-Interscience, New York (1995)). The method of the invention can be carried out under low stringency conditions, for example in the presence of a buffered solution of 30 to 35% formamide, 1 M NaCl and 1 % SDS (sodium dodecyl sulfate) at 37°C followed by a wash in from 1X (0.1650 M NaCl) to 2X (0.33 M NaCl) SSC (salt sodium citrate) at 50°C. The method of the invention can be carried out under moderate stringency conditions, for example in the presence of a buffer solution of 40 to 45% formamide, 1 M NaCl, and 1 % SDS at 37°C, followed by a wash in from 0.5X (0.0825 M NaCl) to 1X (0.1650 M NaCl) SSC at 55°C. The method of the invention can be carried out under high stringency conditions, for example in the presence of a buffered solution of 50% formamide, 1 M NaCl, 1% SDS at 37°C, followed by a wash in 0.1X (0.0165 M NaCl) SSC at 60°C.

If more than one of nucleotide k-mers are being detected simultaneously, the different probes used to detect the different nucleotide k-mers may be labelled with different labels. Probes having different labels are preferable when different nucleotide k-mers are being detected simultaneously in the same volume of sample. When the two or more different nucleotide k-mers are being detected in the same volume of sample, it must be possible to distinguish between the different labels and hence detect the different nucleotide k-mers. For instance, fluorescent molecules that emit different wavelengths of light can be used. A suitable group of fluorescent labels, each of which can be simultaneously detected, is HEX hexachloro fluorescein phosphoramidite (HEX), carboxyfluorescein (FAM), Cy(R)5 and Texas Red(R). Other suitable groups of labels are known in the art.

Any of the probes and supports of the invention may be used in the detection method. The presence or absence of multiple different organisms may be simultaneously detected using array-based methods.

Specific nucleotide k-mers

The inventors have also identified specific k-mer clusters in various organisms, as described in Example 1. These are shown as SEQ ID NOs: 1 to 213 in the sequence listing. Each cluster comprises various k-mers which may be used to distinguish the relevant bacterium from one or more different organisms, such as one or more different bacteria.

The invention therefore provides a nucleotide k-mer capable of distinguishing an organism from one or more different organisms, wherein the k-mer comprises (or consists of) about 10 or more consecutive nucleotides from any one of the sequences shown in SEQ ID NOs: 1 to 213. The k-mer may be any length. The k-mer may comprise at least about 15, at least about 20, at least about 25, at least about 30, or at least about 35 consecutive nucleotides from any one of the sequences shown in SEQ ID NOs: 1 to 213.

The k-mer preferably comprises (or consists of) from about 10 to about 40 consecutive nucleotides from any of the sequences shown in SEQ ID NOs: 1 to 213. The k-mer is preferably from about 11, about 12, about 13, about 14, about 15, about 16, about 17, about 18, or about 19 nucleotides to about 20, about 21, about 22, about 23, about 24, about 25, about 26, about 27, about 28, about 29, about 30, about 31, about 32, about 33, about 34, about 35, about 36, about 37, about 38 or about 39 consecutive nucleotides from any one of SEQ ID NOs: 1 to 213.

Preferred k-mer lengths include, but are not limited to, about 17, about 20, about 25 and about 30 consecutive nucleotides from any one of the sequences shown in SEQ ID NOs: 1 to 213.

The k-mer preferably comprises (or consists of) any of the consecutive nucleotides listed in the last column in Table 3. The locations of the preferred k-mers are listed in this column as SEQ ID NO:/start nucleotide-end nucleotide. For instance, the first preferred k-mer is located at 1/18-34. This means it corresponds to nucleotides 18 to 34 of SEQ ID NO: 1. The second preferred k-mer is located at 2/9-25. This means it corresponds to nucleotides 9 to 25 of SEQ ID NO: 2 and so on.

The nucleotide k-mer of the invention can be associated with a phenotype of the organism. This is discussed in detail above.

Preferably, the k-mer is from one of SEQ ID NOs: 1 to 5 and is capable of distinguishing *Aggregatibacter actinomycetemcomitans* from one or more different organisms, such as all different organisms.

Preferably, the k-mer is from one of SEQ ID NOs: 6 to 10 and is capable of distinguishing *Bacillus anthracis* from one or more different organisms, such as all different organisms.

Preferably, the k-mer is from one of SEQ ID NOs: 11 to 15 and is capable of distinguishing *Bacillus licheniformis* from one or more different organisms, such as all different organisms.

Preferably, the k-mer is from one of SEQ ID NOs: 16 to 20 and is capable of distinguishing *Bacteroides fragilis* from one or more different organisms, such as all different organisms.

Preferably, the k-mer is from one of SEQ ID NOs: 21 to 25 and is capable of distinguishing *Bartonella henselae* from one or more different organisms, such as all different organisms.

Preferably, the k-mer is from one of SEQ ID NOs: 26 to 30 and is capable of distinguishing *Bordetella pertussis* from one or more different organisms, such as all different organisms.

Preferably, the k-mer is from one of SEQ ID NOs: 31 to 35 and is capable of distinguishing *Borrelia burgdorferi* from one or more different organisms, such as all different organisms.

Preferably, the k-mer is from one of SEQ ID NOs: 36 to 40 and is capable of distinguishing *Brucella abortus* from one or more different organisms, such as all different organisms.

Preferably, the k-mer is from one of SEQ ID NOs: 41 to 44 and is capable of distinguishing *Campylobacter jejuni* from one or more different organisms, such as all different organisms.

Preferably, the k-mer is from one of SEQ ID NOs: 45 to 49 and is capable of distinguishing *Chlamydia trachomatis* from one or more different organisms, such as all different organisms.

5 Preferably, the k-mer is from one of SEQ ID NOs: 50 to 54 and is capable of distinguishing *Chlamydophila pneumonia* from one or more different organisms, such as all different organisms.

Preferably, the k-mer is from one of SEQ ID NOs: 55 to 59 and is capable of distinguishing *Clostridium difficile* from one or more different organisms, such as all different organisms.

10 Preferably, the k-mer is from one of SEQ ID NOs: 60 to 64 and is capable of distinguishing *Clostridium perfringens* from one or more different organisms, such as all different organisms.

15 Preferably, the k-mer is from one of SEQ ID NOs: 65 to 69 and is capable of distinguishing from *Enterobacter aerogenes* from one or more different organisms, such as all different organisms.

Preferably, the k-mer is from one of SEQ ID NOs: 70 to 74 and is capable of distinguishing from *Enterococcus faecalis* from one or more different organisms, such as all different organisms.

20 Preferably, the k-mer is from one of SEQ ID NOs: 75 to 79 and is capable of distinguishing from *Enterococcus faecium* from one or more different organisms, such as all different organisms.

Preferably, the k-mer is from one of SEQ ID NOs: 80 to 84 and is capable of distinguishing *Francisella tularensis* from one or more different organisms, such as all different organisms.

25 Preferably, the k-mer is from SEQ ID NO: 85 and is capable of distinguishing *Haemophilus influenza* from one or more different organisms, such as all different organisms.

Preferably, the k-mer is from one of SEQ ID NOs: 86 and 87 and is capable of distinguishing *Helicobacter pylori* from one or more different organisms, such as all different organisms.

30 Preferably, the k-mer is from one of SEQ ID NOs: 88 to 92 and is capable of distinguishing *Klebsiella oxytoca*

Preferably, the k-mer is from one of SEQ ID NOs: 93 to 97 and is capable of distinguishing *Legionella pneumophila* from one or more different organisms, such as all different organisms.

Preferably, the k-mer is from one of SEQ ID NOs: 98 to 102 and is capable of distinguishing *Listeria monocytogenes* from one or more different organisms, such as all different organisms.

5 Preferably, the k-mer is from one of SEQ ID NOs: 103 to 107 and is capable of distinguishing *Moraxella catarrhalis* from one or more different organisms, such as all different organisms.

Preferably, the k-mer is from one of SEQ ID NOs: 108 to 112 and is capable of distinguishing *Mycobacterium avium* from one or more different organisms, such as all different organisms.

10 Preferably, the k-mer is from one of SEQ ID NOs: 113 to 116 and is capable of distinguishing *Mycobacterium bovis* from one or more different organisms, such as all different organisms.

15 Preferably, the k-mer is from one of SEQ ID NOs: 117 to 121 and is capable of distinguishing *Mycoplasma genitalium* from one or more different organisms, such as all different organisms.

Preferably, the k-mer is from one of SEQ ID NOs: 122 to 126 and is capable of distinguishing *Mycoplasma pneumonia* from one or more different organisms, such as all different organisms.

20 Preferably, the k-mer is from one of SEQ ID NOs: 127 to 131 and is capable of distinguishing *Neisseria gonorrhoeae* from one or more different organisms, such as all different organisms.

Preferably, the k-mer is from one of SEQ ID NOs: 132 to 136 and is capable of distinguishing *Neisseria meningitides* from one or more different organisms, such as all different organisms.

25 Preferably, the k-mer is from one of SEQ ID NOs: 137 to 141 and is capable of distinguishing *Porphyromonas gingivalis* from one or more different organisms, such as all different organisms.

30 Preferably, the k-mer is from one of SEQ ID NOs: 142 to 146 and is capable of distinguishing *Proteus mirabilis* from one or more different organisms, such as all different organisms.

Preferably, the k-mer is from one of SEQ ID NOs: 147 to 151 and is capable of distinguishing *Pseudomonas aeruginosa* from one or more different organisms, such as all different organisms.

Preferably, the k-mer is from one of SEQ ID NOs: 152 to 156 and is capable of distinguishing *Salmonella enterica* from one or more different organisms, such as all different organisms.

5 Preferably, the k-mer is from one of SEQ ID NOs: 157 to 161 and is capable of distinguishing *Serratia marcescens* from one or more different organisms, such as all different organisms.

Preferably, the k-mer is from one of SEQ ID NOs: 162 to 166 and is capable of distinguishing *Staphylococcus aureus* from one or more different organisms, such as all different organisms.

10 Preferably, the k-mer is from one of SEQ ID NOs: 167 to 171 and is capable of distinguishing *Staphylococcus epidermidis* from one or more different organisms, such as all different organisms.

15 Preferably, the k-mer is from one of SEQ ID NOs: 172 to 176 and is capable of distinguishing *Staphylococcus haemolyticus* from one or more different organisms, such as all different organisms.

Preferably, the k-mer is from SEQ ID NO: 177 and is capable of distinguishing *Stenotrophomonas maltophilia* from one or more different organisms, such as all different organisms.

20 Preferably, the k-mer is from one of SEQ ID NOs: 178 to 182 and is capable of distinguishing *Streptococcus mutans* from one or more different organisms, such as all different organisms.

Preferably, the k-mer is from one of SEQ ID NOs: 183 to 187 and is capable of distinguishing *Streptococcus pyogenes* from one or more different organisms, such as all different organisms.

25 Preferably, the k-mer is from one of SEQ ID NOs: 188 to 192 and is capable of distinguishing *Streptococcus salivarius* from one or more different organisms, such as all different organisms.

30 Preferably, the k-mer is from one of SEQ ID NOs: 193 to 197 and is capable of distinguishing *Streptococcus sanguinis* from one or more different organisms, such as all different organisms.

Preferably, the k-mer is from one of SEQ ID NOs: 198 to 202 and is capable of distinguishing *Treponema pallidum* from one or more different organisms, such as all different organisms.

35 Preferably, the k-mer is from SEQ ID NO: 203 and is capable of distinguishing *Vibrio cholera* from one or more different organisms, such as all different organisms.

Preferably, the k-mer is from one of SEQ ID NOs: 204 to 208 and is capable of distinguishing *Vibrio parahaemolyticus* from one or more different organisms, such as all different organisms. or

5 Preferably, the k-mer is from one of SEQ ID NOs: 209 to 213 and is capable of distinguishing *Yersinia enterocolitica* from one or more different organisms, such as all different organisms.

10 The preferred k-mers are preferably capable of distinguishing the bacterium of interest (*i.e.* the bacterium from which they derived) from one or more different bacteria, such as all different bacteria. The preferred k-mers are preferably capable of distinguishing the bacterium of interest (*i.e.* the bacterium from which they are derived) from one or more of the different bacteria in Table 3, such as all the different bacteria in Table 3.

15 The invention also provides an oligonucleotide probe which comprises (or consists of) a sequence which is complementary to a nucleotide k-mer of the invention. Probes are discussed in more detail above. The probe preferably comprises (or consists of) a sequence which is complementary to one of the k-mers listed in the last column in Table 3.

The probe is preferably detectably labelled and/or is a molecular beacon probe.

The invention also provides a support having attached thereto an oligonucleotide probe of the invention. The support comprises a chip, pin, array or dipstick.

20 The invention also provides a method for detecting the presence or absence of an organism in a sample, comprising detecting the presence or absence in the sample of a nucleotide k-mer of the invention. The nucleotide k-mer is capable of distinguishing the organism from one or more different organisms, such as all different organisms. The nucleotide k-mer is characteristic of the organism. The nucleotide k-mer is specific for the organism. Any of the combinations of specific organisms and specific k-mers disclosed above may be used in the method.

25 The presence of the nucleotide k-mer is indicative of the presence of the organism in the sample. The absence of the nucleotide k-mer is indicative of the absence of the organism from the sample. The method may comprise detecting two or more k-mers of the invention, such as three or more, four or more or five or more k-mers of the invention. The organism is typically one of the organisms shown in Table 3 and the k-mer is typically derived from one of the cluster sequences in this organism. The k-mer is preferably one of the k-mers listed in the last column in Table 3.

30 The method is preferably for detecting the presence or absence of the organism with a particular phenotype in a sample and the method comprises detecting the presence or absence in the sample of a nucleotide k-mer of the invention which is associated with the phenotype.

Detection methods are discussed in more detail above and any of the embodiments discussed above equally apply to the method of using the specific k-mers of the invention.

Example 1

5

Overview

This Example describes the computational methodology to efficiently identify collections of short regions of DNA (k-mers) from protein-encoding ribosomal genes that can be used to identify bacterial species. Initial findings show that it is possible to identify a relatively small
10 number of k-mers from these collections that are both highly species specific and have high coverage in terms of the number of isolates within the species that can be reliably identified. The results have implications for practical applications such as rapid diagnostic tests for pathogenic bacteria.

15 Methods

DNA sequence data for bacterial isolates

The whole genome sequence data for bacterial isolates were downloaded from the NCBI Nucleotide database (2,312 isolates) and the NCBI Assembly database (13,307 isolates) and
20 entered into the BIGSdb rMLST isolate database. Plasmid sequences were also included where appropriate. A further 71,348 genomes were derived from assembling short-read sequencing data for isolate entries in the European Nucleotide Archive Sequence Read Archive (ENA-SRA) using the Velvet assembly algorithm. In total, 86,967 isolates with associated whole genome data were used in this study, representing 3,062 different species. The species annotation of each
25 isolate was validated using an analysis of the protein-encoding ribosomal genes to a set of reliable isolate references.

Identification of protein-encoding ribosomal genes

The protein-encoding ribosomal genes within the isolates in the dataset were identified
30 through a series of BLAST sequence searches against the BIGSdb rMLST sequence definition database. This is an iterative process known colloquially as scanning and tagging. The sequence definition database is based on 'seed' alleles from the ribosomal genes of reference genomes of many different species and has now grown to include over 6,000 alleles for each ribosomal gene. The genome sequence of each isolate is searched against the rMLST sequence definition
35 database. DNA regions with identical matches to existing alleles and with valid start and stop

codons are automatically assigned an allele identifier and the start/stop positions of the allele are recorded. Non-identical matches can be identified by manual inspection, helped by an algorithm that looks for start and stop codons within 18 nucleotides of the ends of the alignment with the closest matching allele. Allele matches with a high sequence identity (at least 95%) and at least 90% overlap to an existing allele are added to the sequence definition database and can therefore be automatically identified in subsequent scans. Allele matches below these thresholds require additional validation to confirm that the gene in question has been correctly identified. The protein-encoding ribosomal gene sequences of the 86,967 isolate dataset were extracted based on these allele definitions and used to create a dataset called the 'protein-encoding ribosomal DNA sequence dataset'. The whole genomes of the 86,967 isolates were extracted and form a dataset called 'whole genome DNA sequence dataset'.

Single species k-mer library generation

In order to identify short regions of DNA (k-mers) from protein-encoding ribosomal gene sequences that uniquely identify a particular species of bacteria, four computational stages were defined: 1) k-mer extraction, 2) k-mer scanning and species filtering, 3) k-mer property filtering and 4) final k-mer selection. Each stage is described in more detail below and shown in Figure 1.

Figure 1: A flowchart describing the four computational stages required to generate a small single species k-mer library from protein-encoding ribosomal gene DNA.

Stage 1: K-mer Extraction

K-mers of a particular size, N (e.g. 20 nucleotides) were extracted from the protein-encoding ribosomal DNA sequence dataset by sliding a window of N bases along the DNA sequence of each protein-encoding gene and recording each k-mer the first time it was observed. K-mers containing non-standard nucleotides (non-ATGC) were not included. This creates a unique list of 'protein-encoding ribosomal DNA k-mers'.

Stage 2: K-mer Scanning and Species Filtering

The isolate sequences from the whole genome DNA sequence dataset were scanned one after another against the list of protein-encoding ribosomal DNA k-mers. This was achieved by sliding along each contiguous sequence with a window of N bases (the same size as the list of k-mers) and looking to see if the exact k-mer existed in the input k-mer list. Presence of a k-mer resulted in the internal recording of the k-mer and the species of the isolate in question. If the number of species observed for a particular k-mer exceeded one, the k-mer was removed from

the ribosomal k-mer list. This process creates a library of ‘single species k-mers’ with each k-mer in the library associated with a single species definition.

Stage 3: K-mer Property Filtering (optional)

5 Filtering steps can now be applied to the single species k-mer library to remove sequences that have undesirable properties in an experimental context, for example long runs of the same nucleotide that may be difficult to sequence accurately. Another example would be to remove any k-mers that are present in the human genome as these may cause false positive matches in a diagnostic library used to analyse human samples. These steps are optional but
10 should take place before final k-mer selection.

Stage 4: Final K-mer Selection

 At this stage there can be millions of unique k-mers in the single species k-mer library. This may be acceptable for computational applications, but a smaller collection of k-mers (e.g.
15 hundreds or thousands) may be desirable in some experimental contexts e.g. microarrays or genotyping by species-specific primers.

(4A) Total Coverage Single Species K-mers

 To reduce the redundancy in the single species k-mer library, the first step is to identify
20 the k-mers that have total coverage of the isolates in the species. If these are present, the next step is to map the total coverage k-mers back onto the original protein-encoding ribosomal genes for a representative isolate for each species. Any contiguous region of ribosomal sequence containing at least one k-mer is defined as a k-mer cluster. A limited number of clusters can then be selected (for example 5 per species) and a single representative k-mer can be selected from
25 each cluster.

 Using the above 86,967 isolate dataset, we performed stages 1 and 2 for k-mer lengths 17, 20, 25 and 30 and then filtered out any k-mer that matched human DNA in stage 3. Total coverage single species k-mers were then identified in stage 4. Summary statistics for each stage
30 of the process are show in Table 1.

K-mer Length	Unique K-mers from K-mer extraction (Stage 1)	Unique Single Species K-mers (Stage 2)	Unique Single Species K-mers Filtered to Remove Human DNA (Stage 3)	Unique Single Species K-mers with Total Coverage of the Isolate Population (Stage 4)
17	32,479,274	4,794,410	4,098,981	3,343,710

20	37,446,505	25,962,711	25,759,215	20,878,341
25	42,727,248	35,034,225	35,032,769	27,858,947
30	46,063,946	39,036,215	39,036,166	30,432,337

Table 1. Number of k-mers at each stage of the process for four different k-mer lengths.

The total coverage non-human k-mers for the four k-mer lengths were combined and mapped onto a set of protein-encoding ribosomal DNA sequences from representative isolates (one per species). The cluster density of each k-mer cluster was calculated according to the definition below and the number of k-mer lengths present was recorded. The clusters were ranked by 1) number of k-mer lengths present in the cluster (descending) 2) cluster density (descending). The 5 highest ranked k-mer clusters on different ribosomal genes for each species were selected.

Cluster Density Definition: The number of observed k-mers in each cluster divided by the maximum number of possible k-mers for the cluster length (expressed as a percentage).

Example: For the 4 k-mer lengths used in the study (17, 20, 25, 30), a cluster with length 30 can have 1 x 30nt, 6 x 25nt, 11 x 20nt and 14 x 17nt. This equals a maximum of 32 possible k-mers for this cluster length. If all 32 possible k-mers were present the cluster density would be 100%. If 20 different k-mers were found in the cluster, the cluster density would be $20/32 \times 100\% = 62.5\%$.

Identifying a representative k-mer within each k-mer cluster for each k-mer length was performed as follows: Each nucleotide position in the k-mer cluster was assigned a number equal to the total number of k-mers mapping to that position (across all sampled lengths). Each k-mer in the cluster was scored based on the sum of the nucleotide position scores. The representative k-mer for each k-mer length was the single species k-mer with the highest total score for that length.

(4B) Partial Coverage Single Species K-mers

If total coverage k-mers are not present in the single species k-mer library for a particular species, then the minimal number of partial coverage k-mers for that species can be identified by formalising the scenario as a 'set covering problem' and this can be solved by implementing a greedy algorithm. For example, in a simple scenario with 5 isolates in a particular species, k-mer A is found in isolates 1, 2, 3, k-mer B matches isolates 2 and 4 and k-mer C matches isolates 3 and 4 and k-mer D matches isolates 4 and 5. k-mers A and D would be the minimum set of k-mers to achieve 100% coverage of the isolates in the species (5 isolates). The greedy algorithm

selects sets according to the rule that at each stage, choose the set/k-mer that contains the largest number of uncovered elements (matched isolates). In Computer Science terms (as described in Wikipedia, http://en.wikipedia.org/wiki/Set_cover_problem): Universe = {1,2,3,4,5}, Sets S = { {1,2,3}, {2,4}, {3,4}, {4,5} }. We can cover all the elements of the universe with two sets, {

5 {1,2,3}, {4,5} }.

Results

Species Summary (Table 2)

10 Summary of the results for each of the pathogenic species (46) including external database references of a representative genome (NCBI) and the number of selected k-mers at each k-mer length.

Species	k-me rs w/ 17n t	k-me rs w/ 20n t	k-me rs w/ 25n t	k-me rs w/ 30n t	Tot al k- mer s	NCBI Nucleotid e Accession	NCBI Genbank Identifier	NCBI Nucleotide Entry Description
<i>Aggregatibacter actinomycetemcomitans</i>	5	5	5	5	20	CP001733.1	GI:261412053	Aggregatibacter actinomycetemcomitans D11S-1, complete genome
<i>Bacillus anthracis</i>	5	5	5	5	20	AE017334.2	GI:50082967	Bacillus anthracis str. 'Ames Ancestor', complete genome
<i>Bacillus licheniformis</i>	5	5	5	5	20	NC_006322.1	GI:52783855	Bacillus licheniformis DSM 13 = ATCC 14580, complete genome
<i>Bacteroides fragilis</i>	5	5	5	5	20	NC_006347.1	GI:53711291	Bacteroides fragilis YCH46 DNA, complete genome
<i>Bartonella henselae</i>	5	5	5	5	20	NC_005956.1	GI:49474831	Bartonella henselae str. Houston-1 chromosome, complete genome
<i>Bordetella pertussis</i>	5	5	5	5	20	BX470248.1	GI:33591069	Bordetella pertussis strain Tohama I, complete genome
<i>Borrelia burgdorferi</i>	5	5	5	5	20	NC_011728.1	GI:218249165	Borrelia burgdorferi ZS7, complete genome
<i>Brucella abortus</i>	4	5	5	5	19	AM040264.1	GI:82615033	Brucella melitensis biovar Abortus 2308 chromosome I, complete sequence, strain 2308
<i>Campylobacter jejuni</i>	2	3	2	2	9	AL111168.1	GI:30407139	Campylobacter jejuni subsp. jejuni NCTC 11168 complete genome
<i>Chlamydia trachomatis</i>	5	5	5	5	20	AM884176.1	GI:165929961	Chlamydia trachomatis strain L2/434/Bu

								complete genome
<i>Chlamydophila pneumoniae</i>	5	5	5	5	20	AE00216 1.1	GI:120572 10	Chlamydophila pneumoniae AR39, complete genome
<i>Clostridium difficile</i>	5	5	5	5	20	NC_0090 89.1	GI:126697 566	Peptoclostridium difficile 630, complete genome
<i>Clostridium perfringens</i>	5	5	5	5	20	NC_0082 61.1	GI:110798 562	Clostridium perfringens ATCC 13124, complete genome
<i>Enterobacter aerogenes</i>	5	5	5	5	20	NC_0156 63.1	GI:336246 508	Enterobacter aerogenes KCTC 2190 chromosome, complete genome
<i>Enterococcus faecalis</i>	5	5	5	5	20	AE01683 0.1	GI:293501 90	Enterococcus faecalis V583, complete genome
<i>Enterococcus faecium</i>	5	5	5	5	20	NC_0179 60.1	GI:389867 183	Enterococcus faecium DO chromosome, complete genome
<i>Francisella tularensis</i>	5	5	5	5	20	CP000803 .1	GI:156251 972	Francisella tularensis subsp. holarctica FTNF002-00, complete genome
<i>Haemophilus influenzae</i>	1	1	1	1	4	CP000057 .2	GI:156617 157	Haemophilus influenzae 86-028NP, complete genome
<i>Helicobacter pylori</i>	1	1	0	0	2	CP001173 .1	GI:208431 905	Helicobacter pylori G27, complete genome
<i>Klebsiella oxytoca</i>	5	5	5	5	20	NC_0166 12.1	GI:375256 816	Klebsiella oxytoca KCTC 1686, complete genome
<i>Legionella pneumophila</i>	5	5	5	5	20	NC_0094 94.2	GI:295815 281	Legionella pneumophila str. Corby, complete genome
<i>Listeria monocytogenes</i>	5	5	5	5	20	FM24271 1.1	GI:225875 101	Listeria monocytogenes Clip80459 serotype 4b complete genome
<i>Moraxella catarrhalis</i>	5	5	5	5	20	CP008804 .1	GI:672768 103	Moraxella catarrhalis strain 25240, complete

								genome
<i>Mycobacterium avium</i>	5	5	5	5	20	CP000479.1	GI:118163506	Mycobacterium avium 104, complete genome
<i>Mycobacterium bovis</i>	2	4	4	4	14	BX248333.1	GI:31742509	Mycobacterium bovis subsp. bovis AF2122/97 complete genome
<i>Mycoplasma genitalium</i>	5	5	5	5	20	L43967.2	GI:84626123	Mycoplasma genitalium G37, complete genome
<i>Mycoplasma pneumoniae</i>	5	5	5	5	20	NC_000912.1	GI:13507739	Mycoplasma pneumoniae M129 chromosome, complete genome
<i>Neisseria gonorrhoeae</i>	5	5	5	5	20	AE004969.1	GI:59717368	Neisseria gonorrhoeae FA 1090, complete genome
<i>Neisseria meningitidis</i>	3	5	5	5	18	AM421808.1	GI:120865607	Neisseria meningitidis serogroup C FAM18 complete genome
<i>Porphyromonas gingivalis</i>	5	5	5	5	20	NC_010729.1	GI:188993864	Porphyromonas gingivalis ATCC 33277, complete genome
<i>Proteus mirabilis</i>	5	5	5	5	20	NC_010554.1	GI:197283915	Proteus mirabilis strain HI4320, complete genome
<i>Pseudomonas aeruginosa</i>	5	5	5	5	20	NC_011770.1	GI:218888746	Pseudomonas aeruginosa LESB58 complete genome sequence
<i>Salmonella enterica</i>	3	5	4	1	13	NC_003198.1	GI:16758993	Salmonella enterica subsp. enterica serovar Typhi str. CT18, complete genome
<i>Serratia marcescens</i>	5	5	5	5	20	NC_020064.1	GI:440228796	Serratia marcescens FGI94, complete genome
<i>Staphylococcus aureus</i>	2	5	5	4	16	NC_007622.1	GI:82749777	Staphylococcus aureus RF122 complete genome
<i>Staphylococcus epidermidis</i>	5	5	5	5	20	NC_004461.1	GI:27466918	Staphylococcus epidermidis ATCC 12228 chromosome,

								complete genome
<i>Staphylococcus haemolyticus</i>	5	5	5	5	20	AP006716.1	GI:68445725	Staphylococcus haemolyticus JCSC1435 DNA, complete genome
<i>Stenotrophomonas maltophilia</i>	0	1	1	1	3	NC_010943.1	GI:190572091	Stenotrophomonas maltophilia K279a complete genome, strain K279a
<i>Streptococcus mutans</i>	5	5	5	5	20	AP010655.1	GI:254996425	Streptococcus mutans NN2025 DNA, complete genome
<i>Streptococcus pyogenes</i>	5	5	5	5	20	NC_018936.1	GI:409913960	Streptococcus pyogenes A20, complete genome
<i>Streptococcus salivarius</i>	5	5	5	5	20	NC_015760.1	GI:340397867	Streptococcus salivarius CCHSS3 complete genome
<i>Streptococcus sanguinis</i>	5	5	5	5	20	CP000387.1	GI:125496804	Streptococcus sanguinis SK36, complete genome
<i>Treponema pallidum</i>	5	5	5	5	20	NC_016844.1	GI:378974633	Treponema pallidum subsp. pallidum DAL-1, complete genome
<i>Vibrio cholerae</i>	0	1	1	0	2	NC_012578.1 / NC_012580.1	GI:227080237 / GI:227811634	Vibrio cholerae M66-2 chromosome I, complete sequence / Vibrio cholerae M66-2 chromosome II, complete sequence
<i>Vibrio parahaemolyticus</i>	5	5	5	5	20	NC_004603.1	GI:28896774	Vibrio parahaemolyticus RIMD 2210633 chromosome 1, complete sequence
<i>Yersinia enterocolitica</i>	5	5	5	5	20	AM286415.1	GI:122087364	Yersinia enterocolitica subsp. enterocolitica 8081 complete genome
Grand Total	198	211	208	203	820			

K-mer Cluster Summary

Each k-mer cluster is listed in the sequence listing (SEQ ID NOs 1 to 213).

5 Preferred K-mer Summary (Table 3)

Preferred k-mers are listed in the Table below with their sequence length, the k-mer cluster sequence from which they are derived and their location within that sequence. For instance, the first preferred k-mer is located at 1/18-34. This means it corresponds to nucleotides 18 to 34 of SEQ ID NO: 1. The second preferred k-mer is located at 2/9-25. This means it corresponds to nucleotides 9 to 25 of SEQ ID NO: 2 and so on.

Species	k-mer Length	k-mer derived from SEQ ID NO:	k-mer location in relevant SEQ ID NO:
<i>Aggregatibacter actinomycetemcomitans</i>	17	1	1/18-34
<i>Aggregatibacter actinomycetemcomitans</i>	17	2	2/9-25
<i>Aggregatibacter actinomycetemcomitans</i>	17	3	3/26-42
<i>Aggregatibacter actinomycetemcomitans</i>	17	4	4/13-29
<i>Aggregatibacter actinomycetemcomitans</i>	17	5	5/10-26
<i>Bacillus anthracis</i>	17	6	6/21-37
<i>Bacillus anthracis</i>	17	7	7/21-37
<i>Bacillus anthracis</i>	17	8	8/15-31
<i>Bacillus anthracis</i>	17	9	9/22-38
<i>Bacillus anthracis</i>	17	10	10/14-30
<i>Bacillus licheniformis</i>	17	11	11/19-35
<i>Bacillus licheniformis</i>	17	12	12/11-27
<i>Bacillus licheniformis</i>	17	13	13/15-31
<i>Bacillus licheniformis</i>	17	14	14/19-35
<i>Bacillus licheniformis</i>	17	15	15/40-56
<i>Bacteroides fragilis</i>	17	16	16/28-44
<i>Bacteroides fragilis</i>	17	17	17/14-30
<i>Bacteroides fragilis</i>	17	18	18/14-30
<i>Bacteroides fragilis</i>	17	19	19/29-45
<i>Bacteroides fragilis</i>	17	20	20/26-42
<i>Bartonella henselae</i>	17	21	21/134-150
<i>Bartonella henselae</i>	17	22	22/21-37
<i>Bartonella henselae</i>	17	23	23/17-33
<i>Bartonella henselae</i>	17	24	24/108-124
<i>Bartonella henselae</i>	17	25	25/8-24
<i>Bordetella pertussis</i>	17	26	26/22-38
<i>Bordetella pertussis</i>	17	27	27/18-34
<i>Bordetella pertussis</i>	17	28	28/21-37
<i>Bordetella pertussis</i>	17	29	29/25-41
<i>Bordetella pertussis</i>	17	30	30/22-38

<i>Borrelia burgdorferi</i>	17	31	31/12-28
<i>Borrelia burgdorferi</i>	17	32	32/10-26
<i>Borrelia burgdorferi</i>	17	33	33/28-44
<i>Borrelia burgdorferi</i>	17	34	34/16-32
<i>Borrelia burgdorferi</i>	17	35	35/34-50
<i>Brucella abortus</i>	17	36	36/28-44
<i>Brucella abortus</i>	17	37	37/23-39
<i>Brucella abortus</i>	17	38	38/19-35
<i>Brucella abortus</i>	17	40	40/22-38
<i>Campylobacter jejuni</i>	17	41	41/4-20
<i>Campylobacter jejuni</i>	17	43	43/7-23
<i>Chlamydia trachomatis</i>	17	45	45/16-32
<i>Chlamydia trachomatis</i>	17	46	46/9-25
<i>Chlamydia trachomatis</i>	17	47	47/13-29
<i>Chlamydia trachomatis</i>	17	48	48/73-89
<i>Chlamydia trachomatis</i>	17	49	49/40-56
<i>Chlamydophila pneumoniae</i>	17	50	50/31-47
<i>Chlamydophila pneumoniae</i>	17	51	51/28-44
<i>Chlamydophila pneumoniae</i>	17	52	52/120-136
<i>Chlamydophila pneumoniae</i>	17	53	53/17-33
<i>Chlamydophila pneumoniae</i>	17	54	54/91-107
<i>Clostridium difficile</i>	17	55	55/12-28
<i>Clostridium difficile</i>	17	56	56/43-59
<i>Clostridium difficile</i>	17	57	57/24-40
<i>Clostridium difficile</i>	17	58	58/9-25
<i>Clostridium difficile</i>	17	59	59/28-44
<i>Clostridium perfringens</i>	17	60	60/17-33
<i>Clostridium perfringens</i>	17	61	61/48-64
<i>Clostridium perfringens</i>	17	62	62/162-178
<i>Clostridium perfringens</i>	17	63	63/23-39
<i>Clostridium perfringens</i>	17	64	64/66-82
<i>Enterobacter aerogenes</i>	17	65	65/24-40
<i>Enterobacter aerogenes</i>	17	66	66/23-39
<i>Enterobacter aerogenes</i>	17	67	67/27-43
<i>Enterobacter aerogenes</i>	17	68	68/24-40
<i>Enterobacter aerogenes</i>	17	69	69/23-39
<i>Enterococcus faecalis</i>	17	70	70/11-27
<i>Enterococcus faecalis</i>	17	71	71/12-28
<i>Enterococcus faecalis</i>	17	72	72/18-34
<i>Enterococcus faecalis</i>	17	73	73/25-41
<i>Enterococcus faecalis</i>	17	74	74/26-42
<i>Enterococcus faecium</i>	17	75	75/9-25
<i>Enterococcus faecium</i>	17	76	76/32-48
<i>Enterococcus faecium</i>	17	77	77/11-27
<i>Enterococcus faecium</i>	17	78	78/6-22
<i>Enterococcus faecium</i>	17	79	79/19-35

<i>Francisella tularensis</i>	17	80	80/16-32
<i>Francisella tularensis</i>	17	81	81/11-27
<i>Francisella tularensis</i>	17	82	82/9-25
<i>Francisella tularensis</i>	17	83	83/15-31
<i>Francisella tularensis</i>	17	84	84/17-33
<i>Haemophilus influenzae</i>	17	85	85/30-46
<i>Helicobacter pylori</i>	17	87	87/1-17
<i>Klebsiella oxytoca</i>	17	88	88/14-30
<i>Klebsiella oxytoca</i>	17	89	89/10-26
<i>Klebsiella oxytoca</i>	17	90	90/23-39
<i>Klebsiella oxytoca</i>	17	91	91/28-44
<i>Klebsiella oxytoca</i>	17	92	92/17-33
<i>Legionella pneumophila</i>	17	93	93/7-23
<i>Legionella pneumophila</i>	17	94	94/14-30
<i>Legionella pneumophila</i>	17	95	95/30-46
<i>Legionella pneumophila</i>	17	96	96/11-27
<i>Legionella pneumophila</i>	17	97	97/14-30
<i>Listeria monocytogenes</i>	17	98	98/7-23
<i>Listeria monocytogenes</i>	17	99	99/8-24
<i>Listeria monocytogenes</i>	17	100	100/22-38
<i>Listeria monocytogenes</i>	17	101	101/9-25
<i>Listeria monocytogenes</i>	17	102	102/22-38
<i>Moraxella catarrhalis</i>	17	103	103/31-47
<i>Moraxella catarrhalis</i>	17	104	104/37-53
<i>Moraxella catarrhalis</i>	17	105	105/95-111
<i>Moraxella catarrhalis</i>	17	106	106/62-78
<i>Moraxella catarrhalis</i>	17	107	107/22-38
<i>Mycobacterium avium</i>	17	108	108/17-33
<i>Mycobacterium avium</i>	17	109	109/19-35
<i>Mycobacterium avium</i>	17	110	110/6-22
<i>Mycobacterium avium</i>	17	111	111/50-66
<i>Mycobacterium avium</i>	17	112	112/22-38
<i>Mycobacterium bovis</i>	17	115	115/23-39
<i>Mycobacterium bovis</i>	17	116	116/21-37
<i>Mycoplasma genitalium</i>	17	117	117/8-24
<i>Mycoplasma genitalium</i>	17	118	118/17-33
<i>Mycoplasma genitalium</i>	17	119	119/110-126
<i>Mycoplasma genitalium</i>	17	120	120/33-49
<i>Mycoplasma genitalium</i>	17	121	121/132-148
<i>Mycoplasma pneumoniae</i>	17	122	122/26-42
<i>Mycoplasma pneumoniae</i>	17	123	123/38-54
<i>Mycoplasma pneumoniae</i>	17	124	124/56-72
<i>Mycoplasma pneumoniae</i>	17	125	125/70-86
<i>Mycoplasma pneumoniae</i>	17	126	126/14-30
<i>Neisseria gonorrhoeae</i>	17	127	127/12-28
<i>Neisseria gonorrhoeae</i>	17	128	128/19-35

<i>Neisseria gonorrhoeae</i>	17	129	129/18-34
<i>Neisseria gonorrhoeae</i>	17	130	130/12-28
<i>Neisseria gonorrhoeae</i>	17	131	131/14-30
<i>Neisseria meningitidis</i>	17	134	134/36-52
<i>Neisseria meningitidis</i>	17	135	135/21-37
<i>Neisseria meningitidis</i>	17	136	136/16-32
<i>Porphyromonas gingivalis</i>	17	137	137/73-89
<i>Porphyromonas gingivalis</i>	17	138	138/18-34
<i>Porphyromonas gingivalis</i>	17	139	139/8-24
<i>Porphyromonas gingivalis</i>	17	140	140/97-113
<i>Porphyromonas gingivalis</i>	17	141	141/57-73
<i>Proteus mirabilis</i>	17	142	142/16-32
<i>Proteus mirabilis</i>	17	143	143/53-69
<i>Proteus mirabilis</i>	17	144	144/61-77
<i>Proteus mirabilis</i>	17	145	145/25-41
<i>Proteus mirabilis</i>	17	146	146/26-42
<i>Pseudomonas aeruginosa</i>	17	147	147/9-25
<i>Pseudomonas aeruginosa</i>	17	148	148/13-29
<i>Pseudomonas aeruginosa</i>	17	149	149/18-34
<i>Pseudomonas aeruginosa</i>	17	150	150/13-29
<i>Pseudomonas aeruginosa</i>	17	151	151/23-39
<i>Salmonella enterica</i>	17	153	153/3-19
<i>Salmonella enterica</i>	17	154	154/2-18
<i>Salmonella enterica</i>	17	156	156/9-25
<i>Serratia marcescens</i>	17	157	157/11-27
<i>Serratia marcescens</i>	17	158	158/17-33
<i>Serratia marcescens</i>	17	159	159/14-30
<i>Serratia marcescens</i>	17	160	160/10-26
<i>Serratia marcescens</i>	17	161	161/28-44
<i>Staphylococcus aureus</i>	17	163	163/7-23
<i>Staphylococcus aureus</i>	17	165	165/7-23
<i>Staphylococcus epidermidis</i>	17	167	167/13-29
<i>Staphylococcus epidermidis</i>	17	168	168/21-37
<i>Staphylococcus epidermidis</i>	17	169	169/28-44
<i>Staphylococcus epidermidis</i>	17	170	170/10-26
<i>Staphylococcus epidermidis</i>	17	171	171/23-39
<i>Staphylococcus haemolyticus</i>	17	172	172/18-34
<i>Staphylococcus haemolyticus</i>	17	173	173/9-25
<i>Staphylococcus haemolyticus</i>	17	174	174/23-39
<i>Staphylococcus haemolyticus</i>	17	175	175/19-35
<i>Staphylococcus haemolyticus</i>	17	176	176/6-22
<i>Streptococcus mutans</i>	17	178	178/37-53
<i>Streptococcus mutans</i>	17	179	179/7-23
<i>Streptococcus mutans</i>	17	180	180/10-26
<i>Streptococcus mutans</i>	17	181	181/13-29
<i>Streptococcus mutans</i>	17	182	182/16-32

<i>Streptococcus pyogenes</i>	17	183	183/18-34
<i>Streptococcus pyogenes</i>	17	184	184/13-29
<i>Streptococcus pyogenes</i>	17	185	185/9-25
<i>Streptococcus pyogenes</i>	17	186	186/63-79
<i>Streptococcus pyogenes</i>	17	187	187/9-25
<i>Streptococcus salivarius</i>	17	188	188/11-27
<i>Streptococcus salivarius</i>	17	189	189/27-43
<i>Streptococcus salivarius</i>	17	190	190/19-35
<i>Streptococcus salivarius</i>	17	191	191/13-29
<i>Streptococcus salivarius</i>	17	192	192/5-21
<i>Streptococcus sanguinis</i>	17	193	193/8-24
<i>Streptococcus sanguinis</i>	17	194	194/9-25
<i>Streptococcus sanguinis</i>	17	195	195/37-53
<i>Streptococcus sanguinis</i>	17	196	196/14-30
<i>Streptococcus sanguinis</i>	17	197	197/20-36
<i>Treponema pallidum</i>	17	198	198/45-61
<i>Treponema pallidum</i>	17	199	199/34-50
<i>Treponema pallidum</i>	17	200	200/30-46
<i>Treponema pallidum</i>	17	201	201/28-44
<i>Treponema pallidum</i>	17	202	202/22-38
<i>Vibrio parahaemolyticus</i>	17	204	204/25-41
<i>Vibrio parahaemolyticus</i>	17	205	205/56-72
<i>Vibrio parahaemolyticus</i>	17	206	206/18-34
<i>Vibrio parahaemolyticus</i>	17	207	207/14-30
<i>Vibrio parahaemolyticus</i>	17	208	208/23-39
<i>Yersinia enterocolitica</i>	17	209	209/15-31
<i>Yersinia enterocolitica</i>	17	210	210/25-41
<i>Yersinia enterocolitica</i>	17	211	211/10-26
<i>Yersinia enterocolitica</i>	17	212	212/14-30
<i>Yersinia enterocolitica</i>	17	213	213/19-35
<i>Aggregatibacter actinomycetemcomitans</i>	20	1	1/15-34
<i>Aggregatibacter actinomycetemcomitans</i>	20	2	2/8-27
<i>Aggregatibacter actinomycetemcomitans</i>	20	3	3/26-45
<i>Aggregatibacter actinomycetemcomitans</i>	20	4	4/11-30
<i>Aggregatibacter actinomycetemcomitans</i>	20	5	5/11-30
<i>Bacillus anthracis</i>	20	6	6/19-38
<i>Bacillus anthracis</i>	20	7	7/20-39
<i>Bacillus anthracis</i>	20	8	8/20-39
<i>Bacillus anthracis</i>	20	9	9/21-40
<i>Bacillus anthracis</i>	20	10	10/11-30
<i>Bacillus licheniformis</i>	20	11	11/17-36
<i>Bacillus licheniformis</i>	20	12	12/10-29
<i>Bacillus licheniformis</i>	20	13	13/13-32
<i>Bacillus licheniformis</i>	20	14	14/18-37
<i>Bacillus licheniformis</i>	20	15	15/38-57
<i>Bacteroides fragilis</i>	20	16	16/27-46

<i>Bacteroides fragilis</i>	20	17	17/23-42
<i>Bacteroides fragilis</i>	20	18	18/11-30
<i>Bacteroides fragilis</i>	20	19	19/27-46
<i>Bacteroides fragilis</i>	20	20	20/24-43
<i>Bartonella henselae</i>	20	21	21/131-150
<i>Bartonella henselae</i>	20	22	22/21-40
<i>Bartonella henselae</i>	20	23	23/15-34
<i>Bartonella henselae</i>	20	24	24/106-125
<i>Bartonella henselae</i>	20	25	25/7-26
<i>Bordetella pertussis</i>	20	26	26/21-40
<i>Bordetella pertussis</i>	20	27	27/21-40
<i>Bordetella pertussis</i>	20	28	28/17-36
<i>Bordetella pertussis</i>	20	29	29/21-40
<i>Bordetella pertussis</i>	20	30	30/21-40
<i>Borrelia burgdorferi</i>	20	31	31/13-32
<i>Borrelia burgdorferi</i>	20	32	32/8-27
<i>Borrelia burgdorferi</i>	20	33	33/26-45
<i>Borrelia burgdorferi</i>	20	34	34/20-39
<i>Borrelia burgdorferi</i>	20	35	35/33-52
<i>Brucella abortus</i>	20	36	36/17-36
<i>Brucella abortus</i>	20	37	37/21-40
<i>Brucella abortus</i>	20	38	38/20-39
<i>Brucella abortus</i>	20	39	39/20-39
<i>Brucella abortus</i>	20	40	40/21-40
<i>Campylobacter jejuni</i>	20	41	41/2-21
<i>Campylobacter jejuni</i>	20	43	43/3-22
<i>Campylobacter jejuni</i>	20	44	44/16-35
<i>Chlamydia trachomatis</i>	20	45	45/14-33
<i>Chlamydia trachomatis</i>	20	46	46/7-26
<i>Chlamydia trachomatis</i>	20	47	47/11-30
<i>Chlamydia trachomatis</i>	20	48	48/71-90
<i>Chlamydia trachomatis</i>	20	49	49/39-58
<i>Chlamydophila pneumoniae</i>	20	50	50/29-48
<i>Chlamydophila pneumoniae</i>	20	51	51/28-47
<i>Chlamydophila pneumoniae</i>	20	52	52/118-137
<i>Chlamydophila pneumoniae</i>	20	53	53/16-35
<i>Chlamydophila pneumoniae</i>	20	54	54/89-108
<i>Clostridium difficile</i>	20	55	55/11-30
<i>Clostridium difficile</i>	20	56	56/37-56
<i>Clostridium difficile</i>	20	57	57/21-40
<i>Clostridium difficile</i>	20	58	58/7-26
<i>Clostridium difficile</i>	20	59	59/25-44
<i>Clostridium perfringens</i>	20	60	60/15-34
<i>Clostridium perfringens</i>	20	61	61/47-66
<i>Clostridium perfringens</i>	20	62	62/161-180
<i>Clostridium perfringens</i>	20	63	63/14-33

<i>Clostridium perfringens</i>	20	64	64/64-83
<i>Enterobacter aerogenes</i>	20	65	65/24-43
<i>Enterobacter aerogenes</i>	20	66	66/21-40
<i>Enterobacter aerogenes</i>	20	67	67/40-59
<i>Enterobacter aerogenes</i>	20	68	68/23-42
<i>Enterobacter aerogenes</i>	20	69	69/21-40
<i>Enterococcus faecalis</i>	20	70	70/9-28
<i>Enterococcus faecalis</i>	20	71	71/9-28
<i>Enterococcus faecalis</i>	20	72	72/17-36
<i>Enterococcus faecalis</i>	20	73	73/19-38
<i>Enterococcus faecalis</i>	20	74	74/26-45
<i>Enterococcus faecium</i>	20	75	75/7-26
<i>Enterococcus faecium</i>	20	76	76/30-49
<i>Enterococcus faecium</i>	20	77	77/10-29
<i>Enterococcus faecium</i>	20	78	78/5-24
<i>Enterococcus faecium</i>	20	79	79/17-36
<i>Francisella tularensis</i>	20	80	80/17-36
<i>Francisella tularensis</i>	20	81	81/9-28
<i>Francisella tularensis</i>	20	82	82/7-26
<i>Francisella tularensis</i>	20	83	83/14-33
<i>Francisella tularensis</i>	20	84	84/15-34
<i>Haemophilus influenzae</i>	20	85	85/22-41
<i>Helicobacter pylori</i>	20	86	86/1-20
<i>Klebsiella oxytoca</i>	20	88	88/13-32
<i>Klebsiella oxytoca</i>	20	89	89/11-30
<i>Klebsiella oxytoca</i>	20	90	90/22-41
<i>Klebsiella oxytoca</i>	20	91	91/21-40
<i>Klebsiella oxytoca</i>	20	92	92/13-32
<i>Legionella pneumophila</i>	20	93	93/6-25
<i>Legionella pneumophila</i>	20	94	94/12-31
<i>Legionella pneumophila</i>	20	95	95/29-48
<i>Legionella pneumophila</i>	20	96	96/8-27
<i>Legionella pneumophila</i>	20	97	97/10-29
<i>Listeria monocytogenes</i>	20	98	98/8-27
<i>Listeria monocytogenes</i>	20	99	99/7-26
<i>Listeria monocytogenes</i>	20	100	100/20-39
<i>Listeria monocytogenes</i>	20	101	101/8-27
<i>Listeria monocytogenes</i>	20	102	102/20-39
<i>Moraxella catarrhalis</i>	20	103	103/30-49
<i>Moraxella catarrhalis</i>	20	104	104/35-54
<i>Moraxella catarrhalis</i>	20	105	105/93-112
<i>Moraxella catarrhalis</i>	20	106	106/59-78
<i>Moraxella catarrhalis</i>	20	107	107/14-33
<i>Mycobacterium avium</i>	20	108	108/17-36
<i>Mycobacterium avium</i>	20	109	109/14-33
<i>Mycobacterium avium</i>	20	110	110/7-26

<i>Mycobacterium avium</i>	20	111	111/27-46
<i>Mycobacterium avium</i>	20	112	112/18-37
<i>Mycobacterium bovis</i>	20	113	113/21-40
<i>Mycobacterium bovis</i>	20	114	114/21-40
<i>Mycobacterium bovis</i>	20	115	115/21-40
<i>Mycobacterium bovis</i>	20	116	116/20-39
<i>Mycoplasma genitalium</i>	20	117	117/6-25
<i>Mycoplasma genitalium</i>	20	118	118/15-34
<i>Mycoplasma genitalium</i>	20	119	119/109-128
<i>Mycoplasma genitalium</i>	20	120	120/28-47
<i>Mycoplasma genitalium</i>	20	121	121/128-147
<i>Mycoplasma pneumoniae</i>	20	122	122/23-42
<i>Mycoplasma pneumoniae</i>	20	123	123/36-55
<i>Mycoplasma pneumoniae</i>	20	124	124/54-73
<i>Mycoplasma pneumoniae</i>	20	125	125/69-88
<i>Mycoplasma pneumoniae</i>	20	126	126/15-34
<i>Neisseria gonorrhoeae</i>	20	127	127/8-27
<i>Neisseria gonorrhoeae</i>	20	128	128/18-37
<i>Neisseria gonorrhoeae</i>	20	129	129/16-35
<i>Neisseria gonorrhoeae</i>	20	130	130/14-33
<i>Neisseria gonorrhoeae</i>	20	131	131/13-32
<i>Neisseria meningitidis</i>	20	132	132/18-37
<i>Neisseria meningitidis</i>	20	133	133/13-32
<i>Neisseria meningitidis</i>	20	134	134/32-51
<i>Neisseria meningitidis</i>	20	135	135/19-38
<i>Neisseria meningitidis</i>	20	136	136/16-35
<i>Porphyromonas gingivalis</i>	20	137	137/71-90
<i>Porphyromonas gingivalis</i>	20	138	138/17-36
<i>Porphyromonas gingivalis</i>	20	139	139/7-26
<i>Porphyromonas gingivalis</i>	20	140	140/96-115
<i>Porphyromonas gingivalis</i>	20	141	141/56-75
<i>Proteus mirabilis</i>	20	142	142/15-34
<i>Proteus mirabilis</i>	20	143	143/51-70
<i>Proteus mirabilis</i>	20	144	144/55-74
<i>Proteus mirabilis</i>	20	145	145/24-43
<i>Proteus mirabilis</i>	20	146	146/24-43
<i>Pseudomonas aeruginosa</i>	20	147	147/7-26
<i>Pseudomonas aeruginosa</i>	20	148	148/10-29
<i>Pseudomonas aeruginosa</i>	20	149	149/16-35
<i>Pseudomonas aeruginosa</i>	20	150	150/12-31
<i>Pseudomonas aeruginosa</i>	20	151	151/23-42
<i>Salmonella enterica</i>	20	152	152/11-30
<i>Salmonella enterica</i>	20	153	153/1-20
<i>Salmonella enterica</i>	20	154	154/2-21
<i>Salmonella enterica</i>	20	155	155/6-25
<i>Salmonella enterica</i>	20	156	156/6-25

<i>Serratia marcescens</i>	20	157	157/15-34
<i>Serratia marcescens</i>	20	158	158/15-34
<i>Serratia marcescens</i>	20	159	159/21-40
<i>Serratia marcescens</i>	20	160	160/11-30
<i>Serratia marcescens</i>	20	161	161/21-40
<i>Staphylococcus aureus</i>	20	162	162/8-27
<i>Staphylococcus aureus</i>	20	163	163/6-25
<i>Staphylococcus aureus</i>	20	164	164/4-23
<i>Staphylococcus aureus</i>	20	165	165/6-25
<i>Staphylococcus aureus</i>	20	166	166/8-27
<i>Staphylococcus epidermidis</i>	20	167	167/11-30
<i>Staphylococcus epidermidis</i>	20	168	168/20-39
<i>Staphylococcus epidermidis</i>	20	169	169/26-45
<i>Staphylococcus epidermidis</i>	20	170	170/8-27
<i>Staphylococcus epidermidis</i>	20	171	171/21-40
<i>Staphylococcus haemolyticus</i>	20	172	172/16-35
<i>Staphylococcus haemolyticus</i>	20	173	173/12-31
<i>Staphylococcus haemolyticus</i>	20	174	174/18-37
<i>Staphylococcus haemolyticus</i>	20	175	175/20-39
<i>Staphylococcus haemolyticus</i>	20	176	176/6-25
<i>Stenotrophomonas maltophilia</i>	20	177	177/21-40
<i>Streptococcus mutans</i>	20	178	178/36-55
<i>Streptococcus mutans</i>	20	179	179/5-24
<i>Streptococcus mutans</i>	20	180	180/8-27
<i>Streptococcus mutans</i>	20	181	181/11-30
<i>Streptococcus mutans</i>	20	182	182/14-33
<i>Streptococcus pyogenes</i>	20	183	183/17-36
<i>Streptococcus pyogenes</i>	20	184	184/15-34
<i>Streptococcus pyogenes</i>	20	185	185/7-26
<i>Streptococcus pyogenes</i>	20	186	186/52-71
<i>Streptococcus pyogenes</i>	20	187	187/10-29
<i>Streptococcus salivarius</i>	20	188	188/9-28
<i>Streptococcus salivarius</i>	20	189	189/17-36
<i>Streptococcus salivarius</i>	20	190	190/16-35
<i>Streptococcus salivarius</i>	20	191	191/11-30
<i>Streptococcus salivarius</i>	20	192	192/6-25
<i>Streptococcus sanguinis</i>	20	193	193/7-26
<i>Streptococcus sanguinis</i>	20	194	194/7-26
<i>Streptococcus sanguinis</i>	20	195	195/38-57
<i>Streptococcus sanguinis</i>	20	196	196/15-34
<i>Streptococcus sanguinis</i>	20	197	197/16-35
<i>Treponema pallidum</i>	20	198	198/38-57
<i>Treponema pallidum</i>	20	199	199/29-48
<i>Treponema pallidum</i>	20	200	200/30-49
<i>Treponema pallidum</i>	20	201	201/26-45
<i>Treponema pallidum</i>	20	202	202/20-39

<i>Vibrio cholerae</i>	20	203	203/6-25
<i>Vibrio parahaemolyticus</i>	20	204	204/20-39
<i>Vibrio parahaemolyticus</i>	20	205	205/47-66
<i>Vibrio parahaemolyticus</i>	20	206	206/14-33
<i>Vibrio parahaemolyticus</i>	20	207	207/13-32
<i>Vibrio parahaemolyticus</i>	20	208	208/16-35
<i>Yersinia enterocolitica</i>	20	209	209/12-31
<i>Yersinia enterocolitica</i>	20	210	210/23-42
<i>Yersinia enterocolitica</i>	20	211	211/12-31
<i>Yersinia enterocolitica</i>	20	212	212/16-35
<i>Yersinia enterocolitica</i>	20	213	213/22-41
<i>Aggregatibacter actinomycetemcomitans</i>	25	1	1/12-36
<i>Aggregatibacter actinomycetemcomitans</i>	25	2	2/5-29
<i>Aggregatibacter actinomycetemcomitans</i>	25	3	3/25-49
<i>Aggregatibacter actinomycetemcomitans</i>	25	4	4/8-32
<i>Aggregatibacter actinomycetemcomitans</i>	25	5	5/9-33
<i>Bacillus anthracis</i>	25	6	6/17-41
<i>Bacillus anthracis</i>	25	7	7/17-41
<i>Bacillus anthracis</i>	25	8	8/18-42
<i>Bacillus anthracis</i>	25	9	9/18-42
<i>Bacillus anthracis</i>	25	10	10/8-32
<i>Bacillus licheniformis</i>	25	11	11/15-39
<i>Bacillus licheniformis</i>	25	12	12/7-31
<i>Bacillus licheniformis</i>	25	13	13/11-35
<i>Bacillus licheniformis</i>	25	14	14/16-40
<i>Bacillus licheniformis</i>	25	15	15/34-58
<i>Bacteroides fragilis</i>	25	16	16/27-51
<i>Bacteroides fragilis</i>	25	17	17/25-49
<i>Bacteroides fragilis</i>	25	18	18/9-33
<i>Bacteroides fragilis</i>	25	19	19/24-48
<i>Bacteroides fragilis</i>	25	20	20/21-45
<i>Bartonella henselae</i>	25	21	21/128-152
<i>Bartonella henselae</i>	25	22	22/18-42
<i>Bartonella henselae</i>	25	23	23/12-36
<i>Bartonella henselae</i>	25	24	24/99-123
<i>Bartonella henselae</i>	25	25	25/4-28
<i>Bordetella pertussis</i>	25	26	26/18-42
<i>Bordetella pertussis</i>	25	27	27/18-42
<i>Bordetella pertussis</i>	25	28	28/15-39
<i>Bordetella pertussis</i>	25	29	29/18-42
<i>Bordetella pertussis</i>	25	30	30/18-42
<i>Borrelia burgdorferi</i>	25	31	31/11-35
<i>Borrelia burgdorferi</i>	25	32	32/5-29
<i>Borrelia burgdorferi</i>	25	33	33/23-47
<i>Borrelia burgdorferi</i>	25	34	34/19-43
<i>Borrelia burgdorferi</i>	25	35	35/32-56

<i>Brucella abortus</i>	25	36	36/15-39
<i>Brucella abortus</i>	25	37	37/18-42
<i>Brucella abortus</i>	25	38	38/17-41
<i>Brucella abortus</i>	25	39	39/18-42
<i>Brucella abortus</i>	25	40	40/18-42
<i>Campylobacter jejuni</i>	25	42	42/6-30
<i>Campylobacter jejuni</i>	25	44	44/14-38
<i>Chlamydia trachomatis</i>	25	45	45/11-35
<i>Chlamydia trachomatis</i>	25	46	46/5-29
<i>Chlamydia trachomatis</i>	25	47	47/9-33
<i>Chlamydia trachomatis</i>	25	48	48/70-94
<i>Chlamydia trachomatis</i>	25	49	49/37-61
<i>Chlamydophila pneumoniae</i>	25	50	50/28-52
<i>Chlamydophila pneumoniae</i>	25	51	51/28-52
<i>Chlamydophila pneumoniae</i>	25	52	52/116-140
<i>Chlamydophila pneumoniae</i>	25	53	53/14-38
<i>Chlamydophila pneumoniae</i>	25	54	54/85-109
<i>Clostridium difficile</i>	25	55	55/8-32
<i>Clostridium difficile</i>	25	56	56/33-57
<i>Clostridium difficile</i>	25	57	57/18-42
<i>Clostridium difficile</i>	25	58	58/5-29
<i>Clostridium difficile</i>	25	59	59/22-46
<i>Clostridium perfringens</i>	25	60	60/13-37
<i>Clostridium perfringens</i>	25	61	61/42-66
<i>Clostridium perfringens</i>	25	62	62/158-182
<i>Clostridium perfringens</i>	25	63	63/11-35
<i>Clostridium perfringens</i>	25	64	64/60-84
<i>Enterobacter aerogenes</i>	25	65	65/25-49
<i>Enterobacter aerogenes</i>	25	66	66/18-42
<i>Enterobacter aerogenes</i>	25	67	67/34-58
<i>Enterobacter aerogenes</i>	25	68	68/21-45
<i>Enterobacter aerogenes</i>	25	69	69/18-42
<i>Enterococcus faecalis</i>	25	70	70/6-30
<i>Enterococcus faecalis</i>	25	71	71/6-30
<i>Enterococcus faecalis</i>	25	72	72/15-39
<i>Enterococcus faecalis</i>	25	73	73/17-41
<i>Enterococcus faecalis</i>	25	74	74/25-49
<i>Enterococcus faecium</i>	25	75	75/5-29
<i>Enterococcus faecium</i>	25	76	76/28-52
<i>Enterococcus faecium</i>	25	77	77/7-31
<i>Enterococcus faecium</i>	25	78	78/3-27
<i>Enterococcus faecium</i>	25	79	79/14-38
<i>Francisella tularensis</i>	25	80	80/15-39
<i>Francisella tularensis</i>	25	81	81/6-30
<i>Francisella tularensis</i>	25	82	82/5-29
<i>Francisella tularensis</i>	25	83	83/11-35

<i>Francisella tularensis</i>	25	84	84/13-37
<i>Haemophilus influenzae</i>	25	85	85/17-41
<i>Klebsiella oxytoca</i>	25	88	88/11-35
<i>Klebsiella oxytoca</i>	25	89	89/8-32
<i>Klebsiella oxytoca</i>	25	90	90/19-43
<i>Klebsiella oxytoca</i>	25	91	91/19-43
<i>Klebsiella oxytoca</i>	25	92	92/10-34
<i>Legionella pneumophila</i>	25	93	93/4-28
<i>Legionella pneumophila</i>	25	94	94/9-33
<i>Legionella pneumophila</i>	25	95	95/27-51
<i>Legionella pneumophila</i>	25	96	96/6-30
<i>Legionella pneumophila</i>	25	97	97/7-31
<i>Listeria monocytogenes</i>	25	98	98/6-30
<i>Listeria monocytogenes</i>	25	99	99/5-29
<i>Listeria monocytogenes</i>	25	100	100/18-42
<i>Listeria monocytogenes</i>	25	101	101/6-30
<i>Listeria monocytogenes</i>	25	102	102/18-42
<i>Moraxella catarrhalis</i>	25	103	103/29-53
<i>Moraxella catarrhalis</i>	25	104	104/32-56
<i>Moraxella catarrhalis</i>	25	105	105/89-113
<i>Moraxella catarrhalis</i>	25	106	106/56-80
<i>Moraxella catarrhalis</i>	25	107	107/12-36
<i>Mycobacterium avium</i>	25	108	108/14-38
<i>Mycobacterium avium</i>	25	109	109/11-35
<i>Mycobacterium avium</i>	25	110	110/4-28
<i>Mycobacterium avium</i>	25	111	111/25-49
<i>Mycobacterium avium</i>	25	112	112/16-40
<i>Mycobacterium bovis</i>	25	113	113/19-43
<i>Mycobacterium bovis</i>	25	114	114/19-43
<i>Mycobacterium bovis</i>	25	115	115/18-42
<i>Mycobacterium bovis</i>	25	116	116/18-42
<i>Mycoplasma genitalium</i>	25	117	117/4-28
<i>Mycoplasma genitalium</i>	25	118	118/12-36
<i>Mycoplasma genitalium</i>	25	119	119/105-129
<i>Mycoplasma genitalium</i>	25	120	120/23-47
<i>Mycoplasma genitalium</i>	25	121	121/121-145
<i>Mycoplasma pneumoniae</i>	25	122	122/21-45
<i>Mycoplasma pneumoniae</i>	25	123	123/31-55
<i>Mycoplasma pneumoniae</i>	25	124	124/50-74
<i>Mycoplasma pneumoniae</i>	25	125	125/70-94
<i>Mycoplasma pneumoniae</i>	25	126	126/13-37
<i>Neisseria gonorrhoeae</i>	25	127	127/5-29
<i>Neisseria gonorrhoeae</i>	25	128	128/16-40
<i>Neisseria gonorrhoeae</i>	25	129	129/14-38
<i>Neisseria gonorrhoeae</i>	25	130	130/12-36
<i>Neisseria gonorrhoeae</i>	25	131	131/10-34

<i>Neisseria meningitidis</i>	25	132	132/15-39
<i>Neisseria meningitidis</i>	25	133	133/11-35
<i>Neisseria meningitidis</i>	25	134	134/29-53
<i>Neisseria meningitidis</i>	25	135	135/17-41
<i>Neisseria meningitidis</i>	25	136	136/13-37
<i>Porphyromonas gingivalis</i>	25	137	137/70-94
<i>Porphyromonas gingivalis</i>	25	138	138/14-38
<i>Porphyromonas gingivalis</i>	25	139	139/4-28
<i>Porphyromonas gingivalis</i>	25	140	140/94-118
<i>Porphyromonas gingivalis</i>	25	141	141/58-82
<i>Proteus mirabilis</i>	25	142	142/12-36
<i>Proteus mirabilis</i>	25	143	143/45-69
<i>Proteus mirabilis</i>	25	144	144/53-77
<i>Proteus mirabilis</i>	25	145	145/22-46
<i>Proteus mirabilis</i>	25	146	146/21-45
<i>Pseudomonas aeruginosa</i>	25	147	147/5-29
<i>Pseudomonas aeruginosa</i>	25	148	148/6-30
<i>Pseudomonas aeruginosa</i>	25	149	149/14-38
<i>Pseudomonas aeruginosa</i>	25	150	150/9-33
<i>Pseudomonas aeruginosa</i>	25	151	151/21-45
<i>Salmonella enterica</i>	25	152	152/6-30
<i>Salmonella enterica</i>	25	154	154/1-25
<i>Salmonella enterica</i>	25	155	155/1-25
<i>Salmonella enterica</i>	25	156	156/1-25
<i>Serratia marcescens</i>	25	157	157/12-36
<i>Serratia marcescens</i>	25	158	158/11-35
<i>Serratia marcescens</i>	25	159	159/18-42
<i>Serratia marcescens</i>	25	160	160/9-33
<i>Serratia marcescens</i>	25	161	161/19-43
<i>Staphylococcus aureus</i>	25	162	162/5-29
<i>Staphylococcus aureus</i>	25	163	163/3-27
<i>Staphylococcus aureus</i>	25	164	164/2-26
<i>Staphylococcus aureus</i>	25	165	165/4-28
<i>Staphylococcus aureus</i>	25	166	166/6-30
<i>Staphylococcus epidermidis</i>	25	167	167/8-32
<i>Staphylococcus epidermidis</i>	25	168	168/17-41
<i>Staphylococcus epidermidis</i>	25	169	169/23-47
<i>Staphylococcus epidermidis</i>	25	170	170/6-30
<i>Staphylococcus epidermidis</i>	25	171	171/18-42
<i>Staphylococcus haemolyticus</i>	25	172	172/14-38
<i>Staphylococcus haemolyticus</i>	25	173	173/9-33
<i>Staphylococcus haemolyticus</i>	25	174	174/15-39
<i>Staphylococcus haemolyticus</i>	25	175	175/23-47
<i>Staphylococcus haemolyticus</i>	25	176	176/5-29
<i>Stenotrophomonas maltophilia</i>	25	177	177/18-42
<i>Streptococcus mutans</i>	25	178	178/32-56

<i>Streptococcus mutans</i>	25	179	179/3-27
<i>Streptococcus mutans</i>	25	180	180/6-30
<i>Streptococcus mutans</i>	25	181	181/8-32
<i>Streptococcus mutans</i>	25	182	182/11-35
<i>Streptococcus pyogenes</i>	25	183	183/14-38
<i>Streptococcus pyogenes</i>	25	184	184/12-36
<i>Streptococcus pyogenes</i>	25	185	185/5-29
<i>Streptococcus pyogenes</i>	25	186	186/50-74
<i>Streptococcus pyogenes</i>	25	187	187/8-32
<i>Streptococcus salivarius</i>	25	188	188/5-29
<i>Streptococcus salivarius</i>	25	189	189/15-39
<i>Streptococcus salivarius</i>	25	190	190/13-37
<i>Streptococcus salivarius</i>	25	191	191/8-32
<i>Streptococcus salivarius</i>	25	192	192/5-29
<i>Streptococcus sanguinis</i>	25	193	193/5-29
<i>Streptococcus sanguinis</i>	25	194	194/4-28
<i>Streptococcus sanguinis</i>	25	195	195/33-57
<i>Streptococcus sanguinis</i>	25	196	196/12-36
<i>Streptococcus sanguinis</i>	25	197	197/14-38
<i>Treponema pallidum</i>	25	198	198/32-56
<i>Treponema pallidum</i>	25	199	199/27-51
<i>Treponema pallidum</i>	25	200	200/30-54
<i>Treponema pallidum</i>	25	201	201/24-48
<i>Treponema pallidum</i>	25	202	202/18-42
<i>Vibrio cholerae</i>	25	203	203/2-26
<i>Vibrio parahaemolyticus</i>	25	204	204/18-42
<i>Vibrio parahaemolyticus</i>	25	205	205/42-66
<i>Vibrio parahaemolyticus</i>	25	206	206/11-35
<i>Vibrio parahaemolyticus</i>	25	207	207/10-34
<i>Vibrio parahaemolyticus</i>	25	208	208/14-38
<i>Yersinia enterocolitica</i>	25	209	209/9-33
<i>Yersinia enterocolitica</i>	25	210	210/20-44
<i>Yersinia enterocolitica</i>	25	211	211/12-36
<i>Yersinia enterocolitica</i>	25	212	212/13-37
<i>Yersinia enterocolitica</i>	25	213	213/20-44
<i>Aggregatibacter actinomycetemcomitans</i>	30	1	1/10-39
<i>Aggregatibacter actinomycetemcomitans</i>	30	2	2/2-31
<i>Aggregatibacter actinomycetemcomitans</i>	30	3	3/23-52
<i>Aggregatibacter actinomycetemcomitans</i>	30	4	4/6-35
<i>Aggregatibacter actinomycetemcomitans</i>	30	5	5/6-35
<i>Bacillus anthracis</i>	30	6	6/15-44
<i>Bacillus anthracis</i>	30	7	7/15-44
<i>Bacillus anthracis</i>	30	8	8/15-44
<i>Bacillus anthracis</i>	30	9	9/16-45
<i>Bacillus anthracis</i>	30	10	10/5-34
<i>Bacillus licheniformis</i>	30	11	11/12-41

<i>Bacillus licheniformis</i>	30	12	12/5-34
<i>Bacillus licheniformis</i>	30	13	13/8-37
<i>Bacillus licheniformis</i>	30	14	14/13-42
<i>Bacillus licheniformis</i>	30	15	15/28-57
<i>Bacteroides fragilis</i>	30	16	16/25-54
<i>Bacteroides fragilis</i>	30	17	17/23-52
<i>Bacteroides fragilis</i>	30	18	18/6-35
<i>Bacteroides fragilis</i>	30	19	19/22-51
<i>Bacteroides fragilis</i>	30	20	20/19-48
<i>Bartonella henselae</i>	30	21	21/125-154
<i>Bartonella henselae</i>	30	22	22/15-44
<i>Bartonella henselae</i>	30	23	23/10-39
<i>Bartonella henselae</i>	30	24	24/98-127
<i>Bartonella henselae</i>	30	25	25/2-31
<i>Bordetella pertussis</i>	30	26	26/16-45
<i>Bordetella pertussis</i>	30	27	27/15-44
<i>Bordetella pertussis</i>	30	28	28/12-41
<i>Bordetella pertussis</i>	30	29	29/15-44
<i>Bordetella pertussis</i>	30	30	30/15-44
<i>Borrelia burgdorferi</i>	30	31	31/9-38
<i>Borrelia burgdorferi</i>	30	32	32/3-32
<i>Borrelia burgdorferi</i>	30	33	33/20-49
<i>Borrelia burgdorferi</i>	30	34	34/16-45
<i>Borrelia burgdorferi</i>	30	35	35/30-59
<i>Brucella abortus</i>	30	36	36/12-41
<i>Brucella abortus</i>	30	37	37/16-45
<i>Brucella abortus</i>	30	38	38/15-44
<i>Brucella abortus</i>	30	39	39/15-44
<i>Brucella abortus</i>	30	40	40/15-44
<i>Campylobacter jejuni</i>	30	42	42/3-32
<i>Campylobacter jejuni</i>	30	44	44/11-40
<i>Chlamydia trachomatis</i>	30	45	45/8-37
<i>Chlamydia trachomatis</i>	30	46	46/3-32
<i>Chlamydia trachomatis</i>	30	47	47/6-35
<i>Chlamydia trachomatis</i>	30	48	48/68-97
<i>Chlamydia trachomatis</i>	30	49	49/33-62
<i>Chlamydophila pneumoniae</i>	30	50	50/26-55
<i>Chlamydophila pneumoniae</i>	30	51	51/28-57
<i>Chlamydophila pneumoniae</i>	30	52	52/111-140
<i>Chlamydophila pneumoniae</i>	30	53	53/12-41
<i>Chlamydophila pneumoniae</i>	30	54	54/80-109
<i>Clostridium difficile</i>	30	55	55/5-34
<i>Clostridium difficile</i>	30	56	56/30-59
<i>Clostridium difficile</i>	30	57	57/15-44
<i>Clostridium difficile</i>	30	58	58/3-32
<i>Clostridium difficile</i>	30	59	59/19-48

<i>Clostridium perfringens</i>	30	60	60/10-39
<i>Clostridium perfringens</i>	30	61	61/35-64
<i>Clostridium perfringens</i>	30	62	62/153-182
<i>Clostridium perfringens</i>	30	63	63/9-38
<i>Clostridium perfringens</i>	30	64	64/55-84
<i>Enterobacter aerogenes</i>	30	65	65/26-55
<i>Enterobacter aerogenes</i>	30	66	66/16-45
<i>Enterobacter aerogenes</i>	30	67	67/27-56
<i>Enterobacter aerogenes</i>	30	68	68/18-47
<i>Enterobacter aerogenes</i>	30	69	69/16-45
<i>Enterococcus faecalis</i>	30	70	70/4-33
<i>Enterococcus faecalis</i>	30	71	71/4-33
<i>Enterococcus faecalis</i>	30	72	72/12-41
<i>Enterococcus faecalis</i>	30	73	73/14-43
<i>Enterococcus faecalis</i>	30	74	74/23-52
<i>Enterococcus faecium</i>	30	75	75/3-32
<i>Enterococcus faecium</i>	30	76	76/25-54
<i>Enterococcus faecium</i>	30	77	77/4-33
<i>Enterococcus faecium</i>	30	78	78/1-30
<i>Enterococcus faecium</i>	30	79	79/11-40
<i>Francisella tularensis</i>	30	80	80/12-41
<i>Francisella tularensis</i>	30	81	81/3-32
<i>Francisella tularensis</i>	30	82	82/2-31
<i>Francisella tularensis</i>	30	83	83/9-38
<i>Francisella tularensis</i>	30	84	84/10-39
<i>Haemophilus influenzae</i>	30	85	85/15-44
<i>Klebsiella oxytoca</i>	30	88	88/8-37
<i>Klebsiella oxytoca</i>	30	89	89/6-35
<i>Klebsiella oxytoca</i>	30	90	90/17-46
<i>Klebsiella oxytoca</i>	30	91	91/16-45
<i>Klebsiella oxytoca</i>	30	92	92/8-37
<i>Legionella pneumophila</i>	30	93	93/2-31
<i>Legionella pneumophila</i>	30	94	94/6-35
<i>Legionella pneumophila</i>	30	95	95/27-56
<i>Legionella pneumophila</i>	30	96	96/3-32
<i>Legionella pneumophila</i>	30	97	97/5-34
<i>Listeria monocytogenes</i>	30	98	98/4-33
<i>Listeria monocytogenes</i>	30	99	99/3-32
<i>Listeria monocytogenes</i>	30	100	100/15-44
<i>Listeria monocytogenes</i>	30	101	101/4-33
<i>Listeria monocytogenes</i>	30	102	102/15-44
<i>Moraxella catarrhalis</i>	30	103	103/29-58
<i>Moraxella catarrhalis</i>	30	104	104/30-59
<i>Moraxella catarrhalis</i>	30	105	105/85-114
<i>Moraxella catarrhalis</i>	30	106	106/54-83
<i>Moraxella catarrhalis</i>	30	107	107/9-38

<i>Mycobacterium avium</i>	30	108	108/12-41
<i>Mycobacterium avium</i>	30	109	109/8-37
<i>Mycobacterium avium</i>	30	110	110/2-31
<i>Mycobacterium avium</i>	30	111	111/22-51
<i>Mycobacterium avium</i>	30	112	112/14-43
<i>Mycobacterium bovis</i>	30	113	113/16-45
<i>Mycobacterium bovis</i>	30	114	114/16-45
<i>Mycobacterium bovis</i>	30	115	115/15-44
<i>Mycobacterium bovis</i>	30	116	116/15-44
<i>Mycoplasma genitalium</i>	30	117	117/2-31
<i>Mycoplasma genitalium</i>	30	118	118/10-39
<i>Mycoplasma genitalium</i>	30	119	119/101-130
<i>Mycoplasma genitalium</i>	30	120	120/20-49
<i>Mycoplasma genitalium</i>	30	121	121/118-147
<i>Mycoplasma pneumoniae</i>	30	122	122/19-48
<i>Mycoplasma pneumoniae</i>	30	123	123/28-57
<i>Mycoplasma pneumoniae</i>	30	124	124/47-76
<i>Mycoplasma pneumoniae</i>	30	125	125/65-94
<i>Mycoplasma pneumoniae</i>	30	126	126/10-39
<i>Neisseria gonorrhoeae</i>	30	127	127/2-31
<i>Neisseria gonorrhoeae</i>	30	128	128/14-43
<i>Neisseria gonorrhoeae</i>	30	129	129/12-41
<i>Neisseria gonorrhoeae</i>	30	130	130/11-40
<i>Neisseria gonorrhoeae</i>	30	131	131/8-37
<i>Neisseria meningitidis</i>	30	132	132/12-41
<i>Neisseria meningitidis</i>	30	133	133/8-37
<i>Neisseria meningitidis</i>	30	134	134/26-55
<i>Neisseria meningitidis</i>	30	135	135/15-44
<i>Neisseria meningitidis</i>	30	136	136/11-40
<i>Porphyromonas gingivalis</i>	30	137	137/65-94
<i>Porphyromonas gingivalis</i>	30	138	138/12-41
<i>Porphyromonas gingivalis</i>	30	139	139/2-31
<i>Porphyromonas gingivalis</i>	30	140	140/92-121
<i>Porphyromonas gingivalis</i>	30	141	141/56-85
<i>Proteus mirabilis</i>	30	142	142/10-39
<i>Proteus mirabilis</i>	30	143	143/40-69
<i>Proteus mirabilis</i>	30	144	144/51-80
<i>Proteus mirabilis</i>	30	145	145/19-48
<i>Proteus mirabilis</i>	30	146	146/18-47
<i>Pseudomonas aeruginosa</i>	30	147	147/2-31
<i>Pseudomonas aeruginosa</i>	30	148	148/4-33
<i>Pseudomonas aeruginosa</i>	30	149	149/12-41
<i>Pseudomonas aeruginosa</i>	30	150	150/7-36
<i>Pseudomonas aeruginosa</i>	30	151	151/19-48
<i>Salmonella enterica</i>	30	152	152/2-31
<i>Serratia marcescens</i>	30	157	157/10-39

<i>Serratia marcescens</i>	30	158	158/8-37
<i>Serratia marcescens</i>	30	159	159/16-45
<i>Serratia marcescens</i>	30	160	160/6-35
<i>Serratia marcescens</i>	30	161	161/15-44
<i>Staphylococcus aureus</i>	30	162	162/3-32
<i>Staphylococcus aureus</i>	30	164	164/1-30
<i>Staphylococcus aureus</i>	30	165	165/2-31
<i>Staphylococcus aureus</i>	30	166	166/3-32
<i>Staphylococcus epidermidis</i>	30	167	167/6-35
<i>Staphylococcus epidermidis</i>	30	168	168/15-44
<i>Staphylococcus epidermidis</i>	30	169	169/20-49
<i>Staphylococcus epidermidis</i>	30	170	170/3-32
<i>Staphylococcus epidermidis</i>	30	171	171/16-45
<i>Staphylococcus haemolyticus</i>	30	172	172/11-40
<i>Staphylococcus haemolyticus</i>	30	173	173/7-36
<i>Staphylococcus haemolyticus</i>	30	174	174/13-42
<i>Staphylococcus haemolyticus</i>	30	175	175/21-50
<i>Staphylococcus haemolyticus</i>	30	176	176/2-31
<i>Stenotrophomonas maltophilia</i>	30	177	177/16-45
<i>Streptococcus mutans</i>	30	178	178/28-57
<i>Streptococcus mutans</i>	30	179	179/1-30
<i>Streptococcus mutans</i>	30	180	180/3-32
<i>Streptococcus mutans</i>	30	181	181/6-35
<i>Streptococcus mutans</i>	30	182	182/8-37
<i>Streptococcus pyogenes</i>	30	183	183/11-40
<i>Streptococcus pyogenes</i>	30	184	184/10-39
<i>Streptococcus pyogenes</i>	30	185	185/3-32
<i>Streptococcus pyogenes</i>	30	186	186/47-76
<i>Streptococcus pyogenes</i>	30	187	187/6-35
<i>Streptococcus salivarius</i>	30	188	188/2-31
<i>Streptococcus salivarius</i>	30	189	189/12-41
<i>Streptococcus salivarius</i>	30	190	190/10-39
<i>Streptococcus salivarius</i>	30	191	191/5-34
<i>Streptococcus salivarius</i>	30	192	192/3-32
<i>Streptococcus sanguinis</i>	30	193	193/2-31
<i>Streptococcus sanguinis</i>	30	194	194/2-31
<i>Streptococcus sanguinis</i>	30	195	195/28-57
<i>Streptococcus sanguinis</i>	30	196	196/10-39
<i>Streptococcus sanguinis</i>	30	197	197/11-40
<i>Treponema pallidum</i>	30	198	198/29-58
<i>Treponema pallidum</i>	30	199	199/26-55
<i>Treponema pallidum</i>	30	200	200/29-58
<i>Treponema pallidum</i>	30	201	201/22-51
<i>Treponema pallidum</i>	30	202	202/15-44
<i>Vibrio parahaemolyticus</i>	30	204	204/15-44
<i>Vibrio parahaemolyticus</i>	30	205	205/37-66

<i>Vibrio parahaemolyticus</i>	30	206	206/7-36
<i>Vibrio parahaemolyticus</i>	30	207	207/7-36
<i>Vibrio parahaemolyticus</i>	30	208	208/11-40
<i>Yersinia enterocolitica</i>	30	209	209/7-36
<i>Yersinia enterocolitica</i>	30	210	210/18-47
<i>Yersinia enterocolitica</i>	30	211	211/9-38
<i>Yersinia enterocolitica</i>	30	212	212/11-40
<i>Yersinia enterocolitica</i>	30	213	213/18-47

CLAIMS

1. A method for identifying one or more nucleotide k-mers in the genome of an organism which are capable of distinguishing the organism from one or more different organisms, comprising:
(a) extracting all nucleotide k-mers from the genome of the organism or from one or more parts of said genome; and
(b) comparing the nucleotide k-mers extracted in step (a) with the genome(s) of the one or more different organisms and identifying those nucleotide k-mers which do not appear in the genome(s) of the one or more different organisms.
2. A method according to any one of the preceding claims, wherein the nucleotide k-mers are extracted from the parts of the genome which encode the ribosomal protein subunits.
3. A method according to claim 1 or 2, wherein step (a) comprises defining a length for the one or more nucleotide k-mers and extracting all nucleotide k-mers having that length from the genome of the organism.
4. A method according to claim 3, wherein step (a) comprises defining a length of from about 10 to about 40 nucleotides for the one or more nucleotide k-mers and extracting all nucleotide k-mers having that length from the genome of the organism.
5. A method according to any one of the preceding claims, wherein the method further comprises (c) discounting any nucleotide k-mers identified in step (b) which have an undesirable property.
6. A method according to claim 5, wherein the undesirable property is (i) its presence in the human genome, (ii) a homopolymer or repetitive sequence, (iii) the ability to form secondary structure, or (iv) its presence in a contaminating organism.

7. A method according to any one of the preceding claims, wherein the method further comprises (d) analysing the nucleotide k-mers identified in step (b) or step (c) and reducing their number by removing at least some sequence redundancy.
8. A method according to claim 7, wherein step (d) comprises minimising the number of nucleotide k-mers by removing all sequence redundancy.
9. A method according to any one of the preceding claims, wherein the organism is a microorganism, a fungus or a virus and the one or more different organisms are one or more different microorganisms, fungi or viruses.
10. A method according to claim 9, wherein the organism is a bacterium and the one or more different organisms are one or more different bacteria.
11. A method according to claim 10, wherein the organism is a bacterium and the one or more different organisms are one or more bacteria from one or more different genera of bacteria.
12. A method according to claim 10 or 11, wherein the organism is a bacterium and the one or more different organisms are one or more different species of bacteria.
13. A method according to any one of the preceding claims, wherein step (a) comprises extracting all nucleotide k-mers from about two or more different versions of the genome of the organism.
14. A method according to any one of the preceding claims, wherein step (a) comprises extracting all nucleotide k-mers from (i) about two hundred or more or (ii) about two thousand or more different versions of the genome of the organism.
15. A method according to any one of the preceding claims, wherein step (a) comprises extracting all nucleotide k-mers from all known versions of the genome of the organism.
16. A method according to any one of claims 13 to 15, wherein the method is for identifying about 25 or fewer nucleotide k-mers which are present in all of the different versions of the genome of the organism and which are capable of distinguishing the organism from the one or more different organisms.

17. A method according to any one of claims 13 to 15, wherein the method is for identifying only one nucleotide k-mer which is present in all of the different versions of the genome of the organism and which is capable of distinguishing the organism from the one or more different organisms.
18. A method according to any one of the preceding claims, wherein step (b) comprises comparing the nucleotide k-mers extracted in step (a) with two or more versions of the genome of each the one or more different organisms and identifying those nucleotide k-mers which do not appear in the genomes of the one or more different organisms
19. A method according to any one of claims 1 to 17, wherein step (b) comprises comparing the nucleotide k-mers extracted in step (a) with two hundred or more, two thousand or more or all known versions of the genome of each the one or more different organisms and identifying those nucleotide k-mers which do not appear in the genomes of the one or more different organisms.
20. A method according to any one of the preceding claims, wherein the method is for identifying one or more nucleotide k-mers in the genome of an organism which can be associated with a phenotype of the organism, comprising:
- (a) extracting all nucleotide k-mers from the genome of the organism; and
 - (b) comparing the nucleotide k-mers extracted in step (a) with the genome(s) of one or more different organisms which do not display the phenotype and identifying those nucleotide k-mers which do not appear in the genome(s) of the one or more different organisms which do not display the phenotype.
21. A method according to claim 20, wherein the phenotype is pathogenicity, resistance to antibiotics, host specificity, tissue specificity, transmissibility, virulence, antigenicity or biochemical properties.
22. A computer program configured to carry out a method according to any one of the preceding claims.

23. A computer program medium comprising a computer program according to claim 22.
24. A computer for carrying out a method according to any one of claims 1 to 21 or programmed with a computer program according to claim 22.
25. One or more nucleotide k-mers identified using a method according to any one of claims 1 to 21 or a computer program according to claim 22.
26. One more nucleotide k-mers according to claim 25, which are identified from the parts of the genome which encode the ribosomal protein subunits.
27. One or more nucleotide k-mers according to claim 25 or 26, wherein the one or more k-mers can be associated with a phenotype of the organism.
28. An oligonucleotide probe which comprises a sequence which is complementary to a nucleotide k-mer according to any one of claims 25 to 27.
29. A plurality of oligonucleotide probes each of which comprises a sequence which is complementary to one of the nucleotide k-mers in a population of two or more nucleotide k-mers according to any one of claims 25 to 27.
30. An oligonucleotide probe according to claim 28 or a plurality of oligonucleotide probes according to claim 29, wherein a said nucleotide k-mer to which a sequence in a probe is complementary is identified from the parts of the genome which encode the ribosomal protein subunits.
31. An oligonucleotide probe according to claim 28 or 30 or a plurality of oligonucleotide probes according to claim 29 or 30, wherein the probe is or the probes are detectably labelled.
32. An oligonucleotide probe according to any one of claims 28, 30 or 31 or a plurality of oligonucleotide probes according to any one of claims 29 to 31, wherein the probe is a molecular beacon probe or the probes are molecular beacon probes.

33. A support having attached thereto an oligonucleotide probe according to any one of claims 28 or 30 to 32 or a plurality of oligonucleotide probes according to any one of claims 29 to 32.
34. A support according to claim 33, wherein the support comprises a chip, pin, array or dipstick.
35. A method for detecting the presence or absence of an organism in a sample, comprising detecting the presence or absence in the sample of one or more nucleotide k-mers according to claim 25, wherein the presence of the one or more nucleotide k-mers is indicative of the presence of the organism in the sample and the wherein the absence of the one or more nucleotide k-mers is indicative of the absence of the organism from the sample.
36. A method according to claim 35, wherein the nucleotide k-mers are extracted from the parts of the genome which encode the ribosomal protein subunits.
37. A method according to claim 35 or 36, wherein the method is for detecting the presence or absence of an organism with a phenotype in a sample and the method comprises detecting the presence or absence in the sample of one or more nucleotide k-mers according to claim 24.
38. A nucleotide k-mer capable of distinguishing an organism from one or more different organisms, wherein the k-mer comprises about 10 or more consecutive nucleotides from any one of the sequences shown in SEQ ID NOs: 1 to 213.
39. A nucleotide k-mer according to claim 38, wherein the k-mer comprises (i) from about 10 to about 40 consecutive nucleotides from any of the sequences shown in SEQ ID NOs: 1 to 213 or (ii) any of the consecutive nucleotides listed in the last column in Table 3.
40. A nucleotide k-mer according to claim 38 or 39, wherein the k-mer can be associated with a phenotype of the organism.
41. A nucleotide k-mer according to any one of claims 38 to 40, wherein

the k-mer is from one of SEQ ID NOs: 1 to 5 and the organism is *Aggregatibacter actinomycetemcomitans*;

the k-mer is from one of SEQ ID NOs: 6 to 10 and the organism is *Bacillus anthracis*;

the k-mer is from one of SEQ ID NOs: 11 to 15 and the organism is *Bacillus licheniformis*;

the k-mer is from one of SEQ ID NOs: 16 to 20 and the organism is *Bacteroides fragilis*;

the k-mer is from one of SEQ ID NOs: 21 to 25 to 5 and the organism is *Bartonella henselae*;

the k-mer is from one of SEQ ID NOs: 26 to 30 and the organism is *Bordetella pertussis*;

the k-mer is from one of SEQ ID NOs: 31 to 35 and the organism is *Borrelia burgdorferi*;

the k-mer is from one of SEQ ID NOs: 36 to 40 and the organism is *Brucella abortus*;

the k-mer is from one of SEQ ID NOs: 41 to 44 and the organism is *Campylobacter jejuni*;

the k-mer is from one of SEQ ID NOs: 45 to 49 and the organism is *Chlamydia trachomatis*;

the k-mer is from one of SEQ ID NOs: 50 to 54 and the organism is *Chlamydophila pneumonia*;

the k-mer is from one of SEQ ID NOs: 55 to 59 and the organism is *Clostridium difficile*;

the k-mer is from one of SEQ ID NOs: 60 to 64 and the organism is *Clostridium perfringens*;

the k-mer is from one of SEQ ID NOs: 65 to 69 and the organism is from *Enterobacter aerogenes*;

the k-mer is from one of SEQ ID NOs: 70 to 74 and the organism is from *Enterococcus faecalis*;

the k-mer is from one of SEQ ID NOs: 75 to 79 and the organism is from *Enterococcus faecium*;

the k-mer is from one of SEQ ID NOs: 80 to 84 and the organism is *Francisella tularensis*;

the k-mer is from SEQ ID NO: 85 and the organism is *Haemophilus influenza*;

the k-mer is from one of SEQ ID NOs: 86 and 87 and the organism is *Helicobacter pylori*;

the k-mer is from one of SEQ ID NOs: 88 to 92 and the organism is *Klebsiella oxytoca*

the k-mer is from one of SEQ ID NOs: 93 to 97 and the organism is *Legionella pneumophila*;

the k-mer is from one of SEQ ID NOs: 98 to 102 and the organism is *Listeria monocytogenes*;

the k-mer is from one of SEQ ID NOs: 103 to 107 and the organism is *Moraxella catarrhalis*;

the k-mer is from one of SEQ ID NOs: 108 to 112 and the organism is *Mycobacterium avium*;

the k-mer is from one of SEQ ID NOs: 113 to 116 and the organism is *Mycobacterium bovis*;

the k-mer is from one of SEQ ID NOs: 117 to 121 and the organism is *Mycoplasma genitalium*;

the k-mer is from one of SEQ ID NOs: 122 to 126 and the organism is *Mycoplasma pneumonia*;

the k-mer is from one of SEQ ID NOs: 127 to 131 and the organism is *Neisseria gonorrhoeae*;

the k-mer is from one of SEQ ID NOs: 132 to 136 and the organism is *Neisseria meningitides*;

the k-mer is from one of SEQ ID NOs: 137 to 141 and the organism is *Porphyromonas gingivalis*;

the k-mer is from one of SEQ ID NOs: 142 to 146 and the organism is *Proteus mirabilis*;

the k-mer is from one of SEQ ID NOs: 147 to 151 and the organism is *Pseudomonas aeruginosa*;

the k-mer is from one of SEQ ID NOs: 152 to 156 and the organism is *Salmonella enterica*;

the k-mer is from one of SEQ ID NOs: 157 to 161 and the organism is *Serratia marcescens*;

the k-mer is from one of SEQ ID NOs: 162 to 166 and the organism is *Staphylococcus aureus*;

the k-mer is from one of SEQ ID NOs: 167 to 171 and the organism is *Staphylococcus epidermidis*;

the k-mer is from one of SEQ ID NOs: 172 to 176 and the organism is *Staphylococcus haemolyticus*;

the k-mer is from SEQ ID NO: 177 and the organism is *Stenotrophomonas maltophilia*;

the k-mer is from one of SEQ ID NOs: 178 to 182 and the organism is *Streptococcus mutans*;

the k-mer is from one of SEQ ID NOs: 183 to 187 and the organism is *Streptococcus pyogenes*;

the k-mer is from one of SEQ ID NOs: 188 to 192 and the organism is *Streptococcus salivarius*;

the k-mer is from one of SEQ ID NOs: 193 to 197 and the organism is *Streptococcus sanguinis*;

the k-mer is from one of SEQ ID NOs: 198 to 202 and the organism is *Treponema pallidum*;

the k-mer is from SEQ ID NO: 203 and the organism is *Vibrio cholera*;

the k-mer is from one of SEQ ID NOs: 204 to 208 and the organism is *Vibrio parahaemolyticus*; or

the k-mer is from one of SEQ ID NOs: 209 to 213 and the organism is *Yersinia enterocolitica*;

42. An oligonucleotide probe which comprises a sequence which is complementary to a nucleotide k-mer according to any one of claims 38 to 41.

43. An oligonucleotide probe according to claim 42, wherein the probe is detectably labelled and/or is a molecular beacon probe.

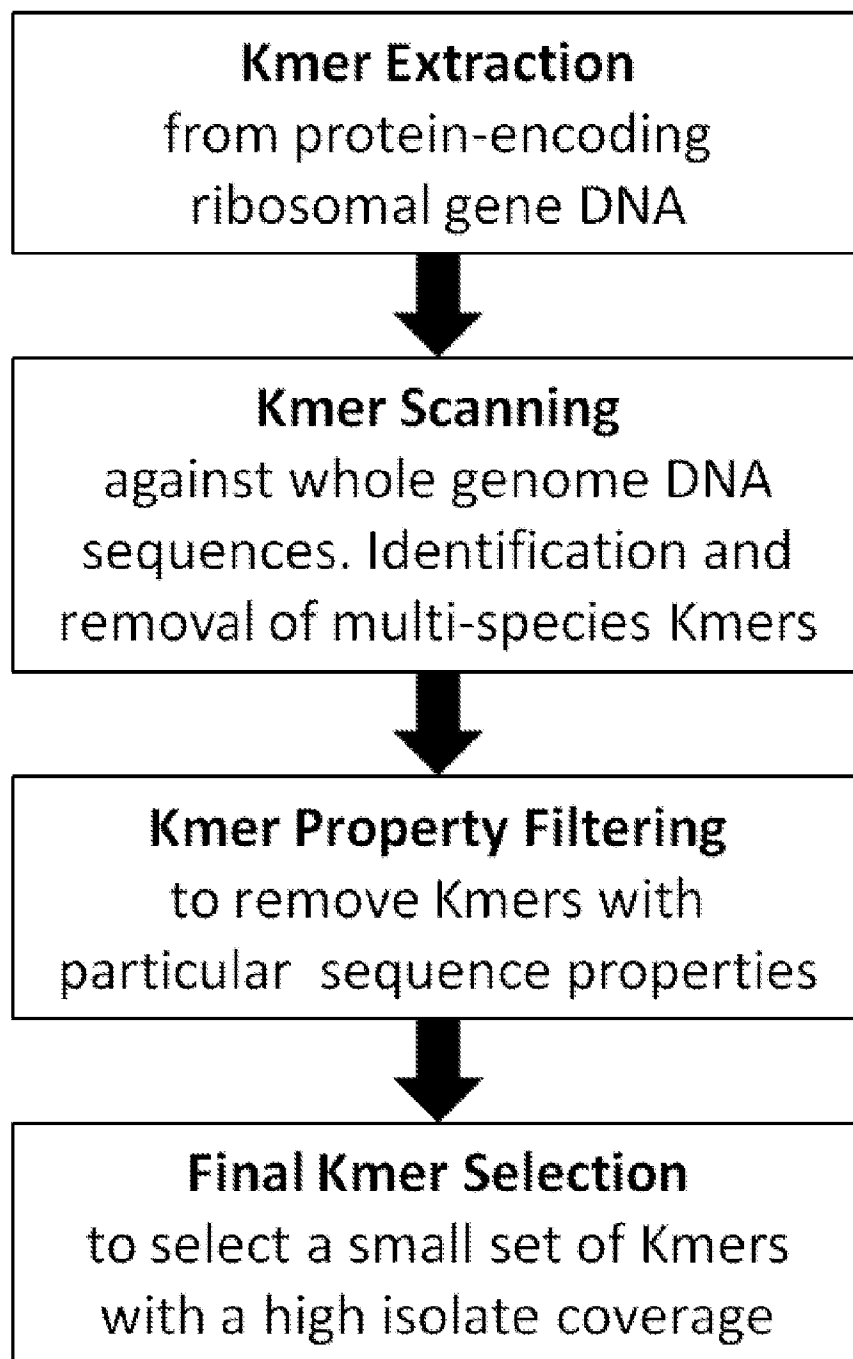
44. A support having attached thereto a oligonucleotide probe according to claim 42 or 43.

45. A support according to claim 44, wherein the support comprises a chip, pin, array or dipstick.

46. A method for detecting the presence or absence of an organism in a sample, comprising detecting the presence or absence in the sample of a nucleotide k-mer according to any one of claims 38 to 41, wherein the presence of the nucleotide k-mer is indicative of the presence of the organism in the sample and the wherein the absence of the nucleotide k-mer is indicative of the absence of the organism from the sample.

47. A method according to claim 46, wherein the method is for detecting the presence or absence of the organism with a particular phenotype in a sample and the method comprises detecting the presence or absence in the sample of a nucleotide k-mer according to claim 40.

Figure 1



INTERNATIONAL SEARCH REPORT

International application No

PCT/GB2016/051802

A. CLASSIFICATION OF SUBJECT MATTER

INV. G06F19/18 G06F19/22
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EPO-Internal

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 2013/175164 A1 (DISCUVA LTD [GB]) 28 November 2013 (2013-11-28) whole document, in particular claim 1 p. 1, examples 1-3	1-24, 35-37
X	----- CHOR BENNY ET AL: "Genomic DNA k-mer spectra: models and modalities", GENOME BIOLOGY, BIOMED CENTRAL LTD., LONDON, GB, vol. 10, no. 10, 8 October 2009 (2009-10-08), page R108, XP021065352, ISSN: 1465-6906, DOI: 10.1186/GB-2009-10-10-R108 whole document, in particular abstract and fig.5 ----- -/--	1-24, 35-37



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

15 September 2016

Date of mailing of the international search report

23/11/2016

Name and mailing address of the ISA/

European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040,
Fax: (+31-70) 340-3016

Authorized officer

Lüdemann, Susanna

INTERNATIONAL SEARCH REPORT

International application No.

PCT/GB2016/051802

Box No. I Nucleotide and/or amino acid sequence(s) (Continuation of item 1.c of the first sheet)

1. With regard to any nucleotide and/or amino acid sequence disclosed in the international application, the international search was carried out on the basis of a sequence listing:
 - a. ☒ forming part of the international application as filed:
 - ☒ in the form of an Annex C/ST.25 text file.
 - ☐ on paper or in the form of an image file.
 - b. ☐ furnished together with the international application under PCT Rule 13~~ter~~.1(a) for the purposes of international search only in the form of an Annex C/ST.25 text file.
 - c. ☐ furnished subsequent to the international filing date for the purposes of international search only:
 - ☐ in the form of an Annex C/ST.25 text file (Rule 13~~ter~~.1(a)).
 - ☐ on paper or in the form of an image file (Rule 13~~ter~~.1(b) and Administrative Instructions, Section 713).
2. ☐ In addition, in the case that more than one version or copy of a sequence listing has been filed or furnished, the required statements that the information in the subsequent or additional copies is identical to that forming part of the application as filed or does not go beyond the application as filed, as appropriate, were furnished.
3. Additional comments:

INTERNATIONAL SEARCH REPORT

International application No.
PCT/GB2016/051802

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:
2. ☒ Claims Nos.: 25-34
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:
see FURTHER INFORMATION sheet PCT/ISA/210
3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

see additional sheet

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.
2. ☐ As all searchable claims could be searched without effort justifying an additional fees, this Authority did not invite payment of additional fees.
3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:
4. ☒ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:
1-24(completely); 35-37(partially)

Remark on Protest

- ☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.
- ☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.
- ☐ No protest accompanied the payment of additional search fees.

INTERNATIONAL SEARCH REPORT

International application No

PCT/GB2016/051802

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 2004/065565 A2 (SCIENCE APPLIC INT CORP [US]) 5 August 2004 (2004-08-05) whole document, in particular claims and fig. 4	1-24, 35-37
X	----- YI-PING PHOEBE CHEN ET AL: "BACTERIAL POPULATION ASSAY VIA K-MER ANALYSIS (EXTENDED ABSTRACT)", PROCEEDINGS OF THE 3RD ASIA-PACIFIC BIOINFORMATICS CONFERENCE, 1 January 2005 (2005-01-01), pages 299-308, XP055301849, DOI: 10.1142/9781860947322_0030 ISBN: 978-1-86094-732-2 the whole document -----	1-24, 35-37

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/GB2016/051802

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 2013175164	A1	28-11-2013	NONE

WO 2004065565	A2	05-08-2004	US 2005050101 A1 03-03-2005
			WO 2004065565 A2 05-08-2004
			WO 2005017488 A2 24-02-2005

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210

This International Searching Authority found multiple (groups of) inventions in this international application, as follows:

1. claims: 1-24(completely); 35-37(partially)

A method for identifying one or more nucleotide k-mers in the genome of an organism which are capable of distinguishing the organism from one or more different organisms, comprising:(a)extracting all nucleotide k-mers from the genome of the organism or from one or more parts of said genome; and(b) comparing the nucleotide k-mers extracted in step (a) with the genome(s) of the one or more different organisms and identifying those nucleotide k-mers which do not appear in the genome(s) of the one or more different organisms.

2-214. claims: 38-47(completely); 35-37(partially)

A nucleotide k-mer capable of distinguishing an organism from one or more different organisms, wherein the k-mer comprises about 10 or more consecutive nucleotides from any one of the sequences shown in SEQ ID NOs: 1 to 213 and oligonucleotide probes and supports and methods based thereon.

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210

Continuation of Box II.2

Claims Nos.: 25-34

Present claims 25-33 encompass compounds (oligonucleotides and support with attached oligonucleotides) defined only by their desired function, contrary to the requirements of clarity of Article 6 PCT, because the result-to-be-achieved type of definition does not allow the scope of the claim to be ascertained. The fact that any compound could be screened by computer does not overcome this objection, as the skilled person would not have knowledge beforehand as to whether it would fall within the scope claimed, except for the compounds according to claims 38-43, which however, lack unity. Undue experimentation/computation would be required to screen compounds randomly. This non-compliance with the substantive provisions is to such an extent, that the search was performed taking into consideration the non-compliance in determining the extent of the search for claim.

The search of said claims will be consequently restricted to compounds according to the inventions listed as inventions 2-214.

The applicant's attention is drawn to the fact that claims relating to inventions in respect of which no international search report has been established need not be the subject of an international preliminary examination (Rule 66.1(e) PCT). The applicant is advised that the EPO policy when acting as an International Preliminary Examining Authority is normally not to carry out a preliminary examination on matter which has not been searched. This is the case irrespective of whether or not the claims are amended following receipt of the search report or during any Chapter II procedure. If the application proceeds into the regional phase before the EPO, the applicant is reminded that a search may be carried out during examination before the EPO (see EPO Guidelines C-IV, 7.2), should the problems which led to the Article 17(2) declaration be overcome.