BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 22 2014, pages 3174–3180
doi:10.1093/bioinformatics/btu515

Sequence analysis

Advance Access publication August 7, 2014

A method for *de novo* nucleic acid diagnostic target discovery

Yeting Zhang[1,2] and Yazhou Sun[2,3,*]

[1]Department of Biology, Pennsylvania State University, University Park, PA 16802, [2]Department of Research, Synblex LLC, 200 Innovation Blvd. State College, PA 16801 and [3]New Jersey Center for Science, Technology & Mathematics, Kean University, Union, NJ 07083, USA

Associate Editor: John Hancock

ABSTRACT

Motivation: A proper target or marker is essential in any diagnosis (e.g. an infection or cancer). An ideal diagnostic target should be both conserved in and unique to the pathogen. Currently, these targets can only be identified manually, which is time-consuming and usually error-prone. Because of the increasingly frequent occurrences of emerging epidemics and multidrug-resistant ‘superbugs’, a rapid diagnostic target identification process is needed.

Results: A new method that can identify uniquely conserved regions (UCRs) as candidate diagnostic targets for a selected group of organisms solely from their genomic sequences has been developed and successfully tested. Using a sequence-indexing algorithm to identify UCRs and a *k*-mer integer-mapping model for computational efficiency, this method has successfully identified UCRs within the bacteria domain for 15 test groups, including pathogenic, probiotic, commensal and extremophilic bacterial species or strains. Based on the identified UCRs, new diagnostic primer sets were designed, and their specificity and efficiency were tested by polymerase chain reaction amplifications from both pure isolates and samples containing mixed cultures.

Availability and implementation: The UCRs identified for the 15 bacterial species are now freely available at http://ucr.synblex.com. The source code of the programs used in this study is accessible at http://ucr.synblex.com/bacterialIdSourceCode.d.zip

Contact: yazhousun@synblex.com

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on March 31, 2014; revised on July 15, 2014; accepted on July 22, 2014

1 INTRODUCTION

Finding the proper target or marker for a specific disease (e.g. an infection or cancer) is essential for its successful diagnosis. An ideal candidate region must be both conserved in and unique to the specific target (e.g. a pathogen species or strain), i.e. a uniquely conserved region, to ensure specific and effective detection. Target-specific nucleic acid diagnostic methods, including hybridization, amplification and sequencing, have been widely used in academia and in health care, food safety and biosecurity industries (Caliendo, 2011; Hoorfar, 2012; Miller and Tang, 2009; Morse, 2012; Neuberger *et al.*, 2008; Yang and Rothman, 2004). Specific yet sensitive oligonucleotides (probes or primers) are essential to all of these assays. Current tools, such as Primer3 (Koressaar and Remm, 2007) and Primer-BLAST (Ye *et al.*, 2012), can only facilitate probe or primer design when a candidate region is provided, but are not able to identify these candidate regions by themselves. So far, no existing tools can identify these candidate regions *de novo*, and these candidate targets can only be selected manually, usually based on the pathogen’s known properties, such as secreted toxins or genes involved in its pathogenesis (Brakstad *et al.*, 1992; Salo *et al.*, 1995). This process is time-consuming and usually error-prone, as the proposed detection primers must be thoroughly examined to ensure that they are universally present within the targeted group (e.g. a pathogenic species or strain) and to rule out high similarity matches in unintended groups. More pressingly, the number of emerging epidemics and cases with multidrug-resistant pathogens has continued to increase in recent years (David and Daum, 2010; Kruse *et al.*, 2013; Mathers *et al.*, 2012; Newell *et al.*, 2010; Snitkin *et al.*, 2012). Early and accurate diagnosis is crucial in these cases, which often necessitate a rapid diagnostic test development. As a result, methods that can automatically identify these candidate regions within a reasonable time frame and have an affordable cost are desirable.

2 MATERIALS AND METHODS

For a specific group of organisms (e.g. a species or strain), the regions that meet the aforementioned requirements as candidate diagnostic target regions are referred to as the uniquely conserved regions (UCRs). Mathematically, UCRs could be defined as a sequence s that is present in the genome x of each member of the selected group G, and is not present in the genome y of each member of G^C. In the context of a given domain D (e.g. bacteria), $G^C = D - G = \{y \mid y \in D \wedge y \notin G\}$. The set of UCRs for G is denoted as S. For a given G, S may be $\emptyset$, as there might not be any UCRs for a group.

$|S|$ has no mathematical upper bound without limiting the length k of each s, so solving S is intractable in the absence of k bounds. By fixing k, S could be solved by comparing the occurrence count of every possible k-mer between G and G^C. The problem, thus, has been reduced to finding a k-mer u that is present in every genome x in G, but not in any genome y in G^C. All uniquely conserved k-mers could be identified by first indexing all k-mer occurrences in G and G^C separately, and then comparing these two indexes to screen for those k-mers that meet the definition, i.e. are UCR k-mers (Fig. 1A). A region where all k-mers (k above a preset value) are uniquely conserved is a UCR.

However, the possible k-mer space is vast for typical k-mers commonly used in primers or probes. Considering only the A, T, C and G four nucleotides, for a given length k, the total number of k-mers is 4^k. Longer k gives better resolution but also means higher computational, storage and transmission costs. The typical primer length is between 18 and 24 (Dieffenbach *et al.*, 1993). With k set to the lower bound 18, the

*To whom correspondence should be addressed.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

© The Author 2014. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com

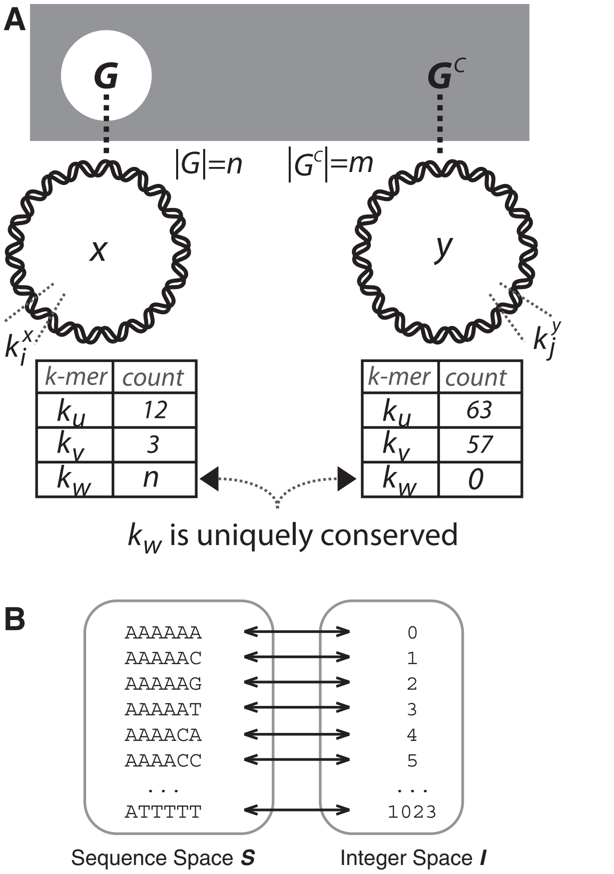


Fig. 1. The algorithm for *de novo* nucleic acid diagnostic target discovery. (**A**) The *k*-mer indexing method for identifying UCRs. The white circle *G* stands for the selected group, the gray box without *G* is G^C and *G* and G^C together is the entire *k*-mer space (e.g. the *k*-mer space of the entire bacteria domain). The occurrence of each *k*-mer in *G* and G^C is indexed, and then the uniquely conserved *k*-mers of *G* can be indentified through comparison. (**B**). The *k*-mer integer mapping method. Showing $k = 6$, when A = 0, C = 1, G = 2, T = 3, following a positional numeral system of radix 4. The one-to-one mapping between sequence space *S* and integer space *I* is thus established, and *S* is therefore naturally ordered by the mapped integer

total space required for a straightforward *k*-mer occurrence index would require at least 1.4 TB of space (6.9×10^{10} index length, 18 bytes for *k*-mer and 2 bytes for occurrence count). This large size has posed a serious challenge for both constructing and screening this index, even for out-of-core (i.e. external memory) algorithms (Vitter, 2008).

To reduce the space requirement, and to facilitate task partitioning in parallel computing, a *k*-mer integer mapping method was used (Fig. 1B). This method maps each DNA sequence of a fixed length *k* to a specific integer by encoding the nucleotide sequence in a positional numeral system of radix 4 (i.e. a quaternary numeral system). Let A = 0, C = 1, G = 2 and T = 3; each DNA sequence of a fixed length *k* could be converted into a unique integer, allowing straightforward lossless encoding and decoding. More precisely, for a sequence *s* of (length *n*) = $b_{n-1} \ldots b_0$, where $b_i \in \{A = 0, C = 1, G = 2, T = 3\}$, the corresponding integer is $\sum_{i=0}^{n-1}(b_i \times 4^i)$.

For example, with the above formula, ACGT = 27 and GTCA = 180. When given $k = 4$, the integers 27 and 180 could be uniquely decoded as ACGT and GTCA, respectively. As a result, the position offset within the occurrence count index was used to implicitly represent the *k*-mer, which reduced the space cost by 10-fold. In addition, the entire index could be partitioned and addressed by the integer order, which greatly simplified partitioning and aggregation in parallelization. Additional steps were also included to consider the reverse complement sequence of a *k*-mer, and the *k*-mers spanning the break points of the circular bacterial nuclear genomes.

After identifying all UCR *k*-mers for a group, the UCRs were assembled from these UCR *k*-mers. The UCR *k*-mer distribution among genes (or genomic regions) was then calculated against a known reference genome annotation. Primers were picked from a gene or region that contains the largest amount of UCRs with the highest density but avoids common household genes across the bacterial domain. These primers were then screened for appropriate product size (0.6 ~ 1 kb), a neutral GC% (46%~54%), the presence of a 3′-GC clamp, a compatible melting temperature between forward and reverse primer pairs and minimal heterodimer/homodimer Δ*G*s. Finally, as for any UCR *k*-mer, the $(k-i)$-mers $(0<i<k-1)$ contained within do not necessarily and often do not meet the UCR definition, all remaining candidate primers were screened against the NCBI nucleotide database to further remove those with high similarity matches in unintended groups, especially close to the 3′ end.

The complete source code for building a background *k*-mer index, UCR identification and primer selection is provided at http://ucr.synblex.com/bacterialIdSourceCode.d.zip. After formatting the sequence source (e.g. the entire collection of bacterial genomes from NCBI) with the included formatter, the background *k*-mer occurrence array could be computed using the BacterialIDkMerIndexBlockBuilder. This Builder can run on machines with as little as 2 GB RAM. A partition information tool for parallel computing and a *k*-mer space distribution analysis tool were also provided. After gathering all of the genomes for a particular group, the conserved *k*-mer occurrence index for this given group could be computed with the included Finder and Counter utilities. Finally, a tool (UniquelyConservedKmerFinder) compares the group's index against the background index to find UCR *k*-mers. Additional scripts were provided to analyze the genome distribution and annotation information of the identified UCR *k*-mers and UCRs as well as screening for optimal primers.

3 RESULTS

A total of 2717 bacterial genomes were obtained from the NCBI nucleotide database in December 2013. Because of the high probability of horizontal gene transfer for plasmids, only genomic DNA was included. The *k*-mer length was selected as 18, which can produce UCR *k*-mers that may be directly used as primers while at the same time maintaining a manageable index size (6.9×10^{10}). The background 18-mer occurrence index was computed, costing ~200 CPU-h (2.4 GHz), and the total uncompressed size was 137 GB. These 18-mer occurrences were largely uniformly distributed across the mapped integer space. Close to 90% of the 18-mers did not exist in the known bacterial genomes, and among those that did exist, only >90% had no more than five occurrences (Fig. 2). This indicates that the bacterial 18-mer space is unsaturated, which means that there are many potential UCR 18-mers.

3.1 Successful UCR identification in selected groups of bacteria

Sixteen groups were selected from these genomes to test this proposed UCR-identification method. This collection consists

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

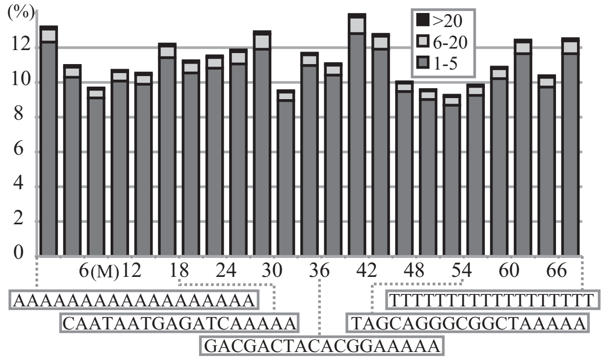


Fig. 2. The *k*-mer occurrence distribution histogram across the integer-mapped *k*-mer space (0∼6.9 × 10^{10}). The *k*-mer representations of some key integer coordinates (e.g. 0, 18 M, 36 M, 54 M) are noted. Each bar represents the percentage of *k*-mer within a given occurrence count range (i.e. 1∼5, 6∼20, >20) at that given interval

Table 1. UCR 18-mer identification result for the selected groups

Group	Role	GS	#G	#UCR
C.trachomatis	Obligate intracellular pathogen	1.0	70	577K
E.coli O157:H7	Pathogen	5.3	4	1799
E.coli	Pathogen/Commensal	5.0	61	0
S.aureus	Pathogen/Commensal	2.7	48	131K
P.aeruginosa	Opportunistic pathogen	6.2	11	502K
S.equi	Equine pathogen	2.1	4	611K
S.gallolyticus	Possible Pathogen/Carcinogen	2.2	3	392K
B.lactis	Probiotic bacterium	1.8	11	474K
L.casei	Probiotic bacterium	2.9	7	39K
L.plantarum	Probiotic bacterium	3.0	6	1674K
L.reuteri	Probiotic bacterium	1.9	5	367K
L.lactis	Probiotic bacterium	2.3	11	109K
P.fluorescens	Free living bacterium	6.5	6	87
S.islandicus	Acidophile and thermophile	2.5	10	829K
S. PCC 6803	Cyanobacteria (model)	3.4	6	2669K
T.thermophilus	Thermophile	1.8	4	352K

GS, genome size (MB); #G, number of genomes in this group; #UCR, number of UCR 18-mers discovered.

of groups that are of both clinical and veterinary importance and are also of industrial and academic interests, including one obligate intracellular human pathogen (*Chlamydia trachomatis*), one enterohemorrhagic foodborne pathogen (*Escherichia coli O157:H7*), the general *E.coli* species containing both pathogenic and commensal strains, one pathogenic/commensal species (*Staphylococcus aureus*), one opportunistic pathogenic species (*Pseudomonas aeruginosa*), one species containing animal pathogens (*Streptococcus equi*), one commonly found commensal bacterium in ruminants and also a human pathogen (*Streptococcus gallolyticus*), five probiotic bacteria (*Bifidobacterium animalis lactis*, *Lactobacillus casei*, *Lactobacillus plantarum*, *Lactobacillus reuteri*, *Lactococcus lactis*), one free-living bacterium and also a model organism (*P.fluorescens*), two extremophiles (*Sulfolobus islandicus*, *Thermus thermophilus*) and one cyanobacterium and also a model organism (*Synechocystis* PCC 6803) (Table 1).

3.2 Association between UCR distribution and bacterial lifestyle

The composition of the genes containing the most identified UCR 18-mers has an obvious correlation with group's lifestyle/characteristics (Table 2, and further details in Supplementary Table S1–S15). The groups that need to interact with host immune systems tend to have membrane/surface proteins that contain the most UCR 18-mers, while free-living groups tend to have housekeeping genes, such as metabolism and expression-related genes, containing the most UCR 18-mers. For example, 80% of the top 10 genes that contain the most UCR 18-mers are membrane/surface proteins in the *C.trachomatis* group, an obligate intracellular pathogen. In contrast, 60% of the top 10 genes in the *T.thermophilus* group are metabolism-related and 80% are in the *P.fluorescens* group. Other genes that frequently contain the most UCR 18-mers are RNA polymerases (e.g. *rpo*) and DNA polymerases (e.g. *pol*). Probiotic bacterial groups, whose top 10 genes contain both housekeeping genes and membrane proteins, need to interact with the host immune systems but do not require a host to survive. As a result, the UCR *k*-mer distribution among genes provided valuable information about the lifestyle/characteristics of a given group.

3.3 Selection and validation of UCR detection primers

Previously, most detection primers and targets were selected based on known properties of the pathogen, such as secreted toxins or genes involved in its pathogenesis. For example, the *ply* gene of the pneumolysin toxin was selected as the detection target for the pathogenic bacteria *Streptococcus pneumoniae* for its involvement in pneumococcal infections and interference with host immune response (Salo *et al.*, 1995). Similarly, the *nuc* gene of the extracellular thermostable nuclease was selected as the target for *S.aureus* (Brakstad *et al.*, 1992). And also the two primer sets that were designed specifically for the Shiga toxin 1 (*stx1*) and Shiga toxin 2 (*stx2*) genes of the *E.coli* O157:H7 (Fratamico *et al.*, 2000). Such primers were widely used; however, whether these primer sets met the requirements as an ideal detection target (i.e. first, conserved in, and second, unique to the specific pathogen species or group) were not examined when they were first proposed. It is therefore interesting to test whether these primer sets meet these requirements, especially now that there is a large number of available bacterial genomes. For the *S.pneumoniae* primer set (Salo *et al.*, 1995), 97% of the available *S.pneumoniae* genomes contain the exact match of the primer sequences. However, for the widely used *S.aureus* primer set (Brakstad *et al.*, 1992), 15% of genomes have mismatches within the five nucleotides relative to the 3′ terminus of the forward primer. Additionally, the *stx1* primer set has only significant matches in 50% of the *E.coli* O157:H7 genomes, and both *stx1* and *stx2* primers have significant matches even in species other than *E.coli*.

As a result, the UCR *k*-mer approach could help to refine the target selection and improve primer specificity and efficiency by systematically satisfying the detection target requirement

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

Table 2. Top 10 genes that have the most UCR 18-mers in each group

C.trachomatis

Excinuclease ABC subunit A, polymorphic outer membrane protein (×4)[a], protein translocase subunit, SWF/SNF family helicase, DNA polymerase III subunit alpha, putative membrane spanning protein (×2).

E.coli O157:H7

O-island #76 region (×3), prophage CP-933T proteins (i.e. tail fiber component, tail fiber protein, tail sheath protein, replication protein, serine acetlyltransferase, stability/partitioning protein), hypothetical protein.

S.aureus

Hypothetical protein SAR0284 (similar to diarrhoeal toxin BceT), Na^+/Pi^-cotransporter protein, phosphohydrolase, quinol oxidase polypeptide I, formate acetyltransferase, cell division protein, glucose-specific phosphotransferase transporter protein, transporter protein, putative Na^+/ Pi^-cotransporter.

P.aeruginosa

Peptide synthase, hemagglutinin, bifunctional prolinedehydrogenase/pyrroline-5-carboxylate dehydrogenase, pyochelin synthetase, hypothetical protein (×6).

S.equi

scpC chemokine protease, essC ESAT-6 secretion system protein, cell surface-anchored pullulanase, polC DNA polymerase III, ATP-dependent exonuclease subunit B, hyaluronate lyase precursor, glycosyl hydrolase family protein, iron transport-associated protein, glycosyl hydrolase family 2 protein, DNA gyrase subunit A.

S.gallolyticus

Cell envelope proteinase A, FtsK/SpoIIIE family protein, extracellular fructan hydrolase, gtfA glucosyltransferase, gtfB glucosyltransferase-T, cell wall bound protein, cell surface protein, adhesin Cna protein B-type domain, alpha-amylase, hypothetical protein.

B.lactis

Fibronectin type III domain-containing protein, ATP-dependent helicase II, sulfatase (×2), glycosyl transferase family 2, collagen adhesion protein, hypothetical protein (×4).

L.casei

bglH, uvrB excinuclease ABC subunit B, hypothetical protein, rmlD, Zn-dependent protease, outer membrane protein (×2), uvrA1 excinuclease ABC subunit A, rmlC.

L.plantarum

polC DNA-directed DNA polymerase III, mucus-binding protein precursor, aapA adherence-associated mucus-binding protein, cell surface protein precursor, rexA ATP-dependent nuclease subunit A, rexB ATP-dependent nuclease subunit B, mfd transcription-repair coupling factor, smc cell division protein, endo-beta-N-acetylglucosaminidase, membrane protein.

L.reuteri

rpoB DNA-directed RNA polymerase subunit beta, rpoC DNA-directed RNA polymerase subunit beta', polC DNA polymerase III, maltose phosphorylase, ileS isoleucyl-tRNA synthetase, transposase (×2), hypothetical protein (×2), ATP-dependent nuclease subunit B.

L.lactis

23S ribosomal RNA, rpoC DNA-directed RNA polymerase subunit beta', rpoB DNA-directed RNA polymerase subunit beta, gyrA DNA gyrase subunit A, ptsI phosphoenolpyruvate-protein phosphotransferase, mfd transcription-repair coupling factor, typA GTP-binding protein, celB PTS system cellobiose-specific transporter subunit IIC, rpoA DNA-directed RNA polymerase subunit alpha, pgi glucose-6-phosphate isomerase.

P.fluorescens

cyoB_2 cytochrome o ubiquinol oxidase subunit I, alpha/beta hydrolase, rsmY ncRNA, glutaredoxin, algE alginate biosynthesis protein, sucA 2-oxoglutarate dehydrogenase E1, ubiB 2-polyprenylphenol 6-hydroxylase, dnaX DNA polymerase III subunits gamma and tau, folC bifunctional protein folylpolyglutamate synthase/dihydrofolate synthase, acnB bifunctional aconitate hydratase2/2-methylisocitrate dehydratase.

S.islandicus

NADH dehydrogenase (quinone), rpoB DNA-directed RNA polymerase subunit B, peptidase S53 (×2), reverse gyrase (×2), alpha-L-rhamnosidase, carb carbamoyl phosphate synthase large subunit, alpha-mannosidase, ileS isoleucyl-tRNA synthetase.

Synechocystis PCC 6803

integrin subunit alpha (×2), integrin subunits alpha/beta4, hypothetical protein (×5), cadherin, nucH extracellular nuclease.

T.thermophilus

ribonucleoside-diphosphate reductase alpha chain, glutamate synthase [NADPH] large chain, sensory transduction histidine kinase, acetyl-coenzyme A synthetase, methyltransferase, acriflavin resistance protein B, hypothetical protein, anaerobic dimethyl sulfoxide reductase chain B, phosphoribosylformylglycinamidine synthase II.

[a](×N) represents N copies of this gene.

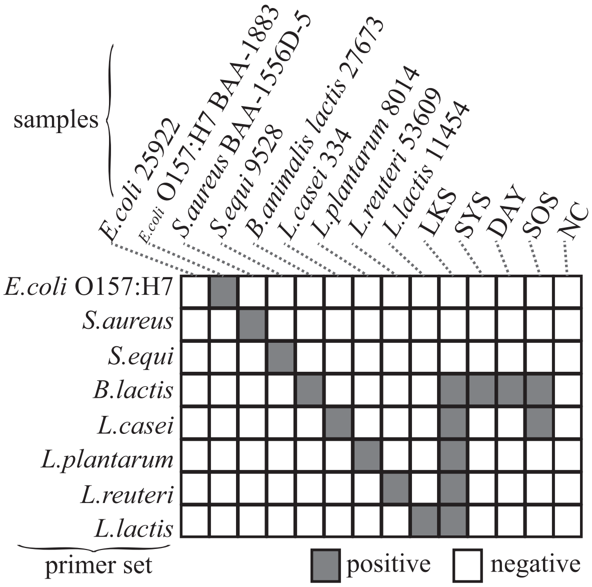
through conditions built into the algorithm (i.e. making sure that the primer sequences are first, conserved in, and second, unique to the specific pathogen species or group). With this method, based on the identified UCR 18-mers, detection primers were selected for 14 of the 15 groups (Table 3). These primers also met additional requirements that facilitate amplification and detection (e.g. having 0.6~1 kb product size, 46~54% GC%, 3′ GC clamps, <3°C melting temperature differences between forward and reverse primer pairs and <10kcal/mole heterodimer/homodimer ΔGs). There are many primer sets satisfying these requirements from within the UCRs for each genome, and one random set for each genome was chosen. The identified UCR *k*-mers for the *P.fluorescens* were too few in number and too far apart to produce any usable primers.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

Table 3. Group-specific primers based on UCR 18-mers

Group	Primer sequence (with location information[a])	T_m(°C)
C.trachomatis	>F-378288 AGCCTATGATCCGGTTACACTAC >R-379264 CGATTTGAATCCTTGGATTACTTCCTC	55.8
E.coli O157:H7	>F-2672459 CAGCAGAGCAAGACAAGCTGTC >R-2673569 AGCTCTTGCTGCATTGCTTCC	58.5
S.aureus	>F-117489 CATAACTGCTAAACGCTCAACGTC >R-118583 CCAGTATTAGGTGTTATTGCAGGTATCG	56.8
P.aeruginosa	>F-117547 GCATGAGTGTCGAGATCCATC >R-118509 ATTGGCCATGTAGCAATCCATG	55.7
S.equi	>F-1661447 CATTTGCTGCTTTCAGGCTGAC >R-1662460 TAGCCGCAATAGACCACAGAG	56.9
S.gallolyticus	>F-782211 CGCAGCTGGTATTACCTATGG >R-783069 CGGAATCGCAATCGCTGTATC	55.8
B.lactis	>F-1797010 GATCAGCAGTGTGAGGAACACG >R-1797901 GTGGCTCGTCATCATAGGGTTC	57.7
L.casei	>F-652180 CTCCTGTTGCCAACTGGTCAGG >R-652756 CTTGGTATTGTCCTGCTGTTGGTGTC	60.0
L.plantarum	>F-1846116 TCCACGCTTAGACAGGTCTTCC >R-1846870 GGATGCTCGAAGAGACTCACC	57.9
L.reuteri	>F-67606 CCCTTGGCATGTTGCTACTCAC >R-68339 CTTCTTTACCAGCAGCTACAGTACC	57.7
L.lactis	>F-120844 CTTGGTGAAGAAGCAGCATCTG >R-121770 GAACATTCCGTCACCAGCTGTTG	58.3
S.islandicus	>F-922564 CCCTACTATGTCCTTCATAGCC >R-922969 CAAACCAATTGTTGGAGGTAGACC	55.1
S. PCC 6803	>F-587255 TTAGTCTATGGGAAACTGTCCG >R-587973 GTTAAAGCTCCTCCTTGCTGAG	54.8
T.thermophilus	>F-1837389 GTCCGTGACCAGGTAGAAGCTC >R-1838051 CCCTTCAACGAGTACTATGTGC	59.0

[a] Location information is relative to the reference genome.



primer set \ samples	*E.coli* 25922	*E.coli* O157:H7 BAA-1883	*S.aureus* BAA-1556D-5	*S.equi* 9528	*B.animalis lactis* 27673	*L.casei* 334	*L.plantarum* 8014	*L.reuteri* 53609	*L.lactis* 11454	LKS	SYS	DAY	SOS	NC
E.coli O157:H7	□	▨	□	□	□	□	□	□	□	□	□	□	□	□
S.aureus	□	□	▨	□	□	□	□	□	□	□	□	□	□	□
S.equi	□	□	□	▨	□	□	□	□	□	□	□	□	□	□
B.lactis	□	□	□	□	▨	□	□	□	□	▨	▨	▨	▨	□
L.casei	□	□	□	□	□	▨	□	□	□	▨	□	□	▨	□
L.plantarum	□	□	□	□	□	□	▨	□	□	▨	□	□	□	□
L.reuteri	□	□	□	□	□	□	□	▨	□	▨	□	□	□	□
L.lactis	□	□	□	□	□	□	□	□	▨	▨	□	□	□	□

▨ positive □ negative

Fig. 3. PCR amplification results using UCR primer sets on pure cultures and mixture samples. Shadowed boxes indicate successful amplification, and white boxes indicate reactions without detectable amplification. The intended target names of the UCR primer sets are listed on the left side of the matrix, and the samples are noted on top of the matrix. ATCC product number is noted where applicable. LKS, SYS, DAY and SOS are dairy products containing live probiotic cultures (Supplementary Material). NC stands for negative control, which is DNA extracted from molecular grade ddH2O

To test the performance of the designed primers, eight sets of primers were used in polymerase chain reaction (PCR) amplifications of DNA extracted from both single strain cultures and materials containing mixed bacterial cultures (Fig. 3). The selected PCR primers are between 22 and 29 bp long, with an average length of 24 bp. The PCR product size is between 576 and 1110 bp, with an average of 887 bp. Single-strain bacterial cultures were ordered from ATCC, and dairy products containing live probiotic cultures were used to represent mixed bacterial culture samples (Supplemental Material). The PCR products were analyzed on a 2100 Bioanalyzer (Supplementary Material), and reactions that had clean band at the expected size were recorded as successful detections. These eight primer sets were able to specifically target the intended strain, even from samples containing multiple bacterial species and non-bacterial organisms.

4 DISCUSSION

The UCR identification method described here was able to successfully and automatically, with reasonable efficiency, identify regions that are both conserved in and unique to a selected bacterial group (a species or strain) solely based on their genomic sequences. These regions are by definition ideal diagnostic targets for nucleic acid hybridization, amplification and/or sequencing. Although high-throughput sequencing has been used in major medical centers to identify and track emerging epidemics and multidrug-resistant bacterial infections (Snitkin *et al.*, 2012), most medical facilities and reference laboratories lack the necessary equipment or personnel to take full advantage of these latest technologies. However, with this method, effective and accurate nucleic acid diagnostic targets could be rapidly identified from the data generated at those major medical centers or government agencies, and can be quickly adopted by most hospitals and laboratories with existing procedures such as PCR.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

The UCRs identified for the 15 bacterial species are now freely available at http://ucr.synblex.com. The source code of the programs used in this study is accessible at http://ucr.synblex.com/bacterialIdSourceCode.d.zip, while an automated pipeline for diagnostic target and primer selection is under development. Hopefully, this new method could rapidly and efficiently generate diagnostic targets that are both specific and sensitive, and ultimately lead to improved bacterial infectious disease management.

4.1 Intended usage and known limitations

A group must be clearly identified before the application of this method. Therefore, this method provides a possibility for large clinical testing centers, pharmaceutical companies and research laboratories to identify more efficient diagnostic markers, PCR primers and other products in managing bacterial infections caused by a known group of bacteria. Additionally, this method might also provide insights for the discovery of the markers of an unknown bacterial infectious agent present in patients with shared symptoms but without a clinically known disease, for example, by identifying UCRs in assembled patient microbiome data.

One significant drawback of this method is that it requires at least a few available genomes within a group to select UCRs and UCR primers. If only one or two genomes are available to represent a group, the properties of the generated UCRs and UCR primers are usually violated. However, the number of available genomes has increased significantly and is continuing to rise, which might be able to alleviate the impact of this drawback.

Additionally, as with any laboratory testing procedures, a test result only provides one piece of information on which the health-care providers make a diagnosis. Furthermore, it is known that pathogenic strains might be present asymptomatically within some individuals (e.g. silent carriers), while some environmental bacteria might cause severe infections in individuals with a compromised immune system. As a result, it is worth reiterating that the presence of a pathogenic strain might not immediately equate clinical symptoms, although such a discovery could provide valuable information about infectious disease dispersion and transmission. In contrast, the absence of all known pathogens, opportunistic or otherwise, is not evidence of an absence of bacterial infection. Therefore, in the latter cases, where infection symptoms are obvious, appropriate antibiotic treatment should be provided at the health-care providers' discretion, especially to individuals with known immunodeficiency.

4.2 Applicability on partially assembled genomes and limitations with raw sequencing reads

This method could be used on partially assembled genomes. However, due to its principle and how the raw sequencing reads were currently generated, this method should not be used on raw sequence data when building the background *k*-mer index. Currently, on most NGS-sequencing platforms (e.g. Illumina, Ion Torrent, 454, etc.), the error rate for each single raw read is relatively high. Only through combining multiple raw reads together at the same location through mapping or *de novo* assembly could a reasonably reliable sequence be achieved. Although error-containing *k*-mers are unlikely to become conserved within a group, this only ensures that one of the UCR conditions (i.e. conserved in a specific group) is not violated. When the background *k*-mer index is built, and if raw sequencing reads are allowed, these error-containing *k*-mers will be calculated in the background *k*-mer index. Subsequently, a random match between a true UCR *k*-mer for one group and an error-containing *k*-mer from another group would automatically rule out this true UCR *k*-mer in the selection process, as this by definition violates the other UCR condition (i.e. unique to a specific group). Currently, with billions of reads generated for each experiment, a mere 3% error-containing reads could still yield millions of error-containing *k*-mers, which would create a substantial amount of noise in the background *k*-mer index, and therefore, limit this method's selection power for *k*-mers unique to a given group. As a result, raw sequencing reads should not be used in the process of building the background *k*-mer index. In contrast, if the background index has already been built with completed genomes, and multiple samples of a new group were sequenced, UCR *k*-mers could be identified directly from the raw sequencing reads with this method. This will be a major functionality in the future release of the fully automated pipeline.

Although partially assembled genomes contain gaps, the error rate for each contig has been reduced to an acceptable level through at least one round of assembly. The gaps (or break points) between contigs will prevent the sequences spanning these gaps from being indexed, and thus, the sequence information close to the gaps would be lost. However, a routine microbial genome sequencing project would typically generate no more than 50–100 large contigs, and these 50–100 gaps represent only a limited amount of lost information, especially when the *k*-mer length is several orders shorter than that of a contig.

4.3 UCRs as a concise reference database in microbiome screening

In a typical microbiome screening, one key step in the analysis pipeline is to identify the sources of the sequences within the sample, usually by comparing these sequences against reference databases. For example, for a metagenomic screening with 16S amplicons, the raw reads could be compared against popular databases such as Silva (Quast *et al.*, 2012), RDP (Cole *et al.*, 2013) or Greengenes (McDonald *et al.*, 2011). For a typical WGS metagenomic screening, the raw reads could be directly compared against all known genomes or the NCBI nr/nt databases. This UCR method is capable of identifying regions that can serve as markers for some of the known bacterial species, as demonstrated in this study. If UCRs could be identified for most clinically relevant species and/or strains, and with these UCRs as a concise reference database, a routine clinical analysis would only need to compare each raw metagenomic read with the relatively few number of identified UCRs instead of the entire known collection of genomes or the NCBI nr/nt databases. As a result, in theory, the computational cost could be drastically reduced.

However, as mentioned earlier, without considering contamination and other human errors, the instrument error rate at the raw sequencing read level is relatively high. A match between a UCR and a raw sequencing read might be owing to such errors.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

Therefore, as in any NGS analyses involving raw reads, only after observing a significant amount of matched raw reads for a given UCR, i.e. achieving a certain coverage threshold, could one confidently conclude the presence of the organism associated with this UCR. Additionally, the fact that only a tiny percentage of organisms have sequenced genomes dictates that UCRs can only be calculated for a limited number of species or groups. As a result, using UCRs as a reference database might be able to drastically reduce computational time, but this is appropriate when only known organisms (e.g. pathogens) are of interest in a microbiome study, such as a routine clinical screening. In other scientific studies, where the information from the previously unknown species is of equal interest, the UCR reference database is of limited usefulness.

4.4 Protein versus nucleic acid space *k*-mer indexing

Although a bacterial genome is mostly protein coding, and the capacity of 8-mer peptide (2.6×10^{10} index length) is comparable with that of the 18-mer DNA/RNA sequence, due to codon degeneracy, and therefore, the existence of synonymous mutations, protein sequences are more conserved than DNA sequences between species. As a result, traditionally, intraspecific studies and studies involving closely related species usually use nucleotide sequences, as they have a higher observable substitution rate, and hence, higher resolution for short evolutionary distances. In contrast, studies involving long evolutionary processes (e.g. over millions of years), usually use protein sequences, as their relatively low substitution rate retains alignable sequence patterns better and is less prone to the effect of multiple hits over a long evolutionary process. With the purpose being identifying diagnostic markers that are capable of differentiating between species or strains, nucleic acid sequence markers are more appropriate. However, for other genomes with an extremely high mutation rate (e.g. viral genomes), nucleic acid UCRs might not be feasible because of the higher intraspecific/interstrain differences. In this case, peptide UCRs would be more appropriate, as their protein sequences are more conserved.

4.5 Possibilities and limitations beyond bacterial genomes

This method was demonstrated with bacterial genomes. Theoretically, this method could also be applicable in eukaryotes. However, the limited number of available genomes within a eukaryotic species has severely restricted applicability. For example, even with the 1000 Fungal Genome Project, there are currently <200 publically available fungal genomes, and on average, there is only one complete genome for each fungal species. Without a substantial number of genomes sampled for each species or strain, there would not be any conservation information for a given group, nor would there be a robust background *k*-mer index. With an increasingly faster genome generation rate, and more and more completed fungal genomes, it is hoped that this limitation could soon be addressed.

Although close to 2000 viral genomes have been published, the application on viruses is more challenging. Because of their high mutation rate, and the resulting low sequence similarity between genomes within the same viral species, we have failed to identify any viral nucleic acid UCRs on a species level (with $k > 15$). Instead, peptide UCRs might be possible and more appropriate. The same algorithm could be adapted to index and screen for uniquely conserved peptide sequences, but validating the results would be significantly more challenging, and viral processing and protein sequencing are beyond the capability of this research group.

ACKNOWLEDGEMENTS

The authors sincerely thank Prof. Qi Ji (Fudan University) for discussions related to the *k*-mer index method. And they also thank the three anonymous reviewers for their critical and constructive comments.

Funding: This work was supported by the Synblex LLC research fund.

Conflict of interest: none declared.

REFERENCES

Brakstad,O.G. *et al.* (1992) Detection of Staphylococcus aureus by polymerase chain reaction amplification of the nuc gene. *J. Clin. Microbiol.*, **30**, 1654–1660.

Caliendo,A.M. (2011) Multiplex PCR and emerging technologies for the detection of respiratory pathogens. *Clin. Infect. Dis.*, **52** (**Suppl. 4**), S326–S330.

Cole,J.R. *et al.* (2013) Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res.*, **42**, D633–D642.

David,M.Z. and Daum,R.S. (2010) Community-associated methicillin-resistant Staphylococcus aureus: epidemiology and clinical consequences of an emerging epidemic. *Clin. Microbiol. Rev.*, **23**, 616–687.

Dieffenbach,C.W. *et al.* (1993) General concepts for PCR primer design. *PCR Methods Appl.*, **3**, S30–S37.

Fratamico,P.M. *et al.* (2000) A multiplex polymerase chain reaction assay for rapid detection and identification of Escherichia coli O157:H7 in foods and bovine feces. *J. Food Prot.*, **63**, 1032–1037.

Hoorfar,J. (2012) Rapid detection, characterization, and enumeration of foodborne pathogens. *APMIS Suppl.*, **133**, 1–24.

Koressaar,T. and Remm,M. (2007) Enhancements and modifications of primer design program Primer3. *Bioinformatics*, **23**, 1289–1291.

Kruse,E.B. *et al.* (2013) Carbapenem-resistant enterobacteriaceae: laboratory detection and infection control practices. *Curr. Infect. Dis. Rep.*, [Epub ahead of print].

Mathers,A.J. *et al.* (2012) First clinical cases of OXA-48-producing carbapenem-resistant Klebsiella pneumoniae in the United States: the “menace” arrives in the new world. *J. Clin. Microbiol.*, **51**, 680–683.

McDonald,D. *et al.* (2011) An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.*, **6**, 610–618.

Miller,M.B. and Tang,Y.W. (2009) Basic concepts of microarrays and potential applications in clinical microbiology. *Clin. Microbiol. Rev.*, **22**, 611–633.

Morse,S.S. (2012) Public health surveillance and infectious disease detection. *Biosecur. Bioterror.*, **10**, 6–16.

Neuberger,A. *et al.* (2008) Clinical impact of a PCR assay for rapid identification of Klebsiella pneumoniae in blood cultures. *J. Clin. Microbiol.*, **46**, 377–379.

Newell,D.G. *et al.* (2010) Food-borne diseases–the challenges of 20 years ago still persist while new ones continue to emerge. *Int. J. Food Microbiol.*, **139** (**Suppl. 1**), S3–S15.

Quast,C. *et al.* (2012) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.*, **41**, D590–D596.

Salo,P. *et al.* (1995) Diagnosis of bacteremic pneumococcal pneumonia by amplification of pneumolysin gene fragment in serum. *J. Infect. Dis.*, **171**, 479–482.

Snitkin,E.S. *et al.* (2012) Tracking a hospital outbreak of carbapenem-resistant Klebsiella pneumoniae with whole-genome sequencing. *Sci. Transl. Med.*, **4**, 148ra116.

Vitter,J.S. (2008) *Algorithms and Data Structures for External Memory*. Now Publishers Inc, Hanover, MA, USA.

Yang,S. and Rothman,R.E. (2004) PCR-based diagnostics for infectious diseases: uses, limitations, and future applications in acute-care settings. *Lancet Infect. Dis.*, **4**, 337–348.

Ye,J. *et al.* (2012) Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics*, **13**, 134.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016