Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Nucleotide 9-mers characterize the type II diabetic gut metagenome

Balázs Szalkai [a], Vince Grolmusz [a,b,*]

[a] PIT Bioinformatics Group, Eötvös University, H-1117 Budapest, Hungary
[b] Uratim Ltd., H-1118 Budapest, Hungary

ARTICLE INFO

Article history:
Received 21 September 2015
Received in revised form 25 February 2016
Accepted 26 February 2016
Available online xxxx

Keywords:
Type II diabetes
Metagenome
Biomarkers
G-C content
Oligomers
Short nucleotide sequences
Tetranucleotides

ABSTRACT

Discoveries of new biomarkers for frequently occurring diseases are of special importance in today's medicine. While fully developed type II diabetes (T2D) can be detected easily, the early identification of high risk individuals is an area of interest in T2D, too. Metagenomic analysis of the human bacterial flora has shown subtle changes in diabetic patients, but no specific microbes are known to cause or promote the disease. Moderate changes were also detected in the microbial gene composition of the metagenomes of diabetic patients, but again, no specific gene was found that is present in disease-related and missing in healthy metagenome. However, these fine differences in microbial taxon- and gene composition are difficult to apply as quantitative biomarkers for diagnosing or predicting type II diabetes. In the present work we report some nucleotide 9-mers with significantly differing frequencies in diabetic and healthy intestinal flora. To our knowledge, it is the first time such short DNA fragments have been associated with T2D. The automated, quantitative analysis of the frequencies of short nucleotide sequences seems to be more feasible than accurate phylogenetic and functional analysis, and thus it might be a promising direction of diagnostic research.

1. Introduction

Metagenomics [1] is rapidly gaining importance in clinical research [2–9], environmental studies [10–12] and biotechnology [13–15]. Numerous complex and reliable methods have been published for the phylogenetic identification of non-cloned short DNA reads from environmental or clinical samples, for example, the similarity-based methods MEGAN [16–18] and MG-RAST [19,20], the marker-gene identifying phylogenetic analyzer AMPHORA [21] and its more user-friendly versions, AMPHORA2 [22] and AmphoraNet [23,24].

These methods use multi-phase, complex approaches to retrieve phylogenetic information from the short read datasets, applying reference database operations in the process.

Surprisingly, it was shown that simple frequency counting of nucleotides or short nucleotide sequences in the metagenomic samples may also imply phylogenetic information.

It has been widely known for a long time that genomic AT/GC ratio is distributed in a wide range in bacterial species, and can be characteristic to some of them [25–27]. The ratio is shown to be influenced by numerous environmental and metabolic factors [28] and also carries phylogenetic information.

The article [29] reports differences in di- and tetranucleotide frequencies among numerous bacterial species, and examines the possible application of these signatures in molecular phylogeny.

Tetranucleotide sequence frequencies were applied in supervised and unsupervised phylogenetic classification, or “binning” in [30].

The work [31] applies conserved gene fragments, each encoding several dozens of amino acids, identified from the Pfam database [32]. The fragments are called “environmental gene tags”, and are used successfully for phylogenetic binning in [31].

The study of [3] investigated the differences in gut metagenomes of diabetic and healthy subjects. The metagenomes were *de novo* assembled, and the bacterial genes were mapped to a metagenomic gene catalog. Genes related to oxidative stress response were found more abundant in the samples originating from diabetic subjects. Additionally, moderate changes in intestinal bacterial composition were detected, but no specific microbes were associated with the metagenomes of the type II diabetes (T2D) patients.

After a very complex selection and filtering process, genome-specific nucleotide markers of length 50 were identified in [33]. The markers were applied for strain/species identification, and also as markers for microbial species that might play a role in T2D and obesity in the data set of [3].

Here we describe a very simple and straightforward approach for finding short nucleotide sequences whose frequencies significantly differ in T2D and healthy metagenomes of the dataset of [3]. We identify several nucleotide 9-mers that may serve as quantitative biomarkers of the pre-diabetic state in the future. To our knowledge, such short sequences have never been found to characterize T2D or any other disease.

We need to clarify that we do not state that the identified 9-mers will generally be applicable as biomarkers for diabetes for all human

[*] Corresponding author.
E-mail addresses: szalkai@pitgroup.org (B. Szalkai), grolmusz@pitgroup.org (V. Grolmusz).

http://dx.doi.org/10.1016/j.ygeno.2016.02.007
0888-7543/© 2016 Elsevier Inc. All rights reserved.

populations. We believe that "enterotype-specific" [34] quantitative biomarkers could be found for each enterotypes by exhaustive searches described in the Methods section, and those enterotype-specific biomarkers could serve as predictors of type 2 diabetes mellitus.

2. Discussion and results

Our results are summarized in Table 1 and in Fig. 1. Table 1 contains twenty 7-, 8- and 9-mers of the highest statistical significance, distinguishing between the diabetic and non-diabetic metagenomes of the study [3].

Table 1 was prepared without considering complementarities between the short nucleotide sequences. Therefore, the complements found with very close frequencies and statistical parameters independently verify our results. It is easy to recognize in Table 1 that TGTGGTA and TACCACA are exact complements. The complement of TCCACAT, ATGTGGA, is almost the prefix of ATGTGGTAC. The complement of TGTGGTACT (line 3) is again the exact complement of AGTACCACA (line 6), just to mention some of the complementarities in the table.

Fig. 1 gives the empirical cumulative distribution functions of the frequency of 9-mer TGTGGTGTA in the diabetic and in the non-diabetic samples. The difference between the expected values (means) of the two distribution is obvious on the figure and is quantified statistically in Table 1.

We also searched for short nucleotide sequences characterizing lean/obese and male/female individuals in the dataset of [3]. Only one short sequence passed even a rather large statistical significance bound of about 0.1 in the lean/obese search, and none in the male/female search (c.f. Tables S1 and S2 and Fig. S1 in the Supplementary material).

The source of the bias in short nucleotide sequence frequencies is most probably due to the difference in the gene- and species composition of diabetic and healthy metagenomes, found in [3,33]. These frequencies could be measured and evaluated more easily than the much more involved characteristics found in [3,33].

3. Methods

Our data source was the set of metagenomes of 345 Chinese subjects, collected by Qin et al. [3] and deposited in the Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra) under accession numbers SRA045646 (145 subjects) and SRA050230 (225 subjects). The assembled data was downloaded from the GigaScience database, GigaDB at http://dx.doi.org/10.5524/100036.

We considered all the possible DNA sequences of length at most 9 (this means over 300,000 possible sequences). For each sequence, we counted the number of exact matches in each raw metagenome. Our aim was to determine whether there are any short DNA fragments whose frequencies differ for diabetic/non-diabetic, lean/obese or female/male individuals. We had to draw the line at 9 nucleotides, because calculating the frequency of longer sequences is computationally more expensive, and the chance of a false positive greatly increases when testing a large number of sequences.

We first defined the frequency of a short DNA fragment for a given metagenome as the number of occurrences (exact matches), divided by the total size, measured in base-pairs (bp), of the metagenome. Additionally – to account for minor mutations – we also included those sequences in the counting process that differed by only one nucleotide, but these were considered with half a weight. So, for example, the final *frequency* of the sequence AAA included not only how many times the sequence AAA occurs in a specific metagenome, but also how many times AAG, CAA, ATA, … occur in that metagenome, except that the number of occurrences for these related DNA fragments was divided by two.

Let ℓ_M denote the length in base-pairs (bp) of a metagenome M. Suppose we want to count the all occurrences of a certain sequence s of length v. If w is a contiguous sequence of the metagenome of length u, then our sequence s could be present as a gap-less subsequence in w at most $u - v$ times: every nucleotide of w could serve as a potential starting position for s except the last $v - 1$ positions of w. Therefore, ℓ_M can serve as an upper bound to the all possible appearances of s in M.

Let $d(s,t)$ be the number of mismatches between the two sequences of same length, s and t (also called the Hamming distance). Let $k_M(s)$ de-

Table 1
Frequencies of 7-, 8- and 9-mers in diabetic vs. non-diabetic samples with the highest significance (training set: Study 1, holdout set: Study 2). The columns of the table are: the sequence itself, the frequencies for diabetic and non-diabetic subjects, the *p*-value for the training and the holdout sets, the corrected *p*-value for the holdout set (multiplied by the factor determined by the Benjamini-Hochberg correction), and the false discovery rate for the fragments so far. Choosing an FDR of about 7% allows us to make 15 discoveries, expecting about 1 of them to be false, but the real FDR should be lower due to strong positive correlation among the tests. It is easy to recognize that TGTGGTA and TACCACA are exact complements. The complement of TCCACAT, ATGTGGA, is almost the prefix of ATGTGGTAC. 9-mer TGTGGTACT (line 3) is the exact complement of AGTACCACA (line 6). One can find further complementarities in the table. These independently found complements with very close frequencies and *p*-values strengthen our findings. More tables (for lean-obese and female-male distributions) are given in the Supplementary material.

Fragment	Diabetic	Non-diabetic	*p* (training set)	*p* (holdout set)	*p* (corrected)	FDR
TGTGGTGTA	4.48E-05	4.71E-05	7.80E-09	0.000296	0.021151852	0.021151852
TGTGCTATC	4.35E-05	4.55E-05	1.87E-08	0.001764	0.063026802	0.063026802
TGTGGTACT	4.01E-05	4.16E-05	9.51E-10	0.001929	0.04594811	0.063026802
TGTGGTA	0.0006214	0.0006428	1.40E-08	0.001937	0.034604001	0.063026802
TGTGGTACA	4.67E-05	4.88E-05	2.97E-08	0.002098	0.029984179	0.063026802
AGTACCACA	4.10E-05	4.24E-05	2.15E-08	0.002246	0.02674947	0.063026802
CCATCTGT	0.0002318	0.0002424	2.14E-08	0.003092	0.031564443	0.063026802
TGCCACATA	5.81E-05	6.13E-05	6.42E-09	0.004678	0.041785626	0.063026802
TGTGGTATG	4.81E-05	5.04E-05	9.19E-09	0.004925	0.03910393	0.063026802
TACCACA	0.0006332	0.0006531	3.38E-08	0.004999	0.035722333	0.063026802
TGTGGAGAT	6.54E-05	6.79E-05	1.52E-08	0.008901	0.05782329	0.063026802
TGTGGTATC	5.04E-05	5.25E-05	1.49E-08	0.011902	0.070875377	0.070875377
ATGGTCTGT	5.85E-05	6.07E-05	1.29E-08	0.012383	0.068067407	0.070875377
GTACCACAT	4.18E-05	4.31E-05	1.06E-08	0.012814	0.065405364	0.070875377
CCACATACT	5.13E-05	5.35E-05	2.44E-08	0.014294	0.068095624	0.070875377
ATGTGGTAC	4.14E-05	4.27E-05	9.50E-09	0.02434	0.108706941	0.108706941
TCTCCACAT	6.97E-05	7.26E-05	1.58E-08	0.07478	0.314335349	0.314335349
ATCTCCACA	6.62E-05	6.84E-05	5.43E-09	0.078516	0.311703978	0.314335349
CTCCACATA	5.58E-05	5.75E-05	2.02E-08	0.257111	0.966993912	0.966993912
TCCACAT	0.0008132	0.0008294	1.92E-08	0.266428	0.951933372	0.966993912

Please cite this article as: B. Szalkai, V. Grolmusz, Genomics (2016), http://dx.doi.org/10.1016/j.ygeno.2016.02.007