Application of shotgun metagenomics to smoked salmon experimentally spiked: Comparison between sequencing and microbiological data using different bioinformatic approaches

Alessandra De Cesare,[1] Chiara Oliveri,[2] Alex Lucchi,[2] Frederique Pasquali,[2] Gerardo Manfreda[2]

[1]Department of Veterinary Medical Sciences, and [2]Department of Agricultural and Food Sciences, Alma Mater Studiorum University of Bologna, Italy

Abstract

The aims of this study were i) to evaluate the possibility to detect and possibly quantify microorganisms belonging to different domains experimentally spiked in smoked salmon at known concentrations using shotgun metagenomics; ii) to compare the sequencing results using four bioinformatic tools. The salmon was spiked with six species of bacteria, including potential foodborne pathogens, as well as *Cryptosporidium parvum, Saccharomyces cerevisiae* and *Bovine alphaherpesvirus* 1. After spiking, the salmon was kept refrigerated before DNA extraction, library preparation and sequencing at 7 Gbp in paired ends at 150 bp. The bioinformatic tools named MG-RAST, OneCodex, CosmosID and MgMapper were used for the sequence analysis and the data provided were compared using STAMP. All bacteria spiked in the salmon were identified using all bioinformatic tools. Such tools were also able to assign the higher abundances to the species *Propionibacterium freudenreichii* spiked at the highest concentration in comparison to the other bacteria. Nevertheless, different abundances were quantified for bacteria spiked in the salmon at the same cell concentration. *Cryptosporidium parvum* was detected by all bioinformatics tools, while *Saccharomyces cerevisiae* by MG-RAST only. Finally, the DNA virus was detected by CosmosID and OneCodex only. Overall, the results of this study showed that shotgun metagenomics can be applied to detect microorganisms belonging to different domains in the same food sample. Nevertheless, a direct correlation between cell concentration of each spiked microorganism and number of corresponding reads cannot be established yet.

Introduction

Shotgun metagenomics has been applied for the detection, identification and characterisation of pathogens in foods (Aw *et al*., 2016; Leonard *et al*., 2015, 2016) and in the food chain environment (Yang *et al*., 2016). It certainly provides an opportunity to survey the diversity and the dynamic abundance of microorganisms, including pathogens, within a food sample in a less biased manner than amplicon sequencing (Forbes *et al*., 2017; Jagadeesan *et al*., 2018), although there are still many drawbacks in terms of standardization and validation of this sequencing strategy.

Performing high-throughput shotgun sequencing of total nucleic acids obtained from foods results in a large and complex data sets that can be used to investigate both taxonomic composition and, potentially, functional capacity of the entire food ecosystem under study (Lindgreen *et al*., 2015). Factors that can affect microorganism identification and abundance include sample handling (Lewandowska *et al*., 2017, Wylezich *et al*., 2018), nucleic acid extraction (Knudsen *et al*., 2016), library preparation (Jones *et al*., 2015) and sequencing platforms (Tremblay *et al*., 2015) but also sequence analyses.

Many EU and global institutions perform sequence analysis by using internal pipelines which are not publicly available or pipelines which are in the public domain but combined in an unknown way. Among the few data analysis tools public available there are MG RAST (Keegan *et al*., 2016), which is public and free (www.mg-rast.org); OneCodex (Minot *et al*., 2015) (www.onecodex.com) and CosmosID (Yan *et al*., 2018) (https://app.cosmosid.com/) which are public but not free for the analysis of many metagenomes; MGMapper (Petersen *et al*., 2017), hosted at the CGE, now call CCMetagen 1.0 (https://cge.cbs.dtu.dk/services/MGmapper/) which is public, free but not always updated in the web version.

To contribute to assess the suitability of shotgun metagenomics to detect a wide range of target microorganisms in foods, a proficiency test (PT) was organised as part of the COMPARE project (www.compare-europe.eu) involving 11 Partners from inside and outside the EU. The aims of the trial were (1) to check to which extent bacteria, viruses and eukaryotes were detected and quantified in the metagenomes obtained by the Participants using their own wet lab procedures for shotgun metagenomics of smoked salmon experimentally spiked; (2) to identify which steps in the wet lab protocols mostly affect the microorganism detection and quantification results. In the study described in this paper three samples of smoked salmon obtained using the same wet lab protocols were analysed using the four bioinformatic tools described above to select the best dataset to provide to the COMPARE PT.

Correspondence: Alessandra De Cesare, Department of Veterinary Medical Sciences, Alma Mater Studiorum-University of Bologna, via Tolara di Sopra 50, 40064 Ozzano dell'Emilia (BO), Italy.
Tel.: +39.051.2097583 - Fax: +39.051.2097852
E-mail: alessandra.decesare@unibo.it

Key words: Shotgun metagenomics, smoked salmon, microbiological hazards, bioinformatic tools.

Contributions: AD data analysis and writing, CO sample processing and data analysis; AL and FP sample processing, GM manuscript editing and review.

Conflict of interests: the authors declare no potential conflict of interests.

Funding: the work was supported by the EU founded project COMPARE (Grant Agreement N° 643476).

Received for publication: 1 August 2019.
Revision received: 10 October 2019.
Accepted for publication: 10 October 2019.

Italian Journal of Food Safety 2019; 8:8462
doi:10.4081/ijfs.2019.8462

Materials and Methods

A total of 0.2 g of cold-smoked salmon were cut in very small pieces and transferred to Nunc screw cap tubes. Subsequently, each tube was kept on ice and spiked with 50 µL of a mock community consisting of bacteria (*i.e., Propionibacterium freudenreichii, Staphylococcus aures, Bacteroides fragilis*, *Escherichia coli*, *Fusobacterium nucleatum* and *Salmonella enterica)* as well as *Cryptosporidium parvum, Saccharomyces cerevisiae* and the heat-inactivated *Bovine alphaherpesvirus* 1 (Table 1). After the spiking, each tube was vortex-mixed and placed at refrigeration temperature. The DNA was extracted using PowerFood® Microbial DNA Isolation kit (MoBio) and

then fragmented and tagged with sequencing indexes and adapters using Nextera XT DNA Library Preparation Kit (Illumina, San Diego, CA). Sequencing was performed using NextSeq500 (Illumina) at 2 ×150 bp, in paired-end mode. The metagenomes were characterized by an average output of 7 Gbp. Filtering and trimming of raw reads and taxonomic classification were performed using four different web data analysis tools represented by MG-RAST, OneCodex, CosmosID and MGMapper. In MG-RAST, the taxonomic classification was performed using the RefSeq reference database (Pruitt *et al*., 2005) as well as Silva LSU, Silva SSU, RDP and Greengenes. In OneCodex, the One Codex database was used and in CosmosID the GenBook database. Finally, for MGMapper the database selected was Silva. The results of abundance of each taxonomic level for each sample were analyzed using the Statistical Analysis of Metagenomic profile Software v 2.0.9 (STAMP) (Parks *et al*., 2014). The statistical differences between the outputs of different bioinformatics tools were not assessed because only three samples were available for each combination of tool/database. The metagenomes of this study are public available in MG-RAST under the study FOOD METAGENOMIC RING TRIAL 2018 with the codes M30, M31 and M32.

Results and Discussion

Ni *et al*. (2013) state that the genome of a single species can be accurately assembled from a complex metagenomic dataset when it shows roughly at least 20-fold coverage, meaning that there are 20-fold sequence data covering that specific genome. According to their calculation at least 7 Gbp of sequencing output is required to enumerate the gene contents of prokaryotes with relative abundance of more than 1% in a microbiome. Therefore, 7 Gbp has been selected as sequencing depth in this study with the aim to correlate the concentration of spiked microorganisms with the abundance of their reads.

The MG-RAST outputs represented by the percentage abundances obtained for each microorganism of the mock community using the databases available in the software tool are summarised in Table 2. According to Petersen *et al*., 2017, within a dataset obtained by shotgun metagenomics, the taxonomic classification of a microorganism can be considered correct when the ratio between the number of reads associated to that microorganism and the total number of reads in the metagenome is >0.1%. Using the RefSeq database, all the bacteria of the mock community were identified and those spiked at higher concentrations were quantified with percentage abundances >10% (Table 2). Nevertheless, the bacteria spiked at the concentration of 50,000,000 cells showed different percentage abundances, ranging between 9.41 and 1.62% (Table 2). Percentage abundances >10% were obtained for *Propionibacterium freudenreichii* also by Silva SSU, RDP and Greengenes. However, using these databases, *S. aureus*, which was also spiked at 500,000,000 as *Propionibacterium freudenreichii*, was quantified at lower abundances, ranging between 2.32 and 6.23% (Table 2). As for RefSeq, using Silva LSU, Silva SSU, RDP and Greengenes the bacteria spiked at the concentration of 50,000,000 cells were quantified with percentage abundances ranging between 6.95 to 0.16% (Table 2). Both *C. parvum* and *S. cerevisiae* were detected using RefSeq, although at abun-

Table 1. Composition of the mock community used to spike the samples of cold smoked salmon and concentration of each microorganism.

Taxon	Amount per subsample (cells/virus gene copies)
Propionibacterium freudenreichii subsp. *freudenreichii* DSM 20271	500,000,000
Staphylococcus aureus subsp. *aureus* NCTC 8325	500,000,000
Bacteroides fragilis NCTC 9343 / DSM 2151	50,000,000
Fusobacterium nucleatum subsp. *nucleatum* ATCC 25586 / DSM 15643	50,000,000
Escherichia coli ATCC 25922	50,000,000
Salmonella enterica subsp. *enterica* serovar *Typhimurium* str. ATCC 14028S / DSM 19587	50,000,000
Cryptosporidium parvum IOWA II isolate	1,000,000
Saccharomyces cerevisiae S288C	5,000,000
Bovine alphaherpesvirus 1 (ds DNA virus)	1,20E+10

Table 2. Abundance values (%) obtained for the microorganisms of the mock community by MG-RAST with the databases RefSeq, Silva LSU, Silva SSU, RDP and Greengenes.

Species	RefSeq	SILVA LSU	SILVA SSU	RDP	GREENGENES
P. freudenreichii	23.08	3.82	14.91	19.14	24.25
S. aureus	10.13	2.32	6.23	3.33	3.10
B. fragilis	9.41	1.54	4.49	6.95	6.92
F. nucleatum	1.62	1.59	1.60	2.80	2.03
E. coli	4.79	2.07	4.32	0.63	1.51
S. Typhimurium	8.72	1.25	1.52	0.16	1.34
C. parvum	0.15	0.01	0.13	ND	ND
S. cerevisiae	0.01	<0.01	<0.01	ND	ND
B. alphaherpesvirus	ND	ND	ND	ND	ND

ND: not detected

dance values very close to the cut off level for correct taxonomic classification. The same result was obtained using Silva LSU and Silva SSU for the parasite, which was not detected using RDP and Greengenes. Similar results were observed for the yeast, which was detected at very low abundances by Silva LSU and Silva SSU. Finally, the DNA virus was not identified by MG-RAST with any database (Table 2).

Since the MG-RAST outputs achieved using RefSeq corresponded to the higher percentage abundances of the microorganisms of the mock community they were compared with the results obtained by MGMapper, CosmosID and OneCodex (Table 3). All these data analysis tools are reference based because the data collected in a well performed metagenomic project are sufficient to characterize the major functions of the microbial communities as well as to identify their taxon (Nielsen *et al*., 2014). The percentage abundances of *Propionibacterium freudenreichii* quantified by CosmosID and OneCodex were 45.65 and 63.34%, respectively, whereas those of other bacteria never exceeded 20% neither for *S. aureus* spiked at a concentration of 500,000,000 cells (Table 3). For the bacteria spiked at the concentration of 50.000.000 cells the detected values were very diverse either within the same bioinformatic tool as well as between them. MGMapper provided the lower percentage abundances for all species of bacteria, whereas CosmosID produced the higher percentages. Besides, it performed very well also for the parasite and the DNA virus. Nevertheless, it was not able to detect the yeast. Both the parasite and the DNA virus were also detected using OneCodex, although at lower abundances in comparison to CosmosID. Besides, OneCodex was not able to detect the yeast neither.

Among the tested bioinformatic tools, OneCodex and CosmosID are the most user friendly in terms of sequence upload and data interpretation. The CosmosID databases are organized phylogenetically and contain hundreds of millions of marker gene sequences. The markers represent both coding and non-coding sequences uniquely identified by taxon and/or distinct nodes of phylogenetic trees. This means that the tree structure was created based on genomic relatedness of organisms rather than predetermined taxonomy based on phenotype. This allows CosmosID to have a high degree of accuracy in identifying microorganisms based on their DNA in metagenomic samples. It also helps identify the closest match to genomes that do not have strain level references in the database (if, for example, they have never been sequenced before). However, as far as quantification results are concern, the high percentage abundances detected using CosmosID for the microorganisms of the mock community are due to the fact that the abundance analysis is done for each domain separately. Therefore, an abundance of 88.74% for *C. parvum* it does not mean that the parasite reads represent the majority of the reads of the metagenome but the majority of the reads assigned to eukaryotes.

One Codex identifies microbial sequences using a "k-mer based" taxonomic classification algorithm as CosmosID and MG-RAST, but it is built on a web-based data platform, using a reference database that currently includes approximately 40,000 bacterial, viral, fungal, and protozoan genomes. Quantitative evaluation of several published microbial detection methods shows that One Codex has the highest degree of sensitivity and specificity (AUC = 0.97, compared to 0.82-0.88 for other methods), both when detecting well-characterized species as well as newly sequenced, "taxonomically novel" organisms (Minot *et al*., 2015).

Besides the facility of use and also speed of analysis of both CosmosID and OneCodex, MG-RAST include data analysis options not available for the other software. Besides in this study MG-RAST was able to detect *Saccharomyces cerevisiae* although the DNA virus was neither detected nor quantified. Using MG-RAST the RefSeq provided the best results. The NCBI's Reference Sequence (RefSeq) collection is a freely accessible database of naturally occurring DNA, RNA, and protein sequences. It is a unique resource because it provides a large, multi-species, curated sequence database representing separate but explicitly linked records from genomes to transcripts and translation products (Pruitt *et al*., 2012). Unlike the sequence redundancy found in the public sequence repositories, the RefSeq collection aims to provide, for each included species, a complete set of non-redundant, extensively cross-linked, and richly annotated nucleic acid and protein records (Pruitt *et al*., 2012).

Even though current computational analysis strategies for metagenomic data rely largely on comparisons to reference genomes, they represent only a fraction of what we know and therefore limit our ability to segregate metagenomic data into coherent biological entities and fail to describe previously unknown species, phages and modules of genetic variation within microbial species (Nielsen *et al*., 2014). A possible alternative is the *de novo* assembly (*i.e.,* assembly without a reference) of genomes from complex metagenomic data, although it is inherently difficult due to many sequence ambiguities that confuse the assembly process. Hence, a typical metagenomic assembly will result in a large set of independent contigs that are not easily aggregated into biological entities.

Yang *et al*., 2016 acknowledge that given appropriate sequencing depth, shotgun metagenomics has great utility for investigating the ecology of foodborne pathogens. Nevertheless, it cannot currently be used for identification and quantification of pathogens for regulatory purposes due to

Table 3 Abundance values (%) obtained for the microorganisms of the mock community by MG-RAST, MGmapper, CosmosID and OneCodex.

Species	MG-RAST RefSeq	MGMapper Silva	CosmosID GenBook	OneCodex
P. freudenreichii	23.08	4.61	45.65	63.34
S. aureus	10.13	0.46	20.01	6.51
B. fragilis	9.41	1.21	18.26	8.51
F. nucleatum	1.62	0.11	6.59	2.29
E. coli	4.79	1.19	0.38	7.80
S. Typhimurium	8.72	0.90	9.73	7.15
C. parvum	0.15	0.01	88.74	0.08
S. cerevisiae	0.01	<0.01	ND	ND
B. alphaherpesvirus	ND	<0.01	7.14	1.43

limitations of the available technology and the incompleteness of bacterial genome databases. Specifically, the misclassification, that is inherent to the read length, the inability to get deep coverage of the pathogenic organisms in the sample due to the existence of other prokaryote and eukaryote DNA within the sample, and the impossibility of obtaining a comprehensive database containing all possible pathogenic organisms of interest invalidates the use of this approach for regulatory purposes.

Conclusions

All in all, our results demonstrate that MG-RAST with the database RefSeq, OneCodex and CosmosID can be used as data analysis tools to detect microorganisms belonging to different domains experimentally spiked in smoked salmon analysed by shotgun metagenomics sequencing. Nevertheless, a direct correlation between cell concentration of each spiked microorganism and number of corresponding reads is still not possible, although bacteria were identified with higher abundances than *C. parvum, S. cerevisiae* and *Bovine alphaherpesvirus*.

References

Aw TG, Wengert S, Rose JB, 2016. Metagenomic analysis of viruses associated with field-grown and retail lettuce identifies human and animal viruses. Int J Food Microbiol 223:50-56.

Forbes JD, Knox NC, Ronholm J, Pagotto F, Reimer A, 2017. Metagenomics: the next culture-independent game changer. Front Microbiol 8:1069.

Jagadeesan B, Gerner-Smidt P, Allard MW, Leuillet S, Winkler A, Xiao Y, Chaffron S, Van Der Vossen J, Tang S, Katase M, McClure P, Kimura B, Ching Chai L, Chapman J, Grant K, 2018. The use of next generation sequencing for improving food safety: translation into practice. Food Microbiol 79:96-115.

Jones MB, Highlander SK, Anderson EL, Li W, Dayrit M, Klitgord N, Fabani MM, Seguritan V, Green J, Pride DT, Yooseph S, Biggs W, Nelson KE, Venter JC, 2015. Library preparation methodology can influence genomic and functional predictions in human microbiome research. Proc Natl Acad Sci US 112:14024-9.

Keegan KP, Glass EM, Meyer F, 2016. MG-RAST, a metagenomics service for analysis of microbial community structure and function. Microb Environ Gen 1399:207-33.

Knudsen BE, Bergmark L, Munk P, Lukjancenko O, Priemé A, Aarestrup FM, Pamp SJ, 2016. Impact of sample type and DNA isolation procedure on genomic inference of microbiome composition. mSystems 1:00095-16.

Leonard SR, Mammel MK, Lacher DW, Elkins CA, 2015. Application of Metagenomic Sequencing to Food Safety: Detection of Shiga Toxin-Producing Escherichia coli on Fresh Bagged Spinach. Appl Environ Microbiol 8123:8183-91.

Leonard SR, Mammel MK, Lacher, DW, Elkins CA, 2016. Strain-level discrimination of Shiga toxin-producing Escherichia coli in spinach using metagenomic sequencing. PloS One 11:0167870.

Lewandowska DW, Zagordi O, Geissberger FD, Kufner V, Schmutz S, Böni J, Metzner KJ, Trkola A, Huber M, 2017. Optimization and validation of sample preparation for metagenomic sequencing of viruses in clinical samples. Microb 5:94-10.

Lindgreen S, Adair KL, Gardner PP, 2015. An evaluation of the accuracy and speed of metagenome analysis tools. Sci Rep 6:33-92.

Minot SS, Krumm N, Greenfield NB, 2015. One codex: A sensitive and accurate data platform for genomic microbial identification. BioRxiv 027607.

Ni J, Yan Q, Yu Y, 2013. How much metagenomic sequencing is enough to achieve a given goal? Sci Rep 3:1968.

Nielsen HB, Almeida M, Juncker AS, Rasmussen S, Li J, Sunagawa S, *et al*., 2014. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat Biotech 32:822-8.

Parks DH, Tyson GW, Hugenholtz P, Beiko RG, 2014. STAMP: statistical analysis of taxonomic and functional profiles. Bioinf 30:3123-4

Petersen TN, Lukjancenko O, Thomsen MCF, Sperotto MM, Lund O, Aarestrup FM, Sicheritz-Pontén T, 2017. MGmapper: reference-based mapping and taxonomy annotation of metagenomics sequence reads. PLoS One 12:0176469.

Pruitt K, Brown G, Tatusova T, *et al*. The Reference Sequence (RefSeq) Database. 2002 Oct 9 [Updated 2012 Apr 6]. In: McEntyre J, Ostell J, editors. The NCBI Handbook [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2002. Chapter 18. Available from: https://www.ncbi.nlm.nih.gov/books/NBK21091/

Pruitt KD, Tatusova T, Maglott DR, 2005. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucl Acid Res 33:501-5.

Pruitt KD, Tatusova T, Brown GR, Maglott DR, 2012.NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucl Acid Res 40:130-5.

Tremblay J, Singh K, Fern A, Kirton E S, He S, Woyke T, Lee J, Chen F, Dangl JL, Tringe SG, 2015. Primer and platform effects on 16S rRNA tag sequencing. Front Microbiol 6:771.

Wylezich C, Papa A, Beer M, Höper D, 2018. A versatile sample processing workflow for diagnostic metagenomics. Sci Rep 8:13108.

Yan Q, Wi YM, Thoendel MJ, Raval YS, Greenwood-Quaintance KE, Abdel MP, Jeraldo PR, Chia N, Patel R, 2019. Evaluation of the CosmosID Bioinformatics Platform for Prosthetic Joint-Associated Sonicate Fluid Shotgun Metagenomic Data Analysis. J Clin Microbiol 57:01182-18.

Yang X, Noyes NR, Doster E, Martin JN, Linke LM, Magnuson RJ, Yang H, Geornaras I, Woerner DR, Jones KL, 2016. Use of metagenomic shotgun sequencing technology to detect foodborne pathogens within the microbiome of the beef production chain. Appl Environ Microbiol 82:2433-43.