Journal of Bacteriology and Parasitology

Orosz, J Bacteriol Parasitol 2017, 8:4
DOI: 10.4172/2155-9597.1000317

Short Communication OMICS International

On the Benefit of Publishing Uncurated Genome Assembly Data

Ferenc Orosz*

Institute of Enzymology, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Budapest, Hungary

***Corresponding author:** Ferenc Orosz, Institute of Enzymology, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Budapest, Hungary, E-mail: orosz.ferenc@ttk.mta.hu

Received date: July 03, 2017; **Accepted date:** August 30, 2017; **Published date:** September 04, 2017

Keywords: Genome sequencing; Parasitology; *Neisseria gonorrhoeae*; Ribosomal RNA

Short Communication

The uncurated genome assembly data often contains DNA contaminations, originated from exotic organisms, introduced during DNA extraction or sequencing. It happens sometimes that it is not removed when the sequence is deposited into public databases such as GenBank or European Nucleotide Archive. Consequently, database searches could lead mistaken results due to these impurities [1,2]. Human DNA is an everyday contamination, from the scientists who extract and sequence the samples [3]. Impurities of human origin and other laboratory contaminants such as *E. coli* and cloning vectors can be effectively eliminated using highly efficient computational filters applied to the draft sequences [4,5]. However, other contaminations, as discussed later in the paper, are more difficult to identify. By the spreading of next-generation sequencing this has become a common problem due to the vast amount of reads which are generally short and of low quality in these projects [6-8].

A further source of contamination can be the pathogens present in the samples used for sequencing. Substantial bacterial contamination is routinely found in existing human-derived clinical RNA-seq datasets that likely arises from environmental sources [9]. Insect and other arthropod sequences were identified when analysing plant transcriptomes [10]. Just the opposite happened when the pathogen genome was found to be contaminated by the host. This was the case e.g. when it was discovered that the genome of the bacteria, *Neisseria gonorrhoeae* included sequences of cow and sheep origin [1].

It was found by me [11] that apicortin, a characteristic protein of apicomplexan parasites but absent in more developed animals (Eumetazoa), was virtually found in an animal genome assembly from the northern bobwhite (*Colinus virginianus*). Thus I decided to systematically investigate this problem: sequences of the apicoplast, an apicomplexan organelle, were used as queries in BLASTN search against nucleotide sequences of various animal groups, searching for possible contaminations. I found that beside the draft genome of the bobwhite [12] that of a bat, *Myotis davidii* [13], contained at least 6 and 17 contigs, respectively, of apicoplast origin. This is a general method for fast identification of genomes contaminated by DNA of apicomplexan origin, which needs limited computation and practically does not give false positives, as any significant hit is a clear indication of contamination. Moreover, by comparing some contaminating sequences with sequences of known apicomplexan parasites I was able to construct phylogenetic trees which show the phylogenetic position of the tentative contaminating species. Although the number of the complete apicomplexan genomes is increasing continuously, there are still not enough to use apicoplast sequences for constructing trees. Thus I used two characteristic genes, often used for phylogenetic analysis and known in many cases, 18S ribosomal RNA and the internal transcribed spacer 1 (ITS-1). I suggested that a second member of the *Nephroisospora* genus exists, which similarly to the first member, *Nephroisospora eptesici*, is hosted by a bat, *Myotis davidii*, and proposed its tentative name as "*Nephroisospora myotisi*". Of course, the christening of the unknown species was not accepted by the strict rules of parasitology require the isolation and taxonomic description of the species.

However, this idea was picked up and developed significantly by Janus Borner and Thorsten Burmester [14]. They pointed out that "while previous approaches have mostly focused on the removal of contaminating sequences, the identification of parasite-derived contaminations may also enable the discovery of novel parasite taxa and shed light on previously unknown host-parasite associations" [14]. The high level of accuracy and sensitivity of next generation sequencing for quantifying genetic material across organismal boundaries gives tremendous potential for pathogen discovery. Previously, the PathSeq program [15] was developed to identify microorganisms by deep sequencing of human tissue, which first subtracts all reads derived from the human host. Of course, this method can be used only in the case of the high-quality genome data as the human genome is. Borner and Burmeister's new departure [14] can be applied in a much broader field.

In the case of wild beasts, it is not possible to avoid infection by parasites before sequencing. E.g., in the above mentioned cases, the kidney of the bat and the muscle of the bobwhite, respectively, contained the parasitic cysts. Unveiled contaminations of animal genomes cause misinterpretation of data; however, if known, parasite-originated sequences can provide useful information. Thus Borner and Burmester [14] suggested that parasite-derived "impurities" mean plentiful information that can help the discovery and identification of novel parasites. They argued "that uncurated assembly data should routinely be made available in addition to the final assemblies" [14]. They showed that sequences of apicomplexan origin were found in many animal transcriptomes and genomes, which indicates apicomplexan infection in the sequenced host. They extracted these sequences from the datasets by a novel bioinformatic pipeline (ContamFinder) and assigned to distinct taxa using phylogenetic methods. (The softwares can be freely downloaded from https://sourceforge.net/projects/contamfinder.) They analysed 920 datasets of which 51 was contaminated and they recognised more than twenty-thousand contigs derived from apicomplexan parasites. The contaminating species were members of various apicomplexan taxa of Haemosporida, Piroplasmida, Coccidia and Gregarinasina. A typical finding was that in the assembly of the superseded genome of *Gorilla gorilla gorilla* (western lowland gorilla) there were sequences that were more than 99.9% identical at the nucleotide level (!) to those of *Plasmodium falciparum*, including the full mitochondrial genome. For other, less investigated parasite species, where no or only a few

molecular data were known previously, these kinds of draft (uncurated) genomes may represent an abundant source of the gene repertoire of parasites.

These results have a significant importance for apicomplexan research. Sequencing of apicomplexans is rather biased to genus of medical or veterinary interest as, first of all, *Plasmodium*, then *Babesia*, *Eimeria*, Toxoplasma etc., while for Gregarinasina, which parasitizes only invertebrates, much less data are available. Analysis of contaminations renders possible the identification or even the discovery of new parasite taxa and enlightens the apicomplexan phylogeny. Moreover, their method can be generalized and also be applied to investigate contaminations by bacteria, viruses and other pathogens. I agree absolutely with their final conclusion that draft genome assembly data should also be made public.

References

1. Merchant S, Wood DE, Salzberg SL (2014) Unexpected cross-species contamination in genome sequencing projects. PeerJ 2: e675.
2. Tao Z, Sui X, Jun C, Culleton R, Fang Q, et al. (2015) Vector sequence contamination of the Plasmodium vivax sequence database in PlasmoDB and in silico correction of 26 parasite sequences. Parasit Vectors 8: 318.
3. Longo MS, O'Neill MJ, O'Neill RJ (2011) Abundant human DNA contamination identified in non-primate genome databases. PLoS ONE 6: e16410.
4. Schmieder R, Edwards R (2011) Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PLoS One 6: e17288.
5. Jun G, Flickinger M, Hetrick KN, Romm JM, Doheny KF, et al. (2012) Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. Am J Hum Genet 91: 839-848.
6. Naccache SN, Greninger AL, Lee D, Coffey LL, Phan T, et al. (2013) The perils of pathogen discovery: origin of a novel parvovirus-like hybrid genome traced to nucleic acid extraction spin columns. J Virol 87: 11966-11977.
7. Laurence M, Hatzis C, Brash DE (2014) Common contaminants in next-generation sequencing that hinder discovery of low-abundance microbes. PLoS One 9: e97876.
8. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, et al. (2014) Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol 12: 87.
9. Strong MJ, Xu G, Morici L, Splinter Bon-Durant S, Baddoo M, et al. (2014) Microbial contamination in next generation sequencing: implications for sequence-based analysis of clinical samples. PLoS Pathog 10: e1004437.
10. Zhu J, Wang G, Pelosi P (2016) Plant transcriptomes reveal hidden guests. Biochem Biophys Res Commun 474: 497-502.
11. Orosz F (2015) Two recently sequenced vertebrate genomes are contaminated with apicomplexan species of the Sarcocystidae family. Int J Parasitol 45: 871-878.
12. Halley YA, Dowd SE, Decker JE, Seabury PM, Bhattarai E, et al. (2014) A draft de novo genome assembly for the northern bobwhite (Colinus virginianus) reveals evidence for a rapid decline in effective population size beginning in the late Pleistocene. PLoS One 9: e90240.
13. Zhang G, Cowled C, Shi Z, Huang Z, Bishop-Lilly KA, et al. (2013) Comparative analysis of bat genomes provides insight into the evolution of flight and immunity. Science 39: 456-460.
14. Borner J, Burmester T (2017) Parasite infection of public databases: a data mining approach to identify apicomplexan contaminations in animal genome and transcriptome assemblies. BMC Genomics 18: 100.
15. Kostic AD, Ojesina AI, Pedamallu CS, Jung J, Verhaak RGW, et al. (2011) PathSeq: software to identify or discover microbes by deep sequencing of human tissue. Nat Biotechnol 29: 393-396.