EMBARGOED until November 18th 2014: 7 am EDT; 12 midday UK time

PeerJ
Your Peers, Your Science
Academic Publishing is Evolving

Unexpected cross-species contamination in genome sequencing projects

As genome sequencing has gotten faster and cheaper, the pace of whole-genome sequencing has accelerated, dramatically increasing the number of genomes deposited in public archives. Although these genomes are a valuable resource, problems can arise when researchers misapply computational methods to assemble them, or accidentally introduce unnoticed contaminations during sequencing.

The first complete bacterial genome, *Haemophilus influenzae*, appeared in 1995, and today the public GenBank database contains over 27,000 prokaryotic and 1,600 eukaryotic genomes. The vast majority of these are draft genomes that contain gaps in their sequences, and researchers often use these draft sequences for future analyses.

Each genome sequencing project begins with a DNA source, which varies depending on the species. For animals, blood is a common source, while for smaller organisms such as insects the entire organism or a population of organisms may be required to yield enough DNA for sequencing. Throughout the process of DNA isolation and sequencing, contamination remains a possibility. Computational filters applied to the raw sequencing reads are usually effective at removing common laboratory contaminants such as *E. coli*, but other contaminants may be more difficult to identify.

In a new study in PeerJ (http://peerj.com), authors from Johns Hopkins University discovered contaminating bacterial and viral sequences in "draft" assemblies of animal and plant genomes that had been deposited in GenBank. These may cause particular problems for the rapidly growing field of microbiome analysis, when sequences labeled as animal in origin actually turn out to be microbial.

In an even more surprising finding, the authors discovered the presence of cow and sheep DNA in the supposedly finished genome of a pathogenic bacterium, *Neisseria gonorrhoeae*. Although deposited in GenBank as a finished genome, the bacterium apparently was a draft genome that was submitted as complete, with erroneous DNA inserted in five places. If taken at face value, this data would appear to be a startling case of lateral gene transfer, but the correct explanation appears to be more mundane.

These findings highlight the importance of careful screening of DNA sequence data both at the time of release and, in some cases, for many years after publication.

###

EMBARGOED until November 18th 2014: 7 am EST; 12 midday UK time (i.e. the date of publication)

PDF of this Press Release:
http://static.peerj.com/pressReleases/2014/PressReleasePeerJ_Merchant.pdf

Link to the Press Preview of the Original Article (this link should only be used BEFORE the embargo ends): http://static.peerj.com/press/previews/2014/11/675.pdf (note: this is an author proof and so may change slightly before publication)

Link to the Published Version of the article (quote this link in your story – the link will ONLY work **after** the embargo lifts): https://peerj.com/articles/675 - your readers will be able to **freely** access this article at this URL.

Citation to the article: Merchant, Wood and Salzberg (2014), Unexpected cross-species contamination in genome sequencing projects. PeerJ 2:e675; DOI 10.7717/peerj.675

###

About PeerJ

PeerJ is an Open Access publisher of peer reviewed articles, which offers researchers a lifetime publication plan, for a single low price, providing them with the ability to openly publish all future articles for free. PeerJ is based in San Francisco, CA and London, UK and can be accessed at https://peerj.com/. PeerJ's mission is to help the world efficiently publish its knowledge.

All works published in PeerJ are Open Access and published using a Creative Commons license (CC-BY 4.0). Everything is immediately available—to read, download, redistribute, include in databases and otherwise use—without cost to anyone, anywhere, subject only to the condition that the original authors and source are properly attributed.

PeerJ has an Editorial Board of almost 900 respected academics, including 5 Nobel Laureates. PeerJ was the recipient of the 2013 ALPSP Award for Publishing Innovation.

PeerJ Media Resources (including logos) can be found at: https://peerj.com/about/press/

###

Media Contacts

Note: If you would like to join the PeerJ Press Release list, visit: http://bit.ly/PressList

For the authors: Steven Salzberg, Johns Hopkins University School of Medicine. Phone: 410-614-6112
Email: salzberg@jhu.edu

For *PeerJ:* email: press@peerj.com , https://peerj.com/about/press/

###

Abstract (from the article):

The raw data from a genome sequencing project sometimes contains DNA from contaminating organisms, which may be introduced during sample collection or sequence preparation. In some instances, these contaminants remain in the sequence even after assembly and deposition of the genome into public databases. As a result, searches of these databases may yield erroneous and confusing results. We used efficient microbiome analysis software to scan the draft assembly of domestic cow, *Bos taurus*, and identify 173 small contigs that appeared to derive from microbial contaminants. In the course of verifying these findings, we discovered that one genome, *Neisseria gonorrhoeae* TCDC-NG08107, although putatively a complete genome, contained multiple sequences that actually derived from the cow and sheep genomes. Our findings illustrate the need to carefully validate findings of anomalous DNA that rely on comparisons to either draft or finished genomes.