Closed and High-Quality Bacterial Genome Sequences of the Oligo-Mouse-Microbiota Community

Quentin Lamy-Besnier,a,b,c Romain Koszul,b Laurent Debarbieux,a Martial Marboutyb

aBacteriophage, Bacterium, Host Laboratory, Department of Microbiology, Institut Pasteur, Paris, France
bUnité Régulation Spatiale des Génomes, Institut Pasteur, CNRS, UMR 3525, Paris, France
cUniversité Paris Descartes, Paris, France

ABSTRACT The Oligo-Mouse-Microbiota (OMM¹²) gnotobiotic murine model is an increasingly popular model in microbiota studies. However, following Illumina and PacBio sequencing, the genomes of the 12 strains could not be closed. Here, we used genomic chromosome conformation capture (Hi-C) data to reorganize, close, and improve the quality of these 12 genomes.

The Oligo-Mouse-Microbiota (OMM¹²) is a murine bacterial synthetic community (12 strains) introduced in axenic mice. The resulting gnotobiotic murine model has been increasingly used for diverse gut microbiota studies across several animal facilities (1–4). This bacterial consortium comprises members from the 5 major phyla naturally present in mice microbiota and reconstitutes its main functionalities, such as metabolism and colonization resistance against pathogens (5).

Despite two rounds of sequencing using PacBio and Illumina technologies (6, 7), one-half of the genomes currently accessible remain made of 2 to 20 contigs (Table 1). In particular, the genome of Bacteroides caecimuris, the most abundant bacterium of the OMM¹² consortium, consists of 19 contigs (5). Improving scaffolding would facilitate a number of downstream analyses, such as prophage prediction.

We used genomic chromosome conformation capture (Hi-C), a technique that uses chemical fixation (formaldehyde) to in vivo cross-link nearby DNA sequences, digestion with restriction enzymes, DNA extremities refilling with biotinylated nucleotides, and finally, proximity ligation, DNA extraction, and deep sequencing. The relative ligation frequencies between nonadjacent DNA segments reflect the average three-dimensional (3D) organization of the genome of interest (8) and also, because of the polymer nature of DNA, the relative distance separating them along the chromosome. The latter property has therefore been exploited to bridge scaffolding gaps resulting, for instance, from the presence of repeated elements (9). Hi-C scaffolding is now commonly used in sequencing projects of large eukaryotic genomes but applies to bacteria as well (10, 11).

A Hi-C metagenomic protocol for mammalian gut samples developed in our laboratory (12) was applied to fecal pellets from the OMM¹² mice bred at Institut Pasteur (protocol 20.173 approved by the veterinary staff and under authorization APAFIS 26874 by the national ethics committee). Libraries were prepared using streptavidin beads to capture biotinylated ligation junction as described previously (13) and sequenced with an Illumina NextSeq 550 system to generate a total of 101,182,905 reads of 35 bp. The quality of the reads was first verified with FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc). Then, for each bacterium, the python library hicstuff (https://github.com/koszullab/hicstuff) was used to generate Hi-C contact maps using the most recent OMM¹² genomes (6). Contigs were reorganized (i.e., scaffolded) based on their relative contact frequencies (14). When necessary, preexisting contigs were split and rearranged to better fit the expected 3D structure, based on the contact data. In order for users to detect junctions.
between contigs, contigs were separated by 10 × N in the final fasta file. We successfully closed the genomes of the 12 strains except for Flavonifractor plautii, where the position of a 70,760-bp region could not be assigned into the main scaffold. The genome of F. plautii therefore remains split in 2 scaffolds. An automatic annotation was then performed using RAST (15). The access to a single scaffold of high quality will greatly improve the accuracy of genomic analyses performed by the community of users of the OMM12 gnotobiotic mice.

Data availability. The 12 reassembled and closed genome sequences of the OMM12 bacteria as well as the FastQ reads have been deposited under the accession numbers provided in Table 1 under the BioProject number PRJNA680355.

ACKNOWLEDGMENTS

We thank Cyril Matthey Doret and Agnès Thierry for assistance during computational and experimental work, respectively.

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication. This research was supported by funding to L.D. and M.M. from ANR-20-CE92-0048 and to R.K. from the European Research Council under the Horizon 2020 Program (ERC grant agreement 771813) and from JPI-EC-AMR STARS ANR-16-JPEC-0003-05. Q.L.-B. is funded by Ecole Doctorale FIRE-Program Bettencourt.

REFERENCES


<table>
<thead>
<tr>
<th>Strain</th>
<th>Genome size (bp)</th>
<th>Previous no. of contigsa</th>
<th>No. of genes</th>
<th>DSM no.</th>
<th>Accession no.</th>
<th>Reads (%)a</th>
</tr>
</thead>
<tbody>
<tr>
<td>Akkermansia muciniphila YL44</td>
<td>2,737,357</td>
<td>1</td>
<td>2,717</td>
<td>26127</td>
<td>CP065322</td>
<td>2.8</td>
</tr>
<tr>
<td>Acutalibacter muris KB18</td>
<td>3,802,913</td>
<td>5</td>
<td>4,271</td>
<td>26090</td>
<td>CP065321</td>
<td>0.002</td>
</tr>
<tr>
<td>Bifidobacterium animalis YL2</td>
<td>2,021,936</td>
<td>2</td>
<td>1,740</td>
<td>26074</td>
<td>CP065311</td>
<td>0.0007</td>
</tr>
<tr>
<td>Bacteroides cacicimus I48</td>
<td>4,800,606</td>
<td>19</td>
<td>4,322</td>
<td>26085</td>
<td>CP065319</td>
<td>55.8</td>
</tr>
<tr>
<td>B lactia coelicoides YLS8</td>
<td>5,128,582</td>
<td>1</td>
<td>5,351</td>
<td>26115</td>
<td>CP065312</td>
<td>20.5</td>
</tr>
<tr>
<td>Enterocloster clostidiiformis YL32</td>
<td>7,157,610</td>
<td>16</td>
<td>7,955</td>
<td>26114</td>
<td>CP065314</td>
<td>4.9</td>
</tr>
<tr>
<td>Clostridium innocuum I46</td>
<td>4,452,146</td>
<td>1</td>
<td>4,966</td>
<td>26113</td>
<td>CP065320</td>
<td>5.6</td>
</tr>
<tr>
<td>Enterococcus faecalis KB1</td>
<td>3,025,655</td>
<td>1</td>
<td>2,919</td>
<td>32036</td>
<td>CP065317</td>
<td>0.1</td>
</tr>
<tr>
<td>Flavonifractor plautii YL31</td>
<td>3,813,715</td>
<td>5</td>
<td>4,049</td>
<td>26117</td>
<td>CP065315</td>
<td>0.38</td>
</tr>
<tr>
<td>Limosilactobacillus reuteri I49</td>
<td>2,063,624</td>
<td>3</td>
<td>1,958</td>
<td>32035</td>
<td>CP065318</td>
<td>0.99</td>
</tr>
<tr>
<td>Mucilaginibacter intestinalis YL27</td>
<td>3,307,069</td>
<td>1</td>
<td>2,957</td>
<td>28989</td>
<td>CP065316</td>
<td>4.5</td>
</tr>
<tr>
<td>Turicimonas muris YL45</td>
<td>2,887,949</td>
<td>20</td>
<td>2,753</td>
<td>26109</td>
<td>CP065313</td>
<td>0.089</td>
</tr>
</tbody>
</table>

a Raw data can be accessed at SRR9907181.

a Garzetti et al. (8).

a Unmapped reads = 4.34 %.

TABLE 1 Description and accession numbers of the OMM12 genomesa


