Supplementary Material

Data

Characterization of genome repeat content

To characterize the repeat content of the three genomes selected for this benchmarking we used the “repeat-match” tool from the MUMmer package [1, 2]. The minimum length threshold for repeat detection was set to 22 bp and the distributions of detected repeat lengths are shown in Supplementary Figure 1.

Preprocessing of real data

All datasets were preprocessed to discard reads arisen from plasmidic DNA or from contaminant organisms. The original reads were mapped on the respective reference sequence (GenBank: AE008692.2, CP001582.1 and U00096.2 for Z. mobilis, H. pylori and E. coli, respectively) using BLAST [3]: only reads aligned with a minimum identity percentage of 95% were selected as input data for the assemblers. We did not require a minimum alignment length to avoid the loss of short paired end sequences.

When necessary, reads and quality scores were extracted from the SFF files using the “sffinfo” tool from the GS FLX data analysis software package.

Sequencing error rate

For the three real datasets we determined the amount of sequencing errors due to the contribution of both nucleotides misincorporations, i.e., mismatches with the reference sequence, and wrong homopolymers length estimate, i.e., gaps (Supplementary Table 1).

Supplementary Table 1: Pyrosequencing error rate for the three datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Mismatch per 100 nt</th>
<th>Gaps per 100 nt</th>
<th>Total error per 100 nt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z. mobilis</td>
<td>0.09</td>
<td>0.24</td>
<td>0.33</td>
</tr>
<tr>
<td>H. pylori</td>
<td>0.05</td>
<td>0.49</td>
<td>0.54</td>
</tr>
<tr>
<td>E. coli</td>
<td>0.06</td>
<td>0.53</td>
<td>0.59</td>
</tr>
</tbody>
</table>

MetaSim simulations

The MetaSim [4] algorithm consists of two main steps: firstly, reads are correctly generated from the input genome and, secondly, they are corrupted with sequencing errors, on the basis of a selectable “error model”. We have created the benchmark datasets using the Z. mobilis, H. pylori and E. coli reference sequences. As requested by MetaSim, we have provided the parameters similar to the features of the real data: number of reads, mean length, percentage of shotgun and paired end reads; moreover, an estimation of mean and standard deviation of fragments sizes is required by this tool.

For the Z. mobilis genome, we performed one simulation using these settings: 76 000 reads, 270 bp expected read length, 35% mate-pair probability, 110 expected mate-pair length, mean of clone sizes 2500 bp and standard deviation 740.

For the H. pylori genome, we performed two simulations: a run of unpaired reads (0% mate-pair probability) and a paired end run. The former were carried out with these settings: 400 000 reads, 250 bp expected read length. For the latter we used the following settings: 75 000 reads, 250 bp expected read length, 40% mate-pair probability, 110 expected mate-pair length, mean of clone sizes 3000 bp and standard deviation 900.

Finally, the options for the E. coli unpaired reads simulation were set to: 110 000 reads, 370 bp expected read length (reads length is higher than the previous datasets because E. coli sequencing was performed using Titanium chemistry). Those for the paired end simulation were set to: 200 000 reads, 315 bp expected read length, 70% mate-pair probability, 135 expected mate-pair length, mean of clone sizes 8000 bp and standard deviation 1200.

MetaSim gives as output the corrupted reads plus their position in the reference; we used this latter in order to recover the “uncorrupted” version of the reads.
Supplementary Figure 1: Distribution of repeat lengths detected in *Z. mobilis* (a), *H. pylori* (b) and *E. coli* (c) genomes.
Characterization of simulated data

In Supplementary Tables 2 - 3 are reported the features of the simulated data with respect to real data, in terms of GC content and read coverage.

**Supplementary Table 2:** GC content of *Z. mobilis*, *H. pylori* and *E. coli* reference genomes and datasets.

<table>
<thead>
<tr>
<th></th>
<th>Reference</th>
<th>Real reads</th>
<th>Simulated reads</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Z. mobilis</em></td>
<td>46.33%</td>
<td>45.64%</td>
<td>46.30%</td>
</tr>
<tr>
<td><em>H. pylori</em></td>
<td>38.97%</td>
<td>39.14%</td>
<td>39.06%</td>
</tr>
<tr>
<td><em>E. coli</em></td>
<td>50.79%</td>
<td>49.13%</td>
<td>50.48%</td>
</tr>
</tbody>
</table>

**Supplementary Table 3:** Percentage of low-coverage (≤4x) bases in real and simulated datasets.

<table>
<thead>
<tr>
<th></th>
<th><em>Z. mobilis</em></th>
<th><em>H. pylori</em></th>
<th><em>E. coli</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Real data</td>
<td>11.857%</td>
<td>0.007%</td>
<td>0.047%</td>
</tr>
<tr>
<td>Simulated data</td>
<td>3.432%</td>
<td>0.002%</td>
<td>0.003%</td>
</tr>
</tbody>
</table>

Distribution of mate-pair distances

With the exception of Newbler, all the assemblers require some information on the distribution of mate-pair distances. Paired end distances were calculated by mapping mates with PASS [5] on the corresponding reference sequence. The distributions of mate-pair distances on the *Z. mobilis*, *H. pylori* and *E. coli* references are shown in Supplementary Figure 2. As the distributions of paired end distances are not Gaussian and characterized by outliers, we used their median value as expected length and $2 \cdot MAD$ (Median Absolute Deviation) as measure of dispersion. When minimum and maximum paired end distances were required by the assembler, we defined them as median $\pm 3 \cdot MAD$ (Supplementary Table 4).

**Supplementary Table 4:** Statistics on mate-pair distribution on *H. pylori*, *Z. mobilis* and *E. coli* references.

<table>
<thead>
<tr>
<th></th>
<th><em>Z. mobilis</em></th>
<th><em>H. pylori</em></th>
<th><em>E. coli</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Median</td>
<td>2529</td>
<td>2893</td>
<td>7891</td>
</tr>
<tr>
<td>MAD</td>
<td>469</td>
<td>457</td>
<td>797</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>938</td>
<td>914</td>
<td>1594</td>
</tr>
<tr>
<td>Min distance</td>
<td>1122</td>
<td>1522</td>
<td>5500</td>
</tr>
<tr>
<td>Max distance</td>
<td>3936</td>
<td>4264</td>
<td>19282</td>
</tr>
</tbody>
</table>

Assembly

**Details on assemblers algorithms**

The assemblers tested in this study belong to two different algorithm strategies: the Overlap-Layout-Consensus (OLC) and the De Bruijn graph (DBG) path finding. OLC assemblers use a strategy based on an overlap graph of reads obtained by an all-against-all pair-wise reads comparison. This graph is then used to construct a raw layout of reads belonging to groups that are further refined with multiple sequence alignment techniques [6]. DBG, on the contrary, works on a K-mer graph where the nodes are represented by fixed-length subsequences extracted from the reads. In this way a single node contains a subsequence that may be present in different reads [6, 7, 8]. In the OLC category we have assessed Newbler (2.3 and 2.5 versions), CABOG, a pre-release version of PCAP (PCAP454), and MIRA. In the DBG category we have evaluated CLC Assembly Cell.

An interesting aspect to take into consideration when choosing a particular assembler for a sequencing project regards the finishing phase. The visualization of aligned reads used to build contigs is what users need in order to check gaps, errors and weakly covered regions. Algorithms based on DBG strategy produce contigs that are generated from a path reconstruction of fragments rather than whole reads and this can lead to artifacts (i.e., reads origin is not necessarily tracked and “read-coherence” is not guaranteed). Consequently, a mapping tool is required to verify contigs consistency and visualize tiled reads.
Supplementary Figure 2: Distribution of mate-pairs on *Z. mobilis* (a), *H. pylori* (b) and *E. coli* (c) references.
Assembly of real data – technical notes

**CABOG** - The real data, available in SFF format, were converted to FRG format with the “sffToCA tool”, contained in the CABOG package. We specified the settings -clear 454 and -trim chop as suggested by the developers (http://sourceforge.net/apps/mediawiki/wgs-assembler/index.php?title=SffToCA).

**CLC** - For the preprocessing of paired end reads we used the “split _sequences” tool contained in the CLC Assembly Cell package.

**MIRA** - For the conversion of reads from SFF to FASTA format we used the “sff_extract” tool (http://bioinf.comav.upv.es/sff_extract/) distributed along with the MIRA 3 package. The assembly mode was set to: --job=denovo,genome,accurate,454.

**PCAP454** - The parameters of the beta version of PCAP454 were set following the instructions of the developer. In particular, the command-line for the paired-end preprocessing is:

```
./bcon.rep.454 file_name linker.fasta -e 30 -f 38 -o 18 -s 600 -p min_dist -q max_dist
```

In the following, the description of options:

- **e N** segment pair score cutoff \(N > 25\) (50)
- **f N** chain score cutoff \(N > 25\) (100)
- **o N** overlap length cutoff \(> 15\) (30)
- **p N** min read pair distance \(> 50\) (900)
- **q N** max read pair distance \(> 500\) (5500)
- **s N** overlap similarity score cutoff \(N > 100\) (1500)

For the assembly step, we did not change the default parameters setting. For the scaffolding step, we used the “makeScaf” tool provided by the developer.

Assembly of simulated data – technical notes

**CABOG** - The real data, available in FASTA format, were converted to FRG format with the “convert-fasta-to-v2.pl” script, also contained in the CABOG package.

**MIRA** - We did not use the “sff_tool” as we made with the real data, but we provided directly mate pairs sequences to the assembler, adding the option -notraceinfo.

**Newbler** - Paired end sequences were supplied as FASTA files of mates, following developer recommendations. In particular, left and right arms were separated from the linker and their headers were formatted in order to contain pairing information (we used custom-made Perl scripts, implemented following the instructions reported in the GS FLX System Software Manual, and the tool “fnafile”, present in 454 package).

Analysis

Distribution of errors and assembly gaps

Supplementary Figures 3 - 5 show the distribution of assembly errors and unreconstructed bases in relation to the structure of the three reference genome sequences.

The coverage distribution of the three datasets, shown by the black track in panel 1 of figures, was determined by aligning the reads to the respective reference sequences using BLAST. Coverage and genome positions are reported on the y and x axis, respectively. Peaks are caused by the alignment of reads in multiple positions and hence represent sparse and tandem repeats. For each read, all alignments with similarity \(\geq 95\%\) (98\% for simulated data) and length \(\geq 90\%\) of the read were considered and the number of reads covering each reference nucleotide was calculated. Finally, the average coverage was reported for every 250 nt. Tandem repeats are indicated as grey bars pointing downward, whose heights are proportional to the repeat sizes. The unreconstructed bases are computed considering the contigs aligned on the references using BLAT [9]. The BLAT output reports, for each contig, a list of blocks, defined as portions of contig-reference alignments containing no gaps. As we considered as reconstructed bases only the reference bases which are covered by these blocks,
Supplementary Figure 3: Representation of read coverage (thin black line in panel 1), tandem repeats (grey bars in panel 2), assembly gaps (black bars in panels 3-8) and insertion/deletion (INS/DEL) errors (grey dots in panels 3-8) of the real (a) and simulated (b) *Z. mobilis* datasets.
Supplementary Figure 4: Representation of read coverage (thin black line in panel 1), tandem repeats (grey bars in panel 2), assembly gaps (black bars in panels 3-8) and insertion/deletion (INS/DEL) errors (grey dots in panels 3-8) of the real (a) and simulated (b) *H. pylori* datasets.
Supplementary Figure 5: Representation of read coverage (thin black line in panel 1), tandem repeats (grey bars in panel 2), assembly gaps (black bars in panels 3-8) and insertion/deletion (INS/DEL) errors (grey dots in panels 3-8) of the real (a) and simulated (b) E. coli datasets.
the resulting unreconstructed bases represent both deletion errors and inter-contig gaps (even though only assembly gaps are actually noticeable because of graph resolution). Assembly gaps (black bars) are displayed for each assembler and their height is proportional to the length of the corresponding unreconstructed regions. In addition, grey circles highlight insertion/deletion errors (INS/DEL).

Supplementary Figures 3 - 5 show that the genomic regions characterized by tandem and sparse repeats correspond to assembly errors and gaps, especially for the *E. coli* and *H. pylori* datasets. The problematic repeats are few in *H. pylori* and, maybe leveraging on the high coverage, the assemblers are able to manage most of the critical situations, though with some exceptions and differences in terms of gaps lengths and number of errors (see Supplementary Figure 4 and Table 4 in the main text for the contig classification). On this dataset, Newbler 2.3 is more accurate than 2.5 version, as it produces more contigs but no errors. CABOG can handle these data effectively, giving only five contigs.

The *E. coli* genome is a hard test due to the great number of exact tandem and sparse repeats and it results in a higher number of assembly gaps and contig deletions (Supplementary Figure 5). The amount of misassembled regions decreases in simulated data (see also Table 4B in the main text). CABOG certainly deserves a special mention, as it has been able to correctly reconstruct difficult regions the other assemblers fail to solve. On the other hand Newbler 2.3 makes very few mistakes, confirming its stringency and precision in the assembly output. It is worth noting that, even in the worst scenario (see *E. coli* on Supplementary Figure 5), the behavior of the assemblers in certain problematic genomic regions is not homogeneous and somehow complementary.

The assemblies of real *Z. mobilis* dataset are characterized by a high number of gaps (Supplementary Figure 3), mainly because of the low coverage provided (10x). In this context, CLC shows a superior performance in terms of reconstructed reference genome if compared to the other assemblers. As reported in the main text (“Contig Analysis” section), the number of misassembled regions decreases in the simulated data, suggesting that repeats are not responsible for all of them and other factors contribute to hampering the assembly of real data (e.g., sequencing errors, uneven coverage and GC bias).

**Contig deletions at a glance**

To better clarify the relationship between genome structure and assembler errors, we focus on a region of *Z. mobilis* (bases 243 800–245 574) that has not been correctly reconstructed by any assembler starting from simulated reads. In particular, the corresponding contigs of MIRA, PCAP454 and CABOG are classified as DEL, while no one of Newbler and PCAP454 contigs spans this region.

In Supplementary Figure 6 are shown the alignments between the DEL contigs of MIRA, PCAP454 and CABOG and the reference through the Artemis Comparison Tool (ACT) [10]. The problematic genomic region is a tandem repeat; Supplementary Figure 6 highlights the repeat copies that are reconstructed in the correspondent contigs (red ribbons) and those that have been deleted in the assemblies (white spaces): CABOG deletion is the longest, as it reconstructs a lowest number of copies, while MIRA reconstructs the highest number of copies, obtaining the shortest deletion.

The results of the analysis performed using Tandem Repeat Finder [11] on this problematic region of the reference sequence and on the correspondent contigs are reported in Supplementary Table 5. The original sequence consists of 5.7 copies of a period of size 312 bp. MIRA fails to reconstruct one copy, PCAP454 two copies and CABOG three copies. CABOG reconstruction slightly differs from the reference in terms of nucleotide composition.

Finally, we focused on the CABOG contig to investigate the reasons of the deletion at read level, using Tablet [12] and Tandem Repeat Database [13]. Supplementary Figure 7A shows the reads tiled on CABOG contig (bases 101 872 – 101 940) and the final consensus, visualized through Tablet. The repeat copies of the reference sequence identified by the Tandem Repeat Database tools are reported in Supplementary Figure 7B; the sequences of the different repeat copies (red bars with highlighted mismatches) and their consensus (blue sequence on the top) are reported only for the portions corresponding to the bases of CABOG contig aligned above them. In the tiled reads mismatches with respect to the contig consensus are highlighted, showing that reads arisen from different repeat copies are wrongly assembled in a single sequence, probably because of the low coverage. Indeed, reads contain no sequencing errors and the mismatches shown represent the true polymorphisms that differentiate between the different repeat copies.
Supplementary Figure 6: Visualization of alignments between the Z. mobilis reference and MIRA (a), PCAP454 (b) and CABOG (c) contigs.
**Supplementary Table 5:** Statistics generated by Tandem Repeat Finder on the *Z. mobilis* reference (bases 243800–245574) and on the correspondent contigs. In Newbler 2.3, Newbler 2.5 and CLC assemblies there are no contigs spanning this region.

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Period size</th>
<th>Copy number</th>
<th>Consensus size</th>
<th>Percent matches</th>
<th>Percent indels</th>
<th>Score</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
<th>Entropy (0-2)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Z. mobilis</em></td>
<td>312</td>
<td>5.7</td>
<td>312</td>
<td>97</td>
<td>0</td>
<td>3145</td>
<td>23</td>
<td>24</td>
<td>32</td>
<td>19</td>
<td>1.97</td>
</tr>
<tr>
<td>Mira</td>
<td>312</td>
<td>4.7</td>
<td>312</td>
<td>96</td>
<td>0</td>
<td>2593</td>
<td>23</td>
<td>24</td>
<td>32</td>
<td>19</td>
<td>1.97</td>
</tr>
<tr>
<td>PCAP454</td>
<td>312</td>
<td>3.7</td>
<td>312</td>
<td>94</td>
<td>0</td>
<td>1489</td>
<td>23</td>
<td>23</td>
<td>32</td>
<td>19</td>
<td>1.98</td>
</tr>
<tr>
<td>CABOG</td>
<td>312</td>
<td>2.7</td>
<td>312</td>
<td>94</td>
<td>0</td>
<td>1489</td>
<td>23</td>
<td>23</td>
<td>32</td>
<td>19</td>
<td>1.98</td>
</tr>
<tr>
<td>Newbler 2.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><em>Not reconstructed</em></td>
</tr>
<tr>
<td>Newbler 2.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><em>Not reconstructed</em></td>
</tr>
<tr>
<td>CLC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><em>Not reconstructed</em></td>
</tr>
</tbody>
</table>

**Supplementary Figure 7:** Comparison between CABOG DEL contig (bases 101,872 – 101,940) (A) and the original copies of the *Z. mobilis* tandem repeat (B).
References


