# Enly: improving draft genomes through reads recycling – Supplementary Material

**Marco Fondi[1*], Valerio Orlandini[1], Giorgio Corti[2^], Marco Severgnini[2], Marco Galardini[1], Alessandro Pietrelli[2], Fabio Fuligni[2,§], Michele Iacono[2,†], Ermanno Rizzi[2], Gianluca De Bellis[2] and Renato Fani[1]**

[1] *Dept. of Evolutionary Biology, Via Madonna del Piano 6, 50143 Sesto Fiorentino,Florence, Italy;*

[2] *Istituto di Tecnologie Biomediche, Consiglio Nazionale delle Ricerche (ITB-CNR), Segrate (MI), Italy;*

[^] *Present address: Institute for Cancer Research and Treatment, Candiolo (TO), Italy*

[§] *Present Address: Unit of Hematopathology, Department of Hematology and Oncological Sciences "L. and A. Seràgnoli", S. Orsola Malpighi Hospital, University of Bologna.*

[†] *Present Address: Roche Diagnostics, Applied Science, Monza, Italy*

*Corresponding author

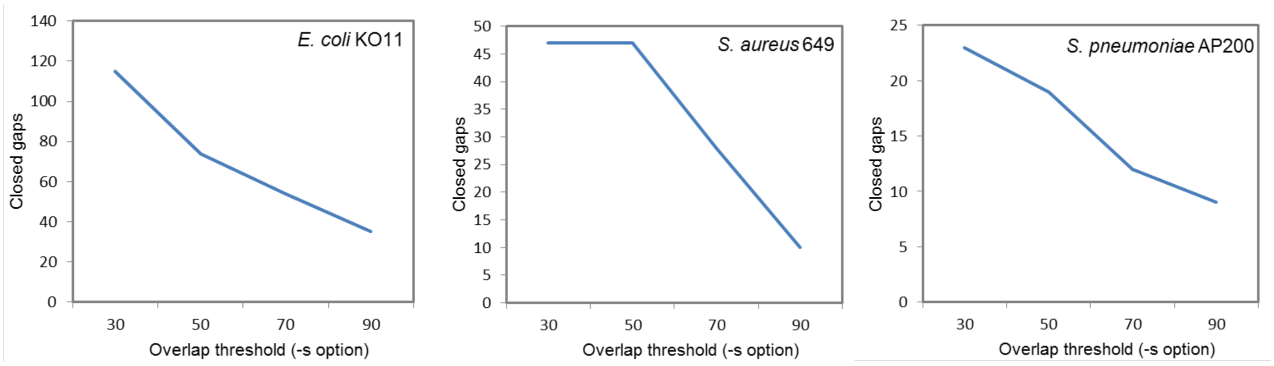
### 1) Parameters used for testing runs.

| Strain | -m | -b | -d | -c | -s | -r | -y | -a |
|---|---|---|---|---|---|---|---|---|
| *E. coli* KO11 | 300 | 400 | 50 | 15 | 30 | No | Yes | P |
| *S. aureus* 649 | 150 | 350 | 50 | 15 | 30 | No | Yes | P |
| *S. pneumoniae* AP200 | 150 | 350 | 50 | 15 | 30 | No | Yes | P |

**Supplementary Table 1:** list of all the parameter values used in the tests in Table 1.

### 2) Relationship between parameters and closed gaps

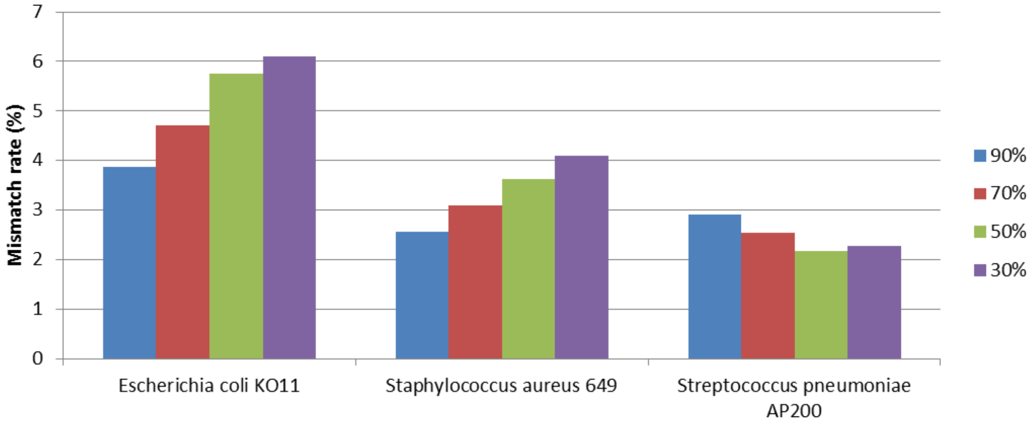
For each of the testing datasets, different values of overlap threshold (between reads and contigs) were chosen and the number of closed gaps recorded. In all the cases, Enly was run until no more bases were added to the draft genome. Concerning *-b*, *-m* and *–d* command line options, values ranging from *a*-100 to *a*+100 bp (where *a* is the average reads length of each reads dataset, see Table 1) were detached for each Enly's run. Intuitively, setting this overlapping threshold to higher values will increase accuracy during reads mapping, although reducing the number of bases added at each cycle. The use of higher thresholds is therefore recommended when a reference genome/scaffold is not available for checking possible chimeric joints. Contigs merging will still be reliable even in the case low overlap thresholds (e.g. 30%) are selected, although extended contigs extremities that are not merged into any (sub)scaffold cannot be checked for the presence of wrongly incorporated sequence.



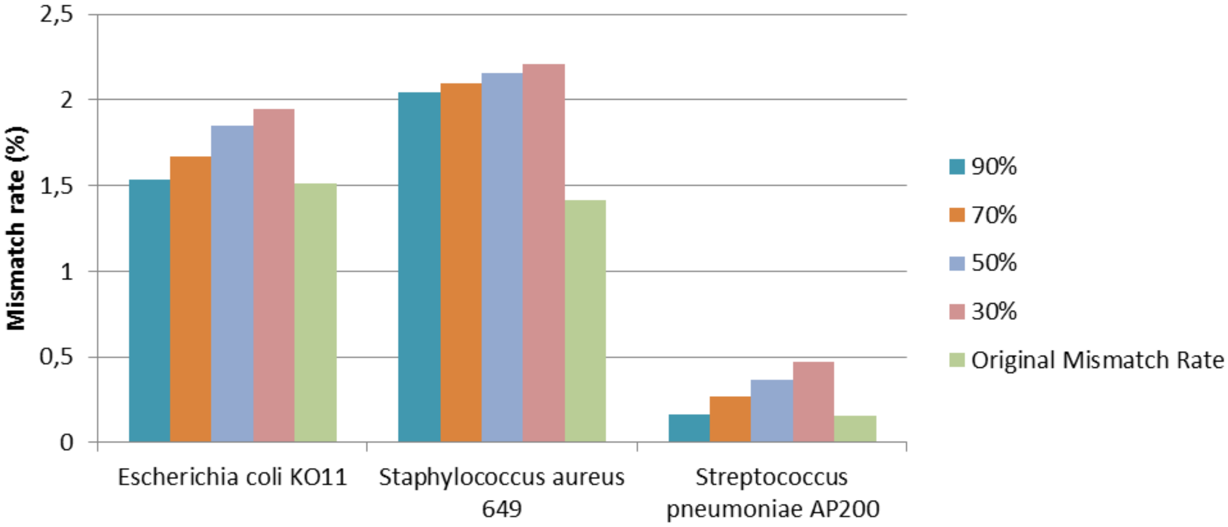
**Supplementary Figure 1 :** N. of closed gaps in relation to different overlap threshold (–*s*) values.

### 3) Relationship between parameters and mismatch rates

Mismatch rates were calculated on the original draft assemblies used as input for Enly and on the enlarged contigs, according to different values of the –*s* parameter). As expected, the mismatch rate of the enlarged part of the assemblies increases with the decrease of the overlap length threshold (-s parameter), ranging from around 6% to slightly more than 2% (Supplementary Figure 2). Anyway, the overall mismatch rate of the whole assembly considered is poorly affected by the slightly higher mismatch rate of the enlarged part (Supplementary Figure 3).



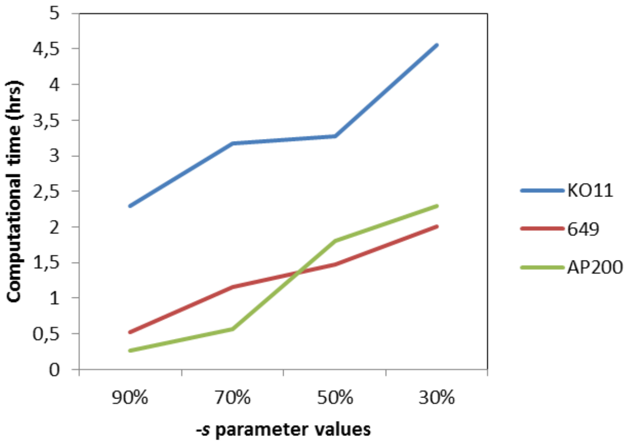
**Supplementary Figure 2:** Mismatch rate of the enlarged part of assemblies contigs at different overlap threshold values.



**Supplementary Figure 3:** Overall mismatch rate of enlarged assemblies in comparison with the mismatch rate of the original draft genomes at different overlap threshold values.

### 3) Relationship between parameters and computational time

Computational time required for the analysis of the test-datasets described in the text (Table 1) at different *-s* values (ranging from 30% to 90%).



**Supplementary Figure 4:** Required computational time for different values of *-s*.

### 4) Tests on reads from other sequencing technologies.

We tested our software on two draft genomes for which multi-platform reads were available, namely Escherichia coli O104 H4 (PMID: 22522955) and *Rhodobacter sphaeroides* 2.4.1 (PMID: 23718773). All tests were performed with *–s* parameter set to 50%. Reference scaffolds were obtained using the corresponding closed genomes as reference.

Tests on *Escherichia coli* O104 H4 were performed using the contigs of Newbler assembly (available at https://raw.github.com/nickloman/benchtop-sequencing-comparison/master/assemblies/newbler/reference/454AllContigs.fna) as input for the pipeline. Results obtained are shown in Supplementary Table 2.

| SRA code | Sequencing platform | N. reads | Av. reads length | N. contigs | Scaffolded contigs | Closed gaps (% in respect to original and scaffolded contigs) | N50 before/after Enly (% variation) |
|---|---|---|---|---|---|---|---|
| SRR388806/7 | 454 | 273520 | 461.633 | 152 | 78 | 14 (9.2% - 18%) | 124176/160769 (+29.5%) |
| SRA048511 | IonTorrent | 4638445 | 121.506 | 152 | 78 | 11 (7.2% - 14.1) | 124176/133494 (+7.5%) |
| SRA048664 | MiSeq | 1766516 | 141.723 | 152 | 78 | 10 (6.5% - 12.8%) | 124176/135621 (+9%) |
| Combined | - | 6678481 | - | 152 | 78 | 26 (17% - 33.3%) | 124176/178453 (+31%) |

**Supplementary Table 2:** Enly results on three datasets from *Escherichia coli* O104 H4 and obtained with three different platforms. "Scaffolded contigs" column refers to the number of contigs of the input draft genome that were scaffolded based on the comparison with the reference one. Indeed, since Enly joins contigs on the basis of their relative position in the input scaffold (i.e. they must be present in the scaffold structure file to be eventually merged), the percentage of closed gaps should be calculated also in respect to this number.

List of closed gaps according to the different datasets:

**IonTorrent**
>contig00041_contig00043
>contig00044_contig00045
>contig00048_contig00049
>contig00060_contig00059
>contig00017_contig00018
>contig00037_contig00038
>contig00064_contig00063
>contig00027_contig00025_contig00026
>contig00070_contig00069
>contig00071_contig00072

**454**
>contig00012_contig00011
>contig00017_contig00016
>contig00021_contig00020
>contig00045_contig00044
>contig00048_contig00049
>contig00054_contig00055
>contig00064_contig00065
>contig00066_contig00087
>contig00038_contig00037
>contig00073_contig00074
>contig00075_contig00076
>contig00060_contig00061_contig00059

**MiSeq**
>contig00017_contig00018
>contig00025_contig00024
>contig00027_contig00026
>contig00060_contig00061
>contig00075_contig00076
>contig00080_contig00079
>contig00081_contig00082
>contig00049_contig00048
>contig00057_contig00058
>contig00074_contig00073
>contig00052_contig00051

Tests on *Rhodobacter sphaeroides* 2.4.1 were performed using the assembly obtained from MiSeq reads (SRA code: SRR522246 - SRR520124 - SRR520123) as input for the pipeline. Using MiSeq on this assembly didn't allow closing any gap. Results obtained are shown in Supplementary Table 2.

| SRA code | Sequencing platform | N. reads | Av. reads length | N.contigs | Scaffolded contigs | Closed gaps (% in respect to original and scaffolded contigs) | N50 before/after Enly (% variation) |
|---|---|---|---|---|---|---|---|
| SRX202807 | IonTorrent | 3134559 | 166.064 | 1280 | 357 | 76 (6% - 21%) | 11087/13246 (+27%) |
| SRX109819/ 47/12/30 | PacBio | 85583 | 822.411 | 1280 | 357 | 6 (0.5% - 1.7%) | 11087/12990 (+15%) |
| Combined | - | 3220142 | 494.3 | 1280 | 357 | 80 (6.25% - 22.4%) | 11087/15376 (+28%) |

**Supplementary Table 3:** Enly's performances on three datasets from *Rhodobacter sphaeroides* 2.4.1 and obtained with two different platforms. "Scaffolded contigs" column refers to the number of contigs of the input draft genome that were scaffolded based on the comparison with the reference one. Indeed, since Enly joins contigs on the basis of their relative position in the input scaffold (i.e. they must be present in the scaffold structure file to be eventually merged), the percentage of closed gaps should be calculated also in respect to this number.

List of closed gaps according to the different datasets:

**PacBio**

>contig00188_contig00780
>contig00206_contig00660
>contig00271_contig00221
>contig00778_contig00735
>contig00855_contig00697
>contig00906_contig00226

**Ion Torrent**

>contig00006_contig00590
>contig00015_contig00266
>contig00022_contig00343
>contig00031_contig00466
>contig00040_contig00276
>contig00051_contig01035
>contig00061_contig01170
>contig00137_contig01244
>contig00139_contig00838
>contig00146_contig00978
>contig00171_contig00636
>contig00176_contig00107
>contig00188_contig00780
>contig00190_contig00227
>contig00221_contig00246
>contig00229_contig01209

```
>contig00236_contig00700
>contig00238_contig00952
>contig00279_contig00398
>contig00287_contig01164
>contig00299_contig00441
>contig00310_contig01017
>contig00338_contig01280
>contig00389_contig00899
>contig00418_contig00059
>contig00425_contig00596
>contig00464_contig00743
>contig00491_contig00524
>contig00503_contig00557
>contig00562_contig00885
>contig00634_contig00986
>contig00638_contig01110
>contig00639_contig01211
>contig00646_contig00851
>contig00670_contig00306
>contig00677_contig00334
>contig00706_contig00204
>contig00763_contig01177
>contig00770_contig00821
>contig00772_contig00445
>contig00776_contig00411
>contig00793_contig01153
>contig00795_contig00998
>contig00797_contig00918
>contig00815_contig01057
>contig00820_contig00972
>contig00895_contig00506
>contig00924_contig00840
>contig00934_contig00241
>contig01003_contig00078
>contig01010_contig00758
>contig01152_contig01224
>contig01222_contig00507
>contig00003_contig00220
>contig00032_contig00542
>contig00251_contig01184
>contig00285_contig00526
>contig00317_contig00848
>contig00529_contig00617_contig00527
>contig00637_contig00094
>contig00833_contig00029
>contig00907_contig00535
>contig00929_contig01123
>contig00970_contig00697_contig00855
>contig00982_contig00050
>contig01196_contig00626_contig00296
>contig00226_contig00906_contig00447
```

```
>contig00242_contig01119_contig01162
>contig01173_contig00569_contig00883_contig00175
```