Executive summary of B. brunii assembly round 1

Gents,

I’ll keep this email short, and if you’re interested in the nuts and bolts I am attaching several files, including a 5 page summary that goes into more depth.

The original draft B. braunii genome was rejected to a large extent because many of the contigs had contamination with DNA from prokaryotic symbionts. Also, the overwhelming majority of the contigs were very short, likely in part due to that contamination, if you assume that symbiont DNA from any 1 species would be present in very low quantities, resulting in a lot of low-coverage reads. As well, it was not obvious that the original assembly had been done using read quality trimming, adapter trimming, or error correction.

The attached results are using a subset of reads that survived a series of rigorous trimming and correction steps. While the original raw reads represented about 166 genome equivalents, the final pollux-corrected reads had only 28 genome equivalents. (As a rule of thumb, you can divide the G.E. by 5 to get the 99% coverage). While this level of coverage doesn’t bode well for a good assembly, I thought it was important to test assembly with the rigorous dataset as a baseline. Those results are found in Bbraunii\_coverage\_RM.xlsx.

As the file assemblystats.xlsx shows, by most measures, the original draft genome is better than the genomes from the smaller but rigourous datasets. What we have learned from these experiments is that in general, both Spades and SOAPdenovo2 show better assemblies with higher k-mer sizes. This trend is seen for #contigs > 5000, largest contig, total length and N50. Considering the low coverage of the rigorous dataset, the assemblies were surprisingly good, compared to the original.

**Strategy for Assembly round 2**

1. Re-do trimming with the following parameters for Trimmomatic:

* SLIDING WINDOW - required quality 10, averaged across 8 positions
* MINLEN = 50 nt

[Macmanes](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3908319/) has shown that a quality cutoff of 20 significantly decreases the number of high frequency k-mers from the read pool. Since high-frequency k-mers are needed for the read correction step, this in turn compromises the number of reads which survive read correction. Decreasing the quality cutoff at the trimmimg step should increase the number of reads which survive correction. As well, decreasing the MINLEN value to 50 from 60 should allow a substantial number of additional reads to survive. Loss of reads is particularly critical for paired end reads, because it only takes 1 bad read to turn a paried end read into a SE read. We should therefore have a much higher number of good PE reads coming through read correction.

2. Estimated insert sizes.

To verify insert sizes, corrected reads were aligned with contigs > 50 kb from the B. braunii draft genome, as shown in the tables above. I will redo these estimates with the round 1 assembly, to hopefully get a better estimate of insert sizes for PE reads. Since SOAPdenovo2 asks you to set an insert size parameter for each set of PE reads, better insert estimates might improve subsequent assemblies.

3. Running Spades with error correction - I have found in the past that Spades works better using its own read correction. For consistency with other programs, I ran Spades using the same Pollux-corrected reads. In round 2, I will try Spades with trimmed reads and error correction, and using Pollux corrected reads with the assembly only option for Spades.

4. The final potential improvement might be found by filtering out corrected reads that match prokaryotic sequences, based on Magicblast. I have been hesitant to do this for fear of biasing the dataset. Ideally, the better read is to filter out contigs with prokaryotic sequences from the final assembly, but we can try it both ways.