Iterative read mapping and assembly allows the use of a more distant reference in metagenome assembly

Bas E. Dutilh, Martijn A. Huynen, Jolein Gloerich and Marc Strous. CMBI/NCMLS and Nijmegen Proteomics Facility, Radboud University Nijmegen Medical Centre; Department of Microbiology, Radboud University Nijmegen; MPI for Marine Microbiology; Centre for Biotechnology, University of Bielefeld.

Introduction

We assemble a quasispecies consensus genome from 32nt metagenomic reads with 91.1% identity to the original reference. Briefly, we used a permissive (BlastN) and a strict (Maq) mapping algorithm and assembled a majority consensus [1]. As the initial assembly better represents the sequenced genomes than the reference genome does, we iterated the mapping and assembly several times.

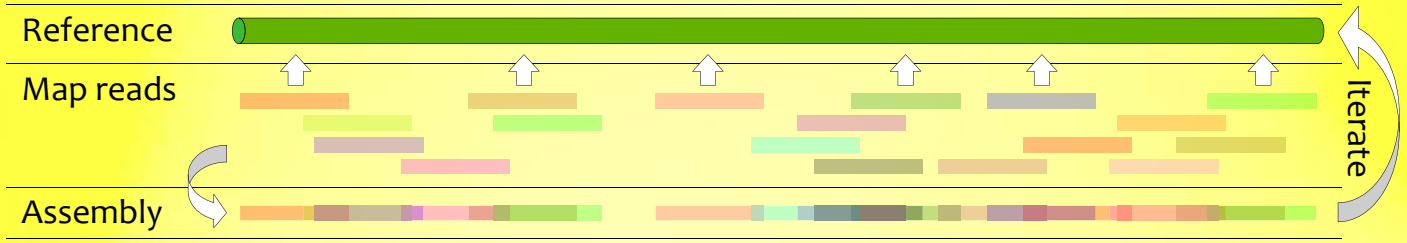

Figure 1. Outline of the iterative mapping and assembly approach.

Results

The number of mapped reads increases, the assembly converges, and the assembly diverges from the original reference with iterations.

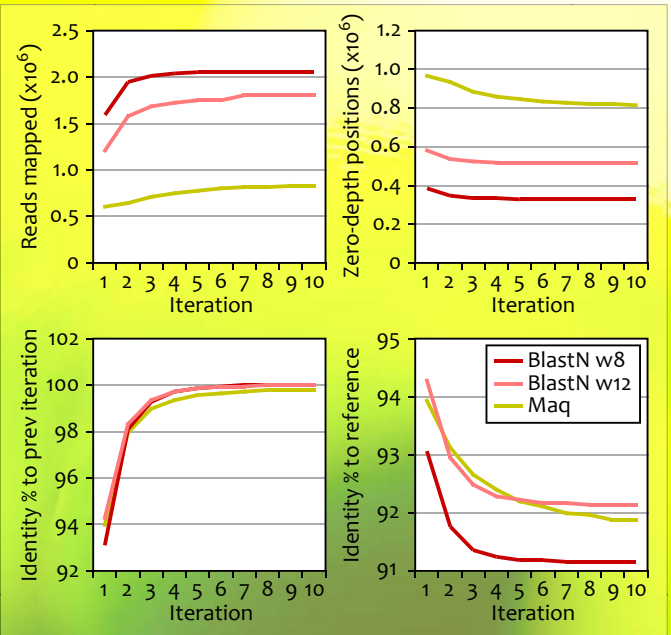

Figure 2. Statistics of the evolving assembly.

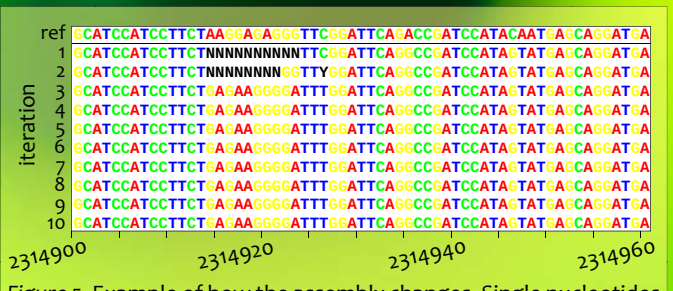

Figure 5. Example of how the assembly changes. Single nucleotides are settled and gaps are filled.

Positive control The assembly converged toward the consensus genome of the sequenced quasispecies: the metagenomic reads were mapped with lower e-value, and the translated ORFs mapped more metaproteomic peptides.

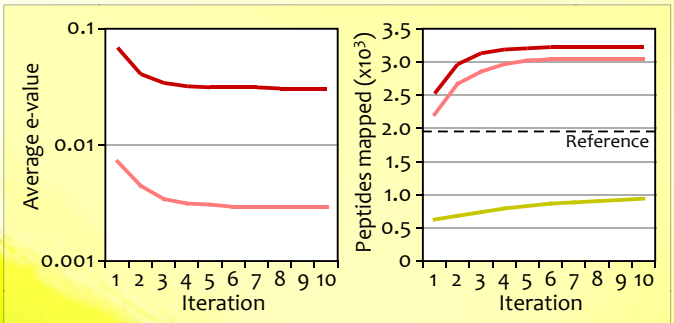

Figure 3. Quality measures: average read mapping quality (e-value; only for Blast), and number of metaproteomic peptides explained.

Negative control We expect only a small fraction of alien reads to be incorporated, depending on the mapping algorithm.

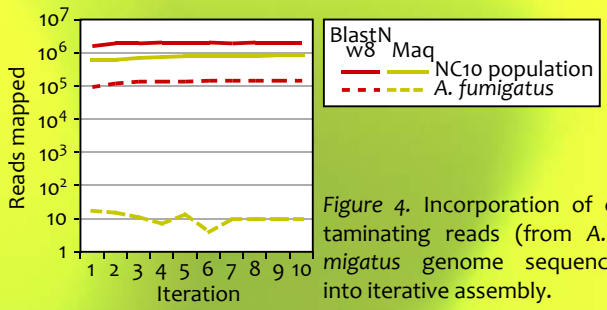

Figure 4. Incorporation of contaminating reads (from *A. fumigatus* genome sequencing) into iterative assembly.

Discussion

The assembly does not represent any existing genome but a consensus sequence that captures the diversity in the sequenced population [2]. It may be considered as our best estimate of the "wild type" genome(s).

[1] B.E. Dutilh, M.A. Huynen and M. Strous (2009), "Increasing the coverage of a metapopulation consensus genome by iterative read mapping and assembly", *Bioinformatics* **25**: 2878-2881.
[2] K.F. Ettwig*, M.K. Butler*, D. Le Paslier, E. Pelletier, S. Mangenot, M.M.M. Kuypers, F. Schreiber, B.E. Dutilh, J. Zedelius, D. de Beer, J. Gloerich, H.J.C.T. Wessels, T.A. van Alen, F. Luesken, M.L. Wu, K.T. van de Pas-Schoonen, H.J.M. Op den Camp, E.M. Janssen-Megens, K.-J. Francoijs, H. Stunnenberg, J. Weissenbach, M.S.M. Jetten and M. Strous (2010), "Nitrite-driven anaerobic methane oxidation by oxygenic bacteria", *Nature* **464**: 543-548. * Authors contributed equally.