Metagenomic proxy assemblies of single cell genomes

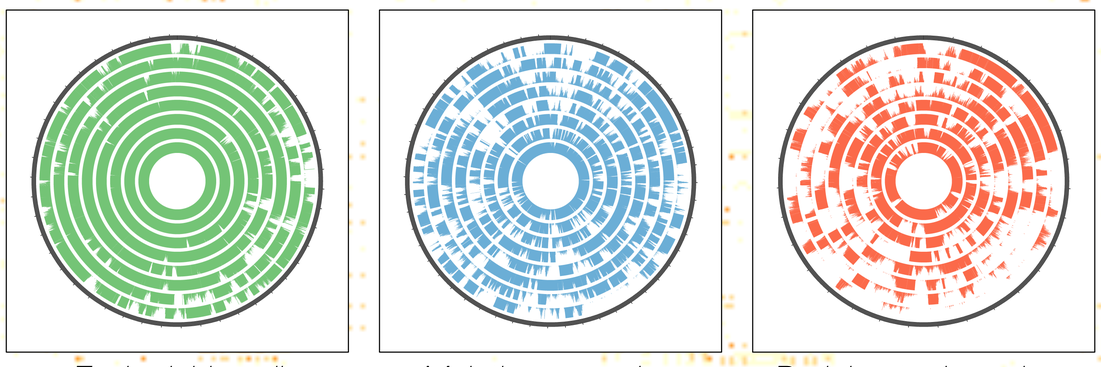
Andreas Bremges[1], Tanja Woyke[2], Alexander Sczyrba[1]
[1]Center for Biotechnology, Bielefeld University, 33615 Bielefeld, Germany
[2]U.S. Department of Energy Joint Genome Institute, Walnut Creek, CA 94598, USA

Studying the microbial dark matter

Over 99% of the microbial species observed in nature cannot be grown in pure culture, making it impossible to study them using classical genomic methods. Metagenomics and single cell genomics are two approaches to study the microbial dark matter.

Single cell genomes tend to be incomplete

Prior to sequencing of a single cell, its DNA needs to be amplified. This usually is done with with multiple displacement amplification (MDA), introducing a tremendous coverage bias. Shown below is the raw read coverage (capped at 10x) for 3 reference strains, 8 single cells per reference, sequenced to ~1,000x coverage each:



Escherichia coli
4.6 Mb, 51% GC

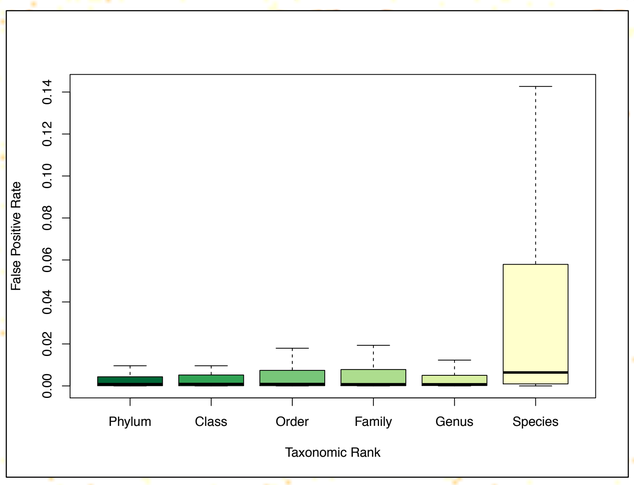
Meiothermus ruber
3.1 Mb, 63% GC

Pedobacter heparinus
5.2 Mb, 42% GC

Poorly amplified regions result in extremely low sequencing coverage or physical sequencing gaps. These parts of the genome cannot be reconstructed in the subsequent assembly step, and therefore genomic information is lost. The completeness of single cell genomes is estimated to range from less than 10% to more than 90% (mean 40%) [Rinke *et al.*, 2013].

Metagenomic "proxy" reads to the rescue

In a metagenome, each genome's coverage is constant and depends only on its abundance. We developed a fast, k-mer based recruitment method to sensitively identify metagenomic proxy reads representing the single cell of interest, using the raw single cell sequencing reads as recruitment seeds. The core of our algorithm is the applied q-gram lemma [Ukkonen, 1992], enabling us to process ~250,000 reads per second on a single core, using 16 GB of RAM.



We estimate the expected false positive rate by applying our recruitment criterion in an all-vs-all fashion on all 3,132 complete microbial genomes. For each taxonomic rank, we select the worst case of all pairwise combinations. The heat map in the background visualizes this on genus level.

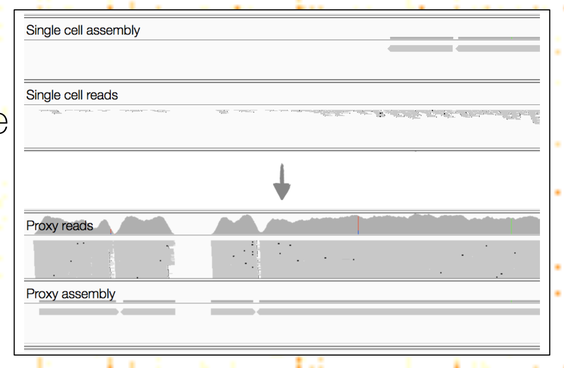
Overall, the false positive rate is neglectably for a k-mer length of 18. We use SPAdes [Bankevich *et al.*, 2012] for the assembly step, and thereafter BWA-MEM [Li, 2013] to map the original single cell reads on the proxy assembly, eliminating the few, and usually very short, contigs without any single cell read mapping to it.

Accurate and contiguous proxy assemblies

With sufficient metagenomic coverage, we can significantly improve the assembly contiguity. Working on the 117 single cells from the Microbial Dark Matter project [Rinke *et al.*, 2013] with a reference, obtained by pooling multiple single cells, and a corresponding metagenome available, 72 proxy assemblies show an increased NG50 value, at least doubling in 7 cases. In all cases, the average nucleotide identity (ANI) between single cell and proxy assemblies is higher than 99%, indicating comparable accuracy.

Case study: Aminacenantes (OP8) single cell

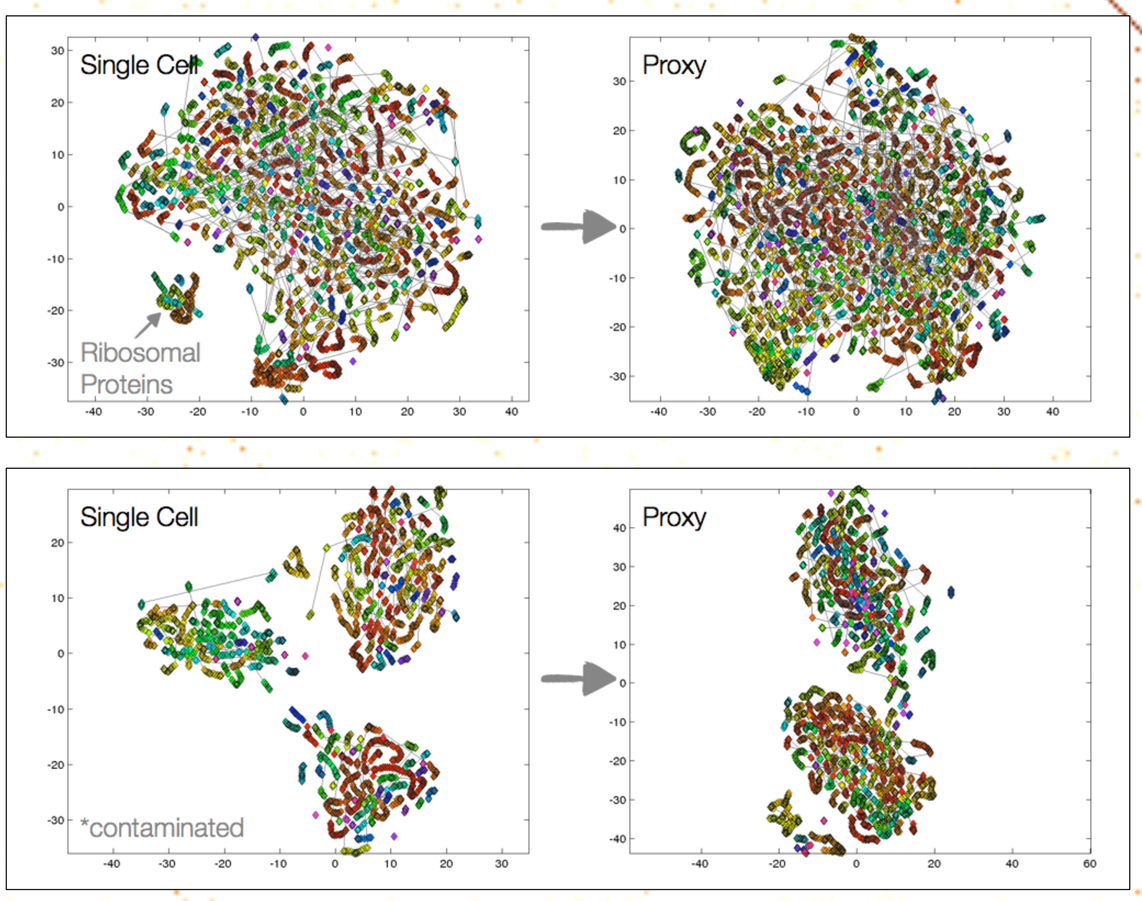
To emphasize the power of our approach, we take a close look at one Aminacenantes (OP8) single cell from [Rinke *et al.*, 2013]. The single cell and a corresponding metagenome were generated from a sample of brackish water in 120m depth in Sakinaw Lake, BC.



We applied our method to identify metagenomic proxy reads and assembled the original single cell reads (top) or their proxy reads (bottom, shown is a 10 Kb low coverage region). Around sparsely scattered single cell reads, that did not make it into the single cell assembly, islands of metagenomic proxy reads are recruited. These proxy reads can easily be assembled into longer contigs, eventually reconstructing previously lost genomic information. While the single cell assembly was estimated to be 53% complete, we estimate this proxy assembly to be 76% complete (marker gene analysis).

Careful contamination screening is needed

Working with single cells, contamination is always present. We apply t-SNE [van der Maaten & Hinton, 2008], a nonlinear dimensionality reduction technique, on the contigs' tetranucleotide frequencies (TNF) to detect and visualize contamination. Please ask me for details!



Acknowledgements and Contact

AB is supported by a fellowship from the CLIB Graduate Cluster Industrial Biotechnology.
Thanks to Barbara Hammer, Jessica Jarett, Markus Lux and Patrick Schwientek for data and discussions.
Correspondence should be addressed to abremges@cebitec.uni-bielefeld.de or @abremges