ANDREAS BREMGES

ASSEMBLING THE MICROBIAL DARK MATTER

BIELEFELD UNIVERSITY
FACULTY OF TECHNOLOGY

Assembling the microbial dark matter

THESIS BY ANDREAS BREMGES

The vast majority of microbial species found in nature has yet to be grown in pure culture, turning metagenomics and – more recently – single cell genomics into indispensable methods to study the microbial dark matter.

I DEVELOPED, APPLIED, AND BENCHMARKED genome assembly protocols for single cell and metagenome sequencing data to access microbial dark matter genomes.

IN THE FIRST PART of my thesis, I propose new algorithms that naturally exploit the complementary nature of single cells and metagenomes to improve the quality of single cell assemblies.

IN THE SECOND PART, I apply advanced metagenome assembly and binning techniques to untangle genomes from metagenomes, eventually reconstructing hundreds of near-complete genomes of process-relevant community members in the biogas microbiome.

Doctoral thesis, submitted to the
FACULTY OF TECHNOLOGY
BIELEFELD UNIVERSITY
for the degree of Dr. rer. nat.

Referees:
DR. ALEXANDER SCZYRBA
PROF. DR. JENS STOYE
PROF. DR. ALFRED PÜHLER

November 2016

Contents

Introduction

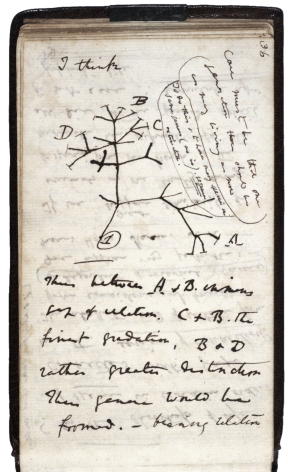


Figure 0.1: **Darwin's tree of life.** Above it, he scribbled "I think."

In 1837, Charles Darwin sketched a small evolutionary tree in his "B" notebook, *Transmutation of Species*, perfectly encapsulating his big idea that all species descend from a common ancestor (Figure 0.1).[1] Since then, generations of scientists have been adding nodes (and edges) to this tree of life.

[1] Darwin, 1837, 1859

DNA sequencing – in particular targeted sequencing of the small subunit ribosomal RNA gene – greatly expanded our view of the tree, currently incorporating three domains of life: Bacteria, Archaea, and Eukarya.[2] However, depictions of the tree of life have largely focused on eukaryotic diversity.[3]

[2] Woese and Fox, 1977; Lane et al., 1985; Woese et al., 1990; Yarza et al., 2014

[3] Hinchliff et al., 2015

A recent study presents a new view of the tree of life by also including 1,011 microorganisms from lineages for which genome sequences were previously unavailable.[4] Bacteria and – to a lesser extent – archaea occupy most of the tree; all eykaryotes are crowded together on one thin branch (Figure 0.2). 68 of 123 major lineages lack an isolated (cultured) representative, thus counting towards the *micobial dark matter*.[5]

[4] Hug et al., 2016

[5] Filée et al., 2005; Rinke et al., 2013

Metagenomics and single cell genomics are essential, culture-independent, and complementary methods to access the genetic makeup of microbial dark matter.[6] An estimated 85–99% of bacteria and archaea cannot be grown in pure culture yet, holding back the search for novel compounds of pharmaceutical or biotechnological relevance, such as new antibiotics or carbohydrate-active enzymes (CAZymes).[7]

[6] Brown et al., 2015; Rinke et al., 2013

[7] Lok, 2015

Antimicrobial resistance is a global threat to public health, but the pace of antibiotic discovery has slowed down.[8] Almost

[8] Lewis, 2013

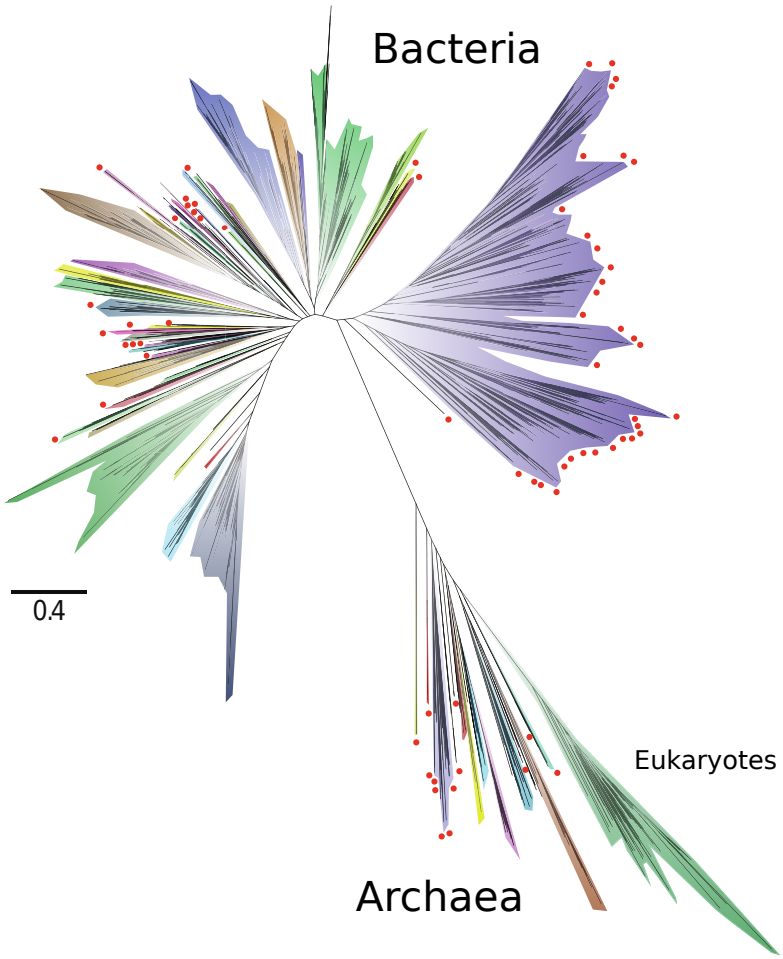


Figure 0.2: **A current view of the tree of life.** The tree – modified from Hug et al., 2016 – includes 92 bacterial and 26 archaeal phyla, and all five eukaryotic supergroups. Red dots highlight the 68 lineages lacking an isolated representative.

all of our antibiotics were sourced from marine or soil-derived actinomycetes, which represent only a fraction of the prokaryotic diversity.[9] Microbial dark matter is therefore perceived as an untapped resource of new antibiotics.[10]

Cellulose, a renewable resource for biofuel production, is notoriously difficult to deconstruct using currently available enzyme technology.[11] In nature, however, a variety of digestive ecosystems – such as the hindgut of higher termites or the cow rumen – are able to efficiently degrade plant biomass.[12]

[9] Demain and Sanchez, 2009; Gallagher et al., 2010; Manivasagan et al., 2014

[10] Wilson et al., 2014

[11] Klemm et al., 2005

[12] Warnecke et al., 2007; Hess et al., 2011

Culture-independent methods continue to expand the catalog of carbohydrate-active genes and genomes for biofuel production.[13]

[13] Morrison et al., 2009; Pope et al., 2010; Hess et al., 2011

Genome sequencing and assembly of culturable microbes turned from a challenge into a routine, primarily due to the advent of long-read sequencing.[14] Assembling microbial dark matter genomes from single cells or metagenomes, on the other hand, is an open question – addressed in my thesis.

[14] Wibberg et al., 2014, 2016; Goodwin et al., 2016

Thesis structure

PART I

First, I disclose limitations and the state of the art in single cell assembly. I benchmark three modern single cell assemblers on real data and perform regression testing of one (Chapter 1). Consequently, I propose new algorithms exploiting metagenome sequencing data to improve the quality of single cell assemblies: MeCorS[15] is a metagenome-enabled error correction method to accurately correct sequencing errors and chimeras in single cell sequencing reads (Chapter 2); kgrep[16] identifies metagenomic "proxy" reads to assemble instead of the original single cell reads and circumvents most challenges of single cell assembly (Chapter 3). Chapter 4 concludes this first part of my thesis.

[15] Bremges et al., 2016

[16] Bremges et al., in prep.

PART II

In the second part, I focus on biogas metagenomics. After briefly reviewing advanced metagenome assembly and binning techniques (Chapter 5), I present the first metagenome assembly of a biogas-producing microbial community from a production-scale biogas plant (Chapter 6).[17] In Chapter 7, I describe how deeper sequencing of more samples enabled a more inclusive assembly of the biogas microbiome. Successive binning of assembled contigs recovered hundreds of near-complete genomes of process-relevant community members.[18] Lastly, I gauge at the value of this genome catalog and advocate the integration of metatranscriptomic, -proteomic, and single cell data (Chapter 8).

[17] Bremges et al., 2015

[18] Stolze et al., 2016

EPILOGUE

I conclude my thesis by motivating the need for systematic benchmarking of methods in metagenomics, as implemented in the *Critical Assessment of Metagenome Interpretation* challenge.[19]

[19] http://cami-challenge.org

Reproducibility statement

Reproducibility is a main principle of the scientific method, yet analyses in psychology and cancer biology revealed that only 39% and 11%, respectively, of published work is reproducible.[20]

[20] Open Science Collaboration, 2015; Begley and Ellis, 2012

To FOSTER REPRODUCIBILITY, all analyses throughout my thesis are performed using free and open-source software (Table 0.1). All (sequencing) data are publicly available.

Software	Version	Reference
BayesHammer	3.6.0	Nikolenko et al., 2013
BEDTools	2.22.0	Quinlan and Hall, 2010
BLAST	2.2.29+	Altschul et al., 1990
BlastKOALA	2.1	Kanehisa et al., 2016b
Bowtie 2	2.2.4	Langmead and Salzberg, 2012
BWA-MEM	0.7.12	Li, 2013
CheckM	1.0.4	Parks et al., 2015
IDBA-UD	1.1.2	Peng et al., 2012
kgrep	0.7.0	Bremges et al., in prep.
Mash	1.1	Ondov et al., 2016
MeCorS	0.4.1	Bremges et al., 2016
MEGAHIT	1.0.5	Li et al., 2015
MetaBAT	0.23.1	Kang et al., 2015
MetaProdigal	2.6.0	Hyatt et al., 2012
Prokka	1.11	Seemann, 2014
QUAST	3.1	Gurevich et al., 2013
Ray Meta	2.3.1	Boisvert et al., 2012
SAMtools	1.1	Li et al., 2009
SPAdes	3.8.0	Bankevich et al., 2012
taxator-tk	1.2.1	Dröge et al., 2015
Trimmomatic	0.33	Bolger et al., 2014

Table 0.1: **Software.** All tool names, used versions, and their primary references.

Complete list of publications

14. Bremges, A., J. Jarett, T. Woyke, and A. Sczyrba. **Metagenomic proxy assemblies of single cell genomes.** *TBA*, in preparation. ★

13. Weimann, A., K. Mooren, J. Frank, P. B. Pope, A. Bremges, and A. C. McHardy. **From genomes to phenotypes: Traitar, the microbial trait analyzer.** *bioRxiv*, 2016. DOI: 10.1101/043315.

12. Maus, I., D. E. Koeck, K. Cibis, S. Hahnke, Y. S. Kim, T. Langer, J. Kreubel, M. Erhard, A. Bremges, S. Off, Y. Stolze, S. Jaenicke, A. Sczyrba, P. Scherer, H. König, W. H. Schwarz, A. Pühler, A. Schlüter, and M. Klocke. **Unraveling the microbiome of a thermophilic biogas plant by metagenome and metatranscriptome analysis complemented by characterization of bacterial and archaeal isolates.** *Biotechnology for Biofuels*, 2016. DOI: 10.1186/s13068-016-0581-3

11. Stolze, Y.*, A. Bremges*, M. Rumming, C. Henke, I. Maus, A. Pühler, A. Sczyrba, and A. Schlüter. **Identification and genome reconstruction of abundant distinct taxa in microbiomes from one thermophilic and three mesophilic production-scale biogas plants.** *Biotechnology for Biofuels*, 2016. DOI: 10.1186/s13068-016-0565-3. ★

10. Ortseifen, V., Y. Stolze, I. Maus, A. Sczyrba, A. Bremges, S. Albaum, S. Jaenicke, J. Fracowiak, A. Pühler, and A. Schlüter. **An integrated metagenome and metaproteome analysis of the microbial community residing in a biogas production plant.** *Journal of Biotechnology*, 2016. DOI: 10.1016/j.jbiotec.2016.06.014.

9. Maus, I., K. G. Cibis, A. Bremges, Y. Stolze, D. Wibberg, G. Tomazetto, J. Blom, A. Sczyrba, H. König, A. Pühler, and A. Schlüter. **Genomic characterization of *Defluviitoga tunisiensis* L3, a key hydrolytic bacterium in a thermophilic biogas plant and its abundance as determined by metagenome fragment recruitment.** *Journal of Biotechnology*, 2016. DOI: 10.1016/j.jbiotec.2016.05.001.

8. Wibberg, D.*, A. Bremges*, T. Dammann-Kalinowski, I. Maus, I. Igeño, R. Vogelsang, C. König, V. M. Luque-Almagro, D. Roldán, A. Sczyrba, C. Moreno-Vivián, R. Blasco, A. Pühler, and A. Schlüter. **Finished genome sequence and methylome of the cyanide-degrading *Pseudomonas pseudoalcaligenes* strain CECT5344 as resolved by single-molecule real-time sequencing.** *Journal of Biotechnology*, 2016. DOI: 10.1016/j.jbiotec.2016.04.008. ★

7. Bremges, A., E. Singer, T. Woyke, and A. Sczyrba. **MeCorS: metagenome-enabled error correction of single cell sequencing reads.** *Bioinformatics*, 2016. DOI: 10.1093/bioinformatics/btw144. ★

6. Belmann, P., J. Dröge, A. Bremges, A. C. McHardy, A. Sczyrba, and M. D. Barton. **Bioboxes: standardised containers for interchangeable bioinformatics software.** *GigaScience*, 2015. DOI: 10.1186/s13742-015-0087-0.

5. Bremges, A., I. Maus, P. Belmann, F. G. Eikmeyer, A. Winkler, A. Albersmeier, A. Pühler, A. Schlüter, and A. Sczyrba. **Deeply sequenced metagenome and metatranscriptome of a biogas-producing microbial community from an agricultural production-scale biogas plant.** *GigaScience*, 2015. DOI: 10.1186/s13742-015-0073-6. ★

4. Kohrs, F., S. Wolter, D. Benndorf, R. Heyer, M. Hoffmann, E. Rapp, A. Bremges, A. Sczyrba, A. Schlüter, and U. Reichl. **Fractionation of biogas plant sludge material improves metaproteomic characterization to investigate metabolic activity of microbial communities.** *Proteomics*, 2015. DOI: 10.1002/pmic.201400557.

3. Wibberg, D., V. M. Luque-Almagro, I. Igeño, A. Bremges, D. Roldán, F. Merchán, L. P. Sáez, I. Guijo, M. I. Manso, D. Macías, P. Cabello, G. Becerra, I. Ibáñez, I. Carmona, M. P. Escribano, F. Castillo, A. Sczyrba, C. Moreno-Vivián, R. Blasco, A. Pühler, and A. Schlüter. **Complete genome sequence of the cyanide-degrading bacterium *Pseudomonas pseudoalcaligenes* CECT5344.** *Journal of Biotechnology*, 2014. DOI: 10.1016/j.jbiotec.2014.02.004.

2. Gillett, A., P. Bergman, R. Parsa, A. Bremges, R. Giegerich, and M. Jagodic. **A silent exonic SNP in *kdm3a* affects nucleic acids structure but does not regulate experimental autoimmune encephalomyelitis.** *PLoS ONE*, 2013. DOI: 10.1371/journal.pone.0081912.

1. Bremges, A.*, S. Schirmer*, and R. Giegerich. **Fine-tunig structural RNA alignments in the twilight zone.** *BMC Bioinformatics*, 2010. DOI: 10.1186/1471-2105-11-222. ★

Acknowledgements

I owe to co-workers, advisors, friends, and family:

Alexander Sczyrba and Jens Stoye for the opportunity to conduct my doctoral studies; guidance, impulses, feedback, tips and tricks, and overall an empowering work athmosphere.

Tanja Woyke, Alfred Pühler, and Andreas Schlüter for inviting me to work on exciting projects, sharing their biological data and expertise, and answering dummy questions.

Robert Giegerich, Jörn Kalinowski, and Ulf Kirse for help when needed, effectively enabling my doctoral studies.

Irena Maus, Peter Belman, and David Laehnemann for being awesome colleagues, contributing data and analyses, and proof-reading parts of my thesis.

CLIB-GC and DiDy for (partial) funding and generous travel funds. I particularly thank the program coordinators, Iris Brune and Roland Wittler, for their support.

Katharina, Benjamin, and Eva for everything – I love you!

Part I

GENOMES FROM SINGLE CELLS

1 *The cutting edge of single cell assembly*

Single cell genomics revolutionized our understanding of biology by bringing the study of genomes to the cellular level. The sequencing of single microbial cells from environmental samples grants access to the genetic makeup of as-yet unculturable bacterial phyla and major archaeal groups.[1]

In 2007, the first single cell genomes – rare and uncultivated members of the TM7 phylum from the human mouth – were amplified and sequenced.[2] Since then, single amplified genomes (SAGs) were generated for more candidate phyla, *e.g.*

- OP11 (from an anoxic spring),[3]
- SR-1 (from human oral mucosa),[4]
- TM6 (from biofilm on a hospital sink),[5]
- OP9 (from a hot spring),[6] and
- JS1 (from marine sediment),[7]

shedding light on their phylogeny and physiology. In the largest (microbial) single cell sequencing study to date, Tanja Woyke and colleagues from the Joint Genome Institute generated 201 SAGs of unculturable microorganisms from diverse environments, uncovering biological phenomena, such as an archaeal-type purine synthesis in Bacteria and complete sigma factors in Archaea.[8]

The current state of the art in single cell genomics has been reviewed extensively, highlighting recent (and mostly technical) advancements.[9] After the physical separation and lysis of an individual cell, its DNA needs to be amplified before it

[1] Ishoey et al., 2008; Stepanauskas, 2012; Clingenpeel et al., 2014a

[2] Marcy et al., 2007; Podar et al., 2007

[3] Youssef et al., 2011

[4] Campbell et al., 2013

[5] McLean et al., 2013

[6] Dodsworth et al., 2013

[7] Nobu et al., 2016

[8] Rinke et al., 2013

[9] Lasken, 2013; Blainey and Quake, 2014; Eberwine et al., 2014; Gawad et al., 2016

is sequenced.[10] Almost all studies to date have used multiple displacement amplification (MDA).[11]

This amplification is heavily biased and leads to highly uneven sequencing depth, including ultra-low coverage regions.[12] To make things worse, chimera formation occurs roughly once per 10 kbp.[13] Alternatives to MDA – such as MALBAC[14] – were developed, but their amplification of microbial genomes is even less reliable than MDA.[15]

[10] Rinke et al., 2014
[11] Lasken, 2007; Gawad et al., 2016
[12] Chitsaz et al., 2011
[13] Rodrigue et al., 2009
[14] Zong et al., 2012
[15] Blainey, 2013; de Bourcy et al., 2014

ASSEMBLING MICROBIAL DARK MATTER GENOMES from single cells therefore remains a bioinformatics challenge. In this chapter, I benchmark state-of-the-art single cell assemblers on real sequencing data to access genome recovery and error rates. Consequently, I emphasize current limitations and advocate the use of metagenomic sequencing data to improve SAG assembly.

1.1 *Reference single amplified genomes*

As a realistic benchmark, I use 24 publicly available SAGs from three bacterial strains: *Escherichia coli* K12-MG1655 (51% *GC*), *Meiothermus ruber* DSM 1279 (63% *GC*), and *Pedobacter heparinus* DSM 2366 (42% *GC*).[16] For each strain, the complete genome sequence is known and eight SAGs were sequenced to a mean coverage of 315×.[17]

I used Bowtie 2[18] to map all SAG reads on the corresponding reference genome and SAMtools[19] to sort the alignment file and calculate mapping statistics (Table 1.1). I used Circos[20] to generate the circular read coverage plots (Figure 1.1).

[16] Clingenpeel et al., 2014a
[17] Clingenpeel et al., 2014b
[18] Langmead and Salzberg, 2012
[19] Li et al., 2009
[20] Krzywinski et al., 2009

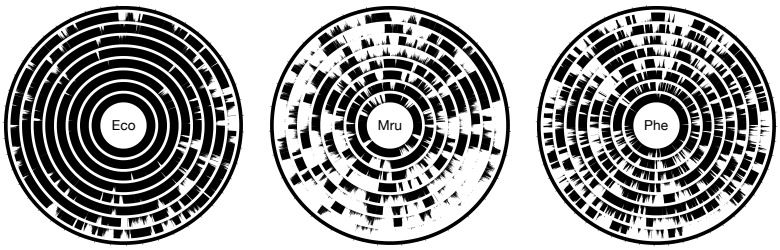


Figure 1.1: **Read coverage.** Circular coverage tracks, capped at 10×, for *E. coli* (Eco), *M. ruber* (Mru), and *P. heparinus* (Phe) SAGs.

Reference	SAG	# reads	# bases	% ref. covered
E. coli	0	9365134	1404770100	96.41
	1	9604918	1440737700	99.37
	2	8811278	1321691700	96.88
	3	8396488	1259473200	92.13
	4	9257066	1388559900	87.91
	6	8609900	1291485000	100.00
	7	8990744	1348611600	100.00
	8	9682468	1452370200	98.25
M. ruber	0	2859916	428987400	88.26
	2	4661806	699270900	88.50
	3	4274040	641106000	45.67
	5	1091400	163710000	74.60
	6	2236560	335484000	47.18
	7	1770260	265539000	54.89
	8	2605244	390786600	73.50
	9	2188386	328257900	62.02
P. heparinus	1	8604456	1290668400	78.26
	3	9064332	1359649800	83.96
	4	7856752	1178512800	55.02
	5	7793844	1169076600	81.44
	6	5194106	779115900	71.07
	7	6025058	903758700	96.93
	8	9106278	1365941700	79.90
	9	7995640	1199346000	76.63

Table 1.1: **Sequencing statistics.** Number of reads, bases, and genome fraction with at least 1× coverage for the *E. coli*, *M. ruber*, and *P. heparinus* SAGs.

The MDA-induced coverage bias varies from SAG to SAG and does not follow any obvious pattern. Genome coverage is the most important variable for *de novo* genome assembly; the 24 reference SAGs are therefore realistic benchmarking data.

1.2 *Assessment of single cell assemblers*

All modern single cell assemblers – IDBA-UD[21], SPAdes[22], and MEGAHIT[23] – use de Bruijn graphs as their underlying data structures. How to apply de Bruijn graphs to genome assembly is well-known and has been reviewed extensively.[24]

One notable characteristic that all three assemblers share, is the use of multiple *k*-mer sizes to increase assembly contiguity

[21] Peng et al., 2012

[22] Bankevich et al., 2012

[23] Li et al., 2015

[24] Compeau et al., 2011; Nagarajan and Pop, 2013

while avoiding misassemblies. Iteratively increasing the k-mer size from small (*e.g.* $k = 21$) to large (*e.g.* $k = 99$), they try to assemble low-coverage regions (small k) as well as resolve genomic repeats (large k).

To BENCHMARK single cell assemblers, I assemble the 24 reference SAGs generated from *E. coli* (Eco), *M. ruber* (Mru), and *P. heparinus* (Phe) with IDBA-UD, MEGAHIT, and SPAdes, using the default settings recommended for SAG assembly. I also include "perfect" assemblies – $Gold(x)$ – for each SAG, generated by extracting all regions of the reference genome with SAG read coverage of at least x, as gold standards to compare against.

Figure 1.2 shows the recovered genome fraction in contigs greater than 500 bp, as determined by QUAST[25], for all SAGs including gold standard assemblies for $x \in 1, 2, 4, 8, 16$.

[25] Gurevich et al., 2013

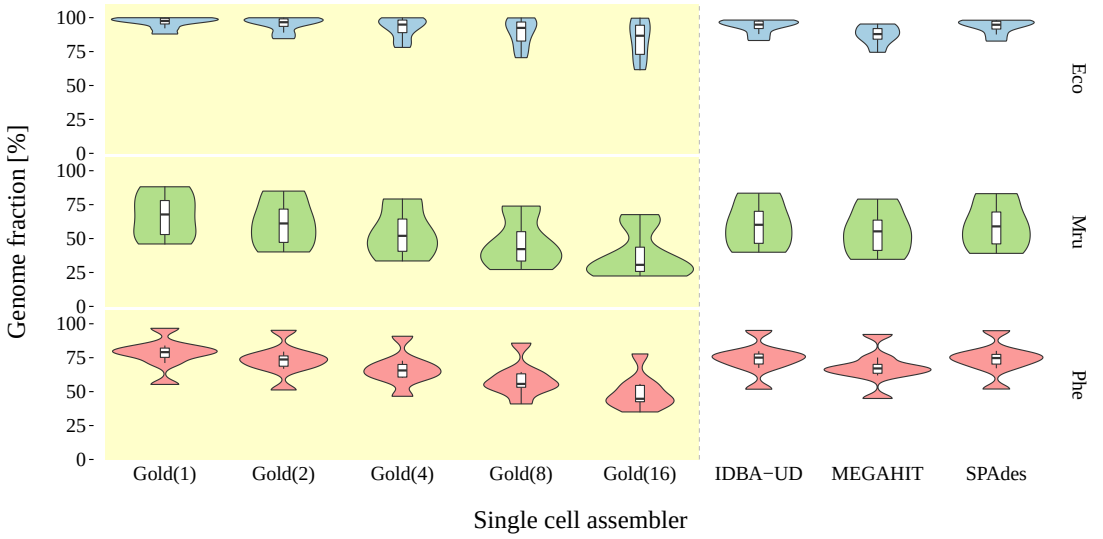


Figure 1.2: **Genome fraction.** Quality assessment with QUAST.

As expected, the recovered genome fraction varies a lot, but differences between assemblers are much smaller than differences between SAGs (and organisms) and can be attributed largely to the quality of the data for the respective SAG. IDBA-UD and SPAdes approach genome recovery rates close to $Gold(2)$,

i.e. they also assemble ultra-low coverage regions of the single cell. MEGAHIT performs slightly worse in this metric – but surprisingly well given that assembling single cell data was only recently added and is still flagged as an experimental feature.

Genome assemblers use heuristic methods to minimize assembly errors while maximizing contiguity.[26] Errors are either local (*i.e.* mismatches or indels) or of larger scale (*i.e.* rearrangements or chimeric contigs).[27]

[26] Earl et al., 2011; Bradnam et al., 2013

[27] Gurevich et al., 2013

Figure 1.3 shows the total amount of such assembly errors per 100 *kbp* assembly for the three assemblers and the (error-free) perfect assembly, $Gold(1)$. SPAdes and IDBA-UD outperform MEGAHIT, and – unsurprisingly – the *M. ruber* (Mru) SAG assemblies contain the most errors.

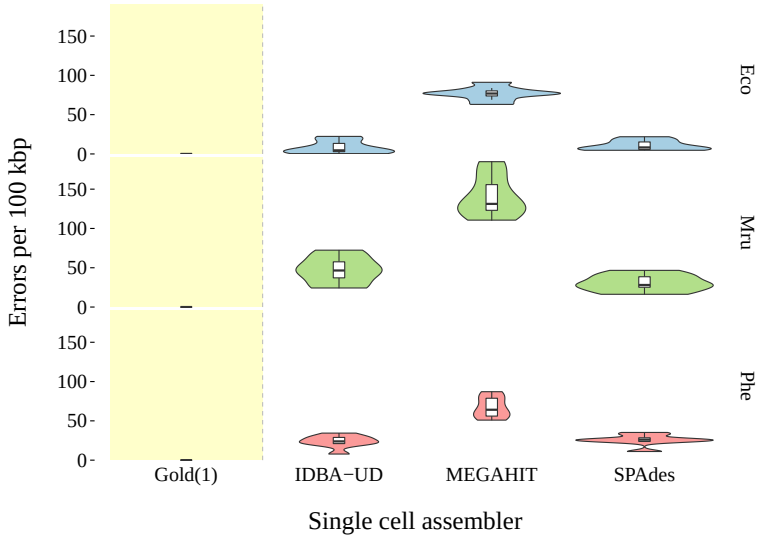


Figure 1.3: **Assembly errors.** Mismatches, indels, and misassemblies count as errors. Quality assessment with QUAST.

Assembly contiguity is usually evaluated by the N50 metric. The N50 value is the length-weighted median contig size, *i.e.* half of the total assembly is contained in contigs of length larger than (or equal to) the N50 value.

There are two problems with this metric, as the N50 (1) is not comparable between assemblies of different lengths, and (2) does

not account for assembly errors, especially misassembly events.

If the reference genome is known, the NGA50 can be used instead; a useful combination of NG50 (normalize by the real genome length instead of assembly size) and NA50 (break contigs at large-scale misassemblies). This metric is also implemented in QUAST and Figure 1.4 shows that assembly contiguity could theoretically be improved for all assemblers and SAGs.

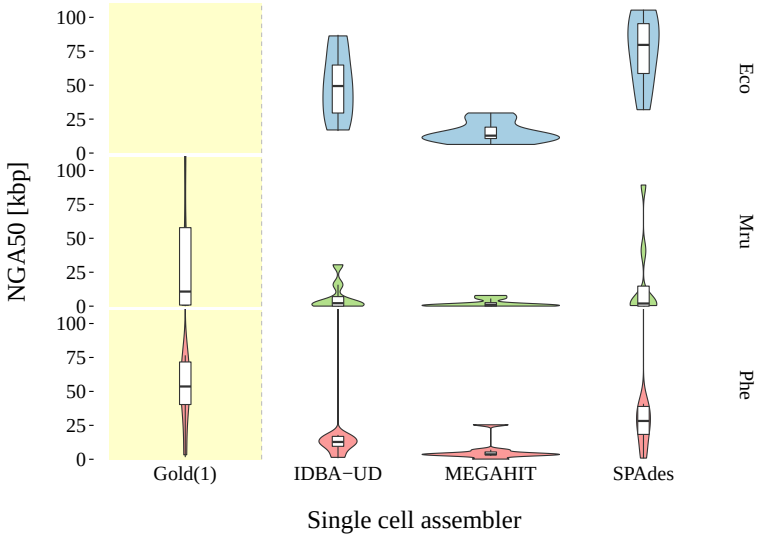


Figure 1.4: **Assembly contiguity.** The NGA50 values for Eco $Gold(1)$ assemblies are literally off the charts. Quality assessment with QUAST.

To CONCLUDE, modern single cell assemblers already do a good job recovering much of the genome – even regions covered only barely –, but there is room for improvement when it comes to error rate and assembly contiguity.

1.3 *Evolution of the SPAdes assembler*

The SPAdes assembler is one of the most widely used assemblers today, probably thanks to its ease of use combined with favorable rankings in genome assembly benchmarks.[28] It has been under active development for more than three years, with – according to the changelog – significant improvements in terms of genome

[28] Magoc et al., 2013; Jünemann et al., 2014

recovery, accuracy, and contiguity (not to mention runtime and memory requirements, which also improved).

INSPIRED BY REGRESSION TESTING in software development, I in retrospect assemble the 24 reference SAGs with each (major) version of the SPAdes assembler and compare the results.

The genome fraction different SPAdes versions assemble into contigs greater than 500 *bp* is almost constant (Figure 1.5).

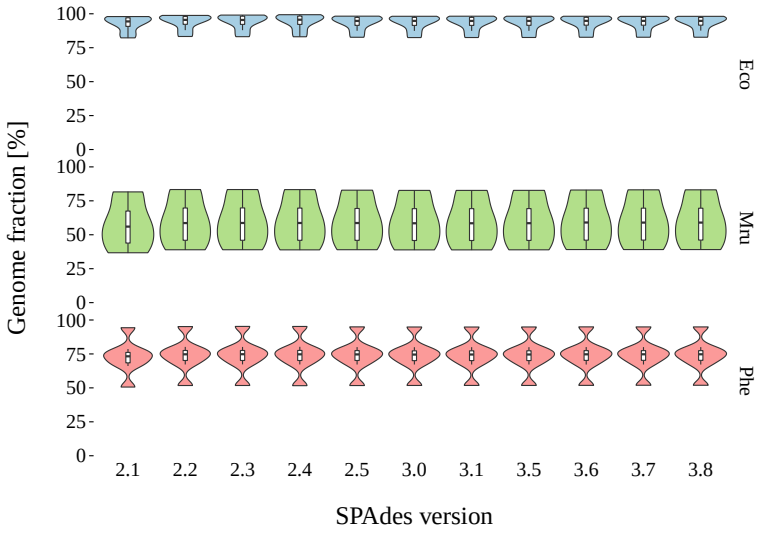


Figure 1.5: **Genome fraction.** Quality assessment with QUAST.

The assembly accuracy, assessed by the sum of mismatches, indels, and misassemblies per 100 *kbp* assembly, is depicted in Figure 1.6; contiguity in terms of NGA50 is shown in Figure 1.7. These results suggest that the developers of SPAdes first focused on improving assembly contiguity at the cost of introducing more errors, and then worked on the latter.

1.4 Conclusions

Single cell genome assembly algorithms matured and reconstruct most of the genome represented by SAG reads. Compared to the gold standard, assembly accuracy and contiguity eventually

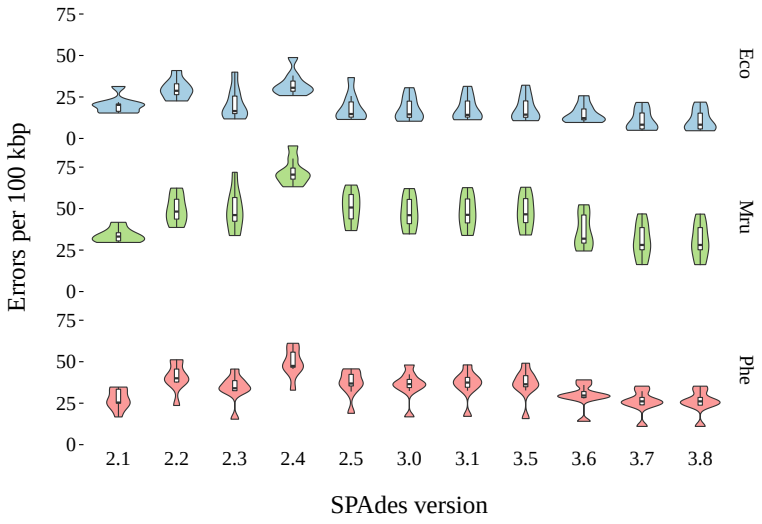


Figure 1.6: **Assembly errors.** Quality assessment with QUAST.

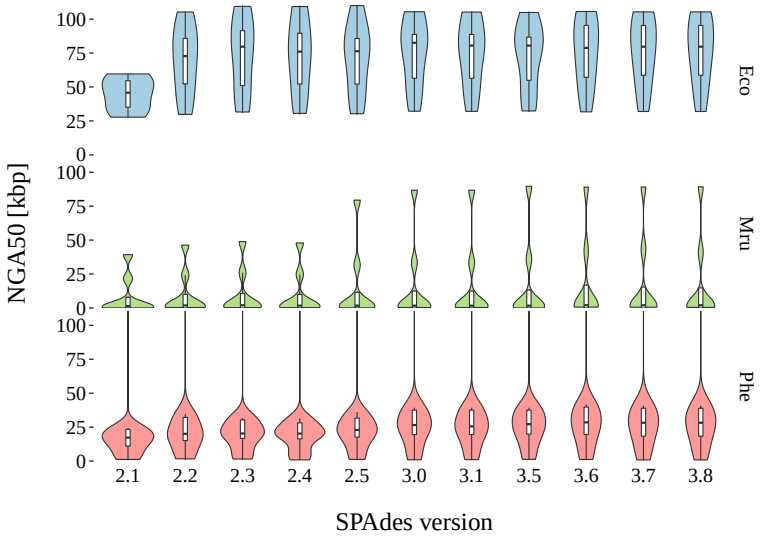


Figure 1.7: **Assembly contiguity.** Quality assessment with QUAST.

increased for more recent versions of the tool of choice, SPAdes. Also, technical advancements, such as an improved DNA amplification with less bias or the direct sequencing of a cell's DNA, will eventually enable better single cell genome assemblies.

REFRAMING A PROBLEM helps to unlock innovation. In the following two chapters, I therefore propose new bioinformatic methods that take a different perspective to improve upon the state-of-the-art in single cell assembly, leveraging another data type that is often available for SAGs: the shotgun metagenome of the environmental sample that the SAG was generated from.

I conceived and developed MECORS and KGREP. Both tools incorporate unbiased metagenomic sequence information to increase the accuracy, contiguity, and genome recovery rate for single cell genomes.

2 *Metagenome-enabled error correction*

Correcting potential errors in sequencing reads prior to assembly usually improves the downstream assembly result.[1] Modern error correction tools typically use algorithms similar to solving the *spectral alignment problem*.[2] Given a set of trusted *k*-mers, they try to find a sequence with minimal corrections such that each *k*-mer on the corrected sequence is trusted. When sequencing isolate-grade genomes, a simple *k*-mer coverage threshold can be used to accurately distinguish between trusted and untrusted *k*-mers.[3]

[1] Laehnemann et al., 2016

[2] Pevzner et al., 2001

[3] Kelley et al., 2010; Song et al., 2014

However, a single cell's DNA needs to be amplified prior to sequencing, as usually accomplished by multiple displacement amplification (MDA).[4] This amplification is heavily biased, leads to uneven sequencing depth throughout the single amplified genome (SAG), and thus revokes the assumption of uniform sequencing depth that most error correction tools make. Only one tool was specifically designed to correct SAG data with uneven sequencing depth: hammer[5], recently refined to BayesHammer[6].

[4] Lasken, 2007

[5] Medvedev et al., 2011

[6] Nikolenko et al., 2013

I PROPOSE MECORS, a metagenome-enabled error correction strategy for single cell sequencing reads.[7] Frequently, single cells and shotgun metagenomes are generated from the same environmental sample, and are methodologically combined *e.g.* to validate metagenome bins with single cell reads or to improve the SAG's assembly contiguity.[8] MECORS takes advantage of largely unbiased metagenomic coverage, enabling it to correct positions with too low a coverage for SAG-only error correction, and to correct chimeric SAG reads through non-chimeric metagenome reads.

[7] Bremges et al., 2016

[8] Hess et al., 2011; Campbell et al., 2013

2.1 *Cartoonesque k-mer relationship*

If a SAG and a shotgun metagenome were generated from one sample, then the organism represented by the single cell is also a member of the microbial community (otherwise it would have been impossible to capture it). Therefore – assuming sufficient metagenomic sequencing coverage – all true genomic *k*-mers in the SAG data have to occur in the metagenome, too. Any *k*-mers without support in the metagenome likely originate from sequencing errors or MDA-induced chimeric junctions in the single cell sequencing reads (Figure 2.1).

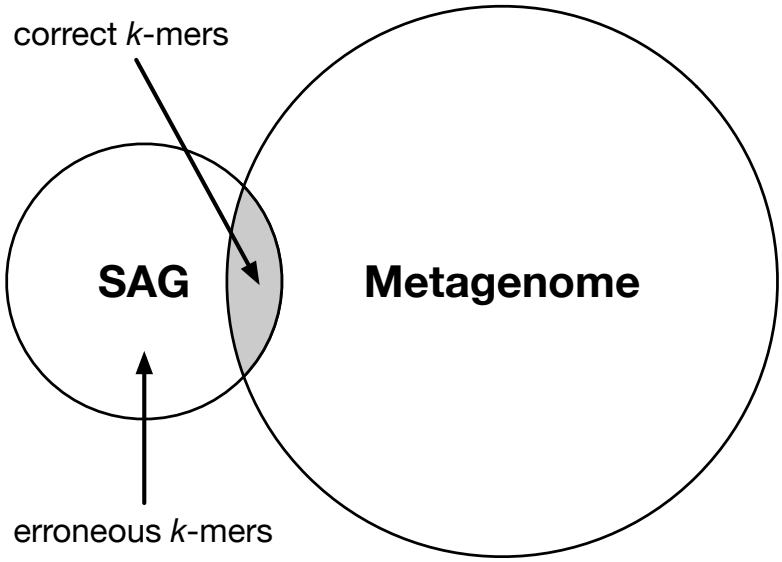


Figure 2.1: **Simplified *k*-mer relationship.** True genomic *k*-mers are shared between a SAG and its accompanying metagenome.

2.2 *Error correction algorithm*

The correction algorithm of MECORS was inspired by fermi[9] and BFC[10], but it does not act on the assumption of uniform sequencing coverage. Instead, it exploits metagenomic sequence information to correct errors resulting from amplification and sequencing, as well as chimeras, even in ultra-low coverage regions of the SAG.

[9] Li, 2012

[10] Li, 2015

MeCorS works in three phases:

1. MeCorS collects all 31-mers (and their reverse complements) occurring in the SAG reads. It uses this information to initialize a hash table with the 31-mers being valid keys.

2. MeCorS scans the accompanying metagenomic reads. For each stored 31-mer, it counts the occurrence of the next (i.e. the 32nd) base in the metagenome and stores the totals in the hash table. This step is largely I/O bound and dominates MeCorS's runtime.

3. MeCorS processes each SAG read by using the 31-mer hash table to check if the 32nd base is sufficiently supported in the metagenome. Untrusted 32nd bases are replaced with the most frequent and trusted 32nd bases from the metagenome.

MeCorS considers a 31-mer trusted if it occurs at least twice in the accompanying metagenome. This coverage threshold were determined empirically (as discussed further down) and the *k*-mer size of 31 for error correction was chosen according to the literature.[11] Both parameters can be adjusted by the user to potentially improve MeCorS's performance for specific data sets.

[11] Li, 2015

The non-chimeric nature of the metagenome reads enables a correction of chimeric SAG reads. Metagenome sequencing is largely unbiased and free of chimeras, while MDA introduces chimeric junctions roughly once per 10 kbp in SAGs.[12]

[12] Lasken and Stockwell, 2007

Chimeric reads contain DNA sequences originating from two different genome regions, say *A* and *B*, with the first part originating from region *A*, the second part from region *B*. A chimeric junction will (in most cases) result in an untrusted 32nd base (from region *B*) when looking at its 31-mer prefix (from region *A*; phase 3 of MeCorS). MeCorS then tries to correct this position of the SAG read by replacing the untrusted 32nd base (*B*) with the most frequent and trusted 32nd base from the metagenome (*A*). MeCorS therefore performs an implicit and thorough write-through correction of chimeric SAG reads, completely rewriting their second parts.

2.3 *Reference SAGs and mock metagenome*

As a realistic benchmark, I used the eight previously described *Escherichia coli* K12-MG1655 reference SAGs.[13] A concomitant *in vitro* mock metagenome consisting of 3 archaeal and 23 bacterial species, including *E. coli* K12-MG1655, was sequenced on the Illumina HiSeq 2000 platform using 2×150 *bp* paired-end sequencing and generating a total of 355,875,608 reads (53 *Gbp*).[14]

I mapped these with BWA-MEM[15] simultaneously against all 26 reference genomes, postprocessed the alignment files with SAMtools[16], and calculated the per-base coverage values (Table 2.1). The relative abundance of *E. coli* is 0.15%, corresponding to a mean per-base coverage of only 20.7×. The taxonomic profile of the mock community, visualized with Krona[17], is shown in Figure 2.2.

[13] Clingenpeel et al., 2014b

[14] Bowers et al., 2015

[15] Li, 2013

[16] Li et al., 2009

[17] Ondov et al., 2011

Taxonomy ID	Phylum	Species	# reads	coverage	abundance
771875	Thermotogae	*Fervidobacterium pennivorans*	39566833	2708.24	19.39%
646529	Firmicutes	*Desulfosporosinus acidophilus*	53202915	1579.64	11.31%
526227	Deinococcus-Thermus	*Meiothermus silvanus*	32231620	1276.02	9.14%
573413	Spirochaetes	*Spirochaeta smaragdinae*	39431130	1255.97	8.99%
582402	Proteobacteria	*Hirschia baltica*	28144226	1181.16	8.46%
717605	Firmicutes	*Thermobacillus composti*	30954326	1046.39	7.49%
767817	Firmicutes	*Desulfotomaculum gibsoniae*	24329020	741.83	5.31%
767434	Proteobacteria	*Frateuria aurantia*	13922996	568.16	4.07%
633147	Actinobacteria	*Olsenella uli*	7839526	552.99	3.96%
768704	Firmicutes	*Desulfosporosinus meridiei*	16066750	487.57	3.49%
583355	Verrucomicrobia	*Coraliomargarita akajimensis*	11810956	467.03	3.34%
694430	Euryarchaeota	*Natronococcus occultus*	12505736	403.70	2.89%
797304	Euryarchaeota	*Natronobacterium gregoryi*	8676937	335.23	2.40%
797302	Euryarchaeota	*Halovivax ruber*	6060380	274.95	1.97%
926566	Acidobacteria	*Terriglobus roseus*	8464573	239.33	1.71%
640132	Actinobacteria	*Segniliparus rotundus*	4886507	225.63	1.62%
644801	Proteobacteria	*Pseudomonas stutzeri*	5448168	174.50	1.25%
160490	Firmicutes	*Streptococcus pyogenes*	1502208	120.44	0.86%
195103	Firmicutes	*Clostridium perfringens*	1461840	66.55	0.48%
203119	Firmicutes	*Clostridium thermocellum*	1542460	59.62	0.43%
882884	Proteobacteria	*Salmonella enterica*	1831145	58.72	0.42%
926556	Bacteroidetes	*Echinicola vietnamensis*	2160987	57.22	0.41%
196627	Actinobacteria	*Corynebacterium glutamicum*	1063668	47.67	0.34%
511145	**Proteobacteria**	***Escherichia coli***	**647555**	**20.65**	**0.15%**
218493	Proteobacteria	*Salmonella bongori*	501312	16.61	0.12%
446468	Actinobacteria	*Nocardiopsis dassonvillei*	6640	0.06	$< 0.01\%$

Table 2.1: **Mock community members.** 26 microbial species.

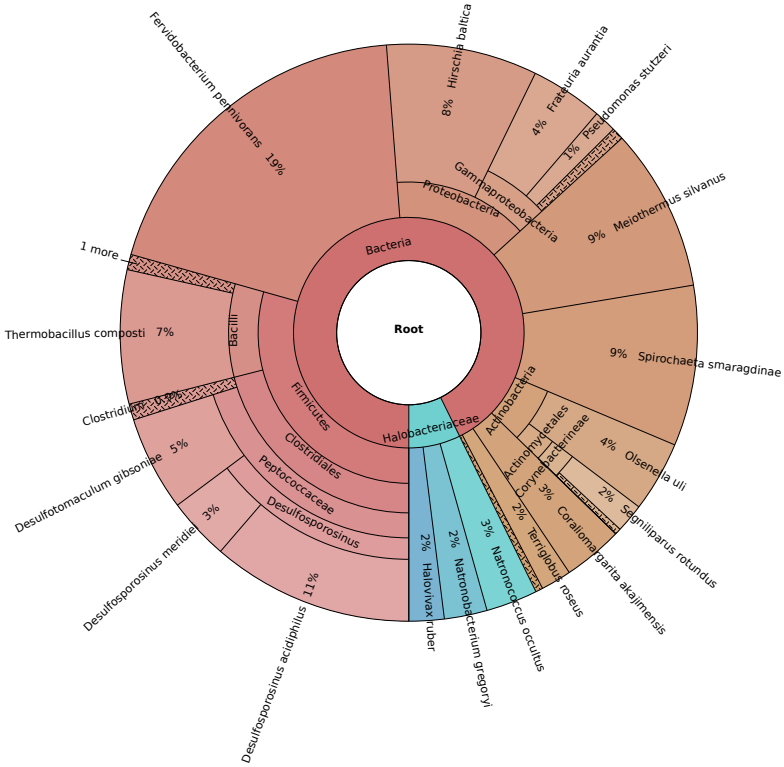


Figure 2.2: **Mock community profile.** 26 microbial species.

2.4 *Performance of SAG error correction*

I evaluated MeCorS along with BayesHammer[18], the state-of-the-art error correction tool for SAG data. I evaluated the performance of read error correction as described in Li, 2015, using BWA-MEM[19] and calculating the same read-based metrics:[20]

[18] Nikolenko et al., 2013

[19] Li, 2013

[20] Li, 2015

> "A read is said to become *better* (or *worse*) if the best alignment of the corrected sequence has more (or fewer) identical bases to the reference genome than the best alignment of the original sequence. The table gives [...] the number of reads mapped *perfectly*, number of *chimeric* reads (i.e. reads with parts mapped to different places), number of corrected reads becoming *better* and the number of corrected reads becoming *worse* than the original reads."

MeCorS corrects more errors than BayesHammer, producing a significantly higher fraction of better and perfect reads after correction (Table 2.2; detailed statistics for each SAG, including runtime and memory usage, are given in Table 2.3).

Program	% perfect	% chimeric	% better	% worse
raw	22.52 ± 1.07	0.73 ± 0.15	–	–
BayesHammer	80.35 ± 8.77	0.77 ± 0.17	71.66 ± 2.12	0.33 ± 0.06
MeCorS	$\mathbf{95.52 \pm 0.43}$	$\mathbf{0.06 \pm 0.02}$	$\mathbf{75.45 \pm 1.11}$	$\mathbf{0.26 \pm 0.03}$

Table 2.2: **Performance of SAG error correction.** Evaluation as described in Li, 2015.

In contrast to BayesHammer, MeCorS also considerably reduces the amount of chimeric SAG reads, likely due to the non-chimeric nature of the metagenome reads. Despite only implicitely correcting chimeric SAG reads, MeCorS reduces the amount of chimeras by one order of magnitude. Chimeric reads originate from amplification errors during MDA and greatly complicate *de novo* SAG assembly.[21] Therefore, I look at the effect of the improved error correction on SAG assembly next.

[21] Nurk et al., 2013

		SAG							
Metric	Program	0	1	2	3	4	6	7	8
Reads	–	9365134	9604918	8811278	8396488	9257066	8609900	8990744	9682468
Perfect	raw	2120932	2179609	1937541	1954244	1872800	2049675	2063874	2183454
	BayesHammer	7656274	8260510	6302861	5970186	6297298	7639715	8068555	8317006
	MeCorS	**8886436**	**9188854**	**8440502**	**7965810**	**8829995**	**8264229**	**8611867**	**9272559**
Chimeric	raw	69568	67625	81938	75509	79443	52246	44983	59265
	BayesHammer	72820	70590	87564	80336	84813	53875	46387	61948
	MeCorS	**5502**	**4593**	**10397**	**4648**	**5257**	**3941**	**3824**	**4889**
Better	raw	–	–	–	–	–	–	–	–
	BayesHammer	6743983	7026478	6156387	5669978	6574627	6244274	6645542	7095951
	MeCorS	**7008163**	**7236195**	**6707136**	**6260408**	**7206731**	**6403639**	**6749255**	**7306961**
Worse	raw	–	–	–	–	–	–	–	–
	BayesHammer	31644	26951	32315	29096	41584	25270	26959	28960
	MeCorS	**25990**	**25304**	**20560**	**27335**	**27390**	**20791**	**21074**	**24001**
Time (h)	BayesHammer	1:43	1:45	1:51	2:04	2:24	1:35	1:38	1:46
	MeCorS	**1:13**	**1:09**	**0:58**	**0:53**	**1:06**	**1:00**	**1:08**	**1:13**
RAM (GB)	BayesHammer	14.90	15.64	13.70	12.34	15.33	14.59	15.72	15.77
	MeCorS	**10.77**	**10.75**	**10.76**	**9.65**	**10.76**	**10.79**	**10.75**	**10.75**

Table 2.3: **Detailed performance.** Time using 16 threads.

2.5 *Effect on SAG assembly*

I used IDBA-UD[22] and SPAdes[23] to assemble raw and corrected SAG reads, and QUAST[24] to evaluate the 48 SAG assemblies.

SPAdes was run with the parameters `--careful` (to minimize the number of mismatches in the final contigs) and `-k 21,33,55,77` (to account for longer SAG sequencing reads; iterating over these four *k*-mer sizes generated assemblies of higher contiguity than the default settings of `-k 21,33,55`, while maintaining a high accuracy).

[22] Peng et al., 2012

[23] version 3.6.0; Bankevich et al., 2012

[24] Gurevich et al., 2013

MeCorS works well with both single cell assemblers, most notably reducing their misassembly rate by half, while providing high sequence contiguity (Figure 2.3).

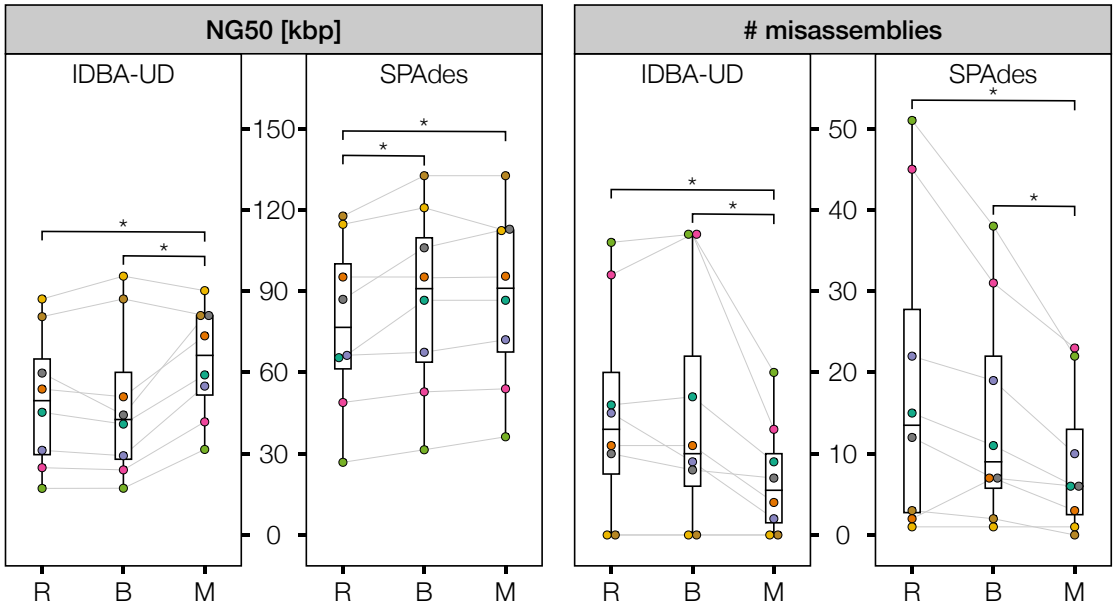


Figure 2.3: **Effect on SAG assembly.** We corrected the raw reads (R) with BayesHammer (B) or MeCorS (M). Statistical significance ($p < 0.05$; two-tailed Wilcoxon signed-rank test) accentuated.

In particular poorly amplified SAGs benefit from metagenome-enabled error correction, yielding improved assembly accuracy and contiguity (Table 2.4; Table 2.5).

While there are subtle differences between the IDBA-UD and SPAdes assemblies, both results demonstrate the large potential of metagenome-enabled error correction.

Metric	Program	SAG 0	1	2	3	4	6	7	8
NG50	R	45284	53863	31250	24834	17196	87102	80574	59754
	B	40924	51004	29256	24023	17246	**95532**	**87102**	44292
	M	**59081**	**73496**	**54946**	**41749**	**31569**	90184	80997	**80997**
# contigs	R	330	342	514	607	654	201	200	303
	B	335	345	530	626	653	**191**	**194**	328
	M	**277**	**272**	**409**	**497**	**512**	200	198	**249**
Largest contig	R	227106	203026	**203098**	141383	102074	**221687**	**232585**	162612
	B	203098	157125	197417	141494	107872	**221687**	178322	139398
	M	**236473**	**203098**	144213	**141579**	**124628**	221683	221683	**236473**
Total length	R	4400079	4587934	4400469	4139052	3940625	4640153	**4639167**	4533515
	B	4402128	**4591089**	4400334	4144742	3934141	**4641409**	4638005	**4538546**
	M	**4408926**	4590112	**4421966**	**4171137**	**3972787**	4639453	4636457	4538141
# misassemblies	R	16	11	15	32	36	**0**	**0**	10
	B	17	11	9	37	37	**0**	**0**	8
	M	**9**	**4**	**2**	**13**	**20**	**0**	**0**	**7**
# mismatches per 100 kbp	R	**4.17**	**3.64**	10.16	17.02	20.41	0.31	0.13	3.16
	B	4.88	3.88	12.50	16.35	20.76	**0.15**	**0.11**	**2.90**
	M	4.38	3.80	**7.93**	**8.86**	**12.82**	2.30	2.39	3.27
# indels per 100 kbp	R	0.32	**0.24**	0.69	1.47	1.09	0.13	**0.09**	0.36
	B	0.32	0.29	0.79	1.52	1.51	0.11	**0.09**	0.34
	M	**0.25**	0.31	**0.53**	**0.90**	**0.85**	**0.09**	**0.09**	**0.18**
Genome fraction (%)	R	93.656	97.182	93.344	87.892	83.101	98.266	98.245	96.104
	B	93.631	97.165	93.304	87.924	82.958	98.211	98.175	96.028
	M	**93.928**	**97.431**	**93.988**	**88.827**	**84.053**	**98.271**	**98.252**	**96.265**
# genes	R	3873	4049	3741	3477	3220	4186	4191	4027
	B	3859	4052	3709	3469	3208	4194	4199	4005
	M	**3931**	**4126**	**3857**	**3593**	**3347**	**4204**	**4202**	**4088**

Table 2.4: **IDBA-UD assembly results**. Quality assessment with QUAST.

Metric	Program	SAG 0	1	2	3	4	6	7	8
NG50	R	65444	95218	66287	48903	26823	114661	117715	86966
	B	**86625**	95218	67436	52817	31448	**120770**	**132608**	105995
	M	**86625**	**95517**	**72055**	**53947**	**36236**	112350	**132608**	**112853**
# contigs	R	447	400	606	718	813	245	233	324
	B	302	**275**	474	594	676	**198**	**185**	**250**
	M	**288**	279	**418**	**534**	**569**	210	213	263
Largest contig	R	203603	**224667**	**218793**	**178300**	113773	269308	268816	223154
	B	**204882**	203257	**218793**	167410	135551	**312119**	269348	**269318**
	M	203394	224320	**218793**	178231	**155221**	268535	**312008**	268327
Total length	R	**4522153**	**4703061**	**4533214**	**4290463**	**4138591**	**4713277**	**4718163**	**4633297**
	B	4443696	4633876	4464269	4233011	4046113	4686582	4689565	4584513
	M	4452849	4645907	4471868	4240428	4065827	4698390	4702810	4600454
# misassemblies	R	15	**2**	22	45	51	**1**	3	12
	B	11	7	19	31	38	**1**	2	7
	M	**6**	3	**10**	**23**	**22**	**1**	**0**	**6**
# mismatches per 100 kbp	R	15.30	11.57	34.70	48.21	50.53	2.84	**2.14**	9.72
	B	12.70	10.30	30.34	40.41	48.42	**1.27**	2.17	**7.66**
	M	**10.41**	**9.32**	**22.69**	**30.86**	**36.47**	5.66	5.21	8.43
# indels per 100 kbp	R	0.89	1.17	2.26	4.48	4.24	0.31	**0.22**	1.00
	B	1.17	1.19	3.16	3.48	4.58	**0.24**	0.35	0.94
	M	**0.64**	**0.95**	**2.24**	**3.30**	**3.28**	0.55	0.31	**0.83**
Genome fraction (%)	R	**94.241**	**97.948**	**94.239**	89.098	84.372	**98.665**	**98.708**	**96.702**
	B	94.050	97.527	93.984	89.133	84.220	98.505	98.446	96.459
	M	94.223	97.629	94.223	**89.543**	**84.798**	98.580	98.541	96.603
# genes	R	3876	4117	3782	3532	3281	4217	**4224**	4081
	B	3898	4124	3805	3562	3300	4211	4219	4093
	M	**3937**	**4133**	**3866**	**3608**	**3390**	**4218**	4220	**4097**

Table 2.5: **SPAdes assembly results**. Quality assessment with QUAST.

2.6 *Metagenome coverage threshold*

By default, MeCorS considers a *k*-mer trusted if it occurs at least twice in the accompanying metagenome. The user can adjust this threshold to potentially improve its performance for specific data sets. Figure 2.4 shows the effect of a parameter sweep for the *E. coli* SAGs and the *in vitro* mock metagenome.

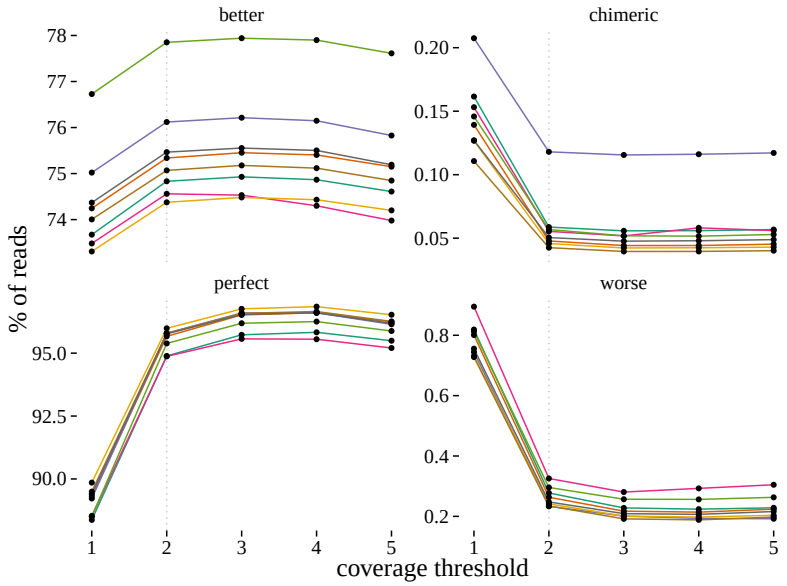


Figure 2.4: **Effect of different coverage thresholds.** MeCorS considers a *k*-mer trusted if it occurs at least twice in the metagenome.

Increasing the *k*-mer coverage threshold from 1 to 2 is the most beneficial, further increasing this threshold only marginally improves results. Above some coverage threshold error correction performance begins to decline, which seems to be dependent on the taget genome's metagenomic coverage. For the *E. coli* SAGs and the concomitant mock metagenome, this turning point seems to be around 4.

I recommend running MeCorS with default settings; they work sufficiently well for most SAG/metagenome combinations.

2.7 Conclusions

It should be noted that such a hybrid error correction of SAG data may result in miscorrection(s) of rare variants. If the captured cell contains a variant that is rare or absent in the corresponding metagenome, correction will be biased towards the most abundant variant in the metagenome sequence. If strain resolution is desired, I suggest polishing the SAG assembly with *e.g.* SEQuel[25] or Pilon[26] using the uncorrected raw data as input reads. In all other cases, SAG assemblies benefit directly from metagenome-enabled error correction via MeCorS.

[25] Ronen et al., 2012
[26] Walker et al., 2014

Uneven genome coverage and chimera formation present the biggest challenges in the downstream processing and analysis of SAG datasets to date. I developed MeCorS for the correction of SAG reads when complementary metagenome datasets are available. Error and chimera correction is essential for improved SAG assembly and demonstrates a powerful application of combined shotgun metagenome and single cell sequencing.

2.8 Software availability

MeCorS is implemented in C and is freely available under the open-source MIT license at:

`https://github.com/abremges/mecors`

3 *Metagenomic proxy assemblies*

Prior to sequencing of a single cell, its DNA needs to be amplified. This usually is done by multiple displacement amplification (MDA), introducing a tremendous coverage bias.[1] Poorly amplified regions result in extremely low sequencing coverage or physical sequencing gaps.[2] Those regions of the genome cannot be reconstructed in the subsequent assembly step and genomic information is lost.[3]

A complementary approach to single cell genomics is metagenomics, *i.e.* the direct sequencing of environmental samples. Frequently, single amplified genomes (SAGs) and shotgun metagenomes are generated from the same environmental sample.[4] In a metagenome, each genome's coverage is (more or less) constant and depends only on its abundance.[5]

I PROPOSE KGREP, a fast, *k*-mer based recruitment method to identify *metagenomic proxy reads* representing the single cell genome of interest (using the raw single cell sequencing reads as recruitment seeds). By assembling metagenomic proxy reads instead of the single cell reads, I circumvent most challenges of single cell assembly, such as the aforementioned coverage bias and chimeric MDA products.[6] In a final step, the original single cell reads are used for quality assessment of the proxy assembly.

A conceptionally similar approach – *mitochondrial baiting and iterative mapping* – has been used to reconstruct complete mitochondrial genomes from sequencing data.[7] My implementation, KGREP, is more flexible and *e.g.* allows to ignore *k*-mers originating from known contaminants, and is significantly faster, which is essential to process large metagenomic data sets.

[1] Lasken, 2007; Chitsaz et al., 2011; Nikolenko et al., 2013

[2] Bankevich et al., 2012; Bremges et al., 2016

[3] Nurk et al., 2013; Clingenpeel et al., 2014a

[4] Hedlund et al., 2014; Bremges et al., 2016

[5] Wooley et al., 2010; Escobar-Zepeda et al., 2015

[6] Lasken and Stockwell, 2007; Gole et al., 2013

[7] Hahn et al., 2013

3.1 *Metagenomic proxy reads*

The declared goal is to identify metagenomic sequencing reads that belong to a genome of interest. Let R be a metagenomic read of length $|R|$, composed of $|R| - k + 1$ k-mers (words of length k). If R belongs to genome G, for which the complete sequence is known, and R contains no sequencing errors, then it shares all of its k-mers with G (Figure 3.1).

Figure 3.1: **Error-free read.** A perfect read (R; green) shares all k-mers with its source genome (G; blue).

Without loss of generality, I can assume that sequencing errors are randomly distributed and that each error introduces at most k erroneous k-mers. The latter holds true for base substitutions and indels (insertions or deletions) of length 1, which happen to represent the vast majority of all sequencing errors in Illumina data.[8] Errors in the first or last $k - 1$ bases of R introduce less than k erroneous k-mers. Thus, the expected number N of shared k-mers between R and G depends on the read length $|R|$ and the sequencing error rate SER:

[8] Minoche et al., 2011; Goodwin et al., 2016

$$N = |R| - k + 1 - E,$$
$$\text{where: } E = \lceil |R| \cdot SER \rceil \cdot k$$

The above mimics Ukkonen's q-gram lemma for approximate string matching within a certain edit distance.[9] I estimate the expected edit distance between R and G as the product of read length and error rate. For Illumina data (the most predominant data type), this currently means read lengths of $\sim$ 150 bp and a per-base error rate of $\varepsilon = 0.1$–0.3%.[10] In other words, I expect less than one error per 150 bp read. Inserting these numbers into the error term's formula gives:

[9] Ukkonen, 1992

[10] Goodwin et al., 2016; Laehnemann et al., 2016

$$E = \lceil 150 \cdot \varepsilon \rceil \cdot k = k$$

I therefore allow an edit distance of 1 or k erroneous k-mers (Figure 3.2). I call a 150 bp metagenomic read a *metagenomic proxy read* if it shares $n \geq N$ k-mers with G. In theory, metagenomic proxy reads resemble isolate-grade genome data with a Poisson-like coverage distribution and I circumvent most challenges of single cell assembly, such as the aforementioned coverage bias and chimeric MDA products, by assembling the proxy reads instead of the original single cell ones.

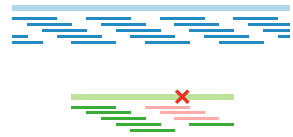


Figure 3.2: **Read with error.** A sequencing error introduces at most k erroneous k-mers.

Figure 3.3: **Starting from SAG reads.** No single cell assembly is needed to recruit metagenomic proxy reads.

This strategy does not require a closed genome sequence G, but it works just as good on the raw single cell sequencing reads instead. Single cell assemblies tend to be incomplete and usually do not include 100% of the reads as single cell sequencing reads are only sparsely scattered throughout low-coverage regions, which makes their assembly difficult or impossible.[11] Therefore, the initial read set contains more information than any SAG assembly and single cell reads should be used as a starting point to recruit metagenomic proxy reads (Figure 3.3).

[11] Nurk et al., 2013; Clingenpeel et al., 2014a

Analogous to above: Let R be a metagenomic read of length $|R|$, composed of $|R| - k + 1$ canonical k-mers, with the error rate SER. If R belongs to the genome G, for which only an arbitrarly large set of sequencing reads G' is known, then it shares at least $|R| - k + 1 - \lceil |R| \cdot SER \rceil \cdot k$ canonical k-mers with the read set G'. I use canonical k-mers to allow strand-neutral comparisons, *i.e.* we only consider the lexicographically smaller k-mer of the forward and reverse complement representation of a k-mer.

Each sequencing error (or chimeric junction) in the read set G' adds at most k erroneous k-mers to the reference (Figure 3.3). However, because errors are randomly distributed and infrequent, the added noise is insignificant as long as k is big enough.

3.2 *The choice of k*

In nucleotide space, *i.e.* with the well-known DNA alphabet $\sum = \{A, C, G, T\}$, there are 4^k possible non-canonical k-mers and the probability to observe a given k-mer K in an *i.i.d.* uniform random (meta)genome sequence X of length $|X|$ is:

$$P(K \in X) = 1 - \left(1 - \frac{1}{4^k}\right)^{|X|-k+1}$$

I can avoid random hits by choosing a suitable large value for k; k-mer sizes of 15 and 18 were suggested for bacterial and human genomes, resprectively.[12] For $k = 15$, the probability of observing a given k-mer in a 5 *Mbp* bacterial genome is 0.46%.

[12] Kelley et al., 2010

The choice of k, *e.g.* for read error correction or *de novo* genome assembly, is usually a tradeoff between sensitivity and specificity.[13] Longer k-mers are *per se* more specific, but to identify metagenomic proxy reads a too large value for k is detrimental (because $E = k$). Therefore, an optimal k is small, but large enough to avoid random k-mer hits between genomes.

[13] Li, 2015; Chikhi and Medvedev, 2014

Closely related genomes in the metagenome, *e.g.* multiple strains of one species or similar species of one genus, deserve my closer attention. Starting with single cell reads from only one strain A_1, kgrep probably recruits reads from strains $A_2, A_3, \ldots, A_n$ present in the metagenome. This might (or might not) pose a challenge for the downstream assembly of metagenomic proxy reads probably representing a strain-mixture.

The average nucleotide identity (ANI) between two genomes is a robust measure of genome relatedness; an ANI value of 95% roughly corresponds to a 70% DNA-DNA reassociation value – a historical definition of bacterial species.[14]

[14] Varghese et al., 2015; Konstantinidis and Tiedje, 2005

To determine suitable values for k (and to quantify the impact of genome relatedness on kgrep), I analyzed 500 publicly available *Escherichia* genomes for which pairwise ANI values are available.[15] The majority of *Escherichia* genome pairs fall in the 96–99% ANI range and constitute of various *Escherichia coli* strains (Figure 3.4).

[15] Ondov et al., 2016

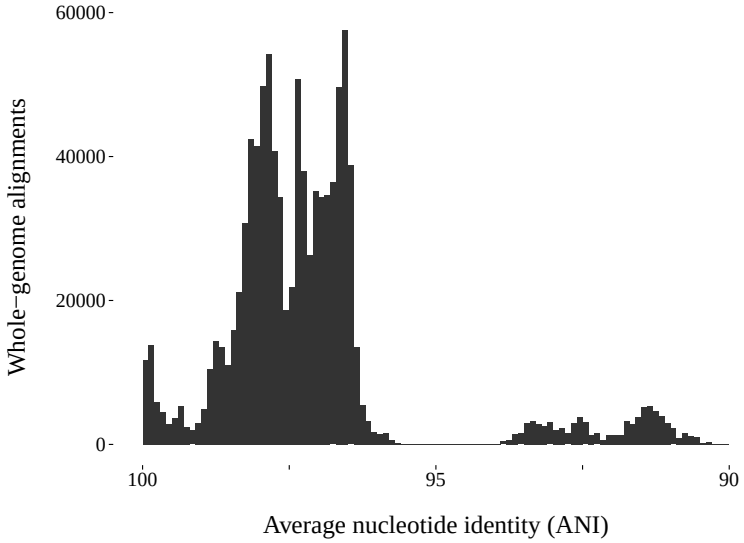


Figure 3.4: **Average nucleotide identity.** The ANI between 500 *Escherichia* genomes.

Let X and Y be two genomes. X should represent the single cell, *i.e.* the recruitment seed, and Y a genome present in the accompanying metagenome. I determine and store all constituent k-mers in X. Then, I simulate "reads" for Y by sliding a window of length 150 *bp* (a typical read length for Illumina data) across the genome sequence. I count how many of these reads KGREP would identify as metagenomic proxy reads within a given (estimated edit) distance $d \in \{0, 1, 2, 3\}$. The number of proxy reads divided by the total number of reads in Y is the cross-genome recruitment rate.

I calculated cross-genome recruitment rates for all 250,000 pairwise combinations in the collection of *Eschericia* genomes for $k \in \{11, 12, \ldots, 17\}$ (Figure 3.5). Increasing the k-mer size reduces the cross-recruitment rate rate for $k \leq 15$, further increasing k shows no effect. This suggests that KGREP should be run with $k \geq 15$ for $d = 1$ and 150 *bp* reads. I observed that the practical difference of choosing *e.g.* 15 or 17 as the k-mer size is negligibly small and empirically selected $k = 17$ as the default value in my implementation, giving good results for all SAG/metagenome combinations in my hands.

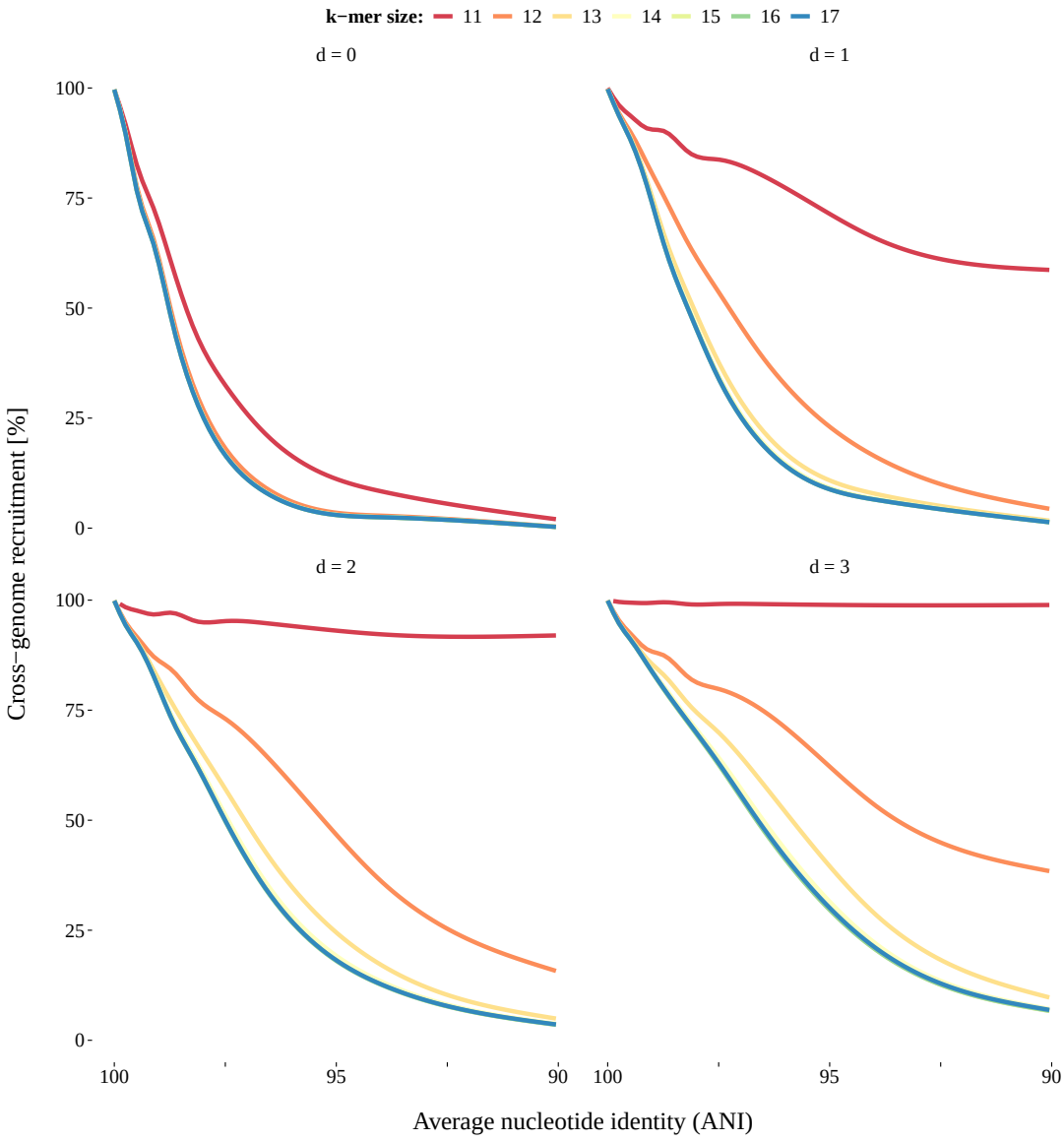


Figure 3.5: **Cross-genome recruitments.** Related to ANI.

3.3 *Precision and recall*

The first benchmark data are the eight *Escherichia coli* SAGs and their concominant *in vitro* mock metagenome from Chapter 2.[16] I estimated the relative abundance of *E. coli* to amount to 0.15%, corresponding to a mean per-base coverage of only 20.7× – too little coverage for a *de novo* proxy assembly, but enough reads to calculate KGREP's precision and recall on this data set:

[16] Clingenpeel et al., 2014a; Bowers et al., 2015

$$\begin{aligned} Precision &= \frac{tp}{tp + fp} \\ &= \% \text{ recruited reads that can be mapped} \end{aligned}$$

$$\begin{aligned} Recall &= \frac{tp}{tp + fn} \\ &= \% \text{ mappable reads that are recruited} \end{aligned}$$

I used Bowtie 2[17] to map all metagenome and the subsets of KGREP-recruited proxy reads against the *E. coli* reference genome and report mapping rates, precision, and recall in Table 3.1.

[17] Langmead and Salzberg, 2012

E. coli	GF [%]	Proxy reads	# mappable	Precision	Recall
SAG 0	96.41	553,838	536,312	0.968	0.662
SAG 1	99.37	575,114	554,551	0.964	0.685
SAG 2	96.88	552,732	529,328	0.958	0.653
SAG 3	92.13	523,956	503,076	0.960	0.621
SAG 4	87.91	528,404	482,075	0.912	0.595
SAG 6	100	599,842	560,217	0.934	0.692
SAG 7	100	602,900	561,757	0.932	0.694
SAG 8	98.25	585,386	548,667	0.937	0.677
Reference	100	523,438	523,116	0.999	0.646

Table 3.1: **Recruitment benchmark.** A total of 810,015 metagenomic reads align against the *E. coli* genome.

A total of 810,015 metagenomic reads can be aligned to the reference – slightly more than previously observed, because I align against one genome instead of all 26 genomes of the mock community –, I count these as condition positive reads. KGREP is very precise. Moderate recall values suggest that KGREP rejects (mappable) reads that contain too many errors.

3.4 *On-the-fly recruitment seed expansion*

A greedy extension of my recruitment strategy enables the reconstruction of genomic regions that otherwise would have been lost. Poorly amplified regions of the single cell introduce physical sequencing gaps, *i.e.* regions without SAG read coverage (Figure 3.6).[18] However, the per-genome coverage in a metagenome is (more or less) constant; the per-base metagenomic coverage depends only on the genome's abundance and the metagenomic sequencing depth.[19] Therefore, metagenomic reads most likely cover all positions of the target genome, also spanning regions missed by single cell sequencing.

[18] Bankevich et al., 2012; Bremges et al., 2016

[19] Wooley et al., 2010; Escobar-Zepeda et al., 2015

Figure 3.6: **Region without SAG read coverage.** Metagenomic proxy reads span regions missed by SAG sequencing.

If R is a metagenomic proxy read and contains m k-mers not present in G', then I add these m novel k-mers to the list of known k-mers from G'. In other words, if a metagenomic proxy read contains unobserved k-mers, potentially spanning (or reaching into) uncovered regions of the single cell, then I add these k-mers to my reference set (Figure 3.7).

Afterwards, I continue evaluating the next metagenomic read with the newly extended set of k-mers representing G' until I have processed all metagenomic reads.

Figure 3.7: **Greedy recruitment seed expansion.** Orange k-mers are added.

3.5 *Contaminating or decontaminating?*

Reagent and laboratory contamination is ubiquitous in laboratory reagents, *i.e.* commonly used DNA extraction kits.[20] Most enzymes are produced by living organisms and therefore include almost inevitably contaminant DNA from the producer. It has been shown that even commercially available MDA reagents frequently contain contaminant DNA, which then is co-amplified with the target DNA.[21] This contamination not only reduces the efficiency of sequencing microbial single cells, but also confounds the analysis of potentially unknown genomes.[22] Recently, automated methods to screen against contamination in genome assemblies became available, but it is still advised to avoid known sources of contamination whenever possible.[23]

[20] Laurence et al., 2014; Lusk, 2014; Salter et al., 2014

[21] Woyke et al., 2011

[22] Woyke et al., 2010; Blainey and Quake, 2011

[23] Lux et al., 2015; Tennessen et al., 2016

METAGENOMIC PROXY READS are immune to contamination exclusively affecting either single cells or the metagenome, *e.g.* contaminant MDA reagents introduce contamination to the SAG, but not the metagenome. Also laboratory contamination during single cell sorting is less of an issue.

However, if metagenomic and single cell DNA is treated with the same reagents, they share contaminant DNA (and therefore contaminant *k*-mers). I solve this problem by allowing to ignore certain *k*-mers known to originate from contaminant DNA to avoid spurious recruitments due to contaminant *k*-mer hits. A curated list of contaminant sequences is bundled with KGREP.

3.6 *Assembling metagenomic proxy reads*

Metagenomic proxy reads can be treated as if they originated from an isolate-grade genome. The metagenomic coverage of a genome is constant and chimeric reads are rare (because metagenomic DNA is usually not amplified with MDA).

I therefore assemble proxy reads with SPAdes[24] in its regular (multi-cell) mode. Afterwards, I map the original SAG reads to the assembled proxy contigs with Bowtie 2[25] and remove the few (and usually very short) contigs without any SAG hits.

[24] Bankevich et al., 2012

[25] Langmead and Salzberg, 2012

3.7 *Aminacenantes (OP8) single cells from Sakinaw Lake*

As a realistic benchmark, I use 24 *Aminacenantes* (OP8) single cells from Sakinaw Lake, for which a reference genome (co-assembled from these 24 SAGs and manually decontaminated) is available.[26] While this OP8 co-assembly was estimated to be 100% complete (based on a marker gene analysis), it is by no means a closed reference genome and some regions of the genome might be missing. SAG sequencing reads originating from these regions cannot be aligned to the reference genome and therefore appear contaminant. Consequently, metagenomic proxy reads recruited from these regions are also unmappable, as are the resulting contigs in the metagenomic proxy assembly.

[26] Rinke et al., 2013

Therefore, I first map all SAG reads to the available reference genome with Bowtie 2[27] and keep all reads that align. Table 3.2 gives the alignment rates for all 24 single cells. I use this curated set of SAG sequencing reads to further benchmark KGREP and metagenomic proxy assemblies.

[27] Langmead and Salzberg, 2012

SAKINAW LAKE in British Columbia, Canada, is a meromictic lake, *i.e.* it has layers of water that do not intermix, which became famous among microbiologists for its richness in candidate phyla.[28] Samples from 120*m* depth were used to generate SAGs and a corresponding deeply sequenced metagenome; the latter was sequenced on the Illumina HiSeq 2000 platform using 2×150 *bp* paired-end sequencing and generating a total of 386,581,812 reads (58 *Gbp*).[29]

[28] Gies et al., 2014; Nobu et al., 2016

[29] Rinke et al., 2013

Based on metagenome read mapping with Bowtie 2[30], I estimate the relative abundance of *Aminacenantes* (OP8) to account to 3.3%, corresponding to a mean per-base coverage of 656.5×.

[30] Langmead and Salzberg, 2012

3.8 *Aminacenantes (OP8) metagenomic proxy assemblies*

I use the filtered OP8 single cell sequencing reads as recruitment seeds to recruit metagenomic proxy reads with KGREP. In the following, *Proxy* denotes the proxy assembly of KGREP-recruited reads without the recruitment seed expansion; *Proxy** means the recruitment seed expansion was enabled. It makes sense to

SAG ID	# raw reads	# mapped reads	% mapped reads
31	21,417,846	16,238,861	75.82
33	17,032,640	15,445,876	90.68
36	10,764,770	10,335,857	96.02
38	23,885,748	22,832,580	95.59
40	14,706,772	10,245,506	69.67
47	16,909,846	13,785,069	81.52
48	17,586,704	16,758,449	95.29
49	25,726,550	22,868,807	88.89
52	9,320,936	8,077,828	86.66
54	24,744,270	21,851,004	88.31
57	24,243,792	17,293,311	71.33
58	11,250,846	10,727,597	95.35
59	22,732,658	21,178,380	93.16
64	10,937,730	10,186,356	93.13
67	25,259,458	21,063,146	83.39
83	17,227,512	15,682,729	91.03
84	10,421,578	9,504,274	91.20
88	3,160,992	2,191,597	69.33
93	26,014,290	12,263,311	47.14
119	35,809,646	29,699,887	82.94
125	36,529,520	26,228,359	71.80
128	8,735,682	7,920,678	90.67
134	8,910,092	2,867,125	32.18
137	6,439,768	2,891,745	44.90

Table 3.2: **SAG read preprocessing.** OP8 single cell sequencing reads mapped against the manually decontaminated co-assembly of 24 SAGs.

iteratively recruit proxy reads with on-the-fly seed expansion enabled; *Proxy*** therefore denotes a two-pass recruitment and *Proxy**** a three-pass recruitment and successive assembly.

I assemble all metagenomic proxy reads into metagenomic proxy assemblies with SPAdes[31] in its regular (multi-cell) mode and the filtered SAG reads with SPAdes in its single-cell mode.

[31] Bankevich et al., 2012

METAGENOMIC PROXY ASSEMBLIES in all flavours recover more of the genome (Figure 3.8), contain less errors (Figure 3.9), and are of higher contiguity (Figure 3.10) than SAG-only assemblies.

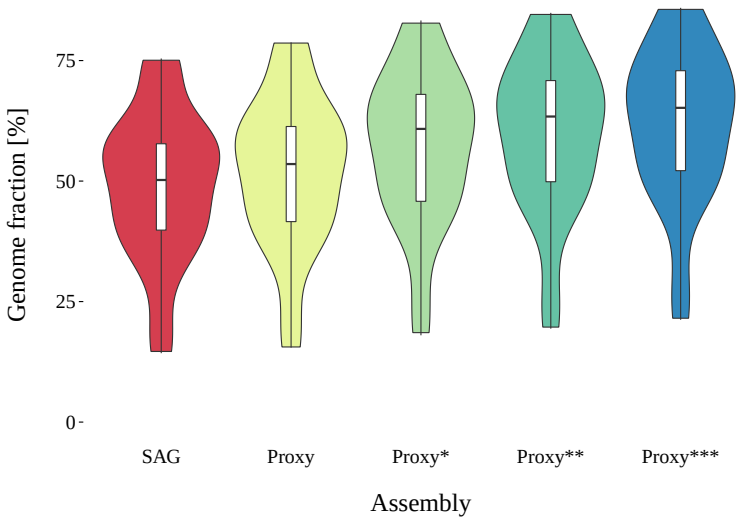


Figure 3.8: **Genome fraction.** Quality assessment with QUAST.

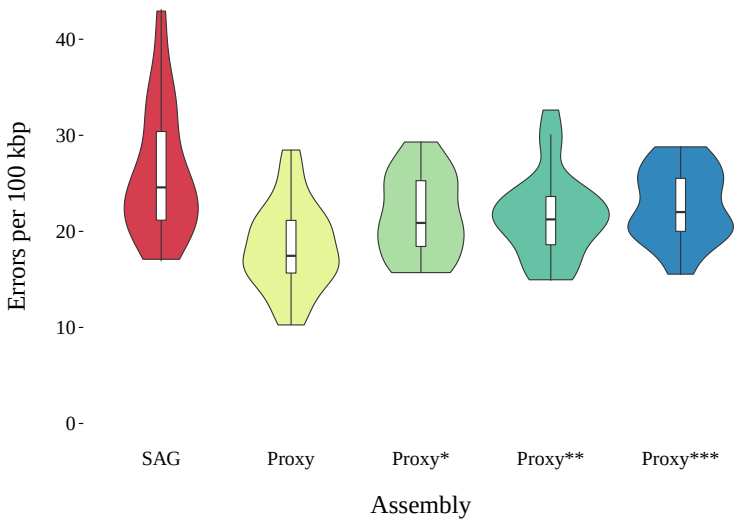


Figure 3.9: **Assembly errors.** Quality assessment with QUAST.

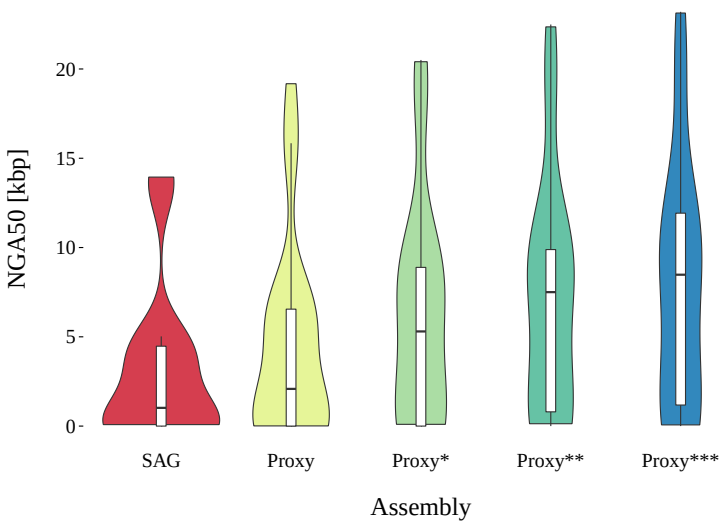


Figure 3.10: **Assembly contiguity.** Quality assessment with QUAST.

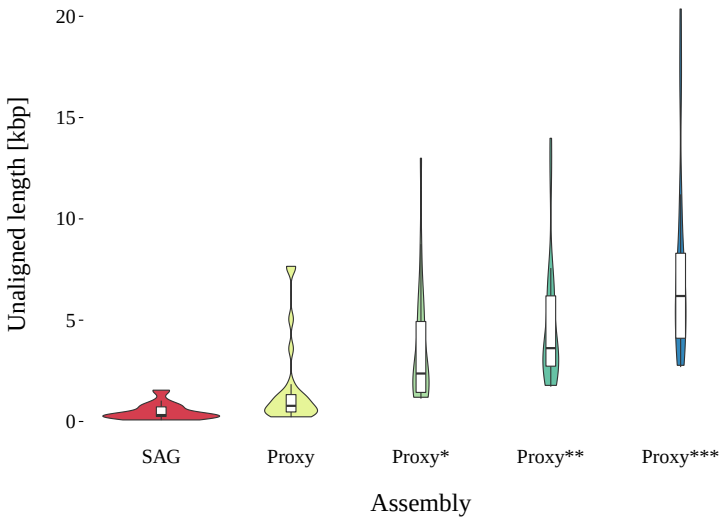


Figure 3.11: **Unaligned length.** Quality assessment with QUAST.

Iteratively extending the recruitment seed seems to further improve assembly results, while the percentage of unaligned contigs – possibly originating from cross-genome recruitments – remains negligibly small (Figure 3.11).

I CONCLUDE that metagenomic proxy assemblies are the better single cell assemblies for these 24 SAGs.

3.9 Conclusions

De novo SAG assemblies greatly suffer from uneven genome coverage and are therefore limited in their contiguity and accuracy. I propose to exploit shotgun metagenomic data to improve the quality of single cell genome assemblies and developed KGREP, a fast, *k*-mer based recruitment method to identify metagenomic proxy reads representing the single cell genome of interest.

I CIRCUMVENT MOST CHALLENGES of single cell assembly by assembling proxy reads instead of the single cell reads. Effectively, the assembly of metagenomic proxy reads enables me to span (or walk into) physical sequencing gaps of the single cell and to reconstruct a more complete representation of the genome. Metagenomic proxy assemblies therefore demonstrate yet another powerful combination of shotgun metagenome and single cell sequencing.

3.10 Implementation details

KGREP requires 2 bits per *k*-mer: one bit to encode the *k*-mer presence (or absence) in the SAG data; a second bit to flag known contaminant *k*-mers to ignore during recruitment. It stores all possible canonical *k*-mers in a bitset of 4^{k-1} byte length (Table 3.3).

Storing the *k*-mer occurences in a bitset (instead of *e.g.* a more space-efficient hash table) enables KGREP to process $\sim 250,000$ reads per second per core! To put this into perspective, KGREP identifies metagenomic proxy reads in the complete 58 *Gbp* Sakinaw Lake metagenome in half an hour.

k-mer size	(14)	15	16	17	18	(19)
Memory [MB]	(64)	256	1,024	4,096	16,384	(65,536)

Table 3.3: **Memory requirements.** KGREP stores a *k*-mer in 2 bits.

When paired-end sequencing data is available, requiring *k*-mer hits in both mates obviously further reduces the cross-genome recruitment rate (and is enabled by default).

3.11 Software availability

KGREP is implemented in C and is freely available under the open-source MIT license at:

```
https://github.com/abremges/kgrep
```

4 *An integrated assembly pipeline*

I presented two approaches that exploit shotgun metagenome data to improve the quality of single cell assemblies: MeCorS[1] and kgrep.[2] However, it is not obvious from the start which strategy works best for new SAG/metagenome combinations. This has been shown to apply for genome assembly in general.[3]

For *e.g.* the biogas microbiome – my research focus in Part II of this thesis – we already have shotgun metagenome data available and wait for 96 single cells to be generated eventually. I cannot predict which approach produces the best results for these – and, in fact, performance might vary by SAG – and therefore suggest to try all.

To facilitate the assembly of a large number of SAGs and their accompanying metagenomes, I implemented YinYang, an integrated and automated assembly pipeline for single cell genomes. Given a list of SAG and metagenome FASTQ files, YinYang produces an array of assemblies for each single cell using SPAdes[4]: (1) a SAG-only assembly, (2) a MeCors-corrected SAG assembly, and (3) metagenomic proxy assemblies via kgrep (Figure 4.1).

To estimate the inclusivity of each assembly, YinYang maps the SAG reads to the assembled contigs with Bowtie 2[5] and SAMtools[6] and reports the overall mapping rate. Optionally, YinYang also runs QUAST[7] to determine basic assembly statistics (*e.g.* the N50 value) and CheckM[8] to estimate genome completeness and possible contamination. All reports are collected in one place for the user to decide which assembly to pick.

[1] Bremges et al., 2016

[2] Bremges et al., in prep.

[3] Earl et al., 2011; Salzberg et al., 2012; Magoc et al., 2013; Bradnam et al., 2013

[4] Bankevich et al., 2012

[5] Langmead and Salzberg, 2012

[6] Li et al., 2009

[7] Gurevich et al., 2013

[8] Parks et al., 2015

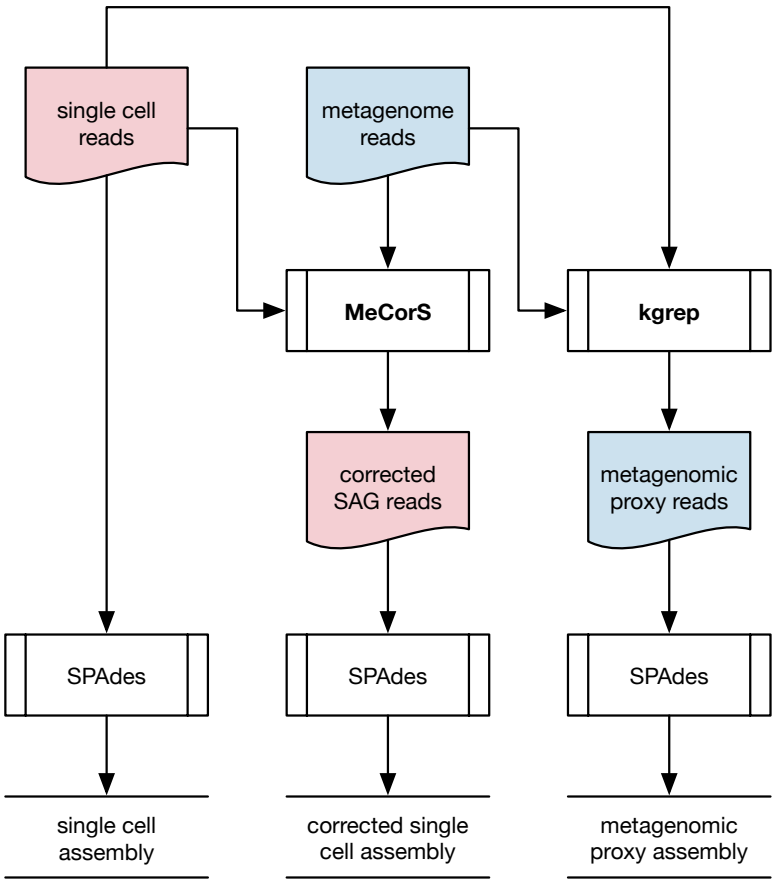


Figure 4.1: **Integrated assembly pipeline.** YinYang is my pipeline to assemble single cell genomes.

On my wish list for YinYang in future versions is the inclusion of a fully automated tool for the decontamination of genome assemblies. ACDC[9] and ProDeGe[10] are the most promising candidates that I will investigate.

[9] Lux et al., 2015
[10] Tennessen et al., 2016

Software availability

YinYang is implemented in Perl and is freely available under the open-source MIT license at:

```
https://github.com/abremges/yinyang
```

Part II

GENOMES FROM METAGENOMES

5 *Metagenome assembly and binning techniques*

If single cell genomics is *Yin*, then metagenomics is *Yang* – they are complementary (rather than opposing) approaches to study the microbial dark matter.[1] In Part II of my thesis, I focus on computational metagenomics.

[1] Ohsawa, 1931

Shotgun metagenomics is a method of choice to analyze the coding potential of whole microbial communities.[2] Untangling individual genomes from metagenomes requires (1) the assembly of metagenome sequencing reads into contigs and (2) the successive grouping of these contigs into genome bins (Figure 5.1).

[2] Sharon and Banfield, 2013

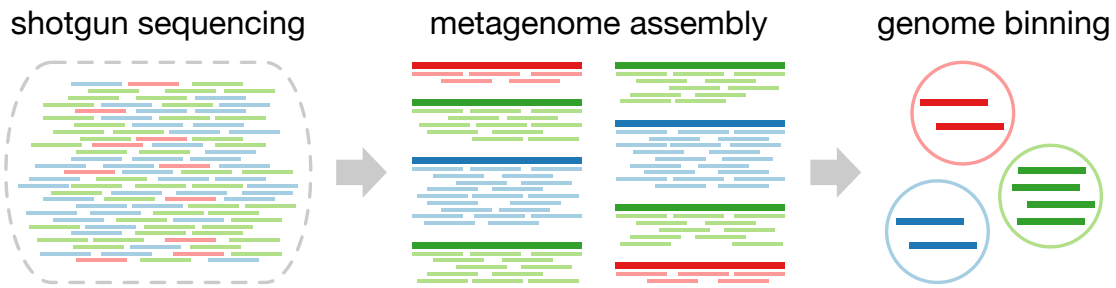


Figure 5.1: **Assembly and binning.** Key challenges in computational metagenomics.

Metagenome assembly is computational challenging because (1) metagenomic datasets are huge and approach terabytes in size, (2) read coverage of different organisms in the environmental sample is non-uniform, and (3) cross-genome repeats (*e.g.* rRNA genes) are longer than typical read lengths and therefore impossible to fully resolve.[3]

[3] Kunin et al., 2008; Hess et al., 2011; Nagarajan and Pop, 2013

Metagenome assemblers address these challenges by *e.g.* massively distributing computation[4], using Bloom filters or succint de Bruijn graphs (to reduce the memory footprint)[5], or adopting ideas proven useful in the assembly of single cells and highly polymorphic diploid genomes.[6]

[4] Boisvert et al., 2012

[5] Chikhi and Rizk, 2012; Li et al., 2015

[6] Prjibelski et al., 2014; Safonova et al., 2015; Nurk et al., 2016

METAGENOME BINNING is the post-assembly taxonomic assignment of contigs into genome bins that enables the study of individual organisms (and their interactions), directly from deeply sequenced metagenomes. Therefore, the task of a binning tool is to assign an identifier to every assembled contig, with each identifier ideally representing a single genome.[7]

Taxonomic binning tools, such as Megan[8] or CARMA[9], act as classifiers and label contigs with taxa from an existing taxonomy, such as the NCBI Taxonomy database.[10]

Unsupervised and reference-free binning tools traditionally use nucleotide composition (in particular tetranucleotide frequencies) to group contigs with similar usage, thus effectively differentiating between contigs of different species.[11] Today, binning tools increasingly leverage additional information to improve genome recovery – even in the presence of multiple genomes from individual species in a sample –, such as paired-end read linkage[12], mean contig coverage[13], per-sample (differential) coverage[14], or combinations thereof.[15]

[7] McHardy and Rigoutsos, 2007

[8] Huson et al., 2007

[9] Krause et al., 2008; Gerlach et al., 2009

[10] Federhen, 2012

[11] Teeling et al., 2004; Dick et al., 2009

[12] Iverson et al., 2012

[13] Wu et al., 2014

[14] Albertsen et al., 2013; Imelfort et al., 2014

[15] Alneberg et al., 2014; Kang et al., 2015

NEAR-COMPLETE GENOME BINS can often be recovered and subsequently mined for their metabolic potential.[16] Nevertheless, all assembly and binning results – even if the presumably "best" tool was chosen – should be inspected carefully by *e.g.* looking at taxonomic assignments of individual contigs, visualizing the underlying differential coverage information, or using an automated method for assessing the quality of metagenome-derived microbial genomes.[17]

[16] Campanaro et al., 2016; Stolze et al., 2016

[17] Albertsen et al., 2013; Parks et al., 2015; Eren et al., 2015

6 *Assembling a biogas-producing community*

Biogas is regarded a clean, renewable, and environmentally compatible energy source.[1] Moreover, the generation of energy from biogas relies on a balanced carbon dioxide cycle. In Germany, there are close to 9,000 biogas plants (BGPs) with a combined electric capacity of over 4,000 MW and a gross electricity production of over 30 TWh per year. They can supply more than nine million households with biogas-based electricity.[2]

[1] Weiland, 2010

[2] German Biogas Association; http://www.biogas.org

The process of biogas production takes place under anaerobic conditions and involves microbial decomposition of organic matter, yielding methane as the main final product of the fermentation process (Figure 6.1). Complex consortia of microorganisms are responsible for biomass decomposition and biogas production.[3] The majority of the participating microbes are still unknown, as is their influence on reactor performance.[4] Since most of the organisms within biogas communities are non-cultivable by today's conventional microbiological techniques, shotgun metagenome sequencing currently is the method of choice to obtain unbiased insights into community composition and the metabolic potential of key community members.

[3] Schlüter et al., 2008; Maus et al., 2016b

[4] Wirth et al., 2012

Here, I describe the first deeply sequenced metagenome and metatranscriptome of an agricultural production-scale biogas plant on the Illumina platform.[5] I assembled the metagenome and *e.g.* reconstructed most genes involved in the methane metabolism, a key pathway involving methanogenesis performed by methanogenic Archaea.

[5] Bremges et al., 2015

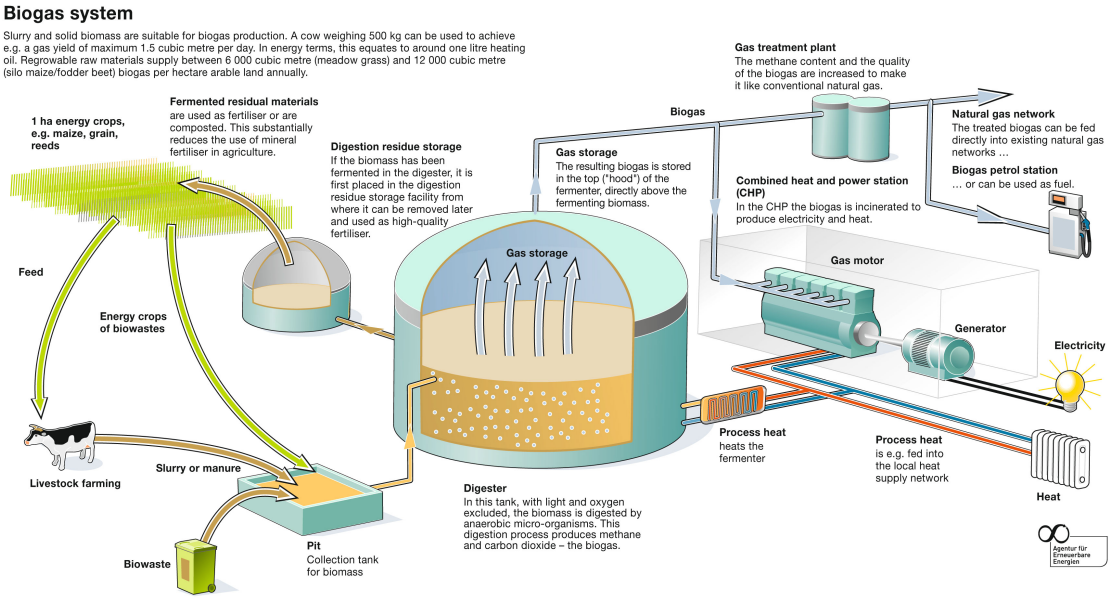


Figure 6.1: **Biogas production.** Overview of the biogas system (anaerobic digestions). Figure courtesy of Renewable Energies Agency, Germany.

6.1 *Digester management and process characterization*

The biogas plant, located in North Rhine Westphalia, Germany, features a mesophilic continuous wet fermentation technology characterized recently.[6] It was designed for a capacity of 537 kW_{el} combined heat and power (CHP) generation. The process comprises three digesters: a primary and secondary digester, where the main proportion of biogas is produced, and a storage tank, where the digestate is fermented thereafter.

THE PRIMARY DIGESTER is fed hourly with a mixture of 72% maize silage and 28% liquid pig manure. The biogas and methane yields at the time of sampling were at 810.5 and 417.8 liters per kg organic dry matter ($l/kg\ oDM$), respectively. After a theoretical retention time of 55 days, the digestate is stored in the closed, non-heated final storage tank. Further metadata are summarized in Table 6.1.

[6] Stolze et al., 2015

Process parameter	Sample
Net volume	$2,041\ m^3$
Dimensions	$6.4\ m$ high, diameter of $21\ m$
Electrical capacity	$537\ kW_{el}$
pH	7.83
Temperature	40 °C
Conductivity	$22.10\ mS/cm$
Volative organic acids (VOA)	$5,327\ mg/l$
Total inorganic carbon (TIC)	$14,397\ mg/l$
VOA/TIC	0.37
Ammoniacal nitrogen	$2.93\ g/l$
Acetic acid	$863\ mg/l$
Propionic acid	$76\ mg/l$
Fed substrates	72% maize silage, 28% pig manure
Organic load	$4.0\ kg\ oDM\ m^{-3}\ d^{-1}$
Retention time	$55\ d$
Biogas yield	$810.5\ l/kg\ oDM$
Methane yield	$417.8\ l/kg\ oDM$

Table 6.1: **Characteristics of the BGP.** Primary digester, sampled on Nov 15, 2010.

6.2 *Sampling and library construction*

Samples from the primary digester of the aforementioned biogas plant were taken in November 2010. Prior to the sampling process, approximately 15 l of the fermenter substrate were discarded before aliquots of 1 l were transferred into clean gastight sampling vessels and transported directly to the laboratory.

For the metagenome, aliquots of 20 g of the fermentation sample were used for total community DNA preparation as described previously.[7]

[7] Schlüter et al., 2008

For the metatranscriptome, a random-primed cDNA library was prepared. Total RNA was first treated with 5′-P dependent Terminator exonuclease to enrich for full-length mRNA carrying 5′ CAP or triphosphate structures. Then, first-strand cDNA was synthesized using a N6 random primer and M-MLV-RNase H reverse transcriptase, and second-strand cDNA synthesis was performed according to the Gubler-Hoffman protocol.[8]

[8] Gubler and Hoffman, 1983

6.3 Metagenomic and metatranscriptomic sequencing

We sequenced one metatranscriptome and two metagenome shotgun libraries on Illumina's Genome Analyzer IIx system, applying the Paired-End DNA Sample Preparation Kit (Illumina Inc.) as described by the manufacturer and generating 2×161 *bp* paired-end reads. On Illumina's MiSeq system, we sequenced three further metagenome shotgun libraries, applying the Nextera DNA Sample Preparation Kit (Illumina Inc.) as described by the manufacturer and generating 2×155 *bp* paired-end reads. Our sequencing efforts, yielding 35 *Gbp* in total, are summarized in Table 6.2.

Library name	Library type	Insert size[1]	Cycles	Reads	Bases
GAIIx, Lane 6	RNA, TruSeq	202 ± 49	2×161	78,752,308	12,679,121,588
GAIIx, Lane 7	DNA, TruSeq	157 ± 19	2×161	54,630,090	8,795,444,490
GAIIx, Lane 8	DNA, TruSeq	298 ± 32	2×161	74,547,252	12,002,107,572
MiSeq, Run A1[2]	DNA, Nextera	173 ± 53	2×155	4,915,698	761,933,190
MiSeq, Run A2[2]	DNA, Nextera[3]	522 ± 88	2×155	1,927,244	298,722,820
MiSeq, Run B1[2]	DNA, Nextera	249 ± 30	2×155	3,840,850	573,901,713
MiSeq, Run B2[2]	DNA, Nextera[3]	525 ± 90	2×155	4,114,304	614,787,564

[1]Insert sizes determined with Picard tools. [2]Partial runs. [3]This Nextera library was sequenced twice.

Table 6.2: **Sequencing statistics.** Overview of the different sequencing libraries.

6.4 Read quality control

Prior to assembly, I used Trimmomatic (Bolger et al., 2014) for adapter removal and moderate quality trimming. After adapter clipping, using Trimmomatic's *Truseq2-PE* and *Nextera-PE* templates, I removed leading and trailing ambiguous or low quality bases (below Phred quality scores of 3). Table 6.3 summarizes the effect on sequencing depth, more than 25 *Gbp* of sequence data passed quality control.

Library type	Reads, raw	post-QC	Bases, raw	post-QC
Metagenome (total)	143,975,438	137,365,053	23,046,897,349	17,267,320,221
Metatranscriptome	78,752,308	73,165,986	12,679,121,588	8,455,809,264

Table 6.3: **Quality control.** Adapter removal and quality trimming.

6.5 *Metagenome assembly and annotation*

I assembled the metagenome with IDBA-UD[9], MEGAHIT[10], and Ray Meta[11], trying a range of *k*-mer sizes from 21 to 61 in steps of 10 for the latter. To estimate the inclusivity of the set of assemblies, I aligned the post-QC sequencing reads to the assembled contigs with Bowtie 2[12] and used SAMtools[13] to convert SAM to BAM, sort the alignment file, and calculate the mapping statistics.

[9] Peng et al., 2012
[10] Li et al., 2015
[11] Boisvert et al., 2012
[12] Langmead and Salzberg, 2012
[13] Li et al., 2009

Based on total assembly size, contiguity, and the percentage of mapped back metagenomic reads, we selected the Ray Meta assembly produced with a *k*-mer size of 31. Here, we assembled approximately 228 *Mbp* in 54,489 contigs greater than 1,000 *bp*, with an N50 value of 9,796 *bp*.[14] 77% (79%) of metagenomic (metatranscriptomic) reads mapped back to this assembly.

[14] It's over 9000!

I used MetaProdigal[15] to predict 250,596 protein-coding genes on the assembled contigs. I blasted the protein sequences of all predicted genes against the KEGG database[16], release 72.0, using Protein-Protein BLAST.[17] Of the 250,596 predicted genes, 191,766 (76.5%) had a match in the KEGG database, using an Evalue cutoff of 10^{-6}. I determined the KEGG Orthology (KO) for each gene by mapping the top-scoring BLAST hit to its orthologous gene in KEGG, resulting in 109,501 genes with an assigned KEGG Orthology. Table 6.4 summarizes these results.

[15] Hyatt et al., 2012
[16] Kanehisa et al., 2014, 2016a
[17] Camacho et al., 2009

Assembly metric	Our assembly
Total size	228,382,457 *bp*
Number of contigs	54,489
N50 value	9,796 *bp*
Largest contig	333,979 *bp*
Mapped DNA reads	105,461,596 (77%)
Mapped RNA reads	57,436,058 (79%)
Predicted genes	250,596
of these, full-length	172,372 (69%)
Match in KEGG Genes	191,766
of these, assigned KO	109,501
of these, in KEGG pathways	61,100

Table 6.4: **Assembly and annotation results.** Minimum contig size of 1,000 *bp*.

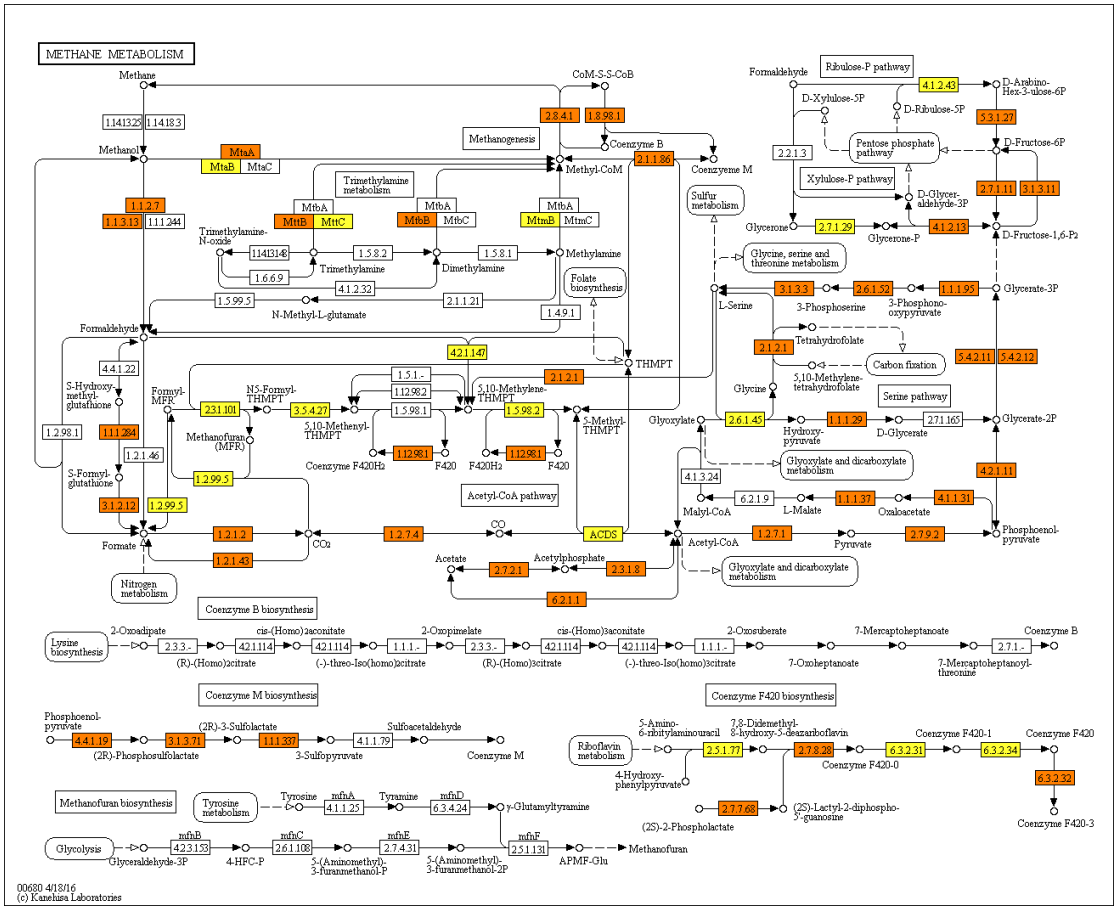


Figure 6.2: **Methane metabolism pathway.** Genes involved in the methane metabolism highlighted: Genes with only metagenomic support in yellow, genes with also metatranscriptomic support in orange.

6.6 *Relating the metagenome and the metatranscriptome*

To illustrate potential use cases, I first counted the number of reads within genes using BEDTools (Quinlan and Hall, 2010) and highlighted metagenomic and metatranscriptomic coverage of the methane metabolism pathway in Figure 6.2. The assembly therefore containes the majority of genes involved in the methane metabolism from our metagenomic data, with accompanying metatranscriptomic data suggesting active gene expression for many.

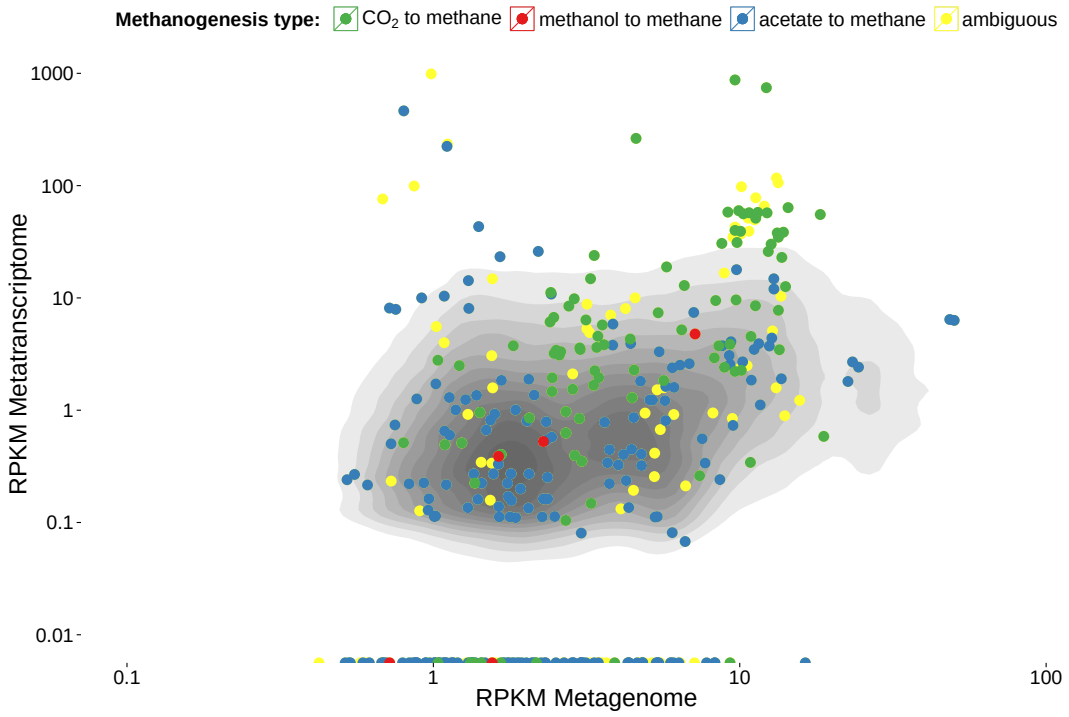


Figure 6.3: **Relating the metagenome and metatranscriptome.** Highlighted are genes involved in methanogenesis; in the background a two-dimensional density estimation for all 250,596 genes.

For a second example, I calculated the reads per kilobase per million mapped reads (RPKM) for each gene as a crude measure for abundance (metagenome) or expression (metatranscriptome). Figure 6.3 relates the two; I accentuated all genes assigned to either of the three known types of methanogenic pathways: CO_2 to methane (96 genes), methanol to methane (5 genes), and acetate to methane (209 genes). 80 common genes are shared between pathway types.

Hydrogenotrophic methanogenesis, *i.e.* the reduction of CO_2 with hydrogen, appears to be highly expressed in the reactor analyzed, which is in agreement with results obtained via metatranscriptome sequencing.[18]

[18] Zakrzewski et al., 2012

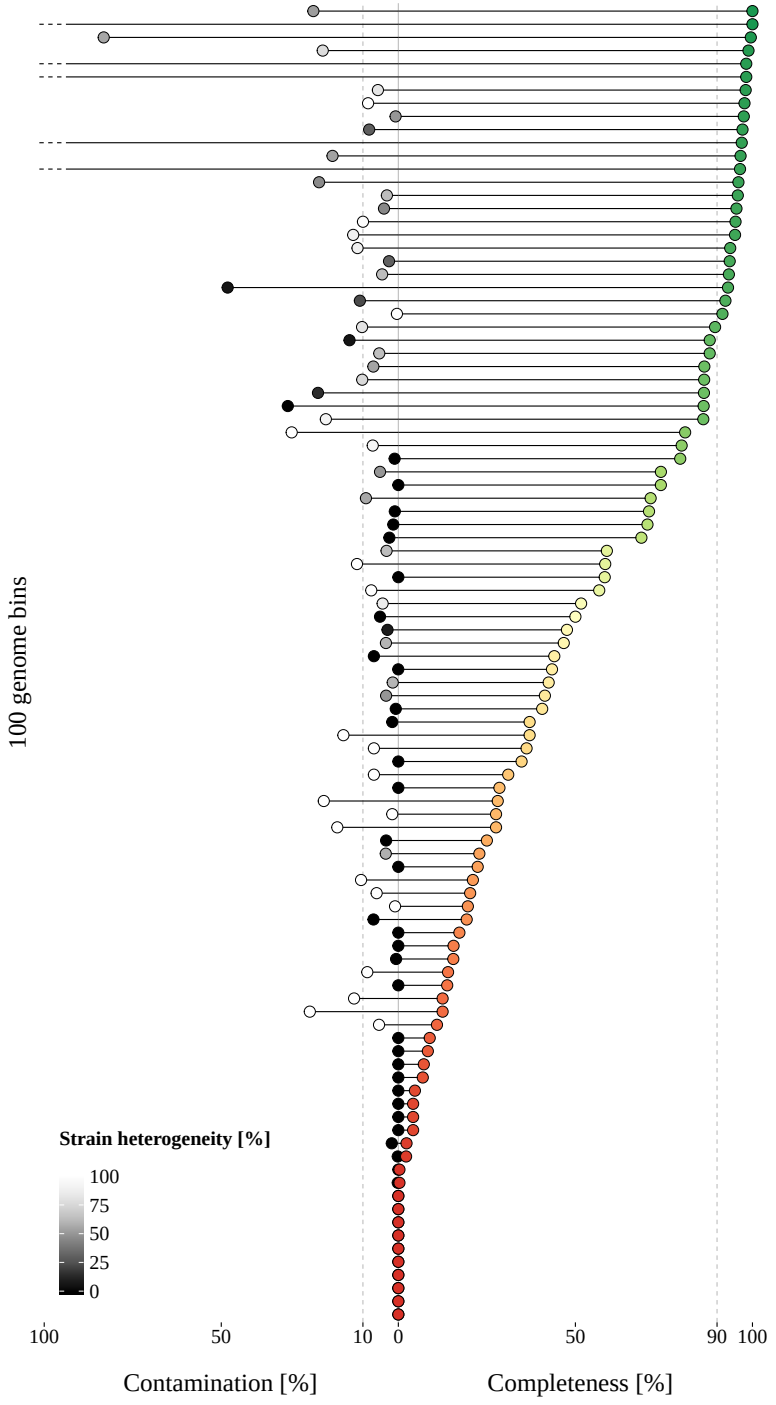


Figure 6.4: **Binning of the metagenome.** One hundred genome bins were generated, but most are either incomplete or contaminated (or both).

6.7 *In retrospect: Metagenome binning*

Extending Bremges et al., 2015, I tried to recover genome bins from our assembly using MetaBAT.[19] MetaBAT is an unsupervised binning tool that always leverages tetranucleotide frequencies and paired-end linkage to group contigs into genome bins. If multiple samples are available, it additionally uses per-sample (differential) coverage information. If only one sample is available – as it is the case here –, it resorts to the mean contig coverage instead.

[19] Kang et al., 2015

One hundred genome bins were generated and I assessed their quality with CheckM.[20] CheckM estimates genome completeness, contamination, and strain heterogeneity.

[20] Parks et al., 2015

First, CheckM places the genome bin onto a fixed phylogenetic reference tree with pplacer[21]) to determine the most likely clade it originates from. Then, it used profile hidden Markov models[22] to search for the clade-specific marker genes and counts their presence (or absence) in the genome bin.

[21] Matsen et al., 2010

[22] Eddy, 1998, 2008

If there are *e.g.* 100 marker genes for a certain clade and the genome bin contains 69 of them, then its estimated genome completeness is 69%. Furthermore, if 42 of the marker genes occur more than once, the bin's estimated contamination is 42%. Pairs of multi-copy marker genes with an average amino acid identity $AAI \geq 90\%$ count towards strain heterogeneity (assuming that genes from different strains are very similar). In other words, if a genome bin appears highly contaminated but has a strain heterogeneity value of 100%, then the entire contamination can be explained by having multiple strain of the same species in one genome bin.

I present each genome bin's completeness, contamination, and strain heterogeneity as a Cleveland dot plot in Figure 6.4. As a rule of thumb, we aim for $\geq$ 90% completeness and $\leq$ 10% contamination.[23] This leaves us with only ten genome bins – we surely can do better (as shown in Chapter 7). Genome binning works more reliable when multiple (related) samples are available that contribute valuable differential coverage information.[24]

[23] Parks et al., 2015; Eren et al., 2015

[24] Turaev and Rattei, 2016

6.8 Conclusions

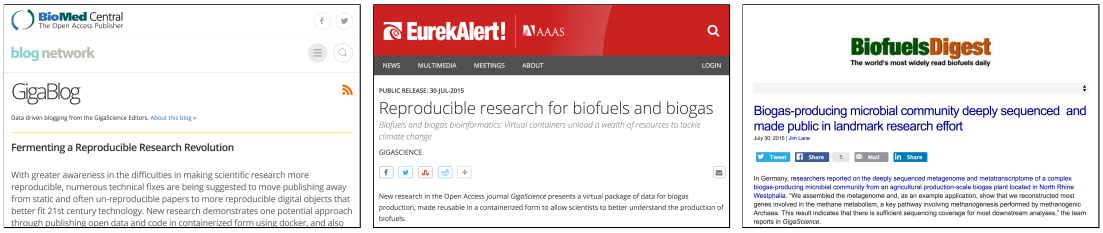
At the time of publication, the sequencing depth was unprecedented for any microbial community from a production-scale biogas plant. We sequenced the metagenome 27× and 19× deeper, respectively, than previous studies applying 454 or SOLiD sequencing (and primarily focusing on community composition).[25] Metatranscriptomic sequencing of total community RNA, 230× deeper than previously reported, complemented our metagenome.[26] We therefore anticipated that the data were of great interest to the biogas research community in general and microbiologists working on biogas-producing microbial communities in particular – even without genome bins.

[25] Jaenicke et al., 2011; Wirth et al., 2012

[26] Zakrzewski et al., 2012

The metagenome assembly has since been used in one applied study to improve the characterization of a metaproteome generated from biogas plant fermentation samples and to investigate the metabolic activity of the microbial community.[27]

[27] Kohrs et al., 2015



(a) ***GigaScience* blog entry.** July 30, 2015, Scott Edmunds.

(b) ***EurekAlert!* press release.** July 30, 2015, *GigaScience*.

(c) **Article in *BiofuelsDigest*.** July 30, 2015, Jim Lane.

Figure 6.5: **Reactions.** Making our analyses reproducible paid off.

What took me by surprise was the press and social media coverage that Bremges et al., 2015 received, largely triggered by the fact that we dockerized all data and analyses, and therefore made our research more accessible and reproducible. The Docker container accompanying our manuscript is available at:

```
https://registry.hub.docker.com/u/metagenomics/2015-biogas-cebitec
```

7 *A genome catalog of the biogas microbiome*

Most members of the biogas microbiome belong to microbial dark matter, *i.e.* they are non-cultivable by today's conventional microbiological techniques. In a pilot study, we sequenced and assembled the metagenome of a biogas-producing microbial community from a production-scale biogas plant, but eventually failed to generate (more than a few) high-quality genome bins.[1]

[1] Bremges et al., 2015

Untangling dozens of near-complete genomes from biogas metagenomes was my dream, turned into reality in the context of our Community Science Program "Biogas-producing microbial communities" at the DOE Joint Genome Institute.[2]

[2] FY 2013; PI: Alex Sczyrba

Extensive metagenome sequencing of four production-scale biogas plants, greatly surpassing our previous efforts, enabled a more inclusive assembly of the biogas microbiome. Successive binning of assembled contigs recovered hundreds of near-complete genomes for process-relevant community members, also comprising the prevalent distinctive phyla Cloacimonetes, Spirochaetes, Fusobacteria, and Thermotogae.[3]

[3] Stolze et al., 2016

Yvonne Stolze and I contributed equally and we both wish to include this joint project in our doctoral theses. Therefore – and to avoid too many overlaps with regards to content –, I primarily describe the bioinformatics part: metagenome assembly and binning.

For laboratory details and the binning-enabled insights into the biology of distinct abundant taxa, please refer to Yvonne's forthcoming thesis or our co-first-authored manuscript in the journal *Biotechnology for Biofuels*.

7.1 *Metagenome sampling and sequencing*

The production of biogas happens usually at mesophilic (34–40 $^{\circ}C$) or thermophilic (55–60 $^{\circ}C$) conditions.[4] Differences in the biogas microbiome have been observed depending on the process conditions, *e.g.* temperature.[5]

[4] Weiland, 2010

[5] Ritari et al., 2012; Ziembińska-Buczyńska et al., 2014

Accounting for availability and accessibility, we selected three mesophilic and one thermophilic industrial biogas plants (BGPs) for metagenomic sequencing. For each BGP, total community DNA was extracted and sequenced in replicates at the DOE Joint Genome Institute. Table 7.1 lists our sequencing efforts, a total of 2.3 billion reads (347.5 *Gbp*) were sequenced.

BGP	# raw reads	# raw bases	# QC'ed reads	# QC'ed bases
1.1	267,749,142	40,162,371,300	256,033,246	38,404,986,900
1.2	289,930,844	43,489,626,600	276,028,796	41,404,319,400
2.1	298,185,500	44,727,825,000	283,504,064	42,525,609,600
2.2	281,693,590	42,254,038,500	277,123,112	41,568,466,800
3.1	242,121,112	36,318,166,800	208,532,304	31,279,845,600
3.2	338,184,952	50,727,742,800	326,116,028	48,917,404,200
4.1	307,971,670	46,195,750,500	288,040,900	43,206,135,000
4.2	290,604,188	43,590,628,200	271,494,384	40,724,157,600
Total	2,316,440,998	347,466,149,700	2,186,872,834	328,030,925,100

Table 7.1: **Sequencing statistics.** We sequenced the metagenomes from four BGPs in replicates.

7.2 *Community structure and similarity*

Metagenome binning techniques are most effective when mutliple samples are available, *e.g.* a time-series or employing different DNA extraction methods, for which a combined metagenome assembly makes sense.[6]

[6] Albertsen et al., 2013; Alneberg et al., 2014; Turaev and Rattei, 2016

I used Mash[7] to quantify the pairwise similarity of the biogas metagenomes prior to assembly. Mash reduces large sequence sets to compressed sketches using the MinHash algorithm.[8] Using these sketches, Mash rapidly estimates pairwise distances between two sets. Table 7.2 gives the pairwise Mash distances.

[7] Ondov et al., 2016

[8] Broder, 1997

BGP	1.1	1.2	2.1	2.2	3.1	3.2	4.1	4.2
1.1	0	0.028	0.062	0.061	0.067	0.067	0.094	0.08
1.2	0.028	0	0.063	0.063	0.067	0.067	0.093	0.08
2.1	0.062	0.063	0	0.024	0.038	0.038	0.099	0.08
2.2	0.061	0.063	0.024	0	0.038	0.039	0.100	0.081
3.1	0.067	0.067	0.038	0.038	0	0.024	0.095	0.077
3.2	0.067	0.067	0.038	0.039	0.024	0	0.095	0.079
4.1	0.094	0.093	0.099	0.100	0.095	0.095	0	0.028
4.2	0.08	0.08	0.08	0.081	0.077	0.079	0.028	0

Table 7.2: **Mash distances.** Pairwise comparisons of 4×2 metagenome read sets using Mash with a *k*-mer size of 21 and a sketch size of 1,000,000.

Pairwise Mash distances are small, indicating a high fraction of similar genomes across samples. Replicates are most similar, closely followed by a high similarity between the metagenome samples of BGP2 and BGP3. BGP4 is the outlier, which I expected because it is the only thermophilic BGP we sampled. Nevertheless, the Mash results support the combined assembly of all samples to facilitate the reconstruction of low-abundance community members and downstream genome binning.

7.3 *Combined metagenome assembly*

I used Ray Meta[9] to co-assemble all metagenome reads, using a *k*-mer size of 31. The gene prediction tool Prodigal[10] was used to predict genes on assembled contigs (Table 7.3).

[9] Boisvert et al., 2012

[10] Hyatt et al., 2012

Total bases	# contigs	N50	Largest contig	# genes
1,488,298,777	330,955	10,556	668,635	1,591,820

Table 7.3: **Assembly results.** Minimum contig size of 1,000 *bp*.

I aligned all metagenome reads to the assembled contigs with Bowtie 2 (Langmead and Salzberg, 2012) and calculated mapping statistics with SAMtools (Li et al., 2009; Table 7.4). Figure 7.1 and 7.2 visualize differential contig coverages between BGPs.

% reads of BGP1	% reads of BGP2	% reads of BGP3	% reads of BGP4
74.83 | 75.14	78.07 | 78.34	81.11 | 81.29	86.53 | 86.50

Table 7.4: **Mapping results.** 80.3% of all reads are included.

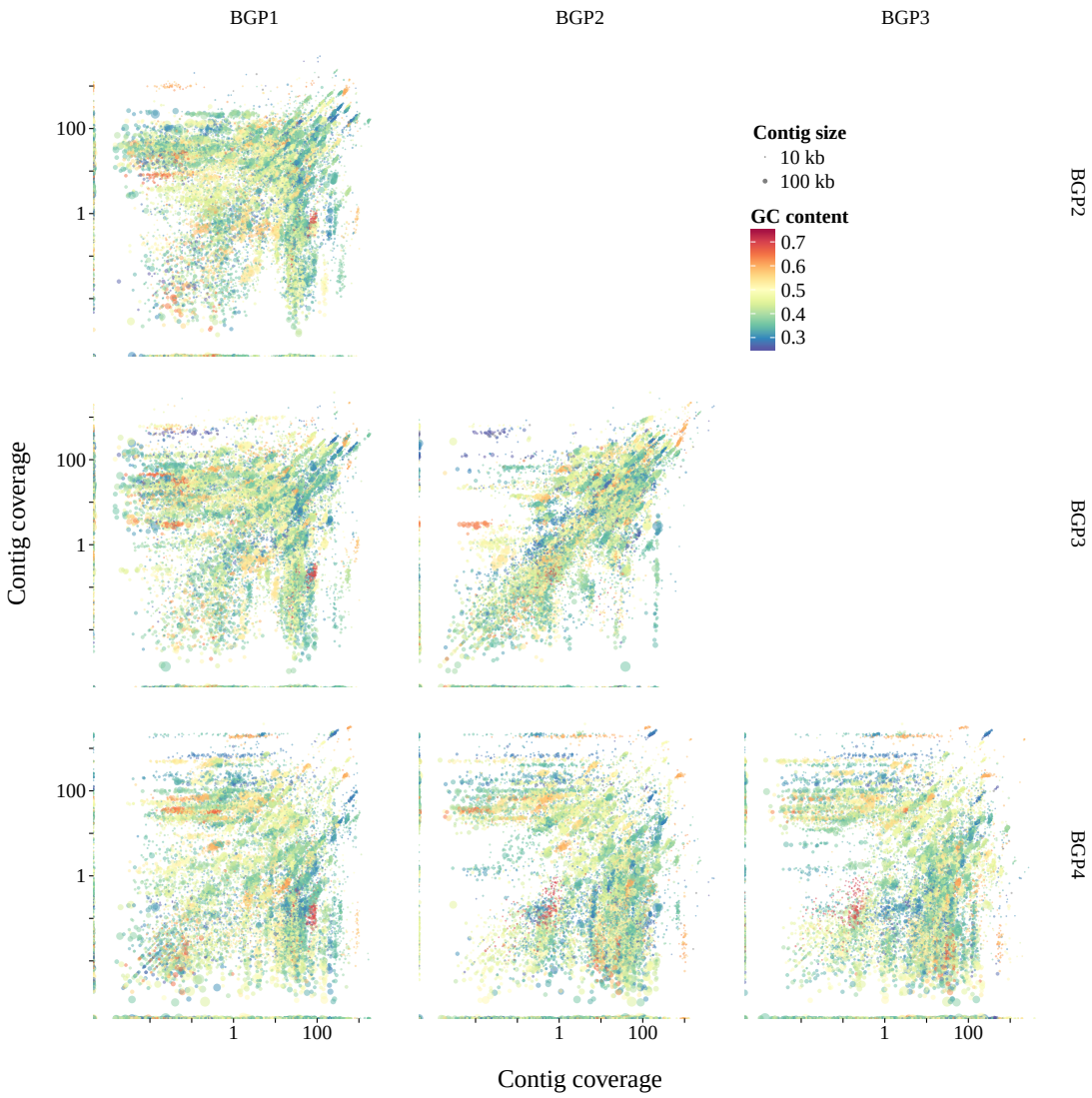


Figure 7.1: **Contig coverage.** Each point is one contig, colored by its GC content.

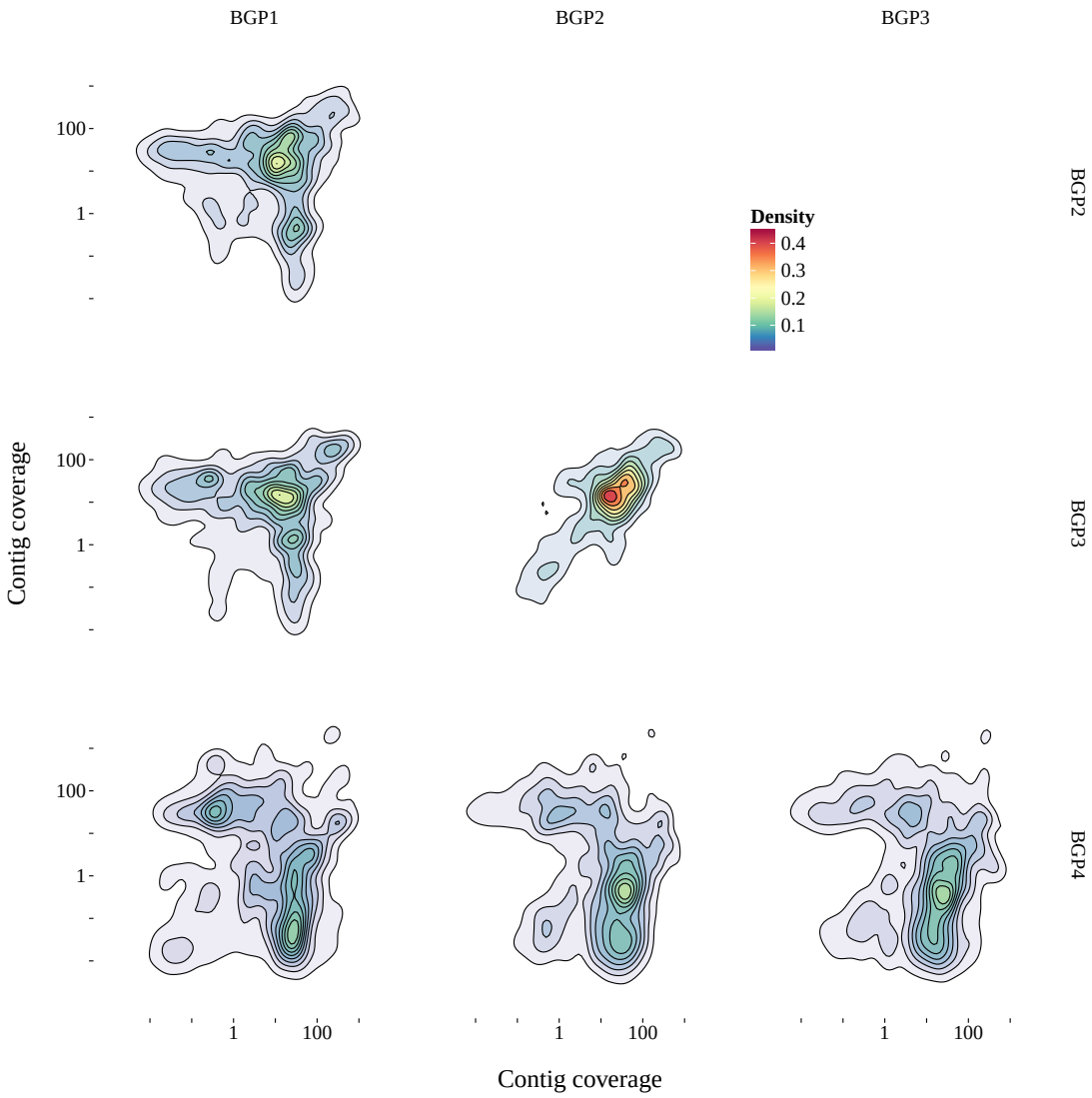


Figure 7.2: **Contig coverage.** Length-weighted 2D density estimation.

7.4 *Metagenome binning*

At the time of analyses, three unsupervised binning tools that leverage differential coverage information became available: GroopM[11], CONCOCT[12], and MetaBat.[13] I ran all tools on our data:

[11] Imelfort et al., 2014
[12] Alneberg et al., 2014
[13] Kang et al., 2015

- GroopM failed to generate any genome bins on our data. I suspect that either our experimental design – metagenomes originating from four different samples – violates (and breaks) GroopM's statistical model, or that it simply was not tested on such large datasets;
- CONCOCT[14] grouped 316,848 contigs (95.7% of all contigs) into 283 genome bins; and
- MetaBAT – in its very specific mode – grouped 72,891 contigs (22% of all contigs) into 532 genome bins.

[14] I used an early version of CONCOCT, recent versions might behave differently.

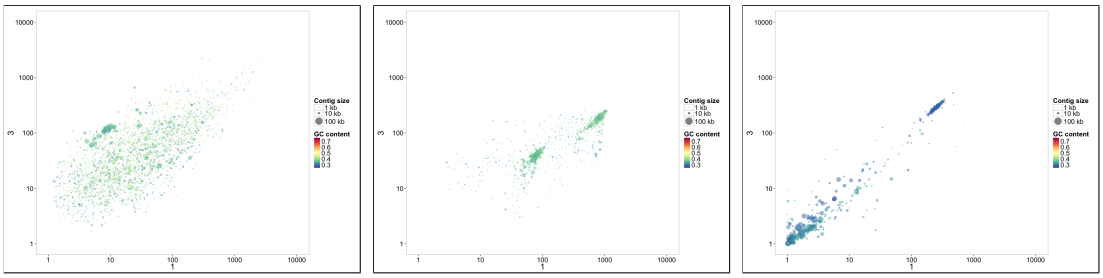


Figure 7.3: **Contig coverages for three CONCOCT bins.** This doesn't look right...

I visually and interactively explored the binning results by generating separate contig coverage plots for each CONCOCT and MetaBAT genome bin. For CONCOCT, I observed many cases for which the contigs's coverages did not align well (Figure 7.3). I did not observe the same effect for any MetaBAT bin.

Following our intuition, we picked the MetaBAT binning results. Even though only 22% of all contigs are binned, the 532 bins contain 62.6% (932 *Mbp*) of the total assembly.

Eventually, our decision was confirmed by a "real" quality assessment with a (then) new tool – CheckM[15] – and I estimated each bin's genome completeness and contamination (Figure 7.5).

[15] Parks et al., 2015

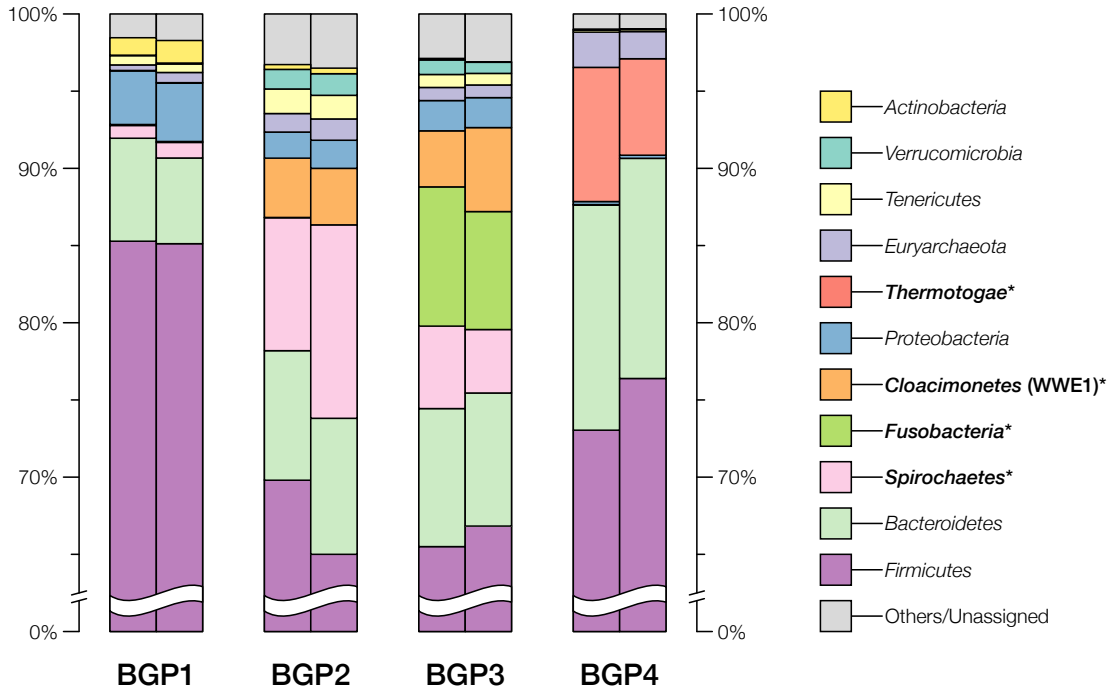


Figure 7.4: **16S-based profile.** Madis Rumming analyzed the 16S amplicon data.

7.5 *Abundant distinct taxa*

16S-based community profiling uncovered four distinct abundant phyla: Thermotogae in BGP4 (the thermophilic BGP); Fusobacteria in BGP3; Spirochaetes in BGP2 and BGP3; and Cloacimonetes in BGP2 and BGP3 (Figure 7.4).[16]

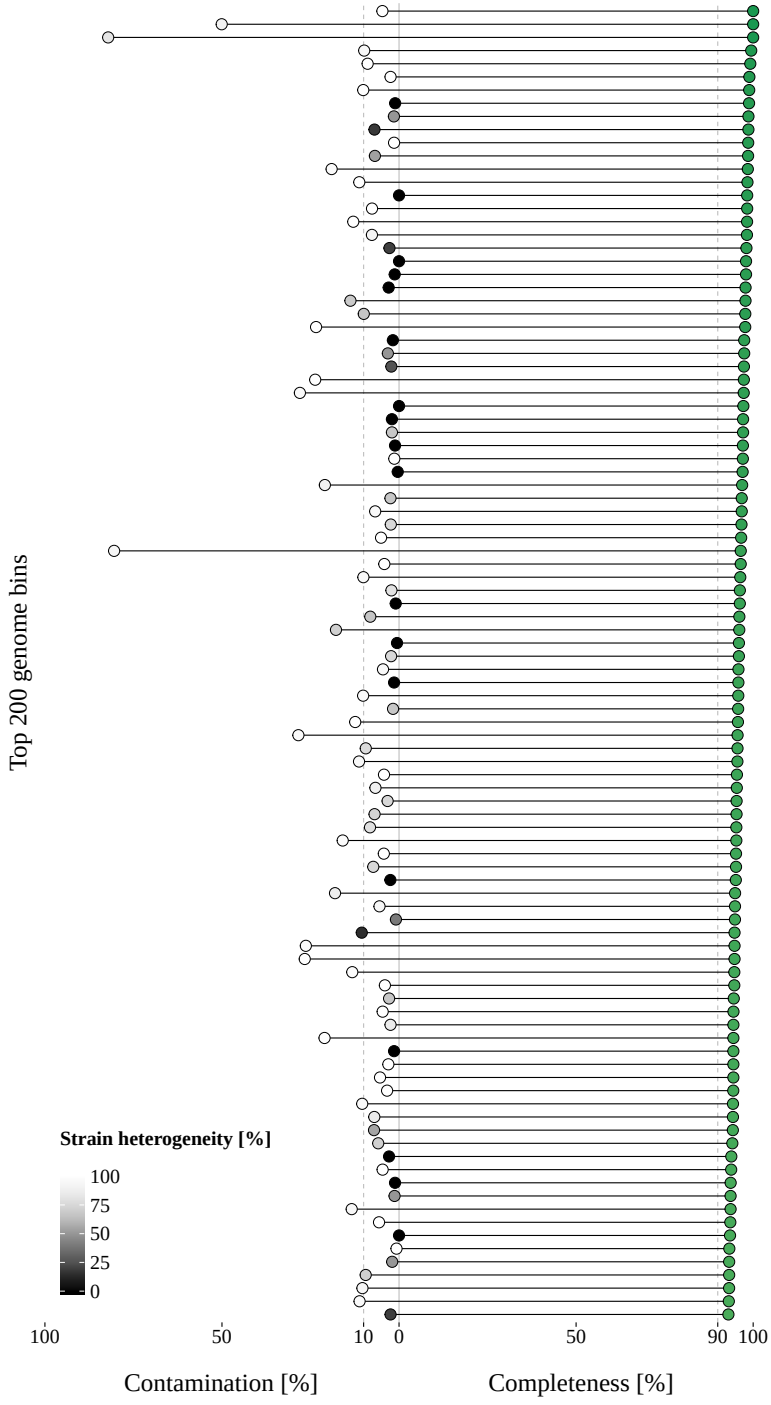
[16] Stolze et al., 2016

I IDENTIFIED GENOME BINS matching these taxa by (A) counting the taxonomic assignments on gene level (which were generated by comparing predicted protein sequences to NCBI's database using the BLASTP mode of DIAMOND[17] and then loading the resulting output file into MEGAN5[18] for taxonomic classification), and (B) running taxator-tk's BLASTN-based binning workflow[19] to additionally assign a taxon label on contig level. These two approaches were largely in agreement, identifying high-confident genome bins for the taxa of interest (Table 7.5).

[17] Buchfink et al., 2015

[18] Huson et al., 2007

[19] Dröge et al., 2015


Top 200 genome bins
Strain heterogeneity [%]
100
75
50
25
0
100
50
10
0
50
90
100
Contamination [%]
Completeness [%]

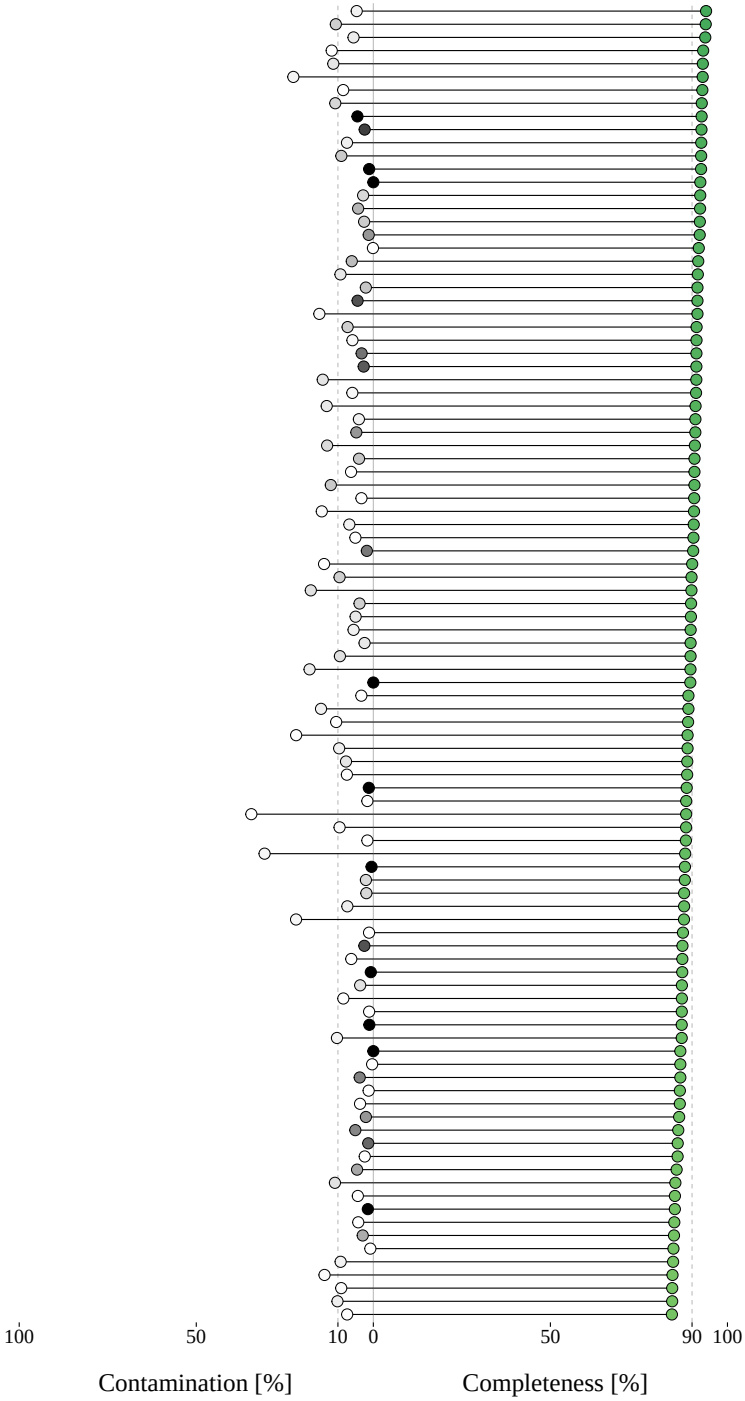


Figure 7.5: **Top 200 genome bins.** A genome catalog of the biogas microbiome.

Bin ID	Assembly size	# contigs	N50
206_Thermotogae	1,904,666	277	8,541
175_Fusobacteria	2,063,893	143	26,189
138_Spirochaetes	2,196,644	86	38,653
244_Cloacimonetes	1,745,914	101	25,062
120_Cloacimonetes	2,265,914	162	18,253

Bin ID	% completeness	% contamination	% strain heterogeneity
206_Thermotogae	82.81	7.37	87.50
175_Fusobacteria	94.38	3.37	100.00
138_Spirochaetes	96.48	4.16	100.00
244_Cloacimonetes	96.70	2.33	75.00
120_Cloacimonetes	95.60	28.42	97.44

Table 7.5: **Genome bins of abundant distinct taxa.** Assembly statistics and quality assessment.

THE GENOME BINS representing the abundant distinct taxa are (with one exception) $\geq$ 90% complete and $\leq$ 10% contaminated. Members of the phylum Cloacimonetes occur in BGP2 and BGP3; I recovered one genome bin from each BGP. The presumably high contamination value for the 120_Cloacimonetes bin is due to strain heterogeneity, *i.e.* different strains of the same species got sorted into this bin.

OF SPECIAL INTEREST is the 206_Thermotogae bin because it matches *Defluviitoga tunisiensis* L3, a strain for which the complete reference genome is known.[20] Comparing the reference genome with our genome bin, we observed that *e.g.* for the known sugar utilization pathways, bin 206_Thermotogae contains all but 13 modules of the closed reference genome (Figure 7.6). This finding further confirmed the applicability and reliability of our binning approach.

[20] Maus et al., 2016a

7.6 Conclusions

Applying state-of-the-art metagenome assembly and binning techniques, I compiled a genome catalog of the biogas microbiome containing hundreds of near-complete genomes.

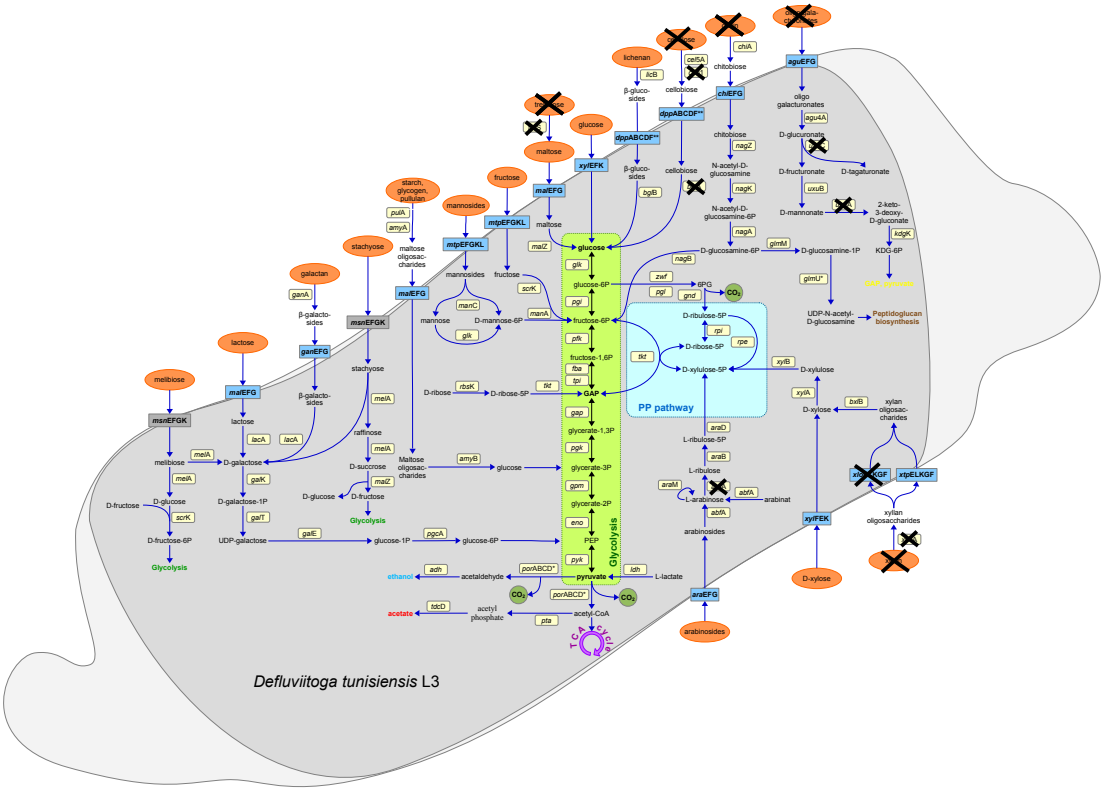


Figure 7.6: **Sugar utilization pathways in the genome bin matching *Defluviitoga tunisiensis* L3.** Metageome binning recovered the near-complete genome sequence containing all but 13 modules. Figure courtesy of Irena Maus, modified from Maus et al., 2016a.

8 Setting the stage for future biogas research

Research of microbial communities residing in industrial biogas plants has long been a focus at the CeBiTec, Bielefeld University.[1] We heralded a new era of assembly-based metagenomics and eventually compiled an exhaustive genome catalog of the biogas microbiome.[2] Metagenome assembly and binning therefore complements the cultivation and sequencing of key players in the biogas-producing microbial community.[3] Combined, these results enable the return to genome-centric analyses and will shape future biogas research at the CeBiTec – and beyond.

[1] Schlüter et al., 2008; Kröber et al., 2009; Jaenicke et al., 2011; Zakrzewski et al., 2012; Eikmeyer et al., 2013

[2] Bremges et al., 2015; Stolze et al., 2016

[3] Maus et al., 2015, 2016a

8.1 Binning-enabled metatranscriptomics

Relating the metagenome and metatranscriptome can identify active members in a microbial community. In our pilot biogas metagenome study, we focused only on distinct abundant taxa in the four biogas plants and identified four of them (represented by five near-complete genome bins).[4] In a follow-up study, we plan to incorporate the corresponding metatranscriptomes to depict the metabolic activity of those genome bins. Metatranscriptomic sequencing was done at the Joint Genome Institute, too; Table 8.1 summarizes these sequencing efforts.

[4] Stolze et al., 2016

To get a first impression, I aligned all metatranscriptome reads to the metagenome assembly – using Bowtie2[5] – and plot each bins's genomic abundance against its expression. Figure 8.1 illustrates this relationship for the Top 200 genome bins, highlighting the ones we focus on and indicating that these are both, abundant and active (part of Yvonne Stolze's PhD project).[6]

[5] Langmead and Salzberg, 2012

[6] Andreas Schlüter's group, CeBiTec, Bielefeld University

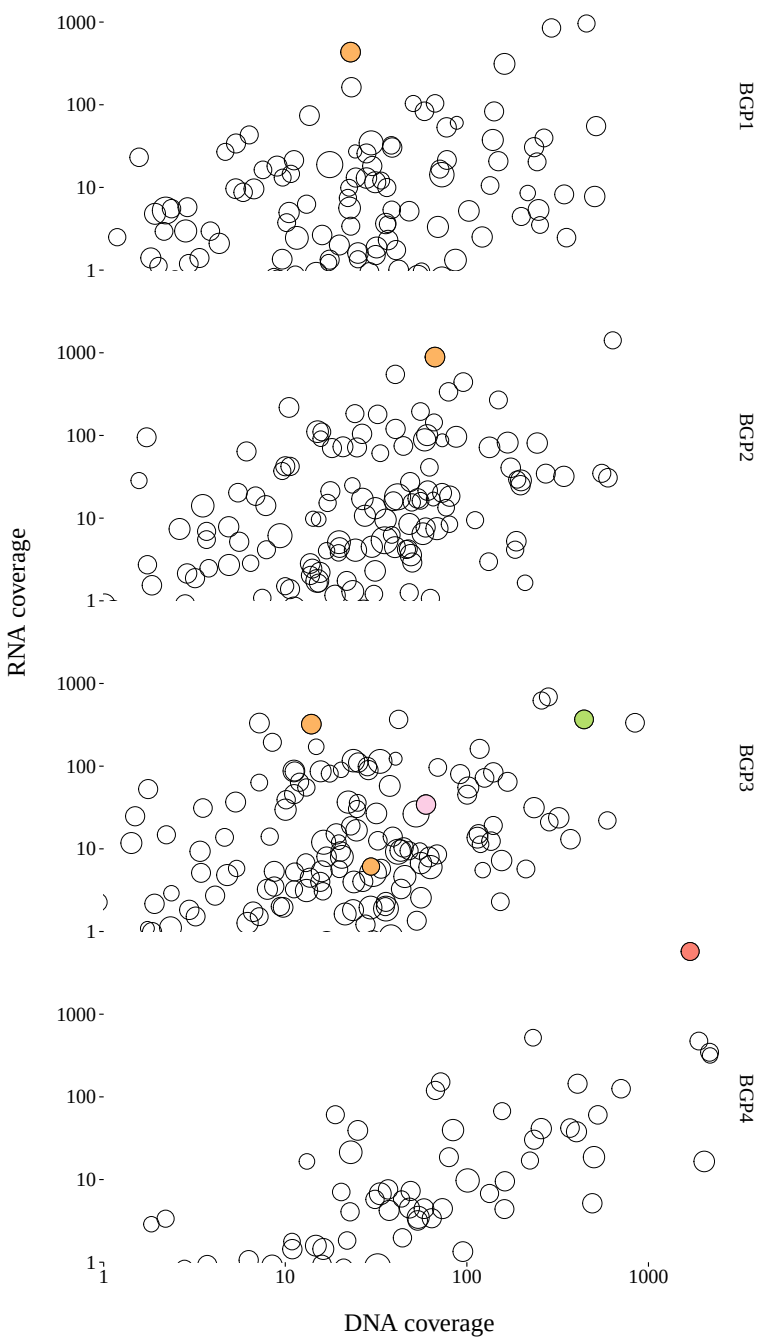


Figure 8.1: **Relating the metagenome and metatranscriptome.** Each circle represents one bin, scaled by its genome size. Distinct abundant taxa from Stolze et al., 2016 in color (see Figure 7.4).

BGP	# reads	# bases	% mapped to assembly
1.1	166,280,226	24,942,033,900	77.76
1.2	223,517,128	33,527,569,200	87.22
2.1	263,289,018	39,493,352,700	80.81
2.2	227,006,018	34,050,902,700	75.50
3.1	261,433,302	39,214,995,300	78.76
3.2	258,702,414	38,805,362,100	78.09
4.1	161,677,326	24,251,598,900	84.74
4.2	233,914,040	35,087,106,000	88.97

Table 8.1: **Metatranscriptome sequencing.** 1,795,819,472 reads (269.4 *Gbp*) were sequenced at the JGI.

8.2 *Integration of other 'omics data*

Due to circumstances beyond our control, we do not yet have single cell sequencing data for biogas-producing community members available. When SAGs are generated eventually, we will (1) assemble them using *e.g.* MeCors or kgrep and thus add dozens of SAG-derived genomes to our catalog; and (2) use SAG reads to validate existing metagenome-derived genomes.[7]

[7] Hess et al., 2011

Integrated metagenome and -proteome analyses will further elucidate the metabolic activity of biogas-producing microbial communities. We have shown that *e.g.* metagenomics complements metaproteomics by significantly improving protein classification rates, but – so far – only scratched the surface.[8]

[8] Kohrs et al., 2015; Ortseifen et al., 2016

8.3 *A focus on Archaea*

The methane metabolism is a key pathway involving methanogenesis performed by methanogenic Archaea, a group for which only a few reference genomes are available.[9] Browsing our biogas genome catalog, I immediately spotted eight near-complete archaeal genome bins. I preliminarily annotated the genome bins with Prokka[10] and BlastKOALA[11] to gauge if methane metabolism pathways are present (they largely are; Figure 8.2).

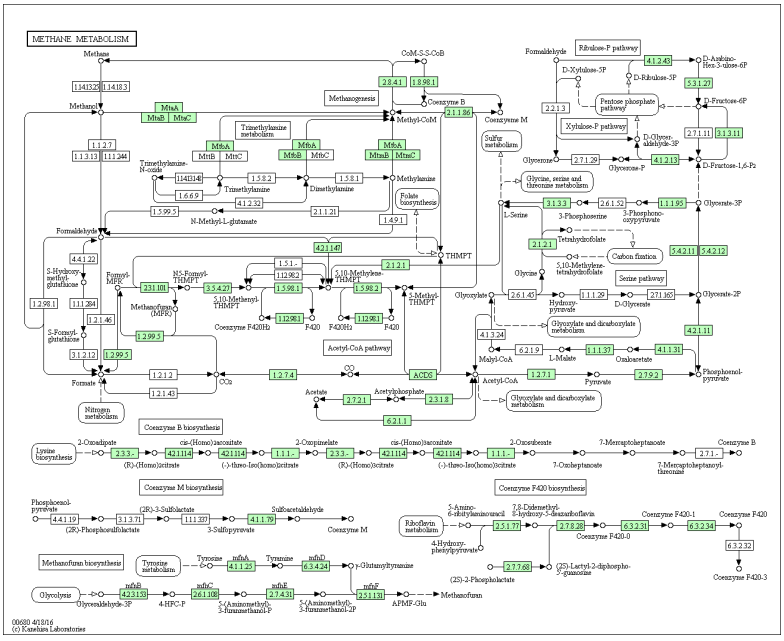
[9] Maus et al., 2012, 2016c

[10] Seemann, 2014

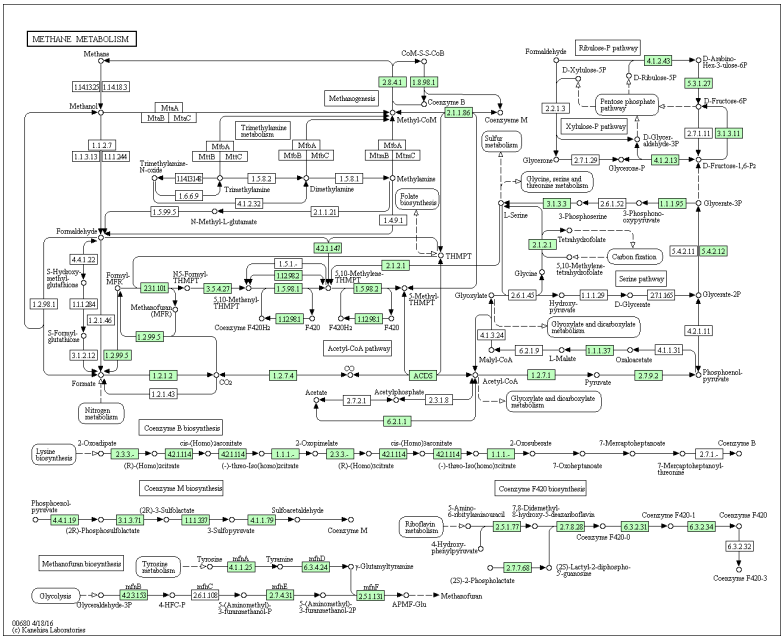
[11] Kanehisa et al., 2016b

Genome-scale metabolic reconstruction of the archaeal subcommunity is (part of) Julia Hassa's PhD project.[12] The foundation for her work is my newly established genome catalog of the biogas microbiome.

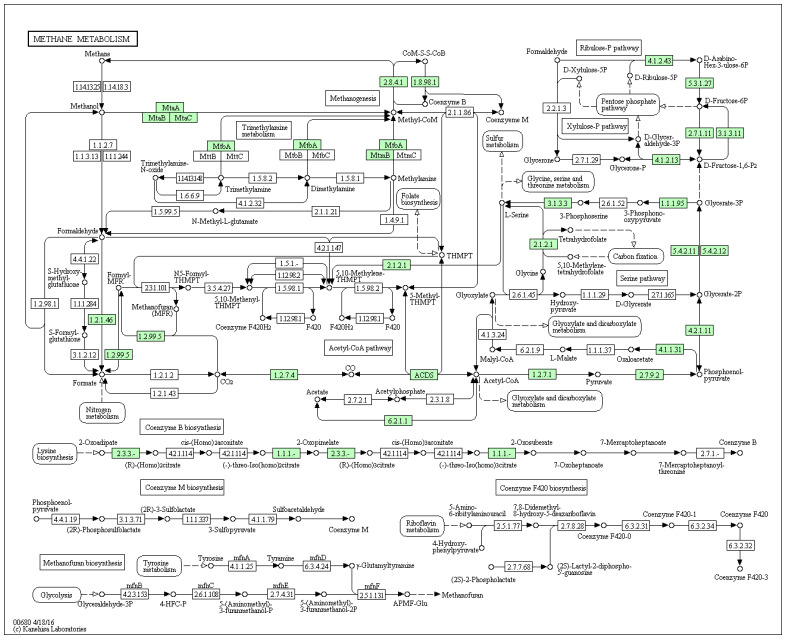
[12] Andreas Schlüter's group, CeBiTec, Bielefeld University



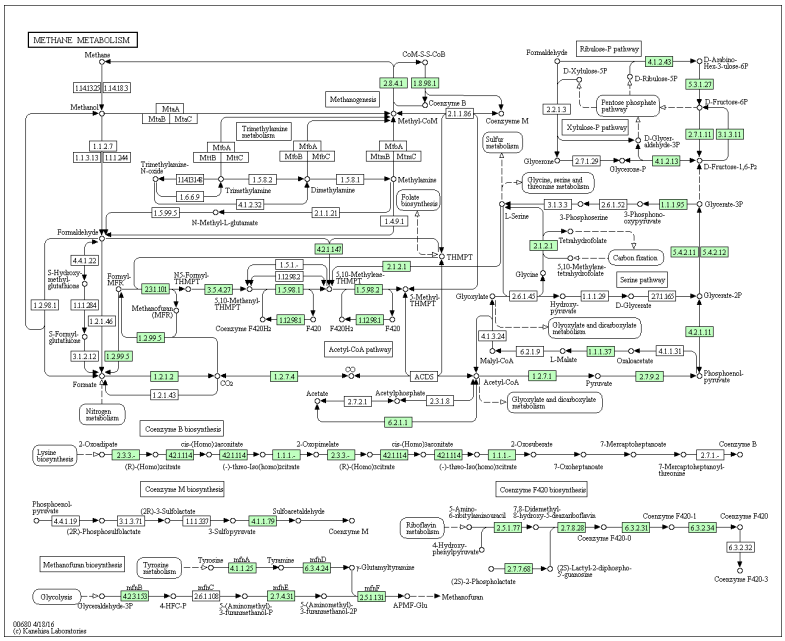
(a) 42_Euryarchaeota; 95.4%



(b) 185_Euryarchaeota; 90.4%

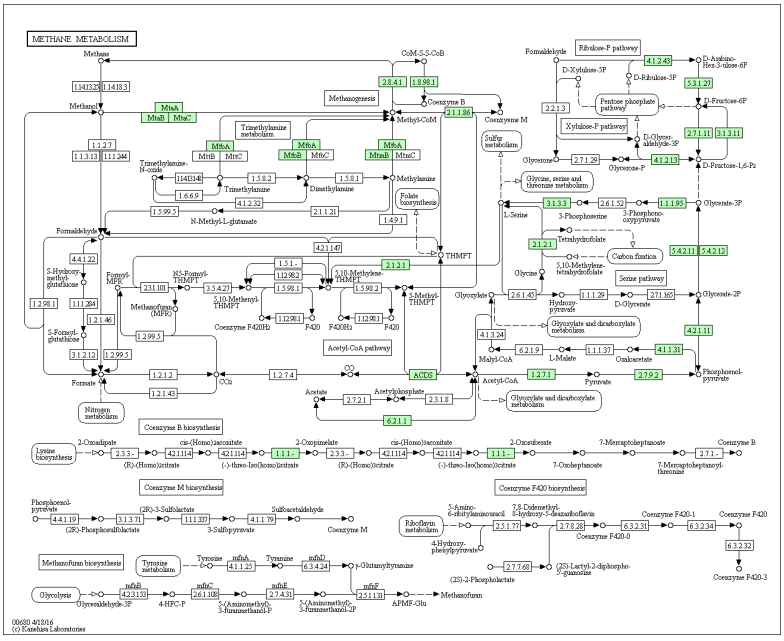


(c) 193_Euryarchaeota; 99.2%

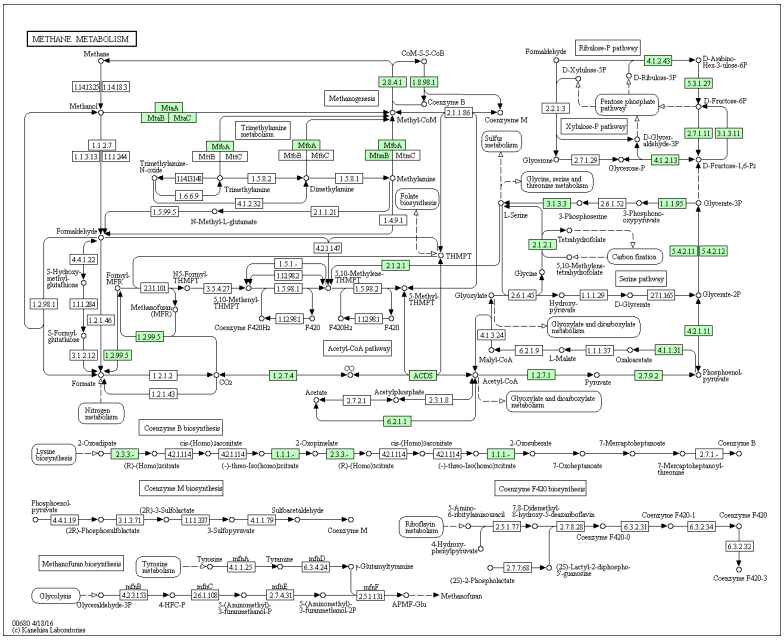


(d) 216_Euryarchaeota; 97.4%

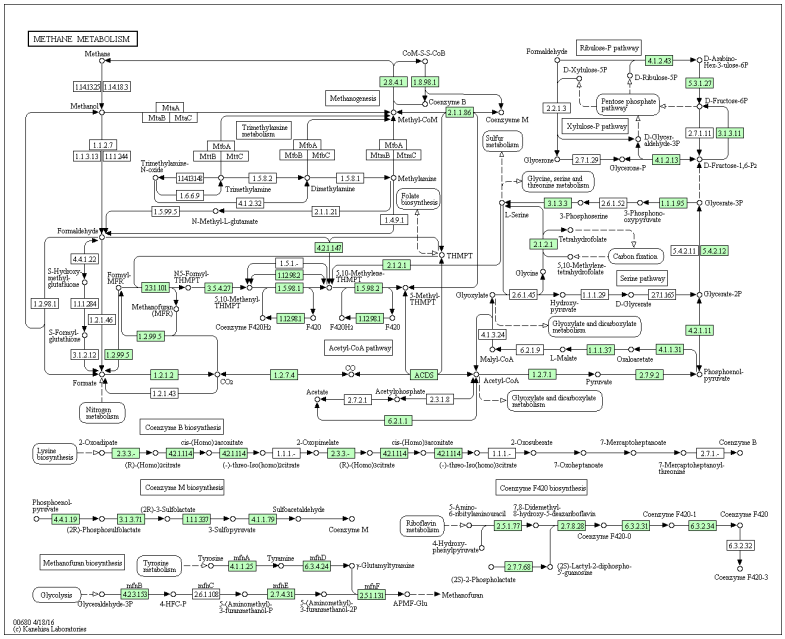
Figure 8.2: **Methane metabolism pathway analyses for eight archaeal genome bins.** Subcaptions give each bin's ID and estimated genome completeness.



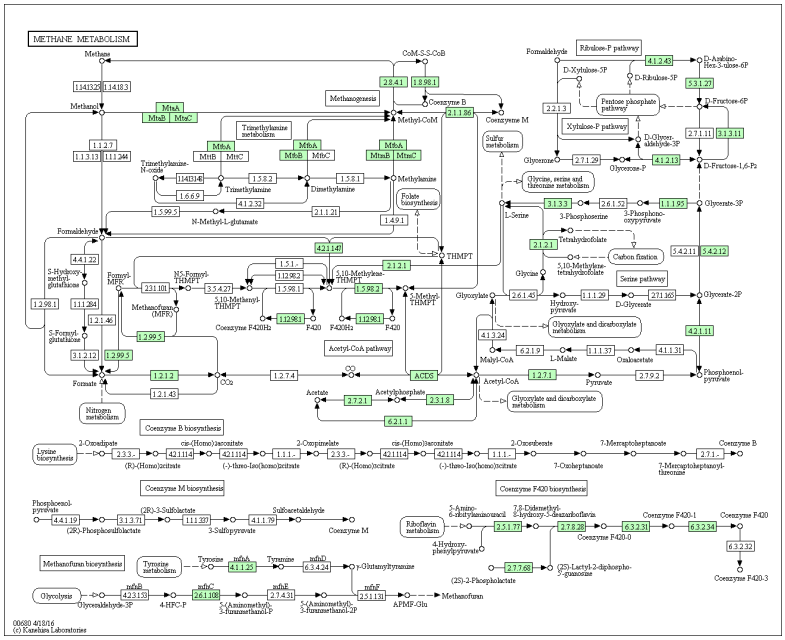
(e) 239_Euryarchaeota; 93.6%



(f) 257_Euryarchaeota; 98.4%



(g) 266_Euryarchaeota; 98.3%



(h) 289_Euryarchaeota; 92.2%

Figure 8.2: **Methane metabolism pathway analyses for eight archaeal genome bins.** Subcaptions give each bin's ID and estimated genome completeness.

Epilogue: The CAMI initiative

Metagenome assembly and successive genome binning is one promising approach to access microbial dark matter genomes.[1] Computational tool development for metagenome assembly and binning is a very active research area and tremendous progress has been achieved during the last years.[2] However, a systematic benchmarking of tools in metagenomics is lacking.

[1] Turaev and Rattei, 2016

[2] Marx, 2016

The Critical Assessment of Metagenomic Information initiative will continuously benchmark tools for metagenome assembly, binning, and profiling. Reproducibility is fostered by using bioboxes, *i.e.* standardised containers for interchangeable bioinformatics software.[3] In the future, researchers will be able to select the most suitable tool for their metagenomic analysis task based on always up-to-date CAMI evaluation results.

[3] Belmann et al., 2015

We started CAMI in 2014 – spearheaded by Alexander Sczyrba, Thomas Rattei, and Alice C. McHardy – and the first results of our evaluations are available at:

`https://data.cami-challenge.org`

Thank you.

Bibliography

Albertsen, M., P. Hugenholtz, A. Skarshewski, K. L. Nielsen, G. W. Tyson, and P. H. Nielsen. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.*, 2013. DOI: 10.1038/nbt.2579.

Alneberg, J., B. S. Bjarnason, I. de Bruijn, M. Schirmer, J. Quick, U. Z. Ijaz, L. Lahti, N. J. Loman, A. F. Andersson, and C. Quince. Binning metagenomic contigs by coverage and composition. *Nat. Methods*, 2014. DOI: 10.1038/nmeth.3103.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 1990. DOI: 10.1016/S0022-2836(05)80360-2.

Bankevich, A., S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham, A. D. Prjibelski, A. V. Pyshkin, A. V. Sirotkin, N. Vyahhi, G. Tesler, M. A. Alekseyev, and P. A. Pevzner. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.*, 2012. DOI: 10.1089/cmb.2012.0021.

Begley, C. G. and L. M. Ellis. Drug development: Raise standards for preclinical cancer research. *Nature*, 2012. DOI: 10.1038/483531a.

Belmann, P., J. Dröge, A. Bremges, A. C. McHardy, A. Sczyrba, and M. D. Barton. Bioboxes: standardised containers for interchangeable bioinformatics software. *GigaScience*, 2015. DOI: 10.1186/s13742-015-0087-0.

Blainey, P. C. The future is now: single-cell genomics of bacteria and archaea. *FEMS Microbiol. Rev.*, 2013. DOI: 10.1111/1574-6976.12015.

Blainey, P. C. and S. R. Quake. Digital MDA for enumeration of total nucleic acid contamination. *Nucleic Acids Res.*, 2011. DOI: 10.1093/nar/gkq1074.

Blainey, P. C. and S. R. Quake. Dissecting genomic diversity, one cell at a time. *Nat. Methods*, 2014. DOI: 10.1038/nmeth.2783.

Boisvert, S., F. Raymond, E. Godzaridis, F. Laviolette, and J. Corbeil. Ray Meta: scalable de novo metagenome assembly and profiling. *Genome Biol.*, 2012. DOI: 10.1186/gb-2012-13-12-r122.

Bolger, A. M., M. Lohse, and B. Usadel. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 2014. DOI: 10.1093/bioinformatics/btu170.

Bowers, R. M., A. Clum, H. Tice, J. Lim, K. Singh, D. Ciobanu, C. Y. Ngan, J. F. Cheng, S. G. Tringe, and T. Woyke. Impact of library preparation protocols and template quantity on the metagenomic reconstruction of a mock microbial community. *BMC Genomics*, 2015. DOI: 10.1186/s12864-015-2063-6.

Bradnam, K. R., J. N. Fass, A. Alexandrov, P. Baranay, M. Bechner, I. Birol, S. Boisvert, J. A. Chapman, G. Chapuis, R. Chikhi, H. Chitsaz, W. C. Chou, J. Corbeil, C. Del Fabbro, T. R. Docking, R. Durbin, D. Earl, S. Emrich, P. Fedotov, N. A. Fonseca, G. Ganapathy, R. A. Gibbs, S. Gnerre, E. Godzaridis, S. Goldstein, M. Haimel, G. Hall, D. Haussler, J. B. Hiatt, I. Y. Ho, J. Howard, M. Hunt, S. D. Jackman, D. B. Jaffe, E. D. Jarvis, H. Jiang, S. Kazakov, P. J. Kersey, J. O. Kitzman, J. R. Knight, S. Koren, T. W. Lam, D. Lavenier, F. Laviolette, Y. Li, Z. Li, B. Liu, Y. Liu, R. Luo, I. Maccallum, M. D. Macmanes, N. Maillet, S. Melnikov, D. Naquin, Z. Ning, T. D. Otto, B. Paten, O. S. Paulo, A. M. Phillippy, F. Pina-Martins, M. Place, D. Przybylski, X. Qin, C. Qu, F. J. Ribeiro, S. Richards, D. S. Rokhsar, J. G. Ruby, S. Scalabrin, M. C. Schatz, D. C. Schwartz, A. Sergushichev, T. Sharpe, T. I. Shaw, J. Shendure, Y. Shi, J. T. Simpson, H. Song, F. Tsarev, F. Vezzi, R. Vicedomini, B. M. Vieira, J. Wang, K. C. Worley, S. Yin, S. M. Yiu, J. Yuan, G. Zhang, H. Zhang, S. Zhou, and I. F. Korf. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience*, 2013. DOI: 10.1186/2047-217X-2-10.

Bremges, A., I. Maus, P. Belmann, F. Eikmeyer, A. Winkler, A. Albersmeier, A. Pühler, A. Schlüter, and A. Sczyrba. Deeply sequenced metagenome and metatranscriptome of a biogas-producing microbial community from an agricultural production-scale biogas plant. *GigaScience*, 2015. DOI: 10.1186/s13742-015-0073-6.

Bremges, A., E. Singer, T. Woyke, and A. Sczyrba. MeCorS: metagenome-enabled error correction of single cell sequencing reads. *Bioinformatics*, 2016. DOI: 10.1093/bioinformatics/btw144.

Bremges, A., J. Jarett, T. Woyke, and A. Sczyrba. Metagenomic proxy assemblies of single cell genomes. *TBA*, in prep.

Broder, A. On the resemblance and containment of documents. In *Proceedings of the Compression and Complexity of Sequences 1997*, 1997.

Brown, C. T., L. A. Hug, B. C. Thomas, I. Sharon, C. J. Castelle, A. Singh, M. J. Wilkins, K. C. Wrighton, K. H. Williams, and J. F. Banfield. Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature*, 2015. DOI: 10.1038/nature14486.

Buchfink, B., C. Xie, and D. H. Huson. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*, 2015. DOI: 10.1038/nmeth.3176.

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden. BLAST+: architecture and applications. *BMC Bioinformatics*, 2009. DOI: 10.1186/1471-2105-10-421.

Campanaro, S., L. Treu, P. G. Kougias, D. De Francisci, G. Valle, and I. Angelidaki. Metagenomic analysis and functional characterization of the biogas microbiome using high throughput shotgun sequencing and a novel binning strategy. *Biotechnol Biofuels*, 2016. DOI: 10.1186/s13068-016-0441-1.

Campbell, J. H., P. O'Donoghue, A. G. Campbell, P. Schwientek, A. Sczyrba, T. Woyke, D. Soll, and M. Podar. UGA is an additional glycine codon in uncultured SR1 bacteria from the human microbiota. *Proc. Natl. Acad. Sci. U.S.A.*, 2013. DOI: 10.1073/pnas.1303090110.

Chikhi, R. and P. Medvedev. Informed and automated k-mer size selection for genome assembly. *Bioinformatics*, 2014. DOI: 10.1093/bioinformatics/btt310.

Chikhi, R. and G. Rizk. Space-efficient and exact de bruijn graph representation based on a bloom filter. In *WABI*, 2012.

Chitsaz, H., J. L. Yee-Greenbaum, G. Tesler, M. J. Lombardo, C. L. Dupont, J. H. Badger, M. Novotny, D. B. Rusch, L. J. Fraser, N. A. Gormley, O. Schulz-Trieglaff, G. P. Smith, D. J. Evers, P. A. Pevzner, and R. S. Lasken. Efficient de novo assembly of single-cell bacterial genomes from short-read data sets. *Nat. Biotechnol.*, 2011. DOI: 10.1038/nbt.1966.

Clingenpeel, S., A. Clum, P. Schwientek, C. Rinke, and T. Woyke. Reconstructing each cell's genome within complex microbial communities – dream or reality? *Front Microbiol*, 2014a. DOI: 10.3389/fmicb.2014.00771.

Clingenpeel, S., P. Schwientek, P. Hugenholtz, and T. Woyke. Effects of sample treatments on genome recovery via single-cell genomics. *ISME J*, 2014b. DOI: 10.1038/ismej.2014.92.

Compeau, P. E., P. A. Pevzner, and G. Tesler. How to apply de Bruijn graphs to genome assembly. *Nat. Biotechnol.*, 2011. DOI: 10.1038/nbt.2023.

Darwin, C. R. *Transmutation of Species*. "B" notebook, 1837.

Darwin, C. R. *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. John Murray, 1859.

de Bourcy, C. F., I. De Vlaminck, J. N. Kanbar, J. Wang, C. Gawad, and S. R. Quake. A quantitative comparison of single-cell whole genome amplification methods. *PLoS ONE*, 2014. DOI: 10.1371/journal.pone.0105585.

Demain, A. L. and S. Sanchez. Microbial drug discovery: 80 years of progress. *J. Antibiot.*, 2009. DOI: 10.1038/ja.2008.16.

Dick, G. J., A. F. Andersson, B. J. Baker, S. L. Simmons, B. C. Thomas, A. P. Yelton, and J. F. Banfield. Community-wide analysis of microbial genome sequence signatures. *Genome Biol.*, 2009. DOI: 10.1186/gb-2009-10-8-r85.

Dodsworth, J. A., P. C. Blainey, S. K. Murugapiran, W. D. Swingley, C. A. Ross, S. G. Tringe, P. S. Chain, M. B. Scholz, C. C. Lo, J. Raymond, S. R. Quake, and B. P. Hedlund. Single-cell and metagenomic analyses indicate a fermentative and saccharolytic lifestyle for members of the OP9 lineage. *Nat Commun*, 2013. DOI: 10.1038/ncomms2884.

Dröge, J., I. Gregor, and A. C. McHardy. Taxator-tk: precise taxonomic assignment of metagenomes by fast approximation of evolutionary neighborhoods. *Bioinformatics*, 2015. DOI: 10.1093/bioinformatics/btu745.

Earl, D., K. Bradnam, J. St John, A. Darling, D. Lin, J. Fass, H. O. Yu, V. Buffalo, D. R. Zerbino, M. Diekhans, N. Nguyen, P. N. Ariyaratne, W. K. Sung, Z. Ning, M. Haimel, J. T. Simpson, N. A. Fonseca, I. Birol, T. R. Docking, I. Y. Ho, D. S. Rokhsar, R. Chikhi, D. Lavenier, G. Chapuis, D. Naquin, N. Maillet, M. C. Schatz, D. R. Kelley, A. M. Phillippy, S. Koren, S. P. Yang, W. Wu, W. C. Chou, A. Srivastava, T. I. Shaw, J. G. Ruby, P. Skewes-Cox, M. Betegon, M. T. Dimon, V. Solovyev, I. Seledtsov, P. Kosarev, D. Vorobyev, R. Ramirez-Gonzalez, R. Leggett, D. MacLean, F. Xia, R. Luo, Z. Li, Y. Xie, B. Liu, S. Gnerre, I. MacCallum, D. Przybylski, F. J. Ribeiro, S. Yin, T. Sharpe, G. Hall, P. J. Kersey, R. Durbin, S. D. Jackman, J. A. Chapman, X. Huang, J. L. DeRisi, M. Caccamo, Y. Li, D. B. Jaffe, R. E. Green, D. Haussler, I. Korf, and B. Paten. Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome Res.*, 2011. DOI: 10.1101/gr.126599.111.

Eberwine, J., J. Y. Sul, T. Bartfai, and J. Kim. The promise of single-cell sequencing. *Nat. Methods*, 2014. DOI: 10.1038/nmeth.2769.

Eddy, S. R. Profile hidden Markov models. *Bioinformatics*, 1998. DOI: 10.1093/bioinformatics/14.9.755.

Eddy, S. R. A probabilistic model of local sequence alignment that simplifies statistical significance estimation. *PLoS Comput. Biol.*, 2008. DOI: 10.1371/journal.pcbi.1000069.

Eikmeyer, F. G., A. Rademacher, A. Hanreich, M. Hennig, S. Jaenicke, I. Maus, D. Wibberg, M. Zakrzewski, A. Pühler, M. Klocke, and A. Schlüter. Detailed analysis of metagenome datasets obtained from biogas-producing microbial communities residing in biogas reactors does not indicate the presence of putative pathogenic microorganisms. *Biotechnol Biofuels*, 2013. DOI: 10.1186/1754-6834-6-49.

Eren, A. M., O. C. Esen, C. Quince, J. H. Vineis, H. G. Morrison, M. L. Sogin, and T. O. Delmont. Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ*, 2015. DOI: 10.7717/peerj.1319.

Escobar-Zepeda, A., A. Vera-Ponce de Leon, and A. Sanchez-Flores. The road to metagenomics: from microbiology to DNA sequencing technologies and bioinformatics. *Front Genet*, 2015. DOI: 10.3389/fgene.2015.00348.

Federhen, S. The NCBI Taxonomy database. *Nucleic Acids Res.*, 2012. DOI: 10.1093/nar/gkr1178.

Filée, J., F. Tetart, C. A. Suttle, and H. M. Krisch. Marine T4-type bacteriophages, a ubiquitous component of the dark matter of the biosphere. *Proc. Natl. Acad. Sci. U.S.A.*, 2005. DOI: 10.1073/pnas.0503404102.

Gallagher, K. A., W. Fenical, and P. R. Jensen. Hybrid isoprenoid secondary metabolite production in terrestrial and marine actinomycetes. *Curr. Opin. Biotechnol.*, 2010. DOI: 10.1016/j.copbio.2010.09.010.

Gawad, C., W. Koh, and S. R. Quake. Single-cell genome sequencing: current state of the science. *Nat. Rev. Genet.*, 2016. DOI: 10.1038/nrg.2015.16.

Gerlach, W., S. Jünemann, F. Tille, A. Goesmann, and J. Stoye. WebCARMA: a web application for the functional and taxonomic classification of unassembled metagenomic reads. *BMC Bioinformatics*, 2009. DOI: 10.1186/1471-2105-10-430.

Gies, E. A., K. M. Konwar, J. T. Beatty, and S. J. Hallam. Illuminating microbial dark matter in meromictic Sakinaw Lake. *Appl. Environ. Microbiol.*, 2014. DOI: 10.1128/AEM.01774-14.

Gole, J., A. Gore, A. Richards, Y. J. Chiu, H. L. Fung, D. Bushman, H. I. Chiang, J. Chun, Y. H. Lo, and K. Zhang. Massively parallel polymerase cloning and genome sequencing of single cells using nanoliter microwells. *Nat. Biotechnol.*, 2013. DOI: 10.1038/nbt.2720.

Goodwin, S., J. D. McPherson, and W. R. McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.*, 2016. DOI: 10.1038/nrg.2016.49.

Gubler, U. and B. J. Hoffman. A simple and very efficient method for generating cDNA libraries. *Gene*, 1983. DOI: 10.1016/0378-1119(83)90230-5.

Gurevich, A., V. Saveliev, N. Vyahhi, and G. Tesler. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, 2013. DOI: 10.1093/bioinformatics/btt086.

Hahn, C., L. Bachmann, and B. Chevreux. Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads–a baiting and iterative mapping approach. *Nucleic Acids Res.*, 2013. DOI: 10.1093/nar/gkt371.

Hedlund, B. P., J. A. Dodsworth, S. K. Murugapiran, C. Rinke, and T. Woyke. Impact of single-cell genomics and metagenomics on the emerging view of extremophile "microbial dark matter". *Extremophiles*, 2014. DOI: 10.1007/s00792-014-0664-7.

Hess, M., A. Sczyrba, R. Egan, T. W. Kim, H. Chokhawala, G. Schroth, S. Luo, D. S. Clark, F. Chen, T. Zhang, R. I. Mackie, L. A. Pennacchio, S. G. Tringe, A. Visel, T. Woyke, Z. Wang, and E. M. Rubin. Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science*, 2011. DOI: 10.1126/science.1200387.

Hinchliff, C. E., S. A. Smith, J. F. Allman, J. G. Burleigh, R. Chaudhary, L. M. Coghill, K. A. Crandall, J. Deng, B. T. Drew, R. Gazis, K. Gude, D. S. Hibbett, L. A. Katz, H. D. Laughinghouse, E. J. McTavish, P. E. Midford, C. L. Owen, R. H. Ree, J. A. Rees, D. E. Soltis, T. Williams, and K. A. Cranston. Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proc. Natl. Acad. Sci. U.S.A.*, 2015. DOI: 10.1073/pnas.1423041112.

Hug, L. A., B. J. Baker, K. Anantharaman, C. T. Brown, A. J. Probst, C. J. Castelle, C. N. Butterfield, A. W. Hernsdorf, Y. Amano, K. Ise, Y. Suzuki, N. Dudek, D. A. Relman, K. M. Finstad, R. Amundson, B. C. Thomas, and J. F. Banfield. A new view of the tree of life. *Nat. Microbiol.*, 2016. DOI: 10.1038/nmicrobiol.2016.48.

Huson, D. H., A. F. Auch, J. Qi, and S. C. Schuster. MEGAN analysis of metagenomic data. *Genome Res.*, 2007. DOI: 10.1101/gr.5969107.

Hyatt, D., P. F. LoCascio, L. J. Hauser, and E. C. Uberbacher. Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics*, 2012. DOI: 10.1093/bioinformatics/bts429.

Imelfort, M., D. Parks, B. J. Woodcroft, P. Dennis, P. Hugenholtz, and G. W. Tyson. GroopM: an automated tool for the recovery of population genomes from related metagenomes. *PeerJ*, 2014. DOI: 10.7717/peerj.603.

Ishoey, T., T. Woyke, R. Stepanauskas, M. Novotny, and R. S. Lasken. Genomic sequencing of single microbial cells from environmental samples. *Curr. Opin. Microbiol.*, 2008. DOI: 10.1016/j.mib.2008.05.006.

Iverson, V., R. M. Morris, C. D. Frazar, C. T. Berthiaume, R. L. Morales, and E. V. Armbrust. Untangling genomes from metagenomes: revealing an uncultured class of marine Euryarchaeota. *Science*, 2012. DOI: 10.1126/science.1212665.

Jaenicke, S., C. Ander, T. Bekel, R. Bisdorf, M. Dröge, K. H. Gartemann, S. Jünemann, O. Kaiser, L. Krause, F. Tille, M. Zakrzewski, A. Pühler, A. Schlüter, and A. Goesmann. Comparative and joint analysis of two metagenomic datasets from a biogas fermenter obtained by 454-pyrosequencing. *PLoS ONE*, 2011. DOI: 10.1371/journal.pone.0014519.

Jünemann, S., K. Prior, A. Albersmeier, S. Albaum, J. Kalinowski, A. Goesmann, J. Stoye, and D. Harmsen. GABenchToB: a genome assembly benchmark tuned on bacteria and benchtop sequencers. *PLoS ONE*, 2014. DOI: 10.1371/journal.pone.0107014.

Kanehisa, M., S. Goto, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.*, 2014. DOI: 10.1093/nar/gkt1076.

Kanehisa, M., Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.*, 2016a. DOI: 10.1093/nar/gkv1070.

Kanehisa, M., Y. Sato, and K. Morishima. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *J. Mol. Biol.*, 2016b. DOI: 10.1016/j.jmb.2015.11.006.

Kang, D. D., J. Froula, R. Egan, and Z. Wang. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ*, 2015. DOI: 10.7717/peerj.1165.

Kelley, D. R., M. C. Schatz, and S. L. Salzberg. Quake: quality-aware detection and correction of sequencing errors. *Genome Biol.*, 2010. DOI: 10.1186/gb-2010-11-11-r116.

Klemm, D., B. Heublein, H. P. Fink, and A. Bohn. Cellulose: fascinating biopolymer and sustainable raw material. *Angew. Chem. Int. Ed. Engl.*, 2005. DOI: 10.1002/anie.200460587.

Kohrs, F., S. Wolter, D. Benndorf, R. Heyer, M. Hoffmann, E. Rapp, A. Bremges, A. Sczyrba, A. Schlüter, and U. Reichl. Fractionation of biogas plant sludge material improves metaproteomic characterization to investigate metabolic activity of microbial communities. *Proteomics*, 2015. DOI: 10.1002/pmic.201400557.

Konstantinidis, K. T. and J. M. Tiedje. Genomic insights that advance the species definition for prokaryotes. *Proc. Natl. Acad. Sci. U.S.A.*, 2005. DOI: 10.1073/pnas.0409727102.

Krause, L., N. N. Diaz, A. Goesmann, S. Kelley, T. W. Nattkemper, F. Rohwer, R. A. Edwards, and J. Stoye. Phylogenetic classification of short environmental DNA fragments. *Nucleic Acids Res.*, 2008. DOI: 10.1093/nar/gkn038.

Krzywinski, M., J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, and M. A. Marra. Circos: an information aesthetic for comparative genomics. *Genome Res.*, 2009. DOI: 10.1101/gr.092759.109.

Kröber, M., T. Bekel, N. N. Diaz, A. Goesmann, S. Jaenicke, L. Krause, D. Miller, K. J. Runte, P. Viehöver, A. Pühler, and A. Schlüter. Phylogenetic characterization of a biogas plant microbial community integrating clone library 16S-rDNA sequences and metagenome sequence data obtained by 454-pyrosequencing. *J. Biotechnol.*, 2009. DOI: 10.1016/j.jbiotec.2009.02.010.

Kunin, V., A. Copeland, A. Lapidus, K. Mavromatis, and P. Hugenholtz. A bioinformatician's guide to metagenomics. *Microbiol. Mol. Biol. Rev.*, 2008. DOI: 10.1128/MMBR.00009-08.

Laehnemann, D., A. Borkhardt, and A. C. McHardy. Denoising DNA deep sequencing data-high-throughput sequencing errors and their correction. *Brief. Bioinformatics*, 2016. DOI: 10.1093/bib/bbv029.

Lane, D. J., B. Pace, G. J. Olsen, D. A. Stahl, M. L. Sogin, and N. R. Pace. Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc. Natl. Acad. Sci. U.S.A.*, 1985.

Langmead, B. and S. L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, 2012. DOI: 10.1038/nmeth.1923.

Lasken, R. S. Single-cell genomic sequencing using Multiple Displacement Amplification. *Curr. Opin. Microbiol.*, 2007. DOI: 10.1016/j.mib.2007.08.005.

Lasken, R. S. Single-cell sequencing in its prime. *Nat. Biotechnol.*, 2013. DOI: 10.1038/nbt.2523.

Lasken, R. S. and T. B. Stockwell. Mechanism of chimera formation during the Multiple Displacement Amplification reaction. *BMC Biotechnol.*, 2007. DOI: 10.1186/1472-6750-7-19.

Laurence, M., C. Hatzis, and D. E. Brash. Common contaminants in next-generation sequencing that hinder discovery of low-abundance microbes. *PLoS ONE*, 2014. DOI: 10.1371/journal.pone.0097876.

Lewis, K. Platforms for antibiotic discovery. *Nat Rev Drug Discov*, 2013. DOI: 10.1038/nrd3975.

Li, D., C. M. Liu, R. Luo, K. Sadakane, and T. W. Lam. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, 2015. DOI: 10.1093/bioinformatics/btv033.

Li, H. Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly. *Bioinformatics*, 2012. DOI: 10.1093/bioinformatics/bts280.

Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:1303.3997*, 2013.

Li, H. BFC: correcting Illumina sequencing errors. *Bioinformatics*, 2015. DOI: 10.1093/bioinformatics/btv290.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 2009. DOI: 10.1093/bioinformatics/btp352.

Lok, C. Mining the microbial dark matter. *Nature*, 2015. DOI: 10.1038/522270a.

Lusk, R. W. Diverse and widespread contamination evident in the unmapped depths of high throughput sequencing data. *PLoS ONE*, 2014. DOI: 10.1371/journal.pone.0110808.

Lux, M., B. Hammer, and A. Sczyrba. Automated contamination detection in single-cell sequencing. *bioRxiv*, 2015. DOI: 10.1101/020859.

Magoc, T., S. Pabinger, S. Canzar, X. Liu, Q. Su, D. Puiu, L. J. Tallon, and S. L. Salzberg. GAGE-B: an evaluation of genome assemblers for bacterial organisms. *Bioinformatics*, 2013. DOI: 10.1093/bioinformatics/btt273.

Manivasagan, P., J. Venkatesan, K. Sivakumar, and S. K. Kim. Pharmaceutically active secondary metabolites of marine actinobacteria. *Microbiol. Res.*, 2014. DOI: 10.1016/j.micres.2013.07.014.

Marcy, Y., C. Ouverney, E. M. Bik, T. Losekann, N. Ivanova, H. G. Martin, E. Szeto, D. Platt, P. Hugenholtz, D. A. Relman, and S. R. Quake. Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *Proc. Natl. Acad. Sci. U.S.A.*, 2007. DOI: 10.1073/pnas.0704662104.

Marx, V. Microbiology: the road to strain-level identification. *Nat. Methods*, 2016. DOI: 10.1038/nmeth.3837.

Matsen, F. A., R. B. Kodner, and E. V. Armbrust. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*, 2010. DOI: 10.1186/1471-2105-11-538.

Maus, I., D. Wibberg, R. Stantscheff, F. G. Eikmeyer, A. Seffner, J. Boelter, R. Szczepanowski, J. Blom, S. Jaenicke, H. Konig, A. Pühler, and A. Schlüter. Complete genome sequence of the hydrogenotrophic, methanogenic archaeon Methanoculleus bourgensis strain MS2(T), isolated from a sewage sludge digester. *J. Bacteriol.*, 2012. DOI: 10.1128/JB.01292-12.

Maus, I., D. Wibberg, R. Stantscheff, Y. Stolze, J. Blom, F. G. Eikmeyer, J. Fracowiak, H. König, A. Pühler, and A. Schlüter. Insights into the annotated genome sequence of Methanoculleus bourgensis MS2(T), related to dominant methanogens in biogas-producing plants. *J. Biotechnol.*, 2015. DOI: 10.1016/j.jbiotec.2014.11.020.

Maus, I., K. G. Cibis, A. Bremges, Y. Stolze, D. Wibberg, G. Tomazetto, J. Blom, A. Sczyrba, H. König, A. Pühler, and A. Schlüter. Genomic characterization of Defluviitoga tunisiensis L3, a key hydrolytic bacterium in a thermophilic biogas plant and its abundance as determined by metagenome fragment recruitment. *J. Biotechnol.*, 2016a. DOI: 10.1016/j.jbiotec.2016.05.001.

Maus, I., D. E. Koeck, K. Cibis, S. Hahnke, Y. S. Kim, T. Langer, J. Kreubel, M. Erhard, A. Bremges, S. Off, Y. Stolze, S. Jaenicke, A. Sczyrba, P. Scherer, H. König, W. H. Schwarz, A. Pühler, A. Schlüter, and M. Klocke. Unraveling the microbiome of a thermophilic biogas plant by metagenome and metatranscriptome analysis complemented by characterization of bacterial and archaeal isolates. *Biotechnol Biofuels*, 2016b. DOI: 10.1186/s13068-016-0581-3.

Maus, I., D. Wibberg, A. Winkler, A. Pühler, A. Schnurer, and A. Schlüter. Complete Genome Sequence of the Methanogen Methanoculleus bourgensis BA1 Isolated from a Biogas Reactor. *Genome Announc*, 2016c. DOI: 10.1128/genomeA.00568-16.

McHardy, A. C. and I. Rigoutsos. What's in the mix: phylogenetic classification of metagenome sequence samples. *Curr. Opin. Microbiol.*, 2007. DOI: 10.1016/j.mib.2007.08.004.

McLean, J. S., M. J. Lombardo, J. H. Badger, A. Edlund, M. Novotny, J. Yee-Greenbaum, N. Vyahhi, A. P. Hall, Y. Yang, C. L. Dupont, M. G. Ziegler, H. Chitsaz, A. E. Allen, S. Yooseph, G. Tesler, P. A. Pevzner, R. M. Friedman, K. H. Nealson, J. C. Venter, and R. S. Lasken. Candidate phylum TM6 genome recovered from a hospital sink biofilm provides genomic insights into this uncultivated phylum. *Proc. Natl. Acad. Sci. U.S.A.*, 2013. DOI: 10.1073/pnas.1219809110.

Medvedev, P., E. Scott, B. Kakaradov, and P. Pevzner. Error correction of high-throughput sequencing datasets with non-uniform coverage. *Bioinformatics*, 2011. DOI: 10.1093/bioinformatics/btr208.

Minoche, A. E., J. C. Dohm, and H. Himmelbauer. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome Biol.*, 2011. DOI: 10.1186/gb-2011-12-11-r112.

Morrison, M., P. B. Pope, S. E. Denman, and C. S. McSweeney. Plant biomass degradation by gut microbiomes: more of the same or something new? *Curr. Opin. Biotechnol.*, 2009. DOI: 10.1016/j.copbio.2009.05.004.

Nagarajan, N. and M. Pop. Sequence assembly demystified. *Nat. Rev. Genet.*, 2013. DOI: 10.1038/nrg3367.

Nikolenko, S. I., A. I. Korobeynikov, and M. A. Alekseyev. BayesHammer: Bayesian clustering for error correction in single-cell sequencing. *BMC Genomics*, 2013. DOI: 10.1186/1471-2164-14-S1-S7.

Nobu, M. K., J. A. Dodsworth, S. K. Murugapiran, C. Rinke, E. A. Gies, G. Webster, P. Schwientek, P. Kille, R. J. Parkes, H. Sass, B. B. Jørgensen, A. J. Weightman, W. T. Liu, S. J. Hallam, G. Tsiamis, T. Woyke, and B. P. Hedlund. Phylogeny and physiology of candidate phylum 'Atribacteria' (OP9/JS1) inferred from cultivation-independent genomics. *ISME J*, 2016. DOI: 10.1038/ismej.2015.97.

Nurk, S., A. Bankevich, D. Antipov, A. A. Gurevich, A. Korobeynikov, A. Lapidus, A. D. Prjibelski, A. Pyshkin, A. Sirotkin, Y. Sirotkin, R. Stepanauskas, S. R. Clingenpeel, T. Woyke, J. S. McLean, R. S. Lasken, G. Tesler, M. A. Alekseyev, and P. A. Pevzner. Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. *J. Comput. Biol.*, 2013. DOI: 10.1089/cmb.2013.0084.

Nurk, S., D. Meleshko, A. Korobeynikov, and P. Pevzner. metaSPAdes: a new versatile de novo metagenomics assembler. *arXiv:1604.03071*, 2016.

Ohsawa, G. *The Unique Principle*. J. Vrin Philosophical Library, 1931.

Ondov, B. D., N. H. Bergman, and A. M. Phillippy. Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics*, 2011. DOI: 10.1186/1471-2105-12-385.

Ondov, B. D., T. J. Treangen, P. Melsted, A. B. Mallonee, N. H. Bergman, S. Koren, and A. M. Phillippy. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.*, 2016. DOI: 10.1186/s13059-016-0997-x.

Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 2015. DOI: 10.1126/science.aac4716.

Ortseifen, V., Y. Stolze, I. Maus, A. Sczyrba, A. Bremges, S. P. Albaum, S. Jaenicke, J. Fracowiak, A. Pühler, and A. Schlüter. An integrated metagenome and -proteome analysis of the microbial community residing in a biogas production plant. *J. Biotechnol.*, 2016. DOI: 10.1016/j.jbiotec.2016.06.014.

Parks, D. H., M. Imelfort, C. T. Skennerton, P. Hugenholtz, and G. W. Tyson. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.*, 2015. DOI: 10.1101/gr.186072.114.

Peng, Y., H. C. Leung, S. M. Yiu, and F. Y. Chin. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, 2012. DOI: 10.1093/bioinformatics/bts174.

Pevzner, P. A., H. Tang, and M. S. Waterman. An Eulerian path approach to DNA fragment assembly. *Proc. Natl. Acad. Sci. U.S.A.*, 2001. DOI: 10.1073/pnas.171285098.

Podar, M., C. B. Abulencia, M. Walcher, D. Hutchison, K. Zengler, J. A. Garcia, T. Holland, D. Cotton, L. Hauser, and M. Keller. Targeted access to the genomes of low-abundance organisms in complex microbial communities. *Appl. Environ. Microbiol.*, 2007. DOI: 10.1128/AEM.02985-06.

Pope, P. B., S. E. Denman, M. Jones, S. G. Tringe, K. Barry, S. A. Malfatti, A. C. McHardy, J. F. Cheng, P. Hugenholtz, C. S. McSweeney, and M. Morrison. Adaptation to herbivory by the Tammar wallaby includes bacterial and glycoside hydrolase profiles different from other herbivores. *Proc. Natl. Acad. Sci. U.S.A.*, 2010. DOI: 10.1073/pnas.1005297107.

Prjibelski, A. D., I. Vasilinetc, A. Bankevich, A. Gurevich, T. Krivosheeva, S. Nurk, S. Pham, A. Korobeynikov, A. Lapidus, and P. A. Pevzner. ExSPAnder: a universal repeat resolver for DNA fragment assembly. *Bioinformatics*, 2014. DOI: 10.1093/bioinformatics/btu266.

Quinlan, A. R. and I. M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 2010. DOI: 10.1093/bioinformatics/btq033.

Rinke, C., P. Schwientek, A. Sczyrba, N. N. Ivanova, I. J. Anderson, J. F. Cheng, A. Darling, S. Malfatti, B. K. Swan, E. A. Gies, J. A. Dodsworth, B. P. Hedlund, G. Tsiamis, S. M. Sievert, W. T. Liu, J. A. Eisen, S. J. Hallam, N. C. Kyrpides, R. Stepanauskas, E. M. Rubin, P. Hugenholtz, and T. Woyke. Insights into

the phylogeny and coding potential of microbial dark matter. *Nature*, 2013. DOI: 10.1038/nature12352.

Rinke, C., J. Lee, N. Nath, D. Goudeau, B. Thompson, N. Poulton, E. Dmitrieff, R. Malmstrom, R. Stepanauskas, and T. Woyke. Obtaining genomes from uncultivated environmental microorganisms using FACS-based single-cell genomics. *Nat Protoc*, 2014. DOI: 10.1038/nprot.2014.067.

Ritari, J., K. Koskinen, J. Hultman, J. M. Kurola, M. Kymäläinen, M. Romantschuk, L. Paulin, and P. Auvinen. Molecular analysis of meso- and thermophilic microbiota associated with anaerobic biowaste degradation. *BMC Microbiol.*, 2012. DOI: 10.1186/1471-2180-12-121.

Rodrigue, S., R. R. Malmstrom, A. M. Berlin, B. W. Birren, M. R. Henn, and S. W. Chisholm. Whole genome amplification and de novo assembly of single bacterial cells. *PLoS ONE*, 2009. DOI: 10.1371/journal.pone.0006864.

Ronen, R., C. Boucher, H. Chitsaz, and P. Pevzner. SEQuel: improving the accuracy of genome assemblies. *Bioinformatics*, 2012. DOI: 10.1093/bioinformatics/bts219.

Safonova, Y., A. Bankevich, and P. A. Pevzner. dipSPAdes: Assembler for Highly Polymorphic Diploid Genomes. *J. Comput. Biol.*, 2015. DOI: 10.1089/cmb.2014.0153.

Salter, S. J., M. J. Cox, E. M. Turek, S. T. Calus, W. O. Cookson, M. F. Moffatt, P. Turner, J. Parkhill, N. J. Loman, and A. W. Walker. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.*, 2014. DOI: 10.1186/s12915-014-0087-z.

Salzberg, S. L., A. M. Phillippy, A. Zimin, D. Puiu, T. Magoc, S. Koren, T. J. Treangen, M. C. Schatz, A. L. Delcher, M. Roberts, G. Marcais, M. Pop, and J. A. Yorke. GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Res.*, 2012. DOI: 10.1101/gr.131383.111.

Schlüter, A., T. Bekel, N. N. Diaz, M. Dondrup, R. Eichenlaub, K. H. Gartemann, I. Krahn, L. Krause, H. Krömeke, O. Kruse, J. H. Mussgnug, H. Neuweger, K. Niehaus, A. Pühler, K. J. Runte, R. Szczepanowski, A. Tauch, A. Tilker, P. Viehöver, and A. Goesmann. The metagenome of a biogas-producing microbial community of a production-scale biogas plant fermenter analysed by the 454-pyrosequencing technology. *J. Biotechnol.*, 2008. DOI: 10.1016/j.jbiotec.2008.05.008.

Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 2014. DOI: 10.1093/bioinformatics/btu153.

Sharon, I. and J. F. Banfield. Genomes from metagenomics. *Science*, 2013. DOI: 10.1126/science.1247023.

Song, L., L. Florea, and B. Langmead. Lighter: fast and memory-efficient sequencing error correction without counting. *Genome Biol.*, 2014. DOI: 10.1186/s13059-014-0509-9.

Stepanauskas, R. Single cell genomics: an individual look at microbes. *Curr. Opin. Microbiol.*, 2012. DOI: 10.1016/j.mib.2012.09.001.

Stolze, Y., M. Zakrzewski, I. Maus, F. Eikmeyer, S. Jaenicke, N. Rottmann, C. Siebner, A. Pühler, and A. Schlüter. Comparative metagenomics of biogas-producing microbial communities from production-scale biogas plants operating under wet or dry fermentation conditions. *Biotechnol Biofuels*, 2015. DOI: 10.1186/s13068-014-0193-8.

Stolze, Y., A. Bremges, M. Rumming, C. Henke, I. Maus, A. Pühler, A. Sczyrba, and A. Schlüter. Identification and genome reconstruction of abundant distinct taxa in microbiomes from one thermophilic and three mesophilic production-scale biogas plants. *Biotechnol Biofuels*, 2016. DOI: 10.1186/s13068-016-0565-3.

Teeling, H., J. Waldmann, T. Lombardot, M. Bauer, and F. O. Glockner. TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics*, 2004. DOI: 10.1186/1471-2105-5-163.

Tennessen, K., E. Andersen, S. Clingenpeel, C. Rinke, D. S. Lundberg, J. Han, J. L. Dangl, N. Ivanova, T. Woyke, N. Kyrpides, and A. Pati. ProDeGe: a computational protocol for fully automated decontamination of genomes. *ISME J*, 2016. DOI: 10.1038/ismej.2015.100.

Turaev, D. and T. Rattei. High definition for systems biology of microbial communities: metagenomics gets genome-centric and strain-resolved. *Curr. Opin. Biotechnol.*, 2016. DOI: 10.1016/j.copbio.2016.04.011.

Ukkonen, E. Approximate string-matching with q-grams and maximal matches. *Theor. Comput. Sci.*, 1992. DOI: 10.1016/0304-3975(92)90143-4.

Varghese, N. J., S. Mukherjee, N. Ivanova, K. T. Konstantinidis, K. Mavrommatis, N. C. Kyrpides, and A. Pati. Microbial species delineation using whole genome sequences. *Nucleic Acids Res.*, 2015. DOI: 10.1093/nar/gkv657.

Walker, B. J., T. Abeel, T. Shea, M. Priest, A. Abouelliel, S. Sakthikumar, C. A. Cuomo, Q. Zeng, J. Wortman, S. K. Young, and A. M. Earl. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE*, 2014. DOI: 10.1371/journal.pone.0112963.

Warnecke, F., P. Luginbuhl, N. Ivanova, M. Ghassemian, T. H. Richardson, J. T. Stege, M. Cayouette, A. C. McHardy, G. Djordjevic, N. Aboushadi, R. Sorek, S. G. Tringe, M. Podar, H. G. Martin, V. Kunin, D. Dalevi, J. Madejska, E. Kirton, D. Platt, E. Szeto, A. Salamov, K. Barry, N. Mikhailova, N. C. Kyrpides, E. G. Matson, E. A. Ottesen, X. Zhang, M. Hernandez, C. Murillo, L. G. Acosta, I. Rigoutsos, G. Tamayo, B. D. Green, C. Chang, E. M. Rubin, E. J. Mathur, D. E. Robertson, P. Hugenholtz, and J. R. Leadbetter. Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature*, 2007. DOI: 10.1038/nature06269.

Weiland, P. Biogas production: current state and perspectives. *Appl. Microbiol. Biotechnol.*, 2010. DOI: 10.1007/s00253-009-2246-7.

Wibberg, D., V. M. Luque-Almagro, I. Igeño, A. Bremges, D. Roldán, F. Merchán, L. P. Sáez, I. Guijo, I. Manso, D. Macías, P. Cabello, G. Becerra, I. Ibáñez, I. Carmona, M. P. Escribano, F. Castillo, A. Sczyrba, C. Moreno-Vivián, R. Blasco, A. Pühler, and A. Schlüter. Complete genome sequence of the cyanide-degrading bacterium Pseudomonas pseudoalcaligenes CECT5344. *J. Biotechnol.*, 2014. DOI: 10.1016/j.jbiotec.2014.02.004.

Wibberg, D., A. Bremges, T. Dammann-Kalinowski, I. Maus, I. Igeño, R. Vogelsang, C. König, V. M. Luque-Almagro, D. Roldán, A. Sczyrba, C. Moreno-Vivián, R. Blasco, A. Pühler, and A. Schlüter. Finished genome sequence and methylome of the cyanide-degrading Pseudomonas pseudoalcaligenes strain CECT5344 as resolved by single-molecule real-time sequencing. *J. Biotechnol.*, 2016. DOI: 10.1016/j.jbiotec.2016.04.008.

Wilson, M. C., T. Mori, C. Rückert, A. R. Uria, M. J. Helf, K. Takada, C. Gernert, U. A. Steffens, N. Heycke, S. Schmitt, C. Rinke, E. J. Helfrich, A. O. Brachmann, C. Gurgui, T. Wakimoto, M. Kracht, M. Crüsemann, U. Hentschel, I. Abe, S. Matsunaga, J. Kalinowski, H. Takeyama, and J. Piel. An environmental bacterial taxon with a large and distinct metabolic repertoire. *Nature*, 2014. DOI: 10.1038/nature12959.

Wirth, R., E. Kovács, G. Maróti, Z. Bagi, G. Rákhely, and K. L. Kovács. Characterization of a biogas-producing microbial community by short-read next generation DNA sequencing. *Biotechnol Biofuels*, 2012. DOI: 10.1186/1754-6834-5-41.

Woese, C. R. and G. E. Fox. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl. Acad. Sci. U.S.A.*, 1977. DOI: 10.1073/pnas.74.11.5088.

Woese, C. R., O. Kandler, and M. L. Wheelis. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. U.S.A.*, 1990. DOI: 10.1073/pnas.87.12.4576.

Wooley, J. C., A. Godzik, and I. Friedberg. A primer on metagenomics. *PLoS Comput. Biol.*, 2010. DOI: 10.1371/journal.pcbi.1000667.

Woyke, T., D. Tighe, K. Mavromatis, A. Clum, A. Copeland, W. Schackwitz, A. Lapidus, D. Wu, J. P. McCutcheon, B. R. McDonald, N. A. Moran, J. Bristow, and J. F. Cheng. One bacterial cell, one complete genome. *PLoS ONE*, 2010. DOI: 10.1371/journal.pone.0010314.

Woyke, T., A. Sczyrba, J. Lee, C. Rinke, D. Tighe, S. Clingenpeel, R. Malmstrom, R. Stepanauskas, and J. F. Cheng. Decontamination of MDA reagents for single cell whole genome amplification. *PLoS ONE*, 2011. DOI: 10.1371/journal.pone.0026161.

Wu, Y. W., Y. H. Tang, S. G. Tringe, B. A. Simmons, and S. W. Singer. MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome*, 2014. DOI: 10.1186/2049-2618-2-26.

Yarza, P., P. Yilmaz, E. Pruesse, F. O. Glockner, W. Ludwig, K. H. Schleifer, W. B. Whitman, J. Euzeby, R. Amann, and R. Rossello-Mora. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat. Rev. Microbiol.*, 2014. DOI: 10.1038/nrmicro3330.

Youssef, N. H., P. C. Blainey, S. R. Quake, and M. S. Elshahed. Partial genome assembly for a candidate division OP11 single cell from an anoxic spring (Zodletone Spring, Oklahoma). *Appl. Environ. Microbiol.*, 2011. DOI: 10.1128/AEM.06059-11.

Zakrzewski, M., A. Goesmann, S. Jaenicke, S. Jünemann, F. Eikmeyer, R. Szczepanowski, W. A. Al-Soud, S. Sørensen, A. Pühler, and A. Schlüter. Profiling of the metabolically active community from a production-scale biogas plant by means of high-throughput metatranscriptome sequencing. *J. Biotechnol.*, 2012. DOI: 10.1016/j.jbiotec.2012.01.020.

Ziembińska-Buczyńska, A., A. Banach, T. Bacza, and M. Pieczykolan. Diversity and variability of methanogens during the shift from mesophilic to thermophilic conditions while biogas production. *World J. Microbiol. Biotechnol.*, 2014. DOI: 10.1007/s11274-014-1731-z.

Zong, C., S. Lu, A. R. Chapman, and X. S. Xie. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science*, 2012. DOI: 10.1126/science.1229164.