Identification of mobile elements in metagenomic data.

Josef W. Moser[2], Samuel M. Gerner[1,3], and Alexandra B. Graf[1,*]

[1] University of Applied Sciences FH Campus Wien, Department Bioengineering, Vienna, Austria

[2] Austrian Centre of Industrial Biotechnology (ACIB), Vienna, Austria

[3] Division of Computational System Biology, Department of Microbiology and Ecosystem Science, University of Vienna, Vienna, Austria

*Corresponding author

Introduction

Metagenomics gives us the unique opportunity to study aspects of the microbial community which were up to now hidden from the scientific eye. Several environments have been extensively sampled in the past few years, and most of all the human body itself. In relation to the human microbiome research, urban metagenomics is still an understudied but not less important field. We, as humans, are embedded in a complex system with which we and our microbiome interacts every day. Not surprisingly, the human exposome, the entirety of environmental influences we are exposed to, is shifting into the spotlight of attention. Scientists as well as decision makers begin to understand the potential for disease prevention that such research harbours.

Considering the microbiological exposome, one of the key factors is how much environmental microbial populations we interact with are influencing our microbiome. One popular and important example is antibiotics resistance (Finley et al. 2013). Such resistances are partly transferred through horizontal gene transfer to pathogenic bacteria, posing a serious health risk (Courvalin 2016). Horizontal gene transfer through mobile elements, such as plasmids, is a very common evolutionary vehicle within the bacterial domain and usually confers genes that give an evolutionary advantage but are not essential for the metabolism of the cell (Courvalin 2016; Stanisich 1984). The spread of degradative pathways and pathogenicity determinants of pathogens is mediated by plasmids as well (Smalla et al. 2015), and might also take influence on human health and well feeling.
Therefore, to investigate how microbial populations in the environment interact with the human microbiome, mobile elements are the key. Several ways exist through which bacteria can exchange genetic material, these include plasmids, phage integration sites, integrons or transposons (de la Cruz & Davies 2000). Plasmids are especially interesting since they also play an important role in the formation of biofilms (Cook & Dunny 2014) CRISPR cluster give additional information about the plasmids and phages the organisms was exposed to (Jackson et al. 2017).

For the analysis and annotation of metagenome samples, mobile elements pose several risks and challenges. In a database alignment they may lead to misclassification, due to their species independent distribution and in de-novo assemblies their repetitive nature interferes with with the assembly process (Smalla et al. 2015). In the years since the next generation sequencing boom, we have discovered that there is a much greater variety in the composition and functioning of microbial genomes. Metagenomes present an even greater pool of unknown species, determining the mobilom of such samples will improve our understanding of microbial populations and may lead to the discovery of unknown replication systems used by these communities (Smalla et al. 2015; Jørgensen et al. 2014).

Aim of the study

- Using assemblies from metagenome samples we want to predict mobile elements, especially focusing on plasmids
- Predict CRISPR regions in metagenome samples, evaluating if assembly based methods or read based methods show greater potential for CRISPR prediction.
- Comparison between cities and estimation of the impact of sequencing parameters on the prediction of mobile elements

In our analysis we first assembled the metagenome data, predicting the presence of plasmids within the assembled dataset. We also scanned for transposons, phage integration sites and integrons, within the assembled as well as the read sequences. In addition, we evaluated the state of phage and plasmid exposure with the identification of CRISPR regions.

Plasmids can be experimentally extracted from metagenome samples and sequenced, which gives a better and more concise result. We were still interested if mobile elements could be extracted from an urban metagenome sample for the following reasons.

- With routine metagenome swabbing of cities, how much of the mobilom can we really see in a standard sample. Is it possible to use such samples in the common sequencing depth and quality for a mobilome prediction?
- Antibiotics resistance prediction for metagenome samples is common, but how well can we relate the predicted antibiotic resistances to horizontal gene transfer?

Preliminary results:

From the tested assemblers (MetaSPAdes (Nurk et al. 2017) , MEGAHIT (Li et al. 2016), MINIA (Chikhi & Rizk 2012), IDBA_UD (Sharan 2014)), MetaSPAdes and MEGAHIT performed best, regarding the common statistical parameters like assembly size, N50, and re-alignment rate. Due to the extensive runtime of MetaSPAdes and time constraints, only assemblies of MEGAHIT were used for generating the preliminary results.

Plasmid prediction was approached by the search for cyclic contigs in the assembly graphs, using the program Recycler (Rozov et al. 2017). The length of the predicted contigs ranges from ~1000 bp to 23191 bp (Boston), and 71650 bp respectively (Sacramento). The mean contig length was found to be around 2000 bp, indicating a strong bias towards shorter plasmids. Some of these contigs could be assigned to plasmid sequences available in NCBI and EBI databases. Potential plasmid candidates in the unassigned contigs, were identified by predicting proteins for all candidate plasmids with Prodigal (Hyatt et al. 2010) and using the program hmmscan to search for plasmid specific protein domains. Additional to the de-novo assembly approach, the quality processed reads were aligned against a database of known plasmids.

The prediction of CRISPR regions in the assembled contigs was performed with MinCED (https://github.com/ctSkennerton/minced). Since CRISPR regions are likely not correctly assembled in Illumina data, due to repeat regions and recurring spacer regions, prediction was performed based on input reads, using CRASS (Skennerton et al. 2013). Results showed significant higher diversity than those proposed by MinCED. Functional assignment of the spacer sequences is still pending in the current state of the project.

SACRAMENTO	MEGAHIT	Recycler	Prodigal/Hmmer	Blast	Crass	MinCED
	assembly size	No. cycl. contigs	Rep Gene	NCBI/EBI DB	CRISPR	CRISPR
			motivs		prediction	prediction
Sample1A	101759459	47	4	5	110	5
Sample1B	126154418	65	8	5	130	17
Sample1C	82746339	68	1	2	102	8
Sample2A	71245988	50	5	1	105	23
Sample2B	86167171	46	4	1	137	9
Sample2C	101684635	41	0	0	127	15
Sample3A	58878533	49	0	0	127	10
Sample3B	808994	0	0	0	0	0
Sample3C	63080666	36	0	1	155	8
Sample4A	87916835	46	0	0	143	12
Sample4B	65034782	37	0	0	97	7
Sample4C	71416966	13	1	0	114	18
Sample5A	87172941	35	4	1	166	6
Sample5B	72564827	25	2	1	146	5
Sample5C	91833860	39	5	1	159	19
Sample6A	71600288	36	1	0	152	8
Sample6B	64259613	43	4	0	132	9
Sample6C	119297299	44	0	1	123	22

BOSTON / WGS	MEGAHIT	Recycler	Prodigal/Hmmer	Blast	Crass	MinCED
	assembly size	No. cycl. contigs	Rep Gene motivs	NCBI/EBI DB	CRISPR	CRISPR
					prediction	prediction
SRR3545898	4511188	12	0	0	7	0
SRR3545910	3124585	16	0	0	15	0
SRR3545919	17170813	26	2	1	51	6
SRR3545934	5141883	11	0	0	44	1
SRR3545941	3749352	13	0	0	15	1
SRR3545948	12329956	22	0	0	39	2
SRR3545955	3497640	12	0	0	26	1
SRR3545963	10077399	24	1	1	40	1
SRR3546351	335977	6	0	0	2	1
SRR3546354	4354388	16	3	2	13	0
SRR3546356	5370801	16	3	2	22	1
SRR3546358	4738060	17	0	0	21	1
SRR3546361	15642223	19	0	0	66	4
SRR3546363	8235861	18	0	0	37	4
SRR3546365	4399465	12	0	0	39	0
SRR3546367	20734110	31	0	0	82	1
SRR3546371	3865138	17	0	0	14	2
SRR3546373	3921175	14	1	0	24	0
SRR3546375	28626893	32	5	3	128	6
SRR3546378	735433	13	2	2	10	0
SRR3546380	12824446	28	0	0	56	4
SRR3546382	4538206	22	1	0	30	1
SRR3546384	414762	12	1	1	3	1
SRR3555059	6554397	14	0	0	52	0

Outlook:

Low abundances of circular sequences in the samples do not allow to make any conclusions at the current state of the project. Therefore, all processing steps, including pre-processing and assembly are constantly re-evaluated. Assemblies and reads will also be evaluated for other mobilome content.

References:

Chikhi, R. & Rizk, G., 2012. Space-Efficient and Exact de Bruijn Graph Representation Based on a Bloom Filter. In *Lecture Notes in Computer Science*. pp. 236–248.

Cook, L.C.C. & Dunny, G.M., 2014. The Influence of Biofilms in the Biology of Plasmids. *Microbiology Spectrum*, 2(5). Available at: http://dx.doi.org/10.1128/microbiolspec.plas-0012-2013.

Courvalin, P., 2016. Why is antibiotic resistance a deadly emerging disease? *Clinical microbiology and infection: the official publication of the European Society of Clinical Microbiology and Infectious Diseases*, 22(5), pp.405–407.

de la Cruz, F. & Davies, J., 2000. Horizontal gene transfer and the origin of species: lessons from bacteria. *Trends in microbiology*, 8(3), pp.128–133.

Finley, R.L. et al., 2013. The scourge of antibiotic resistance: the important role of the environment. *Clinical infectious diseases: an official publication of the Infectious Diseases Society of America*, 57(5), pp.704–710.

Hyatt, D. et al., 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics*, 11, p.119.

Jackson, S.A. et al., 2017. CRISPR-Cas: Adapting to change. *Science*, 356(6333). Available at: http://dx.doi.org/10.1126/science.aal5056.

Jørgensen, T.S. et al., 2014. Hundreds of Circular Novel Plasmids and DNA Elements Identified in a Rat Cecum Metamobilome. *PloS one*, 9(2), p.e87924.

Li, D. et al., 2016. MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* , 102, pp.3–11.

Nurk, S. et al., 2017. metaSPAdes: a new versatile metagenomic assembler. *Genome research*, 27(5), pp.824–834.

Rozov, R. et al., 2017. Recycler: an algorithm for detecting plasmids from de novo assembly graphs. *Bioinformatics* , 33(4), pp.475–482.

Sharan, R., 2014. *Research in Computational Molecular Biology: 18th Annual International Conference, RECOMB 2014, Pittsburgh, PA, USA, April 2-5, 2014, Proceedings*, Springer.

Skennerton, C.T., Imelfort, M. & Tyson, G.W., 2013. Crass: identification and reconstruction of CRISPR from unassembled metagenomic data. *Nucleic acids research*, 41(10), p.e105.

Smalla, K., Jechalke, S. & Top, E.M., 2015. Plasmid Detection, Characterization, and Ecology. *Microbiology spectrum*, 3(1), pp.PLAS–0038–2014.

Stanisich, V.A., 1984. 2 Identification and Analysis of Plasmids at the Genetic Level. In *Methods in Microbiology*. pp. 5–32.