



Universiteit
Leiden
The Netherlands

Metagenomics : beyond the horizon of current implementations and methods

Khachatryan, L.

Citation

Khachatryan, L. (2020, April 28). *Metagenomics : beyond the horizon of current implementations and methods*. Retrieved from <https://hdl.handle.net/1887/87513>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/87513>

Note: To cite this publication please use the final published version (if applicable).

Cover Page



Universiteit Leiden



The handle <http://hdl.handle.net/1887/87513> holds various files of this Leiden University dissertation.

Author: Khachatryan, L.

Title: Metagenomics : beyond the horizon of current implementations and methods

Issue Date: 2020-04-28

**Metagenomics:
Beyond the horizon of current
implementations and methods**

Lusine Khachatryan

This work is part of the research programme "Forensic Science" with project number 727.011.002, which is financed by the Dutch Research Council (NWO).

ISBN: 9789464020892

Cover Artwork: Alessandra Sequeira

Printing: GILDEPRINT, www.gildeprint.nl

© Copyright 2020 by Lusine Khachatryan, all rights reserved.

No part of this book may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, without the prior permission of the author.

Metagenomics: Beyond the horizon of current implementations and methods

PROEFSCHRIFT

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van Rector Magnificus prof.mr. C.J.J.M. Stolker,
volgens besluit van het College voor Promoties
te verdedigen op dinsdag 28 april 2020
klokke 16:15 uur

door

Lusine Khachatryan
geboren te Jermuk, Armenië in 1990

Promotor: Prof. dr. P. de Knijff

Co-promotor: Dr. J. F. J. Laros

Leden promotiecommissie: Prof.dr. A. Geluk
Prof. dr. A. C. M. Kroes
Prof. dr. J. N. Kok¹
Dr. T. Sijen²

¹ Faculty of Electrical Engineering, University of Twente, Enschede, The Netherlands

² Netherlands Forensic Institute, The Hague, The Netherlands

Моему Папочке

Matthew 7:7 New International Version

*"Ask and it will be given to you; seek and you will find;
knock and the door will be opened to you"*

Contents

1	Introduction	11
1.1	Why metagenomics	12
1.2	Metagenomics sequencing data	15
1.2.1	Amplicon sequencing data	15
1.2.2	Whole genome sequencing data	16
1.3	Approaches used in metagenomics	17
1.3.1	Homology-based profiling	18
1.3.2	<i>De novo</i> profiling	23
1.3.3	Mixed profiling	24
1.3.4	Reference-free comparison of metagenomics data	25
1.4	The outline of this thesis	26
2	Taxonomic classification and abundance estimation using 16S and WGS - a comparison using controlled reference samples	27
2.1	Background	28
2.2	Materials and Methods	30
2.2.1	DNA extraction and concentration measurement	30
2.2.2	Metagenomic mixes creation	30
2.2.3	WGS sequencing library creation	31
2.2.4	16S sequencing library creation	31
2.2.5	DNA sequencing	31
2.2.6	Bacterial genomes assembly	32
2.2.7	Regression analysis	32
2.2.8	Analysis using Centrifuge	32
2.2.9	Analysis using MG-RAST	33
2.2.10	Taxa abundance estimation and results evaluation	33
2.2.11	Statistical and correlation analysis	34
2.3	Results and Discussions	35
2.3.1	Individual bacterial genomes assembly	35
2.3.2	Estimation of reference abundances	35
2.3.3	Analysis of bacterial mixes using Centrifuge and MG-RAST	37

2.3.4	Profiling accuracy without considering relative abundances	39
2.3.5	Abundance assignment accuracy	39
2.4	Conclusions	45
2.5	Author Statements	49
2.5.1	Funding information	49
2.5.2	Authors' contributions	49
2.5.3	Acknowledgements	49
2.5.4	Conflicts of interest	49
2.6	Data Availability	49
3	Reference-free resolving of long-read metagenomic data	51
3.1	Background	52
3.2	Materials and Methods	54
3.2.1	Software	54
3.2.2	PacBio data simulation	54
3.2.3	Bioreactor metagenome PacBio sequencing	54
3.2.4	Reads origin checking	55
3.2.5	Bioreactor metagenome PacBio reads assembly	55
3.2.6	Binning procedure	55
3.2.7	Classification for larger sets	57
3.2.8	Data availability	58
3.3	Results	59
3.3.1	Reads classification in artificial PacBio metagenomes	59
3.3.2	PacBio sequencing of bioreactor metagenome	60
3.3.3	Bioreactor metagenome PacBio read classification	60
3.3.4	Assembly of the bioreactor metagenome before and after reads binning	64
3.4	Discussion	65
3.5	Author Statements	67
3.5.1	Funding information	67
3.5.2	Acknowledgements	67
3.5.3	Conflicts of interest	67
4	Determining the quality and complexity of next-generation sequencing data without a reference genome	69
4.1	Background	70
4.2	Materials and Methods	71
4.2.1	kPAL implementation	71
4.2.2	Creating k -mer profiles	71
4.2.3	Measuring pairwise distances	72
4.2.4	Calculating the k -mer balance	72
4.2.5	Statistical analysis	72

4.2.6	Library preparation and sequencing	72
4.2.7	Pre-processing	73
4.2.8	Alignment	73
4.2.9	SGA	74
4.2.10	Data availability	74
4.3	Results and Discussion	75
4.3.1	Principles of kPAL	75
4.3.2	Setting k size	75
4.3.3	Evaluating data quality without a reference	78
4.3.4	Comparative analysis of kPAL performance	82
4.3.5	Detecting data complexity	85
4.4	Conclusions	88
4.5	Appendix	90
4.6	Abbreviations	90
4.7	Competing interests	90
4.8	Authors' contributions	90
4.9	Acknowledgements	91
5	BacTag - a pipeline for fast and accurate gene and allele typing in bacterial sequencing data	93
5.1	Background	94
5.2	Materials and Methods	96
5.2.1	Pipeline implementation	96
5.2.2	Pipeline testing	99
5.2.3	Database	99
5.3	Results	103
5.3.1	Building the preprocessed MLST databases	103
5.3.2	Testing BacTag on artificial data	103
5.3.3	Testing BacTag on real <i>E. coli</i> and <i>K. pneumoniae</i> data	104
5.3.4	Comparing BacTag with web-based tools for <i>E. coli</i> Achtman MLST	106
5.4	Discussion	109
5.5	Conclusions	112
5.6	Abbreviations	113
5.7	Author Statements	113
5.7.1	Acknowledgements	113
5.7.2	Funding information	113
5.7.3	Availability of data and materials	113
5.7.4	Authors' contributions	114
5.7.5	Ethics approval and consent to participate	114
5.7.6	Competing interests	114

6	General discussion and possible future improvement	115
6.1	Who is inhabiting the microbiome?	116
6.2	How complex is the investigated microbiome?	117
6.3	How to compare different metagenomes?	118
6.4	What is the possible pathogenic impact of the metagenome?	119
	Bibliography	121
	Samenvatting	145
	Publications	149
	Acknowledgements	151
	Curriculum vitae	153