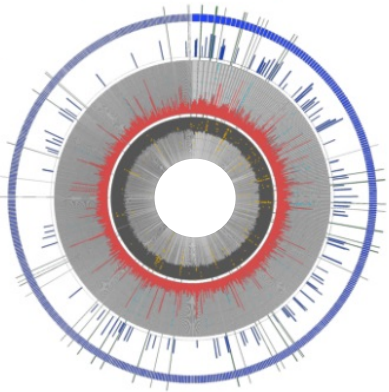
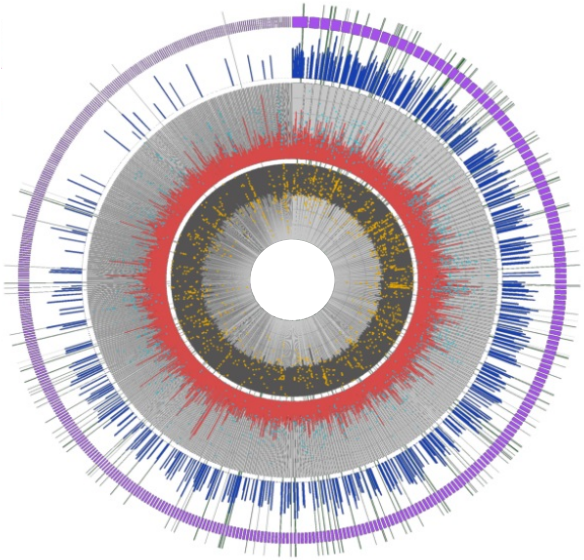
The study of microbial communities: Bioinformatics applications within the UL HPC environment



UL HPC school 2017
13 June 2017

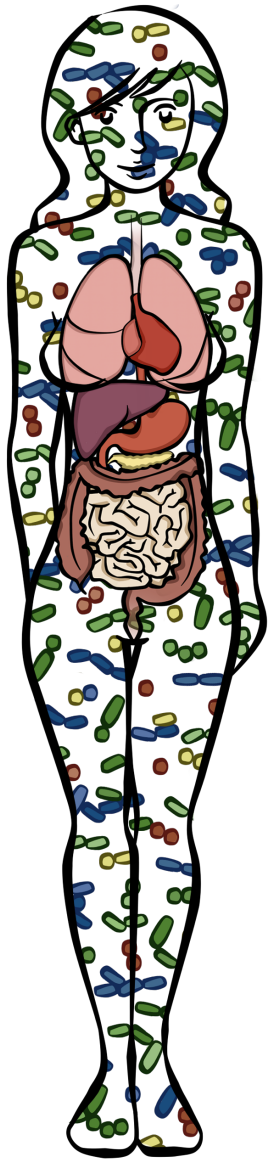
Shaman Narayanasamy
Eco-Systems Biology group of LCSB



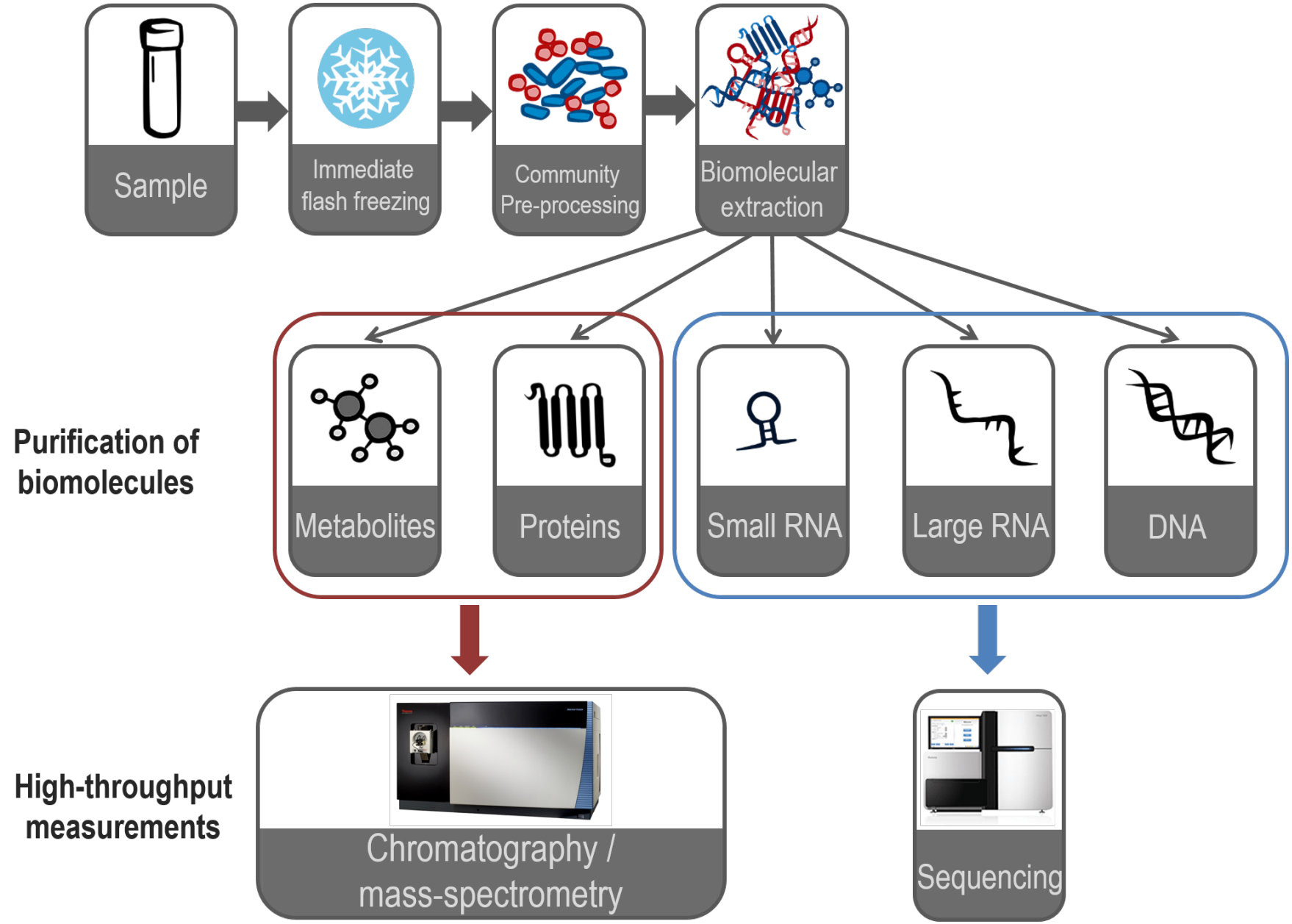
Fonds National de la Recherche Luxembourg

UNIVERSITÉ DU LUXEMBOURG

The subject: microbial communities

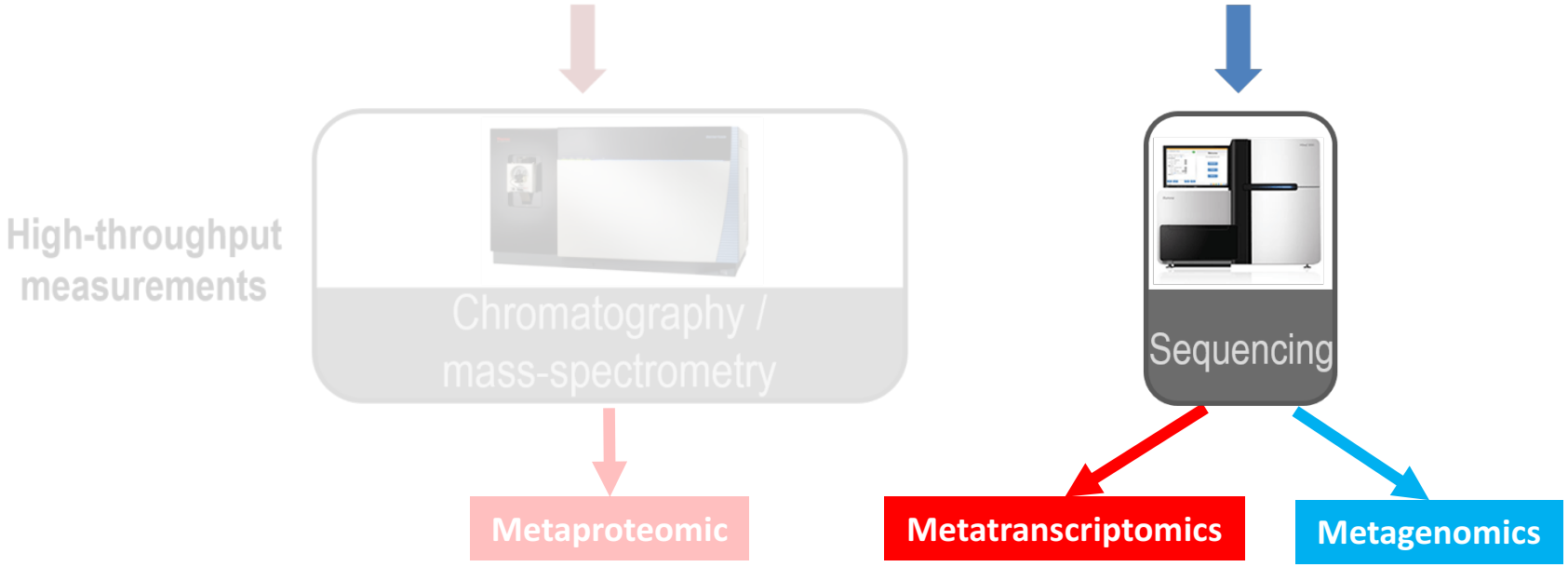


The samples: Biomolecules



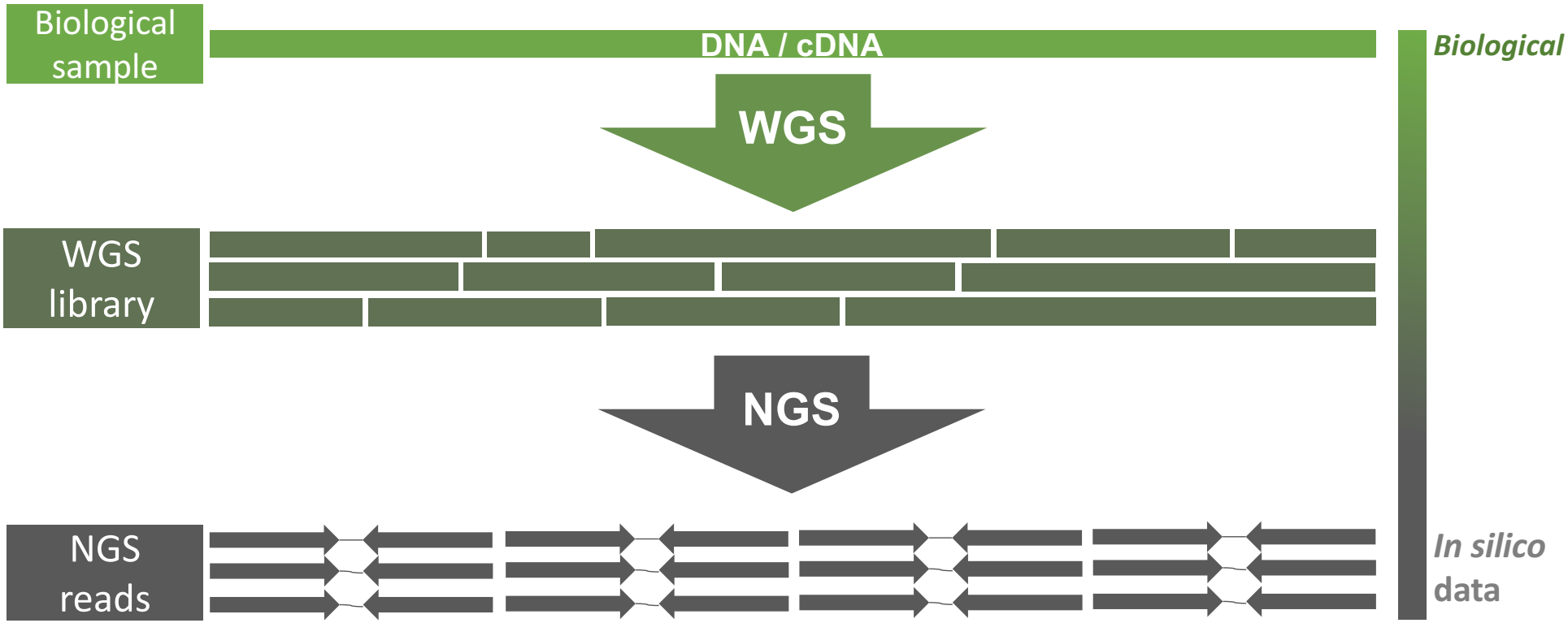
Roume *et al*. ISME J. (2013) **7**:110-21
Roume *et al*. Methods Enzymol. (2013) **531**:219-36

The measurements: High-throughput data



Roume *et al*. ISME J. (2013) **7**:110-21
Roume *et al*. Methods Enzymol. (2013) **531**:219-36

The measurements: Random shotgun sequencing

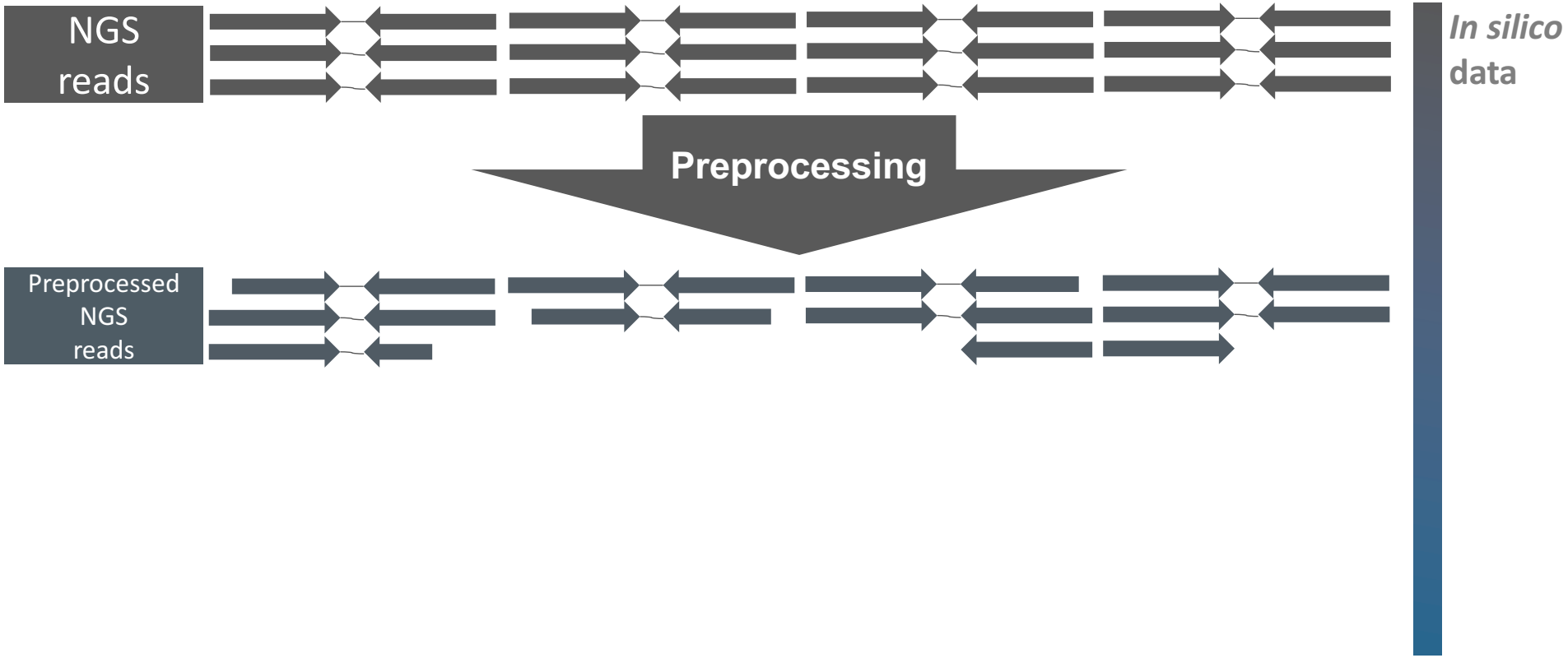


cDNA: complementary DNA
WGS: whole genome shotgun
NGS: next-generation sequencing

The data: Next-generation sequencing (NGS)

```
@HWUSI-EAS664L:30:637UKAAXX:7:1:4970:1139 1:N:0:TACTTCGG
AAANCATTGGGCAATTCAAAAATCTCAAATTCATTCATAAAGTTCTTTCCTTACTGTTTTATGGGTATTAAATCTGCTATATCAGCTTCTATCGATTGATTTA
+
7;8#;@CBACHHE@HHHHHHHDHHHHHFHHHGHHHHGHHBGGGGGHHHHHHHHHHHHBHHHFGGGEEGDGG@GGG>HHHFHBHHHHGGHHHGHHHHHHH8FGGB
@HWUSI-EAS664L:30:637UKAAXX:7:1:4172:1139 1:N:0:TACTTCGG
TATNACAGAGCAATCCCATGGCAAAAATAAAATGCAGGGCCAGTTTGGCTGGTTTAACATAGATGGCATCACCGGTTAAACCACTGGCCACCATGATCCAGCC
+
BBB#BFFFFFIIIIIIIIIIIIIIIHIEIIFIIIIIIIIIIIGIHHGGIGEGGBFGADEB?>CC?@FIH@IEGGGFBEEH>EGFFDFEEE3@AA@BB>EECEEG
@HWUSI-EAS664L:30:637UKAAXX:7:1:4116:1139 1:N:0:TACTTCGG
CGCNGGCAAAAGAGCATCAATTCGAATCTCACTCATAGGCTGTATCCTCATCTACGATATGTATTGTGAATTTTTTCACAACGCGAGTATATCAGAAGAAATA
+
BBB#?EDDDDHGHHHGHHHDHHHHGGDGGGG@DGAGDDGGHHCHHHHHHHHHF>EE<ACACA>CC@EGE3FEBCFFHDBGBD@BG140:*@;@?@=AD@@;.A
@HWUSI-EAS664L:30:637UKAAXX:7:1:7406:1139 1:N:0:TACTTCGG
CCCNAAAACCCCAAAACCCCAAAACCCCAAAACCCCTGTAAATTGAAAATAAATGTCAAAGCCAGTTGTGCTCACTGAAACTCAAGTGCAGCAGTGATGCAAG
+
BB@#BBBACCHHHGHHHHHHHHHGHHHHHDEGGGGHGHGHHHHHHF@HHHFHHFHEHH<FGDECEBCF@GBBECEDF@BBFCBC2A=A?4:=*=;?73>=55=
@HWUSI-EAS664L:30:637UKAAXX:7:1:5688:1139 1:N:0:TACTTCGG
AAANGTAAAGGAGGAGGATTCAGACAAACTTGCGTGCTCTGTCATTTTTGACAGCAAAATACGTTGCATGGACAGGATGGTTTGATACGGTAGTTTACCTTGC
+
??=#;FFEDBGGFGGGGGE?GGGGG?GGGEHFHBHHHHDHHHHHHHHHHH4DDGBGGDGGHHHBHHFHHH3<BBDCC8DE*16=+EAAE8;A@<A=;9B5AAF
@HWUSI-EAS664L:30:637UKAAXX:7:1:14717:1140 1:N:0:TACTTCGG
ATTNCCATTTGTTTTAACTGAATTAACAACTTTATTTCGGTTTTTATTAATCTTTACTGATTTATTCCCGGTTACTTTAGATTTATCATTGCAGCTATTGGAT
+
CCA#AFFFEFHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHGHHHHHHHHHHGHHHHHGHHGHHGHEHHGHHHBGGGBHHHHHHB>GHHHGFHDBG
@HWUSI-EAS664L:30:637UKAAXX:7:1:5080:1140 1:N:0:TACTTCGG
GCCNCACATGTCGAGGGTGTCGCAACCGGTCGCGCCCGCGTCGGTGACGATGTTGGCGTCGCCGTAGAGGATCTTTCCGGCGCGAAGCACGAGCGCGACATCC
+
CAA#AEEFFFHHHHHGGEGGGGGEGHHHHHHHHHHHHHHHHCGGG>GDGGBAGABGGGGCGEEBC@3A=AA>3C<A>EAACACAC2:??3B:,:??########
@HWUSI-EAS664L:30:637UKAAXX:7:1:2474:1140 1:N:0:TACTTCGG
CAANATCATCTTGGTGACGTTGCCAGAAGAAACGCCAGTCTCGCAAGCGGCGGCGCTGCAAGACGACCTGCGACGGGCCAGGGGCGAGCCCTTCGCCTGGGGG
+
BBB#?FFBDFGGGEEEGDGGGFGGGDAGGGGFGGDGDGGGGGGGFEDG@EGEGCEA+A##########################################
```

The process: NGS read preprocessing

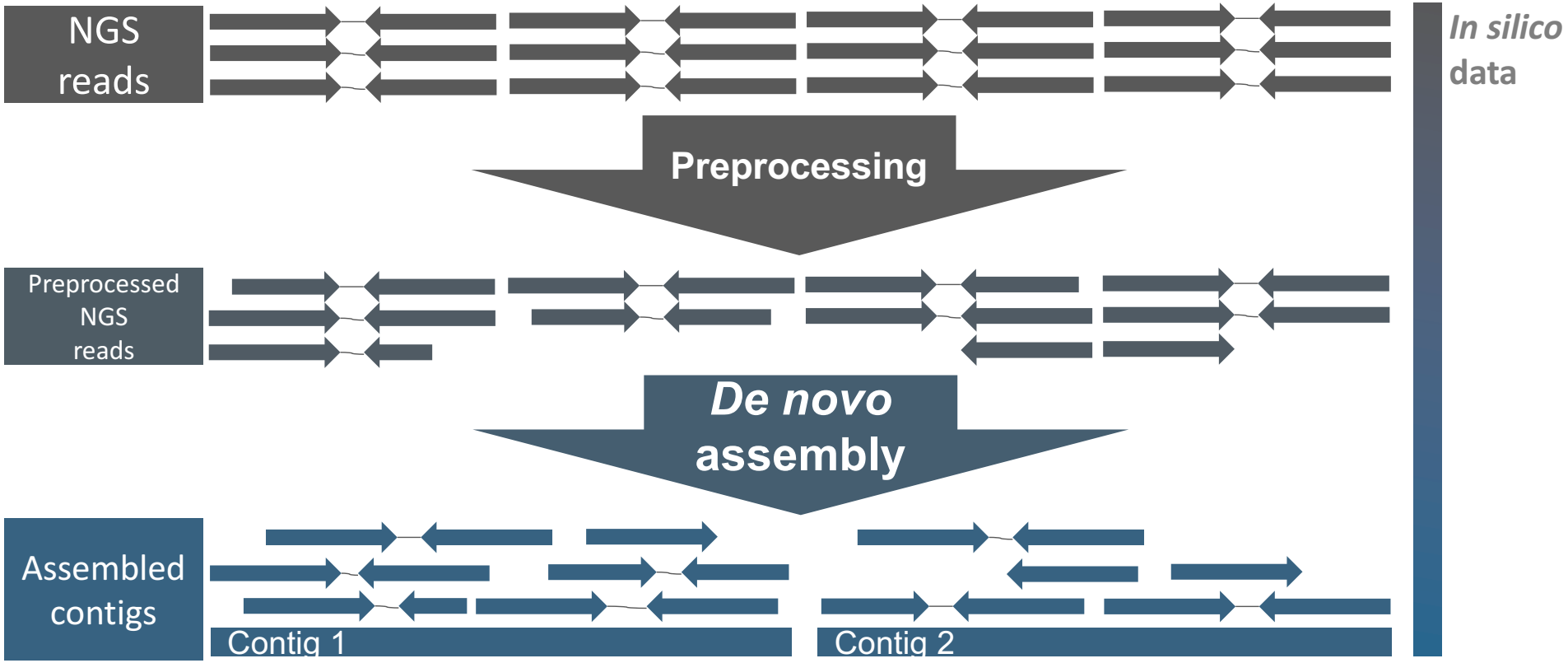


The process: NGS read preprocessing

Preprocessing

Trimmomatic
CutAdapt
SortMeRNA
*BWA
*Bowtie2

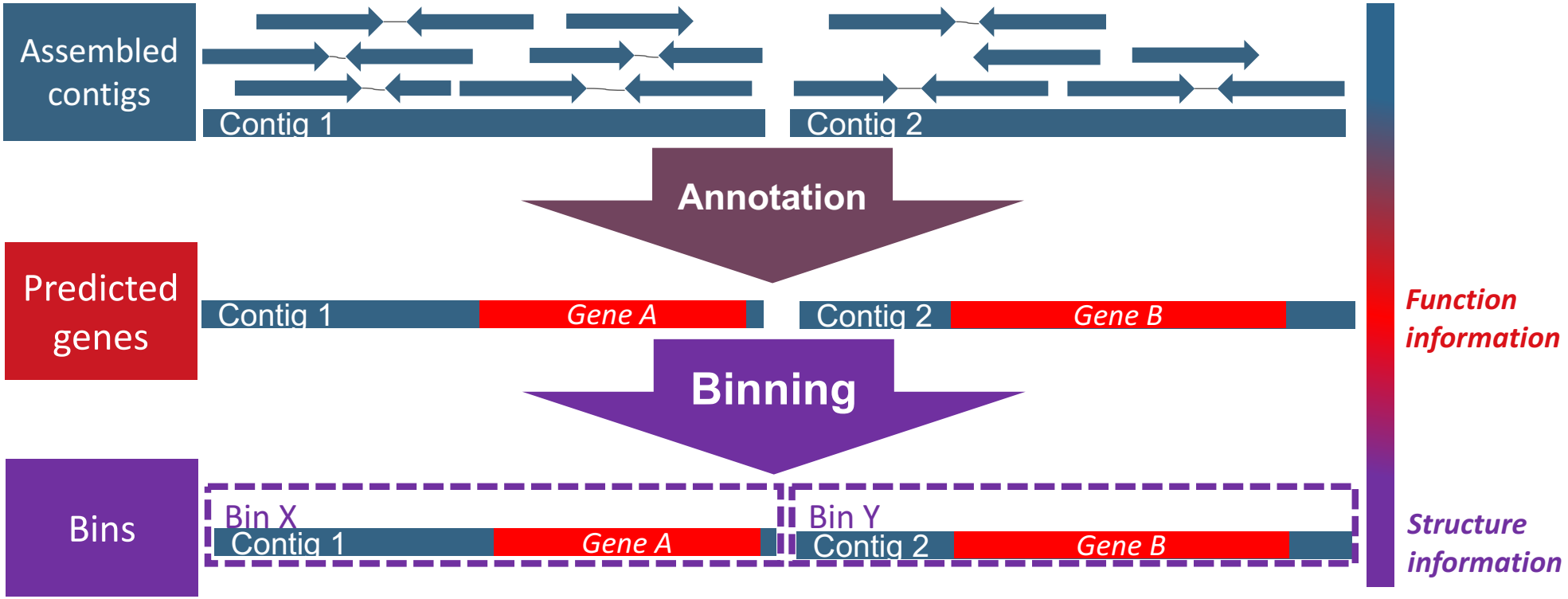
The process: *De novo* assembly



The process: *De novo* assembly

Preprocessing	Assembly
Trimmomatic	IDBA-UD
CutAdapt	MEGAHIT
SortMeRNA	SPAdes
*BWA	AbySS
*Bowtie2	Newbler
	Cap3

The process: Post-assembly analysis



The process: Post-assembly analysis

Preprocessing	Assembly	Post-assembly
Trimmomatic	IDBA-UD	BWA
CutAdapt	MEGAHIT	Bowtie2
SortMeRNA	SPAdes	MaxBin
*BWA	AbySS	dRep
*Bowtie2	Newbler	HMMer
	Cap3	BLASTn
		AMPHORA2
		PhyloPhlan

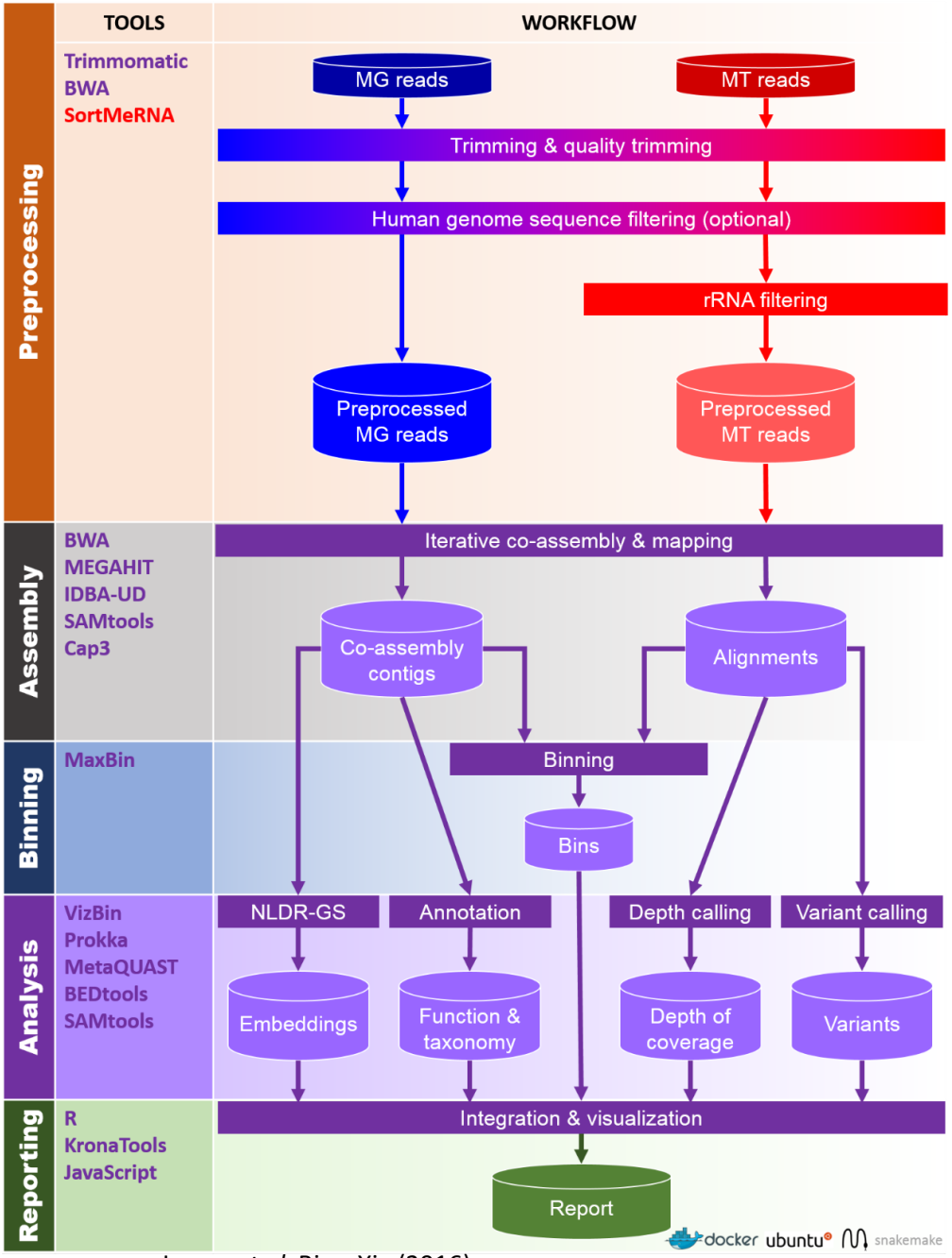
The process: Automation

Preprocessing	Assembly	Post-assembly	Automation
Trimmomatic	IDBA-UD	BWA	Bash
CutAdapt	MEGAHIT	Bowtie2	Make
SortMeRNA	SPAdes	MaxBin	Python
*BWA	AbySS	dRep	Perl
*Bowtie2	Newbler	HMMer	Galaxy
	Cap3	BLASTn	Snakemake
		AMPHORA2	CWL
		PhyloPhlan	Ruffus

The process: Reproducibility

Preprocessing	Assembly	Post-assembly	Automation	Containerization
Trimmomatic	IDBA-UD	BWA	Bash	Docker
CutAdapt	MEGAHIT	Bowtie2	Make	LXD
SortMeRNA	SPAdes	MaxBin	Python	Vigrant
*BWA	AbySS	dRep	Perl	*BioConda
*Bowtie2	Newbler	HMMer	Galaxy	
	Cap3	BLASTn	Snakemake	
		AMPHORA2	CWL	
		PhyloPhlan	Ruffus	

The process: Integrated meta-omics pipeline (IMP)



Original logo by Linda Wampach

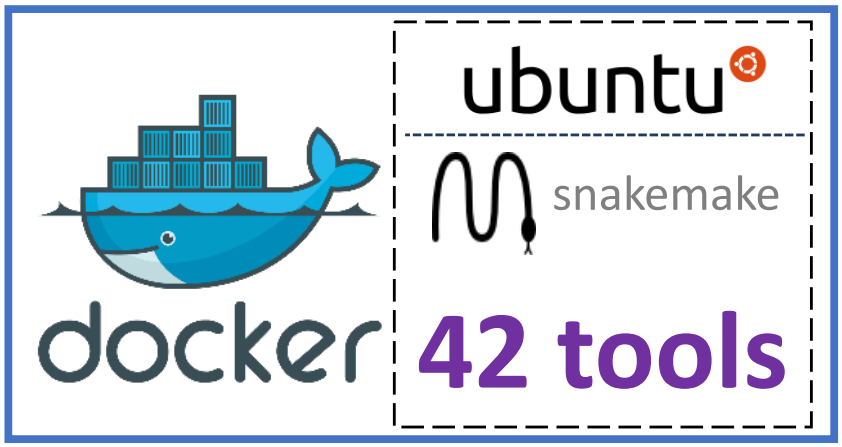
IMP available at:

http://r3lab.uni.lu/web/imp

Narayanasamy, Jarosz *et al*. BioarXiv (2016)

Narayanasamy, Jarosz *et al*. Genome Biology (2016)

The requirements, performance and output: In numbers

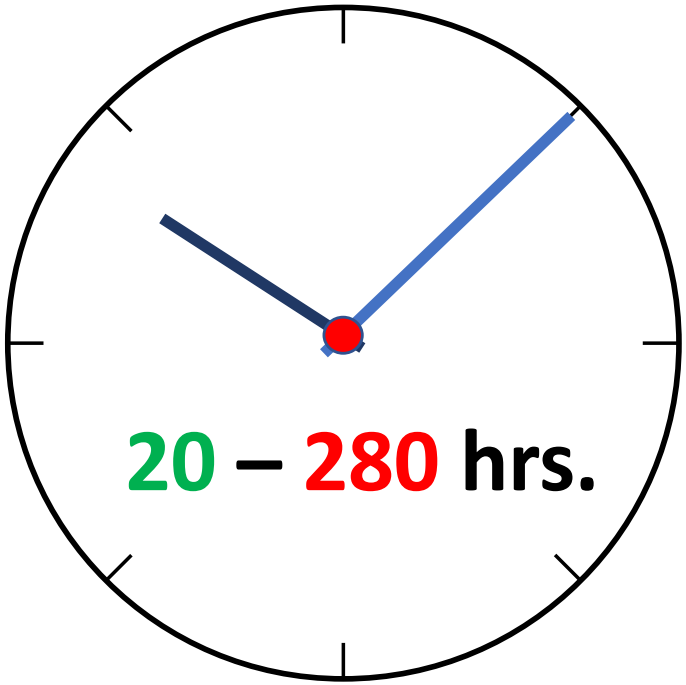


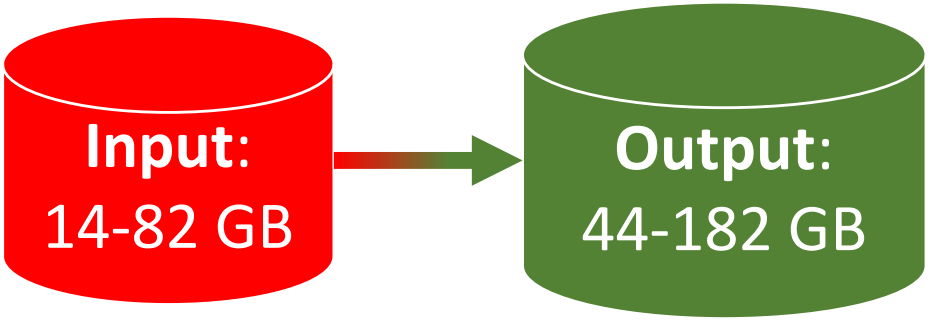
Computing platforms

- 8 cores
- 256 – 1024 GB RAM

amazon web services™

- r3.4xlarge
- 16 cores
- 122 GB





Narayanasamy, Jarosz *et al*. BioarXiv (2016)
Narayanasamy, Jarosz *et al*. Genome Biology (2016)

The outcome: Knowledge on microbial communities

ARTICLE

Received 15 Aug 2014 | Accepted 20 Oct 2014 | Published 26 Nov 2014 DOI: 10.1038/ncomms6603 OPEN

Community-integrated omics links dominance of a microbial generalist to fine-tuned resource usage

Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes

Narayanasamy *et al. Genome Biology* (2016) 17:260
DOI 10.1186/s13059-016-1116-8

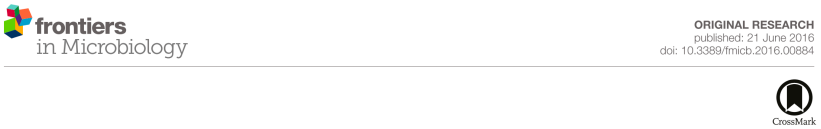
Genome Biology

SOFTWARE Open Access

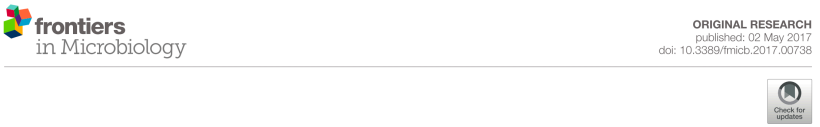
IMP: a pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses

ARTICLE OPEN

Comparative integrated omics: identification of key functionalities in microbial community-wide metabolic networks



Identification, Recovery, and Refinement of Hitherto Undescribed Population-Level Genomes from the Human Gastrointestinal Tract



OPEN ACCESS

Edited by:
Maria Carmen Collado,
Instituto de Agroquímica y Tecnología de Alimentos (CSIC), Spain

Reviewed by:
Carmen Portillo,

Colonization and Succession within the Human Gut Microbiome by Archaea, Bacteria, and Microeukaryotes during the First Year of Life

Muller, Pinel *et al*. Nature Communications (2014)
Roume, Heintz-Buschart *et al*. NPJ Microbiome and Biofilms (2015)
Laczny *et al*. Frontiers in Microbiology (2016)
Heintz-Buschart *et al*. Nature Microbiology (2016)
Narayanasamy, Jarosz *et al*. Genome Biology (2016)
Wampach *et al*. Frontiers in Microbiology (2017)
Kaysen *et al*. Translational Research (accepted)
Muller, Narayanasamy *et al*. Standards in Genomic Sciences (in review)
Wampach, Heintz-Buschart *et al*. (in preparation)
Herold *et al*. (in preparation)
Narayanasamy, Martinez-Arbas *et al*. (in preparation)

The outcome: AcKnowledge the HPC

Acknowledgements

We thank the scientists and technical staff of the Luxembourg Centre for Systems Biomedicine and TGen North, particularly Dr René Hussong and Dr Owain Moss for their assistance and support; Mr Bissen and Mr Di Pentima from the Syndicat Intercommunal à Vocation Ecologique (SIVEC), for their permission to collect samples and access to the monitoring platform of the Schifflange wastewater treatment plant; Joëlle Fritz, Sheri Simmons and Rudi Balling for their comments on the manuscript. Bioinformatic analyses presented in this paper were carried out using the high-performance computing facilities of the University of the Luxembourg. This work was

Acknowledgements

The authors thank the staff of the Clinical and Epidemiological Investigation Center (CIEC) Luxembourg for undertaking the sample and data collection from participants in this study. The authors thank B. Phillips for input and feedback during the conception of the study and S. Collignon, K. Greenhalgh, P. do Rosario Martins Conde and A. Kaysen for technical assistance with biomolecular extraction and quality control. The authors thank D. Baiwir and G. Mazucchelli (CART-GIGA) for measurements and assistance. The *in silico* analysis results presented in this Article were obtained using the high-performance computing facilities of the University of Luxembourg, whose administrators are acknowledged. The authors thank M. Brunkow at the Institute for Systems Biology (ISB), Seattle, who provided project management, and acknowledge support from the Family Genomics Group of Leroy Hood (ISB) for the human whole-genome sequencing data. The present work was

additional analyses were performed on the Gaia cluster of the University of Luxembourg HPC platform [78].

Acknowledgements

We would like to acknowledge John Larsson from SciLifeLab (Sweden) for kindly providing the KEGG-based functional Krona plot scripts. Albi Celaj from the University of Toronto is thanked for supplying the in silico simulated metatranscriptomic data and the corresponding reference genomes. The University of Luxembourg High Performance Computing (HPC) facility is duly thanked for providing and maintaining the computing platform. The Reproducible Research Results (R3) team of the Luxembourg Centre for

Code availability and computational resources

All in-house developed scripts are available from the authors upon request. *In silico* analysis results were obtained using the high performance computing facilities of the University of Luxembourg.[51]

species may allow community-wide control strategies to be formulated where other community-wide phenotypic outcomes may be desirable, e.g., in the human gastrointestinal tract. *In silico* analysis results presented in this paper were obtained using the high performance computing facilities of the University of Luxembourg[51].

ACKNOWLEDGMENT

In silico analyses presented in this paper were carried out using the HPC facilities of the University of Luxembourg (Varrette et al., 2014).

ACKNOWLEDGMENTS

In silico analyses presented in this paper were carried out using the HPC facilities of University of Luxembourg (Varrette et al., 2014). We are deeply grateful for all the parents and infants that have and will participate in the study. We also thank the dedicated clinical staff for participant recruitment, neonatologists of the Clinique pédiatrique and gynecologists, especially Isabelle

The outcome: AcKnowledge the HPC

Acknowledgements

We thank the scientists and technical staff of the Luxembourg Centre for Systems Biomedicine and TGen North, particularly Dr René Hussong and Dr Owain Moss for their assistance and support; Mr Bissen and Mr Di Pentima from the Syndicat Intercommunal à Vocation Ecologique (SIVEC), for their permission to collect samples and access to the monitoring platform of the Schifflange wastewater treatment plant; Joëlle Fritz, Sheri Simmons and Rudi Balling for their comments on the manuscript. Bioinformatic analyses presented in this paper were carried out using the high-performance computing facilities of the University of the Luxembourg. This work was

Acknowledgements

The authors thank the staff of the Clinical and Epidemiological Investigation Center (CIEC) Luxembourg for undertaking the sample and data collection from participants in this study. The authors thank B. Phillips for input and feedback during the conception of the study and S. Collignon, K. Greenhalgh, P. do Rosario Martins Conde and A. Kaysen for technical assistance with biomolecular extraction and quality control. The authors thank D. Baiwir and G. Mazucchelli (CART-GIGA) for measurements and assistance. The *in silico* analysis results presented in this Article were obtained using the high-performance computing facilities of the University of Luxembourg, whose administrators are acknowledged. The authors thank M. Brunkow at the Institute for Systems Biology (ISB), Seattle, who provided project management, and acknowledge support from the Family Genomics Group of Leroy Hood (ISB) for the human whole-genome sequencing data. The present work was

additional analyses were performed on the Gaia cluster of the University of Luxembourg HPC platform [78].

Acknowledgements

We would like to acknowledge John Larsson from SciLifeLab (Sweden) for kindly providing the KEGG-based functional Krona plot scripts. Albi Celaj from the University of Toronto is thanked for supplying the in silico simulated metatranscriptomic data and the corresponding reference genomes. The University of Luxembourg High Performance Computing (HPC) facility is duly thanked for providing and maintaining the computing platform. The Reproducible Research Results (R3) team of the Luxembourg Centre for

Code availability and computational resources

All in-house developed scripts are available from the authors upon request. *In silico* analysis results were obtained using the high performance computing facilities of the University of Luxembourg.[51]

species may allow community-wide control strategies to be formulated where other community-wide phenotypic outcomes may be desirable, e.g., in the human gastrointestinal tract. *In silico* analysis results presented in this paper were obtained using the high performance computing facilities of the University of Luxembourg[51].

ACKNOWLEDGMENT

In silico analyses presented in this paper were carried out using the HPC facilities of the University of Luxembourg (Varrette et al., 2014).

ACKNOWLEDGMENTS

In silico analyses presented in this paper were carried out using the HPC facilities of University of Luxembourg (Varrette et al., 2014). We are deeply grateful for all the parents and infants that have and will participate in the study. We also thank the dedicated clinical staff for participant recruitment, neonatologists of the Clinique pédiatrique and gynecologists, especially Isabelle

And in all **presentations/posters** in international conferences and **PhD theses**!

The experience: Continued improvement

- First impression: Impressed!

- Initial problems:
  - Learning curve
  - File system issues
  - Users “misbehaving”
  - Independent systems (bigbug compute node and storage “boxes”)
  - No dedicated system admin for LCSB

- Improvements over the years:
  - Solved file system issues
  - HPC school
  - Improved documentation
  - Well behaved users
  - Dedicated system admin for LCSB

- Additional request:
  - High-quality logo on HPC website for presentations

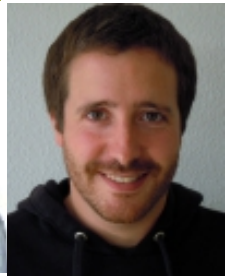
The future: Best practices and improvements

- Best practices:
  - (Try to) Be a good user; attend the HPC school
  - Incorporate cost of HPC into budgets/grants
  - Acknowledge the HPC (manuscripts, presentations)
  - Communicate effectively!
- Future practices and improvements:
  - Integration of independent machines with HPC
  - Reduce reliance on Docker
  - Better data management
  - Software management
  - Software benchmarking
  - *Dedicated personnel within group
  - Continuous learning!

Acknowledgements

Former ESBers:
Emilie Muller
Cedric Laczny
Abdul Sheik
Hugo Roume
Myriam Zeimes



Fonds National de la Recherche Luxembourg

HPC uni.lu UNIVERSITÉ DU LUXEMBOURG

LCSB

uni.lu UNIVERSITÉ DU LUXEMBOURG