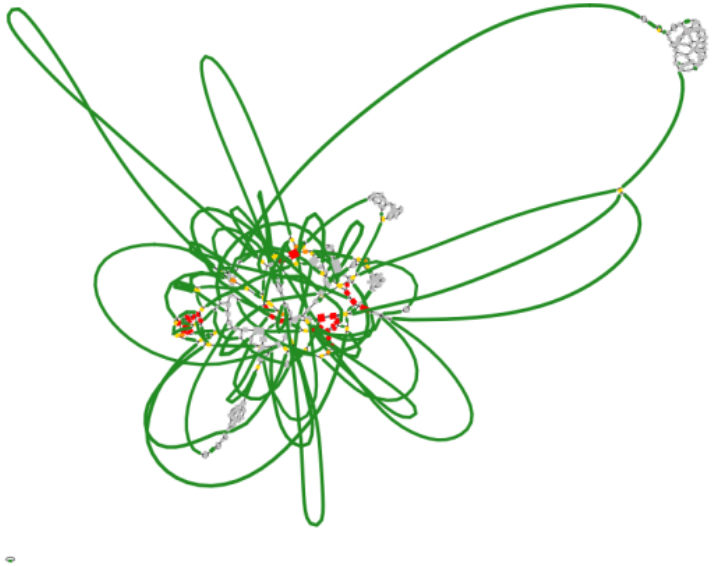
rijksuniversiteit groningen

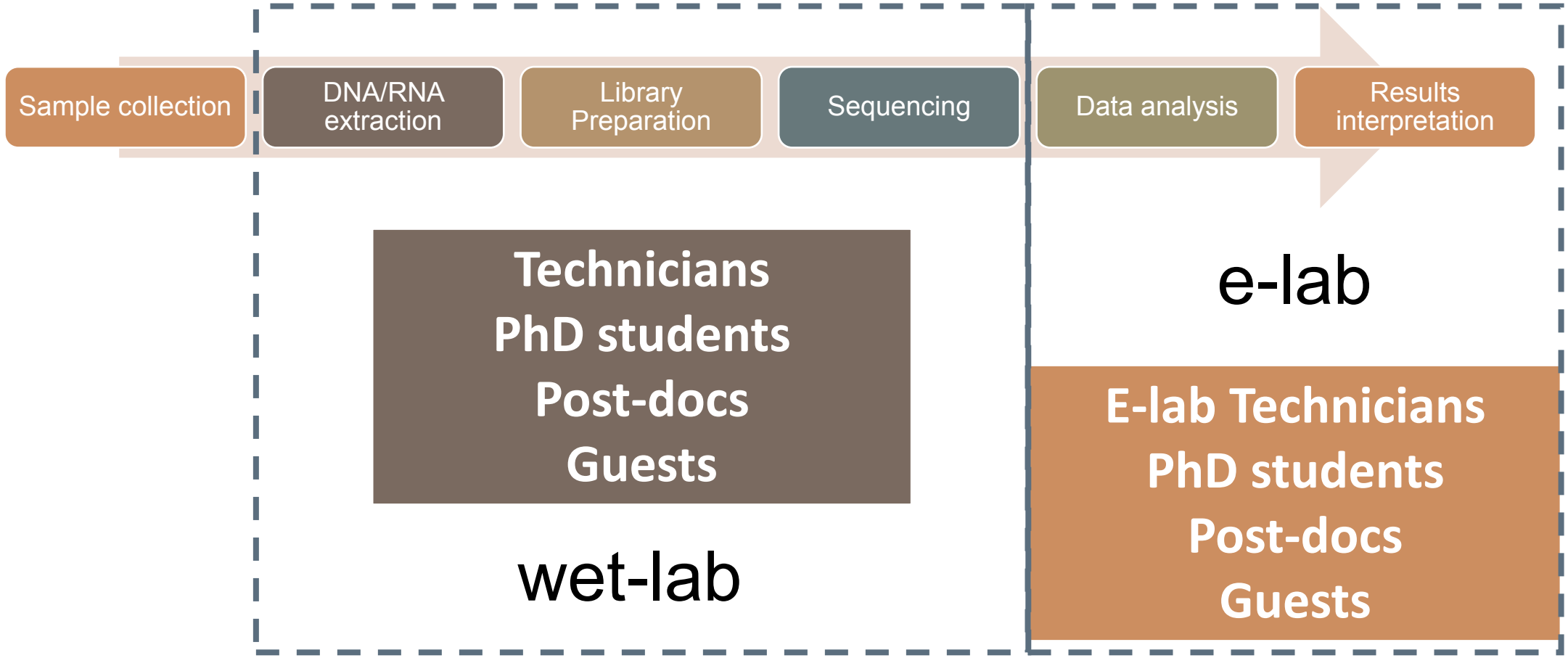
umcg

An update on clinical metagenomics tools

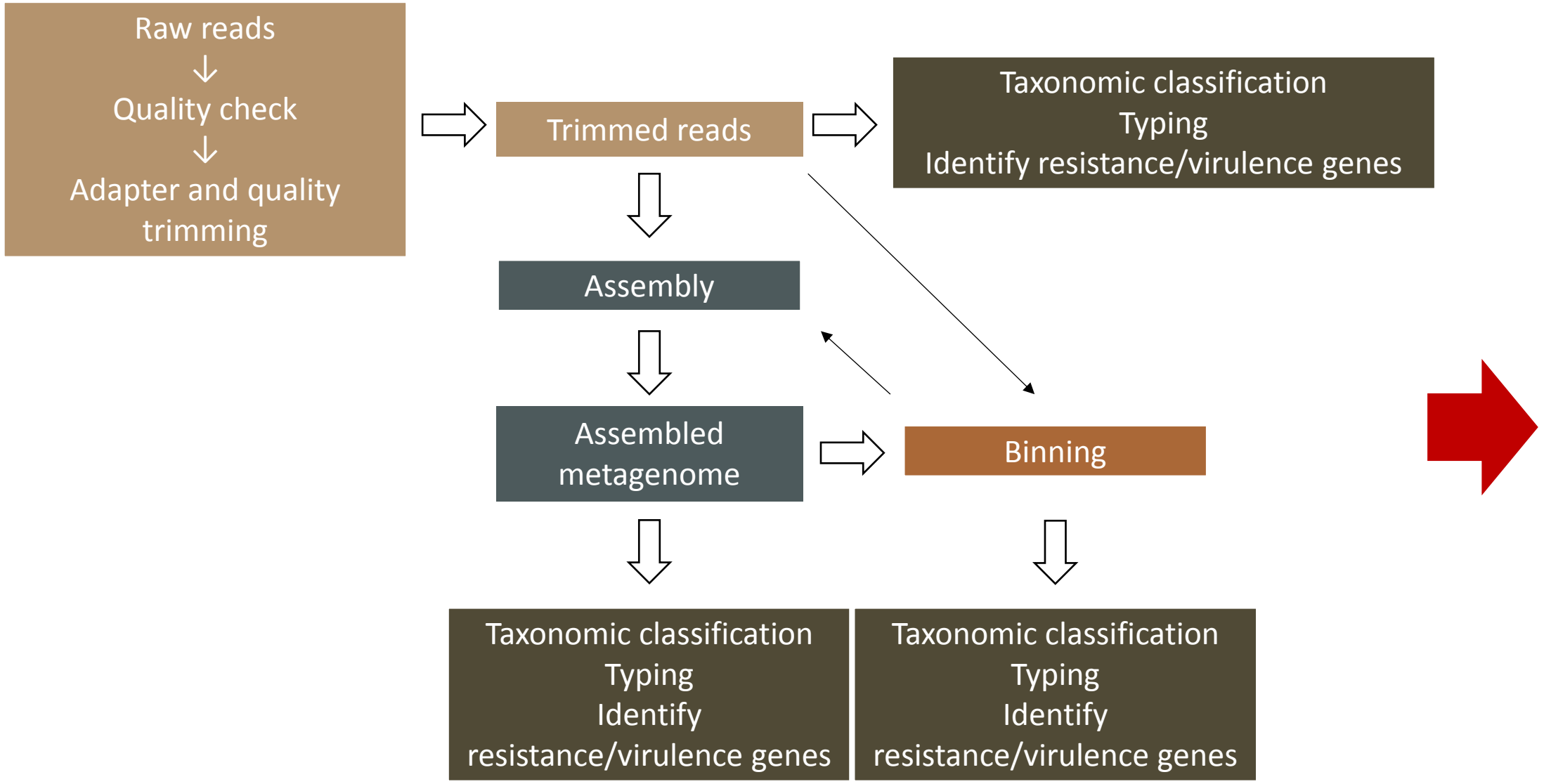


Natacha Couto
Department of Medical Microbiology
University Medical Center Groningen, RUG
Geneva, 17-18th October 2019

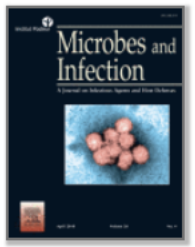
Lab design



ISO 15189 certified NGS for diagnostics
ISO 9001 certification seeking


Raw reads
↓
Quality check
↓
Adapter and quality trimming
Trimmed reads
Taxonomic classification
Typing
Identify resistance/virulence genes
Assembly
Assembled metagenome
Binning
Taxonomic classification
Typing
Identify resistance/virulence genes
Taxonomic classification
Typing
Identify resistance/virulence genes
INTERPRETATION

Microbes and Infection
Volume 20, Issue 4, April 2018, Pages 222-227



Meeting report

Messages from the second International Conference on Clinical Metagenomics (ICCMg2)

Etienne Ruppé [a], Jacques Schrenzel [b]

Show more

https://doi.org/10.1016/j.micinf.2018.02.005

Table 2
Summary of the take-home messages and related key-points of the ICCMg2. CLIA: Clinical Laboratory Improvement Amendments. ISO: International Organization for Standardization.

Message	Key points
Microbiome studies	Push the identification of bacteria up to the strain level. Case-control studies: towards more complex design to address causality. Importance of a biological/clinical expertise along with the bioinformatic and biostatistical analysis
The importance of contaminants in clinical metagenomics	What negative control(s) should be used? How to substract the contaminants from the results?
Towards a universal pipeline and consequences on the nucleic acids extraction	Consider viruses (DNA and RNA), bacteria, antibiotic resistance genes, fungi, parasites in a single pipeline. Extract DNA and RNA. Consider the host's gene expression.
	Several efficient solutions (most unpublished yet) to remove human DNA.
The increasing fastness of clinical metagenomics	Fast results within hours with Nanopore sequencing, yet quality still not optimal.
“New” culprits identified by metagenomic studies	Pathogenicity of unexpected microbes? Already actionable results when conventional methods fail to identify any causative microbe.
Quality	Adapt CLIA or ISO15189 requirements to the clinical metagenomics workflow. Validation of the method: towards a confidence score (like mass spectrometry?) Are clinical parameters the best comparator to validate clinical metagenomics tests?
Antibiotic resistance	EUCAST consultation: the WGS antibiogram not for now, but works well for some couples bacterium-antibiotic. Metagenomics allows to identify new resistance genes. Need for a database of resistance genes and associated metadata. Towards a clinical resistance with clinical metagenomics instead of an antimicrobial resistance?

Contamination

- Nucleic acid extraction kits (kitome)
- Reagents and diluents
- Host
- Post-sampling environment (i.e. airborne particles, index switching, crossovers from past sequencing runs)
- Misclassification related to the classification algorithms used and/or the reference databases available

Schlaberg et al., Arch Pathol Lab Med 2016; Martí, PLOS Comp Biol 2019;15(4):e1006967.

Contamination?

Method	Total number of bacteria identified[a]	True positives[a]	False positives	False negatives	Sensitivity (%)	PPV (%)
Culture/MALDI-TOF	9	9	0	0	100%	100%
MetaPhlAn (BaseSpace)	16	7	9	2	78%	44%
Genius (BaseSpace)	35	8	27	1	89%	23%
Kraken (BaseSpace)	959	7	952	2	78%	1%
Taxonomer (Full Analysis)	4649	8	4641	1	89%	0%
CosmosID	35	8	27	1	89%	23%
Taxonomic Profiling (CLC Genomics Workbench v10.0.1)	17	6	11	3	67%	35%
Best match K-mer spectra (CLC Genomics Workbench v10.0.1)	12	8	4	1	89%	67%
Kraken (Unix)	198	7	191	2	78%	4%
MetaPhlAn2 (Unix)	15	7	6	4	78%	60%
MIDAS (Unix)	34	7	26	2	78%	24%

Table 5. Performance of the different taxonomic classification methods for each sample. Sensitivity and positive predictive value were calculated using culture/MALDI-TOF as standards. [a]Excluding the samples with non-identified anaerobic bacteria (Samples 2 and 5). Abbreviations: PPV, positive predictive value.

Couto et al., Sci Rep 2018;8:13767.

Contamination?

- Most of the studies deal with contamination based on *ad hoc* cut-offs or thresholds…

Table 3 Number of correctly and incorrectly predicted species[a] for different thresholds[b] without clade exclusion. Some methods vastly overpredict the number of species, even when the true number of species is low (in this case the true number of species is 11)

	No cutoff[b]		Cutoff > 0.01 %[b]		Cutoff > 0.1 %[b]		Cutoff > 1 %[b]	
Method	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
CARMA3	11	56	11	4	11	0	10	0
CLARK	11	364	11	25	11	5	11	0
DiScRIBinATE RAPSearch2[c]	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Kraken	11	327	11	25	11	5	11	0
Filtered Kraken	11	14	11	1	11	0	11	0
MEGAN4 BlastN	11	110	11	19	11	3	9	1
MEGAN4 RAPSearch2	11	183	11	41	11	1	9	1
MetaBin	11	561	10	77	10	6	10	1
MetaCV	11	1226	11	232	11	6	10	1
MetaPhyler	11	9	11	9	11	5	7	1
PhymmBL[c]	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
RITA	11	466	10	80	10	10	10	1
TACOA[c]	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
MG-RAST best hit	11	927	10	180	10	36	10	8
MG-RAST LCA	11	476	11	69	11	5	11	1

[a]Using the FW *in vitro* dataset of sequenced reads from 11 species
[b]A cutoff of > × %, for example 0.01 %, would indicate that only species with a predicted abundance of at least x % of the total set of predictions were considered. Correctly predicted species are any of the 11 species that were used to simulate the reads in the dataset, whereas any other predicted species was incorrect
[c]These methods do not predict to the species level at this read length (they require longer read lengths). See additional analyses at other levels of clade exclusion

Tools to tackle contamination

Davis *et al. Microbiome* (2018) 6:226
https://doi.org/10.1186/s40168-018-0605-2

Microbiome

METHODOLOGY Open Access

Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data

Nicole M. Davis[1], Diana M. Proctor[2,3], Susan P. Holmes[4], David A. Relman[1,2,5] and Benjamin J. Callahan[6,7*]

RESEARCH ARTICLE

Recentrifuge: Robust comparative analysis and contamination removal for metagenomics

Jose Manuel Martí *

Institute for Integrative Systems Biology (I^{2}SysBio), Valencia, Spain

* jose.m.marti@uv.es

Davis *et al. Microbiome* (2018) 6:226
https://doi.org/10.1186/s40168-018-0605-2

Microbiome

METHODOLOGY Open Access

Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data

Nicole M. Davis[1], Diana M. Proctor[2,3], Susan P. Holmes[4], David A. Relman[1,2,5] and Benjamin J. Callahan[6,7*]

Consideration

- Total sample DNA (T) is a mixture of two components:
  - Contaminating DNA (C) present in uniform concentration across samples
  - True sample DNA (S) present in varying concentration across samples

$$T = C + S$$

Davis et al. Microbiome (2018) 6:226

First assumption

- *"Sequences from contaminating taxa are likely to have frequencies that inversely correlate with sample DNA concentration."*

$$f_C = C/(C+S) \sim 1/T$$
$$f_S = S/(C+S) \sim 1$$

Not suitable for low-biomass samples:
C ~ *S* or *C* > *S*

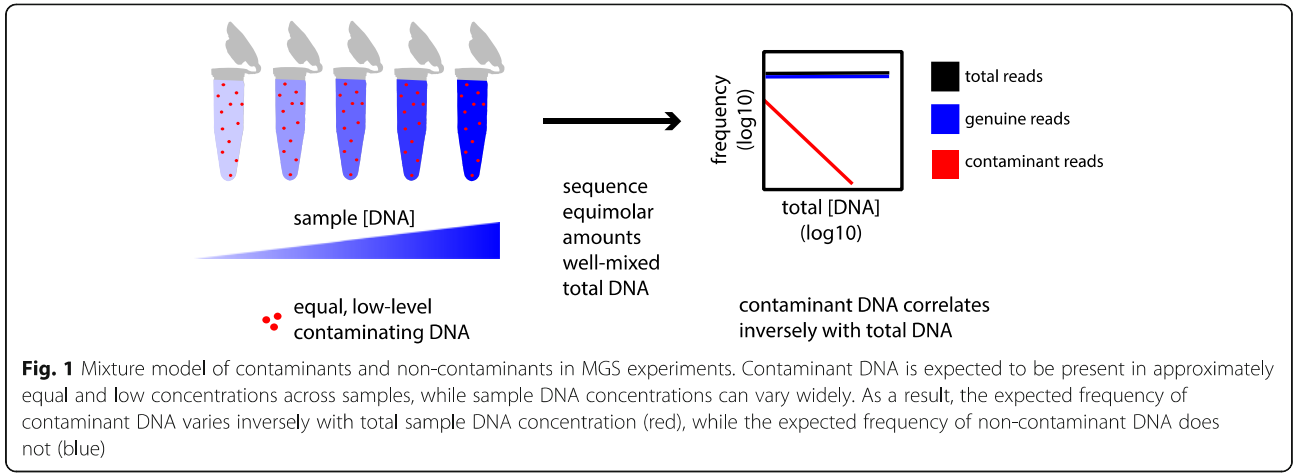
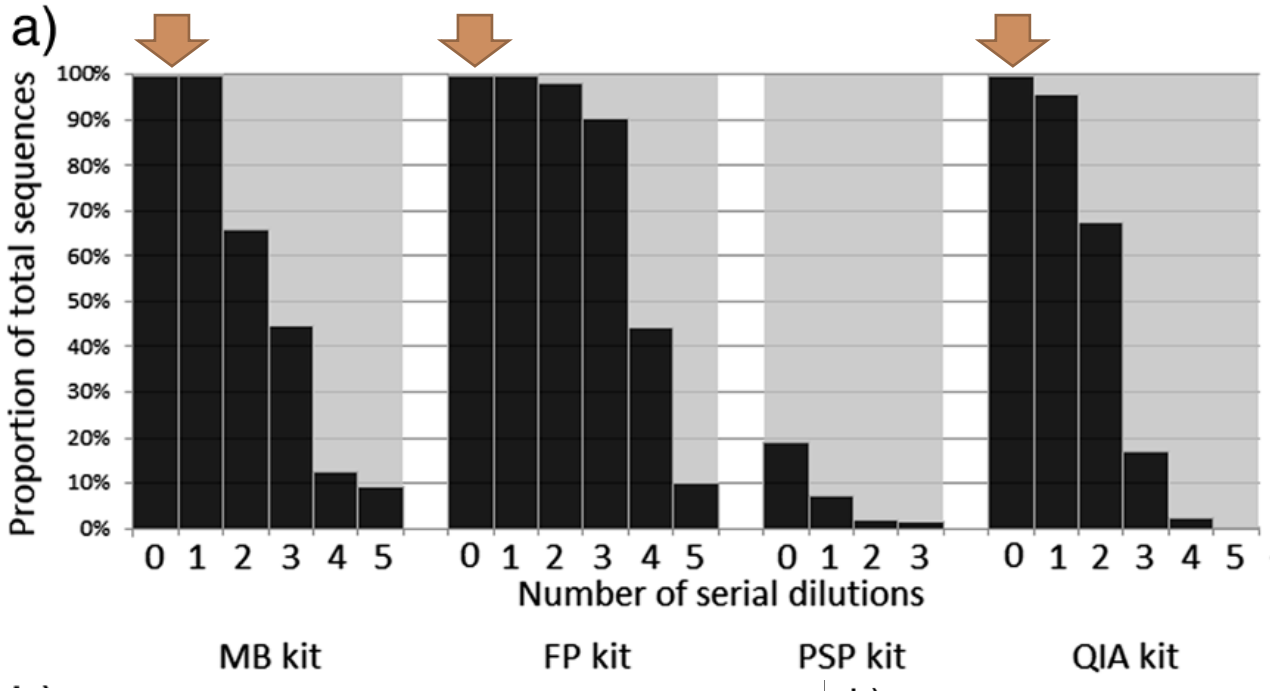
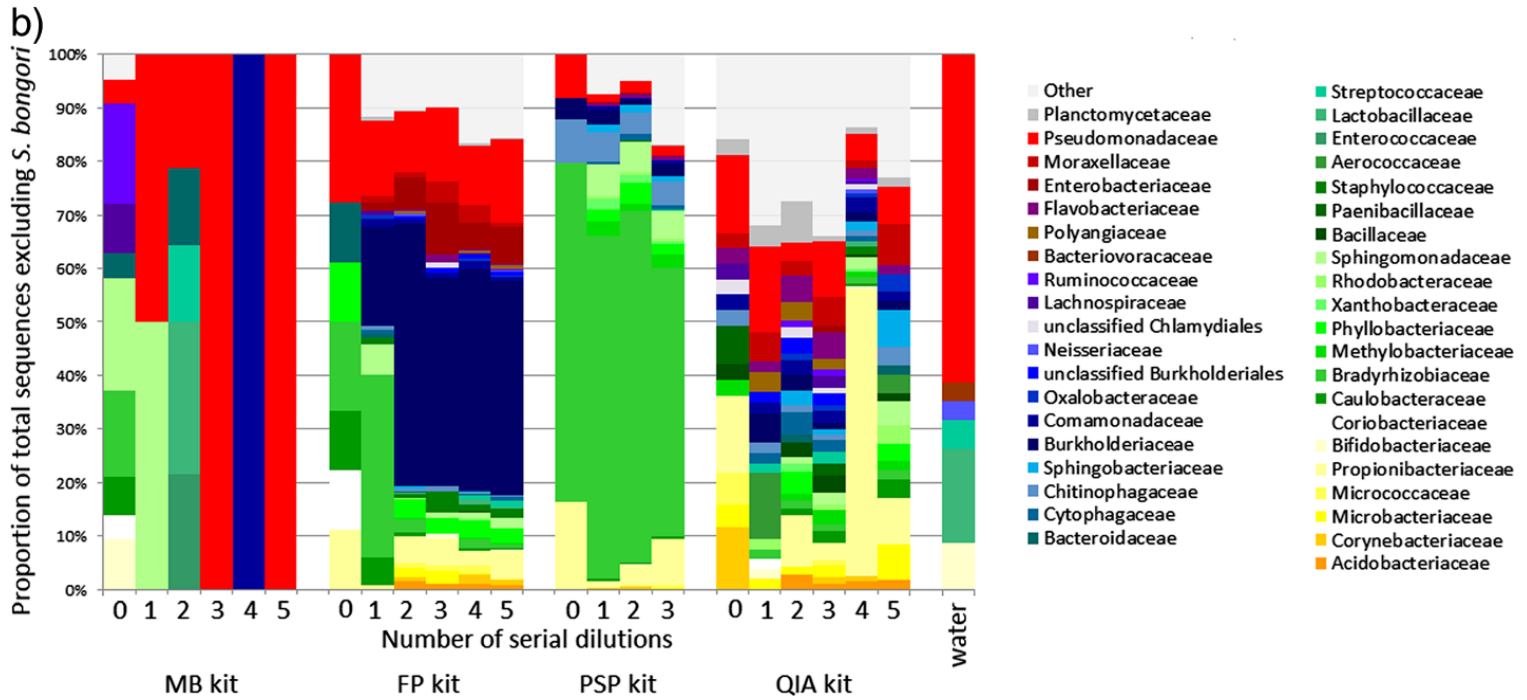


Fig. 1 Mixture model of contaminants and non-contaminants in MGS experiments. Contaminant DNA is expected to be present in approximately equal and low concentrations across samples, while sample DNA concentrations can vary widely. As a result, the expected frequency of contaminant DNA varies inversely with total sample DNA concentration (red), while the expected frequency of non-contaminant DNA does not (blue)

Frequency-based identification

Davis et al. Microbiome (2018) 6:226





Salter et al. BMC Biology 2014, 12:87

Second assumption

- *"Sequences from contaminating taxa are likely to have higher prevalence in control samples than in true samples."*
- *C* negative control > *C* true sample
- Negative control $S \sim 0$,
- True sample $S > 0$

Prevalence-based identification

Davis et al. Microbiome (2018) 6:226.

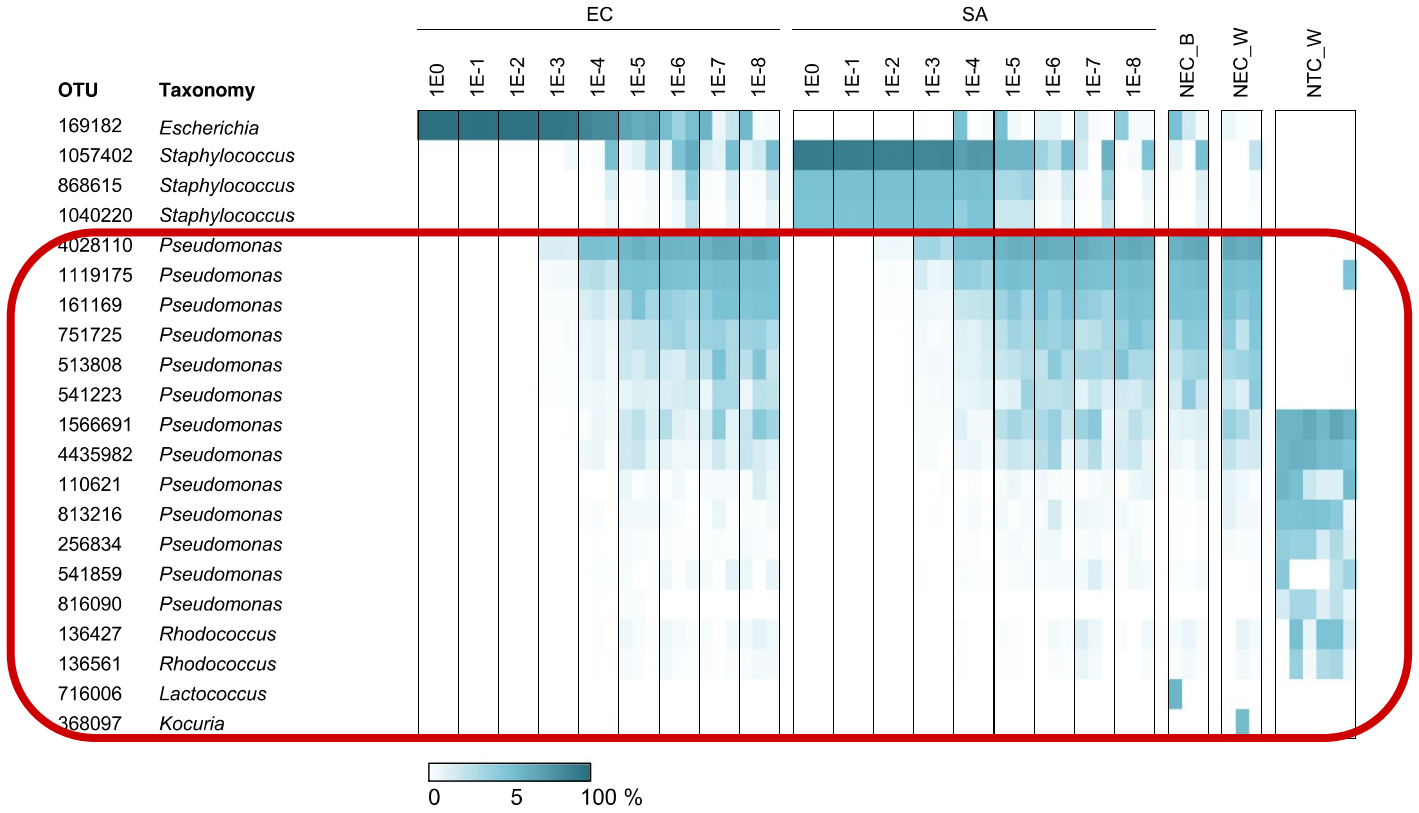


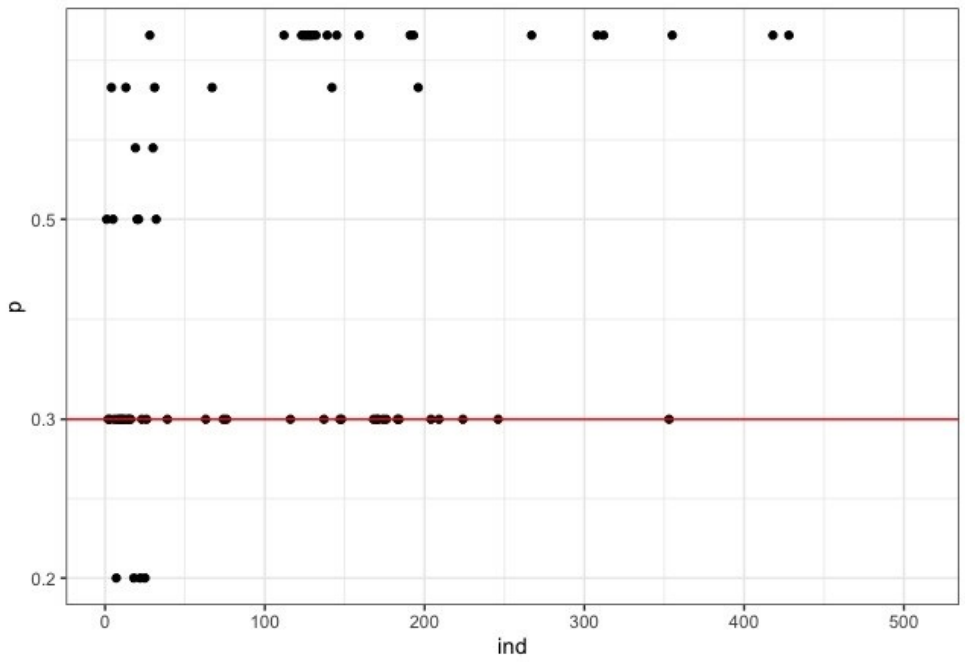
Fig. 3 Relative abundance of predominant OTUs. OTUs with a mean relative abundance >1 % in either samples, negative extraction controls or NTC_W are presented. The proportion is indicated by the scale at the bottom of the plot. Dilutions of the master stock are indicated from 1E0 (no dilution) to 1E-8 (10^{-8}). For EC, SA, NEC_B and NEC_W, the data obtained from DNA extractions performed on three occasions (Exp1–Exp3) are presented from left to right. NTC_W were performed in duplicate for each of the three series. EC, *E. coli*; SA, *S. aureus*. NEC_W, negative extraction controls obtained substituting culture for water; NEC_B, negative extraction controls obtained by substituting culture for lysis buffer; NTC_W, no-template PCR control

Algorithm

- Developed in R
- Contains two modules:
  - *isContaminant* function (score statistic *P,* threshold p>0.01, frequency-based identification, prevalence-based identification, combined-based identification)
  - *isNotContaminant* function (score statistic 1 – *P*, p<0.05, prevalence-based identification) for low biomass samples
  - Requirements:
    - A feature table of the relative abundances or frequencies of sequence features in each sample (e.g., an OTU table) in format .biom, and
    - (1) quantitative DNA concentrations for each sample, often obtained during amplicon or shotgun sequencing library preparation in the form of a standardized fluorescence intensity (e.g., PicoGreen), and/or
    - (2) sequenced negative control samples, preferably DNA extraction controls to which no sample DNA was added.

Davis et al. Microbiome (2018) 6:226

isNotContaminant, threshold < 0.5 or 0.3



Sample number	Culture result (CFU)[a]	Conventional identification (MALDI-TOF)	WGS-based identification	Shotgun metagenomics		
				Kraken[b]	MIDAS[c]	MetaPhlAn[c]
1	10^3 10^3 10	*E. faecium* *S. haemolyticus* *C. glabrata*	*E. faecium* *S. haemolyticus* —	*E. faecium* (34.6%) *S. haemolyticus* (10.1%) —	*E. faecium* (62.0%) *S. haemolyticus* (28.0%) —	*E. faecium* (66.6%) *S. haemolyticys* (27.7%) —
2	10^3 1 Not determined	*E. avium* *E. coli* Anaerobes	—# —# —#	Not identified* Not identified* Several species (29.5%)	Not identified* Not identified* Several species (100.0%)	Not identified* Not identified* Several species (100.0%)
3	1	*S. epidermidis*	—#	*S. aureus* (0.2%)	Not identified*	Not identified*
4	10^3	*S. aureus*	*S. aureus*	*S. aureus* (0.73%)	*S. aureus* (100%)	*S. aureus* (100%)
5	$\geq 10^5 \geq 10^5$ 10^3 10^3 Not determined 10	*E. coli* *K. oxytoca* *S. anginosus* *E. faecalis* Anaerobes *C. albicans*	*E. coli* *K. oxytoca* —# *E. faecalis* —# —#	*E. coli* (9.7%) *K. oxytoca* (0.5%) *S. anginosus* (0.07%) *E. faecalis* (0.3%) Several species (12.7%) —	*E. coli* (6.5%) *K. oxytoca* (0.3%) *S. anginosus* (0.01%) *E. faecalis* (0.9%) Several species (96.7%) —	*E. coli* (8.5%) *K. oxytoca* (0.3%) *Streptococcus* spp. (0.09%) *E. faecalis* (0.7%) Several species (90.4%) —
6	10^3	*E. faecium*	*E. faecium*	*E. faecium* (0.77%)	Not identified*	Not identified*
7	10^2	*S. aureus*	—#	*S. aureus* (82.9%)	*S. aureus* (100%)	*S. aureus* (100%)
8	10^3	*O. intermedium*	*O. intermedium*	*O. anthropi* (21.3%)	*O. intermedium* (99.4%)	*O. intermedium* (99.1%)
9	10^3	*S. aureus*	*S. aureus*	*S. aureus* (22.9%)	*S. aureus* (100%)	*S. aureus* (100%)
10	10^3	*S. marcescens*	—#	*S. marcescens* (64.7%)	*S. marcescens* (99.1%)	*S. marcescens* (100%)

37 species (threshold < 0.3):
Multiple *Staphylococcus* spp.
Serratia sp.
Citrobacter freundii
Clostridium botulinum
Klebsiella pneumoniae
Streptococcus anginosus

4 species (threshold < 0.5):
Staphylococcus epidermidis
Serratia sp.
Citrobacter freundii
Clostridium botulinum

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8	Sample 9	Sample 10	Negative control
Sample type	Peritoneal fluid	Pus (abscess)	Synovial fluid	Synovial fluid	Pus (abscess)	Pus (empyema)	Pus (empyema)	Bone biopsy	Pus (abscess)	Sputum	Water
DNA extraction method	Ultra-Deep Microbiome Prep (Molzym)	Ultra-Deep Microbiome Prep (Molzym)	Ultra-Deep Microbiome Prep (Molzym)	Ultra-Deep Microbiome Prep (Molzym)	Ultra-Deep Microbiome Prep (Molzym)	QIAamp DNA Microbiome Kit (Qiagen)	QIAamp DNA Microbiome Kit (Qiagen)	Micro-DX™ (Molzym)	Micro-DX™ (Molzym)	Micro-DX™ (Molzym)	QIAamp DNA Microbiome Kit (Qiagen)
Total number of reads	5,892,978	9,603,346	8,615,810	6,078,166	8,368,930	2,912,802	1,486,700	6,534,866	6,173,132	7,596,836	1,730,738
Mapped reads against hg19	5,249,063 (89.2%)	7,828.746 (81.6%)	8,254,594 (95.9%)	6,015,945 (99.0%)	309,588 (3.7%)	2,877,066 (98.8%)	922,932 (62.2%)	229,149 (3.5%)	6,081,612 (98.5%)	7,337,832 (96.7%)	1,706,861 (98.9%)
Unmapped reads	632,951 (10.8%)	1,770,558 (18.4%)	355,200 (4.1%)	61,099 (1.0%)	8,052,272 (96.3%)	34,506 (1.1%)	561,772 (37.8%)	6,303,803 (96.5%)	89,922 (1.5%)	235,520 (3.3%)	19,805 (1.2%)

Conclusions

- The classification accuracy is dependent on the number of samples in which a sequence feature appeared (its prevalence) → depends on patterns across samples to identify contaminants (low sensitivity for detecting contaminants that are found in very few samples) → so probably 4-5 samples is not enough to draw any conclusions
- Is not designed to remove cross-contamination → severely affected by this phenomenon

Davis et al. Microbiome (2018) 6:226

RESEARCH ARTICLE

Recentrifuge: Robust comparative analysis and contamination removal for metagenomics

Jose Manuel Martí *

Institute for Integrative Systems Biology (I^2SysBio), Valencia, Spain

* jose.m.marti@uv.es

Martí (2019); PLoS Comput Biol 15(4): e1006967.

Two strategies

1. Accounts for the score level of the classifications in every single step provided by the taxonomic classifier;
2. It uses a removal algorithm that detects and selectively eliminates various types of contamination, including crossovers.

- Supports high-performance classifiers such as Centrifuge, LMAT, CLARK, CLARK-S and Kraken (and Kraken2), but alternative classifiers can also be used.

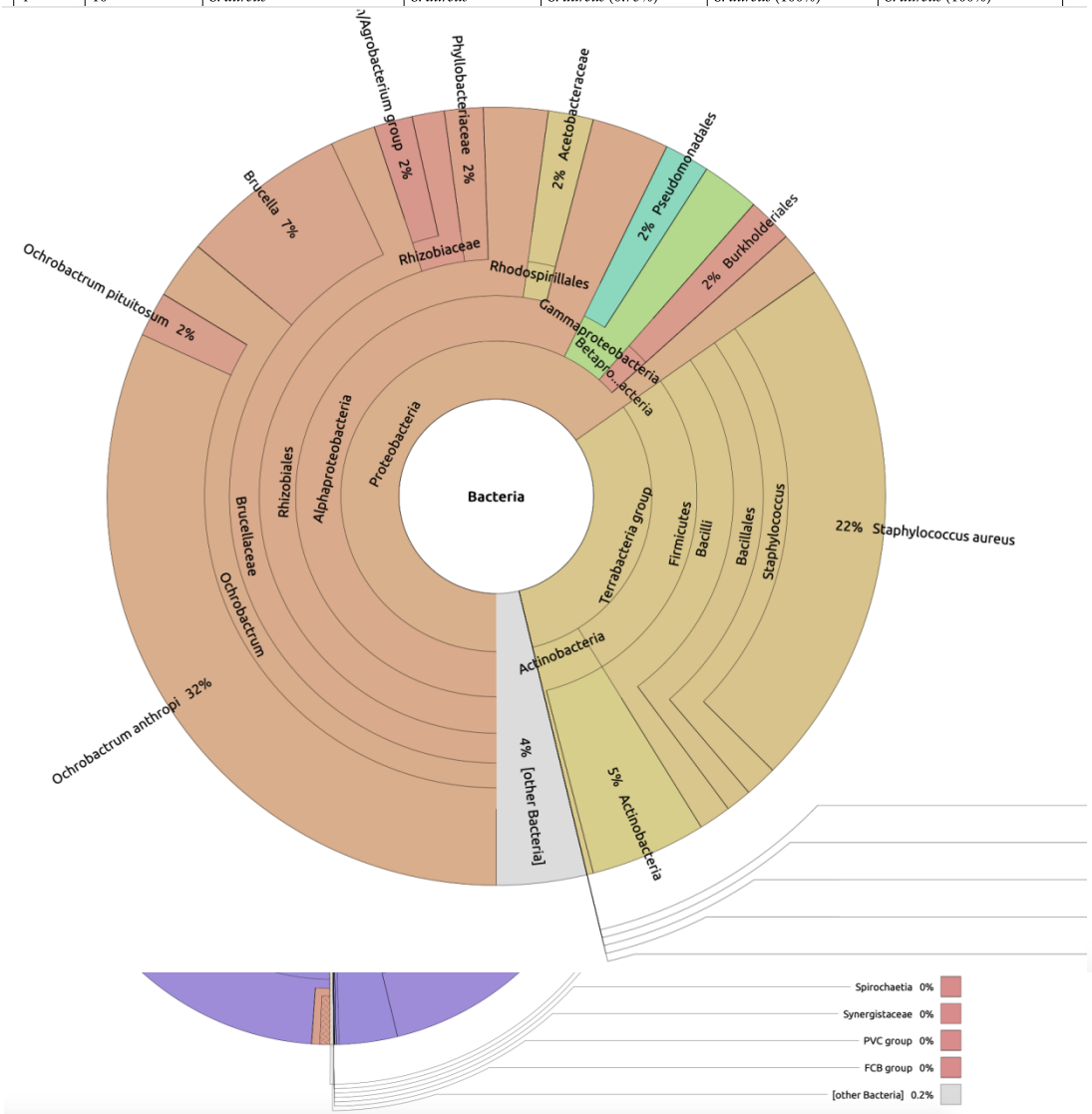
Martí (2019); PLoS Comput Biol 15(4): e1006967.

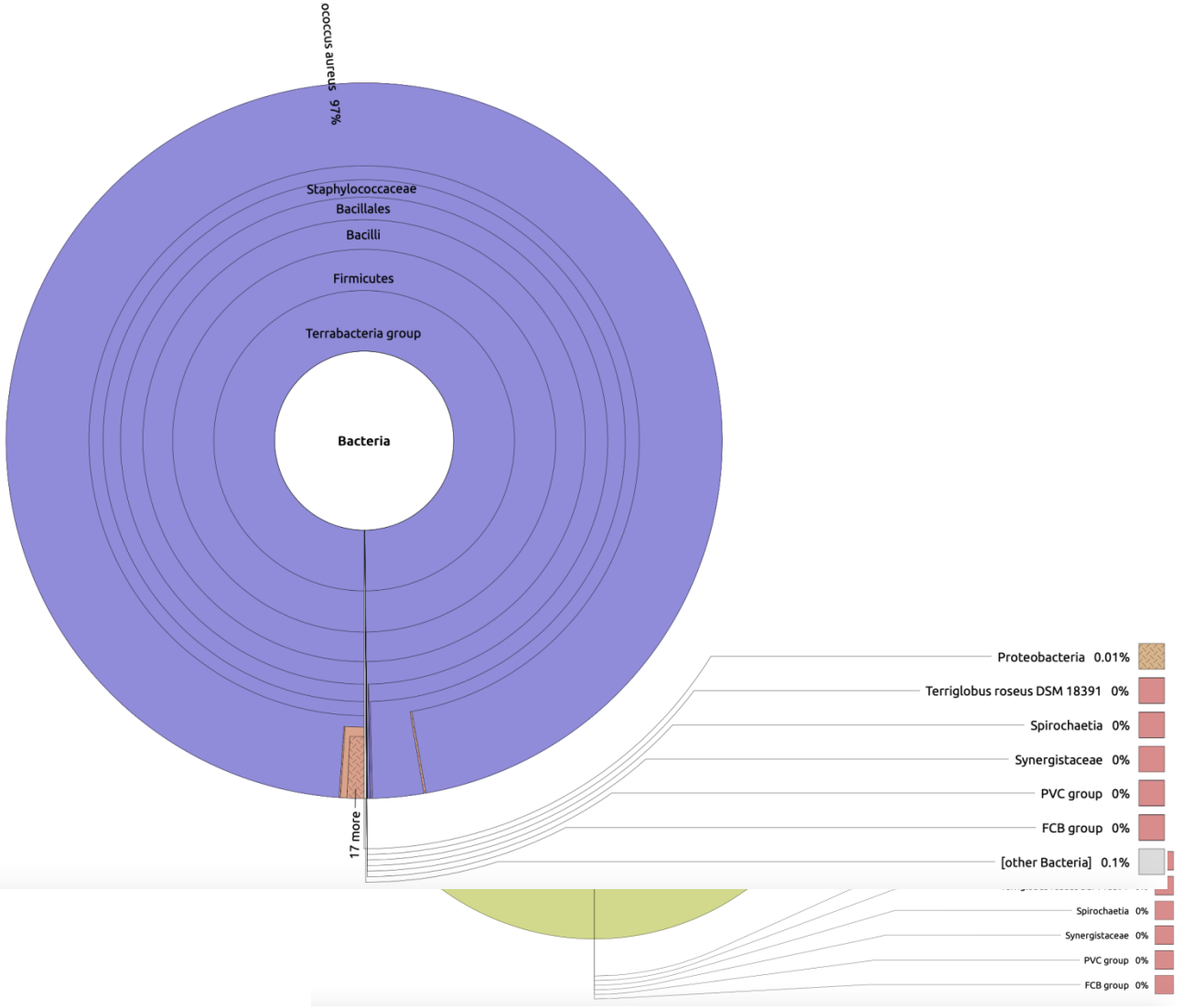
Recentrifuge

- Depending on the relative frequency of the “candidate contaminating taxa” in the control samples and if they are present in other specimens, the algorithm classifies them in contamination level groups: critical, severe, mild, and other.
- Except for the “other”, the candidate contaminants are removed from non-control groups.
- The “other contaminants” group is checked for crossover contamination, so those taxon are eliminated from all samples except for the one or ones selected as the source of “pollution”.

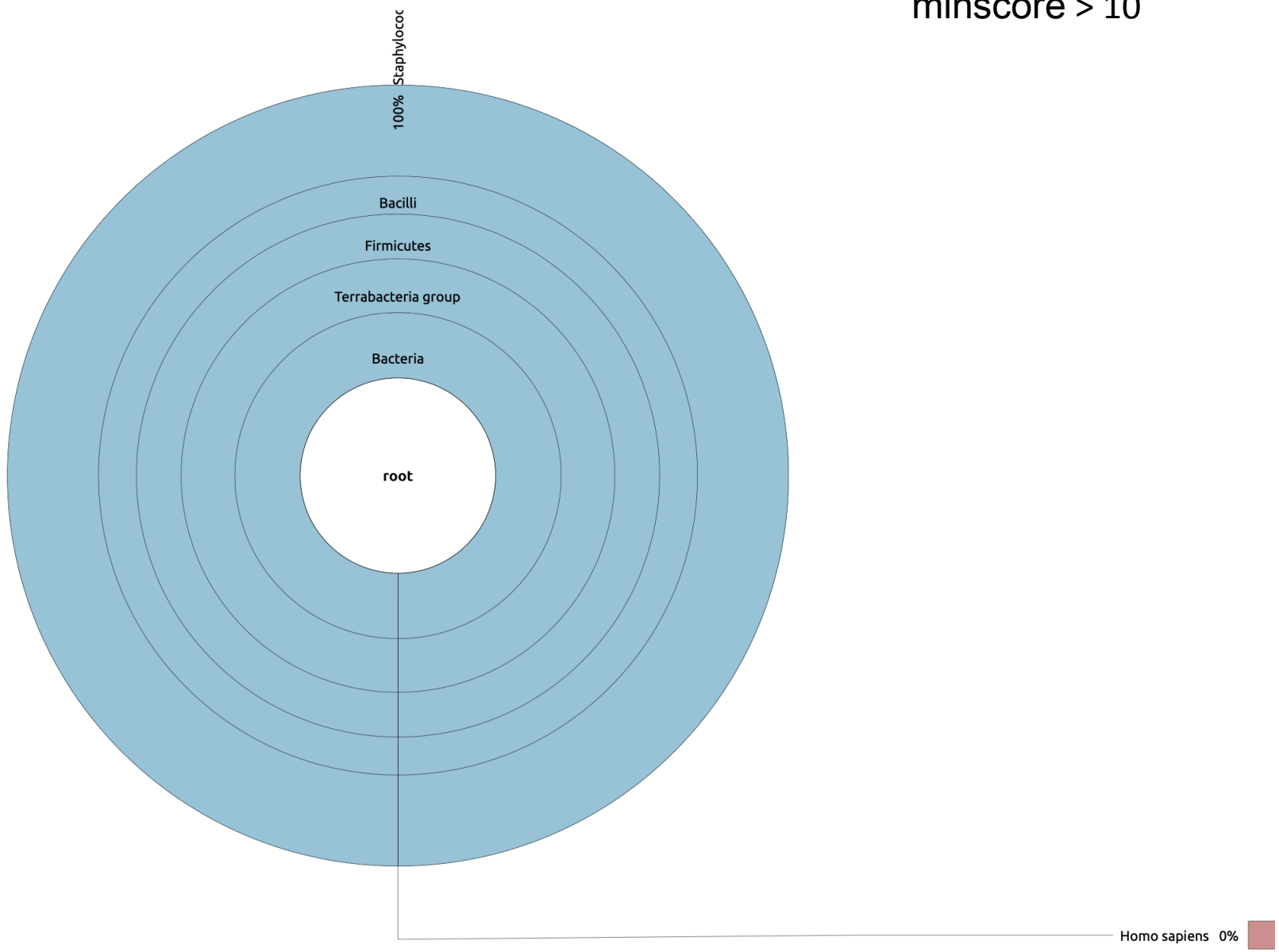
Martí (2019); PLoS Comput Biol 15(4): e1006967.

Sample number	Culture result (CFU)[a]	Conventional identification (MALDI-TOF)	WGS-based identification	Shotgun metagenomics		
				Kraken[b]	MIDAS[c]	MetaPhlAn[c]
1	10^3 10^3 10	*E. faecium* *S. haemolyticus* *C. glabrata*	*E. faecium* *S. haemolyticus* —	*E. faecium* (34.6%) *S. haemolyticus* (10.1%) —	*E. faecium* (62.0%) *S. haemolyticus* (28.0%) —	*E. faecium* (66.6%) *S. haemolyticys* (27.7%) —
2	10^3 1 Not determined	*E. avium* *E. coli* Anaerobes	—[#] —[#] —[#]	Not identified* Not identified* Several species (29.5%)	Not identified* Not identified* Several species (100.0%)	Not identified* Not identified* Several species (100.0%)
3	1	*S. epidermidis*	—[#]	*S. aureus* (0.2%)	Not identified*	Not identified*
4	10^3	*S. aureus*	*S. aureus*	*S. aureus* (0.73%)	*S. aureus* (100%)	*S. aureus* (100%)





minscore > 10


100% Staphylococ
Bacilli
Firmicutes
Terrabacteria group
Bacteria
root
Homo sapiens 0%
24
22
19
17
14
12
Kmer coverage (%)

Conclusion

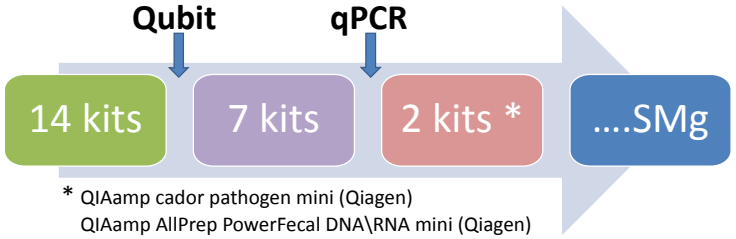
- Recentrifuge performs better than *Decontam* and is much more user friendly
- It can lead to “false contaminants”, but raising the minscore should solve the problem

Conclusion

- New tools for “Decontamination” are available and can be validated for clinical metagenomics
- Always include negative controls for each run and so you can better predict the contaminants using the tools mentioned before

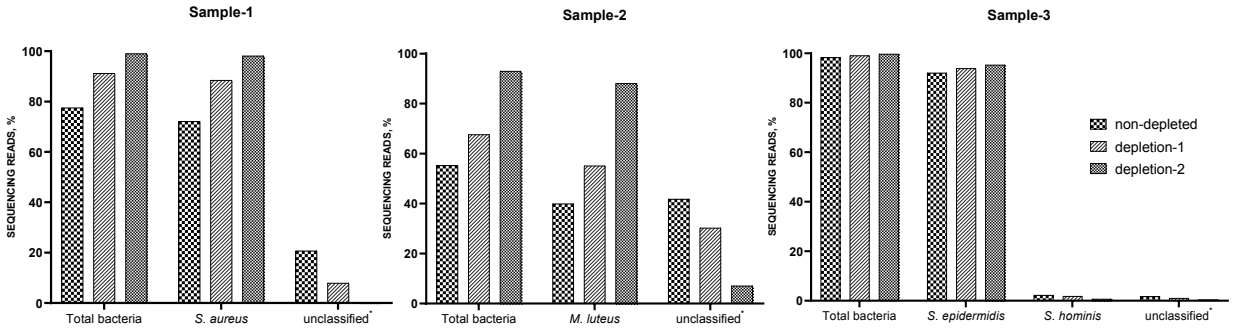
Optimized host depletion methods

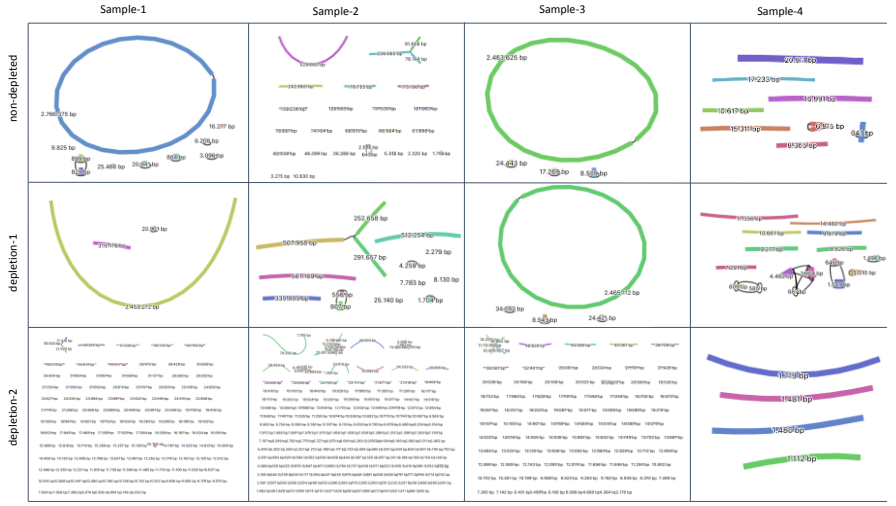
Fungi	*Candida albicans*
Bacteria (gram -)	*Escherichia coli*
Bacteria (gram +)	*Staphylococcus aureus*
RNA virus (non-enveloped)	Echo 18
RNA virus (enveloped)	PDV
DNA virus	PhHV



Nilay Peker

Leonard Schüle





Posters

Evaluation of nucleic acid extraction kits for Shotgun Metagenomic Sequencing

Sample preparation for diagnosis of bloodstream infections by Shotgun Metagenomics

New assembler - Flye

ARTICLES
https://doi.org/10.1038/s41587-019-0072-8
nature biotechnology

Assembly of long, error-prone reads using repeat graphs

Mikhail Kolmogorov[1], Jeffrey Yuan[2], Yu Lin[3] and Pavel A. Pevzner[1*]

Accurate genome assembly is hampered by repetitive regions. Although long single molecule sequencing reads are better able to resolve genomic repeats than short-read data, most long-read assembly algorithms do not provide the repeat characterization necessary for producing optimal assemblies. Here, we present Flye, a long-read assembly algorithm that generates arbitrary paths in an unknown repeat graph, called disjointigs, and constructs an accurate repeat graph from these error-riddled disjointigs. We benchmark Flye against five state-of-the-art assemblers and show that it generates better or comparable assemblies, while being an order of magnitude faster. Flye nearly doubled the contiguity of the human genome assembly (as measured by the NGA50 assembly quality metric) compared with existing assemblers.

Improving outbreak surveillance with rapid- and long-read sequencing

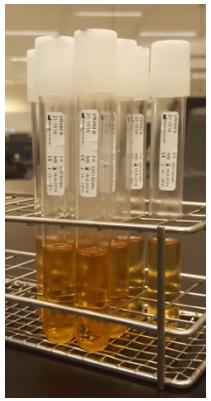
VRE as an example

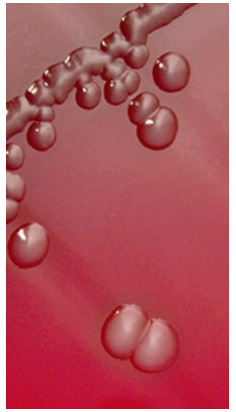
Carolien Doorenboos

Vancomycin-resistant enterococci

- *Enterococcus faecium* and *Enterococcus faecalis*
  - Commensals of human gut
  - Associated with hospital acquired infections (mainly *E. faecium*)
- Vancomycin resistance in *E. faecium* and *E. faecalis* is mediated by the *vanA* and *vanB* gene
- In the Netherlands VRE carriage is an indication for hospital care in isolation to prevent transmission
- High risk patients and wards are routinely screened

Screening for VRE

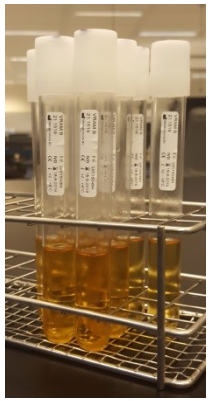




Day 0	Day 1	Day 2-4	Day 3-5	Day 4-6
Rectal swab	Selective liquid broth	RT-PCR Targets: • *E. faecium* • *vanA* • *vanB*	If PCR positive: Oxoid Brilliance VRE (72h)	Axenic culture Resistance testing

VRE screening protocol UMCG

Screening for VRE with the MinION



```
0       1
0  1    11
1  00   1
1  01   0
00 11   0
01 01   0
01 10   11
10 0    11
1       01
```

Day 0	Day 1	Day 2-4

Rectal swab

Selective liquid broth

RT-PCR
Targets:

- *E. faecium*
- *vanA*
- *vanB*

MinION sequencing

(Real time) data analysis

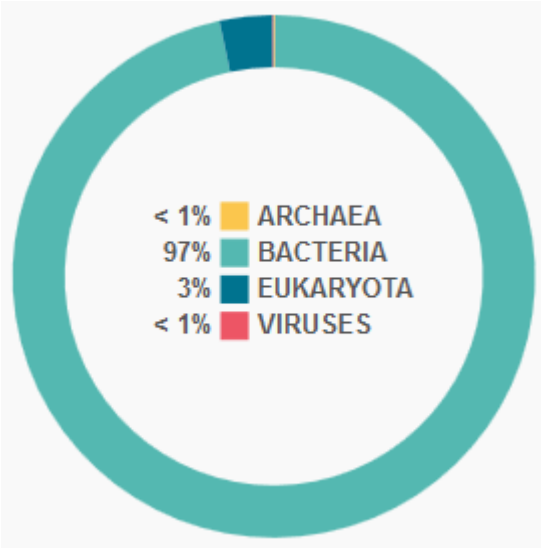
VRE screening protocol UMCG

Data analysis for long read sequencing

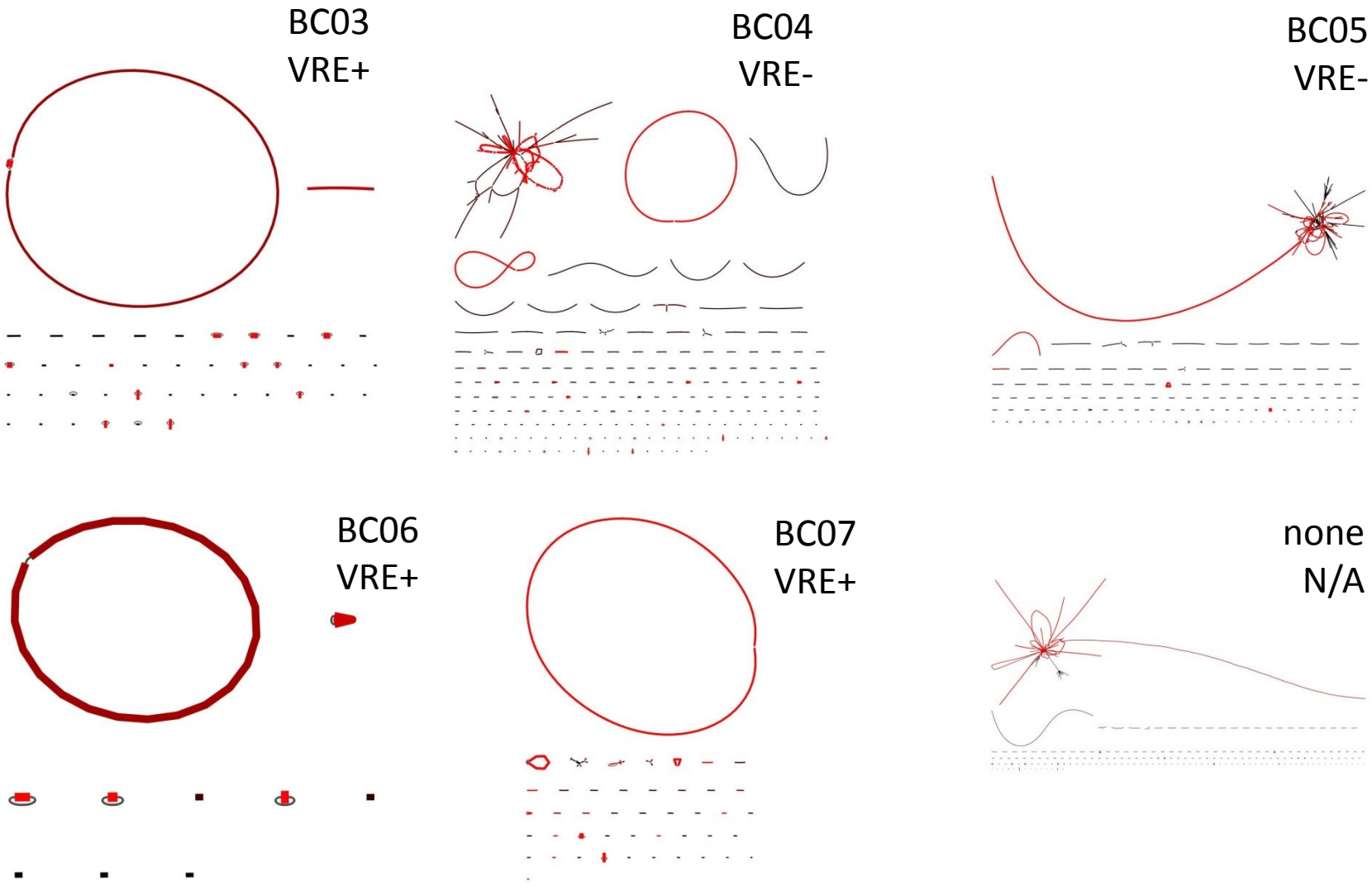
• Basecalling	MinIT/Guppy √
• Demultiplexing	qcat/fastp √
• Trimming of barcodes and adapters	qcat, filtlong √
• Assembly	flye, canu, metaspades √
• Polishing	nanopolish √, medaka, pilon, racon
• Taxonomy assignment	Kraken2 √
• Resistance	ABRicate √
• Phylogeny	Ridom SeqSphere+ √
• Visualization of assembly	Bandage √

First preliminary results

- 5 samples
  - 5/5 *E. faecium* and either *vanA* or *vanB* in RT-PCR
  - 3/5 VRE in culture

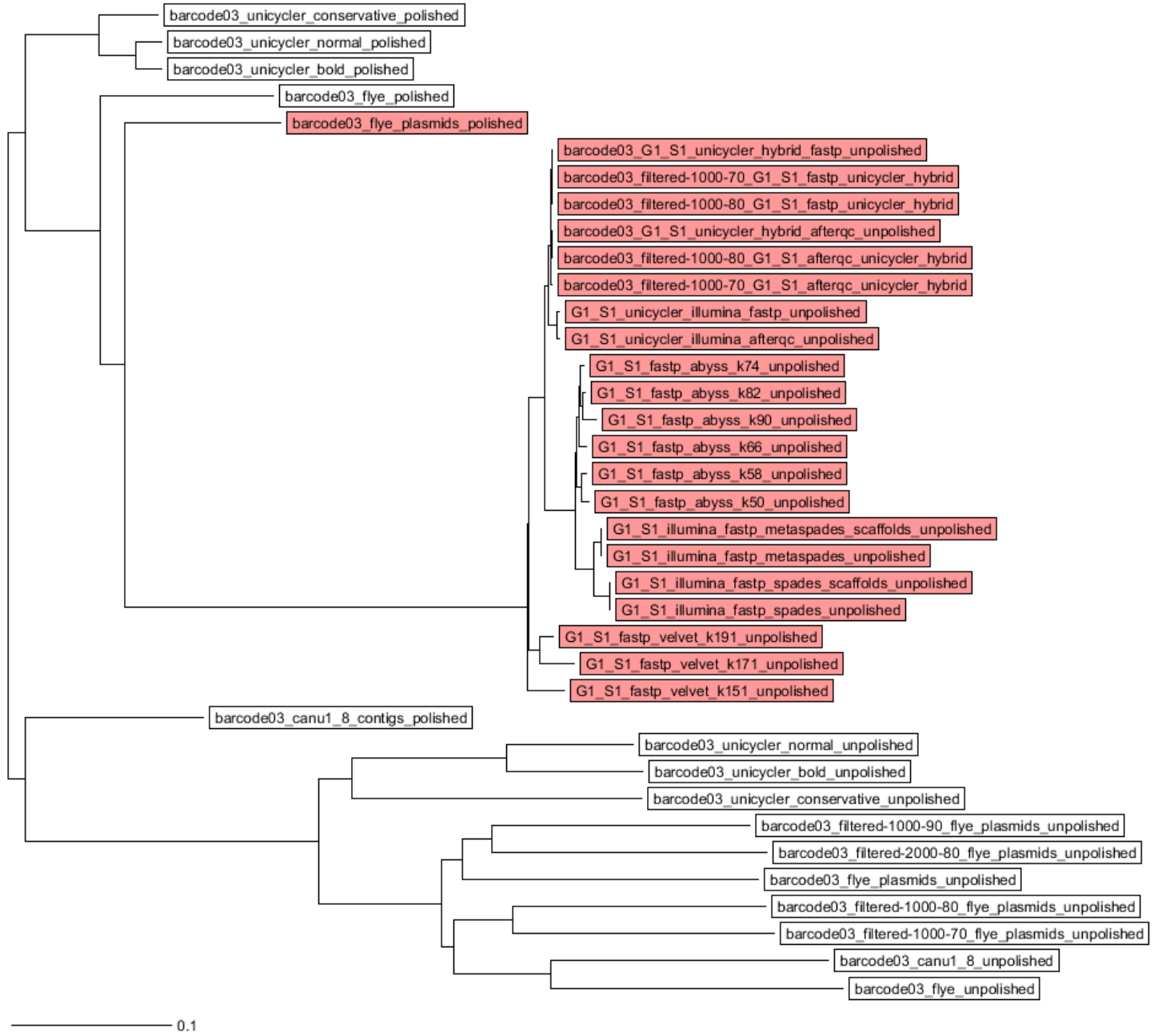


ID	Read Count
BC01	108
BC02	391
BC03	312,069
BC04	435,090
BC05	267,628
BC06	375,240
BC07	574,237
BC08	46
BC09	113
BC10	12
BC11	138
BC12	69
No Barcode	286,337



- Bandage with flye –meta assembly

wgMLST analysis using Ridom SeqSphere+ v6.0



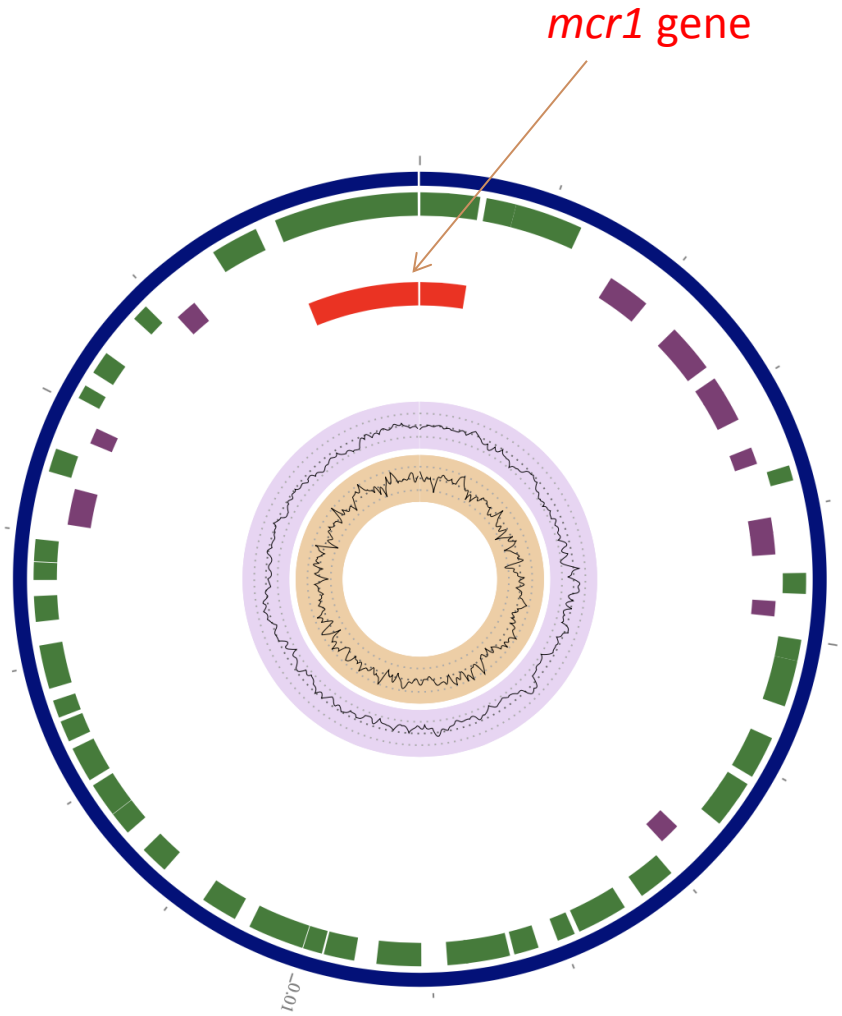
IDENTIFICATION AND CHARACTERIZATION OF VIRUSES DIRECTLY FROM BLOOD PLASMA AND NASAL SWABS FROM PIGS FOR THE EARLY WARNING OF INFECTIOUS DISEASES.

food PRO·TEC·TS
Food production technologies for trans-boundary systems

Leonard Schüle

Found an *mcr1* gene in one sample;
Different assembly strategies:

- Short-read assembly with CLC Genomics Workbench, SPAdes, SPAdes --meta, Megahit
- Short-read assembly with scaffolding using long reads using CLC Genomics Workbench
- Long-read assembly with CANU
- Long-read assembly with Flye --meta --plasmid √



*p*LEO1 18 176 bp

Acknowledgements

- Funding
  - EH-1H project
  - FoodProtects
- MMB-UMCG-RUG
  - Alex W. Friedrich
  - John W. Rossen
  - Erley Lizarazo-Forero
  - Leonard Schuele
  - Nilay Peker
  - Carolien Doorenbos
  - Giuseppe Fleres
  - Inês Mendes
  - Erwin C. Raangs
- Univeristy of Utah and IDbyDNA:
  - Robert Schlaberg
- University of Tubingen:
  - Silke Peter
- University of Munster:
  - Dag Harmsen
  - Allexander Mellman
  - Karola Prier
- Hvidovre Hospital:
  - Henrik Westh
- Örebro University
  - Martin Sundqvist
  - Paula Mölling