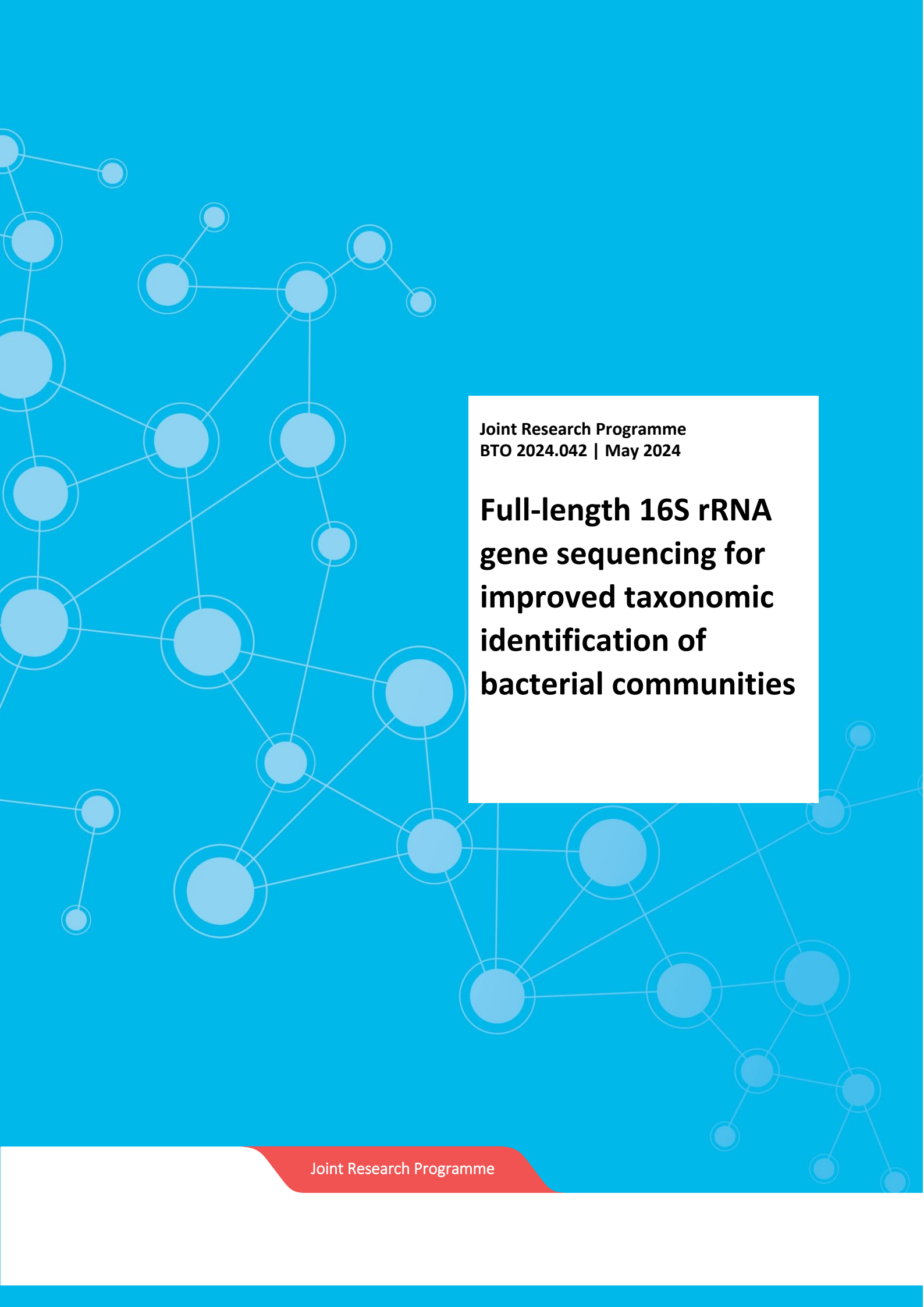




Joint Research Programme
BTO 2024.042 | May 2024

**Full-length 16S rRNA
gene sequencing for
improved taxonomic
identification of
bacterial communities**

Report

Part I: Verkenning brede microbiologische screening tot op soortniveau (Nederlands)

Part II: Species-level identification of bacteria with short- and long-read 16S rRNA gene sequencing methods and analysis software (English)

BTO 2024.042 | May 2024

This research is part of the Joint Research Programme of KWR, the water utilities, and Vewin.

Project number

402045/249

Project manager

Michiel Hootsmans

Client

BTO - Thematical research - Biological safety

Authors

Leo Heijnen, Michiel in 't Zandt and Peer Timmers

Quality Assurance

Paul van der Wielen

Sent to

This report is distributed to BTO-participants.

A year after publication it is public.

Keywords

NGS, 16S rRNA gene, Loop Genomics, QIIME2, species

Year of publishing
2024

More information

Leo Heijnen
T +31 (0)30 606 9743
E leo.heijnen@kwrwater.nl

PO Box 1072
3430 BB Nieuwegein
The Netherlands

T +31 (0)30 60 69 511
E info@kwrwater.nl
I www.kwrwater.nl

KWR

May 2024 ©

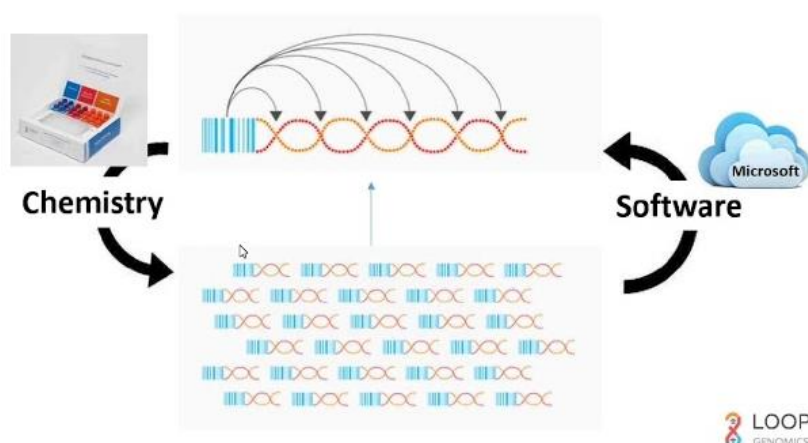
All rights reserved by KWR. No part of this publication may be reproduced, stored in an automatic database, or transmitted in any form or by any means, be it electronic, mechanical, by photocopying, recording, or otherwise, without the prior written permission of KWR.

Managementsamenvatting

Volledige 16S rRNA gen-sequencing geeft een verbeterde taxonomische identificatie van bacteriën en mogelijkheden voor een brede microbiologische screeningsmethode

Auteur(s) Leo Heijnen, Michiel in 't Zandt en Peer Timmers.

Full-length 16S rRNA gen-sequencing blijkt in potentie bruikbaar als “brede microbiologische screening” voor (drink)water. De onderzochte methode, LoopSeq van de firma Loop Genomics is nog niet geschikt om ook bacteriën met lage aantallen (waaronder pathogenen) in het drinkwatermilieu te identificeren. Vervolgonderzoek zal zich richten op toepassingsmogelijkheden van alternatieve full-length 16S rRNA gen-methoden (zoals het recent sterk verbeterde MinION systeem) waarmee mogelijk beter bacteriën kunnen worden geïdentificeerd die in lage aantallen aanwezig zijn.



Principe van de methode van de “Loop Genomics methode” voor het bepalen van de volledige 16S rRNA gen-sequenties.

Belang: volledig beeld microbiologische waterkwaliteit met “brede” screeningsmethode

De analyse van de waterkwaliteit richt zich nu vooral op het aantonen van een beperkt aantal micro-organismen die gebruikt worden als indicatoren voor de aanwezigheid van (fecale) pathogenen of het optreden van nagroei. Met next-generation sequencing (NGS) is het mogelijk om snel veel DNA-sequenties te analyseren. Door NGS-analyses uit te voeren op taxonomische sequenties, zoals het 16S rRNA gen (microbial profiling), is identificatie van een gehele microbiële gemeenschap in één enkele analyse mogelijk. Dit geeft inzicht in de in water aanwezige ongewenste (pathogenen) en gewenste

bacteriën. Dit biedt mogelijkheden voor microbial profiling als universele “brede microbiologische screening”. Tot nu toe vond de analyse echter slechts plaats op een klein deel van het 16S rRNA gen, waardoor identificatie zich beperkt tot genusniveau en niet gaat tot op soortniveau, zodat het identificeren van (opportunistische) pathogenen meestal niet mogelijk is en de toepasbaarheid van de verkregen informatie beperkt blijft. Inmiddels zijn er technieken ontwikkeld om NGS-analyses uit te voeren op het gehele 16S rRNA gen (full-length 16S). Hierbij komt meer taxonomische informatie beschikbaar. Door een dergelijke “brede” NGS-schermingsmethode te ontwikkelen wordt het

mogelijk een volledig beeld te krijgen van de samenstelling van de microbiologische populatie en de aanwezige micro-organismen tot op soortniveau te identificeren. Zo ontstaat meer inzicht in de waterkwaliteit en in het functioneren van drinkwaterzuivering.

Aanpak: eerst verkenning en daarna testen

Er is eerst een verkennende studie uitgevoerd naar de technische mogelijkheden van NGS-methoden voor identificatie van micro-organismen tot op soortniveau. Vervolgens zijn de toepassingsmogelijkheden onderzocht van één geselecteerde veelbelovende methode, de LoopSeq methode van Loop Genomics. Hiervoor zijn analyses uitgevoerd op kunstmatig samengestelde bacteriemengsels en monsters uit de drinkwaterzuivering.

Resultaten: identificatie tot op soortniveau mogelijk bij bacteriën in voldoende concentraties

De analyses met de veelbelovende Loop Genomics methode maken diverse zaken duidelijk:

- Identificatie van bekende bacteriesoorten tot op soortniveau is mogelijk met full-length 16S rRNA gen-sequencing.
- Doordat de meeste bacteriën in drinkwater niet zijn beschreven of gekweekt, zijn de databases beperkt en is identificatie van veel dominante bacteriën uit de drinkwaterzuivering niet mogelijk.
- De Loop Genomics is minder geschikt voor het identificeren van bacteriën met lage aantallen in het drinkwatermilieu.

Toepassing: brede microbiologische screening met microbial profiling is mogelijk

De resultaten van dit onderzoek laten zien dat “brede microbiologische screening” met NGS haalbaar is, maar de methode van Loop Genomics is niet geschikt om daarmee ook micro-organismen te identificeren die in lage aantallen in het drinkwatermilieu aanwezig zijn, zoals ziekteverwekkers. Door recente verbeteringen van het MinION sequencing systeem is dit systeem inmiddels bruikbaar om de sequentie van volledige 16S rRNA genen betrouwbaar te bepalen. Hiermee lijkt dit systeem veelbelovend om toe te kunnen passen voor een “brede microbiologische screening” in het drinkwatermilieu.

Rapport

Dit onderzoek is beschreven in het rapport *Full-length 16S rRNA gene sequencing for improved taxonomic identification of bacterial communities* (BTO 024.042).

Jaar van publicatie
2024

Meer informatie
Leo Heijnen

T +31-654296656
E leo.heijnen@kwrwater.nl

PO Box 1072
3430 BB Nieuwegein
The Netherlands

T +31 (0)30 60 69 511
E info@kwrwater.nl
I www.kwrwater.nl

KWR

May 2024 ©

Alle rechten voorbehouden aan KWR. Niets uit deze uitgave mag - zonder voorafgaande schriftelijke toestemming van KWR - worden vervaelvoudigd, opgeslagen in een geautomatiseerd gegevensbestand, of openbaar gemaakt, in enige vorm of op enige wijze, hetzij elektronisch, mechanisch, door fotokopieën, opnamen, of enig andere manier.

Keywords

NGS, DNA, 16S, bacteriën, populatie, brede screening, Loop Genomics, MinION

Leeswijzer

Het in dit rapport beschreven onderzoek is in twee fasen uitgevoerd. In eerste instantie is er een verkennende bureaustudie uitgevoerd waarin is onderzocht welke mogelijkheden er zijn voor het toepassen van microbial profiling als “brede microbiologische screening” waarbij de samenstelling van de microbiologische populatie tot op soortniveau kan worden geïdentificeerd. De rapportage van de resultaten van deze bureaustudie zijn beschreven in een uitgebreide memo (21 juli 2021) en gedeeld en besproken met de themagroep. Als resultaat van de bureaustudie is er vervolgonderzoek gedefinieerd en uitgevoerd waarin de toepassingsmogelijkheden van potentieel toepasbare “brede screeningsmethoden” is onderzocht. In deze rapportage zijn de resultaten van beide onderdelen beschreven, de hoofdstukken 1 t/m 3 bevatten de resultaten van de bureaustudie (beschreven in het Nederlands) en de hoofdstukken 4 t/m 7 bevatten de resultaten van het (praktisch) onderzoek waarin de mogelijkheden van veelbelovende methoden voor het (drink)wateronderzoek zijn onderzocht (beschreven in het Engels).

Jaar van publicatie
2024

Meer informatie
Leo Heijnen

T +31-654296656
E leo.heijnen@kwrwater.nl

PO Box 1072
3430 BB Nieuwegein
The Netherlands

T +31 (0)30 60 69 511
E info@kwrwater.nl
I www.kwrwater.nl

KWR

May 2024 ©

Alle rechten voorbehouden aan KWR. Niets uit deze uitgave mag - zonder voorafgaande schriftelijke toestemming van KWR - worden vervaelvoudigd, opgeslagen in een geautomatiseerd gegevensbestand, of openbaar gemaakt, in enige vorm of op enige wijze, hetzij elektronisch, mechanisch, door fotokopieën, opnamen, of enig andere manier.

Keywords

NGS, DNA, 16S, bacteriën, populatie, brede screening, Loop Genomics, MinION

Contents

Report 1

Contents	3
1 Inleiding: verkenning mogelijkheden van brede microbiologische screening tot op soortniveau	6
2 Brede microbiologische screening tot op soortniveau	7
2.1 Microbiële identificatie tot op soortniveau als screeningsmethode	7
2.2 Next-generation sequencing-technologieën	8
2.2.1 Illumina sequencing – short-read technologie	10
2.2.2 Uitdagingen en kansen rondom het gebruik van Illumina	10
2.2.3 PacBio – long-read technologie	10
2.2.4 Uitdagingen en kansen rondom het gebruik van PacBio	11
2.2.5 Oxford Nanopore MinION – long-read technologie	11
2.2.6 Uitdagingen en kansen rondom het gebruik van de Oxford Nanopore MinION	12
2.3 Synthetische long-read technologieën voor Illumina sequencing	13
2.3.1 Ontwikkelingen en limitaties van synthetische long-read technologie voor Illumina	13
2.3.2 Loop Genomics long-read kits	15
2.4 Ontwikkelingen in software en databases	16
2.4.1 Software-ontwikkelingen op het gebied van Operational Taxonomic Units (OTUs) en Amplicon Sequence Variants (ASVs)	16
2.4.2 Software-ontwikkelingen voor soortidentificatie met korte DNA-fragmenten	17
2.4.3 Software: kansen voor soortidentificatie met QIIME 2	18
2.4.4 Ontwikkelingen in bestaande databases	18
2.5 Primerkeuze en primerbias	19
2.5.1 Primerkeuze voor het 16S rRNA gen	19
2.5.2 Gedegenereerde 27F-1492R primers voor het 16S rRNA gen	20
2.5.3 Primerkeuze voor het 18S rRNA gen	21
2.5.4 De internal transcribed spacer (ITS) als alternatief voor het 18S rRNA gen	22
2.6 Concrete kansen voor soortidentificatie met 16S en 18S rRNA gen sequencing	22

2.6.1	Validatie van Loop Genomics 16S rRNA gen-kit en de nieuwe Oxford Nanopore chemie	23
2.6.2	Aanvullende mogelijkheden voor soortidentificatie met 16S rRNA gen-sequencing platforms	25
3	Conclusies en aanbevelingen	26
3.1	Conclusies	26
3.2	Aanbevelingen	26
4	Introduction: 16S rRNA gene sequencing for species-level identification	27
4.1	Full-length 16S rRNA gene sequencing for species-level identification	27
4.2	QIIME 2™ species classifier plugin on short 16S rRNA gene v3 reads	28
4.3	Testing approach	29
5	Materials and methods	30
5.1	Selection of samples	30
5.2	Loop Genomics 16S rRNA gene Long Read Kit sample preparation & sequencing	32
5.3	Sequence analysis on the Loop Genomics server	32
5.4	Short fragment 16S rRNA gene sequence analysis	33
6	Results & Discussion	35
6.1	Loop Genomics long-read sequencing technology: general	35
6.2	Sequence identification with the Loop Genomics 16S Long Read kit and provided software	35
6.2.1	General sequence statistics	35
6.3	Mock community: Loop Genomics and short v4 fragment comparison	37
6.3.1	Loop Genomics comparison with short 16S rRNA gene v4 fragment sequencing	40
6.4	Analysis of water, sediment and biofilm samples.	43
6.4.1	Species and genus identification in all samples	43
6.4.2	Samples from production location “De Laak” (Oasen)	45
6.4.3	Samples from water and monitoring wells from water treatment plant “De Lange Vlieter” (WML)	50
6.4.4	Samples from the GOBAM project	51
6.5	Discussion: full-length 16S rRNA gene sequencing using Loop Genomics	51
7	Conclusions and Recommendations	53
7.1	Conclusions	53

7.2	Recommendations	54
8	References	55
8.1	BTO reports, KWR reports and newsletters	56
I	Bijlage: Begrippenlijst	58
II	Identification of the Mock community with “short (v4) sequencing	65
III	Top 25 taxons in samples “De Laak”	66
IV	Top 25 taxons in samples “De lange Vlieter”	69
V	Top 25 taxons in samples from “GOBAM”	72

1 Inleiding: verkenning mogelijkheden van brede microbiologische screening tot op soortniveau

De ontwikkelingen op het gebied van **next-generation sequencing (NGS)** hebben het mogelijk gemaakt om snel veel DNA-sequenties te analyseren. Hierdoor zijn microbiële diversiteitsanalyses de afgelopen jaren een relatief goedkope en standaard methode geworden. Een belangrijke ontwikkeling hierbij is de toepassing van NGS op taxonomische sequenties zoals van het **16S rRNA gen** (aanwezig in alle **prokaryote** organismen, zoals bacteriën) (Figuur 1) of het **18S rRNA gen** (aanwezig in alle **eukaryote** organismen, zoals **schimmels** en **protozoa**).

Met deze technologie is zogeheten 'microbial profiling', de identificatie van gehele microbiële gemeenschappen in één enkele analyse mogelijk. Hierbij vormt **Illumina** sequencing de standaard. Een brede toepassing van deze methode in verschillende BTO onderzoeken (zie onder andere BTO 2020.044 Microbiologische compositie en potentiële metabole processen tijdens drinkwaterdistributie, BTO 2020.206(s) Verwijdering OMV's (organische microverontreinigingen en Microbial profiling met NGS voor drinkwateronderzoek, BTO 2021.054) tijdens bodempassage in het infiltratiebekken De Lange Vlieter & BTO 2017.033 Analyse van de bacteriepopulatie met NGS in water, sediment en biofilm van het leidingnet in Kralingen en Braakman in 2014 en 2015 (Evides Waterbedrijf) heeft laten zien dat microbial profiling waardevolle inzichten kan geven in bijvoorbeeld microbiële veranderingen in verschillende zuiveringsstappen, de distributie en ook de introductie van verontreinigingen in drinkwater. Het lijkt daarnaast een waardevolle techniek om de herkomst van water te bepalen (BTO 2020.206(s) Verwijdering OMV's tijdens bodempassage in het infiltratiebekken De Lange Vlieter & KWR 2021.013 NGS bronopsporing pilot inlaatwater Woerden). Een limitatie van de huidige methoden is dat microbiële identificatie niet voor alle sequenties mogelijk is tot op **soortniveau**, waardoor de resolutie beperkt is. Echter, voor de detectie van (opportunistische) pathogenen of indicatororganismen is identificatie tot op soortniveau en in een aantal gevallen zelfs tot op **stamniveau** gewenst. In genera die ook niet-pathogene soorten bevatten, zoals *Legionella*, is de identificatie tot op soortniveau bijvoorbeeld noodzakelijk om uitspraken te kunnen doen over aan- en afwezigheid van pathogene soorten van dit **genus**.

Vanwege het groot aantal begrippen dat binnen sequencing en sequencing-analyse wordt gebruikt is in dit rapport tevens een begrippenlijst opgenomen (Bijlage I). Alle begrippen die in deze begrippenlijst zijn opgenomen zijn tijdens het eerste gebruik in de tekst vetgedrukt.

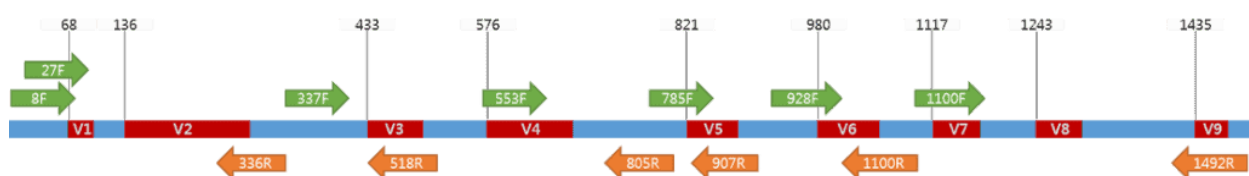
2 Brede microbiologische screening tot op soortniveau

2.1 Microbiële identificatie tot op soortniveau als screeningsmethode

Wanneer identificatie tot op soortniveau met NGS mogelijk is, kan deze techniek worden ingezet als brede screeningsmethode om veranderingen en mogelijke besmettingen van het leidingwater te achterhalen na werkzaamheden of calamiteiten. Ook kan deze brede screening tot op soortniveau inzicht geven in hoeverre en onder welke omstandigheden (opportunistische) pathogenen voorkomen in het Nederlandse drinkwater en welk risico dit mogelijk vormt voor de volksgezondheid. Daarnaast kan mogelijk ook meer inzicht verkregen worden in de potentiële biochemische karakteristieken van de geïdentificeerde soorten, waardoor de aan- of afwezigheid van bepaalde micro-organismen in verschillende zuiveringsstappen kan worden geduid.

Om microbiële identificatie tot op soortniveau mogelijk te maken kan worden gedacht aan I) de aanpassing van huidige **short-read technologieën**, II) de toepassing van **long-read technologieën** waarmee het hele 16S of 18S rRNA gen kan worden gesequenced en III) de aanpassing van analysemethoden (software en databases). Hierna wordt tevens kort ingegaan op de primerkeuze en de effecten van primerkeuzes op diversiteitsanalyses. De primerkeuze is in principe ondergeschikt aan de gekozen sequencing-technologie, maar vormt een belangrijke afweging in elk NGS-experiment.

Het doel van deze verkenning is om de ontwikkelingen en aanpassingen in kaart te brengen die 16S en 18S rRNA gen-gebaseerde identificatie tot op soortniveau mogelijk maken. In deze verkenning wordt grotendeels over 16S rRNA gen-gebaseerde analyses gesproken, aangezien hier het meeste onderzoek aan is verricht. Ook kunnen de methoden voor het 16S rRNA gen op andere (functionele) genen worden toegepast.



Figuur 1 Schematisch overzicht van het 16S rRNA gen.

In het overzicht zijn de 9 variabele regionen in rood gelabeld van V1 t/m V9 en is de nucleotidepositie van het begin van de variabele regio zichtbaar (volgens het referentiegenoom van Escherichia coli). Een aantal veelgebruikte primers voor PCR-gebaseerde analyses zijn bovenaan weergegeven in groen (forward primers) en onder in oranje (reverse primers). Bron: EZ BioCloud.

Tevens gelden voor de identificatie van het 18S rRNA gen van eukaryoten, in de praktijk voornamelijk voor schimmels en protozoa, gelden dezelfde uitdagingen als beschreven voor het 16S rRNA gen. Bovendien geldt dat het 18S rRNA gen minder geschikt is voor classificatiedoeleinden. In de praktijk wordt daarom de combinatie van het 18S rRNA gen en de **Internal Transcribed Spacers (ITSs)** gebruikt. Hier wordt in paragraaf 2.3.2 dieper op ingegaan.

Leeswijzer

In paragraaf 2.2 wordt allereerst ingegaan op de beschikbare sequencing-technologieën en de ontwikkelingen die in de afgelopen jaren plaats hebben gevonden (2.2.1: Illumina sequencing, 2.2.3: PacBio sequencing, 2.2.5: Minlon sequencing). Ook worden de uitdagingen en kansen voor de verschillende methoden omtrent identificatie tot op soortniveau besproken (2.2.2, 2.2.4, 2.2.6). In 2.3 worden concrete kansen rondom synthetische long-read technieken met Illumina sequencing besproken. In 2.4 wordt ingegaan op relevante ontwikkelingen in analysesoftware en databases. Vervolgens wordt in 2.5 specifiek voor lange reads de **primerkeuze** en bijbehorende

biases besproken. Tot slot worden in 2.6 de concrete kansen voor soortidentificatie met 16S en 18S rRNA gen sequencing benoemd.

2.2 Next-generation sequencing-technologieën

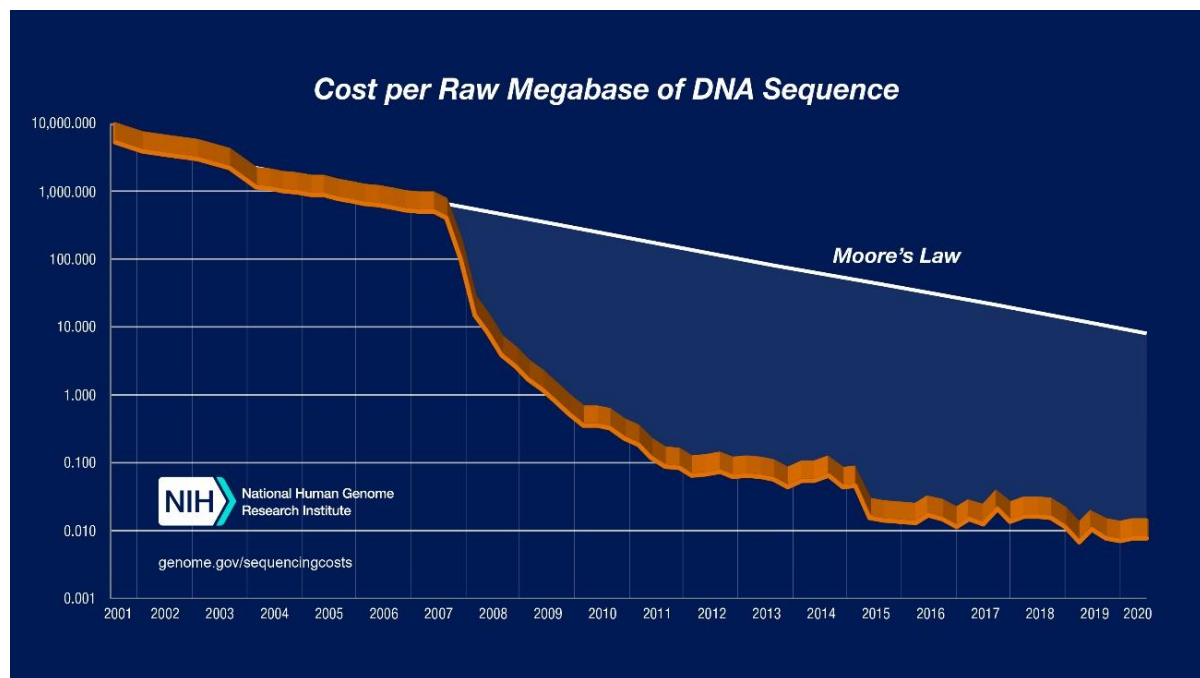
In de afgelopen 15 jaar hebben op het gebied van **DNA-sequencing** grote ontwikkelingen plaatsgevonden. De introductie van next-generation sequencing (NGS), ook wel high-throughput sequencing genoemd, leidde allereerst tot een sterke daling in de kosten (Figuur 2). Daarnaast kwamen tientallen nieuwe methoden op de markt. Al deze NGS-technieken maakten het mogelijk om enorme hoeveelheden DNA (in de orde van giga**basenparen**/Gbp) in één keer te sequencen. Allereerst heeft dit tot de ontwikkeling van **metagenoom sequencing** geleid, waarbij het DNA van gehele monsters kan worden gekarakteriseerd. Daarnaast heeft ook '**amplicon sequencing**', de methode waarbij korte fragmenten DNA met PCR worden geamplificeerd en vervolgens gesequenced, een enorme vlucht genomen.

Metagenoom sequencing, of 'full metagenomics' wordt gebruikt om het volledige (microbiële) **potentieel** van een ecosysteem in kaart te brengen. Belangrijke stappen bij metagenoom sequencing zijn 1) **annotatie**, het proces waarbij de locatie en identiteit van genen op de DNA-sequenties worden geïdentificeerd en 2) metagenoom **binning**, het proces waarbij individuele genomen van micro-organismen worden gegenereerd. Met metagenoom-data kan naast de identiteit ook informatie worden verkregen over de potentiële functies van micro-organismen. Deze techniek staat centraal bij vragen rondom de functie van specifieke micro-organismen in een (eco)systeem. Een van de grote uitdagingen van deze techniek is dat veel sequencing-data en uitgebreide arbeidsintensieve software-analyses nodig zijn om goede uitspraken te kunnen doen. Hier wordt uitgebreid op ingegaan in onderdeel 3 van het overkoepelende BTO-project waarvan deze studie deel uitmaakt.

Bij amplicon sequencing wordt één specifiek gen in grote aantallen gesequenced. Het voordeel van amplicon sequencing is dat met universele genen, zoals het 16S rRNA gen dat aanwezig is in alle bacteriën, relatief snel hele microbiologische populaties in kaart kunnen worden gebracht. Bijkomend voordeel van de NGS-methoden is dat door de hoge throughput meerdere monsters in één keer kunnen worden gekarakteriseerd.

Bijvoorbeeld voor vergelijkingsstudies van verschillende zuiveringsstappen of experimentele condities en veranderingen in microbiële populaties over de tijd is amplicon sequencing ideaal. Om meer uit de amplicon data te kunnen halen, is identificatie tot op soortniveau een grote meerwaarde. Aan de hand van beschikbare informatie over specifieke soorten kan hiermee namelijk veel gerichter en nauwkeuriger uitspraak gedaan worden over potentiële biochemische karakteristieken. De laatste jaren hebben een aantal grote ontwikkelingen op het gebied van sequencing-technologie en sequencing-chemie plaatsgevonden, waardoor soortidentificatie steeds haalbaarder wordt. In deze studie wordt ingegaan op mogelijkheden voor soortidentificatie met amplicon sequencing-strategieën. Voor meer detail over de nieuw ontwikkelde chemie en software bij verschillende sequencing-methoden, zie de Nieuwsbrief Kansen van nieuwe microbiologische methoden voor toepassing in de drinkwatersector van juli 2021.

Illumina en **PacBio** sequencing worden momenteel als standaardmethoden in het onderzoek gebruikt. Daarnaast maakt de **Minlon** van **Oxford Nanopore** de laatste jaren sterke ontwikkelingen door, waardoor de verwachting is dat deze techniek in de toekomst onderdeel zal uitmaken van de vaste DNA sequencing-methoden. In deze vergelijking is **Sanger Sequencing** als standaard eerste generatie sequencing-technologie meegenomen (Tabel 1).



Figuur 2 Overzicht van de sequencing-kosten per ruwe megabase aan DNA.
Data afkomstig van het National Human Genome Research Institute.

Tabel 1 Overzicht van veelgebruikte next-generation sequencing (NGS)-technologieën en Sanger sequencing ter vergelijking.

Technologie	Throughput (Gb)	Read-lengte	Sterke punten	Uitdagingen
Sanger Sequencing ABI3500/3730	0,0003	Tot 1 kb	-Hoge accuraatheid -Lange reads	-Hoge per-base kosten -Lage throughput
Illumina MiSeq	0,3-15	Tot 2x300 basen (paired-end)	-Hoge accuraatheid -Hoge throughput	-Korte reads -Sequencing duurt lang
PacBio RSII	0,5-1	Tot 60 kb (gemiddeld 10 kb)	-Lange reads -Snelle sequencing	-Lage accuraatheid -Hoge concentratie input DNA nodig (>100 ng/μL)
Oxford Nanopore MinION	0,1-1	Gemiddeld >10 kb, tot 1 Mb	-Weinig apparatuur nodig (draagbaar) -Lange reads -Live uitleesbare data	-Lage accuraatheid -Gevoelige chemie -Techniek nog niet geoptimaliseerd

2.2.1 Illumina sequencing – short-read technologie

Illumina biedt een groot aantal van de meest bekendste en ver ontwikkelde sequencing-platformen aan



Figuur 3). Het onderzoekslaboratorium van KWR beschikt over een Illumina MiSeq waarmee op dit moment met grote regelmaat 16S rRNA gen-amplicon libraries worden gerund.

Illumina sequencing is gebaseerd op PCR-amplificatie van individuele stukken DNA vóór de daadwerkelijke sequencing. Doordat er zoveel kopieën van hetzelfde molecuul worden gesequenced, is het signaal veel sterker en is de kans op sequencing-fouten relatief laag. Sequencing vindt plaats door het stapsgewijs inbouwen van **nucleotiden** (A, C, T en G). Bij het inbouwen van een nucleotide wordt door een chemische reactie licht afgegeven dat door een sensor wordt gedetecteerd.

Met een Illumina MiSeq kan 15 Gb aan data per run worden gegenereerd, wat neerkomt op 25 miljoen **reads**. Deze reads kunnen tot 600 basenparen lang zijn met de zogeheten ‘paired-end technologie’ waarbij een DNA-sequentie vanaf beide uiteinden voor 300 basenparen wordt gesequenced. In de praktijk dienen deze beide fragmenten een overlap te hebben om vervolgens met software te kunnen worden omgezet in één DNA-fragment. Daardoor zijn de werkelijke te analyseren fragmenten dus wat korter.

2.2.2 Uitdagingen en kansen rondom het gebruik van Illumina

Illumina biedt met de MiSeq een relatief goedkope sequencer die uitermate geschikt is voor amplicon-gebaseerde studies. De hoge throughput en lage foutenmarge bieden daarnaast voordelen ten opzichte van de huidige beschikbare long-read methoden (zie 2.2.3 t/m 2.2.6).

Voor 16S rRNA gen-sequencing worden doorgaans fragmenten van de V3- en V4-regio gekozen. Dit zijn twee variabele regio's op het 16S rRNA gen die zeer geschikt zijn voor microbiële diversiteits-analyses. Voor deze regio's zijn goede primersets voor PCR-amplificatie beschikbaar. Echter, identificatie tot op soortniveau is hiermee (nog) niet mogelijk. Ook voor 18S analyses speelt een vergelijkbare uitdaging. Mogelijkheden om met Illumina sequencing naar soortniveau te gaan zijn I) de ontwikkeling van synthetische long reads (2.3.1) II) de amplificatie van een langer 16S rRNA gen-fragment en III) de combinatie met long-read technologieën (2.3.2).



Figuur 3 Overzicht van een aantal Illumina sequencing platforms.

Van links naar rechts zijn Illumina platforms te zien met een toenemende throughput. Bron: Enseqlopedia.

2.2.3 PacBio – long-read technologie

Long-read technologieën maken het mogelijk om de hele lengte van het 16S en het 18S rRNA gen te sequencen. De klassieke long-read technologie wordt geleverd door **PacBio** en maakt gebruik van de zogeheten **Single Molecule, Real Time (SMRT)**-technologie (zie ook de Nieuwsbrief: kansen van nieuwe moleculair-microbiologische methoden in de drinkwatersector, juni 2019, paragraaf 2.2.3). Hierbij vindt sequencing plaats zonder PCR-amplificatie. De

methode werkt op basis van DNA-synthese, het proces waarbij een enkelstrengs DNA-streng dubbelstrengs wordt gemaakt door de activiteit van het enzym **DNA polymerase** dat stapsgewijs de verschillende nucleotiden (A, C, T en G) inbouwt. Detectie vindt plaats door een uniek fluorescent label aan de vier verschillende nucleotiden. Het breedst toegepast model is de PacBio RSII (Figuur 4).



Figuur 4 De PacBio RSII.

Bron: Hudson Alpha Genome Sequencing Center.

2.2.4 Uitdagingen en kansen rondom het gebruik van PacBio

De grootste uitdaging bij PacBio RS sequencing is op dit moment de foutmarge van 12-13%. Dit komt vooral door de zeer gevoelige detectie van de fluorescente signalen die niet door PCR zijn versterkt, zoals bijvoorbeeld bij Illumina sequencing het geval is. De hoge foutmarge is voor soortidentificatie op 16S of 18S rRNA gen-niveau problematisch. PacBio biedt tegenwoordig de **Circular Consensus Sequencing (CCS)** aan waarbij de read vóór sequencing wordt geamplificeerd en vervolgens parallel gesequenced wordt. Hierdoor wordt het signaal verhoogd en worden sequencing-fouten sterk gereduceerd. De zogeheten **high fidelity (HiFi)** reads hebben met een door PacBio gerapporteerde accuraatheid van 99.8% een vergelijkbare kwaliteit als Illumina MiSeq reads (minimaal 99.9% accuraatheid).

De hoge aanschaffkosten (> €300.000) van een PacBio-machine maken dat deze methode in de regel wordt uitbesteed voor onderzoek. De sequencingkosten zijn circa 5-10x zo hoog per Gb data (en hoger voor de HiFi methode) vergeleken met Illumina Sequencing. Voor de analyse van een groot aantal monsters is de methode daarmee minder interessant. Vooral voor zeer selecte en specifieke monsters is de methode geschikt.

2.2.5 Oxford Nanopore MinION – long-read technologie

Een interessante ontwikkeling op het gebied van long-read technologie is de in 2012 geïntroduceerde **Oxford Nanopore MinION** die deel uitmaakt van de Oxford Nanopore serie (Figuur 5). De MinION is met een \$1.000 'access fee' zeer betaalbaar. De methode werkt op basis van eiwit **nanoporiën** in een **polymere membraan**. Over de membraan staat een ionische stroom, waardoor er een continue stroom van de in de buffer aanwezige ionen door de poriën plaatsvindt. Wanneer een DNA-molecuul door een nanoporie gaat, zorgt dit voor een verandering in de ionenstroom. Deze verandering is specifiek voor elke nucleotide (A, C, T en G) die het centrum van de porie passeert. Zo kan zonder extra PCR-amplificatie de volgorde van een DNA-molecuul worden bepaald. Het grootste voordeel van de MinION is dat er in principe geen lengtebeperking voor de DNA-sequentie is. In de praktijk zijn DNA-moleculen met een lengte van meer dan 1 miljoen basenparen gesequenced.

In een eerder BTO-onderzoek (BTO 2019.057 Verkenning van de toepassingsmogelijkheden van MinION sequencing) wordt een gedetailleerde introductie van de MinION methode gegeven.



Figuur 5 Overzicht van de Oxford Nanopore platforms.

Centraal in de figuur is de MinION weergegeven. Relevant voor hogere throughput zijn de GridION (voor 5 flow cells) en de PromethION (voor 48 flow cells). Bron: Oxford Nanopore Technologies

2.2.6 Uitdagingen en kansen rondom het gebruik van de Oxford Nanopore MinION

De grote uitdaging van de MinION is het hoge foutenpercentage van de reads. Door ontwikkelingen in de samenstelling van de poriën, de chemie en software voor het “vertalen” van ruwe signalen naar DNA-sequenties ligt deze momenteel zo rond de 10-15%. Ook het optreden van **inserties** (het door sequencing onjuist toevoegen van een nucleotide in een DNA-sequentie) en **deleties** (het door sequencing wegvallen van nucleotiden in een DNA-sequentie) vormen momenteel uitdagingen. Met softwarematige correcties kan de accuraatheid worden verhoogd tot circa 98%, maar vooral de inserties en deleties blijven vooralsnog een uitdaging vormen.

Een verkennend BTO-onderzoek (BTO 2018.044 Toepassingsmogelijkheden van Metatranscriptomics, Microbial profiling en MinION sequencing) heeft eerder de uitdagingen rondom soortidentificatie met het 16S rRNA gen aangetoond. Uit onderzoek met een zogeheten ‘**mock community**’ waarin een bekende hoeveelheid DNA van bekende soorten aanwezig is, blijken de meeste soorten te worden aangetroffen. Onjuiste identificatie betrof vooral soorten van dezelfde genus. Opvallend was dat voor *Escherichia coli* ook andere genera (*Shigella*, *Yersinia*) werden aangetroffen in de data.

Vooralsnog limiteert dit de mogelijkheid om de MinION direct als methode toe te passen voor de analyse van de microbiologische diversiteit. Wel vinden snelle ontwikkelingen plaats die de technologie in rap tempo verbeteren. Deze ontwikkelingen beslaan zowel de chemie (de aangeboden kits) en de software voor basecalling. Met de recent uitgebrachte flow cells (R10.3, geïntroduceerd in 2019) en de **Bonito** CRF analyse-software ontwikkeld door Oxford Nanopore wordt een 98.3% ‘single read accuracy’ geclaimd en door gebruikers zou ook een >98% accuraatheid zijn gerapporteerd. In de praktijk zal >95% accuraatheid haalbaar moeten zijn met gebruik van de huidige beschikbare R9 4.1 en R10.3 chemie. De beste resultaten worden verkregen met de R10.3 chemie.

Ook heeft Oxford Nanopore aangekondigd de nieuwe Q20+ chemie vanaf juni 2021 op de markt te gaan brengen. In eerste instantie zal de nieuwe chemie door een beperkt aantal gebruikers in de praktijk worden onderzocht, KWR heeft zich hiervoor aangemeld. Met de nieuwe chemie is in eerdere onderzoeken van Oxford Nanopore zelf een accuraatheid van meer dan 98% bereikt.

Recentelijk is tevens door Karst en co-auteurs het gebruik van **unique molecular identifiers (UMIs)** geïntroduceerd (Karst et al. 2020). UMIs zijn bekende DNA-sequenties met een bekende lengte van 18 nucleotiden. Deze UMIs worden enerzijds gebruikt voor het clusteren van sequenties en anderzijds voor kwaliteitscontrole en filtering. In de test op een bacteriële ‘mock community’ leverde deze techniek een foutpercentage van slechts 0.0049%. Door de bovenstaande ontwikkelingen die elkaar snel opvolgen kan de MinION binnen enkele jaren een standaard methode worden voor de sequencing van het gehele 16S en 18S rRNA gen met hoge accuraatheid. De grootste limitatie voor implementatie op het moment is dat bovenstaande methoden nog in ontwikkeling zijn, waardoor er nog geen

gestandaardiseerde protocollen beschikbaar zijn. Het is echter waardevol om, zo gauw dit mogelijk is, de nieuwe ontwikkelingen (R10.3 flow cell, Q20+ chemie en Bonito CRF basecalling) toe te passen op relevante drinkwatermonsters om de huidige potentie van de Oxford Nanopore-methode vast te stellen. De optie tot validatie van de nieuwe methode is daarom opgenomen in het validatievoorstel en KWR is in contact met Oxford Nanopore om mee te kunnen draaien in de eerste praktijkonderzoeken.

2.3 Synthetische long-read technologieën voor Illumina sequencing

2.3.1 Ontwikkelingen en limitaties van synthetische long-read technologie voor Illumina

Er zijn verschillende onderzoeken gedaan naar de mogelijkheden om met aanpassingen in het sequencing-protocol en met software-ontwikkelingen zogeheten synthetische 'Illumina long-reads' te genereren.

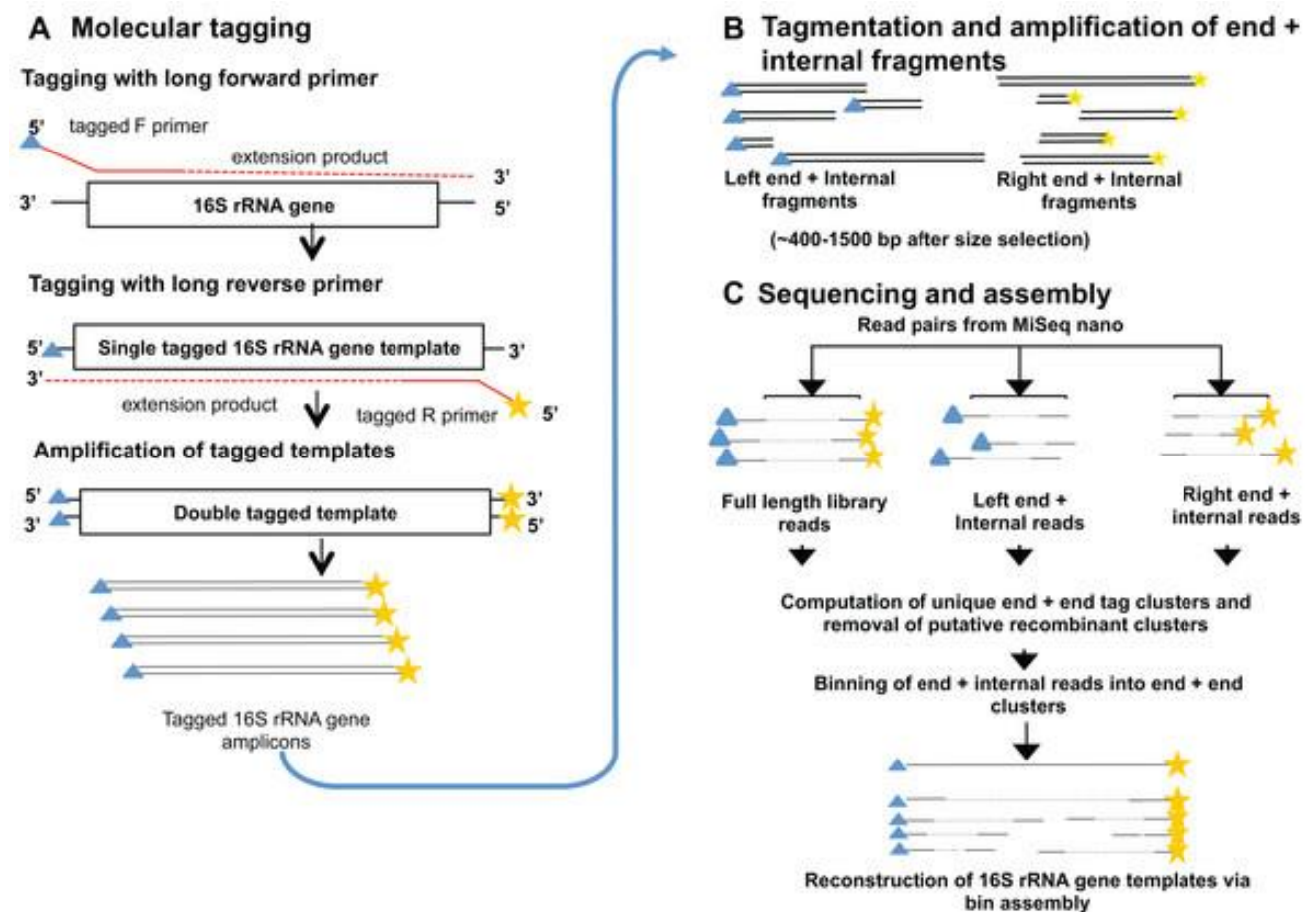
Het principe achter synthetische long reads is beschreven door Stapleton *et al.* en tevens als methode gebruikt voor 16S rRNA gen-gebaseerde analyse door Burke & Darling (Burke and Darling 2016; Stapleton *et al.* 2016). Het principe van een synthetische long read is het softwarematig genereren van volledige 16S rRNA gen-sequenties met short-read sequencing-data. Hieronder wordt het principe in meer detail uitgelegd aan de hand van twee sleutelonderzoeken.

Bij de techniek beschreven door Burke en Darling wordt het 16S rRNA gen getagd met een forward en reverse primer die beiden een toegevoegde unieke en voor de analyse herkenbare **tag** hebben (Figuur 6). Voor meer informatie over primers zie paragraaf 2.4. Het belangrijke principe van de tagging is dat één unieke DNA-sequentie gekoppeld wordt aan één enkel DNA-molecuul. In dit geval wordt dus elk uniek 16S rRNA gen-fragment met unieke tags gelabeld. Dan vindt de amplificatie plaats en wordt een groot aantal kopieën van elk uniek 16S rRNA gen gemaakt.

Vervolgens wordt een deel van de 16S rRNA gen-moleculen met tags gefragmenteerd. Dit proces wordt **tagmentation** (een verkorting van 'tag fragmentation') genoemd. Het resultaat van dit proces zijn fragmenten met een lengte variërend tussen 400 en 1500 basenparen. De 400 basenparenfragmenten beslaan een deel van het 16S rRNA gen, de 1500 basenparen nagenoeg het gehele gen. De fragmenten hebben ofwel de linker sequence tag (blauwe driehoek) of de rechter sequence tag (gele ster). Voor de sequencing worden ook gehele 16S rRNA gen-moleculen meegenomen die zowel de linker als rechter sequence tag hebben. Na de sequencing kan met de informatie van de gehele 16S rRNA genen waarop beide tags liggen de linker en rechter fragmenten die erbij horen worden verzameld. Vervolgens wordt deze sequencing-data met software geassembleerd in lange 16S rRNA gen-sequenties. De tags zijn hierbij cruciaal. De grootste limitatie van deze methode is dat na optimalisatie op de Illumina MiSeq naar schatting ongeveer 80,000 sequences kunnen worden gegenereerd.

Een variatie van de hierboven beschreven methode biedt hiervoor deels een oplossing. Karst *et al.* hebben een hogere throughput methode ontwikkeld waarbij tot een miljoen 16S/18S rRNA gen-sequences kunnen worden gegenereerd zonder het gebruik van 16S rRNA gen-primers (Karst *et al.* 2018). Deze methode werkt op basis van **RNA-extractie** en selectieve opzuivering op basis van grootte van het 16S/18S rRNA gen. Hierna wordt elk fragment uniek getagd. Vervolgens worden hiervan korte fragmenten met dezelfde tags gemaakt. Na sequencing kunnen de korte reads op basis van de labels worden gesorteerd ('**binning**') en kunnen volledige 16S/18S sequenties worden gereconstrueerd. Bestaande uitdagingen zijn dat 16S rRNA gen-fragmenten slechts 24-76% van de reads uitmaken doordat vooral gefragmenteerde **large subunit (LSU)** fragmenten voor interferentie zorgen. In totaal levert een enkele MiSeq run tot ongeveer 55 duizend rRNA reads op, waarvan dus een deel gefragmenteerde LSU-reads zijn. Echter, door de relatief grote hoeveelheid sequencing-data per 16S rRNA gen, is de data van hoge kwaliteit. De door Karst *et al.* ontwikkelde methode is door enkele tientallen studies gebruikt (59 citaties in Crossref), waaronder voor fecale bacteriën in koeienmest (Hagey *et al.* 2019), de identificatie van bacteriestammen in het humane microbioom (Yan *et al.* 2020) en voor het in kaart brengen van de microbiële diversiteit van een boreaal meer in Zweden (Juottonen *et al.* 2020). De methode blijkt daarmee geschikt te zijn voor een brede toepassing. Voor beide methoden geldt dat voor een hogere throughput de Illumina HiSeq aan te bevelen is, waarbij tot 10x

meer rRNA reads worden gesequenced in één enkele run.



Figuur 6 Overzicht van de synthetische long read methode.

De uitleg van de methode is te vinden in paragraaf 2.3.1. Bron: Burke & Darling (2016).

De verschillende stappen van de long read methode in figuur 6 zijn beschreven in Burke & Darling (2016).

Tijdens 'molecular tagging' worden 16S rRNA gen-moleculen getagd met unieke tags in 2 stappen. In stap 1 wordt de 'forward primer' met unieke tag ingebouwd, in stap 2 wordt de 'reverse primer' met unieke tag ingebouwd. De blauwe driehoeken en gele sterren geven de unieke 10 nucleotiden lange tags weer. Vervolgens vindt amplificatie van de dubbel getagde fragmenten plaats.

Tijdens de 'tagmentation' vindt de tagmentatie plaats. Dit is de fragmentatie van het lange 16S rRNA gen-molecuul in kleinere stukken. Deze kleinere stukken bevatten de forward tagged primer (blauwe driehoek) of de reverse tagged primer (gele ster). Voor de sequencing worden aan de binnenkant van de fragmenten de sequencing adapters ingebouwd.

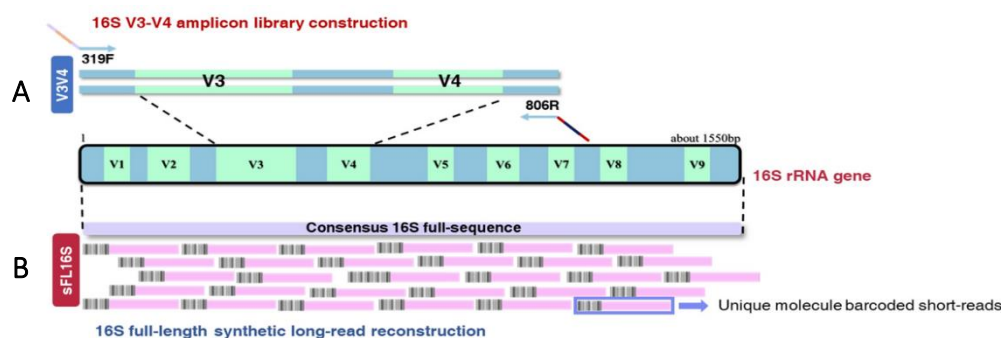
Tot slot vindt sequencing plaats van fragmenten met verschillende lengtes en gehele 16S rRNA moleculen (met beide adapters). Vervolgens kunnen met de informatie van de adapters alle korte 16S-fragmenten die bij eenzelfde molecuul horen worden verzameld en aan elkaar worden gepuzzeld (dit proces heet 'bin assembly').

Naast uitdagingen rond de opschaling van de hierboven beschreven 'synthetic long read' methode zijn de methoden arbeidsintensief, mede doordat er nog geen standaard protocollen zijn ontwikkeld. Dit geeft uitdagingen voor toepassing in huidige onderzoeken. Goede alternatieven zijn commercieel beschikbare kits met een zeer gestroomlijnd protocol. Een long-read methode is in het verleden aangeboden door Illumina (www.illumina.com/science/technology/next-generation-sequencing/long-read-sequencing.html), maar is momenteel niet meer beschikbaar. Recentelijk zijn op basis van de ontwikkelingen van synthetische long-read

technologieën nieuwe commercieel beschikbare long-read kits voor Illumina sequencing ontwikkeld door Loop Genomics. Op de techniek van Loop Genomics wordt hieronder (2.3.2) dieper ingegaan.

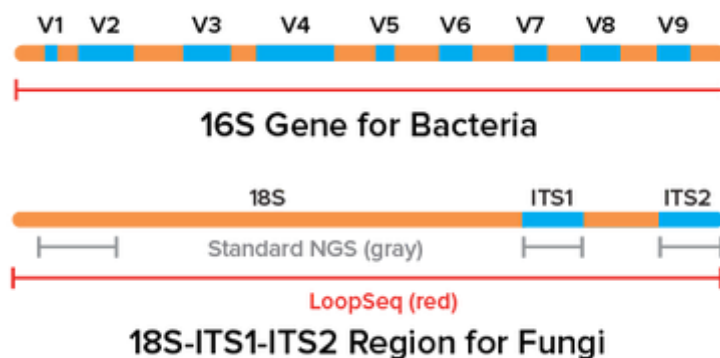
2.3.2 Loop Genomics long-read kits

Recentelijk is door het bedrijf Loop Genomics een methode ontwikkeld voor long read sequencing met Illumina. Deze methode is gebaseerd op de techniek ontwikkeld door Stapleton et al. in 2016 (Stapleton et al. 2016; Callahan et al. 2020). Deze techniek is niet gelimiteerd tot het prokaryote 16S rRNA gen, maar biedt daarnaast ook de mogelijkheid voor sequencing van het eukaryote 18S rRNA gen of zelfgekozen lange DNA-fragmenten tussen 1.000 en 3.000 basenparen te sequencen (en tot 6.000 basenparen met de aangeboden service bij Loop Genomics). Voor het 16S en 18S rRNA gen zijn de aangeboden kits gebaseerd op sequencing van het volledige rRNA-gen (V1-V9 variabele regio voor het 16S rRNA gen (Figuur 7) Ook wordt apart de methode voor sequencing van 18S en de Internal Transcribed Spacer (ITS) die geschikt is voor de identificatie van schimmels aangeboden. Daarnaast kunnen ook eigen specifieke genproducten met een lengte tot 4000-6000 basenparen gebruikt worden, waarmee bijvoorbeeld de aanwezigheid en diversiteit van specifieke functionele genen in een microbiële populatie kan worden bepaald. Één enkele kit is geschikt voor de analyse van 24 monsters die in één MiSeq run kunnen worden geanalyseerd. Dit levert per monster tussen de 5.000 en 10.000 lange 16S rRNA gen-sequenties op. Relevant voor monsters met lage DNA-concentraties is verder dat een minimale hoeveelheid van 2-200 picogram DNA voldoende is voor analyse. Aangezien er geen aanvullende aanpassingen nodig zijn van de bestaande sequencing-methode, bieden de long-read kits een interessant alternatief voor de echte long-read technieken.



Figuur 7 Overzicht van de A) amplicon-methode en B) de long-read methode voor 16S rRNA gen-analyse

Overzicht van A) de V3-V4 amplicon-methode (V3V4) en B) de consensus 16S rRNA gen long-read methode (sFL16S) zoals beschreven in Jeong et al. (2021). Voor de consensusmethode zijn de synthetische long-reads zichtbaar die door de random verdeling uiteindelijk het volledige 16S rRNA gen bedekken.



Figuur 8 Overzicht van de door Loop Genomics getargete gebieden van het 16S rRNA gen en het 18S rRNA gen en de ITS regionen

In de figuur is te zien dat voor het 16S rRNA gen alle variabele regionen (V1 t/m V9) getarget worden. Voor de 18S analyse met de specifieke 18S rRNA gen-kit wordt naast het 18S rRNA gen ook de ITS regio 1 (ITS1) en ITS regio 2 (ITS2) getarget.

Recentelijk is de 16S rRNA gen-sequencing met de Loop Genomics kit gevalideerd met de ZymoBIOMICS Microbial Community DNA Standard (Callahan et al. 2020). In deze mock community zitten 8 bacteriële stammen van *Bacillus subtilis*, *Enterococcus faecalis*, *Escherichia coli*, *Lactobacillus fermentum*, *Listeria monocytogenes*, *Pseudomonas aeruginosa*, *Salmonella enterica*, en *Staphylococcus aureus*. De sequencing-output bevatte reads van gemiddeld 1500 nucleotiden (nt) lengte met hoge accuraatheid (94.6% van de reads zonder fouten) vergelijk met 100-300 nt sequencing (50-80% van de reads zonder fouten). In vergelijking met PacBio circular consensus (CCS), momenteel de gouden standaard in long-read sequencing, lijkt de techniek van Loop Genomics beter te presteren.

Uit een praktijkstudie op 24 fecale monsters bleek dat de methode van Loop Genomics vergeleken met sequencing van de 16S rRNA V3-V4 regio in staat is om misidentificatie te voorkómen (Jeong et al. 2021). Deze misidentificatie ontstaat doordat slechts één of enkele variabele regionen van het 16S rRNA gen worden gekarakteriseerd.

Ook is de methode getest op een aantal watermonsters (aquarium, vijver, regenwater uit een regenton en tuinvijver) (Chung et al. 2020). Hierbij bleken voldoende full-length 16S rRNA gen-sequenties te kunnen worden gegenereerd (>3.5 duizend sequenties voor elk monsters) en voor de meeste, maar niet alle reads was soortidentificatie mogelijk (percentages niet gerapporteerd). Het aantal unieke soorten dat in deze monsters kon worden gedetecteerd varieerde tussen de 118 en 332.

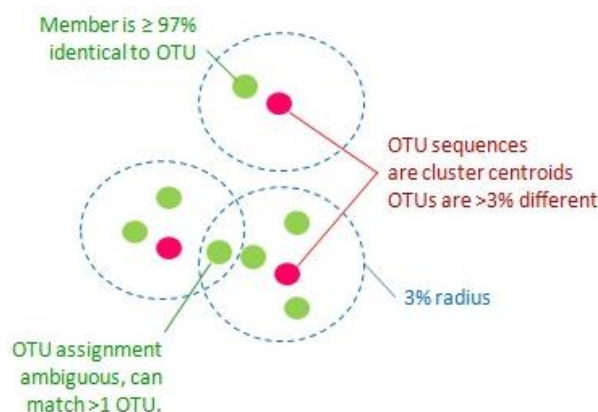
Ook de 18S-ITS kit is in de praktijk getest. In de studie van Callahan et al. (2020) is voor een zestal schimmels uit de ATCC collectie (*Saccharomyces cerevisiae*, *Aspergillus oryzae*, *Candida albicans*, *Trichoderma reesei*, *Kluyveromyces lactis*, *Penicillium chrysogenum*) onderzoek gedaan met het 2.3 kilobasen lange 18S-ITS fragment gegenereerd door de Loop Genomics 18S-ITS kit. Hieruit bleek dat meer dan 80% van de gegenereerde reads overeen kwamen met database-sequences identiek aan deze soorten. Een precieze inschatting hiervan is echter lastig te maken, doordat schimmelcellen vaak een groot aantal kopieën (soms meer dan 100) van het 18S rRNA gen bezitten.

Over de 16S/18S kit is in de praktijk minder bekend. Uit data van Loop Genomics blijkt dat de 16S/18S kit vooral geschikt is voor het meenemen van archaeale diversiteit in de analyses. Met deze methode wordt verder een globaal beeld van de schimmeldiversiteit en de algemene eukaryote diversiteit gegeven (persoonlijk gesprek met Tuval Ben Yehzekel, CEO van Loop Genomics). Anders dan bij de 18S-ITS kit die zich voornamelijk op schimmeldiversiteit richt, zou deze kit dus een meerwaarde kunnen hebben voor de analyse van de totale eukaryote diversiteit, inclusief protozoa.

2.4 Ontwikkelingen in software en databases

2.4.1 Software-ontwikkelingen op het gebied van Operational Taxonomic Units (OTUs) en Amplicon Sequence Variants (ASVs)

16S rRNA gen-gebaseerde amplicon-analyses zijn klassiek gebaseerd op de identificatie van **Operational Taxonomic Units (OTUs)**. Een OTU definieert een groep van nauw verwante 16S rRNA gen-sequenties die in principe sequenties zijn die tot één soort behoren. Om OTUs te bepalen worden in de eerste stap alle sequenties geclusterd om een dataset met representatieve sequenties te vormen (Figuur 9). Bij benadering is elke gevormde OTU-sequentie door de gekozen clustering gelijk aan een 16S rRNA gen-sequentie van een microbiële soort. Aan de hand van een 16S rRNA gen-referentiedatabase, kunnen de OTUs vervolgens worden geclassificeerd. Een voordeel van OTU-gebaseerde methoden is dat sequencing-fouten door de clustering en het genereren van een consensus sequentie grotendeels worden weggewerkt. Een nadeel is dat door deze clustering ook een groot deel aan mogelijk relevante microbiële diversiteit onzichtbaar blijft. Een voorbeeld van een OTU-gebaseerde methode die in verschillende eerdere onderzoeken is toegepast, wordt o.a. gebruikt door de MOTHUR software.



Figuur 9 Grafische weergave voor de vorming van OTUs.

In dit voorbeeld van het UPARSE-OTU algoritme worden 16S rRNA gen-sequenties geclusterd op een identiteit van $\geq 97\%$. In de praktijk worden verschillende clusterwaarden gebruikt, waarbij 97% de laagst gebruikte waarde is. De twee cirkels die aangegeven worden met de rode lijnen vormen twee aparte OTUs. De sequentie aangegeven met de groene lijn vormt een sequentie die twee mogelijke OTUs laat overlappen. Dit geeft een belangrijke uitdaging aan voor de definitie van OTUs.

Recentelijk zijn software tools ontwikkeld die gebruik maken van zogeheten **amplicon sequence variants (ASVs)**. Bij de ASV-methode wordt eerst voor elke unieke sequentie bepaald hoe vaak deze voorkomt. Vervolgens wordt hierop een foutcorrectie toegepast om sequencing-fouten uit de data te halen. Elke ASV wordt vervolgens geclassificeerd aan de hand van een referentiedatabase. Het argument om ASV-gebaseerde methoden te gebruiken is dat dit een grotere sequence-resolutie oplevert en dat de reproduceerbaarheid en vergelijkbaarheid tussen verschillende studies wordt verbeterd. Een voorbeeld van een ASV-gebaseerde methode die in verschillende eerdere onderzoeken is toegepast is Deblur en DADA2 en worden gebruikt door QIIME 2.

Uit vergelijkingsstudies blijkt dat ASV-gebaseerde methoden een hogere resolutie en specificiteit hebben (Prodan et al. 2020). Onderzoeken aan systemen met hoge microbiële diversiteit als drinkwater, oppervlaktewater, afvalwater en sediment kunnen daarom het beste met een ASV-gebaseerde methode worden uitgevoerd. Daarom wordt bij KWR momenteel Deblur in QIIME 2 als standaard analysemethode toegepast. Zie voor een achtergrond op QIIME 2 voor dataverwerking de 'Nieuwsbrief: kansen van nieuwe molecuair-microbiologische methoden in de drinkwatersector', juni 2019, paragraaf 2.3.1.

2.4.2 Software-ontwikkelingen voor soortidentificatie met korte DNA-fragmenten

Een hogere soortidentificatie met korte 16S rRNA gen-fragmenten wordt gegeven door een in 2017 ontwikkelde Bayesian taxonomic classification tool (Gao et al. 2017). Deze tool, voornamelijk ontwikkeld voor studies aan het darm-microbioom, werd getest op een aantal gesimuleerde datasets. Voor korte V4-regio sequences bleek soortidentificatie mogelijk voor 68% van de data. Hogere percentages werden alleen gehaald met grotere 16S rRNA gen-fragmenten ($>95\%$ voor V1-V3 sequences). De identificatie in deze voorbeeldstudies is grotendeels gebaseerd op bekende soorten die goed in referentiedatabases zijn gedocumenteerd. De verwachting is daarom dat deze tool minder goed presteert op bijvoorbeeld drinkwatermonsters, al kan ook hierbij tot 70% soortidentificatie worden gehaald (niet-gepubliceerde data bij KWR).

Voor QIIME 2 is de q2-feature-classifier plugin ontwikkeld en gepubliceerd om taxonomische classificatie uit te voeren (Bokulich et al. 2018). Deze plugin gebruikt **machine-learning** en **alignment**-gebaseerde methoden voor taxonomische classificatie. De methode is daarmee volgens de gebruikers nauwkeuriger dan de andere beschikbare methoden (de 'naive-Bayes', BLAST+-gebaseerde en de VSEARCH-gebaseerde classificatoren). Echter, ook bij deze methode geldt dat voor identificatie tot op soortniveau de accuraatheid maximaal circa 70% is. Hierbij moet wel worden vermeld dat uitgebreider getrainde databases (met meer en accuratere referentiedata) naar verwachting betere resultaten kunnen opleveren.

Voor toepassing in de drinkwatersector is de methode aanvullend veelbelovend, omdat gangbare referentiedatabases relatief weinig drinkwater-gerelateerde sequenties bevatten. Een specifiek op drinkwater getrainde database kan dit probleem verkleinen.

2.4.3 Software: kansen voor soortidentificatie met QIIME 2

Uit bovenstaande bevindingen volgt dat op software- en analysegebied zeker ontwikkelingen zijn gemaakt die identificatie tot op soortniveau met korte DNA-fragmenten verbeteren. Er moet wel worden vastgesteld dat de grote limitatie wordt veroorzaakt door de relatief korte reads zelf. Een belangrijk discussiepunt blijft hierbij of soortidentificatie met korte 16S rRNA gen-reads überhaupt mogelijk en wenselijk is.

Het staat vast dat de grootste ontwikkelingen voor soortidentificatie met korte reads te verwachten zijn met zelflerende databases. Deze zelflerende databases worden momenteel onder andere toegepast in de QIIME 2 pipeline, zoals hierboven is vermeld. Momenteel is bekend dat de zelflerende classificatiemethode goede resultaten geeft voor zogeheten 'mock communities' waarvan de soortsamenvatting van tevoren bekend is. Het gaat hier tevens om soorten waarover veel achtergrondkennis aanwezig is.

Dit is anders voor drinkwatermonsters waarin veel nog onbekende soorten aanwezig zijn. De vraag is of voor deze monsters tevens accurate soortidentificatie kan worden gerealiseerd door de software. Deze vraag biedt een grote uitdaging. Voor het ontwikkelen van een gedegen database is het namelijk noodzakelijk om grote hoeveelheden sequencing-data aan te bieden en te verwerken. Een grotere hoeveelheid input-data geeft namelijk een verder ontwikkelde database die meer zekerheid aan de output geeft. Een huidige beperking binnen het drinkwateronderzoek is dat een goed getrainde database op dit moment nog niet beschikbaar is.

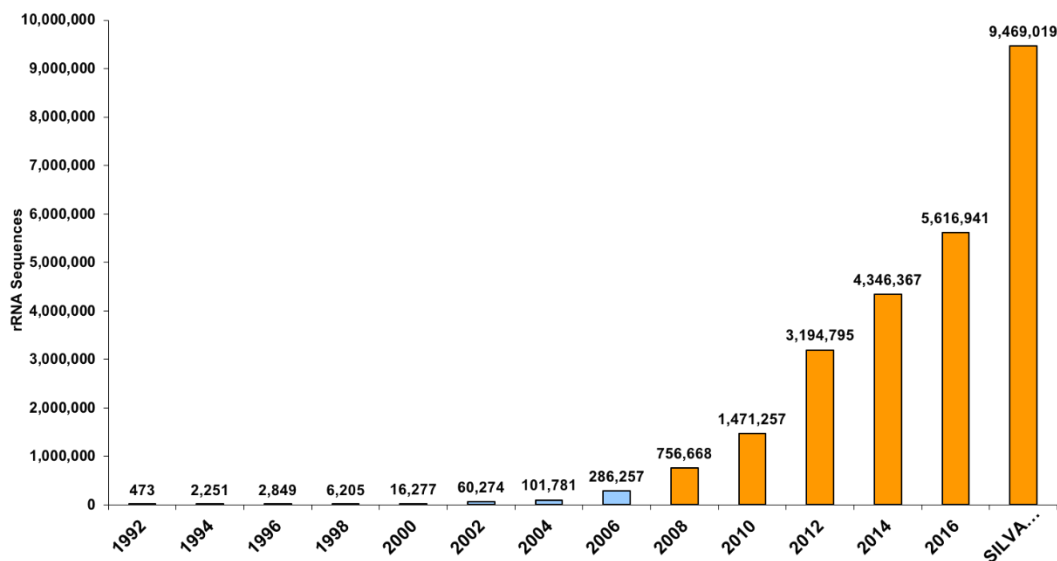
De opzet van deze studie biedt echter een goede mogelijkheid om de potentie van de QIIME 2 software te bepalen. In het voorstel wordt daarom een vergelijking tussen de QIIME-gebaseerde methode met korte reads en de Loop Genomics methode met lange reads van het 16S rRNA gen op dezelfde monsters meegenomen.

2.4.4 Ontwikkelingen in bestaande databases

Referentiedatabases vormen de basis voor microbiële identificatie. Voor verreweg de meeste 16S- en 18S-gebaseerde studies wordt de **SILVA** small subunit (**SSU**: dit beslaat 16S en 18S rRNA gen-sequenties) database gebruikt. Binnen de SILVA databases bestaat de SSU Ref NR 99 database waarin zeer gelijkende sequenties met meer dan 99% identiteit worden opgeschoond en sequenties tevens worden gefilterd op lengte en kwaliteit. In de nieuwste Ref NR 99 database zitten ruim een half miljoen representatieve sequenties, afkomstig uit de complete dataset met 9,4 miljoen sequenties.

De snelle toename in het aantal beschikbare sequenties is te zien in Figuur 10. Dit vormt een belangrijke basis voor een steeds betere en nauwkeurigere taxonomische identificatie van microbiële sequencing-data. Zeker in de afgelopen jaren is ook de hoeveelheid data uit natuurlijke monsters toegenomen. Belangrijk om te realiseren is dat ook voor accurate full-length 16S en/of 18S soortidentificatie de database van cruciaal belang is en blijft. Wanneer een sequentie namelijk niet in de database aanwezig is, kan deze niet worden geclassificeerd.

Een initiatief dat momenteel specifiek loopt voor eukaryote identificatie is het **UniEuk** project. Het doel van dit project is om een taxonomisch framework te ontwikkelen voor eukaryoten. Het ontwikkelen van een dergelijke kwalificatietool is zeer complex, omdat hierbij ook rekening moet worden gehouden met de klassieke classificatie van eukaryoten (denk aan planten, bomen, zoogdieren etc.). Verder is de moleculaire data van bepaalde eukaryote groepen, zoals schimmels, nog beperkt.



Figuur 10 Toename in beschikbare rRNA sequenties in databases.

Blauw: de RDPII database, oranje: alle SILVA small subunit (SSU) sequences na kwaliteitsfiltering. Bron: arb-silva.de.

Een grote kans voor verbeteringen in de databases ligt in de ontwikkeling van specifieke milieu-gebaseerde databases, zoals bijvoorbeeld het geval is bij de 'q2-feature-classifier' plugin van QIIME 2. Deze combinatie vormt in de praktijk een veelbelovende tool voor betere microbiële identificatie, mogelijk zelfs voor short-read data tot op soortniveau.

2.5 Primerkeuze en primerbias

2.5.1 Primerkeuze voor het 16S rRNA gen

De primerkeuze vormt de basis van een 16S rRNA gen-gebaseerd onderzoek. De keuze van een primerset heeft een groot effect op het uiteindelijke resultaat van de analyse. De primers bepalen namelijk in hoge mate in hoeverre de microbiële diversiteit in een DNA-monster kan worden opgepikt.

Binnen het 16S rRNA gen-onderzoek wordt vaak gebruik gemaakt van zogeheten '**universele primersets**' die een grote diversiteit aan micro-organismen kunnen oppikken. Alhoewel de naam suggereert dat deze primers de gehele biodiversiteit van een systeem kunnen detecteren, is het goed om te realiseren dat dit in de praktijk eigenlijk geen juiste aanname is. Universele primers dekken in het algemeen een brede microbiologische diversiteit, maar zullen altijd een bias hebben. Dit houdt in dat een universele primerset bepaalde microbiële groepen zeer goed kan detecteren en andere microbiële groepen zeer beperkt tot niet oppikt. Voor specifiekere onderzoeken naar bepaalde microbiële groepen kunnen zogeheten '**specifieke primersets**' worden gebruikt. Deze primersets zijn echter veelal geoptimaliseerd voor korte amplicon sequences van <500 basenparen. Voor analyses van het gehele 16S rRNA-gen is de keuze beperkt tot de begin- en eindregio van het gen (Figuur 1).

Het klassieke primerpaar voor lange 16S rRNA gen-reads is 27F-1492R, een primerset die initieel ontwikkeld is voor analyse van archaeale diversiteit (DeLong 1992). Deze primerset beslaat 1465 basenparen van het 16S rRNA gen en daarmee ook alle negen variabele regionen. Uit evaluatie van de primerset blijkt deze een relatief brede microbiële diversiteit dekt in vergelijking met verschillende andere ontwikkelde long-read 16S rRNA gen-primersets (Hong et al. 2009; Johnson et al. 2019). Ook voor de analyse van bacteriële diversiteit blijkt de primerset zeer geschikt te zijn. Een modificatie van deze primerset wordt ook als standaard gebruikt voor de 16S rRNA-gen kit van Loop Genomics (zie 2.4.2).

Een alternatieve primerset is in een brede *in silico* primerstudie op basis van softwarematige sequentie-analyses geïdentificeerd (Klindworth et al. 2012). Als referentiedataset is in deze studie gebruik gemaakt van de **non-redundant** SILVA Reference database versie 108 (SSURef 108 NR). Voor 16S rRNA gen long-read sequencing wordt

vanuit deze vergelijkingsstudie de primerset S-D-Bact-0008-a-S-16/S-D-Bact-1492-a-A-16, ook wel bekend als GM3/GM4 aanbevolen. Deze primer dekt net als 27F-1492R de variabele regionen V1 t/m V9 en heeft volgens de *in silico* analyses een specificiteit voor Archaea van 0.2%, voor Bacteria van 77.1% en voor Eukaryota van 0.0%. De primerset is daarmee het meest geschikt voor bacteriële diversiteits-analyses. Uit eerdere prestatie-analyses blijkt de primerset verder 52 van de 59 geteste fyla te dekken (88%). De prestatie van deze primerset is echter nooit experimenteel vergeleken met de 27F-1492R-set, waardoor het niet mogelijk is hier direct uitspraken voor de praktijk over te doen.

Een derde ontwikkelde primerset is de 63F-M1387R. Uit een vergelijkingsstudie met de primerset 27F-1492R blijkt de 63F-M1387R primer relatief beter Alphaproteobacteria en Gammaproteobacteria op te pakken, daar waar de 27F-1492R relatief meer Betaproteobacteria en Firmicutes detecteert (Fredriksson, Hermansson, and Wilén 2013). Voor de detectie, ofwel de aan-/afwezigheid van micro-organismen was er niet direct verschil tussen de primers waarneembaar.

Door de brede toepassing en goede prestaties voor bacteriële diversiteits-analyses kan hieruit geconcludeerd worden dat primerset 27F-1492R geschikt is voor soortidentificatie-studies. Het is desondanks goed om de limitaties van primer-gebaseerde studies in het achterhoofd te houden. Een primer pikt namelijk bepaalde groepen micro-organismen zeer goed op, terwijl andere groepen nauwelijks gedetecteerd worden. Dit wordt de primerbias genoemd. Hieruit volgt ook dat vanuit een puur primer-gebaseerde sequencing-studie weinig te zeggen is over de daadwerkelijke abundantie van bepaalde micro-organismen in een systeem. Daarvoor zijn aanvullende onderzoeken met bijvoorbeeld **kwantitatieve PCR**-technieken (qPCR) of celtellingen nodig.

2.5.2 Gedegenereerde 27F-1492R primers voor het 16S rRNA gen

In verschillende studies zijn variaties op voornamelijk de 27F-primer ontwikkeld om een bredere diversiteit aan bacteriën te kunnen detecteren (). Wanneer deze ontwikkelde primervariaties worden gecombineerd, heet dit een **gedegenereerde ('degenerate') primer**. Gedegenereerde primers bestaan dus uit een mix van primers met verschillende basenparencombinaties die in één enkele PCR-reactie kunnen worden toegepast.

In Tabel 3 is te zien dat bijvoorbeeld voor de gedegenereerde primer 27F-CM in het bijzonder ook enterobacteriën goed kunnen worden gedetecteerd. Dit is bijvoorbeeld relevant binnen het onderzoek naar de aanwezigheid van (opportunistische) pathogenen, waarbij de detectie enterobacteriën van belang is. Interessant om hier te vermelden is dat door Loop Genomics de primer 27F-CM met dus één gedegenereerde positie gebruikt (27F-CM: 5'-AGAGTTTGATC**MT**GGCTCAG-3'; 1492R: 5'-TACCTTGTTACGACTT-3').

Op eenzelfde manier kan naar de detectiegevoeligheid van andere sleutelgroepen en –soorten gekeken worden. Neem bijvoorbeeld *Legionella*-soorten. *Legionella* behoort tot de **klasse** Gammaproteobacteria en het **fylum** Proteobacteria. De gedegenereerde primer 27F-YM zou hiervoor dus een betere detectie kunnen opleveren. Hier is echter momenteel geen validatiestudie van beschikbaar.

Tabel 2. Overzicht van de gedegenereerde 27F primers.

Primer binding site sequence ^a	Phylogenetic group(s) containing the binding site sequence	Exactly matching primer(s) ^b
AGAGTTTGATCCTGGCTCAG	Most <i>Bacteria</i>	1, 2, 3, 4, 5, 6
.....A.....	Many <i>Bacteria</i> , especially enteric bacteria	2, 3, 4, 5, 6
.....T.....	<i>Actinobacteria</i> , some <i>Proteobacteria</i>	3, 4, 6
...A.....T...T...	<i>Chlamydiales</i>	6
.....C.....	<i>Atopobium</i> and chloroplasts	
.....T..	<i>Borrelia</i> spp.	6
.....TA.....	<i>Campylobacteriales</i> and <i>Sphingomonadales</i>	3, 4, 6
..G...C...T.....	<i>Bifidobacteriales</i>	6
..G.....	<i>Thermotogales</i> and <i>Planctomycetales</i>	5

Selectie van data afkomstig uit Frank et al. (2008). De primernamen (indien aanwezig) en -sequenties zijn als volgt, met vetgedrukt gedegenereerde basen: 1) 27F-CC (5'-AGAGTTTGATCCTGGCTCAG-3'); 2) 27F-CM (5'-AGAGTTTGATCMTGGCTCAG-3'); 3) 27F-YM (5'-AGAGTTTGATYMTGGCTCAG-3'); 4) 5'-AGAGTTTGATITGGCTCAG-3'; 5) 5'-AGRGTTTGATCMTGGCTCAG-3'; en 6) 27F-YM+3 (5'-AGAGTTTGATYMTGGCTCAG-3' + 5'-AGAAITTTGATCTTGGTTACAG-3' + 5'-AGAGTTTGATCCTGGCTTAG-3' + 5'-AGGGTTTCGATTCTGGCTCAG-3').

2.5.3 Primerkeuze voor het 18S rRNA gen

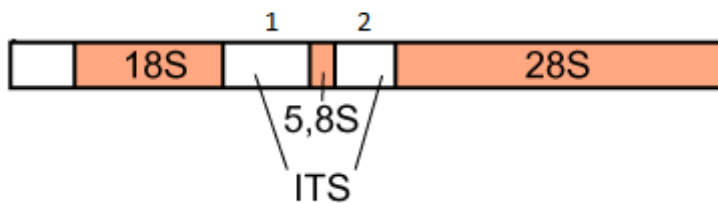
Voor de amplificatie van het 18S rRNA gen gelden dezelfde uitdagingen als voor het 16S rRNA gen (Hugerth et al. 2014). Voor 18S rRNA gen-analyses wordt voor short reads in de praktijk gebruik gemaakt van primers die de V1- V2-, V3-, V4- of V9-regio beslaan (Choi and Park 2020). De analyse van de V4-regio (270-387 basenparen) en de V9-regio (96-134 basenparen) zijn het meest gebruikt. De V4 regio geeft een globaal beeld van de **fylogenetische relatie** tussen eukaryoten; de V9 regio geeft een beter beeld van de diversiteit. Echter, door beperkte databases en de beperkte resolutie van de korte fragmenten is in de praktijk de identificatie van **claden** mogelijk. Dit zijn grotere groepen van organismen met een niet gespecificeerde taxonomische indeling.

In de studie van Hugerth et al. (2014) is een eerste grote poging gedaan om een systematische analyse van primersets te doen om een goede primerkeuze voor verschillende ecosystemen te kunnen maken. In het algemeen geldt voor eukaryote 18S rRNA gen-specifieke primers dat deze vooral geschikt zijn voor bepaalde **taxonomische groepen**. Zo zijn er specifieke primers ontwikkeld die geoptimaliseerd zijn voor schimmels en daardoor andere eukaryote groepen, zoals protozoa, niet of nauwelijks detecteren. In de praktijk geldt daarom dat het noodzakelijk is om voor algemene eukaryote diversiteits-analyses een mengsel van diverse primers te gebruiken. Dit is ook de benadering die voor de Loop Genomics 16S/18S kit is gebruikt.

In de 16S/18S kit worden voor het 16S rRNA gen dezelfde primers gebruikt als in de 16S kit. Aanvullend zijn een drietal extra primers toegevoegd (5'- GATTAAGCCATGCAAGTS-3', 5'-GGTTGATYCTGCCRGRG-3' en 5'-TCYGGTTGATCCTGCC-3') en een tweede reverse primer (5'- GMWACCTTGTTACGACTT -3') om ook archaeale 16S rRNA genen en eukaryote 18S rRNA genen te kunnen amplificeren. Vanuit de literatuur is niet direct bekend in hoeverre de archaeale en eukaryote diversiteit door deze primers wordt gedekt, al is vanuit eerdere communicatie met Loop Genomics bekend dat deze kit als doel heeft een globaal beeld van de diversiteit te geven. Een inschatting van de accuraatheid en de dekkingsgraad van de gebruikte primers kan *in silico* gegeven worden met TestPrime (<https://www.arb-silva.de/search/testprime>). Ondanks het feit dat de kit een globaal beeld geeft, is het wel mogelijk om individuele soorten te identificeren. Het is echter belangrijk om voor deze methode te valideren in hoeverre voldoende identificatie tot op soortniveau mogelijk is.

2.5.4 De internal transcribed spacer (ITS) als alternatief voor het 18S rRNA gen

Wanneer de vraag zich richt op de eukaryote samenstelling en specifiek de schimmelsamenstelling van een systeem, is de 18S-ITS Mycobiome kit beter geschikt. Voor schimmels is bekend dat de internal transcribed spacer (ITS) regio die in deze kit wordt meegenomen een veel gedetailleerder beeld van de diversiteit geeft. In eukaryoten bevinden zich een tweetal ITS sequences die voor diversiteits-analyses worden gebruikt. De ITS1 is gelokaliseerd tussen het 18S en het 5.8S rRNA gen, de ITS2 is gelokaliseerd tussen het 5.8S en het 28S rRNA gen (Figuur 11).



Figuur 11 Overzicht van de internal transcribed spacer (ITS) regio's.

De figuur geeft de locatie van de ITS1 en ITS2 tussen het 18S, 5.8S en het 28S rRNA gen zoals het geval is in de Opisthokonta (de dieren, schimmels en een aantal protisten).

In de 18S-ITS kit wordt gebruik gemaakt van specifieke primers die zijn geoptimaliseerd voor de analyse van de schimmeldiversiteit. Hier wordt in deze bureaustudie echter niet dieper op ingegaan, aangezien hier niet de focus ligt.

2.6 Concrete kansen voor soortidentificatie met 16S en 18S rRNA gen sequencing

NGS soortidentificatie is waardevol als brede screeningsmethode om te achterhalen of er bij werkzaamheden aan het leidingnet of een ongewenst event (zoals lekkage) veranderingen optreden in de microbiologie van het gedistribueerde water. Ook kan een brede screening op soortniveau mogelijk inzicht geven in hoeverre en onder welke omstandigheden (opportunistische) ziekteverwekkers in het Nederlandse drinkwater voor kunnen komen en een mogelijk risico vormen voor de volksgezondheid. Met die informatie kan, indien blijkt dat er sprake is van een onacceptabel risico, gericht onderzoek worden gedaan naar het vóórkomen van dergelijke organismen in het drinkwater. Het is daarbij waardevol wanneer de brede screening snel en zonder benodigde voorkennis kan worden uitgevoerd, zodat uitspraken over de eventuele aanwezigheid van (opportunistische) ziekteverwekkers en indicatororganismen.

In deze studie zijn de mogelijkheden voor soortidentificatie op basis van het 16S en 18S rRNA gen verkend. Hierbij is gekeken naar beschikbare sequencing-platforms, ontwikkelingen in de sequencing-chemie, primerkeuzes en ontwikkelingen in software en databases. Hieruit komen de volgende uitkomsten en aanbevelingen:

1. **A)** Identificatie tot op soortniveau op basis van het 16S en/of 18S rRNA gen → Loop Genomics long-read technologie.
Voordelen: eenvoudig toe te passen in het lab, initiële data-analyse kan worden uitgevoerd met een bijgeleverd protocol en nauwkeurige soortidentificatie is in een aantal studies bewezen.
Nadelen: aanschaf van relatief dure kit, gebruik van specifieke Illumina sequencing kit (MiSeq v3 Paired End 2x300 bp) noodzakelijk, validatie voornamelijk voor 18S rRNA gen-sequencing kits nodig.
B) Indien activiteit 2 niet kan worden uitgevoerd: vergelijking van de Loop Genomics 16S identificatie met de short read-gebaseerde 16S rRNA gen-analyse en soortidentificatie met de in QIIME 2 opgenomen 'q2-feature-classifier' plugin.

Met deze vergelijking is een validatie van de plugin op korte 16S rRNA gen-fragmenten mogelijk. Op basis hiervan kan een uitspraak worden gedaan over de meerwaarde van deze methode.

- Identificatie tot op soortniveau van het 16S rRNA gen met de Oxford Nanopore Q20+ chemie en de R10.3 flow cell op het MinION platform.

Voordelen: goedkoop platform met zeer lange reads met een sterk verbeterde sequencing-foutenmarge

Nadelen: de sequencing-foutenmarge is niet in de praktijk gevalideerd en momenteel is de methode nog sterk in ontwikkeling. Het is momenteel nog niet duidelijk of KWR de mogelijkheid krijgt om op korte termijn (in 2021) de Oxford Nanopore Q20+ chemie toe te passen in het lab.

2.6.1 Validatie van Loop Genomics 16S rRNA gen-kit en de nieuwe Oxford Nanopore chemie

Het voorstel is om in eerste instantie de 16S rRNA gen-kit van Loop Genomics en de nieuwe Q20+ chemie van Oxford Nanopore op 16S rRNA gen-sequences in het lab van KWR te testen. Hiermee kan inzicht worden verkregen in het gemak en eventuele uitdagingen van de toepassing van deze nieuwe methoden. Er wordt in eerste instantie voor gekozen om de methoden alleen op het 16S rRNA gen te testen, omdat hier tevens binnen KWR referentiedata voor aanwezig is.

Voor Loop Genomics wordt voor het valideren van de 16S rRNA gen-kit gekozen, omdat hiervoor diverse validatiestudies zijn uitgevoerd. Door de monsterkeuze vanuit eerder uitgevoerde moleculaire onderzoeken, kan de nieuwe Loop Genomics methode vergeleken worden met de huidige standaard 16S rRNA gen-gebaseerde methode (Tabel 3). Op basis van dit vergelijk kan de mogelijke meerwaarde van de Loop Genomics kit in kaart worden gebracht.

Ook voor het testen van de nieuwe Q20+ chemie van Oxford Nanopore wordt ervoor gekozen om sequences van het gehele 16S rRNA gen te karakteriseren. Dit biedt in de praktijk een goed vergelijk met de methode van Loop Genomics en met de eerder bij KWR toegepaste 16S rRNA gen-gebaseerde methoden.

Aanvullend bestaat de vraag of zelflerende databases in staat zijn om met data van korte fragmenten van het 16S rRNA gen, de methode die nu standaard wordt uitgevoerd, soortidentificatie mogelijk wordt. Goede resultaten zijn bereikt met zogeheten 'mock communities'. Validatie voor milieugebaseerde studies, bijvoorbeeld aan drinkwatermonsters, ontbreken echter nog. In deze studie is daarom het voorstel om voor de monsters waarvan data van korte 16S rRNA gen-fragmenten aanwezig is, deze methode uit te voeren en te vergelijken met de resultaten met de Loop Genomics kit (Tabel 3).

Vervolgens kan op basis van de ervaringen met de nieuwe methoden worden bepaald of deze door KWR kunnen worden uitgevoerd. Indien dit niet het geval is, kan er voor worden gekozen om de uitvoering uit te besteden aan Loop Genomics (in het geval van de Loop Genomics-kit) of om nieuwe ontwikkelingen verder af te wachten (in het geval van Oxford Nanopore). Ook kan op basis van deze ervaringen besloten worden om in de nabije toekomst ook de methoden voor 18S rRNA gen-sequencing en ITS-sequencing te onderzoeken. Aangezien de methoden hiervoor zeer vergelijkbaar zijn, kan een dergelijk onderzoek zich vooral op de potentie binnen het drinkwateronderzoek richten.

Tabel 3 Overzicht monsters voor analyse met de Loop Genomics methode en de Oxford Nanopore Q20+ chemie en, indien haalbaar het vergelijk met short-read data.

Monsterlocatie	Monsterpunten	Monster-nummer	Project-nummer
Dunea, duininfiltratie, GOBAM-project 7 monsters Monsters van zuiveringstraject drinkwater inclusief drinkwater	Snelfilter effluent	LMB-71696-PROCESW	402848-001
	GOBAM effluent	LMB-71695-PROCESW	402848-001
	Infiltratiewater Berkheide	LMB-71888-PROCESW	402848-001
	Korte verblijftijdsplas (plas 37)	LMB-71891-OW	402848-001

	Lange verblijftijdsplas (plas 25-3)	LMB-71889-OW	402848-001
	Verzameld onttrokken water	LMB-71887-PROCESW	402848-001
	Drinkwater	LMB-71886-DW	402848-001
De Lange Vlieter - WML - Onderzoek OMV's 4 monsters Analyse van bodem- en slibmonsters	Bodemmonster locatie 1	LMB-92608-SLIB	402045-132
	Slibmonster #1	LMB-73750-SLIB-A	402045-132
	Ondiep filter (filter 1) WP19	LMB-73746-PROCESW	402045-132
	Diep filter (filter 2) WP19	LMB-73747-PROCESW	402045-132
Drinkwater, biofilm en sediment uit BTO 2020.04 De Laak, Oasen 9 monsters Één drinkwatermonster, één biofilmmonster en één sedimentmonster. Met een triplo van elk monster is de spreiding in de data voor de verschillende methoden in kaart te brengen.	Drinkwater (triplo 1/3)	LMB_60964_DW	402045-038
	Drinkwater (triplo 2/3)		
	Drinkwater (triplo 3/3)		
	Biofilm (triplo 1/3)	LMB_60975_BIOF	402045-038
	Biofilm (triplo 2/3)		
	Biofilm (triplo 3/3)	LMB_60971_DW	402045-038
	Sediment/spuimonster (triplo 1/3)		
	Sediment/spuimonster (triplo 2/3)		
	Sediment/spuimonster (triplo 3/3)		
Mock Community 3 monsters Brengt prestaties van de methode(n) in kaart en dient als standaard	Zymo Mock Community (triplo 1/3)	Aanwezig bij KWR	nvt
	Zymo Mock Community (triplo 2/3)		
	Zymo Mock Community (triplo 3/3)		
Blanco 1 monster Negatieve controle	Blanco	Eigen blanco	nvt

In totaal kunnen met de Loop Genomics kit maximaal 24 monsters worden geanalyseerd. Er is in de huidige opzet rekening gehouden met een aantal triplo monsters. Deze monsters zullen elk in drievoud worden gesequenced om zo de variatie in de methode te kunnen vergelijken. Verder wordt de Zymo Mock Community met een van tevoren bekende microbiële soortsaanblanding meegenomen, om de kracht van de Loop Genomics kit in kaart te kunnen brengen. In totaal wordt met dit experiment de volledige capaciteit van de kit benut.

Voor de MinION sequencing wordt dezelfde opzet gebruikt. De vraag is of het voor de MinION sequencing haalbaar is om alle 24 monsters te sequencen. Indien dit niet het geval is zal er een relevante subselectie worden gemaakt waar ten minste drinkwater, biofilm, sediment en de mock community in zit.

2.6.2 Aanvullende mogelijkheden voor soortidentificatie met 16S rRNA gen-sequencing platforms

Twee aanbevelingen voor identificatie tot op soortniveau die uit dit onderzoek naar voren komen, zijn:

- Identificatie tot op soortniveau op basis van het 16S en/of 18S rRNA gen is mogelijk met het PacBio CCS.
 - o Voordelen: zeer nauwkeurige sequencing en bewezen technologie.
 - o Nadelen: dient te worden uitbesteed, is relatief duur en minder geschikt voor hoge throughput. De methode is zeer geschikt voor de nauwkeurige analyse van een beperkt aantal monsters.
 - o Voor deze methode wordt echter geen directe toepassing in brede screening voorzien.
- Gezien de snelle ontwikkelingen in het veld is het goed om de ontwikkelingen van de verschillende sequencing-methoden nauwlettend in de gaten te houden en in beeld te brengen. Zo kunnen nieuwe, kansrijke moleculaire methoden op tijd worden geïdentificeerd en worden onderzocht om de meerwaarde in het drinkwateronderzoek vast te stellen.

3 Conclusies en aanbevelingen

3.1 Conclusies

- Er zijn verschillende technieken beschikbaar voor NGS-sequencing. De standaard technologie voor korte reads is Illumina, waarvan de Illumina MiSeq bij KWR wordt toegepast. Voor lange reads vormen PacBio en Oxford Nanopore de belangrijke technologieën.
- De ontwikkelingen in Oxford Nanopore MinION sequencing kunnen ertoe leiden dat deze methode binnen enkele jaren standaard wordt toegepast voor soortidentificatie. Aanvullend zijn de kosten voor de hardware laag (ca. €1.000) en zijn de kosten voor de chemie vergelijkbaar met de Illumina MiSeq. Een validatievoorstel met de nieuwe chemie, hardware en software wordt in deze studie gegeven (3.2).
- De Loop Genomics 16S rRNA long read kit biedt een veelbelovend alternatief voor NGS-sequencing en soortidentificatie met het reeds toegepaste Illumina MiSeq platform. Een validatiestudie voor de Loop Genomics 16S rRNA gen-kit wordt in deze studie gegeven (3.2).
- Software-ontwikkelingen dragen bij aan verbeterde soortidentificatie voor NGS-analyses. Voor QIIME 2 is de q2-feature-classifier plugin een belangrijke tool die tot 70% van korte 16S rRNA gen-sequenties op soortniveau kan identificeren.
- Uitgebreidere databases met meer en accuratere referentiedata voor drinkwater-milieus spelen een belangrijke rol in verbeterde soortidentificatie voor NGS-analyses.
- De primerkeuze vormt de basis van een 16S rRNA gen-gebaseerd onderzoek. De keuze van primers bepaalt in hoge mate hoe goed de microbiële diversiteit in een DNA-monster kan worden opgepikt.

3.2 Aanbevelingen

Uit deze bureaustudie volgt de aanbeveling om de methode van Loop Genomics voor de volledige sequencing van het 16S rRNA gen uit te voeren op een 24-tal monsters op het Illumina MiSeq platform. Daarnaast is het voorstel om de nieuwste chemie, hardware en software van Oxford Nanopore voor de MinION te testen om de potentie van deze methode voor toekomstig onderzoek in kaart te brengen. Wanneer het niet haalbaar is om op korte termijn (in 2021) de Oxford Nanopore Q20+ chemie te testen, is het voorstel om de Loop Genomics methode te vergelijken met de prestaties van de q2-classifier-plugin in QIIME 2.

4 Introduction: 16S rRNA gene sequencing for species-level identification

Developments in next generation sequencing (NGS) enable the rapid and high-throughput analysis of large amounts of DNA sequences. As a consequence, microbial diversity analyses have become relatively cheap and mainstream methods in microbiological studies. NGS methods on taxonomic genes, such as the 16S rRNA gene (present in all prokaryotic organisms, i.e. bacteria and archaea) are widely applied in microbial diversity studies (Figure 1).

Current 16S rRNA gene-based techniques enable microbial profiling in which entire microbial communities of tens to hundreds of samples can be sequenced within a single NGS run. Illumina sequencing is the standard sequencing method nowadays in such profiling studies. Due to its sequencing chemistry, Illumina sequencers can sequence DNA molecules with a maximum length of approximately 500 base pairs. In practice, sequences of 200 to 300 base pairs are often used for high-throughput 16S rRNA gene-based studies, due to opportunity to generate overlapping sequences resulting in a higher sequence-quality and accuracy of these lengths. These relative short 16S rRNA gene sequences, however, have a limited taxonomic resolution and hamper microbial identification up to species or even genus level. Therefore, taxonomic identification is generally limited up to genus or family level for most sequenced fragments. In genera that harbor both pathogenic and non-pathogenic microbial species, such as *Legionella*, identification up to species level is necessary to obtain information on the presence and absence of pathogenic species (e.g. pathogenic *Legionella pneumophila* versus other less/non-pathogenic *Legionella* species). In addition, to link the microbial identity to potential metabolic functions, species-level identification can link specific species to the databases of well-studied micro-organisms.

An earlier literature study (described in chapters 1 to 3) highlighted the potential of the Loop Genomics Long Read 16S rRNA kit, in which near-complete 16S rRNA gene sequences can be generated using a sequencing protocol for Illumina sequencing platforms. Full-length 16S rRNA gene sequences do have the resolution for species and in some instances even strain level identification. This method therewith provides a promising tool for microbiological studies of drinking water systems. In addition, this method can be implemented in the KWR lab with the current (Illumina Miseq) infrastructure. Next to the possibility of obtaining near full-length 16S rRNA gene sequences with Loop Genomics, there have also been developments in bioinformatics that would make it better possible to use short-length 16S rRNA gene sequences for species identification. Both approaches are described in the following paragraphs.

4.1 Full-length 16S rRNA gene sequencing for species-level identification

With the Loop Genomics protocol the 16S rRNA gene is amplified with the 27F and 1492R primers that cover the near-complete 16S rRNA gene (Figure 1). Next, the PCR products each receive a unique barcode before the fragments are enzymatically cut into smaller fragments that receive the same unique barcode and that can be sequenced with Illumina chemistry (Figure 2). Due to the unique barcode that comes from the full-length fragment, these fragments can be assembled back into the full-length 16S rRNA gene sequence (Figure 2). After sequencing, the data is uploaded to the Loop Genomics online server where full-length 16S rRNA gene sequences are reconstructed (Figure 2).

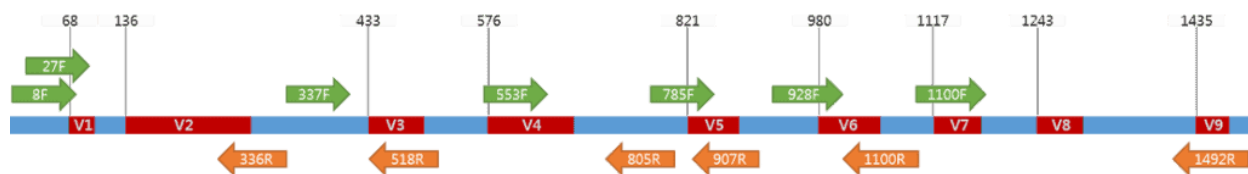


Figure 1. Schematic overview of the 16S rRNA gene.

The overview displays the nine variable regions labeled in red (V1 to V9) and the nucleotide position of the start of each variable region (according to the *E. coli* reference genome). A number of frequently used primers and their positions on the 16S rRNA gene are indicated. Forward primers are labeled in green, reverse primers are labeled in orange.

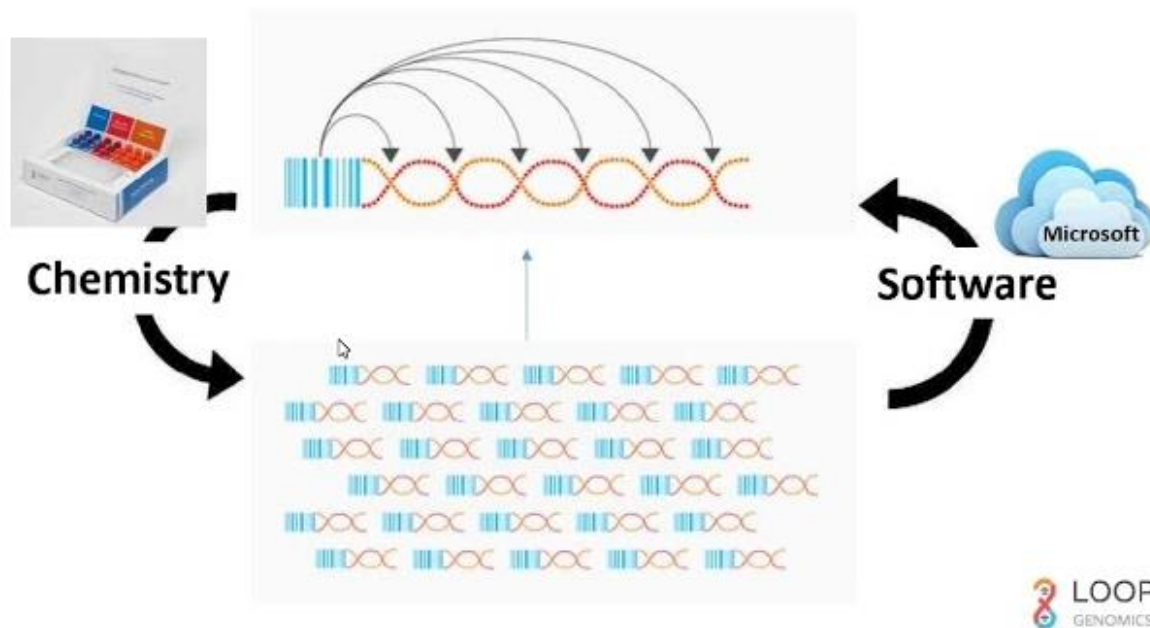


Figure 2. Overview of the Loop Genomics method to obtain full-length 16S rRNA gene sequences. In the first step, the kit chemistry is used to obtain short fragments of full-length 16S rRNA gene PCR products. All short fragments from individual full-length 16S fragments contain a unique barcode sequence that is used to later on assemble the fragments into full-length sequences, called contigs. The short fragments are generated to enable sequencing on the Illumina sequencing platforms. After sequencing an online cloud service provided by Loop Genomics is used to assemble the short, barcoded fragments back into the full-length 16S rRNA gene sequences.

4.2 QIIME 2™ species classifier plugin on short 16S rRNA gene v3 reads

There are also developments in species-level identification methods on short 16S rRNA gene sequences. Within QIIME 2™, a bioinformatics platform to process 16S rRNA gene sequencing data, a species classifier and a pretrained database have been developed that claim to perform species-level identification based on standard short-read data generated from, for example the V3 or V3-V4 region of the 16S rRNA gene. This classifier uses machine learning algorithms for training, a classifier trained on microbial communities from non-saline water and sediment samples appears to be a good candidate. As mentioned earlier, the short-read sequencing methods are still widely used. Therefore, such bioinformatics tools can provide significant advantages to current studies. A major consideration is whether the short fragments provide enough taxonomic resolution for proper species analysis. This species classifier will therefore be tested and compared with the full-length 16S rRNA gene sequencing method to obtain insights in the species identification potential of this tool.

4.3 Testing approach

The use of the different approaches to enable species level identification (Loop Genomics and Qiime2 species classifier) were tested on a mixture of predefined bacterial species (mock community) and on a selection of water samples derived from drinking water systems obtained from various research projects. Results obtained with the standard partial 16SrRNA gene v4 region NGS sequencing (using the standard classifier and the alternative classifier) were compared with Loop Genomics to be able to judge the added value of these methods.

5 Materials and methods

5.1 Selection of samples

The application and added value of the Loop Genomics approach to generate full-length 16S rRNA gene sequences were studied by applying this method on two types of samples and comparing the results obtained with short fragment (v4) 16S rRNA gene sequencing:

- A mock community containing a simple mixture of known bacterial species.
- A selection of relevant samples obtained from drinking water treatment and distribution.

The origin of the used samples is summarized in Table 1. For all samples mentioned short-length (v4) 16S rRNA gene sequencing according to the default protocol at KWR have been performed previously. A short description of each sample location is given below (Table 1). The origin of the samples is described in multiple reports. The samples for Oasen, De Laak have been previously published (Learbuch & Timmers, 2020), as well as the samples from WML “De Lange Vlieter” (Van der Grift et al., 2020) and the samples from the Dunea “GOBAM” project (Timmers et al., 2021).

Table 1. Overview of the samples used for this study. The sample name indicates the name used in this study. The sample type indicates the origin of the sample: water, sediment or biofilm. The sediment samples in this study originate from drinking water (Oasen, De Laak) and from lake sediment (WML, de Lange Vlieter). The lab codes are specific for this study. The project number refers to the previous projects in which short 16S rRNA gene fragment sequencing has been performed.

Sample name	Sample type	Lab code	Project Number
ZymoBIOMICS Microbial Community DNA*	Standard	ZYMO	
Oasen drinking water production plant “De Laak”			
Drinking water*	water	LMB-60964	402045-038
Sediment in distribution after flushing*	Water with sediment	LMB-60971	402045-038
Biofilm in distribution*	biofilm	LMB-60975	402045-038
WML drinking water production plant “De Lange Vlieter”			
OMPs shallow filter 1 WP19	water	LMB-73746	402045-132
OMPs deep filter 2 WP19	water	LMB-73747	402045-132
OMPs soil sample location 1	soil	LMB-92608	402045-132
Dunea “GOBAM” project			
Rapid sand filter effluent	water	LMB-71696	402848-001
GOBAM AOP effluent	water	LMB-71695	402848-001
GOBAM infiltration water Berkheide	water	LMB-71888	402848-001
GOBAM short residence basin 37	water	LMB-71891	402848-001
GOBAM long residence basin 25-9	water	LMB-71889	402848-001
GOBAM abstracted water after dune passage	water	LMB-71887	402848-001
GOBAM drinking water Katwijk	water	LMB-71886	402848-001

**For validation of the Loop Genomics 16S Long Read method triplicate samples were run for Oasen, de Laak water, sediment and biofilm samples as well as for the ZymoBiomics Microbial Community DNA Standard. For the short-read 16S rRNA gene-based study the original datasets with only one run were used.*

The ZymoBIOMICS Microbial Community DNA Standard

The ZymoBIOMICS Microbial Community DNA Standard was used to validate the Loop Genomics kit and analysis pipeline. This mock community is useful to validate species-level identification with the kit and the provided software tools. The ZymoBIOMICS standard contains a total of eight different bacterial species and two different fungal species (Table 2). Since the general 16S rRNA gene primers, used by both Loop Genomics and the standard short 16S rRNA gene fragment approach, only target the prokaryotic 16S rRNA gene, no fungal sequences (containing the 18S instead of 16S rRNA gene) are amplified.

Table 2. The composition of the ZymoBIOMICS Microbial Community DNA Standard with eight bacterial species of which for each species 12% genomic DNA (gDNA) is added to the standard and two fungal species, for which 2% gDNA is added to the standard (not displayed here). The taxonomic classification is displayed from Phylum to Species level. Gram Stain indicates the negative (-) or positive (+) Gram Stain of the species. GC presents the average GC-percentage of the genome.

Phylum	Class	Order	Family	Genus/species	% GC
Proteobacteria	Gammaproteobacteria	Pseudomonadales	<i>Pseudomonadaceae</i>	<i>Pseudomonas aeruginosa</i>	66.2
Proteobacteria	Gammaproteobacteria	Enterobacterales	<i>Enterobacteriaceae</i>	<i>Escherichia coli</i>	56.8
Proteobacteria	Gammaproteobacteria	Enterobacterales	<i>Enterobacteriaceae</i>	<i>Salmonella enterica</i>	52.2
Firmicutes	Bacilli	Lactobacillales	<i>Lactobacillaceae</i>	<i>Lactobacillus fermentum</i>	52.8
Firmicutes	Bacilli	Lactobacillales	<i>Enterococcaceae</i>	<i>Enterococcus faecalis</i>	37.5
Firmicutes	Bacilli	Bacillales	<i>Staphylococcaceae</i>	<i>Staphylococcus aureus</i>	32.7
Firmicutes	Bacilli	Bacillales	<i>Listeriaceae</i>	<i>Listeria monocytogenes</i>	38.0
Firmicutes	Bacilli	Bacillales	<i>Bacillaceae</i>	<i>Bacillus subtilis</i>	43.8
Ascomycota	Saccharomycetes	Saccharomycetales	<i>Saccharomycetaceae</i>	<i>Saccharomyces cerevisiae</i>	38.4
Basidiomycota	Tremellomycetes	Tremellales	<i>Tremellaceae</i>	<i>Cryptococcus neoformans</i>	48.2

Oasen – De Laak

At drinking water production location “De Laak” (Oasen, Lexmond) a previous project studied the influence of drinking water installations on the microbiological drinking water quality (Learbuch & Timmers, 2020). Samples were taken from drinking water, sediment and biofilm in the drinking water distribution system. Here, we analyzed a single drinking water, sediment and biofilm sample in triplicate to test the reproducibility of the Loop Genomics full-length 16S rRNA gene sequencing method on relevant samples and to investigate the potential of the method to perform species-level identification on samples from a drinking water distribution system.

WML – De Lange Vlieter

At drinking water production location “de Lange Vlieter” (WML, Beegden) a previous study investigated the presence and removal of micropollutants in shallow and deep filters (Van der Grift et al., 2020). The samples also include a soil/sediment sample to test whether the Loop Genomics full-length 16S rRNA gene sequencing method is powerful enough to unravel highly complex soil/sediment communities up to species level.

Dunea – GOBAM

The samples from the Dunea GOBAM project are taken at pretreatment location Bergambacht and at subsequent dune infiltration at Berkheide, with the corresponding post-treatment and resulting drinking water. At Bergambacht, water is pretreated from river water. Due to the occurrence of organic micropollutants in the source water, an advanced oxidation set-up was installed to treat the surface water in the pre-purification process (Timmers et al., 2021) The samples taken from Dunea GOBAM represent relevant points in the drinking water treatment train and were therefore selected to investigate whether species-level identification is possible with

long-read 16S rRNA gene sequences. Samples that were included in this study were, in order of appearance in the treatment train:

1. Dunea GOBAM rapid sand filter effluent
2. Dunea GOBAM AOP effluent
3. Dunea GOBAM infiltration water Berkheide
4. Dunea GOBAM short residence basin 37
5. Dunea GOBAM long residence basin 25-9
6. Dunea GOBAM collected abstracted water
7. Dunea GOBAM drinking water

5.2 Loop Genomics 16S rRNA gene Long Read Kit sample preparation & sequencing

Samples were prepared according to the manufacturer's instructions with the following modifications. For the 16S rRNA gene enrichment PCR all samples were diluted to an input concentration of 1,0 ng/μl DNA and amplified with 20 PCR cycles following the PCR protocol of the Loop Genomics method using 16S rRNA gene specific primers 27F/1492R. After PCR amplification, all samples were diluted ten times in elution buffer and analysed with the Qubit dsDNA HS Assay kit (Thermo Fisher, Waltham, MA, USA). Next, the samples were further diluted in EB buffer to target 0,2 ng/μl DNA (Loop Genomics protocol: <0,3 ng/μl DNA). Next, the barcode PCR was performed and PCR products were cleaned up with the SPRIselect Cleanup provided in the Loop Genomics kit. Next, the SYBR barcode calibration was performed. Barcode PCR samples were diluted 40-fold and, subsequently, the barcode calibration protocol was run. All barcode concentrations were initially too high. All samples were therefore diluted 50-fold to achieve a 2000-fold dilution for each sample. All but one sample, with a low DNA template input that did not show any amplification product in the first amplification round, passed the quality criteria to proceed. The barcode amplification was performed on the diluted samples, after which the samples were pooled, cleaned-up with the SPRIselect Cleanup kit. Next, barcode distribution, barcode activation and neutralization and a post-activation SPRIselect cleanup were performed according to the manufacturer's instructions. The pooled library was size checked on a Bioanalyzer with the Agilent High Sensitivity DNA Kit (Agilent, Santa Clara, CA, USA). The library was quantified using the KAPA Library Quantification Kit (Roche, Basel, Switzerland) for undiluted sample up to 100.000 times diluted sample. Based on the results of the 10.000 times and 20.000 times dilutions the library concentration was 2.8 nM. Sequencing was performed on an Illumina MiSeq platform with a 600-cycle MiSeq Reagent Kit v3 (Illumina, San Diego, CA, USA). The selected Library Prep Kit was the "TruSeq HT" with 1 index read and Read Type "Paired End" with 301 cycles for both read 1 and read 2. Denaturation and library preparation were performed according to the manufacturer's instruction. For the PhiX DNA standard a total of 10% was added based on DNA quantities (molarity).

5.3 Sequence analysis on the Loop Genomics server

The raw fastq.gz (generated by the Illumina MiSeq DNA-sequencer) files were uploaded to the Loop Genomics upload server available on the website under "Upload Data". The loop indices were indicated based on the amplification protocol as specified in the manual. The "silva-132" database was selected as default database (software version for April 2022) and the pipeline run was created. The Loop Genomics server is a data analysis pipeline that is used for low-quality base trimming, unique sample barcode demultiplexing, and synthetic long-read reconstruction. Demultiplexing and synthetic long-read reconstruction is a process that enables *de novo* assembly of full-length 16S rRNA gene long-read data after rearranging the short-leads tagged with the same unique barcode. After processing, which took about 2 hours, all results were downloaded from the server. The taxonomy output files were used for further analysis. From this taxonomy file the top 25 species of each sample were exported for further data processing. The output ".fasta" and ".fastq" files were used for obtaining full-length 16S rRNA gene reads for further investigation.

5.4 Short fragment 16S rRNA gene sequence analysis

Short-read partial 16S rRNA gene sequencing was performed in previous studies that are shown in Table 1. The complete procedure used to generate sequences is described previously (Heijnen 2021). In short, 16S rRNA gene fragments (~300 base pairs) targeting the V4 variable region were amplified using the primer set 515F and 806R (Figure 1). Cleaned-up PCR products were provided with sample specific dual-index sequences using a second PCR step and equimolar mixtures of pooled samples were subsequently sequenced on the Illumina MiSeq sequencer at KWR. The indexed sequences were demultiplexed using the software of the Illumina sequencer and sample specific DNA sequences were analyzed with QIIME 2™ (Boyle et al., 2019). This program improves sequencing data quality by first assembling paired end sequences and then performs several quality filtering steps in which errors in the sequencing data are removed (deblur, (Amir et al. 2017)). Next, chimera sequences (artefacts that originate during PCR amplification) are removed. After these steps, the number of amplicon sequence variants (ASVs) is determined. In this step, all unique sequences are clustered as one specific ASV. In the next step, the ASVs are compared to 16S rRNA gene sequences of identified bacteria to identify the ASVs, this process is called phylotyping.

The taxonomic level to which the ASVs can be classified is dependent on the quality and length of the sequence and the similarity between the ASVs and the sequences in the reference database. To study the influence of the classifier on the ability to classify sequences to a lower taxonomic level (genus- or species-level), two different taxonomic classifiers were compared to identify short 16S rRNA gene sequences in this project. Both classifiers use the taxonomic identification of sequences between primers 515F and 806R of the 16S rRNA gene of the Silva taxonomic database. One classifier uses version 138 of the Silva taxonomic database (silva-138-99-515-806-nb-classifier.qza) and the other classifier is pretrained, using machine learning algorithms, with sequences from non-saline water samples (silva_132_515f_806R_ref_water_non_saline_classifier.qza). The ability of both classifiers to classify the sequences is tested by using both classifiers on the same dataset. It should be noted that both these classifiers are still using the “traditional” phylum nomenclature instead of the recently proposed new phylum names such as e.g. “Pseudomonadota” instead of “Proteobacteria” (Panda, Islam Salim, and Sharma 2022). The general sequence statistics of the raw and processed short sequences are summarized in Table 3.

Table 3. General sequence statistics of the samples analysed by short fragment v4 16S rRNA sequencing.

Sample name	Sample nr.	Raw paired-end Reads	High quality Reads (Feature counts after paired-end assembly and Deblur)
ZymoBIOMICS Microbial Community	Zymo-1	166528	94377
ZymoBIOMICS Microbial Community	Zymo-2	151021	85107
Oasen De Laak drinking water	60964	86785	40377
Oasen De Laak sediment flushing	60971	89100	44045
Oasen De Laak biofilm	60975	88532	56359
WML De Lange Vlieter OMPs shallow filter 1 WP19	73746	288648	109097
WML De Lange Vlieter OMPs diep filter 2 WP19	73747	141829	56180
WML De Lange Vlieter OMPs sludge sample 1	73750	252941	94917
WML De Lange Vlieter OMPs soil sample location 1	92608	303418	115624
Dunea GOBAM rapid sandfilter effluent	71696	168175	82486
Dunea GOBAM AOP effluent	71695	169368	93694
Dunea GOBAM infiltration water Berkheide	71888	156838	77691

Dunea GOBAM short residence basin 37	71891	144397	70718
Dunea GOBAM long residence basin 25-9	71889	374768	188536
Dunea GOBAM collected abstracted water	71887	138124	49832
Dunea GOBAM drinking water	71886	135732	51096
Average		178,513	81,884

6 Results & Discussion

6.1 Loop Genomics long-read sequencing technology: general

The long read 16S rRNA gene kit of Loop Genomics provides a method for the reconstruction of full-length 16S rRNA gene sequences based on assemblies of overlapping short 16S rRNA gene fragments generated by Illumina sequencing. The method uses a random approach to fragment full-length PCR products created by the 27F and 1492R primer set (Figure 1). These short sequences can be sequenced on an Illumina sequencing platform and are subsequently assembled to full-length 16S rRNA gene sequences by an online assembly tool, generating both fasta files (only containing nucleotide sequences) and fastq files (containing nucleotide sequences and sequencing quality scores). The challenges of this method include sequencing depth (number of sequences per sample) and sequence identification. In the experiments described here, the Illumina MiSeq was used for the sequencing procedure. This platform provides a relatively low sequencing depth, compared to the HiSeq and NextSeq platforms. Therefore, a maximum of 24 samples can be sequenced to obtain sufficient sequencing depth, although challenges can occur with the detection of rare sequences in highly complex samples that might require deeper sequencing in cases where these rare sequences are relevant. Overall, the MiSeq Reagent Kit v3 that is used for sequencing can produce 44-50 million paired-end reads. From the Loop Genomics run, over 29 million demultiplexed reads were collected, which equals 1.3 million reads per sample, although the number of reads per sample showed large variations (0.4-2.5 million reads per sample). Even the lowest number of reads per sample (0.4 million) should provide sufficient resolution for analysis. After contig reconstruction, the average sample contained 2,158 contigs, but a large range was observed with lower numbers of contigs for samples with lower read counts (762-6,166 contigs per sample). The number of contigs per sample indicates that the method is likely less suitable to identify low-abundant (<0.1% abundance) species in a sample. Furthermore, the identification of the full-length 16S rRNA gene sequences to species level is highly dependent on the reference database. Previous studies have indicated large fractions of uncultured and, therefore, undocumented prokaryotes in surface water, soil/sediment and drinking water samples that have not yet been described as a species (Learbuch & Timmers, 2020; Timmers, 2021 and unpublished data from studies discussed in this report). This is expected to be the most significant challenge with species-level identification of environmental samples.

6.2 Sequence identification with the Loop Genomics 16S Long Read kit and provided software

6.2.1 General sequence statistics

The Loop Genomics processing pipeline documents the progress of the data throughout the analysis process. Documented data includes the demultiplexed reads, which are the total numbers of (short) reads per sample, the number of reconstructed full-length 16S rRNA gene-contigs, the number of unmapped clustered 16S rRNA gene contigs and the number of 16S rRNA gene contigs that could be successfully mapped against the SILVA v132 SSU rRNA gene database. A summarization of this data is provided in Table 4.

Table 4. General sequence statistics of the samples analysed by the Loop Genomics software tool.

Sample name	Demultiplexed Read Count	Full-length 16S contigs	Mapped contigs
ZymoBIOMICS Microbial Community DNA Standard triplicate 1/3	1,678,600	1,353	1,283 (94.8%)
ZymoBIOMICS Microbial Community DNA Standard triplicate 2/3	1,658,128	1,113	1,041 (93.5%)
ZymoBIOMICS Microbial Community DNA Standard triplicate 3/3	1,479,848	2,650	2,538 (95.8%)
Oasen De Laak drinking water triplicate 1/3	2,488,635	1,431	750 (52.4%)
Oasen De Laak drinking water triplicate 2/3	2,029,110	2,321	1,265 (54.5%)
Oasen De Laak drinking water triplicate 3/3	1,330,088	2,052	1,118 (54.5%)
Oasen De Laak sediment flushing triplicate 1/3	790,049	1,068	697 (65.3%)
Oasen De Laak sediment flushing triplicate 2/3	1,052,585	1,747	1,167 (66.8%)
Oasen De Laak sediment flushing triplicate 3/3	632,471	958	641 (66.9%)
Oasen De Laak biofilm triplicate 1/3	419,691	675	454 (67.3%)
Oasen De Laak biofilm triplicate 2/3	1,416,761	1,956	1280 (65.4%)
Oasen De Laak biofilm triplicate 3/3	1,080,175	1,675	1,073 (64.1%)
WML De Lange Vlieter OMPs shallow filter 1 WP19	500,370	717	232 (32.4%)
WML De Lange Vlieter OMPs deep filter 2 WP19	2,138,296	4,013	1,366 (34.0%)
WML De Lange Vlieter OMPs soil sample location 1	1,138,004	1,597	1,042 (65.2%)
Dunea GOBAM rapid sand filter effluent	543,925	879	747 (85.0%)
Dunea GOBAM AOP effluent	1,831,858	2,191	1,877 (85.7%)
Dunea GOBAM infiltration water Berkheide	435,773	627	530 (84.5%)
Dunea GOBAM short residence basin 37	3,336,774	5,337	4,542 (85.1%)
Dunea GOBAM long residence basin 25-9	1,224,483	2,034	1,850 (91.0%)
Dunea GOBAM collected abstracted water	882,382	1,095	655 (59.8%)
Dunea GOBAM drinking water	1,074,224	1,633	507 (31.0%)
Average	1,325,556	1,778	1,147 (64.5%)

On average, 1,325,566 short sequence reads (range 419,691-3,336,774) resulted in the reconstruction of 1,778 full-length 16S rRNA gene sequences per sample (range 675-5,337), demonstrating that on average, a redundancy of 745.4 reads is used to reconstruct one full-length 16S rRNA gene sequence. In the data generated by short fragment v4 16S rRNA sequencing (Table 3) less raw sequence data (average 178.512 reads per sample) was needed to provide taxonomic information from 81.884 high quality sequence reads. This means that with the applied sequence depth it will be possible to detect taxons that have at least a relative abundance of 1 in 1778 (0.05%) on average in the bacterial community with the loop genomics approach. However, in the samples with the lowest sequence depth (627 full-length contigs) only taxons with a relative abundance of 0.16% will be detectable. This demonstrates that a large number of reads are needed to detect rare sequences in a community making Loop Genomics a relatively expensive approach for rare taxon detection.

Full-length 16S rRNA gene sequences are classified by comparing the assembled contig sequences with the sequences in the SILVA 16S rRNA gene reference database where mapping is not possible if the sequence similarity is below the threshold. For the mock communities, over 94% of the contigs were above the threshold (4-7% of the contigs was below the threshold for a hit). For the drinking water, biofilm and sediment samples 31 – 91% of the contigs were above the threshold. Contigs that are below the threshold could indicate not yet described prokaryotic species, but there are few means other than general diversity parameters to unravel this. It is

important to stress that drinking water, drinking water sediment and biofilm samples on average show that roughly half of the contigs are below the threshold. Therefore, a significant part of the prokaryotes detected with Loop Genomics in these samples were not identified to the species level.

Next, the results of the ZymoBIOMICS mock community were analyzed in more detail to investigate the performance and reproducibility of the Loop Genomics method on this community with known bacterial species composition.

6.3 Mock community: Loop Genomics and short v4 fragment comparison

The ZymoBIOMICS mock community was used to study the performance of Loop Genomics and compare the results with standard 16S rRNA gene short v4 fragment (the 515F-806R region) sequencing to classify the sequences present in this simple artificial community containing eight different bacterial and two different fungal species.

Loop Genomics

The results from the triplicate Loop Genomics measurements were averaged and the standard deviation between the triplicates indicated low variation in the data between sequencing runs (Figure 3, Table 5). This indicates a good reproducibility of the Loop Genomics method.

Table 5. Overview of the bacterial species observed in the ZymoBIOMICS mock community analysed in tri-plicate using Loop Genomics

Species	ZymoBIOMICS standard 1/3	ZymoBIOMICS standard 2/3	ZymoBIOMICS standard 3/3	Average	Standard deviation
<i>Bacillus subtilis</i>	17.2%	17.3%	18.0%	17.5%	0.4%
<i>Listeria monocytogenes</i>	13.9%	11.9%	11.3%	12.4%	1.3%
<i>Staphylococcus aureus</i>	12.9%	13.6%	11.0%	12.5%	1.3%
<i>Salmonella enterica</i>	11.4%	12.0%	13.5%	12.3%	1.1%
<i>Lactobacillus fermentum</i>	8.7%	8.9%	10.3%	9.3%	0.9%
<i>Enterococcus faecalis</i>	8.3%	7.4%	7.1%	7.6%	0.7%
<i>Escherichia coli</i>	7.7%	6.7%	7.4%	7.3%	0.5%
<i>Pseudomonas aeruginosa</i>	6.4%	7.2%	7.3%	7.0%	0.5%
<i>Escherichia/Shigella</i> sp.	5.1%	6.1%	6.7%	6.0%	0.8%
<i>Staphylococcus</i> sp.	3.7%	2.5%	2.7%	3.0%	0.6%
<i>Lactobacillus</i> sp.	2.1%	2.3%	1.7%	2.0%	0.3%
<i>Pseudomonas</i> sp.	1.4%	1.6%	1.3%	1.5%	0.2%
<i>Bacillus</i> sp.	0.7%	1.6%	1.3%	1.2%	0.5%
<i>Enterococcus</i> sp.	0.4%	0.6%	0.3%	0.4%	0.2%
<i>Shigella sonnei</i>	0.2%	0.0%	0.1%	0.1%	0.1%
<i>Exiguobacterium</i> sp.	0.1%	0.0%	0.0%	0.0%	0.0%

On average 99.8 % of the reads were correctly identified up to genus level, 85.5% was identified correctly up to species level and 0.2% of the reads were not correctly identified. The incorrectly identified reads included *Shigella sonnei* (in two samples, 0.1% on average) and *Exiguobacterium* sp. (in one sample, 0.03% on average). The detection of *Shigella* on the genus level in 16S rRNA gene sequencing studies on sample with *Escherichia coli* is explainable. *Shigella* and *Escherichia* are closely related and classified to the genus level as a joint *Escherichia-*

Shigella genus due to their close sequence identity (Devanga Ragupathi et al. 2017). However, the identification of 0.1 % of sequences to the pathogenic species of *Shigella sonnei* (instead of *Escherichia coli*) can potentially lead to wrong conclusions about the presence of pathogenic species in samples that contain *E. coli*.

Overall, the relative abundances were best approached for *Salmonella enterica* and *Listeria monocytogenes*.

The ZymoBIOMICS community also contains two fungal species with a gDNA abundance of 2% for each species (*Saccharomyces cerevisiae* and *Cryptococcus neoformans*). As expected, no fungal sequences were detected, since these organisms contain the 18S rRNA gene which is not targeted by the 16S rRNA gene PCR primers used.

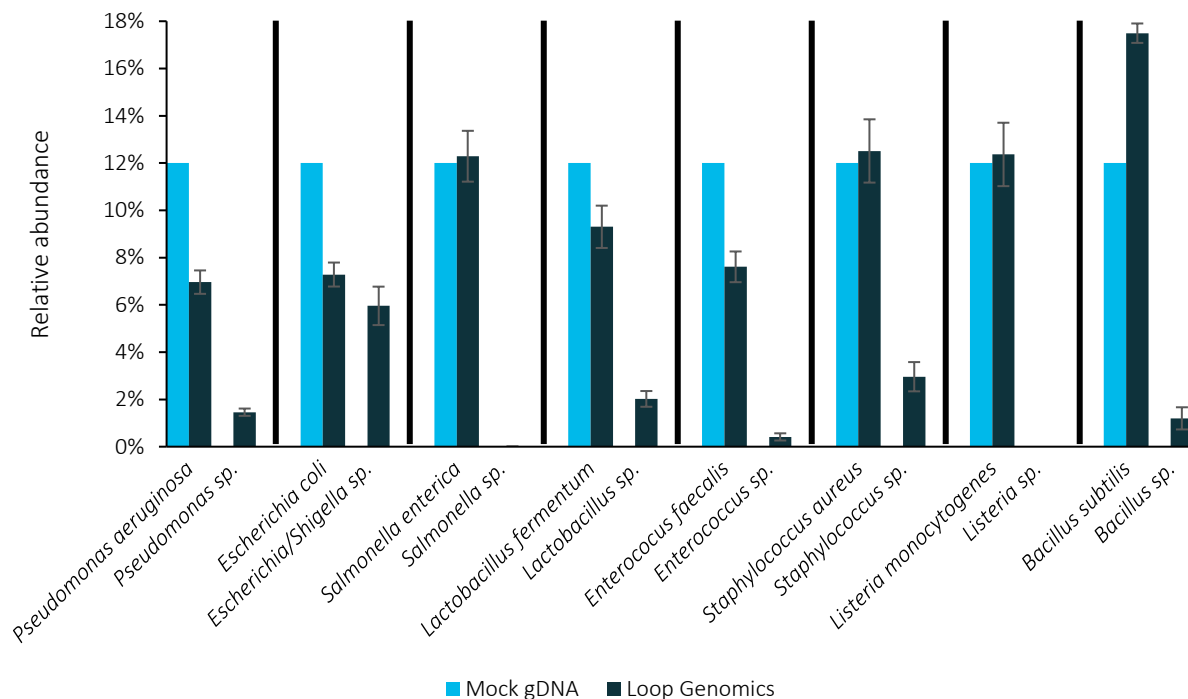


Figure 3. Relative abundances of the eight bacterial species in the ZymoBIOMICS mock community. The Mock gDNA shows the relative abundance of gDNA that was added to the mock community, in this case 12% for each bacterial species. The Loop Genomics results show the average relative abundance of sequences identified to species and genus level for the triplicate samples. Error bars display the standard deviation of the mean.

Next, we investigated biases that could cause differences between gDNA amounts and observed relative abundances in the Loop Genomics datasets. Unravelling such biases is important to increase the accuracy and applicability of the Loop Genomics method. First, we investigated whether the differences between gDNA quantities and observed relative abundances could be caused by species-specific genomic differences. It is known that multiple copies of 16S rRNA genes can be present in one bacterial cell and that these copy numbers can vary between different bacterial species. This is why we investigated whether the 16S rRNA gene copy numbers in the genomes of the eight species influences their relative abundances in the sequencing dataset.

Table 6 shows the mean 16S rRNA gene copy numbers for the eight bacteria species that are present in the ZymoBIOMICS Microbial Community DNA Standard. The 16S rRNA gene copy number data (as reported in the ribosomal database of bacteria and Archaea, *rrnDB* (rrnodb.umms.med.umich.edu)) was used to correct the gDNA amounts in the ZymoBIOMICS Microbial Community DNA Standard (Table 6). This approach accounts for the fact that organisms with a higher 16S rRNA gene copy number in their genomic DNA will result in a higher proportion of sequences in the final sequence library. We observed that the relative abundance of each species in the sequencing data is much closer to the copy number corrected abundance in the gDNA sample of the mock community. As expected, the 16S rRNA gene copy numbers seem to affect the relative abundance outcome of the Loop Genomics full-length 16S rRNA gene sequencing method. The corrected relative abundance for most of the bacterial species was closer to the expected relative abundance of 12.0% than the uncorrected relative abundance. The correction

for 16S rRNA gene copy numbers in the genome can thus provide a better approximation of the abundance of individual species in a sample.

Table 6. 16S rRNA copy numbers of the species present in the ZymoBIOMICS mock community and the copy number corrected relative abundances in this community.

Species	DNA reads uncorrected	Mean rrnDB 16S rRNA gene copy number	gDNA	gDNA corrected
<i>Pseudomonas aeruginosa</i>	7,0%	4.0	12.0%	8.3%
<i>Escherichia coli</i>	7,3%	7.0	12.0%	14.6%
<i>Salmonella enterica</i>	12.5%	7.0	12.0%	14.6%
<i>Lactobacillus fermentum</i> *	9.3%	5.0**	12.0%	10.4%
<i>Enterococcus faecalis</i>	7.6%	4.0	12.0%	8.3%
<i>Staphylococcus aureus</i>	12.5%	5.6	12.0%	11.7%
<i>Listeria monocytogenes</i>	12.4%	5.7	12.0%	11.9%
<i>Bacillus subtilis</i>	17.5%	9.7	12.0%	20.2%

Mean 16S rRNA gene copy numbers for the eight bacterial species in the ZymoBIOMICS Microbial Community DNA Standard as documented in the rrnDB, the gDNA percentage in the ZymoBIOMICS Microbial Community DNA Standard and the 16S rRNA gene copy number corrected gDNA amounts. The corrected gDNA was calculated by multiplying the gDNA amount with the mean rrnDB 16S rRNA gene copy number and dividing this by the total sum of gDNA x 16S copy number values. 16S rRNA gene copy number data was obtained from rrnDB.umms.med.umich.edu on 22-11-2021.

**Limosilactobacillus fermentum* in rrnDB database.

**5,0 selected as value, since only a single data entry reported 1 instead of 5 copy numbers.

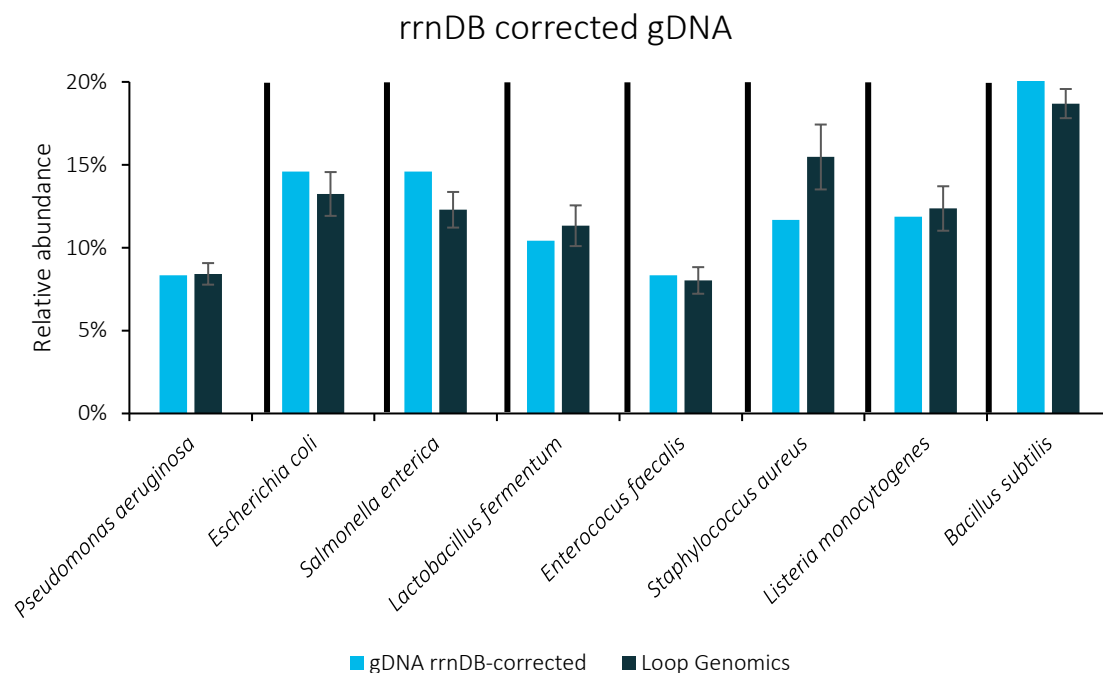


Figure 4. rrnDB-corrected gDNA relative abundances plotted against the total relative abundances of the Loop Genomics method. Loop Genomics read-identification on genus and species level were clustered for better comparability of the data. The Loop Genomics results show the average relative abundance of sequences identified to species and genus level for the triplicate samples. Error bars display the standard deviation of the mean.

The remaining discrepancies between the 16S rRNA gene copy number corrected gDNA quantities and the observed relative abundances of the species in the mock community indicates that other factors in the sequencing procedure also introduce biases.

Overall, the Loop Genomics long-read 16S rRNA gene amplicon sequencing protocol is able to identify all species present in the ZymoBIOMICS mock community with high reproducibility. In addition, the correction for 16S rRNA gene copy numbers provides a useful tool to better approach the true species abundance. It has to be stressed that such copy number corrections can only be performed when the identified species are documented in the rrnDB database. In environmental samples such calculations cannot be made, since the majority of the identified 16S rRNA gene sequences belong to bacterial species for which the rRNA gene copy numbers are unknown. In addition, a DNA extraction bias did not play a role in the mock community sequencing effort but does affect field studies. Such a bias is not addressed by a 16S rRNA gene copy number correction.

6.3.1 Loop Genomics comparison with short 16S rRNA gene v4 fragment sequencing

Duplicate samples of the ZymoBIOMICS mock community were also analysed using a short fragment sequencing approach (the 16S rRNA gene v4 region). The result of classification of the sequences using two classifiers was compared (SILVA v138 classifier and the SILVA v138 classifier pretrained on non-saline water samples (Ben Kaehler, [readytowear](#) at github (Quast et al. 2012; Yilmaz et al. 2013)). The results are summarized in Table 7. All sequences (179484 sequences from two samples) were correctly identified as sequences to the family level and 89.9% were correctly identified to the genus level using both classifiers. Only sequences from *Salmonella enterica* (10.1% of the sequences) were identified only to the family level (*Enterobacteriaceae*), but not to the genus level. Three taxa were identified to the species level using the pretrained classifier on short 16S rRNA gene v4 sequences. However, of these three taxa only *Pseudomonas aeruginosa* was identified to the expected species, whereas *Staphylococcus aureus* and *Enterococcus faecalis* were identified as other, closely related species (*S. warneri* and *E. durans* respectively) by the pretrained classifier (Table 7).

These classifications were also compared with the Loop Genomics classification. The Loop Genomics method correctly identified 85,9% of the sequences to the (correct) species level whereas only 5.8% of the sequences were correctly identified to species level using short 16S rRNA gene v4 sequencing and only with the use of the pretrained SILVA v138 classifier (Table 7).

Table 7. Identification of the ZymoBIOMICS Microbial Community DNA Standard using short fragment (v4) sequencing and two taxonomic classifiers.

Approach (Classifier)	Taxonomic identification			Relative abundance		
	Family	Genus	Species	Mock_1	Mock_2	Average %
Species (gDNA)	Bacillaceae	Bacillus	subtilis			12.0
Short v4 (Silva)	Bacillaceae	Bacillus	—	20.2	19.3	19.8
Short v4 (Pretrained Silva)	Bacillaceae	Bacillus	—	20.2	19.3	19.9
Loop Genomics	Bacillaceae	Bacillus	subtilis			17.5
	Bacillaceae	Bacillus	—			1.2
Species (gDNA)	Lactobacillaceae	Lactobacillus	fermentum			12.0
Short v4 (Silva)	Lactobacillaceae	Lactobacillus	—	15.3	16.7	16.0
Short v4 (Pretrained Silva)	Lactobacillaceae	Lactobacillus	—	15.3	16.7	16.0
Loop Genomics	Lactobacillaceae	Lactobacillus	fermentum			9.3
	Lactobacillaceae	Lactobacillus	—			2.0
Species (gDNA)	Listeriaceae	Listeria	monocytogenes			12.0
Short v4 (Silva)	Listeriaceae	Listeria	—	13.8	12.4	13.1
Short v4 (Pretrained Silva)	Listeriaceae	Listeria	Ambiguous_taxa	13.8	12.4	13.1
Loop Genomics	Listeriaceae	Listeria	monocytogenes			12.4
Species (gDNA)	Enterobacteriaceae	Escherichia	coli			12.0
Short v4 (Silva)	Enterobacteriaceae	Escherichia-Shigella	—	12.8	11.5	12.1
Short v4 (Pretrained Silva)	Enterobacteriaceae	Escherichia-Shigella	—	12.8	11.5	12.1
Loop Genomics	Enterobacteriaceae	Escherichia	coli			7.3
	Enterobacteriaceae	Escherichia-Shigella	—			6.0
Species (gDNA)	Staphylococcaceae	Staphylococcus	aureus			12.0
Short v4 (Silva)	Staphylococcaceae	Staphylococcus	—	12.5	16.5	14.5
Short v4 (Pretrained Silva)	Staphylococcaceae	Staphylococcus	warneri	12.5	16.5	14.5
Loop Genomics	Staphylococcaceae	Staphylococcus	aureus			12.5
	Staphylococcaceae	Staphylococcus	—			3.0
Species (gDNA)	Enterobacteriaceae	Salmonella	Enterica			12.0
Short v4 (Silva)	Enterobacteriaceae	—	—	10.6	9.6	10.1
Short v4 (Pretrained Silva)	Enterobacteriaceae	—	—	10.6	9.6	10.1
Loop Genomics	Enterobacteriaceae	Salmonella	Enterica			12.3
Species (gDNA)	Enterococcaceae	Enterococcus	faecalis			12.0
Short v4 (Silva)	Enterococcaceae	Enterococcus	—	9.0	8.3	8.7
Short v4 (Pretrained Silva)	Enterococcaceae	Enterococcus	Enterococcus durans	9.0	8.3	8.7
Loop Genomics	Enterococcaceae	Enterococcus	faecalis			7.6
	Enterococcaceae	Enterococcus	—			0.4
Species (gDNA)	Pseudomonadaceae	Pseudomonas	aeruginosa			12.0
Short v4 (Silva)	Pseudomonadaceae	Pseudomonas	—	5.9	5.7	5.8

Short v4 (Pretrained Silva)	Pseudomonadaceae	<i>Pseudomonas</i>	<i>aeruginosa</i>	5.9	5.7	5.8
Loop Genomics	Pseudomonadaceae	<i>Pseudomonas</i>	<i>aeruginosa</i>			7.0
	Pseudomonadaceae	<i>Pseudomonas</i>				1.5

Differences in taxonomic identification using the two classifiers are highlighted in grey.

The relative abundance of the identified taxons in the mock community (Figure 5) demonstrates comparable trends in the relative abundancies of the different species between Loop Genomics, short fragment v4 sequencing and the theoretical rrnDB-corrected gDNA. *Lactobacillus* seems to be overrepresented in profiles using short rRNA gene v4 fragment sequencing and *Escherichia* appears to be overrepresented in both short 16S rRNA gene v4 profiles and in loop genomics full-length 16S rRNA gene profiles. Since the same isolated DNA is used for both methods a role for differences in isolation efficiencies can be excluded. It is likely that differences in amplification efficiencies play a role in creating the observed differences in relative abundancies. And, since different primer pairs are used for Loop Genomics and short fragment v4 rRNA gene sequencing, amplification differences between these methods can be expected.

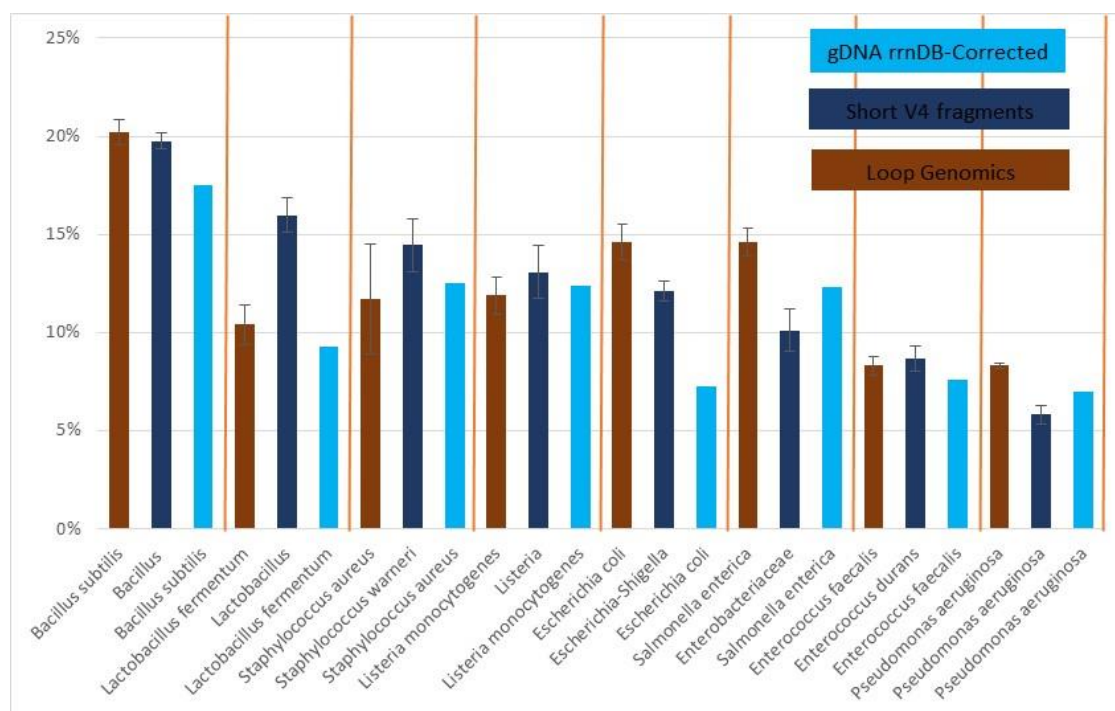


Figure 5. rrnDB-corrected gDNA quantities plotted against the total relative abundances of the Loop Genomics method. Loop Genomics read-identification on genus and species level were clustered for better comparability of the data. The Loop Genomics results show the average relative abundance of sequences identified to species and genus level for the triplicate samples. Error bars display the standard deviation of the mean.

These results demonstrate that:

- Loop Genomics:
 - o Correctly classified 98.8% of the mock community sequences to the genus level.
 - o Correctly classified 85.8% of the mock community sequences to the species level.
 - o Misclassified 0.2% of the mock community sequences (0.2%) to other species than present in the mock community.
- Short (v4) fragment sequencing:
 - o Seven of the eight species present in the mock community are correctly identified to genus level. Only *Salmonella enterica* could not be classified to the genus level, but was correctly classified to the family level (*Enterobacteriaceae*).

- Only one of the eight species in the mock community is identified correctly to species level (*Pseudomonas aeruginosa*), but only when the pretrained Silva classifier was used.
- Comparable trends in fractional abundance of the different bacterial species are observed between Loop Genomics and short fragment (v4) sequencing suggesting limited differences in the introduction of sequence bias.

These results confirm the expectation that full-length 16S rRNA gene sequencing in general enables species level identification of the sequences and demonstrates that only genus level classification can be expected when short (v4) sequencing is used. Although based on limited data, these results also indicate that the classifier used to classify short 16S rRNA gene sequences influences the classification results and that better classification can be expected with the use of a pretrained classifier. This suggests that classification results can be improved by training of the classifiers. Because the apparent better performance of the pretrained classifier, this classifier was used, together with the Loop Genomics method, in the subsequent analysis of sequence data obtained from samples from water treatment processes.

6.4 Analysis of water, sediment and biofilm samples.

Analyses were performed on samples derived from three drinking water treatment plants (Table 1) and the results obtained with the Loop Genomics full-length approach were compared with the results obtained with the standard short fragment v4 16S rRNA gene sequencing approach. A summarization of the most dominant (top 25) taxa in all samples are shown in the tables in Appendix III (“De Laak”), Appendix O (“De Lange Vlieter”) and Appendix V (“GOBAM”). A summary of the species and genus identification is shown and discussed for all analyzed samples in paragraph 6.4.1 and a more detailed description of the results obtained from the samples analyzed from production location “De Laak”, production location “De Lange Vlieter” and “GOBAM” is given in paragraphs 6.4.2 and 6.4.3 respectively.

6.4.1 Species and genus identification in all samples

The results of the classification of the 25 most abundant taxa in the samples were used to compare the ability of Loop Genomics to identify near full-length 16S rRNA gene sequences with short fragment v4 16S rRNA gene (using the pretrained Silva classifier) sequencing to identify the sequences to genus and species level. Taxons are considered as “classified” in cases where the procedure delivers a genus name, a candidate (*Candidatus*) genus name, a species name, a candidate (*Candidatus*) species name. The results of the comparison between Loop Genomics and short fragment v4 16S rRNA gene sequencing are summarized in Table 8.

Table 8. Number of taxons, from the top 25 most abundant taxons, classified to genus or species level

Sample	Species level		Genus level
	% in top 25	nr. of classified taxons	
Location:	De Laak (Appendix III.I)		
Drinking water			
Loop Genomics	67.7%	2	9
short v4 sequencing	50.2%	0	4
Sediment			
Loop Genomics	70.1%	0	9
short v4 sequencing	49.3%	0	1
Biofilm			
Loop Genomics	90.9%	0	8
short v4 sequencing	63.1%	0	2
Location	WML, de Lange Vlieter (Appendix IV)		
Water shallow filter 1 WP19			
Loop Genomics	74.1%	0	15
short v4 sequencing	52.7%	0	4
Water deep filter 2 WP19			
Loop Genomics	55.3%	1	18
short v4 sequencing	60.4%	0	4
Soil sample location 1			
Loop Genomics	55.7%	1	11
short v4 sequencing	30.8%	0	2
Project:		GOBAM project	
GOBAM Rapid sandfilter effluent			
Loop Genomics	88.7%	0	20
short v4 sequencing	72.4%	1	12
GOBAM AOP effluent			
Loop Genomics	91.9%	1	19
short v4 sequencing	83.3%	0	14
GOBAM infiltration Berkheide			
Loop Genomics	87.5%	0	19
short v4 sequencing	69.4%	0	14
GOBAM short residence basin 37			
Loop Genomics	85.9%	0	19
short v4 sequencing	68.2%	1	12
GOBAM long residence basin 25-9			
Loop Genomics	94.2%	2	18
short v4 sequencing	84.6%	1	13
GOBAM collected water after dune passage			
Loop Genomics	74.4%	3	13
short v4 sequencing	59.6%	1	15
Dunea GOBAM drinking water			
Loop Genomics	73.3%	0	14
short v4 sequencing	51.8%	0	4

This comparison shows that there is a large difference in the ability to identify the sequences to genus or species level in all samples. On average 14.8 (range: 8-19) of the 25 most abundant taxa in every sample are identified to the genus level using Loop Genomics whereas 7.8 taxa (range: 2-18) are identified to genus level using short fragment v4 16S rRNA gene sequencing. Only 10 taxa are identified to the species level using Loop Genomics and 4 taxa are identified to species level using short fragment v4 16S rRNA gene sequencing. However, identification to species level appears to be problematic for both approaches since identification to species names from known cultured species appears to be rare. This appears to be mainly the result of the presence of uncultured and unknown bacteria in the dominant fraction of the bacterial communities in these samples.

6.4.2 Samples from production location “De Laak” (Oasen)

To test the performance and reproducibility of the Loop Genomics method in detail, we analyzed triplicate samples from the drinking treatment plant “De Laak”. Samples from drinking water, sediment and biofilm were previously taken from the distribution system and subsequently analyzed with the Loop Genomics full-length 16S rRNA approach and compared with data using short 16S rRNA gene v4 fragment sequencing previously collected on the identical DNA-extract.

Loop Genomics

From the Loop Genomics data the top 25, most dominant, taxonomic groups were selected from each sample and the results are summarized in Figure 6 (drinking water), Figure 7 (sediment) and Figure 8 (biofilm). Overall, the reproducibility between the triplicates was high in all samples as was also observed for the mock community.

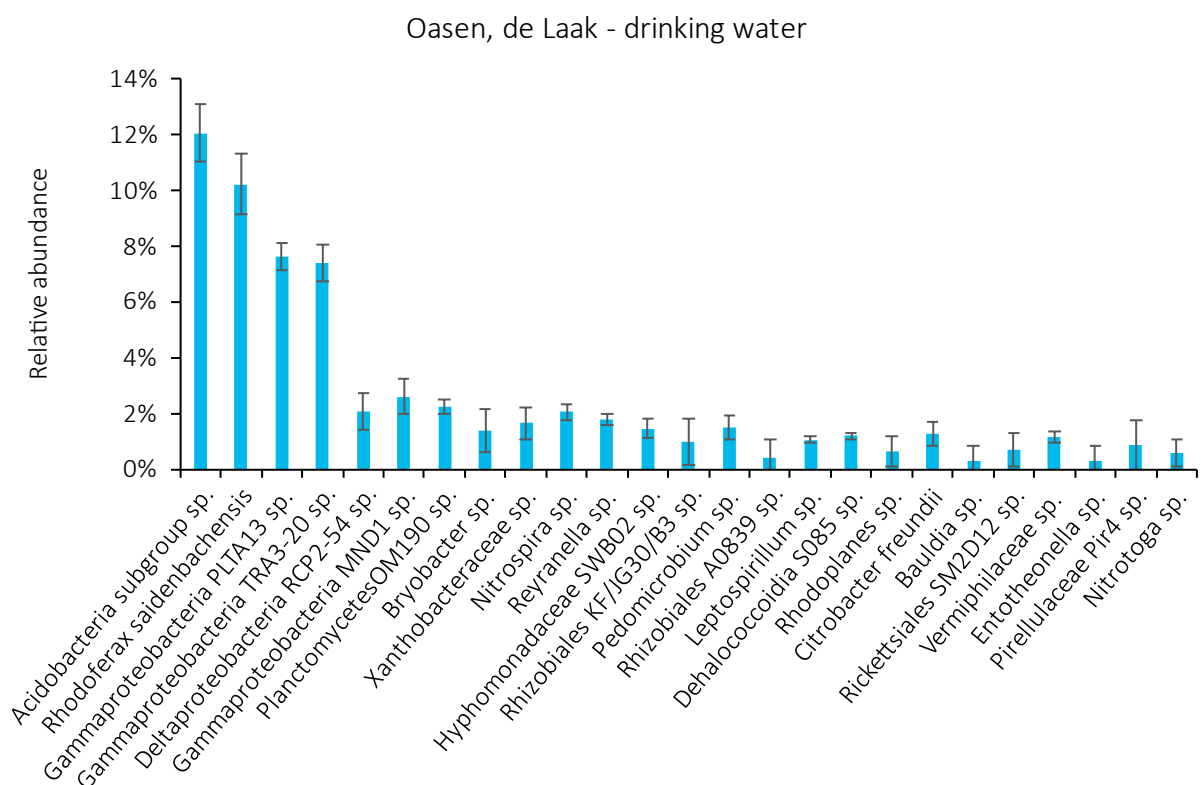


Figure 6. Relative abundance of the top 25 phylogenetic full-length 16S rRNA taxa from the triplicate water samples from Oasen, de Laak. The top 25 included 61-67% of all reads. Clustering is based on the Loop Genomics pipeline. Error bars show standard deviation of the mean.

The results for the drinking water sample indicate a complex microbial community containing four relatively abundant species (each >8% of the community) and 67.7% of the community belonging to the top 25 most abundant taxa (Figure 6). Species-level identification was successfully performed for two out of the 25 most dominant taxonomic groups, *Rhodofarax saidenbachensis* and *Citrobacter freundii*. *R. saidenbachensis* made up 10.2% and *C. freundii* 1.3 % of the reads (together 11.5%) of the dataset. An important point of attention is the

detection of *Citrobacter freundii* in drinking water using this approach. This bacterial species belongs to the bacterial family *Enterobacteriaceae* and is recognized as a coliform. The drinking water produced at “De Laak” is monitored daily for the presence of coliforms and coliform numbers are in general below 1 CFU/100 ml, which makes it unlikely that *C. freundii* is present in such high relative abundance. This indicates that either misclassification or cross contamination during sequencing (also called “cross talking”) might have occurred.

Nine taxonomic groups were identified up to genus level based on the full-length 16S rRNA sequences. These include genera like *Bryobacter* sp., *Nitrospira* sp., *Reyranella* sp., *Pedomicrobium* sp., *Leptosprillum* sp., *Rhodoplanes* sp., *Bauldia* sp., *Entotheonella* sp. and *Nitrotoga* sp., together making up 12.4% of the reads. 39.6% of the reads are mapped to clusters of uncultured bacteria, which indicates the lack of reference data in the Silva database on bacterial species in drinking water. The four most abundant contigs had a relative abundance between 7 and 11 percent and jointly made up 37.4% of the drinking water community from De Laak. Besides *R. saidenbachensis* (10.2%), the other three contigs were identified as *Acidobacteria* subgroup sp. (12.1%), *Gammaproteobacteria* PLTA13 sp. (7.6%) and *Gammaproteobacteria* TRA3-20 sp. (7.4%), which all represent uncultured bacteria mostly obtained from shotgun metagenomics studies in which the whole DNA diversity of a sample is sequenced and documented. Linking the Loop Genomics 16S rRNA gene sequences to the metagenomics data can be useful to obtain information about the functional potential of these not yet identified species.

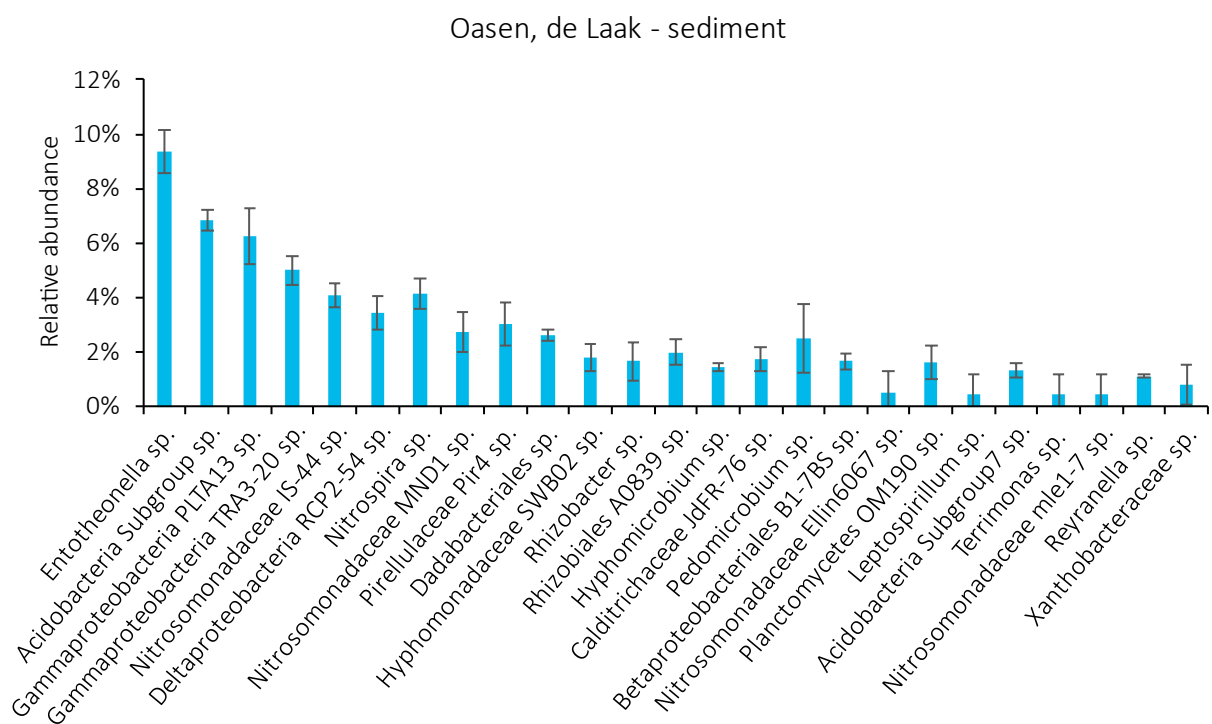


Figure 7. Relative abundance of the top 25 phylogenetic full-length 16S rRNA taxa from the triplicate sediment sample from Oasen, de Laak. The top 25 included 66-71% of all reads. Clustering is based on the Loop Genomics pipeline. Error bars show standard deviation of the mean.

For the sediment samples, no contigs were identified up to species level and nine taxa were identified up to genus level (31.5% of the reads) (Figure 7).

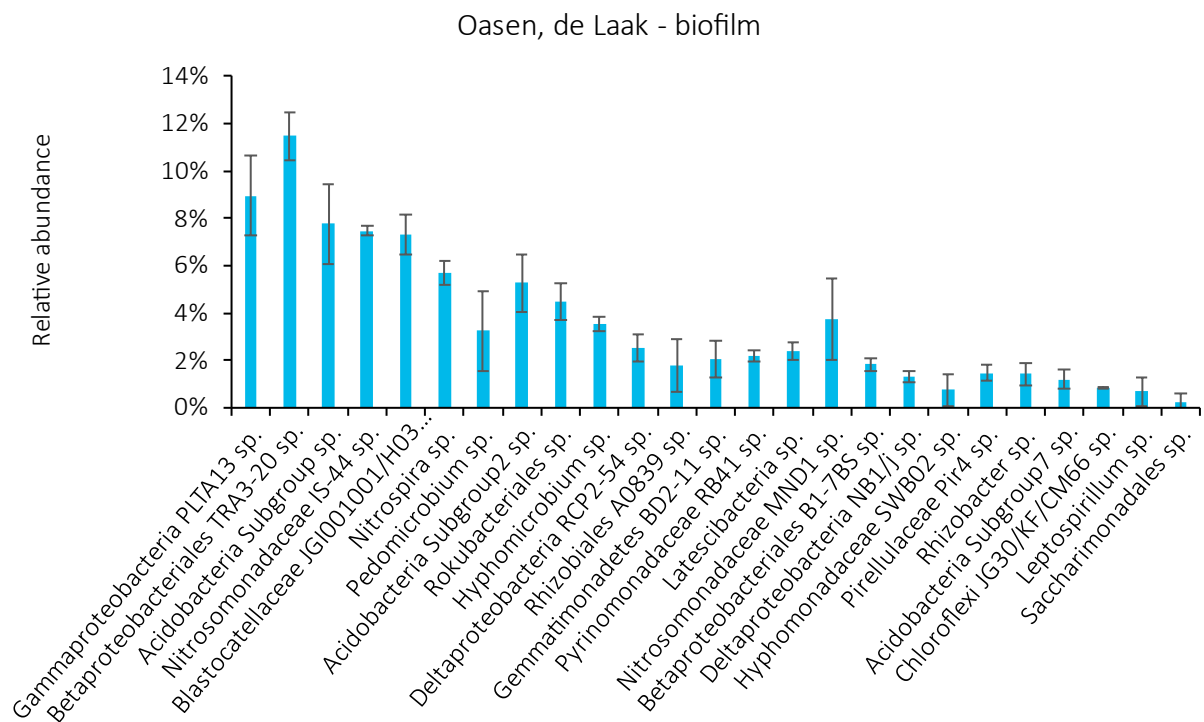


Figure 8. Relative abundance of the top 25 phylogenetic full-length 16S rRNA taxa from the triplicate biofilm sample from Oasen, de Laak. The top 25 included 88-91% of all reads. Clustering is based on the Loop Genomics pipeline. Error bars show standard deviation of the mean.

For the biofilm samples, no contigs were identified up to species level and eight taxa were identified up to genus level (16.2% of the reads)(Figure 8).

Overall, these results stress that limitations in the database, caused by a lack of knowledge about the bacterial species present in these environments, pose the most important limitation for species-level identification of full-length 16S rRNA gene sequences on drinking water, sediment and biofilm samples. This is the result of the availability of mainly bacterial species that have been well-studied (and cultured in most cases) in the used (Silva) and all other available databases. However, to be able to identify rare bacterial species (like pathogens) in these communities with the sequence depth used in these experiments, they have to be present at relative high numbers representing more than 0.05-0.16% relative abundance of the total bacterial community (see 6.2.1).

Comparison Loop Genomics results with short v4 sequencing on samples from production plant “De Laak”.

The same DNA-extracts from drinking water production plant “De Laak” were also analyzed with short 16S rRNA gene v4 fragment sequencing. An overview of the observed top 25 most dominant taxa is shown in Figure 9.

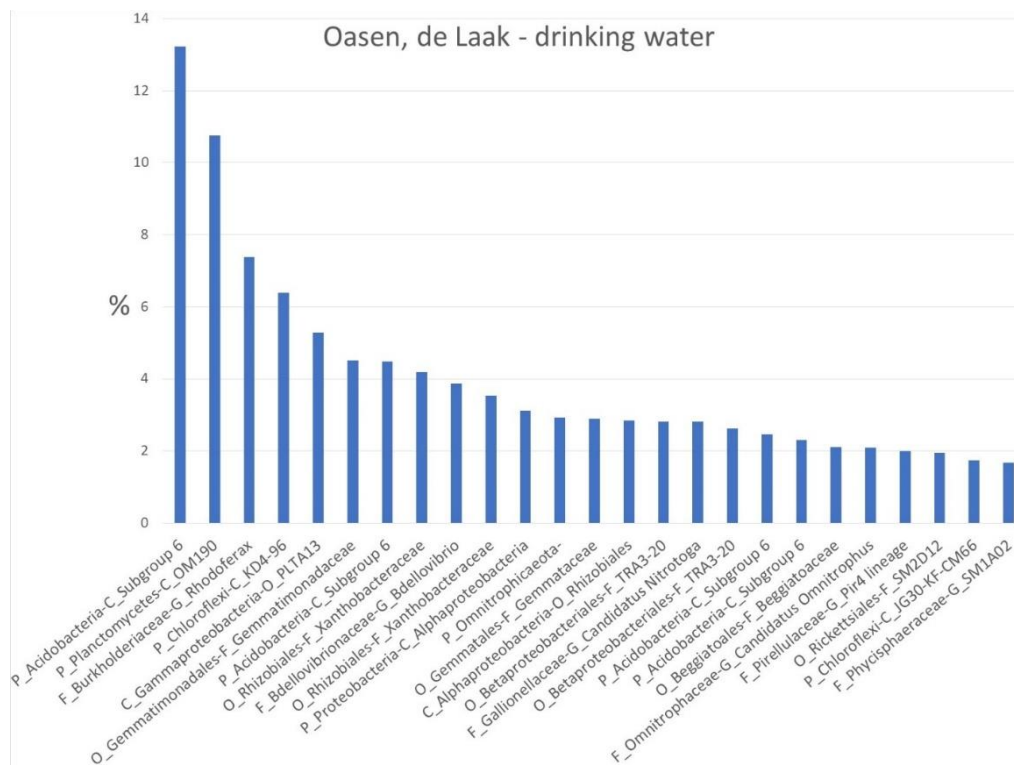


Figure 9. Relative abundance of the top 25 phylogenetic short fragment (v4) sequencing clusters from the drinking water sample from Oasen, de Laak. The top 25 included 49.8% of all reads. Taxonomic identification is based on the pretrained classifier. The following abbreviations are used for the different taxonomic levels: P_ Phylum; C_ Class; _O Order; _F Family; _G Genus; _S Species

From the top 25 most abundant taxa in drinking water, there were no short v4 sequences classified to species level and only four taxa classified to genus (or candidate) level (*Rhodoferrax*, *Bdellovibrio*, *Candidatus-Nitrospira* and *Candidatus-Omnitrophus*). *Rhodoferrax* was identified as the most dominant genus (7.4% of sequences from the top 25 taxa), which is partly in agreement with the Loop Genomics results, where this genus was presumably identified as the species *R. saidenbachensis* (10.2% of sequences from the top 25 taxa). In addition, the most dominant taxon was identified to a subgroup of the *Acidobacteria* phylum using both Loop Genomics and short fragment (v4) sequencing resulting with similar relative abundances (13.1 and 13.2% for Loop Genomics and short fragment sequencing respectively). There is a clear but no perfect overlap between the taxa observed with Loop Genomics and taxa observed with short v4 sequencing. The imperfect overlap between Loop Genomics and short v4 sequencing is not surprising since different primer sets were used to amplify 16S rRNA gene specific sequences for both approaches. It is likely that both primer sets introduce specific biases in the data due to primer specific preferential amplification of 16S rRNA gene sequences of different taxa (Heijnen 2021). The sequences identified to the *C. freundii* species using Loop Genomics were not observed in the short v4 sequencing dataset and the latter dataset also didn't show sequences belonging to the corresponding family (*Enterobacteriaceae*) or order (*Enterobacterales*) of *C. freundii*. A short evaluation of the primers used to generate the Loop Genomics and the short (v4) 16 rRNA amplicons using Testprime 1.0 (<https://www.arb-silva.de/search/testprime>) showed that both primersets target the 16S rRNA gene of *C. freundii* perfectly. And, taking into account that more sequence depth is obtained with short (v4) sequencing (Table 3 and Table 4) makes it unlikely that abundant taxa in Loop Genomics data are missed in short (v4) sequencing data. This makes it unlikely that there are sequences of this coliform bacterium present in this drinking water sample indicating that a contamination, probably caused by cross contamination, was detected using the Loop Genomics approach. This result highlights that care should be taken by next generation sequence analysis, both in the technical and the interpretation part.

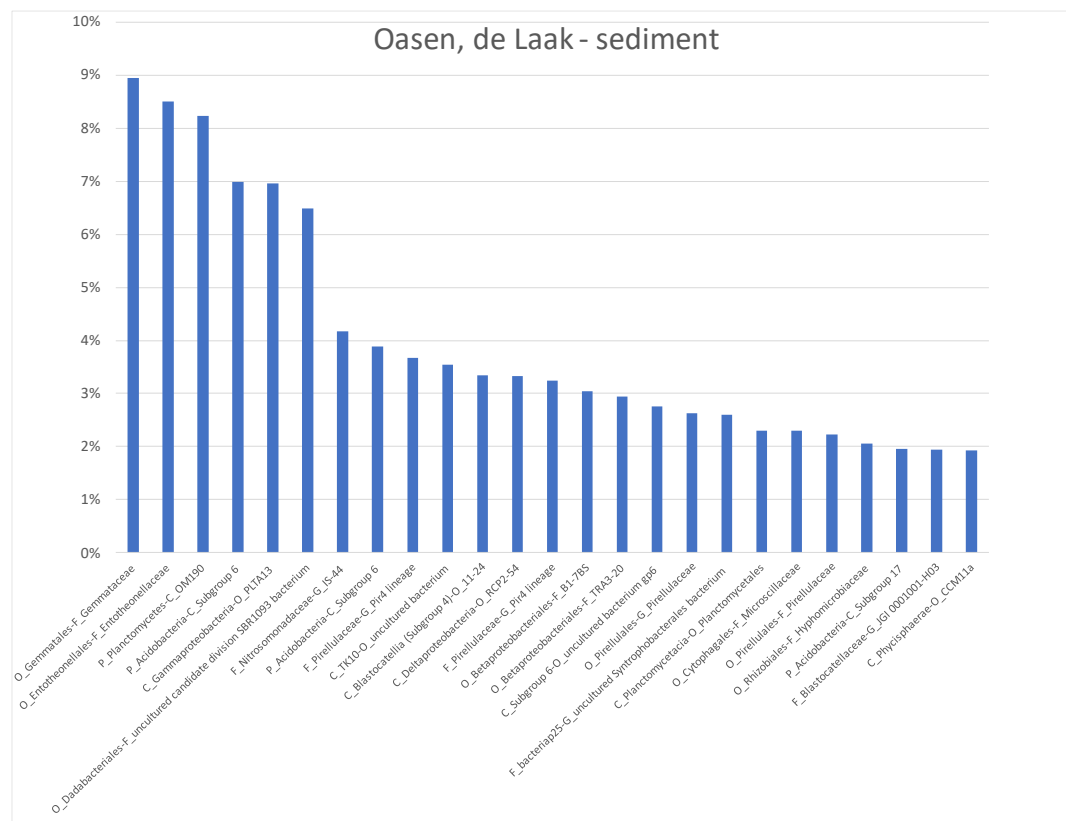


Figure 10. Relative abundance of the top 25 phylogenetic short fragment (v4) sequencing clusters from the sediment sample from Oasen, de Laak. The top 25 included 44.5% of all reads. Taxonomic identification is based on the pretrained classifier (silva_132_515f_806R_ref_water_non_saline_classifier). The following abbreviations are used for the different taxonomic levels: P_ Phylum; C_ Class; _O Order; _F Family; _G Genus; _S Species

In DNA isolated from the sediment of the drinking water distribution system, one taxon was classified to the genus level (*Pedomicrobium*) and no taxons were classified to the species level in the top 25 taxons using short v4 sequences (Figure 10). Except that the most dominant family in short v4 sequences (*Gemmataceae*) was not observed in Loop Genomics data, the overall trends in the observed taxons appears to be comparable between the two sequencing approaches. A short evaluation of the primer sets of both methods shows the presence of mismatches between the Loop Genomics primers and 16S rRNA gene sequences of *Gemmataceae*, whereas the primers used to generate short v4 sequence fragments perfectly match 16S rRNA gene sequences of *Gemmataceae* making this a plausible explanation for this observed difference between the two methods. Differences in the relative abundances of the observed taxons between Loop Genomics and short v4 sequence data are also observed in this sample. These differences are most likely the result of differences in amplification efficiencies between the primer sets used.

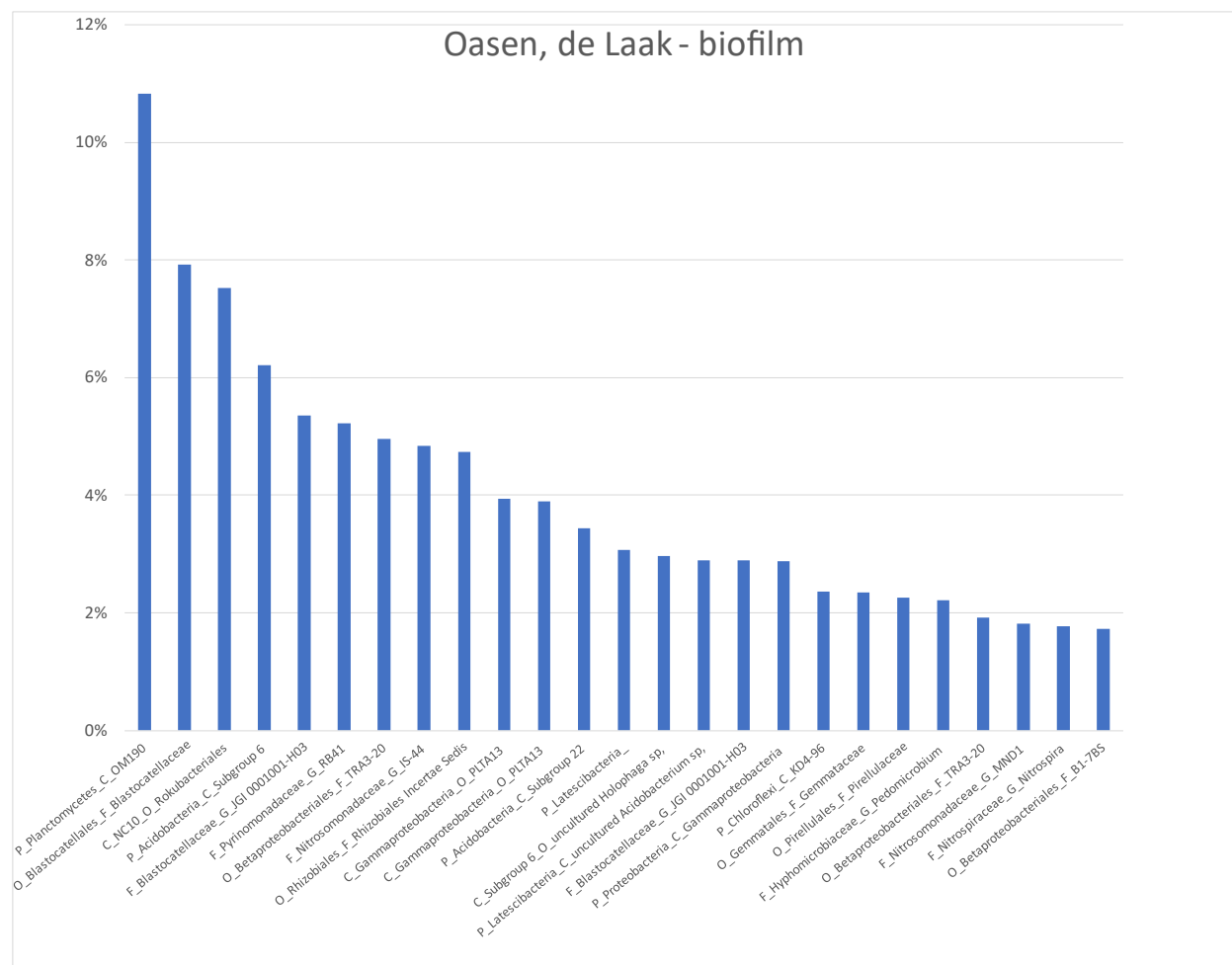


Figure 11. Relative abundance of the top 25 phylogenetic short fragment (v4) sequencing clusters from the biofilm sample from Oasen, de Laak. The top 25 included 63.0% of all reads. Taxonomic identification is based on the pretrained classifier (silva_132_515f_806R_ref_water_non_saline_classifier). The following abbreviations are used for the different taxonomic levels: P_ Phylum; C_ Class; _O Order; _F Family; _G Genus; _S Species

In the top 25 taxa observed in the biofilm sample from production location “De Laak” there were no taxa classified to species level and two taxa classified to the genus level (*Pedomicrobium* and *Nitrospira*) using short v4 sequencing (Figure 11). It is remarkable that the most abundant taxon observed with short v4 sequencing (taxon from the *Planctomycetes* phylum) was not observed with Loop Genomics sequencing. This difference can also be explained by primer-mismatches in the Loop Genomics primers.

6.4.3 Samples from water and monitoring wells from water treatment plant “De Lange Vlieter” (WML)

The top 25 taxa from water and soil from monitoring wells (WPs) of drinking water treatment plant “De Lange Vlieter” are summarized in appendix III. Overall, only two taxa were identified up to species level using Loop Genomics sequencing, but no taxa were identified to the species level with short v4 sequencing. Loop Genomics contigs that could be identified up to species level included *Nitrospira nitrificans* (1.0% of total deep filter 2 WP19 population) and *Pleurocladia lacustris* (2.7% of total soil sample location 1 population). Compared to short v4 sequencing the Loop Genomics approach identified remarkably more taxa to the genus level, confirming the capacity for obtaining a more detailed bacterial classification using full-length 16S rRNA gene sequences. These results confirm the results obtained with the drinking water and sediment samples from the drinking water distribution system of de Laak (paragraph 6.4.1) where Loop Genomics approach was also able to classify more taxa to the species or genus level than the short v4 sequencing method.

A remarkable observation with the samples from de Lange Vlieter is that a relatively large part of the top 25 sequences from the two samples from the filters consist of sequences from archaea (9.6% in water from both

shallow filter 1 and deep filter 2). These archaea are only observed using short v4 sequencing and not with Loop Genomics suggesting that the primers used for Loop Genomics sequencing are not targeting 16S rRNA gene sequences of archaea. Despite this difference between the Loop Genomics and short v4 sequencing results, the other results from both sequencing approaches showed comparable trends.

6.4.4 Samples from the GOBAM project

Only a limited number of sequences were classified to species level in the samples from the GOBAM project (Appendix V) containing water samples from different stages of drinking water treatment using dune infiltration. Loop Genomics resulted in the classification to species level of one taxon (*Bosea vavilviae*) in the effluent of the advanced oxidation process, two taxons (*Methylopumilus turicensis* and *Limnohabitans planktonic*) in a residence basin and three taxons (*Crenothrix polyspora*, *Ferriphaselus Amnicola* and *Methylothera versatilis*) in the abstracted water after dune passage. Short v4 sequencing resulted in classification to species level of only one taxon (*Virgulinella fragilis*), this species was observed in three different samples (effluent rapid sand filtration and in two different residence basins). As in line with the other samples described in the previous paragraph, Loop Genomics resulted to the classification of more taxons to the genus level than short v4 sequencing.

6.5 Discussion: full-length 16S rRNA gene sequencing using Loop Genomics

The results demonstrate that Loop Genomics can be used to:

- I. Generate full-length 16S rRNA gene sequences.
 - A large sequence redundancy (many short read sequences) is needed to assemble many full-length 16S rRNA sequences to be able to detect rare sequences in a community.
- II. Better identify to genus and species level using full-length 16S rRNA sequences compared to short v4 fragment 16S rRNA sequences.
 - This was demonstrated on mock communities and confirmed on samples from drinking water treatment processes.
 - However, species identification is problematic for the far majority of taxons due to limitations of drinking water bacterial and archaeal species in the 16S rRNA gene sequence databases.
- III. Identify bacteria but not archaea. However, Archaea are detected using short v4 fragment 16S rRNA sequencing.

The main challenge for using full-length 16S rRNA gene to classify bacteria to species level is imposed by limitations in the 16S rRNA gene reference databases. The databases that we use at KWR are made by ARB-SILVA (The SILVA ribosomal RNA database project, Max Planck Institute for Marine Microbiology and Jacobs University, Bremen, Germany). The current 16S rRNA gene database which was released in 2019 and named the SILVA 138 SSU (small subunit) database is built-up of a total of nearly 9.5 million 16S rRNA gene sequences, which is a 36% increase compared to the previous SILVA 132 SSU database from 2017. This highlights the rapid increase in sequences present in the database. The current Ref NR 99 database contains 510,984 non redundant sequences and is used as the standard in many studies. In general, this database is the most accurate, up-to-date database available and provides sufficient resolution for the general screening of microorganisms in a wide variety of ecosystems. However, for understudied environments, which mainly include natural, man-modified and manmade environments such as water bodies, soils and sediments, the database still shows limitations. Such environments overall show a very high microbial diversity with many species that have not been studied or cultivated before. Due to the absence of cultured and well-studied species, microbial taxonomy studies are often limited to classification at family to species level. This emphasizes the need for comprehensive identification and characterization of unknown bacteria present in these environments. However, the results obtained on the mock community show that full-length 16S rRNA gene sequences enable the identification of most well-known bacteria to species level using full-length 16S rRNA gene sequences but not with short v4 fragment 16S rRNA gene sequencing. This

suggests that full-length 16S rRNA gene sequences will enable the detection of well-known bacteria, like (opportunistic) pathogens, in these environments. This would make the use of full-length 16S rRNA gene sequencing as a broad pathogen screening method potentially feasible. However, due to the need to sequence short fragments in high redundancy to be able to construct a limited number of full-length 16S rRNA gene sequences, it is expected that it will be costly to use the Loop Genomics approach to detect also the 16S rRNA gene sequences of rare bacteria in a community.

7 Conclusions and Recommendations

7.1 Conclusions

- The Loop Genomics approach enables the identification of sequences to species level in a reproducible way. This was demonstrated by sequence analysis of a mock community containing a mixture of eight characterized bacterial species and two fungi. Almost all bacterial sequences (98.8%) obtained from the mock community samples were correctly identified to genus level and 85.8% of the sequences were correctly identified to the species level.
- Classifying sequences to the species level using Loop Genomics on environmental (water) samples was not very successful, because most of the full-length 16S rRNA gene sequences obtained did not relate to cultivated and described bacterial species, but to bacterial species that have not yet been cultured. This limitation is, thus, not related to the Loop Genomics method to obtain full-length 16S rRNA gene sequences, but to the limitation of the reference database used. This means that any method that produces full-length 16S rRNA gene sequences would face the same problem.
- The Loop Genomics approach results in considerably lower sequencing depth than traditional next generation sequencing of (short parts of) the 16S rRNA genes. In this study the Loop Genomics method needed approximately 1,000,000 reads per sample to construct 1,000 to 2,000 (full-length) 16S rRNA gene sequences, whereas traditional short-read sequencing resulted in 50,000 (short-length) 16S rRNA gene sequences per sample. Consequently, the Loop Genomics method cannot be used to detect low abundant bacterial species in environmental water samples (such as pathogens).
- Short fragment v4 16S rRNA gene sequencing using a specific classifier pretrained on microbial communities from non-saline water and sediment samples did in general not result in classification to the species level with both mock community and environmental (water) samples. It is, therefore, concluded that short fragment V4 sequencing cannot be reliably used to classify sequences to the species level using the currently available classifiers.

7.2 Recommendations

7.2.1 Full length 16S rRNA gene sequencing

This study demonstrates the added value of full-length 16S rRNA sequencing to classify sequences from known bacteria to species level and a large part of unknown bacteria from water samples to the genus level. The species level identification of known bacteria would potentially enable the use of full-length 16S rRNA sequencing as a screening tool for the presence of a wide range of bacterial species, like pathogens. However, the need to generate a high sequence redundancy to be able to construct these full-length sequences makes it an unattractive and expensive method to screen for rare bacteria (such as pathogens) in the drinking water environment. Other sequencing approaches, like Nanopore sequencing (MinION) can sequence much longer DNA-molecules enabling direct sequencing of complete 16S rRNA gene sequences. However, a previous study (Heijnen 2018) demonstrated that in 2018 the quality of these sequences was too low to reliably identify bacterial species. The Oxford Nanopore company has, however, considerably improved the MinION method in the last years by optimizing the membranes used in the sequencing system and the software to translate raw signals to sequence data. These improvement steps have meanwhile resulted in Nanopore sequencing systems that can be used to generate high quality full-length 16S rRNA gene data enabling reliable species identification of bacteria in environmental communities (Zhang et al. 2023).

Due to this quality improvement and the ease of use of MinION sequencing we recommend comparing MinION full-length 16S rRNA gene sequencing with Illumina short fragment 16S rRNA sequencing and Loop Genomics to be able to judge the applicability of MinION full-length 16S rRNA gene sequencing for species identification of known bacteria in communities. When the results show that MinION sequencing can be used to reliably identify microorganisms to the species level, we then recommend to use MinION sequencing in studies where it is important to classify the bacterial community to species level as much as possible.

7.2.2 Expanding the database with cultivated drinking water species

The study also showed that a limitation of using full length 16S rRNA gene sequences to detect drinking water microorganisms to the species level is hindered by the database, because many bacteria in the drinking water environment are not well-described and/or cultivated. It is, therefore, recommended to put more effort in cultivating not-yet described microorganisms from the drinking water environment and analyze their function by performing whole genome sequencing.

8 References

- Amir, A., D. McDonald, J. A. Navas-Molina, E. Kopylova, J. T. Morton, Z. Zech Xu, E. P. Kightley, L. R. Thompson, E. R. Hyde, A. Gonzalez, and R. Knight. 2017. 'Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns', *mSystems*, 2.
- Bokulich, Nicholas A., Benjamin D. Kaehler, Jai Ram Rideout, Matthew Dillon, Evan Bolyen, Rob Knight, Gavin A. Huttley, and J. Gregory Caporaso. 2018. 'Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin', *Microbiome*, 6: 90.
- Burke, Catherine M., and Aaron E. Darling. 2016. 'A method for high precision sequencing of near full-length 16S rRNA genes on an Illumina MiSeq', *PeerJ*, 4: e2492.
- Callahan, Benjamin J, Dmitry Grinevich, Siddhartha Thakur, Michael A Balamotis, and Tuval Ben Yehezekel. 2020. 'Ultra-accurate Microbial Amplicon Sequencing Directly from Complex Samples with Synthetic Long Reads', *bioRxiv*: 2020.07.07.192286.
- Choi, Jaeho, and Jong Soo Park. 2020. 'Comparative analyses of the V4 and V9 regions of 18S rDNA for the extant eukaryotic community using the Illumina platform', *Scientific Reports*, 10: 6519.
- Chung, Nico, Marc W. Van Goethem, Melanie A. Preston, Filip Lhota, Leona Cerna, Ferran Garcia-Pichel, Vanessa Fernandes, Ana Giraldo-Silva, Hee Shin Kim, Evan Hurowitz, Michael Balamotis, Indira Wu, and Tuval Ben-Yehezkel. 2020. 'Accurate Microbiome Sequencing with Synthetic Long Read Sequencing', *bioRxiv*: 2020.10.02.324038.
- DeLong, E F. 1992. 'Archaea in coastal marine environments', *Proceedings of the National Academy of Sciences*, 89: 5685-89.
- Devanga Ragupathi, N. K., D. P. Muthuirulandi Sethuvel, F. Y. Inbanathan, and B. Veeraraghavan. 2017. 'Accurate differentiation of Escherichia coli and Shigella serogroups: challenges and strategies', *New microbes and new infections*, 21: 58-62.
- Fredriksson, Nils Johan, Malte Hermansson, and Britt-Marie Wilén. 2013. 'The choice of PCR primers has great impact on assessments of bacterial community diversity and dynamics in a wastewater treatment plant', *PLOS ONE*, 8: e76431-e31.
- Gao, Xiang, Huaiying Lin, Kashi Revanna, and Qunfeng Dong. 2017. 'A Bayesian taxonomic classification method for 16S rRNA gene sequences with improved species-level accuracy', *BMC Bioinformatics*, 18: 247.
- Hagey, Jill V., Srijak Bhatnagar, Jennifer M. Heguy, Betsy M. Karle, Patricia L. Price, Deanne Meyer, and Elizabeth A. Maga. 2019. 'Fecal Microbial Communities in a Large Representative Cohort of California Dairy Cows', *Frontiers in Microbiology*, 10.
- Heijnen, L. 2018. 'Toepassingsmogelijkheden van Metatranscriptomics, Microbial Profiling en MinION sequencing.', *BTO report*, BTO 2018.044.
- Heijnen, Leo; Timmers, Peer. 2021. 'Microbial profiling met NGS voor drinkwateronderzoek', *BTO report*, BTO 2021.054.
- Hong, SunHee, John Bunge, Chesley Leslin, Sunok Jeon, and Slava S. Epstein. 2009. 'Polymerase chain reaction primers miss half of rRNA microbial diversity', *The ISME Journal*, 3: 1365-73.
- Hugerth, Luisa W., Emilie E. L. Muller, Yue O. O. Hu, Laura A. M. Lebrun, Hugo Roume, Daniel Lundin, Paul Wilmes, and Anders F. Andersson. 2014. 'Systematic Design of 18S rRNA Gene Primers for Determining Eukaryotic Diversity in Microbial Consortia', *PLOS ONE*, 9: e95567.
- Jeong, Jinuk, Kyeongueui Yun, Seyoung Mun, Won-Hyong Chung, Song-Yi Choi, Young-do Nam, Mi Young Lim, Chang Pyo Hong, ChanHyeok Park, Yong Ahn, and Kyudong Han. 2021. 'The effect of taxonomic classification by full-length 16S rRNA sequencing with a synthetic long-read technology', *Scientific Reports*, 11: 1727.
- Johnson, Jethro S., Daniel J. Spakowicz, Bo-Young Hong, Lauren M. Petersen, Patrick Demkowicz, Lei Chen, Shana R. Leopold, Blake M. Hanson, Hanako O. Agresta, Mark Gerstein, Erica Sodergren, and George M. Weinstock. 2019. 'Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis', *Nature Communications*, 10: 5029.
- Juottonen, Heli, Laurent Fontaine, Christian Wurzbacher, Stina Drakare, Sari Peura, and Alexander Eiler. 2020. 'Archaea in boreal Swedish lakes are diverse, dominated by Woese archaeota and follow deterministic community assembly', *Environmental Microbiology*, 22: 3158-71.

- Karst, S. M., M. S. Dueholm, S. J. McIlroy, R. H. Kirkegaard, P. H. Nielsen, and M. Albertsen. 2018. 'Retrieval of a million high-quality, full-length microbial 16S and 18S rRNA gene sequences without primer bias', *Nat Biotechnol*, 36: 190-95.
- Karst, Søren M., Ryan M. Ziels, Rasmus H. Kirkegaard, Emil A. Sørensen, Daniel McDonald, Qiyun Zhu, Rob Knight, and Mads Albertsen. 2020. 'Enabling high-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing', *bioRxiv*: 645903.
- Klindworth, Anna, Elmar Pruesse, Timmy Schweer, Jörg Peplies, Christian Quast, Matthias Horn, and Frank Oliver Glöckner. 2012. 'Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies', *Nucleic Acids Research*, 41: e1-e1.
- Panda, Adyasha, T. Islam Salim, and Gaurav Sharma. 2022. 'Harmonizing Prokaryotic Nomenclature: Fixing the Fuss over Phylum Name Flipping', *mBio*, 13: e00970-22.
- Prodan, Andrei, Valentina Tremaroli, Harald Brolin, Aeilko H. Zwinderman, Max Nieuwdorp, and Evgeni Levin. 2020. 'Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing', *PLOS ONE*, 15: e0227434.
- Quast, Christian, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. 2012. 'The SILVA ribosomal RNA gene database project: improved data processing and web-based tools', *Nucleic Acids Res*, 41: D590-D96.
- Stapleton, J. A., J. Kim, J. P. Hamilton, M. Wu, L. C. Irber, R. Maddamsetti, B. Briney, L. Newton, D. R. Burton, C. T. Brown, C. Chan, C. R. Buell, and T. A. Whitehead. 2016. 'Haplotype-Phased Synthetic Long Reads from Short-Read Sequencing', *PLOS ONE*, 11: e0147229.
- Yan, Yan, Long H. Nguyen, Eric A. Franzosa, and Curtis Huttenhower. 2020. 'Strain-level epidemiology of microbial communities and the human microbiome', *Genome Medicine*, 12: 71.
- Yilmaz, Pelin, Laura Wegener Parfrey, Pablo Yarza, Jan Gerken, Elmar Pruesse, Christian Quast, Timmy Schweer, Jörg Peplies, Wolfgang Ludwig, and Frank Oliver Glöckner. 2013. 'The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks', *Nucleic Acids Res*, 42: D643-D48.
- Zhang, T., H. Li, S. Ma, J. Cao, H. Liao, Q. Huang, and W. Chen. 2023. 'The newest Oxford Nanopore R10.4.1 full-length 16S rRNA sequencing enables the accurate resolution of species-level microbial community profiling', *Appl Environ Microbiol*, 89: e0060523.

8.1 BTO reports, KWR reports and newsletters

BTO and KWR reports

- van der Grift, B., Timmers, P. (2020) Verwijdering OMV's tijdens bodempassage in het infiltratiebekken De Lange Vlieter. BTO 2020.206(s), KWR, Nieuwegein.
- Heijnen, L. (2019). Verkenning van de toepassingsmogelijkheden van MinION sequencing. BTO 2019.057, KWR, Nieuwegein.
- Heijnen, L. (2018). Toepassingsmogelijkheden van: Metatranscriptomics, Microbial profiling en MinION sequencing. BTO 2018.044, KWR, Nieuwegein.
- Learbuch, K. Timmers, P. (2020). Microbiologische compositie en potentiële metabole processen tijdens drinkwaterdistributie. BTO 2020.044, KWR, Nieuwegein.
- Heijnen, L. Timmers, P. (2022) Microbial profiling met NGS voor drinkwateronderzoek. BTO 2021.054, KWR, Nieuwegein
- Timmers, P. (2021). NGS bronopsporing pilot inlaatwater Woerden. KWR 2021.013, KWR.
- Van der Grift B., Timmers P. 2020. Verwijdering OMV's tijdens bodempassage in het infiltratiebekken De Lange Vlieter. BTO 2020.206(s).
- Timmers et al. 2021. Effecten van geavanceerde oxidatie op de waterkwaliteit en microbiologie bij duininfiltratie. KWR 2020.066.

Newsletters

- Heijnen, L., Timmers, P. (juni 2019). Nieuwsbrief: kansen van nieuwe moleculair-microbiologische methoden in de drinkwatersector. KWR, Nieuwegein.

in 't Zandt, M. (juli 2021). Nieuwsbrief: kansen van nieuwe molecuair-microbiologische methoden in de drinkwatersector. *KWR*, Nieuwegein.

I Bijlage: Begrippenlijst

16S rRNA gen	Het 16S rRNA komt voor in prokaryoten (bacteriën en archaea) en wordt gecodeerd door het 16S rRNA gen. Het vormt onderdeel van het ribosoom, het eiwit-synthetiserend organel van een cel.
18S rRNA gen	Het 18S rRNA komt voor in eukaryoten en wordt gecodeerd door het 18S rRNA gen. Het vormt onderdeel van het ribosoom, het eiwit-synthetiserend organel van een cel.
Alignment	Het proces waarbij DNA-sequenties gealigneerd (bij elkaar gezet) worden. Hierbij worden sequenties zo gesorteerd en gegroepeerd dat gelijkende DNA-sequenties van (fragmenten van) genen netjes onder elkaar komen te staan. Het doel van “alignement” is om op een snelle manier vergelijkingen tussen verschillende sequenties te kunnen maken om zo de gelijkenis in kaart te kunnen brengen.
Amplicon sequencing	Het proces waarbij een groot aantal kopieën van eenzelfde DNA-fragment wordt gegenereerd via PCR en vervolgens wordt gesequenced. Een goed voorbeeld hiervan is 16S rRNA sequencing waarbij (meestal) een kort fragment van het 16S rRNA gen wordt geselecteerd, geamplificeerd met PCR en vervolgens wordt gesequenced. Er kunnen ook andere (functionele) genen worden geselecteerd voor amplicon sequencing.
Annotatie	Het proces waarbij met behulp van een referentiedatabase in DNA sequencing-data wordt gezocht naar de aanwezigheid en identiteit van genen. Hier wordt in de regel de geassembleerde (aan elkaar geplakte) DNA-data of zelfs de MAGs voor gebruikt.
ASV	Bij de ASV-methode wordt eerst voor elke unieke sequentie bepaald hoe vaak deze voorkomt. Vervolgens wordt hierop een foutcorrectie toegepast om sequencing-fouten uit de data te halen. Elke ASV wordt vervolgens geclassificeerd aan de hand van een referentiedatabase.
Basenpaar	De eenheid van twee basen die de basis vormt voor het dubbelstrengs DNA-molecuul. Van nature vindt basenparing plaats tussen de nucleotiden A en T en de nucleotiden C en G. In dubbelstrengs RNA vindt eenzelfde basenparing plaats, maar bevindt zich de nucleotide U in plaats van T.
Bias	De vertekening van de data door afwijkingen in de gebruikte methode(n). Bij DNA-sequencing zijn bekende punten waarbij biases optreden de DNA-extractie en de DNA-sequencing.
Binning	Het proces waarbij vanuit geassembleerde (aan elkaar geplakte) DNA-data genomen worden gereconstrueerd. Deze gereconstrueerde

genomen worden in de regel bins of metagenome assembled genomes (MAGs) genoemd.

Bonito

De vaste basecaller van Oxford Nanopore waarmee de ruwe meetdata van de sequencing-machines wordt omgezet in nucleotiden-sequenties.

CCS

Circular consensus sequencing. Ne nucleotidenvolgorde van een DNA-sequentie die bepaald is aan de hand van een aantal sequencing-resultaten van dezelfde DNA-sequentie. Om de consensus-sequentie te maken worden deze kopieën op elkaar gelegd en wordt hieruit voor elke nucleotide op elke positie de meestvoorkomende nucleotide, ofwel de consensus gehaald. Bij CCS wordt dit specifiek gedaan met een circulair DNA-molecuul dat een reeks aan kopieën van eenzelfde DNA-fragment bevat.

Clade(n)

Een groep organismen die allen afstemmen van dezelfde voorouder (inclusief deze voorouder). Een clade wordt ook wel een monofyletische groep genoemd.

Degenerate primer

Een mix van verschillende sequence-varianties van eenzelfde primer die in staat zijn om een brede(re) diversiteit van de doelorganismen te dekken. Bij 16S rRNA gen-sequencing wordt vaak gebruik gemaakt van degenerate primers om een zo breed mogelijke microbiële diversiteit op te kunnen pikken.

Deletie (DNA)

Een mutatie (verandering) in het DNA waarbij de hoeveelheid genetisch materiaal afneemt. Een deletie kan plaatsvinden tijdens het sequencing-proces. Vaak gaat het om puntdeleties waarbij één enkele nucleotide uit een DNA-volgorde wegvalt. Het kan echter ook om grotere stukken DNA gaan.

DNA polymerase

Het enzymcomplex (eiwit-complex) dat betrokken is bij het repliceren van het DNA. Het enzymcomplex werkt met een enkelstrengs DNA-molecuul en plakt hier voor elke nucleotide de complementaire nucleotide aan om zo dubbelstrengs DNA te maken.

DNA synthese

Het proces waarbij nieuwe DNA-sequenties worden gemaakt.

DNA-sequencing

Het proces waarbij van DNA-moleculen nucleotide-data wordt verkregen.

Eukaryoot

Een eukaryoot is een eencellig of meercellig organisme die een nucleus met DNA heeft. Onder de eukaryoten vallen onder andere dieren, planten, schimmels en protozoën.

Fylogenetische relatie

Vaak cladistiek genoemd. Het proces waarbij het evolutionair verwantschap tussen verschillende organismen in kaart wordt gebracht. Dit kan alleen gedaan worden op een eigenschap die aanwezig is in alle onderzochte organismen. Bij bacteriën vormt het

16S rRNA gen hier een goed voorbeeld van.

Fylum	Ook wel stam genoemd. Rang van taxonomische classificatie (niveau boven “Klasse” en onder “Rijk”).
Genus	Ook wel Geslacht genoemd. In <i>Legionella pneumophila</i> is bijvoorbeeld het genus “ <i>Legionella</i> ”. <i>Legionella anisa</i> en <i>L. pneumophila</i> behoren dus tot hetzelfde genus.
HiFi	High Fidelity (voor PacBio sequencing). Proces waarbij door het genereren van een gecirculariseerde kopie van één enkel DNA-fragment een zeer nauwkeurige consensus-sequentie kan worden gegenereerd.
Illumina	Sequencing-platform en methode waarbij de productie van fluorescente lichtsignalen wordt gebruikt voor het in kaart brengen van de DNA-volgorde van een molecuul. De methode werkt met zogeheten “reversible terminated chemistry” waarbij na het inbouwen van één nucleotide allereerst de fluorescente terminator (die verdere replicatie blokkeert) verwijderd dient te worden. Dit gebeurt na de detectie van de fluorescente terminator. Deze fluorescente terminator is uniek voor elke nucleotide en zo kan de DNA-volgorde verkregen worden. Om voldoende signaal te genereren vindt op de sequencing chip een PCR-amplificatiestap (dit proces wordt “bridge amplification” genoemd) plaats om van de aanwezige unieke DNA-strengen grote clusters met kopieën te genereren.
in silico	Een experiment dat uitgevoerd wordt op een computer of via een computersimulatie.
Insertie (DNA)	Een mutatie (verandering) in het DNA waarbij de hoeveelheid genetisch materiaal toeneemt. Een insertie kan plaatsvinden tijdens het sequencing-proces. Vaak gaat het om de insertie van een enkele nucleotide, maar het kan ook gaan om de insertie van langere DNA-fragmenten. (niet te verwarren met duplicatie waarbij stukken DNA in een genoom worden gekopieerd).
ITS	De internally transcribed spacers komen voor tussen het 18S rRNA gen en het 5.8S rRNA gen (ITS1) en tussen het 5.8S rRNA gen en het 28S rRNA gen (ITS2). De ITS-sequenties coderen niet voor RNA-sequenties die in het ribosoom voorkomen.
Klasse	Ook wel “Class” genoemd. Rang van taxonomische classificatie (niveau boven “Orde” en onder “Stam”).
Long-read technologie	Sequencing-technologie waarbij data van lange DNA-sequenties wordt gegenereerd. Deze zijn in de regel in ieder geval langer dan 1,000 basenparen, maar vaak zeker enkele duizenden basenparen tot zelfs

honderdduizenden basenparen.

LSU

LSU staat voor 'large subunit', ofwel het grote fragment van het ribosomale RNA, gecodeerd door het 23S rRNA gen (bacteriën) en het 28S rRNA gen (eukaryoten; schimmels en protozoa).

Machine learning

Het proces waarbij met zelflerende software-tools op een geautomatiseerde manier de verwerking en interpretatie van data wordt geoptimaliseerd.

Metagenoom sequencing

Ook wel "metagenomics" genoemd. Het proces waarbij het volledige DNA van een monster wordt gesequenced.

MinION

DNA-sequencer ontwikkeld door Oxford Nanopore. Het is de kleinste sequencer in de reeks. De sequencing werkt op basis van een eiwit nanopore waarover een ionische (zoutoplossing) spanning staat. Hierdoor is er een constante stroom van ionen door de nanopore. Deze stroom wordt continu gemeten. Wanneer een enkelstrengs DNA-molecuul door deze nanopore gaat, verandert de ionenstroom. Deze verandering is uniek voor elke base en kan door de sensor worden geregistreerd. Vervolgens wordt deze data omgezet naar sequencing-data via het proces dat "basecalling" wordt genoemd.

Mock community

Een kunstmatig samengestelde (microbiële) gemeenschap met daarin bekende ratio's van goed gekarakteriseerde micro-organismen. Voor veel mock communities kan zowel levend celmateriaal als DNA-materiaal besteld worden.

Nanopore

De eiwitporie van een Oxford Nanopore sequencer (zoals bijvoorbeeld de MinION) waardoor de constante ionenstroom plaatsvindt en waardoor ook het enkelstrengs DNA loopt. Aan de nanopore zit een sensor gekoppeld waarmee de ionenstroom wordt gemeten.

NGS

Afkorting van "next-generation sequencing", een verzamelnaam voor sequencing-methoden ontwikkeld na Sanger sequencing. In de praktijk vaak gebruikt om een amplicon-sequencing studie te beschrijven op een NGS-machine.

non-redundant

Bij databases een term om aan te geven dat dubbele of zeer op elkaar gelijkende referentiesequenties uit de database gefilterd zijn om zo de database enigszins gecomprimeerd te houden, wat data-analyse versnelt en ruis in de data-analyse vermindert.

Nucleotide

Bouwsteen van het DNA en RNA.

OTU

Een OTU definieert een groep van nauw verwante 16S rRNA gen-sequenties die in principe (en bij benadering) sequenties zijn die tot één soort behoren.

Oxford Nanopore	De ontwikkelaar van de nanopore-gebaseerde sequencing-technologieën.
PacBio	De ontwikkelaar van de PacBio sequencing-technologieën.
Polymere membraan	Een kunstmatige membraan met vaak zeer vergelijkbare eigenschappen met een natuurlijke membraan die in levende cellen voorkomt. Van een polymere membraan is echter de exacte samenstelling bekend en kan ook de samenstelling van eiwitten in de membraan nauwkeurig worden afgesteld. Polymere membranen worden bijvoorbeeld bij sequencing van Oxford Nanopore gebruikt.
Potentieel (microbieel)	De genetische capaciteit van een micro-organisme om een bepaald proces (een bepaalde serie chemische reacties) uit te voeren. Het potentieel wordt bepaald aan de hand van de aanwezige genen die coderen voor eiwitten/enzymen die de functies uit kunnen voeren.
Primer	De primer is het startpunt voor de PCR amplificatie. Een primer is een kort stuk enkelstrengs DNA, vaak van circa 10-20 nucleotiden lang, dat complementair is aan een sequentie op het DNA van een organisme. Doordat de sequentie complementair is, bindt deze primer sterk op het DNA van het doelorganisme. Een specifieke primer voor bijvoorbeeld het 16S rRNA gen van Legionella-soorten zal er voor zorgen dat alleen het Legionella-16S rRNA gen met PCR wordt geamplificeerd.
Prokaryoot	Een eencellig organisme dat geen nucleus met DNA heeft. Onder de prokaryoten vallen de bacteriën en de archaea.
Protozoa	Een verzamelnaam voor een groep van eencellige, heterotrofe (levend van organisch materiaal) eukaryoten die in staat zijn om zichzelf voort te bewegen.
qPCR	Afkorting van “quantitative PCR” ofwel kwantitatieve PCR, ook wel real-time PCR genoemd. Bij een qPCR kan tijdens de duplicatie van DNA-fragmenten de hoeveelheid gevormde DNA-fragmenten worden gevormd, doordat er gelabelde probes worden toegevoegd die aan specifieke stukken gesynthetiseerd DNA kunnen binden. Het signaal van een qPCR wordt vervolgens als relatief getal uitgedrukt (RFU = relative fluorescent unit) en wordt gekoppeld aan het aantal PCR cycli. Met de data kan worden teruggerekend hoeveel specifieke fragmenten van een bepaald gen zich in het originele monster bevinden.
Read	Een read is een stuk DNA-sequentie dat het resultaat is van DNA-sequencing. Afhankelijk van het sequencing-platform zijn reads relatief kort (Illumina, enkele honderden basenparen) of juist zeer lang (MinION, tot honderdduizenden en zelfs miljoenen basenparen, vooral afhankelijk van hoe lang de aangeboden DNA-moleculen zijn).

RNA-extractie	Het proces waarbij selectief RNA-moleculen uit een monster worden gehaald.
Sanger Sequencing	Ook wel didesoxysequencing genoemd. Het is de meest toegepaste first-generation sequencing-methode. De sequencing bestaat uit een speciale PCR-reactie met didesoxy-nucleotiden (ddNTs). Ook zit aan elke nucleotide een marker die bij de analyse herkend kan worden. In de eerste versie was dit vaak een algemeen label, later werd dit een uniek fluorescent label per nucleotide. Ook bevinden zich in de reactiemix normale nucleotiden. Wanneer een ddNT wordt ingebouwd stopt de DNA-synthese. Dit is dus gebaseerd op kans en kan overal in het DNA-fragment gebeuren. Vervolgens worden alle fragmenten op gel (of automatische capillaire gel) gerund. Omdat bekend is welke ddNT waar is ingebouwd kan vervolgens met de lengte van het fragment worden bepaald welke nucleotide waar zit.
Schimmels	Wetenschappelijke naam: Fungi. Eukaryote organismen die beschikken over een celwand (net als planten). Tot de schimmels behoren zowel meercellige organismen (zoals paddenstoelen) als eencellige organismen (zoals gist).
Short-read technologie	Sequencing-technologie waarbij data van korte DNA-sequenties wordt gegenereerd. Deze zijn in de regel in ieder geval korter dan 1,000 basenparen en in de regel maar enkele honderden basenparen.
SILVA	Handmatig samengestelde en opgeschoonde eferentiedatabase voor het 16S en 18S rRNA gen. De database bevat (nagenoeg) complete gen-sequenties van alle in online databases aanwezige 16S en 18S rRNA gen-reads. Het meest gebruikt zijn de non-redundant (NR) databases waarbij redundantie uit de data is gefilterd om de data-analyse sneller en met minder ruis te laten verlopen.
SMRT	SMRT staat voor Single Molecule Real-Time. Het is een methode waarbij in één reactie één enkel molecuul wordt gesequenced op een vergelijkbare wijze als bij Illumina sequencing. Het aflezen van het inbouwen van nucleotiden gebeurt middels een lasersignaal in een kleine well, een “zero mode waveguide” of ZMW genoemd.
Soort	Ook wel species genoemd. Rang van taxonomische classificatie (niveau boven ondersoort/stam en onder “Genus”).
SSU	Small SubUnit. Afkorting voor het korte fragment van het ribosomaal RNA. In prokaryoten (onder andere bacteriën) wordt het SSU gecodeerd door het 16S rRNA gen; in eukaryoten (onder anderen protozoa) wordt het SSU gecodeerd door het 18S rRNA gen.
Stam	Een stam (“strain” in het Engels) is in de taxonomie de onderverdeling van verschillende sub-soorten. Er zijn bijvoorbeeld verschillende <i>Legionella pneumophila</i> stammen (AA100, JR32, Lp01 et cetera) die allen tot dezelfde soort behoren, namelijk <i>L. pneumophila</i> . De

verschillen tussen deze stammen zijn dus niet groot genoeg om als verschillende soorten te worden gezien. Toch zijn er wel duidelijke verschillen tussen deze stammen in bijvoorbeeld de virulentie of het energiemetabolisme.

Tag (sequencing)

Het principe waarbij één enkele DNA-sequentie wordt gekoppeld met één specifieke DNA-sequentie, waardoor het na sequencing eenvoudig is om de verschillende DNA-sequenties te groeperen. Het principe is hetzelfde als bij UMIs.

Tagmentation

Het proces bij DNA-tagging waarbij het DNA (meestal enzymatisch) in kleinere stukken wordt geknipt om zo op een Illumina-platform gesequenced te kunnen worden. Na het opdelen in kleinere stukken zit de specifieke tag aan één van de kleinere fragmenten. Alleen de fragmenten die een tag bezitten zullen uiteindelijk gesequenced worden.

Taxonomische groep

Ook wel taxonomische rang genoemd. Het is een hiërarchisch niveau voor het indelen van organismen. Voorbeelden van rangen zijn Soort, Geslacht en Familie.

UMI

Bekende DNA-sequenties met een lengte van 18 nucleotiden. Deze sequenties worden gebruikt om na het sequencen een kwaliteitscontrole uit te voeren en vervolgens de read-data te kunnen clusteren.

UniEuk

Taxonomisch raamwerk en referentie-database voor eukaryoten.

Universele primer

Een primer die in staat is om een grote diversiteit aan organismen op te pikken. Vaak wordt er bij universele primers gebruik gemaakt van een mix van verschillende primer-samenstellingen (degenerate primers).

II Identification of the Mock community with “short (v4) sequencing

Mock		Short (v4) sequence classification			Relative abundance (%)		
Community	Classifier	Family	Genus	Species	Mock_1	Mock_2	Average (%)
Bacillus	Silva	Bacillaceae	Bacillus	—	20.2	19.3	19.8
subtilis	Pretrained Silva				20.2	19.3	19.8
Lactobacillus	Silva	Lactobacillaceae	Lactobacillus	—	15.3	16.7	16.0
fermentum	Pretrained Silva				15.3	16.7	16.0
Listeria	Silva	Listeriaceae	Listeria	—	13.8	12.4	13.1
monocytogenes	Pretrained Silva			Ambiguous_taxa	13.8	12.4	13.1
Escherichia	Silva	Enterobacteriaceae	Escherichia-Shigella	—	12.8	11.5	12.1
coli	Pretrained Silva				12.8	11.5	12.1
Staphylococcus	Silva	Staphylococcaceae	Staphylococcus	—	12.5	16.5	14.5
aureus	Pretrained Silva			S. warneri	12.5	16.5	14.5
Salmonella	Silva	Enterobacteriaceae	—	—	10.6	9.6	10.1
enterica	Pretrained Silva				10.6	9.6	10.1
Enterococcus	Silva	Enterococcaceae	Enterococcus	—	9.0	8.3	8.7
faecalis	Pretrained Silva			E. durans	9.0	8.3	8.7
Pseudomonas	Silva	Pseudomonadaceae	Pseudomonas	—	5.9	5.7	5.8
aeruginosa	Pretrained Silva			P. aeruginosa	5.9	5.7	5.8

III Top 25 taxons in samples “De Laak”

III.I Short V4 16S rRNA gene sequencing

	Kingdom	Phylum	Class	Order	Family	Genus	Species	% Sequences
1	Bacteria	Acidobacteria	Subgroup 6	—	—	—	—	6.7%
2	Bacteria	Planctomycetes	OM190	uncultured bacterium	uncultured bacterium	uncultured bacterium	uncultured bacterium	5.4%
3	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Rhodoferrax	—	3.7%
4	Bacteria	Chloroflexi	KD4-96	Ambiguous_taxa	Ambiguous_taxa	Ambiguous_taxa	Ambiguous_taxa	3.2%
5	Bacteria	Proteobacteria	Gammaproteobacteria	PLTA13	Ambiguous_taxa	Ambiguous_taxa	Ambiguous_taxa	2.7%
6	Bacteria	Gemmatimonadetes	Gemmatimonadales	Gemmatimonadales	Gemmatimonadaceae	uncultured	—	2.3%
7	Bacteria	Acidobacteria	Subgroup 6	uncultured proteobacter	uncultured proteobacterium	uncultured proteobacterium	uncultured proteobacterium	2.3%
8	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Xanthobacteraceae	—	—	2.1%
9	Bacteria	Proteobacteria	Deltaproteobacteria	Bdellovibrionales	Bdellovibrionaceae	Bdellovibrio	uncultured bacterium	1.9%
10	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Xanthobacteraceae	uncultured	—	1.8%
11	Bacteria	Proteobacteria	Alphaproteobacteria	uncultured	uncultured bacterium	uncultured bacterium	uncultured bacterium	1.5%
12	Bacteria	Omnitrophicae	uncultured bacterium	uncultured bacterium	uncultured bacterium	uncultured bacterium	uncultured bacterium	1.5%
13	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	—	—	—	1.4%
14	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Gallionellaceae	Candidatus Nitrotoga	uncultured bacterium	1.4%
15	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	TRA3-20	—	—	1.4%
16	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	TRA3-20	metagenome	metagenome	1.3%
17	Bacteria	Planctomycetes	Planctomycetacia	Gemmatales	Gemmataceae	uncultured	uncultured bacterium	1.3%
18	Bacteria	Acidobacteria	Subgroup 6	uncultured bacterium gp1	uncultured bacterium gp6	uncultured bacterium gp6	uncultured bacterium gp6	1.2%
19	Bacteria	Acidobacteria	Subgroup 6	uncultured Holophaga sp	uncultured Holophaga sp,	uncultured Holophaga sp,	uncultured Holophaga sp,	1.2%
20	Bacteria	Proteobacteria	Gammaproteobacteria	Beggiatoales	Beggiatoaceae	uncultured	uncultured bacterium	1.1%
21	Bacteria	Omnitrophicae	Omnitrophia	Omnitrophales	Omnitrophaceae	Candidatus Omnitrophus	uncultured bacterium	1.0%
22	Bacteria	Planctomycetes	Planctomycetacia	Pirellulales	Pirellulaceae	Pir4 lineage	Ambiguous_taxa	1.0%
23	Bacteria	Proteobacteria	Alphaproteobacteria	Rickettsiales	SM2D12	uncultured bacterium	uncultured bacterium	1.0%
24	Bacteria	Chloroflexi	JG30-KF-CM66	—	—	—	—	0.9%
25	Bacteria	Planctomycetes	Phycisphaerae	Phycisphaerales	Phycisphaeraceae	SM1A02	uncultured bacterium	0.8%

Top 25 taxons identified using short v4 16S rRNA sequencing in sample “Drinking Water” (the species or genus identification is highlighted in green).

	Kingdom	Phylum	Class	Order	Family	Genus	Species	% sequences
1	Bacteria	Planctomycetes	Planctomycetacia	Gemmatales	Gemmataceae	uncultured	uncultured bacterium	4.0
2	Bacteria	Entotheonellaeota	Entotheonellia	Entotheonellales	Entotheonellaceae	—	—	3.8
3	Bacteria	Planctomycetes	OM190	uncultured bacterium	uncultured bacterium	uncultured bacterium	uncultured bacterium	3.7
4	Bacteria	Acidobacteria	Subgroup 6	—	—	—	—	3.1
5	Bacteria	Proteobacteria	Gammaproteobacteria	PLTA13	Ambiguous_taxa	Ambiguous_taxa	Ambiguous_taxa	3.1
6	Bacteria	Dadabacteria	Dadabacteria	Dadabacteriales	uncultured candidate division SBR109	uncultured candidate divi	uncultured candidate division SBR109	2.9
7	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Nitrosomonadaceae	IS-44	—	1.9
8	Bacteria	Acidobacteria	Subgroup 6	uncultured proteobacterium	uncultured proteobacterium	uncultured proteobacteri	uncultured proteobacterium	1.7
9	Bacteria	Planctomycetes	Planctomycetacia	Pirellulales	Pirellulaceae	Pir4 lineage	Ambiguous_taxa	1.6
10	Bacteria	Chloroflexi	TK10	uncultured bacterium	uncultured bacterium	uncultured bacterium	uncultured bacterium	1.6
11	Bacteria	Acidobacteria	Blastocatellia (Subgroup 4)	nov-24	uncultured bacterium	uncultured bacterium	uncultured bacterium	1.5
12	Bacteria	Proteobacteria	Deltaproteobacteria	RCP2-54	metagenome	metagenome	metagenome	1.5
13	Bacteria	Planctomycetes	Planctomycetacia	Pirellulales	Pirellulaceae	Pir4 lineage	uncultured Pirellula sp,	1.4
14	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	B1-7B5	uncultured bacterium	uncultured bacterium	1.4
15	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	TRA3-20	—	—	1.3
16	Bacteria	Acidobacteria	Subgroup 6	uncultured bacterium gp6	uncultured bacterium gp6	uncultured bacterium gp6	uncultured bacterium gp6	1.2
17	Bacteria	Planctomycetes	Planctomycetacia	Pirellulales	Pirellulaceae	uncultured	Ambiguous_taxa	1.2
18	Bacteria	Proteobacteria	Deltaproteobacteria	Myxococcales	bacteriap25	uncultured Syntrophobac	uncultured Syntrophobacteriales bacte	1.2
19	Bacteria	Planctomycetes	Planctomycetacia	Planctomycetales	uncultured	—	—	1.0
20	Bacteria	Bacteroidetes	Bacteroidia	Cytophagales	Microscillaceae	uncultured	metagenome	1.0
21	Bacteria	Planctomycetes	Planctomycetacia	Pirellulales	Pirellulaceae	uncultured	uncultured bacterium	1.0
22	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Hyphomicrobiaceae	Pedomicrobium	—	0.9
23	Bacteria	Acidobacteria	Subgroup 17	metagenome	metagenome	metagenome	metagenome	0.9
24	Bacteria	Acidobacteria	Blastocatellia (Subgroup 4)	Blastocatellales	Blastocatellaceae	JGI 0001001-H03	Ambiguous_taxa	0.9
25	Bacteria	Planctomycetes	Phycisphaerae	CCM11a	uncultured bacterium	uncultured bacterium	uncultured bacterium	0.9

Top 25 taxons identified using short v4 16S rRNA sequencing in sample “Sediment in distribution after flushing”.

	Kingdom	Phylum	Class	Order	Family	Genus	Species	% Sequences
1	Bacteria	Planctomycetes	OM190	uncultured bacterium	uncultured bacterium	uncultured bacterium	uncultured bacterium	6.8%
2	Bacteria	Acidobacteria	Blastocatellia (Subgroup NC10	Blastocatellales	Blastocatellaceae	—	—	5.0%
3	Bacteria	Rokubacteria	NC10	Rokubacteriales	Ambiguous_taxa	Ambiguous_taxa	Ambiguous_taxa	4.8%
4	Bacteria	Acidobacteria	Subgroup 6	—	—	—	—	3.9%
5	Bacteria	Acidobacteria	Blastocatellia (Subgroup	Blastocatellales	Blastocatellaceae	JGI 0001001-H03	uncultured bacterium	3.4%
6	Bacteria	Acidobacteria	Blastocatellia (Subgroup	Pyrinomonadales	Pyrinomonadaceae	RB41	uncultured Acidobacterium sp,	3.3%
7	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	TRA3-20	—	—	3.1%
8	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Nitrosomonadaceae	IS-44	—	3.1%
9	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Rhizobiales Incertae Sedis	uncultured	—	3.0%
10	Bacteria	Proteobacteria	Gammaproteobacteria	PLTA13	uncultured proteobacterium	uncultured proteobacterium	uncultured proteobacterium	2.5%
11	Bacteria	Proteobacteria	Gammaproteobacteria	PLTA13	Ambiguous_taxa	Ambiguous_taxa	Ambiguous_taxa	2.5%
12	Bacteria	Acidobacteria	Subgroup 22	—	—	—	—	2.2%
13	Bacteria	Latescibacteria	—	—	—	—	—	1.9%
14	Bacteria	Acidobacteria	Subgroup 6	uncultured Holophaga sp	uncultured Holophaga sp,	uncultured Holophaga sp,	uncultured Holophaga sp,	1.9%
15	Bacteria	Latescibacteria	uncultured Acidobacteriu	uncultured Acidobacteriu	uncultured Acidobacterium sp,	uncultured Acidobacterium sp,	uncultured Acidobacterium sp,	1.8%
16	Bacteria	Acidobacteria	Blastocatellia (Subgroup	Blastocatellales	Blastocatellaceae	JGI 0001001-H03	—	1.8%
17	Bacteria	Proteobacteria	Gammaproteobacteria	—	—	—	—	1.8%
18	Bacteria	Chloroflexi	KD4-96	Ambiguous_taxa	Ambiguous_taxa	Ambiguous_taxa	Ambiguous_taxa	1.5%
19	Bacteria	Planctomycetes	Planctomycetacia	Pirellulales	Pirellulaceae	—	—	1.4%
20	Bacteria	Planctomycetes	Planctomycetacia	Gemmatales	Gemmataceae	uncultured	uncultured bacterium	1.4%
21	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Hyphomicrobiaceae	Pedomicrobium	—	1.4%
22	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	TRA3-20	metagenome	metagenome	1.2%
23	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Nitrosomonadaceae	MND1	—	1.2%
24	Bacteria	Nitrospirae	Nitrospira	Nitrospirales	Nitrospiraceae	Nitrospira	uncultured Green Bay ferromanganous	1.1%
25	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	B1-7BS	uncultured bacterium	uncultured bacterium	1.1%

Top 25 taxons identified using short v4 16S rRNA sequencing in sample “Biofilm in distribution” (the species or genus identification is accentuated in green).

III.II Loop genomics sequencing

	kingdom	phylum	class	order	family	genus	species	% sequences
1	Bacteria	Acidobacteria	unidentified	unidentified	unidentified	Subgroup	Subgroup sp.	12.0%
2	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Rhodoferrax	Rhodoferrax saidenbachii	9.9%
3	Bacteria	Proteobacteria	Gammaproteobacteria	unidentified	unidentified	PLTA13	PLTA13 sp.	8.1%
4	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	unidentified	TRA3-20	TRA3-20 sp.	6.7%
5	Bacteria	Proteobacteria	Deltaproteobacteria	unidentified	unidentified	RCP2-54	RCP2-54 sp.	2.8%
6	Bacteria	Proteobacteria	Betaproteobacteria	Betaproteobacteriales	Nitrosomonadaceae	MND1	MND1 sp.	2.4%
7	Bacteria	Planctomycetes	unidentified	unidentified	unidentified	OM190	OM190 sp.	2.4%
8	Bacteria	Acidobacteria	Acidobacteriia	Solibacterales	Solibacteraceae (Subgroup	Bryobacter	Bryobacter sp.	2.3%
9	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	unidentified	Xanthobacteraceae	Xanthobacteraceae sp.	2.1%
10	Bacteria	Nitrospirae	Nitrospira	Nitrospirales	Nitrospiraceae	Nitrospira	Nitrospira sp.	2.0%
11	Bacteria	Proteobacteria	Alphaproteobacteria	Reyranellales	Reyranellaceae	Reyranella	Reyranella sp.	2.0%
12	Bacteria	Proteobacteria	Alphaproteobacteria	Caulobacteriales	Hyphomonadaceae	SWB02	SWB02 sp.	1.5%
13	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	unidentified	KF/JG30/B3	KF/JG30/B3 sp.	1.5%
14	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Hyphomicrobiaceae	Pedomicrobium	Pedomicrobium sp.	1.2%
15	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	unidentified	A0839	A0839 sp.	1.2%
16	Bacteria	Nitrospirae	Nitrospira	Nitrospirales	Nitrospiraceae	Leptospirillum	Leptospirillum sp.	1.2%
17	Bacteria	Chloroflexi	Dehalococcoidia	unidentified	unidentified	S085	S085 sp.	1.1%
18	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Xanthobacteraceae	Rhodoplanes	Rhodoplanes sp.	1.1%
19	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Citrobacter	Citrobacter freundii	1.1%
20	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Rhizobiales Incertae Sedis	Bauldia	Bauldia sp.	0.9%
21	Bacteria	Proteobacteria	Alphaproteobacteria	Rickettsiales	unidentified	SM2D12	SM2D12 sp.	0.9%
22	Bacteria	Dependentiae	Babeliae	Babeliales	unidentified	Vermiphilaceae	Vermiphilaceae sp.	0.9%
23	Bacteria	Entotheonellaeo	Entotheonellia	Entotheonellales	Entotheonellaceae	Entotheonella	Entotheonella sp.	0.9%
24	Bacteria	Planctomycetes	Planctomycetacia	Pirellulales	Pirellulaceae	Pir4	Pir4 sp.	0.8%
25	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Gallionellaceae	Nitrotoga	Nitrotoga sp.	0.8%

Top 25 taxons identified using Loop Genomics sequencing in sample “Drinking Water” (the species or genus identification is accentuated in green).

	kingdom	phylum	class	order	family	genus	species	% sequences
1	Bacteria	Entothaeonellae	Entothaeonellia	Entothaeonellales	Entothaeonellaceae	Entothaeonella	Entothaeonella sp.	9.6%
2	Bacteria	Acidobacteria	unidentified	unidentified	unidentified	Subgroup	Subgroup sp.	7.0%
3	Bacteria	Proteobacteria	Gammaproteobacteria	unidentified	unidentified	PLTA13	PLTA13 sp.	5.7%
4	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	unidentified	TRA3-20	TRA3-20 sp.	5.3%
5	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Nitrosomonadaceae	IS-44	IS-44 sp.	4.4%
6	Bacteria	Proteobacteria	Deltaproteobacteria	unidentified	unidentified	RCP2-54	RCP2-54 sp.	3.9%
7	Bacteria	Nitrospirae	Nitrospira	Nitrospirales	Nitrospiraceae	Nitrospira	Nitrospira sp.	3.7%
8	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Nitrosomonadaceae	MND1	MND1 sp.	3.6%
9	Bacteria	Planctomycetes	Planctomycetacia	Pirellulales	Pirellulaceae	Pir4	Pir4 sp.	3.2%
10	Bacteria	Dadabacteria	Dadabacteria	unidentified	unidentified	Dadabacteriales	Dadabacteriales sp.	2.9%
11	Bacteria	Proteobacteria	Alphaproteobacteria	Caulobacterales	Hyphomonadaceae	SWB02	SWB02 sp.	2.2%
12	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Rhizobacter	Rhizobacter sp.	1.7%
13	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	unidentified	A0839	A0839 sp.	1.6%
14	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Hyphomicrobiaceae	Hyphomicrobium	Hyphomicrobium sp.	1.6%
15	Bacteria	Calditrichaeota	Calditrichia	Calditrichales	Calditrichaceae	JdFR-76	JdFR-76 sp.	1.6%
16	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Hyphomicrobiaceae	Pedomicrobium	Pedomicrobium sp.	1.4%
17	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	unidentified	B1-7BS	B1-7BS sp.	1.4%
18	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Nitrosomonadaceae	Ellin6067	Ellin6067 sp.	1.4%
19	Bacteria	Planctomycetes	unidentified	unidentified	unidentified	OM190	OM190 sp.	1.3%
20	Bacteria	Nitrospirae	Nitrospira	Nitrospirales	Nitrospiraceae	Leptospirillum	Leptospirillum sp.	1.3%
21	Bacteria	Acidobacteria	unidentified	unidentified	unidentified	Subgroup7	Subgroup7 sp.	1.3%
22	Bacteria	Bacteroidetes	Bacteroidia	Chitinophagales	Chitinophagaceae	Terrimonas	Terrimonas sp.	1.3%
23	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Nitrosomonadaceae	mle1-7	mle1-7 sp.	1.3%
24	Bacteria	Proteobacteria	Alphaproteobacteria	Reynanellales	Reynanellaceae	Reynanella	Reynanella sp.	1.1%
25	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	unidentified	Xanthobacteraceae	Xanthobacteraceae sp.	1.0%

Top 25 taxons identified using Loop Genomics sequencing in sample “Sediment in distribution after flushing” (the species or genus identification is accentuated in green).

	kingdom	phylum	class	order	family	genus	species	% Sequences
1	Bacteria	Proteobacteria	Gammaproteobacteria	unidentified	unidentified	PLTA13	PLTA13 sp.	10.8%
2	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	unidentified	TRA3-20	TRA3-20 sp.	10.6%
3	Bacteria	Acidobacteria	unidentified	unidentified	unidentified	Subgroup	Subgroup sp.	7.5%
4	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Nitrosomonadaceae	IS-44	IS-44 sp.	7.5%
5	Bacteria	Acidobacteria	Blastocatellia (Subgroup)	Blastocatellales	Blastocatellaceae	JGI001001/H03	JGI001001/H03 sp.	7.3%
6	Bacteria	Nitrospirae	Nitrospira	Nitrospirales	Nitrospiraceae	Nitrospira	Nitrospira sp.	5.7%
7	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Hyphomicrobiaceae	Pedomicrobium	Pedomicrobium sp.	5.1%
8	Bacteria	Acidobacteria	unidentified	unidentified	unidentified	Subgroup2	Subgroup2 sp.	4.0%
9	Bacteria	Rokubacteria	NC10	unidentified	unidentified	Rokubacteriales	Rokubacteriales sp.	3.7%
10	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Hyphomicrobiaceae	Hyphomicrobium	Hyphomicrobium sp.	3.5%
11	Bacteria	Proteobacteria	Deltaproteobacteria	unidentified	unidentified	RCP2-54	RCP2-54 sp.	3.1%
12	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	unidentified	A0839	A0839 sp.	3.1%
13	Bacteria	Gemmatimonadetes	unidentified	unidentified	unidentified	BD2-11	BD2-11 sp.	2.9%
14	Bacteria	Acidobacteria	Blastocatellia (Subgroup)	Pyrinomonadales	Pyrinomonadaceae	RB41	RB41 sp.	2.4%
15	Bacteria	unidentified	unidentified	unidentified	unidentified	Latescibacteria	Latescibacteria sp.	2.2%
16	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Nitrosomonadaceae	MND1	MND1 sp.	1.8%
17	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	unidentified	B1-7BS	B1-7BS sp.	1.8%
18	Bacteria	Proteobacteria	Deltaproteobacteria	unidentified	unidentified	NB1/j	NB1/j sp.	1.3%
19	Bacteria	Proteobacteria	Alphaproteobacteria	Caulobacterales	Hyphomonadaceae	SWB02	SWB02 sp.	1.3%
20	Bacteria	Planctomycetes	Planctomycetacia	Pirellulales	Pirellulaceae	Pir4	Pir4 sp.	1.1%
21	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Rhizobacter	Rhizobacter sp.	1.1%
22	Bacteria	Acidobacteria	unidentified	unidentified	unidentified	Subgroup7	Subgroup7 sp.	0.9%
23	Bacteria	Chloroflexi	unidentified	unidentified	unidentified	JG30/KF/CM66	JG30/KF/CM66 sp.	0.9%
24	Bacteria	Nitrospirae	Nitrospira	Nitrospirales	Nitrospiraceae	Leptospirillum	Leptospirillum sp.	0.9%
25	Bacteria	Patescibacteria	Saccharimonadia	unidentified	unidentified	Saccharimonadales	Saccharimonadales sp.	0.7%

Top 25 taxons identified using Loop Genomics sequencing in sample “Biofilm in distribution” (the species or genus identification is accentuated in green).

IV Top 25 taxons in samples “De lange Vlieter”

IV.I Short V4 16S rRNA gene sequencing

							nr. sequences	100. (%)
1	Bacteria	Omnitrophicaeae	uncultured bacterium	uncultured bacterium	uncultured bacterium	uncultured bacterium	uncultured bacterium	29.0%
2	Archaea	Thaumarchaeota	Marine Benthic Group A	uncultured archaeon	uncultured archaeon	uncultured archaeon	uncultured archaeon	5.6%
3	Bacteria	Nitrospirae	Thermodesulfobionia	uncultured	—	—	—	2.2%
4	Bacteria	—	—	—	—	—	—	2.0%
5	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	—	—	1.9%
6	Bacteria	Chloroflexi	Dehalococcoidia	SAR202 clade	uncultured bacterium	uncultured bacterium	uncultured bacterium	1.7%
7	Bacteria	Nitrospirae	Thermodesulfobionia	uncultured	uncultured bacterium	uncultured bacterium	uncultured bacterium	1.5%
8	Archaea	Nanoarchaeae	Nanohaloarchaea	Aenigmarchaeales	uncultured archaeon	uncultured archaeon	uncultured archaeon	1.5%
9	Bacteria	Proteobacteria	Gammaproteobacteria	Methylococcales	Methylomonaceae	Crenothrix	—	1.4%
10	Archaea	Nanoarchaeae	Woesearchaeia	uncultured euryarchaeote	uncultured euryarchaeote	uncultured euryarchaeote	uncultured euryarchaeote	1.4%
11	Proteobacteria	Deltaproteobacteria	Syntrophobacteriales	Syntrophaceae	uncultured	uncultured delta proteobacterium	—	1.3%
12	Archaea	Nanoarchaeae	Woesearchaeia	—	—	—	—	1.3%
13	Archaea	Thaumarchaeota	Nitrososphaeria	Nitrosopumilales	Nitrosopumilaceae	Candidatus Nitrosoarchaeum	uncultured archaeon	1.3%
14	Bacteria	Chloroflexi	Dehalococcoidia	—	—	—	—	0.9%
15	Bacteria	Patescibacteria	Berkelbacteria	uncultured bacterium	uncultured bacterium	uncultured bacterium	uncultured bacterium	0.8%
16	Bacteria	Proteobacteria	Deltaproteobacteria	Syntrophobacteriales	Syntrophaceae	Desulfobacca	uncultured bacterium	0.8%
17	Bacteria	Patescibacteria	Gracilibacteria	Candidatus Peregrinibact	uncultured bacterium	uncultured bacterium	uncultured bacterium	0.8%
18	Bacteria	Acidobacteria	Subgroup 6	—	—	—	—	0.7%
19	Bacteria	Proteobacteria	Deltaproteobacteria	Desulfobacteriales	Desulfobulbaceae	uncultured	uncultured bacterium	0.6%
20	Archaea	Crenarchaeota	Bathyarchaeia	uncultured archaeon	uncultured archaeon	uncultured archaeon	uncultured archaeon	0.6%
21	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Gallionellaceae	Gallionella	Ambiguous_taxa	0.6%
22	Archaea	Diapherotrites	Micrarchaeia	uncultured archaeon	uncultured archaeon	uncultured archaeon	uncultured archaeon	0.6%
23	Bacteria	Patescibacteria	Berkelbacteria	—	—	—	—	0.6%
24	Bacteria	Chloroflexi	—	—	—	—	—	0.6%
25	Archaea	Crenarchaeota	Bathyarchaeia	uncultured archaeon Arc	uncultured archaeon Arc,201	uncultured archaeon Arc,201	uncultured archaeon Arc,201	0.6%

Top 25 taxons identified using short v4 16S rRNA sequencing in sample “Deep Filter 2 WP19” (the species or genus identification is accentuated in green).

	Kingdom	Phylum	Class	Order	Family	Genus	Species	% Sequences
1	Bacteria	Omnitrophicaeae	uncultured bacterium	uncultured bacterium	uncultured bacterium	uncultured bacterium	uncultured bacterium	7.0%
2	Archaea	Thaumarchaeota	Nitrososphaeria	Nitrosopumilales	Nitrosopumilaceae	uncultured archaeon	uncultured archaeon	4.9%
3	Archaea	Nanoarchaeae	Woesearchaeia	—	—	—	—	3.8%
4	Bacteria	Acidobacteria	Subgroup 6	—	—	—	—	2.9%
5	Archaea	Nanoarchaeae	Woesearchaeia	uncultured euryarchaeote	uncultured euryarchaeote	uncultured euryarchaeote	uncultured euryarchaeote	2.9%
6	Bacteria	Proteobacteria	Deltaproteobacteria	DTB120	uncultured bacterium	uncultured bacterium	uncultured bacterium	2.7%
7	Bacteria	—	—	—	—	—	—	2.5%
8	Bacteria	Chloroflexi	Dehalococcoidia	SAR202 clade	uncultured bacterium	uncultured bacterium	uncultured bacterium	2.3%
9	Archaea	Thaumarchaeota	Nitrososphaeria	Nitrosopumilales	Nitrosopumilaceae	Candidatus Nitrosoarchaeum	uncultured archaeon	1.9%
10	Bacteria	Patescibacteria	Gracilibacteria	Candidatus Peregrinibact	uncultured bacterium	uncultured bacterium	uncultured bacterium	1.8%
11	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Gallionellaceae	—	—	1.8%
12	Bacteria	Chloroflexi	Dehalococcoidia	SAR202 clade	—	—	—	1.6%
13	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	—	—	—	1.5%
14	Bacteria	Chloroflexi	Dehalococcoidia	MSBL5	uncultured bacterium	uncultured bacterium	uncultured bacterium	1.5%
15	Bacteria	Rokubacteria	NC10	Methyloirabiliales	Methyloirabilaceae	Candidatus Methyloirabilis	uncultured Firmicutes bacterium	1.5%
16	Bacteria	Chloroflexi	—	—	—	—	—	1.4%
17	Bacteria	Chloroflexi	Dehalococcoidia	—	—	—	—	1.4%
18	Bacteria	Patescibacteria	Gracilibacteria	Candidatus Peribacteria	uncultured bacterium	uncultured bacterium	uncultured bacterium	1.3%
19	Bacteria	Omnitrophicaeae	Omnitrophia	Omnitrophales	Omnitrophaceae	Candidatus Omnitrophus	uncultured bacterium	1.2%
20	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Gallionellaceae	Candidatus Nitrotoga	uncultured bacterium	1.2%
21	Bacteria	Proteobacteria	Deltaproteobacteria	Myxococcales	bacteriap25	Deltaproteobacteria bacterium	Deltaproteobacteria bacterium RIFCSP	1.2%
22	Bacteria	Rokubacteria	NC10	Rokubacteriales	uncultured bacterium	uncultured bacterium	uncultured bacterium	1.2%
23	Bacteria	Nitrospirae	Thermodesulfobionia	uncultured	uncultured bacterium	uncultured bacterium	uncultured bacterium	1.1%
24	Archaea	Thaumarchaeota	Marine Benthic Group A	uncultured archaeon	uncultured archaeon	uncultured archaeon	uncultured archaeon	1.1%

Top 25 taxons identified using short v4 16S rRNA sequencing in sample “Shallow filter 1 WP19” (the species or genus identification is accentuated in green).

Kingdom	Phylum	Class	Order	Family	Genus	Species	% Sequences
						nr. sequences	100. (%)
1	Bacteria	Acidobacteria	Subgroup 6	uncultured prokaryote	uncultured prokaryote	uncultured prokaryote	2.8%
2	Bacteria	Proteobacteria	Gammaproteobacteria	PLTA13	—	—	2.6%
3	Bacteria	Acidobacteria	Subgroup 6	—	—	—	2.5%
4	Bacteria	Chloroflexi	KD4-96	Ambiguous_taxa	Ambiguous_taxa	Ambiguous_taxa	1.9%
5	Bacteria	Chloroflexi	KD4-96	—	—	—	1.7%
6	Bacteria	Chloroflexi	Anaerolineae	Anaerolineales	Anaerolineaceae	uncultured	1.5%
7	Bacteria	Chloroflexi	Anaerolineae	Anaerolineales	Anaerolineaceae	uncultured	1.4%
8	Bacteria	Planctomycetes	Planctomycetacia	Pirellulales	Pirellulaceae	uncultured bacterium	1.3%
9	Bacteria	Nitrospirae	Thermodesulfobivibria	uncultured	uncultured bacterium	uncultured bacterium	1.3%
10	Bacteria	Latescibacteria	—	—	—	—	1.0%
11	Bacteria	Chloroflexi	Dehalococcoidia	MSBL5	uncultured bacterium	uncultured bacterium	1.0%
12	Bacteria	Proteobacteria	Deltaproteobacteria	Sva0485	uncultured prokaryote	uncultured prokaryote	1.0%
13	Bacteria	Chloroflexi	Anaerolineae	Anaerolineales	Anaerolineaceae	uncultured	1.0%
14	Bacteria	Proteobacteria	Deltaproteobacteria	MBNT15	uncultured prokaryote	uncultured prokaryote	1.0%
15	Archaea	Thaumarchaeota	Nitrososphaeria	Nitrosopumilales	Nitrosopumilaceae	Candidatus Nitrosoarchaeum	1.0%
16	Archaea	Crenarchaeota	Bathymarchaeia	uncultured methanogeni	uncultured methanogenic archaea	uncultured methanogenic archaea	0.9%
17	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	SC-I-84	—	0.9%
18	Bacteria	Actinobacteria	Thermoleophilia	Gaiellales	Gaiellaceae	Gaiella	0.9%
19	Archaea	Crenarchaeota	Bathymarchaeia	uncultured euryarchaeota	uncultured euryarchaeote	uncultured euryarchaeote	0.8%
20	Bacteria	—	—	—	—	—	0.8%
21	Bacteria	Proteobacteria	Deltaproteobacteria	MBNT15	Ambiguous_taxa	Ambiguous_taxa	0.7%
22	Bacteria	Rokubacteria	NC10	Rokubacteriales	—	—	0.7%
23	Bacteria	Chloroflexi	KD4-96	uncultured bacterium	uncultured bacterium	uncultured bacterium	0.7%
24	Bacteria	Verrucomicrobia	Verrucomicrobiae	Pedospaerales	Pedospaeraceae	—	0.7%
25	Bacteria	Nitrospirae	P9X2b3D02	uncultured prokaryote	uncultured prokaryote	uncultured prokaryote	0.7%

Top 25 taxons identified using short v4 16S rRNA sequencing in soil sample location 1 (the species or genus identification is accentuated in green).

IV.II Loop genomics sequencing

	kingdom	phylum	class	order	family	genus	species	% Sequences
1	Bacteria	unidentified	unidentified	unidentified	unidentified	Omnitrophicaeota	Omnitrophicaeota sp.	12.3%
2	Bacteria	Proteobacteria	Gammaproteobacteria	Methylococcales	Methylomonaceae	Crenothrix	Crenothrix sp.	4.5%
3	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Acidovorax	Acidovorax sp.	4.0%
4	Bacteria	Proteobacteria	Deltaproteobacteria	Syntrophobacterales	Syntrophaceae	Desulfobacca	Desulfobacca sp.	3.4%
5	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Gallionellaceae	Gallionella	Gallionella sp.	3.4%
6	Bacteria	Nitrospirae	Nitrospira	Nitrospirales	Nitrospiraceae	Nitrospira	Nitrospira sp.	3.4%
7	Bacteria	Proteobacteria	Deltaproteobacteria	unidentified	unidentified	Sva0485	Sva0485 sp.	2.0%
8	Bacteria	Acidobacteria	unidentified	unidentified	unidentified	Subgroup	Subgroup sp.	1.9%
9	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Undibacterium	Undibacterium sp.	1.9%
10	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Nitrosomonadaceae	Ellin6067	Ellin6067 sp.	1.9%
11	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Nitrosomonadaceae	Nitrosomonas	Nitrosomonas sp.	1.5%
12	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Piscinibacter	Piscinibacter sp.	1.5%
13	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Massilia	Massilia sp.	1.4%
14	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Comamonas	Comamonas sp.	1.4%
15	Bacteria	Proteobacteria	Deltaproteobacteria	Desulfuromonadales	Geobacteraceae	Geobacter	Geobacter sp.	1.3%
16	Bacteria	Firmicutes	Clostridia	Clostridiales	Clostridiaceae 1	Clostridium	Clostridium sp.	1.1%
17	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Rhodoferax	Rhodoferax sp.	1.0%
18	Bacteria	Proteobacteria	Alphaproteobacteria	Caulobacterales	Hyphomonadaceae	SWB02	SWB02 sp.	1.0%
19	Bacteria	Actinobacteria	Thermoleophilia	Solirubrobacterales	unidentified	67-14	67-14 sp.	1.0%
20	Bacteria	Acidobacteria	Holophagae	Holophagales	Holophagaceae	Geothrix	Geothrix sp.	1.0%
21	Bacteria	Nitrospirae	Nitrospira	Nitrospirales	Nitrospiraceae	Nitrospira	Nitrospira nitrificans	1.0%
22	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Pseudorhodobacter	Pseudorhodobacter sp.	1.0%
23	Bacteria	Actinobacteria	Actinobacteria	Propionibacteriales	Nocardioidaceae	Nocardioides	Nocardioides sp.	0.9%
24	Bacteria	Proteobacteria	Deltaproteobacteria	unidentified	unidentified	MBNT15	MBNT15 sp.	0.8%
25	Bacteria	Acidobacteria	unidentified	unidentified	unidentified	Subgroup7	Subgroup7 sp.	0.8%

Top 25 taxons identified using Loop Genomics in sample "Deep Filter 2 WP19 (the species or genus identification is highlighted in green).

	kingdom	phylum	class	order	family	genus	species	% Sequences
1	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Gallionellaceae	Gallionella	Gallionella sp.	8.2%
2	Bacteria	Proteobacteria	Deltaproteobacteria	unidentified	unidentified	DTB120	DTB120 sp.	7.8%
3	Bacteria	unidentified	unidentified	unidentified	unidentified	Omnitrophicaeota	Omnitrophicaeota sp.	6.9%
4	Bacteria	Acidobacteria	unidentified	unidentified	unidentified	Subgroup	Subgroup sp.	6.5%
5	Bacteria	Rokubacteria	NC10	unidentified	unidentified	Rokubacteriales	Rokubacteriales sp.	6.0%
6	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	unidentified	TRA3-20	TRA3-20 sp.	4.3%
7	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Sulfuricellaceae	Sulfuricella	Sulfuricella sp.	4.3%
8	Bacteria	Rokubacteria	NC10	Methyloirabiales	Methyloirabillaceae	Methyloirabillis	Methyloirabillis sp.	3.9%
9	Bacteria	Proteobacteria	Deltaproteobacteria	Myxococcales	unidentified	bacteriap25	bacteriap25 sp.	3.9%
10	Bacteria	Proteobacteria	Gammaproteobacteria	Acidiferrobacteriales	Acidiferrobacteraceae	Sulfurifustis	Sulfurifustis sp.	2.6%
11	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Gallionellaceae	Nitrotoga	Nitrotoga sp.	2.2%
12	Bacteria	Patescibacteria	Parcubacteria	unidentified	unidentified	UBA9983	UBA9983 sp.	2.2%
13	Bacteria	Entothionellae	Entothionellia	Entothionellales	unidentified	Entothionellaceae	Entothionellaceae sp.	2.2%
14	Bacteria	unidentified	unidentified	unidentified	unidentified	Latescibacteria	Latescibacteria sp.	1.7%
15	Bacteria	Nitrospirae	Nitrospira	Nitrospirales	Nitrospiraceae	Nitrospira	Nitrospira sp.	1.3%
16	Bacteria	Gemmatimonade	unidentified	unidentified	unidentified	BD2-11	BD2-11 sp.	1.3%
17	Bacteria	Bacteroidetes	Bacteroidia	Sphingobacteriales	unidentified	AKYH767	AKYH767 sp.	1.3%
18	Bacteria	Nitrospirae	Nitrospira	Nitrospirales	Nitrospiraceae	Leptospirillum	Leptospirillum sp.	1.3%
19	Bacteria	Acidobacteria	Acidobacteriia	unidentified	unidentified	Subgroup	Subgroup sp.	1.3%
20	Bacteria	Patescibacteria	unidentified	unidentified	unidentified	Gracilibacteria	Gracilibacteria sp.	0.9%
21	Bacteria	Actinobacteria	Thermoleophilia	Gaiellales	Gaiellaceae	Gaiella	Gaiella sp.	0.9%
22	Bacteria	Proteobacteria	Gammaproteobacteria	Coxiellales	Coxiellaceae	Coxiella	Coxiella sp.	0.9%
23	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Gallionellaceae	Sideroxydans	Sideroxydans sp.	0.9%
24	Bacteria	Bacteroidetes	Ignavibacteria	Kryptoniales	unidentified	BSV26	BSV26 sp.	0.9%
25	Bacteria	Rokubacteria	NC10	Methyloirabiales	Methyloirabillaceae	Sh765B-TzT-35	Sh765B-TzT-35 sp.	0.9%

Top 25 taxa identified using Loop Genomics in sample "Shallow filter 1 WP19" (the species or genus identification is highlighted in green).

	kingdom	phylum	class	order	family	genus	species	% Sequences
1	Bacteria	Nitrospirae	Nitrospira	Nitrospirales	Nitrospiraceae	Nitrospira	Nitrospira sp.	6.4%
2	Bacteria	Acidobacteria	unidentified	unidentified	unidentified	Subgroup	Subgroup sp.	5.8%
3	Bacteria	Bacteroidetes	Bacteroidia	Flavobacteriales	Flavobacteriaceae	Flavobacterium	Flavobacterium sp.	4.3%
4	Bacteria	Cyanobacteria	Oxyphotobacteria	unidentified	unidentified	Chloroplast	Chloroplast sp.	3.4%
5	Bacteria	Acidobacteria	unidentified	unidentified	unidentified	Subgroup7	Subgroup7 sp.	3.1%
6	Bacteria	unidentified	unidentified	unidentified	unidentified	Latescibacteria	Latescibacteria sp.	3.0%
7	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	unidentified	TRA3-20	TRA3-20 sp.	2.8%
8	Bacteria	Cyanobacteria	Oxyphotobacteria	Chloroplast	unidentified	Pleurocladia	Pleurocladia lacustris	2.7%
9	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Nitrosomonadaceae	Ellin6067	Ellin6067 sp.	2.4%
10	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Nitrosomonadaceae	IS-44	IS-44 sp.	1.9%
11	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Nitrosomonadaceae	MND1	MND1 sp.	1.8%
12	Bacteria	Rokubacteria	NC10	unidentified	unidentified	Rokubacteriales	Rokubacteriales sp.	1.7%
13	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	Pseudomonas	Pseudomonas sp.	1.6%
14	Bacteria	Cyanobacteria	Oxyphotobacteria	Synechococcales	Synechococcales Incertae	Schizothrix	Schizothrix sp.	1.6%
15	Bacteria	Proteobacteria	Gammaproteobacteria	unidentified	unidentified	JG36-GS-52	JG36-GS-52 sp.	1.6%
16	Bacteria	Proteobacteria	Gammaproteobacteria	unidentified	unidentified	PLTA13	PLTA13 sp.	1.5%
17	Bacteria	Proteobacteria	Deltaproteobacteria	unidentified	unidentified	NB1/j	NB1/j sp.	1.4%
18	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	unidentified	SC-I-84	SC-I-84 sp.	1.2%
19	Bacteria	Proteobacteria	Deltaproteobacteria	unidentified	unidentified	MBNT15	MBNT15 sp.	1.2%
20	Bacteria	Actinobacteria	Thermoleophilia	Gaiellales	Gaiellaceae	Gaiella	Gaiella sp.	1.1%
21	Bacteria	Planctomycetes	Planctomycetacia	Pirellulales	Pirellulaceae	Pirellula	Pirellula sp.	1.1%
22	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Rhodobacter	Rhodobacter sp.	1.1%
23	Bacteria	Actinobacteria	unidentified	unidentified	unidentified	MB-A2-108	MB-A2-108 sp.	1.1%
24	Bacteria	Planctomycetes	unidentified	unidentified	unidentified	OM190	OM190 sp.	1.0%
25	Bacteria	Proteobacteria	Gammaproteobacteria	Steroidobacteriales	Woeseiaceae	Woeseia	Woeseia sp.	1.0%

Top 25 taxa identified using Loop Genomics in soil sample location 1 (the species or genus identification is highlighted in green).

V Top 25 taxons in samples from “GOBAM”

V.I Short V4 16S rRNA gene sequencing

Kingdom	Phylum	Class	Order	Family	Genus	Species	% Sequences
						nr. sequences	% Sequences
1	Bacteria	Actinobacteria	Frankiales	Sporichthyaceae	hgcl clade	uncultured Candidatus Planktophilia sp	11.0%
2	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Methylophilaceae	Candidatus Methylophilus	6.0%
3	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Limnhabitans	5.8%
4	Bacteria	Actinobacteria	Frankiales	Sporichthyaceae	hgcl clade	actinobacterium SCGC AAA278-O22	3.9%
5	Bacteria	Cyanobacteria	Oxyphotobacteria	Chloroplast	—	—	3.7%
6	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	—	3.4%
7	Bacteria	Actinobacteria	Micrococcales	Microbacteriaceae	Candidatus Limnoluna	uncultured bacterium	3.3%
8	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Polaromonas	3.3%
9	Bacteria	Cyanobacteria	Oxyphotobacteria	Chloroplast	uncultured Cryptomonadaceae	uncultured Cryptomonadaceae	3.2%
10	Bacteria	Planctomycetes	Planctomycetacia	Planctomycetales	—	—	3.0%
11	Bacteria	Actinobacteria	Frankiales	Sporichthyaceae	hgcl clade	—	3.0%
12	Bacteria	Actinobacteria	Frankiales	Sporichthyaceae	Ambiguous_taxa	uncultured bacterium	2.5%
13	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Polynucleobacter	metagenome	2.4%
14	Bacteria	Bacteroidetes	Bacteroidia	Flavobacteriales	Flavobacteriaceae	Flavobacterium	2.3%
15	Bacteria	Actinobacteria	Frankiales	Sporichthyaceae	hgcl clade	uncultured actinobacterium	2.0%
16	Bacteria	Actinobacteria	Frankiales	Sporichthyaceae	Candidatus Planktophilia	Ambiguous_taxa	2.0%
17	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Polynucleobacter	—	1.9%
18	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Limnhabitans	—	1.7%
19	Bacteria	Actinobacteria	Frankiales	Sporichthyaceae	hgcl clade	Actinobacteria bacterium IMCC26103	1.4%
20	Bacteria	Chloroflexi	SL56 marine group	uncultured bacterium	uncultured bacterium	uncultured bacterium	1.3%
21	Bacteria	Cyanobacteria	Oxyphotobacteria	Chloroplast	Virgulinella fragilis	Virgulinella fragilis	1.2%
22	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Extensimonas	1.1%
23	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Polynucleobacter	1.0%
24	Bacteria	Actinobacteria	Frankiales	Sporichthyaceae	hgcl clade	Ambiguous_taxa	1.0%
25	Bacteria	Bacteroidetes	Bacteroidia	Cytophagales	Spirosomaceae	Pseudarcicella	0.9%

Top 25 taxons identified using short v4 16S rRNA sequencing in Rapid filter effluent (GOBAM) (the species or genus identification is highlighted in green).

Kingdom	Phylum	Class	Order	Family	Genus	Species	% Sequences
						nr. sequences	% Sequences
1	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Hyphomicrobiaceae	Hyphomicrobium	40.1%
2	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	—	8.9%
3	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Gallionellaceae	Gallionella	7.8%
4	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Porphyrabacter	6.6%
5	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Hyphomicrobiaceae	Hyphomicrobium	2.6%
6	Bacteria	Cyanobacteria	Oxyphotobacteria	Chloroplast	—	—	2.1%
7	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Methylophilaceae	Candidatus Methylophilus	2.0%
8	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Beijerinckiaceae	Bosea	1.8%
9	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Rhizobacter	1.6%
10	Bacteria	Actinobacteria	Frankiales	Sporichthyaceae	hgcl clade	uncultured Candidatus Pla	1.5%
11	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Xanthobacteraceae	Afipia	1.3%
12	Bacteria	Chlamydiae	Chlamydiae	Parachlamydiaceae	Neochlamydia	metagenome	1.2%
13	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Gallionellaceae	Candidatus Nitrotoga	0.8%
14	Bacteria	Acidobacteria	Blastocatellia (Subgroup)	Blastocatellales	Blastocatellaceae	Blastocatella	0.6%
15	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Hydrogenophaga	0.6%
16	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	TRA3-20	—	0.5%
17	Bacteria	Proteobacteria	Gammaproteobacteria	Beggiatoales	Beggiatoaceae	uncultured	0.5%
18	Bacteria	Actinobacteria	Frankiales	Sporichthyaceae	hgcl clade	actinobacterium SCGC AA	0.4%
19	Bacteria	Actinobacteria	Micrococcales	Microbacteriaceae	Candidatus Limnoluna	uncultured bacterium	0.4%
20	Bacteria	Actinobacteria	Frankiales	Sporichthyaceae	hgcl clade	—	0.4%
21	Bacteria	Proteobacteria	Gammaproteobacteria	Beggiatoales	Beggiatoaceae	uncultured	0.4%
22	Bacteria	Proteobacteria	Gammaproteobacteria	Piscirickettsiales	Piscirickettsiaceae	Candidatus Endoecksteinascidia	0.4%
23	Bacteria	Patescibacteria	Saccharimonadia	uncultured bacterium	uncultured bacterium	uncultured bacterium	0.3%
24	Bacteria	Cyanobacteria	Oxyphotobacteria	Chloroplast	uncultured Cryptomonadaceae	uncultured Cryptomonadaceae	0.3%
25	Bacteria	Proteobacteria	Deltaproteobacteria	SAR324 clade(Marine grc	uncultured soil bacterium	uncultured soil bacterium	0.3%

Top 25 taxons identified using short v4 16S rRNA sequencing in sample “GOBAM effluent” (the species or genus identification is highlighted in green).

Kingdom	Phylum	Class	Order	Family	Genus	Species	% Sequences
Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	—	—	8,7%
Bacteria	Actinobacteria	Actinobacteria	Frankiales	Sporichthyaceae	hgcl clade	uncultured Candidatus Planktophila sp.	8,6%
Bacteria	Bacteroidetes	Bacteroidia	Flavobacteriales	Flavobacteriaceae	Flavobacterium	—	4,8%
Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Limnohabitans	uncultured Limnohabitans sp.	4,6%
Bacteria	Actinobacteria	Actinobacteria	Frankiales	Sporichthyaceae	hgcl clade	actinobacterium SCGC AAA278-022	3,6%
Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Gallionellaceae	Gallionella	—	3,4%
Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Porphyrobacter	Ambiguous_taxa	3,4%
Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Methylophilaceae	Candidatus Methylophilum	uncultured bacterium	3,1%
Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Polaromonas	—	2,8%
Bacteria	Actinobacteria	Actinobacteria	Frankiales	Sporichthyaceae	hgcl clade	—	2,7%
Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Microbacteriaceae	Candidatus Limnoluna	uncultured bacterium	2,5%
Bacteria	Actinobacteria	Actinobacteria	Frankiales	Sporichthyaceae	Ambiguous_taxa	uncultured bacterium	2,2%
Bacteria	Actinobacteria	Alphaproteobacteria	Rhizobiales	Hyphomicrobiaceae	Hyphomicrobium	—	1,7%
Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Polynucleobacter	Ambiguous_taxa	1,7%
Bacteria	Cyanobacteria	Oxyphotobacteria	Chloroplast	—	—	—	1,6%
Bacteria	Cyanobacteria	Oxyphotobacteria	Chloroplast	uncultured Cryptomonadaceae	uncultured Cryptomonadaceae	uncultured Cryptomonadaceae	1,6%
Bacteria	Actinobacteria	Actinobacteria	Frankiales	Sporichthyaceae	Candidatus Planktophila	Ambiguous_taxa	1,5%
Bacteria	Proteobacteria	Gammaproteobacteria	Burkholderiaceae	Burkholderiaceae	Polynucleobacter	—	1,5%
Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Polynucleobacter	metagenome	1,4%
Bacteria	Bacteroidetes	Bacteroidia	Cytophagales	Spirosomaceae	Pseudarcicella	—	1,4%
Bacteria	Bacteroidetes	Bacteroidia	Flavobacteriales	Flavobacteriaceae	Flavobacterium	Bacteroidetes bacterium SCGC AAA027-N21	1,3%
Bacteria	Actinobacteria	Actinobacteria	Frankiales	Sporichthyaceae	hgcl clade	uncultured actinobacterium	1,3%
Bacteria	Planctomycetes	Planctomycetacia	Planctomycetales	uncultured	—	—	1,3%
Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Hydrogenophaga	—	1,2%
Bacteria	Actinobacteria	Actinobacteria	Frankiales	Sporichthyaceae	hgcl clade	Actinobacteria bacterium IMCC26103	1,2%

Top 25 taxa identified using short v4 16S rRNA sequencing in sample “Infiltration water Berkeide” (GOBAM).

Kingdom	Phylum	Class	Order	Family	Genus	Species	% Sequences
1	Bacteria	Actinobacteria	Actinobacteria	Frankiales	Sporichthyaceae	hgcl clade	nr. sequences 100.0%
2	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	uncultured Candidatus Planktophila sp.	9.8%
3	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Methylophilaceae	Candidatus Methylophilum	8.9%
4	Bacteria	Actinobacteria	Actinobacteria	Frankiales	Sporichthyaceae	hgcl clade	3.9%
5	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Limnohabitans	3.9%
6	Bacteria	Cyanobacteria	Oxyphotobacteria	Chloroplast	—	uncultured Limnohabitans sp.	3.7%
7	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Porphyrobacter	3.4%
8	Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Microbacteriaceae	Candidatus Limnoluna	3.0%
9	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Gallionellaceae	Gallionella	2.9%
10	Bacteria	Actinobacteria	Actinobacteria	Frankiales	Sporichthyaceae	hgcl clade	2.9%
11	Bacteria	Bacteroidetes	Bacteroidia	Flavobacteriales	Flavobacteriaceae	Flavobacterium	2.8%
12	Bacteria	Actinobacteria	Actinobacteria	Frankiales	Sporichthyaceae	Ambiguous_taxa	2.4%
13	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Polaromonas	2.3%
14	Bacteria	Actinobacteria	Actinobacteria	Frankiales	Sporichthyaceae	Candidatus Planktophila	1.8%
15	Bacteria	Cyanobacteria	Oxyphotobacteria	Chloroplast	uncultured Cryptomonadaceae	uncultured Cryptomonadaceae	1.7%
16	Bacteria	Actinobacteria	Actinobacteria	Frankiales	Sporichthyaceae	hgcl clade	1.7%
17	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Polynucleobacter	1.6%
18	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Polynucleobacter	1.6%
19	Bacteria	Planctomycetes	Planctomycetacia	Planctomycetales	—	—	1.5%
20	Bacteria	Actinobacteria	Actinobacteria	Frankiales	Sporichthyaceae	hgcl clade	1.3%
21	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Hyphomicrobiaceae	Hyphomicrobium	1.3%
22	Bacteria	Chloroflexi	SL56 marine group	uncultured bacterium	uncultured bacterium	uncultured bacterium	1.0%
23	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Hydrogenophaga	1.0%
24	Bacteria	Cyanobacteria	Oxyphotobacteria	Chloroplast	Virgulinella fragilis	Virgulinella fragilis	0.9%
25	Bacteria	Actinobacteria	Actinobacteria	Frankiales	Sporichthyaceae	hgcl clade	Ambiguous_taxa 0.9%

Top 25 taxa identified using short v4 16S rRNA sequencing in sample “Short residence basin 37” (GOBAM) (the species or genus identification is highlighted in green).

Kingdom	Phylum	Class	Order	Family	Genus	Species	% Sequences
1	Bacteria	Cyanobacteria	Oxyphotobacteria	Chloroplast	Virgulinella fragilis	Virgulinella fragilis	22.1%
2	Bacteria	Bacteroidetes	Bacteroidia	Flavobacteriales	Flavobacteriaceae	Flavobacterium	8.8%
3	Bacteria	Bacteroidetes	Bacteroidia	Flavobacteriales	Crociniomicaceae	Bacteroidetes bacterium SCGC AAA027-N21	7.0%
4	Bacteria	Bacteroidetes	Bacteroidia	Flavobacteriales	Flavobacteriaceae	Fluviicola	5.5%
5	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Limnohabitans	4.4%
6	Bacteria	Actinobacteria	Actinobacteria	Frankiales	Sporichthyaceae	hgcl clade	4.0%
7	Bacteria	Bacteroidetes	Bacteroidia	Cytophagales	Spirosomaceae	Pseudarcicella	3.4%
8	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Sphingorhabdus	2.9%
9	Bacteria	Actinobacteria	Actinobacteria	Frankiales	Sporichthyaceae	hgcl clade	2.8%
10	Bacteria	Cyanobacteria	Oxyphotobacteria	Chloroplast	uncultured Cryptomonadaceae	uncultured Cryptomonadaceae	2.3%
11	Bacteria	Cyanobacteria	Oxyphotobacteria	Chloroplast	metagenome	metagenome	2.2%
12	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Methylophilaceae	Methylophilum	2.1%
13	Bacteria	Actinobacteria	Actinobacteria	Frankiales	Sporichthyaceae	Ambiguous_taxa	2.0%
14	Bacteria	Proteobacteria	Gammaproteobacteria	Burkholderiaceae	Burkholderiaceae	GKS98 freshwater group	1.9%
15	Bacteria	Bacteroidetes	Bacteroidia	Chitinophagales	Chitinophagaceae	Sediminibacterium	1.7%
16	Bacteria	Cyanobacteria	Oxyphotobacteria	Chloroplast	—	—	1.6%
17	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Polaromonas	1.6%
18	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	—	1.6%
19	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Polynucleobacter	1.4%
20	Bacteria	Actinobacteria	Actinobacteria	Frankiales	Sporichthyaceae	hgcl clade	1.1%
21	Bacteria	Bacteroidetes	Bacteroidia	Flavobacteriales	Flavobacteriaceae	Flavobacterium	1.0%
22	Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Microbacteriaceae	Candidatus Limnoluna	0.9%
23	Bacteria	Bacteroidetes	Bacteroidia	Flavobacteriales	Flavobacteriaceae	Flavobacterium	0.9%
24	Bacteria	Cyanobacteria	Oxyphotobacteria	Chloroplast	uncultured bacterium	uncultured bacterium	0.8%
25	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Limnohabitans	0.8%

Top 25 taxa identified using short v4 16S rRNA sequencing in sample “Long residence basin 25-9” (GOBAM) (the species or genus identification is highlighted in green).

Kingdom	Phylum	Class	Order	Family	Genus	Species	% Sequences
						nr. sequences	100.0%
1	Bacteria	Proteobacteria	Gammaproteobacteria	Methylococcales	Methylomonaceae	pLW-20	11.9%
2	Bacteria	Proteobacteria	Gammaproteobacteria	Methylococcales	Methylomonaceae	Crenothrix	9.5%
3	Bacteria	Proteobacteria	Gammaproteobacteria	Methylococcales	Methylomonaceae	Crenothrix	5.4%
4	Bacteria	Proteobacteria	Gammaproteobacteria	Methylococcales	Methylomonaceae	Crenothrix	4.6%
5	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Methylophilaceae	Methylothera	3.6%
6	Bacteria	Omnitrophicae	uncultured bacterium	uncultured bacterium	uncultured bacterium	uncultured bacterium	2.4%
7	Bacteria	Proteobacteria	Gammaproteobacteria	Methylococcales	Methylomonaceae	pLW-20	2.3%
8	Bacteria	Proteobacteria	Gammaproteobacteria	Methylococcales	Methylomonaceae		2.2%
9	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Limnhabitans	1.6%
10	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Rhodoferrax	1.5%
11	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Methylophilaceae	Methylothera	1.2%
12	Bacteria	Nitrospirae	Nitrospira	Nitrospirales	Nitrospiraceae	Nitrospira	1.2%
13	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Gallionellaceae	Sideroxydans	1.2%
14	Bacteria	Acidobacteria	Subgroup 6				1.2%
15	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Gallionellaceae	Ferriphaselus	1.1%
16	Archaea	Thaumarchaeota	Nitrososphaeria	Nitrosopumilales	Nitrosopumilaceae	uncultured archaeon	1.0%
17	Bacteria	Rokubacteria	NC10	Rokubacteriales	uncultured bacterium	uncultured bacterium	1.0%
18	Bacteria						1.0%
19	Bacteria	Omnitrophicae	Omnitrophia	Omnitrophales	Omnitrophaceae	Candidatus Omnitrophus	1.0%
20	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Gallionellaceae	Gallionella	0.9%
21	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Gallionellaceae	Gallionella	0.8%
22	Bacteria	Proteobacteria	Gammaproteobacteria	Methylococcales	Methylomonaceae	Methylobacter	0.8%
23	Bacteria	Omnitrophicae	Omnitrophia	Omnitrophales	Omnitrophaceae	Candidatus Omnitrophus	0.8%
24	Bacteria	Proteobacteria	Gammaproteobacteria	Methylococcales	Methylomonaceae	uncultured	0.7%
25	Archaea	Thaumarchaeota	Nitrososphaeria	Nitrosopumilales	Nitrosopumilaceae	Candidatus Nitrosoarchaeum	0.7%

Top 25 taxons identified using short v4 16S rRNA sequencing in sample "Collected water after dune passage" (GOBAM) (the species or genus identification is highlighted in green).

Kingdom	Phylum	Class	Order	Family	Genus	Species	% Sequences
						nr. sequences	100.0%
1	Bacteria	Omnitrophicae	uncultured bacterium	uncultured bacterium	uncultured bacterium	uncultured bacterium	9.7%
2	Bacteria	Acidobacteria	Subgroup 6				4.8%
3	Bacteria	Omnitrophicae	Omnitrophia	Omnitrophales	Omnitrophaceae	Candidatus Omnitrophus	4.4%
4	Bacteria	Patescibacteria	Gracilbacteria	Candidatus Peregrinibact	uncultured bacterium	uncultured bacterium	2.9%
5	Bacteria						2.7%
6	Bacteria	Rokubacteria	NC10	Rokubacteriales	uncultured bacterium	uncultured bacterium	2.6%
7	Bacteria	Omnitrophicae	Omnitrophia	Omnitrophales	Omnitrophaceae	Candidatus Omnitrophus	2.1%
8	Bacteria	Rokubacteria	NC10	Rokubacteriales	Ambiguous_taxa	Ambiguous_taxa	1.9%
9	Bacteria	Patescibacteria	Gracilbacteria	Candidatus Peribacteria	uncultured bacterium	uncultured bacterium	1.9%
10	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Hyphomicrobiaceae		1.7%
11	Bacteria	Planctomycetes	Brocadia	Brocadiales	Brocadia	Planctomycetes bacterium RIFXYD12_FULL	1.6%
12	Bacteria	Rokubacteria	NC10	Methylomirabilales	Methylomirabilaceae	Sh765B-TzT-35	1.6%
13	Archaea	Thaumarchaeota	Nitrososphaeria	Nitrosopumilales	Nitrosopumilaceae	uncultured archaeon	1.6%
14	Archaea	Nanoarchaeota	Woesearchaeia	uncultured euryarchaeot	uncultured euryarchaeote	uncultured euryarchaeote	1.5%
15	Archaea	Thaumarchaeota	Nitrososphaeria	Nitrosopumilales	Nitrosopumilaceae	Candidatus Nitrosoarchaeum	1.4%
16	Archaea	Nanoarchaeota	Woesearchaeia				1.4%
17	Bacteria	Nitrospirae	Thermodesulfobionria	uncultured	uncultured bacterium	uncultured bacterium	1.1%
18	Bacteria	Planctomycetes	Planctomycetacia	Gemmatales	Gemmataceae	uncultured	1.0%
19	Bacteria	Chloroflexi	Dehalococcoidia	SAR202 clade	uncultured bacterium	uncultured bacterium	1.0%
20	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodospirillales	uncultured		0.9%
21	Bacteria	Omnitrophicae	uncultured Firmicutes ba	uncultured Firmicutes ba	uncultured Firmicutes bacterium	uncultured Firmicutes bacterium	0.9%
22	Bacteria	Omnitrophicae	Omnitrophia	Omnitrophales	Omnitrophaceae	Candidatus Omnitrophus	0.8%
23	Bacteria	Nitrospirae	4-29-1	uncultured bacterium	uncultured bacterium	uncultured bacterium	0.8%
24	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	TRA3-20		0.8%
25	Bacteria	Cyanobacteria	Oxyphotobacteria	Chloroplast			0.8%

Top 25 taxons identified using short v4 16S rRNA sequencing in sample "Drinking water" (GOBAM) (the species or genus identification is highlighted in green)).

V.II Loop genomics sequencing

	kingdom	phylum	class	order	family	genus	species	% Sequences
1	Bacteria	Proteobacteria	Alphaproteobacteria	SAR11 clade	unidentified	Clade	Clade sp.	13.8%
2	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Limnohabitans	Limnohabitans sp.	13.1%
3	Bacteria	Actinobacteria	Actinobacteria	Frankiales	Sporichthyaceae	hgcl	hgcl sp.	12.2%
4	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Methylophilaceae	Methylophilum	Methylophilum sp.	6.0%
5	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Polynucleobacter	Polynucleobacter sp.	5.5%
6	Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Microbacteriaceae	Limnoluna	Limnoluna sp.	5.4%
7	Bacteria	Actinobacteria	Actinobacteria	Frankiales	Sporichthyaceae	Planktophila	Planktophila sp.	5.1%
8	Bacteria	Cyanobacteria	Oxyphotobacteria	unidentified	unidentified	Chloroplast	Chloroplast sp.	5.0%
9	Bacteria	Bacteroidetes	Bacteroidia	Flavobacteriales	Flavobacteriaceae	Flavobacterium	Flavobacterium sp.	3.5%
10	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Polaromonas	Polaromonas sp.	3.5%
11	Bacteria	Chloroflexi	unidentified	unidentified	unidentified	SL56	SL56 sp.	2.5%
12	Bacteria	Actinobacteria	Acidimicrobiia	Microtrichales	Ilumatobacteraceae	CL500-29	CL500-29 sp.	2.3%
13	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Noviherbaspirillum	Noviherbaspirillum sp.	1.5%
14	Bacteria	Bacteroidetes	Bacteroidia	Flavobacteriales	Crocinitomicaceae	Fluviicola	Fluviicola sp.	1.2%
15	Bacteria	Patescibacteria	Parcubacteria	unidentified	unidentified	Kaiserbacteria	Kaiserbacteria sp.	1.1%
16	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Hydrogenophaga	Hydrogenophaga sp.	0.9%
17	Bacteria	Bacteroidetes	Bacteroidia	Cytophagales	Spirosomaceae	Pseudarcicella	Pseudarcicella sp.	0.9%
18	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Ramlibacter	Ramlibacter sp.	0.9%
19	Bacteria	Bacteroidetes	Bacteroidia	Chitinophagales	Chitinophagaceae	Sediminibacterium	Sediminibacterium sp.	0.8%
20	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Rhodoferrax	Rhodoferrax sp.	0.7%
21	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Undibacterium	Undibacterium sp.	0.7%
22	Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Microbacteriaceae	Rhodoluna	Rhodoluna sp.	0.7%
23	Bacteria	Bacteroidetes	Bacteroidia	Sphingobacteriales	unidentified	NS11-12	NS11-12 sp.	0.5%
24	Bacteria	Proteobacteria	Alphaproteobacteria	Rickettsiales	unidentified	Mitochondria	Mitochondria sp.	0.5%
25	Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Microbacteriaceae	Planktoluna	Planktoluna sp.	0.5%

Top 25 taxa identified using Loop Genomics sequencing in Rapid filter effluent (GOBAM) (the species or genus identification is highlighted in green).

	kingdom	phylum	class	order	family	genus	species	% Sequences
1	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Hyphomicrobiaceae	Hyphomicrobium	Hyphomicrobium sp.	47.3%
2	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Gallionellaceae	Gallionella	Gallionella sp.	9.3%
3	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Rhizorhapis	Rhizorhapis sp.	7.0%
4	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Porphyrobacter	Porphyrobacter sp.	6.7%
5	Bacteria	Patescibacteria	Parcubacteria	unidentified	unidentified	Kaiserbacteria	Kaiserbacteria sp.	3.5%
6	Bacteria	Proteobacteria	Alphaproteobacteria	SAR11 clade	unidentified	Clade	Clade sp.	1.7%
7	Bacteria	Actinobacteria	Actinobacteria	Frankiales	Sporichthyaceae	hgcl	hgcl sp.	1.7%
8	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Xanthobacteraceae	Bradyrhizobium	Bradyrhizobium sp.	1.7%
9	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Rhizobacter	Rhizobacter sp.	1.5%
10	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Hydrogenophaga	Hydrogenophaga sp.	1.2%
11	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Beijerinckiaceae	Bosea	Bosea vaviloviae	1.2%
12	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Beijerinckiaceae	Bosea	Bosea sp.	1.1%
13	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	unidentified	TRA3-20	TRA3-20 sp.	1.0%
14	Bacteria	Chloroflexi	unidentified	unidentified	unidentified	SL56	SL56 sp.	0.7%
15	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Gallionellaceae	Nitrotoga	Nitrotoga sp.	0.7%
16	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Sphingopyxis	Sphingopyxis sp.	0.7%
17	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Limnohabitans	Limnohabitans sp.	0.6%
18	Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Microbacteriaceae	Limnoluna	Limnoluna sp.	0.6%
19	Bacteria	Actinobacteria	Actinobacteria	Frankiales	Sporichthyaceae	Planktophila	Planktophila sp.	0.6%
20	Bacteria	Bacteroidetes	Bacteroidia	Flavobacteriales	Flavobacteriaceae	Flavobacterium	Flavobacterium sp.	0.6%
21	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Leptothrix	Leptothrix sp.	0.5%
22	Bacteria	Bacteroidetes	Bacteroidia	Chitinophagales	Chitinophagaceae	Sediminibacterium	Sediminibacterium sp.	0.5%
23	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Beijerinckiaceae	alpha1	alpha1 sp.	0.5%
24	Bacteria	Patescibacteria	Gracilibacteria	unidentified	unidentified	Peribacteria	Peribacteria sp.	0.5%
25	Bacteria	Cyanobacteria	Oxyphotobacteria	unidentified	unidentified	Chloroplast	Chloroplast sp.	0.4%

Top 25 taxa identified using Loop Genomics sequencing in sample "GOBAM effluent" (the species or genus identification is highlighted in green).

green)

	kingdom	phylum	class	order	family	genus	species	% Sequences
1	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Limnohabitans	Limnohabitans sp.	14.7%
2	Bacteria	Proteobacteria	Alphaproteobacteria	SAR11 clade	unidentified	Clade	Clade sp.	10.8%
3	Bacteria	Actinobacteria	Actinobacteria	Frankiales	Sporichthyaceae	hgcl	hgcl sp.	10.4%
4	Bacteria	Bacteroidetes	Bacteroidia	Flavobacteriales	Flavobacteriaceae	Flavobacterium	Flavobacterium sp.	8.9%
5	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Polynucleobacter	Polynucleobacter sp.	6.8%
6	Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Microbacteriaceae	Limnoluna	Limnoluna sp.	4.3%
7	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Polaromonas	Polaromonas sp.	3.6%
8	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Methylophilaceae	Methylopumilus	Methylopumilus sp.	3.0%
9	Bacteria	Actinobacteria	Actinobacteria	Frankiales	Sporichthyaceae	Planktophila	Planktophila sp.	2.8%
10	Bacteria	Cyanobacteria	Oxyphotobacteria	unidentified	unidentified	Chloroplast	Chloroplast sp.	2.6%
11	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Porphyrobacter	Porphyrobacter sp.	2.5%
12	Bacteria	Chloroflexi	unidentified	unidentified	unidentified	SL56	SL56 sp.	2.3%
13	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Gallionellaceae	Gallionella	Gallionella sp.	2.1%
14	Bacteria	Bacteroidetes	Bacteroidia	Chitinophagales	Chitinophagaceae	Sediminibacterium	Sediminibacterium sp.	1.9%
15	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Hyphomicrobiaceae	Hyphomicrobium	Hyphomicrobium sp.	1.9%
16	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Rhodoferrax	Rhodoferrax sp.	1.5%
17	Bacteria	Bacteroidetes	Bacteroidia	Flavobacteriales	Crocinitomicaceae	Fluviicola	Fluviicola sp.	1.5%
18	Bacteria	Actinobacteria	Acidimicrobiia	Microtrichales	Ilumatobacteraceae	CL500-29	CL500-29 sp.	0.9%
19	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Methylophilaceae	Methylotenera	Methylotenera sp.	0.8%
20	Bacteria	Patescibacteria	Parcubacteria	unidentified	unidentified	Kaiserbacteria	Kaiserbacteria sp.	0.8%
21	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Noviherbaspirillum	Noviherbaspirillum sp.	0.8%
22	Bacteria	Bacteroidetes	Bacteroidia	Cytophagales	Spirosomaceae	Pseudarcicella	Pseudarcicella sp.	0.8%
23	Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Microbacteriaceae	Rhodoluna	Rhodoluna sp.	0.8%
24	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Beijerinckiacae	alpha1	alpha1 sp.	0.8%
25	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	unidentified	TRA3-20	TRA3-20 sp.	0.6%

Top 25 taxa identified using Loop Genomics sequencing in sample "Infiltration water Berkheide" (GOBAM) (the species or genus identification is highlighted in green).

	kingdom	phylum	class	order	family	genus	species	% sequences
1	Bacteria	Actinobacteria	Actinobacteria	Frankiales	Sporichthyaceae	hgcl	hgcl sp.	13.6%
2	Bacteria	Proteobacteria	Alphaproteobacteria	SAR11 clade	unidentified	Clade	Clade sp.	13.1%
3	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Limnohabitans	Limnohabitans sp.	12.2%
4	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Polynucleobacter	Polynucleobacter sp.	4.8%
5	Bacteria	Cyanobacteria	Oxyphotobacteria	unidentified	unidentified	Chloroplast	Chloroplast sp.	4.1%
6	Bacteria	Bacteroidetes	Bacteroidia	Flavobacteriales	Flavobacteriaceae	Flavobacterium	Flavobacterium sp.	4.1%
7	Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Microbacteriaceae	Limnoluna	Limnoluna sp.	4.0%
8	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Methylophilaceae	Methylopumilus	Methylopumilus sp.	4.0%
9	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Rhodoferrax	Rhodoferrax sp.	3.5%
10	Bacteria	Actinobacteria	Actinobacteria	Frankiales	Sporichthyaceae	Planktophila	Planktophila sp.	3.1%
11	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Gallionellaceae	Gallionella	Gallionella sp.	2.7%
12	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Porphyrobacter	Porphyrobacter sp.	2.6%
13	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Polaromonas	Polaromonas sp.	2.2%
14	Bacteria	Chloroflexi	unidentified	unidentified	unidentified	SL56	SL56 sp.	1.7%
15	Bacteria	Bacteroidetes	Bacteroidia	Flavobacteriales	Crocinitomicaceae	Fluviicola	Fluviicola sp.	1.4%
16	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Hyphomicrobiaceae	Hyphomicrobium	Hyphomicrobium sp.	1.2%
17	Bacteria	Bacteroidetes	Bacteroidia	Cytophagales	Spirosomaceae	Pseudarcicella	Pseudarcicella sp.	1.1%
18	Bacteria	Bacteroidetes	Bacteroidia	Chitinophagales	Chitinophagaceae	Sediminibacterium	Sediminibacterium sp.	1.1%
19	Bacteria	Actinobacteria	Acidimicrobiia	Microtrichales	Ilumatobacteraceae	CL500-29	CL500-29 sp.	1.1%
20	Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Microbacteriaceae	Rhodoluna	Rhodoluna sp.	1.0%
21	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Hydrogenophaga	Hydrogenophaga sp.	0.9%
22	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Noviherbaspirillum	Noviherbaspirillum sp.	0.8%
23	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Beijerinckiacae	alpha1	alpha1 sp.	0.6%
24	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	unidentified	TRA3-20	TRA3-20 sp.	0.6%
25	Bacteria	Proteobacteria	Alphaproteobacteria	Acetobacteriales	Acetobacteraceae	Roseococcus	Roseococcus sp.	0.5%

Top 25 taxa identified Loop Genomics sequencing in sample "Short residence basin 37" (GOBAM) (the species or genus identification is highlighted in green)).

	kingdom	phylum	class	order	family	genus	species	% Sequences
1	Bacteria	Cyanobacteria	Oxyphotobacteria	unidentified	unidentified	Chloroplast	Chloroplast sp.	22.2%
2	Bacteria	Bacteroidetes	Bacteroidia	Flavobacteriales	Flavobacteriaceae	Flavobacterium	Flavobacterium sp.	19.2%
3	Bacteria	Actinobacteria	Actinobacteria	Frankiales	Sporichthyaceae	hgcl	hgcl sp.	9.8%
4	Bacteria	Bacteroidetes	Bacteroidia	Flavobacteriales	Crocinitomicaceae	Fluviicola	Fluviicola sp.	7.7%
5	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Limnohabitans	Limnohabitans sp.	7.0%
6	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Sphingorhabdus	Sphingorhabdus sp.	4.1%
7	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Polynucleobacter	Polynucleobacter sp.	3.2%
8	Bacteria	Bacteroidetes	Bacteroidia	Cytophagales	Spirosomaceae	Pseudarcicella	Pseudarcicella sp.	2.7%
9	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Methylophilaceae	Methylophilaceae	Methylophilaceae sp.	2.4%
10	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	GKS98	GKS98 sp.	2.2%
11	Bacteria	Bacteroidetes	Bacteroidia	Chitinophagales	Chitinophagaceae	Sediminibacterium	Sediminibacterium sp.	1.9%
12	Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Microbacteriaceae	Rhodoluna	Rhodoluna sp.	1.9%
13	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Polaromonas	Polaromonas sp.	1.6%
14	Bacteria	Proteobacteria	Alphaproteobacteria	SAR11 clade	unidentified	Clade	Clade sp.	1.5%
15	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Limnohabitans	Limnohabitans planktonicus	1.4%
16	Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Microbacteriaceae	Limnoluna	Limnoluna sp.	1.1%
17	Bacteria	Actinobacteria	Actinobacteria	Frankiales	Sporichthyaceae	Planktophila	Planktophila sp.	0.9%
18	Bacteria	Bacteroidetes	Bacteroidia	Sphingobacteriales	unidentified	NS11-12	NS11-12 sp.	0.6%
19	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Methylophilaceae	Methylopumilus	Methylopumilus sp.	0.5%
20	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Hydrogenophaga	Hydrogenophaga sp.	0.4%
21	Bacteria	Bacteroidetes	Bacteroidia	Cytophagales	Cyclobacteriaceae	Algoriphagus	Algoriphagus sp.	0.4%
22	Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Microbacteriaceae	Planktoluna	Planktoluna sp.	0.4%
23	Bacteria	Bacteroidetes	Bacteroidia	Sphingobacteriales	unidentified	env.OP57	env.OP57 sp.	0.4%
24	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Methylophilaceae	Methylopumilus	Methylopumilus turicensis	0.4%
25	Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Microbacteriaceae	Aurantimicrobium	Aurantimicrobium sp.	0.3%

Top 25 taxons identified using Loop Genomics sequencing in sample “Long residence basin 25-9” (GOBAM) (the highest identified taxonomic level is accentuated in green).

	kingdom	phylum	class	order	family	genus	species	relabund
1	Bacteria	Proteobacteria	Gammaproteobacteria	Methylococcales	Methylomonaceae	pLW-20	pLW-20 sp.	14.7%
2	Bacteria	Proteobacteria	Gammaproteobacteria	Methylococcales	Methylomonaceae	Crenothrix	Crenothrix polyspora	13.9%
3	Bacteria	Proteobacteria	Gammaproteobacteria	Methylococcales	Methylomonaceae	Crenothrix	Crenothrix sp.	10.5%
4	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Methylophilaceae	Methylophilaceae	Methylophilaceae sp.	6.1%
5	Bacteria	Nitrospirae	Nitrospira	Nitrospirales	Nitrospiraceae	Nitrospira	Nitrospira sp.	4.0%
6	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Rhodoferrax	Rhodoferrax sp.	3.7%
7	Bacteria	Acidobacteria	unidentified	unidentified	unidentified	Subgroup	Subgroup sp.	2.7%
8	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Gallionellaceae	Gallionella	Gallionella sp.	2.7%
9	Bacteria	unidentified	unidentified	unidentified	unidentified	Omnitrophicaeota	Omnitrophicaeota sp.	1.7%
10	Bacteria	Patescibacteria	Parcubacteria	unidentified	unidentified	UBA9983	UBA9983 sp.	1.5%
11	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Gallionellaceae	Sideroxydans	Sideroxydans sp.	1.5%
12	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Paucibacter	Paucibacter sp.	1.4%
13	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Nitrosomonadaceae	Nitrosomonas	Nitrosomonas sp.	1.1%
14	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Gallionellaceae	Ferriphaselus	Ferriphaselus amnicola	1.1%
15	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Hyphomicrobiaceae	Pedomicrobium	Pedomicrobium sp.	0.8%
16	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Nitrosomonadaceae	MND1	MND1 sp.	0.8%
17	Bacteria	Patescibacteria	Parcubacteria	unidentified	unidentified	Kaiserbacteria	Kaiserbacteria sp.	0.8%
18	Bacteria	Bacteroidetes	Ignavibacteria	Kryptoniales	unidentified	BSV26	BSV26 sp.	0.8%
19	Bacteria	Rokubacteria	NC10	Methylomirabilales	Methylomirabilaceae	Methylomirabilis	Methylomirabilis sp.	0.8%
20	Bacteria	Bacteroidetes	Bacteroidia	Chitinophagales	Chitinophagaceae	Sediminibacterium	Sediminibacterium sp.	0.8%
21	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Methylophilaceae	Methylophilaceae	Methylophilaceae sp.	0.8%
22	Bacteria	Rokubacteria	NC10	unidentified	unidentified	Rokubacteriales	Rokubacteriales sp.	0.6%
23	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	unidentified	TRA3-20	TRA3-20 sp.	0.6%
24	Bacteria	Rokubacteria	NC10	Methylomirabilales	Methylomirabilaceae	Sh765B-TzT-35	Sh765B-TzT-35 sp.	0.6%
25	Bacteria	Proteobacteria	Deltaproteobacteria	unidentified	unidentified	Sva0485	Sva0485 sp.	0.6%

Top 25 taxons identified using Loop Genomics sequencing in sample “Collected water after dune passage” (GOBAM) (the highest identified taxonomic level is accentuated in green).

	kingdom	phylum	class	order	family	genus	species	% Sequences
1	Bacteria	unidentified	unidentified	unidentified	unidentified	Omnitrophicaeota	Omnitrophicaeota sp.	10.8%
2	Bacteria	Acidobacteria	unidentified	unidentified	unidentified	Subgroup	Subgroup sp.	9.9%
3	Bacteria	Patescibacteria	Gracilibacteria	unidentified	unidentified	Peribacteria	Peribacteria sp.	6.7%
4	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Hyphomicrobiaceae	Pedomicrobium	Pedomicrobium sp.	5.1%
5	Bacteria	Rokubacteria	NC10	unidentified	unidentified	Rokubacteriales	Rokubacteriales sp.	4.5%
6	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	unidentified	TRA3-20	TRA3-20 sp.	3.2%
7	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Nitrosomonadaceae	MND1	MND1 sp.	3.2%
8	Bacteria	Proteobacteria	Gammaproteobacteria	unidentified	unidentified	PLTA13	PLTA13 sp.	2.6%
9	Bacteria	Patescibacteria	Parcubacteria	unidentified	unidentified	UBA9983	UBA9983 sp.	2.2%
10	Bacteria	Cyanobacteria	Oxyphotobacteria	unidentified	unidentified	Chloroplast	Chloroplast sp.	2.0%
11	Bacteria	Nitrospirae	Nitrospira	Nitrospirales	Nitrospiraceae	Nitrospira	Nitrospira sp.	2.0%
12	Bacteria	Patescibacteria	Gracilibacteria	unidentified	unidentified	JGI000069/P22	JGI000069/P22 sp.	2.0%
13	Bacteria	Omnitrophicaeot	Omnitrophia	Omnitrophales	Omnitrophaceae	Omnitrophus	Omnitrophus sp.	1.8%
14	Bacteria	Chloroflexi	Dehalococcoidia	unidentified	unidentified	S085	S085 sp.	1.8%
15	Bacteria	Proteobacteria	Gammaproteobacteria	Acidiferrobacterales	Acidiferrobacteraceae	Sulfurifustis	Sulfurifustis sp.	1.8%
16	Bacteria	Rokubacteria	NC10	Methyloirabilales	Methyloirabilaceae	Sh765B-TzT-35	Sh765B-TzT-35 sp.	1.8%
17	Bacteria	Patescibacteria	Parcubacteria	unidentified	unidentified	Kaiserbacteria	Kaiserbacteria sp.	1.6%
18	Bacteria	Patescibacteria	ABY1	unidentified	unidentified	Falkowbacteria	Falkowbacteria sp.	1.6%
19	Bacteria	Proteobacteria	Deltaproteobacteria	Bdellovibrionales	Bdellovibrionaceae	Bdellovibrio	Bdellovibrio sp.	1.6%
20	Bacteria	Proteobacteria	Deltaproteobacteria	Myxococcales	unidentified	bacteriap25	bacteriap25 sp.	1.6%
21	Bacteria	Proteobacteria	Alphaproteobacteria	SAR11 clade	unidentified	Clade	Clade sp.	1.4%
22	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Gallionellaceae	Gallionella	Gallionella sp.	1.2%
23	Bacteria	Proteobacteria	Deltaproteobacteria	unidentified	unidentified	DTB120	DTB120 sp.	1.2%
24	Bacteria	unidentified	unidentified	unidentified	unidentified	Latescibacteria	Latescibacteria sp.	1.2%
25	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Hyphomicrobiaceae	Hyphomicrobium	Hyphomicrobium sp.	1.0%

Top 25 taxons identified using Loop Genomics sequencing in sample "Drinking water" (GOBAM) (the highest identified taxonomic level is accentuated in green).