Who's there? Community analysis by amplicon sequencing

3.4.2017

Anu Mikkonen anu.mikkonen@jyu.fi

JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

Content

- What is microbiome analysis by amplicon sequencing
- How many sequences do you need
- Effect of PCR conditions
- Effect of sequencing platforms
- Analysis platforms and known issues

JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

Take-home message

1. Choose one of the commonly used pipelines and follow it – unless you can fully justify exceptions
   - both *in vitro* and *in silico*
2. Process all your samples in EXACTLY the same way
3. Ion Torrent PGM is a good and more economic alternative for Illumina MiSeq
4. High-throughput sequence data is never perfect

JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

Microbial community analysis – Microbiome analysis

- NOT metagenomics – if you use spesific primers, its not (and shouldn't be called) metagenomics

- Bacterial communities
  - 16S rRNA gene or 16S rRNA
- Bacterial and archaeal communities
  - 16S rRNA gene or 16S rRNA
- Fungal communities
  - ITS1 region (internal transcribed spacer between 18S and 5.8S rRNA genes)

Amplicon sequencing can serve lots of other purposes too!

MOLECULAR ECOLOGY

Molecular Ecology (2012) 21, 3647–3655 doi: 10.1111/j.1365-294X.2012.05545.x

FROM THE COVER

DNA from soil mirrors plant taxonomic and growth form diversity

N. G. YOCCOZ,* K. A. BRÅTHEN,* L. GIELLY,† J. HAILE,‡§ M. E. EDWARDS,¶ T. GOSLAR,** H. VON STEDINGK,¶ A. K. BRYSTING,†† E. COISSAC,† F. POMPANON,† J. H. SØNSTEBØ,†† C. MIQUEL,† A. VALENTINI,† F. DE BELLO,†,‡‡ J. CHAVE,§§ W. THUILLER,† P. WINCKER,¶¶ C. CRUAUD,¶¶ F. GAVORY,¶¶ M. RASMUSSEN,‡ M. T. P. GILBERT,‡ L. ORLANDO‡ C. BROCHMANN,††[1] E. WILLERSLEV,‡[1] and P. TABERLET,†[1]

*Department of Arctic and Marine Biology, University of Tromsø, NO-9037 Tromsø, Norway, †Laboratoire d'Ecologie Alpine, CNRS UMR 5553, Université Joseph Fourier, BP 43, F-38041 Grenoble Cedex 9, France, ‡Centre for GeoGenetics, University of Copenhagen, Øster Voldgade 5-7, 1350 Copenhagen, Denmark, §Murdoch University, Perth, Western Australia 6150, Australia, ¶University of Southampton, Geography and Environment, Southampton SO17 1BJ, UK, **Faculty of Physics, Adam Mickiewicz University, ul. Umultowska 85, 61-614 Poznan, Poland, ††National Centre for Biosystematics, Natural History Museum, University of Oslo, PO Box 1172, Blindern, N-0318 Oslo, Norway, ‡‡Institute of Botany, Czech Academy of Sciences, Dukelská 135, CZ-379 82, Třeboň, Czech Republic, §§Laboratoire Evolution et Diversité Biologique, CNRS UMR 5174, Université Paul Sabatier, F-31062 Toulouse, France, ¶¶Genoscope, CEA, CNRS, UMR 8030, 2 rue Gaston Crémieux, BP 5706, F-91057 Evry cedex, France

Abstract

Ecosystems across the globe are threatened by climate change and human activities. New rapid survey approaches for monitoring biodiversity would greatly advance assessment and understanding of these threats. Taking advantage of next-generation DNA sequencing, we tested an approach we call metabarcoding: high-throughput and simultaneous taxa identification based on a very short (usually <100 base pairs) but informative DNA fragment. Short DNA fragments allow the use of degraded DNA from environmental samples. All analyses included amplification using plant-specific versatile primers, sequencing and estimation of taxonomic diversity. We tested in three steps whether degraded DNA from dead material in soil has the potential of efficiently assessing biodiversity in different biomes. First, soil DNA from eight boreal plant communities located in two different vegetation types (meadow and heath) was amplified. Plant diversity detected from boreal soil was highly consistent with plant taxonomic and growth form diversity estimated from conventional above-ground surveys. Second, we assessed DNA persistence using samples from formerly cultivated soils in temperate environments. We found that the number of crop DNA sequences retrieved strongly varied with years since last cultivation, and crop sequences were absent from nearby, uncultivated plots. Third, we assessed the universal applicability of DNA metabarcoding using soil samples from tropical environments: a large proportion of species and families from the study site were efficiently recovered. The results open unprecedented opportunities for large-scale DNA-based biodiversity studies across a range of taxonomic groups using standardized metabarcoding approaches.

Sapkota and Nicolaisen *BMC Ecology* (2015) 15:3
DOI 10.1186/s12898-014-0034-4

BMC Ecology

METHODOLOGY ARTICLE Open Access

High-throughput sequencing of nematode communities from total soil DNA extractions

Rumakanta Sapkota and Mogens Nicolaisen*

Abstract

Background: Nematodes are extremely diverse and numbers of species are predicted to be more than a million. Studies on nematode diversity are difficult and laborious using classical methods and therefore high-throughput sequencing is an attractive alternative. Primers that have been used in previous sequence-based studies are not nematode specific but also amplify other groups of organisms such as fungi and plantae, and thus require a nematode enrichment step that may introduce biases.

Results: In this study an amplification strategy which selectively amplifies a fragment of the SSU from nematodes without the need for enrichment was developed. Using this strategy on DNA templates from a set of 22 agricultural soils, we obtained 64.4% sequences of nematode origin in total, whereas the remaining sequences were almost entirely from other metazoans. The nematode sequences were derived from a broad taxonomic range and most sequences were from nematode taxa that have previously been found to be abundant in soil such as Tylenchida, Rhabditida, Dorylaimida, Triplonchida and Araeolaimida.

Conclusions: Our amplification and sequencing strategy for assessing nematode diversity was able to collect a broad diversity without prior nematode enrichment and thus the method will be highly valuable in ecological studies of nematodes.

Keywords: Nematode, Community, Next-generation sequencing, SSU, Diversity, 18S, rDNA

JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

What will properly analysed microbiomes tell you

- Who are there…
- …in which proportions of 16S rRNA genes (compared to your other samples, or other samples processed with the same pipeline)
  - Remember! 16S rRNA gene copy numbers vary from 1 to 15 per bacterial genome
- It will not confirm someone is not there or that someone is there
  - sampling depth
  - procedures to ensure lack of contamination)

JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

Proportions of gDNA vs. 16S rRNA genes in ZymoBIOMICS™ Microbial Community DNA Standard

Table 1: Microbial Composition

Species	Theoretical Composition (%)	
	Genomic DNA	16S rRNA[1]
Pseudomonas aeruginosa	12.0	4.6
Escherichia coli	12.0	10.0
Salmonella enterica	12.0	11.3
Lactobacillus fermentum	12.0	18.8
Enterococcus faecalis	12.0	10.4
Staphylococcus aureus	12.0	13.3
Listeria monocytogenes	12.0	15.9
Bacillus subtilis	12.0	15.7
Saccharomyces cerevisiae	2.0	-
Cryptococcus neoformans	2.0	-

[1] The theoretical composition in terms of 16S rRNA gene abundance was calculated from theoretical genomic DNA composition with the following formula: 16S copy number = total genomic DNA (g) × unit conversion constant (bp/g) / genome size (bp) × 16S copy number per genome

nature biotechnology

NATURE BIOTECHNOLOGY | COMPUTATIONAL BIOLOGY | ANALYSIS

日本語要約

Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences

Morgan G I Langille, Jesse Zaneveld, J Gregory Caporaso, Daniel McDonald, Dan Knights, Joshua A Reyes, Jose C Clemente, Deron E Burkepile, Rebecca L Vega Thurber, Rob Knight, Robert G Beiko & Curtis Huttenhower

Affiliations | Contributions | Corresponding author

Nature Biotechnology **31**, 814–821 (2013) | doi:10.1038/nbt.2676
Received 13 May 2013 | Accepted 29 July 2013 | Published online 25 August 2013

Abstract

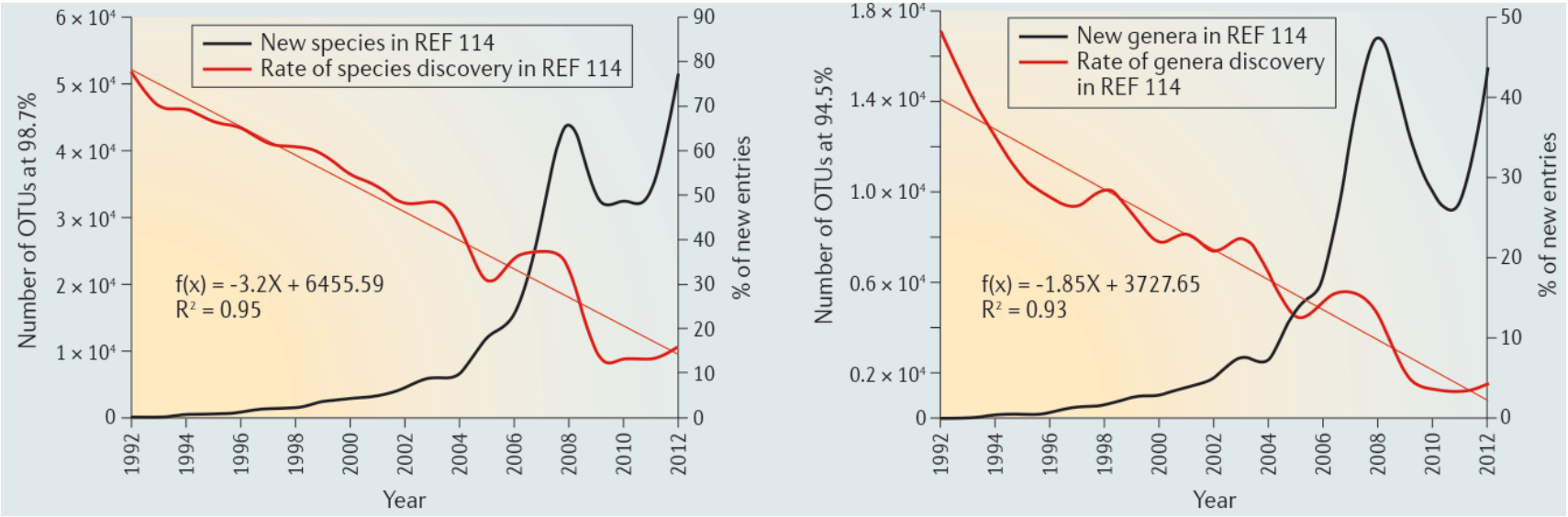
Abstract · Introduction · Results · Discussion · Methods · References · Acknowledgments · Author information · Supplementary information

Profiling phylogenetic marker genes, such as the 16S rRNA gene, is a key tool for studies of microbial communities but does not provide direct evidence of a community's functional capabilities. Here we describe PICRUSt (phylogenetic investigation of communities by reconstruction of unobserved states), a computational approach to predict the functional composition of a metagenome using marker gene data and a database of reference genomes. PICRUSt uses an extended ancestral-state reconstruction algorithm to predict which gene families are present and then combines gene families to estimate the composite metagenome. Using 16S information, PICRUSt recaptures key findings from the Human Microbiome Project and accurately predicts the abundance of gene families in host-associated and environmental communities, with quantifiable uncertainty. Our results demonstrate that phylogeny and function are sufficiently linked that this 'predictive metagenomic' approach should provide useful insights into the thousands of uncultivated microbial communities for which only marker gene surveys are currently available.

Indication of what functions are found (based on literature and e.g. PICRUSt)

JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

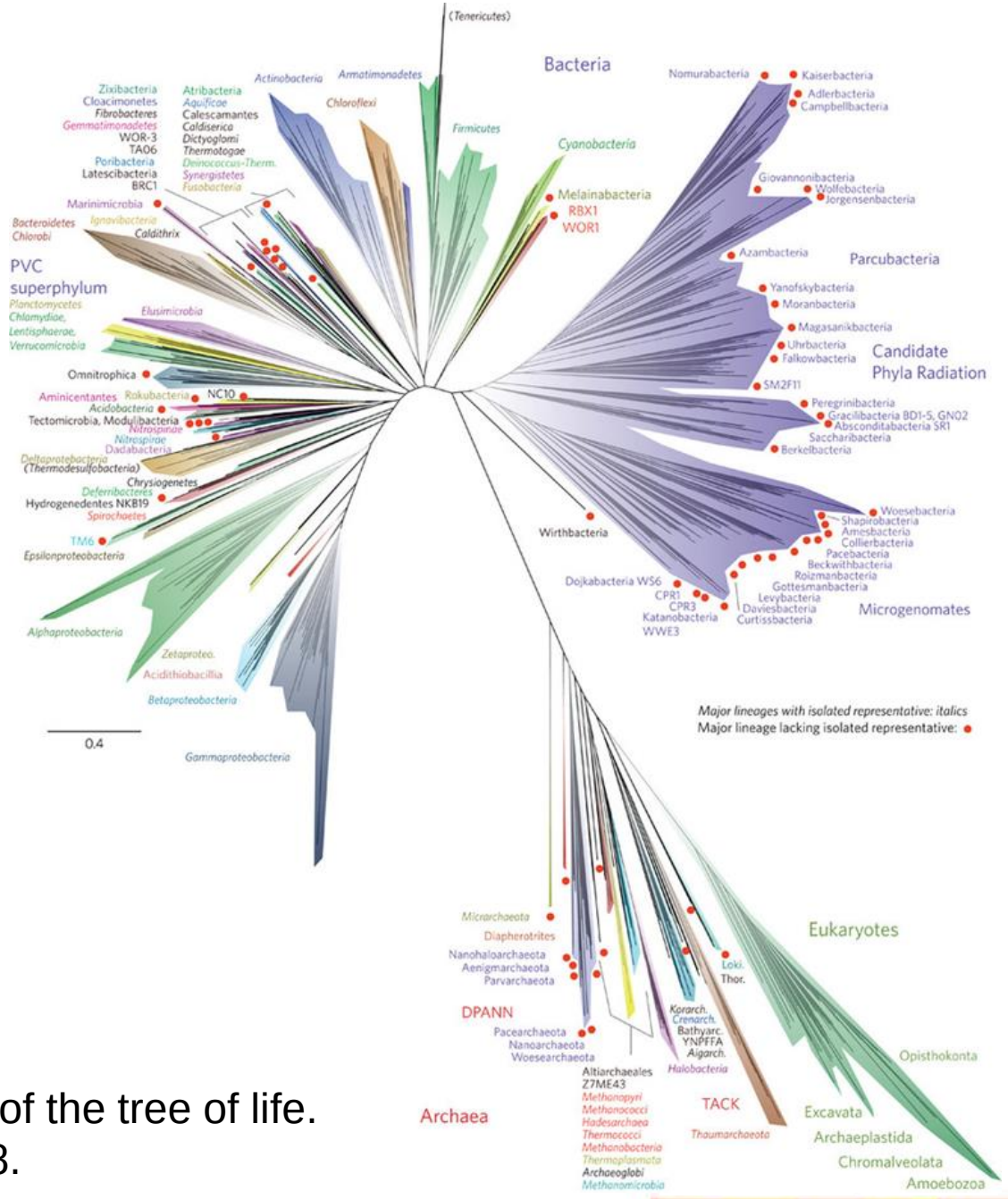
We already know pretty well the existing microbial diversity…



Yarza et al. Nature Reviews Microbiology 2014

JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

…or do we?



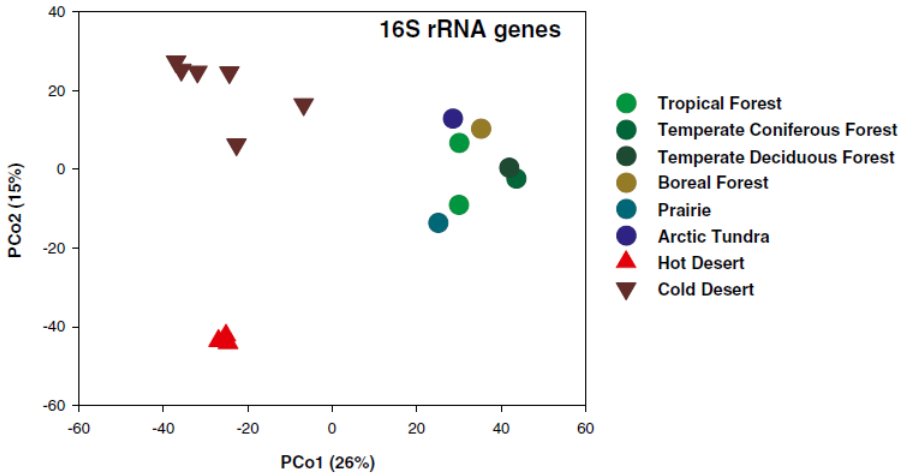
Hug et al. 2016. A new view of the tree of life. Nature Microbiology 1:16048.

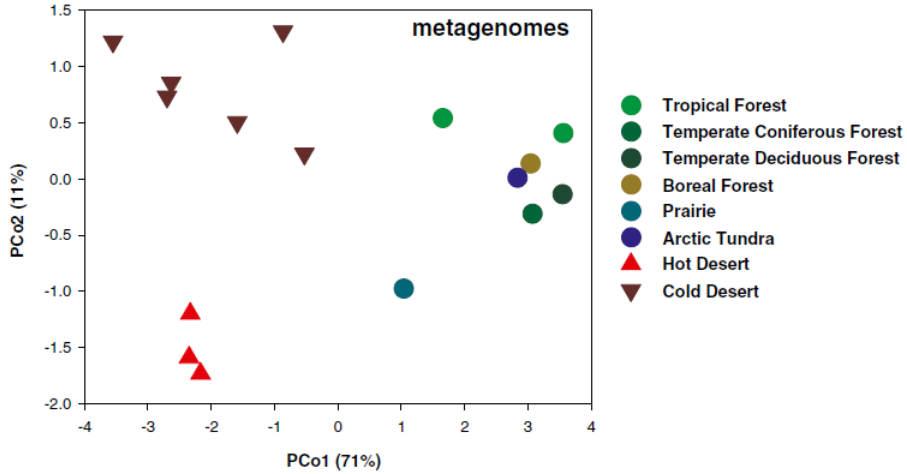
Cross-biome metagenomic analyses of soil microbial communities and their functional attributes

2012

Noah Fierer[a,b,1], Jonathan W. Leff[b], Byron J. Adams[c], Uffe N. Nielsen[d], Scott Thomas Bates[b], Christian L. Lauber[b], Sarah Owens[e,f], Jack A. Gilbert[e,g], Diana H. Wall[h], and J. Gregory Caporaso[e,i]

[a]Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, CO 80309; [b]Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, CO 80309; [c]Department of Biology and Evolutionary Ecology Laboratories, Brigham Young University, Provo, UT 84602; [d]Hawkesbury Institute for the Environment and School of Science and Health, University of Western Sydney, Penrith, NSW 2751, Australia; [e]Institute of Genomic and Systems Biology, Argonne National Laboratory, Argonne, IL 60439; [f]Computation Institute, University of Chicago, Chicago, IL 60637; [g]Department of Ecology and Evolution, University of Chicago, Chicago, IL 60637; [h]Department of Biology and School of Global Environmental Sustainability, Colorado State University, Fort Collins, CO 80523; and [i]Department of Computer Science, Northern Arizona University, Flagstaff, AZ 86011





Microbiome analysis results often cohere with metagenomics data

JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

How many sequences do you need?

- Betadiversiteettiin 1000 ihan hyvin!! Syvempi sekvensointi kertoo osaryhmien betadiversiteetistä paremmin, mutta vielä 100 000 sekvenssillä et voi varmistaa että näytteessä oikeasti on jotain (tähän tarvitaan aseptinen työskentely, rinnakkaiset ja teoria!) tai ei ole jotain (otanta esim. grammasta maata kuitenkin surkean pieni! ja tähän tarvitaan varmistus, että uutto ja monistus toimivat juuri tässä matriisissa juuri tälle ryhmälle)

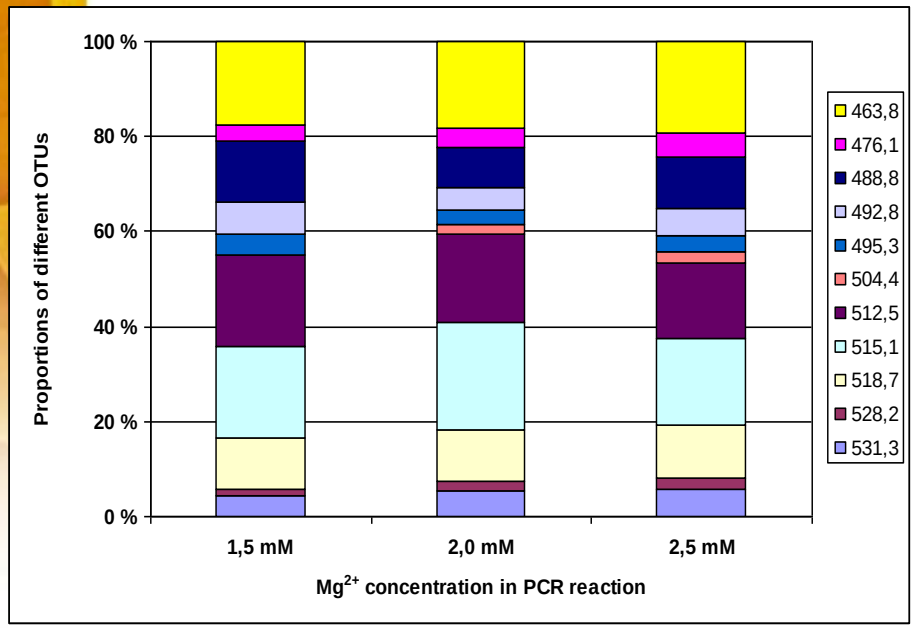
Näytä PCoA vertailut 1000 vs 5000 vs 10000?

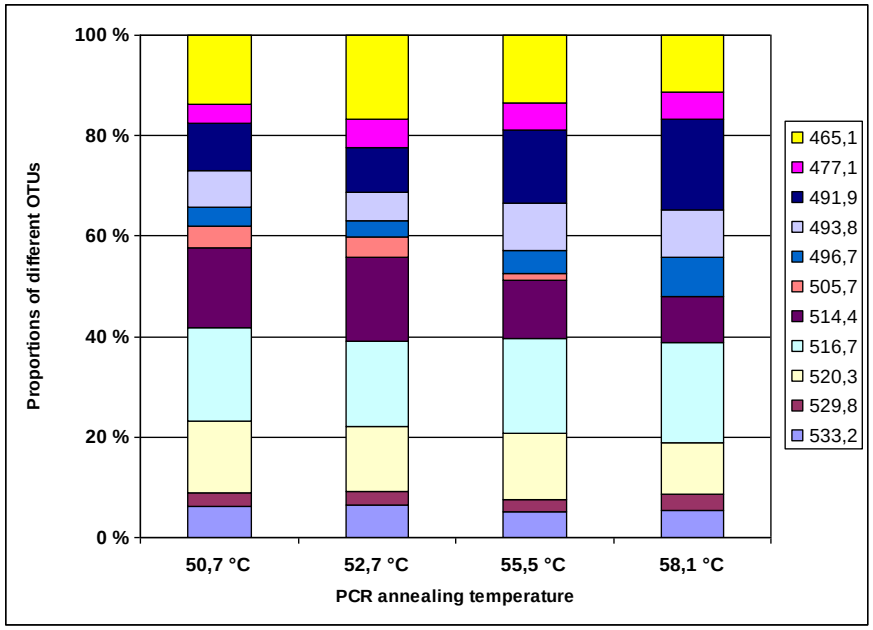
JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

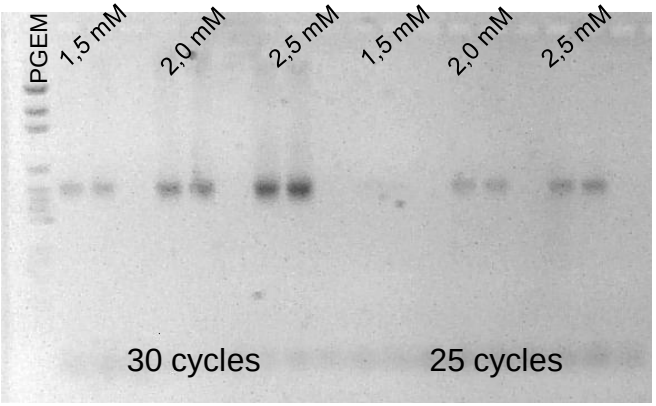
EFFECT OF PCR CONDITIONS

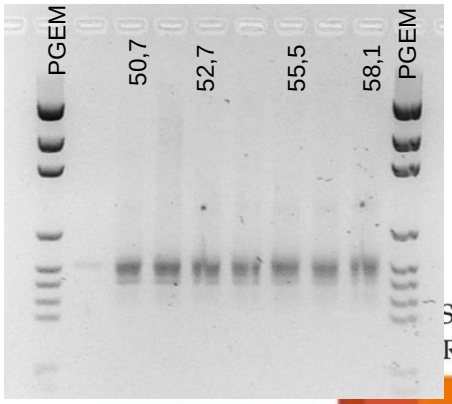
JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

Effect of priming stringency (Mg^{2+} and annealing temperature) on amplified community









SKYLÄN YLIOPISTO
RSITY OF JYVÄSKYLÄ

Effect of polymerase/master mix

- In the work for Mikkonen et al. 2014 FEMS Microbiology Ecology, we never got archea to amplify with a proofreading polymerase

RESEARCH ARTICLE

Bacterial and archaeal communities in long-term contaminated surface and subsurface soil evaluated through coextracted RNA and DNA

Anu Mikkonen, Minna Santalahti, Kaisa Lappi, Anni-Mari Pulkkinen, Leone Montonen & Leena Suominen

Department of Food and Environmental Sciences, Division of Microbiology and Biotechnology, University of Helsinki, Helsinki, Finland

- For amplicon sequencing some laboratories recommend proofreading polymerase, but most (e.g. Fierer lab) use Taq polymerase – and expect any Taq to be the same Taq

JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

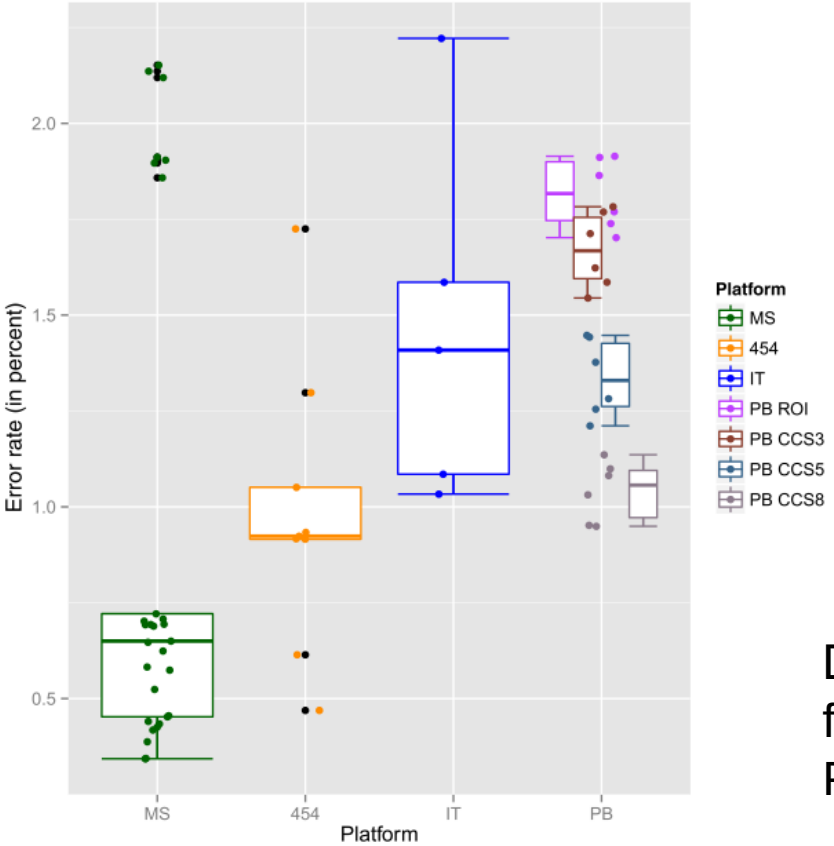
Effect of template concentration

	Water sample 1 (0.047 ng/µl)			Water sample 2 (0,5 ng/µl)		
	1/1	1/10	1/100	1/1	1/10	1/100
Acidobacteria	2,25	1,54	1,15	0,08	0,03	1,37
Actinobacteria	**1,68**	**1,04**	**0,91**	**27,87**	**16,45**	**8,22**
Armatimonadetes	0,21	0,15	0,04	0,00	0,00	0,00
BD1-5	0,76	0,62	0,80	0,00	0,00	0,00
Bacteroidetes	**10,89**	**8,86**	**7,94**	**30,70**	**19,29**	**11,64**
Caldiserica	0,53	0,27	0,62	0,05	0,03	0,00
Candidate_division_KB1	0,02	0,02	0,04	0,00	0,00	0,00
Candidate_division_OD1	13,04	13,01	12,14	0,57	0,79	3,42
Candidate_division_OP11	0,87	1,22	0,82	0,05	0,00	0,00
Candidate_division_OP3	0,30	0,32	0,22	0,03	0,03	0,00
Candidate_division_OP8	0,21	0,25	0,33	0,00	0,03	0,68
Candidate_division_SR1	0,72	1,04	0,73	0,05	0,09	0,00
Candidate_division_TM7	0,38	0,22	0,15	0,00	0,00	0,00
Candidate_division_WS3	0,04	0,12	0,13	0,00	0,00	0,00
Chlorobi	1,38	0,77	1,08	0,05	0,03	0,00
Chloroflexi	9,32	11,24	11,25	0,82	0,52	0,00
Cyanobacteria	4,74	6,37	6,10	0,15	0,09	2,74
Elusimicrobia	0,45	0,47	0,53	0,00	0,00	0,00
Fibrobacteres	0,11	0,05	0,04	0,00	0,00	0,00
Firmicutes	1,63	2,06	1,77	0,05	0,09	0,00
Fusobacteria	0,02	0,00	0,00	0,00	0,00	0,00
GOUTA4	0,00	0,02	0,00	0,00	0,00	0,00
Gemmatimonadetes	0,02	0,02	0,02	0,00	0,00	0,00
Lentisphaerae	0,04	0,02	0,00	0,00	0,00	0,00
Nitrospirae	0,09	0,00	0,00	0,00	0,00	0,00
Planctomycetes	0,13	0,02	0,02	0,00	0,00	0,00
Proteobacteria	**18,70**	**19,20**	**21,53**	**37,36**	**60,76**	**65,75**
Spirochaetae	0,68	0,37	0,15	0,03	0,00	0,68
TA06	0,17	0,05	0,20	0,00	0,00	0,00
TM6	2,50	2,19	1,88	0,00	0,09	0,00
Verrucomicrobia	0,06	0,07	0,15	0,00	0,00	0,00
WD272	0,06	0,07	0,07	0,00	0,00	0,00
unclassified	20,76	19,38	19,85	1,83	1,50	4,79

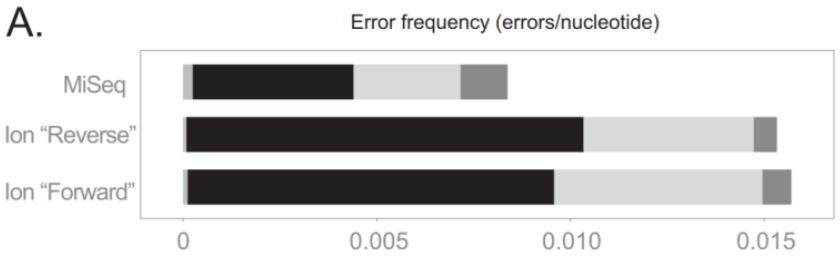
Most groups don't quantify and normalise template DNA concentrations - but it is HIGHLY recommended

JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

Comparison of sequencing platforms



D'Amore et al. BMC Genomics 2016



Salipante et al. AEM 2014

DON'T try to merge OTUs or sequences from different platforms!
Phylotypes are more safe to merge

JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

Comparison of sequencing platforms with current chemistries

- Illumina MiSeq Reagent Kit v3 (2 × 300 bp)
  - outsourced to HY-BI, ~50 € / sample
- Ion Torrent PGM HiQ View template preparation and sequencing kits (1 × 400 bp)
  - in-house at JYU, ~7 € / sample

- Same defined template
  - mock community consisting of known strains
- V1-V2 region of 16S rRNA genes

JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

ZymoBIOMICS™ Microbial Community DNA Standard

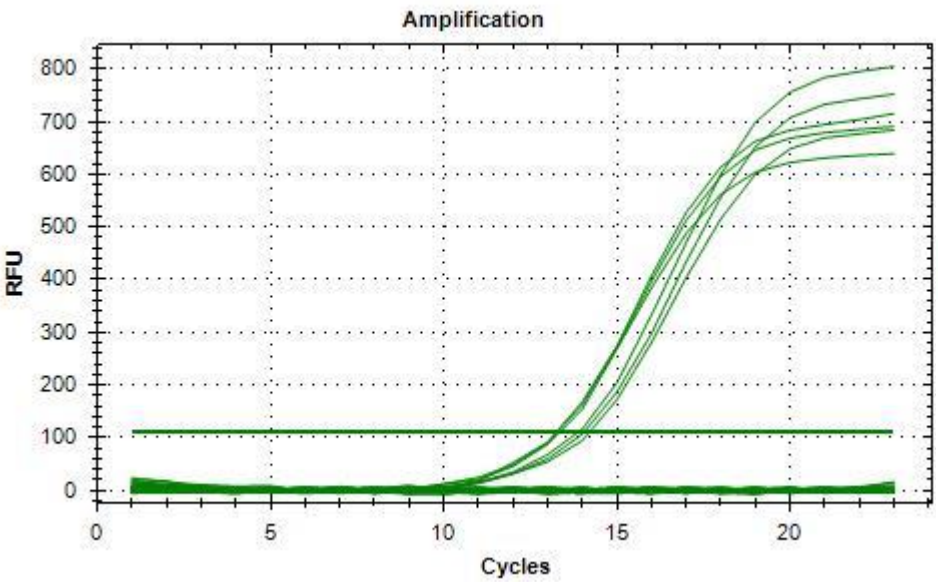
Table 2: Strain Information

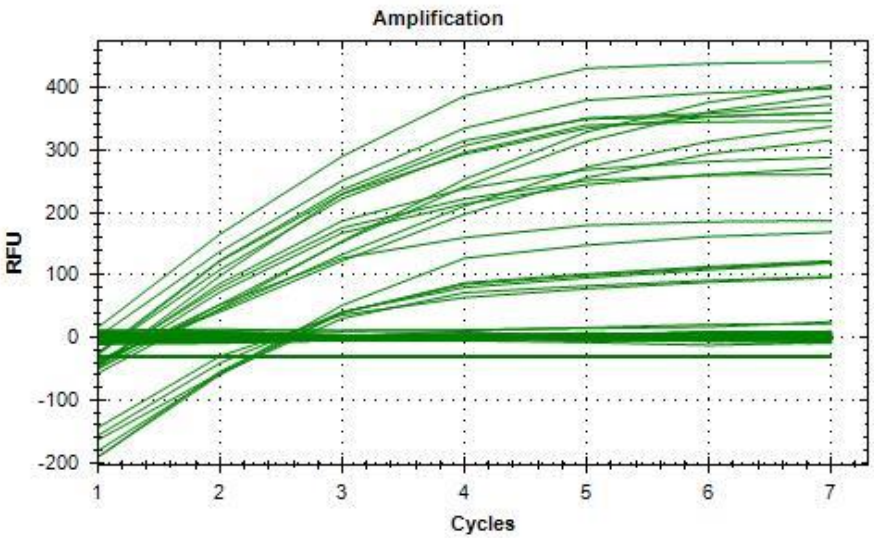
Species	Genome Size (Mb)	Ploidy	GC Content (%)	16/18S Copy Number	Gram Stain
Pseudomonas aeruginosa	6.77	1	66.2	4	-
Escherichia coli	5.47	1	56.8	7	-
Salmonella enterica	4.83	1	52.2	7	-
Lactobacillus fermentum	2.08	1	52.8	5	+
Enterococcus faecalis	3.01	1	37.5	4	+
Staphylococcus aureus	2.93	1	32.7	5	+
Listeria monocytogenes	2.95	1	38.0	6	+
Bacillus subtilis	3.98	1	43.8	8	+
Saccharomyces cerevisiae	13.3	2	38.4	109[1]	Yeast
Cryptococcus neoformans	18.9	2	48.2	60[1]	Yeast

Full 16S rRNA gene sequences available online

JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

Same PCR conditions and good amplification efficiency



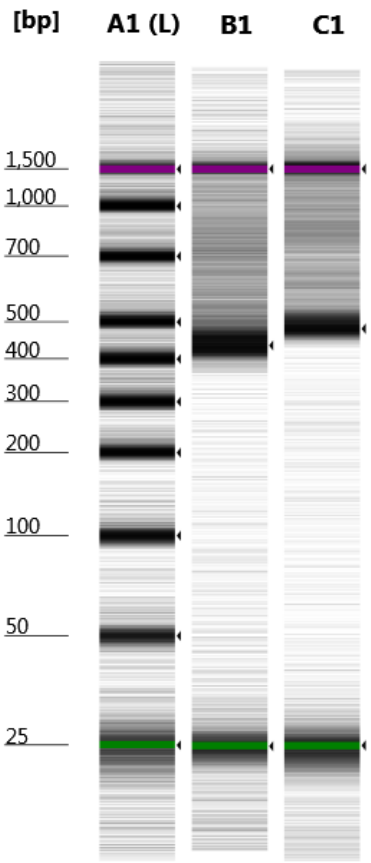


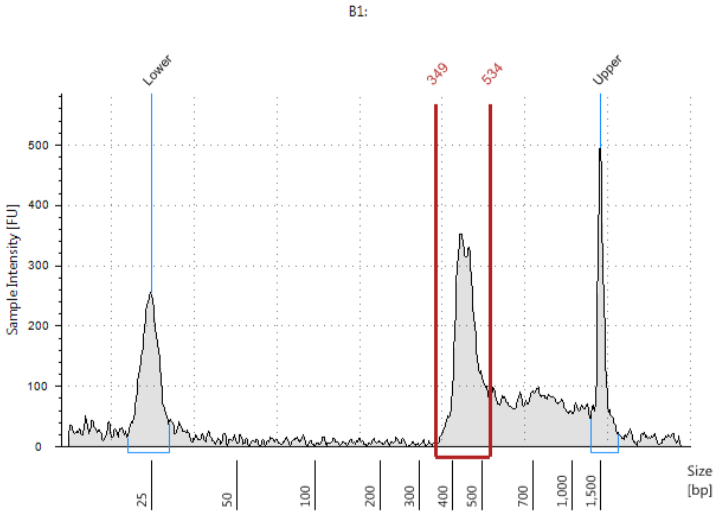
- 1st PCR in triplicate, pooled after amplification
- PGM primers just 27F and 338R
- MiSeq primers had partial TruSeq adapter sequences at 5' ends of 27F and 338R

- 2nd PCR using 2 µl of the pooled 1st PCR product as the template
- Addition of barcodes and sequencing adapters
- 4 barcodes per platform

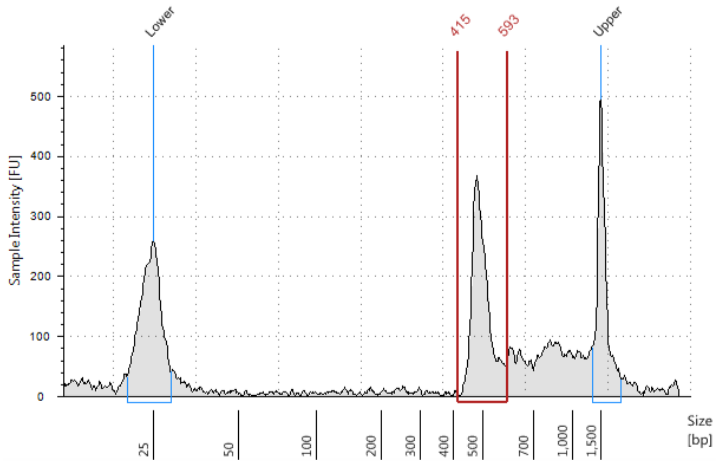
JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

TapeStation analysis of pool quality





PGM pool



MiSeq pool

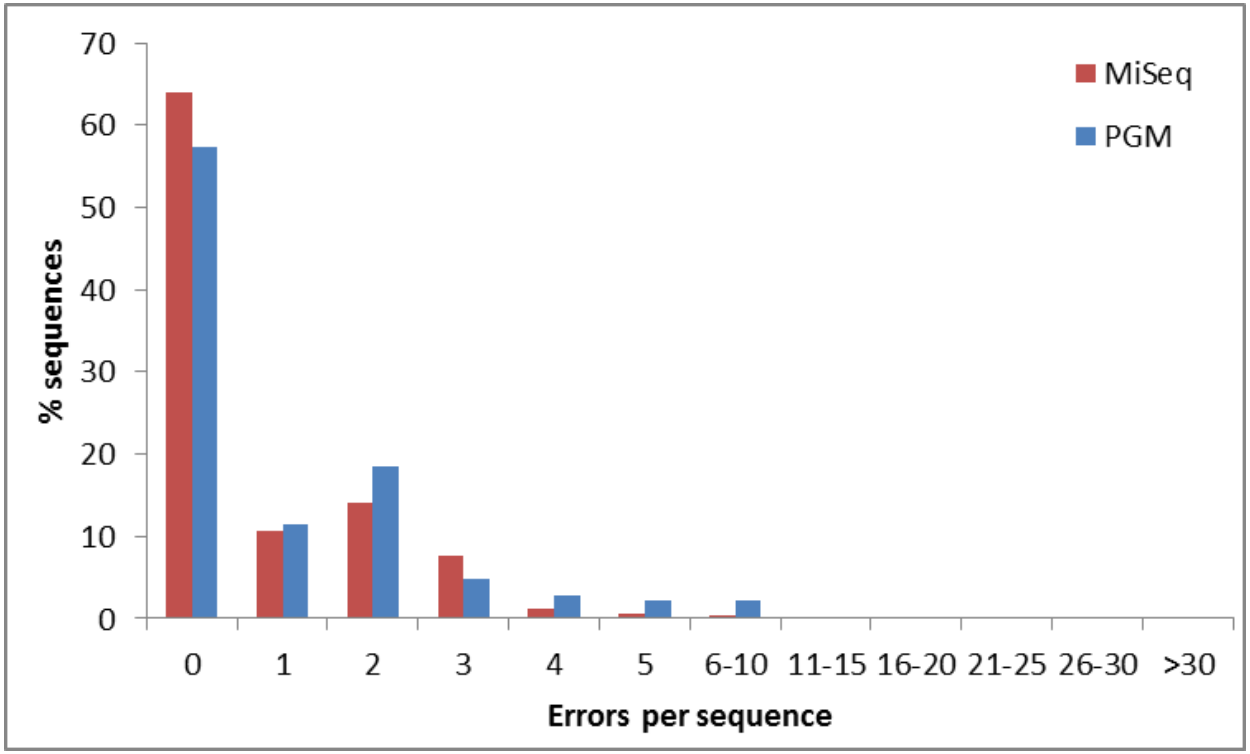
JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

Processing the samples according to mothur standard operating procedures

- 454-SOP for Ion Torrent data – without sequence quality-based filtering
  - Allow one barcode and forward primer error, no quality filtering
  - Minimum length of 200 bp
- MiSeq-SOP for MiSeq data
  - Merge R1&R2, quality filtering embedded
  - Maximum length of 400
- Only the beginning differs, otherwise identical pipeline (alignment and classification based on Silva v. 128, chimera detection with Uchime, clustering with OptiClust)
- Error analysis according to 454-SOP

JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

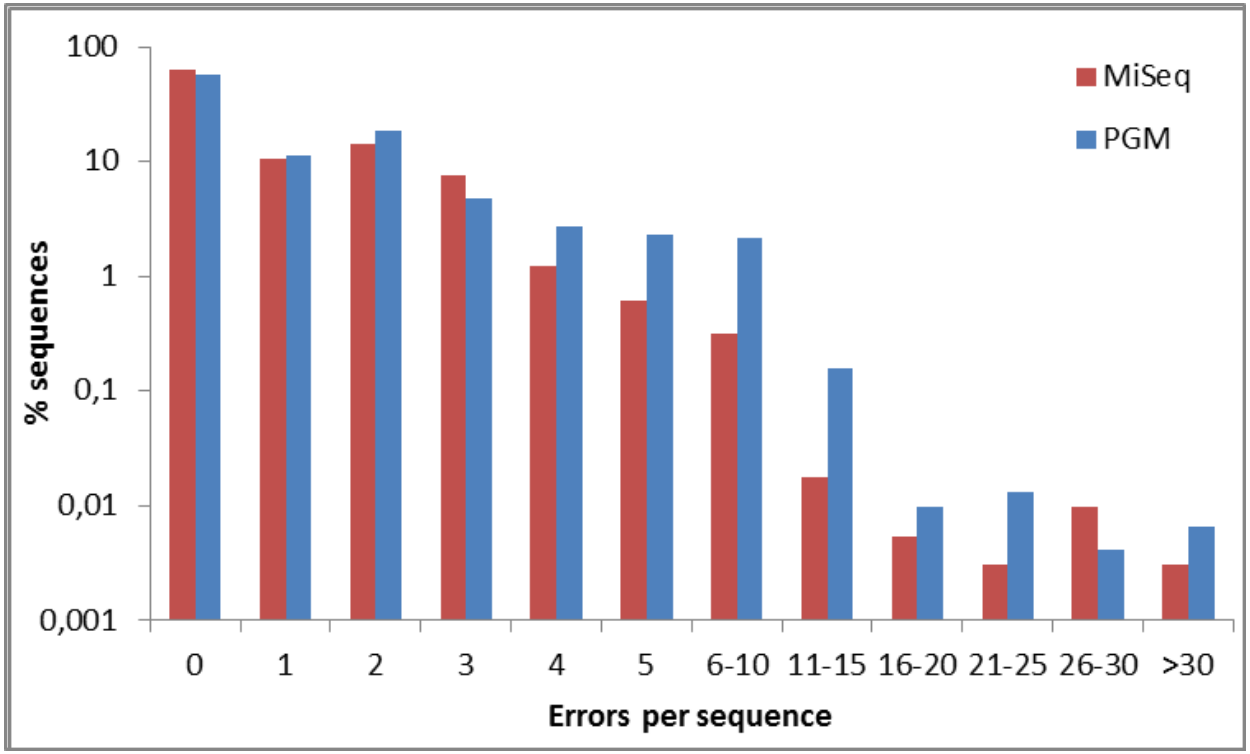
Number of errors per sequence



- MiSeq total error 0.22%
  - 706 790 seqs: 169 920-195 581 per barcode
- PGM total error 0.47%
  - 122 405 seqs: 27 877-32 380 per barcode

JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

Number of errors per sequence



- MiSeq total error 0.22%
  - 706 790 seqs: 169 920-195 581 per barcode
- PGM total error 0.47%
  - 122 405 seqs: 27 877-32 380 per barcode

JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

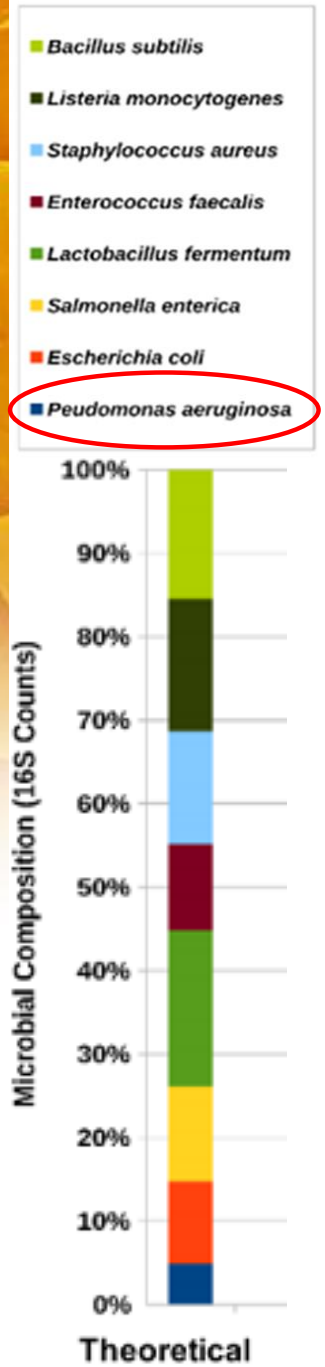
Processing the samples according to CLCbio OTU clustering pipeline

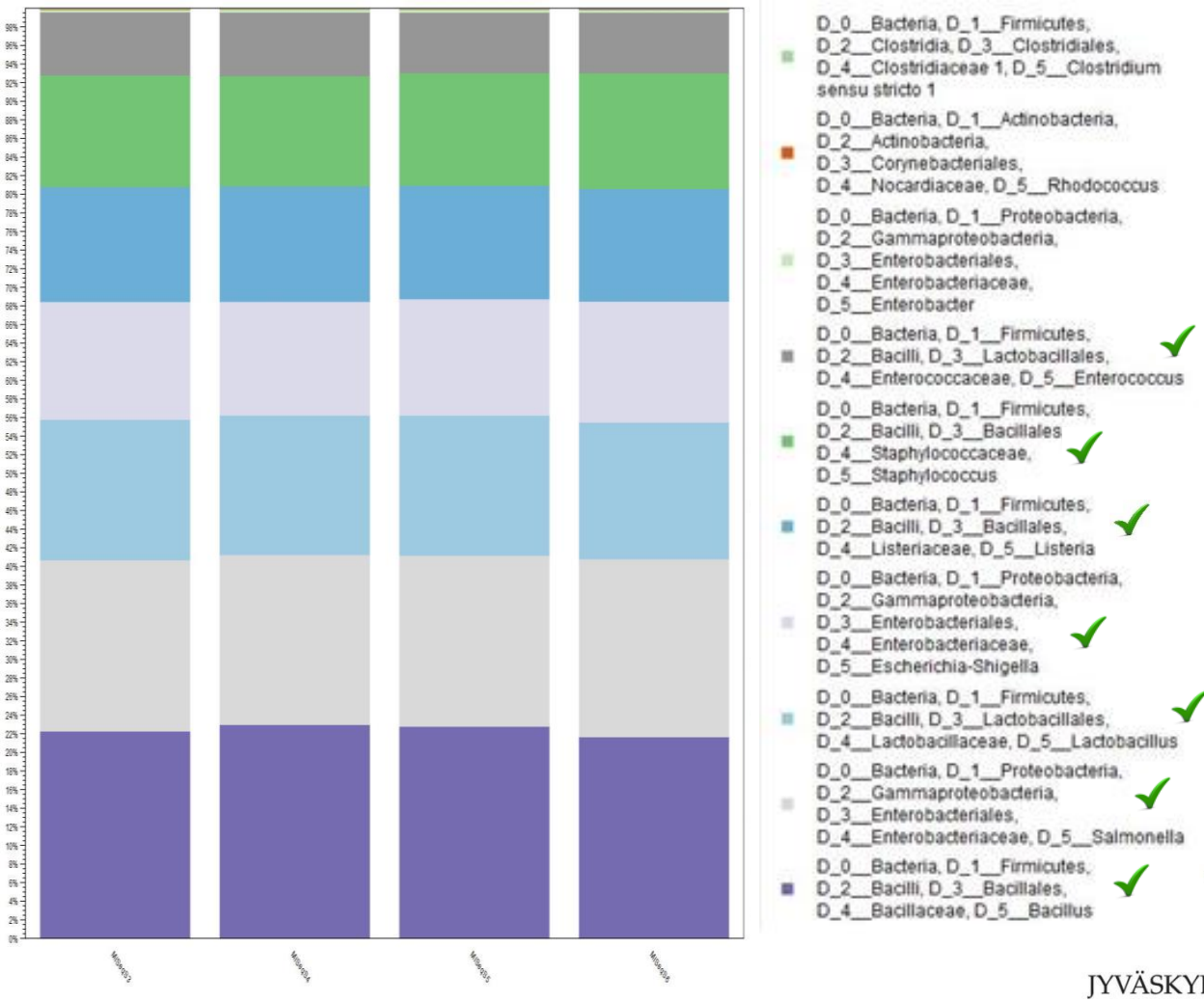
- Identical processing with default parameters:
  1. Quality trimming and primer sequence removal
  2. Trimming to fixed length
  3. OTU clustering based on Silva v.123 reference database, with built-in chimera removal

- Only difference that the MiSeq reads (R1&R2) were merged before first quality trimming

JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

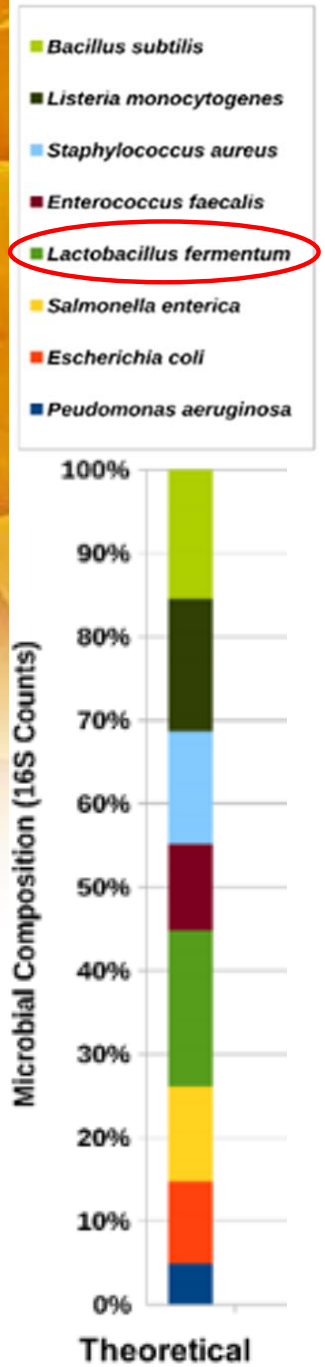
Classification - MiSeq

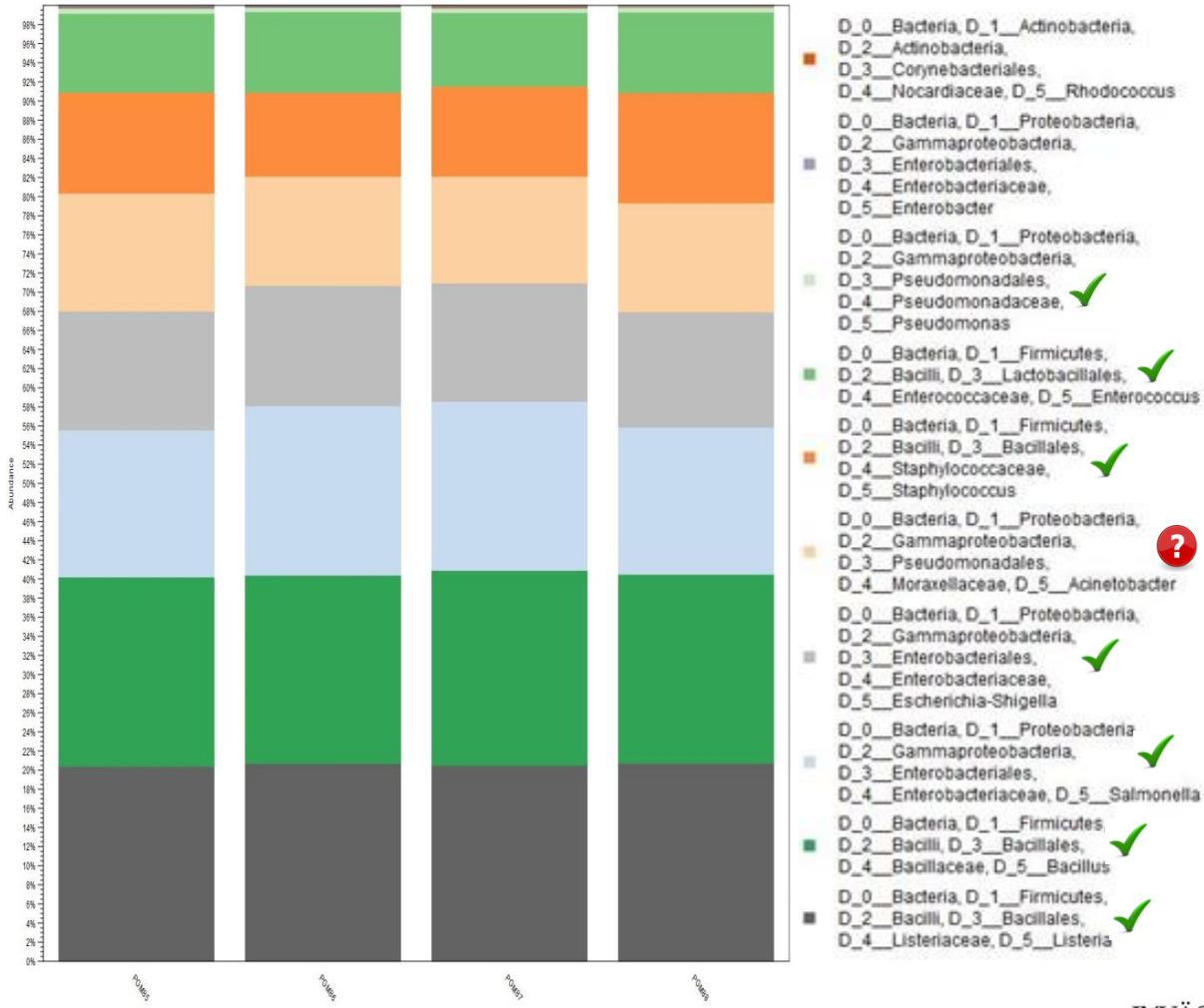




JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

Classification - PGM





JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

Coherent result with mothur

- PGM – 555 OTUs
  - All genera found in 10 most abundant OTUs
  - *Bacillus* split to two OTUs
  - *Lactobacillus* underrepresented
  - Quality filtering could improve the situation?
- MiSeq -118 OTUs
  - All genera found in 8 most abundant OTUs
  - more even ratios
  - Better merge tool could improve the situation?

Mothur, QIIME/QIITA, CLC, USEARCH…
and with which parameters?

ANALYSIS PLATFORMS – KNOWN ISSUES

JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

Merging MiSeq R1 and R2 reads

- USEARCH, followed by quality filtering based on expected errors, works well

http://www.drive5.com/usearch/manual/exp_errs.html

  - but multicore version is not free
- Other tools don't, unless the reads overlap fully (but few of us settle for just V4…)

http://blog.mothur.org/2014/09/11/Why-such-a-large-distance-matrix/

- MiSeq v3 chemistry is definitely not better than v2
- CLC pipeline seems to work pretty well

JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

Quality filtering of PGM reads

- Torrent Suite (server) does quality trimming for ends
- Additional quality window filtering has been used in mothur
- NOT recommended!
  - Significantly discriminated genera Nitrospira, Cupriavidus, Acidivorax, Curvibacter, Ramlibacter, Methylophilus, Zoogloea
- …and not absolutely necessary currently

OTU clustering

- OTU-splitting (OTU count inflation) is a known issue with most OTU clustering algorithms
- Usearch seems to perform best
  - but multicore version is not free
- OptiClust, implemented in mothur v.1.39 (January 2017), works well and efficiently! http://biorxiv.org/content/early/2016/12/23/096537

JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

Take-home message

1. Choose one of the commonly used pipelines and follow it – unless you can professionally fully justify exceptions
   - both *in vitro* and *in silico*
2. To enable comparison, process all your samples in EXACTLY the same way
3. Ion Torrent PGM is a usable and more economic alternative for Illumina MiSeq
4. High-throughput sequence data is never perfect

JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

Questions and discussion

JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ