Pros cons and challenges

Amplicons/marker genes/16S rRNA gene/microbiomics

Open & reproducible microbiome data analysis spring school
Wageningen, The Netherlands, May 28-30, 2018

Gerben DA Hermes, PhD

Laboratory of Microbiology,

Wageningen University & Research

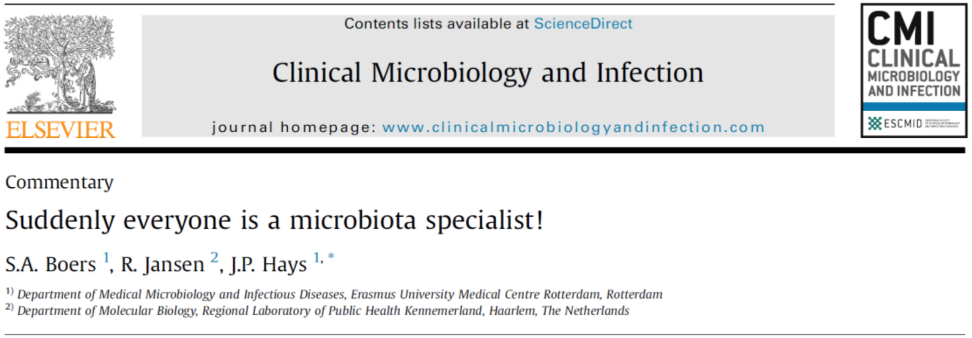
Some learning goals

Understand

- Critical steps in sample prep & data analysis
- Limitations (which conclusions can and can’t you draw)
- Data (biological!) interpretation (day 2/3)

Interpret literature

sense & nonsense



Contents lists available at ScienceDirect

Clinical Microbiology and Infection

journal homepage: www.clinicalmicrobiologyandinfection.com

Commentary

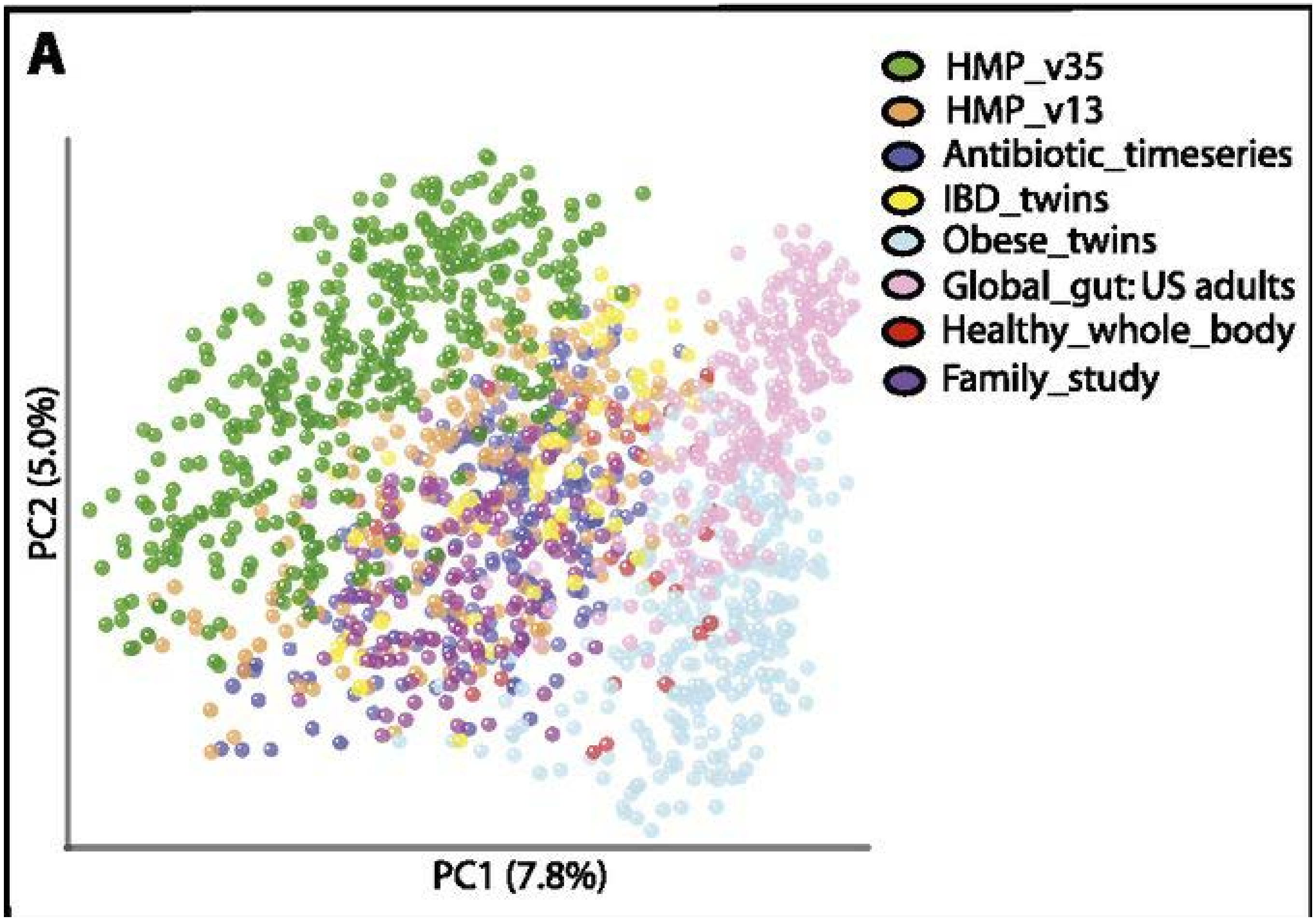
Suddenly everyone is a microbiota specialist!

S.A. Boers [1], R. Jansen [2], J.P. Hays [1,*]

[1] Department of Medical Microbiology and Infectious Diseases, Erasmus University Medical Centre Rotterdam, Rotterdam

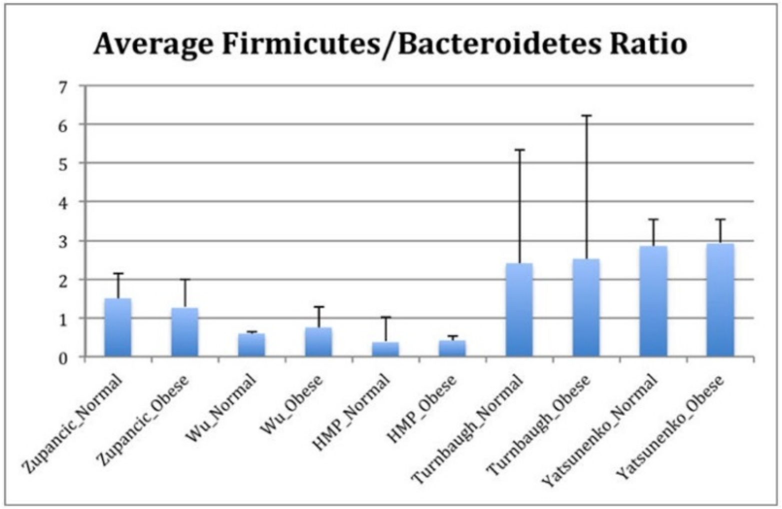
[2] Department of Molecular Biology, Regional Laboratory of Public Health Kennemerland, Haarlem, The Netherlands

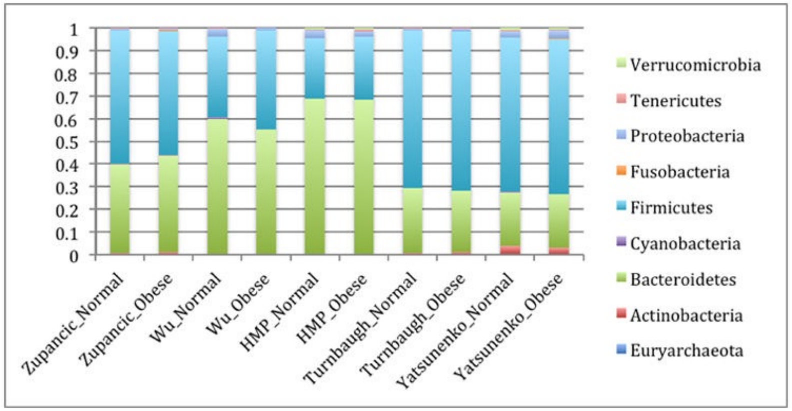
Literature & microbial biomarkers



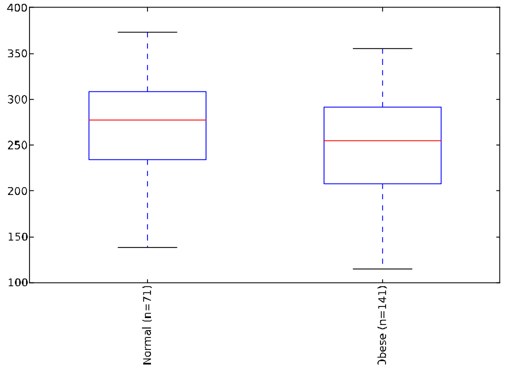
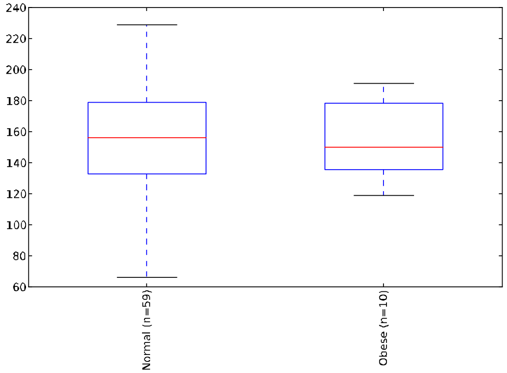
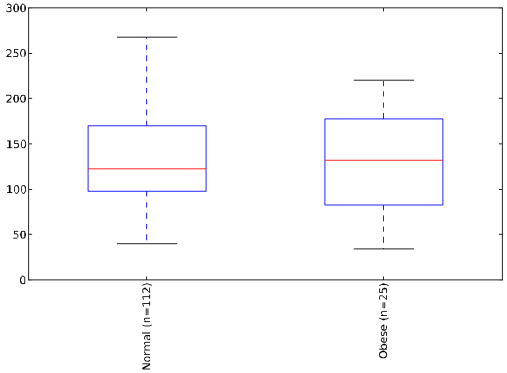
Catherine A. Lozupone et al. Genome Res. 2013;23:1704-1714

Microbial biomarkers for obesity





Walters et al. Meta-analyses of human gut microbes associated with obesity and IBD. FEBS Lett 2014

Study	Observed species plot for normal and obese individuals	p-value
Zupancic		0.015
Wu		0.95
HMP		0.955

WAGENINGEN
UNIVERSITY & RESEARCH

Before sample prep/data generation

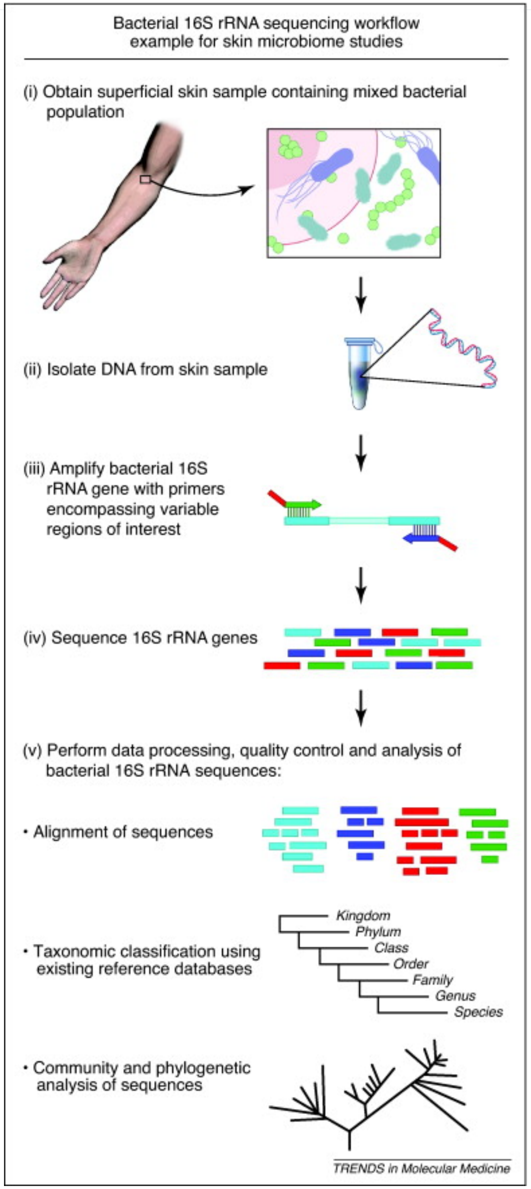
- How will you obtain fecal samples in your study?

- What do you want to do with it?
  - Functional (transcriptome -> quenching)
  - Metabolome
  - *In vitro* fermentations

- Freezing (how fast? What temperature? Freezer space)

- Pre-label tubes

Typical amplicon sequencing workflow



Bacterial 16S rRNA sequencing workflow example for skin microbiome studies

(i) Obtain superficial skin sample containing mixed bacterial population

(ii) Isolate DNA from skin sample

(iii) Amplify bacterial 16S rRNA gene with primers encompassing variable regions of interest

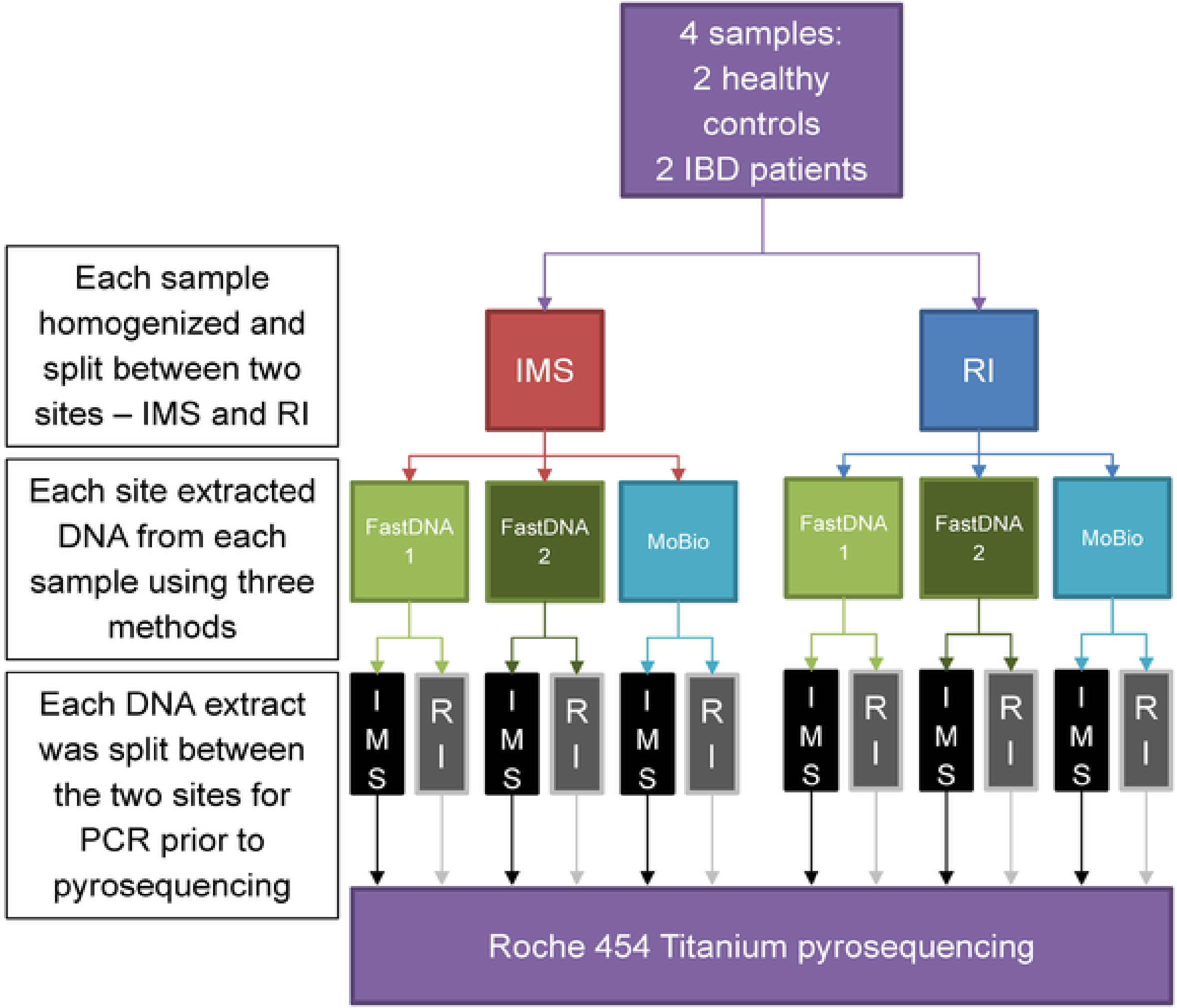
(iv) Sequence 16S rRNA genes

(v) Perform data processing, quality control and analysis of bacterial 16S rRNA sequences:

- Alignment of sequences
- Taxonomic classification using existing reference databases
- Community and phylogenetic analysis of sequences

TRENDS in Molecular Medicine

1. DNA isolation



4 samples:
2 healthy controls
2 IBD patients

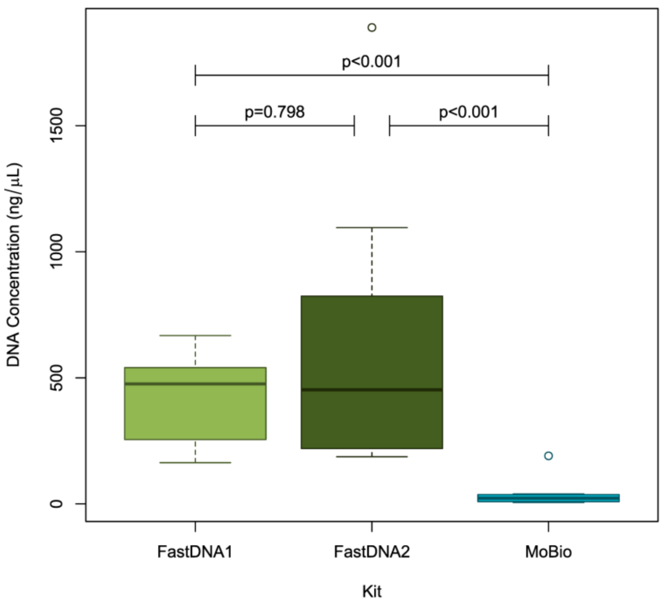
Each sample homogenized and split between two sites – IMS and RI

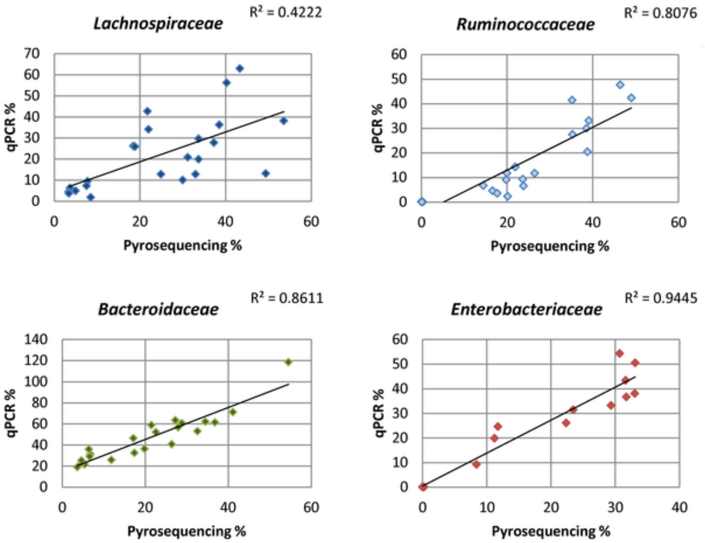
Each site extracted DNA from each sample using three methods

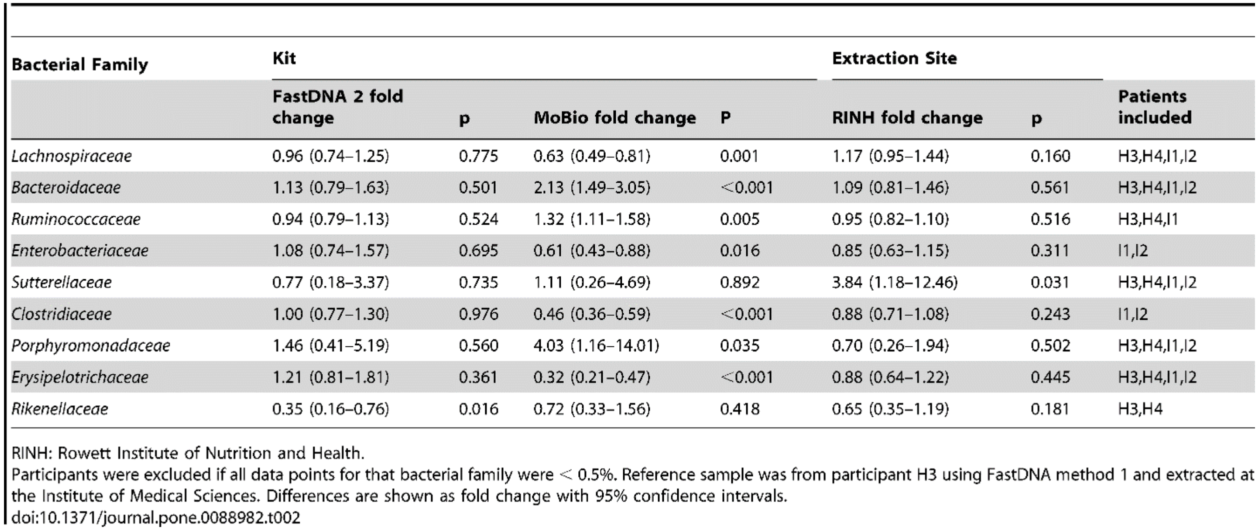
Each DNA extract was split between the two sites for PCR prior to pyrosequencing

Roche 454 Titanium pyrosequencing

Kennedy NA, Walker AW, Berry SH, Duncan SH, Farquarson FM, et al. (2014) The Impact of Different DNA Extraction Kits and Laboratories upon the Assessment of Human Gut Microbiota Composition by 16S rRNA Gene Sequencing. PLOS ONE 9(2): e88982. https://doi.org/10.1371/journal.pone.0088982
http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0088982





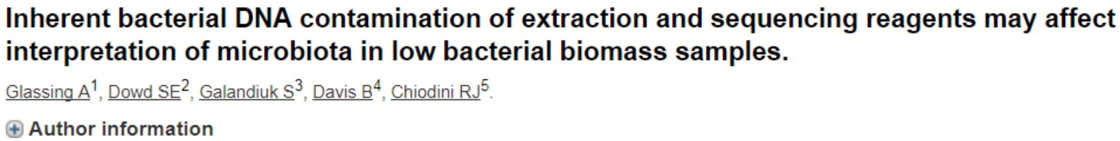


Bacterial Family	Kit				Extraction Site		
	FastDNA 2 fold change	p	MoBio fold change	P	RINH fold change	p	Patients included
Lachnospiraceae	0.96 (0.74–1.25)	0.775	0.63 (0.49–0.81)	0.001	1.17 (0.95–1.44)	0.160	H3,H4,I1,I2
Bacteroidaceae	1.13 (0.79–1.63)	0.501	2.13 (1.49–3.05)	<0.001	1.09 (0.81–1.46)	0.561	H3,H4,I1,I2
Ruminococcaceae	0.94 (0.79–1.13)	0.524	1.32 (1.11–1.58)	0.005	0.95 (0.82–1.10)	0.516	H3,H4,I1
Enterobacteriaceae	1.08 (0.74–1.57)	0.695	0.61 (0.43–0.88)	0.016	0.85 (0.63–1.15)	0.311	I1,I2
Sutterellaceae	0.77 (0.18–3.37)	0.735	1.11 (0.26–4.69)	0.892	3.84 (1.18–12.46)	0.031	H3,H4,I1,I2
Clostridiaceae	1.00 (0.77–1.30)	0.976	0.46 (0.36–0.59)	<0.001	0.88 (0.71–1.08)	0.243	I1,I2
Porphyromonadaceae	1.46 (0.41–5.19)	0.560	4.03 (1.16–14.01)	0.035	0.70 (0.26–1.94)	0.502	H3,H4,I1,I2
Erysipelotrichaceae	1.21 (0.81–1.81)	0.361	0.32 (0.21–0.47)	<0.001	0.88 (0.64–1.22)	0.445	H3,H4,I1,I2
Rikenellaceae	0.35 (0.16–0.76)	0.016	0.72 (0.33–1.56)	0.418	0.65 (0.35–1.19)	0.181	H3,H4

RINH: Rowett Institute of Nutrition and Health.
Participants were excluded if all data points for that bacterial family were < 0.5%. Reference sample was from participant H3 using FastDNA method 1 and extracted at the Institute of Medical Sciences. Differences are shown as fold change with 95% confidence intervals.
doi:10.1371/journal.pone.0088982.t002

Additional considerations

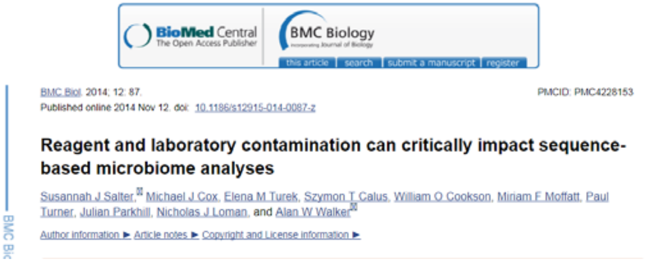
- ~~(Preferably)~~ **Don’t** work with low DNA yield samples
  - Can “contaminate” other samples
- Use negative extraction controls
- Laboratory reagent specific microbiota



Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples.

Glassing A[1], Dowd SE[2], Galandiuk S[3], Davis B[4], Chiodini RJ[5].

⊕ Author information



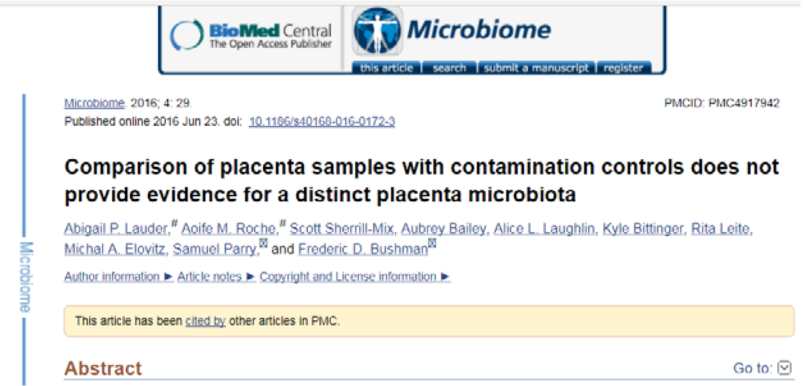
BMC Biol. 2014; 12: 87. PMCID: PMC4228153

Published online 2014 Nov 12. doi: 10.1186/s12915-014-0087-z

Reagent and laboratory contamination can critically impact sequence-based microbiome analyses

Susannah J Salter, Michael J Cox, Elena M Turek, Szymon T Calus, William O Cookson, Miriam F Moffatt, Paul Turner, Julian Parkhill, Nicholas J Loman, and Alan W Walker

Author information ► Article notes ► Copyright and License information ►



Microbiome. 2016; 4: 29. PMCID: PMC4917942

Published online 2016 Jun 23. doi: 10.1186/s40168-016-0172-3

Comparison of placenta samples with contamination controls does not provide evidence for a distinct placenta microbiota

Abigail P. Lauder,# Aoife M. Roche,# Scott Sherrill-Mix, Aubrey Bailey, Alice L. Laughlin, Kyle Bittinger, Rita Leite, Michal A. Elovitz, Samuel Parry, and Frederic D. Bushman

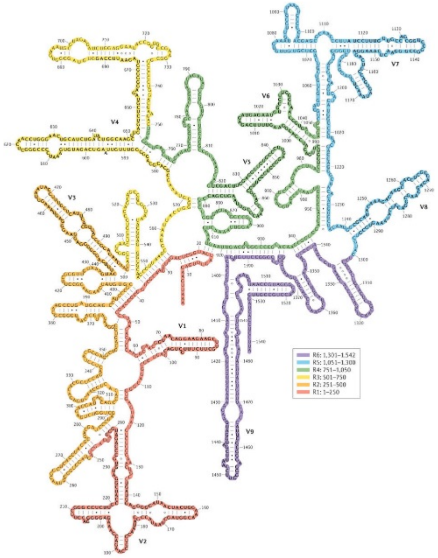
Author information ► Article notes ► Copyright and License information ►

This article has been cited by other articles in PMC.

Abstract Go to:

2. Barcoded PCR:

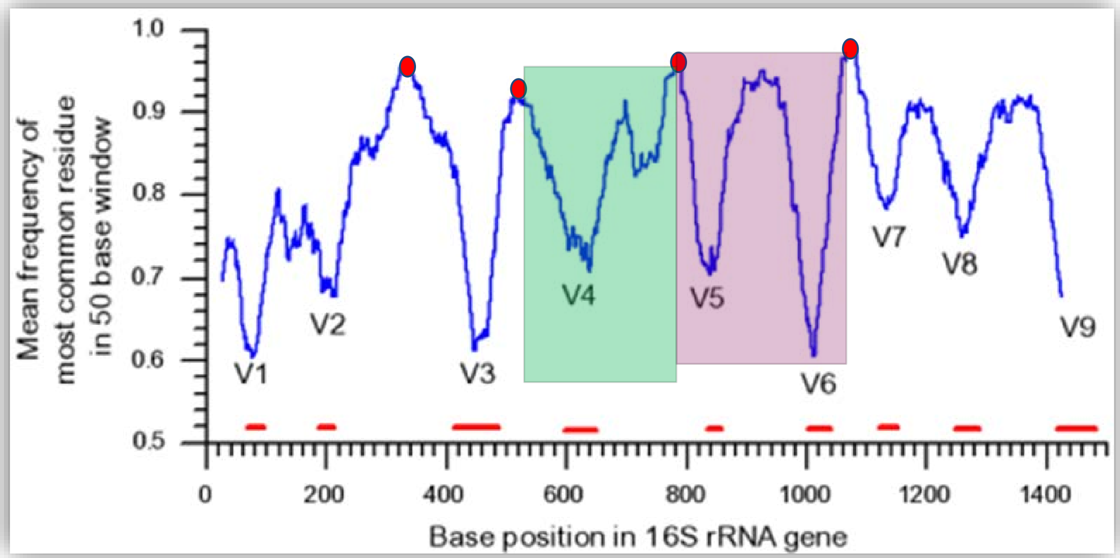
Primers & Region: Coverage, Resolution, bias



Nature Reviews | Microbiology

16S rRNA gene

- **Target for detection & identification of bacteria.**
- **Ideal phylogenetic marker.**
  - **–Universally distributed, functionally constant,**
  - **–Sufficiently conserved, no horizontal transfer (?)**
- **More than 3,000,000 sequences in databases.**
- **Alternating variable and conserved sequence domains.**



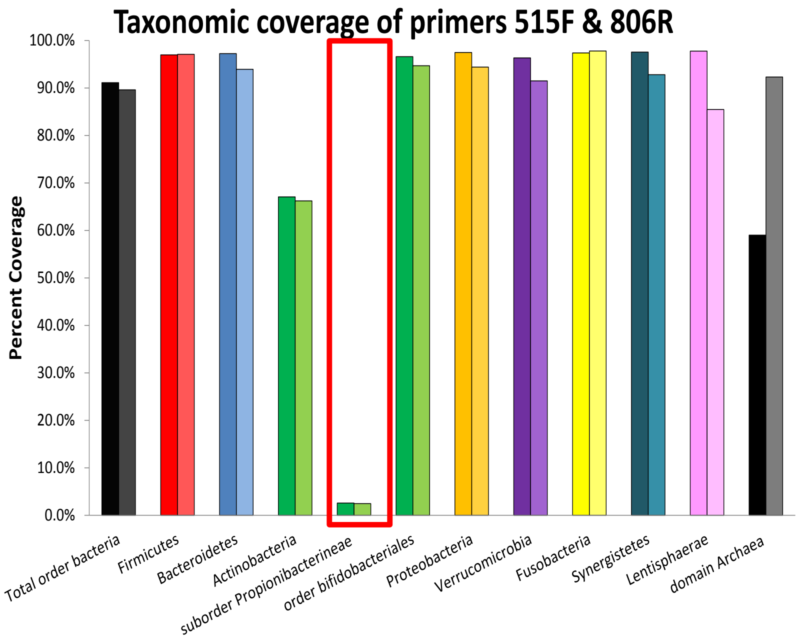
variability

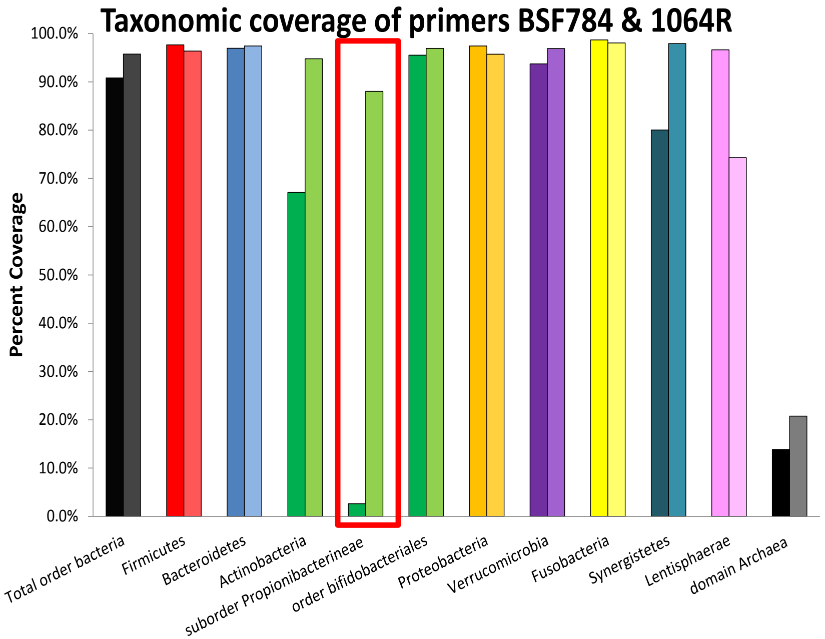
base position

2. Barcoded PCR: Primers

In silico Coverage

Check for your ecosystem of interest





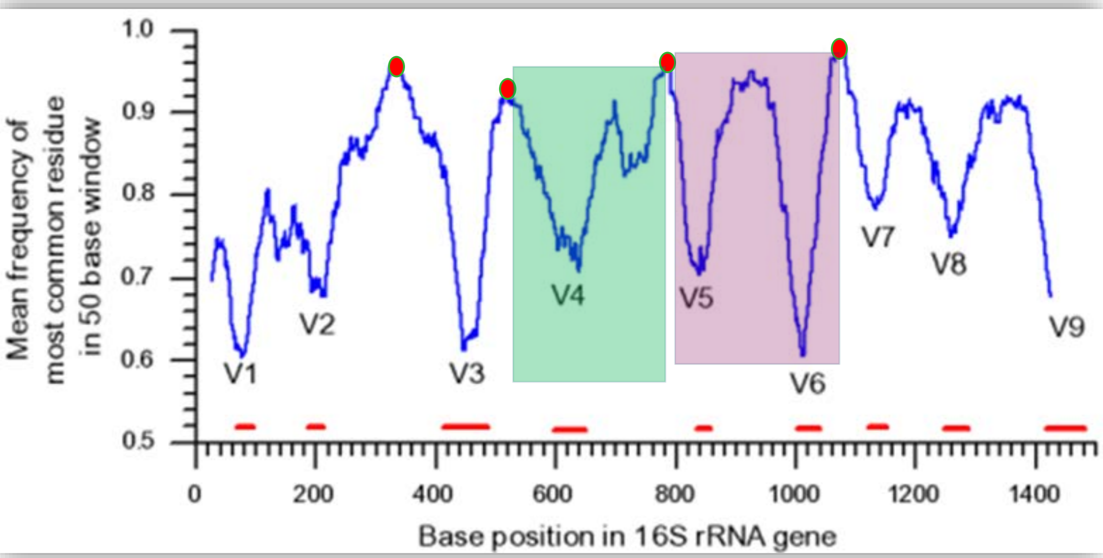
- Americans didn’t have *Bifidobacterium* up to ~2006
- No *Verrucomicrobia* in soil (actually 20%)

2. Barcoded PCR: Primers
Resolution

Unequal sequence conservation

Be very careful with species level classification

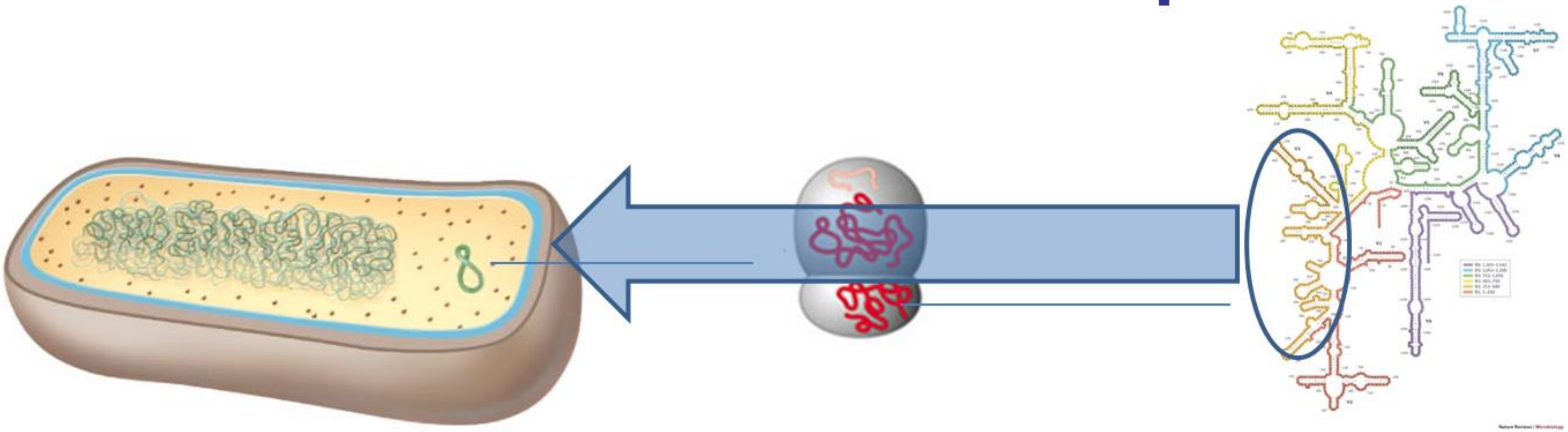
Up to genus level identification is recommended!



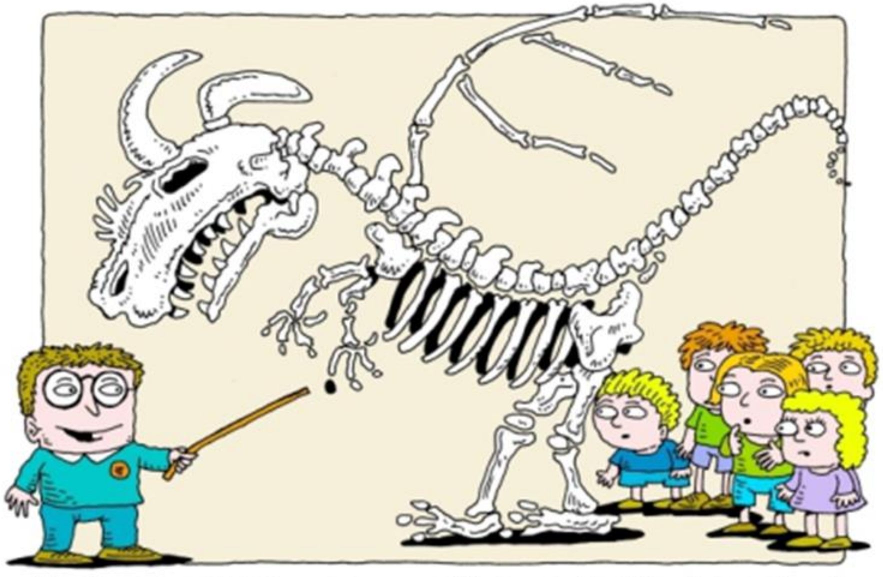
variability

base position

Be careful with functional interpretations!



Nature Reviews | Microbiology

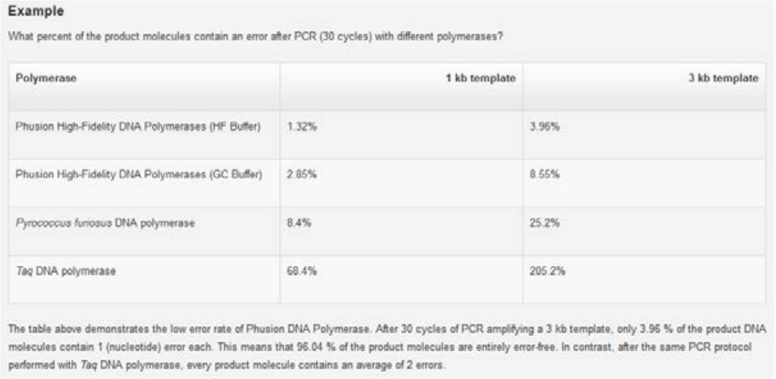


"Luckily we found THIS bone, so we were able to reconstruct the whole creature..."

2. Barcoded PCR: Primers

Bias

- Polymerase
- Primer-dimers
- Amplification bias: sequence of the bacteria themselves (high GC)
- Rate of chimera formation
- Sequencing specific sequence specific error rates
- Etc…

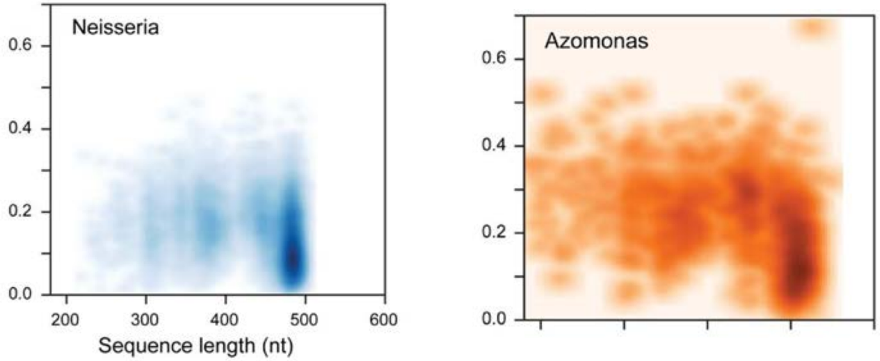


Example

What percent of the product molecules contain an error after PCR (30 cycles) with different polymerases?

Polymerase	1 kb template	3 kb template
Phusion High-Fidelity DNA Polymerases (HF Buffer)	1.32%	3.96%
Phusion High-Fidelity DNA Polymerases (GC Buffer)	2.85%	8.55%
Pyrococcus furiosus DNA polymerase	8.4%	25.2%
Taq DNA polymerase	68.4%	205.2%

The table above demonstrates the low error rate of Phusion DNA Polymerase. After 30 cycles of PCR amplifying a 3 kb template, only 3.96 % of the product DNA molecules contain 1 (nucleotide) error each. This means that 96.04 % of the product molecules are entirely error-free. In contrast, after the same PCR protocol performed with *Taq* DNA polymerase, every product molecule contains an average of 2 errors.



2. Barcoded PCR: Primers

No universal primer (yet ?)

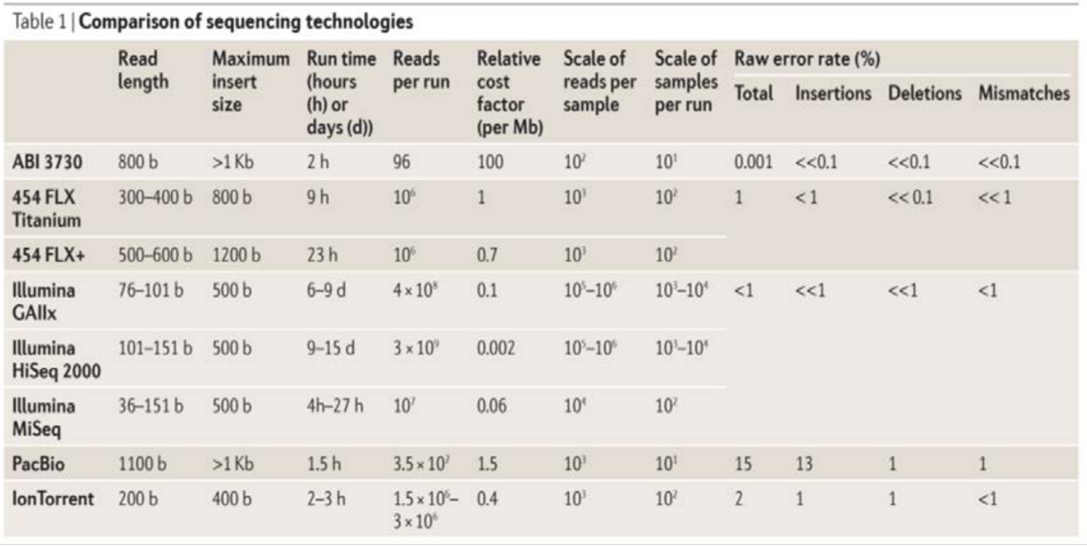
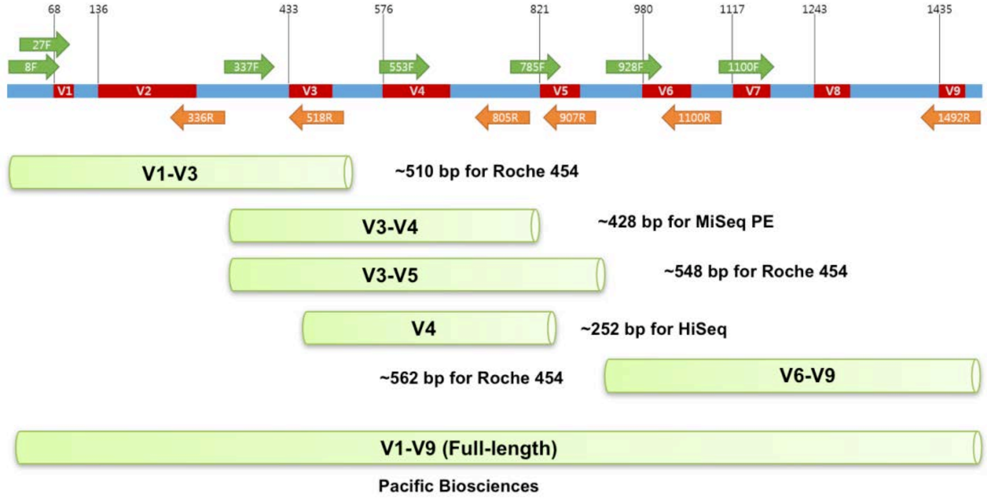


Table 1 | **Comparison of sequencing technologies**

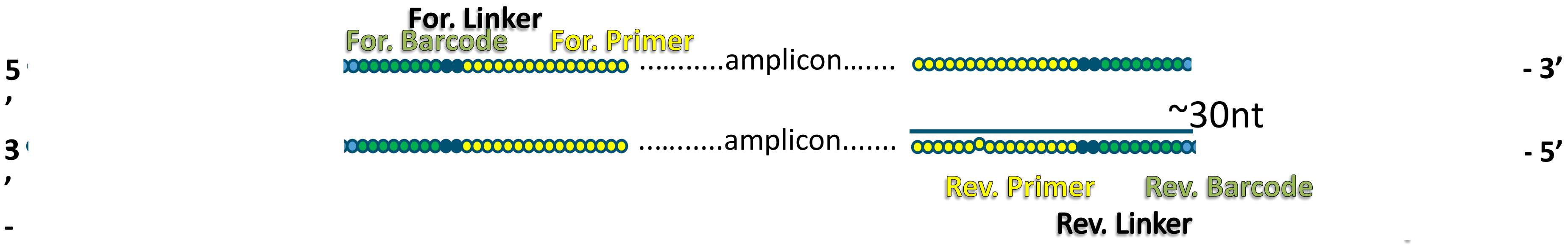
	Read length	Maximum insert size	Run time (hours (h) or days (d))	Reads per run	Relative cost factor (per Mb)	Scale of reads per sample	Scale of samples per run	Raw error rate (%)			
								Total	Insertions	Deletions	Mismatches
ABI 3730	800 b	>1 Kb	2 h	96	100	10^2	10^1	0.001	<<0.1	<<0.1	<<0.1
454 FLX Titanium	300–400 b	800 b	9 h	10^6	1	10^3	10^2	1	< 1	<< 0.1	<< 1
454 FLX+	500–600 b	1200 b	23 h	10^6	0.7	10^3	10^2				
Illumina GAIIx	76–101 b	500 b	6–9 d	4×10^8	0.1	10^5–10^6	10^3–10^4	<1	<<1	<<1	<1
Illumina HiSeq 2000	101–151 b	500 b	9–15 d	3×10^9	0.002	10^5–10^6	10^3–10^4				
Illumina MiSeq	36–151 b	500 b	4h–27 h	10^7	0.06	10^4	10^2				
PacBio	1100 b	>1 Kb	1.5 h	3.5×10^7	1.5	10^3	10^1	15	13	1	1
IonTorrent	200 b	400 b	2–3 h	1.5×10^6– 3×10^6	0.4	10^3	10^2	2	1	1	<1

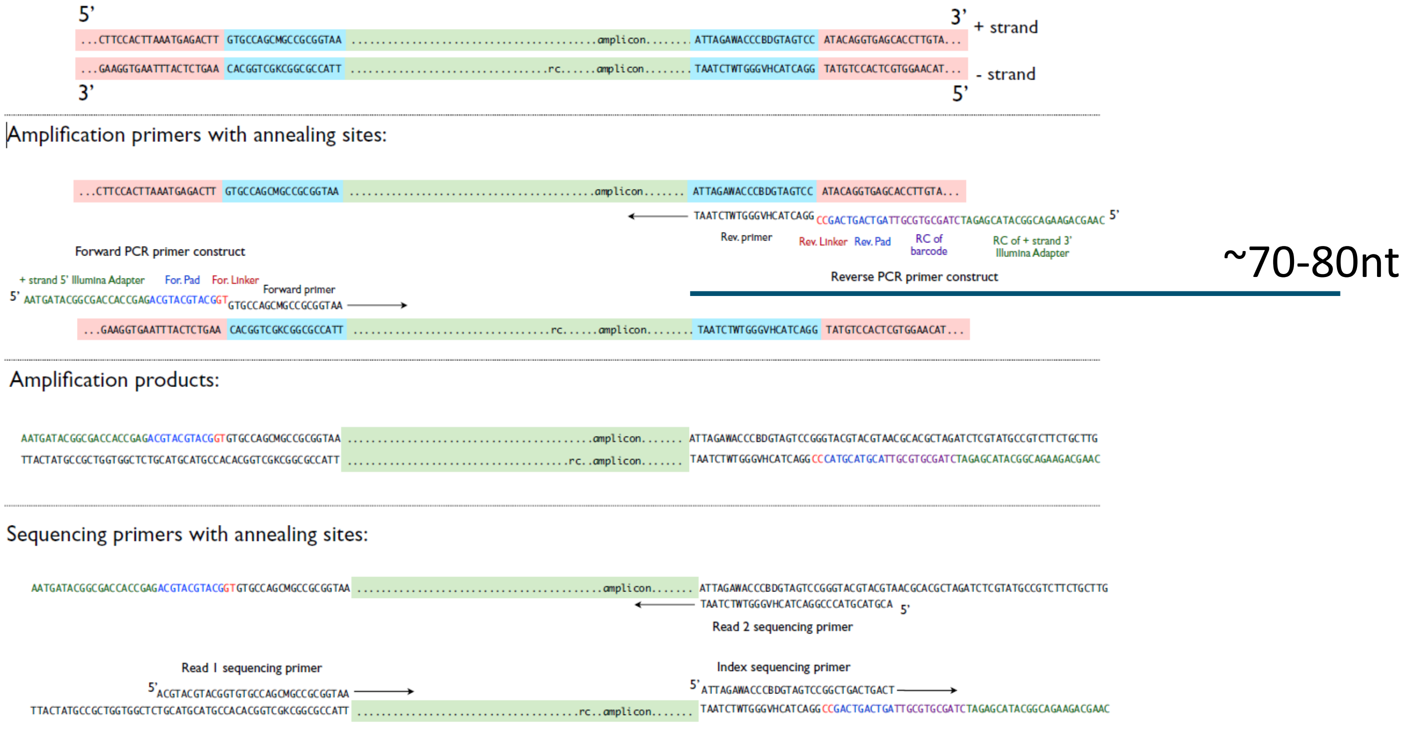
Kuczynski et al., Nature rev. 2011



WAGENINGE
UNIVERSITY & RESEAR

2. Barcoded PCR: barcoding strategy





```
5'                                                                                                                              3'
...CTTCCACTTAAATGAGACTT GTGCCAGCMGCCGCGGTAA ........................................amplicon....... ATTAGAWACCCBDGTAGTCC ATACAGGTGAGCACCTTGTA...   + strand
...GAAGGTGAATTTACTCTGAA CACGGTCGKCGGCGCCATT ...............................rc......amplicon....... TAATCTWTGGGVHCATCAGG TATGTCCACTCGTGGAACAT...   - strand
3'                                                                                                                              5'
```

Amplification primers with annealing sites:

```
...CTTCCACTTAAATGAGACTT GTGCCAGCMGCCGCGGTAA ........................................amplicon....... ATTAGAWACCCBDGTAGTCC ATACAGGTGAGCACCTTGTA...
                                                                                  ←——— TAATCTWTGGGVHCATCAGG CCGACTGACTGATTGCGTGCGATCTAGAGCATACGGCAGAAGACGAAC 5'
                                                                                       Rev. primer       Rev. Linker  Rev. Pad  RC of barcode  RC of + strand 3' Illumina Adapter
     Forward PCR primer construct
                                                                                                            Reverse PCR primer construct
 + strand 5' Illumina Adapter   For. Pad  For. Linker  Forward primer
5' AATGATACGGCGACCACCGAGACGTACGTACGGT GTGCCAGCMGCCGCGGTAA ———→
...GAAGGTGAATTTACTCTGAA CACGGTCGKCGGCGCCATT ...............................rc......amplicon....... TAATCTWTGGGVHCATCAGG TATGTCCACTCGTGGAACAT...
```

~70-80nt

Amplification products:

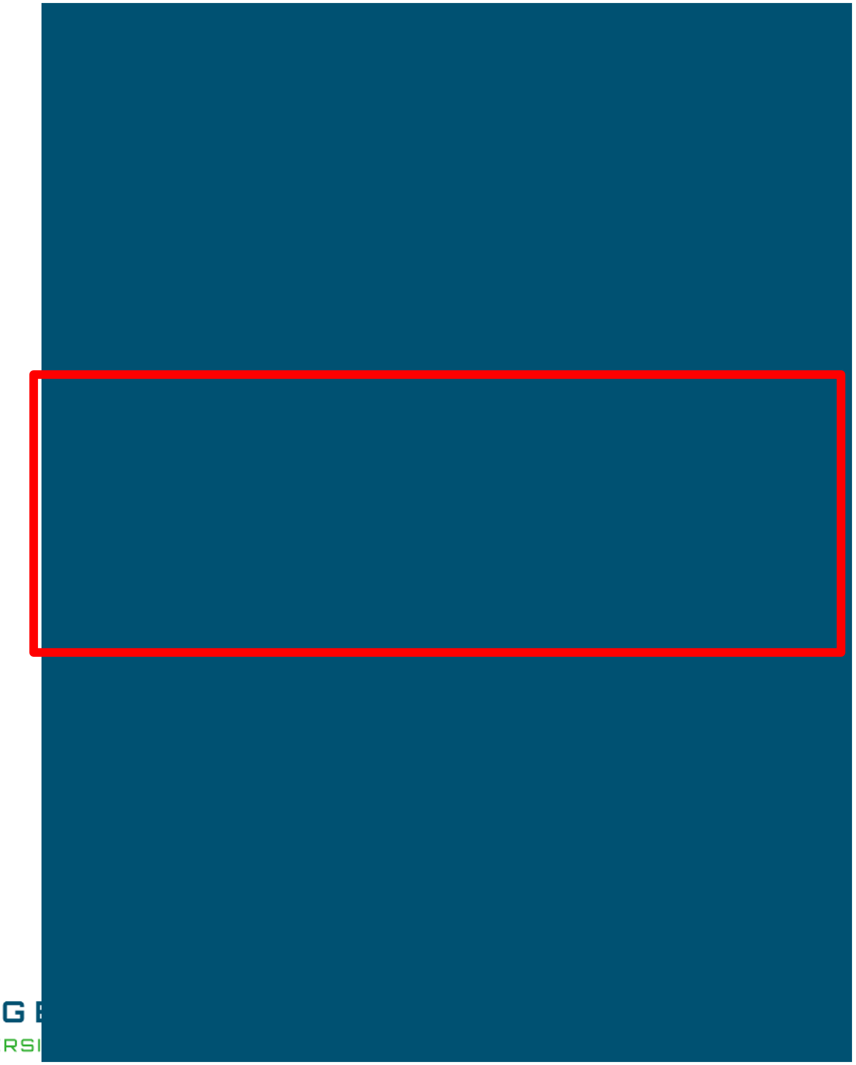
```
AATGATACGGCGACCACCGAGACGTACGTACGGTGTGCCAGCMGCCGCGGTAA ........................................amplicon....... ATTAGAWACCCBDGTAGTCCGGGTACGTACGTAACGCACGCTAGATCTCGTATGCCGTCTTCTGCTTG
TTACTATGCCGCTGGTGGCTCTGCATGCATGCCACACGGTCGKCGGCGCCATT ..................................rc..amplicon....... TAATCTWTGGGVHCATCAGGCCCATGCATGCATTGCGTGCGATCTAGAGCATACGGCAGAAGACGAAC
```

Sequencing primers with annealing sites:

```
AATGATACGGCGACCACCGAGACGTACGTACGGTGTGCCAGCMGCCGCGGTAA ........................................amplicon....... ATTAGAWACCCBDGTAGTCCGGGTACGTACGTAACGCACGCTAGATCTCGTATGCCGTCTTCTGCTTG
                                                                                   ←——— TAATCTWTGGGVHCATCAGGCCCATGCATGCA 5'
                                                                                         Read 2 sequencing primer

                  Read 1 sequencing primer                                               Index sequencing primer
              5'ACGTACGTACGGTGTGCCAGCMGCCGCGGTAA ———→                                 5'ATTAGAWACCCBDGTAGTCCGGCTGACTGACT ———→
TTACTATGCCGCTGGTGGCTCTGCATGCATGCCACACGGTCGKCGGCGCCATT ..................................rc..amplicon....... TAATCTWTGGGVHCATCAGGCCGACTGACTGATTGCGTGCGATCTAGAGCATACGGCAGAAGACGAAC
```

Analysis pipeline



CLC Genomics Workbench

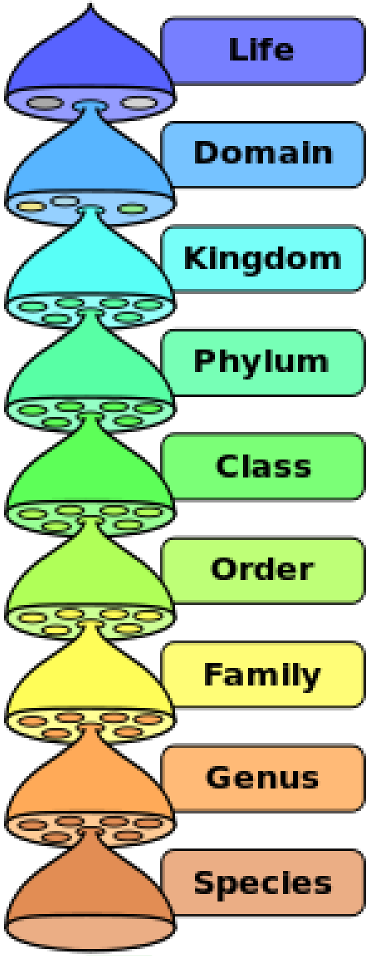
QIIME

Mothur

DADA2

Etc.........

Analysis pipeline
OTU picking



Operational Taxonomic Unit

- traditionally clustered @ 97% -> ‘species’ proxy
- Now : ‘sequences’
- Summarize to genus level

~88%

~92%

~95%

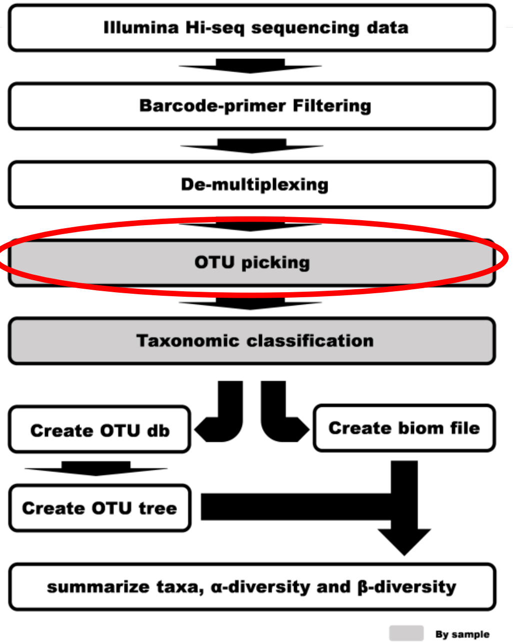
Escherichia Coli & Salmonella >98% similar

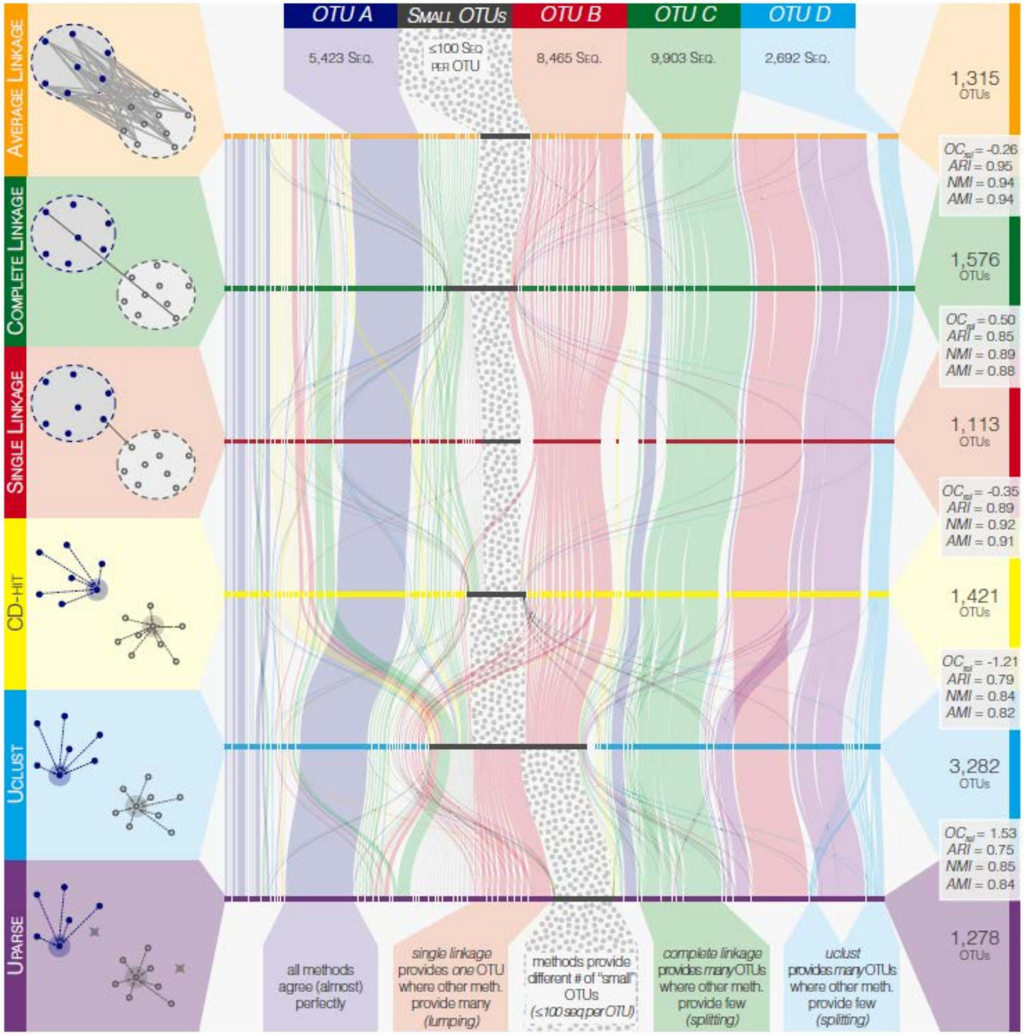
~97%

Horizontal gene transfer

Analysis pipeline:

OTU picking





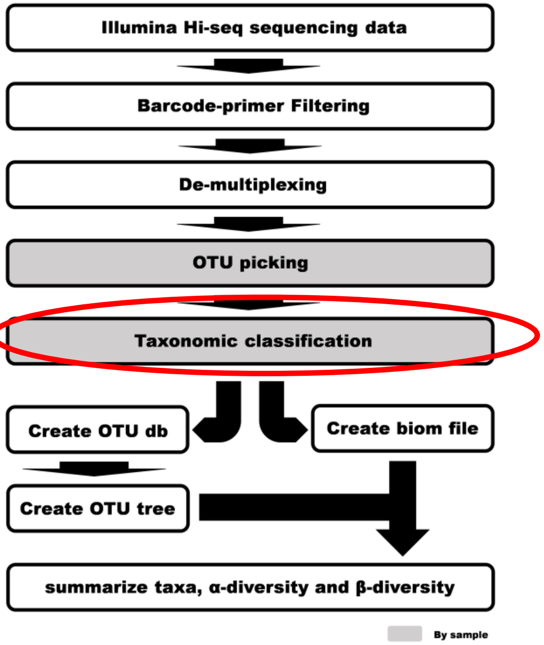
97%

ered

Schmidt et al 2014, Limits to robustness and reproducibility in the demarcation of operational taxonomic units. Env micr.

Analysis pipeline

Taxonomic classification



Different methods of classification -> different name (scoring system)

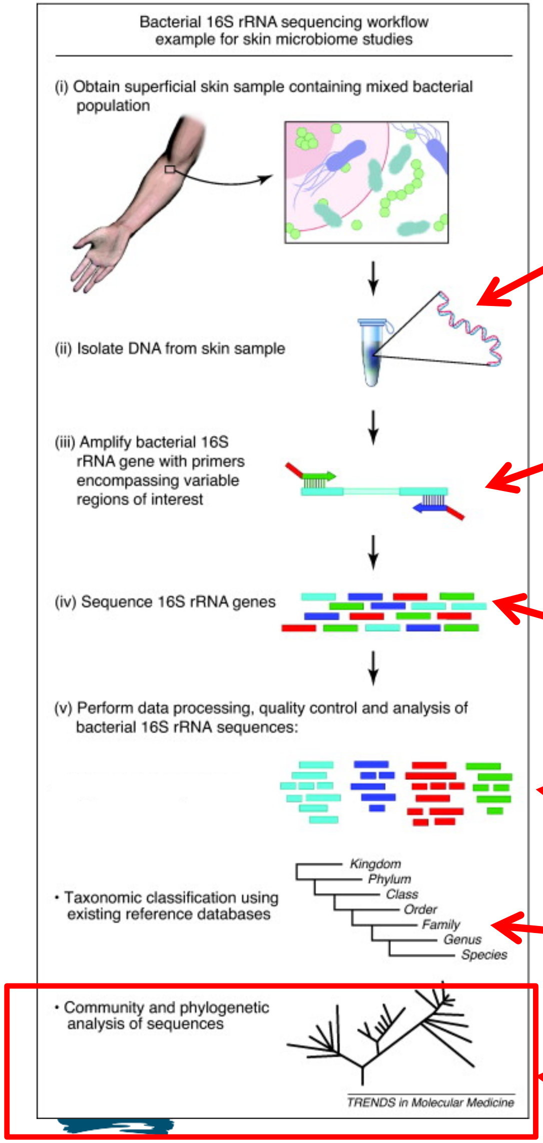
- BLAST
- Kmer
- MisMatches

16s rRNA gene Databases

- Ribosomal Database Project (RDP-II)
- ~~Greengenes~~ (13_05) (poor alignment/outdated)
- SILVA

WAGENINGEN
UNIVERSITY & RESEARCH

Challenges



Bacterial 16S rRNA sequencing workflow example for skin microbiome studies

(i) Obtain superficial skin sample containing mixed bacterial population

(ii) Isolate DNA from skin sample

(iii) Amplify bacterial 16S rRNA gene with primers encompassing variable regions of interest

(iv) Sequence 16S rRNA genes

(v) Perform data processing, quality control and analysis of bacterial 16S rRNA sequences:

- Taxonomic classification using existing reference databases
- Community and phylogenetic analysis of sequences

TRENDS in Molecular Medicine

Important!
Bias (might miss bacteria)
Can overshadow biology

Important!
Bias (eg. coverage, amplification bias)
Can overshadow biology

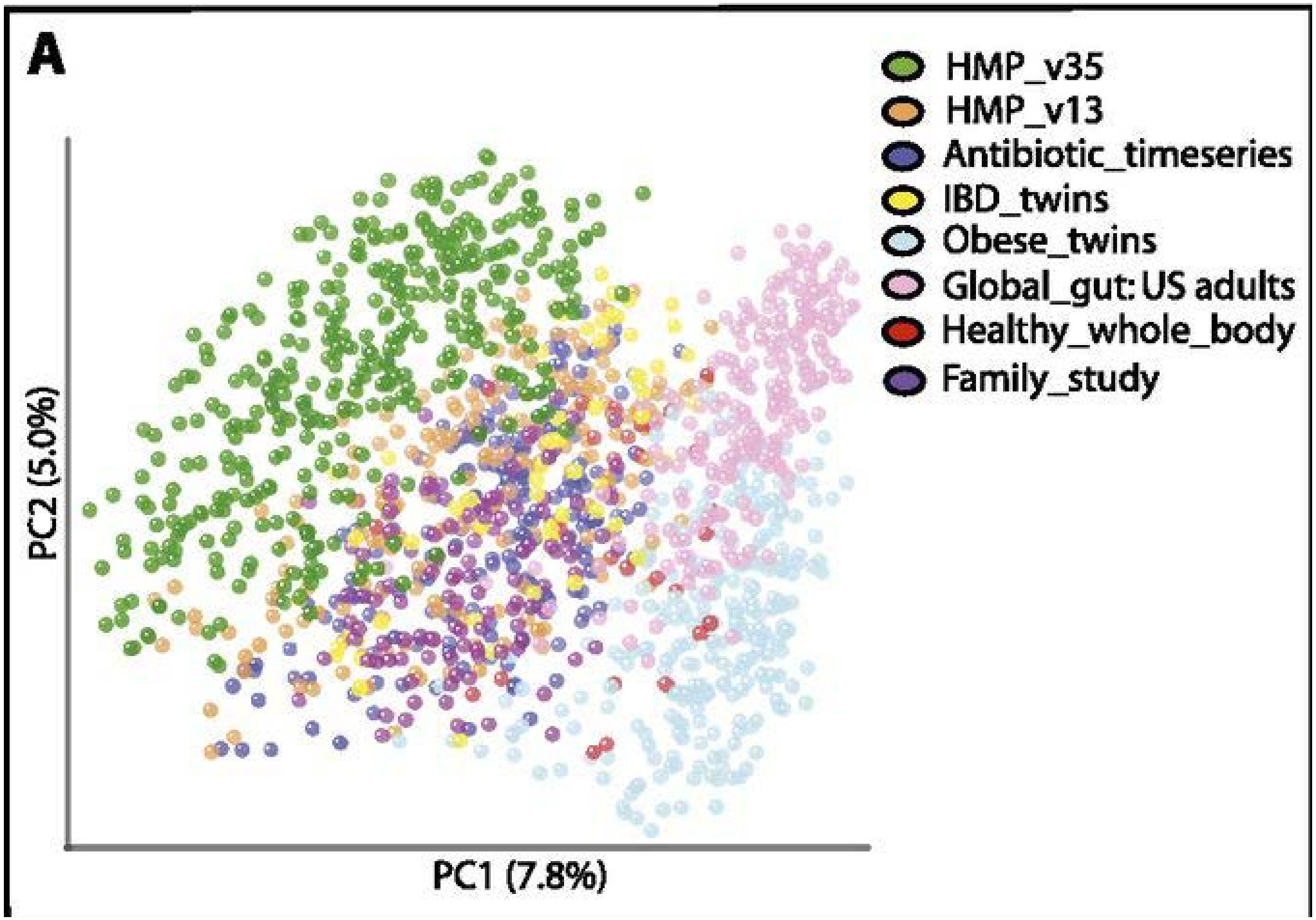
Important!
Bias (Sequencing error (type **and** machine specific))

Important!
Bias (OTU clustering)

Important!
Bias (classifier and database (different names))

The hard part

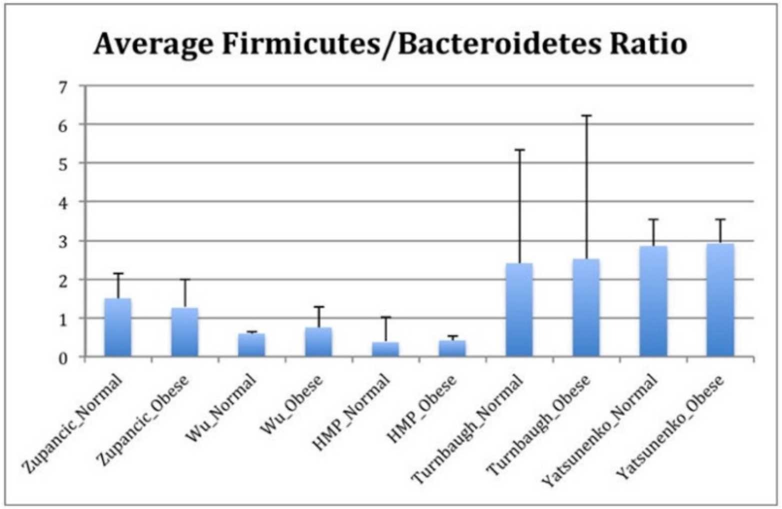
Effects of all these different parameters in literature

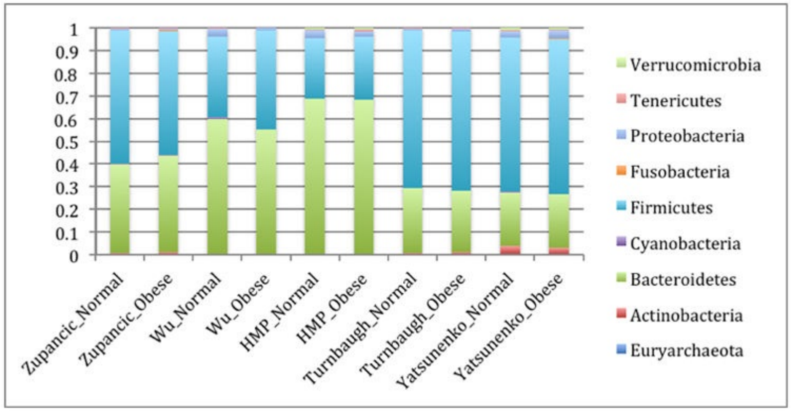


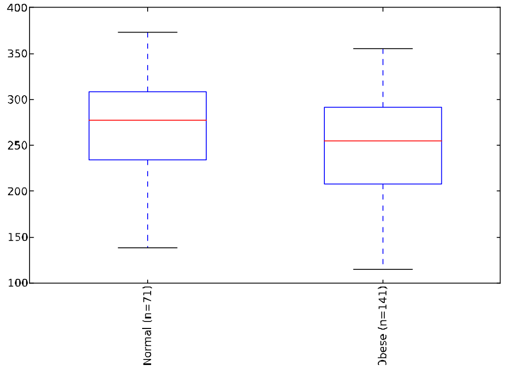
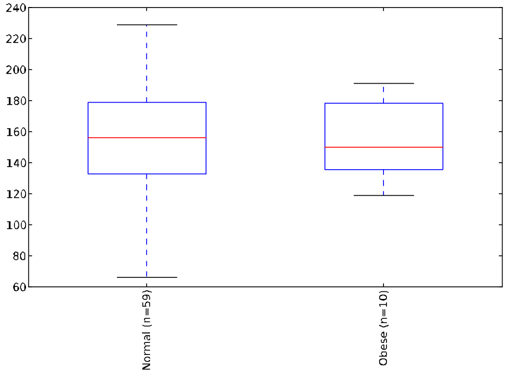
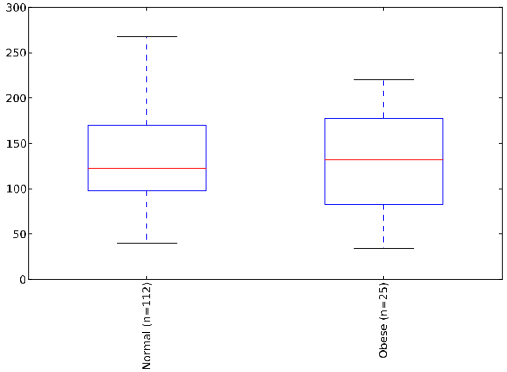
Catherine A. Lozupone et al. Genome Res. 2013;23:1704-1714

WAGENINGEN
UNIVERSITY & RESEARCH

Microbial biomarkers

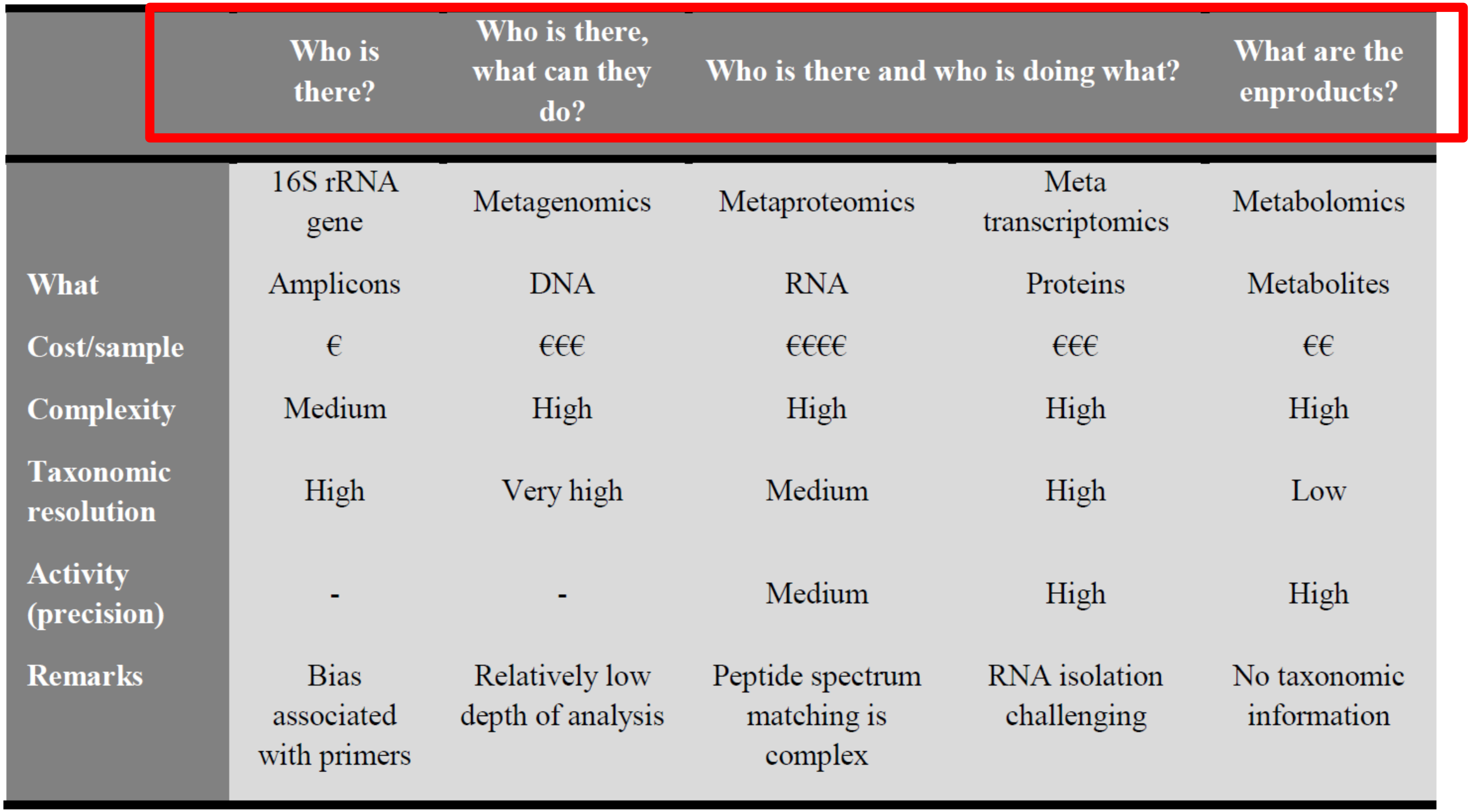




Study	Observed species plot for normal and obese individuals	p-value
Zupancic	 	0.015
Wu	 	0.95
HMP	 	0.955

Walters et al. Meta-analyses of human gut microbes associated with obesity and IBD. FEBS Lett 2014

Pros & cons



	Who is there?	Who is there, what can they do?	Who is there and who is doing what?		What are the enproducts?
	16S rRNA gene	Metagenomics	Metaproteomics	Meta transcriptomics	Metabolomics
What	Amplicons	DNA	RNA	Proteins	Metabolites
Cost/sample	€	€€€	€€€€	€€€	€€
Complexity	Medium	High	High	High	High
Taxonomic resolution	High	Very high	Medium	High	Low
Activity (precision)	-	-	Medium	High	High
Remarks	Bias associated with primers	Relatively low depth of analysis	Peptide spectrum matching is complex	RNA isolation challenging	No taxonomic information

Keep in mind/challenges/interpretation

OTUs

- Prokaryotic species level boundary is still hotly debated
- Species level concept for organisms with mobile genetic elements....
- Function & 16S are ‘only’ correlated (sometimes actually not that much)

- E coli: probiotic (*nissle*), pathogen (O157:H7 acute haemorrhagic colitis, Shiga-toxin) & everything in between (K12)
- Most bacteria are opportunists

Recap