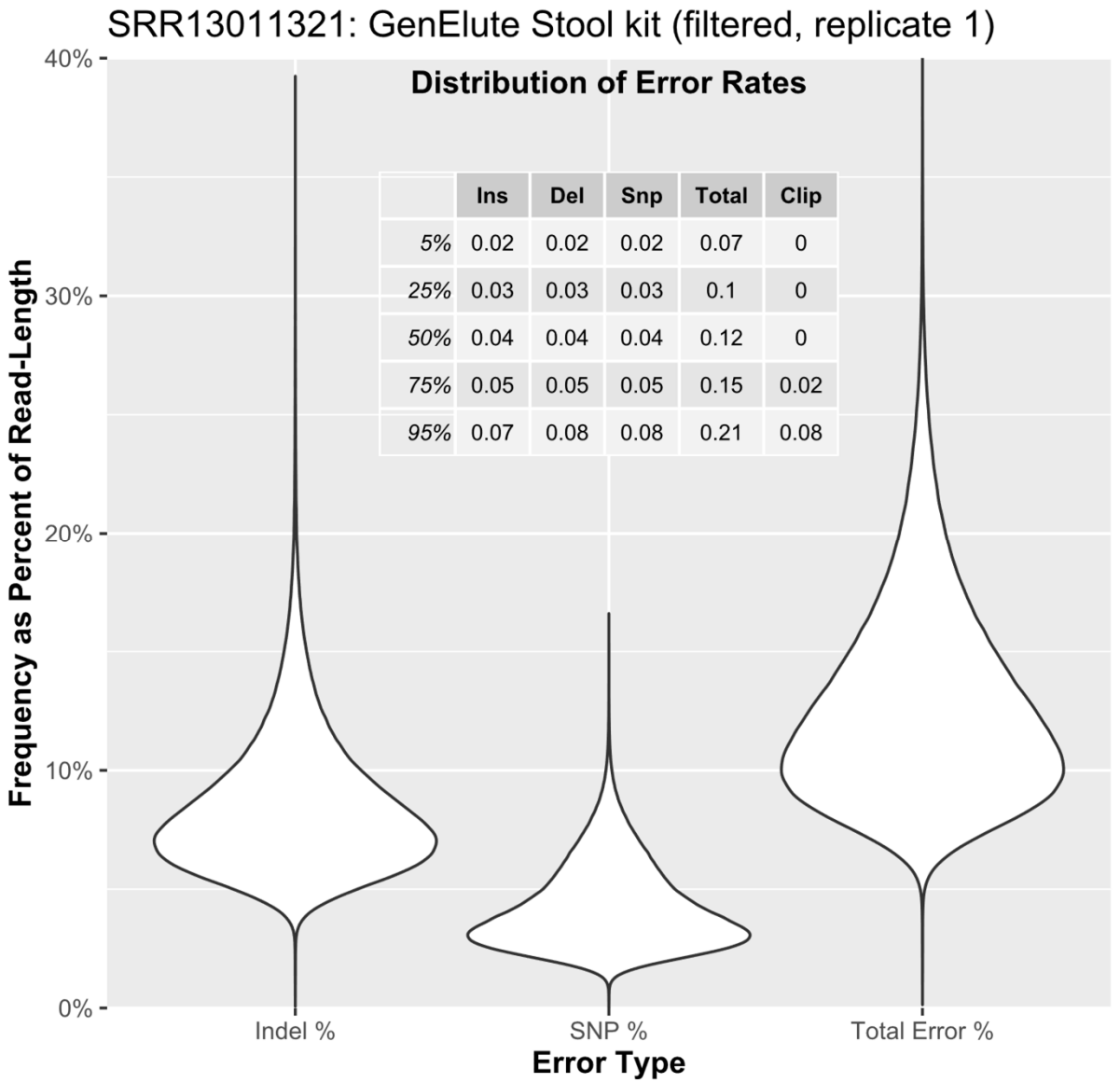
Kick off topic:
Taxon Abundance with Full-Length 16S ONT Reads

Not the HiFi ones, the noisy stuff…

ONT reads still have high error

- Data from Mann, et al. (2023)[1]
  - 16S amplicons with 27F/1492R primers
  - Multiple extraction protocols
    - *Only GenElute shown at left*[2]
    - *Others have nearly identical error rates*
  - Zymo standard, sequenced on MinION
  - R9.4.1 chemistry
  - Base-called with Guppy 3.2.4
  - Reads filtered with NanoFilt
- Median total error rate: 12%
  - 90% of reads between 7-21% error as percent of read-length
  - *Not* including soft-clipping, which remains in some reads even post-filtering.



	Ins	Del	Snp	Total	Clip
5%	0.02	0.02	0.02	0.07	0
25%	0.03	0.03	0.03	0.1	0
50%	0.04	0.04	0.04	0.12	0
75%	0.05	0.05	0.05	0.15	0.02
95%	0.07	0.08	0.08	0.21	0.08

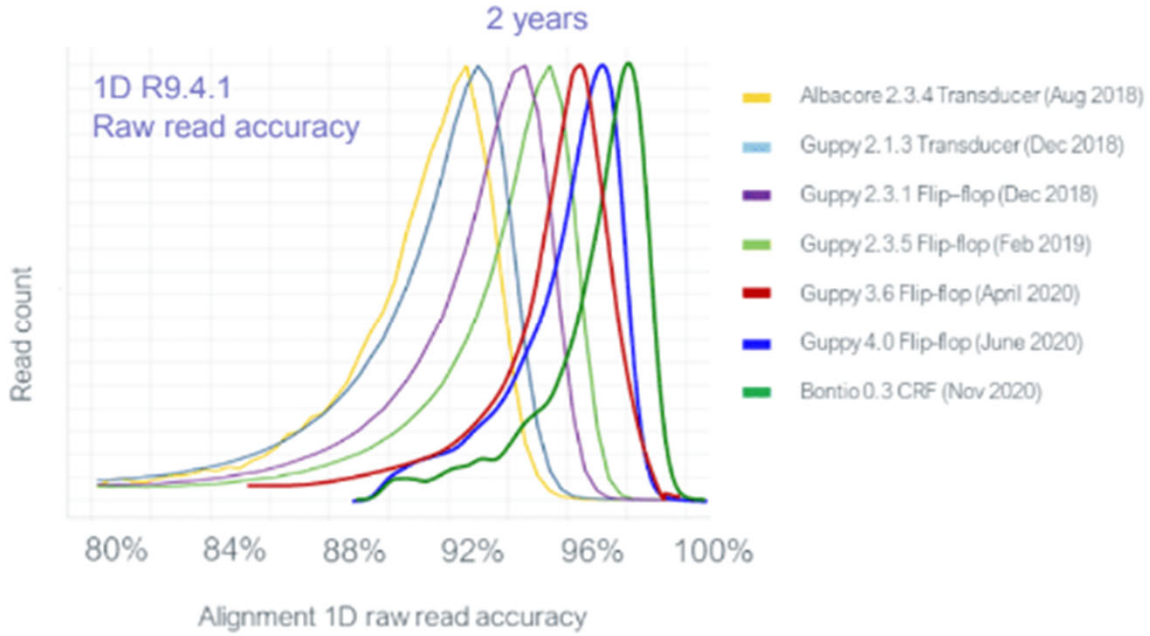
[1]*DOI: 10.1016/j.dib.2021.107036*2

[2]*As a sidenote, one of the kits was the QIAamp DNA Microbiome Kit which appeared to **really** mess up the profile of microbes recovered somehow.*

Errors coming down, but still material

- R10 vs. R9 chemistry:
  - Zhang, et al (2023)[3] reports error rates on Zymo Standard
  - Qiagen DNeasy PowersoilPro, PromethION, Guppy v6,
  - R9.4.1 and R10 chemistry
  - *(Data is non-public)*
- Error Rates[4]:
  - 9% total error with Guppy v6 vs. 11-12% ca. 2020

Chem.	Samp.	SNP	Ins.	Del.	Total
R9.4.1	S1	3.1%	1.5%	2.6%	7.2%
	S2	3.5%	1.5%	2.6%	7.6%
	Zymo	4.5%	1.6%	2.8%	8.8%
	All	3.6%	1.5%	2.7%	7.9%
R10.4.1	S1	1.5%	0.4%	0.6%	2.4%
	S2	1.8%	0.4%	0.6%	2.9%
	Zymo	2.9%	0.6%	0.7%	4.2%
	All	2.1%	0.5%	0.6%	3.2%



Basecalling has been improving, but true error rates still have a ways to go…

Image credit: Clive Brown via @nanopore on ~~Twitter~~X

[3]*DOI: 10.1128/aem.00605-23*

[4]*This is a subset of Table 2 from Zhang, et al. Zymo is the Zymo Standard D6305, S1/S2 are other samples*

16S Taxonomic Profiling: Read-Length vs. Error-Rate

- Relative Abundance from 16S Amplicons:
  - First 16S taxon identification was accomplished by the ancient Romans *citation needed*
- Illumina Reads:
  - ***Short reads*** ⟹ limited power to discriminate at species level (at least without extra effort)
  - ***Low error*** ⟹ classification generally accurate
  - ∗ *Recall/Sensitivity problem*
- Nanopore Reads:
  - ***Full length 16S*** ⟹ Species-level classification
  - ***High error rate*** ⟹ Cross confusion common with near relatives
  - ∗ *Precision/Specificity problem*

› Microbiome. 2020 May 15;8(1):65. doi: 10.1186/s40168-020-00841-w.

Construction of habitat-specific training sets to achieve species-level assignment in 16S rRNA gene datasets

Isabel F Escapa [1 2 3], Yanmei Huang [1 2], Tsute Chen [1 2], Maoxuan Lin [1], Alexis Kokaras [1], Floyd E Dewhirst [1 2], Katherine P Lemon [4 5 6 7]

Affiliations + expand

PMID: 32414415 PMCID: PMC7291764 DOI: 10.1186/s40168-020-00841-w

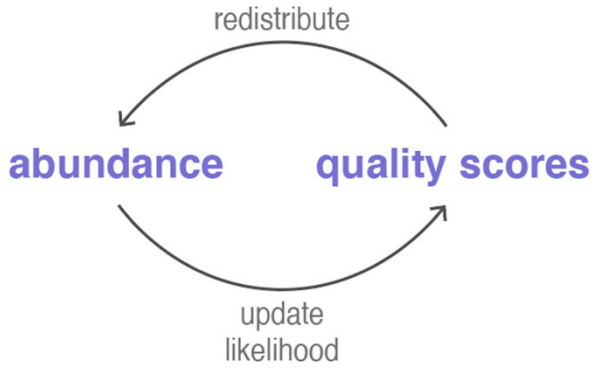
Free PMC article

This is one of the few references that attempts species-level classification based on 16S sequencing using Illumina sequences…

…does it by designing domain-specific database.

Emu: Species-level Taxon ID for full-length 16S reads

The EM algorithm corrects taxon assignment errors by considering prior likelihood that read was mis-assigned to a close relative.



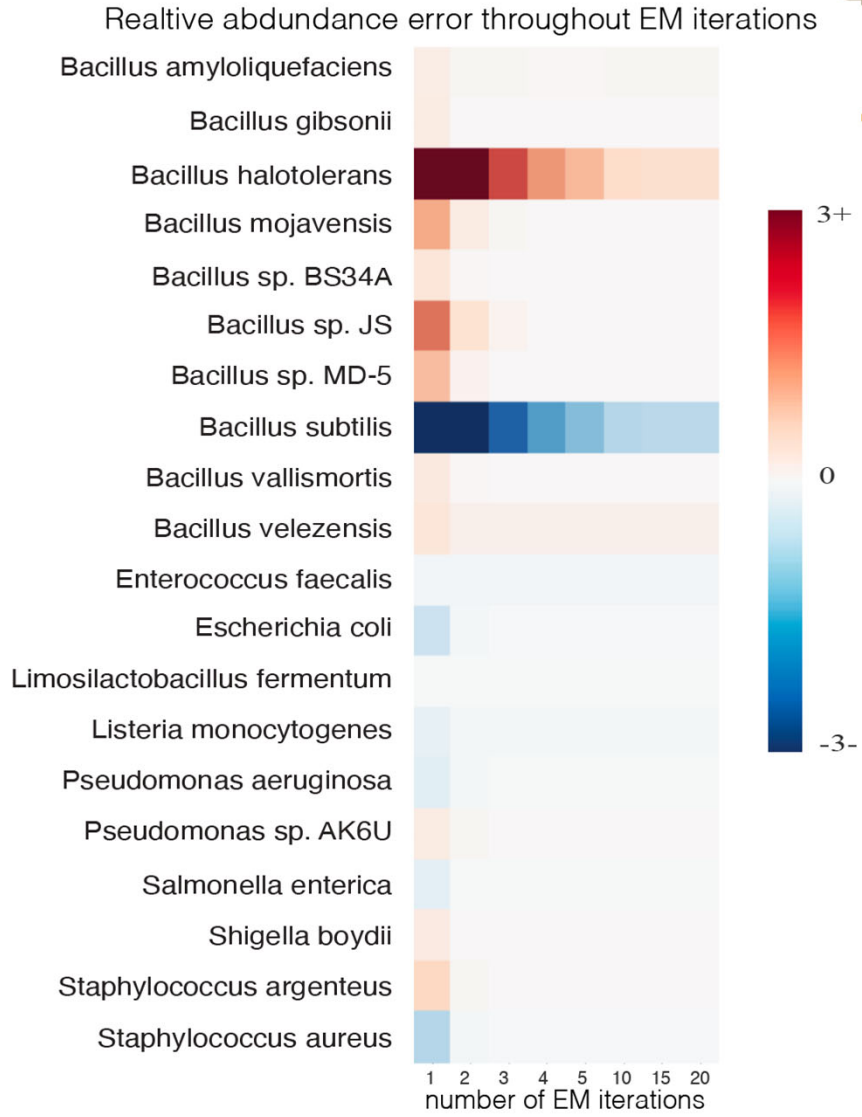
Prob. that read r comes from taxon t.

Likelihood of seeing read r when actual sequence is taxon t.
(Entirely driven by ONT error rates.)

Likelihood function:

$$P(t|r) = \frac{L(r|t)\pi(t)}{\sum_{t \in T} L(r|t)\pi(t)}$$

Prior probability of the read coming from taxon t (i.e. current estimate of abundance profile)



In collaboration with: Alexander Dilthey and Alona Tyshaieva, Institute of Medical Microbiology and Hospital Hygiene, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

Slide Credit: Kristen Curry

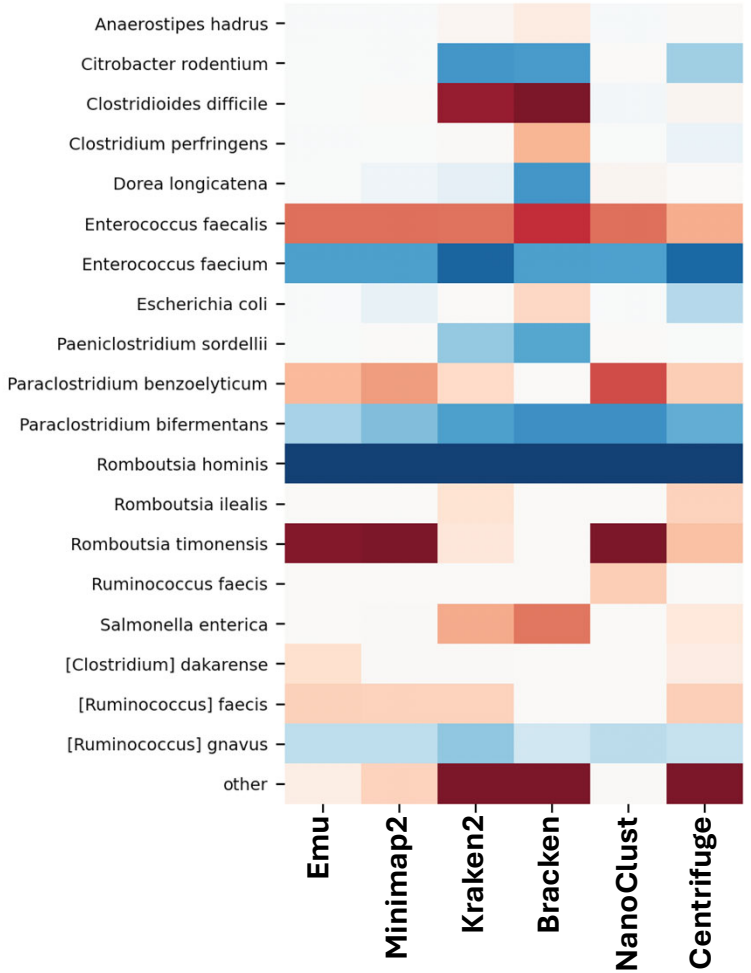
Example Results: Species-Level Relative Abundance

Synthetic Gut Community

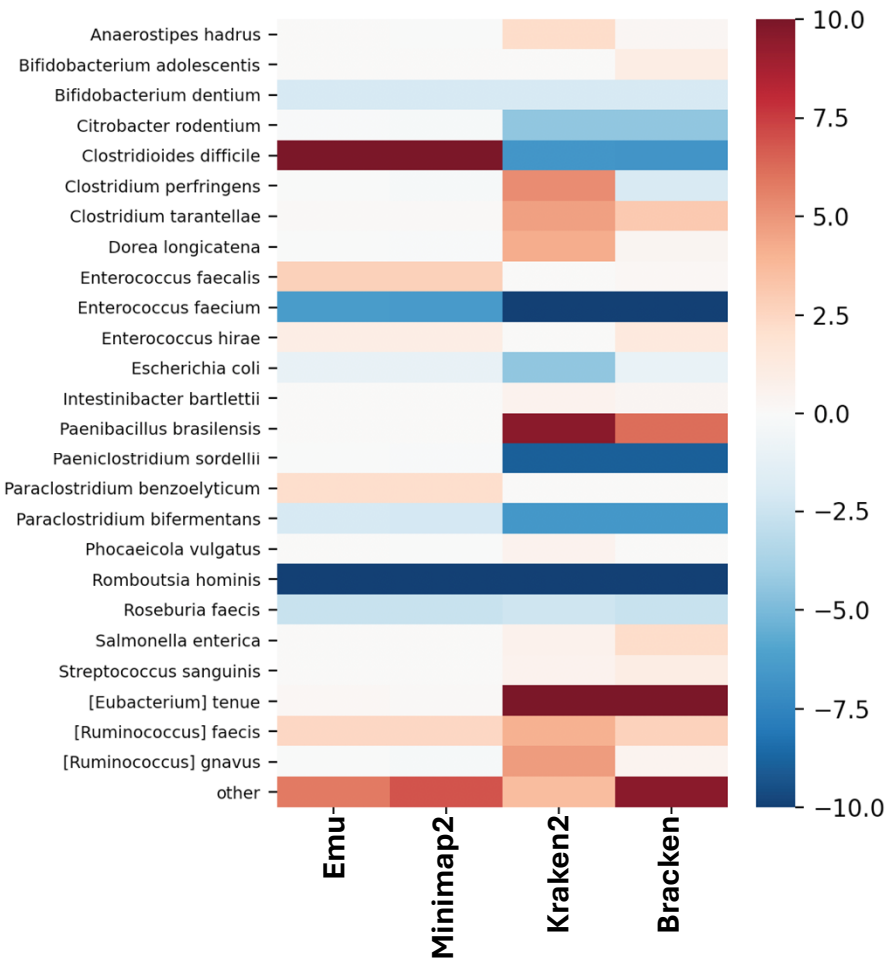
Approximate Ground Truth Abundance

Strain	CFU	Pct
Bifidobacterium dentum ATCC 27678	20000	18.9
Entercocccus faecium S66643	20000	18.9
Citribacter rodentium ATCC 51459	13000	12.3
Gemmiger formicilis ATCC 27749	10000	9.4
Escherichia coli MG1655	10000	9.4
Clostridium perfringens MT676	9000	8.5
Romboutsia hominis FRIFI	7000	6.6
Clostridium leptum ATCC 29065	5000	4.7
Clostridium scindens ATCC 35704	4000	3.7
Ruminococcus gnavus ATCC 29149	3000	2.8
Bacteroides vulgatus PC510	2000	1.9
Clostridium bartlettii DSM 16795	600	0.5
Clostridum innocuum ATCC 14501	440	0.4
Clostridium bifermentans ERIN_30100	300	0.2
Clostridium sordellii ATCC 9714	240	0.2
Bacteroides thetaiotaomicron VPI -5482	200	0.19
Eubacterium hadrum DSM 3319	200	0.19
Clostridium difficile 630	200	0.19
Dorea longicatena DSM 13814	100	0.095
Roseburia faecis DSM 16840 (M72)	4	0.004

Nanopore

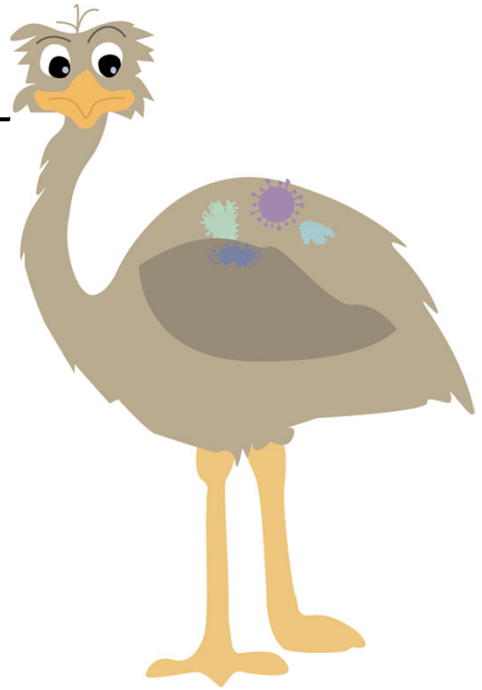


Illumina



...note that Emu does best here but it was also the only purpose-built 16S ONT method at the time (2021)

Emu Conclusions



- **Download:** https://github.com/treangenlab/emu
- Purpose-built for full-length, high-error 16S amplicon reads
  - I.e. ONT reads, especially R9 chemistry, even with old base-caller
  - Was the first algorithm *really* for this
    - Some newer ones now but we haven't evaluated
- Advantages (big):
  - Relative abundance accuracy (★ ★ ★)
  - Lower false positives (★ ★ ★ ★)
- Disadvantages (no big deal):
  - Some extra run time vs. just minimap2
  - Abundance below some minimum (e.g. 0.5%) is set to 0,
    - So not ideal for estimating very-low abundance members
- Extension to shotgun data via Lemur/Magnet[5] (pre-printed)

[5]*Image Credit: DALL-E*