

Khoa luận tốt nghiệp

Đề tài

Đánh giá quy trình phân tích dữ liệu giải trình tự đa hệ gen (*shotgun metagenomics*) và áp dụng phân tích đặc điểm hệ vi sinh vật đường ruột giữa mẹ và trẻ sơ sinh nhiễm trùng huyết

THÔNG TIN

Họ và tên

Tôn Ngọc Minh Quân

MSSV

19180142

Giảng viên hướng dẫn

TS. Phạm Thanh Duy

Nội dung trình bày

1. Tổng quan
2. Mục tiêu
3. Thiết kế thí nghiệm
4. Kết quả
5. Kết luận & Kiến nghị

TỔNG QUAN

Nhiễm trùng huyết sơ sinh và tình hình dịch tễ



Tình trạng da ở cẳng chân và bàn chân phải của bệnh nhân sốc nhiễm trùng với nhiều vùng ban xuất huyết (Manson's Tropical Disease)

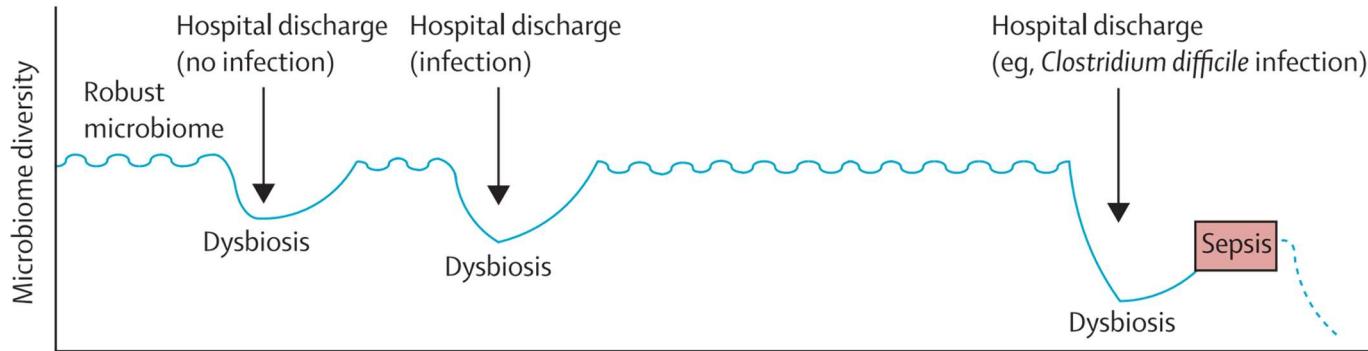
ID bệnh nhân	Nơi thu mẫu	Tác nhân gây bệnh
N_123	NICU	<i>E. cloacae</i>
N_134	NICU	<i>E. cloacae</i>
N_25	NICU, PICU	<i>E. kobei</i>
N_26	NICU	<i>E. kobei</i>
N_28	NICU	<i>E. kobei</i>
N_30	NICU	<i>E. kobei</i>
N_31	NICU	<i>E. kobei</i>
N_33	NICU	<i>E. kobei</i>
N_GMM	Nursery	<i>E. kobei</i>
N_NN	Nursery	<i>E. kobei</i>
N_RK	NICU	<i>E. xiangfangensis, E. mori</i>
N_RPG	NICU	<i>E. kobei</i>
N_SB	NICU	<i>E. kobei</i>
N_ST	NICU	<i>E. kobei</i>

Tác nhân gây nhiễm trùng huyết sơ sinh trong đợt dịch bùng phát ở
bệnh viện Patan, Nepal giai đoạn 5/2016 - 12/2017
(10.1016/j.jhin.2022.03.015)

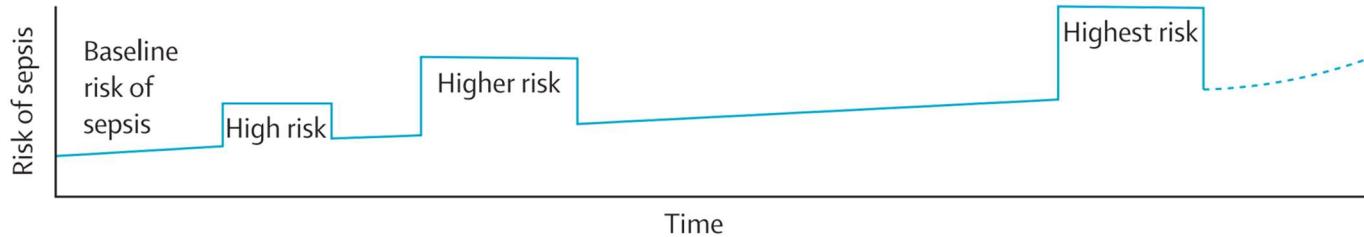
TỔNG QUAN

Nhiễm trùng huyết và hệ vi sinh đường ruột

A



B

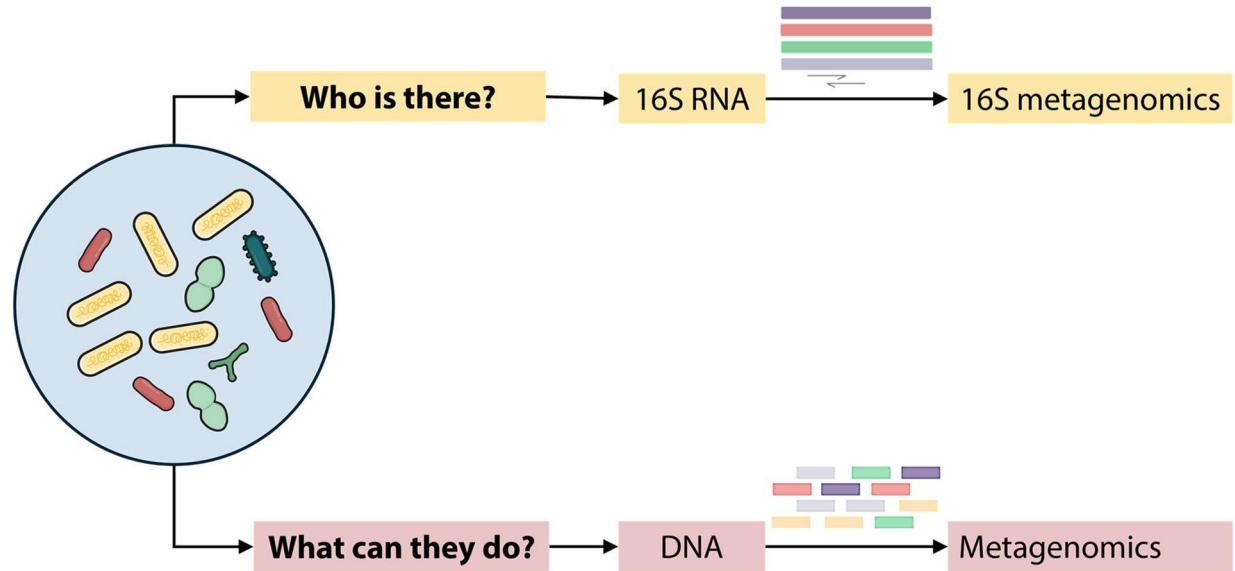


Sơ đồ mô tả mối quan hệ giữa (a) đặc điểm của hệ vi sinh đường ruột đối với nhiễm trùng huyết, và (b) nguy cơ mắc nhiễm trùng huyết giữa các lần rối loạn vi sinh đường ruột (10.1016/S2468-1253(16)30119-4)

TỔNG QUAN

Shotgun metagenomics

- 16S kém hiệu quả trong việc phân loại các loài có **độ tương đồng** cao và dễ sai lệch do bị phụ thuộc vào **thiết kế mồi**.
- Shotgun metagenomics có khả năng cung cấp một **bức tranh tổng thể** của cộng đồng

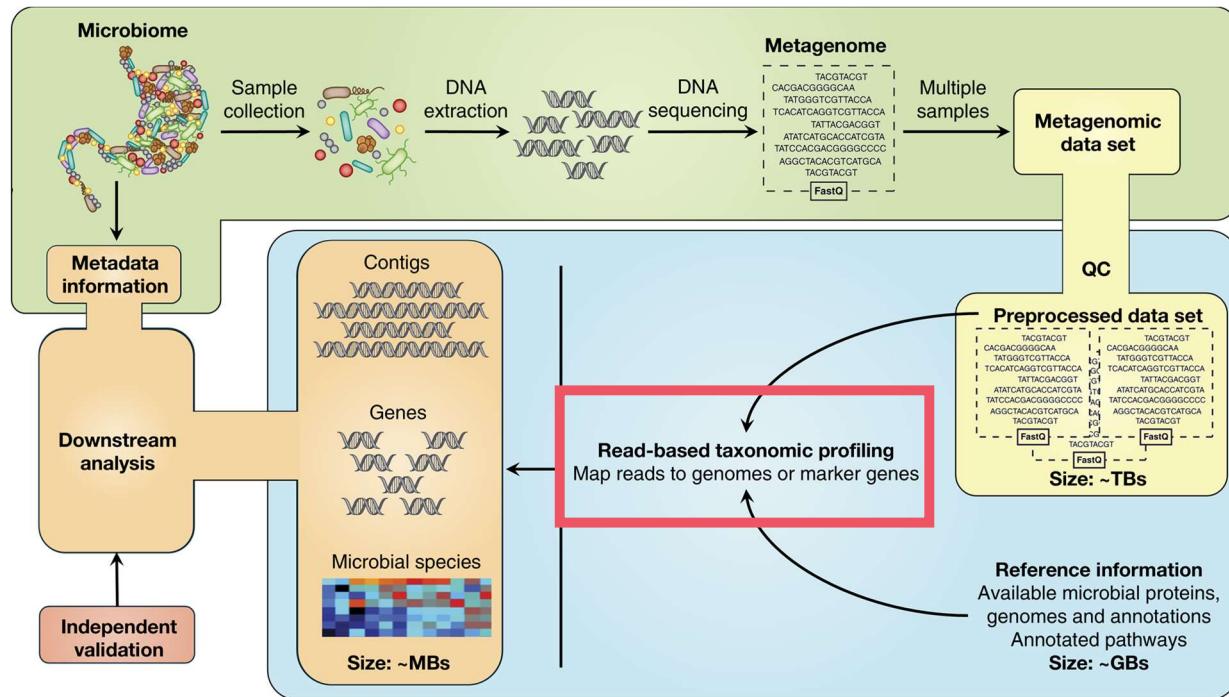


Các phương pháp tiếp cận dữ liệu khác nhau giúp trả lời các câu hỏi liên quan đến đặc điểm của quần thể vi sinh cụ thể (Vẽ lại từ 10/gdq95k)

TỔNG QUAN

Shotgun metagenomics

- Quy trình làm việc với dữ liệu metagenomics rất phức tạp, **dễ bị sai lệch và sai sót ở nhiều bước.**
- Bước **phân tích** bằng công cụ tin sinh học có khả năng gây sai lệch về kết quả.
- Các **công cụ phân loại loài** cho kết quả biến động nhất.
- Cần phải có **tiêu chuẩn hoá**.



Tóm tắt các bước chuẩn bị, thu nhận và xử lý dữ liệu metagenomics
(Vẽ lại từ 10.1038/nbt.3935)

MỤC TIÊU NGHIÊN CỨU

Câu hỏi đặt ra

- Công cụ nào cho khả năng phân loại loài chính xác và ổn định nhất?
- Công cụ nào phản ánh đúng/gần với thành phần vi sinh có trong mẫu nhất?
- Ở các điều kiện sử dụng khác nhau, liệu công cụ đã chọn có còn tốt hay không?
- Kết quả của công cụ trên các bộ mẫu thực tế như thế nào?

CÔNG CỤ NÀO ĐÃ ĐƯỢC CHỌN?

Chúng ta đang có những công cụ nào?

Dựa vào trình tự DNA

DNA - DNA

Kraken2
Bracken

MegaBLAST

KrakenUniq

CLARK-S

sourmash

yacht

PathSeq

Centrifuge

CLARK

DNA markers

MetaPhlAn

mOTUs

Dựa vào trình tự aa

Kraken2X

MMseqs2
Taxonomy

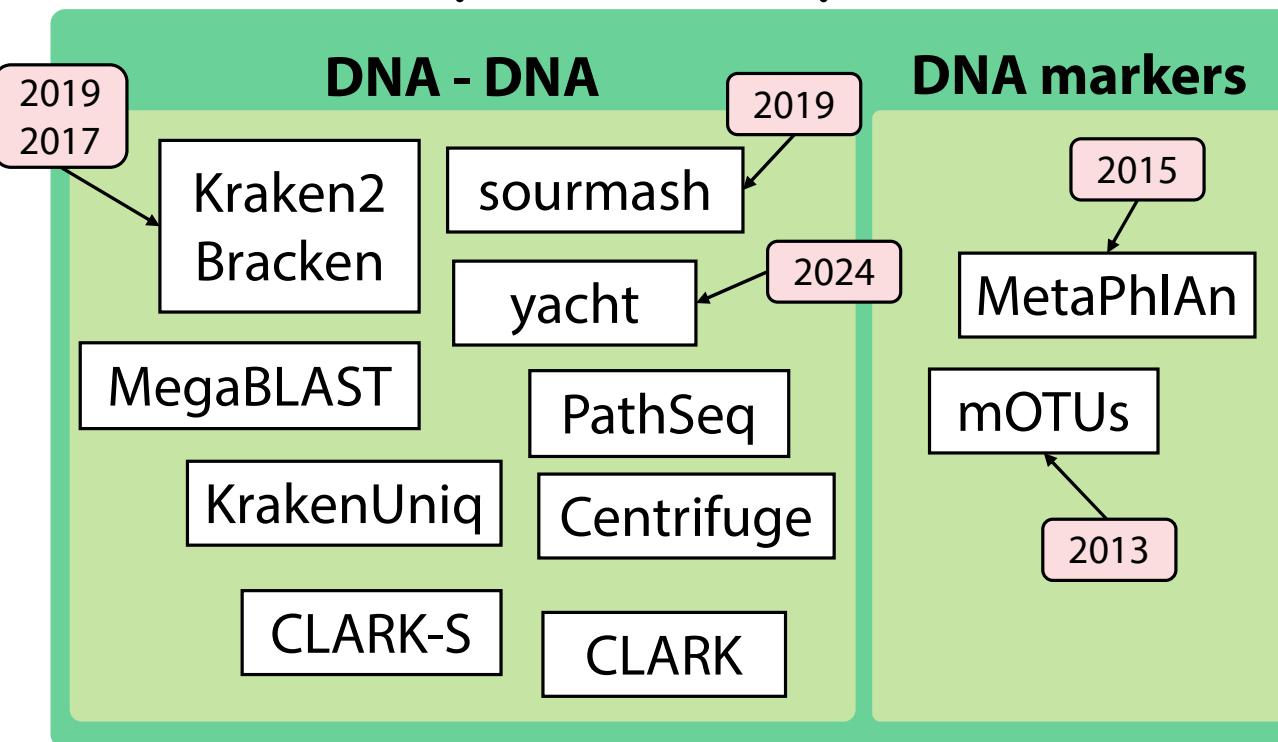
DIAMOND

Kaiju

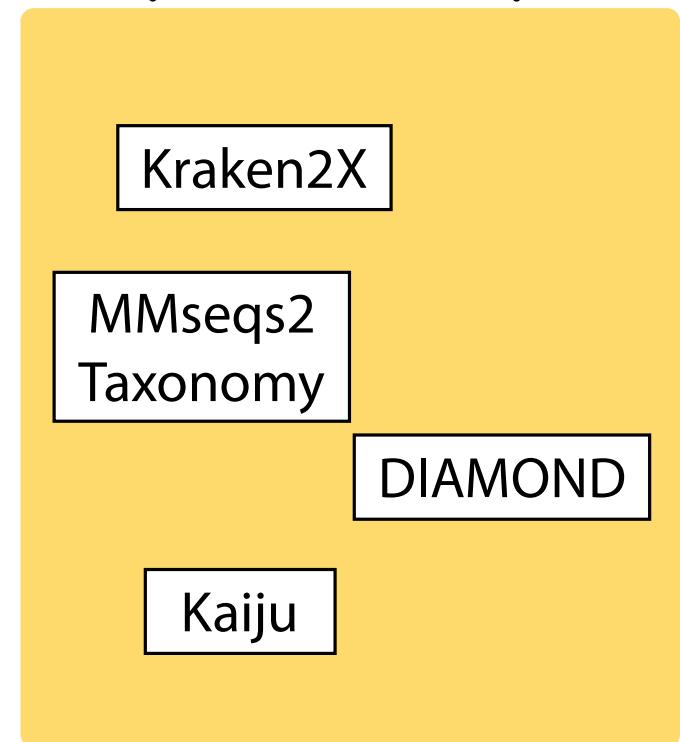
CÔNG CỤ NÀO ĐÃ ĐƯỢC CHỌN?

Chúng ta đang có những công cụ nào?

Dựa vào trình tự DNA



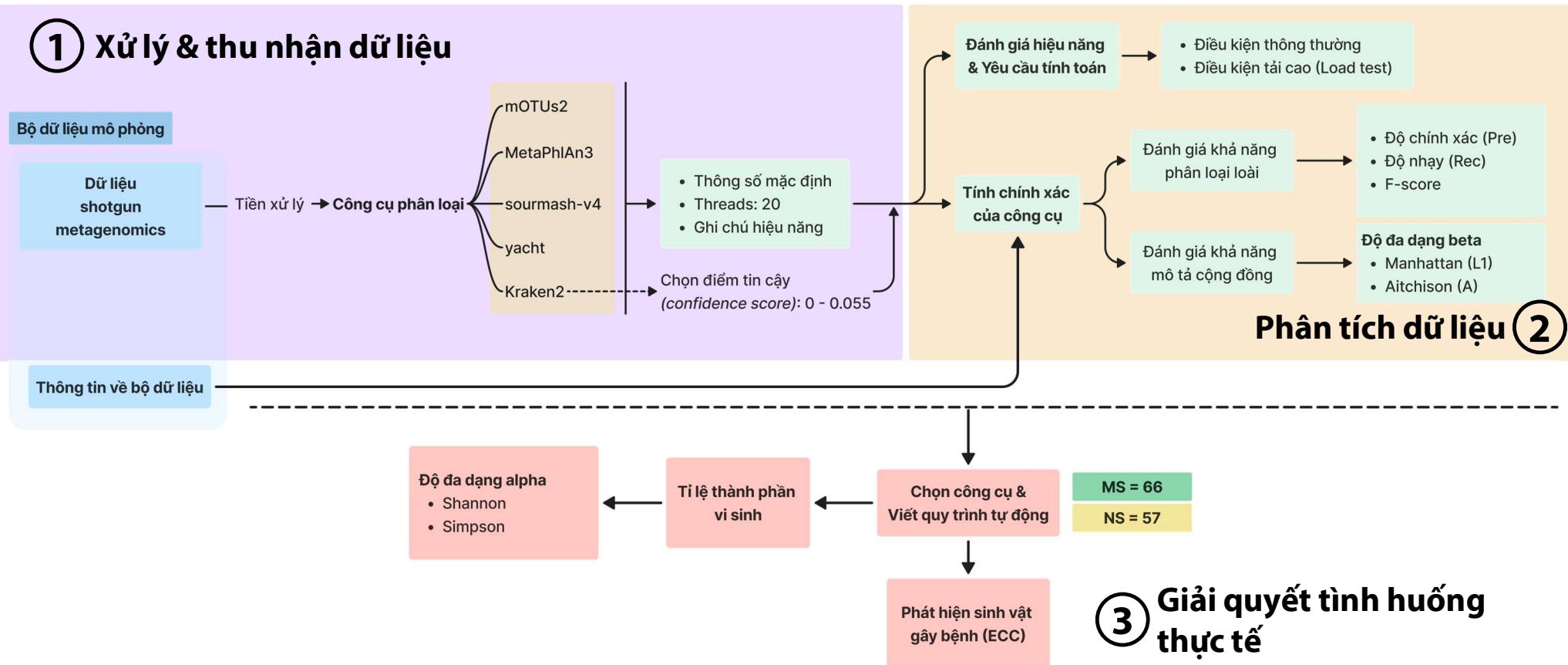
Dựa vào trình tự aa



THIẾT KẾ THÍ NGHIỆM

Thiết kế thí nghiệm

① Xử lý & thu nhận dữ liệu



DỮ LIỆU ĐÃ CHỌN LIỆU CÓ ĐỦ TỐT?

Thông tin bộ dữ liệu mô phỏng

- BioProject: PRJNA747117
(Tourlousse, 2022)
- N = 140
(56 shotgun + 84 amplicon)
- ~160GB
- Mang tính đại diện vi khuẩn đường ruột người.
- Đa dạng về %GC.

TABLE 1 Bacterial species included in the mock communities

Species	Culture collection	Nucleotide accession	Genome size (bp)	GC content (%)	16S rRNA genes	Cell wall (Gram-type) ^a	Relative abundance in DNA mock ^b (%)	Relative abundance in cell mock ^c (%)
<i>Bacteroides uniformis</i>	NBRC 113350	AP019724 – AP019728	4,989,532	46.2	4	–	4.7	5.6
<i>Blautia</i> sp.	NBRC 113351	CP084061	6,247,046	46.7	5	+	4.5	5.6
<i>Enterocloster clostridioformis</i>	NBRC 113352	BJLB01000001 – BJLB01000002	5,687,315	48.9	5	+	5.3	5.6
<i>Parabacteroides distasonis</i>	NBRC 113806	AP019729	5,179,960	45.0	7	–	4.8	5.6
<i>Bacillus subtilis</i> subsp. <i>subtilis</i>	NBRC 13719	AP019714 – AP019715	4,295,305	43.3	10	+	5.2	5.6
<i>Streptococcus mutans</i>	NBRC 13955	AP019720	2,018,796	36.9	5	+	6.9	5.6
<i>Pseudomonas putida</i>	NBRC 14164	AP013070	6,156,701	62.3	7	–	3.9	5.6
<i>Lactobacillus delbrueckii</i> subsp. <i>delbrueckii</i>	NBRC 3202	AP019750	1,910,306	50.1	8	+	3.6	5.6
<i>Escherichia coli</i>	NBRC 3301	CP048439 – CP048440	4,755,096	50.8	7	–	5.6	5.6
<i>Flavonifractor plautii</i>	NBRC 113805	CP084007	4,277,038	60.4	3	+	3.7	5.6
<i>Staphylococcus epidermidis</i>	NBRC 113846	CP084008 – CP084011	2,520,735	32.2	6	+	4.8	5.6
<i>Cutibacterium acnes</i> subsp. <i>acnes</i>	NBRC 113869	CP084017	2,560,907	60.0	3	+	5.0	5.6
<i>Bifidobacterium longum</i>	NBRC 114370	CP084012 – CP084013	2,594,022	60.1	5	+	5.7	5.6
<i>Anaerostipes caccae</i>	NBRC 114412	CP084016	3,284,789	44.5	4	±	5.3	5.6
<i>Ruminococcus gnavus</i>	NBRC 114413	CP084014 – CP084015	3,757,469	42.5	5	+	5.6	5.6
<i>Megasphaera massiliensis</i>	NBRC 114414	CP084019	2,610,024	50.6	7	–	4.8	0 ^b
<i>Megamonas funiformis</i>	NBRC 114415	CP084018	2,464,533	31.5	6	–	3.7	0 ^b
<i>Collinsella aerofaciens</i>	NBRC 114504	CP084004 – CP084006	2,278,612	60.3	5	+	6.2	5.6
<i>Bifidobacterium longum</i> subsp. <i>longum</i>	NBRC 114494	CP084020 – CP084022	2,534,372	60.1	4	+	4.7	5.6
<i>Akkermansia muciniphila</i>	NBRC 114322	CP084201 – CP084202	2,788,458	55.7	3	–	6.0	5.6

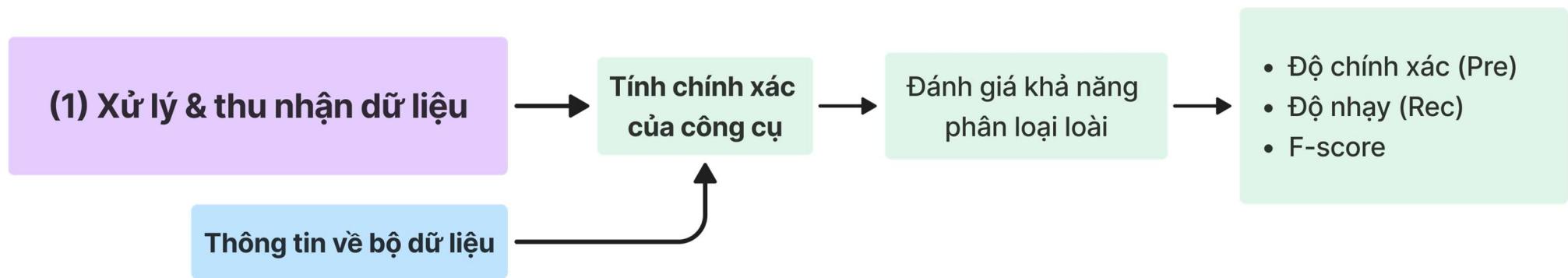
^aThe symbols +, – and ± indicate strains with Gram-positive, Gram-negative and Gram-variable type cell walls, respectively.

^bRelative abundances represent values assigned during formulation of the mock communities, based on quantification of the total DNA content of individual strains prior to mixing.

^c*M. massiliensis* and *M. funiformis* were excluded from the cell mock community.

CÔNG CỤ NÀO PHÂN LOẠI LOÀI CHÍNH XÁC NHẤT?

Thiết kế thí nghiệm



CÔNG CỤ NÀO PHÂN LOẠI LOÀI CHÍNH XÁC NHẤT?

Đánh giá khả năng phân loại loài

- Độ chính xác (Precision/Purity)

$$\text{Pre} = \frac{TP}{TP + FP}$$

- Độ nhạy (Recall/Completeness)

$$\text{Rec} = \frac{TP}{TP + FN}$$

- F_1 -score

$$F_1 = \frac{2 \times \text{Pre} \times \text{Rec}}{\text{Pre} + \text{Rec}}$$

- $F_{0.5}$ -score

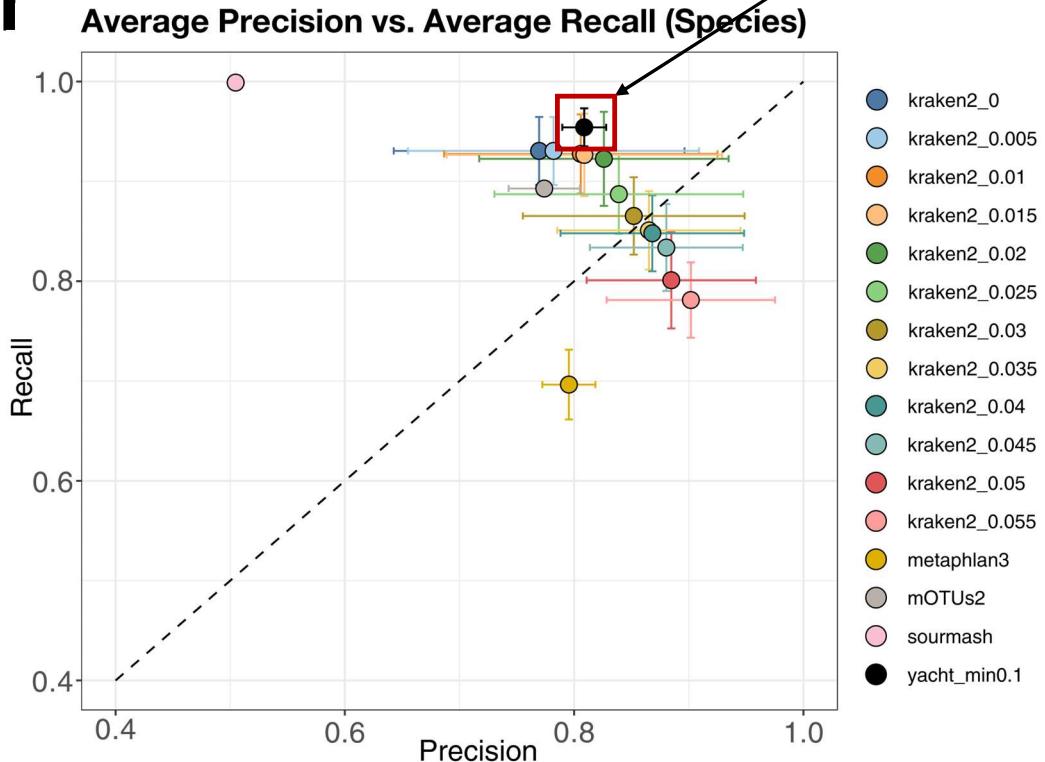
$$F_{0.5} = \frac{(1 + 0,5^2) \times \text{Pre} \times \text{Rec}}{0,5^2 \times \text{Pre} + \text{Rec}}$$

CÔNG CỤ NÀO PHÂN LOẠI LOÀI CHÍNH XÁC NHẤT?

Khả năng phân loại loài

- Điểm càng nghiêng về góc trên bên phải thì kết quả càng tốt

- Pre: $0,81 \pm 0,02$
- Rec: $0,95 \pm 0,02$

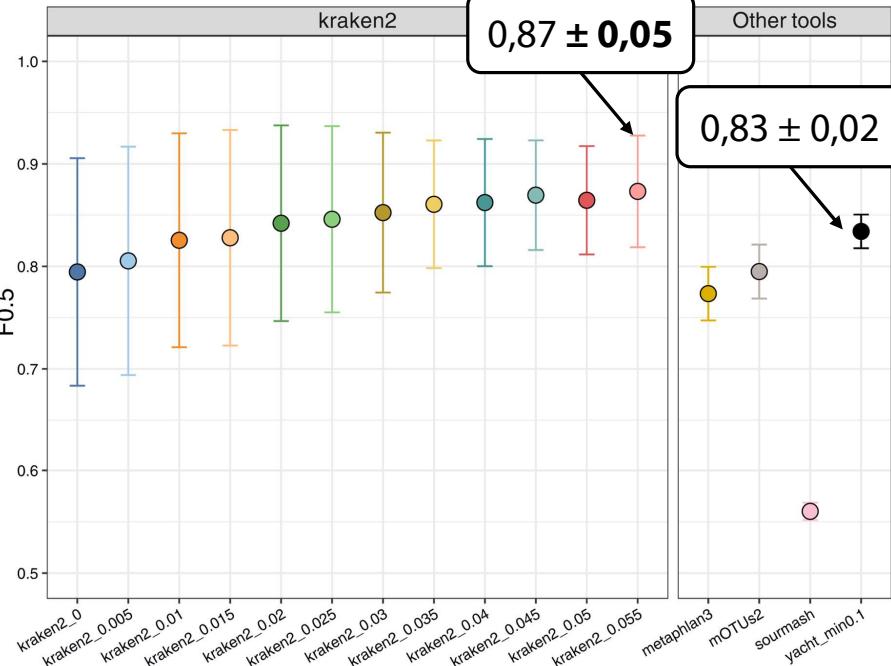


Kết quả tương quan giữa độ chính xác (precision) và độ nhạy (recall) của 16 phương pháp khác nhau đối với bộ mẫu của Tourlousse ($n = 56$) ở mức phân loại loài. Các điểm và thanh sai số chuẩn tương ứng với giá trị trung bình của độ chính xác và độ nhạy.

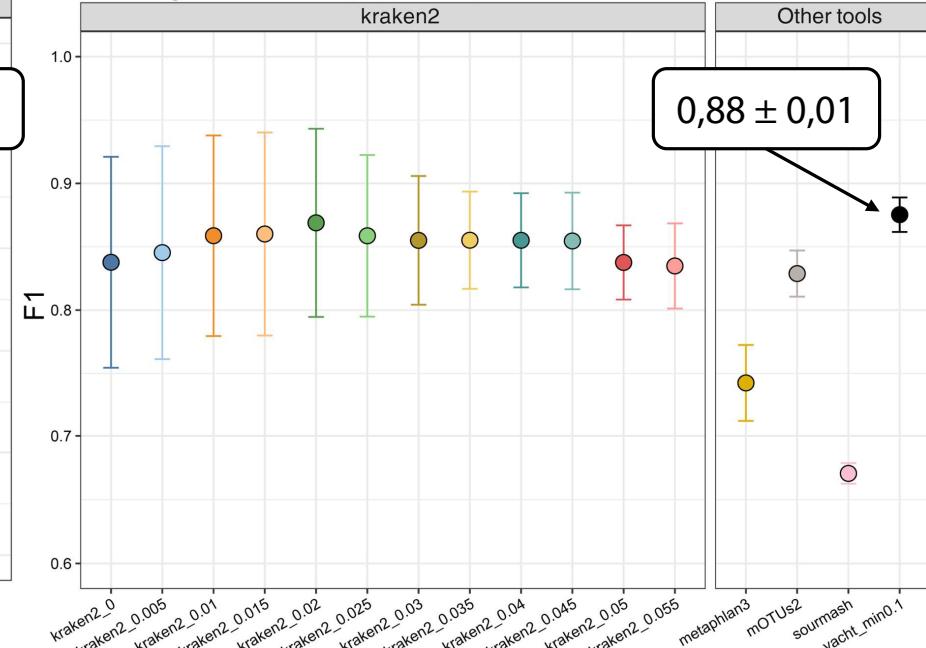
CÔNG CỤ NÀO PHÂN LOẠI LOÀI CHÍNH XÁC NHẤT?

Khả năng phân loại loài

Average F0.5 score (Species)



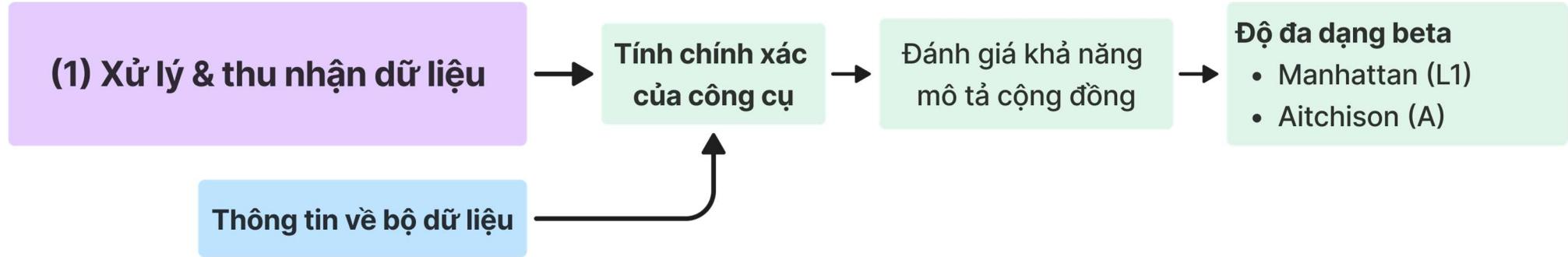
Average F1 score (Species)



Kết quả về giá trị $F_{0.5}$ và F_1 của 16 phương pháp khác nhau đối với bộ mẫu của Tourlousse ($n = 56$) ở mức phân loại loài.
Mỗi công cụ được biểu diễn với giá trị trung bình và thanh sai số chuẩn

SO VỚI THÔNG TIN BAN ĐẦU, CÔNG CỤ NÀO PHẢN ÁNH ĐÚNG THÀNH PHẦN NHẤT?

Thiết kế thí nghiệm



SO VỚI THÔNG TIN BAN ĐẦU, CÔNG CỤ NÀO PHẢN ÁNH ĐÚNG THÀNH PHẦN NHẤT?

Đánh giá khả năng mô tả cộng đồng

- Khoảng cách L1 (Manhattan)

$$L_{1(r)} = \sum_i \left| \left(x_r \right)_i - \left(x_r \right)_i^* \right|$$

- Khoảng cách Aitchison

$$d_{A(r)} \left(\left(x_r \right)_i, \left(x_r \right)_i^* \right)$$

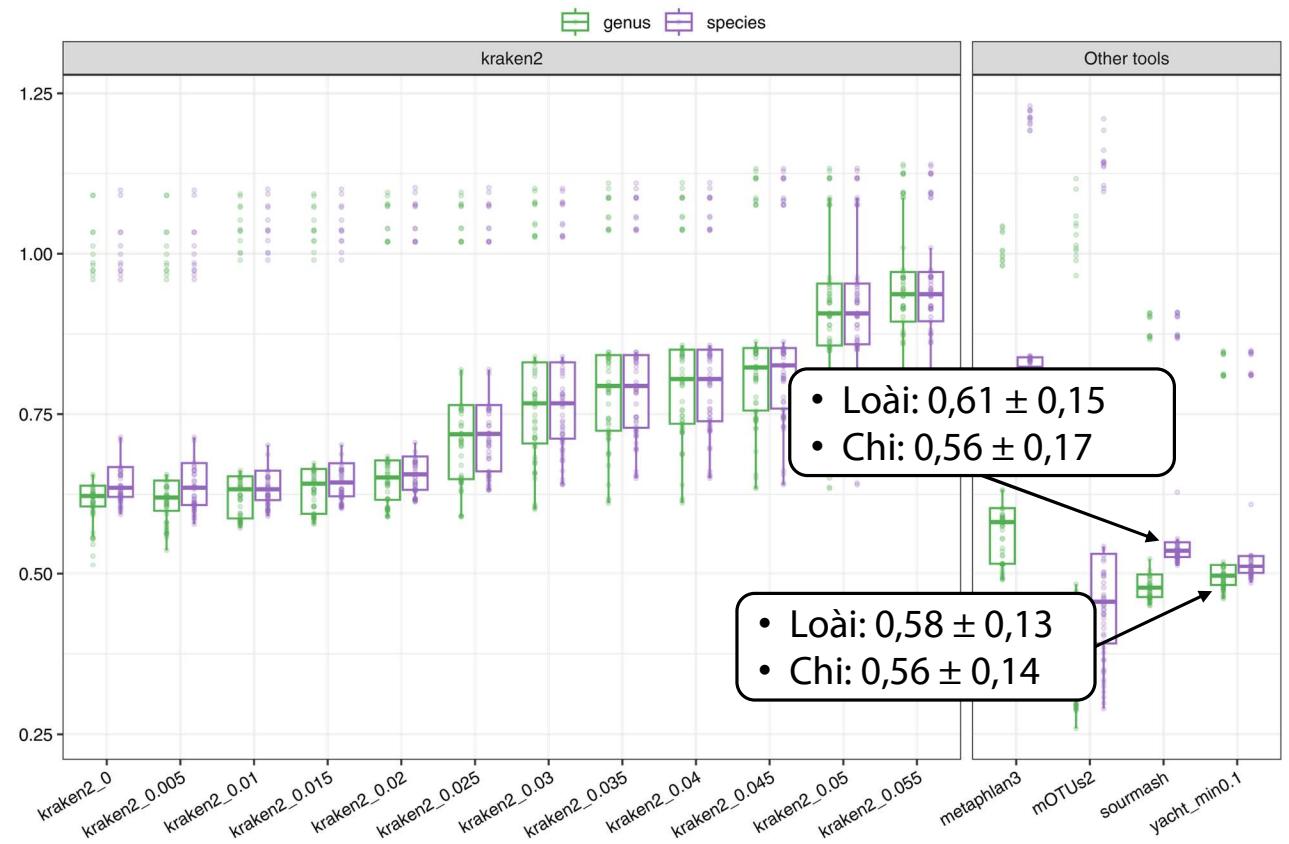
$$= d_{E(r)} \left(\text{clr} \left(\left(x_r \right)_i \right), \text{clr} \left(\left(x_r \right)_i^* \right) \right)$$

$$= \sqrt{\left(\ln \frac{\left(x_r \right)_i}{\sqrt{\left(x_r \right)_i \left(x_r \right)_i^*}} \right)^2 + \left(\ln \frac{\left(x_r \right)_i^*}{\sqrt{\left(x_r \right)_i \left(x_r \right)_i^*}} \right)^2}$$

SO VỚI THÔNG TIN BAN ĐẦU, CÔNG CỤ NÀO PHẢN ÁNH ĐÚNG THÀNH PHẦN NHẤT? Khả năng mô tả lại chính xác thành phần sinh vật

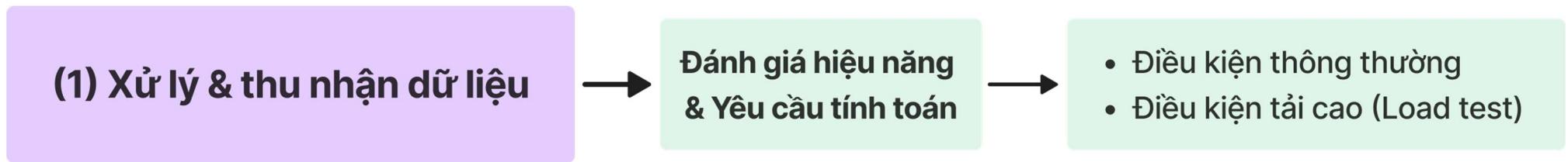
- Giá trị càng lớn, độ chênh lệch so với dữ liệu gốc càng cao.

Kết quả khoảng cách L_1 (Manhattan) của 16 phương pháp khác nhau đối với bộ mẫu của Tourlousse ($n = 56$) ở mức phân loại chi và loài.



Ở ĐIỀU KIỆN SỬ DỤNG KHÁC NHAU, LIỆU CÔNG CỤ CÓ CÒN TỐT HAY KHÔNG?

Thiết kế thí nghiệm



Ở ĐIỀU KIỆN SỬ DỤNG KHÁC NHAU, LIỆU CÔNG CỤ CÓ CÒN TỐT HAY KHÔNG?

Đánh giá hiệu suất & yêu cầu tính toán

- GNU time với format chung là `time -a -f "%U,%S,%E,%M,%P"`
- Chạy ở điều kiện thông thường (lần lượt 140 mẫu).
- Chạy ở điều kiện tải cao: gộp 140 mẫu được cung cấp từ Tourlousse (56 shotgun + 84 amplicon), tiến hành chạy lại các thông số như đánh giá khả năng phân loại loài.

Ở ĐIỀU KIỆN SỬ DỤNG KHÁC NHAU, LIỆU CÔNG CỤ CÓ CÒN TỐT HAY KHÔNG?

Điều kiện thông thường

Công cụ	Tiến hành lần lượt tổng cộng 140 mẫu		
	Bộ nhớ tối đa (GBs)	CPU (%)	Tổng thời gian (h)
Kraken 2_0	94,456	251	5,35
Kraken 2_0.005	94,419	251	5,24
Kraken 2_0.01	94,437	258	5,34
Kraken 2_0.015	94,434	244	5,37
Kraken 2_0.02	94,418	258	5,27
Kraken 2_0.025	94,406	253	5,63
Kraken 2_0.03	94,403	262	5,23
Kraken 2_0.035	94,443	254	6,37
Kraken 2_0.04	94,405	250	5,22
Kraken 2_0.045	94,387	254	6,17
Kraken 2_0.05	94,419	251	5,24
Kraken 2_0.055	94,428	220	5,99
MetaPhlAn 3	4,923	1322	11,62
sourmash	2,859	383	15,17
yacht	15,237	265	71,54
mOTUs2	4,857	155	8,7

Kết quả về hiệu suất ở điều kiện thông thường của các công cụ

Ở ĐIỀU KIỆN SỬ DỤNG KHÁC NHAU, LIỆU CÔNG CỤ CÓ CÒN TỐT HAY KHÔNG?

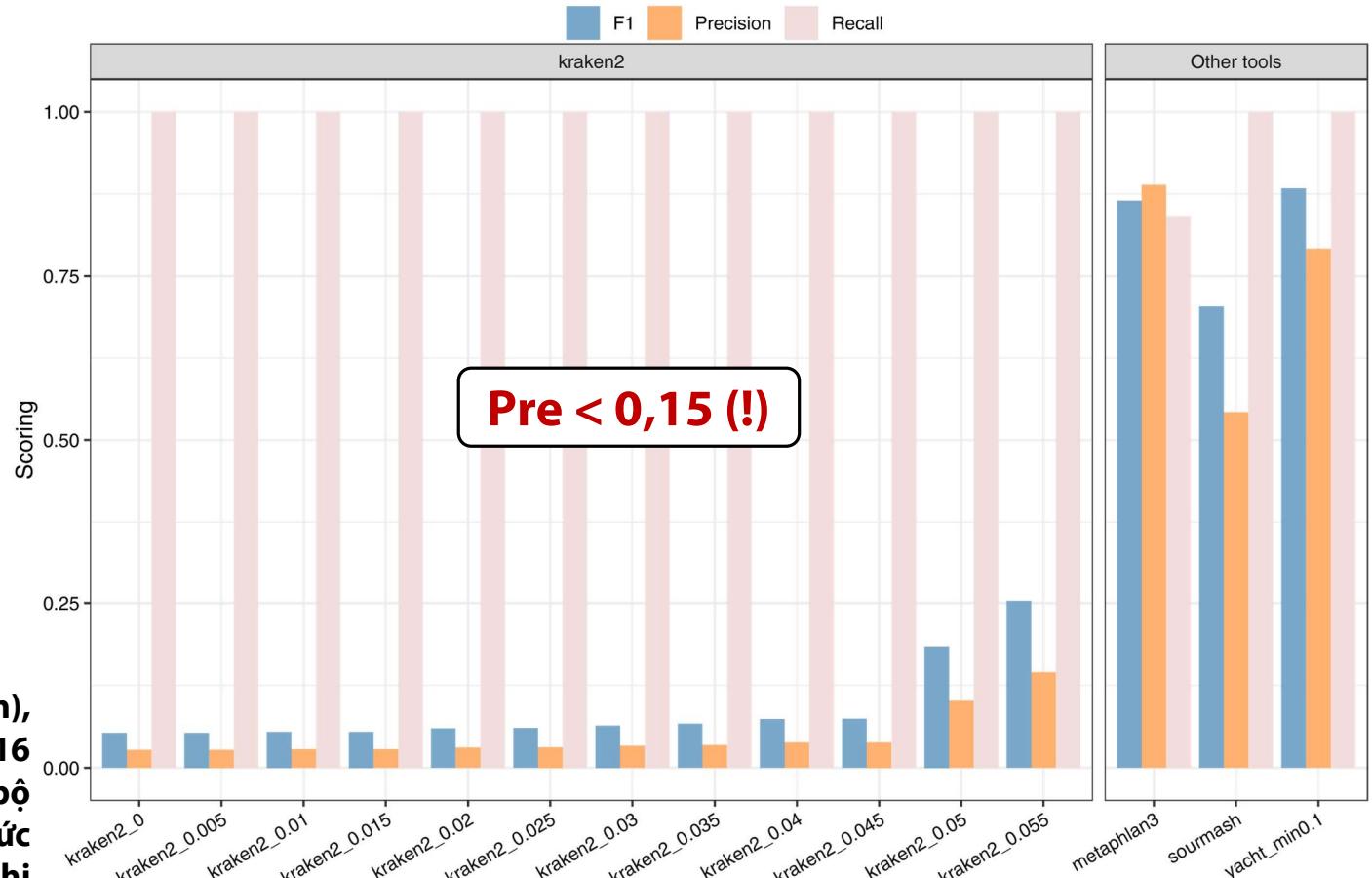
Load test

Kết quả về hiệu suất ở điều kiện tải cao của các công cụ

Công cụ	Tiến hành trên mẫu gộp (~140 GB)		
	Bộ nhớ tối đa (GBs)	CPU (%)	Tổng thời gian (h)
Kraken 2_0	95,12	255	6,71
Kraken 2_0.005	95,21	239	6,61
Kraken 2_0.01	95,1	228	6,77
Kraken 2_0.015	95,05	251	6,66
Kraken 2_0.02	95,15	243	6,61
Kraken 2_0.025	95,16	280	6,78
Kraken 2_0.03	95,15	≈	6,66
Kraken 2_0.035	95,01	264	6,46
Kraken 2_0.04	95,17	238	6,58
Kraken 2_0.045	95,03	285	6,55
Kraken 2_0.05	95,19	249	6,72
Kraken 2_0.055	95,09	282	6,58
MetaPhlAn 3	7,26	291	3,5
sourmash	6,75	195	20,67
yacht	1,41	100	16,9
mOTUs2	-	-	-

Ở ĐIỀU KIỆN SỬ DỤNG KHÁC NHAU, LIỆU CÔNG CỤ CÓ CÒN TỐT HAY KHÔNG?

Load test



Kết quả về độ chính xác (precision),
độ nhạy (recall) và hệ số F1 của 16
phương pháp khác nhau đối với bộ
mẫu của Tourlousse ($n = 1$) ở mức
phân loại chi

CÔNG CỤ NÀO CHÍNH XÁC NHẤT?

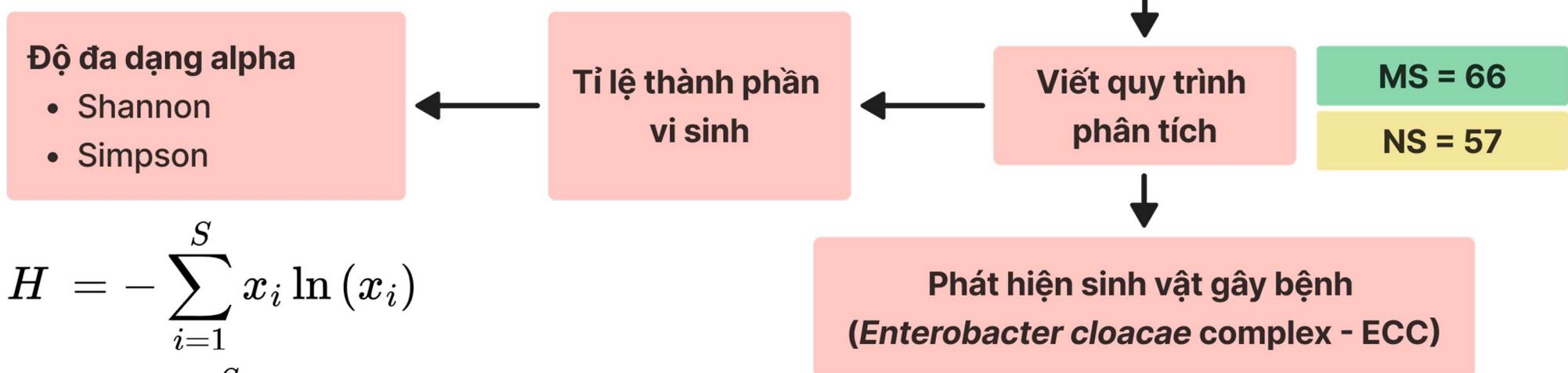
Kết quả đánh giá công cụ

- Công cụ **yacht** có kết quả tốt nhất.
- Các công cụ: yacht, MetaPhlAn 3 và Kraken 2 (hệ số tin cậy 0.05) được chọn cho các bước áp dụng trên mẫu lâm sàng.

KẾT QUẢ THỰC TẾ NHƯ THẾ NÀO?

Thiết kế thí nghiệm

- yacht
- MetaPhiAn 3
- Kraken 2 (hệ số tin cậy 0,05)



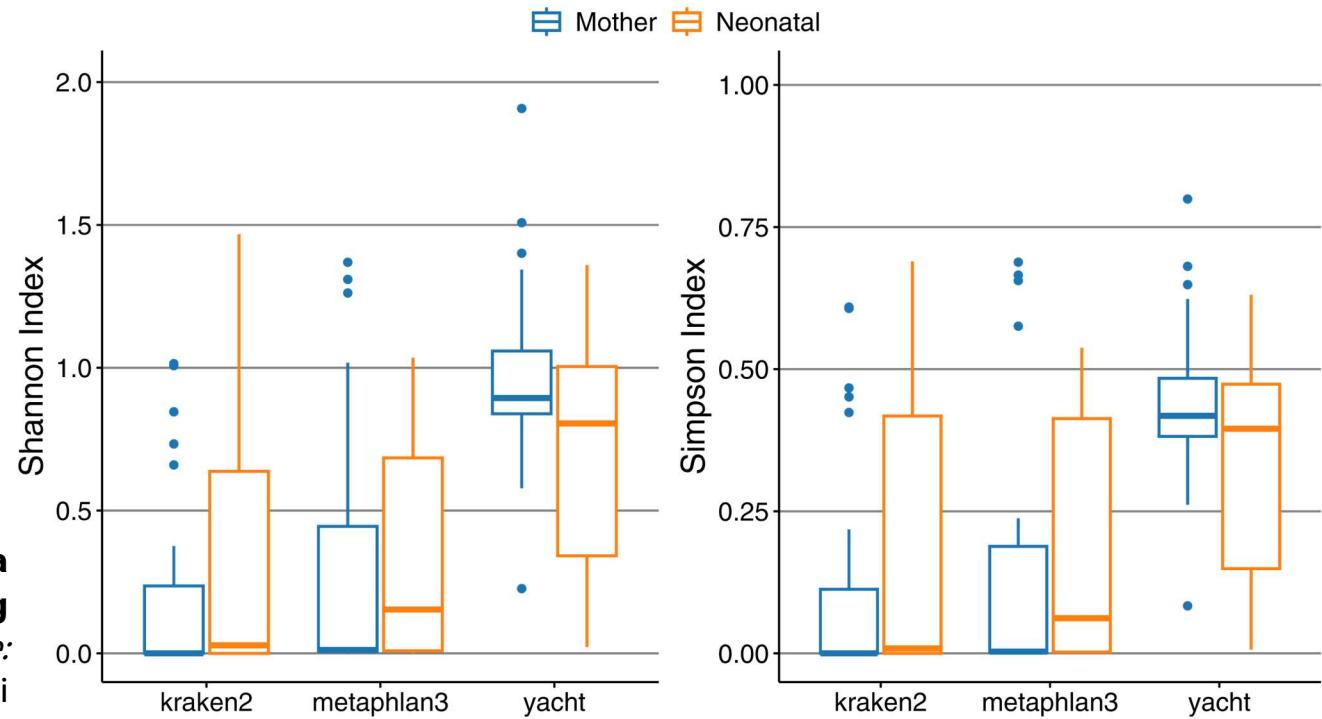
KẾT QUẢ THỰC TẾ NHƯ THẾ NÀO?

Thông tin bộ dữ liệu lâm sàng

- Các trình tự bộ gen vi khuẩn, trình tự dữ liệu metagenomic và dữ liệu lâm sàng được thu nhận từ các nghiên cứu của Đơn vị Nghiên cứu Lâm sàng Đại học Oxford (OUCRU) tại Nepal và Việt Nam.
- Đề tài tập trung vào dữ liệu lâm sàng và dữ liệu metagenome của mẫu phân được thu nhận từ các cặp mẹ - bé từ nghiên cứu "*Tầm soát các yếu tố nguy cơ gây nhiễm trùng máu sơ sinh của trẻ ở khoa hồi sức tích cực mang Enterobacteriaceae sinh men ESBL ở bệnh viện Patan*" từ tháng 03/2016 đến tháng 10/2017.
- Nuôi cấy chọn lọc vi khuẩn Gram âm trên môi trường MacConkey đơn thuần → các khuẩn lạc mọc trên môi trường sẽ được thu nhận lại và đem đi giải trình tự metagenome.

KẾT QUẢ THỰC TẾ NHƯ THẾ NÀO?

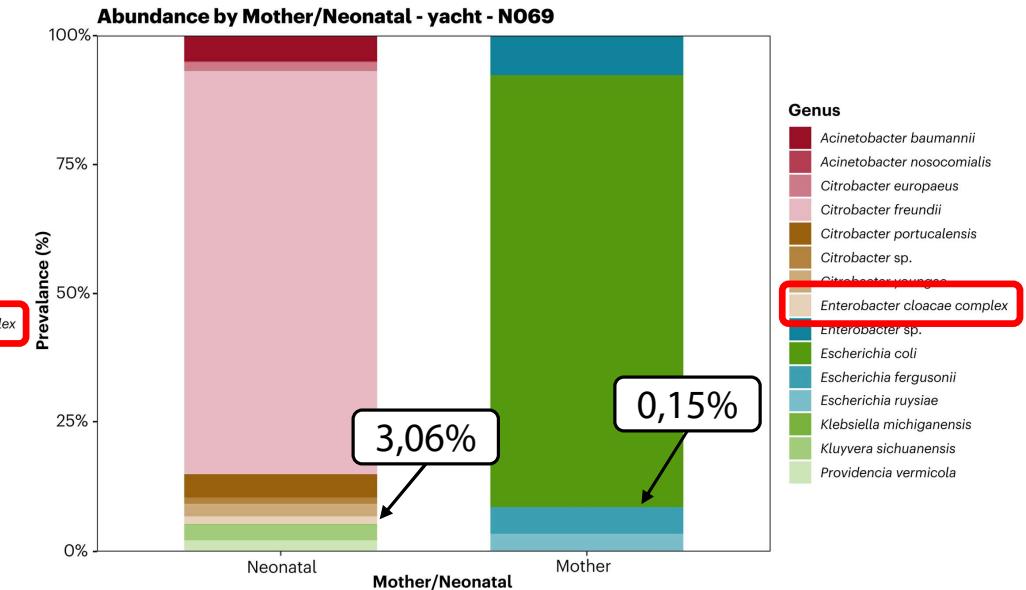
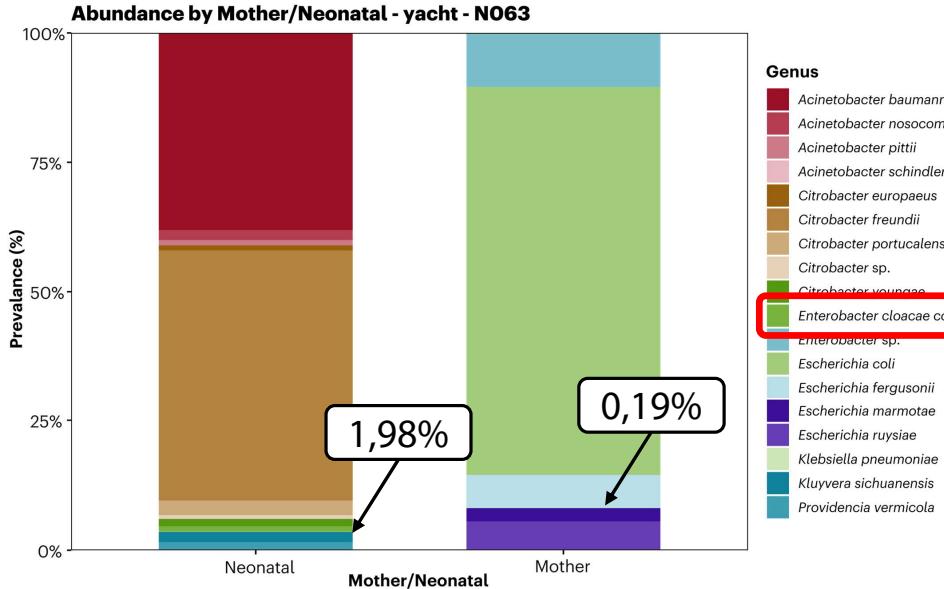
Kết quả mẫu lâm sàng - Độ đa dạng



Kết quả về độ đa dạng alpha của
mẫu mẹ và bé được đánh giá thông
qua chỉ số Shannon và Simpson (mẹ:
 $n = 66$, bé: $n = 57$) ở mức phân loại loài

KẾT QUẢ THỰC TẾ NHƯ THẾ NÀO?

Kết quả mẫu lâm sàng - Phát hiện ECC



Đặc điểm về thành phần vi sinh của hai mẫu mẹ và bé (N063, N069) ở mức phân loại loài được phát hiện bởi công cụ yacht

Kết luận

- **yacht** có kết quả phân loại loài lẩn khả năng mô tả lại thành phần vi sinh chính xác nhất, do đó công cụ này có tiềm năng trở thành một công cụ phân tích cộng đồng vi sinh trong các mẫu shotgun metagenomics đầy hứa hẹn.
- Phát hiện được tác nhân gây nhiễm trùng huyết sơ sinh trong báo cáo trước đây của Quỳnh & Sulochana.
- Ghi nhận được sự biến động về thành phần vi sinh ở bé nhiều hơn so với mẹ.

Kiến nghị

- Mở rộng việc đánh giá các công cụ tin sinh cho những bước tiếp theo để có thể hoàn thiện quy trình phân tích mẫu metagenomics.
- Viết quy trình tự động để người sử dụng có thể thu được kết quả cuối mà không phần phải viết các câu lệnh thủ công.
- So sánh về mặt di truyền đến mức dưới loài cho các cặp mẹ - bé bằng cách kết hợp thêm dữ liệu WGS của ECC đường ruột đã được ghi nhận trong báo cáo trước đây của Quỳnh & Sulochana.

THE END.

Cảm ơn thầy/cô đã theo dõi



OUCRU
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
ĐH QUỐC HỘI HỒ CHÍ MINH



KHOA SINH HỌC –
CÔNG NGHỆ SINH HỌC



Khoa luận tốt nghiệp

Acknowledgement

OUCRU-VN

- Dr Pham Thanh Duy
- Dr Chung The Hao
- Dr Maia Rabaa
- Nguyen Thi Nguyen To

Special thanks to

- Nguyen Pham Nhu Quynh

OUCRU-Nepal

- Dr Sulochana Manandhar
- Dr Abhilasha Karkey

Cambridge University

- Prof. Stephen Baker