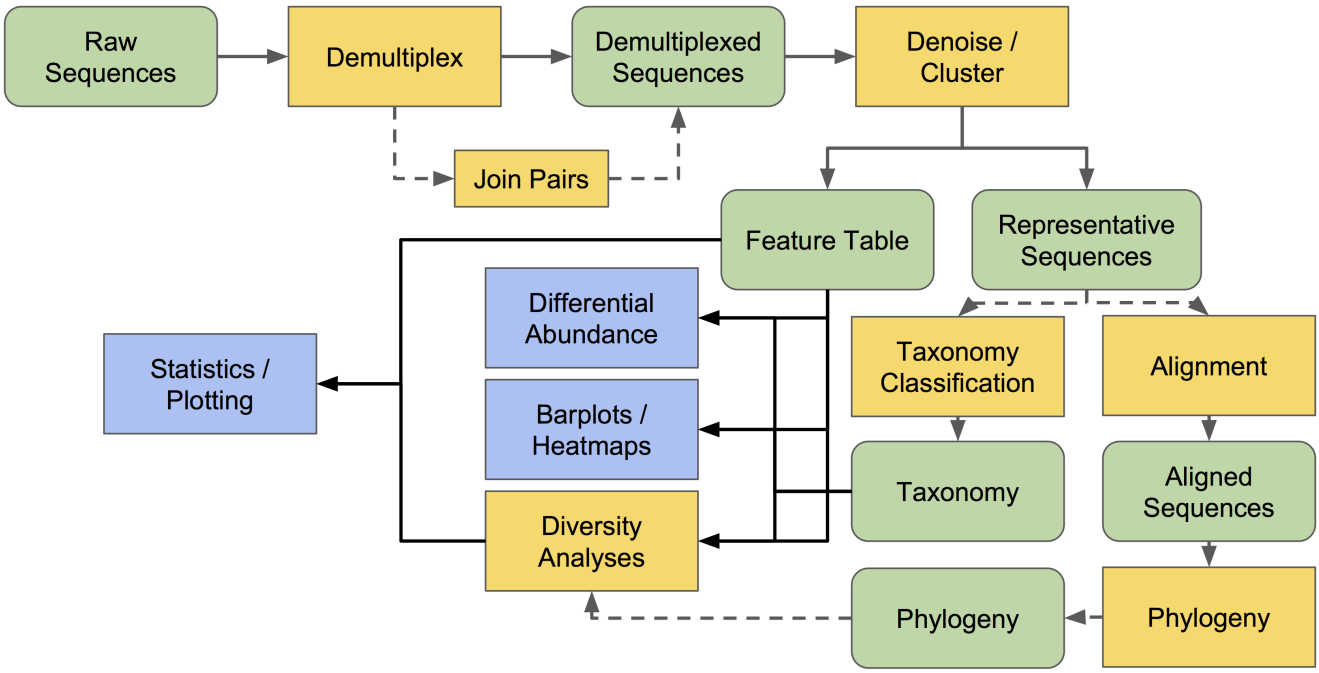
Câu 1

Nên sử dụng kỹ thuật amplicon metagenomics. Vì để xác định thành phần hệ vi sinh nấm có thể sử dụng các vùng gene ITS (*nucler ribosomal internal transcribed spacer*), đây là vùng gene bảo tồn ở nấm giúp phân biệt nấm với các loài khác và cũng đủ phân kỳ để định danh, nghiên cứu về phat sinh loài ở nấm. Bên cạnh đó giúp giảm tiêu hao tài nguyên giải trình tự.

Câu 2

QIIME2

Đây là một bộ dụng cụ được tích hợp sẵn các công cụ cần thiết để phân tích amplicon sequencing.



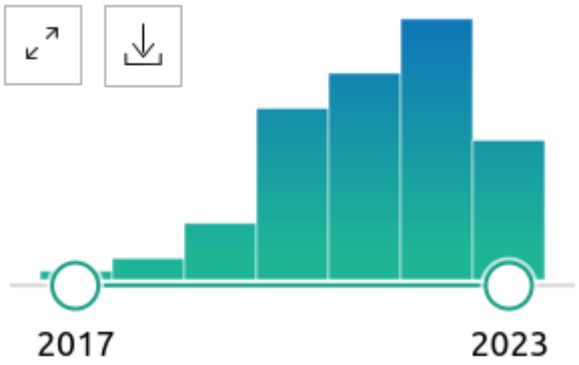
Dữ liệu đầu vào gồm có 3 thành phần:

- Sample Data: Bộ dữ liệu được thu thập các đặc tính của các đối tượng mà có thể liên quan đến các câu hỏi sinh học khi nghiên cứu metagenomic (như tuổi, nơi lấy mẫu, kháng sinh,...)
- FastQ file: Dữ liệu giải trình tự
- Barcode: Dữ liệu về mối liên hệ giữa mẫu và trình tự được giải thông qua barcode (giúp xác định trình tự nào của sample nào).

Dữ liệu đầu vào sẽ được Demultiplex tức trình tự sẽ được link với Sample Data và lưu ở từng file fastq khác nhau tương ứng với các barcode. Sau quá trình này ta sẽ có 1 file Sample Data ban đầu và các file fastq tương ứng với các barcode.

Sau đó trình tự được lọc nhiễu, loại bỏ các bias của phản ứng PCR xác định các OTU hoặc ASV đại diện cho sự hiện diện của đối tượng trong mẫu. Sau quá trình này thu được Feature Table và Các trình tự. Các trình tự

có thể dùng phân tích phát sinh loài, xác định các loài trong mẫu. Bên cạnh đó có thể sử dụng Feature Table để phân tích thêm Diversity α và β,...



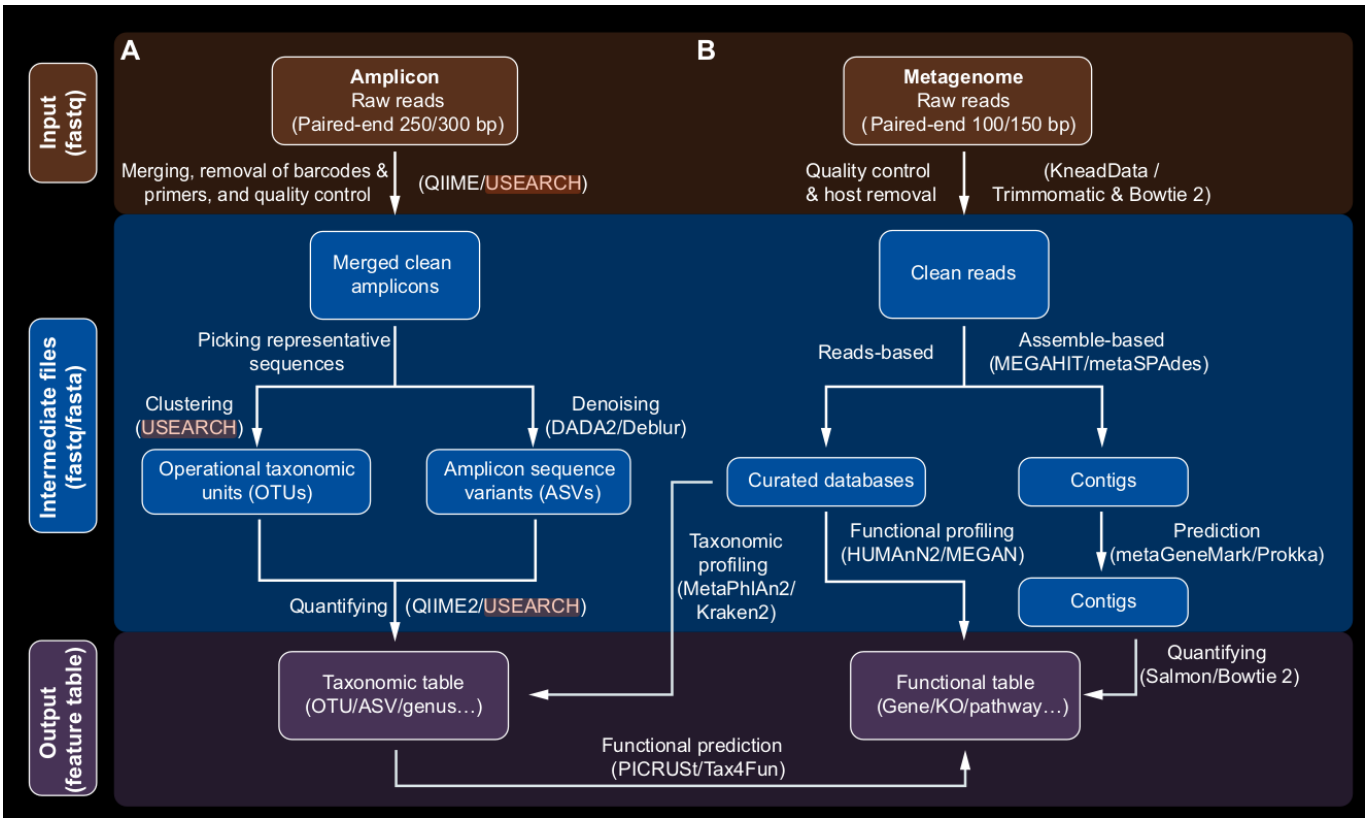
Ưu điểm

- Công cụ linh hoạt, có thể tùy pipline thay đổi linh hoạt dựa trên input và câu hỏi sinh học.
- Có cộng đồng phân tích đông đảo, có thể trao đổi thông tin.
- Mã nguồn mở, miễn phí, dễ dàng tiếp cận, mở rộng.
- Kiểm soát pipeline, nguồn trích dẫn được lưu trong artifact data.

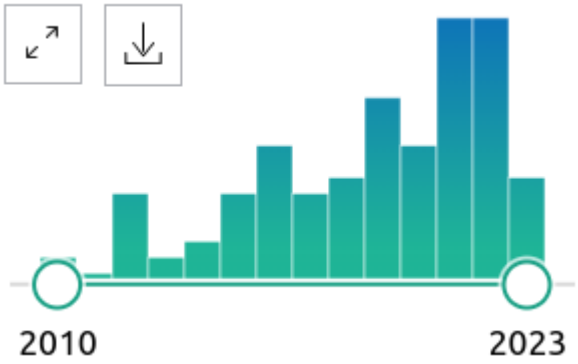
Nhược điểm

- Có những thuật toán thống kê phức tạp, đòi hỏi kiến thức tối thiểu về toán
- Khi có bug, cần trao đổi cần phải chờ phản hồi.

USEARCH-VSEARCH



Cũng tương tự như QUIIME, với USEARCH, mục tiêu đầu hướng tới cũng là chuyển trình tự thô thành Feature Table. Dữ liệu cũng sẽ được kết hợp với nhau và chọn các trình tự đại diện nhằm giảm sự bias về số lượng mẫu trong phân tích. Với USEARCH sẽ tạo thành các OTU bằng các nhóm các trình tự tương đồng, quy về OTU, và các bước downstream có thể phân tích tương tự như QIIME.



Ưu điểm

- Được phát triển trước QIIME
- Miễn phí, mã nguồn mở
- Linh động, có thể tùy chỉnh

Nhược điểm

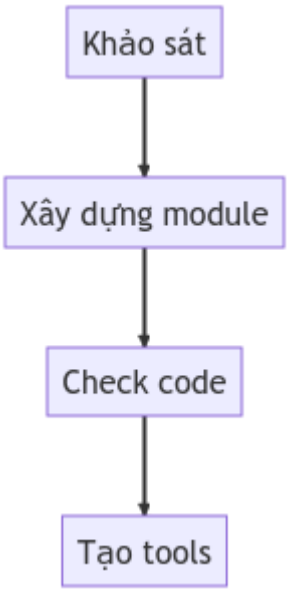
- Cộng đồng nghiên cứu ít (công bố 78 bài báo liên quan trong khi QIIME là 231 bài báo)
- Giao diện trang chủ đơn giản, ít thu hút.

```
 Theo em công cụ QIIME thích hợp hơn với công ty hơn nhờ những tính
năng vượt trội của đó, có tiềm năng phát triển hơn trong tương lai.
Công cụ linh hoạt dễ phân công chia việc, tạo pipline với workflow,
documment chi tiết, cộng động sử dụng đông đảo.
```

Câu 3

Trong các artifact data của QIIME2 thực chất là các file zip. Dữ liệu bên trong dễ dàng được unzip sử để sử dụng, lập trình thêm các công cụ bổ sung bằng python, bash, nextflow,... Ngoài ra phần mềm còn hỗ trợ tương tác trực tiếp thông qua python bằng Jupyter Notebook hoặc Galaxy Workflow. Ở bước nghiên cứu Diversity có thể phát triển thêm các công cụ trực quan hóa dữ liệu có thể webapp, giúp dễ dàng truy cập và sử dụng.

Câu 4



Đầu tiên cần khảo sát thị trường về nhu cầu visualize dữ liệu. Xem những hướng visualize nào đã và đang phát triển, những hướng vẫn còn chưa phát triển. Thu thập, Xây dựng dần các module hỗ trợ trực quan hóa. Check tính năng và sự chính xác của các module. Sau đó lắp ráp các module tạo thành các công cụ web app giúp người dùng dễ dàng truy cập và sử dụng.

Các tiêu chí đánh giá:

- Câu hỏi sinh học có tính lặp lại ở các khách hàng hay không?
- Đã có công cụ đã phát triển chưa? Nếu có công cụ đó đã tối ưu chưa?
- Ứng dụng đã xây dựng có lượt sử dụng bao nhiêu? Số lượng bài báo công bố liên quan?