Basic of Qiime2

ANALYZING MICROBIOME FOR
UNDERSTANDING INVISIBLE

This file was created by Ryo NIWA, Kyoto University

Introduction

- NGSによるアンプリコンシーケンス解析とは
- シーケンスにおけるルール
- Qiime2とは何か
- DADA2による解析 (原理編)
- OTUによる解析 (原理編)
- 多様性の話
- メタ解析の話
- まとめ

NGSによるアンプリコンシーケンス解析とは

1. サンプル中細菌ゲノムの16S rRNA領域などを増幅し、NGSを使用した網羅的配列決定を行う。

2. その配列と既存データベースとの照合により、細菌群衆を調査する。

サンプルの用意

DNA抽出

網羅的にPCRをかける。
それを全てシーケンス。

NGS

DBによる辞書引き
各配列の由来調査

SBSでは1ラン=1サンプルではない

MiSeqを始めとするillumina社のシーケンスプラットフォームでは数百万~数千万ものリードを生み出す。

「タグ分け」、「Demultiplex」と呼ばれる作業はこのために行う。一般的に既知の微生物叢を知りたい場合は**10000**リードあれば十分。新規微生物探索を行う場合は**100000**リードを目安に解析を行う。

応微で食品を探索する場合は10000リードでOKと思われる。

Qiime2で受け付けるファイルは限られている

Qiime2では基本的に

EMP (Earth Microbiome project) protocol

と呼ばれる方法でMultiplexされたFASTQファイル

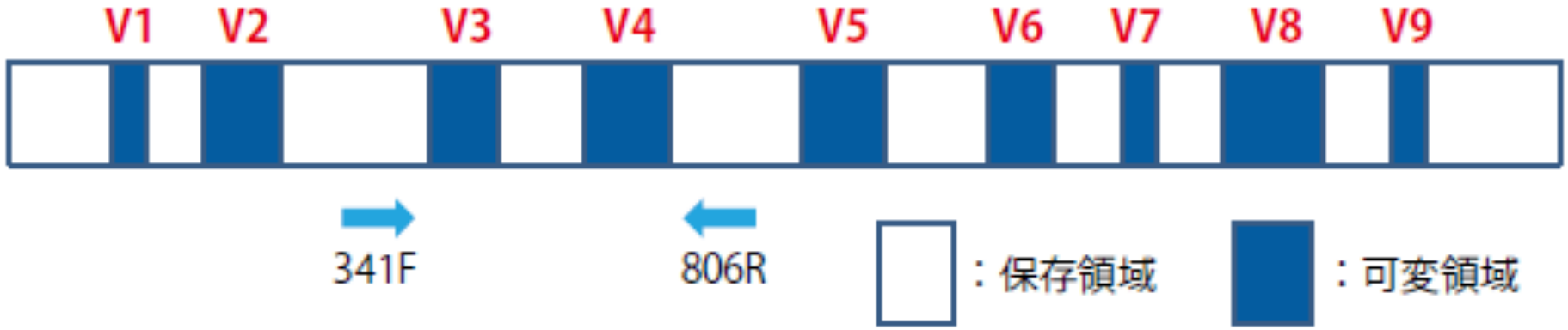
もしくは

Demultiplexed済みのFASTQファイルを受け付ける。

*EMP Protocol*について

例えば16S rRNA解析に**V3-V4領域**を用いる場合を考えると、Primerとしては341Fと806Rを利用することになる。

EMP Protocolでは**V4領域 (もしくはV4+V5)** を用いる。標準が決まっており、PCRの方法まで指定されている。12塩基のBarcodeを515Fに仕込むことになる。



*http://catalog.takara-bio.co.jp/product/basic_info.php?unitid=U100009221

他の方法はQiime2では使えないのか？

例の場合では、341F, 806Rを用いており、341F側に一定数のバーコード配列を仕込むような形になるだろう。

つまり、EMP Protocolではないので、本来Qiime2では受け付けない形式となっている。したがって、Qiime2とは別にDemultiplexedした上でデータをインポートをする必要がある。

この点に留意して解析を進める必要がある！

全体の流れ

サンプルDNAの抽出

J-Bio21
NIPPON STEEL & SUMIKIN Eco-Tech Corporation

16SrRNA領域のPCR増幅

BIOER TECHNOLOGY

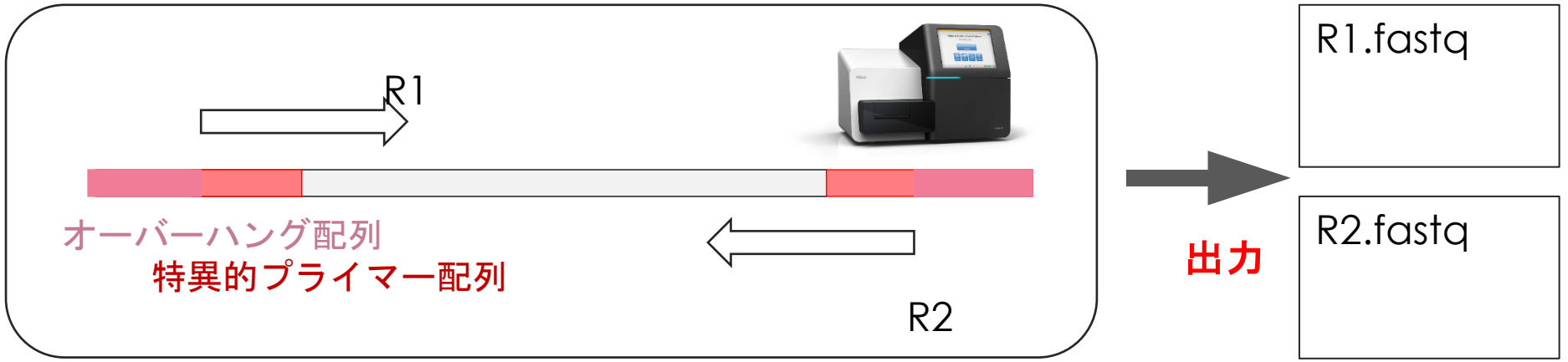
NGSによるシーケンシング

illumina®

配列処理・分類割り当て

NGS (MiSeq) から出力される配列

- MiSeqは両側から配列を301bp読む



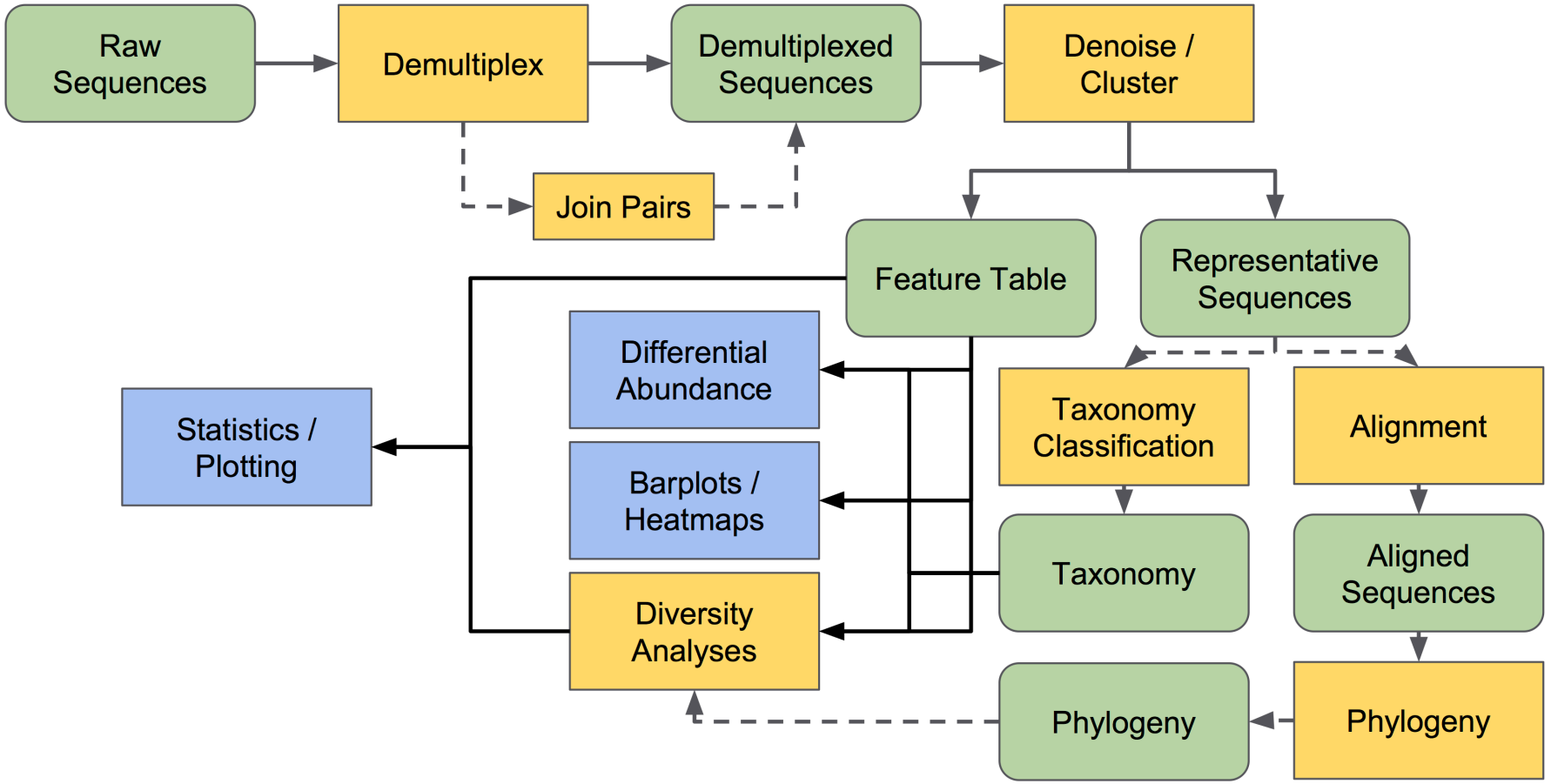
- R1, R2の二つのFASTQファイルが出力

Qiime2とは何か

Qiime2はマイクロバイオーム研究のバイオインフォマティクスツールであり、包括的にツールを納めたパッケージです。

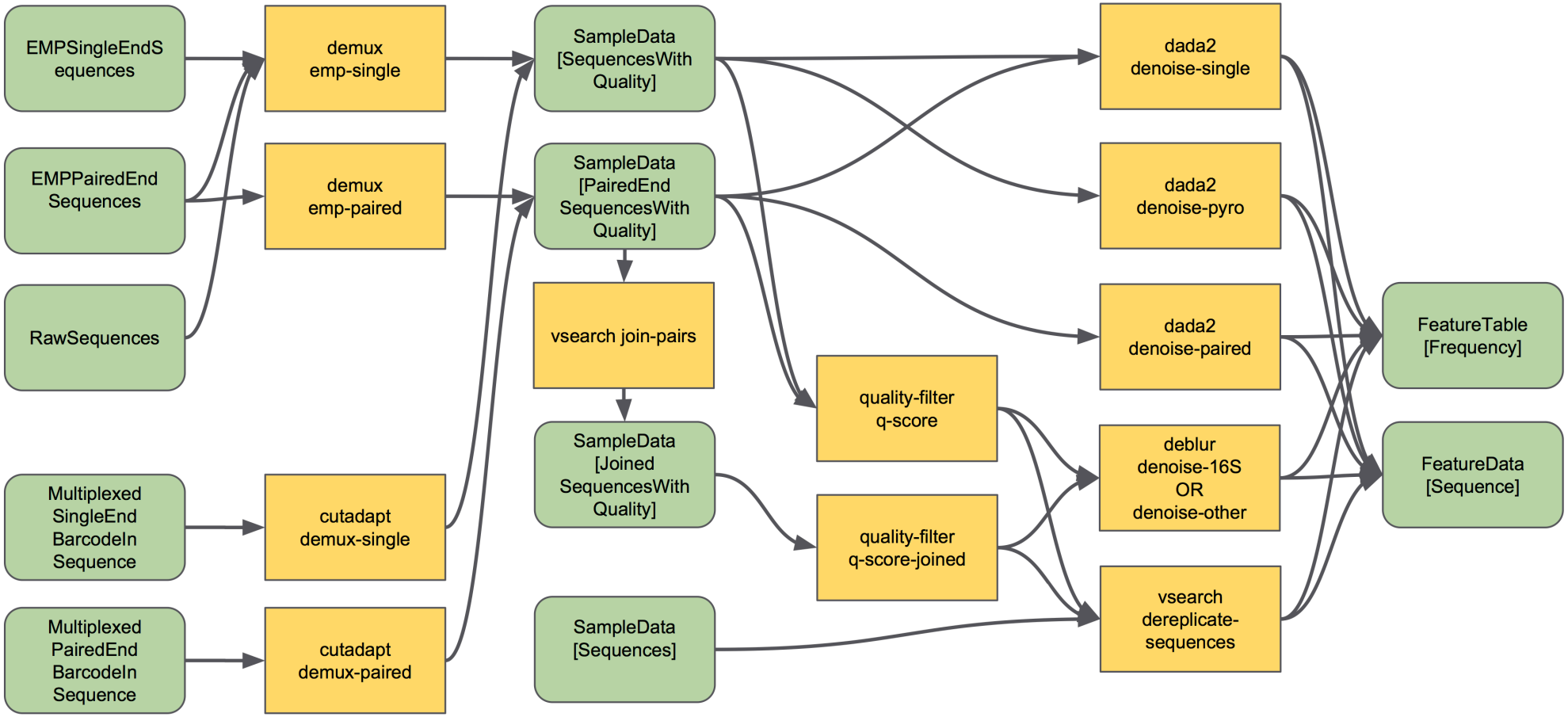
Qiime2は本来複数あるスクリプトを1つに統合したもので、要は1回のダウンロードで解析に必要なすべてのツールを簡単に落とす事ができると考えてください。

Qiime2を使えば全部できる

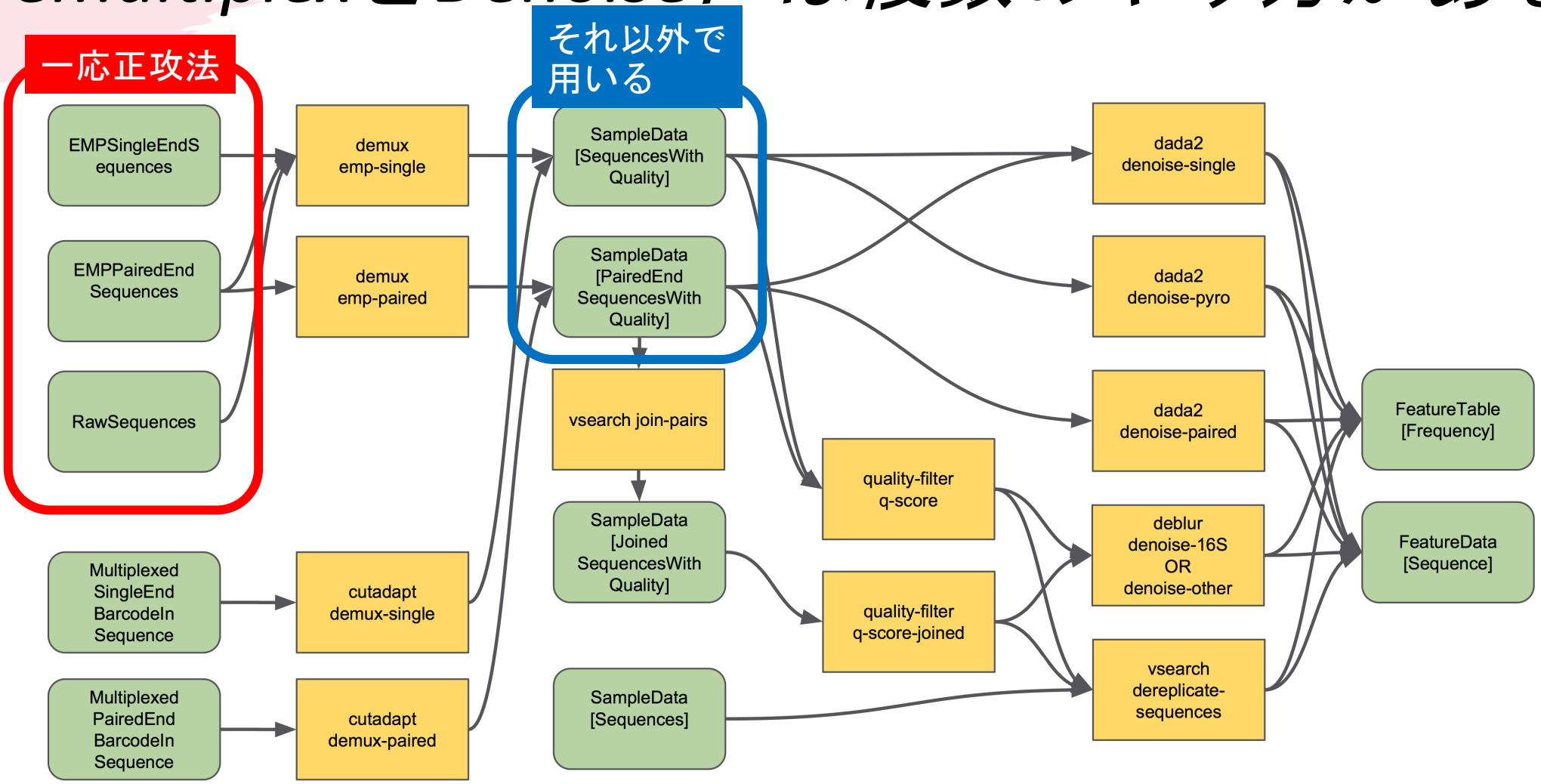


https://docs.qiime2.org/2020.2/tutorials/overview/

DemultiplexとDenoiseには複数のやり方がある



DemultiplexとDenoiseには複数のやり方がある



Minicondaはパッケージ管理ソフト

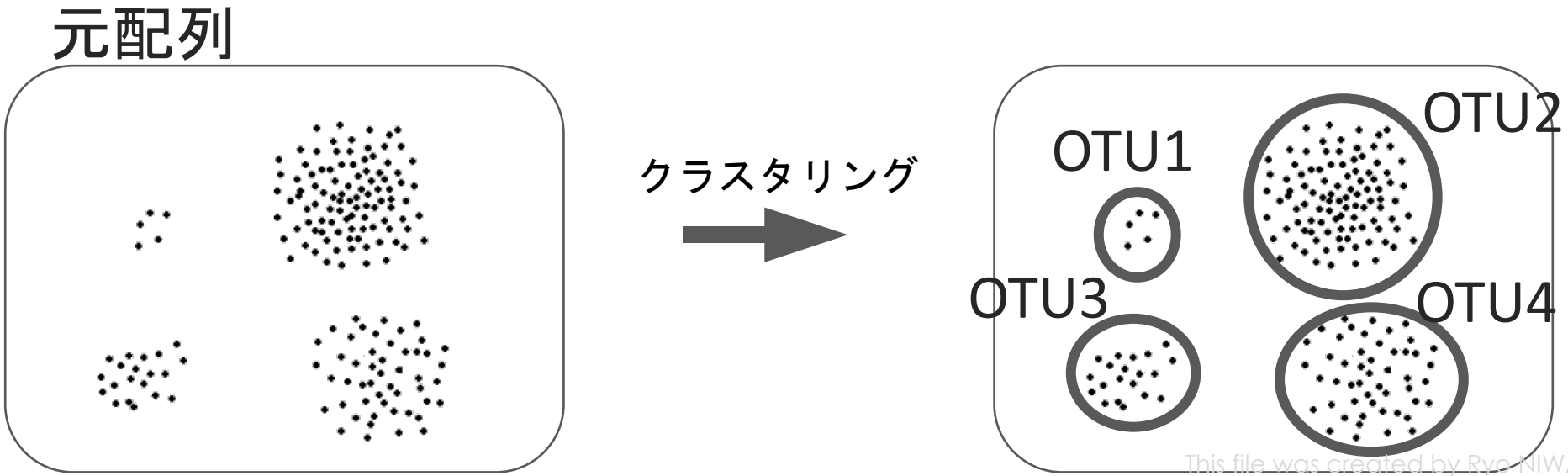
MinicondaはAnacondaの縮小版です。簡単に仮想環境を作る事ができ、Qiime2などのソフトウェアを楽に扱う事ができます。

今回は今後バイオインフォマティクスを利用する上で重要だと思うので、WSL上で作業してもらってますが、多分Anaconda Prompt上でも作業はできるかと思います。

DADA2はOTUとは異なるアルゴリズム

これまでillumina社製のシーケンサーでシーケンスを行なった場合のエラー補正はOTU (Operational Taxonomic Unit) と呼ばれる操作的分類単位で行われてきた。

基本的には3%の閾値で類縁するリード (97%以上同じ配列) をまとめ上げて、代表配列で配列の参照をすることでエラーを減らす。



OTUは微細なスケールの遺伝情報を見逃すこともある

OTUでは低頻度出現配列を排除するアルゴリズムが実装されており、生態学的に重要なニッチを見逃す可能性がある。

そこでDADA2を用いる。DADA2はクオリティーを活用しシーケンスエラーを排除する計算アルゴリズム。

OTUとしてまとめる訳ではないので、**1塩基の違い**まで見つけ出す事ができる。

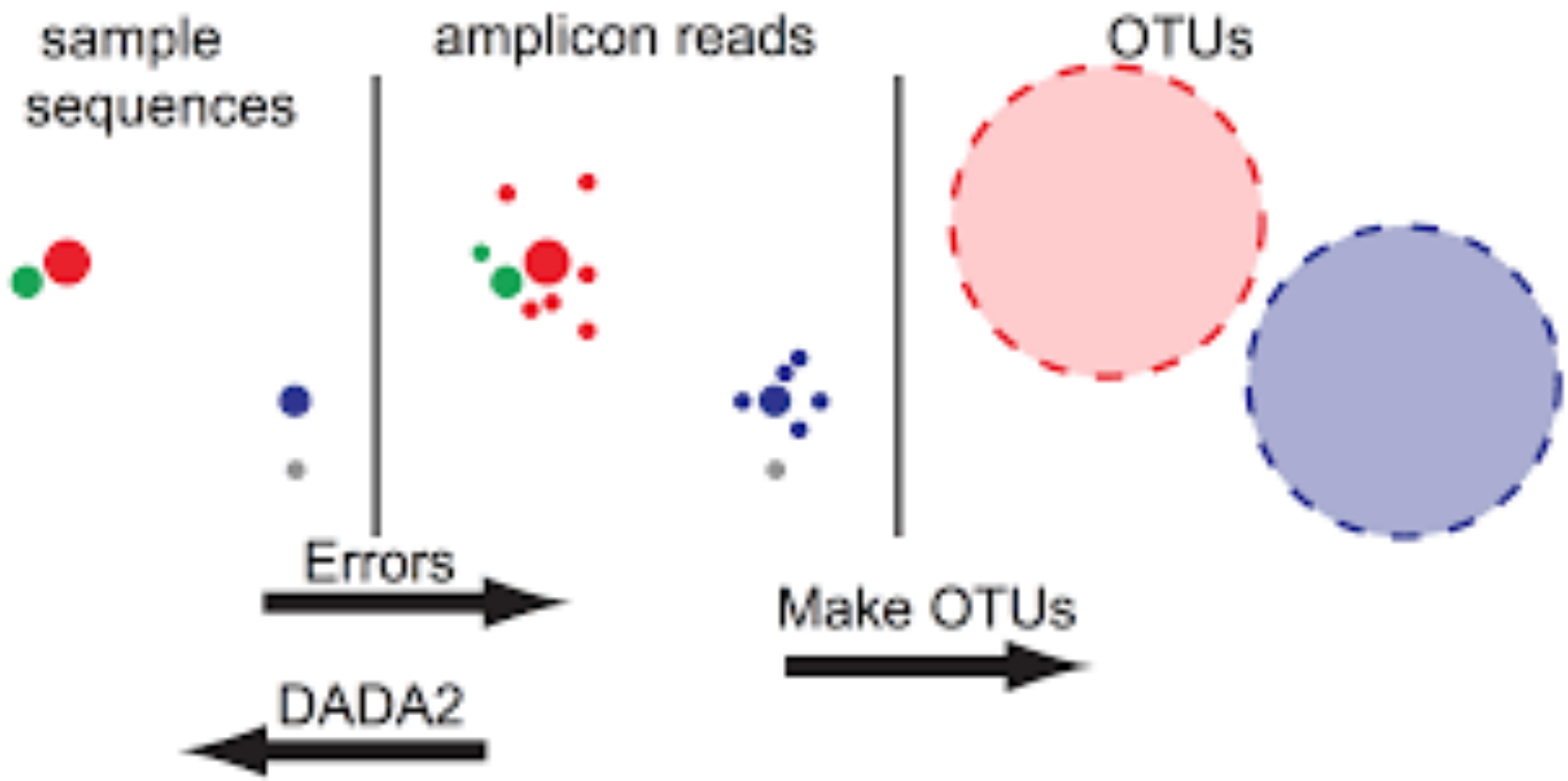
DADA2は統計的処理によりエラーを推定する

DADA2は独自のアルゴリズム (同一配列のシーケンス量、クオリティーなどをパラメータとする) を用いて、1塩基程度の違いのリードを生物学的変異なのかどうか推定できる。

このような方法はOTU法に対してASV (Amplicon Sequencing Variant) 法と呼ばれており、近年のシーケンス技術の根幹となっている。

ちなみにDADA2はキメラ配列の検出などもすべて自動でこなす。

OTU法とASV法の位置関係は逆方向



Bioconductorより引用

ただし、OTU法による解析は今も用いられている

ITSなどを解析する際に、OTUは活躍する。DADA2は16Sに特化した統計解析なので、ITSほどバラエティーに富んだ配列をDADA2でエラー除去すると配列が**除去され過ぎる**。

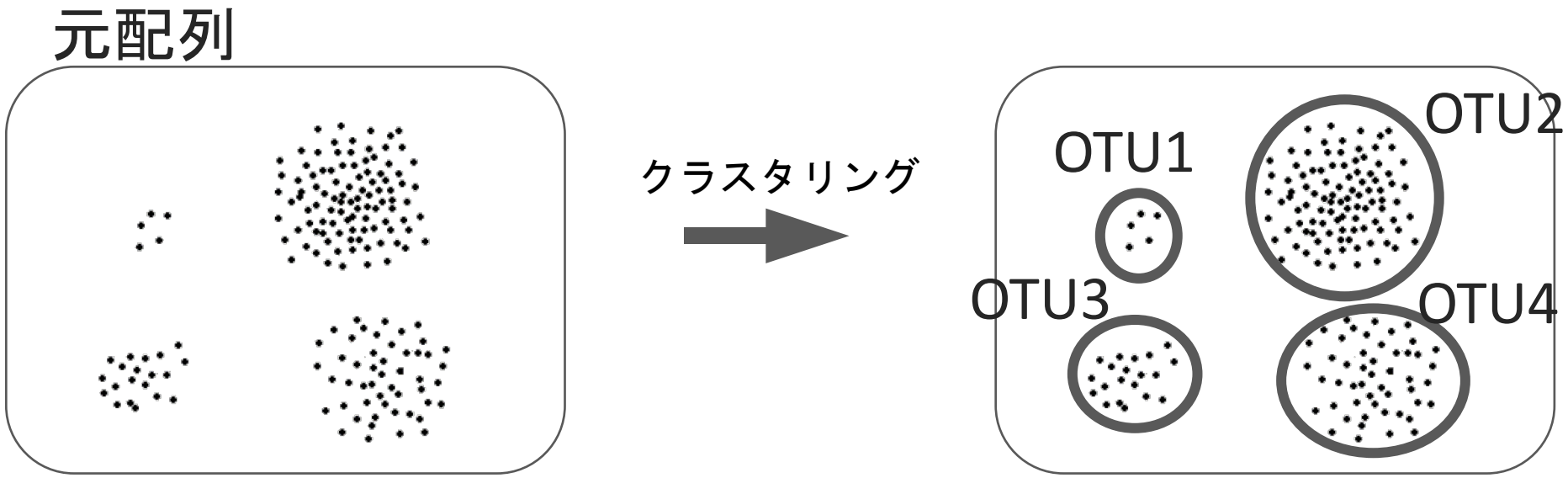
そこで、OTUが用いられる場合があるので、今回はOTUを作成し解析する方法も同時に紹介する。

なお、DADA2でもITS用にチューニングされたものがあり、それ用のプラグインも存在する。

*要参照: https://benjjneb.github.io/dada2/ITS_workflow.html

OTUには3種類の根本的考え方がある

ClaidentをベースにOTU解析を進めてきた当方だが、Claidentの説明書を見てもOTUについては実は詳しく述べられてない。ただし、Qiime2にはOTUの作り方が3種類あり、うち2つは未だによく用いられる方法なので要注意。



*De novo*法、*Open-reference*法、*Close-reference*法

・***de novo* clustering**

インプットとして与えた配列に対してのみクラスタリングを行う方法。例えば、僕が10個の配列を持っていたとしたら、この10個の配列同士で配列類似性を考えてクラスタリングを行います。

・**Closed-reference Clustering**

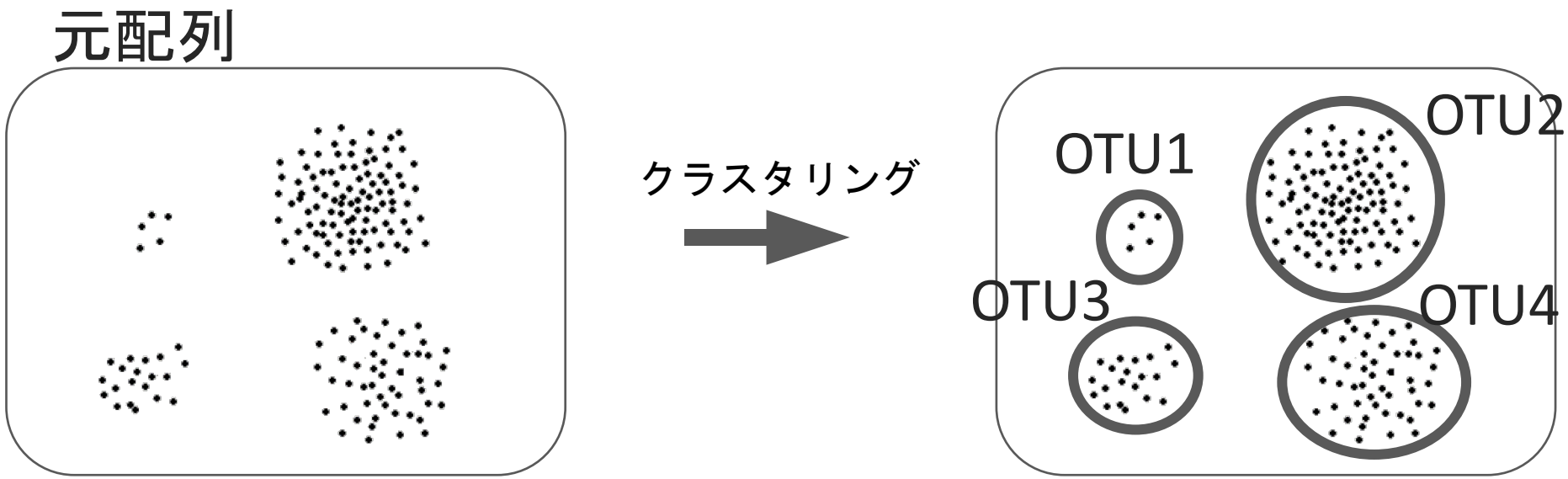
reference sequence collectionというデータベースに対して、配列類似性を見極めて計算を行います。何がClosedなのかというと、データベースとクエリーの配列類似性が閾値に達しなかった場合その配列を捨てられます。つまり、データベースに載っていない配列はカウントされません。

・**Open-reference Clustering**

Open-reference Clusteringでは、Closed-reference Clusteringとは対照的に閾値に達していなくてもカウントされます。この場合は*de novo*法で配列がまとめられます。

De novo法はインプット間での配列類似性！

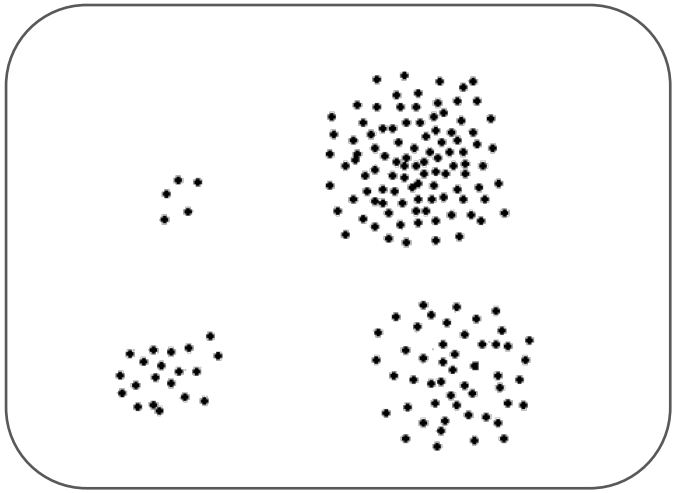
De novo法



Close-reference法はデータベース依存！

Close-reference法

元配列

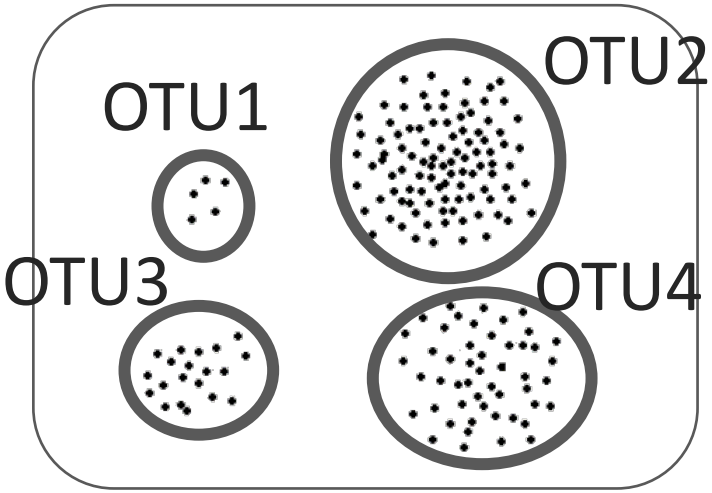


データベースに存在する配列

De novo同様にOTUとしてまとめられる。
→

データベースに存在しない配列

ゴミとして以下の解析には用いない。

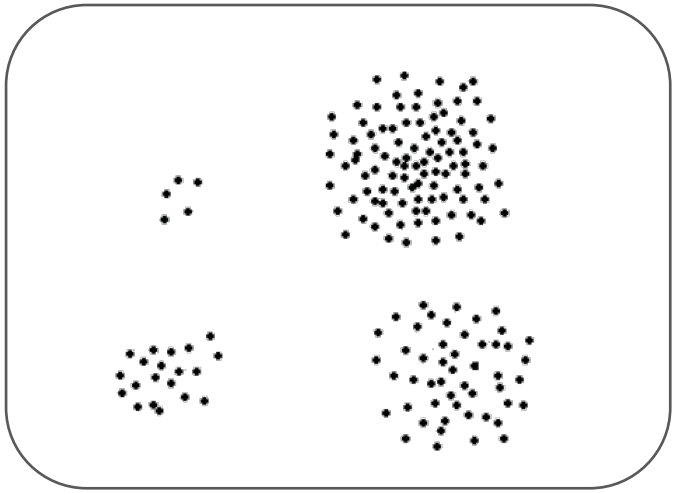


データベースに対してクラスタリング

Open-reference法はデータベース依存！

Open-reference法

元配列



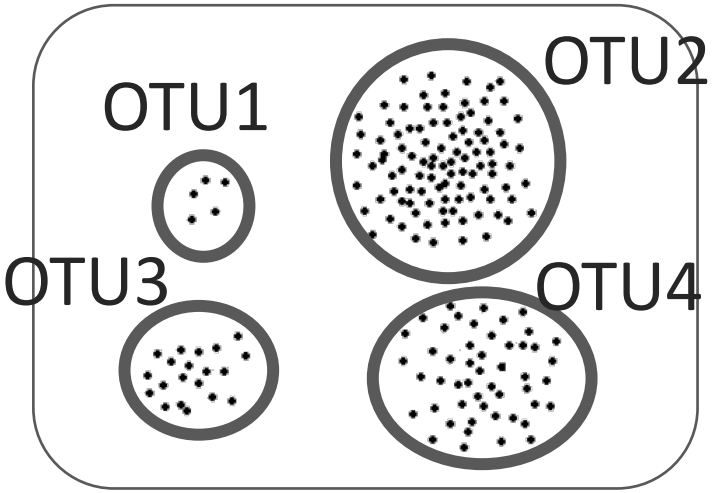
データベースに存在する配列

De novo同様にOTUとしてまとめられる。
→

データベースに存在しない配列

その配列だけがDe novo法にて処理されて残される。

→



データベースに対してクラスタリング

3つの方法の良いところ、ダメなところ

- ***de novo* clustering**

最も古くに考案された方法なので、参考にできる文献がたくさんある。多くのソフトウェアで実装されているので使いやすい。ただ、ゴミもOTUとしてカウントすることは否定できない (？)

- **Closed-reference Clustering**

再現性が取れる。同じデータベースを使えば、誰が行っても同じ結果を得られる。データベースの選択方法次第で、ゴミのカウントが増える。

- **Open-reference Clustering**

De novoとCloseの組み合わせなので新規サンプルにも対応でき、信頼性も担保できる。リファレンスがある領域なら、ベンチマークスコアも高い。

これまでに解説したことは解析の手法そのもの

Qiime2について

DADA2について

OTUについて などなど…

ただ、Qiime2をやって行う解析のゴールは？

- 菌叢を明らかにすること

- 多様性の比較

前回までにやったことは解析の手法そのもの

Qiime2について

DADA2について

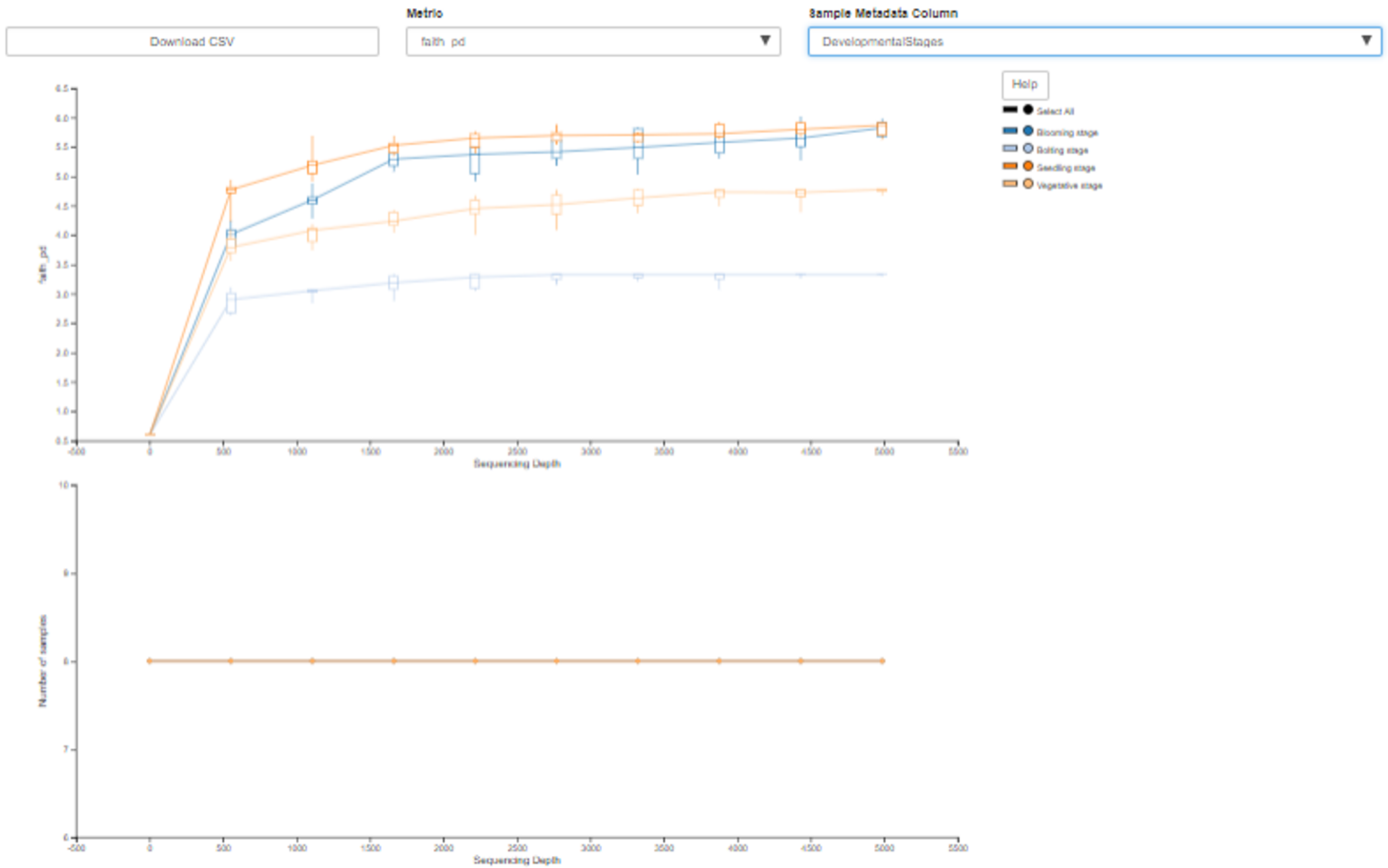
OTUについて などなど…

ただ、Qiime2をやって行う解析のゴールは？

~~**- 菌叢を明らかにすること**~~

- 多様性の比較

多様性解析は少しだけ工夫が必要



シーケンスによりサンプル間のシーケンス深度は異なる。

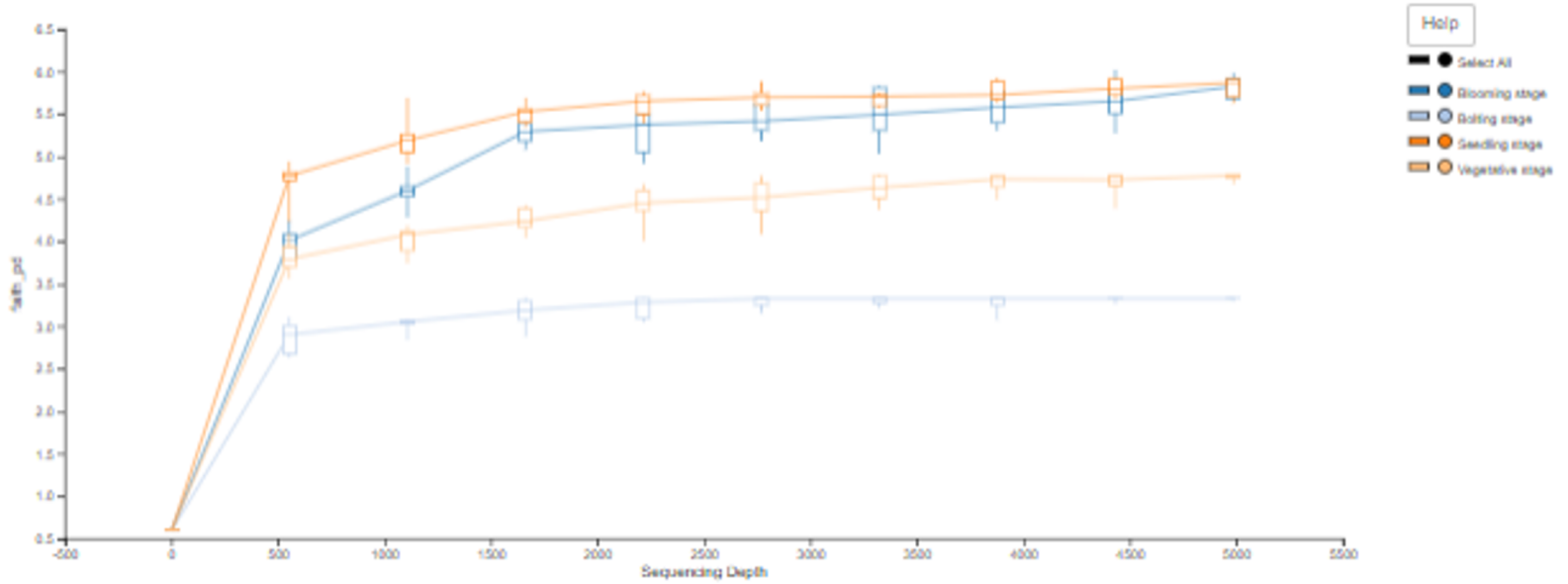
通常シーケンス深度が大きくなるほど、生物多様性は大きくなる。

その多様性はどこかで平衡状態になっているはずだが、単純に比較すると得られたリードが少ないサンプルほど多様性が小さくなるかもしれないので、

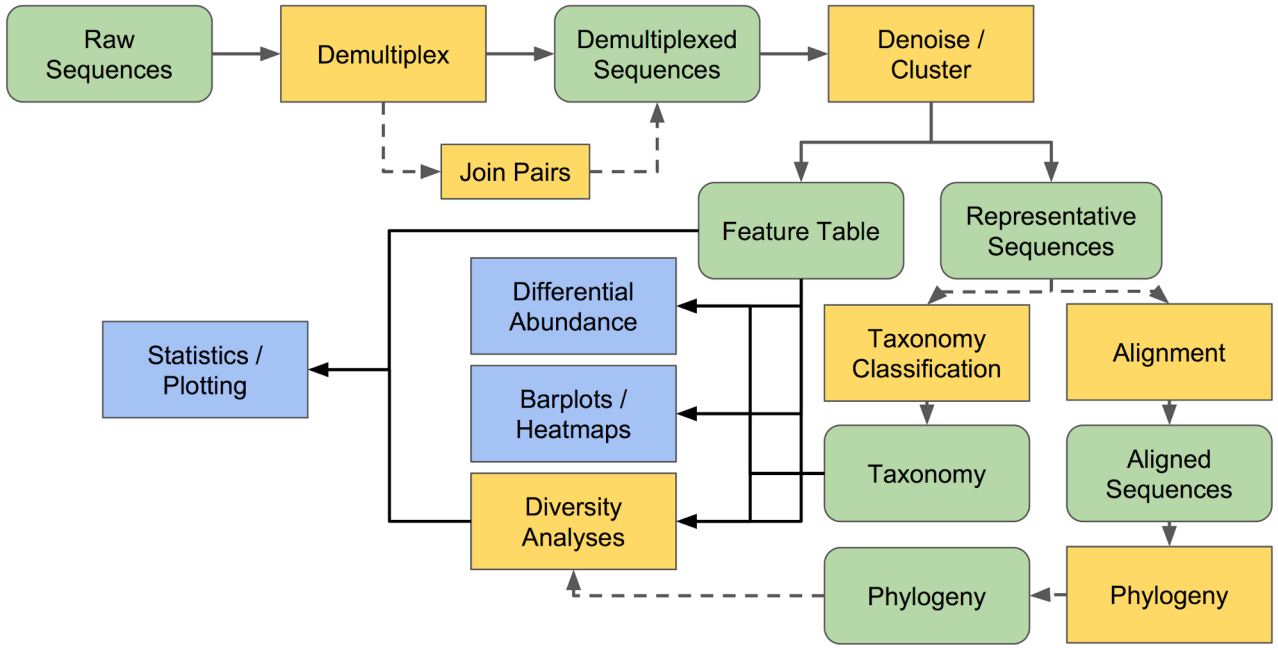
希薄化 (サンプルを薄めて) して解析を行う。

Alpha-rarefactionは多様性解析に有効

サンプルの希薄化はアルファ多様性ベースに行う。
アルファ多様性が上昇しないならシーケンス深度として十分と仮定。
リード数を減らす水準になる。下記なら2500リード以上くらいでサチっている。



今回の解析フロー



前回同様にDenoiseを行った後、配列からPhylogenyの解析を行う。(配列間のIdentityを検索) Phylogenyデータ からDiversity Analysisを行う。アルファ多様性解析を行い、それをベースに希薄化。最終的に**ヒートマップ**等を作成し、サンプルの様子を比較する。

今回はShannon多様性指数を利用したアルファ多様性解析とJaccard距離を利用したベータ多様性解析を行う。

なお、Qiime2では次のような多様性計算が可能。先行研究やトレンドに従い、多様性測定を行うのがベター。

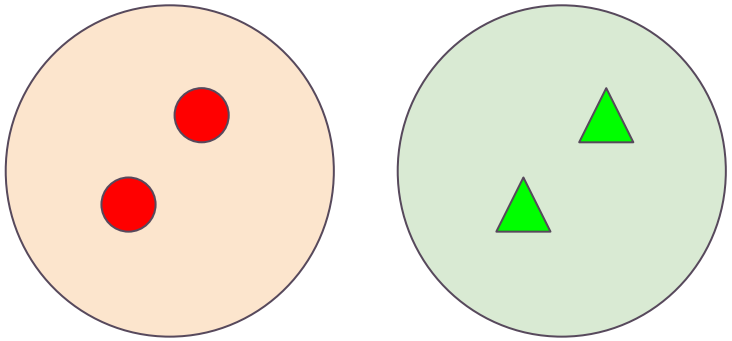
Alpha Diversity
- Shannon's diversity index
- Observed OTUs
- Faith's Phylogenetic Diversity
- Evenness

Beta Diversity
- Jaccard distance
- Bray-Curtis distance
- unweighted UniFrac distance
- weighted UniFrac distance

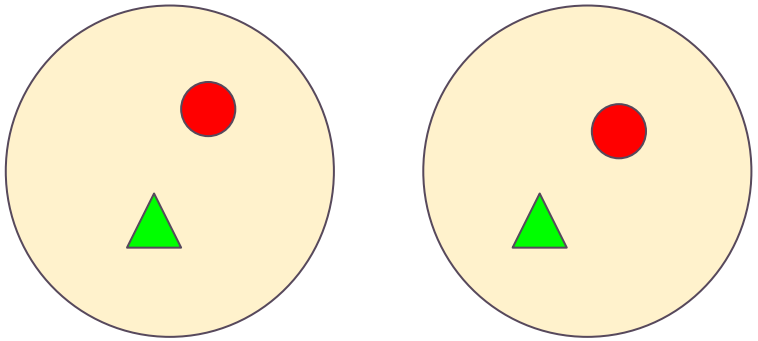
α多様性とβ多様性は比較の対象が異なる

α：ある地域内での多様性
β：ある地域と地域の間での多様性
γ：全体での多様性

α: 1, β: 2, γ: 2

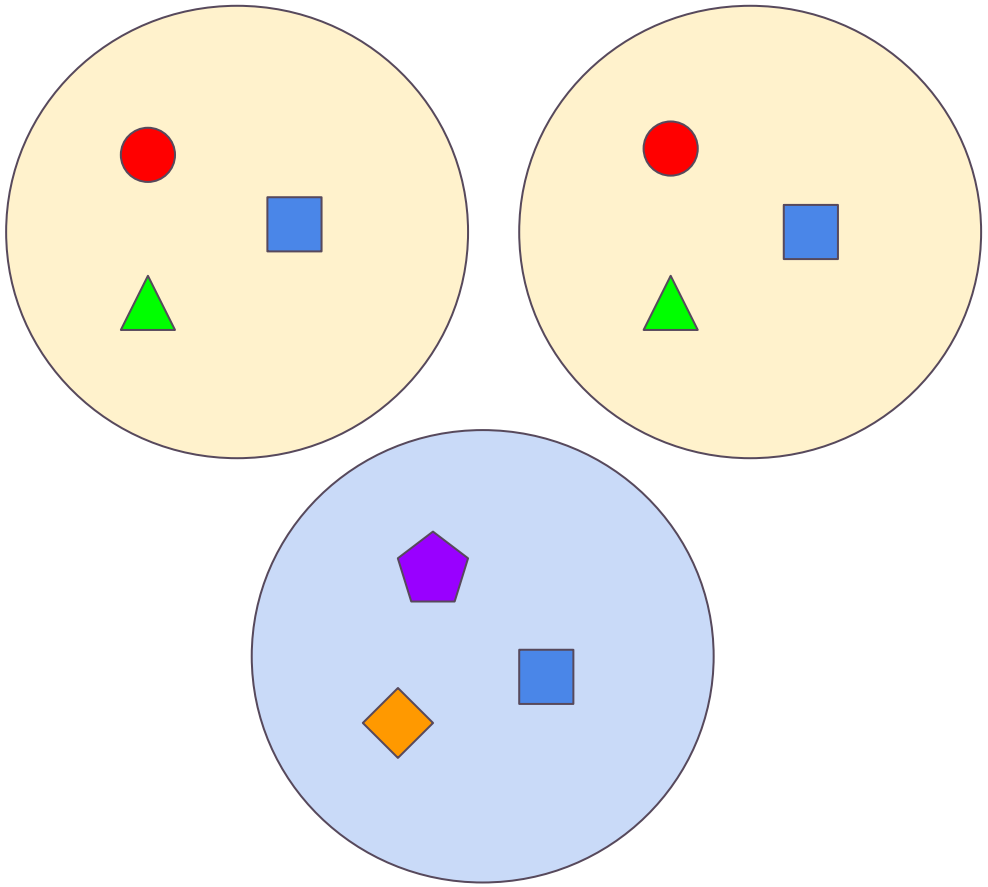


α: 2, β: 1, γ: 2

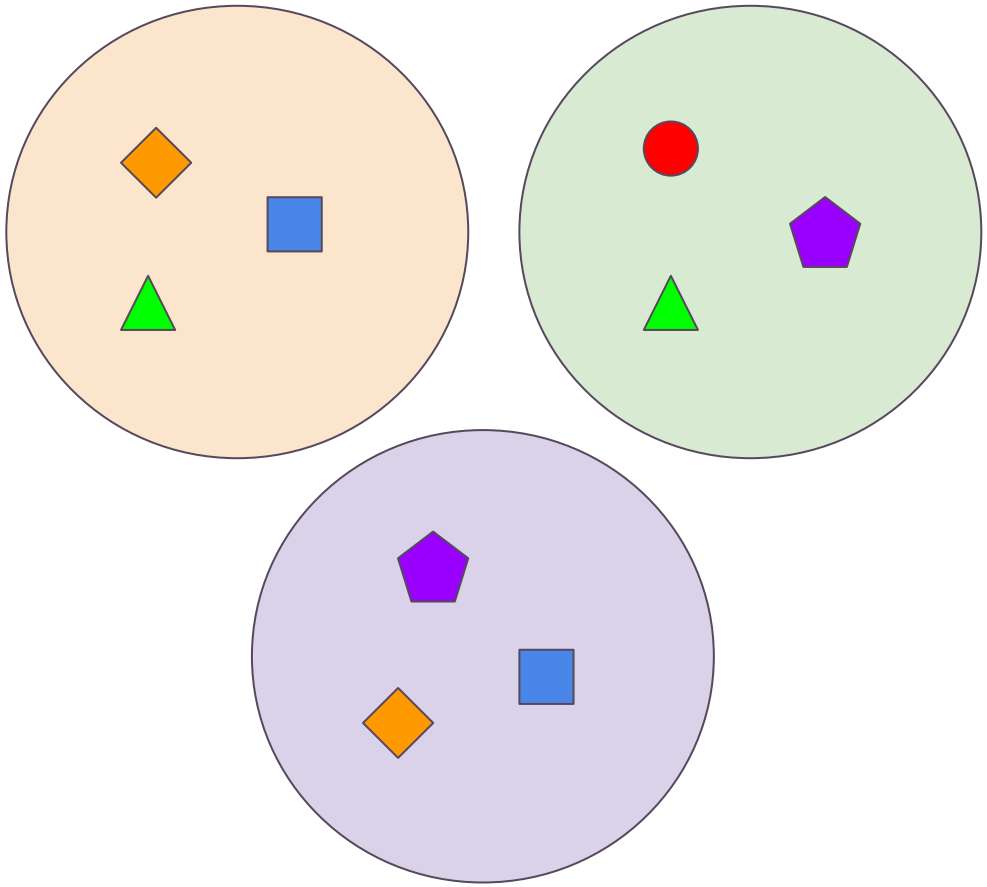


α多様性とβ多様性の考え方の例

α: 3, β: 2, γ: 5

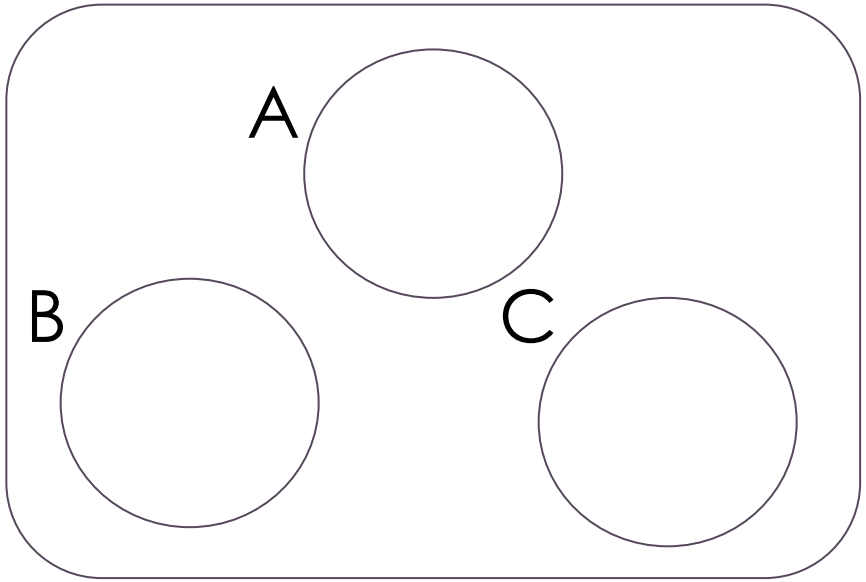


α: 3, β: 3, γ: 5

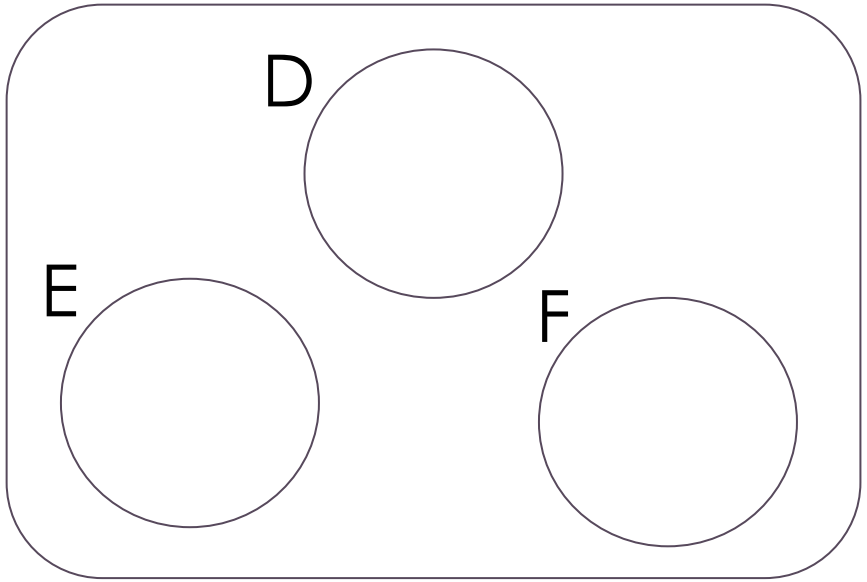


DADA2とDeblur の違いは統計の考え方！

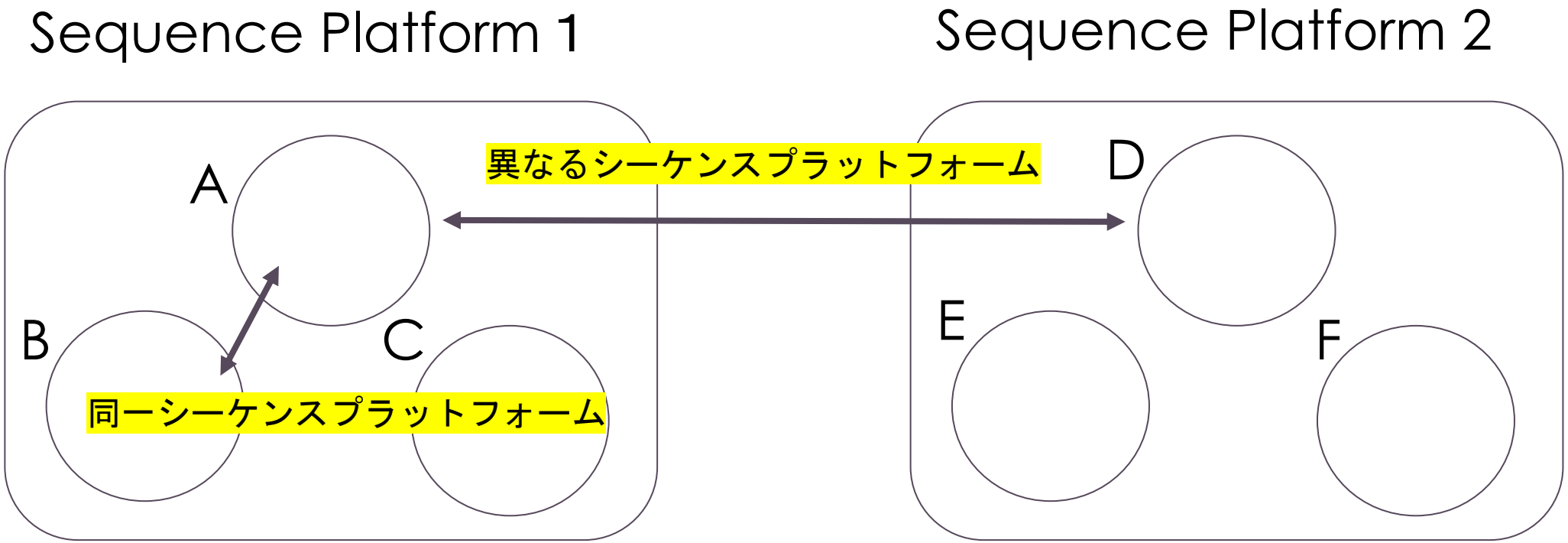
Sequence Platform 1



Sequence Platform 2

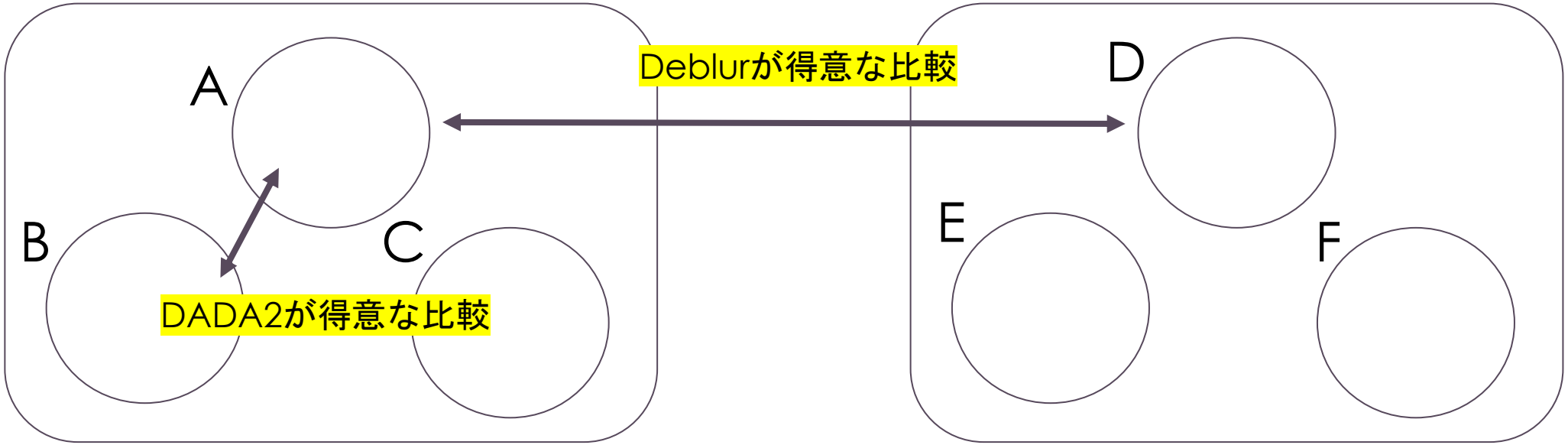


DADA2とDeblur の違いは統計の考え方！



DADA2とDeblur の違いは統計の考え方！

DADA2は同一シーケンシングのエラー寄与率の計算を得意とし、Deblurはサンプルごとのエラー計算を得意とする。



Estaki, M., Jiang, L., Bokulich, N. A., McDonald, D., González, A., Kosciolek, T., Martino, C., Zhu, Q., Birmingham, A., Vázquez-Baeza, Y., Dillon, M. R., Bolyen, E., Caporaso, J. G., & Knight, R. (2020). QIIME 2 enables comprehensive end-to-end analysis of diverse microbiome data and comparative studies with publicly available data. *Current Protocol. Bioinformatics*, 70, e100. doi: 10.1002/cpbi.100

あなたがやりたい解析は？

Qiime2でできることの大抵はここに記載しているかと思いますが、1つのデータをとっても、できることは非常に多いです。

どんな解析をしたいのか、何を見つけたいのかを改めて、考え直して作図、解釈をしていく必要があるのかなと思います。