Metabarcoding Pipeline Building

An introduction with hands on exercises

CUSO – DPEE Activity

Gerhard Thallinger, PhD, Graz University of Technology

Rachel Korn, PhD, Université de Fribourg

Magdalena Steiner, PhD, Agroscope Wädenswil

Starting point

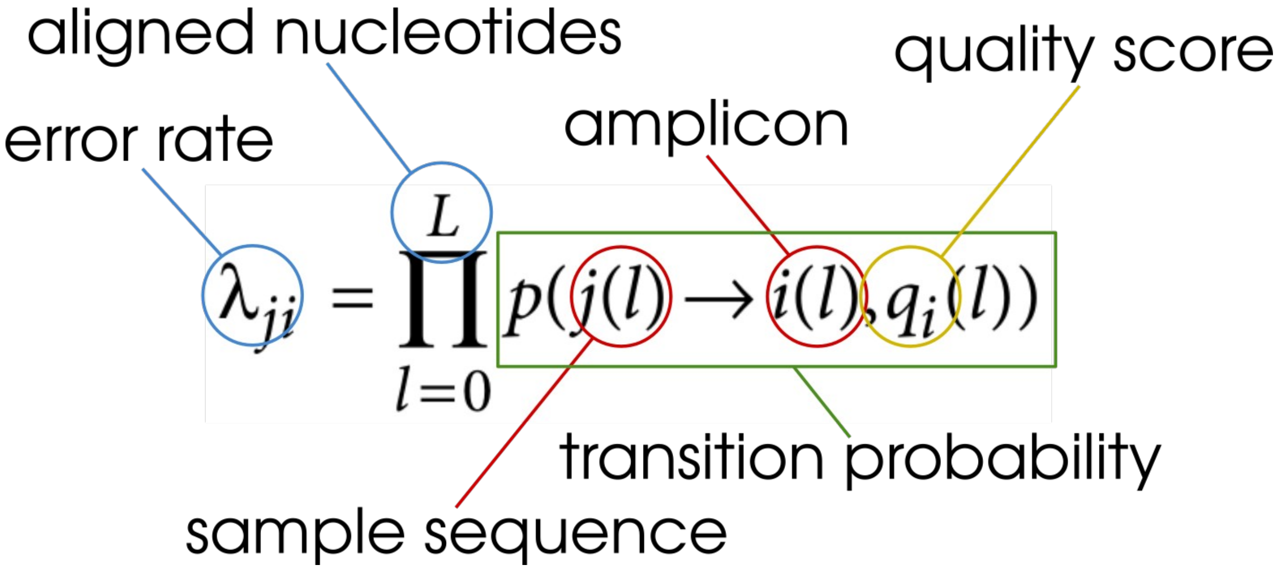
Amplicon Sequencing. **Exactly.** *Version 1.18*

- Demultiplexed samples without technical sequences
- Matching order of forward and reverse reads between FASTQ files

Callahan et al. 2016

Error model

- Alignment
- Error model
  - Transition probability ~ original nucleotide, substituting nucleotide and associated quality score



$$\lambda_{ji} = \prod_{l=0}^{L} \boxed{p(j(l) \rightarrow i(l), q_i(l))}$$

Callahan et al. 2016

Error model

- Abundance *p*-value p_A
  - Excludes all singletons

expected # reads of *j*

abundance

$$p_A(j \to i) = \frac{1}{1-\boxed{\rho_{\text{pois}}(n_j\lambda_{ji},0)}} \sum_{a=a_i}^{\infty} \boxed{\rho_{\text{pois}}(n_j\lambda_{ji},a)}$$

Poisson density function

Callahan et al. 2016

Divisive Amplicon Denoising Algorithm

(1) Unique sequences into one partition

(2) Center = most abundant sequence

(3) All unique sequences are compared to their center, error rates & p_A are calculated

(4) If p_A is small enough, its sequence forms a new partition

(5) All unique sequences join the partition most likely it has produced it

→ back to (3) etc.

Callahan et al. 2016

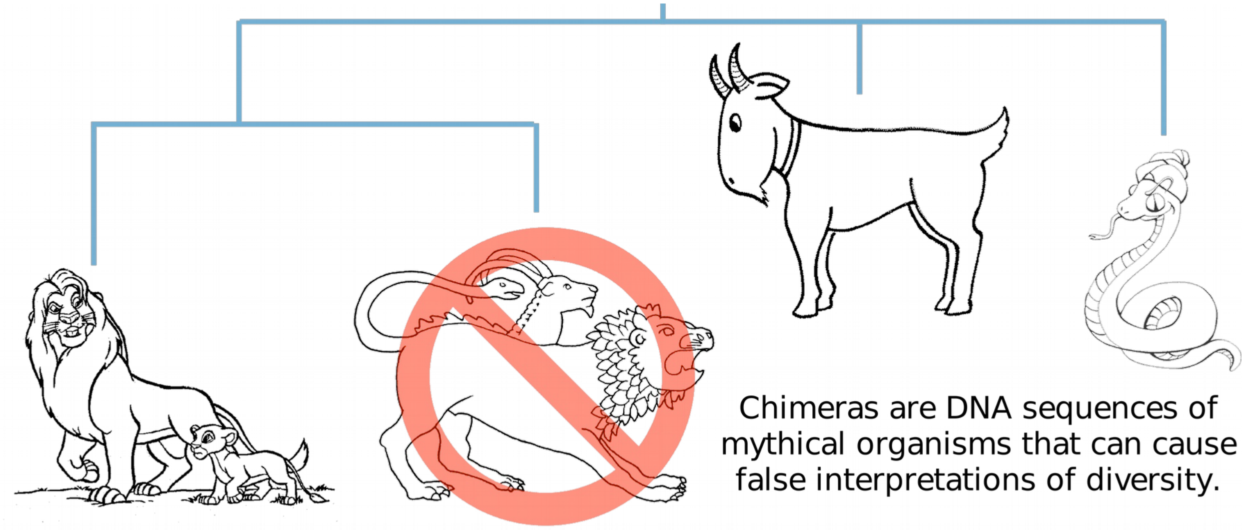
Merge paired reads & construct ASVs

- Aligns the denoised forward reads with the reverse-complement of the corresponding denoised reverse reads → “contigs”
- Default: 12 matching and overlapping nucleotides
- Although not recommended, non-overlapping reads are supported

Callahan et al. 2016

Chimera detection

- If reconstruction of a contig is possible by combining a left- and a right-segment from two more abundant “parent” sequences, it is considered as chimera



Chimeras are DNA sequences of mythical organisms that can cause false interpretations of diversity.

Wright et al.

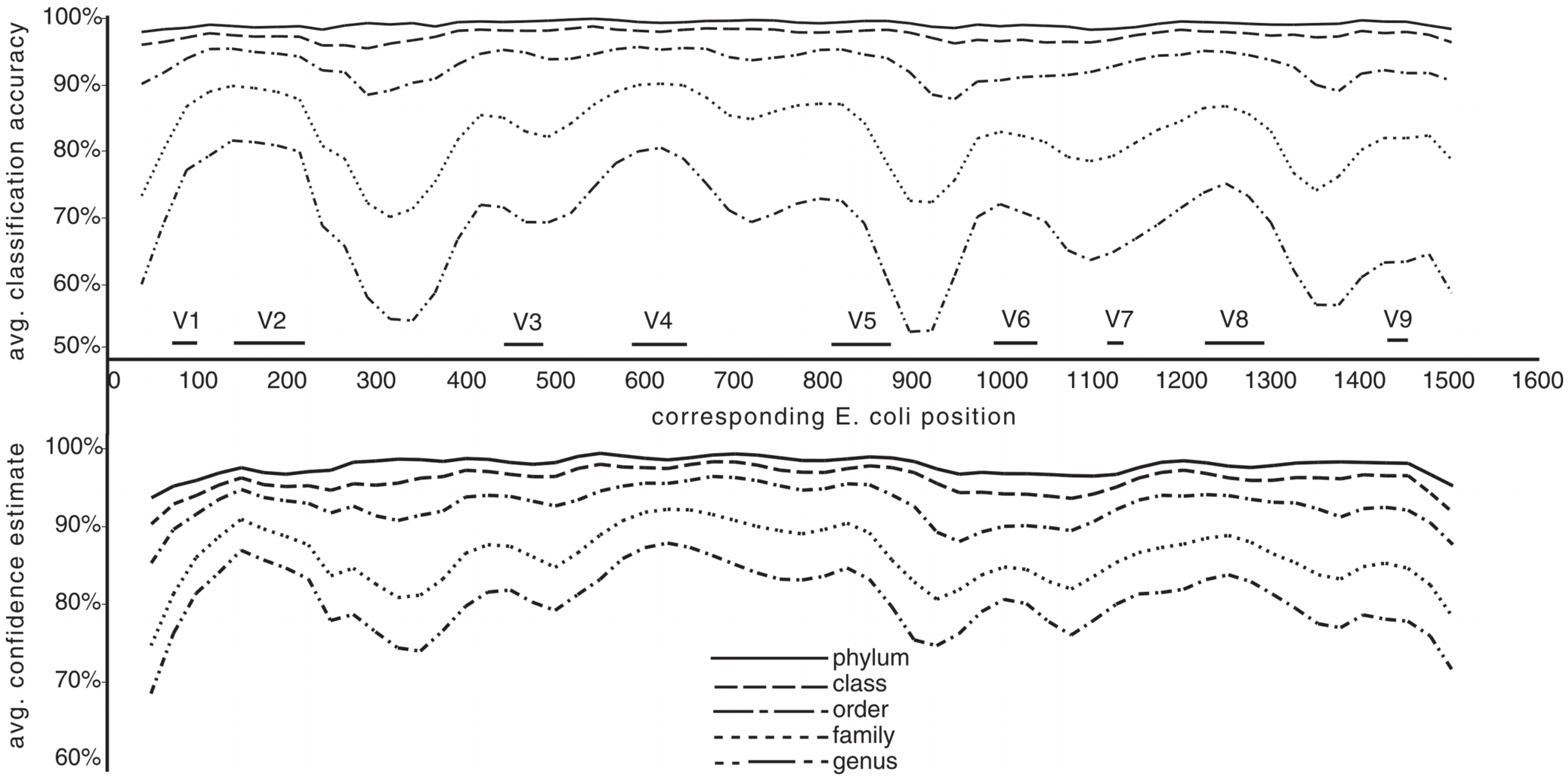
Assign taxonomy

- Classifiers
  - RDP classifier
  - IDTAXA

Wang et al. 2007, Murali et al. 2018

RDP classifier

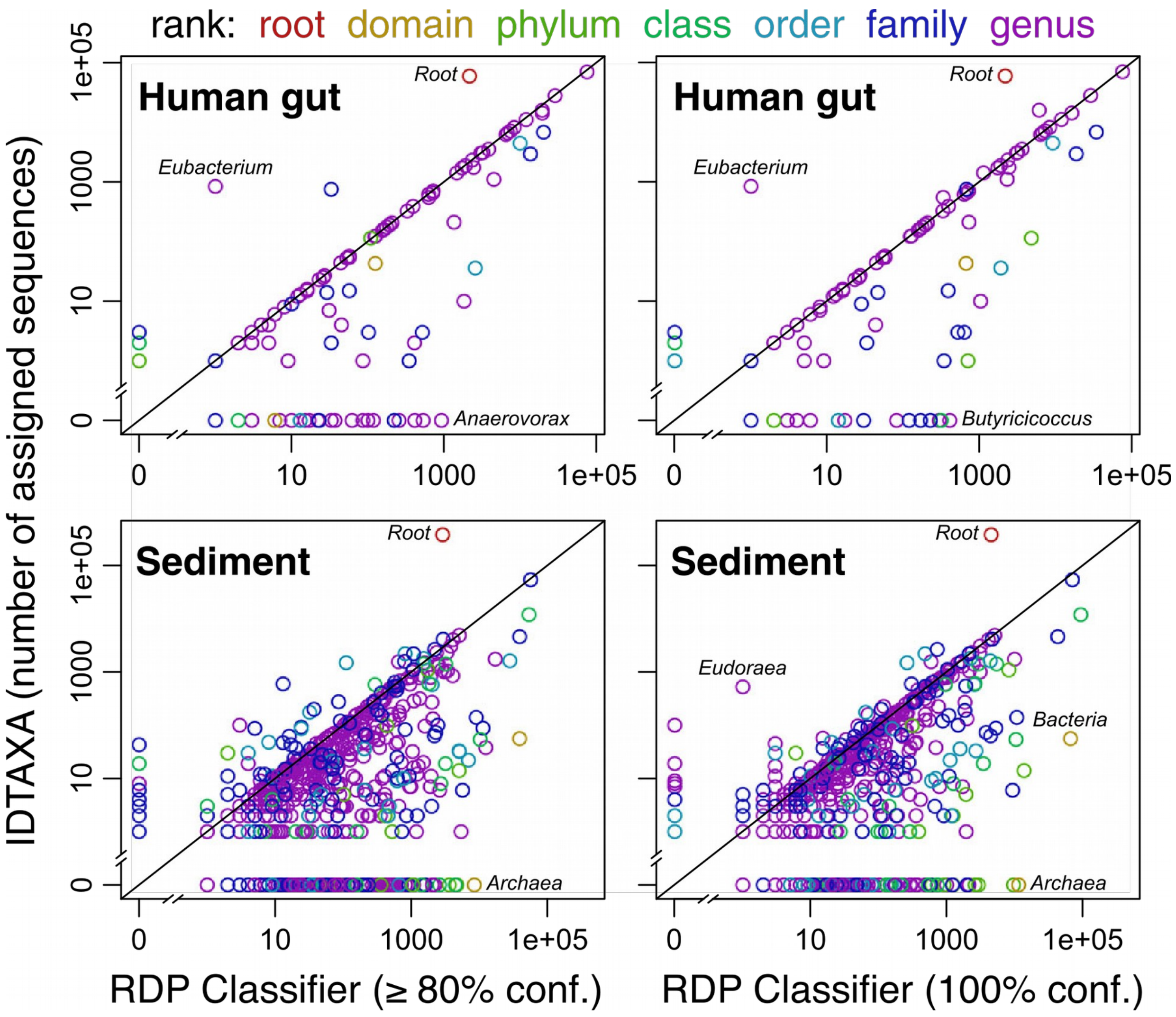
- Ribosomal Database Project Classifier
- Naïve Bayesian classifier



Wang et al. 2007

IDTAXA

- DECIPHER R package
- Hybrid classification combining phylogenetic, distance-based and machine learning techniques
- Reduced over classification



Murali et al. 2018

Assign species

- Exact string matching = 100 % sequence identity necessary