Approaches for natural and synthetic community amplicon data analyses

Ruben Garrido-Oter
Max Planck Institute for Plant Breeding Research

DECRyPT bioinformatics workshop - October 2019

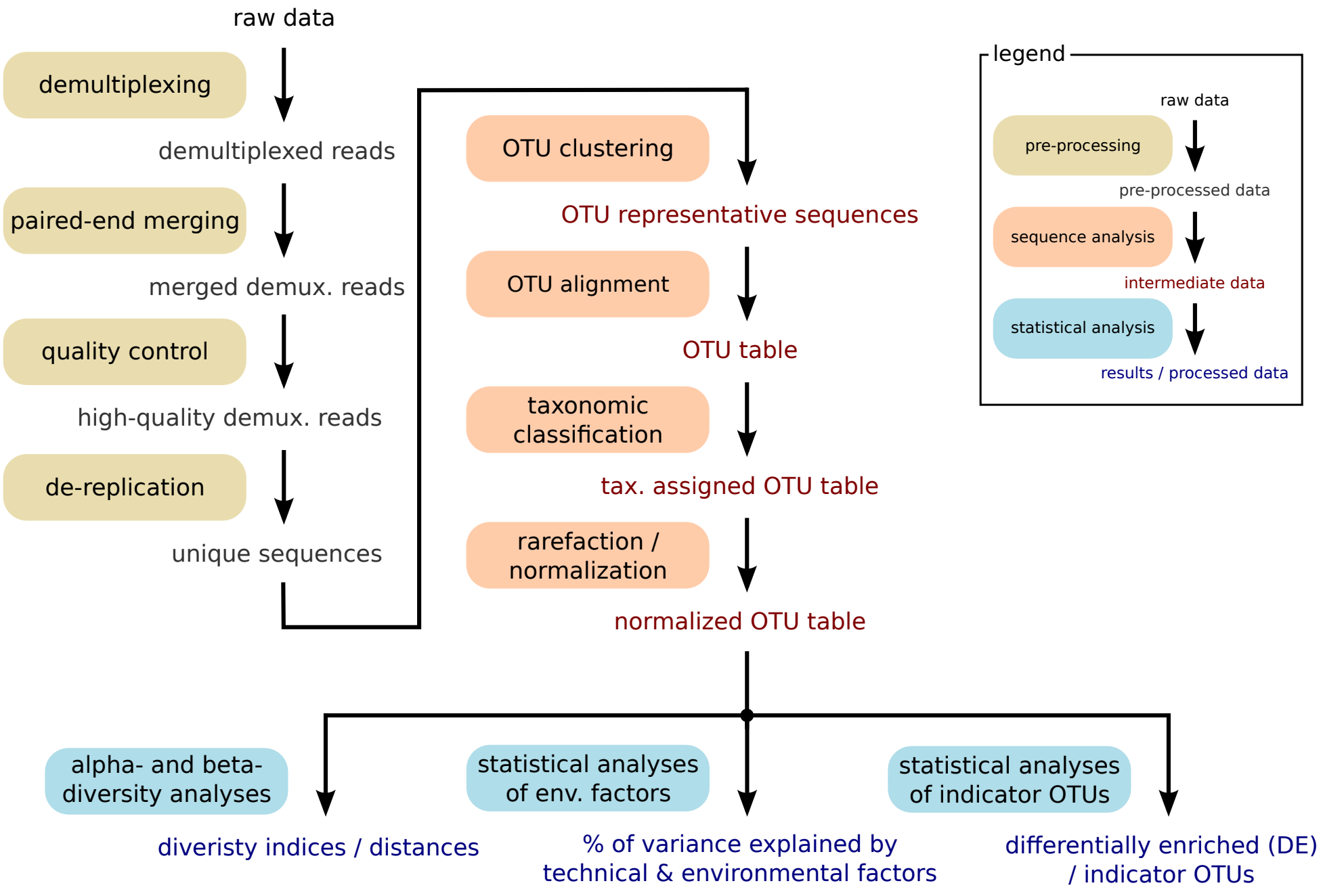
Commonly used toolkits for amplicon sequence analysis

qiime

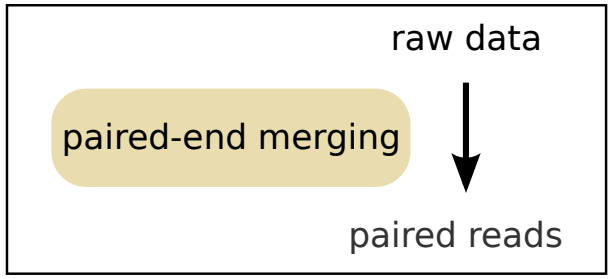
DADA2

Amplicon Sequencing. Exactly. Version 1.12

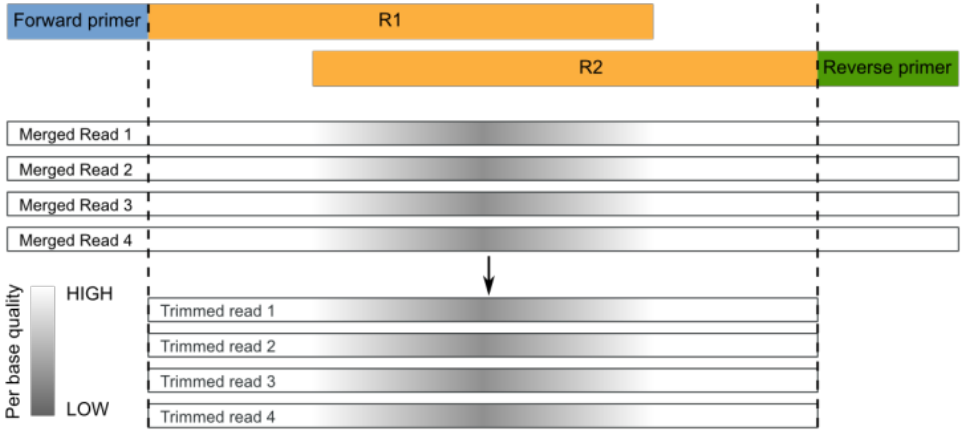
Workflow amplicon data analysis (natural communities)



Merging paired-end reads (Illumina)



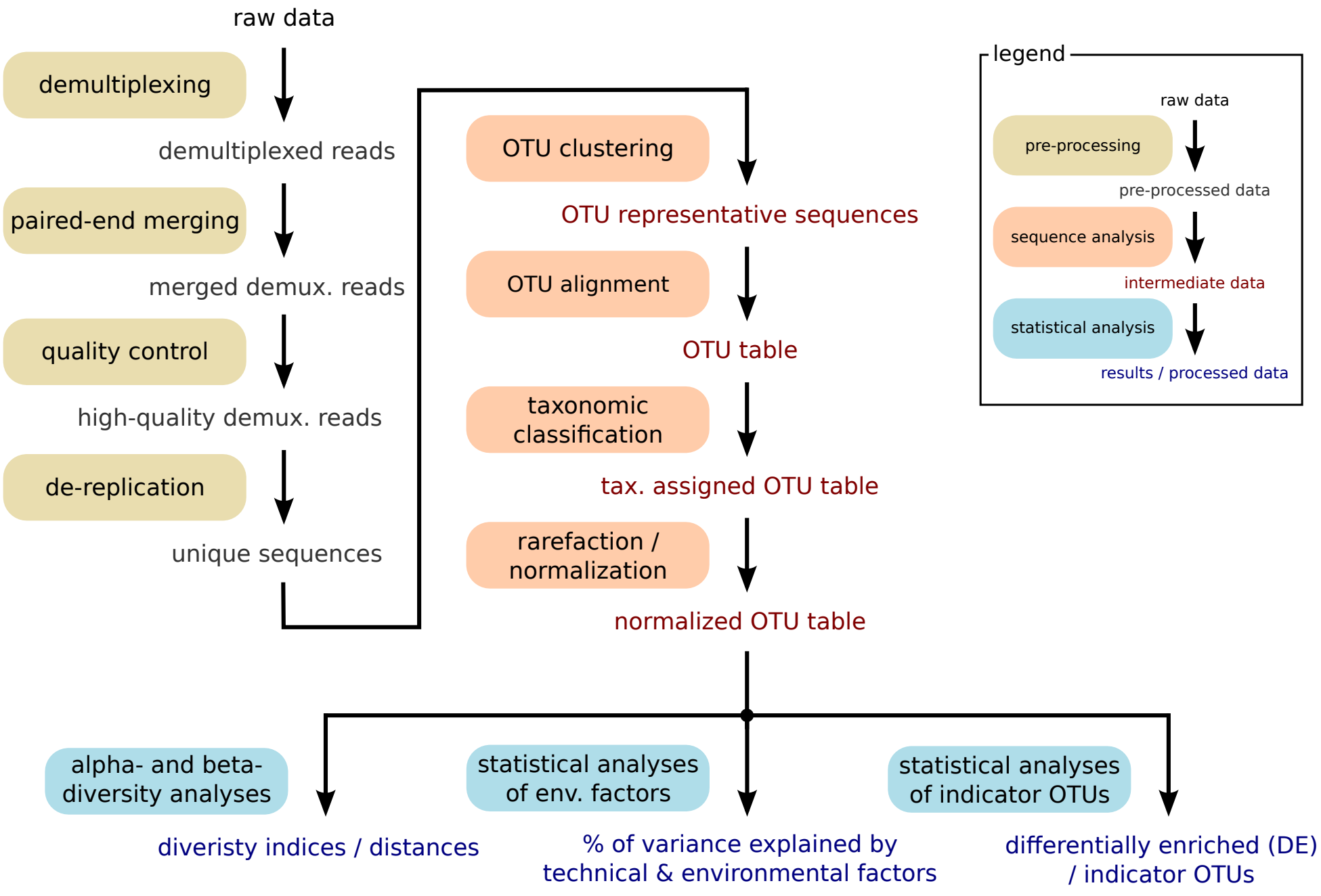
Consists on merging (assembling) paired-ed reads into consensus sequences (and consensus quality scores for downstream filtering)



Poorly overlapping pairs are (generally) discarded for fixed-length markers

For overlapping pairs quality scores need to be recomputed

Workflow amplicon data analysis (natural communities)



Quality filtering of amplicon reads

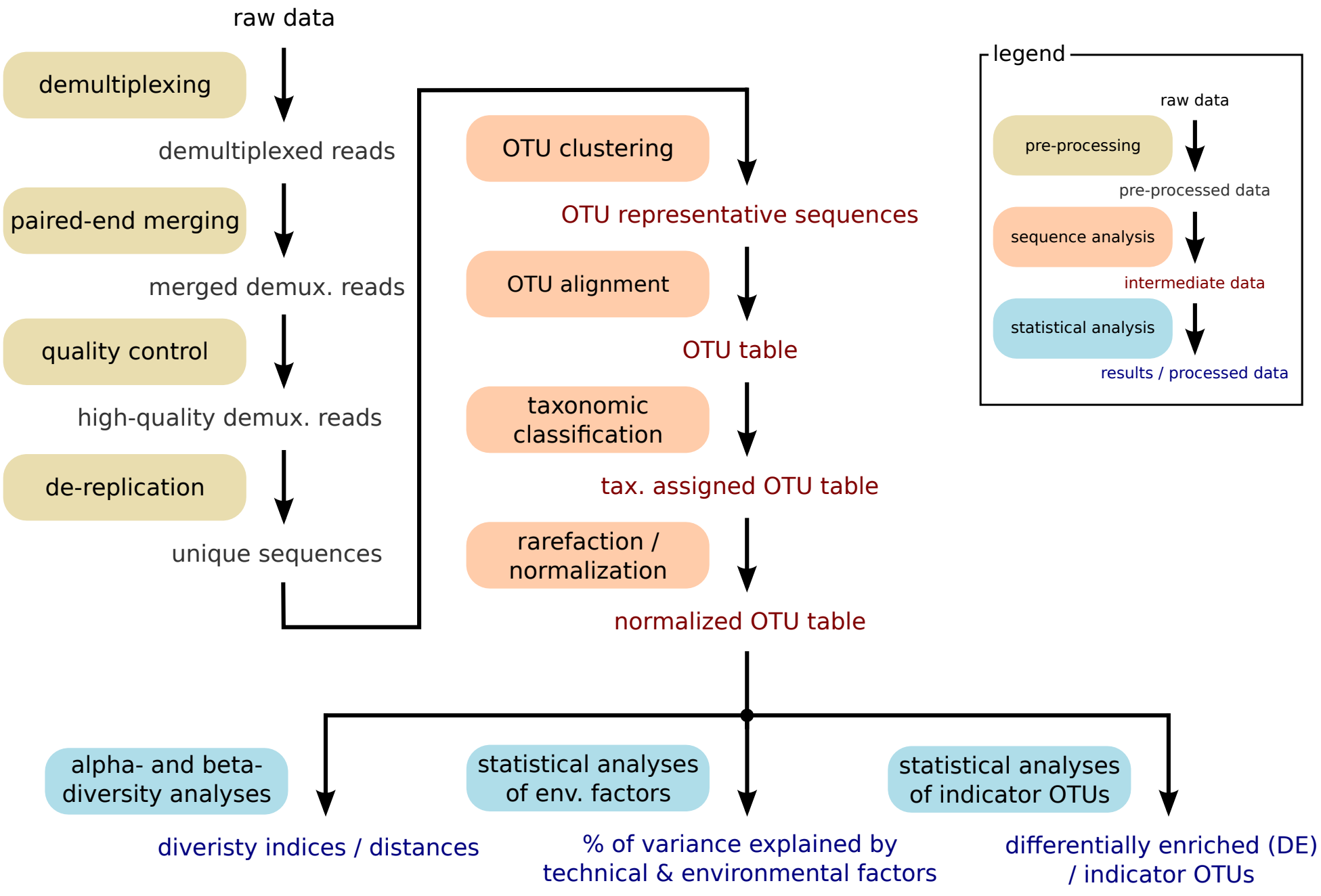
NGS instruments (454, Illumina) indicate the probabilities of sequencing errors using quality (Phred or Q) scores

ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

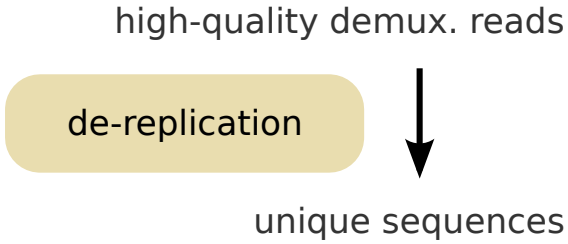
Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

For amplicon data, it is very difficult to distinguish PRC artifacts (SNP errors / chimeras) from sequencing errors

Workflow amplicon data analysis (natural communities)



De-replication

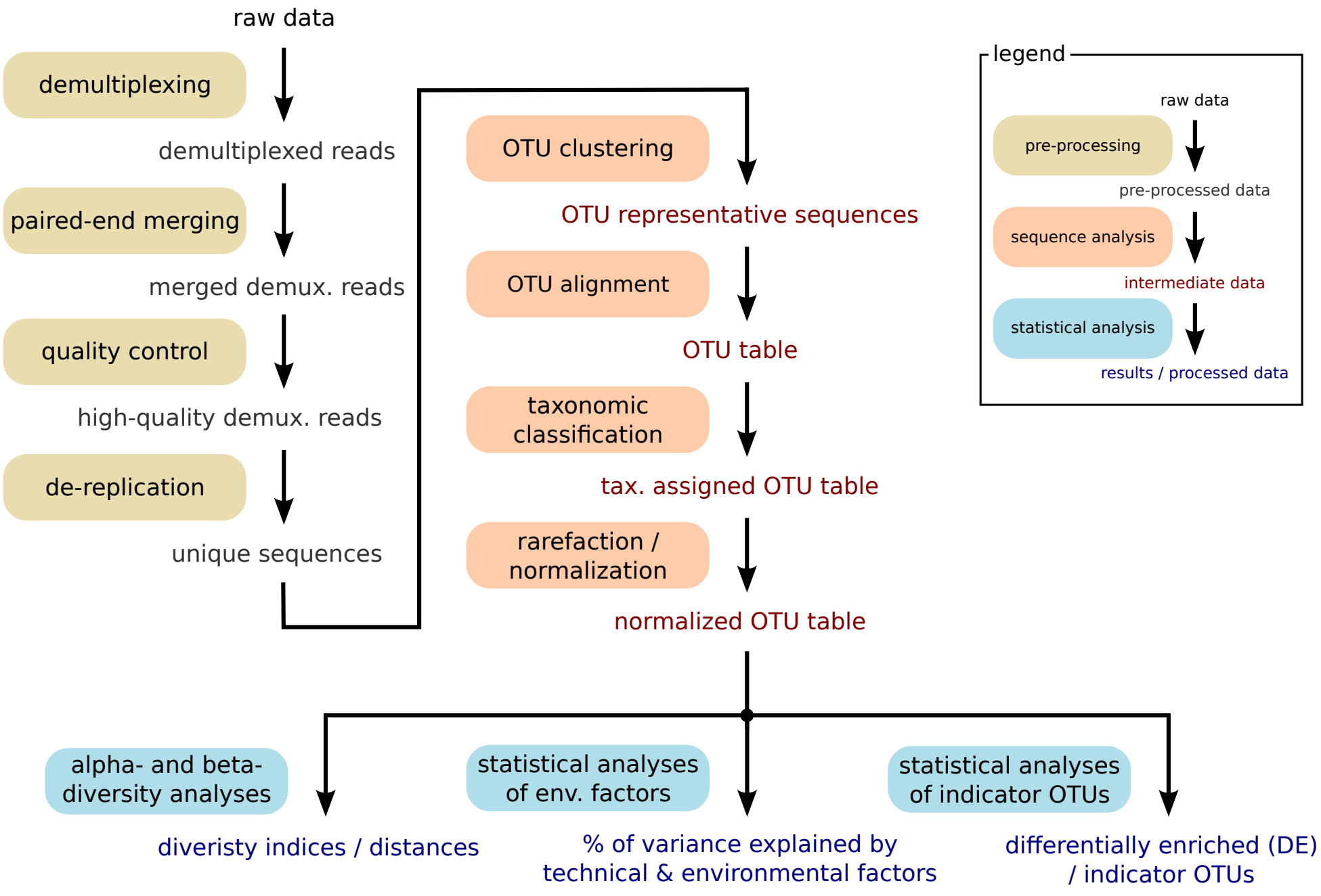


Some steps (e.g. OTU clustering and taxonomic classification of representatives) only need to be done once per set of identical reads

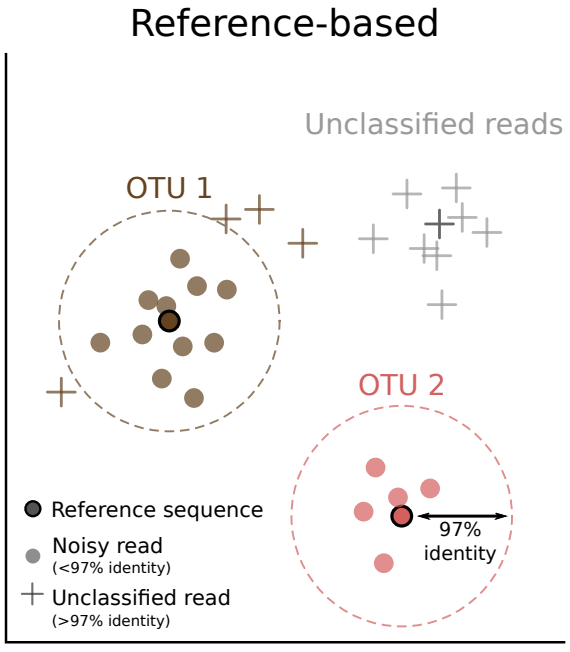
It is generally advisable to remove singletons. The basic assumption is that these are usually artifacts and that errors seldom occur multiple times on the same template sequence

Mapping of non-replicated reads onto OTU representatives is necessary to obtain accurate abundances

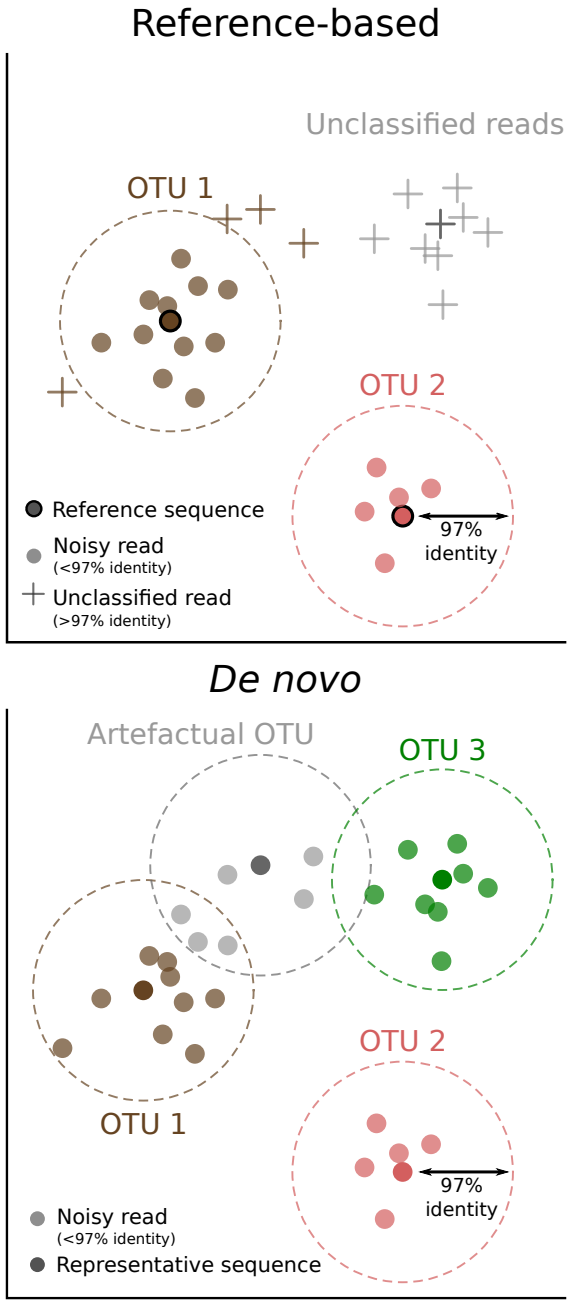
Workflow amplicon data analysis (natural communities)



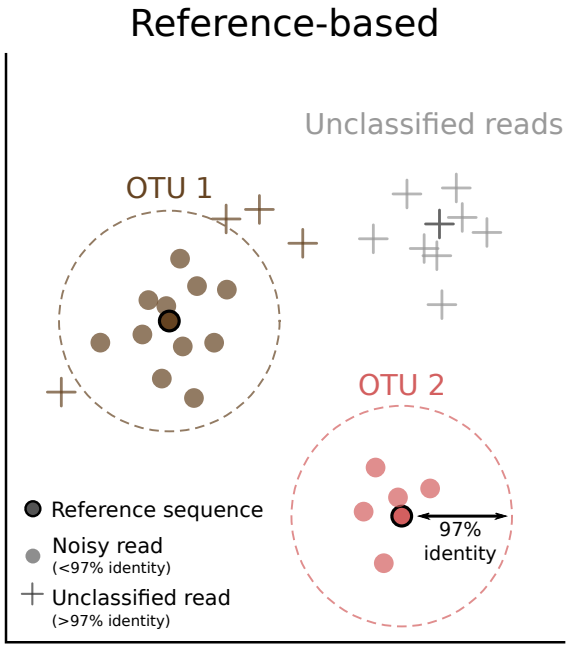
OTU clustering

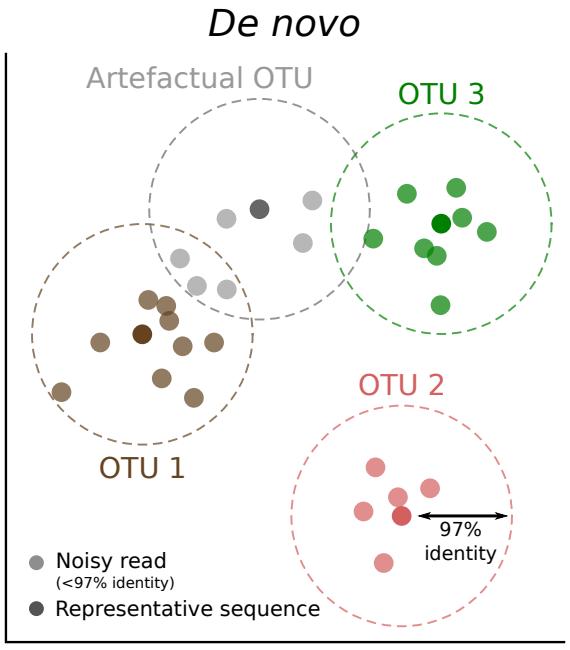


OTU clustering



OTU clustering

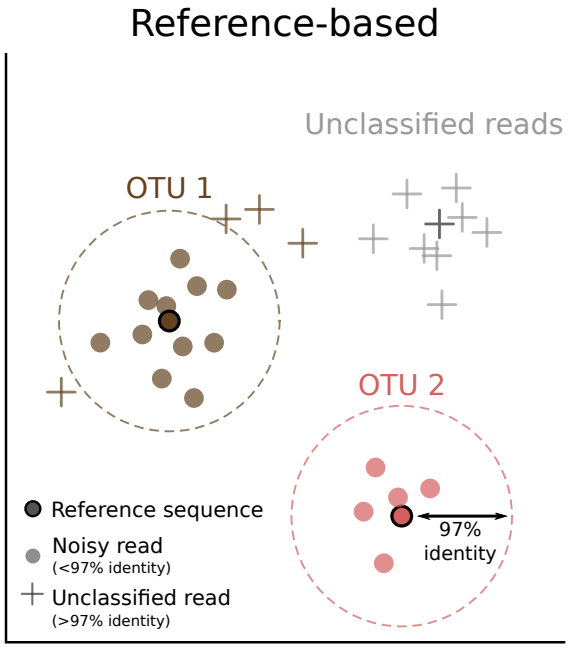


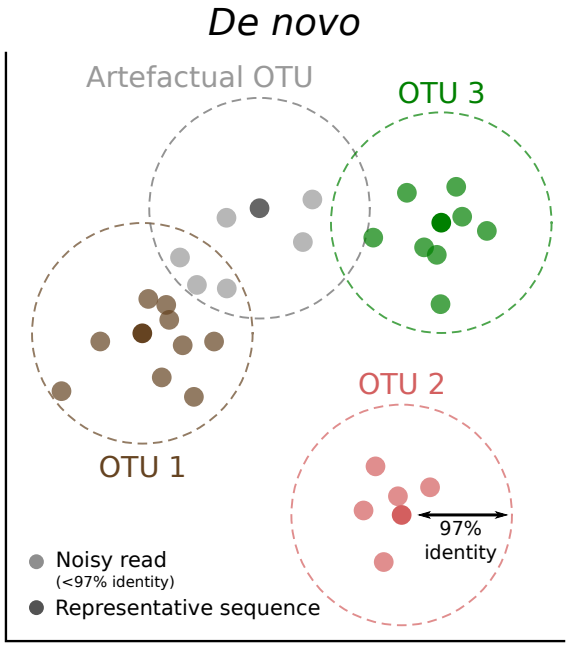


General problems associated with the use of OTUs:

- Low of resolution (typically 97% sequence identity)
- Lack of intrinsic biological meaning
- Inflation of alpha-diversity (artefactual OTUs)
- Low replicability and reproducibility
- Difficult to cross-reference between datasets / studies

OTU clustering





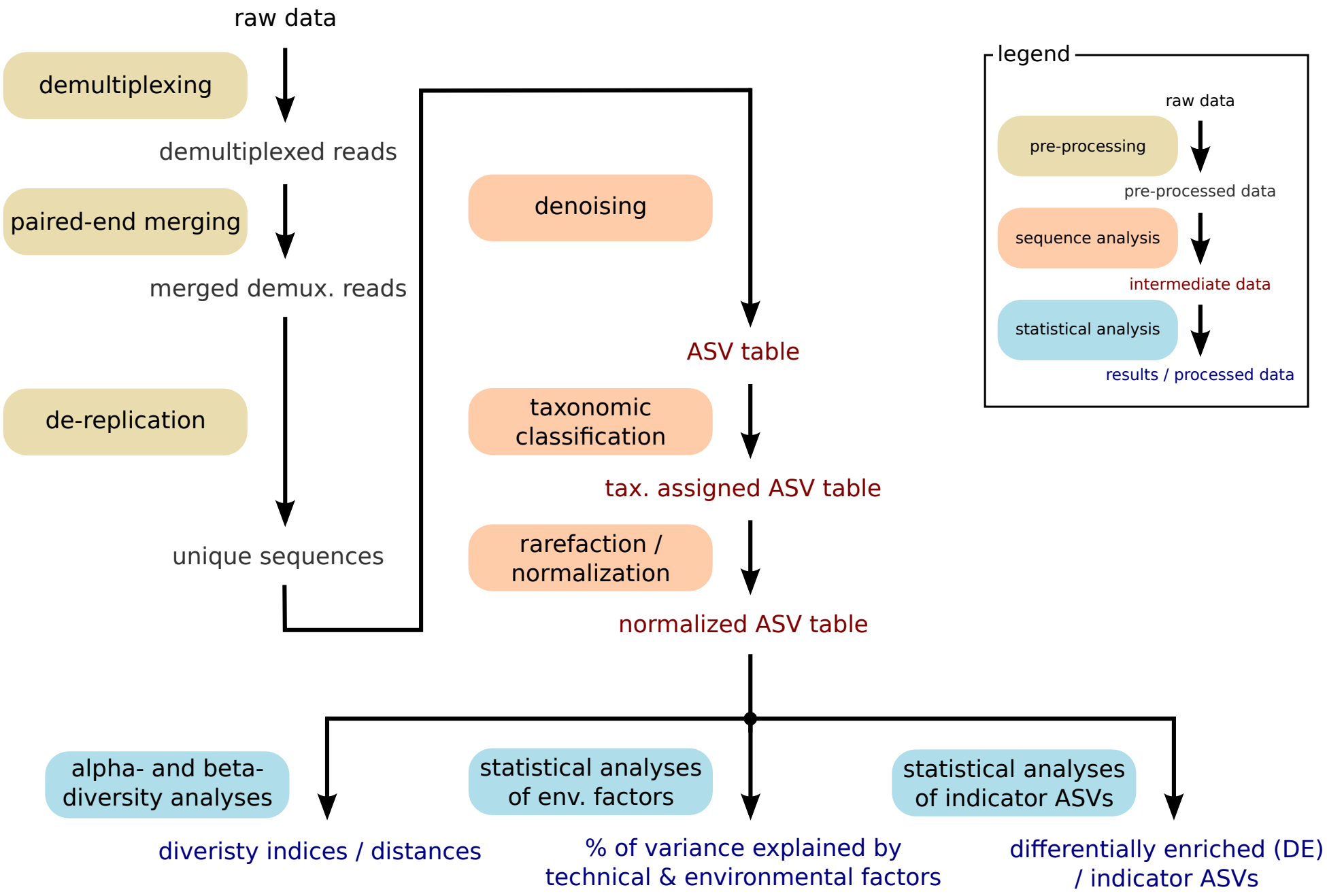
General problems associated with the use of OTUs:

- Low of resolution (typically 97% sequence identity)
- Lack of intrinsic biological meaning
- Inflation of alpha-diversity (artefactual OTUs)
- Low replicability and reproducibility
- Difficult to cross-reference between datasets / studies

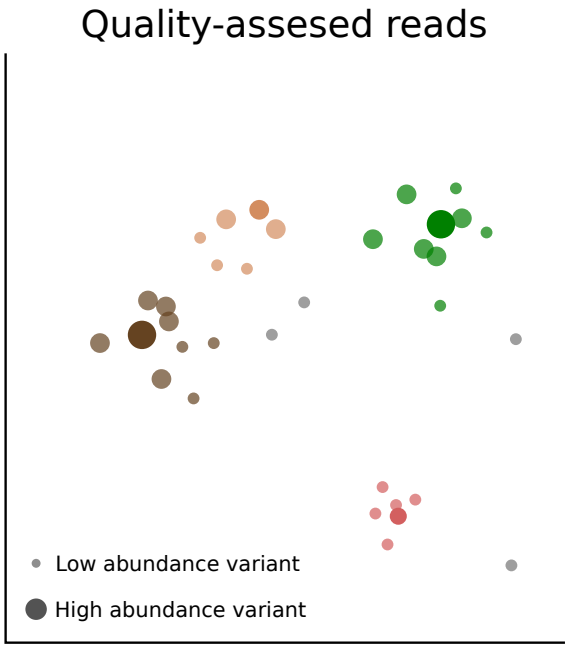
OTU clustering and synthetic community data

- Exhaustive reference database available
- High-resolution is required (large intra-OTU variation)
- Error correction using references (>denoising)

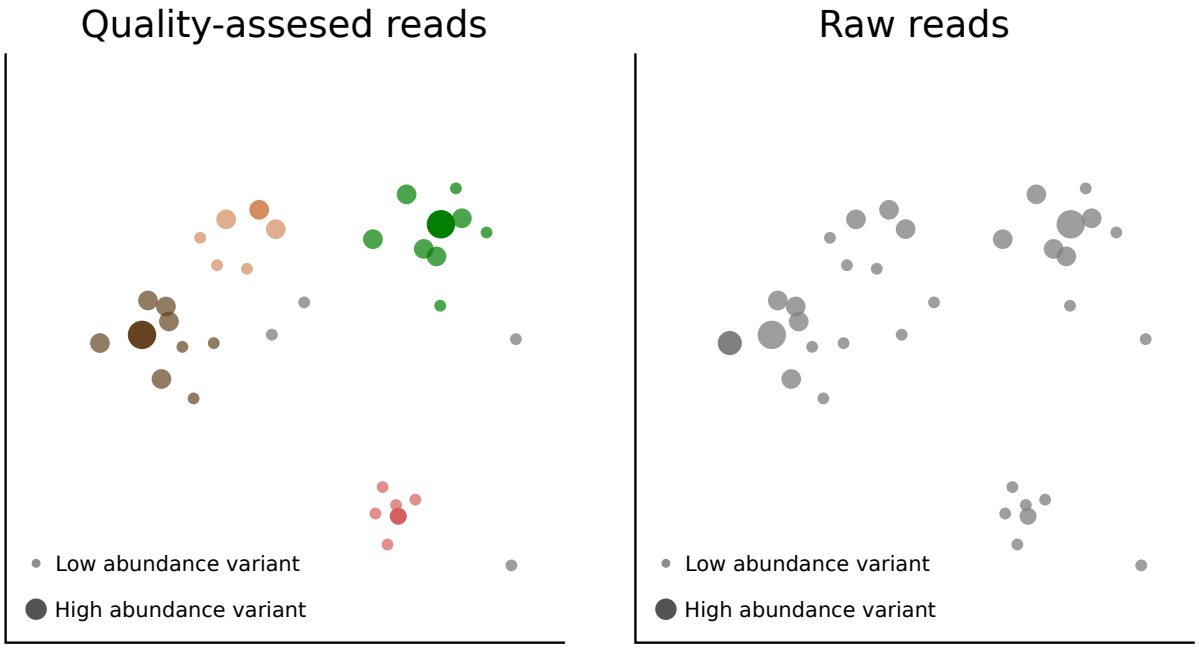
Workflow amplicon data analysis (natural communities)



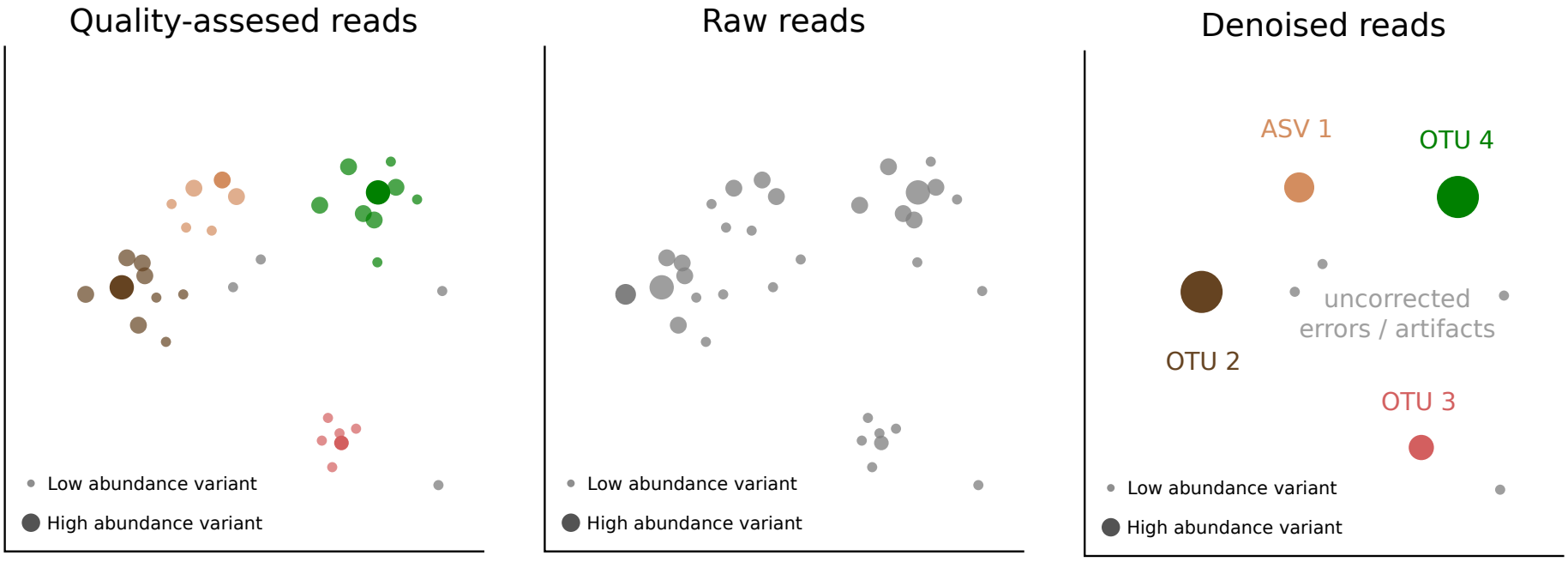
Denoising of amplicon data



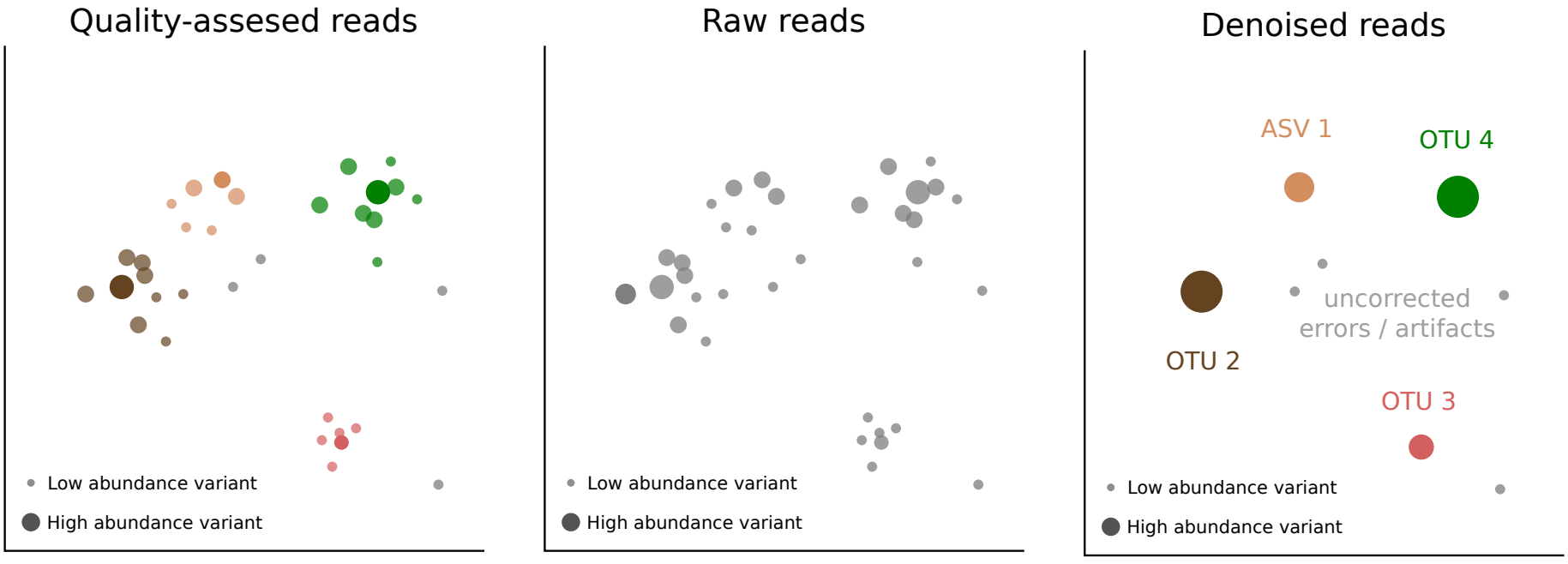
Denoising of amplicon data



Denoising of amplicon data

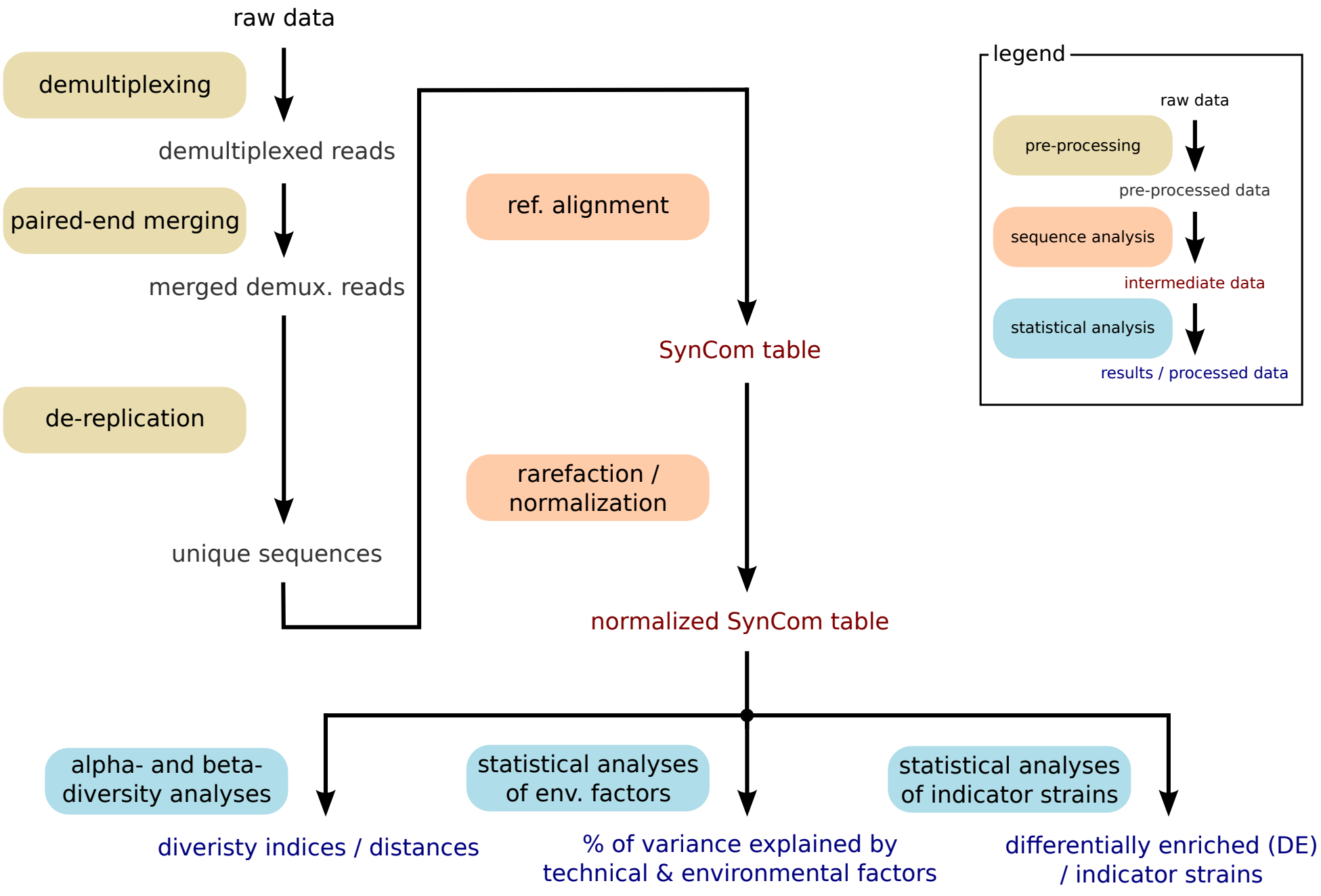


Denoising of amplicon data

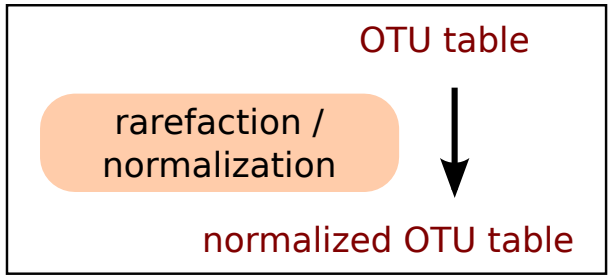


- Better than OTU clustering at removing artifacts
- Higher resolution
- Less dependent on arbitrary parameter choices (e.g. sequence identity)
- Computationally demanding / not fully parallelizable

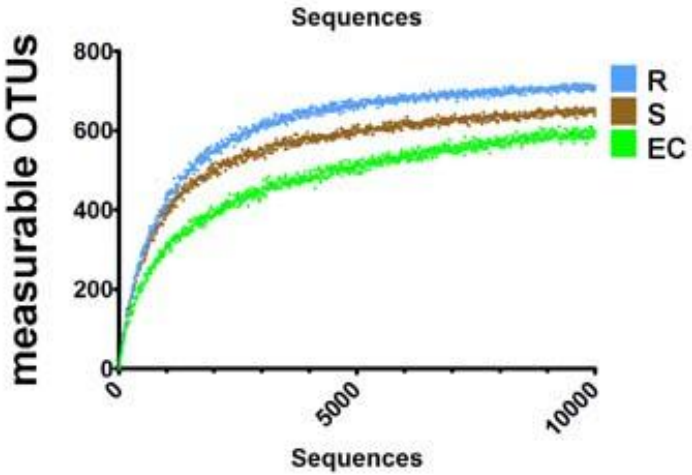
Workflow amplicon data analysis (synthetic communities)



Rarefaction / normalization



- Sample depth in amplicon sequencing data (number of reads per sample) is highly variable.
- When studying complex communities deep samples will capture more diversity
- There are two main strategies to address this issue:
  - sub-sampling or rarefaction, which consists on randomly selecting an equal number of sequences from each sample
  - calculate relative abundances (e.g. by dividing every OTU count by the total sample depth) with or without subsequent transformations
- Both approaches have pitfalls and there is considerable debate as to which strategy is preferable (see e.g. McMurdie *et al.*, 2014 and Weiss *et al.*, 2017)



Original Abundance	A	B
OTU1	62	500
OTU2	38	500
Total	100	1000

Rarefied Abundance	A	B
OTU1	62	50
OTU2	38	50
	100	100

Standard Tests for Difference

P-value	chi-2	Prop	Fisher
Original	0.0290	0.0290	0.0272
Rarefied	0.1171	0.1171	0.1169

Thank you!