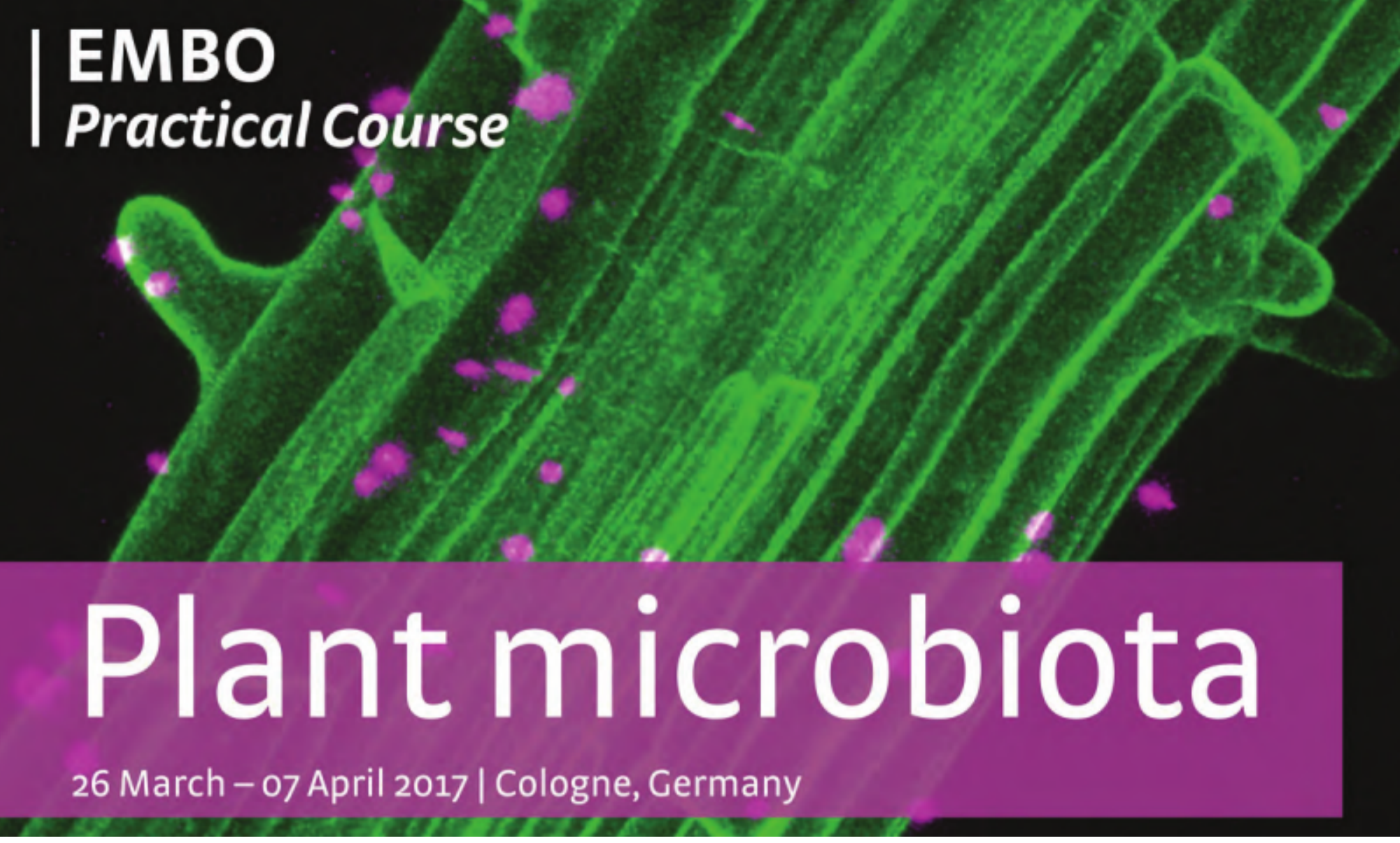
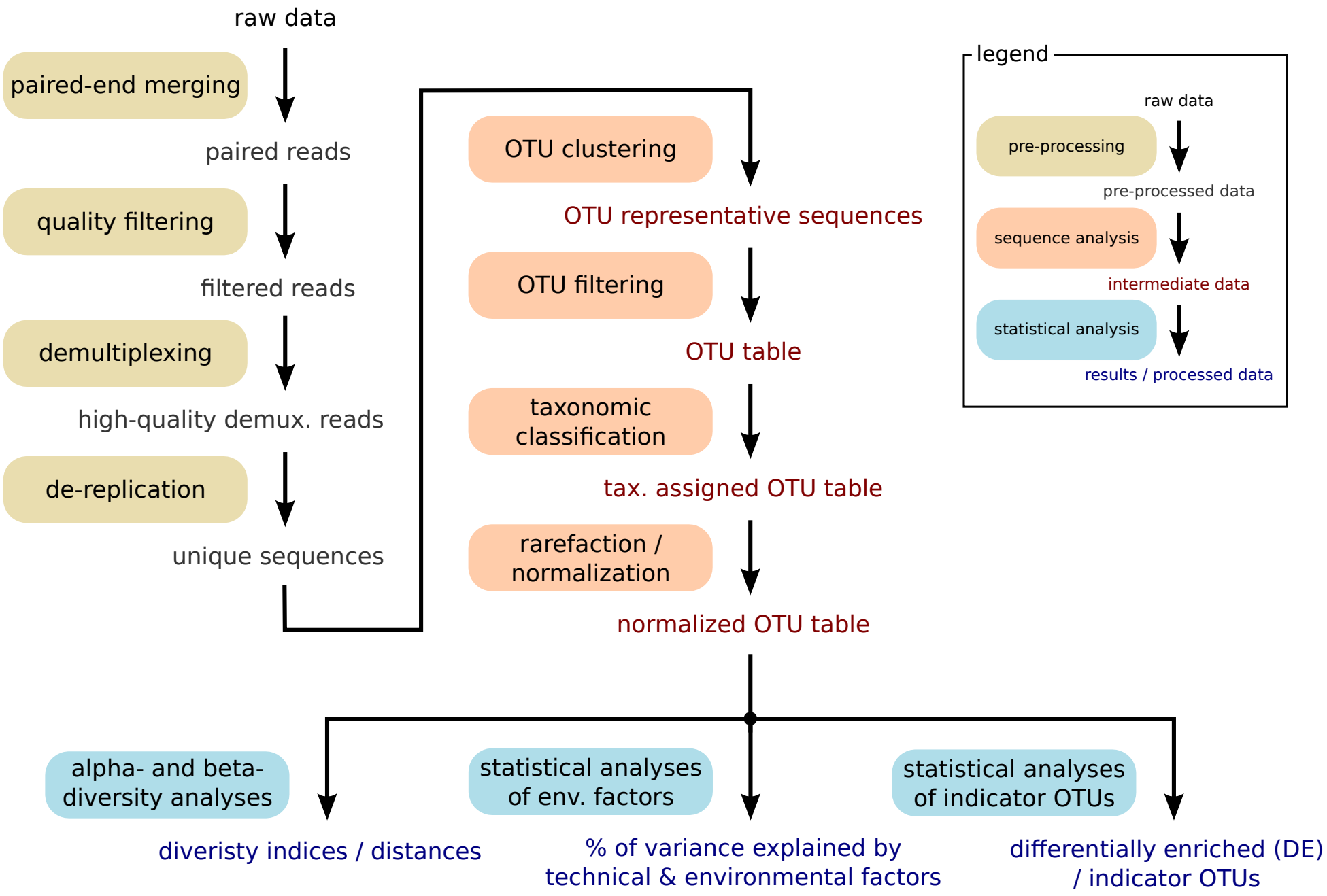


State-of-the-art approaches for amplicon data analysis

Ruben Garrido-Oter

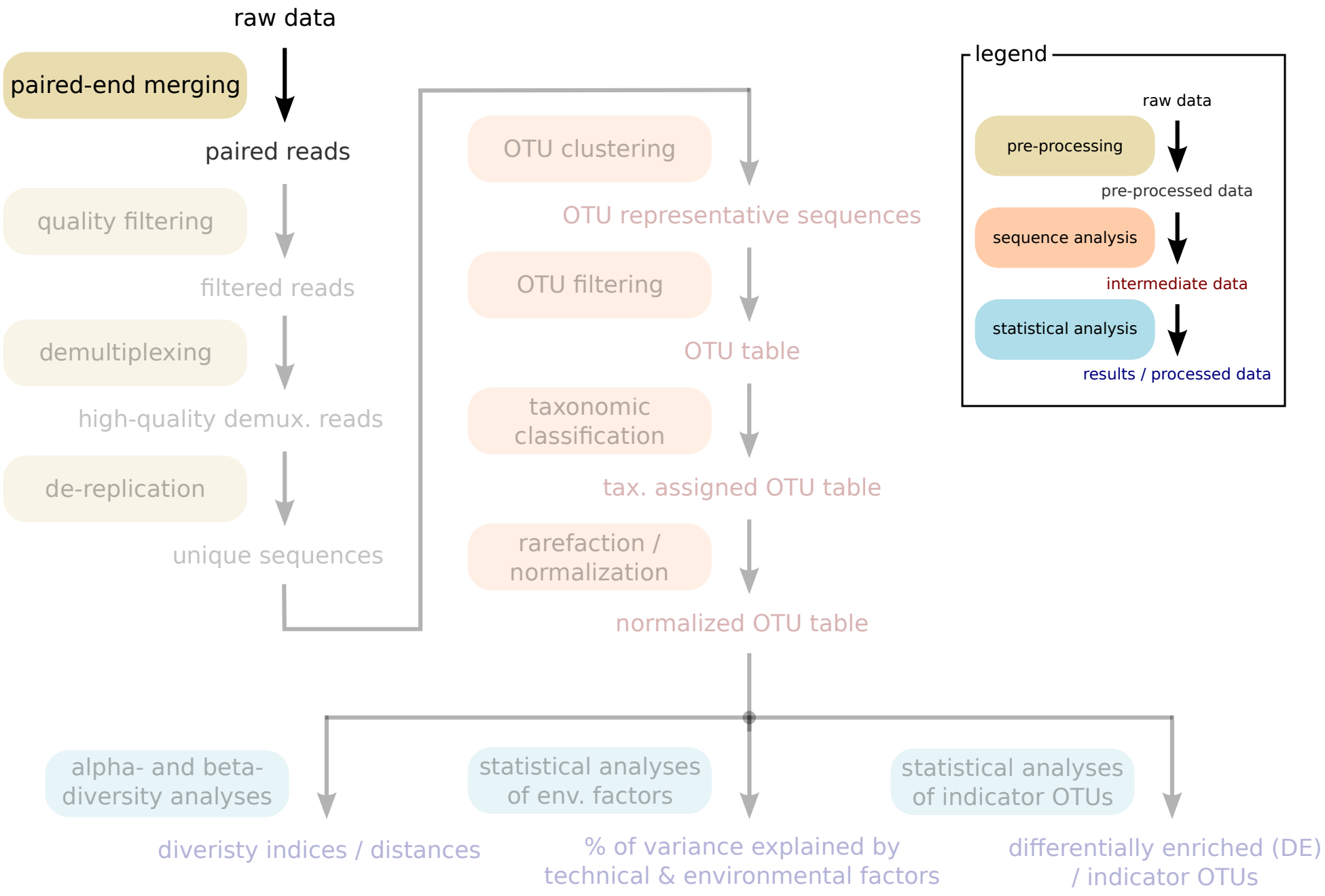
Workflow amplicon data analysis (natural community)



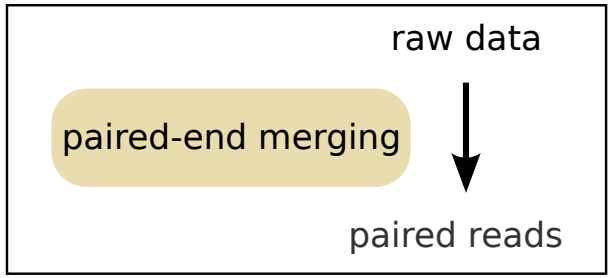
Commonly used toolkits for amplicon sequence analysis

qiime

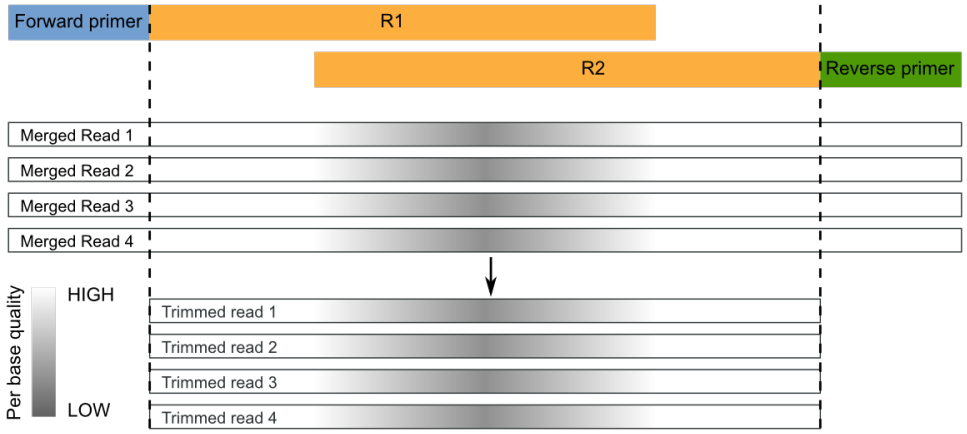
Workflow amplicon data analysis (natural community)



Merging paired-end reads (Illumina)

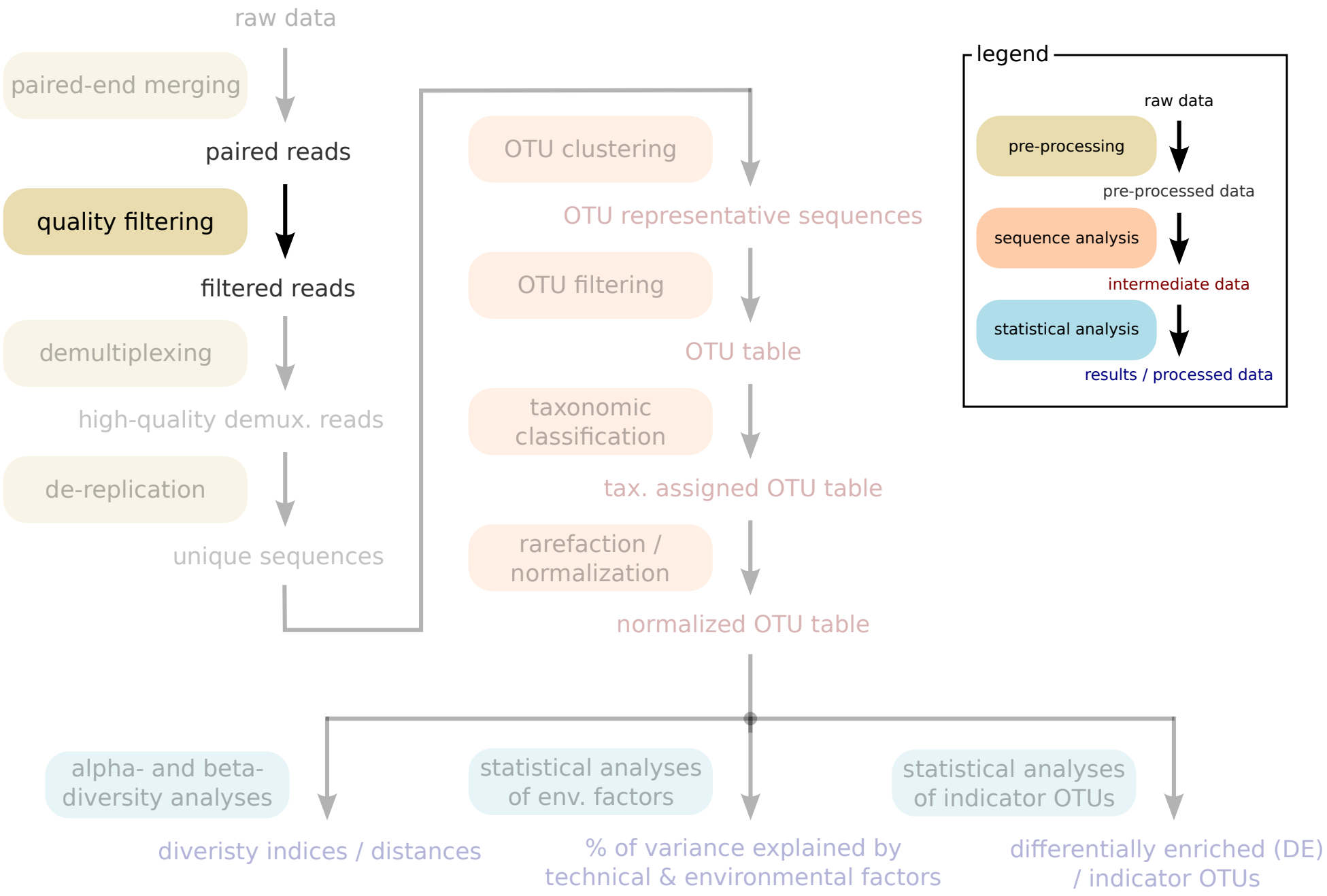


- Consists on merging (assembling) paired-ed reads into consensus sequences (and consensus quality scores for downstream filtering)

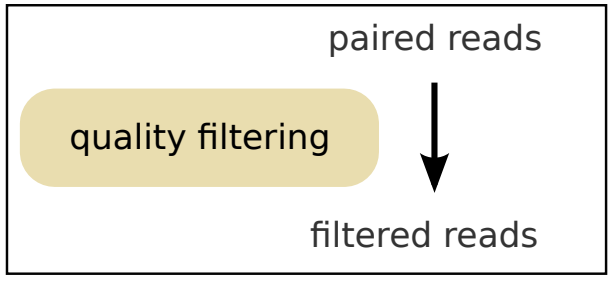


- Poorly overlapping pairs are (generally) discarded for fixed-length markers
- For overlapping pairs quality scores need to be recomputed
- Multiple tools, e.g.:
  - PANDAseq (Masella, *et al.*, 2012)
  - ***join_paired_ends*** in QIIME (uses the fastq-join tool)
  - ***fastq_mergepairs*** command in USEARCH

Workflow amplicon data analysis (natural community)



Quality filtering of amplicon reads



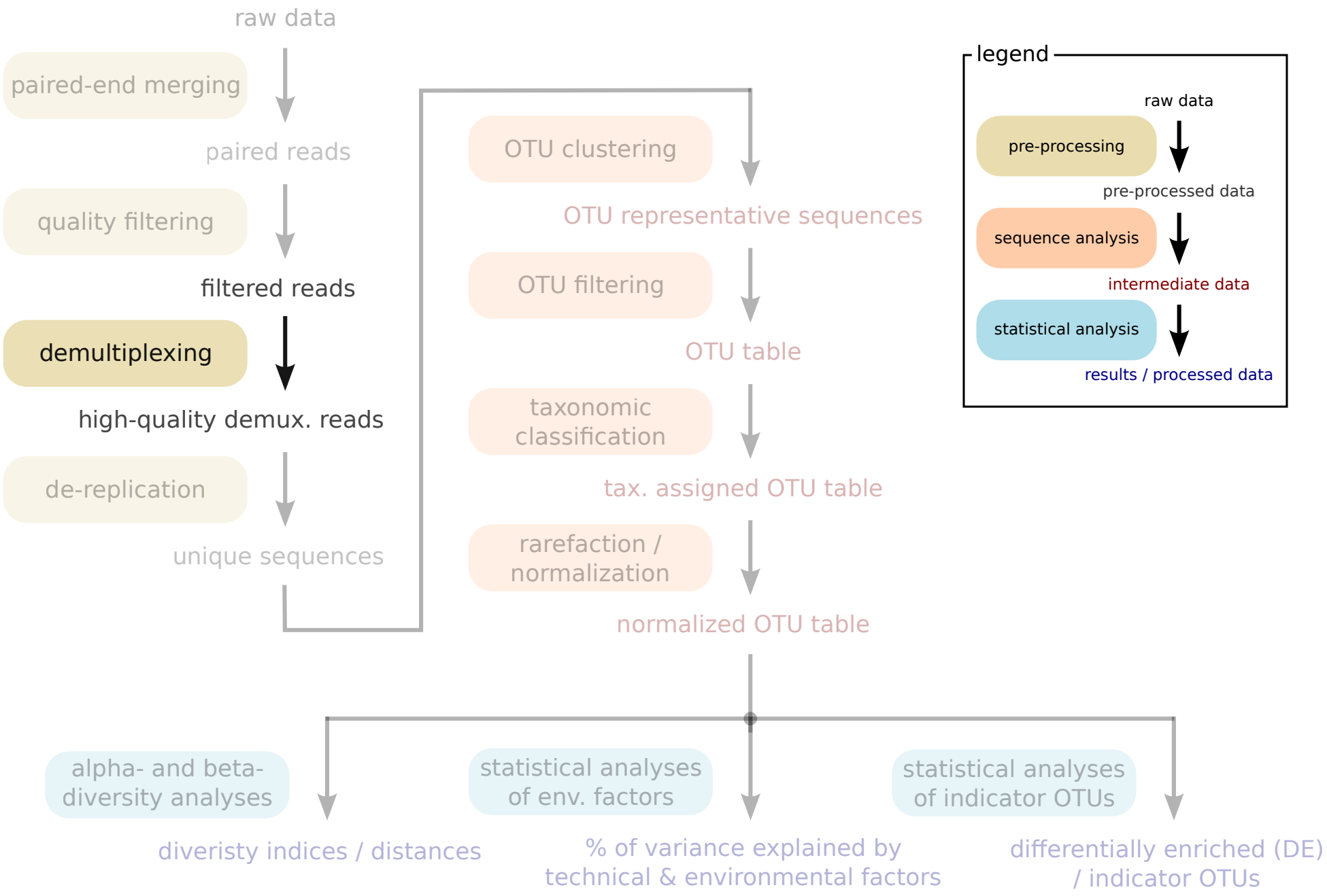
- NGS instruments (454, Illumina) indicate the probabilities of sequencing errors using quality (Phred or Q) scores

ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

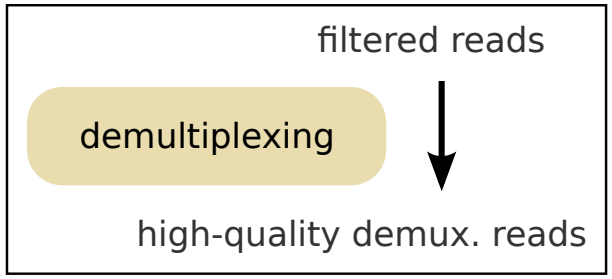
Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :	36	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
- Two ways (mainly) to filter sequencing errors:
  - maximum unacceptable Phred quality score in a read
    e.g. ***split_libraries_fastq*** in QIIME
  - expected number of errors in a read (dataset)
    e.g. ***fastq_filter*** (and ***fastx_learn***) command in USEARCH

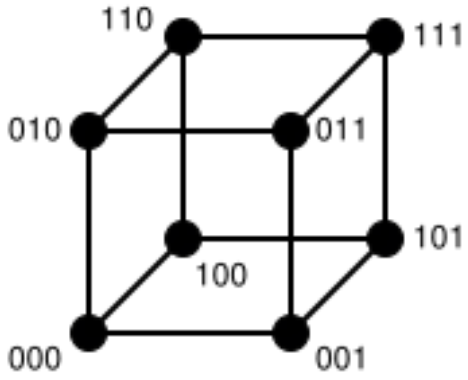
Workflow amplicon data analysis (natural community)



Demultiplexing

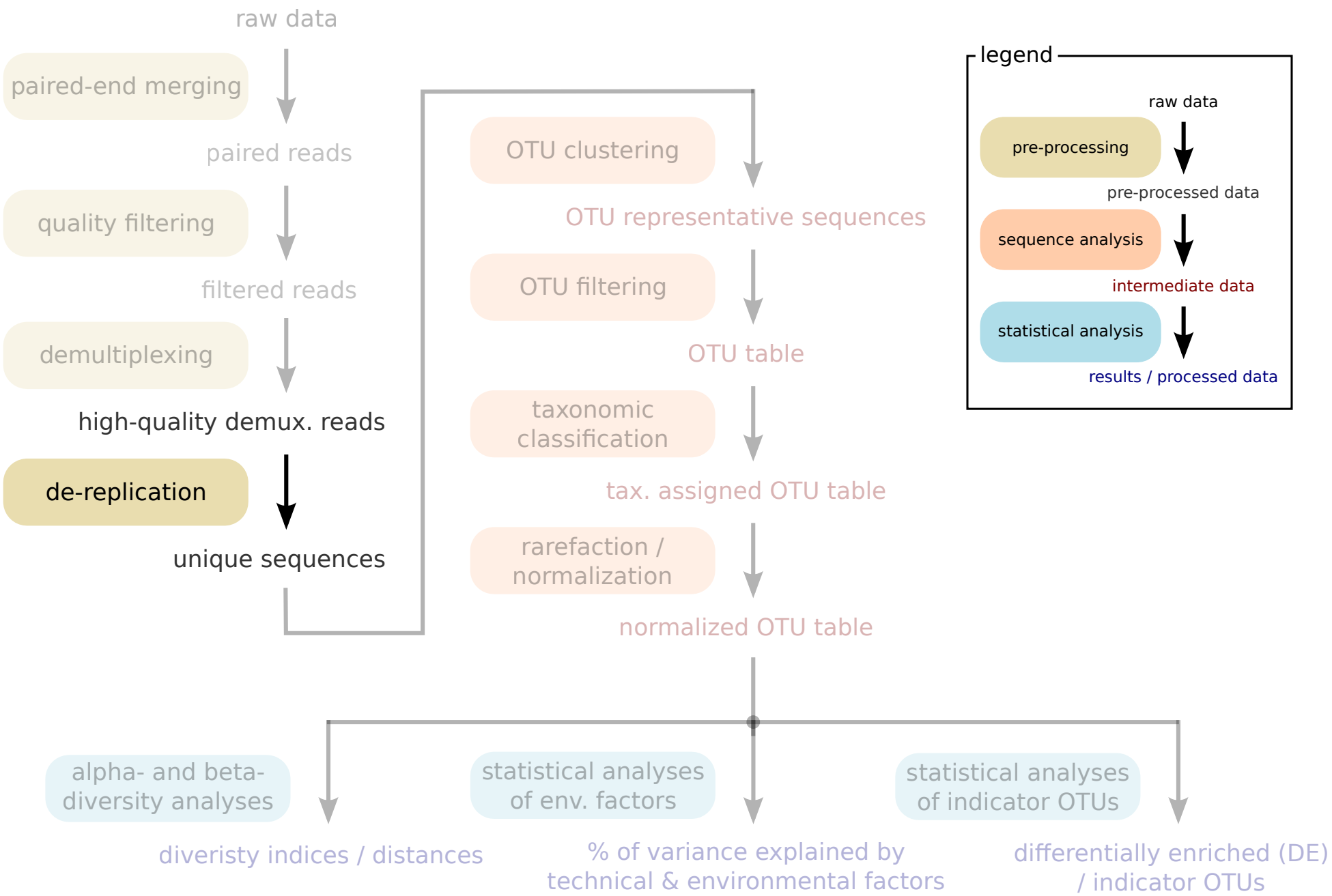


- Assign merged reads to samples using (and removing) error-correcting barcode sequences.

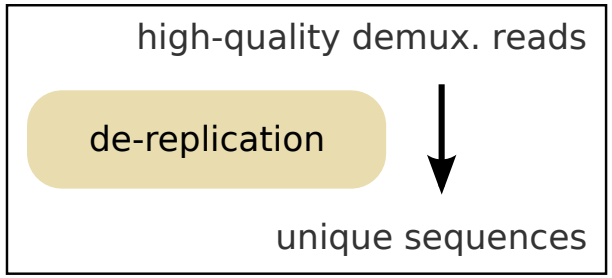


- Barcodes can be added as a tag to the sequence header or as a separate FASTQ file (depending on the instrument)
- Barcodes using Golay codes allow errors (2 for 12 base pair sequences)
- Tools append a sample identifier to the sequence header (and generally use different encodings), e.g.:
  - ***split_libraries_fastq*** script in QIIME (*demultiplex_fasta* for 454)
  - ***derep_fulllength*** script in USEARCH (with *-relabel* option)

Workflow amplicon data analysis (natural community)

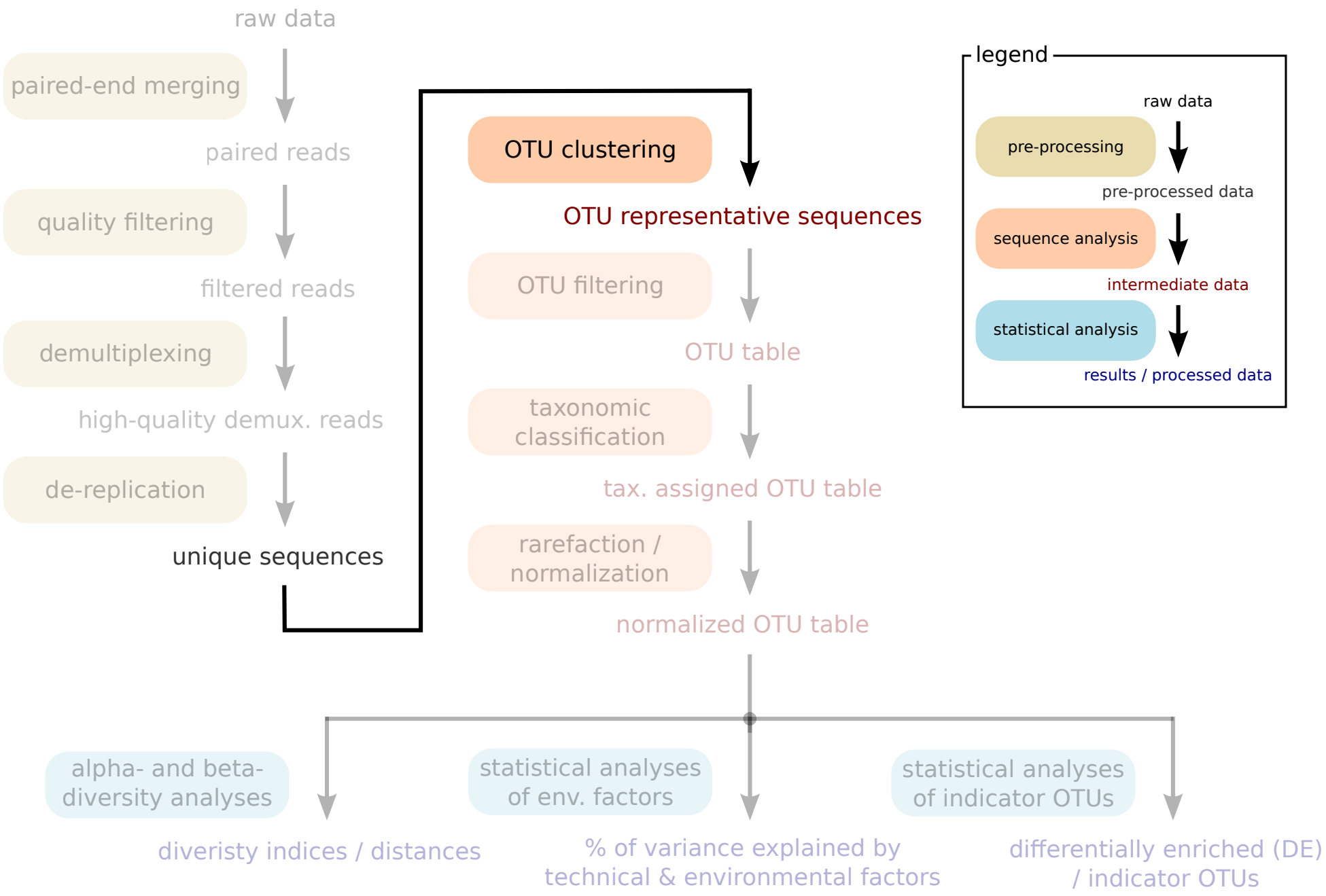


De-replication

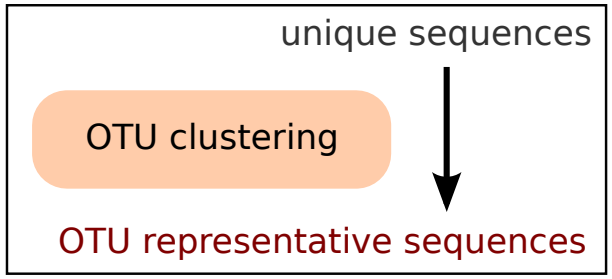


- Some steps (e.g. OTU clustering and taxonomic classification of representatives) only need to be done once per set of identical reads
- It is generally advisable to remove singletons. The basic assumption is that these are usually artifacts and that errors seldom occur multiple times on the same template sequence
- Mapping of non-replicated reads onto OTU representatives is necessary to obtain accurate abundances (***usearch_global*** script in USEARCH)
- Multiple alternatives implemented for full-length dereplication (paired-end Illumina with overlapping reads), e.g.:
  - ***split_libraries_fastq*** script in QIIME (*demultiplex_fasta* for 454)
  - ***derep_fulllength*** script in USEARCH (with *-relabel* option)
- Non-overlapping singleton reads (e.g. ITS) need to be trimmed to equal length *prior* de-replication:
  - ***fastx_truncate*** command in USEARCH)
  - ***truncate_fasta_qual_files*** script in QIIME (outdated)

Workflow amplicon data analysis (natural community)



OTU (Operational Taxonomic Unit) clustering



- OTU clustering ('picking') refers to the process of grouping related sequences into constructs that roughly correspond with bacterial species

- Arbitrary thresholds of sequence similarly are generally used (97%) that are highly dependent on the marker gene (and region within)

- OTUs are useful data constructs in microbial ecology but suffer from various shortcomings, e.g.:

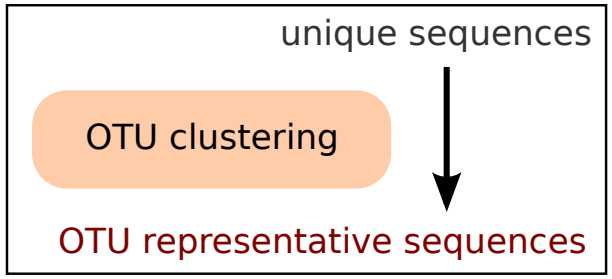
    - OTUs do not correspond to any meaningul biological entity: isolates from the same OTU often have very different genomes

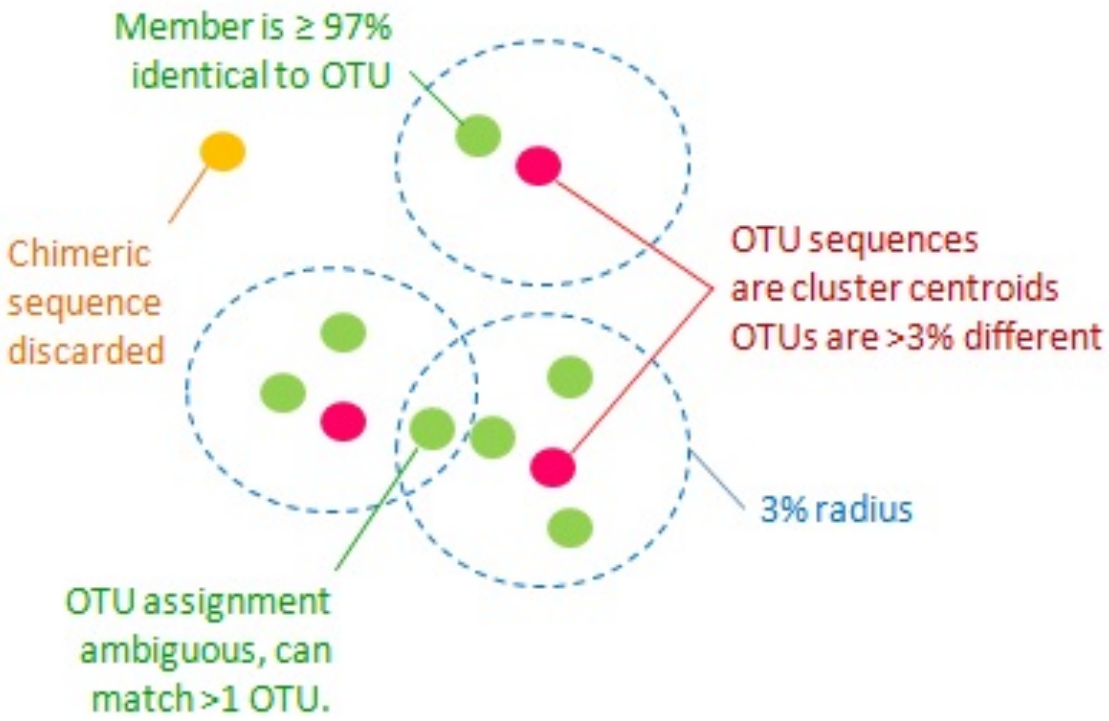
    - there is a taxonomy bias in sequence variation (some OTUs are more homogeneous than others / have lower conservation in the amplified genetic locus)

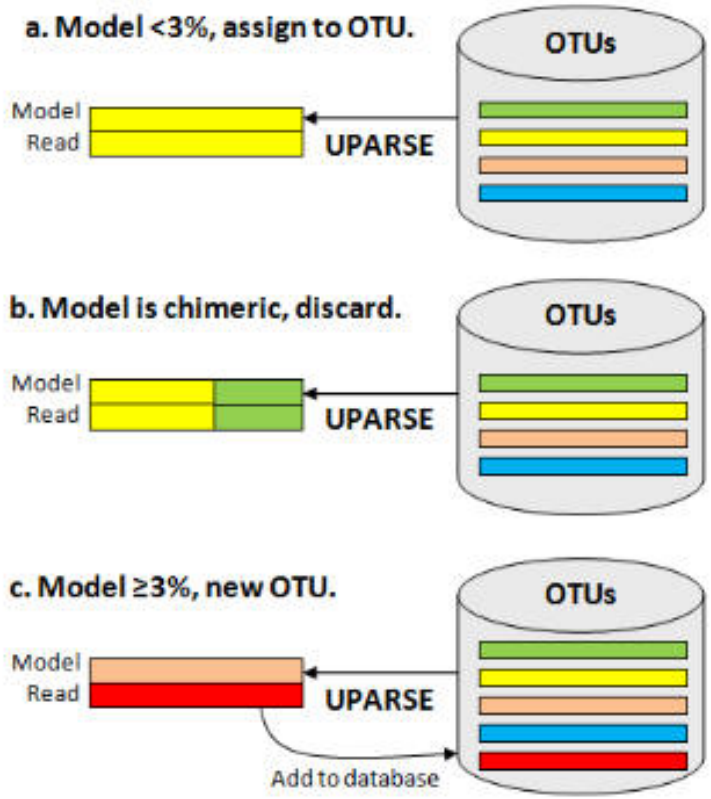
    - sequencing and PCR errors impose a resolution limit (tradeoff between resolution and artefactual OTUs)

OTU (Operational Taxonomic Unit) clustering

- OTU clustering ('picking') refers to the process of grouping related sequences into constructs that roughly correspond with bacterial species

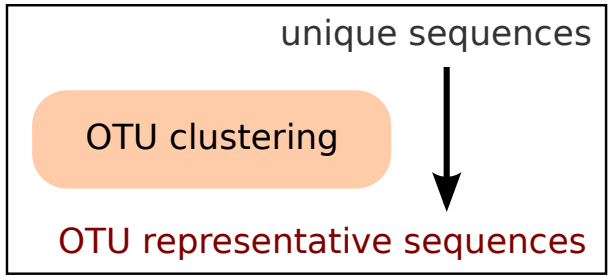






Edgar, 2013

OTU (Operational Taxonomic Unit) clustering



- There are three main strategies for OTU picking:
  - *de novo* OTU clustering

    (do not depend on a reference database)

    e.g. UCLUST (Edgar, 2010), UPARSE (Edgar, 2013) or Swarm (Mahé *et al.*, 2014)

  - closed-reference OTU clustering

    (take advantage of reference database; SynComs)

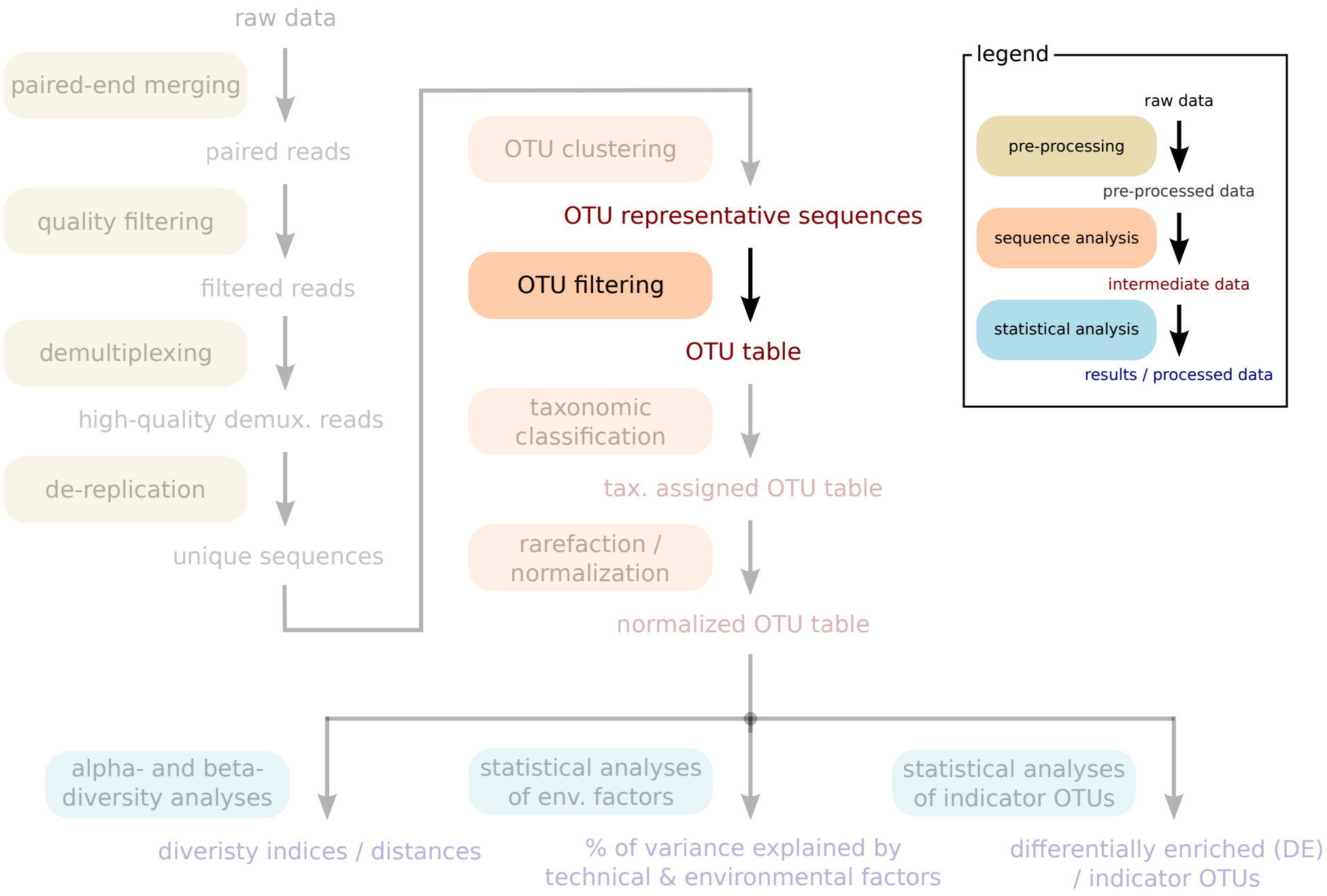
  - hybrid (open-reference) OTU clustering

    (typically perform a first-pass reference iteration, followed by *de novo* clustering of left-out sequences)

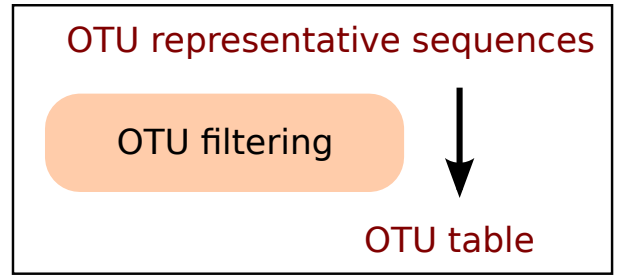
    e.g. SortMeRNA (Kopylova *et al.*, 2012)

- Multiple implementations of these methods in various toolkits (e.g. ***cluster_otus*** command in USEARCH or ***pick_otus*** in QIIME)

Workflow amplicon data analysis (natural community)



OTU filtering



- Filtering of low abundance OTUs assuming these are likely artefacts (depending on the complexity / sequencing depth 1% 0.1% R.A.)

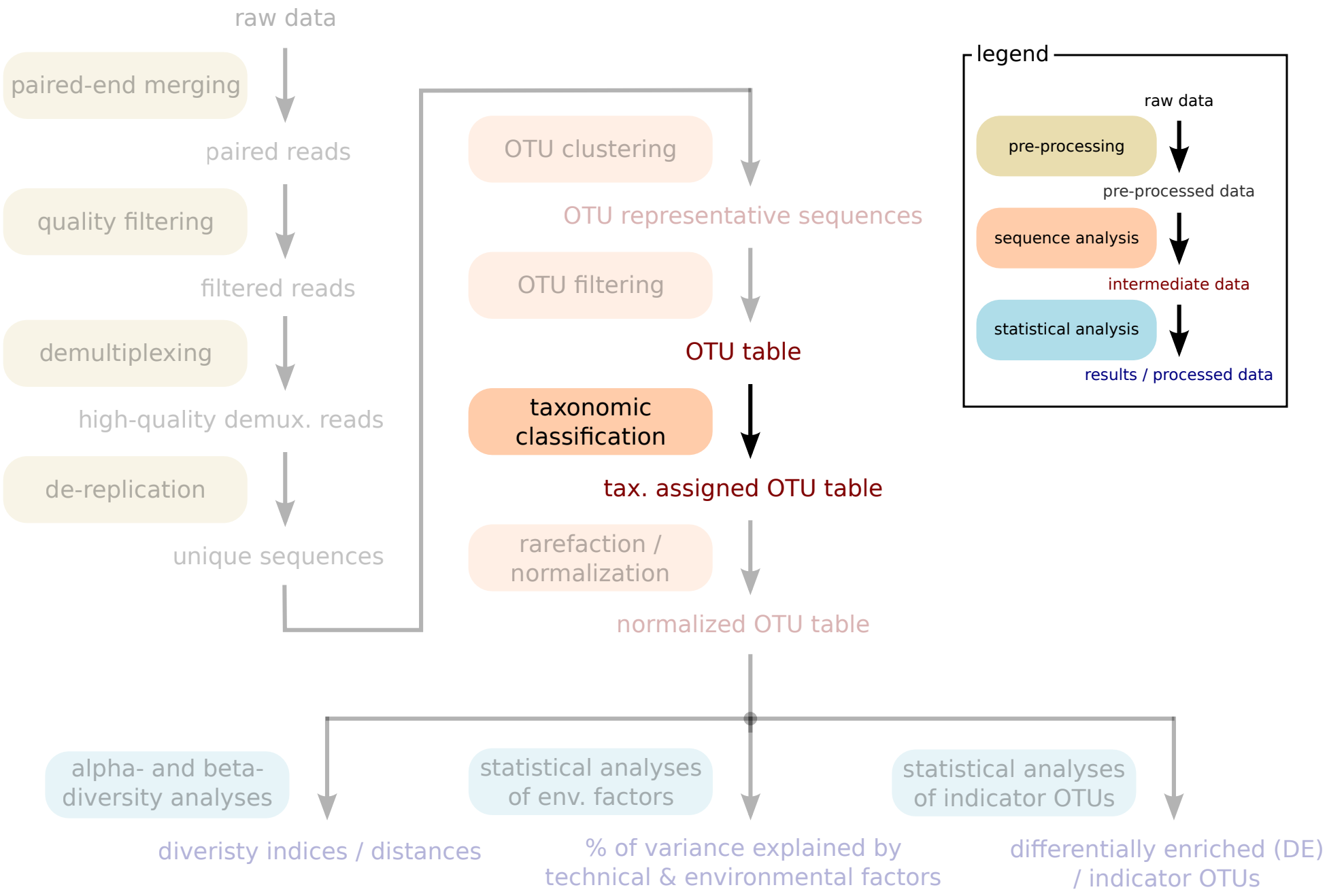
- Chimera detection and removal (artefactual sequences formed from two or more biological sequences joined together e.g. during PCR)

  multiple methods, e.g.: ChimeraSlayer (Haas *et al.*, 2011), DECIPHER (Wright *et al.*, 2012), Perseus (Quince *et al.*, 2011) or UCHIME (Edgar et al., 2011) (***chimera_ref*** command in USEARCH or ***identify_chimeric_seqs*** in QIIME)

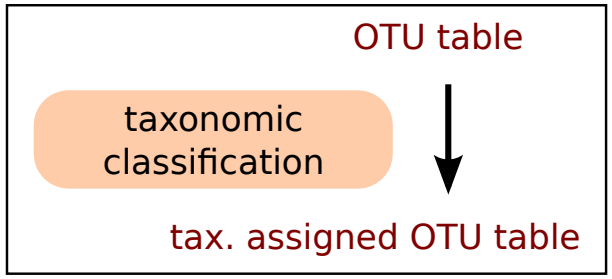
- Filtering nonsense OTUs (generated from prevalent sequencing / PCR artefacts that do not align to public databases of bacteria above a certain threshold; e.g. 75% sequence identity)

  e.g. using the ***usearch_global*** alignment command in USEARCH

Workflow amplicon data analysis (natural community)



Taxonomic classification



- There are multiple algorithms and tools for classification of marker gene sequences, e.g. RDP, Blast, RTAX, UTAX, mothur-*knn*, UCLUST, SortMeRNA, etc.

  QIIME provies a wrapper for many with the ***assign_taxonomy*** script

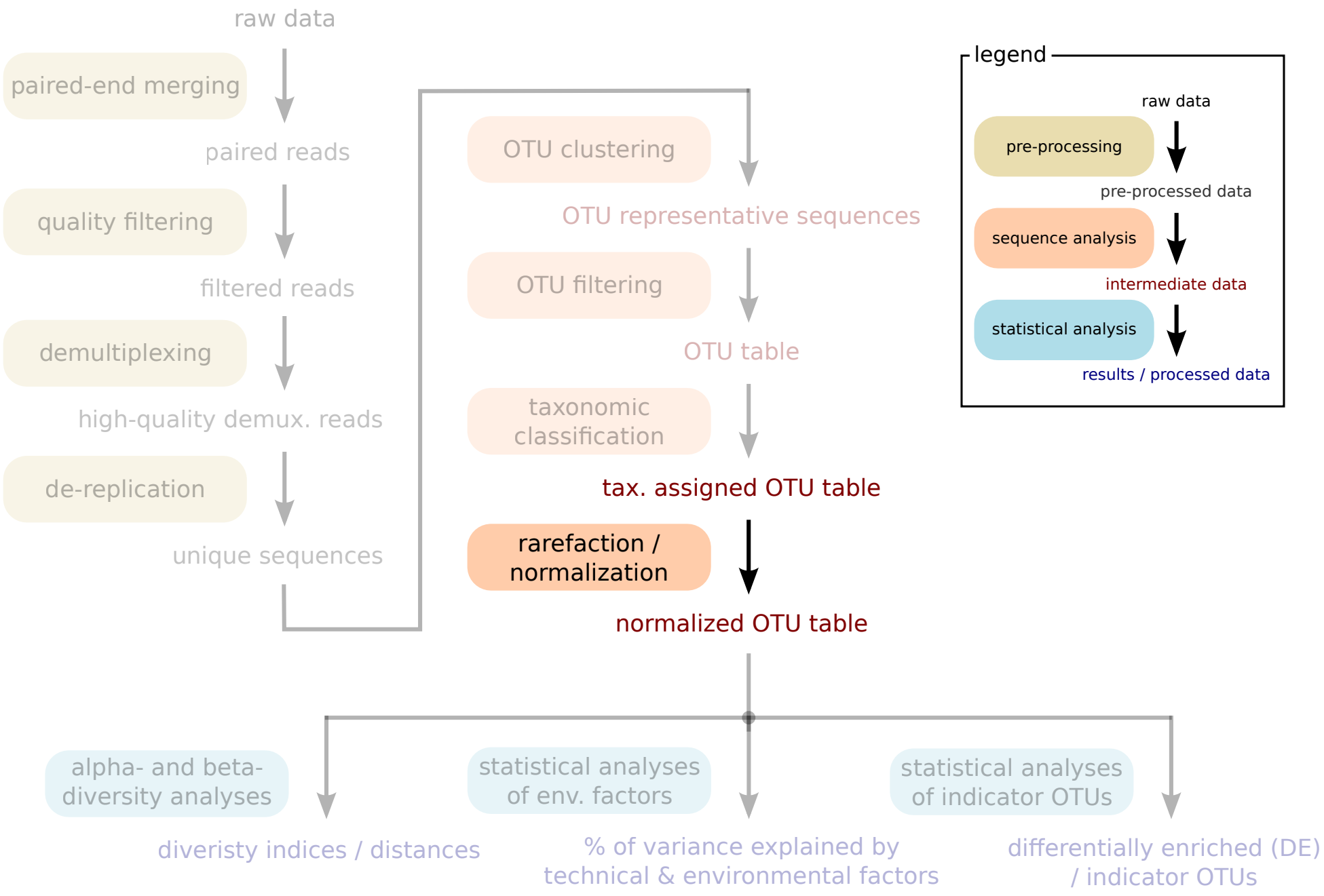
  USEARCH implements UTAX in the commands ***utax*** and ***cluster_otus_utax***

- These methods rely on homology -- results are highly dependent on high similarity with representatives in the database

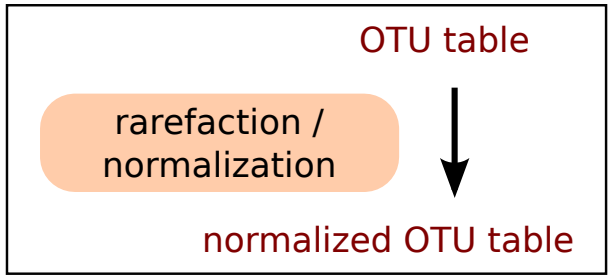
- Reference databases have poor coverage and are highly biased (e.g. towards culturable bacteria)

- Various types of errors and tradeoffs: type I & II errors and overclassification errors (i.e. tax. assigment for novel sequences)

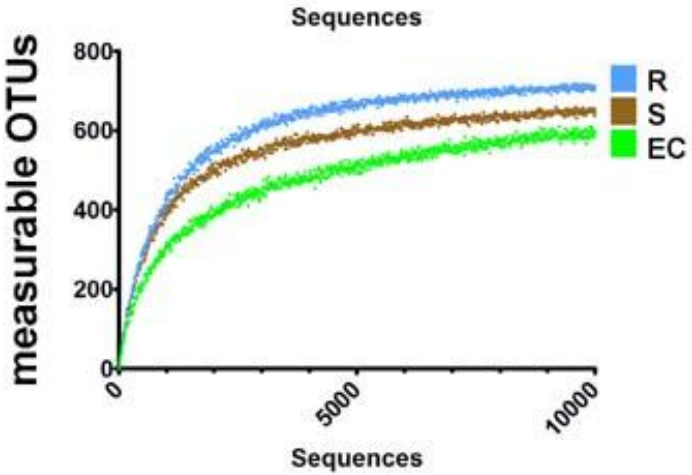
Workflow amplicon data analysis (natural community)



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)



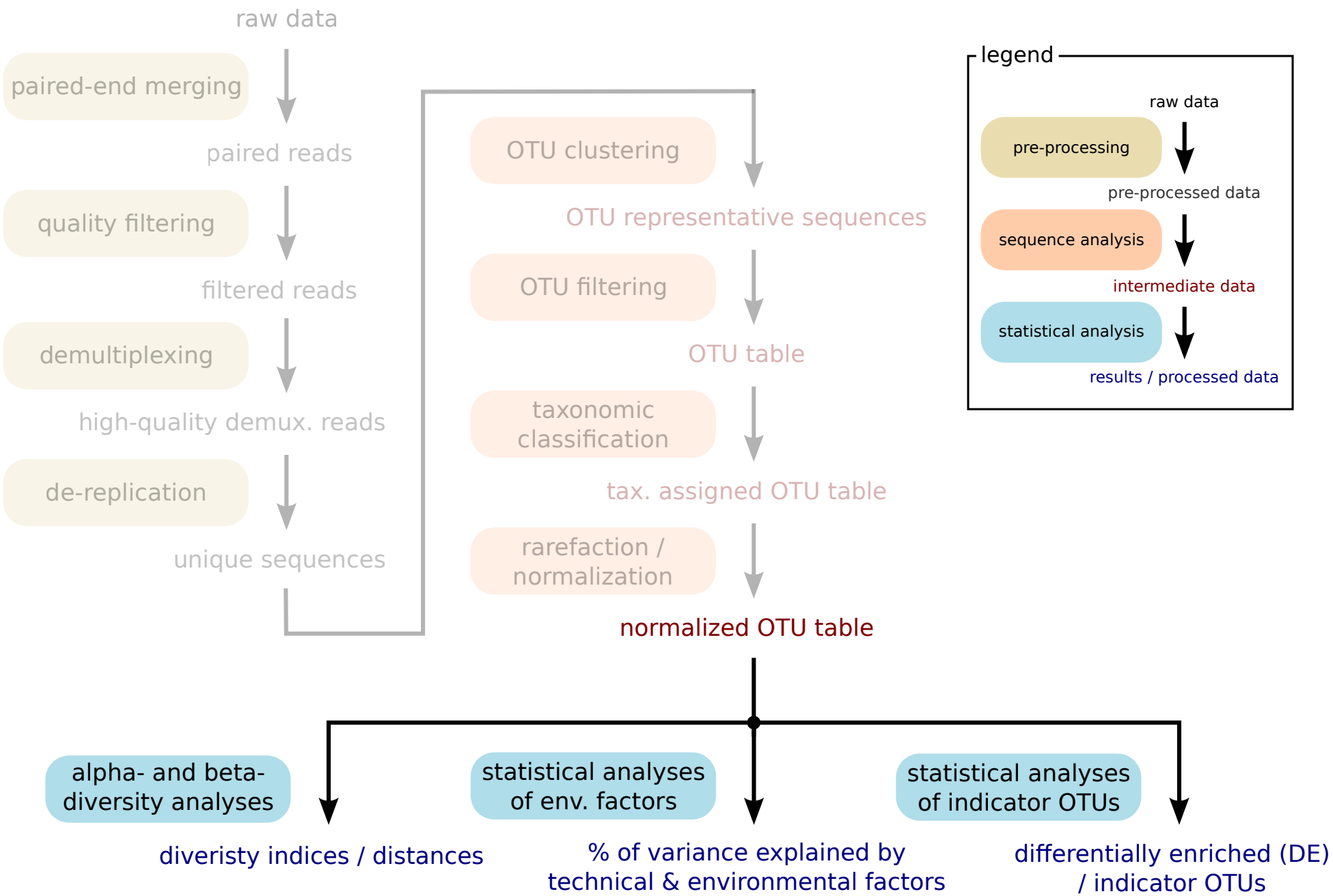
Original Abundance	A	B
OTU1	62	500
OTU2	38	500
Total	100	1000

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

Workflow amplicon data analysis (natural community)



Thank you!