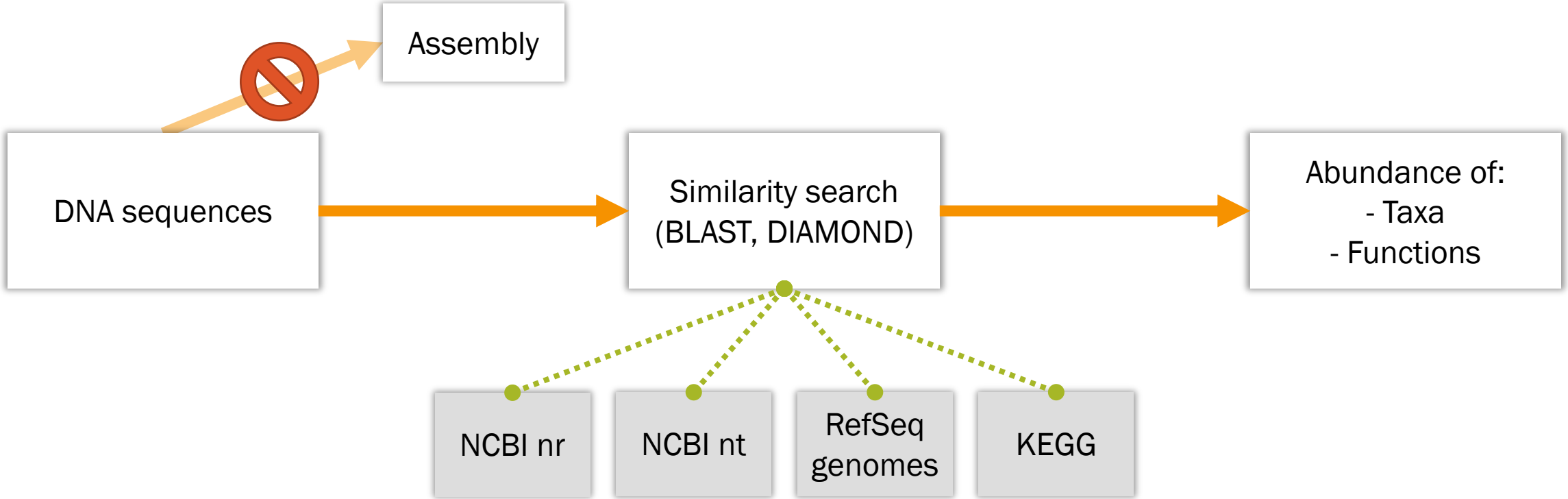
Environmental metagenomics

Read-based analyses

Igor S. Pessi & Antti Karkman, University of Helsinki

What is read-based profiling?



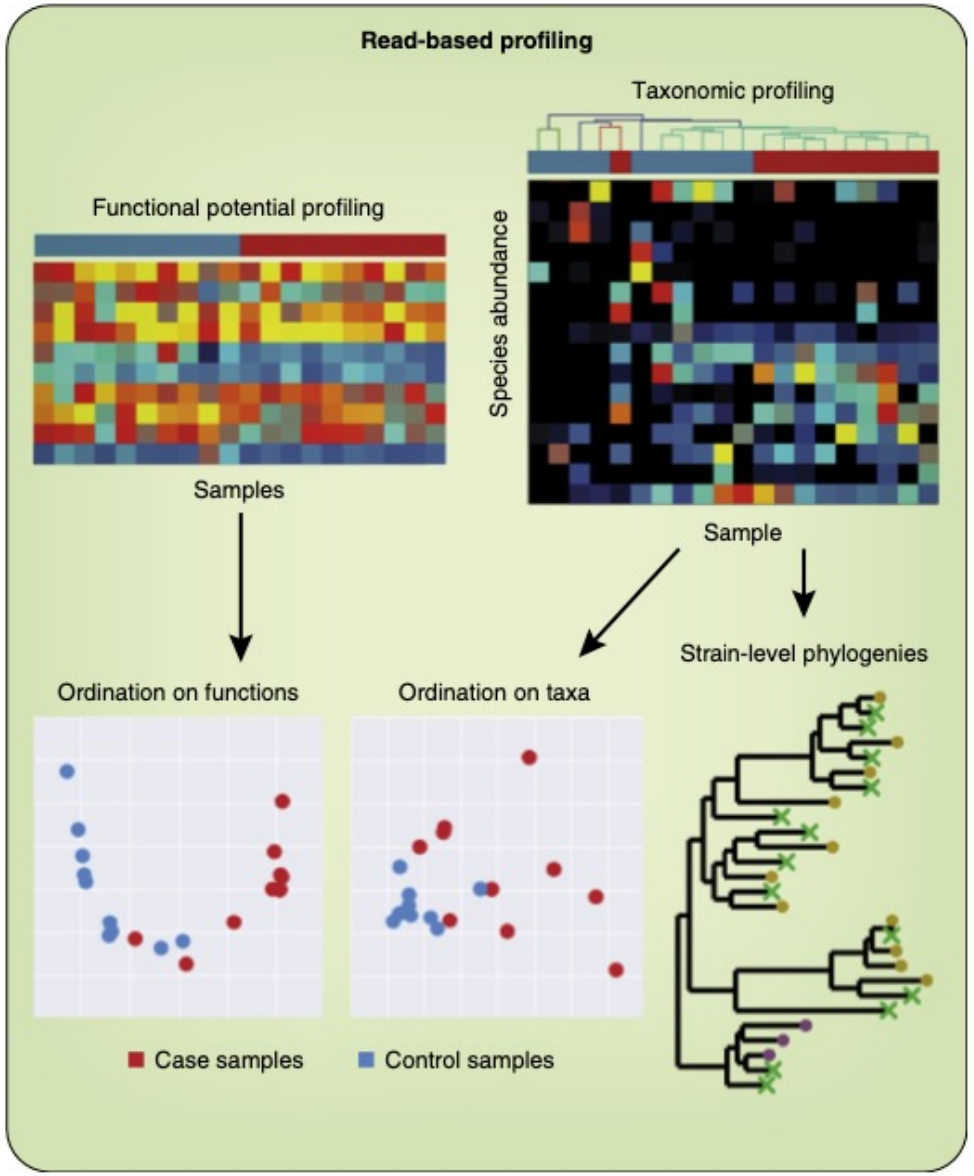
Read-based profiling is

Fast

Quantitative

Somewhat outdated

- Assembly-based are preferred
- Can give interesting preliminary insights
- Usually done as a ”quick-and-dirty” estimate prior to assembly

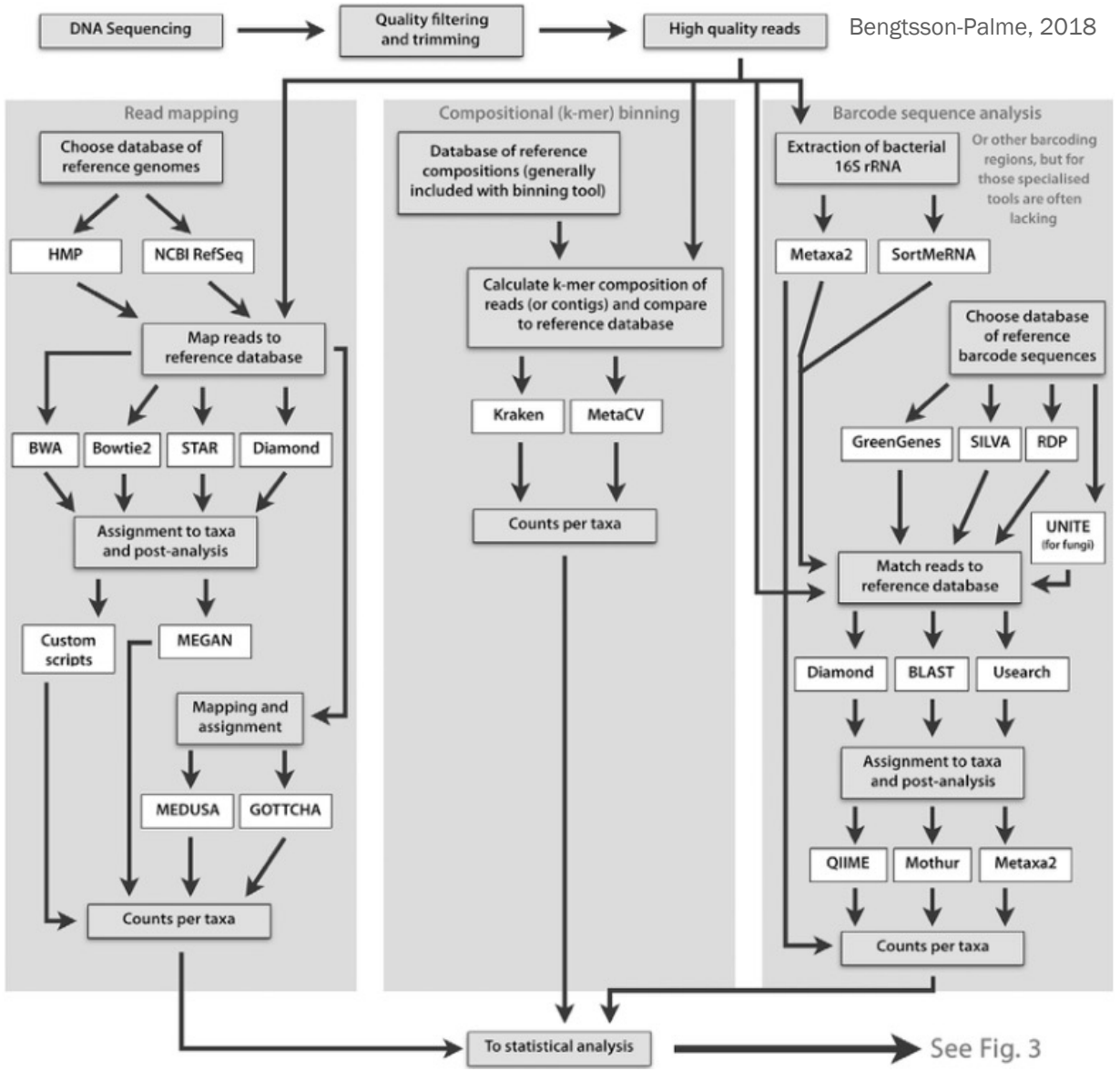


Quince et al. 2017

Approaches to taxonomic profiling

Read mapping and compositional binning

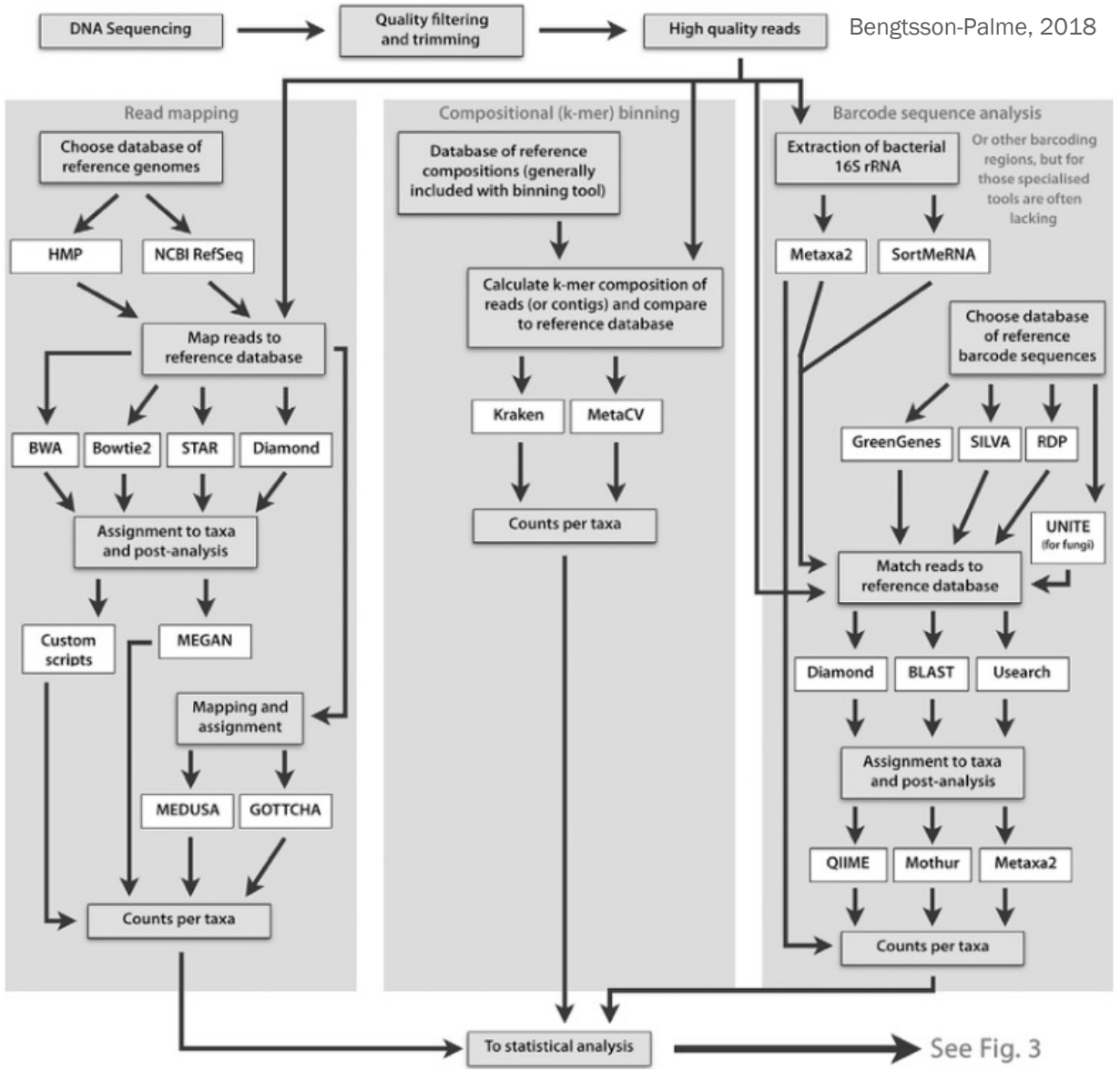
- Analysis of all reads
- Reference database of sequenced genomes
- Mapping: slow, requires lots of CPU and RAM
- Compositional binning: faster but less accurate



Approaches to taxonomic profiling

Barcode sequence analysis

- Analysis of specific barcode genes (e.g. 16S rRNA)
- Curated database of barcode sequences (e.g. SILVA)
- Much faster than the other approaches, but provides lower resolution



Approaches to taxonomic profiling: how to choose?

Analysis of all reads suffer from limited databases of reference genomes

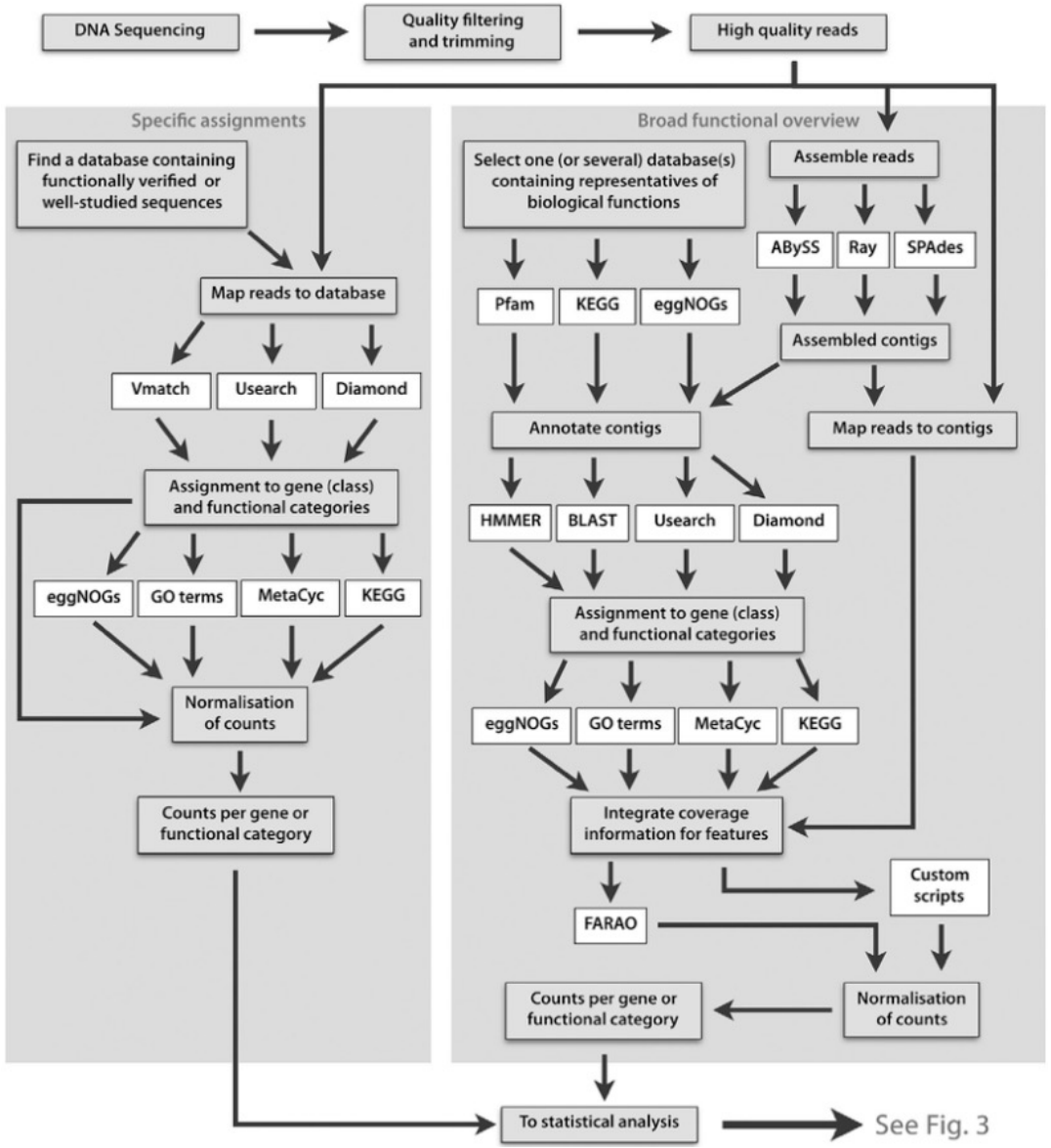
- More suitable for environments that are better described (e.g. human gut)

Analysis of barcode genes suffer from lower resolution

- More suitable for environments with a high fraction of unknown microorganisms (e.g. soil)

Approaches to functional profiling

Bengtsson-Palme, 2018



Broad *versus* specific profiling

- Broad DBs: entire functional universe (e.g. KEGG, PFAM)
- Specific DBs: focusing on one or few processes (e.g. CAZy, CARD)

Approaches to functional profiling: how to choose?

Broad databases give an overview of the functional potential of microbial communities

- Suitable for investigating major differences across environments

Specific databases are often highly curated and can give substrate-level information

- Suitable for investigating e.g. gene variants across environments

Making sense of read-based analyses

Comparative analyses

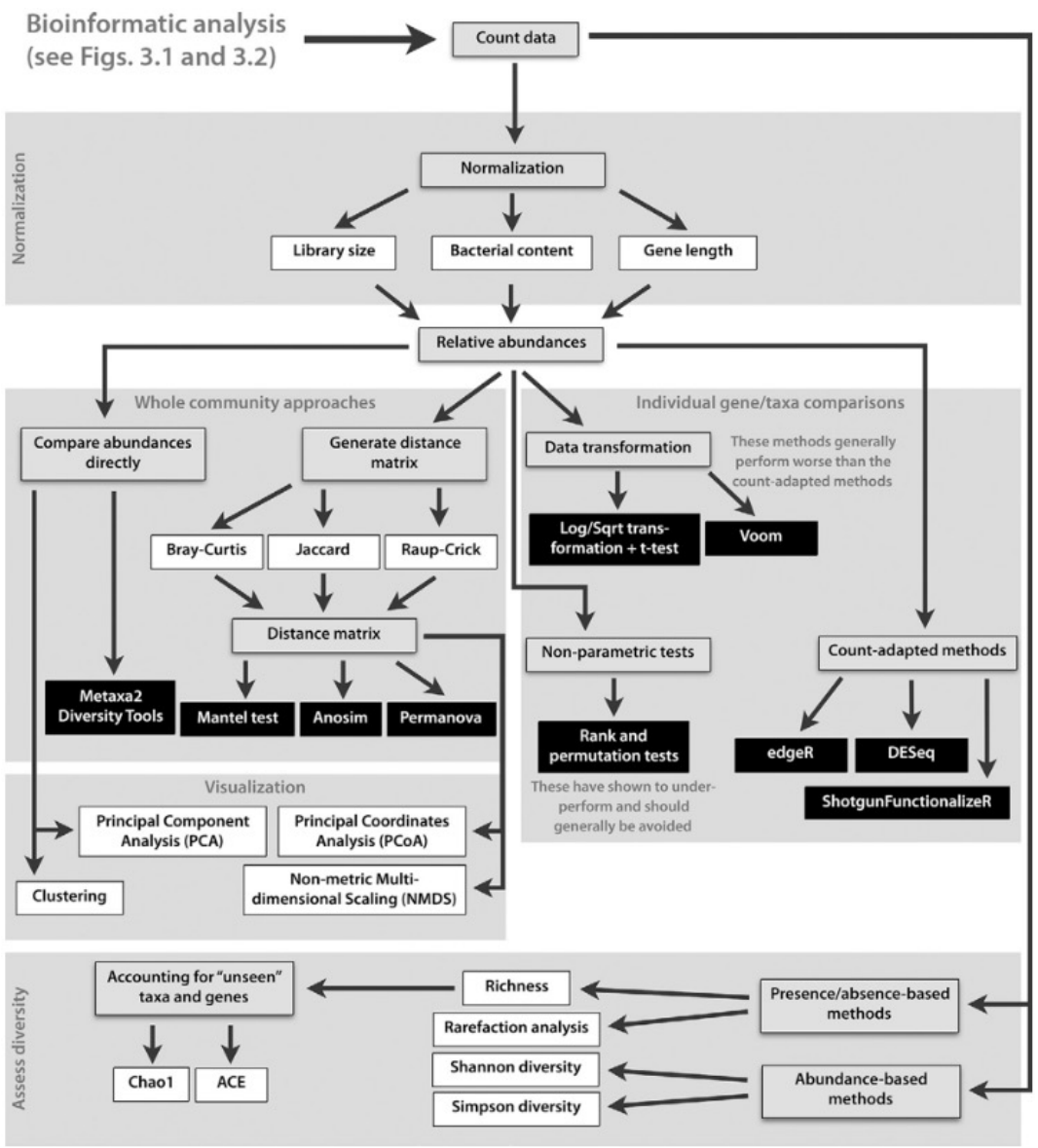
Statistics

- Univariate (e.g. ANOVA of specific genes and taxa)
- Multivariate (e.g. PERMANOVA, ordination/clustering, Mantel test)

Normalization!

- Library size
- Bacterial content (e.g *rpoB* gene)

Bengtsson-Palme, 2018



Pitfalls of read-based analyses

Curation level of the database

- Are sequences verified experimentally to perform the expected function?

Comprehensiveness of the database

- Both taxonomic- and functionally

Speed *versus* sensitivity tradeoff

- E.g. BLAST *versus* DIAMOND

Choice of identity, bitscore/e-value and coverage cutoffs

- No way to generalise for all genes, things have to be checked more or less manually, e.g. by looking at the literature for the gene

Remember: always sanity check!

Especially for unexpected findings:

- Redo with more strict thresholds
- Redo with a different tool (e.g. BLAST *versus* DIAMOND) and database
- Investigate other genes belonging to the same pathway

References and further reading

Quince C. et al. 2017. Shotgun metagenomics, from sampling to analysis. Link

Bengtsson-Palme J. 2018. Strategies for taxonomic and functional annotation of metagenomes. Link

Paliy O. & Shankar V. 2016. Application of multivariate statistical techniques in microbial ecology. Link

Jonsson V. et al. 2016. Statistical evaluation of methods for identification of differentially abundant genes in comparative metagenomics. Link