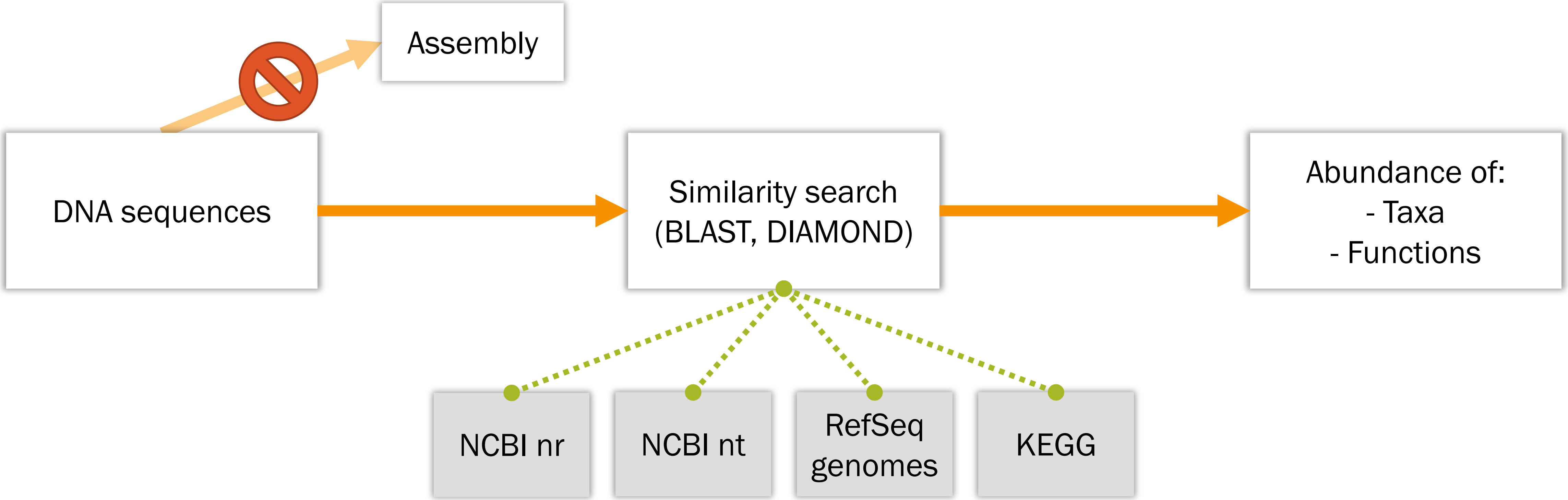
Environmental metagenomics

Read-based analyses: Part 1

Physalia
Courses

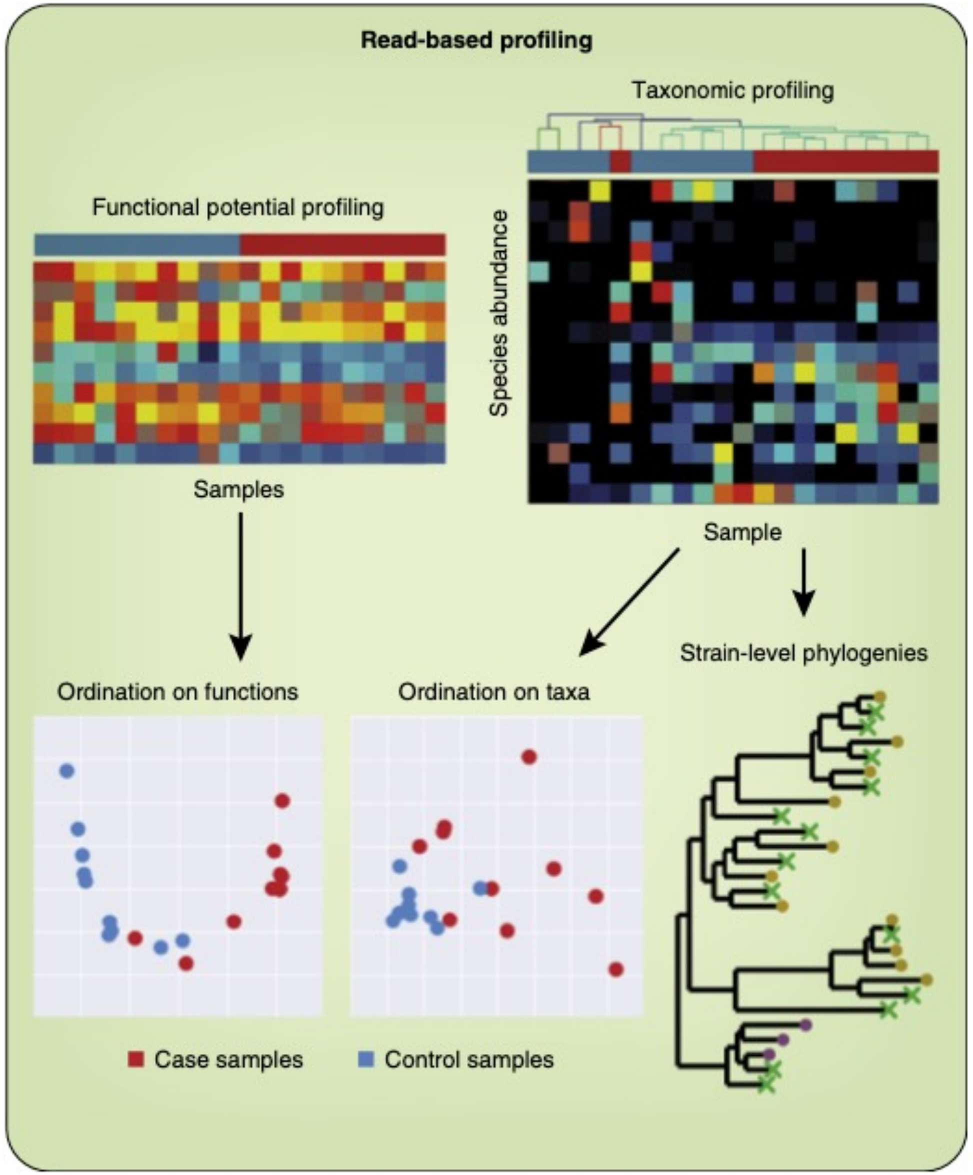
Igor S. Pessi & Antti Karkman, University of Helsinki

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- Somewhat outdated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

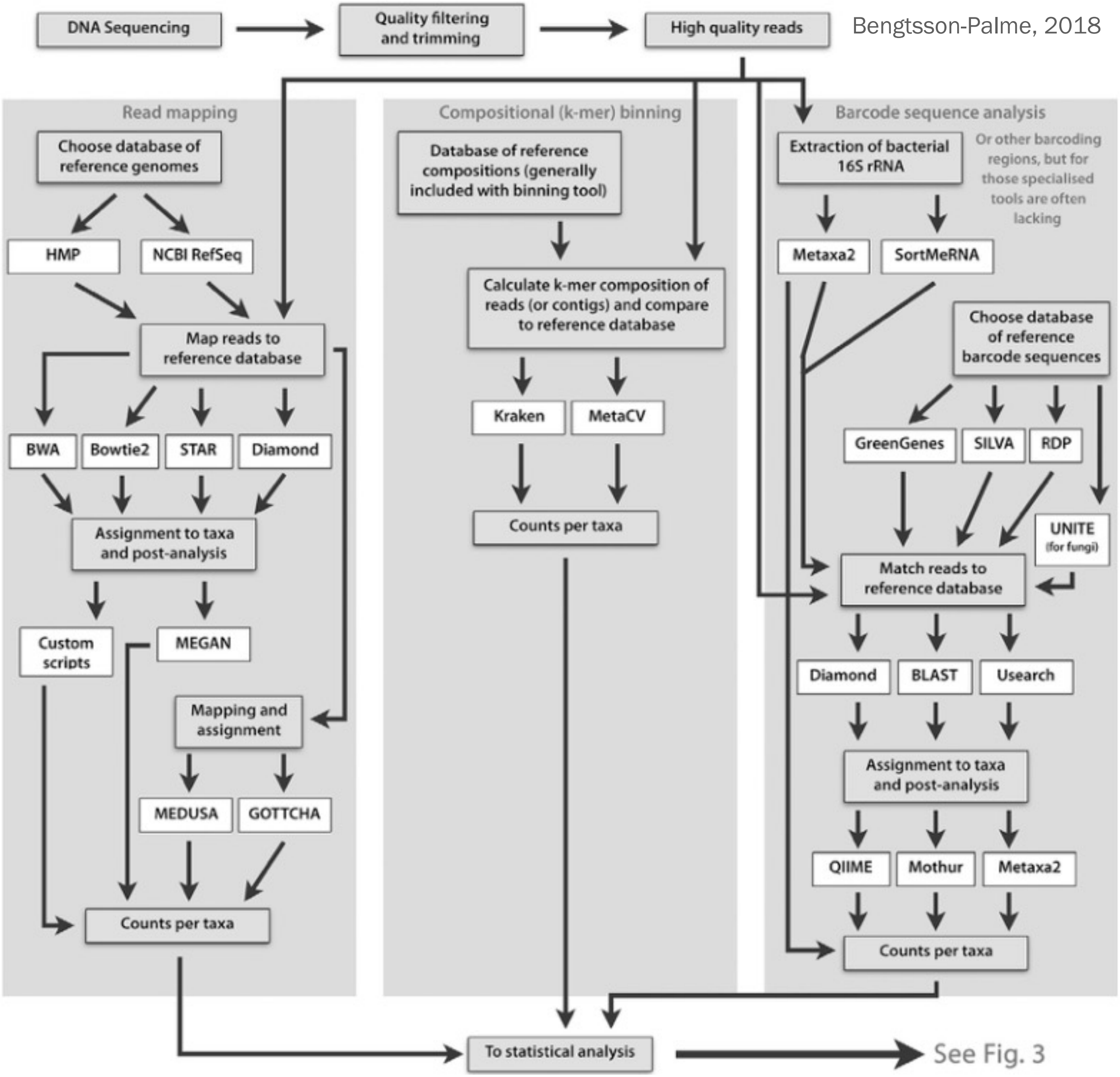


Quince et al. 2017

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



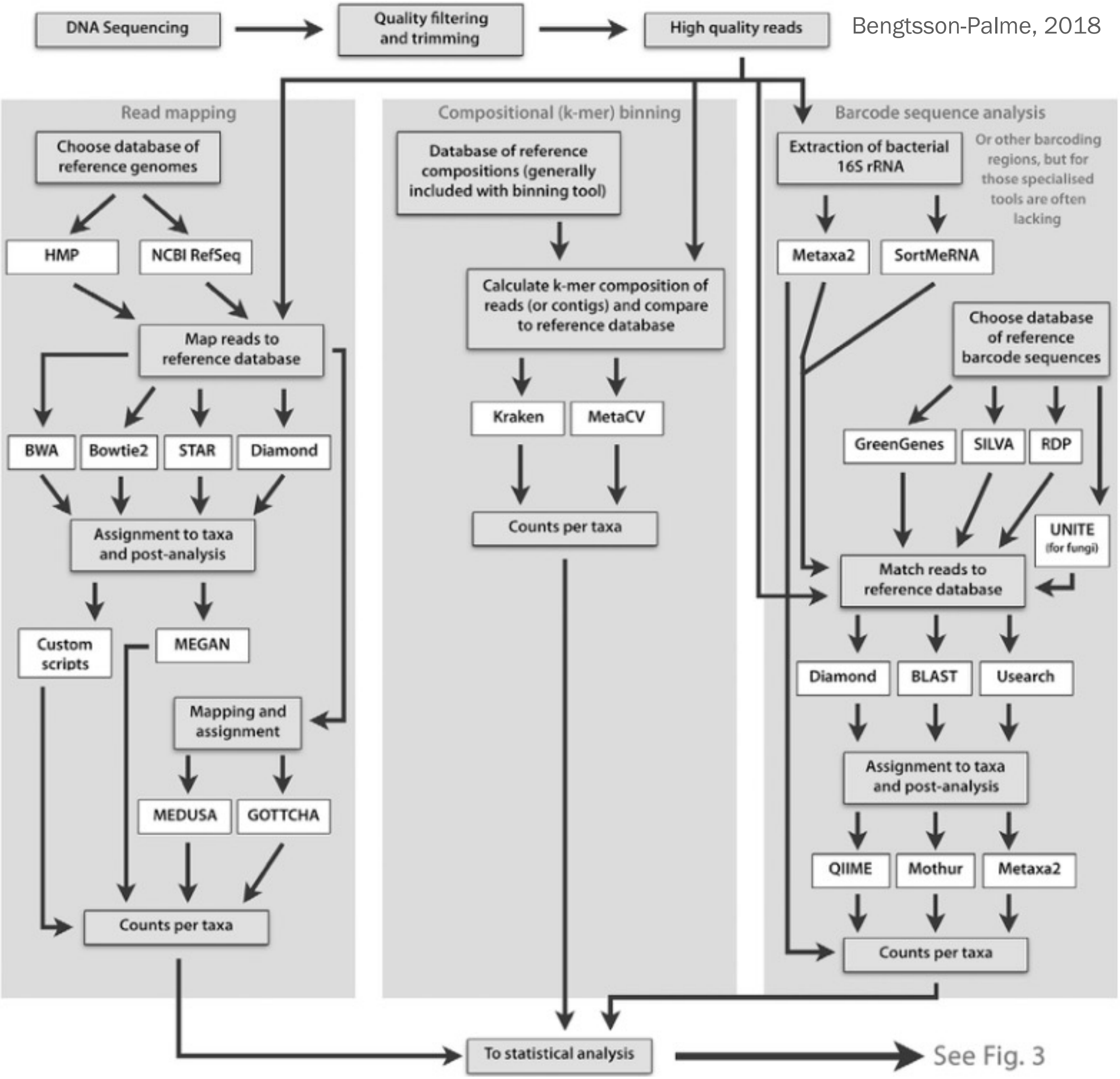
DNA Sequencing
Quality filtering and trimming
High quality reads
Read mapping
Choose database of reference genomes
HMP
NCBI RefSeq
Map reads to reference database
BWA
Bowtie2
STAR
Diamond
Assignment to taxa and post-analysis
Custom scripts
MEGAN
Mapping and assignment
MEDUSA
GOTTCHA
Counts per taxa
Compositional (k-mer) binning
Database of reference compositions (generally included with binning tool)
Calculate k-mer composition of reads (or contigs) and compare to reference database
Kraken
MetaCV
Counts per taxa
Barcode sequence analysis
Extraction of bacterial 16S rRNA
Or other barcoding regions, but for those specialised tools are often lacking
Metaxa2
SortMeRNA
Choose database of reference barcode sequences
GreenGenes
SILVA
RDP
UNITE (for fungi)
Match reads to reference database
Diamond
BLAST
Usearch
Assignment to taxa and post-analysis
QIIME
Mothur
Metaxa2
Counts per taxa
To statistical analysis
See Fig. 3

Bengtsson-Palme, 2018

Approaches to taxonomic profiling

Barcode sequence analysis

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


DNA Sequencing
Quality filtering and trimming
High quality reads
Read mapping
Choose database of reference genomes
HMP
NCBI RefSeq
Map reads to reference database
BWA
Bowtie2
STAR
Diamond
Assignment to taxa and post-analysis
Custom scripts
MEGAN
Mapping and assignment
MEDUSA
GOTTCHA
Counts per taxa
Compositional (k-mer) binning
Database of reference compositions (generally included with binning tool)
Calculate k-mer composition of reads (or contigs) and compare to reference database
Kraken
MetaCV
Counts per taxa
Barcode sequence analysis
Extraction of bacterial 16S rRNA
Or other barcoding regions, but for those specialised tools are often lacking
Metaxa2
SortMeRNA
Choose database of reference barcode sequences
GreenGenes
SILVA
RDP
UNITE (for fungi)
Match reads to reference database
Diamond
BLAST
Usearch
Assignment to taxa and post-analysis
QIIME
Mothur
Metaxa2
Counts per taxa
To statistical analysis
See Fig. 3

Bengtsson-Palme, 2018

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

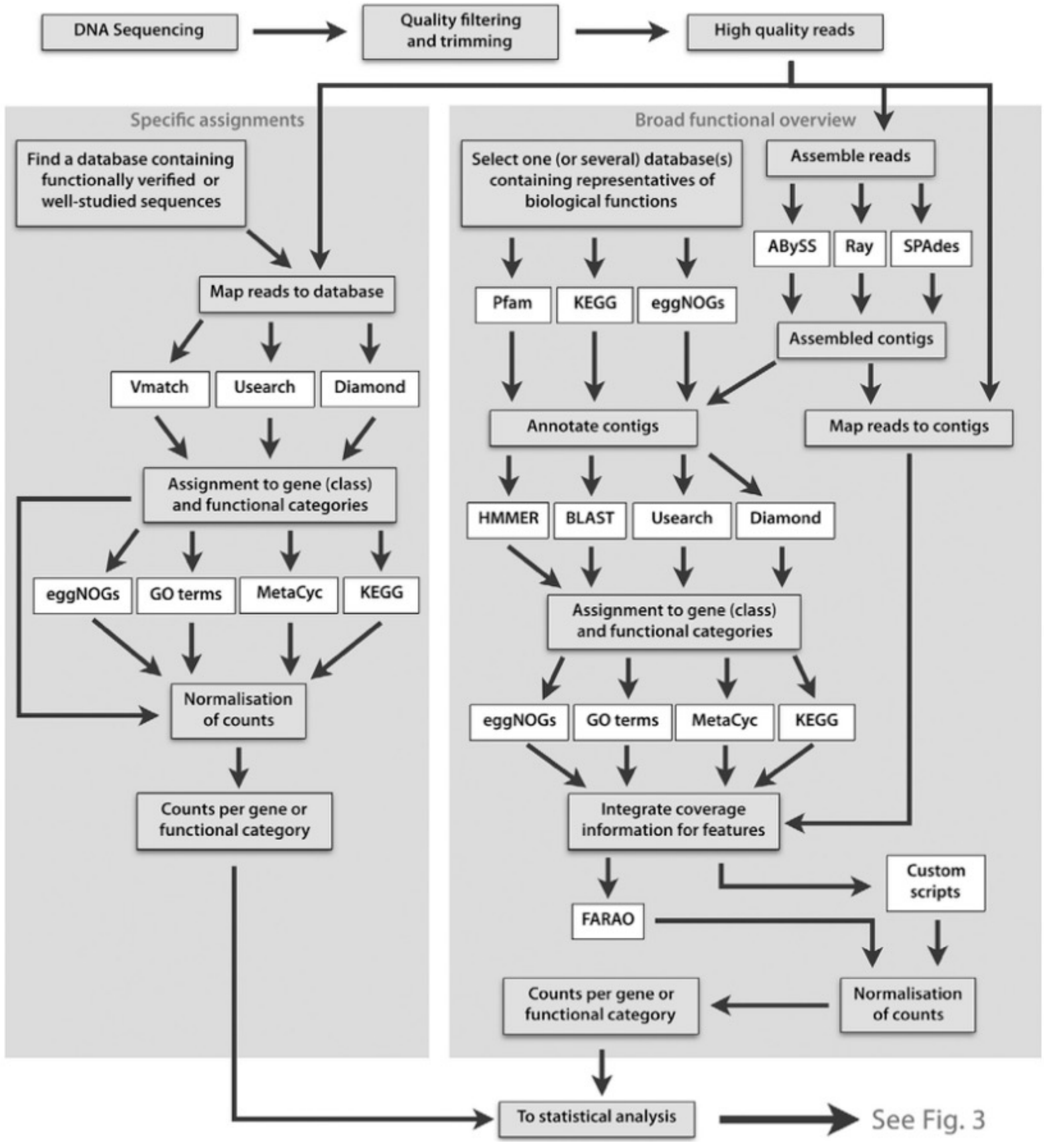
- More suitable for environments with a high fraction of unknown microorganisms (e.g. soil)

Approaches to functional profiling

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)

Bengtsson-Palme, 2018


DNA Sequencing
Quality filtering and trimming
High quality reads
Specific assignments
Find a database containing functionally verified or well-studied sequences
Map reads to database
Vmatch
Usearch
Diamond
Assignment to gene (class) and functional categories
eggNOGs
GO terms
MetaCyc
KEGG
Normalisation of counts
Counts per gene or functional category
Broad functional overview
Select one (or several) database(s) containing representatives of biological functions
Assemble reads
ABySS
Ray
SPAdes
Pfam
KEGG
eggNOGs
Assembled contigs
Annotate contigs
Map reads to contigs
HMMER
BLAST
Usearch
Diamond
Assignment to gene (class) and functional categories
eggNOGs
GO terms
MetaCyc
KEGG
Integrate coverage information for features
Custom scripts
FARAO
Counts per gene or functional category
Normalisation of counts
To statistical analysis
See Fig. 3

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

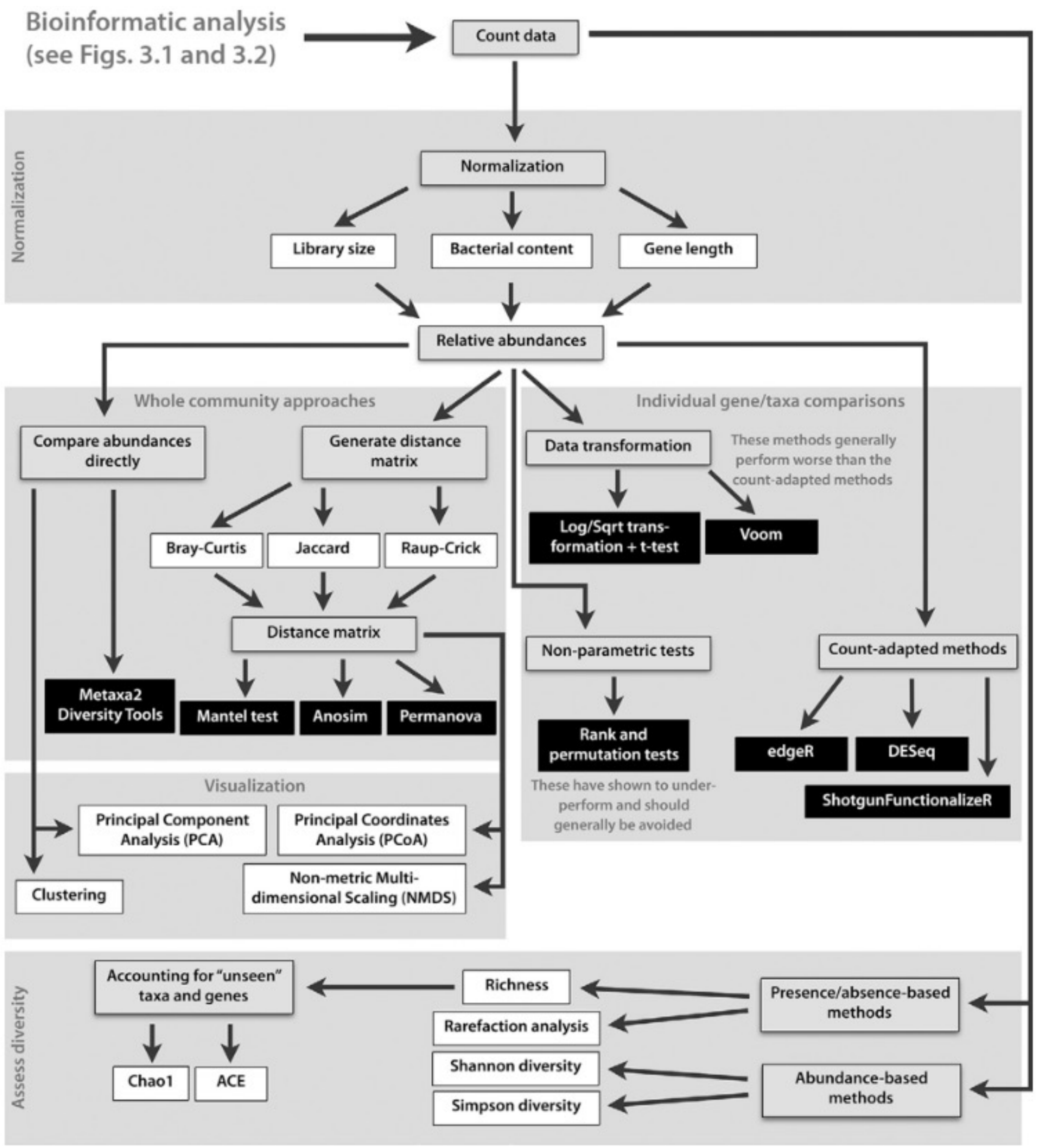
Specific databases are often highly curated and can give substrate-level information

- 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)
    - Pitfalls: data distribution, zero counts
  - Multivariate (e.g. PERMANOVA, ordination/clustering, Mantel test)
- Normalization!
  - Library size
  - Bacterial content (e.g *rpoB* gene)

Bengtsson-Palme, 2018



Bioinformatic analysis (see Figs. 3.1 and 3.2)

Count data

Normalization

Normalization

Library size

Bacterial content

Gene length

Relative abundances

Whole community approaches

Compare abundances directly

Generate distance matrix

Bray-Curtis

Jaccard

Raup-Crick

Distance matrix

Metaxa2 Diversity Tools

Mantel test

Anosim

Permanova

Visualization

Principal Component Analysis (PCA)

Principal Coordinates Analysis (PCoA)

Clustering

Non-metric Multi-dimensional Scaling (NMDS)

Individual gene/taxa comparisons

Data transformation

These methods generally perform worse than the count-adapted methods

Log/Sqrt trans-formation + t-test

Voom

Non-parametric tests

Rank and permutation tests

These have shown to under-perform and should generally be avoided

Count-adapted methods

edgeR

DESeq

ShotgunFunctionalizeR

Assess diversity

Accounting for "unseen" taxa and genes

Chao1

ACE

Richness

Rarefaction analysis

Shannon diversity

Simpson diversity

Presence/absence-based methods

Abundance-based methods

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)
  - Investigate other genes belonging to the same pathway

Tools for read-based analyses

- MEGAN
  - Both taxonomic (all reads) and functional
- METAXA
  - Taxonomic (marker gene, SSU or LSU)
- Humann
- Metaphlan
- MG-RAST

Now let's:

- Take a look at the script together and run it
- Go through one of the samples together in MEGAN
- Use MEGAN to compare the four samples

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