UFRGS
UNIVERSIDADE FEDERAL
DO RIO GRANDE DO SUL

PPGBCM
CBiot/UFRGS

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

CSC

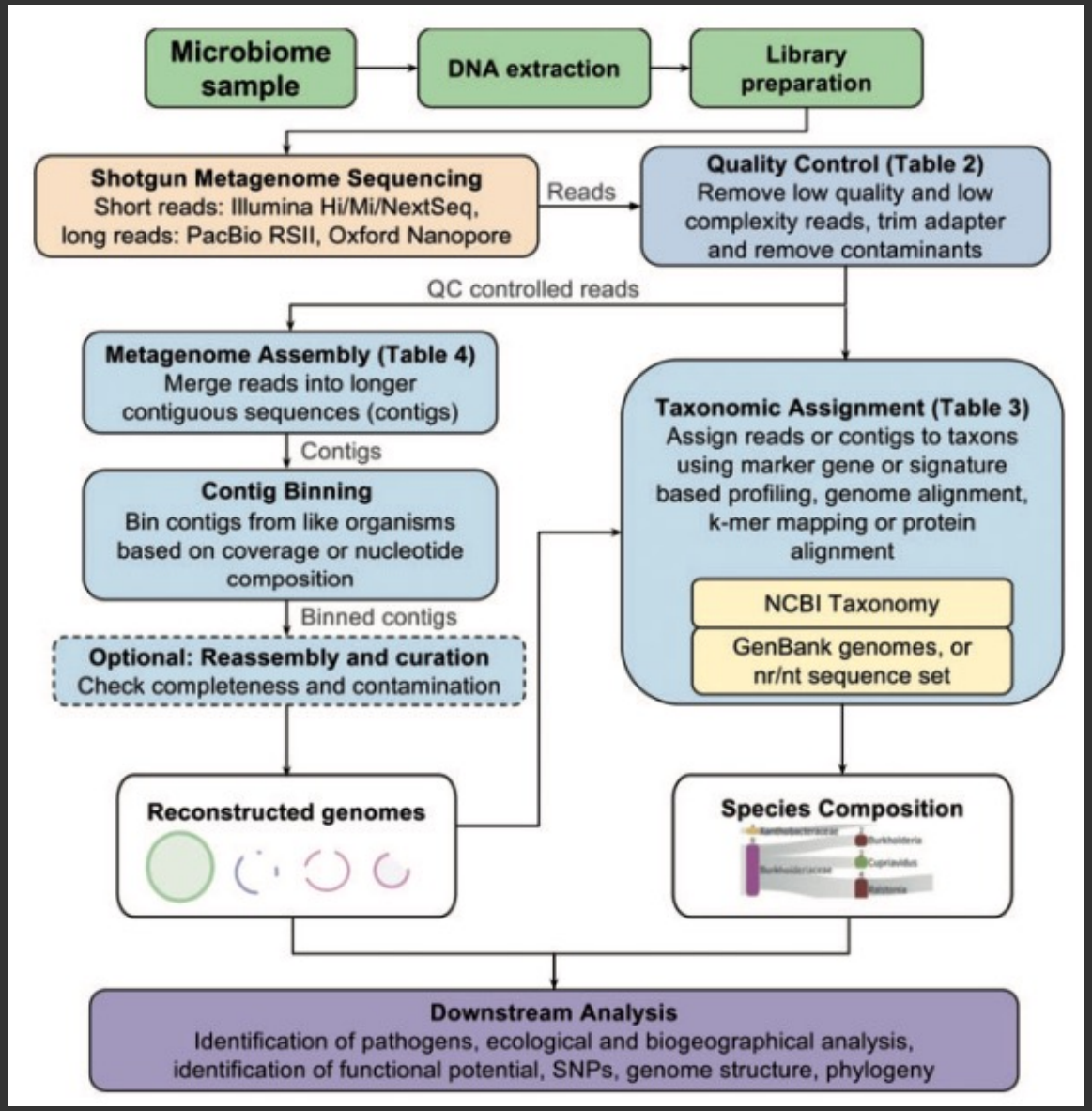
Introdução à metagenômica: curso prático

Obtenção de perfis taxonômicos e funcionais

[igor_pessi@UniversityOfHelsinki]$ _

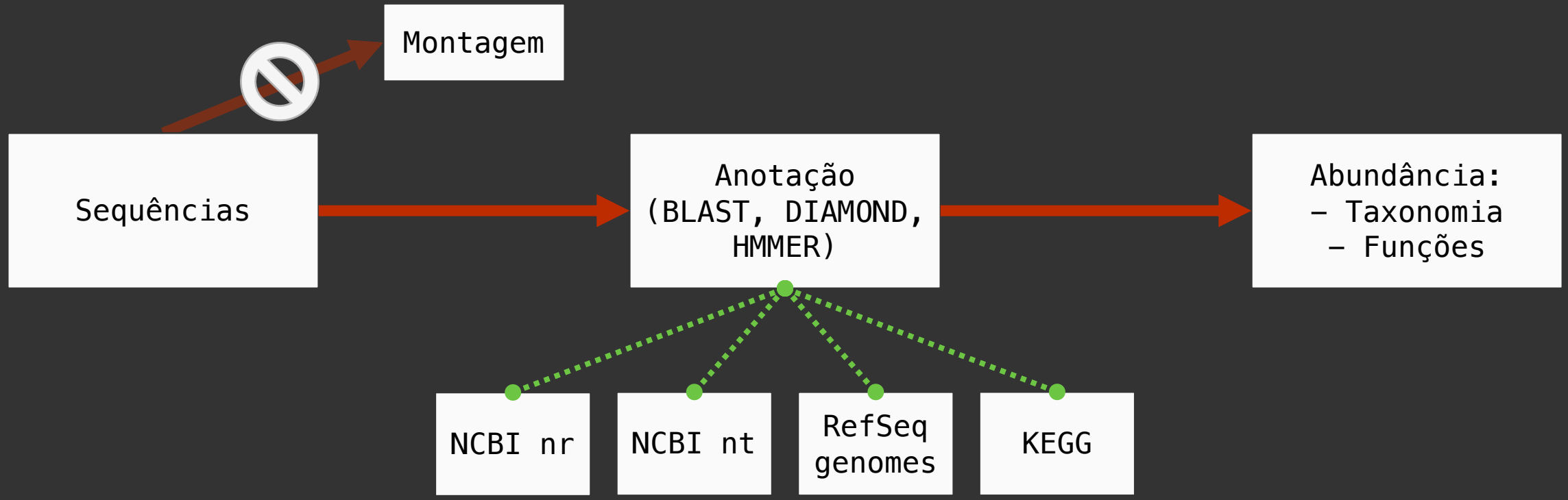
Novembro/2022

Fluxo típico de uma análise metagenômica



Breitwieser et al. (2019)

Perfis baseados em sequências curtas (*read-based profiling*)



Perfis baseados em sequências curtas são

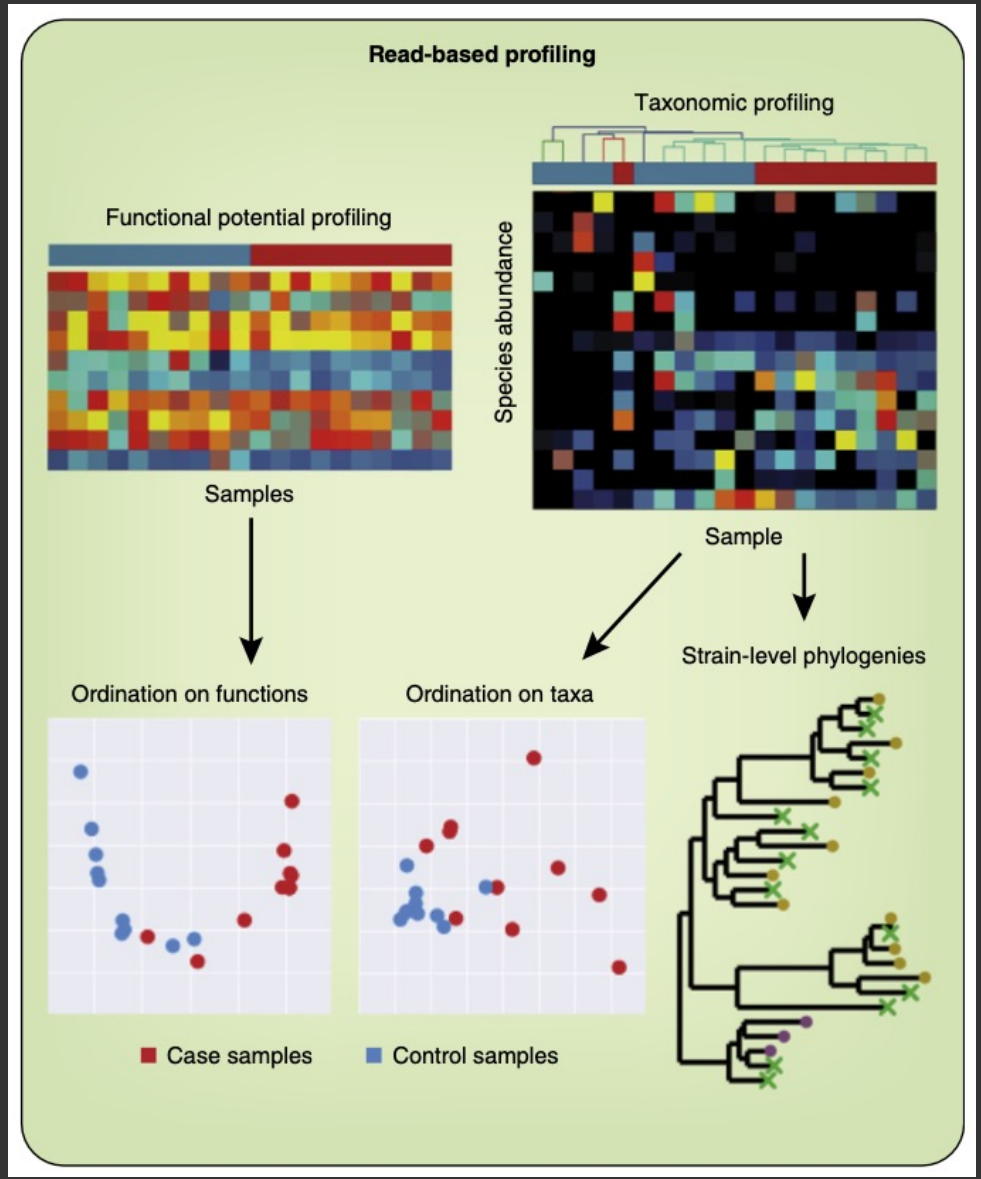
Rápidos de obter

Quantitativos

De certa maneira ultrapassados

Montagem de metagenomas é preferível

Pode gerar uma visão geral preliminar, *quick and dirty*



Quince et al. (2017)

Abordagens para obtenção de perfis taxonômicos

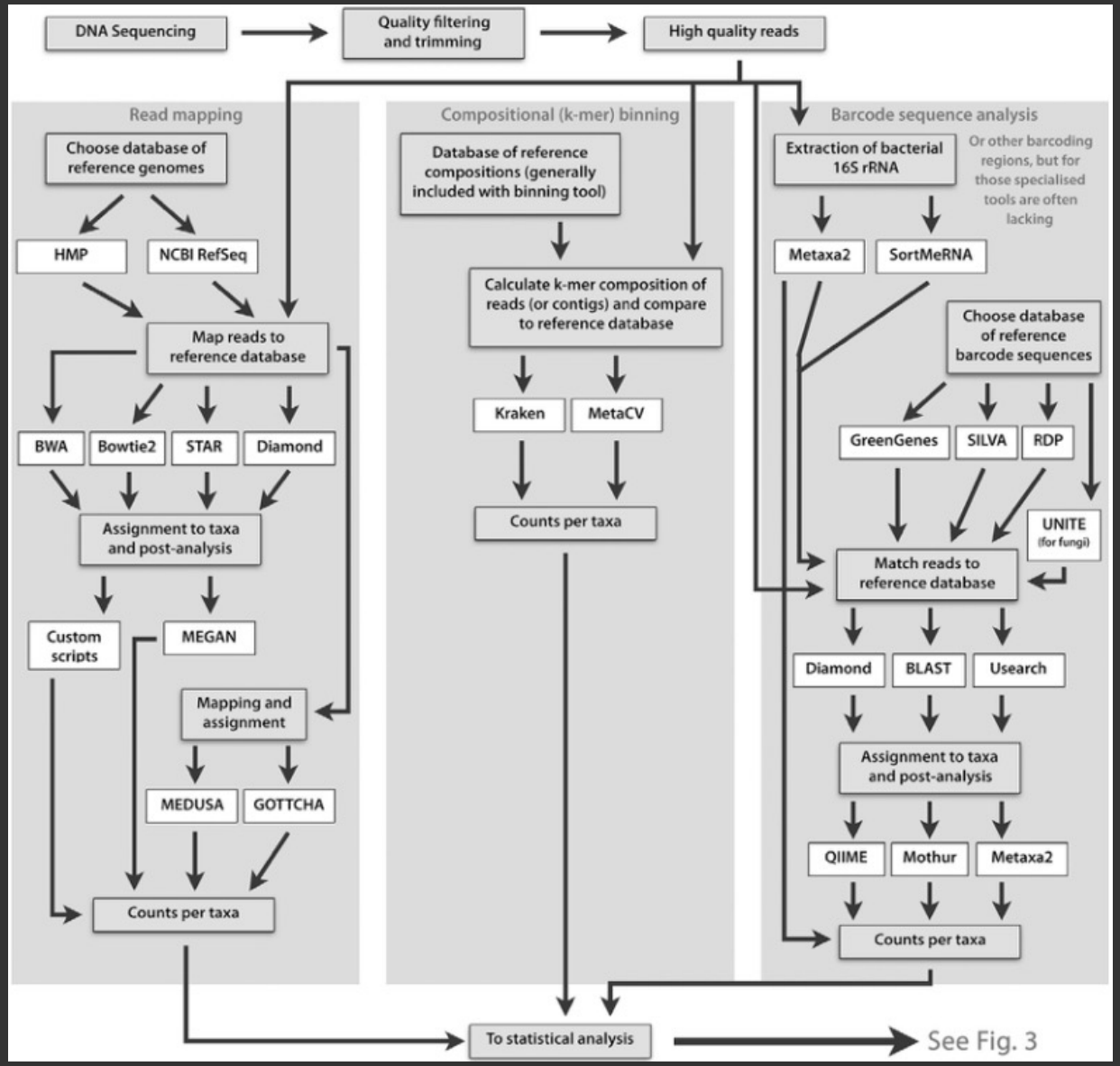
Mapeamento de sequências agrupamento composicional

Analisa todas as sequências

Bancos de genomas referências

Mapeamento: lento, requer bastante processamento e RAM

Agrupamento composicional: mais rápido mas menos preciso

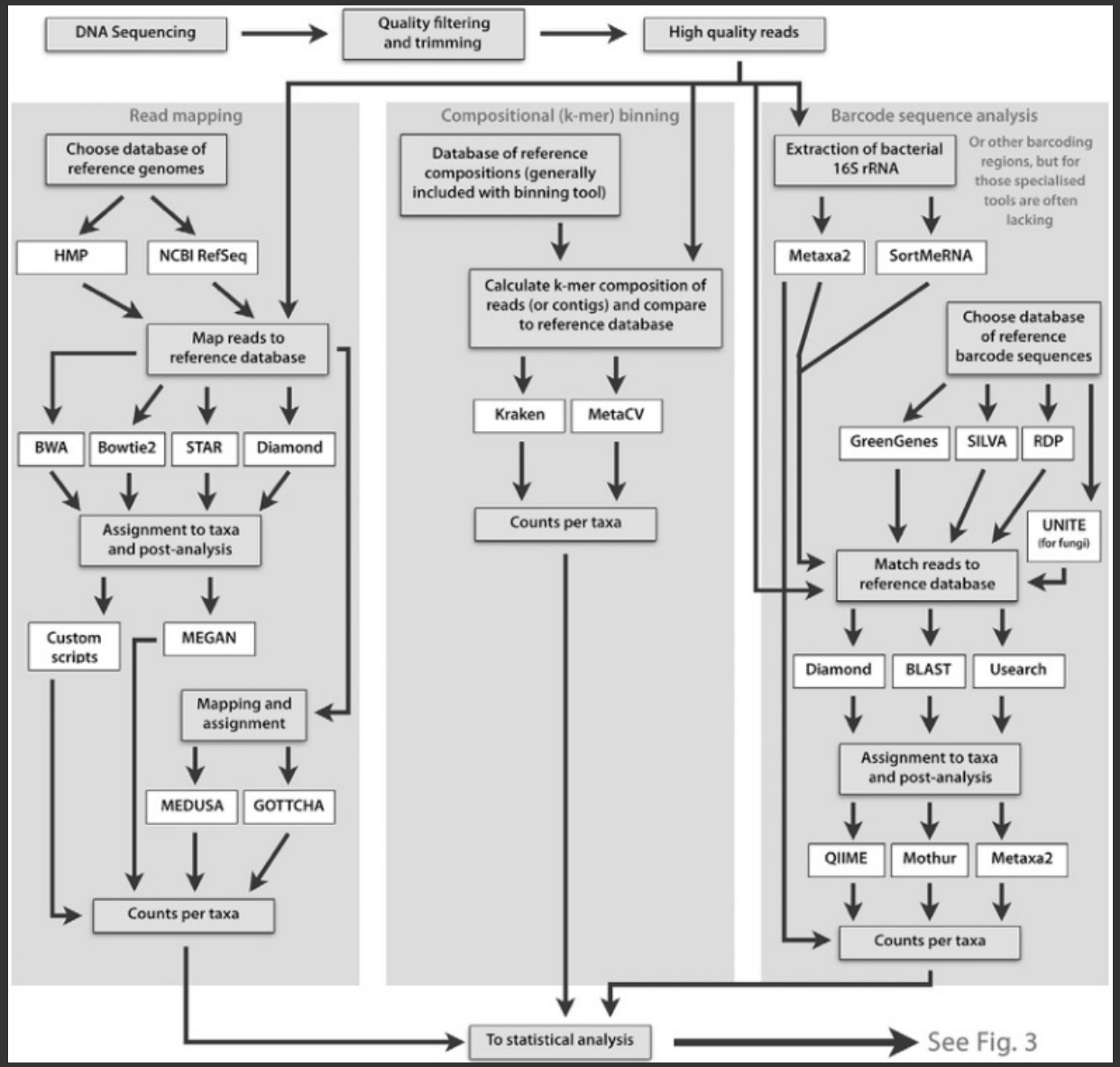


Bengtsson-Palme (2018)

Abordagens para obtenção de perfis taxonômicos

Análise de *barcoding*

- # Analise genes específicos (por exemplo, 16S rRNA)
- # Bancos de sequências curadas (por exemplo, SILVA)
- # Muito mais rápido do que as outras abordagens, mas resolução depende do gene utilizDO



Bengtsson-Palme (2018)

Abordagens para obtenção de perfis taxonômicos: como escolher?

Análise de todas sequências (mapeamento de sequências e agrupamento composicional) sofre com bancos de genomas limitados

```
# Mais adequado para ambientes
  melhor descritos (por exemplo,
  microbioma humano)
```

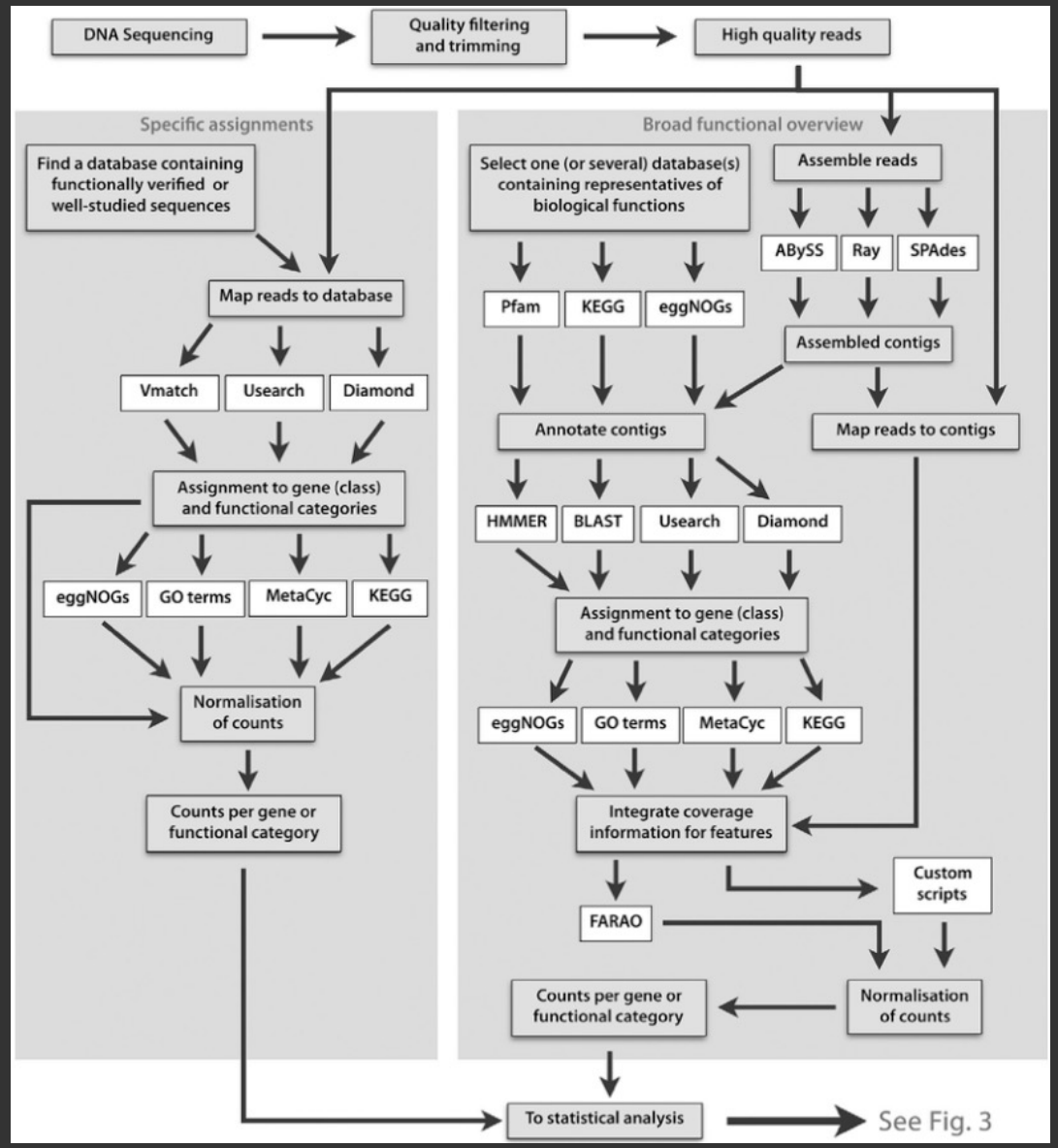
Análise de *barcoding* oferece menor resolução

```
# Mais adequado para ambientes
  com uma alta fração de
  microrganismos desconhecidos
  (por exemplo, solo)
```

Abordagens para obtenção de perfis funcionais

Perfis amplos *vs* específicos

- # Bancos de dados amplos:
  todo universo funcional
  (e.g. KEGG, PFAM)
- # Bancos de dados específicos:
  foco em um ou mais processos
  (e.g. CAZy, CARD)



Bengtsson-Palme (2018)

Abordagens para obtenção de perfis taxonômicos: como escolher?

Bancos de dados amplos fornecem uma visão geral do potencial funcional das comunidades microbianas

```
# Adequado para investigar grandes diferenças entre ambientes
```

Bancos de dados específicos geralmente são melhor curados e podem fornecer informações à nível de substrato

```
# Adequado para investigar variantes de genes em ambientes relacionados
```

Interpretação de perfis taxonômicos e funcionais

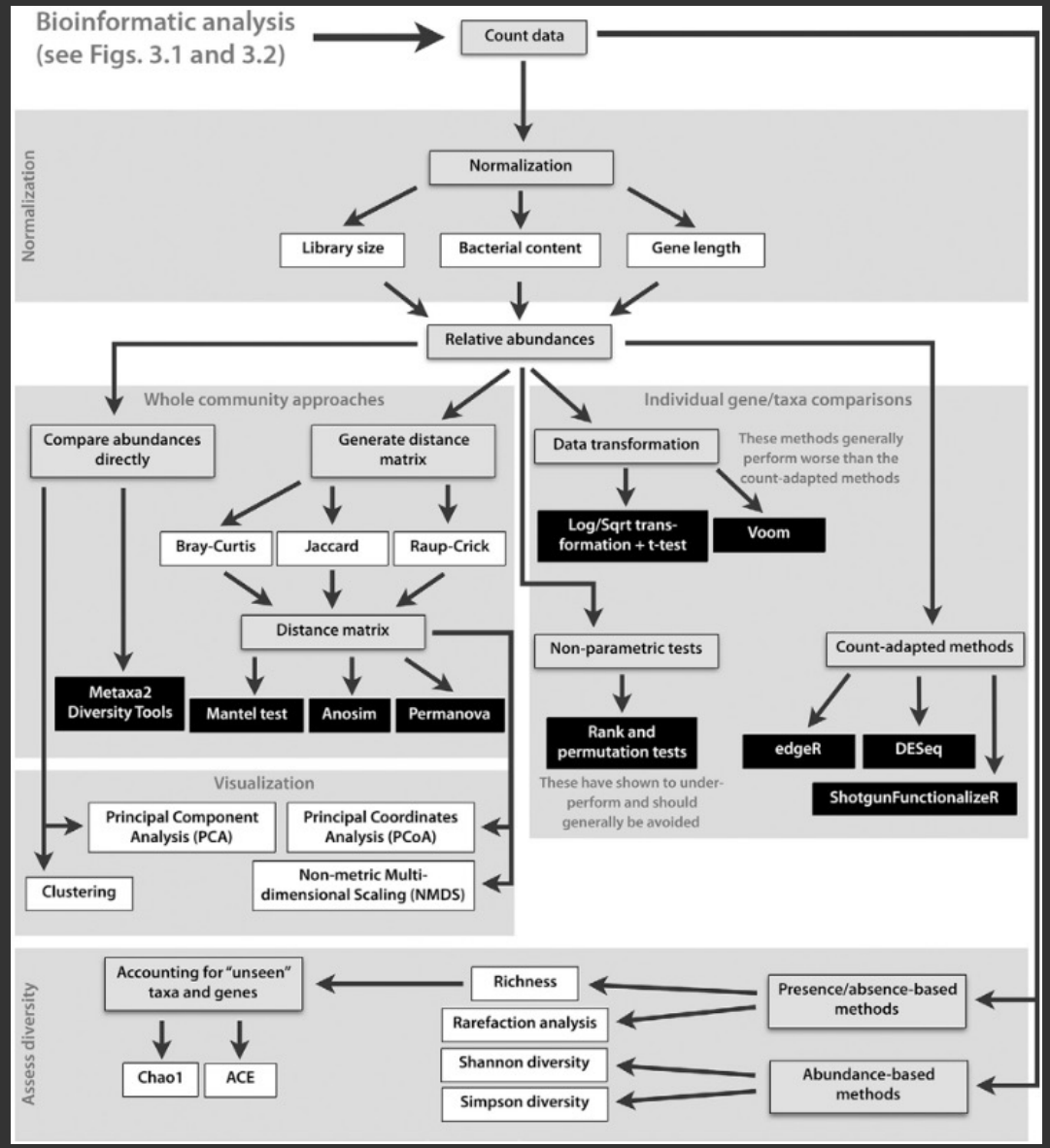
Análises comparativas

Estatística

- # Univariada (e.g. ANOVA individual para cada táxon/gene)
- # Multivariada (e.g. PERMANOVA, ordenamento, teste de Mantel)

Normalização

- # Tamanho da biblioteca
- # Abundância (e.g gene *rpoB*)



Bengtsson-Palme (2018)

Pontos fortes e fracos de perfis baseados em sequências curtas

Abrangência	Fornecer uma imagem agregada da função ou estrutura da comunidade, mas é baseado apenas na pequena fração existente em bancos de dados
Complexidade da comunidade	Pode lidar com comunidades complexas dependendo da profundidade de sequenciamento e cobertura em bancos de dados
Novidade	Não possibilita a resolução de organismos distantes de genomas de referência
Demanda computacional	Pode ser realizado de forma eficiente permitindo grandes meta-análises
Metabolismo à nível de genoma	Normalmente fornece apenas o metabolismo agregado da comunidade; conexões com filogenia só são possíveis no contexto de genomas de referência
Curadoria manual	Geralmente não requer curadoria manual, mas a seleção de genomas de referência a serem usados pode envolver supervisão humana
Integração com genômica	Perfis obtidos não podem ser colocados diretamente no contexto de genomas derivados de isolados cultivados

Quince et al. (2017)

Armadilhas de perfis baseados em sequências curtas

Nível de curadoria do banco de dados

```
# As sequências são verificadas experimentalmente?
```

Abrangência do banco de dados

```
# Tanto taxonômica quanto funcionalmente
```

Troca entre velocidade *vs* sensibilidade

```
# Por exemplo, BLAST vs DIAMOND
```

Escolha de limites de identidade, bitscore, e-value, cobertura

```
# Não é possível generalizar para todos os genes
```

Lembre-se: faça sempre testes de sanidade

Principalmente para resultados novos/não esperados:

```
    # Refaça as análises com limites mais estritos
    # Refaça com programas e bancos de dados diferente
    # Investigue outros genes pertencentes à mesma rota metabólica
```

Alguns exemplos de programas

Table 3. Metagenomic classifiers, aligners and profilers

Tool	Synopsis	Reference	Web site
Kraken	Fast taxonomic classifier using in-memory *k*-mer search of metagenomics reads against a database built from multiple genomes	[64]	https://ccb.jhu.edu/software/kraken/
Kraken-HLL	Extension of Kraken counting unique *k*-mers for taxa and allowing multiple databases		https://github.com/fbreitwieser/kraken-hll
CLARK(-S)	Fast taxonomic classifier using in-memory *k*-mer search of metagenomics reads against a database built from completed genomes. S extension uses spaced *k*-mer seeds for better classification	[65, 66]	http://clark.cs.ucr.edu
Kallisto	Taxonomic profiler using pseudo-alignment with *k*-mers using techniques based on transcript (RNA-seq) quantification	[67]	https://github.com/pachterlab/kallisto
k-SLAM	Taxonomic classifier using database of nonoverlapping *k*-mers in genomes. Reads are split into *k*-mers, and overlaps found by lexicographical ordering are pseudo-assembled	[68]	https://github.com/aindj/k-SLAM
Kaiju	Fast taxonomic classifier against protein sequences using FM-index with reduced amino acid alphabet	[69]	https://github.com/bioinformatics-centre/kaiju
DIAMOND	Protein homology search using spaced seeds with a reduced amino acid alphabet, 2000–20 000 times faster than BLASTX	[70]	https://github.com/bbuchfink/diamond
BLAST+	Highly sensitive nucleotide and translated-nucleotide protein alignment	[61, 71]	https://blast.ncbi.nlm.nih.gov
MEGAN6/CE	Desktop and Web metagenomics analysis suite. Uses BLAST or diamond to match sequences and assigns LCA of matches	[72, 73]	http://ab.inf.uni-tuebingen.de/software/megan6/
DUDes	Top-down assignment of metagenomics reads	[74]	https://sourceforge.net/projects/dudes/
Taxonomer	Web-based metagenomics classifier including binning and visualization	[75]	http://taxonomer.iobio.io/
GOTTCHA	Taxonomic profiler that maps reads against short unique subsequences ('signature') at multiple taxonomic ranks	[76]	http://lanl-bioinformatics.github.io/GOTTCHA/
LMAT(-ML)	K-mer-based taxonomic read classifier using extensive database including draft genomes and eukaryotes. ML (Marker Library) extension reduces RAM requirements by stringent pruning of non-informative and overlapping *k*-mers	[77, 78]	https://sourceforge.net/projects/lmat/
taxator-tk	Uses BLAST or LAST output for binning and taxonomic assignment via overlapping regions and pairwise distance measures	[79]	https://github.com/fungs/taxator-tk
Centrifuge	Fast taxonomic classifier using database compressed with FM-index, database and output format similar to Kraken	[80]	http://ccb.jhu.edu/software/centrifuge/
MetaPhlAn 2	Marker gene-based taxonomic profiler	[81]	https://bitbucket.org/biobakery/metaphlan2
mOTU	Taxonomic profiler based on a set of 40 prokaryotic marker genes	[82]	http://www.bork.embl.de/software/mOTU/
Mash	MinHash-based taxonomic profiler enabling superfast overlap estimations	[83]	http://mash.readthedocs.io
sourmash	Alternative implementation of MinHash algorithm using fast searches with sequence bloom trees for taxonomic profiling	[84]	https://github.com/dib-lab/sourmash
PanPhlAn	Pan-genome-based phylogenomic analysis	[2]	http://segatalab.cibio.unitn.it/tools/panphlan/

Breitwieser et al. (2019)