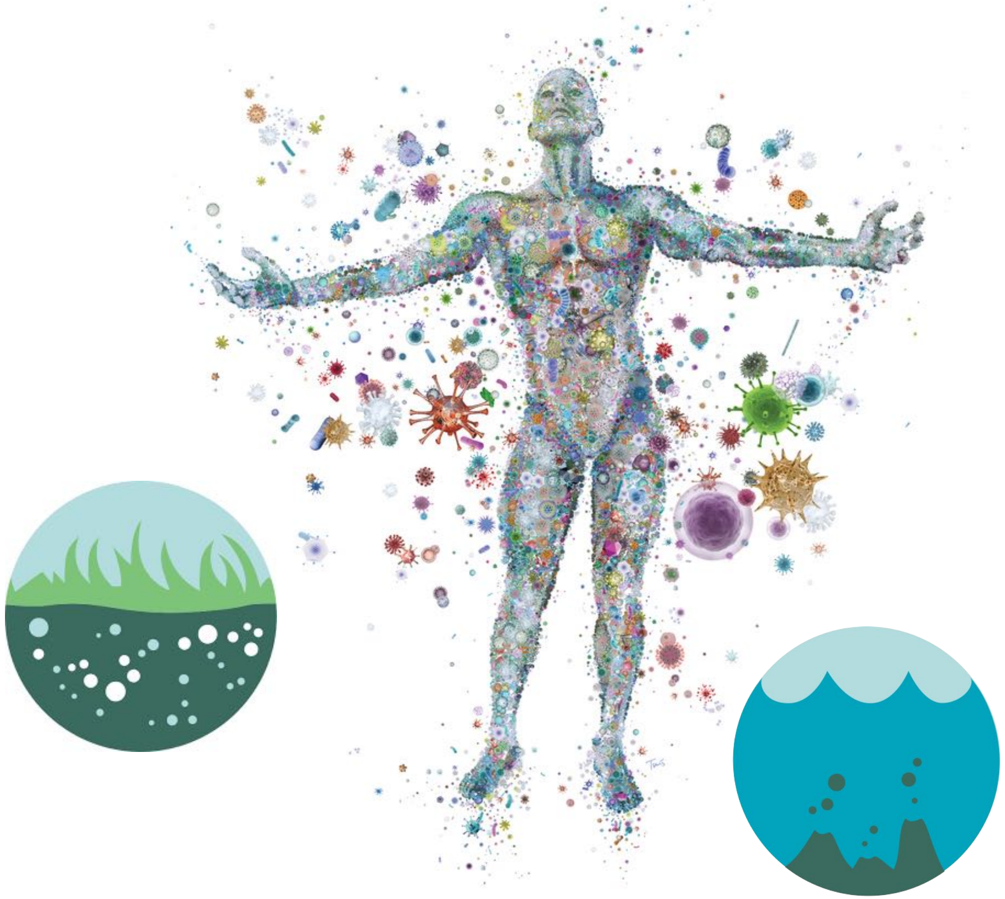
National Institute of Allergy and Infectious Diseases

METAGENOMICS OVERVIEW

Metagenomics

MSB7105

April, 2021



NIAID

NIH National Institute of Allergy and Infectious Diseases

Angelina Angelova, PhD

Bioinformatics and Computational Biosciences Branch (BCBB)

OCICB/OSMO/OD/NIAID/NIH

Today’s instructor

Angelina Angelova, PhD
Metagenomics Analysis Specialist

Bioinformatics and Computational Biosciences Branch (BCBB)
National Institute of Allergies and Infectious Diseases (NIAID)
National Institute of Health (NIH)
Bethesda, MD, USA

Contact instructor:
angelina.angelova@nih.gov
Contact our team:
bioinformatics@niaid.nih.gov

NIH National Institute of Allergy and Infectious Diseases

NIAID

Define ”Metagenomics”

- Metagenomics: Refers to the idea that the collection of genes (the metagenome), obtained directly from a community in its natural habitat (the microbiome), can provide an understanding of the functional and taxonomic traits of the whole community.
- NGS made the field of metagenomics possible
- Metagenomics bypasses the need for isolation or cultivation of individual microbes.
- Allows for exploration of the structure (abundance & identities), interactions, strategies (communication, survival, etc.), functionality and dynamics of a community

Example microbiomes:

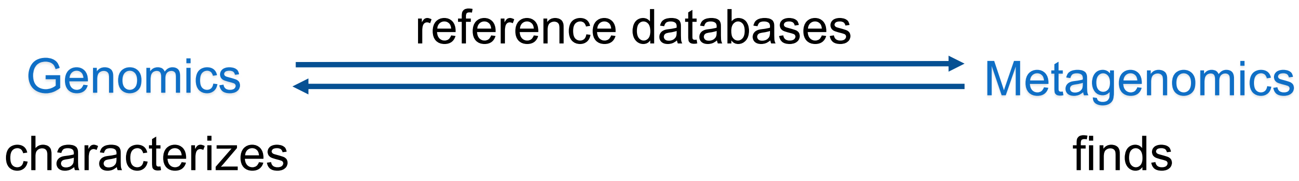
National Institute of Allergy and Infectious Diseases

https://www.ncbi.nlm.nih.gov/books/NBK54011/

Reference genomic databases

A reference genomic databases are a collection of DNA sequences that are idealistic genomic representations of recognized organisms. These sequences are sourced either from individual cultivated organisms (a type strain representing that lineage) or in case of more complex organisms – from multiple organisms from the same species (e.g. human).

- RefDBs allow for
  - Taxonomic characterization of specific species, through identification of conserved genes within that organism's genome (genetic markers).
  - Functional characterization of genes/proteins through understanding of known gene/proteins



reference databases

Genomics

Metagenomics

characterizes

finds

Shotgun VS Amplicon

Shotgun Strategy

ALL the DNA from ALL the genomes within the ENTIRE community, is fragmented to the “bite-size” capacity of a sequencing platform. ALL DNA is sequenced. The sequences are used to explore taxonomic composition *and* functional capacity of the entire community

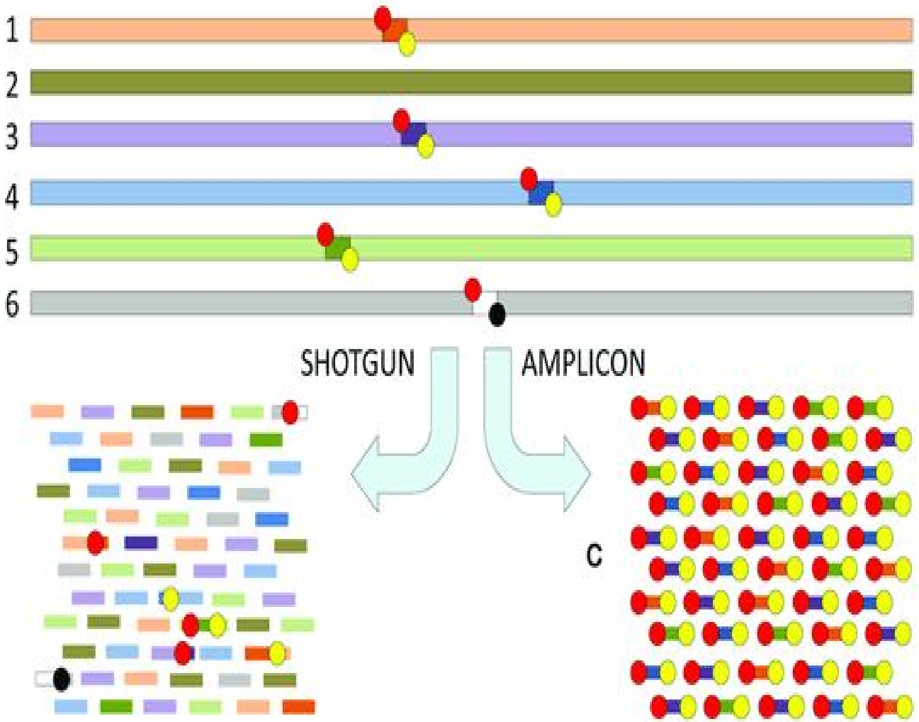
Long-read Strategy

ALL the DNA from ALL the genomes within the ENTIRE community, is sequenced in “large bites”. The sequences are used to explore taxonomic composition *and* functional capacity of the entire community

Common platforms

For long reads:

- ○ PacBio, Nanopore (MinIon)



Common platforms

For Short reads:

- ○ Illumina HiSeq, NextSeq, NovoSeq

Amplicon Strategy

<u>One gene</u> (a marker gene or a fraction of it) from ALL the genes from within ALL the genomes of ALL the organisms in a community, is targeted for amplification. Its sequence is used to explore the taxonomic composition of the entire community.

Common marker genes:

- ➢ For Bacterial & Archaeal organisms:
  - ○ 16S rRNA gene
- ➢ For Eukaryotic organisms:
  - ○ 18S rRNA *gene* (less conserved)
  - ○ ITS: internal transcribed spacer region

Common platforms

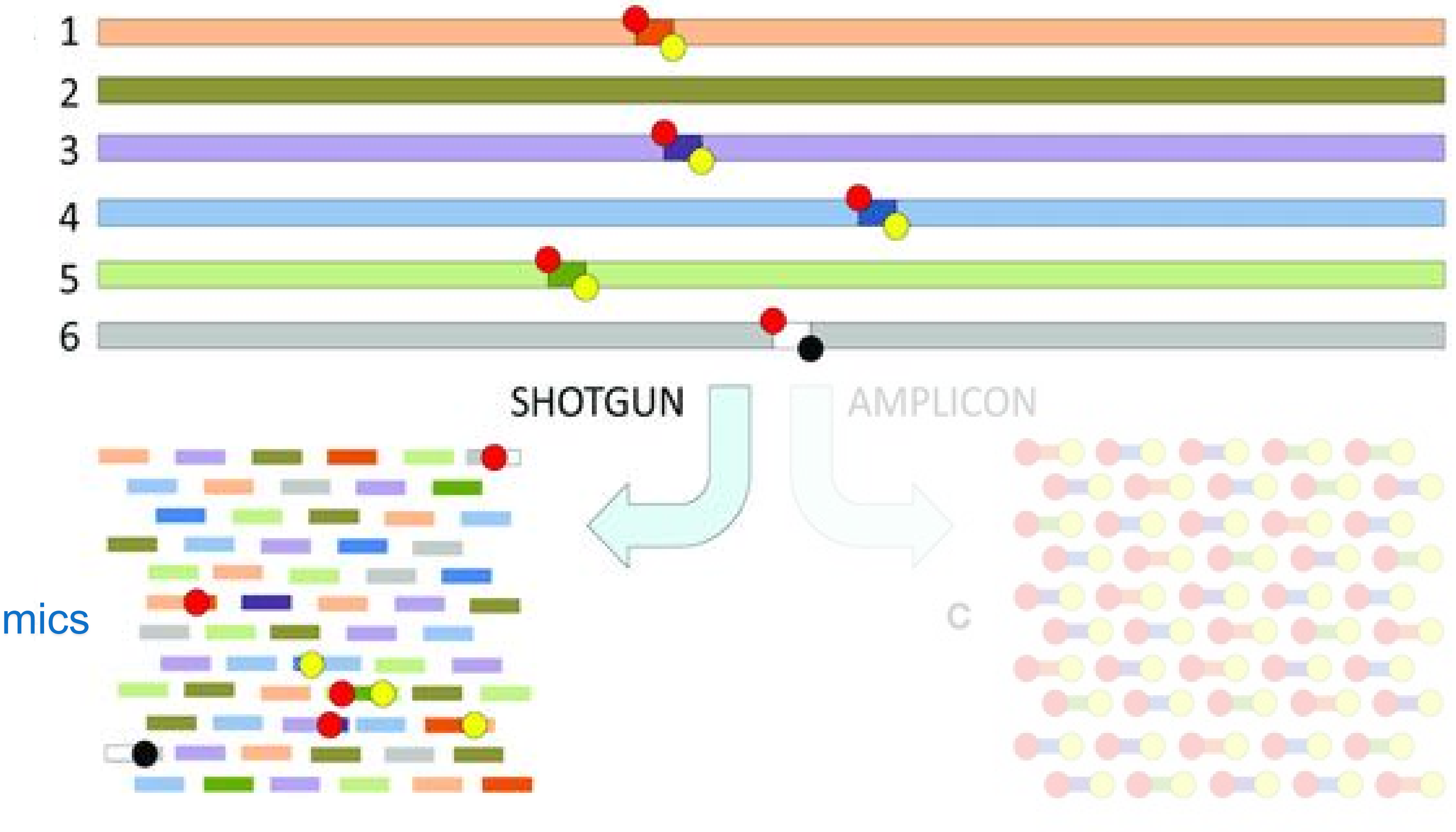
For Short reads:

- ○ Illumina MiSeq, NextSeq
- ○ TermoFisher IonTorrent

National Institute of Allergy and Infectious Diseases

JOURNAL OF CLINICAL MICROBIOLOGY, Sept. 2007, 45 (9), https://jcm.asm.org/content/45/9/2761.short
PNAS April 17, 2012 109 (16) 6241-6246; https://doi.org/10.1073/pnas.1117018109
Nature Biotechnology. Sept. 2017. https://doi.org/10.1038/nbt.3935

Shotgun meta-genomics



Metagenomics

Whole metagenome sequencing (WmGS)

National Institute of Allergy and Infectious Diseases

Down rabbit hole #2

Questions addressable by <u>shotgun </u>sequencing

- What organisms are present in microbiome 1 and in what proportion? (community structure)
- What is the natural variation of microbiome 1?
- How is microbiome 1 different from microbiome 2 in its taxonomic composition?
- Are there more of organism 1 in microbiome 1 than in microbiome B?
- Which microbiome has higher diversity of bugs?
- What is the natural core microbiome (non-variable faction) in microbiome 1 vs microbiome 2?
- How does the diversity of community change with factor 1 or factor 2?
- Which organism responds to factors 1 or 2?

Meta-taxon-omics

+

- **What additional information can we get from shotgun sequencing ?**
- If we have the near-full genomes of all organisms in 2 communities, what kind of questions can we answer?

Questions addressable by shotgun sequencing

- What organisms are present in microbiome 1 and in what proportion? (community structure)
- What is the natural variation of microbiome 1?
- How is microbiome 1 different from microbiome 2 in its taxonomic composition?
- Are there more of organism 1 in microbiome 1 than in microbiome B?
- Which microbiome has higher diversity of bugs?
- What is the natural core microbiome (non-variable faction) in microbiome 1 vs microbiome 2?
- How does the diversity of community change with factor 1 or factor 2?
- Which organism responds to factors 1 or 2?

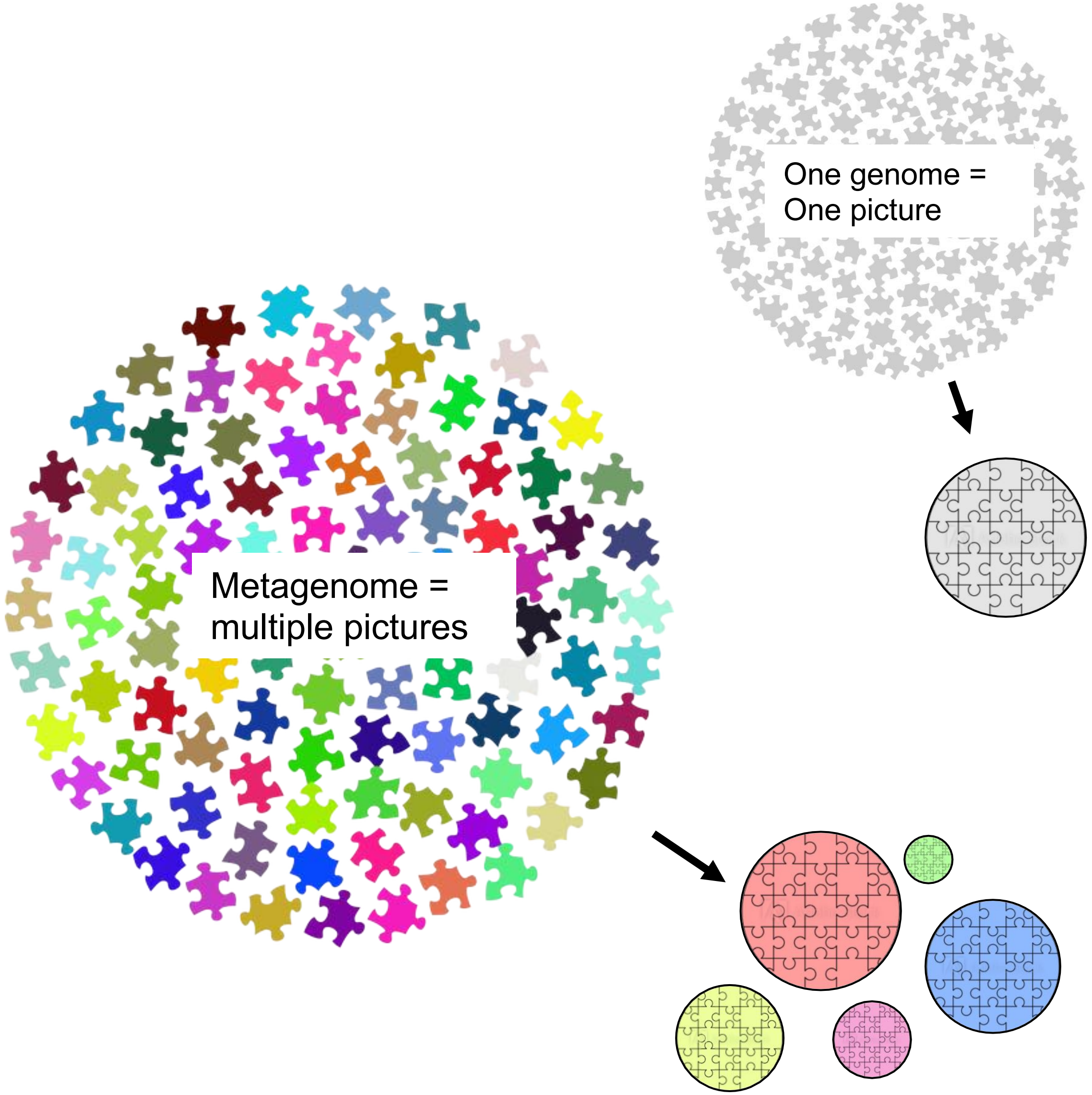
Meta-taxonomics

+

- **What were organisms 1, 2 or 3 capable of doing (functional capacity)?**
- **What is the internal diversity of organism 1 (strains)?**
- **How is the community functionally responding to factor 1?**
- **What fraction of the microbiome are presented by organisms from other domains (e.g. viral, eukaryotic composition)?**
- **What is the reservoir of genes within a community used to degrade substrate 1? What are these genes?**
- **Which community has more genes or higher diversity of genes involved in function A?**

Caveats of shotgun sequencing approach

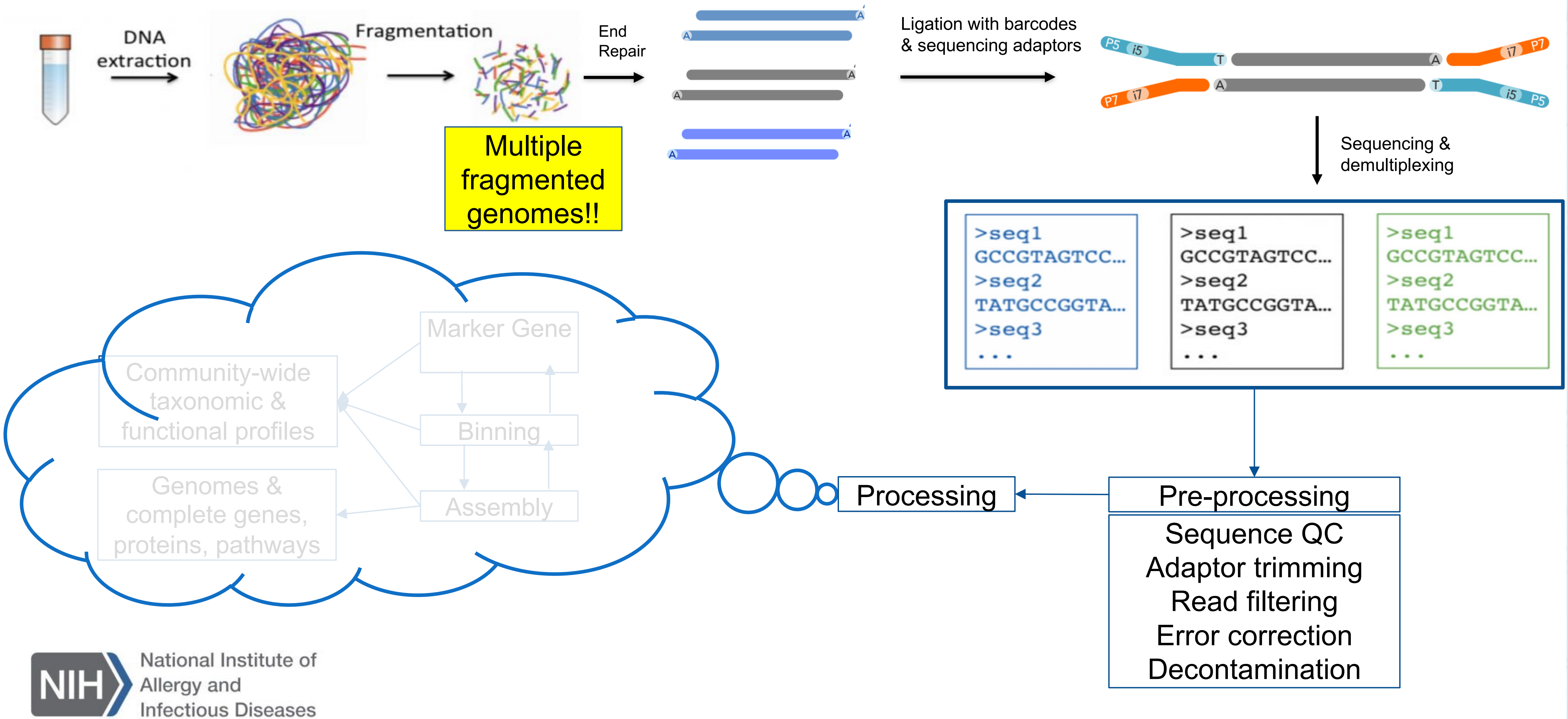
- Contamination from:
  - Organisms with relatively huge genomes (e.g. host DNA)
- Computational requirements
  - Large amounts of data
  - Computationally heavy steps (e.g. assembly)
  - Functional & taxonomic annotations of non-coding sequences
  - Genes with unknown taxonomy & function
- Very complex genes would have limited detectability



National Institute of Allergy and Infectious Diseases

NIAID

Library preparation for shotgun sequencing



DNA extraction
Fragmentation
End Repair
A
Ligation with barcodes & sequencing adaptors
P5
i5
T
A
i7
P7
Sequencing & demultiplexing

Multiple fragmented genomes!!

```
>seq1
GCCGTAGTCC…
>seq2
TATGCCGGTA…
>seq3
. . .
```

```
>seq1
GCCGTAGTCC…
>seq2
TATGCCGGTA…
>seq3
. . .
```

```
>seq1
GCCGTAGTCC…
>seq2
TATGCCGGTA…
>seq3
. . .
```

Marker Gene
Community-wide taxonomic & functional profiles
Binning
Assembly
Genomes & complete genes, proteins, pathways

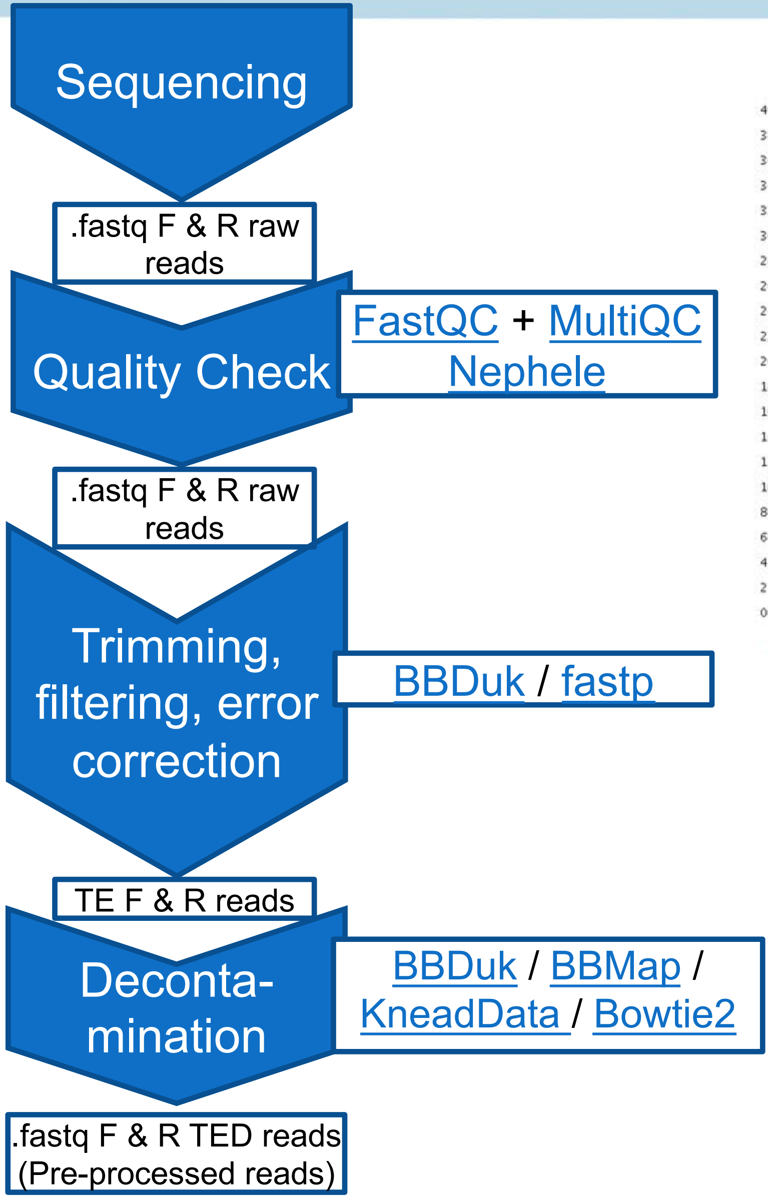
Processing

Pre-processing

Sequence QC
Adaptor trimming
Read filtering
Error correction
Decontamination

National Institute of Allergy and Infectious Diseases

Data pre-processing: QC, trimming & decontamination



Sequencing

.fastq F & R raw reads

Quality Check — FastQC + MultiQC Nephele

.fastq F & R raw reads

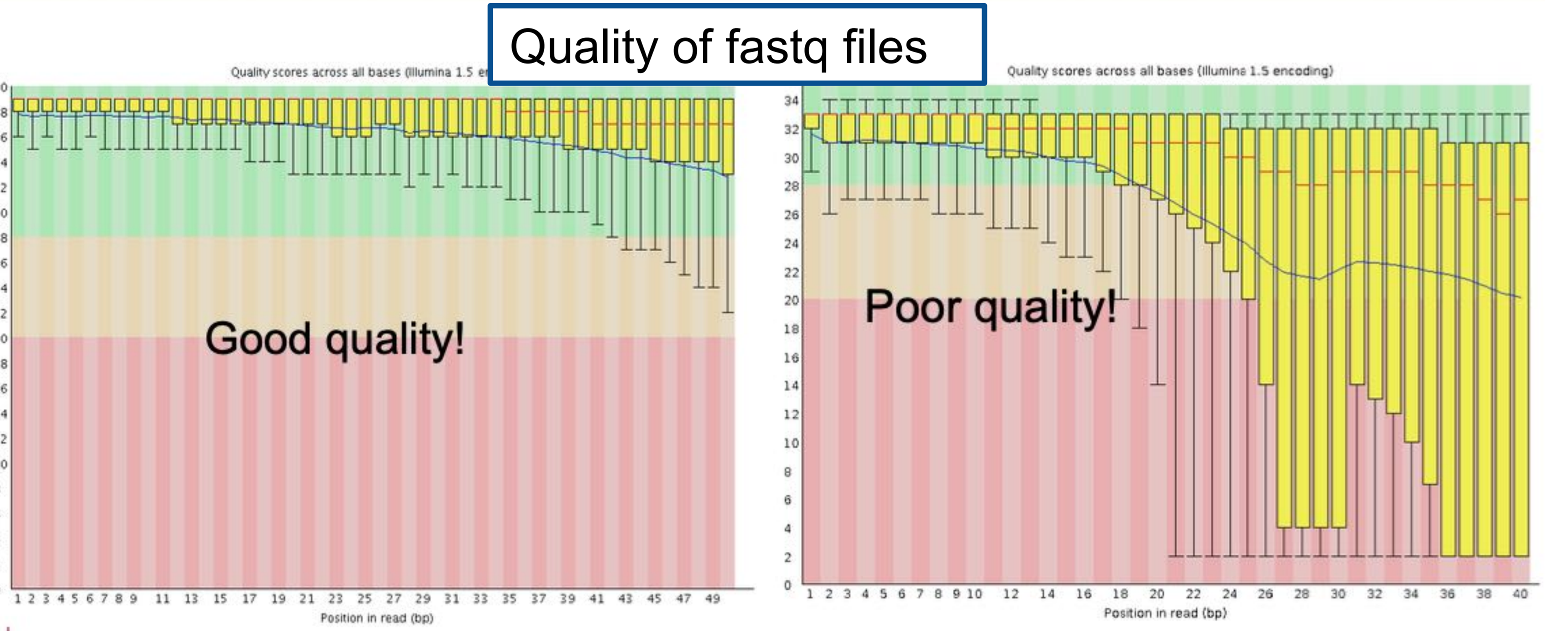
Trimming, filtering, error correction — BBDuk / fastp

TE F & R reads

Deconta-mination — BBDuk / BBMap / KneadData / Bowtie2

.fastq F & R TED reads (Pre-processed reads)

Quality of fastq files



Good quality!

Poor quality!

Adaptor & Quality Trimming & Error correction example:

```
fastp -i  Sample1_R1.fastq.gz -I Sample1_R2.fastq.gz \
    -o Sample1_R1_te.fastq.gz -O Sample1_R2_te.fastq.gz \
    -h fastplog.html -y -c --trim_poly_x -e 10 –w 16 -5 20 -3 15
```

Decontamination example:

```
bbtools bbmap minid=0.95 maxindel=3 bwr=0.16 bw=12 quickmatch fast \
   minhits=2 -Xmx100g  ref=${refHostGenomeDB} \
   in=Sample1_R1_te.fastq.gz    in2=Sample1_R2_te.fastq.gz \
   outu=Sample1_R1_ted.fastq.gz  outu2=Sample1_R2_ted.fastq.gz \
   outm1=Sample1_R1_contam.fq.gz  outm2=Sample1_R2_contam.fq.gz
```

Use reference database of contaminant organism

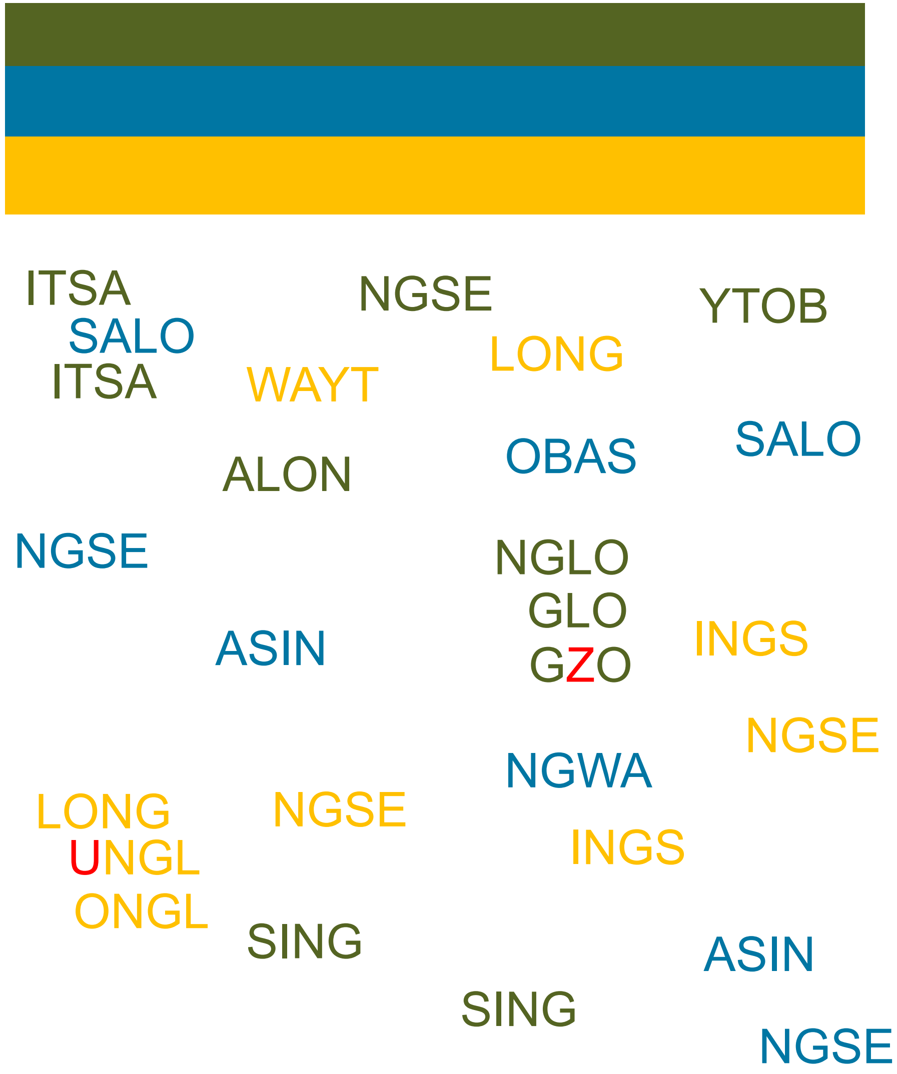
Error correction & decontamination

Error correction:

- *some sequences carry sequencing errors or other sequence-altering artifacts.*
  - *Cause complications during processing (assemblies, mapping, alignments, binning of reads)*
  - Correction algorithms explore the k-mers from the sequences along with their coverage in the dataset and remove highly underrepresented k-mers

Decontamination:

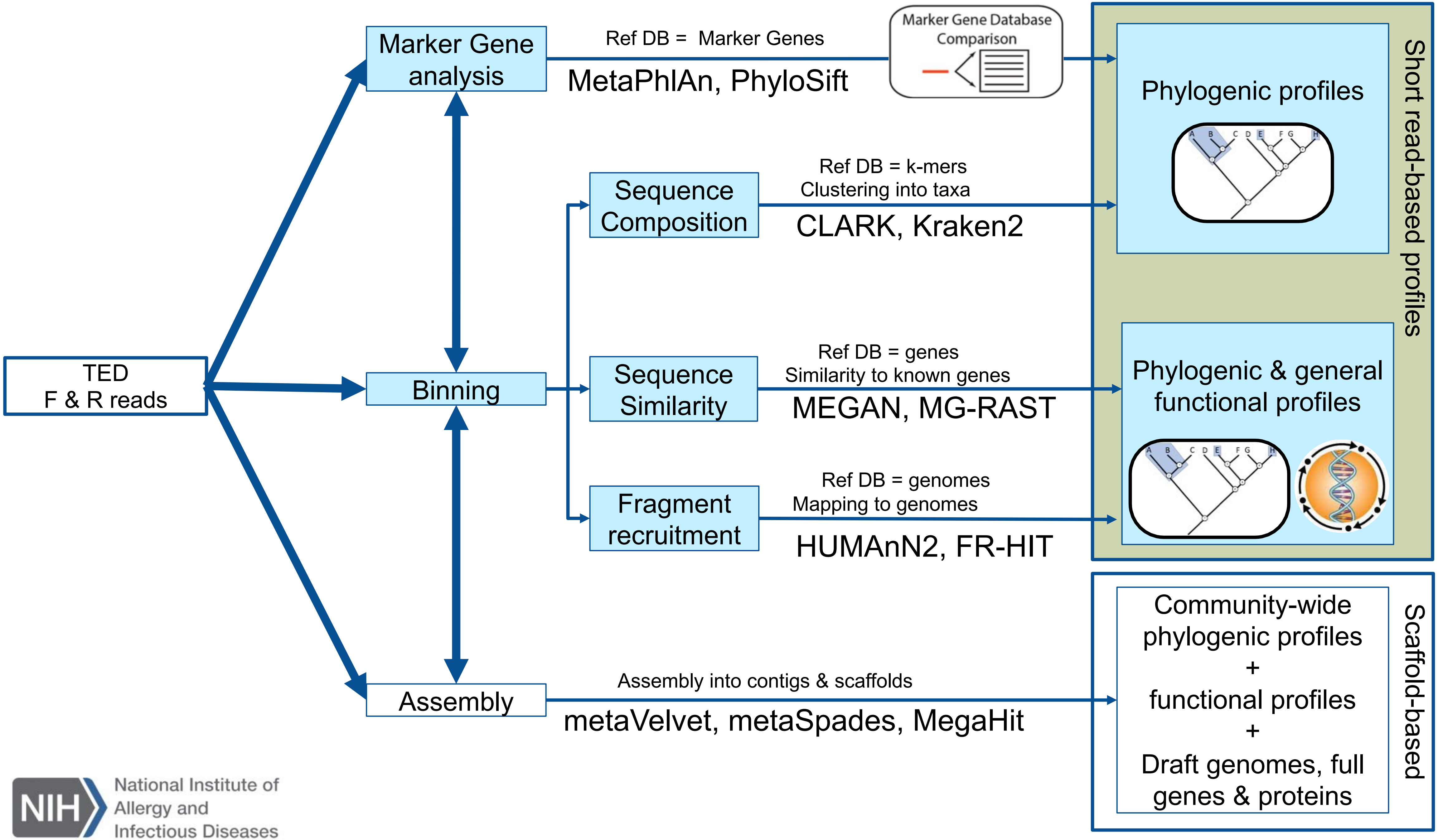
  - Shotgun sequences will contain reads from unwanted genomes (e.g. host DNA, eukaryotic DNA)



https://www.youtube.com/watch?v=OY9Q_rUCGDw

NIH National Institute of Allergy and Infectious Diseases

NIAID



TED
F & R reads

Marker Gene analysis

Ref DB = Marker Genes

MetaPhlAn, PhyloSift

Marker Gene Database Comparison

Binning

Sequence Composition

Ref DB = k-mers
Clustering into taxa

CLARK, Kraken2

Sequence Similarity

Ref DB = genes
Similarity to known genes

MEGAN, MG-RAST

Fragment recruitment

Ref DB = genomes
Mapping to genomes

HUMAnN2, FR-HIT

Assembly

Assembly into contigs & scaffolds

metaVelvet, metaSpades, MegaHit

Short read-based profiles

Phylogenic profiles

Phylogenic & general functional profiles

Scaffold-based

Community-wide phylogenic profiles
+
functional profiles
+
Draft genomes, full genes & proteins

National Institute of Allergy and Infectious Diseases

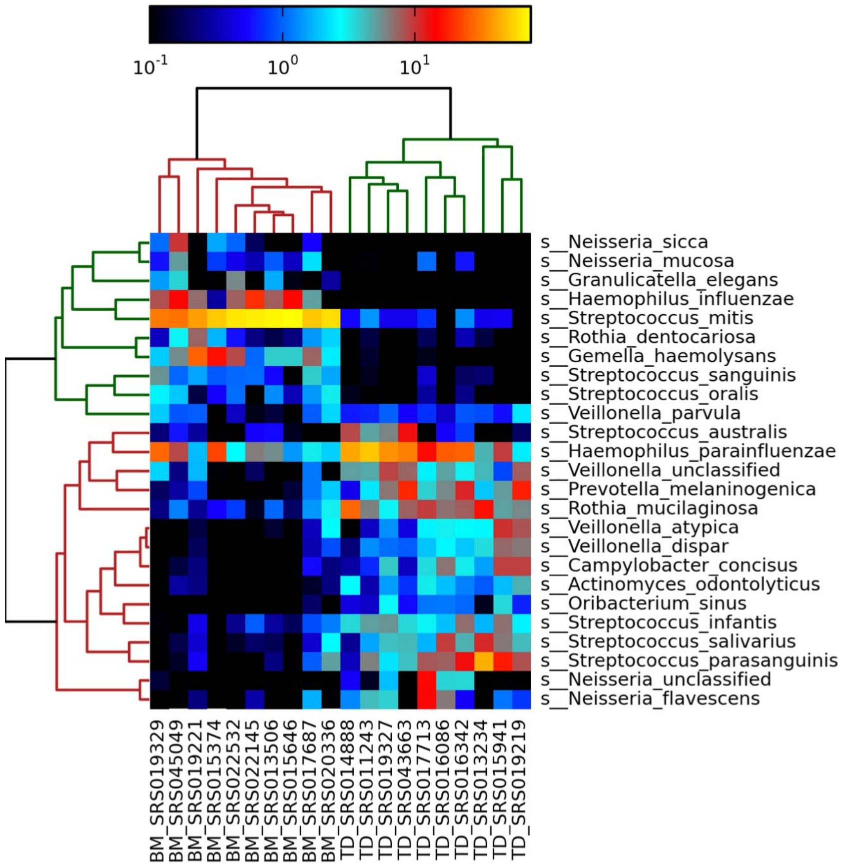
Sharpton 2014. Frontiers in Plant Science. https://doi.org/10.3389/fpls.2014.00209

MetaPhlAn 2.0: Metagenomic Phylogenic Analysis

Short reads Tax ID

Marker gene-based characterization

- Uses bowtie2 to *align* your short shotgun sequences (query) to selected *marker genes* (longer sequences) specific for each clade, identified from ~17K reference genomes from all domains of life. Allows for:
  - Accurate species-level resolution into composition of communities
  - Estimation of organismal relative abundance
- No pre-processing of shotgun reads is required (e.g. error correction, filtering)



```
>metaphlan mergedreads.fasta  --bowtie2db  ~/PATH/to/METAPHLAN_DB \
 --nproc 4  --input_type  fasta  >  Sample_profile.txt

>merge_metaphlan_tables.py  *_profile.txt  >  merged_abund_table.txt

>hclust2.py  --ftop 10  --fdend_width 8  --min 0.1  -l  \
 --in  merged_abund_table.txt  --out merge_abund_heatmap.png
```

https://github.com/biobakery/biobakery/wiki/metaphlan3#create-taxonomic-profiles
Nicola Segata et al. 2012. **Nature Methods**, 8, 811–814

Kraken2: taxonomic sequence classifier

Short reads Tax ID

K-mer based binning

K-mer: a string of sequence with chosen length *k representing sections of a longer sequence*

- Fastest tool for highly accurate binning
- Taxonomic characterization (only ☹)
- Uses your short query to map even shorter k-mer sequences obtained from known genomes.
- Can assign taxonomy to the level of the lowest common ancestor (LCA).
- Great for short metagenomic reads (e.g. shotgun reads)
- Uses standard and custom DBs
- Can be memory demanding depending on DB size

```
>kraken2   --db  KRAKEN_DB   --threads  16   --paired    --out-fmt    paired \
    --fastq-input  Sample1_R1.fastq.gz      Sample1_R2.fastq.gz   \
   --gzip-compressed     --output      kraken_out/Sample1_kraken.txt \
   --report      kraken_out/Sample1_krakenREPORT.txt
```

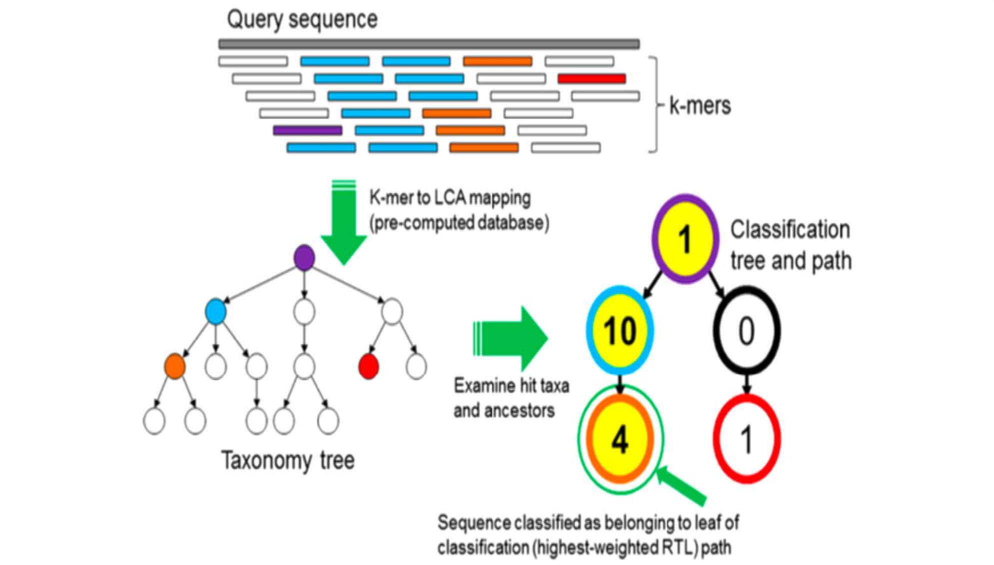


Figure 1 The Kraken sequence classification algorithm. To classify a sequence, each *k*-mer in the sequence is mapped to the lowest common ancestor (LCA) of the genomes that contain that *k*-mer in a database. The taxa associated with the sequence's *k*-mers, as well as the taxa's ancestors, form a pruned subtree of the general taxonomy tree, which is used for classification. In the classification tree, each node has a weight equal to the number of *k*-mers in the sequence associated with the node's taxon. Each root-to-leaf (RTL) path in the classification tree is scored by adding all weights in the path, and the maximal RTL path in the classification tree is the classification path (nodes highlighted in yellow). The leaf of this classification path (the orange, leftmost leaf in the classification tree) is the classification used for the query sequence.

NIH National Institute of Allergy and Infectious Diseases

Wood and Salzberg Genome Biology 2014, 15:R46 http://genomebiology.com/2014/15/3/R46

NIAD

MEGAN6-CE

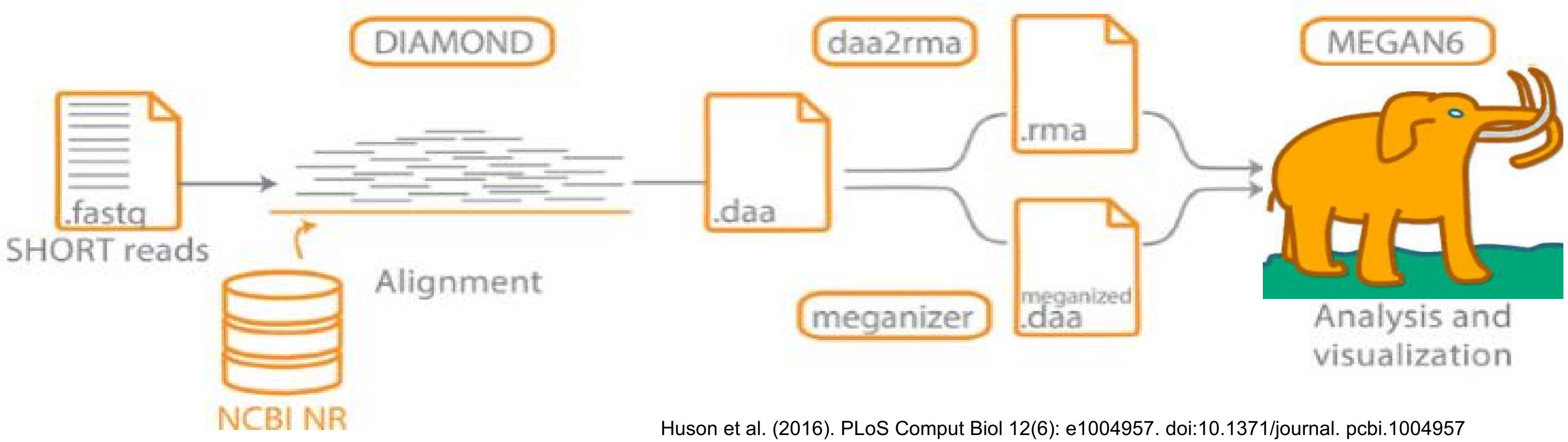
Short reads Tax & Fnc ID

Similarity-based binning

- Uses BLAST or DIAMOND to align shotgun reads to NCBI NR or NT database of proteins
- Allows for
  - **Taxonomic & *functional*** classification of reads to the level of the Lowest Common Ancestor (LCA).
  - Use of Graphical User Interface (GUI) to manipulate, visualize and analyze shotgun data
  - Community characterization & visualization (e.g. alpha & beta diversity, profile plots, networks)
  - Laptop analysis of large metagenomic datasets
  - Analysis of both short and long reads!

```
bbtools   bbmerge.sh   threads=4 \
trimq=15   qtrim=rl   minlength=40 \
in=Sample1_R1.fastq.gz \
in2=Sample1_R2.fastq.gz \
out=Sample1_merged.fasta

diamond blastx --threads 4   \
-d $DMND_ncbiNR_db  \
-q  Sample1_merged.fasta \
-o Sample1.daa     -f 100
```

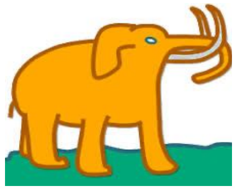


Huson et al. (2016). PLoS Comput Biol 12(6): e1004957. doi:10.1371/journal. pcbi.1004957

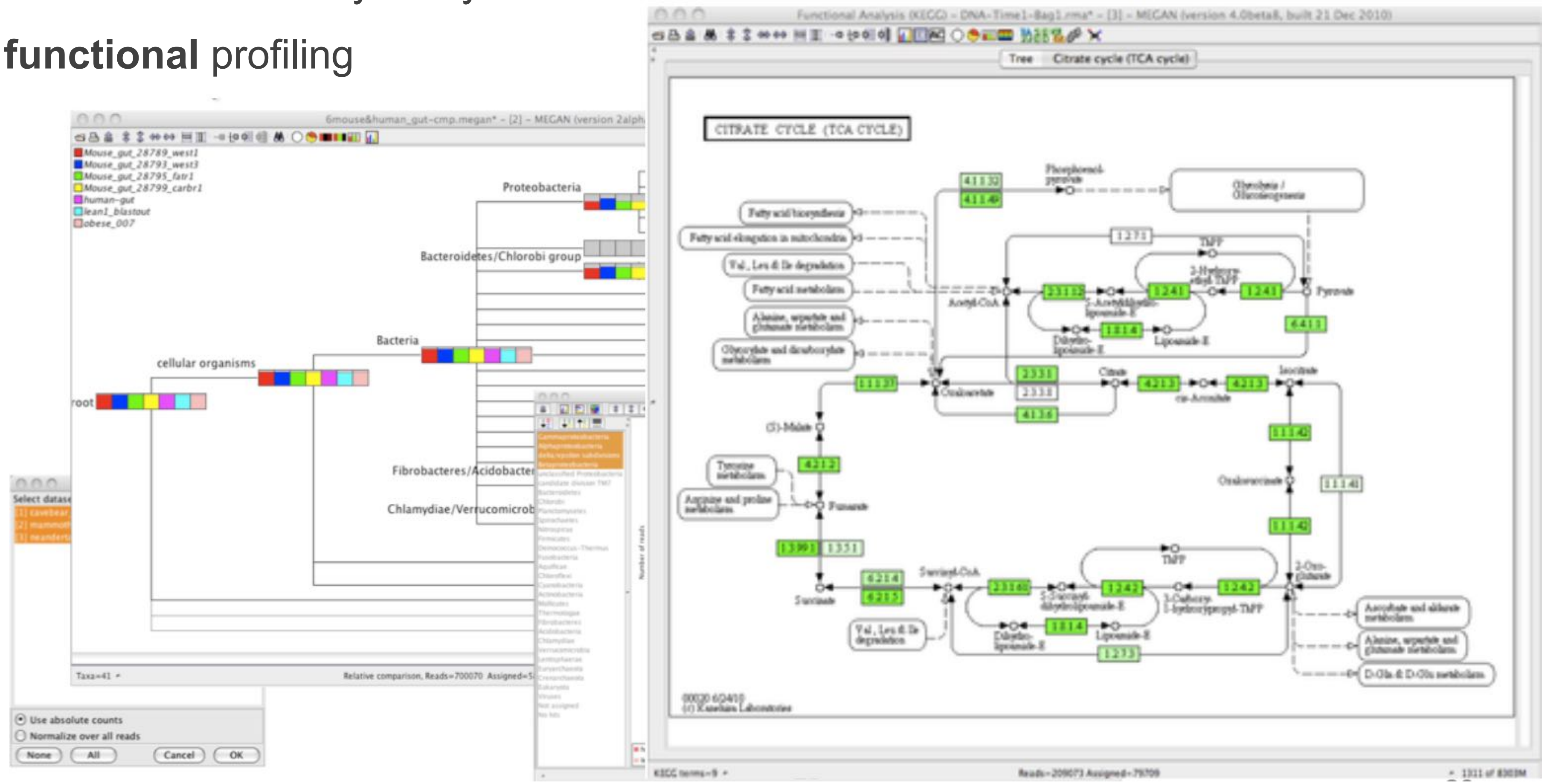
MEGAN6-CE

Short reads Tax & Fnc ID

Similarity-based binning



- Integrates many additional databases (InterPro & GO, eggNOG, KEGG, and SEED)
- Incorporation of metadata for community analysis
- Both **community & functional** profiling
- **Gene-centric assembly**
- User friendly, interactive interface
- Freely available



National Institute of Allergy and Infectious Diseases

NIAID

Huson et al. (2016). PLoS Comput Biol 12(6): e1004957. doi:10.1371/journal. pcbi.1004957

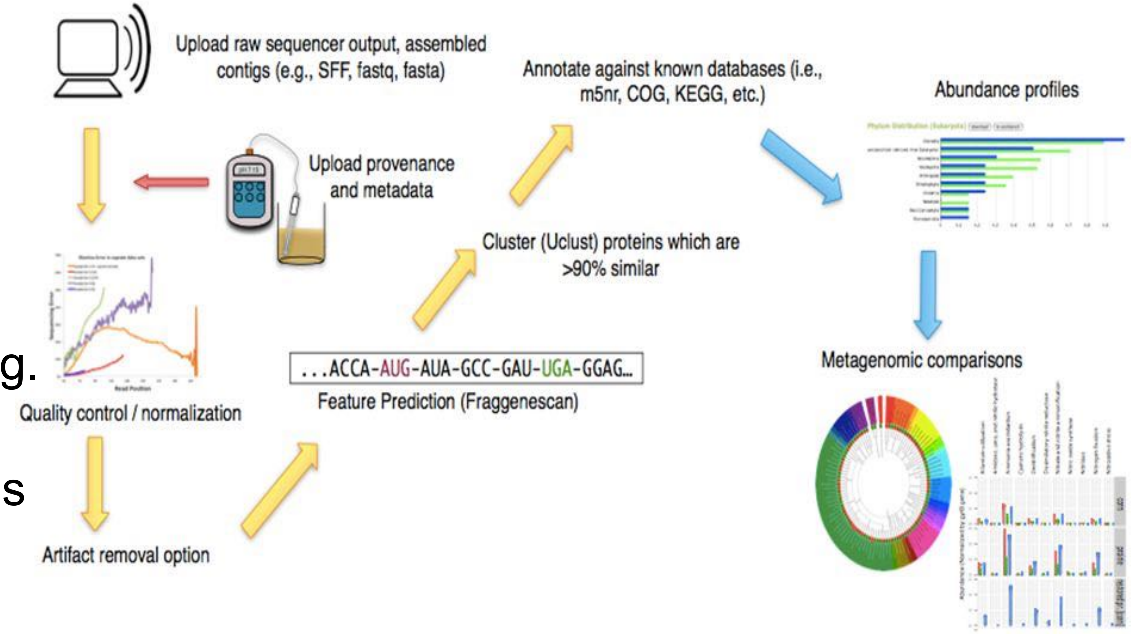
MG-RAST: Metagenomics Analysis Server Rapid Annotations Subsystem Technology

Similarity-based binning

Short reads Tax & Fnc ID

- Entire pipeline of analysis
- Web-based
  - No software installation required
  - No command line use requirements
  - Upload of data required
- Annotation and analysis of metagenomic sequence data (both amplicon & shotgun)
  - Assessment of sequence quality
  - Sequence annotation with multiple databases (e.g. KEGG, GO, NCBI, SEED, UniPort, eggNOG)
  - Post-annotation analyses & visualization pipelines
- Repository for >150K datasets (>23K are publicly available)

MG-RAST
metagenomics analysis server
version 4.0.3
cite us


Upload raw sequencer output, assembled contigs (e.g., SFF, fastq, fasta)

Upload provenance and metadata

Quality control / normalization

Artifact removal option

...ACCA-AUG-AUA-GCC-GAU-UGA-GGAG...

Feature Prediction (Fraggenescan)

Cluster (Uclust) proteins which are >90% similar

Annotate against known databases (i.e., m5nr, COG, KEGG, etc.)

Abundance profiles

Metagenomic comparisons

NIH National Institute of Allergy and Infectious Diseases

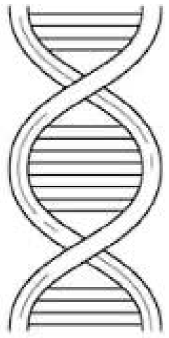
NIAID

Aziz et al. 2008. *BMC Genomics* **9,** 75 (2008). https://doi.org/10.1186/1471-2164-9-75
Cold Spring Harb Protoc. 2010 Jan;2010(1):pdb.prot5368. doi: 10.1101/pdb.prot5368.
https://help.mg-rast.org/user_manual.html

Gene calling on short reads (unassembled)

Short reads Fnc ID

- Accurate, fast & computationally lenient strategy to predict ORFs in **short** reads.
- A lot of genes will only partially present & can missed!
- Traditional gene callers will not work well on short read metagenomic data
- Specialized software use heuristic models of known genes (characteristic-based method), to assign short reads to a functional category.



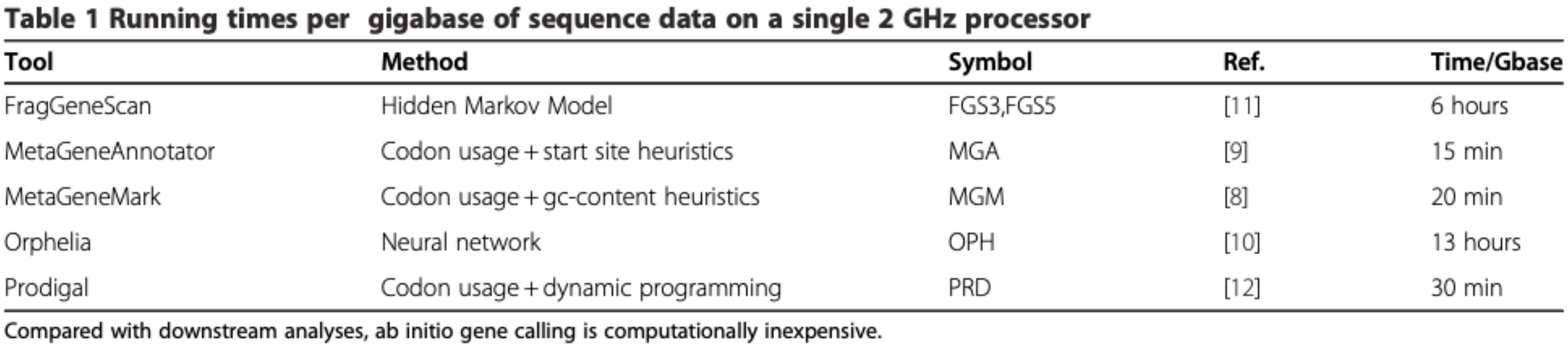


Table 1 Running times per gigabase of sequence data on a single 2 GHz processor

Tool	Method	Symbol	Ref.	Time/Gbase
FragGeneScan	Hidden Markov Model	FGS3,FGS5	[11]	6 hours
MetaGeneAnnotator	Codon usage + start site heuristics	MGA	[9]	15 min
MetaGeneMark	Codon usage + gc-content heuristics	MGM	[8]	20 min
Orphelia	Neural network	OPH	[10]	13 hours
Prodigal	Codon usage + dynamic programming	PRD	[12]	30 min

Compared with downstream analyses, ab initio gene calling is computationally inexpensive.

NIH National Institute of Allergy and Infectious Diseases

Trimble *et al.* *BMC Bioinformatics* **13,** 183 (2012). https://doi.org/10.1186/1471-2105-13-183

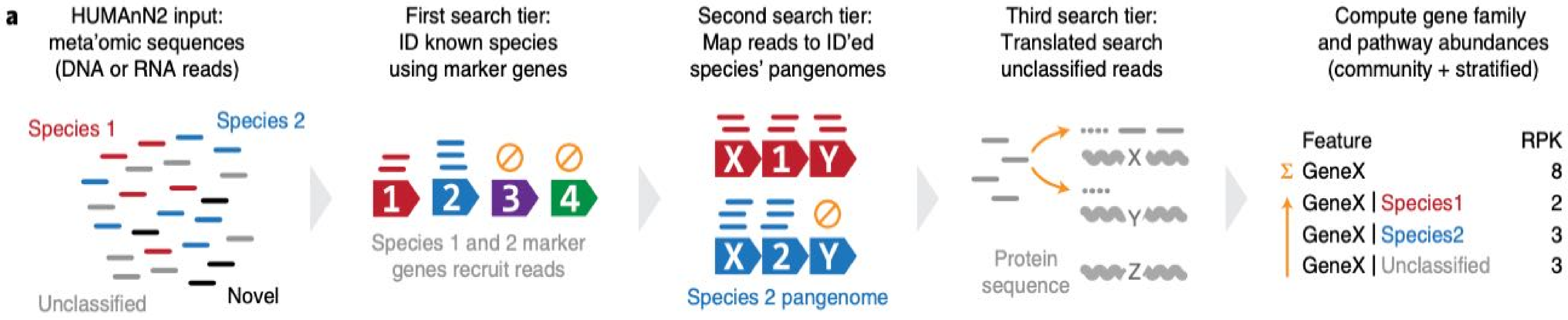
NIAID

HUMAnN 2: HMP Unified Metabolic Analysis Network

Short reads Tax & Fnc ID

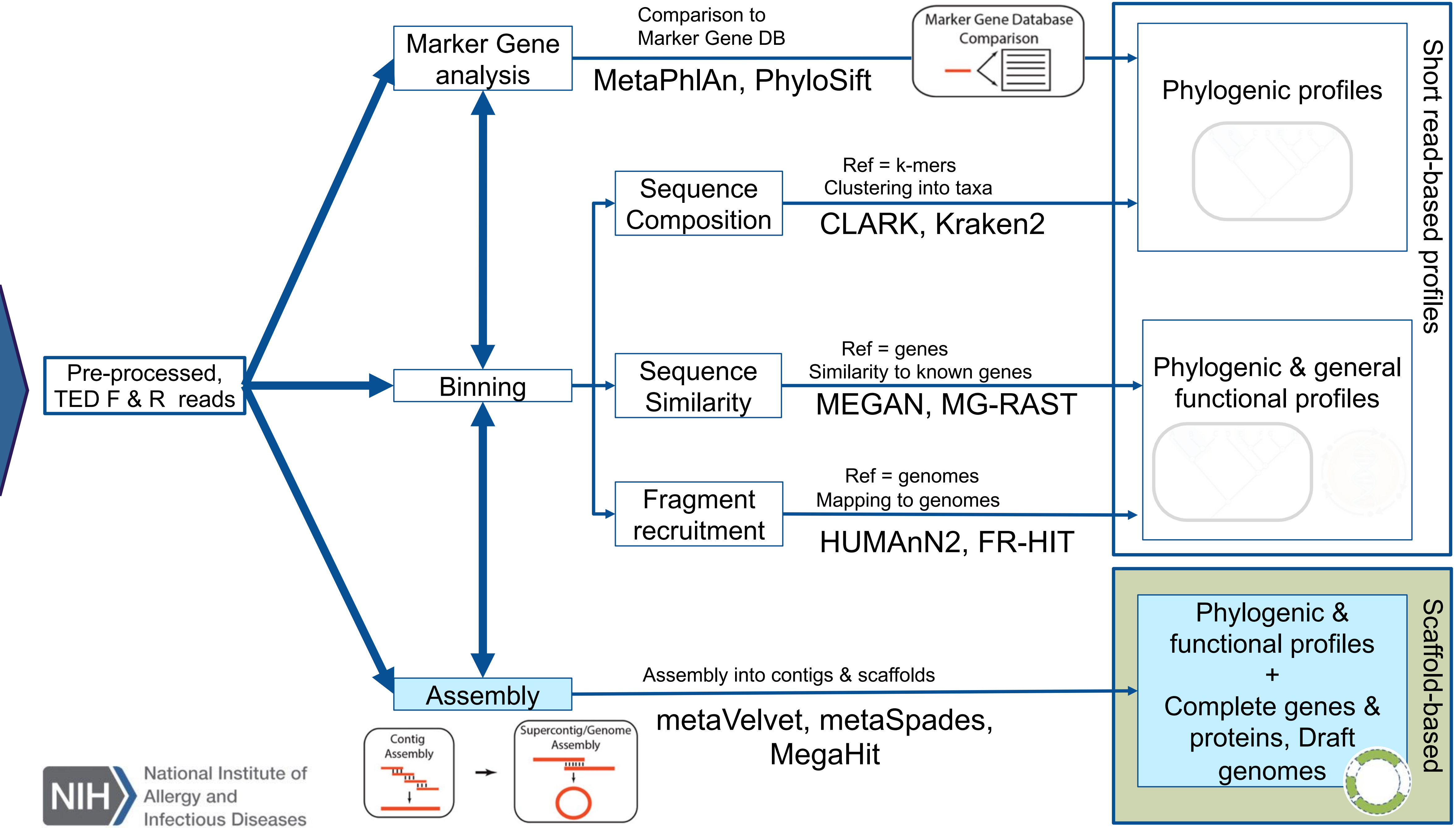
Fragment recruitment

- Entire pipeline of analysis
- Although called “human” the tool is **appropriate for microbiomes of any source**, not just human or host-associated microbiomes.
- Uses short SG reads to identify known microbial species (MetaPhlan2), then:
- Maps all reads to genes sourced from those recognized reference genomes
- Organizes recognized functional genes into pathways based on MetaCyc DB (DIAMOND)
- Determines presence & abundance of each pathway


a
HUMAnN2 input: meta'omic sequences (DNA or RNA reads)
Species 1
Species 2
Unclassified
Novel
First search tier: ID known species using marker genes
1
2
3
4
Species 1 and 2 marker genes recruit reads
Second search tier: Map reads to ID'ed species' pangenomes
X 1 Y
X 2 Y
Species 2 pangenome
Third search tier: Translated search unclassified reads
X
Y
Z
Protein sequence
Compute gene family and pathway abundances (community + stratified)
Feature
RPK
Σ GeneX 8
GeneX | Species1 2
GeneX | Species2 3
GeneX | Unclassified 3

National Institute of Allergy and Infectious Diseases

Franzosa 2018. https://doi.org/10.1038/s41592-018-0176-y
https://huttenhower.sph.harvard.edu/humann



Pre-processed, TED F & R reads

Marker Gene analysis

Comparison to Marker Gene DB

MetaPhlAn, PhyloSift

Binning

Sequence Composition

Ref = k-mers
Clustering into taxa

CLARK, Kraken2

Sequence Similarity

Ref = genes
Similarity to known genes

MEGAN, MG-RAST

Fragment recruitment

Ref = genomes
Mapping to genomes

HUMAnN2, FR-HIT

Assembly

Assembly into contigs & scaffolds

metaVelvet, metaSpades, MegaHit

Phylogenic profiles

Phylogenic & general functional profiles

Short read-based profiles

Phylogenic & functional profiles
+
Complete genes & proteins, Draft genomes

Scaffold-based

National Institute of Allergy and Infectious Diseases

Sharpton 2014. Frontiers in Plant Science. https://doi.org/10.3389/fpls.2014.00209

Assembly Strategies

De novo assembly

- Reference-free (very powerful!)
- Assembly of all organisms
- Assembly of **unknown** organisms
- Miss-assemblies: repetitive or homologous regions produce chimeras, or inaccuracies (large insertions / deletions / inversions) in the assembled genomes
- Example tools: metaVelvet, MegaHit, meta-IDBA, metaSpades
- Deepest exploration of your community

Reference-based assembly

- Closed reference -> Reconstructs only genomes closely related to those in DB
- Uses comparisons to reference genomes -> more reliable assemblies
- Strain-focused
- Miss-assemblies: due to genetic differences between reference and sampled genomes
- Examples: Maq, Bowtie, AMOScmp, MIRA

Hybrid assemblies

- Incorporate both reference-based & *de novo* techniques
- Assemblies incorporate short and long read data (e.g. PacBio)

National Institute of Allergy and Infectious Diseases

NIAID

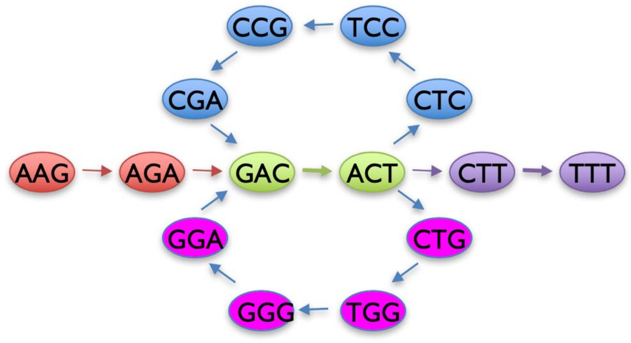
De novo assembly process: De Bruijn graphs

- Uses k-mers to make assembly “possibility” graphs
- Detect and count k-mers out the dataset and tries to build an assembly based on the overlapping of these short sequences
- Best and most commonly used for metagenomics
- Best assemblers: metaVelvet, metaSpades, MegaHit

Reads

AAGA
ACTT
ACTC
ACTG
AGAG
CCGA
CGAC
CTCC
CTGG
CTTT
…

de Bruijn Graph



Potential Genomes

AAGACTCCGACTGGGACTTT

AAGACTGGGACTCCGACTTT

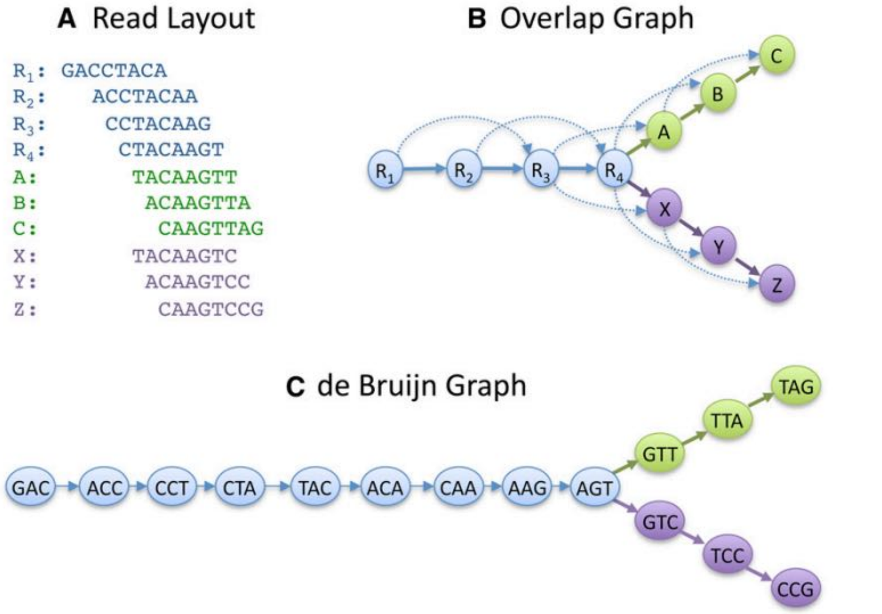


Figure 2. Differences between an overlap graph and a de Bruijn graph for assembly. Based on the set of 10 8-bp reads (*A*), we can build an overlap graph (*B*) in which each read is a node, and overlaps >5 bp are indicated by directed edges. Transitive overlaps, which are implied by other longer overlaps, are shown as dotted edges. In a de Bruin graph (*C*), a node is created for every *k*-mer in all the reads; here the *k*-mer size is 3. Edges are drawn between every pair of successive *k*-mers in a read, where the *k*-mers overlap by *k* − 1 bases. In both approaches, repeat sequences create a fork in the graph. Note here we have only considered the forward orientation of each sequence to simplify the figure.

NIH National Institute of Allergy and Infectious Diseases

https://www.youtube.com/watch?v=OY9Q_rUCGDw
Schatz M C et al. Genome Res. 2010;20:1165-1173
http://sysbio.unl.edu/Teaching/BIOS497897_2014/NGS_Lecture_5_WGA_Assembly_1.pdf

Contigs, Scaffolds, Scaffolding

Contig

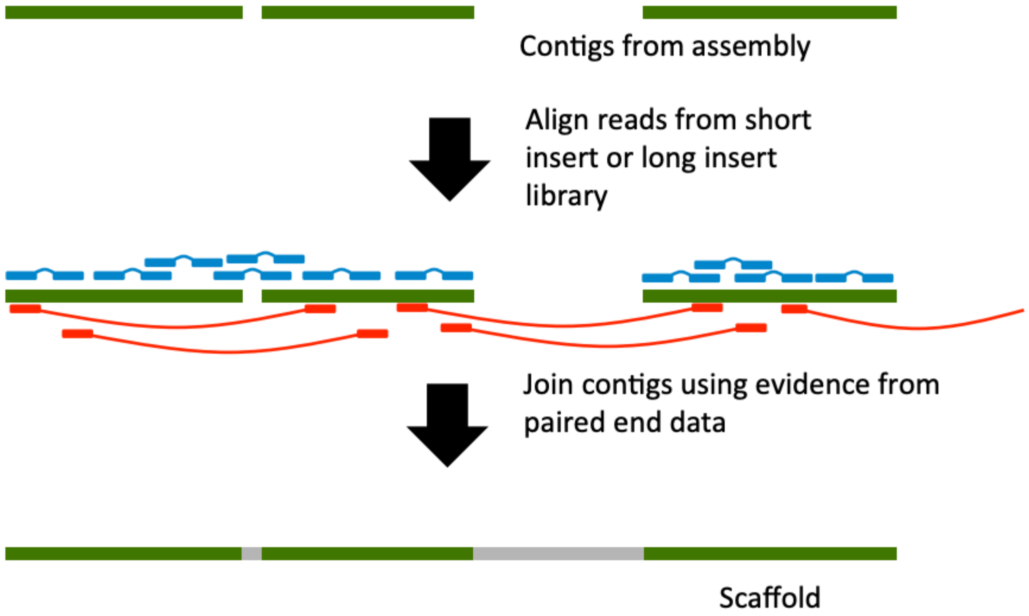
- A contiguous sequence representing the consensus of overlapping sequences (or k-mers), put together during the assembly process
- Often due to missing sequence data, contigs cannot be further extended or connected contiguously.

Scaffolds:

- Due to the paired end-nature of the reads within each contigs, some contigs can be grouped into subsets with known order, orientation and nt distance (scaffolds).

Scaffolding:

- The process of determining contig grouping, order, distance and orientation, by exploiting the PE-nature of the incorporated reads.
  - Metagenomics assembly tools that automatically do scaffolding: metaSPAdes, metaVelvet


Contigs from assembly
Align reads from short insert or long insert library
Join contigs using evidence from paired end data
Scaffold

National Institute of Allergy and Infectious Diseases

Assemble a sentence from tetra-mers

ITSA WAYT LONG

SALO

ALON

NGSE OBAS NGLO

GLO INGS

BASI ASIN YTOB

NGWA

LONG NGSE ONGL

SING

K-mer: a string of sequence with chosen length *k representing sections of the full sequence*

Uncle Iroh’s song (1 genome) is broken up onto k-mers of 4 letters (tetra-mers). Since uncle Iroh performed his song 3 times in the past (3x coverage of the 1 genome), we have k-mers from 3 representations of the lyric.

Can you put together the lyric and discover what Iroh sang?

Help Uncle Iroh recall his favorite song!

National Institute of Allergy and Infectious Diseases

Assembly pipeline

fastp

```
fastp -i  Sample1_R1.fastq.gz -I Sample1_R2.fastq.gz \
    -o Sample1_R1_te.fastq.gz -O Sample1_R2_te.fastq.gz \
    -h fastplog.html -y -c --trim_poly_x -e 10 -w 16 -5 20 -3 15
```

BBTools::
BBmap

```
bbtools bbmap minid=0.95 maxindel=3 bwr=0.16 bw=12 quickmatch fast \
   minhits=2 –Xmx100g  ref=${refHostGenomeDB} \
   in=Sample1_R1_te.fastq.gz    in2=Sample1_R2_te.fastq.gz \
   outu=Sample1_R1_ted.fastq.gz   outu2=Sample1_R2_ted.fastq.gz \
   outm1=Sample1_R1_contam.fq.gz   outm2=Sample1_R2_contam.fq.gz
```

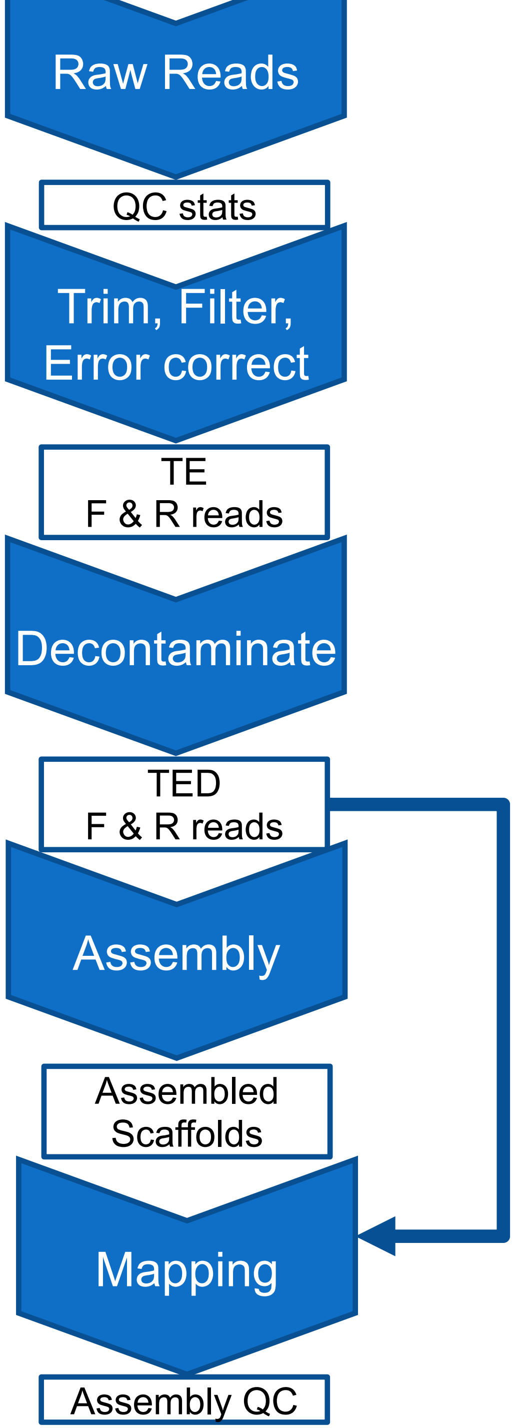
metaVelvet,
metaSpades,
MegaHit

```
metaspades.py       --only-assembler \
        -1    Smp1_R1_ted.fastq.gz        \
        -2    Smp1_R2_ted.fastq.gz        \
        -o    Smp1_assembly    -t    32
```

BWA,
bowtie2,
BBMap

```
>bowtie2-build    Smp1_assembly/scaffolds.fasta    Smp1_assembly/scaffolds.fasa.db

>bowtie2 –sensitive-local    --phred33    -p   42    --no-unal \
       -x   Smp1_assembly/scaffolds.fasta.db     -S    Smp1_assembly.sam \
       -1  Smp1_R1ted.fastq.gz          -2  Smp1_R2ted.fastq.gz
```



Raw Reads

QC stats

Trim, Filter,
Error correct

TE
F & R reads

Decontaminate

TED
F & R reads

Assembly

Assembled
Scaffolds

Mapping

Assembly QC

Assembly QC statistics

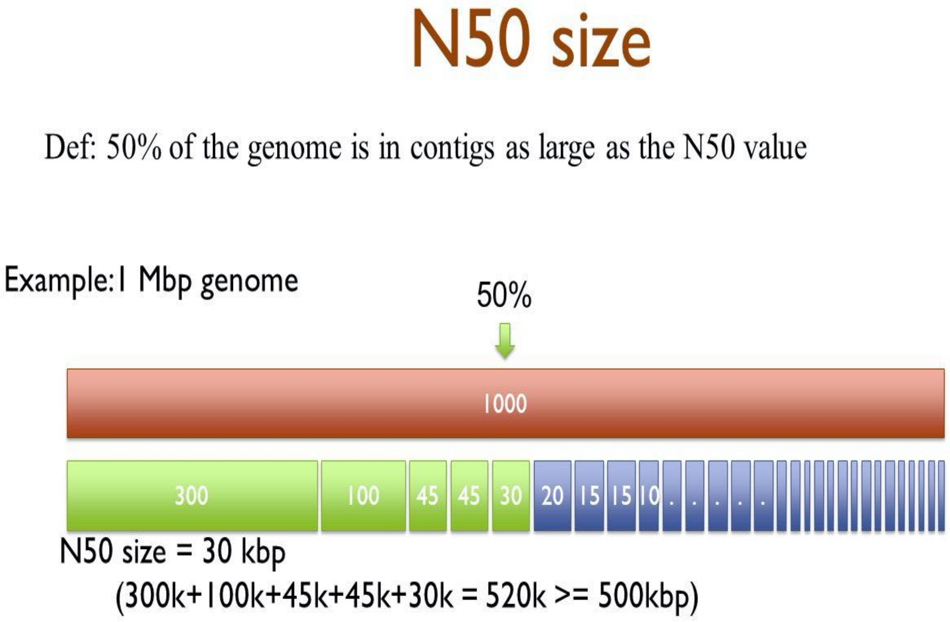
For single organism (genomics):

- Total assembly size (length)
- Number of contigs
- Length of largest contig
- Number of large contigs (e.g. > 50kb)
- Percent reads mapping back to the assembly
- **N50 size**
  - Used to describe the quality of an assembly
  - The length of the shortest contig within the set of largest contigs, comprising at least 50% of the assembly
- **L50**
  - The number of contigs making up 50% of the assembly

For <u>multiple organisms (metagenomics):</u>

- Total assembly size
- Percent reads mapping back to the assembly
- Number of predicted / annotated genes

```
jgi_summarize_bam_contig_depths   Smp1_assembly.bam
-outputDepth     Smp1_assembly_depth.txt
```



N50 size

Def: 50% of the genome is in contigs as large as the N50 value

N50 size = 30 kbp
(300k+100k+45k+45k+30k = 520k >= 500kbp)

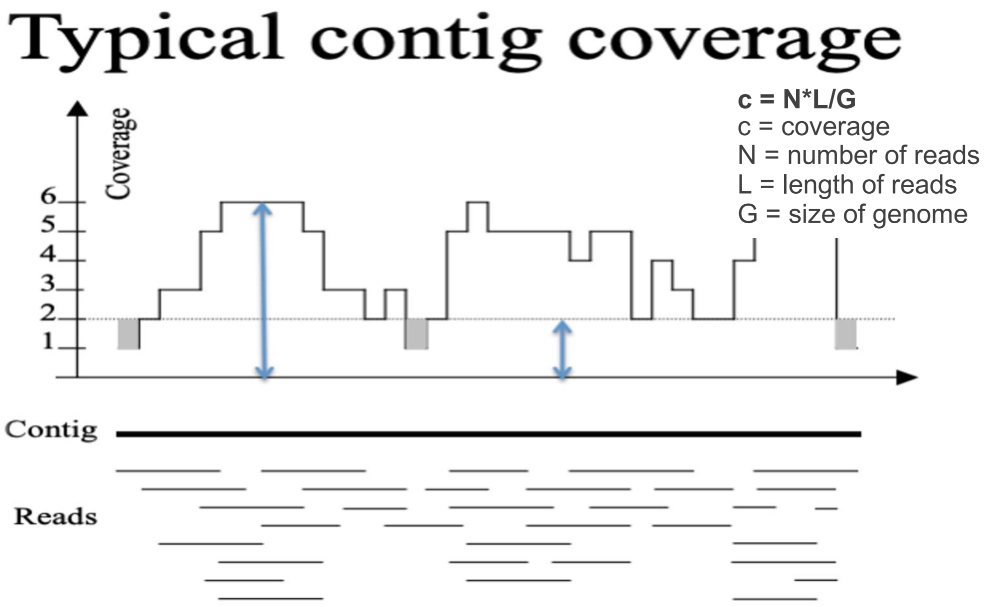
N50 values are only comparable between genomes of same sizes /assemblies of same size!

Rodriguez & Konstantinidies, 2014. ISME. https://doi.org/10.1038/ismej.2014.76
Kunin et a. 2008. Microbiology and Molecular Biology Reviews, Dec. 2008, doi:10.1128/MMBR.00009-08

Metagenomic assembly coverage

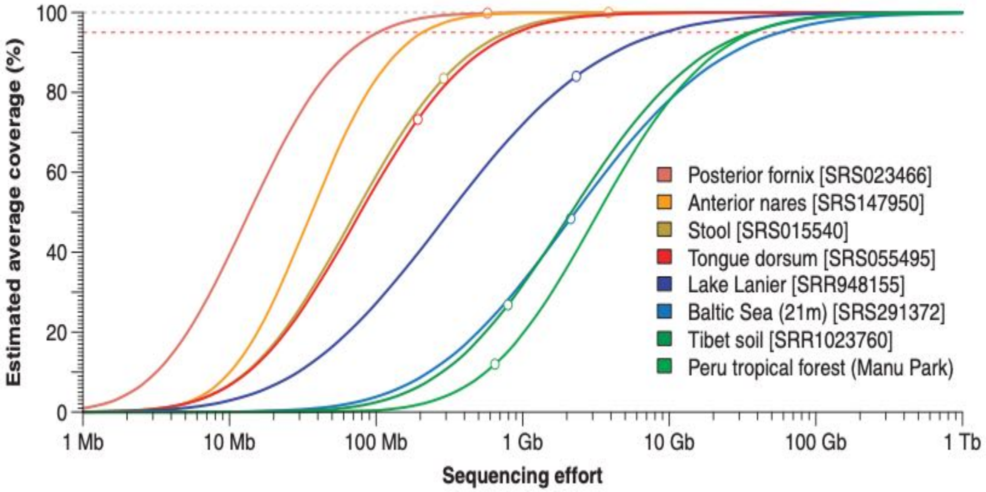
Coverage:

- Hard to predict the sequencing depth (coverage) needed to fully cover all the genomes in a metagenome sample, during sequencing & sufficiently represent all organisms
- Depends community complexity & organismal content
- Post sequencing: Determined by mapping the original processed reads (error-corrected) back to the assembly.
- Sufficient coverage to close a draft genome from a metagenomic dataset is not commonly achieved (complex organismal community).



Typical contig coverage

c = N*L/G
c = coverage
N = number of reads
L = length of reads
G = size of genome



▨	Posterior fornix [SRS023466]
▨	Anterior nares [SRS147950]
▨	Stool [SRS015540]
▨	Tongue dorsum [SRS055495]
▨	Lake Lanier [SRR948155]
▨	Baltic Sea (21m) [SRS291372]
▨	Tibet soil [SRR1023760]
▨	Peru tropical forest (Manu Park)

Rodriguez & Konstantinidies, 2014. ISME.
https://doi.org/10.1038/ismej.2014.76
Kunin et a. 2008. Microbiology and Molecular Biology Reviews, Dec. 2008,
doi:10.1128/MMBR.00009-08

NIH National Institute of Allergy and Infectious Diseases

NIAID

Draft genomes of assembled sequences
(binning)

Long read
Tax ID

An unsupervised method (without the assistance of a database) of clustering scaffolds into taxonomic groups, based on sequence features (GC content, read coverage, etc.) and contig linkage patterns

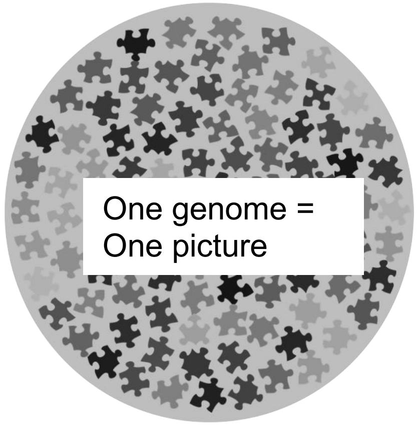
Advantages:

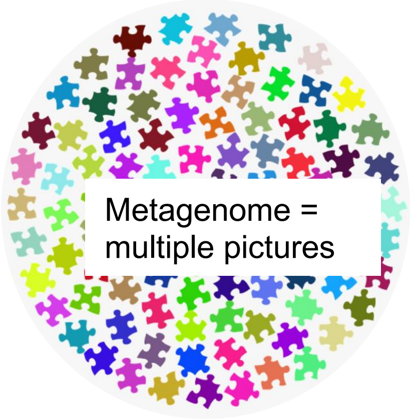
- Improving taxonomic and genomic assignment
- Discover novel taxa (without cultivation)
- Elucidate functional potential of taxa
- Lower risk of false positives

Disadvantages:

- Higher abundance limit for detection
- Inaccuracies with complex communities

- Binning tools: MetaBat, MaxBin, CONCOCT, GroopM





```
metabat2    -i    scaffolds.fasta  \
-o     bins/     --unbinned  -t    8
```

National Institute of
Allergy and
Infectious Diseases

CheckM: Parks. 2015. Genome Res. 2015. 25: doi:10.1101/gr.186072.114

Long read Tax ID

Assessing draft genome quality

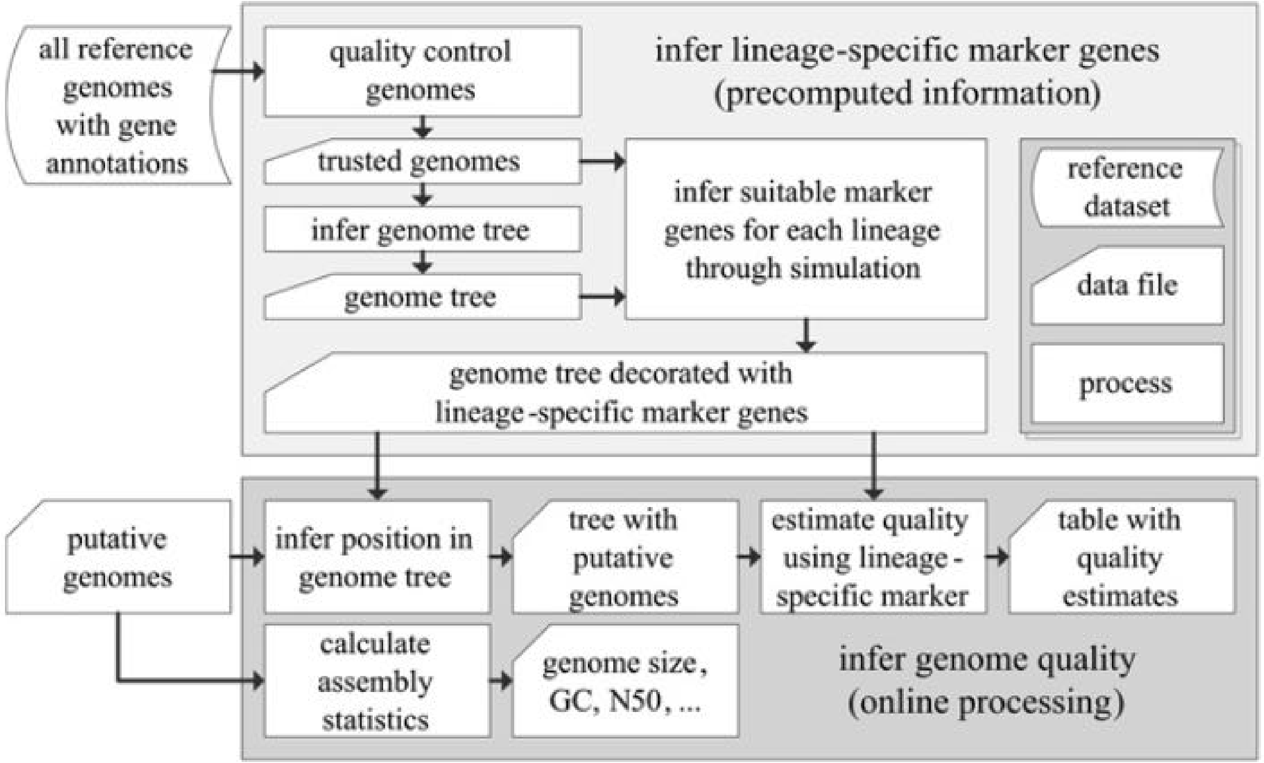
Uses a database with a broad set of *marker genes* with information about their relative position, co-location and distribution throughout their reference genomes, in order to assess characteristics of the draft genomes (bins)

```
>checkm lineage_wf  --pplacer_threads  8  -t 8  --nt  -x   fasta   \
        bins/   --tab_table    checkm_wf/
>checkm qa    -o   2    --tab_table    -f     sum_meta.txt    \
          checkm_wf/lineage.ms    checkm_wf/    -t    4
>checkm tree_qa    -o  2    --tab_table     -f     \
        checkm_wf/tree_qa_results.txt     checkm_wf/
```

Characteristics of the draft genomes (bins)

- **completeness**
- **contamination levels**
- **phylogenic association**

- Allows for **manual bin** exploration
- Allows for **manual bin** curation


all reference genomes with gene annotations
quality control genomes
infer lineage-specific marker genes (precomputed information)
trusted genomes
infer genome tree
genome tree
infer suitable marker genes for each lineage through simulation
reference dataset
data file
process
genome tree decorated with lineage-specific marker genes
putative genomes
infer position in genome tree
tree with putative genomes
estimate quality using lineage-specific marker
table with quality estimates
calculate assembly statistics
genome size, GC, N50, ...
infer genome quality (online processing)

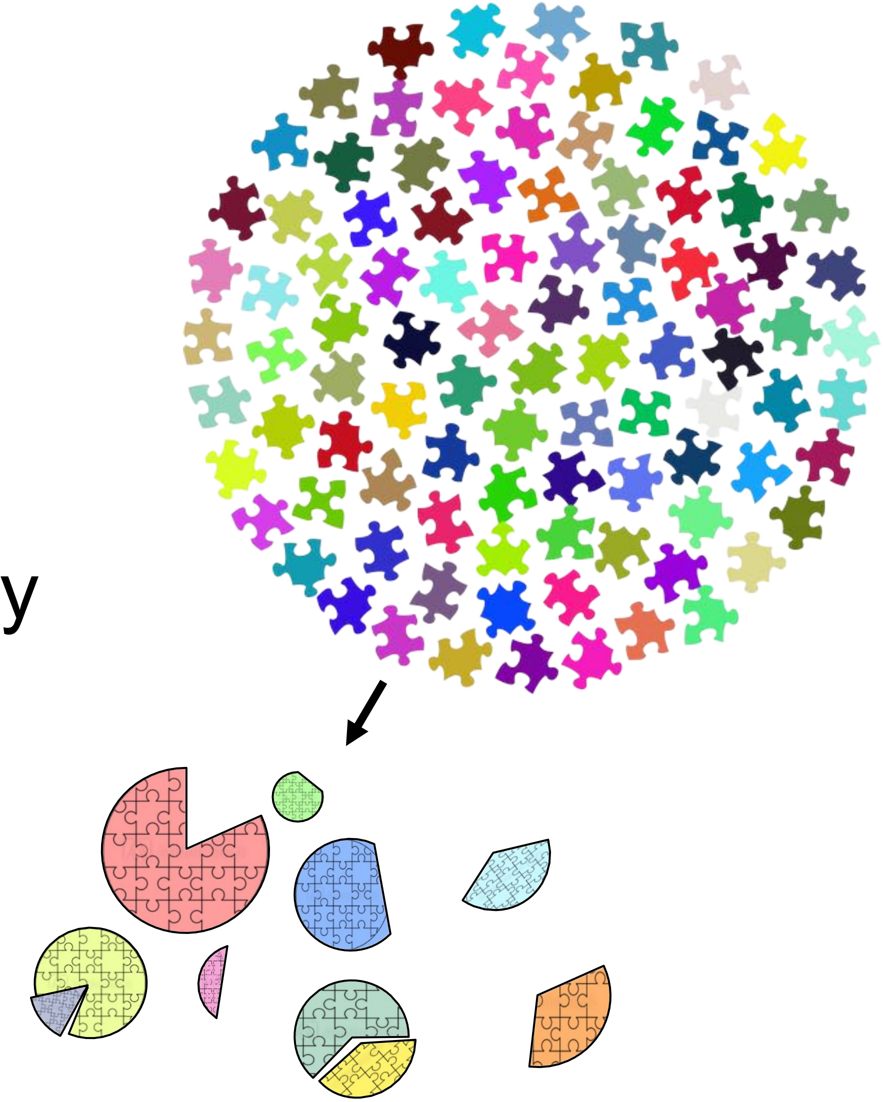
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2848648/
https://ecogenomics.github.io/CheckM/

National Institute of Allergy and Infectious Diseases

NIAID

Challenges in metagenome assembly

- Computationally demanding
- Chimeric assemblies: unrelated genomes may contain similar DNA
- Genomes from same species may harbor genetic differences
- Specialized **meta**genome assembly algorithms needed (single genome assembly algorithms won’t do).



National Institute of Allergy and Infectious Diseases

Gene & functional annotation

Long read Tax & Fnc ID

- Gene predictions algorithms find genes based on different strategies:
  - structural features (e.g. GC content, k-mer content, transcription start/end sites, base occurrence periodicity, etc.)
  - co-location of genes (probabilistic distances for co-location of genes within a (draft) genome / scaffold / long read)
  - masking of non-coding regions (e.g. repeats, junk DNA, TEs)
  - Tools: Prodigal, GeneFragScan, metaEuk, metaErg
- Gene annotation algorithms assign biological relevance to the predicted genes
  - Based on **homology to reference gene databases** of hidden Markov models (HMMs) constructed from empirically explored genes
  - Assigned are gene annotations (gene names, EC numbers, Gene Ontologies, etc.)
  - Tools: metaProkka, InterProScan, GhostKoala, DAVID, EuGene, MG-RAST, Galaxy
- Functional annotation algorithms perform and/or use gene annotations to reconstruct metabolic pathways and predict functional capacity of organism/ communities
  - Tools: MinPath, KEGG Mapper, MG-RAST, Galaxy

https://tinyurl.com/y4orq76q

Metagenomics strategies

Amplicon VS Shotgun

	Amplicon Sequencing	Shotgun Sequencing
Community content & diversity	Yes	Yes
Community dynamics	Yes	Yes
Taxonomic response to factors	Yes	Mostly
Diversity detail	Bacterial / Targeted	Abundant/Larger Genomes
Taxonomic Assignment	Genus level	Strain level
Taxonomic Resolution	Rare species	Abundant-med abund organisms
Core community	Yes	Yes
Taxonomic targeting	Yes (*in situ*)	*In silico*
Functional capacity	Only inferred / Limited	Yes
Introduced biases	PCR & Primers bias	Genome size and complexity
Microbial “dark matter”	Less detectable	Detectable
Variant detection	Only for amplicon	For any gene / region
Computational demand	Less	Rather large
Cost	Cheaper	More expensive

Metadata is just as important as the data itself!

Genomic Standards Consortium http://gensc.org

- ***<u>Metadata is critical</u>*** *to data interpretation & reproducibility.*
- **Metadata Standards** are being implemented by scientific community!
  - to promote standardization of sequence data **and metadata** quality (e.g. ontology, descriptive fields)
  - to promote data discoverability, comparability and reproducibility of studies.

Checklists for Minimum Information about any sequence (MIxS) implement specific requirements for different types of information needed to describe each study and sample (e.g. biome, longitudinal study)

- For (meta)genomic studies: **Minimal Information about a (Meta)Genomic Sequence (MIGS & MIMS)** checklists
- For marker gene studies (e.g. 16S): **Minimal Information about a Marker Sequence (MIMARKS)** checklists

These and other standardization checklists available at: https://gensc.org/mixs/

MIxS

Thank you

Questions

angelina.angelova@nih.gov

bioinformatics@niaid.nih.gov

Acknowledgements

Bioinformatics and Computational Biosciences Branch

Office of Cyber Infrastructure and Computational Biology