Metagenomics

Shinjini Mukherjee
Laboratory of Aquatic Ecology, Evolution and Conservation

Introduction

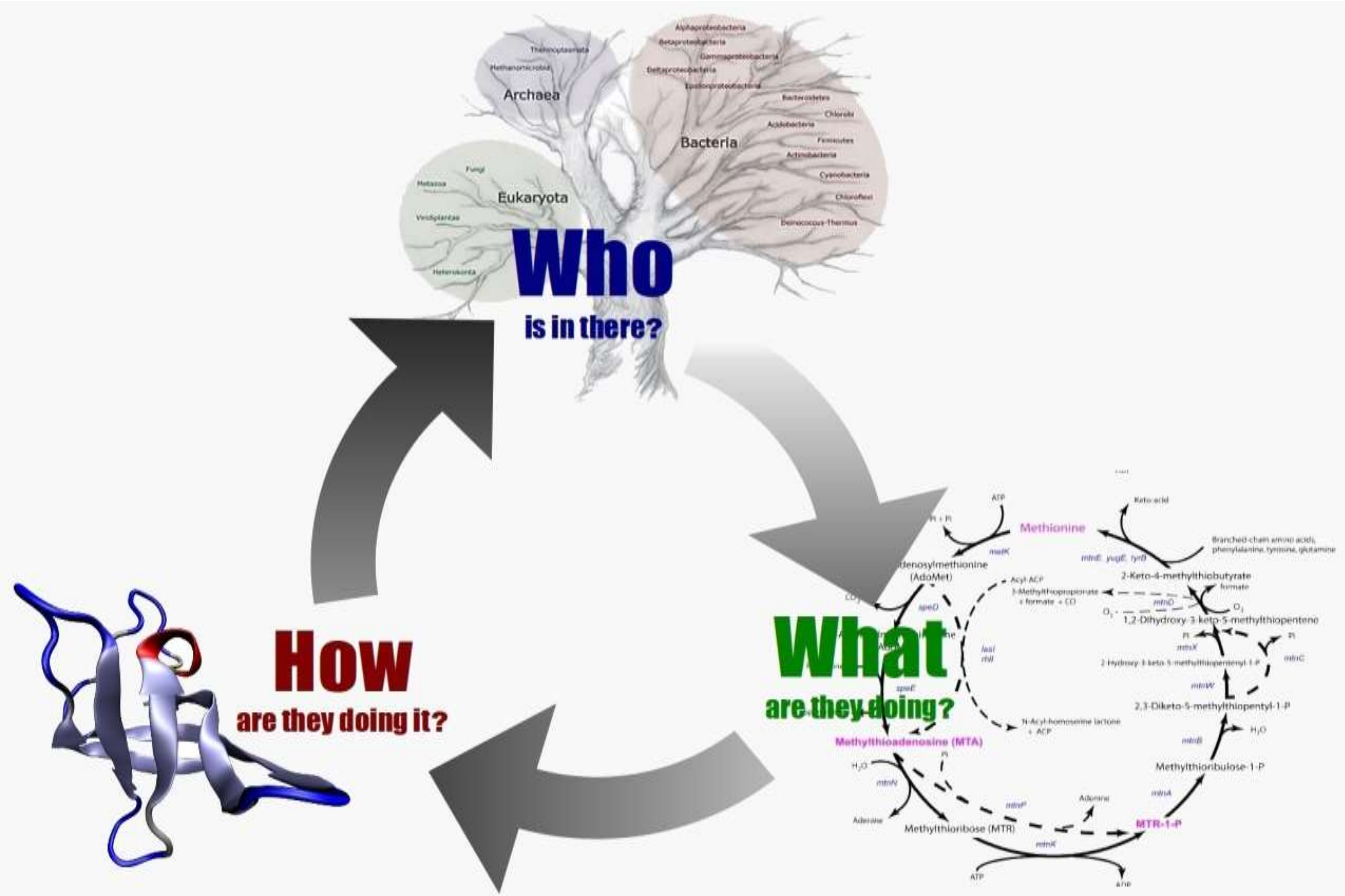
Metagenomics literally means "beyond the genome."

Genome = Parts list of a single species

Metagenome = Parts list of the community

Introduction

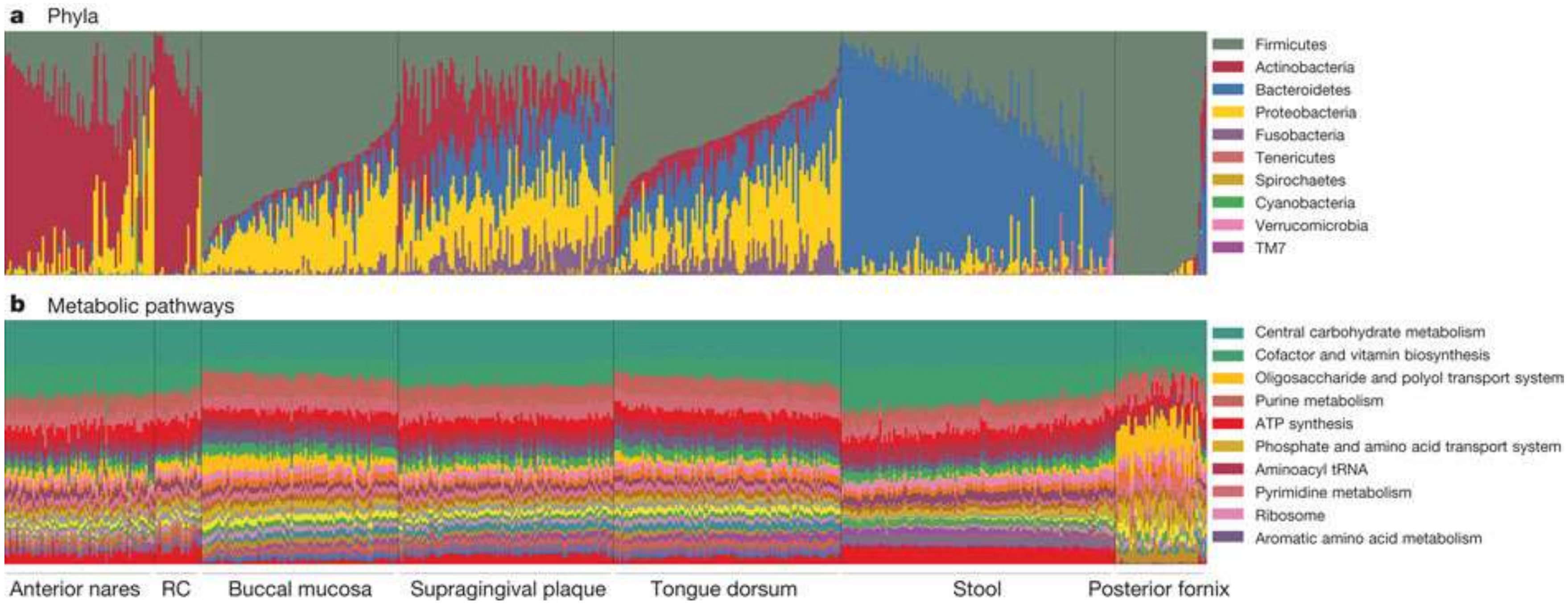
Why metagenomics?



Source: Center for Biological Sequence Analysis, DTU

Introduction

Why metagenomics?



Carriage of microbial taxa varies while metabolic pathways remain stable within a healthy population

C Huttenhower *et al. Nature* **486**, 207-214 (2012)

Introduction

Some fascinating metagenomic investigations

North America
Atlantic Ocean
Pacific Ocean
South America

Global ocean sampling

HMP

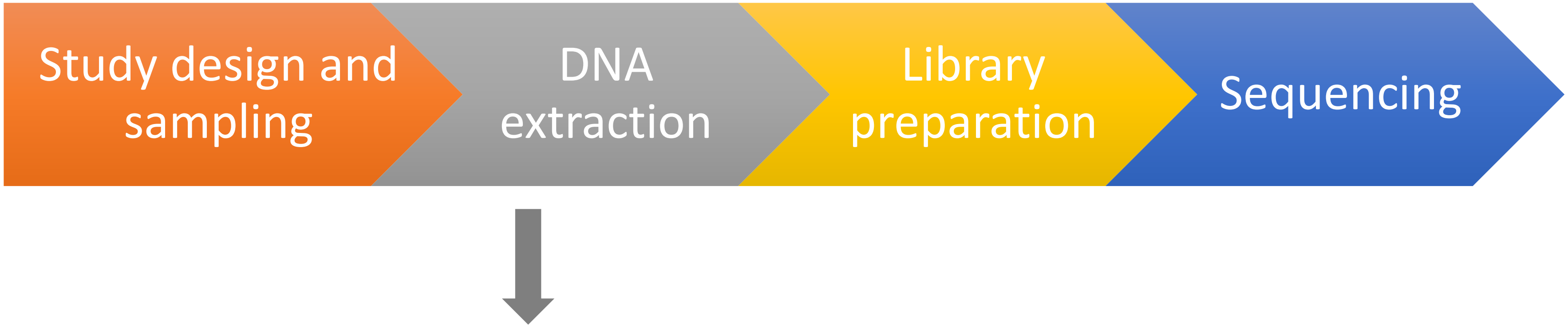
NIH HUMAN MICROBIOME PROJECT

TARA OCEANS

earth microbiome project

I. Shotgun metagenomic sequencing

Processing and sequencing of the samples



Study design and sampling

DNA extraction

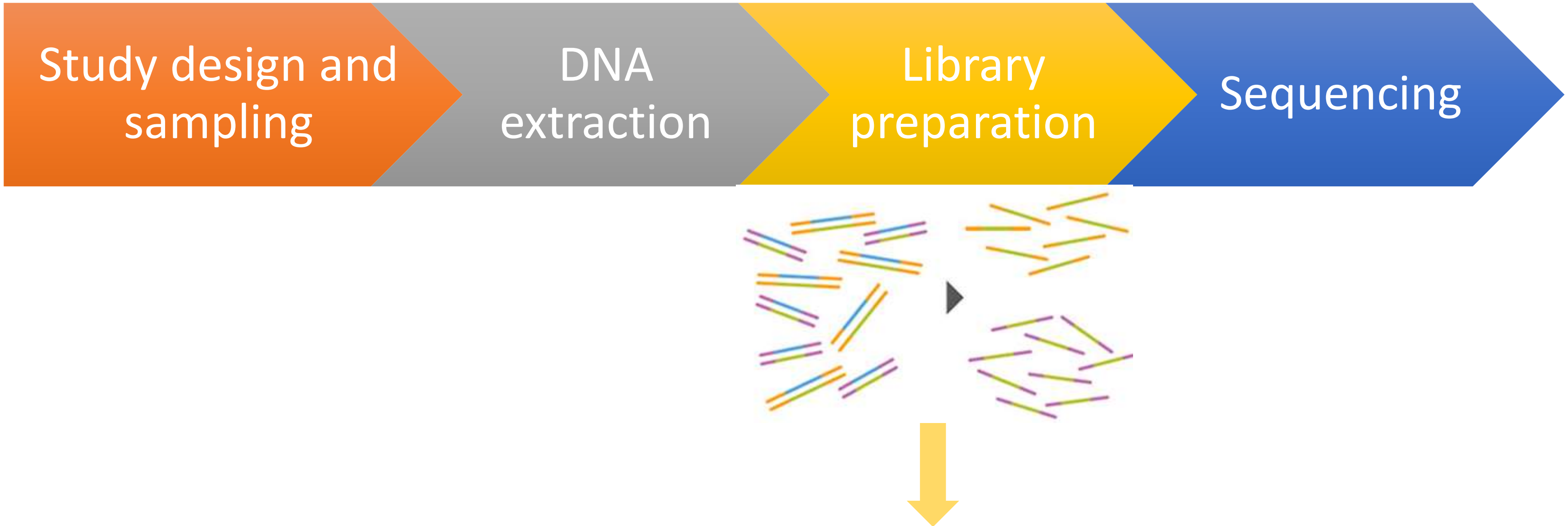
Library preparation

Sequencing

- Extract DNA from biological sample
- For low biomass samples, multiple DNA extractions might be required
- An extra step of enriching prokaryotic DNA might be required

I. Shotgun metagenomic sequencing

Processing and sequencing of the samples



Study design and sampling

DNA extraction

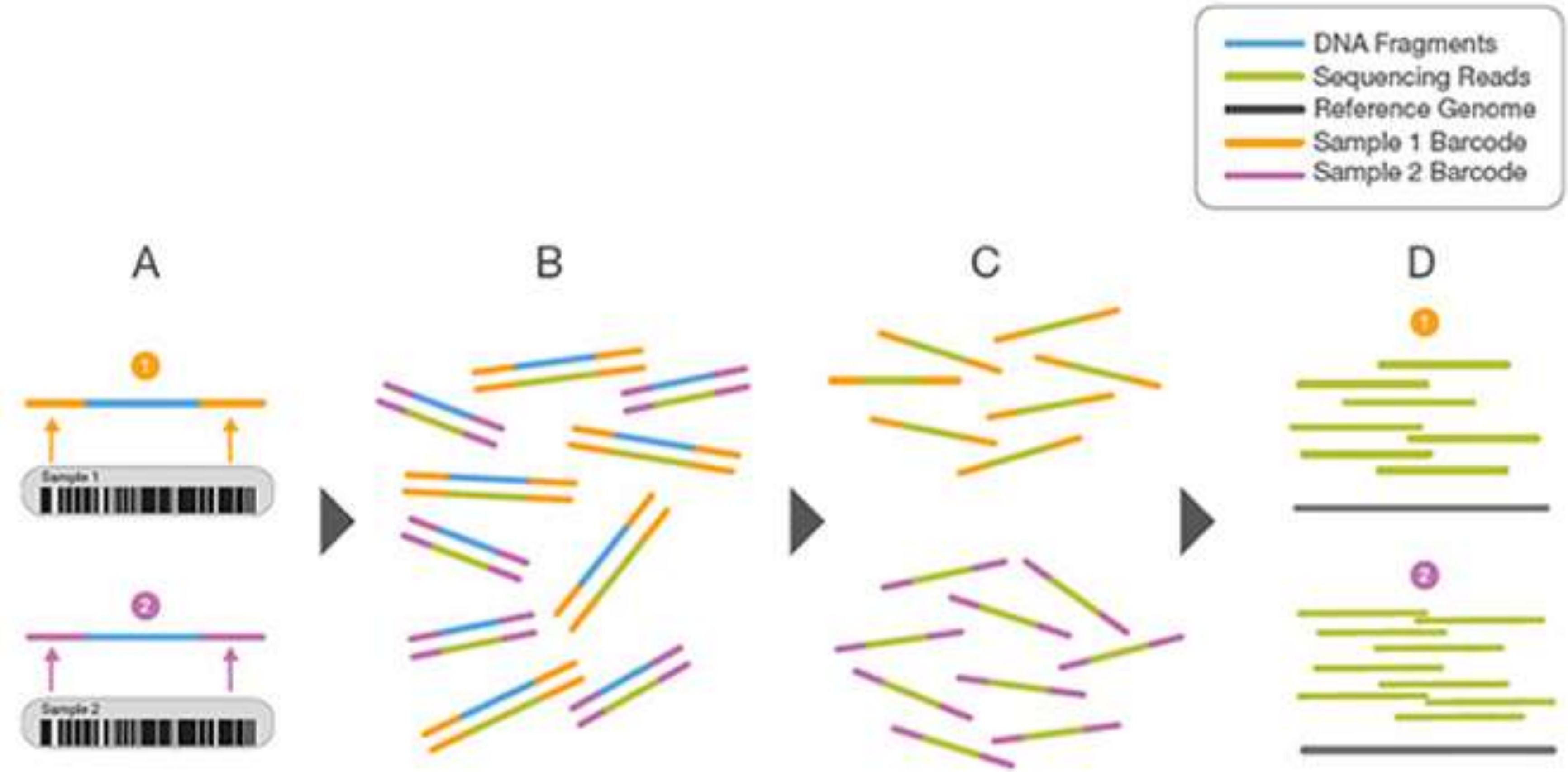
Library preparation

Sequencing

- Equimolar amount of DNA from each sample is used for library preparation
- Multiple methods of library preparation exist, usually differing in the method of fragmentation (e.g., tagmentation, mechanical and enzymatic fragmentation)

I. Shotgun metagenomic sequencing

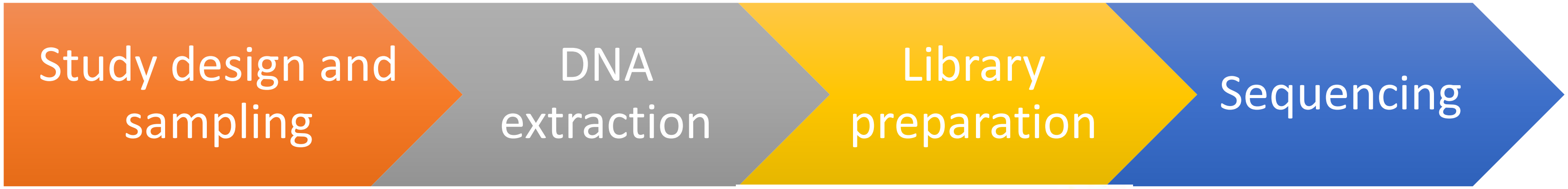
Conceptual overview of multiplexing



Adapted from illumina technology overview

I. Shotgun metagenomic sequencing

Processing and sequencing of the samples

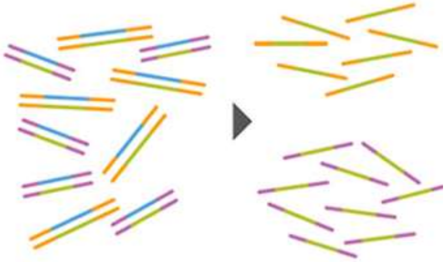


Study design and sampling

DNA extraction

Library preparation

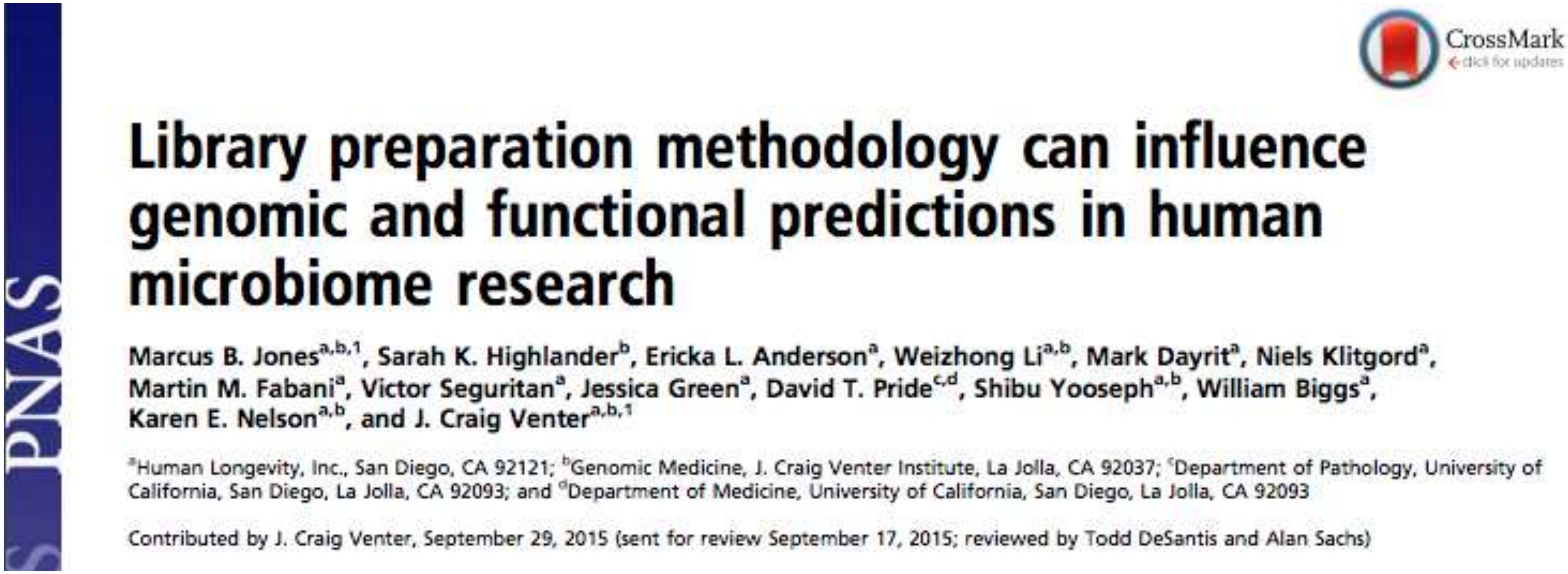
Sequencing



- Some library preparation kits use a PCR step (this can be useful for low biomass samples but can also introduce a bias)
- Illumina Nextera XT and Illumina TruSeq DNA PCR-free kits, and the KAPA Biosystems Hyper Prep PCR and PCR-free systems are among the most commonly used library preparation kits.

I. Shotgun metagenomic sequencing

A word of caution!



PNAS

CrossMark
click for updates

Library preparation methodology can influence genomic and functional predictions in human microbiome research

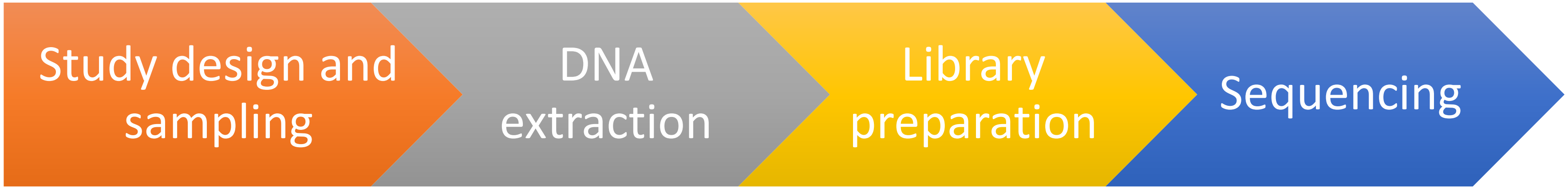
Marcus B. Jones[a,b,1], Sarah K. Highlander[b], Ericka L. Anderson[a], Weizhong Li[a,b], Mark Dayrit[a], Niels Klitgord[a], Martin M. Fabani[a], Victor Seguritan[a], Jessica Green[a], David T. Pride[c,d], Shibu Yooseph[a,b], William Biggs[a], Karen E. Nelson[a,b], and J. Craig Venter[a,b,1]

[a]Human Longevity, Inc., San Diego, CA 92121; [b]Genomic Medicine, J. Craig Venter Institute, La Jolla, CA 92037; [c]Department of Pathology, University of California, San Diego, La Jolla, CA 92093; and [d]Department of Medicine, University of California, San Diego, La Jolla, CA 92093

Contributed by J. Craig Venter, September 29, 2015 (sent for review September 17, 2015; reviewed by Todd DeSantis and Alan Sachs)

I. Shotgun metagenomic sequencing

Processing and sequencing of the samples


Study design and sampling

DNA extraction

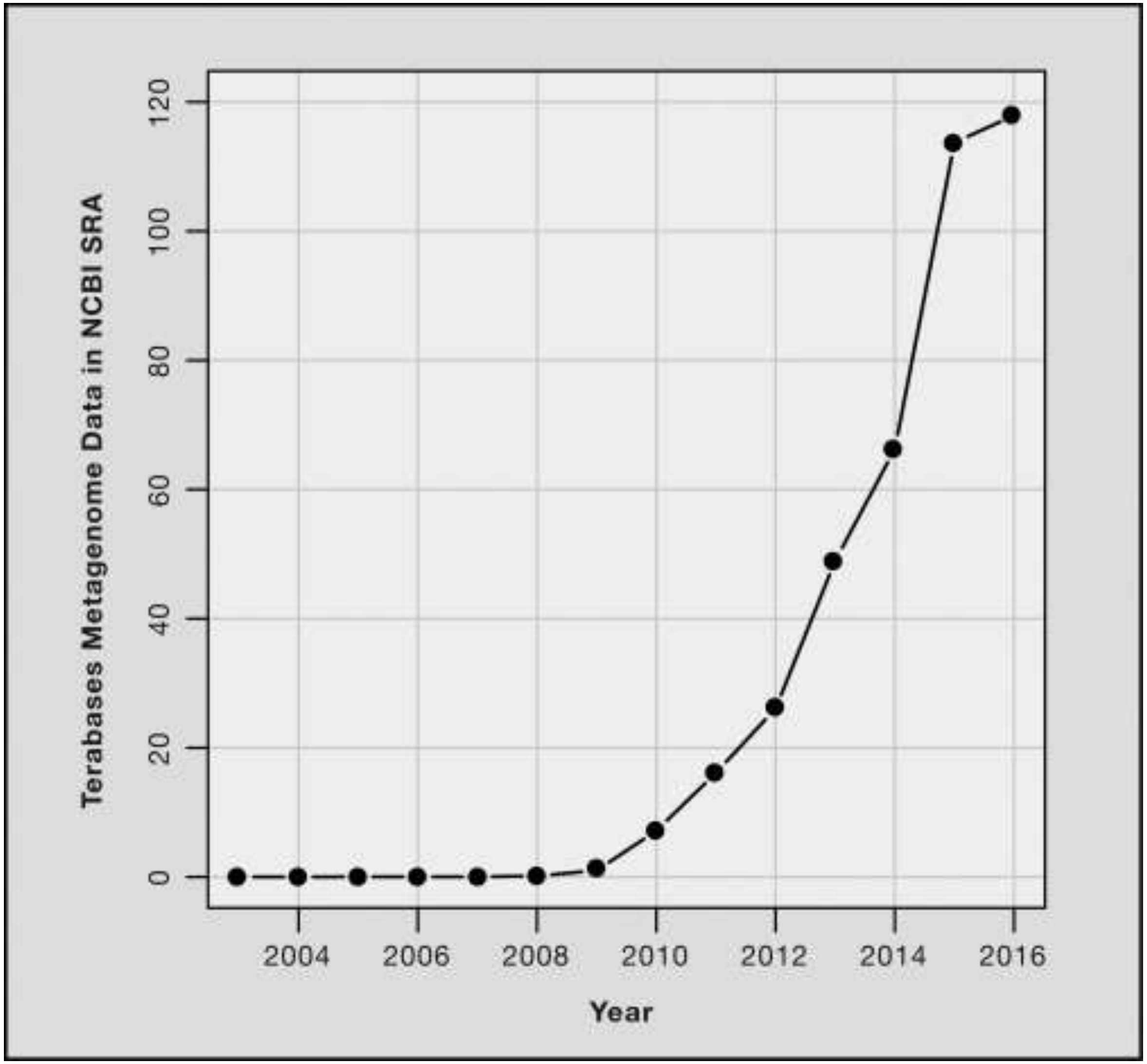
Library preparation

Sequencing

- Amount of “coverage” depends on your scientific question and study system
- Illumina HiSeq 2500 , HiSeq 4000, NextSeq and NovaSeq are recommended as they produce high volumes of sequence data (between 120 Gb and 1.5 Tb per run)

II. Metagenomic data analyses

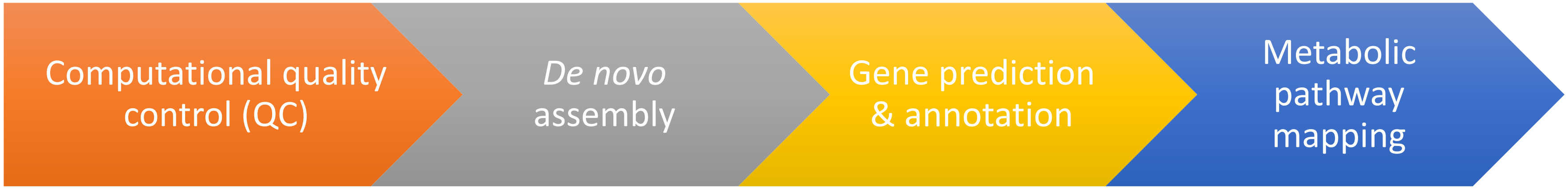
The metagenomic data deluge!



Nayfach and Pollard, Cell 2016

II. Metagenomic data analysis

Overview of steps involved in metagenomic data analysis

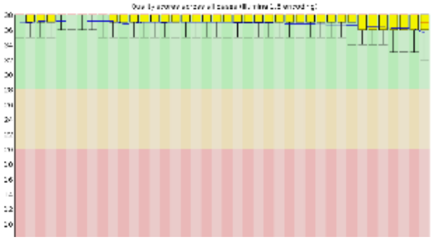


Computational quality control (QC)

De novo assembly

Gene prediction & annotation

Metabolic pathway mapping



- **Quality trimming**: remove bad quality sequences, adapter trimming

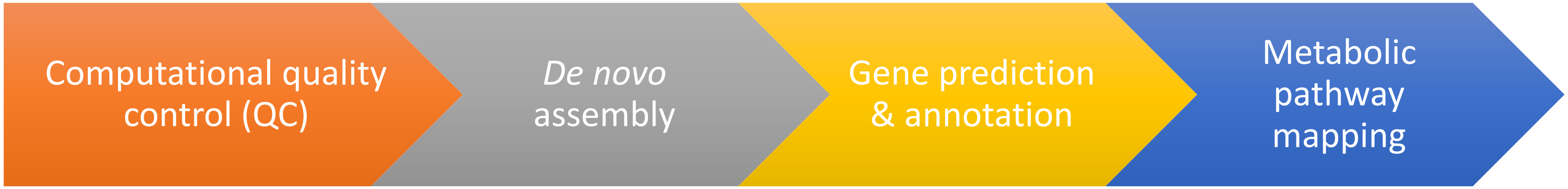
Tools: Trimmomatic, FASTX-Toolkit, cutadapt , sickle, scythe, Picard Tools

- **Removal of non-target DNA:** identifying and removing eukaryotic contamination from microbial metagenomes

Tool: DeconSeq

II. Metagenomic data analysis

Overview of steps involved in metagenomic data analysis

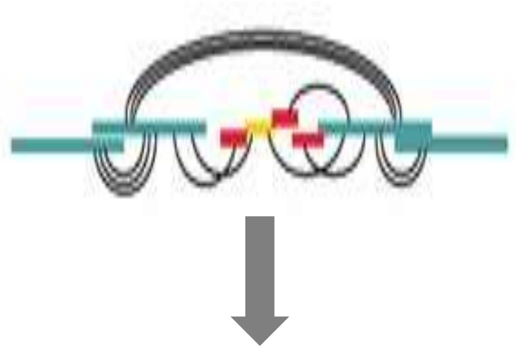


Computational quality control (QC)

De novo assembly

Gene prediction & annotation

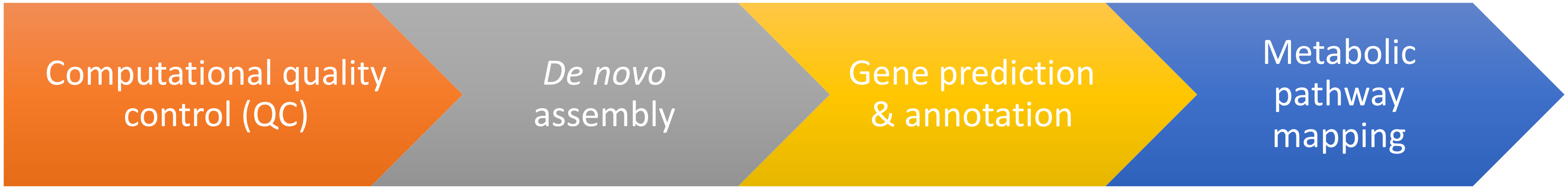
Metabolic pathway mapping



- **Assembly:** merging collinear metagenomic reads from the same genome into a single contiguous sequence (i.e., **contig**)
- ***De novo* assembly:** in the absence of reference genome(s)

II. Metagenomic data analysis

Overview of steps involved in metagenomic data analysis

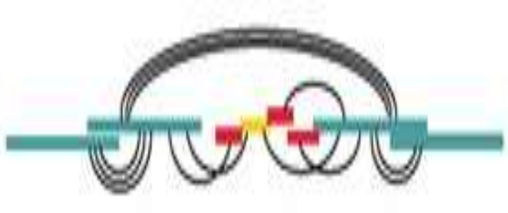


Computational quality control (QC)

De novo assembly

Gene prediction & annotation

Metabolic pathway mapping

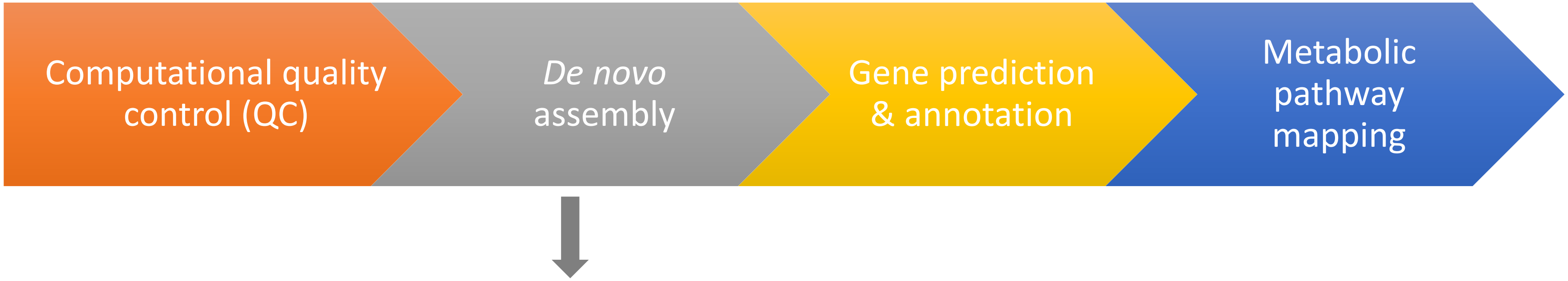


Why assemble?

- Simplify bioinformatic analysis relative to unassembled short metagenomic reads
- Better annoations/homology searches
- Possibility of assembling complete or near complete genomes

II. Metagenomic data analysis

Overview of steps involved in metagenomic data analysis



Computational quality control (QC)

De novo assembly

Gene prediction & annotation

Metabolic pathway mapping

Common strategies for *de novo* metagenomic assembly:

1. **Overlap-layout-consensus** (OLC): finds overlaps (O) among all the reads, carries out a layout (L) of all the reads and overlaps information on a graph and then infers the consensus (C) sequence

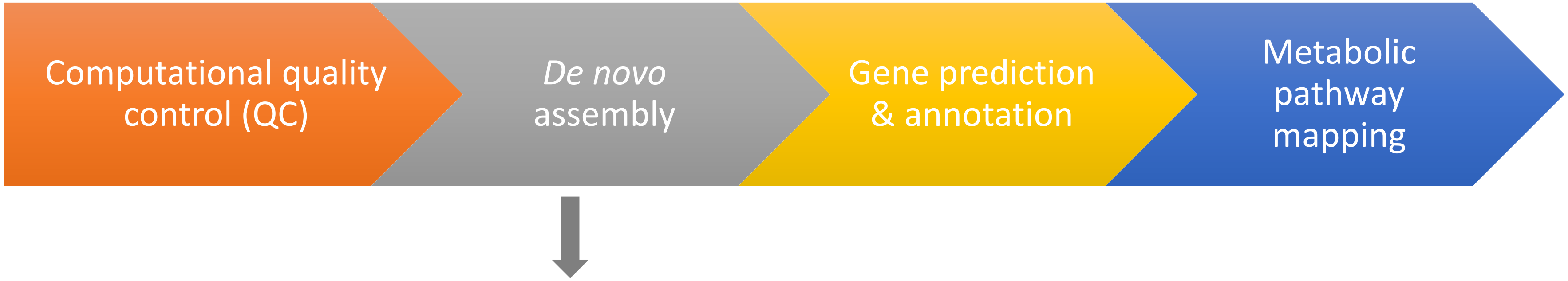
-efficient in handling longer reads, requires significant computational power

2. De Bruijn graph (DBG): chops reads into overlapping substrings of fixed length k (k-mers), these overlapping 'k-mers’ are organized in a graph structure. The assembler tries to findi an Eulerian path – a path through the graph that visits each edge once to infer the consensus sequence

- most commonly used algorithm for metagenomic assembly from short reads

II. Metagenomic data analysis

Overview of steps involved in metagenomic data analysis



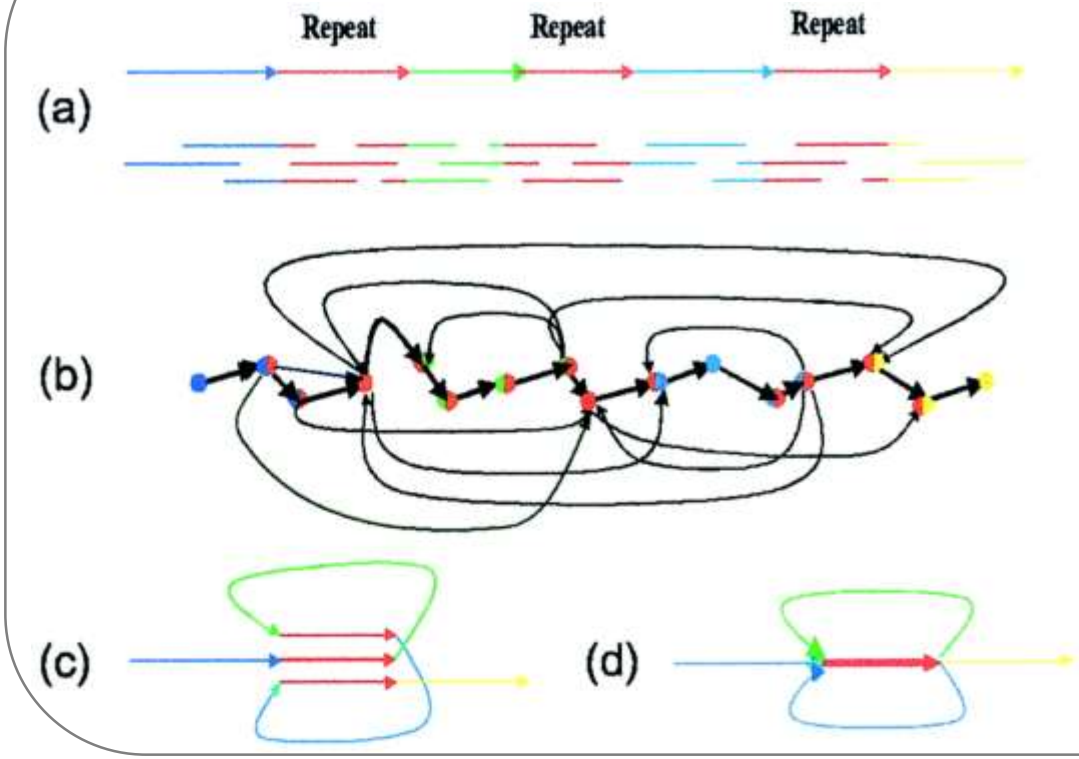
Computational quality control (QC)

De novo assembly

Gene prediction & annotation

Metabolic pathway mapping

De Bruijn graph

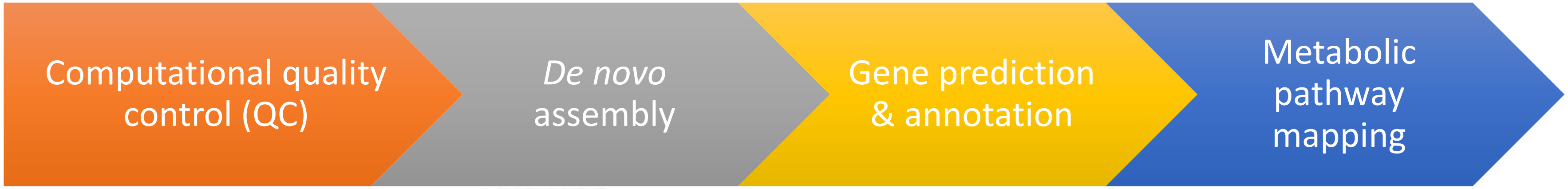


a) DNA sequence with a triple repeat
b) the layout graph
c) construction of the de Bruijn graph by gluing repeats
d) de Bruijn graph

Pavel A. Pevzner et al. PNAS 2001

II. Metagenomic data analysis

Overview of steps involved in metagenomic data analysis

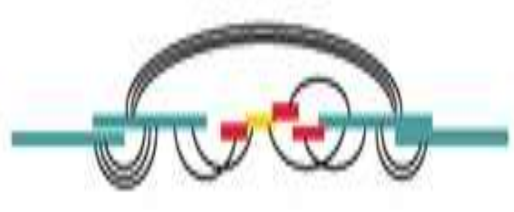


Computational quality control (QC)

De novo assembly

Gene prediction & annotation

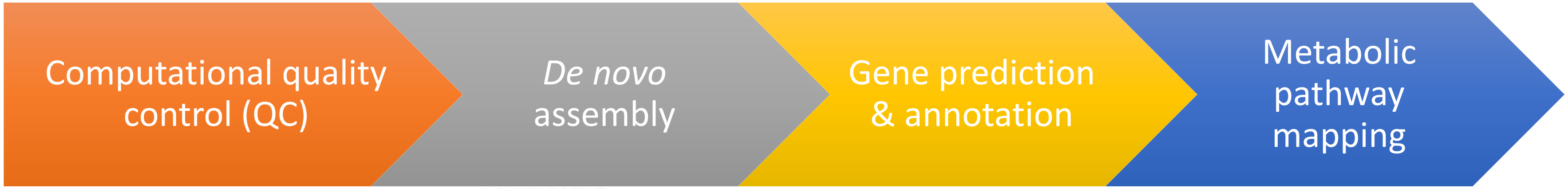
Metabolic pathway mapping



- **Common assembly tools based on DBG**: MEGAHIT, MetaVelvet, Meta-IDBA, metaSPAdes
- **Evaluating assembly quality**: completeness, continuity and propensity to generate chimeric contigs (tool: MetaQUAST)
- Very little community consensus on performance of different assemblers (Assemblathon)
- Choice of assembler depends on biological and technical factors of your study
- Might be a good idea to compare a few assemblers for your dataset

II. Metagenomic data analysis

Overview of steps involved in metagenomic data analysis

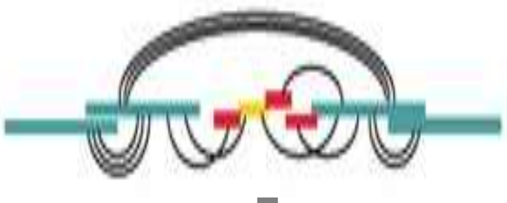


Computational quality control (QC)

De novo assembly

Gene prediction & annotation

Metabolic pathway mapping

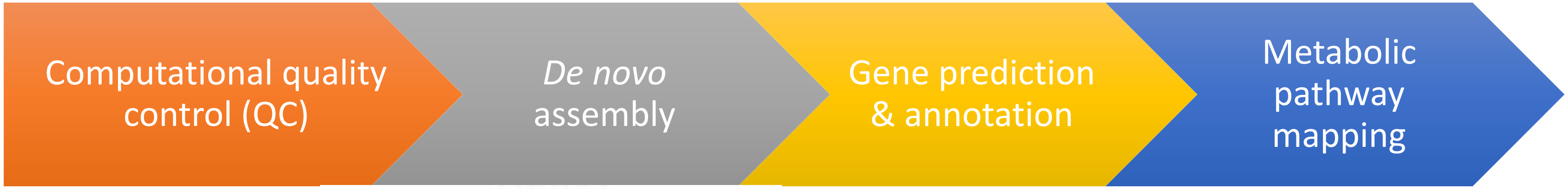


Binning of contigs:

- Grouping of metagenomic contings into “species” or “genomes” by supervised (based on databases of sequenced genomes) or unsupervised (in the absence of sequenced genomes) methods
- In most cases, a large fraction of contigs cannot be mapped to reference genomes as majority of microbial genomes have not been sequenced.
- Unsupervised binning uses similarity metrics based on tertramer (k-mer) frequencies, GC content, DNA sequence coverage and abundance pattern across samples

II. Metagenomic data analysis

Overview of steps involved in metagenomic data analysis

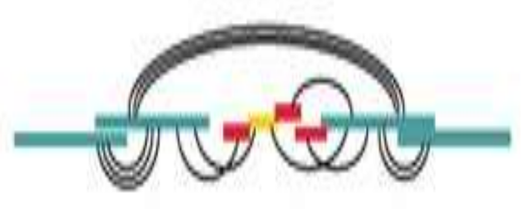


Computational quality control (QC)

De novo assembly

Gene prediction & annotation

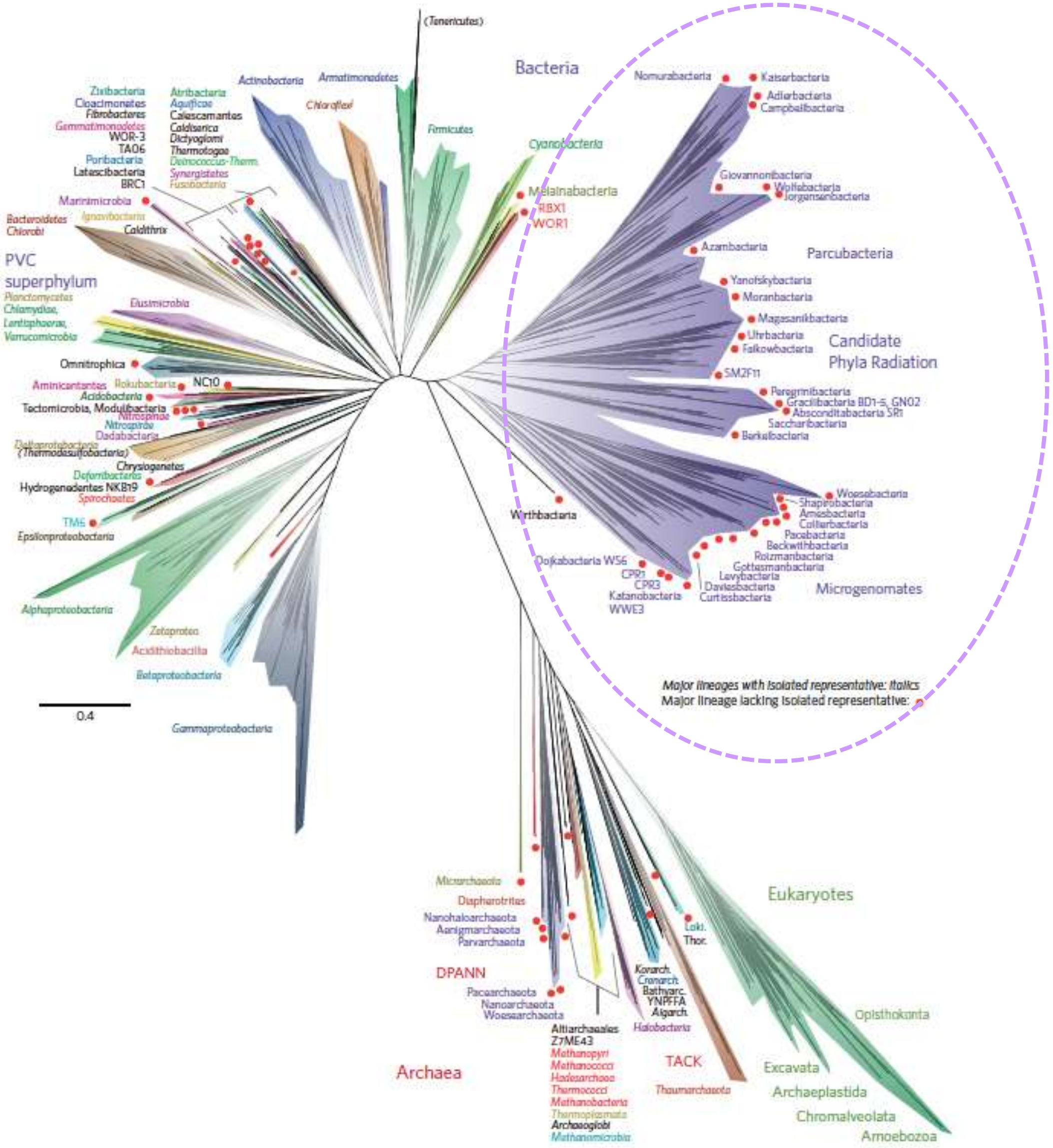
Metabolic pathway mapping



Metagenomic assembled genomes (MAGs)

- Discovery of uncultivable microbes and unraveling the hidden microbial diversity
- The Candidate Phyla Radiation, a new bacterial subdivision identified from MAGs has been added to the Tree of Life.

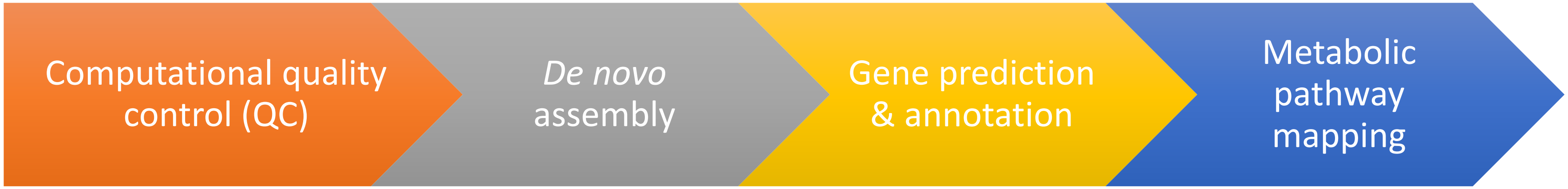
A new view of the tree of life



Hug et al., Nat. Microbiology, 2016

II. Metagenomic data analysis

Overview of steps involved in metagenomic data analysis

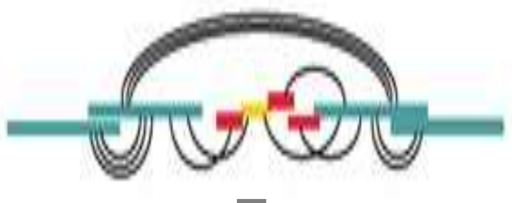


Computational quality control (QC)

De novo assembly

Gene prediction & annotation

Metabolic pathway mapping

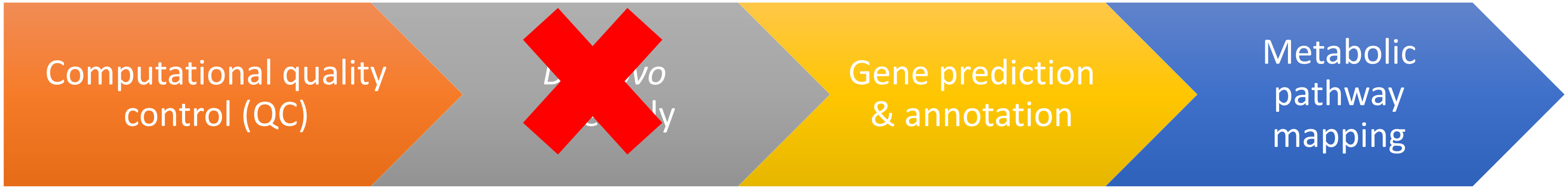


Challenges of *de novo* assembly

- Computationally intensive
- Generation of chimeras, wherein sequences from two distinct genomes are spuriously assembled into a contig
- Repetitive regions within a genome are difficult to assemble
- Sequencing errors are challenging for most assmeblers
- Difficult to assemble rare or low abundance species

II. Metagenomic data analysis

Overview of steps involved in metagenomic data analysis



Computational quality control (QC)

De novo assembly

Gene prediction & annotation

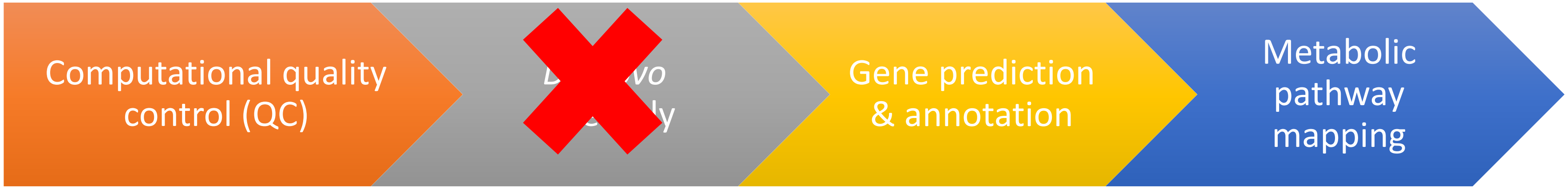
Metabolic pathway mapping

Assembly-free metagenomic profiling:

- Mitigates the problems associated with *de novo* assembly
- Difficult to profile previously uncharacterized microbes
- Potential way of studying well characterized systems (*eg*., human gut) where reference genomes are available
- Tools for taxonomic profiling from unassembled reads: mOTUs, MetaPhlAn, StrainPhlAn, SortMeRNA

II. Metagenomic data analysis

Overview of steps involved in metagenomic data analysis



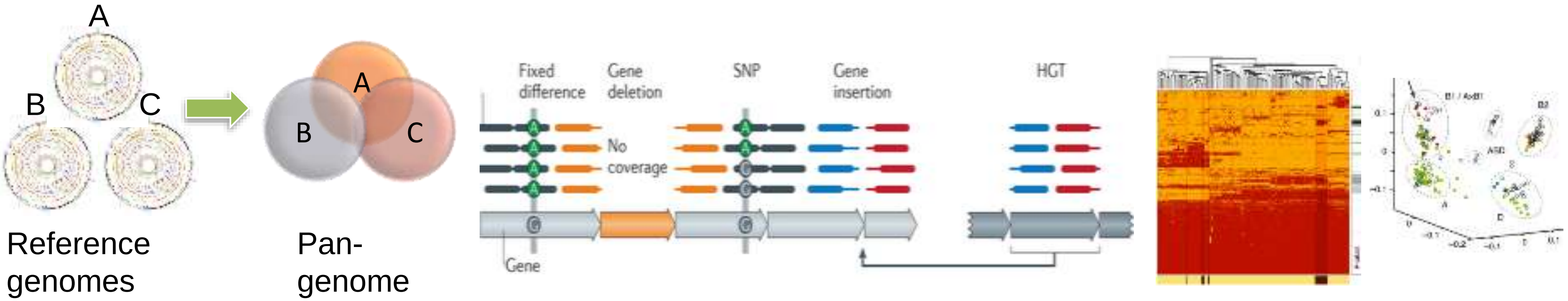
Computational quality control (QC)

~~De novo assembly~~

Gene prediction & annotation

Metabolic pathway mapping

Example workflow for assembly free strain-level metagenomic profiling:



Construction of Pan-genome database

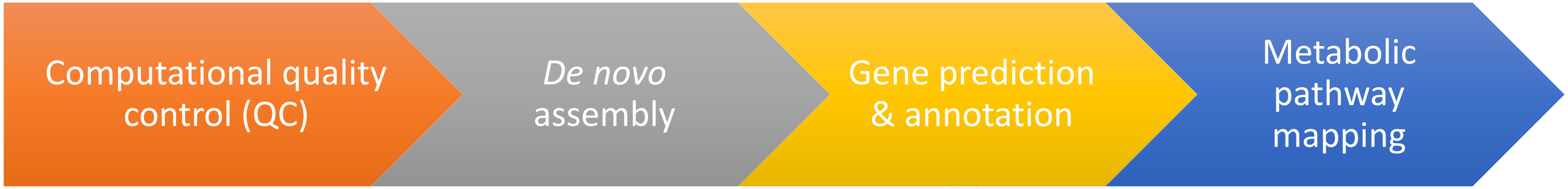
Mapping metagenomic reads against pan-genome database

Generation of strain-specific gene repertoire

Based on workflow of StrainPhlAn and PanPhlAn

II. Metagenomic data analysis

Overview of steps involved in metagenomic data analysis

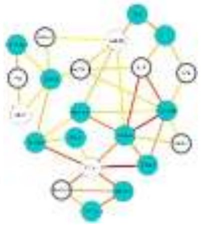


Computational quality control (QC)

De novo assembly

Gene prediction & annotation

Metabolic pathway mapping

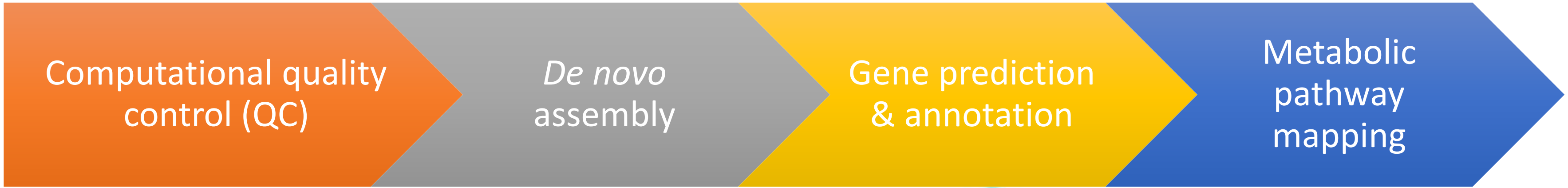


Gene prediction: optimized for metagenomic datasets, includes *ab initio* methods that are able to identify genes having no similarity to ones existing in databases

Tools: FragGeneScan, MetaGeneMark, Glimmer-MG

II. Metagenomic data analysis

Overview of steps involved in metagenomic data analysis

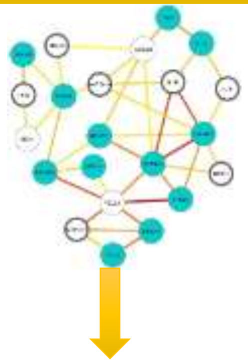


Computational quality control (QC)

De novo assembly

Gene prediction & annotation

Metabolic pathway mapping



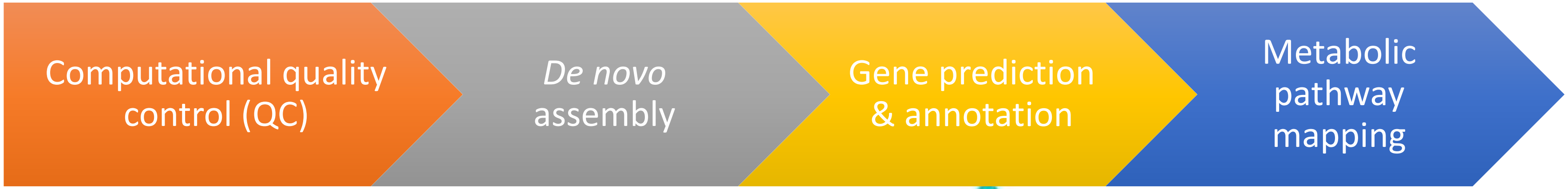
Databases for annotation:

KEGG: Kyoto Encyclopedia of Genes and Genomes contains detailed genomic and chemical information, network information including molecular wiring diagrams (interaction/reaction networks) and hierarchical classifications (relation networks) to represent high-level functions.

FOAM: functional ontology dedicated to classify gene functions relevant to environmental microorganisms based on Hidden Markov Models (HMMs).

II. Metagenomic data analysis

Overview of steps involved in metagenomic data analysis

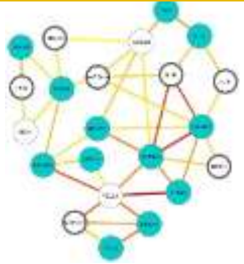


Computational quality control (QC)

De novo assembly

Gene prediction & annotation

Metabolic pathway mapping



Databases for annotation (cont.)..

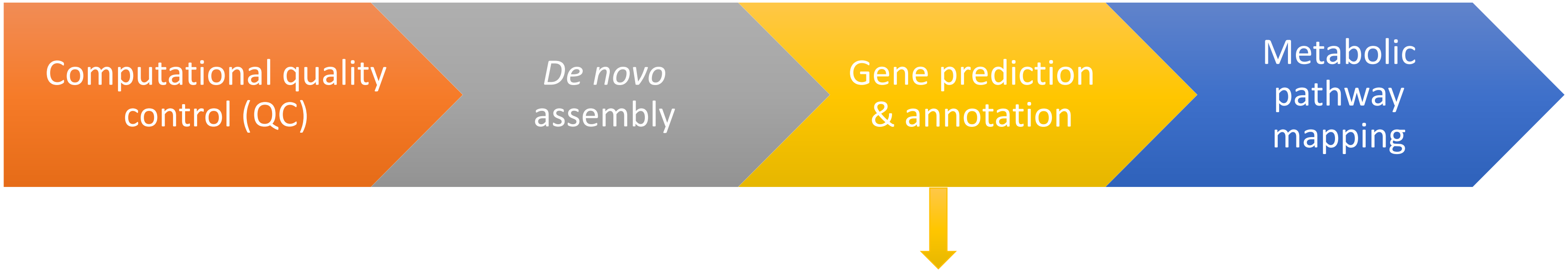
PFAM : collection of protein families, each represented by multiple sequence HMMs.

TIGRFAM : HMMs for protein sequence classification, and associated information designed to support automated annotation of (mostly prokaryotic) proteins.

eggNOG : Orthologous Groups (OGs) of proteins at different taxonomic levels, each with integrated and summarized functional annotations.

II. Metagenomic data analysis

Overview of steps involved in metagenomic data analysis

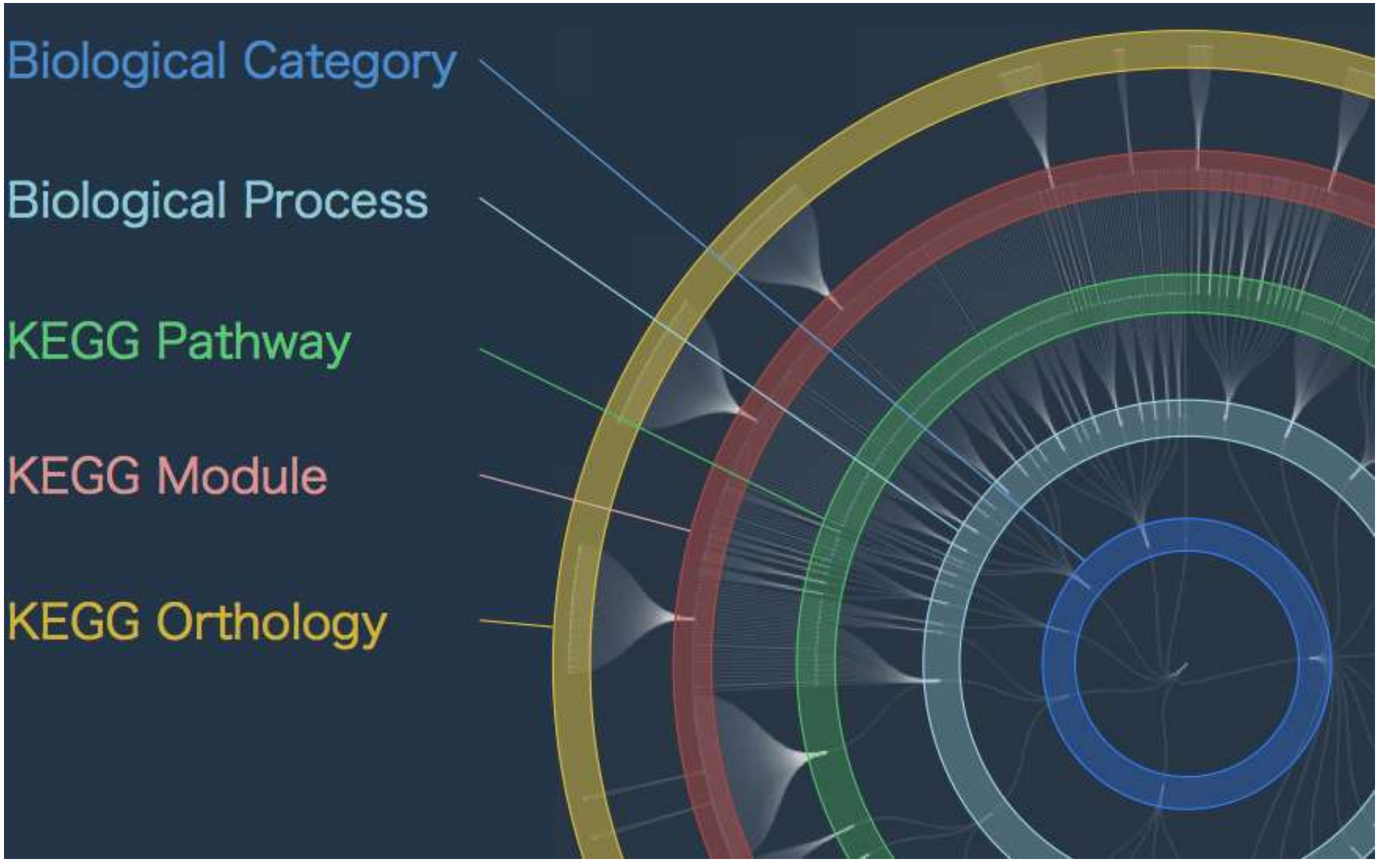


Computational quality control (QC)

De novo assembly

Gene prediction & annotation

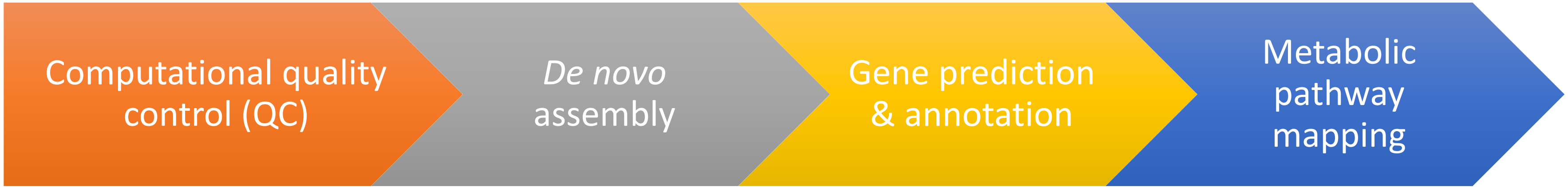
Metabolic pathway mapping



Source: FuncTree

II. Metagenomic data analysis

Overview of steps involved in metagenomic data analysis

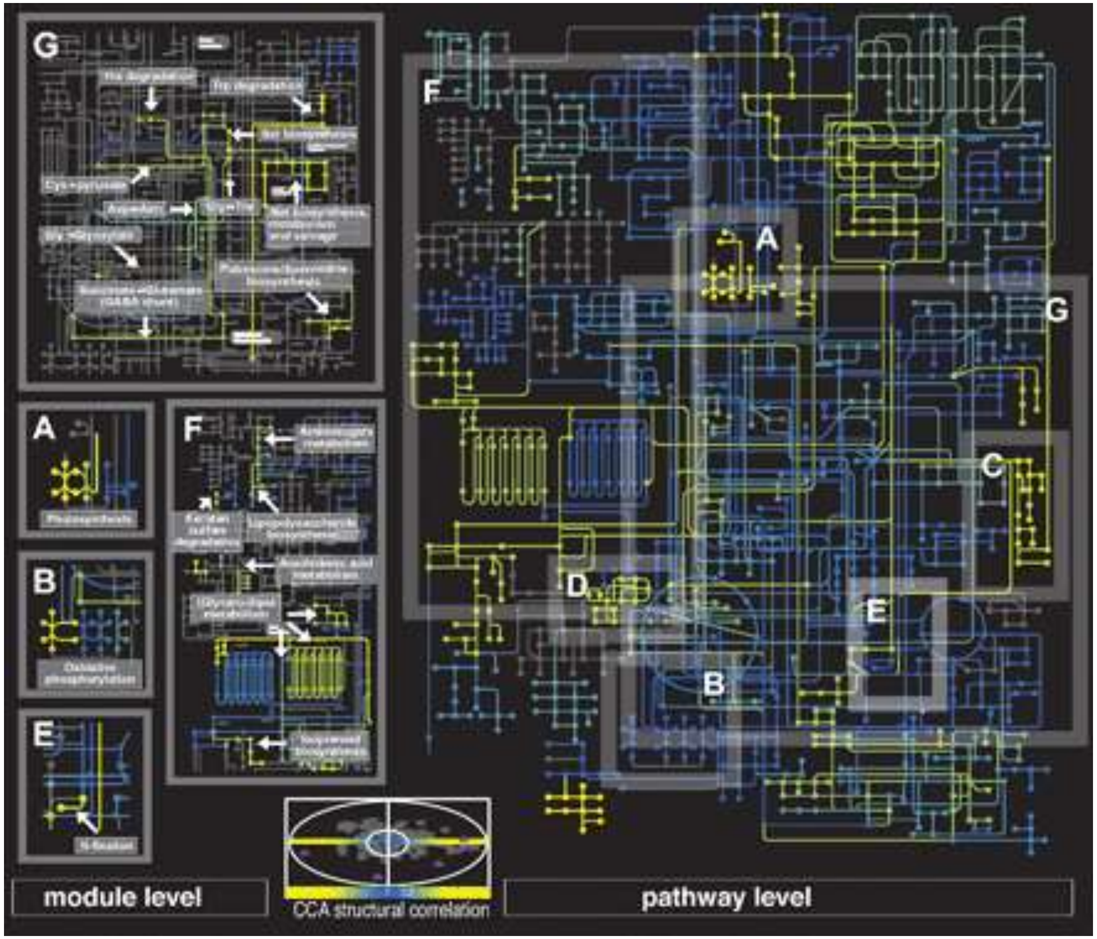


Computational quality control (QC)

De novo assembly

Gene prediction & annotation

Metabolic pathway mapping

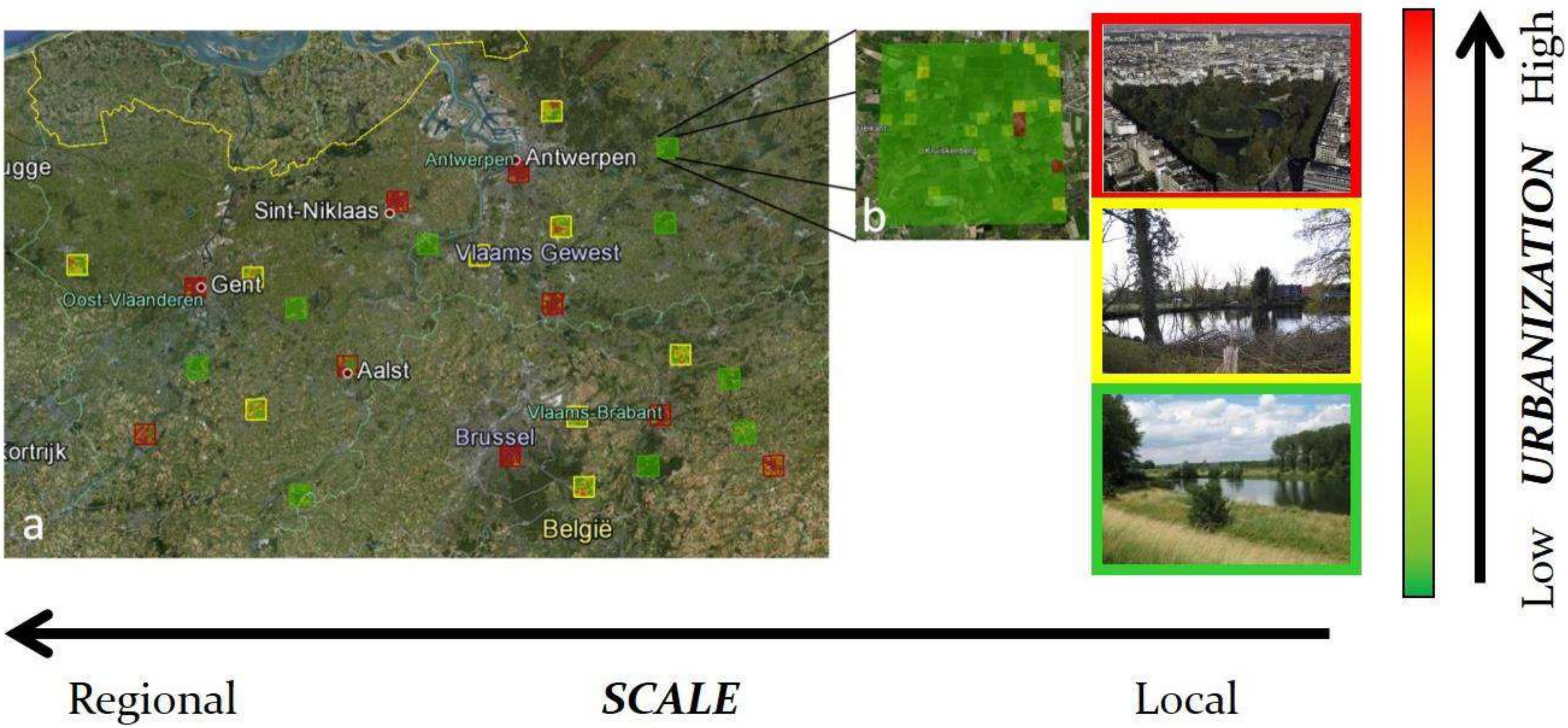


Gianoulis et al, PNAS 2009

III. Post-processing and statistical analysis

- Final output comprises of data matrices of samples versus relative abundance of metagenomic features (genes, modules, pathways, species etc.)
- Major goal of metagenomic studies is to characterize metagenomic features of individual samples and to interpret the correlations between metagenomic features and sample metadata
- Similar statistical tools as applied for metagenetics data analysis can be used for metagenomic feature analysis (eg., constrained and unconstrained multivariate analyses, univariate analyses, heatmaps, networks)
- R packages: vegan, phyloseq, DESeq2, metagenomeSeq etc.

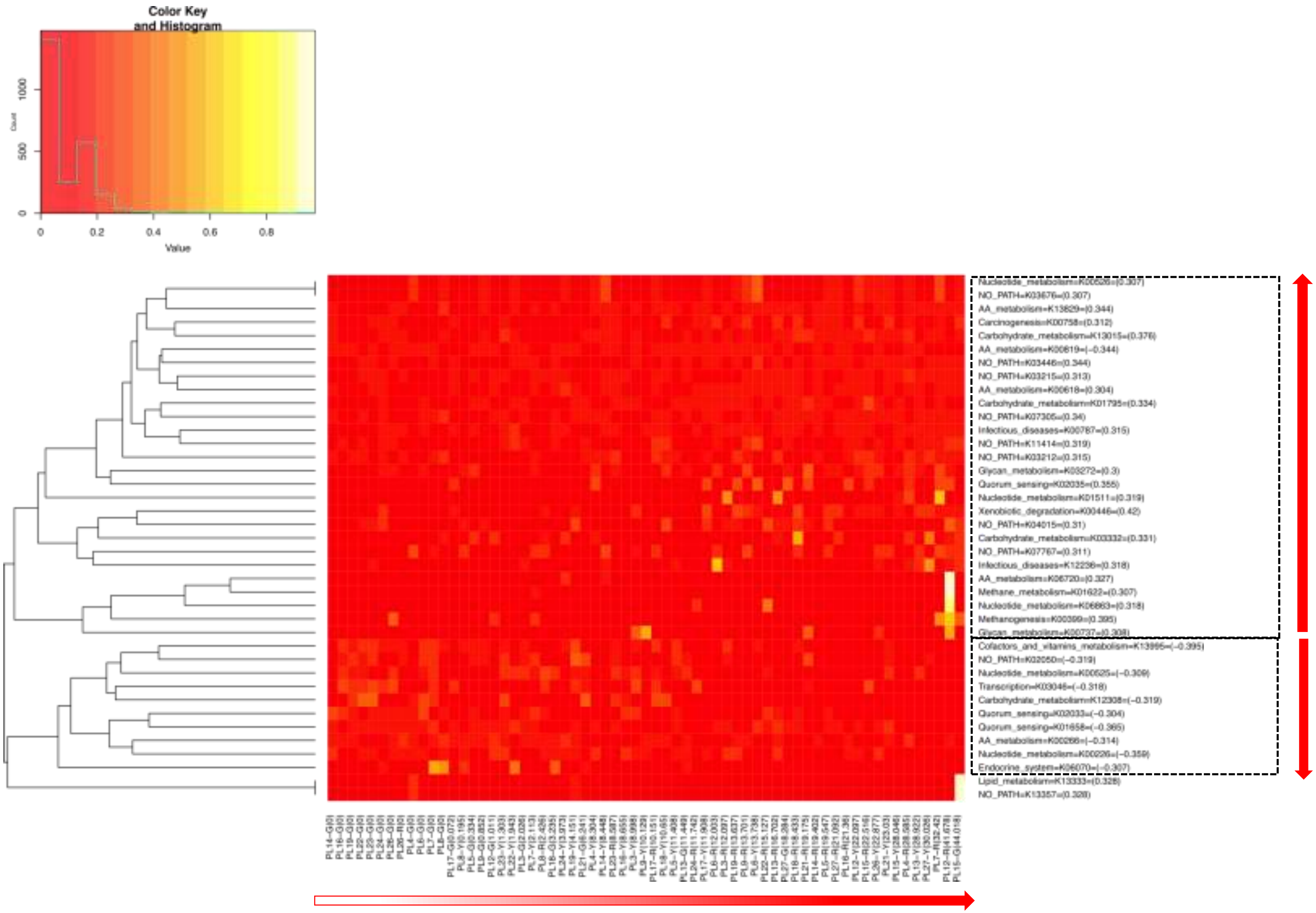
IV. MicroCity: A metagenomic trait-based approach for identifying microbial responses to urbanization



3 step gradient in urbanization × 2 spatial scales (local and regional)

Manuscript under prep.: Shinjini Mukherjee, Youssef Darzi, Caroline Souffreau, Ajay Anand Kumar, Jeroen Raes and Luc De Meester

IV. MicroCity: A metagenomic trait-based approach for identifying microbial responses to urbanization

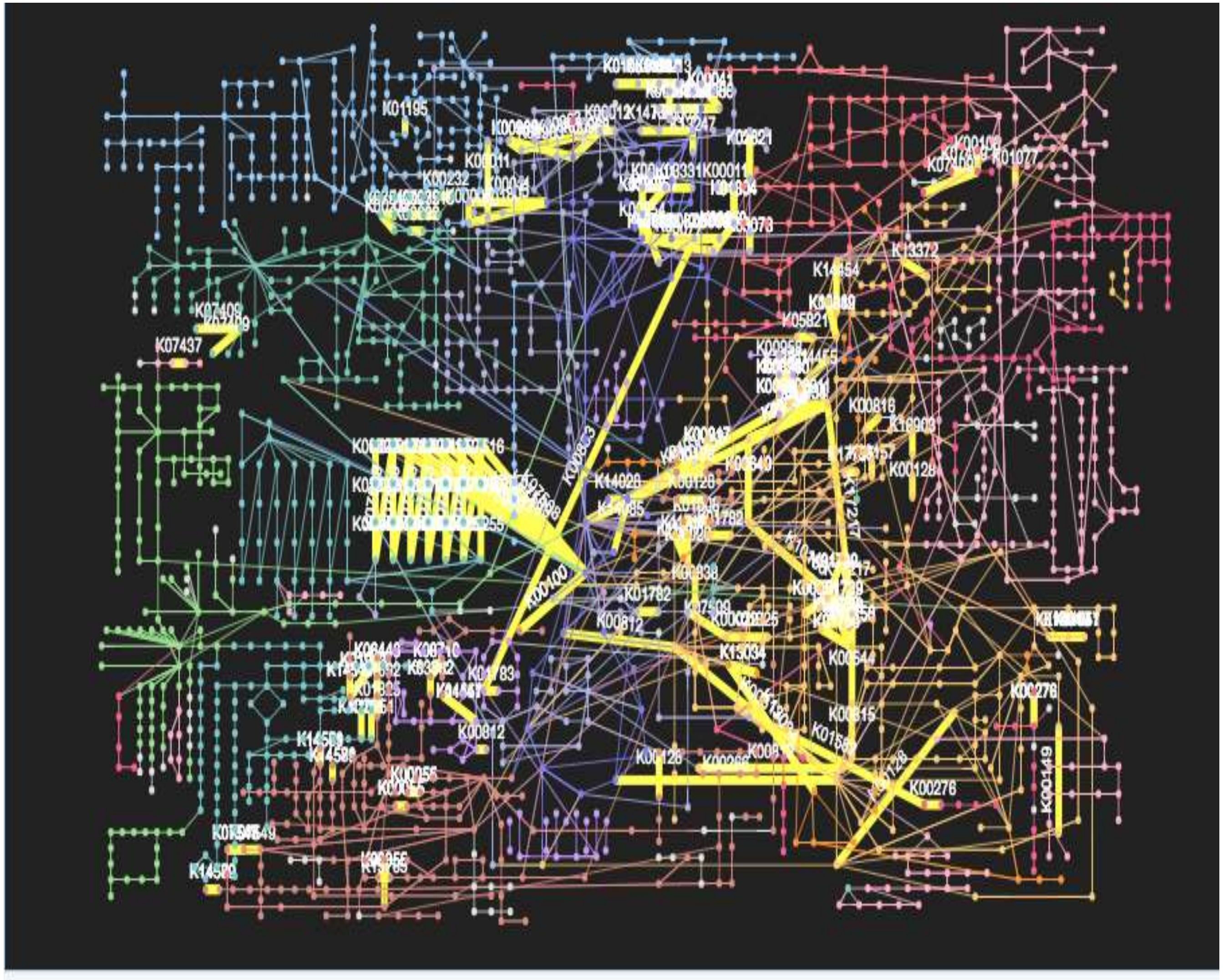


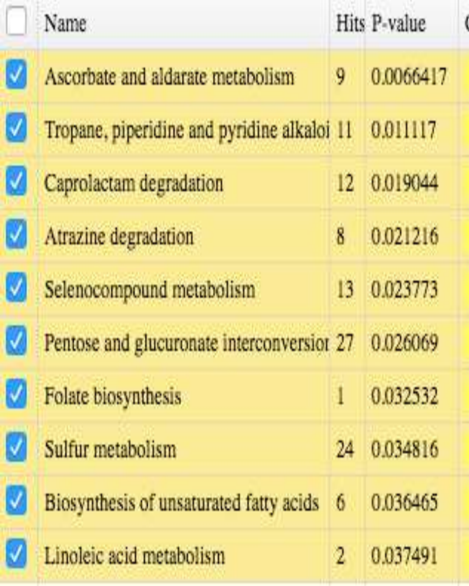
Relative abundance of KEGG Orthologs (Kos)

Urbanization (percentage built-up area)

IV. MicroCity: A metagenomic trait-based approach for identifying microbial responses to urbanization

Pathways enriched in urbanized ponds

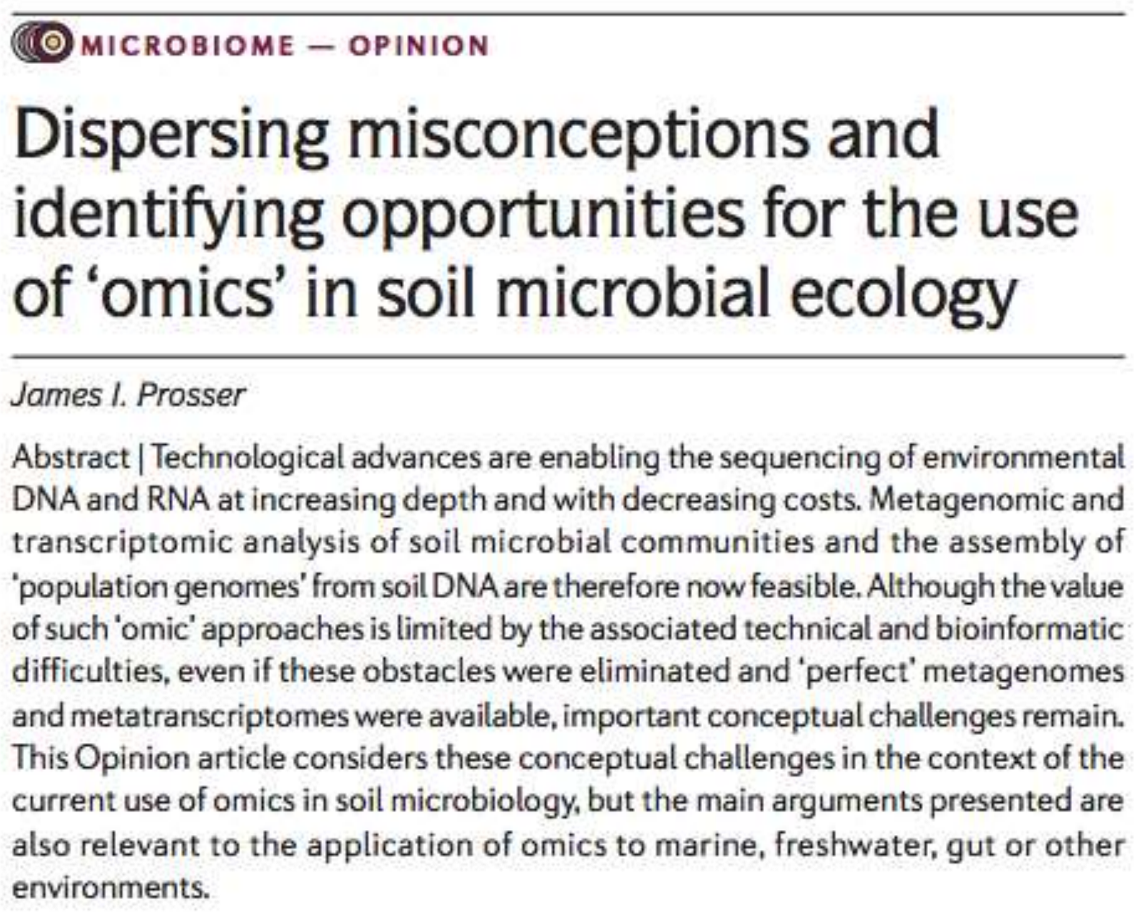




□	Name	Hits	P-value
☑	Ascorbate and aldarate metabolism	9	0.0066417
☑	Tropane, piperidine and pyridine alkaloi	11	0.011117
☑	Caprolactam degradation	12	0.019044
☑	Atrazine degradation	8	0.021216
☑	Selenocompound metabolism	13	0.023773
☑	Pentose and glucuronate interconversior	27	0.026069
☑	Folate biosynthesis	1	0.032532
☑	Sulfur metabolism	24	0.034816
☑	Biosynthesis of unsaturated fatty acids	6	0.036465
☑	Linoleic acid metabolism	2	0.037491

V. Limitations of metagenomics

An excellent read!



MICROBIOME — OPINION

Dispersing misconceptions and identifying opportunities for the use of 'omics' in soil microbial ecology

James I. Prosser

Abstract | Technological advances are enabling the sequencing of environmental DNA and RNA at increasing depth and with decreasing costs. Metagenomic and transcriptomic analysis of soil microbial communities and the assembly of 'population genomes' from soil DNA are therefore now feasible. Although the value of such 'omic' approaches is limited by the associated technical and bioinformatic difficulties, even if these obstacles were eliminated and 'perfect' metagenomes and metatranscriptomes were available, important conceptual challenges remain. This Opinion article considers these conceptual challenges in the context of the current use of omics in soil microbiology, but the main arguments presented are also relevant to the application of omics to marine, freshwater, gut or other environments.

Nature Reviews Microbiology volume 13, pages 439–446 (2015)

Thank you!

The data deluge | The Economist