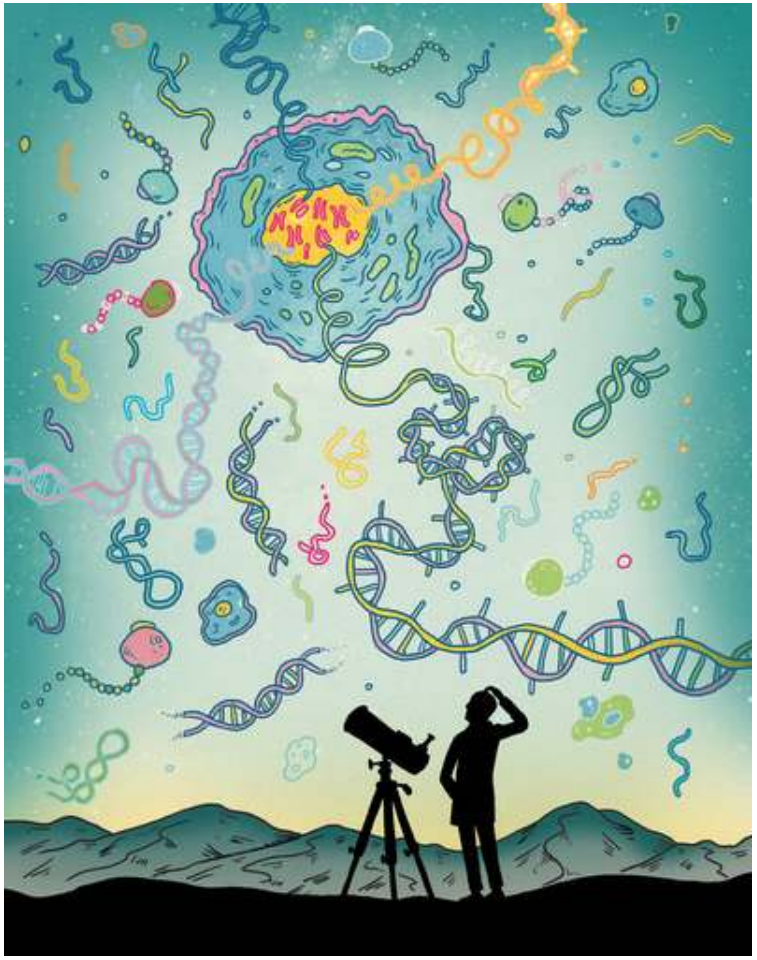
Club Bioinfo IBIS

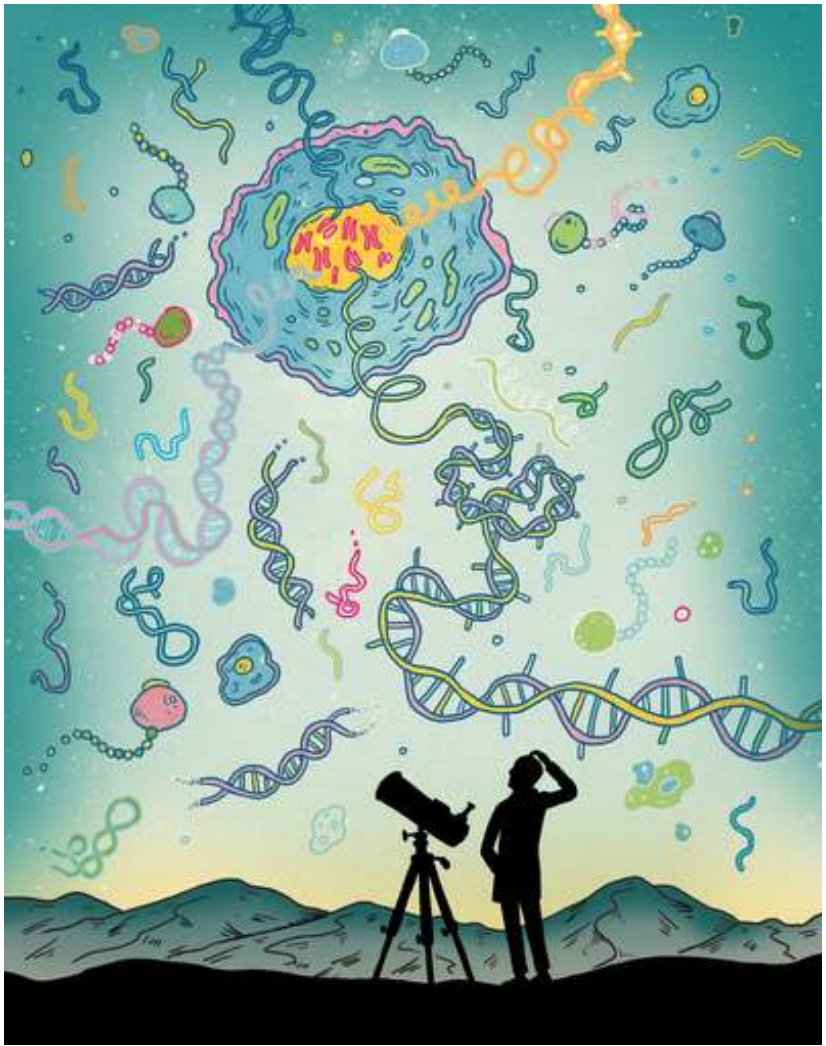
Lire un métagénome c'est retrouver son chemin dans une forêt ?



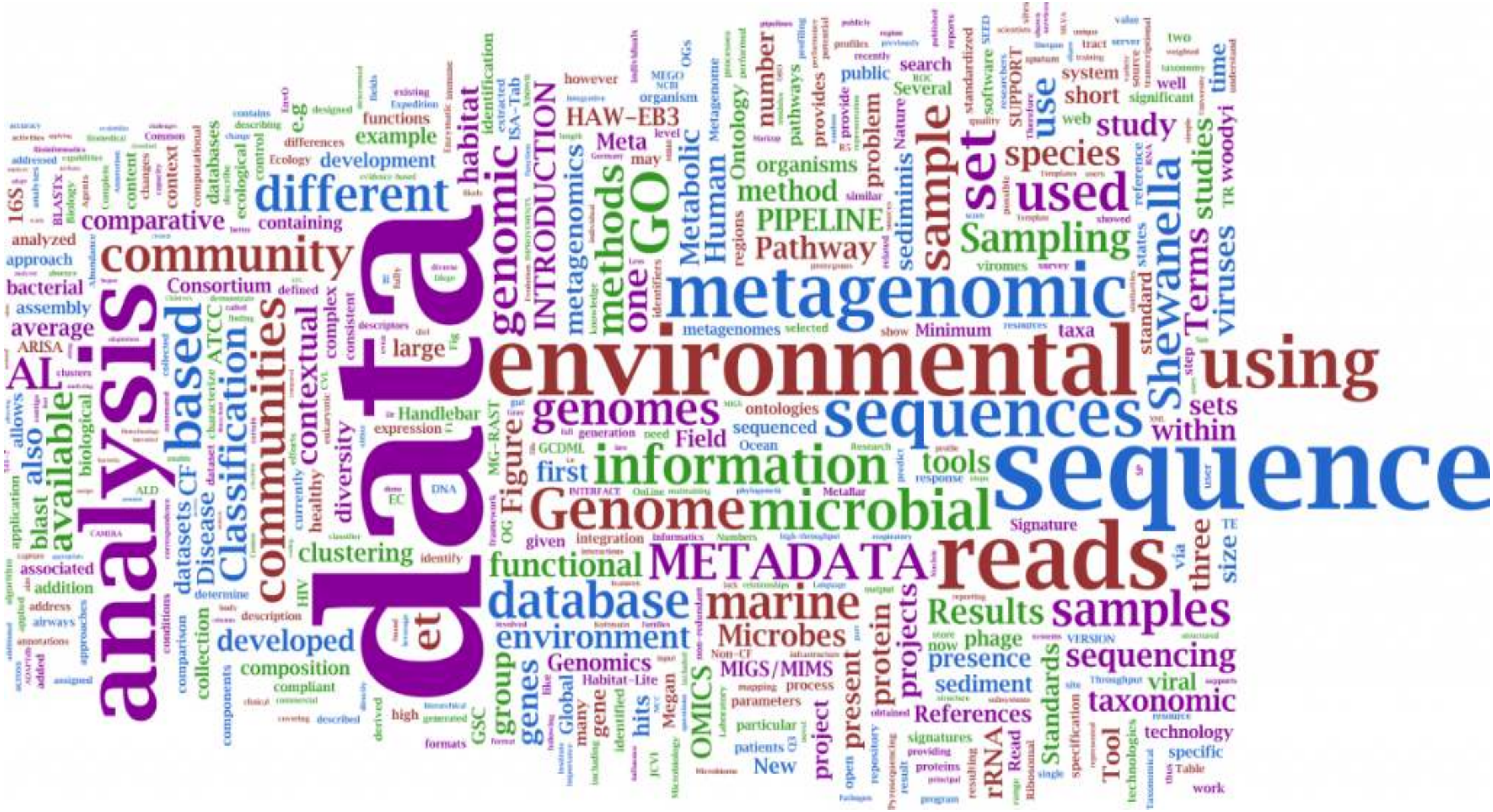
Bachar Cheaib
Laboratoire Derome
Le 07 Avril 2015

Du ciel confortable (bio) à la forêt inconfortable (bioinfo)

NGS



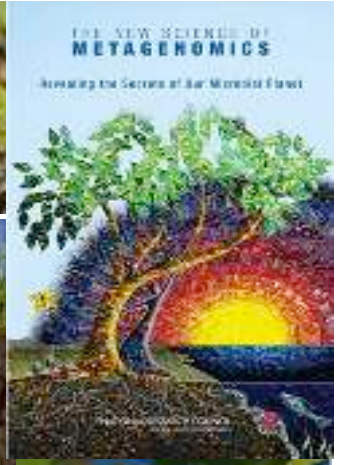
Rendre la complexité simple et lisible



La métagénomique une méthode d'accès aux ressources génétiques

Échantillonnage

culture cellulaire


METAGENOMICS


Metagenomes

NEON Domains

La métagénomique une méthode d'accès aux ressources génétiques

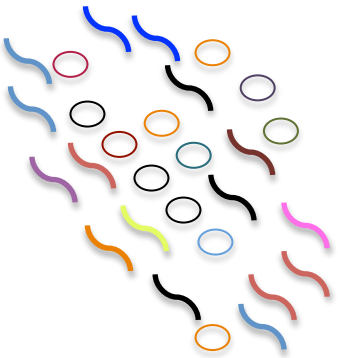
Échantillonnage

↓

Microfiltrage adapté

↓

Extraction de L'ADN total

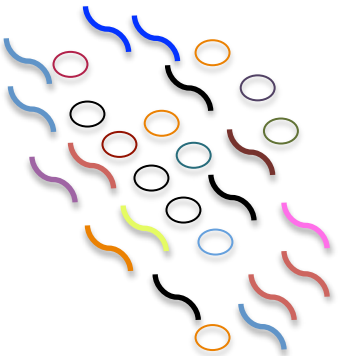


La métagénomique une méthode d'accès aux ressources génétiques

Échantillonnage

Microfiltrage adapté

Extraction de L'ADN total



Coupure

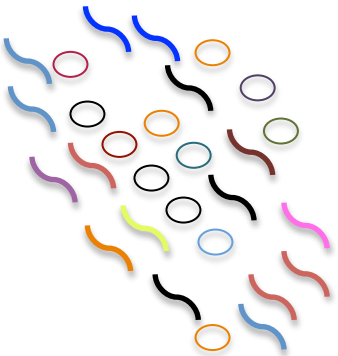
Préparation des librairies amplification

La métagénomique une méthode d'accès aux ressources génétiques

Échantillonnage

Microfiltrage adapté

Extraction de L'ADN total

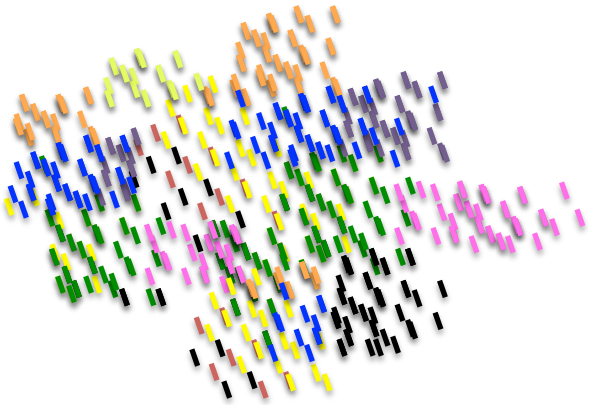


Coupure

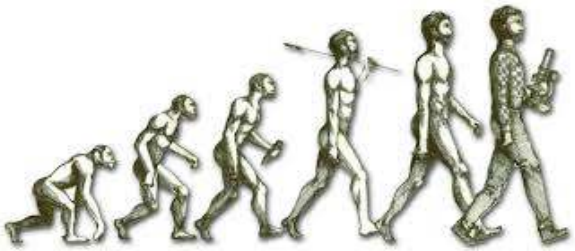
Préparation des librairies amplification

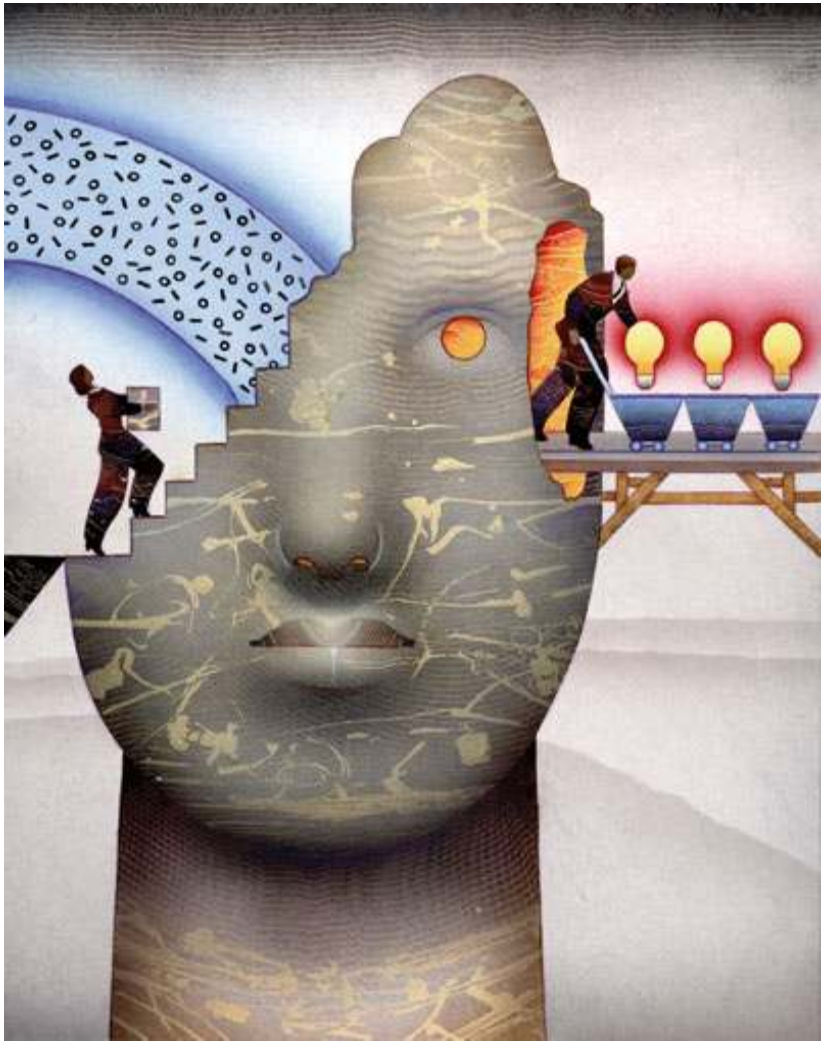
Séquençage massif

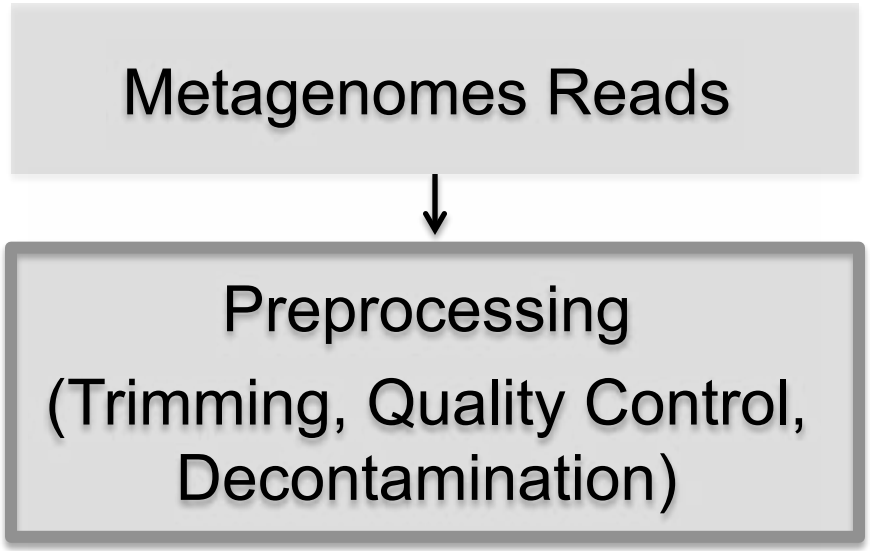
fragments courts

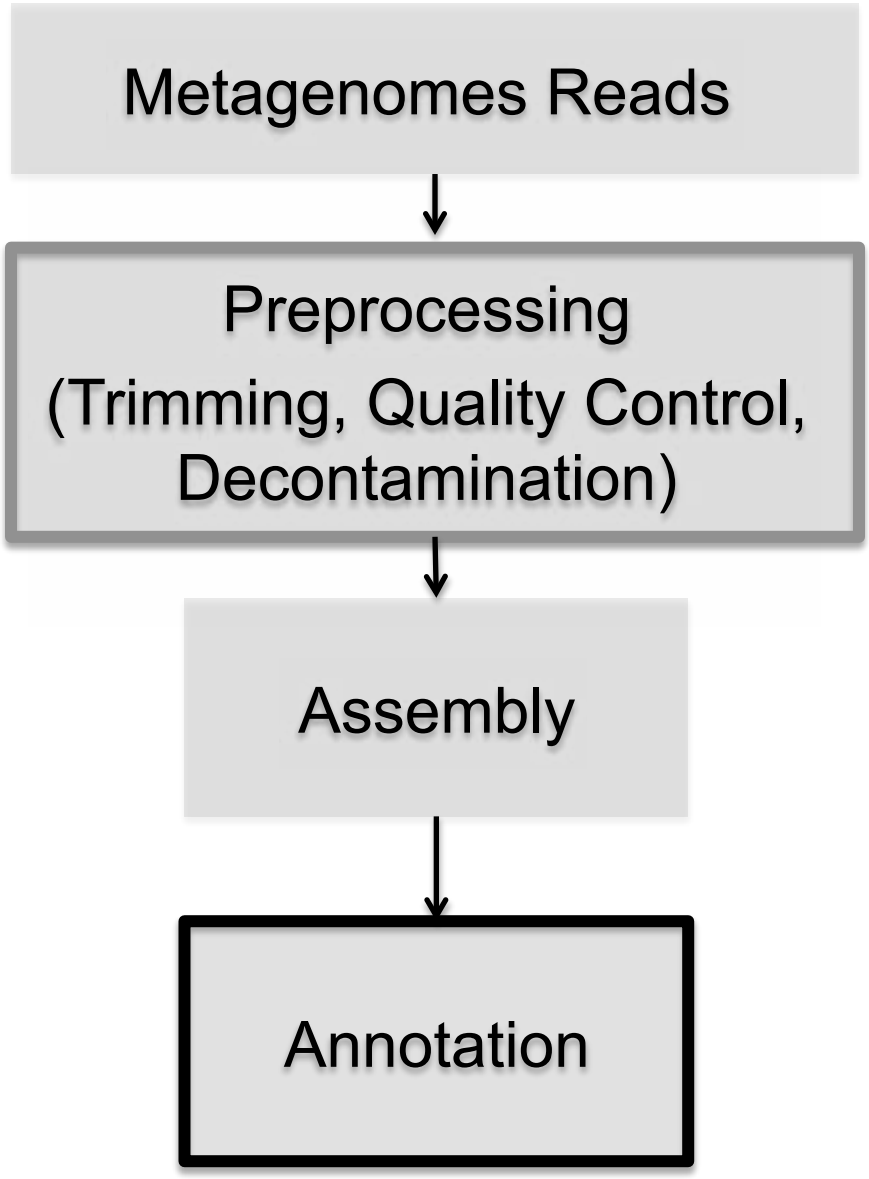


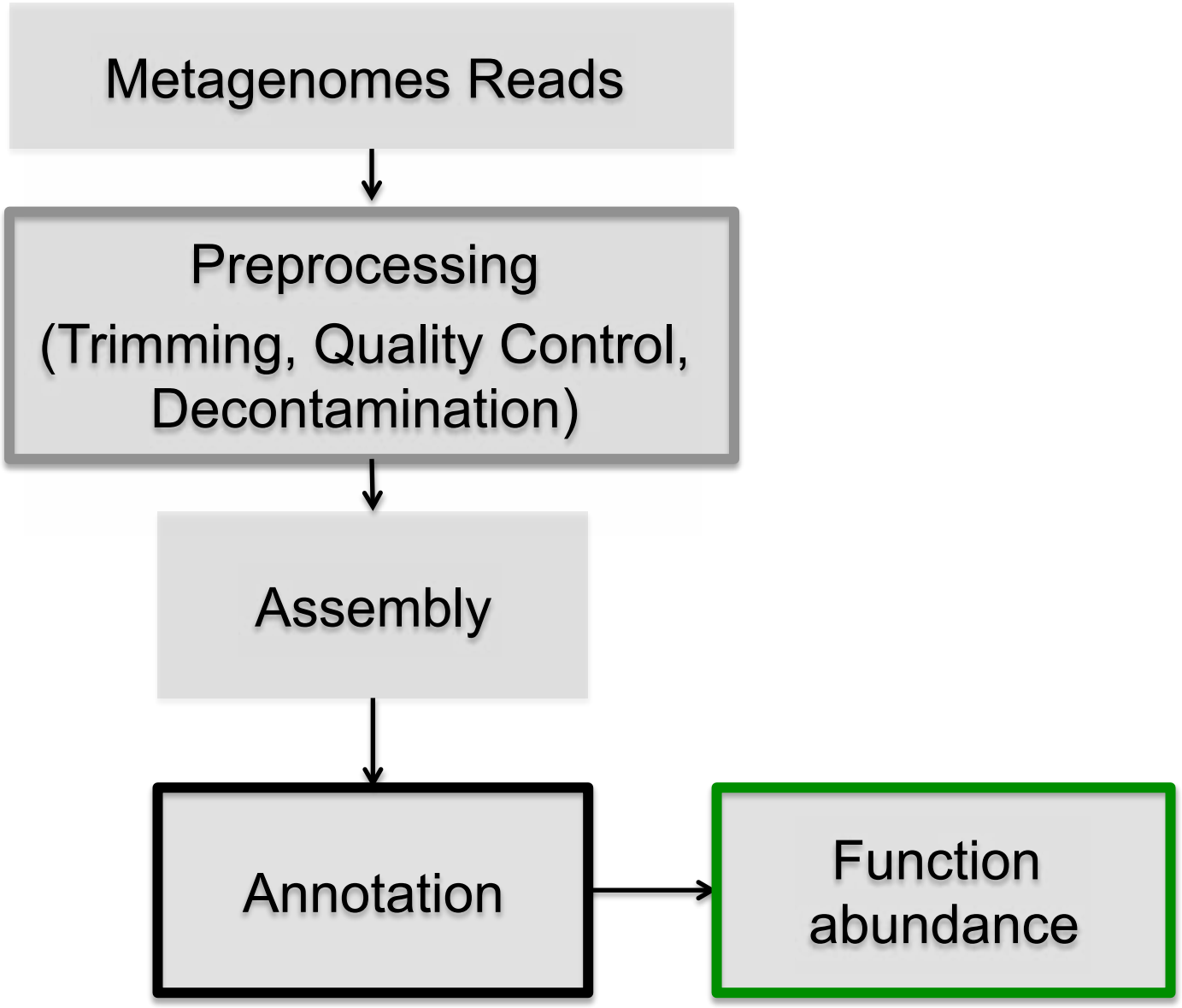
Du séquençage à la bioinformatique

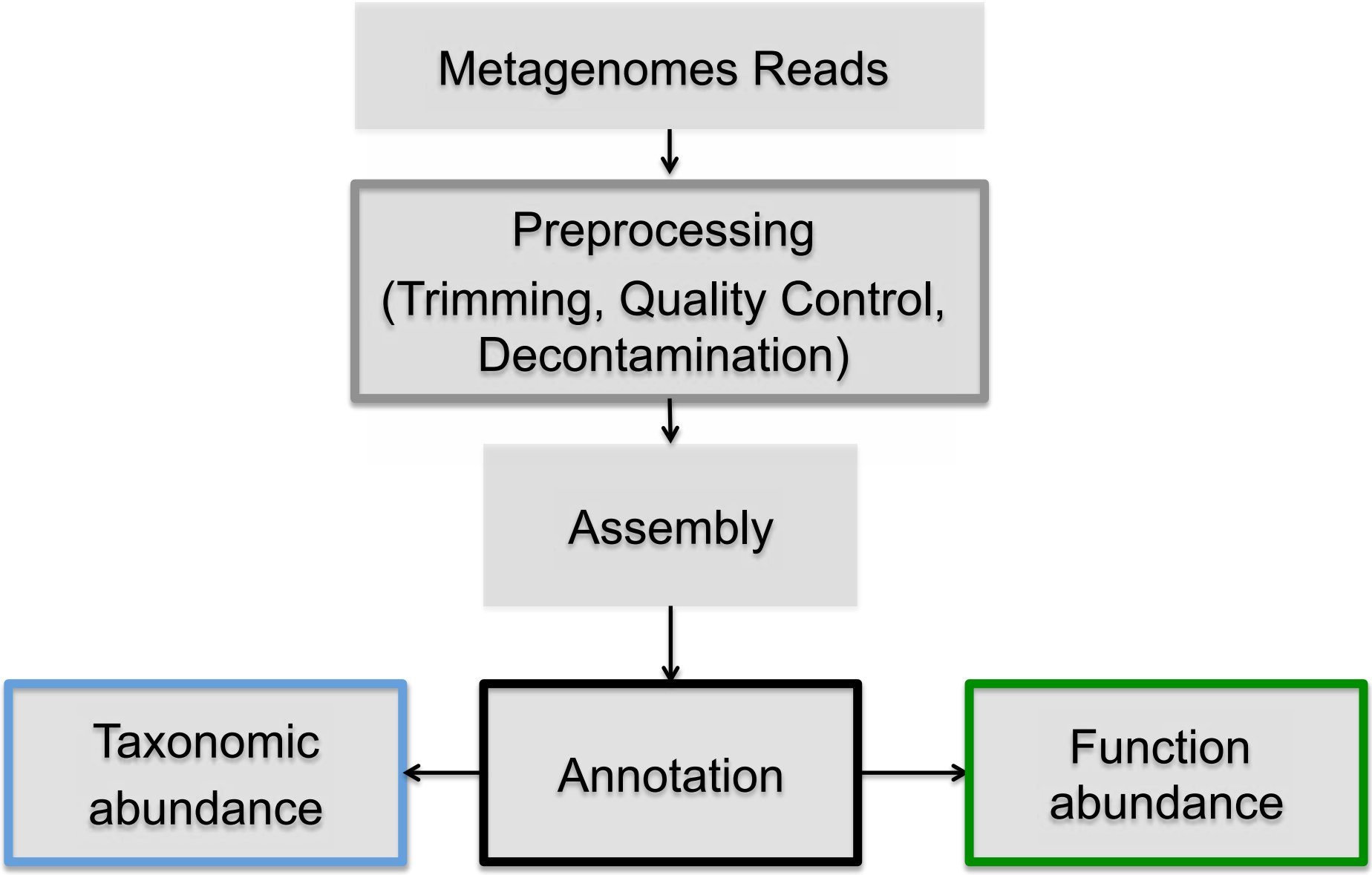


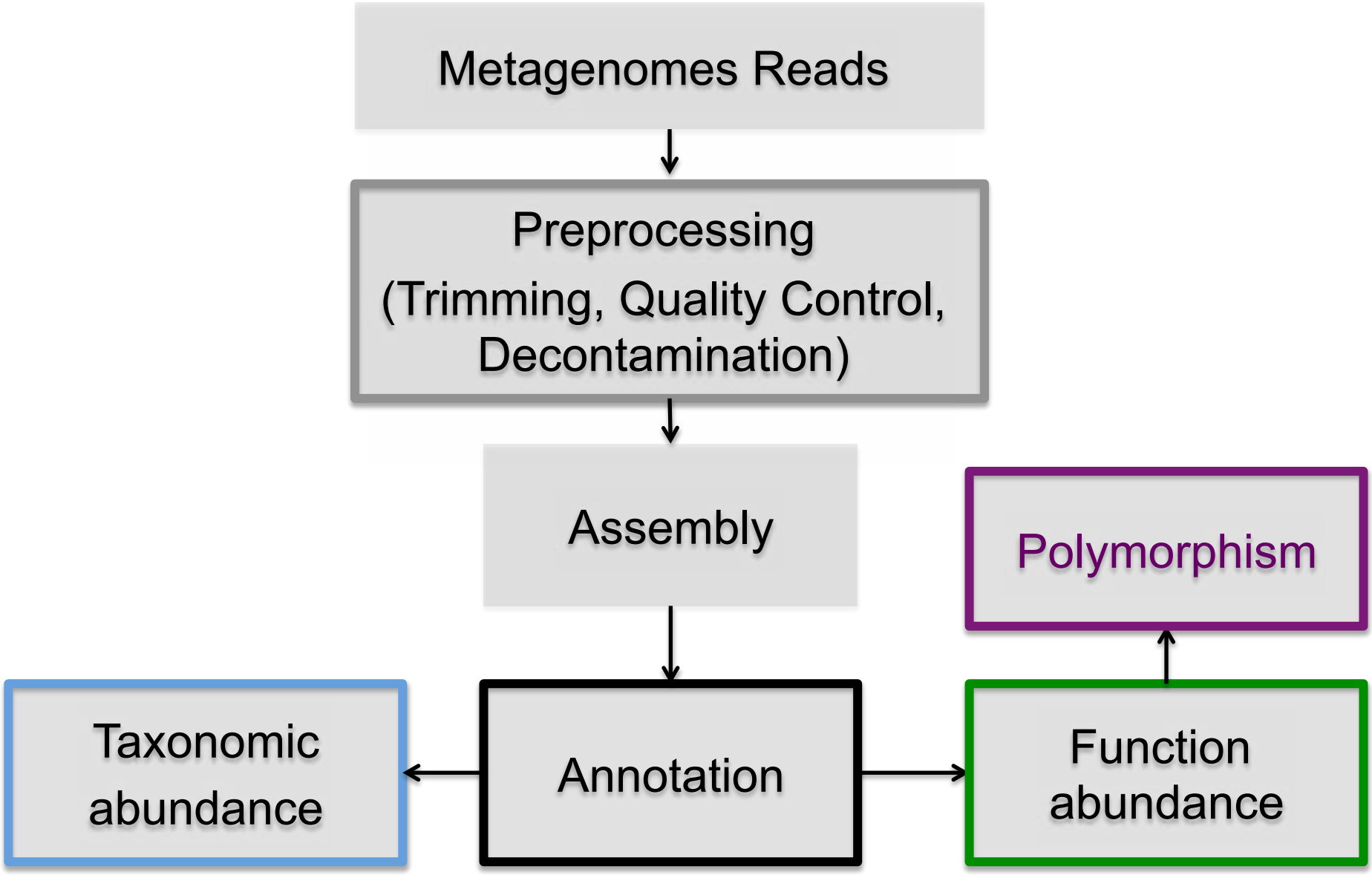


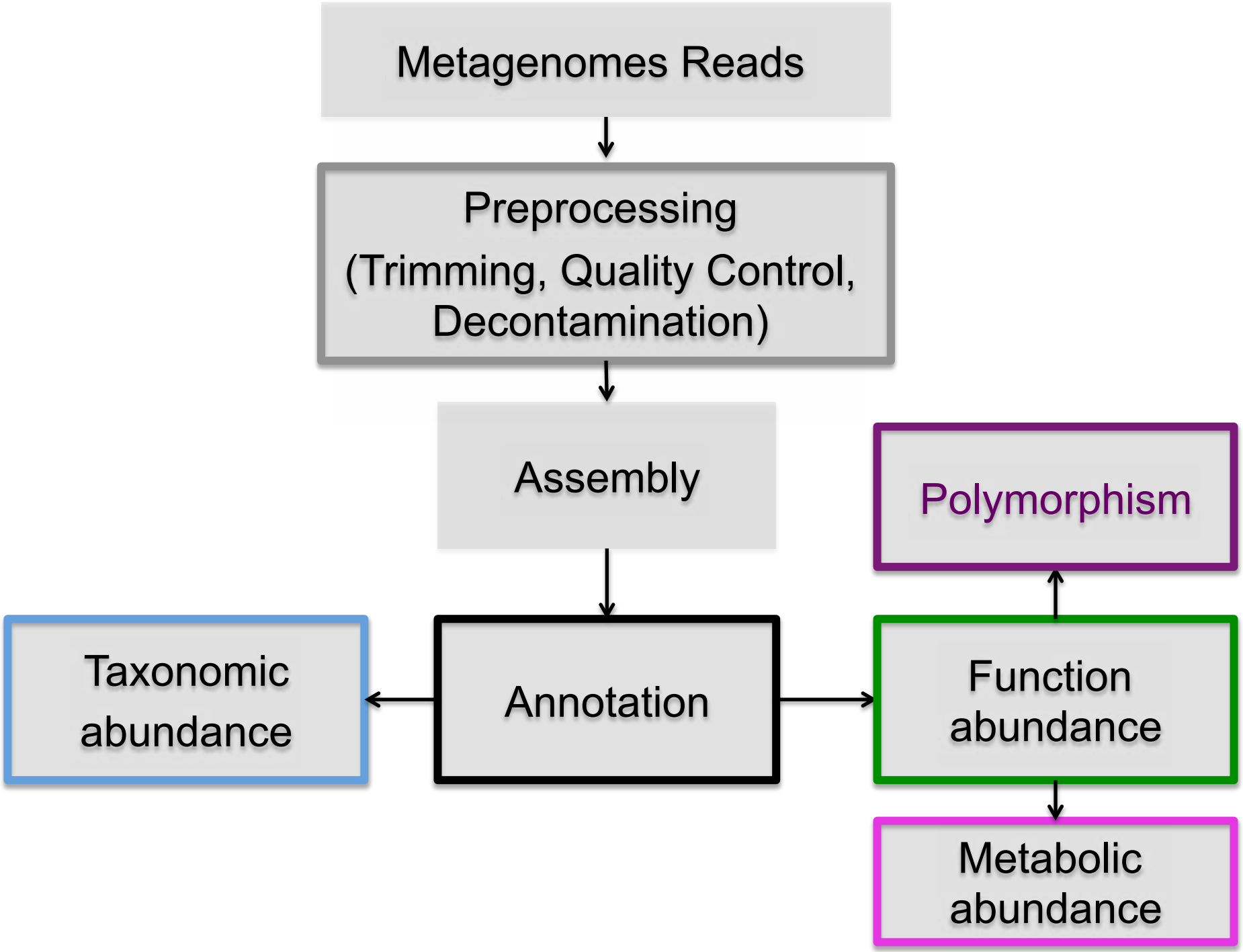

Metagenomes Reads
Preprocessing
(Trimming, Quality Control, Decontamination)


Metagenomes Reads
Preprocessing
(Trimming, Quality Control, Decontamination)
Assembly
Annotation


Metagenomes Reads
Preprocessing
(Trimming, Quality Control, Decontamination)
Assembly
Annotation
Function abundance


Metagenomes Reads
Preprocessing
(Trimming, Quality Control, Decontamination)
Assembly
Taxonomic abundance
Annotation
Function abundance


Metagenomes Reads
Preprocessing
(Trimming, Quality Control, Decontamination)
Assembly
Polymorphism
Taxonomic abundance
Annotation
Function abundance


Metagenomes Reads
Preprocessing
(Trimming, Quality Control, Decontamination)
Assembly
Polymorphism
Taxonomic abundance
Annotation
Function abundance
Metabolic abundance

Before filtering…

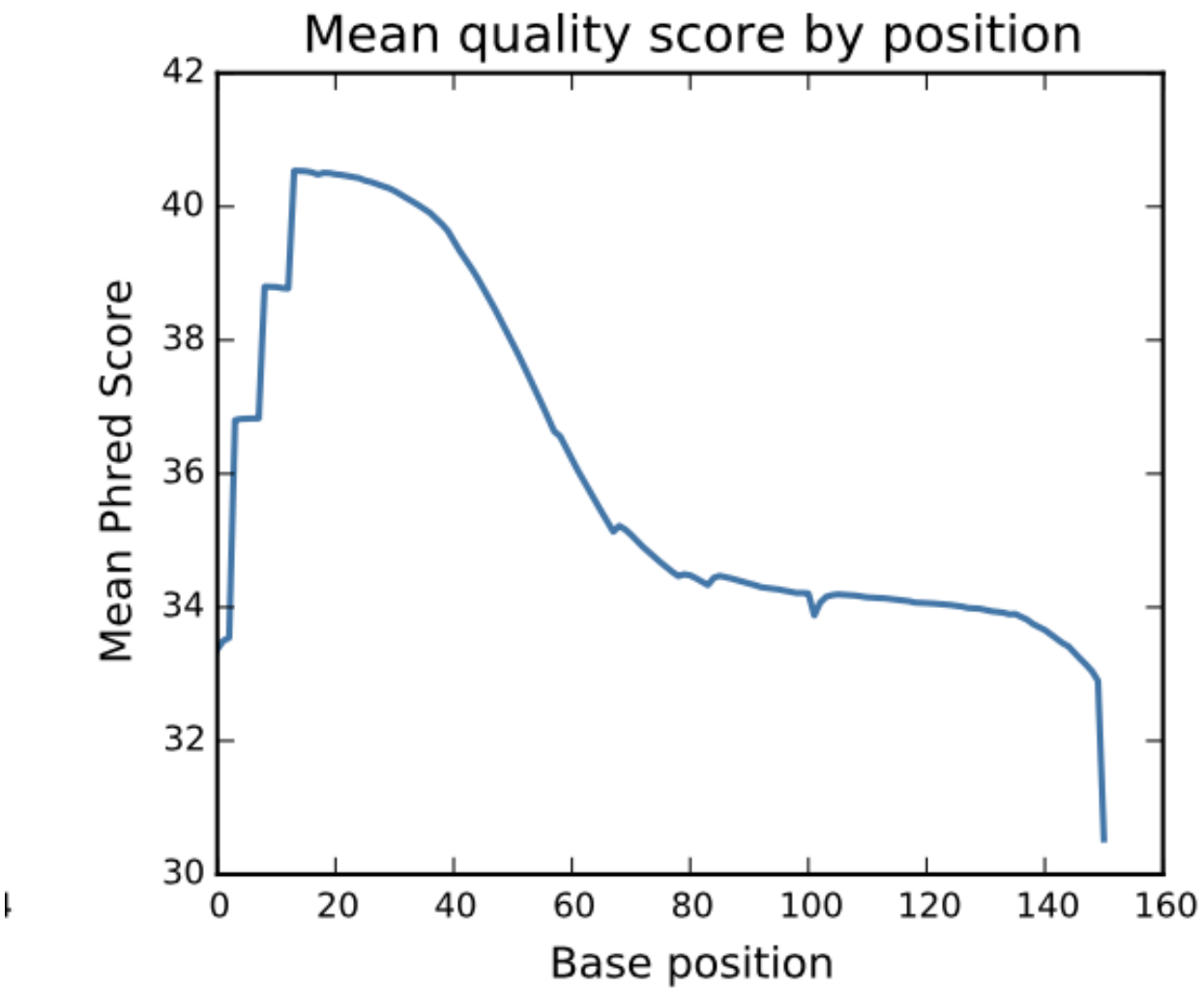
Metagenomes Reads features

Data exploration (i.e **SGA Preqc**)

Per-base error rates
Sequence coverage
Repeat-content
Metagenome size,

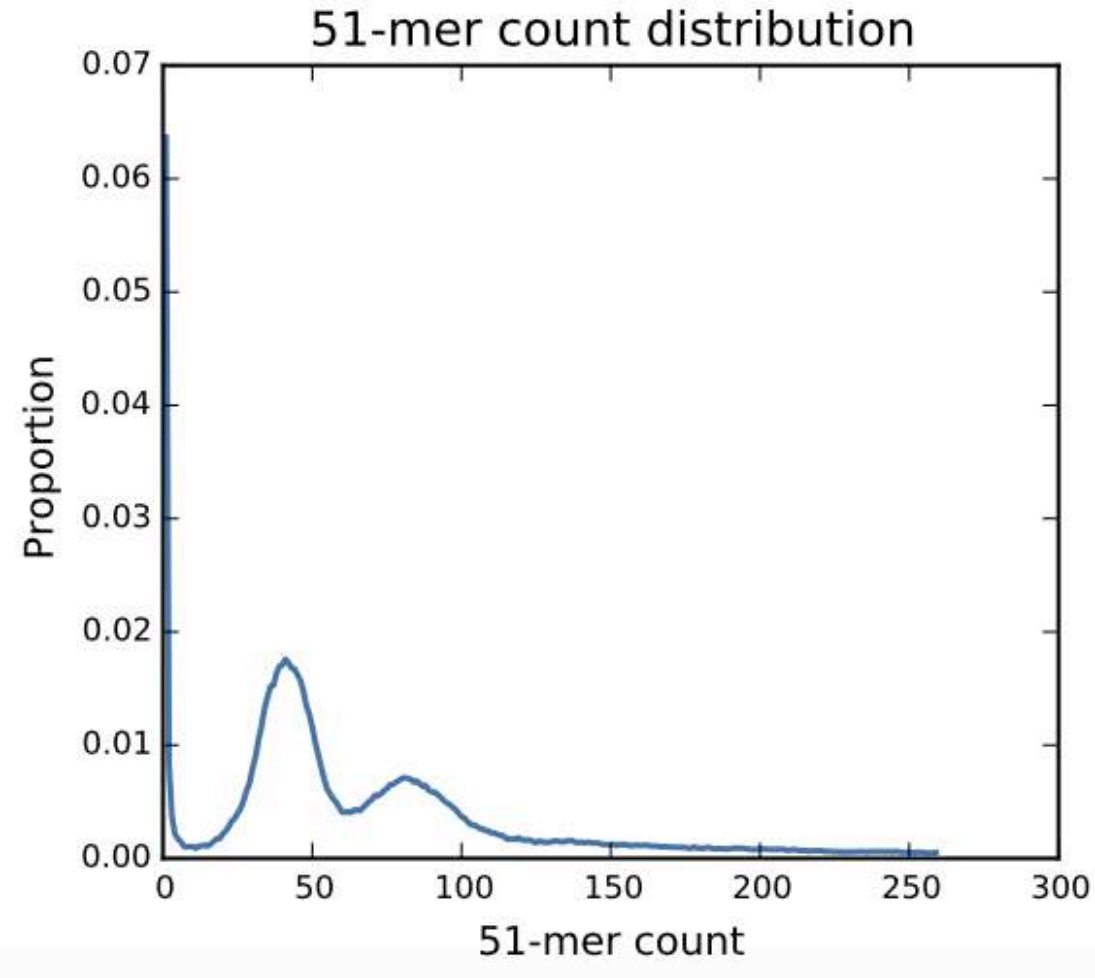
Before filtering…

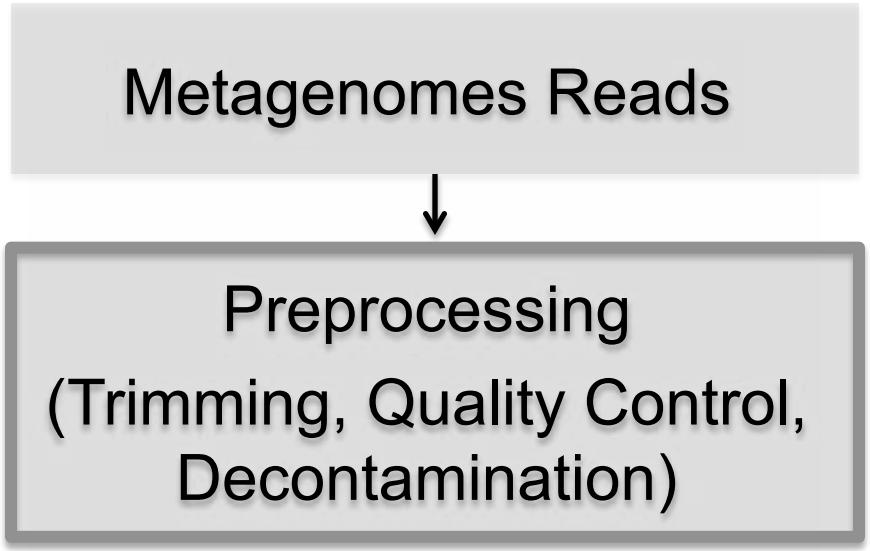
Metagenomes Reads features

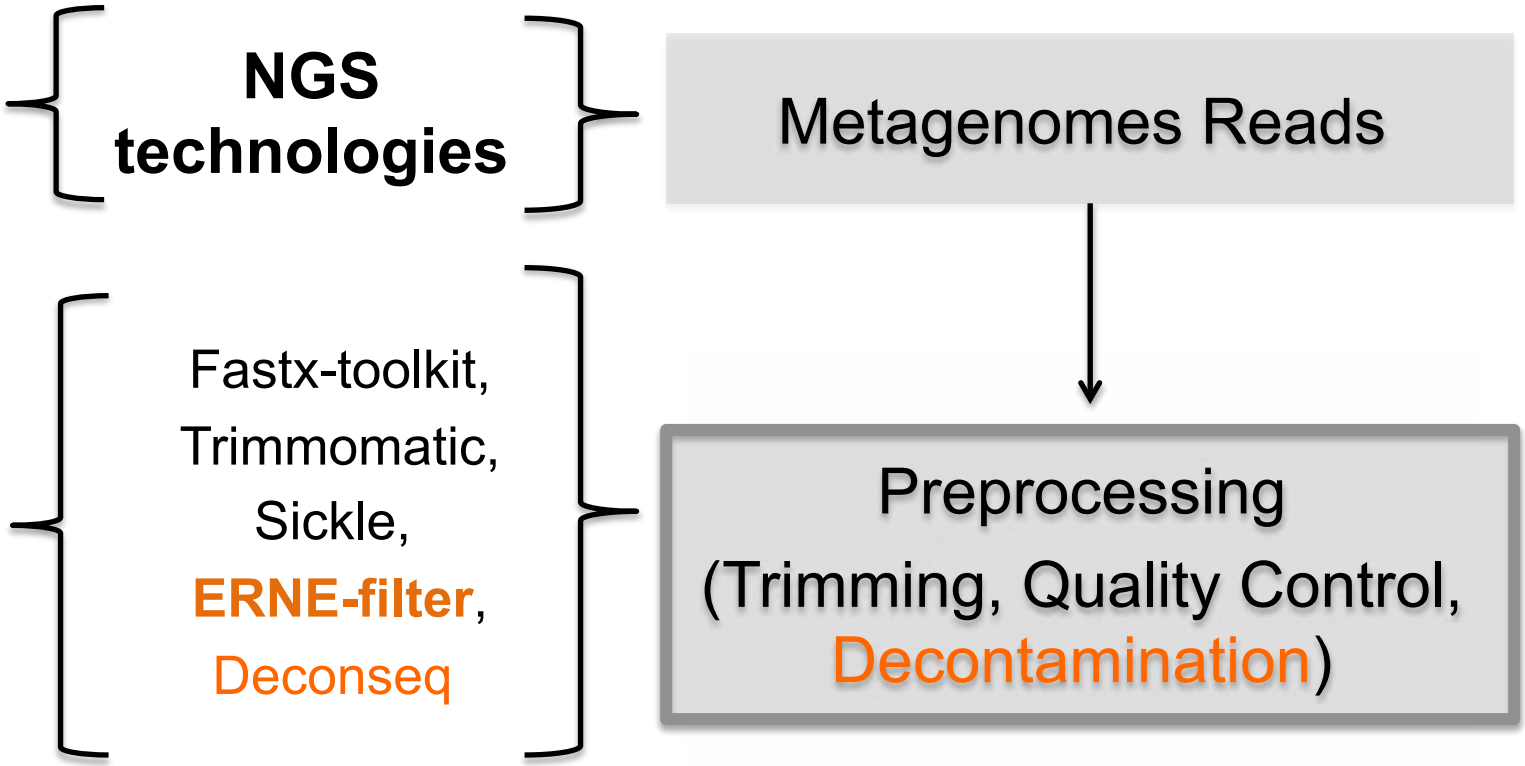


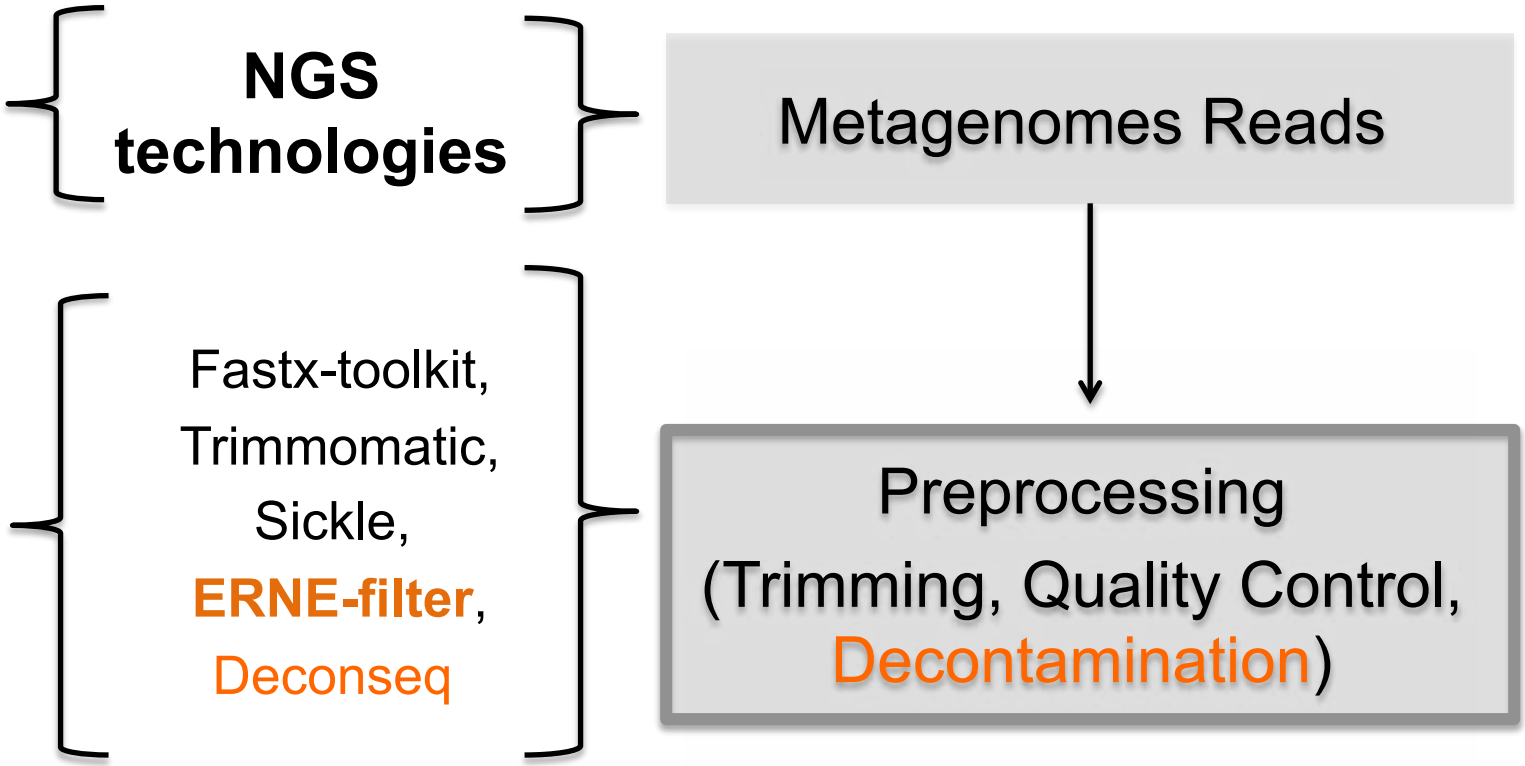
Before filtering…

Metagenomes Reads features

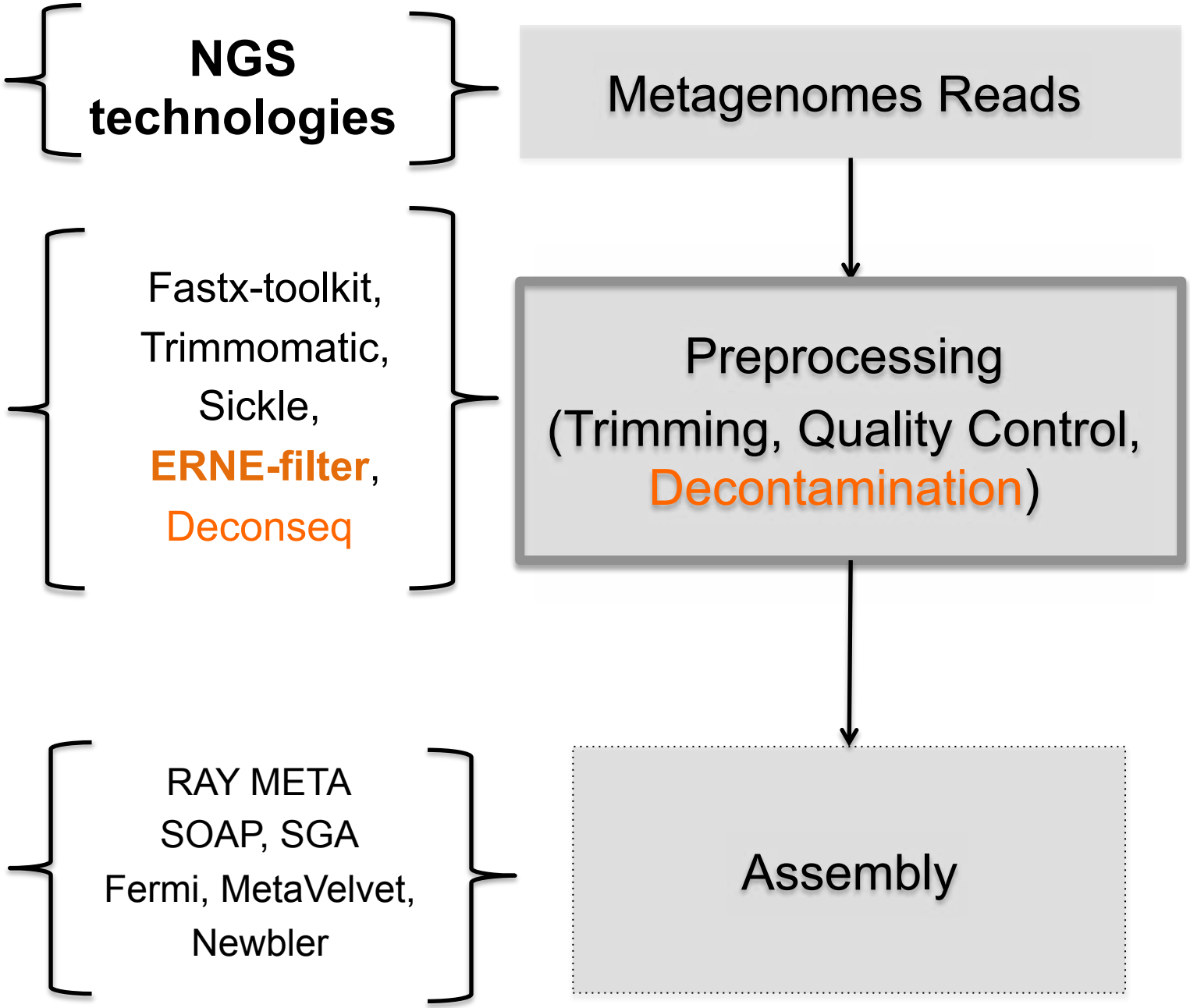


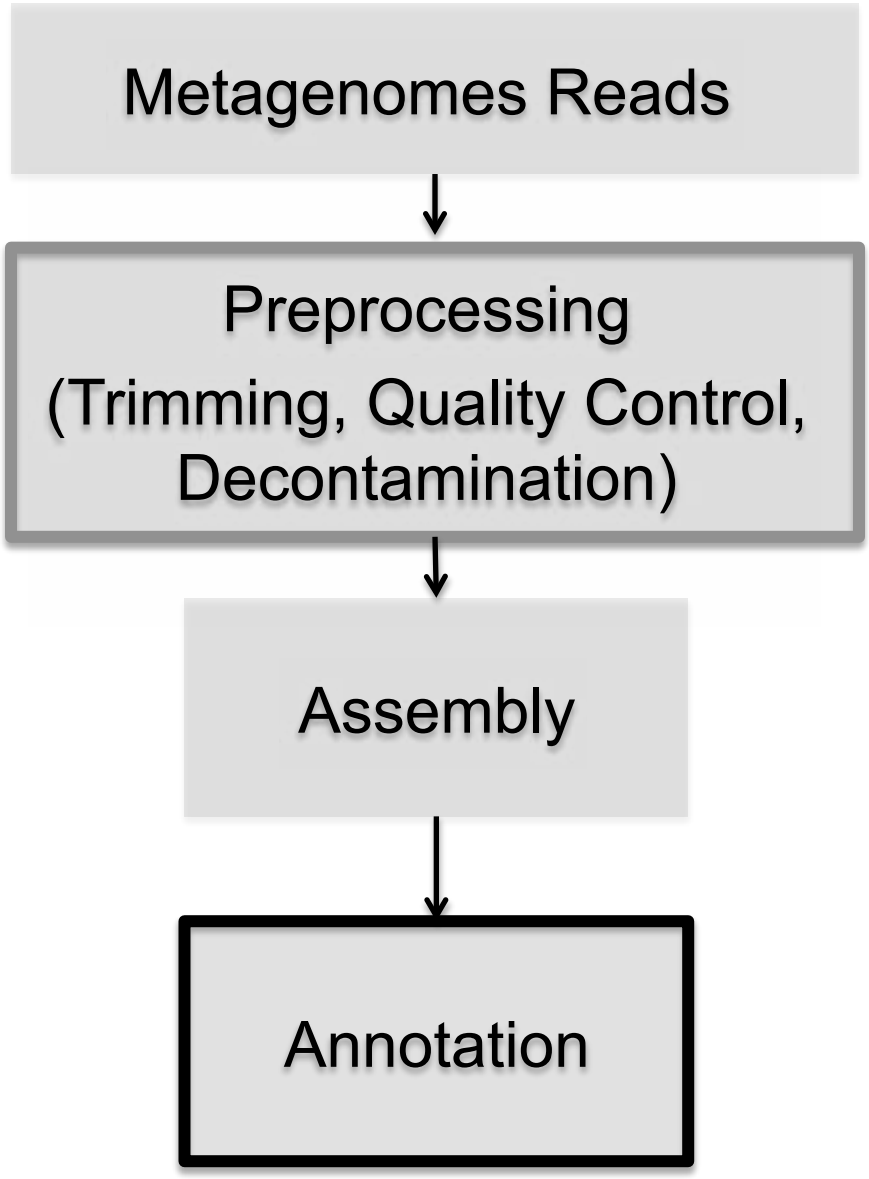

Metagenomes Reads
Preprocessing
(Trimming, Quality Control, Decontamination)


NGS technologies
Metagenomes Reads
Fastx-toolkit,
Trimmomatic,
Sickle,
ERNE-filter,
Deconseq
Preprocessing
(Trimming, Quality Control,
Decontamination)

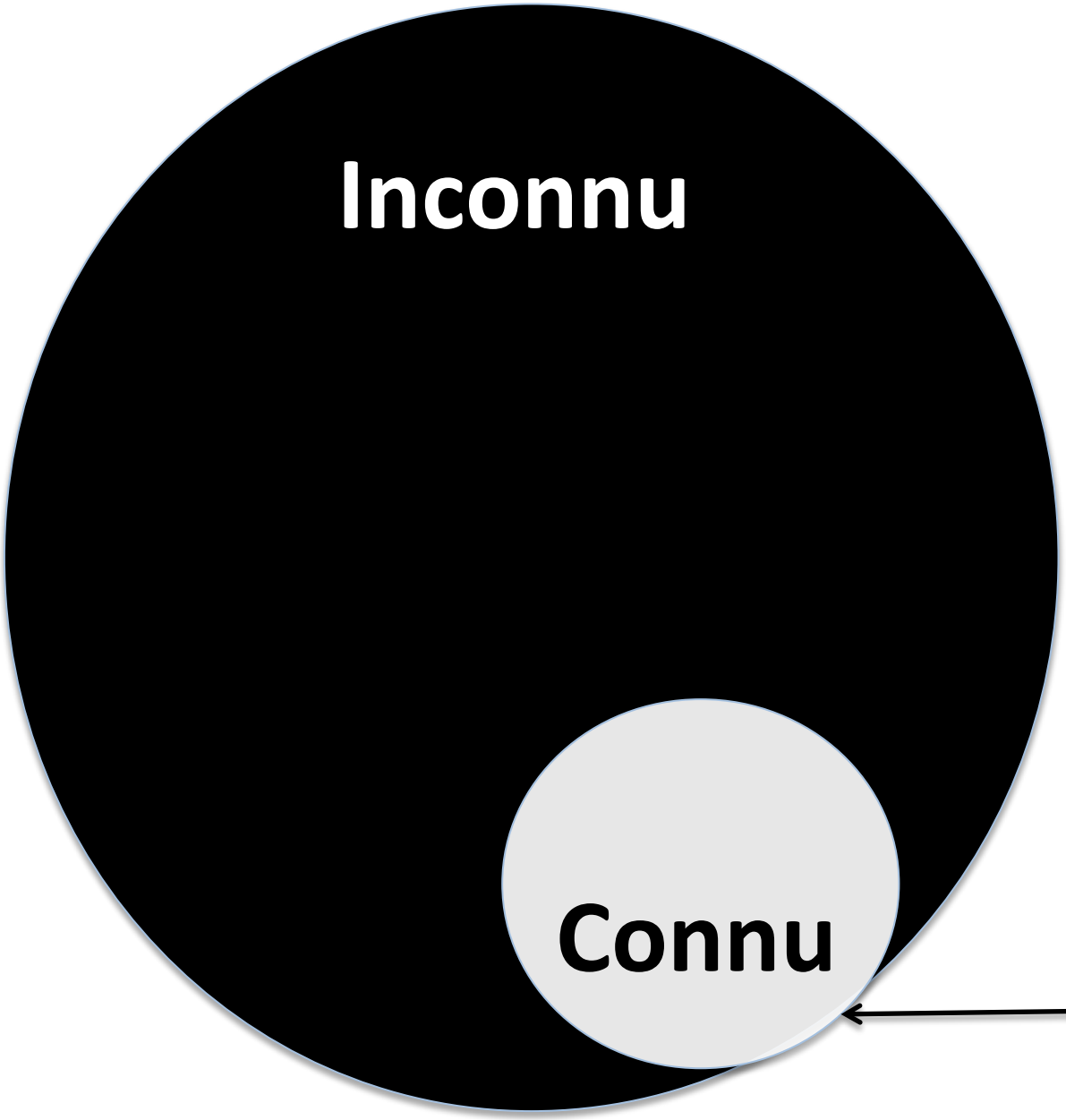


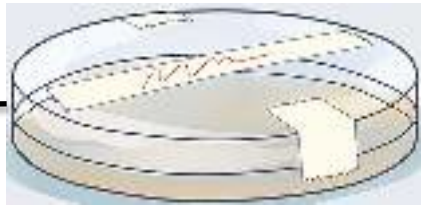
- **Fastx-toolkit** for Short-Reads FASTA/FASTQ (Hannon Lab)
- **Trimmomatic** (Bolger et al 2014, Bioinformatics)
- **Sickle (**Joshi NA and Fass JN. , 2011)
- **ERNE** (Extended Randomized Numerical alignEr) (Fabbro et al, 2013)
- **Deconseq** (DECONtamination of SEQuence data) (Schmeider and Edwards, 2011)


NGS technologies
Metagenomes Reads
Fastx-toolkit,
Trimmomatic,
Sickle,
ERNE-filter,
Deconseq
Preprocessing
(Trimming, Quality Control,
Decontamination)
RAY META
SOAP, SGA
Fermi, MetaVelvet,
Newbler
Assembly

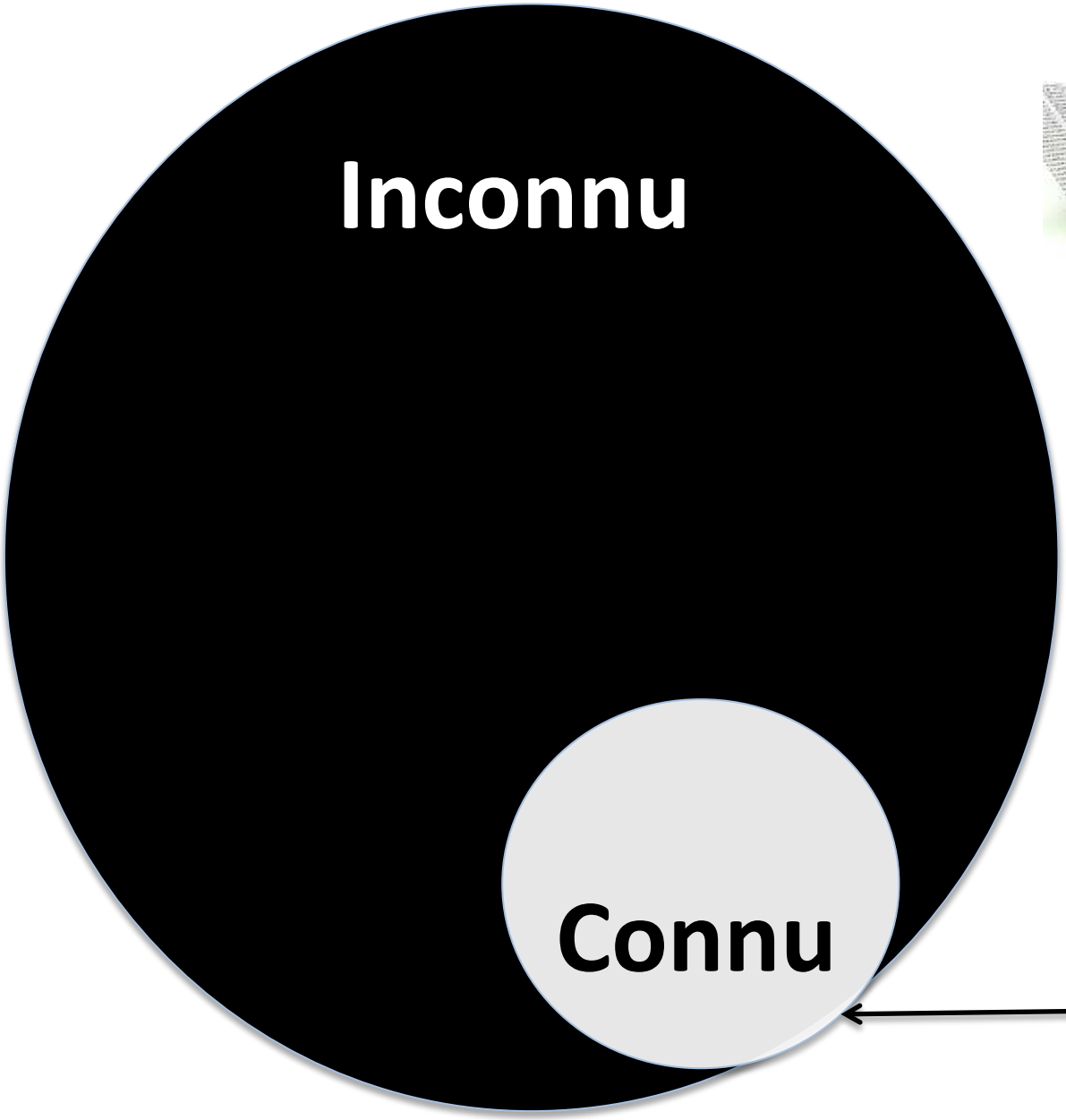

Metagenomes Reads
Preprocessing
(Trimming, Quality Control, Decontamination)
Assembly
Annotation

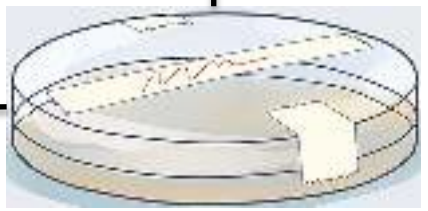
Retrouver le connu et prédire l'inconnu





Retrouver le connu et prédire l'inconnu





Sequence/function/structure

Sequence/function/structure

Generalist

Specialist

Curated/ verified

Not verified/ curated

Curated/ verified

Not verified/ curated

SwissProt
…

GeneBank
EMBL
…

SEED
Model organisms
Databases (FlyBase)

Pfam
ProDom
…

Local databases
…

Functional classification

- **The gene/protein family approach or the Clusters of Orthologous Groups(COG)s, (clustering algorithms based similarity)**
- **The Subsystem approach to genome annotation (Overbeek et al 2005)**
- **Bio-ontology (rarely used in metagenomics)**

Nucleic Acids Research, 2014 1
doi: 10.1093/nar/gku702

FOAM (Functional Ontology Assignments for Metagenomes): a Hidden Markov Model (HMM) database with environmental focus

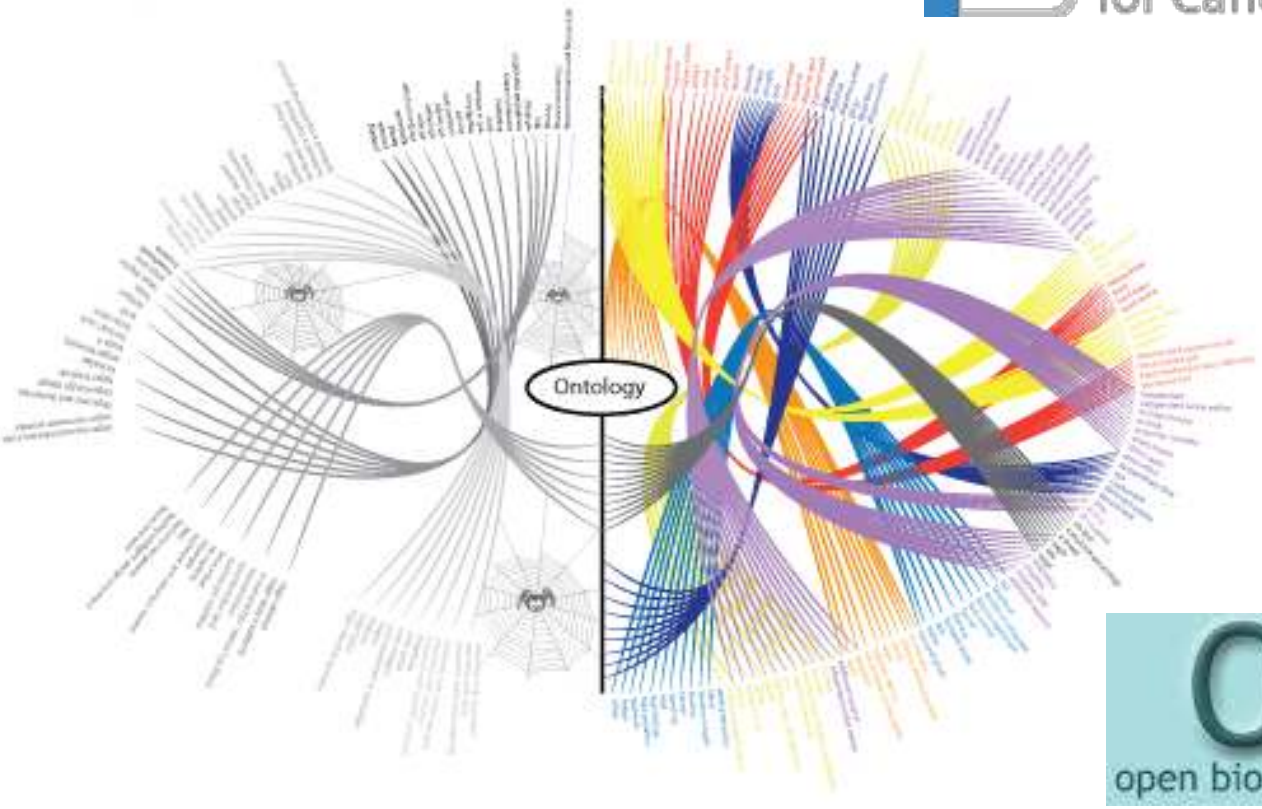
Emmanuel Prestat[1,2,†], Maude M. David[1,†], Jenni Hultman[1], Neslihan Taş[1], Regina Lamendella[1], Jill Dvornik[1], Rachel Mackelprang[1,3], David D. Myrold[4], Ari Jumpponen[2], Susannah G. Tringe[3], Elizabeth Holman[1], Konstantinos Mavromatis[3] and Janet K. Jansson[1,3,5,6,7,*]

http://portal.nersc.gov/project/m1317/FOAM/

Bio-onthologies

GENEONTOLOGY
Unifying Biology

cBioPortal
for Cancer Genomics



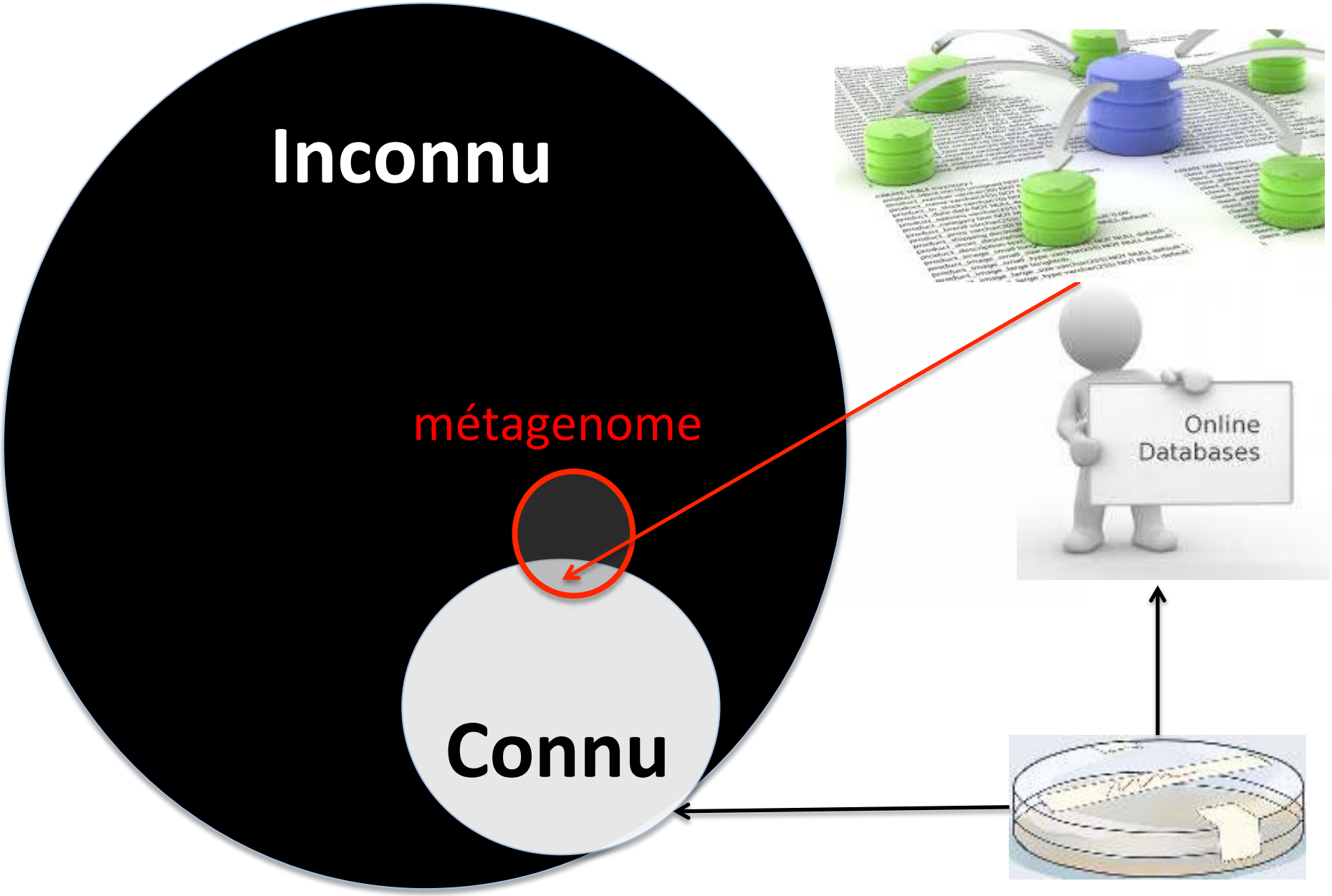
OBO
open biomedical ontologies

BioPAX@
Biological Pathways Exchange

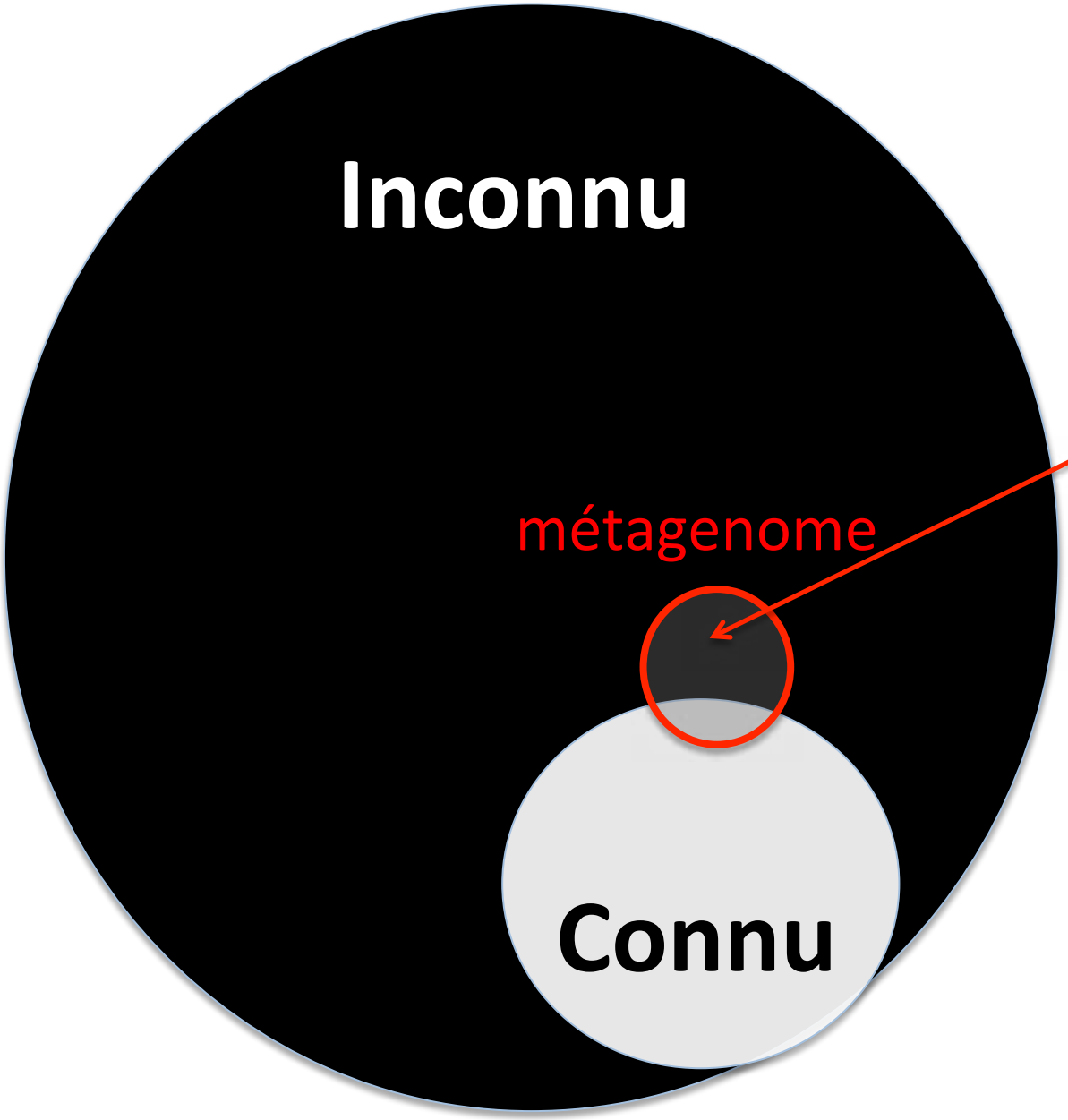
VCELL

THE NATIONAL CENTER FOR
BIOMEDICAL ONTOLOGY

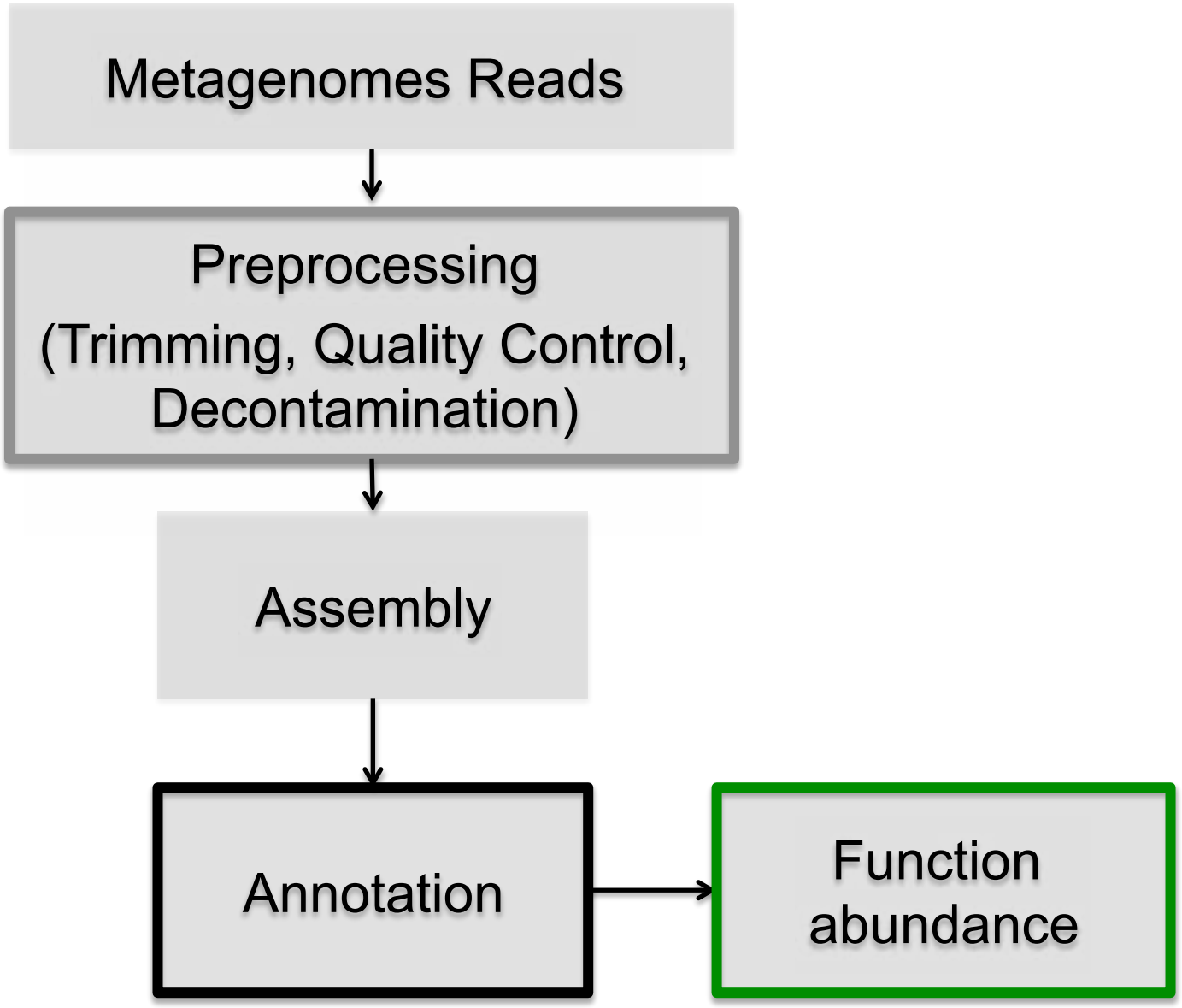
Retrouver le connu et prédire l'inconnu

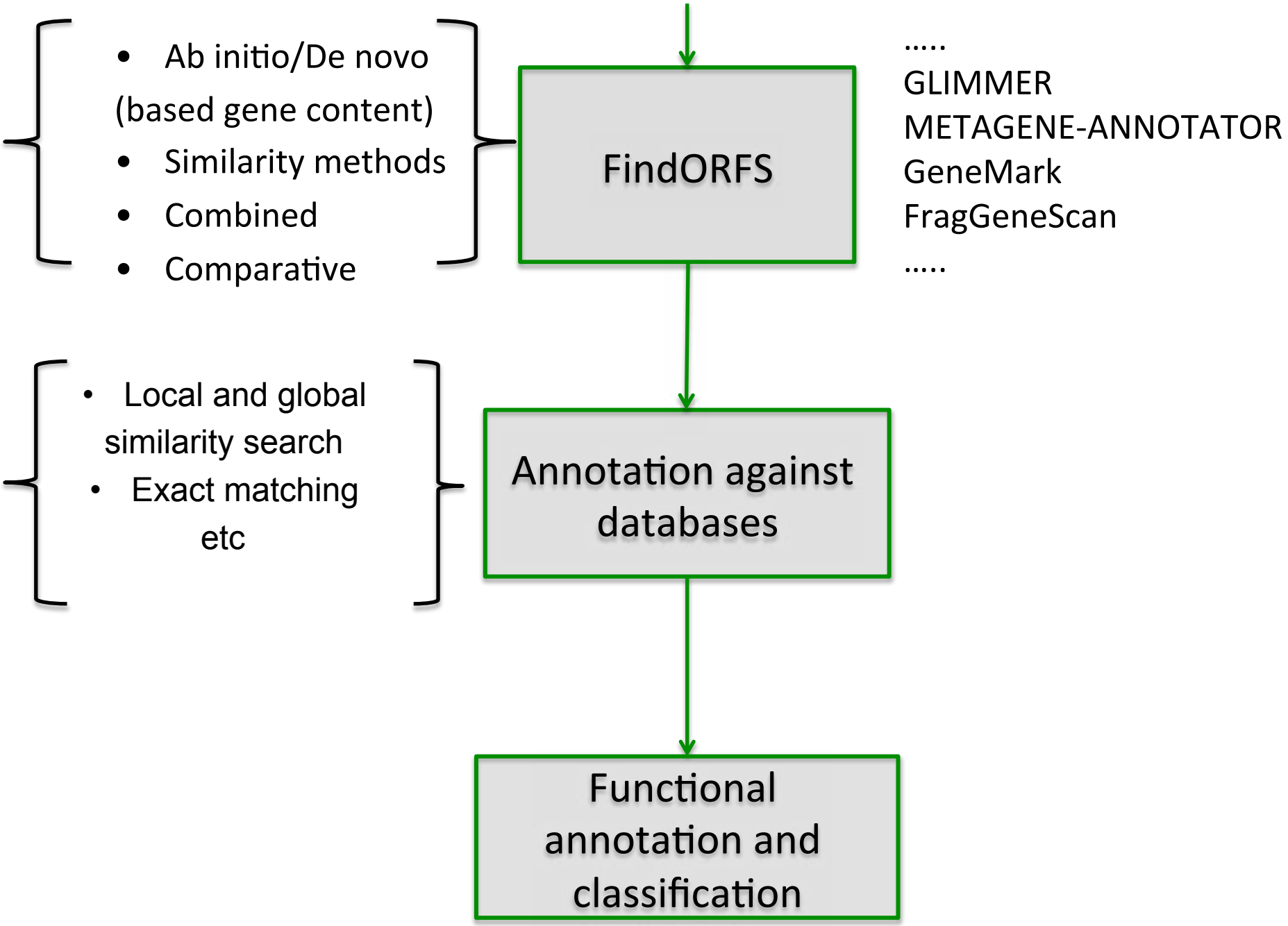


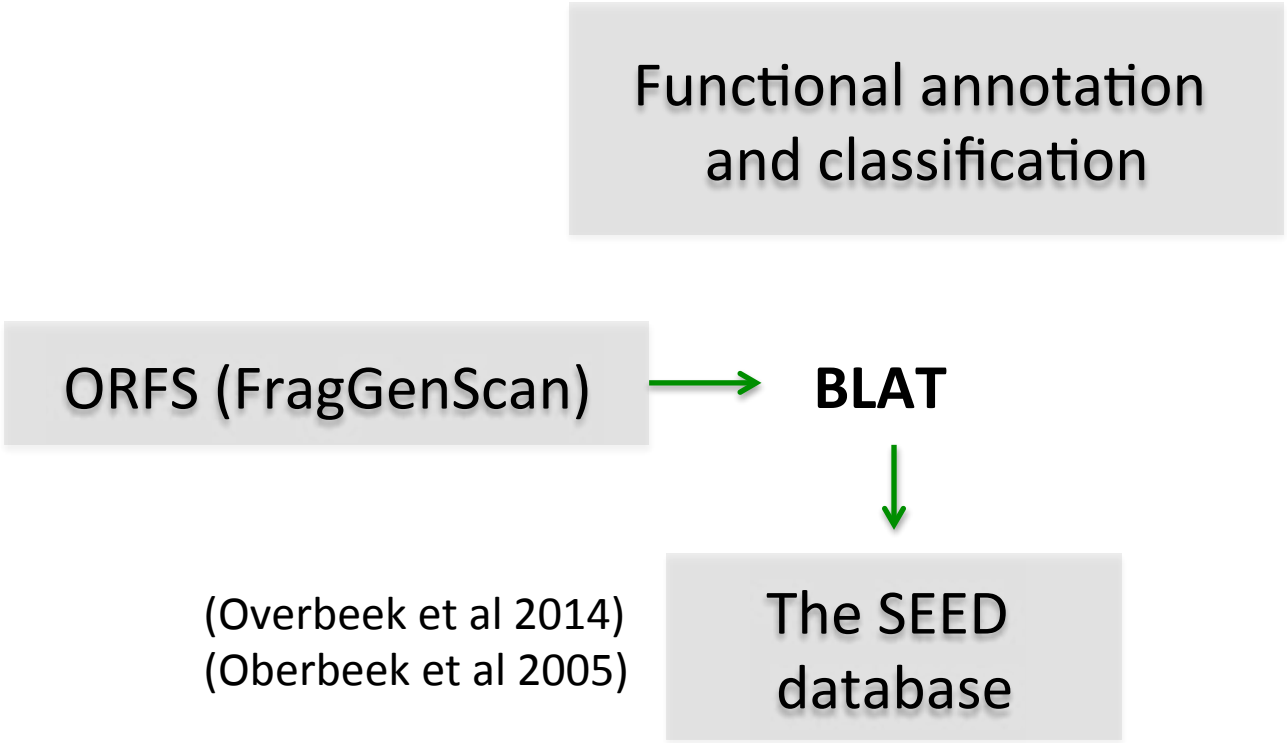
Retrouver le connu et prédire l'inconnu



- Importance de l'algorithmique de graphes
- Assemblage des nouveaux génomes
- Découverte des nouvelles organismes


Metagenomes Reads
Preprocessing
(Trimming, Quality Control, Decontamination)
Assembly
Annotation
Function abundance


• Ab initio/De novo (based gene content)
• Similarity methods
• Combined
• Comparative
FindORFS
.....
GLIMMER
METAGENE-ANNOTATOR
GeneMark
FragGeneScan
.....
• Local and global similarity search
• Exact matching
etc
Annotation against databases
Functional annotation and classification


Functional annotation and classification
ORFS (FragGenScan)
BLAT
The SEED database
(Overbeek et al 2014)
(Oberbeek et al 2005)

Functional annotation and classification

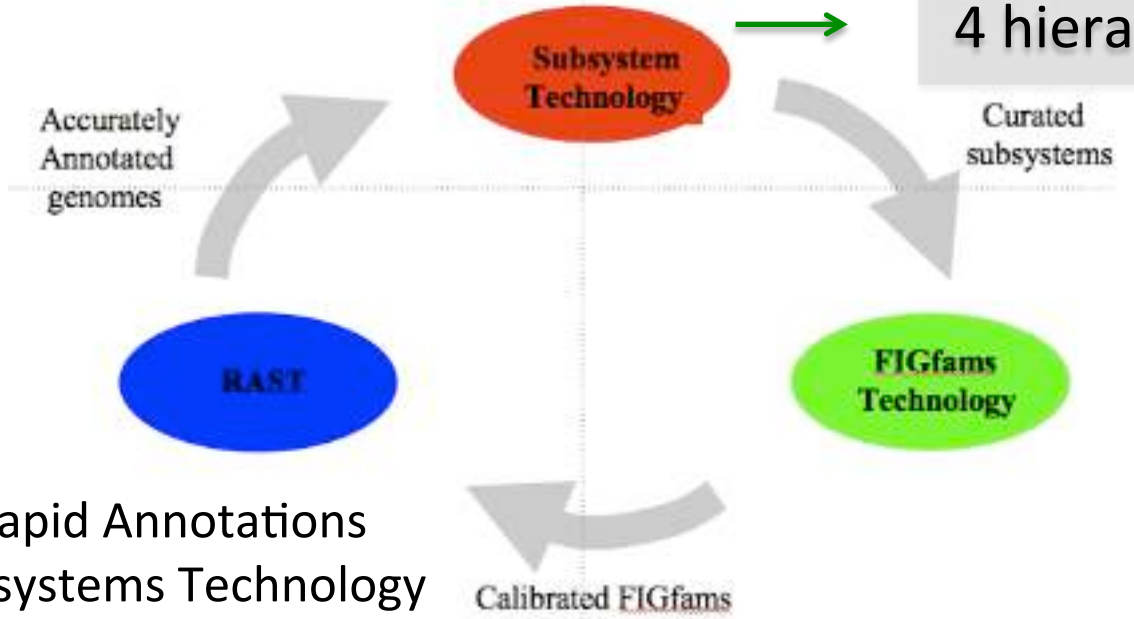
ORFS (FragGenScan) → **BLAT/BLAST**

BLAT/BLAST → The SEED database

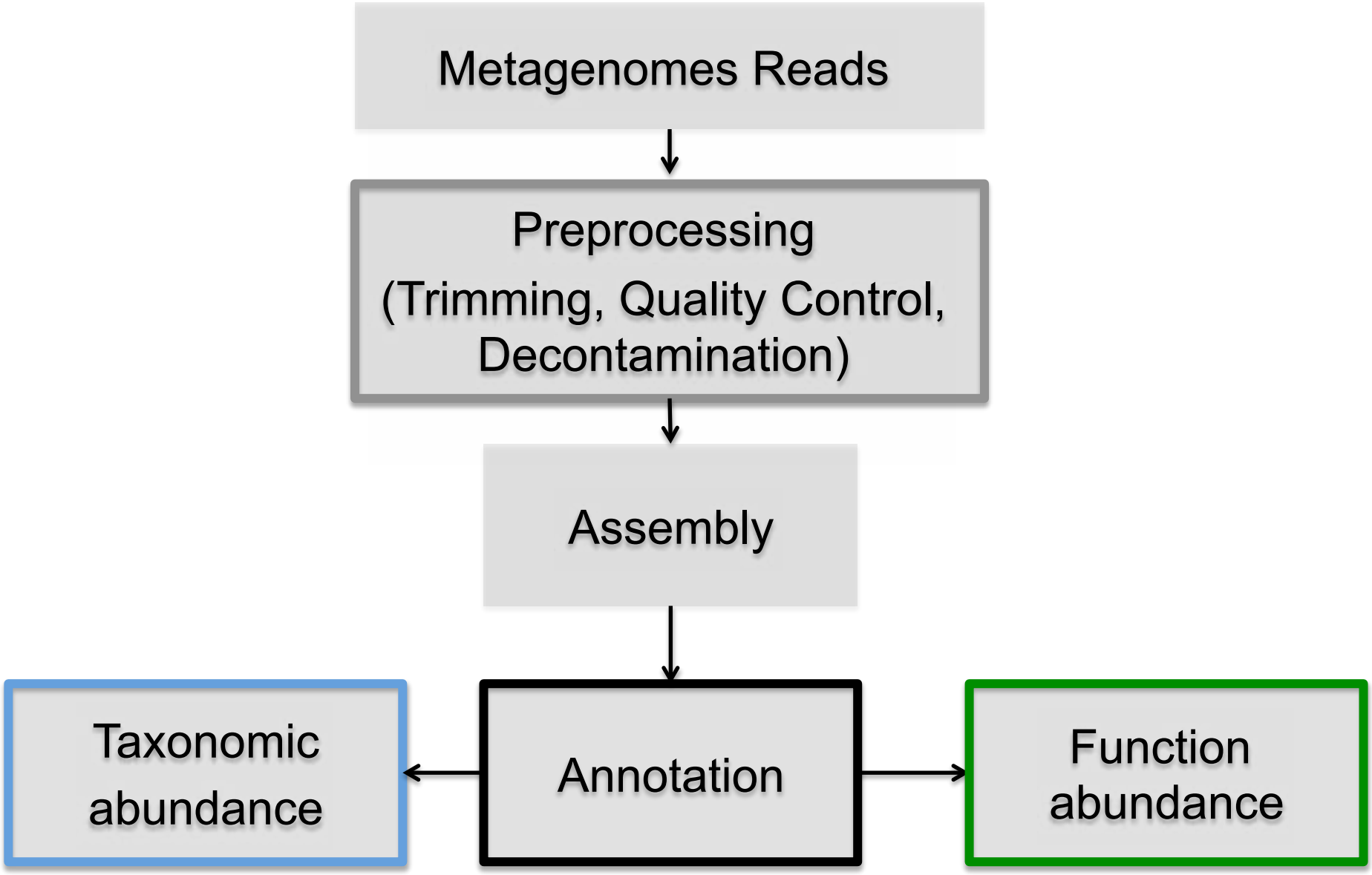
(Overbeek et al 2014)
(Oberbeek et al 2005)

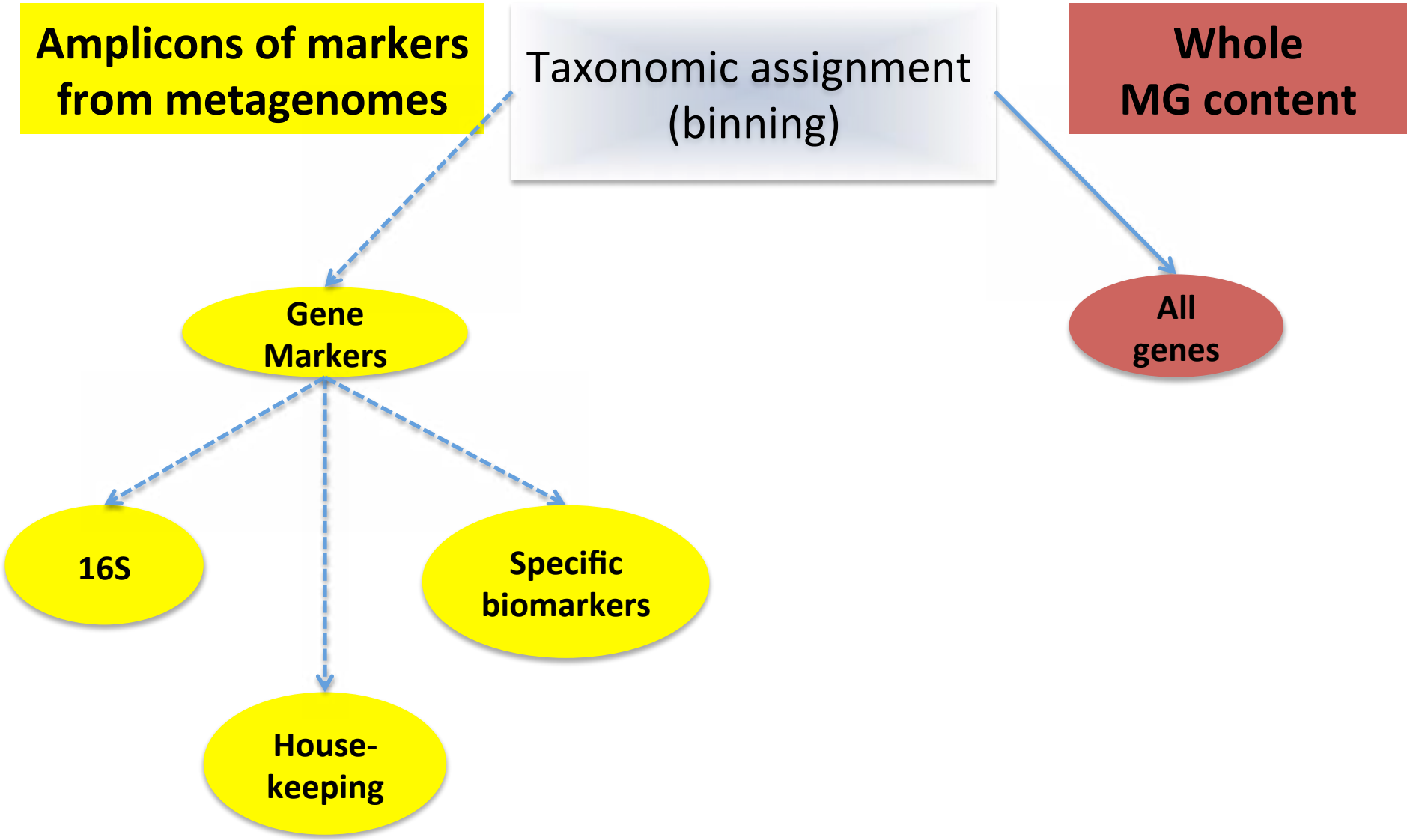
The SEED database → 4 hierarchical levels

4 hierarchical levels → Level 1 (27 categories)



RAST, Rapid Annotations using Subsystems Technology


Metagenomes Reads
Preprocessing
(Trimming, Quality Control, Decontamination)
Assembly
Taxonomic abundance
Annotation
Function abundance


Amplicons of markers from metagenomes
Taxonomic assignment (binning)
Whole MG content
Gene Markers
All genes
16S
Specific biomarkers
House-keeping

Taxonomic assignment (binning)

Whole MG content

All genes

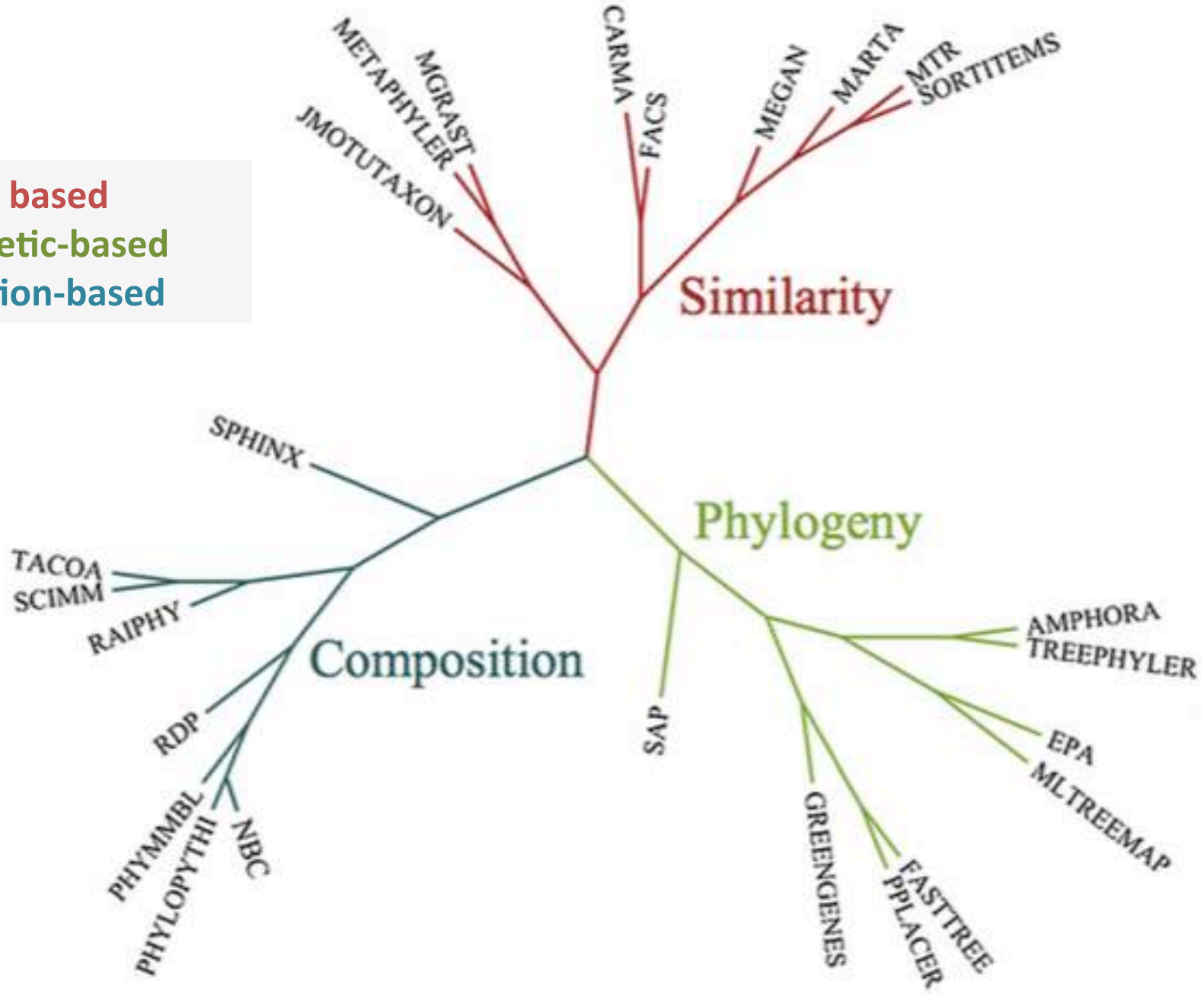
LCA (Lowest Common Ancestor)

Best hit

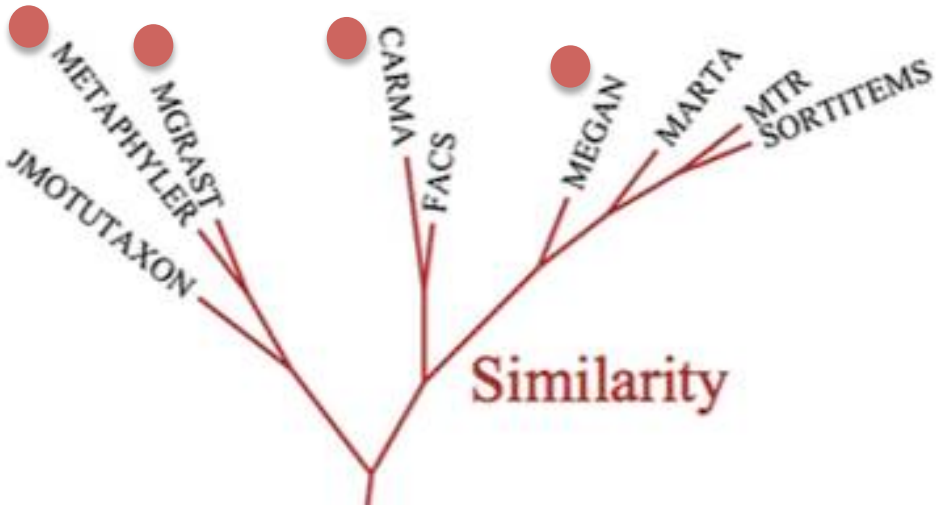
Representative hit

Assignement des taxons

Similarity based
Phylogenetic-based
Composition-based



Bazinet and Cummings BMC Bioinformatics 2012



Bazinet and Cummings, 2012

Program	FACS 269 bp	MetaPhyler 300 bp	CARMA 265 bp	PhyloPythia 961 bp	Mean
Percentage of sequence classified					
CARMA	29.0	93.6	68.7	61.3	63.2
MEGAN	48.4	88.2	90.5	62.2	72.3
MetaPhyler	0.2	80.9	0.5	0.6	20.6
MG-RAST	27.1	29.8	80.2	70.5	51.9
Sensitivity (percentage)					
CARMA	26.7	93.4	68.5	59.8	62.1
MEGAN	42.5	87.9	90.3	61.0	70.4
MetaPhyler	0.1	80.7	0.5	0.5	20.5
MG-RAST	25.0	29.7	80.1	67.2	50.5
Precision (percentage)					
CARMA	92.0	99.7	99.7	97.4	97.2
MEGAN	78.1	99.7	99.8	98.1	93.9
MetaPhyler	84.0	99.7	100.0	83.8	91.9
MG-RAST	92.4	99.8	99.9	95.3	96.9

For Amplicons only, not metagenomes

Taxonomic assignment (binning)

Gene Markers

16S

Mothur (Uclust)

Qiime (Usearch)

OTU (Operational Taxonomic Unit) Clustering

OTUs Profiling based

16S gene markers

Alpha-Diversity estimation

Amplicons of markers from metagenomes

Taxonomic assignment (binning)

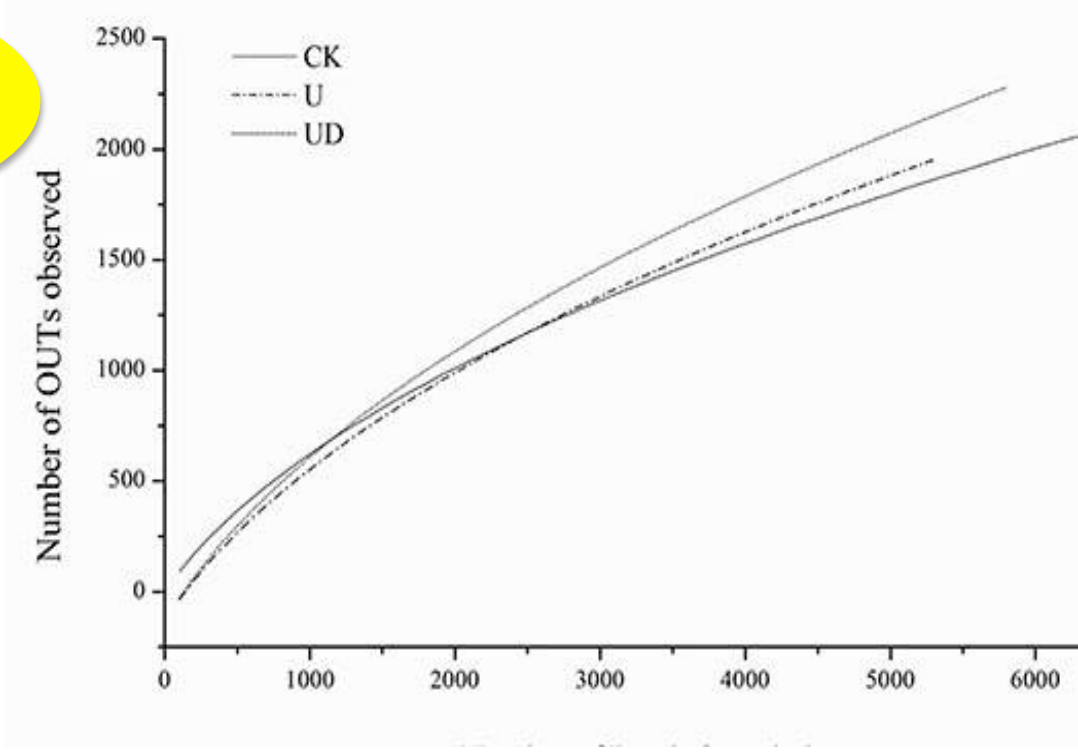
OTUs based Gene Markers

16S

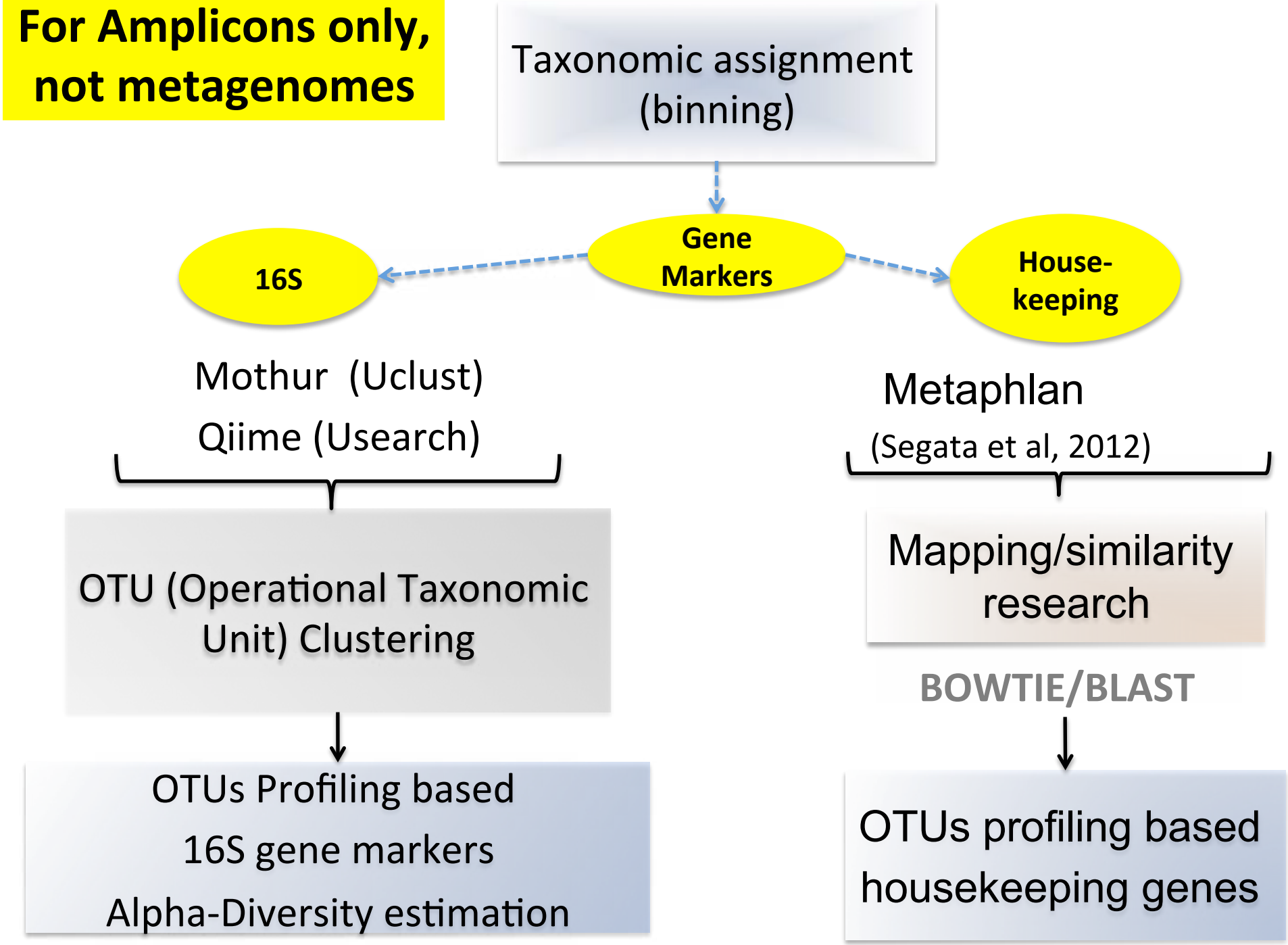
Specific biomarkers

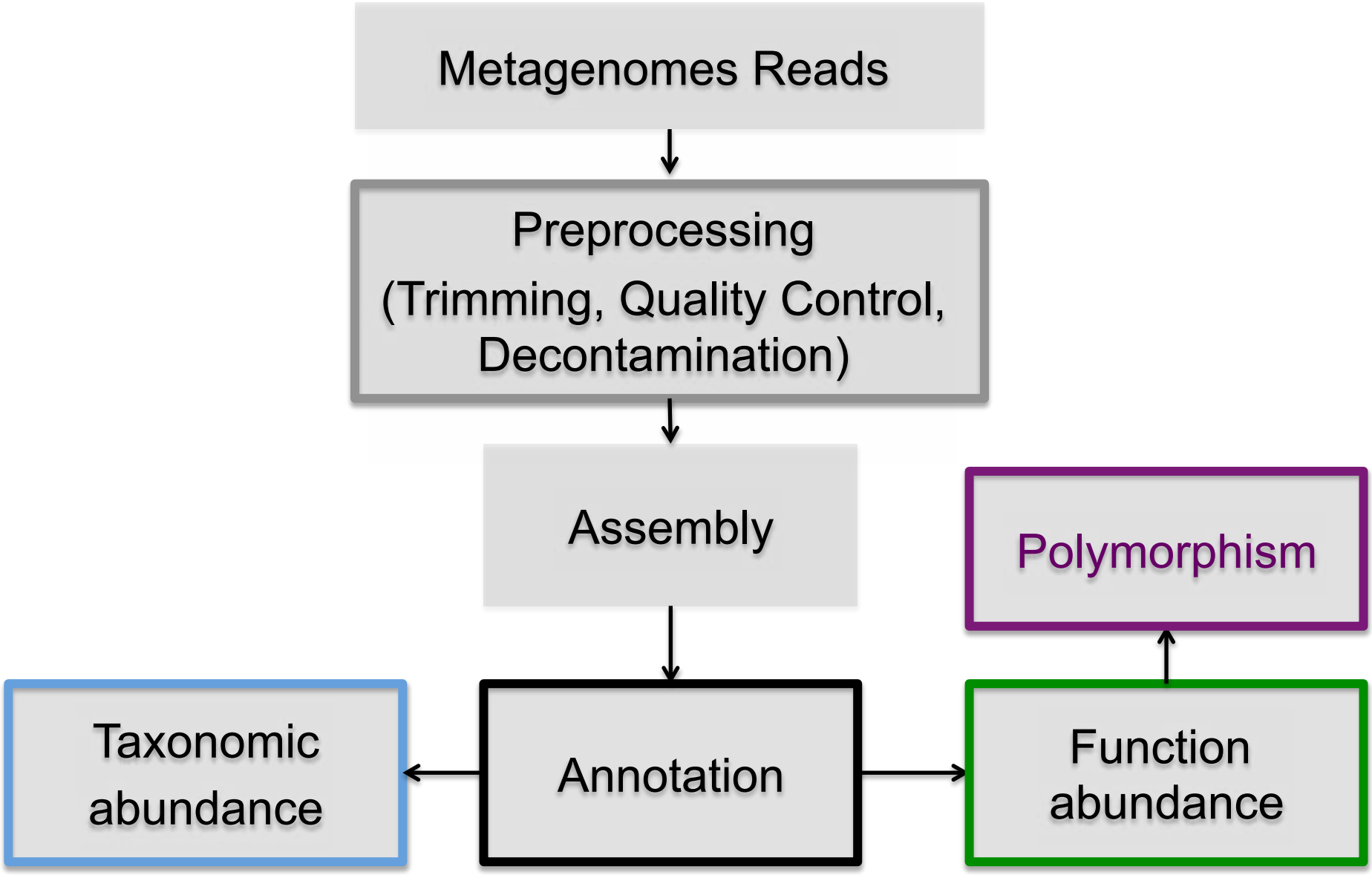
House-keeping

Biodiversity sampling ?
Rarefaction curves



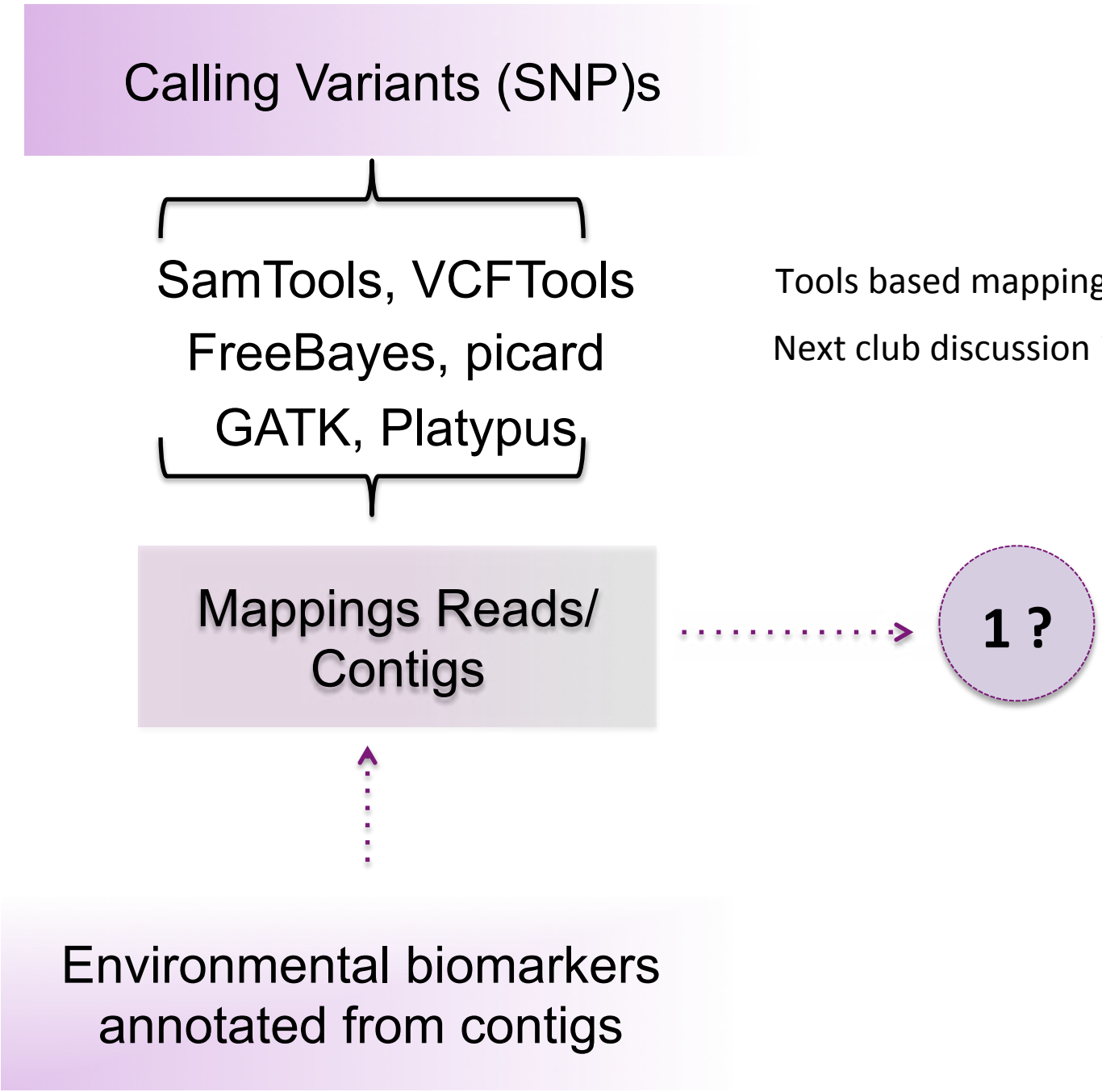
For Amplicons only, not metagenomes




Metagenomes Reads
Preprocessing
(Trimming, Quality Control, Decontamination)
Assembly
Polymorphism
Taxonomic abundance
Annotation
Function abundance

Calling Variants (SNP)s

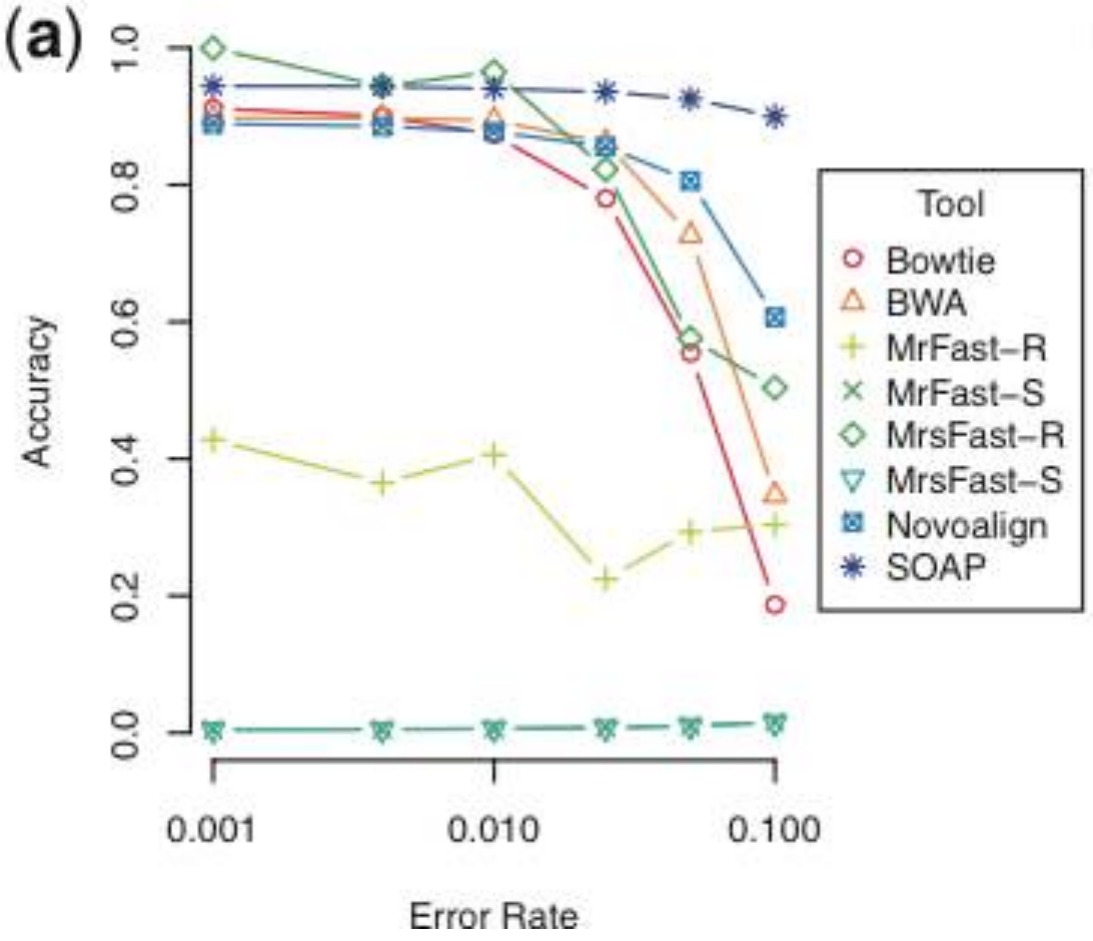
Environmental biomarkers annotated from contigs


Calling Variants (SNP)s
SamTools, VCFTools
FreeBayes, picard
GATK, Platypus
Tools based mapping
Next club discussion ?
Mappings Reads/
Contigs
1 ?
Environmental biomarkers
annotated from contigs

Mappings Reads/ Contigs

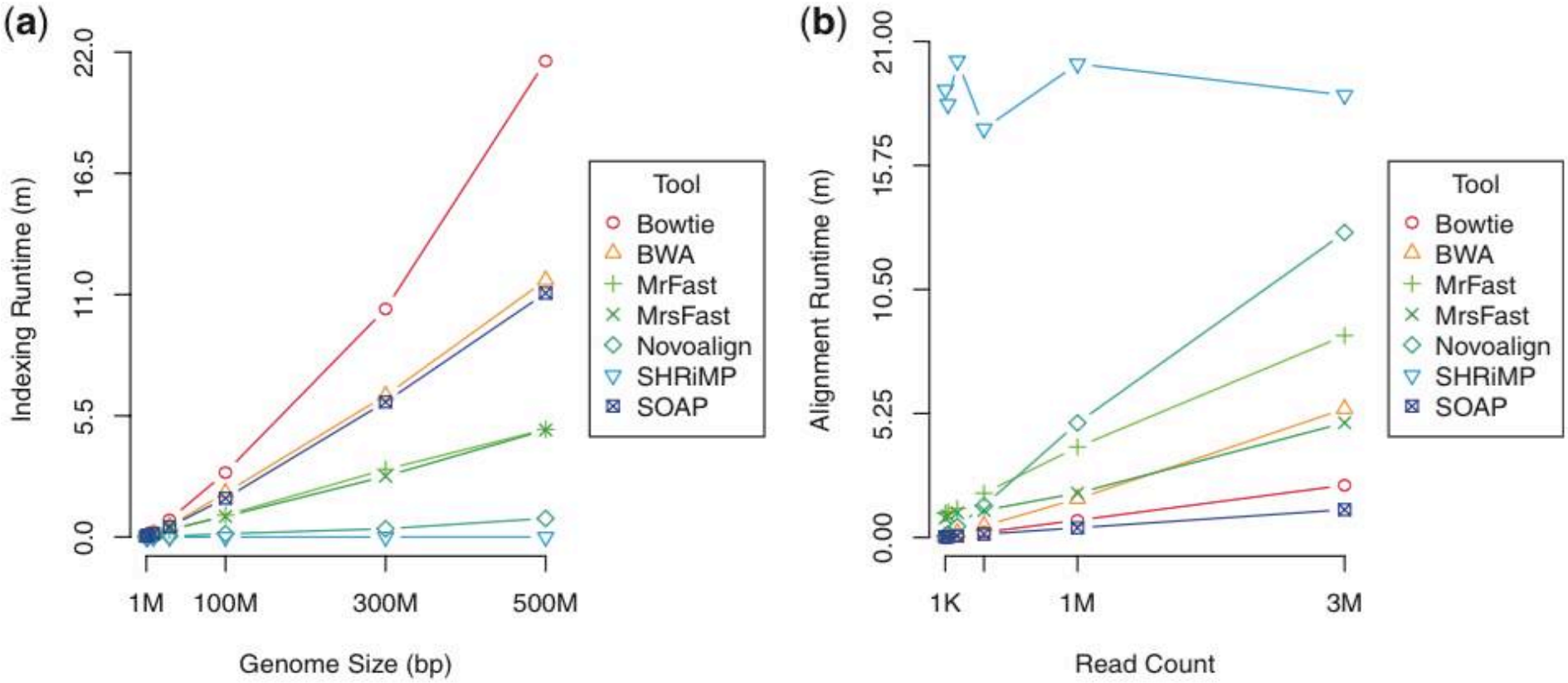
Tools Evaluation

SEAL evaluator : Seal is available as open source at http://compbio.case .edu/seal/

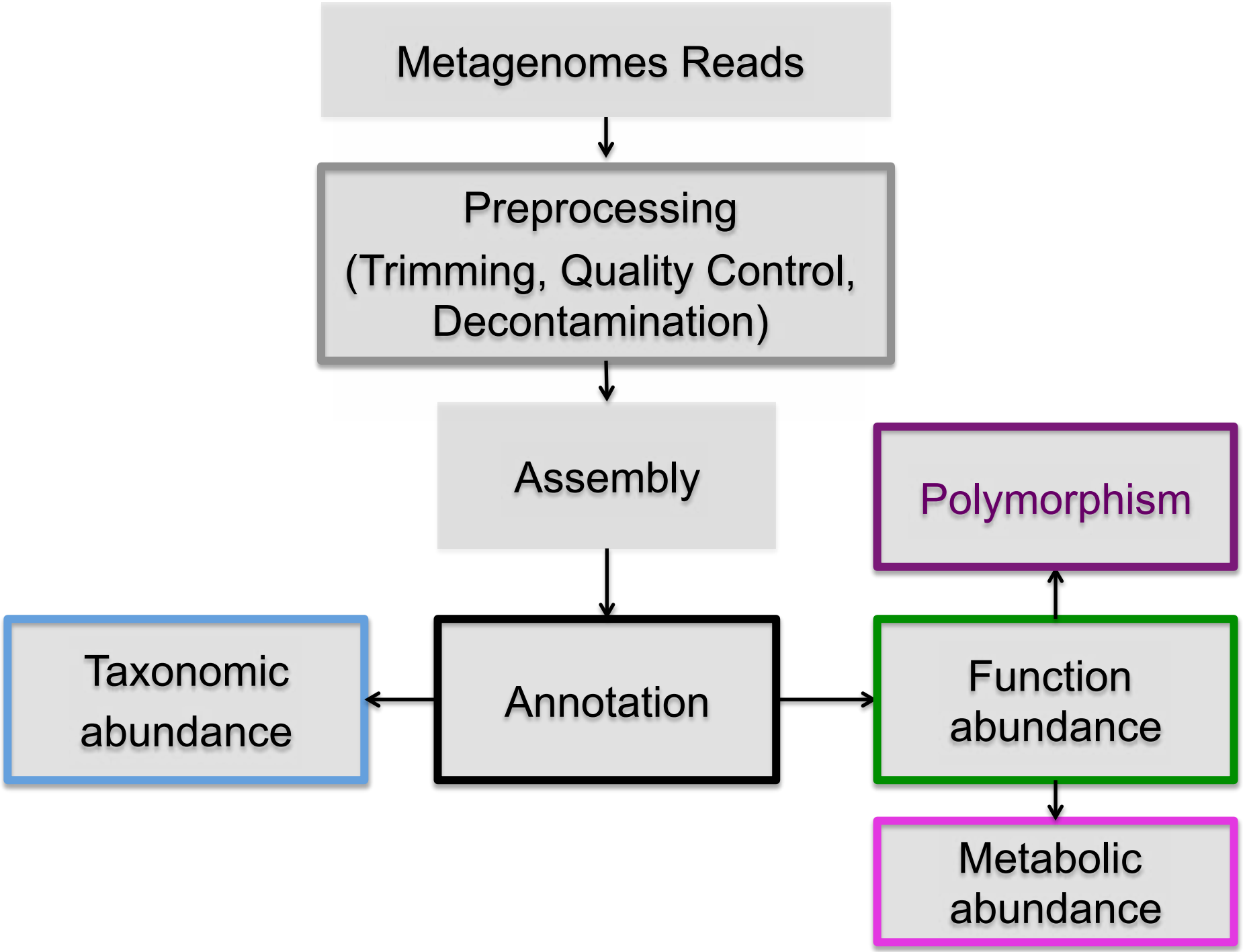


Rufallo et al, 2011

Indexing runtime versus Alignment Runtime



Rufallo et al, 2011

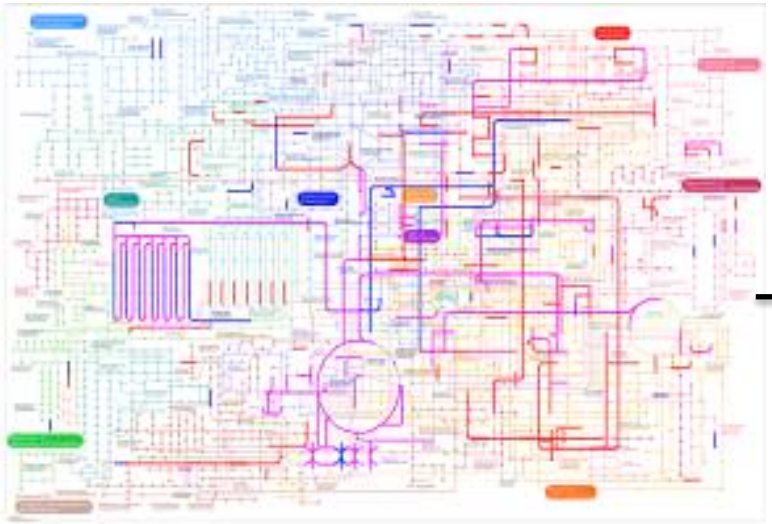

Metagenomes Reads
Preprocessing
(Trimming, Quality Control, Decontamination)
Assembly
Polymorphism
Taxonomic abundance
Annotation
Function abundance
Metabolic abundance

Functional metagenome annotations

↓

Mapping metabolic pathways

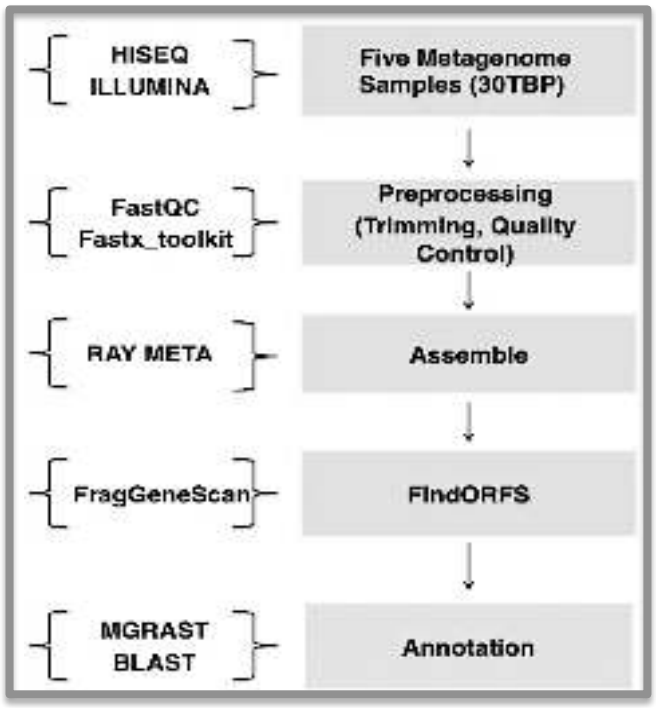
↓ KEGG MAP TOOL



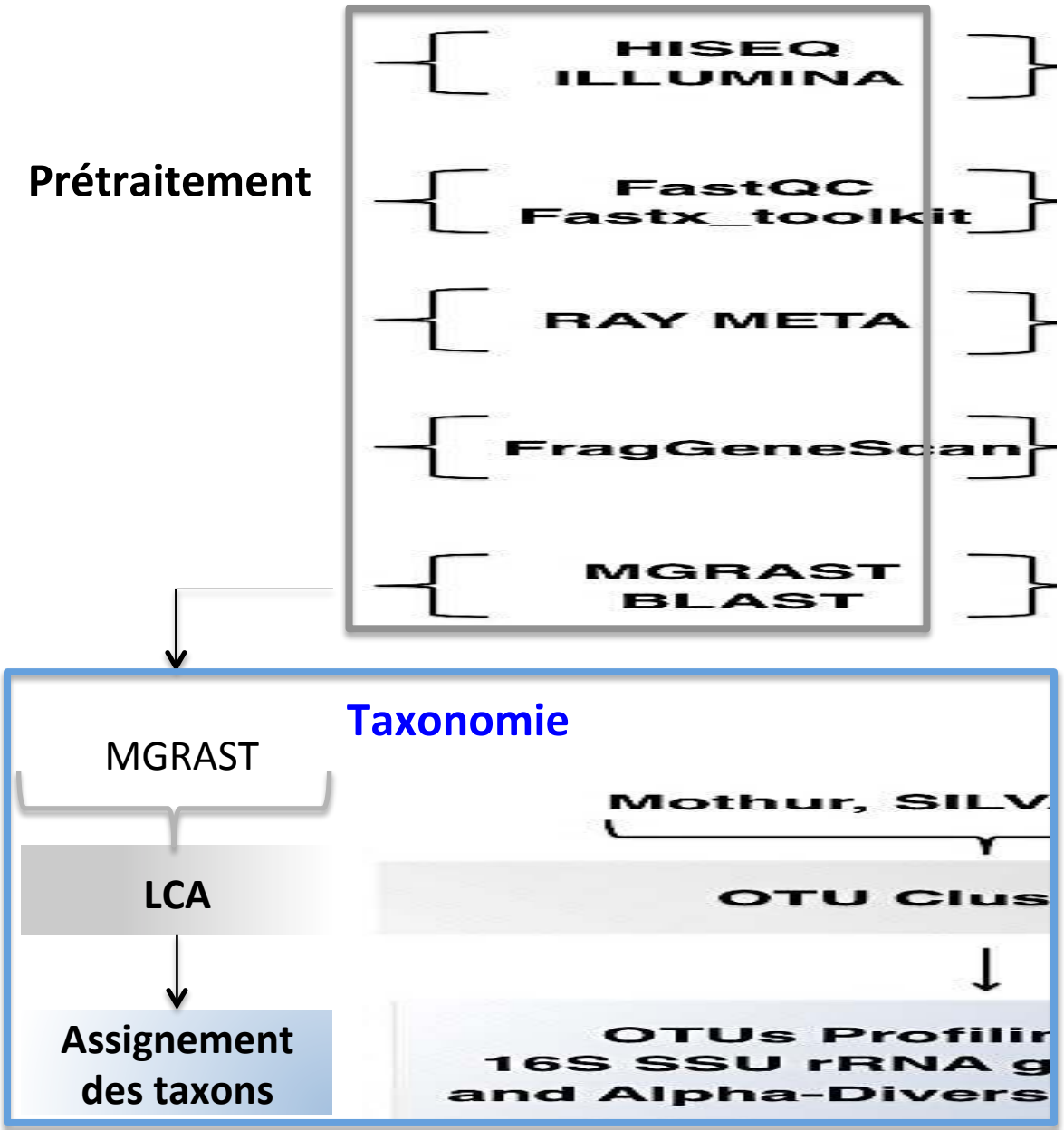
→ Abundance of pathways enzymes

Pipeline d'analyse de "reads" aux métabolites

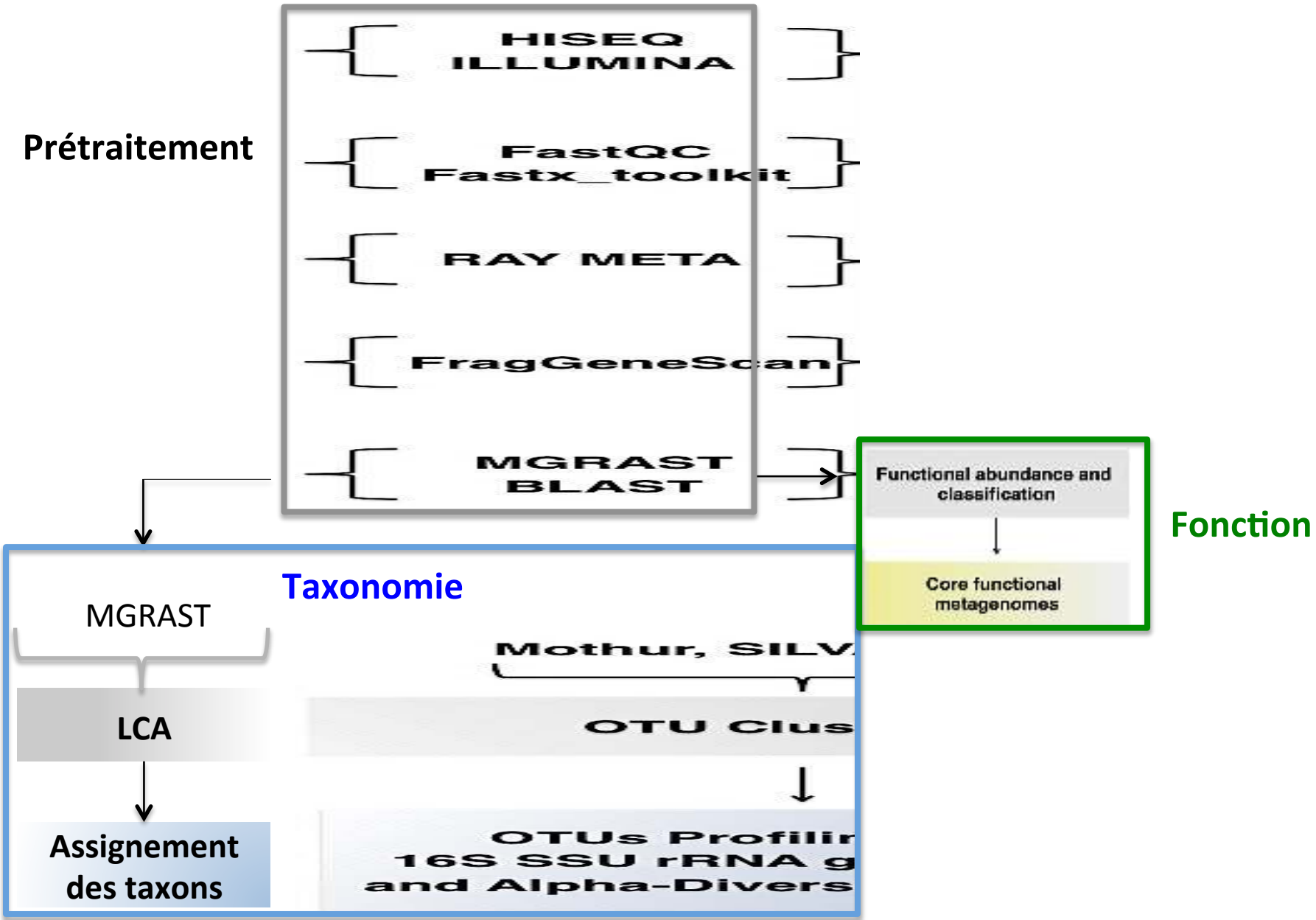
Prétraitement



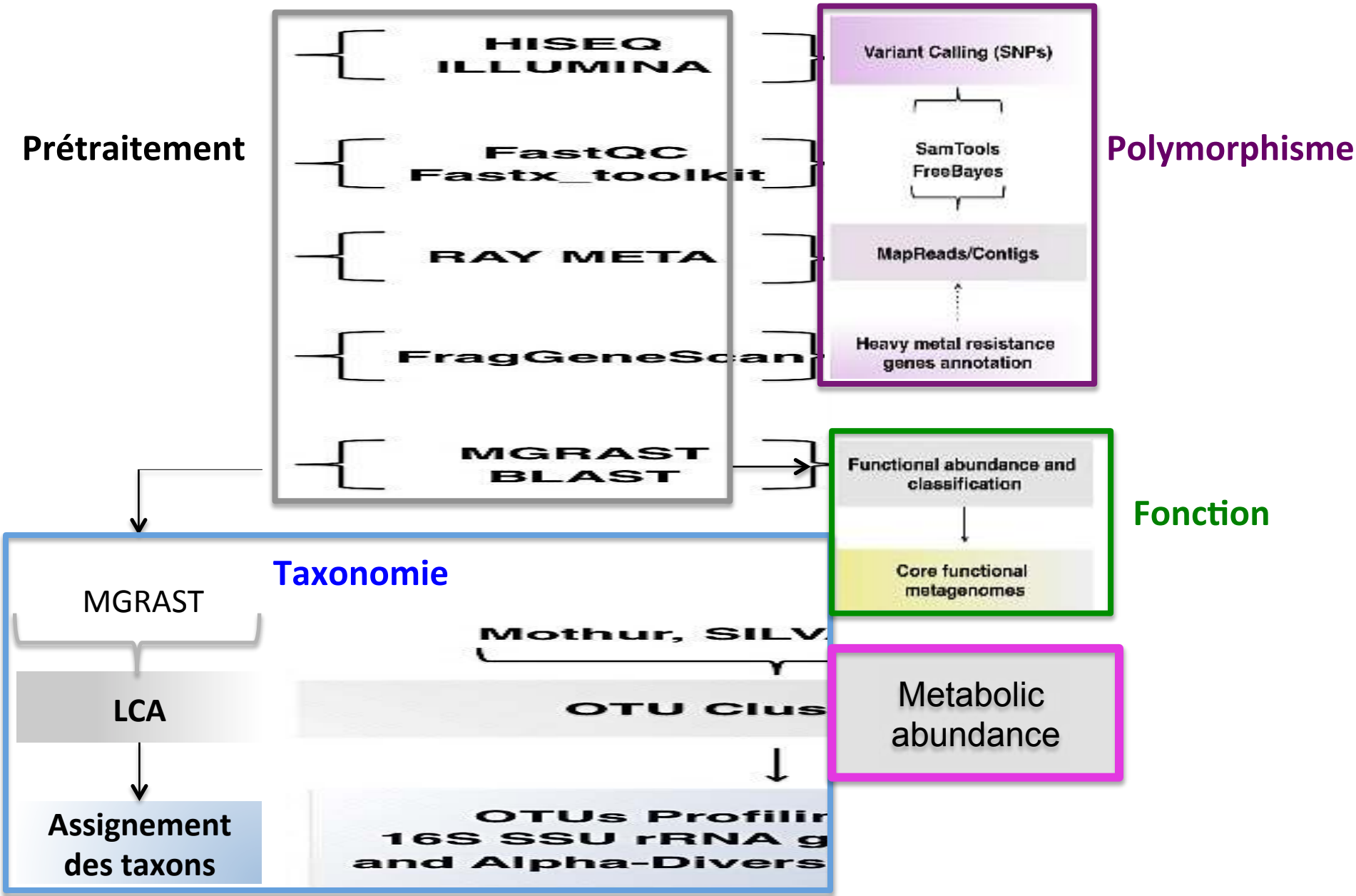
Pipeline d'analyse de "reads" aux métabolites



Pipeline d'analyse de "reads" aux métabolites



Pipeline d'analyse de "reads" aux métabolites



Metagenome webservers and software

- MG-RAST
- METAGENassist
- AMPHORA2

- QIIME (amplicons)
- MOTHUR (amplicons)

- MEGAN

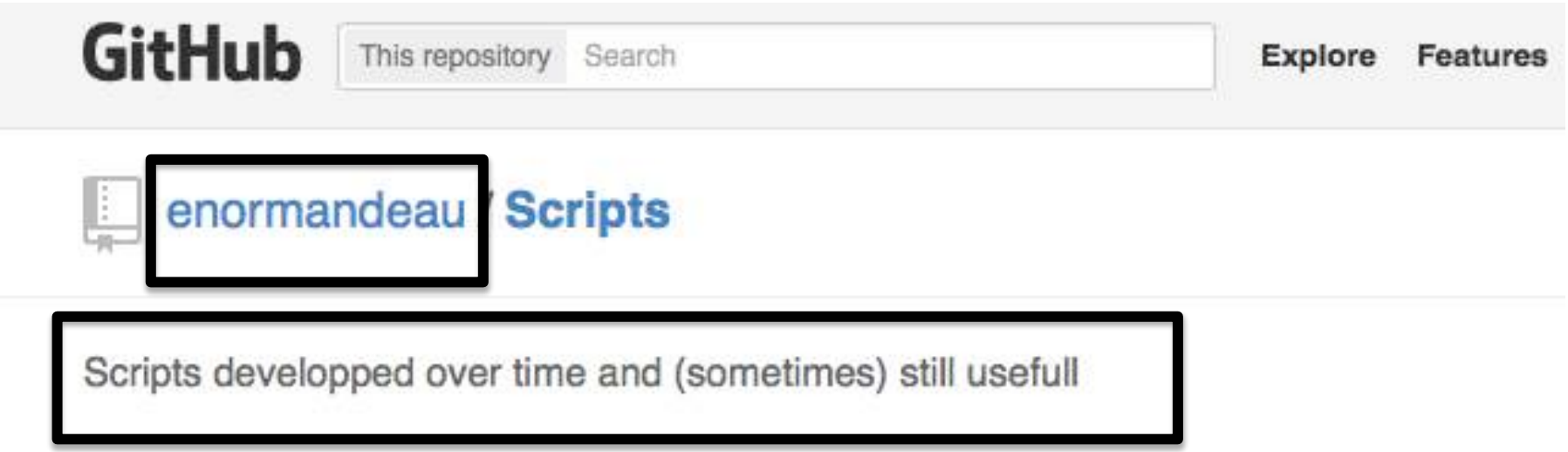
etc

Quelques astuces …

- Chercher des études comparatives ou des logiciels d'évaluation
- Chercher des références pour l'optimisation des paramètres
- Choisir les bonnes ressources d'annotation
- Faire des expériences supplémentaires pour évaluer la pertinence et la fiabilité des méthodes
- Ouvrir la boite noire des outils, logiciels, web serveurs etc.

"Whether you want to uncover the secrets of the universe, ***or*** *you just want to pursue a career in the 21st century,* ***basic computer programming is an essential skil****l to learn."*

Stephen Hawking



"Scripts developed over time and (sometimes) still usefull"

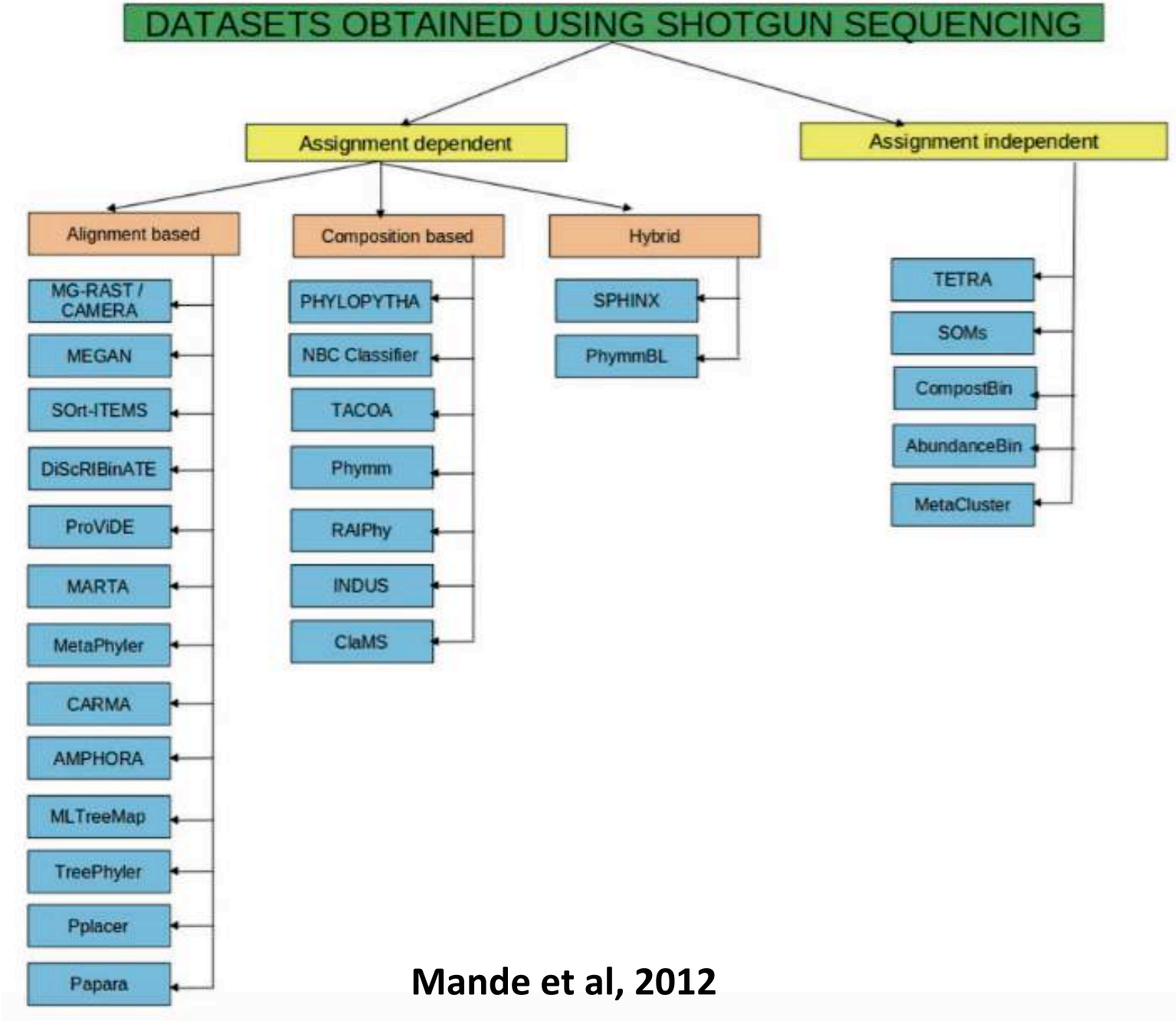
Club of "Biocoders" idea ?

R, Python, PERL, AWK …

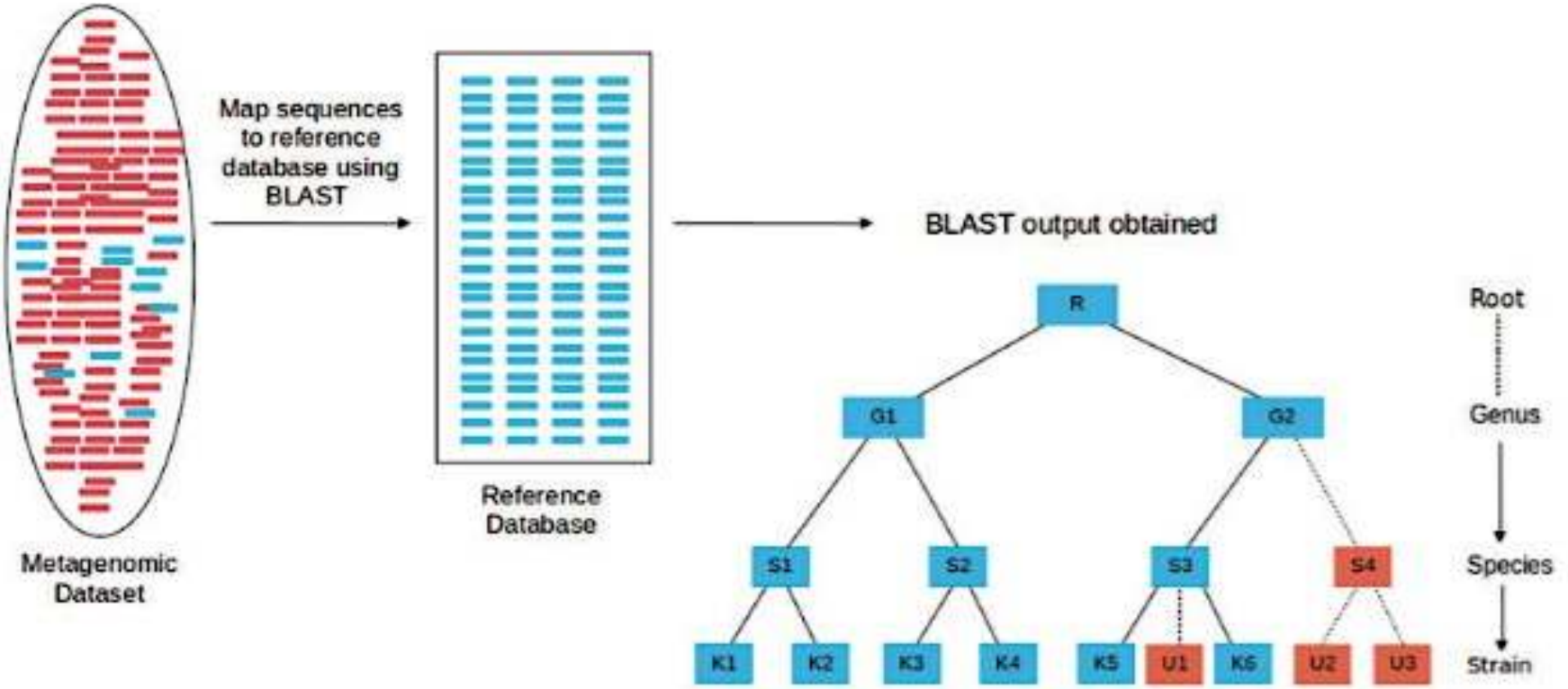
References

- Sharon et Banfield, Science, 2013
- Albertsen et al, Nature biotechnology, 2013
- Iverson et al, Science, 2012
- Yilmaz et al, Nature biotechnology, 2011
- Luo et al., Methods in enzymology, 2013
- Sharpton et al., Frontiers in PLANT science, 2014
- Caporaso et al, Nature Methods, 2010
- Huson et al, Genome Research, 2011


Questions ?



Mande et al, 2012



K=known

U=Unknown

Reads originate from	Significant BLAST Hits	Assingment Strategies	
		Best BLAST Hit Approach	LCA
K1	K1, K2, K3	K1 (√)	G1 (√)
U1	K5, K6	K5 (X)	S3 (√)
U2 and U3	K5, K6	K5 (X)	S3 (X)

Mande et al, 2012