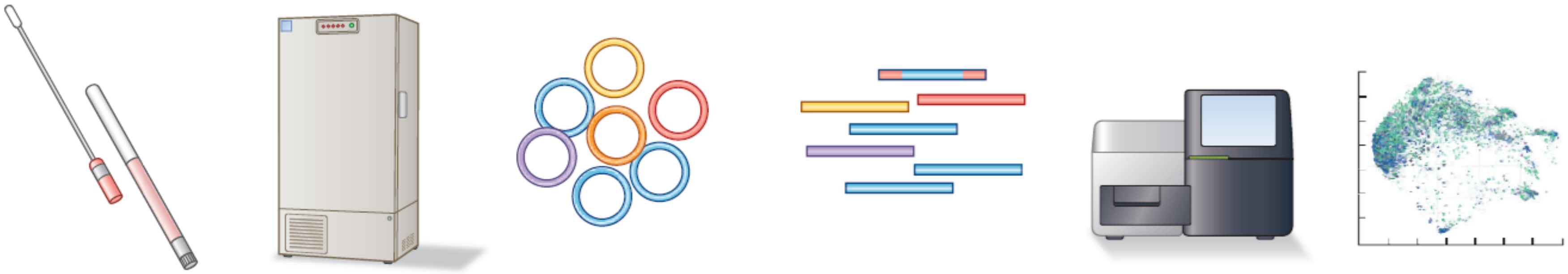
THEME 2

Develop tools for strain level identification and functional analysis.

Alain Pluquet

19.10.2017 • IMI Stakeholder Forum | Microbiome • Brussels, Belgium

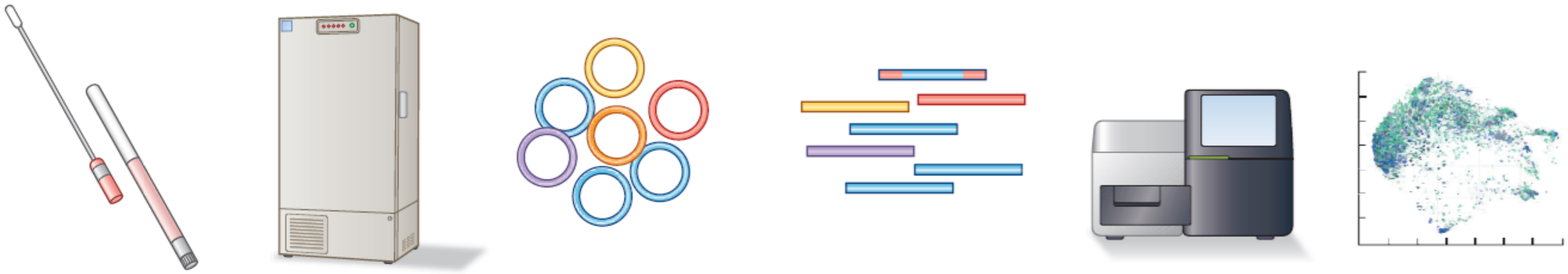
Microbiome workflows : potential errors and biases at each step.



Sample processing step	Sample collection	Sample storage	DNA extraction	Sequencing library preparation	DNA sequencing	Computational analysis
Technical sources of error and bias	• Inadequate sampling • Incomplete sample stabilization • Sampling kit contamination • Mislabeling of samples	• Change in community structure due to differential growth • Degradation of DNA during freeze-thaw cycles	• Differential recovery of DNA from different strains • Extraction kit contamination • Sample swaps during transfer • Sample cross-contamination	• Quantitative amplification bias (PCR efficiency) • Qualitative amplification bias (primer mismatches) • Amplification errors (PCR chimeras, substitution errors) • Reagent contamination • Sample cross-contamination	• Sequencing errors • Run-to-run carryover • Barcode swapping • Demultiplexing errors	• Suboptimal quality control or filtering • Alignment errors • Database errors • Database bias • Batch effects • Failure to flag contaminants

Daryl M Gohl, *Nat. Biotechnol,* doi:10.1038/nbt.3983 (2017)

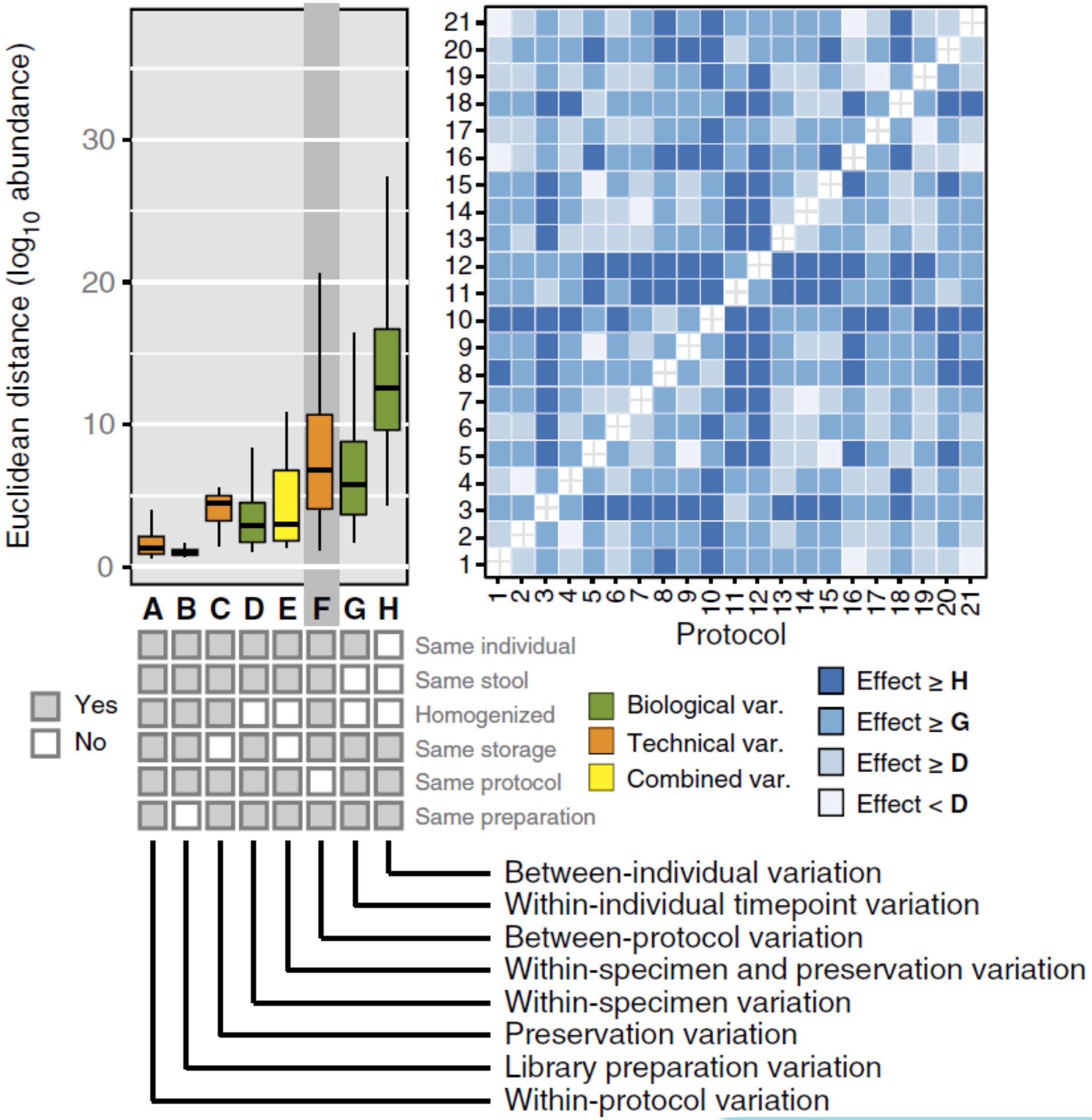
Microbiome workflows : potential errors and biases at each step



Sample processing step	Sample collection	Sample storage	DNA extraction	Sequencing library preparation	DNA sequencing	Computational analysis
Technical sources of error and bias	• Inadequate sampling • Incomplete sample stabilization • Sampling kit contamination • Mislabeling of samples	• Change in community structure due to differential growth • Degradation of DNA during freeze-thaw cycles	• Differential recovery of DNA from different strains • Extraction kit contamination • Sample swaps during transfer • Sample cross-contamination Costea *et al.* (21 labs)	• Quantitative amplification bias (PCR efficiency) • Qualitative amplification bias (primer mismatches) • Amplification errors (PCR chimeras, substitution errors) • Reagent contamination • Sample cross-contamination	• Sequencing errors • Run-to-run carryover • Barcode swapping • Demultiplexing errors	• Suboptimal quality control or filtering • Alignment errors • Database errors • Database bias • Batch effects • Failure to flag contaminants

Daryl M Gohl, *Nat. Biotechnol,* doi:10.1038/nbt.3983 (2017)

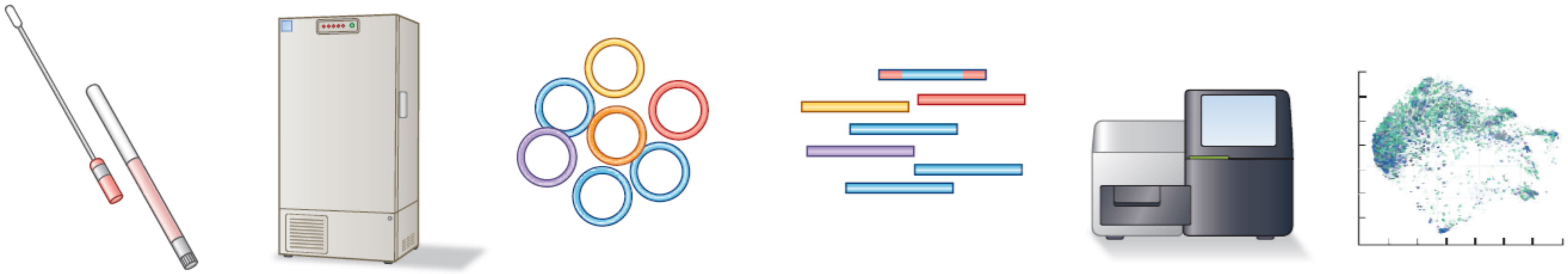
Extraction protocols induce variations, both in taxonomic and functional space



The majority of extraction protocol effects were greater than biological variation within specimens and across time points within the same individual.

Costea, P.I. *et al. Nat. Biotechnol.* http://dx.doi.org/10.1038/nbt.3960 (2017).

Microbiome workflows : potential errors and biases at each step



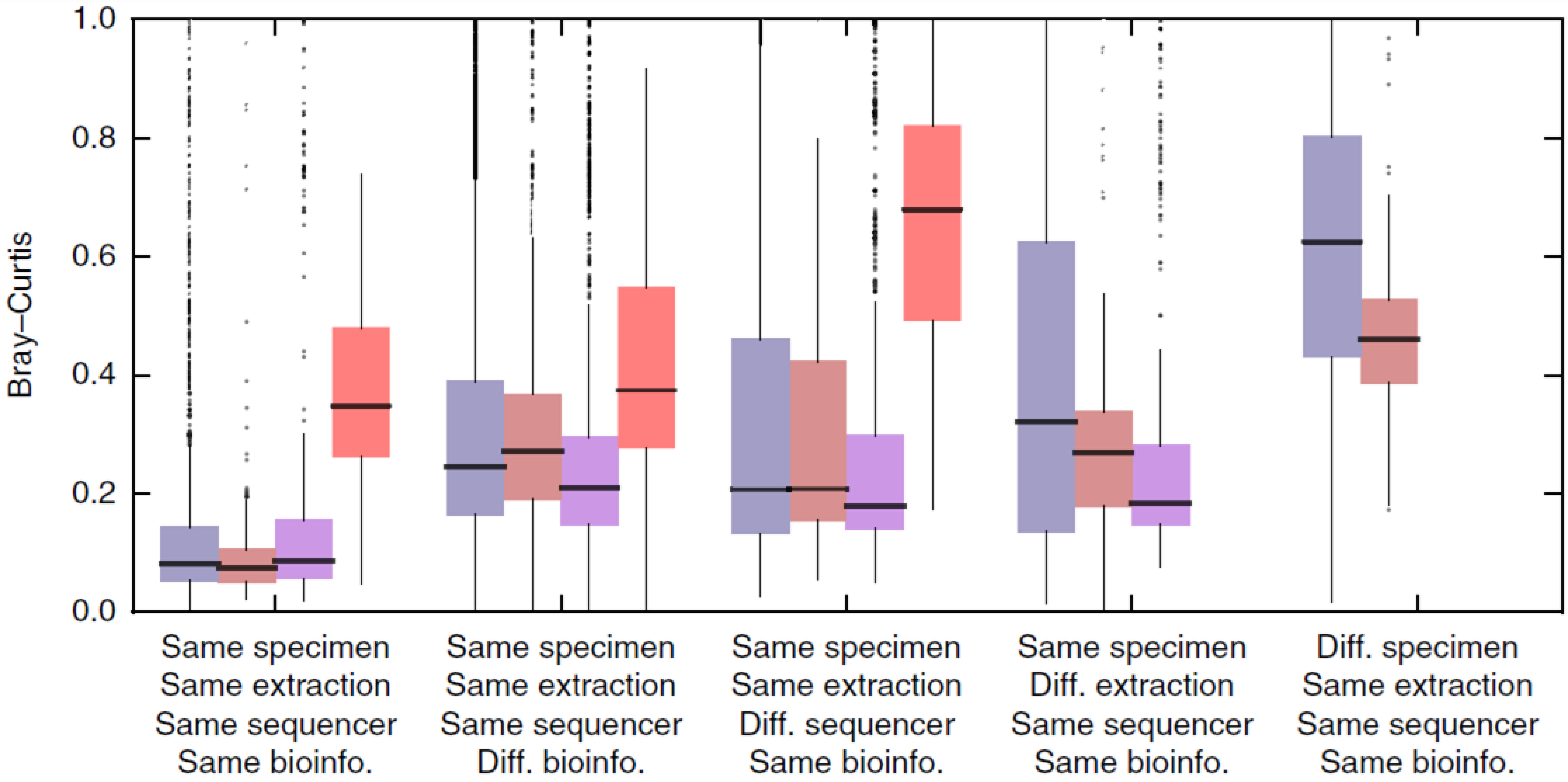
Sample processing step	Sample collection	Sample storage	DNA extraction	Sequencing library preparation	DNA sequencing	Computational analysis
Technical sources of error and bias	• Inadequate sampling • Incomplete sample stabilization • Sampling kit contamination • Mislabeling of samples	• Change in community structure due to differential growth • Degradation of DNA during freeze-thaw cycles	• Differential recovery of DNA from different strains • Extraction kit contamination • Sample swaps during transfer • Sample cross-contamination	• Quantitative amplification bias (PCR efficiency) • Qualitative amplification bias (primer mismatches) • Amplification errors (PCR chimeras, substitution errors) • Reagent contamination • Sample cross-contamination	• Sequencing errors • Run-to-run carryover • Barcode swapping • Demultiplexing errors	• Suboptimal quality control or filtering • Alignment errors • Database errors • Database bias • Batch effects • Failure to flag contaminants

Sinah *et al.*
(15 labs)

Daryl M Gohl, *Nat. Biotechnol,* doi:10.1038/nbt.3983 (2017)

innovative medicines initiative

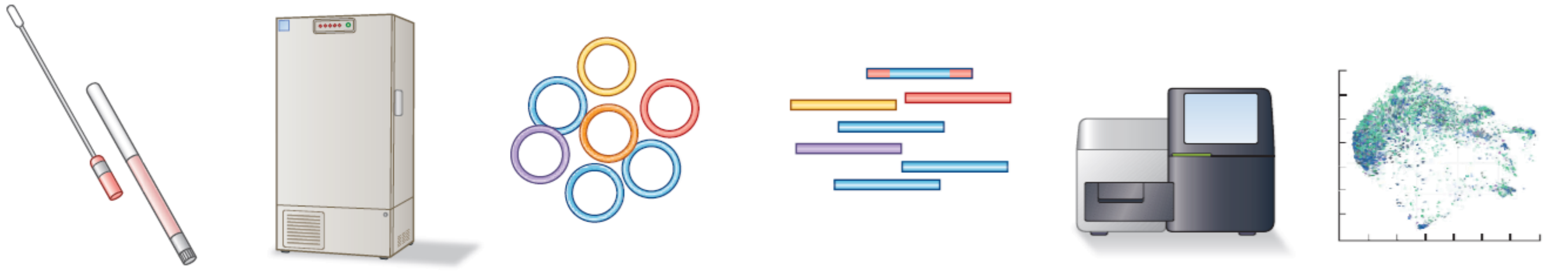
Each step can induce variation of comparable with biological differences.



- Almost any data generation or analysis protocol choice has the potential to yield divergent results.
- However, many potential sources of variation (seq. platform, chemistry, bioinformatics, …) were, when detectable, typically of smaller effect size than phenotypes of clinical interest.

Sinha, R. *et al. Nat. Biotechnol.* http://dx.doi.org/10.1038/nbt.3981 (2017).

Microbiome workflows : potential errors and biases at each step

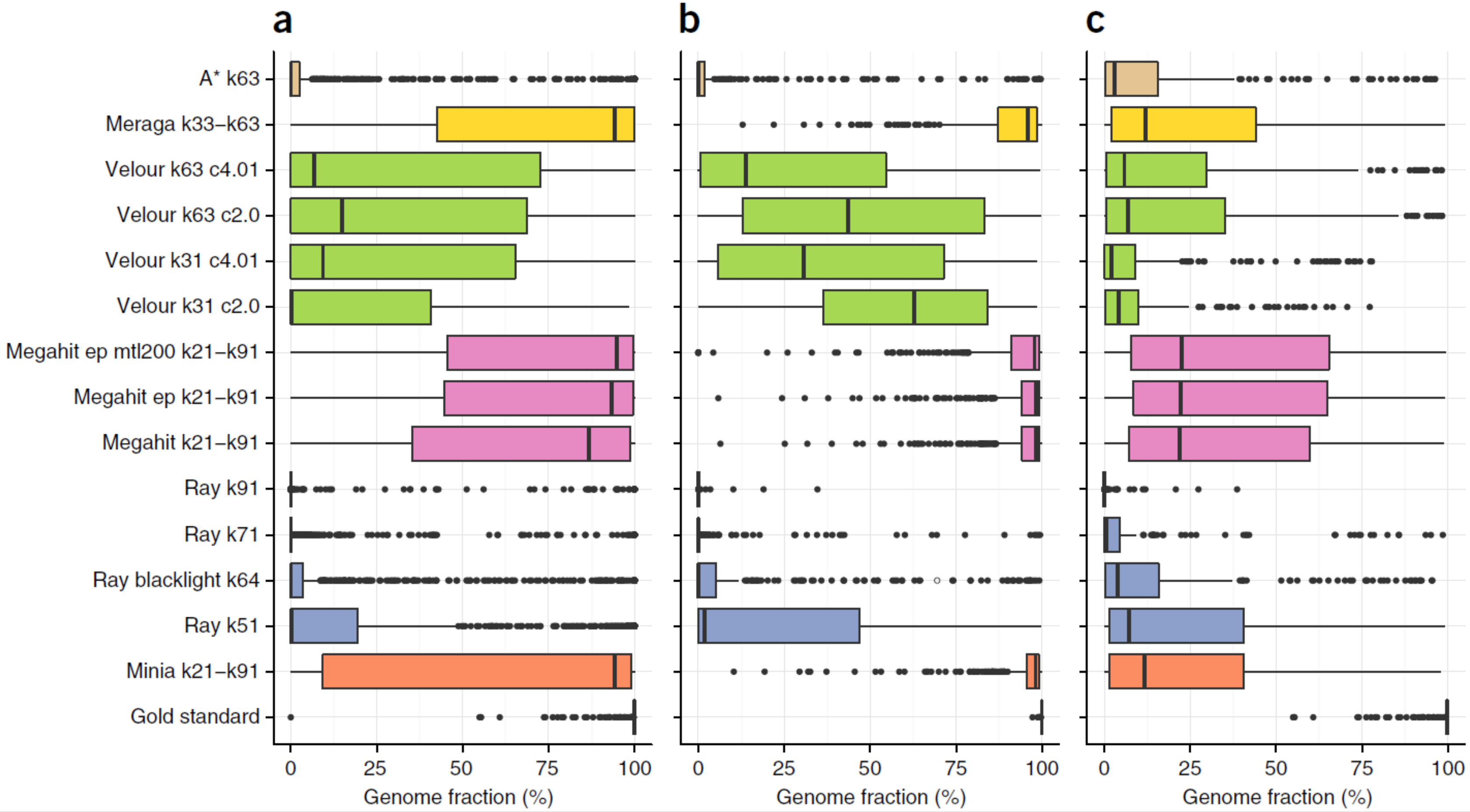


Sample processing step	**Sample collection**	**Sample storage**	**DNA extraction**	**Sequencing library preparation**	**DNA sequencing**	**Computational analysis**
Technical sources of error and bias	• Inadequate sampling • Incomplete sample stabilization • Sampling kit contamination • Mislabeling of samples	• Change in community structure due to differential growth • Degradation of DNA during freeze-thaw cycles	• Differential recovery of DNA from different strains • Extraction kit contamination • Sample swaps during transfer • Sample cross-contamination	• Quantitative amplification bias (PCR efficiency) • Qualitative amplification bias (primer mismatches) • Amplification errors (PCR chimeras, substitution errors) • Reagent contamination • Sample cross-contamination	• Sequencing errors • Run-to-run carryover • Barcode swapping • Demultiplexing errors	• Suboptimal quality control or filtering • Alignment errors • Database errors • Database bias • Batch effects • Failure to flag contaminants

Sczyrba *et al.* (215 submissions)

Daryl M Gohl, *Nat. Biotechnol,* doi:10.1038/nbt.3983 (2017)

Each step can induce variation of comparable with biological differences.



- Lack of consensus about benchmarking.
- Good performances for individual genomes, but substantially affected by related strains.
- Proficient at high taxonomic ranks, with a notable performance decrease below family level.
- Parameter settings affected performance, underscoring their importance for reproducibility.

Sczyrba *et al., Nat. Methods,* doi:10.1038/nmeth.4458 (2017)

Needs and Rationale

Need for tools needed to **identify and quantify microorganisms or genes** present in human samples.

These tools should provide algorithms and user interfaces to work on **individual samples** but also on **group of samples.**

Storage and processing of **metadata** are as important as genomics contents and must be an integral part of the tools.

Need for **integrated platforms**, combining all the necessary algorithms, GUIs, import/export, etc. in a single non-expert friendly environment.

Robustness, reproducibility, automated quality indicators for **end-to-end workflows** are key features to pave the way for **standardization**.

Need for public-private collaborative research

Depending on the application, numerous features can be optimized :

- **fundamental performances** (*e.g.* sampling, sensitivity, specificity, speed).
- **global functionalities** (*e.g.* traceability, reproducibility, controls, ergonomics).

Double advantage of working in a public-private collaborative setting :

1. it offers the possibility to **assemble the best bricks** coming from the two communities in order to build the best tools.
2. it allows **extensive testing** in various and demanding conditions of any new piece of software of interest.

Objectives, deliverables

Overall objectives

Deliver best-in-class bio-informatics solutions :

- As a ***tool*** used by academic or industrial groups which are developing novel diagnostic tests, therapeutic drugs, nutrition products, services, etc.
- As a ***product*** for companies proposing commercial bio-informatics platform and services,
- As ***validated pipelines*** or*,* if possible, *recognized standards* for the microbiome community.

Suggested key deliverables

- **Fully operational end-to-end pipelines** for the specific applications targeted within an IMI microbiome program.

Common to themes 2 and 3 :

- **Standardization recommendations**, especially for regulated contexts (medical, nutrition, …)
- **Benchmarks**.

Open questions

- Bio-informatics only ? Whole workflow ?
- Meta-genomics only ? Meta-proteomics / meta-metabolomics ?
- Application driven tools (*e.g.* correlating microbial, physiological and dietary data) ?

Special thanks to

Tobias RECKER (ILSI)

Wulf FISCHER-KNUPPERTZ (Biocrates)

Rudiger RAUE (Zoetis)

Thank you

Alain Pluquet • BioMérieux Corporate VP
alain.pluquet@biomerieux.com

www.imi.europa.eu
@IMI_JU