Microbial Diversity in the Vaginal Microbiota and its Link to Pregnancy Outcomes

Sean P Kennedy: *Dept of Computational Biology*

Geneva, Switzerland

7th International Conference on Clinical Metagenomics

20th - 21st October 2022

Institut Pasteur

Metagenomic Signatures Team

Sean Kennedy (PI): Microbiologist and bioanalyist. Leading a team focused on microbiota research and analysis tools.

Maria Victoria Bussoletti: Erasmus, biology and biotechnology. Structural annotation of unknown genes in the vaginal microbiota.

Agnes Baud: Python development and bioanalyist. Responsible for much of the analysis of the microbiota community including diversity and statistical comparisons.

Darja Nikatina: Erasmus, health sciences and microbiota. Analysis of mouse microbiota in NASPEC project to develop new antimicrobials.

Kenzo-Hugo Hillion: Bioinformatician and developer. Development of several Python tool and analysis pipelines. (former member)

Keywords: *Microbiota, Clinical Research, Antibiotic Resistance, Analysis Development*

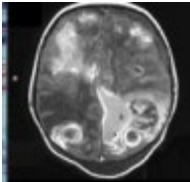# InSPIRe: Innovative Strategies for Perinatal Infection Risk-Reduction
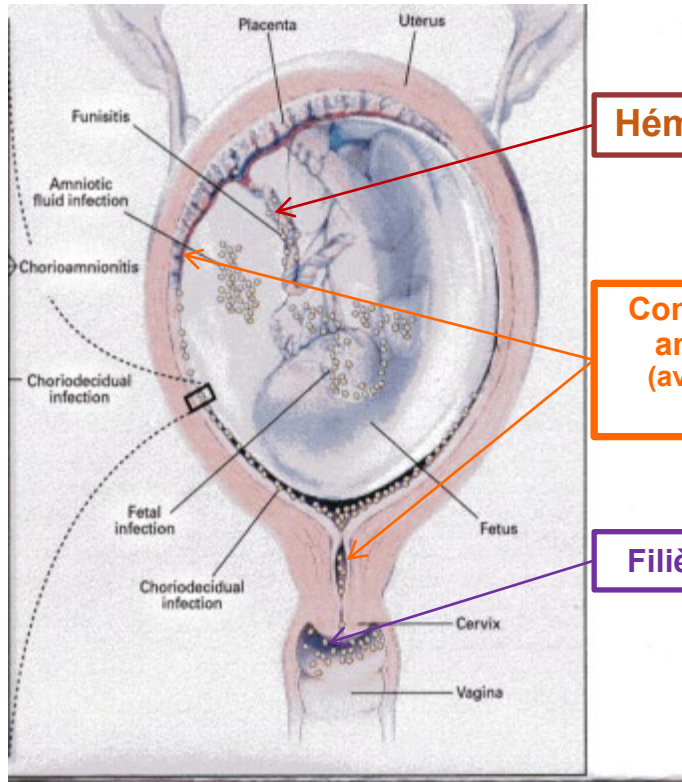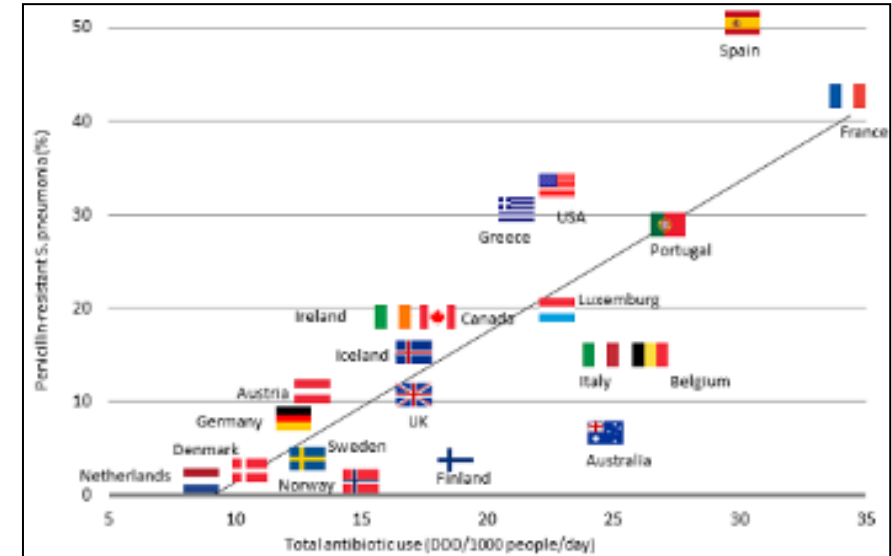

Hématogène

Contamination amniotique (avec ou sans RPM)

Filière génitale

Infection Suspected:
=> Sampling,
=> AtB prophylaxis
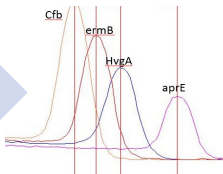
Increasing rates of AtB use and resistance

Test multi-pathogens
- 5+ -plex for detecting GBS and other pathogens
- Sensitivity < 100 copies

Test for resistance
- PoC of GBS with further attribution of resistances
- Technology 'polonies'

InSPIRe is a clinical study with Paris hospitals in order to identify microbiological risk factors for prenatal infection with Group B Streptococcus (GBS) and other potential pathogens and to examine the role different vaginal microbiota communities play in the risk of premature birth.
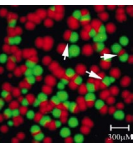
elvesys Innovation Center | Risques & grossesse DHU | ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS | Inserm Institut national de la santé et de la recherche médicale | Institut Pasteur | UNIVERSITÉ PARIS DESCARTES | université PARIS DIDEROT PARIS 7

The Vaginal Microbiota is Usually Dominated by *Lactobacillus sp.*, which Impacts Overall Diversity

Ravel *et al.*, described five **community sequence types (CSTs)** dominated by one of four *Lactobacillus* species resulting in a relatively low-diversity environment.

Gardnerella vaginalis, a fifth CST has been associated with the dysbiotic condition BV with consequent elevated risk during pregnancy. However, this **CST (IV) is highly prevalent in African women** in our study and in others. **Harwich Jr. *et al.*, remarked that both commensal and pathogenic species could be distinguished.

Our results show overall **higher diversity of a range of species to be indicative of higher risk.**
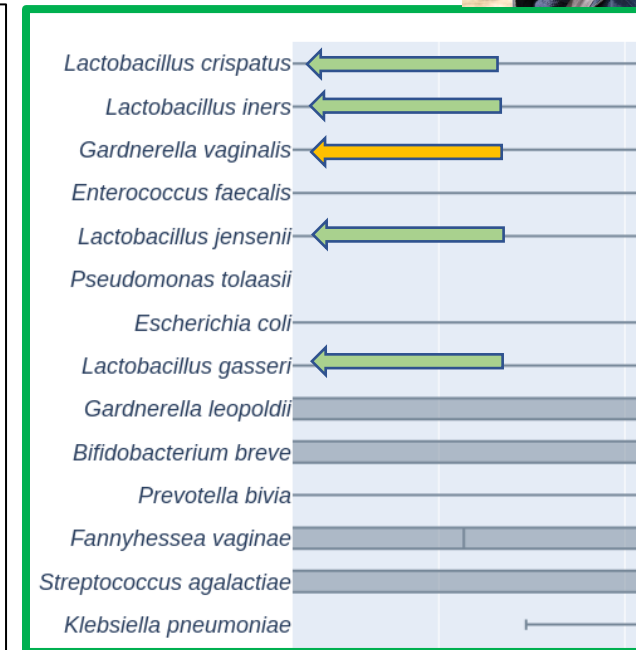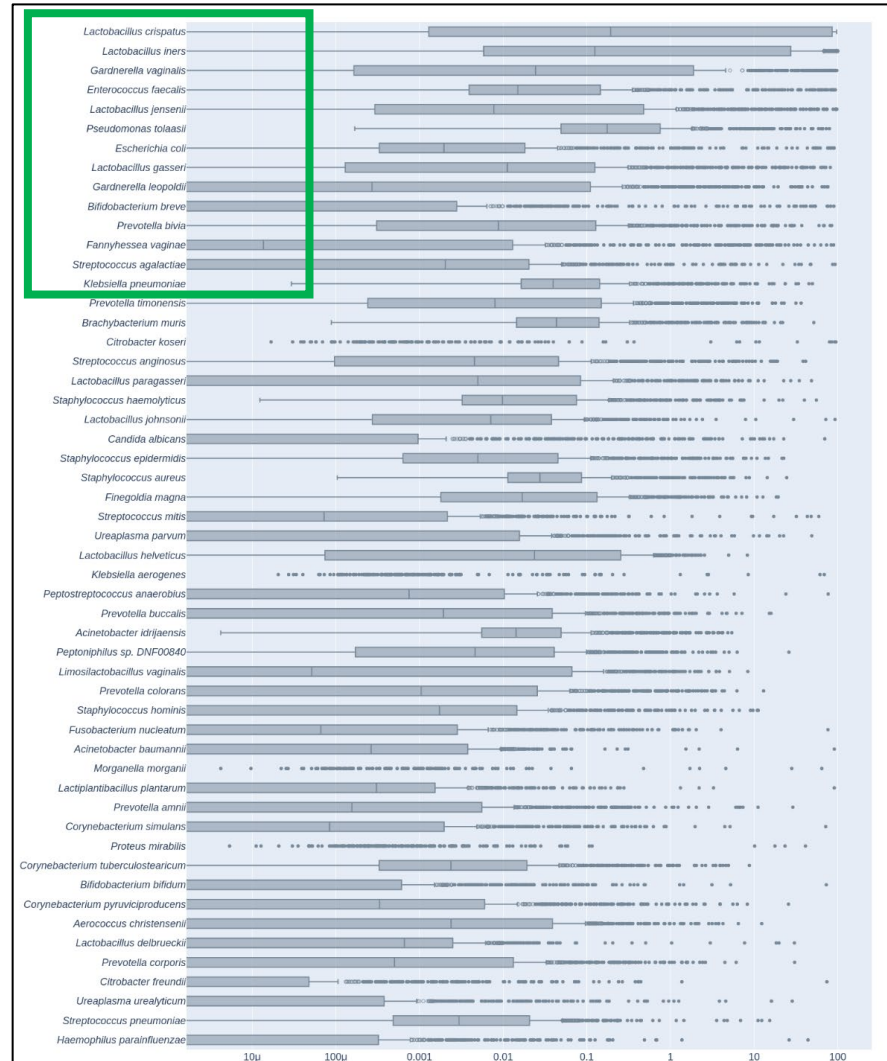
K.K. Jefferson/Virginia Commonwealth University

*Ravel, J. *et al.* Vaginal microbiome of reproductive-age women. *PNAS* **108**, 4680–4687 (2011).
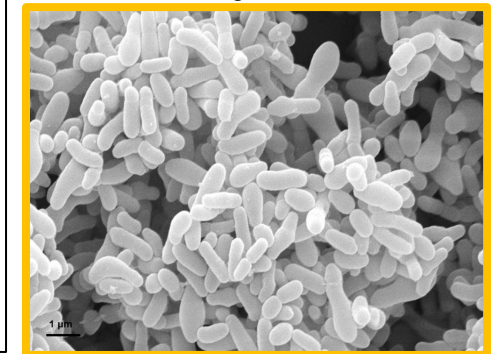
**Harwich, M. D. *et al.* Drawing the line between commensal and pathogenic Gardnerella vaginalis through genome analysis and virulence studies. *BMC Genomics*

Pipeline for Automated Construction of a Project-Specific Database for Metagenomic Analysis

MetaPhlan3[2]:
Marker gene detection

Kraken2[1]:
K-mer mapping algorithm

MetaPhlAn3	Kraken2
+ more specific results **–** no access to the real counts (results in percentage after unknown treatment of the data)	**+** access to the real counts **–** DB too large →not as specific results

marker genes DB built by metaphlan3   *sequencing fastq files*   *large microbial DB*

MetaPhlAn3   Kraken2

count dataframe csv file   *count dataframe csv file*

listing all genus found
+ constructing DB
(sequences from the DB used for Kraken2 + NCBI)

Kraken2

CustomDB (1302 species)   *count dataframe csv file*

Kraken2 analysis using the custom database for the project is most consistent with bacteriology performed on the sample samples and validated the approach.

Vaginal Microbiota: Kraken2 standard database failed to identify *Lactobacillus* sp., the most important component of the vaginal microbiota. MetaPhlAn3 significantly undercounted *Klebsiella* compared with experimental plating results.

[1]Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. *Genome Biol* 2019;**20**:257.
[2]Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, *et al.* MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat Methods* 2015;**12**:902–3.

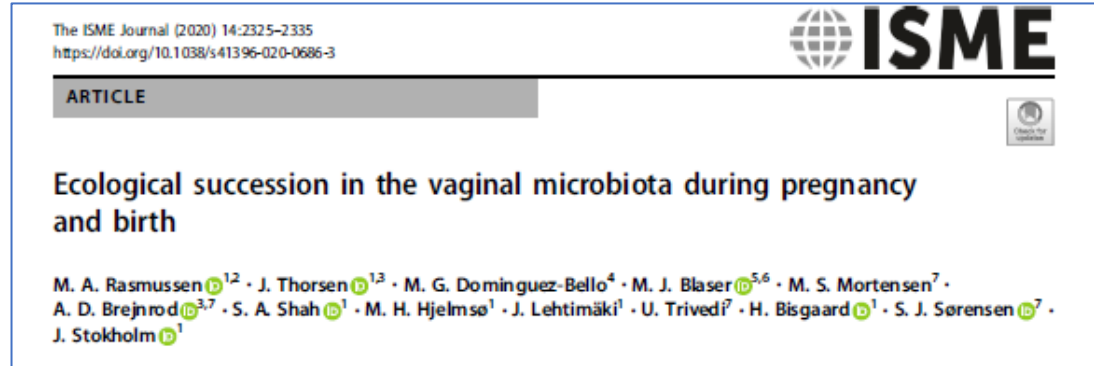# Including Taxonomic Information in Metagenomic Analysis

- **Beta Diversity:**
  - UniFrac is **a distance metric used for comparing biological communities**. It differs from dissimilarity measures such as **Bray-Curtis** dissimilarity in that it incorporates information on the relative relatedness of community members by incorporating phylogenetic distances between observed organisms in the computation.

- **Alpha Diversity:**
  - **Phylogenetic diversity ("PD")** is a measure of biodiversity, based on phylogeny (the tree of life). *Faith (1992)* defined the phylogenetic diversity of a set of species as equal to the sum of the lengths of all those branches on the tree that span the members of the set. The branch lengths on the tree are informative because they count the relative number of new features arising along that part of the tree.

Ecological succession in the vaginal microbiota during pregnancy and birth

M. A. Rasmussen [1,2] · J. Thorsen [1,3] · M. G. Dominguez-Bello[4] · M. J. Blaser [5,6] · M. S. Mortensen[7] · A. D. Brejnrod [3,7] · S. A. Shah [1] · M. H. Hjelmsø[1] · J. Lehtimäki[1] · U. Trivedi[7] · H. Bisgaard [1] · S. J. Sørensen [7] · J. Stokholm [1]

- We compiled the InSPIRe Custom Database of **1302** species that were then used by Kraken2 to classify sequence reads.

- Recovered 16S sequences for these species from in the non-redundant SSU (16/18S) SILVA dataset, supplemented with the NCBI nucleotide database

  - => 1215 species found

- Multiple Sequence Alignment with MAFFT

- Construction of the Phylogenetic Tree with RAxML (100 bootstraps)

- Manual Addition of the 87 missing species

Inclusion of Taxonomic Information Greatly Improves PCoA in Vaginal Microbiota Samples

OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

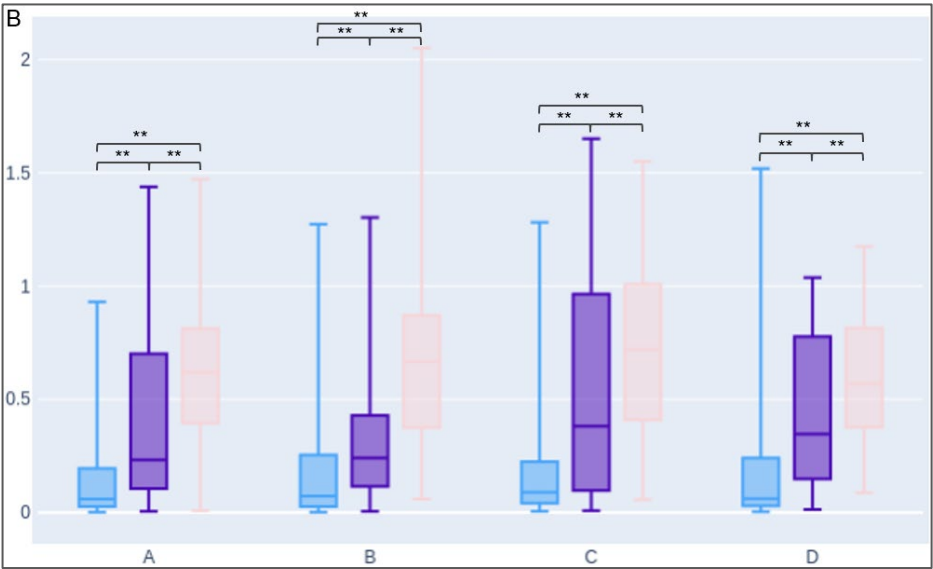## Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible
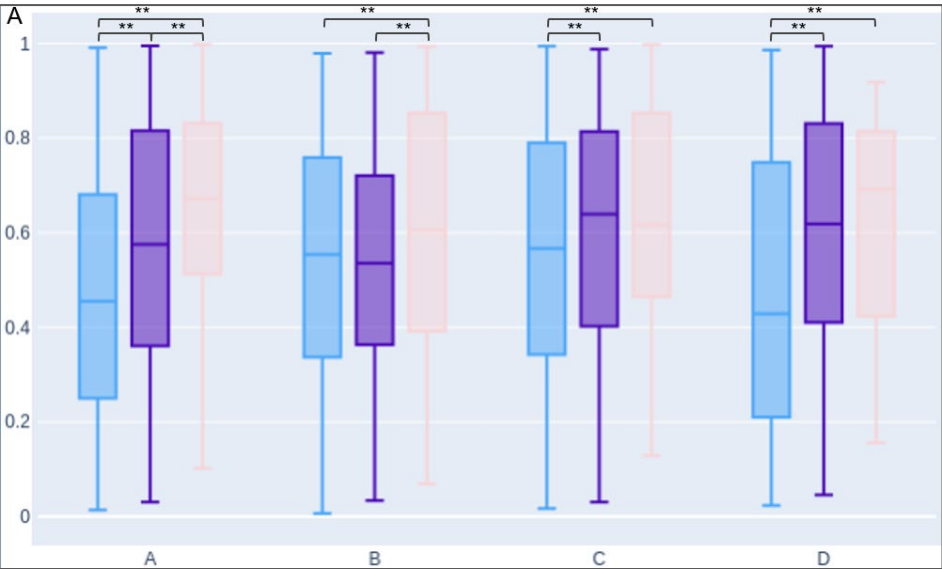
and ignoring taxonomic information

Paul J. McMurdie, Susan Holmes*

Statistics Department, Stanford University, Stanford, California, United States of America

Using our own Unifrac phylogenetic distances to calculate the PCoA matrix, we observe that we can nearly double the amount of information contained in the first two components: 58% versus 31%
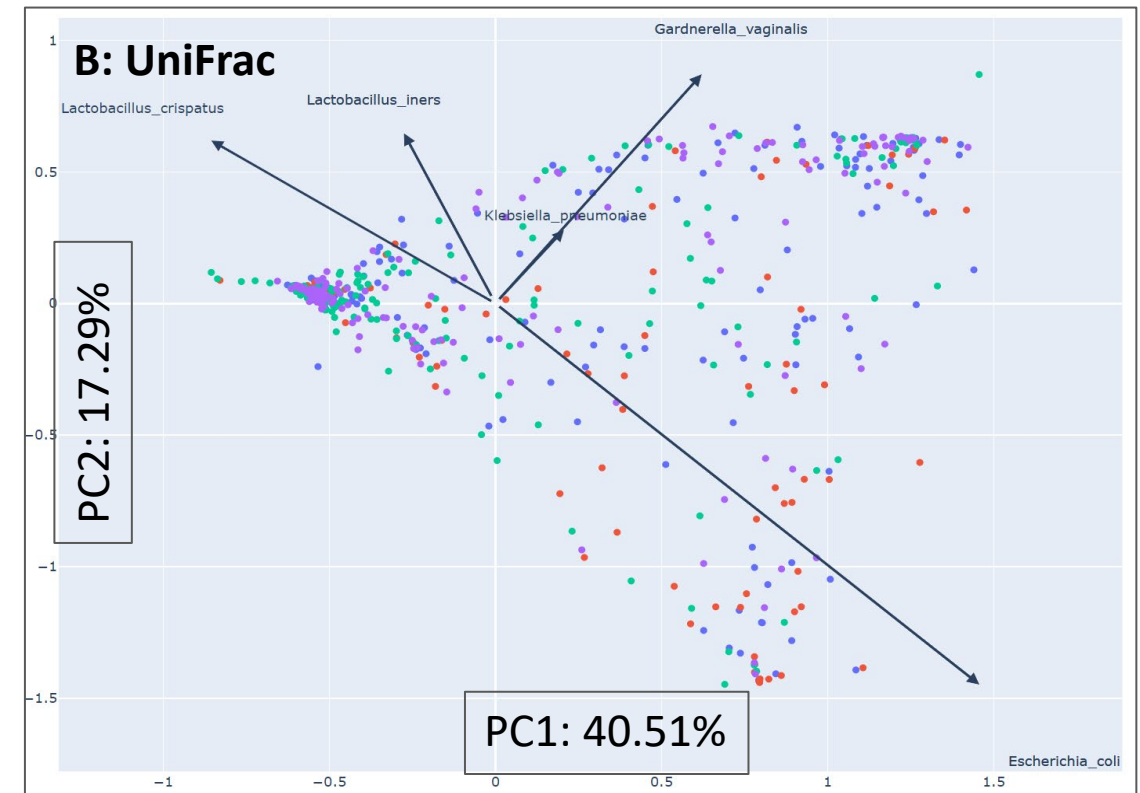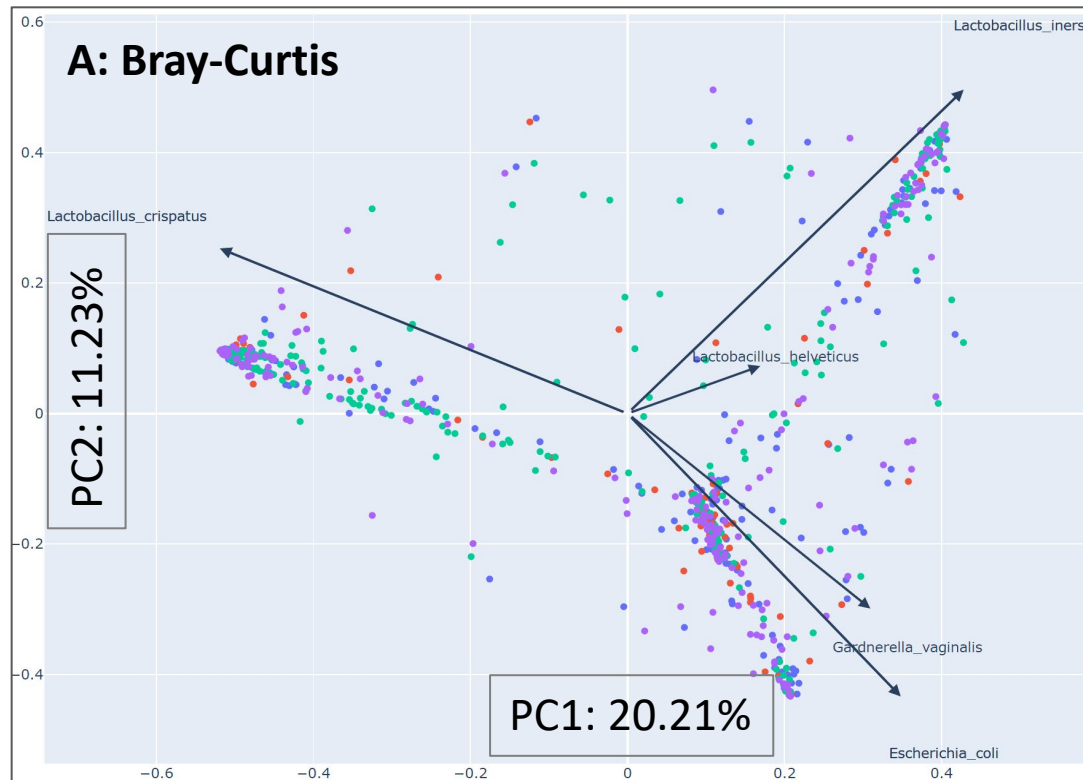
Taxonomic and Quantification Improve Statistical Power

Beta-diversity between samples from the same group of inclusion, dominated by the same species, computed with the Bray-Curtis dissimilarity (A)(Bray and Curtis, 1957) and with Weighted UniFrac (B)

* = p < 0.05; ** = p < 0.01; False Discovery Rate, correction with Benjamini-Hochberg

Legend:
- Lactobacillus crispatus
- Lactobacillus iners
- Gardnerella vaginalis

Group of inclusion

Group A: Full term pregnancy, without premature rupture of membranes (PROM)

Group B: Full term (≥ 37 LMP) pregnancy with PROM

Group C: Premature or threat of premature birth (< 37 LMP)

Group D: PROM before full-term (< 37 LMP)

		BrayCurtis p-values	Weighted UniFrac p-values
A - Gardnerella_vaginalis	A - Lactobacillus_crispatus	5.43e-64	5.45e-223
A - Gardnerella_vaginalis	A - Lactobacillus_iners	5.13e-11	6.75e-50
A - Lactobacillus_crispatus	A - Lactobacillus_iners	9.75e-30	1.34e-156
B - Gardnerella_vaginalis	B - Lactobacillus_crispatus	1.03e-04	3.35e-83
B - Gardnerella_vaginalis	B - Lactobacillus_iners	1.03e-04	2.50e-37
B - Lactobacillus_crispatus	B - Lactobacillus_iners	5.82e-01	2.34e-144
C - Gardnerella_vaginalis	C - Lactobacillus_crispatus	9.32e-08	1.64e-111
C - Gardnerella_vaginalis	C - Lactobacillus_iners	1.58e-02	2.00e-16
C - Lactobacillus_crispatus	C - Lactobacillus_iners	5.53e-03	1.21e-47
D - Gardnerella_vaginalis	D - Lactobacillus_crispatus	8.73e-05	6.56e-21
D - Gardnerella_vaginalis	D - Lactobacillus_iners	6.47e-01	9.54e-05
D - Lactobacillus_crispatus	D - Lactobacillus_iners	9.90e-06	2.57e-20

VCM Dominate Species and Overall Diversity are Linked to Pregnancy Length

a

PD α-diversity

L. crispatus *L. iners* *G. vaginalis* 'Other' non-*Lactobacili*

b

Pregnancy Length

38.2- 40.0
36.1- 38.2
34.4- 36.1
33.2- 34.4
29.7- 33.2
21.7- 29.7

0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

a

12 species variables

ROC fold 0 (AUC = 0.86)
ROC fold 1 (AUC = 0.78)
ROC fold 2 (AUC = 0.83)
Random
Mean ROC (AUC = 0.82 ± 0.04)
± 1 std. dev.

b

Features (species)	Mean weight in estimators	Cumulative Importance
Streptococcus mitis	12.0%	12.0%
Streptococcus agalactiae	10.4%	22.3%
Fusobacterium nucleatum	9.9%	32.3%
Lactobacillus delbrueckii	8.9%	41.2%
Anaerococcus hydrogenalis	8.3%	49.5%
Streptococcus periodonticum	8.2%	57.6%
Peptoniphilus lacrimalis	7.8%	65.5%
Lactobacillus paragasseri	7.8%	73.2%
Limosilactobacillus reuteri	7.1%	80.3%
Lactobacillus amylolyticus	6.9%	87.2%
Anaerococcus tetradius	6.5%	93.8%
Anaerococcus mediterraneensis	6.2%	100.0%

There is a clear difference when analyzing samples by dominate species.

Further, diversity outside of *Lactobacillus* species, and even *Gardnerella vaginalis* is a predicter of pre-term labor risk.
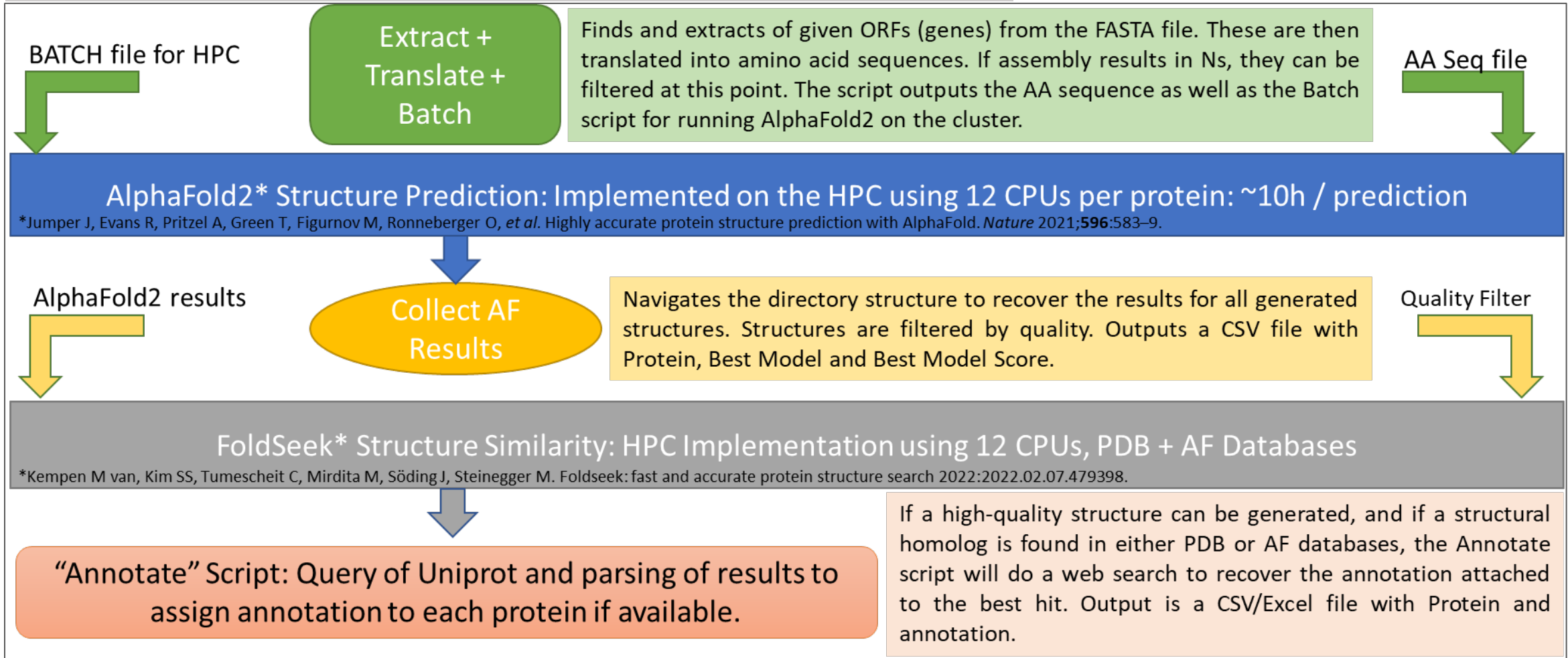
Results from Rasmussen[1] *et al.* showed increased genetic diversity, often independent from the number of species, in the later weeks of pregnancy.

[1]Rasmussen, M. A. *et al.* Ecological succession in the vaginal microbiota during pregnancy and birth. *ISME J* **14**, 2325–2335 (2020).

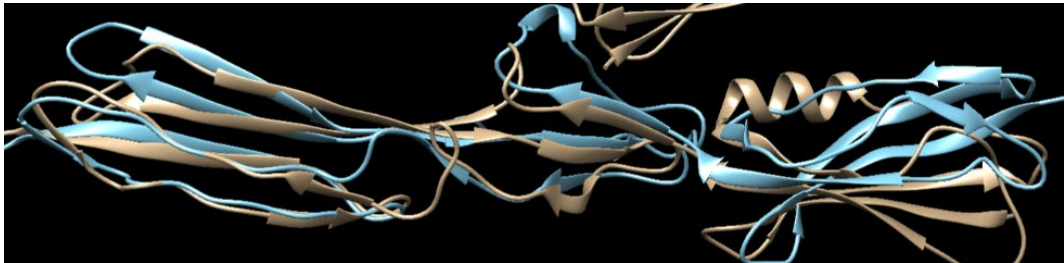# Development of a Pipeline to Harness the HPC Implementation of AlphaFold2

Pipeline for structure prediction and annotation. Input is a Multi-FASTA file.

BATCH file for HPC

Extract + Translate + Batch

Finds and extracts of given ORFs (genes) from the FASTA file. These are then translated into amino acid sequences. If assembly results in Ns, they can be filtered at this point. The script outputs the AA sequence as well as the Batch script for running AlphaFold2 on the cluster.

AA Seq file

AlphaFold2* Structure Prediction: Implemented on the HPC using 12 CPUs per protein: ~10h / prediction
*Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**:583–9.

AlphaFold2 results

Collect AF Results

Navigates the directory structure to recover the results for all generated structures. Structures are filtered by quality. Outputs a CSV file with Protein, Best Model and Best Model Score.

Quality Filter

FoldSeek* Structure Similarity: HPC Implementation using 12 CPUs, PDB + AF Databases
*Kempen M van, Kim SS, Tumescheit C, Mirdita M, Söding J, Steinegger M. Foldseek: fast and accurate protein structure search 2022:2022.02.07.479398.

"Annotate" Script: Query of Uniprot and parsing of results to assign annotation to each protein if available.

If a high-quality structure can be generated, and if a structural homolog is found in either PDB or AF databases, the Annotate script will do a web search to recover the annotation attached to the best hit. Output is a CSV/Excel file with Protein and annotation.

Understanding *L. crispatus* Dominance in the Vaginal Microbiota
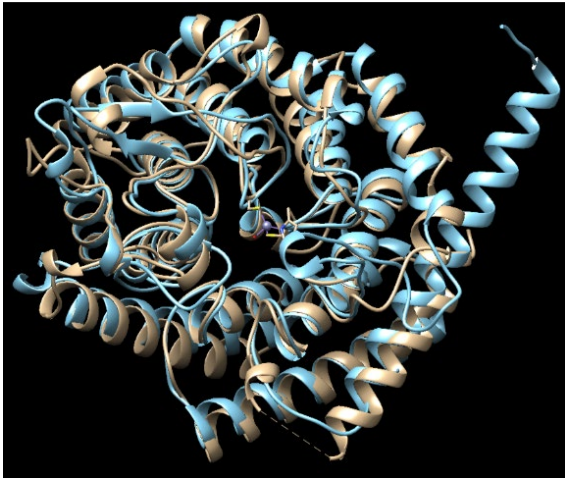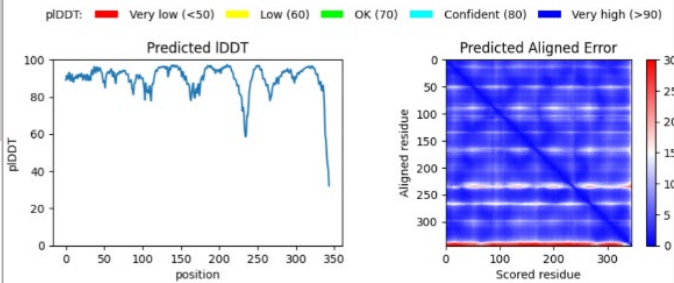

Adhesion

MUB: Mucus Binding Protein, involved in bacterial interaction with mucus and colonization of digestive tract. The binding of full-length MUB to mucus via multiple interactions involving terminal sialylated mucin glycans

Competition

Nisin is the most famous and best studied lantibiotic and it has high antibactericidal activity against a wide range of Gram-positive bacteria

Metabolism

Disaccharide importer: *Lactobacillus* sp. often do not grow in glycogen, are able to utilize glycogen-breakdown products such as maltose, maltotriose, and maltotetraose.

Defense

OxyR: Hydrogen peroxide sensor. Activates the expression of a regulon of hydrogen peroxide-inducible genes such as katG, gor, ahpC, ahpF, oxyS (a regulatory RNA),

Thank you!

Institut Pasteur

Interested in further discussion or collaboration:
sean.kennedy@pasteur.fr