Équipe de recherche
[1] Université Paris-Saclay, INRAE, MaIAGE, 7 8350, Jouy-en-Josas, France

Intra-species diversity in metagenomic datasets

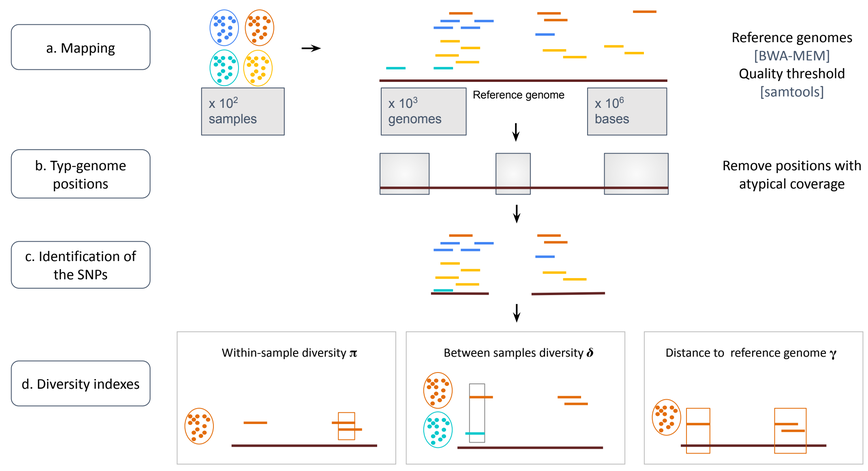
Anne-Laure ABRAHAM[1], Guillaume KON KAM KING[1], Solène PETY[1], Anne-Carmen SANCHEZ[1], Hélène CHIAPELLO[1] and Pierre NICOLAS[1]

1. Résumé

Microbial ecosystems are composed of tens to thousands of species of bacteria, archaea, microbial eukaryotes, and viruses. **Shotgun metagenomic sequencing** has revealed **a high level of intra-species diversity** in several ecosystems. Identifying polymorphisms and reconstructing strains is challenging due to sequencing errors (which must be differentiated from true polymorphisms) and short read length, particularly for species in low abundance. Some approaches aim at resolving strains, either based on selected marker genes or on entire genomes (review [Ventolero, 2022]). These approaches have the advantage of providing precise information on strain contents. However, they are usually limited to species with a high abundance, requiring approximately 5X coverage. Other methods use reads mapped to references to quantify within and between-sample genomic variation, by computing several metrics to compare samples, such as similarity indexes inspired by population genetics (π and F_{ST}) [Costea 2017, Olm 2021], distribution of major allele frequencies [Garud, 2019] or pairwise distance between samples [Podlesny, 2022]. To our knowledge, none of these methods can handle species in very low abundance.

Here, we present **INTERSTICE (INTra-species divERSity in meTagenomIC rEads)**, a new **method for studying intra-species diversity** that is designed to handle **species in low abundance**. The method proposes an estimation of within-sample diversity and between-sample distance, for each species, by adapting to metagenomic samples the computation of indexes used in population genetics : nucleotide diversity π and Nei's standard genetic distance [Nei 1978, 1979]. It first maps metagenomic reads to a complete ecosystem-adapted reference genome catalog (UHGG for human gut microbiota [Almeida, 2021]) and applies stringent quality filters. Diversity indexes are computed only on reads mapped on genomic regions that are conserved at species-level. These regions are determined by analyzing coverage variation across samples (removing regions with atypical profiles) and are designated as the Typ-genome. We applied this method on data from two cohorts: HMP [Huttenhower 2012] (adults) and DIABIMMUNE [Yassour 2016] (longitudinal data on children between 0 and 3 years). With sub-sampled datasets, we assessed the robustness of our metrics with respect to decreasing coverage and confirm that values above 0.001 bp^{-1} require the pairwise comparison of reads on only 10Kbp of the Typ-genome to be reliably estimated. This makes it possible to retrieve information on low abundance species with genome coverage below 0.1X. By analyzing the 747 bacterial species satisfying this minimal criterion, we identify the species with high or low within-sample diversity, the species with rapid lineage turnover, and the species with atypical amount of shared lineages between samples.

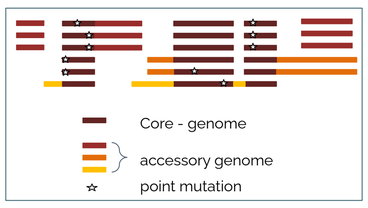
2. Global strategy



A dedicated workflow to compute intra-species intra-sample diversity and between-sample diversity from metagenomic reads. a. After read quality filters and removal of host reads, metagenomic reads are aligned to a database of reference genomes. We remove reads and bases of low quality (low mapping quality, multi-mapped reads, paired incorrectly mapped, base quality below 35). b. For each genome, Typically-conserved-positions (Typ-genome) are computed. c. Allele frequencies are computed for each Typ-genome position covered by reads d. Diversity indices are computed on these conserved positions.

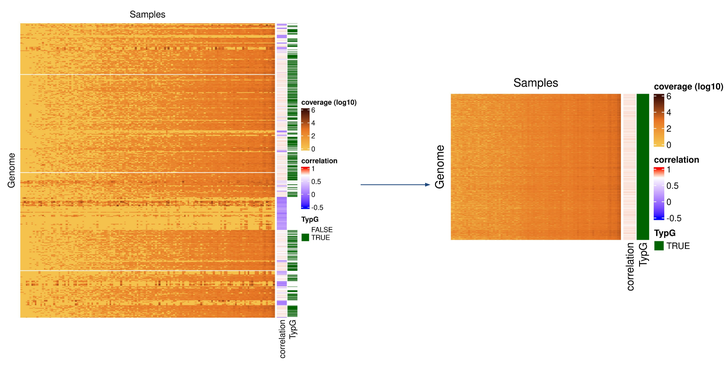
3. Typ-Genome

Objective: identifying conserved position in species



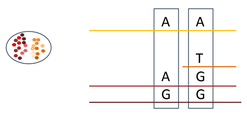
Idea: use coverage across samples to compute Typically-conserved-positions

Method
- Align all samples on reference genome
- Compute coverage on sliding windows - 1000 bases
- Compute mean coverage for each sample
- Identify regions with coverage closest to the mean coverage across samples (pearson correlation)
- Typ-genome = 50% windows with higher correlation



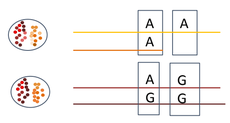
4. Diversity indices

Within-sample diversity



$$\pi = \frac{\sum_{l=1}^{L}\sum_{i\neq j}^{n} d_{ij}}{\sum_{l=1}^{L} n^{[l]}(n^{[l]}-1)/2} I\{n^{[l]} \geq 2\}$$

Between-sample diversity

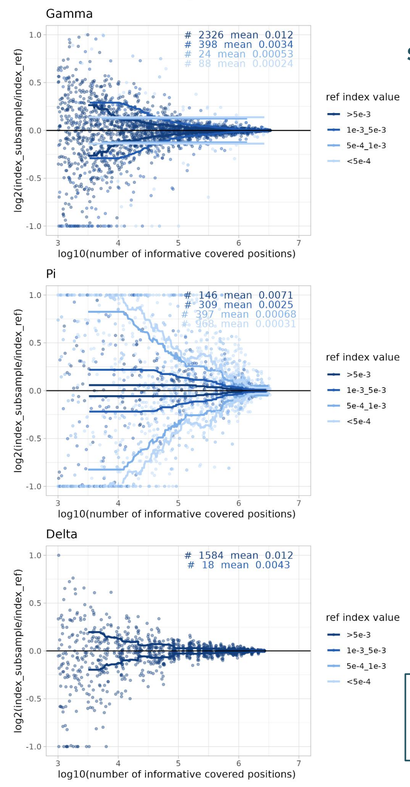


$$\delta = \frac{\sum_{1<l<L}\left(n^{[l,1]}n^{[l,2]} - \sum_{x\in\{a,c,g,t\}} n^{[l,1,x]}n^{[l,2,x]}\right)}{\sum_{1<l<L} n^{[l,1]}n^{[l,2]}} I\{n^{[l,1]} \geq 1, n^{[l,2]} \geq 1\}$$

$$dNei = -ln\frac{1-\delta}{\sqrt{(1-\pi_1)(1-\pi_2)}}$$

π, δ and Nei distance are independent from the reference genome

5. Low abundance species



Subsampling
- 3 samples with high coverage of some species
- Sub-sampling (0.3 to 0.00001, 3 replicates)
- Select species with horizontal coverage at full depth > 50%
- Comparison of indices on sub-sampled datasets with full depth index

Lines: quantile values on a moving window of size 100, with a probability of 0.5.
Index value at reference depth are classified in 4 categories.

Threshold >= 10 000 covered position on Typ-genome

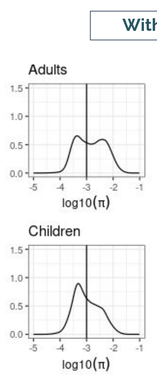
6. Diversity and stability over time?

Reference genome catalogue[Almeida 2021]
- Unified Human Gastrointestinal Genome (UHGG)
- > 200,000 genomes & MAG
- Clustering 95% identity
- 4,644 prokaryotic species

Adults cohort[Huttenhower 2012]
148 samples
High sequencing depth (20-138 M reads)

Children cohort[Yassour 2016]
240 samples - 39 children (0-3 years old)
1-60M reads

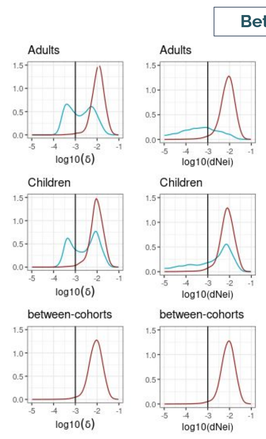
Within-sample diversity



- $\pi < 10^{-3}$: low diversity (probably one lineage)
- $\pi > 10^{-3}$: higher diversity

-> higher diversity for adults cohort at the intra-species level

Between-sample diversity



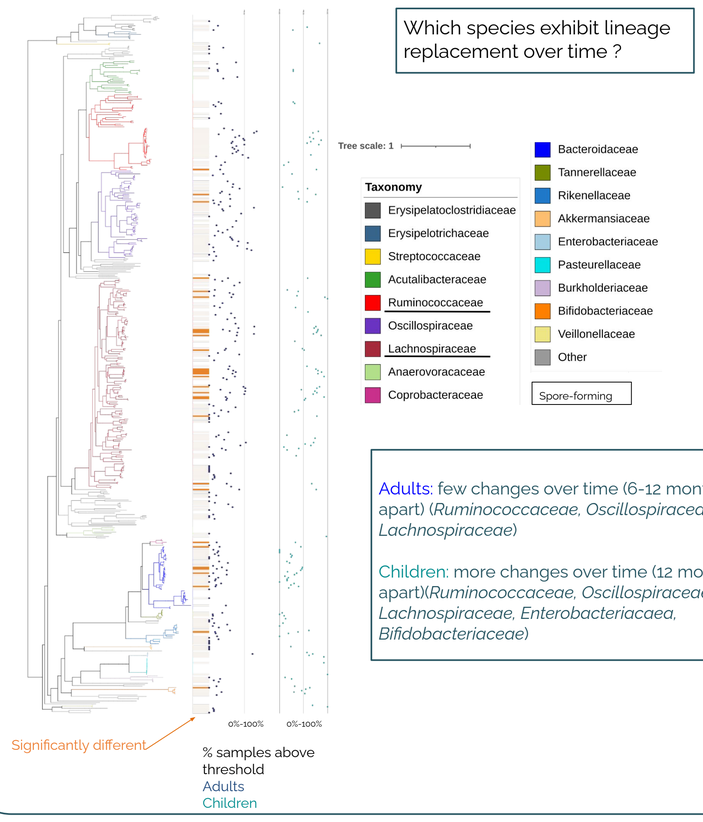
Inter-individual comparison:
- $\delta > 10^{-3}$: different lineages between individuals
- $\delta < 10^{-3}$: few shared lineages between different individuals (diet, probiotics … ?) (for ex. *Streptococcus thermophilus*, *Lactococcus lactis*, *Lactobacillus rhamnosus*….)

Intra-individual comparison:
- $\delta > 10^{-3}$: changes in composition between 2 time points
- $\delta < 10^{-3}$: same composition in lineages between 2 time points

-> stability for adults cohort, changes for children cohort

7. Lineage turn-over

Which species exhibit lineage replacement over time ?



Adults: few changes over time (6-12 months apart) (*Ruminococcaceae, Oscillospiraceae, Lachnospiraceae*)

Children: more changes over time (12 months apart)(*Ruminococcaceae, Oscillospiraceae, Lachnospiraceae, Enterobacteriacaea, Bifidobacteriaceae*)

8. Conclusion

A new method to compute intra-species diversity in metagenomic datasets
- Adaptation indices used in population genetics & computation of Typ-genome
- Validity for low abundance species
- Reference genome database adapted to the ecosystem
- Work on computation time & disk space

Results coherent with literature on human gut microbiota
- Stability of adult gut microbiota over time [Truong 2017, Chen 2021].
  - spore-forming species could have a higher turnover than non-spore-forming species [Browne, 2016]
  - Strains of the same species in different subjects are generally distinct - except few species (diet, medication or probiotics)
- Variation of infant gut microbiota during the first years of life
  - First colonisation: microorganisms transmitted by the mother (*Bifidobacterium*, *Bacteroides* and *Parabacteroides*) [Yassour 2018, Ferretti 2018, Podlesny 2021, Chen 2021]
  - Late colonizing bacteria (including spore-former *Ruminococcaceae, Lachnospiraceae*) could be acquired from the environment or family members [Browne 2016, Podlesny 2021]

Work in progress

- Almeida, A. et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. Nat Biotechnol 39, 105–114 (2021).9.
- Browne, H. P. et al. Culturing of 'unculturable' human microbiota reveals novel taxa and extensive sporulation. Nature 533, 543–546 (2016).
- Chen, D. W. & Garud, N. R. Rapid evolution and strain turnover in the infant gut microbiome. Genome Res 32, 1124–1136 (2022).
- Costea, P. I. et al. metaSNV: A tool for metagenomic strain level analysis. PLOS ONE 12, e0182392 (2017).
- Ferretti, P. et al. Mother-to-Infant Microbial Transmission from Different Body Sites Shapes the Developing Infant Gut Microbiome. Cell Host & Microbe 24, 133-145.e5 (2018).
- Garud, N. R., Good, B. H., Hallatschek, O. & Pollard, K. S. Evolutionary dynamics of bacteria in the gut microbiome within and across hosts. PLOS Biology 17, e3000102 (2019).
- Huttenhower, C. et al. Structure, function and diversity of the healthy human microbiome. Nature 486, 207–214 (2012).
- Nei, M. & Li, W. H. Mathematical model for studying genetic variation in terms of restriction endonucleases. PNAS USA 76, 5269–5273 (1979).
- Nei, M. Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics 89, 583–590 (1978).
- Olm, M. R. et al. inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. Nat Biotechnol 39, 727–736 (2021).
- Podlesny, D. et al. Metagenomic strain detection with SameStr: identification of a persisting core gut microbiota transferable by fecal transplantation. Microbiome 10, 53 (2022).
- Podlesny, D. & Fricke, W. F. Strain inheritance and neonatal gut microbiota development: A meta-analysis. IJMM 311, 151483 (2021).
- Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C. & Segata, N. Microbial strain-level population structure and genetic diversity from metagenomes. Genome Res. 27, 626–638 (2017).
- Ventolero, M. F., Wang, S., Hu, H. & Li, X. Computational analyses of bacterial strains from shotgun reads. Briefings in Bioinformatics 23, bbac013 (2022).
- Yassour, M. et al. Strain-Level Analysis of Mother-to-Child Bacterial Transmission during the First Few Months of Life. Cell Host & Microbe 24, 146-154.e4 (2018).
- Yassour, M. et al. Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability. Science Translational Medicine 8, 343ra81-343ra81 (2016).

Centre
Île-de-France – Jouy-en-Josas – Antony

Domaine de Vilvert 78352Jouy-en-Josas Cedex
contact : anne-laure.abraham@inrae.fr
www.inrae.fr

Acknowledgments

We are grateful to the **INRAE MIGALE bioinformatics facility** (MIGALE, INRAE, 2020. Migale bioinformatics Facility, doi: 10.15454/1.5572390655343293E12) for providing computing and storage resources.