# Intra-species diversity in metagenomic datasets

Anne-Laure Abraham, Guillaume Kon Kam King, Solène Pety, Anne-Carmen Sanchez, Hélène Chiapello, Pierre Nicolas

# 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 Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France

Corresponding Author: anne-laure.abraham@inrae.fr

**Keywords**

Metagenomic, intra-species diversity, human gut microbiota

**Abstract**

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 by Ventolero *et al.* [9]). 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 ($\pi$ and $F_{ST}$) [2, 7], distribution of major allele frequencies [3] or pairwise distance between samples [8]. 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 $\pi$ and Nei's standard genetic distance [5,6]. It first maps metagenomic reads to a complete ecosystem-adapted reference genome catalog (UHGG for human gut microbiota [1]) 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 [4] (adults) and DIABIMMUNE [10] (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.

## References


1. Almeida A, Nayfach S, Boland M, Strozzi F, Beracochea M, Shi ZJ, et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. Nat Biotechnol. janv 2021;39(1):105-14.
2. Costea PI, Munch R, Coelho LP, Paoli L, Sunagawa S, Bork P. metaSNV: A tool for metagenomic strain level analysis. PLOS ONE. 28 juill 2017;12(7):e0182392.
3. Garud NR, Good BH, Hallatschek O, Pollard KS. Evolutionary dynamics of bacteria in the gut microbiome within and across hosts. PLOS Biology. 23 janv 2019;17(1):e3000102.
4. Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, Chinwalla AT, et al. Structure, function and diversity of the healthy human microbiome. Nature. juin 2012;486(7402):207-14.
5. Nei M, Li WH. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci U S A. oct 1979;76(10):5269-73.
6. Nei M. ESTIMATION OF AVERAGE HETEROZYGOSITY AND GENETIC DISTANCE FROM A SMALL NUMBER OF INDIVIDUALS. Genetics. 20 juill 1978;89(3):583-90.
7. Olm MR, Crits-Christoph A, Bouma-Gregson K, Firek BA, Morowitz MJ, Banfield JF. inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. Nat Biotechnol. juin 2021;39(6):727-36.
8. Podlesny D, Arze C, Dörner E, Verma S, Dutta S, Walter J, et al. Metagenomic strain detection with SameStr: identification of a persisting core gut microbiota transferable by fecal transplantation. Microbiome. 25 mars 2022;10(1):53.
9. Ventolero MF, Wang S, Hu H, Li X. Computational analyses of bacterial strains from shotgun reads. Briefings in Bioinformatics. 1 mars 2022;23(2):bbac013.
10. Yassour M, Vatanen T, Siljander H, Hämäläinen AM, Härkönen T, Ryhänen SJ, et al. Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability. Science Translational Medicine. 15 juin 2016;8(343):343ra81-343ra81.