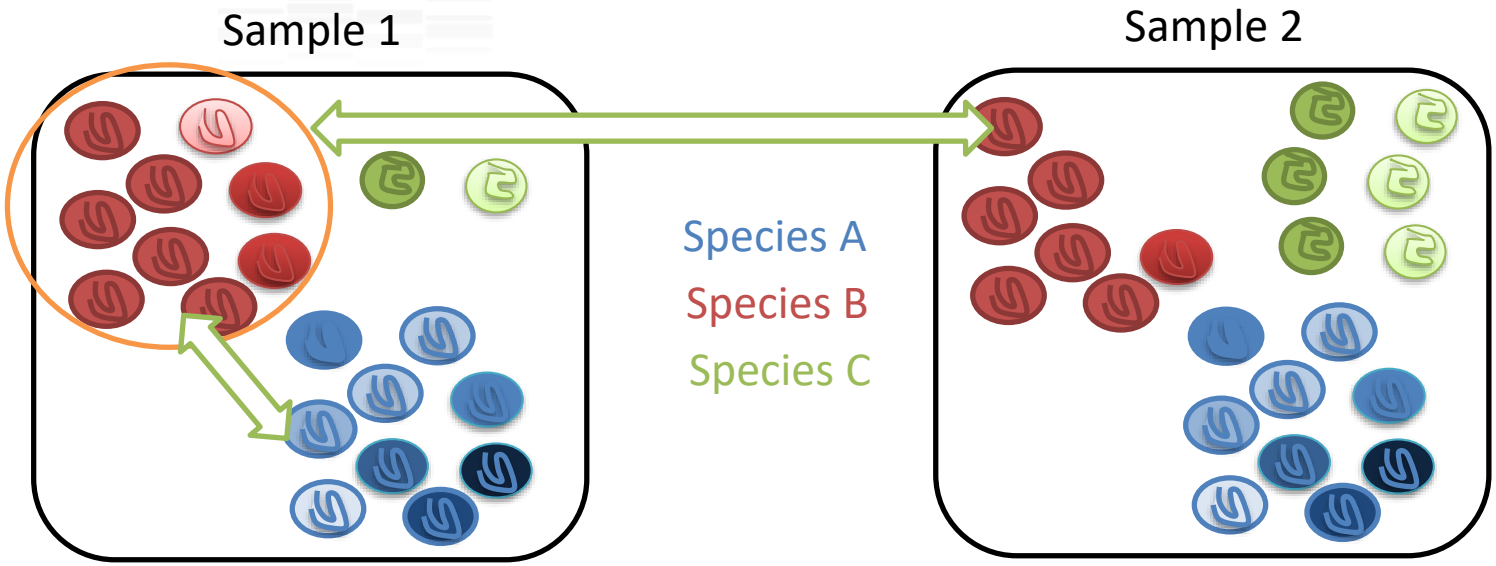
Intraspecific polymorphism in metagenomic shotgun datasets

Anne-Laure Abraham, Hélène Chiapello, Pierre Nicolas

AG du département MIA/NUMM – 22 mai 2019

Intraspecies polymorphism in metagenomic datasets

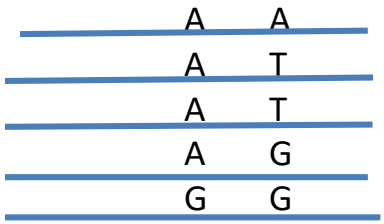


Intra species diversity

Inter species comparisons (same sample)
Inter sample comparisons (sample species)

Two estimators of nucleotide diversity in population genetics

Genetic variation in population is determined by factors such as **migration, population size, mutation rate and natural selection**.



L=100
n=5

L = sequence Length (number of aligned nucleotides)
n = number of sequences
$d_{i,j}$ = number of differences between sequences i and j

Number of segregating sites

S=2

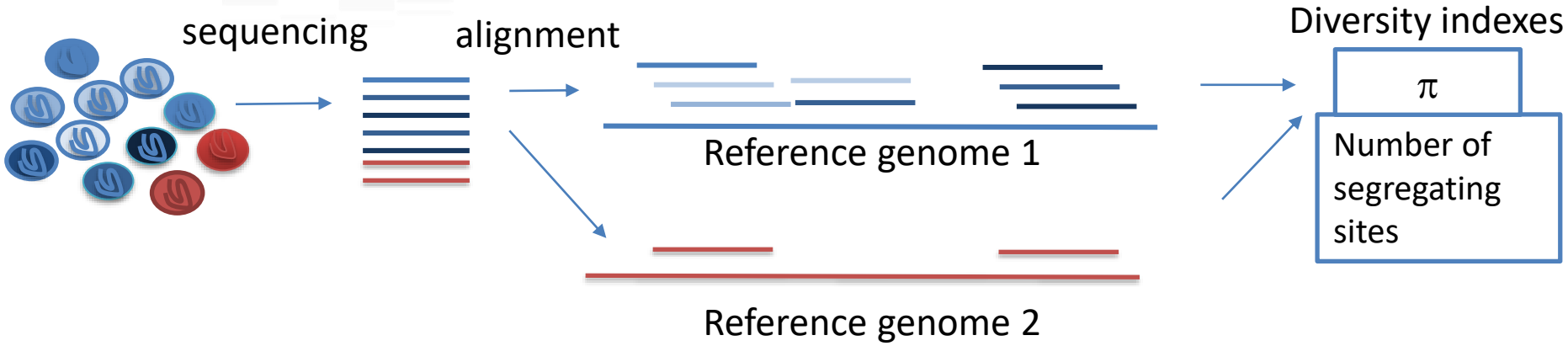
Pairwise nucleotide diversity

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

$$\pi = \frac{1}{100} \times \frac{1}{10} \times 12 = 0{,}012$$

Number of segregating sites (S) & pairwise nucleotide diversity (π) are the **two main summary statistics** used in **population genetics** to measure the amount of genetic variation in a population.
Usually π intra species ≤ 0.03

Challenges in polymorphism study in metagenomic datasets



1- Variable coverage between species (high / low abondance species)

2- Variable coverage along the genome (intrinsic randomness, core/pan genome)

3- Sequencing error vs true polymorphism, especially for low abundant species

4- Choice of reference genomes

Strategy

sequencing

Read quality filters

alignment

2-Alignment tool & parameters?

1-Ecosystem choice?

fastq_quality_trimmer

3-Quality threshold?

Base quality filters

- **Identify true polymorphisms** in metagenomic datasets
- Adaptation of indexes to **metagenomic data**

Samtools Mpileup

Allelic frequencies

4- Index computation?

Diversity indexes

Strategy

sequencing

Read quality filters

alignment

2-Alignment tool & parameters?

1-Ecosystem choice?

fastq_quality_trimmer

3-Quality threshold?

Base quality filters

Choice of an ecosystem for the proof of concept

Caracteristics:

- Many samples sequenced
- A lot of reference genomes
- Expertise of local microbiologists colleagues

Samtools Mpileup

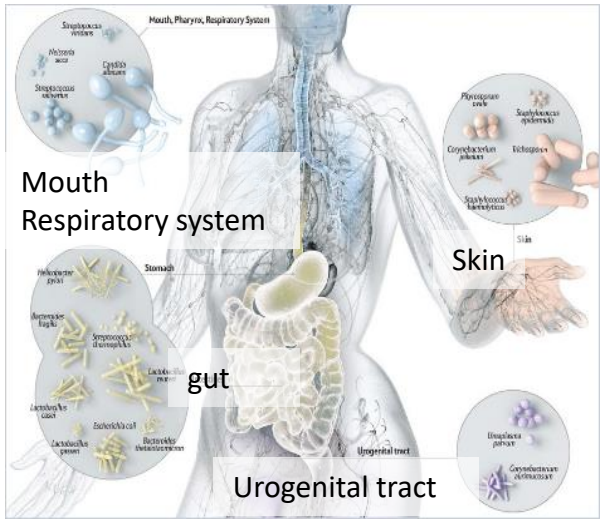
Allelic frequencies

4- Index computation?

Diversity indexes

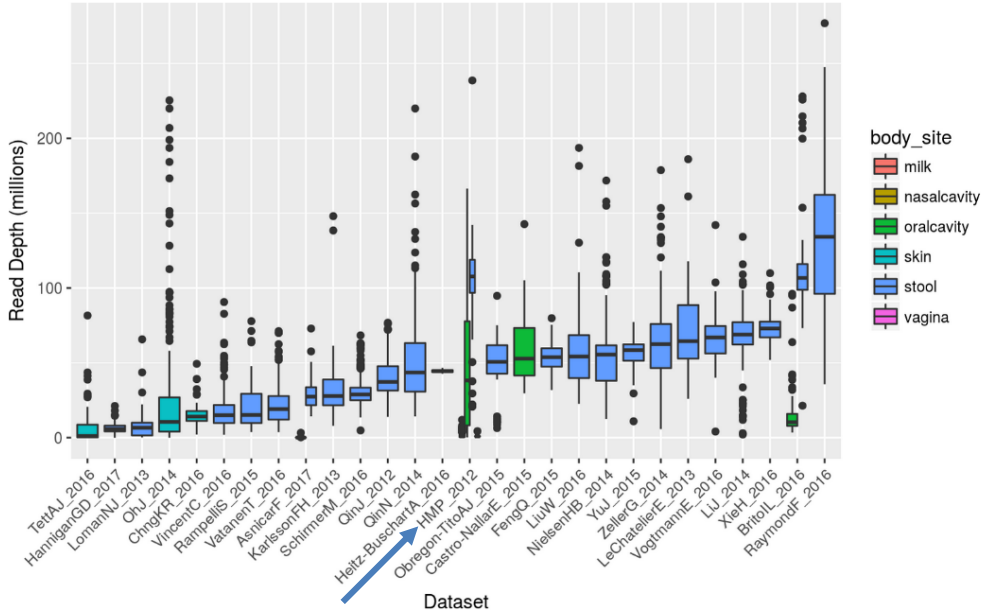
Ecosystem choice

- Lots of reference genomes
  - Gut: 1 161 species assemblied from 1 267 samples (Plaza Onate, Bioinformatics, 2019)
  - All human microbiota: 4 930 species assemblied from 9 428 samples (Pasoli, Cell, 2019)
- Many samples sequenced
  - 28 metagenomic studies – human microbiota (Pasoli, Nature methods, 2017)



749 samples, gut & oral cavity

(HMP consortium, Nature 2012)



Strategy

sequencing

Read quality filters

alignment

2-Alignment tool & parameters?

1-Ecosystem choice?

fastq_quality_trimmer

3-Quality threshold?

Base quality filters

Identify polymorphisms in metagenomic datasets

Usually > 95% identity between strains of the same species

Allow alignment with 95% identity

Test of 3 tools: Bowtie, Bowtie2, BWA-MEM

4 samples

4 reference genomes of Bacteroides ovatus

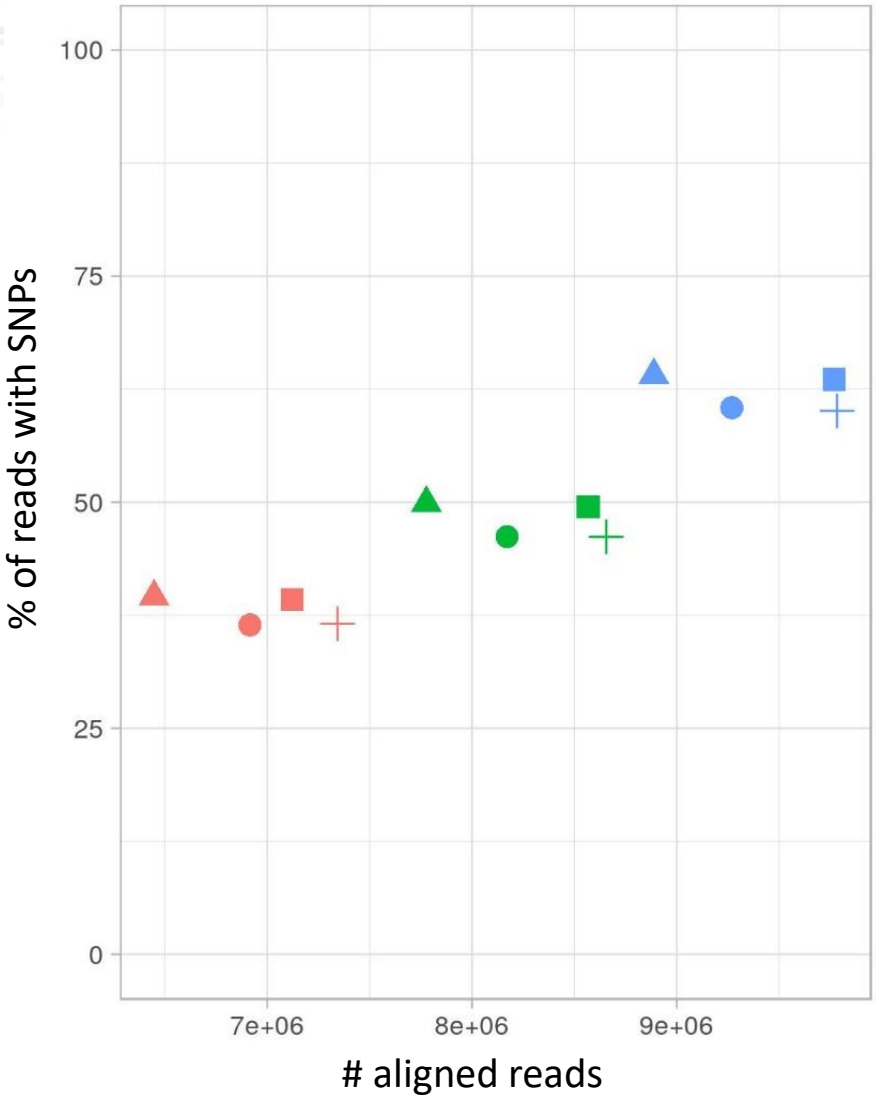
Samtools Mpileup

Allelic frequencies

4- Index computation?

Diversity indexes

Sensitivity of alignment tools to polymorphisms



aligned reads : 8
% of reads with SNPs: 2/8x100

mapper
- Bowtie
- Bowtie2
- BWA

GenomeName
- Bacteroides_ovatus_3_8_47FAA_3_8_47FAA
- Bacteroides_ovatus_ATCC_8483
- Bacteroides_ovatus_CL02T12C04_CL02T12C04
- Bacteroides_ovatus_CL03T12C18_CL03T12C18

Same results for 4 other samples

BWA-MEM align more reads, allow more SNP
BWA-MEM performs local alignment

Strategy

sequencing

Read quality filters

alignment

2-Alignment tool & parameters?

1-Ecosystem choice?

fastq_quality_trimmer

3-Quality threshold?

Base quality filters

Filter sequencing error/ converve true polymorphism

Test of base quality threshold:

5 base quality threshold [0-38]

5 reference genomes of Bacteroides ovatus

1 sample

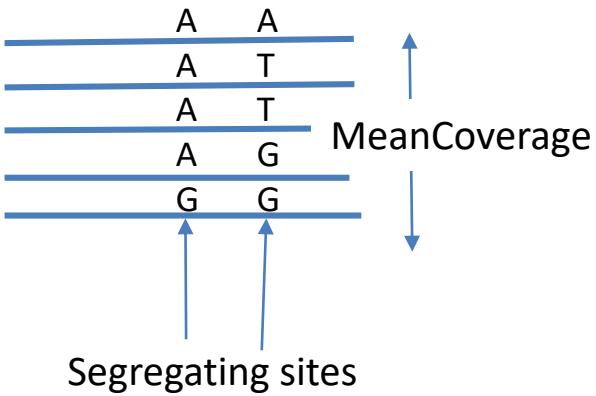
Samtools Mpileup

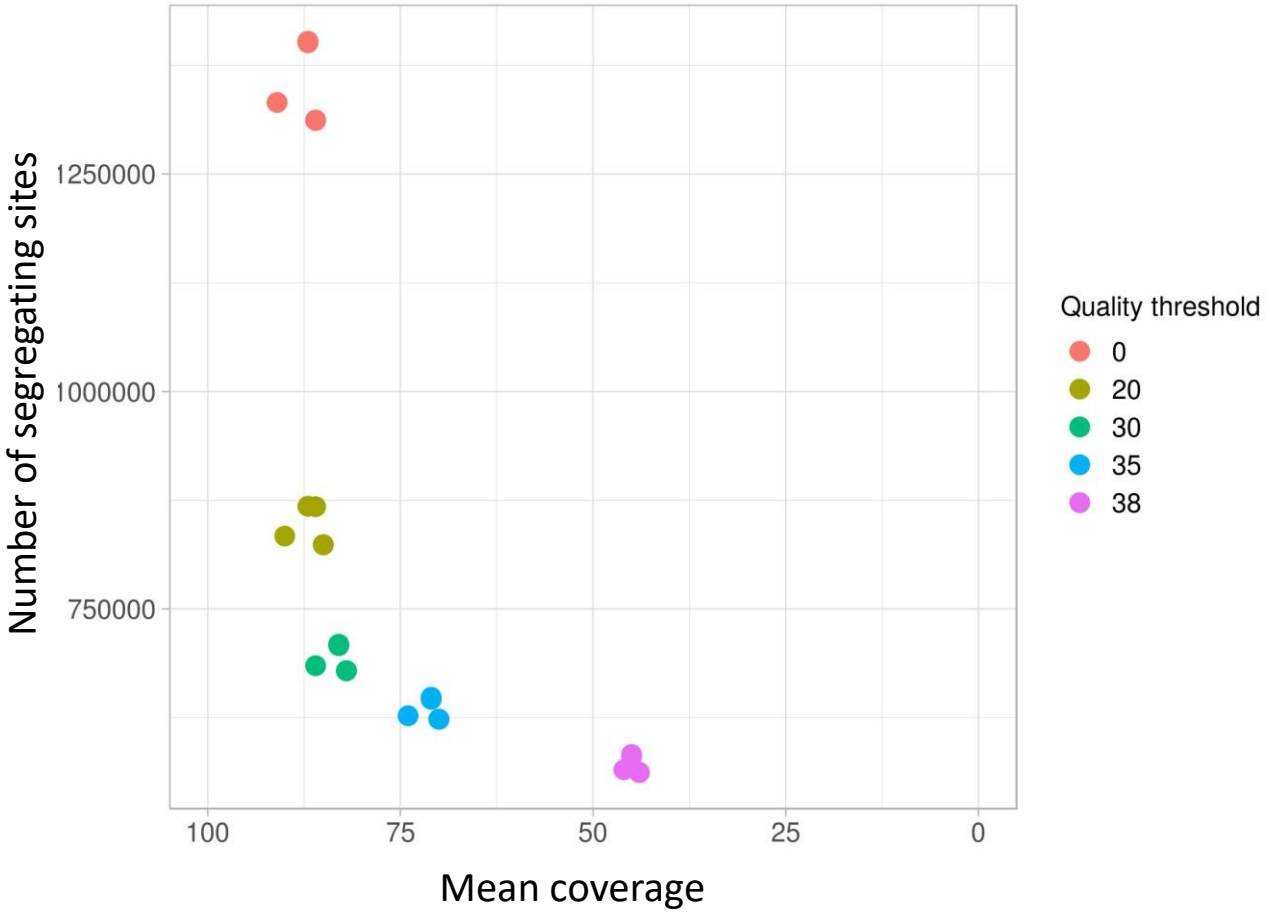
Allelic frequencies

4- Index computation?

Diversity indexes

Can we rely on base quality to filter sequencing errors?





S is very sensitive to sequencing errors since it gives the same weight to each detected SNP
Quality filtering is usefull to remove some sequencing error
Quality threshold:35

Strategy

sequencing

Read quality filters

alignment

2-Alignment tool & parameters?

1-Ecosystem choice?

fastq_quality_trimmer

3-Quality threshold?

Base quality filters

Allelic frequencies

Samtools Mpileup

Adaptation of indexes to metagenomic data

- Index independent from reference genome
- Index robust for low abundant species
- Minimal coverage to compute diversity indexes?

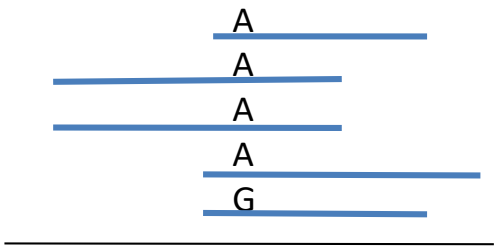
4- Index computation?

Diversity indexes

A new π index for diversity evaluation in metagenomics data

Pairwise nucleotide diversity

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



n_l = number of nucleotides at position l

$n_{l,x}$ = number of nucleotide x at position l

L=1

n_l=5

For each site:

$$\pi_l = 1 - \sum_{x\in\{a,c,g,t\}} \frac{n_{l,x}(n_{l,x}-1)}{n_l(n_l-1)}$$

For the whole genome: Pairwise nucleotide diversity

$$\pi = \frac{1}{L}\sum_{l=1}^{L} \pi_l$$

π_u

unweighted, depending on coverage (nl)

$$\pi_u = \frac{1}{\sum_{l=1}^{L} I\{n_{l\geq 2}\}}\sum_{l=1}^{L} \pi_l \; I\{n_{l\geq 2}\}$$

π_w

Weighted, independant from coverage (nl)

$$\pi_w = \frac{1}{\sum_{l=1}^{L} \frac{n_l(n_l-1)}{2} I\{n_{l\geq 2}\}}\sum_{l=1}^{L} \pi_l \frac{n_l(n_l-1)}{2} I\{n_{l\geq 2}\}$$

π is independant from the reference genome

Test of minimal coverage to compute pi

π should be robust regarding reference genome choice
can π be computed for low abundance genomes?

2 samples
Several coverages tested from 100 X to 0.005 X

alignment

5 Bacteroidetes ovatus genomes

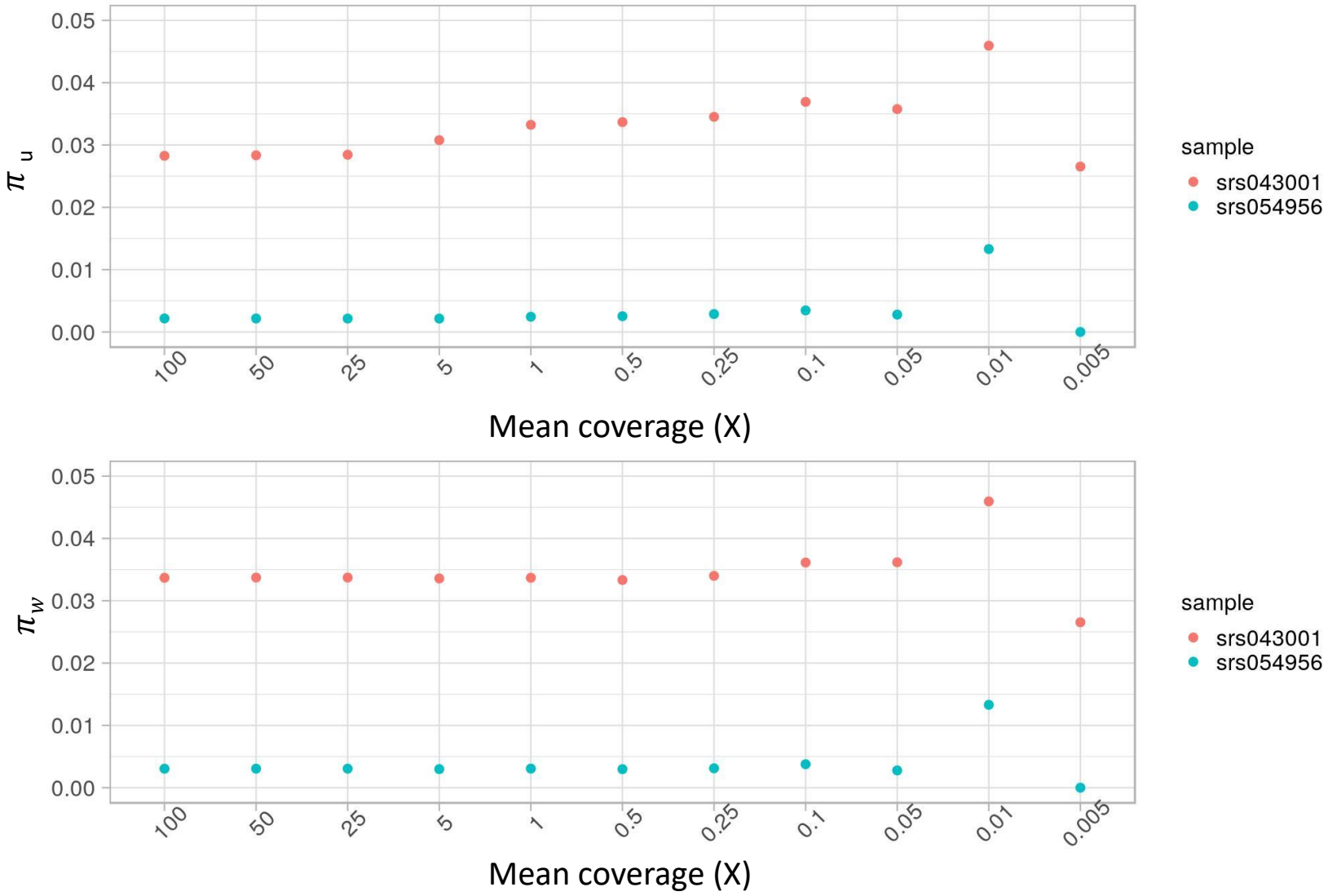
A
T
T
G
G

Allelic frequencies

π

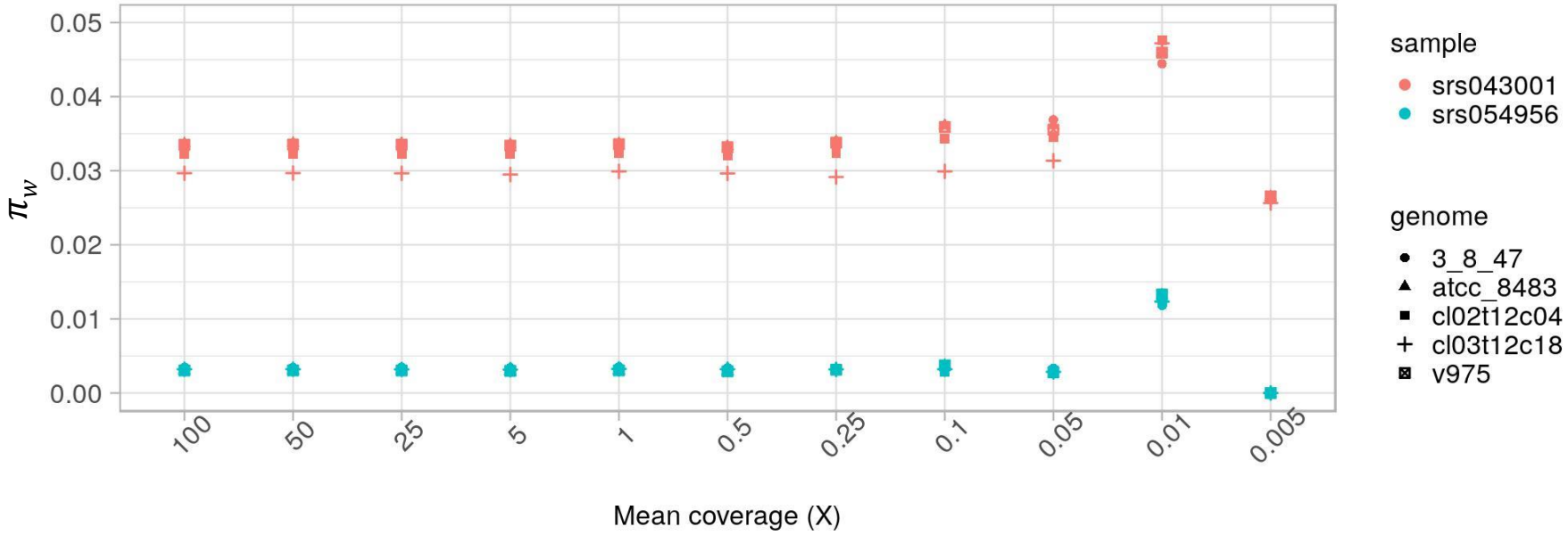
- Samples and species chosen because of high coverage (185X to 253X)
- Reference genomes chosen because of completeness (< 30 contigs)

Minimal coverage for pi computation



π_w is more robust when coverage varies and is adapted for low abundance species
0.1X means ~60,000bp of overlapping reads for a 6Mbp genome which is enough to estimate a π (a binomial proportion)

Reference genome choice for pi computation



π_w is relatively robust regarding reference genome choice

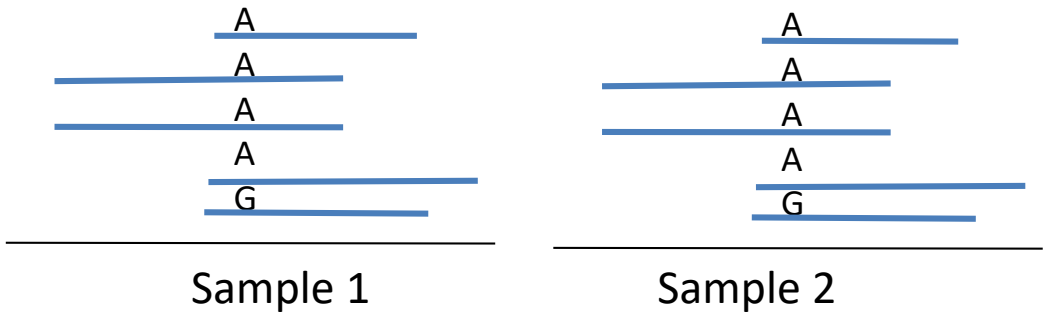
Conclusions

- First preliminary results of intraspecies diversity study
- First results of π computation :
  - π independent from reference genome
  - π adapted for low abondance species
- Whole pipeline under construction

Perspectives

- Intra-sample diversity improvement
  - Compute π on a conserved part of the genome (core genome) to have more robust results
  - Test of the whole pipeline & large scale computation

- Compute inter-sample divergence



- Analyses of the patterns of polymorphism described by intra-sample diversity and inter-sample divergence (networks).

- Website ?

Thanks to

MaIAGE
Mathématiques et Informatique Appliquées du Génome à l'Environnement

StatInfOmics team

Pierre Nicolas
Hélène Chiapello

Gwenaëlle André-Leroux
Jean-François Gibrat
Cyprien Guérin
Thomas Lacroix
Mahendra Mariadassou
Sandra Plancade
Sophie Schbath

Migale Plateform

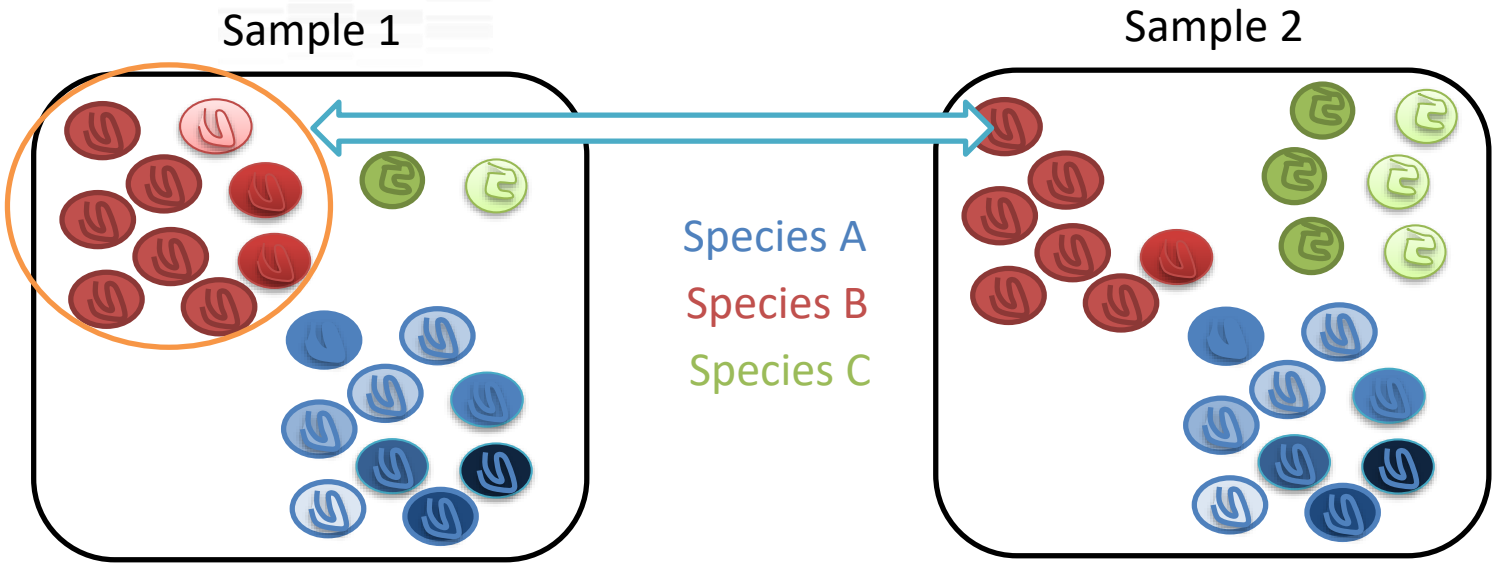
Valentin Loux

Sandra Dérozier
Olivier Inizan
Véronique Martin
Olivier Rué
Valérie Vidal
Cédric Midoux

INRA
SCIENCE & IMPACT

Merci !

Intraspecies polymorphism in metagenomic datasets



Intra species diversity

In one sample

Between 2 samples

Inter species comparisons (same sample)
Inter sample comparisons (sample species)

Inter sample comparison (pairs of samples)
Inter species comparison (pairs of sample)

Intraspecies polymorphism in human microbiota

A supprimer ?

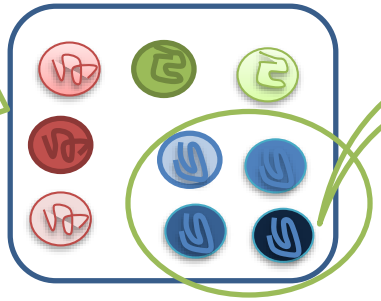
- Genetic diversity
  - 10.3 M SNP in 252 gut sample mapped on 1497 reference genomes (Schloissning Nature 2013)
  - 64.3% of species represented by at least 2 strains (Truong Genome Research 2017)
  - Only 3.67% of strains shared among individuals compared to 35.31% species shared between 2 individuals on average (Truong Genome Research 2017)
- Microbiota transmission
  - Evidence of oral-fecal transmission, with increased level for colorectal cancer and rheumatoid arthritis patients (Schmidt eLife 2019)
  - Evidence of transmission of microbiota from mother to child (Ferretti Cell Host & Microbe 2018)
- Several study at the strain level to identify pathogen or for personalized medicin
  - À compléter
- Few study on strain diversity

3 outils d'alignement testés

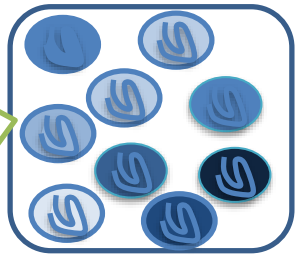
Outil	ref	lien	mismatches	Alignements multiples
Bowtie	Langmead B, Trapnell C, Pop M, Salzberg SL. Genome Biol 10:R25.	http://bowtie-bio.sourceforge.net	3 au total	1 alignement au hasard si plusieurs possibles
Bowtie 2	Langmead B, Salzberg S.. Nature Methods. 2012, 9:357-359.	http://bowtie-bio.sourceforge.net/bowtie2	1 dans la graine (graine : 16 nt)	1 alignement au hasard si plusieurs possibles
BWA short reads	Li H. and Durbin R. (2009). Bioinformatics, 25:1754-60.	http://bio-bwa.sourceforge.net/	Graine exacte : 16 nt Pénalité de mismatch	Suppression des alignements multiples

Intraspecies polymorphism in metagenomic datasets

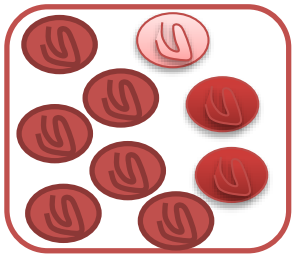
Metagenomic sample



Microbial diversity
Several species



Species A

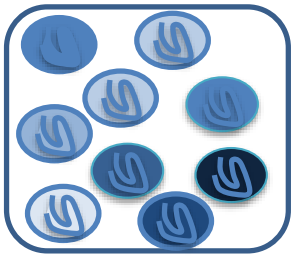


Species B

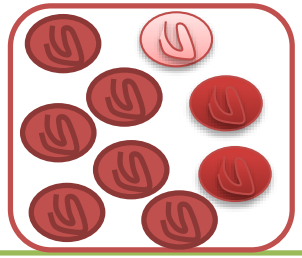
What is the natural diversity for each species?

Compare natural diversity between species?

Compare natural diversity between samples for each species?



Species A



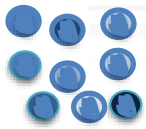
Species B

Diversity for different samples?

Diversity between ecosystems?

Intra species polymorphisms

Polymorphism study in metagenomic datasets – several approches



sample → sequencing → Metagenomic reads

MetaSNV

StranPhlan

Alignment BWA

- 5X coverage
- > 40% of genome covered

APPROCHES EN METAG - SOUCHES

PanPhlan

A
T
T
G
G

→ Allelic frequencies
Mpileup

Combination StrainPhlan - PanPhlan

Pairwise distance per species
Nucleotide diversity
Fixation index

- Lots of reference genome for human microbiota
- Challenges :
  - Sequencing error vs SNP
  - low coverage