Microbial Data Science

Universidad Andrés Bello

CBIB
CENTER FOR BIOINFORMATICS & INTEGRATIVE BIOLOGY

# Aproximaciones genómicas y bioinformáticas para el estudio de patógenos emergentes y reemergentes

**Juan A. Ugalde**

Centro de Bioinformática y Biología Integrativa

Facultad de Ciencias de la Vida

Universidad Andrés Bello

GENE2DIS
GENOMIC EPIDEMIOLOGY IN EMERGENT DISEASES

XVII Jornadas Científicas 2025. Instituto de Saludo Pública de Chile.

# Gene2DIS Project

**Cecilia Vial**
Director

**Rafael Araos**
Co-Director

**Pablo Vial**
Co-investigator

**Johanna Acevedo**
Co-investigator

**Marcela Ferres**
Co-investigator

**Katia Abarca**
Co-investigator

**Cecilia Poli**
Co-investigator

**Thomas Weitzel**
Co-investigator

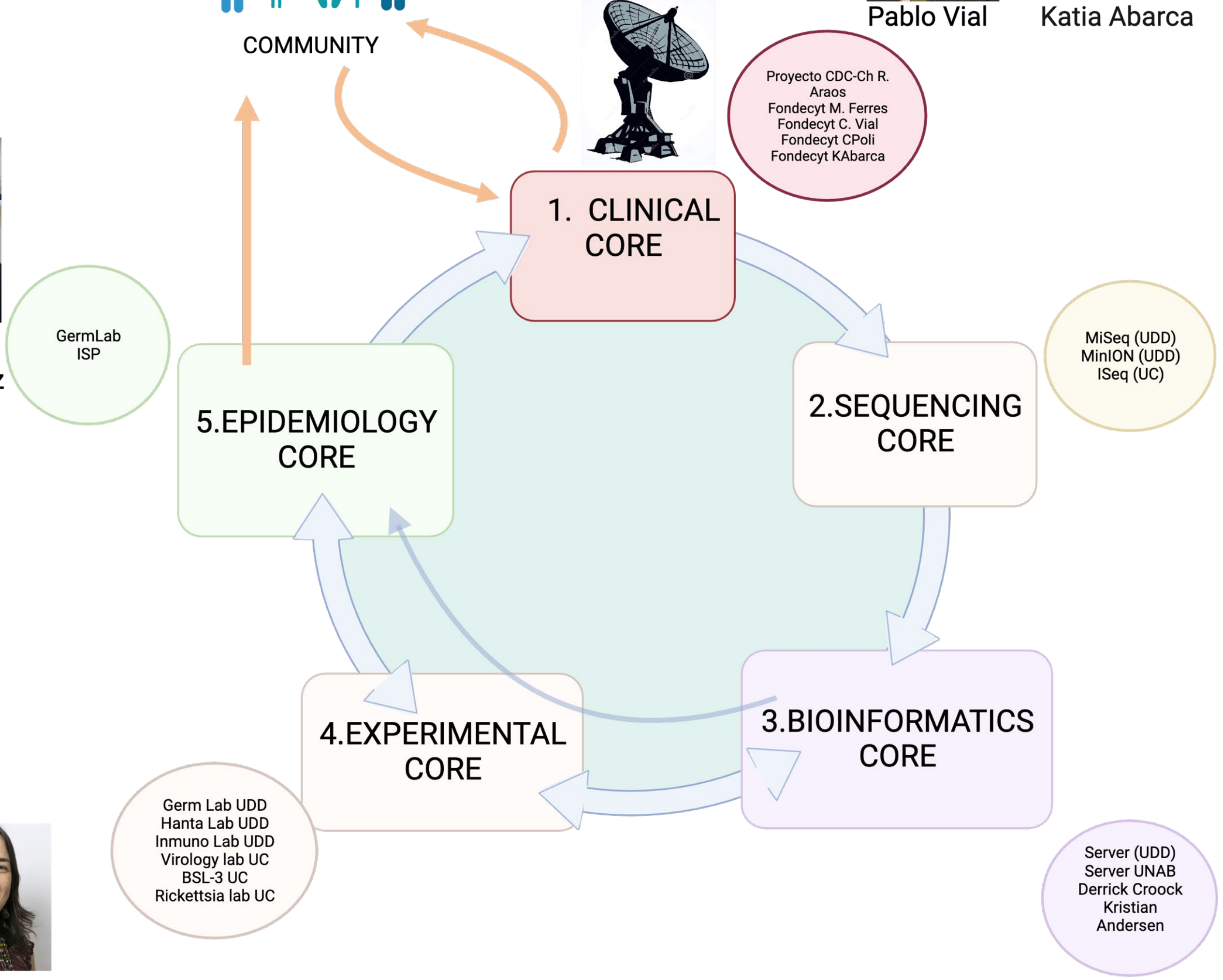
**Juan Ugalde**
Co-investigator

PONTIFICIA
UNIVERSIDAD
CATÓLICA
DE CHILE

ICIM UDD
INSTITUTO DE CIENCIAS E INNOVACIÓN EN MEDICINA
Facultad de Medicina
Clínica Alemana - Universidad del Desarrollo

Universidad
Andrés Bello®
Conectar • Innovar • Liderar

# Collaborative work



Proyecto CDC-Ch R. Araos
Fondecyt M. Ferres
Fondecyt C. Vial
Fondecyt CPoli
Fondecyt KAbarca

MiSeq (UDD)
MinION (UDD)
ISeq (UC)

Server (UDD)
Server UNAB
Derrick Croock
Kristian Andersen

Germ Lab UDD
Hanta Lab UDD
Inmuno Lab UDD
Virology lab UC
BSL-3 UC
Rickettsia lab UC

GermLab
ISP


Pablo Vial


Katia Abarca


Thomas Weitzel


CeciliaVial


Marcela Ferres


Juan Ugalde


Rafael Araos


Johanna Acevedo


Jorge Fernandez


Cecilia Vial


Marcela Ferres


Cecilia Poli

# Bioinformatics and Data Science

- Bioinformatics tools and pipelines for the analysis of microbial genome datasets.
- Phylogenomics and comparative genomics of microbial pathogens.
- Microbiome studies (human and environmental)

Microbial Data Science

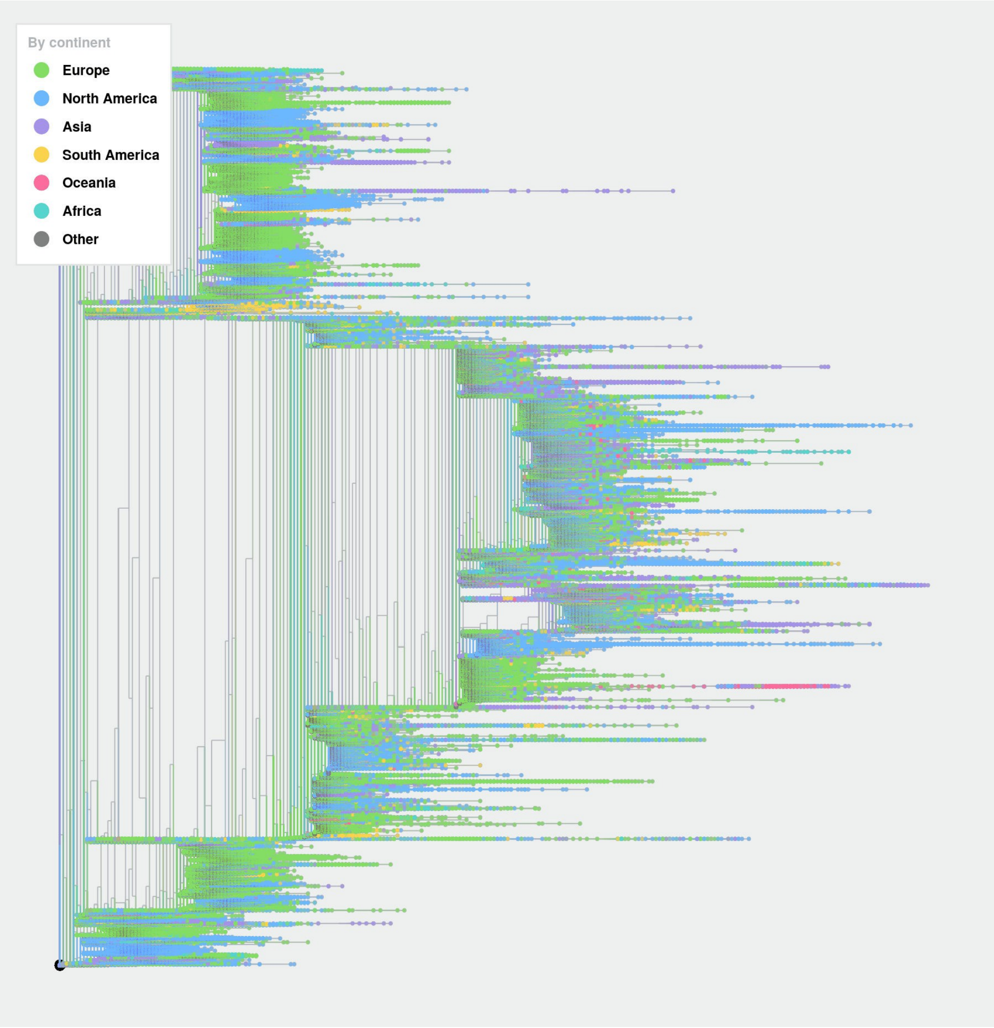
Universidad Andrés Bello

CENTER FOR BIOINFORMATICS & INTEGRATIVE BIOLOGY

# SARS-CoV-2

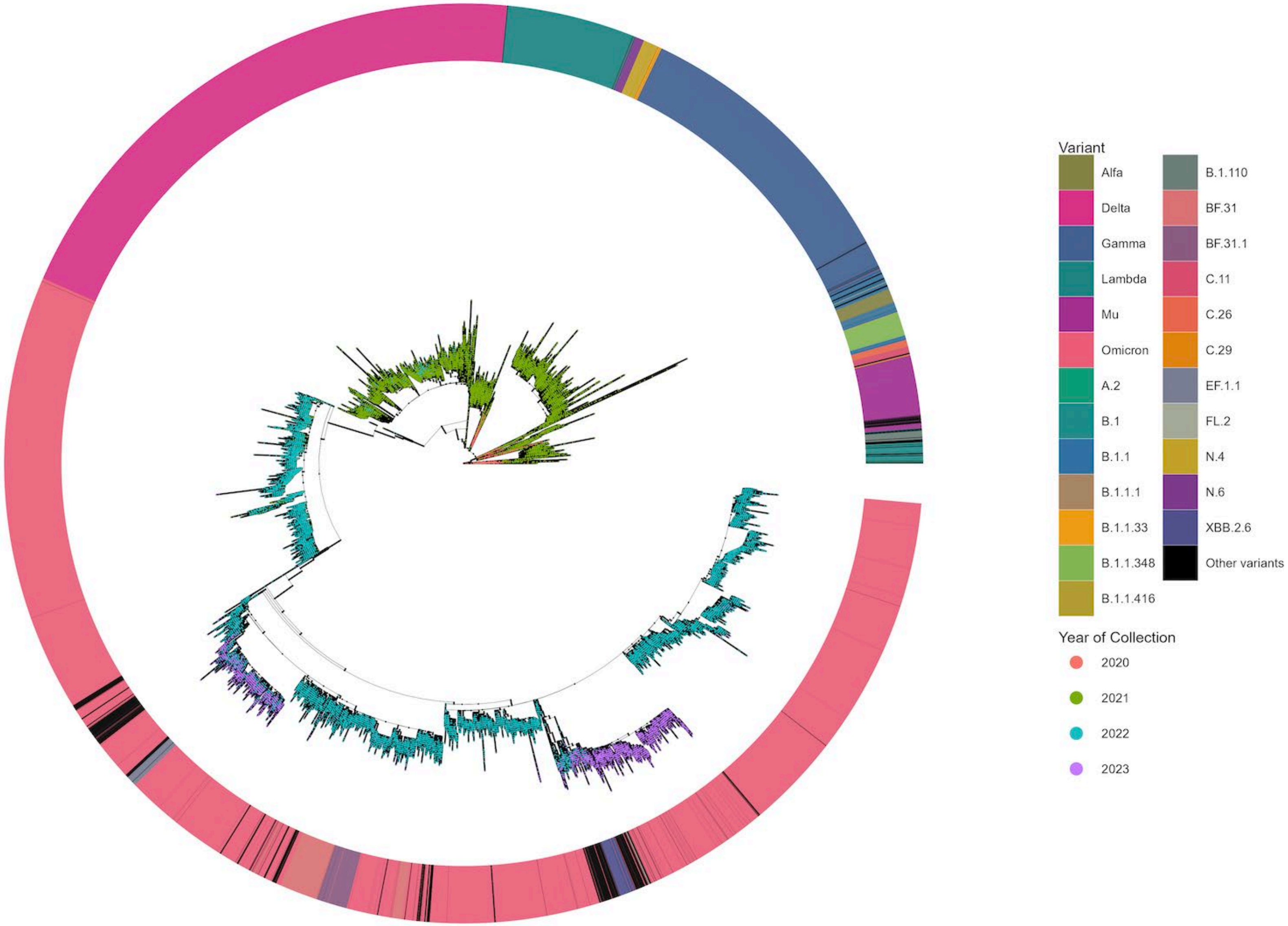
# Opportunities for phylogenomics analysis of SARS-CoV-2 in Chile

- **World:** 12,643,193 genomes
- **South America:** 407,407genomes
- **Chile:** 45,000 genomes



Shu, Y., & McCauley, J.(2017). https://doi.org/10.2807/1560-7917

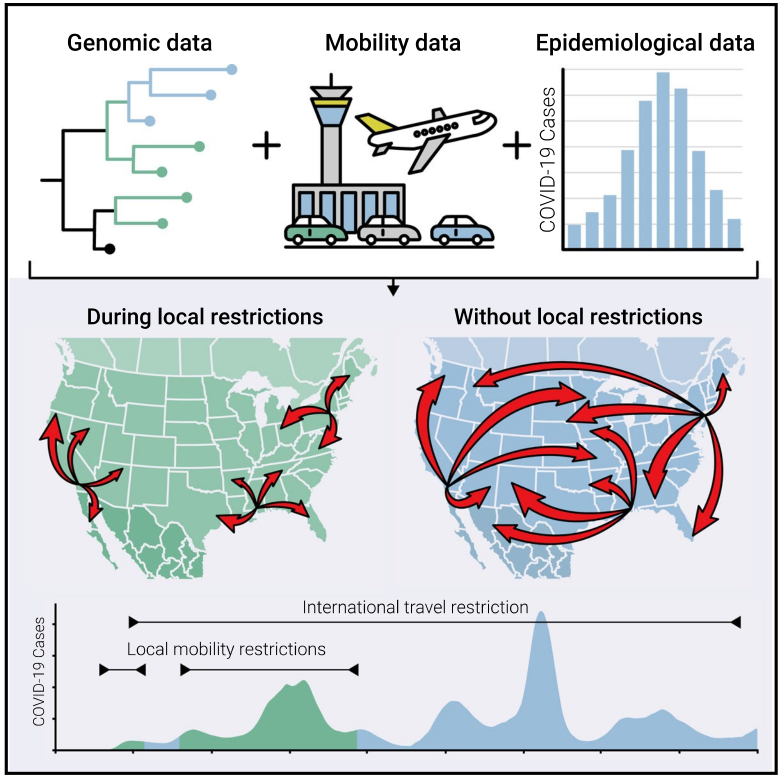
# Phylogenomics of SARS-CoV-2 in Chile



Ignacio Ramo

# Genomic surveillance reveals dynamic shifts in the connectivity of COVID-19 epidemics

## Graphical abstract



## Authors

Nathaniel L. Matteson, Gabriel W. Hassler,
Ezra Kurzban, ...,
Abraham Campos-Romero, Shirlee Wohl,
Mark Zeller

## Correspondence

natem@scripps.edu (N.L.M.),
andersen@scripps.edu (K.G.A.),
zellerm@scripps.edu (M.Z.)

## In brief

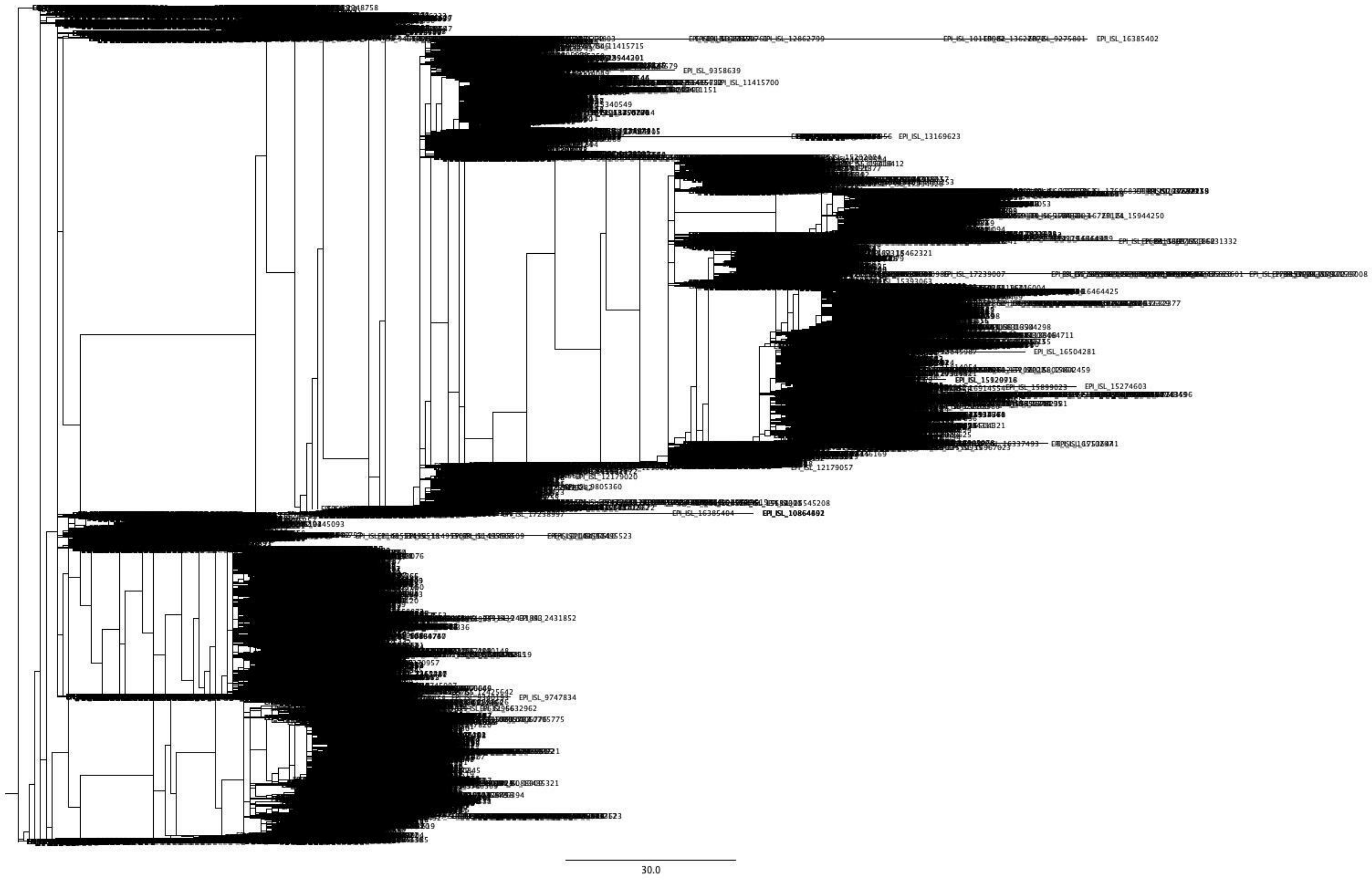
Genomic surveillance, paired with mobility and epidemiological data, quantifies the impact of local and international travel restrictions on SARS-CoV-2 transmission. Both phylogenetic and mobility analyses indicate that collaborative interventions are more effective than targeted border closures at reducing the transmission of SARS-CoV-2 between highly connected locations.

## Highlights

- Phylogenetic similarity of virus populations suggests connectivity between locations
- COVID-19 mandates contained the spread of SARS-CoV-2 in the US
- The lifting of mandates enabled SARS-CoV-2 to spread further as travel increased
- Border closures to non-essential travel minimally impacted cross-border transmission

ANDERSENlab
The Scripps Research Institute

# Full South America phylogenomic tree



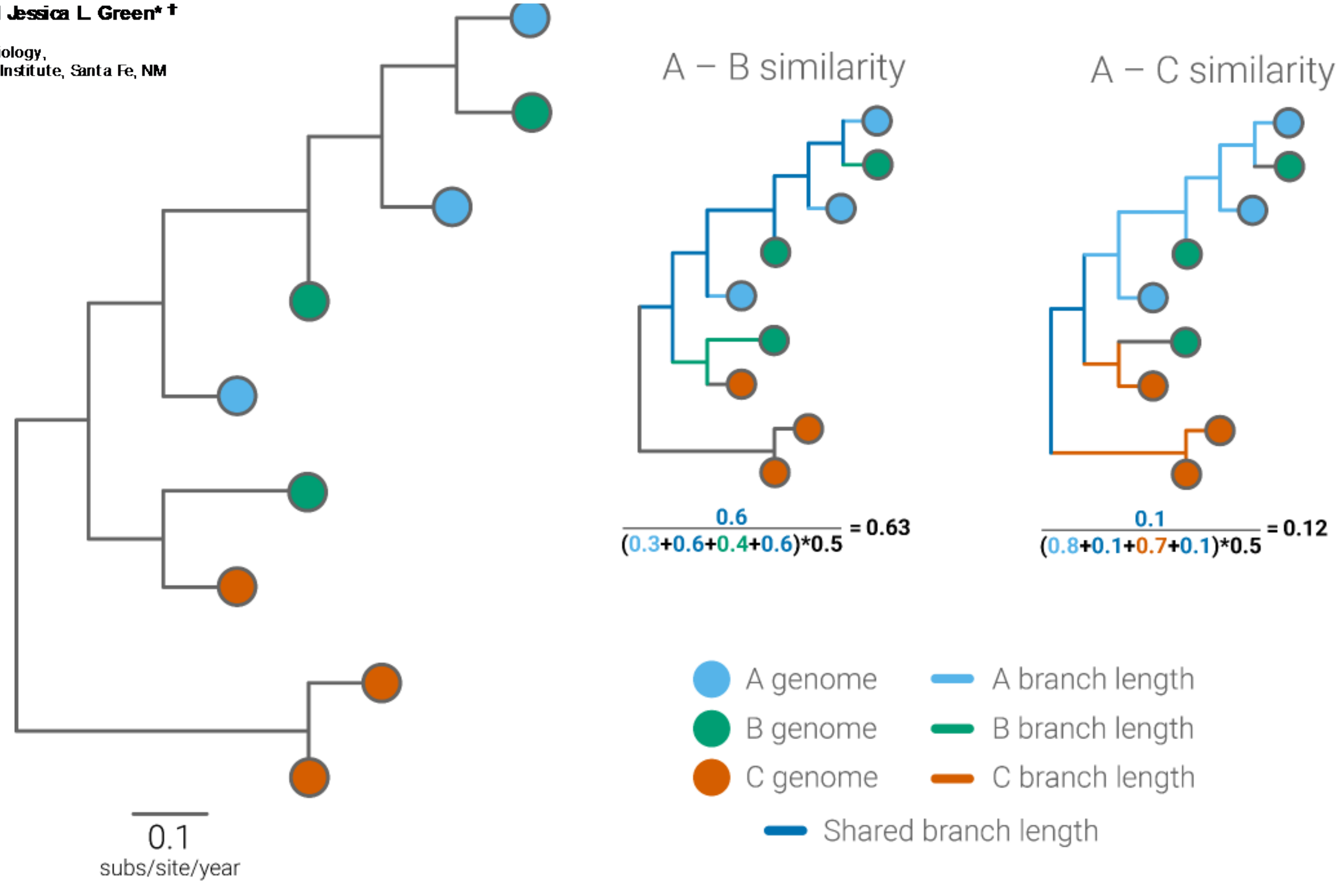
# SARS-CoV-2 phylogenetic distances between countries

## Microbes on mountainsides: Contrasting elevational patterns of bacterial and plant diversity

Jessica A. Bryant[*,†], Christine Lamanna[‡], Hélène Morlon[*], Andrew J. Kerkhoff[§], Brian J. Enquist[‡,¶,‖], and Jessica L. Green[*,†]

[*]Center for Ecology and Evolutionary Biology, University of Oregon, Eugene, OR 97403; [‡]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721; [§]Departments of Biology and Mathematics, Kenyon College, Gambier, OH 43022; [¶]Santa Fe Institute, Santa Fe, NM 87501; and Center for Applied Biodiversity, Science Conservation International, Arlington, VA 22202

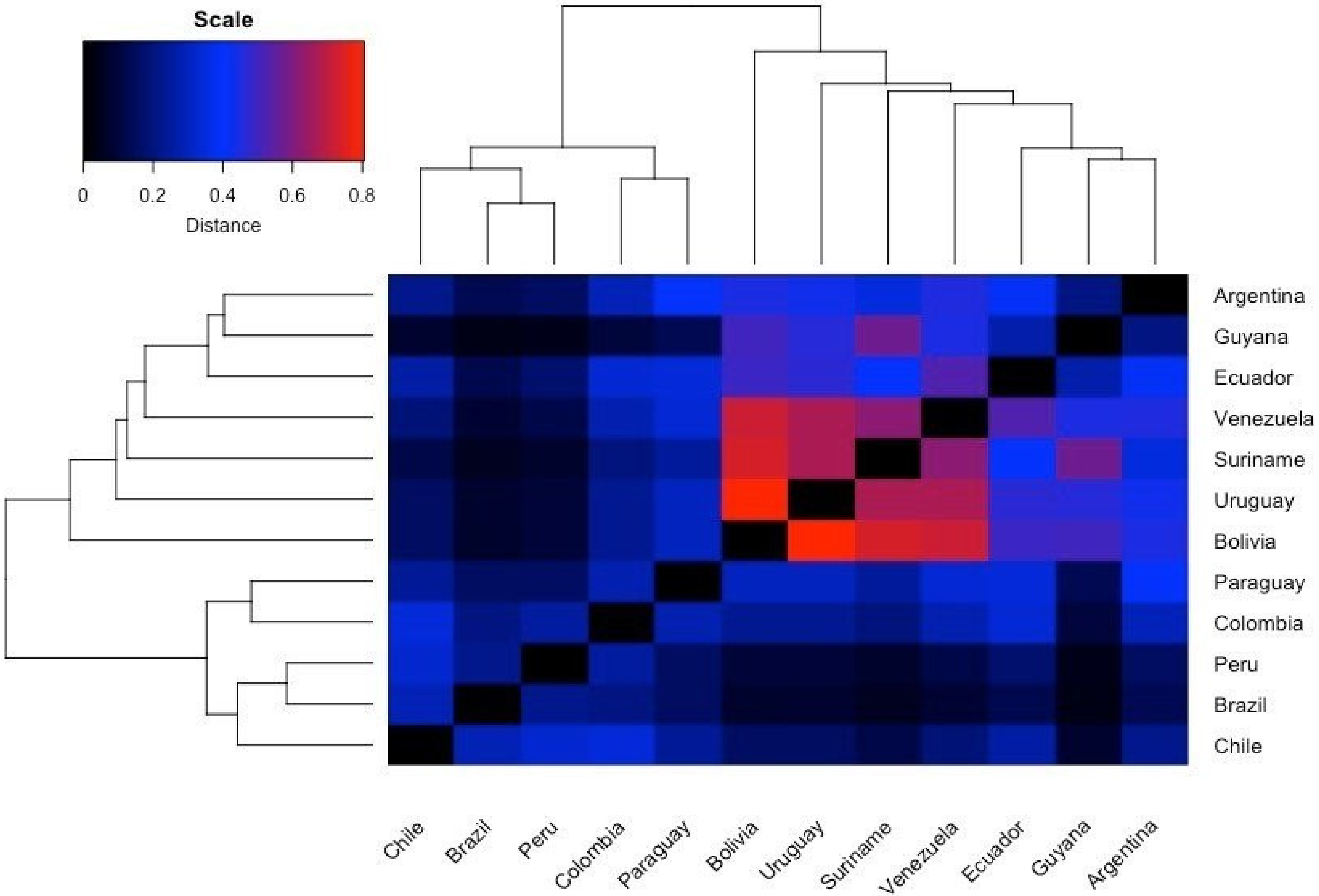
**Phylosor Analysis (Beta-diversity analysis)**



*Code soon available in*

# SARS-CoV-2 phylogenetic distances between countries

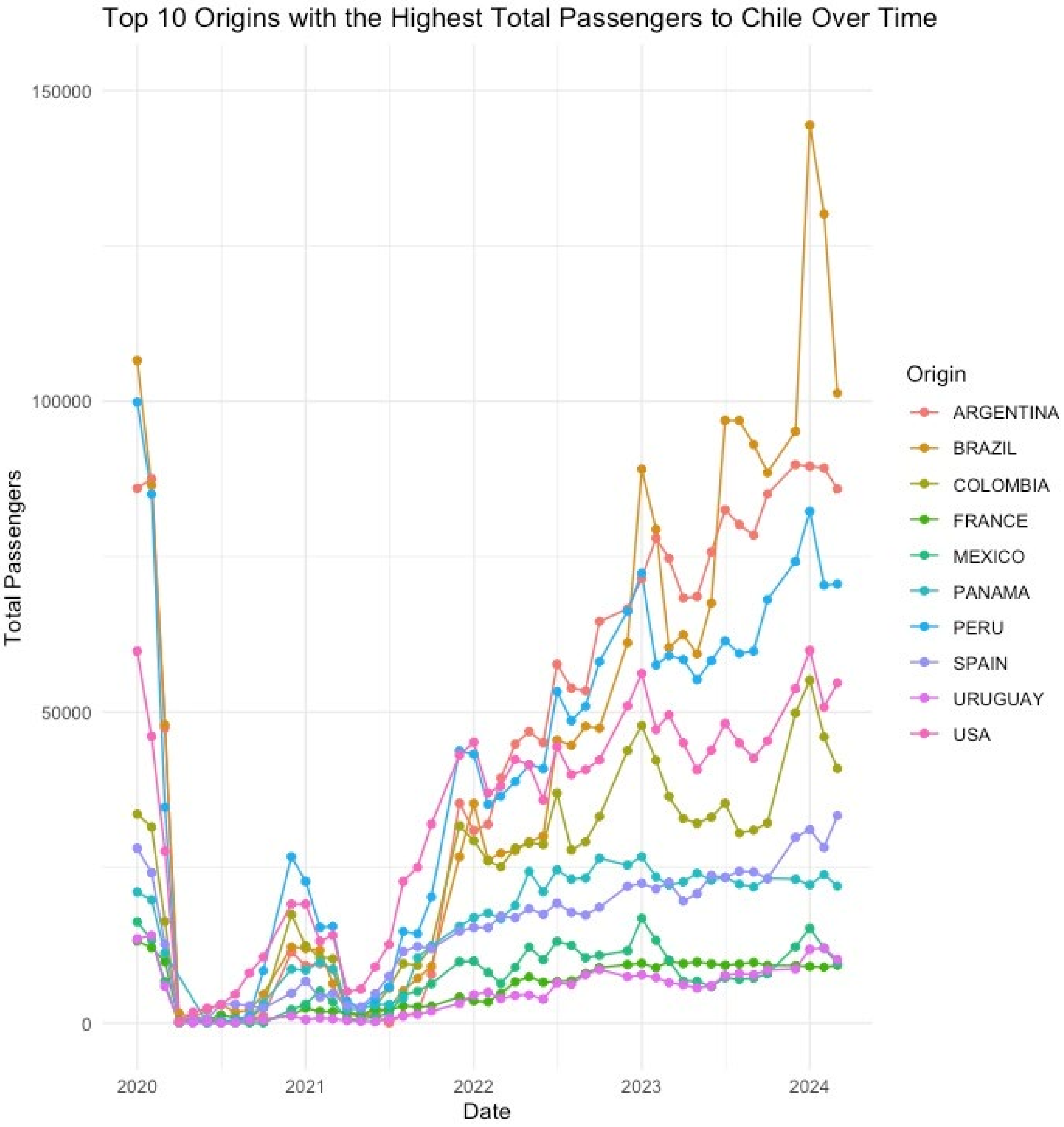
**Phylosor Analysis (Beta-diversity analysis)**



# Spatial phylodynamics of BA.4 subvariant

- 1,045 genomes of the BA.4 subvariant
- Reconstruction with BEAST, HYK model
- 1 million iterations, with sampling every 10,000 generations

# Air travel to Chi



# Next steps

- Modelling at the South America level using air travel information
- Modelling within Chile, using cell phone data (collaboration with Leo Ferres, Data Science Institute, UDD)

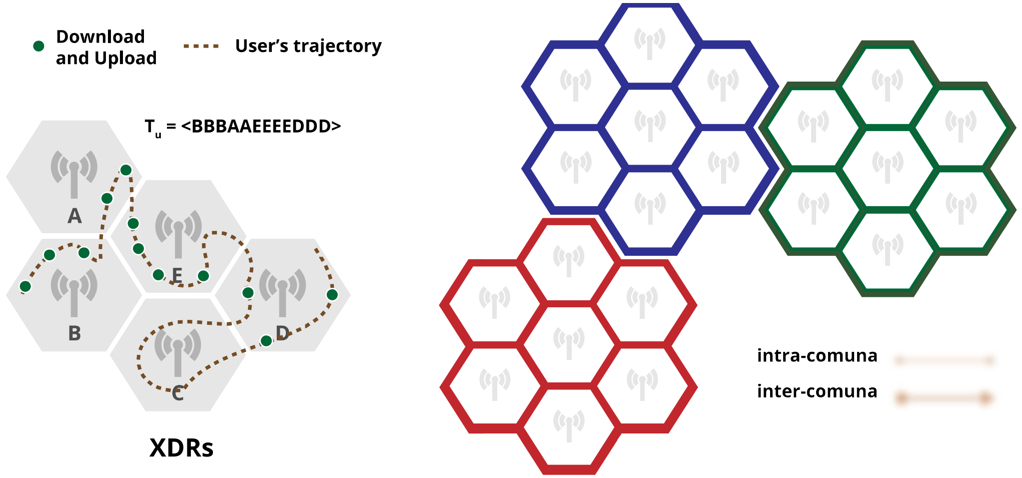
scientific **data**

Check for updates

**DATA DESCRIPTOR**

## A dataset to assess mobility changes in Chile following local quarantines

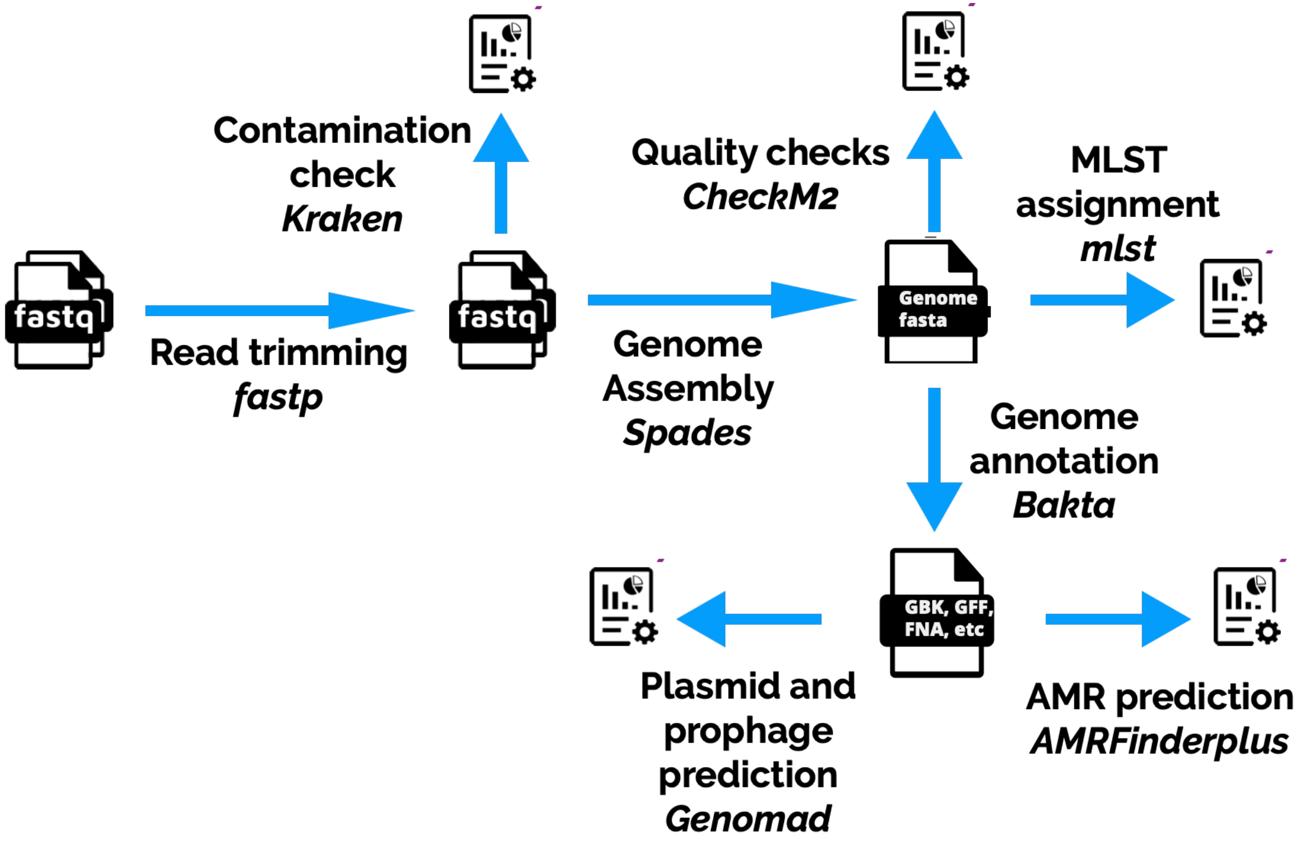
Luca Pappalardo[1], Giuliano Cornacchia[1,2], Victor Navarro[3], Loreto Bravo[3,4] & Leo Ferres[3,4,5] ✉



# Microbial Genomics

# Development and applications of tools for microbial genomics

- Support for ONT and Illumina
- Support for contig annotation
- Upcoming updates!



**https://github.com/gene2dis/mgap**

# *Streptococcus pyogenes*

- More than 275 *emm* types, based on the gene coding for the M protein.
- There is an association between human development index (HDI) and strain diversity.
- *emm1* is the most frequent in invasive infections in high-income countries, and is considered highly virulent.
- For example, in countries such as China, Hong Kong, UK, *emm12* and *emm1* are the most common causes of scarlet fever.

# Emergence of dominant toxigenic M1T1 *Streptococcus pyogenes* clone during increased scarlet fever activity in England: a population-based molecular epidemiological study

*Nicola N Lynskey*, Elita Jauneikaite*, Ho Kwong Li, Xiangyun Zhi, Claire E Turner, Mia Mosavie, Max Pearson, Masanori Asai, Ludmila Lobkowicz, J Yimmy Chow, Julian Parkhill, Theresa Lamagni, Victoria J Chalker, Shiranee Sriskandan*

## Summary

**Background** Since 2014, England has seen increased scarlet fever activity unprecedented in modern times. In 2016, England's scarlet fever seasonal rise coincided with an unexpected elevation in invasive *Streptococcus pyogenes* infections. We describe the molecular epidemiological investigation of these events.

**Interpretation** A dominant new *emm*1 *S pyogenes* lineage characterised by increased SpeA production has emerged during increased *S pyogenes* activity in England. The expanded reservoir of $\mathrm{M1_{UK}}$ and recognised invasive potential of *emm*1 *S pyogenes* provide plausible explanation for the increased incidence of invasive disease, and rationale for global surveillance.

Article https://doi.org/10.1038/s41467-024-47929-7

# Rapid expansion and international spread of $M1_{UK}$ in the post-pandemic UK upsurge of *Streptococcus pyogenes*

Received: 12 January 2024

Accepted: 15 April 2024

Ana Vieira[1,2,3], Yu Wan[1,3,4], Yan Ryan[5], Ho Kwong Li[1,2], Rebecca L. Guy[4], Maria Papangeli[1,2], Kristin K. Huse[1,2], Lucy C. Reeves[1,2], Valerie W. C. Soo[1,2], Roger Daniel[5], Alessandra Harley[5], Karen Broughton[5], Chenchal Dhami[5], Mark Ganner[5], Marjorie A. Ganner[5], Zaynab Mumin[5], Maryam Razaei[5], Emma Rundberg[5], Rufat Mammadov[5], Ewurabena A. Mills[1,2], Vincenzo Sgro[1], Kai Yi Mok[1], Xavier Didelot[6], Nicholas J. Croucher[7,8], Elita Jauneikaite[3,7,8], Theresa Lamagni[3,4], Colin S. Brown[3,4], Juliana Coelho[3,4,5] & Shiranee Sriskandan[1,2,3]

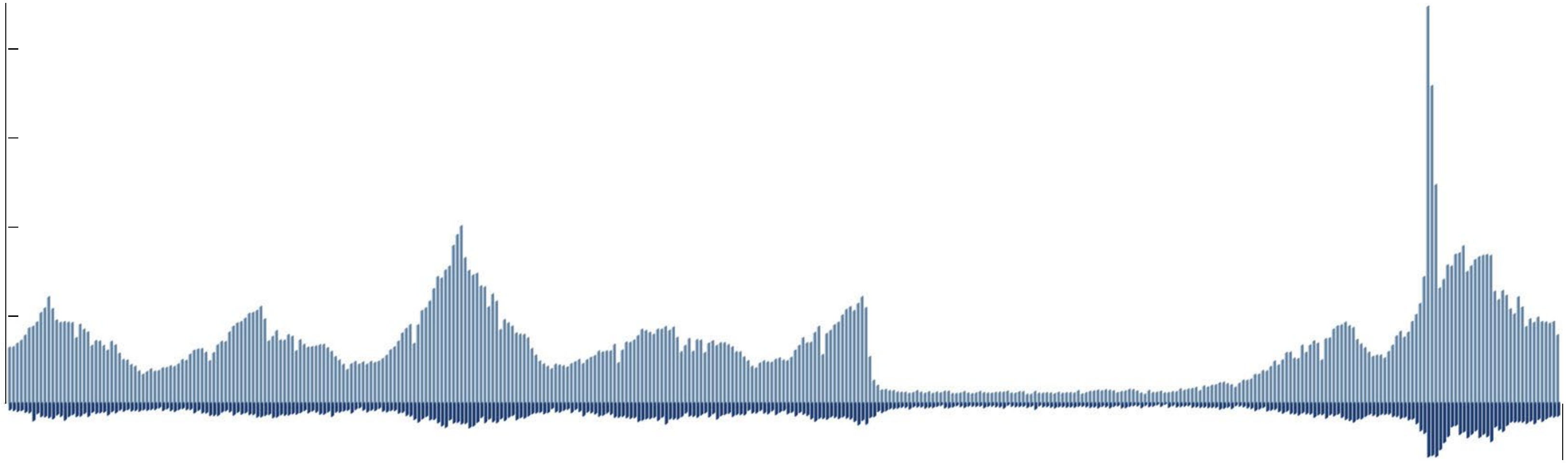
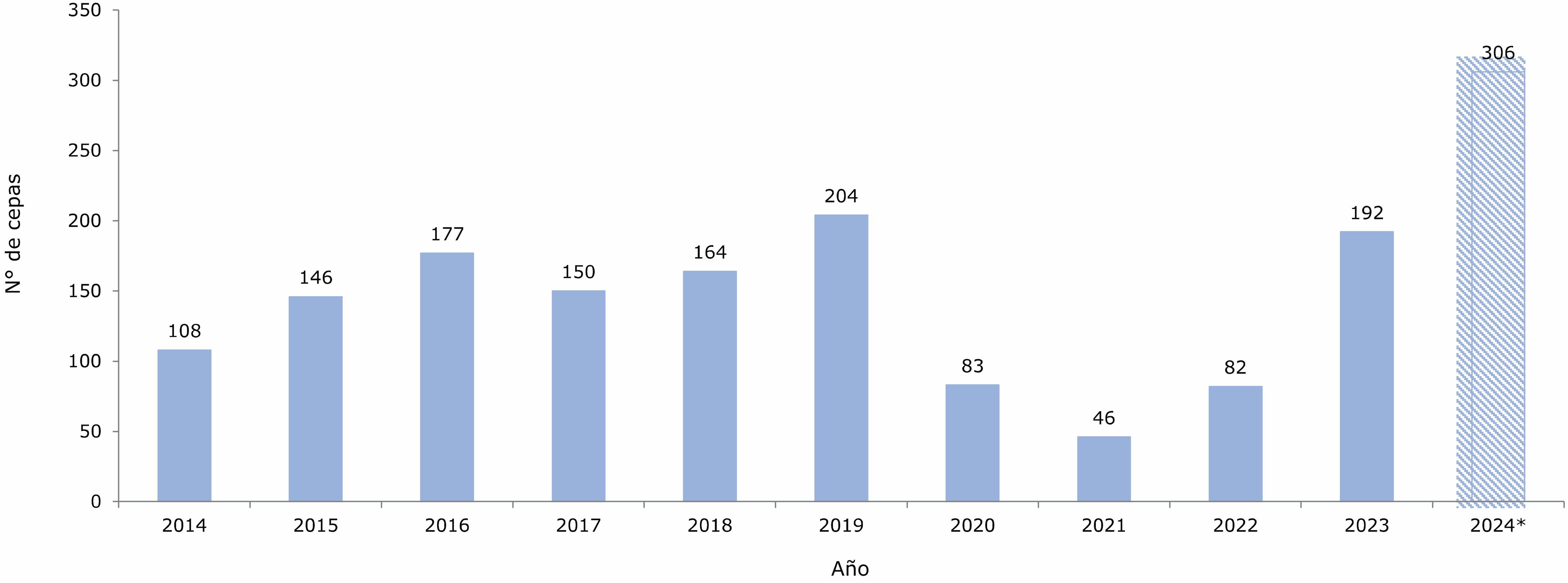


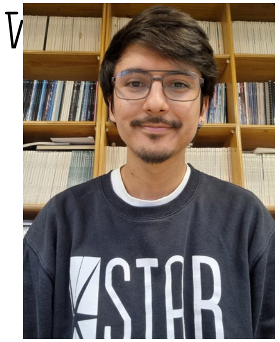
**Figura 1.** Número de cepas confirmadas de *S. pyogenes*, según año. Chile, 2014-2024*.



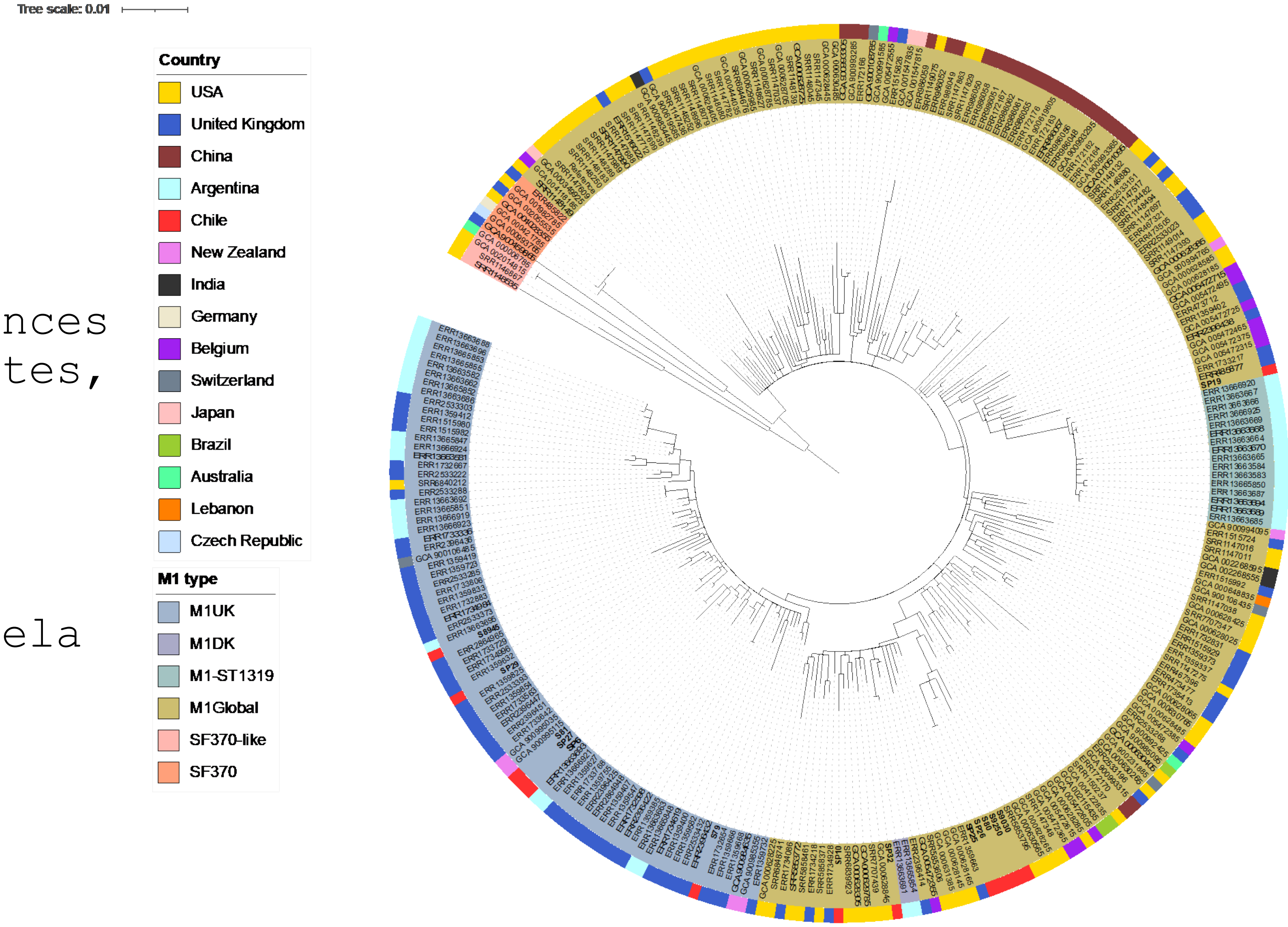
*: Datos provisorios al 7 de junio 2024
Fuente: Laboratorio de Agentes de Meningitis Bacteriana. Departamento de Laboratorio Biomédico.
Instituto de Salud Pública de Chile.

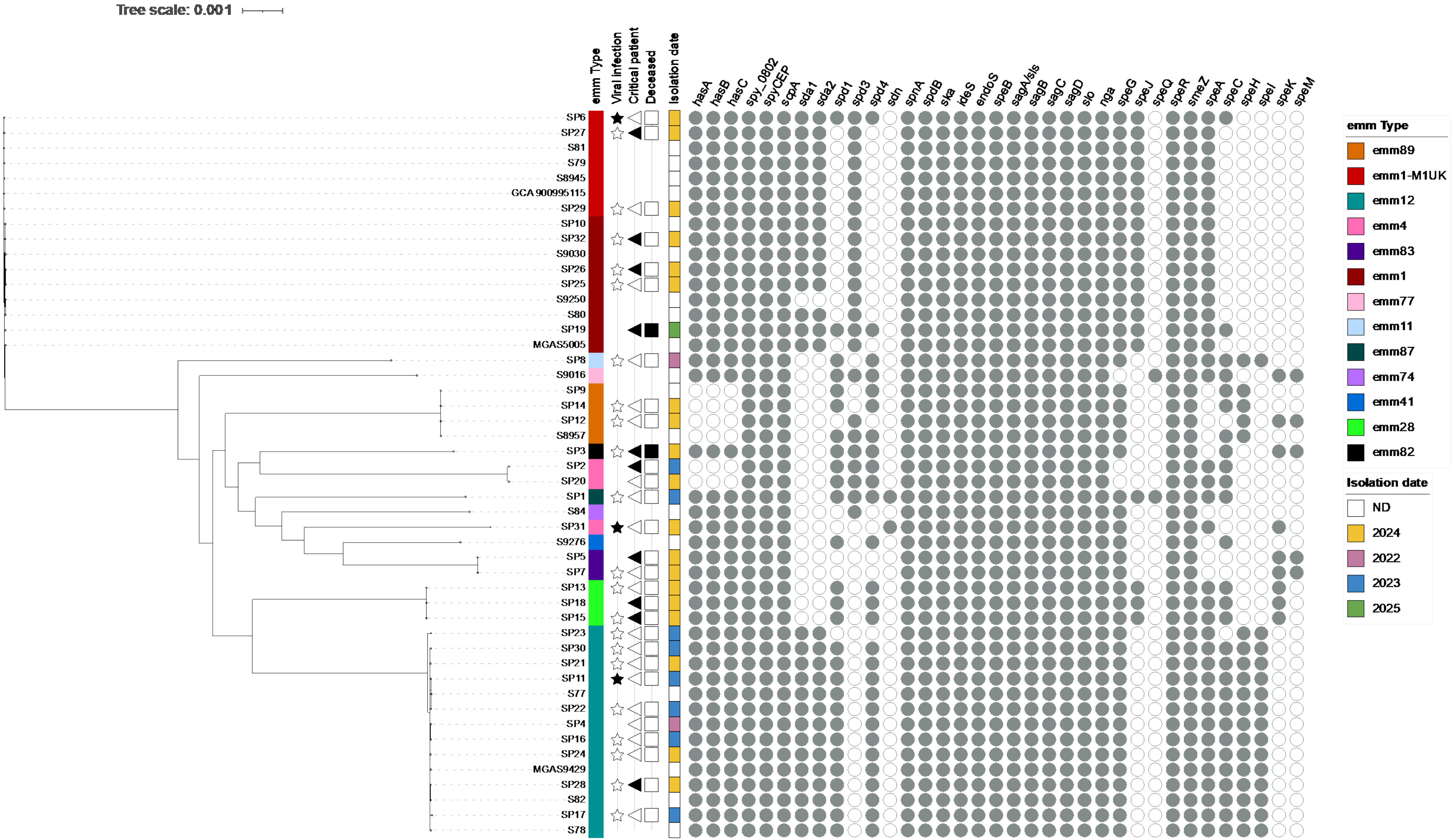
# *Streptococcus pyogenes* genomics in Chile

- Generated 32 sequences from chilean isolates, from one medical center
- Combined with 7 isolates from PUC (thanks to Dr. Aniela W !)



Nicolas Canales



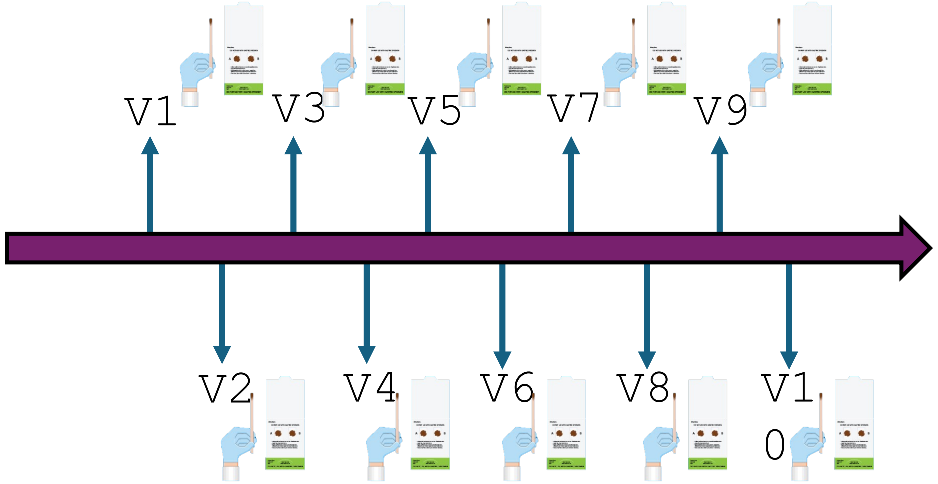

Tree scale: 0.001
emm Type
Viral infection
Critical patient
Deceased
Isolation date
hasA
hasB
hasC
spy_0802
spyCEP
scpA
sda1
sda2
spd1
spd3
spd4
sdn
spnA
spdB
ska
ideS
endoS
speB
sagA/sls
sagB
sagC
sagD
slo
nga
speG
speJ
speQ
speR
smeZ
speA
speC
speH
speI
speK
speM
SP6
SP27
S81
S79
S8945
GCA 900995115
SP29
SP10
SP32
S9030
SP26
SP25
S9250
S80
SP19
MGAS5005
SP8
S9016
SP9
SP14
SP12
S8957
SP3
SP2
SP20
SP1
S84
SP31
S9276
SP5
SP7
SP13
SP18
SP15
SP23
SP30
SP21
SP11
S77
SP22
SP4
SP16
SP24
MGAS9429
SP28
S82
SP17
S78
emm Type
emm89
emm1-M1UK
emm12
emm4
emm83
emm1
emm77
emm11
emm87
emm74
emm41
emm28
emm82
Isolation date
ND
2024
2022
2023
2025

# Metagenomics and antibiotic resistance

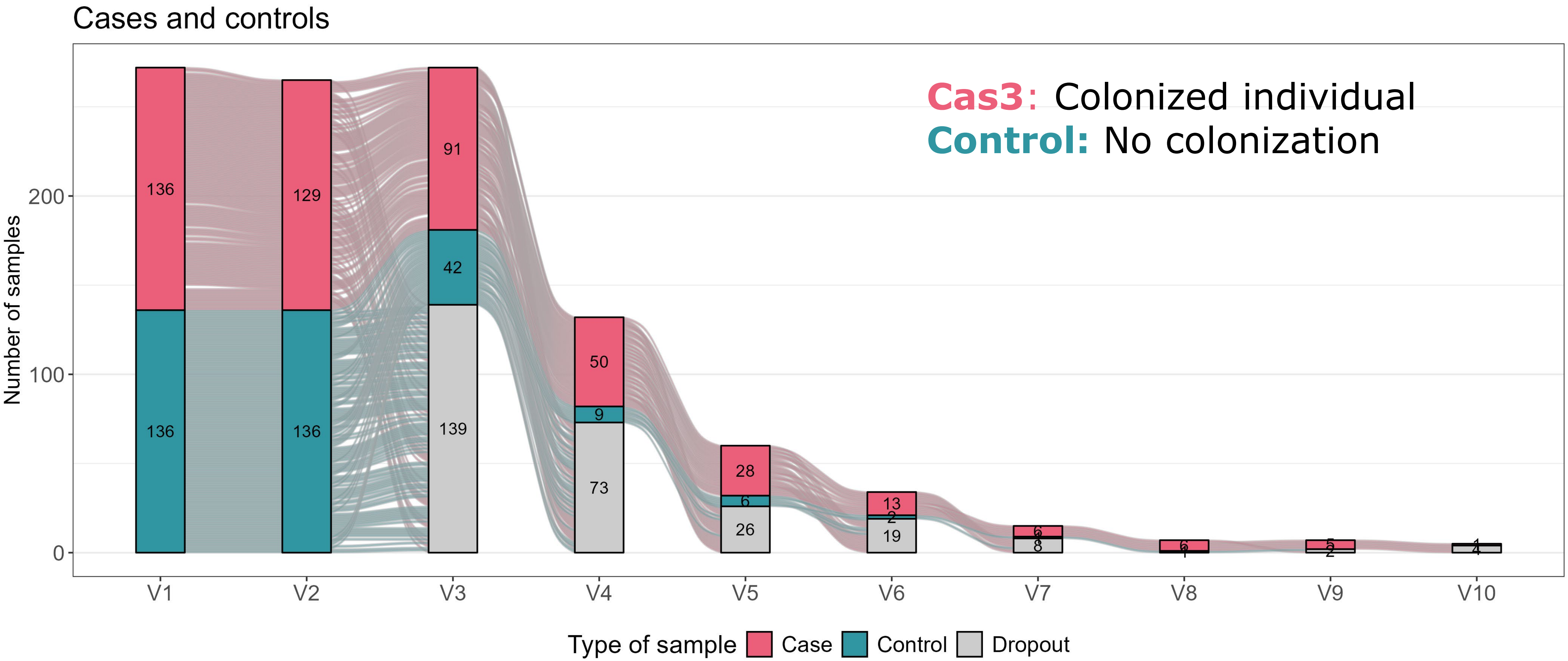
# Gut microbiome and antimicrobial resistant microorganisms

- Colonization study: 798 samples from rectal swabs, from 272 individuals.
- All individuals were hospitalized, and samples were collected every 7 days.
- For each available sample, cultures were performed to evaluate for Gram-negative bacilli resistant to ceftazidime and ciprofloxacine

Rafael

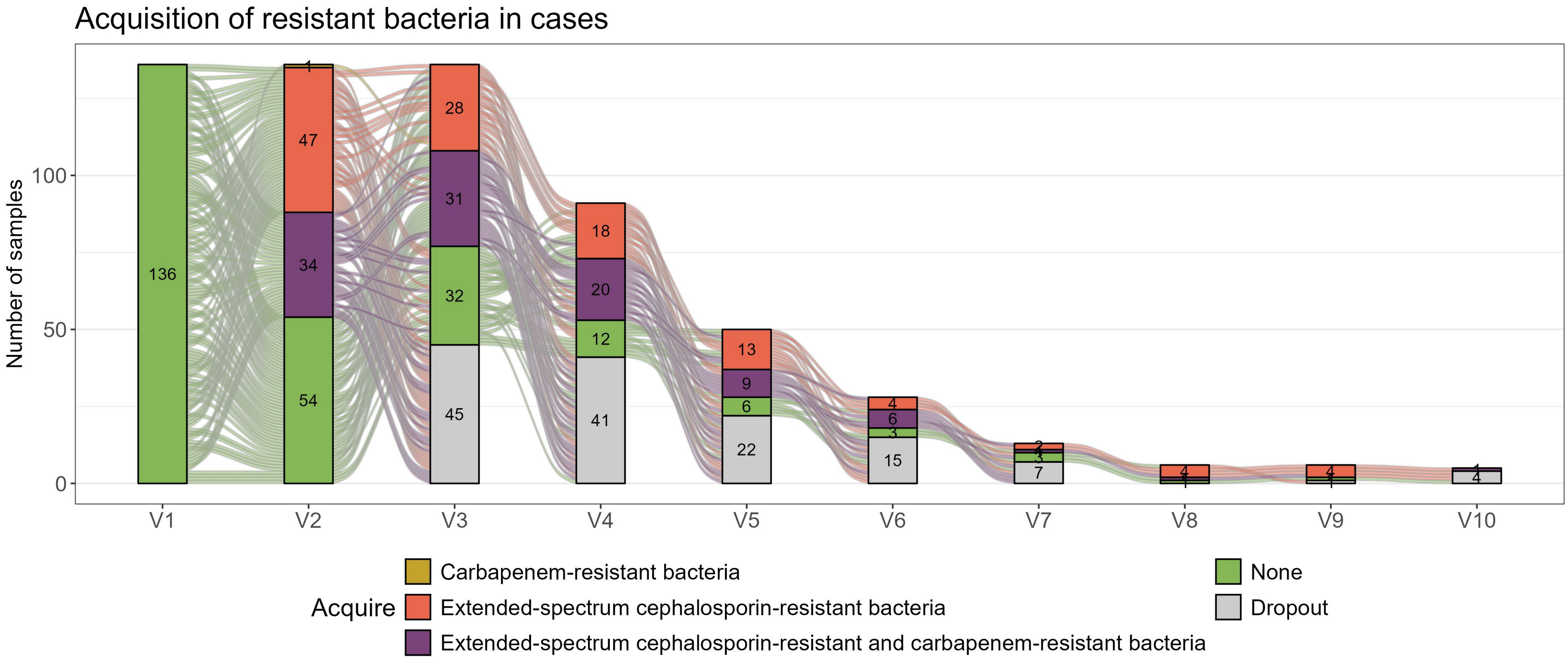


# Samples for most patients until the third visit


Cases and controls

**Cas3:** Colonized individual
**Control:** No colonization

# The majority of the colonization was found during the second and third visit


Acquisition of resistant bacteria in cases

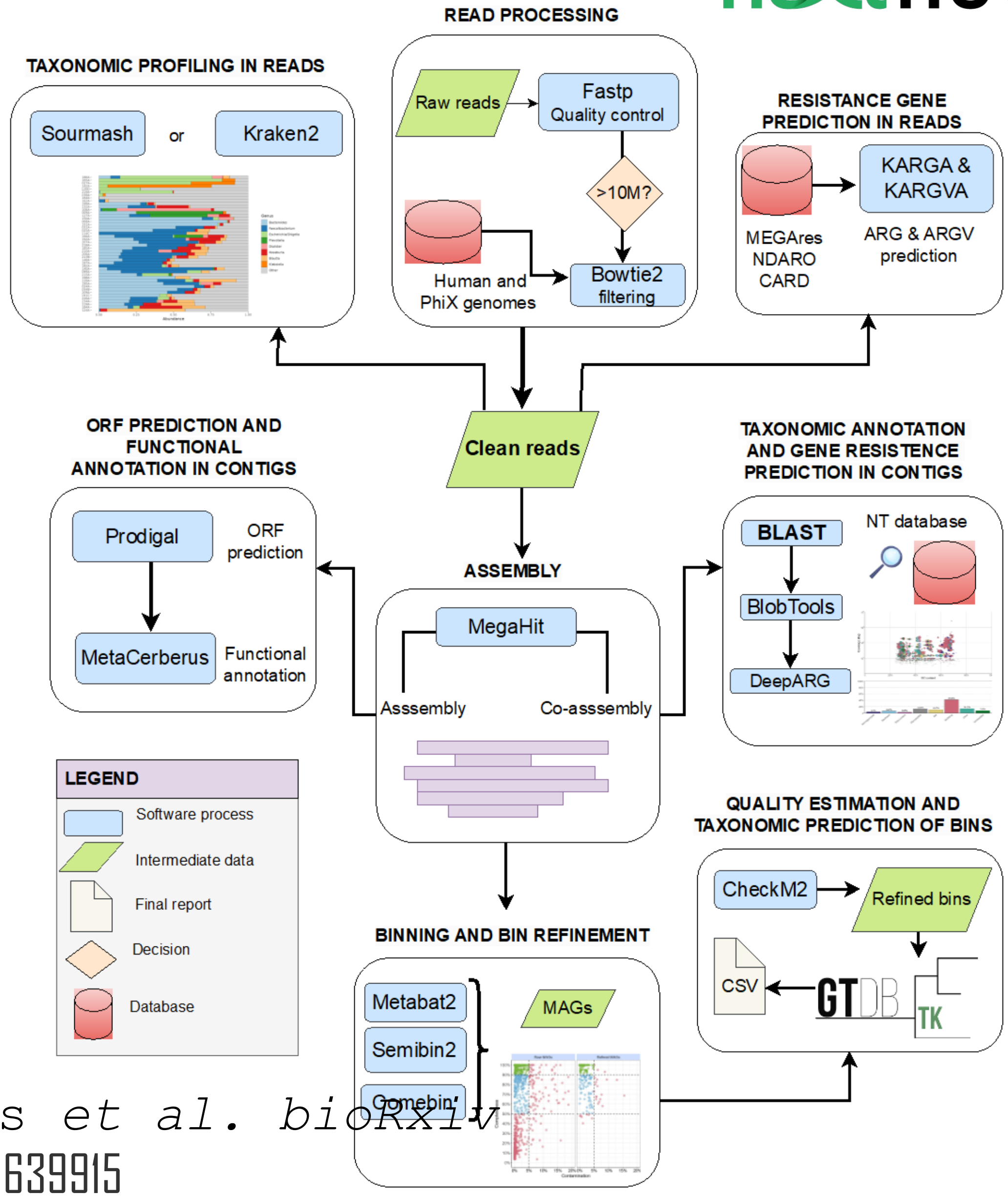
# Metagenomic sequencing

- DNA extraction from 798 cards
- Shotgun metagenomic sequencing. 40 million PE reads (2x150). ~9 Gb per sample.
- Total number of reads: **97,433 millions of reads**
- Total size of the dataset: **14.7 Tb**

# Bioinformatic workflow for metagenomic analysis

nextflow

- Reproducible and portable pipeline for metagenomic analysis (with an emphasis on AMR analysis).
- *https://github.com/gene2dis/bugbuster*



**READ PROCESSING**

Raw reads → Fastp Quality control → >10M? → Bowtie2 filtering

Human and PhiX genomes → Bowtie2 filtering

**TAXONOMIC PROFILING IN READS**

Sourmash or Kraken2

**RESISTANCE GENE PREDICTION IN READS**

MEGAres NDARO CARD → KARGA & KARGVA

ARG & ARGV prediction

Clean reads

**ORF PREDICTION AND FUNCTIONAL ANNOTATION IN CONTIGS**

Prodigal – ORF prediction

MetaCerberus – Functional annotation

**ASSEMBLY**

MegaHit

Asssembly – Co-asssembly

**TAXONOMIC ANNOTATION AND GENE RESISTENCE PREDICTION IN CONTIGS**

BLAST → BlobTools → DeepARG

NT database

**LEGEND**

Software process

Intermediate data

Final report

Decision

Database

**BINNING AND BIN REFINEMENT**

Metabat2

Semibin2

Comebin

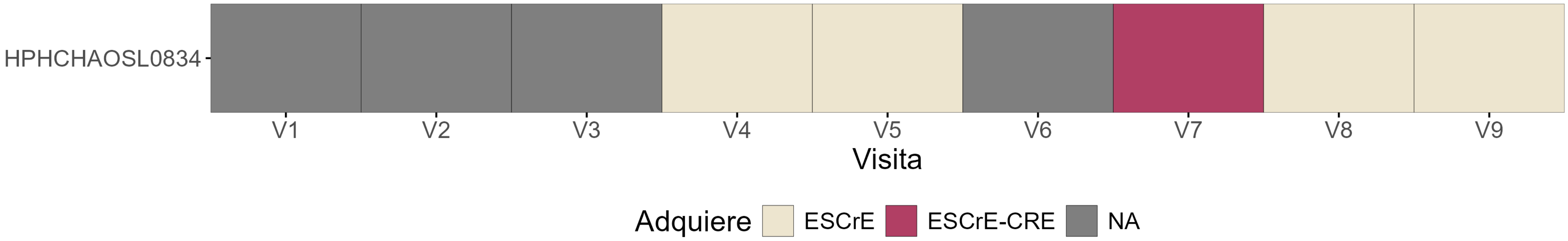
MAGs

**QUALITY ESTIMATION AND TAXONOMIC PREDICTION OF BINS**

CheckM2 → Refined bins → GTDB TK → CSV

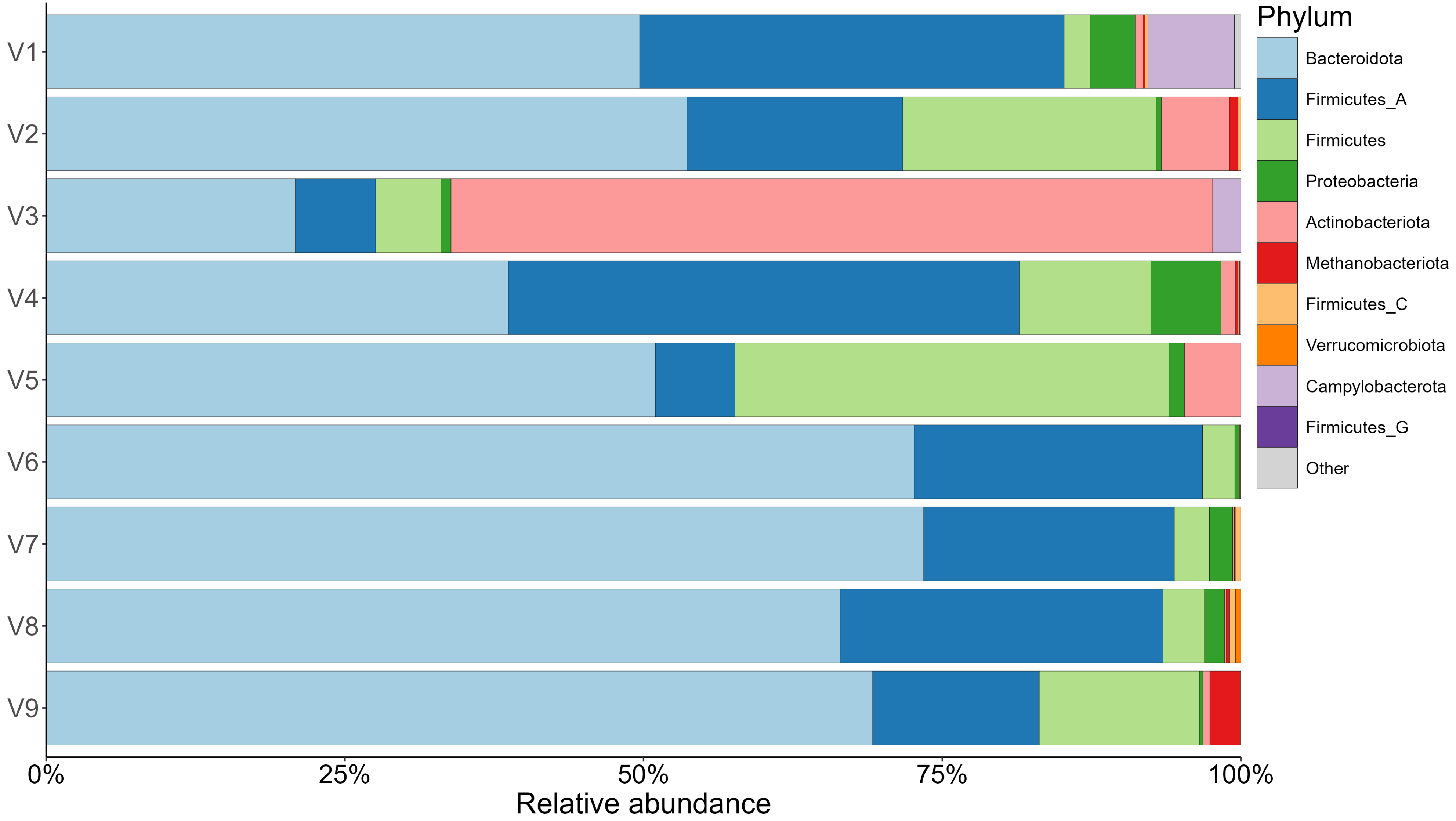
Fuentes *et al. bioRxiv* 2025.02.24.639915

# Example: One subject with multiple visits

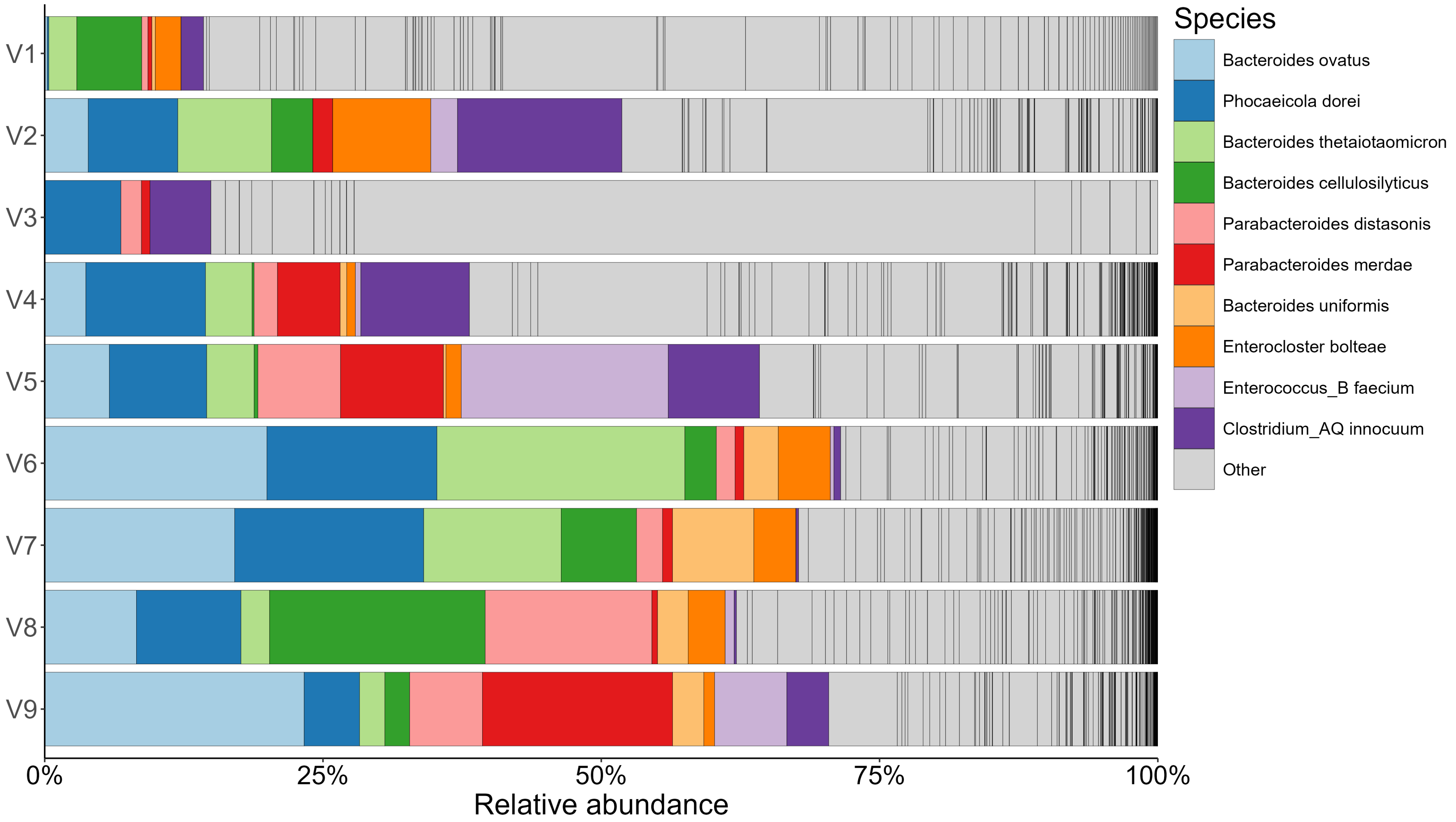


- ESCrE = Extended-spectrum cephalosporin-resistant bacteria
- ESCrE-CRE = Extended-spectrum cephalosporin-resistant and carbapenem-resistant bacteria

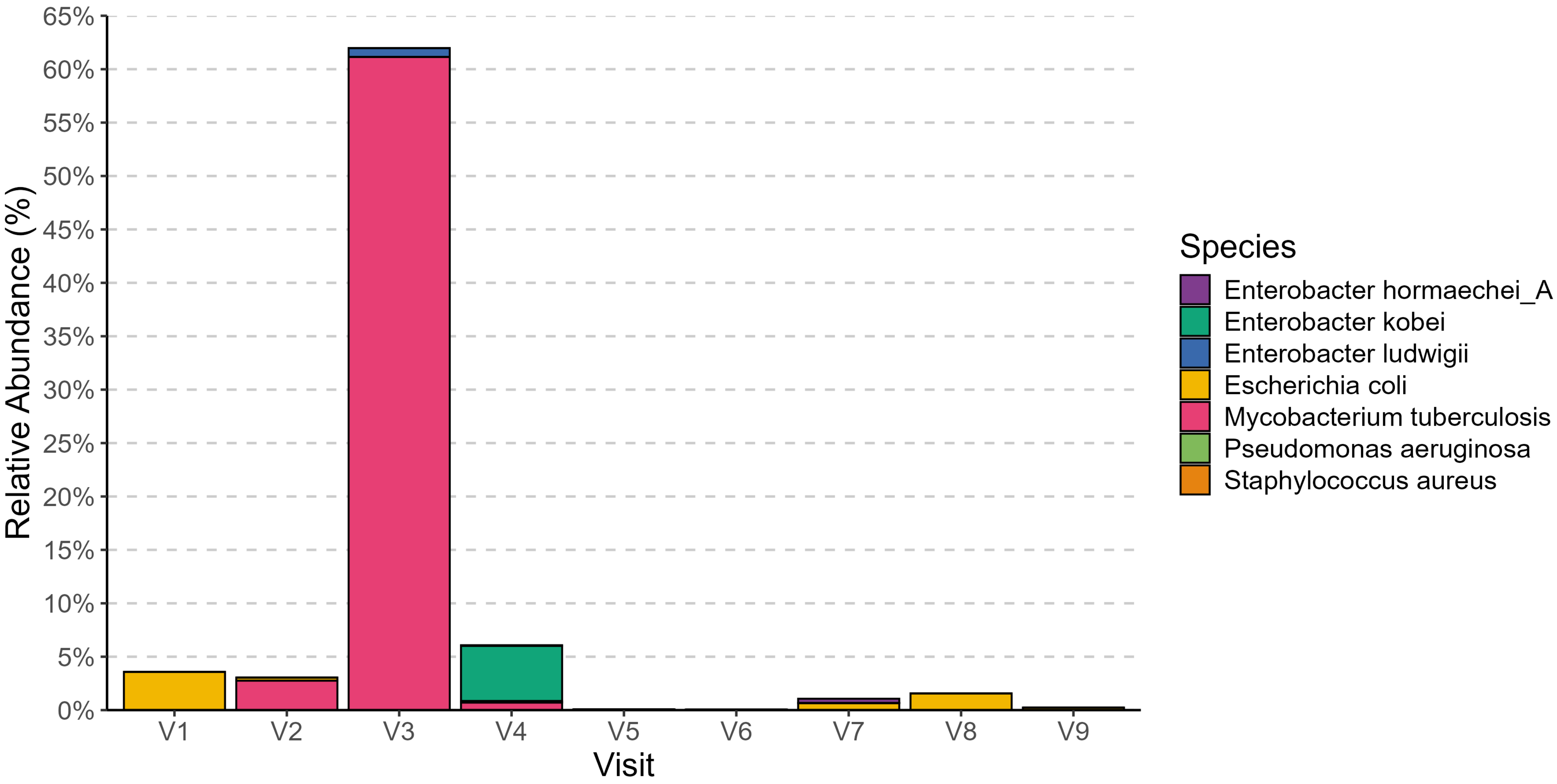
# At the phylum level, the community appears to be more stable from the sixth visit



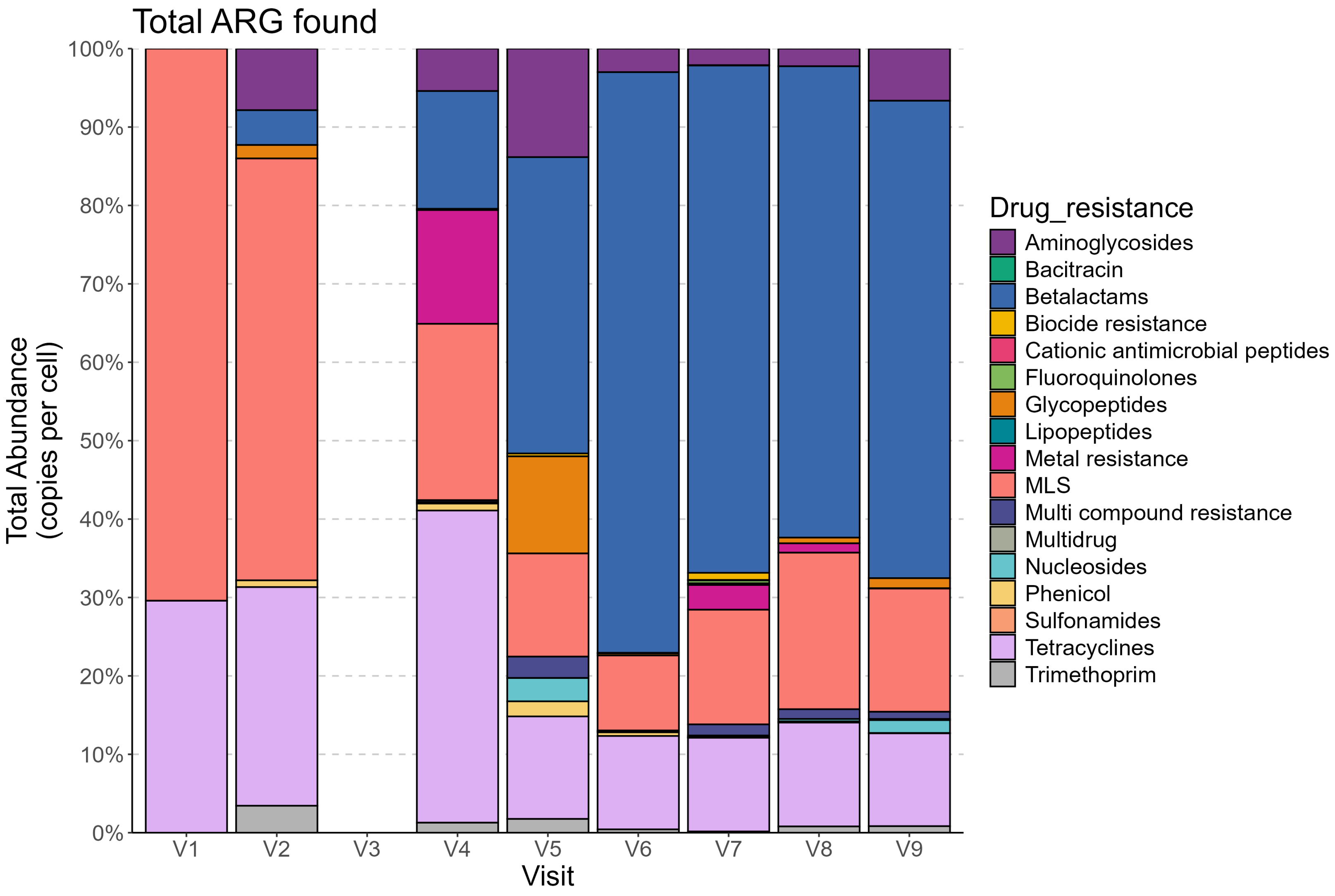
# At the species level, we noticed a particular species dominated on V3



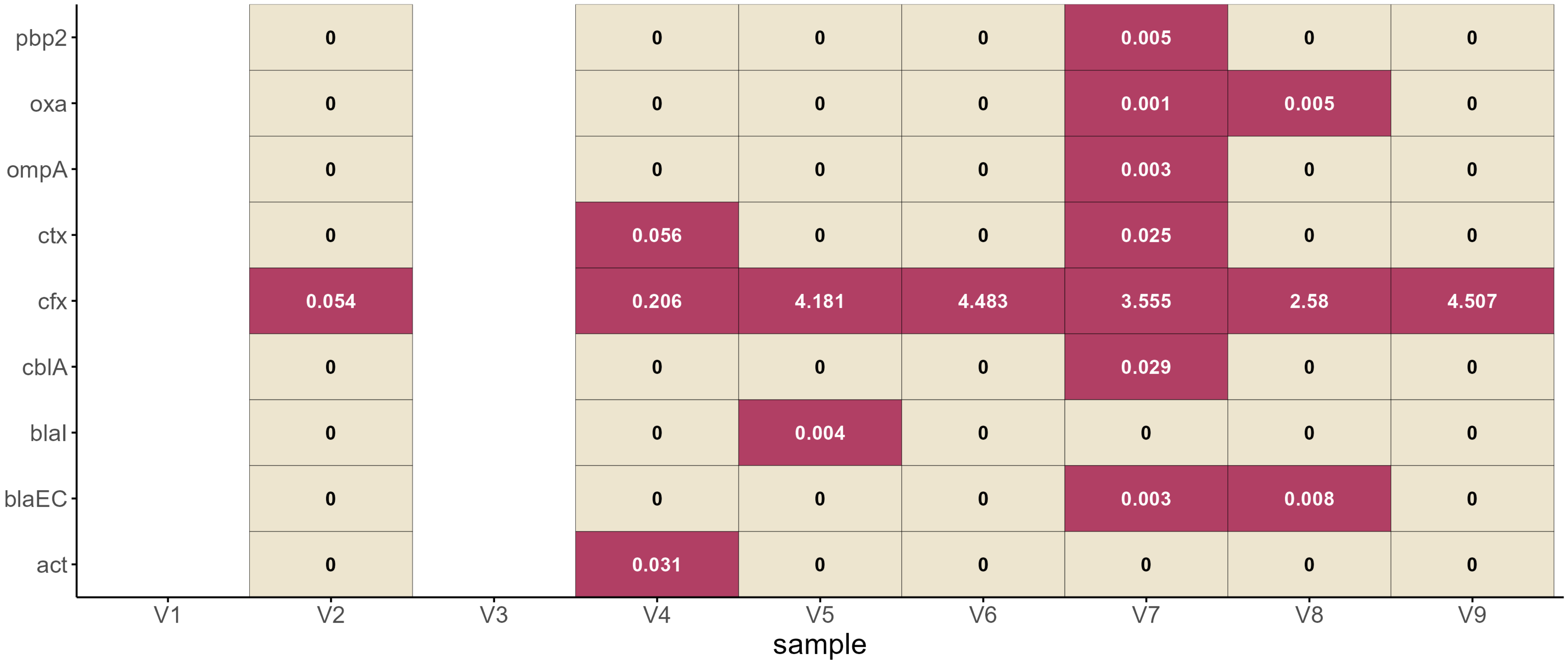
# *Mycobacterium tuberculosis* was the most abundant species in V3



# Relative abundance and diversity of antimicrobial resitance categories changes across visits

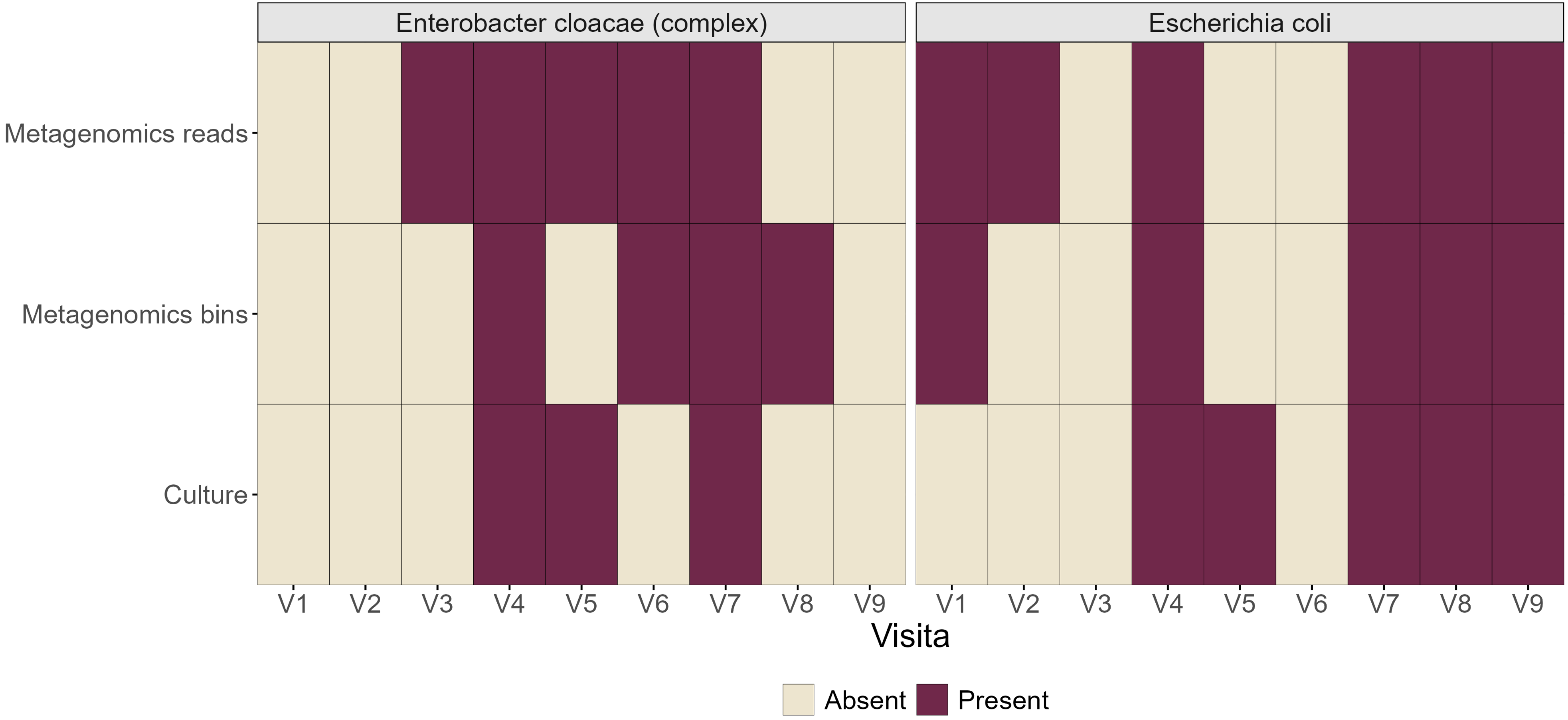


# Beta-lactamase are more abundant from visit 5

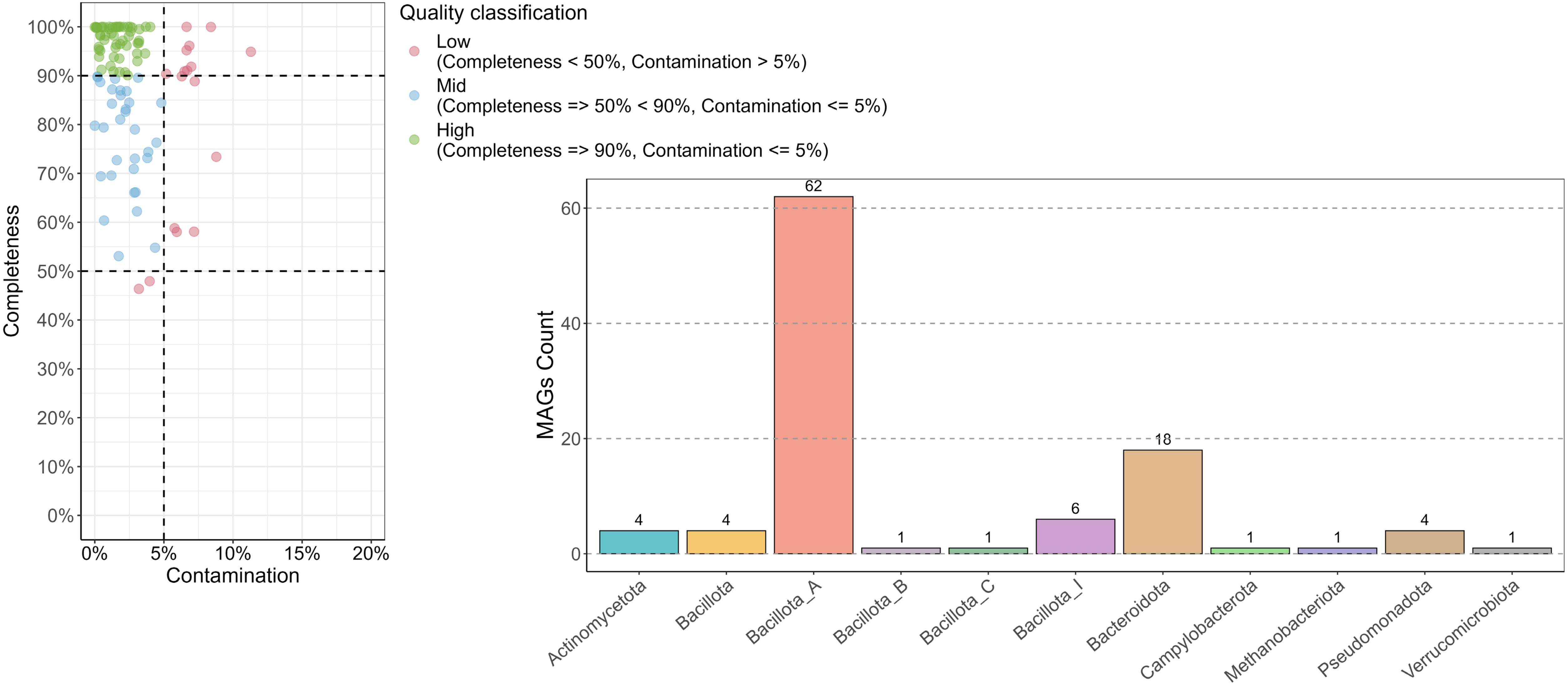


| | V1 | V2 | V3 | V4 | V5 | V6 | V7 | V8 | V9 |
|---|---|---|---|---|---|---|---|---|---|
| pbp2 | | 0 | | 0 | 0 | 0 | 0.005 | 0 | 0 |
| oxa | | 0 | | 0 | 0 | 0 | 0.001 | 0.005 | 0 |
| ompA | | 0 | | 0 | 0 | 0 | 0.003 | 0 | 0 |
| ctx | | 0 | | 0.056 | 0 | 0 | 0.025 | 0 | 0 |
| cfx | | 0.054 | | 0.206 | 4.181 | 4.483 | 3.555 | 2.58 | 4.507 |
| cblA | | 0 | | 0 | 0 | 0 | 0.029 | 0 | 0 |
| blaI | | 0 | | 0 | 0.004 | 0 | 0 | 0 | 0 |
| blaEC | | 0 | | 0 | 0 | 0 | 0.003 | 0.008 | 0 |
| act | | 0 | | 0.031 | 0 | 0 | 0 | 0 | 0 |

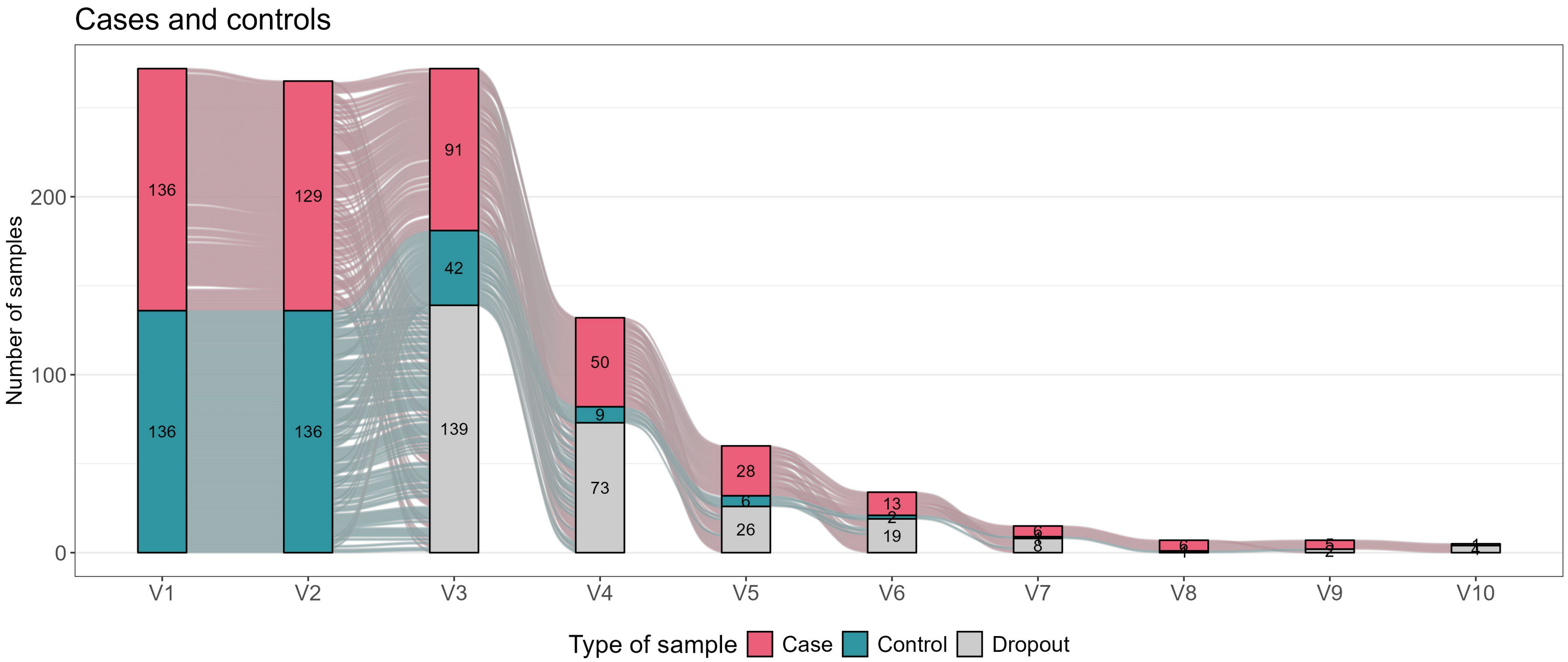
# Using metagenomics we detect the same species that were found in cultures



# From a single individual, across multiple visits, we were able to generate 103 metagenome-assembled genomes (MAGs)



# Next steps



- Full characterization of the dataset
- Changes in species populations over time
- Reconstruction of mobile elements
- Comparison of MAGs with isolate

# Acknowledgments

Microbial
Data Science

ANID
Agencia
Nacional de
Investigación y
Desarrollo

GENE 2 DIS
GENOMIC EPIDEMIOLOGY
IN EMERGENT DISEASES

Agencia
Nacional de
Investigación
y Desarrollo
Ministerio de Ciencia,
Tecnología, Conocimiento
e Innovación
Gobierno de Chile

**Fondecyt**
Fondo Nacional de Desarrollo
Científico y Tecnológico

CDC
CENTERS FOR DISEASE
CONTROL AND PREVENTION