

Building an Updated Replicon Database for PlasmidFinder

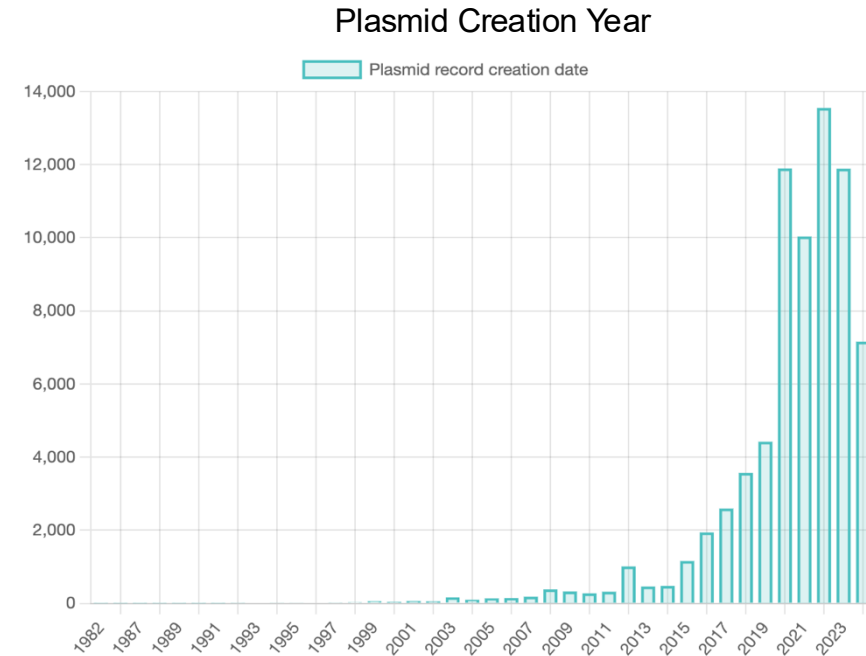
Alessandro Caula

DTU, Research Group for Genomic Epidemiology

Why Extend PlasmidFinder's Database?

Current: PlasmidFinder's database holds **488** replicon sequences, from Lozano 2012 and Carattoli 2014 .

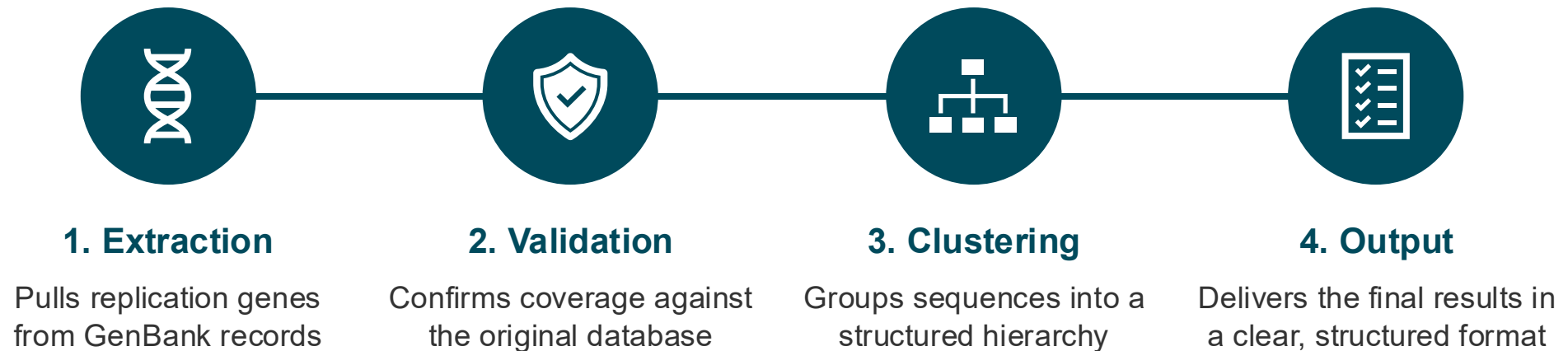
Gap: It hasn't grown since 2014, missing a decade of new plasmid sequencing.



**Extend replicon database and improve PlasmidFinder tool,
without losing anything it already does.**

Pipeline overview

From raw GenBank records to a validated, organized replicon database.



Extraction & Filtering

About 381,394 records

GenBank plasmid records
collected as the starting pool

Download

Extraction

125,791 Replicon candidates

Gene features matching rep
gene keywords, plus exclusion
tokens and a length filter

32,799 sequences, 92,992 removed

Exact duplicate sequences removed

Deduplication

Containment

24,254 sequences, 8,545 removed

Sequences contained within a longer
one removed

Clustering

We need to separate near duplicates of existing PlasmidFinder replicons from genuinely new diversity.



Cluster PF (488) together with the novel sequences: **24,742 total sequences**



We tested both **DBSCAN** and **CD-HIT**



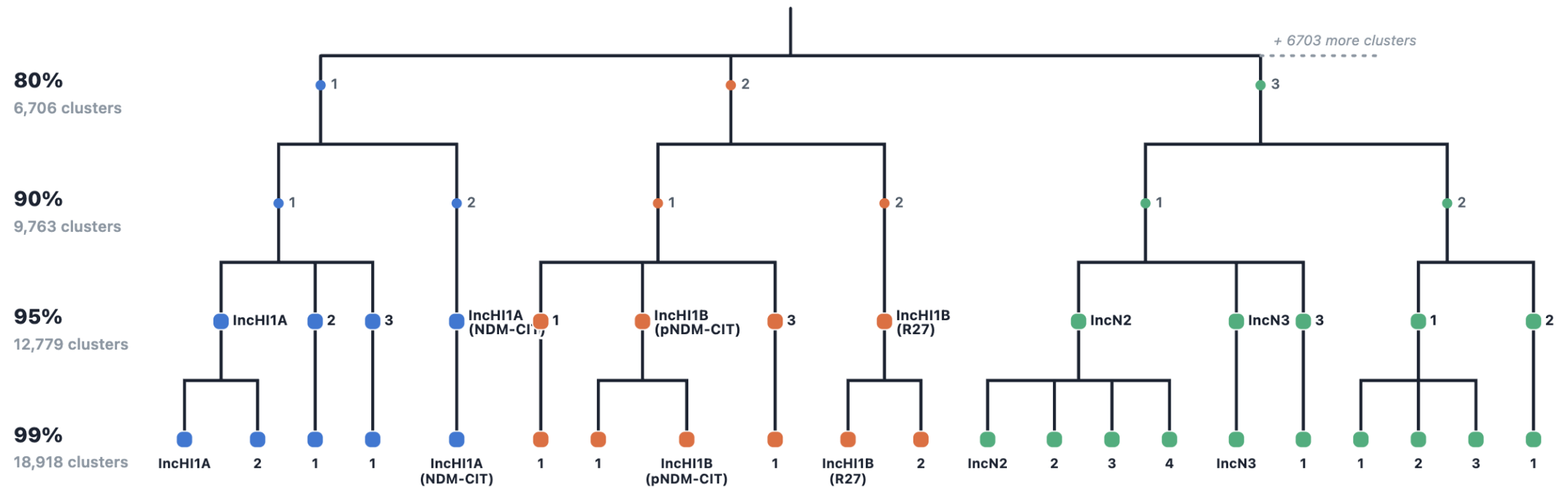
DBSCAN showed a *bridging problem*, chains of similar sequences linked unrelated replicon types into one big cluster



CD-HIT was chosen. It grouped PlasmidFinder replicons of the same type together far more consistently, while correctly keeping different types apart

Building a Hierarchy with CD-HIT

Group replicons by similarity at four thresholds, from broad clusters at 80% identity down to fine sub-variants at 99%, like a **taxonomic lineage**.



PlasmidFinder Output

PlasmidFinder will still let users choose which database to use, the original curated one or this new, computationally derived one.

Enterobacteriales Hide

Plasmid factor	↕	Identity	↕	Query / Template length	↕	Contig	↕	Position in contig	↕	Accession no.	↕
> IncFII(pECLA)		100.00 %		747 / 747		CP047730.1 product=RepB_family_plasmid_replication_initiator_protein		3..749		CP001919	
> IncHI2A		100.00 %		630 / 630		CP047730.1 product=RepB_family_plasmid_replication_initiator_protein		1087..1716		BX664015	

New Format: for a query hit, report which cluster it falls into at each resolution.

Enterobacteriales Hide

Plasmid factor	↕	Identity	↕	Query / Template length	↕	Contig	↕	Position in contig	↕	Accession no.	↕	h80	↕	h90	↕	h95	↕	h99	↕
> IncFII(pECLA)		100.00 %		747 / 747		CP047730.1 product=RepB_family_plasmid_replication_initiator_protein		3..749		CP001919		105		1		IncFII(pECLA)		IncFII(pECLA)	
> IncHI2A		100.00 %		630 / 630		CP047730.1 product=RepB_family_plasmid_replication_initiator_protein		1087..1716		BX664015		120		1		IncHI2A		IncHI2A	

Next Steps

1

Integrate the taxonomic data for the new replicons

2

Format the database with a defined standardized structure

3

Adapt the PlasmidFinder tool for the new database format

Thank you!

Alessandro Caula
DTU, Research Group for Genomic Epidemiology