Bacterial Strain Taxonomy for Genomic Surveillance

chewBBACA and the creation of cgMLST schemas

Mário Ramirez and Rafael Mamede
Faculty of Medicine of the University of Lisbon

23th October 2025

What is chewBBACA?

Intended Learning Objectives

Specific objectives of this session:

1. Learn about key concepts used in wg/cgMLST and their implementation in chewBBACA.
2. Learn about the functionalities provided by chewBBACA.
3. Learn how to apply chewBBACA to:
    - Create schemas for any bacterial species.
    - Determine the allelic profiles of strains of interest.
    - Evaluate schemas and allele calling results to generate interactive reports that enable a comprehensive analysis of the results.

Outline

This session consists of the following elements

1. Brief explanation of wg/cgMLST.
2. Overview of the functionalities provided by chewBBACA for wg/cgMLST.
3. Using chewBBACA to create a schema for a species of interest.
4. Using chewBBACA to classify and differentiate bacterial lineages.

Whole- and core-genome MLST (wg/cgMLST)

- Expands the MLST concept to hundreds or thousands of loci for increased resolution.

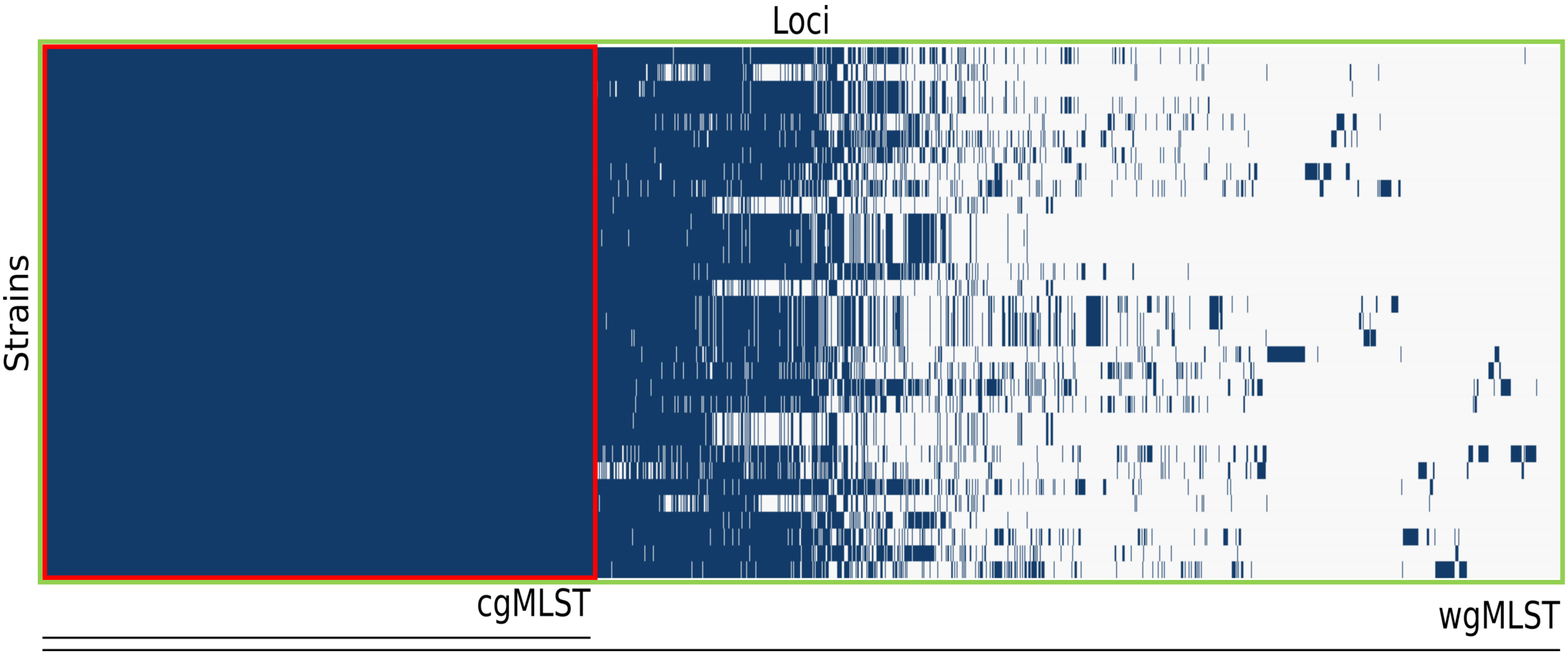
MLST

	l_1	l_2	l_3	l_4	l_5	l_6	l_7
s_1	1	3	1	1	2	2	1
s_2	1	3	1	7	2	2	1
s_3	2	3	1	9	3	2	1
s_4	1	2	1	4	3	2	1
s_5	2	2	1	4	4	2	7
s_6	2	1	2	4	4	2	7

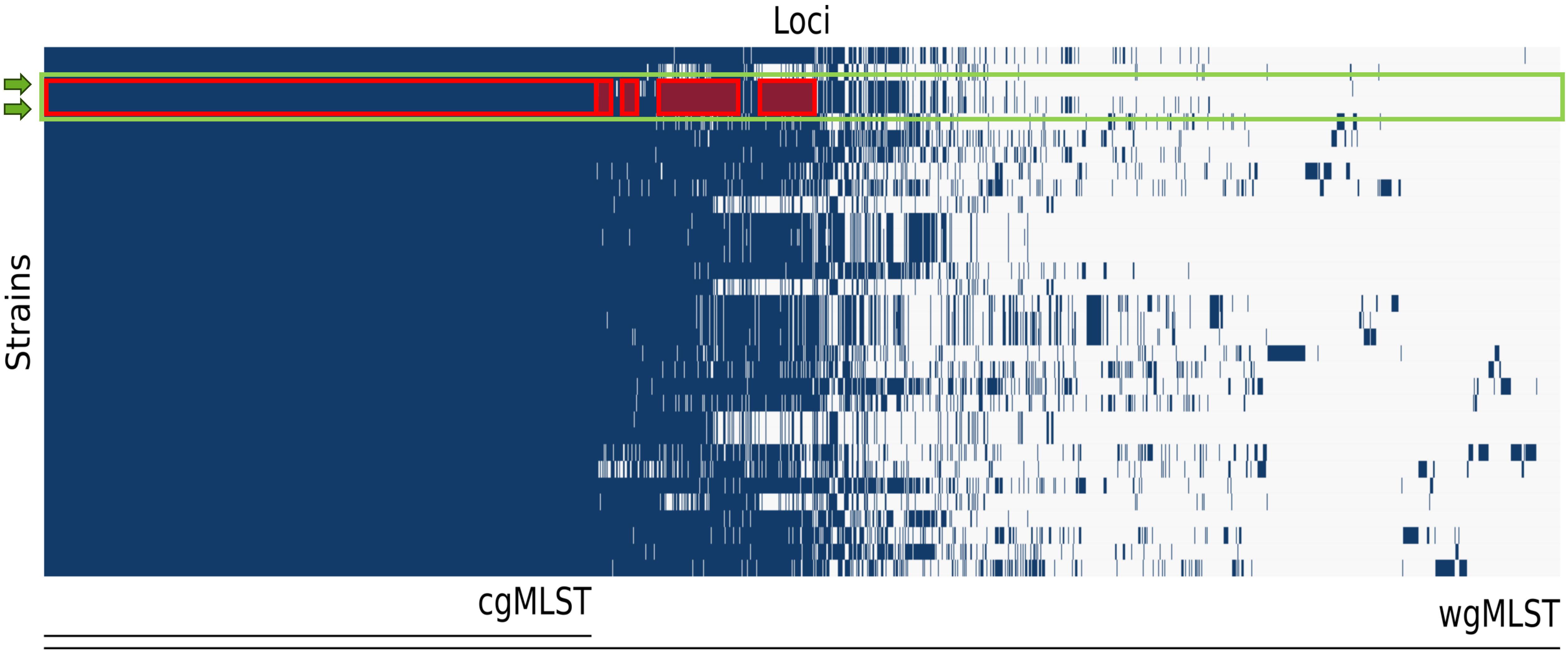
wg/cgMLST

	l_1	l_2	l_3	l_4	l_5	l_6	l_7		l_n
s_1	1	3	1	1	2	2	1	...	1
s_2	1	3	1	7	2	2	1	...	1
s_3	2	3	1	9	3	2	1	...	5
s_4	1	2	1	4	3	2	1	...	5
s_5	2	2	1	4	4	2	7	...	3
s_6	2	1	2	4	4	2	7	...	2

Whole- and core-genome MLST (wg/cgMLST)

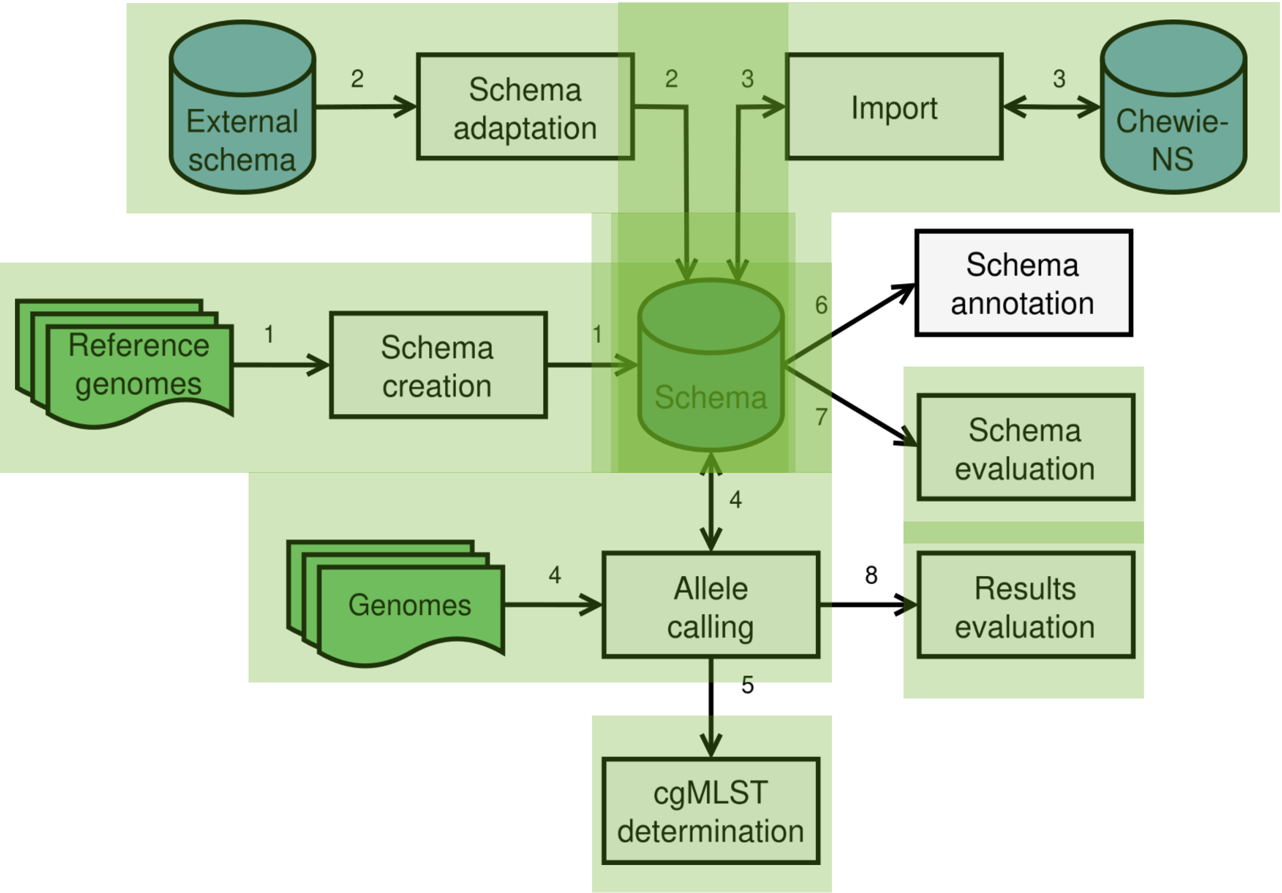


wgMLST as a stepping stone for dynamic cgMLST



chewBBACA - Overview

- chewBBACA is a software suite for the creation and evaluation of wg/cgMLST schemas and results.



1. Schema creation from a set of reference genomes.
2. Adapting schemas from external platforms (e.g. BIGSdb, EnteroBase) for usage with chewBBACA.
3. Importing ready-to-use schemas from Chewie-NS.
4. Determine the allelic profiles of strains of interest.
5. Determine the set of core loci to define a cgMLST schema.
6. Create interactive reports to view and analyse allele calling results.
7. Create interactive reports to explore the diversity of the loci in a schema.

chewBBACA – standalone classification of ORFs

Independent of network access

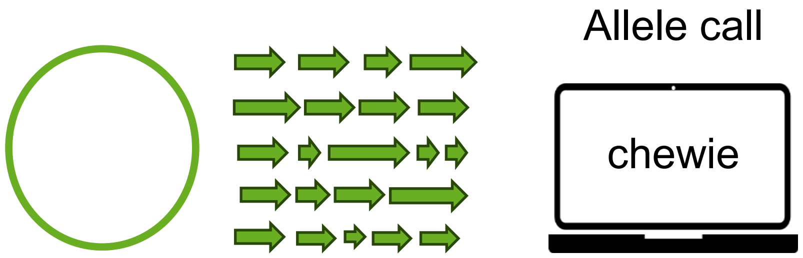
chewie

No data is requested or sent to a central server when using chewBBACA

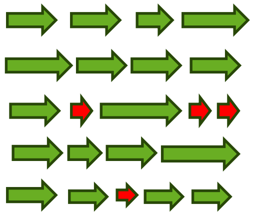
Data requests and submissions only under user direction and in specific circumstances

Can run on a laptop

Pyrodigal



chewBBACA’s loci and alleles are based on Pyrodigal’s identification of ORFs



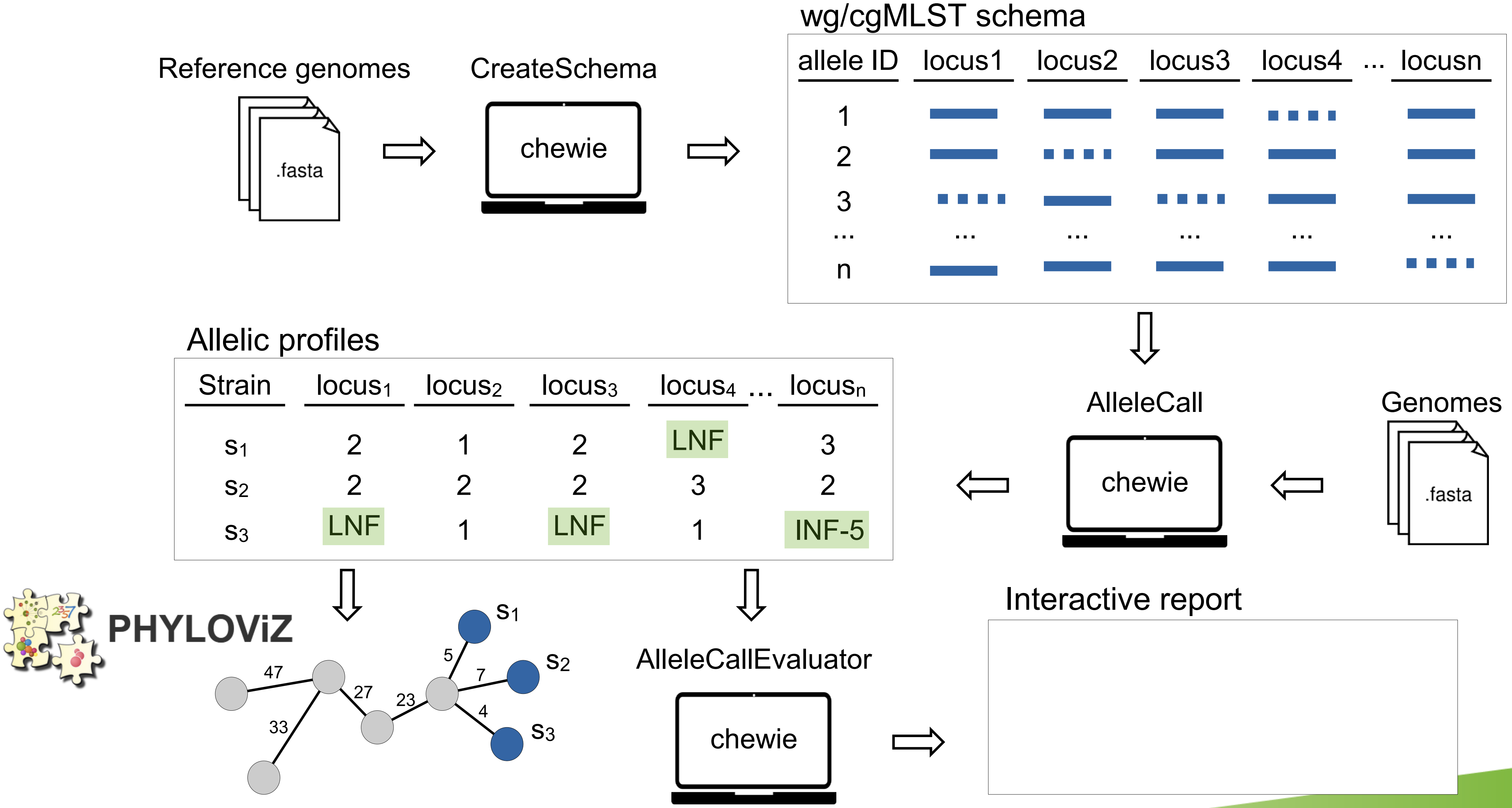
Minimal size of the ORF to be considered by chewBBACA (201nt)

Blast Score Ratio (BSR) to be considered a valid locus allele (0.6)

Maximum deviation from mode to be considered a valid locus allele (20%)

Rasko et al. 2005. BMC Bioinformatics 6:2; Larralde. 2022. Journal of Open Source Software 7:4296.

chewBBACA – workflow example



Reference genomes

.fasta

CreateSchema

chewie

wg/cgMLST schema

allele ID	locus1	locus2	locus3	locus4	...	locusn
1	▬▬▬	▬▬▬	▬▬▬	▪ ▪ ▪ ▪		▬▬▬
2	▬▬▬	▪ ▪ ▪ ▪	▬▬▬	▬▬▬		▬▬▬
3	▪ ▪ ▪ ▪	▬▬▬	▪ ▪ ▪ ▪	▬▬▬		▬▬▬
...	...	...	...	...		...
n	▬▬▬	▬▬▬	▬▬▬	▬▬▬		▪ ▪ ▪ ▪

AlleleCall

chewie

Genomes

.fasta

Allelic profiles

Strain	locus$_1$	locus$_2$	locus$_3$	locus$_4$	...	locus$_n$
s$_1$	2	1	2	LNF		3
s$_2$	2	2	2	3		2
s$_3$	LNF	1	LNF	1		INF-5

PHYLOViZ

AlleleCallEvaluator

chewie

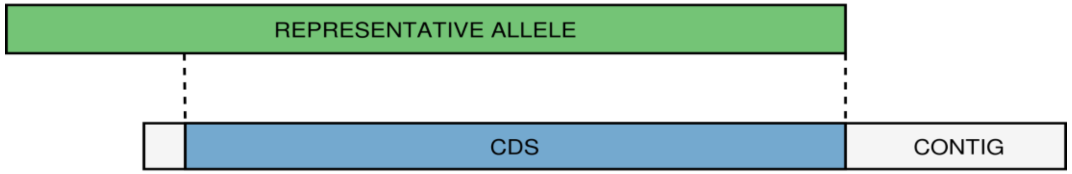
Interactive report

chewBBACA – loci classifications

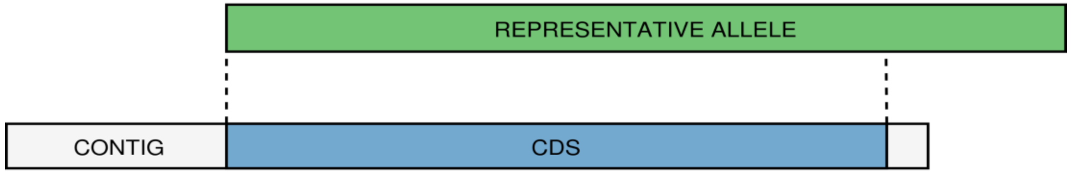
Allelic profiles

Strain	locus$_1$	locus$_2$	locus$_3$	locus$_4$ …	locus$_n$
s$_1$	2	1	2	LNF	3
s$_2$	2	2	2	3	2
s$_3$	LNF	1	LNF	1	INF-5

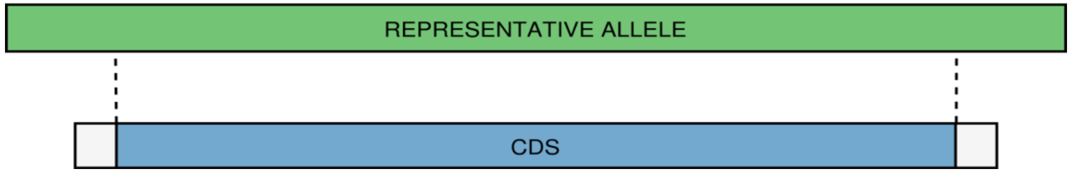
PLOT5



PLOT3



LOTSC

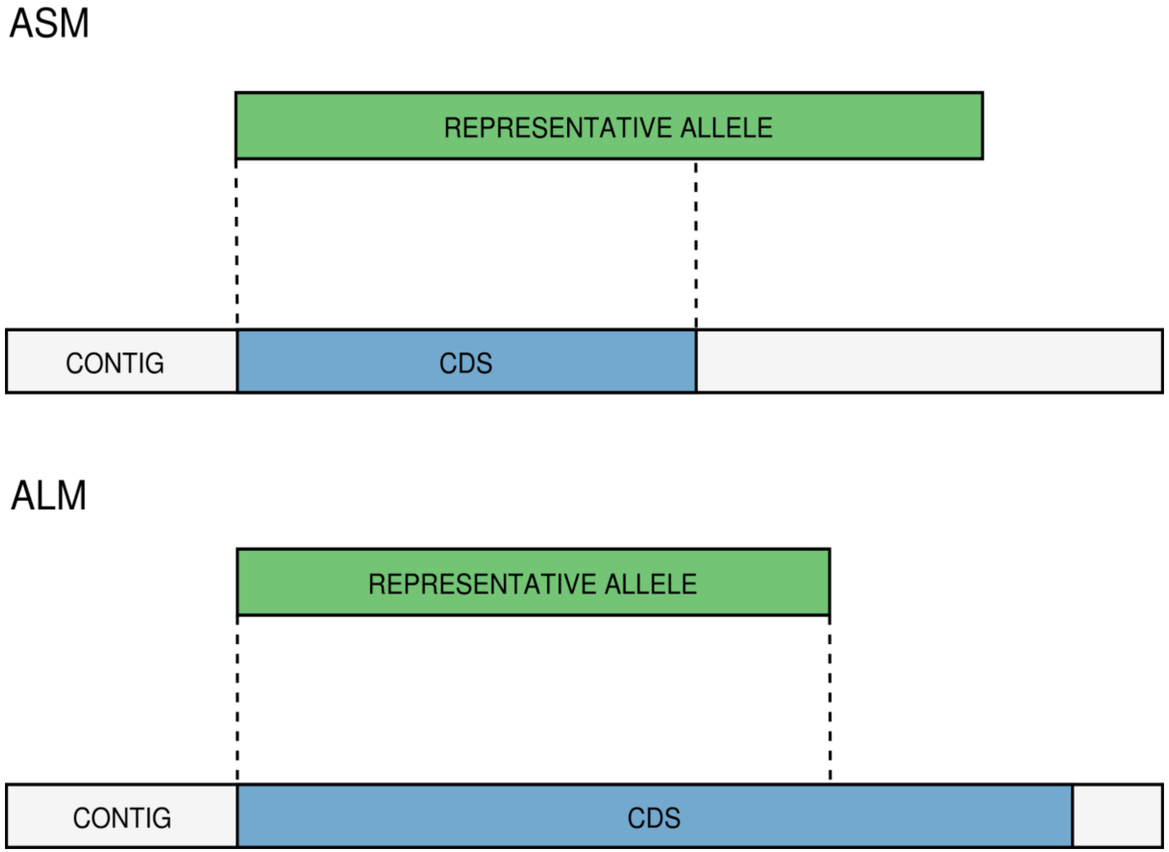


- chewBBACA assigns integer allele identifiers when it finds loci alleles that are in the schema.
- chewBBACA infers novel alleles (INF-) and adds them to schemas.
- The Locus Not Found (LNF) class is assigned when a locus is not found in a genome.
- Other classifications are assigned when a locus may have an uncertain allele

PLOT – Possible Locus on Tip of contig

LOTSC - Locus on Tip of a small contig

chewBBACA – loci classifications

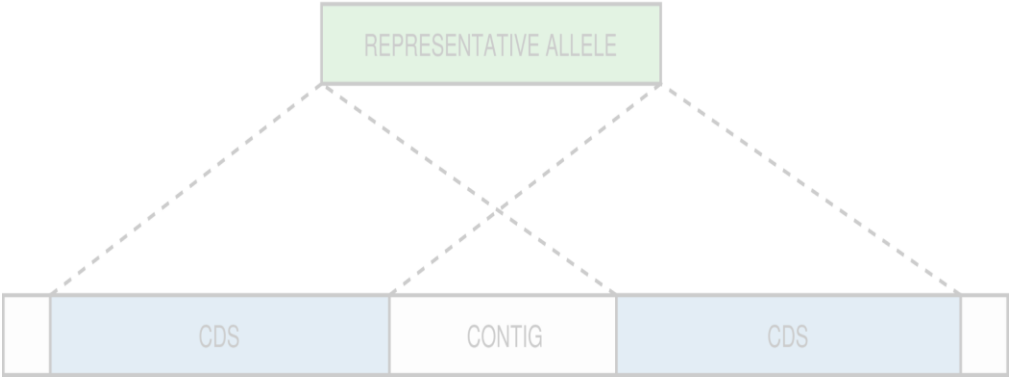


ASM – Allele Smaller than Mode

ALM – Allele Larger than Mode

chewBBACA – loci classifications

NIPH/NIPHEM

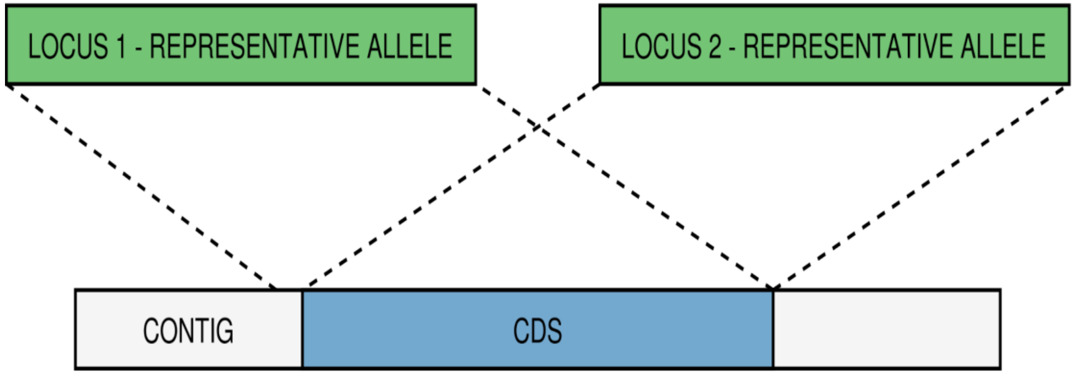


NIPH – Non-Informative Paralogous Hit

NIPHEM – Non-Informative Paralogous Hit Exact Match

Potential paralogues or orthologues in the **query genome**

PAMA



PAMA - Paralogous Match

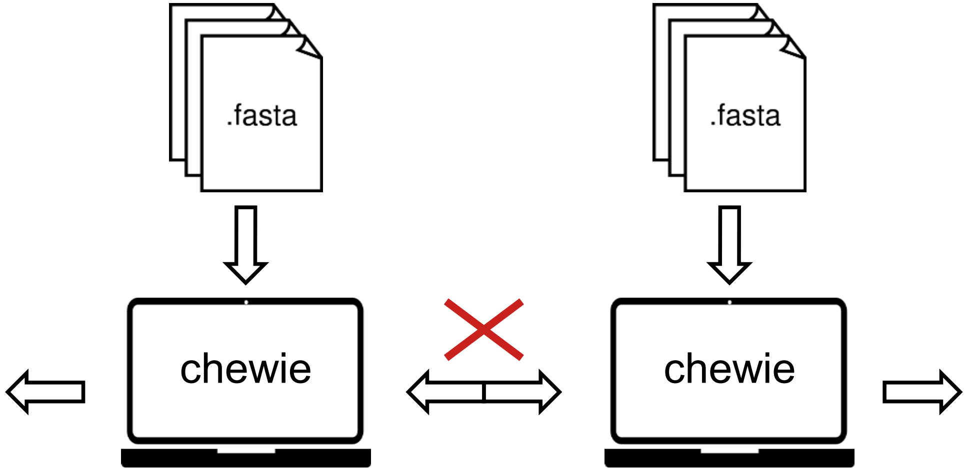
Loci in the **schema** that are potential paralogues or orthologues

Sharing wg/cgMLST schemas and results

- Well-known platforms that implement wg/cgMLST (e.g. BIGSdb, EnteroBase) centralize data analysis to guarantee the consistency of the results.
- chewBBACA allows users to perform their analyses locally, offering greater scalability and complying with stricter data privacy policies.
  - However, results generated by different users may not be directly comparable.

Allelic profiles

Strain	locus$_1$	locus$_2$	locus$_3$
s$_1$	2	1	INF-3
s$_2$	2	2	2
s$_3$	LNF	INF-2	1

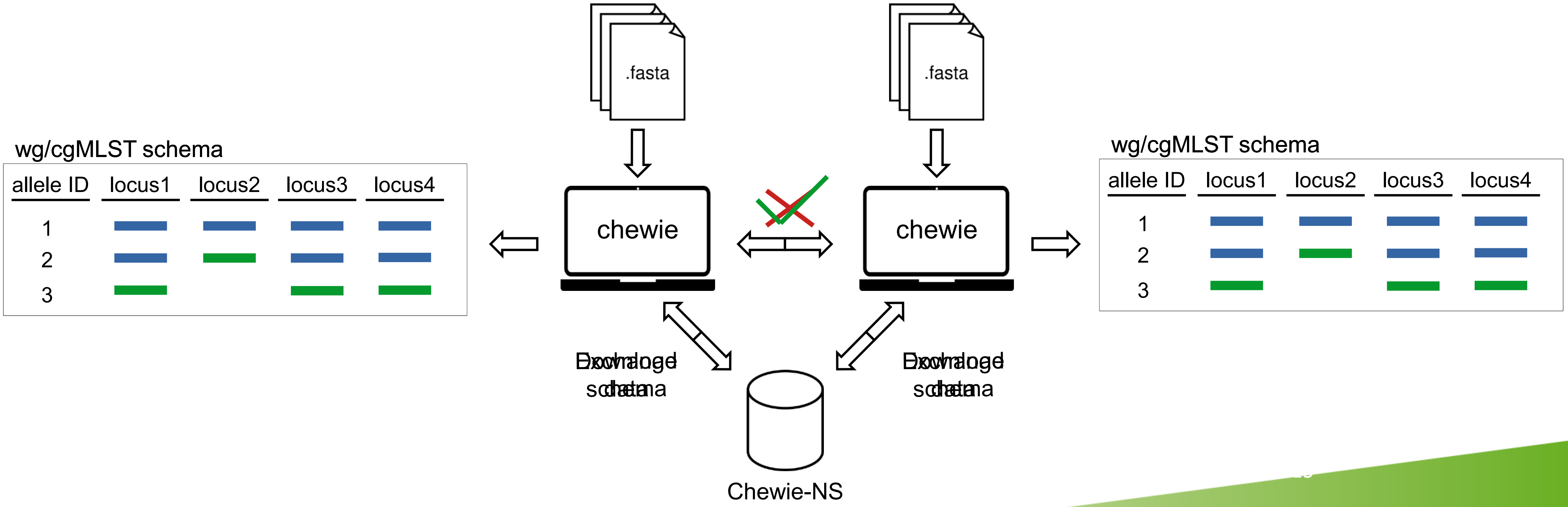


Allelic profiles

Strain	locus$_1$	locus$_2$	locus$_3$
s$_1$	2	1	INF-1
s$_2$	2	2	2
s$_3$	LNF	INF-1	1

Sharing wg/cgMLST data - Chewie-NS

- The Chewie-NS web server was created to offer a simple solution to store and distribute wg/cgMLST schemas.
- Additionally, it integrates with chewBBACA, allowing users to synchronize their schemas to ensure the consistency and comparability of the results.



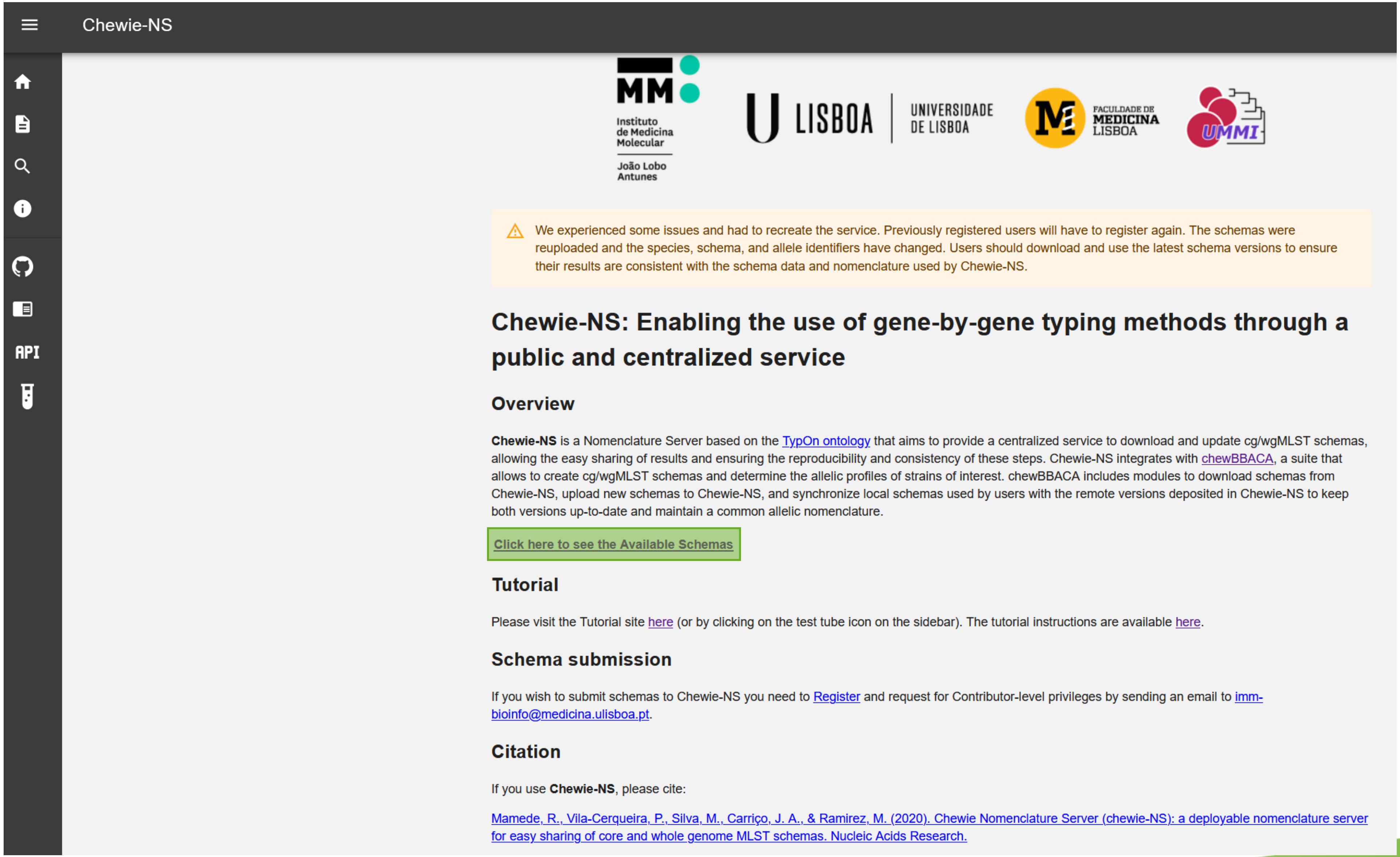
wg/cgMLST schema

allele ID	locus1	locus2	locus3	locus4
1	🟦	🟦	🟦	🟦
2	🟦	🟩	🟦	🟦
3	🟩		🟩	🟩

wg/cgMLST schema

allele ID	locus1	locus2	locus3	locus4
1	🟦	🟦	🟦	🟦
2	🟦	🟩	🟦	🟦
3	🟩		🟩	🟩

Chewie-NS: https://chewbbaca.online/



⚠ We experienced some issues and had to recreate the service. Previously registered users will have to register again. The schemas were reuploaded and the species, schema, and allele identifiers have changed. Users should download and use the latest schema versions to ensure their results are consistent with the schema data and nomenclature used by Chewie-NS.

Chewie-NS: Enabling the use of gene-by-gene typing methods through a public and centralized service

Overview

Chewie-NS is a Nomenclature Server based on the TypOn ontology that aims to provide a centralized service to download and update cg/wgMLST schemas, allowing the easy sharing of results and ensuring the reproducibility and consistency of these steps. Chewie-NS integrates with chewBBACA, a suite that allows to create cg/wgMLST schemas and determine the allelic profiles of strains of interest. chewBBACA includes modules to download schemas from Chewie-NS, upload new schemas to Chewie-NS, and synchronize local schemas used by users with the remote versions deposited in Chewie-NS to keep both versions up-to-date and maintain a common allelic nomenclature.

Click here to see the Available Schemas

Tutorial

Please visit the Tutorial site here (or by clicking on the test tube icon on the sidebar). The tutorial instructions are available here.

Schema submission

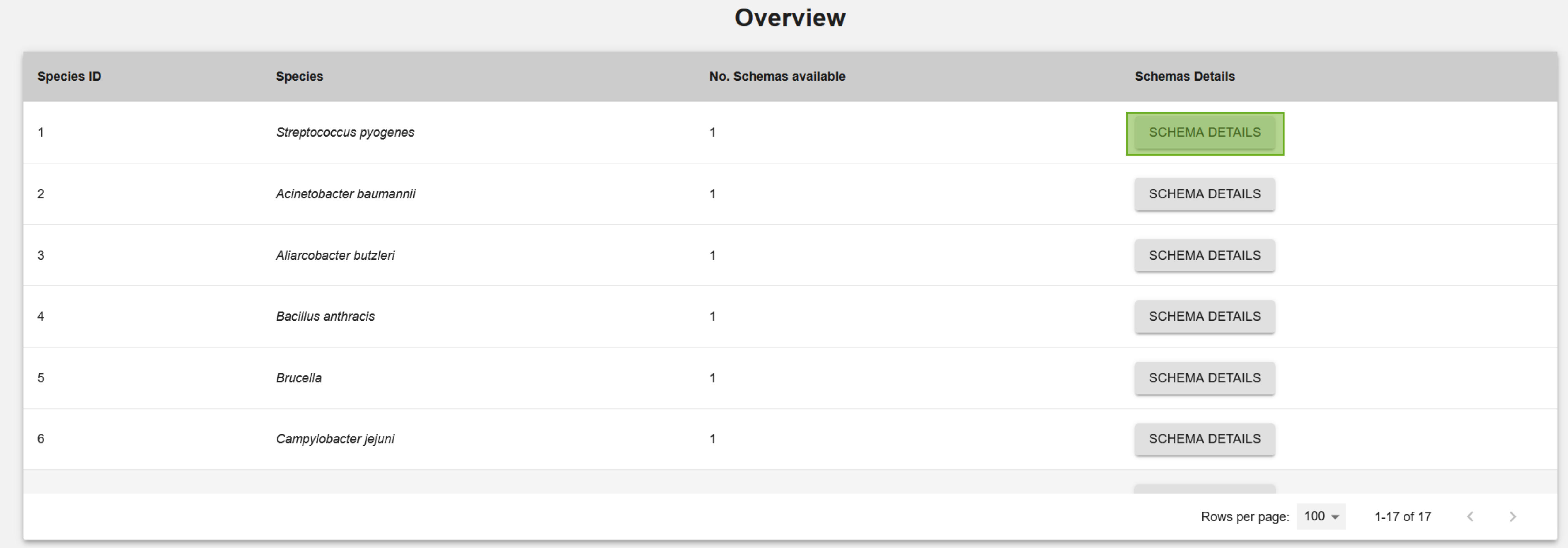
If you wish to submit schemas to Chewie-NS you need to Register and request for Contributor-level privileges by sending an email to imm-bioinfo@medicina.ulisboa.pt.

Citation

If you use **Chewie-NS**, please cite:

Mamede, R., Vila-Cerqueira, P., Silva, M., Carriço, J. A., & Ramirez, M. (2020). Chewie Nomenclature Server (chewie-NS): a deployable nomenclature server for easy sharing of core and whole genome MLST schemas. Nucleic Acids Research.

Chewie-NS: a repository of existing schemas



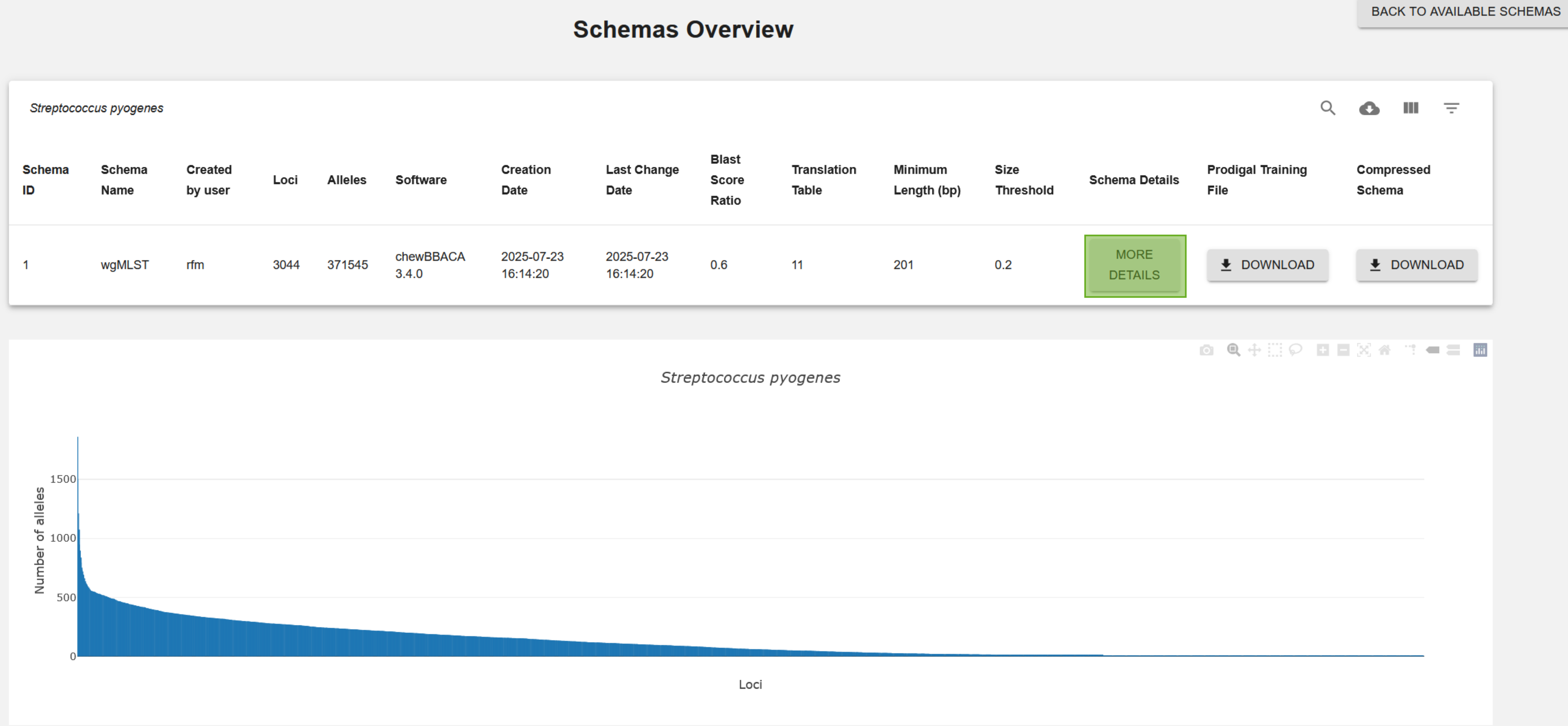
Overview

Species ID	Species	No. Schemas available	Schemas Details
1	*Streptococcus pyogenes*	1	SCHEMA DETAILS
2	*Acinetobacter baumannii*	1	SCHEMA DETAILS
3	*Aliarcobacter butzleri*	1	SCHEMA DETAILS
4	*Bacillus anthracis*	1	SCHEMA DETAILS
5	*Brucella*	1	SCHEMA DETAILS
6	*Campylobacter jejuni*	1	SCHEMA DETAILS

Rows per page: 100

1-17 of 17

Chewie-NS: store schema information



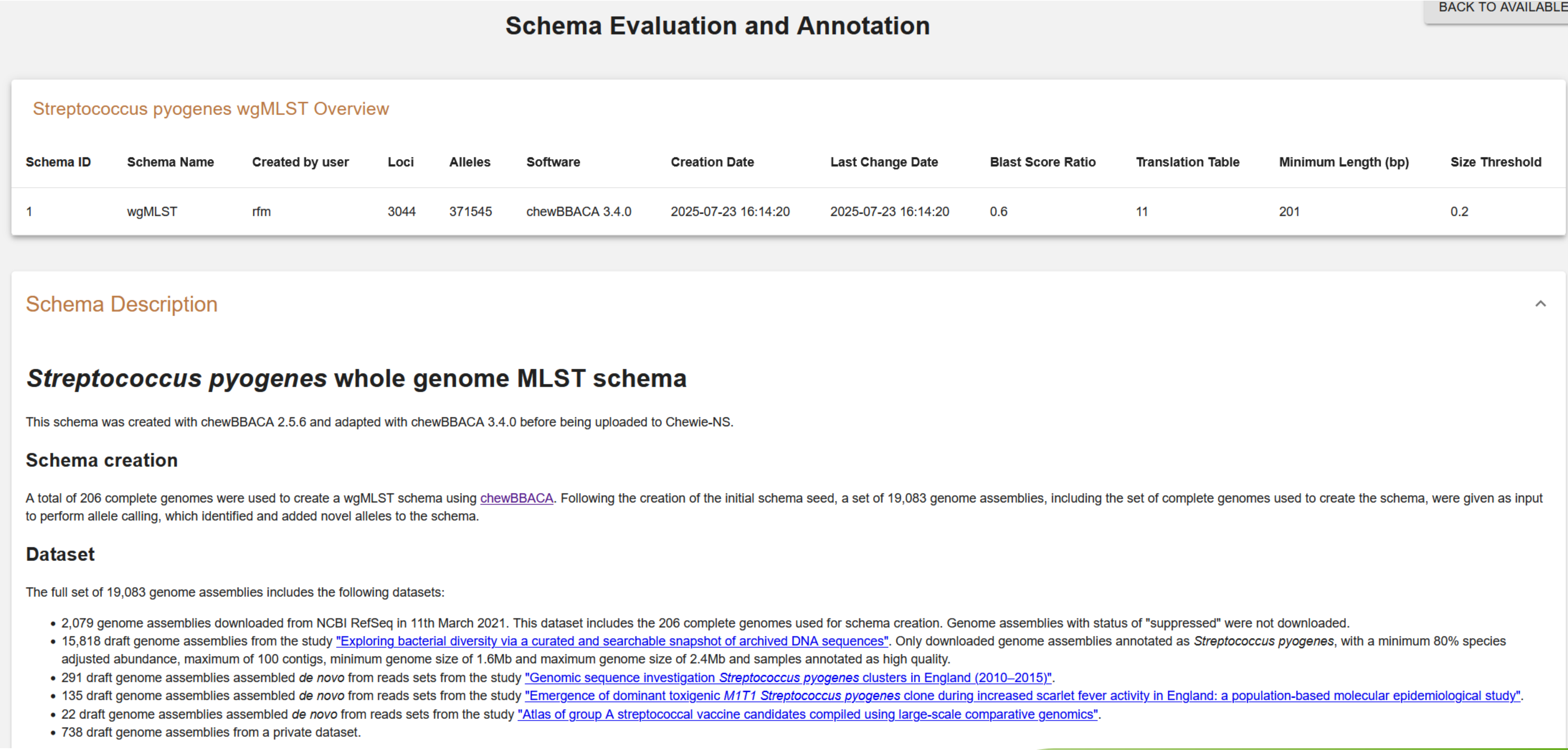
BACK TO AVAILABLE SCHEMAS

Schemas Overview

Streptococcus pyogenes

Schema ID	Schema Name	Created by user	Loci	Alleles	Software	Creation Date	Last Change Date	Blast Score Ratio	Translation Table	Minimum Length (bp)	Size Threshold	Schema Details	Prodigal Training File	Compressed Schema
1	wgMLST	rfm	3044	371545	chewBBACA 3.4.0	2025-07-23 16:14:20	2025-07-23 16:14:20	0.6	11	201	0.2	MORE DETAILS	DOWNLOAD	DOWNLOAD

Chewie-NS: schema description



Schema Evaluation and Annotation

Streptococcus pyogenes wgMLST Overview

Schema ID	Schema Name	Created by user	Loci	Alleles	Software	Creation Date	Last Change Date	Blast Score Ratio	Translation Table	Minimum Length (bp)	Size Threshold
1	wgMLST	rfm	3044	371545	chewBBACA 3.4.0	2025-07-23 16:14:20	2025-07-23 16:14:20	0.6	11	201	0.2

Schema Description

Streptococcus pyogenes whole genome MLST schema

This schema was created with chewBBACA 2.5.6 and adapted with chewBBACA 3.4.0 before being uploaded to Chewie-NS.

Schema creation

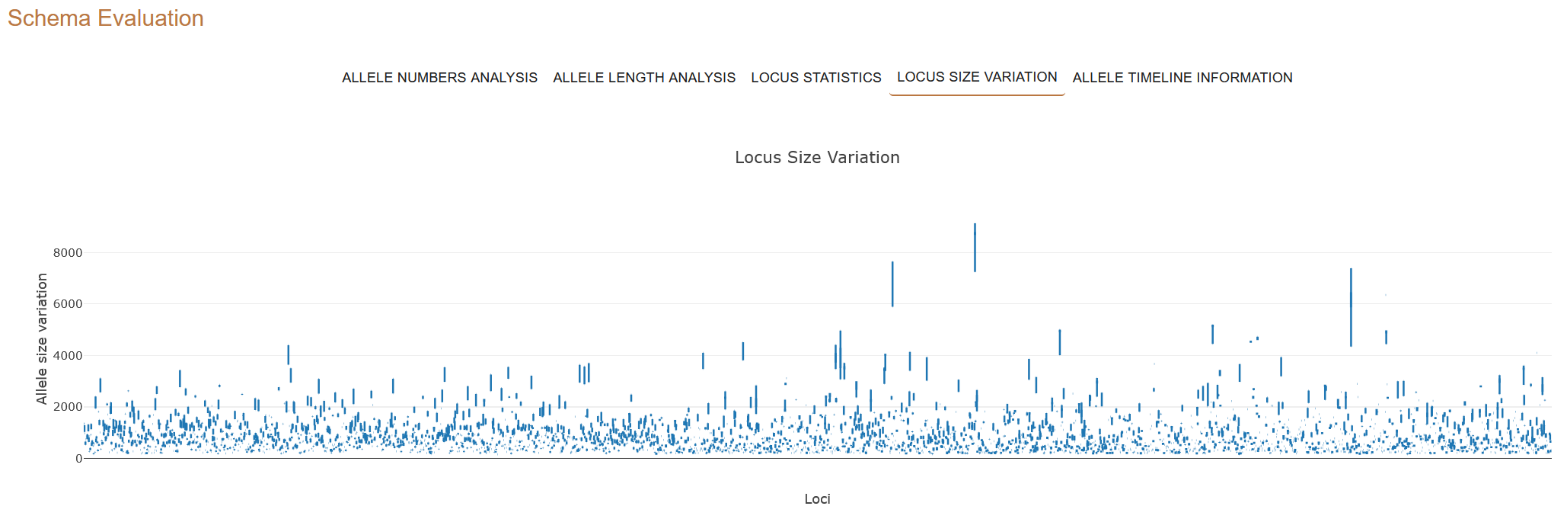
A total of 206 complete genomes were used to create a wgMLST schema using chewBBACA. Following the creation of the initial schema seed, a set of 19,083 genome assemblies, including the set of complete genomes used to create the schema, were given as input to perform allele calling, which identified and added novel alleles to the schema.

Dataset

The full set of 19,083 genome assemblies includes the following datasets:

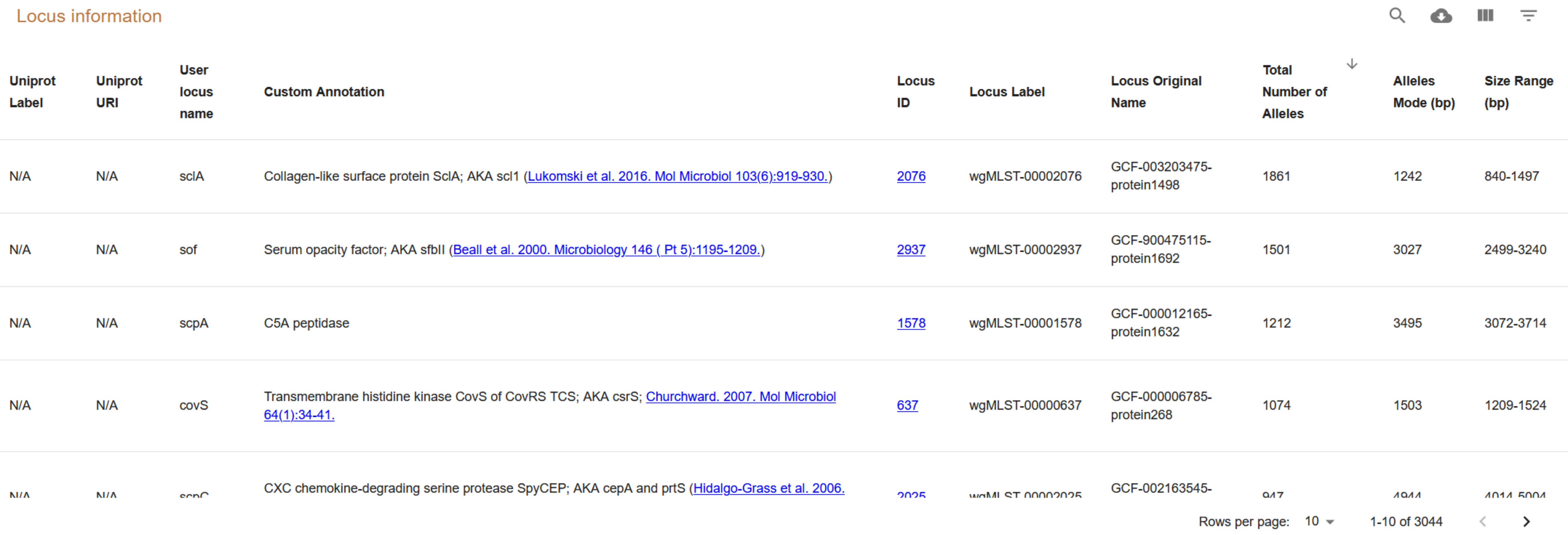
- 2,079 genome assemblies downloaded from NCBI RefSeq in 11th March 2021. This dataset includes the 206 complete genomes used for schema creation. Genome assemblies with status of "suppressed" were not downloaded.
- 15,818 draft genome assemblies from the study "Exploring bacterial diversity via a curated and searchable snapshot of archived DNA sequences". Only downloaded genome assemblies annotated as *Streptococcus pyogenes*, with a minimum 80% species adjusted abundance, maximum of 100 contigs, minimum genome size of 1.6Mb and maximum genome size of 2.4Mb and samples annotated as high quality.
- 291 draft genome assemblies assembled *de novo* from reads sets from the study "Genomic sequence investigation *Streptococcus pyogenes* clusters in England (2010–2015)".
- 135 draft genome assemblies assembled *de novo* from reads sets from the study "Emergence of dominant toxigenic *M1T1 Streptococcus pyogenes* clone during increased scarlet fever activity in England: a population-based molecular epidemiological study".
- 22 draft genome assemblies assembled *de novo* from reads sets from the study "Atlas of group A streptococcal vaccine candidates compiled using large-scale comparative genomics".
- 738 draft genome assemblies from a private dataset.

Chewie-NS: schema statistics



Schema Evaluation

Chewie-NS: locus information



Locus information

Uniprot Label	Uniprot URI	User locus name	Custom Annotation	Locus ID	Locus Label	Locus Original Name	Total Number of Alleles	Alleles Mode (bp)	Size Range (bp)
N/A	N/A	sclA	Collagen-like surface protein SclA; AKA scl1 (Lukomski et al. 2016. Mol Microbiol 103(6):919-930.)	2076	wgMLST-00002076	GCF-003203475-protein1498	1861	1242	840-1497
N/A	N/A	sof	Serum opacity factor; AKA sfbII (Beall et al. 2000. Microbiology 146 (Pt 5):1195-1209.)	2937	wgMLST-00002937	GCF-900475115-protein1692	1501	3027	2499-3240
N/A	N/A	scpA	C5A peptidase	1578	wgMLST-00001578	GCF-000012165-protein1632	1212	3495	3072-3714
N/A	N/A	covS	Transmembrane histidine kinase CovS of CovRS TCS; AKA csrS; Churchward. 2007. Mol Microbiol 64(1):34-41.	637	wgMLST-00000637	GCF-000006785-protein268	1074	1503	1209-1524
N/A	N/A	scpC	CXC chemokine-degrading serine protease SpyCEP; AKA cepA and prtS (Hidalgo-Grass et al. 2006.	2025	wgMLST-00002025	GCF-002163545-	947	4944	4014-5004

Rows per page: 10 1-10 of 3044