

# Introduction to core genome MLST (cgMLST)

Federica Gigliucci

Bioinformatics course,  
11-12 July 2019

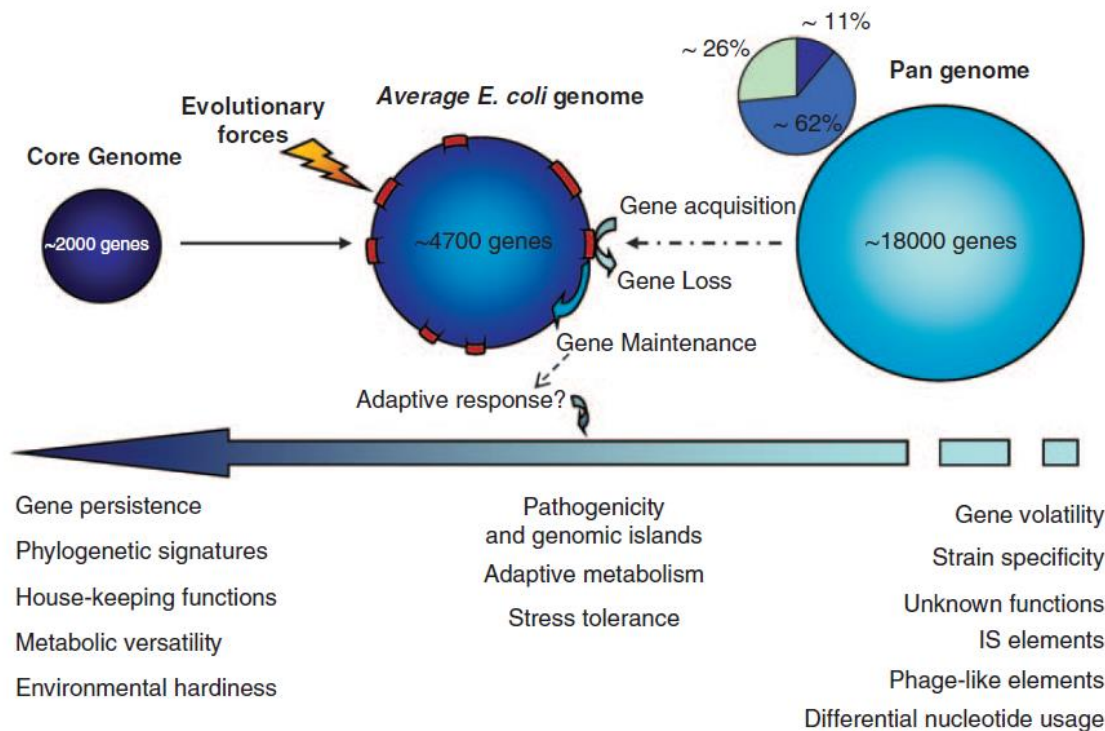


Istituto Superiore di Sanità, Dep. of Food Safety, Nutrition and Veterinary Public Health  
European Union and National Reference Laboratory for *E. coli*, Rome, Italy



# The *E. coli* pangenome

## Genomic plasticity



Van Elsas J.D. et al., 2011

Pangenome

Whole genome

Core genome

Accessory genome

Housekeeping



# Applying MLST to *E. coli*

---

## Conventional MLST

7 housekeeping genes

Low sensitivity

Good for phylogenetic analysis

High robustness

Not good enough for outbreak investigation

## MLST from WGS data



whole genome (**wgMLST**) – set of loci present in at least one strain



core genes (**cgMLST**) – set of loci present in the 95% of the strains



housekeeping genes (**7-genesMLST**)

# Public database hosting MLST schemas

Enterobase

## Available Databases

<p><b>Salmonella</b> Strains:214177</p> <p>Assembled</p> <ul style="list-style-type: none"><li>Legacy:5555</li><li>From NGS:203822</li><li>In Progress:715</li></ul> <p>Schemas</p> <ul style="list-style-type: none"><li>Achtman 7 Gene MLST:212531</li><li>cgMLST V2 + HierCC:206311</li><li>rMLST:206323</li><li>wgMLST:206278</li></ul> <p>Database Home <a href="#">↗</a></p>	<p><b>Escherichia/Shigella</b> Strains:111098</p> <p>Assembled</p> <ul style="list-style-type: none"><li>Legacy:9612</li><li>From NGS:191458</li><li>In Progress:582</li></ul> <p>Schemas</p> <ul style="list-style-type: none"><li>Achtman 7 Gene MLST:110289</li><li>cgMLST V1 + HierCC:190704</li><li>rMLST:100654</li><li>wgMLST:100418</li></ul> <p>Database Home <a href="#">↗</a></p>	<p><b>Clostridioides</b> Strains:14155</p> <p>Assembled</p> <ul style="list-style-type: none"><li>From NGS:14155</li><li>In Progress:0</li></ul> <p>Schemas</p> <ul style="list-style-type: none"><li>cgMLST V1 + HierCC:14141</li><li>Griffiths 7 Gene:13914</li><li>rMLST:14125</li><li>wgMLST:14149</li></ul> <p>Database Home <a href="#">↗</a></p>
<p><b>Vibrio</b> Strains:6981</p> <p>Assembled</p> <ul style="list-style-type: none"><li>From NGS:9981</li><li>In Progress:0</li></ul> <p>Schemas</p> <ul style="list-style-type: none"><li>rMLST:8928</li></ul> <p>Database Home <a href="#">↗</a></p>	<p><b>Yersinia</b> Strains:3617</p> <p>Assembled</p> <ul style="list-style-type: none"><li>Legacy:831</li><li>From NGS:2738</li><li>In Progress:1</li></ul> <p>Schemas</p> <ul style="list-style-type: none"><li>Achtman 7 Gene:3337</li><li>cgMLST V1 + HierCC:2733</li><li>McNally 7 Gene:2969</li><li>rMLST:2737</li><li>wgMLST:2751</li></ul> <p>Database Home <a href="#">↗</a></p>	<p><b>Helicobacter</b> Strains:2425</p> <p>Assembled</p> <ul style="list-style-type: none"><li>From NGS:2425</li><li>In Progress:1</li></ul> <p>Schemas</p> <ul style="list-style-type: none"><li>rMLST:2425</li></ul> <p>Database Home <a href="#">↗</a></p>
<p><b>Moraxella</b> Strains:1053</p> <p>Assembled</p> <ul style="list-style-type: none"><li>Legacy:418</li><li>From NGS:837</li><li>In Progress:0</li></ul> <p>Schemas</p> <ul style="list-style-type: none"><li>Achtman 7 Gene:1954</li><li>rMLST:837</li></ul> <p>Database Home <a href="#">↗</a></p>		



Need Help? Not sure where to start? [Click here](#) to read the documentation  
Any Questions or comments? please post to our [issue tracker](#) (@BBSrcUK)  
[Support for users of the old 7-gene MLST site - click here](#)



Istituto Superiore di Sanità, Dep. of Food Safety, Nutrition and Veterinary Public Health  
European Union and National Reference Laboratory for *E. coli*, Rome, Italy



# chewBBACA: assembly based allele-calling of cgMLST

---

Developed by INNUENDO (EFSA-funded project)

Based on cgMLST scheme developed by Enterobase

*E. coli* scheme by chewBBACA: 2360 curated loci

<https://github.com/B-UMMI/chewBBACA>

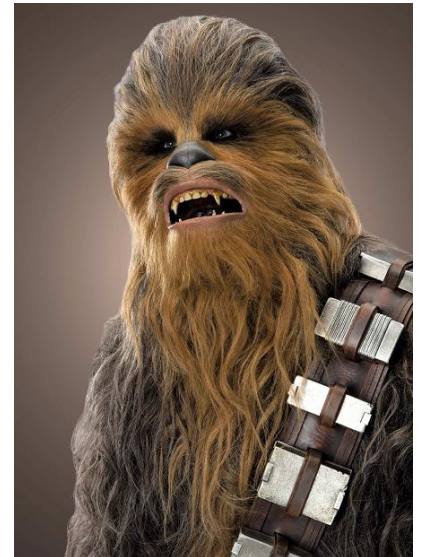
## MICROBIAL GENOMICS

Methods paper template

### chewBBACA: A complete suite for gene-by-gene schema creation and strain identification

Mickael Silva<sup>1</sup>, Miguel Machado<sup>1</sup>, Diogo N. Silva<sup>1</sup>, Mirko Rossi<sup>2</sup>, Jacob Moran-Gilad<sup>3,4</sup>, Sergio Santos<sup>1</sup>, Mario Ramirez<sup>1</sup> and João André Carriço<sup>1\*</sup>

1 Instituto de Microbiologia, Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, Lisbon, Portugal 2 Department of Food Hygiene and Environmental Health, Faculty of Veterinary Medicine, University of Helsinki, Finland 3 Faculty of Health Sciences, Ben-Gurion University of the Negev, Beer Sheva, Israel 4 Public Health Services, Ministry of Health, Jerusalem, Israel



# chewBBACA: assembly based allele-calling of cgMLST

---

## chewBBACA: BSR-Based Allele Calling Algorithm

Developed by INNUENDO (EFSA-funded project)

- It works on pre-assembled contigs (.fasta)
- Complete coding sequence (CDS) identified by Prodigal – BLASTp search – for each genome query
- Blast comparison between CDS of blastp-db Vs alleles of the cgMLST scheme
- BLAST Score Ratio (BSR) > 0.6 to identify the allele. The 0.6 value is related to a DNA identity of 80%

# chewBBACA results – Statistics

---

Genome	EXC	INF	LNF	PLOT	NIPH	ALM	ASM
NC_017162.fna	892	2319	1909	0	104	5	37
NC_011586.fna	1563	1697	1809	0	116	6	75

The column headers stand for:

- **EXC** - alleles which have exact matches (100% DNA identity) with previously identified alleles
- **INF** - inferred new alleles using Prodigal CDS predictions
- **LNF** - loci not found. No alleles were found for the number of loci in the schema shown. This means that, for those loci, there were no BLAST hits or they were not within the BSR threshold for allele assignment.
- **PLOT** - possible loci on the tip of the query genome contigs (see image below). A locus is classified as *PLOT* when the CDS of the query genome has a BLAST hit with a known larger allele that covers the CDS sequence entirely and the unaligned regions of the larger allele exceeds one of the query genome contigs ends. This could be an artifact caused by genome fragmentation resulting in a shorter CDS prediction by Prodigal. To avoid locus misclassification, loci in such situations are classified as *PLOT*.

# chewBBACA results – Statistics

Genome	EXC	INF	LNF	PLOT	NIPH	ALM	ASM
NC_017162.fna	892	2319	1909	0	104	5	37
NC_011586.fna	1563	1697	1809	0	116	6	75

The column headers stand for:

- **NIPH** - non-informative paralogous hit (see image below). When  $\geq 2$  CDSs in the query genome match one locus in the schema with a BSR  $> 0.6$ , that locus is classified as *NIPH*. This suggests that such locus can have paralogous (or orthologous) loci in the query genome and should be removed from the analysis due to the potential uncertainty in allele assignment (for example, due to the presence of multiple copies of the same mobile genetic element (MGE) or as a consequence of gene duplication followed by pseudogenization). A high number of NIPH may also indicate a poorly assembled genome due to a high number of smaller contigs which result in partial CDS predictions. These partial CDSs may contain conserved domains that match multiple loci. This classification takes precedence over *PLOT* classification.
- **ALM** - alleles 20% larger than length mode of the distribution of the matched loci (CDS length  $>$  (locus length mode + locus length mode \* 0.2)) (see image below). This determination is based on the currently identified set of alleles for a given locus.
- **ASM** - similar to *ALM* but for alleles 20% smaller than length mode distribution of the matched loci (CDS length  $<$  (locus length mode - locus length mode \* 0.2)).

**A high number of PLOT, ASM, ALM and/or NIPH usually indicates bad quality or contaminated assemblies.**



# chewBBACA on ARIES

**Galaxy / ARIES** Analyze Data Workflow Visualize Shared Data Admin Help User

Tools  ×

--- COMMON TOOLS ---

- Get Data
- Send Data
- Lift-Over
- Text Manipulation
- Filter and Sort
- Join, Subtract and Group
- Convert Formats
- Extract Features
- Fetch Sequences
- Fetch Alignments
- Statistics
- Graph/Display Data
- GraPhlAn
- IRIDA

---PHYLOGENY TOOLS---

- Phylogenetics
  - PopPUNK POPulation Partitioning Using Nucleotide Kmers
  - chewBBACA BSR-Based Allele Calling Algorithm**
  - ClustalW multiple sequence alignment program for DNA or proteins

**chewBBACA BSR-Based Allele Calling Algorithm (Galaxy Version 2.0)** Versions Options

Select function

AlleleCall: Perform allele call for target genomes

CreateSchema: Create a gene by gene schema based on genomes

**AlleleCall: Perform allele call for target genomes**

SchemaEvaluator: Tool that builds an html output to better navigate/visualize your schema

TestGenomeQuality: Analyze your allele call output to refine schemas

ExtractCgMLST: Select a subset of loci without missing data (to be used as PHYLOViZ input)

RemoveGenes: Remove a provided list of loci from your allele call output

Choose reference

schema\_Clostridioides\_difficile\_agMLST

minimum BSR score

0.6

taxon

Escherichia coli

**chewBBACA** stands for "BSR-Based Allele Calling Algorithm". The "chew" part could be thought of as "Comprehensive and Highly Efficient Workflow" but at this ; add extra coolness to the software name. This tool is in beta test.

The development of the tools have been supported by INNUENDO project (<https://www.innuendoweb.org>) co-funded by the European Food Safety Authority (EFSA identifying and characterizing microbial and chemical hazards) and by the ONEIDA project (LISBOA-01-0145-FEDER-016417) co-funded by FEEI - "Fundos Eurr Lisboa 2020" and by national funds from FCT - "Fundação para a Ciência e a Tecnologia" and BacGenTrack (TUBITAK/0004/2014) [FCT/ Scientific and Technolo; Kurumu, TÜBITAK].

# chewBBACA on ARIES

The screenshot shows the Galaxy / ARIES interface for the 'chewBBACA BSR-Based Allele Calling Algorithm (Galaxy Version 2.0)'. The tool is configured with the following settings:

- Select function:** AlleleCall: Perform allele call for target genomes
- Selection of genome files (fasta):** A list of five files: 450: Strain5\_contigs.fasta, 449: Strain4\_contigs.fasta, 448: Strain3\_contigs.fasta, 447: Strain2\_contigs.fasta, and 446: Strain1\_contigs.fasta.
- Which schema would you like to use as a reference?:** System reference (selected).
- Choose reference:** A dropdown menu showing 'schema\_chewBBACA\_cgMLST\_V4' as the selected option. Below it, a search bar contains 'm' and a list of other schemas is visible, including 'schema\_Clostridioides\_difficile\_agMLST', 'schema\_Clostridioides\_difficile\_cgMLST', 'schema\_enterobase\_V4\_called', and 'schema\_chewBBACA\_cgMLST\_V4' (highlighted in blue).
- Execute:** A blue button with a checkmark and the text 'Execute' is visible at the bottom left of the tool panel.

cgMLST scheme for *E. coli*: **Schema\_chewBBACA\_cgMLST\_V4**

This block shows a close-up of the 'minimum BSR score' and 'taxon' fields. Two red arrows point to the input fields:

- minimum BSR score:** The input field contains the value '0.6'.
- taxon:** The dropdown menu is set to 'Escherichia coli'.

At the bottom, there is a blue 'Execute' button with a checkmark.

# chewBBACA results on ARIES

## Statistics

Genome	EXC	INF	LNF	PLOT	NIPH	ALM	ASM
Genome	EXC	INF	LNF	PLOT	NIPH	ALM	ASM
ED1032_contigs	3543	16	4007	0	27	3	5
ED1088_contigs	3348	77	4120	3	16	2	35
ED1089_contigs.fasta	3105	116	4263	6	11	16	84
ED1104_contigs.fasta	3493	4	4055	1	13	5	30
ED1105_contigs.fasta	3433	12	4098	1	14	4	39

## Contigs info

FILE	b0073.fasta	b0074.fasta	b0075.fasta
ED1032_contigs	scaffold_0&199417-198324&-	scaffold_0&200988-199415&-	LNF
ED1088_contigs	NODE_1_length_228150_cov_40.8159_ID_1&198543-197450&-	NODE_1_length_228150_cov_40.8159_ID_1&200114-198541&-	LNF
ED1089_contigs.fasta	NODE_1_length_227956_cov_19.4419_ID_1&197229-196136&-	NODE_1_length_227956_cov_19.4419_ID_1&198800-197227&-	LNF
ED1104_contigs.fasta	NODE_4_length_186376_cov_34.6136_ID_7&29626-30717&+	NODE_4_length_186376_cov_34.6136_ID_7&28055-29626&+	LNF
ED1105_contigs.fasta	NODE_4_length_186376_cov_40.795_ID_7&29626-30717&+	NODE_4_length_186376_cov_40.795_ID_7&28055-29626&+	LNF

## Alleles

### Target genome file names

### Allele call data for loci present in the schema

FILE	b0073.fasta	b0074.fasta	b0075.fasta	b0076.fasta	b0077.fasta	b0078.fasta
ED1032_contigs	10	11	LNF	460	13	2
ED1088_contigs	10	11	LNF	3	13	2
ED1089_contigs.fasta	10	11	LNF	3	13	2
ED1104_contigs.fasta	10	11	LNF	3	13	2
ED1105_contigs.fasta	10	11	LNF	3	13	2

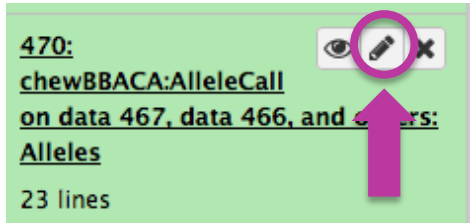
**File to use for cluster analysis**

## Logging info

## Repeated loci

# Mentalist Distance Matrix tool calculates the number of allelic differences between strains

470:  
chewBBACA:AlleleCall  
on data 467, data 466, and others:  
Alleles  
23 lines



## Edit dataset attributes

Attributes Convert Datatypes Permissions

### Change datatype

Change datatype

#### New Type

tsv

tabu

dbnsfp.tabular

tabular

## Galaxy / ARIES

Analyze Data Workflow Visualize Shared Data Admin Help User

### Tools

search tools

--- COMMON TOOLS ---

[Get Data](#)

[Send Data](#)

[Lift-Over](#)

[Text Manipulation](#)

[Filter and Sort](#)

[Join, Subtract and Group](#)

[Convert Formats](#)

[Extract Features](#)

[Fetch Sequences](#)

[Fetch Alignments](#)

[Statistics](#)

### MentaliST Distance Matrix (Galaxy Version 0.1.9)

Options

#### MentaliST Calls

470: chewBBACA:AlleleCall on data 467, data 466, and others: Alleles

Execute

mentalist\_distance <input.tsv>



# Distance Matrix with allelic differences

	Strain10_c ontigs.fasta	Strain11_c ontigs.fasta	Strain12_c ontigs.fasta	Strain13_c ontigs.fasta	Strain14_c ontigs.fasta	Strain15_c ontigs.fasta	Strain16_c ontigs.fasta	Strain17_c ontigs.fasta	Strain18_c ontigs.fasta	Strain19_c ontigs.fasta	Strain1_co ntigs.fasta	Strain20_c ontigs.fasta	Strain21_c ontigs.fasta	Strain22_c ontigs.fasta	Strain2_co ntigs.fasta	Strain3_co ntigs.fasta	Strain4_co ntigs.fasta	Strain5_co ntigs.fasta	Strain6_co ntigs.fasta	Strain7_co ntigs.fasta	Strain8_co ntigs.fasta	Strain9_co ntigs.fasta
Strain10_c_ontigs.fasta	0	434	174	1647	1649	1647	1665	1650	1646	1640	536	2235	2237	2232	536	538	536	158	166	70	0	0
Strain11_c_ontigs.fasta	434	0	531	1737	1739	1737	1752	1737	1740	1729	820	2242	2241	2234	820	821	819	514	523	444	434	434
Strain12_c_ontigs.fasta	174	531	0	1643	1645	1643	1663	1646	1644	1636	536	2237	2239	2234	537	539	537	168	40	182	174	174
Strain13_c_ontigs.fasta	1647	1737	1643	0	9	5	193	62	63	156	1648	2231	2232	2230	1647	1650	1648	1637	1643	1648	1647	1647
Strain14_c_ontigs.fasta	1649	1739	1645	9	0	8	192	62	59	154	1650	2232	2233	2231	1649	1652	1650	1639	1645	1650	1649	1649
Strain15_c_ontigs.fasta	1647	1737	1643	5	8	0	191	59	61	153	1648	2231	2232	2230	1647	1650	1648	1637	1643	1648	1647	1647
Strain16_c_ontigs.fasta	1665	1752	1663	193	192	191	0	198	193	192	1664	2233	2236	2231	1663	1666	1664	1658	1665	1663	1665	1665
Strain17_c_ontigs.fasta	1650	1737	1646	62	62	59	198	0	69	159	1650	2231	2233	2229	1649	1652	1650	1639	1645	1651	1650	1650
Strain18_c_ontigs.fasta	1646	1740	1644	63	59	61	193	69	0	152	1647	2231	2233	2230	1646	1649	1647	1638	1644	1647	1646	1646
Strain19_c_ontigs.fasta	1640	1729	1636	156	154	153	192	159	152	0	1639	2229	2232	2230	1638	1641	1639	1630	1637	1639	1640	1640
Strain1_c_ontigs.fasta	536	820	536	1648	1650	1648	1664	1650	1647	1639	0	2239	2241	2235	2	4	3	501	530	539	536	536
Strain20_c_ontigs.fasta	2235	2242	2237	2231	2232	2231	2233	2231	2231	2229	2239	0	226	302	2238	2238	2238	2235	2237	2235	2235	2235
Strain21_c_ontigs.fasta	2237	2241	2239	2232	2233	2232	2236	2233	2233	2232	2241	226	0	237	2240	2240	2240	2237	2239	2237	2237	2237
Strain22_c_ontigs.fasta	2232	2234	2234	2230	2231	2230	2231	2229	2230	2230	2235	302	237	0	2234	2234	2234	2232	2234	2232	2232	2232
Strain2_c_ontigs.fasta	536	820	537	1647	1649	1647	1663	1649	1646	1638	2	2238	2240	2234	0	4	3	501	531	539	536	536
Strain3_c_ontigs.fasta	538	821	539	1650	1652	1650	1666	1652	1649	1641	4	2238	2240	2234	4	0	5	503	533	541	538	538
Strain4_c_ontigs.fasta	536	819	537	1648	1650	1648	1664	1650	1647	1639	3	2238	2240	2234	3	5	0	501	531	539	536	536
Strain5_c_ontigs.fasta	158	514	168	1637	1639	1637	1658	1639	1638	1630	501	2235	2237	2232	501	503	501	0	160	166	158	158
Strain6_c_ontigs.fasta	166	523	40	1643	1645	1643	1665	1645	1644	1637	530	2237	2239	2234	531	533	531	160	0	175	166	166
Strain7_c_ontigs.fasta	70	444	182	1648	1650	1648	1663	1651	1647	1639	539	2235	2237	2232	539	541	539	166	175	0	70	70
Strain8_c_ontigs.fasta	0	434	174	1647	1649	1647	1665	1650	1646	1640	536	2235	2237	2232	536	538	536	158	166	70	0	0
Strain9_c_ontigs.fasta	0	434	174	1647	1649	1647	1665	1650	1646	1640	536	2235	2237	2232	536	538	536	158	166	70	0	0

Threshold of maximum **10 - 15 allelic differences** to consider *E. coli* strains related

# Mentalist Tree tool uses the matrix with allelic differences to build a phylogenetic tree

The screenshot displays the Galaxy / ARIES web interface. The top navigation bar includes 'Galaxy / ARIES', 'Analyze Data', 'Workflow', 'Visualize', 'Shared Data', 'Admin', 'Help', and 'User'. On the left, a 'Tools' sidebar lists various categories: 'COMMON TOOLS' (Get Data, Send Data, Lift-Over, Text Manipulation, Filter and Sort, Join, Subtract and Group, Convert Formats, Extract Features, Fetch Sequences, Fetch Alignments, Statistics, Graph/Display Data, GraPhlAn, IRIDA) and 'PHYLOGENY TOOLS' (Phylogenetics, MLST 7 Loci, MLST Scans genomes against PubMLST schemes, MLST List, SRST2.7 loci, MentaLIST MLST Analysis 0.2.3, MentaLIST MLST Analysis, MentaLIST Tree, MentaLIST Distance Matrix, MentaLIST MLST Analysis single-end reads). Three red arrows point to 'Phylogenetics', 'MLST 7 Loci', and 'MentaLIST Tree'. The main workspace shows the 'MentaLIST Tree (Galaxy Version 0.1.9)' tool configuration. It includes a 'MentaLIST Distance Matrix' dropdown menu with the value '473: MentaLIST Distance Matrix on data 470' and an 'Execute' button. Below the configuration, the command 'mentalist\_distance <input.tsv>' is visible.

