# Introduction to core genome MLST (cgMLST)

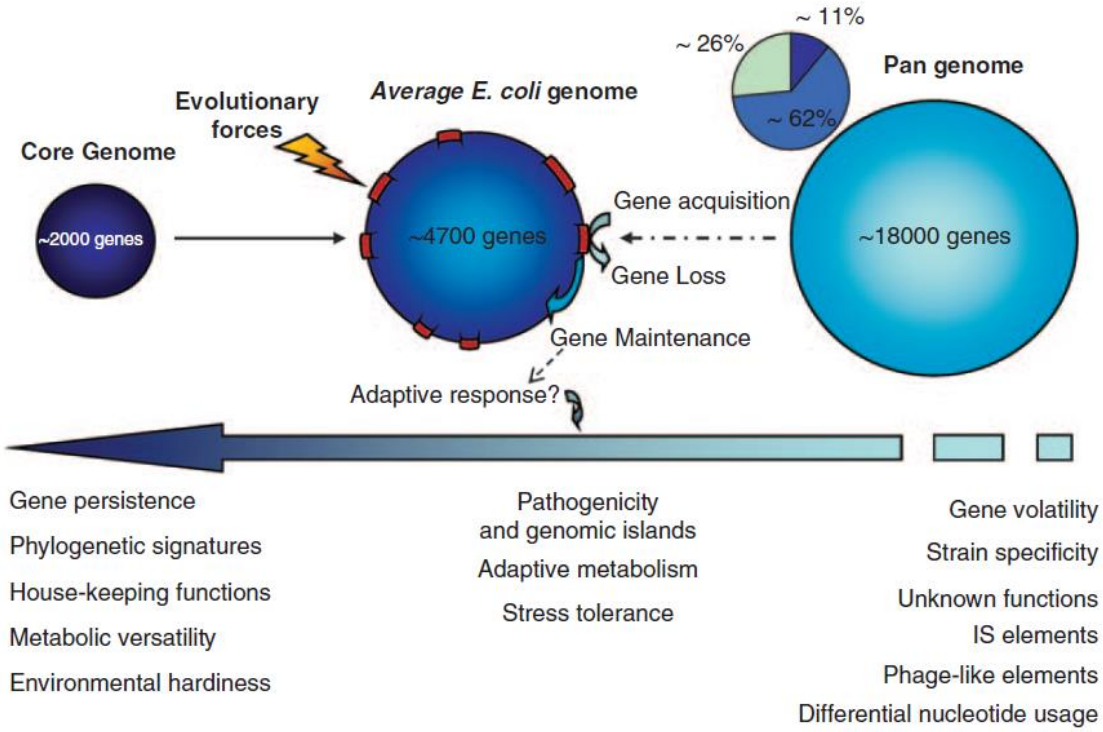
Federica Gigliucci

Bioinformatics course,
19-20 October 2020

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** Housekeeping **Accessory genome**

# Applying MLST to *E. coli*

## Conventional MLST

7 housekeeping genes

Low sensitivity — Good for phylogenetic analysis

High robustness — Not good enough for outbreak investigation

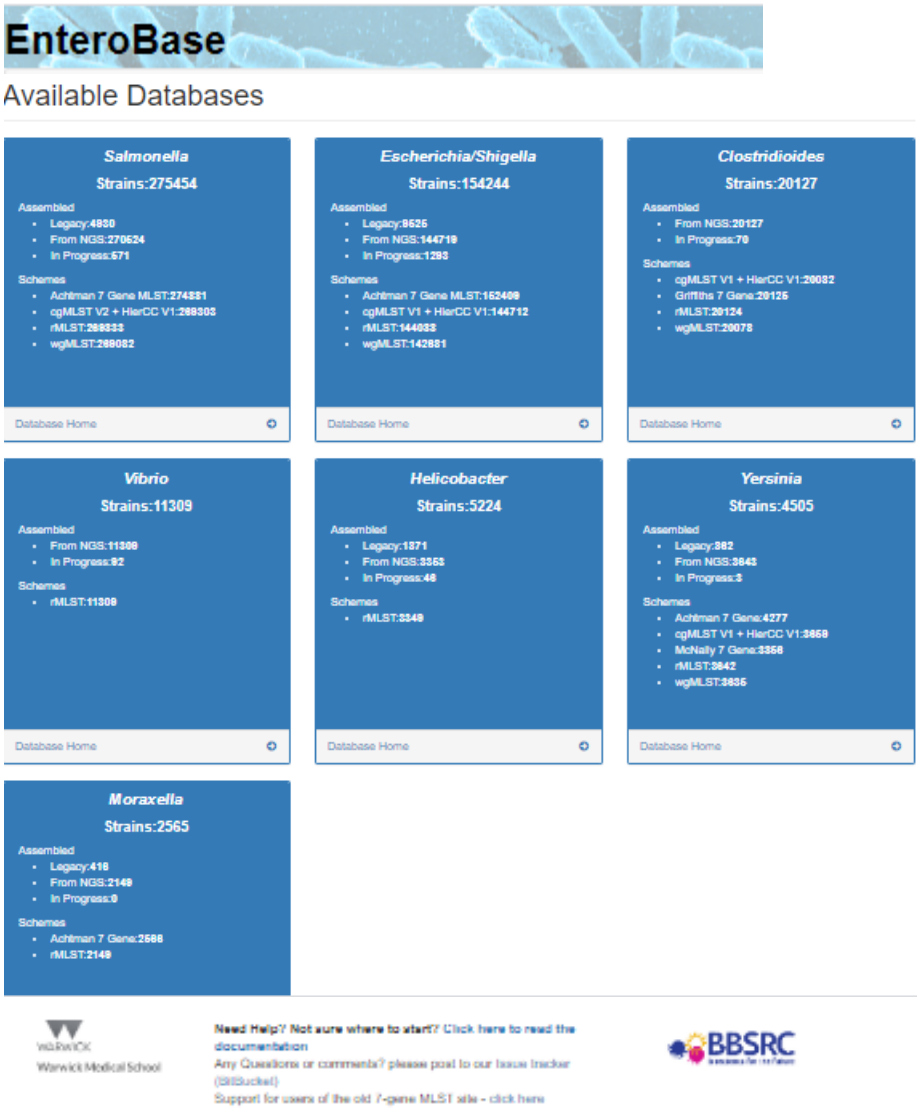
## MLST from WGS data

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

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

housekeeping genes (7-genesMLST)

# Public database hosting MLST schemes



# chewBBACA: assembly based allele-calling of cgMLST

Developed by INNUENDO (EFSA-funded project)

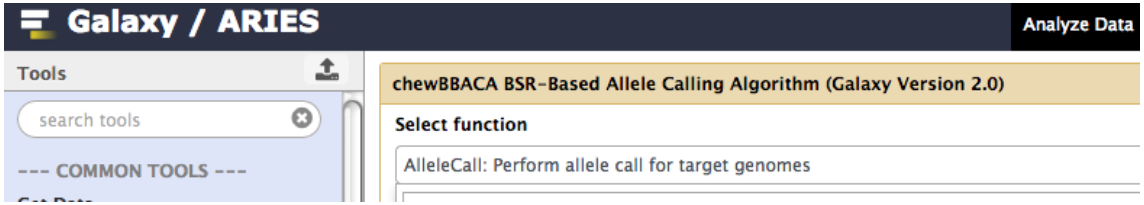
**Based on cgMLST scheme developed by Enterobase**

***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[1], Miguel Machado[1], Diogo N. Silva[1], Mirko Rossi[2], Jacob Moran-Gilad[3,4], Sergio Santos[1], Mario Ramirez[1] and João André Carriço[1]*

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 ≥ 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 my 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 results – Statistics

| Genome | EXC | INF | LNF | PLOT | NIPH | ALM | ASM |
|---|---|---|---|---|---|---|---|
| Genome | EXC | INF | LNF | PLOT | NIPH | ALM | ASM |
| ED0655-phantastic_contigs.fasta | 2349 | 0 | 6 | 0 | 3 | 0 | 2 |
| ED0696-phantastic_contigs.fasta | 2350 | 0 | 6 | 0 | 3 | 0 | 1 |
| ED0812-phantastic_contigs.fasta | 2338 | 0 | 9 | 3 | 4 | 0 | 6 |
| ED0813-phantastic_contigs.fasta | 2347 | 0 | 6 | 0 | 4 | 0 | 3 |
| ED0840-phantastic_contigs.fasta | 2345 | 0 | 6 | 1 | 4 | 0 | 4 |
| ED0884-phantastic_contigs.fasta | 2347 | 0 | 4 | 2 | 3 | 0 | 4 |
| ED0918-phantastic_contigs.fasta | 2351 | 0 | 3 | 0 | 3 | 0 | 3 |
| ED1000-phantastic_contigs.fasta | 2346 | 0 | 7 | 0 | 3 | 0 | 4 |
| ED1001-phantastic_contigs.fasta | 2341 | 0 | 8 | 2 | 4 | 0 | 5 |
| ED1029-phantastic_contigs.fasta | 2350 | 0 | 6 | 0 | 3 | 0 | 1 |
| ED1049-phantastic_contigs.fasta | 2351 | 0 | 4 | 0 | 4 | 0 | 1 |
| ED1152-phantastic_contigs.fasta | 2347 | 0 | 5 | 2 | 4 | 0 | 2 |
| ED1232-phantastic_contigs.fasta | 2346 | 0 | 6 | 3 | 3 | 0 | 2 |
| ED1269-phantastic_contigs.fasta | 2304 | 0 | 48 | 1 | 3 | 1 | 3 |
| ED1273-phantastic_contigs.fasta | 2349 | 0 | 4 | 0 | 4 | 0 | 3 |
| ED1301-phantastic_contigs.fasta | 2346 | 0 | 7 | 0 | 4 | 1 | 2 |
| ED1304-phantastic_contigs.fasta | 2343 | 0 | 10 | 3 | 2 | 0 | 2 |
| ED1308-phantastic_contigs.fasta | 2341 | 0 | 14 | 1 | 3 | 0 | 1 |
| ED1319-phantastic_contigs.fasta | 2339 | 0 | 8 | 0 | 4 | 0 | 9 |
| EF0453-phantastic_contigs.fasta | 1867 | 0 | 318 | 86 | 1 | 4 | 84 |

**1867 alleles called out of 2360 loci of the cgMLST scheme**

# 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 |

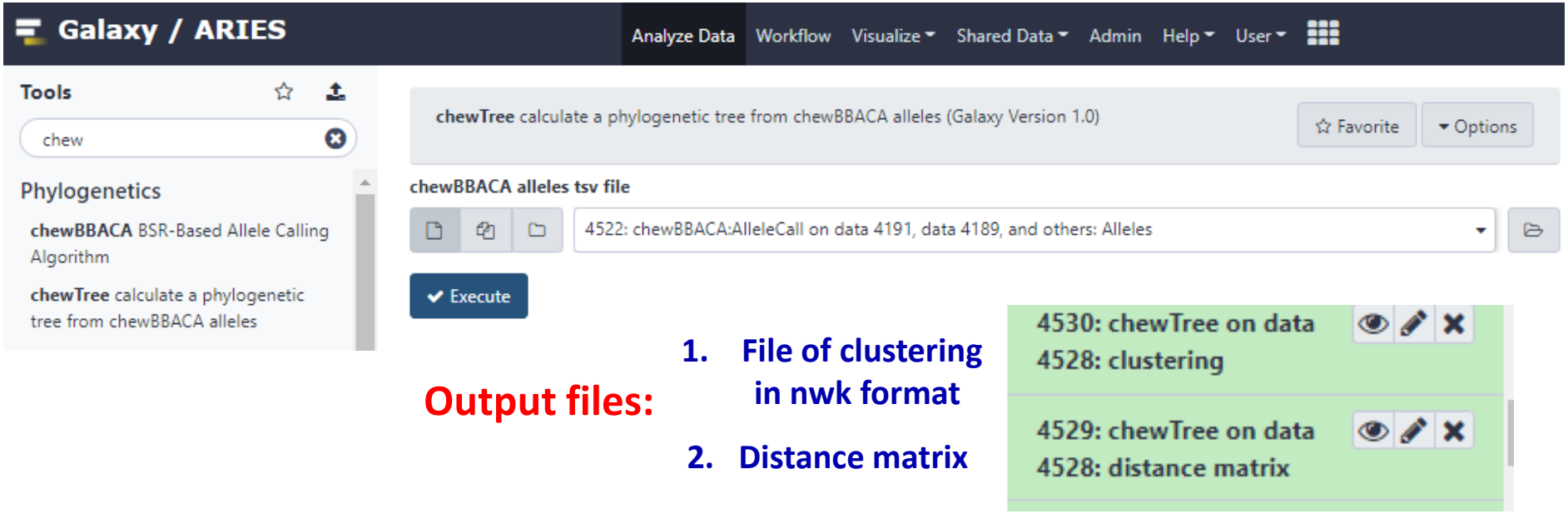
**File to use for cluster analysis**

## Logging info

## Repeated loci

# Cluster analysis - chewTree on ARIES

## Calculate a phylogenetic tree from chewBBACA alleles




- Pairwise comparison of the allelic profiles to calculate the distance between strains
- For each pair of samples, the alleles not found or not correctly assigned to any locus were filter out from the pairwise comparison
- The pairwise comparison is considered as reliable when alleles corresponding to at least 80% of loci were called.
- The genomes with less than 80% of the alleles called are filtered out from the analysis

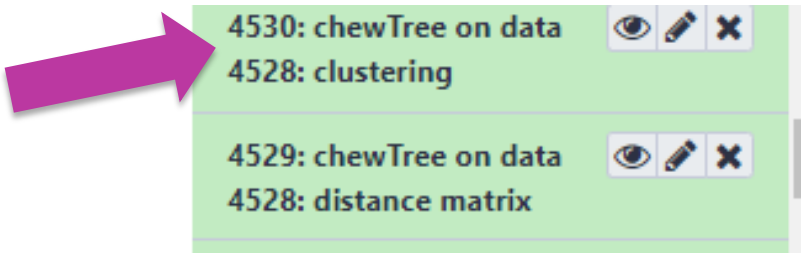
# Distance Matrix with allelic differences

| | Strain10_contigs.fasta | Strain11_contigs.fasta | Strain12_contigs.fasta | Strain13_contigs.fasta | Strain14_contigs.fasta | Strain15_contigs.fasta | Strain16_contigs.fasta | Strain17_contigs.fasta | Strain18_contigs.fasta | Strain19_contigs.fasta | Strain1_contigs.fasta | Strain20_contigs.fasta | Strain21_contigs.fasta | Strain22_contigs.fasta | Strain2_contigs.fasta | Strain3_contigs.fasta | Strain4_contigs.fasta | Strain5_contigs.fasta | Strain6_contigs.fasta | Strain7_contigs.fasta | Strain8_contigs.fasta | Strain9_contigs.fasta |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Strain10_contigs.fasta** | 0 | 434 | 174 | 1647 | 1649 | 1647 | 1665 | 1650 | 1646 | 1640 | 536 | 2235 | 2237 | 2232 | 536 | 538 | 536 | 158 | 166 | 70 | 0 | 0 |
| **Strain11_contigs.fasta** | 434 | 0 | 531 | 1737 | 1739 | 1737 | 1752 | 1737 | 1740 | 1729 | 820 | 2242 | 2241 | 2234 | 820 | 821 | 819 | 514 | 523 | 444 | 434 | 434 |
| **Strain12_contigs.fasta** | 174 | 531 | 0 | 1643 | 1645 | 1643 | 1663 | 1646 | 1644 | 1636 | 536 | 2237 | 2239 | 2234 | 537 | 539 | 537 | 168 | 40 | 182 | 174 | 174 |
| **Strain13_contigs.fasta** | 1647 | 1737 | 1643 | 0 | 9 | 5 | 193 | 62 | 63 | 156 | 1648 | 2231 | 2232 | 2230 | 1647 | 1650 | 1648 | 1637 | 1643 | 1648 | 1647 | 1647 |
| **Strain14_contigs.fasta** | 1649 | 1739 | 1645 | 9 | 0 | 8 | 192 | 62 | 59 | 154 | 1650 | 2232 | 2233 | 2231 | 1649 | 1652 | 1650 | 1639 | 1645 | 1650 | 1649 | 1649 |
| **Strain15_contigs.fasta** | 1647 | 1737 | 1643 | 5 | 8 | 0 | 191 | 59 | 61 | 153 | 1648 | 2231 | 2232 | 2230 | 1647 | 1650 | 1648 | 1637 | 1643 | 1648 | 1647 | 1647 |
| **Strain16_contigs.fasta** | 1665 | 1752 | 1663 | 193 | 192 | 191 | 0 | 198 | 193 | 192 | 1664 | 2233 | 2236 | 2231 | 1663 | 1666 | 1664 | 1658 | 1665 | 1663 | 1665 | 1665 |
| **Strain17_contigs.fasta** | 1650 | 1737 | 1646 | 62 | 62 | 59 | 198 | 0 | 69 | 159 | 1650 | 2231 | 2233 | 2229 | 1649 | 1652 | 1650 | 1639 | 1645 | 1651 | 1650 | 1650 |
| **Strain18_contigs.fasta** | 1646 | 1740 | 1644 | 63 | 59 | 61 | 193 | 69 | 0 | 152 | 1647 | 2231 | 2233 | 2230 | 1646 | 1649 | 1647 | 1638 | 1644 | 1647 | 1646 | 1646 |
| **Strain19_contigs.fasta** | 1640 | 1729 | 1636 | 156 | 154 | 153 | 192 | 159 | 152 | 0 | 1639 | 2229 | 2232 | 2230 | 1638 | 1641 | 1639 | 1630 | 1637 | 1639 | 1640 | 1640 |
| **Strain1_contigs.fasta** | 536 | 820 | 536 | 1648 | 1650 | 1648 | 1664 | 1650 | 1647 | 1639 | 0 | 2239 | 2241 | 2235 | 2 | 4 | 3 | 501 | 530 | 539 | 536 | 536 |
| **Strain20_contigs.fasta** | 2235 | 2242 | 2237 | 2231 | 2232 | 2231 | 2233 | 2231 | 2231 | 2229 | 2239 | 0 | 226 | 302 | 2238 | 2238 | 2238 | 2235 | 2237 | 2235 | 2235 | 2235 |
| **Strain21_contigs.fasta** | 2237 | 2241 | 2239 | 2232 | 2233 | 2232 | 2236 | 2233 | 2233 | 2232 | 2241 | 226 | 0 | 237 | 2240 | 2240 | 2240 | 2237 | 2239 | 2237 | 2237 | 2237 |
| **Strain22_contigs.fasta** | 2232 | 2234 | 2234 | 2230 | 2231 | 2230 | 2231 | 2229 | 2230 | 2230 | 2235 | 302 | 237 | 0 | 2234 | 2234 | 2234 | 2232 | 2234 | 2232 | 2232 | 2232 |
| **Strain2_contigs.fasta** | 536 | 820 | 537 | 1647 | 1649 | 1647 | 1663 | 1649 | 1646 | 1638 | 2 | 2238 | 2240 | 2234 | 0 | 4 | 3 | 501 | 531 | 539 | 536 | 536 |
| **Strain3_contigs.fasta** | 538 | 821 | 539 | 1650 | 1652 | 1650 | 1666 | 1652 | 1649 | 1641 | 4 | 2238 | 2240 | 2234 | 4 | 0 | 5 | 503 | 533 | 541 | 538 | 538 |
| **Strain4_contigs.fasta** | 536 | 819 | 537 | 1648 | 1650 | 1648 | 1664 | 1650 | 1647 | 1639 | 3 | 2238 | 2240 | 2234 | 3 | 5 | 0 | 501 | 531 | 539 | 536 | 536 |
| **Strain5_contigs.fasta** | 158 | 514 | 168 | 1637 | 1639 | 1637 | 1658 | 1639 | 1638 | 1630 | 501 | 2235 | 2237 | 2232 | 501 | 503 | 501 | 0 | 160 | 166 | 158 | 158 |
| **Strain6_contigs.fasta** | 166 | 523 | 40 | 1643 | 1645 | 1643 | 1665 | 1645 | 1644 | 1637 | 530 | 2237 | 2239 | 2234 | 531 | 533 | 531 | 160 | 0 | 175 | 166 | 166 |
| **Strain7_contigs.fasta** | 70 | 444 | 182 | 1648 | 1650 | 1648 | 1663 | 1651 | 1647 | 1639 | 539 | 2235 | 2237 | 2232 | 539 | 541 | 539 | 166 | 175 | 0 | 70 | 70 |
| **Strain8_contigs.fasta** | 0 | 434 | 174 | 1647 | 1649 | 1647 | 1665 | 1650 | 1646 | 1640 | 536 | 2235 | 2237 | 2232 | 536 | 538 | 536 | 158 | 166 | 70 | 0 | 0 |
| **Strain9_contigs.fasta** | 0 | 434 | 174 | 1647 | 1649 | 1647 | 1665 | 1650 | 1646 | 1640 | 536 | 2235 | 2237 | 2232 | 536 | 538 | 536 | 158 | 166 | 70 | 0 | 0 |

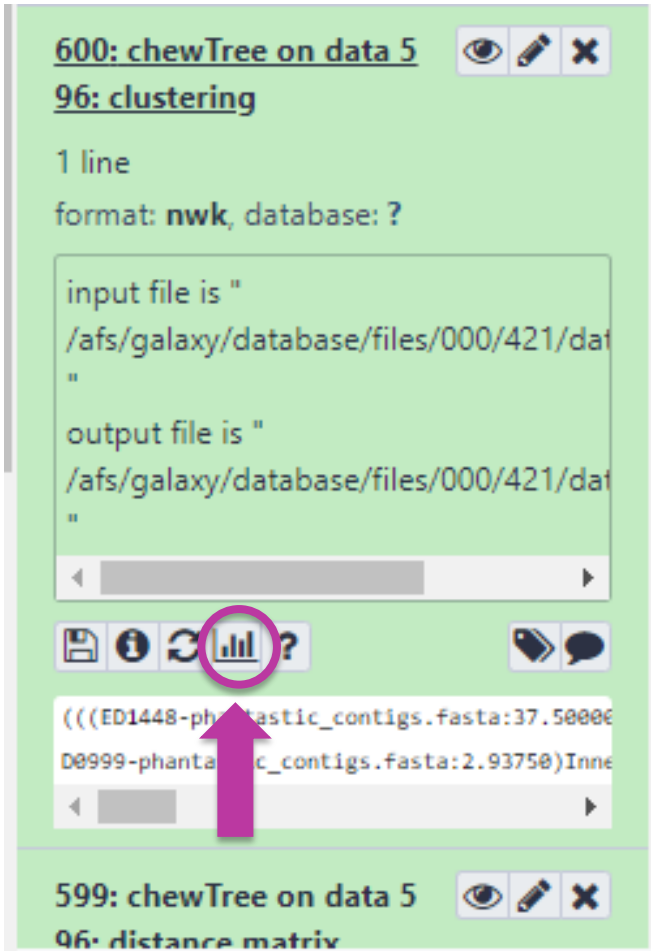
**Threshold of maximum 10 - 15 allelic differences to consider *E. coli* strains related**

# Phylogenetic tree visualization on ARIES

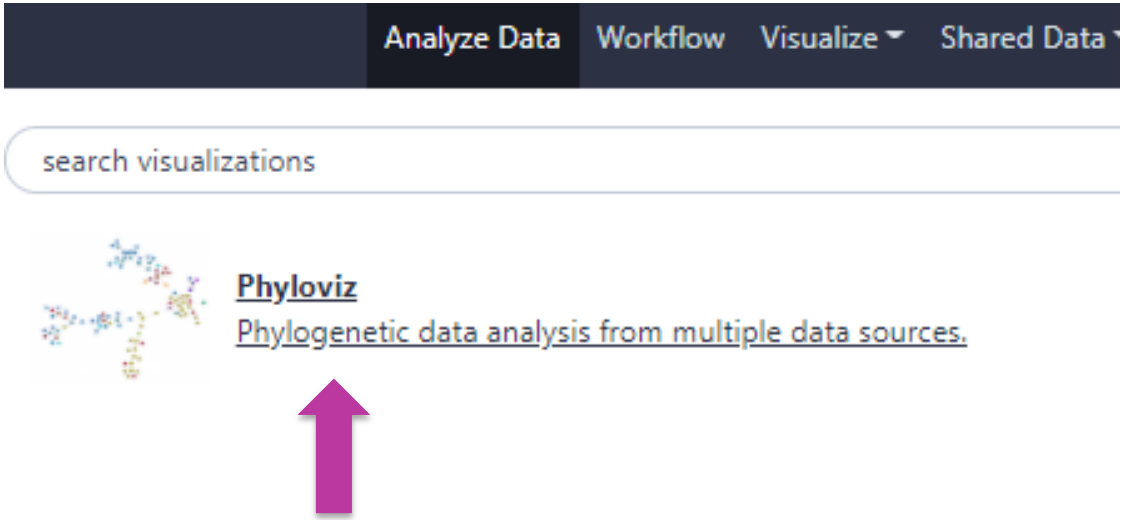
1



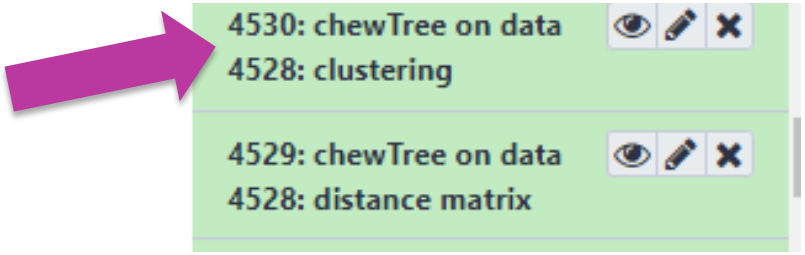

2



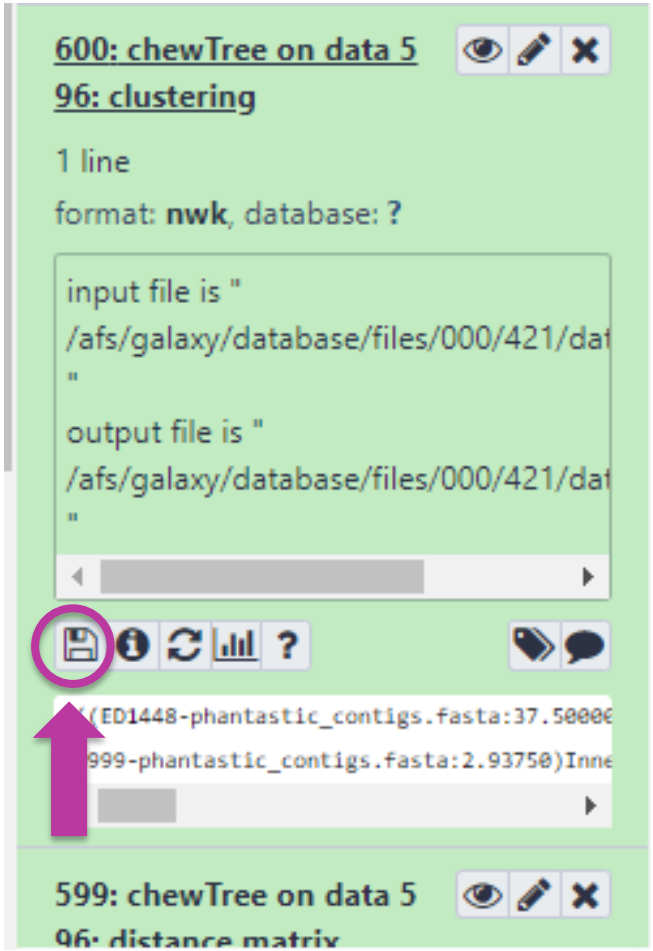

3



# 1 bis




# 2 bis




**Downloading the output of clustering in nwk format to visualize the tree outside from ARIES**