# **Supplementary materials**

# **Tables**

**Supplementary Table 1** Threshold values for each bacterial species and for each subset of genes when using the complete linkage method with the percentiles presented in Figure 4 and Supplementary Figure 3. The threshold refers to the distance between isolates of the same CT, which is defined as the proportion of genes that disagree on their allelic assignment.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Scheme** | **No. of genes** | **0.5th percentile** | **1th percentile** | **1.5th percentile** |
| *Francisella  tularensis* | 1147 | 0.002 | 0.003 | 0.004 |
| 133 | 0 | 0.008 | 0.015 |
| 93 | 0 | 0 | 0.011 |
| 53 | 0 | 0 | 0 |
| **Mean** | **3E-04** | **0.002** | **0.006** |
| **Variance** | **5E-07** | **9E-06** | **4E-05** |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Scheme** | **No. of genes** | **0.1th percentile** | **0.5th percentile** | **1th percentile** |
| *Legionella pneumophila* | 1521 | 0.001 | 0.009 | 0.020 |
| 1030 | 0.001 | 0.011 | 0.024 |
| 930 | 0.001 | 0.011 | 0.025 |
| 830 | 0.001 | 0.012 | 0.025 |
| 730 | 0.001 | 0.012 | 0.026 |
| 630 | 0.002 | 0.013 | 0.029 |
| 530 | 0.002 | 0.013 | 0.030 |
| 430 | 0 | 0.014 | 0.033 |
| 330 | 0 | 0.015 | 0.036 |
| 230 | 0 | 0.017 | 0.039 |
| 130 | 0 | 0.015 | 0.038 |
| 30 | 0 | 0 | 0.033 |
| **Mean** | **0.001** | **0.012** | **0.030** |
| **Variance** | **5E-07** | **2E-05** | **3E-05** |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Scheme** | **No. of genes** | **0.5th percentile** | **1th percentile** | **1.5th percentile** |
| *Clostridioides difficile* | 2270 | 0.004 | 0.006 | 0.007 |
| 2169 | 0.005 | 0.006 | 0.007 |
| 2069 | 0.005 | 0.006 | 0.008 |
| 1969 | 0.005 | 0.007 | 0.008 |
| 1869 | 0.005 | 0.007 | 0.009 |
| 1769 | 0.005 | 0.007 | 0.008 |
| 1669 | 0.005 | 0.008 | 0.009 |
| 1569 | 0.006 | 0.008 | 0.010 |
| 1469 | 0.006 | 0.008 | 0.010 |
| 1369 | 0.007 | 0.009 | 0.010 |
| 1269 | 0.006 | 0.009 | 0.011 |
| 1169 | 0.007 | 0.009 | 0.012 |
| 1069 | 0.007 | 0.010 | 0.012 |
| 969 | 0.007 | 0.010 | 0.013 |
| 869 | 0.008 | 0.012 | 0.014 |
| 769 | 0.009 | 0.012 | 0.016 |
| 669 | 0.009 | 0.013 | 0.016 |
| 569 | 0.011 | 0.014 | 0.018 |
| 469 | 0.011 | 0.015 | 0.019 |
| 369 | 0.014 | 0.019 | 0.022 |
| 269 | 0.015 | 0.022 | 0.026 |
| 169 | 0.018 | 0.024 | 0.030 |
| 69 | 0.014 | 0.029 | 0.043 |
| **Mean** | **0.008** | **0.012** | **0.015** |
| **Variance** | **1E-05** | **4E-05** | **7E-05** |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Scheme** | **No. of genes** | **0.1th percentile** | **0.5th percentile** | **1th percentile** |
| *Acinetobacter baumannii* | 2390 | 0.005 | 0.015 | 0.021 |
| 2347 | 0.006 | 0.015 | 0.021 |
| 2247 | 0.006 | 0.016 | 0.022 |
| 2147 | 0.006 | 0.016 | 0.022 |
| 2047 | 0.006 | 0.017 | 0.023 |
| 1947 | 0.006 | 0.017 | 0.024 |
| 1847 | 0.006 | 0.017 | 0.025 |
| 1747 | 0.006 | 0.018 | 0.026 |
| 1647 | 0.007 | 0.019 | 0.027 |
| 1547 | 0.007 | 0.019 | 0.028 |
| 1447 | 0.007 | 0.020 | 0.028 |
| 1347 | 0.007 | 0.021 | 0.030 |
| 1247 | 0.007 | 0.022 | 0.030 |
| 1147 | 0.008 | 0.022 | 0.032 |
| 1047 | 0.008 | 0.023 | 0.033 |
| 947 | 0.008 | 0.024 | 0.035 |
| 847 | 0.008 | 0.025 | 0.037 |
| 747 | 0.008 | 0.025 | 0.037 |
| 647 | 0.009 | 0.026 | 0.039 |
| 547 | 0.009 | 0.029 | 0.042 |
| 447 | 0.009 | 0.029 | 0.045 |
| 347 | 0.009 | 0.029 | 0.046 |
| 247 | 0.008 | 0.028 | 0.049 |
| 147 | 0.007 | 0.034 | 0.054 |
| 47 | 0 | 0.021 | 0.064 |
| **Mean** | **0.007** | **0.022** | **0.034** |
| **Variance** | **3E-06** | **3E-05** | **1E-04** |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Scheme** | **No. of genes** | **0.5th percentile** | **1th percentile** | **1.5th percentile** |
| *Klebsiella pneumoniae* | 2358 | 0.009 | 0.015 | 0.018 |
| 2100 | 0.010 | 0.016 | 0.020 |
| 2000 | 0.010 | 0.017 | 0.021 |
| 1900 | 0.011 | 0.017 | 0.022 |
| 1800 | 0.011 | 0.018 | 0.022 |
| 1700 | 0.011 | 0.018 | 0.023 |
| 1600 | 0.012 | 0.019 | 0.024 |
| 1500 | 0.012 | 0.020 | 0.025 |
| 1400 | 0.013 | 0.021 | 0.026 |
| 1300 | 0.013 | 0.022 | 0.027 |
| 1200 | 0.013 | 0.023 | 0.028 |
| 1100 | 0.014 | 0.023 | 0.029 |
| 1000 | 0.015 | 0.024 | 0.031 |
| 900 | 0.016 | 0.026 | 0.033 |
| 800 | 0.016 | 0.028 | 0.035 |
| 700 | 0.017 | 0.029 | 0.037 |
| 600 | 0.018 | 0.030 | 0.040 |
| 500 | 0.018 | 0.032 | 0.042 |
| 400 | 0.020 | 0.033 | 0.045 |
| 300 | 0.020 | 0.037 | 0.050 |
| 200 | 0.025 | 0.045 | 0.055 |
| 100 | 0.030 | 0.050 | 0.070 |
| **Mean** | **0.015** | **0.025** | **0.033** |
| **Variance** | **3E-05** | **8E-05** | **2E-04** |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Scheme** | **No. of genes** | **1th percentile** | **8.5th percentile** | **9th percentile** | **9.5th percentile** |
| *Enterococcus faecium* | 1423 | 0.013 | 0.077 | 0.082 | 0.085 |
| 1088 | 0.016 | 0.097 | 0.103 | 0.108 |
| 988 | 0.016 | 0.106 | 0.111 | 0.116 |
| 888 | 0.017 | 0.114 | 0.118 | 0.124 |
| 788 | 0.018 | 0.123 | 0.129 | 0.135 |
| 688 | 0.019 | 0.134 | 0.140 | 0.145 |
| 588 | 0.019 | 0.145 | 0.151 | 0.158 |
| 488 | 0.020 | 0.162 | 0.168 | 0.174 |
| 388 | 0.021 | 0.175 | 0.183 | 0.188 |
| 288 | 0.024 | 0.188 | 0.194 | 0.201 |
| 188 | 0.021 | 0.218 | 0.234 | 0.239 |
| 88 | 0.023 | 0.239 | 0.250 | 0.273 |
| **Mean** | **0.019** | **0.148** | **0.155** | **0.162** |
| **Variance** | **1E-05** | **0.002** | **0.003** | **0.003** |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Scheme** | **No. of genes** | **0.1th percentile** | **0.5th percentile** | **1th percentile** |
| *Listeria monocytogenes* | 1701 | 0.002 | 0.014 | 0.020 |
| 1692 | 0.002 | 0.014 | 0.020 |
| 1592 | 0.002 | 0.014 | 0.021 |
| 1492 | 0.002 | 0.015 | 0.022 |
| 1392 | 0.002 | 0.016 | 0.023 |
| 1292 | 0.002 | 0.016 | 0.024 |
| 1192 | 0.003 | 0.017 | 0.025 |
| 1092 | 0.003 | 0.017 | 0.026 |
| 992 | 0.002 | 0.018 | 0.027 |
| 892 | 0.002 | 0.019 | 0.028 |
| 792 | 0.003 | 0.020 | 0.029 |
| 692 | 0.003 | 0.020 | 0.032 |
| 592 | 0.003 | 0.022 | 0.032 |
| 492 | 0.002 | 0.022 | 0.035 |
| 392 | 0.003 | 0.023 | 0.036 |
| 292 | 0.003 | 0.024 | 0.038 |
| 192 | 0.005 | 0.026 | 0.042 |
| 92 | 0 | 0.022 | 0.043 |
| **Mean** | **0.002** | **0.019** | **0.029** |
| **Variance** | **1E-06** | **1E-05** | **5E-05** |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Scheme** | **No. of genes** | **0.1th percentile** | **0.5th percentile** | **1th percentile** |
| *Staphylococcus aureus* | 1861 | 0.020 | 0.032 | 0.039 |
| 1860 | 0.020 | 0.032 | 0.039 |
| 1760 | 0.022 | 0.034 | 0.040 |
| 1660 | 0.022 | 0.034 | 0.042 |
| 1560 | 0.022 | 0.036 | 0.043 |
| 1460 | 0.023 | 0.037 | 0.045 |
| 1360 | 0.024 | 0.038 | 0.046 |
| 1260 | 0.025 | 0.039 | 0.048 |
| 1160 | 0.026 | 0.041 | 0.049 |
| 1060 | 0.026 | 0.042 | 0.051 |
| 960 | 0.027 | 0.044 | 0.053 |
| 860 | 0.028 | 0.045 | 0.056 |
| 760 | 0.030 | 0.047 | 0.058 |
| 660 | 0.030 | 0.050 | 0.061 |
| 560 | 0.032 | 0.054 | 0.064 |
| 460 | 0.033 | 0.057 | 0.070 |
| 360 | 0.033 | 0.058 | 0.072 |
| 260 | 0.035 | 0.062 | 0.077 |
| 160 | 0.038 | 0.069 | 0.088 |
| 60 | 0.033 | 0.083 | 0.100 |
| **Mean** | **0.028** | **0.047** | **0.057** |
| **Variance** | **3E-05** | **2E-04** | **3E-04** |

**Supplementary Table 2** Threshold values for each bacterial species and for each subset of genes when using the single linkage method with the percentiles presented in Supplementary Figure 4. The threshold refers to the distance between isolates of the same CT, which is defined as the proportion of genes that disagree on their allelic assignment.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Scheme** | **No. of genes** | **0.1th percentile** | **0.5th percentile** | **1th percentile** |
| *Francisella  tularensis* | 1147 | 0 | 0.002 | 0.003 |
| 133 | 0 | 0 | 0.008 |
| 93 | 0 | 0 | 0 |
| 53 | 0 | 0 | 0 |
| **Mean** | **0** | **3E-04** | **0.002** |
| **Variance** | **0** | **5E-07** | **9E-06** |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Scheme** | **No. of genes** | **0.1th percentile** | **0.5th percentile** | **1th percentile** |
| *Legionella pneumophila* | 1521 | 0.001 | 0.009 | 0.020 |
| 1030 | 0.001 | 0.011 | 0.024 |
| 930 | 0.001 | 0.011 | 0.025 |
| 830 | 0.001 | 0.012 | 0.025 |
| 730 | 0.001 | 0.012 | 0.026 |
| 630 | 0.002 | 0.013 | 0.029 |
| 530 | 0.002 | 0.013 | 0.030 |
| 430 | 0 | 0.014 | 0.033 |
| 330 | 0 | 0.015 | 0.036 |
| 230 | 0 | 0.017 | 0.039 |
| 130 | 0 | 0.015 | 0.038 |
| 30 | 0 | 0 | 0.033 |
| **Mean** | **0.001** | **0.012** | **0.030** |
| **Variance** | **5E-07** | **2E-05** | **3E-05** |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Scheme** | **No. of genes** | **0.01th percentile** | **0.05th percentile** | **0.1th percentile** |
| *Clostridioides difficile* | 2270 | 4E-04 | 0.002 | 0.003 |
| 2169 | 5E-04 | 0.002 | 0.002 |
| 2069 | 5E-04 | 0.001 | 0.002 |
| 1969 | 0.001 | 0.002 | 0.003 |
| 1869 | 0.001 | 0.002 | 0.003 |
| 1769 | 0.001 | 0.002 | 0.003 |
| 1669 | 0.001 | 0.002 | 0.003 |
| 1569 | 0.001 | 0.002 | 0.003 |
| 1469 | 0.001 | 0.002 | 0.003 |
| 1369 | 0.001 | 0.002 | 0.003 |
| 1269 | 0.001 | 0.002 | 0.003 |
| 1169 | 0.001 | 0.003 | 0.003 |
| 1069 | 0.001 | 0.002 | 0.004 |
| 969 | 0.001 | 0.002 | 0.004 |
| 869 | 0 | 0.002 | 0.003 |
| 769 | 0 | 0.003 | 0.004 |
| 669 | 0 | 0.003 | 0.004 |
| 569 | 0 | 0.004 | 0.004 |
| 469 | 0 | 0.002 | 0.004 |
| 369 | 0 | 0.003 | 0.005 |
| 269 | 0 | 0.004 | 0.007 |
| 169 | 0 | 0.006 | 0.006 |
| 69 | 0 | 0 | 0 |
| **Mean** | **4E-04** | **0.002** | **0.003** |
| **Variance** | **1E-07** | **1E-06** | **2E-06** |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Scheme** | **No. of genes** | **0.01th percentile** | **0.05th percentile** | **0.1th percentile** |
| *Acinetobacter baumannii* | 2390 | 0 | 0.003 | 0.005 |
| 2347 | 0 | 0.003 | 0.006 |
| 2247 | 0 | 0.003 | 0.006 |
| 2147 | 0 | 0.003 | 0.006 |
| 2047 | 0 | 0.003 | 0.006 |
| 1947 | 0 | 0.003 | 0.006 |
| 1847 | 0 | 0.003 | 0.006 |
| 1747 | 0 | 0.003 | 0.006 |
| 1647 | 0 | 0.003 | 0.007 |
| 1547 | 0 | 0.003 | 0.007 |
| 1447 | 0 | 0.003 | 0.007 |
| 1347 | 0 | 0.004 | 0.007 |
| 1247 | 0 | 0.003 | 0.007 |
| 1147 | 0 | 0.003 | 0.008 |
| 1047 | 0 | 0.004 | 0.008 |
| 947 | 0 | 0.004 | 0.008 |
| 847 | 0 | 0.004 | 0.008 |
| 747 | 0 | 0.004 | 0.008 |
| 647 | 0 | 0.005 | 0.009 |
| 547 | 0 | 0.004 | 0.009 |
| 447 | 0 | 0.004 | 0.009 |
| 347 | 0 | 0.003 | 0.009 |
| 247 | 0 | 0.004 | 0.008 |
| 147 | 0 | 0 | 0.007 |
| 47 | 0 | 0 | 0 |
| **Mean** | **0** | **0.003** | **0.007** |
| **Variance** | **0** | **1E-06** | **3E-06** |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Scheme** | **No. of genes** | **0.1th percentile** | **0.5th percentile** | **1th percentile** |
| *Klebsiella pneumoniae* | 2358 | 0.002 | 0.009 | 0.015 |
| 2100 | 0.002 | 0.010 | 0.016 |
| 2000 | 0.002 | 0.010 | 0.017 |
| 1900 | 0.002 | 0.011 | 0.017 |
| 1800 | 0.002 | 0.011 | 0.018 |
| 1700 | 0.002 | 0.011 | 0.018 |
| 1600 | 0.003 | 0.012 | 0.019 |
| 1500 | 0.003 | 0.012 | 0.020 |
| 1400 | 0.002 | 0.013 | 0.021 |
| 1300 | 0.002 | 0.013 | 0.022 |
| 1200 | 0.003 | 0.013 | 0.023 |
| 1100 | 0.003 | 0.014 | 0.023 |
| 1000 | 0.003 | 0.015 | 0.024 |
| 900 | 0.003 | 0.016 | 0.026 |
| 800 | 0.003 | 0.016 | 0.028 |
| 700 | 0.003 | 0.017 | 0.029 |
| 600 | 0.003 | 0.018 | 0.030 |
| 500 | 0.004 | 0.018 | 0.032 |
| 400 | 0.005 | 0.020 | 0.033 |
| 300 | 0.003 | 0.020 | 0.037 |
| 200 | 0.005 | 0.025 | 0.045 |
| 100 | 0 | 0.030 | 0.050 |
| **Mean** | **0.003** | **0.015** | **0.025** |
| **Variance** | **1E-06** | **3E-05** | **8E-05** |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Scheme** | **No. of genes** | **0.5th percentile** | **1th percentile** | **1.5th percentile** |
| *Enterococcus faecium* | 1423 | 0.008 | 0.013 | 0.017 |
| 1088 | 0.010 | 0.016 | 0.019 |
| 988 | 0.010 | 0.016 | 0.020 |
| 888 | 0.011 | 0.017 | 0.021 |
| 788 | 0.011 | 0.018 | 0.023 |
| 688 | 0.012 | 0.019 | 0.023 |
| 588 | 0.012 | 0.019 | 0.024 |
| 488 | 0.012 | 0.020 | 0.027 |
| 388 | 0.013 | 0.021 | 0.028 |
| 288 | 0.014 | 0.024 | 0.031 |
| 188 | 0.011 | 0.021 | 0.032 |
| 88 | 0.011 | 0.023 | 0.023 |
| **Mean** | **0.011** | **0.019** | **0.024** |
| **Variance** | **2E-06** | **1E-05** | **2E-05** |

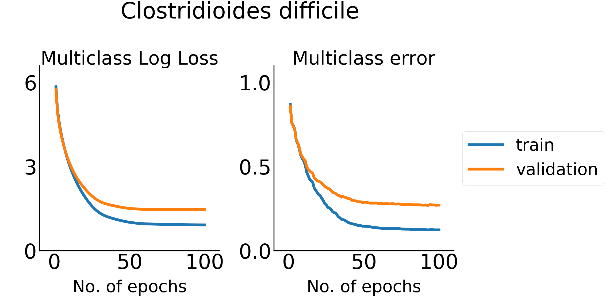
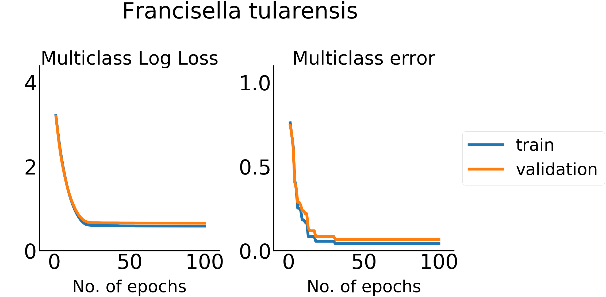
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Scheme** | **No. of genes** | **0.05th percentile** | **0.1th percentile** | **0.5th percentile** |
| *Listeria monocytogenes* | 1701 | 0.001 | 0.002 | 0.014 |
| 1692 | 0.001 | 0.002 | 0.014 |
| 1592 | 0.001 | 0.002 | 0.014 |
| 1492 | 0.001 | 0.002 | 0.015 |
| 1392 | 0.001 | 0.002 | 0.016 |
| 1292 | 0.001 | 0.002 | 0.016 |
| 1192 | 0.001 | 0.003 | 0.017 |
| 1092 | 0.001 | 0.003 | 0.017 |
| 992 | 0.001 | 0.002 | 0.018 |
| 892 | 0.001 | 0.002 | 0.019 |
| 792 | 0.001 | 0.003 | 0.020 |
| 692 | 0.001 | 0.003 | 0.020 |
| 592 | 0.002 | 0.003 | 0.022 |
| 492 | 0 | 0.002 | 0.022 |
| 392 | 0 | 0.003 | 0.023 |
| 292 | 0 | 0.003 | 0.024 |
| 192 | 0 | 0.005 | 0.026 |
| 92 | 0 | 0 | 0.022 |
| **Mean** | **7E-04** | **0.002** | **0.019** |
| **Variance** | **3E-07** | **1E-06** | **1E-05** |

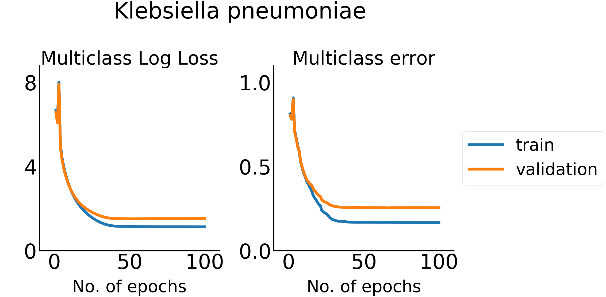
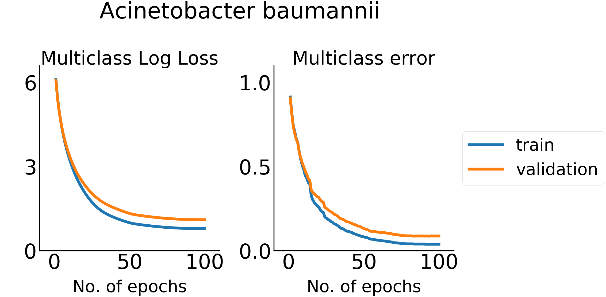
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Scheme** | **No. of genes** | **0.005th percentile** | **0.01th percentile** | **0.05th percentile** |
| *Staphylococcus aureus* | 1861 | 0.001 | 0.004 | 0.017 |
| 1860 | 0.001 | 0.004 | 0.017 |
| 1760 | 0.001 | 0.004 | 0.017 |
| 1660 | 0.001 | 0.004 | 0.017 |
| 1560 | 0.001 | 0.004 | 0.018 |
| 1460 | 0.001 | 0.004 | 0.018 |
| 1360 | 0.001 | 0.004 | 0.019 |
| 1260 | 0.002 | 0.005 | 0.020 |
| 1160 | 0.001 | 0.005 | 0.020 |
| 1060 | 0.001 | 0.005 | 0.021 |
| 960 | 0.001 | 0.005 | 0.022 |
| 860 | 0.001 | 0.006 | 0.022 |
| 760 | 0.001 | 0.005 | 0.024 |
| 660 | 0.002 | 0.006 | 0.024 |
| 560 | 0.002 | 0.007 | 0.025 |
| 460 | 0.002 | 0.007 | 0.026 |
| 360 | 0.003 | 0.008 | 0.028 |
| 260 | 0 | 0.008 | 0.027 |
| 160 | 0 | 0.006 | 0.025 |
| 60 | 0 | 0 | 0.017 |
| **Mean** | **0.001** | **0.005** | **0.021** |
| **Variance** | **4E-07** | **3E-06** | **1E-05** |

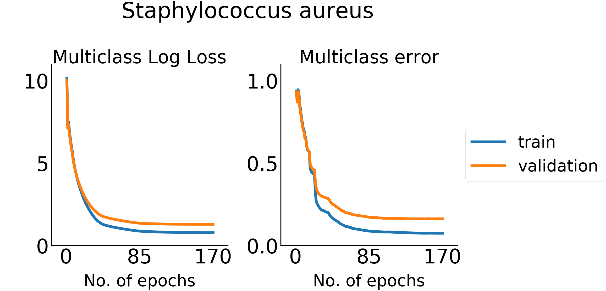
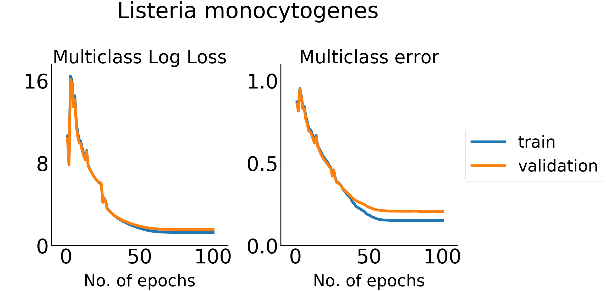
**Supplementary Table 3** Percentage of 1/2/3/4-size clusters (CTs with 1,2,3 and 4 related isolates respectively)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Scheme** | **Singleton CTs (%)** | | **2-size CTs (%)** | **3-size CTs (%)** | **4-size CTs (%)** |
| *Francisella tularensis* | | 75.9 | 10.3 | 6.9 | 2.1 |
| *Legionella pneumophila* | | 60.4 | 18.5 | 4.2 | 4.8 |
| *Clostridioides difficile* | | 74.4 | 15.6 | 3.6 | 1.9 |
| *Acinetobacter baumannii* | | 52.6 | 30.0 | 5.9 | 3.4 |
| *Klebsiella pneumoniae* | | 65.9 | 17.6 | 5.0 | 3.3 |
| *Enterococcus faecium* | | 61.5 | 17.8 | 5.5 | 3.2 |
| *Listeria monocytogenes* | | 64.8 | 16.1 | 6.7 | 3.7 |
| *Staphylococcus aureus* | | 70.4 | 14.2 | 5.5 | 3.1 |

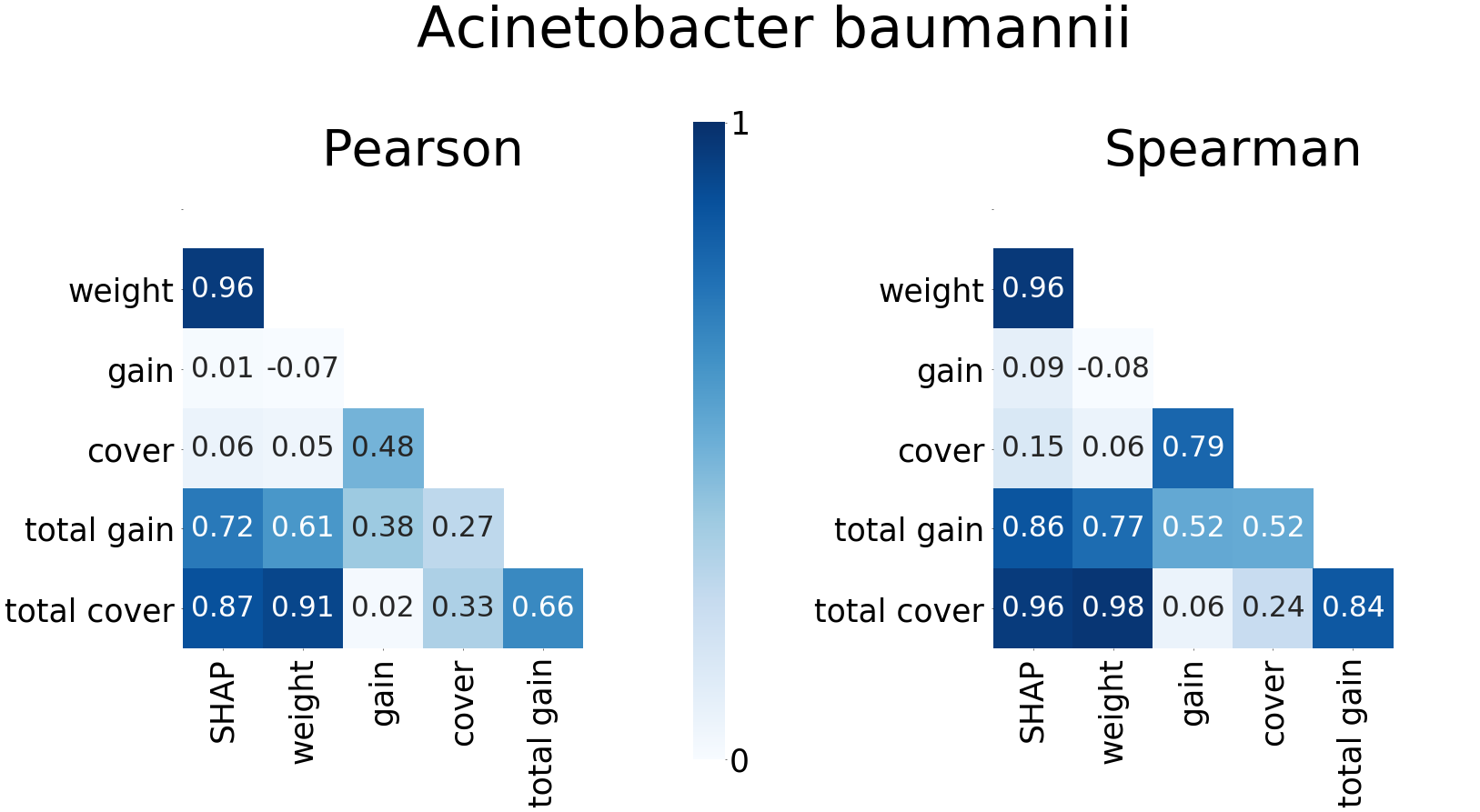
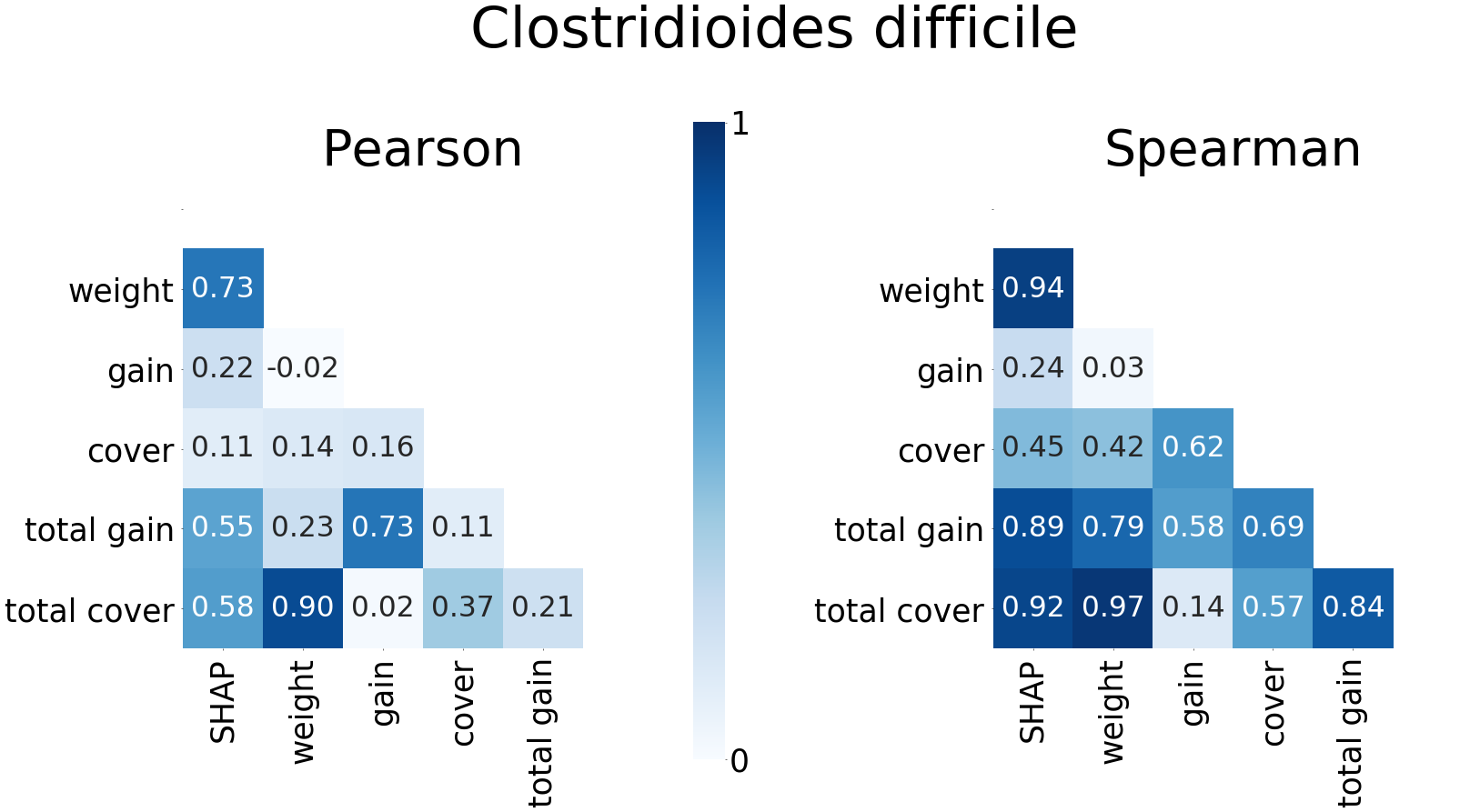
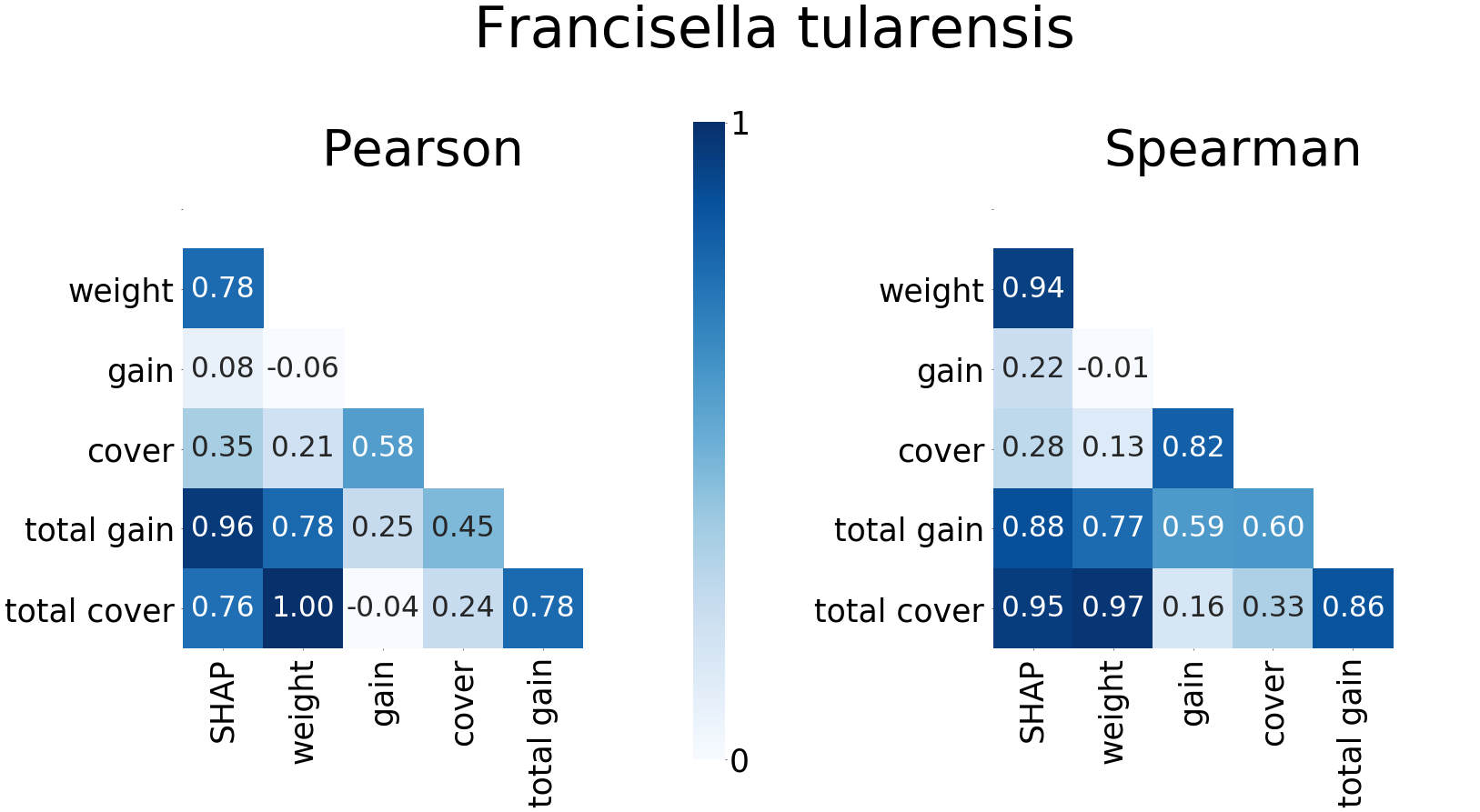
# **Figures**

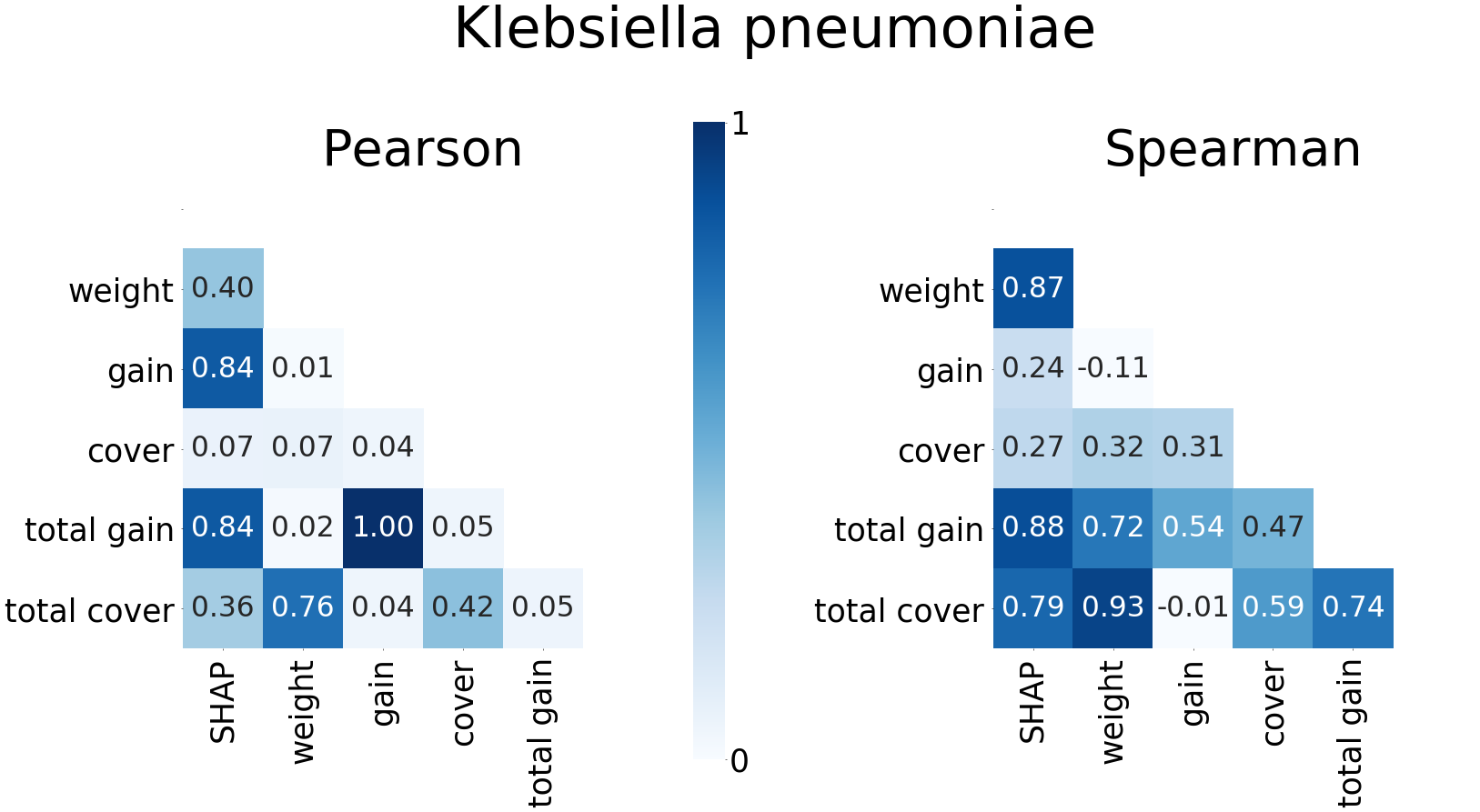


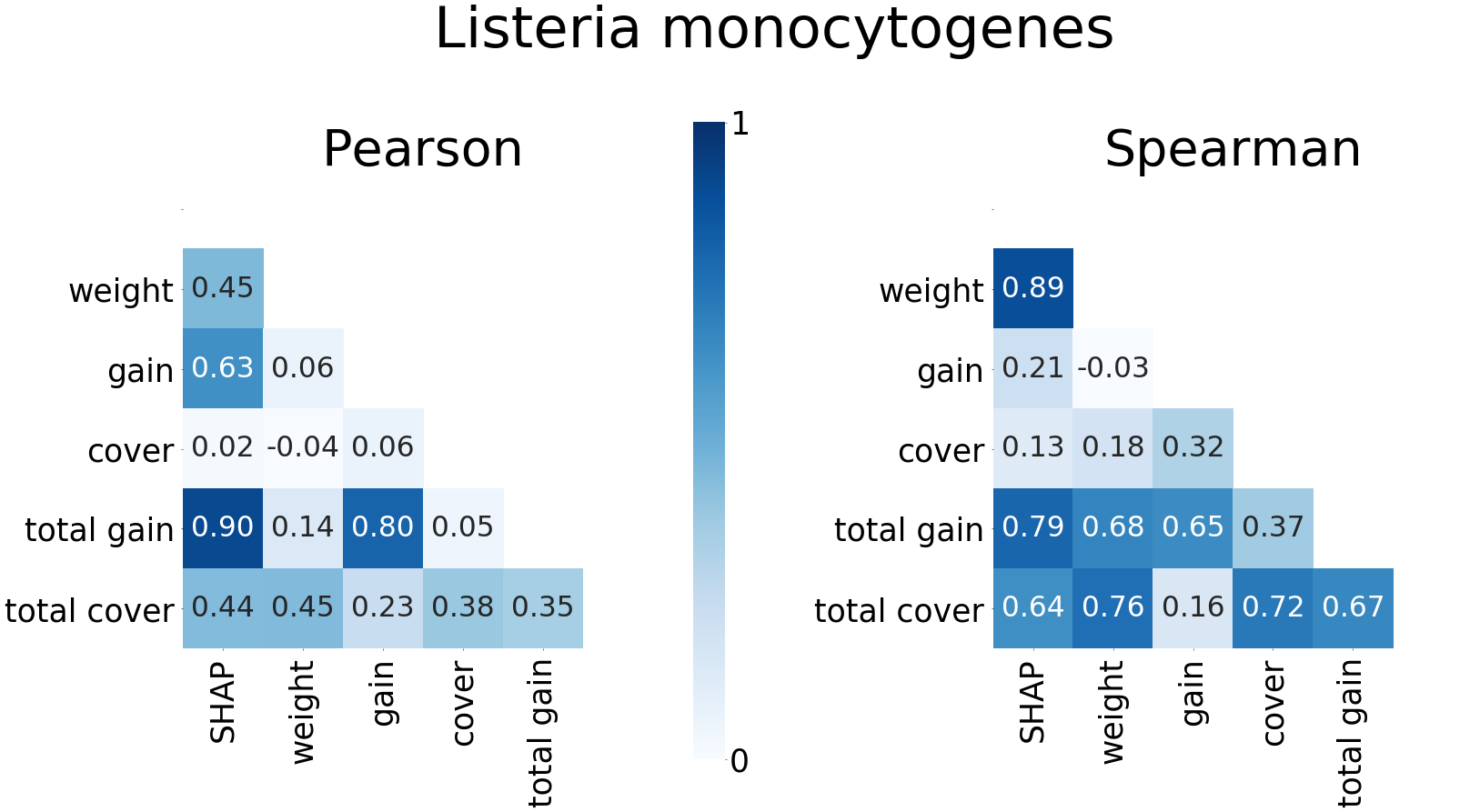


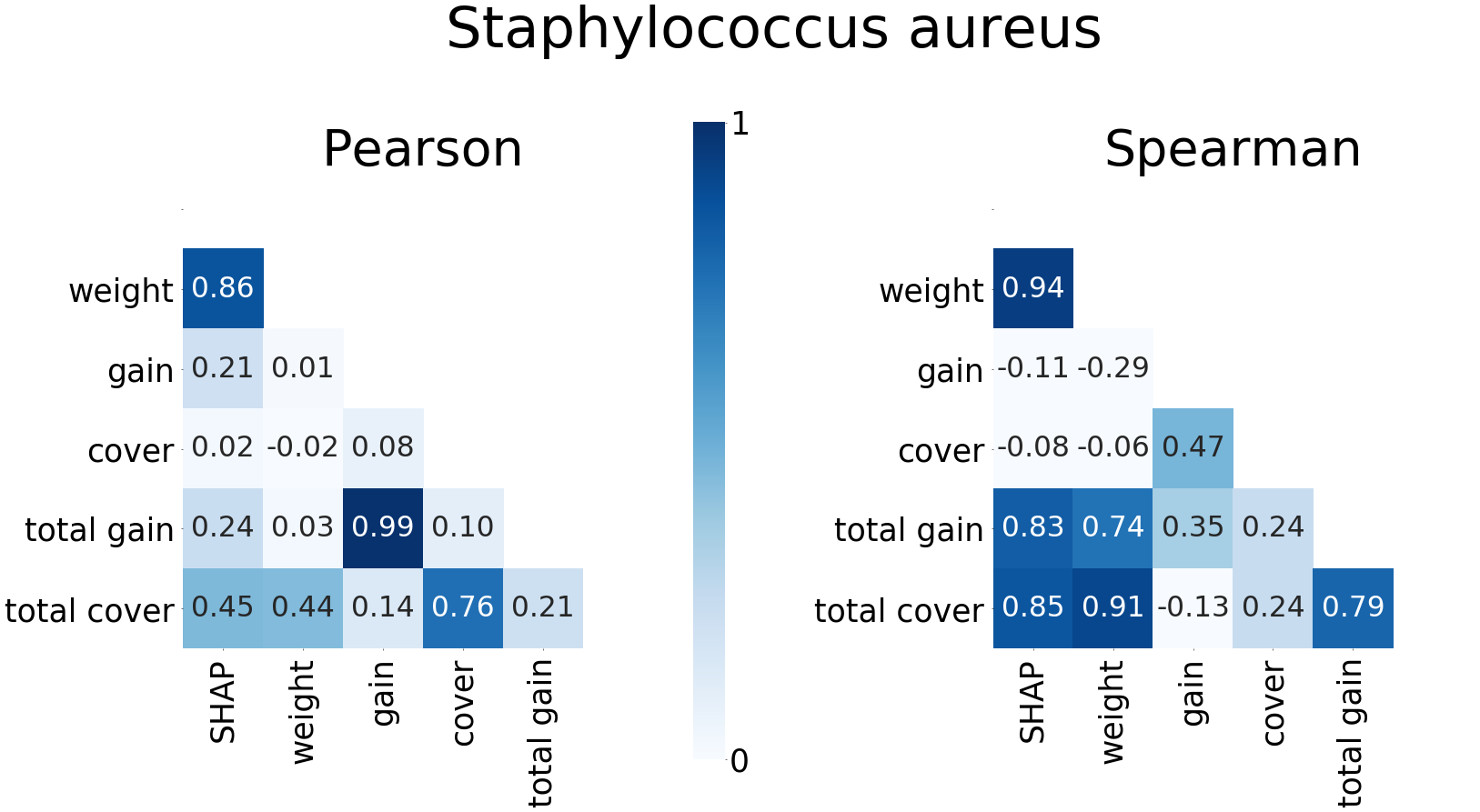


**Supplementary Fig. 1** Metrics of the XGBoost model along the training process. Multiclass Log Loss (objective function) and multiclass error values on the train and validation sets are presented as a function of the training epochs.

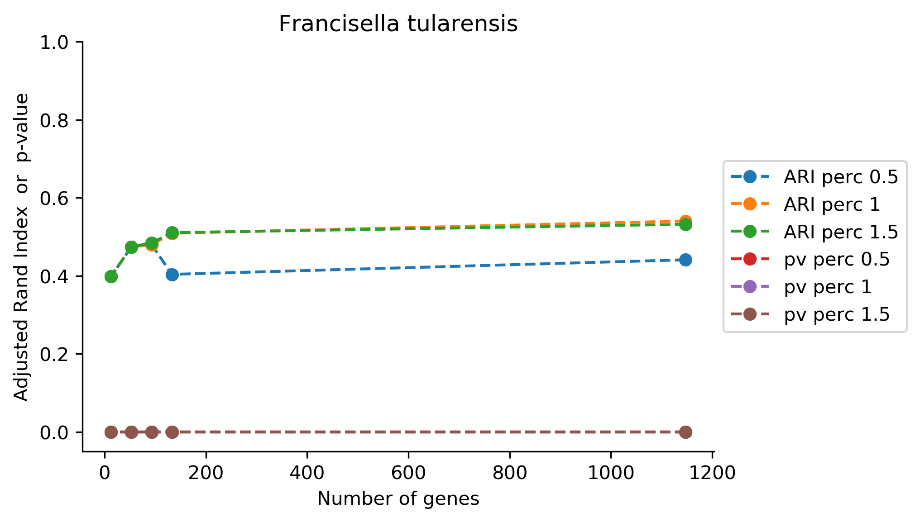


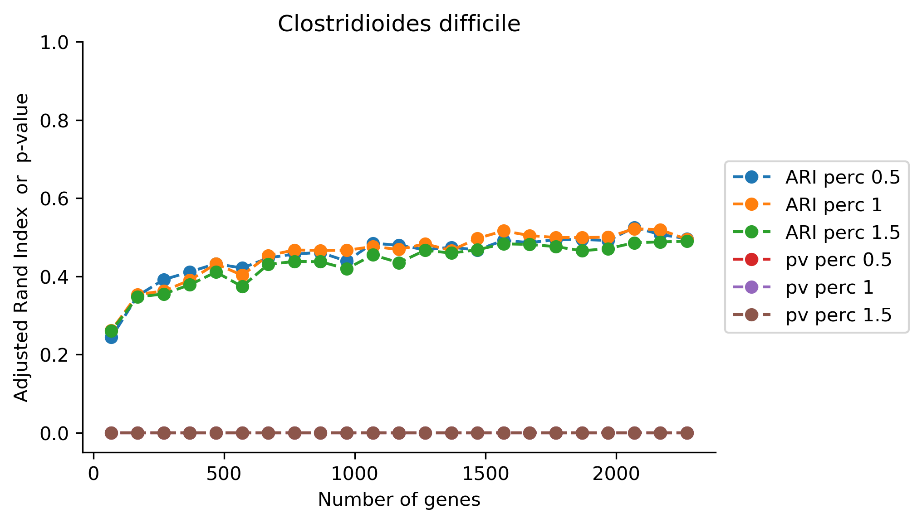


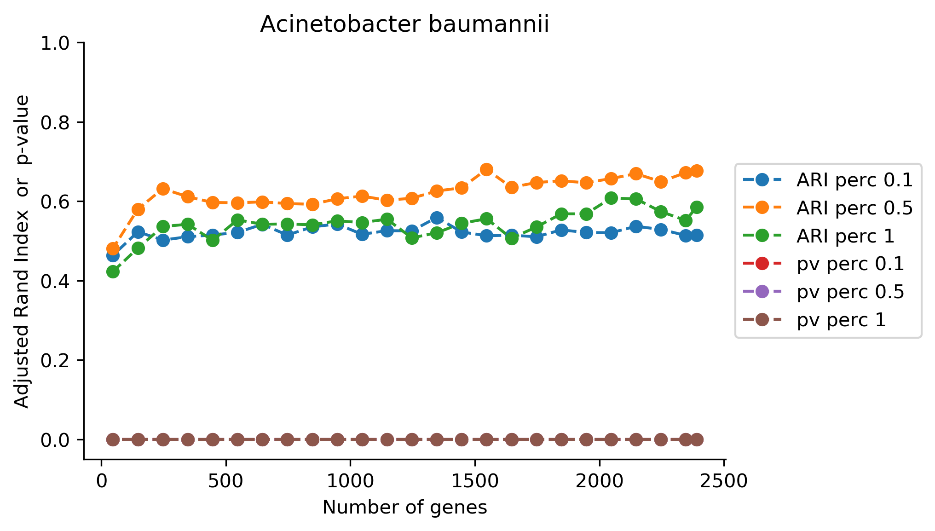


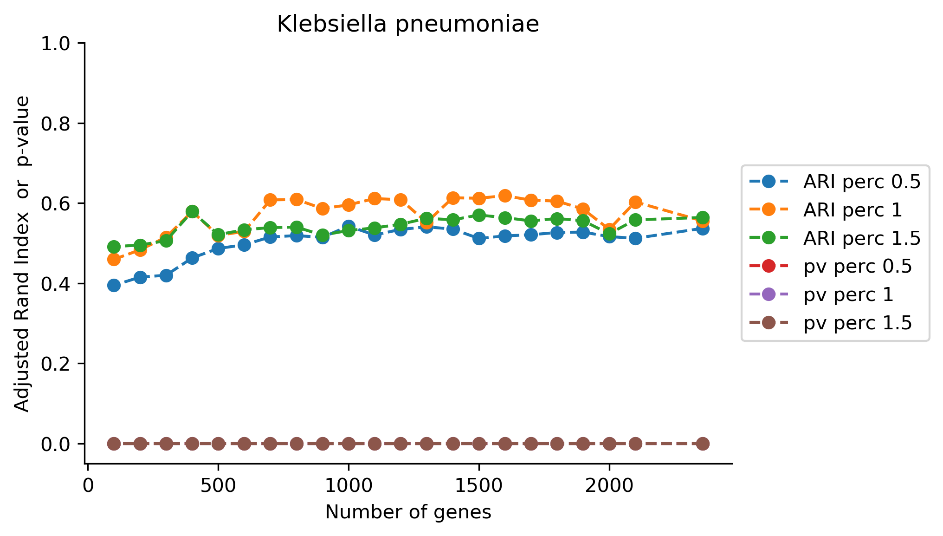


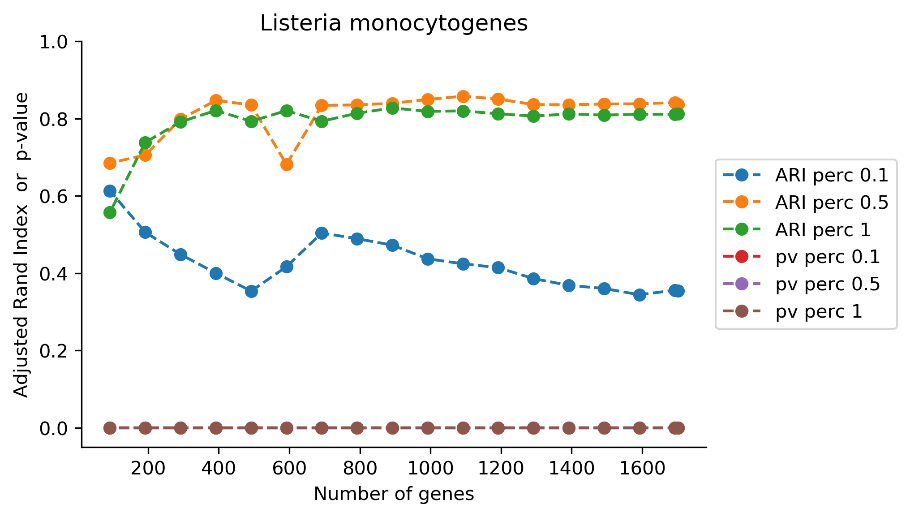
**Supplementary Fig. 2** Pearson’s and Spearman’s correlations between gene importance values computed by six different measures. The measures are the mean magnitude of the SHAP values, weight, average gain, average cover, total gain, and total cover.

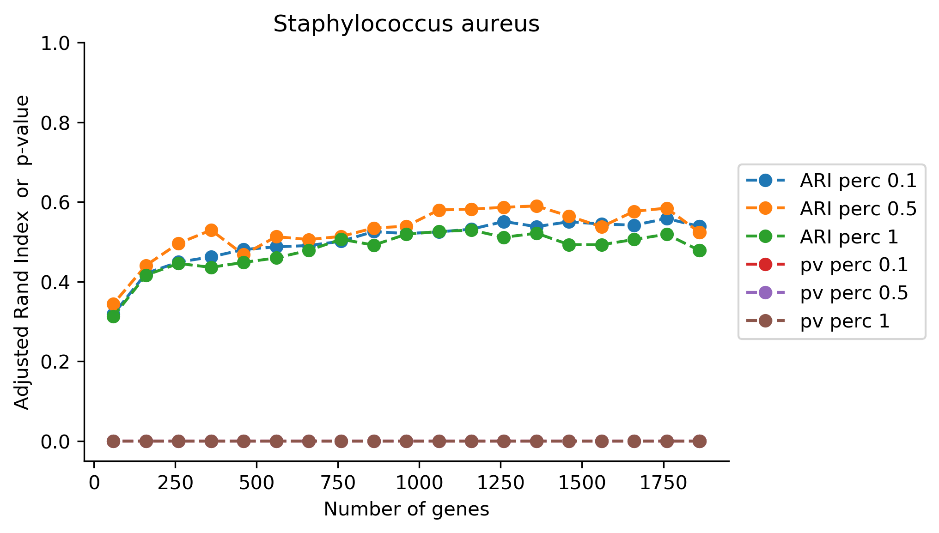




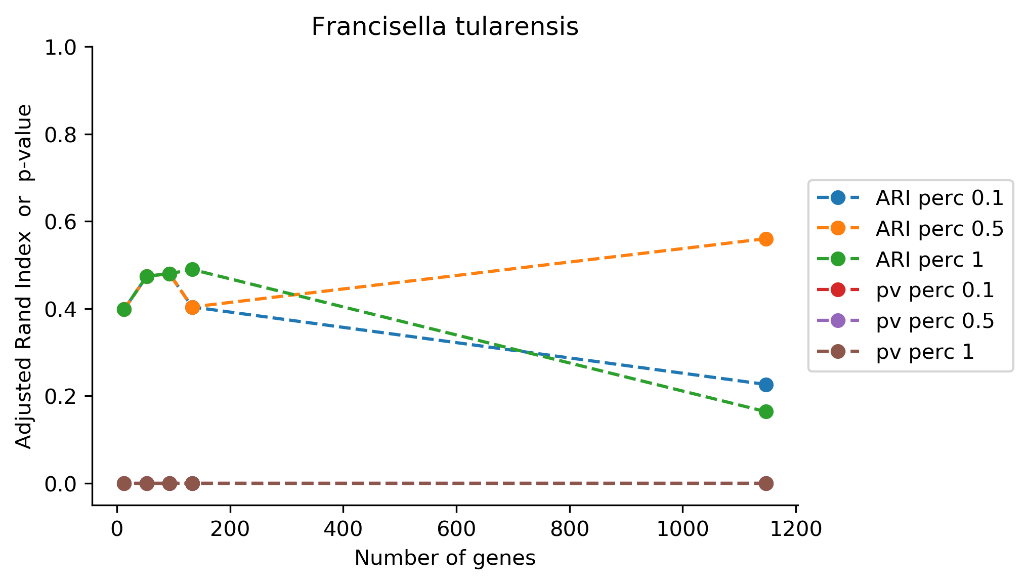


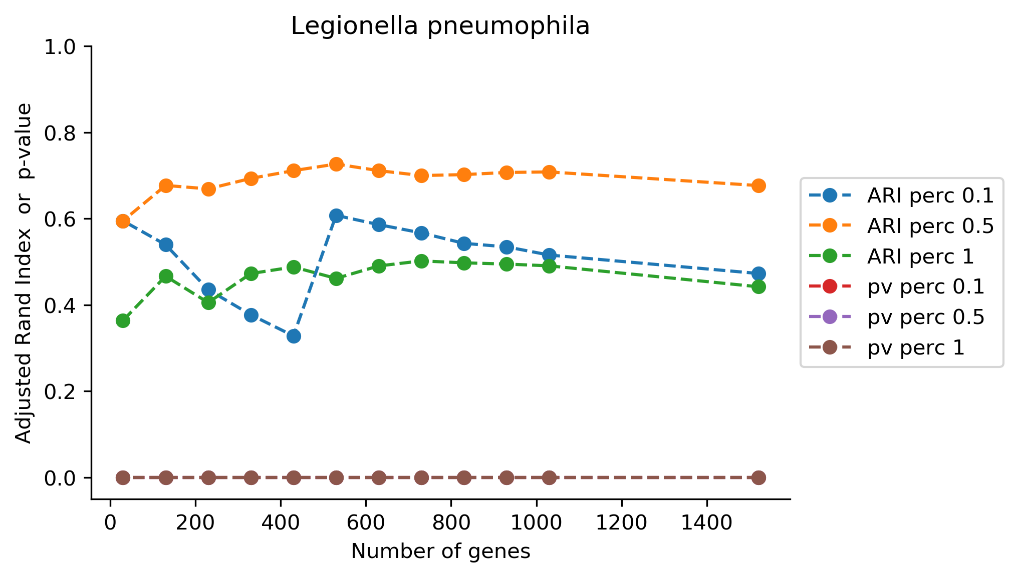


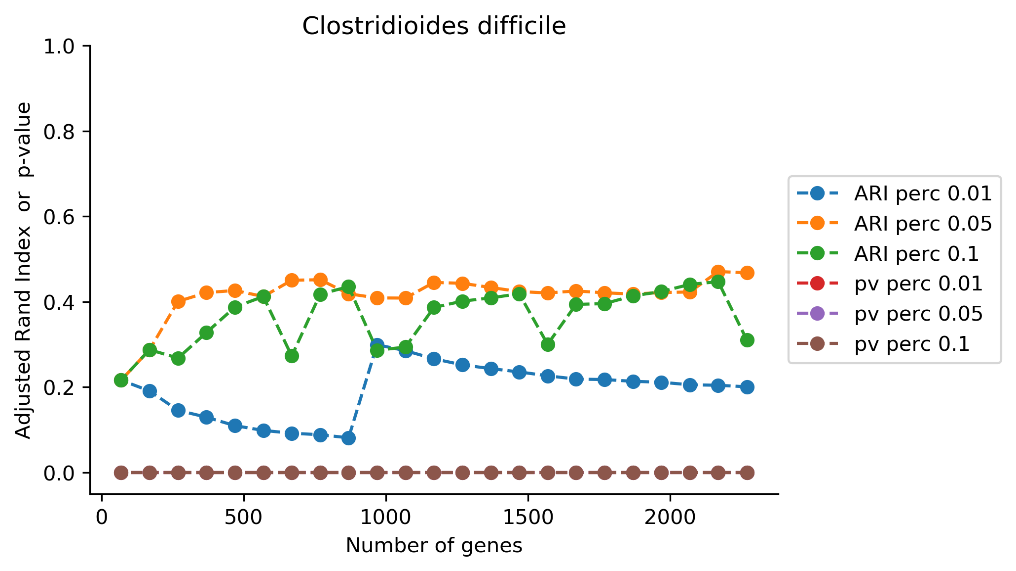


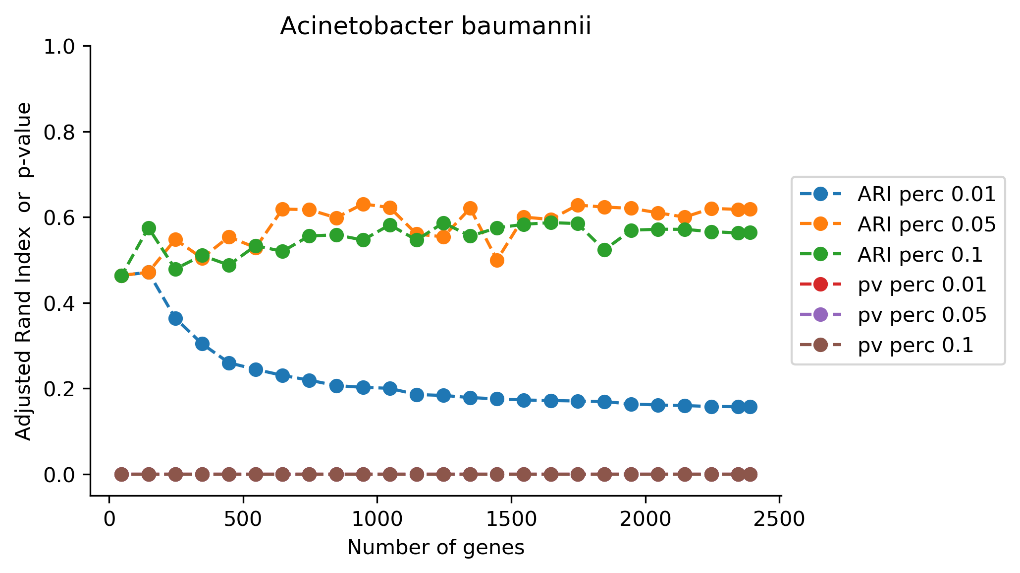


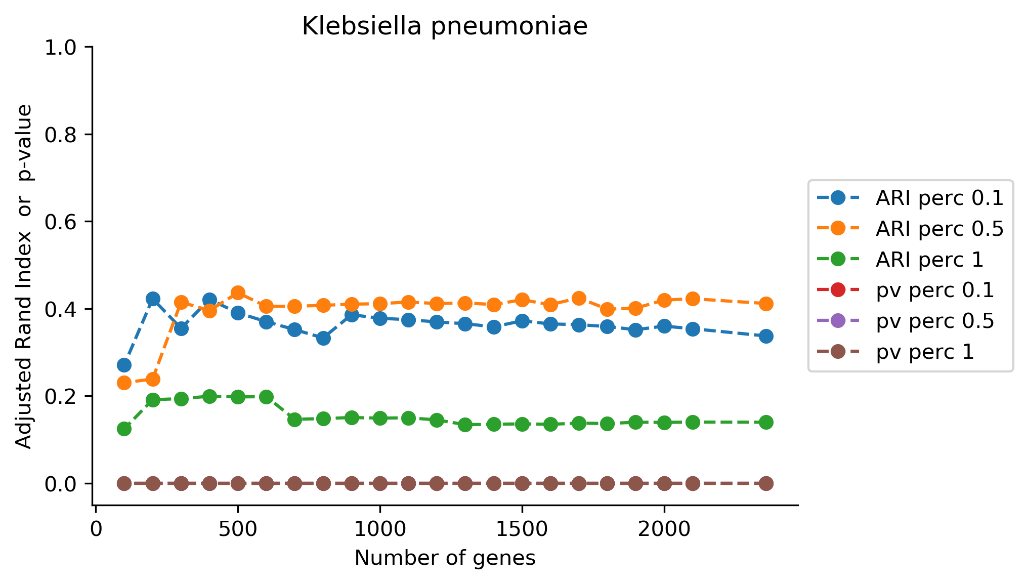
**Supplementary Fig. 3** Adjusted Rand Index and p-value computed for each subset of most informative genes when using the complete linkage method. We present the results obtained with the best percentile and its two adjacent percentiles (predecessor and successor) in the search space of*Algorithm 1*. The search space included the following percentiles of distances’ distribution: [0.005, 0.01, 0.05, 0.1, 0.5, 1, 1.5, 2, … , 10].

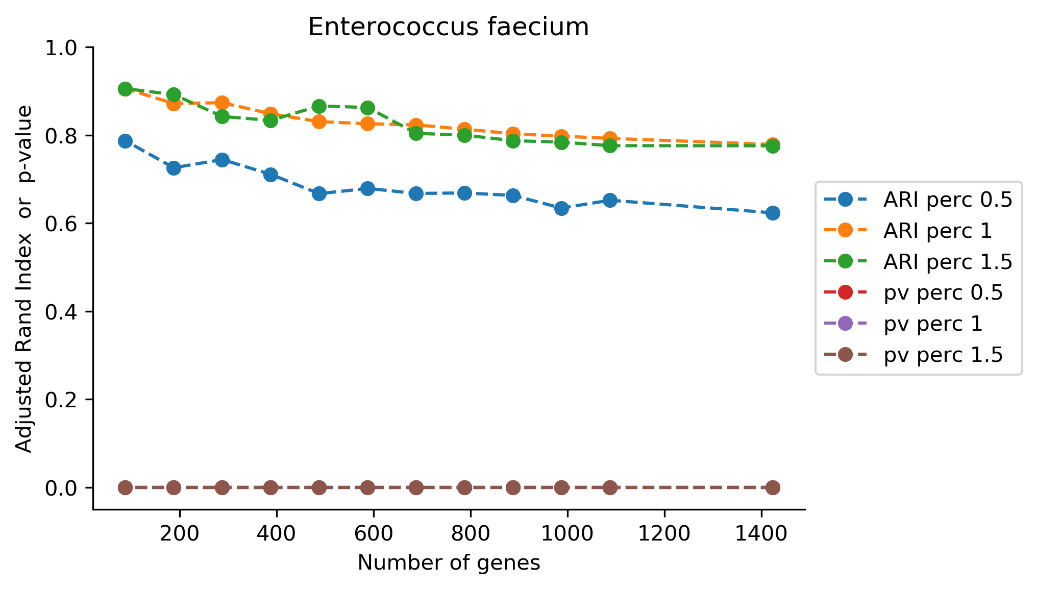


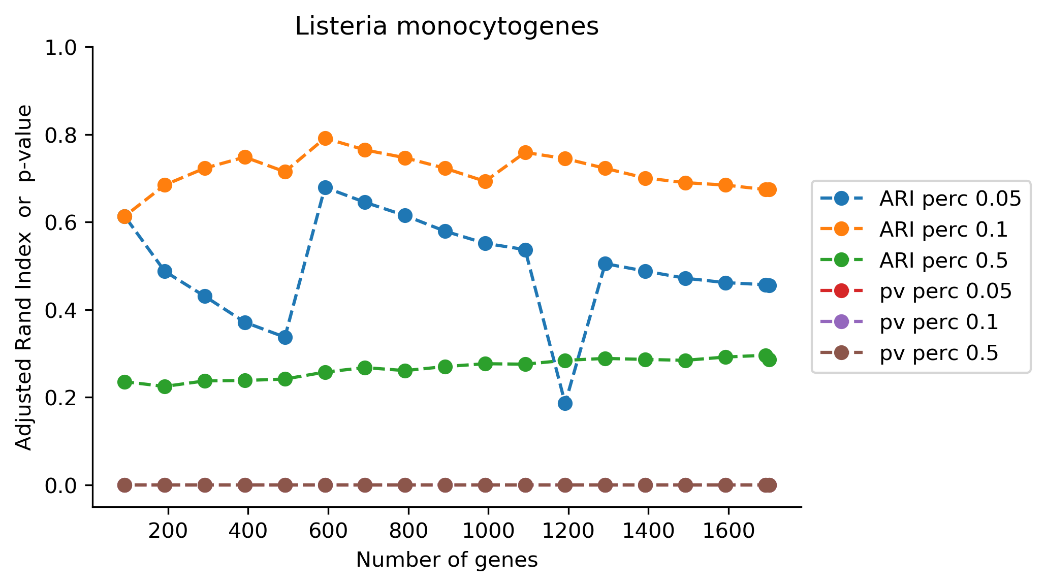


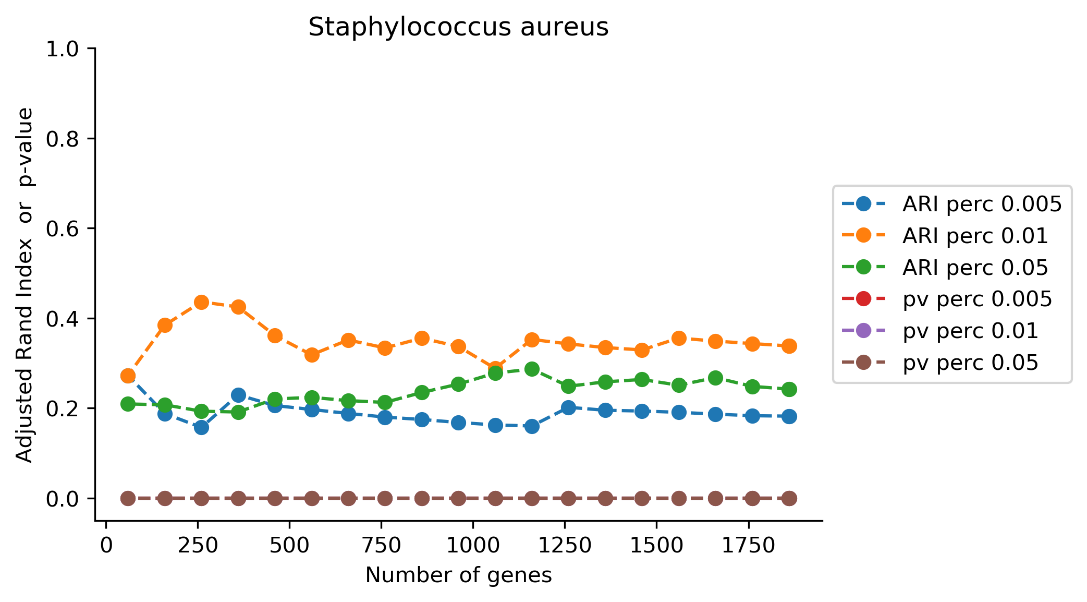




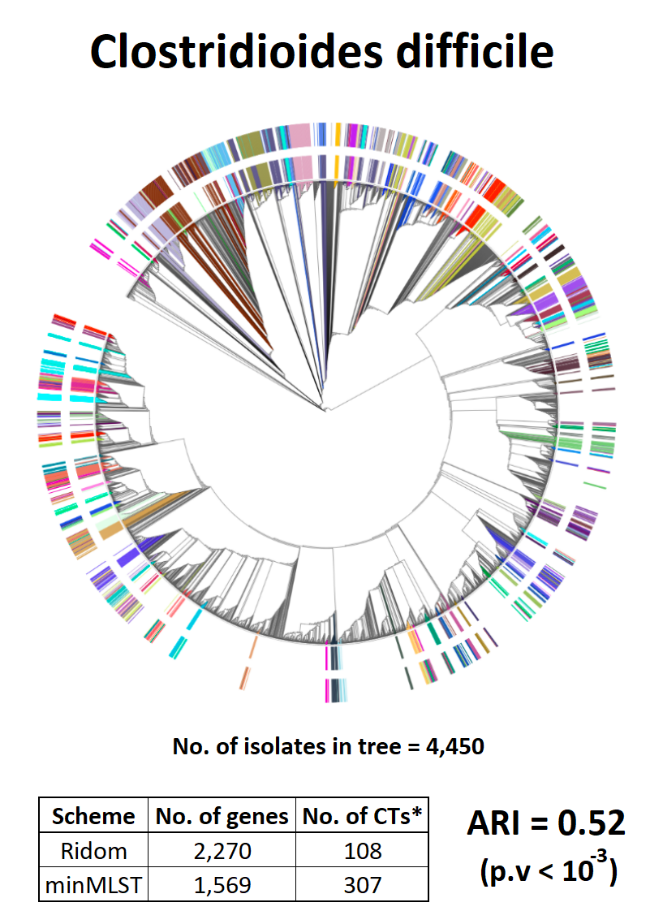
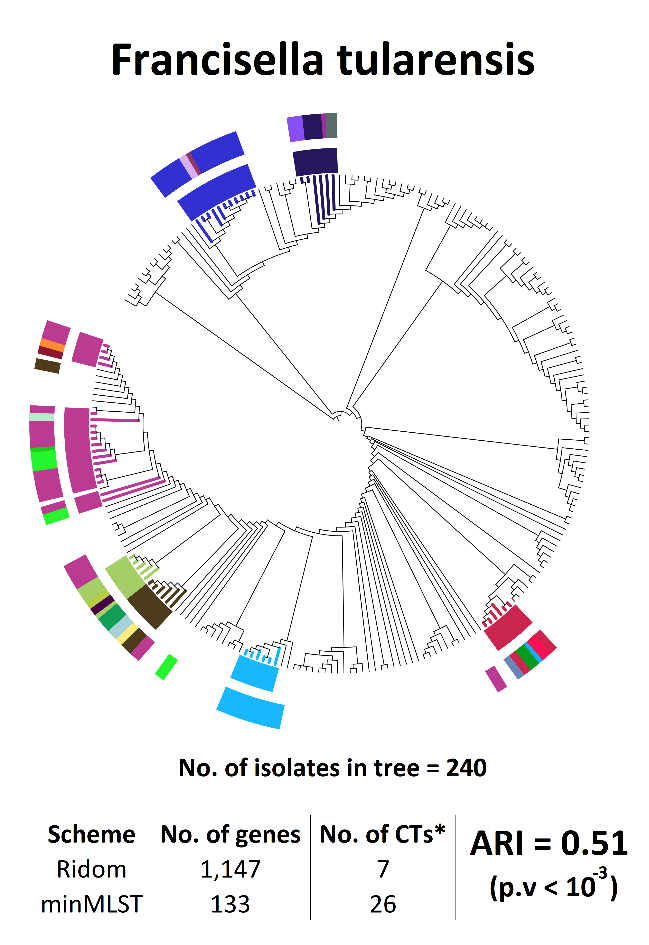


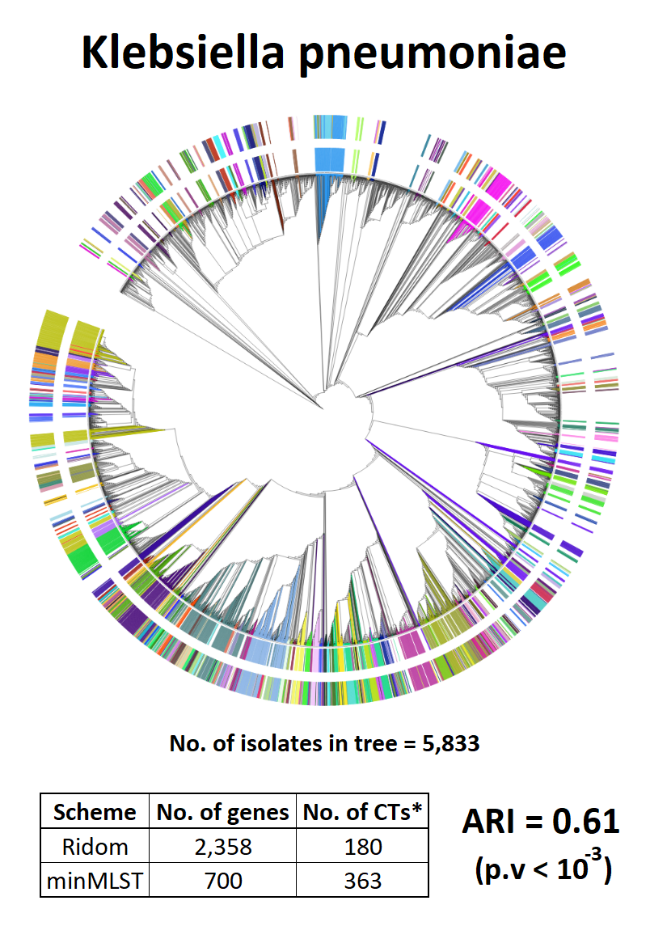
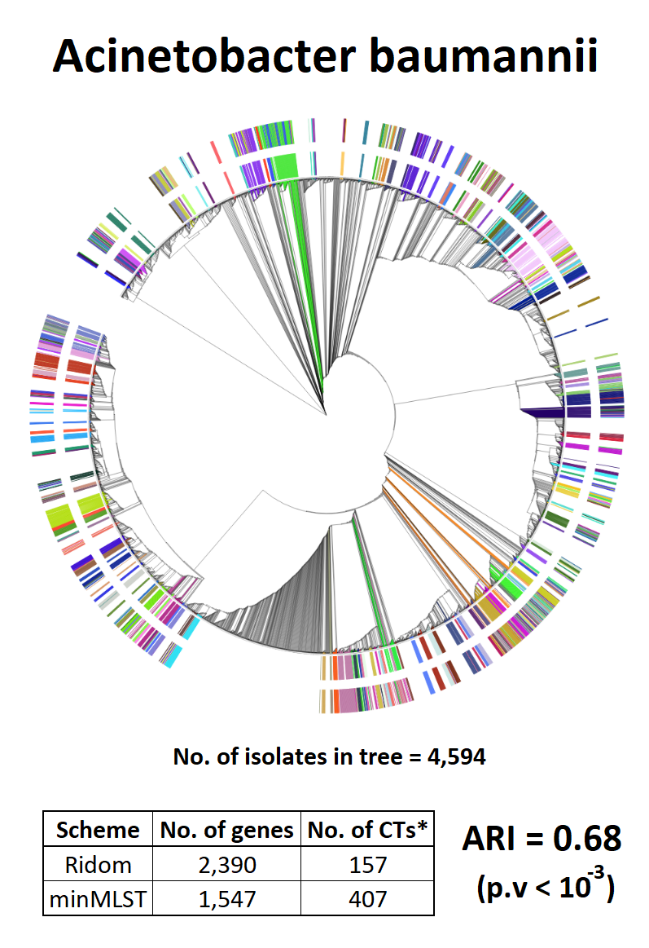




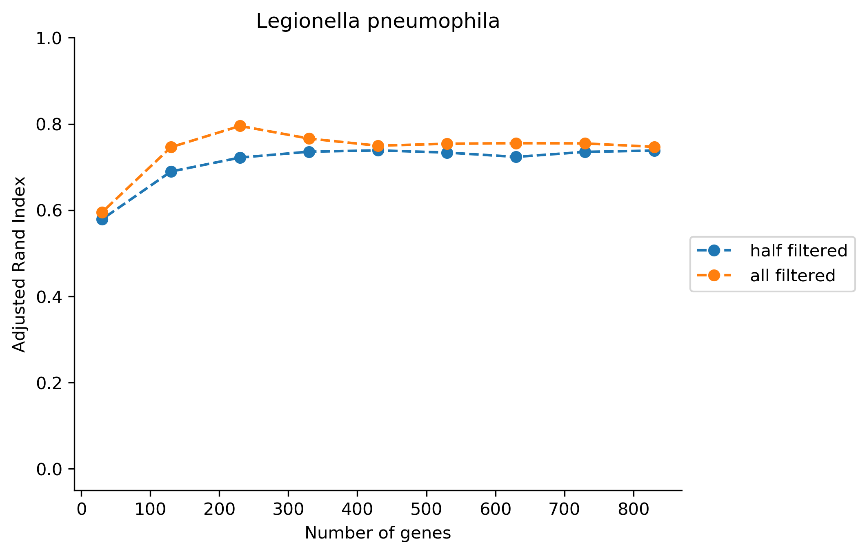


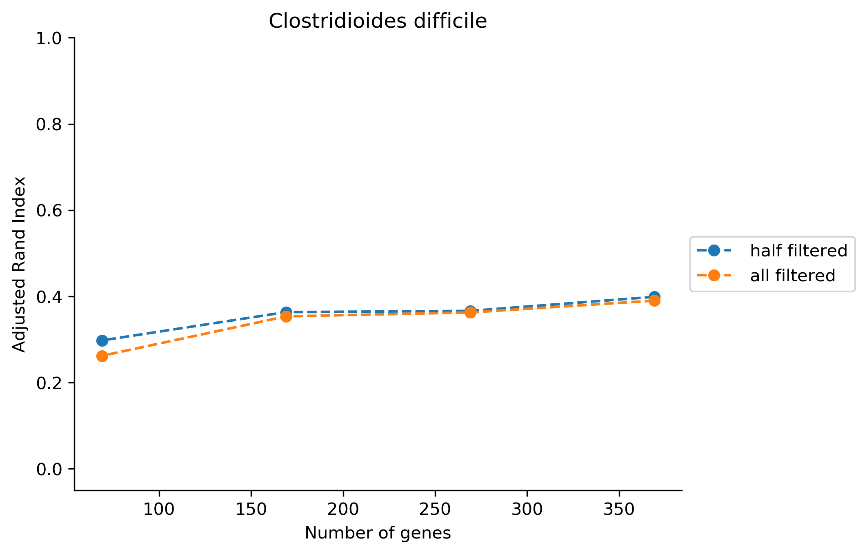
**Supplementary Fig. 4** Adjusted Rand Index and p-value computed for each subset of most informative genes when using the single linkage method. We present the results obtained with the best percentile and its two adjacent percentiles (predecessor and successor) in the search space of*Algorithm 1*. The search space included the following percentiles of distances’ distribution: [0.005, 0.01, 0.05, 0.1, 0.5, 1, 1.5, 2, … , 10].

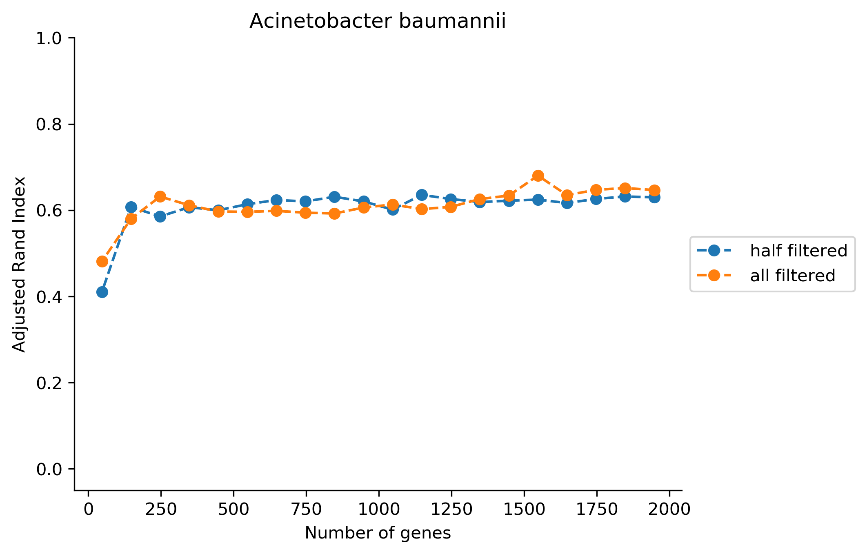


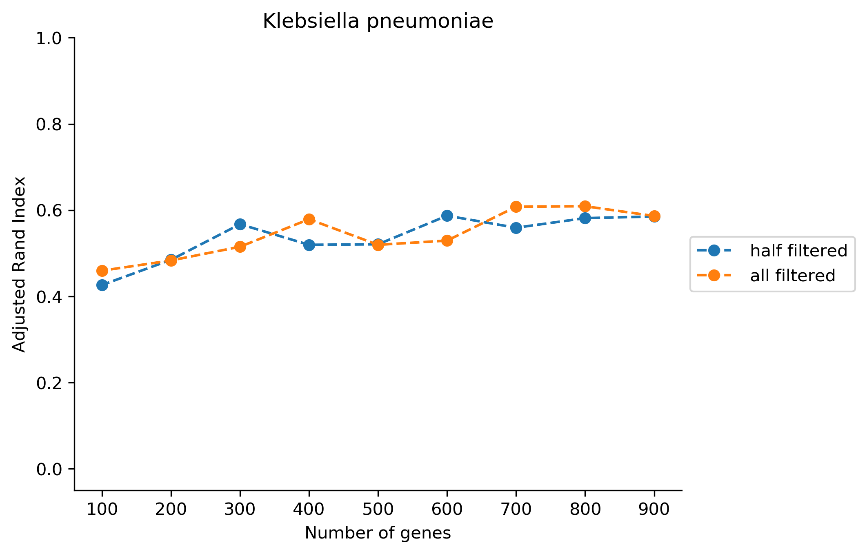


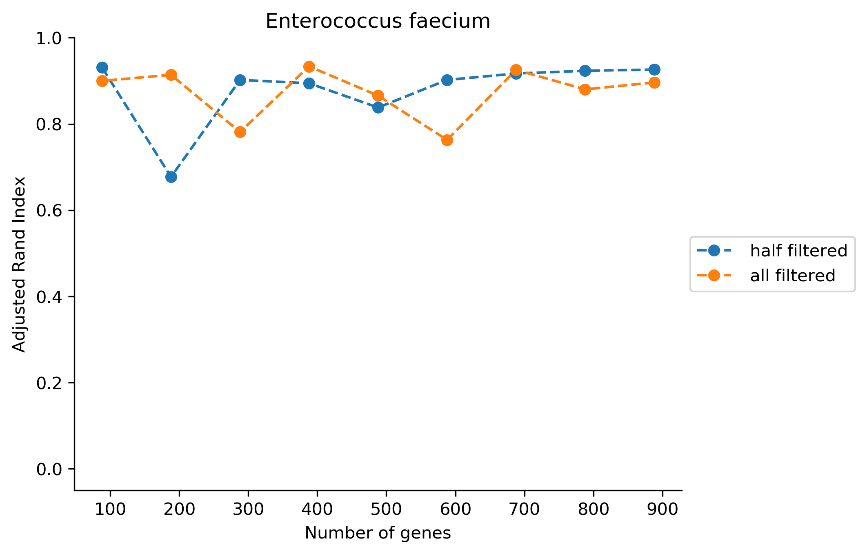
**Supplementary Fig. 5** Phylogenetic trees reconstructed by the GrapeTree tool using the full cgMLST scheme. The inner colors’ ring represents the CTs predefined by *Ridom SeqSphere+* based on all core genes, the outer colors’ ring represents the CTs induced by *minMLST* based on a minimal subset of informative genes. To keep a clear view, only CTs with more than 5 related isolates are colored in the inner ring, whereas the outer ring shows the corresponding typing results of these isolates according to *minMLST* (logic for color settings is depicted in **Supplementary Algorithm 1**). ***\*No. of CTs*** – In *Ridom*’s scheme, it refers to the number of CTs with more than 5 related isolates, i.e. the number of CTs presented in the inner colors’ ring. In *minMLST*’s scheme, it refers to the number of CTs defined for these isolates according to *minMLST*, i.e. the number of CTs presented in the outer colors’ ring. The and p-value were calculated based on the typing results of all isolates.

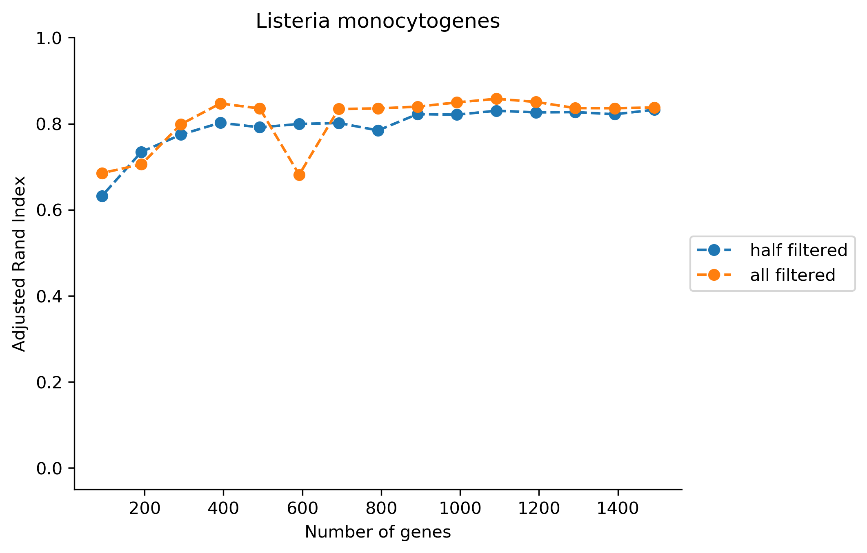


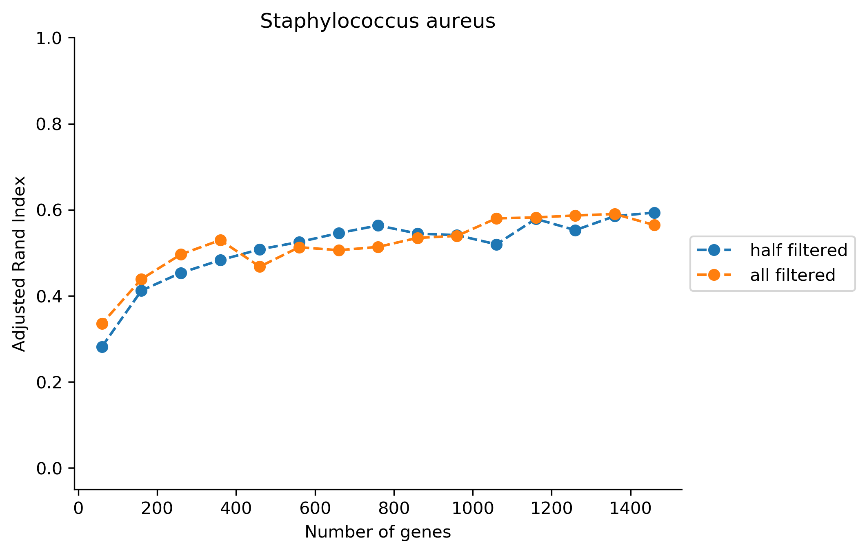




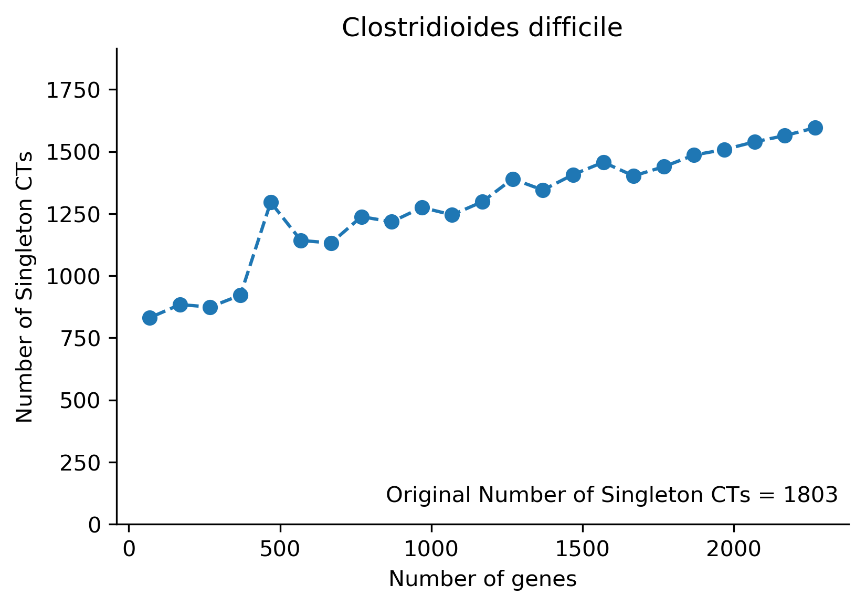
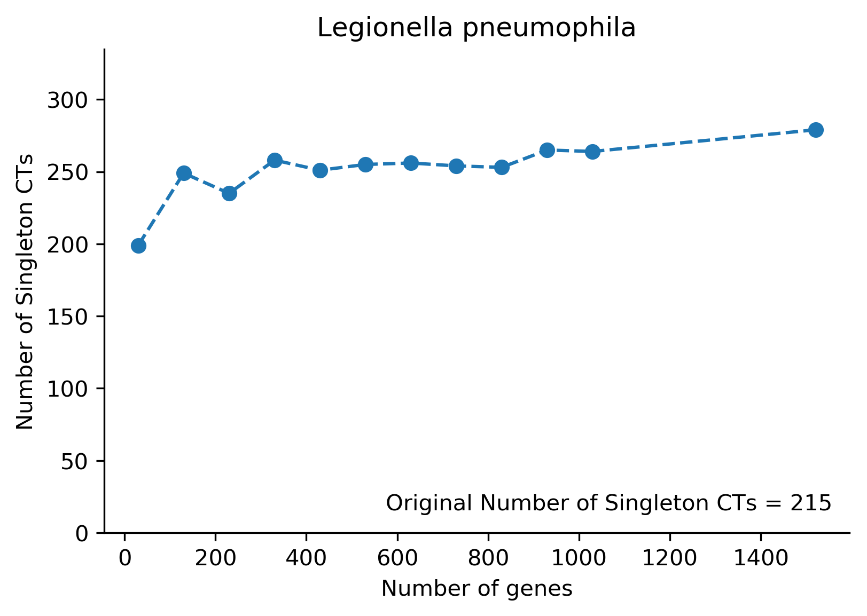
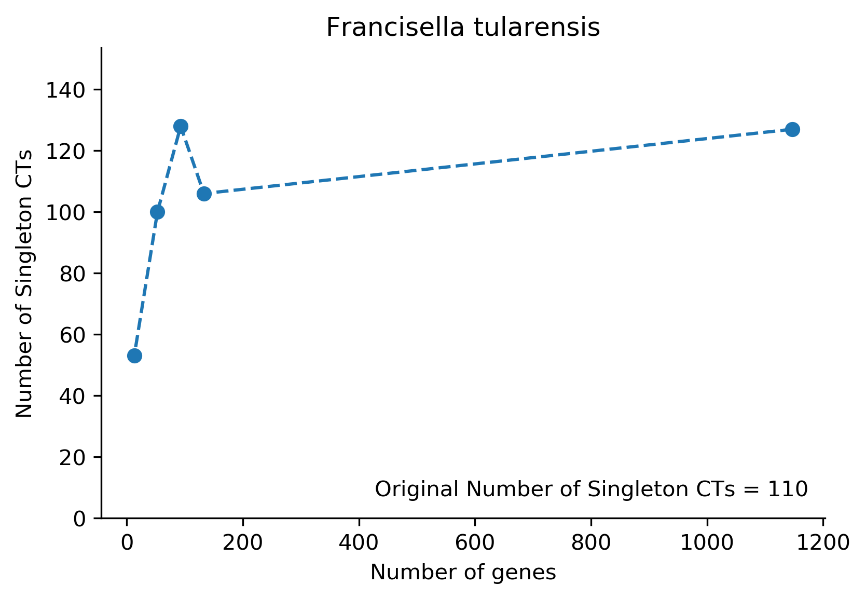


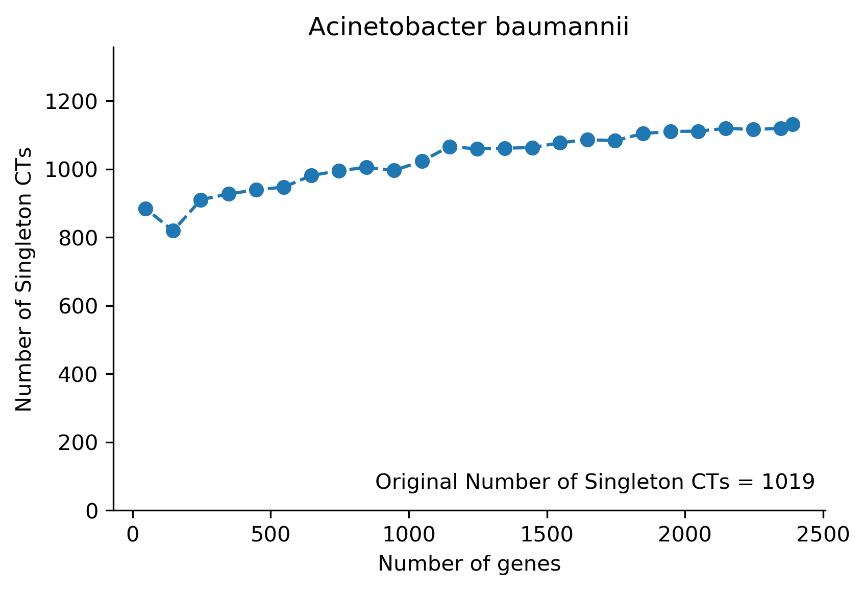


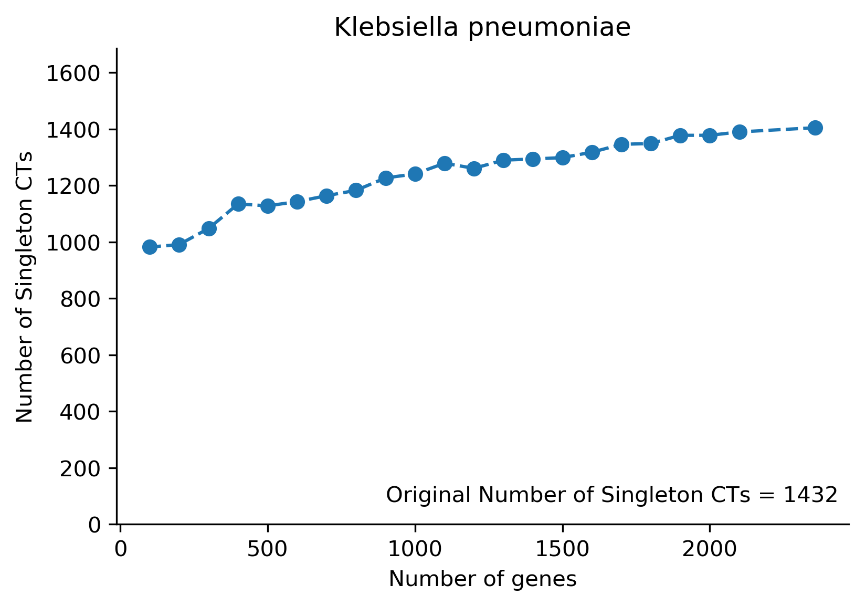


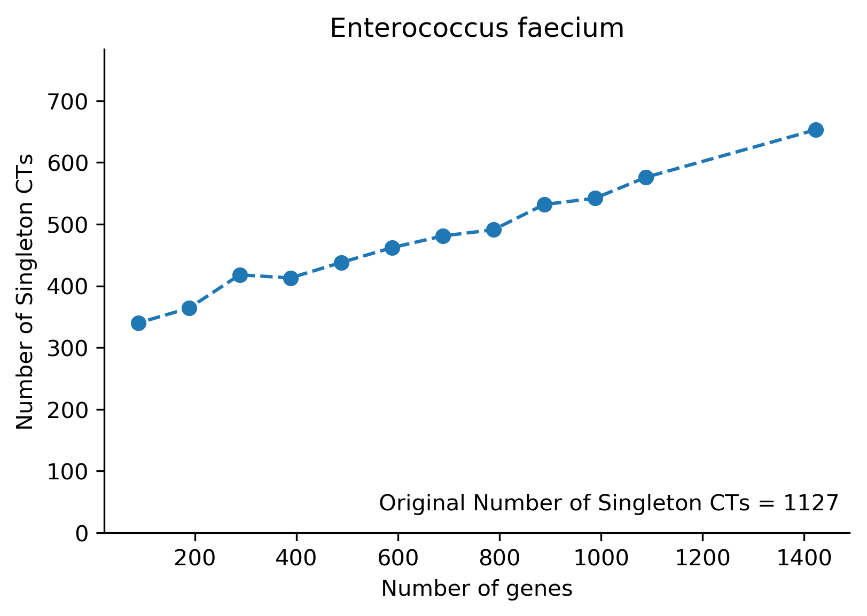


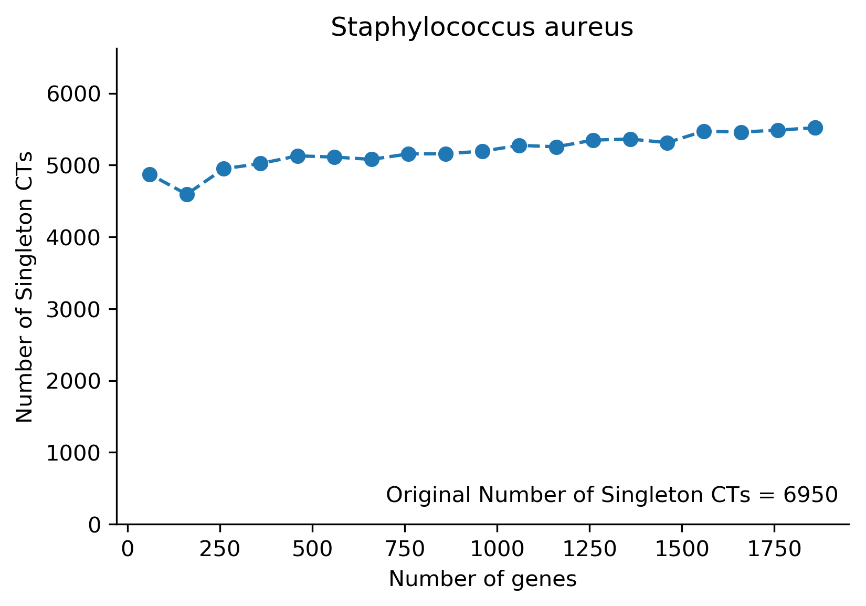
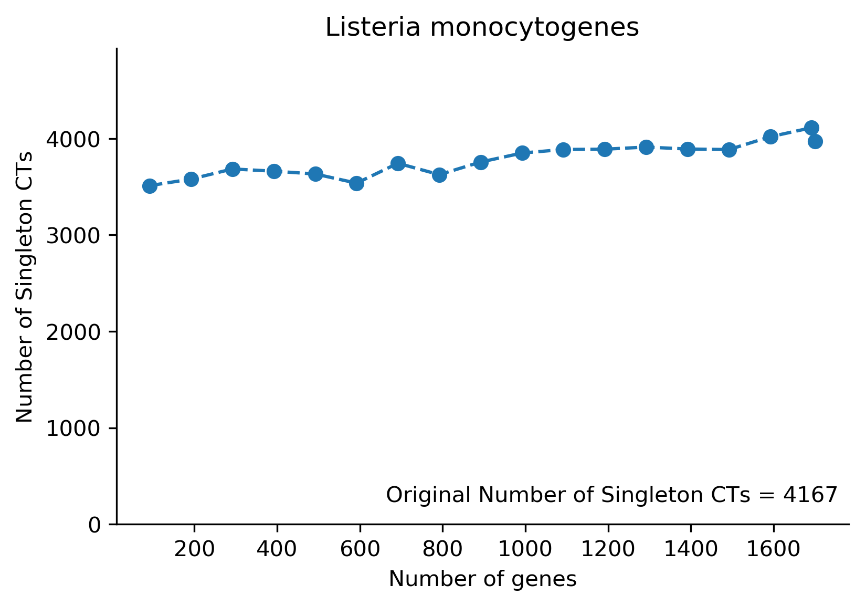
**Supplementary Fig. 6** Adjusted Rand Index computed for each subset of most informative genes when the XGBoost model is trained without singletons (denoted as ‘all filtered’) vs. trained with half of the singletons selected randomly (denoted as ‘half filtered’). We present the results obtained with the best percentile found by *minMLST* when using the complete linkage method for clustering. The results include only subsets of genes that are comparable between the two models, i.e., subsets that are consist of the same number of informative genes. Specifically, for *Francisella tularensis*, only 38 informative genes were found by the ‘half filtered’ model, resulting in only one overlapping point with the ‘all filtered’ model (of 13 genes). Therefore, its results were omitted.





****





**Supplementary Fig. 7** Number of singleton CTs (cluster types containing a single related isolate) generated by *minMLST* with each subset of most informative genes when using the complete linkage method. We present the results obtained with the best percentile found by *minMLST* for each bacterial scheme. We also mention the number of singleton CTs in the original scheme predefined by *Ridom*.

# **Files**

**Supplementary File 1** Gene importance values for all bacterial species, computed by six different measures: *the mean magnitude of the SHAP values*, *weight*, *average gain*, *average cover*, *total gain*, and *total cover*.

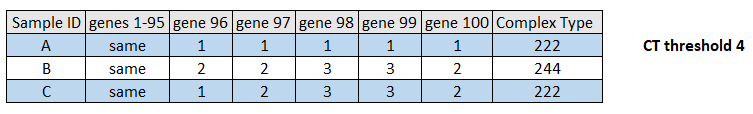
# **Algorithms**

**Supplementary Algorithm 1** Pseudocode of color settings

|  |
| --- |
| :   * – a bacterial scheme in a matrix format with **m** rows and **n+1** columns: **n-1** columns  correspond to genes, column **n** contains the predefined CT (denoted as ), and column **n+1** contains the CT induced by *minMLST* based on a minimal subset of genes (denoted as ).  Each row **m** represents an allelic profile of a single isolate. * – the minimal number of related isolates required per predefined CT in order to be colored. |
| 1: For each in : |
| 2: If has more than related isolates in : |
| 3: = a unique color selected randomly. |
| 4: Else: |
| 5: = None. |
| 6: For each in : |
| 7: = None. |
| 8: For each in : |
| 9: = the isolates relate to |
| 10: = all the that isolates from are related to. |
| 11: = the most common among isolates from . |
| 12: For each in : |
| 13: if = : |
| 14: = . |
| 15: Else: |
| 16: = a unique color selected randomly. |
| 17: For each isolate in : |
| 18: = the predefined CT of isolate . |
| 19: = the CT induced by *minMLST* for isolate . |
| 20: = . |
| 21: = . |

# **Ridom SeqSphere+**

As described in the *Ridom* [*SeqSphere+*](https://www.ridom.de/seqsphere)official documentation[[1]](#footnote-2), every sample submitted to the cgMLST.org nomenclature server is assigned to a static complex type (CT), i.e. cluster-type, that lumps together samples with similar cgMLST allelic profiles. Each CT is represented by a numeric value and has a CT founder, which is the first sample that was submitted and assigned to this CT by the system. When a new sample is submitted, it will be assigned to an existing CT only if its distance from the CT founder does not exceed a pre-defined threshold, otherwise, a new CT will be established, and the sample will become the founder of the new CT. The pre-defined threshold is species-specific and stands for the maximal number of different alleles allowed between a submitted sample and an established CT founder. In case a submitted sample is within the boundaries of several CT founders, it will be assigned to the CT with the lowest numeric value that not necessarily belongs to the closest CT founder. This algorithm guarantees that an exact identical allelic profile that is submitted multiple times will always get the same CT assigned, regardless of the order of submissions or the time when the CT founder was established. However, some other issues in CT assignments may stem from the above algorithm; First, the CT assignment of samples is influenced by the order of submission since it is determined with respect to the allelic profile of the first submitted CT founder, from which the distance is being measured. Moreover, due to the tie-breaking rule of selecting the lower CT numeric value, two different CT may fall together as illustrated in **Supplementary Figure 7**. More examples are provided in the official documentation of *Ridom* [*SeqSphere+*](https://www.ridom.de/seqsphere)[[2]](#footnote-3).



**Supplementary Fig. 7** An example that illustrates the issue caused by the tie-breaking rule of selecting the lowest CT numeric value. Given a threshold of 4, sample C will get the CT value of sample A (distance =4), though it is much more similar to sample B (distance =1).

1. <https://www.ridom.de/u/Core_Genome_MLST_Complex_Type.html> [↑](#footnote-ref-2)
2. <https://www.ridom.de/u/Core_Genome_MLST_Complex_Type.html> [↑](#footnote-ref-3)