

A Modest Proposal

for making the *Genomes* of *GenBank* beneficial to the PUBLICK
and preventing them from being a Burthen
to the *Curators* and *Taxonomists*

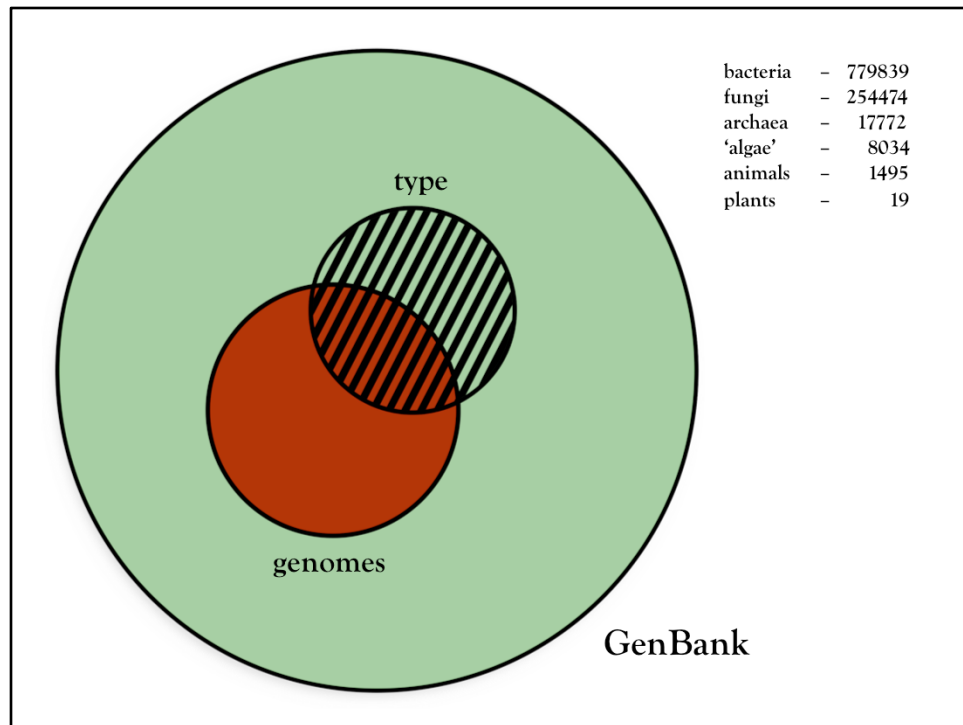
Apologies to Jonathan Swift for cribbing his title. Talk first given at the NCBI IEB seminar, January 2015. And again at the NCBI Genomic Taxonomy Workshop, 12 May 2015.

About that which we cannot align, we must remain silent.

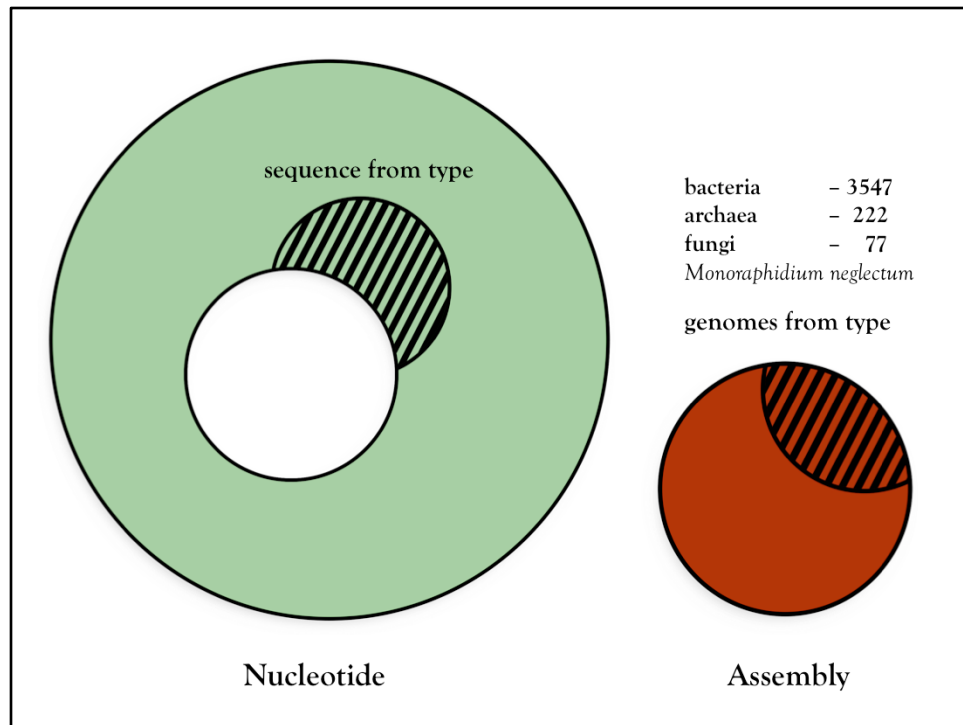
*All things being simple,
every genome within a clade should be statistically equivalent
with respect to every genome outside of the clade.*

like Yi qi of the Scansoriopterygidae, a carnivorous flying dinosaur

three mantras.



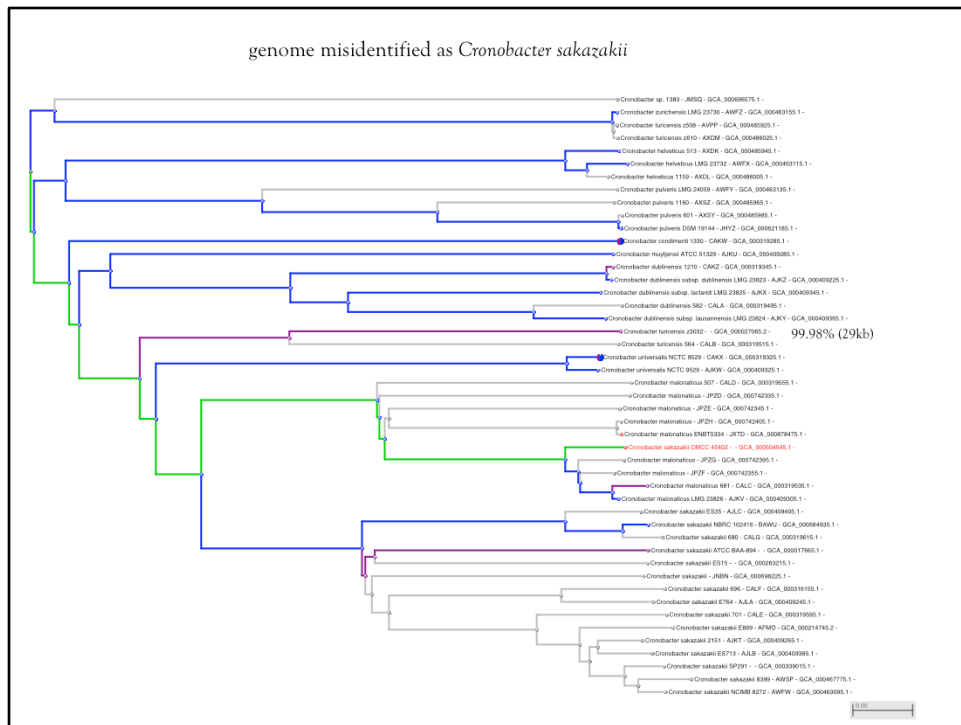
Sequences in GenBank. Some are components of genome submissions & some are taken from type specimens. Some are both & some are neither. Most of our sequences from type are from the cultured microbes, where type material is easily obtained from the culture collections.



For counting purposes, it is convenient to separate the genome submissions from the rest of GenBank. We have just over 4000 genomes from type, roughly 2/3 from WGS projects and 1/3 complete sequences. The proxytype analysis described below involves blasting the remaining sequence from type in GenBank against all of our genomes (from type and not from type).

100.		type	701108	Cronobacter malonaticus LMG 23826
99.95 (96% 95%)			1190578	Cronobacter malonaticus
99.94 (97% 95%)			1190558	Cronobacter malonaticus
99.92 (96% 93%)			871738	Cronobacter sakazakii CMCC 45402
99.84 (98% 96%)			506628	Cronobacter malonaticus 681
98.46 (92% 90%)			1190598	Cronobacter malonaticus
98.42 (92% 89%)			1190538	Cronobacter malonaticus
98.40 (92% 92%)			1190528	Cronobacter malonaticus
98.34 (89% 87%)			506668	Cronobacter malonaticus 507
94.85 (85% 95%)			347428	Cronobacter sakazakii E899
94.79 (90% 88%)		type	1063258	Cronobacter sakazakii NBRC 102416
94.78 (89% 91%)			701208	Cronobacter sakazakii ES35
94.75 (87% 88%)			506768	Cronobacter sakazakii 680
94.75 (90% 88%)			701188	Cronobacter sakazakii ES713
94.75 (87% 90%)			380658	Cronobacter sakazakii ES15
94.72 (89% 87%)			538458	Cronobacter sakazakii SP291
94.72 (90% 87%)			773168	Cronobacter sakazakii NCIMB 8272
94.72 (88% 89%)			701068	Cronobacter sakazakii 2151
94.69 (87% 80%)			506728	Cronobacter sakazakii 701
94.69 (89% 87%)			18468	Cronobacter sakazakii ATCC BAA-894
94.69 (90% 85%)			782428	Cronobacter sakazakii 8399
94.68 (89% 79%)			501968	Cronobacter sakazakii 696
94.64 (89% 89%)			1094758	Cronobacter sakazakii
94.60 (89% 89%)			701048	Cronobacter sakazakii E764
93.95 (88% 90%)		type	701128	Cronobacter universalis NCTC 9529
93.94 (87% 85%)		type	506278	Cronobacter universalis NCTC 9529
92.42 (87% 84%)			506588	Cronobacter turicensis 564
92.37 (90% 86%)			249078	Cronobacter turicensis z3032

ANI neighboring table for the type genome of *Cronobacter malonaticus*. Figures in parentheses gives the fraction of each genome which contribute to the pairwise alignment. Note the presence of a putative misidentified genome (assembly id 871738, GenBank CP006731) submitted as *Cronobacter sakazakii*. Spaces introduced in display to highlight clades with statistically equivalent scores.



Position of the putative misidentification in the kmer tree (lineage trace in green, label in red). All trees in this presentation are neighbor-joining trees of pairwise kmer statistics, imaged in the Genome Workbench.

Lineage traces in blue lead to genomes from type (and proxtype). lineages in purple are RefSeq reference & representative genomes. We do not have a genome from type for *Cronobacter turicensis*, but 29kb of GenBank sequence from type identifies GCA_000027065 as the proxtype genome for this species (at 99.98% identity). There is a YouTube web tutorial on using the genome workbench tree viewer –

<https://youtu.be/dei0Z6HS6OM>

non-genomic type sequences in GenBank

947 *Agrobacterium fabrum*
715 *Bacillus subtilis*
469 *Desulfovibrio vulgaris*
421 *Bacillus alcalophilus*
391 *Legionella pneumophila*
...
75 *Bacillus subtilis* subsp. *spizizenii*
...
5 560 species
4 787 species
3 1004 species
2 1648 species
1 4552 species
0 2935 species

Counts of sequence from type in GenBank (excluding components of genomes from type). We do have genomes from type for the top species on this list. The 4552 species with a single type sequence in GenBank will be almost entirely 16S sequences. Many of the 2935 species without type sequences in GenBank will be legacy names without standing in the current nomenclature.

how identical is co-identical?

4 pairs of genomes < 90% identical.
5 pairs of genomes 96% - 99% identical.
8 pairs of genomes 99% - 99.9% identical.
112 pairs of genomes 99.9% - 99.99% identical.
135 pairs of genomes 99.99% - 99.999% identical.
10 pairs of genomes > 99.999% identical.

There are 274 species for which we have more than one genome from type, often sequenced by different labs, from different culture collections. The median ANI is just over 99.99% identity, but there is a wide range of values.

81.77%	<i>Streptomyces aureofaciens</i> NRRL B-2657	(18.35%) GCA_000719175
	<i>Streptomyces aureofaciens</i> NRRL B-2183	(20.09%) GCA_000720845
83.13%	<i>Pseudomonas cremoricolorata</i> DSM 17059	(54.39%) GCA_000425745
	<i>Pseudomonas cremoricolorata</i> NBRC 16634	(56.75%) GCA_000730545
83.36%	<i>Streptomyces sclerotialis</i> NRRL B-2317	(36.92%) GCA_000719105
	<i>Streptomyces sclerotialis</i> NRRL ISP-5269	(31.52%) GCA_000720555
89.21%	<i>Acinetobacter lwoffii</i> CIP 64.10	(79.32%) GCA_000369105
	<i>Acinetobacter lwoffii</i> NCTC 5866	(82.94%) GCA_000248355
96.27%	<i>Piscirickettsia salmonis</i> LF-89	(79.81%) GCA_000297215
	<i>Piscirickettsia salmonis</i> ATCC VR-1361	(78.96%) GCA_000300295
96.80%	<i>Bacillus subtilis</i> subsp. <i>spizizenii</i> TU-B-10	(89.73%) GCA_000227465
	<i>Bacillus subtilis</i> subsp. <i>spizizenii</i> ATCC 6633	(94.87%) GCA_000177595
97.98%	<i>Brucella inopinata</i> BO1	(91.32%) GCA_000662015
	<i>Brucella inopinata</i> BO1	(87.03%) GCA_000182725
98.20%	<i>Actinobacillus pleuropneumoniae</i> str. 4074	(90.56%) GCA_000167095
	<i>Actinobacillus pleuropneumoniae</i> str. 4074	(91.75%) GCA_000178495
98.35%	<i>Eubacterium rectale</i> ATCC 33656	(73.93%) GCA_000020605
	<i>Eubacterium rectale</i> DSM 17629	(78.28%) GCA_000209935
99.03%	<i>Eggerthia cateniformis</i> OT 569	(87.13%) GCA_000340375
	<i>Eggerthia cateniformis</i> DSM 20559	(87.94%) GCA_000422605
99.45%	<i>Edwardsiella tarda</i> ATCC 15947	(90.67%) GCA_000264805
	<i>Edwardsiella tarda</i> NBRC 105688	(92.78%) GCA_000341505
99.47%	<i>Mycobacterium intracellulare</i> ATCC 13950	(95.87%) GCA_000172115
	<i>Mycobacterium intracellulare</i> ATCC 13950	(97.20%) GCA_000277125
99.87%	<i>Cronobacter universalis</i> NCTC 9529	(94.01%) GCA_000319325
	<i>Cronobacter universalis</i> NCTC 9529	(98.19%) GCA_000409325
99.88%	<i>Prevotella disiens</i> ATCC 29426	(99.60%) GCA_000467875
	<i>Prevotella disiens</i> JCM 6334	(95.61%) GCA_000613345
99.88%	<i>Bifidobacterium angulatum</i> DSM 20098	(99.77%) GCA_000156635
	<i>Bifidobacterium angulatum</i> LMG 11039	(99.94%) GCA_000741065
99.88%	<i>Prochlorothrix hollandica</i> PCC 9006	(58.29%) GCA_000332315
	<i>Prochlorothrix hollandica</i> PCC 9006	(97.04%) GCA_000341585
99.89%	<i>Streptomyces clavuligerus</i> ATCC 27064	(72.29%) GCA_000148465
	<i>Streptomyces clavuligerus</i> ATCC 27064	(98.16%) GCA_000154925

These are the most divergent values, including all of the cases with less than 99.9% identity. One was from inappropriate annotation in the taxonomy database – ATCC 6633 is not a type strain of *Bacillus subtilis* subsp. *spizizenii*. Others are assemblies with contamination – note that GCA_000332315 is almost twice the size of GCA_000154925 (both of which are from PCC 9006). Others are problems with strains in the culture collections – NRRL B-2657 & NRRL B-2183 both claim to be type strains of *Streptomyces aureofaciens*. Neither of these are actually type, and NRRL B-2657 is likely from a different species.

Types of Types	
type strain	(P) deposited in at least two different culture collections
neotype strain	(P) replacement culture for a type that has been lost
holotype	(BZ) single name-bearing type specimen
paratype	(BZ) other specimens in the original type series
neotype	(BZ) single name-bearing specimen designated when holotype is lost
allotype	(Z) designated specimen of opposite sex to the holotype
epitype	(B) other specimen designated in subsequent type series
isotype	(B) duplicate specimen of the holotype
syntype	(BZ) one of a name-bearing series of type specimens
isosyntype	(B) duplicate specimen of a syntype
lectotype	(BZ) syntype subsequently designated as the single name-bearing type
paralectotype	(B) syntype specimens not designated as lectotype
hapanotype	(B) protest culture/collection collectively designated as holotype
type material	(-) type material of unknown type
reference material	(-) non-type material reserved for arbitrary uses
culture from type	(-) living culture derived from type material

Types of type annotated in the NCBI taxonomy database, taken from the bacterial (P), botanical (B) and zoological (Z) codes of nomenclature. Figure from (Federhen, 2015)


[news](#)
[about](#)
[contact](#)
[help](#)

[Advanced search](#)
[StrainInfo Projects](#)
[log in](#)
[register](#)

Taxon Passport

Escherichia coli

overview

species *Escherichia coli*

parent taxon *Escherichia sp.*

type strain [ACM 1803 T](#), [AJ 2617 T](#), [ATCC 11775 T](#), [BCRC 10675 T](#), [BTCC US/41 T](#), [CAPM 6101 T](#), [CCM 5172 T](#), [CCRC 10675 T](#), [CCT 0547 T](#), [CCT 1357 T](#), [CCTM 2067 T](#), [CCTM La 2067 T](#), [CCUG 24 T](#), [CCUG 29300 T](#), [CD88 964 T](#), [CECT 515 T](#), [CGMCC 1.2389 T](#), [CGMCC 1.2993 T](#), [CIP 54.8 T](#), [CIP 54.8T^T](#), [CN 4382 T](#), [CNCTC 6859 T](#), [CNCTC Eck 206/59 T](#), [CNCTC Eck 58/59 T](#), [DSM 30083 T](#), [DSMZ 30083 T](#), [ELI 50 T](#), [F. Oerskov T](#), [FIRDI 675 T](#), [GISK 240001 T](#), [IAM 12119 T](#), [ICMP 15663 T](#), [JCM 1649 T](#), [K. Lincoln T](#), [Kauffmann US/41 T](#), [KCTC 2441 T](#), [LMD 54.8 T](#), [LMG 2092 T](#), [LRA 73.08.009 T](#), [NBIMCC 3398 T](#), [NBRC 102203 T](#), [NCAIM 8.01874 T](#), [NCC8 54008 T](#), [NCDO 1989 T](#), [NCFB 1989 T](#), [NCIB 11943 T](#), [NCIMB 11943 T](#), [NCTC 9001 T](#), [NCTC 9001 US/41^T](#), [NCTC 9001^T](#), [NZRM 3309 T](#), [PCM 172 T](#), [PCM 321 T](#), [R. Sakazaki T](#), [SSIC U 5/41 T](#), [strain U 5/41 T](#), [U 5-41 T](#), [US/41 T](#), [USCC 2054 T](#), [USCC 2520 T](#), [VTT E-94564 T](#), [WDCM 00090 T](#), [Y. Kosako 82039 T](#)

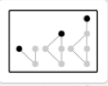
16S rRNA gene [AB681728 \(LTP: X80725\)](#)

external links [Catalogue of Life[®]](#), [DSMZ[®]](#), [LPSN: J.P. Euzéby[®]](#), [NCBI[®]](#), [WikiSpecies[®]](#)

genome projects [Escherichia coli str. K-12 substr. W3110](#)
[Escherichia coli CFT073](#)
[Escherichia coli APEC O1](#)
[Escherichia coli O157:H7 str. Sakai](#)
[Escherichia coli UT189](#)
[Escherichia coli S36](#)
[Escherichia coli str. K-12 substr. MG1655](#)

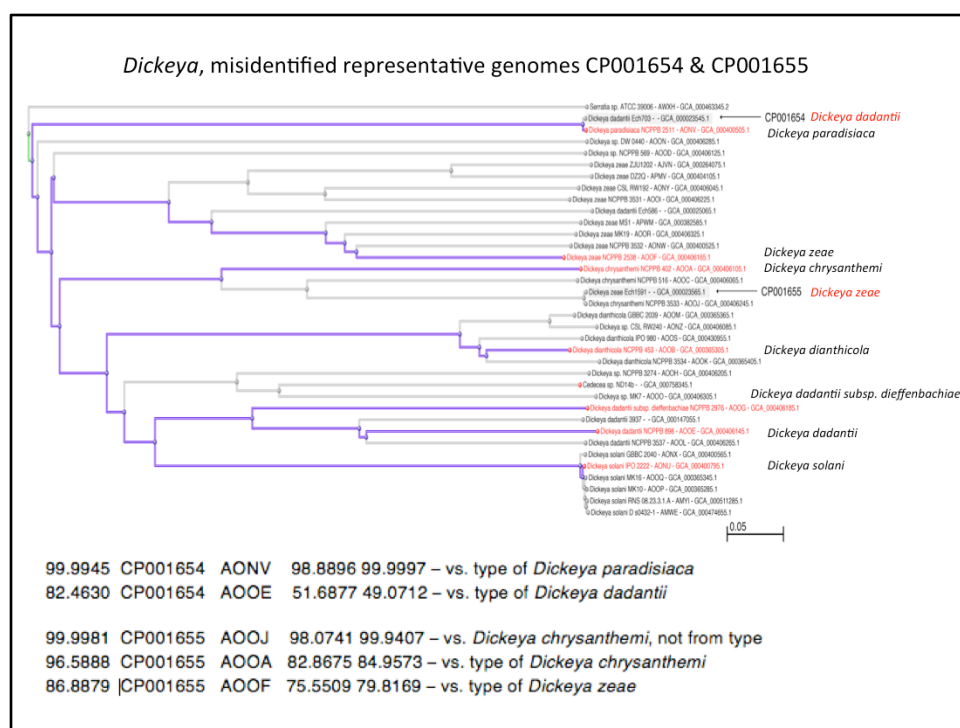
Wiki [Wikipedia Article](#), [Wikispecies Article](#)

search StrainInfo [Find all strains](#) [Find all type strains](#)



SeqRank workflow

Types for *Escherichia coli* in StrainInfo – data derived from the corresponding culture collections.



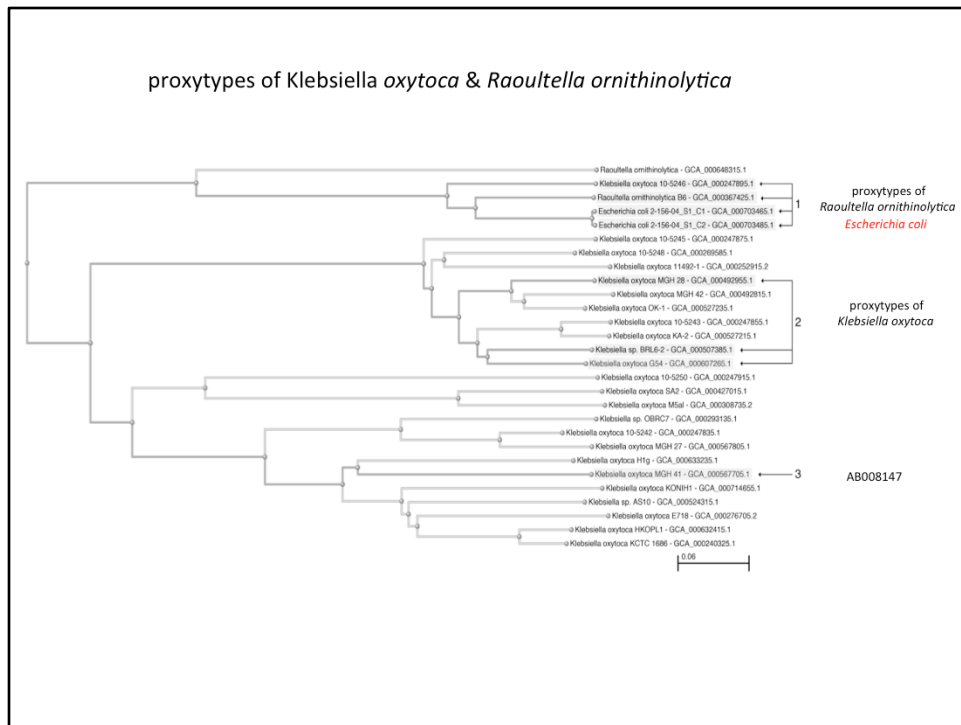
genomes from type in the genus *Dickeya*, with two misidentified genomes. CP001654 was submitted as *Dickeya dadantii*, but is nearly identical to the type genome of *Dickeya paradisiaca*.

CP001655 was submitted as *Dickeya zeae*, but is actually *Dickeya chrysanthemi*.

These two GenBank entries have been UNVERIFIED.

LOCUS	CP001654	4679450 bp	DNA	circular	BCT 10-MAR-2015
DEFINITION	UNVERIFIED: Dickeya dadantii Ech703, complete genome.				
ACCESSION	CP001654				
VERSION	CP001654.1 GI:242129517				
DBLINK	BioProject: PRJNA33069				
	BioSample: SAMN02598494				
KEYWORDS	UNVERIFIED.				
SOURCE	Dickeya dadantii Ech703				
ORGANISM	Dickeya dadantii Ech703				
	Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae; Dickeya.				
REFERENCE	1 (bases 1 to 4679450)				
AUTHORS	Lucas,S., Copeland,A., Lapidus,A., Glavina del Rio,T., Dalin,E., Tice,H., Bruce,D., Goodwin,L., Pitluck,S., Chertkov,O., Brettin,T., Detter,J.C., Han,C., Larimer,F., Land,M., Hauser,L., Kyrpides,N., Mikhailova,N., Balakrishnan,V., Glasner,J. and Perna,N.T.				
CONSRM	US DOE Joint Genome Institute				
TITLE	Complete sequence of Dickeya dadantii Ech703				
JOURNAL	Unpublished				
REFERENCE	2 (bases 1 to 4679450)				
AUTHORS	Lucas,S., Copeland,A., Lapidus,A., Glavina del Rio,T., Tice,H., Bruce,D., Goodwin,L., Pitluck,S., Chertkov,O., Brettin,T., Detter,J.C., Han,C., Larimer,F., Land,M., Hauser,L., Kyrpides,N., Mikhailova,N., Balakrishnan,V., Glasner,J. and Perna,N.T.				
CONSRM	US DOE Joint Genome Institute				
TITLE	Direct Submission				
JOURNAL	Submitted (24-JUN-2009) US DOE Joint Genome Institute, 2800 Mitchell Drive B310, Walnut Creek, CA 94598-1698, USA				
COMMENT	GenBank staff is unable to verify source organism provided by the submitter.				
	CP001654 was submitted as Dickeya dadantii Ech703, but it is 99.9988% identical to the genome from Dickeya paradisiaca NCPPB 2511 (WGS entry AONV00000000.1), the type strain of Dickeya paradisiaca				

The UNVERIFIED entry for *Dickeya dadantii* Ech703. UNVERIFIED entries are indexed in Entrez and exchanged with the INSDC, but are not present in the default BLAST databases.



Two genomes misidentified as *Escherichia coli* appeared in the *Raoultella/Klebsiella oxytoca* clade. We don't have genomes from type for either *Klebsiella oxytoca* or *Raoultella ornithinolytica*, but we can assign proxytype genomes based on sequences from type in GenBank. AB008147 is anomalous – it claims to be type sequence from *Klebsiella oxytoca*, but does not behave as expected.

proxytype computation for *Raoultella ornithinolytica*

	AF129447	AF303652	JX397961	AY134479	AF303618	DQ227492	AY307386	AY134485	AB008149
	ATCC 31898	ATCC 31898	DSM 7464	DSM 7464	ATCC 31898	DSM 7464	ATCC 31898	DSM 7464	JCM 6096
	rpoB	parC	tuf	tuf	gyrA	oriC	blaORN1a	atpD	groES/EL
99.505 6638/6671	Raoultella ornithinolytica B6	496	319	515	789	379	289	1128	801
99.505 6638/6671	Escherichia coli 2-156-04_S1_C2	494	319	515	789	380	289	1128	803
99.505 6638/6671	Escherichia coli 2-156-04_S1_C1	494	319	515	789	380	289	1128	803
99.49 6637/6671	Klebsiella oxytoca 10-5246	496	319	515	789	380	289	1125	801
97.692 6517/6671	Raoultella ornithinolytica	487	318	510	784	380	287	1051	795
94.549 4753/5027	Klebsiella oxytoca MGH 41	289	507	779	349	261			759
94.49 4750/5027	Klebsiella oxytoca 10-5242	290	507	778	351	261			759
94.47 4749/5027	Klebsiella oxytoca MGH 27	290	507	778	350	261			759
94.45 4748/5027	Klebsiella sp. AS10	289	505	777	352	259			759
94.317 5228/5543	Klebsiella oxytoca HCOPL1	473	289	507	779	352	261		759
94.299 5227/5543	Klebsiella oxytoca KCTC 1686	472	289	507	779	352	261		759
94.281 5226/5543	Klebsiella oxytoca KONH1	473	289	507	779	351	261		759
94.245 5224/5543	Klebsiella oxytoca E718	471	288	507	779	351	261		759
94.245 5224/5543	Klebsiella oxytoca	473	289	507	779	352	258		759
94.231 4737/5027	Klebsiella oxytoca MSal		291	505	775	352	257		757
94.028 4912/5224	Enterobacter aerogenes KCTC 2190	471	497	764	354	258			772
94.028 4912/5224	Enterobacter aerogenes EA509E	470	497	764	354	258			775
93.92 5206/5543	Klebsiella oxytoca 10-5250	478	291	506	777	349	257		755
93.496 4701/5028	Klebsiella oxytoca 11492-1	472	288		768	349	260		760
93.477 4700/5028	Klebsiella oxytoca 10-5245	474	288		768	349	259		761
93.455 4612/4935	Enterobacter cloacae ECNH2	474		499	756	341			763
93.389 4549/4871	Enterobacter lignolyticus SCF1	466	290	504	770				755
93.325 5173/5543	Klebsiella varicola At-22	469	286	497	764	340	261		762
93.29 4435/4754	Klebsiella oxytoca SA2	473	291	505	752	257			757
93.289 5171/5543	Klebsiella pneumoniae 342	470	286	496	763	339	262		761
93.212 4600/4935	Enterobacter asburiae L1	473	497	751	341				755
93.166 4867/5224	Citrobacter freundii CFWH1	473		501	771	353	242		762
93.147 4866/5224	Klebsiella pneumoniae subsp. pneumoniae MGH 7857i	465	495	763	337	262			762
93.126 5162/5543	Klebsiella pneumoniae subsp. pneumoniae KPS-1	470	286	494	761	340	261		762
93.109 4864/5224	Klebsiella pneumoniae subsp. pneumoniae Kp13	466		495	763	334	262		762
93.09 4863/5224	Klebsiella pneumoniae subsp. pneumoniae 1084	466		495	763	336	262		761
93.07 4862/5224	Klebsiella pneumoniae	466		495	761	336	262		762
93.051 4861/5224	Klebsiella pneumoniae subsp. pneumoniae H511286	465		495	762	336	262		762
93.051 4861/5224	Klebsiella pneumoniae CG43	465		495	763	336	261		762
93.051 4861/5224	Klebsiella pneumoniae ATCC BAA-2146	465		495	762	336	262		762
93.032 4860/5224	Klebsiella pneumoniae 30684/NIST258_2	465		495	762	335	262		762
93.032 4860/5224	Klebsiella pneumoniae 30660/NIST258_1	465		495	762	335	262		762
93.015 4887/5254	Enterobacter cloacae subsp. cloacae ATCC 13047	469	285	497	753	343			756

We list nine sequences from type for *Raoultella ornithinolytica* (ribosomal RNAs were excluded from this computation) Counts in the table are the number of identities in the strongest BLAST HSP.

These are accumulated by target genome and sorted by %identity. The top four genomes are equally good proxytypes for *Raoultella ornithinolytica* (at 99.5% identity) Only one of these was actually submitted with the name *Raoultella ornithinolytica*.

proxytype computation for *Klebsiella oxytoca*

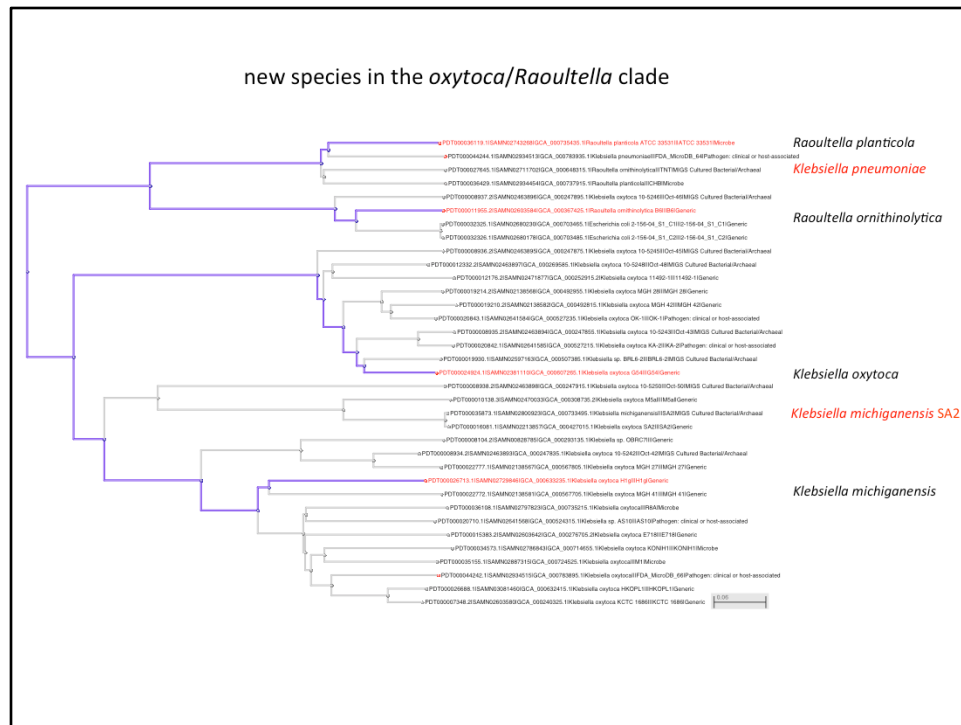
	EU010882	DQ227490	AF525464	(AB008147)	AF008433	AF767363	EU010028	JK424964	AF767380	AF475377	AF052257	JK425223	JK425093	JK425349	EU010005	EU010109	U77442	AY770849	
	ATCC 13382	D504-1175	ATCC 13382	ATCC 13382	ATCC 13382	ATCC 13382	ATCC 13382	LMG 3055	ATCC 13382	ATCC 13382	LMG 3055	LMG 3055	LMG 3055	LMG 3055	ATCC 13382	ATCC 13382	ATCC 13382	ATCC 13382	
	pyrG	oriC	aba	groEL	gmk	rpoB	hcrA	atpD	hsh	bla OXY	gmk	rnfB	gyrB	rpoB	hcrA	rpoB	rpoB	gyrB	
99.883 10306/10318	Klebsiella sp. BRL6-2	306	289	463	1872	178	1029	633	642	670	1056	589	615	742	637	642	501	512	502
99.746 10392/10318	Klebsiella oxytoca MGH 28	306	287	462	1873	178	1026	632	642	670	1056	588	615	742	636	639	500	511	502
99.746 10392/10318	Klebsiella oxytoca G54	306	289	463	1873	178	1027	633	642	670	1053	589	615	740	636	639	500	511	501
99.716 10180/10318	Klebsiella oxytoca MGH 42	306	287	464	1873	178	1026	632	642	669	1052	588	615	742	636	637	500	511	502
99.667 10184/10318	Klebsiella oxytoca OK 1	306	286	462	1873	178	1026	632	642	670	1051	588	615	742	636	637	500	511	502
99.657 10183/10318	Klebsiella oxytoca 10-1245	305	286	462	1873	177	1028	632	642	667	1053	588	613	740	637	638	500	512	501
99.648 10182/10318	Klebsiella oxytoca 11482.1	305	287	461	1872	178	1026	632	641	670	1053	588	613	740	635	636	500	510	501
99.648 10182/10318	Klebsiella oxytoca 10-1248	306	287	462	1872	178	1026	633	641	670	1054	588	613	740	635	639	499	510	501
99.609 10178/10318	Klebsiella oxytoca KA 2	305	288	461	1873	178	1026	632	642	670	1051	589	615	739	635	636	499	510	502
99.609 10178/10318	Klebsiella oxytoca 10-1243	305	288	461	1873	178	1026	632	642	670	1051	589	615	739	635	636	499	510	502
95.234 9131/10318	Klebsiella oxytoca HPGPL1	287	281	624	1915	172	1003	624	636	636	913	575	599	704	622	605	487	495	488
95.205 9128/10318	Klebsiella oxytoca K09M1	286	281	624	1912	171	1003	623	636	637	915	574	599	702	622	606	487	495	487
95.195 9127/10318	Klebsiella oxytoca ICTC 2086	287	281	624	1915	172	1002	624	636	636	913	575	599	704	621	605	486	494	488
95.195 9127/10318	Klebsiella oxytoca	286	278	622	1912	172	1003	624	636	637	915	575	599	704	622	605	487	495	487
95.175 9125/10318	Klebsiella sp. AS20	285	279	624	1912	172	1003	624	636	636	915	575	599	701	622	605	487	495	487
95.097 9117/10318	Klebsiella oxytoca E718	285	281	623	1913	171	1001	624	635	636	912	575	599	704	620	606	487	493	487
94.981 9121/9603	Klebsiella oxytoca MGH 41	286	281	625	1927	172	1001	624	637	635	912	575	702	621	602	486	495	487	
94.949 9118/9603	Klebsiella sp. ORMC7	286	281	620	1914	172	1001	624	635	634	914	575	703	620	608	487	493	485	
94.949 9118/9603	Klebsiella oxytoca 10-1242	286	281	622	1914	172	1003	624	636	634	907	575	702	622	608	487	495	484	
94.929 9116/9603	Klebsiella oxytoca MGH 27	286	281	622	1914	171	1003	624	636	634	907	574	702	622	608	487	495	484	
94.887 9112/9603	Klebsiella oxytoca H12	285	281	621	1914	173	1000	624	636	636	914	573	701	620	604	486	493	485	
94.156 9012/9603	Klebsiella oxytoca S42	288	279	598	1874	176	996	627	630	626	907	575	695	619	588	486	495	478	
94.293 9055/9603	Klebsiella oxytoca M5a1	288	279	598	1874	176	995	627	630	626	907	575	692	618	589	485	495	475	
93.825 9010/9603	Klebsiella oxytoca 10-1250	289	279	592	1874	177	987	626	633	622	902	575	694	613	579	482	486	474	
93.542 6142/6566	Enterobacter aerogenes UCI 48	279	262	1805	166	986	599	552	552	673	618	571	489	495	452				
93.542 6142/6566	Enterobacter aerogenes UCI 16	279	262	1805	166	986	599	552	552	673	618	571	489	495	452				
93.542 6142/6566	Enterobacter aerogenes UCI 15	279	262	1805	166	986	599	552	552	673	618	571	489	495	452				
93.542 6142/6566	Enterobacter aerogenes MGH-78	279	262	1805	166	986	599	552	552	673	618	571	489	495	452				
93.512 6141/6566	Enterobacter aerogenes UCI 46	279	262	1805	165	986	599	551	551	673	618	571	489	495	452				
93.512 6141/6566	Enterobacter aerogenes UCI 47	279	262	1805	165	986	599	551	551	673	618	571	489	495	452				
93.451 6136/6566	Enterobacter aerogenes MGH 62	279	262	1808	166	987	599	552	552	667	619	570	489	495	451				
93.451 6136/6566	Enterobacter aerogenes MGH 11	279	260	1808	166	987	598	552	552	670	619	569	489	495	452				
93.405 6133/6566	Enterobacter aerogenes UCI 45	279	262	1807	166	984	599	552	552	671	617	572	487	491	451				
93.405 6133/6566	Enterobacter aerogenes UCI 28	279	262	1807	166	984	599	552	552	671	617	572	487	491	451				
93.405 6133/6566	Enterobacter aerogenes UCI 27	279	262	1807	166	984	599	552	552	671	617	572	487	491	451				
93.36 6130/6566	Enterobacter aerogenes F635	277	261	1809	166	985	598	552	552	670	618	570	487	494	452				
93.313 6133/7001	Citrobacter koseri ATCC BAA-895	272	229	1756	993	617	611	537	537	570	662	616	487	492	447				
93.276 7297/7823	Enterobacter aerogenes ICTC 2128	279	262	1807	166	986	599	608	551	549	671	619	572	487	495	451			
93.212 7292/7823	Enterobacter aerogenes SA3208	279	262	1805	165	985	599	606	551	549	673	617	571	489	494	452			
92.91 1766/6206	Enterobacter cloacae P101	264		1776	987	588	610		528	554	650	613	484	488					

We list more sequences from type for *Klebsiella oxytoca*, and in this case there is a single best candidate for proxytype – the WGS genome ARVT, submitted as *Klebsiella* sp. BRL6-2. Note that AB008147(fourth column, in parenthesis) behaves differently from the rest of these sequences. The strongest hits for this sequence appear in the middle of the table, peaking at *Klebsiella oxytoca* MGH 41.

proxytype computation for *Enterococcus faecium*

100 -		AF124222	X87180	X87181	(AF029771)	(AF029770)	AB075741	AF515643	HQ611243	AJ843271	(U021375)	GU457278	AF375581	AF417582	AL295305	AJ843428	AJ843486
		ATCC 19434	ATCC 19434	ATCC 19434	ATCC 19434	ATCC 19434	JCM 5804	DSM 20477	ATCC 19434	LMG 11423	CP 103014	ATCC 19434	LMG 11423	ATCC 19434	LMG 11423	LMG 11423	LMG 11423
		ifA	ITS	ITS	IS485	IS485	enterocin	rpoB	rpoB	uspA	gruEL	argD	tmhNA	gruEL	235	pheS	rpoA
		751 bp	344 bp	94 bp	798 bp	135 bp	137 bp	823 bp	1177 bp	1102 bp	757 bp	718 bp	318 bp	1904 bp	2854 bp	455 bp	700 bp
99.901	15082/15087	Enterococcus faecium ICT-ET128	746	218	91	713	195	196		1177	1102	757	718	318	1904	455	700
99.9	17893/18011	Enterococcus faecium EnGen0007	746	342	91	753	195	196		1177	1102	757	718	318	1904	2910	455
99.876	14516/14534	Enterococcus sp. GMD4E	746	152			39			1177		757		318	1903	2909	455
99.848	13115/13139	Enterococcus faecium UC10237	746	220		714	157			1177	1094	757	718	318	1904	455	700
99.843	17786/17814	Enterococcus faecium EnGen0175	746	343	91	753	195			1177	1091	757	718	318	1904	2910	455
99.842	14576/14599	Enterococcus durans FB29-CNAB-4	746	343	91	753	195			1177	1091	757	718	318	1904		455
99.84	17485/17513	Enterococcus faecium EnGen0031	746	343	91	753	195			1177	1091	757	718	318	1904	2910	455
99.837	17785/17814	Enterococcus faecium EnGen0149	746	340	91	753	195			1177	1091	757	718	318	1904	2909	455
99.83	14420/14448	Enterococcus faecium LA48-2	746	269	91					1177	1094	757	718	318	1904	2909	455
99.828	17980/18011	Enterococcus faecium ATCC 8459	746	342	91	753	195	193		1177	1094	757	718	318	1903	2910	455
99.826	17783/17814	Enterococcus faecium EnGen0018	746	342	91	753	42			1176	1090	757	718	318	1904	2910	455
99.82	17782/17814	Enterococcus faecium E1071	746	270	91		157			1177	1094	757	718	318	1904	2909	455
99.819	17676/17710	Enterococcus faecium NREL B-2354	746	343	91	753	195	193		1177	1094	757	718	318	1904	2909	455
99.819	11599/11620	Enterococcus sp. GMD3E	746				39			1177		757		318	1903		455
99.814	17687/17720	Enterococcus faecium EnGen0178	746	341			51			1177	1093	756	717	318	1903	2909	455
99.804	17779/17814	Enterococcus faecium NG1	746	341	91		51			1177	1093	756	717	318	1903	2910	455
99.8	17476/17513	Enterococcus faecium EnGen0016	746	341	91		40			1177	1093	756	717	318	1903	2910	454
99.799	13403/13430	Enterococcus faecium CRL1879	746	266	91	753	195	192		1177	1094	757	718	318	1904		455
99.798	17778/17814	Enterococcus faecium EnGen0030	746	342	91		40			1177	1093	756	717	318	1903	2910	455
99.793	14424/14448	Enterococcus faecium EnGen0234	746				40			1177	1093	756	717	318	1903	2910	455
99.792	17777/17814	Enterococcus faecium EnGen0377	746	342	91		40			1177	1089	757	718	318	1904	2910	455
99.792	17777/17814	Enterococcus faecium EnGen0376	746	342	91		40			1177	1089	757	718	318	1904	2910	455
99.792	17777/17814	Enterococcus faecium EnGen0321	746	342	91		40			1177	1089	757	718	318	1904	2910	455
99.792	17777/17814	Enterococcus faecium EnGen0180	746	342	91		40			1177	1089	757	718	318	1903	2910	455
99.789	17476/17513	Enterococcus faecium DO	746	342	91		41			1177	1093	756	717	318	1903	2910	454
99.787	17776/17814	Enterococcus faecium EnGen0375	746	342	91		40			1177	1089	757	718	318	1904	2909	455

One more illustrative proxytype calculation. In this case AF515643 doesn't make the cut with any of the genomes in the table. HQ611243 is also from rpoB, and behaves as expected. his turned out to be a strain mix-up in the submitter's lab, and has been suppressed.



After the publication of (Federhen 2015) we looked at same region again. Several changes had occurred.

A type genome for *Raoultella planticola* appeared, along with another misidentified genome (*Klebsiella pneumoniae* this time).

In addition, someone resequenced strain SA2 at higher coverage, and submitted it as *Klebsiella michiganensis* instead of *Klebsiella oxytoca*.

proxytype calculation for *Klebsiella michiganensis*

		JQ269337 strain W14 834 bp rpoB	JQ070300 strain W14 1395 bp 16S	JQ990329 strain W14 625 bp. gyrA	JQ284304 strain W14 978 bp gyrB
99.347 3804/3829	<i>Klebsiella oxytoca</i> H1g	831 834 0	1382 1392 0	626 626 0	965 977 1
99.19 3798/3829	<i>Klebsiella oxytoca</i> MGH 41	830 834 0	1382 1392 0	622 626 0	964 977 1
99.138 2416/2437	<i>Klebsiella</i> sp. OBRC7	830 834 0		622 626 0	964 977 1
99.112 3795/3829	<i>Klebsiella oxytoca</i> KCTC 1686	828 834 0	1382 1392 0	621 626 0	964 977 1
99.086 3794/3829	<i>Klebsiella oxytoca</i> KONIH1	829 834 0	1382 1392 0	620 626 0	963 977 1
99.06 3793/3829	<i>Klebsiella</i> sp. AS10	829 834 0	1382 1392 0	621 626 0	961 977 1
99.06 3793/3829	<i>Klebsiella oxytoca</i> MGH 27	829 834 0	1380 1392 0	621 626 0	963 977 1
99.034 3792/3829	<i>Klebsiella oxytoca</i> HKOPL1	829 834 0	1378 1392 0	621 626 0	964 977 1
99.034 3792/3829	<i>Klebsiella oxytoca</i> 10-5242	829 834 0	1378 1392 0	622 626 0	963 977 1
99.034 3792/3829	<i>Klebsiella oxytoca</i>	829 834 0	1380 1392 0	621 626 0	962 977 1
98.903 3787/3829	<i>Klebsiella oxytoca</i> E718	827 834 0	1378 1392 0	620 626 0	962 977 1
97.806 3745/3829	<i>Klebsiella oxytoca</i> SA2	809 834 0	1382 1392 0	609 626 0	945 977 1
97.806 3745/3829	<i>Klebsiella michiganensis</i>	809 834 0	1382 1392 0	609 626 0	945 977 1
97.728 3742/3829	<i>Klebsiella oxytoca</i> M5al	809 834 0	1382 1392 0	609 626 0	942 977 1
97.257 3723/3828	<i>Klebsiella oxytoca</i> 10-5250	801 833 0	1379 1392 0	607 626 0	936 977 1
96.911 3702/3820	<i>Klebsiella</i> sp. BRL6-2	805 834 0	1375 1393 2	610 626 0	912 967 0
96.859 3700/3820	<i>Klebsiella oxytoca</i> MGH 42	805 834 0	1374 1393 2	609 626 0	912 967 0
96.859 3700/3820	<i>Klebsiella oxytoca</i> MGH 28	805 834 0	1374 1393 2	609 626 0	912 967 0
96.859 3700/3820	<i>Klebsiella oxytoca</i> G54	804 834 0	1375 1393 2	610 626 0	911 967 0
96.859 3700/3820	<i>Klebsiella oxytoca</i> 10-5248	805 834 0	1374 1393 2	610 626 0	911 967 0
96.859 3700/3820	<i>Klebsiella oxytoca</i> 10-5243	805 834 0	1373 1393 2	610 626 0	912 967 0
96.754 3696/3820	<i>Klebsiella oxytoca</i> 10-5245	804 834 0	1368 1393 2	611 626 0	913 967 0
96.702 3694/3820	<i>Klebsiella oxytoca</i> OK-1	805 834 0	1368 1393 2	609 626 0	912 967 0
96.702 3694/3820	<i>Klebsiella oxytoca</i> KA-2	805 834 0	1367 1393 2	610 626 0	912 967 0
96.181 2745/2854	<i>Enterococcus faecalis</i> SDVK1A	799 834 0	1365 1394 2	581 626 0	
96.181 2745/2854	<i>Citrobacter freundii</i> UCI 32	798 834 0	1366 1394 2	581 626 0	

The proxytype calculation for *Klebsiella michiganensis*, based on four sequences published with the original description.

ANI_cutoff values in Taxonomy

<i>Acetobacter pasteurianus</i>		92.5
<i>Acinetobacter pittii</i>		92.5
<i>Aeromonas allosaccharophila</i>		95.0
<i>Aeromonas veronii</i>		94.0
<i>Bacillus subtilis</i>		93.3
<i>Delftia acidovorans</i>		98.5
<i>Delftia tsuruhatensis</i>		98.5
<i>Helicobacter pylori</i>		91.0
<i>Klebsiella michiganensis</i>		93.5
<i>Listeria monocytogenes</i>		92.4
<i>Mycobacterium africanum</i>		99.9
<i>Mycobacterium bovis</i>		99.9
<i>Mycobacterium tuberculosis</i>		99.9
<i>Prochlorococcus marinus</i>		78.0
<i>Raoultella ornithinolytica</i>		96.5
<i>Raoultella planticola</i>		96.5
<i>Rhodococcus fascians</i>		80.0

Some current species concepts span much more (or much less) than the default rule-of-thumb 96% ANI. We can set this value on a species-by-species basis in the taxonomy database.

(2) <i>Escherichia coli</i>	→ <i>Raoultella ornithinolytica</i>
(1) <i>Klebsiella pneumoniae</i>	→ <i>Raoultella planticola</i>
(1) <i>Salmonella enterica</i>	→ <i>Raoultella planticola</i>
(1) <i>Shigella boydii</i>	→ <i>Klebsiella michiganensis</i>
(1) <i>Shigella sonnei</i>	→ <i>Klebsiella michiganensis</i>
(2) <i>Raoultella ornithinolytica</i>	→ <i>Raoultella planticola</i>
(1) <i>Raoultella ornithinolytica</i>	→ <i>Raoultella terrigena</i>
(1) <i>Klebsiella oxytoca</i>	→ <i>Raoultella ornithinolytica</i>
(1) <i>Klebsiella sp.</i>	→ <i>Klebsiella oxytoca</i>
(43) <i>Klebsiella oxytoca</i>	→ <i>Klebsiella michiganensis</i>
(2) <i>Klebsiella sp.</i>	→ <i>Klebsiella michiganensis</i>

Counts of genomes with taxonomic updates in the new model. The first set is egregious misidentifications, where the submitted identification falls outside of the *Raoultella*/*Klebsiella* clade. The second set is misidentifications within the clade. And the final set contains updates necessitated by the description of a new species (*Klebsiella michiganensis* Saha et al. 2013, pmid 23053492).

```

LOCUS      JHQH01000000      39 rc      DNA      linear      BCT 18-JUN-2014
DEFINITION Raoultella planticola, whole genome shotgun sequencing project.
ACCESSION  JHQH00000000

SOURCE     Raoultella planticola (previously Raoultella ornithinolytica)

COMMENT    ##Taxonomic-Update-Data-START##
           Query          :: GCA_000648315.1
           New Taxon      :: Raoultella planticola
           Subject        :: GCA_000734435.1
           Status         :: type
           Identity       :: 99.440%
           Query Coverage :: 90%
           Subject Coverage :: 90%

           Query          :: GCA_000648315.1
           Old Taxon      :: Raoultella ornithinolytica
           Subject        :: GCA_000367425.1
           Status         :: proxytype
           Identity       :: 96.234%
           Query Coverage :: 86%
           Subject Coverage :: 91%

```

This genome was submitted as *Raoultella ornithinolytica*, but should actually be identified as *Raoultella planticola*. The structured comment presents the ANI neighboring data that supports this taxonomic correction.

```

LOCUS      JNPC01000000          57 rc    DNA        linear    BCT 11-JUN-2014
DEFINITION Raouttella ornithinolytica, whole genome shotgun sequencing
project.
ACCESSION  JNPC00000000

SOURCE     Raouttella ornithinolytica (previously Escherichia coli)

COMMENT    ##Proxytype-Designation-Data-START##
           Taxon          :: Raouttella ornithinolytica
           Assembly       :: GCA_000703465.1
           Accessions     :: AF129447 AF303652 JX397961 AY134479 AF303618
                           DQ227492 AY307386 AY134485 A008149
           Identity       :: 99.505%
           Counts         :: 6638/6671

           ##Taxonomic-Update-Data-START##
           Query          :: GCA_000703465.1
           New Taxon      :: Raouttella ornithinolytica
           Subject        :: GCA_000703465.1
           Status         :: proxytype
           Identity       :: 100.000%
           Query Coverage :: 100%
           Subject Coverage :: 100%

           Query          :: GCA_000703465.1
           Old Taxon      :: Escherichia coli
           Subject        :: GCA_000690815.1
           Status         :: type
           Identity       :: 82.327%
           Query Coverage :: 38%
           Subject Coverage :: 43%

```

In this case, the misidentified genome itself is one of the co-proxytype genomes of the destination species name (*Raouttella ornithinolytica*). We present the supporting data for the proxytype designation in a separate structured comment. After that, the first block of the taxonomic update comment is a self-comparison.

```

LOCUS      AYM101000000          68 rc   DNA       linear   BCT 18-APR-2014
DEFINITION Klebsiella michiganensis Hlg, whole genome shotgun sequencing project.
ACCESSION  AYM100000000

SOURCE     Klebsiella michiganensis (previously Klebsiella oxytoca)

COMMENT    ##Taxonomic-Publication-Data-START##
           Name      :: Klebsiella michiganensis
           Authority  :: Saha et al. 2013
           Effective Date :: 06-OCT-2012
           Valid Date  :: 01-MAR-2013
           Type        :: strain 14; DSM 2544; ATCC BAA-2403
           Accessions  :: JQ070300 JQ269337 JQ284304 JQ990329
           Authors     :: Saha,R., Farrance,C.E., Verghese,B.,
                        Hong,S. and Donofrio,R.S.
           Title       :: Klebsiella michiganensis sp. nov., A New
                        Bacterium Isolated from a Tooth Brush Holder
           Journal     :: Curr. Microbiol. 66 (1), 72-78 (2013)
           pmid        :: 23053492

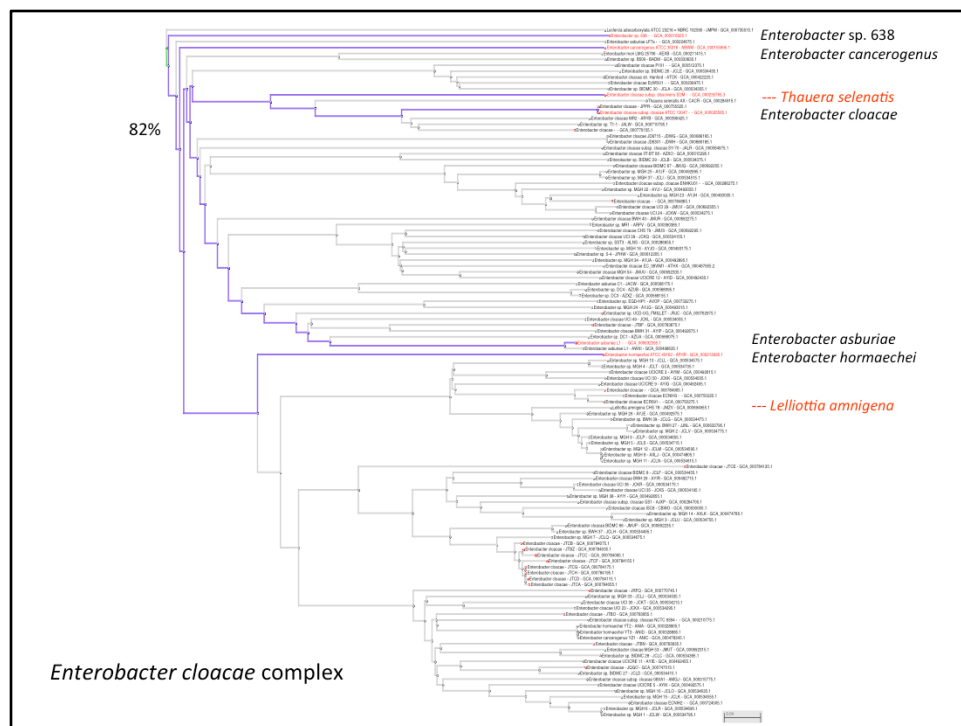
           ##Proxytype-Designation-Data-START##
           Taxon       :: Klebsiella michiganensis
           Assembly    :: GCA_000633235.1
           Accessions  :: JQ070300 JQ269337 JQ284304 JQ990329
           Identity    :: 99.347%
           Counts      :: 3804/3829

           ##Taxonomic-Update-Data-START##
           Query       :: GCA_000633235.1
           New Taxon   :: Klebsiella michiganensis
           Subject     :: GCA_000633235.1
           Status      :: proxytype
           Identity    :: 100.000%
           Query Coverage :: 100%
           Subject Coverage :: 100%

           Query       :: GCA_000633235.1
           Old Taxon   :: Klebsiella oxytoca
           Subject     :: GCA_000507385.1
           Status      :: proxytype
           Identity    :: 92.764%
           Query Coverage :: 84%
           Subject Coverage :: 85%

```

Here is the supporting information for the designation of a proxytype genome for a species that is newly described in the literature. It would be useful to display some structured data about the new species.



The *Enterobacter cloacae* complex spans 82% ANI, a very broad range. These are medically significant bacteria, and clinicians are trained to call everything here *Enterobacter cloacae*. A few genomes from type were submitted with different names (*Enterobacter asburiae*, *Enterobacter hormaechei*) and *Enterobacter cancerogenus* just outside of the *cloacae* complex. There are two interlopers. The WGS genome CACR (submitted as *Thauera selenatis*) was massively contaminated with sequence from *Enterobacter cloacae*, and has been suppressed. The WGS genome JMZV (submitted as *Lelliottia amnigena*) appears to be a simple misidentification.

proxytype table for <i>Lelliottia anguina</i>														
97.022 12347/12726	Enterobacter sp. 638	628 641.0	625 637.0	1241 1320	1281 1350.0	1401 1402.0	1787 1830.0	598 615.0	913 926.0	1612 1714.1	1445 1449.1	726 742.0		
94.231 7841/8321	Enterobacter cloacae P301	608 641.0	586 637.0			1381 1402.0	1747 1830.17	560 615.0	895 923.0		1424 1449.1	638 724.0		
94.198 7841/8324	Enterobacter cloacae EcW501	606 641.0	586 637.0			1381 1402.0	1747 1830.17	560 615.0	898 926.0		1424 1449.1	637 724.0		
94.128 7860/8293	Enterobacter aerogenes EA1509E	602 641.0	584 637.0			1379 1402.0	1738 1814.10	546 612.0	887 926.0		1421 1449.1	630 712.0		
93.767 7789/8423	Enterobacter isolatus C1	602 638.0	587 637.0	1137 1309.0		1379 1402.0	1751 1829.16		896 926.0	116 133.1	1421 1449.1			
93.777/8302	Enterobacter sp. MGII 16	599 638.0	586 637.0	1137 1320.0		1374 1402.0	1756 1830.18		897 926.0		1421 1449.1			
93.699 7217/7697	Klebsiella pneumoniae subsp. pneumoniae	600 642.0	578 637.0			1374 1402.0	1732 1831.11		888 926.0		1420 1449.1	620 710.0		
93.686 7217/7697	Klebsiella pneumoniae subsp. pneumoniae MGH 2675R	600 642.0	579 637.0			1374 1402.0	1733 1831.11		888 926.0		1420 1449.1	617 710.0		
93.658 7295/7789	Enterobacter cloacae CH 79		589 637.0	1134 1320.0		1378 1396.0	1760 1830.18		897 926.0	113 131.2	1424 1449.1			
93.658 7312/7697	Klebsiella pneumoniae subsp. pneumoniae KPN9127	600 642.0	577 637.0			1373 1402.0	1731 1831.8		888 926.0		1415 1449.1	619 710.0		
93.653 7864/8429	Enterobacter cloacae UCCE 12	599 638.0	586 637.0	1135 1320.0		1380 1396.0	1762 1830.18		896 926.0	113 131.2	1423 1449.1			
93.647 7208/7697	Klebsiella pneumoniae subsp. pneumoniae KPN9208	600 642.0	577 637.0			1374 1402.0	1732 1831.11		888 926.0		1420 1449.1	617 710.0		
93.647 7208/7697	Klebsiella pneumoniae subsp. pneumoniae KPN9110	600 642.0	577 637.0			1374 1402.0	1732 1831.11		888 926.0		1420 1449.1	617 710.0		
93.647 7208/7697	Klebsiella pneumoniae subsp. pneumoniae KPN911	600 642.0	577 637.0			1374 1402.0	1732 1831.11		888 926.0		1420 1449.1	617 710.0		
93.647 7208/7697	Klebsiella pneumoniae 30684/NIST28.2	600 642.0	577 637.0			1374 1402.0	1732 1831.11		888 926.0		1420 1449.1	617 710.0		
93.634 7207/7697	Klebsiella pneumoniae subsp. pneumoniae G613	600 642.0	578 637.0			1373 1402.0	1731 1831.11		888 926.0		1420 1449.1	617 710.0		
93.634 7207/7697	Klebsiella pneumoniae subsp. pneumoniae KPN9204	600 642.0	577 637.0			1374 1402.0	1732 1831.11		888 926.0		1420 1449.1	617 710.0		
93.632 7100/7679	Klebsiella pneumoniae subsp. pneumoniae NTJH K250	599 642.0	578 637.0			1373 1402.0	1735 1831.8		888 926.0		1419 1449.1	618 710.0		
93.621 7208/7697	Klebsiella pneumoniae JH45	600 642.0	577 637.0			1373 1402.0	1732 1831.11		888 926.0		1420 1449.1	617 710.0		
93.593 7187/7679	Klebsiella pneumoniae CG43	598 642.0	577 637.0			1373 1402.0	1734 1831.8		888 926.0		1419 1449.1	618 710.0		

31

Lelliottia amnigena NBRC 105700 ANI neighboring table

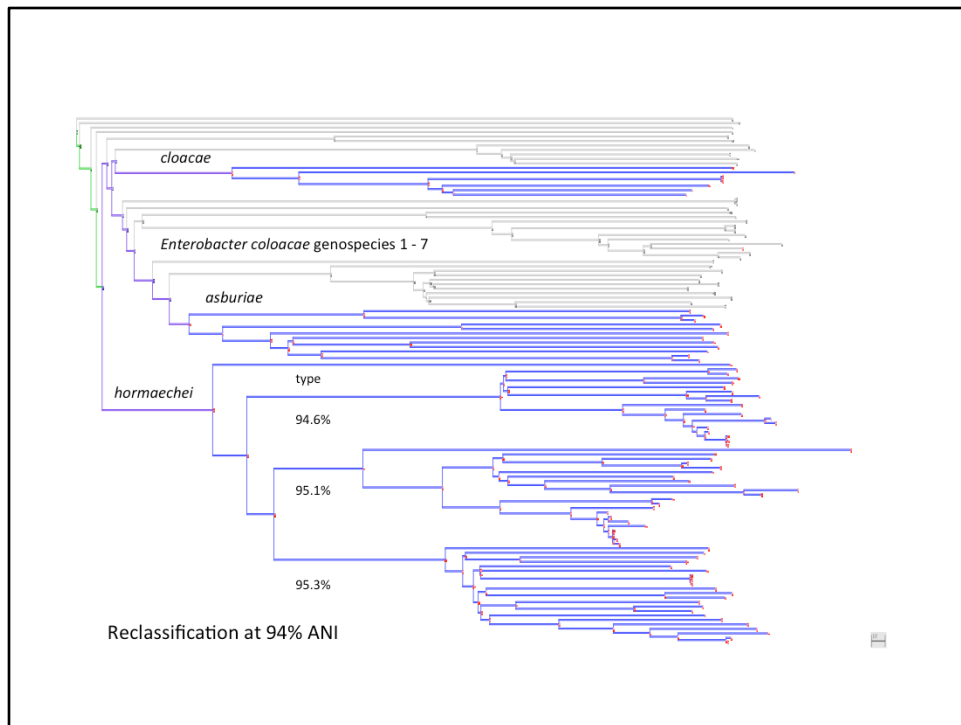
ANiB	<i>L. amnigena</i> NBRC 105700	<i>Enterobacter</i> sp. 638	<i>E. hormaechei</i> ATCC 49162	<i>L. amnigena</i> CHS 78	<i>E. asburiae</i> L1	<i>E. cloacae</i> ATCC 13047	<i>E. cloacae</i> EcWSU1	<i>E. cancerogenus</i> ATCC 35316	<i>E. aerogenes</i> EA1509E	<i>K. pneumoniae</i> MGH 78578
<i>L. amnigena</i> NBRC 105700	---	93.64	82.25	82.26	82.45	82.19	82.11	82.25	77.67	77.68
<i>Enterobacter</i> sp. 638	93.65	---	82.44	82.53	82.57	82.27	82.29	82.37	77.83	77.89
<i>E. hormaechei</i> ATCC 49162	82.22	82.20	---	94.12	87.40	86.66	86.16	86.13	78.56	79.27
<i>L. amnigena</i> CHS 78	82.35	82.50	94.24	---	87.47	86.90	86.29	86.19	78.93	79.55
<i>E. asburiae</i> L1	82.50	82.55	87.45	87.46	---	88.41	88.26	86.42	78.93	79.30
<i>E. cloacae</i> ATCC 13047	82.07	82.09	86.60	86.72	88.26	---	87.47	85.58	78.45	79.60
<i>E. cloacae</i> EcWSU1	82.17	82.19	86.21	86.20	88.20	87.58	---	85.75	78.42	78.73
<i>E. cancerogenus</i> ATCC 35316	82.22	82.27	86.18	86.18	86.41	85.76	85.79	---	78.68	79.17
<i>E. aerogenes</i> EA1509E	77.85	77.80	78.66	78.81	78.81	78.46	78.35	78.64	---	84.51
<i>K. pneumoniae</i> MGH 78578	77.73	77.67	79.20	79.39	79.12	79.23	78.58	78.97	84.33	---

unpublished draft genome from Fujita Nobuyuki, ANiB from JSpecies

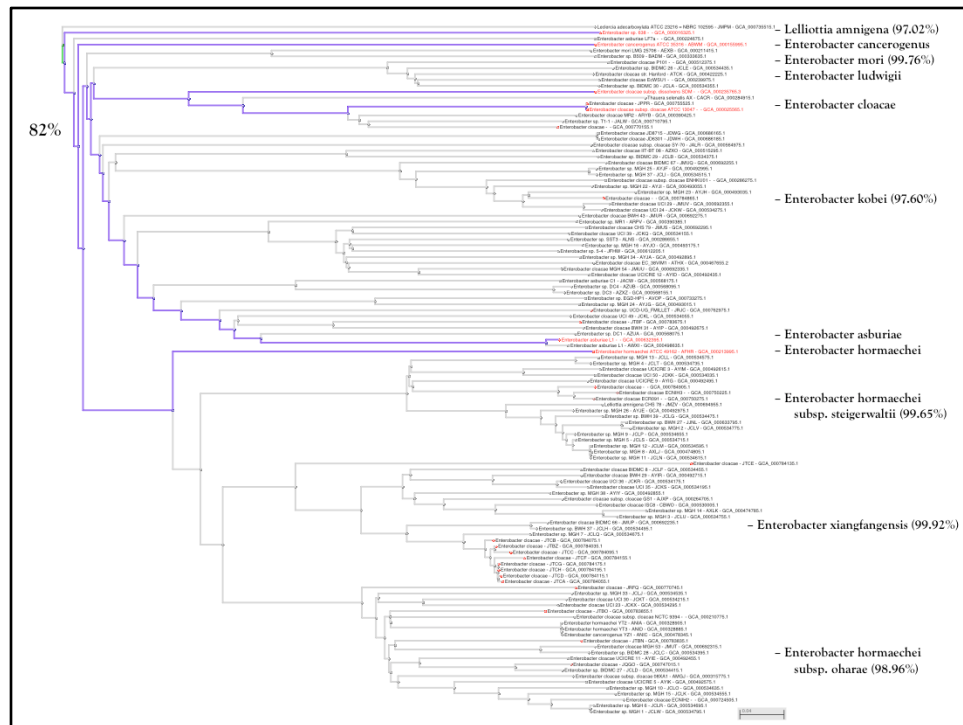
Fortuitously, a bioproject was registered in 2012 for the type genome of *Lelliottia amnigena*. This genome has not been finished, but the submitter agreed to share the ANI neighboring statistics of their current assembly.

The type genome of *Lelliottia amnigena* is indeed closest to *Enterobacter* sp. 638 (at 93.64% ANI). Note that our current genome submitted as *Lelliottia amnigena* is closest to the type of *Enterobacter hormaechei* (at 94.2% ANI).

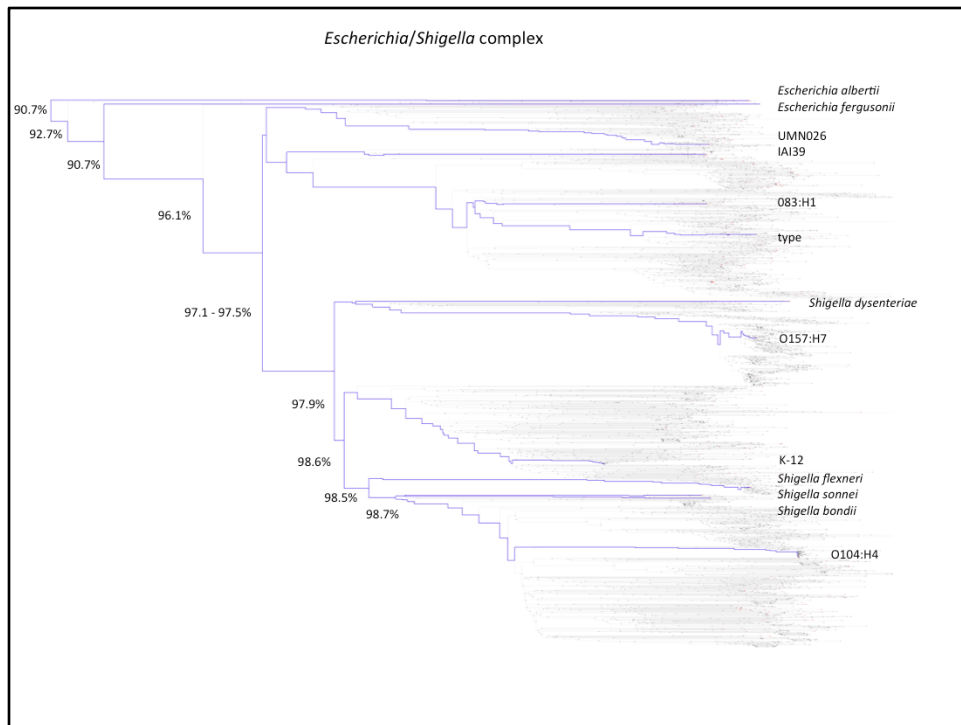
This provides strong support for the proxtype analysis outlined above.



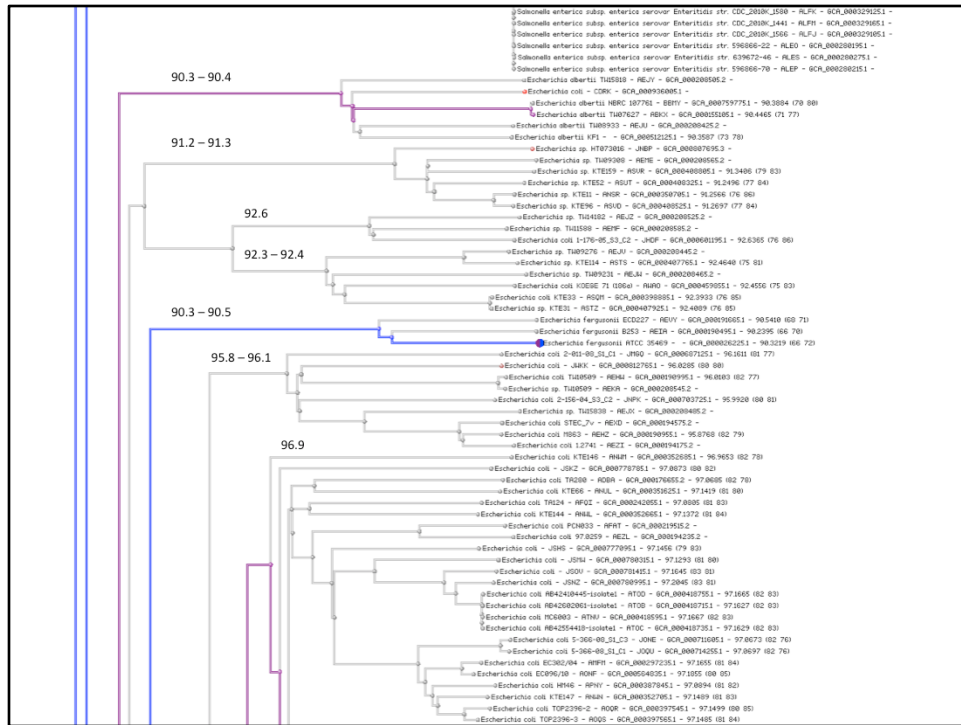
Enterobacter cloacae complex partitioned at 94% ANI with respect to the existing genomes from type. Enterobacter cloacae sensu stricto is represented by a very small clade. Several clades of genomes in the cloacae complex are less than 94% ANI from the three type genomes, and could be given informal clade names.



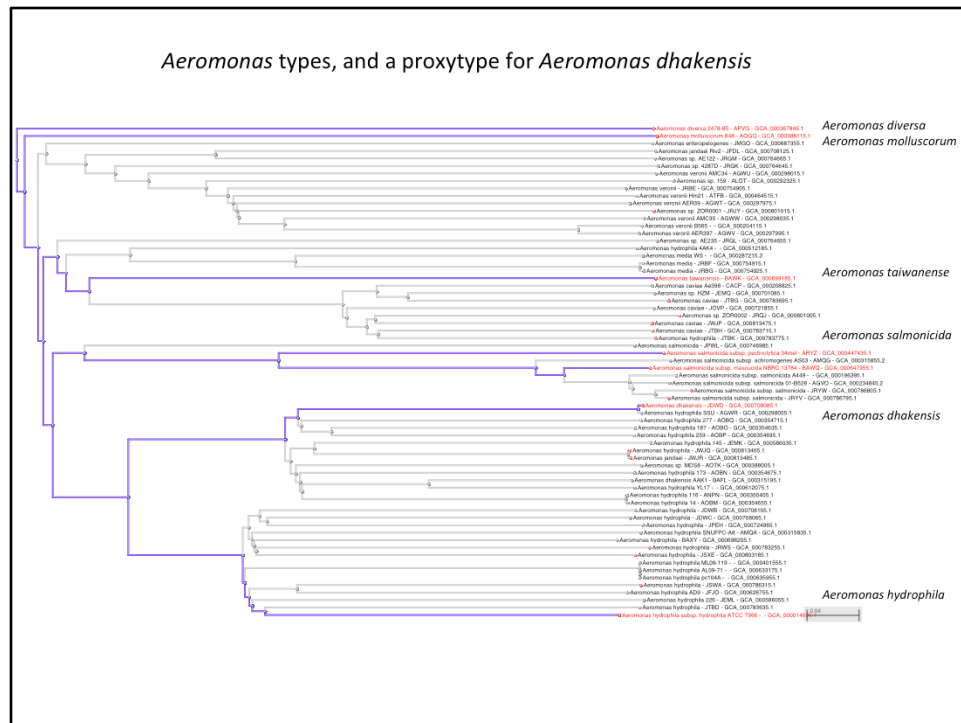
If we compute proxytypes for all of the species of *Enterobacter* without genomes from type, formal names can be assigned to several more clades. These include *Enterobacter mori*, *Enterobacter kobei*, *Enterobacter xiangfangensis*, and two subspecies of *Enterobacter hormaechei* that are effectively but not validly published. Note that the species *Enterobacter hormaechei* is paraphyletic in this labeling.



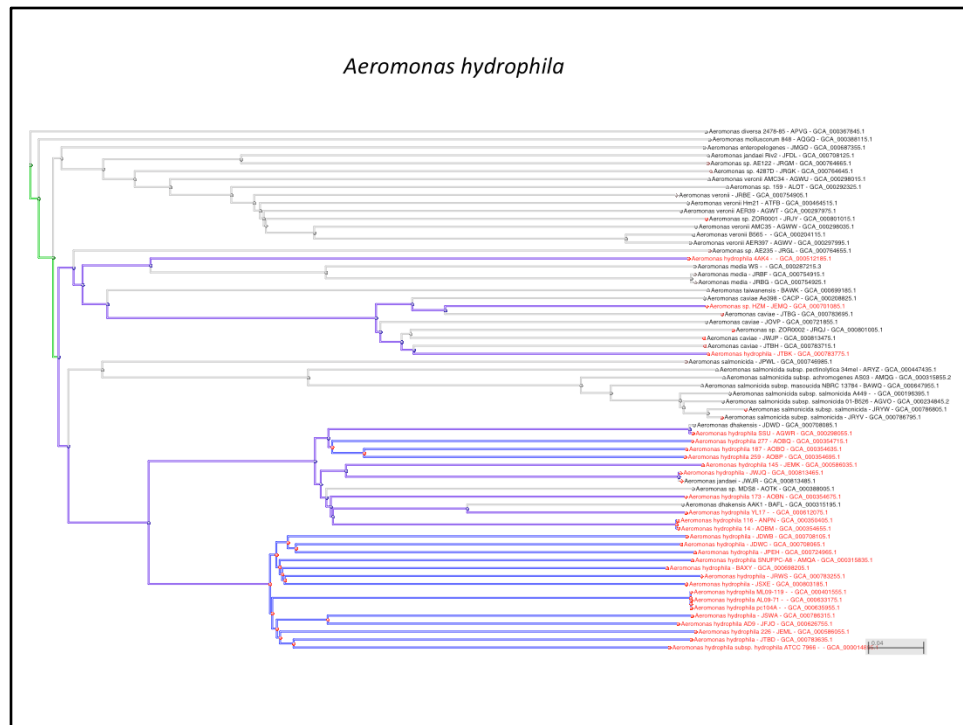
Some species are known to have taxonomic issues that can only be resolved in the literature. *Escherichia coli* contains all four species of *Shigella*, whose names have been conserved by a ruling of the Judicial Commission. This figure shows the 3000 genomes of *E. coli* in the kmer tree, together with *Shigella* and the sibling species *E. fergusonii* & *E. albertii*. The vast majority of *coli* genomes cluster within 97% ANI, with a small clade at 96.1% ANI. There is an ANI/kmer statistic anomaly at the top of the figure, highlighted on the next slide.



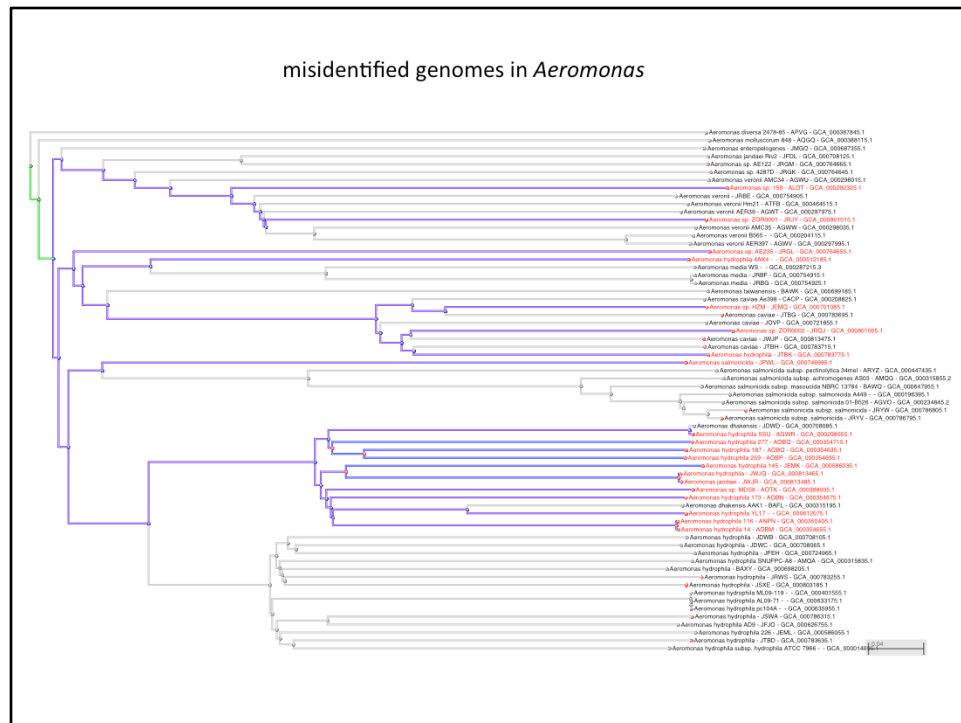
The clade of genomes between *Escherichia albertii* and *Escherichia fergusonii* in this figure show an ANI/kmer anomaly. The kmer scores place them further from *Escherichia coli* than the ANI scores do. Some as yet undetermined characteristic of these genomes affects the ANI & kmer statistics differently. This situation is rare.



Aeromonas has become something of a test case for ANI-based genomic taxonomy in the literature. Here are the genomes from type for *Aeromonas* in GenBank as of 2014.



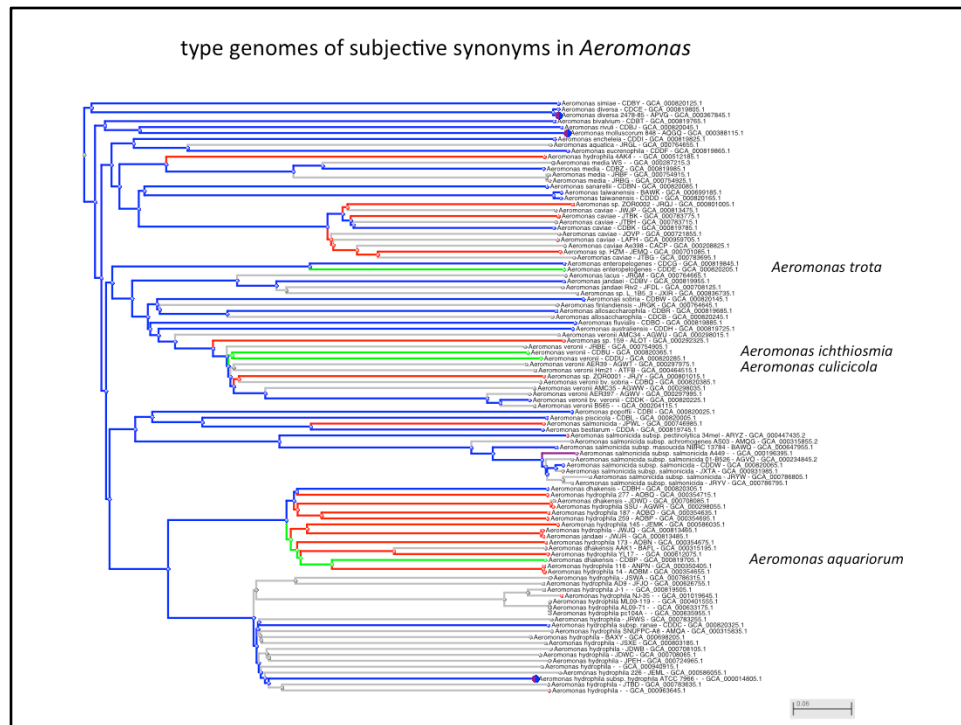
This is the kmer tree distribution of genomes submitted as *Aeromonas hydrophila*.



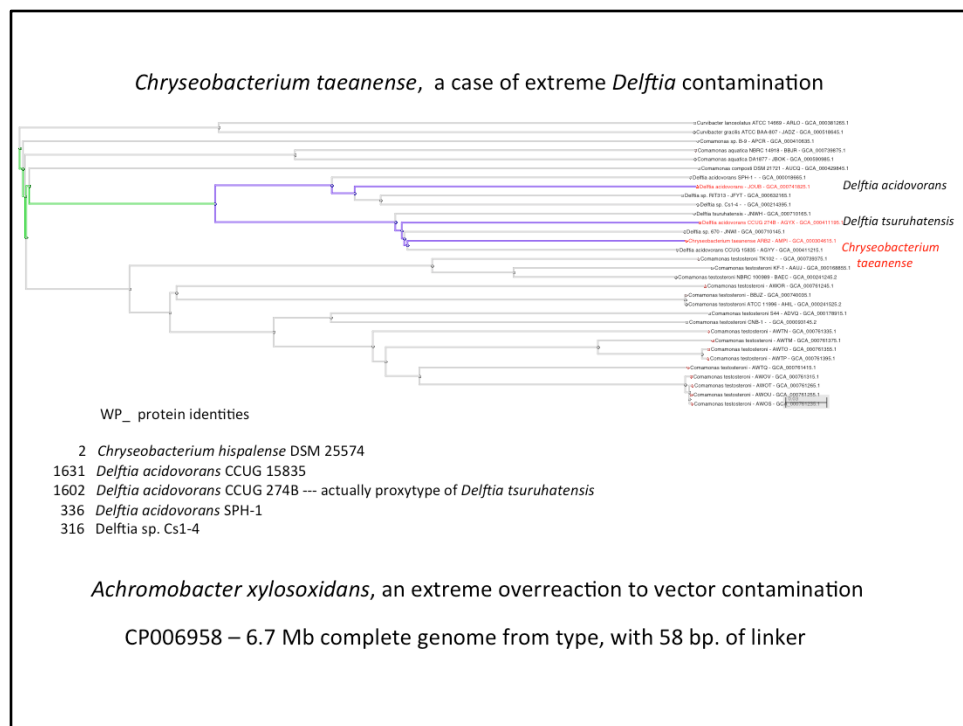
These are the *Aeromonas* genomes which would be given new taxonomic identifications based on the ANI analysis.

The large group at the bottom of the figure get new names due to the promotion of *Aeromonas hydrophila* subsp. *dhakensis* to species level.

(If those had been submitted with trinomials to begin with that would have happened automatically.)

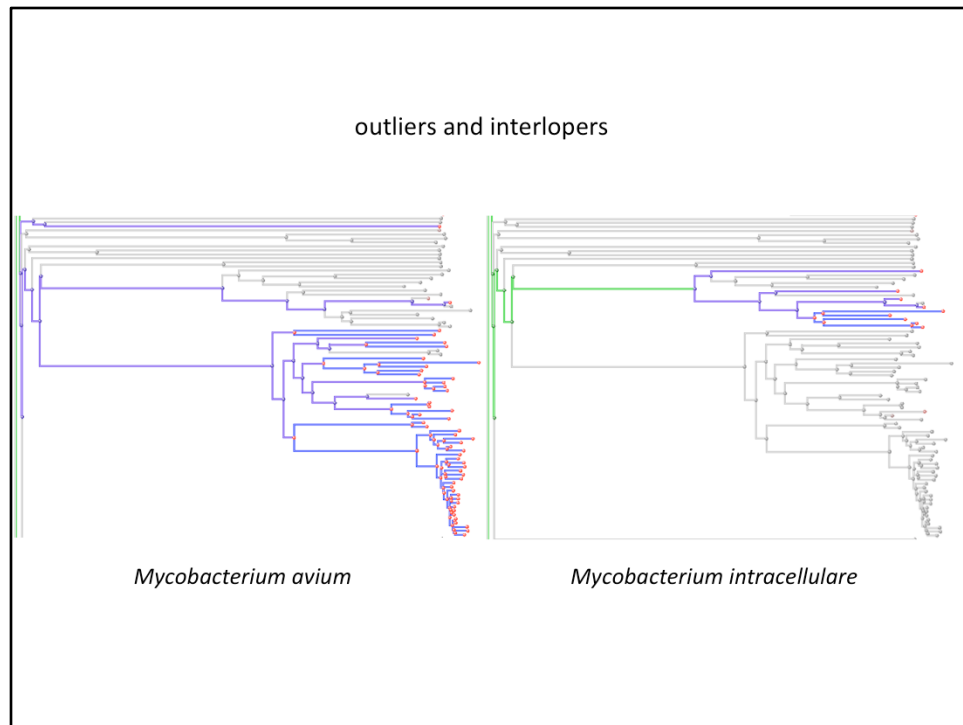


We flag type material from heterotypic synonyms separately from the primary type material. Traces leading to genomes from type strains of heterotypic synonyms are shown in green above. These placements in the kmer tree (and the corresponding ANI neighboring tables) confirm that it is reasonable to synonymize *Aeromonas trota* with *Aeromonas enteropelogenes*, to synonymize both *Aeromonas ichthiosmia* & *Aeromonas culicicola* with *Aeromonas veronii*, and to synonymize *Aeromonas aquariorum* with *Aeromonas dhakensis*.

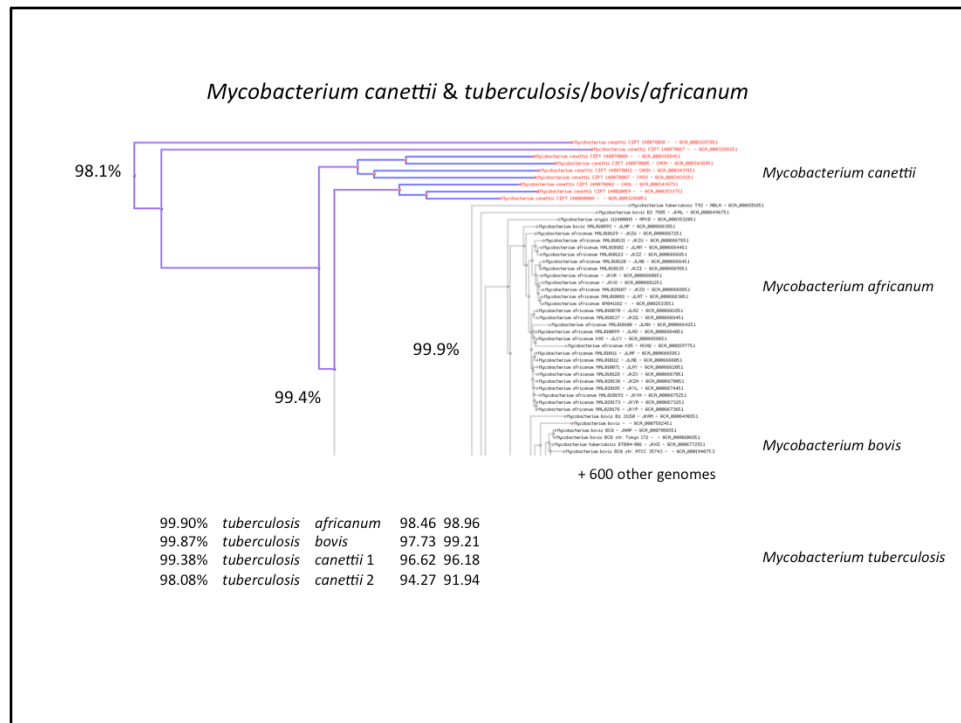


The elephant in the room here is contamination – we have never done a very good job of screening our bacterial genomes for contamination from other genomes. We identified several cases of massive contamination by investigating type genomes which appeared to be misplaced in the kmer tree. The putative type genome for *Chryseobacterium taeanense* had two proteins with sequences identical to *Chryseobacterium hispalense*, but 3800 proteins with sequences identical to *Delftia* proteins. Turns out *Delftia* is a common contaminant of sequencing reagents. We currently have no type sequences from *Delftia*, but we can assign proxytype genomes for *Delftia acidovorans* & *Delftia tsuruhatensis*.

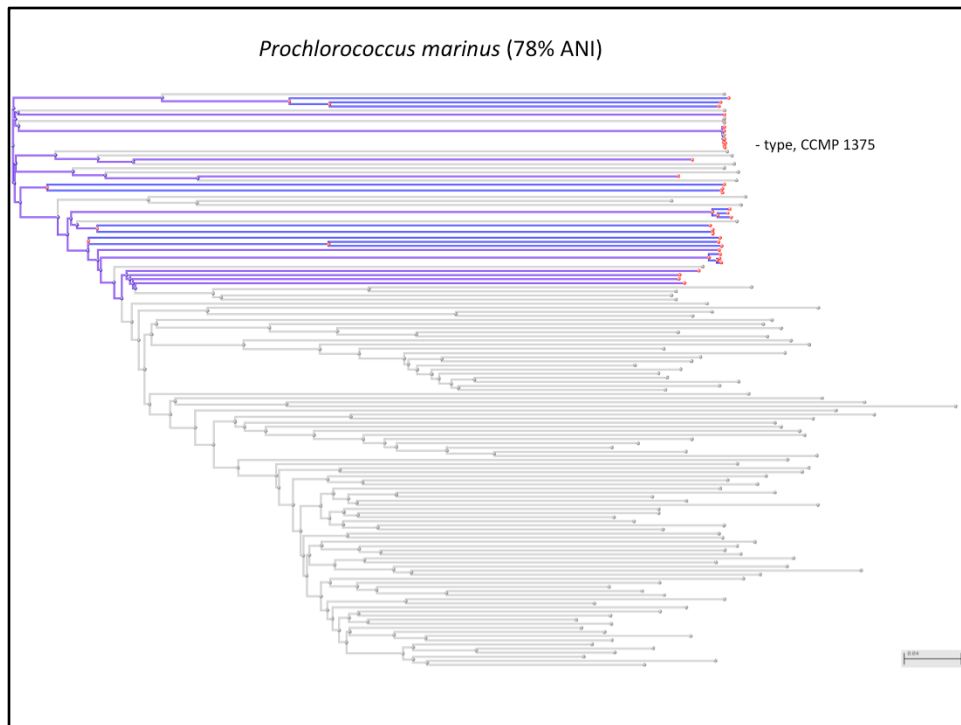
(The proxytype genome for *Delftia tsuruhatensis* was submitted as *Delftia acidovorans*, and would be renamed under the new system.) We have found several type genomes with enough contamination to misplace them in the kmer tree (as above, and with *Thauera selenatis* on slide 25). On the other extreme, a complete type genome was suppressed for 58 bp of linker when we could not get a response from the submitter. It has since been UNVERIFIED with a misc_feat comment on the linker contamination.



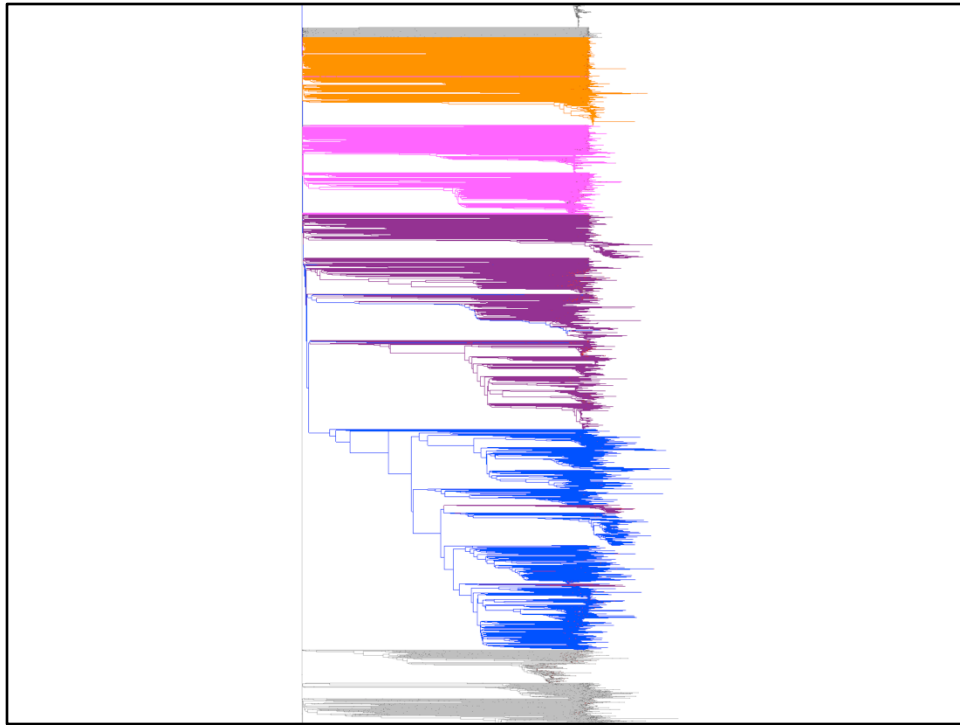
The *Mycobacterium avium*/intracellulare complex separates cleanly into clades in the kmer tree, with a few outliers & interlopers which may be misidentifications.



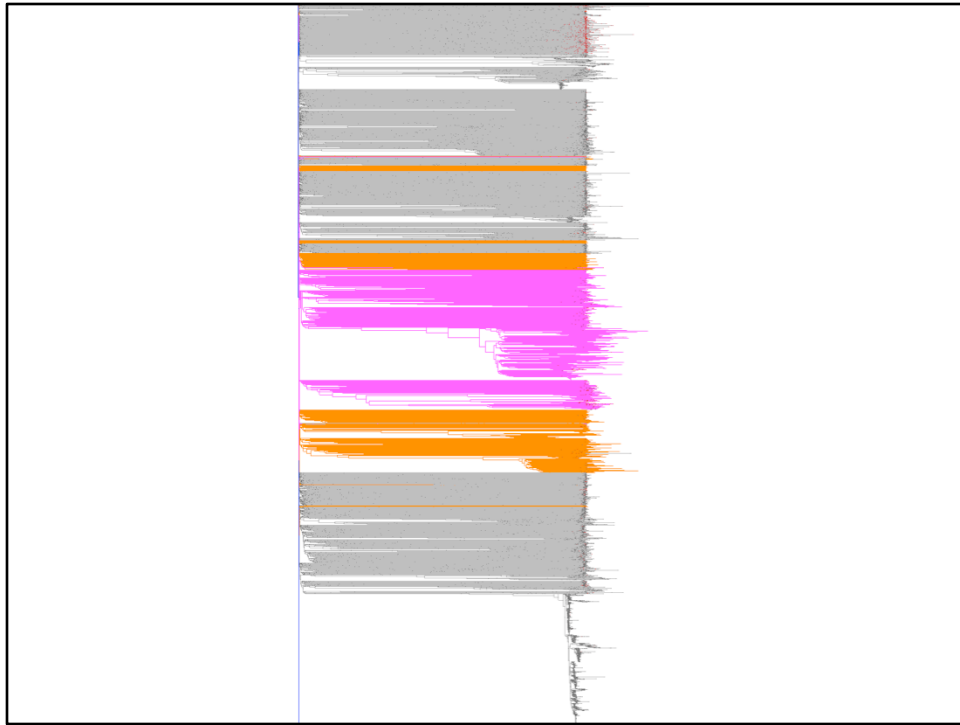
The *Mycobacterium tuberculosis* complex, on the other hand, is a known taxonomic problem case. *Mycobacterium canettii* is a paraphyletic species which is effectively but not validly published. *Mycobacterium africanum*, *bovis* & *tuberculosis* separate more or less cleanly into clades, but all of the genomes are within 99.9% ANI of each other. Thankfully, there are very few cases of species with serious taxonomic issues like this, and we can handle these as special cases.



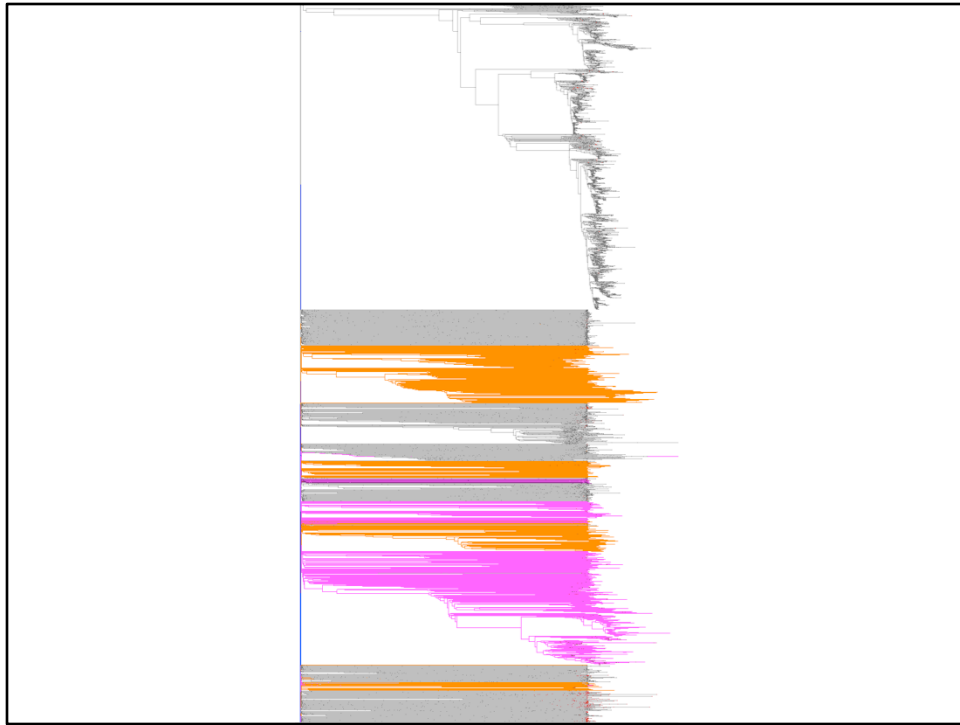
A few species span very very large ranges of ANI.



In the genome workbench it is easy to add metadata to the tree via attribute files. This tree shows the nested (and prioritized) queries genus=Escherichia > family=Enterobacteriaceae > class=Gammaproteobacteria > phylum=Proteobacteria. The purple patches nested within Escherichia are Shigella (aka STEC – shigatoxigenic Escherichia coli). The rest next nicely in spite of their placement near the spine of the tree, where we wouldn't expect to have much confidence in the topology.



Other clades of gammaproteobacteria (pink) and proteobacteria (orange) map elsewhere with respect to the deep branches of the kmer tree.



and this is the last of them.
Firmicutes above.