Comparative genomic and metagenomic investigations of the *Corynebacterium tuberculostearicum* species complex reveals potential mechanisms underlying associations to skin health and disease

Rauf Salamzade[1], Mary Hannah Swaney[1], Lindsay R. Kalan[1,2]
[1]Department of Medical Microbiology and Immunology, School of Medicine and Public Health, University of Wisconsin, Madison, WI, USA
[2]Department of Medicine, School of Medicine and Public Health, University of Wisconsin, Madison, WI, USA

Address for Correspondence: Lindsay Kalan, 1550 Linden Dr, 6155 Microbial Sciences Building, Madison, WI, USA, 53706. lkalan@wisc.edu

ABSTRACT

Corynebacterium are a diverse genus and dominant member of the human skin microbiome. Recently, we reported that the most prevalent *Corynebacterium* species found on skin – including *Corynebacterium tuberculostearicum* and *Corynebacterium kefirresidentii* – comprise a narrow species complex despite the diversity of the genus. Here, we apply high-resolution phylogenomics and comparative genomics to describe the structure of the *C. tuberculostearicum* species complex. We find this species complex is missing a fatty acid biosynthesis gene family which is often found in multi-copy in approximately 99% of other *Corynebacterium* species. Conversely, this species complex is enriched for multiple genetic traits, including a gene encoding for a collagen-like peptide. Further, through metagenomic investigations, we find that one species within the complex, *C. kefirresidentii*, increases in relative abundance during atopic dermatitis flares and show that most members of this species possess a colocalized set of putative virulence genes.

INTRODUCTION

Several species of *Corynebacterium* have been isolated from human skin, often associated with diseases such as diphtheria and regarded as opportunistic pathogens[1,2]. Only recently, with the application of metagenomics to comprehensively survey the skin microbiome, have we begun to develop a better understanding of this understudied genus and assess mechanisms underlying their basis as core constituents of skin microbiomes[3–5].

Regarded as a lipophilic and disease-associated species, *C. tuberculostearicum* was first named and associated with cases of diphtheria in 1984[1], later found to be associated with rhinosinusitis[6], and more recently shown to elicit specific inflammatory signaling pathways of skin cells[5]. While it was noted that *C. tuberculostearicum* could be closely related to other *Corynebacterium* species when it was first validated as a species in 2004[7], this analysis was based on 16S sequences which provide limited resolution for such investigations. Later work suggested that misclassification of *C. tuberculostearicum* is likely prevalent by clinical biochemical assays and that the species is potentially underreported[8].

We recently found that the three most prevalent species of *Corynebacterium* in healthy skin microbiomes from a metagenomic survey in our lab[9], currently classified as *Corynebacterium tuberculostearicum*, *Corynebacterium kefirresidentii*, and *Corynebacterium aurimucosum* type E in GTDB[10], belong to a narrow clade, which includes at least one additional species[11]. Because the average nucleotide identity (ANI) between genomes in this clade is relatively high and the lowest value observed was at 88%, we refer to the clade as the *C. tuberculostearicum* species complex[10,11]. While *C. aurimucosum* and *C. tuberculostearicum* have been validated as species[7,12], reassessment of genomes classified as these species in NCBI's GenBank database by GTDB suggests that some are mislabelled[10]. *C. kefirresidentii* was only recently proposed as a new species and has not yet been validated by the International Committee on Systematics of Prokaryotes[13].

Isolates of *C. kefirresidentii* sampled from healthy human skin are lipophilic similar to *C. tuberculostearicum* and also grow better in the presence of compounds commonly found in sweat[9]. This finding is in accordance with our earlier observation that members of the *C. tuberculostearicum* species complex are often found at body sites regarded as moist[11]. In this study, we apply comparative genomics and metagenomics analytics to develop a better understanding of the genetic factors underlying the prevalence of the *C. tuberculostearicum* species complex on healthy skin and determine they lack key metabolic enzymes likely rendering them lipid-dependent. In addition, we assess the biogeographical distribution of individual species within the complex across different skin body sites and whether they associate with atopic dermatitis flares.

RESULTS

A high-resolution phylogeny of the *C. tuberculostearicum* species complex

A set of 26 genomes belonging to the *C. tuberculostearicum* species complex were used for phylogenomic investigation of the complex (Table S1). This set of genomes includes 22 genomes we had previously described as part of the complex[11]. In addition, we added three

recently reported metagenome-assembled genomes (MAGs) from skin predicted to correspond to novel *Corynebacterium* species[14] and the genome of a *C. kefirresidentii* we recently isolated from human skin[9]. All 26 genomes, including a total of 8 MAGs, were regarded as high completion and showed low rates of contamination by CheckM analysis[15] (Table S1).

A set of 1,250 single-copy-core (SCC) orthologs was identified amongst the 26 genomes belonging to the species complex and used to construct a high-resolution phylogeny following filtering of sites predicted to be affected by recent or ancestral recombination[16,17] (Figure 1). This phylogeny was found to be highly concordant with species classifications for the genomes by GTDB-tk[18] run with GTDB release 207[10], but not their taxonomic designations on GenBank (Table S1).

ANI between genomes belonging to different species ranged from a lower limit of ~88%[11] to nearly 95% (Figure S1). This suggests recent speciation occurring within the species complex, in particular within the subclade containing the species *Corynebacterium* sp900539985, *C. aurimucosum* type E*,* and *C. tuberculostearicum*, where genomes from different species can exhibit ANI greater than 94% to each other, which is slightly under the 95% ANI threshold commonly used to delineate species[19,20]. In addition, the ANI between genomes *C. spp.* SMGC_38 and *C. spp*. SMGC_337 was estimated as being between 93% and 95%, suggesting recent speciation between them. These two newly reported species[14], exhibited greater than 90% ANI to *C. kefirresidentii* genomes (Figure S1). The MAG proposed to correspond to a novel *Corynebacterium* species[14], *Corynebacterium spp.* SMGC_92, was found to exhibit high ANI to an established but currently unnamed species in GTDB, *C.* sp900539985, and designated as belonging to this species. Due to close phylogenetic placement and ANI estimates exceeding 95%, we further propose that GTDB species *C. tuberculostearicum* and *C. tuberculostearicum* type C should simply be regarded as one species (*C. tuberculostearicum*).

Overall, the phylogeny constructed using the core-genome with sites predicted to be affected by recombination removed, was well supported as indicated by high bootstrap values at ancestral nodes to species and sets of species (Figure 1). However, we were unable to confidently resolve whether *C. sp900539985* are more related to *C. tuberculostearicum*, as ANI analysis suggests, or *C. aurimucosum* (Figure 1, S1).

Unique genes and the biogeographic distribution of the *C. tuberculostearicum* species complex

To address potential biases predicting prevalence of the *C. tuberculostearicum* species complex associated with the metagenomic survey performed in our laboratory[4], we used StrainGE[21] to profile the presence of representative *Corynebacterium* genomes amongst published metagenomic datasets profiling the healthy skin microbiome[3] including a study profiling the skin microbiome in atopic dermatitis, a common pediatric allergic disease[22]. In concordance with our previous findings[11], we observed that all three species from the complex represented in the StrainGE database are highly prevalent in these additional metagenomic datasets relative to other *Corynebacterium* species (Figure 2A; Table S2). We further observed that *C. tuberculostearicum* and *C. kefirresidentii* were differentially abundant at distinct body sites. While *C. tuberculostearicum* were prevalent and found in high abundances at foot

115 associated body sites, in particular the toe web space, *C. kefirresidentii* were more prevalent in
116 nasal and surrounding alar crease sites (Figure 2B).
117 To understand the basis of the enrichment of the *C. tuberculostearicum* species
118 complex, we performed comparative genomics to determine if particular homolog groups were
119 found more commonly or less commonly than expected within the complex relative to the rest of
120 the genus (Table S3). A representative set of 187 *Corynebacterium* genomes was selected
121 based on dereplication[23] of those designated as belonging to the genus in GTDB at 95% ANI, to
122 approximate selection of a single representative genome per species.
123 We identified only two homolog groups which were statistically significant for being
124 absent in the species complex but highly conserved throughout the genus. One of these was
125 annotated as a short-chain dehydrogenase reductase and was absent in all four representative
126 genomes of the *C. tuberculostearicum* species complex but found at a median copy count of
127 three in 98.9% other *Corynebacterium*. Searching for highly conserved domains on the
128 consensus sequence of the homolog group revealed a more precise annotation of 3-oxoacyl-
129 ACP reductase FabG[24], thus suggesting a role in fatty acid biosynthesis[25].
130 A larger set of proteins were identified as being statistically enriched in the *C.*
131 *tuberculostearicum* species complex in relation to other *Corynebacterium*. We filtered for 26
132 homolog groups which were found to be nearly core to the species complex (found in $\geq$95% of
133 the 26 *C. tuberculostearicum* genomes). While the majority of these homolog groups lacked
134 reliable functional annotations, we identified three to putatively correspond to acetyltransferases
135 and an additional three to be factors in cellular regulation or translation. The largest homolog
136 group, found in only 3.8% of *Corynebacterium* outside of the species complex, was predicted to
137 encode a collagen-like peptide. Collagen-like proteins have previously been suggested to
138 function as adhesins of pathogens to host cells[26–29] and could correspond to an analogous
139 mechanism in the *C. tuberculostearicum* species complex for attaching to skin cells.
140
141 **Species specific partitioning into distinct moist body sites and association with atopic**
142 **dermatitis**
143
144 Because prior taxonomic classifications of genomes were incongruent with their
145 phylogenomic placement, it was previously impossible to accurately associate individual species
146 within the *C. tuberculostearicum* species complex with skin-related diseases. With our resolved
147 view of the species complex, we thus reinvestigated the metagenomic dataset from Byrd *et al.*
148 to assess the relative abundance of *C. tuberculostearicum*, *C. kefirresidentii*, and type E *C.*
149 *aurimucosum* at different stages of atopic dermatitis for a pediatric cohort. We found that *C.*
150 *kefirresidentii* was more frequently detected during flares, in comparison to baseline (p=1.5e-4)
151 and healthy control microbiomes (p=4.0e-4). Furthermore, in certain metagenomes sampled
152 during a flare, *C. kefirresidentii* exhibited a substantial increase in relative abundance (Figure
153 3A; S2).
154 To further affirm the association of *C. kefirresidentii* with atopic dermatitis flares, we also
155 applied an alignment-based approach in which we aligned metagenomic reads to the single-
156 copy core genes of the complex and measured coverage of species specific alleles. *C.*
157 *kefirresidentii* was more evidently present at higher abundances primarily in metagenomes from
158 flares and post-flare samplings (Figure 3B), which also tended to have the lowest viable

159 sequencing data after filtering metagenomes for quality, adapter sequences, and human
160 contamination (Figure S3). We further developed a ratio statistic to assess the abundance of *C.*
161 *kefirresidentii* to *C. tuberculostearicum* and type E *C. aurimucosum* individually within each
162 metagenome. *C. kefirresidentii* were more abundant than type E *C. aurimucosum* and *C.*
163 *tuberculostearicum* combined in 52.7% of metagenomes survey from flares, which was
164 significantly more than for baseline (25.7%; *p*=2.3e-4) and healthy control (21.1%; *p*=3.8e-5)
165 metagenomes (Figure 3C).
166 To understand what potential genetic factors might be contributing to the unique
167 association of *C. kefirresidentii* with atopic dermatitis flares, we performed comparative
168 genomics to identify traits which were lost or gained in their genomes relation to the closely
169 related *C. tubercuostearicum* and type E *C. aurimucosum* species. We found a total of 97
170 homolog groups were either enriched or depleted in *C. kefirresidentii* relative to *C.*
171 *tubercuostearicum* and type E *C. aurimucosum* (Table S4). *C. tubercuostearicum* and type E *C.*
172 *aurimucosum* feature a core thermonuclease that is homologous to the virulence factor *nuc*[30]
173 from *Staphylococcus aureus*, but missing in all *C. kefirresidentii.* In contrast, two distinct co-
174 located sets of genes, one predicted to encode for thiamine biosynthesis and the other encoding
175 molybdopterin-binding enzymes, were enriched in *C. kefirresidenttii* but missing or significantly
176 lacking in *C. tubercuostearicum* and type E *C. aurimucosum* genomes. Of note, molybdenum
177 enzymes have recently been discussed as important virulence factors for infection in a variety of
178 diverse pathogens[31,32].
179
180 **DISCUSSION**
181
182 The *C. tuberculostearicum* species complex features the three major species of
183 *Corynebacterium* found on human skin[11]. Despite this, fewer than 30 representative genomes
184 across all species belonging to the complex are present in the most recent GTDB release and
185 eight of the 26 genomes (30.8%) used in this study were MAGs. The lack of genomes for the
186 species complex, despite its prevalence on healthy skin, might be in part due to difficulties with
187 culturing. For instance, we recently found that a single *C. kefirresidentii* isolate grows best when
188 high concentrations of compounds found in both sebum and sweat are available[9].
189 Here we determine that the *C. tuberculostearicum* species complex appears to be
190 missing a fatty acid biosynthesis gene family that is often found in multi-copy in other
191 *Corynebacterium*, potentially explaining their lipophilic nature[7,8]. However, other
192 *Corynebacterium* species outside of this species complex
193 have also been characterized as lipophilic[9,33], suggesting convergent evolution with different
194 genetic factors underlying the shared trait in diverged species across the genus. Comparative
195 genomics further highlighted several genes encoding proteins that are specific to the species
196 complex, though the majority of these lack any functional annotation and future molecular
197 characterization will be essential to further our functional understanding of them. Interestingly,
198 one of these genes, core to the species yet present in less than 4% of other *Corynebacterium*,
199 encodes for a collagen-like peptide and could be involved in adherence to skin.
200 We further identify an increase in relative abundance of *C. kefirresidentii* in
201 metagenomes sampled during atopic dermatitis flares. Comparative genomics between *C.*
202 *kefirresidentii* with *C. tuberculostearicum* and *C. aurimucosum* type E genomes found that

203 multiple genes differentiate *C. kefirresidentii* from its two neighboring species. Most striking are
204 a set of five colocated molybdopterin-associated enzymes, which are commonly associated with
205 virulence[31,32].
206 Two ATCC isolates (ATCC-35693 / 2628 LB and ATCC-35694 / FPSA)[34], designated as
207 *C. tuberculostearicum* and originating from the original study first describing the species taken
208 from leprosy patients[1], have more recently been used in studies to associate the species to
209 rhinosinusitis[6] and elicitation of a specific host inflammatory signaling pathway[5]. To our
210 knowledge, these isolates lack genomes however and might currently be misclassified since
211 other species of *Corynebacterium* have been reported to similarly produce tuberculostearic acid,
212 the trait initially used to differentiate the species from others in the genus[7,12,35]. Identifying
213 genetic markers to distinguish species within the *C. tuberculostearicum* species complex and
214 continued efforts in whole genome and metagenomic sequencing will thus be critical to correctly
215 associate the contributions of particular species within the complex to diseases.
216
217
218 **ACKNOWLEDGMENTS**
219
220 This work was supported by grants from the National Institutes of Health (NIAID U19AI142720
221 and NIGMS R35GM137828 [L.R.K]). The content is solely the responsibility of the authors and
222 does not necessarily represent the official views of the National Institutes of Health. The authors
223 gratefully acknowledge J.Z. Alex Cheong, Dr. Nicole Lane Starr, Lucas van Dijk, Dr. Abigail
224 Manson, and Dr. Ashlee Earl for helpful discussions.
225
226 **METHODS**
227
228 **Genome selection and taxonomic reclassification**
229 We had previously identified a set of 22 genomes through an interactive process in
230 which we first identified the clade using phylogenomics as consisting of the five GTDB defined
231 species: *C. tuberculostearicum*, *C. tuberculostearicum*_C, *C. aurimucosum*_E, and *C.*
232 *kefirresidentii* (using release R202)[11]. Five additional genomes belonging to the clade were
233 identified in NCBI which were missing in GTDB release R202 and validated to belong to one of
234 the five species through GTDB-tk classification[18]. In our earlier study, we also found that three
235 recently published MAGs from skin microbiomes, proposed novel species of *Corynebacterium*
236 (based on GTDB release 88)[14], had >88% ANI to one of the prior defined 22 genomes. We thus
237 included these three genomes, in addition to the genome for a *C. kefirresidentii* isolate, LK1134,
238 we had isolated from healthy skin and profiled the growth dynamics off across different
239 concentrations of artificially synthesized sebum and sweat[9].
240 Systematic reclassification on the total set of 26 genomes proposed to belong to the *C.*
241 *tuberculostearicum* species complex was performed using GTDB-tk2 with the latest GTDB
242 release R207. All genomes were assigned to one of the five previously mentioned GTDB
243 species, except for the MAGs *C. spp.* SMGC_38 and *C. spp.* SMGC_337[14], which appear to still
244 correspond to novel species of close relation to *C. kefirresidentii*. Because MAGs were included,
245 the contamination and completeness of genomes was assessed using CheckM (v1.2.1) (Table
246 S1).

247 For comparative genomics of the *C. tuberculostearicum* species complex against other
248 species from the *Corynebacterium* genus, all 1,296 genomes in the genus from GTDB R207
249 were downloaded[10]. Genome dereplication was performed using dRep[23] with a secondary ANI
250 filter of 95% using fastANI[19]. This retained a set of 187 distinct genomes, including four
251 representatives of the *C. tuberculostearicum* species complex.
252
253 **Homolog group inference and functional annotation**
254 Gene calling was performed on genomes using prodigal[36] and OrthoFinder (v2.5.4)[16]
255 was subsequently used to perform homolog group detection independently amongst the set of
256 26 *C. tuberculostearicum* genomes and the genus-wide set of 187 representative
257 *Corynebacterium* genomes. For both analyses, the course set of initial orthogroups following
258 MCL were used instead of phylogenetic hierarchical orthogroups.
259 Functional annotation was only performed for homolog groups which comparative
260 genomics investigations suggested were significantly more prevalent or lacking within certain
261 phylogenetic clades. Annotation was primarily performed using individual protein sequences
262 from such homolog groups via the EggNOG mapper webserver run with default settings[37].
263 Consolidation of annotations was performed across proteins belonging to each homolog group.
264 For homolog groups predicted to be enriched or depleted in the *C. tuberculostearicum* species
265 complex relative to other *Corynebacterium*, we further attempted annotation by generating
266 consensus amino acid sequences for each homolog group and using NCBI's conserved domain
267 search[24] as well as Phyre2 to assess structural similarity to structurally characterized proteins[38].
268
269 **Phylogenomics of the *C. tuberculostearicum* species complex and the representative**
270 **selection of *Corynebacterium***
271 We identified 1,250 homolog groups as single-copy orthologs between the 26 *C.*
272 *tuberculostearicum* species complex genomes. Protein alignments were constructed using
273 MUSCLE (v5)[39] and converted to codon alignments using PAL2NAL[40]. RAxML (v8.2.12) was
274 then used to construct an initial maximum likelihood phylogeny on a concatenation of codon
275 alignments (1,239,366 bp) with the GTRCAT model and 1,000 bootstraps. Core codon
276 alignments were projected onto and ordered according to the reference genome *C.*
277 *kefirresidentii* FDAARGOS_1055 and, together with the RAxML phylogeny, used as input for
278 ClonalFrameML (v1.12)[17] to infer sites affected by recombination by first using a simple model,
279 with parameter emsim set to 100 and kappa (the relative rate of transitions to transversions) set
280 to 5.09664. For reference projection, gaps between single-copy orthologs were set to be the
281 length of distances in between corresponding genes in the reference genome and codon
282 alignments were reverse complemented if the gene was predicted to occur on the antistrand.
283 Non-core sites were specified as the gaps placed in between single copy core orthologs so as
284 to appropriately model the strength of linkage between them. Results from running
285 ClonalFrameML with a simple model were used to initialize values for R/theta, 1/delta, and nu
286 (set to 0.126997, 0.0027686, and 0.0583416, respectively) and run ClonalFrameML with a per-
287 branch model with the embranch_dispersion parameter set 0.1, to allow for greater dispersion in
288 parameters amongst branches. The initial alignment was filtered for sites predicted to be
289 affected by recombination and subsequently subset for 15,936 sites which lacked ambiguity for
290 any of the 26 genomes. This alignment was finally used to construct a high-resolution,

291 maximum likelihood phylogeny of the species complex using RAxML with the same modeling
292 strategy as applied initially.
293 For phylogeny construction of the full genus, we used GToTree[41] (v1.6.36) with the set
294 of single-copy gene sets specific to Actinomycetota (formerly Actinobaceria) and FastTree2
295 phylogeny inference[42].
296
297 **Statistical testing for enrichment or depletion of homolog groups in clades for**
298 **comparative genomics analyses**
299 Assessment of enrichment or depletion of homolog groups for two sets of genomes was
300 carried out using a two-sided permutation test based on mean copy-count differences with
301 100,000 resamples. This allowed us to detect homolog groups which were not merely
302 differentially present, but also those which had greater copy-counts in one set of genomes
303 relative to the other. A generalizable program for this testing is provided in the Github repository:
304 https://github.com/Kalan-Lab/Salamzade_etal_CtuberCompGen. Multiple testing correction was
305 performed for both such comparative genomics analyses using the Benjamini-Hochberg
306 procedure for P-value adjustment to control the false discovery rate. For homolog groups found
307 to be enriched in the four *C. tuberculostearicum* species complex representative genomes
308 relative to the complementary set of 183 *Corynebacterium* genomes, we further required that
309 they map to homolog groups deemed to be ≥95% near-core amongst the 26 *C.*
310 *tuberculostearicum* species complex genomes (there was a total of 1,871 near-core homolog
311 groups).
312
313 **Metagenomic detection of *Corynebacterium* strains and species**
314 Three independent sets of metagenomes profiling the microbiome of healthy skin and
315 skin affected by atopic-dermitis were processed using a pipeline previously described[4]. Briefly,
316 metagenomes were processed for quality and human contamination using fastp[43] and
317 KneadData (https://github.com/biobakery/kneaddata). Mock and negative control metagenomes
318 from the Swaney *et al.* 2022 study were excluded[4] as were 22 metagenomes from the Oh *et al.*
319 2016 dataset marked as "human depletion" or "whole genome amplification" under "Notes
320 (excluded for comparisons)" in their spreadsheet describing all samples[3]. Deeply sequenced
321 metagenomes from Oh *et al.* 2016 were randomly downsampled to 25 million read pairs using
322 seqtk (https://github.com/lh3/seqtk). In addition, as described previously[4], five metagenomes
323 from the Byrd *et al*. 2017 dataset[22] (3 flare and 2 post-flare) were excluded due to insufficient
324 sequencing data following processing. All symmetric metagenomes (e.g. metagenomes
325 corresponding to sampling the left and right sides of a body site) from the Oh *et al.* 2016 and
326 Byrd *et al.* 2017 metagenomes were individually considered.
327 Detection of strains or representative genomes was performed using StrainGST[21] with a
328 custom database constructed from complete *Corynebacterium* genomes from NCBI, as
329 described previously[11]. Briefly, this database featured 164 distinct representative genomes
330 following dereplication using programs for StrainGST (0.1+191.g7fcfbcd.dirty) database
331 creation[21]. These genomes were run through GTDB-tk with database release R207[10,18] to
332 generate consistent taxonomic classifications. One genome regarded as *Corynebacterium* in
333 RefSeq[44] at the time of database creation was classified as not part of the genus by GTDB-tk
334 and not observed in any of the metagenomic datasets. Another genome was classified as a

335 novel *Corynebacterium spp.* not yet given an identifier in GTDB R207. We counted the number
336 of distinct species detected in metagenomes from the three datasets and illustrated them as a
337 multi-bar graph on the phylogeny of 187 *Corynebacterium* by species name. Because *C.*
338 *tuberculostearicum* and *C. tuberculostearicum* type C classified genomes were found in this
339 study to phylogenetically group as one, we considered them as a single species and mapping of
340 detection counts for the *C. tuberculostearicum* representative genome in StrainGST was
341 allowed to the *C. tuberculostearicum* type C representative genome in the set of 187 genomes
342 used for phylogeny construction and comparative genomics. The species *C. sp900539985* from
343 the *C. tuberculostearicum* species complex, which was represented in the phylogeny of 187
344 genomes, was not represented in the StrainGST database as it lacks a complete genome
345 currently. Further, there was a single representative strain / genome for each of the three
346 species from the complex represented in the StrainGST database: *C. tuberculostearicum*, *C.*
347 *aurimucosum_E*, and *C. kefirresidentii*. The relative abundance estimation for these three
348 species was thus computed based on the relative abundance computed for the corresponding
349 representative strains / genomes by StrainGST (*rapct*) after updating to version 1.3.3, to
350 incorporate updates to how relative abundance was being computed relative to earlier versions.
351 *Confirmation of StrainGST results using an independent approach:* StrainGST is a k-mer
352 based approach to assess strain presence in metagenomes and to complement it we used an
353 independent and complementary alignment based approach to gauge whether *C. kefirresidentii*
354 were enriched during atopic dermatitis flares in the Byrd *et al.* 2017 metagenomes. For this
355 approach, we aligned metagenomic reads individually to 4,514 representative genes from 367
356 homolog groups which comparative genomics had revealed were single-copy orthologs and
357 featured at least 20 sites where *C. kefirresidentii* had a core allele that was not observed in the
358 complementary set of genomes in the *C. tuberculostearicum* species complex. Reads were
359 aligned as unpaired using Bowtie 2[45] with parameters "--very-sensitive-local --no-unal -a -x".
360 Reads were aligned as unpaired because gene sequences are short and pairs can thus align
361 poorly. Additionally, all alignments for each read are requested because multiple representative
362 genes were considered per homolog group. We then processed alignments to each
363 representative gene where coverage at 1X was not observed for 90% of the sites. Following this
364 preliminary scan for checking that genes were loosely covered, we assessed alignments to the
365 gene met either of the following two criteria: (i) $\geq$99% identity with length $\geq$60 bp or (2) $\geq$95%
366 identity with length $\geq$100 bp for the core alignment region (excluding flanking regions to account
367 for alignments hanging off representative gene edges). We also allowed a maximum of 5 indelic
368 sites within the core alignment region for the alignment to be considered. Alignments which met
369 these criteria were then further processed and retained if they had the best mapping score
370 observed for a particular read (in the case of ties, all alignments with the best Bowtie 2 mapping
371 score were retained). Finally, allele counts were tallied for each metagenome at each site in
372 homolog group codon alignments based on mapping of the query base to the reference /
373 representative gene position in the codon alignments. Only positions in reads which had greater
374 than 30 base quality were considered.
375 Results from determining the presence of alleles at sites along homolog group codon
376 alignments for each metagenome were processed to identify the proportion of *C. kefirresidentii*
377 specific alleles covered by at least three reads. The mean coverage of these well-covered *C.*
378 *kefirresidentii* specific alleles were computed and normalized for sequencing depth through

379 dividing by a metagenome specific scalar value (Figure S3). This scalar value corresponded to
380 the ratio of the number of bases the metagenome featured in relation to the metagenome with
381 the fewest number of bases in the final forward read files following processing.
382 We additionally identified *C. tuberculostearicum* and type E *C. aurimucosum* specific
383 alleles along the same set of 367 core homolog groups and quantified the mean coverage of
384 species specific alleles within each metagenome (inclusive of coverage at alleles which were
385 uncovered). This allowed us to compute a ratio statistic measuring the abundance of *C.*
386 *kefirresidentii* to *C. tuberculostearicum* and type E. *C. aureimucosum* with a pseudocount of 1
387 added to both the numerator and denominator to avoid dividing by zero. A value greater than 1
388 for this statistic thus indicates that *C. kefirresidentii* are more prevalent than *C.*
389 *tuberculostearicum* and type E. *C. aureimucosum* within an individual metagenome; whereas, a
390 value less than 1 supports the opposite. Our code for metagenomic alignment and calculation of
391 the ratio statistic is provided in the Github repository: https://github.com/Kalan-
392 Lab/Salamzade_etal_CtuberCompGen.
393 *Statistical testing*: To test whether StrainGST-based detection of *C. kefirresidentii* was
394 significantly more prevalent within metagenomes in atopic dermatitis flare metagenomes relative
395 to baseline and healthy control metagenomes, we used the two-sided Fisher's exact test. The
396 two-sided Fisher's exact test was also used to statistically assess whether the ratio statistic was
397 larger than 1 in metagenomes sampled during atopic dermatitis flares relative to baseline and
398 healthy control metagenomes.
399
400 **Visualizations**
401 iTol was used for illustrating phylogenetic trees and associated features[46]. The R library
402 ggplot2[47] was used to visualize heatmaps, scatterplots, and bar graphs. Circleator[48] was used to
403 generate an overview of conservation and showcase genes highlighted by comparative
404 genomics along the chromosome of *C. kefirresidentii* FDAARGOS_1055.
405
406 **REFERENCES**

407 1. Brown, S., Lanéelle, M. A., Asselineau, J. & Barksdale, L. Description of Corynebacterium
408 tuberculostearicum sp. nov., a leprosy-derived Corynebacterium. *Ann. Microbiol.* **135B**,
409 251–267 (1984).
410 2. Bernard, K. The genus corynebacterium and other medically relevant coryneform-like
411 bacteria. *J. Clin. Microbiol.* **50**, 3152–3158 (2012).
412 3. Oh, J. *et al.* Temporal Stability of the Human Skin Microbiome. *Cell* **165**, 854–866 (2016).
413 4. Swaney, M. H., Sandstrom, S. & Kalan, L. R. Cobamide Sharing Is Predicted in the Human
414 Skin Microbiome. *mSystems* e0067722 (2022).
415 5. Altonsy, M. O. *et al.* Corynebacterium tuberculostearicum, a human skin colonizer, induces
416 the canonical nuclear factor-κB inflammatory signaling pathway in human skin cells.
417 *Immun. Inflamm. Dis.* **8**, 62–79 (2020).
418 6. Abreu, N. A. *et al.* Sinus microbiome diversity depletion and Corynebacterium
419 tuberculostearicum enrichment mediates rhinosinusitis. *Sci. Transl. Med.* **4**, 151ra124
420 (2012).
421 7. Feurer, C. *et al.* Taxonomic characterization of nine strains isolated from clinical and
422 environmental specimens, and proposal of Corynebacterium tuberculostearicum sp. nov.
423 *Int. J. Syst. Evol. Microbiol.* **54**, 1055–1061 (2004).
424 8. Hinic, V. *et al.* Corynebacterium tuberculostearicum: a potentially misidentified and

425 multiresistant Corynebacterium species isolated from clinical specimens. *J. Clin. Microbiol.*
426 **50**, 2561–2567 (2012).
427 9. Swaney, M. H., Nelsen, A., Sandstrom, S. & Kalan, L. R. Sweat and sebum preferences of
428 the human skin microbiota. doi:10.1101/2022.08.13.503869.
429 10. Parks, D. H. *et al.* GTDB: an ongoing census of bacterial and archaeal diversity through a
430 phylogenetically consistent, rank normalized and complete genome-based taxonomy.
431 *Nucleic Acids Res.* (2021) doi:10.1093/nar/gkab776.
432 11. Salamzade, R. *et al.* IsaBGC provides a comprehensive framework for evolutionary
433 analysis of biosynthetic gene clusters within focal taxa. *bioRxiv* 2022.04.20.488953 (2022)
434 doi:10.1101/2022.04.20.488953.
435 12. Yassin, A. F., Steiner, U. & Ludwig, W. Corynebacterium aurimucosum sp. nov. and
436 emended description of Corynebacterium minutissimum Collins and Jones (1983). *Int. J.*
437 *Syst. Evol. Microbiol.* **52**, 1001–1005 (2002).
438 13. Blasche, S., Kim, Y. & Patil, K. R. Draft Genome Sequence of Corynebacterium
439 kefirresidentii SB, Isolated from Kefir. *Genome Announc.* **5**, (2017).
440 14. Saheb Kashaf, S. *et al.* Integrating cultivation and metagenomics for a multi-kingdom view
441 of skin microbiome diversity and functions. *Nat Microbiol* **7**, 169–179 (2022).
442 15. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM:
443 assessing the quality of microbial genomes recovered from isolates, single cells, and
444 metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
445 16. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative
446 genomics. *Genome Biol.* **20**, 238 (2019).
447 17. Didelot, X. & Wilson, D. J. ClonalFrameML: efficient inference of recombination in whole
448 bacterial genomes. *PLoS Comput. Biol.* **11**, e1004041 (2015).
449 18. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify
450 genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2019).
451 19. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High
452 throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries.
453 *Nat. Commun.* **9**, 1–8 (2018).
454 20. Olm, M. R. *et al.* Consistent Metagenome-Derived Metrics Verify and Delineate Bacterial
455 Species Boundaries. *mSystems* **5**, (2020).
456 21. van Dijk, L. R. *et al.* StrainGE: a toolkit to track and characterize low-abundance strains in
457 complex microbial communities. *Genome Biol.* **23**, 74 (2022).
458 22. Byrd, A. L. *et al.* Staphylococcus aureus and Staphylococcus epidermidis strain diversity
459 underlying pediatric atopic dermatitis. *Sci. Transl. Med.* **9**, (2017).
460 23. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate
461 genomic comparisons that enables improved genome recovery from metagenomes through
462 de-replication. *ISME J.* **11**, 2864–2868 (2017).
463 24. Lu, S. *et al.* CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res.*
464 **48**, D265–D268 (2020).
465 25. Gago, G., Diacovich, L., Arabolaza, A., Tsai, S.-C. & Gramajo, H. Fatty acid biosynthesis in
466 actinomycetes. *FEMS Microbiol. Rev.* **35**, 475–497 (2011).
467 26. Chen, S.-M. *et al.* Streptococcal collagen-like surface protein 1 promotes adhesion to the
468 respiratory epithelial cell. *BMC Microbiol.* **10**, 320 (2010).
469 27. Yu, Z., An, B., Ramshaw, J. A. M. & Brodsky, B. Bacterial collagen-like proteins that form
470 triple-helical structures. *J. Struct. Biol.* **186**, 451–461 (2014).
471 28. Bachert, B. A. *et al.* A Unique Set of the Burkholderia Collagen-Like Proteins Provides
472 Insight into Pathogenesis, Genome Evolution and Niche Adaptation, and Infection
473 Detection. *PLoS One* **10**, e0137578 (2015).
474 29. Qiu, Y., Zhai, C., Chen, L., Liu, X. & Yeo, J. Current Insights on the Diverse Structures and
475 Functions in Bacterial Collagen-like Proteins. *ACS Biomater Sci Eng* (2021)

476 doi:10.1021/acsbiomaterials.1c00018.
477 30. Berends, E. T. M. *et al.* Nuclease expression by Staphylococcus aureus facilitates escape
478 from neutrophil extracellular traps. *J. Innate Immun.* **2**, 576–586 (2010).
479 31. Zhong, Q., Kobe, B. & Kappler, U. Molybdenum Enzymes and How They Support Virulence
480 in Pathogenic Bacteria. *Front. Microbiol.* **11**, 615860 (2020).
481 32. Williams, M. J. *et al.* bis-Molybdopterin guanine dinucleotide is required for persistence of
482 Mycobacterium tuberculosis in guinea pigs. *Infect. Immun.* **83**, 544–550 (2015).
483 33. Tauch, A. *et al.* Complete genome sequence and analysis of the multiresistant nosocomial
484 pathogen Corynebacterium jeikeium K411, a lipid-requiring bacterium of the human skin
485 flora. *J. Bacteriol.* **187**, 4671–4682 (2005).
486 34. Institute of Medicine (US) Committee on Resource Sharing, Berns, K. I., Bond, E. C. &
487 Manning, F. J. *The American Type Culture Collection*. (National Academies Press (US),
488 1996).
489 35. Funke, G., Osorio, C. R., Frei, R., Riegel, P. & Collins, M. D. Corynebacterium confusum
490 sp. nov., isolated from human clinical specimens. *Int. J. Syst. Bacteriol.* **48 Pt 4**, 1291–1296
491 (1998).
492 36. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site
493 identification. *BMC Bioinformatics* **11**, 119 (2010).
494 37. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J.
495 eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain
496 Prediction at the Metagenomic Scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).
497 38. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. E. The Phyre2 web
498 portal for protein modeling, prediction and analysis. *Nat. Protoc.* **10**, 845–858 (2015).
499 39. Edgar, R. C. MUSCLE v5 enables improved estimates of phylogenetic tree confidence by
500 ensemble bootstrapping. *bioRxiv* 2021.06.20.449169 (2021)
501 doi:10.1101/2021.06.20.449169.
502 40. Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence
503 alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**, W609–12
504 (2006).
505 41. Lee, M. D. GToTree: a user-friendly workflow for phylogenomics. *Bioinformatics* **35**, 4162–
506 4164 (2019).
507 42. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood
508 trees for large alignments. *PLoS One* **5**, e9490 (2010).
509 43. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor.
510 *Bioinformatics* **34**, i884–i890 (2018).
511 44. Li, W. *et al.* RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with
512 protein family model curation. *Nucleic Acids Res.* **49**, D1020–D1028 (2021).
513 45. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods*
514 **9**, 357–359 (2012).
515 46. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new
516 developments. *Nucleic Acids Res.* **47**, W256–W259 (2019).
517 47. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer, New York, NY, 2009).
518 48. Crabtree, J. *et al.* Circleator: flexible circular visualization of genome-associated data with
519 BioPerl and SVG. *Bioinformatics* **30**, 3125–3127 (2014).

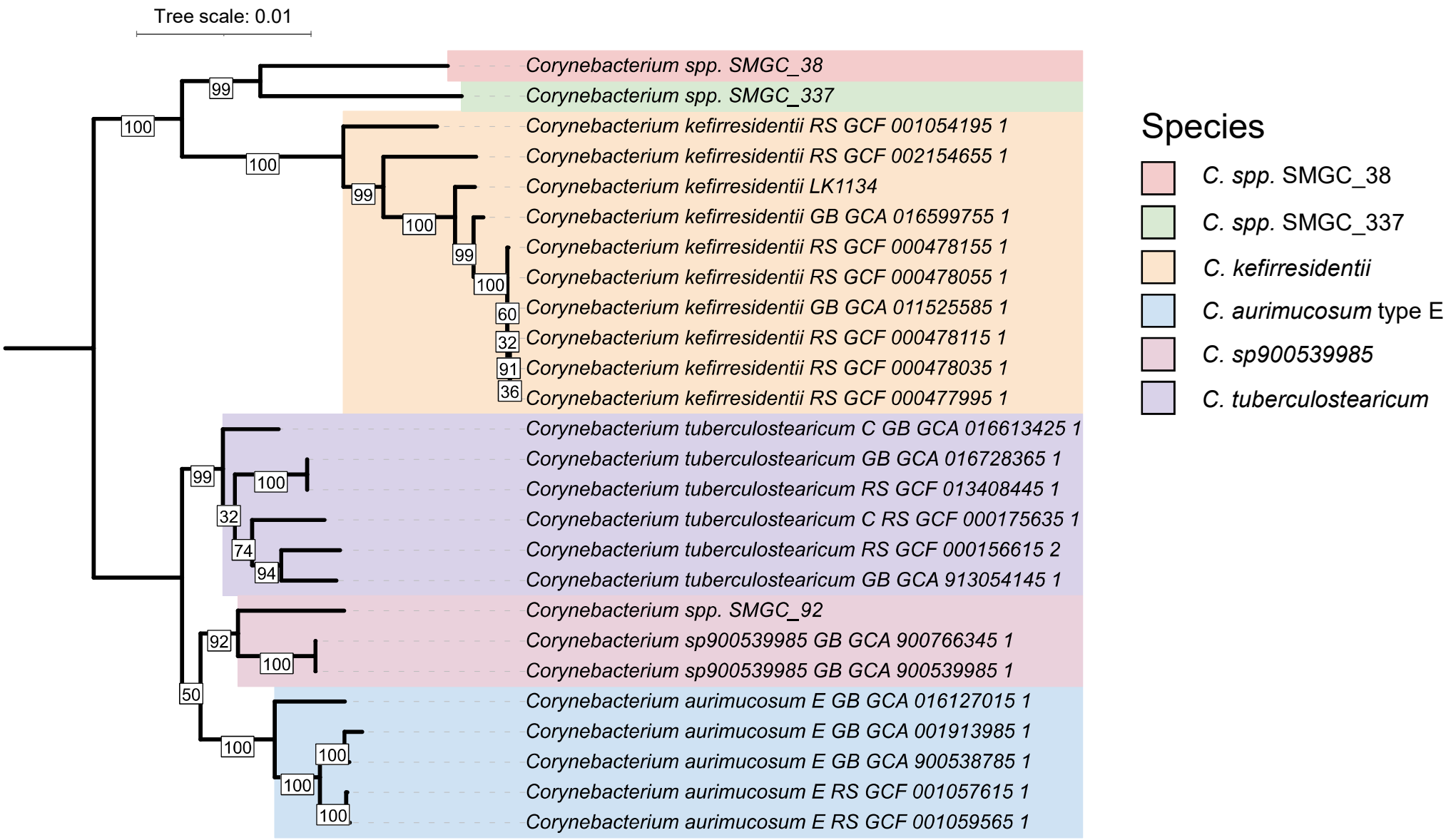

Figure 1: A high-resolution phylogenomic view of the *C. tuberculostearicum* species complex. A maximum-likelihood phylogeny of the *C. tuberculostearicum* species complex, constructed from 15,936 sites predicted to not be impacted by recombination and not exhibit ambiguity in any of the 26 genomes, is shown after midpoint rooting. Bootstrap values are indicated at inner-nodes.

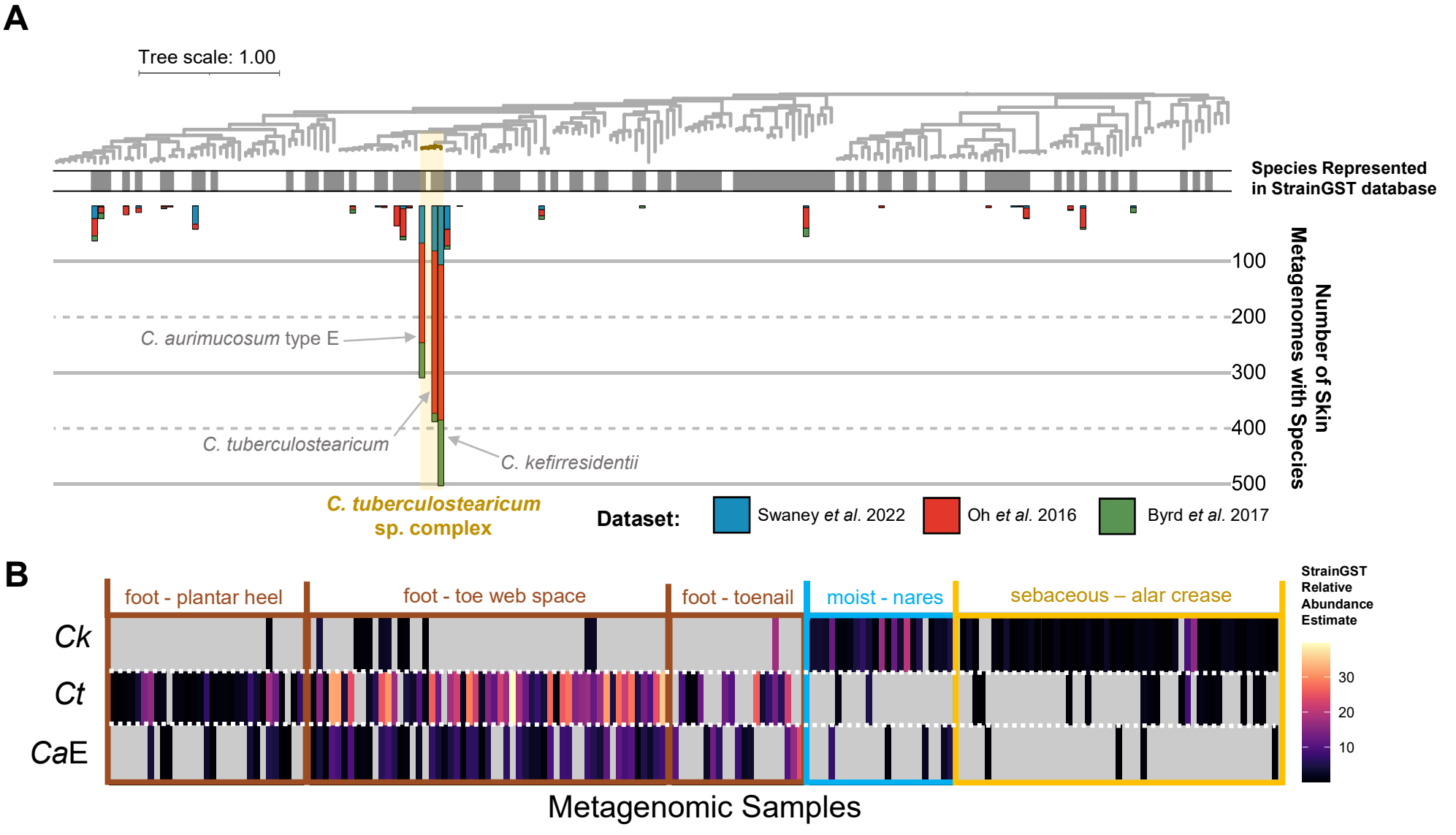

Figure 2: The *C. tuberculostearicum* species complex across skin metagenomic datasets and body sites. A) A phylogeny of the 187 *Corynebacterium* is shown with the *C. tuberculostearicum* species complex highlighted in gold. A single track is shown for species which were represented in the StrainGST database which we used to detect strains/species within metagenomes. Bars correspond to the number of skin-related metagenomic samples which were found to feature a species, with colors representing the study the metagenomic. **B**) The estimated relative abundance of *C. kefirresidentii* (*Ck*), *C. tuberculostearicum* (*Ct*), and *C. aurimucosum* type E (*Ca*E) is shown across Oh *et al.* 2016 and Swaney *et al.* 2022 metagenomes from select body sites in adults. Metagenomes without any of the three species detected are not shown.

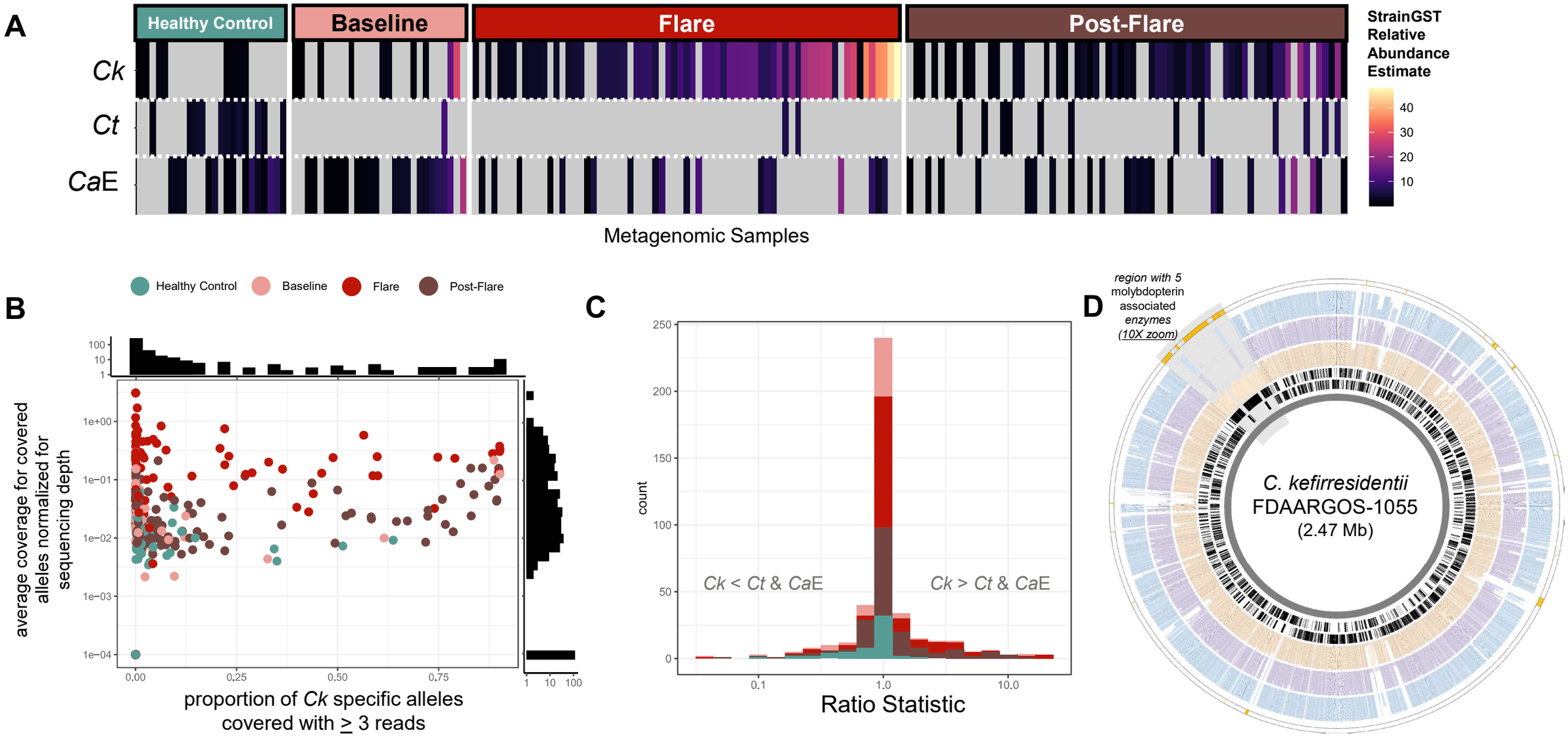

Figure 3: Increase in *C. kefirresidentii* relative abundance in metagenomes sampled during atopic dermatitis flares. A) The estimated relative abundance of *Corynebacterium* species in the species complex is shown across all metagenomes from Byrd *et al.* 2017 with at least one species from the complex detected. **B**) Average coverage of *C. kefirresidentii* specific alleles along 367 single-copy-orthologs of the *C. tuberculostearicum* species is shown. Only sites/alleles which were covered by at least three reads are regarded. **C**) The ratio of the total average coverage for *C. kefirresidentii* specific alleles (inclusive of uncovered alleles) to the sum of total average coverage for *C. tuberculostearicum* specific alleles and *C. aurimucosum* type E specific alleles is shown. A pseudo-count of 1 was added to both the numerator and denominator. **D)** Genomic location of homolog groups identified as enriched in *C. kefirresidentii* genomes relative to *C. tuberculostearicum* and *C. aurimucosum* type E genomes along the chromosome of *C. kefirresidentii* FDAARGOS_1055 is shown as the outermost track in gold. The most inner two rings in black show the location of coding genes on the reverse and forward strands. The following rings (from inner to outer) show the conservation of homologs of the gene in *C. kefirresidentii* (orange), *C. tuberculostearicum* (purple), and *C. aurimucosum* type E (blue) genomes. The co-located set of genes found to be conserved in 80% of *C. kefirresidentii* featuring five molybdopterin associated enzymes is highlighted in grey and zoomed in at 10X.