# A non-adaptive demographic mechanism for genome expansion in *Streptomyces*

Mallory J. Choudoir[a†#], Marko J. Järvenpää[b‡], Pekka Marttinen[b], and *Daniel H. Buckley[b#]

[a]School of Integrative Plant Science, Cornell University, Ithaca, NY, USA

[b]Department of Computer Science, Helsinki Institute for Information Technology HIIT, Aalto University, Espoo, Finland

[†]Present address: Department of Microbiology, University of Massachusetts Amherst, Amherst, MA, USA

[‡]Present address: Department of Biostatistics, University of Oslo, Oslo, Norway

#Address correspondence to Mallory J. Choudoir, mchoudoir@umass.edu or Daniel H. Buckley, dhb28@cornell.edu

MC and DB conceived and designed the study. MC generated and analyzed the data. MJ and PM performed the simulations. MC and DB wrote the paper.

**Running title:** Non-adaptive mechanism for genome expansion

**Abstract:** 250 words

**Main text:** 4401 words

**Abstract**

The evolution of microbial genome size is driven by gene acquisition and loss events that occur at scales from individual genomes to entire pangenomes. The equilibrium between gene gain and loss is shaped by evolutionary forces, including selection and drift, which are in turn influenced by population demographics. There is a well-known bias towards deletion in microbial genomes, which promotes genome streamlining. Less well described are mechanisms that promote genome expansion, giving rise to the many microbes, such as *Streptomyces*, that have unusually large genomes. We find evidence of genome expansion in *Streptomyces* sister-taxa, and we hypothesize that a recent demographic range expansion drove increases in genome size through a non-adaptive mechanism. These *Streptomyces* sister-taxa, NDR (northern-derived) and SDR (southern-derived), represent recently diverged lineages that occupy distinct geographic ranges. Relative to SDR genomes, NDR genomes are larger, have more genes, and their genomes are enriched in intermediate frequency genes. We also find evidence of relaxed selection in NDR genomes relative to SDR genomes. We hypothesize that geographic range expansion, coupled with relaxed selection, facilitated the introgression of non-adaptive horizontally acquired genes, which accumulated at intermediate frequencies through a mechanism known as genome surfing. We show that similar patterns of pangenome structure and genome expansion occur in a simulation that models the effects of population expansion on genome dynamics. We show that non-adaptive evolutionary phenomena can explain expansion of microbial genome size, and suggest that this mechanism might explain why so many bacteria with large genomes can be found in soil.

**Importance**

Most bacterial genomes are small, but some are quite large, and differences in genome size are ultimately driven by the interplay of gene gain and loss dynamics operating at the population level. The evolutionary forces that favor genome size reduction are well known, but less understood are the forces that drive genome expansion. It is generally assumed that large genomes are adaptive because they favor metabolic versatility. However, we find evidence in *Streptomyces* for a non-adaptive mechanism of genome expansion driven by horizontal gene transfer. We hypothesize that historical range expansion decreased the strength of selection acting these genomes. Relaxed selection allowed many newly acquired genes (which would normally be lost to deletion) to accumulate, leading to increased genome size. *Streptomyces* have large genomes that contain a remarkable diversity of antibiotic producing gene clusters, and genome expansion has likely contributed to the evolution of these traits.

## Introduction

Microbial genomes are extraordinarily dynamic. Genome size varies considerably, and gene content in strains of the same species can differ dramatically, giving rise to the pangenome. The pangenome concept has transformed our understanding of evolutionary processes in diverse taxa (1–4). The pangenome is the entire collection of genes in a microbial species, and is subdivided into core genes present in all strains, dispensable or accessory genes present in some strains, and strain-specific or unique genes (5, 6). Rates of gene acquisition and gene loss determine the individual genome size, and consequently, pangenome composition is shaped by evolutionary mechanisms that alter gene frequencies in microbial populations (7–9).

69 Genome size varies by four orders of magnitude ($10^4$–$10^7$ kb) in eukaryotic organisms and two
70 orders of magnitude in prokaryotic organisms (from less than 150 kb in certain endosymbionts to
71 over 10 Mb for some free-living bacteria) (10). Unlike eukaryotes, whose genomes contain large
72 portions of non-coding DNA, prokaryotic gene content is directly related to genome size because
73 bacterial and archaeal taxa have high coding density (11, 12). While microbial genomes are
74 constantly in flux, deletion rates are approximately three-fold greater than rates of gene
75 acquisition (13). Multiple factors contribute to the strong deletion bias in microbial genomes,
76 including selection for efficiency, "use it or lose it" purging of nonessential genes, and genetic
77 drift (14–17).
78
79 Because of this tendency towards deletion, microbial genome reduction has been examined in
80 greater detail than genome expansion. For example, the evolutionary mechanisms driving
81 genome reduction in obligate pathogens like *Rickettsia* and symbionts like *Buchnera* in aphids
82 are well described (17, 18). The transition from a free-living to a host-associated lifestyle
83 involves substantial loss of superfluous genes, and generations of vertical transmission in small
84 asexual populations leads to gene inactivation and deletion accelerated by genetic drift (16).
85 Alternatively, genome streamlining leads to reduction of both genome and cell size through
86 selection for increased metabolic efficiency in free-living microbes with large populations (15).
87 Genome streamlining is historically associated with marine oligotrophic *Pelagibacter* (14, 19)
88 but has more recently been described for soil-dwelling *Verrucomicrobia* (20).
89
90 Large genomes are frequent among terrestrial free-living microbes, and must be the product of
91 evolutionary forces that drive genome expansion. A common, though relatively untested,

92 hypothesis to explain large genomes is that high environmental heterogeneity (a characteristic of
93 terrestrial habitats) selects for metabolic versatility afforded by gene gain, and thereby drives
94 genome expansion (21, 22). For example, massive gene acquisition and adaptation to alkaline
95 conditions caused genome expansion in the myxobacterium *Sorangium cellulosum*, which at 15
96 Mb is one of the largest known bacterial genomes (23). Mechanisms of gene gain include
97 duplication or horizontal gene transfer (HGT), and large genomes are enriched in functional
98 genes acquired from phylogenetically distant origins (24). Much of the evolution of gene
99 families can be attributed to HGT rather than duplication events (25, 26), and HGT is a major
100 driver of genome expansion (27, 28). While HGT-mediated gene acquisition occurs with great
101 frequency, microbial genomes remain relatively small, and genome size tends to be fairly
102 conserved within a species.
103
104 Gene frequencies at the population-level are governed by selection and drift, and these
105 evolutionary forces determine whether a newly acquired gene will be purged from the
106 pangenome or whether it will sweep to fixation. The strength of selection and drift varies
107 inversely, and their relative contributions are determined by a gene's selection coefficient and
108 effective population size ($N_e$) (29, 30). Drift can exert large effects on populations with small $N_e$,
109 but these effects decline as $N_e$ increases and selection intensifies. Our ability to disentangle the
110 contributions of selection and drift to pangenome dynamics are complicated by the fact that it
111 remains difficult to estimate microbial $N_e$ (31, 32) and to delimit microbial population and
112 species boundaries (33–35). Another complication is that demographic models often include the
113 simplifying expectation that $N_e$ is invariable over time.
114

115 Rapid changes in population size are typical in the evolutionary histories of many microbial
116 species, and fluctuations in $N_e$ such as population bottlenecks or expansions can have profound
117 impacts on contemporary patterns of genomic diversity. For example, the population structure
118 for many pathogenic bacterial lineages is exemplified by episodes of rapid expansion of clonal
119 complexes repeated across space and time (36–38). Microbial population expansions can also be
120 linked to ecological or geographical range expansions (39–42). For instance, demographic
121 expansion in the oral bacteria *Streptococcus mutans* coincides with the origin of human
122 agricultural practices (41).
123
124 We find evidence for post-glacial range expansion in the genus *Streptomyces*, and these species
125 exhibit several of the genetic characteristics described in plant and animal species whose
126 biogeography was influenced by Pleistocene glaciation (43, 44). By examining *Streptomyces*
127 isolated from sites across North America, we observed genetic evidence for dispersal limitation,
128 a latitudinal gradient in taxonomic richness, and a latitudinal gradient in genetic diversity (45,
129 46). We also identified recently diverged sister-taxa comprising a more genetically diverse
130 southern-derived (SDR) clade and a more homogenous northern-derived (NDR) clade, which
131 occupied discrete geographic ranges spanning the boundary of glaciation (47). We further
132 observed larger genomes in the northern clade compared to the southern clade.
133
134 We hypothesize that genome expansion in NDR is a consequence of demographic change driven
135 by post-Pleistocene range expansion. Here, we evaluate the effects of historical range expansion
136 on lineage divergence, genome size, and pangenome structure, and assess these data in the
137 context of the genome surfing hypothesis. Genome surfing is a non-adaptive mechanism which

138 describes the introgression of horizontally acquired genes facilitated by relaxed selection and
139 amplified by geographic expansion (48). We hypothesize that range expansion, coupled with
140 relaxed selection, dampened gene loss thereby facilitating an increase in non-adaptive,
141 intermediate frequency genes in the NDR pangenome. We infer gene gain and loss dynamics by
142 evaluating patterns of shared gene content between strains. We predict that the contribution of
143 drift is greater in NDR compared to SDR, and determine the relative strength of selection by
144 comparing genome-wide rates of amino acid substitution between clades. Finally, we evaluate
145 our hypothesis by modeling population expansion under a regime of relaxed selection and ask
146 whether these demographic conditions increase retention of horizontally acquired genes at
147 intermediate frequencies, ultimately causing genome expansion.
148

149 **Results**

150 *Streptomyces sister-taxa*

151 We sequenced the genomes of 20 *Streptomyces* strains isolated from ecologically similar
152 grasslands sites across the United States (Table S1, Table S2). These genomes derive from sister-
153 taxa comprising a northern-derived (NDR) and southern-derived clade (SDR), which originate
154 from sites spanning the historical extent of glaciation (Figure S1, see (45)). These sister-taxa
155 represent closely related but genetically distinct microbial species. Genomes within NDR share
156 97.8 ± 1.3% (mean ± SD) ANI and those within SDR share 97.6 ± 0.1% (mean ± SD) ANI,
157 while inter-clade genomic ANI is 93.0 ± 0.14% (mean ± SD). An ANI of 93–96% is typically
158 indicative of taxonomic species boundaries (49, 50). For comparative purposes, we also
159 sequenced the genomes of four strains that co-localized with the sister-taxa. The closest

160 taxonomic neighbor to these 24 strains is *Streptomyces griseus* subsp. *griseus* NBRC 13350,
161 although all strains share < 95% ANI with this type strain (Figure S1).
162
163 *Genomic attributes and gene content*
164 NDR genomes are larger (8.70 ± 0.23 Mb, mean ± SD) than SDR genomes (7.87 ± 0.19 Mb,
165 mean ± SD), and this difference is significant (Mann Whitney U test; $P < 0.0001$) (Figure 1a).
166 NDR genomes also have also have more orthologous protein-coding gene clusters (hereby
167 referred to as genes) (7,775 ± 196 genes, mean ± SD) than SDR genomes (7,093 ± 205 genes,
168 mean ± SD), and this difference is also significant (Mann Whitney U test; $P < 0.0001$) (Figure
169 1b). As expected, there is a strong positive correlation between genome size and gene content
170 ($R^2 = 0.95$, $P < 0.0001$), but coding density did not differ between clades (Figure S2). SDR
171 genomes are more genetically diverse than NDR. Nucleotide diversity ($\pi$) across conserved,
172 single-copy core genes is greater in SDR than NDR, and this difference is significant (Mann
173 Whitney U test; $P < 0.0001$) (see (47)). Finally, NDR genomes have slightly lower genome-wide
174 GC content (71.50 ± 0.087%, mean ± SD) than SDR genomes (71.62 ± 0.11%, mean ± SD), and
175 this difference is significant (Mann Whitney U test; $P = 0.017$) (Figure 1c). Shared gene content
176 between strains correlates with genomic similarity as measured by ANI (NDR: $R^2 = 0.82$, $P <$
177 $0.0001$; SDR: $R^2 = 0.64$, $P < 0.0001$) (Figure 2). However, gene content varies more in NDR
178 than in SDR, and there is a significant interaction between genomic similarity and clade with
179 respect to gene content shared between strains (Table S3). This interaction comes from shared
180 gene content between strains increasing more rapidly over recent phylogenetic timescales in
181 NDR compared to SDR (Figure 3).
182

183 *Pangenome structure and dynamics*

184 The 24 *Streptomyces* genomes (Table S2) contain 22,055 total orthologous protein-coding gene
185 clusters (i.e., genes), and 42% (9,285 genes) are strain-specific. All 24 genomes share 3,234
186 (2,778 single-copy) genes, which represent 40–48% of the total gene content per strain. While
187 NDR has a smaller core genome than SDR (4,234 and 4,400 genes, respectively), its pangenome
188 is larger (13,681 genes in NDR versus 12,259 genes in SDR) and contains a greater number of
189 clade-specific genes (5,647 genes unique to NDR versus 4,308 genes unique to SDR) (Figure 3,
190 Figure 4, Table S4).

191

192 For most microbial species, pangenome frequency distributions are U-shaped, reflecting high
193 proportions of both strain-specific genes and core genes (51). While the pangenome structures of
194 our *Streptomyces* sister-taxa generally conform to this shape, the NDR pangenome is enriched in
195 intermediate frequency accessory genes relative to SDR (Figure 4). The proportion of
196 intermediate-low frequency (i.e., present in 3–5 strains) accessory genes is higher in NDR than
197 in SDR (19% of total genes for NDR versus 9.2% of total genes for SDR) (Table S4), and this
198 difference is statistically significant (two proportion z-test; $P < 0.0001$). Conversely, the
199 proportion of intermediate-high frequency (i.e., present in 6–8 strains) accessory genes is
200 equivalent (6.9% of total genes for NDR versus 7.2% of total genes for SDR; two proportion z-
201 test; $P = 0.26$) (Table S4).

202

203 Next, we determined if genes across different gene pools, binned according to their pangenome
204 frequencies, differed in genetic attributes including per-gene GC content and codon usage bias.
205 GC content differs between gene pools for both NDR and SDR pangenomes (ANOVA; $F_{3, 25932}$

206 = 267.5, *P*-value < 0.0001) (Figure S3). In general, GC content is greater in high frequency and
207 core genes compared to rare and intermediate frequency genes for both sister-taxa. Codon usage
208 bias as measured by the effective number of codons (ENC) (52) also differs between gene pools
209 for both NDR and SDR pangenomes (ANOVA; $F_{3, 21624} = 1862.7$, *P*-value < 0.0001) (Figure
210 S4). Rare and intermediate frequency genes exhibit less overall codon bias compared to high
211 frequency and core genes, which tend to use codons more preferentially.
212
213 *Historical population demography*
214 Due to founder effects occurring at the edge of an expanding population, $N_e$ is dramatically
215 reduced during geographic range expansion (53). Consequently, relaxed selection will
216 accompany range expansion since the contribution of selection scales directly with $N_e$. Based on
217 the theory of neutral molecular evolution, which states that selection on synonymous sites is
218 negligible (54), the ratio of non-synonymous to synonymous amino acid substitutions ($K_A/K_S$)
219 reflects the relative strength of selection acting on a sequence. When assessed at the level of
220 single-copy genes conserved between the sister taxa (2,444 genes), we observe that genome-wide
221 $K_A/K_S$ tends to be higher in NDR than in SDR (Figure 5), and this difference is significant
222 (Mann-Whitney U test; $P < 0.0001$). This result indicates that selection is weaker and genetic
223 drift stronger in NDR relative to SDR.
224
225 We used a population model (modified from (55)) to determine whether demographic expansion
226 could produce increased intermediate gene frequencies and result in genome expansion. We
227 simulated gene gain and loss events in a population undergoing exponential growth over 100
228 generations, and determined changes in pangenome structure and genome size. To approximate

229 relaxed selection during the population expansion, we imposed a fitness penalty for newly
230 acquired genes that scaled inversely with population size. At the beginning of expansion, most
231 genes were present at high frequencies due to strong founder effects (Figure S5, top and middle).
232 During the expansion, we observed a transient enrichment of intermediate frequency genes
233 within the pangenome (Figure S5, top and middle). Total gene content also increased during
234 population expansion due to relaxed selection pressure when $N_e$ was small, which allowed for
235 the persistence of newly HGT-acquired genes. Genome size stabilized when $N_e$ reached
236 maximum size, and selection pressure balanced HGT-mediated gene gain with simultaneous
237 gene loss (Figure S5, bottom).
238
239 **Discussion**
240 We have hypothesized that the biogeography of our *Streptomyces* sister-taxa is explained by
241 historical demographic change driven by geologic and climatic events that occurred in the late
242 Pleistocene (46, 47). Following the last glacial maxima, North American plant and animal
243 species rapidly colonized glacial retreat zones, and the genetic consequences of post-glacial
244 expansion are well documented and include northern-ranged populations with low diversity that
245 established vast geographic extent (43, 44). We hypothesize that the recent common ancestor of
246 NDR and SDR inhabited southern glacial refugia prior to the last glacial maxima (LGM). Post
247 glaciation, NDR dispersed northward and colonized the latitudinal range it occupies today (see
248 (46)). We previously described patterns of gene flow, genomic diversity, and ecological
249 adaptation in these sister-taxa, with both adaptive and non-adaptive processes likely reinforcing
250 lineage divergence (47). Here, we evaluate the outcomes of historical range expansion on sister-
251 taxa pangenome structure and genome size.

252

253 Expanding populations experience repeated founder effects as individuals along the leading edge
254 disperse and colonize new landscapes, creating spatial patterns of genetic diversity akin to
255 genetic drift (53). Allele surfing, or gene surfing, is a non-adaptive mechanism that propagates
256 rare alleles along an expanding edge such that neutral, or even deleterious, variants 'surf' to
257 higher frequencies than would be expected under population equilibrium (56–58). When applied
258 to expanding microbial populations, gene surfing can facilitate genome surfing, a neutral
259 mechanism acting at the pangenome level that causes rare genes to surf to higher frequencies
260 independent of natural selection (48). Below, we outline how historical range expansion and
261 genome surfing could give rise to genome expansion in *Streptomyces*.

262

263 Genome surfing is most likely to occur in microbial populations with intermediate levels of
264 dispersal and in taxa capable of HGT. Bacteria in the genus *Streptomyces* are ubiquitous in soil
265 and produce desiccation and starvation resistant spores which are easily disseminated (59),
266 making them ideal for studying patterns of biogeography dependent on dispersal limitation.
267 Rates of HGT in *Streptomyces* are among the highest estimated across a range of bacterial
268 species (60–62). In many instances, HGT events occurred in ancestral lineages creating patterns
269 of shared genetic ancestry and reticulate evolution in many extant *Streptomyces* species (63). We
270 previously observed a distance decay relationship between sites up to 6,000 km apart, indicative
271 of dispersal limitation at intermediate spatial scales that allows detection of geographic patterns
272 of diversity across the sampled range (45, 46). We also found evidence of restricted gene flow
273 between the core genomes of NDR and SDR (47).

274

275 Since NDR and SDR sister-taxa share a recent common ancestor (Figure S1), they must also
276 share a common ancestral genome size. Hence, differences in genome size accompanying
277 lineage divergence resulted from either genome expansion in NDR or genome reduction in SDR.
278 Given that changes in genome size are ultimately the result of gene gain and loss, we first
279 evaluated differences in shared gene content between NDR and SDR strains. We find greater
280 variability in shared gene content in NDR compared to SDR (Figure 2, Figure 3). This result
281 suggests relative gene content stability for SDR and gene content instability for NDR, most
282 notably in recent phylogenetic history (Figure 3). Likewise, the pangenome of NDR exceeds that
283 of SDR by over 1,000 genes. Evidence suggests that during range expansion, founders at the
284 expansion edge disperse into new habitats and acquire genes from local gene pools
285 asymmetrically at unequal rates, and gene flow is almost exclusively from local to invading
286 genomes (64). These data are consistent with the observation that that NDR has a larger, more
287 diverse, and more dynamic pangenome than SDR due to introgression from local gene pools.
288 Regardless of their origin, most novel horizontally-acquired genes are neutral or nearly neutral
289 (65). In most situations, selection will balance gene gain with gene deletion, and genome size
290 will remain relatively constant.
291
292 Genetic diversity in individuals at the leading edge of an expanding population is dramatically
293 reduced, and their genomes experience relaxed selection pressure due to consecutive population
294 bottlenecks and low $N_e$ (66). We find that NDR has lower genetic diversity (47) and higher rates
295 of $K_A/K_S$ across its core genome relative to SDR (Figure 5), which is consistent with the
296 prediction that NDR has experienced a period of relaxed selection relative to SDR. A positive
297 correlation is observed between GC content and selection pressure on microbial genomes (67,

298 68), and genome expansion in *Chlamydia* has been linked to relaxed selection resulting in a
299 decrease in genome-wide GC content (69). We likewise observe a decrease in genome-wide GC
300 content in NDR relative to SDR (Figure 1). Relaxed selection pressure in NDR would mitigate
301 the natural bias towards deletion and permit genes acquired by HGT to persist in the genome,
302 regardless of their adaptive coefficient. Microbial sectoring that accompanies geographic range
303 expansion (70) would then allow these newly acquired genes to accumulate at intermediate
304 frequencies in the pangenome. The fact that NDR has larger overall genome size and that relative
305 selection pressure is lower in NDR than SDR, is contrary to the predictions of the metabolic
306 versatility hypothesis of large genomes.
307
308 We hypothesize that relaxed selection and drift caused genome expansion in NDR. While these
309 same mechanisms are known to promote genome reduction in endosymbionts and obligate
310 pathogens (17, 18), it is important to recognize that these outcomes are not contradictory (Figure
311 6). Genome size is regulated by rates of gene gain and loss, the selective coefficient for each
312 gene in the genome, and the strength of selection. Endosymbionts and obligate intracellular
313 pathogens have small population sizes and accordingly, relaxed selection and stronger drift.
314 Relaxed selection pressure should lessen deletion bias. But under these conditions, host
315 compensation for microbial gene function radically alters selective coefficients of core genes,
316 thereby favoring genome reduction, and slightly deleterious mutations accumulate over time via
317 Muller's ratchet (71, 72). In addition, rates of HGT from non-host sources are essentially zero,
318 since there is little opportunity for endosymbionts to interact with other microbial cells, resulting
319 in a one way track to genome erosion. In contrast, for free-living microbes relaxed selection
320 pressure should bring about genome expansion by shifting the selective coefficients of accessory

genes towards neutral. For example, genome expansion in *Chlamydia* was driven by relaxed selection, recombination, and introgression (69). In this way small population size can favor genome erosion in endosymbionts, while also favoring genome expansion in free-living organisms (Figure 6). Meanwhile, free-living organisms that have large population sizes and high selection pressure will experience high rates of deletion that purge unnecessary genes in order to promote genome streamlining (14, 15).

Newly acquired genes tend to occur at low frequency in a population unless they provide an adaptive benefit (73), while adaptive genes will increase rapidly in frequency to join the core genome. These dynamics are believed to explain the characteristic U-shape of pangenome gene frequency distributions (51, 74). Deviations from U-shape expectations, including increased intermediate frequency genes, can result from changes in selection coefficients of genes or under conditions where HGT exceeds deletion rates (75). Alternatively, negative frequency dependent selection can cause highly beneficial genes to occur at low and intermediate frequencies (76, 77). A large portion of rare genes in microbial pangenomes are hypothetical proteins or genes of unknown function acquired through HGT (78, 79). For both NDR and SDR, approximately 60% of unique-rare genes (i.e., present in 1–2 strains) are annotated as hypothetical proteins. Nearly half of the 2,596 genes in NDR's intermediate-low frequency gene pool (i.e., present in 3–5 strains) are also hypothetical genes. Furthermore, intermediate-low frequency genes are similar to rare frequency genes in regards to GC content (Figure S3) and codon usage (Figure S4). These data are consistent with our hypotheses that NDR intermediate frequency genes represent evolutionarily recent HGT-gene acquisitions, which increased in frequency as a result of genome surfing.

344

345 HGT-mediated genome expansion supplies a reservoir of novel genetic material for the evolution
346 of gene families (25, 26), biosynthetic pathways (80), and formation of new metabolic networks
347 (81). Hence, the metabolic versatility of large genomes might be a classic example of an
348 evolutionary spandrel (82), an adaptive trait associated with large genomes that originated not
349 because of selection for versatility, but rather because the acquisition of diverse metabolic
350 pathways is a byproduct of non-adaptive evolutionary process that cause genome expansion.
351

352 We show that pangenome analysis of *Streptomyces* sister-taxa verifies several predictions of the
353 hypothesis that genome expansion within this clade was enabled by non-adaptive evolutionary
354 processes, most likely driven by late Pleistocene demography. We hypothesize that small
355 effective population size and relaxed selection, a consequence of geographic range expansion,
356 allowed for genes newly acquired by HGT to increase in frequency within the NDR pangenome
357 as a result of genome surfing. Further amplifying this effect is introgression of genes from local
358 gene pools encountered following dispersal into new environments. Non-adaptive genome
359 expansion is inherently a non-equilibrium process driven by a transient period of relaxed
360 selection, and population stabilization will re-impose selection pressures that favor deletion. At
361 this point, intermediate frequency genes will either be lost to deletion or fixed if they provide
362 adaptive benefits, and these processes will shift the pangenome structure back to U-shaped
363 expectations. These insights highlight the importance of considering population demography and
364 the profound influence of historical contingency on contemporary patterns of microbial genome
365 diversity.
366

367 **Material and Methods**

368 *Streptomyces isolation and genomic DNA extraction*

369 The strains in this study belong to a larger culture collection of *Streptomyces* isolated from
370 surface soils (0–5 cm) spanning sites across the United States (see (45, 46)) (Table S1). To
371 minimize the effects of environmental filtering in driving patterns of microbial diversity, we
372 selected sample locations with similar ecologies including meadow, pasture, or native grasslands
373 dominated by perennials and with moderately acidic to neutral soils (pH 6.0 ± 1.0, mean ± SD).

374

375 *Streptomyces* strains were isolated by plating air-dried soils on glycerol-arginine agar (pH 8.7)
376 plus cycloheximide and Rose Bengal (83, 84) as previously described (60). Genomic DNA was
377 extracted with a standard phenol/chloroform/isoamyl alcohol protocol from 72 h liquid cultures
378 grown at 30˚C with shaking in yeast extract-malt extract medium (YEME) + 0.5% glycine (59).

379

380 *Genome sequencing, assembly, and annotation*

381 Genome sequencing, assembly, and annotation is previously described (see (47)). Briefly, we
382 used the Nextera DNA Library Preparation Kit (Illumina, San Diego, CA, USA) to prepare
383 sequencing libraries. Genomes were sequenced on an Illumina HiSeq2500 instrument with
384 paired-end reads (2 x 100 bp). Genomes were assembled with the A5 pipeline (85) and annotated
385 with RAST (86). This generated high quality draft genome assemblies with over 25X coverage
386 and estimated completeness > 99% as assessed with CheckM (87). We used ITEP and MCL
387 clustering (inflation value = 2.0, cutoff = 0.04, maxbit score) (88) to identify orthologous
388 protein-coding gene clusters (i.e., genes). Genome sequences are available through NCBI under
389 BioProject PRJNA401484 accession numbers SAMN07606143–SAMN07606166.

390

391 *Phylogeny*

392 Phylogenetic relationships were reconstructed from whole genome alignments. We used Mugsy
393 (89) to generate multiple genome nucleotide alignments and trimAl v1.2 (90) for automatic
394 trimming of poorly aligned regions. Maximum likelihood (ML) trees were built using the
395 generalized time reversible nucleotide substitution model (91) with gamma distributed rate
396 heterogeneity among sites (GTRGAMMA) in RAxML v7.3.0(92), and bootstrap support was
397 determined following 20 ML searches with 100 inferences using the RAxML rapid bootstrapping
398 algorithm (93). Average nucleotide identity (ANI) was calculated from whole genome nucleotide
399 alignments using mothur (94).

400

401 *Pangenome and population genetics analyses*

402 The pangenome was determined from the gene content of 24 *Streptomyces* genomes (Table S2).
403 Strains in this collection were initially chosen for whole genome sequencing based on their
404 genetic similarity at house-keeping loci (see (46)). Subsequent analyses focused on recently
405 diverged sister-taxa clades of 10 genomes each, the northern-derived (NDR) and southern-
406 derived (SDR) lineages. Gene content patterns between strains and pangenome gene frequency
407 distributions were determined from gene presence/absence data.

408

409 Gene-level attributes across gene pools were determined from the average of all nucleotide
410 sequences within an orthologous protein-coding gene cluster (see above). GC content was
411 calculated for each gene using the R package Biostrings (95). Codon usage bias was calculated
412 for each gene using the R package cordon (96). Clade-level population genetic traits were

evaluated using 2,778 single-copy genes conserved across all 24 genomes. For each core gene, nucleotide sequences were aligned using MAFFT v.7 (97), and Gblocks (98) removed poorly aligned positions. PAL2NAL (99) generated codon alignments, and SNAP (100) calculated intra-clade non-synonymous ($K_A$) and synonymous ($K_S$) substitution rates (values > 2 were filtered prior to plotting and statistical analysis).

*Demographic simulation*

We assumed that the SDR pangenome approximates the gene frequency distribution of the last common ancestor of NDR and SDR. For the starting generation 0, we used the model from Marttinen *et al.* (55) to simulate a population of sequences and learn parameter values for rates of gene acquisition and deletion that produced the frequency distribution for SDR. To model range expansion demographics (i.e., severe bottleneck followed by exponential growth), we sampled 5 strains from generation 0 as the founding population for the subsequent generation, and simulated this for 100 generations. The simulated population had a growth rate of 5% per generation until a maximum of 100 individuals was reached. We varied the initial sizes of the founding population as well as the growth rate, and observed qualitatively similar results.

The model included gene acquisition events and deletion events similar to Marttinen *et al.* (55) but modified to allow for multiple changes. Instead of acquisitions/deletions happening independently, there were k=20 simultaneous acquisitions/deletions per strain per generation. The previous model (55) included a multiplicative fitness penalty of 0.99 for each gene exceeding a pre-specified genome size threshold. During the expansion, we relaxed the penalty for excess genes to $0.99^{(\text{current size/max size})}$ allowing for genome size variation.

436

437 **Data Availability**

438 *Streptomyces* genome sequences are available through NCBI under BioProject PRJNA401484

439 accession numbers SAMN07606143–SAMN07606166.

440

441 **Acknowledgements**

442 This work was supported by the National Science Foundation under Grant No. DEB-1456821

443 awarded to Daniel H. Buckley.

444

445 **References**

446 1. Vernikos G, Medini D, Riley DR, Tettelin H. 2015. Ten years of pan-genome analyses.

447 Curr Opin Microbiol 23:148–154.

448 2. Rasko DA, Rosovitz MJ, Myers GSA, Mongodin EF, Fricke WF, Gajer P, Crabtree J,

449 Sebaihia M, Thomson NR, Chaudhuri R, Henderson IR, Sperandio V, Ravel J. 2008. The

450 pangenome structure of *Escherichia coli:* comparative genomic analysis of *E. coli*

451 commensal and pathogenic isolates. J Bacteriol 190:6881–6893.

452 3. Reno ML, Held NL, Fields CJ, Burke PV, Whitaker RJ. 2009. Biogeography of the

453 *Sulfolobus islandicus* pan-genome. Proc Natl Acad Sci U S A 106:8605–8610.

454 4. Lefébure T, Bitar PDP, Suzuki H, Stanhope MJ. 2010. Evolutionary dynamics of complete

455 *Campylobacter* pan-genomes and the bacterial species concept. Genome Biol Evol 2:646–

456 655.

457 5. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R. 2005. The microbial pan-

458 genome. Curr Opin Genet Dev 15:589–594.

459 6. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV,
460 Crabtree J, Jones AL, Scott Durkin A, DeBoy RT, Davidsen TM, Mora M, Scarselli M, Ros
461 IM y., Peterson JD, Hauser CR, Sundaram JP, Nelson WC, Madupu R, Brinkac LM,
462 Dodson RJ, Rosovitz MJ, Sullivan SA, Daugherty SC, Haft DH, Selengut J, Gwinn ML,
463 Zhou L, Zafar N, Khouri H, Radune D, Dimitrov G, Watkins K, O'Connor KJB, Smith S,
464 Utterback TR, White O, Rubens CE, Grandi G, Madoff LC, Kasper DL, Telford JL,
465 Wessels MR, Rappuoli R, Fraser CM. 2005. Genome analysis of multiple pathogenic
466 isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome." Proc
467 Natl Acad Sci U S A 102:13950–13955.
468 7. McInerney JO, McNally A, O'Connell MJ. 2017. Why prokaryotes have pangenomes.
469 Nature Microbiology 2:17040.
470 8. Brockhurst MA, Harrison E, Hall JPJ, Richards T, McNally A, MacLean C. 2019. The
471 ecology and evolution of pangenomes. Curr Biol 29:R1094–R1103.
472 9. Azarian T, Huang I-T, Hanage WP. 2020. Structure and Dynamics of Bacterial Populations:
473 Pangenome Ecology, p. . *In* Tettelin, H, Medini, D (eds.), The Pangenome: Diversity,
474 Dynamics and Evolution of Genomes. Springer, Cham (CH).
475 10. Lynch M. 2006. Streamlining and simplification of microbial genome architecture. Annu
476 Rev Microbiol 60:327–349.
477 11. Fraser CM, Eisen J, Fleischmann RD, Ketchum KA, Peterson S. 2000. Comparative
478 genomics and understanding of microbial biology. Emerg Infect Dis 6:505–512.
479 12. Kuo C-H, Moran NA, Ochman H. 2009. The consequences of genetic drift for bacterial
480 genome complexity. Genome Res 19:1450–1454.
481 13. Puigbò P, Lobkovsky AE, Kristensen DM, Wolf YI, Koonin EV. 2014. Genomes in

482 turmoil: quantification of genome dynamics in prokaryote supergenomes. BMC Biol 12:66.
483 14. Viklund J, Ettema TJG, Andersson SGE. 2012. Independent genome reduction and
484 phylogenetic reclassification of the oceanic SAR11 clade. Mol Biol Evol 29:599–615.
485 15. Giovannoni SJ, Cameron Thrash J, Temperton B. 2014. Implications of streamlining theory
486 for microbial ecology. ISME J 8:1553–1565.
487 16. Mira A, Ochman H, Moran NA. 2001. Deletional bias and the evolution of bacterial
488 genomes. Trends Genet 17:589–596.
489 17. Moran NA. 2002. Microbial minimalism: Minireview genome reduction in bacterial
490 pathogens. Cell 108:583–586.
491 18. McCutcheon JP, Moran NA. 2011. Extreme genome reduction in symbiotic bacteria. Nat
492 Rev Microbiol 10:13–26.
493 19. Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D, Bibbs L, Eads J,
494 Richardson TH, Noordewier M, Rappé MS, Short JM, Carrington JC, Mathur EJ. 2005.
495 Genome streamlining in a cosmopolitan oceanic bacterium. Science 309:1242–1245.
496 20. Brewer TE, Handley KM, Carini P, Gilbert JA, Fierer N. 2016. Genome reduction in an
497 abundant and ubiquitous soil bacterium "*Candidatus* Udaeobacter copiosus." Nat Microbiol
498 2:16198.
499 21. Konstantinidis KT, Tiedje JM. 2004. Trends between gene content and genome size in
500 prokaryotic species with larger genomes. Proc Natl Acad Sci U S A 101:3160–3165.
501 22. Dini-Andreote F, Andreote FD, Araújo WL, Trevors JT, van Elsas JD. 2012. Bacterial
502 genomes: habitat specificity and uncharted organisms. Microb Ecol 64:1–7.
503 23. Han K, Li Z-F, Peng R, Zhu L-P, Zhou T, Wang L-G, Li S-G, Zhang X-B, Hu W, Wu Z-H,
504 Qin N, Li Y-Z. 2013. Extraordinary expansion of a *Sorangium cellulosum* genome from an

505 alkaline milieu. Sci Rep 3:2101.
506 24. Cordero OX, Hogeweg P. 2009. The impact of long-distance horizontal gene transfer on
507 prokaryotic genome size. Proc Natl Acad Sci U S A 106:21748–21753.
508 25. Lerat E, Daubin V, Ochman H, Moran NA. 2005. Evolutionary origins of genomic
509 repertoires in bacteria. PLoS Biol 3:e130.
510 26. Treangen TJ, Rocha EPC. 2011. Horizontal transfer, not duplication, drives the expansion
511 of protein families in prokaryotes. PLoS Genet 7:e1001284.
512 27. Bohlin J, Brynildsrud OB, Sekse C, Snipen L. 2014. An evolutionary analysis of genome
513 expansion and pathogenicity in *Escherichia coli*. BMC Genomics 15:882.
514 28. Tsai Y-M, Chang A, Kuo C-H. 2018. Horizontal gene acquisitions contributed to genome
515 expansion in insect-symbiotic *Spiroplasma clarkii*. Genome Biol Evol 10:1526–1532.
516 29. Wright S. 1931. Evolution in Mendelian populations. Genetics 16:97–159.
517 30. Kimura M. 1968. Evolutionary rate at the molecular level. Nature 217:624–626.
518 31. Bobay L-M, Ochman H. 2017. The Evolution of Bacterial Genome Architecture. Front
519 Genet 8:72.
520 32. Rocha EPC. 2018. Neutral theory, microbial practice: Challenges in bacterial population
521 genetics. Mol Biol Evol 35:1338–1347.
522 33. Roselló-Mora, Ramon, Amann, Rudolf. 2001. The species concept for prokaryotes. FEMS
523 Microbiol Rev 25:39–67.
524 34. Achtman M, Wagner M. 2008. Microbial diversity and the genetic nature of microbial
525 species. Nat Rev Microbiol 6:431–440.
526 35. Shapiro BJ. 2019. What Microbial Population Genomics Has Taught Us About Speciation,
527 p. 31–47. *In* Polz, MF, Rajora, OP (eds.), Population Genomics: Microorganisms. Springer

528 International Publishing, Cham.
529 36. Smith NH, Dale J, Inwald J, Palmer S, Gordon SV, Hewinson RG, Smith JM. 2003. The
530 population structure of *Mycobacterium bovis* in Great Britain: clonal expansion. Proc Natl
531 Acad Sci U S A 100:15271–15275.
532 37. Achtman M. 2004. Population structure of pathogenic bacteria revisited. Int J Med
533 Microbiol 294:67–73.
534 38. Nübel U, Dordel J, Kurt K, Strommenger B, Westh H, Shukla SK, Zemlicková H, Leblois
535 R, Wirth T, Jombart T, Balloux F, Witte W. 2010. A timescale for evolution, population
536 expansion, and spatial spread of an emerging clone of methicillin-resistant *Staphylococcus*
537 *aureus.* PLoS Pathog 6:e1000855.
538 39. Wirth T, Hildebrand F, Allix-Béguec C, Wölbeling F, Kubica T, Kremer K, van Soolingen
539 D, Rüsch-Gerdes S, Locht C, Brisse S, Meyer A, Supply P, Niemann S. 2008. Origin,
540 spread and demography of the *Mycobacterium tuberculosis* complex. PLoS Pathog
541 4:e1000160.
542 40. Takuno S, Kado T, Sugino RP, Nakhleh L, Innan H. 2012. Population genomics in bacteria:
543 a case study of *Staphylococcus aureus*. Mol Biol Evol 29:797–809.
544 41. Cornejo OE, Lefebure T, Pavinski 2. Paulina, Lang P, Richards 2. Vincent P., Eilertson K,
545 Do T, Beighton D, Zeng L, Ahn S-J, Burne RA, Siepel A, Bustamante CD, Stanhope MJ.
546 Evolutionary and population genomics of the cavity causing bacteria *Streptococcus mutans*.
547 Evolution 30:881–893.
548 42. Montano V, Didelot X, Foll M, Linz B, Reinhardt R, Suerbaum S, Moodley Y, Jensen JD.
549 2015. Worldwide population structure, long-term demography, and local adaptation of
550 *Helicobacter pylori.* Genetics 200:947–963.

551 43. Hewitt G. 1996. Some genetic consequences of ice ages, and their role in divergence and
552 speciation. Biol J Linn Soc 3:247–276.
553 44. Hewitt GM. 2004. Genetic consequences of climatic oscillations in the Quaternary. Philos
554 Trans R Soc Lond B Biol Sci 359:183–195.
555 45. Andam CP, Doroghazi JR, Campbell AN, Kelly PJ, Choudoir MJ, Buckley DH. 2016. A
556 latitudinal diversity gradient in terrestrial bacteria of the genus *Streptomyces*. MBio
557 7:e02200–15.
558 46. Choudoir MJ, Doroghazi JR, Buckley DH. 2016. Latitude delineates patterns of
559 biogeography in terrestrial *Streptomyces*. Environ Microbiol 18:4931–4945.
560 47. Choudoir MJ, Buckley DH. 2018. Phylogenetic conservatism of thermal traits explains
561 dispersal limitation and genomic differentiation of *Streptomyces* sister-taxa. ISME J
562 12:2176–2186.
563 48. Choudoir MJ, Panke-Buisse K, Andam CP, Buckley DH. 2017. Genome surfing as driver of
564 microbial genomic diversity. Trends Microbiol 25:624–636.
565 49. Kim M, Oh H-S, Park S-C, Chun J. 2014. Towards a taxonomic coherence between average
566 nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of
567 prokaryotes. Int J Syst Evol Microbiol 64:346–351.
568 50. Ciufo S, Kannan S, Sharma S, Badretdin A, Clark K, Turner S, Brover S, Schoch CL,
569 Kimchi A, DiCuccio M. 2018. Using average nucleotide identity to improve taxonomic
570 assignments in prokaryotic genomes at the NCBI. Int J Syst Evol Microbiol 68:2386–2392.
571 51. Haegeman B, Weitz JS. 2012. A neutral theory of genome evolution and the frequency
572 distribution of genes. BMC Genomics 13:1.
573 52. Wright F. 1990. The “effective number of codons” used in a gene. Gene 87:23–29.

574 53. Slatkin M, Excoffier L. 2012. Serial founder effects during range expansion: a spatial
575 analog of genetic drift. Genetics 191:171–181.
576 54. Kimura M. 1983. The Neutral Theory of Molecular Evolution. Cambridge: Cambridge
577 University Press.
578 55. Marttinen P, Croucher NJ, Gutmann MU, Corander J, Hanage WP. 2015. Recombination
579 produces coherent bacterial species clusters in both core and accessory genomes. Microbial
580 Genomics 1.
581 56. Edmonds CA, Lillie AS, Cavalli-Sforza LL. 2004. Mutations arising in the wave front of an
582 expanding population. Proc Natl Acad Sci U S A 101:975–979.
583 57. Travis JMJ, Münkemüller T, Burton OJ, Best A, Dytham C, Johst K. 2007. Deleterious
584 mutations can surf to high densities on the wave front of an expanding population. Mol Biol
585 Evol 24:2334–2343.
586 58. Chuang A, Peterson CR. 2016. Expanding population edges: theories, traits, and trade-offs.
587 Glob Chang Biol 2:494–512.
588 59. Kieser, T, Bibb, MJ, Buttner, MJ, Charter, KF, Hopwood, DA. 2000. Practical *Streptomyces*
589 Genetics. John Innes Foundation, Norwich, UK.
590 60. Doroghazi JR, Buckley DH. 2010. Widespread homologous recombination within and
591 between *Streptomyces* species. ISME J 4:1136.
592 61. Doroghazi JR, Buckley DH. 2014. Intraspecies comparison of *Streptomyces pratensis*
593 genomes reveals high levels of recombination and gene conservation between strains of
594 disparate geographic origin. BMC Genomics 15:970.
595 62. Cheng K, Rong X, Huang Y. 2016. Widespread interspecies homologous recombination
596 reveals reticulate evolution within the genus *Streptomyces*. Mol Phylogenet Evol 102:246–

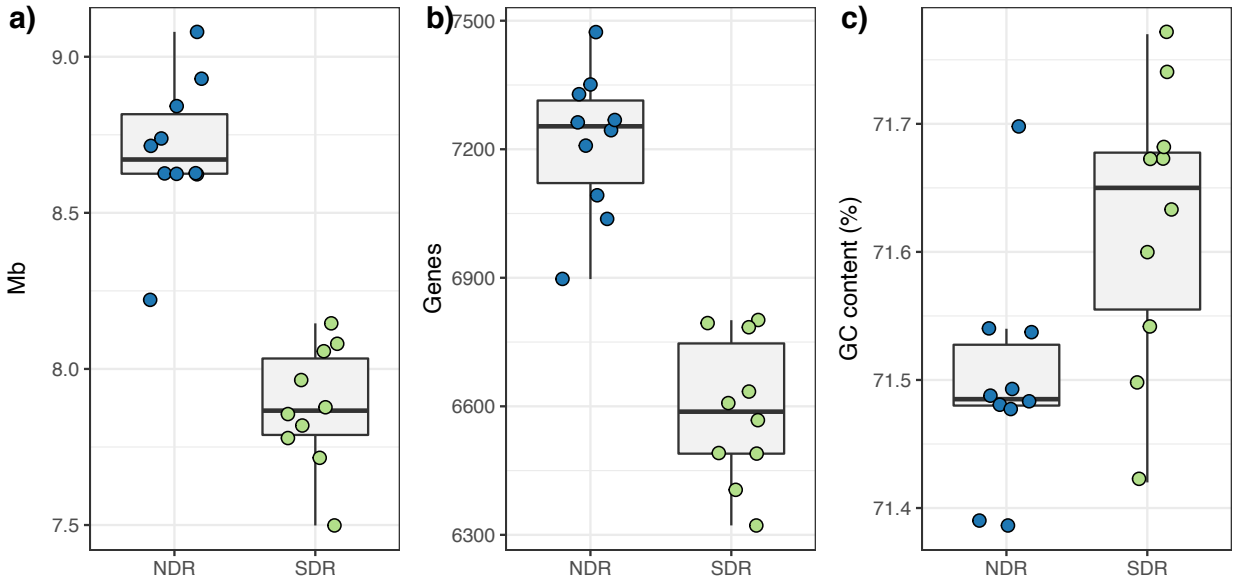
597 254.
598 63. Andam CP, Choudoir MJ, Vinh Nguyen A, Sol Park H, Buckley DH. 2016. Contributions
599 of ancestral inter-species recombination to the genetic diversity of extant *Streptomyces*
600 lineages. ISME J 10:1731–1741.
601 64. Currat M, Ruedi M, Petit RJ, Excoffier L. 2008. The hidden side of invasions: massive
602 introgression by local genes. Evolution 62:1908–1920.
603 65. Gogarten JP, Townsend JP. 2005. Horizontal gene transfer, genome innovation and
604 evolution. Nat Rev Microbiol 3:679–687.
605 66. Excoffier L, Foll M, Petit RJ. 2009. Genetic consequences of range expansions. Annu Rev
606 Ecol Evol Syst 40:481–501.
607 67. Hildebrand F, Meyer A, Eyre-Walker A. 2010. Evidence of selection upon genomic GC-
608 content in bacteria. PLoS Genet 6:e1001107.
609 68. Raghavan R, Kelkar YD, Ochman H. 2012. A selective force favoring increased G+C
610 content in bacterial genes. Proc Natl Acad Sci U S A 109:14504–14507.
611 69. Bohlin J. 2015. Genome expansion in bacteria: the curious case of *Chlamydia trachomatis.*
612 BMC Res Notes 8:512.
613 70. Hallatschek O, Hersen P, Ramanathan S, Nelson DR. 2007. Genetic drift at expanding
614 frontiers promotes gene segregation. Proc Natl Acad Sci U S A 104:19926–19930.
615 71. Moran NA. 1996. Accelerated evolution and Muller's rachet in endosymbiotic bacteria.
616 Proc Natl Acad Sci U S A 93:2873–2878.
617 72. Rispe C, Moran NA. 2000. Accumulation of deleterious mutations in endosymbionts:
618 Muller's ratchet with two levels of selection. Am Nat 156:425–441.
619 73. Kuo C-H, Ochman H. 2009. The fate of new bacterial genes. FEMS Microbiol Rev 33:38–

620 43.
621 74. Lobkovsky AE, Wolf YI, Koonin EV. 2013. Gene frequency distributions reject a neutral
622 model of genome evolution. Genome Biol Evol 5:233–242.
623 75. Domingo-Sananes MR, McInerney JO. 2019. Selection-based model of prokaryote
624 pangenomes. bioRxiv doi.org/10.1101/782573
625 76. Cordero OX, Ventouras L-A, DeLong EF, Polz MF. 2012. Public good dynamics drive
626 evolution of iron acquisition strategies in natural bacterioplankton populations. Proc Natl
627 Acad Sci U S A 109:20059–20064.
628 77. McNally A, Kallonen T, Connor C, Abudahab K, Aanensen DM, Horner C, Peacock SJ,
629 Parkhill J, Croucher NJ, Corander J. 2019. Diversification of colonization factors in a
630 multidrug-resistant *Escherichia coli* lineage evolving under negative frequency-dependent
631 selection. MBio 10:e00644–19.
632 78. Mira A, Klasson L, Andersson SGE. 2002. Microbial genome evolution: sources of
633 variability. Curr Opin Microbiol 5:506–512.
634 79. Mira A, Martín-Cuadrado AB, D'Auria G, Rodríguez-Valera F. 2010. The bacterial pan-
635 genome:a new paradigm in microbiology. Int Microbiol 13:45–57.
636 80. Boucher Y, Doolittle WF. 2000. The role of lateral gene transfer in the evolution of
637 isoprenoid biosynthesis pathways. Mol Microbiol 37:703–716.
638 81. Pál C, Papp B, Lercher MJ. 2005. Adaptive evolution of bacterial metabolic networks by
639 horizontal gene transfer. Nat Genet 37:1372–1375.
640 82. Gould SJ, Lewontin RC. 1979. The spandrels of San Marco and the Panglossian paradigm:
641 a critique of the adaptationist programme. Proc R Soc Lond B Biol Sci 205:581–598.
642 83. El-Nakeeb MA, Lechevalier HA. 1963. Selective isolation of aerobic Actinomycetes. Appl

643 Microbiol 11:75–77.
644 84. Ottow JCG. 1972. Rose Bengal as a selective aid in the isolation of fungi and actinomycetes
645 from natural sources. Mycologia 64:304.
646 85. Tritt A, Eisen JA, Facciotti MT, Darling AE. 2012. An integrated pipeline for *de novo*
647 assembly of microbial genomes. PLoS One 7:e42304.
648 86. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S,
649 Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil
650 LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O,
651 Vonstein V, Wilke A, Zagnitko O. 2008. The RAST server: Rapid annotations using
652 subsystems technology. BMC Genomics 9:75.
653 87. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing
654 the quality of microbial genomes recovered from isolates, single cells, and metagenomes.
655 Genome Res 25:1043–1055.
656 88. Benedict MN, Henriksen JR, Metcalf WW, Whitaker RJ, Price ND. 2014. ITEP: an
657 integrated toolkit for exploration of microbial pan-genomes. BMC Genomics 15:8.
658 89. Angiuoli SV, Salzberg SL. 2011. Mugsy: fast multiple alignment of closely related whole
659 genomes. Bioinformatics 27:334–342.
660 90. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. 2009. trimAl: a tool for automated
661 alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972–1973.
662 91. Tavare S. 1986. Some probabilistic and statistical problems in the analysis of DNA
663 sequences. Lectures Math Life Sci 17: 57-86 58. Warscheid T, Braams J (2000)
664 Biodeterioration of stone: a review. Int Biodeterior Biodegradation 46:343–368.
665 92. Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses

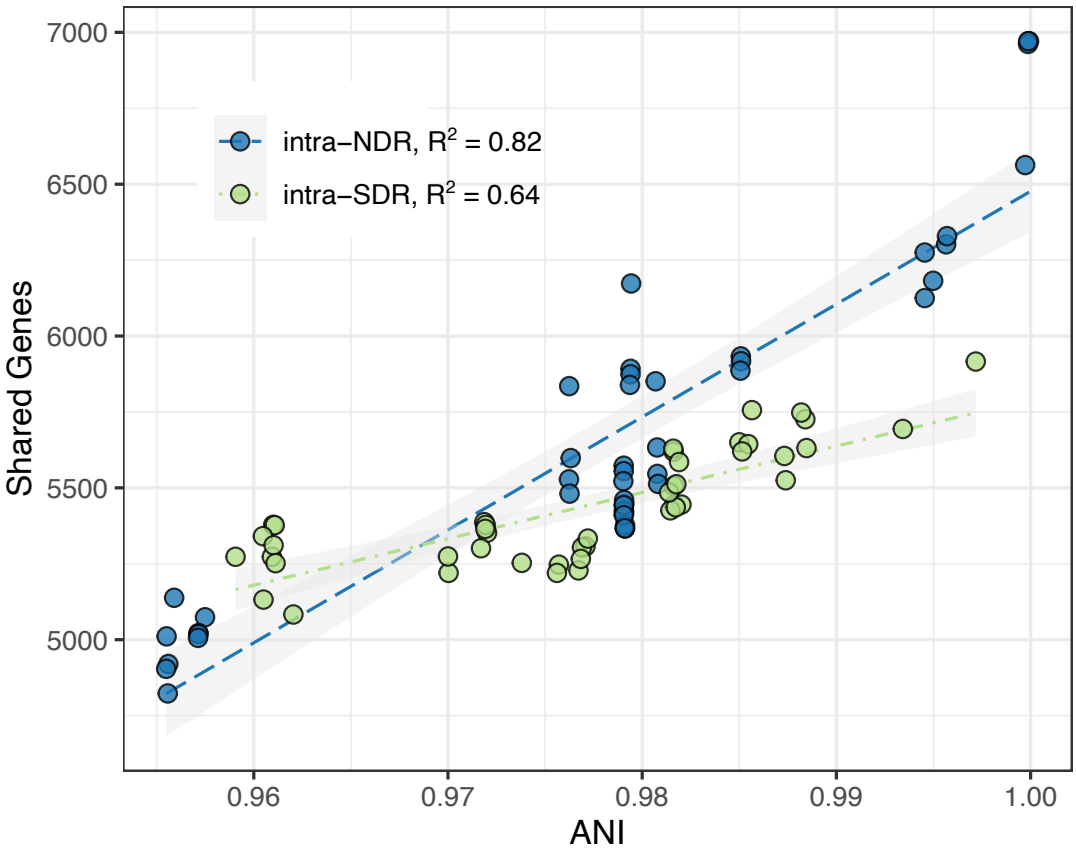
666 with thousands of taxa and mixed models. Bioinformatics 22:2688–2690.
667 93. Stamatakis A, Hoover P, Rougemont J, Renner S. 2008. A rapid bootstrap algorithm for the
668 RAxML web servers. Syst Biol 57:758–771.
669 94. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA,
670 Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ,
671 Weber CF. 2009. Introducing mothur: open-source, platform-independent, community-
672 supported software for describing and comparing microbial communities. Appl Environ
673 Microbiol 75:7537–7541.
674 95. Pagès HA, Gentleman P, DebRoy R. 2020. Biostrings: Efficient manipulation of biological
675 strings. R package version 2.59.
676 96. Elek A, Kuzman M, Vlahoviček K. 2019. coRdon: codon usage analysis and prediction of
677 gene expressivity. R package version 1.8.0.
678 97. Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7:
679 improvements in performance and usability. Mol Biol Evol 30:772–780.
680 98. Talavera, Gerard, Castresana, Jose. 2007. Improvement of phylogenies after removing
681 divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol
682 56:564–577.
683 99. Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence
684 alignments into the corresponding codon alignments. Nucleic Acids Res 34:W609–12.
685 100. Korber BT. 2000. HIV Signature and Sequence Variation Analysis, Chapter 4, pages 55–72.
686 *In* Allen G. Rodrigo and Gerald H. Learn (ed.), Computational Analysis of HIV Molecular
687 Sequences. Dordrecht, Netherlands: Kluwer Academic Publishers.
688

689 **Figure Legends**




690

691 **Figure 1.** Genomic attributes of NDR and SDR sister-taxa. NDR genomes are larger, have more
692 genes, and have lower GC content compared to SDR genomes. Plots show the distributions of
693 genome size in Mb (a), number of genes (b), and genome-wide GC content (%) (c) for
694 *Streptomyces* sister-taxa. Boxplots show the clade-level medians, interquartile ranges, and 1.5
695 times interquartile ranges. Colored circles illustrate the values for individual genomes belonging
696 to the NDR clade (blue) or the SDR clade (green).




697

698 **Figure 2.** Genomic similarity versus shared gene content for NDR and SDR. Differences in
699 shared gene content across increasing average nucleotide identity (ANI) are greater within the
700 NDR clade compared to the SDR clade (Table S3). Circles show pairwise comparisons of the
701 number of shared genes between two strains versus ANI and are colored by clade according to
702 the legend. Dashed lines show linear regressions, and the shaded area is the 95% confidence
703 interval.

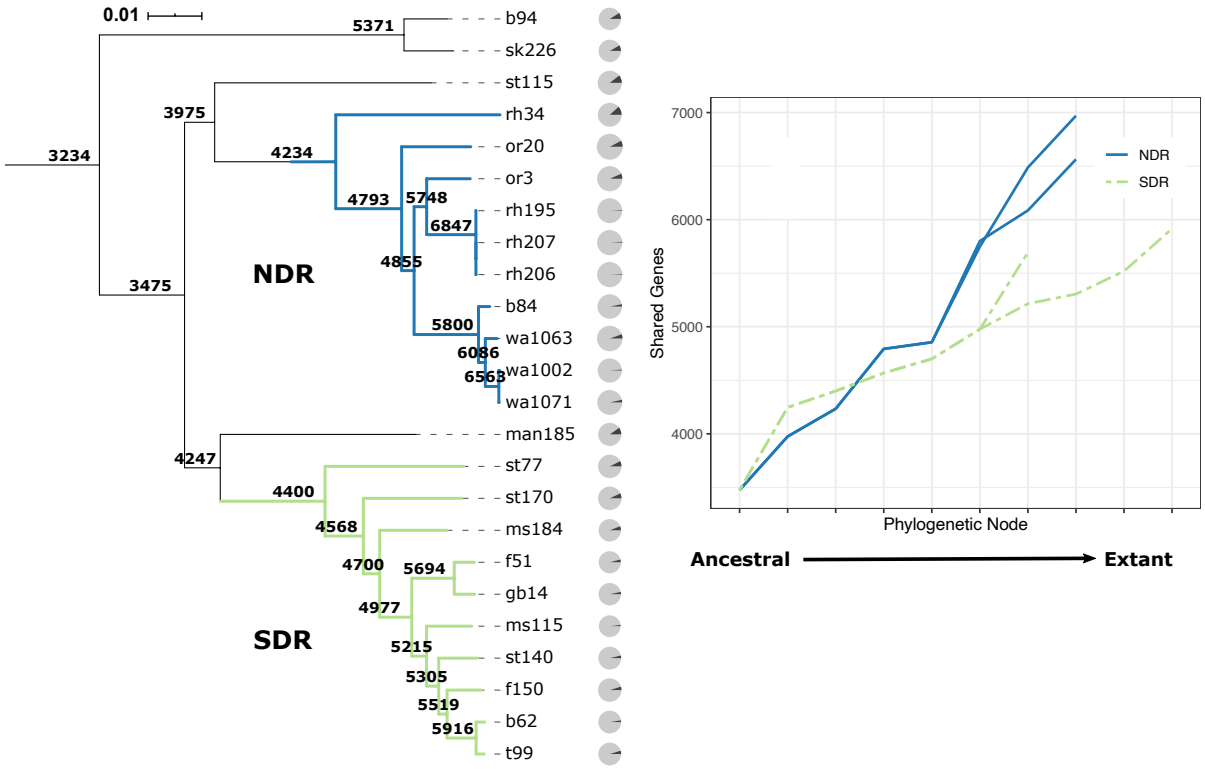
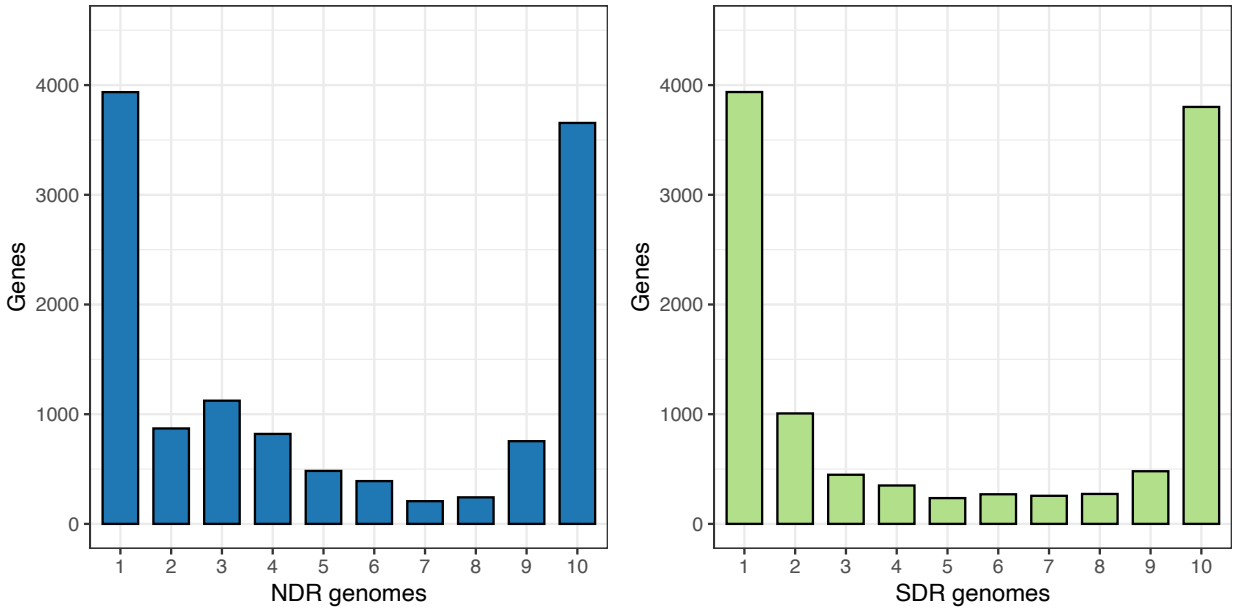



**Figure 3.** Presence/absence of genes across phylogeny. Gene content changes more rapidly across ancestral phylogenetic nodes for NDR genomes compared to SDR genomes. Tree is made from whole genome nucleotide alignments, and the scale bar shows nucleotide substitutions per site (see Figure S1). Branch colors reflect clade membership. Phylogenetic nodes are labeled with the number of genes conserved in all members of descendent nodes. Gray pie charts at tree tips show the portion of total genes per genome that are strain-specific (black slice). Right panel plots the differences in gene content across the phylogeny beginning at the shared ancestral node and ending with extant taxa at the terminal tips for NDR (blue-solid) and SDR (green-dashed) lineages. Multiple lines represent monophyletic lineages.



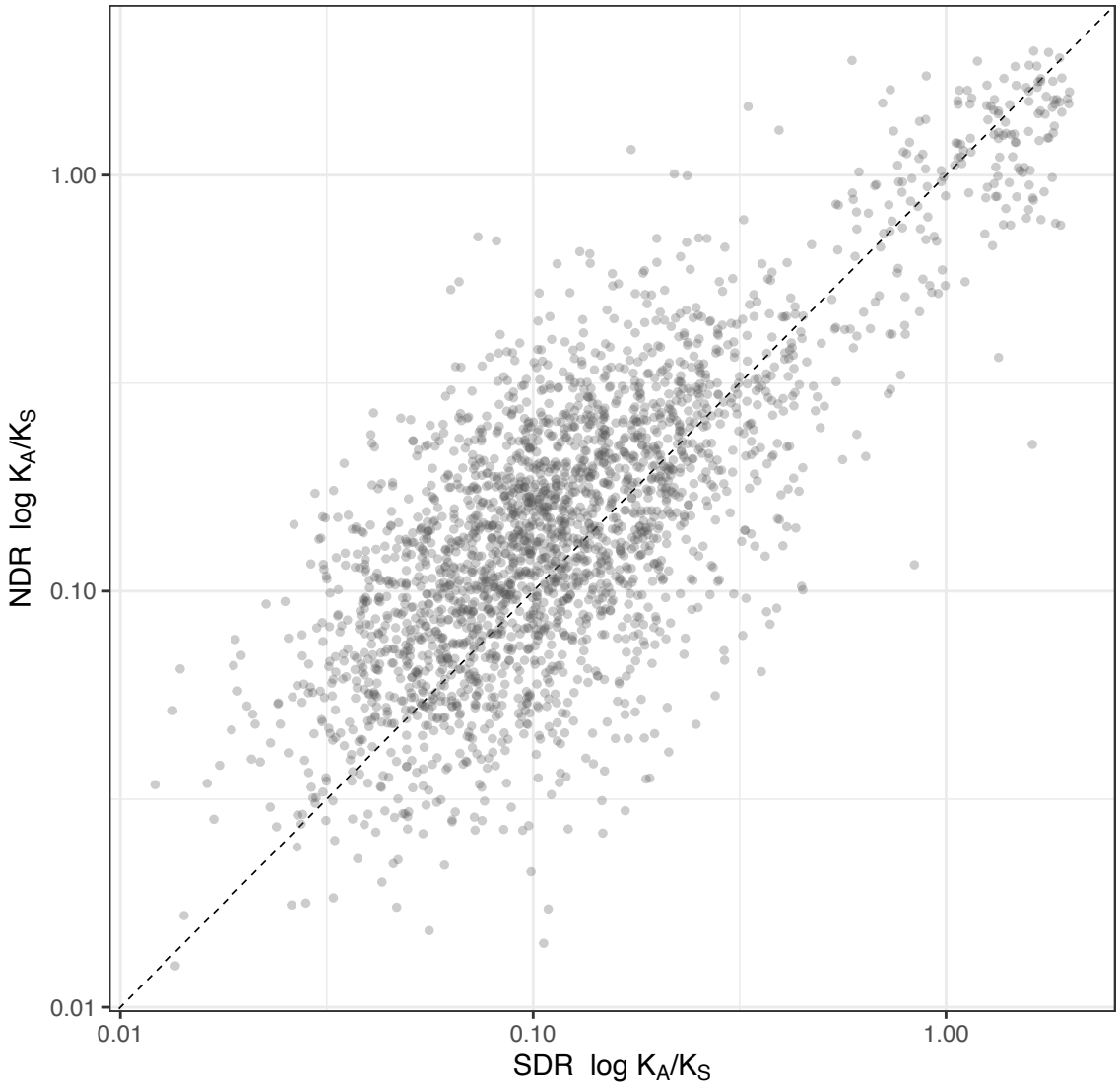

714

715 **Figure 4.** Pangenome gene frequency distributions. NDR genomes are enriched in intermediate

716 frequency genes. Plots show the pangenome gene frequency distributions for NDR (left) and

717 SDR (right). Bars show the population-level sums of genes present in 1–10 genomes. See Table

718 S3 for raw values and proportions.




719

720 **Figure 5.** $K_A/K_S$ values between the NDR and SDR sister-taxa core genome. NDR core genes
721 have, on average, greater rates of non-synonymous to synonymous amino acid substitutions
722 compared to SDR core genes. Circles plot clade-level rates of non-synonymous to synonymous
723 amino acid substitutions ($K_A/K_S$) for each of 2,444 single-copy core genes for NDR (y-axis) and
724 SDR (x-axis). Axes are logarithmic scale. The black dashed line is a slope of 1, and points along
725 this line are genes with equal $K_A/K_S$ mean values in both clades. $K_A/K_S$ is proportional to the
726 relative strength of genetic drift and inversely proportional to the relative strength of selection.

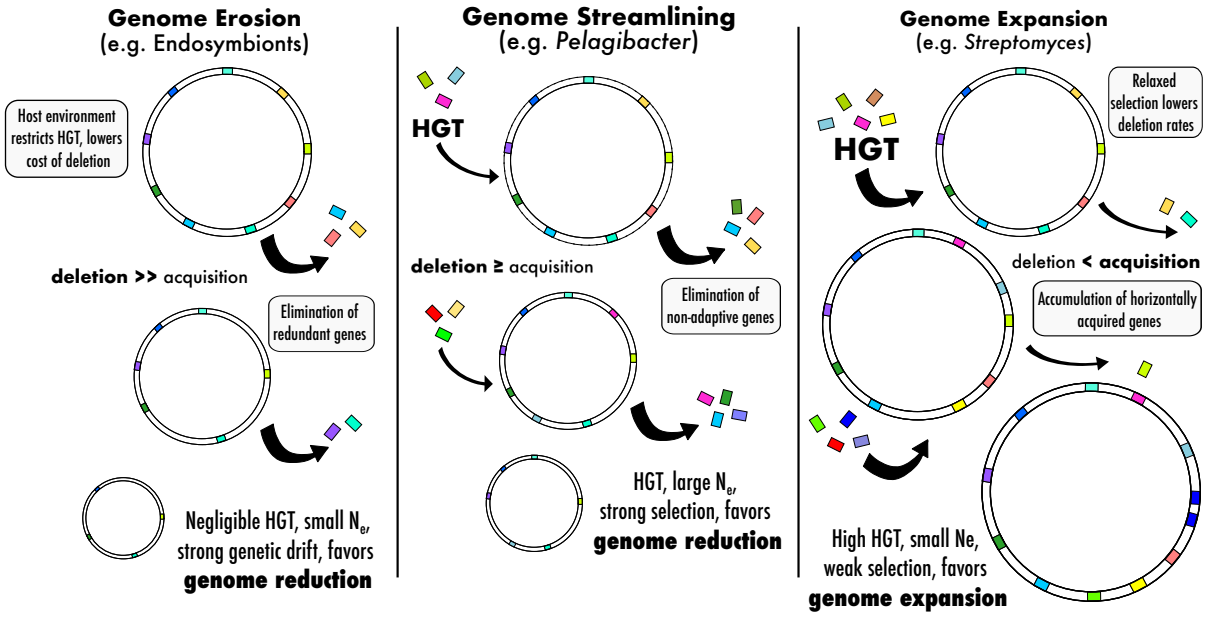



**Figure 6.** Conceptual overview of the evolutionary processes and demographic conditions that support changes in genome size. Genome erosion (left) in endosymbionts is the result of small $N_e$ and strong genetic drift, with host compensation lowering costs of deletion while restricting gene flow (HGT). Genome streamlining (middle) in free-living microbes with large populations like *Pelagibacter* involves strong selection and elimination of non-adaptive genes. Genome expansion (right) in *Streptomyces* is facilitated by high rates of HGT and relaxed selection, allowing for the accumulation of non-adaptive genes and ultimately larger genomes.

## Supplemental Material

**Table S1.** Strain sample location and metadata.

**Table S2.** *Streptomyces* genomic attributes and metadata.

**Table S3.** Shared gene content and genomic similarity linear model summary.

**Table S4.** Pangenome frequency distributions.

**Figure S1**. Whole genome phylogeny and map of sample locations.

742 **Figure S2.** Genome-wide gene density plots.

743 **Figure S3.** Mean GC content across gene frequency pools.

744 **Figure S4.** Mean codon bias across gene frequency pools.

745 **Figure S5.** Demographic simulation.