Title: Highly-resolved within-species dynamics in the human facial skin microbiome

Authors: Jacob S. Baker[1,2], Evan Qu[1,2], Christopher P. Mancuso[1,2], A. Delphine Tripp[1,2,3], Arolyn Conwill[1,2], Tami D. Lieberman[1,2,4,5*]

Affiliations:

[1]Institute for Medical Engineering and Sciences, Massachusetts Institute of Technology; Cambridge, MA 02139, USA.

[2]Department of Civil and Environmental Engineering, Massachusetts Institute of Technology; Cambridge, MA 02139, USA.

[3]Department of Systems Biology, Harvard University; Cambridge, MA 02138, USA.

[4]Broad Institute of MIT and Harvard; Cambridge, MA 02139, USA.

[5]Ragon Institute of MGH, MIT, and Harvard; Cambridge, MA 02139, USA.

*Corresponding author. Email: tami@mit.edu

Abstract: Human facial skin microbiomes (FSMs) on adults are dominated by just two bacterial species, *Cutibacterium acnes* and *Staphylococcus epidermidis.* Underlying this apparent simplicity, each FSM harbors multiple strains of both species whose assembly dynamics on individuals are unknown. Here, we use 4,055 isolate genomes and 360 metagenomes to trace the dynamics of strains on individuals and their transmission. Strains are shared amongst family members of all ages, but each individual harbors unique strain consortia. Strain stability changes upon formation of the adult-type FSM: *S. epidermidis* lineage turnover slows, and the rate of *C. acnes* colonization increases before stabilizing, suggesting this transitional window could facilitate engraftment of therapeutic strains. Our work reveals previously undetectable community dynamics and informs the design of therapeutic interventions.

One-Sentence Summary: Previously undetectable gain and loss of individual strains underlies the species stability of important microbiome bacteria.

Main Text:

Although human facial skin is constantly exposed to the diverse microorganisms of our surrounding environments, only select species and strains are able to colonize and persist (*1–5*). Notably, two globally prevalent skin dwelling species, *Cutibacterium acnes* and *Staphylococcus epidermidis,* predominate in nearly every adult facial skin microbiome (FSM) sampled to date (together representing about ~80% of the average individual's relative abundance) (*4*, *6*). The

universality of this composition across individuals and its simplicity relative to gut communities makes mechanistic study of this site's ecology more easily attainable.

Despite commonality at the species level and the potential for transmission across individuals, each person's FSM is composed of strains (groups within a species that share common ancestry (*7*, *8*) which are distinct from those found on other individuals (*5*). This strain-level individuality indicates the presence of unknown ecological barriers which hinder the migration of new strains, as evidenced by the difficulty of achieving durable engraftment of topically applied *C. acnes (9)*. Determining the factors that facilitate successful colonization amidst unknown ecological barriers is of critical importance for developing durable probiotics, predicting the consequences of antibiotic interventions, and contextualizing the differences between healthy and diseased microbiomes (*10*, *11*).

Intraspecies dynamics in microbial communities have remained difficult to measure due to the limited resolution of metagenomics when intraspecies diversity is high (*12*). Here, we use thousands of cultures from individual bacterial cells to obtain whole-genomes and avoid these limitations. This resolution enables us to cluster intraspecies populations into extremely closely-related clades that originate on individual people, which we refer to as lineages (*13*, *14*). Cells from a lineage are separated by less than 100 point mutations across their entire genomes, reflecting their recent origin on an individual person, and are related by direct or indirect transmission (*7*, *15–17*).

From a cohort of families we describe the species and lineage-level dynamics of the FSM during the transition to adulthood and beyond. While birth and infancy have received extensive attention as critical periods for the assembly of human microbiomes (*18–21*), the role of the years between infancy and adulthood in facilitating strain colonization is underexplored. There is a well-documented transformation of the skin during this time, during which lipid secretion surges (*22*) and bacterial density in the FSM increases more than 10,000-fold per cm^2 (*23–26*). We ask whether this transition is associated with changes in intraspecies composition, alters the stability of already colonizing lineages, or enhances the capacity of migrating lineages to successfully colonize. We amassed a unique collection comprising 2,025 *S. epidermidis* and 2,030 *C. acnes* isolates from 57 individuals, supplemented with associated metagenomes. This extraordinary resolution and depth enables analyses traversing taxonomic levels from species down to individual genotypes, providing mechanistic insights into the intraspecies dynamics associated with niche expansion in the FSM.

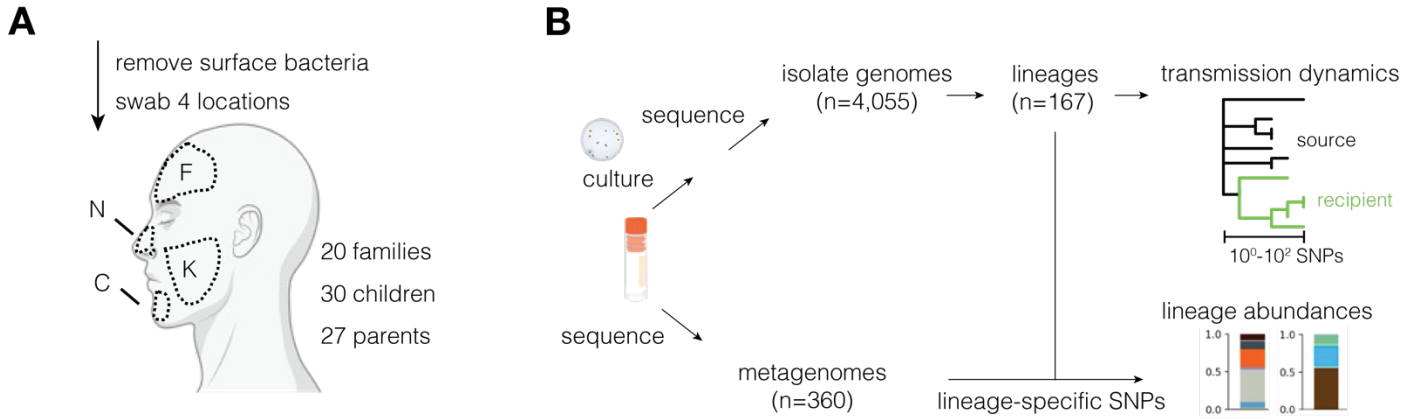

Fig. 1. Integrating isolate genomes with metagenomics to resolve within-species facial skin microbiomes at the lineage level. (A) We collected samples of the Facial Skin Microbiomes (FSMs) from four facial skin sites from 57 people belonging to 20 distinct nuclear families. **(B)** From 8 of these nuclear families (top), we obtained 4,055 high-quality genomes clustered in 167 distinct *C. acnes* and *S. epidermidis* lineages, and used their phylogenetic topologies to analyze transmission events. Concurrently (bottom) we sequenced metagenomic samples from all subjects and combined these reads with information obtained from isolate lineages (lineage-specific SNPs) to identify lineages in metagenomics tubes across samples for which we had no isolates, as well as other metrics of intraspecies and interspecies diversity.

Species-level formation of the adult-type FSM

We collected facial skin microbiome samples from children and their parents at a single K-8 school across a 2.5 year period (Fig. 1). Individuals were sampled opportunistically on 5 different dates, and not every family member was sampled at the same time. In total, we sampled 30 children (aged 5 to 15) and 27 parents (aged 34 to 61). Basic health and demographic information was acquired via a questionnaire (Methods); given the modest size of our cohort, this information was primarily used for hypothesis generation.

Four facial locations were sampled from each subject using a pre-wet swab (forehead, nose, cheek, and chin). To limit collection of surface contaminants, each subject's face was washed using a gentle cleansing cloth, rinsed with a saline wipe, and air dried. This process enabled repopulation of the surface from sebaceous secretions during the drying period (*1*, *27*). From each sample, we profiled the entire community using metagenomic sequencing (mean non-human reads/sample = 4.5×10^6), enabling study of both species-level and within-species diversity. We found only minor compositional differences amongst facial sites (Fig. S1), consistent with previous results (*4*, *13*); we therefore grouped metagenomic reads across sites for each timepoint (herein called 'samples') to increase sensitivity to lowly abundant intraspecies groups.

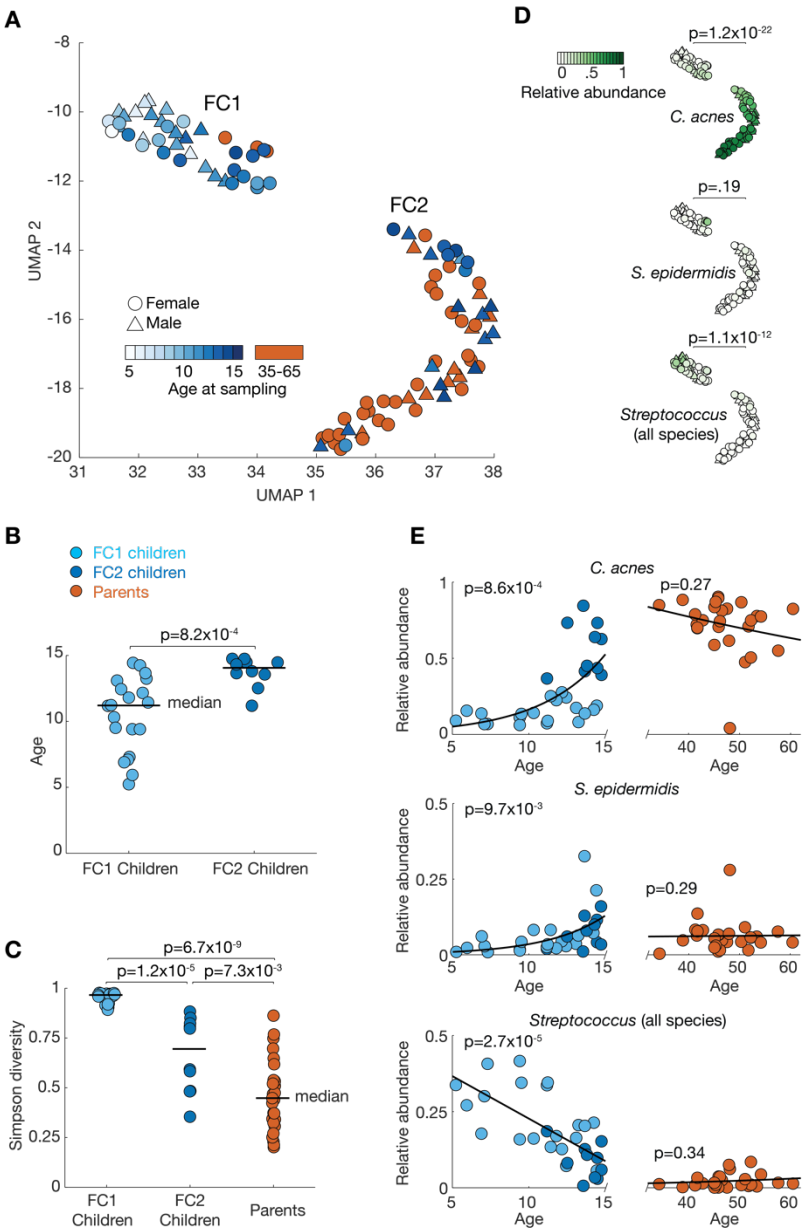

Fig. 2. Humans transition to the low-diversity adult-type FSM many years after birth. (A) UMAP projection of species-level communities (each dot represents one subject at one time point) shows a clear separation into 2 clusters which we call 'Facial Cutotypes' (FCs). **(B)** Child subjects classified as having the adult-type FSM (FC2 Children) are older than FC1 children, but lack a clear age threshold. P-value shows the result of a two-sided ranksum test. **(C)** Species-level diversity is lost when children transition from FC1 to FC2, and similarly within FC2 between children and parents. P-values show results of two-sided ranksum tests. **(D)** The relative abundance of *C. acnes* and *Streptococcus* species, but not *S. epidermidis*, are significantly higher in FC2 samples. P-values show results of two-sided Kolmogorov-Smirnov tests and are not corrected for multiple hypotheses. **(E)** The relative abundances of *C. acnes, S. epidermidis,* and *Streptococcus* are each significantly dynamic in children and do not change significantly in parents. Trendlines are first-degree exponentials for *C. acnes* and *S. epidermidis* in children, and linear elsewhere– P-values are for linear correlation (Pearson).

Consistent with the expectation of dramatic changes to the skin during development (*24*, *25*), younger children and adults have distinct species-level FSM communities. Clustering of relative species abundances from metagenomics using UMAP (*28*) and HDBSCAN (*29*) revealed two distinct groups, herein called Facial Cutotypes (FCs) (Fig. 2A, Methods). FC1 communities are found primarily on samples from younger children (median =11.4 years). Only one outlier parent had an FC1 community (11PA); this individual remained FC1 across each of their three timepoints (Fig. S2A) and harbored an unusually high abundance of *S. epidermidis*. Samples from all other parents, and most older children, have FC2 communities (median age among samples from children =14.1 years). To categorize subjects by FC, we classified children as FC2 ('FC2 Children' in figures) if at least 1 of their samples is classified as FC2 (Fig. S2B). Children with only FC1 samples were classified as FC1 ('FC1 Children' in figures) and all parents, including the outlier parent, were analyzed ('Parents' in figures). While FC2 children are significantly older than FC1 children ($P<8.2x10^{-4}$, Fig. 2B) there is not a clear boundary. It is possible that sampling more individuals might have revealed a continuum rather than a dichotomy; regardless, the clear separation between FC1 and FC2 found here enables us to clearly delineate which children have not yet begun developing an adult-like FSM.

The transition to the adult-type FC2 community is distinguished by a dramatic loss of species-level alpha diversity (Fig. 2C), driven by the dynamics of only a few key species (Fig. S2C-D). Consistent with previous results (*26*) the relative abundance of *Cutibacterium acnes* is significantly higher in the adult-type FC2 community (mean 15.9% vs 72.4%, $P=1.2x10^{-22}$; Fig. 2D) and increases with age in children ($P=8.6x10^{-4}$; Fig. 2E). The more diverse FC1 communities are dominated by various *Streptococcus* species, which decrease with age in children. *Staphylococcus epidermidis* is found at similar relative abundance in FC1 and FC2 communities (7.6% vs 5.9%, P=.19). However, the absolute abundance of *S. epidermidis* was significantly higher in parents than either FC1 or FC2 children (Fig. S2E, P<.05), highlighting shifts in microbial density during the transition to adulthood.

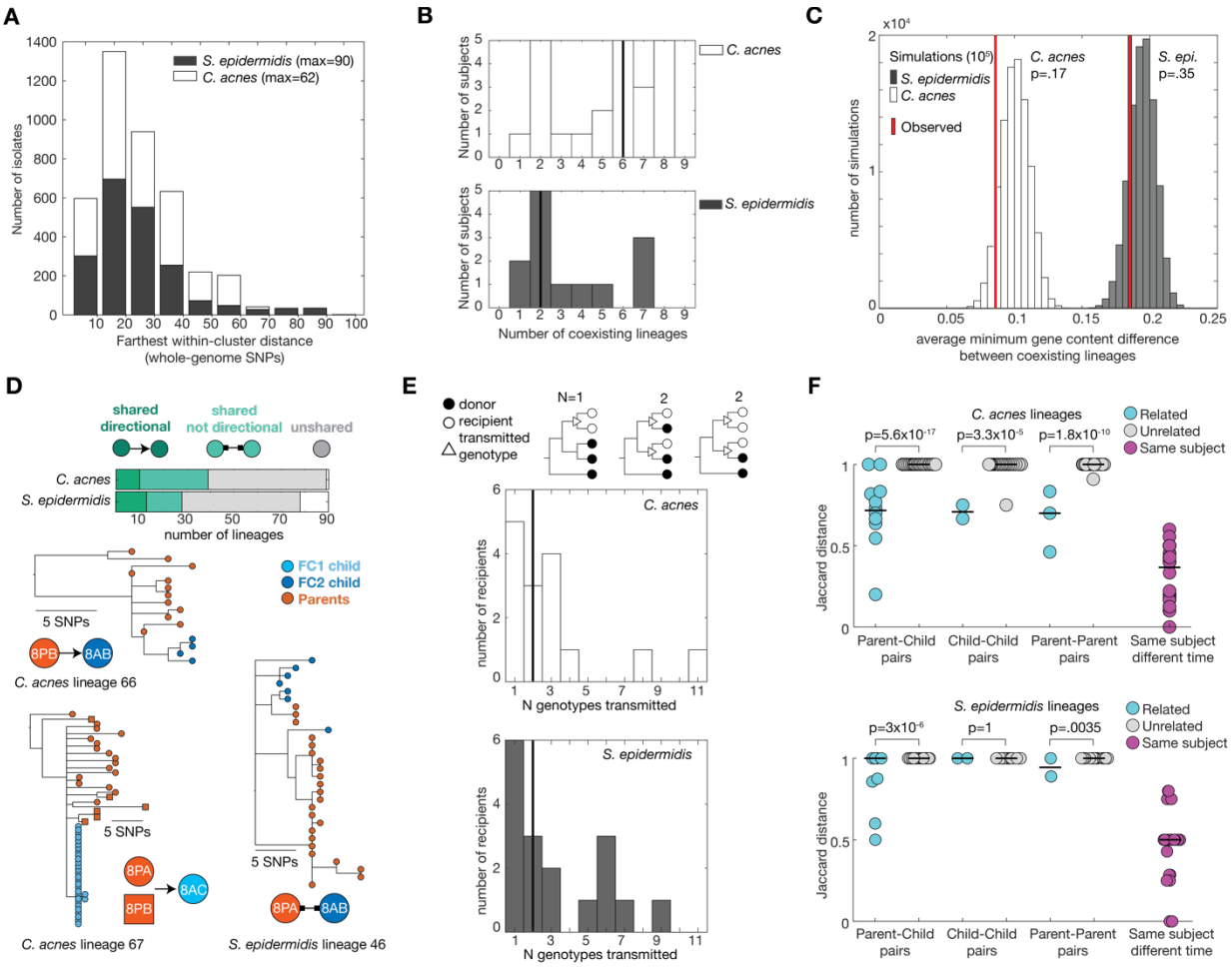

Fig. 3. Multiple cells from lineages establish populations during transmission. (A) The farthest observed within-lineage distance for either species is only 90 SNPs across the entire genome, indicating a most recent common ancestor within the lifespan of an individual person. **(B)** For subjects with >25 isolates and >70% lineage-level assignment for a given species, we almost always observe multiple co-existing lineages in both species in metagenomics and/or isolate data (Fig. S3). Vertical bars represent the median. **(C)** The most similar genomes between co-existing lineages (for either species) are not more different than chance, which would be expected under metabolic niche partitioning. **(D)** (upper) Many lineages of both species are shared between two or more family members, but only a minority (dark green) have directionality which can be confidently inferred from one or more sources (lower left); the rest either have insufficient numbers of genotypes to be sure or have more complex topologies (lower right). **(E)** The minimum number of transmitted genotypes required to explain the diversity of lineages is often greater than 1 (median 2, vertical bars), suggesting the migration of multiple cells. Lineages shared by three or more individuals are counted once for each recipient. **(F)** Jaccard distances of lineage composition between pairs of subjects shows that lineages are shared extensively within and between parents and their children, rarely shared among unrelated children at the same school, and much more rarely shared between unrelated parents. These results indicate that exposure is predictive of transmission. Notably, parents of the same family share only a fraction of their lineages despite exposure. Horizontal bars represent the median. P-values show results of two-sided ranksum tests and are not corrected for multiple hypotheses.

Neutral lineage-level coexistence on individuals

We next characterized intraspecies diversity and dynamics at the lineage level. As determining lineage membership requires whole-genome resolution, we employed a high-throughput, culture-based, single-organism approach. We focused on 8 family units for which we had samples from at least 3 subjects, including at least one parent and one child. From each timepoint and individual, we obtained whole genomes from up to 133 *C. acnes* and 186 *S. epidermidis* single-colony isolates, derived using mildly selective growth conditions (Methods). We clustered 4,055 isolates into 167 lineages, each with at least 3 isolates (2,030 and 2,025 isolates, 89 and 78 lineages for *C. acnes* and *S. epidermidis,* respectively) using an approach that guarantees isolates within a lineage are more closely related to one another than any are to isolates from other lineages (Methods). While we observed evidence of gene content among isolates from the same lineage, we focus here on single-nucleotide polymorphisms because of their clock-like accumulation over time. Isolates from the same lineage are separated by ≤ 62 and ≤ 90 SNPs across the whole genome for *C. acnes* and *S. epidermidis*, respectively (Fig. 3A), reflecting a common ancestor within the lifespan of an individual person (*7*).

To better detect lowly abundant lineages missed by culturing but shared between subjects and get finer measurements of lineage abundances, we also assessed the composition of samples using metagenomic data (Methods). This also enabled us to query intraspecies diversity, though at a coarser resolution, from subjects from whom no isolates were obtained. Metagenomic-inferred lineages and their abundances correspond well with isolate-inferred abundances in subjects with sufficient data for both (Fig. 3B, Fig. S3). Across all subjects, we find that multi-lineage colonization is common, with up to 9 lineages of either species coexisting on a person simultaneously (medians: *C. acnes*: 6; *S. epidermidis*: 2; threshold of 1% relative abundance within a species to be called as present). The large number of coexisting lineages for both species suggests at least temporary openness of the FSM to new colonizations.

To test if co-existing lineages of the same species fill distinct functional niches, we examined the differences in gene content and phylogenetic differences between co-existing lineages. If lineages with similar functionality are less likely to coexist on a person, we would expect to see overdispersion– a greater dissimilarity between pairs of lineages on a person than - that expected from a random sampling of lineage. For both species, we find no signal for overdispersion for either gene content or phylogenetic distance (Fig. 3C and Fig. S4), indicating that closely-related lineages are able to co-exist. These results are consistent with a neutral model of lineage coexistence (*30*) and suggest that functional niche partitioning is not the major driver of lineage co-existence for *C. acnes* and *S. epidermidis* on facial skin.

Co-transmission of multiple genotypes in shared lineages

Lineages are extensively shared within members of the same family. Across all lineages, 44% and 46% are shared between 2 or more people for *C. acnes* and *S. epidermidis*, respectively, as detected from metagenomic data (Methods). Lineages are overwhelmingly shared within families; only 2 lineages are shared between non-family members (both *C. acnes*). In both cases

151 of between-family sharing, the lineages found at low (<5%) abundance on one of the two
152 individuals implicated, potentially indicating experimental or computational error. Many
153 lineages are shared between parents (12/35 of *C. acnes* and 1/24 of *S. epidermidis* lineages on
154 parents in families with 2 parents sampled), indicating that lineages can be shared in an
155 exposure-dependent manner as opposed to being strictly vertical.
156 To understand the directionality and number of cells involved in transmission, we turned
157 to isolate-based lineage phylogenies (Fig. 3D). Directionality, as evidenced by an individual's
158 isolates having a younger most recent common ancestor (MRCA) than that of another
159 individual's isolates, was inferrable for only a minority of lineages with sufficient isolates (Fig.
160 S5A, Methods). A parent was the donor in the majority of inferred directional transmissions (Fig.
161 S5B-C, 9/10 for *C. acnes* and 9/13 for *S. epidermidis*). In contrast, the majority of shared
162 lineages had complex topologies, emerging from the involvement of more than two individuals
163 or transmission of multiple genotypes (Fig 3D, Lineage 67). We conservatively estimate a
164 median of at least two transmitted genotypes (cells) per recipient per lineage for both species (up
165 to 11 transmitted genotypes, Fig 3E, Methods). The number of transmitted genotypes is
166 dependent on the number of recovered colonies for *C. acnes* (Fig S6), further supporting the
167 conservative nature of these estimates. The predominance of complex topologies in *C. acnes* is
168 consistent with our previous work indicating that each sebaceous follicle is independently
169 colonized, creating an archipelago-like community structure which maintains coexisting
170 genotypes by preventing direct competition (*13*). It is unclear if there is an anatomical basis for
171 the coexistence of multiple transmitted *S. epidermidis* genotypes, given the unknown fine-scale
172 spatial structure of *S. epidermidis* populations on human skin (*27*, *31*). *C. acnes* and *S.*
173 *epidermidis* have significantly different within-lineage topologies, with *C. acnes* having a more
174 comb-like structure than *S. epidermidis*, which could indicate that populations of *S. epidermidis*
175 are more efficiently mixed (Fig. S5E, Methods). Regardless, the absence of single-celled
176 bottlenecks for either species suggests the possibility of a spatially structured environment that
177 facilitates stable transmission and co-existence of closely-related genotypes.
178 The apparent frequency of transmission between family members indicates that lineage-
179 level communities could, in theory, converge within families over time. However, the FCMs of
180 family members do not homogenize, and related individuals retain lineages that are undetectable
181 on other family members (Fig. S7). Furthermore, cohabitating parental pairs have significant
182 dissimilarity at the lineage level for both species (Fig. 3G). In the next sections, we dissect the
183 species-specific assembly dynamics that establish these person-specific communities.
184

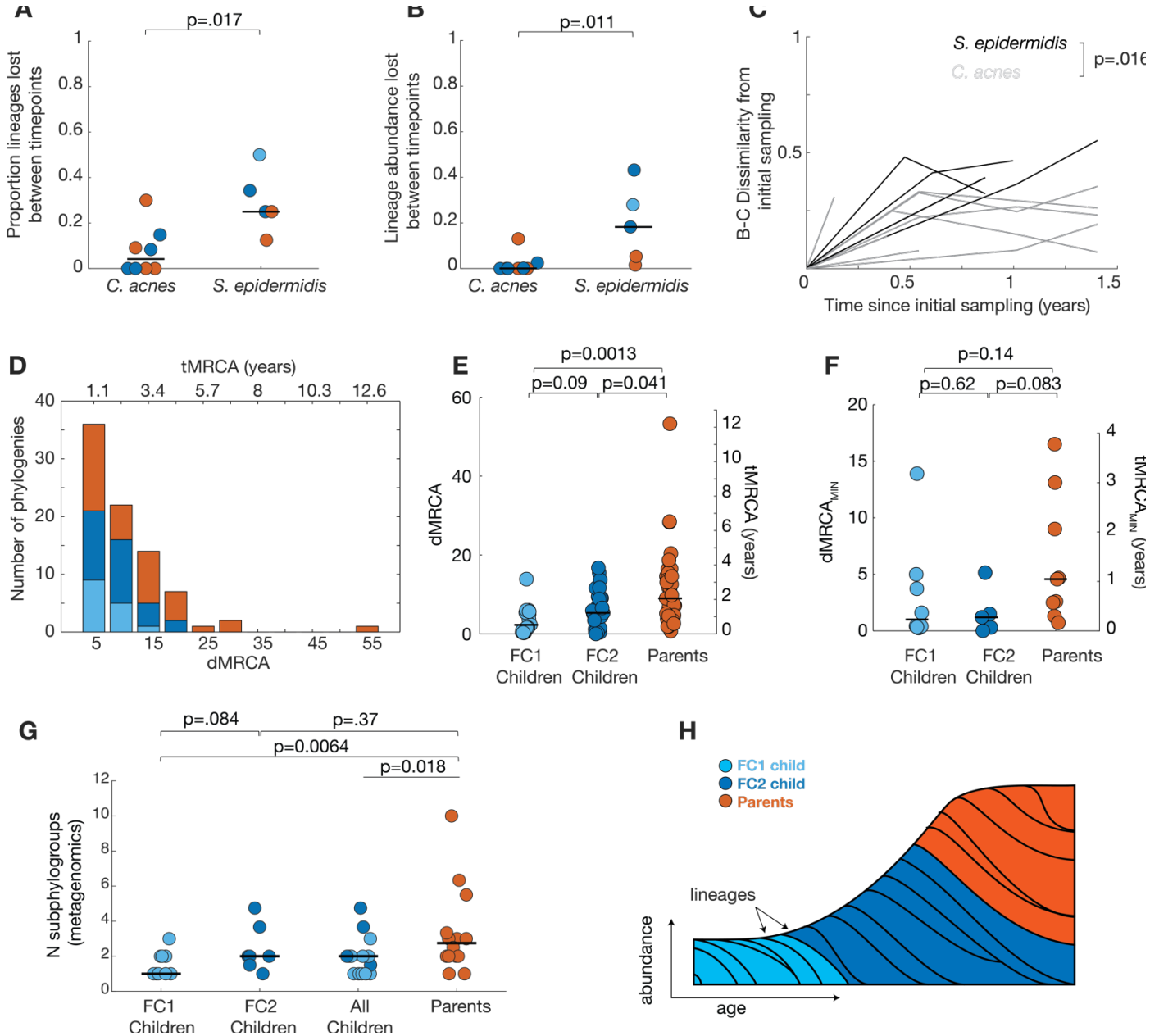

Figure 4: Fast cycles of *S. epidermidis* lineage turnover are continuous and slow down over time. (A) For subjects with sufficient data across multiple timepoints, lineage-level communities from metagenomics show that the proportion of lineages and **(B)** total abundance of lineages lost between timepoints is significantly higher in *S. epidermidis* than *C. acnes.* P-values show results of two-sided ranksum tests. **(C)** Over our observation window of ~1.5 years, individuals' *S. epidermidis* lineage-level communities turn over faster than *C. acnes.* P-value shows result of a two-sided Kolmogorov-Smirnov test of the maximum dissimilarity values for each subject. **(D)** The age (tMRCA) of *S. epidermidis* lineages indicate that all colonizations are recent– no lineage has more than ~12.6 years since its most recent common ancestor and an average age of ~ 1.6 years. **(E)** While relatively recent, lineages on parents are older than both FC1 and FC2 children. These age estimates are not an artifact of sampling depth (Fig. S11). **(F)** The most recently colonized lineage per-subject ($tMRCA_{MIN}$), is similar across groups, indicating similar rates of new colonizations. **(G)** Intraspecies richness, as indicated by the number of coexisting sub-phylotypes of *S.*

epidermidis, is higher on parents than children, potentially due to an increased carrying capacity. P-values show results of two-sided ranksum tests. **(H)** Our observations suggest a model in which lineages of *S. epidermidis* are continuously gained and lost, but tend to persist for longer on adults, who have higher carrying capacity of both *S. epidermidis* cells and *S. epidermidis* diversity. The area between each set of lines represents the abundance of a lineage, and each lineage is colored by the time window in which it was acquired.

S. epidermidis lineages are continuously acquired and increasingly stable over time

Despite an overall trend of long-term persistence of *S. epidermidis* on individuals, we observe many turnover events of *S. epidermidis* lineages within the short time windows assessed here (max interval length = 1.5 years). Compared to *C. acnes*, we find that a significantly higher proportion of lineages and total abundance of lineage-level communities are lost across timepoints in *S. epidermidis.* (Fig 4A-B). This dynamism results in *S. epidermidis* communities more rapidly turning over than those in *C. acnes* (Fig. 4C). We observed similar fluctuations in both parents and children at the phylotype level (Fig. S8 and S9), suggesting that *S. epidermidis* are continually acquired and lost throughout life stages.

To understand *S. epidermidis* lineage stability beyond the limit of our sampling windows, we turned to SNP-level lineage phylogenetics and inferred the age of the MRCA for each lineage on each person. We first measured the rate at which *S. epidermidis* accumulates mutations using data from all subjects with longitudinal data grouped together, obtaining a value of 4.5 mutations/genome/year (Methods, Fig. S10). This value agrees well with those estimated previously from the literature (Range 4-10) (*14*, *32–34*). We then used this rate to estimate lineage tMRCAs, the times since the most recent common ancestor of each lineage on each subject.

Supporting a model of *S. epidermidis* lineage instability, all values of tMRCA were significantly smaller than their subjects' ages, with a maximum value of tMRCA of 12.2 years (Fig. 4D). The median tMRCA of 1.2 years indicates frequent *S. epidermidis* turnover and is not due to lineages with few isolates (Fig. S11). While low tMRCAs could alternatively reflect a recent selective sweep or neutral bottleneck that purged diversity (*14*), we do not find evidence of frequent adaptive evolution, as indicated by parallel evolution or ratios of nonsynonymous to synonymous mutations (Fig. S12). Together, these results indicate that *S. epidermidis* lineages generally persist on individuals for a duration much shorter than their human host's lifespan.

Notably, tMRCAs on parents were significantly older than those on either FC1 or FC2 children (Fig. 4E, median of 2.1 for parents vs. 0.52 and 1.2 years for FC1 and FC2 children, respectively). We note that this is not a trivial result from colonization at birth, as all children in our study were over 5 years of age. Thus, despite the capacity for constant turnover, *S. epidermidis* lineage colonization is generally more stable on parents. In contrast, we found no significant difference in the min tMRCA on a given subject ($tMRCA_{min}$) across groups (Fig. 4F). As $tMRCA_{min}$ reflects the age of the most recently acquired simple (non multi-cell) lineage acquisition per subject, these results support a model in which *S. epidermidis* lineages arrive on subjects of all ages at similar rates, with enhanced stability of individual lineages on parents.

These two results from tMRCA analyses– enhanced stability of lineages on parents and no difference in the acquisition of new lineages– together suggest that more lineages could coexist on adults than children at any given time. Indeed, we find that adults have more lineages on average than children, though this is not statistically significant (median of 2 vs 3; P=.06; Fig S13). When including subjects without cultured representatives via analysis at the sub-phylotype level (taxonomic grouping between lineage and phylotype-level, Fig. S14), we find a significant enrichment in intraspecies richness on parents (Fig. 4G). Together, our results indicate that *S. epidermidis* lineages are acquired and lost from at least childhood through adulthood, with longer persistence on adults (Fig. 4H).

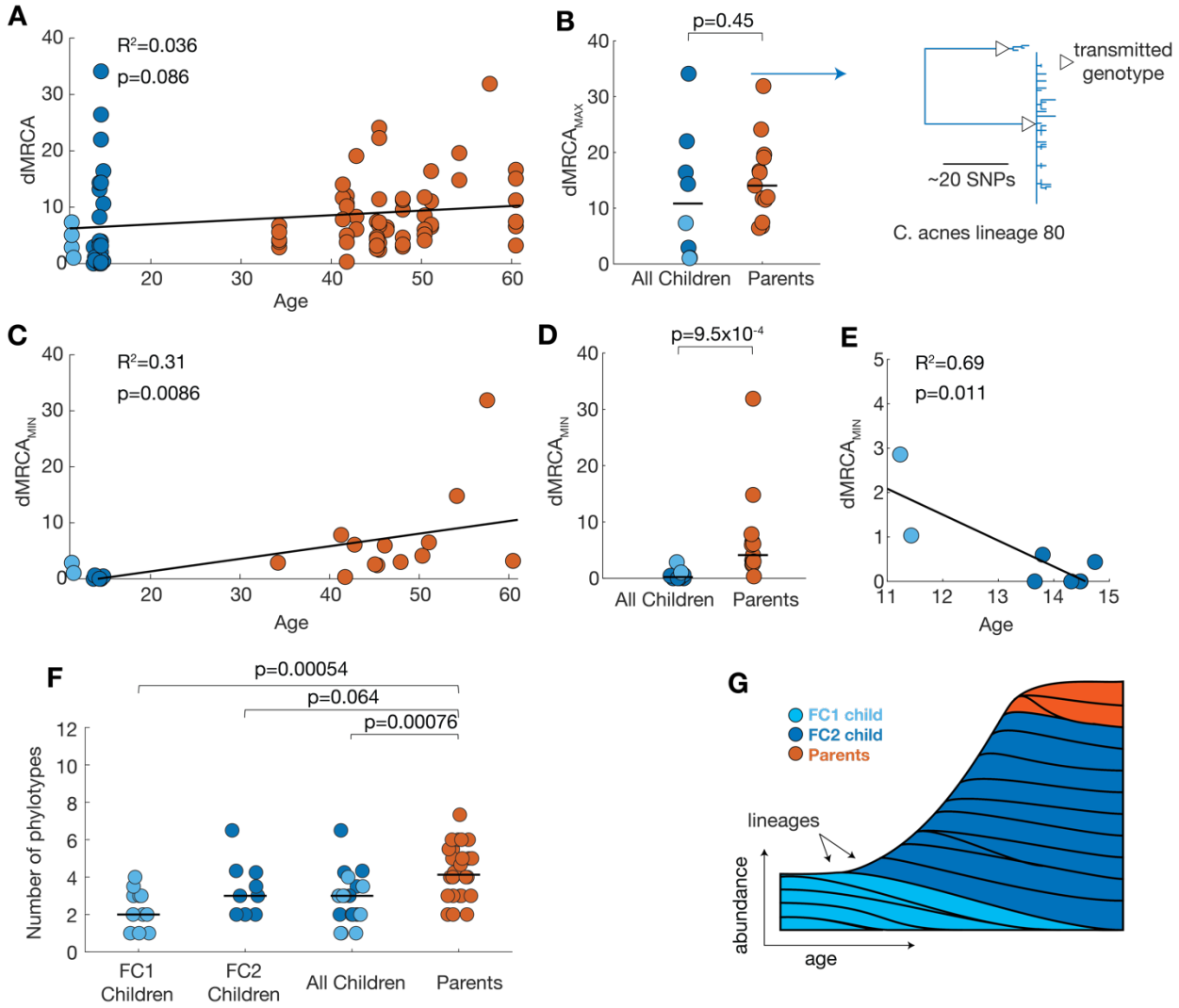

Fig. 5. Rapid acquisition of *C. acnes* lineages during transition to FC2 suggests that population expansion temporarily alleviates ecological barriers to colonization. (A) The dMRCA of *C. acnes* lineages is not significantly correlated with age. **(B)** Moreover, the oldest lineage per subject ($dMRCA_{MAX}$) is not different between parents and children. The most extreme value from any child (inset phylogeny) appears to show two distinct co-existing sub-lineages, which could result from a recent transmission of two cells (triangles) despite the apparently long dMRCA. P-value shows result of two-sided ranksum test. **(C)** The age of diversity in the youngest lineage (most recently colonized) per subject, $dMRCA_{MIN}$, is significantly correlated with age, suggesting a decreasing rate of new colonizations in parents. **(D)** Parents' most recently colonized lineages are significantly older than those in all children, indicating less recent colonization. P-value shows result of two-sided ranksum test. **(E)** In contrast to **(C)** **,** the opposite trend is apparent in children, with older children more recently colonized than younger ones. **(F)** Adult *C. acnes* communities have a larger number of co-existing phylotypes. P-values show results of two-sided ranksum tests. **(G)** Illustrative model of *C. acnes* lineage dynamics throughout the lifespan. We propose that lineages are acquired most readily during the transition to an adult FSM and alongside a population expansion. The area between each set of lines represents the abundance of a lineage, and each lineage is colored by the time window in which it was acquired. All correlation coefficients and P-values are for lineage (Pearson) correlations.

228 ***C. acnes* is readily acquired during population expansion**

229 In contrast to *S. epidermidis*, *C. acnes* lineages were remarkably stable across individuals
230 during the time intervals studied (Fig. 4C, P=.01, Methods). Lineages with turnover had lower
231 confidence at detected timepoints than stable lineages (Fig. S15), suggesting these lineages may
232 appear gained or lost due to fluctuations near the limit of detection. *C. acnes* communities
233 remained more stable than *S. epidermidis* communities when examining all subjects (including
234 those without isolates) using the broader taxonomic level of phylotypes (Fig. S8 and S9). This
235 observation, despite the added statistical power from the larger number of phylotypes in *C. acnes*
236 (9 vs 4), furthers evidence for the robustness of *C. acnes* colonizations over time.

237 Despite this stability, *C. acnes'* lineage dMRCAs did not significantly correlate with host
238 age or FC type (Fig. 5A), and we were unable to identify a significant molecular clock signal to
239 infer tMRCAs. The absence of a correlation between dMRCA and age could theoretically
240 emerge from on-person selective sweeps, which purge accumulated diversity. However, we
241 observe no evidence of selection in *C. acnes* (Fig. S12), consistent with prior results from our
242 group (*13*). The consistency of dMRCAs across the lifespan may be due to recently co-
243 transmitted *C. acnes* genotypes causing lineages to appear as old on recipients as their sources
244 (Fig. 5B); in cases of multi-cell transmission, the MRCA of a recipient's lineage diversity could
245 have emerged on the source individual, potentially even before a subject was born. Accordingly,
246 we find deeply branching clades coexisting on young individuals (Fig. 5B, inset) and no
247 significant difference in $dMRCA_{max}$ between children and parents (Fig. 5B). Multi-clade
248 transmission may also play a role in preventing a significant molecular clock signal for *C. acnes*
249 here (Fig. S10) and in previous data (*13*) for converting values of dMRCA to tMRCA.

250 We therefore turned to $dMRCA_{min}$ to assess the age of the most recently acquired simple
251 lineage acquisition per subject. Across all subjects, *C. acnes* $dMRCA_{min}$ was significantly
252 correlated with age (Fig. 5C). Moreover, children had significantly lower values of $dMRCA_{min}$
253 than parents ($p=9.5x10^{-4}$, Fig. 5D), consistent with more recent acquisitions. When considering
254 only child subjects, $dMRCA_{min}$ was lower on FC2 subjects and negatively correlated with age
255 (p=.011; Fig. 5E). Together, these trends support a model in which *C. acnes* lineages are
256 acquired at a higher rate during the transition to FC2, concurrent with the increase in the size of
257 the *C. acnes* population.

258 This influx of new lineages during the transition to an FC2 community could result in
259 either replacement or addition to preexisting lineages. Consistent with a model of addition, and
260 with an increased carrying capacity, the number of phylotypes is higher on parents than on
261 children (Fig. 5F). While we were unable to obtain statistically significant assessments of *C.*
262 *acnes* within-species diversity between FC1 and FC2 children due to the low abundance of *C.*
263 *acnes* on FC1 children (Fig. S2E), we observe trends in agreement with a model of *C. acnes*
264 lineage addition during adolescence.

265 Altogether, these results demonstrate distinct on-person dynamics for *C. acnes* and *S.*
266 *epidermidis* lineages. The rate of *S. epidermidis* lineage turnover is higher throughout the
267 lifespan. While we cannot determine the basis of this difference, it is interesting to note that *S.*

epidermidis lineages have a larger and more dynamic accessory genome (Fig. S16) and may therefore undergo more between-lineage competition. Another notable difference is that only *C. acnes* lineages appear to be most rapidly around the transition to an adult-like facial skin microbiome. The more dramatic changes to *C. acnes* population size during this timespan (10^5-fold vs. 10^2-fold) (*26*) may be responsible for this difference.

Discussion

Resolving assembly dynamics has remained difficult for intraspecies communities with multiple lineages because of the technical limitations of metagenomics and short time course data. Here, we applied single-organism genomics and evolutionary reconstruction to detail on-person intraspecies diversity at whole-genome level, infer dynamics beyond the limits of our sampling window, and build models of community assembly for the two most dominant bacteria of the skin microbiome. We report that transmission between family members is frequent but incomplete– with many lineages unshared despite ample opportunity. Coexisting lineages display no strong evidence for niche-partitioning, implicating neutral forces like priority effect in determining the success of lineages and genotype (*30*). We show that while both *C. acnes* and *S. epidermidis* lineages are gained and lost throughout life, *C. acnes* lineages are more stable than *S. epidermidis* lineages on adults and are acquired at particularly high rates during the population expansion occurring in adolescence. The importance of this transitional period for *C. acnes* strain acquisition highlights the critical role of life stages beyond birth and infancy for strain acquisition.

For *C. acnes* the findings of incomplete lineage sharing and the primacy of adolescence acquisition can potentially be rationalized in terms of its biology. *C. acnes* growth occurs primarily within sebaceous follicles (pores), and the population within each pore is dominated by the descendents of a neutral single cell bottleneck event (*13*). Pores which briefly become accessible (due to physical disturbance, phage predation or immune clearing) are more likely to be seeded by lineages that are abundant on that person, rather than lowly-abundant migrating lineages (*30*). During adolescence, increased sebum production (*22*) supports a larger population on the surface (*26*) , which could increase the likelihood that a newly opened pore *C. acnes* is colonized by a migrating lineage– and therefore alleviate priority effects.

Noted differences between species underscore that the rules of colonization for different human commensals depend on the particulars of their niches. Unlike *C. acnes*, relatively little is known about the fine-scale spatial structure of *S. epidermidis* populations. Adolescence may be a weaker force for *S. epidermidis* colonization than for *C. acnes* because the change in absolute abundance during this time window is relatively marginal (10^2-fold vs. 10^5-fold, respectively) (*26*), but other factors may be at play. More work will be needed to understand forces that shape *S. epidermidis* person-specificity at the lineage level despite the potential for continual turnover.

Both of these species have been considered as potential probiotics (*9*, *35*), and we anticipate that the dynamics described here will be helpful in the design of microbiome therapies involving these species. Notably, recent studies have raised the possibility that some strains of *C.*

acnes may be protective against acne vulgaris (*9*, *36*). If some or all *C. acnes* are beneficial for preventing acne, providing colonization resistance to infection (*37*) , or treating other skin conditions, the openness of the skin microbiome to *C. acnes* colonization during adolescence suggests that probiotics applied during this time window may be more likely to result in long-lasting efficacy. Finally, it is tempting to contemplate whether recently acquired lineages interact with the host immune system differently, and we anticipate that future work will explore whether the lineage-level dynamics described here are consequential for *acne vulgaris* or other diseases.

Acknowledgments: We thank the participants of the study and Courtney Dickinson, Stephanie Friedhoff, Michael Hirsch, Sarah Zuckermann, and Joshua Schuler for assistance with study recruitment. We thank Tucker Lynn for experimental assistance and the BioMicroCenter at MIT for assistance with DNA sequencing. We thank Otto Cordero, Greg Fournier, and members of the Lieberman Lab for feedback on the manuscript.

This work was funded by a pilot grant from the MIT Center for Microbiome Informatics and Therapeutics, by a Smith Family Foundation Award for Excellence in Biomedical Research, and NIH grant 1DP2GM140922 (all to TDL).

Author contributions: Conceptualization: JSB, TDL; Methodology: JSB, EQ, CPM, ADT, AC, TDL; Investigation: JSB, EQ, CPM, ADT, AC, TDL; Funding acquisition: TDL; Writing – original draft: JSB, TDL; Writing – review & editing: JSB, EQ, CPM, ADT, AC, TDL.

Competing interests: Authors declare that they have no competing interests.

Data and materials availability: Sequencing data is available on the NCBI Sequence Read Archive under Bioproject #SUB14054091. All code needed to reproduce the analysis in this work can be found at https://github.com/jsbaker760/highres_dynamics.

Materials and Methods

<u>Sample collection</u>

We opportunistically sampled subjects at a single K-12 School at five events over the course of 2.5 years. Consent was obtained from each individual under MIT IRB 1802230066. For children, both parental consent and child assent were obtained. In total, this work includes samples from 57 individuals belonging to 23 nuclear families, in addition to two researchers who were collecting samples. Before sampling, subjects filled out a questionnaire with basic health and demographic information, including the time since their most recent antibiotic use. This questionnaire was used primarily to identify potential outliers.

345 Each time a subject was sampled, swabs were collected from four locations (Forehead,
346 Nose, Cheek, and Chin) on the left-hand side of the face. Each set of four sampling sites is
347 referred to as a "kit" in the methods. To collect skin-dwelling bacteria while reducing the
348 collection of incidental surface-dwelling bacteria, debris, and cosmetics, we first firmly wiped all
349 four skin sites with a cleansing wipe (Cetaphil, DPCI: 037-12-0886), ensuring that a different
350 side of the wipe was used for each site. We then used saline cotton wipes (Winner #504021),
351 using the same technique, to remove any residue from the cleansing wipe. After a 1-minute
352 drying period, samples were taken from the forehead, nose, cheek and chin, in that order. During
353 the drying period, examiners recorded whether the subject had facial hair and noted the presence
354 of acne lesions and categorized the subject's skin at the time of sampling as dry, oily, or
355 combination. For each sample, a swab (Puritan, REF#25-1506 1PF 100) was first dipped into
356 sterile PBS, and then applied with light pressure with upward and downward strokes while
357 moving swabs laterally and rotating the swab. Each swab was immediately placed into 1.5mL
358 cryotubes containing 1000mL of storage solution (20:80 glycerol:PBS with .05% w/v Cysteine
359 which had been vacuum filtered at 0.22um and reduced in an anaerobic chamber). Upon
360 completion of a sampling kit, all samples were vortexed for 60s at maximum speed and placed
361 on dry ice. Samples were transferred to long-term storage in a -80°C freezer the same day they
362 were collected.
363
364 Isolating colonies
365 We isolated colonies on both LB agar plates cultured in anaerobic conditions (to enrich
366 for *Cutibacterium* species) and TSB agar plates cultured in aerobic media (to enrich for
367 *Staphylococcus* species). For each sampling kit cultured, tubes were thawed and resuspended by
368 vortexing for 60s at max speed. A serial dilution was prepared using 100µL of this solution and
369 plated onto petri dishes with disposable spreaders. After growth for 3 days aerobically (TSB
370 plates) or up to 7 days anaerobically (LB plates), colonies originating from each tube were
371 selected from the most dilute plate or plates possible to limit competition. When few isolates
372 could be obtained for the initial round of culturing, particularly common in younger subjects
373 with a lower abundance of bacteria, additional serial dilutions were made using lower dilutions.
374 Each colony was picked into a 1000µL of liquid medium (either LB or TSB) and grown
375 for 3 days aerobically or up to 7 days anaerobically at 37°C. Colonies were picked either using a
376 disposable loop or a PIXL colony picking robot (Singer Instruments, PIX-001). When picking by
377 hand, we randomly selected colonies which were far enough apart to avoid cross-contamination.
378 When using the PIXL to pick colonies, we implemented size, circularity, and radius filters to
379 randomly pick colonies that were sufficiently far apart while limiting cross-contamination. These
380 cultures were used for both DNA extraction and storage. For each round of culturing, we used
381 two types of negative controls to ensure the absence of contaminating bacteria. First, an entire
382 unused sampling tube from the sampling event poured onto a single petri dish to control for
383 batch contamination. Second, a plate with an aliquot of the PBS used during serial dilutions, after
384 each sample had been plated, to control for reagent contamination during plating. The first type

385 of negative controls never showed growth indicative of bacterial contamination in the collection
386 tubes. When contaminants were identified in the second type of negative controls due to reagent
387 contamination, all plates were discarded and the experiment was repeated.
388
389 DNA preparation for isolate genomes
390 Cells were spun down at 4200g for 10 minutes and the supernatant was removed. Pellets
391 were resuspended in 100µL of PBS and lysozyme (Sigma #L6876) was added to reach a final
392 concentration of 1000U/mL. Plates were sealed and vortexed until the suspension was
393 homogenous (at least 2 minutes), then incubated at 37°C overnight. To each well, 2µL of
394 20ug/µL proteinase K (NEB, #P8107S) and 2µL of 0.8% SDS solution were added, and plates
395 were then incubated at 55°C for an additional 3 hours, after which the cell suspensions had no
396 apparent turbidity. DNA was purified from these cell lysates using the standard protocol in the
397 DNA extraction kits (Zymo Research #D3010), except that DNA was resuspended in sterile
398 water in the final step. DNA was quantified using SYBR Safe (Thermo, #S33102) and
399 normalized the concentration of each sample to between 1 and 5 ng/µL.To prepare libraries for
400 Illumina sequencing, the tagmentation-based plexWell DNA library prep kits (Seqwell,
401 #PW096) was used according to the manufacturer's protocol.
402
403 DNA preparation for metagenomes
404 Lysates were prepared as above. All reagents were sterilized, vacuum filtered, and
405 autoclaved prior to use. DNA was purified using Purelink gDNA extraction kits (Thermo,
406 #K182104A) as they improved yield in these low-biomass samples. The only modification to the
407 Purelink protocol was elution of DNA into 10µL of sterile water three consecutive times,
408 incubating the plate for 2 minutes at 37°C each time. DNA libraries were prepared using
409 HackFlex (*38*) with the following modifications: (1) the tagmentation was scaled to start with
410 10µL of DNA; (2) the tagmentation-stop step was skipped; (3) we used standard Illumina
411 barcodes for plates MG1-3 and UDI primers (20091660) for plates MG4-7 (Supplemental table
412 3); (4) we used KAPA HiFi master mix (Roche, 07958935001) and 19 cycles of PCR.
413
414 DNA Sequencing and quality filtering
415 All samples were sequenced using the Illumina platform, on a variety of instruments and
416 either 75bp or 150bp reads. Some libraries were re-sequenced to achieve higher depth. Isolates
417 that met all below filters had a median depth of ($8.6x10^5$ reads). Metagenomic samples that met
418 all below filters had a median depth of $3.1x10^6$ reads. Details for each sample are found in
419 (**tables S2** and S**3**). Raw reads were filtered and trimmed as previously described.
420
421 Species and genus- level microbiome composition from metagenomics
422 All reads from metagenomics samples were de-duplicated with the function dedupe.sh (-
423 Xmx100g) from BBMAP (v39) (*39*). Kraken2 (*40*) was used (v2.0.7-beta, standard db built
424 01/18/2019) to classify reads and Bracken (v2.5) (*41*) was used to estimate abundances at the

species level. Immediately after classification, reads labeled as *Homo sapiens* were discarded. Reads labeled as *Aeromonas sp.* were also removed because they were suspected to emerge from contamination due to their high abundance in negative control samples. These samples were additionally filtered using culture-based lineage information before UMAP clustering (see Filtering and selecting metagenomics samples**).**

Metagenomic reads from samples in the same kit were combined since we found minimal compositional differences (Fig. S1). Bracken abundances were averaged and re-normalized, resulting in 101 kit samples.

To determine the Facial Cutotype (FC) category of each subject, kit samples were additionally filtered before clustering as follows: Abundances of taxa beginning with "Human" or "*Toxoplasma*" and all taxa which were found at <.5% across all samples were removed from every sample and abundances were re-normalized. The pairwise distances between samples were calculated using maximum-coordinate (Chebychev) distance. Two clusters were identified using this distance matrix and UMAP with the following parameters: min_dist=.1, spectral initiation, n_epochs=10000, n_components=2, cluster_detail=adaptive, method=MEX. Cluster FC2 contains samples from older children and all parents except 11PA (Fig 2A). Cluster FC1 contained samples from younger children and parent 11PA.

Child subjects were categorized as an "FC2 Child" if any of their time point samples were clustered into FC2. Subjects were only categorized as an "FC1 child" if all samples from that subject were FC1. Subject 11PA was included as a parent for downstream analyses despite their aberrant composition. To obtain per-subject average abundances, Bracken abundances from kits were averaged and re-normalized as above.

<u>Calculating pairwise distances for clustering isolates</u>

We first filtered isolates that might have significant levels of contamination from other species. We reasoned that high-quality samples would contain the core genes for that species, but not core genes for different species. Genomes were assembled for each isolate using Spades (v.3.15.3, --isolate -k 21,33,55,77,99,127 --phred-offset 33) (*42*) and then annotated with prokka (v 1.14.6, --compliant --force --mincontiglen 500) (*43*). In total, 4621 isolate genome assemblies were able to be annotated. CDHIT (v 4.8.1, -p .9) (*44*) was then used to cluster the homologous protein-coding genes of these 4621 assemblies in addition to 79 publicly available reference genomes from NCBI (28 *C. acnes*, 35 *S. epidermidis*, 10 *S. capitis*, and 6. *S. saccharolyticus*). Core genes for each of these reference species were defined as CDHIT gene clusters which were present in 100% of reference genomes for that species, but not found in other reference genomes. Samples with >87% of *C. acnes* core genes and <10% of the core genomes of *S. epidermidis* were included as *C. acnes*. For *S. epidermidis*, samples with >90% of the *S. epidermidis* core gene clusters and <1% of the core genes of *S. capitis* and *C. acnes* and at least 2x mean coverage from spades were included. This results in 2300 *C. acnes* and 2172 *S. epidermidis* isolates for further analysis

464 Pairwise distances between samples were calculated by comparing their SNPs relative to
465 a reference genome. First, each *C. acnes* isolate was aligned to the Pacnes_C1 reference genome,
466 and each *S. epidermidis* sample was aligned to SepidermidisATCC12228 using bowtie2 (v2.2.6 -
467 -very-sensitive --threads 8 -X 2000 --no-mixed --dovetail -x). Samtools mpileup (v.1.5; -q30 -x -
468 s -O -d3000) was used to identify candidate SNPs for each representative isolate. For every
469 potential position in each sample, a nucleotide call was assigned to that position based on the
470 major allele nucleotide. To filter out low-quality polymorphisms and positions across lineages,
471 we first marked allele calls as ambiguous ("N") in a sample if the major allele frequency was
472 below 0.79, the FQ score produced by samtools was above -45, coverage was below 3. We then
473 masked positions across samples if the median coverage across isolates was less than 3, or 15%
474 or fewer of isolates had a non-ambiguous allele at that position. 11 *C. acnes* isolates were
475 removed because >50% of positions were N. Distance was calculated as the sum of positions
476 where both samples had different nucleotides and neither were N.
477
478 <u>Clustering isolates into lineages</u>
479 To cluster genomes which passed the above filters into closely-related clades consistent
480 with on-person evolution or recent transmission, we implemented a two-step process of running
481 clustering algorithms across a wide range of parameters and choosing with the most isolates in
482 self-consistent clusters. Clusters are considered consistent if and only if all isolates in each
483 cluster are more closely related to each other than to any isolate outside of their clusters. The
484 consistency criteria was inspired by expectations in our data set; namely that samples emerge
485 from unrelated families, such that large clusters of related strains with intermediate distances are
486 not expected. A parameter scanning process was required given the absence of solved methods
487 for clustering generally (*45*) and in this use case. Initial parameter ranges were chosen based on
488 the fact that the typical molecular clock range bacteria *in vivo* of about 1 mutation/core-
489 genome/year (*7*) and that we expect isolates from a family to have a single-celled ancestor within
490 the past 50-100 years.
491
492 *Initialization*
493 Sequencing reads for some isolates showed evidence of low-level contamination by
494 isolates from other strains. While low-level contamination only modestly impacts mutation
495 calling when performing within-lineage analysis (see: **Within lineage SNPs**), strict call-masking
496 when using a single reference genome across divergent isolates can artificially decrease inter-
497 isolate distances. To overcome this challenge, we used a network property called the clustering
498 coefficient (*46*), which represents how many of one's neighbors are neighbors with one another.
499 Only isolates with a clustering coefficient above .92 at a 91 SNP cutoff (*C. canes*) and a
500 clustering coefficient equal to 1 at a 90 SNP cutoff for *S. epidermidis* were included.
501
502 *Clustering*

503 Clustering was performed with a two step process: initial cluster generation and greedy
504 addition. Initial cluster generation was performed over a range of parameters for both ANCHOR,
505 a custom algorithm and the established method DBSCAN (*47*). The custom algorithm ANCHOR
506 was designed to produce self-consistent clusters while dealing with noisy distance metrics
507 emerging from low level contamination. ANCHOR is initialized by defining anchor points, each
508 of which represents a cluster to which isolates will later be added if they are within a given
509 distance (C1). Anchor points are distributed such that they are at least 2C+1 away from each
510 other as follows: The first anchor point is chosen. Then, and after each new anchor point is
511 added, the isolate furthest from all anchors is considered as a candidate anchor. Each candidate
512 anchor is added to the list of anchors if and only if it is at least 2C+1 away from every already
513 chosen anchor. When no further isolates can be added while meeting this condition, the
514 initialization step completes. We then added the isolates which could be unambiguously assigned
515 to these clusters (< C from only one cluster).
516
517 *Greedy addition*
518 The effects of the aforementioned noise, and as well as potential processes of accelerated
519 mutation like hypermutation, mean that isolates may be < C from more than one anchor even if
520 they are of high quality. We therefore implemented a greedy addition process to identify and
521 include these isolates. Greedy addition was attempted for all instances of clustering (both
522 algorithms, both species, across many parameters), which resulted in self consistent clusters, a
523 high number of clustered isolates (at least 1800 for *C. acnes* and 1600 for *S. epidermidis*) and a
524 number of lineages consistent with our early clustering attempts (96-150 for *C. acnes* and 100-
525 200 for *S. epidermidis*). We note that lineage number cutoffs are higher than the reported number
526 of lineages as we allowed singleton clusters at this step. at least
527 We next added unclustered isolates to each of these instances as follows: First, all
528 unclustered isolates are sorted in ascending order by their distance to the nearest cluster. Then,
529 for each of these, we added it to its nearest cluster and checked that the clusters are still self
530 consistent and this isolate is not >75 SNPs away from any isolate in its cluster. This process
531 terminates when no isolate can be added without violating either of these properties.
532 We then chose the self-consistent clustering instance which contained the highest number
533 of clustered isolates. Clusters with fewer than three isolates were excluded from all downstream
534 analyses as they could not confidently be distinguished from contamination, though clusters with
535 two isolates were used to build the PHLAME classifier. The full range of parameters testeed
536 used can be found in the github repository associated with this paper. For *S. epidermidis*,
537 DBSCAN (eps=90, minpts=3) gave the optimal result and ANCHOR (C1=91) gave the optimal
538 result for *C. acnes*. For *C. acnes*, 2164 of 2289 starting genomes were clustered in 98 lineages
539 with >2 isolates. For *S. epidermidis*, 2025 of 2172 starting genomes were clustered in 78 lineages
540 with >2 isolates. To confirm that we had not overclustered isolates and artificially broken up
541 related isolates into distinct lineages, we manually confirmed that nearby clusters were not
542 systematically from the same family (Fig. S17).

543
544 Constructing inter-lineage phylogenies
545 A single representative isolate from each lineage was chosen by picking the isolate in
546 each lineage with the highest average alignment coverage across the Pacnes_C1 and
547 SepidermidisATCC12228 reference genomes, respectively. To ensure a wide collection of
548 phylogenetic diversity for *C. acnes*, we also included a representative isolate from each of the 53
549 *C. acnes* lineages defined in (*13*), as well as 197 publicly available *C. acnes* isolates from NCBI.
550 Because the raw reads from some publicly available isolates were not available, we first
551 simulated reads from these genomes using wgsim (v.0.3.1-r13; -e 0.0 -d 500 -N 500000 -1 150 -2
552 150 -r 0.0 -R 0.0 -X 0.0) before aligning all isolates to their respective reference genome
553 (bowtie2 v.2.2.6; -X 2000 --no-mixed --dovetail). After alignment, samtools mpileup (v.1.5; -q30
554 -x -s -O -d3000) was used to identify candidate SNPs for each representative isolate. To filter out
555 low-quality polymorphisms and positions across lineages, we first marked allele calls as
556 ambiguous ("N") in a sample if the major allele frequency was below 0.85, the FQ score
557 produced by samtools was above -30, the coverage across either the forward or reverse reads was
558 below 2, or if greater than 50% of reads at a position supported an indel. We then masked
559 positions across samples if the median coverage across isolates was less than 3, or if more than
560 5% of isolates had an ambiguous allele at that position. We finally removed positions for which
561 the correlation of non-reference alleles within 300bp of each other across isolates was above
562 0.75, which suggests that these alleles were acquired through recombination or other horizontal
563 inheritance events. Positions which passed these filters and still contained at least one
564 polymorphism across isolates were used to generate a maximum-likelihood phylogenetic tree
565 under a GTR model of nucleotide substitution using RaXML (v. 8.2.12; -p 060782 -m
566 GTRCAT).
567
568 Detecting lineages and phylotypes from metagenomics
569 To identify lineages from metagenomics data, we first built a database for each species
570 by aligning all 2164 *C. acnes* and 2025 *S. epidermidis* isolate genomes which passed quality
571 filtering and clustering to the Pacnes_C1 and SepidermidisATCC12228 reference genomes,
572 respectively, using bowtie2 (v.2.2.6; -X 2000 --no-mixed --dovetail). Duplicate reads were
573 filtered out using samtools markdup (v.1.15; -d 100 --mode s). Samtools mpileup (v.1.5; -q30 -x
574 -s -O -d3000) was used to identify candidate polymorphisms. PHLAME (v.0.1) (*48*) was used
575 subsequently to build a database of lineage-specific alleles (common to all isolates of a lineage,
576 but not found in any other lineage) within core genome regions (found in >90% of isolates).
577 Alleles were additionally included if <10% of the isolates in a lineage were missing an
578 alignment position entirely, but the remaining isolates had a unanimous allele at that position.
579 For each sample, the number of reads covering each informative position (containing a
580 lineage-specific SNP) both with and without the allele were extracted and used to determine the
581 presence and frequency of lineages in each metagenomic sample. PHLAME avoids false positive
582 detections of lineages that occur when only a subset of any set of lineage-specific alleles are

covered by reads, which may instead represent the presence of an unknown but closely-related lineage in a sample. Briefly, PHLAME models the number of reads covering a set of lineage-specific SNPs as coming from a zero-inflated negative binomial (ZINB) distribution with rate parameter λ, overdispersion parameter α, and zero-inflation parameter π. The π parameter represents the proportion of lineage-specific SNPs that are systematically missing from a sample. Gibbs sampling, implemented using JAGS v.4.3.0, was used to obtain posterior distributions over these 3 parameters by sampling their full conditional distributions according to the following hierarchical formulation of the ZINB distribution, where x_i represents the set of allele specific counts:

Likelihoods:

$$x_i \sim Poisson(\lambda_i z_i)$$

$$z_i \sim Bernoulli(1 - \pi)$$

$$\lambda_i \sim Gamma(\alpha, \beta)$$

Priors:

$$\beta = \frac{p}{1-p}$$

$$\alpha \sim Gamma(a, b)$$

$$p \sim Beta(1,1); \pi \sim Beta(1,1)$$

Parameters for a prior for the overdispersion parameter α were obtained empirically for each sample using the total number of reads at SNP-containing positions y_i, where ${s_{y_i}}^2$ is the sample variance of y_i.

$$a \sim Gamma(a = \frac{y_i}{{s_{y_i}}^2} - y_i, b = 1)$$

This modeling was performed if the total number of reads supporting lineage-specific alleles in the sample was >10, otherwise the frequency of that lineage was set to 0. Each chain was run for 100,000 iterations with the first 10% discarded as burn-in. In order for a lineage to be detected, 50% of the posterior density over π was required to be below 0.35, and the lower bound of the 95% highest posterior density (HPD) interval of π was required to be <0.1. To increase sensitivity (and therefore limit false observations of gains and losses), if we additionally counted lineage detections where >50% of the posterior density of π was below <0.55 if that lineage was detected using the stricter parameters on the same subject at another timepoint.

The relative abundance of each lineage was calculated as the maximum likelihood estimate for λ for the allele-specific reads divided by the maximum likelihood estimate for λ for all reads at the same positions. Because each lineage in a sample is modeled independently without reference to others, the sum frequency of lineages in a sample can sum to <1 or rarely,

621 >1. If the sum frequency of calls was >1, we re-normalized frequencies down to 1 (the max sum
622 lineage frequency before this step was 1.05).
623 We also grouped lineages into major phylogenetic clades, deemed phylotypes, to search
624 for patterns in intraspecies diversity in subjects without isolates (which therefore are not
625 represented in the lineage database) and because some metagenomic samples were not sequenced
626 to a high enough depth to achieve complete lineage-level resolution. *C. acnes* phylotypes were
627 defined following an existing single locus strain-typing (SLST) scheme (*49*). To ensure a wide
628 collection of phylogenetic diversity for *C. acnes*, we also included four publicly available
629 isolates from *C. acnes* SLST L (ribotype III) into the reference database at the phylotype level
630 (NCBI strain names: Cacnes_PMH7, Cacnes_PMH5, Cacnes_JCM_18919,
631 Cacnes_JCM_18909). Because the raw reads for these genomes were not publicly available, we
632 first simulated reads from each genome using wgsim (v.0.3.1-r13; -e 0.0 -d 500 -N 500000 -1
633 150 -2 150 -r 0.0 -R 0.0 -X 0.0) before inputting into the PHLAME pipeline. Four *S. epidermidis*
634 phylotypes were determined manually by cutting at the longest branch lengths of the inter-
635 lineage phylogeny (which matches other *S. epidermidis* phylogenies; Fig. S14). Because there
636 were only four major *S. epidermidis* phylotypes, of which one (phylotype 4) included >75% of
637 the lineages, we also defined a more resolved sub-phylotype level for *S. epidermidis* by picking
638 the set of monophyletic clades that both split up phylotype 4 and led to the highest average
639 percent assigned across concatenated metagenomics samples (Fig. S14). Phylotypes and sub-
640 phylotypes were detected in a similar way to lineages, using instead a database of specific SNPs
641 common to all isolates in a phylotype but not found in any other phylotypes.
642
643 <u>Filtering and selecting metagenomics samples</u>
644
645 *Filtering by sample*
646 Lineage-level abundances from the preceding section were used to filter out potentially
647 contaminated metagenomics samples. Lineages were defined as "potentially contaminating
648 lineages" if they were found in any sample from any handler (researcher who collected or
649 processed primary samples) at >3% abundance. All tubes in which the total abundance of
650 potentially contaminating lineages was >5% were excluded (11 tubes). Next, samples from
651 subject 8PB from time point 1 were excluded since their questionnaire indicated that they were
652 currently taking topical antibiotics.
653 Samples with <50000 reads assigned at the genus-level from Bracken, across sequencing
654 runs, were excluded. Abundance data for samples which were sequenced twice or more was
655 averaged and re-normalized, for a total of 360 kit samples passing filters.
656
657 *Excluding lineages*
658 The abundances of two types of lineages were excluded from metagenomics samples: (1)
659 lineages for which we only had 1 or 2 corresponding isolates and (2) lineages which appeared to
660 be contaminants from mock community samples processed on the same plates. *C. acnes* lineages

661 43 and 116 and *S. epidermidis* lineages 6 and 45 were used in high-biomass mock communities
662 (isolated from 1AA, 4AA, and 5PA) on plates MG1, MG2, and MG3 (Sup. Fig. S18) and
663 identified in many samples. We therefore conservatively removed their abundances. In total, the
664 abundance data of 87 *C. acnes* and 76 *S. epidermidis* lineages with sufficient corresponding
665 isolate data and no evidence of contamination were used.
666
667 Lineage sharing across subjects
668 To identify lineages shared across subjects, the combined set of lineage calls was used to
669 determine how many lineages ever found on a subject are also found on other subjects. Only
670 subject/time point samples for which >70% of abundance could be assigned were used. Only
671 lineages with an inferred abundance of >=1% in were included as present to limit spurious calls
672 from computational or experimental errors. All lineages found across timepoints are included– a
673 lineage is 'shared' between two subjects even if it is found in a single time point for each.
674
675 Within lineage SNPs
676 Reference genomes were assembled for each lineage to identify mutations in genes
677 present in the flexible gene content of lineages but absent in the reference genomes Pacnes_C1
678 or SepidermidisATCC12228. First, the same number of reads from each sample (equal to the
679 number of reads in the sample with the least reads) were concatenated into individual files for
680 each lineage. Then, Spades was used (-m 500 -k 21,33,55,77 --phred-offset 33 --careful) to
681 assemble these reads into contigs. We then manually applied a minimum-coverage filter for
682 contigs in each assembly, which was usually at half the coverage of the longest contig.
683 In order to identify outgroups for each clade, the average pairwise distance between
684 lineages was obtained using the same distance matrix as the one used to cluster lineages above.
685 For each lineage, the 5 nearest lineages were identified, and the highest coverage sample from
686 each of these lineages was used as the outgroup. Reads from each sample (all isolates from the
687 lineage and all 5 outgroup samples) were realigned to the lineage's co-assembly for downstream
688 SNP identification and phylogeny construction using bowtie, as above.
689 To identify *de novo* SNPs in each lineage and build accurate phylogenies, calls were first
690 strictly filtered using the same general process as above to identify positions where most samples
691 have high-quality calls. To initially filter low-quality calls in the alignments to lineage co-
692 assemblies, we considered calls with a FQ score <30, major allele frequency < .75 or coverage
693 <4 as "N". Next, we excluded positions >40% of samples have an N in that position. To build
694 phylogenies, nucleotide calls at accepted positions were less strictly filtered to retain information
695 at these highly informative positions. Major-allele nucleotides were used as calls provided they
696 had a major allele frequency of 70% and a coverage of at 5 reads (forward or reverse).
697 To identify outgroup positions, we first considered positions where at least 2 outgroup
698 samples had the same call and there were no other alleles. If there were any remaining calls
699 without outgroup calls, the reference genome call was used. If there are any positions which are
700 N in the reference genome, these would be replaced by the modal call from the ingroup. In

701 addition to excluding low-quality positions and calls, we also excluded recombinant positions.
702 We excluded regions where SNPs occurred within a 500bp sliding window with a covariation
703 >.75. We constructed maximum-parsimony phylogenies for each lineage using DNApars.
704
705 Identifying genes within and across lineage assemblies
706 First, clade co-assemblies were annotated using Bakta (v1.6.1) (*50*). To identify which of
707 these genes were the same across lineage co-assemblies, Roary (v3.13.0, -i 90) (*51*) was used to
708 cluster protein sequences into homologous clusters. Genes from different lineage genomes were
709 considered the same gene if they were in the same Roary cluster.
710
711 Subject dMRCAs
712 To calculate $dMRCA_{SUBJECT}$ for a given lineage (Figs. 4 and 5), we only considered
713 instances where each subject had at least 5 isolates in that lineage. Each subject's isolates were
714 considered independently; as such, per-subject lineage dMRCAs ($dMRCA_{SUBJECT}$) were
715 calculated as the mean number of *de novo* mutations excluding mutations common to all
716 subjects' isolates.
717
718 Directionality of transmission
719 Directionality was only determined for lineages from which: (1) at least 2 subjects had at
720 least 3 isolates each in a lineage; (2) at least one subject had a dMRCA ($dMRCA_{SUBJECT}$) which
721 was less than the dMRCA of the clade (recipient); (3) and at least one subject's $dMRCA_{SUBJECT}$
722 was the same as the root of the lineage (source). Where we list each instance of inferrable
723 directionality in supplementary material (Fig. S5), we only show instances in which only
724 parent(s) or child(ren) are the source and /or recipient. For example, a clear instance of parent-to-
725 child transmission could include an instance where either parent could be the source, but the
726 child is clearly the recipient.
727
728 Adaptation analysis
729 To search for significantly mutated genes, we first counted how many times each gene
730 was mutated. We excluded mutations which appeared to be the result of recombination (see:
731 Within-lineage SNPs). We then computed the probability of observing at least as many
732 mutations as the number observed according to a Poisson distribution as done previously (*13*).
733 This calculation takes into account the differences in gene length, codon distribution, and
734 mutational spectrum across lineages by taking the average of each of these factors across
735 lineages with mutation in the gene under consideration. To measure the false-discovery rate, we
736 used the Benjamini-Hochberg procedure where each gene was treated as a separate hypothesis.
737 Calculations of dN/dS for each gene were also computed as previously described in (*13*).
738 SNPs in individual lineages and in homologous gene clusters across lineages are listed in table
739 S6.
740

741 Number of genotype transmissions

742 Since it was often not possible to determine which of the subjects were the source of the
743 lineage and which was the donor, we conservatively estimated the number of transmitted
744 genotypes with at least 5 isolates from each of at least 2 subjects using a parsimony assumption
745 as previously reported (*52*). For each lineage, the smallest number of genotypes required to
746 explain the diversity on all subjects was calculated for each potential source. The instance
747 (source) with the smallest total number of genotype transmission was chosen for each lineage.
748 The number of inferred transmitted genotypes for each lineage are listed with associated
749 metadata in table S4.
750

751 Molecular clock estimation

752 Molecular clock signals from individual people were calculated for all instances where a
753 subject had at least 5 isolates in a given lineage from at least 2 timepoints. For each of *C. acnes*
754 lineage, there was no significant correlation of the number of *de novo* mutations in isolates
755 versus time since initial sampling at a Benjamini-Hochberg false discovery rate (FDR) of 1% for
756 *C. acnes.* For *S. epidermidis,* 3 lineages had significant clock signals for *S. epidermidis* after
757 FDR correction. The estimated rates for *S. epidermidis* ranged from 4.8 to 13
758 SNPs/genome/year. Since these estimates varied by a factor of ~3, we also estimated a molecular
759 clock for *S. epidermidis* across lineages and subjects. For each lineage and for each subject with
760 at least 5 isolates at two timepoints, we plotted the average root-to-tip distance of isolates from
761 each timepoint against the time elapsed since the first sampled timepoint. Together, a significant
762 molecular clock signal of 4.51 mutations/genome/year was estimated, which was slightly slower
763 and therefore more conservative than the slowest of the three individually calculated clock rates
764 and consistent with published rates for this species (*14*, *32–34*).
765
766
767
768
769
770
771
772
773
774
775
776
777

References

1. E. K. Costello, C. L. Lauber, M. Hamady, N. Fierer, J. I. Gordon, R. Knight, Bacterial community variation in human body habitats across space and time. *Science*. **326**, 1694–1697 (2009).

2. M. R. Olm, C. T. Brown, B. Brooks, B. Firek, R. Baker, D. Burstein, K. Soenjoyo, B. C. Thomas, M. Morowitz, J. F. Banfield, Identical bacterial populations colonize premature infant gut, skin, and oral microbiomes and exhibit different in situ growth rates. *Genome Res.* **27**, 601–612 (2017).

3. T. A. Harris-Tryon, E. A. Grice, Microbiota and maintenance of skin barrier function. *Science*. **376**, 940–945 (2022).

4. J. Oh, A. L. Byrd, C. Deming, S. Conlan, NISC Comparative Sequencing Program, H. H. Kong, J. A. Segre, Biogeography and individuality shape function in the human skin metagenome. *Nature*. **514**, 59–64 (2014).

5. J. Oh, A. L. Byrd, M. Park, NISC Comparative Sequencing Program, H. H. Kong, J. A. Segre, Temporal Stability of the Human Skin Microbiome. *Cell*. **165**, 854–866 (2016).

6. Human Microbiome Project Consortium, Structure, function and diversity of the healthy human microbiome. *Nature*. **486**, 207–214 (2012).

7. T. D. Lieberman, Detecting bacterial adaptation within individual microbiomes. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **377**, 20210243 (2022).

8. T. Van Rossum, P. Ferretti, O. M. Maistrenko, P. Bork, Diversity within species: interpreting strains in microbiomes. *Nat. Rev. Microbiol.* **18**, 491–506 (2020).

9. B. Paetzold, J. R. Willis, J. Pereira de Lima, N. Knödlseder, H. Brüggemann, S. R. Quist, T. Gabaldón, M. Güell, Skin microbiome modulation induced by probiotic solutions. *Microbiome*. **7**, 95 (2019).

10. B. Koskella, L. J. Hall, C. J. E. Metcalf, The microbiome beyond the horizon of ecological and evolutionary theory. *Nat Ecol Evol*. **1**, 1606–1615 (2017).

11. T. S. B. Schmidt, J. Raes, P. Bork, The Human Gut Microbiome: From Association to Modulation. *Cell*. **172**, 1198–1215 (2018).

12. Z. Shen, L. Robert, M. Stolpman, Y. Che, K. J. Allen, R. Saffery, A. Walsh, A. Young, J. Eckert, C. Deming, Q. Chen, S. Conlan, K. Laky, J. M. Li, L. Chatman, S. S. Kashaf, NISC Comparative Sequencing Program, VITALITY team, H. H. Kong, P. A. Frischmeyer-Guerrerio, K. P. Perrett, J. A. Segre, A genome catalog of the early-life human skin microbiome. *Genome Biol.* **24**, 252 (2023).

13. A. Conwill, A. C. Kuan, R. Damerla, A. J. Poret, J. S. Baker, A. D. Tripp, E. J. Alm, T. D. Lieberman, Anatomy promotes neutral coexistence of strains in the human skin

813 microbiome. *Cell Host Microbe*. **30**, 171–182.e7 (2022).

814 14. F. M. Key, V. D. Khadka, C. Romo-González, K. J. Blake, L. Deng, T. C. Lynn, J. C. Lee,
815 I. M. Chiu, M. T. García-Romero, T. D. Lieberman, On-person adaptive evolution of
816 Staphylococcus aureus during treatment for atopic dermatitis. *Cell Host Microbe*. **31**, 593–
817 603.e7 (2023).

818 15. F. Hildebrand, T. I. Gossmann, C. Frioux, E. Özkurt, P. N. Myers, P. Ferretti, M. Kuhn, M.
819 Bahram, H. B. Nielsen, P. Bork, Dispersal strategies shape persistence and evolution of
820 human gut bacteria. *Cell Host Microbe*. **29**, 1167–1176.e9 (2021).

821 16. K. S. Xue, S. J. Walton, D. A. Goldman, M. L. Morrison, A. J. Verster, A. B. Parrott, F. B.
822 Yu, N. F. Neff, N. A. Rosenberg, B. D. Ross, D. A. Petrov, K. C. Huang, B. H. Good, D. A.
823 Relman, Prolonged delays in human microbiota transmission after a controlled antibiotic
824 perturbation. *bioRxiv* (2023), doi:10.1101/2023.09.26.559480.

825 17. M. Valles-Colomer, A. Blanco-Míguez, P. Manghi, F. Asnicar, L. Dubois, D. Golzato, F.
826 Armanini, F. Cumbo, K. D. Huang, S. Manara, G. Masetti, F. Pinto, E. Piperni, M.
827 Punčochář, L. Ricci, M. Zolfo, O. Farrant, A. Goncalves, M. Selma-Royo, A. G. Binetti, J.
828 E. Becerra, B. Han, J. Lusingu, J. Amuasi, L. Amoroso, A. Visconti, C. M. Steves, M.
829 Falchi, M. Filosi, A. Tett, A. Last, Q. Xu, N. Qin, H. Qin, J. May, D. Eibach, M. V. Corrias,
830 M. Ponzoni, E. Pasolli, T. D. Spector, E. Domenici, M. C. Collado, N. Segata, The person-
831 to-person transmission landscape of the gut and oral microbiomes. *Nature*. **614**, 125–135
832 (2023).

833 18. Y. C. Lou, M. R. Olm, S. Diamond, A. Crits-Christoph, B. A. Firek, R. Baker, M. J.
834 Morowitz, J. F. Banfield, Infant gut strain persistence is associated with maternal origin,
835 phylogeny, and traits including surface adhesion and iron acquisition. *Cell Rep Med*. **2**,
836 100393 (2021).

837 19. M. Yassour, E. Jason, L. J. Hogstrom, T. D. Arthur, S. Tripathi, H. Siljander, J. Selvenius,
838 S. Oikarinen, H. Hyöty, S. M. Virtanen, J. Ilonen, P. Ferretti, E. Pasolli, A. Tett, F. Asnicar,
839 N. Segata, H. Vlamakis, E. S. Lander, C. Huttenhower, M. Knip, R. J. Xavier, Strain-Level
840 Analysis of Mother-to-Child Bacterial Transmission during the First Few Months of Life.
841 *Cell Host Microbe*. **24**, 146–154.e4 (2018).

842 20. D. W. Chen, N. R. Garud, Rapid evolution and strain turnover in the infant gut microbiome.
843 *Genome Res.* **32**, 1124–1136 (2022).

844 21. S. Nayfach, B. Rodriguez-Mueller, N. Garud, K. S. Pollard, An integrated metagenomics
845 pipeline for strain profiling reveals novel patterns of bacterial transmission and
846 biogeography. *Genome Res.* **26**, 1612–1625 (2016).

847 22. P. E. Pochi, J. S. Strauss, D. T. Downing, Age-related changes in sebaceous gland activity.
848 *J. Invest. Dermatol.* **73**, 108–111 (1979).

849 23. J. Oh, S. Conlan, E. C. Polley, J. A. Segre, H. H. Kong, Shifts in human skin and nares
850 microbiota of healthy children and adults. *Genome Med.* **4**, 77 (2012).

24. J. Park, N. H. Schwardt, J.-H. Jo, Z. Zhang, V. Pillai, S. Phang, S. M. Brady, J. A. Portillo, M. A. MacGibeny, H. Liang, M. Pensler, S. J. Soldin, J. A. Yanovski, J. A. Segre, H. H. Kong, Shifts in the Skin Bacterial and Fungal Communities of Healthy Children Transitioning through Puberty. *J. Invest. Dermatol.* **142**, 212–219 (2022).

25. A. M. Schneider, Z. T. Nolan, K. Banerjee, A. R. Paine, Z. Cong, S. L. Gettle, A. L. Longenecker, X. Zhan, G. W. Agak, A. M. Nelson, Evolution of the facial skin microbiome during puberty in normal and acne skin. *J. Eur. Acad. Dermatol. Venereol.* **37**, 166–175 (2023).

26. J. J. Leyden, K. J. McGinley, O. H. Mills, A. M. Kligman, Age-related changes in the resident bacterial flora of the human face. *J. Invest. Dermatol.* **65**, 379–381 (1975).

27. E. M. Acosta, K. A. Little, B. P. Bratton, J. G. Lopez, X. Mao, A. S. Payne, M. Donia, D. Devenport, Z. Gitai, Bacterial DNA on the skin surface overrepresents the viable skin microbiome. *Elife*. **12** (2023), doi:10.7554/eLife.87192.

28. L. McInnes, J. Healy, N. Saul, L. Großberger, UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **3**, 861 (2018).

29. L. McInnes, J. Healy, S. Astels, hdbscan: Hierarchical density based clustering. *J. Open Source Softw.* **2**, 205 (2017).

30. S. P. Hubbell, *The Unified Neutral Theory of Biodiversity and Biogeography* (Princeton University Press, 2001).

31. W. Zhou, M. Spoto, R. Hardy, C. Guan, E. Fleming, P. J. Larson, J. S. Brown, J. Oh, Host-Specific Evolutionary and Transmission Dynamics Shape the Functional Diversification of Staphylococcus epidermidis in Human Skin. *Cell*. **180**, 454–470.e18 (2020).

32. U. Nübel, M. Nachtnebel, G. Falkenhorst, J. Benzler, J. Hecht, M. Kube, F. Bröcker, K. Moelling, C. Bührer, P. Gastmeier, B. Piening, M. Behnke, M. Dehnert, F. Layer, W. Witte, T. Eckmanns, MRSA transmission on a neonatal intensive care unit: epidemiological and genome-based phylogenetic analyses. *PLoS One*. **8**, e54898 (2013).

33. B. C. Young, T. Golubchik, E. M. Batty, R. Fung, H. Larner-Svensson, A. A. Votintseva, R. R. Miller, H. Godwin, K. Knox, R. G. Everitt, Z. Iqbal, A. J. Rimmer, M. Cule, C. L. C. Ip, X. Didelot, R. M. Harding, P. Donnelly, T. E. Peto, D. W. Crook, R. Bowden, D. J. Wilson, Evolutionary dynamics of *Staphylococcus aureus* during progression from carriage to disease. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 4550–4555 (2012).

34. T. Golubchik, E. M. Batty, R. R. Miller, H. Farr, B. C. Young, H. Larner-Svensson, R. Fung, H. Godwin, K. Knox, A. Votintseva, R. G. Everitt, T. Street, M. Cule, C. L. C. Ip, X. Didelot, T. E. A. Peto, R. M. Harding, D. J. Wilson, D. W. Crook, R. Bowden, Within-host evolution of Staphylococcus aureus during asymptomatic carriage. *PLoS One*. **8**, e61319 (2013).

35. C. Guan, P. J. Larson, E. Fleming, A. P. Tikhonov, S. Mootien, T. H. Grossman, C. Golino,

J. Oh, Engineering a “detect and destroy” skin probiotic to combat methicillin-resistant Staphylococcus aureus. *PLoS One*. **17**, e0276795 (2022).

36. S. Fitz-Gibbon, S. Tomida, B.-H. Chiu, L. Nguyen, C. Du, M. Liu, D. Elashoff, M. C. Erfe, A. Loncaric, J. Kim, R. L. Modlin, J. F. Miller, E. Sodergren, N. Craft, G. M. Weinstock, H. Li, Propionibacterium acnes strain populations in the human skin microbiome associated with acne. *J. Invest. Dermatol.* **133**, 2152–2160 (2013).

37. J. Claesen, J. B. Spagnolo, S. F. Ramos, K. L. Kurita, A. L. Byrd, A. A. Aksenov, A. V. Melnik, W. R. Wong, S. Wang, R. D. Hernandez, M. S. Donia, P. C. Dorrestein, H. H. Kong, J. A. Segre, R. G. Linington, M. A. Fischbach, K. P. Lemon, A Cutibacterium acnes antibiotic modulates human skin microbiota composition in hair follicles. *Sci. Transl. Med.* **12** (2020), doi:10.1126/scitranslmed.aay5445.

38. D. Gaio, K. Anantanawat, J. To, M. Liu, L. Monahan, A. E. Darling, Hackflex: low-cost, high-throughput, Illumina Nextera Flex library construction. *Microb Genom*. **8** (2022), doi:10.1099/mgen.0.000744.

39. BBMap. *SourceForge* (2022), (available at https://sourceforge.net/projects/bbmap/).

40. D. E. Wood, J. Lu, B. Langmead, Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).

41. J. Lu, F. P. Breitwieser, P. Thielen, S. L. Salzberg, Bracken: estimating species abundance in metagenomics data. *PeerJ Comput. Sci.* **3**, e104 (2017).

42. A. Bankevich, S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham, A. D. Prjibelski, A. V. Pyshkin, A. V. Sirotkin, N. Vyahhi, G. Tesler, M. A. Alekseyev, P. A. Pevzner, SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).

43. T. Seemann, Prokka: rapid prokaryotic genome annotation. *Bioinformatics*. **30**, 2068–2069 (2014).

44. W. Li, A. Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*. **22**, 1658–1659 (2006).

45. J. Kleinberg, "An impossibility theorem for clustering" in *Proceedings of the 15th International Conference on Neural Information Processing Systems* (MIT Press, Cambridge, MA, USA, 2002), *NIPS’02*, pp. 463–470.

46. D. J. Watts, S. H. Strogatz, Collective dynamics of “small-world” networks. *Nature*. **393**, 440–442 (1998).

47. M. Ester, H.-P. Kriegel, J. Sander, X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise" in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (AAAI Press, 1996), *KDD’96*, pp. 226–231.

924 48. *phlame_beta* (Github; https://github.com/quevan/phlame_beta).

925 49. C. F. P. Scholz, A. Jensen, H. B. Lomholt, H. Brüggemann, M. Kilian, A novel high-
926 resolution single locus sequence typing scheme for mixed populations of Propionibacterium
927 acnes in vivo. *PLoS One*. **9**, e104199 (2014).

928 50. O. Schwengers, L. Jelonek, M. A. Dieckmann, S. Beyvers, J. Blom, A. Goesmann, Bakta:
929 rapid and standardized annotation of bacterial genomes via alignment-free sequence
930 identification. *Microb Genom*. **7** (2021), doi:10.1099/mgen.0.000685.

931 51. A. J. Page, C. A. Cummins, M. Hunt, V. K. Wong, S. Reuter, M. T. G. Holden, M. Fookes,
932 D. Falush, J. A. Keane, J. Parkhill, Roary: rapid large-scale prokaryote pan genome
933 analysis. *Bioinformatics*. **31**, 3691–3693 (2015).

934 52. T. D. Lieberman, D. Wilson, R. Misra, L. L. Xiong, P. Moodley, T. Cohen, R. Kishony,
935 Genomic diversity in autopsy samples reveals within-host dissemination of HIV-associated
936 Mycobacterium tuberculosis. *Nat. Med.* **22**, 1470–1474 (2016).

937