Supplementary Information

Supporting Text

16S rDNA sequence of KPA171202

All sequenced *P. acnes* genomes encode three copies of 16S rRNA genes, which are identical within each isolate, except KPA171202. Based on the KPA171202 genome (Bruggemann *et al.*, 2004), one copy of the 16S rRNA gene has one nucleotide difference from the other two identical copies of RT1. However, this mutation was never observed in our 16S rDNA dataset. We amplified, cloned and sequenced multiple clones of 16S rDNA gene from KPA171202 and did not find a sequence harboring this mutation. Therefore, we believe that KPA171202 also has three identical copies of 16S rDNA.

Comparison of *P. acnes* strain distribution to other human microbiome datasets

To determine whether the *P. acnes* ribotypes and their relative abundances measured in this study are unique to pilosebaceous units, we applied a similar analysis to the microbiome 16S rDNA data from the Human Microbiome Project (HMP) and the data from Grice *et al*. (2009). Both datasets were obtained from healthy subjects. The relative abundance of the major ribotypes in healthy subjects from our study was largely similar to that found in these two datasets despite the fact that they were sampled from different anatomical sites (Figure S2). RT6 (6.3%) was found to be more abundant than RT4 and RT5 combined (2.8%) in the HMP data, similar to those found in our normal cohort where RT6 represents 4.8% and RT4 and RT5 combined represent 1.2% of the clones. The same five main microbiome types were observed in the two datasets (Figure S5). This also suggests that our sampling and analysis of the microbiome were comparable to other studies.

Genome clustering and phylogenetic tree

The *recA* gene has been widely used to classify *P. acnes* strains into four known types: IA, IB, II and III (McDowell *et al.*, 2008; McDowell *et al.*, 2005). The phylogenetic tree of the 71 genomes based on the SNPs in the core genome matched the *recA* types perfectly except one isolate, HL097PA1. Most of the genomes with ribotypes 1, 4, 5 and 532 were grouped to *recA* Type IA clade, which can be further divided into subclades IA-1 and IA-2. Clade IA-2 is composed of mostly RT4 and RT5. RT4 and most of RT5 genomes seem to belong to the same lineage with very similar genome sequences. All the isolates with ribotypes 3, 8 and 16, who share the mutation of T1007C in the 16S rDNA gene, were grouped to *recA* Type IB clade. Most of the RT3 genomes form a subclade IB-2 and RT8 genomes form a subclade by themselves, IB-1, which was highly associated with acne. Notably, RT2 and RT6, who share T854C mutation, seem to have a more distant phylogenetic relationship to other ribotypes, and were grouped to the *recA* Type II clade. This is consistent with previous studies (Lomholt and Kilian, 2010; McDowell *et al.*, 2005). We did not find *P. acnes* isolates with *recA* type III in our samples.

We further analyzed the associations of *P. acnes* lineages with health and disease states. There was a clear shift of the association strength of the clades with acne along the phylogenetic tree (Figure S6). The three sequenced ribotypes identified as being strongly associated with acne (RT4, RT5, and RT8) were found at one end of the tree in clades IA-2 and IB-1, while the RT6 identified as being associated with normal skin was at the other end of the tree at the tip of clade II (Figure S6).

Antibiotic resistance

P. acnes ribotypes 4, 5 and 10 have a single nucleotide substitution G1058C in the 16S rDNA sequences, which has previously been shown to confer increased resistance to tetracycline (Ross *et al.*, 1998a; Ross *et al.*, 2001). In addition to the substitution in the 16S rDNA sequences, we found that all the strains of RT4 and RT5 that we sequenced have a nucleotide substitution in the 23S rDNA sequences, which confers increased resistance to a different class of antibiotics, erythromycin and clindamycin (Ross *et al.*, 1997; Ross *et al.*, 1998b). We experimentally confirmed that these isolates except two that were unculturable were resistant to tetracycline, erythromycin and clindamycin.

We examined whether the enrichment of these ribotypes in the acne group could be due to antibiotic treatment. However, in our study only a small percentage of the subjects harboring ribotypes 4, 5 or 10 were treated with antibiotics (Table S2). Eighteen of the 29 subjects who harbored any of these three ribotypes gave reports on both past and current treatments. Among them, 50% (9/18) of the subjects were never treated; 33% (6/18) were treated with retinoids; 11% (2/18) were treated with antibiotics in the past, and 5.6% (1/18) were treated with both antibiotics and retinoids in the past. Therefore, the theory of selection by antibiotic treatment is not favored in our study. Previous surveys of antibiotic resistant strains in acne patients demonstrated that previous use of antibiotics did not always result in the presence of resistant strains and that some patients without previous use of antibiotics harbored resistant strains (Coates *et al.*, 2002; Dreno *et al.*, 2001). Our observation in this study is consistent with these previous studies. Nonetheless, we cannot completely rule out a possible influence of past

antibiotic exposure, other acne therapies, or an altered cutaneous environment, on the distribution of ribotypes in individuals.

CRISPR spacer sequences

Although more similar to the GC content of *P. acnes* genomes, four unique spacer sequences found in strains of RT2 and RT6 have the best matches to the genome of *Clostridium leptum*, a commensal bacterium in the gut microbiota (Table 2). On the 55Kb plasmid harbored in HL096PA1 and other RT4 and RT5 genomes, there is also a large cluster of 35 genes that are identical to the genes found in *C. leptum*, including the Tad locus. While *P. acnes* is also found in the gut microbiota, it is unclear how these genes were horizontally transferred between *P. acnes* and *C. leptum* or possibly from a progenitor organism. Further investigation into the relationships between the microbes at different body sites will be crucial to determine how genetic materials are transferred between different microbiomes and how this mechanism affects the health state of the host.

Materials and Methods

Metagenomic DNA extraction, PCR amplification, cloning and 16S rDNA sequencing

Metagenomic DNA extraction – Individual microcomedones were isolated from the adhesive nose strip using sterile forceps and placed in a 2 mL sterile microcentrifuge tube filled with ATL buffer (Qiagen) and 0.1 mm diameter glass beads (BioSpec Products, Inc., Bartlesville, OK). Cells were lysed using a beadbeater for 3 minutes at 4,800 rpm at room temperature. After centrifugation at 14,000 rpm for 5 minutes, the supernatant was retrieved and used for genomic DNA extraction using QIAamp DNA Micro Kit (Qiagen). The manufacturer protocol for

extracting DNA from chewing gum was used. Concentration of the genomic DNA was determined by NanoDrop 1000 Spectrophotometer.

16S rDNA PCR amplification, cloning and sequencing – Most of the metagenomic samples were amplified in triplicate using 16S rDNA specific primers with the following sequences: 27f-MP 5'-AGRGTTTGATCMTGGCTCAG-3' and 1492r-MP 5'-TACGGYTACCTTGTTAYGACTT-3'. PCR reactions contained 0.5 U/μL Platinum Taq DNA Polymerase High Fidelity (Invitrogen), 1X Pre-mix E PCR buffer from Epicentre Fail-Safe PCR system, 0.12 μM concentration of each primer 27f-MP and 1492r-MP, and Sigma PCR grade water. One microliter of DNA (ranging from 0.2 - 10 ng total) was added to each reaction. The G-Storm GS4 thermocycler conditions were as following: initial denaturation of 96°C for 5 minutes, and 30 cycles of denaturation at 94°C for 30 seconds, annealing at 57°C for 1 minute, and extension at 72°C for 2 minutes, with a final extension at 72°C for 7 minutes. Following amplification, an A-tailing reaction was performed by the addition of 1 U of GOTaq DNA Polymerase directly to the amplification reaction and incubation in the thermocycler at 72°C for 10 minutes.

The three PCR amplification reactions from each source DNA were pooled and gel purified (1.2% agarose gel stained with SYBR Green fluorescent dye). The 1.4 Kb product was excised and further purified using the Qiagen QIAquick Gel Extraction kit. The purified product was cloned into OneShot *E. coli* cells using TOPO TA cloning kit from Invitrogen.

White colonies were picked into a 384-well tray containing terrific broth, glycerol, and kanamycin using a Qpix picking robot. Each tray was prepared for sequencing using a magnetic

bead prep from Agilent and sequenced with 1/16th Big Dye Terminator from ABI. Sequencing was done with a universal forward, universal reverse, and for a subset, internal 16S rDNA primer 907R with sequences of TGTAAAACGACGGCCAGT (forward), CAGGAAACAGCTATGACC (reverse), and CCGTCAATTCCTTTRAGTTT (907R). Sequence reactions were loaded on ABI 3730 machines from ABI on 50 cm arrays with a long read run module.

A slightly different PCR and cloning protocol without automation was used for several initial samples as described below. 16S rDNA was amplified using universal primers 8F (5'-AGAGTTTGATYMTGGCTCAG-3') and 1510R (5'-TACGGYTACCTTGTTACGACTT-3') (Gao *et al.*, 2007). Thermocycling conditions were as following: initial denaturation step of 5 minutes at 94°C, 30 cycles of denaturation at 94°C for 45 seconds, annealing at 52°C for 30 seconds and elongation at 72°C for 90 seconds, and a final elongation step at 72°C for 20 minutes.

PCR products were purified using DNA Clean and Concentrator Kit (Zymo Research). Subsequently, the 16S rDNA amplicons were cloned into pCR 2.1-TOPO vector (Invitrogen). One-Shot TOP-10 Chemically Competent *E. coli* cells (Invitrogen) were transformed with the vectors and plated on selective media. Individual positive colonies were picked and inoculated into selective LB liquid medium. After 14 hours of incubation, the plasmids were extracted and purified using PrepEase MiniSpin Plasmid Kit (USB Corporation) or Zyppy Plasmid Miniprep Kit (Zymo Research). The clones were sequenced bidirectionally using Sanger sequencing method with 1/8th chemistry using ABI 3730 sequencer (Applied Biosystems Inc.).

P. acnes isolation and culturing

Sample culture plate – Microcomedones on the inner surface of the nose strip were mashed and scraped using a sterile loop (Fisherbrand, Pittsburgh, PA), and plated onto a blood agar plate (Teknova Brucella Agar Plate with Hemin and Vitamin K, Teknova, Hollister, CA). The plates were incubated at 37°C for 5 - 7 days anaerobically using the AnaeroPack System (Mitsubishi Gas Chemical Company, Tokyo, Japan).

Isolation and culturing of individual strains - Colonies with the macroscopic characteristics of *P. acnes* were picked from each sample plate and were streaked onto A-media plates (Pancreatic Digase of Casine, Difco yeast extract, glucose, KH_2PO_4, $MgSO_4$, Difco Agar, and water). These first-pass plates were then incubated anaerobically at 37°C for 5 - 7 days. As the second pass, single isolated colonies were picked from the first-pass plates and streaked onto new A-Media plates. These plates were then incubated anaerobically at 37°C for 5 - 7 days. The colonies on these plates were picked for culturing, genotyping, and genome sequencing in the subsequence steps.

Genotyping of the P. acnes isolates – each isolate was analyzed by PCR amplification of the 16S rDNA gene. The ribotypes were determined based on the full length sequences. Isolates with desired ribotypes were selected for future culturing and genome sequencing.

Genomic DNA extraction of P. acnes isolates - Isolates were grown in 5 mL of Clostridial medium under anaerobic conditions at 37°C for 5 - 7 days. Cultures were pelleted by

centrifugation and washed with 3 mL phosphate buffer saline (PBS). The same protocol used for the metagenomic DNA extraction was used for extracting the genomic DNA of the isolates.

Metagenomic shotgun sequencing and analysis

Metagenomic DNA samples from microcomedone samples from 22 individuals with normal skin were pooled and sequenced using Roche/454 FLX. The average read length was 236 bp. The sequencing was limited with 13,291 sequence reads. Sequence reads were aligned against the NCBI's non-redundant database using BLAST. Species assignment was based on 97% identity and 100% of the read length aligned.

Assembly, alignment and editing of 16S rDNA sequences

Assembly and alignment - Base calling and quality were determined with Phred (Ewing and Green, 1998; Ewing *et al.*, 1998) using default parameters. Bidirectional reads were assembled and aligned to a core set of NAST-formatted sequences (rRNA16S.gold) using AmosCmp16Spipeline and NAST-ier, which are from the Microbiome Utilities Portal of the Broad Institute (http://microbiomeutil.sourceforge.net/). These tools in turn use Amoscmp (Pop *et al.*, 2004), Mummer (Kurtz *et al.*, 2004), Lucy (Chou and Holmes, 2001), BLAST (Altschul *et al.*, 1990) and CdbTools (http://compbio.dfci.harvard.edu/tgi/software/). Suspected chimeras were identified using ChimeraSlayer and WigeoN (Haas *et al.*, 2011). Sequences with at least 90% bootstrap support for a chimeric breakpoint (ChimeraSlayer) or containing a region that varies at more than the 99% quantile of expected variation (WigeoN) were removed from further analysis.

Quality screening - For diversity analysis of the *P. acnes* population, sequences with at least 99% identity over 1,400 nucleotides to *P. acnes* KPA171202 (Bruggemann *et al.*, 2004) 16S rDNA were trimmed to positions 29-1483 (numbering based on the *E. coli* system of nomenclature (Brosius *et al.*, 1978)). Sequences without full coverage over this region were excluded from further strain level analysis. Chimera screening, as described above, resulted in removal of less than 0.35% of the sequences. This may be an under-estimation of the chimeras, since the majority of sequences differ by only 1 or 2 nucleotides. Low quality sequences were excluded, defined as more than 50 nucleotides between positions 79 and 1433 with Phred quality scores of less than 15. To allow detailed strain-level analysis, the data were extensively manually edited. Chromatograms were visually inspected at all bases with a Phred quality score < 30, and appropriate corrections were applied. For analysis at the species level, the 16S rDNA sequences were not manually edited. Chimera screening of assembled sequences resulted in removal of less than 0.65% of the sequences. Aligned sequences were trimmed to *E. coli* equivalent positions 29-1483 (Brosius *et al.*, 1978). Sequences without full coverage over this region were excluded from further analysis.

Sequence editing - Nearly 62,000 Sanger sequence reads representing the 26,446 assembled *P. acnes* sequences were mapped to the RT1 sequence in CONSED (Gordon, 2003; Gordon *et al.*, 1998). Comprehensive semi-manual editing of the large number of sequences was made feasible by their very high pairwise similarities: a median of only one nucleotide change from RT1 per sequence (three nucleotide changes prior to editing). Editing was facilitated by the use of scripts and the custom navigation feature of CONSED allowing single click jumps to sites requiring inspection. Chromatograms were inspected for all low quality (Phred < 30) bases that differed

from RT1, and corrected as needed, including many commonly occurring sequence errors. In order to minimize the effect of base mis-incorporation and chimera, specific base differences from RT1 occurring in less than 4 sequences (frequency < 0.00015) were considered unreliable and reverted to the corresponding RT1 base. Ribotypes were assigned for the resulting sequences based on 100% identity.

16S rDNA sequence analysis

OTUs and taxonomy assignments – QIIME (Caporaso *et al.*, 2010b) was used to cluster the sequences into OTUs using 99% identity cutoff, furthest neighbor, and UCLUST (Edgar, 2010). Representative sequences (most abundant) were selected and aligned using PYNAST (Caporaso *et al.*, 2010a) to the greengenes database. Taxonomy was assigned using RDP method (Cole *et al.*, 2009). The alignment was filtered with the lanemask provided by greengenes, and a phylogenetic tree was built using FastTree (Price *et al.*, 2009).

Wilcoxon test on the top ten ribotypes - For each sample, the number of clones of each of the top ten ribotypes was normalized by the total number of *P. acnes* clones of the sample. The normalized counts were used to test the significance in enrichment between the acne group and the normal group. The function wilcox_test in the R program (http://www.R-project.org) was used to calculate the p-values.

Microbiome type assignments – Microbiome types were assigned based on the largest clades seen when samples were clustered using thetayc similarity in MOTHUR (Schloss *et al.*, 2009) (Figures 2 and S4) or hierarchical clustering (Eisen *et al.*, 1998) (Figure S5).

Assigning ribotypes to datasets of HMP and Grice et al. 2009 - Sequences were assigned to a ribotype if they met the following criteria. First, there was a single best match. Second, it covered the range required to discriminate between the top 45 ribotypes (58-1388). Third, there were no Ns at discriminatory positions. Lastly, there were no more than ten non-discriminatory differences.

The HMP 16S rDNA Sanger sequence dataset was downloaded with permission from the HMP Data Analysis and Coordination Center. It has 8,492 *P. acnes* sequences from 14 subjects and nine body sites (retroauricular crease, anterior nares, hard palate, buccal mucosa, throat, palatine tonsils, antecubital fossa, saliva, and subgingival plaque). More details on the dataset can be found at http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000228.v2.p1. In this dataset, low quality bases (Phred quality < 20) were converted to Ns, and 26% of the sequences were not assigned due to excessive Ns or Ns at ribotype discriminatory sites. Less than 1% was unresolved due to equal best matches or greater than ten mismatches to RT1.

The dataset from Grice *et al.* (2009) is available at NCBI (GenBank accession numbers GQ000001 to GQ116391). It has 22,378 *P. acnes* sequences from ten subjects and 21 skin sites (buttock, elbow, hypothenar palm, volar forearm, antecubital fossa, axillary vault, gluteal crease, inguinal crease, interdigital web space, nare, plantar heel, popliteal fossa, toe web space, umbilicus, alar crease, back, external auditory canal, glabella, manubrium, occiput, and retroauricular crease). Three percent of the sequences were unassigned due to greater than ten mismatches to RT1, and 1.6% was unassigned due to equal best matches.

For comparison purpose, our unedited 16S rDNA sequences were assigned to ribotypes by the same method described above and the result is shown in Figure S2. Less than 0.6% of the sequences were unassigned due to greater than ten mismatches to RT1, and 1.7 % was unassigned due to equal best matches.

Whole genome shotgun sequencing, assembly and annotation of 66 *P. acnes* isolates

Genome HL096PA1 - The genome was sequenced using Roche/454 FLX at the UCLA Genotyping and Sequencing Core. A total of 590,054 sequence reads were generated with an average read length of 230 bp. Of these, 433,896 were assembled into two contigs, a circular main chromosome of 2,494,190 bp and a linear plasmid of 55,585 bp. Assembly was accomplished by a combination of PHRAP/CONSED (Gordon *et al.*, 1998) and GSMAPPER (Roche) with extensive manual editing in CONSED. GeneMark v2.6r (Borodovsky and McIninch, 1993) and GLIMMER v2.0 (Salzberg *et al.*, 1998) were used to performed *ab initio* protein coding gene prediction. tRNAScan-SE 1.23 was used for tRNA identification and RNAmmer was used for predicting ribosomal RNA genes (5S, 16S, and 23S). Genome annotation results were based on automated searches in public databases, including Pfam (http://pfam.jouy.inra.fr/), KEGG (http://www.genome.jp/kegg), and COG (http://www.ncbi.nlm.nih.gov/COG/). Manual inspection of the annotation was also performed.

Genomes of the other 65 isolates - The genomes were sequenced using Illumina/Solexa Genome Analyzer IIx and annotated by the Genome Center of Washington University at St. Louis.

Assembly: Each genomic DNA sample was randomly sheared and an indexed library was constructed using standard Illumina protocols. Twelve uniquely tagged libraries were pooled and run on one lane of a GAIIx flowcell and paired end sequences were generated. Following deconvolution of the tagged reads into the separate samples, datasets were processed using BWA (Li and Durbin, 2009) quality trimming at a q10 threshold. Reads trimmed to less than 35bp in length were discarded and the remaining reads were assembled using oneButtonVelvet, an optimizer program that runs the Velvet assembler (Zerbino and Birney, 2008) numerous times over a user supplied k-mer range while varying several of the assembler parameters and optimizing for the assembly parameter set which yields the longest N50 contig length.

Annotation: Coding sequences were predicted using GeneMark v3.3 (Borodovsky and McIninch, 1993) and GLIMMER v2.13 (Salzberg *et al.*, 1998). Intergenic regions not spanned by GeneMark and GLIMMER were aligned using BLAST against NCBI's non-redundant database and predictions were generated based on protein alignments. tRNA genes were determined using tRNAscan-SE 1.23 and non-coding RNA genes were determined by RNAmmer-1.2 and Rfam v8.0. The final gene set was processed through a suite of protein categorization tools consisting of Interpro, psort-b and KEGG. The gene product naming comes from the BER pipeline (JCVI). A more detailed standard operating protocol (SOP) can be found at http://hmpdacc.org/doc/sops/reference_genomes/annotation/WUGC_SOP_DACC.pdf.

71 *P. acnes* genome analysis and comparison

Identification of the core regions of P. acnes genomes - The “core” regions were defined as genome sequences that are present in all 71 genomes. *P. acnes* KPA171202 was used as the

reference genome. Each of the other 70 genome sequences (a series of contigs in most of the genomes and two complete genomes) was mapped to the reference genome using Nucmer (Kurtz *et al.*, 2004). All the 70 ".coords" output files of Nucmer program were analyzed to identify overlap regions based on the KPA171202 coordinates using a Perl script. Finally, "core" sequences were extracted based on the genome sequence of KPA171202 with the coordinates calculated above. On average, 90% (ranging from 88% to 92%) of the genomes were included in the core regions.

Identification of SNPs in the core regions – Single nucleotide polymorphisms (SNPs) were identified by using "show-snps" utility option of the Nucmer program (Kurtz *et al.*, 2004) with the default settings. Genome sequence of *P. acnes* KPA171202 was used as the reference genome. All the 70 ".snps" output files of Nucmer program were analyzed to identify unique SNP positions based on the KPA171202 coordinates using a Perl script. The SNPs in the core regions were further analyzed to construct a phylogenetic tree.

Phylogenetic tree construction - The 71 concatenated sequences of the 96,887 SNP nucleotides in the core regions were used to construct a phylogenetic tree of the *P. acnes* genomes. The evolutionary distance of the core regions among the genomes was inferred using the Neighbor-Joining method (Saitou and Nei, 1987). The bootstrap tree inferred from 1,000 replicates was taken. Branches corresponding to partitions reproduced in less than 80% bootstrap replicates were collapsed. Figure 3 shows only the topology. In Figure S6, the tree was drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the p-distance method and

are in the units of the number of nucleotide differences per site. This tree shows the comparison based on only the core regions. The distance does not represent the true evolutionary distance between different genomes, since the non-core regions of each genome were not considered here. All positions containing gaps and missing data were eliminated. Evolutionary analysis was conducted using MEGA5 (Tamura *et al.*, 2007).

Gene content comparison - In order to assess the conservation of gene content across the 71 genomes, protein coding genes in all the genomes were clustered using UCLUST (Edgar, 2010) by first sorting by decreasing length then clustering each sequence to an existing seed sequence if it had at least 90% nucleotide identity over its entire length, otherwise it became a new seed. For visualization, the data were reformatted to columns and rows representing genes and genomes, respectively. One or more copies of the genes in a genome were treated as present. Gene columns were sorted by their position based on the coordinates of the HL096PA1 genome, a fully finished genome with a 55Kb plasmid. Genome rows were sorted by their positions in the SNP-based Neighbor Joining tree described above.

Identification of CRISPR/Cas – CRISPRFinder (Grissa *et al.*, 2007) was used to identify the CRISPR repeat-spacer sequences. The annotation of HL110PA3 was used for BLAST alignment in order to identify the presence of CRISPR/Cas structure and CRISPR repeat-spacer sequences in strains of HL001PA1, HL060PA1, HL082PA2, HL103PA1, HL106PA1, HL110PA4 and J139. Each spacer sequence was annotated by BLAST alignment against NCBI's non-redundant nucleotide database and the reference genomic sequences database (refseq_genomic).

Sequence coverage analysis – MAQ (Li *et al.*, 2008) was used to map the raw sequence reads from Illumina/Roche platform to the reference genomes. Briefly, "map" command was used for mapping, and "assemble" command was used for calling the consensus sequences from read mapping, then "cnd2win" command was used to extract information averaged in a tilling window. A window size of 1,000 bp was used. Randomly selected 1 million reads were used for mapping. This accounted for approximately 40X coverage for all the genomes except HL096PA2, HL096PA3, HL097PA1 and HL099PA1, which had approximately 55X to 75X coverage. BWA (Li and Durbin, 2010) was used to map the raw sequence reads from Roche/454 platform to the reference genome HL096PA1. The average coverage was calculated in 1,000 bp window.

Quantitative PCR

Quantitative PCR (qPCR) targeting *TadA* on the plasmid (Locus 3) and housekeeping genes *Pak* and *RecA* on the chromosome was performed using the genomic DNA extracted from the *P. acnes* isolates. LightCyler 480 High Resolution Melting Master kit was used (Roche Diagnostics GmbH, Mannheim, Germany). Each 10 µL reaction solution was consisted of 5 µL master mix (2X concentrate), 1 µL 25 mM $MgCl_2$, 0.5 µL 4 µM forward and reverse primers, and DNA template. Four qPCR runs were performed on Roche LightCycler 480. Primer sequences for *TadA* are 5'-GATAATCCGTTCGACAAGCTG-3' (forward) and 5'-ACCCACCACGATGATGTTT-3' (reverse). Primer sequences for *pak* are 5'-CGACGCCTCCAATAACTTCC-3' (forward) and 5'-GTCGGCCTCCTCAGCATC-3' (reverse). Primer sequences for *recA* are 5'-CCGGAGACAACGACAGGT-3' (forward) and 5'-

GCTTCCTCATACCACTGGTCATC-3' (reverse). All samples were run in duplicates in each qPCR run, except the second run, which was not duplicated. Thermocycling conditions were as following: initial activation step of 10 minutes at 95^{o}C; 50 amplification cycles with each consisting of 10 seconds at 95^{o}C, 15 seconds at 65^{o}C in the first cycle with a stepwise 0.5^{o}C decrease for each succeeding cycle, and 30 seconds at 72^{o}C; and final melting curve step starting at 65^{o}C and ending at 99^{o}C with a ramp rate of 0.02 ^{o}C/s and acquisition rate of 25/^{o}C. DNA concentration standards were run in duplicates. Copy number ratios of genes were calculated based on the concentrations of the genes on the plasmid and chromosome.

Data Availability

16S rDNA sequences have been deposited at GenBank under the project ID 46327. Whole genome shotgun sequences and annotations of the *P. acnes* strains have been deposited at GenBank under the accession numbers ADWB00000000, ADWC00000000, ADWF00000000, ADWH00000000, ADWI00000000, ADXP00000000, ADXQ00000000, ADXR00000000, ADXS00000000, ADXT00000000, ADXU00000000, ADXW00000000, ADXX00000000, ADXY00000000, ADXZ00000000, ADYA00000000, ADYB00000000, ADYC00000000, ADYD00000000, ADYE00000000, ADYF00000000, ADYG00000000, ADYI00000000, ADYJ00000000, ADYK00000000, ADYL00000000, ADYM00000000, ADYN00000000, ADYO00000000, ADYP00000000, ADYQ00000000, ADYR00000000, ADYS00000000, ADYT00000000, ADYU00000000, ADYV00000000, ADYW00000000, ADYX00000000, ADYY00000000, ADYZ00000000, ADZA00000000, ADZB00000000, ADZC00000000, ADZD00000000, ADZE00000000, ADZF00000000, ADZG00000000, ADZH00000000, ADZI00000000, ADZJ00000000, ADZK00000000, ADZL00000000, ADZM00000000,

ADZN00000000, ADZO00000000, ADZP00000000, ADZQ00000000, ADZR00000000, ADZS00000000, ADZT00000000, ADZV00000000, ADZW00000000, CP003293, and CP003294.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215:403-10.

Brosius J, Palmer ML, Kennedy PJ, Noller HF (1978) Complete nucleotide sequence of a 16S ribosomal RNA gene from Escherichia coli. *Proc Natl Acad Sci U S A* 75:4801-5.

Caporaso JG, Bittinger K, Bushman FD, DeSantis TZ, Andersen GL, Knight R (2010a) PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics* 26:266-7.

Chou HH, Holmes MH (2001) DNA sequence quality trimming and vector removal. *Bioinformatics* 17:1093-104.

Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ, *et al.* (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res* 37:D141-5.

Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26:2460-1.

Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* 95:14863-8.

Gordon D (2003) Viewing and editing assembled sequences using Consed. *Curr Protoc Bioinformatics* Chapter 11:Unit11 2.

Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26:589-95.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754-60.

Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* 18:1851-8.

McDowell A, Perry AL, Lambert PA, Patrick S (2008) A new phylogenetic group of Propionibacterium acnes. *J Med Microbiol* 57:218-24.

McDowell A, Valanne S, Ramage G, Tunney MM, Glenn JV, McLorinan GC, *et al.* (2005) Propionibacterium acnes types I and II represent phylogenetically distinct groups. *J Clin Microbiol* 43:326-34.

Pop M, Phillippy A, Delcher AL, Salzberg SL (2004) Comparative genome assembly. *Brief Bioinform* 5:237-48.

Price MN, Dehal PS, Arkin AP (2009) FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol* 26:1641-50.

Ross JI, Eady EA, Cove JH, Jones CE, Ratyal AH, Miller YW, *et al.* (1997) Clinical resistance to erythromycin and clindamycin in cutaneous propionibacteria isolated from acne patients is associated with mutations in 23S rRNA. *Antimicrob Agents Chemother* 41:1162-5.

Ross JI, Eady EA, Cove JH, Ratyal AH, Cunliffe WJ (1998b) Resistance to erythromycin and clindamycin in cutaneous propionibacteria is associated with mutations in 23S rRNA. *Dermatology* 196:69-70.

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution* 4:406-25.

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, *et al.* (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* 75:7537-41.

Table S1. Six phyla and 42 genera found in pilosebaceous units.

Phylum	Genus	Phylum	Genus
Actinobacteria	Actinobaculum	Bacteroidetes	Chryseobacterium
	Corynebacterium		Niastella
	Gordonia		Parabacteroides
	Kocuria		Prevotella
	Microbacterium	Proteobacteria	Caulobacteraceae
	Propionibacterium		Citrobacter
Firmicutes	Anaerococcus		Cupriavidus
	Anoxybacillus		Delftia
	Bacillus		Diaphorobacter
	Enterococcus		Haemophilus
	Erysipelothrix		Klebsiella
	Finegoldia		Massilia
	Gemella		Neisseriaceae
	Lactobacillus		Novosphingobium
	Paenibacillus		Pelomonas
	Peptoniphilus		Phyllobacterium
	Peptostreptococcaceae		Ralstonia
	Ruminococcaceae		Shigella
	Staphylococcus		Sphingomonas
	Streptococcus		Stenotrophomonas
Fusobacteria	Fusobacterium	Cyanobacteria	Streptophyta

Table S2. Past and current treatments of the subjects.

Group	Acne		Normal	
Number of subjects	49		52	
	with RT4, RT5, or RT10	without RT4, RT5, and RT10	with RT4, RT5, or RT10	without RT4, RT5, and RT10
Number of subjects in each subgroup	20	29	9	43
Subjects reported on current treatment	14	25	8	31
no treatment	10	21	8	30
antibiotics	0	2	0	0
retinoids	3	2	0	0
antibiotics and retinoids	0	0	0	1
unknown	1	0	0	0
Subjects reported on past treatment	12	22	8	31
no treatment	5	16	6	28
antibiotics	2	4	0	1
retinoids	4	1	1	2
antibiotics and retinoids	1	0	1	0
unknown	0	1	0	0

Table S3. P-values calculated using four different statistical tests of non-random distribution between groups show that *P. acnes* population structures in acne patients and normal individuals were significantly different.

Method	Unweighted Unifrac	Mothur Anosim	Parsimony	Mothur Amova
Top10RT_byDiseaseState	0.0096	0.045	0.051	0.069
Top10RT_byGender	0.55	0.78	0.90	0.79
Top10RT_byRandom[a]	0.55	0.40	0.83	0.64
Top110RT_byDiseaseState	0.00098	0.042	0.050	0.061
Top110RT_byGender	0.63	0.72	0.80	0.82
Top110RT_byRandom[a]	0.67	0.40	0.49	0.51

The tests were performed for two different sets of ribotypes: the top ten most abundant ribotypes and the top 110 most abundant ribotypes (with 9 or more clones), and for three pairs of groups: acne vs. normal skin; male vs. female; or random pairing. The analyses were based on thetayc sample to sample distances and did not consider distance between ribotypes, i.e. only ribotype counts. All tests were based on 100,000 iterations and were performed using MOTHUR (Schloss *et al.*, 2009).

[a]random assignment of subjects to groups A and B (repeated five times and averaged)

Table S4. Summary of genes encoded in loci 1, 2, 3.

Gene annotation category	Locus 1	Locus 2	Locus 3
Hypothetical protein	3	14	33
ABC transporter	1	2	
Site-specific recombinase	2		
N-acetylmuramoyl-L-alanine amidase	1		
CobQ/CobB/MinD/ParA nucleotide binding domain protein		1	3
Yag1E		1	1
Sag protein family		3	
CAAX amino terminal protease family protein		1	
Single-strand binding protein		1	
Tad protein family			8
Rcp protein family			2
Sigma subunit, Ecf family/anti-sigma factor			2
Flp-1			1
RepA (replication protein)			1
ResA (resolvase)			1
Plasmid stabilization system protein			1
Permease (putative)			1
Ribonuclease E (putative)			1
Ribbon-helix-helix protein, copG family (putative)			1
Not annotated (no homologue found)			18
Total number of genes in the locus	7	23	74

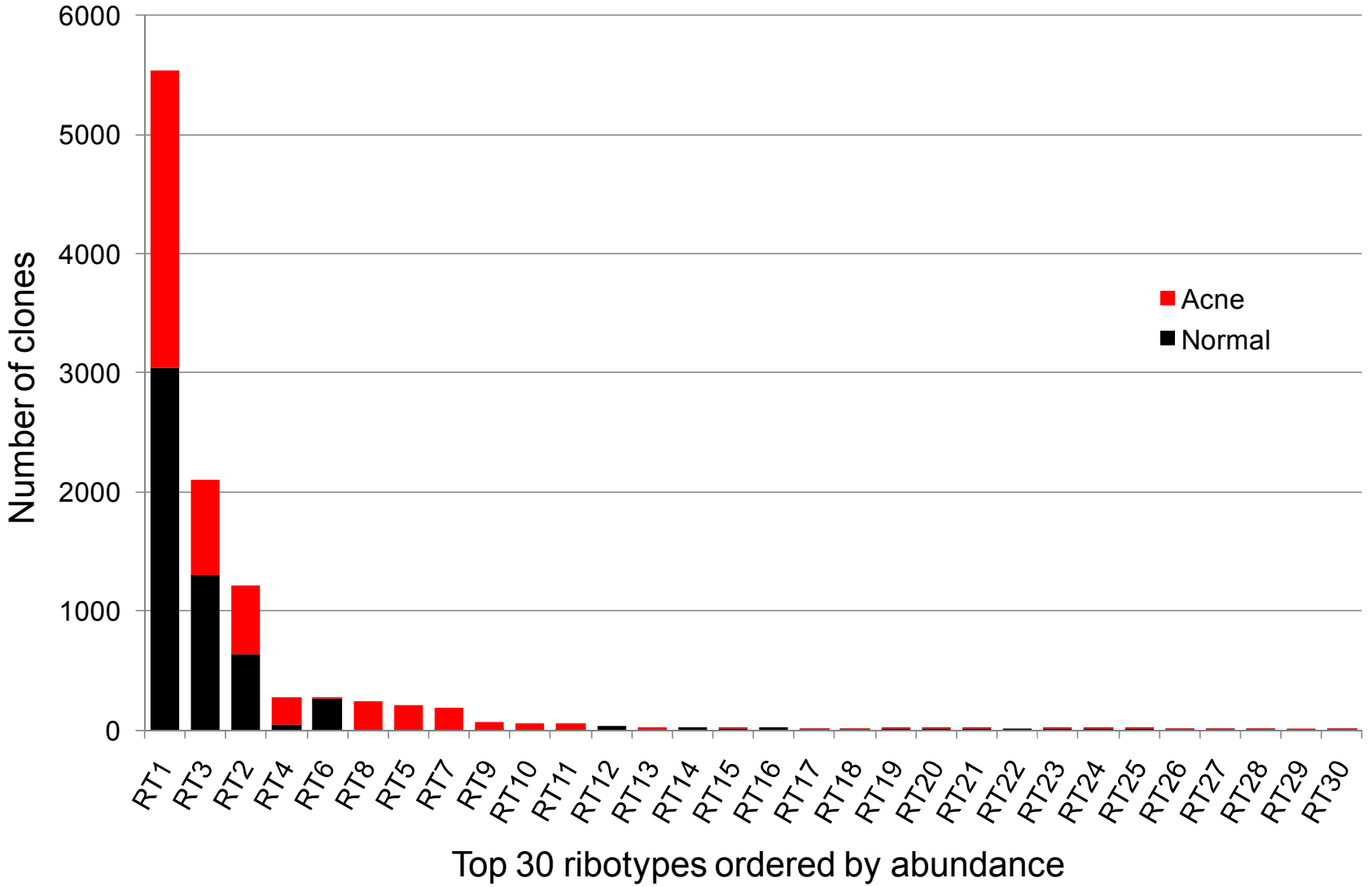

Figure S1. Rank abundance of *P. acnes* ribotypes shows a distribution similar to that seen at the higher taxonomic levels. A few highly abundant ribotypes and a large number of rare ribotypes were observed in the samples. Some ribotypes were highly enriched in acne patients. Only top 30 most abundant ribotypes are shown for graphing purpose.

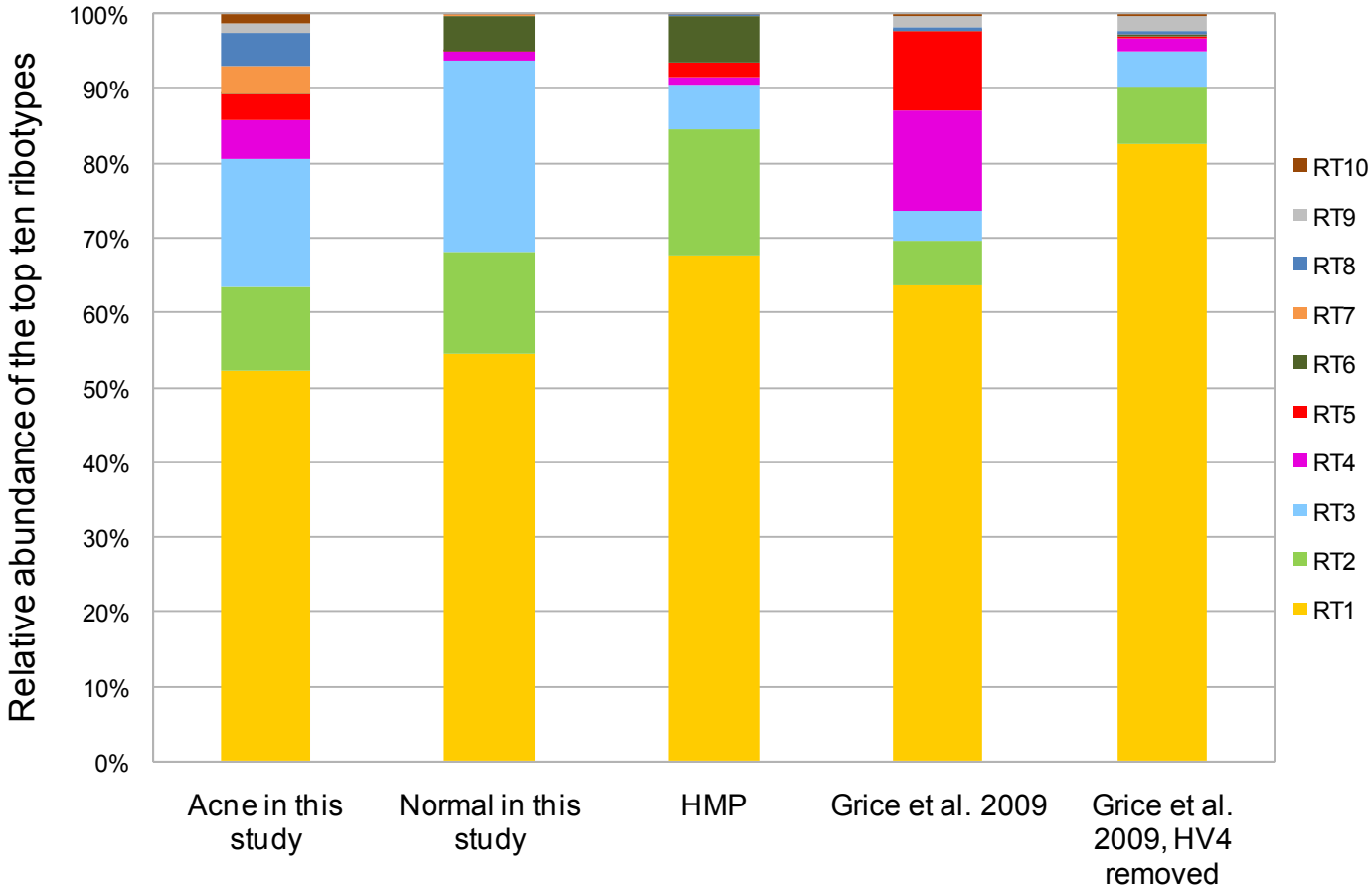

Figure S2. The most abundant *P. acnes* ribotypes in pilosebaceous units were also abundant at other body sites. The major ribotypes found in acne patients and normal individuals from this study were compared to the datasets from the HMP and Grice et al. (2009). The top three ribotypes are the most abundant ones in different datasets. The excess RT4 and RT5 seen in the dataset by Grice et al. (2009) was due to one subject, HV4, whose *P. acnes* strain population was dominated by these two ribotypes at every skin site sampled. After removal of this subject, the ribotype distribution is similar to the HMP samples and the normal skin samples in this study. RT6 is also found abundant in the HMP dataset, which were collected from healthy individuals.

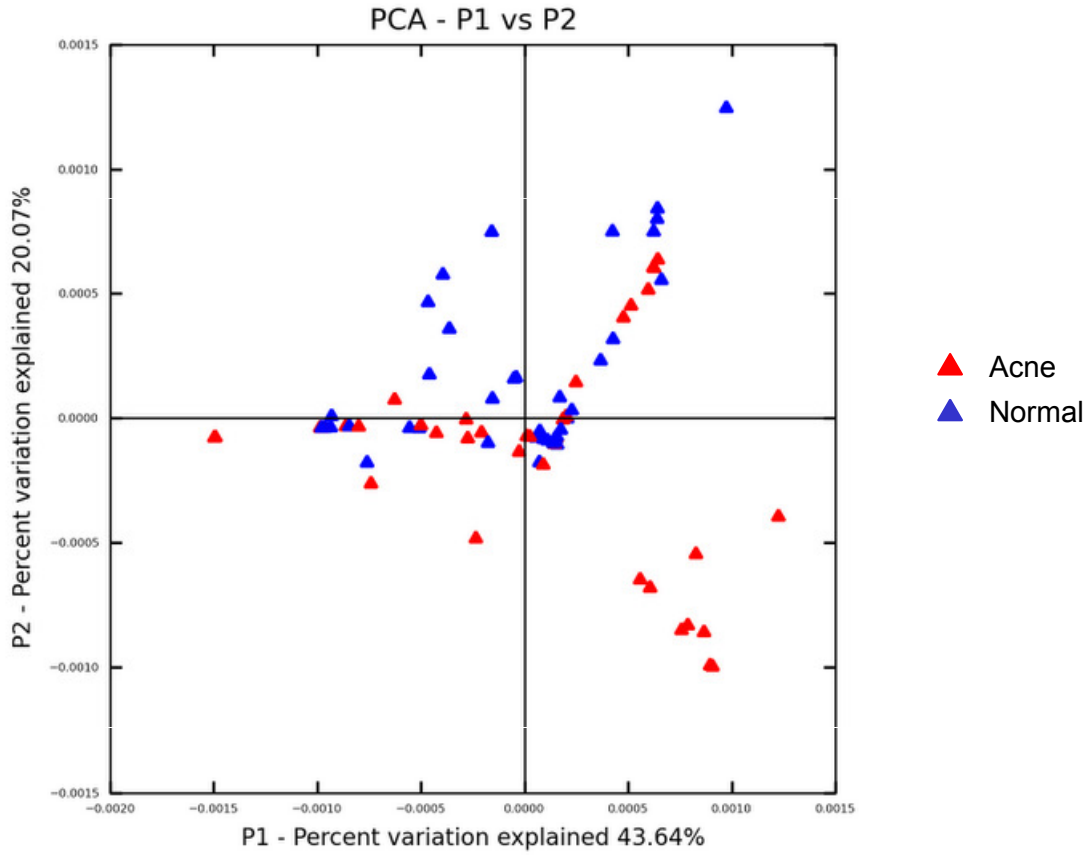

Figure S3. *P. acnes* population structures differ in acne and normal skin. *P. acnes* populations from samples were clustered using principal coordinates analysis of the weighted UniFrac distance matrix for the top ten most abundant ribotypes. The principal coordinate 1 (P1) explains 43.64% of the variation and P2 explains 20.07% of the variation. Analysis was performed using QIIME (Caporaso et al., 2010b).

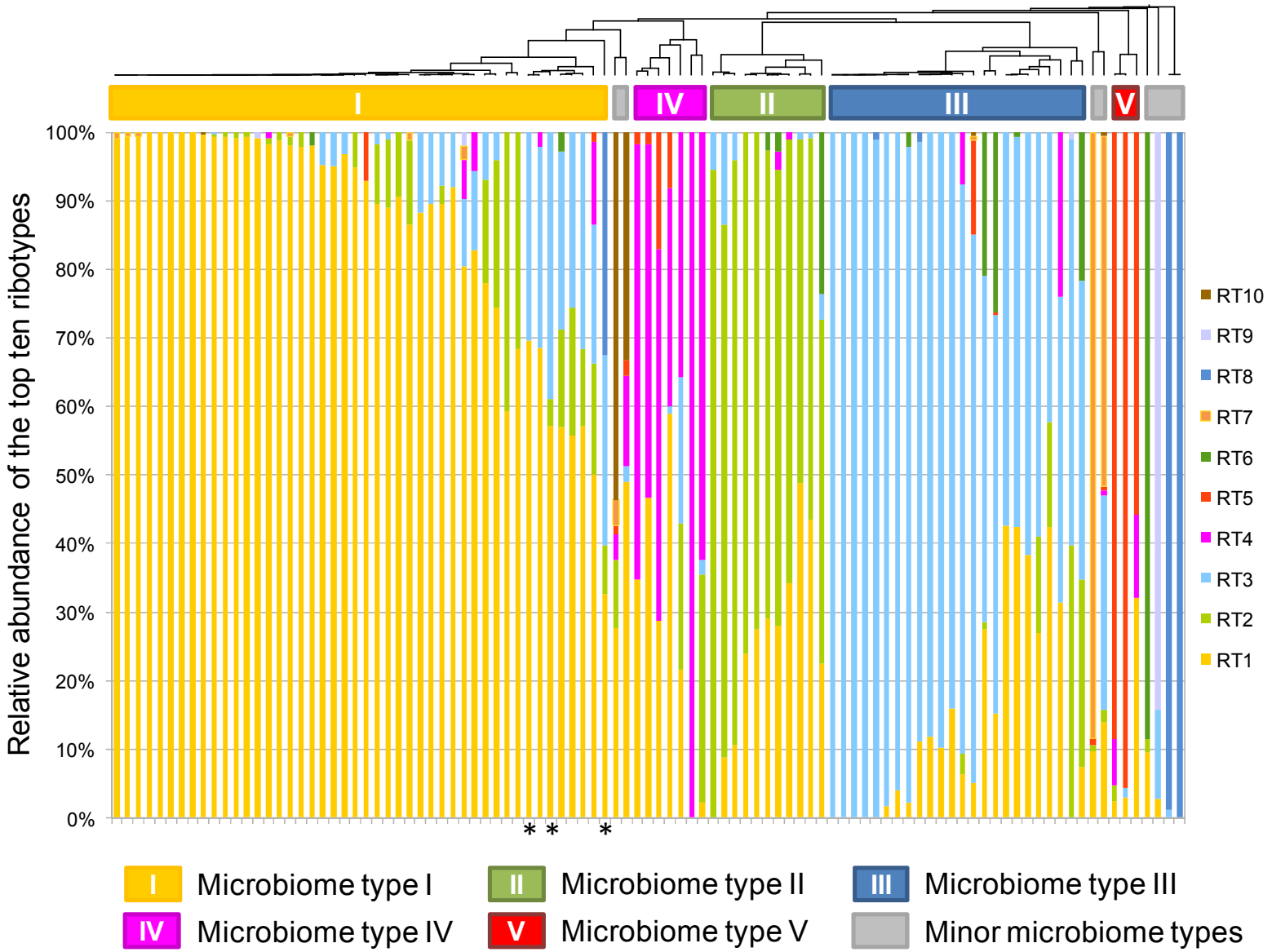

Figure S4. Distribution of the top ten most abundant *P. acnes* ribotypes in all samples without separating the two groups of acne and normal skin. Each column represents the percentage of the top ten ribotypes identified in each sample. When all samples were clustered, we observed the same five major microbiome types at the *P. acnes* strain level. This suggests that the microbiome classification does not depend on the states of the disease. Only three out of 99 samples were clustered differently compared to the one shown in Figure 2 (marked with asterisks). Two samples (one from acne, one from normal skin) with fewer than 50 *P. acnes* 16S rDNA sequences are not shown.

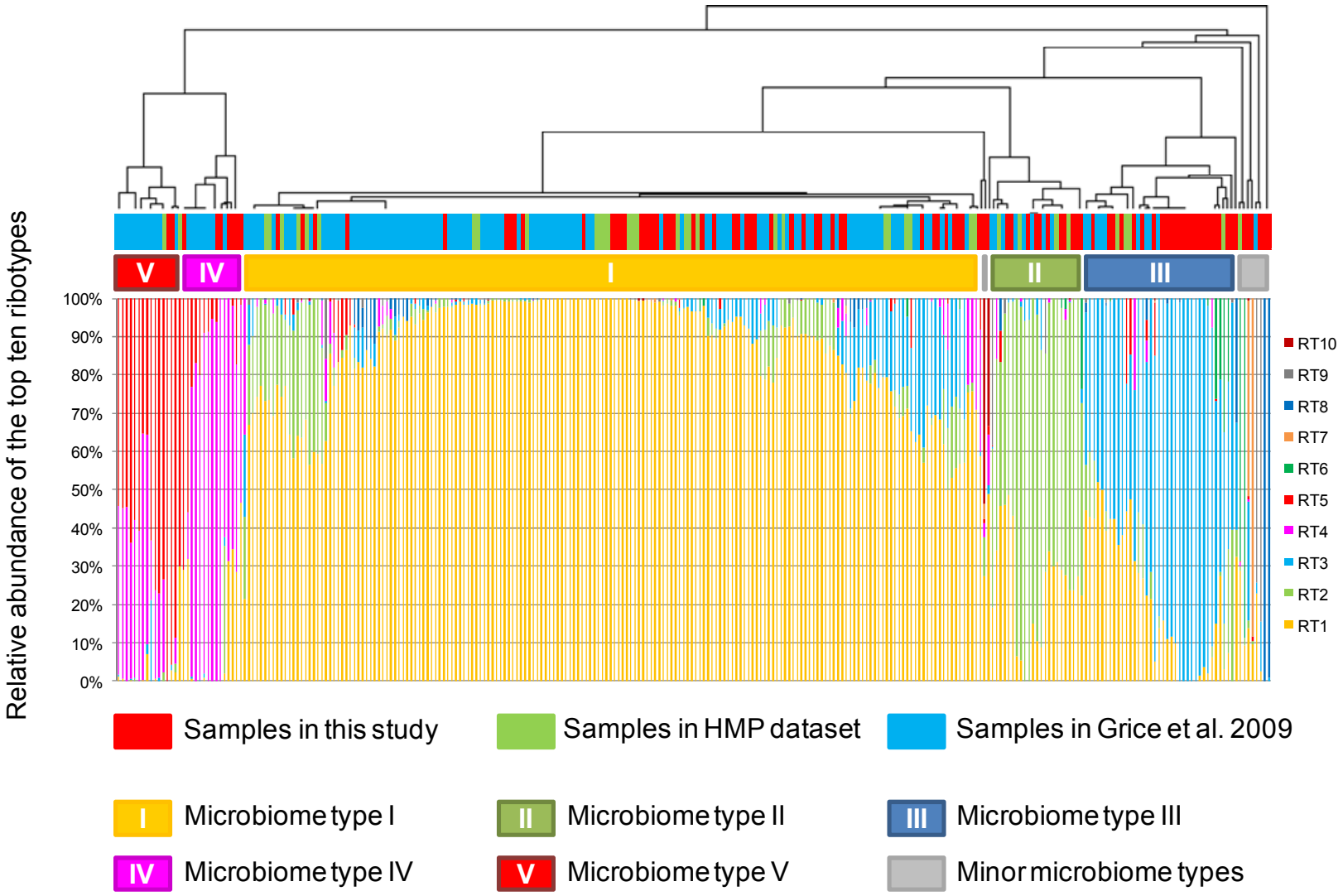

Figure S5. The same five major microbiome types were observed in multiple datasets. Samples from this study, HMP and Grice et al. (2009) were clustered together based on the top ten most abundant *P. acnes* ribotypes. In total 284 samples were included. Each column represents the percentage of the top ten ribotypes identified in each sample. Both HMP samples and samples from Grice et al. (2009) were collected from healthy individuals, therefore the percentage of microbiome types IV and V are under-represented in this analysis. Samples with fewer than ten sequences of the top ten ribotypes were not included.

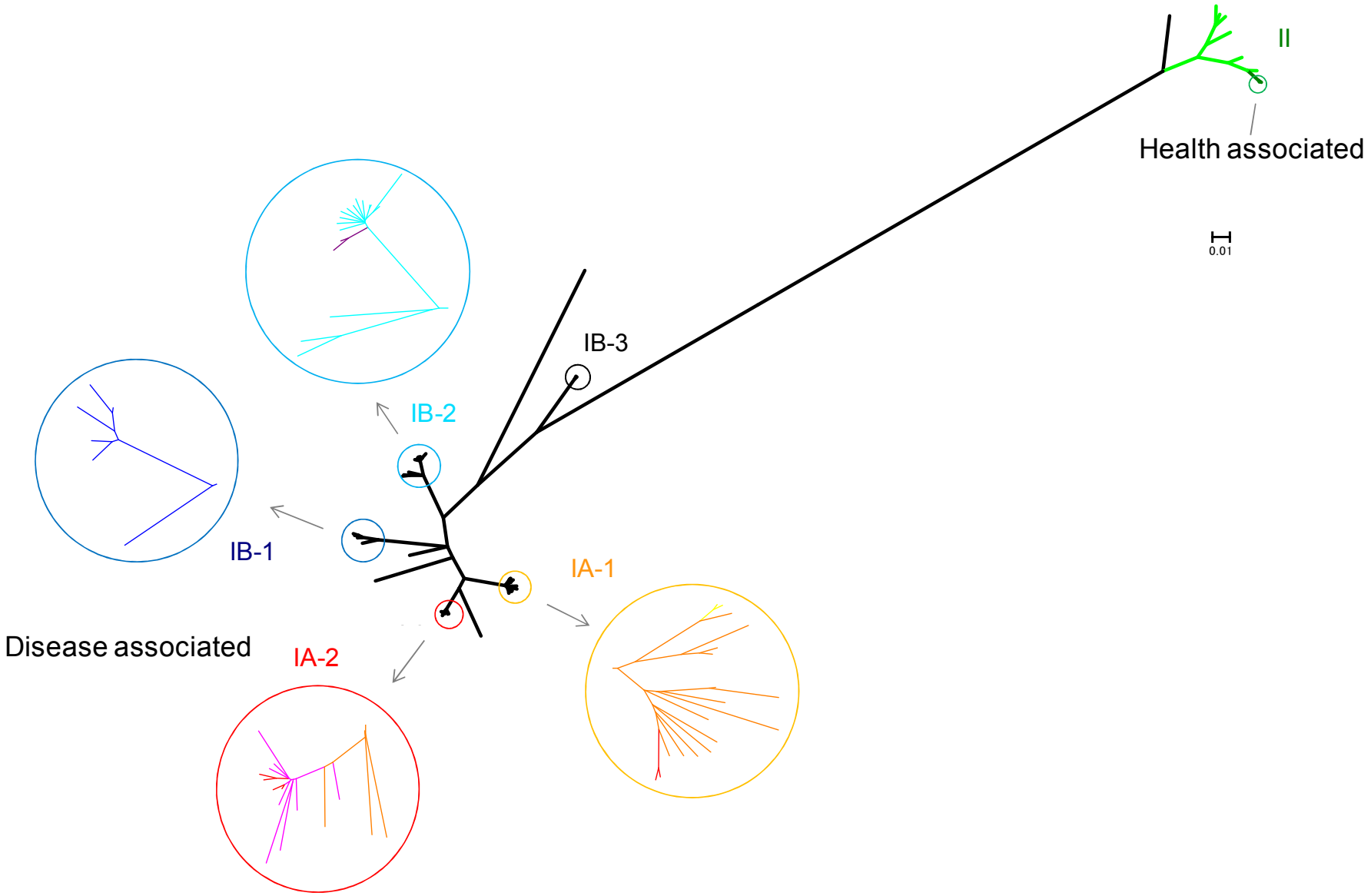

Figure S6. Phylogenetic tree constructed based on the 96,887 SNPs in *P. acnes* core genome shows that the 71 genomes cluster into distinct clades, consistent with *recA* types that have been used to classify *P. acnes* strains. The 16S ribotypes of the genomes represent the relationship of the lineages to a large extent. At one end of the tree, clades IA-2 and IB-1 mainly consist of the ribotypes enriched in acne, and at the other end of the tree, RT6 in clade II was mainly found in healthy subjects. Bootstrap test with 1,000 replicates were performed. The distances between the branches were calculated based on the SNPs in the core genome and do not represent the non-core regions of each genome. The enlarged branches were colored according to the 16S ribotypes as shown in Figure 3.

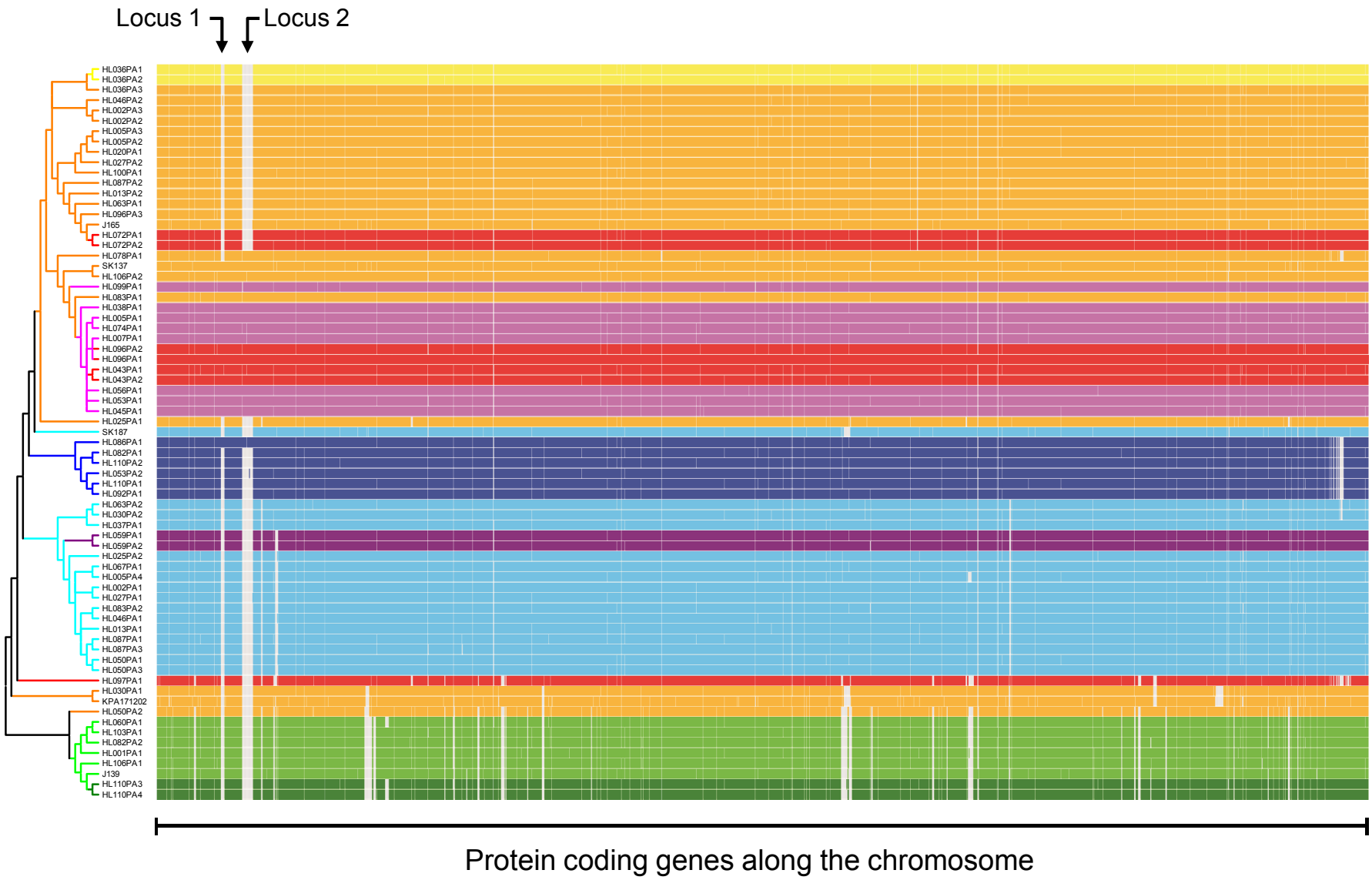

Figure S7. Genome comparison of 71 *P. acnes* strains shows that the genomes of RT4 and RT5 are distinct from others (extension of Figure 3). All the predicted open reading frames (ORFs) encoded on the chromosome are shown. Each row represents a *P. acnes* genome colored according to the ribotypes. Rows are ordered by the phylogeny calculated based on the SNPs in *P. acnes* core genome. Only the topology is shown. Columns represent ORFs in the genomes and are ordered by their positions along the finished genome HL096PA1.

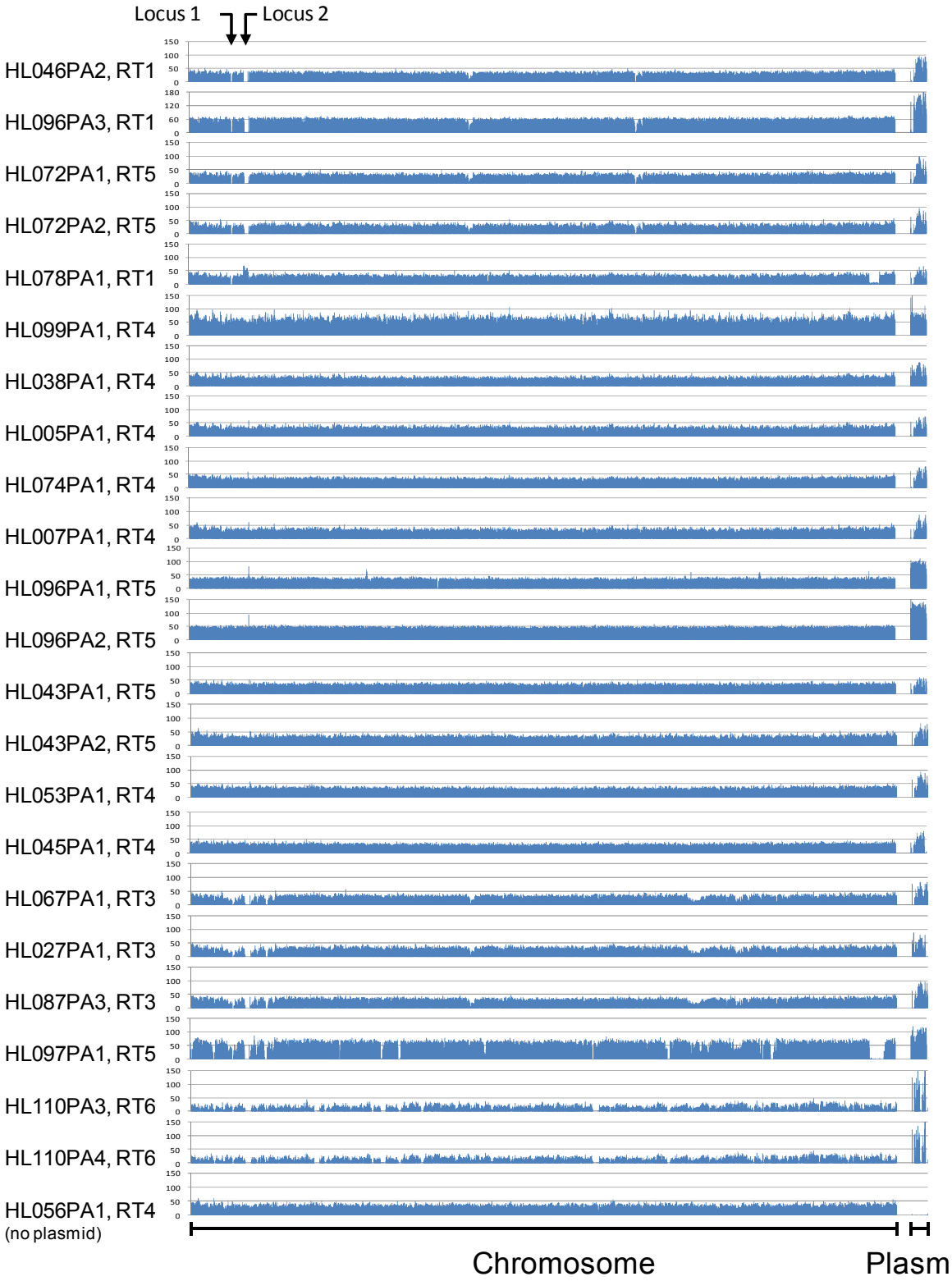

Figure S8. Sequence coverage comparison between the chromosome and the plasmid region in all genomes harboring a putative plasmid shows that the copy number of plasmid ranges from 1 to 3 per genome. X-axis represents the DNA sequences along the chromosome based on the coordinates of the finished genome HL096PA1, followed by plasmid sequences. Y-axis represents the sequence coverage. The genomes were in the same order as in Figure 3, except HL056PA1 (as a negative control).

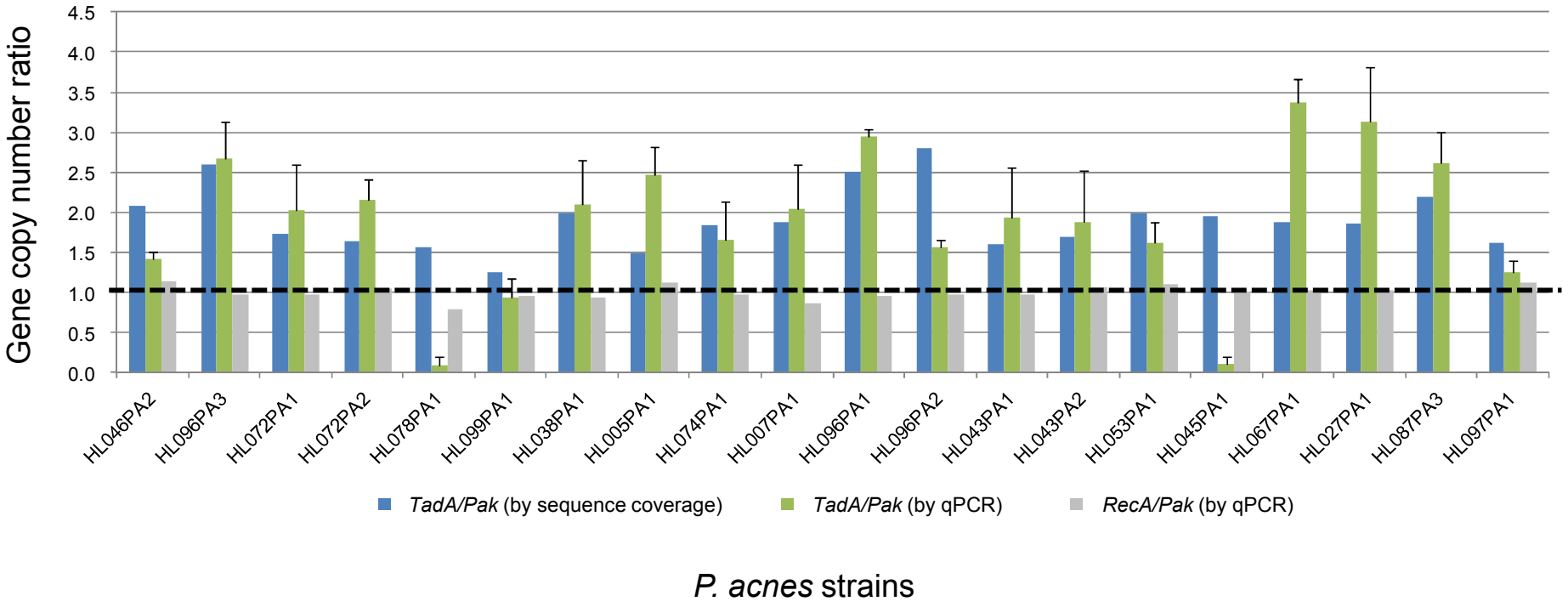

Figure S9. Quantitative PCR (qPCR) confirmed that the copy number of plasmid in each genome is 1-3 as predicted from sequence coverage comparison. *Pak* and *RecA* are housekeeping genes located on the chromosome and *TadA* is a conserved gene in the Tad locus located on the plasmid. The copy number ratio between *TadA* and *Pak* ranges from 1 to 3 in genomes, while the ratio between *RecA* and *Pak* is 1 in all the genomes. The *TadA* gene in HL078PA1 and HL045PA1 had amplification in late cycles in qPCR, thus the copy numbers could not be analysed correctly. Conventional PCR confirmed the amplification of *TadA* in these two strains, while other strains without the plasmid showed no amplification (data not shown).

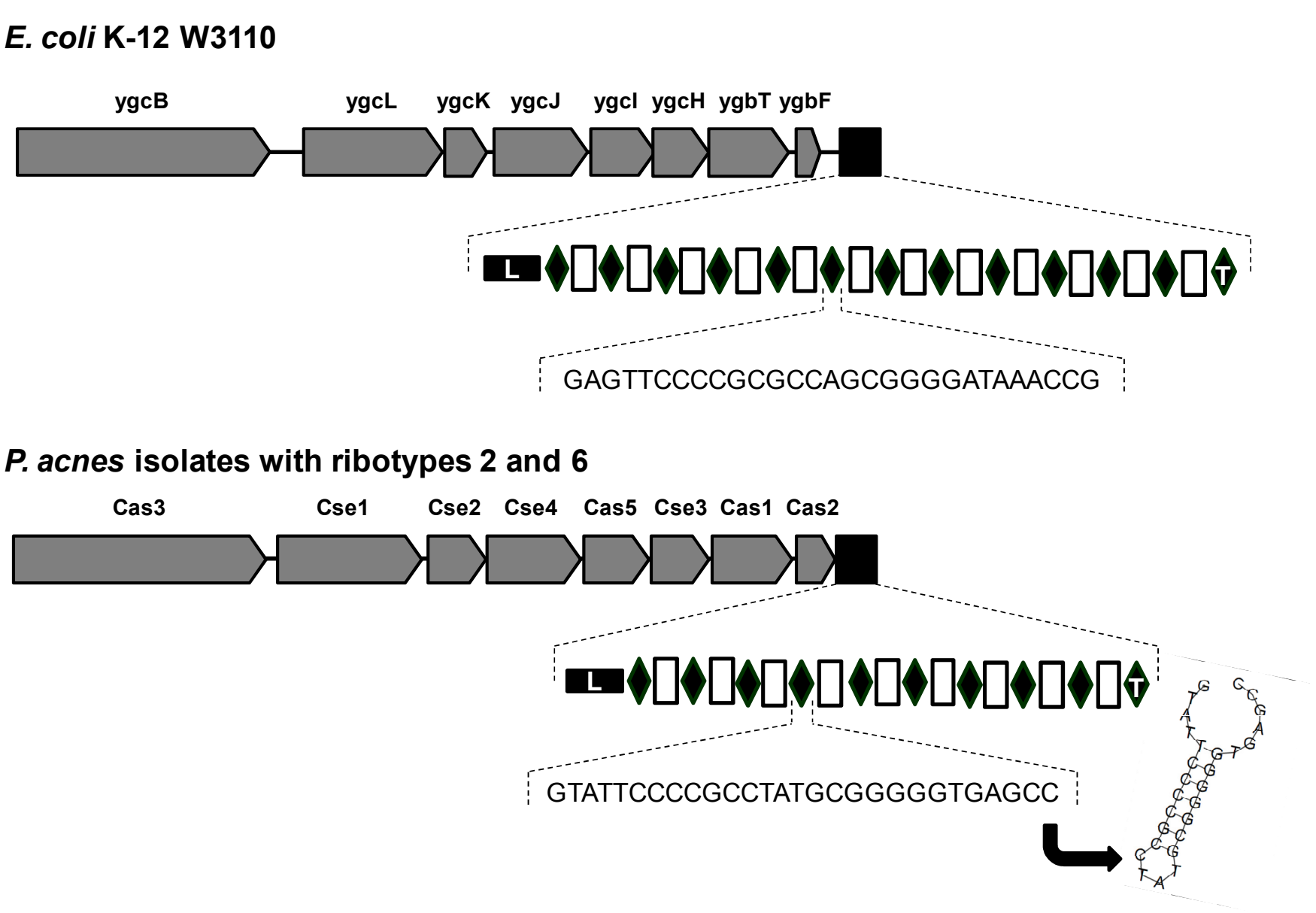

Figure S10. Comparison of the CRISPR/*cas* systems in *P. acnes* and *E. coli*. All the *P. acnes* CRISPR/*cas* systems found in isolates of RT2 and RT6 are homologous to the CRISPR systems in *E. coli* and *Streptococcus thermophilus* CRISPR4 (not shown).