ARTICLE

doi:10.1038/nature13786

Biogeography and individuality shape function in the human skin metagenome

Julia Oh[1], Allyson L. Byrd[1], Clay Deming[1], Sean Conlan[1], NISC Comparative Sequencing Program†, Heidi H. Kong[2]* & Julia A. Segre[1]*

The varied topography of human skin offers a unique opportunity to study how the body's microenvironments influence the functional and taxonomic composition of microbial communities. Phylogenetic marker gene–based studies have identified many bacteria and fungi that colonize distinct skin niches. Here metagenomic analyses of diverse body sites in healthy humans demonstrate that local biogeography and strong individuality define the skin microbiome. We developed a relational analysis of bacterial, fungal and viral communities, which showed not only site specificity but also individual signatures. We further identified strain–level variation of dominant species as heterogeneous and multiphyletic. Reference–free analyses captured the uncharacterized metagenome through the development of a multi–kingdom gene catalogue, which was used to uncover genetic signatures of species lacking reference genomes. This work is foundational for human disease studies investigating inter–kingdom interactions, metabolic changes and strain tracking, and defines the dual influence of biogeography and individuality on microbial composition and function.

Human skin harbours an abundant microbial ecosystem with bidirectional metabolic exchanges supporting symbiotic and commensal processes. The skin's surface consists of diverse microenvironments with distinct pH, temperature, moisture, sebum content and topography[1]. These niche-specific physiological differences influence the resident bacteria[2,3] and fungi[4]; oily surfaces like the forehead support lipophilic bacteria that differ from dry, low biomass sites like the forearm. In turn, microbial sensing and signalling mechanisms, metabolic pathways, or immunogenic features are likely to exhibit site-specificity to sustain host interactions. Similar to the distribution of skin microbes, skin disorders often present in a site-specific manner, such as atopic dermatitis (eczema) in arm and leg creases or psoriasis on the elbows and knees. Inter-kingdom and inter-species microbial interactions may exacerbate disease severity[5] or facilitate transitions from opportunistic to pathogenic. Although skin physiology is a dominant force, individuals retain unique elements of microbial profile and community organization. Here, we explore the complex skin microbial biogeography, integrating broad physiological characteristics with individual discriminatory attributes.

Studies based on phylogenetic marker genes (for example, bacterial 16S ribosomal RNA gene or fungal internal transcribed spacer (ITS) regions) have studied core taxonomic characteristics of different skin sites and disease states. However, such approaches survey kingdoms in isolation and provide limited information into an ecosystem's functionality. Metagenomic shotgun sequencing interrogates the full complement of DNA present in a sample, enabling characterization of both a community's functional capacity and genomes for which no targeted amplicon strategies exist. Several large-scale studies have used metagenomics to examine bacterial or viral communities of the healthy gut and other body sites[6–8], or taxonomic and functional differences in type 2 diabetes[9,10]. To date, a systematic metagenomic investigation of human skin is lacking. The physiological heterogeneity and variable microbial biomass of the skin pose unique technical and analytical challenges for metagenomic studies. Each site on the human skin is constrained by ecological properties such as host microenvironment, yet possesses a distinct biogeography that significantly influences microbial diversity, composition and biomass[2–4,11].

We present the first systematic, multi-site metagenomic study of human skin. We determined the composition and function of the healthy skin microbiome using direct shotgun sequencing of 15 individuals at 18 clinically relevant sites, which included diverse skin microenvironments (dry, moist, sebaceous or toenail, Extended Data Fig. 1). Our dual approach incorporated reference-based and reference-free methods to characterize the metagenome. We present new insights into the larger community of skin microorganisms, including DNA viruses, lower eukaryotes, bacteria and subspecies of dominant bacteria. We defined how functional capacity varies by body site and created a multi-kingdom, skin-associated gene catalogue. Using new analytic approaches, we identified metagenomic 'clusters' representing species without known references. Our study demonstrates that biogeography and individuality significantly shape a community's functional and taxonomic characteristics and provides a framework for human studies investigating inter-kingdom interactions, metabolic changes and pathogen expansion in disease.

Skin sampling and data characteristics

263 specimens were collected from 15 healthy adults (9 males, 6 females) from 18 defined anatomical skin sites (Supplementary Table 1). We modified previous clinical sample acquisition, DNA isolation and library preparation to generate shotgun metagenomic sequence data from skin sites, which varied in biomass and composition. For example, human-derived DNA accounted for 19.4 ± 6.7% to 98.2 ± 0.1% of reads, reflecting the difference between stratified, cornified plantar heel skin and nucleated inner nostril epithelium, respectively (Extended Data Fig. 2a). Microbial sequencing yields and estimated coverage also varied with skin physiological features ('microenvironment'), such that low-diversity, higher-biomass sebaceous sites generally achieving greater coverage (maximum 81.0 ± 7.0%) than high-diversity, lower-biomass dry or moist sites (minimum 38.0 ± 5.7%, Extended Data Fig. 2c). We obtained a total of 289 gigabase pairs (Gbp) of non-human, quality filtered Illumina

[1]Translational and Functional Genomics Branch, National Human Genome Research Institute, NIH, Bethesda, Maryland 20892, USA. [2]Dermatology Branch, Center for Cancer Research, National Cancer Institute, NIH, Bethesda, Maryland 20892, USA.

*These authors contributed equally to this work.

†A list of authors and affiliations appears at the end of the paper.

©2014 Macmillan Publishers Limited. All rights reserved

microbial sequence reads (Extended Data Fig. 2a–c, Supplementary Table 1).

Phylogenetic profiles of skin microbes

To explore the relative abundances of skin microbiota across kingdoms, we performed a relational analysis mapping filtered reads to 2,342 bacterial, 389 fungal, 1,375 viral and 67 archaeal genomes. To validate taxonomic assignments, we compared our metagenomic data with 16S and ITS sequencing of the same samples, which showed high concordance (Extended Data Fig. 3, Supplementary Tables 3–5). While recognizing that fungal and viral genomes are more sparsely represented in reference databases, bacteria predominated at most sites (Fig. 1a–c, Extended Data Figs 1, 4a, Supplementary Table 6) and comprised the bulk of phylogenetic diversity with fungi and viruses contributing relatively fewer species. Fungi, primarily *Malassezia globosa* and *M. restricta*, were a lower fraction (3.9 ± 5.0%), except near the ears and forehead, which had a higher fungal presence (external auditory canal, 16.8 ± 5.1%; retroauricular crease 7.5 ± 4.2%; glabella 7.1 ± 4.0%). The feet had low fungal representation (plantar heel, 0.7 ± 0.2%; toenail 0.5 ± 0.3%; toe web 0.3 ± 0.1%), despite high diversity observed in amplicon-based studies. Archaea were nearly absent on skin, but DNA viruses were abundant at specific sites, with marked interpersonal variation. Note, RNA viruses are not interrogated by these methods and probably represent uncharacterized diversity. The nares and adjacent alar crease showed significant viral representation (51.0 ± 11.8% and 54.6 ± 9.3%), compared to 9.9 ± 1.0% at other sites. Interestingly, a few individuals had sites that were dominated by viruses (up to 96%). These 'blooms' contained *Propionibacterium* or *Staphylococcus* bacteriophage and/or potential human viral pathogens (molluscum contagiosum, human papillomavirus, and Merkel cell polyomavirus), although skin sites were free of clinical lesions. Communities were shaped primarily by the microenvironment, in which differential abundance of stereotypical taxa such as *Propionibacterium acnes*, commensal staphylococci, *Corynebacterium* and *Propionibacterium* phage contributed most significantly to variation both between and within individuals (Fig. 1d).

To compare skin with other body sites, we analysed 552 Human Microbiome Project (HMP) metagenomic samples obtained from the anterior nares, posterior fornix (vagina), retroauricular crease, stool, supragingival plaque and tongue dorsum (Fig. 1b, Extended Data Fig. 4b, Supplementary Tables 6, 7)[12]. Our skin samples were similar to those of the HMP in community membership and structure of all kingdoms ($P > 0.05$). However, retroauricular crease samples from our study had greater fungal abundance than HMP (7.5% versus 3.4%), probably reflecting differences in nucleic acid extraction techniques, which we optimized to recover fungal DNA. Fungi were relatively scarce at non-skin sites. Similar to skin sites with phage co-occurring with their host bacteria, *Lactobacillus* phage was observed in the posterior fornix with marked interpersonal variation. Viruses were found in low abundance in the mouth, but *Streptococcus* phage was nearly universal, present in 99.2% of samples (mean abundance 1.2 ± 0.1%). Overall, the human body is rich in both bacterial and non-bacterial taxa, with site-specific fungal enrichment and viral blooms.

Individuality underlies biogeography

Differential manifestations of phenotypes including disease susceptibility, antibiotic response, drug metabolism or even weight gain are likely to be influenced by an individual's exclusive microbial community features. We explored whether we could classify individuals based on unique taxonomic signatures across their body. We used random forests, which incorporates interactions of both rare and abundant taxa, to identify key taxa that might differentiate individuals (Supplementary Table 8). Surprisingly, low-abundance taxa shared across skin sites discriminated individuals (Fig. 2). For example, the strongest discriminatory feature was Merkel cell polyomavirus, present in low abundance at all skin sites within one individual, regardless of site. Several taxa could also be discriminatory on an individual level; *Gardnerella vaginalis* and *Streptococcus pyogenes* were host-specific across all skin sites, in addition to taxa that probably represent transient populations (for example, Acheta domesticus densovirus).

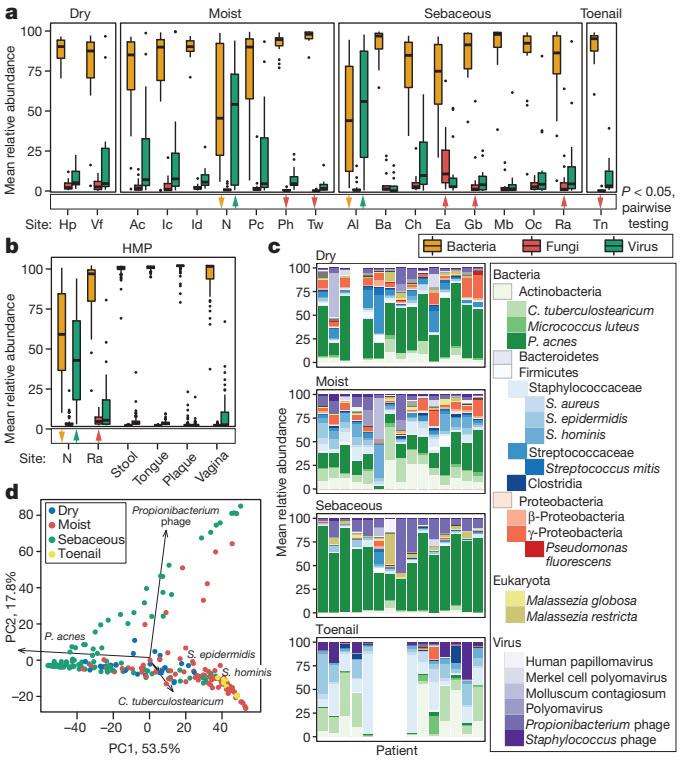


Figure 1 | Multi-kingdom relative abundances are strongly shaped by skin microenvironment. **a**, Boxplots of mean relative abundance of different kingdoms by site. Black lines indicate median; boxes first and third quartiles. Triangles indicate significance (adjusted $P < 0.05$, Kruskal–Wallis post-hoc test) for over- (up) or under- (down) representation in a majority of pairwise comparisons between sites. Hp (hypothenar palm), Vf (volar forearm), Ac (antecubital crease), Ic (inguinal crease), Id (interdigital web space), N (nares), Pc (popliteal crease), Ph (plantar heel), Tw (toeweb space), Al (alar crease), Ba (back), Ch (cheek), Ea (external auditory canal), Gb (glabella), Mb (manubrium), Oc (occiput), Ra (retroauricular crease), Tn (toenail). **b**, Kingdoms in HMP body sites. **c**, Consensus relative abundance plots of major skin taxa by microenvironment. *C., Corynebacterium; P., Propionibacterium; S., Staphylococcus.* **d**, Communities cluster primarily by microenvironment with sebaceous regions most distinct in principal components (PC) analysis.

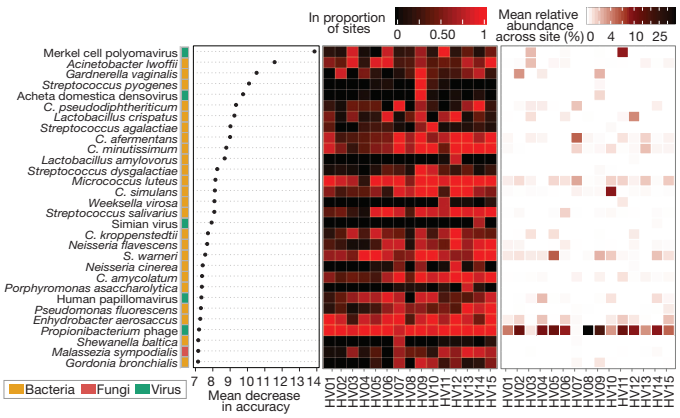


Figure 2 | Individual-specific signatures are typically low abundance but shared across most sites. Left, variable importance plot of most discriminatory taxa from random forests analysis. For each individual, centre, proportion of the 18 sites in which each taxa is present, and right, mean relative abundance of that taxa across sites.

©2014 Macmillan Publishers Limited. All rights reserved

With our multi-kingdom taxonomy, we could differentiate our 15 individuals with >80% accuracy (19.3% error). The increased error estimates based upon kingdom-specific analyses (21.8%, bacteria; 74%, fungi; 41.2%, viruses) underscores the importance of understanding the full phylogenetic diversity of a community. Such approaches are relevant in identifying discriminatory features in disease states or assessing longitudinal community stability in which individuals may be identifiable by microbial features. While site-specificity serves as an overarching constraint on community composition, we observed a remarkable range of individual signatures within the skin biogeography.

Strain heterogeneity in skin symbionts

We further explored individual signatures by examining strain-level variation; subspecies within a clade can possess different properties of transmissibility, virulence, antibiotic resistance, or metabolism[13]. To investigate strain-level heterogeneity, we focused on two common skin commensals with well-documented sequence variation, *P. acnes* and *S. epidermidis*. Using a reference-based approach that leveraged both single nucleotide polymorphisms and larger variants (Extended Data Fig. 5, Supplementary Tables 2, 9–12), we identified phylogenetically 'most similar' strains based on differentiating genomic features. To reduce false discovery, we characterized both strain and a more conservative subtype level that represents phylogenetically similar strain groups (Fig. 3a, b, Extended Data Figs 5, 6).

Given the extensive strain-level diversity observed for both species, our results suggest that individual and microenvironment differentially shape subspecies variation. *P. acnes* strains were more individual- than site-specific (Fig. 3c, e); 11/12 *P. acnes* subtypes were differentially abundant between individuals whereas only one differed between microenvironments (Fig. 3g). In contrast, *S. epidermidis* strains were significantly more site-driven with diminished inter-individual variation (Fig. 3d, f); nearly all subtypes were differentially abundant between sites (Fig. 3h) with subtype 'B' particularly dominant in the foot and toenail (Fig. 3b). These results strongly suggest that *P. acnes* and *S. epidermidis* communities are heterogeneous and multiphyletic, properties that probably vary by species and niche. Further analyses of this resolution will be powerful in determining genetic variation across time, topography and disease. In summary, our systematic analysis of microbial community composition has described a remarkable dynamism spanning inter-kingdom partnerships down to sub-species variability, characteristics that are driven both by broad ecological constraints and an individual's unique carriage.

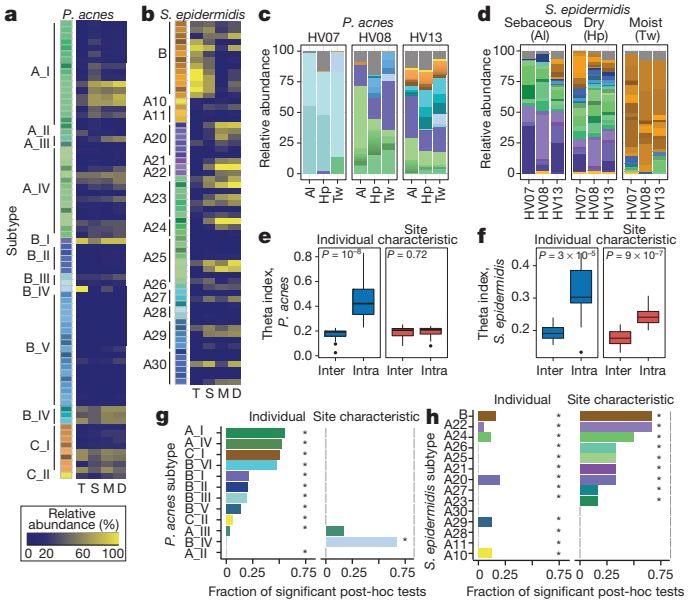


Figure 3 | *Propionibacterium acnes* and *Staphylococcus epidermidis* are heterogeneous and multiphyletic at the strain level. a, b, Reference genomes used for *P. acnes* (**a**) and *S. epidermidis* (**b**). Leftmost bar shows subtypes (phylogenetically similar genomes) as colour groups. Adjacent heat map shows mean relative abundance by skin microenvironment. D, dry; M, moist; S, sebaceous; T, toenail. **c, d**, Select relative abundance plots; strain colours as in **a**, **b. e, f**, *P. acnes* subtypes differ more significantly between individuals than skin microenvironment with the converse observed for *S. epidermidis*. Boxplots of Yue–Clayton theta indices calculate similarity between ('inter') or within ('intra') individuals/microenvironments ($\theta = 1$ means identical). Black lines indicate median, boxes show first and third quartiles. *P* value, Wilcoxon rank-sum test. **g, h**, Bar charts show *P. acnes* and *S. epidermidis* subtypes that differ by microenvironment or individual. Length of bar represents the fraction of post-hoc tests significant for each comparison; 105 comparisons for individual; 6 for microenvironment. *$P < 0.05$, adjusted Kruskal–Wallis test.

Biogeography shapes functional diversity

While taxonomy yields important insight into community organization, metagenomics also enables analysis of a community's collective functional potential. Whereas previous studies reported that most metabolic pathways are evenly distributed across body sites[12], we observed a modest decrease in metabolic diversity that occurred in tandem with lower taxonomic diversity in sebaceous sites (Fig. 4a). Investigating this concept of core functionality, we determined that only 30% (44/148) of modules were 'core' irrespective of site (present in ≥ 2/3 samples), representing processes essential to microbial growth and metabolism (Extended Data Fig. 7, Supplementary Tables 13–15). Extensive variability was observed within subclasses of major pathways, particularly transport systems (sulphate, glutamate, aspartame, L- or branched amino acids and sorbitol) and putrescine/spermidine biosynthesis and transport, which were typically absent in sebaceous regions, attesting to the chemical diversity likely to be present at higher-complexity sites. Conversely, most eukaryotic pathways were more prevalent in sebaceous sites (cell cycle, DNA replication, transcription, translation, protein degradation and vitamin D2 biosynthesis, a fungi-produced phytonutrient). Thus, although a strong functional core exists, this core metagenome can vary tremendously, reflecting functional diversification of skin microenvironments. Future studies with transcriptional profiling will probably reveal additional functional variance *in vivo*.

Modules present across all sites were typically low abundance and associated with uncharacterized biomolecular functions and metabolism. 88% of modules were differentially abundant in at least one microenvironment (adjusted $P < 0.05$, Supplementary Tables 13, 15), suggesting that functional capacity is driven primarily by biogeography. Principal components identified modules that discriminate microenvironments (Fig. 4c). Sebaceous sites (PC1) are distinguished by overrepresentation of glycolysis and related components (ATP and GTP generation) and NADH dehydrogenase I. Toenail samples differed primarily by the presence of different energy production components, such as conversion of oxaloacetate to fructose-6-phosphate, and ATPase and ATP synthase. Dry sites were characterized by the presence of citrate cycle modules. Covariance analysis imputing pathway abundance to select species suggested that *P. acnes* and *M. restricta* are likely candidates to drive some niche-specific metabolism, given their abundance in sebaceous sites (Fig. 4d, Extended Data Fig. 8).

With increasing concerns of antibiotic-resistant microorganisms, we explored the reservoir of antibiotic resistance genes in the skin. Although skin is physically compartmentalized from other body sites, cross-inoculation remains a risk factor. For example, the nares can harbour methicillin-resistant *Staphylococcus aureus* (MRSA)[14] underlying skin and soft tissue infections. Strain crosstalk between oral, lung and skin sites may underlie recurrent infections in immunocompromised patients[15]. Here, we identified presence/absence of well-characterized resistance gene families as pioneered for the gut[16] and soil[17]. We observed significant variability across individuals and resistance types (Extended Data Fig. 9, Supplementary Table 16). Certain antibiotic classes were highly host-specific, such as multi-antimicrobial extrusion (MATE) efflux pumps (Fig. 4e). In an example of site-specific dominance, lincosamide

©2014 Macmillan Publishers Limited. All rights reserved

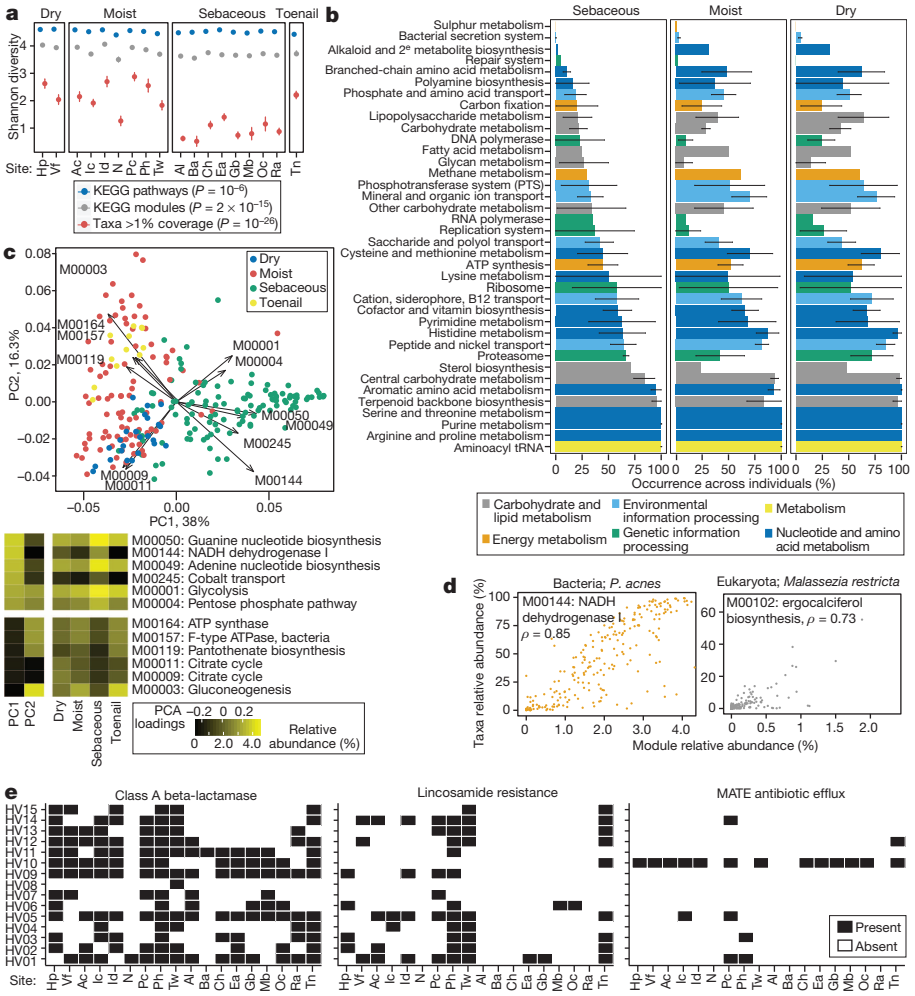


Figure 4 | Functional capacity varies by microenvironment. **a**, Shannon diversity of functional pathways and taxonomy by site; *P* value, Kruskal–Wallis test between microenvironments. Error bars, standard error of the mean. **b**, Microenvironments possess different core modules; 'core' means occurrence in more than 2/3 of samples. Error bars show variation within a class of modules (full version in Extended Data) that may arise from a unique specialization for that microenvironment. **c**, PCA shows clustering by microenvironment, with strong separation of sebaceous, dry and toenail modules. Heat maps: left, loadings for the first two PCs; right, mean relative abundances for modules with the greatest variation by microenvironment. **d**, A module's taxonomic origin can be imputed by Spearman correlation (ρ; adjusted $P \leq 2 \times 10^{-16}$) with *P. acnes* and *M. restricta* relative abundances. **e**, Presence of select antibiotic resistance gene families by individual and site.

resistance showed significant representation in three foot sites but was generally absent in sebaceous regions. Finally, certain families were broadly represented across samples, such as class A beta-lactamases, rRNA methyltransferases, efflux mechanisms, or quinolone resistance. Thus, carriage of antibiotic resistance families demonstrated both site- and individual-specificity, although we note that resistance activity may differ *in vivo*.

Insights into microbial dark matter

Our reference-based analysis showed a large variable fraction of reads (2–96%) unmapped to reference genomes, most frequently originating from decreased bacterial assignments (Supplementary Table 6, Extended Data Fig. 10a). Such uncharacterized sequences likely originate from both taxa with no representative reference and intraspecies pangenomic variation, which can represent significant gene content[13]. Using reference-free methods to capture this 'dark matter' of the skin metagenome, we created a skin gene catalogue that we then used to identify previously uncharacterized taxa in the skin. Such resources will be invaluable for downstream analyses, enabling *in silico* prediction and synthesis of genes and pathways that are over- or underrepresented in, for example, disease states.

The inherent variation in skin community complexity and human DNA admixture presents new challenges in reference-free methodologies; variable microbial load and taxonomic diversity across sites affect sequencing depth and coverage. To account for this variability, we devised an adaptive and iterative strategy (Extended Data Fig. 10b, c) that optimizes assembly on a per-sample basis (Fig. 5a, Supplementary Table 17). We then established the first multi-kingdom skin microbial gene catalogue using both fungal and bacterial prediction models. Of 5.92 million open reading frames (ORFs), 75.7% could be reconstructed as bacterial and 15.9% as eukaryotic, consistent with our taxonomic analyses (Fig. 5b, Supplementary Table 18). Large numbers of KEGG (Kyoto Encyclopedia of Genes and Genomes) hypothetical genes (25.7% of bacterial, 48.3% of eukaryotic) are likely to represent pangenomic loci of characterized taxonomies, for example, *P. acnes* and *M. globosa*, based on association without pathway annotation. In support of their authenticity, ORFs with no identifiable homologues (7.9%) were typically longer than classified ORFs (Fig. 5b, inset). Less than 1% of ORFs were assigned to Archaea and viruses (which require unique prediction models), possibly reflecting integrative viruses or overlap in gene prediction models.

Finally, we used our gene catalogue to identify microbial species and pangenomic content independently of reference genomes. Under the assumption that genes from one genome covary in abundance across samples owing to physical linkage, we created metagenomic 'clusters'[9,10] by correlating gene abundances across samples (Supplementary Table 18).

©2014 Macmillan Publishers Limited. All rights reserved

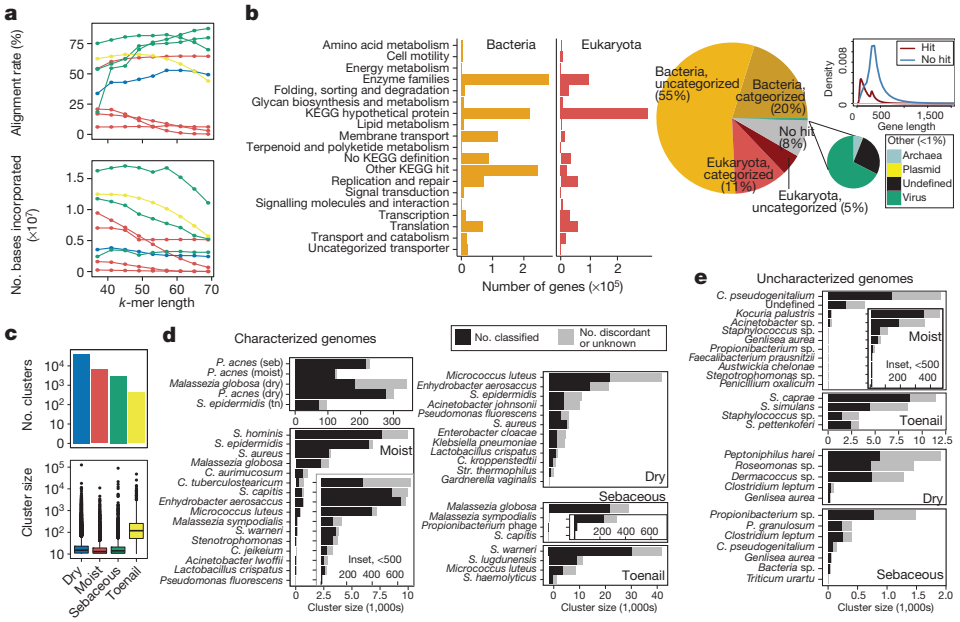


Figure 5 | Reconstruction of metagenomic dark matter with reference-free methods. **a**, Per-sample iterative assembly with variable *k*-mers (nucleotide words of length *k*) optimizes assembly quality as assessed by metrics such as % reads mapping back to assembly (left) and the number of bases incorporated (right). Colours are as in **c**. **b**, Skin gene catalogue was mapped to the NCBI non-redundant (nr) database and KEGG to identify kingdom and functional category. Density plot compares length of genes with and without homology; gene length was typically larger for unmapped genes. **c**, Metagenomic clusters represent genes that covary in abundance across samples within a microenvironment; boxplots show cluster sizes; histograms show number of clusters ($\log_{10}$ scale). **d**, A lowest common ancestor (LCA) was assigned to a cluster with >50% consensus taxonomy. Bar length indicates the total number of 'genes' in a cluster; black represents the number of genes mapping to the LCA. Grey represents ambiguous or unannotated genes. 'Characterized' indicates that a reference genome exists for that species; for **e**, 'Uncharacterized genomes', no reference exists. Seb, sebaceous; tn, toenail.

Most resultant clusters were relatively small, but others contained hundreds of thousands of predicted ORFs, which probably represent both genes and gene fragments. High-complexity dry sites had the most clusters and whereas toenails had the fewest, their median gene recruitment was significantly larger (Fig. 5c). To strengthen the reliability of our metagenomic clusters, we required clusters to share >50% consensus taxonomy at the species level and uncovered large clusters of fungi, bacteria and viruses (Fig. 5d). *M. globosa*, *P. acnes* and *S. epidermidis* had very large clusters, consistent with their high abundance in skin. In addition to clusters representing referenced genomes, we also identified multiple uncharacterized genomes (Fig. 5e), most commonly species of common genera in the skin, including *Corynebacterium*, *Propionibacterium* and *Staphylococcus*. In summary, leveraging reference-free approaches, we identified previously undefined elements of the human skin microbiota. While dominant species or pathogens are targeted for sequencing, metagenomic studies reveal remarkable additional taxonomic and thereby functional diversity.

Conclusions

The healthy skin metagenome possesses surprising taxonomic and functional diversity dependent on both biogeography and individuality. In contrast to other body sites like the gut, the skin has markedly higher viral and fungal representation. For most individuals, common skin species exist as a heterogeneous mix of strains, raising questions of whether transitions to a pathogenic state are mono- or multiphyletic, and how strain heterogeneity affects disease incidence or severity. Significant decreases in community diversity are a hallmark of a disease state[18]; whether such shifts occur at all taxonomic levels down to the subspecies awaits investigation. Our reference-based toolkit for multi-kingdom analyses and strain differentiation is broadly applicable to ecosystems with a well-characterized sequence space. Our reference-free resources, generated by adaptive assemblies, enable interrogation of the significant uncharacterized proportion of the metagenome, even identifying species without reference genomes.

From a therapeutic perspective, the metagenome represents a rich resource for synthetic biology approaches to modify and transplant endogenous elements to other communities. Studies of metabolic capacity, pathogenicity islands and virulence genes in disease states, with our catalogue from healthy skin, will uncover biomarkers associated with transmission, recurrence and severity of disease. Finally, characterization and tracking of surprisingly pervasive antibiotic resistance elements will remain clinically relevant, as skin sites can serve as a taxonomic and genetic reservoir for pathogens. We envision a new therapeutic landscape leveraging unique metagenomic profiles with tailored clinical interventions that reshape our microbial communities.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 3 June; accepted 21 August 2014.

1. Grice, E. A. & Segre, J. A. The skin microbiome. *Nature Rev. Microbiol.* **9,** 244–253 (2011).
2. Grice, E. A. *et al.* Topographical and temporal diversity of the human skin microbiome. *Science* **324,** 1190–1192 (2009).
3. Costello, E. K. *et al.* Bacterial community variation in human body habitats across space and time. *Science* **326,** 1694–1697 (2009).
4. Findley, K. *et al.* Topographic diversity of fungal and bacterial communities in human skin. *Nature* **498,** 367–370 (2013).
5. Peters, B. M. & Noverr, M. C. *Candida albicans-Staphylococcus aureus* polymicrobial peritonitis modulates host innate immunity. *Infect. Immun.* **81,** 2178–2189 (2013).
6. Arumugam, M. *et al.* Enterotypes of the human gut microbiome. *Nature* **473,** 174–180 (2011).
7. De Vlaminck, I. *et al.* Temporal response of the human virome to immunosuppression and antiviral therapy. *Cell* **155,** 1178–1187 (2013).
8. Handley, S. A. *et al.* Pathogenic simian immunodeficiency virus infection is associated with expansion of the enteric virome. *Cell* **151,** 253–266 (2012).
9. Qin, J. *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490,** 55–60 (2012).
10. Karlsson, F. H. *et al.* Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature* **498,** 99–103 (2013).
11. Grice, E. A. *et al.* A diversity profile of the human skin microbiota. *Genome Res.* (2008).

©2014 Macmillan Publishers Limited. All rights reserved

12. The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486,** 207–214 (2012).
13. Tettelin, H. *et al.* Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial 'pan-genome'. *Proc. Natl Acad. Sci. USA* **102,** 13950–13955 (2005).
14. von Eiff, C., Becker, K., Machka, K., Stammer, H. & Peters, G. Nasal carriage as a source of *Staphylococcus aureus* bacteremia. *N. Engl. J. Med.* **344,** 11–16 (2001).
15. Oh, J. *et al.* The altered landscape of the human skin microbiome in patients with primary immunodeficiencies. *Genome Res.* (2013).
16. Sommer, M. O. A., Dantas, G. & Church, G. M. Functional characterization of the antibiotic resistance reservoir in the human microflora. *Science* **325,** 1128–1131 (2009).
17. Forsberg, K. J. *et al.* The shared antibiotic resistome of soil bacteria and human pathogens. *Science* **337,** 1107–1111 (2012).
18. Kong, H. H. *et al.* Temporal shifts in the skin microbiome associated with disease flares and treatment in children with atopic dermatitis. *Genome Res.* **22,** 850–859 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank D. Schoenfeld, A. Pradhan, M. Park and G. Bouffard for their efforts. We also thank members of the Segre laboratory and M. C. Udey for their discussions. This work was supported by National Institutes of Health (NIH) NHGRI and NCI Intramural Research Programs and in part by 1K99AR059222 (H.H.K.). This study used the high-performance computational capabilities of the NIH Biowulf Linux cluster. Sequencing was funded by grants from the National Institutes of Health (1UH2AR057504-01 and 4UH3AR057504-02).

Author Contributions J.O., H.H.K. and J.A.S. designed the study. H.H.K. collected patient samples. C.D. prepared the clinical samples for sequencing, which was carried out by the members of the NIH Intramural Sequencing Center Comparative Sequencing program. J.O., A.L.B. and S.C. analysed sequence data. J.O., H.H.K. and J.A.S. drafted the manuscript. All authors read and approved the final version of the manuscript.

Author Information Data deposition is with the SRA and all sequences can be accessed under BioProject 46333. Human subject clinical data are deposited with dbGaP phs000266. Analysis workflow is available at https://github.com/julia0h/skinmetagenome.git. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to H.H.K. (konghe@mail.nih.gov) or to J.A.S. (jsegre@mail.nih.gov).

NISC Comparative Sequencing Program

Betty Barnabas[1], Robert Blakesley[1], Gerry Bouffard[1], Shelise Brooks[1], Holly Coleman[1], Mila Dekhtyar[1], Michael Gregory[1], Xiaobin Guan[1], Jyoti Gupta[1], Joel Han[1], Shi-ling Ho[1], Richelle Legaspi[1], Quino Maduro[1], Cathy Masiello[1], Baishali Maskeri[1], Jenny McDowell[1], Casandra Montemayor[1], James Mullikin[1], Morgan Park[1], Nancy Riebow[1], Karen Schandler[1], Brian Schmidt[1], Christina Sison[1], Mal Stantripop[1], James Thomas[1], Pamela Thomas[1], Meg Vemulapalli[1] & Alice Young[1]

[1]NIH Intramural Sequencing Center, National Human Genome Research Institute, Bethesda, Maryland 20852, USA.

©2014 Macmillan Publishers Limited. All rights reserved

METHODS

Subject recruitment and sampling. Healthy male and female volunteers of 23 to 39 years of age without chronic skin diseases were recruited from the Washington DC metropolitan region, USA, between June 2011 and May 2013. This natural history study was approved by the Institutional Review Board of the National Human Genome Research Institute (http://www.clinicaltrials.gov/ct2/show/NCT00605878). All subjects provided written informed consent before participation. Subjects provided medical and medication history and underwent a physical examination. Exclusion criteria included history of chronic medical conditions, including chronic dermatologic diseases, and use of antimicrobial medication (antibiotic or antifungal treatments) 1 year before sampling. Cleansing with only non-antibacterial cleansers was allowed during the 7 days before sample collection. To maximize microbial load, no bathing, shampooing or moisturizing was permitted within 24 h of sample collection[15], which we have previously observed produces no discernible shifts in the overall diversity and structures of skin communities.

18 skin sites representing diverse physiological characteristics and sites of predilection for specific dermatologic diseases were sampled: moist (antecubital crease, inguinal crease, interdigital web space, nares, popliteal crease, plantar heel, toe web space), dry (hypothenar palm, volar forearm), sebaceous (alar crease, back, cheek, external auditory canal, glabella, manubrium, occiput, retroauricular crease), and toenail (Extended Data Fig. 1). Additional unmatched samples excluded from statistical analyses included samples extracted with the NEBNext Microbiome DNA Enrichment Kit (NEB), axillary vault (moist), bacterial and fungal mock communities[19], samples that were whole-genome-amplified before library creation, and samples from disease patients (STAT3-hyper IgE, SH). To obtain sufficient DNA from defined anatomical skin sites with low and variable microbial biomass, we modified clinical sample acquisition methods using a swab-scrape-swab procedure, in which a defined anatomical skin area was swabbed with a swab (Catch-All Sample Collection Swabs, Epicentre) pre-moistened with yeast cell lysis buffer (MasterPure Yeast DNA Purification Kit, Epicentre), scraped via sterile disposable surgical blade, and swabbed with the same swab again. Residuals from the scalpel and swab were collected into lysis buffer. Nares and external auditory canal sites were sampled via swabbing with pre-moistened swabs that were then placed into lysis buffer. Toenail samples were cut with sterilized nail clippers and placed into lysis buffer. All samples were stored at −80 °C until extraction. Samples were then incubated in yeast cell lysis buffer (MasterPure Yeast DNA Purification Kit, Epicentre) and treated with Readylyse (Epicentre) for 30 min at 37 °C, then mechanically disrupted using 5 mm stainless steel beads (Qiagen) in a Tissuelyser (Qiagen) for 2 min, 30 Hz. Samples were incubated for 30 min at 65 °C, placed on ice for 5 min, and debris spun down after treatment with MPC protein precipitation reagent. Samples were combined with 350 µl of 100% ethanol and column purified using the Invitrogen PureLink Genomic DNA. Finally, samples were eluted in 30 µl of water (MoBio).

Sample sequencing. Because of low bioburden typical of skin samples, Illumina libraries were created using Nextera library preparation. Briefly, 1–50 ng of extracted DNA was used as input into the transposome fragmentation step. Manufacturer's protocol was followed with the exception of using 10 cycles of PCR. 1–10 ng of extracted DNA was used as input according to manufacturers' recommended protocol (Qiagen Repli-G Mini). Libraries were then sequenced with 2 × 100 bp paired end reads on an Illumina HiSeq at the NIH Intramural Sequencing Center with a target of 15 or 50 million clusters, depending on the microbial diversity of that site and the human DNA admixture. To ascertain that the Nextera approach resulted in minimal sequencing bias, we calculated expected distribution of breaks as represented by the expected frequency of pentamers starting a read for four different genomes, with high correlation with a standard Illumina prep. Moreover, expected versus observed frequencies of species in sequencing of the bacterial mock community were closely matched.

In total, we obtained 7.4 billion reads (289 Gbp) of non-human, quality-filtered paired-end and singleton reads (median 9.5 million reads (893 Mbp) per sample, mean insert size 145 ± 2 bp). Sequencing data were processed to remove low quality reads and any read pairs in which at least one read mapped to the human hg19 human reference. Nextera adaptor sequences were trimmed, if necessary, using Crossmatch 1.090518 (http://www.phrap.org) and custom scripts. Bases with quality score below 20 were trimmed, and reads <50 bp length were removed. Sequencing depth varied by site with estimated *k*-mer coverage ranging from 38.0 ± 5.7% to 81.0 ± 7.0% based on the accumulation of unique DNA substrings, or *k*-mers. Rarefaction curves were generated using Khmer v0.7.1[20] with a 20× coverage cut-off. Briefly, reads were split into *k*-mers, compared to a *k*-mer coverage table and kept only if the median *k*-mer coverage was below the cutoff. Resulting curves showed the coverage of *k*-mer space as a function of sequencing effort. Median insert size was estimated from a subsample of paired reads that match hg19. Post sequence quality control, samples with >20 million reads remaining were subsampled to 10 million paired end reads, and singletons were discarded. HMP data from the anterior nares, retroauricular crease, stool, posterior fornix, tongue dorsum and supragingival plaque were obtained from ftp://public-ftp.hmpdacc.org and subsampled to 1 million reads for taxonomic comparisons.

Amplicon processing. To validate our taxonomic assignments, normalize for sequencing levels, and reduce false positives, we also compared our results with matched bacterial 16S and fungal ITS amplicon sequencing. 159 matched 16S rRNA and 92 matched ITS1 samples were processed as previously described[15]. Briefly, the V1-V3 region of the 16S rRNA gene was amplified using the barcoded 27F and 534R and the ITS1 with 18SF and 5.8S-1R primers. Amplicon libraries were sequenced on a 454 GS FLX (Roche) instrument using titanium chemistry. 16S rRNA and ITS1 samples were processed using the mothur pipeline[21] as previously described[15]. Briefly, 454 flow gram data were denoised, error-trimmed, and chimaeric sequences removed. 16S sequences were classified using RDP training set 9 and ITS1 using a custom ITS1 database[4]. *Staphylococcus* and *Malassezia* genera were classified to the species level using pplacer[22] with custom databases.

Reference-based taxonomic and functional classification. We compiled a list of complete and draft microbial reference genomes of 2,342 bacterial, 389 fungal, 1,375 viral, and 67 archaeal genomes from the National Center for Biological Information (NCBI, http://www.ncbi.nlm.nih.gov), the Human Microbiome Project (HMP, http://www.hmpdacc.org), the Saccharomyces Genome Database (SGD, http://www.yeastgenome.org), the Fungal Genome Initiative (FGI, http://www.broadinstitute.org), FungiDB (http://fungidb.org), and internally sequenced genomes (Supplementary Table 2). Where multiple genomes for a reference were available, we selected complete over draft genomes. Reads not matching hg19 + hg19 rRNA were mapped to this genome collection using bowtie2's[23] —very-sensitive parameter retrieving the top 10 hits. Reads mapping to multiple genomes were then reassigned to a 'most likely' genome using Pathoscope v1.0[24], which uses a Bayesian framework to examine each read's sequence and mapping quality within the context of a global reassignment. Read hit counts were then normalized by genome length and scaled to sum to one. To reduce the likelihood of recovering spurious genomes, we also calculated genome coverage for each genome hit using the genomeCoverageBed tool in the Bedtools suite[25]. For relative abundance and diversity calculations, genomes with coverage <1 were removed to decrease low-abundance false positives, providing a measure of normalization for sequencing depth.

To assess the accuracy of our taxonomic classifications and our estimation of community diversity, we compared taxonomic assignments of bacteria and fungi to 16S and ITS amplicon results, as well as to the output from a bacterial and archaeal mapping tool, Metaphlan[26]. We observed high correlations extending to the species level for bacterial sequences (Extended Data Fig. 3, Supplementary Tables 2–4). Concordance of non-*Malassezia* fungal species was lower, presumably due to the relative paucity of sequenced fungal genomes. We used the Shannon diversity index as well as species observed for diversity comparisons for bacterial classifications. All taxonomies were reconstructed to the species level, combining hits to multiple strain subtypes. The coverage cutoff of 1 was chosen as an inflection point for species accumulation and as a point of concordance between diversity estimates derived from other approaches.

We characterized the representation of functional gene groups in the skin using the KEGG Orthology gene pathway (KO) and module (MO) annotations[27], calculating corresponding abundances and coverages using the HMP Unified Metabolic Analysis Network (HUMAnN)[28]. We note that functional diversity is probably underestimated in the absence of viral pathways in the KEGG database. We mapped reads to the 2013.10.14 KEGG release using USEARCH v7.0[29] e-value <0.01, -accel 0.5 as described[28]. The top 10 hits were then processed with HUMANN v0.99[28]. To define genetic carriage of resistance profiles in the skin, antibiotic resistance genes from the Antibiotic Resistance Genes Database (ARDB)[30] were clustered based on sequence similarity to produce families of unique short sequence markers using ShortBRED (J. Kaminski, N. Segata, E. Franzoza and C. Huttenhower, unpublished). Reads were then mapped to the top marker using USEARCH v7.0, minimum alignment length 20, percent identity 95%. A family (resistance gene) was called present if at least one gene of that family was represented with a non-zero median of all its markers (median number of hits to its markers >0). Each family was normalized by the number of the hits, the marker length, and the length of the original protein sequence. We considered only presence/absence for a more conservative assessment. We note that while antibiotic resistance genes are typically classified with respect to a particular species, from metagenomic data it is difficult to impute an organism of origin because families can be encoded on plasmids (for example, NP_040465, a tetracycline efflux pump).

Reference-based strain mapping. Accurate, *de novo* identification of single nucleotide polymorphisms (SNPs), used in metagenomic strain tracking of high-biomass stool samples, typically requires 100× coverage for robust identification[31]. Given strain variance due to differential representation and sequencing depth, we developed a reference-based approach, assessing feasibility and accuracy with computational simulations of communities of mixed complexity. For bacteria *Propionibacterium acnes* and *Staphylococcus epidermidis*, we created custom, species-specific reference

©2014 Macmillan Publishers Limited. All rights reserved

databases incorporating all complete and draft genomes present for those species from NCBI, totalling 78 and 61, respectively (Supplementary Table 2). To visualize relationships between the strains, all SNPs identified in core regions were used to create dendrograms with the program PhyML 3.0[32]. Strains were assigned to a subtype based on phylogenetic distance, for example, we defined 12 subtypes for *P. acnes* and 14 for *S. epidermidis*.

For each respective set of reference genomes, we identified first, SNPs unique to each strain in regions shared in all genomes ('core'), and second, larger regions that are partially shared or unique to a strain ('non-core', Supplementary Table 2). We mapped reads to each database using bowtie2 with stringent parameters (–score-min L,-0.6,0.006), allowing zero mismatches and as many hits as genomes in the database. Read assignment using Pathoscope was performed as described, except theta_prior, an option that controls the proportion of non-unique reads that are assigned to a genome, was set to 10×10^{88} (most genomes permitted). Normalization was performed as described above.

Because Pathoscope can reassign reads to closely related genomes rather than an actual target genome that may or may not be present in a sample, we evaluated the ability of Pathoscope to accurately reassign reads to very similar sub-strains by first, assessing sensitivity of complex staggered mixtures of synthetic communities, and second, demonstrating the presence of unique genomic loci that allow discrimination between subtypes. First, synthetic communities were created with 6, 12, or 18 genomes per community, with 50,000, 100,000, or 500,000 reads sampled per genome for an even mix, as well as a staggered community to estimate accuracy in abundance calling. 15 random synthetic communities for each even genome group, and 5 for staggered, were created and mapped to the full genome set. Sensitivity was calculated from the expected versus observed abundances. Second, we identified SNPs unique to each genome in 'core' regions of the genome (defined as shared between all reference genomes in species-specific database) using nucmer[33] and custom scripts. nucmer was also used to identify 'non-core' regions in each of the genomes. Simulated reads were then mapped to strains based upon: (1) consensus SNPs, (2) non-core region variants, or (3) full genomes to identify what variants are shared between sites/individuals. In simulations, core SNPs had the highest sensitivity, but whole genomes, which incorporate both core and non-core elements, were best able to identify closest neighbour strains (Extended Data Fig. 5, Supplementary Table 9). Although we have supported our results using SNPs (Supplementary Table 10), mapping to whole genomes provided clear advantages if an exact reference strain is not present *in vivo*, which is likely given the limited number of fully sequenced genomes. In absence of an exact reference, our approach robustly defines most similar strains based on differentiating genomic features.

Adaptive iterative *de novo* assembly. Assembly efficacy varies depending on the site's unique features of community complexity, typically defined by microenvironment, and sequencing depth, which is affected by biomass and human DNA admixture. To optimize assembly parameters, individual samples were assembled using a wide *k*-mer range in Velvet[34], and contigs greater than 300 bp in length were analysed. To examine assembly efficacy, reads were remapped to assemblies using bowtie2 —sensitive. 'Adaptive' denotes that each sample was assembled using *k*-mers ranging from 37–69. A quality score was calculated using % paired or singleton reads realigning to the assembly, the number of bases incorporated into the assembly, and number of contigs >300 bp. The assembly with the highest quality score was used for subsequent analysis. 'Iterative' denotes subsequent steps in which unaligned reads from remapping were then pooled to improve recovery of rare genes that may represent genomes unique to an individual. We found that pooling by individual produced higher quality assemblies than pooling by site (Supplementary Table 17). This observation supported our insight that while site can shape the major features of a community, species and strains are shared within an individual. To improve assembly quality and reduce computational burden, digital normalization[20], which reduces error by removing redundant data and performs similarly to non-normalized data (Extended Data Figure 10c), was applied on pooled samples before assembly. We used two-pass normalization to 20× then 5× with variable coverage and assembled with adaptive *k*-mer selection. Finally, unaligned reads from pooled individual assemblies were pooled and subsampled 1:10 before normalization and variable assembly.

To create a multi-kingdom skin microbial gene catalogue, genes were predicted from contigs using two models, MetaGeneMark[35], which incorporates multiple bacterial models, and Augustus[36] with a *Ustilago maydis* model as a phylogenetically near neighbour to *Malassezia*, the most predominant skin fungi. To account for cases where both fungal and bacterial genes were called for the same contig, we adopted a filtering methodology by which each contig was assigned to a kingdom using blastn against our microbial database, or where no blastn hit was available, a blastx against nr using USEARCH. Discordant calls not resolved by blastn/x filtration were marked ambiguous or assigned to whichever caller generated a prediction. A non-redundant catalogue was constructed using UCLUST with sequence identity cut-off of 0.95 and a minimum coverage cutoff of 0.9 for shorter sequences. This final catalogue contained 5,922,920 putative bacterial and fungal genes.

During this process, we also observed that many short contigs (<1,000 bp) produced no putative genes. To circumvent losing partial genes or genes unidentifiable by our prediction models, we revised our gene catalogue to first retrieve contigs <1,000 bp, then call genes on contigs >1,000 bp as previously described. To assess the abundance of genes, reads were aligned to the gene catalogue with Bowtie2 — sensitive and counts per gene were normalized by length.

Putative metagenomic clusters, based on covariance of gene abundances across samples, were formed as described[10]. Genes from the same genome are assumed to co-vary in relative abundance across subjects due to physical linkage; therefore such clusters can serve as a proxy for unknown organisms or known organisms with variable gene content. We clustered gene abundances across samples, grouped by site characteristic both to improve segregation of clusters and reduce computational burden. To reduce false positives and computational complexity, we required genes to be present in at least 20% of samples for a given site characteristic. The abundances of these genes across samples were then clustered using the Markov clustering algorithm implemented in MCL[37] with a Spearman correlation coefficient of 0.85 and inflation parameter set to 2. Cluster parameters varying presence to 40% presence across samples, correlation coefficients to 0.80 and 0.90, and inflation parameters of 4 produced similar results. For toenail, 40% presence and clustering at 80% was performed due to computational limitations imposed by site complexity. Clusters were taxonomically annotated by blastx-ing each gene in a cluster to nr as previously described, and as a strict requirement against false binning, clusters with at least 50% of genes mapping to the same phylogenetic group at the species, genus, and/or family level were retained as a metagenomic 'cluster'. Clusters with the same consensus taxonomy were merged at the genus and species level; family level analysis showed minimal improvements in consensus (Supplementary Table 18). Because a typical microbial genome contains thousands of genes, we speculate that many of these represent gene fragments that did not pass our stringent redundancy thresholds. While our variable sequencing depth likely precludes recovery of complete genomes from such a metagenomic linkage analysis, we identified large clusters of taxonomically related groups of covarying genes for both characterized and uncharacterized species.

Statistical analysis. All statistical analyses were performed in the R software. Data are represented as mean ± standard error of the mean unless otherwise indicated. For all boxplots, black centre lines represent the median and box edges the first and third quartiles. 'e' in scientific notation refers to 10 ×, for example, 10e5 represents 10×10^5. Spearman correlations (ρ) of non-zero values were used for all correlation coefficients. The nonparametric tests Wilcoxon rank-sum and Kruskal–Wallis were used to determine statistically significant differences between microbial populations, and to identify significant inter-category comparisons, we used a post-hoc multiple comparison test, implemented by the kruskalmc test in the pgirmess package. Unless otherwise indicated, *P* values were adjusted for multiple comparisons using the p.adjust function in R using method = "fdr"[38]. Statistical significance was ascribed to an alpha level of the adjusted *P*-values ≤ 0.05. Site characteristics were treated as separate groups where indicated based on spatial physiological differences between these different body niches[2]. Similarity between samples was assessed using the Yue–Clayton theta similarity index[39] with relative abundances of species, sub-strains, or shared genomic variants. The theta coefficient assesses the similarity between two samples based on (1) number of features in common between two samples, and (2) their relative abundances with $\theta = 0$ indicating totally dissimilar communities and $\theta = 1$ identical communities. To avoid repeated measures, samples belonging to an individual were averaged before statistical comparisons between site characteristic when using summary metrics such as means, diversity, or theta indices.

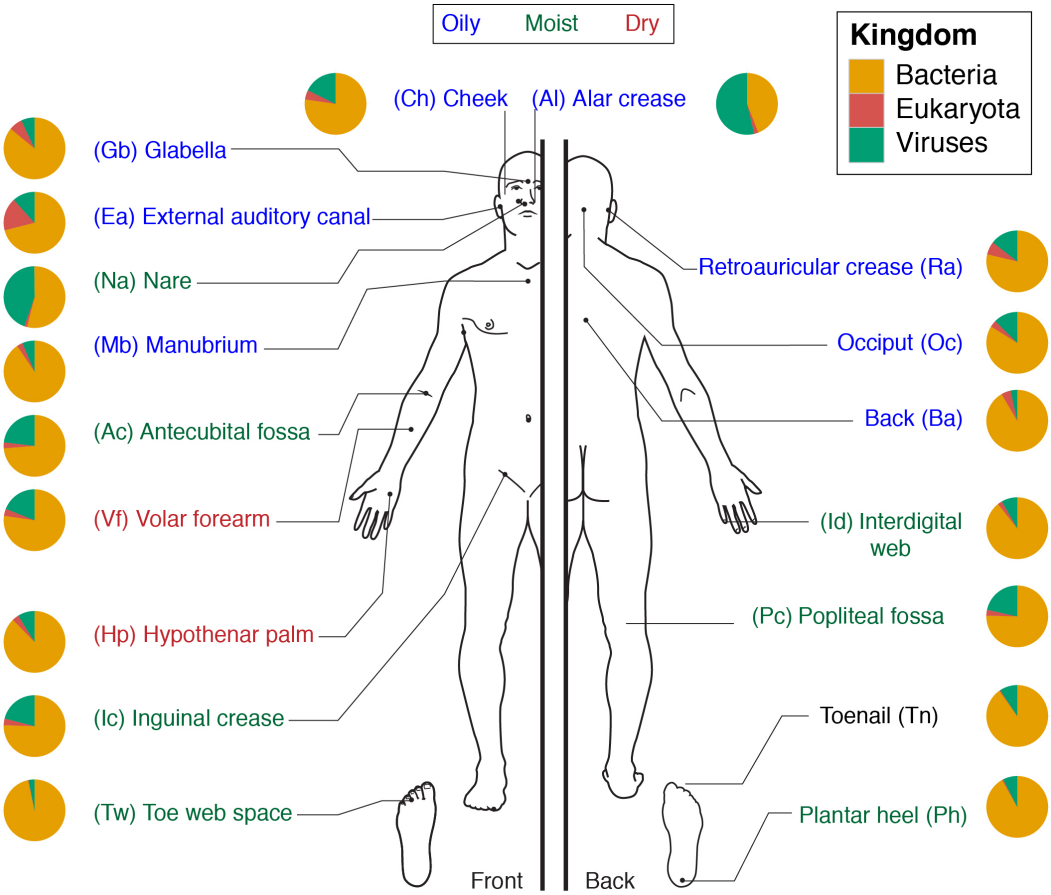
Supervised random forest models to identify discriminatory taxa and modules was implemented with the randomForest package in R[40]. This analysis was enabled by our multi-site sampling strategy, as using a single or few sites lacks statistical power to detect low abundance features. Mean decrease in accuracy denotes the normalized difference in the classification accuracy when that variable is included versus when data are randomly permuted, that is, to what degree inclusion of this predictor in the model reduces classification error. Model accuracy was calculated using the out-of-bag (oob) error estimate, which is an approximation of how frequently an individual is misclassified.

19. Jumpstart Consortium Human Microbiome Project Data Generation Working Group. Evaluation of 16S rDNA-based community profiling for human microbiome research. *PLoS ONE* **7,** e39315 (2012).
20. Howe, A. C. *et al.* Tackling soil diversity with the assembly of large, complex metagenomes. *Proc. Natl Acad. Sci. USA* **111,** 4904–4909 (2014).
21. Schloss, P. D. *et al.* Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* **75,** 7537–7541 (2009).

©2014 Macmillan Publishers Limited. All rights reserved

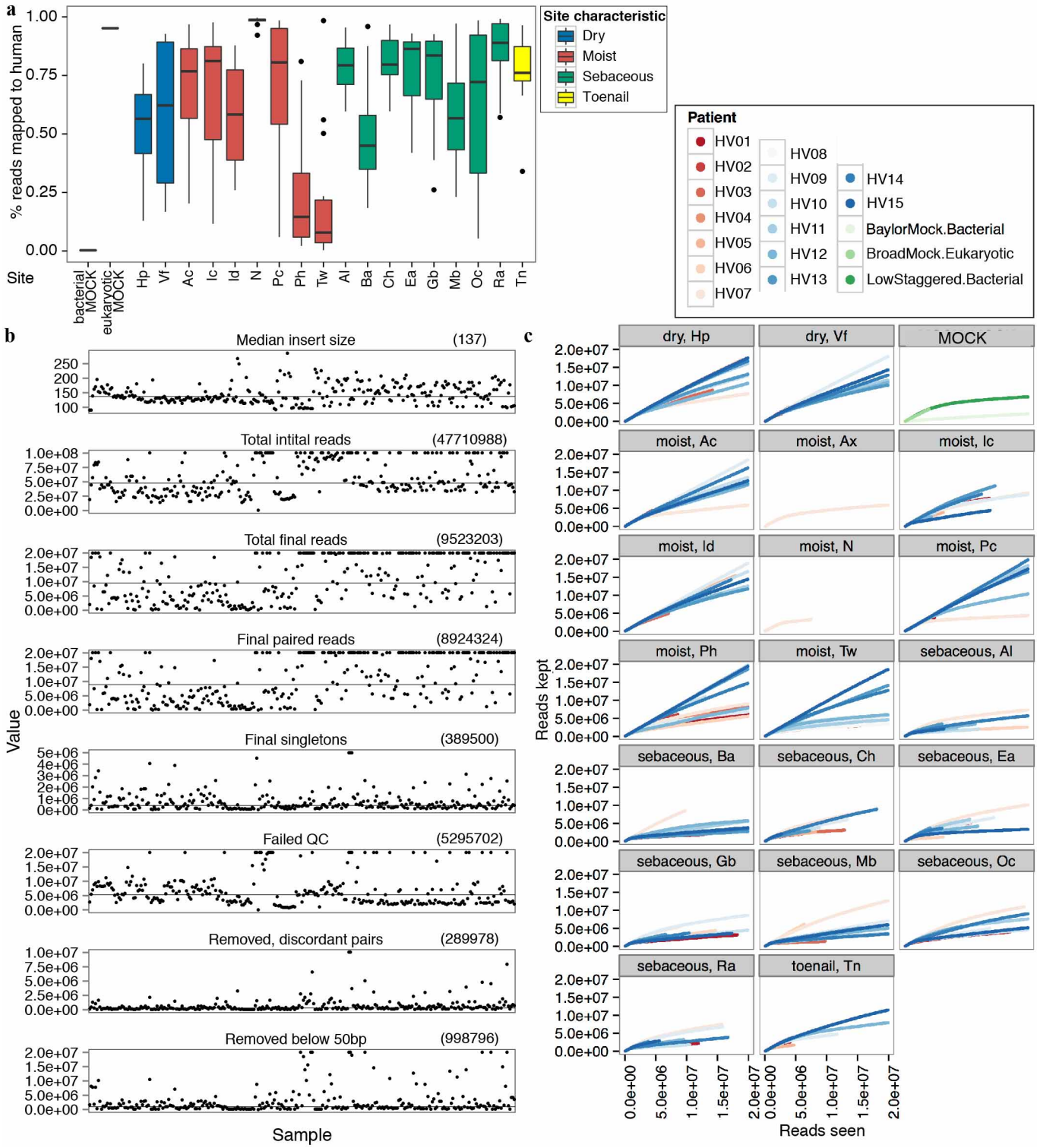
22. Matsen, F. A., Kodner, R. B. & Armbrust, E. V. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics* **11,** 538 (2010).
23. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9,** 357–359 (2012).
24. Francis, O. E. *et al.* Pathoscope: Species identification and strain attribution with unassembled sequencing data. *Genome Res.* **23,** 1721–1729 (2013).
25. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).
26. Segata, N. *et al.* Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods* **9,** 811–814 (2012).
27. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28,** 27–30 (2000).
28. Abubucker, S. *et al.* Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLOS Comput. Biol.* **8,** e1002358 (2012).
29. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26,** 2460–2461 (2010).
30. Liu, B. & Pop, M. ARDB—Antibiotic Resistance Genes Database. *Nucleic Acids Res.* **37,** D443–D447 (2009).
31. Schloissnig, S. *et al.* Genomic variation landscape of the human gut microbiome. *Nature* **493,** 45–50 (2013).
32. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59,** 307–321 (2010).
33. Delcher, A. L., Phillippy, A., Carlton, J. & Salzberg, S. L. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res.* **30,** 2478–2483 (2002).
34. Namiki, T., Hachiya, T., Tanaka, H. & Sakakibara, Y. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Res.* **40,** e155 (2012).
35. Zhu, W., Lomsadze, A. & Borodovsky, M. *Ab initio* gene identification in metagenomic sequences. *Nucleic Acids Res.* **38,** e132 (2010).
36. Stanke, M., Schöffmann, O., Morgenstern, B. & Waack, S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* **7,** 62 (2006).
37. van Dongen, S. & Abreu-Goodger, C. in *Bacterial Molecular Networks* (eds Helden, J., Toussaint, A. & Thieffry, D.) *Methods in Molecular Biology* Vol. 804, pp. 281–295 http://link.springer.com/protocol/10.1007/978-1-61779-361-5_15 (Springer, 2012).
38. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57,** 289–300 (1995).
39. Yue, J. C. & Clayton, M. K. A similarity measure based on species proportions. *Comm. Stat. Theory Methods* **34,** 2123–2131 (2005).
40. Liaw, A. & Wiener, M. Classification and regression by randomForest. *R News* **2,** 18–22 (2002).

©2014 **Macmillan Publishers Limited. All rights reserved**



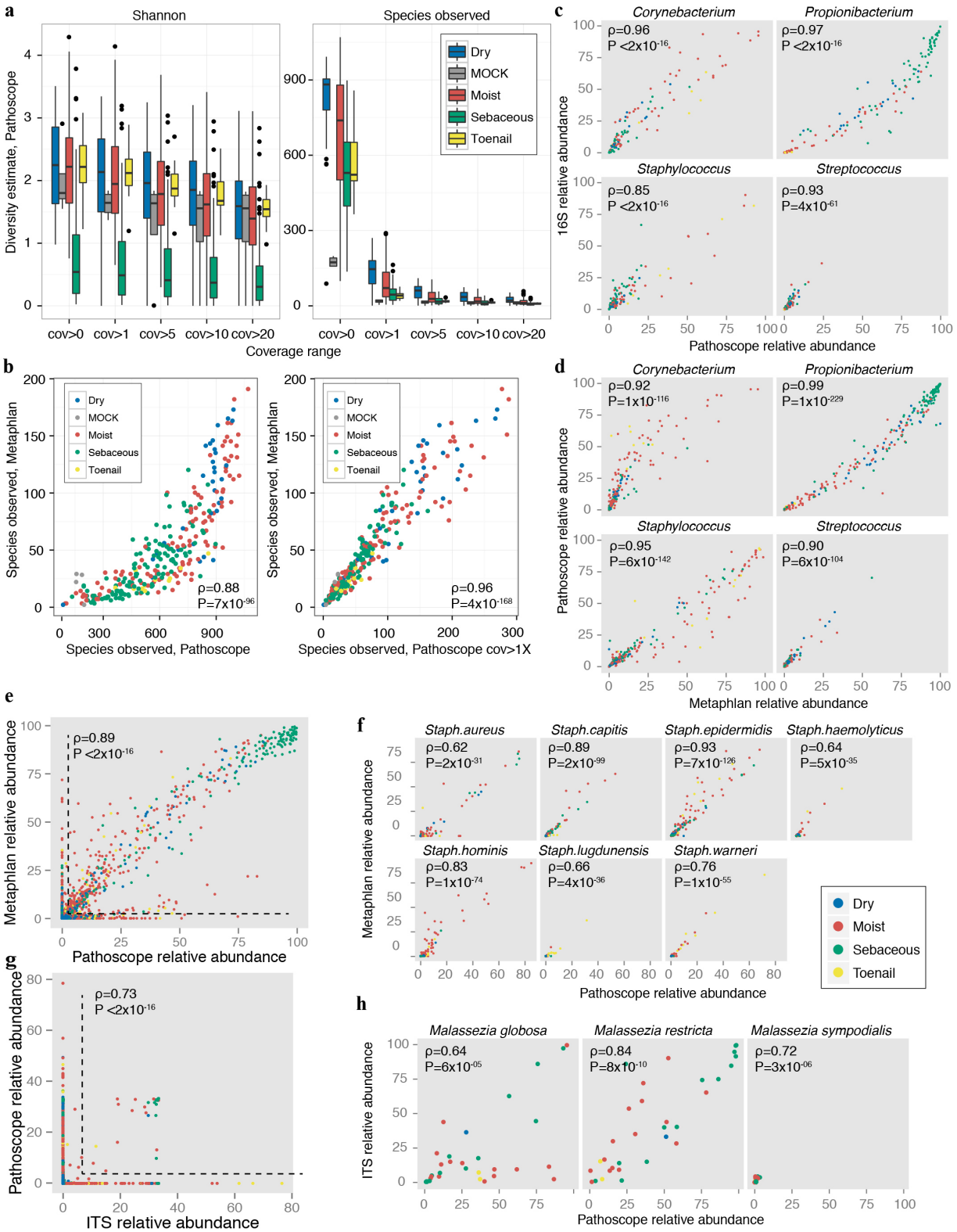
Extended Data Figure 1 | The 18 selected skin sites and their location on the human body. These sites represent three microenvironments: sebaceous (blue), dry (red), and moist (green). Toenail (black) is a site that does not fall under these major microenvironments and is treated separately. Pie charts represent consensus relative abundance of the kingdoms Bacteria, Eukaryota (Fungi), and virus from multi-kingdom mapping.

©2014 Macmillan Publishers Limited. All rights reserved



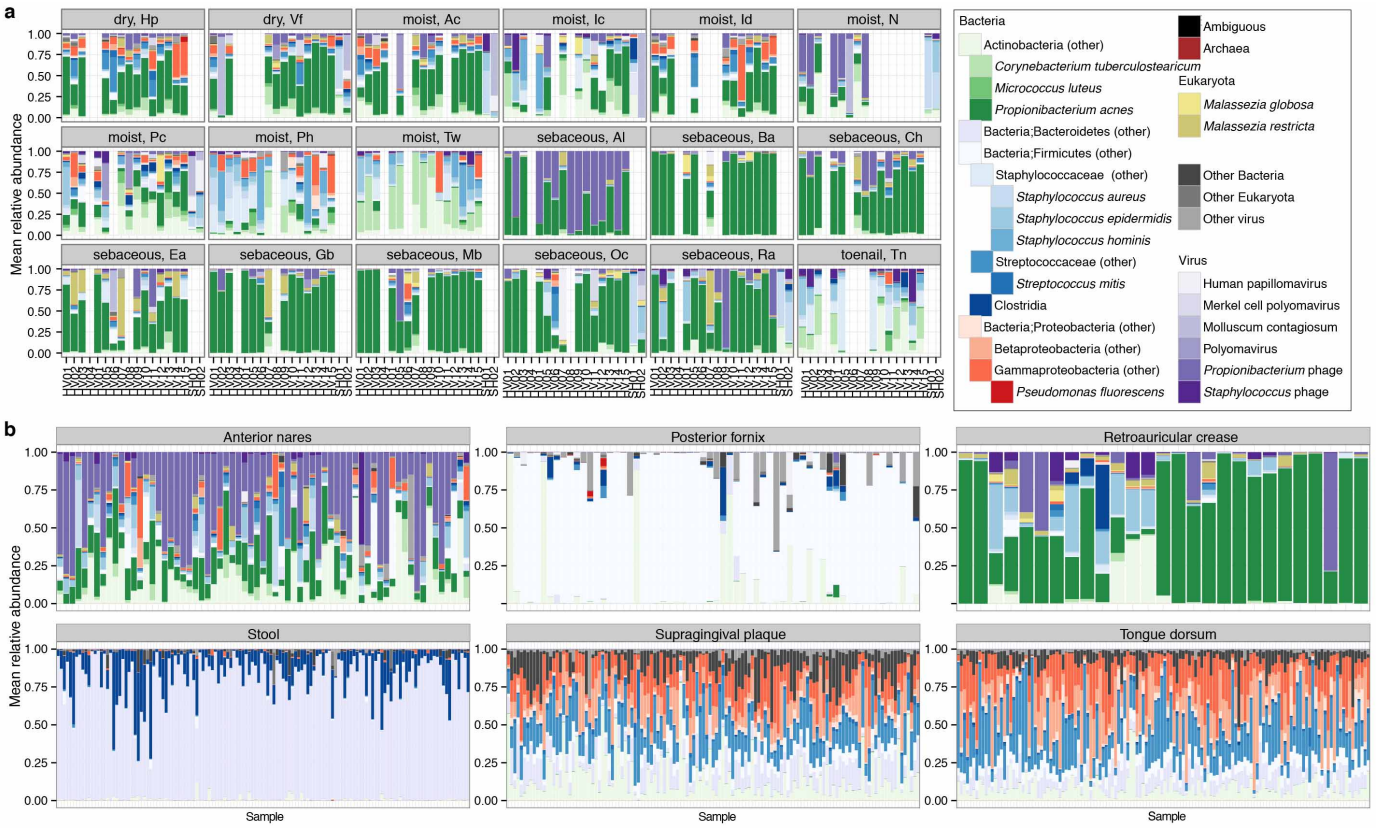
Extended Data Figure 2 | Per-sample read statistics. Additional samples (bacterial and eukaryotic mock communities) are shown. **a**, Boxplots (line indicates median; boxes represent first and third quartiles) show, for each site, % reads mapping to human hg19 that are discarded before analysis. Sites are coloured by site characteristic. **b**, Samples are ordered by label. Lines indicate the median value for that statistic; value is in parenthesis. **c**, Estimate of sequencing coverage. Reads seen is the number of reads in a sample sampled. Reads are then split into 20-mers, compared to a *k*-mer coverage table and kept only if the median *k*-mer coverage is below 20×. Curves are grouped by site, coloured by individual as indicated.

©2014 Macmillan Publishers Limited. All rights reserved



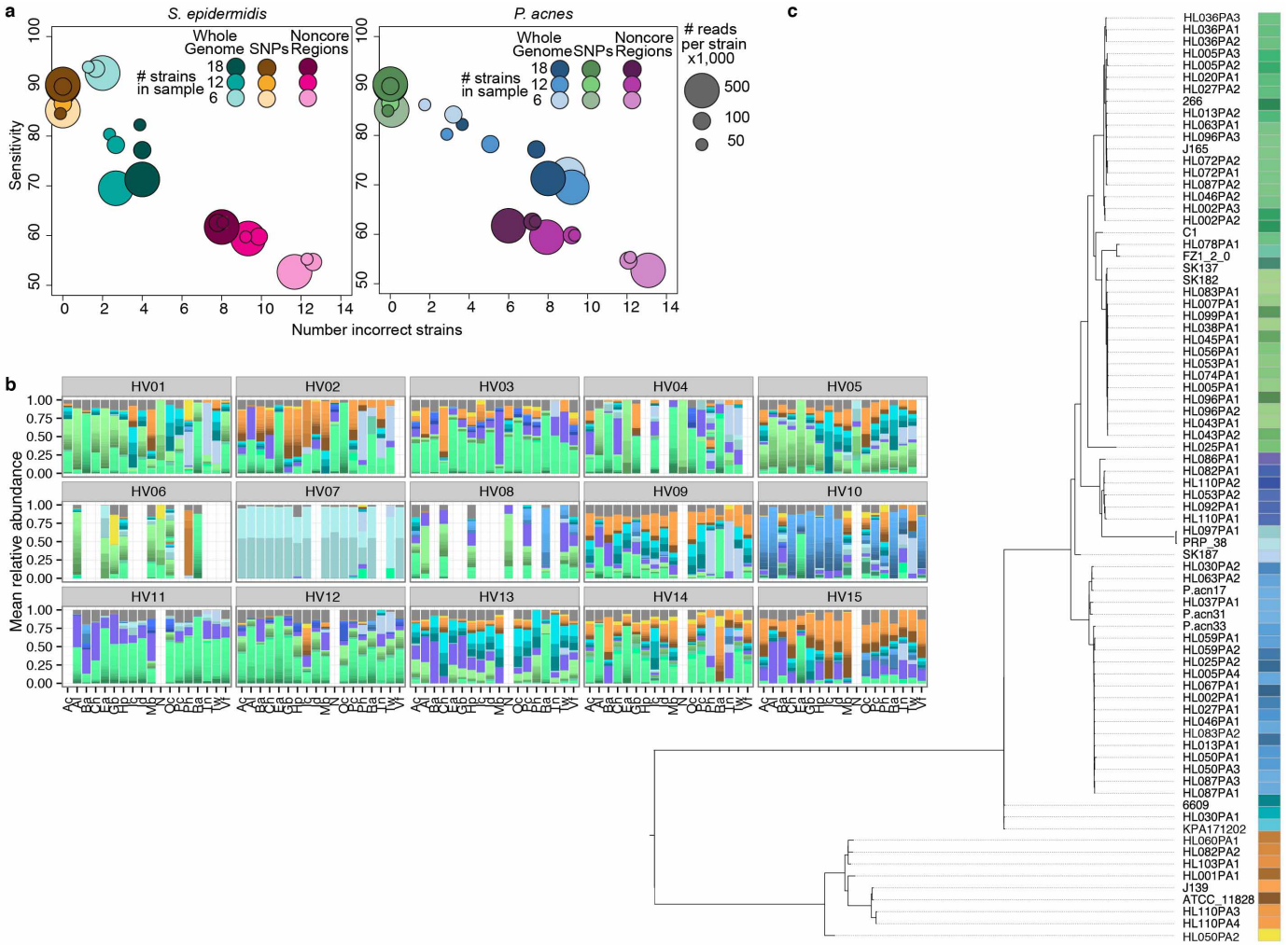
Extended Data Figure 3 | Validation of taxonomic classifications. **a**, Bacterial sample community diversity as a function of genome coverage for two diversity metrics, the Shannon index that measures the richness and evenness of the community (left), and number of species observed (right). Genome coverage is defined as for each genome hit, the % of genome covered by reads. Boxplots show the range of diversity values for all samples, segregated by microenvironment. Black lines indicate median; boxes represent first and third quartiles. As coverage cut-offs increase, diversity estimates drop sharply. **b**, Comparisons of bacterial community diversity for Metaphlan-derived classifications versus custom bacterial Pathoscope-derived classifications. Each point represents a different sample, coloured by microenvironment. With no coverage cut-offs (left), Pathoscope may overestimate diversity, which is reduced by setting a minimum 1× coverage requirement. Spearman correlation (ρ) and corresponding P values are shown. Pathoscope-derived relative abundances versus relative abundances derived from **c**, 16S amplicon sequencing, **d**, Metaphlan genus-level, **e**, Metaphlan-species level (ρ and P value are calculated for non-zero abundance taxa), **f**, Metaphlan, staphylococcal species, **g**, ITS1 amplicon sequencing, genus (ρ and P value are calculated for non-zero abundance taxa), and **h**, ITS1 amplicon sequencing, *Malassezia* species.

©2014 Macmillan Publishers Limited. All rights reserved



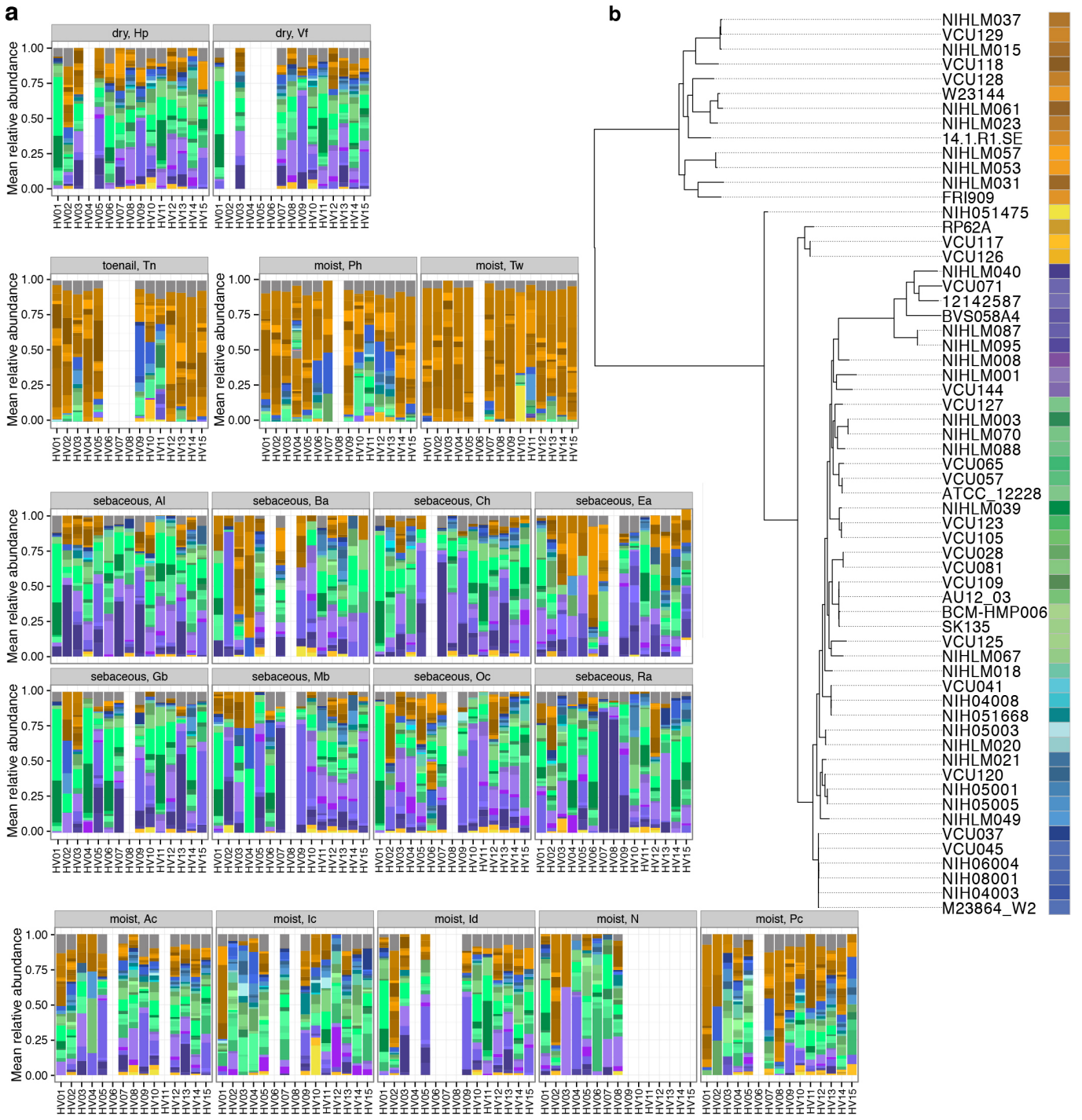
Extended Data Figure 4 | Full taxonomic classifications for all healthy volunteers (HV), all sites. To aid visualization of site- and individual-specific similarities, samples are grouped by site/microenvironment for each individual. Relative abundances of the most abundant skin taxa for each super-kingdom are shown. **b**, Taxonomic re-classification of major sites sampled by the Human Microbiome Project. Samples are from the anterior nares and retroauricular crease (skin), tongue dorsum and supragingival plaque (oral), stool, and posterior fornix (vaginal). Relative abundances of the most abundant taxa for each kingdom in the skin, for comparison, are shown.

©2014 Macmillan Publishers Limited. All rights reserved



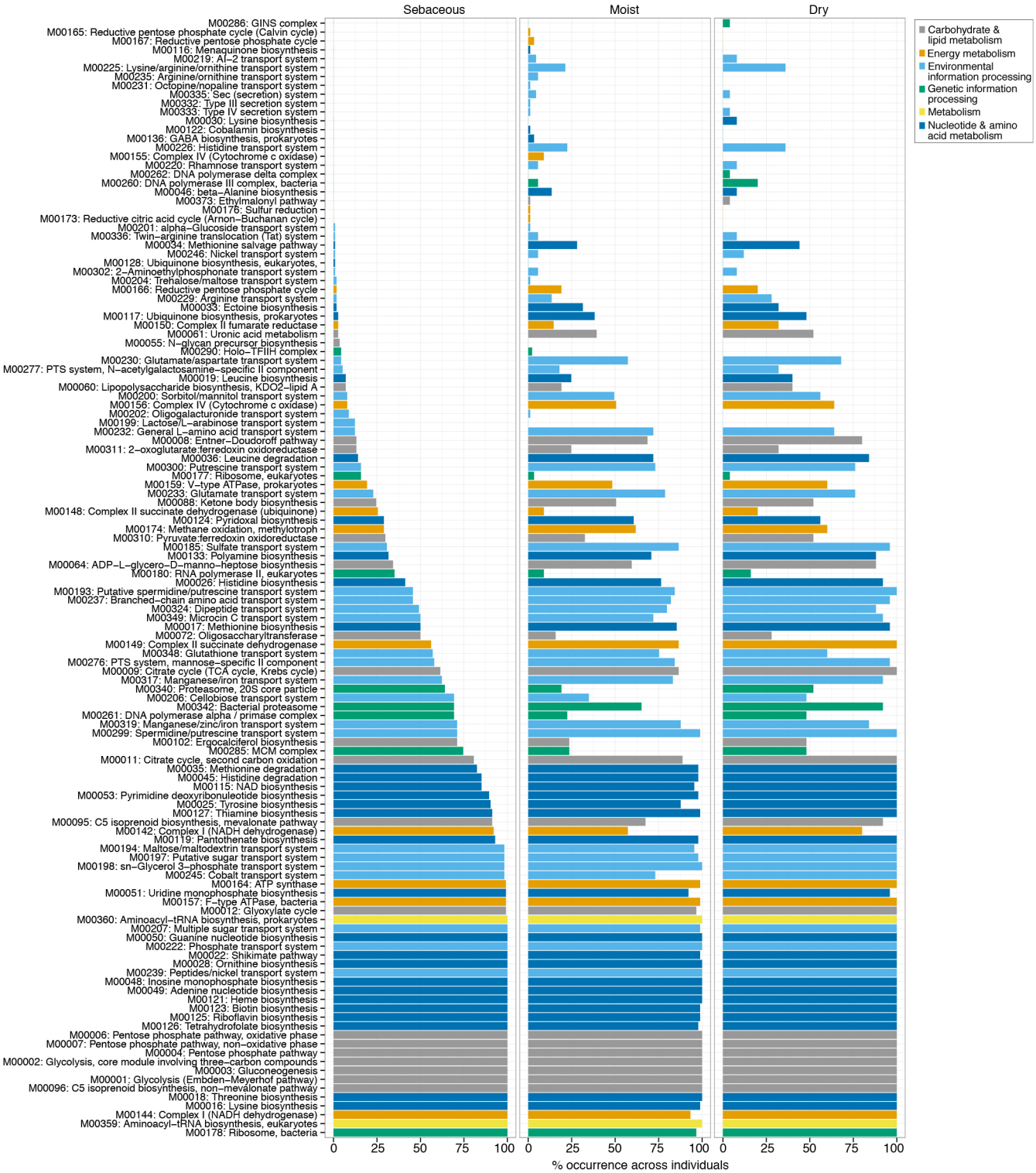
Extended Data Figure 5 | Strain-level classification based on reference genomes show sub-species heterogeneity for dominant skin taxa. **a**, Simulations to assess sensitivity of Pathoscope-based mapping to SNPs, non-core regions, or whole genomes. Synthetic communities were created with 6, 12, or 18 genomes per community. Sizes of circles reflect the number of reads sampled from each genome, for example, 50,000, 100,000, or 500,000 reads per genome. 15 random synthetic communities for each genome group were created and mapped to SNPs, non-core regions, or the full genome set. Sensitivity is calculated from the expected versus the observed abundances. **b**, Full strain-level assignments for samples with relative abundances of closest related *Propionibacterium acnes* strains, by individual. **c**, Dendrograms of strain similarity. Trees were generated using core SNPs; genomes were aligned with nucmer to identify core regions, and then SNPs within these core regions were identified by calculating all pairwise differences between genomes. Bar of colours indicates delineations of subtypes where phylogenetically more similar genomes are in similar colours; for example, we defined 12 subtypes for *P. acnes*.

©2014 Macmillan Publishers Limited. All rights reserved



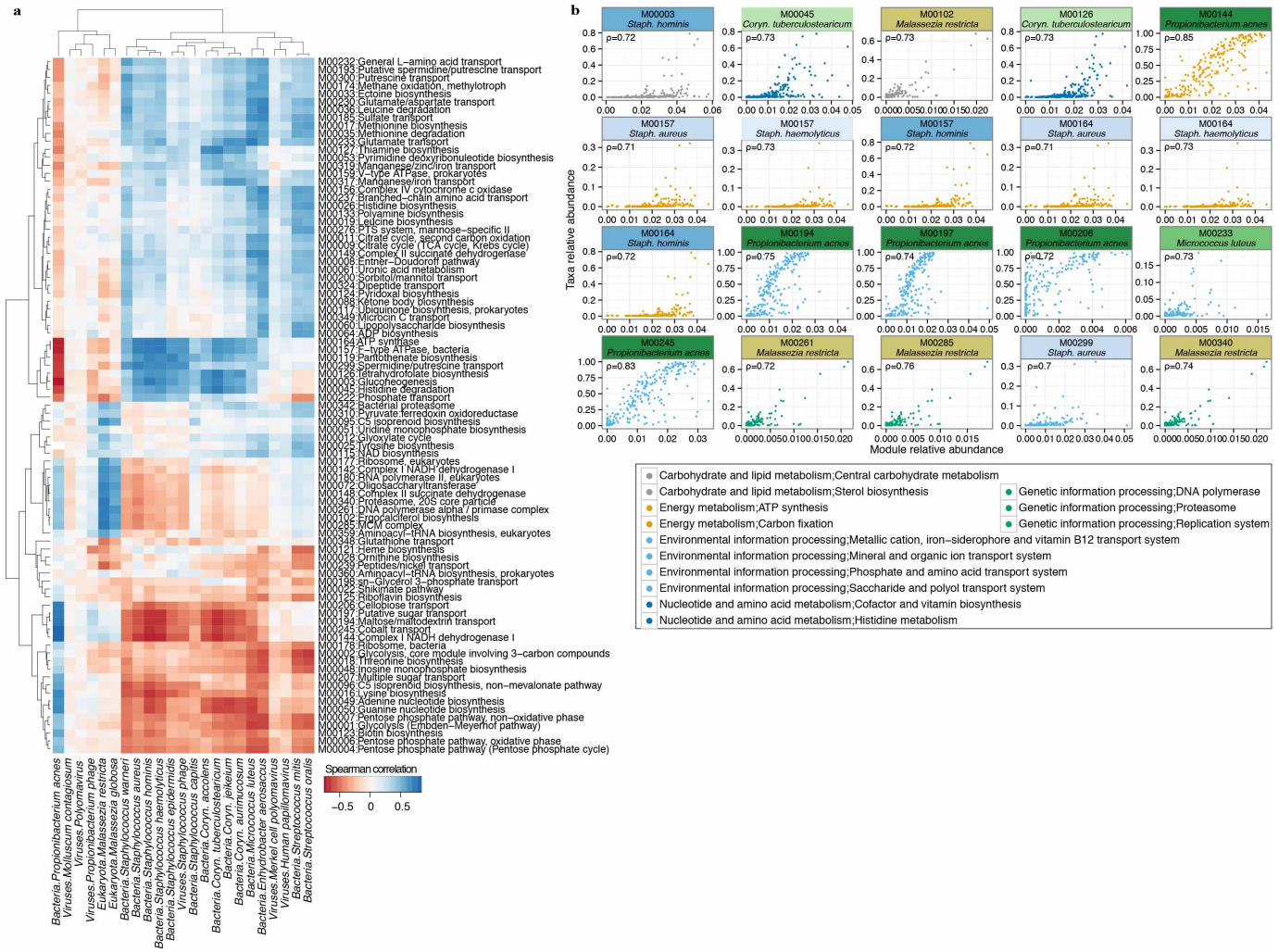
Extended Data Figure 6 | Strain-level classification for *Staphylococcus epidermidis*. **a**, Full strain-level assignments for samples by microenvironment. **b**, Description is as in Extended Data Fig. 5c. We defined 14 subtypes for *S. epidermidis*.

©2014 Macmillan Publishers Limited. All rights reserved



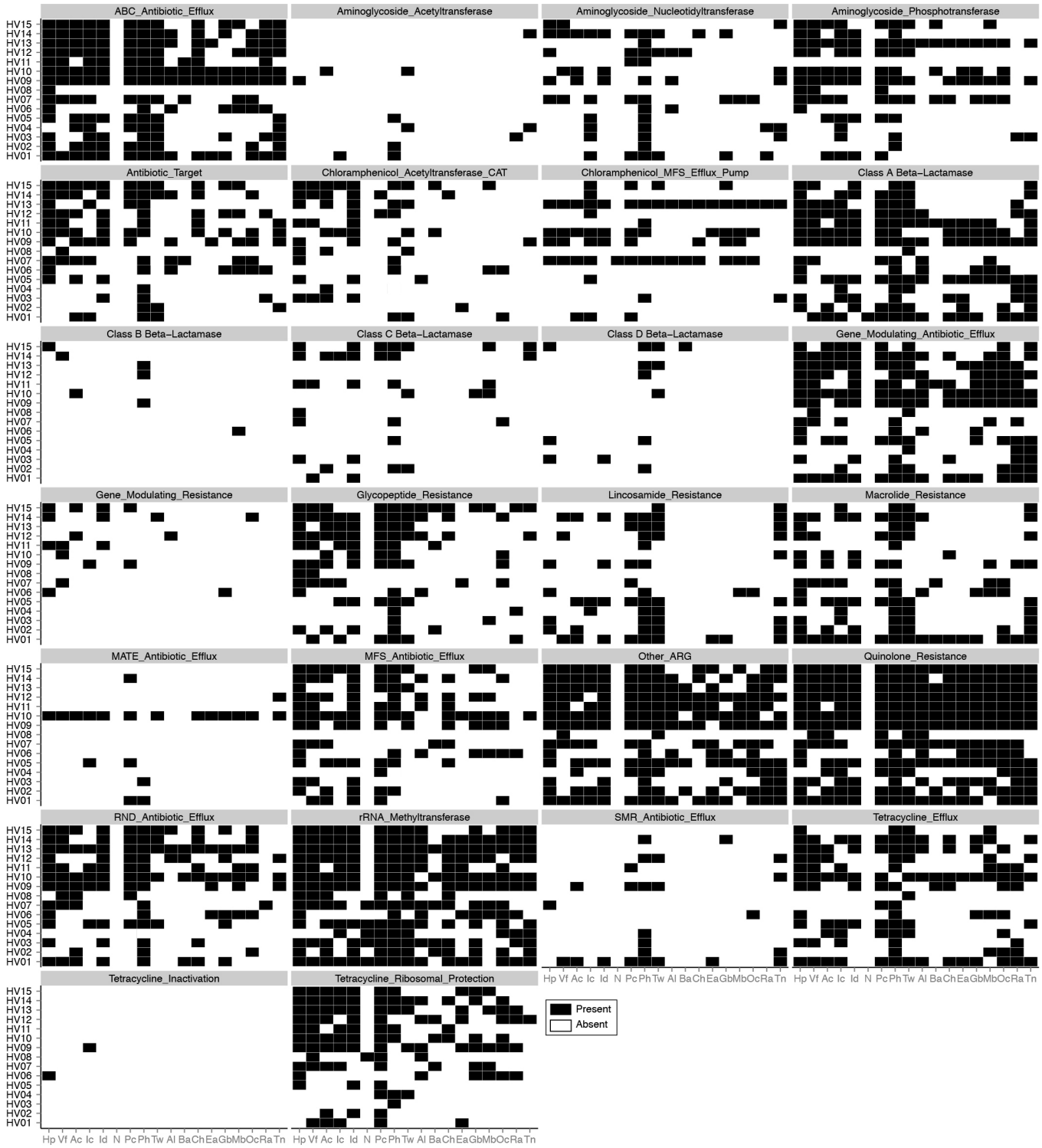
Extended Data Figure 7 | Full version of coreness of different module categories across skin microenvironment. A module is defined as core if occurring in >2/3 of samples for that class. Major KEGG module descriptors are shown in the different colours. Height of bars reflects the proportion of samples that a module occurs in.

©2014 Macmillan Publishers Limited. All rights reserved



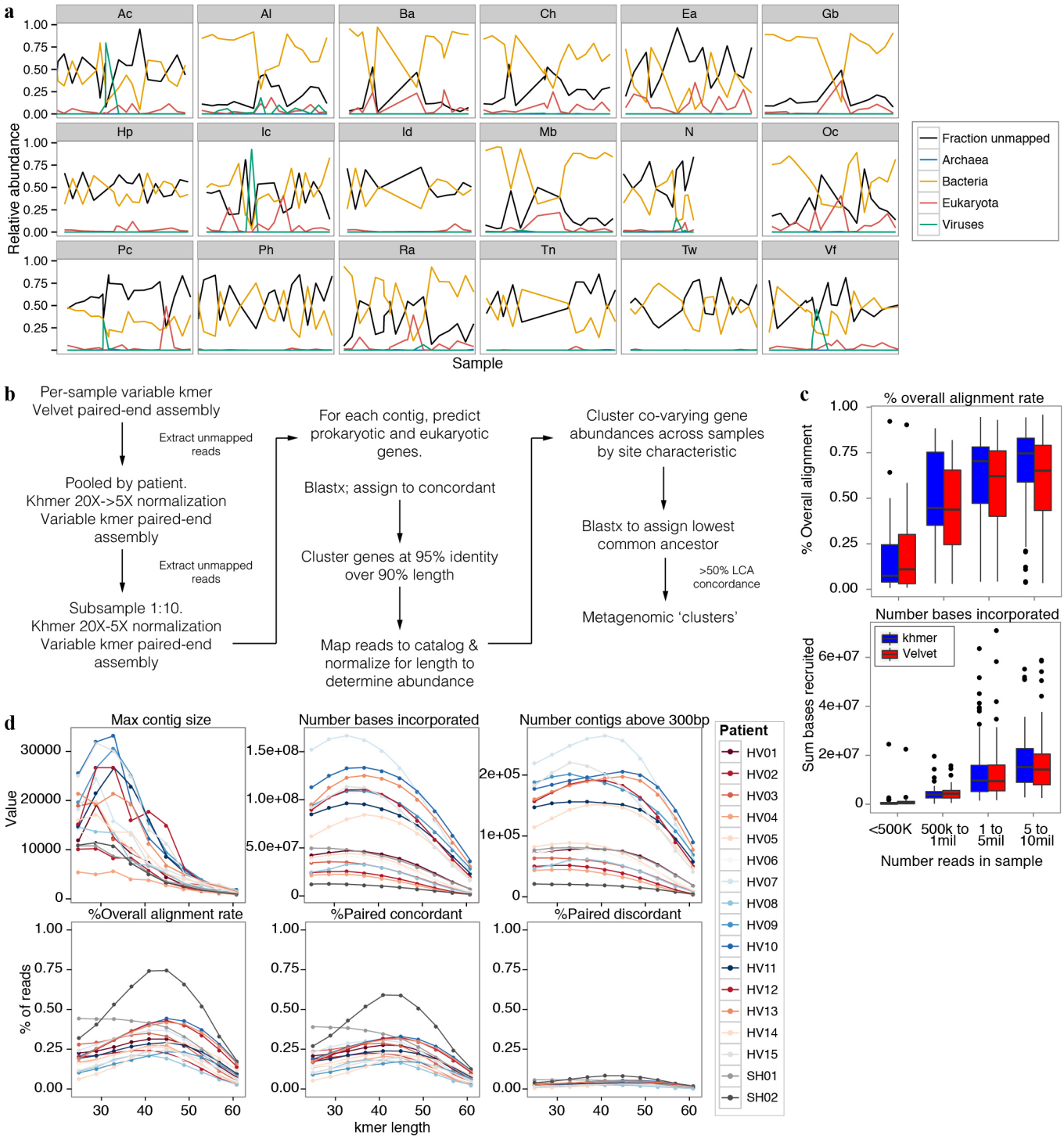
Extended Data Figure 8 | Correlation analysis of module abundance with species abundance to infer a module's taxonomic origin. Spearman correlation (ρ) was calculated with corresponding P value for taxa with relative abundance >0.5% and modules with greater than 0.05% relative abundance. *Coryn., Corynebacterium.* **a**, Unsupervised clustering of correlation coefficients. Species from the same genera clustering together may suggest a shared contribution of a pathway. **b**, Most significantly correlated taxa; colours represent broad KEGG classes. Adjusted $P < 2 \times 10^{-16}$.

©2014 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 9 | Antibiotic resistance profiles in the skin. Reads were mapped to a short marker database consensus created from the ARDB database, which catalogues publicly available resistance genes. Genes are grouped into broad resistance classes; a resistance category is called present (black; absent = white) if at least one gene from its family is present.

©2014 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 10 | Reference-free analysis of skin metagenome with adaptive iterative assembly, gene catalogue, and metagenomic clusters. **a**, Tracking unclassified reads. Fraction unmapped reads refers to the fraction of total reads passing quality control that do not map to the major super kingdoms Archaea, Bacteria, Eukaryota, and viruses. Samples are ordered by label and are divided by site. **b**, Assembly, gene-calling, and clustering workflow. **c**, Assembly efficacy varies significantly by *k*-mer depending on the site's unique features of community complexity and sequencing depth, which is most affected by that site's human DNA admixture. Assembly statistics are shown for samples pooled by individual, which produced higher quality assemblies than pooling by site. Because of large pool size, khmer digital normalization was used before Velvet assembly. % overall alignment rate indicates the total % of reads that map back to that sample's assembly for each *k*-mer. % paired concordant indicates the fraction paired reads (of overall, not of % paired) in which both pairs of a mate map back to an assembly; discordant is where one mate of a pair does not map, or maps to a different contig. Contigs are then assessed by the maximum assembly size, the number of bases that are used in the assembly, and the number of contigs above a threshold of 300 bp. **d**, Effect of khmer digital normalization on individual sample assembly. Digital normalization + Velvet assembly performs similarly to Velvet assembly alone.

©2014 Macmillan Publishers Limited. All rights reserved