nature medicine

LETTERS

https://doi.org/10.1038/s41591-018-0211-7

Expanded skin virome in DOCK8-deficient patients

Osnat Tirosh[1], Sean Conlan[1], Clay Deming[1], Shih-Queen Lee-Lin[1], Xin Huang[1], NISC Comparative Sequencing Program[2], Helen C. Su[3], Alexandra F. Freeman[3], Julia A. Segre[1,6]* and Heidi H. Kong[4,5,6]*

Human microbiome studies have revealed the intricate interplay of host immunity and bacterial communities to achieve homeostatic balance. Healthy skin microbial communities are dominated by bacteria with low viral representation[1–3], mainly bacteriophage. Specific eukaryotic viruses have been implicated in both common and rare skin diseases, but cataloging skin viral communities has been limited. Alterations in host immunity provide an opportunity to expand our understanding of microbial–host interactions. Primary immunodeficient patients manifest with various viral, bacterial, fungal, and parasitic infections, including skin infections[4]. Dedicator of cytokinesis 8 (DOCK8) deficiency is a rare primary human immunodeficiency characterized by recurrent cutaneous and systemic infections, as well as atopy and cancer susceptibility[5]. DOCK8, encoding a guanine nucleotide exchange factor highly expressed in lymphocytes, regulates actin cytoskeleton, which is critical for migration through collagen-dense tissues such as skin[6]. Analyzing deep metagenomic sequencing data from DOCK8-deficient skin samples demonstrated a notable increase in eukaryotic viral representation and diversity compared with healthy volunteers. De novo assembly approaches identified hundreds of novel human papillomavirus genomes, illuminating microbial dark matter. Expansion of the skin virome in DOCK8-deficient patients underscores the importance of immune surveillance in controlling eukaryotic viral colonization and infection.

Viruses are a significant and abundant component of the human microbiome[7,8]. Advances in sequencing purified viral-like particles, shotgun metagenomics, and metatranscriptomics have provided the ability to unearth novel viruses[7,9–11]. In previous human skin microbiome studies, healthy skin demonstrated low viral representation[1–3], with bacteriophage dominating the skin virome. Known skin-associated eukaryotic DNA viruses include human papilloma viruses (HPV), human polyomaviruses, herpesviruses, and molluscum contagiosum virus (MCV)[2,12–14].

While landmark studies have shown that microbiota activate and educate host immunity[15–18], the role of the immune system in shaping microbial communities and its contribution to disease is less well characterized. Studying patients with primary immunodeficiency (PID) provides a unique perspective on the degree to which altered immunity may influence the human microbiome and how, in turn, microbiota may interact with the host to drive disease. Skin-associated bacterial and fungal communities in patients with PID displayed increased ecological permissiveness with decreased site specificity and temporal stability[19]. We sought to investigate the potential eukaryotic virome composition and stability in a population of patients with PID with a high frequency of severe, recalcitrant cutaneous and systemic viral infections. DOCK8 deficiency is a rare autosomal recessive, combined immunodeficiency within the spectrum of hyper-IgE syndromes[5]. In addition to recurrent cutaneous and sinopulmonary infections, these patients can suffer from severe food allergies with anaphylaxis, eczematous dermatitis with bacterial skin infections, mucocutaneous candidiasis, and cancer susceptibility. The cutaneous viral-associated manifestations of DOCK8 deficiency include molluscum contagiosum, herpesvirus infections, and warts (HPV infections) along with atopic dermatitis (Fig. 1a and ref. [5]).

To explore viral diversity capable of colonizing human skin, we collected a total of 188 skin samples from multiple body sites of 27 DOCK8-deficient patients (17 adults and 10 children; Supplementary Table 1) from which we obtained 281 gigabases (Gb) of microbial DNA sequence data (Supplementary Table 2). These shotgun metagenomics microbial reads were mapped to a multikingdom reference database and compared with similar skin metagenomics data from five representative healthy volunteers (three adults and two children[1,20]). Skin microbiome of healthy adults and children is dominated by bacteria with low viral abundance (8.7% mean phage abundance and 6.4% mean eukaryotic viral abundance; Supplementary Table 3, also in refs [3,20]). By contrast, the skin of DOCK8-deficient patients displayed a significantly higher relative abundance of eukaryotic viruses (mean relative abundance 92.6 ± 9.2, $P = 0.00061$; Fig. 1b, Supplementary Fig. 1a, and Supplementary Table 4). Three DOCK8-deficient patients (patients D8_pt23, D8_pt24, and D8_pt27) with >25% of metagenomics reads mapping to the bacterial kingdom also had less severe systemic complications associated with DOCK8 deficiency in the first two decades of life. Skin-associated bacterial and fungal communities of DOCK8-deficient patients demonstrated similar diversity and composition when compared with healthy controls, based on shotgun metagenomics, bacterial amplicon (V1–V3 of 16S rRNA gene), and fungal amplicon (ITS1 region) datasets (Supplementary Fig. 1b,c and ref. [19]).

Five viral families were present on the skin of DOCK8-deficient patients, with a predominance of *Papillomaviridae*, *Polyomaviridae*, and *Poxviridae* (Fig. 1c,d and Supplementary Table 5). *Papillomaviridae* (4 HPV genera with 48 species, 149 types) was abundantly observed

[1]Translational and Functional Genomics Branch, National Human Genome Research Institute, NIH, Bethesda, MD, USA. [2]A full list of members and their affiliations appears at the end of the paper. [3]Laboratory of Clinical Immunology and Microbiology, National Institute of Allergy and Infectious Diseases, NIH, Bethesda, MD, USA. [4]Center for Cancer Research, National Cancer Institute, NIH, Bethesda, MD, USA. [5]Dermatology Branch, National Institute of Arthritis and Musculoskeletal and Skin Diseases, NIH, Bethesda, MD, USA. [6]These authors contributed equally: Julia A. Segre and Heidi H. Kong. *e-mail: jsegre@nhgri.nih.gov; konghe@mail.nih.gov

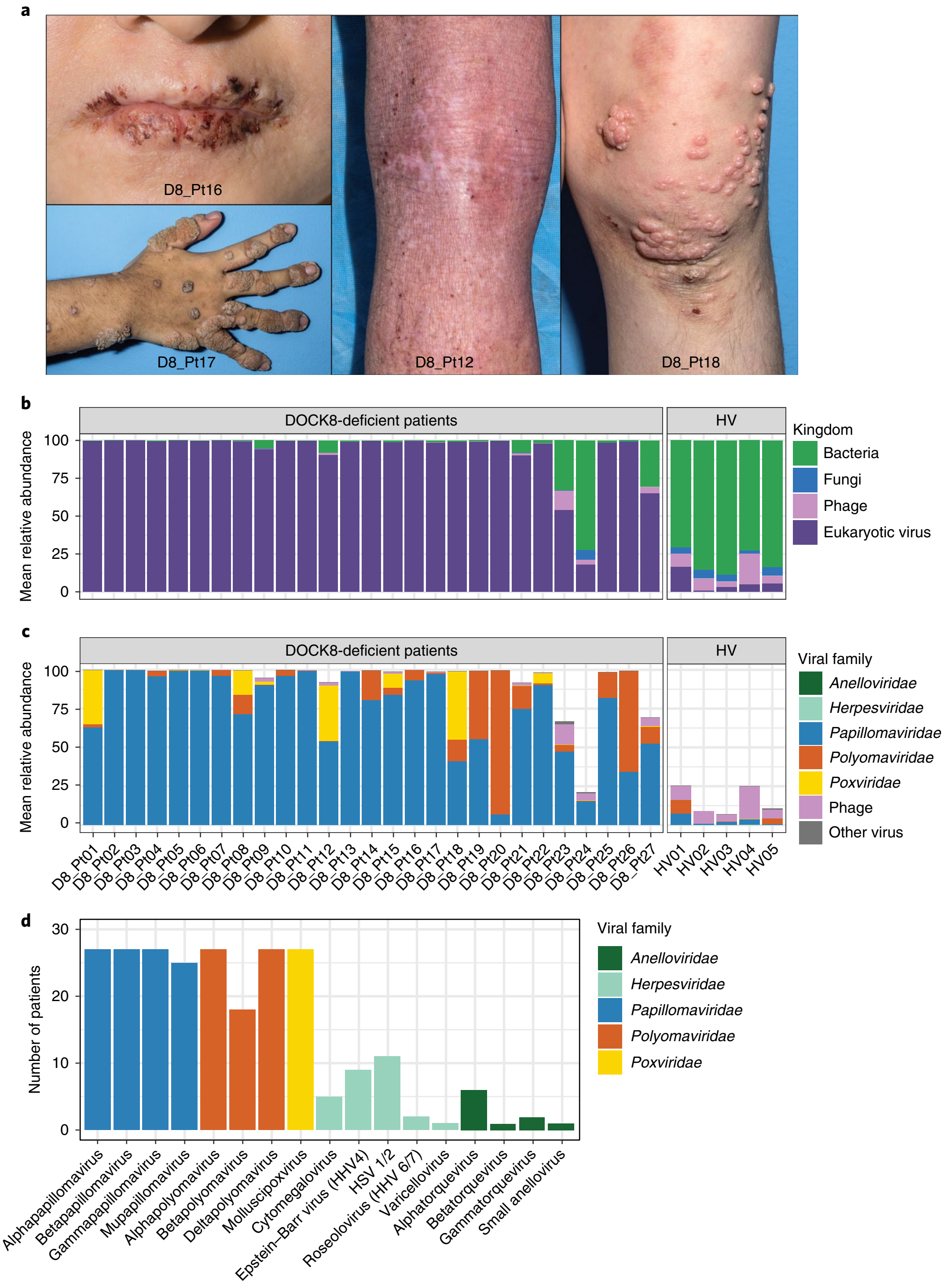

Fig. 1 | Skin microbiome of DOCK8-deficient patients is dominated by eukaryotic viruses. a, Clinical spectrum of dermatologic features of DOCK8-deficient patients: herpes simplex viral infection on the lips of D8_Pt16 (top left); warts on the hand of D8_Pt17 (bottom left); severe eczema behind the knee of D8_Pt12 (middle); and large molluscum contagiosum lesions on the knee of D8_Pt18 (right). **b**, Skin microbiome of DOCK8-deficient patients ($n=27$) and healthy volunteers (HV) ($n=5$) classified at the kingdom level. Shotgun metagenomics data presented as mean relative abundance of total mapped microbial reads (normalized to genome length) from all sampled body sites per patient. **c**, Mean relative abundance of viral families identified on skin of DOCK8-deficient patients compared with HVs. **d**, Number of DOCK8-deficient patients harboring specific eukaryotic DNA virus families.

in a majority of patients; *Polyomaviridae* (11 of 13 known species from 3 genera) was highly abundant in three patients; and *Poxviridae* (MCV) in two individuals (Fig. 1c , Supplementary Fig. 2a–c, and Supplementary Table 6). *Herpesviridae* (5 genera, 8 species including cytomegalovirus, Epstein–Barr virus, and herpes simplex virus 1), *Anelloviridae* (6 genera, 30 species), and bacteriophage were found in low relative abundances with a wide range in prevalence (Supplementary Fig. 2d and Supplementary Tables 7 and 8). In comparison with the high relative abundances of viruses on the skin of DOCK8-deficient patients, oropharyngeal swabs and stool samples from these patients did not demonstrate high viral abundances (data not shown).

Our reference-based analyses revealed a large but variable fraction of unmapped reads (10.3–98.8%, mean 46.7%). To explore our hypothesis that unmapped reads represented novel viral genomes, we performed reference-free de novo assembly of filtered reads from each skin sample after subtracting mapped bacterial and fungal reads. From this large dataset, 10,435 contigs were eukaryotic viruses with high coverage, of which 8,631 contigs were from the *Papillomaviridae* family. To identify novel viral genomes, all (near-complete) papillomavirus genomes (6–8 kilobase (kb)) were screened against the reference Papillomavirus Episteme (PaVe) database and then clustered at 95%, which resulted in 250 nonredundant HPV genomes (Supplementary Fig. 3). Incorporating these HPV genomes into our reference genome database contributed to a significant increase in mapped reads, from 53.3 ± 20.1 to 72.3 ± 18.8% ($P < 2.2 \times 10^{-16}$). At one extreme, the percentage of mapped reads increased from 1.2 to 92% with inclusion of new HPV genomes, illuminating skin microbial 'dark matter' (Supplementary Fig. 4a and Supplementary Table 9). Three additional eukaryotic viral contigs were novel genotypes of single-stranded DNA (ssDNA) Torque teno virus, a genus of *Anelloviridae* (Supplementary Fig. 4b), with 76–82% nucleotide identity with the closest reference genome for the gene used for phylogenetic classification (Supplementary Fig. 4c)[21].

HPV genomes are classified on the basis of the nucleotide sequence of the gene encoding the capsid protein L1, the most conserved protein in the *Papillomaviridae* family. HPV genomes of the same genera share ≥60% L1 similarity; of the same species share ≥71% L1 similarity; and of the same types share ≥90% L1 similarity[22,23]. Of the 250 assembled and clustered HPV genomes, 205 shared 71–89% identity to any reference HPV L1 sequence, thus representing novel HPV types. The remaining 45 genomes showed 60–70% identity to any reference HPV L1 sequence but shared >71% similarity among themselves, representing 45 novel types of a new HPV species (Supplementary Table 10). Of the 250 novel HPV genomes, 229 belonged to the gamma genus of HPV; 19 to the beta genus; and 2 to the mu genus (Fig. 2a and Supplementary Fig. 5). Previously, the mu genus of HPV included only three referenced members from three distinct species, and our genomes supplemented one of the species with two additional types.

When mapping skin microbiome reads to previously existing reference HPV genomes, the mean relative abundance of the beta genus was significantly higher than the gamma genus ($P = 0.0014$, Wilcoxon rank-sum test) on the skin of DOCK8-deficient patients. With the addition of the novel HPV genomes, beta and gamma genera were similarly abundant (Fig. 2b and Supplementary Fig. 6a,b). Overall, novel HPV types outnumbered reference HPV types detected on DOCK8-deficient patients, both in children and in adults (Fig. 2c; $P = 3.3 \times 10^{-4}$ and 1×10^{-6}, respectively). The total number of HPV types detected on each patient increased significantly on addition of novel HPV genomes ($P = 8.2 \times 10^{-10}$; Fig. 2c and Supplementary Table 11), regardless of patient age or site of sampling, with one patient colonized with over 345 different HPV types (Fig. 2d and Supplementary Fig. 6c). Intraindividual (site-to-site) similarity exceeded interindividual similarity significantly at the same site for DOCK8-deficient patients, for both HPV community membership and structure ($P = 0.00051$; Fig. 2e). The trend of viral intraindividual similarity rather than site tropism was similarly observed in healthy volunteers[3].

To assess the longitudinal stability of viral communities on the skin of DOCK8-deficient patients, we collected skin swabs from two body sites (popliteal crease, Pc; retroauricular crease, Ra) of seven patients over time. Time intervals between samplings varied from 1 to 60 months. The vast majority of the shotgun metagenomics reads derived from these samples mapped to *Papillomaviridae* (Fig. 3a). Patient D8_Pt04 had a significant increase in the relative abundance of MCV at the second time point on both skin sites (Fig. 3a), which correlated with clinical findings (Fig. 3b). Intraindividual similarity between time points was higher than interindividual similarity, reflecting the overall stability of the virome ($P = 0.00032$) (Fig. 3c). However, there were notable changes in five of the seven patients in that they both gained and lost at least five HPV types (Fig. 3d). Some of these HPV types were at modest levels but decreased from both sites simultaneously—for example, D8_Pt04 lost HPV-9 on both Pc and Ra, decreasing from 0.3 to 0% and 0.8 to 0%, respectively; and D8_Pt21 gained HPV-15, with relative abundance increasing from 0.07 to 0.3% and 0 to 0.5% on Pc and Ra, respectively (Supplementary Table 12).

While severe recalcitrant cutaneous DNA viral infections in DOCK8 deficiency suggested an expanded DNA virome, the increased incidence of viral infections in general prompted us to examine the RNA virome. We overcame the technical hurdle of extracting sufficient biomass for sequencing from noninvasive skin samples and collected a total of 207 skin and nares swabs from DOCK8-deficient patients; within-patient samples were combined based on skin physiology and/or anatomical proximity to achieve sufficient RNA for sequencing (see Methods for details). We generated 31 Gb of rRNA-subtracted, microbial RNA-seq data from four combined skin sites of 26 DOCK8-deficient patients (Supplementary Table 13). The majority of viral RNA reads mapped to DNA viruses, mostly *Papillomaviridae* and *Polyomaviridae* (Supplementary Table 14), which may indicate active viral replication. The relative abundances of the genera of polyomaviruses and papillomaviruses detected in the RNA and DNA datasets were highly similar (Supplementary Fig. 7a). The nares of DOCK8-deficient patients revealed colonization with several RNA viral families, including members of *Coronaviridae* (coronavirus species), *Picornaviridae* (rhinovirus, enterovirus, and kobuvirus species), *Orthomyxoviridae* (influenza), and *Paramyxoviridae* (parainfluenza) (Fig. 4a). Using reference-free de novo assembly of viral and unmapped RNA reads, we identified contigs with 85–97% identity to distinct reference human rhinovirus (HRV) genomes, in multiple samples from three different patients. From these, we constructed two full genomes of putative novel rhinovirus variant and genotype (Rhinovirus A genome detected in a single patient and a Rhinovirus B genome detected in two other patients), with 89 and 80% nucleotide identity to the closest genotype (*A60* and *B93*), respectively, at the VP1 gene used for phylogenetic assignment (Supplementary Fig. 7b). One DOCK8-deficient patient exhibited high relative abundances of rubella virus, whose genome carried signature mutations of the vaccine strain. Among the nine patients who had RNA viruses detected in their nares, five had undergone clinical laboratory testing for respiratory pathogens and three of these five had positive results that directly corresponded to clinical laboratory results (D8_Pt16, D8_Pt19, D8_Pt21). The other two patients had negative clinical tests and notably lower relative abundances (less than 0.5%) of rhinovirus (D8_Pt20) and parainfluenza (D8_Pt20, D8_Pt21).

DOCK8-deficient patients exhibited marked phenotypic heterogeneity. We investigated whether the presence of cutaneous warts or molluscum contagiosum lesions was associated with the relative abundances of HPV and MCV viral reads, respectively. The mean percentages of HPV and MCV reads were statistically significantly higher in the presence of any cutaneous warts (mean 89% HPV reads) and any molluscum contagiosum lesions (mean 8% MCV reads), respectively, compared with absence of these cutaneous lesions (Fig. 4b,c). However, even in the absence of any clinical warts, the DOCK8-deficient patients harbored high relative abundances of HPV reads (mean 52%). In contrast, the mean relative abundances of MCV reads in patients without clinical molluscum contagiosum lesions were quite low (mean, 0.4%). The differences in the subclinical presence of HPV versus MCV viruses suggests

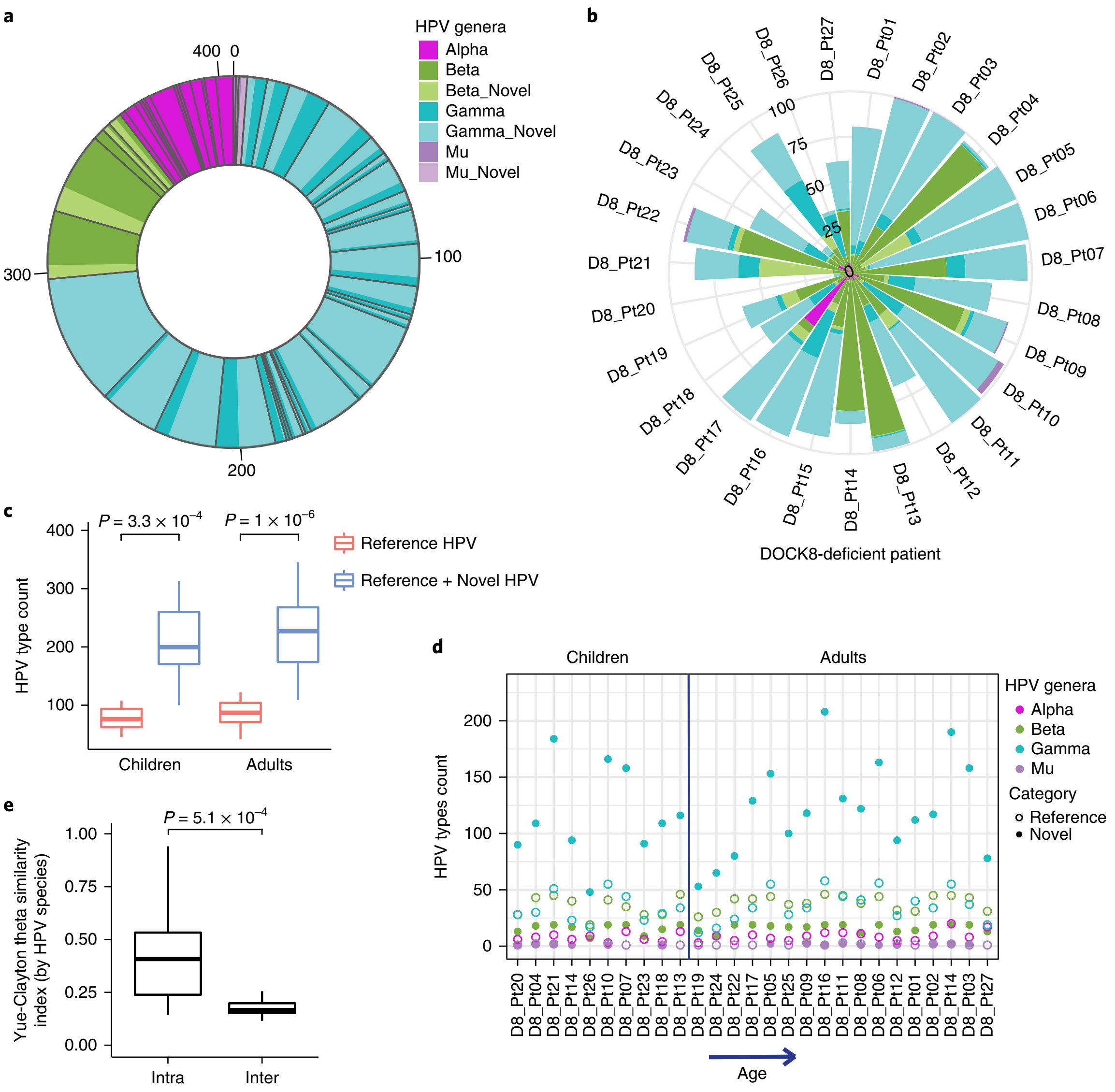

Fig. 2 | HPV diversity on skin of DOCK8-deficient patients. a, A total of 405 HPV types on the skin of all DOCK8-deficient patients ($n=27$), categorized as reference (155) or novel (250) types classified by species and genus. Each genus is depicted as a different color; pie slices represent different species within a genus and lighter shades represent new HPV types within the species. Numbers around the pie refer to the HPV type count. **b**, Skin microbiome of individual DOCK8-deficient patients: the mean relative abundances of HPV types incorporating novel HPV genomes. Relative abundance scale is on the axis of D8_Pt26. Color code as in **a**. **c**, Boxplots showing that mapping skin shotgun metagenomics reads to reference database, which includes novel HPV genomes, significantly increases the number of HPV types detected in children ($n=10$) and adults ($n=17$) throughout all sampled skin sites ($n=188$ biologically independent samples, two-sided Wilcoxon rank-sum test). **d**, Number of HPV types classified at the genus level detected in each patient. **e**, Boxplots showing mean theta similarity index of inter- and intrapersonal pairwise similarity of skin sites of DOCK8-deficient patients ($n=188$ biologically independent samples from all skin sites, two-sided Wilcoxon rank-sum test), comparing HPV species detected at relative abundance > 0.1%. Theta value of 1 indicates identical HPV community structure. All boxplots are the median with the interquartile range, and error bars are the 1.5 times interquartile range (whiskers).

biological distinctions between these eukaryotic viruses, with patient skin demonstrating less susceptibility to MCV carriage than to HPV carriage.

The presence and contribution of viruses to the human microbiome have been underappreciated. This shotgun metagenomics study of a rare patient population has elucidated the selective increases in the diverse viral communities capable of inhabiting human skin, which would not be possible by analyzing healthy individuals owing to their low viral abundances. The expansion of the skin virome in this immunodeficient patient cohort underscores the importance of immune surveillance in controlling host eukaryotic viral colonization and infection. The infectious complications in these immunodeficient patients are similar to the infection risks in immunosuppressed patients undergoing hematopoietic stem cell transplants. The viral reactivation observed in the setting of hematopoietic stem cell transplants has been described in virome studies in the gut and blood[21,24,25].

Our metagenomics analyses in this unique patient cohort have facilitated our ability to identify novel viruses and investigate microbial dark matter. Previously, studying rare patient populations led to the discovery of the first two beta HPV members (HPV-5 and HPV-8) in patients with the rare disorder epidermodysplasia verruciformis, a recessive disease with an increased susceptibility to HPV infections and squamous cell carcinoma[26]. Expanding knowledge of

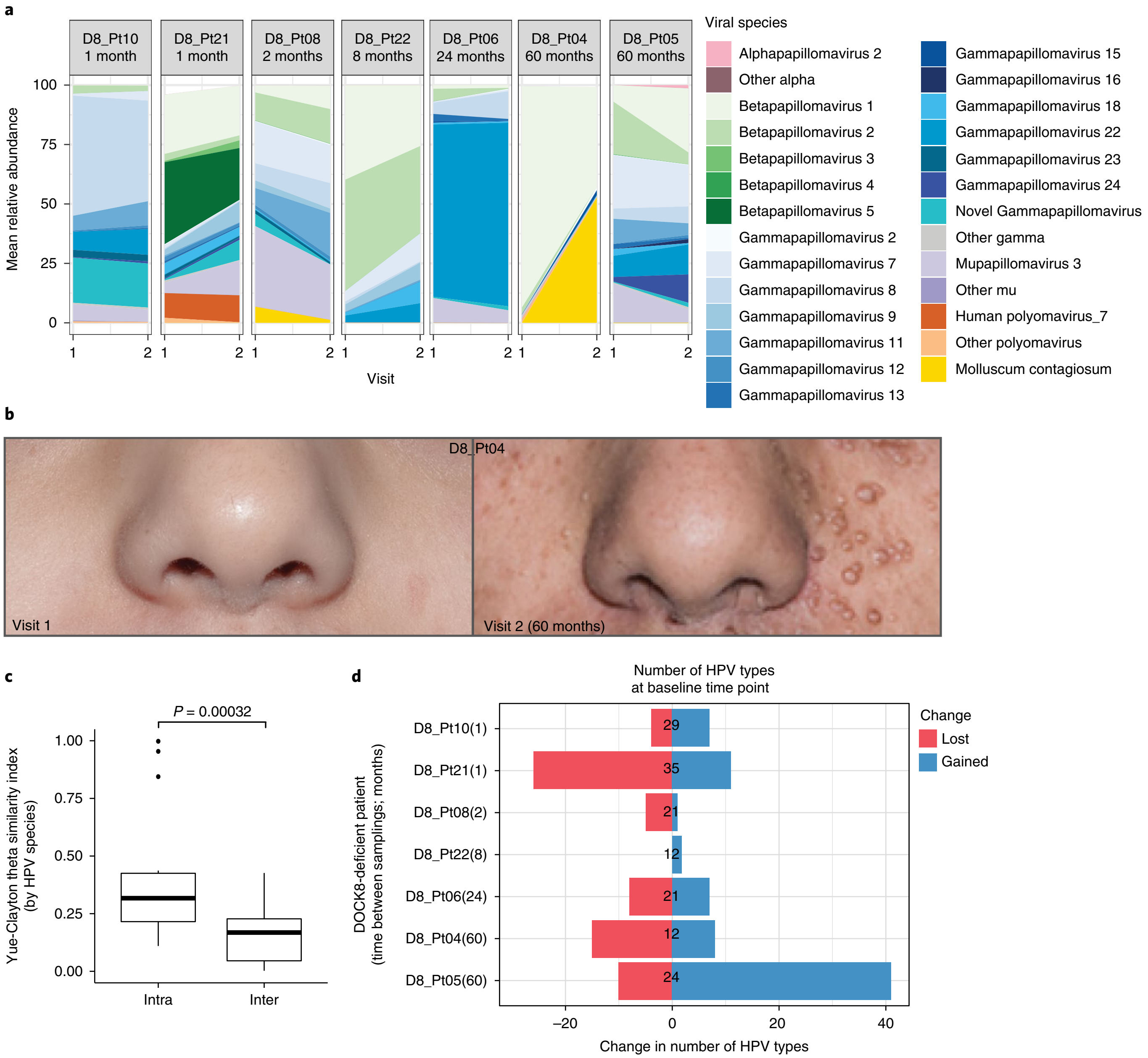

Fig. 3 | Longitudinal stability of the skin virome of DOCK8-deficient patients. a, Mean relative abundances of two skin sites for the three most abundant viral families (*Papillomaviridae*, *Polyomaviridae*, *Poxviridae*) shown at the species level in patients ($n = 7$) sampled longitudinally. Baseline and follow-up time points (first and second visit) are represented by 1 and 2, respectively. Patients ordered by months between samplings (from shortest to longest). **b**, Absence (visit 1, baseline) and presence (visit 2, follow-up, 60 months) of molluscum contagiosum on the face of D8_Pt4. **c**, Boxplots showing mean theta similarity index of inter- and intrapersonal pairwise similarity of patients at baseline and follow-up time points ($n = 28$ biologically independent samples, two-sided Wilcoxon rank-sum test), comparing HPV species detected in relative abundances >0.1%. All boxplots are the median with the interquartile range, and the error bars are the 1.5 times interquartile range (whiskers) and outliers (points). **d**, Longitudinal changes include gain and loss of HPV types (for types with relative abundance > 0.1%) and are patient specific. Duration of time between sampling time points (measured in months) is shown in parentheses next to each patient. Numbers on bars represent count of HPV types at first baseline sampling.

HPV diversity may promote the development of computational tools for predicting and assigning unknown sequences from metagenomics surveys and of molecular assays for rapid identification of viruses with oncogenic potential, such as some members of the HPV community. As DOCK8-deficient patients are susceptible to squamous cell cancers, future studies may be able to address whether HPVs detected on DOCK8-deficient skin have oncogenic potential.

Studying the microbiome with shotgun metagenomics has enabled trans-kingdom explorations. The Human Microbiome Project has focused attention on the beneficial role played by human-associated bacteria; might there also be commensal or beneficial viruses? Recent elegant studies have elucidated that the virome plays a role in the outcome of an immune response to a vaccine challenge[27]. Some evidence of mutualistic interactions between viruses and plants, insects, and even mammals exists[28], but the underlying mechanisms of these interactions still need to be explored. This study reveals another aspect of the intricate interactions between the immune system and the microbiome, demonstrating how a

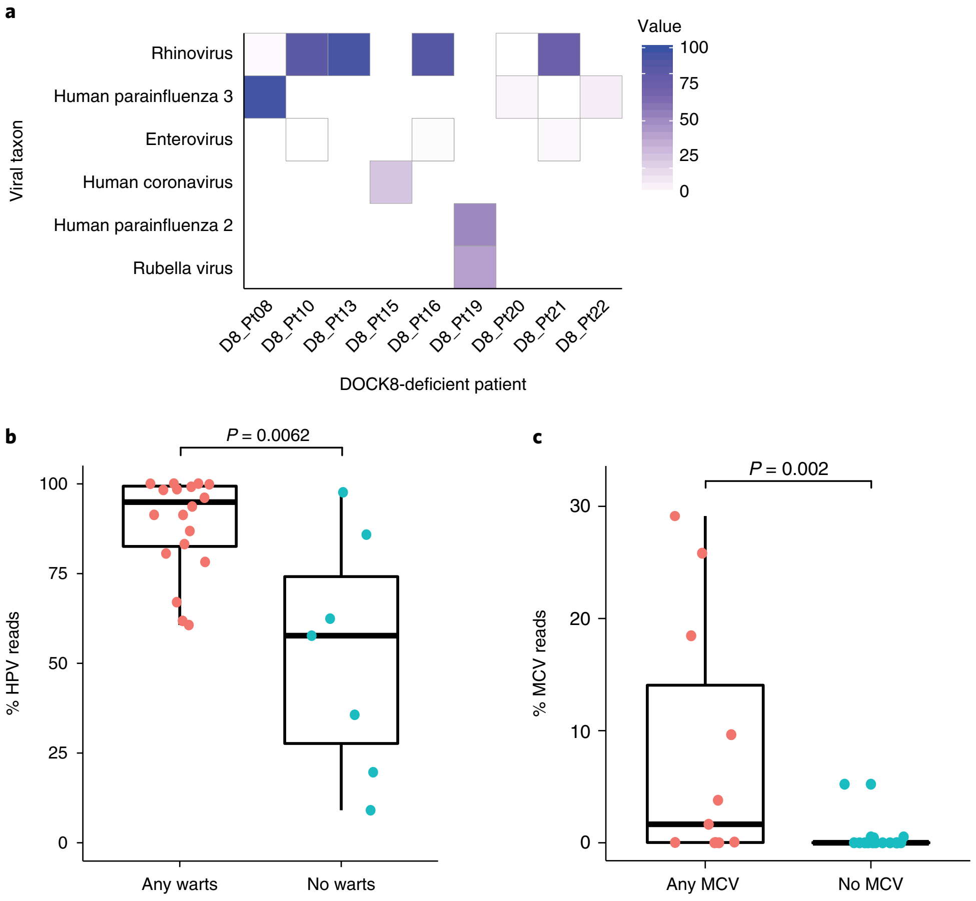

Fig. 4 | Human RNA viruses in DOCK8-deficient patients and comparison between viral presence in sequencing data and cutaneous lesions. **a**, Relative abundance heatmap of human RNA viral species detected in the nares of DOCK8-deficient patients. **b**, Boxplots comparing mean percentage of HPV reads and patients that presented with any warts on their skin ($n = 27$ patients, two-sided Wilcoxon rank-sum test). **c**, Boxplots comparing mean percentage of MCV reads and patients that presented with any molluscum contagiousum lesions on their skin ($n = 27$ patients, two-sided Wilcoxon rank-sum test). All boxplots are the median with the interquartile range, and error bars are the 1.5 times interquartile range (whiskers).

specific underlying immunodeficiency may affect the trans-kingdom equilibrium, thus pointing to the role that the immune system plays in shaping microbial communities in the human body.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41591-018-0211-7.

Received: 6 June 2018; Accepted: 5 September 2018;
Published online: 5 November 2018

References

1. Oh, J. et al. Biogeography and individuality shape function in the human skin metagenome. *Nature* **514**, 59–64 (2014).
2. Hannigan, G. D. et al. The human skin double-stranded DNA virome: topographical and temporal diversity, genetic enrichment, and dynamic associations with the host microbiome. *mBio* **6**, e01578-15 (2015).
3. Oh, J. et al. Temporal stability of the human skin microbiome. *Cell* **165**, 854–866 (2016).
4. Picard, C. et al. Primary immunodeficiency diseases: an update on the classification from the International Union of Immunological Societies Expert Committee for Primary Immunodeficiency 2015. *J. Clin. Immunol.* **35**, 696–726 (2015).
5. Zhang, Q. et al. Combined immunodeficiency associated with DOCK8 mutations. *N. Engl. J. Med.* **361**, 2046–2055 (2009).
6. Zhang, Q. et al. DOCK8 regulates lymphocyte shape integrity for skin antiviral immunity. *J. Exp. Med.* **211**, 2549–2566 (2014).
7. Paez-Espino, D. et al. Uncovering Earth's virome. *Nature* **536**, 425–430 (2016).
8. de la Cruz Pena, M. J. et al. Deciphering the human virome with single-virus genomics and metagenomics. *Viruses* **10**, e113 (2018).
9. Reyes, A. et al. Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature* **466**, 334–338 (2010).
10. Shi, M. et al. The evolutionary history of vertebrate RNA viruses. *Nature* **556**, 197–202 (2018).
11. Shi, M. et al. Redefining the invertebrate RNA virosphere. *Nature* **540**, 539–543 (2016).
12. Foulongne, V. et al. Human skin microbiota: high diversity of DNA viruses identified on the human skin by high throughput sequencing. *PLoS ONE* **7**, e38499 (2012).
13. Schowalter, R. M., Pastrana, D. V., Pumphrey, K. A., Moyer, A. L. & Buck, C. B. Merkel cell polyomavirus and two previously unknown polyomaviruses are chronically shed from human skin. *Cell Host Microbe* **7**, 509–515 (2010).
14. Chen, X., Anstey, A. V. & Bugert, J. J. Molluscum contagiosum virus infection. *Lancet Infect. Dis.* **13**, 877–888 (2013).
15. Naik, S. et al. Compartmentalized control of skin immunity by resident commensals. *Science* **337**, 1115–1119 (2012).
16. Jiang, X. et al. Skin infection generates non-migratory memory CD8 + T(RM) cells providing global skin immunity. *Nature* **483**, 227–231 (2012).
17. Meisel, J. S. et al. Commensal microbiota modulate gene expression in the skin. *Microbiome* **6**, 20 (2018).
18. Gilbert, J. A. et al. Current understanding of the human microbiome. *Nat. Med.* **24**, 392–400 (2018).
19. Oh, J. et al. The altered landscape of the human skin microbiome in patients with primary immunodeficiencies. *Genome Res.* **23**, 2103–2114 (2013).

20. Byrd, A. L. et al. Staphylococcus aureus and Staphylococcus epidermidis strain diversity underlying pediatric atopic dermatitis. *Sci. Transl. Med.* **9**, eaal4651 (2017).
21. Kowarsky, M. et al. Numerous uncharacterized and highly divergent microbes which colonize humans are revealed by circulating cell-free DNA. *Proc. Natl Acad. Sci. USA* **114**, 9623–9628 (2017).
22. de Villiers, E. M., Fauquet, C., Broker, T. R., Bernard, H. U. & zur Hausen, H. Classification of papillomaviruses. *Virology* **324**, 17–27 (2004).
23. Bernard, H. U. et al. Classification of papillomaviruses (PVs) based on 189 PV types and proposal of taxonomic amendments. *Virology* **401**, 70–79 (2010).
24. De Vlaminck, I. et al. Temporal response of the human virome to immunosuppression and antiviral therapy. *Cell* **155**, 1178–1187 (2013).
25. Legoff, J. et al. The eukaryotic gut virome in hematopoietic stem cell transplantation: new clues in enteric graft-versus-host disease. *Nat. Med.* **23**, 1080–1085 (2017).
26. de Jong, S. J. et al. Epidermodysplasia verruciformis: inborn errors of immunity to human beta-papillomaviruses. *Front. Microbiol.* **9**, 1222 (2018).
27. Reese, T. A. et al. Sequential infection with common pathogens promotes human-like immune gene expression and altered vaccine response. *Cell Host Microbe* **19**, 713–719 (2016).
28. Roossinck, M. J. Move over, bacteria! Viruses make their mark as mutualistic microbial symbionts. *J. Virol.* **89**, 6532–6535 (2015).

Acknowledgements

This study utilized the high-performance computational capabilities of the NIH Biowulf Linux cluster (http://hpc.nih.gov). We thank M. Park, P. Thomas, A. Young, S. Phang, A. Pradhan, V. Pillai, J. Fekecs, NIH Patient Photography, and W.-I. Ng for underlying efforts, and the Segre and Kong laboratories for helpful discussions. We appreciate the participation of patients and their families. This work was supported by the National Human Genome Research Institute, National Institute of Allergy and Infectious Diseases, National Cancer Institute, and National Institute of Arthritis and Musculoskeletal and Skin Diseases Intramural Research Programs.

Author contributions

O.T., S.C., H.C.S., A.F.F., J.A.S., and H.H.K. contributed to the design and conception of the study. Sequencing was carried out by NISC. O.T., S.C., C.D., S.-Q.L.-L., and X.H. performed the experiments and analyses. O.T., J.A.S., and H.H.K. drafted the manuscript. All authors revised the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41591-018-0211-7.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to J.A.S. or H.H.K.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

NISC Comparative Sequencing Program

Beatrice B. Barnabas[7], Gerard G. Bouffard[7], Shelise Y. Brooks[7], Holly Marfani[7], Lyudmila Dekhtyar[7], Xiaobin Guan[7], Joel Han[7], Shi-ling Ho[7], Richelle Legaspi[7], Quino L. Maduro[7], Catherine A. Masiello[7], Jennifer C. McDowell[7], Casandra Montemayor[7], James C. Mullikin[7], Morgan Park[7], Nancy L. Riebow[7], Karen Schandler[7], Chanthra Scharer[7], Brian Schmidt[7], Christina Sison[7], Sirintorn Stantripop[7], James W. Thomas[7], Pamela J. Thomas[7], Meghana Vemulapalli[7] and Alice C. Young[7]

[7]NIH Intramural Sequencing Center, National Human Genome Research Institute, Bethesda, MD, USA

Methods

Subject recruitment and sampling. Twenty-seven DOCK8-deficient patients (17 adults and 10 children; mean age 16.0 ± 6.1 years) were recruited to participate in a study approved by the institutional review board of the National Human Genome Research Institute (www.clinicaltrials.gov/ct2/show/NCT00605878) that complied with all ethical regulations. Written informed consent was obtained from all adult patients and parents or guardians of all participating children. Patients were diagnosed with DOCK8 deficiency based on targeted gene sequencing and immunoblotting[5,6]. Subjects provided medical and medication history and underwent a physical examination. For shotgun metagenomics sequencing, eight skin sites representing diverse physiological characteristics were sampled: moist (antecubital crease, Ac; inguinal crease, Ic; popliteal crease, Pc; plantar heel, Ph); dry (hypothenar palm, Hp; volar forearm, Vf); sebaceous (manubrium, Mb; retroauricular crease, Ra). For RNA-seq, swabs from adjacent anatomical sites or from sites with similar physiological characteristics were combined to provide sufficient biomass. The sites sampled and combined were Ra and occiput (Oc) representing the head, Ac/Pc/Ic representing moist creases, Ph and toe web (Tw) representing the foot, and nares (N). Patient metadata are presented in Supplementary Table 1. Negative control swabs were collected for each patient visit.

Sample processing and sequencing. DNA isolation and library preparation to generate shotgun metagenomic sequence data from skin sites were performed as previously described with a target of 15 million to 50 million 2 × 125-bp reads on an Illumina HiSeq[1]. In total, for DNA metagenomics sequencing from 27 patients sampled at up to 8 body sites, we obtained 202 skin swabs and 2.74 billion reads (or 281 Gb) of nonhuman, quality filtered, paired-end reads (median 2.23 million reads; or 217 Mb, per sample, Supplementary Table 2).

Samples for RNA extraction were collected in Yeast Cell Lysis Buffer (Lucigen) and treated with proteinase K (Invitrogen) for 5 min at 37 °C. Buffer AVL, from QIAamp Viral RNA Mini Kit (Qiagen), was added along with linear acrylamide (Invitrogen) at room temperature for 15 min and processed according to the manufacturer's instructions. Extracted RNA was DNAse (Qiagen)–treated on column and eluted with 40 µl of DEPC-treated water. RNA was fluorometrically quantitated on the Qubit (ThermoFisher), and cDNA was synthesized and amplified using Ovation RNA-seq system V2 (Nugen). From 26 patients sampled at 9 body sites, we obtained 207 skin swabs, combined as described to 116 samples and 329 million reads (or 31 Gb) of nonhuman, quality filtered reads, non-rRNA paired-end reads (median 717,000 reads; or 65 Mb, per sample; Supplementary Tables 13 and 15). To remove rRNA reads, all RNA microbial reads were mapped against the SILVA database of bacterial, eukaryotic, and archaeal rRNA using SortMeRNA[29] (Supplementary Table 15) and non-rRNA reads were written out and used for further taxonomic classification. DNA and RNA preparations varied in biomass and composition. For example, human-derived DNA and RNA accounted for 74 and 53% of reads on average, respectively (Supplementary Tables 2 and 14). For RNA-Seq, 73% of reads on average represented rRNA (eukaryotic/bacterial) (Supplementary Table 15).

Amplicon sequencing. DOCK8-deficient patients' sequencing libraries were prepared based on a previously described strategy[30]. The V1–V3 region of the 16S rRNA gene was amplified using the modified primers 27F- (5′-AGAGTTTGATCCTGGCTCAG-3′) and 534R- (5′-ATTACCGCGGCTGCTGG-3′). The ITS1 region was amplified using the modified primers 18S-F (5′-GTAAAAGTCGTAACAAGGTTTC-3′) and 5.8S-1R (5′-GTTCAAAGAYTCGATGATTCAC-3′). 16S and ITS1 amplicons were generated and processed as previously described[30,31] and sequenced on an Illumina MiSeq instrument. Amplicon sequencing data were processed using the mothur pipeline[32] as previously described[19,31]. Briefly, 16S sequences were preprocessed to remove primers and barcodes and subsampled to 5,000 sequences per sample. Chimeras from PCR artifacts were identified and removed using VSEARCH in mothur[33,34]. Remaining sequences were classified to the genus level using RDP training set 16. ITS1 amplicon sequences were trimmed to 200 bp and subsampled to 5,000 reads per sample in mothur. Sequences were then classified to genus level using the *k*-nearest neighbor classifier against an updated custom ITS1 database[35]. Phylotype analysis for both 16S and ITS1 amplicon sequences was performed according to mothur MiSeq SOP (https://www.mothur.org/wiki/MiSeq_SOP).

Reference-based taxonomic classification. Taxonomic classifications were performed as previously described[1]. Quality processed reads not matching hg19 human reference were mapped against a database of 2,349 bacterial, 389 fungal, 4,695 viral, and 67 archaeal reference genomes using Bowtie 2 (version 2.3.2)–very-sensitive parameter[36]. Read hit counts were normalized by genome size. To reduce the effect of low abundance misclassifications, we used a genome coverage cutoff of ≥1 for relative abundance and diversity calculations.

Reference-free de novo assembly and contig analysis. Metagenomic reads not mapping to bacterial, fungal, or archaeal reference genomes were exported using the –un-conc parameter of Bowtie 2 (ref. [36]). DNA or RNA reads from each sample separately (mean of 6 and 1.2 million reads per sample, respectively, representing 90% of mean microbial DNA reads and 80% of RNA microbial reads per sample), were used in de novo assembly of viral genomes using SPAdes genome assembler (version 3.9), with –meta parameter for DNA data and –rna parameter for RNA designed to handle metagenomics or RNA sequencing data, respectively, with k-mer lengths of 21, 33, 55, 77 (refs [37,38]).

A total of 349,919 contigs were assembled from the filtered DNA reads (Supplementary Table 16). Only contigs larger than 750 nucleotides (nt) were selected for further quality filtration using PRINSEQ[39] to discard low entropy contigs (-lc_threshold 70) and extreme GC content (below 25% or above 75%). The 227,924 filtered contigs were assigned to a taxon based on the BLASTN best hit (BLAST + suite v2.6.0 (refs [40,41])) against the nucleotide database (requiring that the best hit cover ≥ 25% of the contig). From the total, 10,435 contigs were assigned to eukaryotic viruses, 1,189 contigs were assigned to phage, 125,755 contigs had nonviral assignments (mostly bacteria), and 90,545 contigs were unassigned under these parameters but were reassigned using a more flexible set of parameters (best hit, but without a threshold for query coverage) and were mostly bacterial.

Of the 10,435 viral contigs, 8,631 belonged to the *Papillomaviridae* family. To identify divergent papillomavirus genomes, we eliminated the coverage requirement and identified another 183 contigs as likely *Papillomaviridae* on the basis of contig size (~7 kb) and L1 typing. A similar search for divergent *Polyomaviridae* returned no additional contigs. From 8,914 *Papillomaviridae* contigs, we subsetted 3,725 that were within a size range of 6–8 kb, which is similar to the size of an average HPV genome (~7.9 kb). These genomes were masked using an HPV reference downloaded from the PaVe database (ref. [42] and www.pave.niaid.nih.gov) to remove all regions ≥ 500 nt with more than 90% identity by BLASTN. Genomes with ≥ 4,000 nt unmasked were kept, resulting in 2,189 genomes. *Papillomaviridae* have circular genomes, which is problematic for alignment-based clustering algorithms, so the Mash algorithm[43] was used instead to build a graph of 95% identical contigs that were then clustered using SPICi[44]. This threshold was set to remove redundant genomes but still retain genomes that are distant enough to represent distinct HPV types. This clustering resulted in 208 cluster seeds and 42 singleton genomes, yielding a final pool of 250 nonredundant HPV genomes. Using BLASTN to compare these 250 nonredundant genomes with themselves, we verified that their nucleotide identity between cluster representatives did not exceed 90%.

The remaining 1,804 contigs assigned to non-*Papillomaviridae* eukaryotic viruses were clustered at 99% using CD-HIT[45,46] and masked against our viral database (masking parameters as previously described). This resulted in 23 *Poxviridae* contigs that were mostly masked, 6 *Herpesviridae* contigs (size range 750–1800 nt), all assembled from the same sample (patient D8_Pt21) that mapped to cytomegalovirus genomes not included in our database, and 4 *Anelloviridae* contigs (Supplementary Fig. 4b). Three of the *Anelloviridae* contigs (in the genus *Torque teno virus*) were assembled from two samples from patient D8_Pt13, with contig sizes of 1,980, 1,270, and 1,599 nt. The fourth contig was assembled from a sample from a different patient (contig size 1,357 nt, D8_Pt11). These contigs were amplified by PCR using targeted primers from the original DNA extracted from the skin swab and sequences verified with Sanger sequencing. Owing to low coverage of these contigs, we were not able to amplify and assemble a complete genome of these putative novel anelloviruses. A novel anellovirus genotype has been defined as the ORF1 gene having < 90% shared nucleotide identity when aligned to any reference genotype, whereas designation as a novel anellovirus species requires < 50% sequence identity. Assembled contigs were aligned to reference genomes in our viral database using MUSCLE[47] v3.8.31 to determine the phylogenetic relationships of the viruses (Supplementary Fig. 4c). The resulting alignment was used to construct a maximum-likelihood phylogenetic tree, using FastTree[48] v2.1.9 and visualized with FigTree (http://tree.bio.ed.ac.uk/software/figtree/).

Unassigned contigs (90,545) were assigned to best hit using the BLASTN/BLASTX modes against the nonredundant (nt/nr) databases, and most were partially matched to bacteria or phage.

RNA reads filtered from rRNA, human, bacterial, and fungal reads were assembled de novo in the same manner. However, since rRNA reads represented a significant portion of the data, some of the samples had a very low number of reads (the lowest sample having only 35,331 filtered reads). As a consequence, the assembly yielded a large number of contigs (625,306) with only about one-third passing quality filtration (206,371) (Supplementary Table 17). Of these, 122,771 were assigned according to our contig analysis pipeline, and only 24,217 were a hit to a viral genome. Almost all viral contigs were assigned to DNA viruses (mainly *Papillomaviridae*), in accordance with the high abundance we detected in read mapping from both DNA and RNA data. A total of 1,458 contigs were assigned to RNA viruses. We detected several contigs from the rhinovirus family, in patients D8_Pt16, D8_Pt21, and D8_Pt22. Contigs from patient D8_Pt16 were identified as a variant of *Rhinovirus A60*, and contigs from patients D8_Pt21 and D8_Pt22 were identified as a variant of *Rhinovirus B97*. The *Picronaviridae* study group defines novelty of a rhinovirus genotype by the coding sequence of the conserved VP1 gene, with slight differences between the species; an HRV type should have at least 13% (HRV-A) or 12% (HRV-B) nucleotide divergence from all other HRV types. Nucleotide divergence lower than this threshold has been used to assign new variants of a genotype[49,50]. With primers designed at terminal contig sequences and cDNA derived from the original patient RNA, we amplified across the gaps, Sanger sequenced, and assembled two complete novel rhinovirus variants and genotypes (~7 kb), as classified by the conserved VP1 protein[49,50] (Supplementary Fig. 7b). Contigs that were unassigned at first were blasted again with adjusted parameters in BLASTN and BLASTX modes against the nonredundant databases and were predominantly partially matched to bacterial or eukaryote genomes and proteins.

Taxonomic classification of novel HPV genomes. Taxonomic classification of papillomaviruses is based on nucleotide similarity of the L1 gene[51]. The family *Papillomaviridae* contains 49 genera (with 5 genera representing human papillomaviruses), each of which is further divided into several species. To be designated as a new type, a single papillomavirus type cannot share >90% similarity to any other known papillomavirus type in the L1 sequence. Papillomavirus types within a species share 71–89% nucleotide identity within the L1 gene and members of the same genus share >60% L1 sequence identity[22,23]. Taxonomic classification of the 250 novel HPV genomes assembled from reads and meeting the size criteria described above (6–8 kb) was performed using the L1 taxonomy tool available on the PaVe website (https://pave.niaid.nih.gov/#analyze/l1_taxonomy_tool) (Supplementary Table 10). Of the 250 genomes, 205 were depicted as novel types, exhibiting less than 90% identity to any reference HPV genome, and 45 genomes were depicted as members of a single novel species, as their L1 sequence shared 60–70% similarity with any reference HPV and 71–89% identity when compared among themselves. All 45 members of the novel species were verified by targeted PCR to the L1 region from the original DNA extracted from the skin swabs, as well as additional 10 genomes, classified as novel HPV types, that were not represented in clusters (and were part of the 42 singleton genomes in the clustering process).

Phylogenetic tree construction. The nucleotide sequences of the HPV L1 gene were extracted from all novel HPV genomes using NCBI's orfFinder tool[52]. All HPV L1 nucleotide sequences of reference HPV genomes (as determined by the International Human Papillomavirus Reference Center) were downloaded from the PaVe database (ref. [42] and www.pave.niaid.nih.gov). All L1 sequences were aligned using MUSCLE v3.8.31 (ref. [47]) and the resulting alignment was used to construct a maximum-likelihood phylogenetic tree, using FastTree v2.1.9 (ref. [48]). The tree was visualized with FigTree (http://tree.bio.ed.ac.uk/software/figtree/).

SNP analysis of MCV. MCV abundance, subtype, and coverage were determined using Clinical Pathoscope[53] and a database of five complete MCV genomes (U60315, KY040275, KY040276, KY040277, KY040274). MCV-1 SNPs were called against U60315 using Snippy (https://github.com/tseemann/snippy) with default parameters on samples with at least 15 times coverage of the MCV-1 genome (49 samples, 10 subjects). Core SNPs were called using the snippy-core program (244 core SNP positions). A histogram of SNP positions identified uneven SNP coverage, indicative of recombination. Gubbins[54] was used to analyze recombination blocks between MCV-1 genomes and identified one block centered on the MC054L gene, which encodes a putative IL-18 binding protein[55,56]. Of the two DOCK8-deficient patients who did not display clonal carriage, D8_Pt08 showed two distinct MCV-1 strains, each exclusive to distinct body sites, and D8_Pt15 exclusively carried MCV-1 at two body sites and MCV-2 at three other body sites, demonstrating the potential for localized skin colonization from distinct subtypes. A phylogenetic tree was built based on SNPs outside these recombinant regions.

Statistical analysis. All statistical analyses were performed using R software. Data are represented as mean ± s.e.m. unless otherwise indicated. Spearman correlations of nonzero values were used for all correlation coefficients. For all boxplots, center lines represent the median, lower and upper box limits represent the first and third quartiles, respectively (interquartile range), whiskers represent the maximal values up to 1.5 times interquartile range, and all values beyond this range are defined as outliers. The nonparametric Wilcoxon rank-sum test was used to determine statistically significant differences between microbial populations. Unless otherwise indicated, P values were adjusted for multiple comparisons using the Bonferroni or FDR correction (by applying the p.adjust function in R using method = 'bonferroni' or 'fdr'). Statistical significance was ascribed to an alpha level of the adjusted $P \leq 0.05$. Similarity between samples was assessed using the Yue–Clayton theta similarity index with relative abundances of HPV species. The theta coefficient assesses the similarity between two samples based on (i) number of features in common between two samples, and (ii) their relative abundances, with $\theta = 0$ indicating totally dissimilar communities and $\theta = 1$ identical communities[57].

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The sequencing data and genome assemblies for this study are linked to the NCBI BioProject ID PRJNA471898.

References

29. Kopylova, E., Noe, L. & Touzet, H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28**, 3211–3217 (2012).
30. Fadrosh, D. W. et al. An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform. *Microbiome* **2**, 6 (2014).
31. Jo, J. H. et al. Diverse human skin fungal communities in children converge in adulthood. *J. Invest. Dermatol.* **136**, 2356–2363 (2016).
32. 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).
33. Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C. & Knight, R. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27**, 2194–2200 (2011).
34. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**, e2584 (2016).
35. Findley, K. et al. Topographic diversity of fungal and bacterial communities in human skin. *Nature* **498**, 367–370 (2013).
36. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
37. Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
38. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).
39. Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**, 863–864 (2011).
40. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
41. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
42. Van Doorslaer, K. et al. The papillomavirus episteme: a central resource for papillomavirus sequence data and analysis. *Nucleic Acids Res.* **41**, D571–D578 (2013).
43. Ondov, B. D. et al. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).
44. Jiang, P. & Singh, M. SPICi: a fast clustering algorithm for large biological networks. *Bioinformatics* **26**, 1105–1111 (2010).
45. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
46. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
47. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
48. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
49. McIntyre, C. L., Knowles, N. J. & Simmonds, P. Proposals for the classification of human rhinovirus species A, B and C into genotypically assigned types. *J. Gen. Virol.* **94**, 1791–1806 (2013).
50. Simmonds, P. et al. Proposals for the classification of human rhinovirus species C into genotypically assigned types. *J. Gen. Virol.* **91**, 2409–2419 (2010).
51. de Villiers, E. M. Cross-roads in the classification of papillomaviruses. *Virology* **445**, 2–10 (2013).
52. Wheeler, D. L. et al. Database resources of the National Center for Biotechnology. *Nucleic Acids Res.* **31**, 28–33 (2003).
53. Byrd, A. L. et al. Clinical PathoScope: rapid alignment and filtration for accurate pathogen identification in clinical samples using unassembled sequencing data. *BMC Bioinformatics* **15**, 262 (2014).
54. Croucher, N. J. et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res.* **43**, e15 (2015).
55. Lopez-Bueno, A., Parras-Molto, M., Lopez-Barrantes, O., Belda, S. & Alejo, A. Recombination events and variability among full-length genomes of co-circulating molluscum contagiosum virus subtypes 1 and 2. *J. Gen. Virol.* **98**, 1073–1079 (2017).
56. Xiang, Y. & Moss, B. Molluscum contagiosum virus interleukin-18 (IL-18) binding protein is secreted as a full-length form that binds cell surface glycosaminoglycans through the C-terminal tail and a furin-cleaved form with only the IL-18 binding domain. *J. Virol.* **77**, 2623–2630 (2003).
57. Yue, J. C. & Clayton, M. K. A similarity measure based on species proportions. *Comm. Stat. – Theor. M.* **34**, 2123–2131 (2005).

nature research

Corresponding author(s): Julia A. Segre and Heidi H. Kong

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
□	☒	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	□	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
□	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	no software was used
Data analysis	Bowtie2 (version 2.3.2), Clinical Pathoscope, SPAdes genome assembler (version 3.9), BLAST+ suite v2.6.0, SPICi, Mash, CD-HIT, PRINSEQ

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The sequencing data for this study are linked to the NCBI BioProject ID PRJNA471898

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences □ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Since our IRB-approved natural history protocol recruits and follows rare primary immunodeficiency syndromes, sample sizes were based on specimen availability. We targeted a sample size of n>10 to account for sample variability.
Data exclusions	We did not exclude any data.
Replication	Not applicable. 27 human DOCK8-deficient patients were enrolled and sampled. For a small number of patients, a second sampling was performed and shown in Figure 4.
Randomization	Not applicable. This was a natural history study.
Blinding	Not applicable. This was a natural history study.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study	
☒	□	Unique biological materials
☒	□	Antibodies
☒	□	Eukaryotic cell lines
☒	□	Palaeontology
☒	□	Animals and other organisms
□	☒	Human research participants

Methods

n/a	Involved in the study	
☒	□	ChIP-seq
☒	□	Flow cytometry
☒	□	MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	27 participants, males and females, ages 6-30 years. All with DOCK8-deficiency due to mutation in the DOCK8 gene which was verified in target gene sequencing or immuno-blotting.
Recruitment	Clinical study personnel who oversee genetic diagnosis and disease management referred DOCK8-deficient patients who were interested in participating in a microbiome study. Any DOCK8-deficient patient who consented was included in sampling and analysis. The authors affirm that human research participants provided informed consent for publication of the images in Figures 1 and 3.