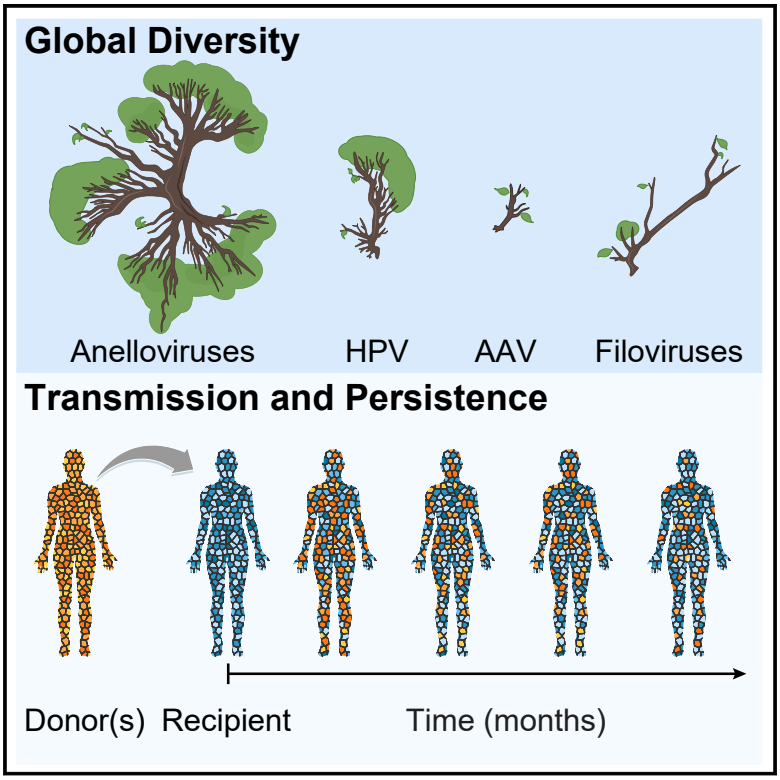
Resource

Cell Host & Microbe

Global genome analysis reveals a vast and dynamic anellovirus landscape within the human virome

Graphical abstract



Authors

Cesar A. Arze, Simeon Springer, Gytis Dudas, ..., Roger J. Hajjar, Kristian G. Andersen, Nathan L. Yozwiak

Correspondence

nyozwiak@ringtx.com

In brief

Anelloviruses comprise a major component of the healthy human virome. Arze et al. describe the extensive diversity of anellovirus genomes (the "anellome") in the blood of transfusion donors and recipients and uncover the dynamics of anellovirus transmission and persistence over several months after transfusion.

Highlights

- Anellovirus genomes assembled from longitudinal blood-transfusion cohorts
- Co-infections are common, with a median of six anellovirus lineages per subject
- Transmitted anellovirus lineages were observed up to 260 days post-transfusion
- Recombination is a key driver in anellovirus genomic diversification

Arze et al., 2021, Cell Host & Microbe *29*, 1305–1315
August 11, 2021
https://doi.org/10.1016/j.chom.2021.07.001

CellPress

Cell Host & Microbe

CellPress
OPEN ACCESS

Resource

Global genome analysis reveals a vast and dynamic anellovirus landscape within the human virome

Cesar A. Arze,[1] Simeon Springer,[1] Gytis Dudas,[2] Sneha Patel,[1] Agamoni Bhattacharyya,[1] Harish Swaminathan,[1] Carlo Brugnara,[3,4] Simon Delagrave,[1] Tuyen Ong,[1] Avak Kahvejian,[1,5] Yann Echelard,[1,5] Erica G. Weinstein,[1,5] Roger J. Hajjar,[1,5] Kristian G. Andersen,[6] and Nathan L. Yozwiak[1,7,*]

[1]Ring Therapeutics, Cambridge, MA 02139, USA
[2]Gothenburg Global Biodiversity Centre, Gothenburg 413 19, Sweden
[3]Department of Laboratory Medicine, Boston Children's Hospital, Harvard Medical School, Boston, MA 02115, USA
[4]Department of Pathology, Harvard Medical School, Boston, MA 02115, USA
[5]Flagship Pioneering, Cambridge, MA 02142, USA
[6]Scripps Research Translational Institute, La Jolla, CA 92037, USA
[7]Lead contact
*Correspondence: nyozwiak@ringtx.com
https://doi.org/10.1016/j.chom.2021.07.001

SUMMARY

Anelloviruses are a ubiquitous component of healthy human viromes and remain highly prevalent after being acquired early in life. The full extent of "anellome" diversity and its evolutionary dynamics remain unexplored. We employed in-depth sequencing of blood-transfusion donor(s)-recipient pairs coupled with public genomic resources for a large-scale assembly of anellovirus genomes and used the data to characterize global and personal anellovirus diversity through time. The breadth of the anellome is much greater than previously appreciated, and individuals harbor unique anellomes and transmit lineages that can persist for several months within a diverse milieu of endemic host lineages. Anellovirus sequence diversity is shaped by extensive recombination at all levels of divergence, hindering traditional phylogenetic analyses. Our findings illuminate the transmission dynamics and vast diversity of anelloviruses and set the foundation for future studies to characterize their biology.

INTRODUCTION

Viruses make up the most abundant component of the biosphere, yet the human virome remains largely understudied despite improved sequencing methods for detecting and analyzing viral sequences (Stulberg et al., 2016). Anelloviruses are the major eukaryotic virus constituents of the human virome, detectable in the blood throughout life (Tyschik et al., 2018), as well as in many biological samples, such as saliva, urine, and bile, suggesting broad tropism (Kaczorowska and van der Hoek, 2020; Hijikata et al., 1999; Okamoto et al., 2000; Takahashi et al., 2000). Since their discovery in 1997 (Nishizawa et al., 1997), numerous studies have surveyed anellovirus presence accompanying various diseases, but the scientific consensus is that there is no association between these viruses and human disease (Koonin et al., 2021). Anelloviruses cause chronic low-level viremia that is believed to be controlled by competent host immune systems (Freer et al., 2018) with titers that have been observed to increase in the blood of patients on immunosuppressive drugs (Görzer et al., 2014), prompting their emergence as a potential biomarker of immune status and organ transplant rejection (De Vlaminck et al., 2013).

Most healthy humans are infected with anelloviruses and co-infections with multiple unique lineages are common (Bal et al., 2018; Moustafa et al., 2017; Rani et al., 2016; Young et al., 2015); these lineages make up one's "personal anellome." The *Anelloviridae* family contains three genera of human anelloviruses: *Alphatorquevirus* (torque teno viruses [TTVs]), *Betatorquevirus* (torque teno mini viruses [TTMVs]), and *Gammatorquevirus* (torque teno midi viruses [TTMDVs]), which share an average of ~35% pairwise identity at the nucleotide level within a genus. Their single-stranded circular DNA genome (ranging from ~2.0–3.9 kb) contains a non-coding region (with a conserved ~150-base section) and a coding region with overlapping open reading frames (ORFs), the largest of which, ORF1 (~700–800 aa), is the predicted viral capsid protein (Takahashi, 1998). Genome replication occurs via a rolling-circle mechanism common to other circular DNA viruses and employs host polymerases. The full extent of anellovirus diversity across humans, the "global anellome," and the mechanism(s) that contribute to their diversification, however, remain largely uncharacterized.

Major questions also remain concerning how the anellome persists in healthy individuals, how individual lineages are acquired and lost over time, and the determinants of transmission through routes such as blood transfusions. Previous studies have shown a marked increase in anellovirus abundance after transfusions (Kapoor et al., 2015), indicating that blood transfusions could alter one's personal anellome, but the nature and duration of these fluctuations are poorly understood.

To investigate the diversity, transmission, and evolution of the human anellome over time, we developed a targeted anellovirus sequencing method that allows for high-scale profiling of anelloviruses directly from human samples and used it to study longitudinal samples from a blood-transfusion cohort consisting of donor(s)-recipient pairs. We assembled a large-scale sequence dataset and investigated the kinetics and transmissibility of the anellome within and between individuals. Our findings reveal the enormity of human anellovirus diversity globally and within individuals, such that anelloviruses occupy most of their potential evolutionary space with limited constraints on diversity and transmission. We show that complex and uniquely individual anellomes can transmit and persist for months in human individuals and that the transmission of anellovirus lineages is independent of sequence similarity to pre-existing resident anellovirus populations. Finally, we show evidence for extensive recombination in anelloviruses, pointing to a mechanism that may account for the immense diversity in this virus family.

RESULTS

Generation of a vastly expanded anellovirus genomic dataset

Unbiased metagenomic sequencing studies have consistently identified anelloviruses as a major component of the human blood virome (Moustafa et al., 2017; Tisza et al., 2020) and have attempted to describe the anellome despite low to moderate yields of anellovirus sequences. We sought to comprehensively uncover the vastness and diversity of both the global and personal anellome by developing a targeted enrichment method to improve sequencing yields. To detect and trace transmission of anellovirus lineages and explore the evolutionary characteristics of these viruses, we screened blood and serum samples from the National Heart, Lung, and Blood Institute's (NHLBI) longitudinal Transfusion-Transmitted Viruses Study (TTVS). We found a rich landscape of anelloviruses, recovering tens to hundreds of unique lineages from each subject.

To increase our ability to obtain high-coverage anellovirus genomes, we designed a targeted rolling-circle amplification (RCA) enrichment method, Anelloscope, utilizing degenerate amplification primers (Table S1) that covered conserved regions of the genome. We validated yield gains by measuring the difference in anellovirus genomic sequences recovered between standard RCA (Niel et al., 2005) and Anelloscope. We found that Anelloscope led to a 1,046- to 52,812-fold increase in coverage when benchmarked on serum samples from the TTVS cohort (Table S2).

To exhaustively identify unique anellovirus sequences in the TTVS cohort, we applied Anelloscope to 67 individuals comprising 128 samples, 53 from donors and 75 from recipients across five time points (one pre-transfusion, four post-transfusion; Table S3). In total, we produced 501.5 Gbp of sequence data, of which 152.9 Gbp were derived from anelloviruses (Table S4). The remaining sequences were identified as human sequences and removed during decontamination steps. We assembled 1,656 high-quality anellovirus contigs (median length = 2,916 bp, min length = 2,190 bp, max length = 4,917 bp) from the generated data. Additionally, we curated publicly available anellovirus genomes to create a supplementary dataset of 445 sequences. In total, we produced a dataset containing 2,101 anellovirus sequences for further downstream analysis, nearly tripling the number of known anellovirus sequences.

Anelloviruses occupy an expansive evolutionary space

We performed both phylogenetic and multidimensional scaling (MDS) analyses on our dataset to assess the overall diversity of human anelloviruses, putting the anelloviruses identified in this study in the context of those previously discovered. We found that our sequences from the TTVS cohort greatly expanded on known diversity within the anellovirus family.

We first examined whether our dataset altered the previously established phylogenetic relationships or taxonomic classifications of anelloviruses. We constructed a maximum-likelihood phylogeny of all 2,101 ORF1 sequences and found that those derived from the TTVS cohort clustered into the three previously established genera (Lefkowitz et al., 2018) (Figure 1A). Furthermore, we observed growth in sequences in each of the genera, *Alphatorque-*, *Betatorque-*, and *Gammatorqueviruses*, by 28%, 27%, and 15%, respectively. More marked was the phylogenetic branch length contribution per new sequence we added in the *Betatorque-* (0.114 substitutions per amino acid site per sequence) and *Gammatorque-* (0.148) genera compared with *Alphatorqueviruses* (0.039) (branches colored black in Figure 1A). These findings indicate that the latter's diversity is mostly explored, whereas each *Betatorquevirus* and *Gammatorquevirus* sequence, on average, adds substantial previously unseen diversity to the phylogenetic tree.

Given the significant anellovirus diversity we observed, we next sought to formally compare anelloviruses with other viruses by using multidimensional scaling (MDS) analysis. We used MDS to avoid the clonal evolutionary assumptions made by phylogenetic analysis (Nei and Kumar, 2000) and account for previously hypothesized recombination (Fahsbender et al., 2017; Lefeuvre et al., 2009; Leppik et al., 2007; Martin et al., 2011; Worobey, 2000) in anelloviruses. We selected eight representative viruses, including examples of viruses that are well known for their capsid diversity (de Villiers et al., 2004), ability to recombine (Dudas and Rambaut, 2016; Holmes et al., 1999b; Smyth et al., 2012; Worobey et al., 1999), or important vectors for gene therapy (Colella et al., 2018), and analyzed their surface proteins alongside our anellovirus dataset (Figure 1B). By quantifying the area occupied by all sequences projected in MDS space we generated a simple diversity metric that reflects total sequence diversity and shows the relationships between sequences without relying on methods like phylogenetic trees that cannot accommodate recombination. The distance between the two furthest points in each MDS plot will be approximately proportional (as much as stress, the differences between pairwise and MDS distances,

CellPress
OPEN ACCESS

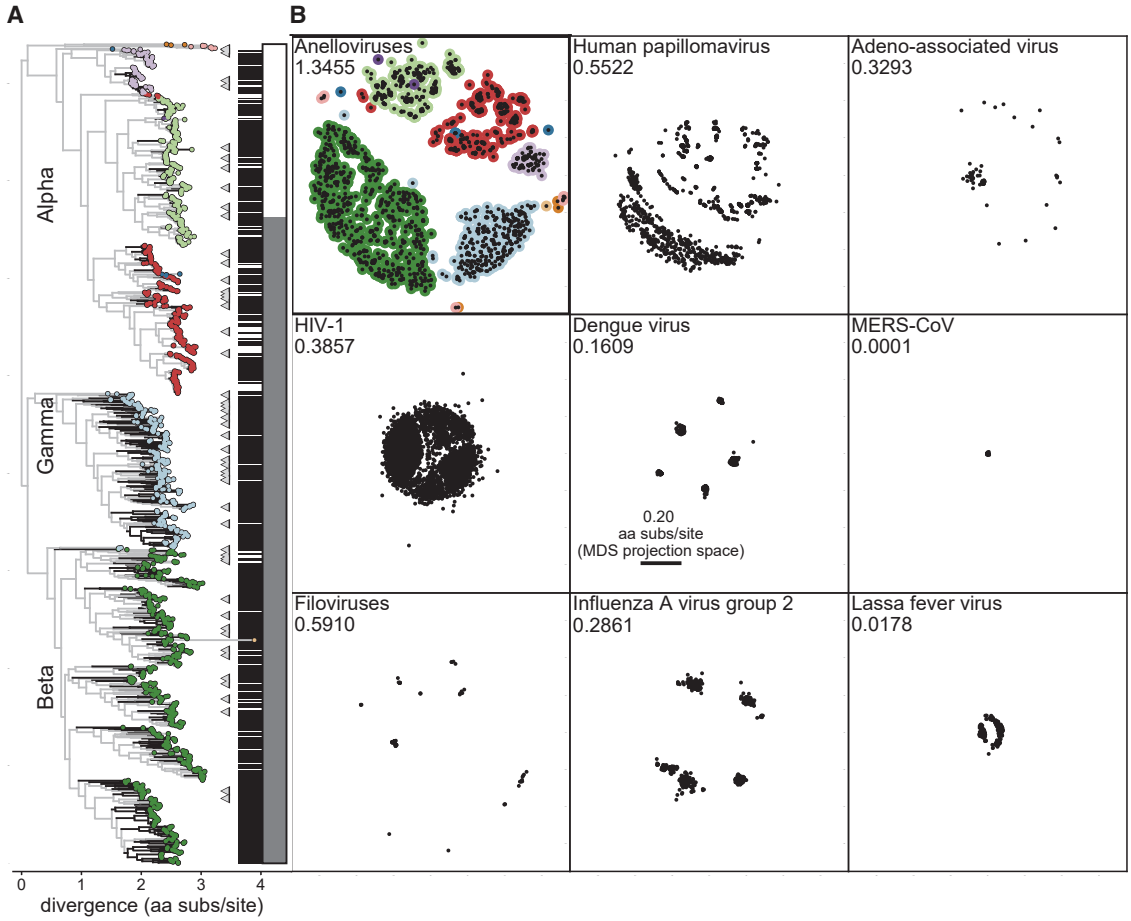


Figure 1. Mapping the extent of anellovirus diversity

(A) Maximum-likelihood phylogeny of anellovirus ORF1 amino acid sequences (n = 2,101). Tips are colored according to agglomerative clustering of pairwise amino acid distances to produce ten arbitrary clusters that are colored to ease navigation between the tree and (B). Gray branches connect previously published sequences to the root, and black branches represent sequences reported in this study. Triangles to the right of the tree indicate reference sequences retrieved from RefSeq. Black dashes to the right of the tree indicate the positions and volume of new sequences. The gray bar to the right displays the total amount of sequences derived from study samples as a proportion of the entire dataset used in the analysis.

(B) Multidimensional scaling (MDS) analysis of 1,575 anellovirus ORF1 amino acid sequences (points are colored as in A) compared with eight other viral surface proteins: 2,627 human papillomavirus (HPV) L1, 86 adeno-associated virus (AAV) capsid, 3,000 human immunodeficiency virus 1 (HIV1) env, 3,000 dengue virus envelope, 425 Middle East-associated respiratory syndrome coronavirus (MERS-CoV) Spike, 3,000 influenza A virus HA (group 2, subtypes H3, H4, H7, H10, and H14), 64 homologous filovirus (encompassing genera Ebolavirus, Dianlovirus, Marburgvirus, and Cuevavirus) GP, and 632 Lassa fever virus GPC protein sequences. MDS plots are shown on the same scale; the scale bar represents 0.2 aa substitutions per site in MDS projection space. Numbers presented above each facet indicate the area occupied by the outermost points of the MDS projection. See also Figure S1.

can be minimized) to the proportion of amino acid sites at which the two most distantly related sequences differ. In anelloviruses, we found that this computed measurement was three to four times larger than any of the eight viruses we compared against (Figure 1B, number above each facet), lending support to the hypothesis that extraordinary diversity is a hallmark of anelloviruses (Kaczorowska and van der Hoek, 2020).

Next, to investigate the anellovirus evolutionary space at amino acid sites, we determined how many of the possible amino acid states are occupied by ORF1 sequences. We tabulated the number of unique amino acids occurring at each position in an alignment of all ORF1 sequences and compared these measurements against those from the surface proteins of the eight viral families that were selected for MDS analysis. As expected, we observed the highest amounts of amino acid diversity in anelloviruses, concentrated in the hypervariable regions (HVRs) (Figure S1, left column, residues ~300 to ~500) of the ORF1 protein that is hypothesized to be involved in host immune evasion (de Villiers and zur Hausen, 2009). Yet, we found that the average unique amino acids per site were also elevated across the entire ORF1 sequence, in all three genera, suggesting that the diversity observed is widespread over the entire protein and not just localized to the HVRs. In comparisons against other viruses, only HIV-1 env exhibited equal or greater diversity than anelloviruses, primarily driven by its hypervariable loops (Figure S1, right column). Together, these findings offer a detailed picture of the extent of anellovirus diversity and indicate that sources of the diversity observed may be widespread across the entire genome.

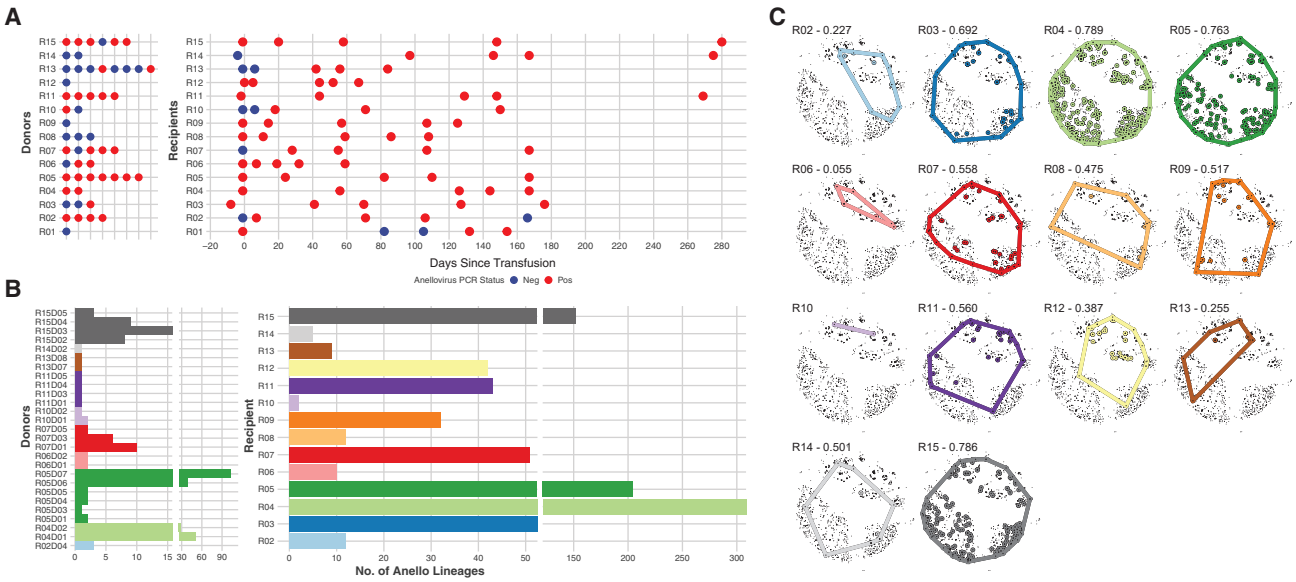


Figure 2. Characterization of the personal anellome

(A) Overview of study design. Fifteen recipients paired with one or more donors received a blood transfusion after surgery. Donor pools were unique to each of the recipients receiving a blood transfusion. Samples were collected both before and after transfusion over a period of 280 days. The corresponding rows from the left panel (donors) are paired with the corresponding row in the right panel (recipients) to define donor pools and their recipients. Pan-anellovirus PCR-positive samples are colored in red, whereas PCR-negative samples are in blue.

(B) The number of unique anellovirus lineages identified per individual. Each donor time point where lineages were isolated is included to indicate the levels of anellovirus in each sample. All five recipient time points are aggregated to show the total number of lineages in each recipient.

(C) Anellovirus diversity per transfusion recipient. MDS analysis demonstrates within-subject anellovirus diversity that spans the space of overall known anellovirus diversity. Convex hulls depict the amount of the diversity space encompassed in each subject set. Numbers presented above each facet indicate the fraction of area occupied by the convex hull (outermost points) of the patient's anelloviruses in comparison with the area of the convex hull of all anelloviruses sampled. See also Figures S2 and S3.

Blood-transfusion donors and recipients reveal an abundance of anellovirus lineages

Our sequencing data allowed us to determine the diversity of personal anellomes. To identify and approximate the number of lineages in each anellome, we surveyed the 128 donor and recipient samples. We found that all blood donors and transfusion recipients had diverse infections with multiple co-circulating anellovirus lineages.

To detect the presence of anellovirus DNA in specimens from the TTVS cohort, we conducted pan-anello PCR assays (Ninomiya et al., 2008) (Figure 2A) on all donor and recipient samples. We detected anelloviruses in 53% of donors (33/53 samples) and 86% of recipients (65/75 samples) and found at least one positive sample in each donor-recipient set. We observed anellovirus presence as far out as 260 days post-transfusion (Figure 2A, recipients 13–15), and one of these lineages was identified as donor derived via sequencing data (Figure 3).

To estimate the number of distinct anellovirus lineages in each anellome, we used our Anelloscope method to recover anellovirus sequences from each subject of the TTVS cohort. We used anellovirus ORF1 sequences as a unique marker that could be easily identified with length filters and a well-conserved motif in the N22 region (Nishizawa et al., 1997) (Figure S2A). We observed a median of six distinct lineages in all subjects from the TTVS cohort, with three individuals containing over 100 lineages across their five associated time points (Figure 2B). We also found that the median number of distinct lineages increased over 4-fold to 27 when we examined just recipients, suggesting that increasing numbers of lineages were most likely elevated by the introduction of donor-derived lineages during blood transfusions.

Individuals harbor a diverse personal anellome

Although we observed a large number of anelloviruses present in each personal anellome, the extent of the diversity of these identified lineages was yet to be characterized. It remained unknown whether preferences for closely related lineages within individuals existed or if anelloviruses covering the spectrum of diversity were present. By analyzing data from the TTVS cohort, we also explored whether the diversity within the anellome is restricted in evolutionary space and found that anellomes have extensive diversity with lineages drawn across the entire anellovirus family. Very rarely did we observe lineages that were shared across multiple subjects (Figure S3A). These findings point toward an even higher prevalence of distinct anellovirus lineages in anellomes than previously reported (Al-Qahtani et al., 2016) and that anelloviruses may inhabit a majority of all individuals.

To contextualize the diversity of the personal anellome in the broader context of the global anellome, we investigated the diversity of anelloviruses identified in each subject from the TTVS cohort and compared them with the diversity of all anelloviruses by using MDS projections (Figure 2B). We found that subjects with the most lineages (recipients 4, 5, and 15; Figure 2B, facets 3, 4, and 15) captured most of the total diversity observed in our global anellome (Figure 1B, facet 1). In

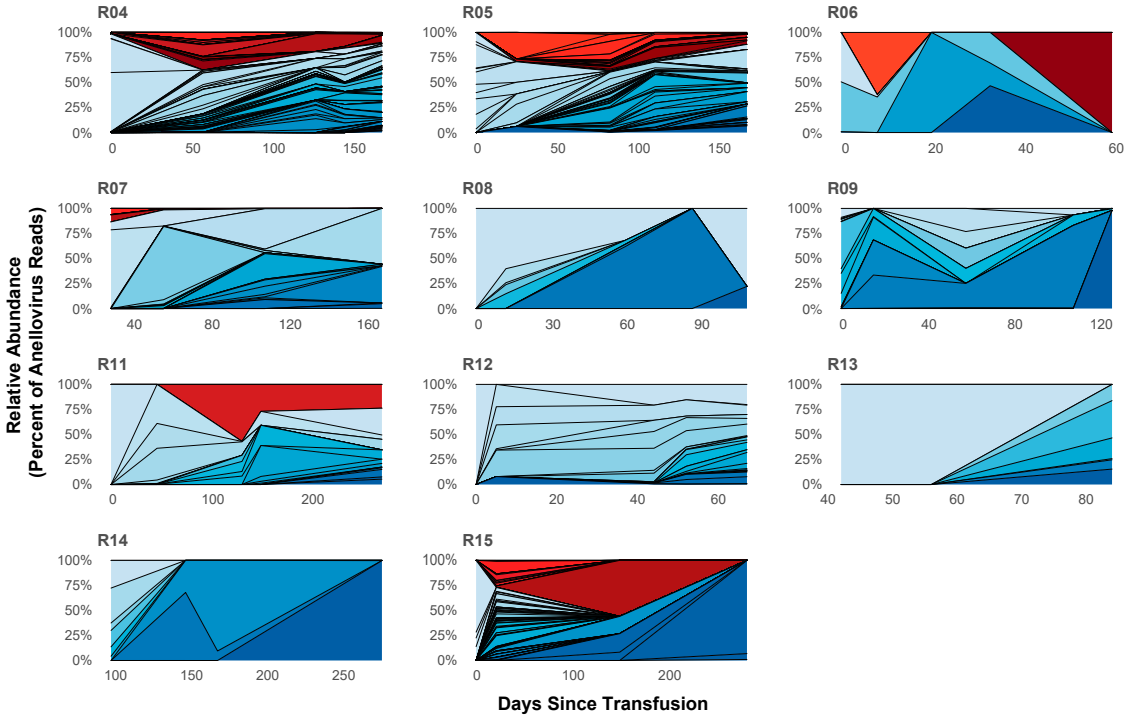


Figure 3. Multiple anellovirus lineages are transmitted via blood transfusion

Tracking of anellovirus relative abundance over the course of the longitudinal TTVS study after blood transfusion. Lineages colored in shades of red denote transmitted lineages from the donor(s), whereas shades of blue indicate resident lineages.

individuals with fewer lineages, however, we also observed that personal anellomes captured the breadth of global diversity (recipients 3, 7–9, 11, and 14), suggesting that even in subjects with smaller anellomes diversity remains high (Figure 2B, facets 2, 6–8, 10, and 13). Additionally, we summarized the total occupied areas of the projected 2D analysis space by using the same summary diversity statistic computed in our global analysis and found that in recipients with large personal anellomes this statistic was only 34% higher than in recipients with smaller personal anellomes (Figure 2B, number above each facet). In the cases where we observed smaller, personal anellomes (recipients 2, 6, 10, 13; Figure 2B, facets 1, 5, 9, and 12), we found that the diversity statistic was on average 80% lower than in recipients with larger anellomes. These results suggest that in most individuals, the breadth of anellovirus diversity is large and uncoupled from the distinct number of lineages present.

Marked and persistent shifts in the anellome after blood transfusion

To investigate the uptake of new anellovirus lineages into anellomes, we identified and tracked individual transmitted lineages in 15 blood-transfusion recipients from the TTVS cohort over time. We found that anellovirus transmission occurred consistently in most subjects and that most blood-transfusion recipients contained at least one lineage that was transmitted from one or more donors (Table S5). By searching for the presence of donor transmitted lineages in recipients and measuring the proportion of sequencing reads that mapped to these lineages, we also found that transmitted lineages can be found at least nine months post-transfusion (Figure 3), consistent with previous reports (Abbas et al., 2019; Bédarida et al., 2017).

We mapped sequence data from longitudinal recipient samples from the TTVS cohort to anellovirus lineages recovered from paired donor samples and classified these lineages as being transmitted if they were present in one or more time points. We found evidence of at least one transmitted lineage in 8/15 recipients. We observed that the maximum number of lineages transmitted between one set of donors and a recipient was 53 (Table S5, rows with subject R04) with a median of five transmitted lineages across all recipients. These findings suggest that anellovirus transmission via blood transfusion is highly permissive, similarly to other blood-borne viruses, such as HIV-1 (Des Jarlais et al., 1988). In addition, we identified six donor lineages that also existed in recipient's personal anellome pre-transfusion, indicating that these lineages could be examples of "re-dosed." In total, out of 194 anellovirus lineages from our donors, we identified 133 lineages that we classified as transmitted to a paired recipient (Table S5).

To investigate the persistence of transmitted anellovirus donor lineages in a recipient's anellome over time, we used the mapped sequence data to donor anelloviruses to calculate the relative proportion of each donor lineage at each recipient time point and used these measurements to approximate the duration of observable infection over the course of the study. We found that in recipients with transmitted lineages, the lineages frequently persisted over 100 days (Figure 3). In a subset of recipients (4, 5, 6, 11, 15; Figure 3, facets 1–3, 7, and 11) we observed marked shifts of anellomes, with resident anellovirus

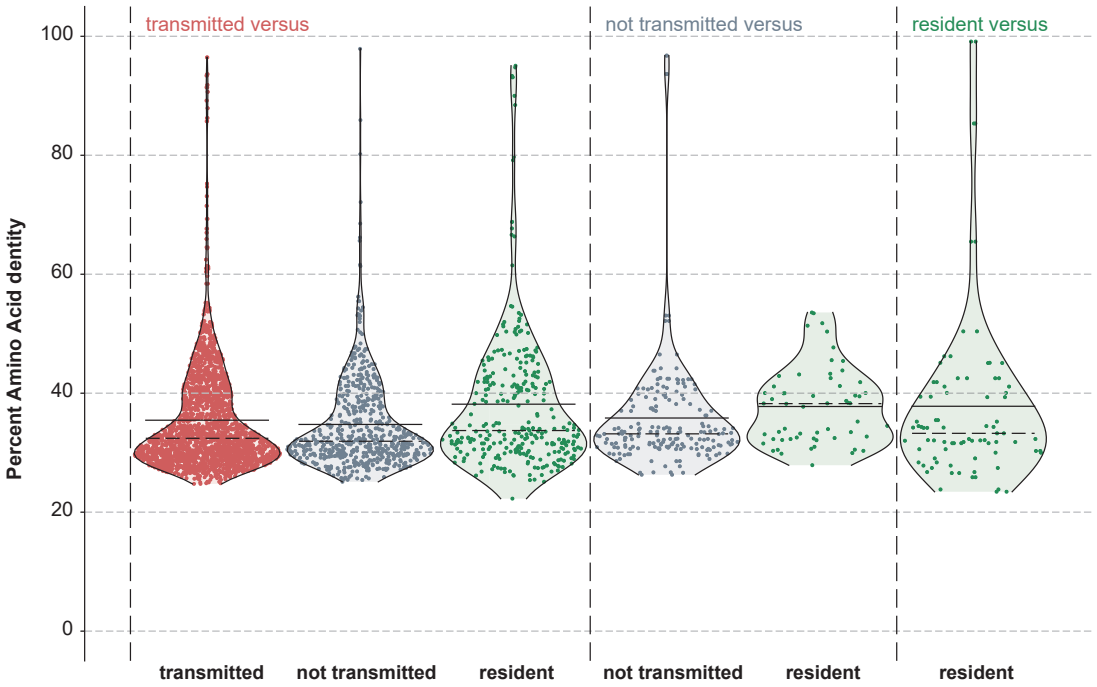


Figure 4. Donor-recipient anellovirus sequence similarity pre-transfusion does not influence transmission

Comparison of pairwise distances between different subsets of anelloviruses isolated from transfusion subjects. Horizontal solid and dashed lines indicate the mean and median, respectively, amino acid identity between comparisons between each group. The similarity of anelloviruses between the donor and those in the recipient prior to transfusion does not predict transmissibility.

lineages decreasing in proportion over time, whereas transmitted lineages increased (Figure 3). These findings suggest that transmitted anellovirus lineages can change the architecture of anellomes after blood transfusion.

Anelloviruses transmit independently of genomic similarity to resident lineages

Next, we compared the pairwise distances of anellovirus subsets to determine if sequence similarity predicts the transmissibility of lineages from donor anellomes to recipient anellomes. By analyzing these sequence similarities stratified into three categories, we found that the genetic similarity of anellovirus lineages to lineages in the resident anellome did not affect transmissibility.

We measured pairwise sequence similarities between ORF1 sequences derived from the anellovirus lineages binned into each of the three categories (“transmitted,” “not transmitted,” and “resident”; Figure 4). We observed that the comparisons between the transmitted lineages and resident lineages were not significantly elevated when contrasted against the comparisons between not transmitted and resident lineages (median similarity of 32.2% versus 32.4%, respectively). Across all comparisons between all three categories, we observed an average median amino acid identity of 32.4% (Figure 4; black dotted lines), suggesting that transmission of lineages occurs irrespective of similarity to resident lineages in recipients.

Recombination is a key driver of anellovirus evolution

Having observed vast diversity in both the global and personal anellomes, we explored what mechanisms might be responsible for creating genetic diversity in anelloviruses. Recent studies have suggested that recombination may be a key contributor to this diversification (Leppik et al., 2007; Worobey, 2000); thus, we investigated its potential by inspecting fragments of highly similar, unambiguously aligned, ORF1 nucleotide sequences. Within these fragments, we searched for hallmarks of recombination: excess repeat mutations (homoplasies) (Smith and Smith, 1998), shifting topologies in constructed phylogenetic trees (Holmes et al., 1999a), and low linkage disequilibrium (LD) measurements (Meunier and Eyre-Walker, 2001). We found strong evidence of recombination across all analyses, suggesting that it may be a key driver of diversity in anelloviruses.

To identify potential regions of unlinked evolution that are suggestive of recombination, we identified several clusters of anellovirus lineages that were closely enough related at the nucleotide level to allow for the creation of a succession of phylogenies tiling across the ORF1 sequence in 500 nucleotide fragments. By visualizing these phylogenies (Figure 5A), we found a pattern of shifting topologies in each subsequent phylogeny and observed several cases where a lack of consistency between neighboring fragments derived from a single anellovirus lineage occurred, suggesting that these fragments most likely originated from a recombination event. We observed this phenomenon occurring in all three genera, indicating that recombination events are not limited to a specific subset of lineages. Furthermore, we used GARD (Kosakovsky Pond et al., 2006) to naively look for the number and distribution of recombination breakpoints in these closely related ORF1 sequences and found that for each genus alignment at least seven trees explain the data better than a

CellPress
OPEN ACCESS

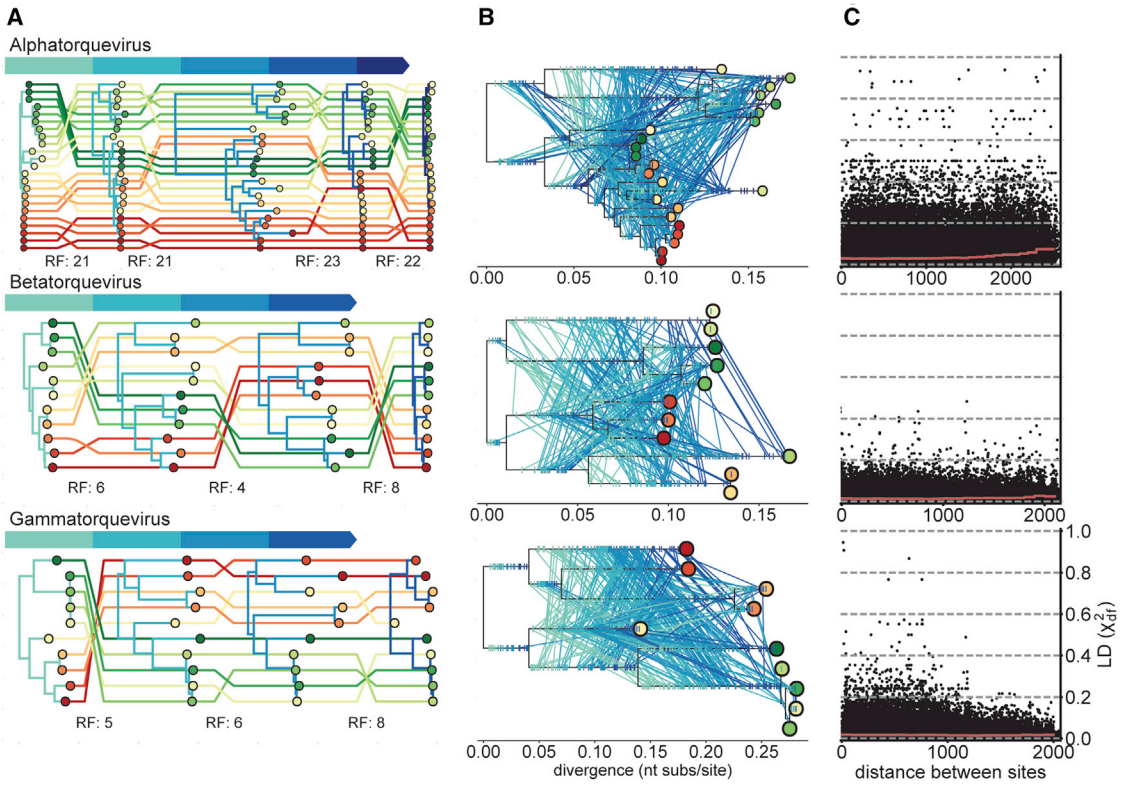


Figure 5. Anellovirus recombination contributes to extensive diversity

(A) Tangled chain of midpoint-rooted phylogenies inferred from 500 nucleotide fragments of anellovirus ORF1; the position of each lineage in successive phylogenies is shown with lines colored by their relative position in the first phylogeny. Labels along the bottom of the trees for each genus indicate the Robinson-Foulds (RF) distance between trees of neighboring fragments. Unlinked evolution across the genome is strong evidence of recombination.

(B) Evidence of recombination in anellovirus through excessive homoplasies. Ancestral sequence reconstruction between sequences within 80% identity of each other at the nucleotide level shows rampant repeat mutations—each line connects identical mutations occurring on different branches, and fractions along the length of the branch indicate the relative position of the mutation in the genome. Ticks on branches indicate mutations that occur uniquely on the branch in question. Branches are colored according to the fraction of all mutations that are homoplasies with the highest values (all mutations are homoplasies, i.e., none are unique) highlighted in white. See also Figures S4 and S5.

single tree (Figure S4). These results indicate that anellovirus recombination is present and prevalent, even at lowest levels of divergence.

Next, we looked for signs of recombination in the NCRs of the anellovirus genome that are under different evolutionary constraints than the ORF1 protein. Using the 5′ untranslated region (UTR) as a proxy, we first established the levels of diversity in the NCR by computing pairwise sequence similarities across all lineages in our dataset. We found that the 5′ UTR demonstrated the highest median pairwise similarity of 83% compared with 46% in full sequences, 31% in ORF1 protein sequences, and 33% in ORF2 protein sequences (Figure S3B). The results indicate that the 5′ UTR (and the NCR by proxy) contains a higher level of conservation than other regions of the genome and that anelloviruses might exhibit lower levels of recombination in NCRs as a result of non-uniform selection that enforces conservation. To confirm this hypothesis, we investigated recombination tracts in the NCR of the *Alphatorquevirus* genus and found evidence of widespread recombination (Figure S5A). Moreover, we observed that these events occurred across the entire diversity of the genus and were not limited to just closely related lineages (Figure S5B). Next, we searched for evidence of recombination events occurring in *Betatorquevirus* and *Gammatorquevirus* lineages by examining alignments in the 5′ UTR (acting as a proxy for the NCR) across all three genera. We observed that five lineages assigned to the genus *Gammatorquevirus* are more closely related to *Betatorquevirus* lineages, indicating that recombination is occurring in the 5′ UTR with respect to the rest of the genome. These findings suggest that recombination likely occurs across the entire genome and not just the regions that we would expect to be under evolutionary pressure to escape immune recognition (Kakkola et al., 2008).

To measure the frequency of recombination events occurring across the anellovirus genome, we next sought to examine the abundance of homoplasies found in phylogenies from all three genera. We reconstructed ancestral sequences at the nucleotide level for each genus by using our set of closely related lineages and observed frequent homoplasies throughout each of the three reconstructed trees (Figure 5B). We found that roughly a third of all mutations reconstructed on each phylogeny were homoplasies (ranging from 28.5% to 33.3%). We selected these sequences because of their low levels of divergence, and the

presence of such a high frequency of homoplasies strongly suggests that recombination rates in anellovirus are a key driver of generating the large amount of genetic diversity observed (Worobey, 2000).

To confirm our phylogenetic-based findings and provide further evidence of recombination, we next quantitatively summarized recombination across the entire ORF1 sequence by correlating LD measurements with genomic distance within each genus. For each lineage, stratified across genus, we computed the LD values between polymorphic sites and found that between neighboring sites, these values averaged near 0. LD values near 0 between adjacent loci point to effectively free recombination (Figure 5C). This is observed across all genera, once again reinforcing that recombination is not isolated to one specific genera of anelloviruses.

The high levels of recombination we observed over the entire genome indicate that carefully understanding and selecting the models of evolution used when examining the diversity and relationships between lineages is crucial. Furthermore, these results provide additional detail on the breadth and scope of recombination events across the entire genome. The effects of recombination, the diversity it drives and how this impacts persistence and host immune evasion warrant further study.

DISCUSSION

In this study, we explored the anellomes of blood-transfusion recipients and their matching donors and found a vast, dynamic, and unique anellovirus landscape that suggests that each individual harbors a distinctive personal anellome. The diversity of anelloviruses has previously been noted (Moustafa et al., 2017; Tisza et al., 2020), but here we observed a depth and complexity that has not previously been characterized as such.

Anellovirus evolution is highly reticulate

Our analysis of anellovirus sequences point to frequent recombination in both the study cohort presented here and public data. We observed frequent co-infections with multiple distinct lineages of anelloviruses that would provide ample opportunity for recombination to occur within individuals, although a deeper understanding of anellovirus replication is lacking to explain the exact mechanism through which recombination occurs. When detected, evidence for recombination is clearest at low levels of divergence between closely related ORF1 sequences from donor-recipient pairs, within-patient sequence clusters, or more conserved regions of the genome. Beyond these scenarios, it is not possible to infer homologous sites with any certainty given the diversity of the ORF1 sequences. However, we observed instances of recombination across distinct anellovirus genera, suggesting a mechanism sensitive to sequence similarity.

We cannot dismiss other mechanisms that generate mutations, because we do not comprehensively understand anellovirus replication or ORF1's function and the role of its HVR. For instance, some viruses possess genomic regions that evolve neutrally (Park et al., 2015) or contain non-coding features with preserved ORFs (DeRisi et al., 2019). If the HVR of ORF1 is similarly functionally unconstrained, the reliance on the host machinery for replication opens them up to mutagenesis via a wide variety of host DNA editing mechanisms, such as error-prone DNA repair or cytidine deaminases of the APOBEC family.

The evidence we present here for recombination in anellovirus evolution has implications for future analyses because strictly clonal models of evolution cannot adequately infer the relationships or distances between sequences, and no region of the genome is entirely free of recombination. Simple passaging experiments with and without co-infection could be illuminating about the ways in which sequence diversity is produced and maintained should a suitable system for experimental propagation of human anelloviruses become available.

Transmission dynamics and persistence of anelloviruses

To understand why some anellovirus lineages in blood donors transmit and others do not, we asked whether the similarity of donor lineages to resident lineages in a recipient confers an increase or decrease in transmissibility. We found that the similarity of donor lineages to the host anellome seems to have little effect on transmission (Figure 3). In fact, we found instances of donor lineages that successfully transmitted despite high sequence similarity (>90%) to resident lineages, which raises the intriguing possibility of re-infection and suggests that potential therapeutic vectors derived from anelloviruses could be effectively re-dosed. The rates and frequency of anellovirus transmission via other non-iatrogenic routes, such as respiratory and fecal-oral, were not evaluated in this study but are likely to occur readily as evidenced in part by the ubiquitous acquisition of anelloviruses in the first year of life (Tyschik et al., 2018).

Targeted anellovirus sequencing enabled us to differentiate and track hundreds of unique anellovirus lineages over time. We found a high prevalence of co-infections with multiple lineages and observed both resident and transmitted lineages that persisted for the duration of this study, highlighting the need for lengthier longitudinal studies to examine the full potential of anellovirus infectivity. The persistence of newly transmitted lineages via blood transfusion suggests that an intravenously delivered therapeutic could be a potent vehicle for delivery.

Our work describes the transmission of individual anellovirus lineages, and future research could further explore the factors that influence transmissibility differences between lineages, such as viral loads, host immune status, and non-structural anellovirus proteins. For instance, quantifying lineage concentrations within subjects in a longitudinal transfusion cohort could resolve the infective dose of distinct lineages. Exploring the permissibility that leads to an environment with a high number of co-infections in light of our evidence indicating that sequence determinism alone does not factor into transmissibility could point to co-infections of specific lineages being key to transmission. Additionally, future longitudinal studies will need to examine lineages that appear in low abundance or are missing at specific time points to rule out whether those lineages have fallen below the limit of detection. Lastly, this work focused on the anellovirus capsid sequences because it is the most easily identifiable, but several other viral proteins and genomic regions should be explored.

Anellovirus diversity could provide opportunities for unique applications

Our work has confirmed that anellovirus infections are prevalent, but more importantly, it has uncovered a depth of abundance and transmissibility that could potentially be explored in an array of translational applications. The anellovirus diversity documented in the TTVS cohort indicates that their commensalism and persistent evasion of the immune system could be harnessed in medical applications, including potential delivery of therapeutic payloads as well as a biomarker of organ transplantation. For example, anellovirus abundance has been proposed as a surrogate for immune system competence in post-transplant individuals (Focosi et al., 2010; Maggi et al., 2008). As anellovirus amplification and sequencing methods improve, sensitive and rapid diagnostic assays could be utilized to improve outcomes for immunosuppressed individuals.

The lack of appropriate *in vitro* propagation methods has thus far limited many facets of anellovirus research (Focosi et al., 2016). The vast number of unique capsid sequences points to a multitude of potential vectors for delivering therapeutic payloads, but important considerations remain. We report that anellovirus diversity is driven in part by recombination, so vectorization would need to mitigate this process. This could potentially be achieved through the use of a replication incompetent vector that would be unable to recombine with co-infected wild-type anelloviruses and would also alleviate unwanted viral shedding of recombinants. An anellovirus therapeutic platform that could overcome these issues could open the door to leveraging the anellome as potential vector candidates.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Sample collection
  - Multiply primed RCA
  - Pan-Anello PCR
  - Illumina library preparation and sequencing
  - Nanopore library preparation and sequencing
  - Retrieval of public anellovirus sequences
  - Genomic anellovirus sequence identification
  - Illumina sequence quality control
  - Nanopore sequence quality control and mapping
  - Genome assembly
  - Genome annotation
  - Genera classification
  - Anellovirus proportion estimation
  - Multidimensional scaling
  - Phylogenetic analysis
  - Recombination analyses
  - Pairwise sequence identity analysis
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.chom.2021.07.001.

ACKNOWLEDGMENTS

This study utilized TTVS research materials obtained from the NHLBI Biological Specimen and Data Repository Information Coordinating Center and does not necessarily reflect the opinions or views of the TTVS or the NHLBI.

AUTHOR CONTRIBUTIONS

Study initiation, N.L.Y., K.G.A., and E.G.W.; conceptualization, N.L.Y., K.G.A., C.A.A., E.G.W., and S.S.; sample collection, processing, and data generation, S.S., A.B., and S.P.; data analysis, C.A.A., G.D., and H.S.; writing, C.A.A., N.L.Y., K.G.A., G.D., H.S., S.D, R.H., E.G.W., Y.E., A.K., C.B., and T.O.

DECLARATION OF INTERESTS

C.A.A., S.S., S.P., A.B., H.S., S.D., R.J.H., T.O., and N.L.Y. are employees of, and hold equity in, Ring Therapeutics. K.G.A is a paid consultant and holds equity in Ring Therapeutics. G.D. is a paid consultant of Ring Therapeutics. A.K., Y.E., and E.G.W. are employees of Flagship Pioneering, which provides funding to Ring Therapeutics. A.K., Y.E., and E.G.W. hold equity in Ring Therapeutics. A.K. is a board member of Ring Therapeutics. C.A.A., S.S., N.L.Y., K.G.A., E.G.W., A.K., S.D., Y.E., and R.J.H. are inventors on patent applications related to the research described here.

Received: January 14, 2021
Revised: April 23, 2021
Accepted: June 11, 2021
Published: July 27, 2021

REFERENCES

Abbas, A.A., Young, J.C., Clarke, E.L., Diamond, J.M., Imai, I., Haas, A.R., Cantu, E., Lederer, D.J., Meyer, K., Milewski, R.K., et al. (2019). Bidirectional transfer of Anelloviridae lineages between graft and host during lung transplantation. Am. J. Transplant. *19*, 1086–1097.

Al-Qahtani, A.A., Alabsi, E.S., Abuodeh, R., Thalib, L., El Zowalaty, M.E., and Nasrallah, G.K. (2016). Prevalence of anelloviruses (TTV, TTMDV, and TTMV) in healthy blood donors and in patients infected with HBV or HCV in Qatar. Virol. J. *13*, 208.

Andrews, S. (2019). FastQC: a quality control tool for high throughput sequence data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res *37*, W202–W208.

Bal, A., Sarkozy, C., Josset, L., Cheynet, V., Oriol, G., Becker, J., Vilchez, G., Sesques, P., Mallet, F., Pachot, A., et al. (2018). Metagenomic next-generation sequencing reveals individual composition and dynamics of anelloviruses during autologous stem cell transplant recipient management. Viruses *10*, 633.

Bédarida, S., Dussol, B., Signoli, M., and Biagini, P. (2017). Analysis of Anelloviridae sequences characterized from serial human and animal biological samples. Infect. Genet. Evol. *53*, 89–93.

Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Sayers, E.W. (2012). GenBank. Nucleic Acids Res. *41*, D36–D42.

Biomatters. (2021). Geneious prime (Geneious).

Broad Institute (2018). Picard tools. http://broadinstitute.github.io/picard/.

Bushnell, B. (2014). BBMap: a fast, accurate, splice-aware aligner (Lawrence Berkeley National Laboratory).

Calus, S.T., Ijaz, U.Z., and Pinto, A.J. (2018). NanoAmpli-Seq: a workflow for amplicon sequencing for mixed microbial communities on the nanopore sequencing platform. GigaScience *7*, giy140.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). Blast+: architecture and applications. BMC Bioinformatics *10*, 421.

Colella, P., Ronzitti, G., and Mingozzi, F. (2018). Emerging issues in AAV-mediated in vivo gene therapy. Mol. Ther. Methods Clin. Dev. *8*, 87–104.

de Villiers, E.M., Fauquet, C., Broker, T.R., Bernard, H.U., and zur Hausen, H. (2004). Classification of papillomaviruses. Virology *324*, 17–27.

de Villiers, E.-M., and zur Hausen, H. (2009). TT viruses–the still elusive human pathogens. Preface. Curr Top Microbiol Immunol *331*, v–vi.

De Vlaminck, I., Khush, K.K., Strehl, C., Kohli, B., Luikart, H., Neff, N.F., Okamoto, J., Snyder, T.M., Cornfield, D.N., Nicolls, M.R., et al. (2013). Temporal response of the human virome to immunosuppression and antiviral therapy. Cell *155*, 1178–1187.

DeRisi, J.L., Huber, G., Kistler, A., Retallack, H., Wilkinson, M., and Yllanes, D. (2019). An exploration of ambigrammatic sequences in narnaviruses. Sci. Rep. *9*, 17982.

Des Jarlais, D.C., Friedman, S.R., and Stoneburner, R.L. (1988). HIV infection and intravenous drug use: critical issues in transmission dynamics, infection outcomes, and prevention. Rev. Infect. Dis. *10*, 151–158.

Didelot, X., and Wilson, D.J. (2015). ClonalFrameML: efficient inference of recombination in whole bacterial genomes. PLOS Comput. Biol. *11*, e1004041.

Dudas, G., and Rambaut, A. (2016). MERS-CoV recombination: implications about the reservoir and potential for adaptation. Virus Evol *2*, vev023.

Edgar, R.C. (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics *26*, 2460–2461.

Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics *32*, 3047–3048.

Fahsbender, E., Burns, J.M., Kim, S., Kraberger, S., Frankfurter, G., Eilers, A.A., Shero, M.R., Beltran, R., Kirkham, A., McCorkell, R., et al. (2017). Diverse and highly recombinant anelloviruses associated with Weddell seals in Antarctica. Virus Evol *3*, vex017.

Focosi, D., Antonelli, G., Pistello, M., and Maggi, F. (2016). Torquetenovirus: the human virome from bench to bedside. Clin. Microbiol. Infect. *22*, 589–593.

Focosi, D., Maggi, F., Albani, M., Macera, L., Ricci, V., Gragnani, S., Di Beo, S., Ghimenti, M., Antonelli, G., Bendinelli, M., et al. (2010). Torquetenovirus viremia kinetics after autologous stem cell transplantation are predictable and may serve as a surrogate marker of functional immune reconstitution. J. Clin. Virol. *47*, 189–192.

Freer, G., Maggi, F., Pifferi, M., Di Cicco, M.E., Peroni, D.G., and Pistello, M. (2018). The virome and its major component, anellovirus, a convoluted system molding human immune defenses and possibly affecting the development of asthma and respiratory diseases in childhood. Front. Microbiol. *9*, 686.

Görzer, I., Haloschan, M., Jaksch, P., Klepetko, W., and Puchhammer-Stöckl, E. (2014). Plasma DNA levels of torque teno virus and immunosuppression after lung transplantation. J. Heart Lung Transplant. *33*, 320–323.

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. *59*, 307–321.

Hasegawa, M., Kishino, H., and Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. *22*, 160–174.

Hedrick, P.W., and Thomson, G. (1986). A two-locus neutrality test: applications to humans, E.coli and lodgepole pine. Genetics *112*, 135–156.

Hijikata, M., Takahashi, K., and Mishiro, S. (1999). Complete circular DNA genome of a TT virus variant (isolate name SANBAN) and 44 partial ORF2 sequences implicating a great degree of diversity beyond genotypes. Virology *260*, 17–22.

Holmes, E.C., Urwin, R., and Maiden, M.C. (1999a). The influence of recombination on the population structure and evolution of the human pathogen Neisseria meningitidis. Mol. Biol. Evol. *16*, 741–749.

Holmes, E.C., Worobey, M., and Rambaut, A. (1999b). Phylogenetic evidence for recombination in dengue virus. Mol. Biol. Evol. *16*, 405–409.

Hrazdilová, K., Slaninková, E., Brožová, K., Modrý, D., Vodička, R., and Celer, V. (2016). New species of torque teno miniviruses infecting gorillas and chimpanzees. Virology *487*, 207–214.

Huerta-Cepas, J., Serra, F., and Bork, P. (2016). ETE 3: reconstruction, analysis, and visualization of phylogenomic data. Mol. Biol. Evol. *33*, 1635–1638.

Hunter, J.D. (2007). Matplotlib: a 2D graphics environment. Comput. Sci. Eng. *9*, 90–95.

Kaczorowska, J., and van der Hoek, L. (2020). Human anelloviruses: diverse, omnipresent and commensal members of the virome. FEMS Microbiol. Rev. *44*, 305–313.

Kakkola, L., Bondén, H., Hedman, L., Kivi, N., Moisala, S., Julin, J., Ylä-Liedenpohja, J., Miettinen, S., Kantola, K., Hedman, K., and Söderlund-Venermo, M. (2008). Expression of all six human torque teno virus (TTV) proteins in bacteria and in insect cells, and analysis of their IgG responses. Virology *382*, 182–189.

Kapoor, A., Kumar, A., Simmonds, P., Bhuva, N., Singh Chauhan, L., Lee, B., Sall, A.A., Jin, Z., Morse, S.S., Shaz, B., et al. (2015). Virome analysis of transfusion recipients reveals a novel human virus that shares genomic features with hepaciviruses and pegiviruses. mBio *6*, e01466–e01415.

Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res *30*, 3059–3066.

Koonin, E.V., Dolja, V.V., and Krupovic, M. (2021). The healthy human virome: from virus–host symbiosis to disease. Curr. Opin. Virol. *47*, 86–94.

Kosakovsky Pond, S.L., Posada, D., Gravenor, M.B., Woelk, C.H., and Frost, S.D. (2006). GARD: a genetic algorithm for recombination detection. Bioinformatics *22*, 3096–3098.

Lefeuvre, P., Lett, J.M., Varsani, A., and Martin, D.P. (2009). Widely conserved recombination patterns among single-stranded DNA viruses. J. Virol. *83*, 2697–2707.

Lefkowitz, E.J., Dempsey, D.M., Hendrickson, R.C., Orton, R.J., Siddell, S.G., and Smith, D.B. (2018). Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV). Nucleic Acids Res *46*, D708–D717.

Leppik, L., Gunst, K., Lehtinen, M., Dillner, J., Streker, K., and de Villiers, E.M. (2007). In vivo and in vitro intragenomic rearrangement of TT viruses. J. Virol. *81*, 9346–9356.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv, 1303.3997 https://arxiv.org/abs/1303.3997.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094–3100.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics *25*, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Maggi, F., Ricci, V., Bendinelli, M., Nelli, L.C., Focosi, D., Papineschi, F., Petrini, M., Paumgardhen, E., and Ghimenti, M. (2008). Changes in $CD8_{+}57_{+}$ T lymphocyte expansions after autologous hematopoietic stem cell transplantation correlate with changes in torquetenovirus viremia. Transplantation *85*, 1867–1868.

Martin, D.P., Biagini, P., Lefeuvre, P., Golden, M., Roumagnac, P., and Varsani, A. (2011). Recombination in eukaryotic single stranded DNA viruses. Viruses *3*, 1699–1738.

Martin, S. (2021). Alvis. https://github.com/SR-Martin/alvis.

Maynard Smith, J., and Smith, N.H. (1998). Detecting recombination from gene trees. Mol. Biol. Evol. *15*, 590–599.

Meunier, J., and Eyre-Walker, A. (2001). The correlation between linkage disequilibrium and distance: implications for recombination in hominid mitochondria. Mol. Biol. Evol. *18*, 2132–2135.

Moustafa, A., Xie, C., Kirkness, E., Biggs, W., Wong, E., Turpaz, Y., Bloom, K., Delwart, E., Nelson, K.E., Venter, J.C., and Telenti, A. (2017). The blood DNA virome in 8,000 humans. PLoS Pathog *13*, e1006292.

Muhire, B.M., Varsani, A., and Martin, D.P. (2014). SDT: a virus classification tool based on pairwise sequence alignment and identity calculation. PLoS One *9*, e108277.

Nei, M., and Kumar, S. (2000). Molecular Evolution and Phylogenetics (Oxford Press).

Niel, C., Diniz-Mendes, L., and Devalle, S. (2005). Rolling-circle amplification of Torque teno virus (TTV) complete genomes from human and swine sera and identification of a novel swine TTV genogroup. J. Gen. Virol. *86*, 1343–1347.

Ninomiya, M., Takahashi, M., Nishizawa, T., Shimosegawa, T., and Okamoto, H. (2008). Development of PCR assays with nested primers specific for differential detection of three human anelloviruses and early acquisition of dual or triple infection during infancy. J. Clin. Microbiol. *46*, 507–514.

Nishimura, Y., Watai, H., Honda, T., Mihara, T., Omae, K., Roux, S., Blanc-Mathieu, R., Yamamoto, K., Hingamp, P., Sako, Y., et al. (2017). Environmental viral genomes shed new light on virus-host interactions in the ocean. mSphere *2*, e00359–16.

Nishizawa, T., Okamoto, H., Konishi, K., Yoshizawa, H., Miyakawa, Y., and Mayumi, M. (1997). A novel DNA virus (TTV) associated with elevated transaminase levels in posttransfusion hepatitis of unknown etiology. Biochem. Biophys. Res. Commun. *241*, 92–97.

Novocraft. (2019). Novocraft Novoalign. http://novocraft.com/.

Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P.A. (2017). metaSPAdes: a new versatile metagenomic assembler. Genome Res *27*, 824–834.

Okamoto, H., Nishizawa, T., Tawara, A., Peng, Y., Takahashi, M., Kishimoto, J., Tanaka, T., Miyakawa, Y., and Mayumi, M. (2000). Species-specific TT viruses in humans and nonhuman primates and their phylogenetic relatedness. Virology *277*, 368–378.

Park, D.J., Dudas, G., Wohl, S., Goba, A., Whitmer, S.L.M., Andersen, K.G., Sealfon, R.S., Ladner, J.T., Kugelman, J.R., Matranga, C.B., et al. (2015). Ebola virus epidemiology, transmission, and evolution during seven months in Sierra Leone. Cell *161*, 1516–1526.

Parks, D. (2020). CompareM. https://github.com/dparks1134/CompareM.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: machine learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

Qiu, J., Kakkola, L., Cheng, F., Ye, C., Söderlund-Venermo, M., Hedman, K., and Pintel, D.J. (2005). Human circovirus TT virus genotype 6 expresses six proteins following transfection of a full-length clone. J. Virol. *79*, 6505–6510.

R Core Team (2013). R: a language and environment for statistical computing (R Foundation for Statistical Computing).

Rani, A., Ranjan, R., McGee, H.S., Metwally, A., Hajjiri, Z., Brennan, D.C., Finn, P.W., and Perkins, D.L. (2016). A diverse virome in kidney transplant patients contains multiple viral subtypes with distinct polymorphisms. Sci. Rep. *6*, 33327.

Robinson, D.F., and Foulds, L.R. (1981). Comparison of phylogenetic trees. Math. Biosci. *53*, 131–147.

Rognes, T., Flouri, T., Nichols, B., Quince, C., and Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. PeerJ *4*, e2584.

Schmieder, R., and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. Bioinformatics *27*, 863–864.

Sedlazeck, F.J., Rescheneder, P., and von Haeseler, A. (2013). NextGenMap: fast and accurate read mapping in highly polymorphic genomes. Bioinformatics *29*, 2790–2791.

Shen, W., Le, S., Li, Y., and Hu, F. (2016). SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PLoS One *11*, e0163962.

Smyth, R.P., Davenport, M.P., and Mak, J. (2012). The origin of genetic diversity in HIV-1. Virus Res *169*, 415–429.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312–1313.

Stulberg, E., Fravel, D., Proctor, L.M., Murray, D.M., LoTempio, J., Chrisey, L., Garland, J., Goodwin, K., Graber, J., Harris, M.C., et al. (2016). An assessment of US microbiome research. Nat. Microbiol. *1*, 15015.

Takahashi, K. (1998). Partial ~2.4-kb sequences of TT virus (TTV) genome from eight Japanese isolates: diagnostic and phylogenetic implications. Hepatol. Res. *12*, 111–120.

Takahashi, K., Iwasa, Y., Hijikata, M., and Mishiro, S. (2000). Identification of a new human DNA virus (TTV-like mini virus, TLMV) intermediately related to TT virus and chicken anemia virus. Arch. Virol. *145*, 979–993.

Tisza, M.J., Pastrana, D.V., Welch, N.L., Stewart, B., Peretti, A., Starrett, G.J., Pang, Y.-Y.S., Krishnamurthy, S.R., Pesavento, P.A., McDermott, D.H., et al. (2020). Discovery of several thousand highly diverse circular DNA viruses. eLife *9*, e51971.

Tyschik, E.A., Rasskazova, A.S., Degtyareva, A.V., Rebrikov, D.V., and Sukhikh, G.T. (2018). Torque teno virus dynamics during the first year of life. Virol. J. *15*, 96.

Van Rossum, G., and Drake, F.L. (2009). Python 3 Reference Manual (CreateSpace).

Weaver, S., Shank, S.D., Spielman, S.J., Li, M., Muse, S.V., and Kosakovsky Pond, S.L. (2018). DataMonkey 2.0: a modern web application for characterizing selective and other evolutionary processes. Mol. Biol. Evol. *35*, 773–777.

Wick, R. (2018a). filtlong. https://github.com/rrwick/Filtlong.

Wick, R. (2018b). Porechop. https://github.com/rrwick/Porechop.

Wickham, H. (2016). Ggplot2: Elegant Graphics for Data Analysis (Springer-Verlag).

Wood, D.E., and Salzberg, S.L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol *15*, R46.

Woodcroft, B.J., Boyd, J.A., and Tyson, G.W. (2016). OrfM: a fast open reading frame predictor for metagenomic data. Bioinformatics *32*, 2702–2703.

Worobey, M. (2000). Extensive homologous recombination among widely divergent TT viruses. J. Virol. *74*, 7666–7670.

Worobey, M., Rambaut, A., and Holmes, E.C. (1999). Widespread intra-serotype recombination in natural populations of dengue virus. Proc. Natl. Acad. Sci. USA *96*, 7352–7357.

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. *39*, 306–314.

Young, J.C., Chehoud, C., Bittinger, K., Bailey, A., Diamond, J.M., Cantu, E., Haas, A.R., Abbas, A., Frye, L., Christie, J.D., et al. (2015). Viral metagenomics reveal blooms of anelloviruses in the respiratory tract of lung transplant recipients. Am. J. Transplant. *15*, 200–209.

Zhang, W., Wang, H., Wang, Y., Liu, Z., Li, J., Guo, L., Yang, S., Shen, Q., Zhao, X., Cui, L., et al. (2016). Identification and genomic characterization of a novel species of feline anellovirus. Virol. J. *13*, 146.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Human serum	BioLINCC	HLB01910909a
Chemicals, peptides, and recombinant proteins		
T7 endonuclease I	NEB	Cat# M0302
dNTPs	NEB	Cat# N0447L
phi29 DNA polymerase	NEB	Cat# M0269L
g-TUBE	Covaris	Cat# 520079
NEBNext FFPE Repair Mix	NEB	Cat# M6630
NEBNext Ultra II End Repair/dA-Tailing Module	NEB	Cat# E7546
NEBNext Quick Ligation Module	NEB	Cat# E6056
Agencourt AMPure XP Beads	Beckman Coulter	Cat# A63880
PureLink Pro 96 Viral RNA/DNA Purification Kit	Life Technologies Corporation	Cat# 12280096A
Critical commercial assays		
Nextera Flex Kit	Illumina	Cat# 20018705
iSeq 100 i1 Reagent v. 2	Illumina	Cat# 20031374
IDT for Illumina Nextera DNA UD Indexes Set A	Illumina	Cat# 20027213
NSQ 500/550 Hi Output KT v. 2.5 (300 CYS)	Illumina	Cat# 20024908
Ligation Sequencing Kit	Oxford Nanopore Technologies	Cat# SQK-LSK109
D5000 ScreenTape	Agilent	Cat# 5067-5588
PCR Master Mix	Sigma Aldrich	Cat# 11636103001
MinION Flow Cell	Oxford Nanopore Technologies	Cat# FLO-MIN106; Cat# FLO-MIN107
Deposited data		
Raw sequence data	this paper	BioProject: PRJNA679286
Analyzed data	this paper	https://doi.org/10.5281/zenodo.4810962
Oligonucleotides		
Universal Anello primers	Ninomiya et al., 2008	Table S1
12 Anello-specific primers	this paper	Table S1
Software and algorithms		
Kraken v. 1.1.1	Wood and Salzberg, 2014	https://ccb.jhu.edu/software/kraken/
BLAST+ v. 2.6.0+	Camacho et al., 2009	https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastDocs&DOC_TYPE=Download
FastQC v. 0.11.8	Andrews, 2019	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
MultiQC v. 1.7	Ewels et al., 2016	https://multiqc.info/
BBMap v. 38.21	Bushnell, 2014	https://sourceforge.net/projects/bbmap/
NextGenMap v. 0.5.5	Sedlazeck et al., 2013	https://cibiv.github.io/NextGenMap/
BWA v. 0.7.17-r1188	Li, 2013	http://bio-bwa.sourceforge.net/
SAMtools v. 0.1.19-44428cd	Li et al., 2009	http://www.htslib.org/
Picard v. 2.18.26	Broad Institute, 2018	https://broadinstitute.github.io/picard/
MinKNOW v. 19.05.0	n/a	https://nanoporetech.com/
porechop v. 0.2.4	Wick, 2018b	https://github.com/rrwick/Porechop
filtlong v. 0.2.0	Wick, 2018a	https://github.com/rrwick/Filtlong
minimap2 v. 2.18-r1015	Li, 2018	https://github.com/lh3/minimap2

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Alvis v. 1.2	Martin, 2021	https://github.com/SR-Martin/alvis
Geneious Prime v. 2021.1.1	Biomatters, 2021	https://www.geneious.com/prime/
SPAdes v. 3.13.0	Nurk et al., 2017	https://cab.spbu.ru/software/spades/
PRINSEQ v. 0.20.4	Schmieder and Edwards, 2011	http://prinseq.sourceforge.net/
usearch v. 11.0.667_i86linux32	Edgar, 2010	https://www.drive5.com/usearch/
ccfind v. 1.4.5	Nishimura et al., 2017	https://github.com/yosuken/ccfind
OrfM v. 0.7.1	Woodcroft et al., 2016	https://github.com/wwood/OrfM
Python v. 3.8.3	Van Rossum and Drake, 2009	https://www.python.org/download/releases/3.0/
SeqKit v. 0.10.0	Shen et al., 2016	https://bioinf.shenwei.me/seqkit/
MEME Suite v. 5.1.1	Bailey et al. 2009	https://meme-suite.org/meme/
Novoalign v. 3.09.00	Novocraft, 2019	http://www.novocraft.com/products/novoalign/
R v. 3.5.1	R Core Team, 2013	https://www.r-project.org/
ggplot2 v. 3.3.2	Wickham, 2016	https://github.com/tidyverse/ggplot2
MAFFT v. 7.407	Katoh et al., 2002	https://mafft.cbrc.jp/alignment/software/changelog.html
Scikit-learn v. 0.23.2	Pedregosa et al., 2011	https://scikit-learn.org/stable/
matplotlib v. 3.3.4	Hunter, 2007	https://matplotlib.org/stable/index.html
RAxML v. 8.2.11	Stamatakis, 2014	https://cme.h-its.org/exelixis/web/software/raxml/
PhyML v. 3.3.20180621	Guindon et al., 2010	http://www.atgc-montpellier.fr/phyml/
ETE v. 3.1.2	Huerta-Cepas et al., 2016	http://etetoolkit.org/
ClonalFrameML v. 1.11-3-g4f13f23	Didelot and Wilson, 2015	https://github.com/xavierdidelot/ClonalFrameML
Datamonkey Server - GARD	Weaver et al., 2018	https://www.datamonkey.org/
Sequence Demarcation Tool (SDT) v. 1.2	Muhire et al., 2014	http://web.cbio.uct.ac.za/~brejnev/
CompareM v. 0.1.2	Parks, 2020	https://github.com/dparks1134/CompareM
Other		
iSeq 100	Illumina	Cat# 20021532
NextSeq 550	Illumina	Cat# SY-415-1002
MinION Mk1B	Oxford Nanopore Technologies	Cat# MIN-101B
Human reference genome NCBI build 37, GRCh37	Genome Reference Consortium	http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/
Resource website for this publication	this paper	https://doi.org/10.5281/zenodo.4810962

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Nathan L. Yozwiak (nyozwiak@ringtx.com).

Materials availability

No new unique reagents were generated by this study. Raw data and code created as part of the study can be found on public resources and public data repositories as specified in the data and code availability section.

Data and code availability

Raw sequence data and anellovirus sequences generated for this study can be found under NCBI BioProject: PRJNA679286. Analytical code and data used to generate tables and figures can be found at https://github.com/ring-therapeutics/anellome_paper.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Serum samples were obtained from the TTVS (accession no. HLB01910909a). The study protocol for this trial was approved by the institutional biosafety committees and review boards at each site, and patients signed written informed consent and release of medical information forms before screening. Further information in regard to cohort size, gender, age, and disease status can be found in Table S1.

METHOD DETAILS

Sample collection

Each set of longitudinal recipient samples was paired with one or more unique donor samples. Nucleic acids were extracted from 200 μL serum with a purelink viral DNA/RNA kit from Invitrogen. The samples were processed according to manufacturer's protocol with an increase to 60 min for the proteinase K incubation. Samples were eluted in 50 μL of nuclease-free water.

Multiply primed RCA

Degenerate amplification primers were designed to cover well-conserved regions based on alignments of Anelloviridae genomes that were generated from published genomes on PubMed and those isolated from metagenomic databases. Primers were protected by two thiophosphate modifications between each of the last three nucleotides at the 3′. The RCA (termed Anello-RCA) contained a premix of 12 Anello-specific primers (Table S1) at a final concentration of 0.4 μM each, 1× phi29 DNA polymerase buffer (New England Biolabs), 2 μL DNA sample, and dH_20 in a final volume of 10 μL. The DNA mixture was then denatured at 95°C for 3 min and then cooled to 4°C, before being put on ice. The denatured sample was then added to 10 μL of the amplification solution which contained the 12 Anello specific primers at a final concentration of 0.4 μM each, 1× phi29 DNA polymerase buffer (New England Biolabs), 200ng/ul bovine albumin serum, 1mM dNTPs, 2U/μL phi29 polymerase, and dH_20. The sample was incubated at 30°C for 20 h followed by inactivation of the enzyme at 65°C for 10 min. All incubations were done in a Mastercycler X50s (Eppendorf). The final product was then diluted to 50 μL by adding nuclease free water.

Pan-Anello PCR

The presence of *Anelloviridae* in serum samples was tested by PCR with pan-anello primers developed by Ninomiya et al. (2008). 10 μL of sample was added to 1× PCR Master Mix (Sigma-Aldrich, Basel Switzerland) and the 4 degenerate primers at a final concentration of 1 μM each in a final volume of 25 μL. Positive samples were identified by the presence of the 128 base pair band in a 2% agarose gel.

Illumina library preparation and sequencing

Post RCA DNA was diluted to a total volume of 50 μL to reduce viscosity of the samples and then the concentration of DNA was assessed by Qubit. Nextera Flex (Illumina, San Diego CA, USA) kit was used to prepare the samples for sequencing following the manufacturer's protocol for 100–500 ng input. Library QC was carried out with D5000 screen tape on an Agilent Tapestation 4200. All libraries were then sequenced on either an Illumina iSeq 100 or a NextSeq 550. Replicates for multiple samples were sequenced over the course of the study to verify the effectiveness of the amplification protocol and, when necessary, to validate samples with low yields (Table S4)

Nanopore library preparation and sequencing

Post RCA DNA was debranched and fragmented to 20 kb sized fragments following NanoAmpli-Seq (Calus et al., 2018) protocol. 4.5 μg of RCA material was diluted in 65 μL of nuclease-free water and treated with 2 μL of T7 endonuclease I (New England Biolabs) for 5 min at RT. The reaction was then loaded in a g-TUBE (Covaris) and centrifuged at 1800 rpm for 4 min. The g-TUBE was then reversed, and the centrifugation process was repeated. An additional round of T7 endonuclease and g-TUBE was performed before the mixture was then cleaned up with SPRI beads at a ratio of 1.8× with a final elution in 20 μL of nuclease-free water. The concentration of DNA was assessed by Qubit. The samples were then library prepared with the SQK-LSK-109 (Oxford nanopore technologies) kit following manufacturer's protocol. Libraries were loaded onto a R9.4 (FLO-MIN106) or R9.5 (FLO-MIN107) flow cell and placed onto the MinION Mk1B (Oxford nanopore technologies) and run for 48 h. Only flow cells that passed the manufacturer's flow cell check test were used.

Retrieval of public anellovirus sequences

Publicly available anellovirus sequences were retrieved from NCBI's GenBank (Benson et al., 2012) repository with the following query terms: the anellovirus taxonomy identifier *txid687329[Organism:exp]*, a length filter between 2,500 and 5,000 bp (*("2500"[SLEN] : "5000"[SLEN])*), and removing any patented sequences, *NOT patent[WORD]*. Sequences here were further filtered down with a custom Python (Van Rossum and Drake, 2009) script to filter out any non-human anellovirus retrieved resulting in a final set of 454 anellovirus sequences.

Genomic anellovirus sequence identification

An initial read-out of anellovirus content in raw sequencing reads were generated by Kraken (Wood and Salzberg, 2014) with default parameters against a custom in-house constructed anellovirus database. These resultant classified sequences were further verified by NCBI's BLASTN (Camacho et al., 2009), against the nucleotide database (Benson et al., 2012) with default parameters, to confirm that the output from kraken were valid anellovirus sequences.

Illumina sequence quality control

Raw sequencing reads were subjected to quality control utilizing FastQC (Andrews, 2019) on each paired-end read set to measure various statistics in regard to each sequencing run. All of the FastQC generated reports were aggregated into a single report with MultiQC (Ewels et al., 2016). Metrics from these reports influenced parameter selection to quality control steps further downstream during analysis.

Low quality sequence data and common adapters were removed with bbduk (Bushnell, 2014) with the following parameters: *ktrim=r*, *k=23*, *mink=11*, *tpe=t*, *tbo=t*, *qtrim=rl*, *trimq=20*, *minlength=50*, *maxns=2*. The supplied contaminant file was assembled by pulling target contaminant sequences from NCBI GenBank covering several bacterial species as well as human genetic elements to be removed. An accession list containing specific sequences is provided in the supplemental information.

Next, human sequences were removed in two passes with both NextGenMap (Sedlazeck et al., 2013) and BWA (Li, 2013; Li and Durbin, 2009) against the GRCh37/hg19 build of the human reference genome. NextGenMap was run with parameters *–affine*, *-s 0.7*, and *-p*, and BWA was run with default parameters. Mapped reads output in SAM file format were converted to paired-end FASTQ format with both SAMtools (Li et al., 2009) and Picard's (Broad Institute, 2018) SamToFastq utility configured with the parameter VALIDATION_STRINGENCY="silent".

rRNA contaminants and common laboratory bacterial contaminants were removed with bbmap (Bushnell, 2014) with the following parameters: *minid=0.95*, *bwr=0.16*, *bw=12*, *quickmatch=t*, *fast=t*, *minhits=2*. An accounting of all reference sequences screened against can be found in the provided supplementary data.

Finally, we de-duplicated the short read data passing all QC and decontamination steps to speed up and aid in genome assembly quality by using clumpify (Bushnell, 2014) configured with the parameter *dedupe=t*.

Nanopore sequence quality control and mapping

Nanopore reads were base called and demultiplexed with MinKNOW software. Adapter sequences were trimmed with porechop (Wick, 2018a) with default parameters followed by quality and length filtering using filtlong with parameters *–min_length 2000 –keep_percent 90* (Wick, 2018b). Reads passing quality control were mapped to anellovirus contig sequences derived from cohort subject R04 with minimap2 (Li, 2018) with the following parameters: *-cx map-ont*. The resultant PAF file was both visualized in Alvis (Martin, 2021) and parsed to identify best hits to the reference contig sequences and these reads were further analyzed with pairwise alignments in Geneious (Biomatters, 2021) with the MAFFT alignment plug-in with the G-INS-i algorithm. These long reads were used to validate the assembled short-reads and to verify that these contigs were not chimeras.

Genome assembly

Trimmed, decontaminated and de-duplicated sequencing data were assembled with metaSPAdes (Nurk et al., 2017) skipping the error correction module via the use of the *–only-assembler* parameter. Assembled contigs were filtered with PRINSEQ lite (Schmieder and Edwards, 2011) configured with parameters *out_format 1*, *-lc_method dust*, and *lc_threshold 20*. Contigs assembled from each sample were clustered at 99.5% similarity to remove any duplicate sequences via the VSEARCH software's *cluster_fast* algorithm (Rognes et al., 2016). Putative complete, circular genomes were recovered from assembled contigs with ccfind (Nishimura et al., 2017) with all parameters set to defaults, producing 190 candidate circular genomes.

Genome annotation

ORF sequences were called from assembled contigs with OrfM (Woodcroft et al., 2016) with parameters configured to print stop codons (*-p*) and print ORF's in the same frame as a stop codon (*-s*) and constrained to ORF sequences no shorter than 50 amino acids (*-m 150*).

Predicted ORF sequences were further filtered with seqkit's *seq* and *grep* utilities (Shen et al., 2016) to subdivide ORF sequences into ORF1, ORF2 and ORF3. ORF1 sequences were identified by filtering ORF sequences with seqkit *grep* for those no shorter than 600 amino acids (*-m 600*) and seqkit *grep* to search just sequence data (*-s*), enable regex pattern searching (*-r*) and by querying for conserved motif YNPX^2DXGX^2N (*-p "YNP.{2}D.G.{2}"*). ORF2 sequences were identified with conserved motif WX^7HX^3CXCX^5H previously identified in literature (Takahashi et al., 2000) through seqkit's *grep* utility (*-p "W.{7}H.{3}C.C.{5}H"*).

In addition to ORF1 and ORF2, a third open reading frame (ORF3) was predicted near the 3′ end of ORF1 in 471 anellovirus genomes in the TTVS dataset. ORF3 uses a STOP codon downstream from the one used by ORF1 and its reading frame is different from that of ORF1 and ORF2. A protein in the ORF3 reading frame, labelled ORF2/3, has previously been characterized in human anelloviruses (Qiu et al., 2005) and studies on anelloviruses infecting other species, such as seals, cats, and gorillas (Fahsbender et al., 2017; Zhang et al., 2016; Hrazdilová et al., 2016) have shown evidence for ORF3. However, little is known about the functional significance of the protein encoded by this open reading frame. Parsing the 471 ORF3 sequences (median length: 68 aa; minimum length: 50 aa; maximum length: 159 aa) through MEME (Bailey et al., 2009) revealed the presence of two previously unknown and

highly conserved motifs located near the 3′ end of ORF3. Motif 1 (26 aa) was observed in 467 out of the 471 sequences (99%) while Motif 2 (5 aa) was observed in 463 out of the 471 sequences (98%) (Figures S2B and S2C).

ORF sequences identified as ORF1, ORF2 or ORF3 frequently contained peptides upstream of the canonical start codon as per the functionality of OrfM. These sequences were trimmed to the proper start and stop codons via an in-house written python script that searched for the first methionine located from the 5′ end and in the cases of ORF1 the start codon was predicted by first locating the arginine-rich region and locating the first methionine upstream. In some cases, a non-canonical start codon was predicted as the ORF1 start codon by searching for the amino acids threonine-proline-tryptophan or threonine-alanine-tryptophan just upstream of the arginine-rich region.

Genera classification

Binning individual anellovirus sequences into one of the three genera was accomplished by homology search of ORF1 sequences using tblastx against a custom in-house typing database consisting of 728 curated anellovirus sequences. Top hits with suitable coverage across the majority of ORF1 sequences were used to classify assembled sequences.

Anellovirus proportion estimation

Estimates of the proportion of individual anellovirus lineages in each sample/longitudinal timepoint for donor-recipient datasets were estimated by identifying the unique set of lineages present across each donor sample by clustering ORF1 sequences at 97.5% similarity with the USEARCH software (Edgar, 2010) and the *cluster_fast* algorithm. These unique donor-derived anellovirus lineages were then searched for in recipient longitudinal samples by mapping the derived short read sequencing data against them with Novoalign software (Novocraft, 2019) with the following parameters: *-H 15*, *-l 30*, *-t 500*, *-r Random*, *-g 50*, *-x 6*, *-F STDFQ*.

The resulting BAM mapping files were used to calculate relative anellovirus proportion estimates for each donor lineage by custom script with the formula below:

$$\text{anellovirus lineage relative proportion} = \frac{\text{reads mapped to lineage}}{\text{total mapped anellovirus reads}} \times 100$$

The relative proportions of all donor lineages in each donor-recipient dataset were collated together into one tab-delimited file for further downstream analysis. Steam graph figures (Figure 3A) depicting anellovirus proportion shifts over time in subjects were generated with R (R Core Team, 2013) using the ggplot2 (Wickham, 2016, p. 2).

Multidimensional scaling

Publicly available and newly described anellovirus sequences were split into *Alphatorquevirus*, *Betatorquevirus*, and *Gammatorquevirus* (689, 619, and 271 sequences, respectively) and trimmed to the ORF1 region. ORF1 sequences were translated and aligned with MAFFT (FFT-NS-i ×1000 setting) (Katoh et al., 2002) and pairwise distances (as percent dissimilarity) between amino acid sequences computed. All three alignments (alpha, beta, and gamma) were then consensus aligned with MAFFT (G-INS-i setting).

As a point of comparison against other viral surface proteins and to help interpretability of MDS plots shown in Figure 1 the following datasets were downloaded from GenBank: human papillomavirus (all HPV, including diverged HPV type 41) late protein (L1), adeno-associated virus (AAV, all diversity found in humans) capsid protein, Dengue virus (all known serotypes) envelope protein, Middle East respiratory syndrome-associated coronavirus (MERS-CoV, all known diversity) Spike protein (S), homologous *Filoviridae* (encompassing genera Marburgvirus, Dianlovirus, Cuevavirus, and Ebolavirus) glycoprotein (GP) protein, and Lassa fever (all known diversity) virus glycoprotein complex (GPC) protein. In brief, surface protein sequences were extracted, translated, and aligned through an iterative procedure of aligning sequences with MAFFT, exclusion of sequences with apparent long insertions relative to other sequences and repeated alignment until at least 1% of amino acid alignment columns were identical. Additional datasets for influenza A virus group 2 haemagglutinin (HA) sequences were downloaded from Influenza Research Database and human immunodeficiency virus-1 (HIV-1) env sequences from Los Alamos National Laboratory pre-made alignment sequence database. Sequences were translated and aligned with MAFFT (auto setting) and down-sampled to 3,000 sequences. The choice of viruses to be represented was a mixture of viruses with known extensive surface protein diversity, readily available datasets, and relevance. HPV, HIV-1, influenza, and Lassa fever virus datasets were chosen because sequence data were numerous and surface proteins known to be diverse within their respective fields. The choice of Dengue virus was largely motivated by both availability of sequence data and its recognized division into distinct serotypes. Adeno-associated virus was chosen because of its use as a gene vector platform with identified limited antigenic diversity. Filoviruses and MERS-CoV were chosen entirely arbitrarily out of conveniently available datasets, with the former being of general interest and the latter being known to have very limited diversity.Multidimensional scaling (MDS) was applied to all viral protein sequences to project them into two dimensions with Scikit-learn (Pedregosa et al., 2011). Agglomerative clustering was additionally applied to pairwise amino acid distances of anelloviruses with Scikit-learn to get 10 (arbitrarily chosen for ease of comparison and visualization) clusters. MDS-projected sequences were visualized with matplotlib (Hunter, 2007) and colored by assigned clusters in the case of anelloviruses.

Phylogenetic analysis

The combined anellovirus amino acid sequence dataset of 2,101 translated ORF1 sequences was used to reconstruct a maximum likelihood phylogenetic tree with RAxML (CAT sequence evolution model, BLOSUM62 substitution matrix) (Stamatakis, 2014).

Recombination analyses

Translationally aligned sequences of the three anellovirus genera were grouped into clusters where all members were at least 80% identical to another member at nucleotide level. This resulted in 28 clusters with more than 10 members (23 clusters of *Alphatorquevirus*, four of *Betatorquevirus* and one of *Gammatorquevirus*). A single representative of each genus was chosen for a closer analysis, giving clusters with 23 *Alphatorquevirus*, 11 *Betatorquevirus*, and 10 *Gammatorquevirus* sequences.

Next, sequences within each cluster were realigned with MAFFT (the slower but more accurate E-INS-i setting) to improve the alignments. Then, each alignment was split into 500 nucleotide fragments and phylogenies inferred from each fragment with PhyML (HKY+Γ_4 substitution model) (Guindon et al., 2010; Hasegawa et al., 1985; Yang, 1994) and midpoint rooted. Phylogenies derived from neighboring fragments were then displayed in a tangled chain where each taxon is tracked through successive trees. Robinson-Foulds distances (Robinson and Foulds, 1981) between neighboring trees were computed with the ETE 3 toolkit (Huerta-Cepas et al., 2016).

The same cluster alignments, undivided, were used to infer single trees with PhyML (HKY+Γ_4 substitution model). Each tree and alignment were then used to reconstruct the mutations that occurred across the tree with ClonalFrameML v. 1.11-3-g4f13f23 (Didelot and Wilson, 2015) with kappa set to 2.0. For every mutation that was reconstructed to only occur once in the tree the branch where the mutation occurred was marked with ticks and every mutation that was inferred to occur more than once in the tree was indicated by a line connecting the mutation to its identical counterparts (i.e., reversions are considered separately) elsewhere in the tree.

To statistically confirm excessive amounts of homoplasies and alternating topologies seen in cluster alignments, we additionally analyzed the same data with GARD (Kosakovsky Pond et al., 2006) by using the Datamonkey server (Weaver et al., 2018). Each alignment was found to contain at least six statistically supported breakpoints: Akaike information criterion (AIC) support for baseline (single tree) versus best model (multiple breakpoints) was 522.79 (*Alphatorqueviruses*, seven breakpoints), 361.43 (*Betatorqueviruses*, eight breakpoints), and 339.47 (*Gammatorqueviruses*, seven breakpoints). The hypervariable region located in the middle of ORF1 was consistently identified as having the shortest fragments and therefore a likely recombination hotspot. However, this could also reflect the lack of functional constraint in the region which would result in accumulation of more diversity and thus vastly improved statistical power to detect recombination there compared to other regions of ORF1.

Finally, we inferred the decay of LD by using the χ^2_{df} statistic (Hedrick and Thomson, 1986), which behaves identically to the more common r^2 statistic for biallelic loci. To this end we used the genus-wide alignments with 689 *Alpha-*, 619 *Beta-*, and 271 *Gammatorquevirus* sequences. Alignment columns with fewer than 10% valid sites (A, C, T, or G) were ignored, as were sites where the minority variant was at lower than 5% frequency. LD measured between pairs of variable sites was then plotted against the distance between sites with mean LD calculated in windows 100 nucleotides long.

Because complete circularized genomes were available for a number of anelloviruses we also sought to gauge the degree of reticulate evolution in non-coding parts of the genome. To this end, we aligned complete genomes of each genus (22 *Alpha-*, 467 *Beta-*, and 23 *Gammatorquevirus*) and extracted the non-coding regions followed by ancestral state reconstruction by using ClonalFrameML, as described in the preceding paragraphs. To identify putative recombination tracts, we looked for repeat mutations (homoplasies) occurring in clusters of at least three mutations within ten nucleotides of each other. Figures S5A–S5C show these tracts for *Alphatorquevirus*.

Pairwise sequence identity analysis

Pairwise identity comparisons at the nucleotide and amino acid levels of all anellovirus lineages derived from the TTVS cohort were computed with Sequence Demarcation Tool (SDT) (Muhire et al., 2014). Each set of sequences was run through SDT with the *mafft* option to specify alignment with MAFFT. Output identity measurements were parsed with custom R (R Core Team, 2013) scripts (included in code repository) to generate Figure S3B.

Pairwise AAI comparisons between anellovirus lineages were computed with the CompareM toolkit (Parks, 2020). All 1,656 lineages derived from the TTVS cohort were first split into their own individual FASTA files with the *seqkit split* command with the *-i* parameter to split by sequence identifier. The directory containing these FASTA files was used as input to CompareM's *aai_wf* command to compute the mean AAI values between each lineage. The resulting CSV file was fed into the custom R scripts (included in code repository) to generate Figure S3A and other mean AAI-related metrics

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical analyses were conducted in Python and R as described in the method details section.