The electronic tree of life (eTOL): a net of long probes to characterize the human microbiome from RNA-seq data

Xinyue Hu
University of Edinburgh
Jürgen Haas
University of Edinburgh
Richard Lathe (richard.lathe@ed.ac.uk)
University of Edinburgh

Research Article

Keywords: Archaea, Bacteria, BLAST, brain, disease, Fungi, microbiome, RNA-seq, Tree of Life, virus

Posted Date: December 29th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1208849/v1

Abstract

Background

Microbiome analysis generally requires PCR-based or metagenomic shotgun sequencing, sophisticated programs, and large volumes of data. Alternative approaches based on widely available RNA-seq data are constrained because of sequence similarities between the transcriptomes of microbes/viruses and those of the host, compounded by the extreme abundance of host sequences in such libraries. Current approaches are also limited to specific microbial groups. There is a need for alternative methods of microbiome analysis that encompass the entire tree of life.

Results

We report a method to specifically retrieve non-human sequences in human tissue RNA-seq data. For cellular microbes we used a bioinformatic 'net', based on filtered 64-mer small subunit rRNA sequences across the Tree of Life (the 'electronic tree of life', eTOL), to comprehensively (98%) entrap all non-human rRNA sequences present in the target tissue. Using brain as a model, retrieval of matching reads, re-exclusion of human-related sequences, followed by contig building and species identification, is followed by reconfirmation of the abundance and identity of the corresponding species groups. We provide methods to automate this analysis. A variant approach is necessary for viruses. Again, because of significant matches between viral and human sequences, a 'stripping' approach is essential. In addition, contamination during workup is a potential problem, and we discuss strategies to circumvent this issue. To illustrate the versatility of the method, we report the use of the eTOL methodology to unambiguously identify exogenous microbial and viral sequences in human tissue RNA-seq data across the entire tree of life including Archaea, Bacteria, Chloroplastida, basal Eukaryota, Fungi, and Holozoa/Metazoa, and discuss the technical and bioinformatic challenges involved.

Conclusions

This generic methodology may find wider application in microbiome analysis including diagnostics.

Introduction

There is growing interest in the role that the microbiome plays in health and disease. Microbes are ubiquitous and are present in all body compartments including the gut, lung, oral and nasal cavities, as well as in body tissues including liver and kidney, as highlighted by the Human Microbiome Project [1, 2] (https://hmpdacc.org/).

Microbiome characterization is generally based on two key techniques – ribosomal RNA (rRNA) analysis, and metagenomics (reviewed in [3–10]. In the first, rRNA or rDNA sequences are amplified using a suite of

PCR primers, sequenced, and compared against the database. Because this depends on the use of short PCR primers, the method may lack specificity – risking amplification of unrelated sequences – in addition to missing any species whose rRNA sequence diverges from the primers. In addition, differences in abundance require quantitative PCR amplification of each sample followed by deep sequencing, and small differences in abundance (e.g., 2–4-fold changes) are difficult to detect.

The second method is based on metagenomic analysis through shotgun sequencing and genome assembly. This requires many-fold more data, often reaching Tb levels, and requires dedicated tools to remove human sequences and to assemble contigs. Moreover, because high sequencing depth is necessary, assembly-based methods are restricted to highly abundant members of the microbiome. Moreover, metagenomics does not easily address differential abundance. Both methods require extensive wet-lab work and can require machine-learning tools to unravel the true extent of the microbiome [11].

An ancillary technique based on short *k*-mers (generally 31-mers), such as Kraken [12] and CLARK [13], can be employed in conjunction with both metagenomics and RNA-seq, but also requires careful interpretation.

Each of these techniques has advantages and disadvantages. In addition to relatively high demands on data processing, and sometimes low selectivity, the different methodologies have often given conflicting results. We illustrate this through studies on the brain.

Like other tissues, the brain carries a burden of endogenous microbes and viruses [14], although these have not been well characterized, and some have questioned whether there is indeed a brain microbiome [15]. However, an increasing body of evidence suggests that microbes readily enter the brain (e.g. [16], can be detected by in situ immunohistochemistry of brain [17], and that infection may play a role in neurodegenerative disorders such as Alzheimer disease (AD) [18]. Nevertheless, there has been extensive debate about which microbes are present. Chronic inflammation and infection caused by spirochetes have been suggested to contribute to the slow progression of AD [19]. *Chlamydia pneumoniae* shows associations with late-onset AD [20], and other bacteria such as Proteobacteria, Actinobacteria, and Firmicutes, as well as Fungi such as *Malassezia*, *Alternaria*, and *Candida* spp., have been reported [21]. An important PCR-based study revealed multiple bacterial species in AD brain [22]. Other work has focused on periodontal pathogens such as *Porphyromonas gingivalis* [23]. However, in other studies very few microbes of this class (Bacteriodetes) were found, and other key species implicated such as spirochetes and *Chlamydia* (see above) were not well represented [21, 22]. Furthermore, beyond specific target groups, the relative abundances of these different species have not been established.

In addition, the majority of studies to date have focused on select microbial groups such as bacteria or fungi. Several classes of cellular microbes across the known Tree of Life have not been widely studied to date, including Archaea, Amoebozoa, Chloroplastida, and Eukaryota. This relative dearth of such broader analysis extends more generally beyond our target tissue (the brain) to tissues analyzed in the majority of endeavors.

Many viruses are also present in human tissues. Herpes simplex virus 1 (HSV-1) sequences were discovered in AD brain around 30 years ago [24]. Infections with viruses such as HSV-1 are widespread in the population; these generally remain in a silent (latent) form life-long, but may be reactivated because of stress, inflammation, or other factors, leading to proliferation and localized damage. Other viruses, notably human herpesvirus 6A (HHV-6A) and 7 (HHV-7), have also been suggested to be associated with AD. Readhead *et al.* used a modified ViromeScan workflow [25] and found that the abundance of these two viruses among 515 viral species was increased in the transcriptomes of AD brain in three of four cohorts compared to normal brain. For specific viruses, unique 31-mers were generated with Jellyfish [26], which can efficiently count *k*-mers up to 31 nt in length efficiently, and BLAST, which was used to filter the 31-mers with homology to human sequences. The RNA-seq reads that are possible human or bacterial sequences were filtered and then mapped to the filtered 31-mers. The viral abundance analysis was performed with 31-mer count matrices [25].

However, Chorlton [27] challenged these results, suggesting that uncorrected *P* values were used in multiple testing, low-complexity sequences were abundant in the datasets, local alignment by Bowtie2 required only a relatively low number of matching bases, and the BM tagger (ftp://ftp.ncbi.nlm.nih.gov/pub/agarwala/bmtagger/) had a 2% false negative rate when filtering human-derived reads. Also, hepatitis C virus and eradicated variola virus were found in 100% and 97.5% of samples, respectively, using the modified ViromeScan. KrakenUniq [28], which performs efficient *k*-mer counts in metagenomics and can better identify false-positive reads, found few HHV-6A reads and no HHV-7 reads were detected [27]. Allnut *et al.* used digital droplet PCR (ddPCR) to amplify specific regions of HHV-6A and HHV-6B in 708 brain sections [29]. PathSeq [30], which has high specificity and sensitivity in distinguishing between human and non-human sequences, was also used as a complementary method with RNA-seq data, which contained part of the cohorts also used by Readhead *et al.*, to screen for pathogens from more than 25,000 microbes, containing 118 human viruses. Neither of these methods found associations between HHV-6 and AD [29], and the true contribution of herpes and other viruses to human brain disease remains unknown.

In addition to viruses, endogenous retroviruses and retroelements constitute a further class of replicative elements that might also contribute to human disease.

Key objectives in the present report have been to devise methods that (i) extend to the entire tree of life, (ii) can accurately determine both the identity and abundance of microbes including viruses and retroelements in human tissues, (iii) are applicable to widely available RNA-seq datasets such as those listed at the National Center for Biotechnology Information (NCBI) database of sequence read archives (which now contain over 25 million Terabases of open-access sequence information https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?), (iv) do not require sophisticated computer expertise or dedicated computer programs other than those that are widely available and/or freely downloadable, and (v) allow a pictorial representation of the microbiome that facilitates comparative interpretation of the results.

We report the development of two related methods. First, an electronic tree of life (eTOL) approach based on a 'net' of 16S/18S rRNA sequences across all cellular lifeforms to comprehensively retrieve all non-human sequences. Second, a 'stripping' method based on viral genomes to unambiguously detect key viral species. In this methodology paper we focus on technical and bioinformatic issues – case studies are presented that illustrate the versatility of the method; application of the eToL methodology to the human microbiome in select tissues will be presented elsewhere (in preparation).

Methodology Development

Our methodology is centrally based on BLAST screening of RNA-seq libraries using a suite of filtered probes (64-mers for microbial rRNA and whole-genome viral sequences). We present here the detailed rationale behind this approach and discuss relevant technical issues.

Cellular microbes: the electronic Tree of Life (eToL) approach

According to the three-domain system, cellular life can be classified into Archaea, Bacteria, and Eukaryota [31] . To address the full diversity of cellular lifeforms we consulted the Open Tree of Life (OToL; https://tree.opentreeoflife.org/), a National Science Foundation (NSF) collaborative effort across 10 institutions that synthesizes phylogenetic trees based on sequence and taxonomic data [32-34] . For completeness, competing phylograms include LifeMap {National Center for Biotechnology Information (NCBI) version; http://lifemap-ncbi.univ-lyon1.fr/}, and we refer to this where appropriate. Because OToL is not definitive on the placement of bacteria, we followed Schulz *et al.* [35] for a recent re-evaluation of the evolutionary phylogeny of bacteria, and synthesized a compromise ToL that includes all taxonomic groups, extending from Archaea and Bacteria to include Amoebozoa, Chloroplastida, Fungi, basal Eukaryota, and Holozoa/Metazoa (**Figure 1**).

From this tree 126 species were selected that were judged to be equally divergent from each other as estimated by their spacing/position on the tree, taking into account the relative diversity of Fungi in particular. Full-length 16S/18S sequences were downloaded from NCBI. To address the diversity of these sequences, and potential duplications within this dataset, we built a phylogenetic tree using Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/). After exclusion of very close relatives (near-duplicates), we recompiled a list of 120 rRNA sequences that span the full diversity of cellular lifeforms (**Table S1** in the supplementary data online). These sequences were then aligned using MUSCLE (v3.8.31) [36] , and the IQTree tool (v1.6.6) was used to build a maximum likelihood tree with the aligned sequences [37] . To find the best-fit model and estimate the phylogenetic support of the nodes, the ModelFinder (MFP) option [38] , and -bb option [39] with 1000 bootstrap replicates were chosen, respectively. The phylogenetic

tree was modified and rooted with the outgroup bacteria using iTOL tool (v6) [40] , as shown in **Figure 2**.

As will be noted from the figure, it is not entirely consistent with current trees for cellular organisms, and some phylogenetically distinct classes showed intermingling in their precise 16S/18S rRNA sequences. However, the objective of this work was not to address the phylogenetic relationships between the diverse species, but instead to devise a means to detect/extract, as far as possible, a comprehensive compendium of non-human sequences in human tissue. We therefore proceeded to use this collection of 120 rRNA sequences as a route towards evaluating the true complexity of cellular lifeforms.

Databases

The RNA-seq datasets used in this work are listed in **Table S2**.

Probe length and specificity

Probing for microbes in RNA-seq data has generally exploited the fact that a typical cell contains many thousands of ribosomes (**Box 1**), whereas housekeeping transcripts are less well represented, often by a factor of 100 or more. However, a central problem in using rRNA probes to detect specific microbial matches in human RNA-seq data is that many, if not the majority, of full-length sequence probes detect significant (according to conventional criteria regarding the likelihood of detection by chance) matches, albeit often partial, in human rRNA. The problem is accentuated with long probes (bacterial 16S rRNA is ~1.5 kb in length, eukaryotic 18S rRNA is ~1.9 kb in length; for comparison, bacterial 23S large subunit rRNA is ~3 kb, and eukaryotic 28S rRNA is ~5 kb in length). As probe length increases, so too does the likelihood of finding matches in human sequences. Conversely, using short sequences (in the range of 20–30 nucleotides) for PCR or *k*-mer analysis risks losing specificity, even though (at high stringency) these should be unique in the human genome/transcriptome, but this does not take into account the diversity of polymorphisms in the human population, *de novo* mutations, and sequencing errors. Given that *in silico* analysis by BLAST is formally equivalent to wet-lab probing by nucleic acid hybridization, we based our design on previous calculations that a minimum probe length of

62 nt is required, at a biologically plausible/significant level of 85% identity, to detect a unique sequence in a random collection of nucleotides of the size of the human genome (ca 3 x 10e9 nt) with a likelihood of 0.1 of encountering a similar sequence by chance [41] . For simplicity, we adopted a probe length of 64 nt.

rRNA hypervariable or constant regions

rRNA genes of bacteria and fungi, in particular, are known to contain hypervariable regions [42, 43] as well as conserved regions. Although the use of hypervariable regions has the advantage that it can identify exact species (or groups of species), the drawback is that probes based exclusively on hypervariable regions are likely to miss other species that have a slightly different sequence, arguing against pre-selection. By contrast, probes based entirely on conserved regions may detect species irrespective of their class. As a further argument against pre-selection of regions, rRNAs contain a high degree of secondary structure that can (unpredictably) impede reverse transcriptase-mediated copying into cDNA, and some pre-designed probes may find few matches in RNA-seq archives. To illustrate, the number of RNA-seq reads matching different regions of *E. coli* 16S rRNA in a test dataset differed by a factor of 100 (not presented). As a working compromise, we generated probe sequences without any pre-selection based on knowledge of the target sequence (e.g., of variable versus conserved regions). Probe redundancy (a potential outcome of random design) is addressed in the sections below.

Probe generation and nomenclature

We generated 64-mer probes from the 120 rRNA sequences (~10 per kb, noting that not all available sequences are complete). The non-overlapping 64-mer sequences were devised as probes by semi-random selection using probe.py script. The version of python was v3.6.3 [44] . This generated a list of 1323 probes. In naming the probes we aimed to devise a simple, but easily remembered, nomenclature. Probes were therefore prefixed by a single letter for each of the major domains as follows: A, Archaea; B, Bacteria; C, Chloroplastida (algae and plants), D, Amoebozoa; E0, basal Eukaryota (that may constitute a clade of their own); F, Fungi; and H, Holozoa/Metazoa. Group G was not allocated, and may be retained should any new branches of lifeforms to be discovered (if there are any). We use the term Chloroplastida to denote all organisms containing chloroplast-related organelles in preference to the term Viridiplantae because the latter implies that algae are plants [45] .

Probes were thus named >X (code of ToL domain and number)_abridged species name_rRNA (16S or 18S)_probe number, for example, A_Hsalinarum_16S1.

BLAST analysis

Analysis is based on the basic local sequence alignment search tool, BLAST/BLAST+ [46, 47] , that is now widely accepted as the gold standard for detecting significant sequence similarities. In this work we carefully explored different settings for similarity detection, including wordsize (default = 28), match/mismatch parameters (default = 1,−2), and gap costs (linear), and in no case did this improve selectivity. We acknowledge that the default settings for evaluating sequence similarity have been carefully optimized by NCBI staff and their advisers, and we identified no obvious improvements. The basic settings (optimize for highly similar sequences) are that BLAST detects 64-mer sequences with 4 mismatches (94% identity), but also delects 28-mer sequences with no mismatches and 38-mer sequences with a single mismatch (which could be as low as 50% identity over the full length of the probe; values correct at time of writing, October 2021; these settings may evolve with updates of the NCBI website). However, we observed that coverage for the majority of matches detected (80–90%) were generally in the range 70–100% and, because a second screening step is applied, the default settings were retained. Cut-off scores are discussed in the sections below.

Removal of probes matching human sequences

Although 64-mer probes based on rRNA are less likely (because of their length) than full-length rRNA sequences to find matches in human databases by chance, rRNA is relatively conserved across species, and many of the probes found matches in human sequences. The 64-mers were therefore screened for probes that were were significantly similar to human sequences (default megablast task and -entrez_query "Homo sapiens {Organism}") by BLASTn v2.11.0+ with nt database [48] . About 300 probes found matches in human sequence databases (default parameters), and these probes were discarded. The final list of probes ($N = 1017$) without matches in the human database was generated in Fasta format and is given in **Table S3**.

Determining the number of matching reads in SRAs

A script was devised to count the number of matches to each probe according to our selection criteria (see also below). Readcounts after filtering can also be determined by accessing the hit table csv file and using the COUNTIF function of Microsoft Excel in the format =COUNTIF(A1:A5000,"probename"), a similar function is available in Apache Open Office (https://www.openoffice.org/welcome/credits.html) and other widely accessible spreadsheet programs.

Refiltering is necessary: sequence similarity does not follow rules of logic

A central issue encountered in this work is that sequence similarity, as assessed by BLAST, does not follow the rules of basic logic/mathematics. Mathematics dictates that, if A = B, and B = C, then A = C. Logic dictates that if A > B, and B > C, then A > C. The same rules do not apply to similarity between nucleotide (or protein) sequences. If sequence A is significantly similar (σ) to sequence B, i.e., A σ B, and B σ C, one may not conclude that A σ C. Conversely, if A is not similar (/σ) to B, and B σ C, one may not conclude that A /σ C. This is because, in three sequences A–C, sequence A (e.g., the probe) may not be similar to B (e.g., human sequences), but there may be an intermediate sequence C that is significantly similar to both A and B. Applied to human RNA-seq data, we observed that a large number of BLAST matches retrieved from human tissue are still of human origin. For this reason second-round filtering of all matches detected is necessary. In addition, representative human genome(s) and transcriptomes at NCBI do not encompass the full diversity of polymorphic variants that are present across the human population, and variant sequences may be encountered that achieve a threshold significant match with a filtered (i.e., no matches in human) probe despite being of human origin.

All matches retrieved from human RNA-seq libraries were therefore revalidated to exclude human sequences. This involves (i) retrieving the match sequences, (ii) searching again for homologies to human with downloaded nt database (ftp://ftp.ncbi.nlm.nih.gov/blast/db/; accessed 20 May 2021). Because the retrieved sequences are longer than 64 nt, the threshold score is adjusted according to the mean readsize in each dataset. The cutoffs were as follows. If the bitscore is >160 (MSBB) or >100 (Miami) or >126 (Rockefeller), the sequence was considered to be human in origin, and was discarded. For other datasets (liver, skin, and 17 brain samples a cutoff of >150 was applied; cutoffs for other databases

were >160 for the Mount Sinai Brain Bank and >150 for the Edinburgh Brain Bank, not presented). Three scripts were generated for this step, EDDIE_ToL.sh, Abundance_ToL.py and Abundance_count.py. Finally, duplicate reads detected with different probes (if any) were filtered to allocate reads to a single probe showing the highest sequence similarity, and the number of confirmed non-human reads were counted for each probe (Abundance_count.py).

Normalization: housekeeping transcripts

Because different tissue samples contain different amounts of RNA, sometimes partially degraded, and different RNA-seq protocols have different efficiencies of conversion to cDNA for sequencing, we normalized the number of readcounts obtained to the amount of biological material in the SRA. For this purpose we relied on three different housekeeping genes: phosphoglycerate kinase (*PGK1*), hydroxymethyl-CoA reductase (*HMGCR*), and neuron-specific enolase (*NSE*). An earlier study accurately assessed the number of transcripts per cell of *PGK1*-driven GFP is 46–64 transcripts per cell [49] (**Box 2**). We found no significant differences between *PGK1* and *NSE* levels in these datasets, whereas *HMGCR* expression was lower. Importantly, different target tissues will require their own panel of housekeeping gene probes. Of note, the expression levels of *PGK1* and *NSE1* in human brain appear to be constant as a function of age (https://hbatlas.org/pages/development).

As before, 64-mer probes were devised, and BLAST searching was employed to determine the number of host cells that each SRA corresponds to. To minimize statistical fluctuation the mean of *PGK* (two probes) and *NSE1* (two probes) was adopted for normalization. The number of host cells thus equals the mean readcounts of *PGK1* and *NSE* probes divided by 50, and in each case the raw readcount was normalized to the number of microbe transcripts per host cell by dividing the by the estimated number of host cells. Other tissue types will require dedicated housekeeping gene probes (discussed in **Box 2**).

Visualization and display of readcounts: heatmapping

An objective of this work was to generate a pictorial description of the distribution of (filtered) sequence matches in a given sample. Two methods were used, both based on the sequential order of probes in the taxonomic groups A (Archaea) to H (Holozoa). In the first, Morpheus software at the Broad Institute of MIT and Harvard (Cambridge, MA, USA; https://software.broadinstitute.org/morpheus/) was used for the display either with or without conversion to log2. All matches below a cutoff value (specified in the figures) are shown in blue, and all values between cutoff and a maximum value (dictated by the experiment in hand) are shown on a white to red scale where white = cutoff and red = maximum (or above). This presentation is outlined in **Figure 4**. In the second, the normalized data for each probe (reads per host cell) were summed for each organism, converted to log2 scale, and heatmaps plotted using the pheatmap package (v1.0.12) of R. Because of the wider range of abundances for retroelements and viruses, these can require different displays (these are specified in all cases).

rRNA depletion in RNA-seq libraries

Ribosomal RNA represents ~80% of the total RNA in each cell. Because most SRA studies in the data repositories have focused on endogenous human transcripts rather than on exogenous microbes, it is commonplace to use protocols to deplete rRNA. Precipitation with LiCl and/or selection of poly(A) RNA using immobilized oligo(dT) are widely used, but rarely remove more than 50–80% of rRNA. To address this we probed different datasets based on raw RNA-seq versus selection for polyA$^+$ RNA before sequencing. As shown in **Figure 5A**, there was no clear difference in the abundance of microbial rRNA transcripts between raw and polyA+ SRA datasets. This is likely to reflect (i) inefficiency of rRNA removal, (ii) different kits are likely to have different efficiencies of removal, and (iii) the extent of variation is comparable to that seen when the same tissue is worked up using alternative protocols in different laboratories.

An alternative approach is to specifically deplete host rRNA in a sample through hybridization to specific complementary oligonucleotides [50] . This method can achieve >90% removal of host (e.g., human) rRNA but does not remove other rRNAs such as those of exogenous microbes. No adjustment for rRNA removal was therefore performed.

Redundancy check

Because all our probes for cellular microbes are based on 16S/18S rRNA, we expected that some probes in the collection would be similar to others, perhaps because they identify highly conserved regions. A concern was therefore that the patterns we observe might to some extent reflect the degree of conservation across different species, as reflected by the number of probes in the collection with similar sequences, rather than the abundance of a target species. To address this, we catenated the entire probe collection into a single file and queried it by BLAST with the individual probes. Of the 1017 probes, 634 (62.3%) were unique within the collection, 100 (9.8%) found 2 matches (i.e., detected one similar sequence), and 122 (12.0%) gave 3–5 matches, and the remainder gave >5 matches (15.9%) (**Figure 6A**). We refer to the number of matches identified by each probe as 'redundancy'. Nevertheless, inspection revealed that the 'next best' matches were on average 48.8 nt in coverage, with 2.5 mismatches, falling to homology stretches as short as 31-mers with one mismatch, arguing that the extent of sequence overlap/duplication within the collection is relatively low.

However, to formally address whether the signals detected in human tissues might reflect the number of probe/probe similarities, we plotted the extent of 'redundancy' against total readcounts across the cohort. As shown in **Figure 6B**, there was a very slight correlation between redundancy and total readcount (trendline). However, the signals with the highest readcounts were at the lower end of the redundancy scale. **Figure 6C–E** shows the outcome of normalizing the readcount for each probe to the redundancy of the probe, indicating that normalization does not have a major effect on the overall pattern. In **Figure 6F** we examined the effect of sequentially deleting readcounts for probes with high reduncancy {20 or above (20+), 10 or above (10+), 5 or above (5+)} and then values for probes with reduncancies of 2–5, leaving only signals generated by unique probes within the collection. As shown, removal of redundant probes led to some degradation of the signal, but (i) many highly abundant signals remained despite restriction to unique probes, and (ii) the remaining signals continued to encompass the entire tree of life. However, coverage was slightly compromised, and probing the collection of known human commensals/pathogens (N = 104) with the 634 'unique' probes (not presented) failed to detect 6 species (5.8%; compared to only 2% missed with the complete probe list, next section).

One potential way to avoid this issue would be to generate a far larger collection of random probes, and then to discard probes with partial overlaps/duplications in the probe collection, and this could be done in the future. Overall, our observations argue that the overall pattern of microbe signals detected is not dependent on probe redundancy. Because

our primary intention was to retrieve the maximum number of non-human sequences, we elected not to remove partially overlapping sequences from the probelist because this could compromise retrieval. A drawback is that different probes might potentially retrieve the same sequences. However, this ambiguity is resolved by separate confirmation with 23S/28S sequences and contig building for identification (see below).

The probe collection is largely comprehensive

To address whether the filtered probe collection is sufficiently comprehensive, we used it to screen a manually assembled list of rRNAs for known human-associated species. A list of 104 organisms known to be present in human samples was assembled according to Archaea found in human gut [51] , bacteria (listed in Wikipedia: https://en.wikipedia.org/wiki/Pathogenic_bacteria), fungi (https://en.wikipedia.org/wiki/Pathogenic_fungus), helminths [52] , and a selection of rarer species assembled manually. Where two or more closely related species were listed, a single representative species was selected, noting that some species with similar names may in fact show significant divergence. Where the corresponding rRNA sequence was not available, the sequence from a closely related species was selected. In each case 16S/18S rRNA sequences were downloaded from nucleotide database of NCBI. This collection was named the PATHLIST (**Table S4**). Probing PATHLIST using the probe collection found matches in 98% of cases, and only missed two species (Discussion), indicating that the eToL v1.0 collection is largely sufficient for our purposes. Subsequent versions of eToL will include these 'missing species'.

Specificity and cross-matching

To determine the extent of cross-matching between probes designed from the different taxonomic groups (A–H), the 1000+ probes were used to screen the PATHLIST dataset of human-associated microbes. Because very few known human-associated microbes fall into the groups C (Chloroplastida), D (Amoebozoa, and E0 (basal Eukaryota), these were not included. As shown in **Figure 5B**, other than some overlaps between Fungi and Holozoa (as expected, given their relateness), there was little cross-matching between groups, attesting to the group-specificity of the probes in the collection.

Species identification, contig generation, and diversity

The eToL net collects a compendium of non-human sequences (confirmed by second-round filtering) in a target tissue. For identification, filtered matching reads for key signals of interest were downloaded, contigs were generated using one of several online tools (e.g., CAP3 assembly at the Rhone-Alpes Bioinformatics Pole PRABI-Doua; http://doua.prabi.fr/cgi-bin/run_cap3; also EG assembler https://www.genome.jp/tools-bin/eassembler4.cgi?status=seqclean&pmode=all [53] ; a summary of available tools is given at https://onlinetoolweb.com/contig-assembly-online-tool/), and their abundances determined by BLAST of each contig generated from human tissue against the collection of sequence matches, retaining only sequences with 100% (or near-100%) identity to the probe. Where necessary key contigs are used as probes to retrieve additional sequences from the same libraries. For sequence identification, BLAST at NCBI allows retrieval of the closest homologs. However, an average of 4000 matches were retrieved across 20 SRA datasets, but this can be as high as 16000. Computation time for contig building using EGassembler is estimated at 30 minutes for 4000 matches, rising to 5 h for 16000 matches [53] (computation time is approximately proportional to the square of the number of sequences). By contrast, contig assembly by phylogenetic group (A–H, where the mean number of matches for each bacteria and fungi was ~1000) is significantly faster and this number of sequences can be assembled into contigs in ~5 minutes [53] . Nevertheless, because the microbial sequences retrieved are not monophyletic, and show divergence within a single individual sample and between different samples, as illustrated in **Box 3**; high-stringency contig building risks excluding important contributors to the microbiome and, as in all such microbiome analyses, caution is necessary in interpretation.

Validation with 23S/28S and mitochondrial DNA

Because this approach employs a large number of probes (>1000), any observed high or low abundance in a particular tissue SRA could occur by chance (Bonferroni correction). To circumvent this issue the same protocols were used to devise fresh probes from 23S/28S rRNA for the same (or the closest available) key species signal of interest, and these were then used to reprobe the SRAs for matches. Again, all matches were filtered against human sequences, contigs assembled, followed by species identification.

An alternative approach that is valid for eukaryotic microbes including Fungi, but not for Bacteria or Archaea, is to employ mitochondrial DNA (mtDNA) sequences that have been 'stripped' (below) against human sequences.

Viruses: whole-genome 'stripping'

Because viruses have no ribosomes, we turned to whole-genome sequences. For viruses where detailed transcriptomes are available, both in latency and in productive infection, we recommend the use of 64-mer probes based on abundant transcripts. However, transcriptome data are only available for a minority of virus types, and for general applications we based our methodology on intact viral genomes. Nevertheless, these also contain internal homologies with human sequences (see below). For methodology development we based our analysis on the report in Neuron by Readhead et al. [25] where they used a *k*-mer method to screen human brain SRAs for 515 viruses. Although this method is prone to false positives (Introduction) it appears to be immune to false negatives. For methodology development we therefore selected the top 20 viruses in terms of readcounts, representing >99.9% of all matches found in human brain (B. Readhead, pers. comm.).

To identify matching sequences in the human genome/transcriptome, complete genomes for these 20 key viruses were used to search human sequences in the NCBI databases (BLAST/nucleotide collection (nr/nt), organism name = *Homo sapiens*). This revealed many matches, some relatively close (**Table S5**). To refine this approach we applied a 'stripping' method, as follows. All regions of sequence similarity between the viral genomes and human sequences identified by BLAST were deleted from the viral genome. In addition, low-complexity regions were removed. The first-round stripped genomes were then searched again against *H. sapiens*, revealing further matches. Four rounds of stripping were necessary to remove all homologies with the human genome/transcriptome. Because some viruses are present as integrated copies in around 1% of the human population (in particular with homology to HHV-6A, HHV-6B, and potentially HHV-7) [54] , these are present in the NCBI databases of 'human' sequences, and these were manually curated to remove key matching sequences (notably human telomeric repeats that are present in the viral genomes). The stripped genomes are given in **Table S6**.

Retroelements

The principal retroelements in the human genome are long and short interspersed nuclear elements (LINEs and SINEs, respectively). LINE activation has been reported in neurological disease [55] . LINEs belong to several subfamilies, and representative 64-mer probes were devised for each of these. SINE are generally relatively short highly conserved sequences related to human 7SL RNA, and probes were included to cover SINE elements. The human genome also includes several classes of human endogenous retroviruses (HERVs) that have been implicated in diverse disorders including neurodegenerative disease [56] , and we included probes to assess their abundance. Following the same nomenclature scheme as before, probes for retroelements including HERVs were allocated domain code R.

Recognizing and excluding contamination

Microbiome identification through rRNA sequencing and metagenomics is prone to various types of contamination that can obscure true signals [57-59] . Contamination may be classified into two principal catagories (Table 1A). Type 1 contamination includes contamination of the sample and reagents used to examine it, whereas type II concerns *in vivo* biocontamination, as discussed below.

(i) Reagent contamination: type 1A

Molecular biology reagents used to work up samples for sequence analysis are often contaminated with microorganisms. Salter *et al.* reported that sample dilution by a factor of 10^3 to 10^4 was necessary before contaminants represent 50% of the signal [57] . In terms of RNA-seq from PCR amplified material, this means that 0.1% of the signals could originate from contaminating material. In a series of 1000 signals, one or more may therefore arise from reagent contamination. Common bacterial contaminant species are given in Table 1, Table S1, and Table S2 of [57] , and Figure 5 of [60] . In addition, organisms commonly encountered in tap water may be consulted (Table S2). Although duplicate and/or blank samples worked up independently have been recommended [59] , this is not possible with pre-existing RNA-seq data. In addition, workup and sequencing of blank samples generally fails because instrument settings reject very low numbers of sequence reads (Azenta Life Sciences/GENEWIZ, authorized personal communication).

(ii) Sample contamination: type 1B

This is an important issue because, if samples are not dissected under fully sterile conditions, they may become contaminated by exogenous organisms, for example airborne spores and microbes from human skin. These latter are likely to contaminate some samples, and analysis is complicated by the fact that the skin microbiome is very diverse and differs according to body site, individual, and geographical location [61-63] . One possible way to tackle this issue is to exclude known human skin microbes. In the present work we subtracted brain datasets against a series of RNA-seq datasets for human skin (Results). Caution is necessary because, for example, a common skin fungus, *Malassezia*, has been directly implicated in human diseases such as psoriasis and has been reliably been detected in other human tissues [64] .

(iii) *In vivo* biocontamination: types 2A and 2B

This falls into two subtypes. Type 2A concerns contamination *in vivo* through life-long exposure to environmental agents. For example, we have observed signals in RNA-seq data corresponding to barley (*Hordeum vulgare*, not presented). Although airborne contamination of samples is not excluded, we suspect that *in vivo* contamination may take place. For example, inhaled ultrafine manganese oxide particles readily enter the central nervous system [65] . In mice exposed to microparticles (5 µm) and macroparticles (20 µm) in drinking water, both types of particle entered body tissues [66] , and environmental exposure to both small (<2.5 µm) and large (2.5–10 µm) particles has been associated with cognitive decline in human [67] . Barley pollen (25 µm) is in the same size range; over the course of a lifetime it is possible that these might also enter the circulation including brain vasculature, and from there into the brain itself. A similar route of infection could apply to microbial spores.

Type 2B contamination concerns contamination of the target tissue *in vivo* before sampling. For example, in studying a body tissue obtained postmortem, the cause of death should be taken into consideration. To illustrate, in many elderly patients (the principal source of postmortem tissues) death is often precipitated by severe infection, often pulmonary, and it

is possible if not likely that microbes enter the circulation and are thus present in diverse body tissues, independently of any disease process under investigation (Table 1A).

(iii) Strategies to exclude contamination

Key recommendations are summarized in Table 1 B. However, there is likely to be substantial overlap between microbes that are representative of the natural microbiome in human tissues and common contaminants of human origin such as from the skin.

(iv) Viruses and contamination

The problem of virus contamination is less severe because analysis is based on RNA-seq data, and contamination with mammalian cells expressing virus transcripts is thought to be unlikely. By contrast, contamination with viral genomes is possible, but these (particularly for DNA viruses) may be recognized because genomic reads are unlikely to correspond to the viral transcriptome. Viral reads were therefore mapped to the viral transcriptome to determine their authenticity (not presented).

Microbes and viruses: how many cells/genomes are being detected?

This analysis is based, for cellular organisms, on the number of copies of rRNA transcripts in each sample. The question therefore arises of how many cells are present, which in turn depends on the number of ribosomes (or viral transcripts) per cell. Because this parameter introduces a further complexity, further discussion is provided in **Boxes 1 and 2**.

Scripts

The scripts developed in this study are available at github (https://github.com/xinyuehu12/ToL).

Results

Tree of life

To generate the probe collection for cellular microbes, 120 key organisms were selected from Open Tree of Life Project and other sources to represent, as far as possible, the full spectrum of cellular organisms. The key organisms cover the domains of Archaea (A), Bacteria (B0–6), Chloroplastida (C1–4), Amoebozoa (D), basal Eukaryota (E0), fungi (F0–6) and Holozoa/Metazoa (H0–3) (Figure 1). To address the distribution of these organisms across the Tree of Life, the 16S/18S sequences corresponding to these 120 organisms were downloaded and used to build a phylogenetic tree (Figure 2). Because Archaea and Eukaryota are more related, the Bacteria group was used as outgroup [68]. This tree has relatively good statistical support because the bootstrap values for most of nodes are over 70. The nodes that have weak support (bootstrap value <70) are shown in gradient colors, ranging from red (7) to blue (69). The 16S rRNA sequences form two monophyletic groups, Archaea and Bacteria, whereas only one monophyletic group, Holozoa, was observed within the 18S rRNA sequences of eukaryotes. The species, especially from Chloroplastida and Amoebozoa, were not clustered well (Figure 2), and their nodes show weak support.

We then devised 64-mer probes for these species (the rationale is presented in the Methodology section), as shown in the pipeline (Figure 3), and filtered them against human sequences. To validate this approach, we checked whether the probe collection detects known human pathogens/commensals. With few exceptions, all organisms in the PATHLIST were detected in the probe collection, generally with multiple matches. The mean number of matches between the probe collection and each sequence in the PATHLIST was 25.8, and the median was 15. Only two species were not detected, *Leishmania donovani* and *Ascaris lumbricoides*. These data suggest that the probe collection covers around 98% of species for which sequences are available (future editions of eToL will be revised to include any missing lineages).

We also developed a uniform method of display based on heat-mapping, as illustrated in Figure 4.

A concern is that rRNA depletion (e.g., through polyA RNA selection) in RNA-seq libraries might compromise detection. However, this was not found to be a systematic problem, and robust microbe signals were detected in both polyA+ and unselected RNA-seq datasets (Figure 5). A further concern is whether the probe collection we have developed contains probes that significantly overlap with each other, and thus constitute partial duplicates. As described in Methodology, this was carefully addressed (Figure 6). The whole probe collection detected 98% of a test list of human pathogens and commensals, whereas selection of the unique probes detected only 92%. We will address this issue in further refinements, but the current analysis was continued with the complete list of probes.

Excluding contamination

Contamination is increasingly recognized to be a problem in microbiome analysis (reviewed by de Goffau *et al.* 2018). We therefore adapted our protocols to address this issue. A signature of contamination is

that different samples processed in parallel have consistent signals in all samples. Multi-positive signals were deleted from both the Miami and Rockefeller datasets. As shown in **Figure S1**, many of the signals in the Miami dataset may be contaminants, whereas the same issue was not encountered with the Rockefeller dataset (**Figure S1**).

Mutiple independent datasets: the brain has its own microbiome

The second strategy was to only consider signals that are present in independent datasets from the same tissue. We therefore screened 17 independent RNA-seq datasets from human brain. As shown in Figure 7A, the patterns were substantially conserved despite entirely independent workup.

To rigorously confirm that the microbial signals detected in brain do not arise through contamination, we sequentially subtracted the signals from the 17 independent brain samples against tapwater, and then against human skin. Specifically, if any signal appeared in any sample of tapwater or skin, all brain values were set to zero. Both procedures markedly depleted the brain microbial signal. However, multiple signals remained (Figure 7B). This argues that brain has its own microbiome that differs from that of skin, and is unlikely to represent either type 1 or type 2 contamination.

Tissue differences: liver versus brain

Previous studies highlighted likely overlaps between the brain and skin microbiomes (e.g., [21, 22]), which may be unsurprising because both tissues have an epithelial developmental origin. Subtraction of brain against skin signals could conceal species that are both (i) present in skin, and (ii) are opportunistic invaders of the CNS. We therefore focused on a different tissue, liver. As shown in **Figure S2**, the microbiome profiles for brain and liver display many similarities, but also some evident differences. The identities of key differentials were established by contig building and probing of NCBI datasets (**Figure S3**). In the samples analyzed, *Malassezia* spp. were found to be brain-specific, whereas *Staphylococcus aureus* appeared to be liver-specific (**Table S7**). The differential presence of key species was confirmed by second-round reprobing using 23S/28S-based probes (not presented).

Heterogeneity

We addressed whether individual probes are generally detecting specific organisms, or clusters of organisms. We observe that, even using these highly selective probes, we identify clusters of organisms that are not monophyletic. We illustrate this in **Box 3** through a case study. It is important to note that each probe detects a cluster of related species rather than a unique species (**Box 3**), and the exact identity of each sequence group retrieved from human tissue must be revalidated by 23S/28S analysis.

Viruses and retroelements

In terms of readcounts, viruses were less abundant than the other microbes (Figure 8A). Adenovirus C was the most abundant in these samples. Other viruses such as HSV-1, CMV, HHV-6A and TTV were also

present in some individuals (not presented), but their overall abundance was low (comprehensive analysis using this methodology is in preparation).

We also screened, using 64-mer probes, the abundance of select retroelements and endogenous retroviruses in tissue samples. As shown in Figure 8B, these are very well represented, although at present their potential contributions to health and disease remain unknown.

How many microbes are there in brain?

Readcounts, normalized on a per cell basis, do not immediately indicate whether microbes in brain are (relative to host cells) rare or abundant. We therefore considered further normalization factors including the number of rRNA copies that are present in typical microbial cells of different types (**Box 1**) to calculate how many cells/genomes are present in normal human brain; we elaborate on this point in the Discussion section and in **Box 2**.

Discussion

We report a new method to comprehensively analyze the entire microbiome of human tissue samples from transcriptomic (RNA-seq) data. The method comprises a 'net' of >1000 probes that covers all organisms from the known spectrum of lifeforms – the electronic Tree of Life (eToL). The method reported here is not intended to replace other established methodologies, but to provide an alternative that does not require any dedicated computer programs beyond those that are already widely available to the community (e.g., BLAST and BLAST+). Primarily based on 16S/18S sequences from cellular organisms, the approach has been extended to cover viruses and retroelements. Although computationally intensive, it is less demanding than metagenomics, and avoids the problems of non-selectivity encountered in earlier studies. Moreover, for the first time it addresses the entire ToL rather than selected subgroups of microbes.

Our study is based on a spectrum of cellular organisms (N = 120) that span the entire ToL (Figures 1 and 2), 20 viruses that are reported to be particularly abundant in our target tissue (brain), and 11 types of retroelements. For cellular species, the maximum likelihood phylogenetic tree shows the relationships between the rRNA sequences of the different cellular organisms, but species from domains such as Chloroplastida, Amoebozoa, and basal Eukaryota were not well clustered. Some nodes have weak statistical support although the overall statistical support is good. This may be because only partial rRNA sequences were available for some species. In addition, the branch of *Enterocytozoon bieneusi* (F1_Ebieneusi_18S) is long (not presented), which means it has high divergence compared to the other species, emphasizing the exceptional diversity of the fungi. However, the robustness of multiple alignment tools plays an important role in the accuracy of the phylogenetic tree. Although MAFFT has been reported to have good alignment accuracy, the MUSCLE tool was found to have better performance in reconstructing trees in our study. Other aligners may also be used to align such distant sequences [69]. To improve the accuracy of the alignments, rRNA secondary structures may also need to be considered [70].

Earlier studies of microbe analysis encountered problems including non-specificity and crossreactivity with human sequences. For example, the Jellyfish tool used by Readhead *et al.* counts *k*-mers with maximum length of 31 bases [25]. However, 31-mers may lack specificity. The *K* value (the frequency of probes aligning to target sequences by chance) can be used to calculate the probe length that is necessary for accurate target detection. By considering the background of incomplete but above-threshold matches (K_b value), the minimum probe length for finding unique matches in a typical mammalian genomic library was calculated to be 62 nt at 85% sequence identity and a K_b of 0.1 [41]. Therefore, our eToL approach was based on 64-mer (or longer) probes, that have been carefully filtered to remove any sequences matching human sequences. The 64-mers were semi-randomly selected because, if we used a sliding window method to produce overlapping 64-mers, the computation time required to process the many thousands of probes generated would become prohibitive.

The collection of probes (eToL) is designed as a 'net' to entrap all non-human microbial sequences. For this reason the identity of each probe does not indicate the exact species present, and retrieval of sequences from human tissue is necessary for species identification. However, the matches to our probes discovered in RNA-seq data unambiguously confirmed the presence of microorganisms because all matches were double-checked to exclude human sequences. Therefore, the false positive rate of detecting human sequences was reduced to near-zero.

The 64-mer probe collection for cellular organisms is believed to be largely comprehensive because 98% of known human pathogens and commensals were detected by the probe collection. In future editions of the ToL probe list will include probes for missing species such as *Leishmania donovani* and *Ascaris lumbricoides*, as well for species that are less well represented in the probe list.

We took strict precautions to recognize and exclude potential contaminants of either type 1 or type 2 (Table 1). This revealed that some RNA-seq datasets are potentially diluted by contaminant species inadvertantly introduced during sample preparation and workup. Subtraction of matches against likely sources of contamination including human skin and tap water reduced but did not eliminate the microbiome signals, arguing that key species are indeed present in human brain (in preparation). To confirm the identity of these brain-resident species, the matches were retrieved from brain, the exact species identified, and then further validated by 23S/28S rRNA analysis, and for eukaryotic species, analysis of mitochondrial DNA (in preparation).

For viruses, we report that the false positives in the study of Readhead *et al.* may be explained, in part, by homologies between viral and human sequences. For example, HHV-6A, 6B, and 7 (that were asserted to be increased in AD) have pronounced matches with human telomeric DNA repeats, HHV-3 and HHV-8 contain sequences similar to human thymidylate synthase (*TYMS*), the HHV-4 genome has matches to human interleukin 10 receptor variant (*IL10RV*), and the genome of variola virus (the agent of smallpox, also detected in human brain by Readhead *et al.* contains homologies to human ribonucleotide reductase subunits (*RRM1* and *RRM2B*) (**Table S5**). The major virus type in human brain was identified to be adenovirus type C (transcript mapping will be reported elsewhere).

Overall, this work reveals that a remarkable diversity of microbes are present in brain (and liver) samples. All major taxonomic groups are represented. In addition to bacteria and fungi, as previously reported, we report that microbes ranging from Archaea to Amoebozoa, Chloroplastida, basal Eukaryota, and Holozoa/Metazoa are present in human brain. Few viruses were encountered, the majority being of the adenovirus C class.

Cautious application of the eToL method may go some way towards answering the question of whether there is indeed a brain microbiome [15]. Because this is a Methodology paper, we do not report systematically on the identities of the target organisms or experimental confirmation (in preparation). However, in terms of microbes/genomes per cell, we estimate that Archaea are present at 10^{-5} microbes per host cell, Bacteria (0.14), Amoebozoa (0.01), basal Eukaryota (0.01), and Fungi (0.05), of which Bacteria and Fungi constitute >50% of the total microbial burden (Figure 7).

The approach presented here, unlike other strategies, has the advantage that all cellular organisms across the ToL can be addressed in a single screen. For viruses, the genome-stripping method offers a route to ensure that only viral sequences are detected. In addition, the method has advantages over other strategies in terms of specificity, and employs widely available analytical tools (BLAST and BLAST+). Simple modifications to the protocol make it applicable to other species such as non-human primates and rodents where whole-genome sequence data are available for stripping. The methodologies presented here may find broad application in the analysis of microbes and viruses in widely available RNA-seq data for human tissues, and thereby enhance our understanding of the role of the microbome in human physiology in health and disease. In addition, given that RNA-seq data can now be obtained for under $400, the method may find application in the bioanalysis of human samples such as skin, oral, nasal, and pulmonary samples, as well as of blood, urine, and cerebrospinal fluid.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

All relevant data and materials, including data repositories and sequence read archive identification numbers, are provided in the manuscript and/or in the supplementary material online.

Competing interests

The authors declare no competing interests

Funding

This project was funded in part by the Benter Foundation.

Authors' contributions

R.L. conceptualized the project. X.H. devised the scripts and performed data analysis in conjunction with R.L. The main manuscript was written by R.L. with assistance from X.H. All authors have contributed to the final manuscript. Overall project supervision was jointly by J.H. and R.L.

Acknowledgements

We thank staff at NCBI, particularly A. Wayne, for ongoing support and advice, and B. Readhead (Arizona) for communicating unpublished data. Sequence searching was performed either online at NCBI (we thank NCBI for making this service available) or locally at the Edinburgh Compute and Data Facility (ECDF) Linux Compute Cluster (http://www.ecdf.ed.ac.uk/) ('EDDIE') at the University of Edinburgh. We thank Amanda Warr (Roslin Institute) and Shelley Allen (Bristol) for critical comments on the manuscript. In addition, we wish to acknowledge the major contribution that Alison Daniels has made to this work. Part of this work was submitted in partial fulfillment of the requirements for the award of a postgraduate degree in Bioinformatics at the University of Edinburgh (X.H.).

Boxes

Box 1. Transcripts per microbial cell/infected host cell

Cellular microbes: ribosomal copy number depends on cell type and growth rate

Readcounts of rRNA provide an indicator of how many microbes are present, but there is no direct one-to-one relationship between rRNA readcounts and microbial cell number. The absolute abundance of ribosomes in each cell (ribosomes per cell, RBPC) depends on both growth rate and cell type/size. For bacteria such as *Escherichia coli*, RBPC values can be up to 70 000 during periods of rapid growth, but as low as a few thousand in poor growth conditions (http://book.bionumbers.org/how-many-ribosomes-are-in-a-cell/). In another bacterium, *Sphingomonas*, RBPC can be as low as 200 [71] under poor growth conditions. We assume that bacteria in human solid tissues grow very slowly, and a compromise estimate of 2000 RBPC has been adopted. The same value has been assumed for Archaea because they resemble bacteria in terms of size.

Eukaryotic cells are generally much larger, contain more ribosomes, and for yeast under fast-growing conditions 200 000 RBPC have been reported [72, 73] . However, as in *E. coli*, ribosome content depends on growth rate [74] . We therefore applied a similar reduction for what we expect to be very slow-growing cells, giving an estimate of 10000 RBPC (fivefold greater than in bacteria), noting that the exact RBPC values will depend on the species. Although this value could apply to other eukaryotes, it may be an underestimate because, for example, some large mammalian cells can have over a million ribosomes [75] ; however, this extreme high value is unlikely to be representative of microbes more generally.

Virus-infected cells

The problem in estimating how many genomes are present per host cell becomes more acute for viruses because we must distinguish between latent/quiescent infection, where viral transcript counts are expected to be similar to those of endogenous housekeeping genes, and virulent replication, where there can be thousands of genomes/transcripts per cell. For this reason detection of viral sequences in RNA-seq data is likely to detect highly replicative viruses, but could potentially miss low-level latency-related transcripts from quiescent viruses. This point is discussed in more detail in **Box 2**.

Box 2. Normalization to endogenous transcripts and detection limits

A key question concerns the extent of coverage of each sequence read archive (SRA). To address this we used *PGK1* probes to normalize RNA-seq data. *PGK1* stands out because Kempe *et al.* [49] carefully measured the level of expression of *PGK1*-driven GFP and accurately measured 46–64 transcripts per cell. In support, it is estimated that ~12 000 genes are expressed per typical cell, and between 360 000 and 10^6 mRNAs are present: the average number of transcripts, per gene, is thus in the range 30–83, strictly comparable to the estimate of ca 50 *PGK1* transcripts per cell determined by Kempe *et al.*

Screening of RNA-seq libraries revealed that *PGK1* match numbers are in the general range of 10 to over 250 per library, with a median of 152, arguing that each sequence library roughly equates to the sequence content of ~3 cells, although more recent RNA-seq datasets appear to be a little larger (10 cells). This conclusion is substantiated by

considerations of SRA file sizes (3–10 GB). Very approximately, an uncompressed file containing 1 MB of information contains 1 megabase of nucleotides, and a file containing 1 GB of information contains ~1 gigabase of nucleotides. The entire transcriptome of a cell, at 600 000 mRNAs per cell, and average mRNA length = 2.2 kb, equates to 1 320 000 000 nucleotides (1.3 GB of data). However, rRNA removal is usually far from complete and, in addition, 1/3 of the information in a typical SRA file is not sequence data (each entry contains details of the specific read in question). To cover the (non-rRNA) transcriptome of a single cell therefore requires a minimum of 2.5 GB of information in FASTA format. Although they may be partly compressed in some formats, consideration of SRA file size (3–10 GB) is consistent with the interpretation that a single SRA equates to the transcriptome of a small number of cells. A legitimate concern is therefore that poorly abundant (but biologically relevant) microbes/viruses may not be detected (see below).

Another consideration is that rRNAs tend to contain regions of secondary structure that reverse transcriptase may find difficult to copy into cDNA. However, this is unpredictable, and we have made no explicit allowance for this factor.

Average read length is a further consideration. Using 64-mer probes, if the mean read length is 150 nt, then (if mRNAs are randomly sampled) some 40% of reads will not sufficiently overlap with the probe, but this falls to ~30% for 250 nt reads. However, this factor is identical for housekeeping genes and for microbial transcripts, and does not affect normalization.

Although our normalization is based on *PGK1* transcripts at 50 per host cell, even using two independent probes for *PGK1* we observed some statistical fluctuation. To dampen this effect we considered two other genes. The first, hydroxymethyl glutaryl-CoA reductase (*HMGCR*) is another housekeeping gene, but transcript levels were on average fivefold lower, making it unsuitable. Instead we turned to a neuron-specific housekeeping gene, neuron-specific enolase (*NSE*, also known as *ENO2*). It can legitimately be argued that, because *NSE* is neuron-specific, it is not entirely representative of our target tissue (brain) where neurons only constitute a little over one half of all host cells. However, we saw excellent agreement between *NSE* and *PGK1* transcripts, and for simplicity our results are presented following normalization to the mean number of *PGK1* and *NSE* transcripts per cell (or per 100 host cells), assuming that 1 host cell = (mean {*PGK1* and *NSE1*})/50. Although this probe combination is probably useful for normalizing RNA-seq data from

brain, in other tissues it will be necessary to carefully choose the most appropriate panel of housekeeping genes for designing normalization probes.

Calculated detection sensitivity

If current sequence libraries represent circa 10 cells (above), and cellular microbes contain 2000 ribosomes (**Box 1**), then detection of a single rRNA transcript (detection limit) means that the corresponding microbe is present at a level of of 0.0005 microbial cells per 10 host cells, which is very sensitive. For lytic viruses, where 1000+ transcripts per infected cell may be present, a similar level of detection (0.0005 infected cells per 10 host cells) is likely to apply. The exception is for latent viruses. For example, herpes viruses express latency transcripts of various types (LAT in the case of HSV-1), and the abundance of such transcripts is assumed to be low, possibly comparable to the level of housekeeping gene expression (50 transcripts per cell). If the RNA-seq dataset represents the transcriptome of at most 10 cells, then the detection of a single LAT transcript would represent 0.02 infected cells per 10 host cells, and this is likely to constitute the lower level of detection. However, multiple viruses are present in human tissues at low levels, where they persist throughout a lifetime [14] , and it remains to be determined whether latent/quiescent infection such viruses compromises cell function; the lower sensitivity may therefore not be a drawback in the analysis of human physiology in health and disease.

Box 3. Case study: signals detected are not monophyletic

To illustrate the complexity of computer searching, we report the case of searching the first liver SRA with a 64-mer probe, B3_AcidobacteriaKBS96_16S_8. This found 44 matches, of which eight were 100% identical, whereas the major class (36) contained one or more mismatches. Contig building generated 6 contigs, of which two were highly abundant, the others less so. Sequence comparisons and tree building (ClustalOmega) revealed that they were divergent in sequence (**Figure I**). To extend the two most abundant contigs, these were used to reprobe the SRA, and matches were used to generate contigs again. This gave two contigs of 496 and 336 nt. The former was 97% identical to an uncultured Betaproteobacteria clone, whereas the latter was 100% identical to a Gammaproteobacteria species, *Acinetobacter*.

Using the major contigs to search a second liver SRA revealed that the major contig from liver 1 found closely related, but not 100% identical, sequences in liver 2, whereas the second most abundant contig in liver 1 found no close relatives in liver 2. This illustrates two points: (i) the exact species present in a tissue sample are not monophyletic, but represent a spectrum of related microbes; and (ii) what is true of one tissue sample may not be true of a near-identical sample from a different individual. The best approach may be to address multiple samples from different individuals, and identify the commonalities.

Figure I. Matches obtained with a single probe highlight the most abundant species (*contigs 3 and 5) in the first liver sample, contig building (Consensus 1 and Consensus 2, respectively) and divergence in a second sample.

Tables

Table 1. Types of contamination and strategies to exclude them

A Types of contamination

Class	Type	Comments
1A	Molecular biology reagents	If contaminated, the same signal will be present in all samples
1B	Sample contamination during dissection	Expect environmental contaminants such as spores, pollens, and skin microbes (caution, skin microbes have been implicated in several human diseases)
2A	Lifelong *in vivo* biocontamination from blood and the environment	Environmental contaminants such as microparticles of the same size as spores and pollens have been demonstrated to enter tissues; microbes of a similar size rapidly enter human tissues
2B	*In vivo* biocontamination: perimortem	When analyzing a tissue in relation to human disease, contamination may arise from *in vivo* dissemination of microbes (e.g., respiratory disease) unrelated to the primary disorder; invasion of diseased tissue may be a consequence rather than a cause of tissue degeneration

B. Strategies to exclude contaminants

Method	Strategy	Comments
Negative controls	Exclude all signals present in blank workups	Negative controls alone are not sufficient to detect all contaminating species. In addition, RNA-seq data rarely have blank controls because these are rejected as errors by the sequencing instrument
Common contaminants	Consider excluding common contaminant species such as those listed in Salter *et al.* [57] and Sanabria *et al.* [60]	Caution is urged because common contaminant microbes may themselves be the cause of disease

Within-batch consistency	Exclude signals that are present at similar levels in all samples	Caution is urged because, if applied to gut or lung, this would exclude many of the major species that are known to be present
Between-batch consistency	Only include signals that are present (and at different levels) in independent datasets from the same tissue	
Differential signals	Only include differential signals between, for example, disease samples versus controls	Consistent differential signals point umbiguously to species that are not contaminants
Microheterogeneity	High-resolution strain/substrain mapping. Contaminants introduced during sample work-up are likely to be of the same specific genotypes in different samples, whereas true signals are most likely heterogeneous in their exact sequences	Download sequences from different samples of the target tissue and prepare phylogenetic trees

References

1. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI The human microbiome project. Nature 2007;449:804–810.
2. Integrative HMP (iHMP) Research Network Consortium The integrative human microbiome project. Nature 2019;569:641–648.
3. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N Shotgun metagenomics, from sampling to analysis. Nat Biotechnol 2017;35:833–844.
4. Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, Gonzalez A, Kosciolek T, McCall LI, McDonald D, Melnik AV, Morton JT, Navas J, Quinn RA, Sanders JG, Swafford AD, Thompson LR, Tripathi A, Xu ZZ, Zaneveld JR, Zhu Q, Caporaso JG, Dorrestein PC Best practices for analysing microbiomes. Nat Rev Microbiol 2018;16:410–422.
5. Osman MA, Neoh HM, Ab Mutalib NS, Chin SF, Jamal R 16S rRNA gene sequencing for deciphering the colorectal cancer gut microbiome: current protocols and workflows. Front Microbiol 2018;9:767.
6. Breitwieser FP, Lu J, Salzberg SL A review of methods and databases for metagenomic classification and assembly. Brief Bioinform 2019;20:1125–1136.

7. Fricker AM, Podlesny D, Fricke WF What is new and relevant for sequencing-based microbiome research? A mini-review. J Adv Res 2019;19:105–112.
8. Bharti R, Grimm DG Current challenges and best-practice protocols for microbiome analysis. Brief Bioinform 2021;22:178–193.
9. Gao B, Chi L, Zhu Y, Shi X, Tu P, Li B, Yin J, Gao N, Shen W, Schnabl B An introduction to next generation sequencing bioinformatic analysis in gut microbiome studies. Biomolecules 2021;11:530.
10. Yen S, Johnson JS Metagenomics: a path to understanding the gut microbiome. Mamm Genome 2021;32:282–296.
11. Moreno-Indias I, Lahti L, Nedyalkova M, Elbere I, Roshchupkin G, Adilovic M, Aydemir O, Bakir-Gungor B, Santa Pau EC, D'Elia D, Desai MS, Falquet L, Gundogdu A, Hron K, Klammsteiner T, Lopes MB, Marcos-Zambrano LJ, Marques C, Mason M, May P, Pašiæ L, Pio G, Pongor S, Promponas VJ, Przymus P, Saez-Rodriguez J, Sampri A, Shigdel R, Stres B, Suharoschi R, Truu J, Truicã C0, Vilne B, Vlachakis D, Yilmaz E, Zeller G, Zomer AL, Gómez-Cabrero D, Claesson MJ Statistical and machine learning techniques in human microbiome studies: contemporary challenges and solutions. Front Microbiol 2021;12:635781.
12. Wood DE, Salzberg SL Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol 2014;15:R46.
13. Ounit R, Lonardi S Higher classification sensitivity of short metagenomic reads with CLARK-S. Bioinformatics 2016;32:3823–3825.
14. Lathe R, St Clair D From conifers to cognition: microbes, brain and behavior. Genes Brain Behav 2020;19:e12680.
15. Link CD Is there a brain microbiome? Neurosci Insights 2021;16:26331055211018709.
16. Coelho C, Camacho E, Salas A, Alanio A, Casadevall A Intranasal Inoculation of Cryptococcus neoformans in mice produces nasal infection with rapid brain dissemination. mSphere 2019;4:e00483-19.
17. Pisa D, Alonso R, Fernandez-Fernandez AM, Rabano A, Carrasco L Polymicrobial infections in brain tissue from Alzheimer's disease patients. Sci Rep 2017;7:5559.
18. Itzhaki RF, Lathe R, Balin BJ, Ball MJ, Bearer EL, Braak H, Bullido MJ, Carter C, Clerici M, Cosby SL, Del TK, Field H, Fulop T, Grassi C, Griffin WS, Haas J, Hudson AP, Kamer AR, Kell DB, Licastro F, Letenneur L, Lovheim H, Mancuso R, Miklossy J, Otth C, Palamara AT, Perry G, Preston C, Pretorius E, Strandberg T, Tabet N, Taylor-Robinson SD, Whittum-Hudson JA Microbes and Alzheimer's disease. J Alzheimers Dis 2016;51:979–984.
19. Miklossy J Alzheimer's disease - a neurospirochetosis. Analysis of the evidence following Koch's and Hill's criteria. J Neuroinflammation 2011;8:90.
20. Balin BJ, Hammond CJ, Little CS, Hingley ST, Al-Atrache Z, Appelt DM, Whittum-Hudson JA, Hudson AP Chlamydia pneumoniae: an etiologic agent for late-onset dementia. Front Aging Neurosci 2018;10:302.

21. Alonso R, Pisa D, Fernández-Fernández AM, Carrasco L Infection of fungi and bacteria in brain tissue from elderly persons and patients with Alzheimer's disease. Front Aging Neurosci 2018;10:159.
22. Emery DC, Shoemark DK, Batstone TE, Waterfall CM, Coghill JA, Cerajewska TL, Davies M, West NX, Allen SJ 16S rRNA next generation sequencing analysis shows bacteria in Alzheimer's post-mortem brain. Front Aging Neurosci 2017;9:195.
23. Dominy SS, Lynch C, Ermini F, Benedyk M, Marczyk A, et al. Porphyromonas gingivalis in Alzheimer's disease brains: evidence for disease causation and treatment with small-molecule inhibitors. Sci Adv 2019;5:eaau3333.
24. Jamieson GA, Maitland NJ, Craske J, Wilcock GK, Itzhaki RF Detection of herpes simplex virus type 1 DNA sequences in normal and Alzheimer's disease brain using polymerase chain reaction. Biochem Soc Trans 1991;19:122S.
25. Readhead B, Haure-Mirande JV, Funk CC, Richards MA, Shannon PSHV, Sano M, Liang W, Beckmann ND, Price ND, Reiman EM, Schadt EE, Ehrlich ME, Gandy S, Dudley JT Multiscale analysis of independent Alzheimer's cohorts finds disruption of molecular, genetic, and clinical networks by human herpesvirus. Neuron 2018;99:64–82.
26. Marçais G, Kingsford C A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 2011;27:764–770.
27. Chorlton SD Reanalysis of Alzheimer's brain sequencing data reveals absence of purported HHV6A and HHV7. J Bioinform Comput Biol 2020;18:2050012.
28. Breitwieser FP, Baker DN, Salzberg SL KrakenUniq: confident and fast metagenomics classification using unique k-mer counts. Genome Biol 2018;19:198.
29. Allnutt MA, Johnson K, Bennett DA, Connor SM, Troncoso JC, Pletnikova O, Albert MS, Resnick SM, Scholz SW, De Jager PL, Jacobson S Human herpesvirus 6 detection in Alzheimer's disease cases and controls across multiple cohorts. Neuron 2020;105:1027–1035.
30. Kostic AD, Ojesina AI, Pedamallu CS, Jung J, Verhaak RG, Getz G, Meyerson M PathSeq: software to identify or discover microbes by deep sequencing of human tissue. Nat Biotechnol 2011;29:393–396.
31. Woese CR, Kandler O, Wheelis ML Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci U S A 1990;87:4576–4579.
32. Redelings BD, Holder MT A supertree pipeline for summarizing phylogenetic and taxonomic information for millions of species. PeerJ 2017;5:e3058.
33. Rees JA, Cranston K Automated assembly of a reference taxonomy for phylogenetic data synthesis. Biodivers Data J 2017:e12581.
34. McTavish EJ, Hinchliff CE, Allman JF, Brown JW, Cranston KA, Holder MT, Rees JA, Smith SA Phylesystem: a git-based data store for community-curated phylogenetic estimates. Bioinformatics 2015;31:2794–2800.
35. Schulz F, Eloe-Fadrosh EA, Bowers RM, Jarett J, Nielsen T, Ivanova NN, Kyrpides NC, Woyke T Towards a balanced view of the bacterial tree of life. Microbiome 2017;5:140.

36. Edgar RC MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004;32:1792–1797.
37. Nguyen LT, Schmidt HA, von HA, Minh BQ IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 2015;32:268–274.
38. Kalyaanamoorthy S, Minh BQ, Wong TKF, von HA, Jermiin LS ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods 2017;14:587–589.
39. Hoang DT, Chernomor O, von HA, Minh BQ, Vinh LS UFBoot2: improving the ultrafast bootstrap approximation. Mol Biol Evol 2018;35:518–522.
40. Letunic I, Bork P Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res 2021;49:W293-W296.
41. Lathe R Synthetic oligonucleotide probes deduced from amino acid sequence data. Theoretical and practical considerations. J Molec Biol 1985;183:1–12.
42. Neefs JM, Van de Peer Y, De RP, Goris A, De WR Compilation of small ribosomal subunit RNA sequences. Nucleic Acids Res 1991;19 Suppl:1987-2015.
43. Huse SM, Dethlefsen L, Huber JA, Mark WD, Relman DA, Sogin ML Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. PLoS Genet 2008;4:e1000255.
44. Van Rossum G, Fred LJ: *Python 3 Reference Manual*. Scotts Valley CA: CreateSpace; 2009.
45. Adl SM, Simpson AG, Farmer MA, Andersen RA, Anderson OR, Barta JR, Bowser SS, Brugerolle G, Fensome RA, Fredericq S, James TY, Karpov S, Kugrens P, Krug J, Lane CE, Lewis LA, Lodge J, Lynn DH, Mann DG, McCourt RM, Mendoza L, Moestrup O, Mozley-Standridge SE, Nerad TA, Shearer CA, Smirnov AV, Spiegel FW, Taylor MF The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. J Eukaryot Microbiol 2005;52:399–451.
46. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ Basic local alignment search tool. J Mol Biol 1990;215:403–410.
47. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL BLAST+: architecture and applications. BMC Bioinformatics 2009;10:421.
48. Zhang Z, Schwartz S, Wagner L, Miller W A greedy algorithm for aligning DNA sequences. J Comput Biol 2000;7:203–214.
49. Kempe H, Schwabe A, Crémazy F, Verschure PJ, Bruggeman FJ The volumes and transcript counts of single cells reveal concentration homeostasis and capture biological noise. Mol Biol Cell 2015;26:797–804.
50. Kraus AJ, Brink BG, Siegel TN Efficient and specific oligo-based depletion of rRNA. Sci Rep 2019;9:12281.
51. Gaci N, Borrel G, Tottey W, O'Toole PW, Brugère J-F Archaea and the human gut: new beginning of an old story. World J Gastroenterol 2014;20:16062–16078.
52. Hewitson JP, Maizels RM Vaccination against helminth parasite infections. Expert Rev Vaccines 2014;13:473–487.

53. Masoudi-Nejad A, Tonomura K, Kawashima S, Moriya Y, Suzuki M, Itoh M, Kanehisa M, Endo T, Goto S EGassembler: online bioinformatics service for large-scale processing, clustering and assembling ESTs and genomic DNA fragments. Nucleic Acids Res 2006;34:W459-W462.
54. Gravel A, Hall CB, Flamand L Sequence analysis of transplacentally acquired human herpesvirus 6 DNA is consistent with transmission of a chromosomally integrated reactivated virus. J Infect Dis 2013;207:1585–1589.
55. Terry DM, Devine SE Aberrantly high levels of somatic LINE-1 expression and retrotransposition in human neurological disorders. Front Genet 2019;10:1244.
56. Küry P, Nath A, Créange A, Dolei A, Marche P, Gold J, Giovannoni G, Hartung HP, Perron H Human endogenous retroviruses in neurological diseases. Trends Mol Med 2018;24:379–394.
57. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, Turner P, Parkhill J, Loman NJ, Walker AW Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol 2014;12:87.
58. Merchant S, Wood DE, Salzberg SL Unexpected cross-species contamination in genome sequencing projects. PeerJ 2014;2:e675.
59. de Goffau MC, Lager S, Salter SJ, Wagner J, Kronbichler A, Charnock-Jones DS, Peacock SJ, Smith GCS, Parkhill J Recognizing the reagent microbiome. Nat Microbiol 2018;3:851–853.
60. Sanabria A, Hjerde E, Johannessen M, Sollid JE, Simonsen GS, Hanssen AM Shotgun-metagenomics on positive blood culture bottles inoculated with prosthetic joint tissue: a proof of concept study. Front Microbiol 2020;11:1687.
61. Oh J, Byrd AL, Deming C, Conlan S, Kong HH, Segre JA Biogeography and individuality shape function in the human skin metagenome. Nature 2014;514:59–64.
62. Oh J, Byrd AL, Park M, Kong HH, Segre JA Temporal stability of the human skin microbiome. Cell 2016;165:854–866.
63. Cho HW, Eom YB Forensic analysis of human microbiome in skin and body fluids based on geographic location. Front Cell Infect Microbiol 2021;11:695191.
64. Abdillah A, Ranque S Chronic diseases associated with Malassezia yeast. J Fungi (Basel) 2021;7:855.
65. Elder A, Gelein R, Silva V, Feikert T, Opanashuk L, Carter J, Potter R, Maynard A, Ito Y, Finkelstein J, Oberdörster.G. Translocation of inhaled ultrafine manganese oxide particles to the central nervous system. Environ Health Perspect 2006;114:1172–1178.
66. Deng Y, Zhang Y, Lemos B, Ren H Tissue accumulation of microplastics in mice and biomarker responses suggest widespread health risks of exposure. Sci Rep 2017;7:46687.
67. Weuve J, Puett RC, Schwartz J, Yanosky JD, Laden F, Grodstein F Exposure to particulate air pollution and cognitive decline in older women. Arch Intern Med 2012;172:219–227.
68. Eme L, Spang A, Lombard J, Stairs CW, Ettema TJG Archaea and the origin of eukaryotes. Nat Rev Microbiol 2017;15:711–723.

69. Liu K, Linder CR, Warnow T Multiple sequence alignment: a major challenge to large-scale phylogenetics. PLoS Curr 2010;2:RRN1198.
70. Letsch HO, Kück P, Stocsits RR, Misof B The impact of rRNA secondary structure consideration in alignment and tree reconstruction: simulated data and a case study on the phylogeny of hexapods. Mol Biol Evol 2010;27:2507–2521.
71. Fegatella F, Lim J, Kjelleberg S, Cavicchioli R Implications of rRNA operon copy number and ribosome content in the marine oligotrophic ultramicrobacterium Sphingomonas sp. strain RB2256. Appl Environ Microbiol 1998;64:4433–4438.
72. Warner JR The economics of ribosome biosynthesis in yeast. Trends Biochem Sci 1999;24:437–440.
73. Yamaguchi M, Namiki Y, Okada H, Mori Y, Furukawa H, Wang J, Ohkusu M, Kawamoto S Structome of Saccharomyces cerevisiae determined by freeze-substitution and serial ultrathin-sectioning electron microscopy. J Electron Microsc (Tokyo) 2011;60:321–335.
74. Nomura M Regulation of ribosome biosynthesis in Escherichia coli and Saccharomyces cerevisiae: diversity and common principles. J Bacteriol 1999;181:6857–6864.
75. Finka A, Sood V, Quadroni M, de Los Rios P, Goloubinoff P Quantitative proteomics of heat-treated human cells show an across-the-board mild depletion of housekeeping proteins to massively accumulate few HSPs. Cell Stress Chaperones 2015;20:605–620.

Figures

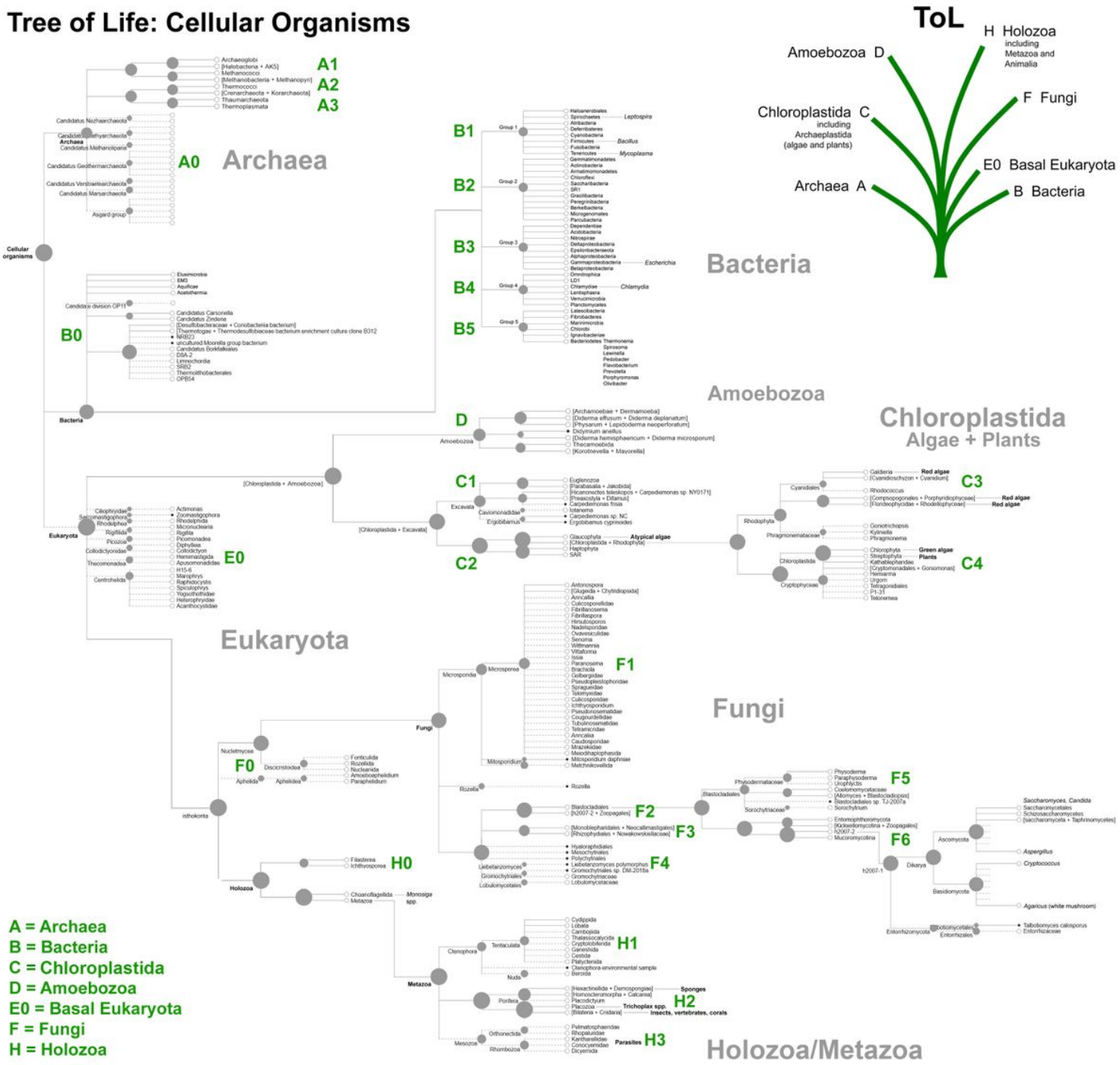

Figure 1

The phylogenetic tree for cellular microbes. The phylogenetic tree was obtained from Open Tree of Life project (modified from https://tree.opentreeoflife.org/) extended to include a recent bacterial consensus tree (main text for details).

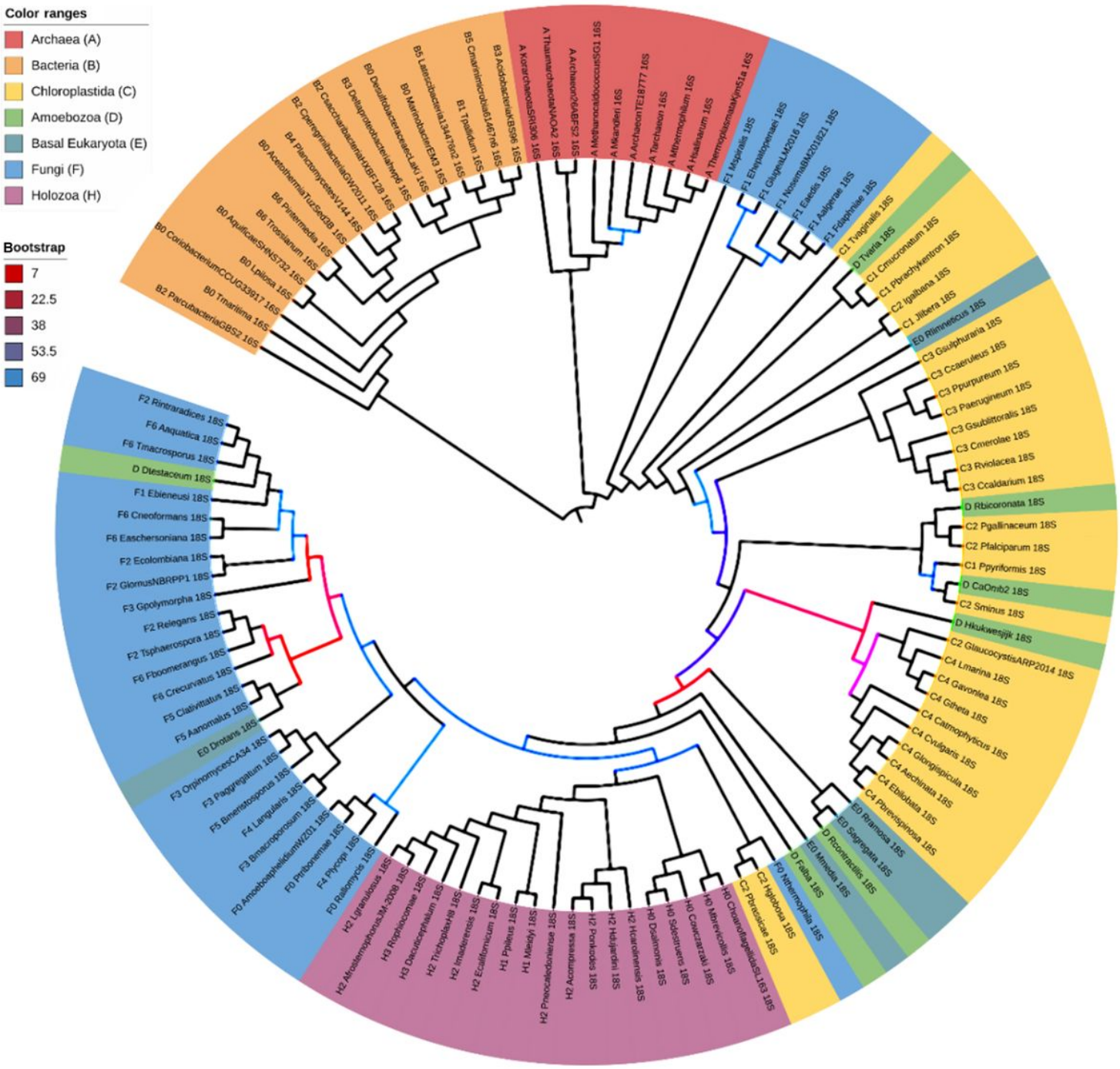

Figure 2

Phylogenetic tree built with rRNA sequences for the key 120 organisms. 16S rRNA sequences for Archaea (red) and Bacteria (orange), and 18S rRNA sequences for Chloroplastida (yellow), Amoebozoa (green), basal Eukaryota (grey blue), Fungi (blue) and Holozoa (purple). Bootstrap values of most of the nodes are over 70 (black), indicating strong phylogenetic support. Nodes with bootstrap values lower than 70 have weaker phylogenetic support and are shown in a color gradient from red (low value) to blue (high value. Note that the placements of Fungi, Chloroplastida, and Amoebozoa are intermingled, whereas the other groups are tightly clustered.

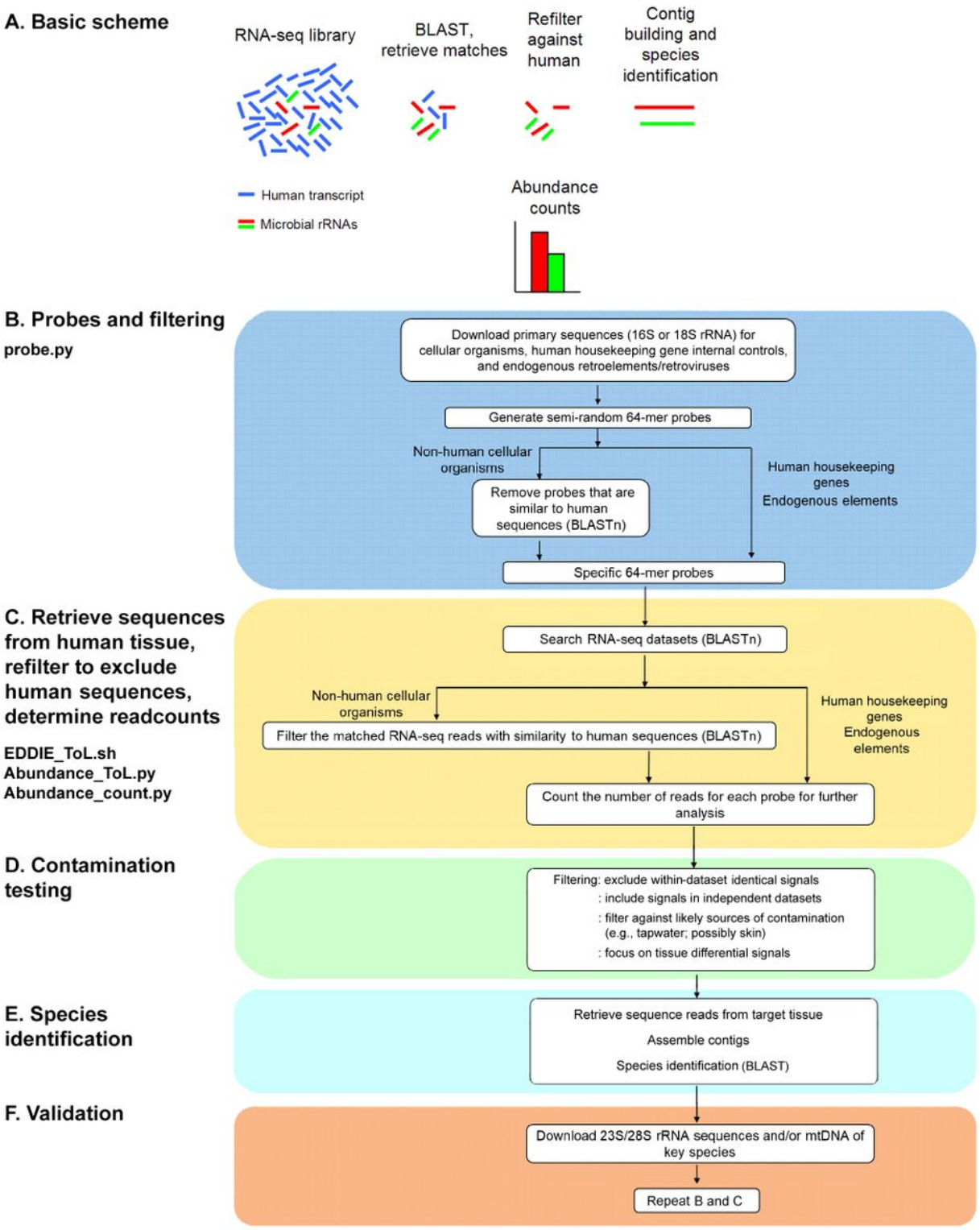

Figure 3

eToL workflow pipeline. The probe.py was used to download primary sequences and devise probes. The probes were aligned to RNA-seq datasets, and the matches (reads) were filtered and counted by EDDIE_ToL.sh, Abundance_ToL.py and Abundance_count.py.

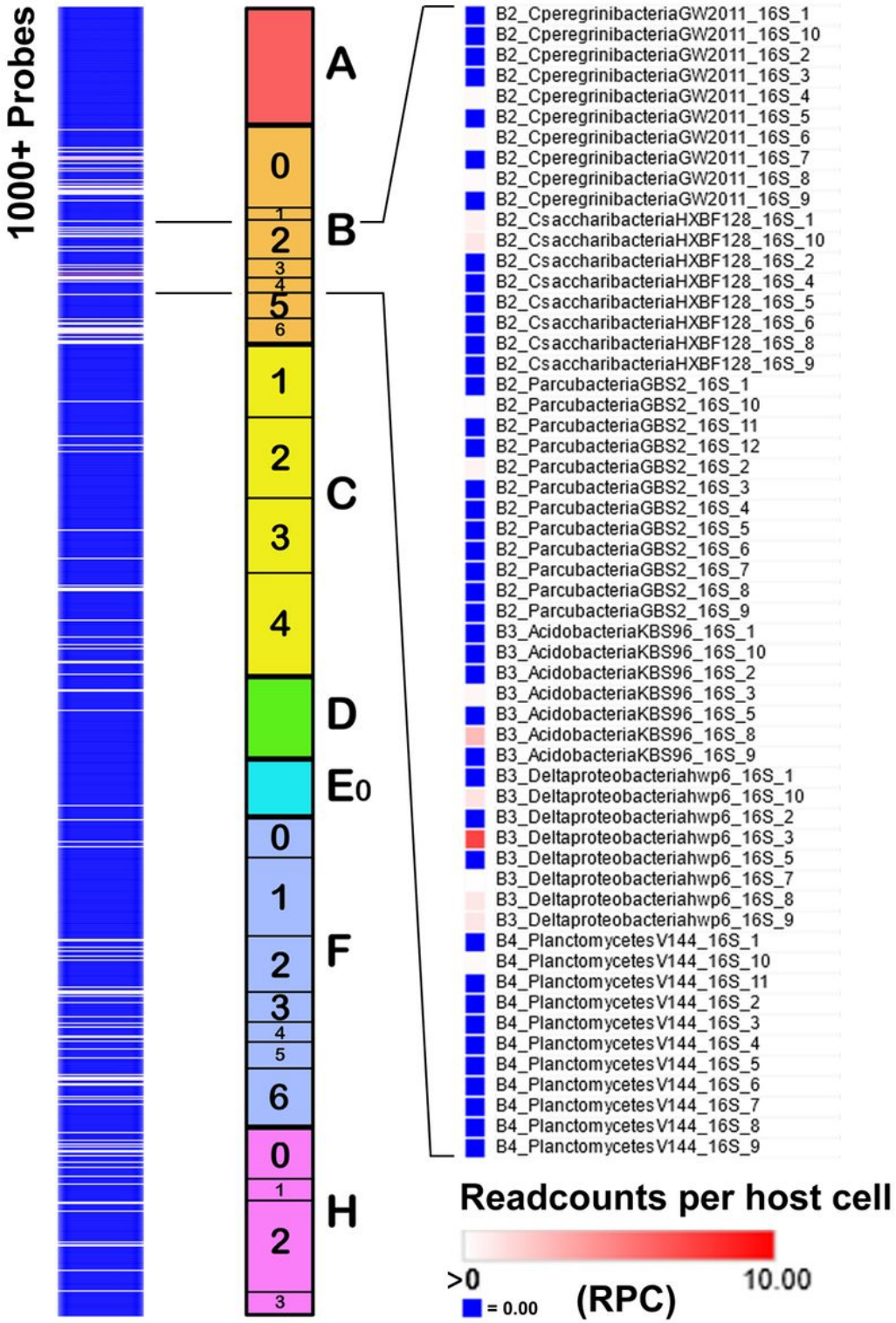

Figure 4

Visualization of the cellular microbiome profile. (Left) Standard display (first brain sequence read archive listed in the supplementary material) generated with Morpheus where the number of matches for the 1000+ probes in a constant order A–H (Center) are colored such that blue = 0, white = lower cutoff (generally any value >0) and red ≥ upper bound (differs between experiments). All values are normalized

to the number of host cells determined by readcounts for two housekeeping genes. (Right) Magnification of the B2–B4 region showing matches for individual probes.

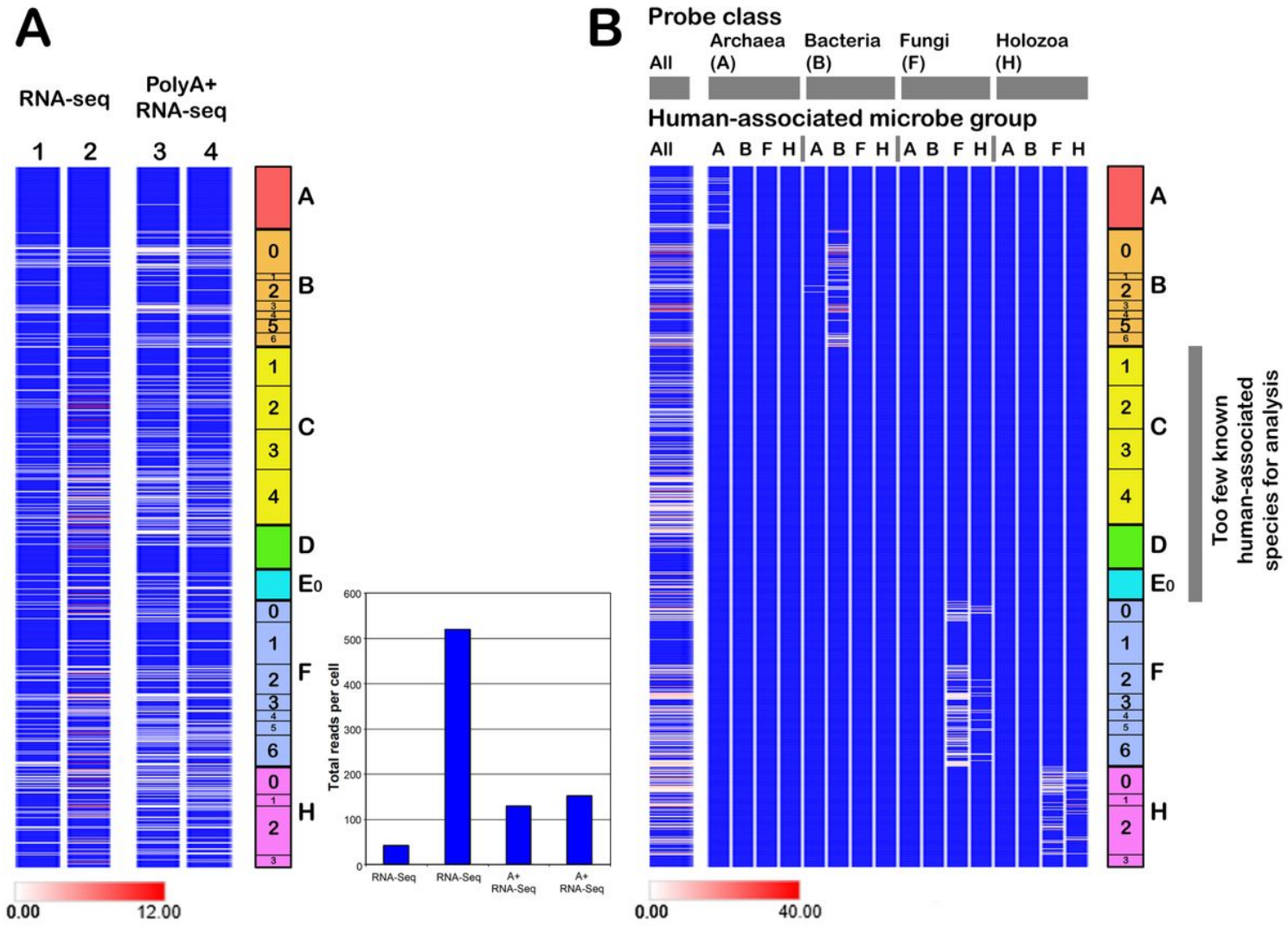

Figure 5

Analysis of RNA-seq of total RNA versus polyA+ RNA, and class specificity of the different probes. (A) Two brain (cortex) polyA+ RNA datasets were available for study, these were matched with two (cortex) total RNA datasets. Although the total number of microbial matches spanned a 10-fold range (inset), as expected given independent sample preparation and sequencing methods, the figure shows that polyA+ RNA selection does not remove all rRNA, and the number of overall matches (inset) for the two polyA+ datasets was intermediate between the two matched total RNA datasets. (B) (Left) The whole probe collection was used to probe the list of 100+ human-associated microbes (PATHLIST). (Right) Probes for classes A (Archaea), B (Bacteria), F (Fungi), and H (Holozoa) were then separately used to probe PATHLIST subclassified into the same groups. Classes C (Chloroplastida), D (Amoebozoa), and E0 (basal Eukaryota) were not studied because too few human-associated species are known. The figure demonstrates that each probe class principally detects species of the same class, although some cross-matching was observed between Fungi and Holozoa as expected because of evolutionary relationships.

Class C–E0 probes failed to find matches in classes A, B, F, and H (panel B), but some cross-matching is expected because these probes found matches in the complete PATHLIST (left).

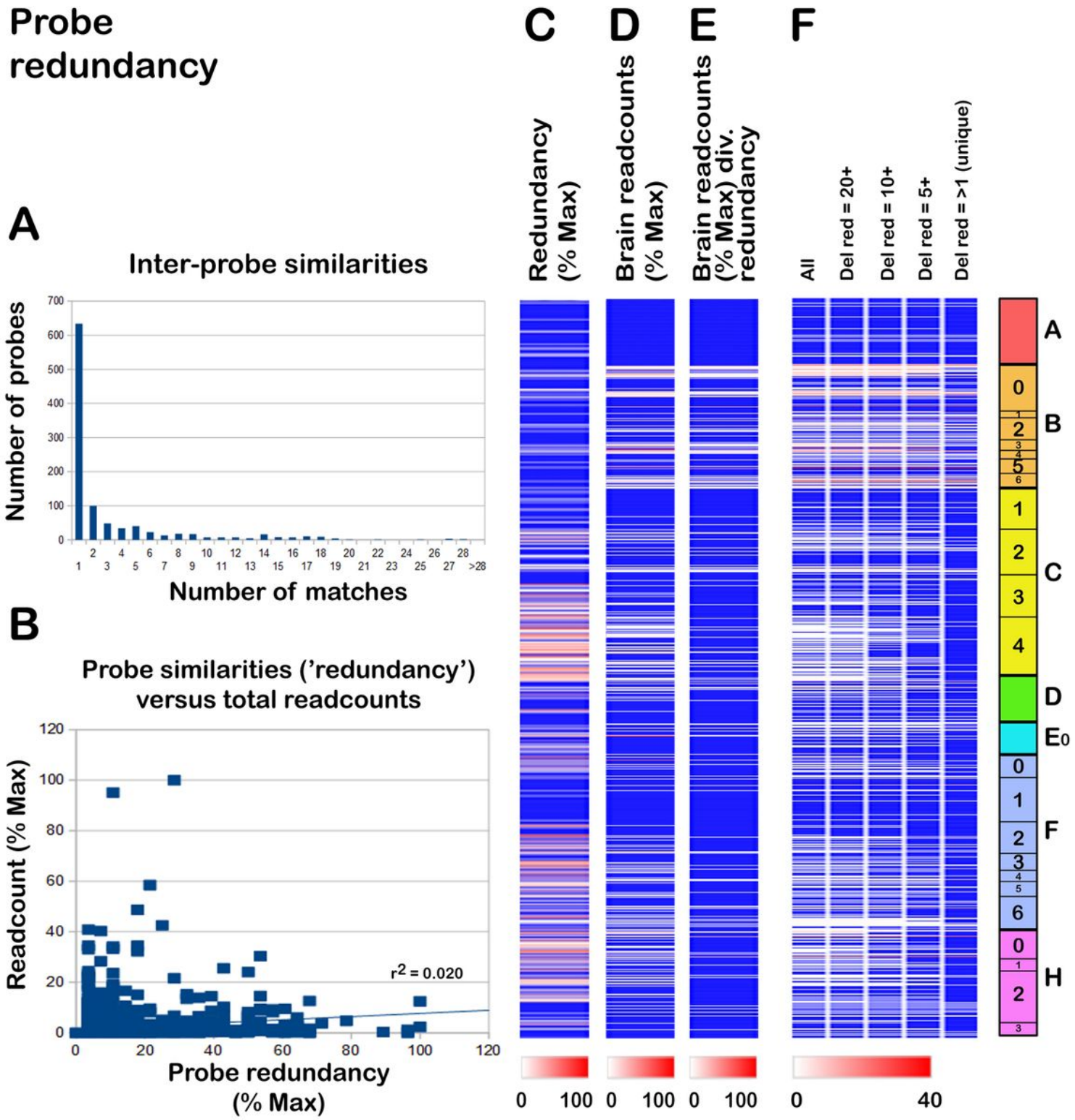

Figure 6

Overlaps within the probe collection ('redundancy'). (A) The collection was compared to itself, revealing that over 60% were unique. (B) To determine how probe overlaps might affect sequence detection, we

compared the total readcounts (brain) to the extent of probe redundancy (number of overlaps in the probe collection), demonstrating that high readcounts do not correlate with redundancy. The r^2 of the trendline (Microsoft Excel) was 0.02, showing that only 2% of the observed variation in numbers of matches can be ascribed to probe redundancy. (C,D) Side-by-side comparison of probe redundancy (C) with brain readcounts (D), demonstrating no obvious correlation. (E) As in (D), but the readcount scores have been divided by the redundancy of each probe, showing some changes, but conservation of the overall pattern. (F) Effect of removing probe signals with different levels of redundancy (≥20, ≥10, ≥5, and >1). Although the overall profile was retained, this degraded many signals, indicating that the complete probe list is preferable for comprehensive retrieval.

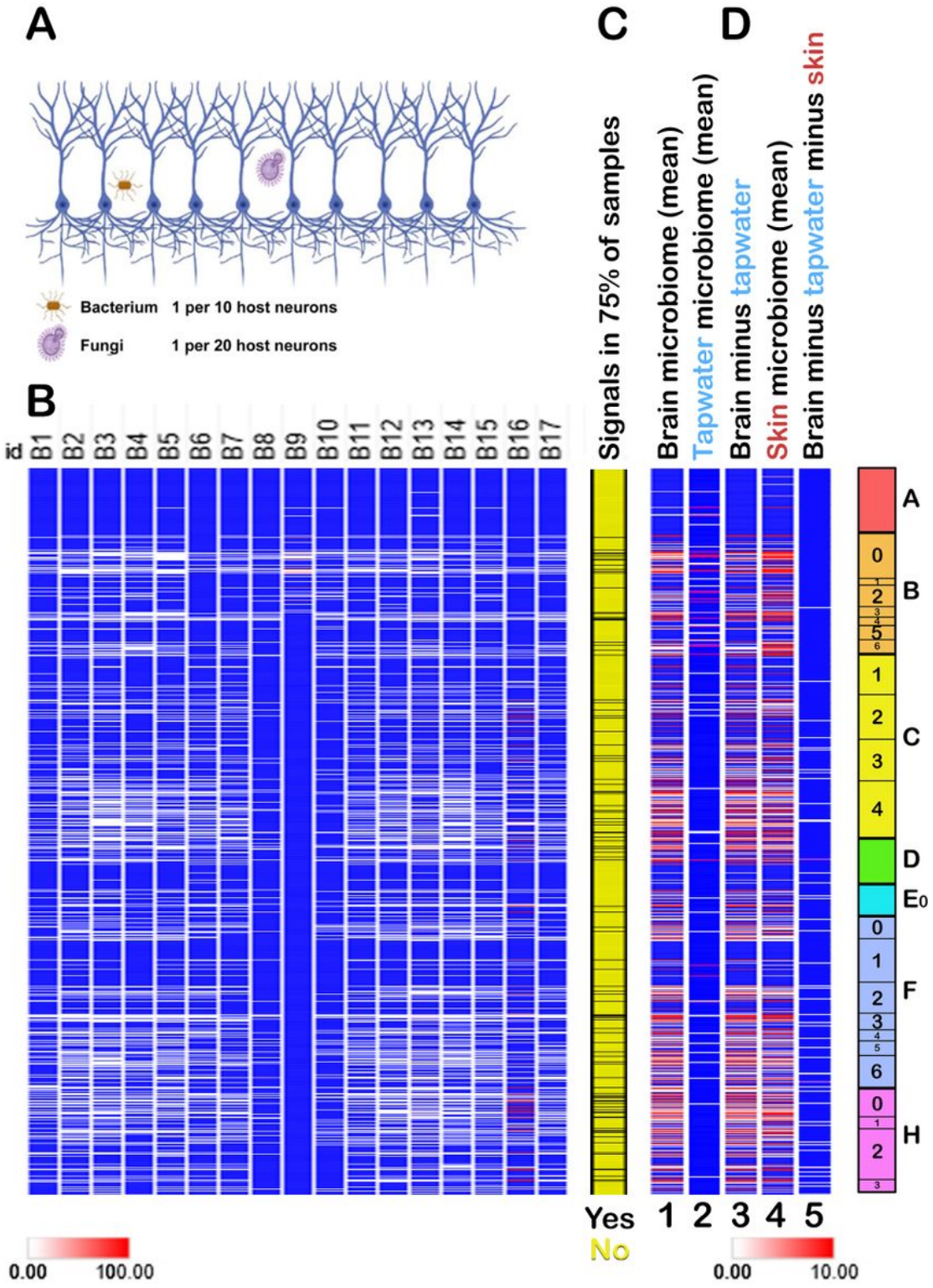

Figure 7

The profile of the cellular microbiome in human brain. (A) Schematic of the relative representation of microbes in normal brain (cortex): overall readcounts from 17 independent brain samples (below) indicate that bacteria and fungi are the major species in brain. (B) Profiles of readcounts in 17 independent brain samples showing different patterns in different brain RNA-seq datasets (where B9 and B16 have distinctive patters) but overall conservation of the profile. Estimated abundances from the

calculated numbers of rRNA per cell in the different organisms are as follows Archaea (10^{-5} microbes per host cell), Bacteria (0.14), Chloroplastida (0.06, but type 2 contamination has not yet been excluded, **Table 1**), Amoebozoa (0.01), basal Eukaryota (0.01), Fungi (0.05), Holozoa (0.05, possibly because of cross-matching with Fungi, **Figure 5B**). Bacteria and Fungi represent 41% and 13% of the total burden (together >50%). (C) Signals that are present in >75% of all samples. (D) The brain has its own microbiome. (Lane 1) The mean brain microbiome profile from 17 independent samples. (Lane 2) The mean profile in tapwater. (Lane 3) Brain profile where all signals also detected in tapwater have been removed. (Lane 4) Mean skin microbiome profile. (Lane 5) Brain profile where all signals also detected in tapwater and skin have been removed, showing degradation of the signal. Although type 1 contamination cannot be formally excluded here, there may be overlaps between the brain and skin microbiomes (Discussion). However, despite attenuation of the signal, the subtraction demonstrates that there are microbial signals in brain that do not occur in skin. Panel (A) was created at Biorender.com.

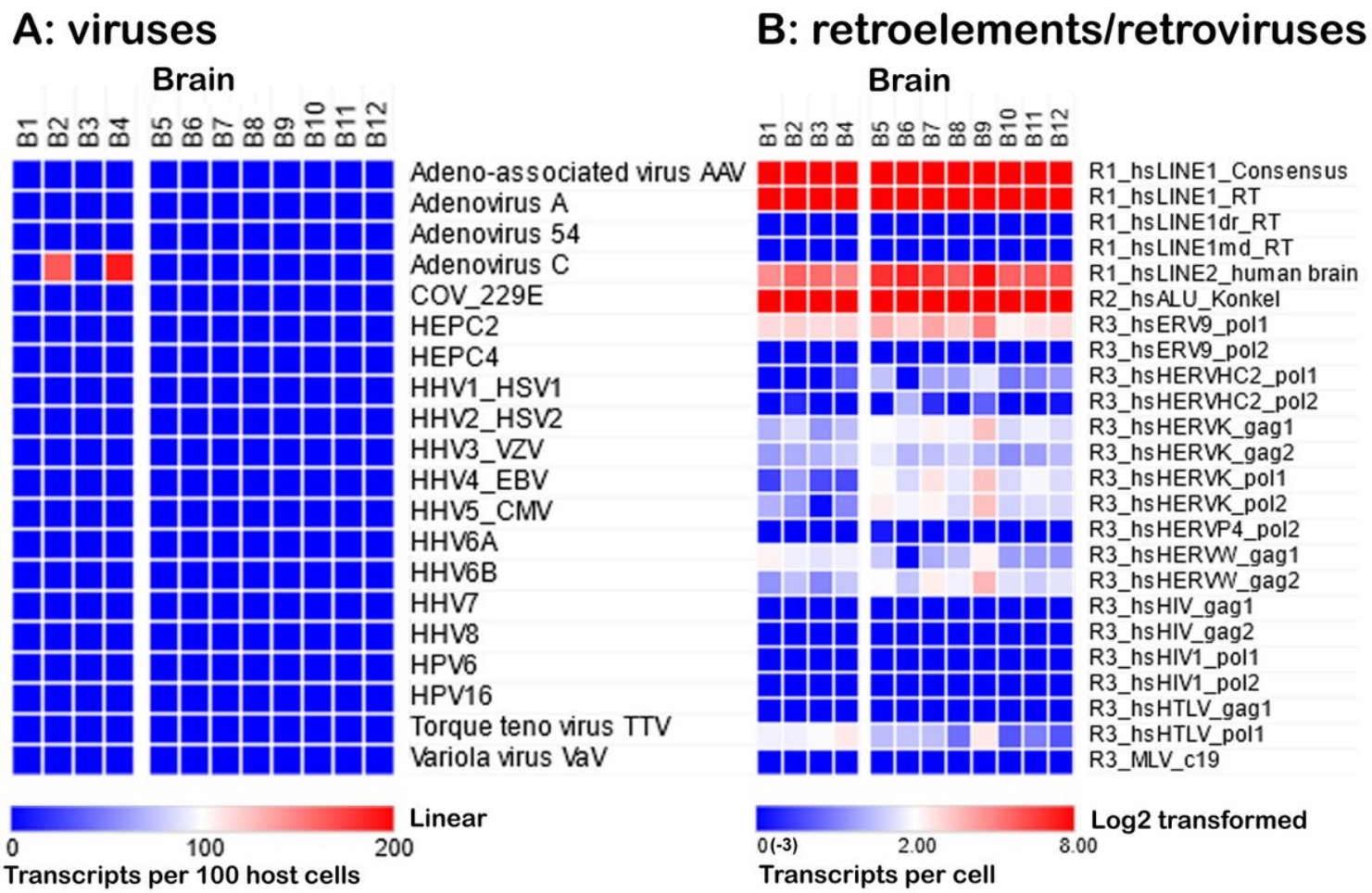

Figure 8

Viruses and retroelements in brain. (A) Screening of 12 normal brain samples (Miami and Rockefeller datasets) with the stripped (to remove all sequences similar to human) genomes corresponding to the top 20 viruses (>99% of brain matches in Readhead et al. 2018; text for details) revealed matches only for adenovirus type C. (B) Screening for retroelements and endogenous retroviruses showing that transcripts

for LINE and SINE elements are highly abundant, whereas endogenous retrovirus transcripts are much less abundant (8–128-fold). All samples were HIV1-negative.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Huetal.SUPPLEMENTARYMATERIALONLINE.pdf
- Huetal.SUPPLEMENTARYTABLES6STRIPPEDVIRUSESv5.doc
- floatimage9.jpeg