Circulating microbial content in myeloid malignancy patients carries diagnostic and prognostic potential

Thomas LaFramboise (Thomas.LaFramboise@case.edu)
Case Western Reserve University

Jakob Woerner
Case Western Reserve University

Yidi Huang
Case Western Reserve University

Stephan Hutter
Munich Leukemia Laboratory

Jesús Sánchez
Centro de Investigación del Cáncer

Janet Wang
Case Western Reserve University

Yimin Wang
Case Western Reserve University

Michael Aaby
Case Western Reserve University

Daniel Schnabel
Case Western Reserve University

Wanying Xu
Case Western Reserve University

Jaroslaw Maciejewski
Cleveland Clinic https://orcid.org/0000-0002-6837-4346

Mehmet Koyuturk
Case Western Reserve University

Torsten Haferlach

Article

Keywords: myeloid malignancy, diagnostic and prognostic tools, Epstein-Barr

Posted Date: April 13th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-380836/v1

Version of Record: A version of this preprint was published at Nature Communications on February 24th, 2022. See the published version at https://doi.org/10.1038/s41467-022-28678-x.

Circulating microbial content in myeloid malignancy patients carries diagnostic and prognostic potential

Jakob Woerner[1], Yidi Huang[1], Stephan Hutter[2], Jesús María Hernández Sánchez[3], Janet Wang[1], Yimin Wang[1], Daniel Schnabel[1], Michael Aaby[1], Wanying Xu[1], Mehmet Koyuturk[4], Jaroslaw Maciejewski[5], Torsten Haferlach[2], Thomas LaFramboise[1]*.

[1]Department of Genetics and Genome Sciences, Case Western Reserve University, Cleveland, USA

[2]Munich Leukemia Laboratory, Munich, Germany

[3]Centro de Investigación del Cáncer, Salamanca, Spain

[4]Department of Computer Science, Case Western Reserve University, Cleveland, USA

[5]Department of Translational Hematology & Oncology Research, Cleveland Clinic Foundation, Cleveland, USA

*Correspondence:

Thomas LaFramboise

Department of Genetics and Genome Sciences

10900 Euclid Avenue

Case Western Reserve University

Cleveland, OH 44106

TXL80@case.edu

ABSTRACT

Although recent work has characterized the microbiome in solid tumors, microbial content in hematological malignancies is not well-characterized. Here we analyzed existing deep DNA sequence data from the blood and bone marrow of 1,870 patients with myeloid malignancies, along with healthy controls, for bacterial, fungal, and viral content. After strict quality filtering, we find evidence for dysbiosis in disease cases, and distinct microbial signatures among diagnoses. In patients with low-risk myelodysplastic syndrome, we provide evidence that Epstein-Barr infection status refines risk stratification into more precise categories than the current standard. Motivated by these observations, we construct machine-learning classifiers that can discriminate among disease subtypes based solely on bacterial content. Our study highlights the potential of the circulating microbiome as a diagnostic and prognostic tool.

INTRODUCTION

Myeloid malignancies are diseases that result from abnormal proliferation or lack of differentiation in myeloid progenitor cells. This class of neoplasms includes acute myeloid leukemia (AML) as well as other diseases that can progress to AML such as myelodysplastic syndrome (MDS), characterized by dysplastic changes of hematopoietic progenitor cells, and myeloproliferative neoplasm (MPN), an over-proliferation of cells. Patients with characteristics of both MDS and MPN are given an MDS/MPN diagnosis. The annual incidence rate of myeloid malignancy is approximately 8 per 100,000 in Europe[1], for example, but is much higher among the elderly[2]. While survival has improved, it is highly variable among the different disease subtypes. In the US, AML five-year survival is estimated at around 27%, and treatment options become increasingly limited and ineffective with increasing patient age. A better understanding of the factors that influence disease outcomes and response to treatment is needed.

Meanwhile, evidence is growing for relationships between human cancers and the microbiome. In the blood cancer realm, some B-cell lymphomas have been associated with the bacteria *Helicobacter pylori*[3], while T-cell leukemia and Burkitt's lymphoma have long been known to be caused by viral infections (human T-cell lymphotropic virus 1[4] and Epstein-Barr virus[5], respectively). Recent work has investigated the relationships between the microbiome and clinical features in myeloid malignancy patients, though these studies have almost exclusively analyzed the gut microbiome. For instance, multiple studies have demonstrated that intestinal microbiota composition can predict survival in stem cell transplant patients[6,7]. The connection between the gut and the bone marrow is well-established[8], and therefore an impact of the gut microbiome on blood cancer and its treatment[9] is conceptually and empirically rational. However, the microbiome at the actual tumor site of myeloid malignancy – bone marrow and peripheral blood – remains unexplored. This stands in contrast to solid tumors, where microbiome research has recently been directed toward the tumor site itself[10-13]. A survey of >1500 tumors revealed distinct microbiome compositions for each of seven tumor types[13]. A

similar study[14] of >18,000 solid tumor and matched normal blood samples was able to find microbial signatures in both the solid tissue and blood that could accurately predict tumor type, and the blood signatures could differentiate between cancer patients and healthy individuals.

Traditionally, human blood and bone marrow have been considered to be normally sterile, and therefore microbiome analysis of these entities would only be performed when deleterious infection was suspected. However, evidence is now accumulating for a normal blood microbiome in healthy individuals[15]. It is believed that microbiota in circulation is partially derived directly from the gut through bacterial translocation[16], and therefore the established ability of intestinal flora to predict patient outcomes may also be valid for blood. Given this and the increasingly acknowledged relationships between microbial communities and response to treatment, we hypothesized analogous relationships between the bone marrow microbiome in myeloid malignancy patients and disease characteristics. To this end, we extracted bacterial, fungal, and viral sequence from deep shotgun sequencing of DNA in the bone marrow and blood of 1,870 myeloid malignancy patients, as well as in the bone marrow of healthy donors. Our goal was to elucidate relationships between microbial content/abundance and clinical features, including diagnosis and patient outcomes.

RESULTS

Patient overview and microbiome ascertainment

The patient cohort comprised 1,870 patients diagnosed with myeloid malignancy. Bone marrow (*n* = 1,756) or peripheral blood (*n* = 114) was taken at diagnosis and sent to the Munich Leukemia Laboratory between 2005 and 2017 for work-up. Diagnoses included MDS (*n* = 640), AML (*n* = 612), MPN (*n* = 354), and MDS/MPN (*n* = 264). Patient characteristics are provided in Table 1. Bone marrow samples from 12 healthy donors were also processed at the same site. DNA extracted from all samples was subjected to whole-genome sequencing, initially with the goal of comprehensively profiling somatic mutations in human DNA. Human genome average depth of coverage ranged between 76.8X and 183.8X (median 97.5X). We used PathSeq[17] – a tool that has been used in prior studies of tumor and blood microbes in cancer[14] – to identify reads derived from bacterial, fungal, and viral DNA. As shotgun metagenomic sequencing is known to be highly prone to artifacts (particularly for low-biomass samples such as blood), we followed strict filtering procedures. Briefly, two broad categories of reads were removed from consideration. First, we curated from the literature[14,18-21] a large list of 165 known problematic genera and 89 known problematic species (Methods; Supplementary Table 1). Reads that aligned to any of these taxa were omitted from downstream analyses. Second, we manually inspected the breadth of genome coverage in species. Species showing read alignments only to very focal regions of their genomes indicate that the reads were derived from cryptic human sequences[18], and thus reads aligning to such species were also filtered out. These extremely strict filtering steps removed 184,919,804 of 185,938,531 reads (99.45%) mapping to unique genera and 128,135,955 of 129,323,801 reads

(99.08%) mapping to unique species. Using the remaining reads, the abundance/dose of a taxon in a sample was quantified as the number of reads from the sample unambiguously aligning to the taxon, normalized to the number of human reads sequenced in the sample (see Methods).

Microbial landscape differs between cases and controls and among diagnoses

The sequencing and filtering protocols yielded means of 48.8 fungal reads, 120.9 viral reads, and 3,835.5 bacterial reads per sample (Figure 1a). To visualize differences and similarities among the samples, we generated t-distributed stochastic neighbor embedding (t-SNE) plots from the relative abundances of genera (Figure 1b), which showed clear grouping of normal control samples. All but one of the control samples has a first t-SNE coordinate greater than 25, whereas only four of 1,870 (0.2%) cases do (Fisher's exact $P < 2.2 \times 10^{-16}$). This suggests that microbe composition in the bone marrow of healthy individuals distinguishes them from disease cases. Furthermore, some grouping of patients by diagnoses was observed. For instance, more than half (52%) of MDS patients had second t-SNE coordinates below -10, while only 12.4% of patients with other diagnoses did (Fisher's exact $P < 2.2 \times 10^{-16}$). These observations raise the potential for microbe content to differentiate among diagnoses. Interestingly, in these t-SNE plots samples did not seem to group by age, sex, or blood/bone marrow status (Supplementary Figure 1), suggesting that these factors are not strongly associated with microbial content.

We next calculated, for each pair of samples, the genus-level Bray-Curtis dissimilarity, which measures how different each pair is regarding microbial content. Normal controls are far more similar to one another than to the case samples (Figure 1c). Interestingly, the same holds true for each diagnosis – that is, patients with the same diagnosis on average have lower dissimilarity than patients with different diagnoses (Figure 1d). We next computed the first two principal coordinates based on these dissimilarity measures. The resulting plots showed four distinct clusters (Figure 1e). The four diagnoses were not randomly dispersed among the four clusters. Instead, we observed strong enrichment in specific clusters for certain diagnoses (Figure 1f) (chi-squared test $P < 2.2 \times 10^{-16}$), providing further evidence that the microbial landscape may carry diagnostic information in myeloid malignancy. Cluster membership also showed strong association with various karyotypic features, particularly normal karyotype, complex karyotype, and trisomy 8 (chi-squared $P = 2.4 \times 10^{-6}$, 3.2×10^{-3}, and 0.016, respectively). For example, cluster 2 was enriched for normal karyotype patients, clusters 3 and 4 for complex karyotype, and cluster 1 for trisomy 8 (Figure 1g). No relationship was observed between the clusters and either age or sex ($P = 0.10$ and 0.33, respectively; Supplementary Figure 2).

Human herpesviruses prevalent among myeloid malignancy patients, with diagnostic and prognostic implications

Although viruses have not been widely implicated in myeloid malignancies, they have been linked to patient outcomes and do have established roles in some lymphoid-lineage malignancies, as noted above. In our cohort, viral reads were detected in 1,346 (72.0%) cases. Particularly

prevalent were torque teno viruses, which are extremely common in humans and are not conclusively linked with any disease[22], and human herpesviruses (Figure 2a). In the latter category, human betaherpesviruses 5 (human cytomegalovirus, HCMV) and 6 (roseolovirus), as well as *human gammaherpesvirus 4* (Epstein-Barr virus, EBV) occur at the highest dosage levels. In contrast, in the normal controls viral sequence was detected only at very low levels and species could only be assigned in four of the 12 samples (Figure 2b). Among cases, EBV and HCMV were by far the most frequently detected, found in 640 (34.2%) and 311 (16.6%) patients, respectively (Figure 2c).

Differences in overall viral infection were observed among the four diagnoses, with MDS patients showing the highest prevalence (78.0% *vs.* 68.9% for all other diagnoses; Fisher's exact test $P = 2.87 \times 10^{-5}$) as well as the highest dosage (median 14.0 *vs.* 6.9; Wilcoxon $P = 1.12 \times 10^{-8}$). In MDS patients, worse overall survival was associated with EBV presence, even after adjusting for age (Figure 2d), suggesting potential for EBV as a prognostic biomarker in MDS. None of the other diagnoses (AML, MPN, or MDS/MPN) showed an association between EBV and overall survival.

The current standard for risk stratification of MDS patients is the Revised International Prognostic Scoring System (IPSS-R)[23], which uses five risk categories. To explore whether the addition of EBV status could improve survival prediction, we tested for its association with survival within IPSS-R categories, finding that higher EBV status was able to refine risk prediction within the low-risk IPSS-R group. In aggregate, low-risk patients had survival outcomes between very low-risk and intermediate-risk patients, as expected, but low-risk patients with EBV infection were statistically indistinguishable from intermediate-risk patients, and those without EBV infection were indistinguishable from very low-risk patients (Figure 2e). We could detect no impact of EBV on survival in the higher risk IPSS categories, likely because any effect of EBV is overwhelmed by the strongly deleterious impact of the risk-defining characteristics (unfavorable blood cell counts and/or cytogenetic abnormalities).

EBV is frequently integrated into the host genome in known EBV-associated tumor types[24], though it is unclear whether these integrations promote malignancy. Given the paired-read nature of the sequencing data, we were able to identify human genome-mapped reads whose mates mapped to the EBV genome, yielding information regarding putative integration sites in our patient cohort. Among the 640 patients with EBV infection, 19 showed evidence of integration, with numbers of putative integration sites ranging from 1-17 per patient (Supplementary Table 2). Overall, 44.2% of integration sites were within gene bodies, all intronic. The only recurrently integrated gene was long non-coding RNA *LINC00486*, which has been identified as a recurrent hepatitis B virus integration site in the liver cancer intrahepatic cholangiocarcinoma[25].

Fungal prevalence is highest in MDS and is dominated by *Trichosporon asahii*

As with viruses, fungal infection was found in a higher proportion of MDS patients than those with other diagnoses (63.3% *vs.* 56.7%; Fisher's exact test P = 0.0064). *Trichosporon asahii* was the most commonly observed fungal species, present in 343 (18.3%) patients. *Trichosporon* infection is commonly reported in patients with acute leukemia[26] and MDS[27], and is a known contributor to mortality in hematological malignancy patients[28]. However, we found no association between *Trichosporon asahii* infection and overall survival.

Landscape of the bacteriome in circulation

The composition of bacteria present in human circulation is known to differ substantially from that in the gut. While gut bacteria are dominated by the phyla Bacteroidetes and Firmicutes[29], studies of the normal blood microbiome consistently demonstrate dominance of Proteobacteria and Actinobacteria[15]. We confirmed the latter composition in our healthy controls (Figure 3a), with Proteobacteria and Actinobacteria together comprising between 98.8% and 99.9% of all bacterial reads in each sample. Proteobacteria was somewhat more dominant (46.1%-77.4%) than Actinobacteria (21.3%-53.7%). The bacterial landscape in disease cases was substantially different from normal controls (Figure 3b). Proteobacteria was generally more dominant in cases (median relative abundance 91.3% *vs.* 61.1% in controls, Wilcoxon P = 8×10^{-7}), and its range was much larger, from 4.9% to 99.9%. Furthermore, Firmicutes and Bacteroidetes appear at relative abundances greater than 2% in some case samples (38 and 178 patients, respectively) but no control samples.

Studies of solid tumors have reported reduced microbial diversity as compared to matched-tissue controls[11,30]. Consistent with these reports, we find reduced α-diversity in cases compared to controls, as measured by the Shannon diversity index (see Methods), at all taxonomic levels save class (Figure 3c). For instance, the median α-diversity at the phylum level is 0.33 for cases and 0.70 for controls (Wilcoxon P = 4.1×10^{-6}); at the species level it is 2.1 for cases and 2.8 for controls (Wilcoxon P = 1.3×10^{-4}).

The reduced diversity in cases raises the question of whether there is competition and cooperation among the various bacterial taxa in patients with myeloid malignancy. To investigate, we tested for correlation/anticorrelation between all pairs of phyla and all pairs of classes. This was assessed both with regard to presence/absence (that is, whether two taxa tend to appear together more or less frequently than would be expected by chance) and abundance (whether the abundances are statistically correlated) (Figure 3d). Assessment of statistical significance, however, is not straightforward in this setting since assumptions of independence are violated. This effect is well known in studies measuring statistically significant mutational concordance/discordance in tumors[31], and renders the use of approaches such as Fisher's exact test prone to false positive discoveries of concordance. As such, we adopted a permutation approach to assess significance of concordance/discordance of taxa (see Methods). In our

analyses, all significant pairs showed positive correlation. Conceivably, this could suggest synergy among different bacterial entities. It could also indicate bacteriemia from a common source. For instance the strong positive correlation between Bacteroidetes and Firmicutes, both in terms of presence/absence and abundance, could be the result of varying degrees of intestinal barrier permeability ("leaky gut") as these are the two most common gut phyla, comprising 90% of microbiota there[32].

Relationship between bacteriome and clinical characteristics

The wide range of Actinobacteria-Proteobacteria ratios in our disease cohort led us to inquire whether the ratio corresponded with clinical characteristics. We observed strong association between diagnosis and Proteobacteria relative abundance (age-adjusted ANOVA $P = 7.1 \times 10^{-7}$; also visible in the horizontal bar at bottom of Figure 3b). In particular, AML had the highest Proteobacteria relative abundance (median 95.0% *vs.* 89.5% in non-AML cases; Wilcoxon $P < 2.2 \times 10^{-16}$). AML patients also tended to have the lowest bacterial α-diversity (Figure 3c) and richness (Figure 3e) but, interestingly, the highest overall bacterial abundance (median 6048 *vs.* 3738 for non-AML cases; Wilcoxon $P < 2.2 \times 10^{-16}$).

Given the observed differences in microbial content among patient diagnoses (Figure 1f; Figure 3c,e), we reasoned that microbial taxa might be able to classify patients by diagnosis. To this end, we constructed machine learning classifiers to diagnose disease from blood/bone marrow bacterial genus abundances. The genus level was chosen as a compromise between reduced resolution at the higher taxonomic levels and the overabundance of classifying taxa on the species level. For each diagnosis we developed a binary classifier to distinguish it from all others. Briefly, random forest[33] classifiers were constructed using a randomly selected training subset of the patient cohort. Classifier performance was assessed using the test subset comprised of the remaining samples (see Methods for details). The classifiers were best able to distinguish AML patients using bacterial content (average area under receiver operating characteristic curve (AUROC) = 0.87, 95% CI 0.84 - 0.90), though considerable separability was also achieved for MDS (AUROC = 0.84, 95% CI 0.81 - 0.88) and, to a lesser degree, MDS/MPN (AUROC = 0.75, 95% CI 0.70 - 0.80) and MPN (AUROC = 0.79, 95% CI 0.75 - 0.83) (Figure 4a). The performances of our machine learning classifiers provide evidence for diagnostic potential of microbial content in circulation.

In addition to providing an algorithm to assign classes (diagnoses in our case) based on features (abundances of bacterial genera), a random forest also assigns a measure of variable importance (VI) to each classifying feature. The VI of a feature measures the deterioration of accuracy resulting from obscuring that feature. In our setting, VI can therefore indicate the strength of association of a bacterial genus with each diagnosis. Among the genera with the highest VI (Figure 4b,c) are Dermabacter and Kytococcus, both of which are known as opportunistic pathogens frequently affecting immunocompromised individuals[34,35]. The species *Kytococcus schroeteri* in

particular has been found in multiple patients with myeloid malignancies[36]. Many other genera that contribute to discrimination among diagnoses (Figure 4b,c) are known causes of bacteremia, e.g., Staphylococcus, Streptomyces, Rothia, Gordonia, and Pandoraea, among others.

CONCLUSION

Here we have reported results of the first, to our knowledge, survey of microbial content in the blood and bone marrow of myeloid malignancy patients. We have catalogued bacterial, fungal, and viral content in circulation for 1,870 disease cases and 12 healthy controls, all processed and sequenced at the same center. Our overarching aims were to investigate the prognostic and diagnostic potential of the circulating microbiome in myeloid malignancies.

Interestingly, we did not observe any strong differences in microbial content between patient samples taken from peripheral blood and those taken from bone marrow, suggesting that the microbes and/or their DNA are transported freely into and out of the marrow through the bone vessels and sinusoids. However, it is unknown if the species detected here are living or active, and, if living, whether they are extracellular or have entered human cells. Even active microbes may be in circulation only transiently, likely translocating from the gut, skin, or mouth[37,38]. Strong effects of age and sex on microbial content were also not detected.

Our study provided strong evidence for substantial dysbiosis in the circulating microbiome of myeloid malignancy patients. The patients had significant shifts in dominant bacterial phyla as compared to healthy controls, and a reduction in α-diversity. This dysbiosis may partially be explained by intestinal permeability in some patients. Intestinal permeability is present in myeloid malignancy patients even before treatment[39], and has been recently implicated in leukemia development in mice[40,41]. All samples analyzed here were taken at diagnosis, and therefore microbial content would not be influenced by therapy.

In 1954 Ludwig Gross hypothesized, based on mouse experiments, a viral cause for human leukemia[42]. Although the hypothesis has not been validated for most leukemias (with some exceptions), recent work has shown associations between viral content and leukemia patient outcomes[43,44]. In our cohort, cases had a much higher viral dose than controls, largely from herpesviruses EBV and HCMV. Viruses and fungi was most prevalent in MDS patients, and our analysis revealed evidence of EBV involvement as a potential prognostic marker in MDS. Additional research is required to determine whether the high levels of viral infection among disease cases play a causal role or are instead a consequence of the immunosuppressive effects of the disease.

On the bacterial level, the landscape differed significantly among the four diagnoses, with multiple taxa showing significant differences among the diagnoses. AML patients had significantly higher bacterial dose but lower diversity, perhaps reflecting the dominance of Proteobacteria in AML patients. These observations motivated us to develop machine learning

classifiers to predict diagnosis from microbial content. The classifiers showed the ability to separate each diagnosis from the other three, demonstrating promise for further refinement of this approach. Our results here are analogous to recent microbe-based classifiers that were able to distinguish between different stages of some solid tumors[14]. Similarly, bacterial taxa have shown differing prevalence among breast cancer subtypes[13]. In our setting, the differences in bacterial signature among diagnoses could be the result of varying degrees of immunocompromise among the four entities leading to differential ability to combat bacteremia, though certainly many other explanations could be postulated.

The strengths of our study include a large cohort of disease cases, all processed sequenced at the same facility along with 12 healthy controls. We used shotgun metagenomic sequencing, which has advantages over 16S sequencing, including better taxonomic resolution, with typically higher revealed diversity[45]. It also enables ascertainment of viral and fungal content along with bacterial content. We took a very aggressive approach to data filtering, taking care to remove all reads that were likely artifactual results of contamination or mis-mapping of human reads. The fact that our filtering approach removed more than 99% of the data serves as a warning that it is critically important to take great care in microbiome studies (particularly those analyzing low-microbial biomass samples).

Our study also had weaknesses. It did not include an external validation cohort owing to the considerable expense in performing shotgun sequencing in a large cohort of myeloid malignancy patients with disparate diagnoses. It should be noted, however, that our diagnostic classifier was validated to some degree by building it on a subset of the cohort and testing it on the remaining samples. Ideally, we would have included technical controls[46] in our study, to account for sources of contamination specific to our experimental protocols. Since the samples were originally collected in clinical practice for mutation profiling, such controls were not available. Nonetheless, our computational approach was conservative, omitting a large list of known reagent and kit contaminants along with other artifacts.

In conclusion, this report serves as an initial baseline for future studies of the microbiome in circulation of myeloid malignancy patients. As growing evidence emerges that response to treatment may be influenced through gut microbiome perturbations, the results reported here may shed light on the potential for analogous manipulation of the blood microbiome to favorably impact patient outcomes[47]. Future work to replicate our findings in separate cohorts is crucial, and functional experiments are needed to determine whether and how microbes influence the course of disease. Such experiments will shed light on the potential of bacteria, fungi, and viruses to serve as biomarkers in myeloid malignancies, and may suggest treatment options for a subset of patients.

METHODS

Diagnosis, sample acquisition and processing, and whole-genome sequencing

All 1,870 cases were diagnosed using cytomorphology, immunophenotyping, cytogenetics, and molecular genetics following World Health Organization (WHO) guidelines. All patients gave their written informed consent for scientific evaluations. The study was approved by the Internal Review Board and adhered to the tenets of the Declaration of Helsinki. Additionally, 12 bone marrow samples from healthy donors were included as controls. Complete cytogenetic data according to ISCN nomenclature[48] is available for all patients by request at Munich Leukemia Laboratory.

For whole genome sequencing (WGS), peripheral blood or bone marrow aspirates were processed using the TruSeq DNA PCR-free library prep kit and 150 bp paired-end sequences were generated on a NovaSeq 6000 or HiSeqX instrument (Illumina, San Diego, CA). Fastq generation and read alignment to the human reference genome were performed using Illumina's BaseSpace platform (whole genome sequencing app 5.0.0).

Identification and quantification of microbial reads

Whole-genome bam files were converted back to fastq files using the GATK4[49] SamToFastq tool. The resulting fastq files were then aligned to the hg19 build of the human genome using bwa mem[50], yielding bam files that served as input into PathSeq[17], distributed as part of GATK 4.0.6.0. Briefly, PathSeq first removes all human genome-aligned reads, then aligns those remaining to an NCBI database of known microbial reference genomes. Default options were used with parameters --min-clipped-read-length 70 and --is-host-aligned true. Required reference files (microbe-fasta, microbe-bwa-image, and taxonomy-file) were downloaded as part of the GATK resource bundle. The output of PathSeq provides, for each taxon and patient, counts of reads that could be unambiguously assigned to that taxon. After filtering steps (see below), abundance/dose of taxon *i* in individual *j* was quantified as

$$6.4 \text{ billion } \times \frac{\text{number of reads aligning unambiguously to taxon } i \text{ in individual } j}{\text{number of reads aligning to the human genome in individual } j}.$$

The rationale for this measure is that it estimates the number of bases of the taxon DNA present per human cell, since there are ~6.4 billion bases of human sequence per human cell.

Quality filtering

Reads deemed unambiguously aligned by PathSeq were subjected to two filtering steps. First, we curated a list of genera and species that were reported in the literature as being problematic for various reasons, including: i) contamination in commercially available kits and reagents; ii) common low-read levels across tumor types; iii) anticorrelation between measured abundance and analyte concentration; iv) high frequency in negative blanks; and v) artifactual human sequence within species reference genomes. Second, alignments of species were manually examined for their locations in the microbe genome. Species that had reads only aligning to focal regions of their genome were flagged as problematic. We removed all reads unambiguously

aligned to these problematic taxa and propagated the removal up the taxonomic tree. For example, if a problematic genus had 20 unambiguously aligned reads in that patient, then 20 reads would also be removed from that genus' parent family, order, class, and phylum. Furthermore, all daughter species of the genus and their reads would be removed from further analysis.

Computing microbial landscape characteristics

Let r_{ij} denote the number of unambiguous reads from taxon *i* in individual *j*. Then the relative abundance for the taxon in that individual is computed as r_{ij}/T_j, where T_j denotes the total number of reads in that individual that map unambiguously to a taxon at the same taxonomic level as taxon *i*. The t-SNE coordinates were generated from the matrix giving the relative abundance of each genus for each sample. A series of pre-processing steps was first implemented as suggested by Kobak et al.[51], then FIt-SNE (FFT-accelerated Interpolation-based t-SNE)[52], a variant of t-SNE algorithm, was used to generate the coordinates.

Bray-Curtis dissimilarity statistics were calculated for each pair of samples using abundance. For each pair of samples *x* and *y*, Bray–Curtis dissimilarity is calculated across *n* genera as

$$BC_{xy} = \frac{\sum_i^n |x_i - y_i|}{\sum_i^n (x_i + y_i)}$$

where x_i and y_i denote the abundance of genus *i* in sample *x* and *y,* respectively. Then the dissimilarity measures were used to generate principal coordinates of this neighborhood matrix using the pcnm function in the vegan package (version 2.5-6).

The Shannon index for a taxonomic level was calculated for patient *s* as

$$H_s = -\sum_i^t p_i \ln p_i$$

where p_i is the proportion of unambiguous reads at the taxonomic level that map to taxon *i* within patient *s,* and *t* is the total number of taxa observed in patient *s* within the taxonomic level.

Statistical Analyses

All statistical analyses were performed using R version 4.0.3. Reported P-values are two-sided. To assess the significance of concordance/discordance between all pairs of taxa, the presence/absence of all *n* taxa within the same taxonomic level was represented as an *n* x 1870 matrix, where the rows represent the *n* taxa, the columns represent the 1,870 patients, and entry (*i*, *j*) is 1 if patient *j* has detected presence of taxon *i* and 0 otherwise. The odds ratio for each pair of taxa is computed in the observed data. To determine the statistical significance of each odds

ratio, we first repeatedly permuted the data matrix in a manner that keeps the row and column sums (the total number of patients harboring each taxon, and the total number of observed taxa within each patient, respectively) constant. In this way, we preserve taxonomic richness for each patient and overall frequency of each taxon. Permutations were performed using the permatfull function in the vegan package, using seed 2021 and parameters fixedmar = "both", shuffle = "both", mtype = "prab", and times = 100, and odds ratios computed for all pairs in each permutation. The P-values corresponding to each odds ratio *x* in the observed data is computed as the proportion of permuted odds ratios as or more extreme as *x*. The Q-values were computed from each observed P-value by dividing the average number of permuted P-values lower than or equal to the observed p-value (false discoveries) by the total number of observed P-values lower than or equal to the observed P-value (discoveries).

Survival analysis was performed using the R packages survival (3.2-7) and survminer (0.4.8). Age-adjusted hazard ratios and corresponding confidence intervals and p-values were obtained by fitting Cox proportional hazards regression models using the coxph function.

Calling EBV integration sites

To identify putative EBV integration sites, all read pairs with one end mapping to the human genome and the other mapping to the EBV genome were flagged. The mapped position in the human genome was reported as the putative EBV integration site.

Diagnostic classifier

A diagnostic classifier to predict one diagnosis against all others was built by training a random forest using the randomForest R package (4.6-14) with default parameters. The model was trained on the bacterial genus abundances in 70% of the samples, and performance was assessed on the remaining 30%. ROC curves were built for the classification model using R package PRROC (1.3.1). Here a sample is deemed to have the diagnosis in question if the proportion of trees classifying it as such exceeds a threshold. The ROC curves are generated by assessing sensitivity and false positivity at each value of the varying threshold. AUROC was used to evaluate performance on the test set using the ROCR R package (v1.0-11). To test whether the model assessments were robust, the process of training random forest models on 70% of samples and assessing performance on 30% (keeping the relative proportions of diagnoses equivalent in training and test sets) was repeated 1000 times for each diagnosis. 95% confidence intervals around the mean of the AUROCs from the 1000 classifiers were determined as the 2.5^{th} and 97.5^{th} percentiles from the 1000 AUROCs calculated.

REFERENCES

1. Visser, O.*, et al.* Incidence, survival and prevalence of myeloid malignancies in Europe. *Eur J Cancer* **48**, 3257-3266 (2012).
2. Craig, B.M., Rollison, D.E., List, A.F. & Cogle, C.R. Underreporting of myeloid malignancies by United States cancer registries. *Cancer Epidemiol Biomarkers Prev* **21**, 474-481 (2012).
3. de Martel, C.*, et al.* Global burden of cancers attributable to infections in 2008: a review and synthetic analysis. *Lancet Oncol* **13**, 607-615 (2012).
4. Matutes, E. Adult T-cell leukaemia/lymphoma. *J Clin Pathol* **60**, 1373-1377 (2007).
5. Brady, G., MacArthur, G.J. & Farrell, P.J. Epstein-Barr virus and Burkitt lymphoma. *J Clin Pathol* **60**, 1397-1402 (2007).
6. Peled, J.U.*, et al.* Microbiota as Predictor of Mortality in Allogeneic Hematopoietic-Cell Transplantation. *The New England journal of medicine* **382**, 822-834 (2020).
7. Taur, Y.*, et al.* The effects of intestinal tract bacterial diversity on mortality following allogeneic hematopoietic stem cell transplantation. *Blood* **124**, 1174-1182 (2014).
8. Santisteban, M.M., Kim, S., Pepine, C.J. & Raizada, M.K. Brain-Gut-Bone Marrow Axis: Implications for Hypertension and Related Therapeutics. *Circ Res* **118**, 1327-1336 (2016).
9. Andermann, T.M.*, et al.* The Microbiome and Hematopoietic Cell Transplantation: Past, Present, and Future. *Biol Blood Marrow Transplant* **24**, 1322-1340 (2018).
10. Riquelme, E.*, et al.* Tumor Microbiome Diversity and Composition Influence Pancreatic Cancer Outcomes. *Cell* **178**, 795-806 e712 (2019).
11. Mukherjee, P.K.*, et al.* Bacteriome and mycobiome associations in oral tongue cancer. *Oncotarget* **8**, 97273-97289 (2017).
12. Corning, B., Copland, A.P. & Frye, J.W. The Esophageal Microbiome in Health and Disease. *Curr Gastroenterol Rep* **20**, 39 (2018).
13. Nejman, D.*, et al.* The human tumor microbiome is composed of tumor type-specific intracellular bacteria. *Science* **368**, 973-980 (2020).
14. Poore, G.D.*, et al.* Microbiome analyses of blood and tissues suggest cancer diagnostic approach. *Nature* **579**, 567-574 (2020).
15. Castillo, D.J., Rifkin, R.F., Cowan, D.A. & Potgieter, M. The Healthy Human Blood Microbiome: Fact or Fiction? *Front Cell Infect Microbiol* **9**, 148 (2019).
16. Wang, L.*, et al.* Methods to determine intestinal permeability and bacterial translocation during liver disease. *Journal of immunological methods* **421**, 44-53 (2015).
17. Walker, M.A.*, et al.* GATK PathSeq: A customizable computational tool for the discovery and identification of microbial sequences in libraries from eukaryotic hosts. *Bioinformatics* (2018).
18. Breitwieser, F.P., Pertea, M., Zimin, A.V. & Salzberg, S.L. Human contamination in bacterial genomes has created thousands of spurious proteins. *Genome research* **29**, 954-960 (2019).
19. Fan, S.*, et al.* Next-generation sequencing of the cerebrospinal fluid in the diagnosis of neurobrucellosis. *International journal of infectious diseases : IJID : official publication of the International Society for Infectious Diseases* **67**, 20-24 (2018).
20. Salter, S.J.*, et al.* Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC biology* **12**, 87 (2014).

21. Stinson, L.F., Keelan, J.A. & Payne, M.S. Identification and removal of contaminating microbial DNA from PCR reagents: impact on low-biomass microbiome analyses. *Letters in applied microbiology* **68**, 2-8 (2019).
22. Okamoto, H. History of discoveries and pathogenicity of TT viruses. *Current topics in microbiology and immunology* **331**, 1-20 (2009).
23. Greenberg, P.L.*, et al.* Revised international prognostic scoring system for myelodysplastic syndromes. *Blood* **120**, 2454-2465 (2012).
24. Xu, M.*, et al.* Genome-wide profiling of Epstein-Barr virus integration by targeted sequencing in Epstein-Barr virus associated malignancies. *Theranostics* **9**, 1115-1124 (2019).
25. Li, M.*, et al.* Characterization of hepatitis B virus DNA integration patterns in intrahepatic cholangiocarcinoma. *Hepatology research : the official journal of the Japan Society of Hepatology* **51**, 102-115 (2021).
26. Kontoyiannis, D.P.*, et al.* Trichosporonosis in a tertiary care cancer center: risk factors, changing spectrum and determinants of outcome. *Scandinavian journal of infectious diseases* **36**, 564-569 (2004).
27. Odabasi, Z.*, et al.* Beta-D-glucan as a diagnostic adjunct for invasive fungal infections: validation, cutoff development, and performance in patients with acute myelogenous leukemia and myelodysplastic syndrome. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* **39**, 199-205 (2004).
28. Suzuki, K.*, et al.* Fatal Trichosporon fungemia in patients with hematologic malignancies. *European journal of haematology* **84**, 441-447 (2010).
29. Lozupone, C.A., Stombaugh, J.I., Gordon, J.I., Jansson, J.K. & Knight, R. Diversity, stability and resilience of the human gut microbiota. *Nature* **489**, 220-230 (2012).
30. Ferreira, R.M.*, et al.* Gastric microbial community profiling reveals a dysbiotic cancer-associated microbiota. *Gut* **67**, 226-236 (2018).
31. Canisius, S., Martens, J.W. & Wessels, L.F. A novel independence test for somatic alterations in cancer shows that biology drives mutual exclusivity but chance explains most co-occurrence. *Genome biology* **17**, 261 (2016).
32. Rinninella, E.*, et al.* What is the Healthy Gut Microbiota Composition? A Changing Ecosystem across Age, Environment, Diet, and Diseases. *Microorganisms* **7**(2019).
33. Chen, X. & Ishwaran, H. Random forests for genomic data analysis. *Genomics* **99**, 323-329 (2012).
34. Schaub, C.*, et al.* Relevance of Dermabacter hominis isolated from clinical samples, 2012-2016: a retrospective case series. *Diagnostic microbiology and infectious disease* **98**, 115118 (2020).
35. Blennow, O., Westling, K., Froding, I. & Ozenci, V. Pneumonia and bacteremia due to Kytococcus schroeteri. *Journal of clinical microbiology* **50**, 522-524 (2012).
36. Amaraneni, A., Malik, D., Jasra, S., Chandana, S.R. & Garg, D. Kytococcus schroeteri Bacteremia in a Patient with Hairy Cell Leukemia: A Case Report and Review of the Literature. *Case reports in infectious diseases* **2015**, 217307 (2015).
37. Whittle, E., Leonard, M.O., Harrison, R., Gant, T.W. & Tonge, D.P. Multi-Method Characterization of the Human Circulating Microbiome. *Frontiers in microbiology* **9**, 3266 (2018).

38. Paisse, S.*, et al.* Comprehensive description of blood microbiome from healthy donors assessed by 16S targeted metagenomic sequencing. *Transfusion* **56**, 1138-1147 (2016).
39. Sundstrom, G.M., Wahlin, A., Nordin-Andersson, I. & Suhr, O.B. Intestinal permeability in patients with acute myeloid leukemia. *European journal of haematology* **61**, 250-254 (1998).
40. Meisel, M.*, et al.* Microbial signals drive pre-leukaemic myeloproliferation in a Tet2-deficient host. *Nature* **557**, 580-584 (2018).
41. Ye, H.*, et al.* Subversion of Systemic Glucose Metabolism as a Mechanism to Support the Growth of Leukemia Cells. *Cancer cell* **34**, 659-673 e656 (2018).
42. Gross, L. Is leukemia caused by a transmissible virus? A working hypothesis. *Blood* **9**, 557-573 (1954).
43. Elmaagacli, A.H.*, et al.* Early human cytomegalovirus replication after transplantation is associated with a decreased relapse risk: evidence for a putative virus-versus-leukemia effect in acute myeloid leukemia patients. *Blood* **118**, 1402-1412 (2011).
44. Guan, H., Miao, H., Ma, N., Lu, W. & Luo, B. Correlations between Epstein-Barr virus and acute leukemia. *Journal of medical virology* **89**, 1453-1460 (2017).
45. Ranjan, R., Rani, A., Metwally, A., McGee, H.S. & Perkins, D.L. Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochemical and biophysical research communications* **469**, 967-977 (2016).
46. Eisenhofer, R.*, et al.* Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations. *Trends in microbiology* **27**, 105-117 (2019).
47. Davar, D.*, et al.* Fecal microbiota transplant overcomes resistance to anti-PD-1 therapy in melanoma patients. *Science* **371**, 595-602 (2021).
48. International Standing Committee on Human Cytogenomic Nomenclature, McGowan-Jordan, J., Simons, A. & Schmid, M. *ISCN : an international system for human cytogenomic nomenclature (2016)*.
49. McKenna, A.*, et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297-1303 (2010).
50. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
51. Kobak, D. & Berens, P. The art of using t-SNE for single-cell transcriptomics. *Nature communications* **10**, 5416 (2019).
52. Linderman, G.C., Rachh, M., Hoskins, J.G., Steinerberger, S. & Kluger, Y. Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. *Nature methods* **16**, 243-245 (2019).

ACKNOWLEDGEMENTS

This work was supported by US National Institutes of Health grants R01LM013067 and R21CA249138 (to T.L.) and the Torsten Haferlach Leukamiediagnostik Stiftung.

TABLES

Table 1: Patient Characteristics

			mean age (years)	karyotypic lesions (proportion)						
	n	sex (% female)	(1st-3rd quartiles)	normal	-5/-5q	-7	trisomy 8	-Y*	complex	median survival (years)
AML	612	45.8%	63.41 (54.38,75.20)	0.367	0.077	0.088	0.103	0.066	0.203	1.41
MDS	640	42.7%	71.07 (66.50,78.00)	0.609	0.189	0.025	0.044	0.114	0.028	6.14
MDS/MPN	264	40.2%	74.71 (70.78,80.72)	0.712	0.008	0.042	0.131	0.045	0.027	5.36
MPN	354	36.7%	62.79 (53.90,73.75)	0.555	0.012	0.009	0.046	0.039	0.028	18.64
overall	1870	42.2%	67.51 (60.85,77.10)	0.533	0.095	0.046	0.076	0.074	0.086	5.71

* calculated only from male patients

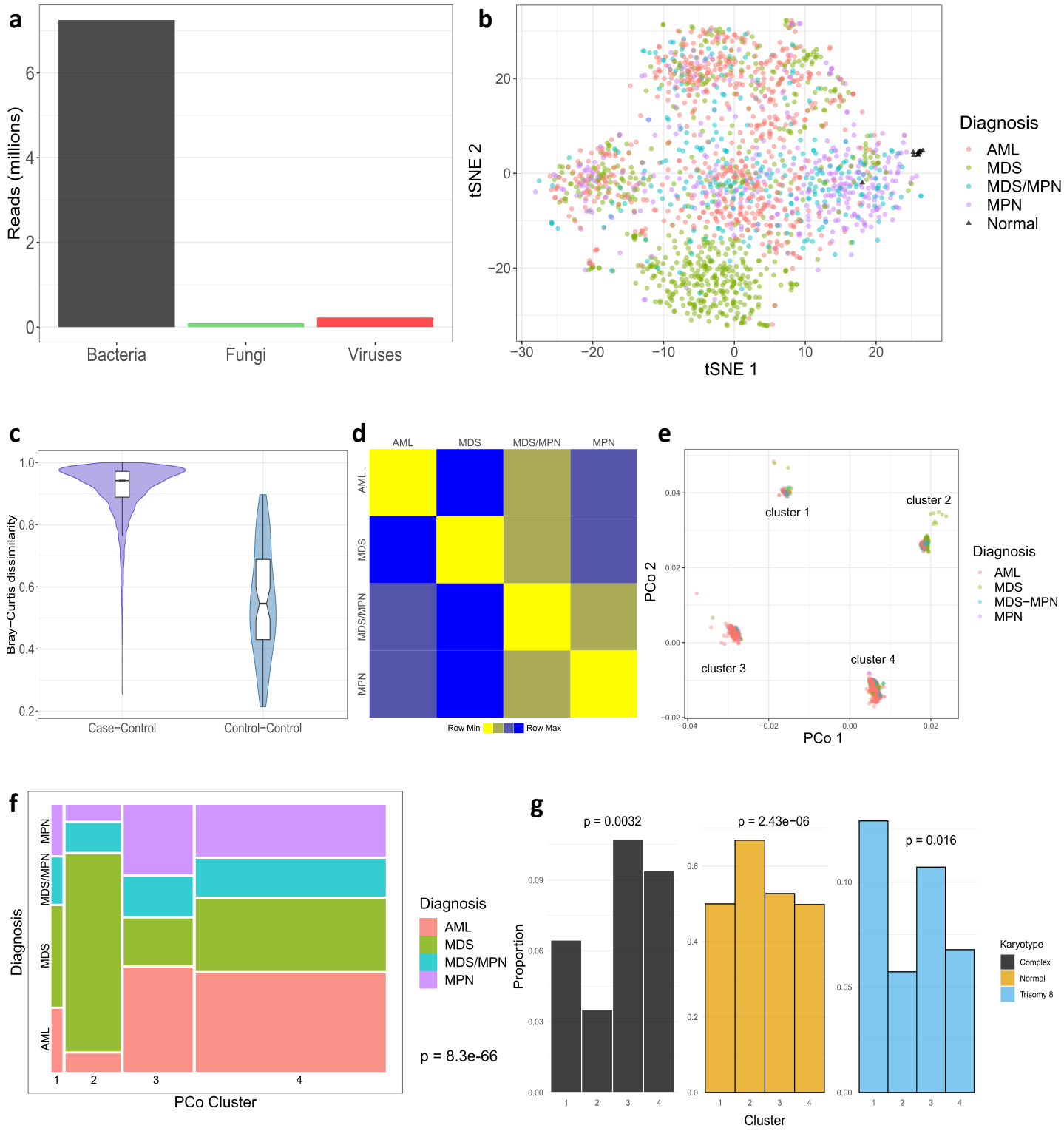

Figure 1: Landscape of microbial content in circulation. a) Barplot showing total numbers of reads for each of the three kingdoms. b) t-SNE plot colored by case/control status (controls shown as black triangles) and diagnosis. c) Bray-Curtis dissimilarity measures, on the genus level, for all case-control pairs (left) and all pairs of control samples (right). d) Heatmap representing the average of all Bray-Curtis dissimilarity measures between sample pairs from the indicated groups. Squares are colored according to rank in the row (yellow = most similar, blue = least similar). e) The first two principal coordinates, on the genus level, colored by diagnosis as in panel b. For clarity, two outliers (an MDS patient and an AML patient) are omitted. f) Mosaic plot indicating the proportion of the patient cohort in each cluster/diagnosis pair. The area of each rectangle (colored by diagnosis) is proportional to number of patients in the corresponding diagnosis and cluster. g) Barplots indicating proportion of patients with complex karyotype, normal karyotype, and trisomy 8 in each principal coordinate cluster.

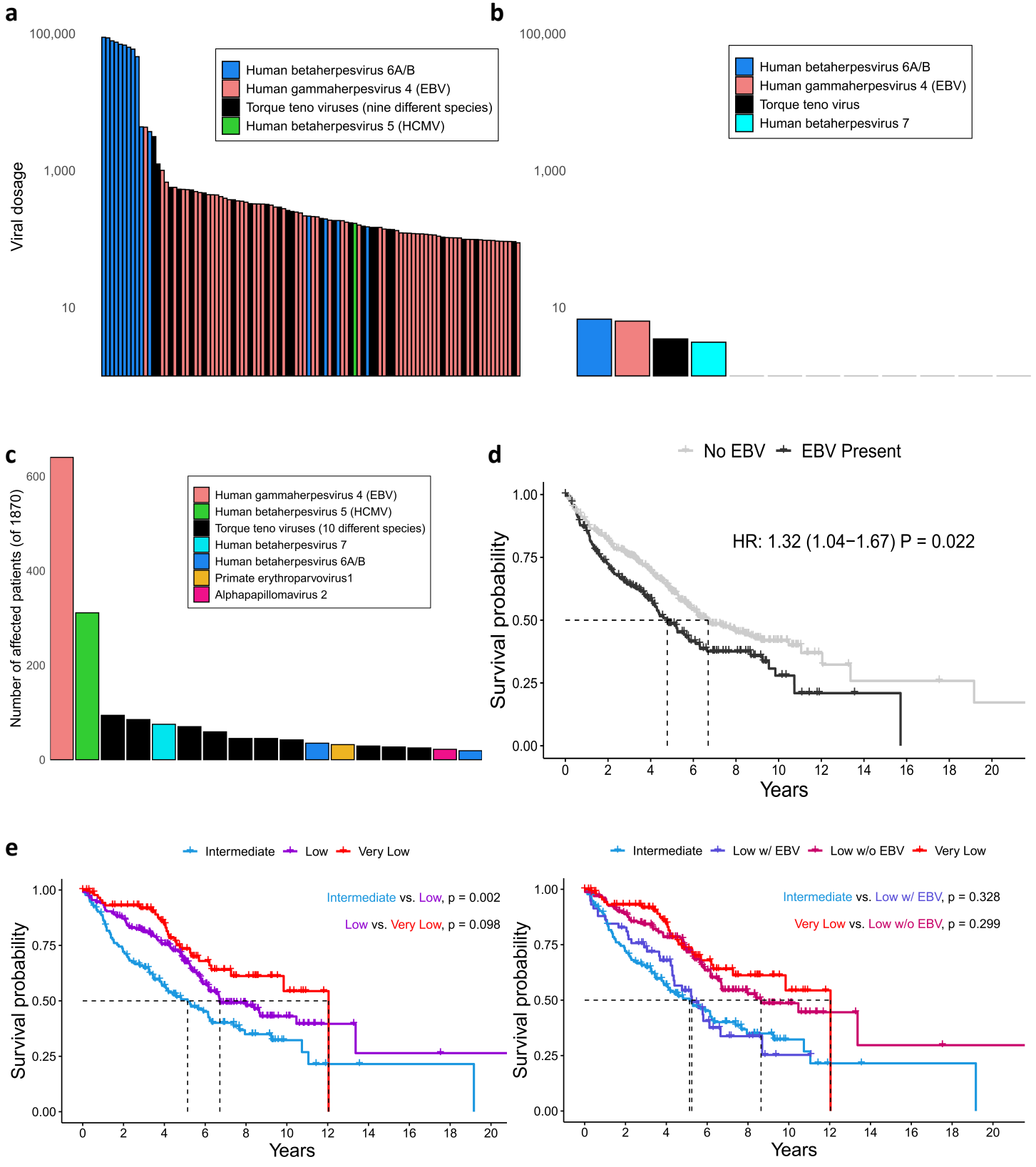

Figure 2: Circulating viral content is associated with clinical characteristics. a) The 100 highest dosages of viral species in individual patients. Each bar represents the dosage of the corresponding virus in a single patient. b) All controls are shown with their corresponding detected viruses, on the same (logarithmic) scale as panel a) for comparison. Only the leftmost four samples had any detectable viral species. c) The prevalence of viral species (those found in >1% of cases are shown). d) Presence of EBV is associated with worse survival in MDS patients (HR and P-value are age-adjusted). e) The left panel shows Kaplan-Meier curves for intermediate, low, and very low IPSS-R categories. In the right panel, the low category is stratified by EBV status. Low-risk patients with and without EBV become statistically indistinguishable from the intermediate-risk and very low-risk categories, respectively.

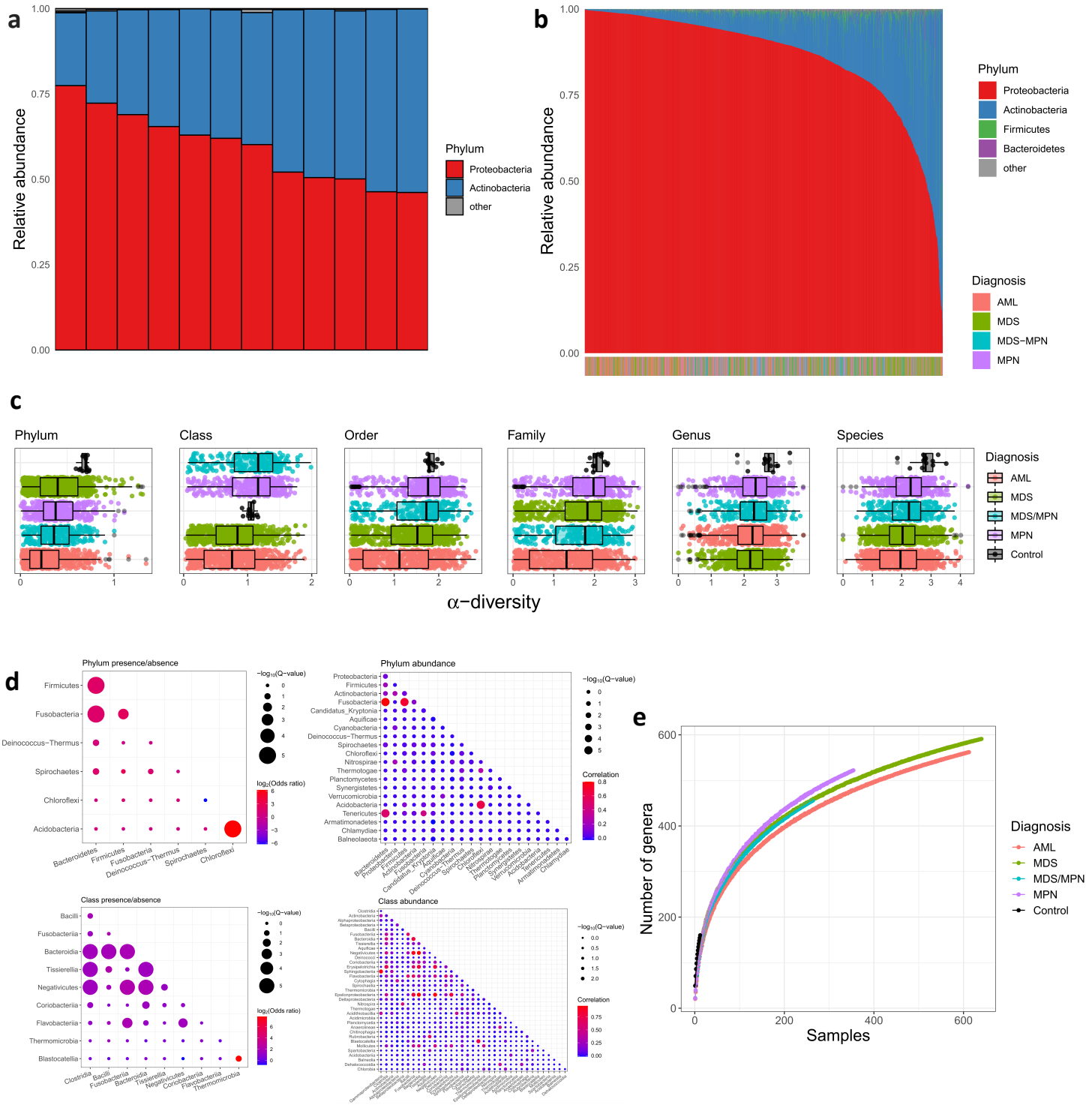

Figure 3: The bacterial landscape in the bone marrow/blood of myeloid malignancy patients and controls. a) Relative abundances of phyla are represented by a colored bar for each of the 12 control bone marrow samples. b) The 1870 colored bars, one for each patient, are ordered left to right by decreasing *Proteobacteria* relative abundance. The diagnosis of each patient is indicated in the horizontal color bar at the bottom (the enrichment of AML patients among the *Proteobacteria*-dominant samples is apparent by the color shift at the left side of the bar). c) α-diversity of each sample within each taxonomic level, stratified by case/control status and diagnosis. Boxplots are ordered top to bottom in decreasing median α-diversity. d) Plot showing pairwise concordance/discordance of taxa, at the phylum (top) and class (bottom) levels, both with regard to presence/absence (left) and abundance (right). Sizes of the circles indicate statistical significance, and color indicates strength and direction of association (odds ratio or Pearson correlation). Only taxa with significant ($Q < 0.1$) concordance/discordance with at least one other taxon are shown. e) Rarefaction plot showing number of genera as a function of number of patients, stratified by diagnosis. For each patient number *n*, a random sample of *n* patients was drawn from each diagnosis, 500 times. Solid curves represent the mean across the 500 replicates. For control samples, sampling is performed exhaustively (that is, all possible subsets of *n* individuals are selected for each $n = 1,2,\ldots,12$).

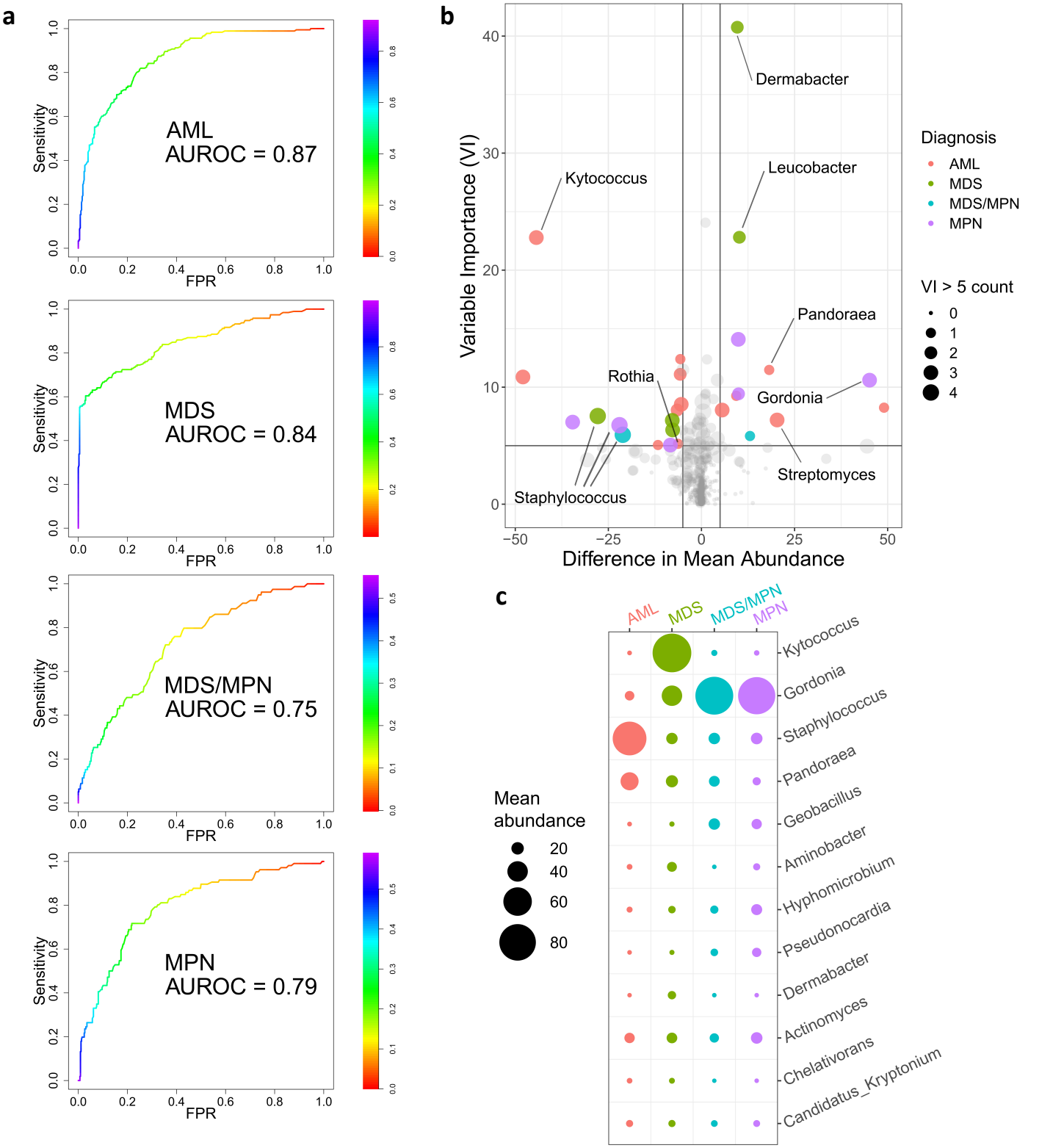

Figure 4: Bacterial composition differs among diagnoses. a) ROC curves showing, for each diagnosis, the performance on the test set (randomly-selected 30% of samples) of binary random forest classifier trained on the training set (remaining 70%). The AUROC values shown are averaged across 1000 random 70%/30% splits. b) Volcano plot showing enrichment/depletion of bacterial genera in specific diagnoses. Here horizontal axis indicates differences in mean abundance (diagnosis of interest – all others), and variable importance is shown on the vertical axis. Point size indicates number of diagnoses (0 = smallest, 4 = largest) for which the corresponding genus has variable importance > 5. Points with mean abundance difference > 5 and variable importance > 5 are colored by corresponding diagnosis. Points of interest are labeled with their corresponding genera. (VI = variable importance). c) Mean abundances, in each diagnosis, of the genera that are among the top five in variable importance for at least one of the diagnoses. Circle size indicates the average abundance in the corresponding diagnosis.

Figures

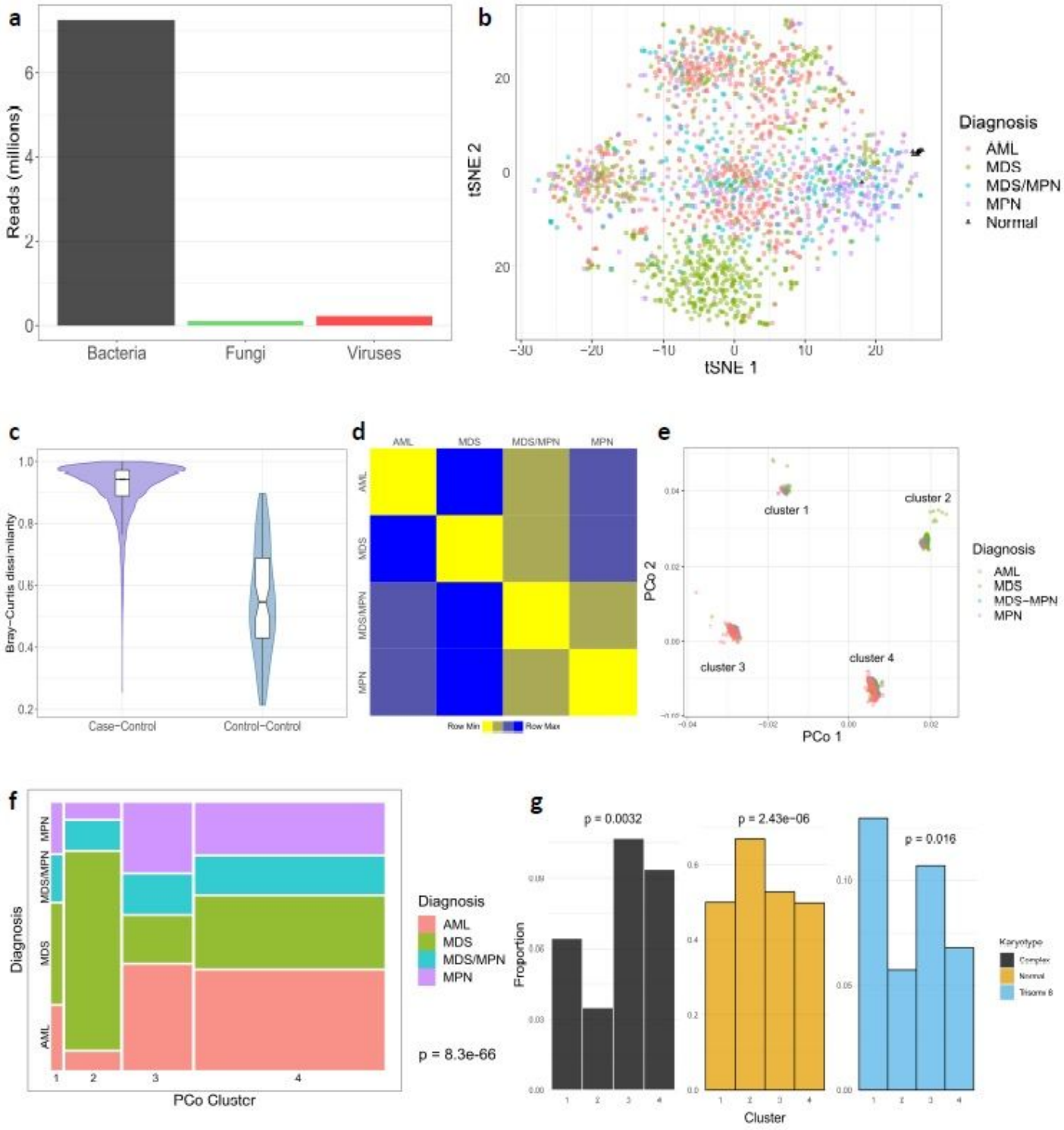

Figure 1

Landscape of microbial content in circulation. a) Barplot showing total numbers of reads for each of the three kingdoms. b) t-SNE plot colored by case/control status (controls shown as black triangles) and diagnosis. c) Bray-Curtis dissimilarity measures, on the genus level, for all case-control pairs (left) and all

pairs of control samples (right). d) Heatmap representing the average of all Bray-Curtis dissimilarity measures between sample pairs from the indicated groups. Squares are colored according to rank in the row (yellow = most similar, blue = least similar). e) The first two principal coordinates, on the genus level, colored by diagnosis as in panel b. For clarity, two outliers (an MDS patient and an AML patient) are omitted. f) Mosaic plot indicating the proportion of the patient cohort in each cluster/diagnosis pair. The area of each rectangle (colored by diagnosis) is proportional to number of patients in the corresponding diagnosis and cluster. g) Barplots indicating proportion of patients with complex karyotype, normal karyotype, and trisomy 8 in each principal coordinate cluster.

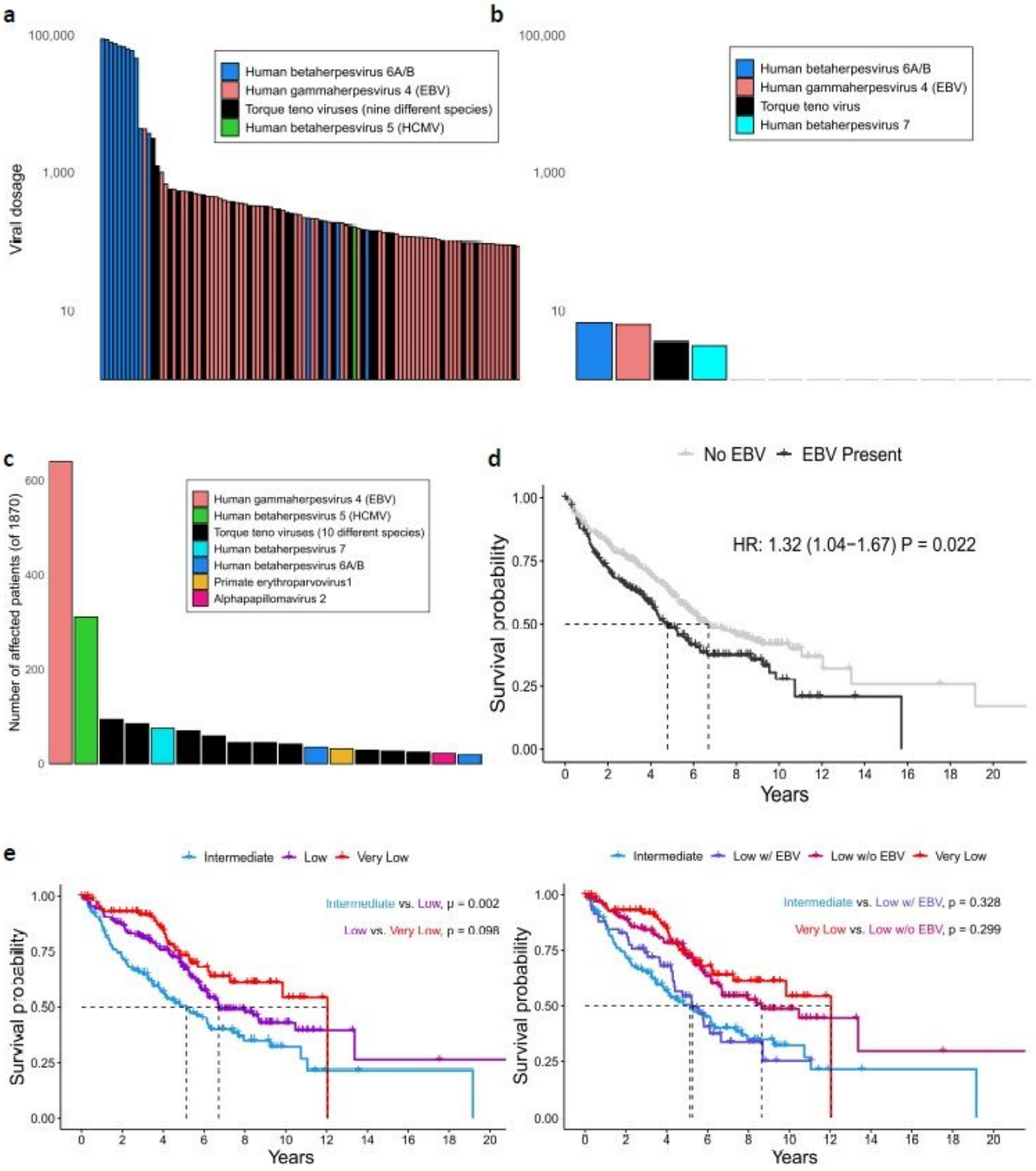

Figure 2

Circulating viral content is associated with clinical characteristics. a) The 100 highest dosages of viral species in individual patients. Each bar represents the dosage of the corresponding virus in a single patient. b) All controls are shown with their corresponding detected viruses, on the same (logarithmic) scale as panel a) for comparison. Only the leftmost four samples had any detectable viral species. c) The prevalence of viral species (those found in >1% of cases are shown). d) Presence of EBV is associated

with worse survival in MDS patients (HR and P-value are age-adjusted). e) The left panel shows Kaplan-Meier curves for intermediate, low, and very low IPSS-R categories. In the right panel, the low category is stratified by EBV status. Low-risk patients with and without EBV become statistically indistinguishable from the intermediate-risk and very low-risk categories, respectively.

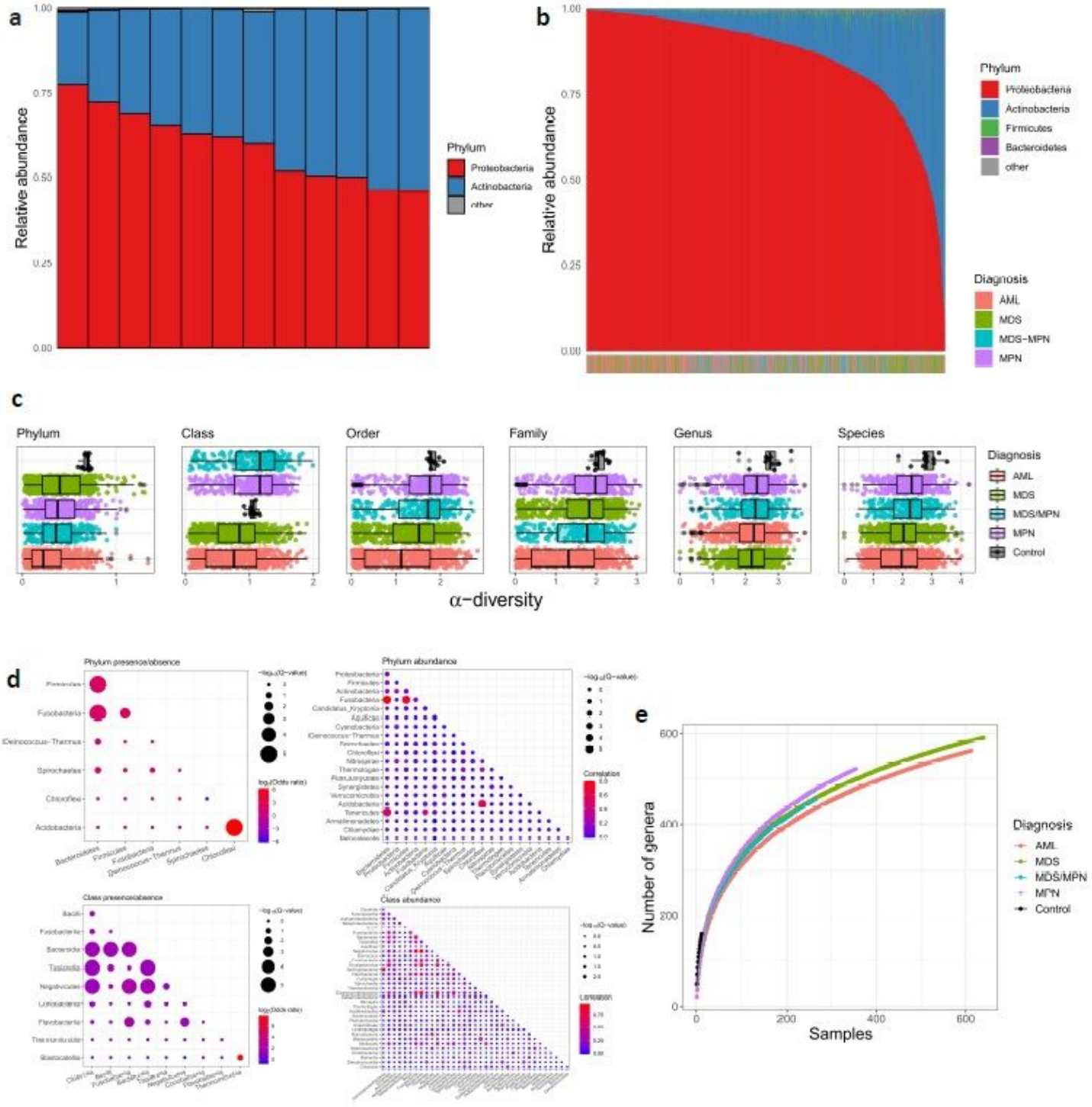

Figure 3

The bacterial landscape in the bone marrow/blood of myeloid malignancy patients and controls. a) Relative abundances of phyla are represented by a colored bar for each of the 12 control bone marrow samples. b) The 1870 colored bars, one for each patient, are ordered left to right by decreasing Proteobacteria relative abundance. The diagnosis of each patient is indicated in the horizontal color bar at the bottom (the enrichment of AML patients among the Proteobacteria-dominant samples is apparent by the color shift at the left side of the bar). c) α-diversity of each sample within each taxonomic level, stratified by case/control status and diagnosis. Boxplots are ordered top to bottom in decreasing median α-diversity. d) Plot showing pairwise concordance/discordance of taxa, at the phylum (top) and class (bottom) levels, both with regard to presence/absence (left) and abundance (right). Sizes of the circles indicate statistical significance, and color indicates strength and direction of association (odds ratio or Pearson correlation). Only taxa with significant ($Q < 0.1$) concordance/discordance with at least one other taxon are shown. e) Rarefaction plot showing number of genera as a function of number of patients, stratified by diagnosis. For each patient number n, a random sample of n patients was drawn from each diagnosis, 500 times. Solid curves represent the mean across the 500 replicates. For control samples, sampling is performed exhaustively (that is, all possible subsets of n individuals are selected for each n = 1,2,…,12).

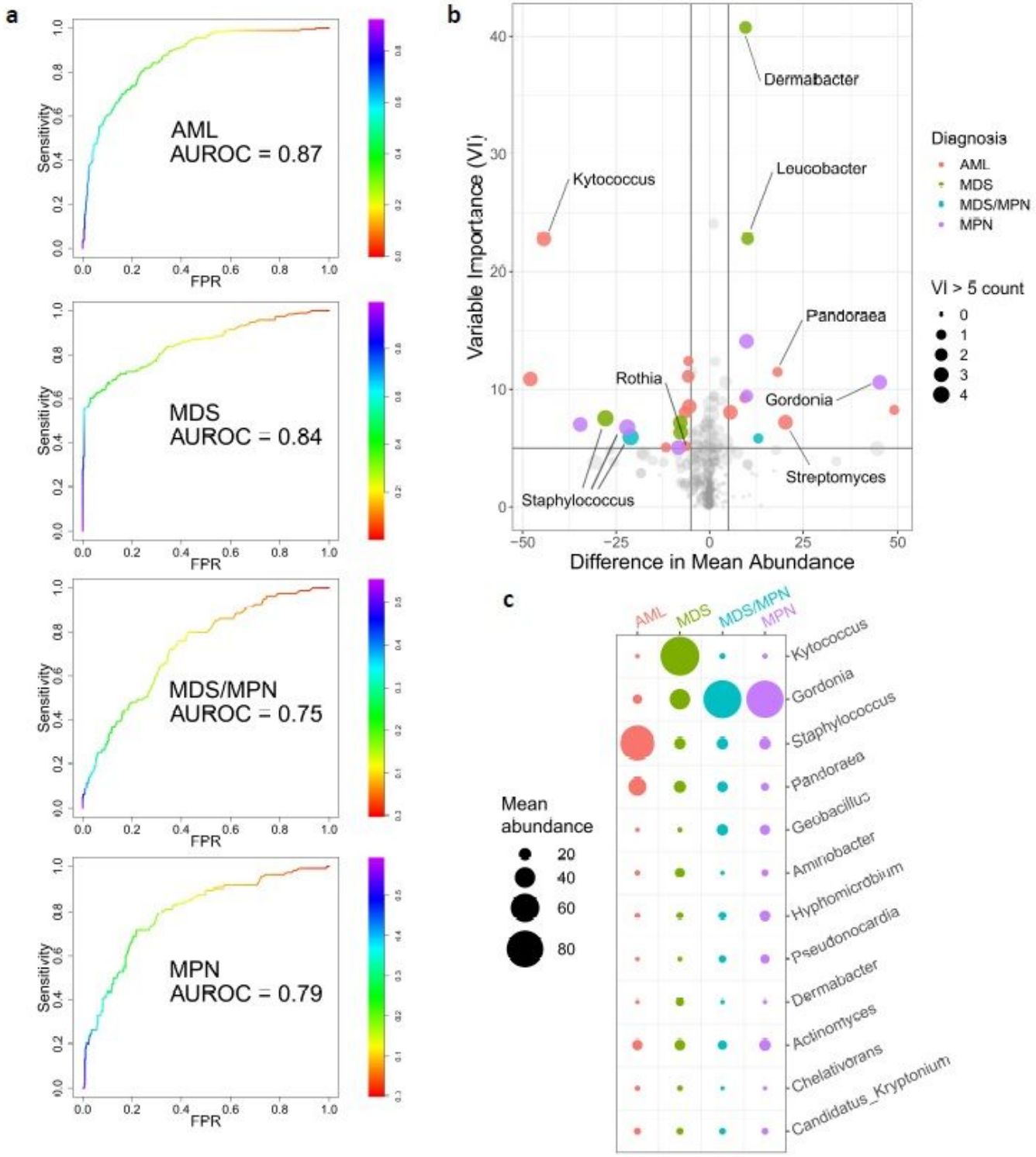

Figure 4

Bacterial composition differs among diagnoses. a) ROC curves showing, for each diagnosis, the performance on the test set (randomly-selected 30% of samples) of binary random forest classifier trained on the training set (remaining 70%). The AUROC values shown are averaged across 1000 random 70%/30% splits. b) Volcano plot showing enrichment/depletion of bacterial genera in specific diagnoses. Here horizontal axis indicates differences in mean abundance (diagnosis of interest – all others), and

variable importance is shown on the vertical axis. Point size indicates number of diagnoses (0 = smallest, 4 = largest) for which the corresponding genus has variable importance > 5. Points with mean abundance difference > 5 and variable importance > 5 are colored by corresponding diagnosis. Points of interest are labeled with their corresponding genera. (VI = variable importance). c) Mean abundances, in each diagnosis, of the genera that are among the top five in variable importance for at least one of the diagnoses. Circle size indicates the average abundance in the corresponding diagnosis.

Supplementary Files

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

- suppfigure1pdf.pdf
- suppfigure2pdf.pdf
- SuppTable1.pdf
- SuppTable2.xlsx