# The defensome of complex bacterial communities

Angelina Beavogui[1], Auriane Lacroix[1], Nicolas Wiart[2], Julie Poulain[1,3], Tom O. Delmont[1,3], Lucas Paoli[4,5], Patrick Wincker[1,3], Pedro H. Oliveira[1, #]

[1]Génomique Métabolique, Genoscope, Institut François Jacob, Commissariat à l'Energie Atomique (CEA), CNRS, Université Evry, Université Paris-Saclay, 2 Rue Gaston Crémieux, 91057 Evry, France.

[2]Genoscope, Institut François Jacob, CEA, Université Paris-Saclay, 2 Rue Gaston Crémieux, 91057 Evry, France.

[3]Research Federation for the Study of Global Ocean Systems Ecology and Evolution, FR2022 / Tara GOsee, Paris, France.

[4]Department of Biology, Institute of Microbiology and Swiss Institute of Bioinformatics, ETH Zürich, Zürich 8093, Switzerland

[5]Institut Pasteur, Université Paris Cité, INSERM U1284, Molecular Diversity of Microbes lab, Paris, France

[#] To whom correspondence should be addressed. Pedro H. Oliveira (pcoutool@genoscope.cns.fr)

Keywords: defense systems; phage-bacteria arms race; metagenome assembled genomes; defense islands; environmental defensomes

# Abstract

Bacteria have developed various defense mechanisms to avoid infection and killing in response to the fast evolution and turnover of viruses and other genetic parasites. Such pan-immune system (or *defensome*) encompasses a growing number of defense lines that include well-studied innate and adaptive systems such as restriction-modification, CRISPR-Cas and abortive infection, but also newly found ones whose mechanisms are still poorly understood. While the abundance and distribution of defense systems is well-known in complete and culturable genomes, there is a void in our understanding of their diversity and richness in complex microbial communities. Here we performed a large-scale in-depth analysis of the defensomes of 7,759 high-quality bacterial population genomes reconstructed from soil, marine, and human gut environments. We observed a wide variation in the frequency and nature of the defensome among large phyla, which correlated with lifestyle, genome size, habitat, and geographic background. The defensome's genetic mobility, its clustering in defense islands, and genetic variability was found to be system-specific and shaped by the bacterial environment. Hence, our results provide a detailed picture of the multiple immune barriers present in environmentally distinct bacterial communities and set the stage for subsequent identification of novel and ingenious strategies of diversification among uncultivated microbes.

# Introduction

Bacteria are under constant threat of infection by a variety of genetic parasites such as bacteriophages (henceforth called phages)[1]. As a result of this strong selective pressure, they have evolved multiple sophisticated defense mechanisms capable of regulating the flux of genetic information spread by mobile genetic elements (MGEs) via horizontal gene transfer (HGT)[2–4]. The complete set of a bacterial defense systems' repertoire can be designated as their *defensome*. Several bacterial defense systems have been discovered and extensively discussed in the literature, revealing two major groupings based on their components and modes of action: innate (non-specific) and adaptive immune systems[5,6]. Typical examples of innate immunity include prevention of phage adsorption[7], restriction-modification (R–M) systems that use methylation to recognize self from non-self-DNA[8], and abortive infection (Abi), in which the infected cell commits suicide before the invading phage can complete its replication cycle[9]. Recent efforts to *de-novo* identify microbial defense systems resulted in the discovery of several additional innate immune mechanisms with a wide range of genetic architectures[3,4], highlighting the strong selective pressure imposed by genetic parasites on microbial communities. Adaptive immune systems, on the other hand, are so far exclusively represented by clustered, regularly interspaced short palindromic repeats (CRISPR)-Cas, a family of defense systems that provides acquired immunity through the acquisition of short DNA sequences from MGEs that are incorporated into the host genome as spacers[10]. Large-scale efforts for defense system mapping have been recently propelled by the development of bioinformatic tools such as DefenseFinder[11] and PADLOC[12] that rely on a profuse collection of HMM profiles and specific decision rules for each known defense system. Such mapping has been mainly conducted in bacterial species from reference genome databases (e.g., NCBI RefSeq) that are known to overrepresent acute / common human pathogens and organisms that can largely be cultivated in laboratory[11–13]. While extremely insightful, such studies provide a limited snapshot of the bacterial defensome, as they miss the uncharted fraction of environmental microbial diversity that remains uncultured.

The current global Earth microbiome has been estimated at approximately $5\times10^{30}$ prokaryotic cells[14] scattered throughout a wide range of environments, including deep oceanic and continental subsurfaces, upper oceanic sediment, soil, and oceans as the most densely populated cases. In many environments, 99% of microbes are yet uncultured[15], while cultured representatives belong overwhelmingly to the phyla Actinobacteria, Bacteroidetes, Firmicutes, and Proteobacteria. For nearly 4 billion years, bacteriophages have co-evolved with bacteria, with recent estimates pointing to the presence of ~$10^{31}$ viral particles in the biosphere[16], and up to $10^{23}$ infection events per second taking place just in the global ocean[17].

During the last decade, extensive progress in high-throughput sequencing technologies and computational methods enabled culture-independent genome-resolved metagenomics to recover draft or complete metagenome-assembled genomes (MAGs)[18–20]. The latter have advanced our understanding on the diversity, abundance, and functional potential of microbiota and phageome composition and corresponding ratios across different environments. A healthy adult human gut for example, is a reservoir for ~$4 \times 10^{13}$ bacterial cells (mostly Firmicutes and Bacteroidetes)[21], and low ($10^{-3}$-1) virus-to-prokaryote ratios (VPRs)[22]. In contrast, marine ecosystems typically show larger VPRs (between $8 \times 10^{-3}$-$2.15 \times 10^{3}$, mean of 21.9), followed by soil environments which show the largest ratios (between $2 \times 10^{-3}$ - $8.2 \times 10^{3}$, mean of 704) (reviewed in [23]). We hypothesize that the strong VPR dynamics across temporal and spatial scales is likely to profoundly shape the defensome arsenal across biomes.

In this study, we conducted a large-scale in-depth investigation on the abundance, distribution, and diversity of the defensome in complex bacterial communities from three key environments: soil, marine, and the human gut. We tested the association between defensome and different mechanisms of genetic mobility, the former's colocalization in defense islands, and assessed the mutational landscape of high-frequency single nucleotide polymorphisms (SNPs) and insertion-deletions (indels) across defense gene families. These results provide a unique view of the interplay between microbial communities and their phage invaders, and will pave the way to the identification of hitherto unknown defense systems and

1 / or other phage-resistance mechanisms across the enormous diversity of yet-uncultivated
2 microbial populations.

# Results

## Abundance and distribution of defensomes in bacterial MAGs

We performed a defensome mapping across a large dataset of 7,759 high-quality (≥ 90% completeness, ≤ 5% contamination/redundancy, see Methods) soil, marine, and human gut MAGs[24–26] (**Fig. 1a**, **Supplementary Tables 1-4**, **Supplementary Fig. 1**). For this purpose, we used the DefenseFinder pipeline[11], which relies on a comprehensive collection of hidden Markov models (HMM) protein families and genetic organization rules targeting all major defense system families described in the literature (**Fig. 1b**).

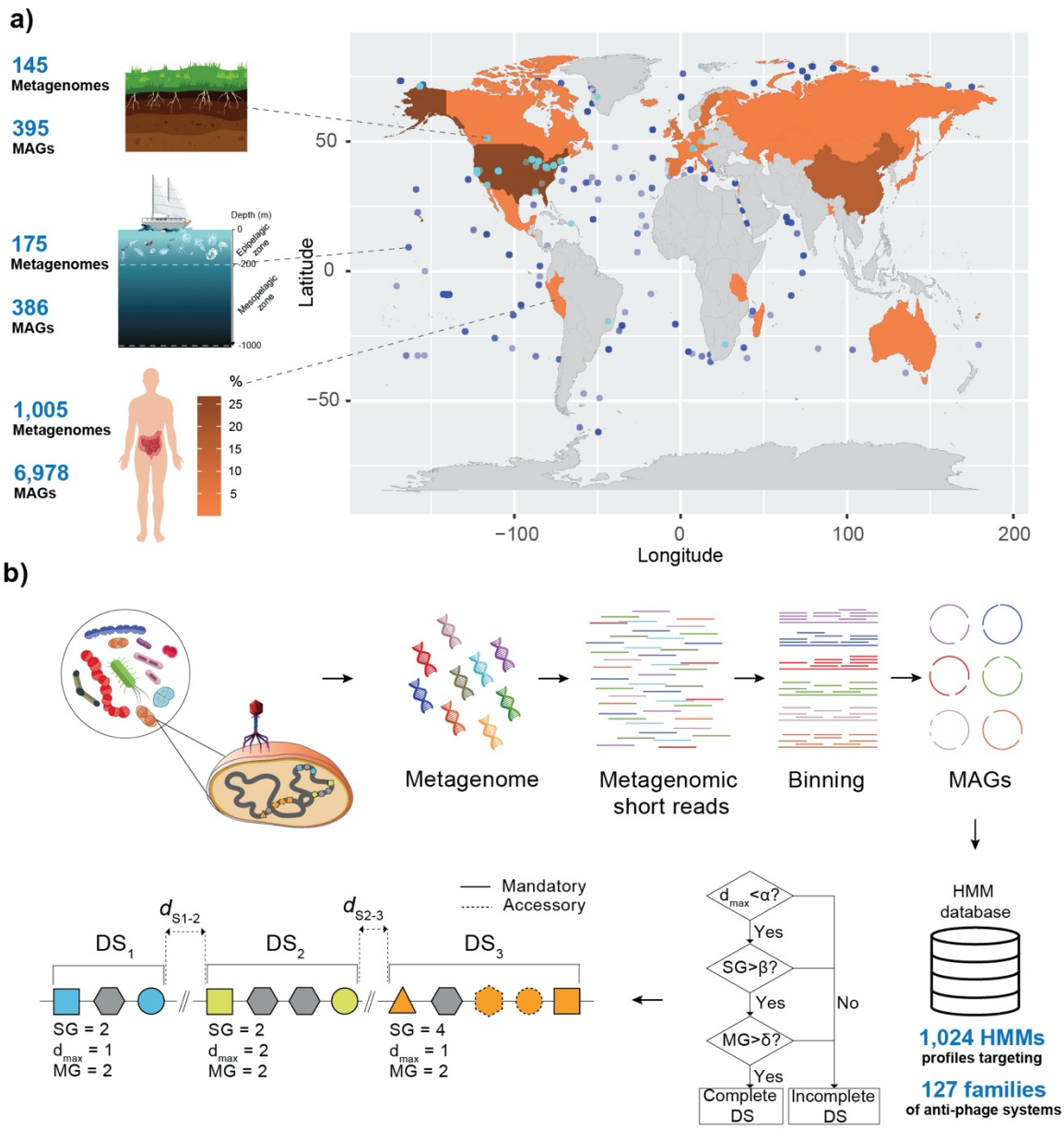


**Fig. 1**. Defensome analysis. (*a*) Our analyses focused on 7,759 high-quality near-complete MAGs recovered from three distinct ecosystems: soil, marine, and human gut[24–26]. The geographical distribution of soil and marine sample collection sites is shown in the world map, as well as the percentage of human gut samples recovered from each country (shown as colored heatmap). Our dataset includes at least 385 Genera (corresponding to a total of 7,593 MAGs) and 25 Classes (corresponding to a total of 93 MAGs) not previously covered in a recent study focusing on the defensome of the NCBI RefSeq prokaryotic database[11]. (*b*) A collection of 1,024 HMM profiles targeting 127 families of anti-MGE defense systems from DefenseFinder, was used to query the entire MAG dataset. Briefly, this was performed by means of genetic organization rules allowing for two types of genetic components: "mandatory" and "accessory" (as described previously[11]). Given the wide diversity of genetic organization of anti-MGE systems, rules were written differently for different types of systems. Shown is an example of a genomic region containing three defense systems (DS1-DS3), respectively characterized by a total sum of genes (SG), a maximum distance between defense genes ($d_{max}$), and a given number of mandatory genes (MG), which will allow disentangling between complete or incomplete defense systems based on established thresholds ($\alpha$, $\beta$, $\delta$).

Throughout this manuscript we will refer to *complete* anti-MGE defense systems as those whose currently described genetic organization has been experimentally shown to confer anti-MGE activity. Such concept of defense system *completeness* is expected to evolve in the future (in particular for the recently described cryptic large multigenic systems), as more details will emerge regarding their functional intra-operability. In the case of defense genes, they can either belong to complete defense systems, or classify as *solitary*, i.e., those often shuttled by HGT or arising from genetic erosion of complete defense systems. Of note, the solitary nature of defense genes does not necessarily preclude its functional activity or even implication in anti-MGE defense roles, as it has been previously shown for solitary bacterial methyltransferases (MTases)[27].

In this study we found 43,263 defense systems and 764,507 defense genes pertaining to a total of 70 defense families across our full MAG dataset (**Supplementary Tables 3**, **4**). The relative distribution of defense systems differed considerably across environments, with R–M, CRISPR-Cas and the SoFIC AMPylase being the most predominant (**Fig. 2a**). When the distribution of total defense genes was represented instead, we observed multiple solitary genes / incomplete systems (e.g., Gabija, Gao_Qat / Gao_Mza, or Dodola) consistently present across most MAGs (**Supplementary Fig. 2**). The latter suggests either non-defensive roles or genetic erosion of complete systems similarly to previous observations in complete genomes[13,27]. While defense system distribution across soil and human gut MAGs followed a

typical binomial distribution (with most genomes encoding between 3-4 defense systems), that observed in genomes from marine environments was geometric-like, with most MAGs (~65%) showing a limited defensome (**Fig. 2b**). Such observations are in agreement with recent observations describing a $10^3$ times lower effective rate of HGT in marine bacteria compared with gut bacteria, with soil bacteria occupying an intermediate position between the former two[28].

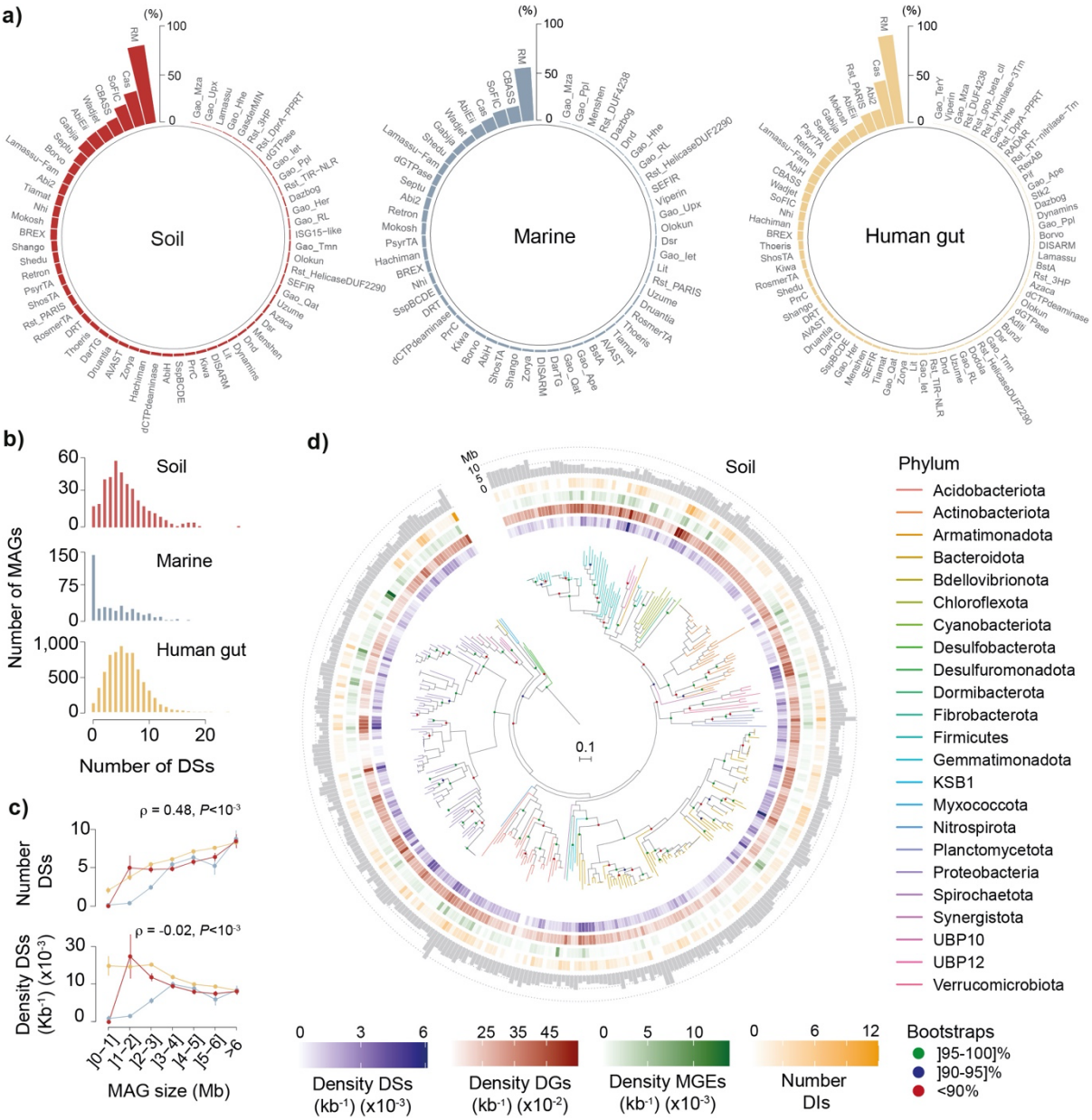



**Fig. 2**. Abundance and distribution of defense systems in MAGs. (*a*) Percentage of soil, marine and human gut MAGs harboring each family of defense system. (*b*) Distribution of number of defense systems (DSs, per MAG) across environments. (*c*) Variation of number and density (per MAG and per kb) of defense systems (DSs) with MAG size (Mb) for each biome. Error bars represent standard deviations of the mean. (*d*) Phylogenetic

representation of 373 soil MAGs, their corresponding phyla, density (per kb) of defense systems (DSs, purple), defense genes (DGs, red), MGEs (green), and number of defense islands (DIs, yellow). Distribution of MAG sizes (Mb) are shown as outer layer barplots. All data corresponds to analyses performed in assemblies with values of $N_{50} \geq 100$ kb.

Similarly to what has been described for R–M systems[13], we observed positive correlations between the total number of defense systems and MAG size and concomitant negative correlations between the density of defense systems and size (**Fig. 2c**). Such trends can be explained by the fact that bacteria with larger genomes typically engage in more HGT[13,2], thus requiring a more abundant and diverse defensive arsenal. No qualitative differences were observed when the analyses shown in **Figs. 2a-c** were performed using MAG assemblies having values of $N_{50} \geq 200$ and 300 kb to control for the effect of contiguity (**Supplementary Fig. 3**).

The density of defense systems (per MAG and per kb) differed widely among clades, from none (largely in intracellular bacteria and obligatory endosymbionts) to more than $8 \times 10^{-3}$ in *Phascolarctobacterium* (human gut) and ~$1.5 \times 10^{-2}$ in *Elsteraceae* (soil) and UBA9040 (marine) environments (**Fig. 2d**, **Supplementary Fig. 4a**, **Supplementary Tables 3, 5**). No MAG was entirely devoid of defense genes, with maximum densities (per MAG and per kb) ~$8.5 \times 10^{-2}$ across the different biomes (**Fig. 2d**, **Supplementary Fig. 4a**, **Supplementary Table 3**). When defense systems were split according to its mechanism of action, R–M, Abi, and potential Abi systems were the most prevalent across biomes (**Supplementary Fig. 4b**, **Supplementary Table 6**), similarly to recent observations[29].

Apart from MAG size, the abundance of defense genes was expected to depend on phylogenetic depth, as deeper lineages accumulate more events of HGT exchanges presumably leading to defensome buildup. We ran stepwise linear regression analyses to assess the role of these variables in explaining the variance of the defensome (**Supplementary Table 7**). These showed that MAG size had the strongest direct effect on defensome abundance, and that phylogenetic depth had a significant but less important explanatory role.

We found in our dataset multiple occurrences of ligand binding WYL domains and protein interaction CARD-like domains (**Supplementary Figs. 4c-e**), with previously demonstrated regulatory activity of phage defense systems, namely BREX, CRISPR-Cas, CBASS and gasdermins[30–33]. Interestingly, we found here a significant colocalization between these domains and multiple defense genes belonging to additional families involved in regulated cell death, such as Lamassu, RosmerTA, and Rst_PARIS. Very few WYL and CARD-like domains were found in genes from marine MAGs (< 0.75% of the dataset), in agreement with the latter's more limited defensome. The patterns of colocalization differed across genomes recovered from the soil and human gut environments (**Supplementary Figs. 4d-e)**. For example, WYL preferentially colocalized with CBASS and RosmerTA, respectively in soil and human gut environments. We also found in the Bacteroidetes bacterium UBA1952, instances of an operon with some similarity to the recently described *Pedobacter rhizosphaerae* CARD-encoding defense system[33]. In particular, UBA1952 codes for a VapC-like nuclease of the PIN domain superfamily presumably operating as effector, and the SMC-like RecN with ATPase function (**Supplementary Fig. 4c**).

Hence, bacterial MAGs possess a diverse repertoire of defense systems (being defense genes essentially ubiquitous), and the patterns of their distribution are very diverse and dependent on genome size and taxonomy. Moreover, defense genes pertaining to systems typically implicated in regulated cell death mechanisms preferentially colocalize with WYL and CARD-like domains and change according to environment.

## The interplay between defensome repertoire and bacterial biogeography

Fluctuations in microbial community composition are a function of a large ensemble of diverse biotic and abiotic drivers. Factors such as pH (and other physicochemical parameters), temperature, nutrient availability, or pollution can fundamentally reshape the spatiotemporal dynamics of soil / marine bacterial and viral communities[34–37]. In parallel, multiple variables such as host lifestyle, nutritional needs, genetics, age, medication, urbanization, and the

impact of westernization are known to significantly impact the human gut microbiome and virome[38,39]. Concurrent with this dynamic interplay between environmental filtering and phage-bacteria antagonistic and/or mutualistic coevolutionary interactions, one expects concomitant changes in defensome composition. This prompted us to examine how the defensome's abundance and diversity correlated with bacterial biogeography. The top five most represented Classes in our dataset for each environment are Gammaproteobacteria (soil, marine, human gut), Alphaproteobacteria (soil, marine), Dehalococcoidia (marine), Bacteroidia (soil, human gut), and Clostridia (human gut) (**Supplementary Table 2**). Such different patterns in species richness, and relative phylogenetic diversity across environments, are expected to impact genetic flux and concomitantly, defensome profiles.

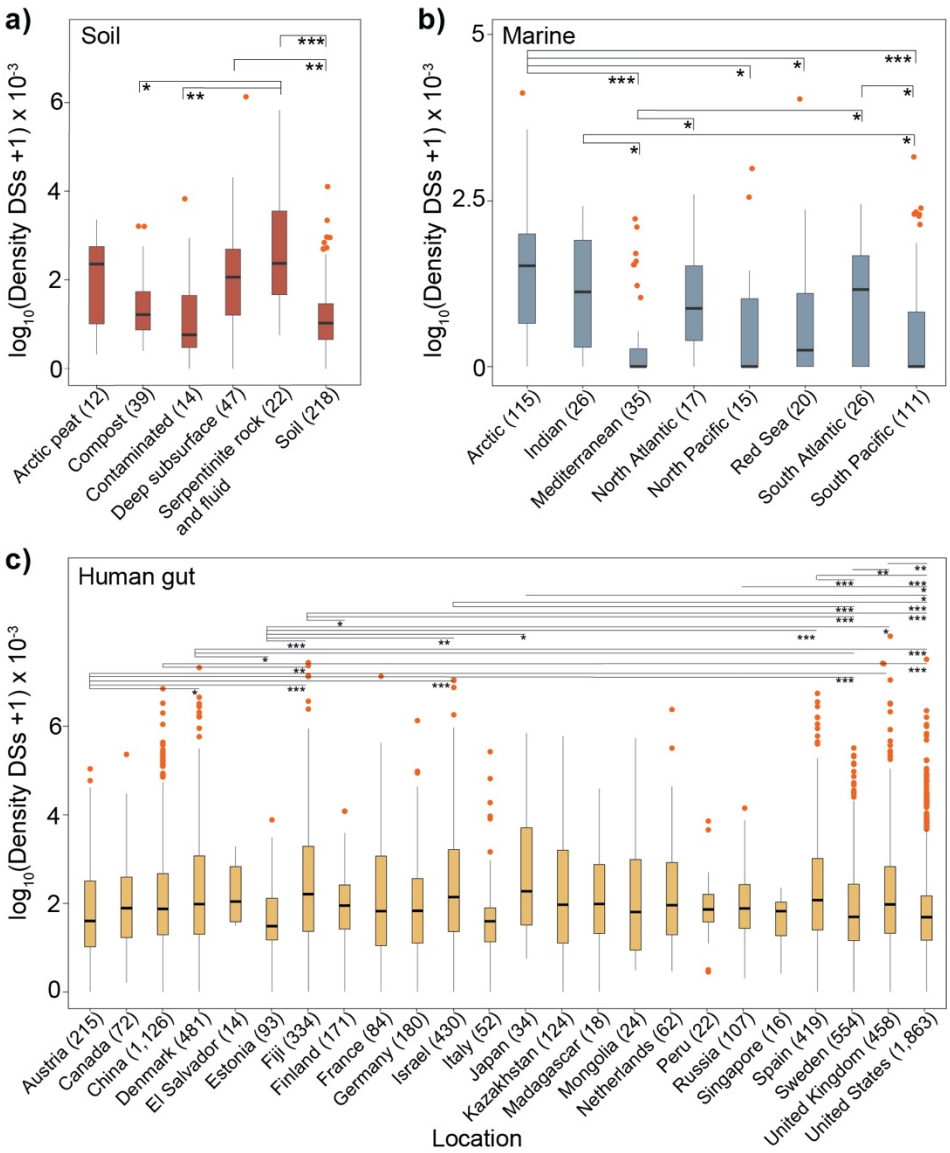


**Fig. 3**. Defensome variation across different ecological and geographical backgrounds. Defense system (DS) density (per MAG per kb) across distinct ecological (soil, marine) (*a*, *b*) and geographical (human gut) (*c*) contexts. Error bars represent standard deviations and Mann–Whitney–Wilcoxon test *P* values are indicated. $^{*}P < 0.05$; $^{**}P < 10^{-2}$; $^{***}P < 10^{-3}$. Number of MAGs analyzed are shown in parentheses.

In soil environments, the highest and lowest densities of defense systems were respectively observed in MAGs recovered from serpentine-hosted ecosystems and contaminated or regular soils (**Fig. 3a**). These observations are consistent with the fact that serpentine environments are among the most challenging niches on Earth, characterized by low cellular abundances, limited microbial diversity, high VPRs[40,41], and consequently, the likely need for additional anti-MGE systems. Conversely, contaminated and regular surface soils impose a type of environmental stress (namely chemical and UV radiation) that is expected to push phage–bacterium interaction from parasitism to mutualism[42–44]. The latter should provide bacterial hosts with diversified competitiveness and environmental adaptability while allowing prophages to avoid direct exposure to the stressor. Interestingly, while R–M, CRISPR-Cas and SoFIC were prevalent in MAGs recovered from almost all types of soil, arctic peat (richer in Bacteroidales) stands out as an outlier with a high abundance (≥ 50%) of AbiEii and BREX (**Supplementary Fig. 5a**, **Supplementary Table 8**). While it remains unclear which processes drive the overrepresentation of these particular defense families in MAGs recovered from arctic peat, the latter could be explained by the cell's need for a second layer of resistance under conditions of high VPRs (see below), or eventually to enforce cooperation between individuals, or even with MGEs[45,46].

In marine MAGs we observed prevalence of R–Ms, but also of the abortive infection system CBASS and the less-known standalone SoFIC. The highest defense system densities were found in MAGs originating from the arctic ocean (**Fig. 3b**). Such increased defensive repertoire fits previous observations describing high VPRs and virus-to-bacteria contact rates in sea ice brine compared to seawater[47,48]. Following our observations for ice peat soil, we also found a particularly high abundance (~28%) of the AbiEii system in arctic ocean MAGs (**Supplementary Fig. 5b**, **Supplementary Table 8**). The overall low defensome abundance

and diversity in the Mediterranean Sea can be due to the latter's conditions of seasonal oligotrophic conditions, higher temperature (>13°C), and lower concentrations of inorganic nutrients N and P compared to waters of similar depth in open oceans, leading to very low VPRs[49].

To what concerns human gut MAGs, the difference in amplitude in defense system densities across different countries is more subtle (albeit often significant) and harder-to-interpret compared to other environments. While there is a strong trend in the literature supporting a gradual reduction in microbial diversity (and subsequent disruption of metavirome profiles) concomitant with westernization[50], the latter did not translate into a clear cut geographical trend in regards to the defensome (**Fig. 3c**, **Supplementary Fig. 5c**).

When defense systems were split according to its mechanism of action, their variation in density across distinct ecological and geographical backgrounds was kept qualitatively the same, at least for the most abundant mechanisms (R–M, Abi, and potential Abi systems) (**Supplementary Fig. 6**).

Hence, not only the microbiome but also its defensome is dramatically shaped by different ecological and geographical constraints. Higher densities of defense systems were found in MAGs recovered from particularly challenging biomes such as serpentine soils or the arctic itself, in line with the high VPRs described in such environments.

## The genetic mobility of bacterial MAG defensomes

Cellular defense genes typically propagate by HGT, in a process frequently mediated by MGEs. Physical colocalization between defense genes and MGEs allows for an efficient strategy to modulate and / or resolve potential conflicts in the interactions between the host and the MGE itself. In this context, a growing number of MGE-encoded defense systems or defense genes have been described in several bacteria, particularly involving the most well studied ones (R–Ms, Abi, CRISPR-Cas) and major families of MGEs (phages, plasmids, integrons, ICEs / IMEs)[13,51,52]. Yet, there is a paucity of data on the genetic mobility of the

defensome in complex bacterial environmental communities. We consistently observed more defense genes in MGEs than in chromosomes (excluding MGEs), irrespective of the environment (**Fig. 4a**, **Supplementary Fig. 7a**, **Supplementary Tables 9, 10**). This is in line with current evidence that MGE-encoded defense systems protect their host cells as a side-effect of their action to protect the MGE from other MGEs[51]. When MGEs were split according to family (excluding integrons which are rare in the human gut microbiota[53]), there was a slight trend for higher colocalization of defense genes with ICEs / IMEs irrespective of the environment (**Fig. 4b**, **Methods**), in agreement with recent observations[52]. When integrons were included for comparison, they showed the highest colocalization densities with defense genes in the human gut (**Supplementary Fig. 7b**), a result that should be taken cautiously given its low statistical power.

A further split of defense genes according to their corresponding family, allowed us to evaluate the former's over- or underrepresentation across MGE classes (**Fig. 4c**, **Supplementary Fig. 7c**). The results put into evidence a few curious aspects of defensome mobility. The first is that irrespectively of the environment, plasmids generally carry a higher than expected by random chance number of defense genes across a large breadth of defense families when compared to other MGE classes. This observation aligns with the fact that plasmids typically allow for a high genetic plasticity and can sustain large gene exchange networks throughout phylogenetically diverse communities[54].

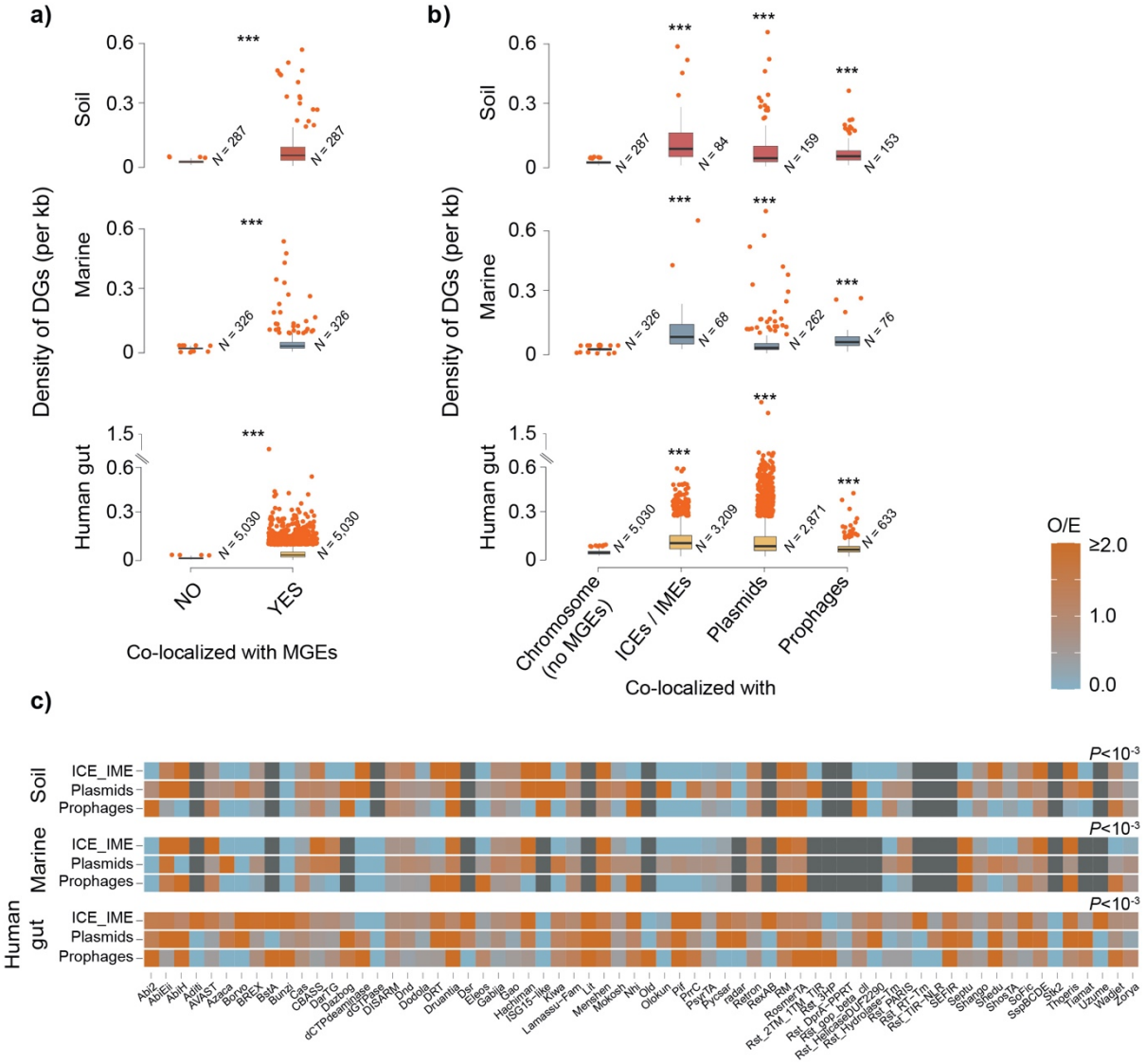



**Fig. 4**. The genetic mobility of the defensome. (*a*) Box plots of the genomic colocalization of defense genes and MGEs. Number of MAGs analyzed are shown as *N* values. (*b*) Box plots of the genomic colocalization of defense genes with plasmids, prophages, and ICEs / IMEs. Error bars represent standard deviations and Mann–Whitney–Wilcoxon test *P* values are indicated. (*c*) Heatmap of observed / expected (O / E) ratios of colocalization between genes belonging to distinct defense families and MGEs. Expected values were obtained by multiplying the total number of genes pertaining to a given defense family by the fraction of defense genes of that family assigned to each MGE. $^{*}P < 0.05$; $^{**}P < 10^{-2}$; $^{***}P < 10^{-3}$, $\chi^2$-test. Dark gray squares represent absence of colocalization.

The second aspect relates to the highly heterogeneous landscape of combinations defense family / MGE class across multiple environments. This reflects the dynamic interplay between a multitude of parameters, including the density and phylogenetic composition of host cells and MGEs present in the community, habitat structure, and environmental pressures. These results also suggest that certain defense genes / systems favor different classes of MGEs for

their shuttling, in a likely dynamic and multilayered interplay with shifting allegiances. Overall, these data shows that a wide range of defense families is carried by MGEs presumably favoring their selfish spread, and that different associations defense family / MGE class are favored across distinct biomes.

## Encoded functional potential of defense islands and defensome colocalization

Defense genes are typically carried in MGEs by HGT. The former may allow the MGE to be kept in the host by promoting addiction, but on the reverse side of the coin, may carry beneficial traits capable of positive epistatic interactions with the resident host functions. To conciliate these two scenarios, defense genes tend to cluster in so-called defense islands, i.e., high-turnover sinks of genetic diversity, that may serve as catalysts of novel defensive strategies. Therefore, we queried the abundance of such islands and their content. We found 12,890 defense islands in 6,217 MAGs (**Supplementary Fig. 8a**, **Supplementary Table 11a**, **Methods**), with a similar size distribution across environments (median ~ 17 genes) (**Fig. 5a**), suggesting that there is an optimal size range for these defense sinks. Defense gene density in defense islands was significantly lower in marine environments, followed by soil and human gut (**Fig. 5b**). The latter is in line with the above observations on a limited defensome in marine MAGs when compared with other environments. Defense islands' anti-MGE content was very diverse (**Fig. 5c**), with several defense families being overrepresented compared to regions outside defense islands (e.g., Hachiman, R–M, Thoeris) while others being underrepresented (e.g., PsyrTA, ShosTA, Zorya) (**Supplementary Fig. 8b**).

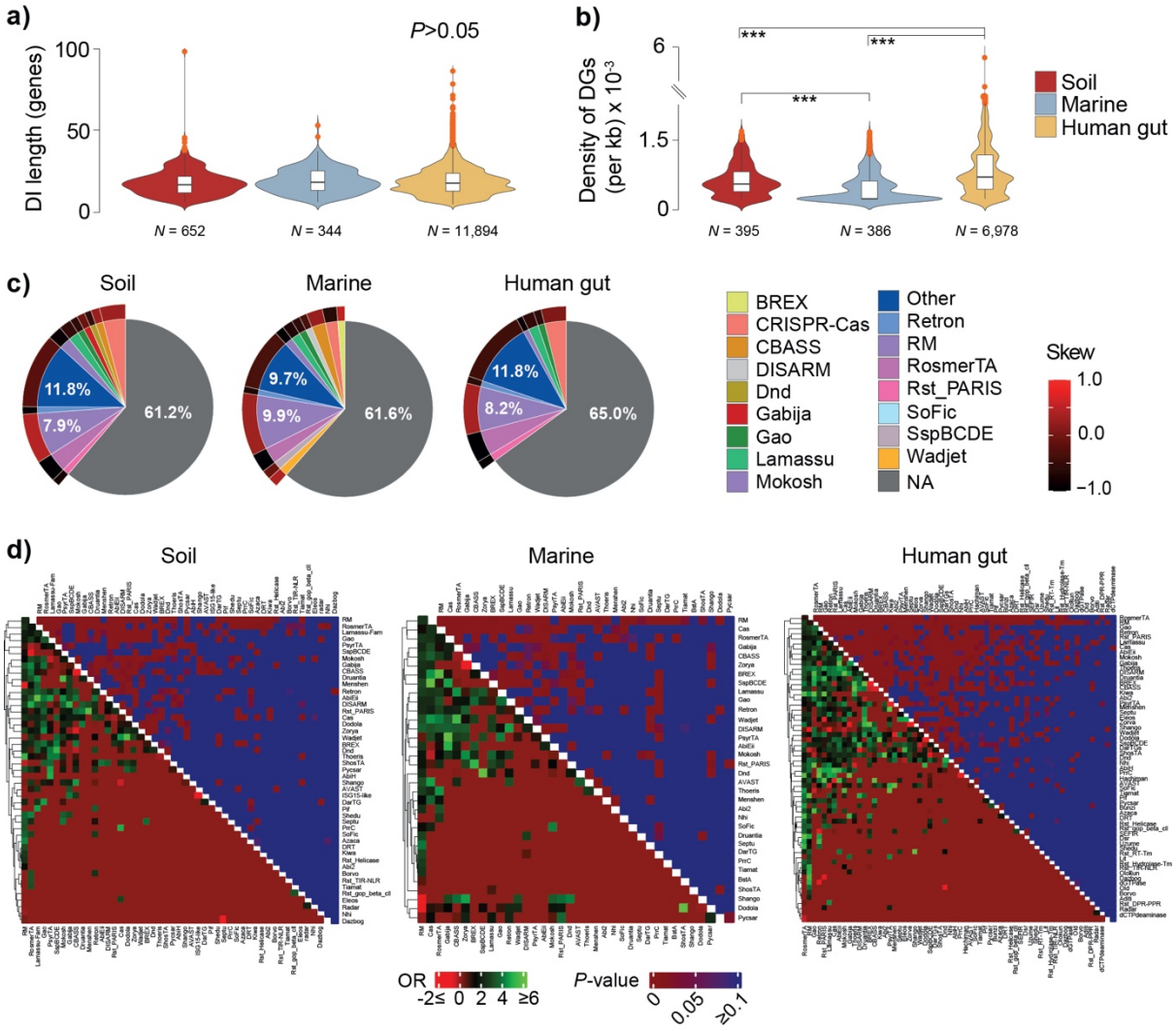



**Fig. 5**. The MAG defense island repertoire. (*a*) Defense island (DI) length distribution (given in genes) in soil, marine, and human gut MAGs. (*b*) Density distribution of defense genes (DGs) (per MAG per kb) across each environment. (*c*) Pie-plots of the relative abundance (%) of gene content in defense islands. Colored slices correspond to defense genes (those with a relative abundance < 1% were merged as 'Other'), and gray slices (NA) correspond to genes not classified as defensive by DefenseFinder. The outer layer corresponds to the skew ratio between genes belonging to complete and incomplete systems given by $\frac{\#genes\ belonging\ to\ complete\ systems - \#genes\ belonging\ to\ incomplete\ systems}{\#genes\ belonging\ to\ complete\ systems + \#genes\ belonging\ to\ incomplete\ systems}$. (*d*) Defense families' odds ratio (OR) of colocalization in defense islands (bottom heatmaps) and associated Fisher's exact test *P* value (upper heatmaps) for the three environments. To eliminate the confounding (inflating) effect of colocalized genes pertaining to the same system, we only considered solitary genes or those pertaining to independent defense systems distanced of 5 genes or less.

This bias for certain defense families to locate in defense islands, suggests either positive epistatic interactions with vicinal genes, or a preferential shuttling by a particular family of MGEs. Despite its diversity, defense families are largely similar across environments and

skewed towards incomplete systems, pointing as expected, towards a high gene turnover at defense islands (**Fig. 5c**). Interestingly, the large majority (~63%) of defense islands' gene content was not predicted to have a defensive role. A COG classification of such 'non-defensive' genes revealed a high prevalence of functions linked to replication / recombination / repair and transcription (**Supplementary Fig. 8c**). The latter can be at least partially explained by the fact that defense genes are often shuttled by MGEs, which rely on such functions for target selectivity, insertion, and excision. The above COG categories and the most abundant defense families (R–M and CRISPR-Cas for soil / human gut; R–M, CBASS and RosmerTA for marine biomes) remained unchanged even when considering defense systems (instead of genes) as the main counting unit in the definition of defense islands (see **Methods**) (**Supplementary Fig. 8d**, **Supplementary Table 11b**).

Since MGEs have different distribution patterns, we quantified the frequency of colocalization of defensome families (≤ 5 genes apart) in defense islands compared to regions outside the latter (**Fig. 5d**, **Supplementary Table 12**, **Methods**). In line with their abundance, frequent shuttling by MGEs and defensive role, R–Ms significantly colocalized with most other defense families in defense islands irrespectively of the environment. Inversely, R–Ms showed a preference to colocalize with genes pertaining to Menshen, Shango and Dodola families outside defense islands. Interestingly, and despite their general underrepresentation in defense islands (**Supplementary Fig. 8b**), genes pertaining to families such as PsyrTA and Zorya showed significant colocalization with other defense families inside defense islands. Conversely, defense families significantly overrepresented in defense islands (e.g., Hachiman) (**Supplementary Fig. 8b**), rarely colocalized with other families. Upon splitting our dataset according to biogeographical zones, and despite the subsequent decrease in statistical power, the colocalization trends of the most abundant defense families still hold qualitatively (**Supplementary Fig. 9**, **Supplementary Table 13**). These observations point for the possibility of previously unappreciated epistatic interactions between selected families of defense genes / systems in defense islands.

Hence, we found approximately 11% of the defensome concentrated in defense islands, an environment-dependent highly heterogeneous distribution of defense families in such regions, a large proportion of 'non-defensive' functions, and a significant colocalization of a subset of families of defense.

## The genetic variability of the defensome

The coevolutionary dynamics between defenses and counter-defenses contributes to an endless process of genetic diversification and evolution of sequence specificity, that can take place through point mutations, recombination, gene duplications, replication slippage, or transposition[55]. Such panoply of diversification processes has been particularly studied in well-described innate immune systems like R–Ms, and can take the form of, for example, target recognition domain swapping in Type I *hsdS* subunits, or phase variability of Type III *modH* genes. However, there is a void in our current understanding on the extent to which differences in selection strength act across distinct defensome gene families. To this end, we performed metagenome read recruitment over defensome genes, assessed the frequency and type of short variants found, and used this information to pinpoint consistently fast or slow evolving genes across environments (see **Methods** for further details).

We observed multiple defense gene families with higher-than-expected values of SNP + indel density across multiple biomes (**Fig. 6a,** only defense families for which at least one defense gene showing a O / E ratio ≥ 1.5 per environment are shown, **Supplementary Table 14**). Genes such as *dolB*, *mzaA*, and *sspH* were among this 'high-mutation frequency' subset irrespective of the environment, while others like *druA*, *zorA*, or *letA* were environment-specific. The results were qualitatively similar when all defense families were included (**Supplementary Fig. 10**). The range of SNP + indel densities differed considerably across defense gene families (**Fig. 6b**). Mutation types were also profoundly affected by the environment (and thus by population structure). For example, indels and non-synonymous SNPs were consistently more abundant in marine than in soil MAGs, even when comparing

across same defense families (e.g., *dolB*, *mzaA* or *sspH*) (**Fig. 6b**). While most variants found was intragenic, *sspH* and particularly *zorA* had as much as 25% of variants located in the first 200 bp upstream the annotated start codons, suggesting potential regulatory effects. The rapid turnover of defense gene repertoires in bacteria, many of which in MGEs, can be followed by selection for the former's conservation or loss in a cell. To investigate the action of natural selection on the defensome gene families showing the highest frequency of variants, we computed the ratio of nonsynonymous over synonymous substitution rates (dN/dS) for the pools of orthologous defense genes within our MAG dataset. Similar to previous observations for CRISPR-Cas and R–M gene families[56,13], all defense genes analyzed were found to be under strong purifying selection (dN/dS<<1; **Supplementary Fig. 11a**, **Supplementary Table 15**). The preferential purge of nonsynonymous mutations by natural selection contributes to maintain the defensive functions of these genes and can be reconciled with a scenario of high turnover, if the selection pressure on the system fluctuates in time, i.e., if these genes alternate periods of strong purifying selection and periods of relaxed selection (e.g., as a result of competition with other defense systems, or during strong selection for HGT). Interestingly, despite their overall negative selection, we observed relatively high levels of divergence and positive selection in certain portions of their sequences (**Supplementary Fig. 11b**). The latter matched, for example, PFAM domains with predicted AAA+ ATPase activity (PF07724 / PF10431 in DolB, and to a less extent PF00004 / PF17862 in IetA), an *ftsH*-like extracellular domain (PF06480 in IetA), and a Sigma70-like non-essential domain (PF04546 in MzaA).

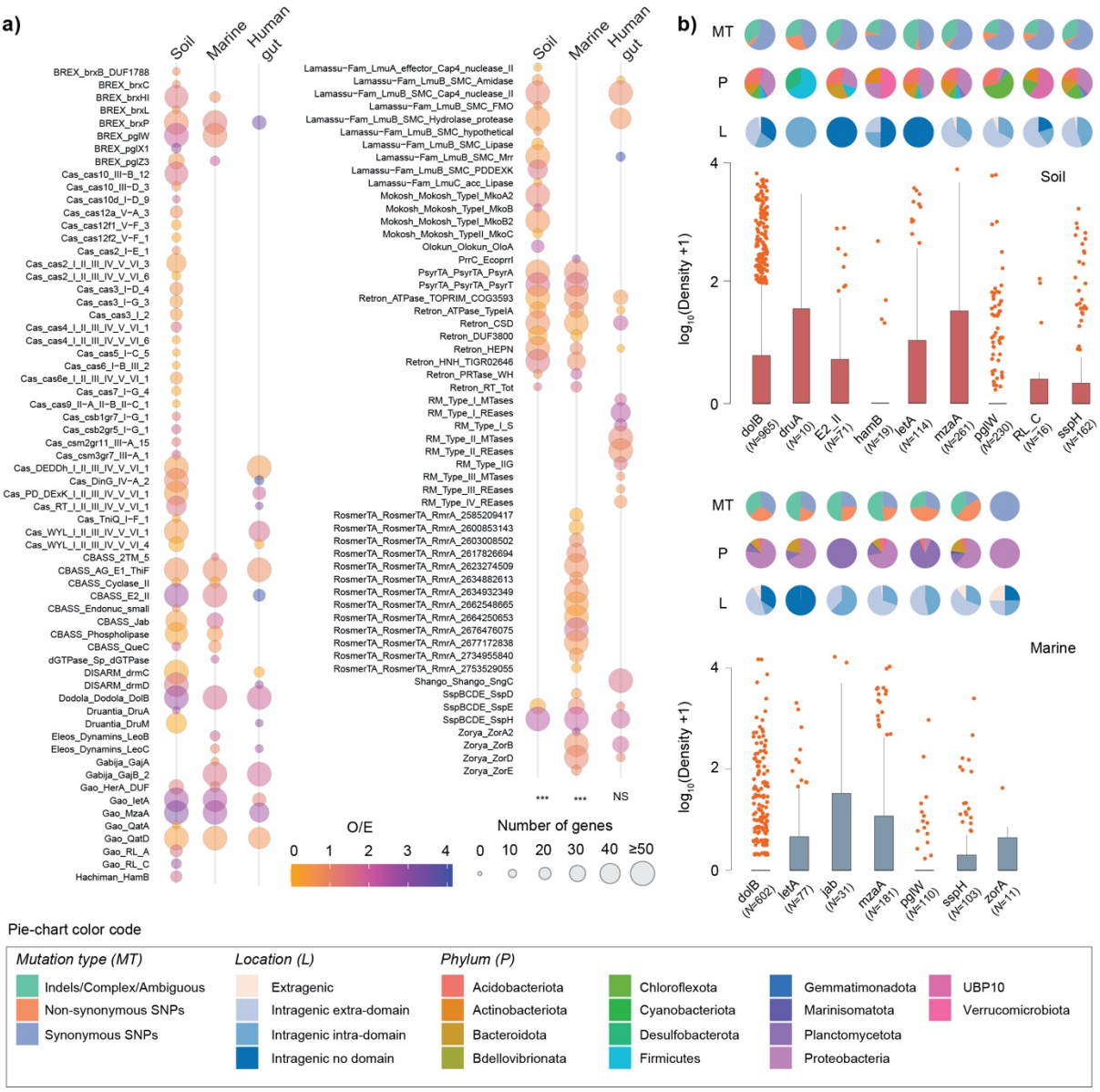



**Fig. 6**. The genetic variability of the defensome. (*a*) 90 metagenomes (30 for each environment) having a broad representativity in terms of sampling sites (soil and marine) and countries (human gut), as well as in terms of presence of most defense families previously identified by DefenseFinder were selected. Shown in circles are the observed / expected (O / E) ratios of number of defense gene families harboring high-frequency SNPs + indels (≥25% at the variant position) in their gene body (including 200 bp upstream the start codon). Expected values were obtained by multiplying the total number of genes pertaining to a given defense family by the fraction of defense genes of that family harboring high-frequency alleles. Circle radius corresponds to the total number of defense genes analyzed per family. To ease visualization, we limited the figure to only defense families for which at least one defense gene showed an O / E ratio ≥ 1.5 per environment. The complete representation is shown in **Supplementary Fig. 10**. *$P < 0.05$; **$P < 10^{-2}$; ***$P < 10^{-3}$, $\chi^2$-test. NS: not significant. (*b*) Density distribution of SNPs + indels for a selection of defense gene families showing the highest O / E values in (*a*). Information on mutation type, location and phylum are indicated in pie-plots. The number of genes analyzed is shown in parentheses.

# Discussion

In this study, we present a large-scale analysis of the abundance and diversity of defensomes of genomes/species from complex microbial communities and three representative biomes: soil, marine and human gut. Our results on the quantification of the defensome in marine MAGs lend support to a scenario of a limited defense arsenal in this environment (**Fig. 2b**). The latter can be accounted by a variety of potential explanations namely: *i*) the fact that oligotrophic open oceans typically show an overrepresentation of clades characterized by heavily streamlined genomes[57] (e.g., Dadabacteria, Chloroflexota) (**Supplementary Fig 1**, **Supplementary Table 2**), and thus, more likely to opt for more transient defense systems and little metabolic plasticity to better cope with the limiting environment of the surface ocean; *ii*) the dominantly planktonic lifestyle and low cell-density in the marine environment (at least for the free-living fractions accounted for in our MAG dataset) which may in itself, or through a reduced frequency of HGT, contribute to a more limited anti-MGE arsenal; *iii*) the fact that the large majority of HMMs currently available to detect defense systems were initially developed on the basis of genetic data that overrepresents not only cultivable bacteria, but also lineages (e.g., *Escherichia*, *Bacillus*, *Pseudomonas*) that are more distantly related to those that make up the global ocean microbiome (**Supplementary Table 3**). On a broader view, our results qualitatively match those recently obtained for RefSeq genomes in terms of the most abundant systems (R–Ms, CRISPR-Cas) and overall diversity of families identified[11]. The enhanced granularity offered by our cross-environment comparison revealed a few curious differences at the level of preferential 'second line' defense families. One of such differences concerns SoFIC and CBASS which are present in roughly 20% of soil and marine MAGs (mainly in Chitinophagales and Caulobacterales), but considerably less predominant (~8%) in human gut MAGs (mainly in Verrucomicrobiales, Enterobacterales, and Bacteroidales) (**Fig. 2a**). Inversely, the abortive infection system Rst_PARIS is present in 20% of human gut MAGs (mainly in Bacteroidales) but is virtually absent in soil or marine environments (**Fig. 2a**). While R–Ms (and to a lesser extent CRISPR-Cas) are largely ubiquitous, our results are supportive

of a 'second line' of defense systems (SoFIC, CBASS, Rst_Paris, etc) that is also mostly non-species-specific, differentially favored across distinct environments, and privileged by distinctive strategies of genetic mobility (**Fig. 4c**, see below). As we move down the ladder of defense system abundance, we face an increasing variety of cryptic, presumably highly specialized, and more species / population-specific systems. By further splitting our dataset into sub-environments or by geographic location, we observed significant differences in defense system abundance (**Fig. 3**). And while the increased densities observed at serpentine systems and across the arctic ocean can be explained by the extreme conditions experienced at such environments and a subsequent phage-bacteria imbalance, the more subtle variations in defense system density in human gut MAGs across multiple countries and the panoply of confounding variables associated, preclude the identification of more explanatory scenarios.

Higher densities of defense genes were consistently observed in (or in the close vicinity) of MGEs compared to those found in the chromosome (excluding MGEs) (**Fig. 4a**). Such colocalization facilitates the rapid acquisition and / or diversification of the defensome to provide resistance against multiple other MGEs. It was recently suggested that the carrying of certain defense systems by MGEs by a given bacterial host, may not always relate with the latter's need for protection, but instead on the best interest of the MGE itself in order to overcome or displace antagonistic MGEs[51]. Our observation of a complex and heterogeneous distribution of defense gene families across different classes of MGEs supports such hypothesis and suggests an exploitation of MGEs by defense genes / systems for purposes other than host defense. It ultimately highlights the need to better understand the molecular, and evolutionary interactions between the threesome host-phage-mobilome.

Genes acquired by HGT, and MGEs in particular, tend to integrate in a small number of chromosome hotspots to decrease the fitness cost of their integration. Successive rounds of integration / excision / partial deletion of MGEs when accompanied by the co-option of defense genes / systems may result in the formation of defense islands. While initially thought that the latter were merely "genomic junkyards" in which the defense genes that are frequently acquired via HGT accumulate because insertion in these regions is unlikely to be deleterious,

it has now become clear that there is a specific selective advantage in such clustering of genes, such as functional cooperation between different defensive modules and generation of novel functions. When compared across environments, defense islands did not show significant differences in terms of size, relative abundances of major defense families, or at the topmost abundant COG functional categories for genes classified as 'non-defensive' (**Fig. 5a, c, Supplementary Fig. 8c**). While many of these genes seem to encode factors involved in genetic mobility, others have hitherto unknown functions. In this line, an interesting next step would be to build upon our precise delimitation of defense islands in such a large and phylogenetically diverse MAG dataset and use a previously developed colocalization framework[3] to leverage the identification of novel defense systems. A significant overrepresentation of several defense families (e.g., Hachiman, R–M, Thoeris) was observed in defense islands (compared to non-island regions). Yet, for certain of these families, such overabundance was not synonymous of a higher likelihood of colocalization with the remainder of the defensome (and vice-versa). These observations point for the possibility of previously unappreciated epistatic interactions or increased probability of functional diversification for a selected subset of families of defense genes / systems in defense islands. In this regard, the extent to which non-canonical HGT mechanisms (e.g., gene transfer agents, nanotubes, membrane vesicles) and MGE-independent mechanisms of diversification (e.g., homologous recombination) respectively shape the movement of defense genes and the evolution of defense islands remains unclear.

Under the Red Queen evolutionary dynamics, the coevolution between opposing hosts and parasites portrays evolution as a never-ending evolutionary arms-race between defense and counter-defense strategies. Such antagonistic coevolution pervades evolutionary change through multiple ingenious strategies, including: *i*) point mutations in phage DNA recognition sites to reduce the likelihood of restriction by R–M systems[58]; *ii*) phase-variation / inversions / point mutations in MTases, REases or S modules leading to altered R–M systems' specificity[55,59]; *iii*) ON/OFF switch in CRISPR immunity through mutations in *cas* genes[60]; among others. Thus, not only turnover and recombination, but also rapid sequence evolution

of certain defense genes / systems through mutation are key factors shaping the host-parasite evolutionary trajectory. Such diversification strategies are a function of the size and the diversity of the defensome gene pool in a bacterial population, and will shape how the latter remains evolutionarily responsive to temporally or spatially variable selection imposed by phages. Different defense genes are expected to evolve at different rates. For example, significant differences in purifying selection have been described across different Types of R–M REases and MTases, highlighting distinct signatures of adaptive evolution[13]. To gain a birds-eye-view on potentially coexisting subpopulations bearing substantial defense gene-level diversity, we built upon a metagenome read recruitment approach. This allowed us to identify a subset of defense genes having a higher-than-expected frequency of SNPs + indels, globally evolving under strong purifying selection, and a heterogeneous landscape of mutation types profoundly affected by the environment (and thus by population structure). Whereas for some of these genes we can point out determinants capable of explaining such observations – namely the presence of domains known for their predisposition to genetic variation (e.g., the motility-associated *motA* domain[61] in *zorA*, or the *ftsk* translocase domain[62] in *sspH*) –, the lack of substantial functional and mechanistic insights on the remaining ones (and on their systems) precludes further meaningful ascertainments.

It is important to appreciate that our computational analysis is challenged by a few difficult-to-control confounding variables and limitations that are worth discussing. The first, concerns the imbalance in our dataset between the number of samples recovered from each biome, as well as their geographic distribution. While the number of soil and marine MAGs analyzed was essentially the same, human gut MAGs were roughly one order of magnitude greater. From the geographic standpoint, marine samples have a global representation, but soil and human gut microbiome data are greatly biased towards the US and China. These observations highlight a critical need for thorough geographic sampling, more global representation of participants in microbiome studies, and a fairer access to genomics resources especially in resource-poor countries. A second confounding variable, likely more relevant, concerns the fact that MAG binning methods using short reads tend to miss certain low-abundance or

difficult to resolve MGE families. The fact that defense genes are often carried or colocalize with MGEs, necessarily indicates that our results *i*) may have a bias in the ratio of defense genes inside versus outside the mobilome, and *ii*) are most likely a partial underestimated picture of the real defensome abundance. Future inclusion of long-read data will enable reference-quality genome reconstruction from metagenomes, and further improve our observations. Third, our observations are not representative of all bacterial communities and are likely influenced by characteristics of the sampled environments. Still, the stringent dataset filtering used in our study in terms of MAG completeness and N50 (with associated controls shown in **Supplementary Fig. 1**), together with previous demonstration on the accuracy of MAG size estimates (that are part of our dataset) compared with reference genomes[26], makes us have good reasons to think that our analyses constitute a reasonable proxy of the defense landscape diversity carried by such populations, and of the complex interplay underlying their interactions at the intra- and inter-environment level. Lastly, while this study provides novel and intriguing insights into the defensome co-localization, it does not address the specific mechanisms and interactions between different systems, nor the interplay with phage counter-defense strategies[63,64].

The efforts recently undertaken to identify novel defense mechanisms in typically easily cultivable bacteria must now be followed by initiatives to expand the search to uncultivated microbes in complex microbial communities, to understand how such mechanisms collaborate or antagonize with one another, how they co-opt or are co-opted by MGEs, and how they are shaped by the surrounding environment. Our work provides a first steppingstone in such a direction.

# Methods

## Data

In this study we built upon a large dataset of 7,759 high-quality soil, marine, and human gut MAGs[24–26] (**Supplementary Table 1**). These MAGs were filtered on the basis of the Minimum Information about a Metagenome Assembled Genome (MIMAG) standard (≥ 90% completeness, ≤ 5% contamination/redundancy, ≥ 18 tRNA genes and presence of at least one class of 5S, 16S and 23S rRNA genes). When not clearly stated in the original study, we performed identification of rRNA genes using both Infernal[65] v1.1.4 (options: -Z 1000 --hmmonly --cut_ga --noali –tblout) and RNAmmer[66] v1.2 (options: -S bac -m tsu,ssu,lsu -h -f -gff) (**Supplementary Table 3**). Since defense systems are often *i*) multigenic and *ii*) clustered in defense islands, we further selected for highly contiguous MAGs to more accurately reflect the defensome abundance and distribution. In particular, we selected assemblies having values of $N_{50}$ ≥ 100 kb (corresponding to at least the top 99.5 % best assemblies), and repeated the analyses for $N_{50}$ ≥ 200 and 300 kb (chosen upon visual inspection of the density distribution) (**Supplementary Fig. 1**) to account for the effect of contiguity in our observations. MAG annotation was performed with PROKKA[67] v1.14.5 (default parameters).

## Identification of anti-MGE defense genes, systems and islands

MAGs were queried for anti-MGE defense genes / systems using DefenseFinder[11] v1.0.8 (option: --preserve-raw). The current version of this tool allows for the screening of 1,024 genes pertaining to 127 families of anti-MGE defense systems. Defense islands were defined as arrays of defense genes (or defense systems) separated from one another by 10 genes or less and with a minimum of 5 genes pertaining to at least 3 different defense families. Functional annotation of 'non-defensive' genes was performed with eggNOG-mapper[68] v.2.1.9 (default parameters). To test for colocalization of defense families in defense islands, we computed their odds ratio and associated Fisher's exact test *P* value. For this purpose, we considered all colocalized defense genes distanced of 5 genes or less both inside and outside

defense islands. Genes belonging to the same defense system are necessarily colocalized, so we deliberately eliminated such hits to avoid inflating same system colocalization frequencies. To determine the presence of putative defense system regulators harboring WYL or CAspase Recruitement Domains (CARD), all MAG proteomes were scanned against the Pfam-HMMs PF13280 (WYL) and PF00619 (CARD) using HMMER3[69] and a cut-off *e*-value of 0.01.

## Identification of mobile genetic elements

Classification of contigs as belonging to chromosomes or plasmids was performed using PlasClass[70] v.0.1.1 and PlasFlow[71] v.1.1 (both with default parameters). Plasmid hits were selected as those with a score greater than or equal to 0.7. Integrons were identified using IntegronFinder[72] v.2.0.1 (option --local_max). Prophages were detected with Virsorter2[73] v.2.2.3 (options --include-groups dsDNAphage,ssDNA --min-length 5000 --min-score 0.5). Despite recent evidence for phage satellites carrying defense systems[74], we deliberately excluded them from our analyses, mainly due to the very few examples of experimentally validated satellites (particularly in non-cultivable bacteria), which precludes the development of robust detection tools and an accurate evaluation of their classification. Integrative Conjugative Elements (ICEs) and Integrative Mobilizable Elements (IMEs) were detected with ICEfinder[75] v.2.6.32-696.10.2.el6.x86_64 (default parameters). All MGE hits matching multiple families were not considered in the analyses (~2.8% of the total MGE dataset detected). While MGE carriage by other MGEs (e.g., integrons by plasmids) is indeed expected, we deliberately eliminated such hits to avoid the confounding effects of their co-occurrence on the defensome analyses.

## Phylogenetic analyses

For phylogenetic tree construction we took for each MAG a concatenate of 15 ribosomal proteins (L2, L3, L4, L5, L6, L14, L16, L18, L22, L24, S3, S8, S10, S17, S19), aligned them

with MAFFT[76] v7.490 (options: --maxiterate 1000 -globalpair) (soil, marine) or Muscle[77] v.5.1 (option: -super5) (human gut), and trimmed poorly aligned regions with BMGE[78] v2.0 (option: -t AA). To avoid plotting poorly supported branches, MAGs harboring less than 50% of the abovementioned ribosomal list were omitted from the phylogenetic representations (> 95 % had the expected number of proteins across the three environments). The trees were computed by maximum likelihood with RaxML[79] v8.2.12 (options: raxmlHPC-PTHREADS-AVX -f a -m PROTGAMMAAUTO -N autoMRE -p 12345 -x 12345) (soil, marine) or iqtree2[80] v2.2.6 (options: -nt 56 -cmax 15 -bb 1000 -alrt 1000 -m TESTNEW -safe) (human gut) (**Supplementary Table 5**). The phylogenetic depth was defined as the average root-to-tip distance, and was computed as the diagonal mean of the phylogenetic variance–covariance matrix of each tree, using the vcv.phylo function in the R package "ape".

## Variant analysis of the defensome

To evaluate which defense gene families are preferential targets for increased genetic diversity (SNPs + indels), we selected 90 metagenomes (30 for each environment with similar sequencing depth) having a broad representativity in terms of sampling sites (soil and marine) and countries (human gut), as well as in terms of presence of most defense families that were characteristic to each environment. Fragment recruitment was performed by mapping metagenomic reads from each sample against the ensemble of defense genes (including 200 bp upstream of the start codon) pertaining to the previously selected 90 metagenomes using BWA-MEM[81] v.0.7.17 (default parameters). Genetic variants were identified from aligned reads with FreeBayes[82] v1.1.0 (options: freebayes-parallel -p 1 -P 0 -C 1 -F 0.025 --min-repeat-entropy 1.0 -q 13 -m 60 --strict-vcf --f) and a subsequent filtering step was performed to select only genes (including upstream regions) containing variants having a minimum frequency of 25% supported by at least 10 reads. A minimum of 10 genes per defense family per environment was considered in the analysis. Alignments were visualized using IGV v.2.14.1. Finally, SNPGenie[83] v1.0 (options: --vcfformat=4 --snpreport --fastafile --gtffile --

outdir) was used for variant classification. For each environment, we computed the observed / expected (O / E) ratio of defense genes harboring high-frequency alleles across all defense families. Expected values were obtained by multiplying the total number of genes pertaining to a given defense family by the fraction of defense genes of that family harboring high-frequency alleles.

## Analysis of substitution rates

All-against-all BLASTP searches were performed on the sets of defense genes scanned in the genomes (default settings, *e*-value $<10^{-3}$). Clustering was performed using the SILIX package[84] v.1.3.0 using a minimum identity threshold of 80% and default values for the remaining parameters. Singletons were eliminated from our data set. The remaining protein sequences (putative orthologs) were reverse-translated to the corresponding DNA sequences using PAL2NAL[85] v14. Pairwise rates of non-synonymous substitutions (dN), synonymous substitutions (dS) and ω (dN/dS) were computed using the KaKs_Calculator[86] v.2.0 implementing the Yang-Nielsen[87] and Nei-Gojobori[88] methods. Estimations yielding dS > 1 (corresponding to situations of substitution saturation and representing 0.2 % of the total data) were discarded to improve the quality of estimation of ω.

## Statistical and graphical analyses of data

All statistical and graphical analyses were conducted using R v.4.3.1. Geographical representation of metagenome sampling locations was generated using the *mapdata* package. Visualization of genomic contexts was performed with the package *gggenes*. Colocalization heatmaps were created using the ComplexHeatmap package. Stepwise linear regression analyses were performed by using the *step* function from the *stats* package.

## Data availability

All data supporting the findings of this study are available within the article and its supplementary files. Source data are provided with this paper.

## Code availability

Wrapper scripts supporting all key analyses of this work are publicly available at https://github.com/oliveira-lab/Defensome.

# References


1. Haudiquet, M., De Sousa, J. M., Touchon, M. & Rocha, E. P. C. Selfish, promiscuous and sometimes useful: how mobile genetic elements drive horizontal gene transfer in microbial populations. *Philos. Trans. R. Soc. B Biol. Sci.* **377**, 20210234 (2022).
2. Oliveira, P. H., Touchon, M. & Rocha, E. P. C. Regulation of genetic flux between bacteria by restriction–modification systems. *Proc. Natl. Acad. Sci.* **113**, 5658–5663 (2016).
3. Doron, S. *et al.* Systematic discovery of antiphage defense systems in the microbial pangenome. *Science* **359**, eaar4120 (2018).
4. Millman, A. *et al.* An expanded arsenal of immune systems that protect bacteria from phages. *Cell Host Microbe* **30**, 1556-1569.e5 (2022).
5. Bernheim, A. & Sorek, R. The pan-immune system of bacteria: antiviral defence as a community resource. *Nat. Rev. Microbiol.* **18**, 113–119 (2020).
6. Mayo-Muñoz, D., Pinilla-Redondo, R., Birkholz, N. & Fineran, P. C. A host of armor: Prokaryotic immune strategies against mobile genetic elements. *Cell Rep.* **42**, 112672 (2023).
7. Rostøl, J. T. & Marraffini, L. (Ph)ighting phages: How bacteria resist their parasites. *Cell Host Microbe* **25**, 184–194 (2019).
8. Kobayashi, I. Behavior of restriction-modification systems as selfish mobile elements and their impact on genome evolution. *Nucleic Acids Res.* **29**, 3742–3756 (2001).
9. Lopatina, A., Tal, N. & Sorek, R. Abortive infection: Bacterial suicide as an antiviral immune strategy. *Annu. Rev. Virol.* **7**, 371–384 (2020).
10. Koonin, E. V. & Makarova, K. S. Origins and evolution of CRISPR-Cas systems. *Philos. Trans. R. Soc. B Biol. Sci.* **374**, 20180087 (2019).
11. Tesson, F. *et al.* Systematic and quantitative view of the antiviral arsenal of prokaryotes. *Nat. Commun.* **13**, 2561 (2022).
12. Payne, L. J. *et al.* PADLOC: a web server for the identification of antiviral defence systems in microbial genomes. *Nucleic Acids Res.* **50**, W541–W550 (2022).
13. Oliveira, P. H., Touchon, M. & Rocha, E. P. C. The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts. *Nucleic Acids Res.* **42**, 10618–10631 (2014).
14. Whitman, W. B., Coleman, D. C. & Wiebe, W. J. Prokaryotes: The unseen majority. *Proc. Natl. Acad. Sci.* **95**, 6578–6583 (1998).
15. Hofer, U. The majority is uncultured. *Nat. Rev. Microbiol.* **16**, 716–717 (2018).
16. Mushegian, A. R. Are there $10^{31}$ virus particles on Earth, or more, or fewer? *J. Bacteriol.* **202**, (2020).
17. Suttle, C. A. Marine viruses — major players in the global ecosystem. *Nat. Rev. Microbiol.* **5**, 801–812 (2007).
18. Tyson, G. W. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37–43 (2004).
19. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nat. Methods* **11**, 1144–1146 (2014).
20. Chen, L.-X., Anantharaman, K., Shaiber, A., Eren, A. M. & Banfield, J. F. Accurate and complete genomes from metagenomes. *Genome Res.* **30**, 315–333 (2020).
21. Sender, R., Fuchs, S. & Milo, R. Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLOS Biol.* **14**, e1002533 (2016).
22. Sausset, R., Petit, M. A., Gaboriau-Routhiau, V. & De Paepe, M. New insights into intestinal phages. *Mucosal Immunol.* **13**, 205–215 (2020).
23. Parikka, K. J., Le Romancer, M., Wauters, N. & Jacquet, S. Deciphering the virus-to-prokaryote ratio (VPR): insights into virus-host relationships in a variety of ecosystems. *Biol. Rev.* **92**, 1081–1100 (2017).
24. Nayfach, S. *et al.* A genomic catalog of Earth's microbiomes. *Nat. Biotechnol.* **39**, 499–509 (2021).
25. Almeida, A. *et al.* A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.* **39**, 105–114 (2021).
26. Paoli, L. *et al.* Biosynthetic potential of the global ocean microbiome. *Nature* **607**, 111–118 (2022).
27. Oliveira, P. H. & Fang, G. Conserved DNA methyltransferases: A window into fundamental mechanisms of epigenetic regulation in bacteria. *Trends Microbiol.* **29**, 28–40 (2021).
28. Sheinman, M. *et al.* Identical sequences found in distant genomes reveal frequent horizontal transfer across the bacterial domain. *eLife* **10**, e62719 (2021).
29. Rousset, F. & Sorek, R. The evolutionary success of regulated cell death in bacterial immunity. *Curr. Opin. Microbiol.* **74**, 102312 (2023).
30. Makarova, K. S., Anantharaman, V., Grishin, N. V., Koonin, E. V. & Aravind, L. CARF and WYL domains: ligand-binding regulators of prokaryotic defense systems. *Front. Genet.* **5**, (2014).
31. Blankenchip, C. L. *et al.* Control of bacterial immune signaling by a WYL domain transcription factor. *Nucleic Acids Res.* **50**, 5239–5250 (2022).
32. Picton, D. M. *et al.* A widespread family of WYL-domain transcriptional regulators co-localizes with diverse phage defence systems and islands. *Nucleic Acids Res.* **50**, 5191–5207 (2022).
33. Wein, T. *et al. CARD-like domains mediate anti-phage defense in bacterial gasdermin systems*. http://biorxiv.org/lookup/doi/10.1101/2023.05.28.542683 (2023) doi:10.1101/2023.05.28.542683.

34. Nogales, B., Lanfranconi, M. P., Piña-Villalonga, J. M. & Bosch, R. Anthropogenic perturbations in marine microbial communities. *FEMS Microbiol. Rev.* **35**, 275–298 (2011).
35. Mojica, K. D. A. & Brussaard, C. P. D. Factors affecting virus dynamics and microbial host - virus interactions in marine environments. *FEMS Microbiol. Ecol.* **89**, 495–515 (2014).
36. Williamson, K. E., Fuhrmann, J. J., Wommack, K. E. & Radosevich, M. Viruses in soil ecosystems: an unknown quantity within an unexplored territory. *Annu. Rev. Virol.* **4**, 201–219 (2017).
37. Mishra, A., Singh, L. & Singh, D. Unboxing the black box—one step forward to understand the soil microbiome: A systematic review. *Microb. Ecol.* **85**, 669–683 (2023).
38. Martiny, J. B. H. *et al.* Microbial biogeography: putting microorganisms on the map. *Nat. Rev. Microbiol.* **4**, 102–112 (2006).
39. Hanson, C. A., Fuhrman, J. A., Horner-Devine, M. C. & Martiny, J. B. H. Beyond biogeographic patterns: processes shaping the microbial landscape. *Nat. Rev. Microbiol.* **10**, 497–506 (2012).
40. Twing, K. I. *et al.* Serpentinization-influenced groundwater harbors extremely low diversity microbial communities adapted to high pH. *Front. Microbiol.* **8**, (2017).
41. Thomas, E., Anderson, R. E., Li, V., Rogan, L. J. & Huber, J. A. Diverse viruses in deep-sea hydrothermal vent fluids have restricted dispersal across ocean basins. *mSystems* **6**, e00068-21 (2021).
42. Drew, G. C., Stevens, E. J. & King, K. C. Microbial evolution and transitions along the parasite–mutualist continuum. *Nat. Rev. Microbiol.* **19**, 623–638 (2021).
43. Huang, D. *et al.* Enhanced mutualistic symbiosis between soil phages and bacteria with elevated chromium-induced environmental stress. *Microbiome* **9**, 150 (2021).
44. Tang, X. *et al.* Lysogenic bacteriophages encoding arsenic resistance determinants promote bacterial community adaptation to arsenic toxicity. *ISME J.* **17**, 1104–1115 (2023).
45. Dy, R. L., Przybilski, R., Semeijn, K., Salmond, G. P. C. & Fineran, P. C. A widespread bacteriophage abortive infection system functions through a Type IV toxin–antitoxin mechanism. *Nucleic Acids Res.* **42**, 4590–4605 (2014).
46. Trubl, G. *et al.* Active virus-host interactions at sub-freezing temperatures in Arctic peat soil. *Microbiome* **9**, 208 (2021).
47. Wells, L. E. & Deming, J. W. Modelled and measured dynamics of viruses in Arctic winter sea-ice brines. *Environ. Microbiol.* **8**, 1115–1121 (2006).
48. Collins, R. E. & Deming, J. W. Abundant dissolved genetic material in Arctic sea ice Part II: Viral dynamics during autumn freeze-up. *Polar Biol.* **34**, 1831–1841 (2011).
49. Danovaro, R., Manini, E. & Dell'Anno, A. Higher abundance of bacteria than of viruses in deep Mediterranean sediments. *Appl. Environ. Microbiol.* **68**, 1468–1472 (2002).
50. Segata, N. Gut microbiome: westernization and the disappearance of intestinal diversity. *Curr. Biol.* **25**, R611–R613 (2015).
51. Rocha, E. P. C. & Bikard, D. Microbial defenses against mobile genetic elements and viruses: Who defends whom from what? *PLOS Biol.* **20**, e3001514 (2022).
52. Botelho, J. Defense systems are pervasive across chromosomally integrated mobile genetic elements and are inversely correlated to virulence and antimicrobial resistance. *Nucleic Acids Res.* **51**, 4385–4397 (2023).
53. Buongermino Pereira, M. *et al.* A comprehensive survey of integron-associated genes present in metagenomes. *BMC Genomics* **21**, 495 (2020).
54. Redondo-Salvo, S. *et al.* Pathways for horizontal gene transfer in bacteria revealed by a global map of their plasmids. *Nat. Commun.* **11**, 3602 (2020).
55. Saravanan, M., Vasu, K. & Nagaraja, V. Evolution of sequence specificity in a restriction endonuclease by a point mutation. *Proc. Natl. Acad. Sci.* **105**, 10344–10347 (2008).
56. Takeuchi, N., Wolf, Y. I., Makarova, K. S. & Koonin, E. V. Nature and intensity of selection pressure on CRISPR-associated genes. *J. Bacteriol.* **194**, 1216–1225 (2012).
57. Giovannoni, S. J., Cameron Thrash, J. & Temperton, B. Implications of streamlining theory for microbial ecology. *ISME J.* **8**, 1553–1565 (2014).
58. Pleška, M. & Guet, C. C. Effects of mutations in phage restriction sites during escape from restriction–modification. *Biol. Lett.* **13**, 20170646 (2017).
59. Atack, J. M., Guo, C., Yang, L., Zhou, Y. & Jennings, M. P. DNA sequence repeats identify numerous Type I restriction-modification systems that are potential epigenetic regulators controlling phase-variable regulons; phasevarions. *FASEB J.* **34**, 1038–1051 (2020).
60. Watson, B. N. J., Steens, J. A., Staals, R. H. J., Westra, E. R. & Van Houte, S. Coevolution between bacterial CRISPR-Cas systems and their bacteriophages. *Cell Host Microbe* **29**, 715–725 (2021).
61. Mohawk, K. L., Poly, F., Sahl, J. W., Rasko, D. A. & Guerry, P. High frequency, spontaneous *motA* mutations in *Campylobacter jejuni* strain 81-176. *PLoS ONE* **9**, e88043 (2014).
62. Diez, A. A., Farewell, A., Nannmark, U. & Nyström, T. A mutation in the *ftsK* gene of *Escherichia coli* affects cell-cell separation, stationary-phase survival, stress adaptation, and expression of the gene encoding the stress protein UspA. *J. Bacteriol.* **179**, 5878–5883 (1997).
63. Srikant, S., Guegler, C. K. & Laub, M. T. The evolution of a counter-defense mechanism in a virus constrains its host range. *eLife* **11**, e79549 (2022).
64. Ho, P. *et al.* Bacteriophage antidefense genes that neutralize TIR and STING immune responses. *Cell Rep.* **42**, 112305 (2023).
65. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).

66. Lagesen, K. *et al.* RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* **35**, 3100–3108 (2007).
67. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
68. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).
69. Eddy, S. R. A probabilistic model of local sequence alignment that simplifies statistical significance estimation. *PLoS Comput. Biol.* **4**, e1000069 (2008).
70. Pellow, D., Mizrahi, I. & Shamir, R. PlasClass improves plasmid sequence classification. *PLOS Comput. Biol.* **16**, e1007781 (2020).
71. Krawczyk, P. S., Lipinski, L. & Dziembowski, A. PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic Acids Res.* **46**, e35–e35 (2018).
72. Néron, B. *et al.* IntegronFinder 2.0: Identification and analysis of integrons across bacteria, with a focus on antibiotic resistance in *Klebsiella*. *Microorganisms* **10**, 700 (2022).
73. Guo, J. *et al.* VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* **9**, 37 (2021).
74. Rousset, F. *et al.* Phages and their satellites encode hotspots of antiviral systems. *Cell Host Microbe* **30**, 740-753.e5 (2022).
75. Liu, M. *et al.* ICEberg 2.0: an updated database of bacterial integrative and conjugative elements. *Nucleic Acids Res.* **47**, D660–D665 (2019).
76. Katoh, K. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).
77. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
78. Criscuolo, A. & Gribaldo, S. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol. Biol.* **10**, 210 (2010).
79. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
80. Minh, B. Q. *et al.* IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
81. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. (2013) doi:10.48550/ARXIV.1303.3997.
82. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. (2012) doi:10.48550/ARXIV.1207.3907.
83. Nelson, C. W., Moncla, L. H. & Hughes, A. L. SNPGenie: estimating evolutionary parameters to detect natural selection using pooled next-generation sequencing data. *Bioinformatics* **31**, 3709–3711 (2015).
84. Miele, V., Penel, S. & Duret, L. Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC Bioinformatics* **12**, 116 (2011).
85. Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**, W609–W612 (2006).
86. Wang, D., Zhang, Y., Zhang, Z., Zhu, J. & Yu, J. KaKs_Calculator 2.0: A toolkit incorporating gamma-series methods and sliding window strategies. *Genomics Proteomics Bioinformatics* **8**, 77–80 (2010).
87. Yang, Z. & Nielsen, R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* **17**, 32–43 (2000).
88. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* (1986) doi:10.1093/oxfordjournals.molbev.a040410.

## Acknowledgements

This work was supported by the Genoscope, the Commissariat à l'Énergie Atomique et aux Énergies Alternatives (CEA), France Génomique (ANR-10-INBS-09–08), and by the Interdisciplinary Center MICROBES of the University Paris-Saclay, as part of the France 2030 program ANR-11-IDEX-0003. L.P. is supported by a European Molecular Biology Organization Postdoctoral Fellowship (EMBO ALTF 100-2023). We thank Eduardo P. C. Rocha (Institut Pasteur, Paris) for critical reading of the manuscript, and Hadrien Guichard (CEA, Genoscope) for initial efforts in defensome analyses. We acknowledge Dr. João Botelho (CBGP, Universidad Politécnica de Madrid) and the two additional anonymous reviewers for their careful reading and insightful comments and suggestions.

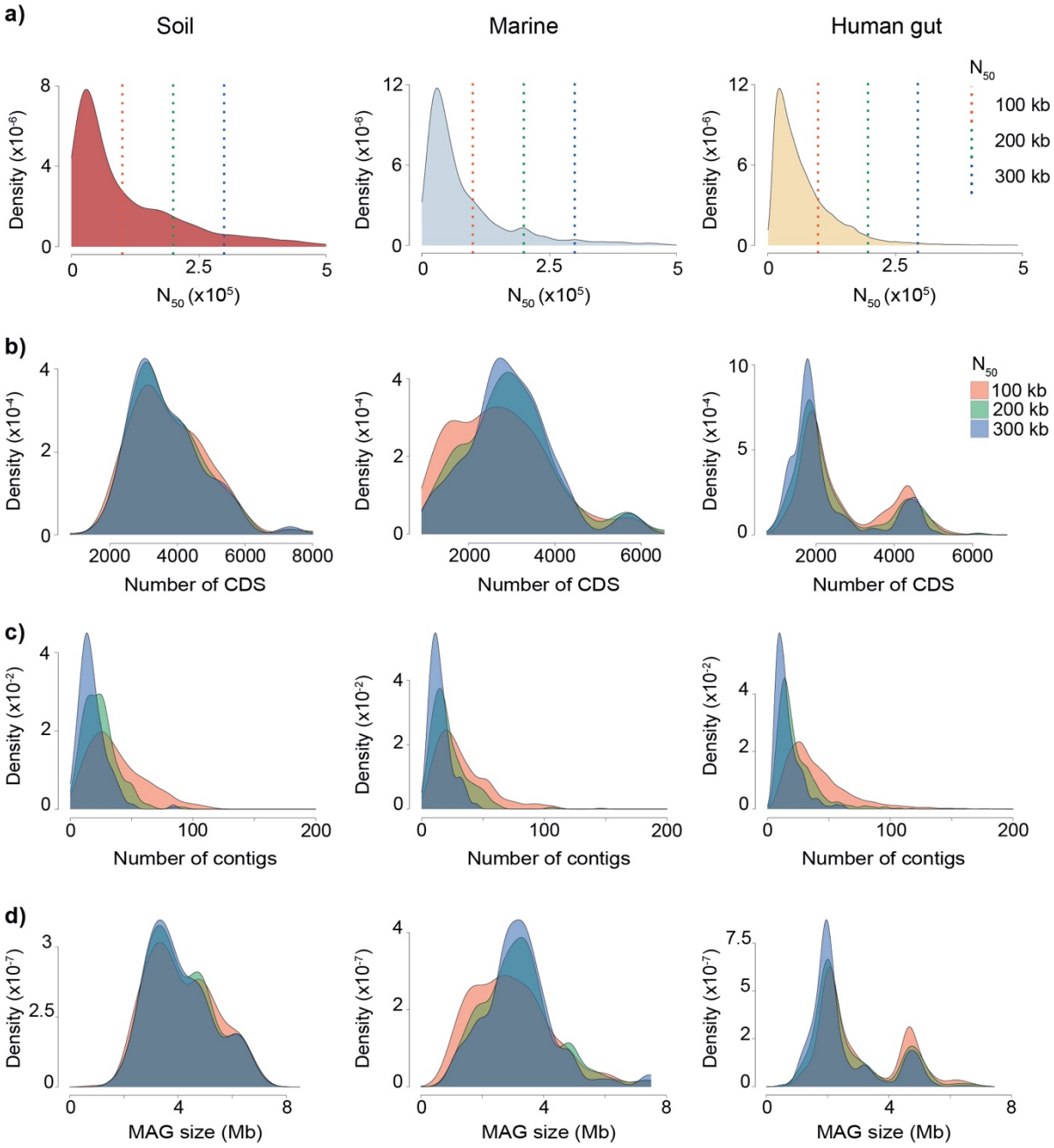
## Author Contributions

P.H.O supervised the project. A.B. and P.H.O. designed the computational methods. A.B. and A.L. performed most of the computational analyses and developed most of the scripts that support the analyses. A.B., A.L., N.W., J.P., T.O.D., L.P., P.W. and P.H.O. analysed the data. A.B. and P.H.O. wrote the manuscript with additional information inputs from other co-authors.

## Competing interests

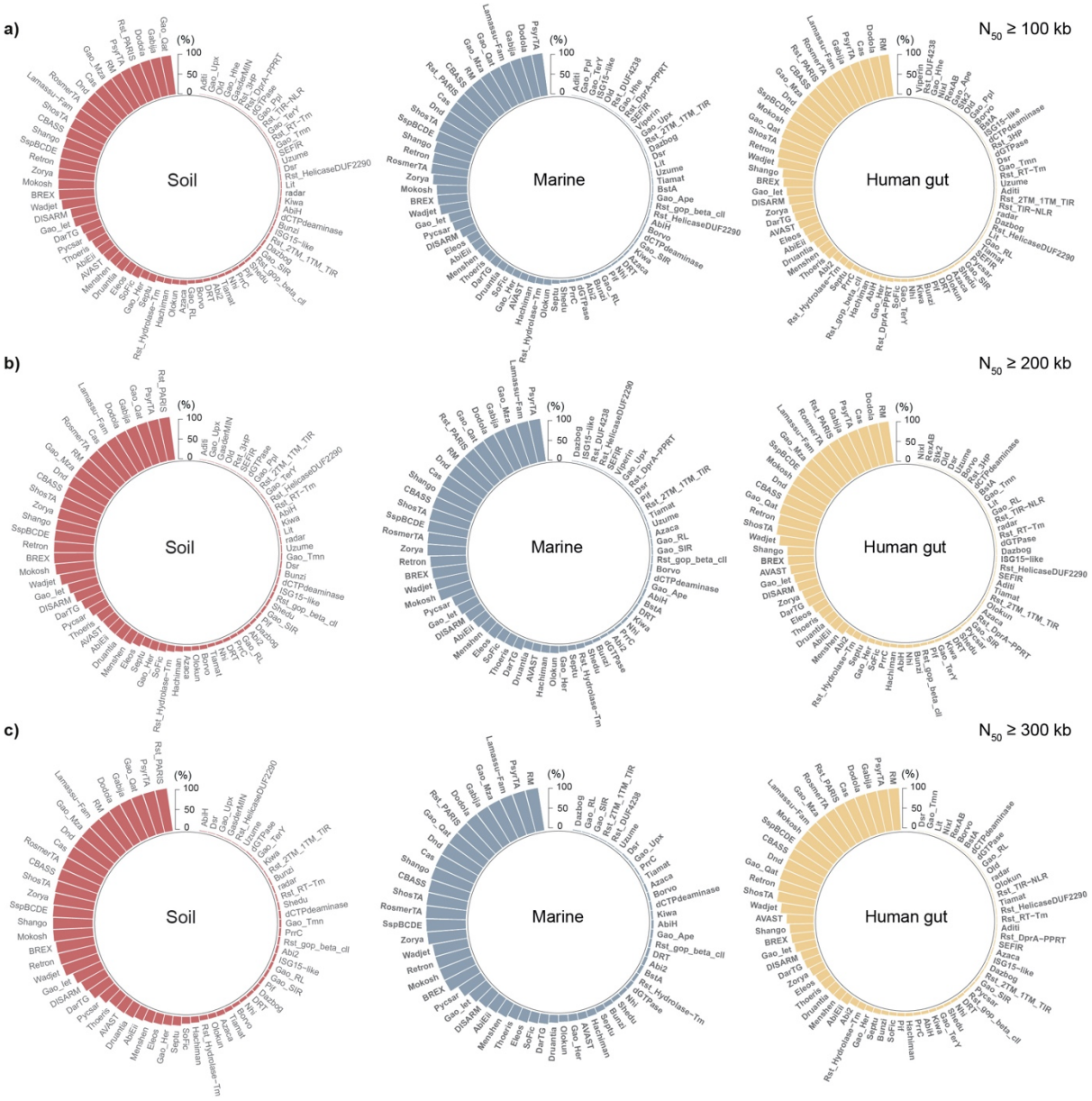
The authors declare no competing interests.

1 ## Supplementary Figures

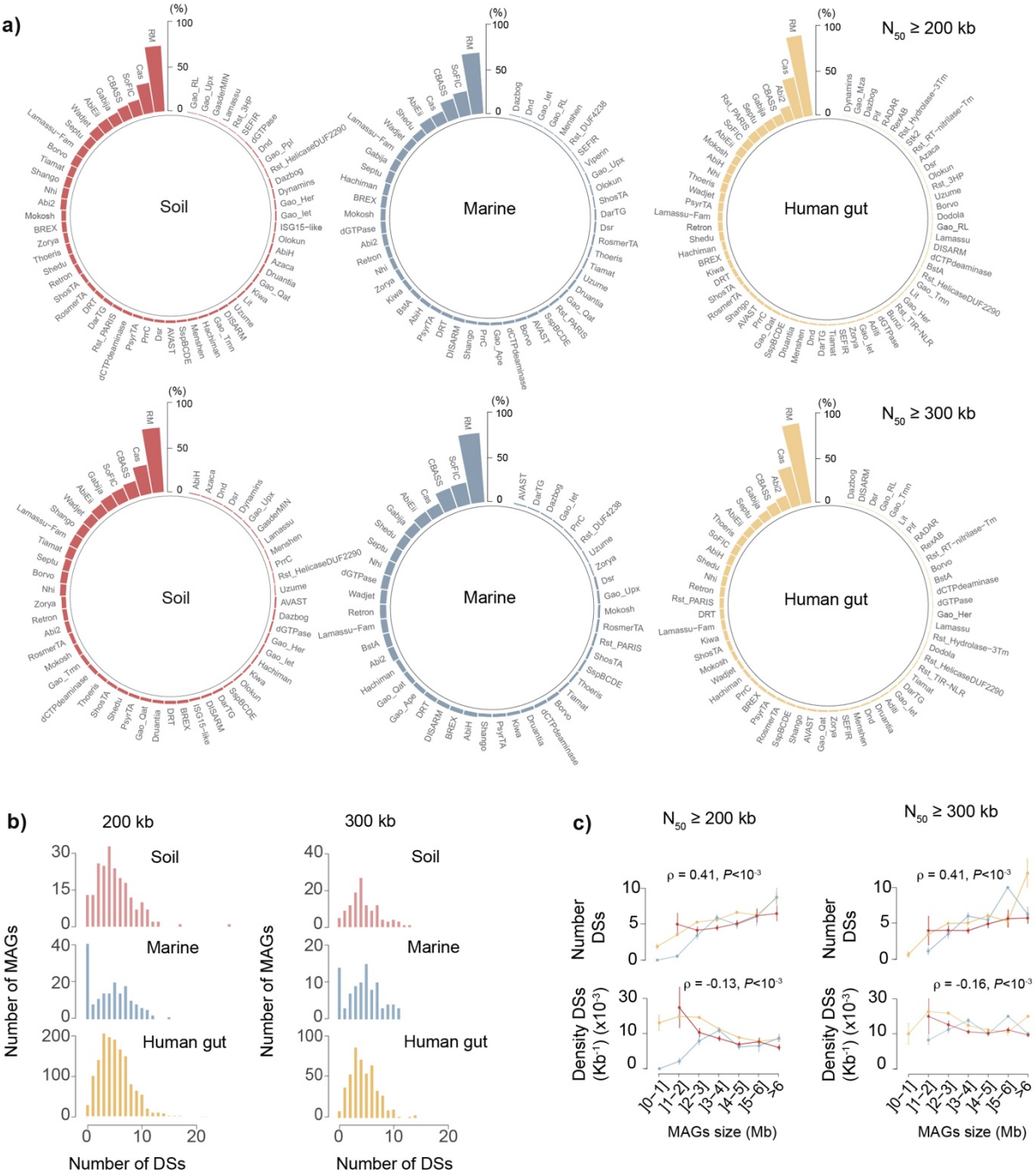



2

3 **Supplementary Fig. 1**. Density distributions of key MAG parameters according to environment: (*a*) $N_{50}$; (*b*) number
4 of CDS; (*c*) number of contigs; and (*d*) size.

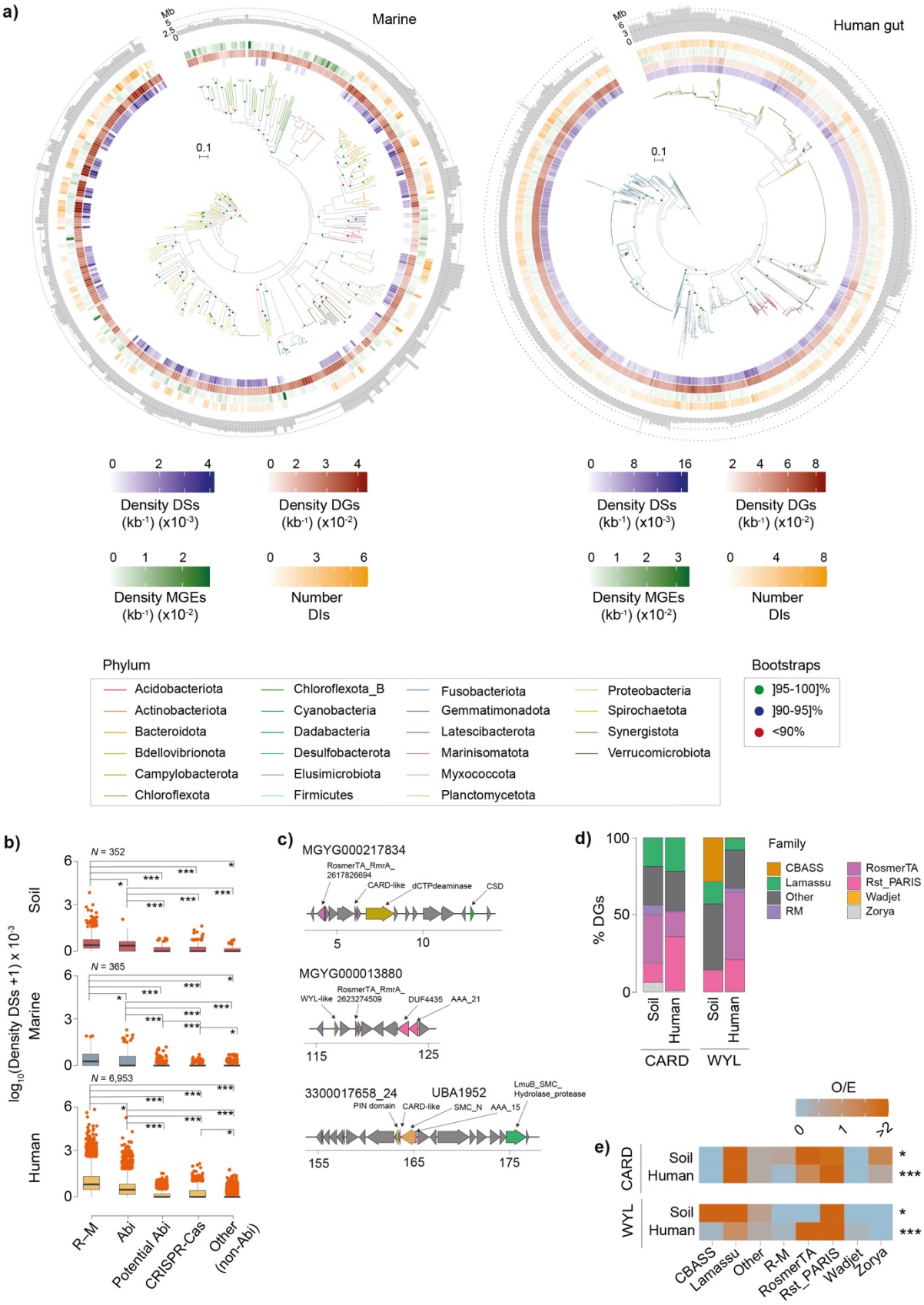



**Supplementary Fig. 2**. Percentage of soil, marine and human gut MAGs harboring each family of defense genes. Shown are assemblies with values of $N_{50}$ ≥ 100, 200, and 300 kb.




**Supplementary Fig. 3**. Distribution of defense systems in MAGs. (*a*) Percentage of soil, marine and human gut MAGs harboring each family of defense systems. Shown are assemblies with values of $N_{50}$ ≥ 200 and 300 kb. (*b*) Distribution of number of defense systems (DSs, per MAG) across environments for assemblies with values of $N_{50}$ ≥ 200 and 300 kb. (*c*) Variation of number and density (per kb) of defense systems (DSs) with MAG size (Mb) for assemblies with values of $N_{50}$ ≥ 200 and 300 kb and for each environment. Error bars represent standard deviations.

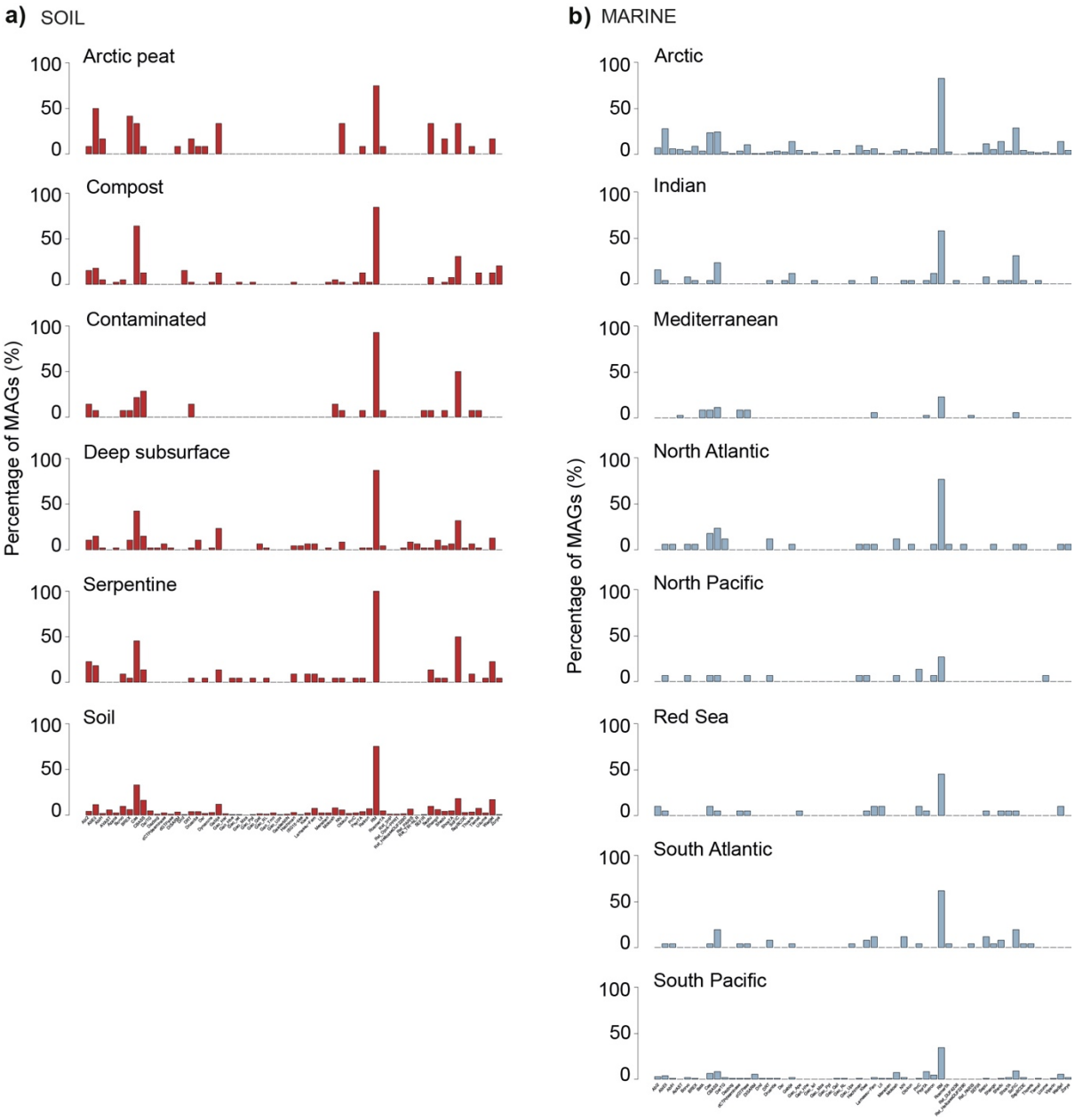
1


a)
Marine
Human gut
Mb
0.1
Density DSs (kb-1) (x10-3)
Density DGs (kb-1) (x10-2)
Density MGEs (kb-1) (x10-2)
Number DIs
Phylum
Acidobacteriota
Actinobacteriota
Bacteroidota
Bdellovibrionota
Campylobacterota
Chloroflexota
Chloroflexota_B
Cyanobacteria
Dadabacteria
Desulfobacterota
Elusimicrobiota
Firmicutes
Fusobacteriota
Gemmatimonadota
Latescibacterota
Marinisomatota
Myxococcota
Planctomycetota
Proteobacteria
Spirochaetota
Synergistota
Verrucomicrobiota
Bootstraps
]95-100]%
]90-95]%
<90%
b)
N = 352
N = 365
N = 6,953
Soil
Marine
Human
log10(Density DSs +1) x 10-3
R–M
Abi
Potential Abi
CRISPR-Cas
Other (non-Abi)
c)
MGYG000217834
RosmerTA_RmrA_ 2617826694
CARD-like
dCTPdeaminase
CSD
MGYG000013880
WYL-like
RosmerTA_RmrA_ 2623274509
DUF4435
AAA_21
3300017658_24
UBA1952
PIN domain
CARD-like
SMC_N
AAA_15
LmuB_SMC_ Hydrolase_protease
d)
% DGs
Family
CBASS
Lamassu
Other
RM
RosmerTA
Rst_PARIS
Wadjet
Zorya
Soil
Human
CARD
WYL
e)
O/E
CBASS
Lamassu
Other
R-M
RosmerTA
Rst_PARIS
Wadjet
Zorya

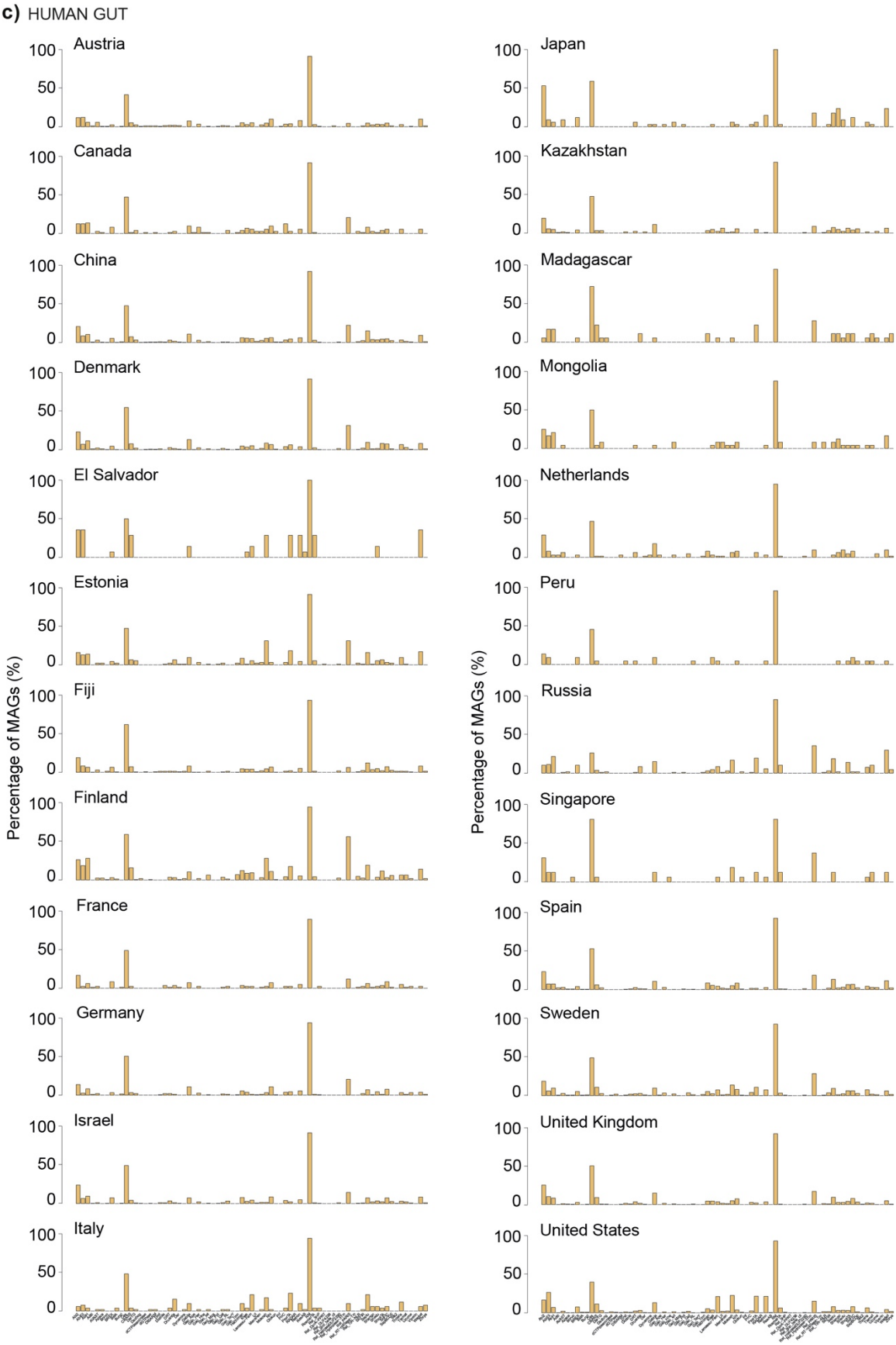
2

**Supplementary Fig. 4**. Phylogenetic distribution of the defensome and association with types of defense mechanism and CARD / WYL domains. (*a*) Phylogenetic representation of 386 marine and 6,369 human gut MAGs, their corresponding phyla, density (per kb) of defense systems (DSs, purple), defense genes (DGs, red), MGEs (green), and number of defense islands (DIs, yellow). Distribution of MAG sizes (Mb) are shown as outer layer barplots. (*b*) Defense system (DS) density (per MAG per kb) split per biome and underlying defense mechanism (R–M, Abi, potential Abi, CRISPR-Cas, and other (non-Abi)). Error bars represent standard deviations and Mann–Whitney–Wilcoxon test *P* values are indicated. $^{*}P < 0.05$; $^{**}P < 10^{-2}$; $^{***}P < 10^{-3}$. (*c*) Representative instances of colocalized (less than 10 genes apart) WYL and CARD domain-containing genes with defense genes in the human MAGs MGYG000217834 and MGYG000013880. Also shown is a chromosomal region of the Bacteroidetes bacterium UBA1952 sharing similarity with a recently described *Pedobacter rhizosphaerae* CARD-encoding defense system[33]. (*d*) Stacked barplots showing the relative abundance of families of defense genes that colocalize with WYL- and CARD-like domains. (*e*) Heatmap of observed / expected (O / E) ratios of colocalization between genes belonging to distinct defense families and WYL and CARD-like domains. Expected values were obtained by multiplying the total number of genes pertaining to a given defense family by the fraction of defense genes of that family colocalized with a given domain. *P* values correspond to the $\chi^2$-test. To avoid performing analyses with weak statistical power, we omitted the marine MAG dataset due to their low WYL and CARD domain abundance (detected in as much 0.75% of the dataset).

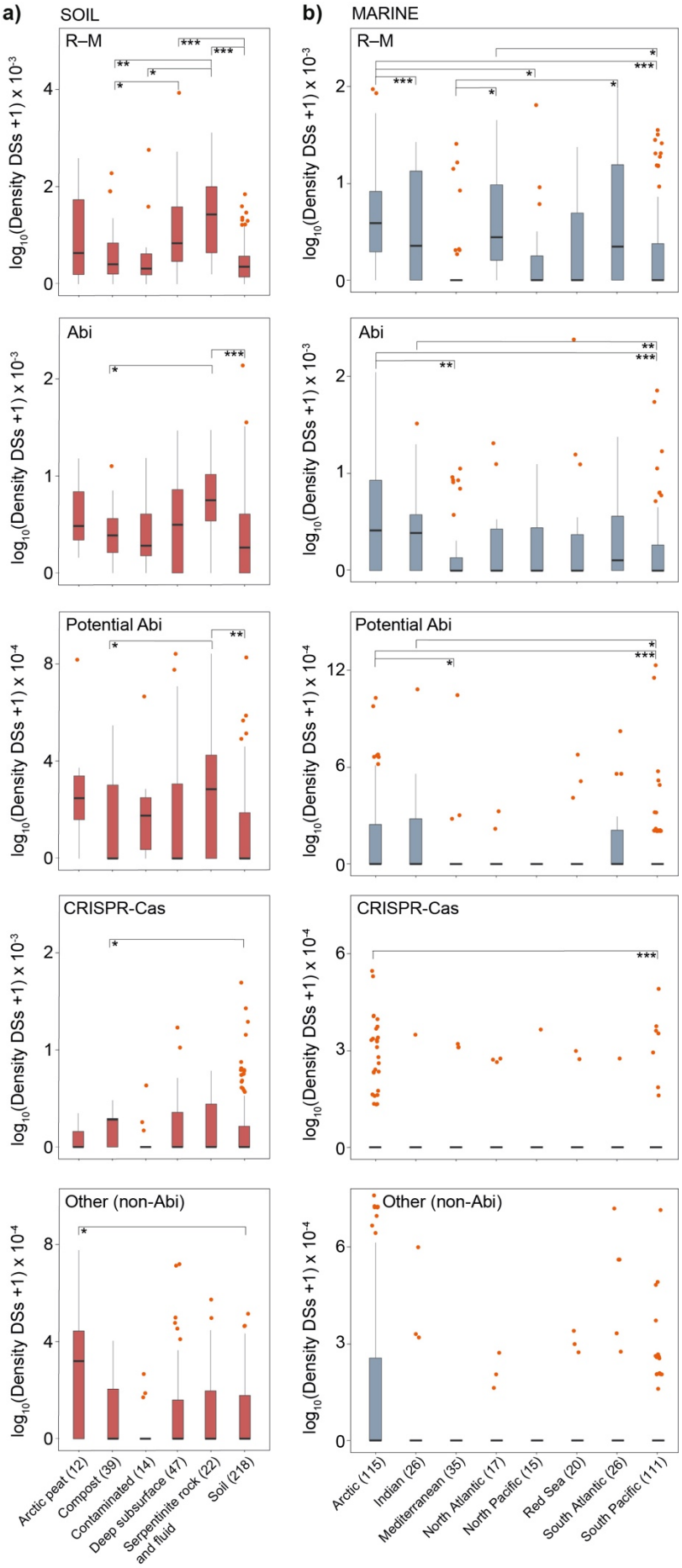
1


a) SOIL
Arctic peat
Compost
Contaminated
Deep subsurface
Serpentine
Soil
Percentage of MAGs (%)
100
50
0
b) MARINE
Arctic
Indian
Mediterranean
North Atlantic
North Pacific
Red Sea
South Atlantic
South Pacific
Percentage of MAGs (%)

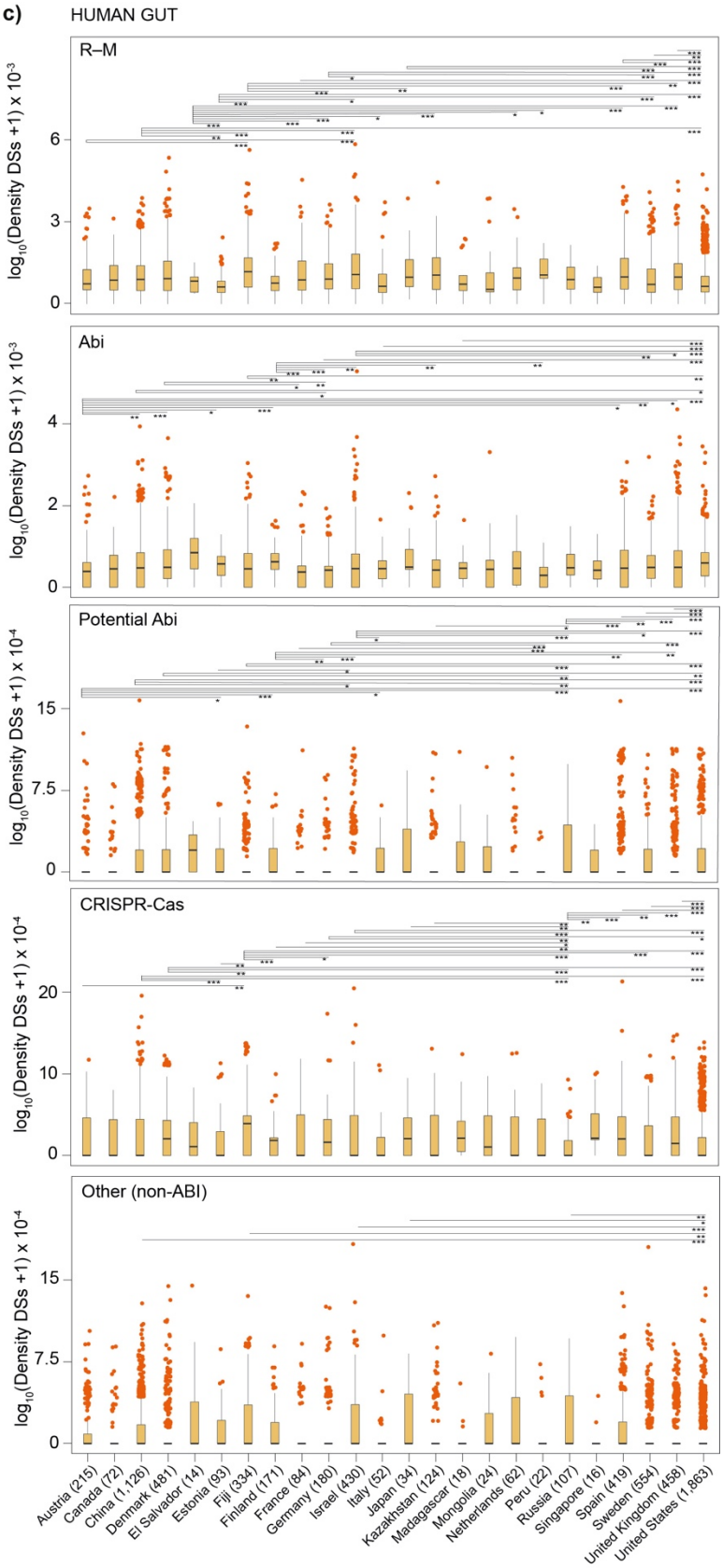
2



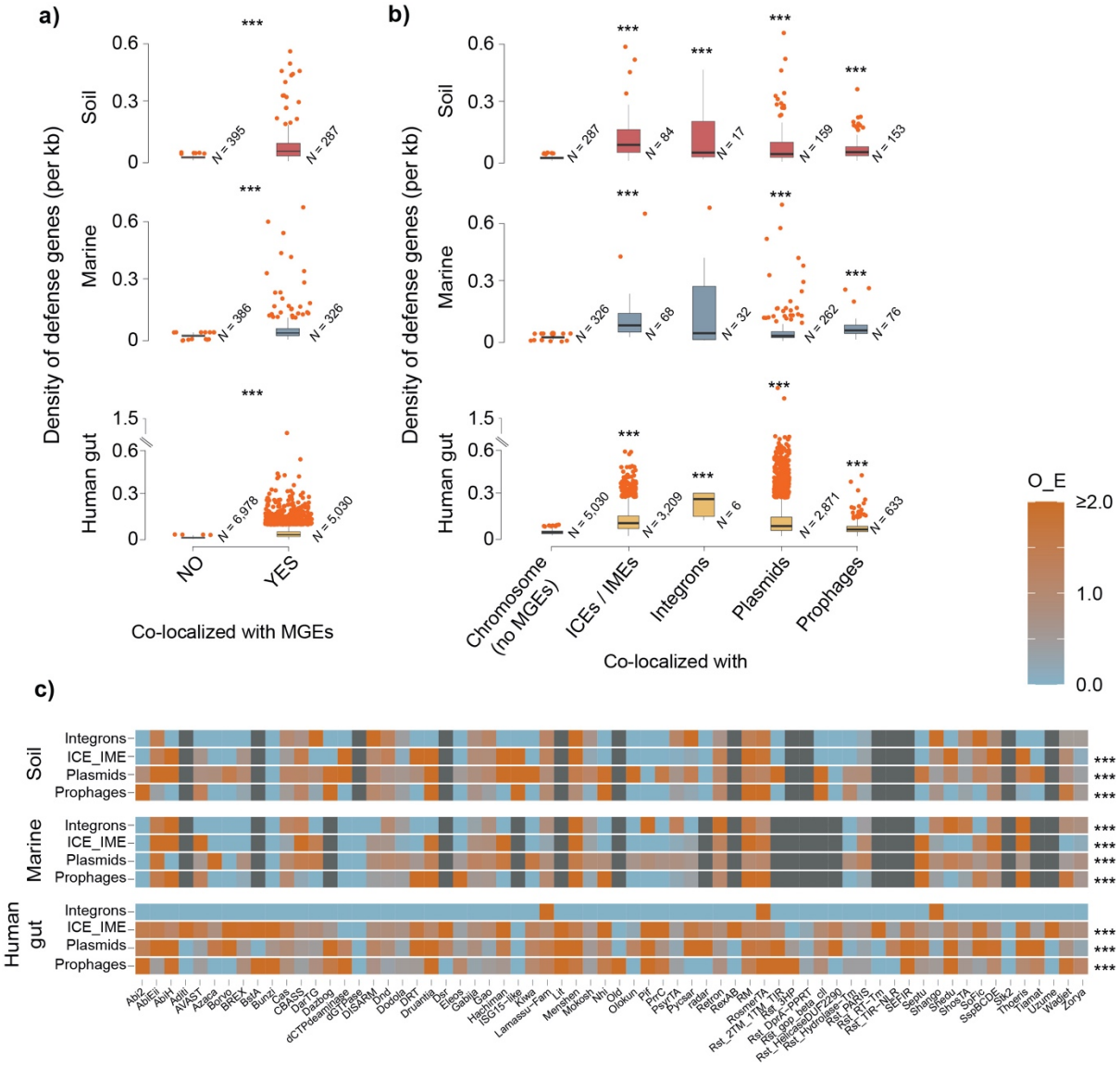

**Supplementary Fig. 5**. Percentage of (*a*) soil, (*b*) marine and (*c*) human gut MAGs harboring each family of defense system across different ecological and geographical backgrounds.


a)
SOIL
b)
MARINE
R–M
Abi
Potential Abi
CRISPR-Cas
Other (non-Abi)
log10(Density DSs +1) x 10-3
log10(Density DSs +1) x 10-4
Arctic peat (12)
Compost (39)
Contaminated (14)
Deep subsurface (47)
Serpentinite rock (22)
and fluid
Soil (218)
Arctic (115)
Indian (26)
Mediterranean (35)
North Atlantic (17)
North Pacific (15)
Red Sea (20)
South Atlantic (26)
South Pacific (111)

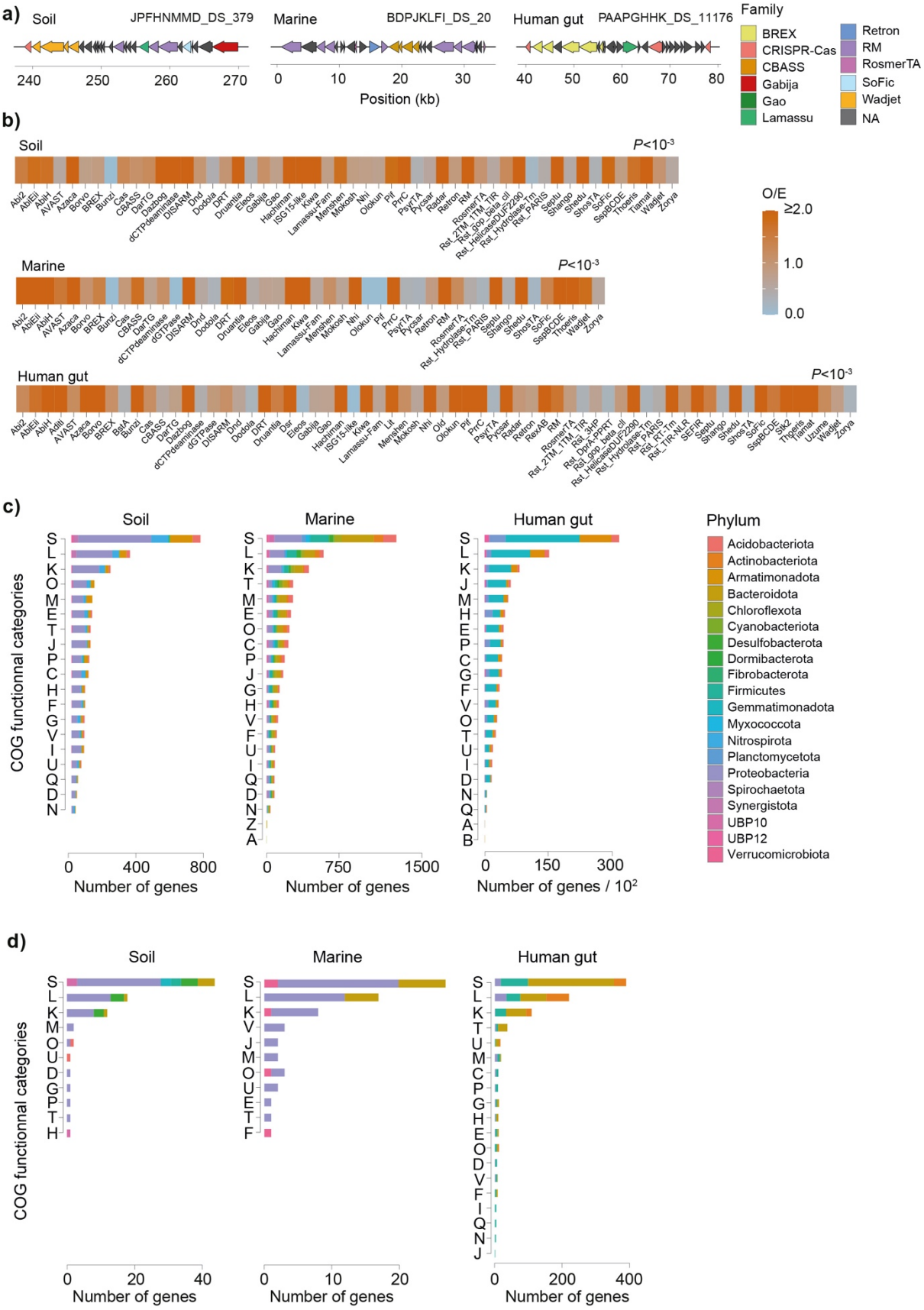
1




**Supplementary Fig. 6**. Defense system (DS) density (per MAG per kb) split according to defense mechanism (R–M, Abi, potential Abi, CRISPR-Cas, and other (non-Abi)), and different ecological (soil, marine) (*a*, *b*) and geographical (human gut) (*c*) contexts. Error bars represent standard deviations and Mann–Whitney–Wilcoxon test *P* values are indicated. **P* < 0.05; ***P* < $10^{-2}$; ****P* < $10^{-3}$.

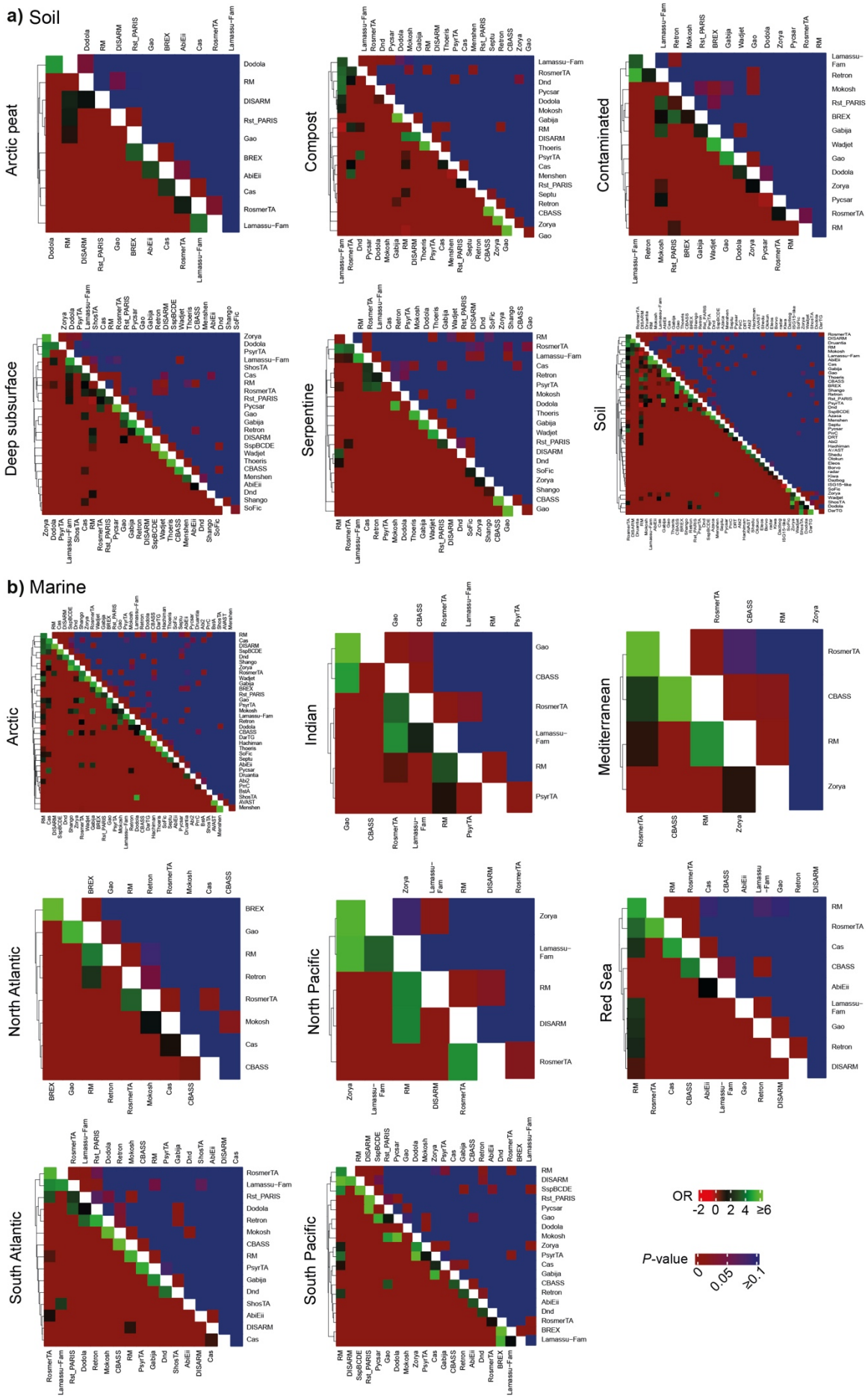



**Supplementary Fig. 7**. The genetic mobility of the defensome. Reproduction of **Fig. 4** of the manuscript with the inclusion of integrons. *$P < 0.05$; **$P < 10^{-2}$; ***$P < 10^{-3}$, $\chi^2$-test.

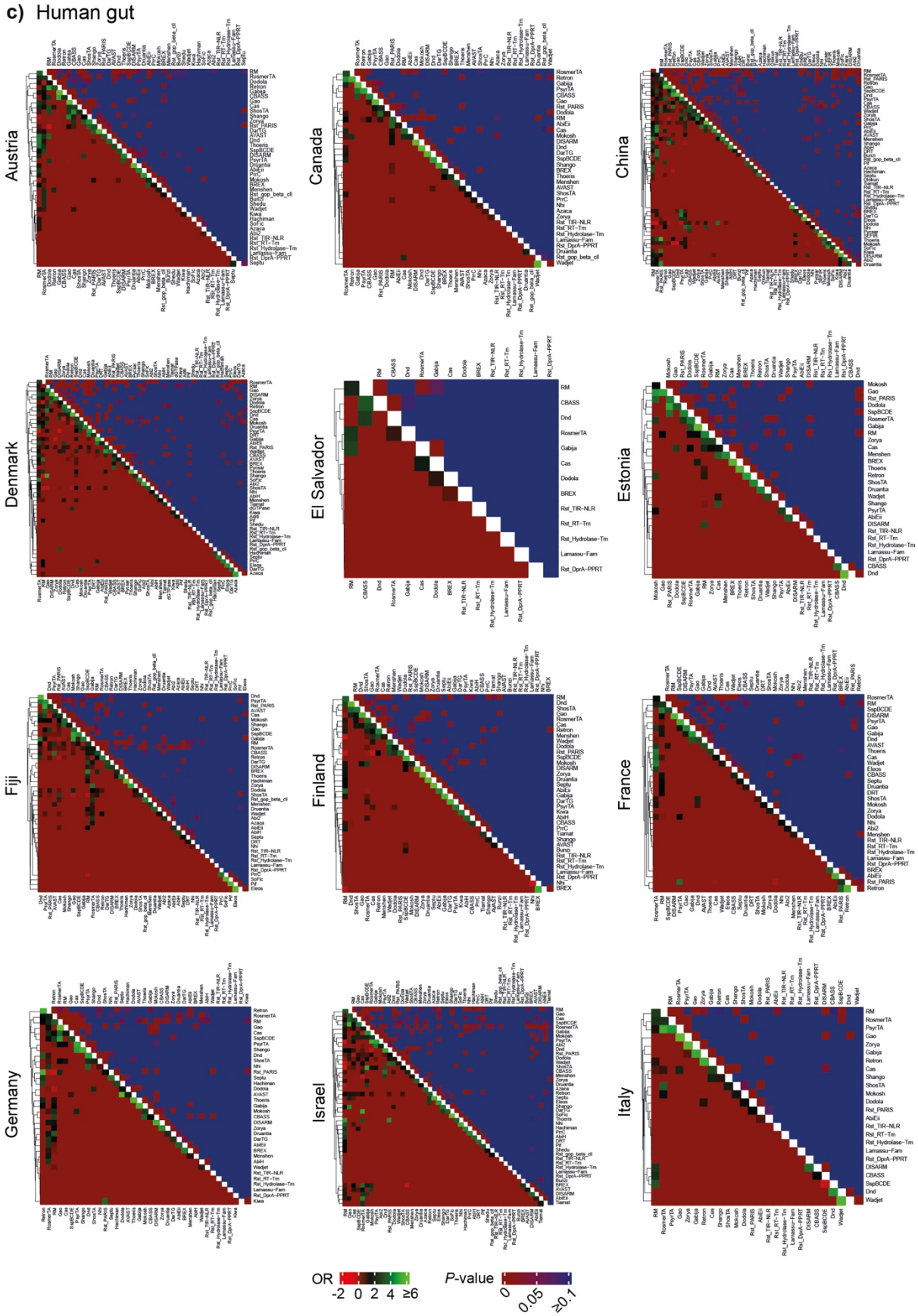


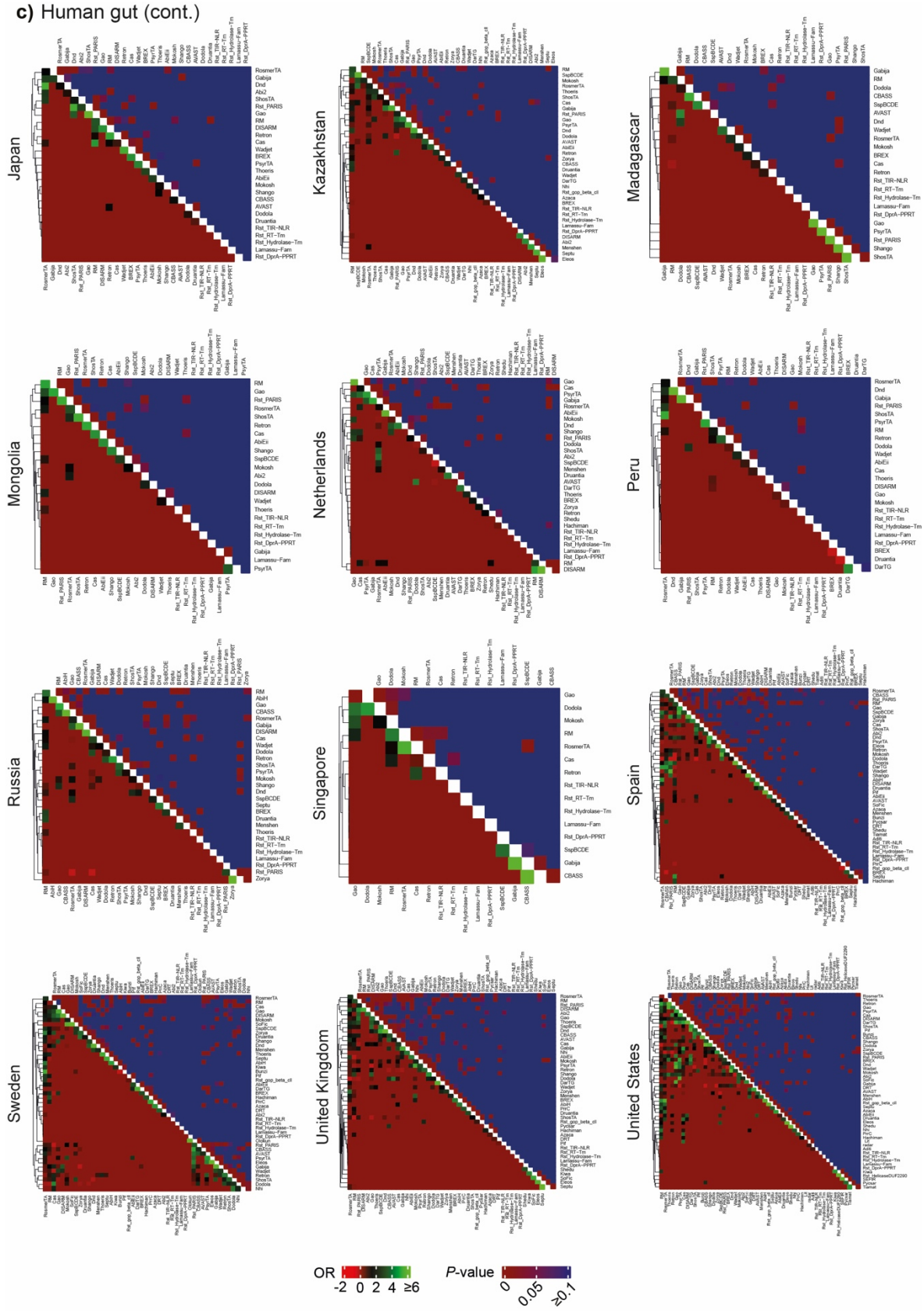

**Supplementary Fig. 8**. The MAG defense island repertoire. (*a*) Representative examples of defense islands for each environment, illustrating the high diversity of defense families present. (*b*) Heatmap of observed/expected (O / E) ratios of colocalization between genes belonging to distinct defense families and defense islands. Expected

values were obtained by multiplying the total number of genes pertaining to a given defense family by the fraction of defense genes of that family assigned to a defense island. *P* values correspond to the $\chi^2$-test. (*c*) COG functional annotation assessed by Eggnog-mapper of the ensemble of non-defensive genes in defense islands. (*d*) COG functional annotation of the ensemble of non-defensive genes in defense islands when complete defense systems were used as counting units for the classification of defense islands (see Methods).

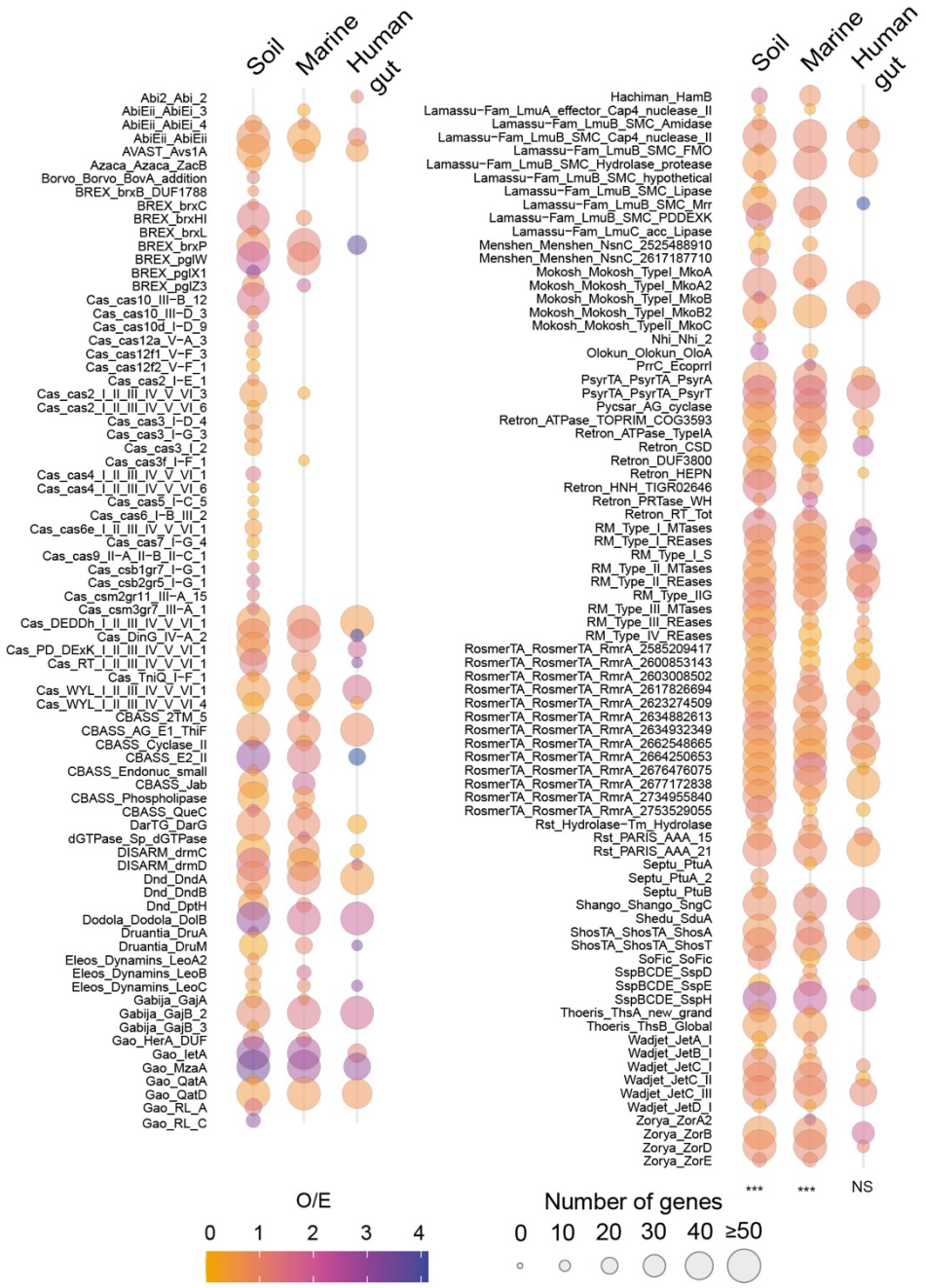

a) Soil
Arctic peat
Compost
Contaminated
Deep subsurface
Serpentine
Soil
b) Marine
Arctic
Indian
Mediterranean
North Atlantic
North Pacific
Red Sea
South Atlantic
South Pacific
OR
-2 0 2 4 ≥6
P-value
0
0.05
≥0.1

1

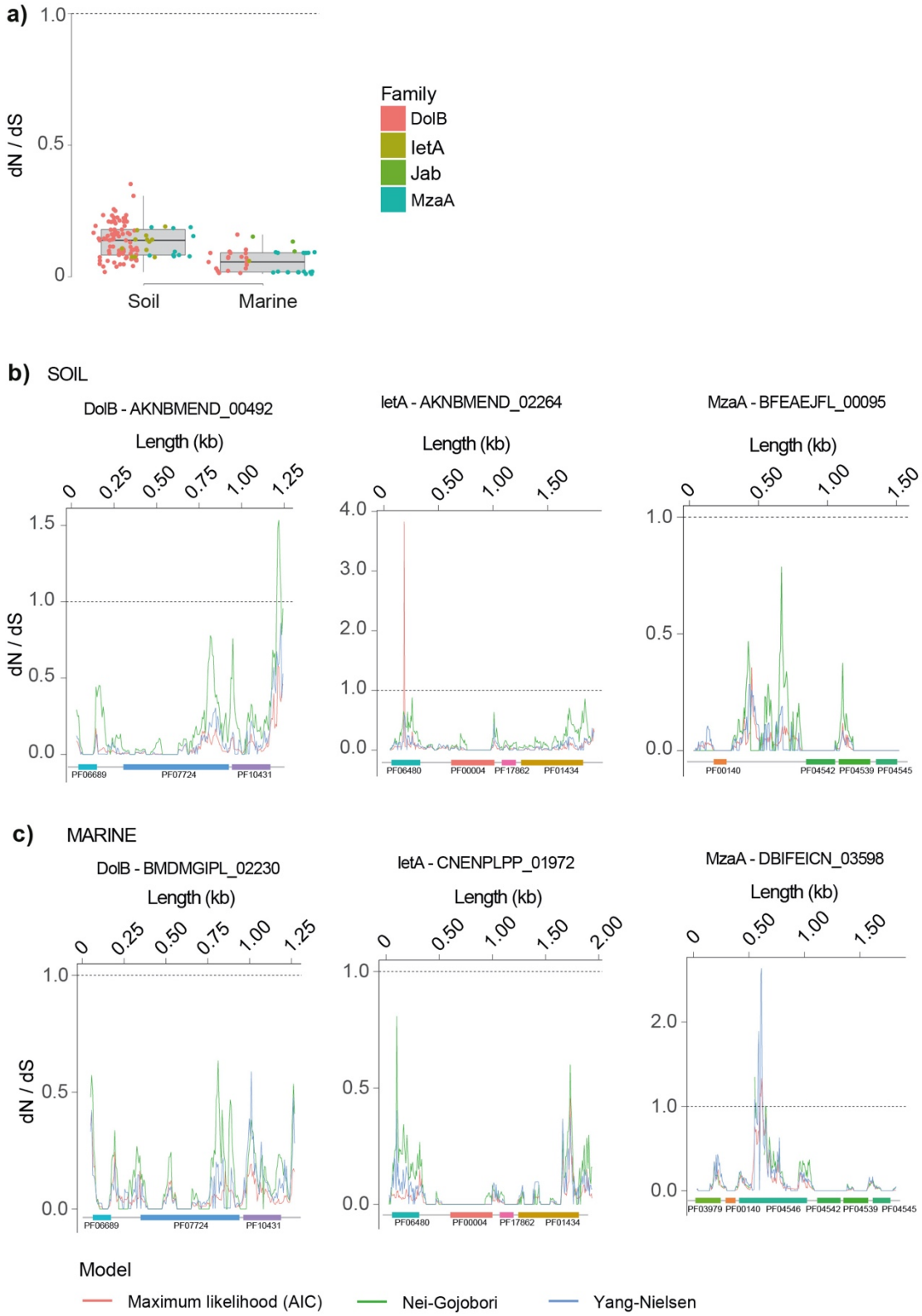

c) Human gut
Austria
Canada
China
Denmark
El Salvador
Estonia
Fiji
Finland
France
Germany
Israel
Italy
OR
-2 0 2 4 ≥6
P-value
0 0.05 ≥0.1




**Supplementary Fig. 9**. Defense families' odds ratio (OR) of colocalization in defense islands (bottom heatmaps) and associated Fisher's exact test *P* value (upper heatmaps) for different ecological (soil, marine) (*a*, *b*) and geographical (human gut) (*c*) contexts.




**Supplementary Fig. 10**. The genetic variability of the defensome. (*a*) 90 metagenomes (30 for each environment) having a broad representativity in terms of sampling sites (soil and marine) and countries (human gut), as well as in terms of presence of most defense families previously identified by DefenseFinder were selected. Shown in circles are the observed / expected (O / E) ratios of number of defense gene families harboring high-frequency (≥ 25% of coverage at the variant position) SNPs + indels positions in their gene body (including 200 bp upstream the start codon). No thresholds on O / E ratio were introduced. Expected values were obtained by multiplying the total number of genes pertaining to a given defense family by the fraction of defense genes of that family harboring high-frequency alleles. Circle radius corresponds to the total number of defense genes analyzed per family. All defense families are represented irrespectively of their O / E ratio span.




**Supplementary Fig. 11**. Evolution of defense genes. (*a*) Variation in global dN/dS given by the Nei-Gojobori (NG) and Yang-Nielson (NY) methods for a selection of defense genes shown to harbor a significantly higher frequency of SNPs + Indels. (*b*) Across gene profiles of dN/dS given by the Maximum Likelihood (Akaike Information Criterium), Nei-Gojobori (NG), and Yang-Nielson (NY) methods for a selection of three defense genes simultaneously present in soil and marine environments. PFAM domains are shown as colored rectangles.