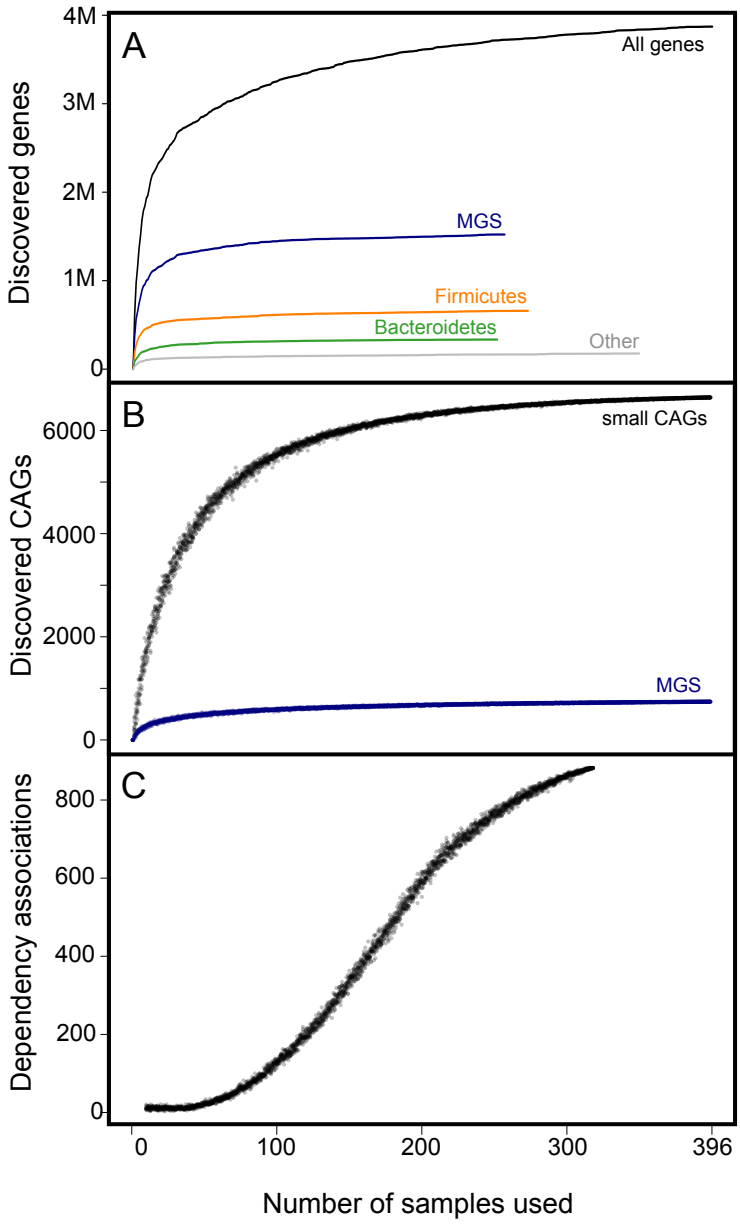
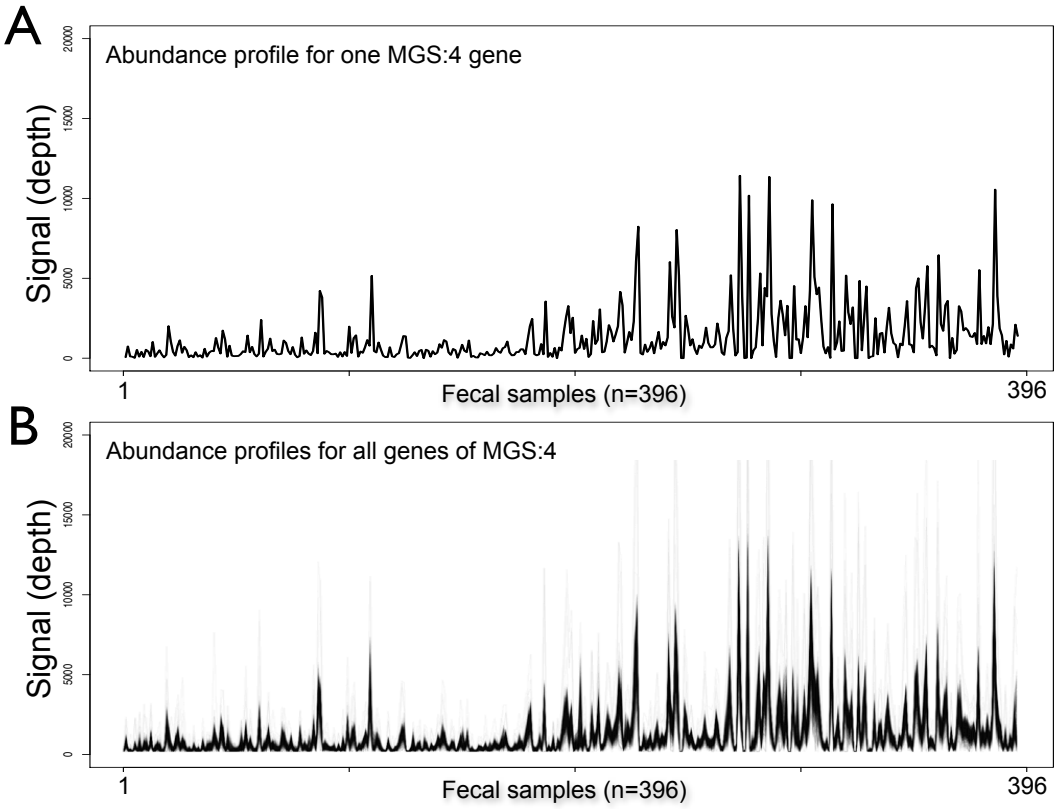
SUPPLEMENTARY FIGURES

Supplementary Figure 1



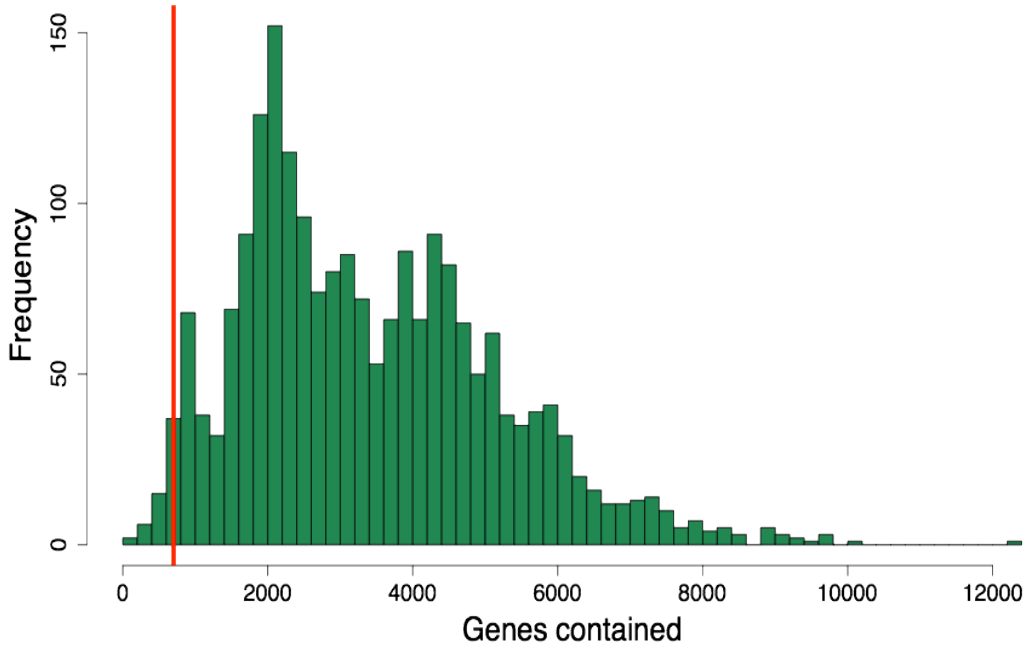
Accumulation curves. A) The five curves show the count of genes for the different types. The last sample, on average, discovers only 584 new genes or 0.02%. Reordering the samples resulted in almost identical cumulative curves (not shown). B) The number of small CAGs (semitransparent black) or MGS (semitransparent blue) found three or more times in random subset of samples of the indicated sample sizes (x-axis). C) The number of significantly dependency associations identified in a random set of samples (from independent individuals) of the indicated sample sizes (1 – 318). In B) and C) 10 independent random drawings were made for each sample size.

Supplementary Figure 2



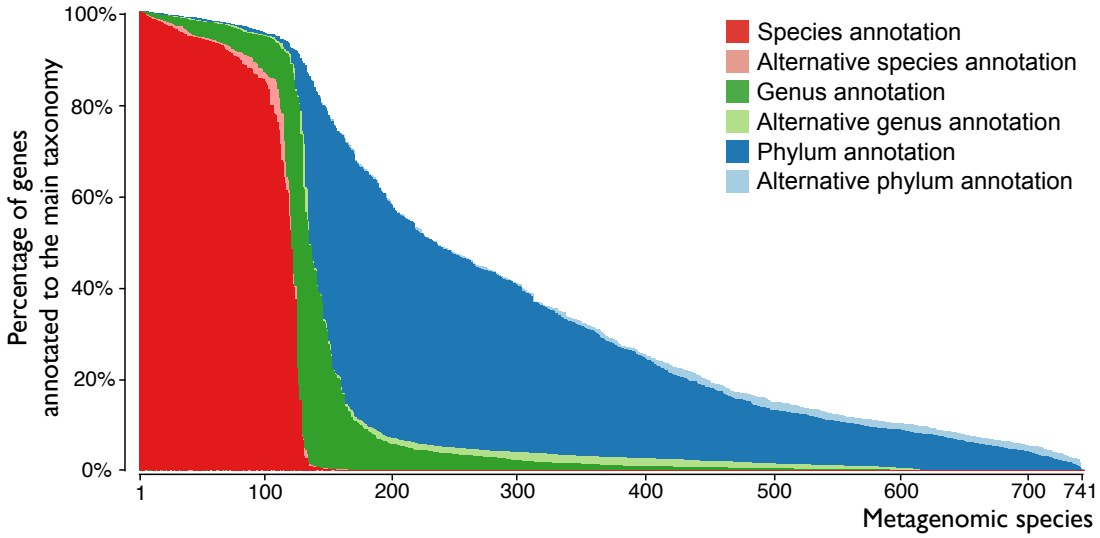
Gene profiles of MGS:4. A) The abundance profile of a single gene from MGS:4 (*E. coli*) across 396 samples. B) The abundance profile for all of the 3,523 genes of MGS:4 (shown as 3,523 semi-transparent lines). The median Pearson correlation coefficient between the abundance profiles of the MGS:4 genes was 0.98.

Supplementary Figure 3



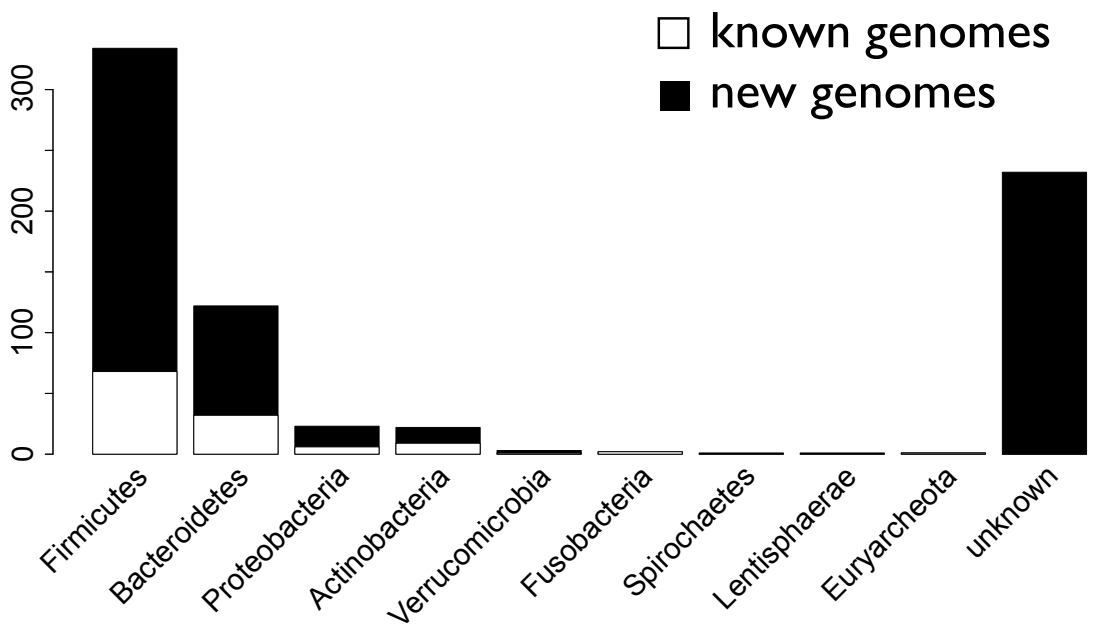
Number of genes encoded by complete prokaryote genomes. The vertical red line indicates the 700 genes threshold between MGS and small CAGs. Gene numbers from all complete prokaryotes in the NCBI genome browser (http://www.ncbi.nlm.nih.gov/genome/browse/) are shown.

Supplementary Figure 4



Bar plot showing the taxonomical consistency of the MGS. The percentage of genes with the most common gene-wise taxonomical annotation for the given MGS is indicated in red, green and blue for species, genus and phylum level annotation, respectively. The percentage of genes annotated to an alternative species, genus or phylum is indicated in light-red, light-green and light-blue, respectively. The area above the bars indicates the percentage of genes without taxonomy annotation. On average, only 1.8% of the genes in an MGS are more similar to alternative species and 518 MGS have no species level similarity to any previously sequenced genome. For the remaining genes, no taxonomical assignment at the indicated level was found. Species, genus and phylum level taxonomical annotation was defined as best sequence match with 95%, 85% and 75% identity over ≥ 100 bp (for details see Methods and Supplementary Data 2).

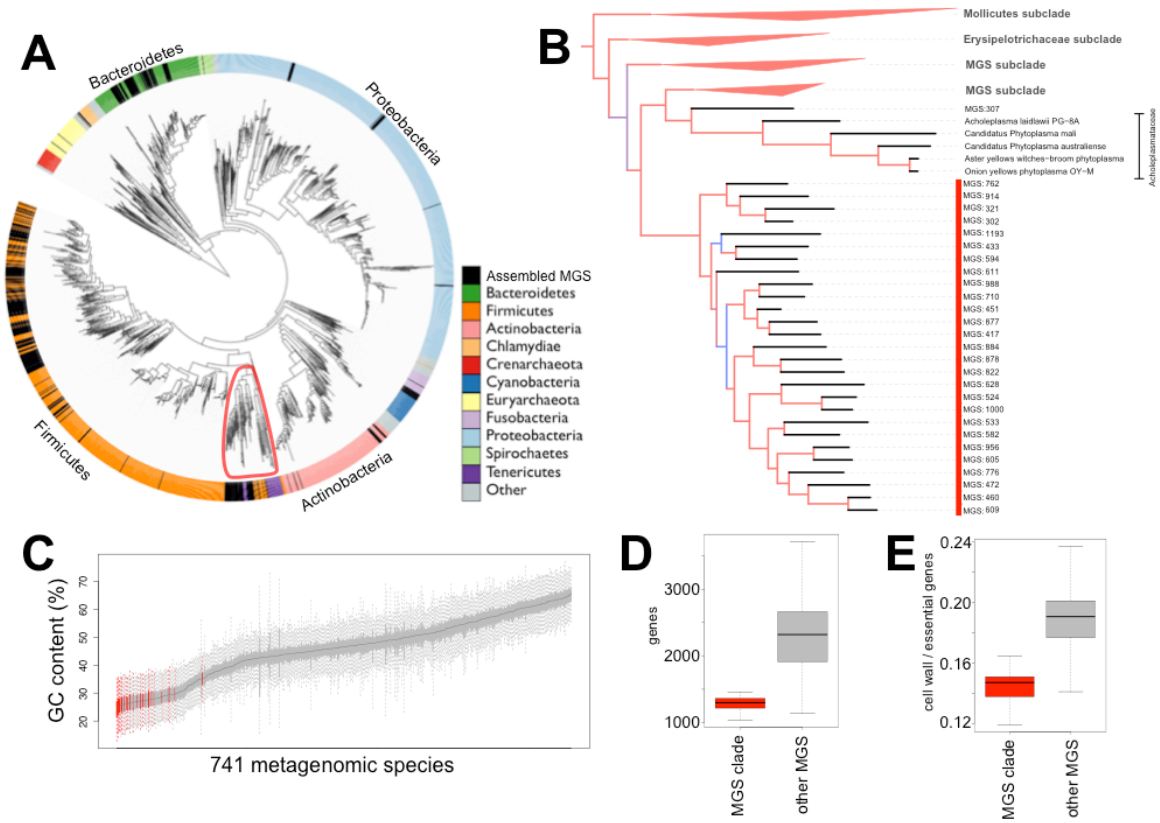
Supplementary Figure 5



Bar plot showing the number of known and previously unsequenced MGS for the indicated phylum. The MGS were assigned to a phylum if 90% of the genes annotated (best hit >75% identity over ≥ 100 bp) indicated the same phylum and more than 100 genes were annotated (for details see Supplementary Data 2).

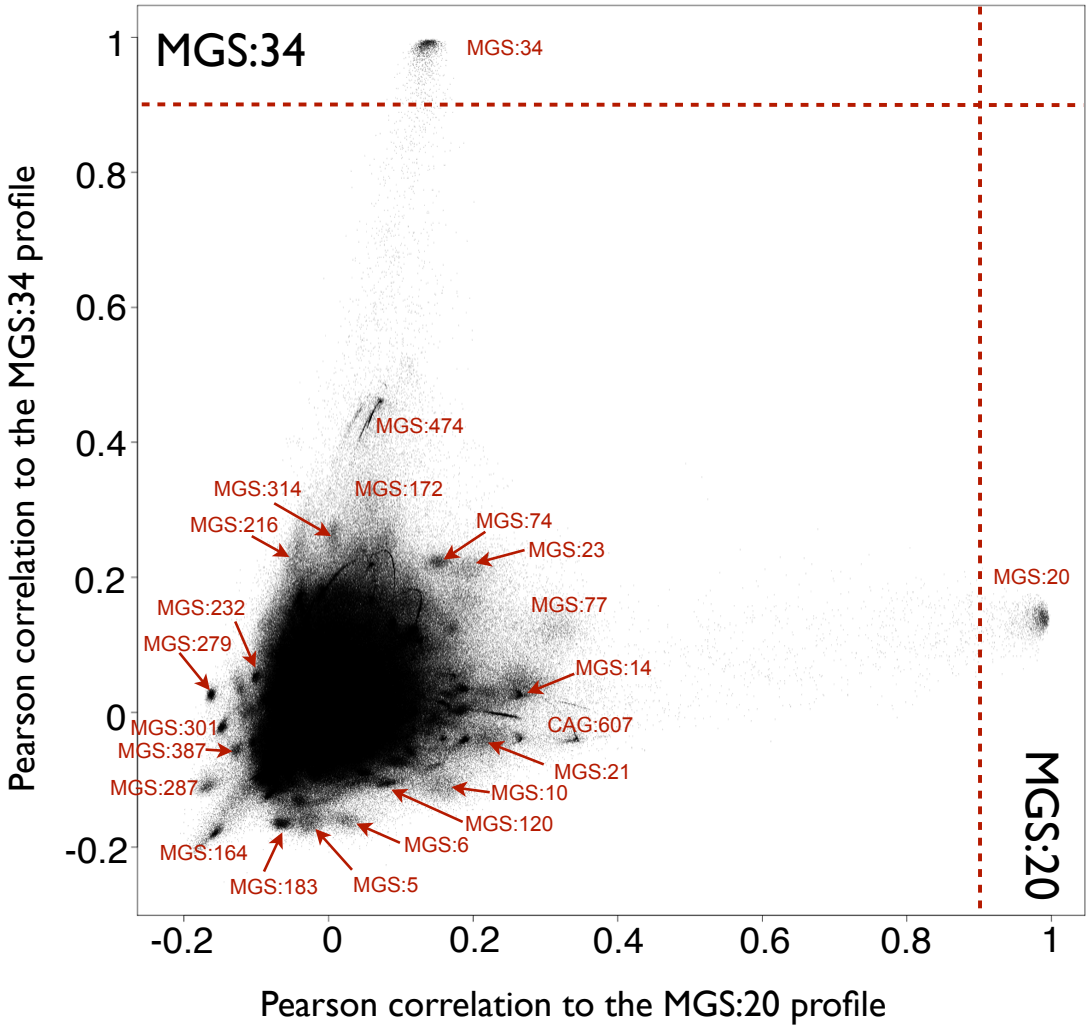
Nature Biotechnology: doi:10.1038/nbt.2939

Supplementary Figure 6



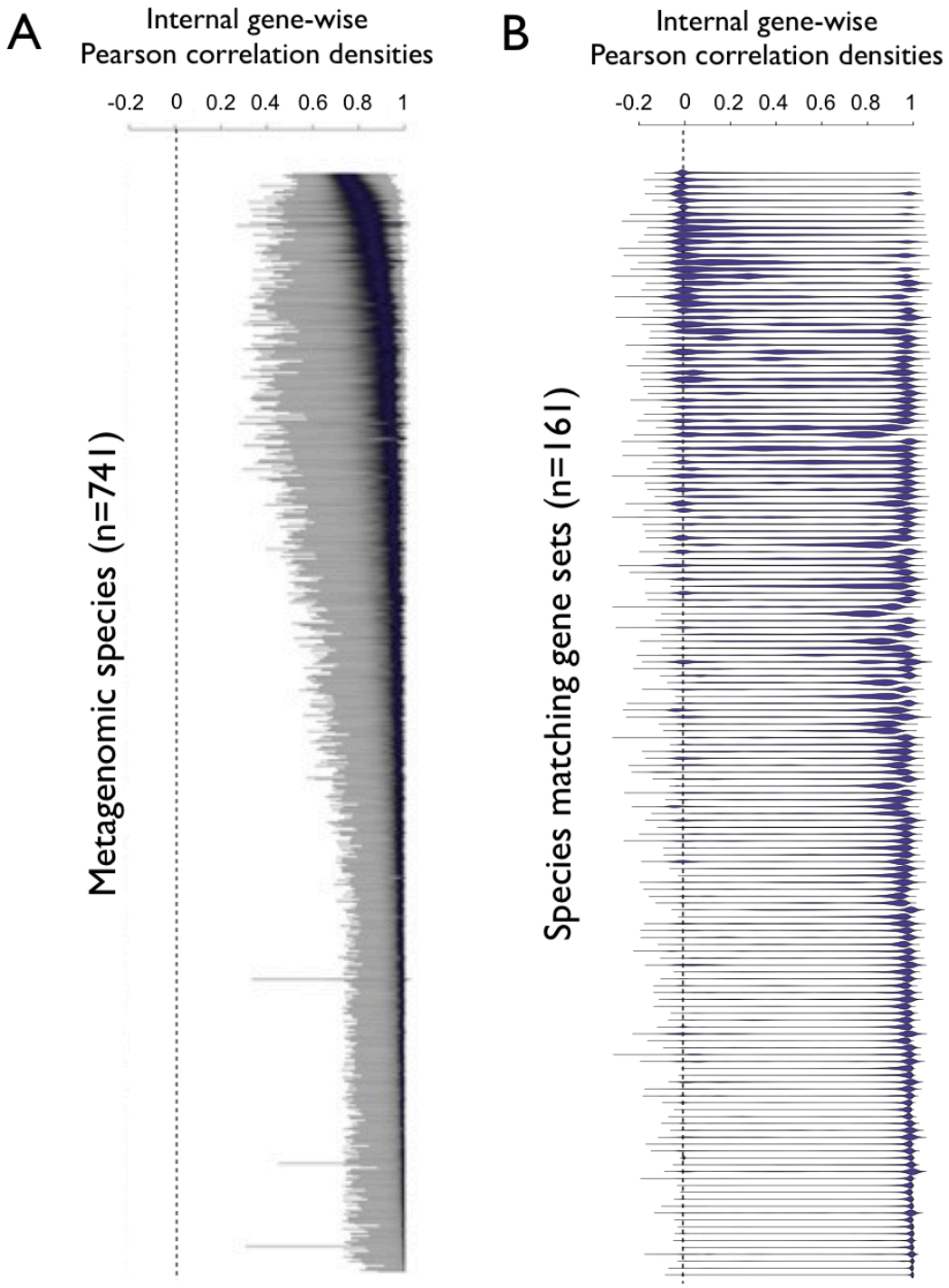
Phylogenetic analysis of MGS augmented assemblies. A) Phylogeny of 337 assemblies (passing 5 or more HMP criteria) plus 1,637 reference genomes including 296 HMP microbiome gastrointestinal tract reference genomes[51]. The coloured ring shows the taxonomy of the reference genomes: green: Bacteroidetes, orange: Firmicutes, light pink: Actinobacteria, light orange: Chlamydia, red: Crenarchaeota, dark blue: Cyanobacteria, yellow: Euryarchaeota, light purple: Fusobacteria, light blue: Proteobacteria, light green: Spirochaetes, purple: Tenericutes and in black: CAG assemblies. The phylogenetic tree was created using the approximate maximum likelihood method implemented in FastTree on an alignment of 40 marker proteins, and visualized using ITOL[45]. The clade marked by the red ellipsoid is shown in B. B) Sub-tree of A, containing the *Tenericutes*, some *Firmicutes* and a clade of 27 CAGs (indicated with a red bar). The branches are coloured by bootstrap support values, where red shows values of 0.95 or higher and blue below 0.95. The MGS clade only consists of MGS augmented assemblies and forms a sister group to the family Acholeplasmataceae (class: Mollicutes). The assembly quality of the species in this clade is comparable to other MGS augmented assemblies (Supplementary Data 3). C) Box-plot of the distribution of gene-wise GC content of MGS. CAGs belonging to the clade shown in B are indicated in red and demonstrate low CG content. D) Box-plot of the gene content of CAG assemblies for the indicated groups. E) Box-plot showing the ratio between the number of cell wall and other essential genes[35] for the assemblies as indicated (Wilcoxon rank sum test, $P = 6 \times 10^{-16}$).

Supplementary Figure 7



Scatter plot showing the Pearson correlation coefficients between the abundance profile of the 3.9M catalogue genes (points) and the abundance profile of MGS:20 (x-axis) and MGS:34 (y-axis). Genes belonging to MGS:20 (n = 2119) and MGS:34 (n = 1799) are defined as genes with correlation coefficients exceeding 0.9 (Pearson) on the x and y axis, respectively (see Methods). In addition to the axis defining MGS, several other MGS are visible as distinct gene clouds at the periphery of the main gene cloud. The IDs of the most visible MGS are indicated.

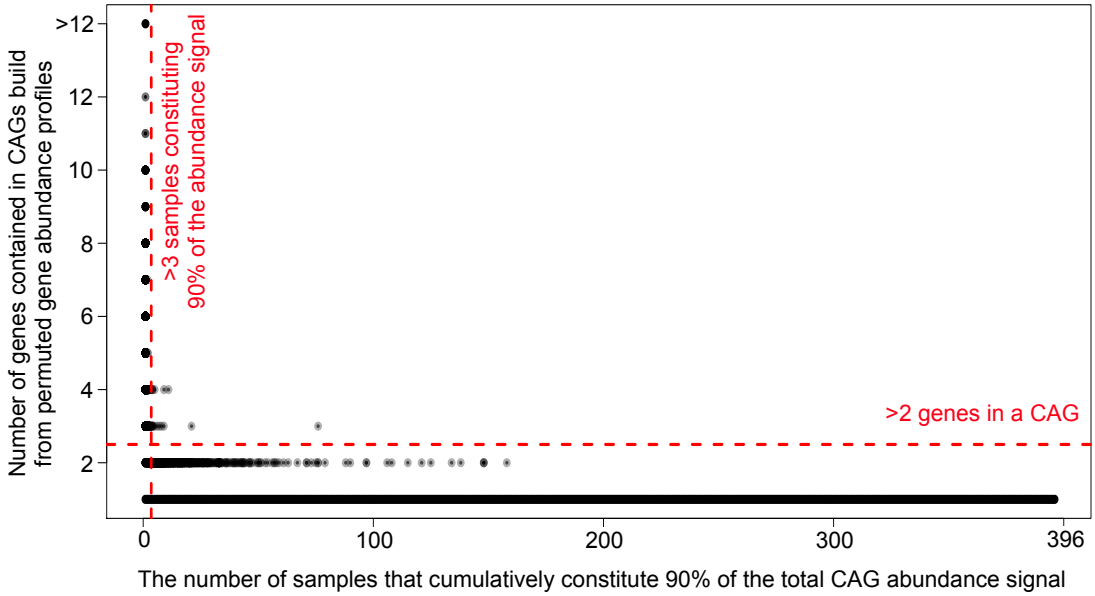
Supplementary Figure 8



Violin plots showing the densities of the internal gene abundance Pearson correlation coefficients for gene sets defined by A) MGS and B) 'reference species gene sets'. The thickness of the horizontal blue 'violins' (lines) indicates the densities of the distribution of Pearson correlation coefficients between the genes within a given gene set. The 'reference species gene sets are defined as sets, that share species level taxonomical assignment by sequence match to a reference genome (best hit, 95% identity over 100 bp or better). The horizontal scale is the same in the two plots. The 'reference species gene sets' and MGS are ordered vertically by the median PCC.

Nature Biotechnology: doi:10.1038/nbt.2939

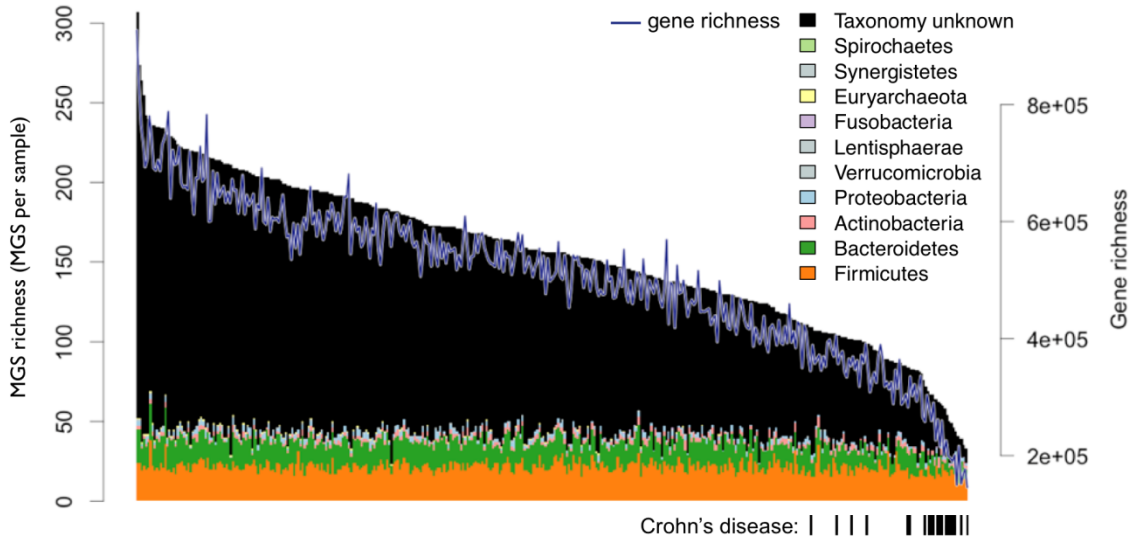
Supplementary Figure 9



Canopy clustering on permuted abundance profiles. The result of an exhaustive co-abundance binning of a gene-wise shuffled abundance matrix is shown. The size (number of genes) and minimal number of samples that constitute 90% of the total abundance signal from the resulting 1,840,781 random CAGs are shown. Only 18 CAGs escape the QC filter indicated with red dashed lines. All of these contained 3 or 4 genes and were observed in a few samples. 1,539,760 of the random CAGs contained 1 gene and 799 contained more than 12 genes. For all of the latter 90% or more of the abundance signal originated from only one sample. The estimated number of randomly occurring CAGs in the non-permuted canopy clustering (*i.e.* the real data) was very low and only expected among the rare and very small CAGs (FDR ~10% for CAGs with 3 or 4 genes).

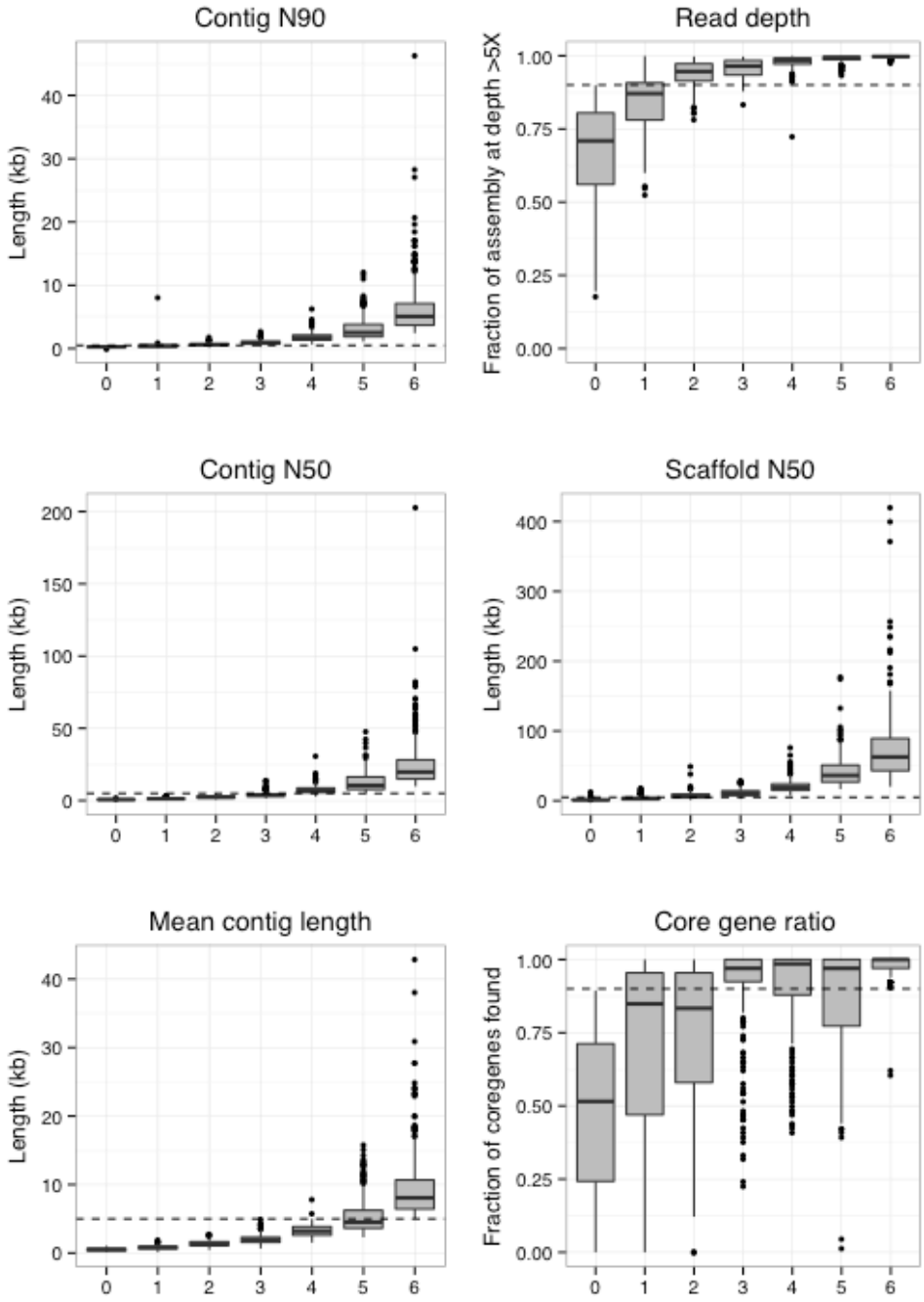
Nature Biotechnology: doi:10.1038/nbt.2939

Supplementary Figure 10



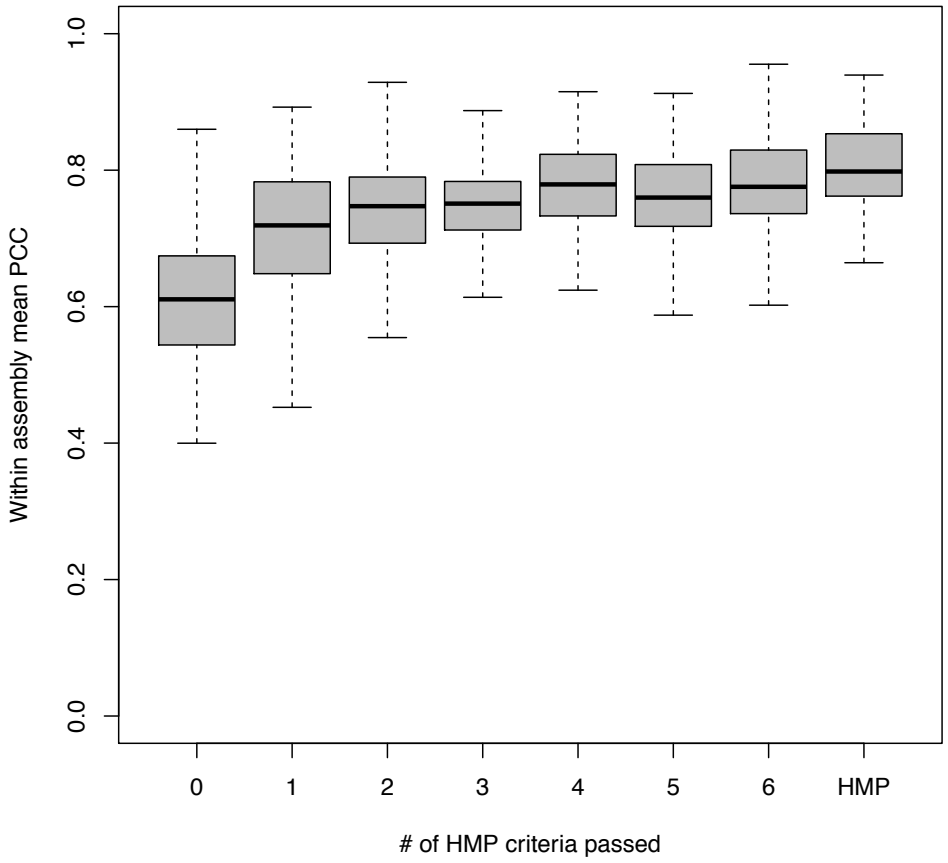
MGS richness across the 396 samples. The height of the bars indicates the number of MGS found in each sample (left axis) and the bar colour shows the phylum level taxonomy of the MGS that represents known species (range: 25 to 81, mean: 50). Black indicates the number of MGS without species level taxonomical annotation. The blue line indicates the sample-wise gene richness (right axis). The PCC between the MGS richness and the gene richness is 0.96 and only 0.55 for the taxonomically known species. Rectangles below the bar plot indicate samples from individuals with Crohn's disease.

Supplementary Figure 11



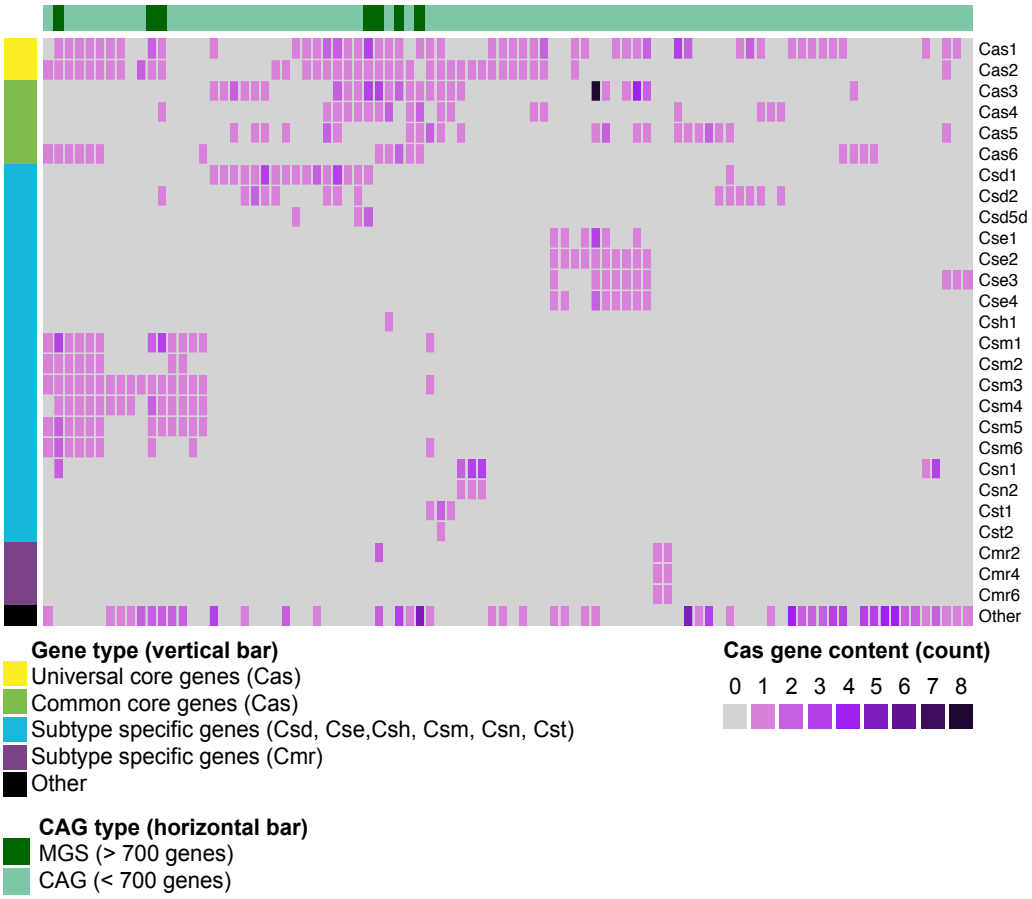
Statistics of the 741 MGS augmented assemblies and visualization of HMP high quality draft genome criteria. The assemblies are divided by how many HMP criteria they pass (x-axis) where passing six criteria equals a high quality assembly. The horizontal dashed lines represent the HMP thresholds for the particular criteria. The lower and upper hinges correspond to the 25th and 75th percentiles, the whiskers represents the 1.5 * Inter-Quartile Range (IQR) extending from the hinges and the dots represents outliers from these. The two assemblies in the "Core gene ratio" panel that pass six criteria but only identified 60% of the core genes are archaeal organisms and they pass the archaea core gene ratios criteria.

Supplementary Figure 12



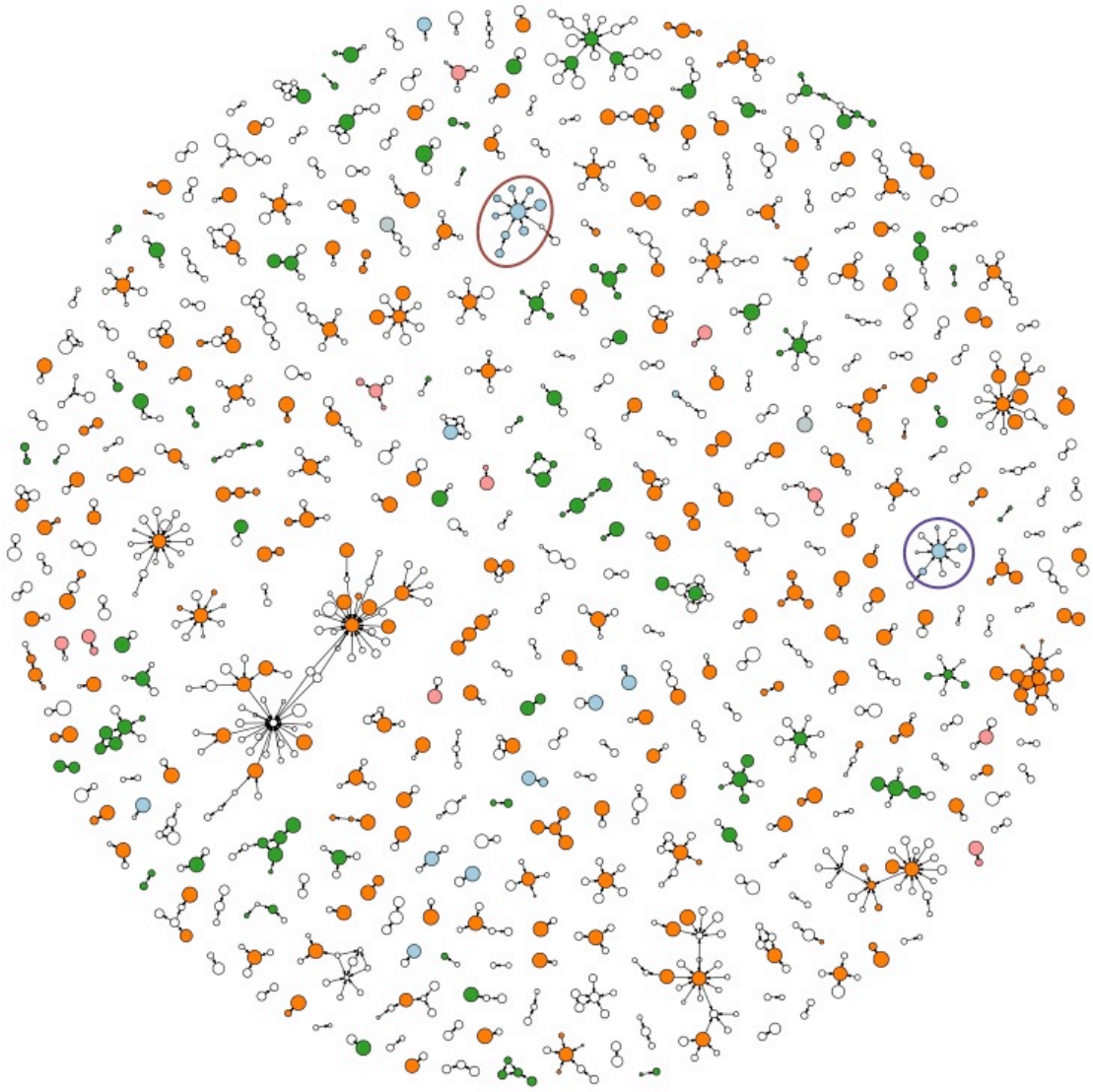
Intra-assembly correlation of Tetra Nucleotide Frequency z-scores (TNF-z) to assembly-specific TNF-z median profiles. The TNF-z profile of all scaffolds within an assembly was correlated using the Pearson Correlation Coefficient (PCC) to the median TNF-z profile of the particular assembly. The figure shows the distribution of assembly mean PCC binned by the number of HMP criteria that the particular assemblies passed (0-6, 6 equals high quality draft) and the 296 HMP gastrointestinal tract reference genomes. The high quality assemblies show similar average PCCs to their median profile as the HMP reference genomes, indicating a similar coherency of the MGS high quality assemblies as the HMP reference genomes created using standard growth, sequencing and assembly techniques.

Supplementary Figure 13



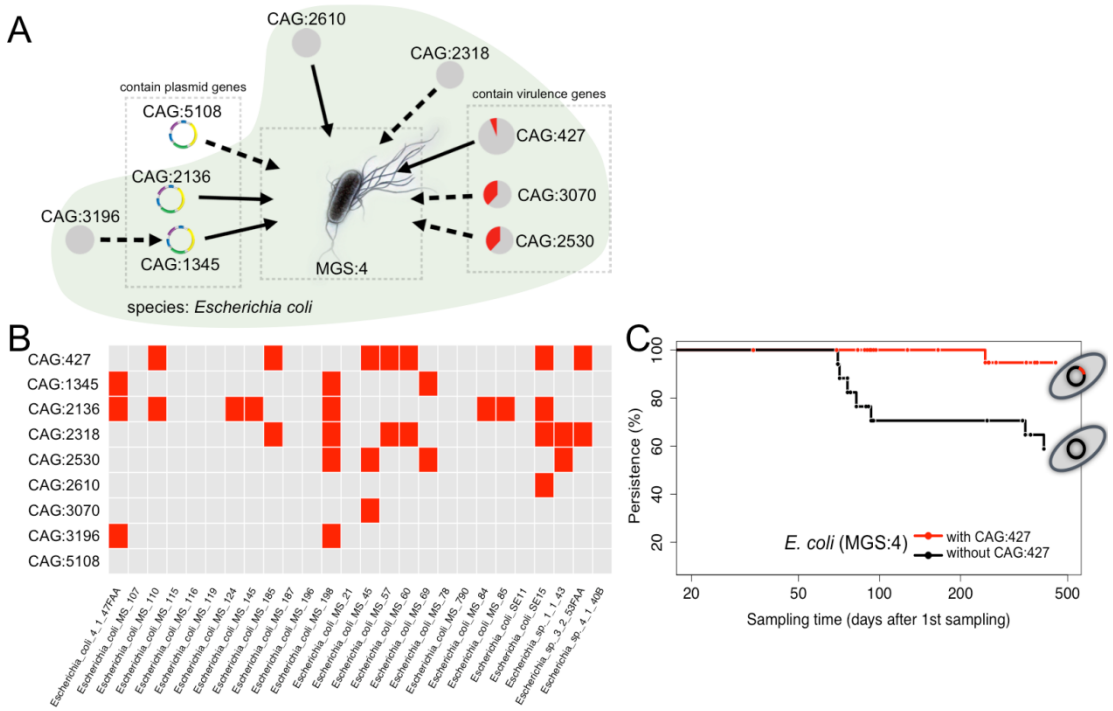
Heatmap showing CRISPR-associated (Cas) genes annotated for MGS and CAGs that were found enriched for CRISPR related genes. The rows show the occurrence of Cas genes in 83 Cas enriched MGS and CAGs (columns). The colour coding at the left corresponds to the subtypes of Cas genes and the colours on top indicate the CAG type. The CAGs cluster according to subtypes of Cas genes they contain.

Supplementary Figure 14



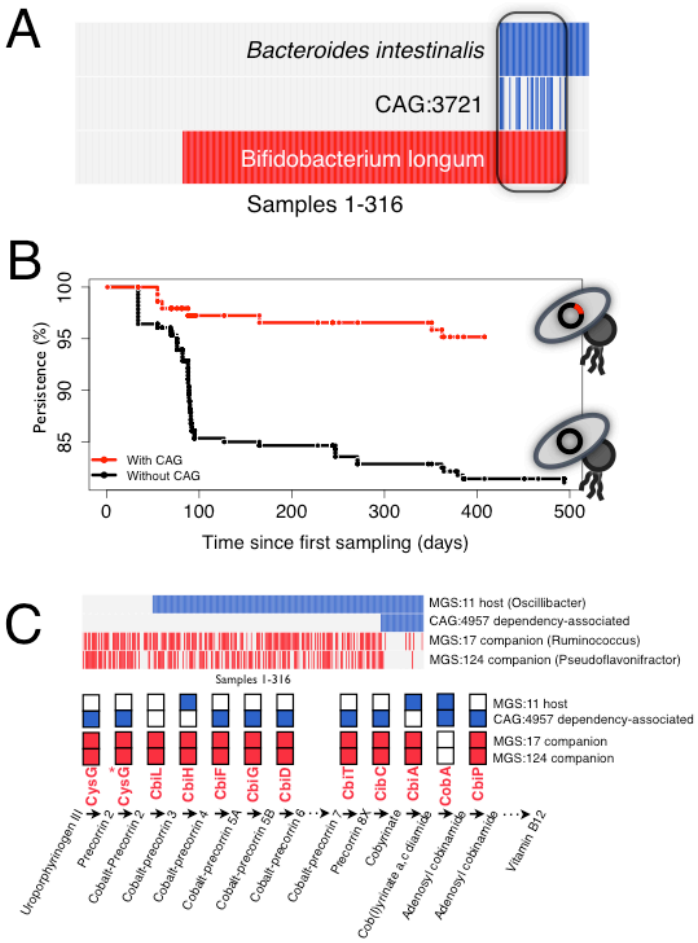
A global dependency-association network. The network shows 882 highly significant ($P < 10^{-10}$ after Bonferroni correction, Fishers exact test) and directional dependency-associations among 287 MGS and 918 small CAGs. Arrows indicate the dependency-associations among CAGs (circles). The size of the circles indicates the number of genes in a specific CAG and the phylum level gene annotation is indicated by colour (green: Bacteroidetes, orange: Firmicutes, blue: Proteobacteria, pink: Actinobacteria). The blue circle indicates the S. *wadsworthensis* (MGS:135) centred sub-network shown in Fig. 5b and the red circle the *E. coli* (MGS:4) centred sub-network in Supplementary Fig. 15.

Supplementary Figure 15



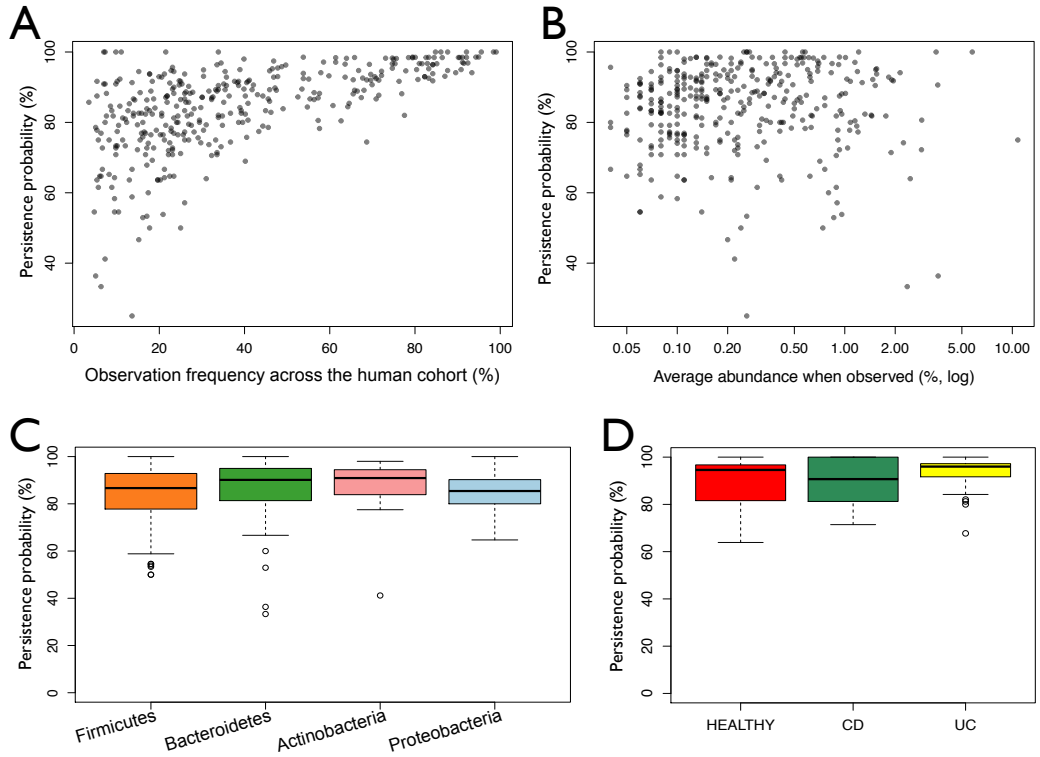
E. coli dependency-associated CAGs. **A**) The dependency-association sub-network centred around *E. coli* (MGS:4). Arrows show directional dependency-associations. The green background colouring indicates CAGs dominated by genes with species level similarity to *E. coli*. The proportion of virulence genes is indicated by pie charts for CAG:427, CAG:3070 and CAG:2530. Together, these three CAGs contain 15% of all the virulence genes found in the entire gene catalogue. CAG:5108, CAG:2136 and CAG:1345 are significantly enriched for plasmid genes. **B**) Identification of the nine MGS:4 dependency-associated CAGs across 25 HMP reference genomes of *E .coli* isolated from the human intestinal tract[51]. Red rectangles indicate a sequence match with >90% identity over >90% of the gene length between the indicated CAG and HMP reference genome. **C**) The cumulative persistence of (MGS:4) with or without the dependency-associated CAG:427 as observed from longitudinal samplings of 34 individuals. Points indicate sampling time of the second sample (in days) relative to the first sampling from the same individual.

Supplementary Figure 16



Adaptations to the co-existence or absence of a companion species. **A**) Sample-wise detections of *Bacteroides intestinalis* (MGS:315) and CAG:3721 shown in blue bars and the occasional companion species *Bifidobacterium longum* (MGS:69) in red bars. CAG:3721 is significantly associated to the co-existence of the two species ($P = 2 \times 10^{-9}$). **B**) Interval censored Kaplan-Meier curves showing the cumulative persistence of species that live in co-existence with a companion species is shown as a function of dependency-associated CAGs. Points indicate sampling time (in days) of the second of two longitudinal samples relative to the first sampling. The curves show the joint observation of 18 inter-species relationships across 73 individuals, where a CAG coincide with substantially increased persistence of the hosting MGS. The annual effect of carrying a positive CAG was 29% as estimated by Bayesian modelling (95% credible interval: 17 to 41 percentage points). **C**) Top panel shows the sample-wise detections of the host (MGS:11, *Oscillibacter*), the dependency-associated CAG:4957 and the two companion species (MGS:17, Ruminococcus like and MGS:124, Pseudoflavonifractor like) as indicated. The detection of CAG:4957 is significantly anti-correlated to the detections of the companion species (MGS:17: $p = 0.0007$ and CAG124: $p = 0.002$). Lower panel show the genetic potential for the anaerobic corrin ring part of the Vitamin B12 biosynthesis pathway[32] for the indicated CAGs as filled boxes. The precorrin 2 to cobalt-precorrin 2 step (marked with *) may be catalysed by both CysG and CbiK. Enzymes catalyzing steps between cobalt-precorrin 6 and 7 lagging experimentally verification are not shown[32]. A possible role for CAG:4957 is to compensate for the biosynthetic potential of the companion species in their absence.

Supplementary Figure 17



Persistence probabilities of MGS. A) Persistence probabilities (estimated from the 2 × 73 re-sampled individuals) as a function of the observation frequency across 318 independent human samples. B) Persistence probabilities as a function of the average abundance across the samples where the MGS was detected. C) Box-plot showing the persistence probabilities for MGS assigned to the four main phyla. D) The MGS persistence probabilities across the patient groups: Healthy, Crohn's disease (CD) and ulcerative colitis (UC). Persistence probabilities were estimated for MGS observed in 5 or more individuals.

SUPPLEMENTARY NOTES

Supplementary Note 1: The majority of 'reference species gene sets' demonstrates incoherent gene abundance profiles

32% of the 3.9 M gene catalogue could be assigned taxonomy at phylum level by similarity to known microbial organisms (best hit with over 75% identity over 100 bp). The majority of these resemble Firmicutes and Bacteroidetes genes (57% and 28%, respectively). However, only 10% of the catalogue genes could be assigned taxonomy at species level (best hit with over 95% identity over 100 bp). Using this criterion, 161 'reference species gene sets', with more than 700 genes assigned to the same species, were defined. Here, these gene sets serve as representatives of a reference genome based structuring of the metagenomics data. Correlation analysis of the abundance profiles of the genes within the 161 'reference species gene sets', show that many of these gene sets do not behave as a coherent entity (Supplementary Fig. 8B). Hence, 88 of the 'reference species gene sets' have significant sub-populations of genes that do not correlate with the bulk of the genes, i.e. 25% or more of the genes have a Pearson correlation coefficient (PCC) < 0.5. The genes within 56 of these sets could be identified as members of multiple distinct MGS, which in turn include additional genes without similarity to these reference genomes. For 12 'reference species gene sets' almost no internal gene correlations could be found (less than 25% of the genes have PCC of 0.5 or more). The highly inconsistent 'reference species gene sets' are found across the major phyla: Firmicutes, Bacteroidetes, Proteobacteria and Actinobacteria, but the Bacteroidetes gene sets stand out as particularly incoherent with an average within gene set PPC of 0.5. At genus level Bacteroides, Faecalibacterium and Prevotella are the most incoherent groups, all with an internal PCC average below 0.5. This level of inconsistency is problematic both because the annotation does not reflect the biological organization of the system and because association between inconsistent 'reference species gene sets' and clinical data, may be misleading or fail to identify an underlying organism. Further work will be required to clarify the reasons for the inconsistency, but we suggest that CAGs may be more suitable than the homology-based gene sets both for understanding the gut microbial communities and their association to health and disease.

Supplementary Note 2: The MGS to small CAGs distinction

The distinction, between the 741 MGS with more than 700 genes and the smaller CAGs, should roughly identify the CAGs that represent cellular microbial species. This division is not clear-cut and some core genomes may fall below and some clone specific variants and mobile elements may exceed the 700-gene threshold. The threshold was chosen based on a combination of expectations and observations. While some bacterial genomes have been reported to contain very low number of genes[52] most known bacterial genomes encode more than some 1,000 genes, whereas most phages and plasmids have less that 500 genes (Supplementary Fig. 3). Prior knowledge therefore suggests a threshold somewhere between 500 and 1000 genes[53]. In addition, three independent observations in our data suggest a threshold around 700 genes. First, the observed bimodal size distribution of the CAGs shown in Fig. 2a narrows around 700 genes and therefore suggest a natural distinction around 700 genes. Second, a significant enrichment for genes essential to bacterial life, detected by homology to the *Bacillus subtilis* essential gene set[35] was found primarily in CAGs with more than 700 genes (Fishers exact test: $P << 1 \times 10^{-100}$). Finally, if small CAGs represent genetic heterogeneity of biological organisms or bacteriophages they should

depend on an organism to proliferate. The number of small CAGs with statistically determined dependency-associations to MGS is highest at the 700-gene threshold and drops for higher thresholds. With this threshold the odds ration for small CAGs to MGS dependency-associations is 12.7.

Supplementary Note 3: MGS profiles are coherent in an independent sample series

To demonstrate that the MGS behave as general biological entities the coherence of abundance profiles for the MGS genes was investigated in 115 independent human stool samples that were not used for the clustering[17]. In this independent sample set the median gene-wise Pearson correlation coefficient for intra MGS gene profiles was as high as 0.98. Hence, the MGS indeed appear to be general descriptors of coherent genetic entities across similar microbial systems.

Supplementary Note 4: The MGS profiles are distinct and robust

Individual gene abundance profiles of a typical MGS are highly coherent and distinct from the profiles of the genes not included in that particular MGS. Consequently, relaxing the gene inclusion criterion from PCC > 0.9 to PCC > 0.8 relative to the median profile of the MGS only extends the MGS gene set on average by 5%. Similarly, raising the inclusion criterion to PCC > 0.95 reduces the number of genes included by 17 % on average. Hence, the co-abundance based clustering is robust to changes in parameters that determine the cluster boundary and importantly the MGS are separated from other genes. This feature is so strong that more than 30 MGS stand out as distinct and dense gene clouds in a two-dimensional Pearson scatter-plot (Supplementary Fig. 7), even when the plan/dimensions of the plot is not targeting the separation of these specific MGS. In addition, even the dependency-associated CAGs are clearly distinct from their hosts, thus on average the PCC between the host and the dependency-associated CAGs is 0.46 (+/- 0.2).

Supplementary Note 5: The MGS richness is in concordance with gene richness and indicative of Crohn's disease

The number of different species, commonly known as the species richness, is an important measure of an ecological system, partially because it is believed to reflect the general health and stability of a system[17]. Obviously, this measure depends on proper detection of species in the ecological system, and we propose the MGS count as an estimate. Across our cohort of stool samples this number ranges from 33 to 307 with a mean of 155. The number of MGS with species level annotation (shown as coloured bars in Supplementary Fig. 10) is more constant across the cohort than the number of unknown species (shown with black bars). Importantly, sample-wise gene-richness correlates significantly better to the MGS richness estimate (PCC = 0.97), than to the taxonomically known species richness (PCC = 0.55) or to richness estimates based on reference genome detection (see Methods, PCC = 0.52). In addition, MGS richness significantly associates the occurrence of Crohn's disease (t-test, $P = 4\text{x}10^{-15}$), much stronger than does species richness derived by sequence similarity to known reference genomes (Crohn's disease: $P = 0.09$). While association between species richness and Crohn's disease has been reported before[54], the MGS richness measure clearly enhances the correlation.

Supplementary Note 6: Common species comprise genes for protection against ROS and Vitamin B_{12} metabolism

Comparisons of MGS between any pair of individuals show an average overlap of 50% (±12%). A substantial part of the shared MGS belongs to a core set of 31 MGS that were detected in at least 90% of the samples and together they account for 25% of the abundance signal. Of these, 18 have clear similarity to taxonomically known species, and, hence, there is an overrepresentation of taxonomically known species among the core MGS (odds ratio 3.0). The most common MGS across the sample series are the *Blautia wexlerae* (MGS:9) and *Bacteroides vulgatus* (MGS:6) which were found in 395 and 392 of the 396 samples, respectively. *B. vulgatus* (MGS:6) is in addition the most abundant species across the samples, matched by on average 6% of the mapped reads and is the dominating species in 58 samples.

A number of orthologous groups (eggNOG)[30] are found significantly more frequently among common species than in less common species (Wilcoxon rank sum test, $P < 1x10^{-15}$, Supplementary Data 8) and suggests a set of functions important for bacterial existence in the human gut. These include enzymes for protection against reactive oxygen and vitamin B_{12} biosynthesis.

Supplementary Note 7: Only very few contigs may be the result of chimeric assemblies

For the assembly of the data we used the MOCAT pipeline[22] which performs a revision of the initial assembly that specifically tries to identify and break chimeric contigs (see Methods). However, to assess the rate of potential chimeric contigs we re-mapped all reads to the assemblies using bwa (bwa-0.7.5)[36] and calculated the number of bases per contig that was not bridged by properly paired read pairs. The properly paired read-pairs are read-pairs that map in the expected orientations and with the expected insert length. Absence of these in regions of an assembly indicates a point of miss-assembly and a potential chimera in metagenomic assemblies. Across all contigs we found that only a small fraction (0.0058) had one or more bases that were not covered by properly paired reads, pointing to a very low contig chimera rate. Among the cross dependency-associated CAG contigs we only found 31 of 7,966 (0.0039) without proper paired read coverage.

Supplementary Note 8: 21% of the abundance-uncorrelated genes can be linked to the MGS

1.2M catalogue genes, with abundance profiles exceeding the filtering criteria (more than 3 samples must constitute 90% of the total abundance signal) did not segregate with any CAG. The detection rate of these abundance uncorrelated genes was however comparable to that of the correlated genes, as their sequence coverage and re-detection rate in paired samples from the same individual, were similar to that of the correlated genes. Although, the abundance-uncorrelated genes on average were detected in significantly fewer samples than the correlated genes (mean: 50 and 93 samples, respectively) the bulk of these were detected in a sufficient number of samples to allow these to be segregated, if they were correlated to other genes.

These abundance-uncorrelated genes are in contrast to the clustered genes significantly underrepresented in essential genes. Interestingly, we found that, genes involved in antibiotic resistance, with the exception of vancomycin resistance, had distinct single gene abundance profiles. This is in line with the fact that most antibiotic genes, except vancomycin resistance genes, are known to single-handedly

provide antibiotic resistance and suggests that some genes may be highly dynamic and perhaps are best understood non-contextually, at the single gene level.

21 % of these abundance-uncorrelated genes, however, can be linked to an MGS by shared sequence contigs in at least one sample, indicating that some of these genes may be clone or strain specific genes of the species. In support of this, these contig-extended genes are likewise significantly underrepresented in essential genes (encoding *Bacillus subtilis* COGs[35], Fisher's exact test $P = 0.002$).

For instance the very common *Bacteroides vulgatus* (MGS:6) comprises 2,271 genes but can be contig extended to include additional 326 genes, across all samples. The average sample however, only comprises 161 of these genes, and the abundance profiles of these genes show little correlation to the *B. vulgatus* (MGS:6) profile (mean PCC = 0.3). This abundance profile inconsistency of the contig extended genes may to some extend resemble the inconsistencies observed for 'reference species gene sets' and as such illustrate the difference between example based gene sets and CAGs.

Supplementary Note 9: Co-existence associated CAGs

For the microbial species the presence of other companion species in the community may be a major factor to which they may adapt. Such adaptations may be indicated by significantly increased occurrence of specific dependency-associated CAGs in samples where a companion species is also found. A subset of 66 dependency-associated CAGs does exactly that (Supplementary Fig. 16A shows an example) and these are therefore candidates for adaptations to co-existence. In 18 of these relationships the dependency-associated CAGs coincide with significantly enhanced persistence probabilities of the hosting MGS when found jointly with a companion species (Supplementary Fig. 16B, Supplementary Data 12). The companion species on the other hand appear to be only marginally affected by the presence of the dependency-associated CAG, with only a slight but insignificant increased persistence.

This set of co-existence associated CAGs is very significantly enriched in genes encoding parts of the TonB complex that is important for extracellular sensing and that in *Pseudomonas aeruginosa* has been associated with biofilm formation and quorum sensing[55]. As an example, the *Odoribacter splanchnicus* (MGS:225) dependency-associated CAG:3500 contains genes that encode a 'TonB-dependent receptor plug protein', a 'two-component sensor histidine kinase' and a 'transcriptional response regulator rprY', proteins that have been reported in signal transduction, chemotaxis and quorum sensing[55,56]. Furthermore, the set of co-existence associated CAGs are enriched in the broad-spectra acriflavin resistance proteins and conjugation-coupling factor proteins.

In contrast, to the co-existence associated CAGs, another set of dependency-associated CAGs were significantly absent in samples where specific companion species were found (Supplementary Fig. 16C and Supplementary Data 12). The MGS:11 (Oscillibacter) has several such associated CAGs. In particular, CAG:4957 encodes 16 proteins that are orthologous to proteins in two companion species (MGS:17, Ruminococcus like and MGS:124, Pseudoflavonifractor like), but not to any proteins in the hosting MGS:11. Seven of these proteins are in the anaerobic corrin ring biosynthesis part of the Vitamin B_{12} pathway (Supplementary Fig. 16C lower panel). Hence a possible role for CAG:4957 is to compensate for the biosynthetic potential of the companion species when they are absent.

Additional references for Supplementary information

51. Nelson, K. E. *et al*. A catalog of reference genomes from the human microbiome. *Science* **328,** 994–9 (2010).
52. Juhas, M., Eberl, L. & Glass, J. I. Essence of life: essential genes of minimal genomes. *Trends Cell Biol*. **21,** 562–8 (2011).
53. Toussaint, A. & Chandler, M. Prokaryote genome fluidity: toward a system approach of the mobilome. *Methods Mol. Biol*. **804,** 57–80 (2012).
54. Manichanh, C. *et al*. Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. *Gut* **55,** 205–11 (2006).
55. Abbas, A., Adams, C., Scully, N., Glennon, J. & O'Gara, F. A role for TonB1 in biofilm formation and quorum sensing in Pseudomonas aeruginosa. *FEMS Microbiol. Lett*. **274,** 269–78 (2007).
56. Wolanin, P., Thomason, P. & Stock, J. Histidine protein kinases: key signal transducers outside the animal kingdom. *Genome Biol*. **3,** reviews3013.1–reviews3013.8 (2002).