SUPPLEMENTARY RESULTS

Physiological response to endurance exercise in highly trained athletes

Since endurance racing causes great physiological stress on the body, not surprisingly, all of the horses showed above-average concentrations for total bilirubin**,** creatine, creatine kinase (CK), aspartate transaminase and serum amyloid A after the race (Suppl Table S11). The elevated creatine level and CK activity likely reflected higher membrane permeability associated with muscle protein leakage to the blood coupled with severe inflammation signaling [1]. Similarly, substantial muscle turnover resulted in the increase in the production of creatine in athletes [2].

The study of the adaptive metabolism response to endurance exercise through metabolomics pinpointed that the concentration of lactate (a proxy for glycolytic stress and disturbances in cellular homeostasis [3]) was significantly increased after the race, as well as the levels of fatty acids from lipoproteins and of certain amino acids, namely alanine, branched amino acids such as leucine, valine and *iso*-valerate, glutamate, glutamine and aromatic amino acids such as tyrosine and phenylalanine (Suppl Table s9). Moreover, ketone bodies were slightly increased after the race (*i.e.,* acetoacetate and acetate; Suppl Table S11). Accordingly, the non-esterified fatty acids and the β-hydroxybutyric concentrations showed similar patterns, indicating increased lipolysis. The elevated level of lipid utilization during endurance exercise was confirmed by the greater mitochondrial β-oxidation activity, reflected by increased blood acylcarnitine concentrations (from 6.06 ± 1.85 to 42.38 ± 13.02 µmol/L; Suppl Table S10).

The gut microbiome contains vast and individual-specific genetic content

We sought to determine the frequency of occurrence of each gene on a sample-by-sample basis. The gene catalog contained 295,948 (1.17%) singletons. On average, 2.5% of the genes in each sample were singletons (standard deviation 30.58%). Moreover, the gene abundances and the percentages of the genes shared among samples indicated that most of the identified genes were at lower abundance/prevalence within the individual samples (Suppl Fig S1).

Supplementary note on the microbial gut catalog taxonomic assignation

Taxonomic inference relying on eggNOG database yielded an annotation for 16,795,097 genes, but only 3,205,480 (12%) could be classified down to the genus level. Significant discrepancies were found however while attempting to cross-validate this orthology-based annotation with available 16S rRNA sequences (Suppl Table S4). Indeed, the most prominent genera were not classified on that basis. For instance, orthology-based inference performed poorly at classifying sequences of the most abundant genera such as *Treponema*, *Clostridium* XIVa, *Prevotella, Ruminococcus* and *Fibrobacter* (Suppl Fig S7), which altogether accounted for most of the gut microbiota composition (56% of 16S rRNA reads), were not classified.

Univariate approach to investigate the metabolic responses between individuals with divergent cardiovascular fitness

Coupled with the DIABLO multivariate method, we applied a univariate approach to the most correlated datasets, that is, dominant microbial phylotypes composition, KOs and CAZy profiles. This univariate method unveiled overall differential changes of 588 genera, 1,451 KOs and 48 CAZy classes (DESeq2, adjusted $p < 0.05$; Suppl Table S16-18, respectively). As with the DIABLO approach, less fit horses were enriched with multiple phylotypes from the

Lachnospiraceae family (Suppl Table S16 and Fig S8a). Since *Lachnospiraceae* have the ability to ferment complex carbohydrates and produce SCFA [4], not surprisingly, the latter microbiomes reflected significantly high abundance of enzymes with catalytic activities of breaking down starches (GH32, GH77) and highly complex and variable plant cell wall polysaccharides (GH3, GH5, GH28, GH31, GH44, GH51, GH73 and GH74; adj. $p < 0.05$; Suppl Fig S8b and Table S18). Contrastingly, more fit individuals showed a proliferation of microbial symbiotic diversity and an unpredictable multi-kingdom community dominated by Proteobacteria (*Thauera, Polynucleobacter, Smithella, Rhodopsudomonas, Desulfatibacillum*), Verrucomicrobia (*Akkermansia, Rubrutakeam, Luteolibacter, Haloferula, Roseimicrobioum, Lentisphaera*) and Planctomycetes phyla (*Pirellula, Anaerohalosphaera, Limihaloglobus, Roseimaritima*), as well as enteric fungi (*Ophiocordyceps, Pseudogymnoascus, Cryptococcus, Trichoderma, Coniochaeta*), algae (*Emiliania and Porphyra*) and archaea (*Methanothrix, Thermococcus, Methanosphaera*; Suppl Table S16). Of note, 360 non-core genera were upregulated in more fit individuals. Supporting the role for host epithelial and mucosal glycans in dictating aspects of microbial community composition and activity, individuals with improved cardiovascular fitness expressed enzymes that cleave animal glycosaminoglycans (PL8, PL12, GH18, GH20, GH29, GH33, GH88 and GH95) coupled with glycosidases that might act in a concerted manner to cleave every carbohydrate substructure in the polymer into fermentable monosaccharide (GH6, GH39, GH54, GH82, GH84, GH99; Suppl Fig S8b and Table S18). Not surprisingly, KO functions related to cellular energy metabolism, such as glycolysis and the TCA cycle, as well as whole-body substrate metabolism, such as amino acid breakdown and fat oxidation were significantly raised in fit participants. Precisely, these microbiomes encoded KOs directly associated with glycan metabolism (K03279, K19715) along with SCFA production, *i.e.,* genes within propanoate metabolism (K00382), methane metabolism (K00202) and pyruvate metabolism (K00029, K00024), as well as amino acid metabolism (K00024, K00128, K01777, K18049, K03852, K22476, K03279, K01744), mitochondrial biogenesis (K03593, K07152) and PPAR signaling pathway (K00029, K01596, K01897; Suppl Table S17). Together with the amino acid metabolism, we observed elevated levels of *iso*-valerate in plasma ($p = 0.01212$, Wilcoxon rank test), suggesting their capacity to ferment protein of dietary host and microbial origin. The top KEGG enriched functional modules in less fit participants were largely redundant and based on transport and metabolism, alongside genetic information processing and signaling.

Thus, univariate analysis supported how microbial rare species might be drivers of specialized functions, like mucus degradation and enhance the adaptations to endurance. On the contrary, this metabolic competitive advantage was not observed in individuals harboring higher abundance of dominant and core Firmicutes with highly CAZymes redundant functionality in the gut.

The metagenome-assembled genomes (MAGs) slightly capture the complexity of the natural population

To uncover whether the recovery of MAGs were appropriate proxies for the whole microbial communities or whether they only captured the most abundant microorganisms, MAGs were constructed from the metagenomic sequencing obtained from the 11 samples described above using the same metagenomic ATLAS pipeline v. 2.4.4 [5].

The mean percent completeness and contamination of the 372 non-redundant MAGs was 83% ± 12% and 1.7% ± 2.05%, respectively (Suppl Fig S9a-c). They were then subsequently classified into taxa using the Genome Taxonomy Database Toolkit (GTDB-Tk) [6]. All 372 MAGs were

classified at the kingdom level (361 to bacteria and 11 to archaea), however, only 289 and 41 MAGs were classified at genus and species levels, respectively (Suppl Fig. S9d and Table S7). Genome sizes ranged from 0.42 to 88.97 Mb (Suppl Fig S9e). Completeness was not significantly correlated with genome size ($r^2 = 0.17$, $p = 0.55$; Fig. 7f). All MAGs together totaled 53,149 contigs (34% of total assembly) with an average assembly N50 value of 4.5 Kb (Suppl Fig S9h) and CG content of 44.49% (Suppl Table S7). The two highest contig NG50s were about 0.233 to 0.459 Mb. Overall, MAGs missed 94.5% and 97.7% of the population core and variable genes of the catalog, respectively, on average.

At the taxonomic level, MAGs did not capture the complexity and diversity observed in the gene catalog, nor mirrored the data obtained from the 16S rRNA gene amplification. As specified above, the MAG repertoire brough read classification rates up to 30% of the data, suggesting that the remaining reads were likely from low-abundance bacterial and archaeal species, difficult-to-assemble genomes, as well as, the fungal, protozoan and viral genomes. That said, our results provide a reference guide for the horse community. We compared our 372 MAGs to the publicly available MAGs in horses to date [7,8]. None of our MAGs show complete identity to any other publicly available MAG in horse, although three MAGs were more than 99% similar to another MAG published in horse, namely MAG1, MAG241 and MAG322 (Suppl Fig S3e).

We then mapped our MAG proteins to the CAZymes database. A total of 103 active enzyme families were identified. The most abundant enzyme classes were the carbohydrate degrading GHs ($n = 52$) and GTs ($n = 34$) (Suppl Table S19). The phylogenetic tree based on genus level showed that most of the MAGs have GHs enzymes, which are a complex and widespread group of enzymes able to degrade polysaccharides (Suppl Fig S10). Bacteroidetes, however, captured greater CAZyme diversity, as they showed an increased number of CAZYmes families and also these families contain more enzymes than others (Suppl Fig S10). KOs profiles ($n = 7,186$) are reported in SuppTable S20.

The association between cardiovascular fitness and gut metagenome revalidated by MAGs

Knowing that the 372 nearly-complete MAGs only represented a small fraction of the global metagenome community, we wanted to assess if they could at least provide key data related to the ecological and genetic mechanisms responsible for the diversity patterns observed between individuals with divergent cardiovascular capacity. Consistent with the dominant phylotypes, samples from individuals with different cardiovascular fitness could be distinguished by MAGs composition (Suppl Fig S11a). That is, the ordination plot based on the β-diversity showed that samples from individuals with similar cardiovascular capacity tended to cluster closely ($p = 0.005$, $R^2 = 0.1543$, PerMANOVA of Bray-Curtis distances), suggesting that the 372 reliably capture the overall structure of the gut microbial community. Conversely, the α-diversity indices comparison indicated that all samples had similar levels of diversity ($p = 0.9212$ for Shannon and inverse Simpson, Wilcoxon rank-sum test) regardless of the cardiovascular capacity. Having found that MAG composition was segregated into two clusters, we established that this compositional dissimilarity did not have any functional capacity (Suppl Fig S11b), probably due to the poor annotations of metagenomic sequencing data at the species level.

We then focused on those MAGs and their functional capacities enriched between individuals with different cardiovascular fitness using the DESeq2 package. The abundance of 52 MAGs was

statistically different between groups (adjusted $p < 0.05$ with DESeq2; Suppl Table S25 and Fig S11c). Changes in MAGs community composition were most pronounced in individuals with improved cardiovascular fitness with increases in the relative abundance of Verrucomicrobia (*e.g., Akkermansia* spp.) and Bacteroidetes, and decreases in the relative abundance of Fibrobacteres and Spirochaetes (Suppl Table S21). Nevertheless, overall CAZymes and KOs profiles were unchanged by group (Suppl Fig S11d) raising awareness that our MAG collection likely underrepresented the whole metagenome and was unable to find distinct metabolic affinities between participants.

SUPPLEMENTARY FIGURES

Figure S1 - Description of the first gut microbial gene catalog in horses: contribution of sequencing depth and sample sources to the gene content

(a) Total contig length (Gb) for each sample; (b) Rarefaction curve showing the number of genes from the total gene catalog that could be discovered with increasing numbers of raw sequencing reads. A rarefaction curve for each individual is depicted; (c) Association of predicted gene number with the sequencing depth; (d) Number of predicted genes per sample; (e) Flower plot indicating the number of genes shared between all samples, and those specific in each individual; (f) Treemap showing taxonomic ranking of the main taxa using Kaiju. The size of each box is proportional to the number of reads assigned to each taxon with respect to the entire dataset. The placement of boxes is arbitrary with respect to boxes within the same taxonomic rank and does not correspond to any form of phylogeny or relatedness; (g) Venn diagrams illustrating the overlap among the 1,146 dominant phylotypes in the microbiome (green) and the dominant bacterial genera detected by 16S rRNA gene sequencing (purple)

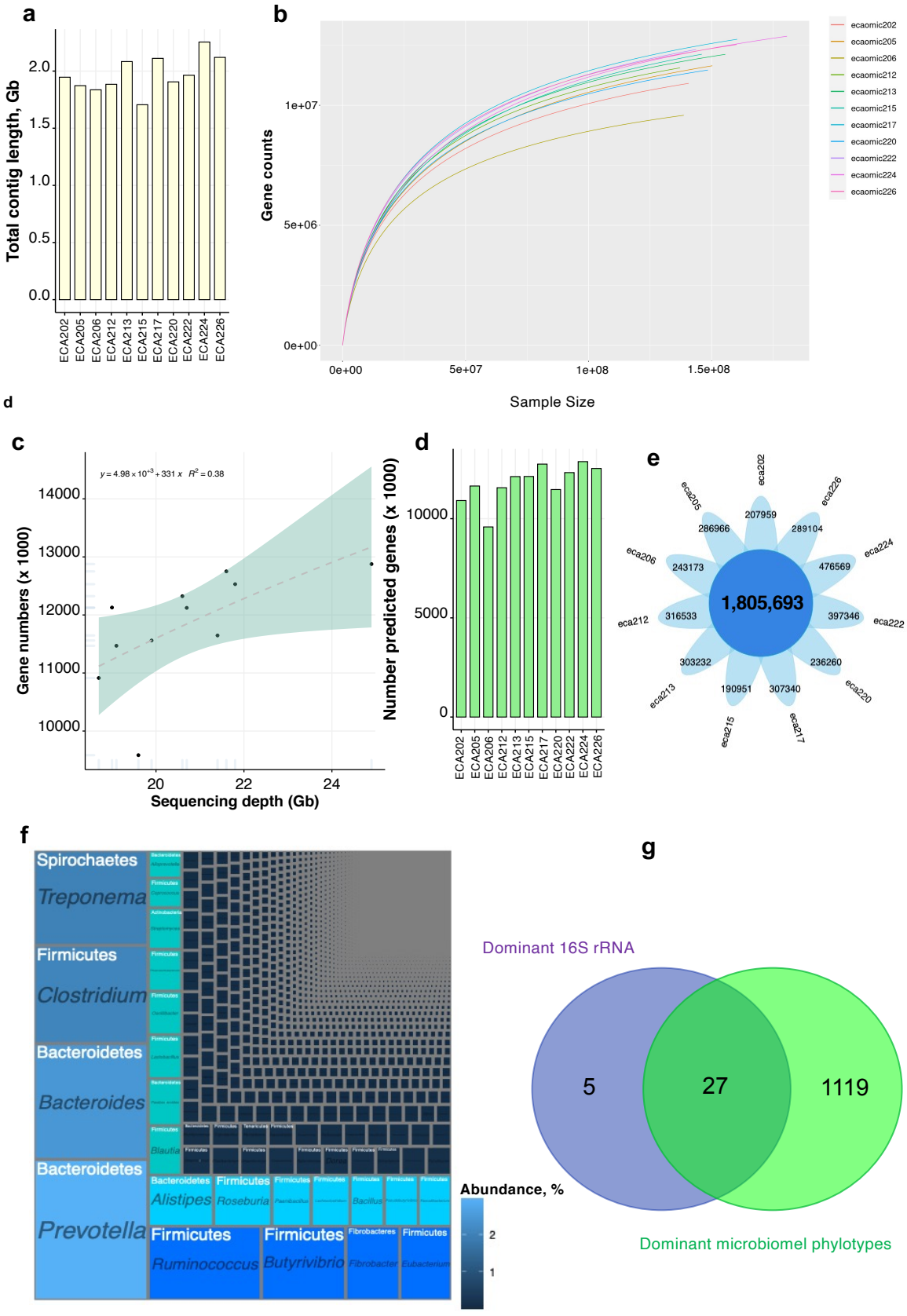

a
Total contig length, Gb
2.0
1.5
1.0
0.5
0.0
ECA202
ECA205
ECA206
ECA212
ECA213
ECA215
ECA217
ECA220
ECA222
ECA224
ECA226
b
Gene counts
1e+07
5e+06
0e+00
0e+00
5e+07
1e+08
1.5e+08
Sample Size
ecaomic202
ecaomic205
ecaomic206
ecaomic212
ecaomic213
ecaomic215
ecaomic217
ecaomic220
ecaomic222
ecaomic224
ecaomic226
d
c
$y = 4.98 \times 10^{-3} + 331\ x\ \ R^2 = 0.38$
Gene numbers (x 1000)
14000
13000
12000
11000
10000
20
22
24
Sequencing depth (Gb)
d
Number predicted genes (x 1000)
10000
5000
0
ECA202
ECA205
ECA206
ECA212
ECA213
ECA215
ECA217
ECA220
ECA222
ECA224
ECA226
e
eca202
eca205
eca226
eca206
eca224
eca212
eca222
eca213
eca220
eca215
eca217
207959
286966
289104
243173
476569
1,805,693
316533
397346
303232
236260
190951
307340
f
Spirochaetes
Treponema
Firmicutes
Clostridium
Bacteroidetes
Bacteroides
Bacteroidetes
Prevotella
Bacteroidetes
Alistipes
Firmicutes
Roseburia
Firmicutes
Ruminococcus
Firmicutes
Butyrivibrio
Fibrobacteres
Fibrobacter
Firmicutes
Eubacterium
Abundance, %
2
1
g
Dominant 16S rRNA
5
27
1119
Dominant microbiomel phylotypes

Figure S2 - Functional diversity of the first horse gut gene catalog

(a) Frequency of KEGG pathways in the gut gene catalog estimated from the orthologous groups (KOs). The horizontal bars represent the absolute number of genes found in each of the KEGG pathways; (b) Frequency of total carbohydrate-active enzymes modules (CAZymes) estimated in the gut gene catalog. CAZymes are colored according to the class they belong to

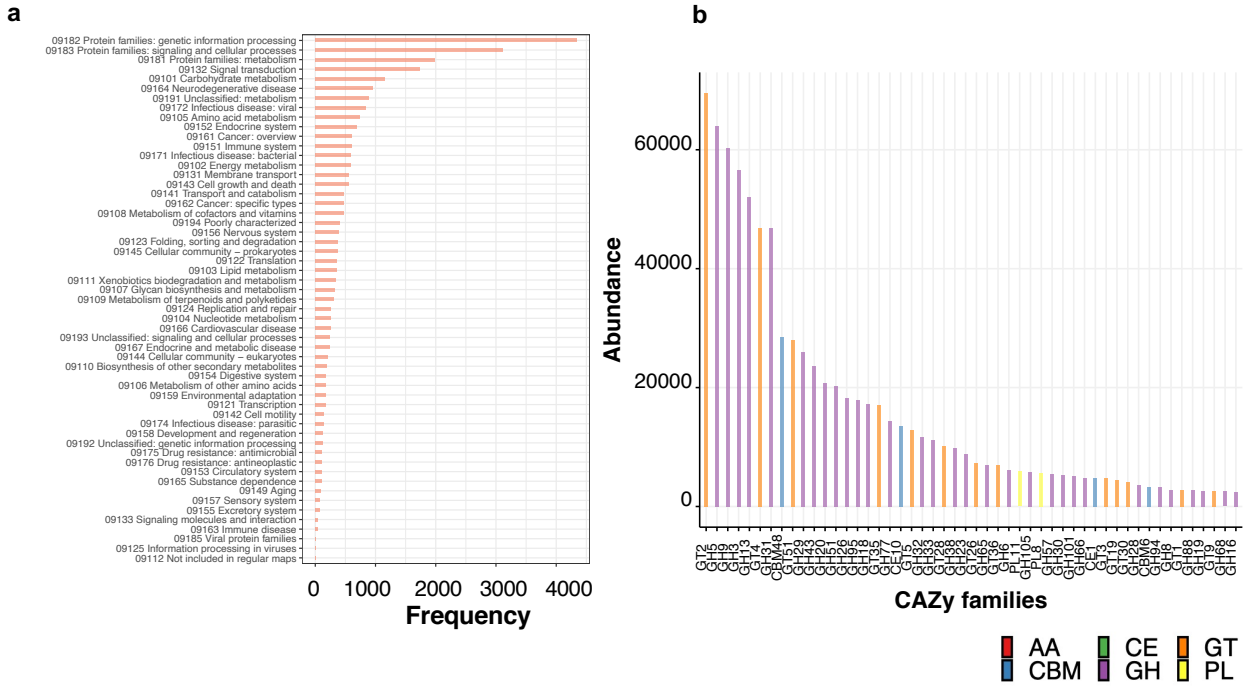


Figure S3 - Taxonomic annotation and phylogenetic tree of 372 metagenome-assembled genomes (MAGs).

(a) Phylogenetic tree of the 372 MAGs belonging to bacteria and archaea. The outer cycle boxplots represent their abundance. Boxplots are colored according to phyla; (b) Number of gut MAGs for each phyla used to generate the MAG catalog, colored according to whether they represent uncultured or cultured genomes; (c) Circular stacked bar plot of the phyla detected for each individual in the cohort; (d) Top 15 MAGs abundance across samples (left) alongside the overlaid abundance histogram and boxplot to describe their distribution in the cohort; (e) The CGView (Circular Genome Viewer) software family was used for generating genome maps of *Streptococcus equinus* (MAG322) and the MAG E1_64 [8], which shared 99% of similarity

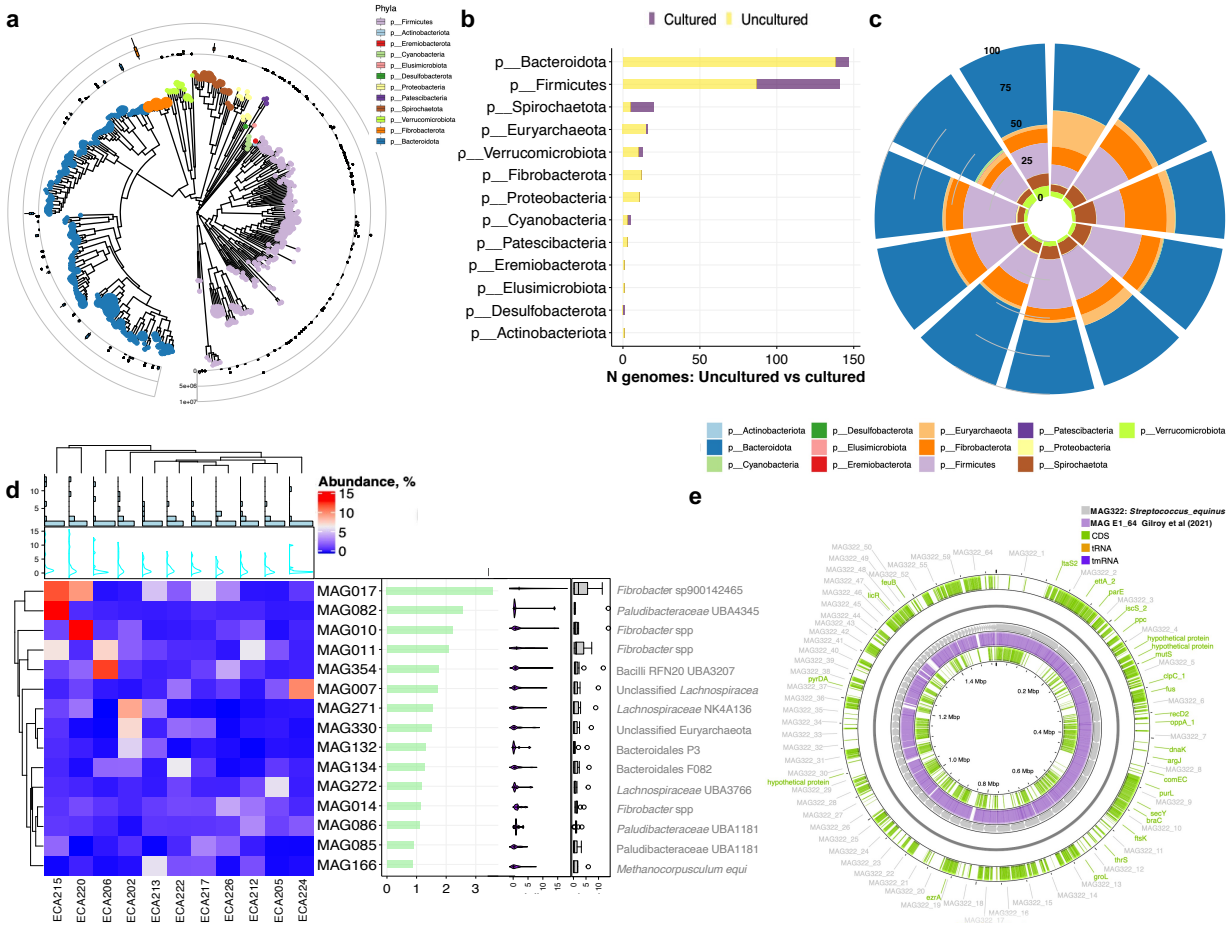


Figure S4 - Landscape of gut metagenome in athletes: two meta-communities with different microbial diversity and distinct metagenomic functions

(a) Principal coordinate analysis (PCoA) ordination analysis (Bray-Curtis distance) of dominant phylotypes composition; (b-c) PCoA ordination analysis (Bray-Curtis distance) of the CAZymes and KOs profiles estimated from the gut gene catalog, respectively. In all cases, colors indicate community classification, the community type 1 (red color) and community type 2 (blue color)

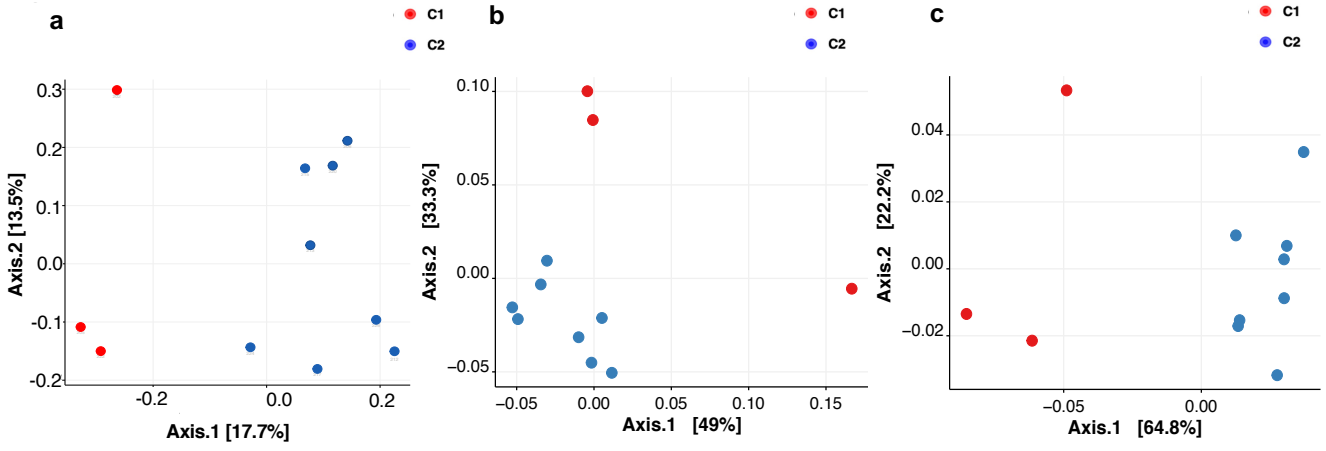


Figure S5 - Effect of cardiovascular fitness on the gut microbiome based on 16S rRNA amplicon sequencing. Validation of results with an independent dataset of 22 athletes

(a) Boxplot representation of the composite of cardiovascular capacity in the validation set: "L" corresponds to "low", M" to "Medium" and "H" to "high". The composite of the cardiovascular capacity was made by combining the post-exercise heart rate, the cardiac recovery time and the average speed during the race. adjusted *p* values from Wilcoxon rank-sum tests; (b) PCoA ordination analysis (Bray-Curtis distance) of the relative abundances of ASV estimated with the 16S rRNA sequencing in the validation set. Points denote individual samples, which are colored according to the clustering group: H (red), M (green) and L (blue); (c) Bray-Curtis distance of the gut microbial ASVs estimated by 16S rRNA gene between the Low, Medium and High groups in the validation set. Boxes show median and interquartile range, and whiskers indicate 5th to 95th percentile. adjusted *p* values from Tukey's Honest significant differences tests; (d) Multivariate sparse partial least-squares discriminant analysis (sPLS-DA) loading plot showing the contributing genera towards the separation between individuals with better cardiovascular fitness (red color) and less fit individuals (blue color). Bar length indicates loading coefficient weight of selected genus, ranked by importance, bottom to top

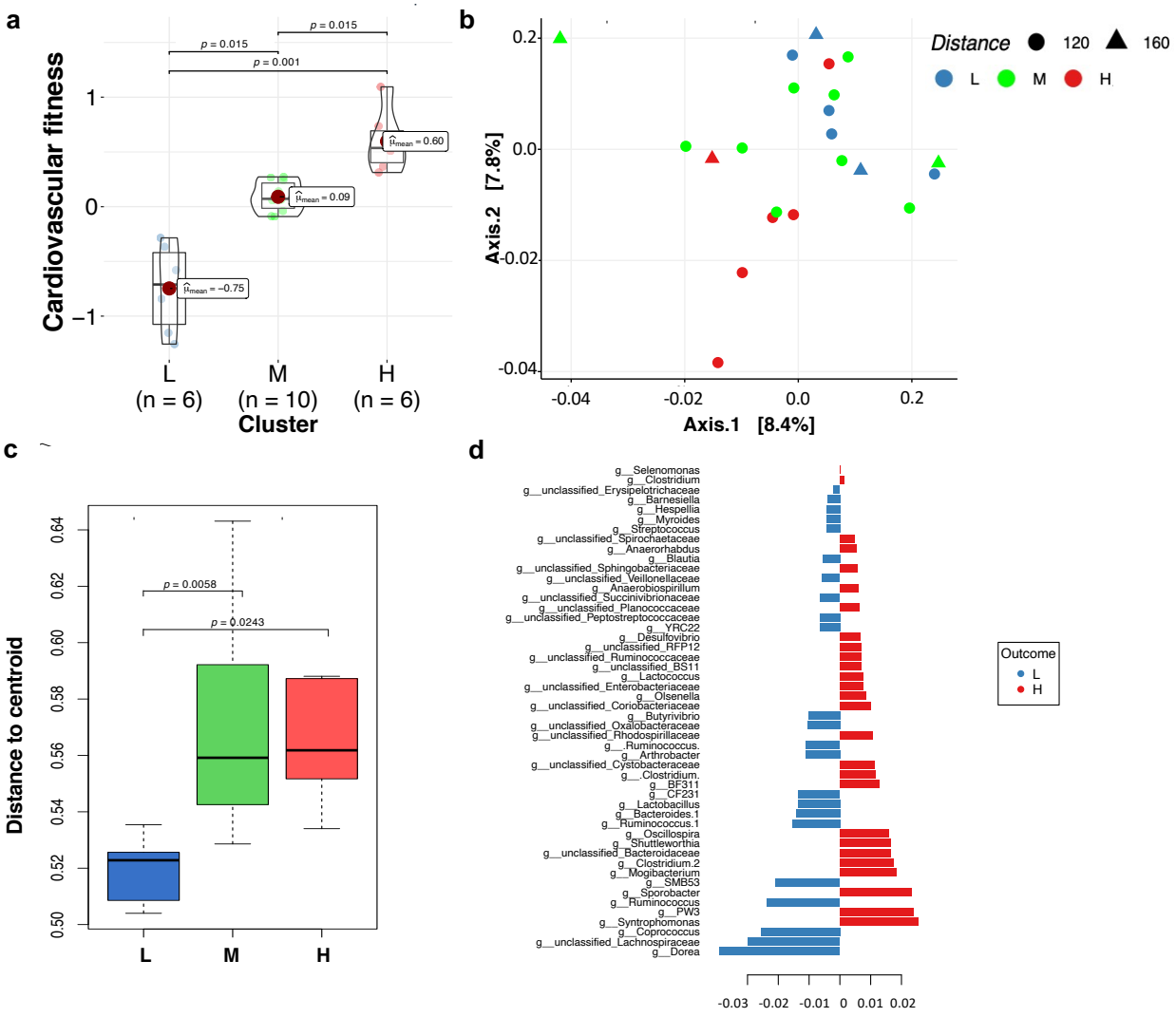


Figure S6 - Interconnection between cardiovascular fitness and gut microbiome composition and functionality

(a-c) DIABLO sample plot demonstrating discrimination between groups based on dominant phylotypes data, mitochondrial related genes transcriptome and carbohydrate-active enzymes (CAZy) data; (d) Mitochondrial-related genes contributing to separation along with component 1 of DIABLO sample plot; (e) Gut microorganisms loads contributing to separation along with component 1 of DIABLO sample plot; In all cases, bar length indicates loading coefficient weight of selected feature, ranked by importance, bottom to top and colors indicate community classification, the community type 1 (red color) and community type 2 (blue color)

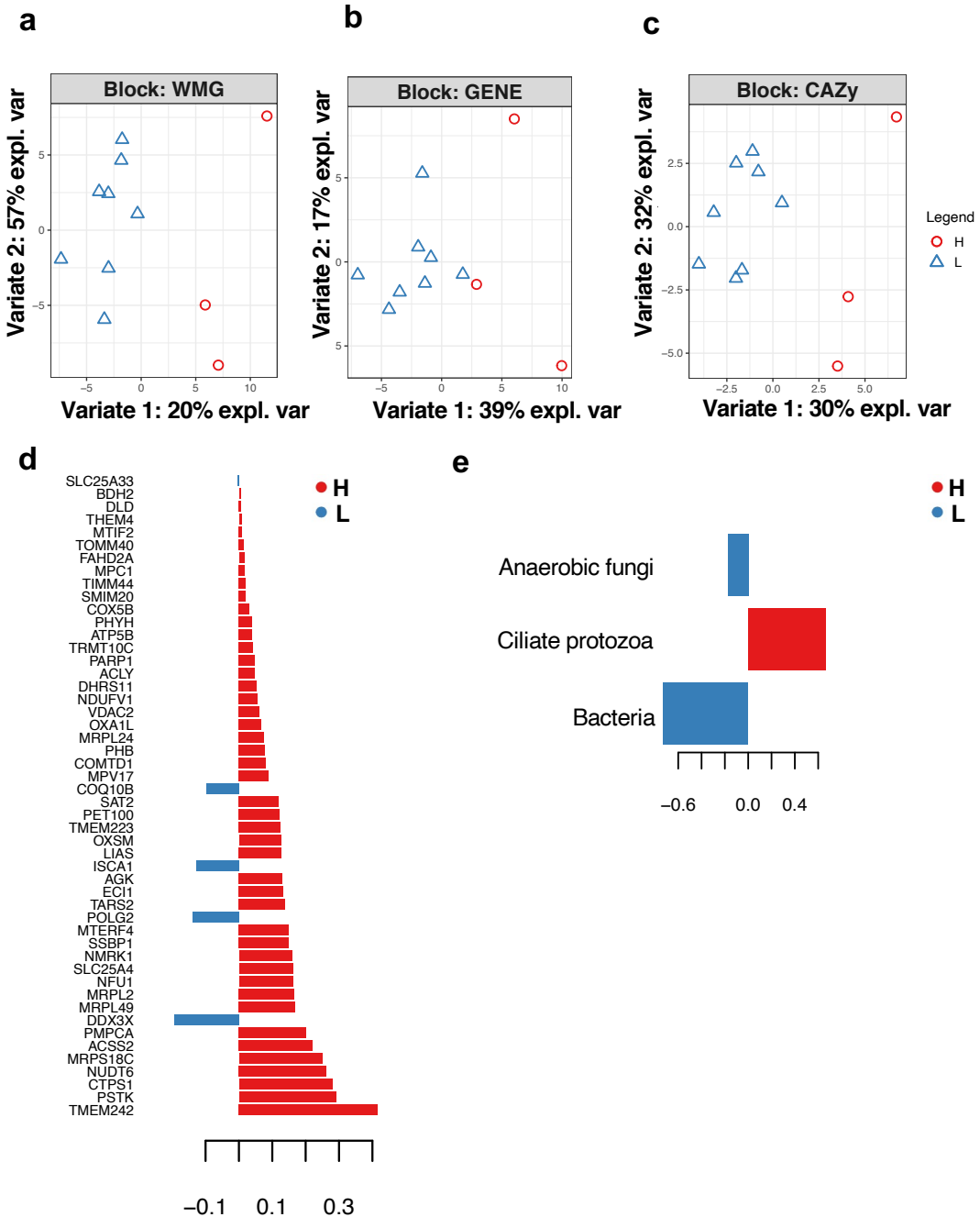


Figure S7 - Taxonomic classification of the first horse gut gene catalog using eggNOG
(a) Taxonomic bar plots of the dominant genera in the gut gene catalog according to the evolutionary genealogy of genes non-supervised orthologous groups (eggNOG). Colors denote microbial genera; (b) Taxonomic bar plots of the dominant genera in the gut according to the 16S rRNA amplicon dataset and the Greengenes 16S rRNA Database (release 13.8)

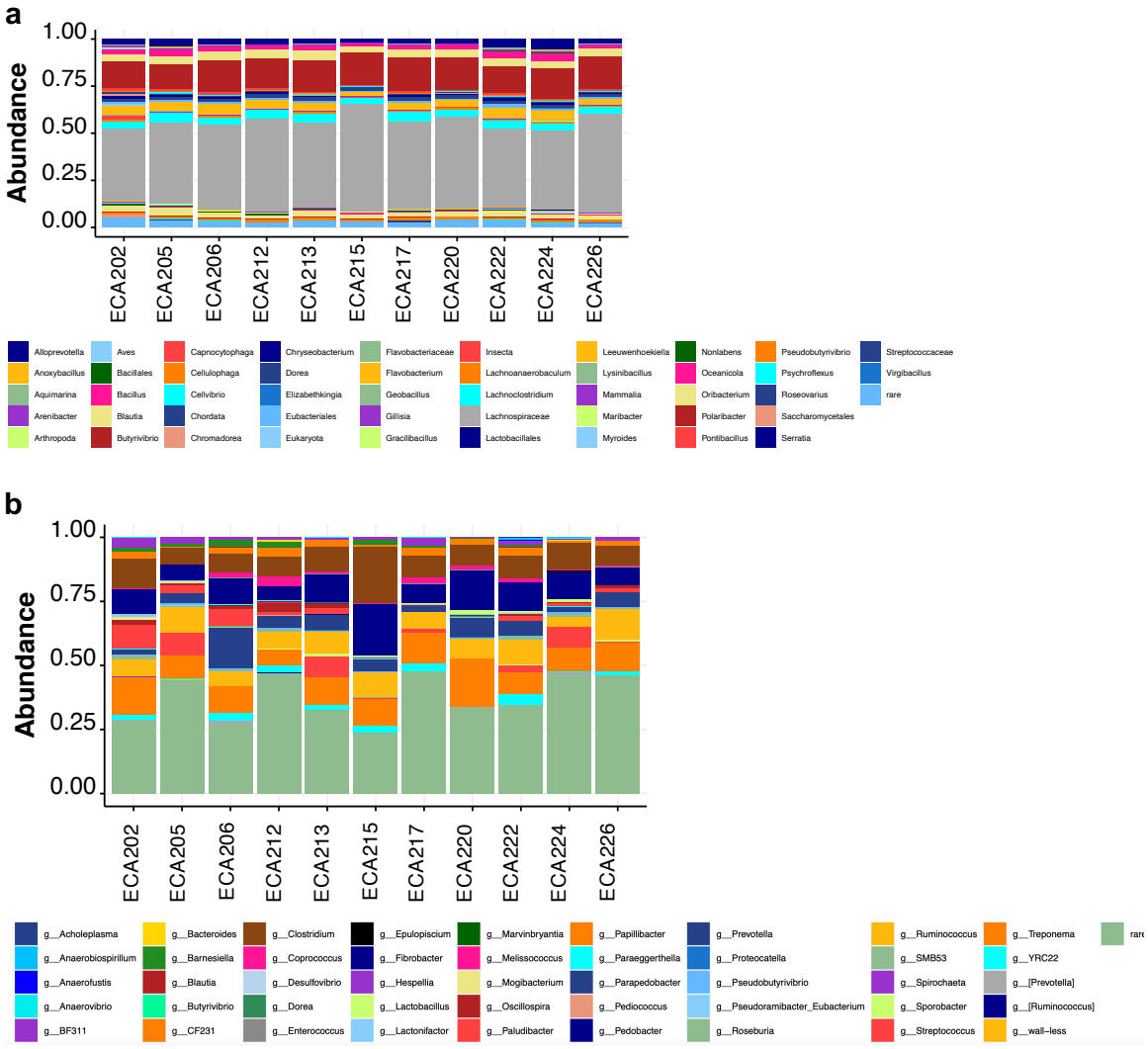


Figure S8 - Interconnection between cardiovascular fitness, microbiome composition and functionality

(a) Dot plot representation of log-transformed fold change of dominant phylotypes that were significantly different between the individuals with reduced and improved cardiovascular capacity. The logs of fold changes lying above 0 indicate that phylotypes were more abundant in individuals with reduced cardiovascular capacity than those with improved cardiovascular fitness. By contrast, the negative logs of fold changes lying between indicate that the phylotypes abundance was lower in individuals with reduced cardiovascular capacity compared to more fit participants. Dots are colored by phyla; (b) Dot plot representation of log-transformed fold change of CAZymes that were significantly different between the individuals with reduced and improved cardiovascular capacity. The logs of fold changes lying between 0 and 1.5 indicate that CAZymes were more abundant in less fit individuals. By contrast, the logs of fold changes lying between 0 and -1.5 indicate that the CAZymes abundance was lower in individuals with reduced cardiovascular capacity compared to more fit individuals. Dots are colored by classes.

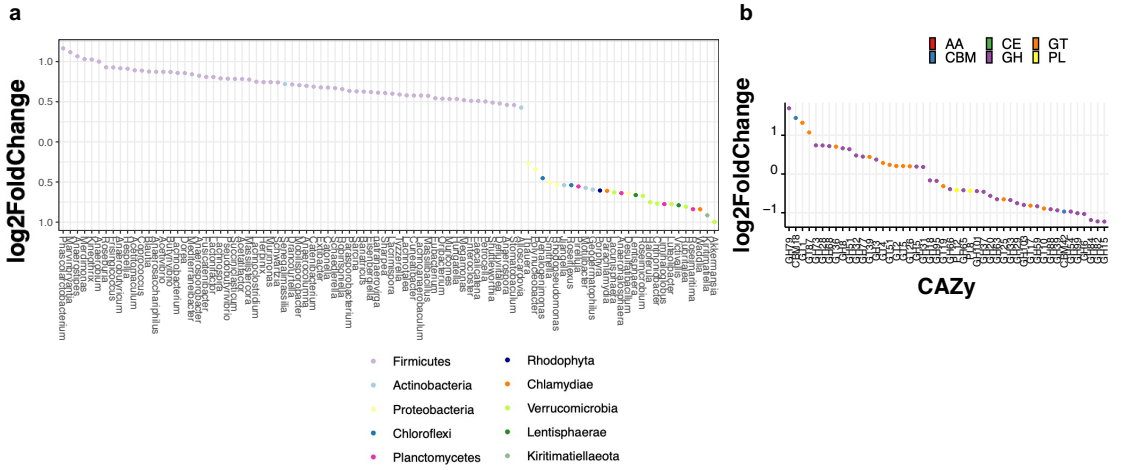


Figure S9 - Characteristics of the 372 metagenome-assembled genomes (MAGs)
(a-b) Distribution of completeness and contamination quality metrics, respectively. The 372 MAGs are overlaid on each boxplot, showing the minimum value, first quartile, median, third quartile and maximum value; (c) Relationship between MAG completeness and contamination. Dot size is proportional to the strain heterogeneity and it is colored according to the phylum; (d) Sankey diagram showing the numbers of MAGS with completeness between 50-90% or > 90%. Only a relatively small fraction of the MAGs (73 out of 372) was annotated at genus level; (e) Frequency of the contig length (Kb); (f) Relationship regression between MAG size (Kb) and percentage of completeness; (g-h) Distribution of the number of contigs and N50 value (Kb), respectively. The 372 MAGs are overlaid on each boxplot, showing the minimum value, first quartile, median, third quartile and maximum value.

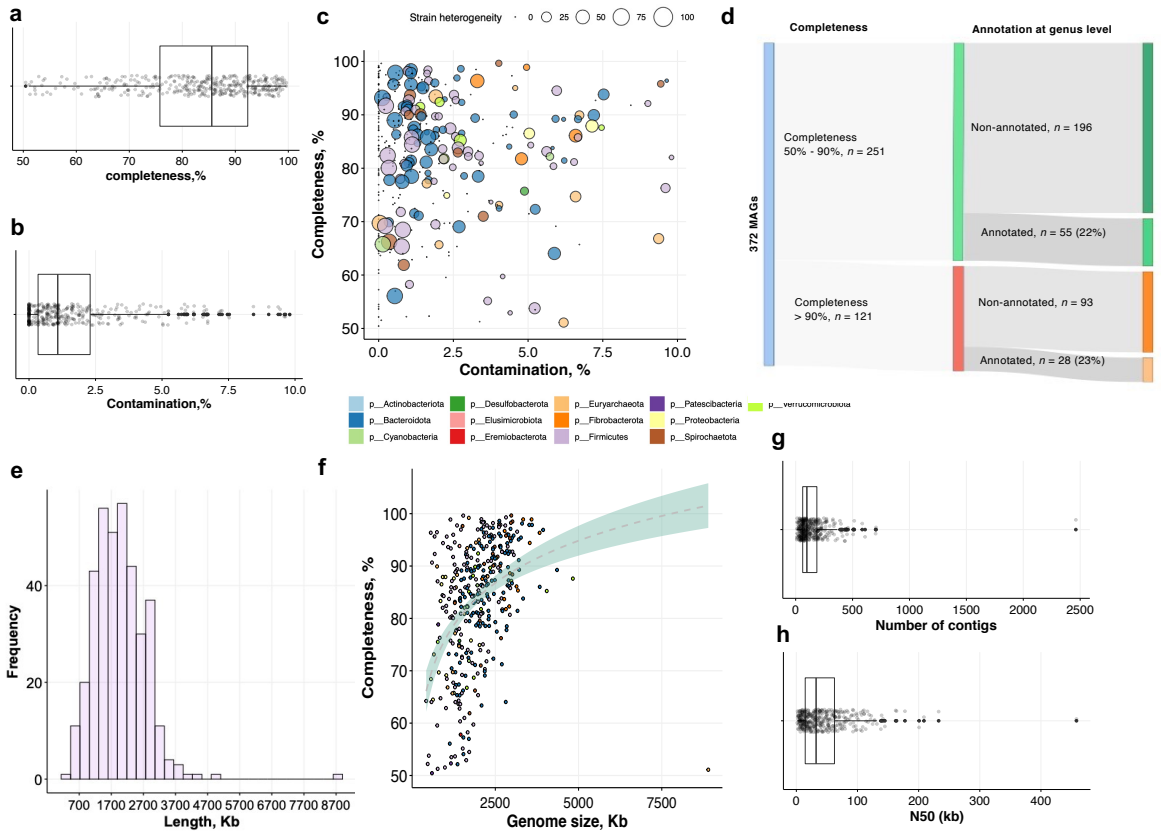


Figure S10 - The CAZy classes are similarly distributed across MAGs

Phylogenetic tree of the MAGs. Each MAG is colored according to its phylum. CAZymes abundance data was integrated into the phylogenetic tree context. CAZymes are colored by class.

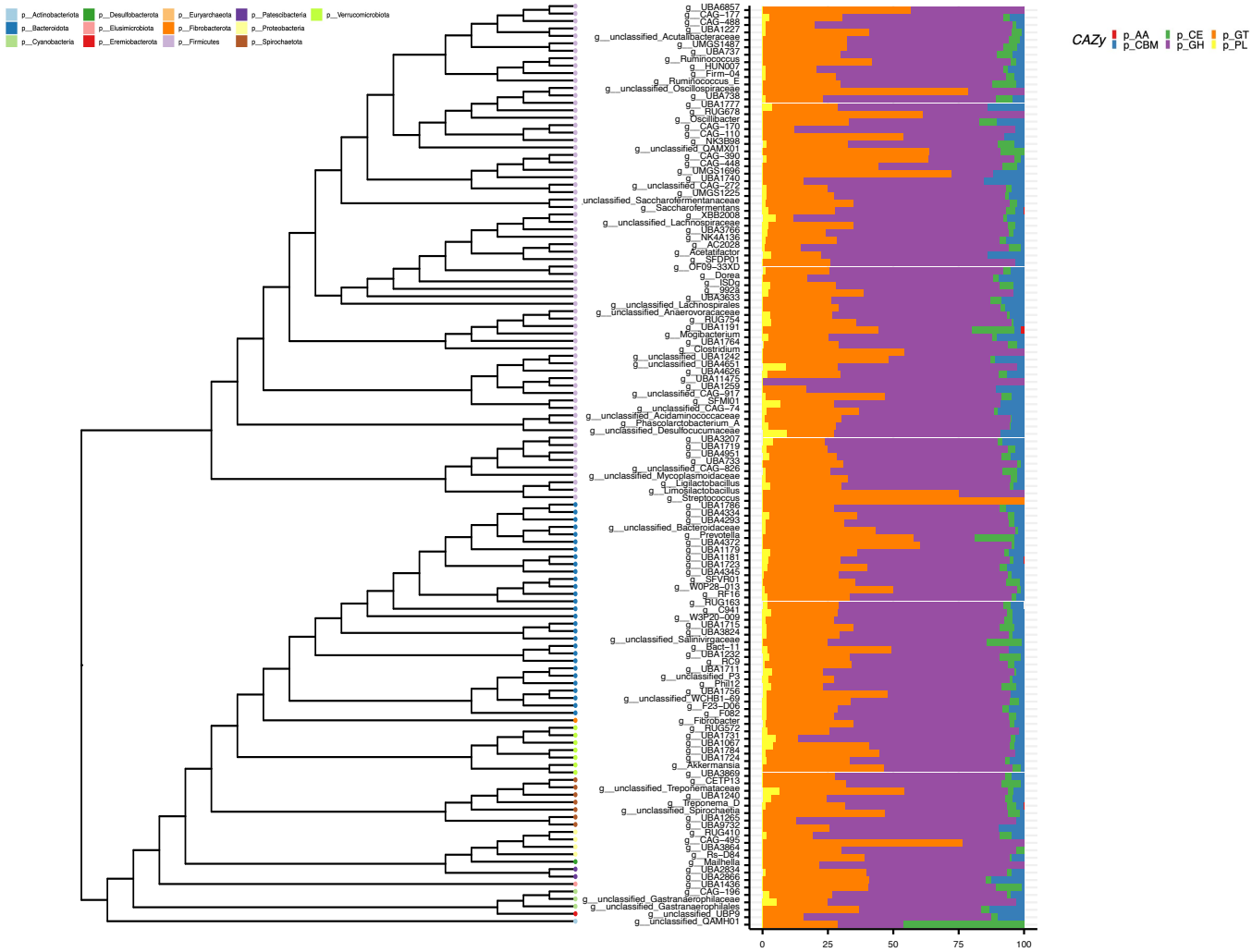
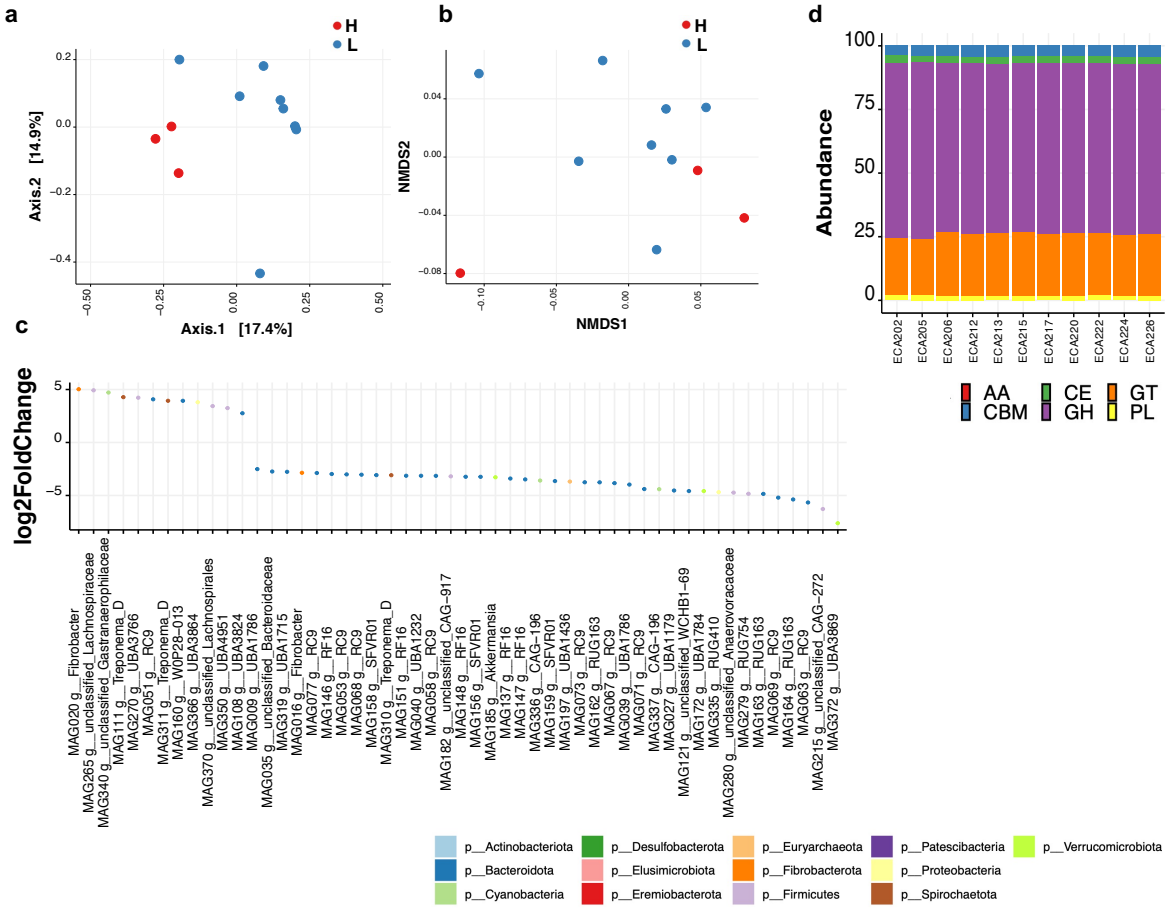


Figure S11 - The metagenome-assembled genomes do not capture the functional complexity of the natural population

(a-b) Principal coordinate analysis (PCoA) ordination analysis (Bray-Curtis distance) of CAzymes and KOs estimated from MAGs, respectively; (c) Dot plot representation of log-transformed fold change of MAGs that were significantly different between the individuals with reduced and improved cardiovascular capacity. The logs of fold changes lying between 0 and 6 indicate that MAGs were more abundant in individuals with low cardiovascular fitness than high. By contrast, the logs of fold changes lying between 0 and -6 indicate that the MAG abundance was lower in less fit participants compared to more fit horses. Dots are colored by phyla; (d) Distribution of the fraction of CAZymes abundance estimated from MAGs in each sample. CAZymes are colored according to their class



SUPPLEMENTARY TABLES

Table S1 - Metadata of horses recruited in the experiment

Table S2 - Sequencing data and assembly statistics for each sample and quality assessment for the gene catalog

Table S3 - Top 25% most dominant phylotypes sorted by their abundance and ubiquitousness found in more than half of the samples. Taxonomic profiling was done with Kaiju. For each individual, the phylotype taxonomic assignment and their counts are depicted

Table S4 - Number of ASVs and relative abundance of genera detected by 16S rRNA gene sequencing

Table S5 - Abundance of the CAZymes, classes and their estimated mechanisms in the gene catalog

Table S6 - List of acquired antimicrobial resistance (AMR) genes observed in this study

Table S7 - Assembly statistics of the 372 metagenome assembled genomes (MAGs)

Table S8 - List of differentially expressed mitochondrial-related genes
Gene description, gene localizations and gene types were determined according to IPA (https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis). Columns from N from AG show the expression values obtained by subtracting the T0 (baseline) from the T1 (post-exercise) expression matrix values, *i.e.*, by calculating the ratio between T1 and T0 log scaled expression values from the two matrices

Table S9 - Relative abundance of metabolites obtained from blood of the 11 horses under study collected before and after the endurance ride

Table S10 - Free carnitine and various acylcarnitine values (µmol/L ± standard deviation) pre- and post-exercise measured in serum from 11 horses under study

Table S11 - Biochemical parameters obtained from the blood of the 11 horses under study collected before and after the endurance race

Table S12 - Summary of the significant correlations between environmental and host variables and microbial community beta diversity ordination on NMDS plots. +/- designates the direction of the association between the fit variable and NMDS axis. Square correlation coefficients are depicted as estimators of the goodness of fit together with the empirical *p*-values for each variable

Table S13 - Metadata of the horses recruited in the validation set

Table S14 - ASV taxonomic assignments and ASV counts for the 22 horses in the validation set based on 16S rRNA sequencing

Table S15 - Fecal pH and fecal short chain fatty acids measurements in the 11 horses under study before the endurance race, as well as concentrations of bacteria, ciliate protozoa and anaerobic fungi in the feces of the 11 horses under study

Table S16 - Comparison of differential abundance of microbial phylotypes between individuals with different cardiovascular capacity. Differences were calculated between the two groups based on the DESeq2 model followed by Benjamini and Hochberg multiple test correction. The table shows the phylotypes that significantly varied in abundance between the less fit individuals and more fit individuals. The information relative to the log fold change and adjusted *p*-value are provided

Table S17 - Comparison of differential abundance of KOs and their associated pathways between individuals with different cardiovascular fitness. Differences were calculated between the two groups based on the DESeq2 model followed by Benjamini and Hochberg multiple test correction. The table shows the KOs that significantly varied in abundance between the less fit individuals and more fit individuals. The information relative to the log fold change and adjusted *p*-value are provided

Table S18 - Comparison of differential abundance of CAZymes between individuals with different cardiovascular capacity. Differences were calculated between the two groups based on the DESeq2 model followed by Benjamini and Hochberg multiple test correction. The table shows the CAZymes that significantly varied in abundance between the less fit individuals and more fit individuals. The information relative to the log fold change and adjusted *p*-value are provided

Table S19 - Distributions of CAZymes in the 372 MAGs

Table S20 - Distributions of KOs in the 372 MAGs

Table S21 - Comparison of differential abundance of MAGs between pathways between individuals with different cardiovascular capacity. Differences were calculated between the two groups based on the DESeq2 model followed by Benjamini and Hochberg multiple test correction. The table shows the MAGs that significantly varied in abundance between the less fit individuals and the more fit individuals. The information relative to the log fold change and adjusted *p*-value are provided

SUPPLEMENTARY REFERENCES

1. Capomaccio, S. *et al.* Microarray analysis after strenuous exercise in peripheral blood mononuclear cells of endurance horses. *Anim. Genet.* **41**, 166–175 (2010).
2. O'Donovan, C. M. *et al.* Distinct microbiome composition and metabolome exists across subgroups of elite Irish athletes. *J. Sci. Med. Sport* **23**, 63–68 (2020).
3. Hawley, J. A., Lundby, C., Cotter, J. D. & Burke, L. M. Maximizing Cellular Adaptation to Endurance Exercise in Skeletal Muscle. *Cell Metab.* **27**, 962–976 (2018).
4. Sorbara, M. T. *et al.* Functional and Genomic Variation between Human-Derived Isolates of Lachnospiraceae Reveals Inter- and Intra-Species Diversity. *Cell Host Microbe* **28**, 134-146.e4 (2020).
5. Kieser, S., Brown, J., Zdobnov, E. M., Trajkovski, M. & McCue, L. A. ATLAS: A

Snakemake workflow for assembly, annotation, and genomic binning of metagenome sequence data. *bioRxiv* 1–8 (2019) doi:10.1101/737528.

6. Chaumeil, P. A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: A toolkit to classify genomes with the genome taxonomy database. *Bioinformatics* **36**, 1925–1927 (2020).
7. Youngblut, N. D. *et al.* Large scale metagenome assembly reveals novel animal-associated microbial genomes, biosynthetic gene clusters, and other genetic diversity. *bioRxiv* **5**, 1–15 (2020).
8. Gilroy, R., Leng, J., Ravi, A., Adriaenssens, E. M. & Oren, A. Metagenomic investigation of the equine faecal microbiome reveals extensive taxonomic and functional diversity. (2021).