1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses

Yuanqiang Zou, Wenbin Xue, Guangwen Luo, Ziqing Deng, Panpan Qin, Ruijin Guo, Haipeng Sun, Yan Xia, Suisha Liang, Ying Dai, Daiwei Wan, Rongrong Jiang, Lili Su, Qiang Feng, Zhiye Jie, Tongkun Guo, Zhongkui Xia, Chuan Liu, Jinghong Yu, Yuxiang Lin, Shanmei Tang, Guicheng Huo, Xun Xu, Yong Hou, Xin Liu, Jian Wang, Huanming Yang, Karsten Kristiansen, Junhua Li, Huijue Jia, and Liang Xiao

Reference genomes are essential for metagenomic analyses and functional characterization of the human gut microbiota. We present the Culturable Genome Reference (CGR), a collection of 1,520 nonredundant, high-quality draft genomes generated from >6,000 bacteria cultivated from fecal samples of healthy humans. Of the 1,520 genomes, which were chosen to cover all major bacterial phyla and genera in the human gut, 264 are not represented in existing reference genome catalogs. We show that this increase in the number of reference bacterial genomes improves the rate of mapping metagenomic sequencing reads from 50% to >70%, enabling higher-resolution descriptions of the human gut microbiome. We use the CGR genomes to annotate functions of 338 bacterial species, showing the utility of this resource for functional studies. We also carry out a pan-genome analysis of 38 important human gut species, which reveals the diversity and specificity of functional enrichment between their core and dispensable genomes.

The human gut microbiota refers to the all the microorganisms that inhabit the human gastrointestinal tract. Diverse roles of the gut microbiota in human health and disease have been recognized. Metagenomic studies have transformed our understanding of the taxonomic and functional diversity of the human microbiota, but more than half of the sequencing reads from a typical human fecal metagenome cannot be mapped to existing bacterial reference genomes. The lack of high-quality reference genomes has become an obstacle for high-resolution analyses of the human gut microbiome. Although the previously reported Integrated Gene Catalog (IGC) has enabled metagenomic, metatranscriptomic and metaproteomic analyses, the gap between compositional and functional analyses can only be filled by individual bacterial genomes. Genes that co-occur among samples can be clustered into metagenomic linkage groups, metagenomic clusters and metagenomic species, whose annotation depends on alignment to the limited number of existing reference genomes. Other metagenomics-based analyses of the gut microbiome—for example, single nucleotide polymorphisms (SNPs), indels and copy number variations—rely on the coverage and quality of reference genomes.

Despite the rapid increase in the number of sequenced bacterial and archaeal genomes, reference genomes for gut bacteria are underrepresented. It is estimated that <4% of the bacterial genomes in the US National Center for Biotechnology Information (NCBI) database belong to the human gut microbiota. Rather, the focus has been on clinically relevant pathogenic bacteria, which are overrepresented in the microbial databases. The first catalog of 178 reference bacterial genomes for the human microbiota was reported by the Human Microbiome Project (HMP) in 2010. To date, the HMP has sequenced >2,000 microbial genomes cultivated from human body sites, 437 of which are gut microbiota (data accessed 8 September 2017). However, the number of reference gut bacterial genomes is still far from saturated.

We present a reference catalog of genomes of cultivated human gut bacteria (named the CGR), established by culture-based isolation of >6,000 bacterial isolates from fecal samples of healthy individuals. The CGR comprises 1,520 nonredundant, high-quality draft bacterial genomes, contributing at least 264 new reference genomes to the gut microbiome. After inclusion of CGR genomes, the mapping rate of selected metagenomic datasets improved from around 50% to over 70%. In addition to improving metagenomic analyses, the CGR will improve functional characterization and pan-genomic analyses of the gut microbiota at high resolution.

Results
Expanded catalog of gut bacterial genomes. We obtained 6,487 bacterial isolates from fresh fecal samples donated by 155 healthy volunteers by using 11 different media under anaerobic conditions. The CGR comprises 1,520 nonredundant, high-quality draft bacterial genomes, contributing at least 264 new reference genomes to the gut microbiome. After inclusion of CGR genomes, the mapping rate of selected metagenomic datasets improved from around 50% to over 70%. In addition to improving metagenomic analyses, the CGR will improve functional characterization and pan-genomic analyses of the gut microbiota at high resolution.

The human gut microbiota refers to the all the microorganisms that inhabit the human gastrointestinal tract. Diverse roles of the gut microbiota in human health and disease have been recognized. Metagenomic studies have transformed our understanding of the taxonomic and functional diversity of the human microbiota, but more than half of the sequencing reads from a typical human fecal metagenome cannot be mapped to existing bacterial reference genomes. The lack of high-quality reference genomes has become an obstacle for high-resolution analyses of the human gut microbiome. Although the previously reported Integrated Gene Catalog (IGC) has enabled metagenomic, metatranscriptomic and metaproteomic analyses, the gap between compositional and functional analyses can only be filled by individual bacterial genomes. Genes that co-occur among samples can be clustered into metagenomic linkage groups, metagenomic clusters and metagenomic species, whose annotation depends on alignment to the limited number of existing reference genomes. Other metagenomics-based analyses of the gut microbiome—for example, single nucleotide polymorphisms (SNPs), indels and copy number variations—rely on the coverage and quality of reference genomes.

Despite the rapid increase in the number of sequenced bacterial and archaeal genomes, reference genomes for gut bacteria are underrepresented. It is estimated that <4% of the bacterial genomes in the US National Center for Biotechnology Information (NCBI) database belong to the human gut microbiota. Rather, the focus has been on clinically relevant pathogenic bacteria, which are overrepresented in the microbial databases. The first catalog of 178 reference bacterial genomes for the human microbiota was reported by the Human Microbiome Project (HMP) in 2010. To date, the HMP has sequenced >2,000 microbial genomes cultivated from human body sites, 437 of which are gut microbiota (data accessed 8 September 2017). However, the number of reference gut bacterial genomes is still far from saturated.

We present a reference catalog of genomes of cultivated human gut bacteria (named the CGR), established by culture-based isolation of >6,000 bacterial isolates from fecal samples of healthy individuals. The CGR comprises 1,520 nonredundant, high-quality draft bacterial genomes, contributing at least 264 new reference genomes to the gut microbiome. After inclusion of CGR genomes, the mapping rate of selected metagenomic datasets improved from around 50% to over 70%. In addition to improving metagenomic analyses, the CGR will improve functional characterization and pan-genomic analyses of the gut microbiota at high resolution.

Expanded catalog of gut bacterial genomes. We obtained 6,487 bacterial isolates from fresh fecal samples donated by 155 healthy volunteers by using 11 different media under anaerobic conditions. The CGR comprises 1,520 nonredundant, high-quality draft bacterial genomes, contributing at least 264 new reference genomes to the gut microbiome. After inclusion of CGR genomes, the mapping rate of selected metagenomic datasets improved from around 50% to over 70%. In addition to improving metagenomic analyses, the CGR will improve functional characterization and pan-genomic analyses of the gut microbiota at high resolution.
conditions (Supplementary Fig. 1a and Supplementary Table 1). Notably, more than half of the isolates were cultured from MPYG medium (Supplementary Fig. 1b). All the isolates were subjected to 16S rRNA gene amplicon sequencing analysis, and 1,759 nonredundant isolates that provided broad coverage of the phylogenetic tree were selected for whole-genome sequencing (Supplementary Fig. 1c and Supplementary Table 2). After de novo assembly of the next-generation sequencing reads, we identified 104 isolates that contained more than one genome. These assembled sequences were then parsed into 212 genomes using our in-house pipeline (Supplementary Table 3). Briefly, multi-genomes were split at scaffold level on the basis of G+C content versus sequencing depth. The closest reference genomes for the spilt scaffolds were determined on the basis of average nucleotide identity (ANI), and the mis-split scaffolds were mapped back to their closest reference genome to get the final split genome (see Methods). In total, we obtained a collection of 1,867 newly assembled genomes, 1,520 (81.4%) of which fulfilled the HMP’s criteria for high-quality draft genomes and exceeded 95% genome completeness and less than 5% contamination as evaluated by CheckM. The genome sizes and G+C contents of CGR ranged from 0.2 to 7.9 Mbp and 26.56% to 64.28%, respectively. A total of 5,749,641 genes were predicted from the annotation data (Supplementary Table 4).

Taxonomic annotation of CGR was carried out using a self-constructed, efficient ANI-based pipeline (Supplementary Fig. 2). The 1,520 high-quality genomes were classified into 338 species-level clusters (ANI ≥ 95%), a species delineation corresponding to 70% DNA–DNA hybridization), which covered all the major phyla of the human gut microbiota, including Firmicutes (211 clusters, 796 genomes), Bacteroidetes (60 clusters, 447 genomes), Actinobacteria (54 clusters, 235 genomes), Proteobacteria (10 clusters, 36 genomes) and Fusobacteria (3 clusters, 6 genomes) (Fig. 1a and Supplementary Table 5). Among these 338 clusters, 134 clusters (corresponding to 264 genomes) were not annotated to any present reference genomes in NCBI (Fig. 1a), and 50 clusters did not fall within any sequenced genera (Supplementary Table 5). To corroborate the presence of novel species in CGR, we carried out additional taxonomic identification using 16S rRNA gene analysis. A species was recognized as novel if its 16S rRNA gene sequence had < 98.7% similarity with known species in the EzBioCloud database (see Methods). Overall, we identified 350 distinct bacterial species (based on operational taxonomic units), including 149 candidate novel species, 42 of which represent candidate novel genera. These results underscore the value of the individual reference genomes provided by the CGR.

Despite the variation of individual microbiota at the genus level, the CGR identified bacterial populations with broad diversity, covering eight out of nine core genera in the Chinese gut microbiota. More than 80 species were novel in comparison with the previously sequenced species from a reported 1,000 cultured bacterial species from the human gastrointestinal tract (Supplementary Fig. 3a). Moreover, the CGR successfully identified 38 genera that were of low relative abundance (<1%) according to the IGC, which is a large catalog of reference genes derived from a collection of ~1,250 metagenomic samples from individuals on three continents (Supplementary Fig. 3b). Among them, 7 genera were identified with more than 20 genomes (Bifidobacterium, Collinsella, Coprobacillus, Dorea, Streptococcus, Prevotella and Parabacteroides). The CGR also identified another 9 genera that were not detected.

Fig. 1 | Phylogenetic tree of 1,520 isolated gut bacteria based on whole-genome sequences. The 1,520 high-quality genomes in CGR are classified into 338 species-level clusters (ANI ≥ 95%) based on their whole-genome sequences. Bacterial species from Firmicutes are colored in orange; Bacteroidetes, blue; Proteobacteria, green; Actinobacteria, violet; Fusobacteria, gray. Novel genera and species are highlighted by red and orange branches, respectively. The bar on the outermost layer indicates the number of genomes archived in each cluster. Rhizobium selenitireducens ATCC BAA 1503 was used as an outgroup for phylogenetic analysis.
by IGC (Butyricoccus, Butyricimonas, Catenibacterium, Dietma,
Erysipelotroxidium, Megamonas, Melissosoccus, Peptoclostridium
and Vagococcus) (Supplementary Fig. 3b). These results underscore
the contribution of the CGR to the existing database of gut bacterial
whole genomes.

**Improvement in metagenomic and SNP analyses.** The existing
reference genomes for metagenomic sequence mapping are far from
saturated. For example, the genomes or draft genomes of bacteria
and archaea used in a recent study allowed mapping of less than
half of the sequences in the fecal metagenome\(^6\). To illustrate the
value of the CGR to metagenomic analyses, we performed sequence
mapping using previous metagenomic data\(^7\) with or without CGR.
For Chinese samples, the read mapping rate in the original study
that used the IGC dataset (3,449 reference genomes from IGC\(^9\))
was 52.00%, which was significantly improved to 76.88% after the
inclusion of the CGR dataset (Fig. 2a and Supplementary Table 6).
Since all the samples in the CGR were from China, it is reasonable
to assume that this dataset contributes substantially to the
Chinese fecal metagenome. To evaluate the contribution of the
CGR to the mapping of non-Chinese metagenomes, we carried
out a similar analysis using metagenomic data from American,
Spanish and Danish fecal samples. Notably, the metagenomic read
mapping ratios of these samples all increased substantially (Fig. 2a),
although to a lesser extent compared with that of Chinese samples
(Supplementary Fig. 4a). The improvement of mapping rates in
both Chinese and non-Chinese samples indicates that the CGR cov-
ers a considerable number of gut bacterial species shared by people
between these countries. To reveal the improvement of gene and
protein diversity enabled by the CGR, we compared the gene and
protein cumulative curve based on genomes used in a previous IGC
study and after addition of the CGR (Supplementary Fig. 4b,c). The
number of gene and protein families increased with inclusion of
the first 1,500 genomes, but more or less plateaued at around 3,000
genomes. The addition of our CGR genomes led to a substantial
increase in the number of added gene and protein families as a
function of genome number. A total of 373,555 gene clusters and
149,945 protein clusters were added by inclusion of the CGR, cor-
responding to a 22% and 16% increase in known gene and protein
sequence diversity, respectively.

To further illustrate the utility of the CGR, we used it to analyze
gut microbiome SNPs in a cohort of 250 samples from the TwinsUK
registry\(^\circ\). We generated a new set of 282 nonredundant representa-
tive genomes from the CGR (see Methods, Supplementary Fig.
5 and Supplementary Table 7), which number nearly doubled the
152 reference genomes used in the original TwinsUK analysis\(^\circ\). To
highlight the new reference genomes identified by analysis with the
existing genomes and the CGR genomes, we performed an ANI-
based alignment of the 282 genomes with the previously reported
152 genomes. Among the 192 newly added reference genomes, 85
were classified species while 107 were unclassified species (Fig. 2b).
A high SNP density was found in Streptococcus mutans
(Cluster 274) and unclassified Firmicutes (Cluster 6),
Eubacterium rectale
(Cluster 217), and
Vagococcus
(Cluster 21).

A high SNP density was found in

**Functions of gut microbiome bacteria.** To better elucidate func-
tions of the gut microbiota, we annotated gene functions in 1,520
CGR genomes using KEGG (the Kyoto Encyclopedia of Genes and
Genomes)\(^14\). Functional pathways at KEGG level 2 pathways involved in carbohydrate and amino acid metabolism
are abundant in all isolated strains, suggesting that these are core
functions of the gut microbiota (Supplementary Fig. 6). We also
analyzed KEGG level 3 pathways and focused on those enriched
at the phylum Fusobacteria, Bacteroidetes and Proteobacteria, the main
phyla of gram-negative bacteria. Genes involved in glycan degra-
dation (ko00531 and ko00511) were abundant in the genomes of the
Bacteroidetes phylum. This observation is consistent with the
notion that members of Bacteroidetes are prominent human gut
symbionts that help degrade glycans in the diet and the gut mucosa\(^15\).
The members of the Bacteroidetes also possess a high proportion of
genomes involved in sphingolipid metabolism (ko00600), glycosphin-
golipid biosynthesis (ko00601, ko00602 and ko00604) and steroid
hormone biosynthesis (ko00140). Sphingolipids and hormone bio-
synthesis are ubiquitous in eukaryotic cells but not present in most
bacteria. These results suggest that members of the Bacteroidetes

---

**Fig. 2 | Contribution of CGR to metagenomic and SNP analyses.** a. The
read mapping ratio of a previous metagenomic analysis (IGCR) was
significantly improved by CGR (IGCR + CGR) in fecal samples from
Chinese (\(n = 368, P = 6 \times 10^{-78}\)), American (\(n = 139, P = 2 \times 10^{-17}\)), Spanish
(\(n = 320, P = 4 \times 10^{-35}\)) and Danish (\(n = 109, P = 4 \times 10^{-17}\)) individuals.
The significance of improvement was determined by two-side Wilcoxon
rank-sum test. IGCR, 3,449 reference genomes used in the IGC study\(^7\);
CGR, 1,520 reference genomes generated in this study. Each box plot
illustrates the estimated median (center line), upper and lower quartiles
(box limits), 1.5 x interquartile range (whiskers), and outliers (points) of
the read mapping ratio. b. Reference genomes for SNP analysis generated in
previous study\(^7\) (IGCR, green) and current study (CGR, blue). The
unclassified species of reference genomes in CGR are highlighted in violet.
not only participate in energy metabolism in the gut, but may also act in sphingolipid and hormone signaling in mamalian cells. The Proteobacteria showed relatively high abundance in genes involved in degradation of xenobiotics (ko01220), possibly contributing to the degradation of environmental chemicals and pharmaceuticals in the gut.

The signal transduction system (two-component system, KEGG level 2 pathway) were ubiquitous in the genera Paenibacillus, Bacillus, Klebsiella, Escherichia, Citrobacter and Enterobacter, which are also presented in environmental niches, such as soil and water. The abundant signal transduction and xenobiotics degradation systems allow these genera to sense and respond to various stresses and toxic substance presented in natural environments. Cell motility (chemotaxis, ko02030; flagellar assembly, ko02040) was conserved in the genera Roseburia, Paenibacillus, Bacillus, Escherichia, Citrobacter and Enterobacter, but varied within the genera Clostridium and Eubacterium.

Next we investigated functions and pathways that are annotated in the KEGG database, but not categorized as KEGG pathways (Fig. 3b and Supplementary Table 9). Virulence factors and antibiotic resistance genes were annotated using the Virulence Factors Database (VFDB) and Comprehensive Antibiotic Resistance Database (CARD), respectively. Virulence factors and antibiotic resistance are clinically relevant and are abundant in the Proteobacteria phylum, suggesting that this phylum may be a reservoir for opportunistic pathogens. We examined the distribution of genes involved in responses to stresses frequently encountered by gut bacteria: oxygen tolerance and acid resistance. Oxygen tolerance was reflected by the number of genes encoding catalase and peroxide dismutase, two detoxification enzymes that scavange reactive oxygen species produced during aerobic respiration. As expected, the facultative anaerobic bacteria in the genera Paenibacillus, Bacillus, Klebsiella, Escherichia, Citrobacter and Enterobacter were more oxygen tolerant. In addition to the previously reported Bacteroides fragilis, other members of Bacteroidetes also showed moderate oxygen tolerance. Notably, bacteria in the Bacteroidetes phylum and the Bifidobacterium genus generally lacked acid resistance genes, suggesting that potential probiotics based on these organisms may suffer impaired survival in the acidic stomach environment after oral administration. Finally, we examined the distribution of six bacterial functions in the CGR that might have beneficial effects on human health. Amino acid and vitamin B synthesis genes were widely present in various gut bacteria, suggesting that gut microbiota might be an alternative source for nutrients that are sparse in vegetarian diets. Genes encoding bacterial bile salt hydrolases, which transform primary bile acids into secondary bile acids in the human intestine, were also ubiquitous in most gut bacteria. Genes encoding β-galactosidases, which might attenuate problems associated with lactose intolerance, were relatively abundant in the phylum Bacteroidetes. Genes involved in bacteriocin synthesis in gut bacteria were relatively rare and did not show phylum- or genus-specific distribution.

Core and pan-genomes of underrepresented gut bacteria. We carried out a pan-genome analysis of 36 species or clusters that contain more than ten genomes, as well as two other species enriched in healthy controls compared with patients with type 2 diabetes in previous studies. Fecalibacterium prausnitzii (cluster 63, seven genomes) and butyrate-producing bacterium SS3_4 (cluster 45, nine genomes). These clusters covered the phyla Firmicutes, Bacteroidetes, Actinobacteria and Proteobacteria (Supplementary Fig. 7a and Supplementary Table 8a). The pan-genome of a cluster can be defined as the sum of the core genes and dispensable genes (including unique genes and accessory genes) of all the members within that cluster. Our pan-genome analysis showed that Eubacterium rectale (cluster 37) contained the lowest proportion of core genes (12%); the remaining genes fell into accessory and unique genomes (38% and 40%, respectively). In contrast, Eubacterium 3_1 (cluster 6) contained the largest proportion of core genes (53%) (Supplementary Fig. 7b). The pan-genome fitting curves showed that most clusters in Bacteroidetes displayed an ‘open’ pan-genome and had a relatively large pan-genome size, with Bacteroides vulgatus
In contrast, COG categories enriched in the dispensable genome included translation, ribosomal structure and biogenesis (J), energy production and conversion (C), amino acid transport and metabolism (E), coenzyme transport and metabolism (F), nucleotide transport and metabolism (G), and replication, recombination and repair (L).
in two recent studies. Although there was some overlap between the novel species archived by CGR and in these two studies, the CGR contains 659 additional genomes (representing 209 clusters or species). Our cultivation methods can be applied to expand the CGR until it is saturated with the genomes of culturable gut bacteria. After that, single-cell sequencing can be used to investigate genomes of unculturable bacteria, with an overall aim of defining a saturated set of reference genomes of gut microbiota to underpin a better understanding of gut microbiome biology.

We applied out CGR genome dataset to assign functions to gut bacteria. For example, we found that virulence factors and antibiotic resistance genes are enriched in *Klebsiella*, *Escherichia*, *Citrobacter* and *Enterobacter*, which are opportunistic pathogens frequently isolated in clinical samples. The abundance of signal transduction and cell motility genes in these bacteria could further contribute to their pathogenicity. Notably, the Proteobacteria also possess abundant genes for degradation of xenobiotics, which might affect drug metabolism of patients in drug therapy. In line with this, a recent study reported that intratumor Proteobacteria can metabolize chemotherapeutic drugs into inactive forms and thus attenuate the efficacy of cancer therapies. The genes involved in beneficial functions such as glycan degradation and vitamin B synthesis are enriched in the *Bacteroides* genus, consistent with its mutualistic role in the human gut. Notably, we found that *Bacteroides* species contain a considerable number of genes involved in sphingolipid and steroid hormone synthesis, suggesting their potential for modulating signaling in mammalian cells. In support of this, a recent study reported that *Bacteroides fragilis* can take advantage of sphingolipid signaling to enable symbiosis in the intestine. It is noteworthy that genes involved in glycan degradation and sphingolipid metabolism were also enriched in the genus *Bifidobacterium*, another well-known gut commensal microbe. However, genes involved in both pathways were not abundant in the *Prevotella* genus, suggesting a distinct function of *Prevotella* compared with other members of the Bacteroidetes phylum. This might account for observed negative correlations between the relative abundances of *Prevotella* and *Bacteroides* in the gut microbiota. The potential role of gut bacteria in metabolism of estrogens has long been recognized, but detailed mechanistic studies are still lacking. It will be interesting to explore the implication of this unique function of gut bacteria in hormone-related health or disease. The CGR also enabled the identification of several potential bacteriocin-producing bacteria strains, which merit further verification.

The CGR will improve metagenomic analyses, genome variation analyses, functional characterization and pan-genome analyses. The isolated gut bacteria strains have been deposited in the China National GeneBank (CNGB) and may be useful for studies that aim to alter microbiota functions, as novel probiotics, or for verification of disease-associated bacterial markers.

**Online content**

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability, and associated accession codes are available at [https://doi.org/10.1038/s41587-018-0008-8](https://doi.org/10.1038/s41587-018-0008-8).

Received: 26 October 2017; Accepted: 13 December 2018; Published online: 4 February 2019

**References**

1. Wang, J. & Jia, H. Metagenome-wide association studies: fine-mining the microbiome. Nat. Rev. Microbiol. 14, 508–522 (2016).
2. Lynch, S. V. & Pedersen, O. The human intestinal microbiome in health and disease. N. Engl. J. Med. 375, 2369–2379 (2016).
3. Qin, J. et al. A human gut microbial gene catalog established by metagenomic sequencing. Nature 464, 59–65 (2010).
4. Sunagawa, S. et al. Metagenomic species profiling using universal phylogenetic marker genes. Nat. Methods 10, 1196–1199 (2013).
5. Methé, B. A. et al. A framework for human microbiome research. Nature 486, 221–222 (2012).
6. Li, J. et al. An integrated catalog of reference genes in the human gut microbiome. Nat. Biotechnol. 32, 834–841 (2014).
7. Qin, J. et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature 490, 55–60 (2012).
8. Karlsson, F. H. et al. Gut metagenome in European women with normal, impaired and diabetic glucose control. Nature 498, 99–103 (2013).
9. Le Chatelier, E. et al. Richness of human gut microbiome correlates with metabolic markers. Nature 500, 541–546 (2013).
10. Nielsen, H. B. et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat. Biotechnol. 32, 822–828 (2014).
11. Schlossing, S. et al. Genomic variation landscape of the human gut microbiome. Nature 493, 45–50 (2013).
12. Hu, Y. et al. Metagenome-wide analysis of antibiotic resistance genes in a large cohort of human gut microbiota. Nat. Commun. 4, 2151 (2013).
13. Greenblum, S., Carr, R. & Borodovsky, M. Extensive strain-level copy number variation across human gut microbiome species. Cell 160, 583–594 (2015).
14. Nelson, K. E. et al. A catalog of reference genomes from the human microbiome. Science 328, 994–999 (2010).
15. Zhang, J. et al. A phyo-functional core of gut microbiota in healthy young Chinese cohorts across lifestyles, geography and ethnicities. ISME J. 9, 1979–1990 (2015).
16. Rajilic-Stojanovic, M. & de Vos, W. M. The first 1000 cultured species of the human gastrointestinal microbiota. FEMS Microbiol. Rev. 38, 996–1047 (2014).
17. Xie, H. et al. Shotgun metagenomics of 250 adult twins reveals genetic and environmental impacts on the gut microbiome. Cell Systems 3, 572–584.e573 (2016).
18. Kaneshia, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28, 27–30 (2000).
19. Koropatkin, N. M., Cameron, E. A. & Martens, E. C. How glycan metabolism shapes the human gut microbiota. Nat. Rev. Microbiol. 10, 323–335 (2012).
20. Chen, L. et al. VFDB: a reference database for bacterial virulence factors. Nucleic Acids Res. 33, D325–D328 (2005).
21. Jia, B. et al. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. Nucleic Acids Res. 45(D1), D566–D573 (2017).
22. Sund, C. J. et al. The *Bacteroides fragilis* transcriptome response to oxygen and H2O2: the role of OxyR and its effect on survival and virulence. Mol. Microbiol. 67, 129–142 (2008).
23. Aw, W. & Fukuda, S. Understanding the role of the gut ecosystem in diabetes mellitus. J. Diabetes Investig. 9, 5–12 (2018).
24. Aw, W. & Fukuda, S. Toward the comprehensive understanding of the gut ecosystem via metabolomics-based integrated omics approach. Semin. Immunopathol. 37, 5–16 (2015).
25. Medini, D., Donati, C., Tettelin, H., Masignani, V. & Rappuoli, R. The microbial pan-genome. Curr. Opin. Genet. Dev. 15, 589–594 (2005).
26. Louis, P. & Flint, H. J. Diversity, metabolism and microbial ecology of butyrate-producing bacteria from the human large intestine. FEMS Microbiol. Lett. 294, 1–8 (2009).
27. Van den Abbeele, P. et al. Butyrate-producing Clostridium cluster XIIa species specifically colonize mucins in an in vitro gut model. ISME J. 7, 949–961 (2013).
28. Louis, P., Young, P., Hollt, G. & Flint, H. J. Diversity of human colonic butyrate-producing bacteria revealed by analysis of the butyryl-CoA:acetate CoA-transferase gene. Environ. Microbiol. 12, 304–314 (2010).
29. de Vries, L. E. et al. The gut as reservoir of antibiotic resistance: microbial diversity of tetracycline resistance in mother and infant. PLoS One 6, e21644 (2011).
30. Tatusov, R. L. et al. The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4, 41 (2003).
31. Lagier, J.-C. et al. Culture of previously uncultured members of the human gut microbiota by cultureuromics. Nat. Microbiol. 1, 16203 (2016).
32. Browne, H. et al. Culture of ‘unculturable’ human microbiota reveals novel taxa and extensive sporulation. Nature 533, 543–546 (2016).
33. Guentzel, M. N., Escherichia, Klebsiella, Enterobacter, Serratia, Citrobacter and Proteus. Medical Microbiology Ch. 25 (University of Texas Medical Branch, Galveston, Texas, USA, 1996).
34. Josenhans, C. & Suerbaum, S. The role of motility as a virulence factor in bacteria. Int. J. Med. Microbiol. 291, 605–614 (2002).
35. Gotth, Y. et al. Two-component signal transduction as potential drug targets in pathogenic bacteria. Curr. Opin. Microbiol. 13, 232–239 (2010).
36. Geiler, L. T. et al. Potential role of intratumor bacteria in mediating tumor resistance to the chemotherapeutic drug gemcitabine. Science 357, 1156–1160 (2017).
Acknowledgements
We gratefully acknowledge colleagues at BGI-Shenzhen for DNA extraction, library construction, sequencing and discussions. This research was supported by the National Natural Science Foundation of China (grants 81670606 and 81673850), the Shenzhen Municipal Government of China (JCYJ20160229172757249 and JCYJ20170818111103886).

Author contributions
H.J., J.L., L.X., Y.Z. and W.X. conceived and designed the project. H.J., J.L. and L.X monitored the project. Y.Z., W.X., Y.X., Z.D. and C.Y. collected samples and performed experiments. Y.Z., Y.X., Z.D., B.G., H.S., X.Y., G.L., Q.F., Z.J., C.J., T.G., Y.H., X.L., J.W., H.Y., S.T., G.H., C.L., Z.X., J.Y., K.K., L.X. and H.J. analyzed and interpreted the data. Y.Z. and Z.D. wrote the paper. K.K., H.J. and L.X. revised the paper. All authors commented on the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41587-018-0008-8.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to J.L., H.J. or L.X.

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
Whole-genome sequencing and de novo assembly. DNA extraction. Isolates cultivated to stationary phase were centrifuged at 7,227 g at 4 °C for 10 min, and the resulting pellets were resuspended in 1 ml of Tris-ENDTA. For bacterial cell lysis, 50 µl of 10% SDS and 10 µl of proteinase K (20 mg/ml) were added, and the solution was incubated at 55 °C in a water bath for 2 h. The released genomic DNA was extracted using the phenol-chloroform method.

Genome sequencing and assembly. Paired-end libraries with an insert size of 500 bp were constructed and sequenced on Illumina Hiseq 2000 platform to obtain about 100 x clean data for each sample. The reads were assembled using SOAPdenovo 2.04 to form scaffolds from which the rRNA genes were extracted by RNAmmer 1.2. An in-house pipeline was used to obtain the best assembly containing an orthologous design, investigate LMDLJi15 (arguments of SOAPdenovo) and a single-factor design to investigate K (argument of SOAPdenovo) by comprehensively considering contig average length, longest scaffold and rRNA score. Libraries with an insert size of 240 bp were constructed and sequenced on the ionProton platform, which produced about 100 x clean data for each sample. The reads were assembled through SPAdes (version 3.1.0) to form scaffolds.

Assessment of genome quality. Six high-quality draft assembly criteria from the Human Microbiome Project (HMP)44 and rRNA (5s, 16s and 23s) completeness were adopted to ensure the assembly quality. The criteria are (i) 90% of the genome assembly must be included in contigs >500 bp, (ii) 90% of the assembled bases must be at >5x read coverage, (iii) the contig N50 must be >5 kb, (iv) scaffold N50 must be >20 kb, (v) average contig length must be >5 kb, and (vi) >90% of the core genes44-45 must be present in the assembly.

Splitting for multi-genome isolates. The multi-genomes in isolates were initially identified using CheckM19 (contamination >5%) and confirmed by manual inspection of the plot of G+C percentage vs. sequencing depth. An in-house pipeline was developed to split the scaffolds of multi-genomes into single genomes. Briefly, scaffolds in multi-genomes were first split on the basis of G+C percentage vs. sequencing depth values using the dbscan function of R (package "fpc"). The "complete" and "contamination" of split genomes were checked using CheckM. For split genomes with "complete" >100% or "contamination" >15%, an additional species-designating pipeline was used to obtain their closest reference (with ANI value >90%). Finally, the mis-split scaffolds in each split genome were mapping back to the closest reference genome using BLASTn (+e 1e-2 -F P -b 100 - k -a 8 -m 8) against the Antibiotic Resistance Genes Database (ARDB) (both query and subject match length higher than 40%, with identity higher than the ARDB-recommended thresholds), the Virulence Factor Database (VFDB16,17) (query match length higher than 50%, with identity higher than 85%), and the bacteriocin database (downloaded from BAGEL3, with identity higher than 60%). Annotation of genes against the Comprehensive Antibiotic Resistance Database (CARD18) was performed using Resistance Gene Identifier available as a downloadable command-line tool in the download section of the CARD website using default parameters.

Mapping ratio of metagenomic samples. The metagenomic reads were first aligned to the reference genomes of IGC3 (3,449 sequenced prokaryotic genomes from IGC3) using SOAP263 (default parameters, except -m 100 -x 1000 -r 1 -l 30) and the unmapped reads were then aligned to the newly sequenced genome of CGR. The read mapping ratio of different mapping strategies was calculated, and the difference between samples was determined by Wilcoxon test in R.

Analysis of gene and protein diversity. Gene clusters. 5,749,641 genes in the 1,759 CGR genomes and 11,330,042 genes in 3,449 IGCR genomes were clustered using CD-HIT66 with default parameters, except -c 0.95 -d 0.95 -m 1000 - i 1 30 -v 5 -s c 0.95. The unmapped reads were then aligned to the newly sequenced genome of CGR. The read mapping ratio of different mapping strategies was calculated, and the difference between samples was determined by Wilcoxon test in R.
clusters was drawn according to the sample name alphabetically with IGCR at the front part and CGR at the latter part.

SNP identification and similarity score. 1,520 genomes from the CGR were aligned with the sequenced reads from the 250 TwinsUK samples using SOAP2 with identity ≥90%. Representative genomes used for SNP analysis were identified according to three criteria described previously17. The resulting 282 genomes (Supplementary Table 7) that fulfilled these criteria were used as references for SNP calling using SAMtools (frequency >1% and supported by ≥4 reads) as previously described17,20. The reference genomes used in a previous study17 (152 genomes) were compared with that from CGR of this study (282 genomes) to identify shared and new reference genomes using ANI ≥95% as a threshold (species level).

Pan genome analysis for 38 cluster. Clusters containing more than ten genomes (from CGR and NCBI), as well as Prevotella prausnitzii (seven genomes) and butyrate-producing bacteria SS3_3 (nine genomes), were used for pan-genome analysis using the Bacterial Pan Genome Analysis tool (BPGA) pipeline26. The set of genes shared by all the members of cluster was defined as core genes, while genes partially shared in members (accessory genes) and unique to single members (unique genes) in a cluster were defined as dispensable gene67. The pan-genome fitting curves of 38 clusters were generated by the BPGA workflow and plotted in R (v.3.3.3). The functions of genes in the pan-genomes of 38 clusters were annotated by KEGG and ARDB, using arguments identical to those used for functional annotation of genomes. The acetyl-CoA-to-butyrate biosynthesis pathway was generated according to a previous study37, and the associated enzymes were identified according to the functional annotation and BLAST to the NCBI protein database (cut-off e−5, identity ≥70%, coverage ≥70%). The COG database24 was also used to identify the functional distribution in the core and dispensable sections via the BPGA pipeline. The significance of the difference between COG distribution in core and dispensable genomes was examined using Wilcoxon test as implemented in R (v.3.3.3).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability
The assembly draft genomes and annotation information for the 1,520 CGR strains are deposited in the NCBI under accession code PRJNA482748, and these data are also available in the China National GeneBank (CNGB) Nucleotide Sequence Archive (CNSA; accession code CNP0000126). All bacterial strains in the CGR have been deposited in the CNGB, a nonprofit, public-service-oriented organization in China. The accession code for each strain is given in Supplementary Table 5 (Genebank_id). Researchers can explore strain information and request strains via http://ebiobank.cngb.org/index.php?g=Content&m=Hq1&n=sample_5&id=393#.

References
40. Giraffa, G., Rossetti, L. & Neviani, E. An evaluation of chelex-based DNA purification protocols for the typing of lactic acid bacteria. J. Microbiol. Methods 42, 173–184 (2000).
41. Luo, R. et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience 1, 18 (2012).
42. Lagesen, K. et al. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 35, 3100–3108 (2007).
43. Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19, 455–477 (2012).
44. Callister, S. J. et al. Comparative bacterial proteomics: analysis of the core genome concept. PLoS One 3, e1542 (2008).
45. Li, L., Stoeckert, C. J., Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13, 2178–2189 (2003).
46. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 25, 1043–1055 (2015).
47. Richter, M. & Rosselló-Móra, R. Shifting the genomic gold standard for the prokaryotic species definition. Proc. Natl Acad. Sci. USA 106, 19126–19131 (2009).
48. Qin, Q. L. et al. A proposed genus boundary for the prokaryotes based on genomic insights. J. Bacteriol. 196, 2210–2215 (2014).
49. Kim, O. S. et al. Introducing ExElxon-e: a prokaryotic 16S rRNA gene sequence database with phylogenies that represent uncultured species. Int. J. Syst. Evol. Microbiol. 62, 716–721 (2012).
50. Schloss, P. D. et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl. Environ. Microbiol. 75, 7537–7541 (2009).
51. Yarza, P. et al. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nat. Rev. Microbiol. 12, 635–645 (2014).
52. Mende, D. R., Sunagawa, S., Zeller, G. & Bork, P. Accurate and universal delineation of prokaryotic species. Nat. Methods 10, 881–884 (2013).
53. Löytynoja, A. Phylogeny-aware alignment with PRANK. Methods Mol. Biol. 1079, 155–170 (2014).
54. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25, 1972–1973 (2009).
55. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313 (2014).
56. Letunic, I. & Bork, P. Interactive tree of life (iTOL):v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. 44(W1), W242–W245 (2016).
57. Besemer, J. & Borodovsky, M. GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. Nucleic Acids Res. 33, W451–W454 (2005).
58. Ye, Y., Choi, J. H. & Tang, H. RAPSearch: a fast protein similarity search tool for short reads. BMC Bioinformatics 12, 159 (2011).
59. Du, J. et al. KEGG-PATH: Kyoto encyclopedia of genes and genomes-based pathway analysis using a path analysis model. Mol. Biosyst. 10, 2441–2447 (2014).
60. Liu, B. & Pop, M. ARDB—Antibiotic Resistance Genes Database. Nucleic Acids Res. 37, D443–D447 (2009).
61. Chen, L., Zheng, D., Liu, B., Yang, J. & Jin, Q. VFDB 2016: hierarchical and refined dataset for big data analysis—10 years on. Nucleic Acids Res. 44(D1), D694–D697 (2016).
62. van Heel, A. J., de Jong, A., Montalbán-López, M., Kok, J. & Kuipers, O. P. BAGEL3: Automated identification of genes encoding bacteriocins and (non-) bactericidal posttranslationally modified peptides. Nucleic Acids Res. 41, W448–W453 (2013).
63. Li, R. et al. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25, 1966–1967 (2009).
64. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22, 1658–1659 (2006).
65. Hauser, M., Mayer, C. E. & Söding, J. kClust: fast and sensitive clustering of large protein sequence databases. BMC Bioinformatics 14, 248 (2013).
66. Chaudhari, N. M., Gupta, V. K. & Dutta, C. BPGA—an ultra-fast pan-genome algorithm for short reads. BMC Bioinformatics 12, 455–477 (2011).
67. Bosi, E. et al. Comparative genome-scale modelling of Staphylococcus aureus strains identifies strain-specific metabolic capabilities linked to pathogenicity. Proc. Natl Acad. Sci. USA 113, E3801–E3809 (2016).
68. Vital, M., Howe, A. C. & Tieje, J. M. Revealing the bacterial butyrate synthesis pathways by analyzing (meta)genomic data. MBio 5, e00889 (2014).
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted
- Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), indicating how they were calculated
- Clearly defined error bars
- State explicitly what error bars represent (e.g. SD, SE, CI)

Software and code

Policy information about availability of computer code

Data collection
- NCBI (19 Nov, 2014) , Ezbiocloud (23 Oct, 2017), KEGG (version 76), ARDB (Version 1.1), VFDB (2012), bacteriocin database (2013), CARD (2013)

Data analysis
- SOAPdenovo2(v r240), RNAMMer (v1.2), SPAdes (v3.1.0), CheckM (v1.0.12), specI (v1.0), prank (v.150803), trimal (v1.4.1), RAxML (v8.2.9), Genemark (v2.6r), RAPSearch (v2.23), BLASTp (v2.2.26), BPGA (v.1.3), RStudio (v1.0.153), R (v.3.3.3)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.
Data

Policy information about availability of data
All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:
- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The assembly draft genome and annotation information of the 1,520 strains have been deposited into the NCBI under accession number PRJNA482748 and the data are also available in the CNGB Nucleotide Sequence Archive (CNSA: https://db.cngb.org/cnsa; accession number CNP0000126).

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- Life sciences
- Behavioural & social sciences
- Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| Sample size | This study collected 155 feces, isolated 6,487 bacterial clones, and generated 1,520 high quality genomes, which is the largest dataset to date for gut bacterial reference genomes. |
| Data exclusions | No data were excluded from analyses. |
| Replication | not applicable; no biological experiments were involved. |
| Randomization | not applicable; no biological experiments were involved. |
| Blinding | not applicable; no biological experiments were involved. |

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

| Study description | Briefly describe the study type including whether data are quantitative, qualitative, or mixed-methods (e.g. qualitative cross-sectional, quantitative experimental, mixed-methods case study). |
| Research sample | State the research sample (e.g. Harvard university undergraduates, villagers in rural India) and provide relevant demographic information (e.g. age, sex) and indicate whether the sample is representative. Provide a rationale for the study sample chosen. For studies involving existing datasets, please describe the dataset and source. |
| Sampling strategy | Describe the sampling procedure (e.g. random, snowball, stratified, convenience). Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient. For qualitative data, please indicate whether data saturation was considered, and what criteria were used to decide that no further sampling was needed. |
| Data collection | Provide details about the data collection procedure, including the instruments or devices used to record the data (e.g. pen and paper, computer, eye tracker, video or audio equipment) whether anyone was present besides the participant(s) and the researcher, and whether the researcher was blind to experimental condition and/or the study hypothesis during data collection. |
| Timing | Indicate the start and stop dates of data collection. If there is a gap between collection periods, state the dates for each sample cohort. |
| Data exclusions | If no data were excluded from the analyses, state so OR if data were excluded, provide the exact number of exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established. |
| Non-participation | State how many participants dropped out/declined participation and the reason(s) given OR provide response rate OR state that no participants dropped out/declined participation. |
| Randomization | If participants were not allocated into experimental groups, state so OR describe how participants were allocated to groups, and if allocation was not random, describe how covariates were controlled. |
All studies must disclose on these points even when the disclosure is negative.

**Ecological, evolutionary & environmental sciences study design**

Briefly describe the study. For quantitative data include treatment factors and interactions, design structure (e.g. factorial, nested, hierarchical), nature and number of experimental units and replicates.

**Research sample**

Describe the research sample (e.g. a group of tagged Passer domesticus, all Stenocereus thurberi within Organ Pipe Cactus National Monument), and provide a rationale for the sample choice. When relevant, describe the organism taxa, source, sex, age range and any manipulations. State what population the sample is meant to represent when applicable. For studies involving existing datasets, describe the data and its source.

**Sampling strategy**

Note the sampling procedure. Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.

**Data collection**

Describe the data collection procedure, including who recorded the data and how.

**Timing and spatial scale**

Indicate the start and stop dates of data collection, noting the frequency and periodicity of sampling and providing a rationale for these choices. If there is a gap between collection periods, state the dates for each sample cohort. Specify the spatial scale from which the data are taken.

**Data exclusions**

If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.

**Reproducibility**

Describe the measures taken to verify the reproducibility of experimental findings. For each experiment, note whether any attempts to repeat the experiment failed OR state that all attempts to repeat the experiment were successful.

**Randomization**

Describe how samples/organisms/participants were allocated into groups. If allocation was not random, describe how covariates were controlled. If this is not relevant to your study, explain why.

**Blinding**

Describe the extent of blinding used during data acquisition and analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.

**Field work, collection and transport**

Describe the study conditions for field work, providing relevant parameters (e.g. temperature, rainfall).

State the location of the sampling or experiment, providing relevant parameters (e.g. latitude and longitude, elevation, water depth).

Describe the efforts you have made to access habitats and to collect and import/export your samples in a responsible manner and in compliance with local, national and international laws, noting any permits that were obtained (give the name of the issuing authority, the date of issue, and any identifying information).

Describe any disturbance caused by the study and how it was minimized.

**Reporting for specific materials, systems and methods**

| Materials & experimental systems | Methods |
|---------------------------------|---------|
| n/a                             | n/a     |
| ☒ Unique biological materials | ☒ ChIP-seq |
| ☒ Antibodies                    | ☒ Flow cytometry |
| ☒ Eukaryotic cell lines         | ☒ MRI-based neuroimaging |
| ☒ Palaeontology                 |         |
| ☒ Animals and other organisms   |         |
| ☒ Human research participants   |         |
Unique biological materials

**Policy information about availability of materials**

**Obtaining unique materials**

Describe any restrictions on the availability of unique materials OR confirm that all unique materials used are readily available from the authors or from standard commercial sources (and specify these sources).

**Antibodies**

**Antibodies used**

Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number.

**Validation**

Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer’s website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.

**Eukaryotic cell lines**

**Policy information about cell lines**

**Cell line source(s)**

State the source of each cell line used.

**Authentication**

Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.

**Mycoplasma contamination**

Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.

**Commonly misidentified lines**

(See ICLAC register)

Name any commonly misidentified cell lines used in the study and provide a rationale for their use.

**Palaeontology**

**Specimen provenance**

Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information).

**Specimen deposition**

Indicate where the specimens have been deposited to permit free access by other researchers.

**Dating methods**

If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.

Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

**Animals and other organisms**

**Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research**

**Laboratory animals**

For laboratory animals, report species, strain, sex and age OR state that the study did not involve laboratory animals.

**Wild animals**

Provide details on animals observed in or captured in the field; report species, sex and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.

**Field-collected samples**

For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.

**Human research participants**

**Policy information about studies involving human research participants**

**Population characteristics**

This study recruited 155 participants from China, including 79 male and 76 females; Among them 2 were age 1-10, 21 were age 11-20, 109 were age 21-50, and 23 were age >50; Please also see population characteristics in Supplementary Figure 1a.

**Recruitment**

All participants were recruited in ShenZhen, China; Fecal samples were collected from 155 healthy donors not taking any drugs during the last months prior to sampling;
### ChIP-seq

**Data deposition**

- Confirm that both raw and final processed data have been deposited in a public database such as GEO.
- Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

**Data access links**

For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.

**Files in database submission**

Provide a list of all files available in the database submission.

**Genome browser session**

(e.g. UCSC)

Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.

### Methodology

**Replicates**

Describe the experimental replicates, specifying number, type and replicate agreement.

**Sequencing depth**

Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.

**Antibodies**

Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.

**Peak calling parameters**

Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.

**Data quality**

Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.

**Software**

Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.

### Flow Cytometry

**Plots**

- The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- All plots are contour plots with outliers or pseudocolor plots.
- A numerical value for number of cells or percentage (with statistics) is provided.

**Methodology**

**Sample preparation**

Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.

**Instrument**

Identify the instrument used for data collection, specifying make and model number.

**Software**

Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.

**Cell population abundance**

Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.

**Gating strategy**

Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.

Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

### Magnetic resonance imaging

**Experimental design**

**Design type**

Indicate task or resting state; event-related or block design.
| Design specifications | Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials. |
|-----------------------|---------------------------------------------------------------------------------------------------------------------------------|
| Behavioral performance measures | State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects). |

### Acquisition

| Imaging type(s) | Specify: functional, structural, diffusion, perfusion. |
|------------------|--------------------------------------------------|
| Field strength | Specify in Tesla |
| Sequence & imaging parameters | Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle. |
| Area of acquisition | State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined. |
| Diffusion MRI |  |
| Used |  |
| Not used |  |

### Preprocessing

| Preprocessing software | Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.). |
|------------------------|------------------------------------------------------------------------------------------------------------------------------------------|
| Normalization | If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization. |
| Normalization template | Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized. |
| Noise and artifact removal | Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration). |
| Volume censoring | Define your software and/or method and criteria for volume censoring, and state the extent of such censoring. |

### Statistical modeling & inference

| Model type and settings | Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation). |
|-------------------------|------------------------------------------------------------------------------------------------------------------------------------------|
| Effect(s) tested | Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used. |
| Specify type of analysis: |  |
| Whole brain |  |
| ROI-based |  |
| Both |  |
| Statistical type for inference | (See Eklund et al. 2016) Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods. |
| Correction | Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo). |

### Models & analysis

| n/a | Involved in the study |
|-----|-----------------------|
|     | Functional and/or effective connectivity |
|     | Graph analysis |
|     | Multivariate modeling or predictive analysis |
| Functional and/or effective connectivity | Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information). |
| Graph analysis | Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.). |
| Multivariate modeling and predictive analysis | Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics. |