doi:10.1038/nature12198

1. Supplementary results

1.1. Associations between gut microbiota, glucose control and medication

Women with T2D who used metformin had increased levels of Enterobacteriaceae (i.e. *Escherichia*, *Shigella*, *Klebsiella* and *Salmonella*) and decreased levels of *Clostridium* and *Eubacterium* (Supplementary Fig. 5a and Supplementary Table 13). The abundance of *E. coli* correlated significantly with the levels of glucagon-like peptide 1 (GLP-1) (Supplementary Fig. 1b); interestingly, metformin has been shown to increase plasma GLP-1 levels[1]. Previous studies also showed increased *E. coli* and Proteobacteria in the faecal microbiota of diabetic patients, but no information about medication was provided in these reports[2,3]. A different pattern appeared when comparing women with good or poor blood glucose control, characterized by an increase in Lactobacillales, mainly *Streptococcus* species, and a decrease in species belonging to *Bacteroides*, *Eubacterium* and *Clostridium* in women with high HbA1c (Supplementary Fig. 5b and Supplementary Table 14). We found no differentially abundant MGCs in microbiota from T2D women with and without family history of diabetes; with or without medication; and with good or poor blood glucose control (Supplementary Fig. 5c,d).

At the functional level, different reporter pathways were significantly associated with the use of metformin and the degree of glucose control. In women taking metformin, the most enriched pathways included KOs for glutathione metabolism (e.g. glutathione synthase and reductase, *gshB* and *gor* genes), bacterial secretion (type I, II, III and VI) and *Vibrio cholera* pathogenic cycle (Supplementary Table 15). These results are in agreement with

the increased levels of Enterobacteriaceae associated with the use of metformin and indeed the functions correlated with the genera *Escherichia*, *Shigella, Yersinia*, and *Salmonella*. In the women with poor blood glucose control, we found significantly enriched KOs in phosphotransferase system (PTS) transporters (functions for the transport of glucose and lactose, which correlated mainly with *Collinsella* abundance), glutathione metabolism, defence against host immune system (*Staphylococcus aureus* infections, resistance to antimicrobial peptides), and several two-component systems (bacterial sensory pathways for sensing and responding to phosphate and nitrogen limitation, nitrogen assimilation and metabolism, multidrug efflux, antibiotic resistance, and outer membrane stress) (Supplementary Table 16 and Supplementary Fig. 17). Most of these functions correlated with enterobacteria but a small number correlated with *Eggerthella* (nitrogen assimilation and trimethylamine N-oxide metabolism), which contains opportunistic pathogens and was increased in Chinese T2D patients[3]. Spearman's rank correlation coefficients and P values for the correlations of genera abundance with KEGG KOs abundance are given in Supplementary Table 21.

The alterations in species and functions associated with metformin and glucose control might not be directly linked to T2D pathogenesis but might be a consequence of treatment and increased glucose availability in the intestinal environment.

1.2. Analysis of Chinese metagenomes with our bioinformatics pipeline

To determine if our bioinformatics pipeline was applicable to other metagenomic datasets, we analysed the metagenomic data from the study recently published by Qin et al[3], focusing on the association between T2D and the gut microbiota in Chinese subjects. Sequence data were downloaded from the Sequence Read Archives SRA045646 and

SRA050230. A total of 344 samples were included in the analysis and all data were processed according to the methods that were used to analyse the sequence data from our cohort (Methods online). Species composition was determined using the 2382 non-redundant reference genomes from the NCBI and HMP databases (Methods online). MGC composition was determined by aligning the Chinese sequence data to the gene catalogue obtained as described in Methods online and using the MGCs defined in our study as described in Methods online.

We observed that Bacteroidetes was the most abundant phylum (62±20%, (SD)) in the Chinese cohort, followed by Firmicutes (28±16%, (SD)) and Proteobacteria (5.5±8.2%, (SD)) (Supplementary Fig. 6). This distribution was different from that of our cohort, which was dominated by Firmicutes, then Bacteroidetes and Actinobacteria (Supplementary Fig. 7). Importantly, the distribution of phyla in our cohort was similar to that of a previous metagenome study of Europeans also performed by Qin et al[4]. Since the two studies by Qin et al[3,4] used similar protocols for DNA extraction and sequencing, the differences observed between our and the Chinese study likely reflect relevant biological dissimilarities and not methodological differences. However, it cannot be ruled out that other factors such as phenotyping of patients, DNA extraction and sequencing might differ between the two cohorts and thus may contribute to differences in findings. It should be noted that a recent study demonstrated large differences in the microbiota of different populations from different geographic locations[5]. Actinobacteria and Verrucomicrobia were less abundant in the Chinese cohort than in our cohort, while Fusobacteria was not a dominant phylum in our cohort (Supplementary Fig. 6 and 7). The most prevalent genera, species and genomes were thus different in the two cohorts (Supplementary Fig 6-9), and *Bacteroides* dominated in Chinese faecal communities, although *Faecalibacterium*

prausnitzii was the most abundant species in both cohorts (Supplementary Fig. 8a and 9a). Therefore, the Chinese and European populations clustered separately in principal component analysis plots of species and MGCs abundance (Supplementary Fig. 10).

Several species and MGCs were differentially abundant in Chinese T2D patients and healthy controls (Supplementary Fig. 11-12 and Supplementary Tables 17-18). In agreement with Qin et al[3], we found a number of clostridial MGCs significantly enriched in T2D Chinese metagenomes (Supplementary Table 18). Among these, two *Clostridium clostridioforme* MGCs (*C. clostridioforme*_262 and *C. clostridioforme*_346,) were also increased in T2D metagenomes from our cohort (Supplementary Table 9). We also found increased levels of three *Akkermansia* and two *Bacteroides* MGCs in Chinese T2D metagenomes (Supplementary Table 18), while *Lactobacillus gasseri*_361 was increased in T2D metagenomes from both cohorts (Supplementary Table 9 and 18). *Eubacterium*, *F. prausnitzii*, *Roseburia*, and other clostridial MGCs were significantly depleted in Chinese T2D metagenomes, as also shown by Qin et al[3]. The MGC *Roseburia*_272 in particular was the most significantly depleted in T2D subjects of both our and Qin's cohorts (Supplementary Tables 9 and 18). Furthermore, we found no *Haemophilus* MGCs but several *Haemophilus* species depleted in T2D Chinese metagenomes. Correlation analysis showed that the MGCs significantly enriched in Chinese T2D metagenomes positively correlated with fasting blood glucose (FBG) and HbA1c, while the MGCs significantly depleted in Chinese T2D metagenomes negatively correlated with FBG and HbA1C (Supplementary Fig. 13).

We classified T2D and healthy Chinese subjects with RF models based on species and MGCs as described in Methods online. We found a maximum AUC of 0.82 for the classification based on MGCs (Supplementary Fig. 14 and Supplementary Table 19). The

30 most important species and MGCs for the predictive models are shown in Supplementary Fig. 15, and they differed from those for our cohort (Fig. 3b,c).

2. Supplementary Figures

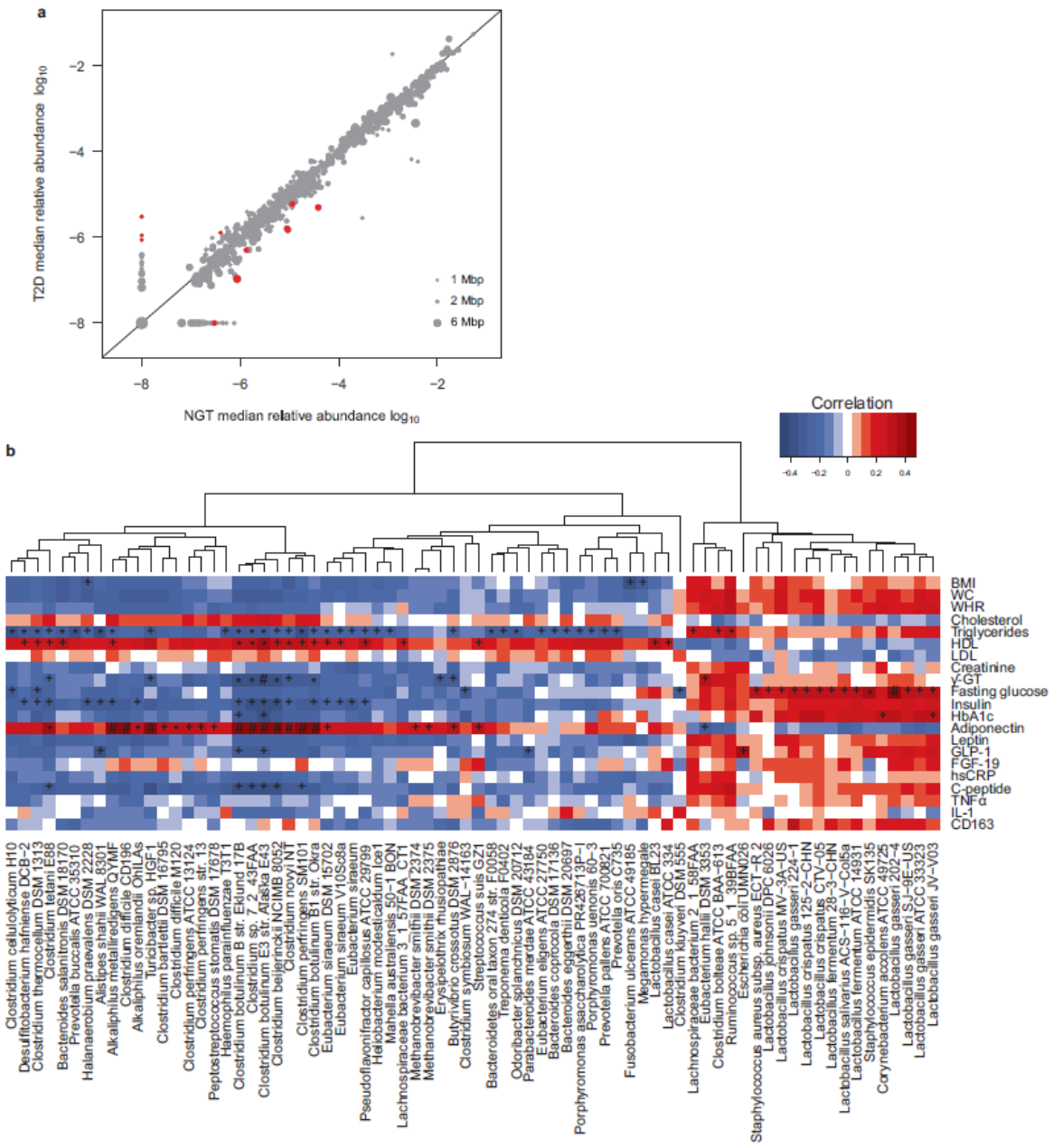

Figure S1 | Species abundance is associated with T2D and clinical biomarkers. a, Scatter plot of median species abundance in T2D (n=53) and NGT (n=43) subjects. Grey points represent species not differentially abundant between groups whereas red points represent species differentially abundant (Adj. P<0.05, Wilcoxon rank sum test). **b,** Spearman's rank correlation of clinical data from the total cohort (n=145) and species abundance + Adj. P<0.05; * Adj. P<0.01; # Adj. P< 0.001. Spearman's rank correlation coefficients and P values for the correlations are listed in Supplementary Table 7.

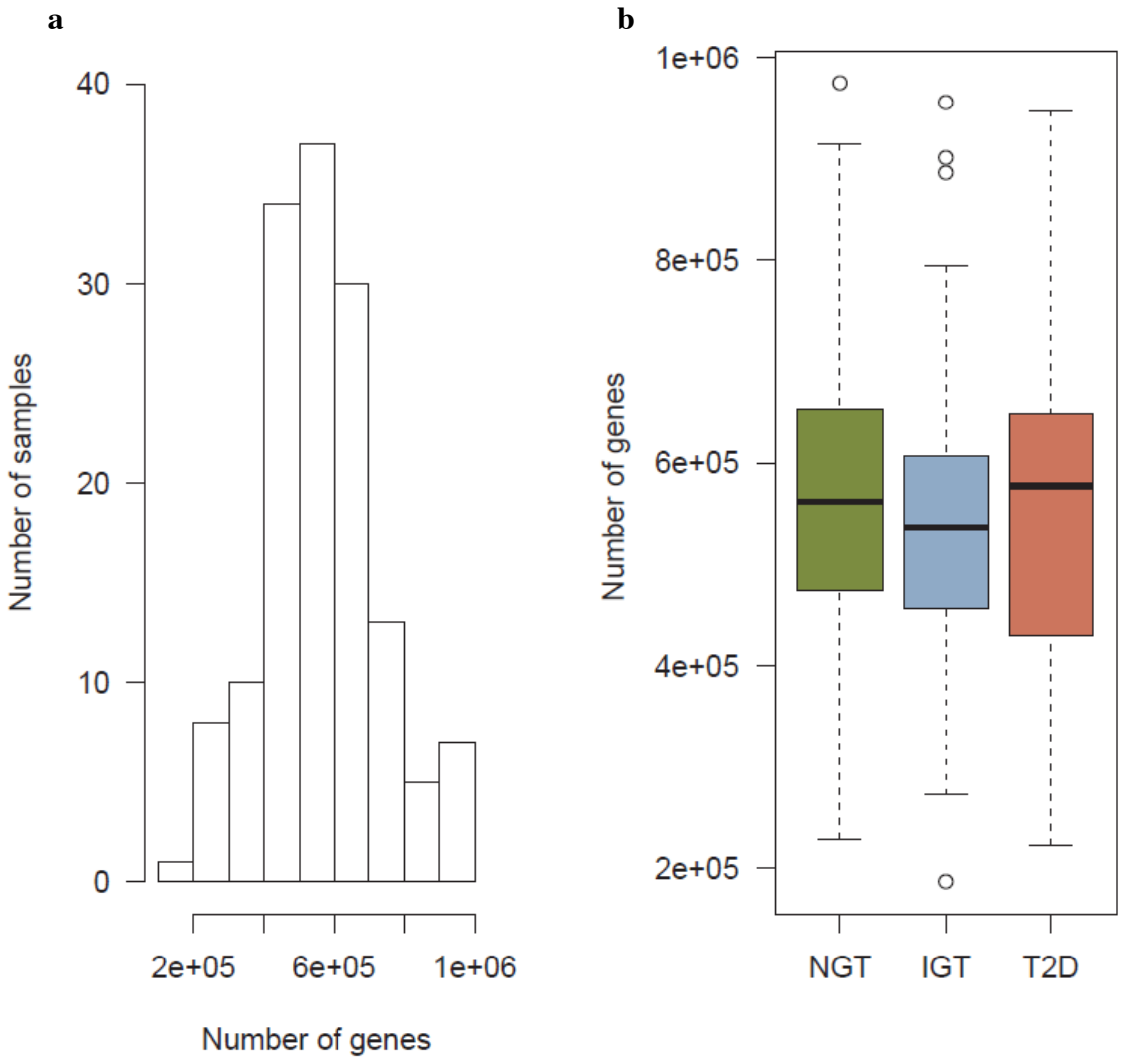

Figure S2 | Number of genes found in each individual and across clinical groups. a, Histogram of the number of genes in each individual. A minimum of two reads aligning to a gene was used to identify a gene. **b,** Number of genes in the three clinical groups shows that the number of genes does not differ with disease state. Boxes denote the interquartile range (IQR) between the first and third quartiles and the line within denotes the median; whiskers denote the lowest and highest values within 1.5 times IQR from the first and third quartiles, respectively. Circles denote data points beyond the whiskers.

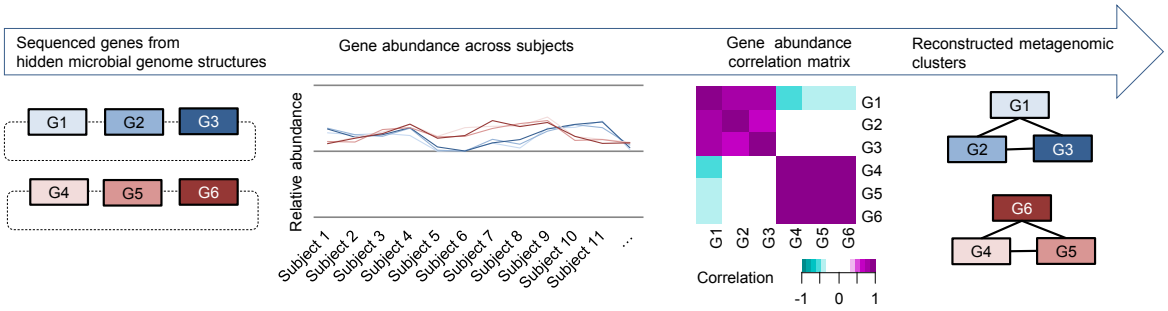

Figure S3 | Schematic diagram showing how metagenomic clusters (MGCs) were defined. By using the assumption that genes in the same genome should have similar abundance in a sample, genes that co-occur were clustered. Highly correlated genes in the matrix G1-3 and G4-6 cluster into two groups. The colors in the correlation matrix are not supposed to match with the rest; the colors in this matrix are indicated in the inset key where magenta means strong correlation.

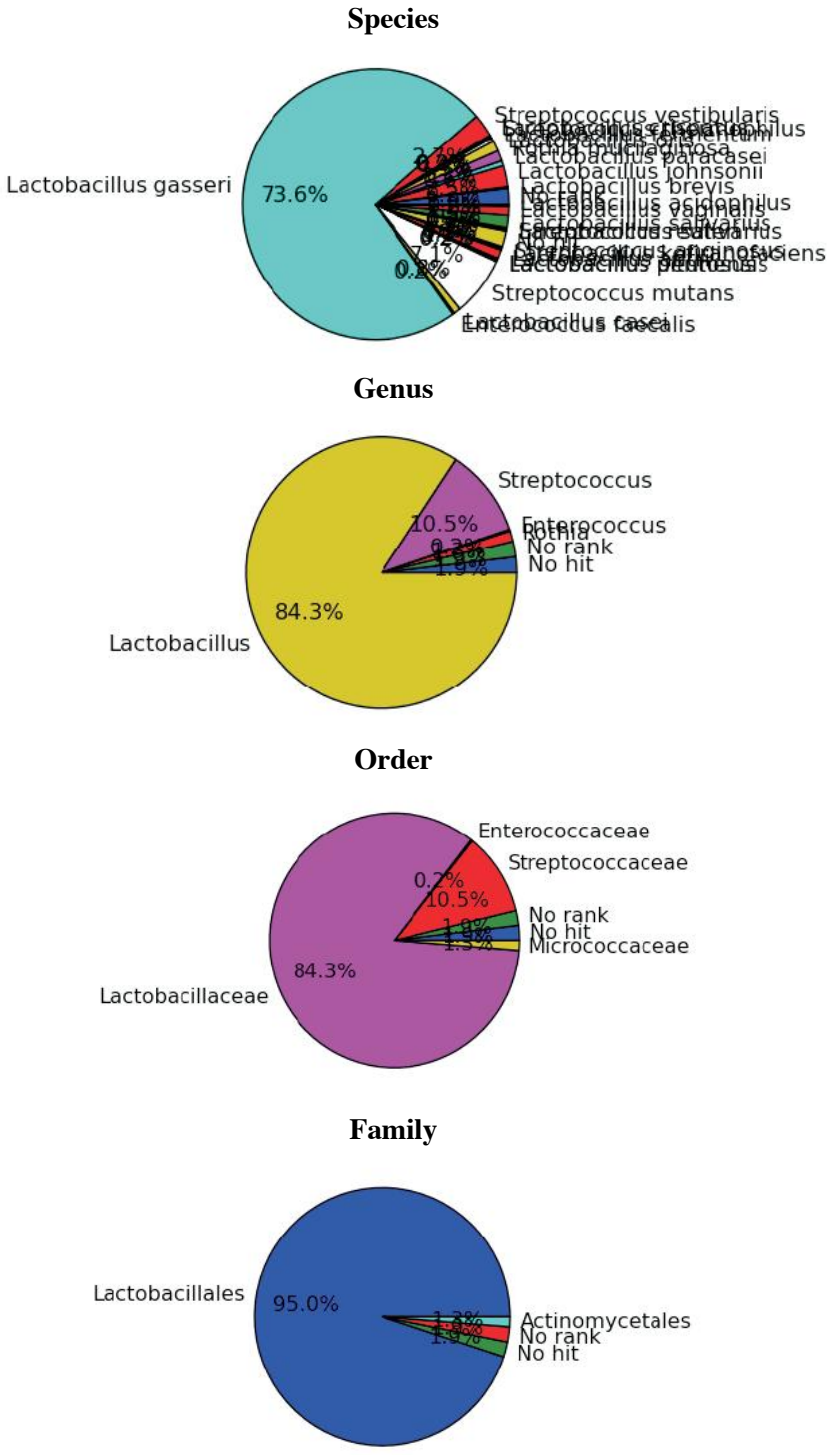

Figure S4 | Determining the lowest common ancestor (LCA) for a metagenomic cluster. Blast results for the metagenomic cluster Lactobacillus.gasseri_361 at different taxonomic levels. Blast results were used to determine the LCA by requiring that at least 50% of genes belonged to the same taxonomic entity.

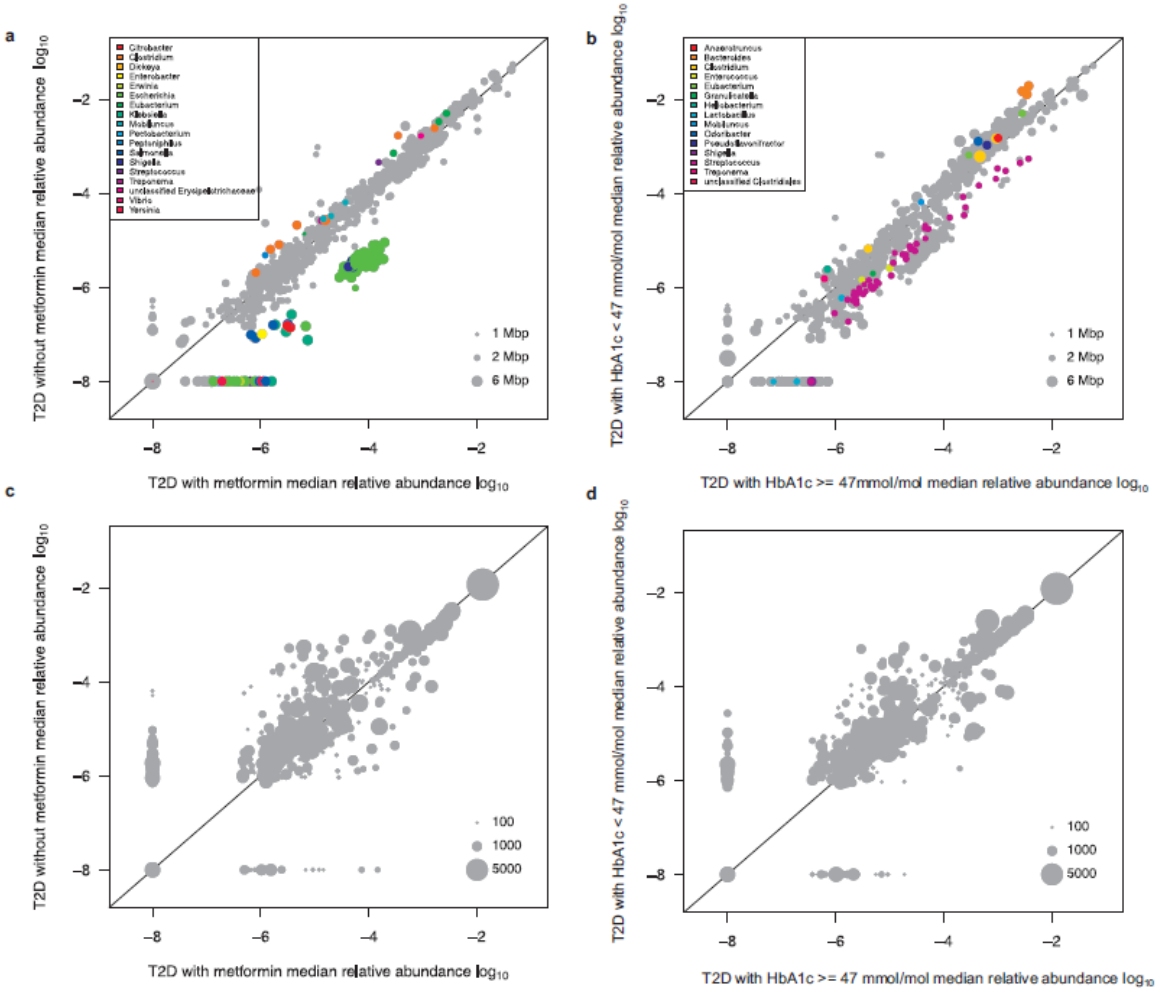

Figure S5 | Associations between gut microbiota composition, metformin and diabetic HbA1c levels in T2D subjects. a, Species abundance in T2D subjects with (n=20) or without (n=33) metformin. The list of species differentially abundant is given in Supplementary Table 13. **b**, Species abundance in T2D subjects with good (n=31) or poor (n=22) glucose control (<47 mmol/mol HbA1c indicates good control). The list of species differentially abundant is given in Supplementary Table 14. **c**, MGC abundance in T2D subjects with or without metformin. **d**, MGC abundance in T2D subjects with good or poor glucose control. Coloured species are differentially abundant (Adj. P <0.05, Wilcoxon rank sum test) and their genera are indicated by colour.

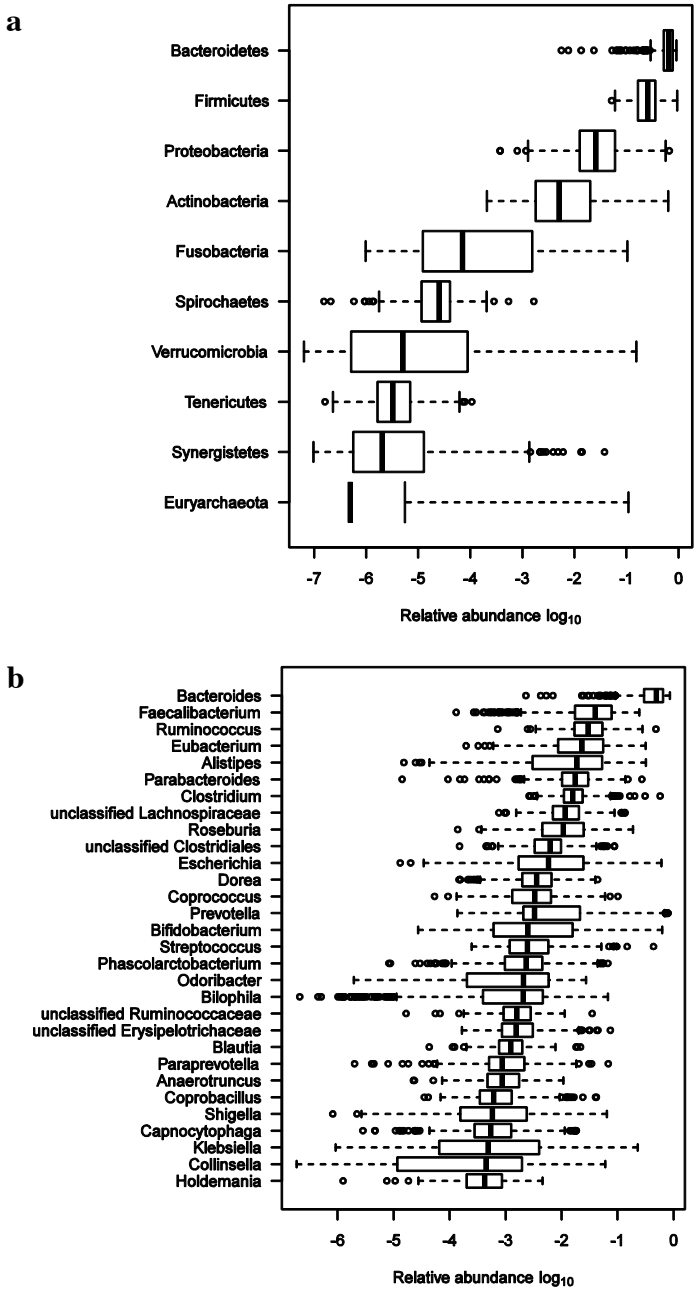

Figure S6 | Relative abundance of bacterial phyla and genera in Chinese metagenomes (n=344). a, 10 most abundant phyla. **b,** 30 most abundant genera. Boxes denote the interquartile range (IQR) between the first and third quartiles and the line within denotes the median; whiskers denote the lowest and highest values within 1.5 times IQR from the first and third quartiles, respectively. Circles denote data points beyond the whiskers.

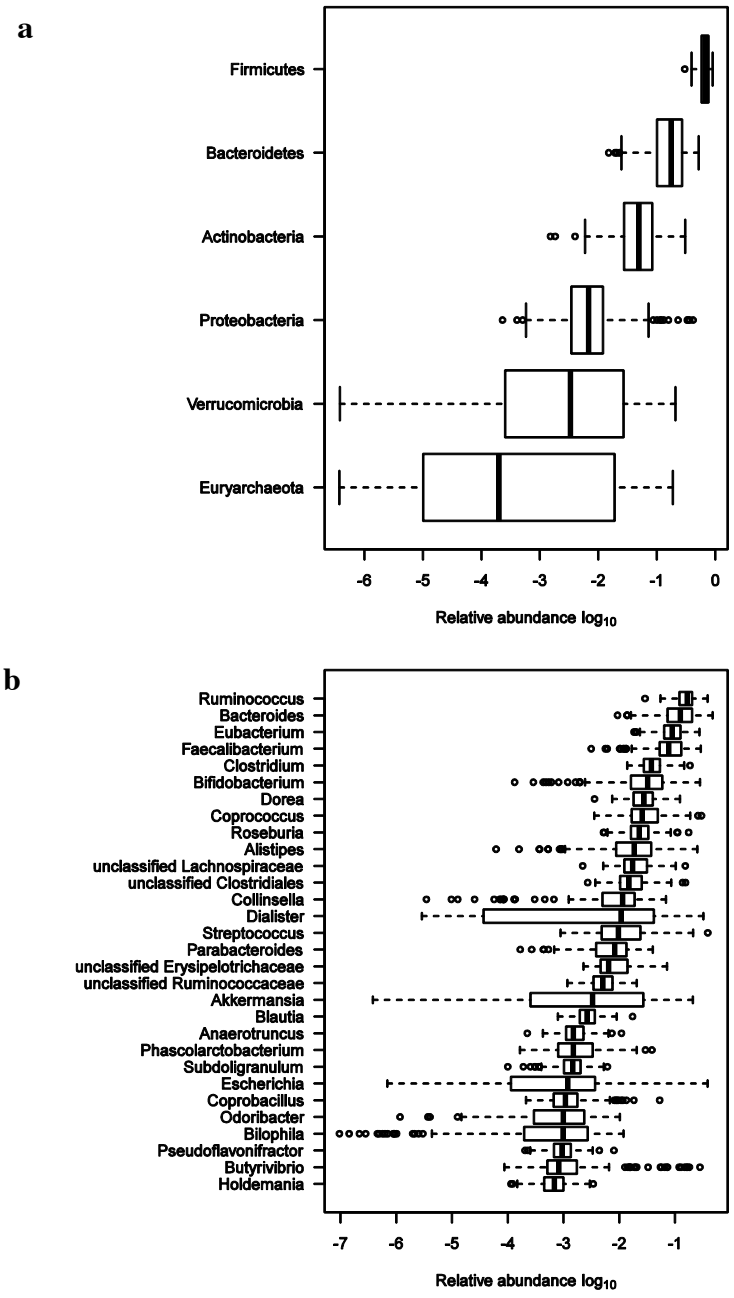

Figure S7 | Relative abundance of bacterial phyla and genera in our cohort (n=145). a, 6 most abundant phyla. **b,** 30 most abundant genera. Boxes denote the interquartile range (IQR) between the first and third quartiles and the line within denotes the median; whiskers denote the lowest and highest values within 1.5 times IQR from the first and third quartiles, respectively. Circles denote data points beyond the whiskers.

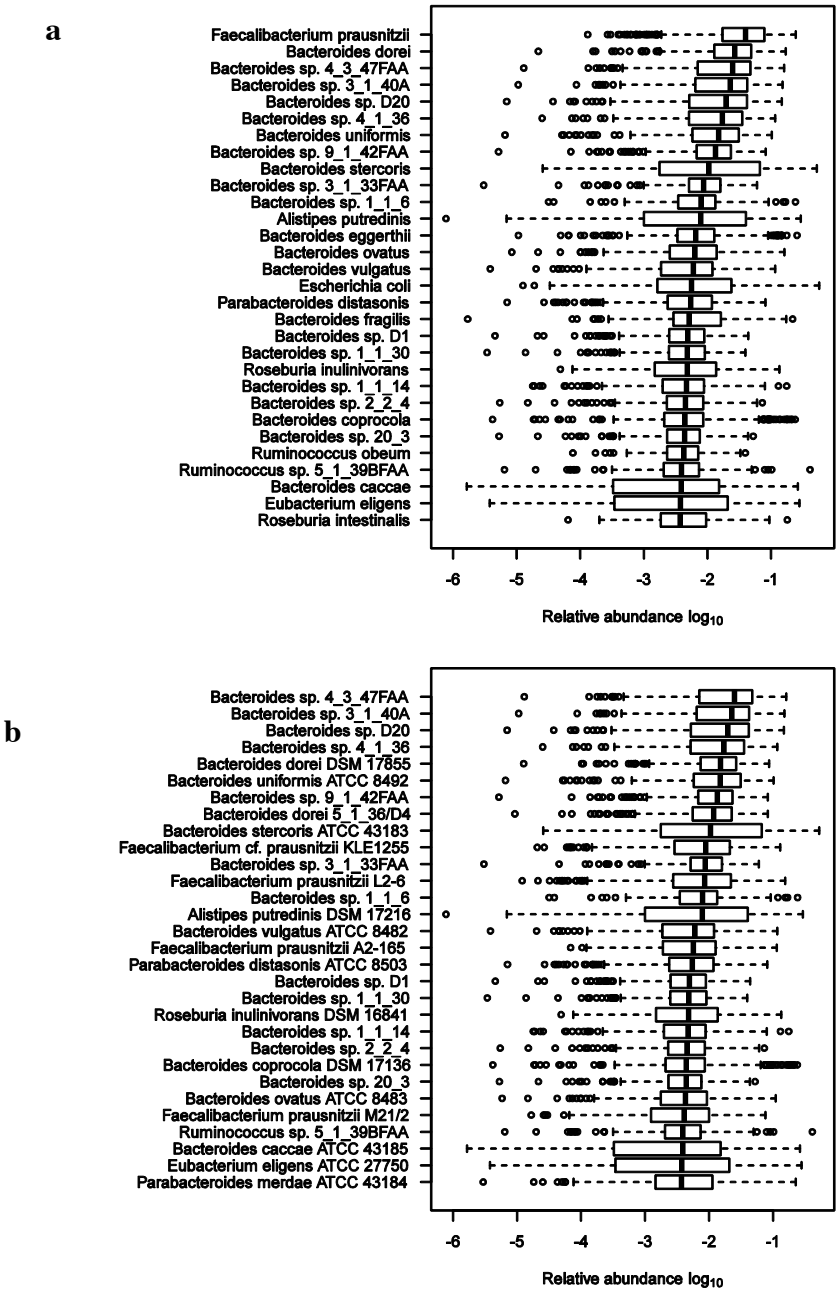

Figure S8 | Relative abundance of bacterial species and genomes in Chinese metagenomes (n=344). a, 30 most abundant species. **b,** 30 most abundant genomes. Boxes denote the interquartile range (IQR) between the first and third quartiles and the line within denotes the median; whiskers denote the lowest and highest values within 1.5 times IQR from the first and third quartiles, respectively. Circles denote data points beyond the whiskers.

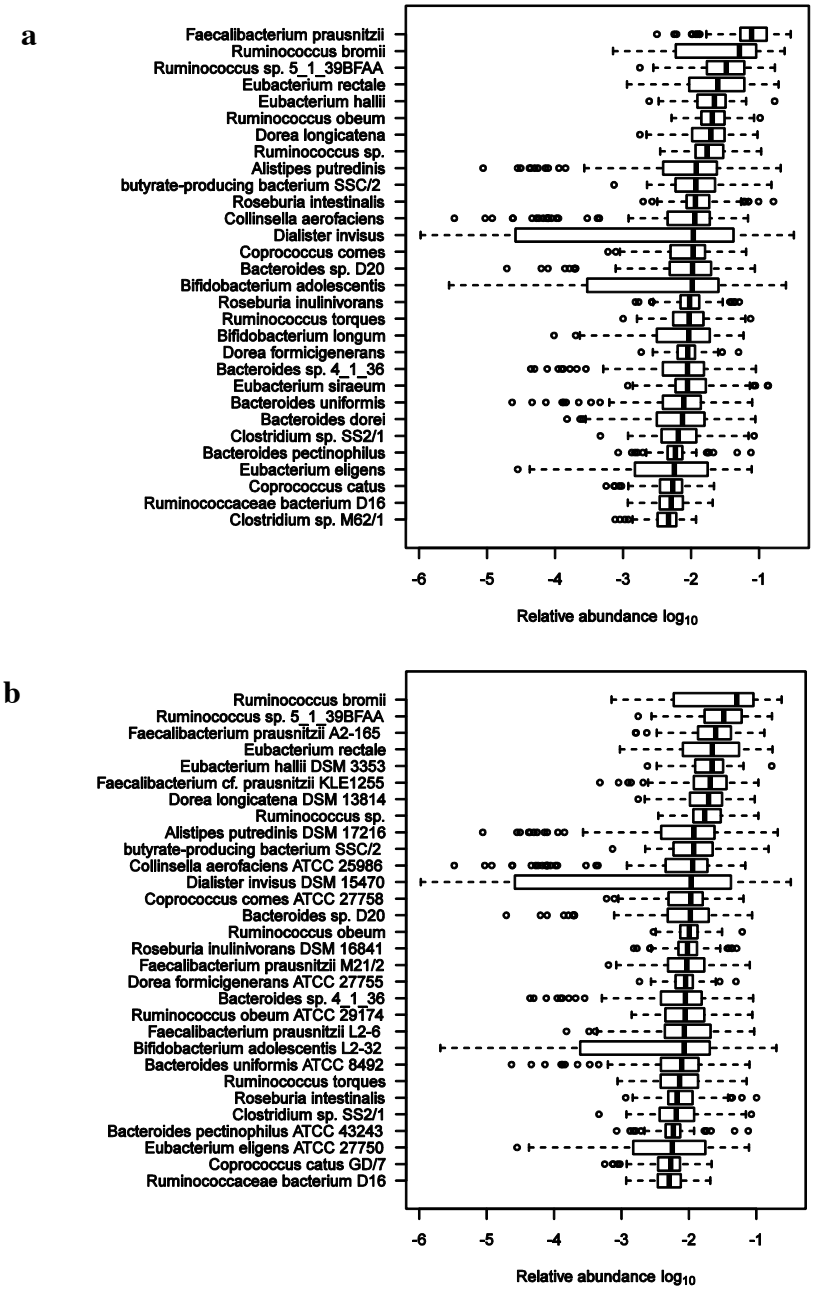

Figure S9 | Relative abundance of bacterial species and genomes in our cohort (n=145). a, 30 most abundant species. **b,** 30 most abundant genomes. Boxes denote the interquartile range (IQR) between the first and third quartiles and the line within denotes the median; whiskers denote the lowest and highest values within 1.5 times IQR from the first and third quartiles, respectively. Circles denote data points beyond the whiskers.

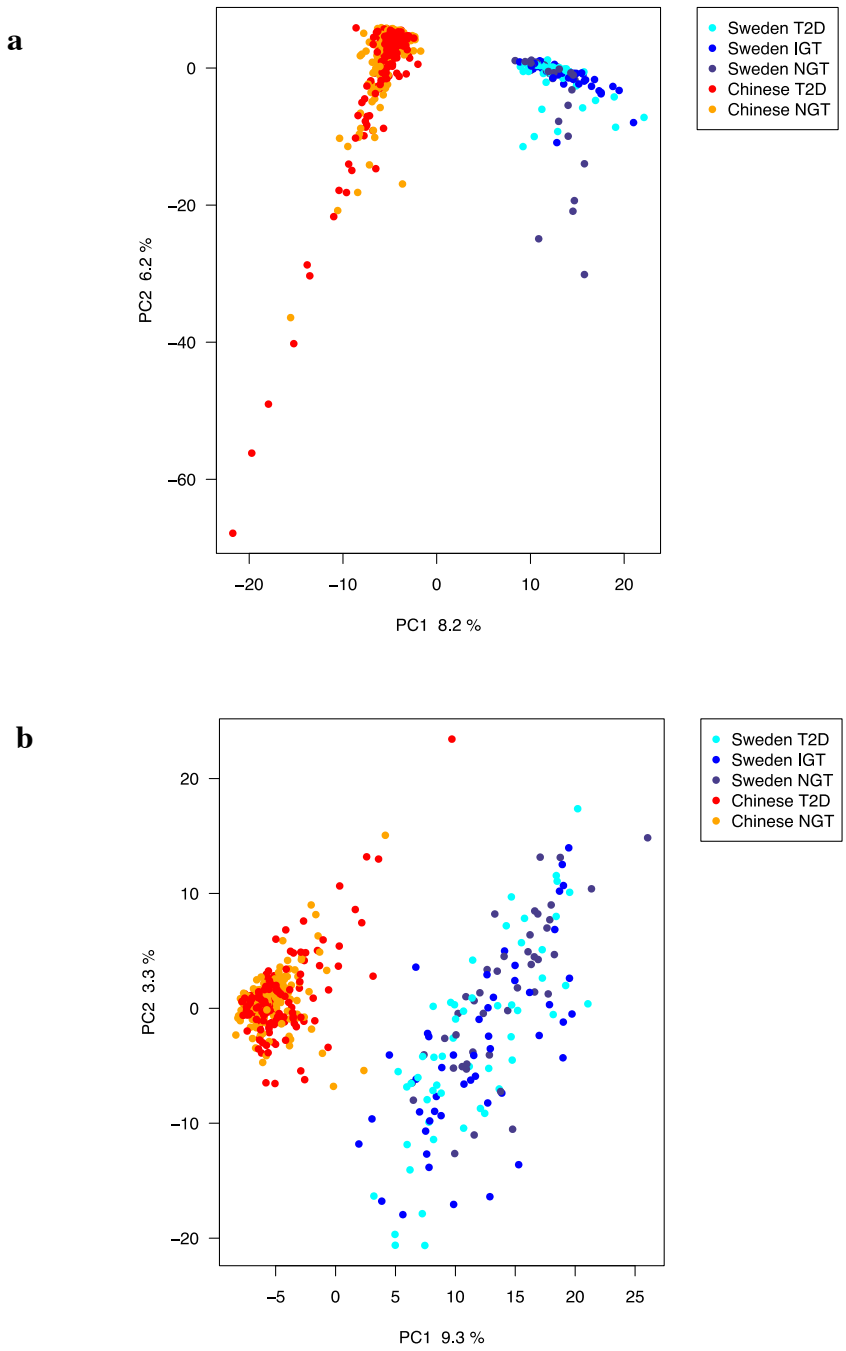

Figure S10 | Principal component analysis of microbial species and MGC abundance. a, Shared species with a maximum abundance above 1e-5 were included in the PCA analysis of the two cohorts showing a clear separation of Chinese and European (Swedish) subjects. b**,** PCA analysis of MGCs also showing a clear separation between the two cohorts.

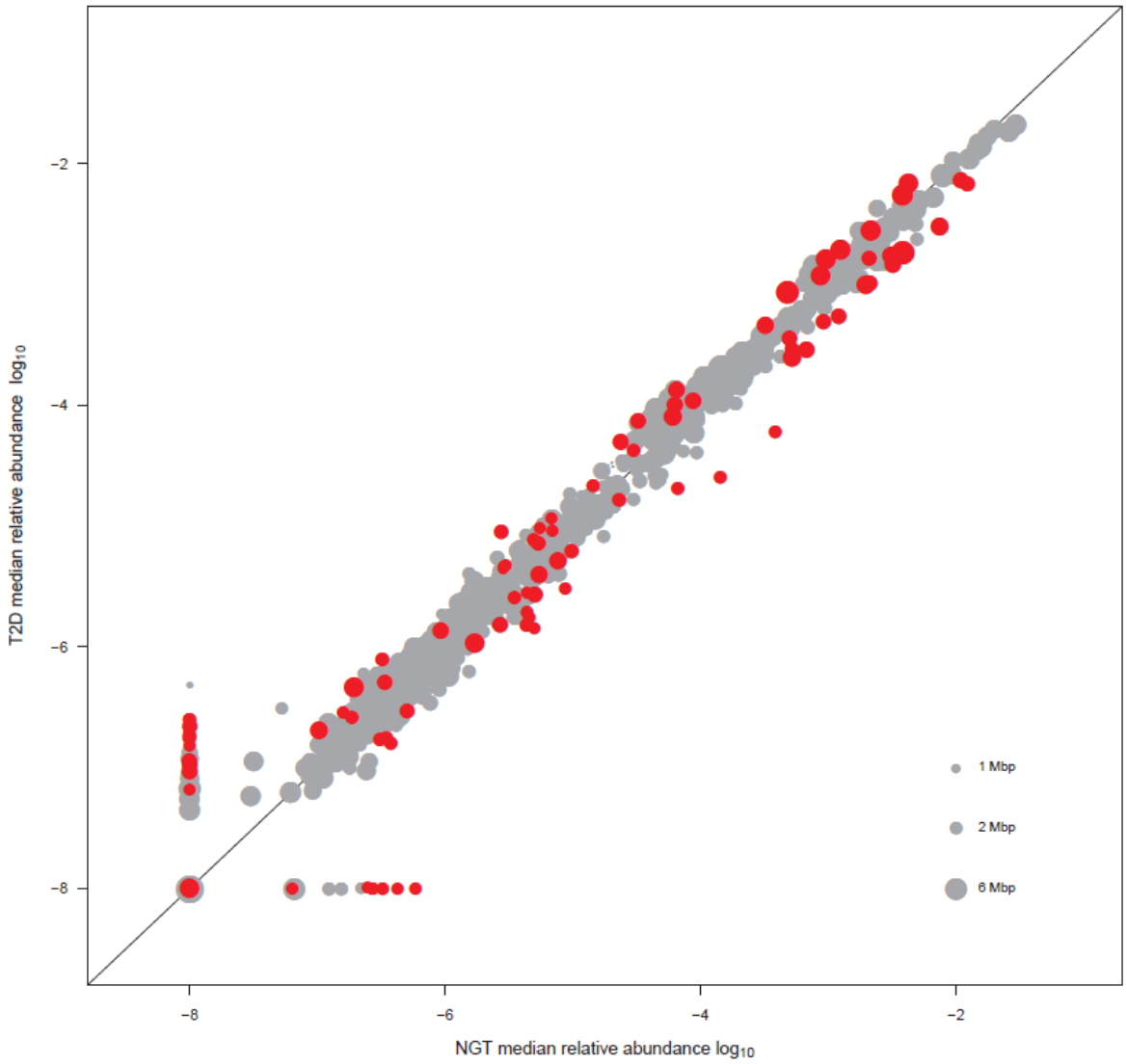

Figure S11 | Species abundance in T2D and control Chinese metagenomes (n=344). Scatter plot of median species abundance in T2D and control subjects. Grey points represent a species not differentially abundant between groups whereas red points represent species differentially abundant (Adj. P<0.05). The identities of species differentially abundant in T2D and control Chinese subjects are listed in Supplementary Table 17.

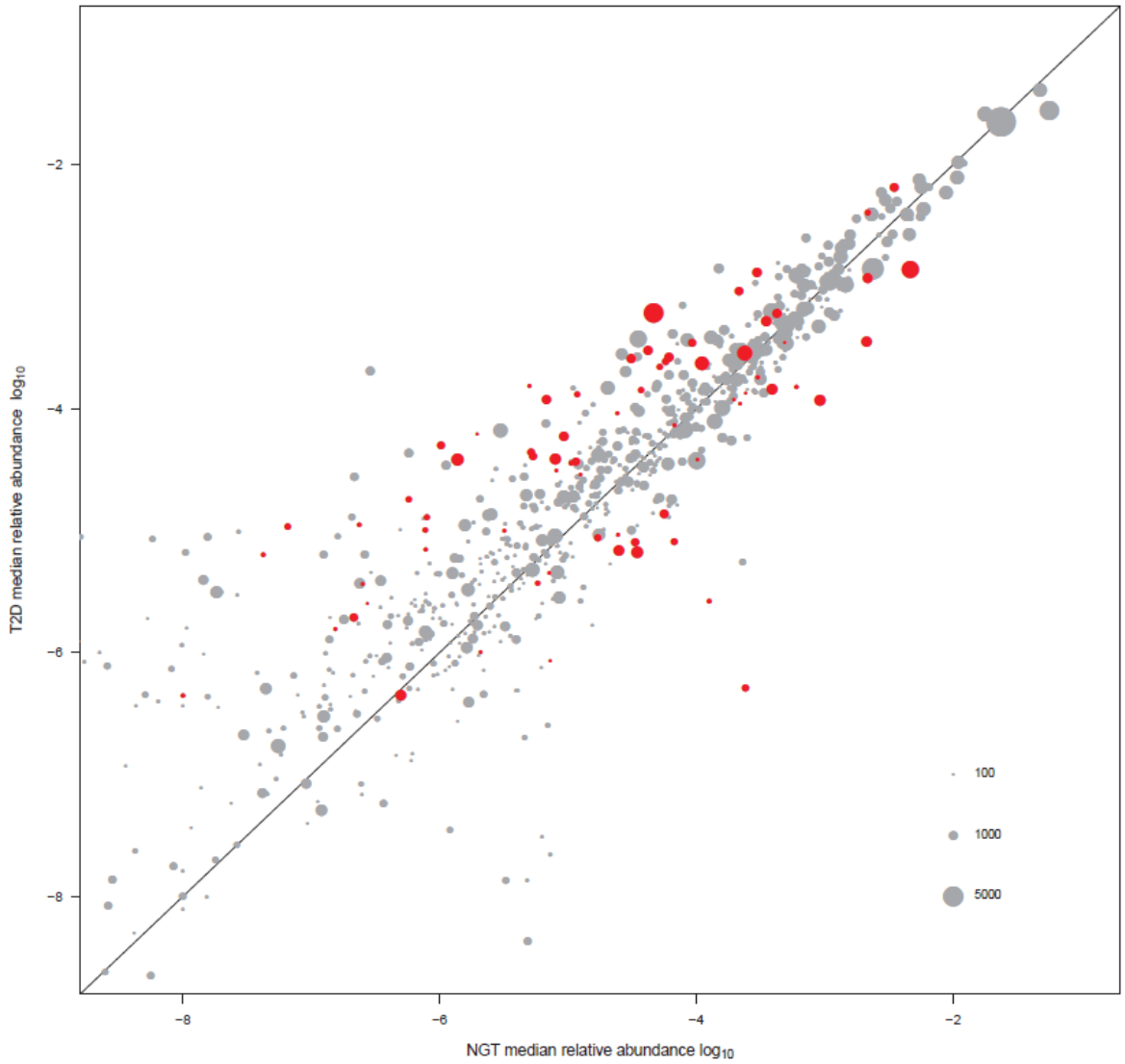

Figure S12 | MGC abundance in T2D and control Chinese metagenomes (n=344). Scatter plot of median MGC abundance in T2D and control subjects. Grey points represent a species not differentially abundant between groups whereas red points represent species differentially abundant (Adj. P<0.05). The identities of MGCs differentially abundant in T2D and control Chinese subjects are listed in Supplementary Table 18.

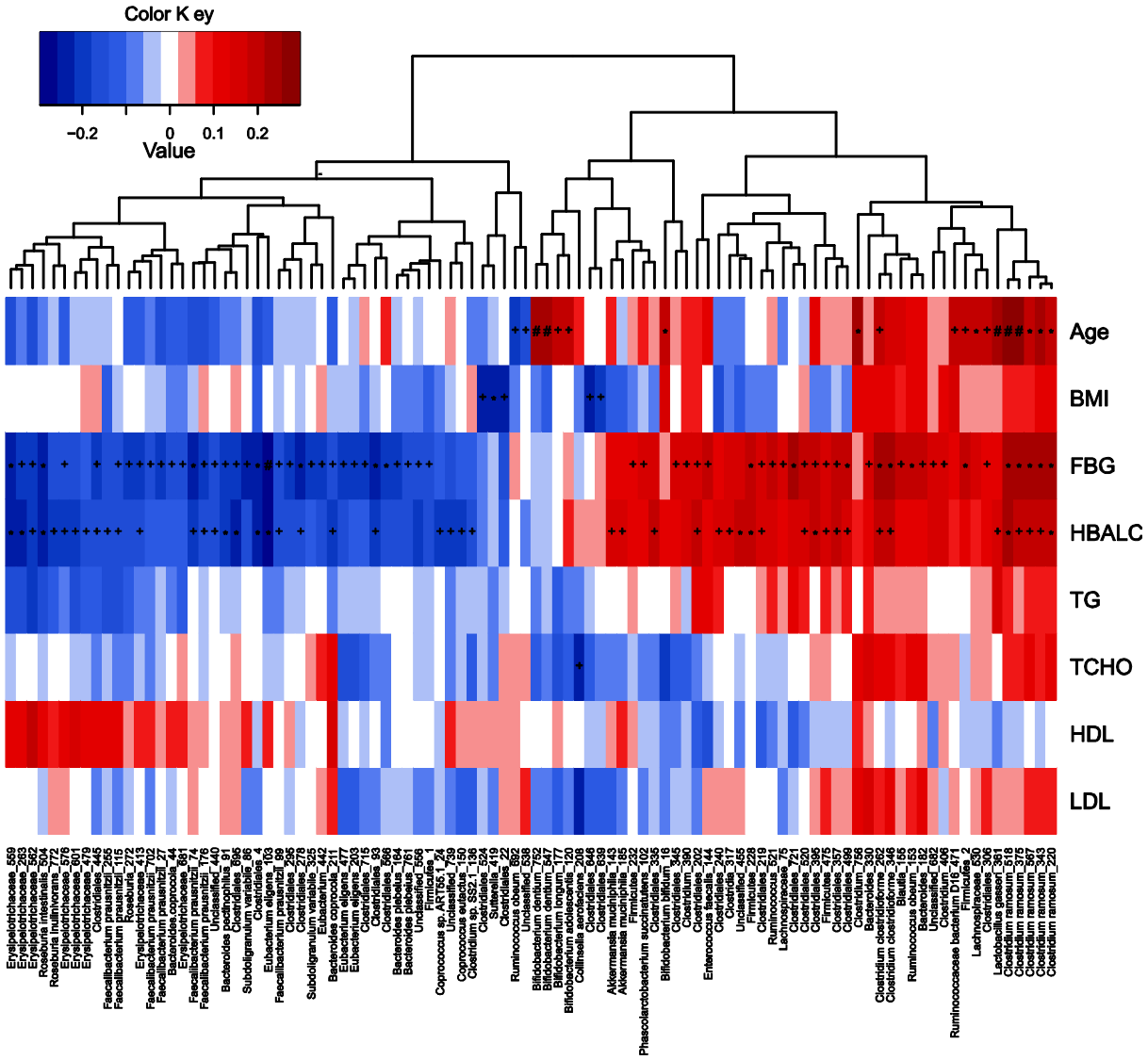

Figure S13 | Spearman's rank correlation of clinical data and MGC abundance for the Chinese cohort (n=344). + Adj. P<0.05; * Adj. P<0.01; # Adj. P< 0.001. HBALC, glycosylated haemoglobin HbAlc; TG, triglycerides; TCHO, total cholesterol.

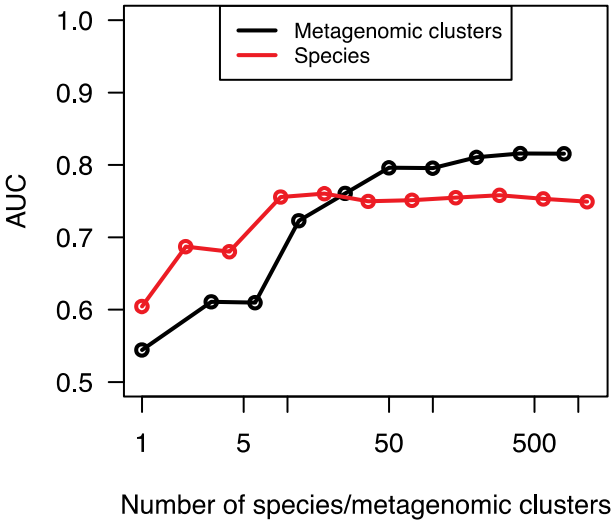

Figure S14 | Classification performance of random forest models for the Chinese metagenomes (n=344). The models use either species or MGC abundance. Performance is assessed by area under the receiver-operating characteristic curve (AUC) with different numbers of explanatory variables, ordered by importance. The AUC values for the two models including increasing numbers of species and MGCs are given in Supplementary Table 19.

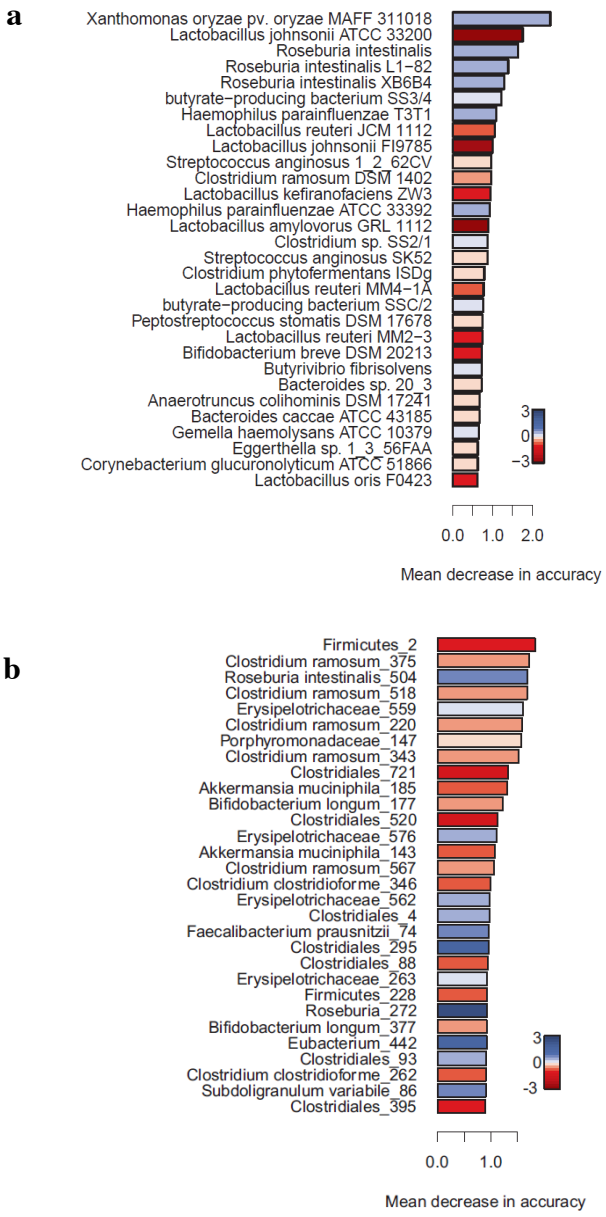

Figure S15 | Important species and MGCs in the predictive models for the classification of T2D and controls in the Chinese cohort. a, 30 most important species and **b**, 30 most important MGCs. Bar length indicates the importance of the variable and colours represent enrichment in T2D (red shades) or in NGT (blue shades).

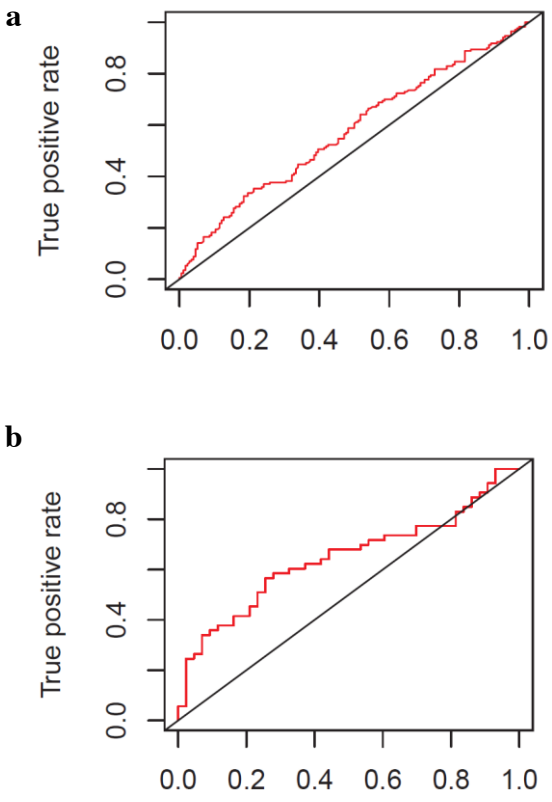

Figure S16 | Performance of MGC random forest models trained on one population for the classification of T2D individuals in a different population. a, MGC model trained on our population and used to classify T2D individuals from the Chinese population (AUC=0.58). **b,** MGC model trained on the Chinese population and used to classify T2D individuals from our population (AUC=0.66).

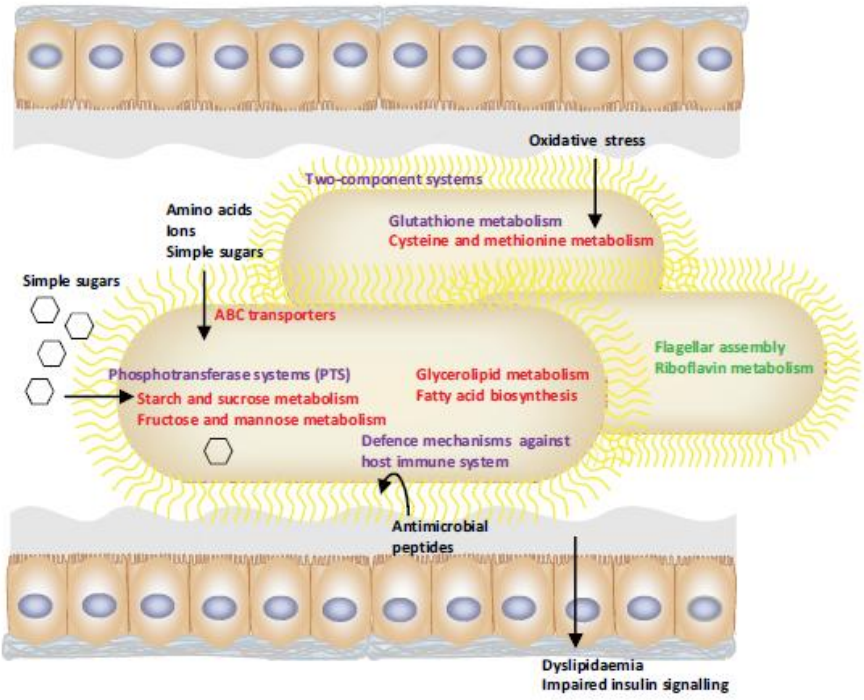

Figure S17 | Reporter pathways associated with NGT, T2D and poor glucose control. Functions enriched in NGT are indicated in green; functions enriched in T2D are indicated in red; functions enriched in subjects with poor glucose control (≥47 mmol/mol HbA1c) are indicated in purple.

Supplementary Tables

Supplementary Table 1. Characteristics of 70-year-old women with type 2 diabetes (T2D), impaired (IGT) and normal (NGT) glucose tolerance.

	T2D (n=53)	IGT (n=49)	NGT (n=43)	P value
Age, years	70.5±0.1	70.5±0.1	70.3±0.1	0.46
Body mass index (BMI), kg/m^2	28.4±0.7	26.9±0.6	25.8±0.7	0.017
Waist, cm	94.2±1.4	88.8±1.2	84.1±1.4	3.7e-06
HbA1c, mmol/mol	47.1±1.3	37.3±0.6	36.5±0.4	1.5e-15
HbA1c ≥ 47 mmol/mol, n (%)	22 (41)	0 (0)	0 (0)	1.7e-10
Serum fasting insulin, mU/L	12.70[a]±1.94	8.94[a]±0.77	6.97[a]±0.53	5.6e-06
Serum C-peptide, nmol/L	0.96[a]±0.08	0.84[a]±0.04	0.67[a]±0.04	0.00025
Serum HDL cholesterol, mmol/L	1.62±0.07	1.79±0.08	1.96±0.08	0.0058
Serum triglycerides, mmol/L	1.26[a]±0.11	1.19[a]±0.14	0.96[a]±0.08	0.017
Serum triglycerides >1.7 mmol/L, n (%)	14 (26)	11 (22)	1 (2)	0.0055
Statin treatment, n (%)	26 (49)	16 (33)	10 (23)	0.027
Insulin treatment, n (%)	6 (11)	0 (0)	0 (0)	0.0044
Oral antidiabetic medication, n (%)	22 (41)	0 (0)	0 (0)	1.7e-10

Differences between groups were analysed with linear regression for continuous variables after log transformation of skewed variables and Chi-squared test for categorical variables. Values are mean ± standard error of the mean if not stated otherwise. [a]Geometric mean.

Supplementary Table 2. Change in glucose tolerance status over a mean of 5.6 years follow-up.

Classification at re-examination	Classification at baseline		
	T2D, n	IGT, n	NGT, n
T2D, n=53	47	6	0
IGT, n=49	0	31	18
NGT, n=43	0	6	37

Supplementary Tables 3-21 are provided in a separate excel file.

3. References

1 Maida, A., Lamont, B. J., Cao, X. & Drucker, D. J. Metformin regulates the incretin receptor axis via a pathway dependent on peroxisome proliferator-activated receptor-alpha in mice. *Diabetologia* 54, 339-349, doi:10.1007/s00125-010-1937-z (2011).

2 Larsen, N. *et al.* Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults. *PLoS ONE* 5, e9085, doi:10.1371/journal.pone.0009085 [doi] (2010).

3 Qin, J. *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, doi:10.1038/nature11450 (2012).

4 Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464, 59-65, doi:nature08821 [pii]
10.1038/nature08821 [doi] (2010).

5 Yatsunenko, T. *et al.* Human gut microbiome viewed across age and geography. *Nature* 486, 222-227, doi:nature11053 [pii]
10.1038/nature11053 [doi] (2012).