doi:10.1038/nature11450

Supplementary Methods

Sample collection and DNA extraction

All 368 faecal samples from 368 Chinese individuals living in the south of China, were collected by Shenzhen Second People's Hospital, Peking University Shenzhen Hospital and Medical Research Center of Guangdong General Hospital, including 345 samples for MGWAS and an additional set of 23 samples for T2D classification. The patients who were diagnosed with Type 2 Diabetes Mellitus according to the 1999 WHO criteria[38] constituted the case group in this study, and the rest non-diabetic individuals were taken as the control group. Detailed clinical information of all sample donors was collected and presented in Supplementary Table 1. The patients and healthy controls were asked to provide a frozen faecal sample. Fresh faecal samples were obtained at home, and samples were immediately frozen by storing in a home freezer for less than 1d. Frozen samples were transferred to BGI-Shenzhen, and then stored at -80°C until analysis.

A frozen aliquot (200 mg) of each fecal sample was suspended in 250 µl of guanidine thiocyanate, 0.1 M Tris (pH 7.5) and 40 µl of 10% N-lauroyl sarcosine. DNA was extracted as previously described[39]. DNA concentration and molecular weight were estimated using a nanodrop instrument (Thermo Scientific) and agarose gel electrophoresis, respectively.

DNA library construction and sequencing

DNA library construction was performed following the manufacturer's instruction (Illumina). We used the same workflow as described elsewhere to perform cluster generation, template hybridization, isothermal amplification, linearization, blocking and denaturation, and hybridization of the sequencing primers.

We constructed one paired-end (PE) library with insert size of 350bp for each samples, followed by a high-throughput sequencing to obtain around 20 million PE reads. The reads length for each end is 75bp-90bp (75bp and 90bp read length in stage I samples; 90bp read length for stage II samples). High quality reads were extracted by filtering low quality reads with 'N' base, adapter contamination or human DNA contamination from the Illumina raw data. On average, the proportion of high quality reads in all samples was about 98.1%, and the actual insert size of our PE library ranges from 313bp to 381bp.

Gene catalogue construction

Gene catalogue updating. Employing the same parameters that were used to construct the MetaHIT gene catalogue[40], we performed de novo assembly and gene prediction for the high quality reads of 145 samples in stage I using SOAPdenovo v1.06[41] and GeneMark v2.7[42], respectively. All predicted genes were aligned pairwise using BLAT and genes, of which over 90% of their length can be aligned to another one with more than 95% identity (no gaps allowed), were removed as redundancies, resulting in a non-redundant gene catalogue comprising of 2,088,328 genes. This gene catalogue from our Chinese samples was further combined with the previously constructed MetaHIT gene catalogue[40], by removing redundancies in the same manner. At last, we obtained an updated gene catalogue that contains 4,267,985 genes.

Taxonomic assignment of genes. Taxonomic assignment of the predicted genes was performed using an in-house pipeline. In our analysis, we collected the reference microbial genomes from IMG database (v3.4, see the full list in Supplementary Table 3), and then aligned all 4.3 million genes onto the reference genomes. Based on the comprehensive parameter exploration of sequence similarity across phylogenetic

ranks by MetaHIT enterotype paper[43], we used the 85% identity as the threshold for genus assignment, as well as another threshold of 80% of the alignment coverage. For each gene, the highest scoring hit(s) above these two thresholds was chosen for the genus assignment. For the taxonomic assignment at the phylum level, the 65% identity was used instead.

Functional annotation. We aligned putative amino acid sequences, which translated from the updated gene catalogue, against the proteins/domains in eggNOG (v3.0) and KEGG databases (release 59.0) using BLASTP (e-value ≤1e-5). Each protein was assigned to the KEGG orthologue group (KO) or eggNOG orthologue group (OG) by the highest scoring annotated hit(s) containing at least one HSP scoring over 60 bits. For the remaining genes without any annotation in eggNOG database, we identified novel gene families based on clustering all-against-all BLASTP results using MCL with an inflation factor of 1.1 and a bit-score cutoff of 60[44]. Using this approach, we identified 7,042 novel gene families (≥20 proteins) from the updated gene catalogue.

Quantification of metagenome content

Computation of relative gene abundance. The high quality reads from each sample were aligned against the gene catalogue by SOAP2 using the criterion of "identity > 90%". In our sequence-based profiling analysis, only two types of alignments could be accepted: i). an entire of a paired-end read can be mapped onto a gene with the correct insert-size; ii). one end of the paired-end read can be mapped onto the end of a gene, only if the other end of read was mapped outside the genic region. In both cases, the mapped read was counted as one copy.

Then, for any sample S, we calculated the abundance as follows:

Step 1: Calculation of the copy number of each gene:

$$b_i = \frac{x_i}{L_i}$$

Step 2: Calculation of the relative abundance of gene i

$$a_i = \frac{b_i}{\sum_j b_j} = \frac{\frac{x_i}{L_i}}{\sum_j \frac{x_j}{L_j}}$$

a_i: The relative abundance of gene i in sample S.

L_i: The length of gene i.

x_i: The times which gene i can be detected in sample S (the number of mapped reads).

b_i: The copy number of gene i in the sequenced data from sample S.

Estimation of profiling accuracy. We used the method developed by Audic and Claverie (1997)[45] to assess the theoretical accuracy of the relative abundance estimates. Given that we have observed x_i reads from gene i, as it occupied only a small part of total reads in a sample, the distribution of x_i is approximated well by a Poisson distribution. Let us denote N the total reads number in a sample, so $N = \sum_i x_i$. Suppose all genes are the same length, so the relative abundance value a_i of gene i simply is $a_i = x_i/N$. Then we could estimate the expected probability of observing y_i reads from the same gene i, is given by the formula below,

$$P(a_i'|a_i) = P(y_i|x_i) = \frac{(x_i + y_i)!}{x_i!\, y_i!\, 2^{(x_i+y_i+1)}}$$

Here, $a_i' = y_i/N$ is the relative abundance computed by y_i reads (See the original paper[45] for details). Based on this formula, we then made a simulation by setting the value of a_i from 0.0 to 1e-5 and N from 0 to 40 million, in order to compute the 99% confidence interval for a_i' and to further estimate the detection error rate (Supplementary Figure 3).

Construction of gene, KO, and OG profile

The updated gene catalogue contains 4,267,985 non-redundant genes, which can be classified into 6,313 KOs and 45,683 OGs (including 7,042 novel gene families). We first removed genes, KOs or OGs that were present in less than 6 samples across all 145 samples in stage I. To reduce the dimensionality of the statistical analyses in MGWAS, in the construction of gene profile, we identified highly correlated gene pairs and then subsequently clustered these genes using a straightforward hierarchical clustering algorithm. If the Pearson correlation coefficient between any two genes is >0.9, we assigned an edge between these two genes. Then, the cluster A and B would not be clustered, if the total number of edges between A and B is smaller than |A|*|B|/3, where |A| and |B| are the sizes of A and B, respectively. Only the longest gene in a gene linkage group was selected to represent this group, yielding a total of 1,138,151 genes. These 1,138,151 genes and their associated measures of relative abundance in 145 stage I samples were used to establish the gene profile for the association study.

For the KO profile, we utilized the gene annotation information of the original 4,267,985 genes and summed the relative abundance of genes from the same KO. This gross relative abundance was taken as the content of this KO in a sample to generate the KO profile of the samples. The OG profile was constructed using the same methods.

Bio-diversity analysis

Within-sample diversity. Based on the gene profile, we calculated the within-sample (alpha) diversity to estimate the gene richness of a sample using Shannon index:

$$H' = -\sum_{i=1}^{S} a_i \ln a_i$$

Where S is the number of genes and a_i is the relative abundance of gene i as defined above. A high alpha diversity indicates a high richness of genes in the sample.

Rarefaction curve analysis. To assess the gene or gene family richness in our cohorts, we generated a rarefaction curve. For a given number of individual samples, we performed a randomized sampling 100 times in the T2D patients group or non-diabetic control group, respectively. Further, we randomly selected the given number of individual samples and then calculated the total number of genes/families that could be identified from these samples. Only the genes with ≥2 mapped reads and gene families with ≥10 mapped reads were determined to be present in a sample to eliminate the incorrectly identification. Actually, the conclusion that the gene/families richness of T2D patients group is higher than that of non-diabetic group was not changed under different thresholds.

Enterotypes identification

The genus relative abundance profile was constructed using the same methods as the KO/OG profile. After that, the genus profile was used for identifying enterotypes from our Chinese samples. We used the same identification method as described in the original paper of enterotypes[43]. In this study, samples were clustered using Jensen-Shannon distance and was then illustrated by PCA (Principal Component Analysis) graph that was implemented in "ade4" package in R software[46].

Statistical analysis of MGWAS

PERMANOVA analysis. In this study, the PERMANOVA (Permutational multivariate analysis of variance)[47] was used to assess the effect of different covariate, such as enterotypes, T2D, age, gender and BMI, on all types of profiles. We performed the analysis using the method implemented in R package - "vegan"[48], and the permuted *P*-value was obtained by 10,000 times permutations.

Population stratification. To correct population stratifications of our metagenome -wide data, we used a modified version of the EIGENSTRAT method[49] allowing the use of covariance matrices estimated from abundance levels instead of genotypes. However, as much of the signal in our data might be driven by the combined effect of many genes and not by just a few genes as assumed in GWAS studies, we modified the method further by replacing each PC axis with the residuals of this PC axis from a regression to T2D state. The number of PC axes of EIGENSTAT was determined by Tracy-Widom test at a significance level of $P<0.05$[50].

Statistical hypothesis test on profiles. In stage I, to identify the association between the metagenome profile and T2D, a two-tailed Wilcoxon rank-sum test was used in the profiles that were adjusted for non-T2D-realted population stratifications. Then, while examining the stage I markers in stage II, a one-tailed Wilcoxon rank-sum test was used instead. Because the T2D is the primary factor impacting on the profile of examined gene markers in stage II, we didn't adjust the population stratification for these genes.

Estimating the false discovery rate (FDR) and the power. Instead of a sequential *P*-value rejection method, we applied the "qvalue" method proposed in a previous study[51] to estimate the FDR. In our MWAS, the statistical hypothesis tests were

performed on a large number of features of the gene, KO and OG profiles. Given that a FDR was obtained by the qvalue method[52], we estimated the power P_e for a given *P*-value threshold by the formula below,

$$P_e = \frac{N_e(1 - FDR_e)}{N(1 - \pi_0)}$$

Here, π_0 is the proportion of null distribution *P*-values among all tested hypotheses; N_e is the number of *P*-values that were less than the *P*-value threshold; N is the total number of all tested hypotheses; FDR_e is the estimated false discovery rate under the *P*-value threshold.

Taking the gene profile as an example, the estimated FDR and power for gene markers of stage II are shown in Fig. 1c.

Identification of MetaHIT IBD-associated markers

To identify the IBD-associated gene and OG markers for the 124 MetaHIT samples[40], firstly, we performed stratified sampling to these samples and obtained a subgroup of 25 IBD patients and 47 control samples (see the following table).

	No. of Samples	Nation	Gender (M/F)	Age (mean±sd)	BMI (mean±sd)
IBD patients	25	Spain=25	10/15	44.8±10.8	24.6±4.2
Controls	47	Spain=14,Denmark=33	19/28	48.6±10.4	24.7±4.1
				P=0.154	*P*=0.906

Then, we calculated the gene and OG relative abundance profiles for these samples by the methods as described before. Using a two-tailed Wilcoxon rank-sum test, we identified 151,039 IBD-associated gene markers ($P<0.01$, corresponding to 4.7% FDR) and 7,680 IBD-associated OG markers ($P<0.05$, 9.7% FDR).

Identification of Metagenomic Linkage Group (MLG)

The clustering method for identifying MLG. In the present study, we devised a concept of metagenomic linkage group (MLG), which could facilitate the taxonomic description of metagenomic data from whole-genome shotgun sequencing. To identify MLG from the set of T2D-associated gene markers, we developed an in-house software that comprises three steps as indicated below:

Step 1: The original set of T2D-associated gene markers was taken as initial sub-clusters of genes. It should be noted that in the establishment of the gene profile we had constructed gene linkage groups to reduce the dimensionality of the statistical analysis. Accordingly, all genes from a gene linkage group were considered as one sub-cluster.

Step 2: We applied the Chameleon algorithm to combine the sub-clusters exhibiting a minimal similarity of 0.4 using dynamic modeling technology and basing selection on both interconnectivity and closeness[53]. The similarity here is defined by the product of interconnectivity and closeness (we used this definition in the whole analysis of MLG identification). We term these clusters semi-clusters.

Step 3: To further merge the semi-clusters established in step 2. In this step, we first updated the similarity between any two semi-clusters, and then performed a taxonomic assignment for each semi-cluster (see the method below). Finally, two or more semi-clusters would be merged into a MLG if they satisfied both of the following two requirements: a) the similarity values between the semi-clusters were > 0.2; b) all these semi-clusters were assigned from the same taxonomy lineage.

Taxonomic assignment for MLGs. All genes from one MLG were aligned to the reference microbial genomes (IMG database, v3.4) at the nucleotide level (by BLASTN) and the NCBI-nr database (Feb. 2012) at the protein level (by BLASTP). The alignment hits were filtered by both the e-value (< 1×10^{-10} at the nucleotide level and

< 1×10^{-5} at the protein level) and the alignment coverage (>70% of a query sequence). From the alignments with the reference microbial genomes, we obtained a list of well-mapped bacterial genomes for each MLG and ordered these bacterial genomes according to the proportion of genes that could be mapped onto the bacterial genome, as well as the average identity of the alignments. The taxonomic assignment of a MLG was determined by the following principles: 1) if more than 90% of genes in this MLG can be mapped onto a reference genome with a threshold of 95% identity at the nucleotide level, we considered this particular MLG to originate from this known bacterial species; 2) if more than 80% of genes in this MLG can be mapped onto a reference genome with a threshold of 85% identity at the both nucleotide and protein levels, we considered this MLG to originate from the same genus of the matched bacterial species; 3) if the 16S rDNA sequences can be identified from the assembly result of a MLG, we performed the phylogenetic analysis by RDP-classifier[54] (bootstrap value > 0.80) and then defined the taxonomic assignment for the MLG if the phylotype from 16S sequences was consistent with that from genes.

Advanced-assembly for MLGs. To reconstruct the potential bacterial genomes, we designed an additional process of advanced-assembly for each MLG, which was implemented in four steps.

Step 1: Taking the genes from a MLG as a seed, we identified samples that contain the seed with the highest abundance among all samples, and then selected the paired-end reads from these samples that could be mapped onto the seed (including the paired-end read that only one end could be mapped). The lower limit of the coverage of these paired-end reads is 50× in no more than 5 samples, which is computed by dividing the total size of selected reads by the total length of the seed.

Step 2: A de novo assembly was performed on the selected reads in step 1 by using the SOAPdenovo with the same parameters used for the construction of the gene

catalogue.

Step 3: To identify and remove the mis-assembled contigs probably caused by contaminated reads, we applied a composition-based binning method. Contigs whose GC content value and sequencing depth value were distinct from the other contigs of the assembly result were removed, as they might be wrongly assembled due to various reasons.

Step 4: Taking the final assembly result from step 3 as a seed, we repeated the procedure from step 2 until that there were no further distinct improvements of the assembly (in detail, the increment of total contig size was less than 5%).

MLG-based analysis

Validation of MLG methods. The performance of our MLG identification methods was evaluated by following steps: 1) in our quantified gene result, the rarely present genes (present in <6 samples) were filtered at first; 2) based on the taxonomic assignment result in the updated gene catalogue, we identified a set of gut bacterial species by the criteria of containing 1,000~5,000 unique mapped genes, with the similarity threshold of 95%. In this step, we manually removed the redundant strains in one species and also discarded the genes that could be mapped onto more than one species. Ultimately, 130,065 genes from 50 gut bacterial species were identified as a test set for validating the MLG method; 3) the standard MLG method described above was performed on the test set. For each MLG, we computed the percentage of genes that were not from the major species as an error rate, which were showed in Supplementary Table 9.

Relative abundance estimating of MLGs. We estimated the relative abundance of a MLG in all samples by using the relative abundance values of genes from this MLG. For this MLG, we first discarded genes that were among the 5% with the highest and

lowest relative abundance, respectively, and then fitted a Poisson distribution to the rest. The estimated mean of the Poisson distribution was interpreted as the relative abundance of this MLG. At last, the profile of MLGs among all samples was obtained for the following analyses.

The co-occurrence network of MLGs. We calculated the Spearman's rank correlation coefficient between MLGs based on the profile of these MLGs. A network was then constructed by using the method implemented in Cytoscape v2.8[55]. In the network, the edges denoted the correlation between two MLGs, under the criterion that Spearman's rank correlation coefficient > 0.40 (blue line of the edge) or < -0.40 (red line of the edge). The size of nodes was proportional to the gene number of the MLG, and the color of nodes denoted the taxonomic assignment of the MLG.

Functional description of identified markers

Functional analyses based on KO markers. Functional analysis was performed mainly on KEGG Orthologue (KO) markers, which had detailed information on biological pathway and module. The percentages of KO markers belong to each KEGG category (the KEGG Class at level 2) out of total T2D-enriched or control-enriched KO markers were designated as comparison parameter. Fisher's exact test was used to calculate the significance level (Supplementary Figure 9).

We then studied the T2D-associated KO markers at the pathway or module level. In the KEGG category of membrane transport, relative module classes were checked. For example, sugar-related and branch-chain-amino-acid-related membrane transport functions were notably from the set of T2D-enriched KOs. KO markers that belong to metabolism of cofactor and vitamins were further checked one by one in KEGG map to identify the type of reaction, such as the biosynthesis, degradation or reversible reaction.

To validate the relationship between cell motility related KO markers and butyrate-producing bacteria, the Spearman's rank correlation coefficient was calculated between the profile of these KO markers and the genera. Only the relationships with the correlation coefficient above 0.5 or below -0.5 were showed in Supplementary Table 11a. In addition, we checked the taxonomic composition of genes that were annotated to these cell-motility-related KOs. At the genus level, we listed the top 3 mostly assigned genera for each KO, which was showed in Supplementary Table 11b. The method of taxonomic assignment was introduced previously.

The functions that were not described in KEGG pathway or modules were checked manually. In detail, the drug resistance related KO markers were screened by key words screening, like penicillin, macrolide, multidrug, streptomycin, chloramphenicol and lactamase et al. Oxidative stress resistance related KO markers were also screened by key words of catalase, nitric oxide reductase, glutathione reductase, peroxidase, peroxiredoxin.

With regard to some special functions indicated by our MLGs, for example, the butyrate production and sulfate reduction, we also searched the homologue genes in our gene catalogue corresponding to such functions, and then took the whole of these homologue genes as a functional group. Please see the example of butyrate-CoA gene identification in the next paragraph.

Identification, phylogenetic and taxonomic analyses of butyrate-CoA genes. Using amino acid sequence of butyryl-CoA:acetate CoA-transferase from *Roseburia hominis A2-183* as reference, we found 37 genes in the updated gene catalogue that covered > 70% of the length of the reference sequence and were above the similarity threshold >70% using BLASTP. The taxonomic assignment of these genes had been done in the gene catalogue. Multiple sequence alignment of these 37 genes were performed at the amino acid level and the aligned amino acid sequences were then

translated back to nucleic acid sequence for phylogenetic tree construction (neighbor-joining method) where the 4-hydroxybutyrate CoA transferase gene from *Anaerostipes caccae L1-92* was chosen as the out-group. Differences of the relative abundance of the butyryl-CoA: acetate CoA-transferase genes between T2D patients and healthy individuals were tested using Wilcoxon rank-sum test in all samples (Supplementary Figure 10).

T2D classification by gut microbial markers

Maximum Relevance Minimum Redundancy (mRMR) feature selection framework. To establish a T2D classification by gut metagenomic markers, we adopted an mRMR method to perform a feature selection[56]. We used the "sideChannelAttack" package of the R software to perform the incremental search and found 344 sequential markers sets. For each sequential set, we estimated the error rate by a leave-one-out cross-validation (LOOCV) of linear discrimination classifier. The optimal selection of marker sets was the one corresponding to the lowest error rate. In the present study, we made the feature selection on a set of 52,484 T2D-associated gene markers. We finally selected a set of 50 gut microbial gene markers as the optimal selection for T2D classification.

Receiver Operator Characteristic (ROC) analysis. We applied the ROC analysis to assess the performance of the T2D classification based on metagenomic markers. Using on the 50 gut metagenomic markers selected by mRMR method, the support vector machine (SVM) classifier (realized by the "e1071" package of R software) with leave-pair-out cross-validation (LPOCV) advocated for analysis of small-sample biological datasets[57], was used to generate ROC curve. The same method was also applied on the clinical datasets. By using the "pROC" package of R software[58], we then computed the 95% confidence interval (CI) of the AUC with 10,000 bootstrap

replicates to assess the variability of the measure.

Definition of T2D index. To evaluate the effect of the gut metagenome on T2D, we defined and computed the T2D index for each individual on the basis of the selected 50 gut metagenomic markers by mRMR method. For each individual sample, the T2D index of sample j that denoted by I_j was computed by the formula below:

$$I_j^d = \sum_{i \in N} A_{ij}$$

$$I_j^n = \sum_{i \in M} A_{ij}$$

$$I_j = \left(\frac{I_j^d}{|N|} - \frac{I_j^n}{|M|} \right) \times 10^6$$

Where A_{ij} is the relative abundance of marker i in sample j. N is a subset of all T2D-enriched markers in these 50 selected gut metagenomic markers. M is a subset of all control-enriched markers in these 50 selected gut metagenomic markers. And $|N|$ and $|M|$ are the sizes of these two sets.

References

38 Alberti, K. G. & Zimmet, P. Z. Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus provisional report of a WHO consultation. *Diabet Med* **15**, 539-553, doi:10.1002/(SICI)1096-9136(199807)15:7<539::AID-DIA668>3.0.CO;2-S (1998).

39 Manichanh, C. *et al.* Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. *Gut* **55**, 205-211, doi:gut.2005.073817 [pii]

10.1136/gut.2005.073817 (2006).

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

10.1038/nature08821 (2010).

41 Li, R. & Zhu, H. De novo assembly of the human genomes with massively parallel short read sequencing. *Genome Res* (2009).

42 Zhu, W., Lomsadze, A. & Borodovsky, M. Ab initio gene identification in metagenomic sequences. *Nucleic acids research* **38**, e132, doi:10.1093/nar/gkq275 (2010).

43 Arumugam, M. *et al.* Enterotypes of the human gut microbiome. *Nature* **473**, 174-180, doi:10.1038/nature09944 (2011).

44 Dongen, v. Graph Clustering by Flow Simulation. *PhD thesis* (2000).

45 Audic, S. & Claverie, J. M. The significance of digital gene expression profiles. *Genome Res* **7**, 986-995 (1997).

46 Dray, S. & Dufour, A. The ade4 Package: Implementing the Duality Diagram for Ecologists. *Journal of Statistical Software* **22**, 1-20, doi:citeulike-article-id:3973170 (2007).

47 McArdle, B. H. & Anderson, M. J. Fitting Multivariate Models to Community Data: A Comment on Distance-Based Redundancy Analysis. *Ecology* **82**, 290-297 (2001).

48 Zapala, M. A. & Schork, N. J. Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 19430-19435, doi:10.1073/pnas.0609333103 (2006).

49 Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics* **38**, 904-909, doi:10.1038/ng1847 (2006).

50 Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS genetics* **2**, e190, doi:10.1371/journal.pgen.0020190 (2006).

51 Storey, J. D. A direct approach to false discovery rates. *Journal of the Royal Statistical Society - Series B: Statistical Methodology* **64**, 479-498 (2002).

52 Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 9440-9445, doi:10.1073/pnas.1530509100 (2003).

53 Karypis, G. & Kumar, V. Chameleon: hierarchical clustering using dynamic modeling. *Computer* **32**, 68-75 (1999).

54 Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* **73**, 5261-5267, doi:AEM.00062-07 [pii]

10.1128/AEM.00062-07 (2007).

55 Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-2504, doi:10.1101/gr.1239303

13/11/2498 [pii] (2003).

56 Peng, H., Long, F. & Ding, C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans Pattern Anal Mach Intell* **27**, 1226-1238, doi:10.1109/TPAMI.2005.159 (2005).

57 Airola, A., Pahikkala, T., Waegeman, W., Baets, B. D. & Salakoski, T. A comparison of AUC estimators in small-sample studies. (2010).

58 Robin, X. *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77, doi:10.1186/1471-2105-12-77 (2011).

Supplementary Figures

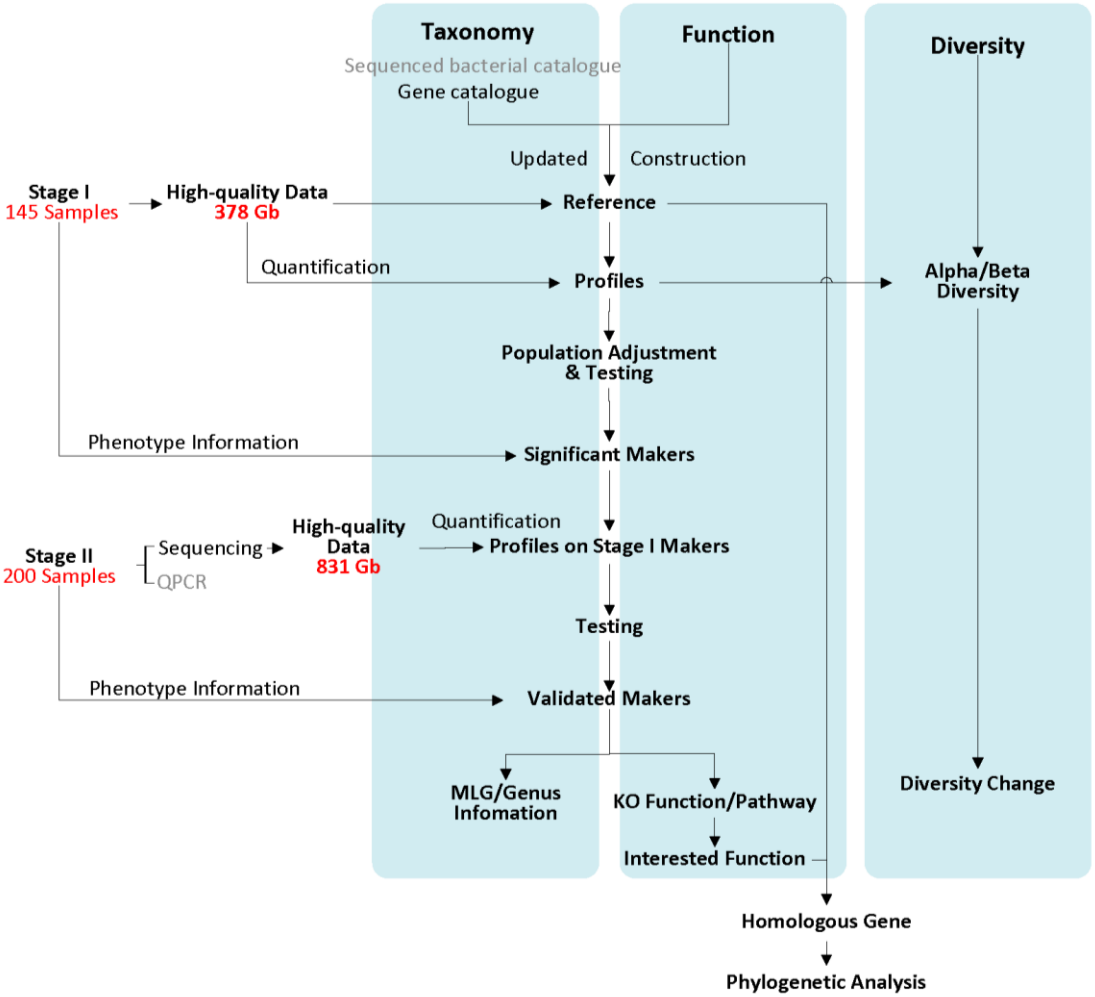


Figure S1| the overall strategy of MGWAS used in our study.

The text with grey colors indicated some alternative choices, but not used in this study.

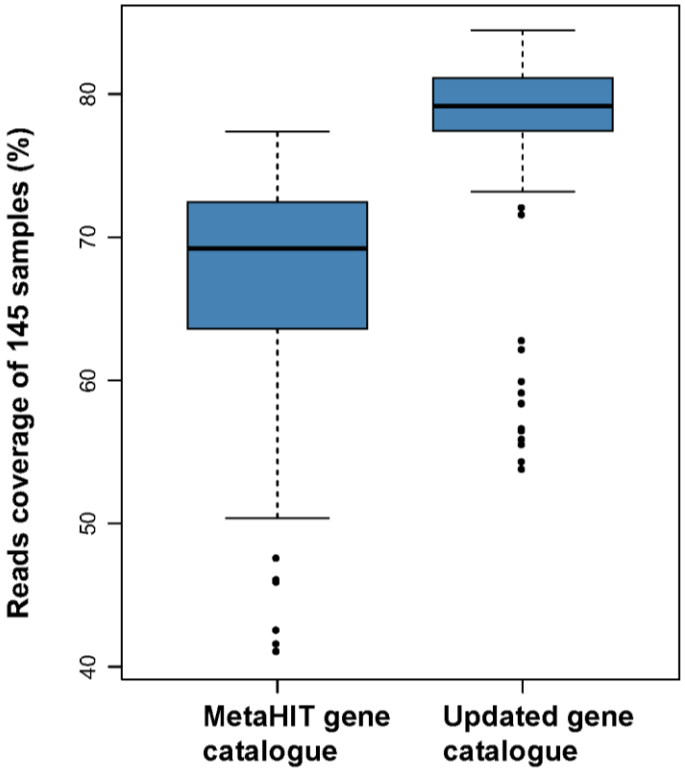


Figure S2| the coverage of sequencing reads in the MetaHIT gene catalogue and the updated gene catalogue.

The high-quality sequencing reads of 145 Chinese samples was mapped onto the MetaHIT gene catalogue (3.3 million genes) and the updated gene catalogue in this study (4.3 million genes).

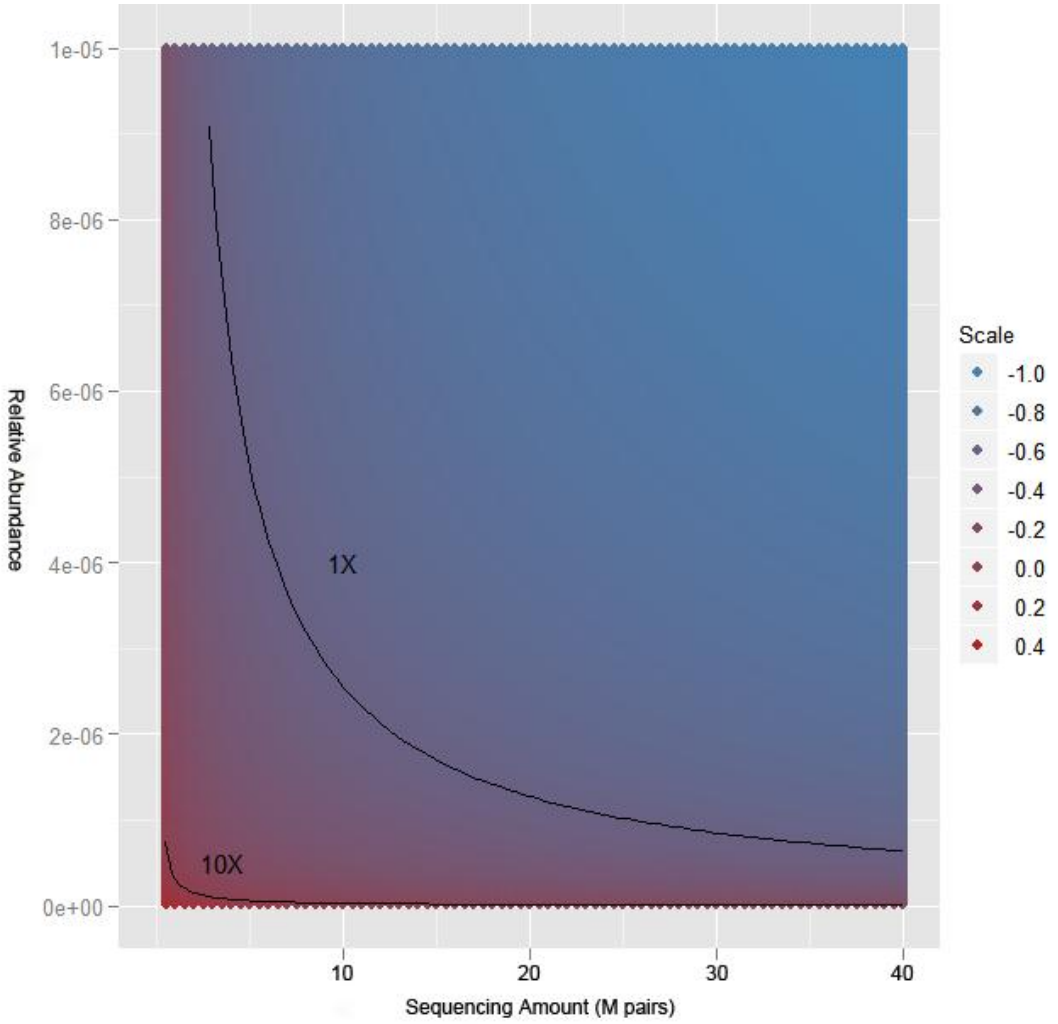


Figure S3| Detection error rate distribution of relative abundance profiles in different sequencing amount.

The X axis represents the sequencing amount of a sample, which was defined as the number of paired-end reads, and the Y axis represents the relative abundance of a gene. The 99% confidence interval (CI) of the relative abundance was estimated and the detection error rate was defined as the ratio of the interval width to the relative abundance itself. The scaled detection error rate, transformed by $\log_{10}(\log_{10}(1+x))$, was used to color all the points, with warmer color representing larger detection error rate. Two indifference curves were added: detection error rate that fall to the upper right of the curves would be less than 1 and 10, respectively.

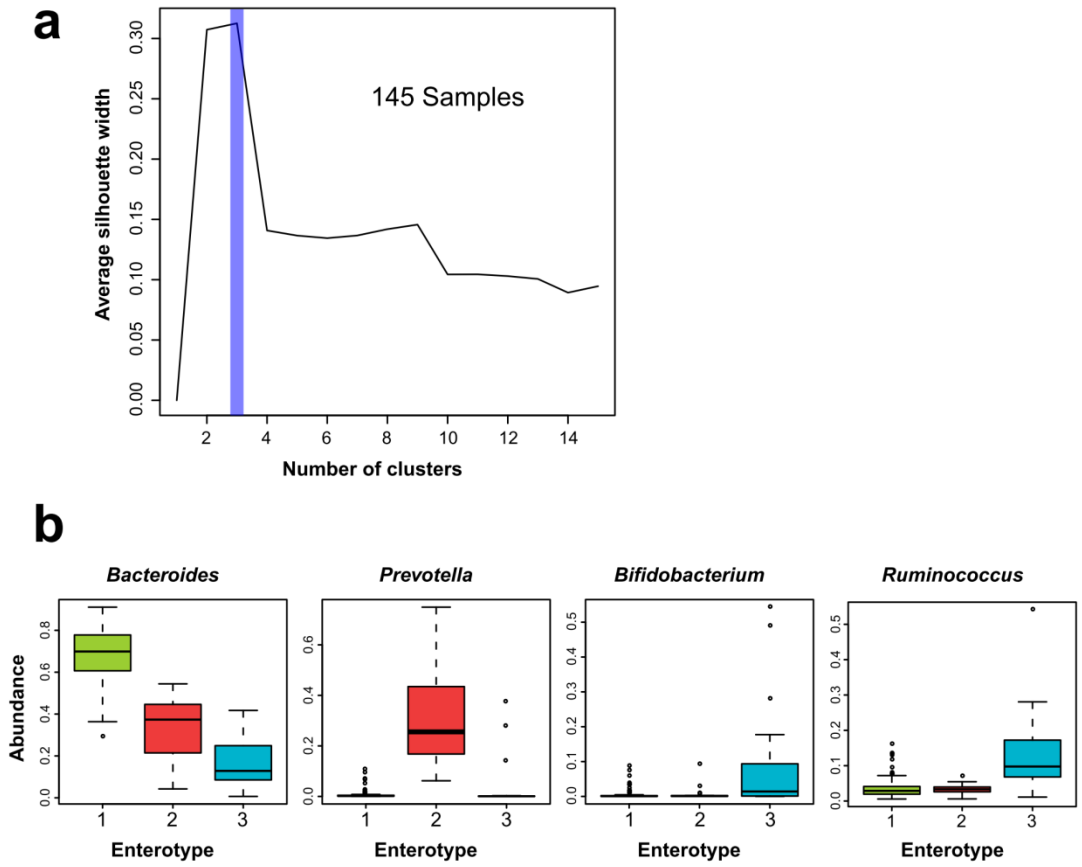
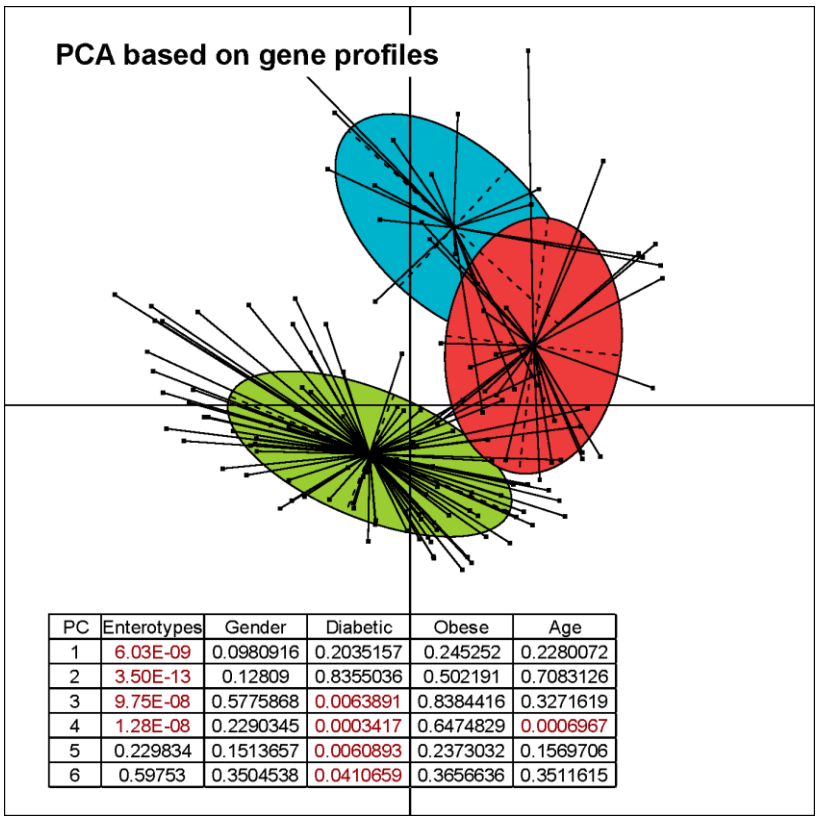


Figure S4| Enterotypes of the gut microbiome of Chinese population.

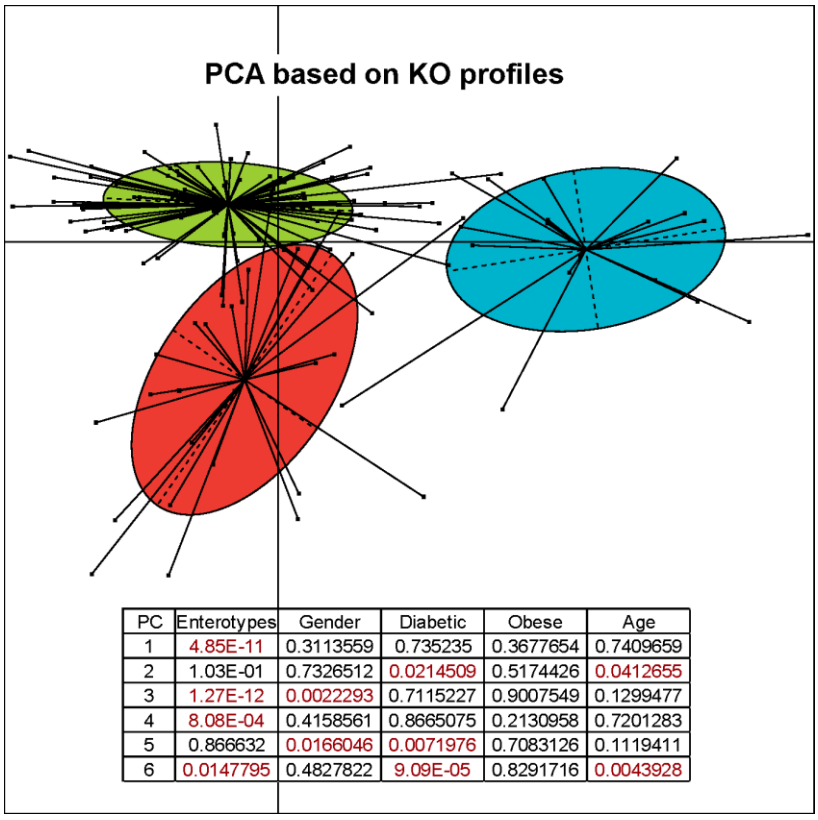
(a), average silhouette width was used to determinate the optimal number of clusters. **(b)**, abundance of the main contributed genera of each enterotypes.

a.



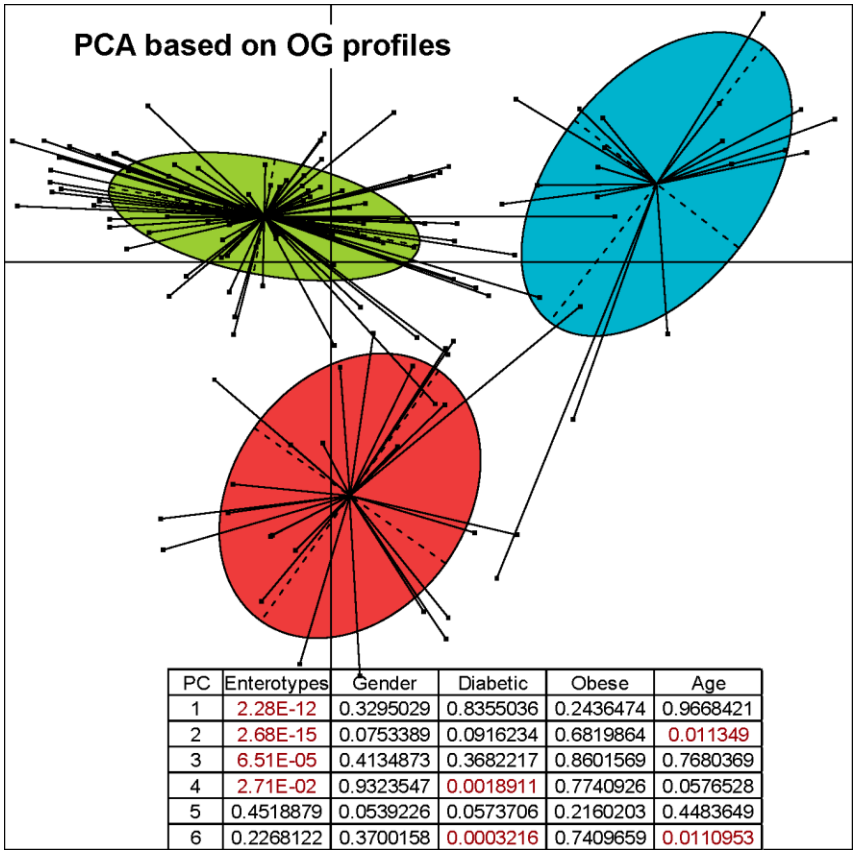
PC	Enterotypes	Gender	Diabetic	Obese	Age
1	6.03E-09	0.0980916	0.2035157	0.245252	0.2280072
2	3.50E-13	0.12809	0.8355036	0.502191	0.7083126
3	9.75E-08	0.5775868	0.0063891	0.8384416	0.3271619
4	1.28E-08	0.2290345	0.0003417	0.6474829	0.0006967
5	0.229834	0.1513657	0.0060893	0.2373032	0.1569706
6	0.59753	0.3504538	0.0410659	0.3656636	0.3511615

b.



PC	Enterotypes	Gender	Diabetic	Obese	Age
1	4.85E-11	0.3113559	0.735235	0.3677654	0.7409659
2	1.03E-01	0.7326512	0.0214509	0.5174426	0.0412655
3	1.27E-12	0.0022293	0.7115227	0.9007549	0.1299477
4	8.08E-04	0.4158561	0.8665075	0.2130958	0.7201283
5	0.866632	0.0166046	0.0071976	0.7083126	0.1119411
6	0.0147795	0.4827822	9.09E-05	0.8291716	0.0043928

c.



PC	Enterotypes	Gender	Diabetic	Obese	Age
1	2.28E-12	0.3295029	0.8355036	0.2436474	0.9668421
2	2.68E-15	0.0753389	0.0916234	0.6819864	0.011349
3	6.51E-05	0.4134873	0.3682217	0.8601569	0.7680369
4	2.71E-02	0.9323547	0.0018911	0.7740926	0.0576528
5	0.4518879	0.0539226	0.0573706	0.2160203	0.4483649
6	0.2268122	0.3700158	0.0003216	0.7409659	0.0110953

d.

PC	Enterotypes	Gender	Diabetic	Obese	Age
1	2.38E-20	0.033832	0.960569	0.166462	0.275406
2	5.10E-17	0.583258	0.006623	0.807636	0.010483
3	0.111214	0.938941	0.125364	0.415892	0.238878
4	0.512009	0.569129	0.535923	0.323268	0.624889
5	0.144566	0.216468	0.000168	0.594392	0.005285
6	0.958097	0.394827	0.004891	0.801509	0.449390

Figure S5| PCA results of gene, KO and OG profiles.

In stage I, these PCA figures were generated in the gene profile **(a)**, KO profile **(b)**, OG profile **(c)** and genus profile **(d)**, respectively. The method of enterotype classification was described in Supplementary Methods. Here, the three enterotypes were labeled and grouped in these PCA figures. In addition, the top six principle components (PCs)

were tested for correlations with some known factors, such as T2D (diabetic), BMI (obese), gender, enterotypes and age. Note that the PCA figure at the genus level is depicted as Fig. 1a in the main text.

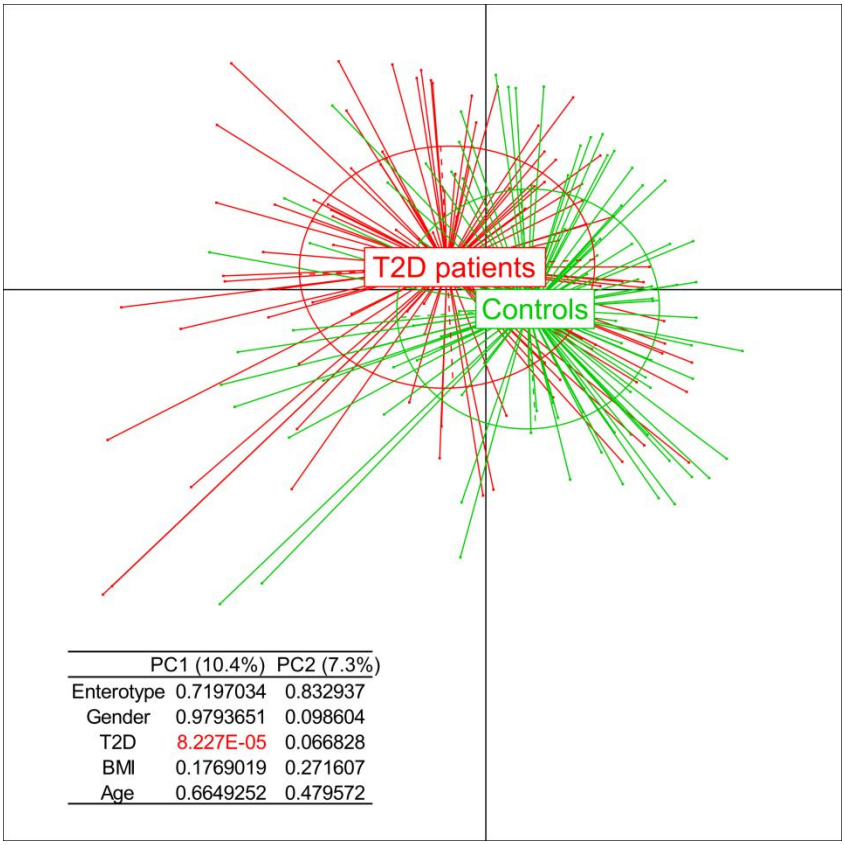


Figure S6| Validating the T2D-associated genes in an independent sample set of stage II.

278,168 gene markers that significantly associated with T2D in stage I, was quantified in stage II samples. Then, we performed a PCA analysis to see the subpopulation of these stage II samples. The first two principle components (PCs) were tested for the correlation with known factors. The T2D disease state was the primary significant factor to explain the different composition of these gut microbial genes.

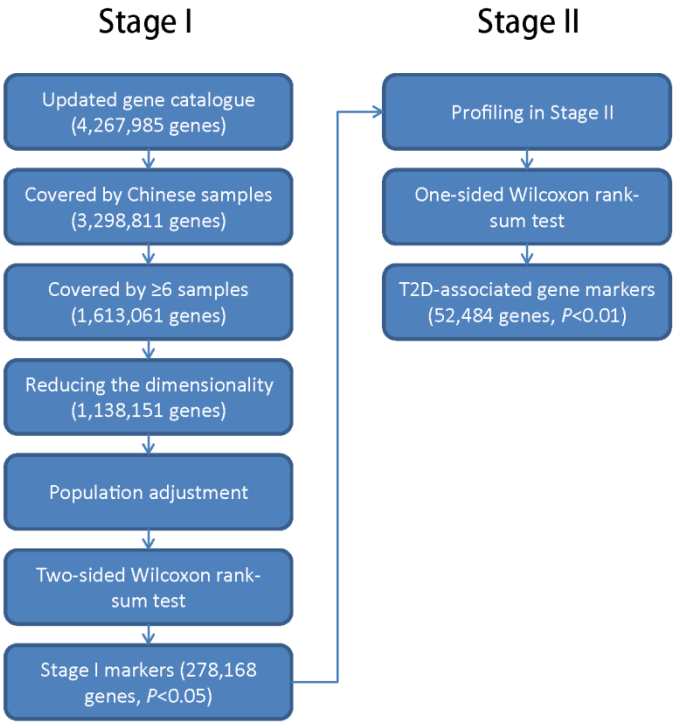
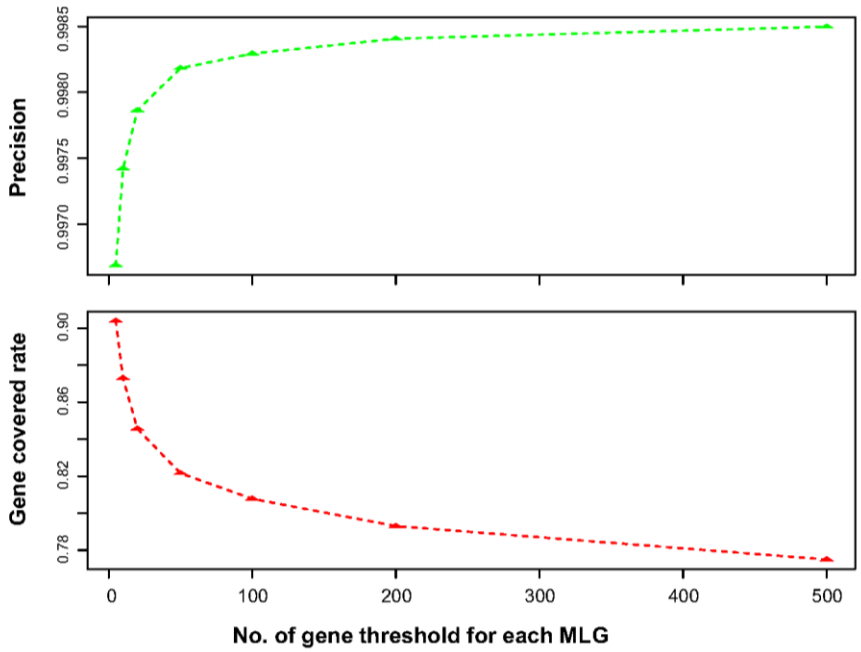


Figure S7| the detailed pipeline of statistical analysis in the gene profile in a two-stage MGWAS.

a



b

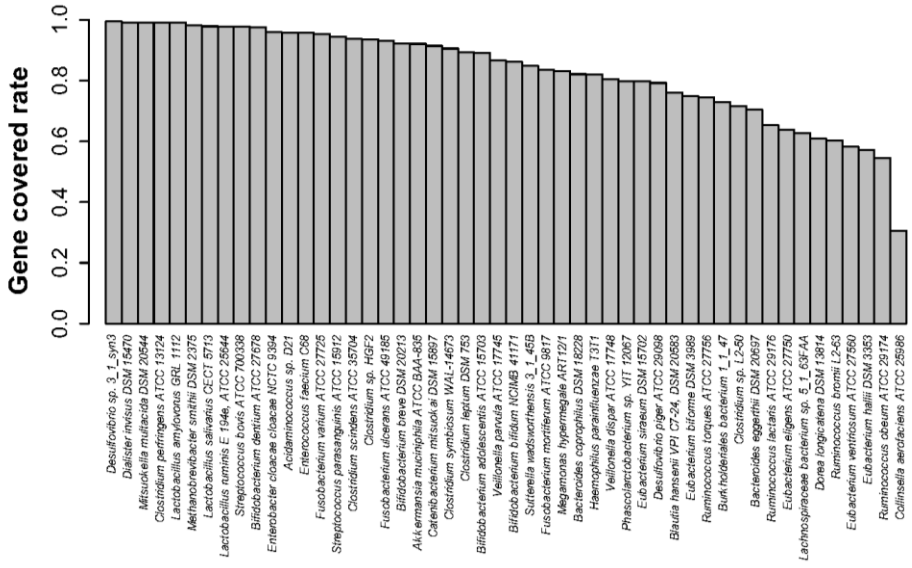


Figure S8| the validation result of our method for identifying MLGs.

To evaluate our MLG method, we customized a subset of 130,605 genes from 50 sequenced bacterial genomes. And then we compared the MLG results on this gene set and the known bacterial species information. **(a)**, at different thresholds of minimal gene number in a MLG, we computed the precision of these MLGs and the

gene coverage of original genes. **(b)**, we identified MLGs with the threshold of minimal gene number 100. This figure showed the coverage of its genome genes by our identified MLGs.

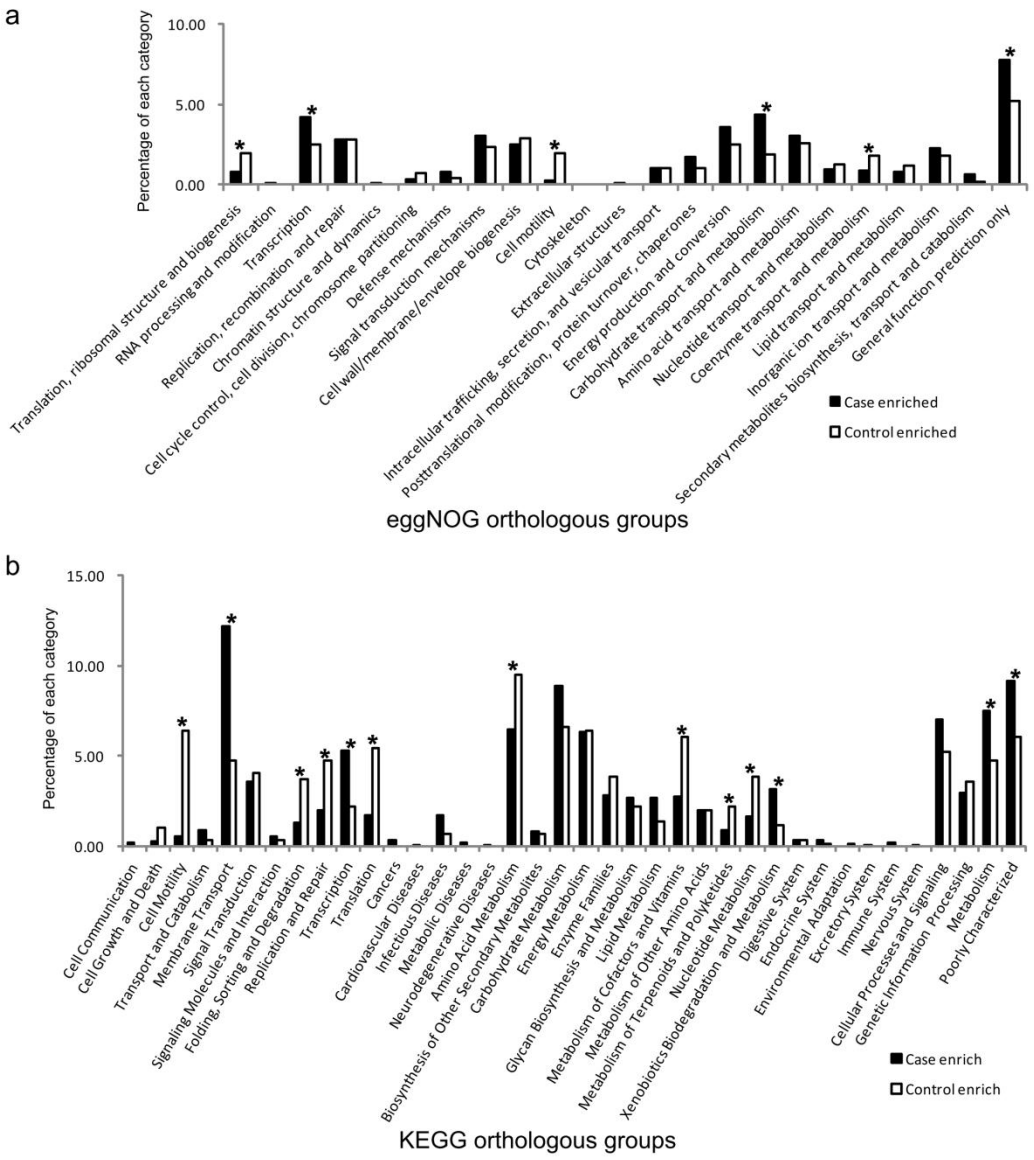
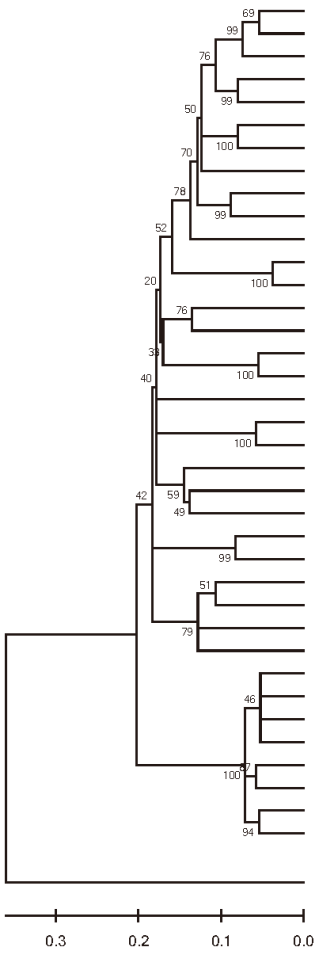


Figure S9| the distribution of functional categories for T2D-associated OG markers and KO markers.

(a), Comparison between the T2D-enriched and control-enriched OG markers on 25 OG functional categories. **(b)**, Comparison between the T2D-enriched and control -enriched KO markers on level 2 of KEGG functional category.



Gene ID	Gene Length	*P*-values (Wilcox's test, n=344)	Enrichment (1=T2D, 0=Control)	Source organism	% similarity
3863709	1,341	0.013704773	0	*Roseburia*	91.0
4163923	1,395	0.000314307	0	*Roseburia inulinivorans*	99.1
2835707	1,341	0.009106995	0	*Unclassified*	
921917	1,341	0.162896123	0	*Roseburia*	80.0
3733262	1,341	0.071250008	0	*Roseburia*	79.0
81490	1,035	0.164718274	0	*Eubacterium*	87.1
2601218	1,341	0.006035311	0	*Eubacterium rectale*	99.0
2230827	1,341	0.053000702	1	*Unclassified*	
187732	1,377	3.09283E-07	0	*Roseburia intestinalis*	98.3
3572277	1,341	0.122132664	0	*Roseburia*	82.4
478953	1,341	0.031505946	0	*Unclassified*	
1455857	1,341	0.011518047	1	*Unclassified*	
1486003	1,341	0.157236178	1	*Unclassified*	
344421	1,353	0.195708142	0	*Unclassified*	
2140427	1,341	0.496714865	0	*Unclassified*	
2216163	1,077	0.032918437	0	*Lachnospiraceae*	89.5
2843340	1,341	0.000396764	0	*Clostridium sp. SS2/1*	99.9
2255468	1,176	0.179364422	0	*Unclassified*	
2808066	1,182	0.478478632	1	*Eubacterium*	89.8
3980578	1,401	0.097585439	1	*Eubacterium hallii*	99.6
4180720	1,197	0.167587354	0	*Unclassified*	
24886	1,344	0.148139967	0	*Unclassified*	
3868351	1,221	0.004893985	0	*Unclassified*	
1796986	1,356	0.301977485	1	*Unclassified*	
3868130	1,374	0.058223323	1	*Unclassified*	
409284	1,344	0.234334752	0	*Clostridiales*	82.6
1842060	963	0.013237227	0	*Clostridiales sp. SS3/4*	94.4
3253537	1,404	0.022823984	1	*Clostridium symbiosum*	100.0
4102489	1,344	0.003417221	0	*Clostridium*	85.2
2872837	1,347	0.136246544	0	*Faecalibacterium prausnitzii*	98.1
3902729	1,347	0.001773163	0	*Faecalibacterium*	90.5
3972236	1,347	0.239400489	0	*Faecalibacterium*	91.4
3145940	1,347	0.000485572	0	*Faecalibacterium cf. prausnitzii*	99.0
2927660	1,347	0.000381794	0	*Faecalibacterium prausnitzii*	98.5
3308604	1,347	0.11947201	0	*Faecalibacterium*	88.7
2995486	1,347	0.183264357	0	*Faecalibacterium prausnitzii*	98.7
3241344	1,347	0.132473869	0	*Faecalibacterium*	90.1
				4-hydroxy-butyrate-CoA transferase from Anaerostipes caccae L1-92	

Figure S10| 37 butyryl-CoA: acetate CoA-transferase genes were identified in our updated gene catalogue.

The numbers of samples out of 344 samples (stage I & stage II) that each gene was presented in are listed. Genes occurring in less than 6 samples were excluded from statistical test and no *P*-value is given. Possible source organisms of each gene were determined based on sequence comparison to the bacterial reference genome at nucleotide level and NCBI-nr database at protein level. The ones colored in red are known species of butyrate-producing bacteria isolated from human colon. The phylogenetic tree was constructed using the neighbor-joining method, based on nucleic acid sequences that were translated back from aligned protein sequences. Bootstrap values, each expressed as a percentage of 1,000 replications, are listed at

the nodes. 4-hydroxybutyrate CoA transferase gene sequence from *Anaerostipes caccae L1-92* was chosen as the out group. The gross relative abundance of these 37 butyrate-CoA transferase genes were significantly higher in healthy controls (P = $3.2x10^{-6}$, Wilcoxon rank-sum test).

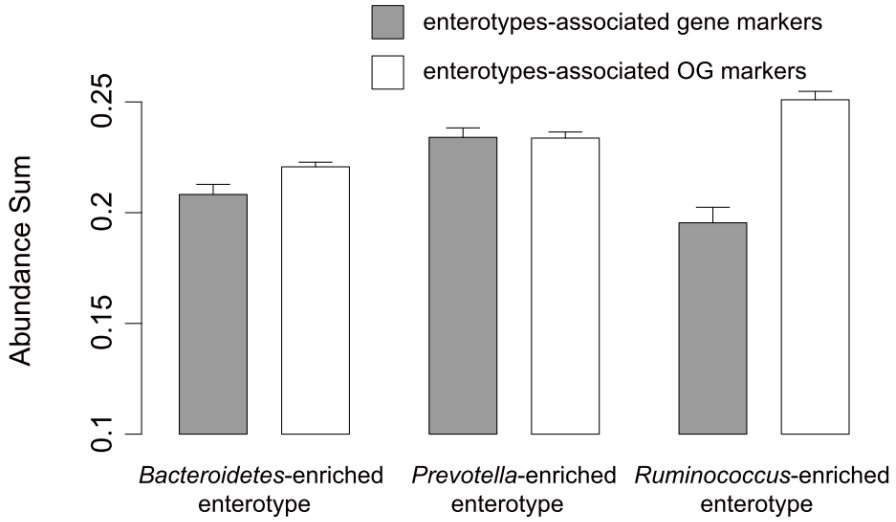


Figure S11| the gross relative abundance of the enterotypes-associated markers.

The enterotypes of each sample was determined by clustering all 344 samples (Supplementary Table 2, see Supplementary Methods for enterotypes identification). To identify enterotypes-associated marker, we performed the two-stage MGWAS analysis to the samples (145 samples in stage I and 199 in stage II) using the same methods and parameters as T2D, which identifing 117,209 gene markers (stage II P<0.01, 1.3% FDR) and 8,676 OG markers (stage II P<0.05, 5.7% FDR).

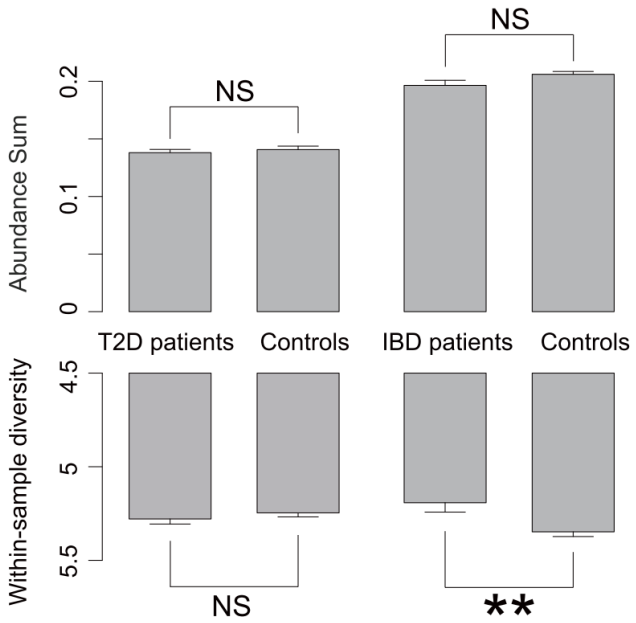
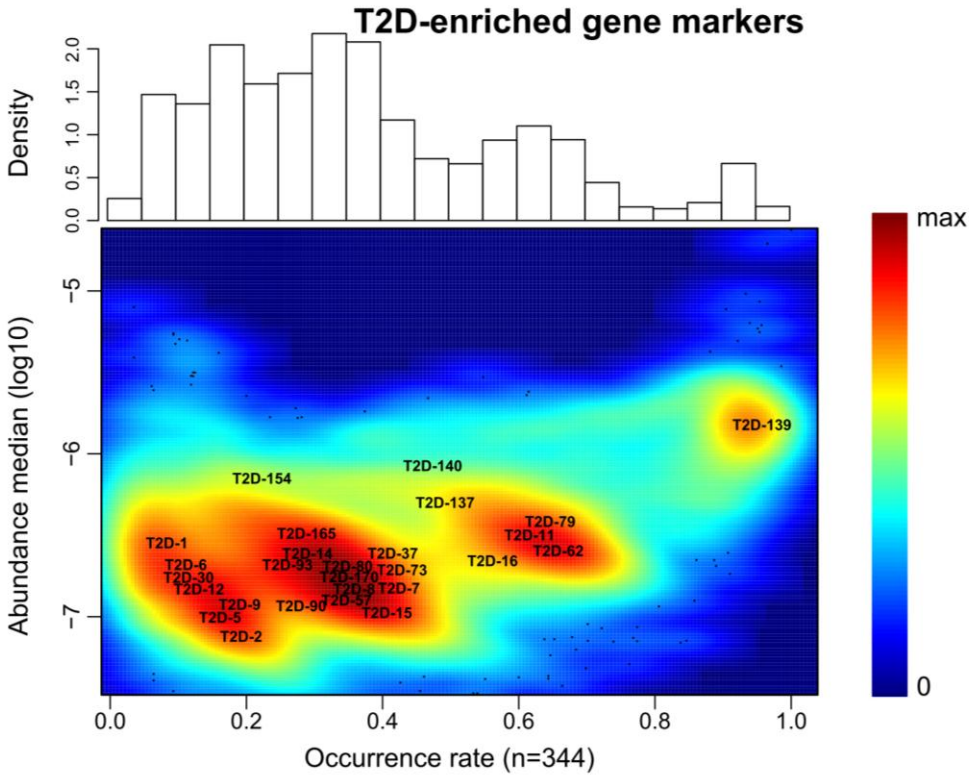


Figure S12| an ecological comparison between T2D/IBD patients and controls.

This figure showed an ecological comparison between T2D patients and controls (170 vs. 174 samples), as well as the MetaHIT IBD patients (n=25) and controls (n=99), based the OG profile. The upward bars denoted the gross relative abundance of the T2D/IBD-associated OG markers for each sample. The downward bars denoted the within-sample diversity (the Shannon Index) in each group. The statistical significance was computed by Student's t-test (** P<0.01).

a.


T2D-enriched gene markers
Density
0.0
0.5
1.0
1.5
2.0
Abundance median (log10)
−5
−6
−7
max
0
T2D-139
T2D-154
T2D-140
T2D-137
T2D-79
T2D-11
T2D-62
T2D-16
T2D-165
T2D-1
T2D-14
T2D-37
T2D-93
T2D-80
T2D-73
T2D-6
T2D-30
T2D-170
T2D-12
T2D-8
T2D-7
T2D-9
T2D-90
T2D-57
T2D-15
T2D-5
T2D-2
0.0
0.2
0.4
0.6
0.8
1.0
Occurrence rate (n=344)

b.

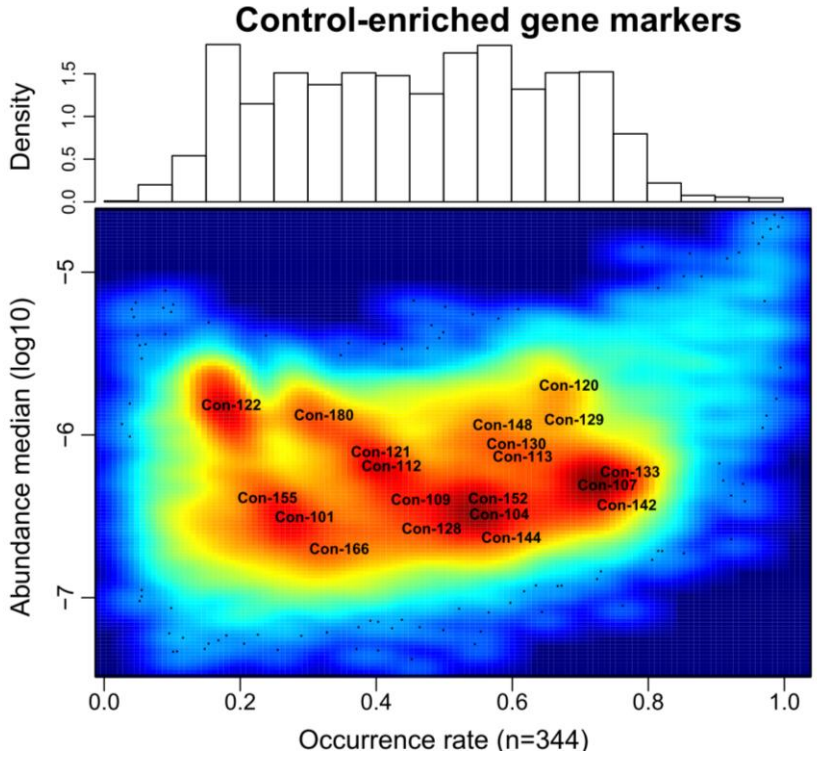

Control-enriched gene markers
Density
0.0
0.5
1.0
1.5
Abundance median (log10)
−5
−6
−7
Con-120
Con-122
Con-180
Con-148
Con-129
Con-130
Con-121
Con-113
Con-112
Con-133
Con-107
Con-155
Con-109
Con-152
Con-142
Con-104
Con-101
Con-128
Con-144
Con-166
0.0
0.2
0.4
0.6
0.8
1.0
Occurrence rate (n=344)

Figure S13| 2-Dimensional histogram plotted of T2D-associated gene markers.

The T2D-associated gene markers were divided into two classes: control-enriched marker and T2D-enriched marker. For each class of gene markers, we computed the occurrence rate and the median relative abundance of each gene and perform a 2-Dimensional histogram to show the distribution of these genes. From this figure, we could see that the control-enriched gene markers were mostly present in high occurrence rate and high relative abundance. In contrast, the T2D-enriched gene markers were quite diverse and most of them are present in low occurrence rate.

Note: Since the genes from the same MLG were linked together by similar abundance in different samples, we labeled the MLGs on this figure to show the abundance and occurrence rate of them.

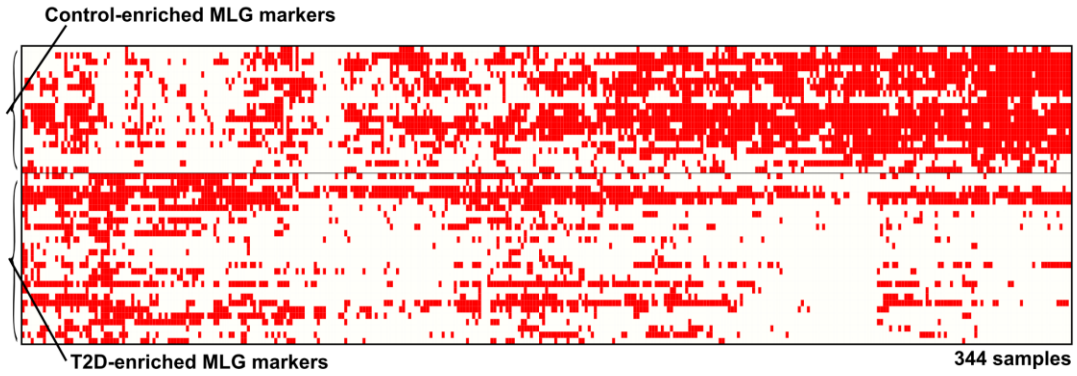


Figure S14| Presence of the T2D-associated MLGs markers in all samples.

The control-enriched markers had a higher occurrence rate than the T2D-enriched markers.

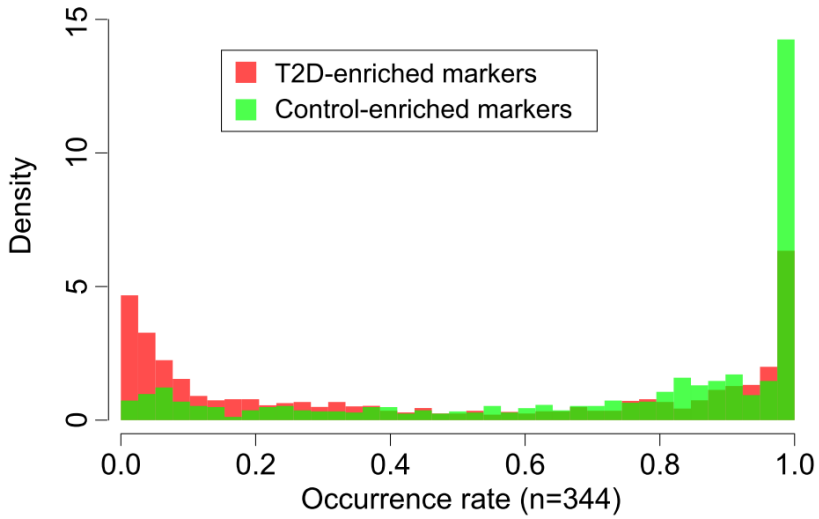


Figure S15| Occurrence rate distribution of T2D-associated OG markers.

This density histogram showed a comparison of the occurrence rate distribution between T2D-enriched OG markers and control-enriched OG markers in all samples.

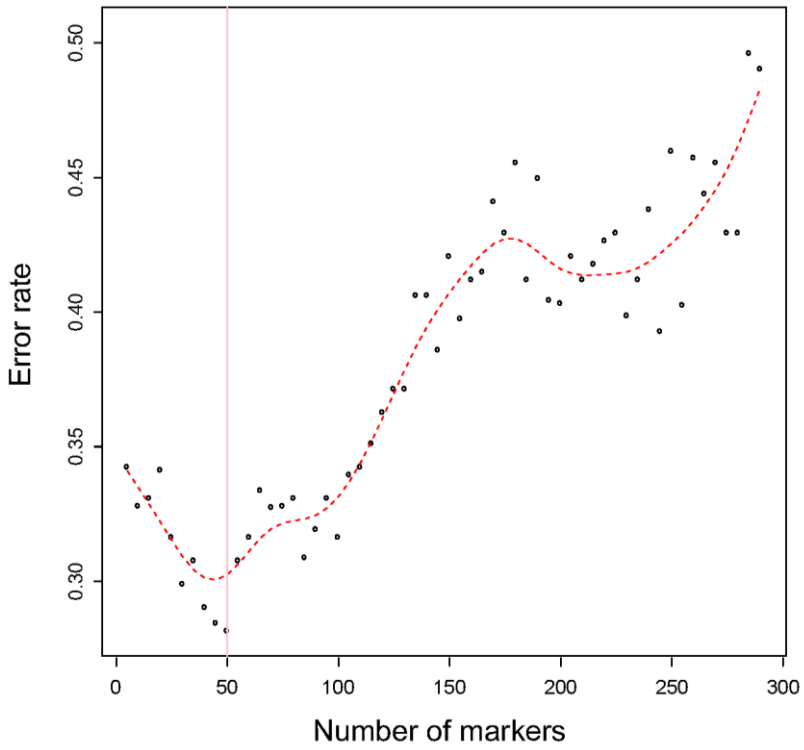


Figure S16| Estimating the optimum number of markers.

We performed incremental search in T2D-associated gene markers by the minimum redundancy maximum relevance (mRMR) methods (see Supplementary Methods for detail), and generated sequential number of subsets. For each subset, the error rate was then estimated by a leave-one-out cross-validation (LOOCV) of a linear discrimination classifier. The optimum (lowest error rate) subset contains 50 gene markers.