ARTICLE

doi:10.1038/nature11450

A metagenome-wide association study of gut microbiota in type 2 diabetes

Junjie Qin[1]*, Yingrui Li[1]*, Zhiming Cai[2]*, Shenghui Li[1]*, Jianfeng Zhu[1]*, Fan Zhang[3]*, Suisha Liang[1], Wenwei Zhang[1], Yuanlin Guan[1], Dongqian Shen[1], Yangqing Peng[1], Dongya Zhang[1], Zhuye Jie[1], Wenxian Wu[1], Youwen Qin[1], Wenbin Xue[1], Junhua Li[1], Lingchuan Han[3], Donghui Lu[3], Peixian Wu[3], Yali Dai[3], Xiaojuan Sun[2], Zesong Li[2], Aifa Tang[2], Shilong Zhong[4], Xiaoping Li[1], Weineng Chen[1], Ran Xu[1], Mingbang Wang[1], Qiang Feng[1], Meihua Gong[1], Jing Yu[1], Yanyan Zhang[1], Ming Zhang[1], Torben Hansen[5], Gaston Sanchez[6], Jeroen Raes[7,8], Gwen Falony[7,8], Shujiro Okuda[7,8], Mathieu Almeida[9], Emmanuelle LeChatelier[9], Pierre Renault[9], Nicolas Pons[9], Jean-Michel Batto[9], Zhaoxi Zhang[1], Hua Chen[1], Ruifu Yang[1,10], Weimou Zheng[1], Songgang Li[1], Huanming Yang[1], Jian Wang[1], S. Dusko Ehrlich[9], Rasmus Nielsen[6], Oluf Pedersen[5,11,12], Karsten Kristiansen[1,13] & Jun Wang[1,5,13]

Assessment and characterization of gut microbiota has become a major research area in human disease, including type 2 diabetes, the most prevalent endocrine disease worldwide. To carry out analysis on gut microbial content in patients with type 2 diabetes, we developed a protocol for a metagenome-wide association study (MGWAS) and undertook a two-stage MGWAS based on deep shotgun sequencing of the gut microbial DNA from 345 Chinese individuals. We identified and validated approximately 60,000 type-2-diabetes-associated markers and established the concept of a metagenomic linkage group, enabling taxonomic species-level analyses. MGWAS analysis showed that patients with type 2 diabetes were characterized by a moderate degree of gut microbial dysbiosis, a decrease in the abundance of some universal butyrate-producing bacteria and an increase in various opportunistic pathogens, as well as an enrichment of other microbial functions conferring sulphate reduction and oxidative stress resistance. An analysis of 23 additional individuals demonstrated that these gut microbial markers might be useful for classifying type 2 diabetes.

Type 2 diabetes (T2D), which is a complex disorder influenced by both genetic and environmental components, has become a major public health issue throughout the world[1,2]. Currently, research to parse the underlying genetic contributors to T2D is mainly through the use of genome-wide association studies (GWAS) focusing on identifying genetic components in the organism's genome[3,4]. Recently, research has indicated that the risk of developing T2D may also involve factors from the 'other genome', that is, the 'intestinal microbiome' (also termed the gut metagenome)[5].

Previous metagenomic research on the gut metagenome, primarily using 16S ribosomal RNA[6] and whole-genome shotgun (WGS) sequencing[7], has provided an overall picture of commensal microbial communities and their functional repertoire. For example, a catalogue of 3.3 million human gut microbial genes were established in 2010 (ref. 8) and, of note, a more extensive catalogue of gut microorganisms and their genes were published later[9,10]. Recent research on the gut metagenome has changed our understanding of human disease and its potential medical impact as many studies have reported. From the perspective of both taxonomic and functional composition, the gut microbiota might be linked to and contribute to many complex diseases[11]. For example, several studies have indicated that obesity is associated with an increase in the phylum Firmicutes and a relatively lower abundance of the phylum Bacteroidetes[7,12–16]. Crohn's disease research has revealed that patients had a significant reduction in the overall diversity of the gut microbiota[17] and had changes in microbial composition[18], and a T2D study showed that the proportion of the phylum Firmicutes and the class Clostridia in the gut of patients was significantly reduced[19]. However, more work is required to gain detailed information about gut microbial compositional changes and their associated impact with these types of diseases, and additional tools are required to find ways to determine associated changes easily and rapidly.

To reach these initial goals, we devised and carried out a two-stage case-control metagenome-wide association study (MGWAS) based on deep next-generation shotgun sequencing of DNA extracted from the stool samples from a total of 345 Chinese T2D patients and non-diabetic controls. From this we pinpointed specific genetic and functional components of the gut metagenome associated with T2D (Supplementary Fig. 1). Our data provide insight into the characteristics of the gut metagenome related to T2D risk, a paradigm for future studies of the pathophysiological role of the gut metagenome in other relevant disorders, and the potential usefulness for a gut–microbiota-based approach for assessment of individuals at risk of such disorders.

Construction of a gut metagenome reference

To identify metagenomic markers associated with T2D, we first developed a comprehensive metagenome reference gene set that included genetic information from Chinese individuals and T2D-specific gut microbiota, as the currently available metagenomic reference (the MetaHIT gene catalogue) did not include such data. We

[1]BGI-Shenzhen, Shenzhen 518083, China. [2]Shenzhen Second People's Hospital, The First Affiliated Hospital of Shenzhen University, Shenzhen 518035, China. [3]Peking University Shenzhen Hospital, Shenzhen 518036, China. [4]Medical Research Center of Guangdong General Hospital, Guangdong Academy of Medical Sciences, Guangzhou 510080, China. [5]The Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health Sciences, University of Copenhagen, DK-2100 Copenhagen, Denmark. [6]Department of Integrative Biology and Department of Statistics, University of California Berkeley, Berkeley, CA 94820, USA. [7]Department of Structural Biology, VIB, 1050 Brussels, Belgium. [8]Department of Applied Biological Sciences (DBIT), Vrije Universiteit Brussel, 1050 Brussels, Belgium. [9]Institut National de la Recherche Agronomique, 78350 Jouy en Josas, France. [10]State Key Laboratory of Pathogen and Biosecurity, Beijing Institute of Microbiology and Epidemiology, Beijing 100071, China. [11]Institute of Biomedical Sciences, University of Copenhagen & Faculty of Health Science, University of Aarhus, DK-8000 Aarhus, Denmark. [12]Hagedorn Research Institute, DK-2820 Gentofte, Denmark. [13]Department of Biology, University of Copenhagen, DK-2200 Copenhagen, Denmark.
*These authors contributed equally to this work.

carried out WGS sequencing on individual faecal DNA samples from 145 Chinese individuals (71 cases and 74 controls, Supplementary Table 1) and obtained an average of 2.61 gigabases (Gb) (15.8 million) paired-end reads for each, totalling 378.4 Gb of high-quality data that was free of human DNA and adaptor contaminants (Supplementary Table 2). We then performed *de novo* assembly and metagenomic gene prediction for all 145 samples. We integrated these data with the MetaHIT gene catalogue, which contained 3.3 million genes that were predicted from the gut metagenomes of individuals of European descent, and obtained an updated gene catalogue with 4,267,985 predicted genes. A total of 1,090,889 of these genes were uniquely assembled from our Chinese samples, which contributed 10.8% additional coverage of sequencing reads when comparing our data against that from the MetaHIT gene catalogue alone (Supplementary Fig. 2).

Having a more complete gene reference, we carried out taxonomic assignment and functional annotation for the updated gene catalogue using 2,890 reference genomes (IMG v3.4; Supplementary Table 3), KEGG (Release 59.0) and eggNOG databases (v3). Here, 21.3% of the genes in the updated catalogue could be robustly assigned to a genus, which covered 26.4%–90.6% (61.2% on average) of the sequencing reads in the 145 samples (Supplementary Methods); the remaining genes were likely to be from currently undefined microbial species. For assessment at a functional level, we identified 6,313 KEGG orthologues and 38,641 eggNOG orthologue groups in the updated gene catalogue, which covered 47.1% and 60.9%, respectively, of the genes in the catalogue. In addition, 14.0% of genes that were not mapped to eggNOG orthologue groups could be clustered into 7,042 novel gene families; however, these do not yet have any functional annotation information, but were still included (as in-house eggNOG orthologue groups) in our analyses. For each metagenomic sample, on average, 48.7% and 68.8% sequencing reads were covered, respectively, by these KEGG orthologues- and eggNOG orthologue groups-annotated genes.

Marker identification using a two-stage MGWAS

To define T2D-associated metagenomic markers, we devised and carried out a two-stage MGWAS strategy. Using a sequence-based profiling method, we quantified the gut microbiota in the 145 samples for use in stage I. On average, with the requirement that there should be ≥ 90% identity, we could uniquely map 77.4 ± 0.6% (mean ± s.e.m.; $n = 145$) paired-end reads to the updated gene catalogue (Supplementary Fig. 2 and Supplementary Table 2). To normalize the sequencing coverage, we used relative abundance instead of the raw read count to quantify the gut microbial genes (Supplementary Methods). With nearly 16 million sequencing reads on average per sample, our sequence-based profiling method could reliably detect very low-abundance genes. For example, given a gene with a real relative abundance of 1×10^{-6}, the detected value ranged from 0.7×10^{-6} to 1.5×10^{-6} based on a theoretical estimation (Supplementary Fig. 3). To facilitate the subsequent statistical analyses at both genetic and functional levels, we further defined and prepared three types of profiles using the quantified gene results: (1) a gene profile; (2) a KEGG orthologues profile; and (3) an eggNOG orthologue groups profile (Supplementary Methods).

We investigated the subpopulations of the 145 samples in these different profiles. Applying the same identification method as used in the MetaHIT study[20], we identified three enterotypes in our Chinese samples (Supplementary Figs 4 and 5). A principal component analysis (PCA) showed that these three enterotypes were primarily made up of several highly abundant genera, including *Bacteroides*, *Prevotella*, *Bifidobacterium* and *Ruminococcus* (Fig. 1a). However, we found no significant relationship between enterotype and T2D disease status ($P = 0.29$, Fisher's exact test). We examined the top five principal components (P value in Tracy–Widom test < 0.05 and contribution >3%): the first and second principal components were significantly correlated with enterotype ($P < 0.001$, Kruskal–Wallis test), and the fifth principal component was significantly correlated with

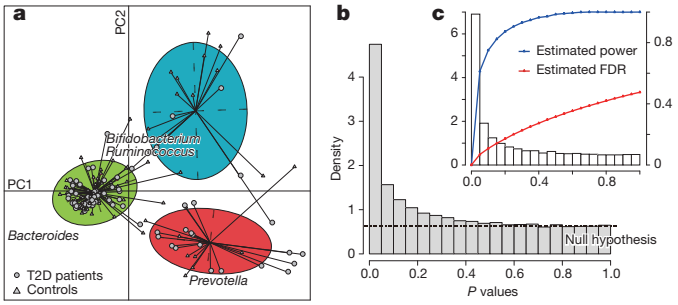

Figure 1 | Identification of T2D-associated markers from gut metagenome. **a**, The T2D patients ($n = 71$) and controls ($n = 74$) from stage I were plotted on the first two principal components of the genus profile. Lines connect individuals determined to have the same enterotype (using the PAM clustering method of refs 20,36), and coloured circles cover the individuals near the centre of gravity for each cluster ($<1.5\sigma$). The top four genera as the main contributors to these clusters were determined and plotted by their loadings in these two components. **b**, Density histogram showing the *P*-value distribution of all genes tested in stage I. The horizon line represents the distribution of *P* values under the null hypothesis. **c**, Density histogram showing the *P*-value distribution of genes in stage II, which were identified from stage I. The blue and red curves denote the estimated statistical power and false discovery rate (FDR), respectively, for a particular *P* value.

T2D ($P < 0.001$, Wilcoxon rank-sum test; Supplementary Fig. 5d), indicating that T2D, in addition to enterotype, was a determining factor in explaining the gut microbial differences in our samples. The third and fourth principal components, however, did not correlate with any known factors.

We then corrected for population stratification, which might be related to the non-T2D-related factors. For this we analysed our data using a modified EIGENSTRAT method[21]; however, unlike what is done in a GWAS subpopulation correction, we applied this analysis to microbial abundance rather than to genotype. For gene profile, after adjustment, we found that the effects that correlated with non-T2D-related factors disappeared (Supplementary Table 4). A Wilcoxon rank-sum test was done on the adjusted gene profile to identify differential metagenomic gene content between the T2D patients and controls. The outcome of our analyses showed a substantial enrichment of a set of microbial genes that had very small *P* values, as compared with the expected distribution under the null hypothesis (Fig. 1b), indicating that these genes were true T2D-associated gut microbial genes.

To validate the significant associations identified in stage I, we carried out the stage II analysis using an additional 200 Chinese individuals (one of these samples had a very low within-sample diversity, which was probably owing to the presence of a high fraction of *Escherichia* and *Klebsiella*, and was therefore excluded in later analyses; Supplementary Tables 1 and 2). We also used WGS sequencing in stage II and generated a total of 830.8 Gb sequence data with 23.6 million paired-end reads on average per sample. We then assessed the 278,167 stage I genes that had *P* values < 0.05 and found that the majority of these genes still correlated with T2D in these stage II study samples (Supplementary Fig. 6). We next controlled for the false discovery rate (FDR) in the stage II analysis, and defined a total of 52,484 T2D-associated gene markers from these genes corresponding to a FDR of 2.5% (stage II *P* value < 0.01; Fig. 1c, Supplementary Fig. 7 and Supplementary Table 5).

We applied the same two-stage analysis using the KEGG orthologues and eggNOG orthologue groups profiles and identified a total of 1,345 KEGG orthologues markers (stage II $P < 0.05$ and 4.5% FDR) and 5,612 eggNOG orthologue groups markers (stage II $P < 0.05$ and 6.6% FDR) that were associated with T2D (Supplementary Tables 6 and 7).

Development of a metagenomic linkage group

To reduce and structurally organize the abundant metagenomic data and to enable us to make a taxonomic description, we devised the

generalized concept of metagenomic linkage group (MLG) in lieu of a species concept for a metagenome. Here a MLG is defined as a group of genetic material in a metagenome that is probably physically linked as a unit rather than being independently distributed; this allowed us to avoid the need to completely determine the specific microbial species present in the metagenome, which is important given there are a large number of unknown organisms and that there is frequent lateral gene transfer (LGT) between bacteria. Using our gene profile, we defined and identified a MLG as a group of genes that co-exists among different individual samples and has a consistent abundance level and taxonomic assignment (Supplementary Methods).

To assess the reliability of our MLG identifying method, we first constructed a subset of bacterial genes from the updated metagenome gene catalogue ($n = 130{,}605$) that were independently derived from 50 known gut bacterial species (Supplementary Methods). We used a threshold for the minimum gene number for a MLG of 100, above which all 50 bacterial species could be identified with an average genome coverage of 83.0% and with an accuracy in the taxonomic classification of genes in the constructed subset of 99.8% (Supplementary Fig. 8 and Supplementary Table 8).

We identified 47 MLGs in the T2D-associated gene markers, which covered 84.4% of these markers (Supplementary Table 9). Of these, 17 MLGs could be assigned to known bacterial species on the basis of strong alignment sequence similarity with sequenced bacterial genomes at the nucleotide level (Table 1). Using the taxonomic characterization from these MLGs, we found that almost all of the MLGs enriched in the control samples were from various butyrate-producing bacteria, including *Clostridiales* sp. SS3/4, *Eubacterium rectale*, *Faecalibacterium prausnitzii*, *Roseburia intestinalis* and *Roseburia inulinivorans*. By contrast, most of T2D-enriched MLGs were from opportunistic pathogens, such as *Bacteroides caccae*, *Clostridium hathewayi*, *Clostridium ramosum*, *Clostridium symbiosum*, *Eggerthella lenta* and *Escherichia coli*, which have previously been reported to cause or underlie human infections such as bacteraemia and intra-abdominal infections[22–25]. Of interest, the known mucin-degrading species *Akkermansia muciniphila* and sulphate-reducing species *Desulfovibrio* sp. 3_1_syn3 were also enriched in T2D samples. The MLGs that were of unknown species origin will be of interest for isolation and analysis in future studies to obtain information on their relevant taxonomy.

A co-occurrence network on these MLGs was generated to assess potential relationships between the T2D-associated gut bacteria (Fig. 2a and Supplementary Methods). In this result, some types of butyrate-producers, from clostridial cluster XIVa and IV, showed a positive correlation with one another and were negatively correlated with a group of the T2D-enriched bacteria from *Clostridium*, which may indicate an antagonistic relationship between these different clostridial clusters. Another interesting finding was the presence of a small MLG from *Haemophilus parainfluenzae*, which is not a butyrate-producer but was significantly enriched in the control samples, even in an independent analysis comparing the coverage of its sequenced bacterial genome (the highest genome coverage in all samples was 94.5%; $P < 0.001$ between case and control groups, Student's *t*-test). In the co-occurrence network, this MLG was clearly separate from the cluster of butyrate producers, and may have an unknown antagonistic relationship with a T2D-enriched bacterium that is unknown but appears closely related to the *Subdoligranulum* genus. These data presented various patterns indicating relationships between the T2D-associated gut bacteria and suggested it may be important to determine, in a case-by-case manner, the different roles gut bacteria may have in maintaining or interacting with their environment.

Functional characterization related to T2D

Using the T2D-associated KEGG orthologues and eggNOG orthologue groups markers, we assessed the potential microbial functional roles in the gut microbiota of T2D patients. In general T2D-enriched markers were typically involved in the KEGG categories of membrane transport ($P < 0.001$, Fisher's exact test). This result is consistent with

Table 1 | The list of T2D-associated MLGs that could be assigned to previously known phylotypes

MLG ID	No. of genes	*P* values*		Odds ratios (95% CI)†	Taxonomy assignment (level)	Percentage similarity‡
		Stage I	Stage II			
T2D-enriched						
T2D-154	337	0.0014	2.54×10^{-4}	1.52 (1.05, 2.19)	*Akkermansia muciniphila*	98.2
T2D-140	148	3.97×10^{-4}	0.0029	1.50 (1.15, 1.97)	*Bacteroides intestinalis*	98.2
T2D-139	3,386	0.0013	2.11×10^{-4}	1.66 (1.26, 2.20)	*Bacteroides* sp. 20_3	99.3
T2D-11	5,113	4.16×10^{-8}	7.58×10^{-5}	5.89 (1.39, 25.0)	*Clostridium bolteae*	99.4
T2D-5	2,378	4.21×10^{-5}	1.97×10^{-6}	23.1 (2.08, 257)	*Clostridium hathewayi*	99.3
T2D-80	2,381	1.30×10^{-4}	1.41×10^{-5}	1.68 (0.97, 2.89)	*Clostridium ramosum*	99.8
T2D-57	821	4.00×10^{-7}	2.21×10^{-5}	2.62 (1.14, 6.03)	*Clostridium* sp. HGF2	99.6
T2D-15	2,492	4.74×10^{-5}	2.97×10^{-4}	1.13 (0.88, 1.44)	*Clostridium symbiosum*	99.6
T2D-1	949	6.01×10^{-4}	0.0036	1.41 (0.93, 2.13)	*Desulfovibrio* sp. 3_1_syn3	98.0
T2D-7	1,056	6.01×10^{-4}	2.80×10^{-4}	1.57 (0.95, 2.58)	*Eggerthella lenta*	99.6
T2D-137	425	6.71×10^{-7}	0.0012	1.72 (1.16, 2.57)	*Escherichia coli*	99.0
T2D-165	131	0.0096	0.0017	1.46 (1.07, 1.99)	*Alistipes (genus)*	99.5§
T2D-12	364	4.52×10^{-6}	8.04×10^{-8}	2.22 (1.12, 4.40)	*Clostridium (genus)*	91.0
T2D-8	5,272	7.08×10^{-10}	9.95×10^{-6}	1.12 (0.86, 1.45)	*Clostridium (genus)*	88.8
T2D-93	1,590	2.01 × 10-4	0.0020	1.84 (1.03, 3.29)	*Parabacteroides (genus)*	80.5§
T2D-62	2,584	7.63×10^{-6}	6.88×10^{-4}	2.41 (1.43, 4.08)	*Subdoligranulum (genus)*	98.7§
T2D-2	2,430	3.14×10^{-5}	0.0019	4.06 (1.28, 12.9)	*Lachnospiraceae (family)*	97.3§
Control-enriched						
Con-107	1,677	1.12×10^{-7}	0.0018	1.44 (1.13, 1.84)	*Clostridiales* sp. SS3/4	98.0
Con-112	232	0.0064	1.99×10^{-4}	1.51 (1.13, 2.03)	*Eubacterium rectale*	97.6
Con-129	1,440	0.0033	0.0010	1.55 (1.19, 2.00)	*Faecalibacterium prausnitzii*	98.2
Con-166	273	3.80×10^{-5}	1.94×10^{-4}	1.25 (0.93, 1.69)	*Haemophilus parainfluenzae*	94.8
Con-121	3,507	6.11×10^{-5}	4.90×10^{-6}	3.10 (1.92, 5.03)	*Roseburia intestinalis*	98.9
Con-113	345	2.85×10^{-4}	9.72×10^{-4}	1.45 (1.11, 1.89)	*Roseburia inulinivorans*	98.2
Con-120	116	1.90×10^{-4}	5.41×10^{-4}	1.55 (1.17, 2.06)	*Eubacterium (genus)*	89.0
Con-130	670	0.0134	0.0018	1.59 (1.21, 2.08)	*Faecalibacterium (genus)*	89.4
Con-131	202	8.99×10^{-4}	0.0017	1.58 (1.16, 2.15)	*Faecalibacterium (genus)*	96.9
Con-133	1,555	3.43×10^{-5}	0.0015	1.52 (1.15, 2.01)	*Erysipelotrichaceae (family)*	66.9§
Con-109	378	0.0135	1.67×10^{-4}	1.41 (1.09, 1.83)	*Clostridiales (order)*	87.0

* The stage I *P* value was calculated after adjustment for population structures, stage II *P* value was one-side.
† Calculated by logistic model.
‡ Similarity at nucleic acid level or, when marked with § at the protein level.

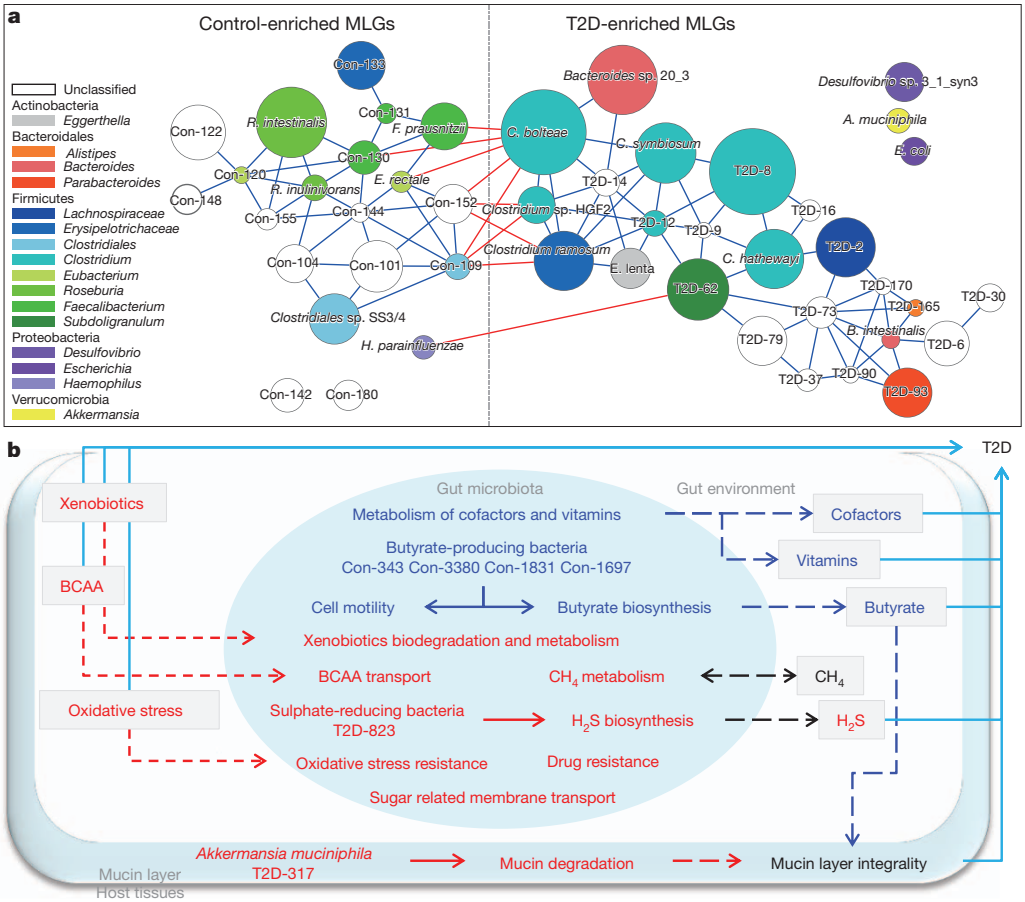

Figure 2 | Taxonomic and functional characterization of gut microbiota in T2D. **a**, A co-occurrence network was deduced from 47 MLGs that were identified from 52,484 gene markers. Nodes depict MLGs with their ID displayed in the centre. The size of the nodes indicates gene number within the MLG. The colour of the nodes indicates their taxonomic assignment. Connecting lines represent Spearman correlation coefficient values above 0.4 (blue) or below −0.4 (red). **b**, A schematic diagram showing the main functions of the gut microbes that had a predicted T2D association. Red text denotes enriched functions in T2D patients; blue text denotes depleted functions in T2D patients; black text denotes an uncertain functional role relative to T2D. The dashed line arrows point to the inference that was not detected directly but reported by previous studies.

the previous findings in studies of inflammatory bowel disease and obese patients[26]. By contrast, control-enriched markers were frequently involved in cell motility and metabolism of cofactors and vitamins ($P < 0.002$; Supplementary Fig. 9).

At the module or pathway level, the gut microbiota of T2D patients was functionally characterized with our T2D-associated markers and showed enrichment in membrane transport of sugars, branched-chain amino acid (BCAA) transport, methane metabolism, xenobiotics degradation and metabolism, and sulphate reduction. By contrast, there was a decrease in the level of bacterial chemotaxis, flagellar assembly, butyrate biosynthesis and metabolism of cofactors and vitamins (Fig. 2b and Supplementary Table 10; see Supplementary Fig. 10 for the detailed information on butyrate-CoA transferase). Some important functions, including butyrate biosynthesis and sulphate reduction, coincided with the T2D-associated bacteria identified in the MLG analysis. The butyrate-producing bacteria seemed to be the primary contributors to the cell motility functions (Supplementary Table 11), potentially indicating some functional enrichment might be related to the presence of specific species enrichment.

We found that seven of the T2D-enriched KEGG orthologues markers were related to oxidative stress resistance, including catalase (K03781), peroxiredoxin (K03386), Mn-containing catalase (K07217), glutathione reductase (NADPH) (K00383), nitric oxide reductase (K02448), putative iron-dependent peroxidase (K07223), and cytochrome *c* peroxidase (K00428), but none of the identified control-enriched KEGG orthologues markers had similar types of function. This may indicate that the gut environment of a T2D patient is one that stimulates bacterial defence mechanisms against oxidative stress (Supplementary Table 10). Similarly, we found 14 KEGG orthologues markers related to drug resistance that were greatly enriched in T2D patients, further supporting that T2D patients may have a more hostile gut environment, and the medical histories of these patients may reflect this (Supplementary Table 10).

T2D-related dysbiosis in gut microbiota

In light of the above MGWAS result and an additional PERMANOVA[27] (permutational multivariate analysis of variance) analysis that clearly showed that T2D was a significant factor for explaining the variation in the examined gut microbial samples (Supplementary Table 12), we deduced that the gut microbiota in T2D patients featured dysbiosis, which is a state where the balance of the normal microbiota has been disturbed. However, the degree of this T2D-related dysbiosis was moderate, because only $3.8 \pm 0.2\%$ (mean $\pm$ s.e.m.; $n = 344$) of the gut microbial genes (at the relative abundance level) were associated with T2D in an individual. Additionally, we did not observe a significant difference in the within-sample diversity between T2D and control groups (Fig. 3a). Specifically, the degree of gut microbiota change in T2D was not as substantial as that seen in inflammatory bowel disease (from the MetaHIT samples[8]; see Fig. 3a) or enterotypes (Supplementary Fig. 11). A similar result using the eggNOG orthologue groups profile supported the same conclusion (Supplementary Fig. 12).

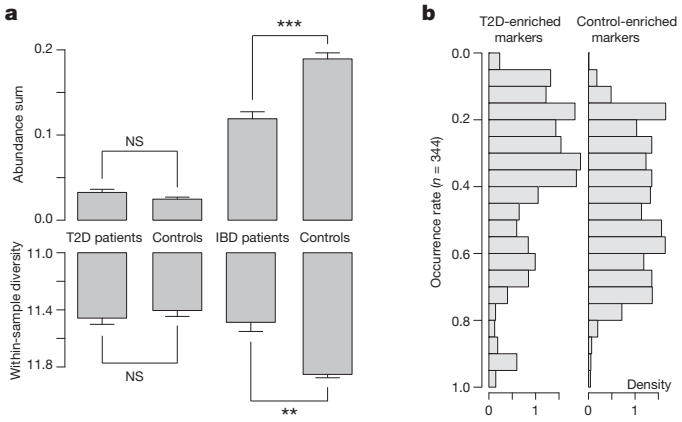

Figure 3 | Gut microbiota of T2D patients show a moderate degree of dysbiosis. **a**, An ecological comparison between T2D patients ($n = 170$) and control ($n = 174$) in all samples, as well as inflammatory bowel disease (IBD) patients ($n = 25$) and control ($n = 99$) from published MetaHIT samples[8]. The upward bars denote the gross relative abundance of the T2D-associated gene markers for each sample and the same value computed on the inflammatory-bowel-disease-associated gene markers (see Supplementary Methods). The downward bars denote the within-sample diversity (calculated using the Shannon index) in each group. For an individual sample, a lower proportion of gut microbiota was implicated in T2D disease and there was no significant difference in the within-sample diversity between the T2D patients and control as compared with the distinct difference seen in the inflammatory bowel disease analysis. $^{**}P < 0.01$; $^{***}P < 0.001$ (Student's *t*-test); NS, not significant; and the error bar denote standard error. **b**, A density histogram showing a comparison of the occurrence rate distribution between T2D-enriched gene markers and control-enriched gene markers in all samples ($n = 344$). The threshold of mapped read number for gene identification is ≥ 2.

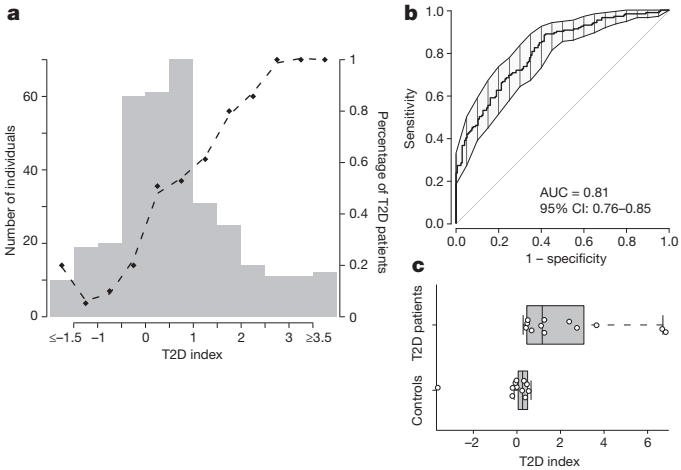

Figure 4 | A trial classification of T2D using gut microbial gene markers. **a**, A classifier to identify T2D individuals was constructed using 50 gene markers selected by mRMR, and then, for each individual, a T2D index was calculated to evaluate the risk of T2D. The histogram shows the distribution of T2D indices for all individuals, in which values less than −1.5 and values greater than 3.5 were grouped. For each bin, the black dots show the proportion of T2D patients in the population of that bin (*y* axis on the right). **b**, The area under the ROC curve (AUC) of gut-microbiota-based T2D classification. The black bars denote the 95% confidence interval (CI) and the area between the two outside curves represents the 95% CI shape. **c**, The T2D index was computed for an additional 11 Chinese T2D samples and 12 non-diabetic controls. The box depicts the interquartile range (IQR) between the first and third quartiles (25th and 75th percentiles, respectively) and the line inside denotes the median, whereas the points represent the T2D index in each sample.

To characterize ecologically the gut bacteria involved in the T2D-related dysbiosis, we compared, in all individual samples, the distribution of the occurrence rate of both T2D-associated gene and function markers, and these showed the same pattern, which was that the control-enriched markers had a higher occurrence rate on average than the T2D-enriched markers (Fig. 3b and Supplementary Figs 13–15). This may be because the beneficial bacteria lost in the T2D gut were universally present, whereas some of the harmful bacteria that appeared in the T2D gut were diverse, and thus had less overall abundance within the human population.

Gut-microbiota-based T2D classification

To exploit the potential ability of T2D classification by gut microbiota, we developed a T2D classifier system based on the 50 gene markers that we defined as an optimal gene set by a minimum redundancy–maximum relevance (mRMR) feature selection method (Supplementary Fig. 16 and Supplementary Table 13). For intuitive evaluation of the risk of T2D disease based on these 50 gut microbial gene markers, we computed a T2D index (Supplementary Methods), which correlated well with the ratio of T2D patients in our population (Fig. 4a), and the area under the receiver operating characteristic (ROC) curve was 0.81 (95% confidence interval 0.76–0.85) (Fig. 4b), indicating the gut-microbiota-based T2D index could be used to classify T2D individuals accurately.

We validated the discriminatory power of our T2D classifier using an independent study group: 11 T2D patients and 12 non-diabetic controls. In this assessment analysis, the top eight samples with the highest T2D index were all T2D patients (Fig. 4c and Supplementary Table 14); the average T2D index between case and control was significantly different ($P = 0.004$, Student's *t*-test). Overall, our cross-sectional study in overt T2D indicated that it would be worthwhile to test more extensively gut-microbiota-based classifiers in future longitudinal studies for their ability to identify subsets of the population that are at high risk for progressing to clinically defined T2D.

Discussion

T2D is a heterogeneous and multifactorial disease, influenced by a number of different genetic and environmental factors. By applying the standard two-stage GWAS strategy to design and carry out a MGWAS to identify disease-associated metagenomic markers, the present study highlights how the gut microbial composition, traditionally considered to be factors of environmental origin[12], differs between T2D patients and non-diabetic control subjects in a Chinese population.

We first established an updated human microbial gene reference set, adding information from both a new ethnicity and from T2D patients, which will be a useful resource for future metagenomic analyses. We also developed the concept of a MLG, which provided various types of taxonomic information from whole-genome shotgun data, including bacterial species-specific regions on a chromosome, and mobile genetic elements, such as plasmids and bacteriophages. Thus, a MLG can provide metagenomic species-level information even for unknown species, instead of requiring traditional taxonomic classification approaches based on sequence composition or similarity[28,29]. The use of species-level information allows assessment of the relationships between the T2D-associated bacteria. For example, we identified what appears to be an antagonistic relationship between beneficial bacteria and harmful bacteria, highlighted by the large populations of clostridial clusters. These species-level analyses also showed various patterns: for example, the MLG from *Haemophilus parainfluenzae* in the control samples could be inferred, under these circumstances, to be beneficial; however, on the basis of relationship patterns, it was quite distinct from the other inferred beneficial bacteria, indicating that *H. parainfluenzae* may have a different type of impact in this specific biological context (Fig. 2a).

Our findings indicated that T2D patients had only a moderate degree gut bacterial dysbiosis; however, functional annotation analyses indicated a decline in butyrate-producing bacteria, which may be metabolically beneficial, and an increase in several opportunistic

pathogens. Importantly, the abundance of these categories of opportunistic pathogens seemed to be quite diverse among our Chinese study participants. Such changes in the intestinal bacteria composition have recently been reported for colorectal cancer patients[30] and ageing population[31]. Thus, a general picture is emerging where butyrate-producing bacteria seem to have a protective role against several types of diseases. Additionally, our finding of a general dysbiosis in T2D patients raises the possibility that there is a 'functional dysbiosis', rather than there being a specific microbial species that has a direct association with T2D pathophysiology. Furthermore, given that other intestinal diseases show a loss of butyrate-producing bacteria with a commensurate increase in opportunistic pathogens, it is possible that dysbiosis that results in a disordered, rather than directional, alteration of gut microbial composition may itself have a role in increasing the susceptibility to a variety of diseases.

Our analysis of bacterial gene functions indicating there was an increase in functions relating to gut oxidative stress response is also of interest, given that previous studies have shown that a high oxidative stress level is related to a predisposition for diabetic complications[32]. Finally, our findings that gut metagenomic markers are able to differentiate between T2D cases and controls with a higher level of specificity than similar analyses based on human genome variation[33] raises the possibility for a mode of monitoring gut health and a complementary approach for risk assessment of this common disorder.

METHODS SUMMARY

Sample collection and DNA extraction. Faecal samples were obtained from 368 volunteers (345 samples for MGWAS and 23 additional samples for T2D classification) after signing an informed consent form. The sampling procedure was approved by the Ethical Committee for Clinical Research from the Peking University Shenzhen Hospital, Shenzhen Second People's Hospital and Medical Research Center of Guangdong General Hospital. The individuals had not received any antibiotic treatment within 2 months before sample collection. The samples were frozen immediately and underwent DNA extraction using standard methods[34].

Sequencing and data processing. Illumina GAIIx and HiSeq 2000 were used to sequence the samples. We constructed a paired-end library with insert size of ~350 base pairs for every sample. Adaptor contamination and low-quality reads were discarded from the raw reads, and the remaining reads were filtered to eliminate human host DNA based on the human genome reference (hg18).

Full Methods and associated references are available in the Supplementary Information.

Received 30 August 2011; accepted 27 July 2012.
Published online 26 September 2012.

1. Wellen, K. E. & Hotamisligil, G. S. Inflammation, stress, and diabetes. *J. Clin. Invest.* **115,** 1111–1119 (2005).
2. Risérus, U., Willett, W. C. & Hu, F. B. Dietary fats and prevention of type 2 diabetes. *Prog. Lipid Res.* **48,** 44–51 (2009).
3. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447,** 661–678 (2007).
4. Scott, L. J. *et al.* A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* **316,** 1341–1345 (2007).
5. Musso, G., Gambino, R. & Cassader, M. Interactions between gut microbiota and host metabolism predisposing to obesity and diabetes. *Annu. Rev. Med.* **62,** 361–380 (2011).
6. Eckburg, P. B. *et al.* Diversity of the human intestinal microbial flora. *Science* **308,** 1635–1638 (2005).
7. Turnbaugh, P. J. *et al.* A core gut microbiome in obese and lean twins. *Nature* **457,** 480–484 (2009).
8. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464,** 59–65 (2010).
9. The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486,** 207–214 (2012).
10. The Human Microbiome Project Consortium. A framework for human microbiome research. *Nature* **486,** 215–221 (2012).
11. Vijay-Kumar, M. *et al.* Metabolic syndrome and altered gut microbiota in mice lacking Toll-like receptor 5. *Science* **328,** 228–231 (2010).
12. Bäckhed, F. *et al.* The gut microbiota as an environmental factor that regulates fat storage. *Proc. Natl Acad. Sci. USA* **101,** 15718–15723 (2004).
13. Ley, R. E. *et al.* Obesity alters gut microbial ecology. *Proc. Natl Acad. Sci. USA* **102,** 11070–11075 (2005).
14. Zhang, H. *et al.* Human gut microbiota in obesity and after gastric bypass. *Proc. Natl Acad. Sci. USA* **106,** 2365–2370 (2009).
15. Bäckhed, F., Manchester, J. K., Semenkovich, C. F. & Gordon, J. I. Mechanisms underlying the resistance to diet-induced obesity in germ-free mice. *Proc. Natl Acad. Sci. USA* **104,** 979–984 (2007).
16. Turnbaugh, P. J. *et al.* An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444,** 1027–1031 (2006).
17. Manichanh, C. *et al.* Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. *Gut* **55,** 205–211 (2006).
18. Joossens, M. *et al.* Dysbiosis of the faecal microbiota in patients with Crohn's disease and their unaffected relatives. *Gut* **60,** 631–637 (2011).
19. Larsen, N. *et al.* Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults. *PLoS ONE* **5,** e9085 (2010).
20. Arumugam, M. *et al.* Enterotypes of the human gut microbiome. *Nature* **473,** 174–180 (2011).
21. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genet.* **38,** 904–909 (2006).
22. Woo, P. C. Y. *et al.* Bacteremia due to *Clostridium hathewayi* in a patient with acute appendicitis. *J. Clin. Microbiol.* **42,** 5947–5949 (2004).
23. Elsayed, S. & Zhang, K. Bacteremia caused by *Clostridium symbiosum*. *J. Clin. Microbiol.* **42,** 4390–4392 (2004).
24. McClean, K. L., Sheehan, G. J. & Harding, G. K. Intraabdominal infection: a review. *Clin. Inf. Dis.* **19,** 100–116 (1994).
25. Brook, I. Clostridial infection in children. *J. Med. Microbiol.* **42,** 78–82 (1995).
26. Greenblum, S., Turnbaugh, P. J. & Borenstein, E. Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. *Proc. Natl Acad. Sci. USA* **109,** 594–599 (2012).
27. McArdle, B. H. & Anderson, M. J. Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology* **82,** 290–297 (2001).
28. Yang, B. *et al.* Unsupervised binning of environmental genomic fragments based on an error robust selection of l-mers. *BMC Bioinformatics* **11** (suppl. 2), S5 (2010).
29. Krause, L. *et al.* Phylogenetic classification of short environmental DNA fragments. *Nucleic Acids Res.* **36,** 2230–2239 (2008).
30. Wang, T. *et al.* Structural segregation of gut microbiota between colorectal cancer patients and healthy volunteers. *ISME J.* **6,** 320–329 (2012).
31. Biagi, E. *et al.* Through ageing, and beyond: gut microbiota and inflammatory status in seniors and centenarians. *PLoS ONE* **5,** e10667 (2010).
32. Kashyap, P. & Farrugia, G. Oxidative stress: key player in gastrointestinal complications of diabetes. *Neurogastroenterol. Motil.* **23,** 111–114 (2011).
33. Lyssenko, V. *et al.* Clinical risk factors, DNA variants, and the development of type 2 diabetes. *N. Engl. J. Med.* **359,** 2220–2232 (2008).
34. Godon, J. J., Zumstein, E., Dabert, P., Habouzit, F. & Moletta, R. Molecular microbial diversity of an anaerobic digestor as determined by small-subunit rDNA sequence analysis. *Appl. Environ. Microbiol.* **63,** 2802–2813 (1997).
35. Li, S. *et. al.* Type 2 diabetes gut metagenome (microbiome) data from 368 Chinese samples. *GigaScience* http://dx.doi.org/10.5524/100036 (2012).
36. Wu, G. D. *et al.* Linking long-term dietary patterns with gut microbial enterotypes. *Science* **334,** 105–108 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank L. Goodman for editing the manuscript and providing comments. This research was supported by the Ministry of Science and Technology of China, 863 program (2012AA02A201), the National Natural Science Foundation of China (30890032, 30725008, 30811130531, 31161130357), the Shenzhen Municipal Government of China (ZYC200903240080A, BGI20100001, CXB201108250096A, CXB201108250098A), the Danish Strategic Research Council grant (2106-07-0021), the Ole Rømer grant from Danish Natural Science Research Council, the Solexa project (272-07-0196), and the European Commission FP7 grant HEALTH-F4-2007-201052. The Lundbeck Foundation Centre for Applied Medical Genomics in Personalised Disease Prediction, Prevention and Care (LuCamp, www.lucamp.org). The Novo Nordisk Foundation Center for Basic Metabolic Research is an independent Research Center at the University of Copenhagen partially funded by an unrestricted donation from the Novo Nordisk Foundation (http://www.metabol.ku.dk). We are also indebted to many additional faculty and staff of BGI-Shenzhen who contributed to this work.

Author Contributions The project idea was conceived and the project was designed by Ju.W., K.K., O.P., R.N. and S.D.E.; J.Q., Y.L., Sh.L. and Ju.W. managed the project. F.Z., Z.C., R.X., Su.L., L.H., D.L., P.W., Y.D., X.S., Z.L., A.T., S.Z., M.W., Q.F. and T.H. performed sample collection and clinical study. Wen.Z., M.G., J.Y., Y.Z. and W.X. performed DNA experiments. Ju.W., K.K., O.P., R.N., S.D.E., J.Q., Y.L., Sh.L. and J.Z. designed the analysis. J.Q., Y.L., Sh.L., J.Z., Su.L., Y.G., Y.P., D.S., X.L., W.C., D.Z., Y.Q., M.Z., Z.Z., Z.J., G.S., J.L., J.R., S.O., H.C. and W.W. performed the data analysis. J.Q., Sh.L., J.Z., Y.G., Y.P., M.A., E.L., P.R., N.P. and J.-M.B. worked on metagenomic linkage group method. J.Q., D.S., Su.L., Y.Q., J.R., G.F. and S.O. did the functional annotation analyses. J.Q., Sh.L., D.S., J.Z., Y.P. and Y.L. wrote the paper. Ju.W., O.P., K.K., R.N., S.D.E., Ji.W., H.Y., So.L., Wei.Z. and R.Y. revised the paper.

Author Information The raw Illumina read data of all 368 samples has been deposited in the NCBI Sequence Read Archive under accession numbers SRA045646 and SRA050230. The assembly data, updated metagenome gene catalogue, annotation information, and MGLs are published in the *GigaScience* database, *Giga*DB[35]. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to Ju.W. (wangj@genomics.org.cn).