doi:10.1038/nature12198

Gut metagenome in European women with normal, impaired and diabetic glucose control

Fredrik H. Karlsson[1]*, Valentina Tremaroli[2]*, Intawat Nookaew[1], Göran Bergström[2], Carl Johan Behre[2], Björn Fagerberg[2], Jens Nielsen[1] & Fredrik Bäckhed[2,3]

Type 2 diabetes (T2D) is a result of complex gene–environment interactions, and several risk factors have been identified, including age, family history, diet, sedentary lifestyle and obesity. Statistical models that combine known risk factors for T2D can partly identify individuals at high risk of developing the disease. However, these studies have so far indicated that human genetics contributes little to the models, whereas socio-demographic and environmental factors have greater influence[1]. Recent evidence suggests the importance of the gut microbiota as an environmental factor, and an altered gut microbiota has been linked to metabolic diseases including obesity[2,3], diabetes[4] and cardiovascular disease[5]. Here we use shotgun sequencing to characterize the faecal metagenome of 145 European women with normal, impaired or diabetic glucose control. We observe compositional and functional alterations in the metagenomes of women with T2D, and develop a mathematical model based on metagenomic profiles that identified T2D with high accuracy. We applied this model to women with impaired glucose tolerance, and show that it can identify women who have a diabetes-like metabolism. Furthermore, glucose control and medication were unlikely to have major confounding effects. We also applied our model to a recently described Chinese cohort[4] and show that the discriminant metagenomic markers for T2D differ between the European and Chinese cohorts. Therefore, metagenomic predictive tools for T2D should be specific for the age and geographical location of the populations studied.

The composition of the gut microbiota differs among geographical locations, and between elderly people, in whom T2D incidence is high, and younger subjects[6–10]. In addition, studies of T2D are complicated by the heterogeneous manifestations and mixed aetiology of the disease, and confounded by the effects of age, gender, degree of glucose control and concomitant treatment. In this study, we examined the composition and function of the faecal microbiota in a well-characterized population of 70-year-old European women to minimize sources of variation. Our cohort was selected using a stratified randomized method from a population-based screening sample[11,12] and classified into three similarly sized subgroups: women who had T2D ($n = 53$), impaired glucose tolerance (IGT; $n = 49$) or normal glucose tolerance (NGT; $n = 43$) (Methods and Supplementary Tables 1–3). Genomic DNA was extracted from faecal samples using a standard procedure[13] and sequenced on Illumina HiSeq 2000. In total, we obtained 453 gigabases (Gb) of paired-end reads, with an average of 3.1 ± 1.8 Gb (mean ± s.d.) for each sample (Supplementary Table 4).

To determine the composition of the gut microbiota, we aligned filtered Illumina reads to 2,382 non-redundant reference genomes obtained from the NCBI and HMP databases (http://www.hmpdacc.org) (Supplementary Table 5) using our recently published MEDUSA platform[5]. We compared the composition of T2D and NGT communities and observed increases in the abundance of four *Lactobacillus* species and decreases in the abundance of five *Clostridium* species in the T2D group (adjusted $P < 0.05$, Wilcoxon rank sum test) (Supplementary Fig. 1a and Supplementary Table 6). In the total cohort, *Lactobacillus* species correlated positively with fasting glucose and HbA1c (glycosylated haemoglobin), a long-term measure of blood glucose control (adjusted $P < 0.05$, Spearman correlation). By contrast, *Clostridium* species correlated negatively with fasting glucose, HbA1c, insulin, C-peptide and plasma triglycerides, and positively with adiponectin and HDL (Supplementary Fig. 1b and Supplementary Table 7). These correlations are relevant for T2D because high triglycerides and low HDL levels are components of the dyslipidaemia typically found in T2D, whereas serum levels of the insulin-sensitizing hormone adiponectin are reduced in people at risk of T2D (ref. 14). Importantly, these *Lactobacillus* and *Clostridium* species did not correlate with body mass index (BMI), waist circumference or waist-to-hip ratio (WHR) (Supplementary Fig. 1b).

To identify microbial species independently of reference genomes and fully exploit the information contained in the metagenomic data, we performed *de novo* assembly of filtered sequence data. The total length of the assembly was 13.6 Gb, from which 18.6 million genes with a length longer than 100 base pairs (bp) could be predicted. We created a non-redundant gene catalogue for our cohort and merged it with the MetaHIT gene catalogue[15]. The merged gene catalogue was used to align reads. The faecal microbiota of NGT, IGT and T2D women contained similar numbers of genes (Supplementary Fig. 2). We clustered these genes based on their profile across samples with the assumption that genes from the same genome should have a similar abundance within each subject. We considered only genes that were shared among at least 10 subjects (2.9 million genes) and calculated the correlation coefficient across subjects. We clustered sets of genes with high correlation between them (Pearson rho > 0.85) and defined these sets as metagenomic clusters (MGCs) (Supplementary Fig. 3). The 800 largest MGCs contained at least 104 genes, and 550,188 genes were included in total (Supplementary Table 8; distribution of the number of genes in MGCs shown in Fig. 1a).

To determine the phylogenetic origin of the MGCs, we blasted the genes in each cluster against the NCBI non-redundant catalogue and determined the lowest common ancestor (LCA) by requiring that at least 50% of the genes had a best hit to the same phylogenetic group (Supplementary Fig. 4). This analysis showed that 36% of the MGCs had an LCA at the species level (Fig. 1b), and that MGCs with an LCA at the order level (30%) were mainly Clostridiales (98%) and few Bacteroidales (2%). The Clostridiales order is very diverse and reference genomes might be lacking in public databases, thus explaining the difficulty of the taxonomic characterization.

We tested the abundance of the 800 largest MGCs in NGT and T2D samples, and found 26 clusters to be differentially abundant between the two groups (adjusted $P < 0.05$, Wilcoxon rank sum test) (Fig. 1c and Supplementary Table 9). The MGCs most significantly enriched in

[1]Department of Chemical and Biological Engineering, Chalmers University of Technology, SE-412 96 Gothenburg, Sweden. [2]The Wallenberg Laboratory and Sahlgrenska Center for Cardiovascular and Metabolic Research, Department of Molecular and Clinical Medicine, Institute of Medicine, University of Gothenburg, SE-413 45 Gothenburg, Sweden. [3]Novo Nordisk Foundation Center for Basic Metabolic Research, Section for Metabolic Receptology and Enteroendocrinology, Faculty of Health Sciences, University of Copenhagen, Copenhagen DK-2200, Denmark.
*These authors contributed equally to this work.

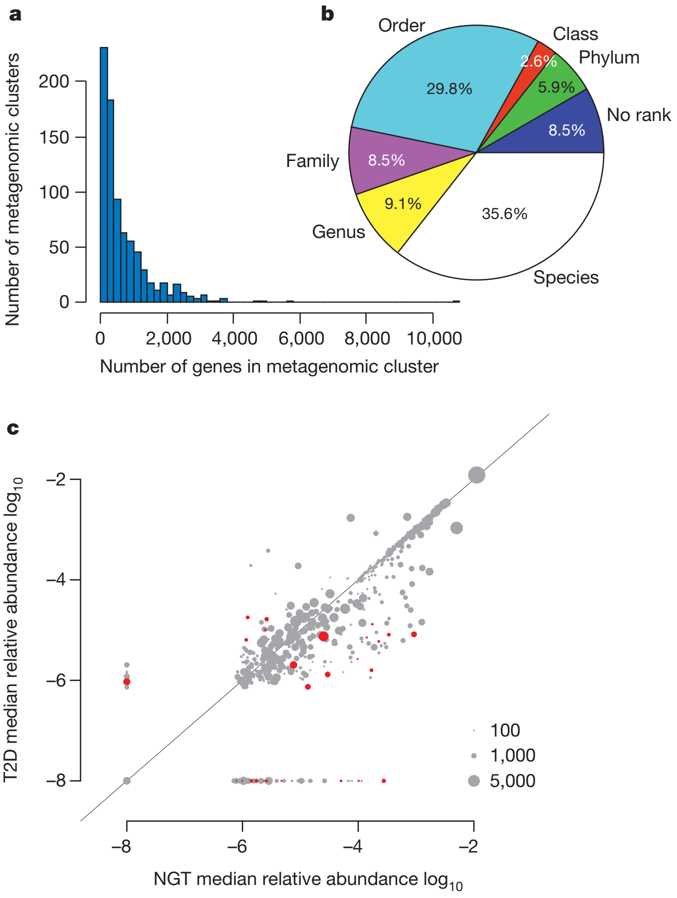

Figure 1 | Definition of MGCs and identification of differentially abundant MGCs in T2D and NGT. **a**, Histogram of the number of genes in the 800 largest MGCs, all with more than 100 genes. **b**, Pie chart of the taxonomic annotation level of MGCs. **c**, Scatter plot of median MGC abundance in T2D ($n = 53$) and NGT ($n = 43$) women. Grey points represent MGCs not differentially abundant between groups, and red points represent differentially abundant MGCs (adjusted $P < 0.05$, Wilcoxon rank sum test).

T2D women were a Clostridiales identified at order level and two *Clostridium clostridioforme*. Two other MGCs were enriched in T2D microbiota, and were identified at species levels as *Lactobacillus gasseri* and *Streptococcus mutans*. *C. clostridioforme* correlated positively with triglyceride and C-peptide levels, whereas *L. gasseri* correlated positively with fasting glucose and HbA1c (Fig. 2 and Supplementary Table 10). Twenty-one MGCs were significantly depleted in T2D, including *Roseburia* (that is, *Roseburia*_272), two unknown *Clostridium* species, several Clostridiales, two *Eubacterium eligens*, Coriobacteriaceae and one *Bacteroides intestinalis*. In the total cohort, the clostridial MGCs correlated negatively with C-peptide, insulin and triglyceride levels, whereas *B. intestinalis* correlated negatively with insulin and waist circumference (Fig. 2 and Supplementary Table 10). These results largely agree with those obtained from the species-based analyses (Supplementary Fig. 1).

To test whether the microbiota composition can identify diabetes status, we trained a random forest model in a training set of the NGT and T2D subjects using the profiles of species and MGCs. We evaluated its performance using a tenfold cross-validation approach and scored the predictive power in a receiver operating characteristic (ROC) analysis. The discriminatory power of species and MGCs was calculated as the area under the ROC curve (AUC). T2D was identified more accurately with MGCs (highest AUC = 0.83) than with microbial species (highest AUC = 0.71) (Fig. 3a and Supplementary Table 11). The increased AUC for the MGC-based model can be explained by the fact that MGCs also provide taxonomical and functional information for unknown species. Therefore, the MGC-based method has the advantage that it can also be applied when reference genomes are missing. When BMI, WHR and waist circumference were used for predicting T2D, we obtained a maximum AUC of 0.70 for waist circumference (BMI, AUC = 0.58; WHR, AUC = 0.60), thus showing that the composition of the microbiota determined by MGCs correlates better with diabetes than these known T2D risk factors[16]. Importantly, the T2D score obtained based on MGCs is similar to other published scores that combine several known risk factors for diabetes development (for example, the FINDRISC score, validated in several countries[1]).

L. gasseri had the highest score for the identification of T2D women in both models (species and MGCs; Fig. 3b, c). Lactobacilli and clostridia were among the ten most important bacteria in the species model (Fig. 3b), whereas *Roseburia*, several Clostridiales, *B. intestinalis*, *C. clostridioforme* and Coriobacteriaceae were among the ten most important clusters in the model based on MGCs (Fig. 3c). The two models indicated different bacterial groups as most discriminant for T2D identification, but the bacteria identified by the MGC model had higher scores than those identified by the species model (Fig. 3b, c). Notably, the MGC model identified *Roseburia* and *Faecalibacterium prausnitzii* as highly discriminant for T2D. These bacteria are known human gut colonizers and butyrate producers[17], and have been linked to improved insulin sensitivity and diabetes amelioration in studies of the human faecal microbiota[18,19].

Many patients are not diagnosed with T2D until cardiovascular complications are apparent[20], but IGT and other metabolic defects often appear before T2D develops[21]. We used our random forest model trained for the discrimination of NGT and T2D individuals to stratify the 49 IGT women of the cohort: 10 IGT women were included in the NGT subgroup, whereas 34 were included in the T2D subgroup (5 could not be classified, as the probability of being either NGT or T2D was 0.5 ± 0.02; Fig. 4a). The characteristics of the two subgroups stratified according to the faecal metagenome showed that plasma levels of triglycerides and C-peptide were significantly higher in the subgroup identified as T2D than in the subgroup identified as NGT (Fig. 4b, c).

To characterize microbial functions, we annotated all of the genes in our catalogue to the KEGG database (version 59). We used the reporter feature algorithm[22,23] in combination with the KEGG metabolic network, pathway annotations and the information about relative gene abundance to identify reporter pathways (that is, pathways with significantly differentially abundant KEGG orthologues) that were associated with T2D and NGT status. We found that NGT and T2D communities had different functional composition and several reporter pathways were differentially abundant in T2D and NGT women (Supplementary Table 12). The pathways that showed the highest scores for enrichment in T2D metagenomes included KEGG orthologues for starch and glucose metabolism (39 out of 46), fructose and mannose metabolism (37 out of 49), and ABC transporters for amino acids, ions and simple sugars (123 out of 174) (Supplementary Table 12). These results are in agreement with previous studies showing an increase in microbial functions for energy metabolism and harvest in the obese microbiome[2,3]. Other metabolic pathways containing KEGG orthologues enriched in women with T2D included glycerolipid metabolism and fatty acid biosynthesis. Pathways for cysteine and methionine metabolism were also enriched in T2D; these pathways are related to glutathione synthesis and may be important for response to oxidative stress.

Similar to our observations, genes related to membrane transporters and oxidative stress resistance were enriched also in the Chinese T2D metagenome[4]. In our study, microbial functions enriched in NGT women were related to flagellar assembly and riboflavin metabolism (Supplementary Table 12). Interestingly, the metagenome of healthy individuals in the Chinese cohort was also enriched in functions related to flagellar assembly and metabolism of cofactors and vitamins[4].

We next investigated whether the composition and functionality of the gut microbiota is influenced by factors other than prevalent T2D,

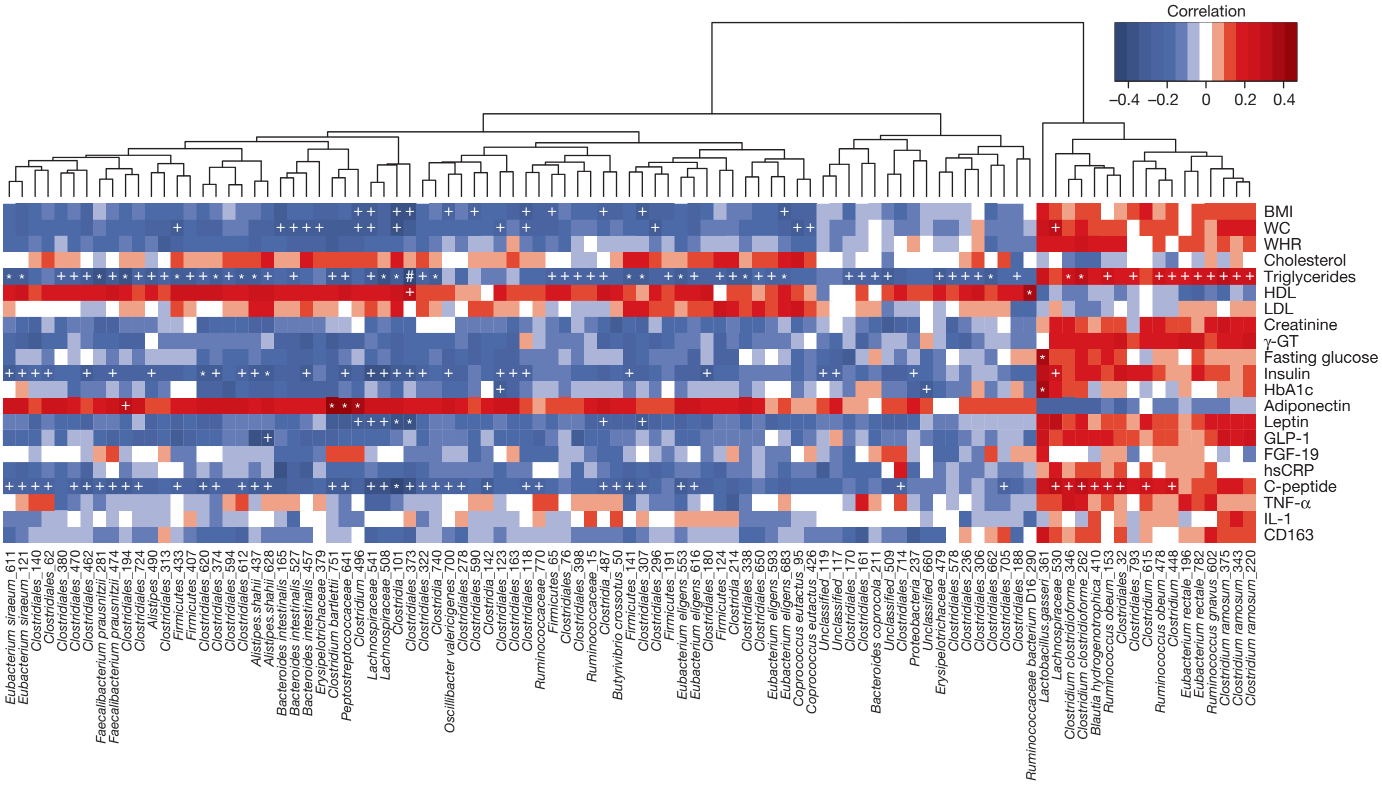

Figure 2 | Associations of MGCs with clinical biomarkers. Spearman's rank correlation coefficients and P values for the correlations are listed in Supplementary Table 10. $n = 145$; '+' denotes adjusted $P < 0.05$; '*' denotes adjusted $P < 0.01$; '#' denotes adjusted $P < 0.001$. FGF-19, fibroblast growth factor 19; γ-GT, γ-glutamyltransferase; GLP-1, glucagon-like peptide 1; HDL, high-density lipoprotein; hsCRP, high-sensitivity C-reactive protein; LDL, low-density lipoprotein; WC, waist circumference.

such as family history of diabetes, medication (that is, statins and metformin) or degree of blood glucose control (as measured by HbA1c levels; $<47\ \mathrm{mmol\ mol}^{-1}$ indicates good control). We observed that several microbial species and gene functions were associated with metformin and glucose control (Supplementary Results, Supplementary Fig. 5a, b and Supplementary Tables 13–16). However, these associations should not have a major confounding effect on the model for the discrimination of T2D women based on faecal microbiota composition as only two of the species included in our model were affected by the use of metformin (*Clostridium botulinum* B str. Eklund 17B and *Clostridium* sp. 7_2_43FAA; Supplementary Table 13) and two others were affected by poor glucose control (*Clostridium thermocellum* DSM 1313 and *Streptococcus* sp. C150; Supplementary Table 14). Furthermore, these associations were not identified using an MGC-based approach (Supplementary Fig. 5c, d).

The similarities between the results obtained in our study and those reported previously[4] clearly underline the link between gut microbiota and T2D. To confirm further the similarities independent of methodological differences, we analysed the Chinese metagenomic data with our bioinformatics platform and compared the results with those from our cohort (Supplementary Results, Supplementary Figs 6–15 and Supplementary Tables 17–19).

The Chinese and European populations clustered separately in principal component analysis plots of species and MGCs abundance (Supplementary Fig. 10), which may be a result of different genetics and/or dietary habits. However, in agreement with the previous study[4] we observed that *Clostridium clostridioforme* MGCs were increased whereas *Roseburia*_272 was decreased in T2D metagenomes from both cohorts (Supplementary Tables 9 and 18). *C. clostridioforme* is a mixture of three opportunistic pathogens (*C. bolteae*, *C. hathewayi* and *C. clostridioforme*) that have been associated with bacteraemia and infections in humans[24], whereas *Roseburia* contains gut bacteria able to produce butyrate. Gut microbiota transplantations from lean donors to recipients with metabolic syndrome have recently been shown to increase *Roseburia* and butyrate levels together with improved insulin sensitivity[18], thus suggesting the importance of butyrate-producing bacteria for blood glucose regulation in humans.

We also found increased *Lactobacillus* species and MGCs in both T2D cohorts (Supplementary Tables 6, 9, 17 and 18). Increased *Lactobacillus* levels in T2D patients were also observed in another small study[25], which used 16S ribosomal DNA pyrosequencing to analyse the microbiota composition of T2D patients and healthy men. A positive correlation between *Lactobacillus* abundance and blood glucose levels was shown[25], in agreement with our study (Fig. 2 and Supplementary Fig. 1b). The increase in *Lactobacillus* could be a consequence of increased glucose levels in the intestine, as increased lactobacilli resulting from increased salivary glucose have been measured in children with insulin-dependent diabetes mellitus[26].

We used the MGCs identified in our study to train a new random forest model on Chinese metagenomes, and used this model to classify Chinese subjects into T2D and controls. We observed an AUC of 0.82 (Supplementary Fig. 14 and Supplementary Table 19), which is in line with the value of 0.81 reported previously[4] and similar to the value reported for the classification of NGT and T2D women in our cohort (0.83; Fig. 3a and Supplementary Table 11). We observed that the most discriminatory MGCs differed between the Chinese subjects and our cohort (Fig. 3b, c and Supplementary Fig. 15). In particular, *Akkermansia* did not contribute to the classification of T2D women in our cohort, whereas *Lactobacillus* did not contribute to the classification of T2D patients in the Chinese population, thus suggesting that classifiers for T2D based on species are population specific. However, it should be noted that, in contrast to our homogenous cohort (70-year-old women), the T2D population in the previous study was older and included more men than the control population,

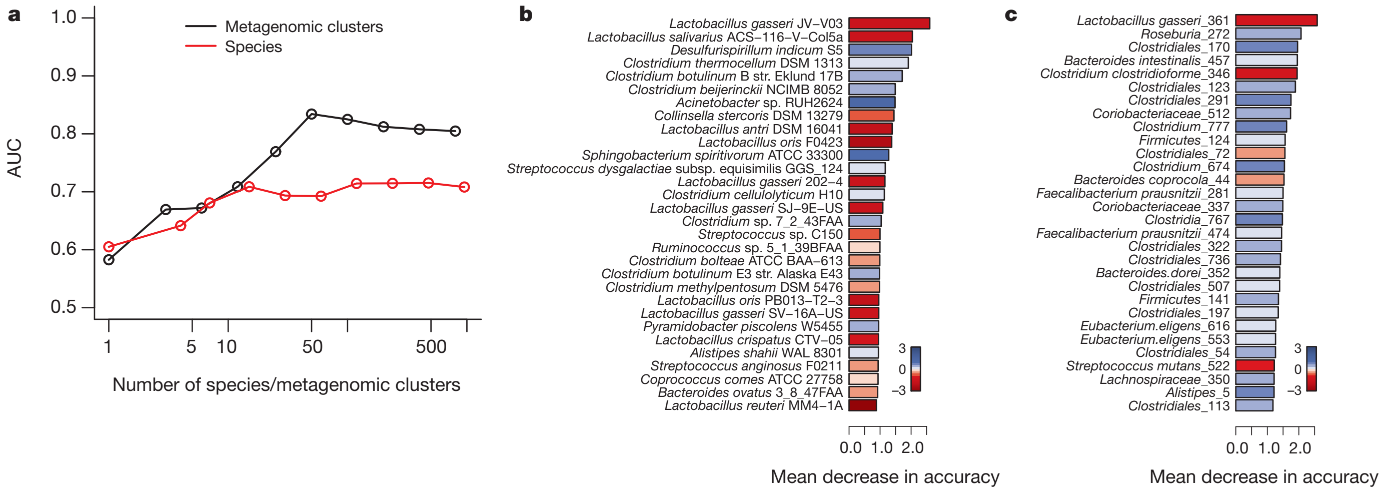

Figure 3 | Classification of T2D status by abundance of species and MGCs. **a**, Classification performance of a random forest model using species or MGC abundance assessed by area under the receiver-operating characteristic curve (AUC). The performance was explored for different numbers of explanatory variables, ordered in importance. **b**, The 30 most discriminant species in the model using 915 species and discriminating between NGT and T2D women. **c**, The 30 most discriminant MGCs in the model using all 800 MGCs and discriminating between NGT and T2D women. The bar lengths in **b** and **c** indicate the importance of the variable, and colours represent enrichment in T2D (red shades) or NGT (blue shades).

which may affect the results. We also tested whether an MGC model trained on one population could classify T2D individuals from the other population. The MGC model based on our cohort had an AUC of 0.58 for the classification of Chinese T2D subjects, whereas the model based on the Chinese cohort had an AUC of 0.66 for the classification of T2D women in our cohort (Supplementary Fig. 16). These AUC values are lower than the values found both in our work and the previous study[4].

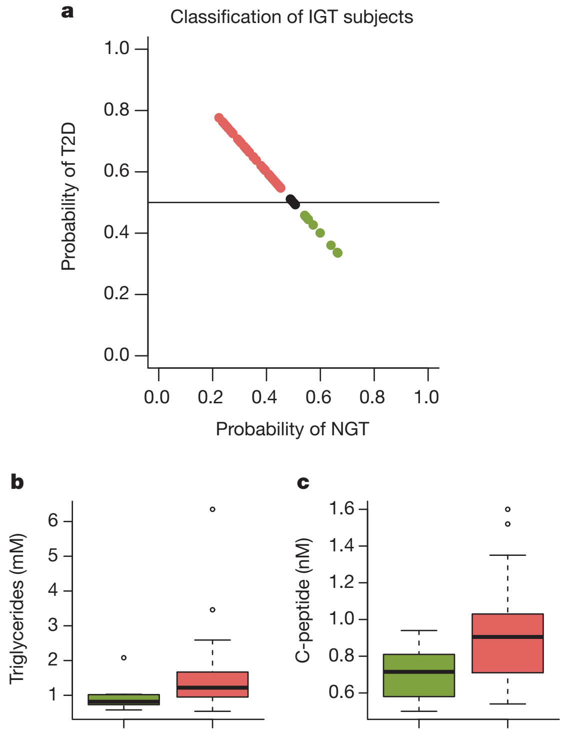

Figure 4 | Stratification of IGT women based on gut microbiota profiles. **a**, Use of the MGC model trained for discriminating NGT and T2D to classify IGT women ($n = 49$) as either NGT (green) or T2D (red). **b**, **c**, IGT women predicted to be T2D had higher triglyceride levels ($P = 0.019$, Wilcoxon rank sum test) (**b**) and higher C-peptide levels ($P = 0.030$, Wilcoxon rank sum test) (**c**). Boxes denote the interquartile range between the first and third quartiles, and the line within denotes the median; whiskers denote the lowest and highest values within 1.5 times interquartile range from the first and third quartiles, respectively. Circles denote data points beyond the whiskers.

In summary, we characterized the faecal metagenome of 70-year-old European women with T2D, IGT and NGT, and investigated the role of metformin on the microbiome. We also developed the concept of MGCs, which allows DNA that has not previously been sequenced to be included in the analysis. We showed that MGCs identify T2D more accurately than species, indicating that several important gut species still need to be characterized. In addition, we classified women with IGT into subgroups with T2D- or NGT-like metabolism on the basis of their faecal microbiome; this classification offers a potentially new approach to identify individuals at high risk of developing T2D.

Our results are concordant with the recent report showing associations between the gut microbiota and T2D in Chinese individuals[4], despite differences in age. Both studies suggest that functional alterations of the gut microbiome, possibly reflecting changes in the intestinal environment of T2D patients, might be directly linked to T2D development. Although it is likely that the same microbial-encoded functions contribute to disease in different populations, we observed that the most discriminatory MGCs differed between our European T2D subgroup and the Chinese T2D cohort. This observation underscores the need to sample human populations and perform parallel studies in different continents. It also indicates that the development of T2D metagenomic predictive tools and diagnostic biomarkers should be specific to the populations studied.

METHODS SUMMARY

Sample collection. 70-year-old women were recruited using a protocol approved by the ethics committee at Sahlgrenska University Hospital, and were included if they had T2D, IGT or NGT. Exclusion criteria were chronic inflammatory disease and treatment with antibiotics during the preceding 3 months. The subjects were given material and written instructions for providing faecal samples at home.

DNA extraction and sequencing. Genomic DNA was extracted using standard methods[13] and sequenced on Illumina HiSeq 2000. Libraries for each sample were prepared with a fragment length of approximately 300 bp. Low-quality reads and reads mapping to human DNA were removed from the raw data.

Full Methods and any associated references are available in the online version of the paper.

Received 13 December 2012; accepted 17 April 2013.
Published online 29 May 2013.

1. Noble, D., Mathur, R., Dent, T., Meads, C. & Greenhalgh, T. Risk models and scores for type 2 diabetes: systematic review. *BMJ* **343,** d7163 (2011).

2. Turnbaugh, P. J. *et al.* An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444,** 1027–1031 (2006).
3. Turnbaugh, P. J. *et al.* A core gut microbiome in obese and lean twins. *Nature* **457,** 480–484 (2009).
4. Qin, J. *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490,** 55–60 (2012).
5. Karlsson, F. H. *et al.* Symptomatic atherosclerosis is associated with an altered gut metagenome. *Nature Commun.* **3,** 1245 (2012).
6. Mueller, S. *et al.* Differences in fecal microbiota in different European study populations in relation to age, gender, and country: a cross-sectional study. *Appl. Environ. Microbiol.* **72,** 1027–1033 (2006).
7. Biagi, E. *et al.* Through ageing, and beyond: gut microbiota and inflammatory status in seniors and centenarians. *PLoS ONE* **5,** e10667 (2010).
8. Claesson, M. J. *et al.* Composition, variability, and temporal stability of the intestinal microbiota of the elderly. *Proc. Natl Acad. Sci. USA* **108,** 4586–4591 (2011).
9. Yatsunenko, T. *et al.* Human gut microbiome viewed across age and geography. *Nature* **486,** 222–227 (2012).
10. De Filippo, C. *et al.* Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proc. Natl Acad. Sci. USA* **107,** 14691–14696 (2010).
11. Brohall, G., Behre, C. J., Hulthe, J., Wikstrand, J. & Fagerberg, B. Prevalence of diabetes and impaired glucose tolerance in 64-year-old Swedish women: experiences of using repeated oral glucose tolerance tests. *Diabetes Care* **29,** 363–367 (2006).
12. Fagerberg, B., Kellis, D., Bergstrom, G. & Behre, C. J. Adiponectin in relation to insulin sensitivity and insulin secretion in the development of type 2 diabetes: a prospective study in 64-year-old women. *J. Intern. Med.* **269,** 636–643 (2011).
13. Salonen, A. *et al.* Comparative analysis of fecal DNA extraction methods with phylogenetic microarray: effective recovery of bacterial and archaeal DNA using mechanical cell lysis. *J. Microbiol. Methods* **81,** 127–134 (2010).
14. Gaetti-Jardim, E. Jr, Marcelino, S. L., Feitosa, A. C., Romito, G. A. & Avila-Campos, M. J. Quantitative detection of periodontopathic bacteria in atherosclerotic plaques from coronary arteries. *J. Med. Microbiol.* **58,** 1568–1575 (2009).
15. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464,** 59–65 (2010).
16. Wang, Y., Rimm, E. B., Stampfer, M. J., Willett, W. C. & Hu, F. B. Comparison of abdominal adiposity and overall obesity in predicting risk of type 2 diabetes among men. *Am. J. Clin. Nutr.* **81,** 555–563 (2005).
17. Louis, P., Young, P., Holtrop, G. & Flint, H. J. Diversity of human colonic butyrate-producing bacteria revealed by analysis of the butyryl-CoA:acetate CoA-transferase gene. *Environ. Microbiol.* **12,** 304–314 (2010).
18. Vrieze, A. *et al.* Transfer of intestinal microbiota from lean donors increases insulin sensitivity in subjects with metabolic syndrome. *Gastroenterology* **143,** 913–916 (2012).
19. Furet, J. P. *et al.* Differential adaptation of human gut microbiota to bariatric surgery-induced weight loss: links with metabolic and low-grade inflammation markers. *Diabetes* **59,** 3049–3057 (2010).
20. Lundberg, V., Stegmayr, B., Asplund, K., Eliasson, M. & Huhtasaari, F. Diabetes as a risk factor for myocardial infarction: population and gender perspectives. *J. Intern. Med.* **241,** 485–492 (1997).
21. Vendrame, F. & Gottlieb, P. A. Prediabetes: prediction and prevention trials. *Endocrinol. Metab. Clin. North Am.* **33,** 75–92 (2004).
22. Oliveira, A. P., Patil, K. R. & Nielsen, J. Architecture of transcriptional regulatory circuits is knitted over the topology of bio-molecular interaction networks. *BMC Syst. Biol.* **2,** 17 (2008).
23. Patil, K. R. & Nielsen, J. Uncovering transcriptional regulation of metabolism by using metabolic network topology. *Proc. Natl Acad. Sci. USA* **102,** 2685–2689 (2005).
24. Finegold, S. M. *et al.* Clostridium clostridioforme: a mixture of three clinically important species. *Eur. J. Clin. Microbiol. Infect. Dis.* **24,** 319–324 (2005).
25. Larsen, N. *et al.* Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults. *PLoS ONE* **5,** e9085 (2010).
26. Karjalainen, K. M., Knuuttila, M. L. & Kaar, M. L. Salivary factors in children and adolescents with insulin-dependent diabetes mellitus. *Pediatr. Dent.* **18,** 306–311 (1996).

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

Acknowledgements We would like to acknowledge R. Perkins for editing the manuscript and M.-L. Ekholm, B. Jannemark, G. Östergren-Lundén, H. Kling Bäckhed, C. Schmidt and U. Prahl for technical assistance. This study was supported by the Swedish Research Council, the NovoNordisk foundation, Torsten Söderberg's foundation, Ragnar Söderberg's foundation, Swedish Diabetes foundation, Swedish Heart Lung Foundation, IngaBritt och Arne Lundbergs foundation, Chalmers foundation, Bioinformatics Infrastructure for Life Sciences (BILS), Knut and Alice Wallenberg foundation, the Swedish Foundation for Strategic Research, AstraZeneca R&D Mölndal, Sweden and the regional agreement on medical training and clinical research (ALF) between Region Västra Götaland and Sahlgrenska University Hospital. The computations were performed on Chalmers C3SE computing resources.

Author Contributions C.J.B., B.F. and F.B. conceived and designed the project. F.H.K., V.T., C.J.B. and G.B. performed the experiments. F.H.K., I.N. and J.N. analysed the sequence data. All authors contributed to writing and editing the manuscript.

Author Information Gut metagenome sequences have been deposited in the Sequence Read Archive (sra) under accession code ERP002469. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.N. (nielsenj@chalmers.se) or F.B. (fredrik.backhed@wlab.gu.se).

METHODS

Study design and recruitment. During 2001–2003, all 64-year-old women in Gothenburg, Sweden, were invited to take part in a screening examination[11] that included anthropometric measurements and a 75-g standardized oral glucose tolerance test, which was repeated in those without NGT. The World Health Organization (WHO) criteria[27] were used for the definitions of diabetes mellitus, IGT and NGT. In the screened cohort of 2,595 women, 9.5% had diabetes mellitus and 14.4% had IGT[11]. As described previously in a study of development of diabetes[12], similarly sized randomized groups with diabetes, IGT and NGT underwent a more detailed baseline examination with a re-examination after more than 5 years. T2D was defined as glutamic acid decarboxylase antibodies $<$ 4.6 units[28].

Women were included in the present sub-study if they had T2D, IGT or NGT. Exclusion criteria were chronic inflammatory disease and treatment with antibiotics during the preceding 3 months. The re-examination took place in 2009 and included questionnaires about current and previous diseases, current medication and smoking habits. Anthropometric measurements were made and blood pressure was recorded. The subjects had fasted overnight and venous blood samples were obtained for measurement of cardiovascular risk factors. We also collected information on medication and glucose control, as well as extensive biometric and plasma measurements. The diagnosis of T2D and IGT at this re-examination was also based on WHO recommendations[27]. All subjects received both written and oral information before they gave their consent to participate in the study. The protocol was approved by the ethics committee at Sahlgrenska University Hospital. After recruitment in the study, three subjects were excluded as they had increased levels of glutamic acid decarboxylase antibodies, indicating type 1 diabetes, and one subject could not be included owing to technical problems with the sequencing.

The characteristics of the included subjects are shown in Supplementary Table 1. The change in glucose tolerance status after a mean of 5.6 years of follow-up is shown in Supplementary Table 2. Supplementary Table 3 lists the biometric and plasma measurements, the country of birth, and the number of years lived in Sweden at the time of first examination for each woman.

The subjects were given material and written instructions for providing faecal samples at home. The samples were produced and transferred to the laboratory one day after the examination. Samples were stored at $-80\,^{\circ}\mathrm{C}$ until extraction. Methods for processing faecal samples and isolation of metagenomic DNA have been described previously[13]. DNA concentration was measured with a Nanodrop instrument (Thermo Scientific) and quality was assessed by agarose gel electrophoresis.

Sequencing. All samples were sequenced in the Illumina HiSeq2000 instrument at GATC Biotech with up to 10 samples pooled in one lane. Libraries were prepared with a fragment length of approximately 300 bp. Paired-end reads were generated with 100 bp in the forward and reverse directions.

Data quality control. The length of each read was trimmed with FASTX from the 3′ end of the read using a quality threshold of 20. Read pairs with either reads shorter than 35 bp were removed. Reads that aligned to the human genome (NCBI version 37) (alignment with Bowtie[29], using -n 2 -l 35 -e 200 -best -p 8 -chunkmbs 1024 -X 600 -tryhard) were also removed. This set of high-quality reads was then used for further analysis.

Alignment to reference genomes and taxonomical analysis. The 2,382 microbial reference genomes were obtained from the National Center for Biological Information (NCBI) and Human Microbiome Project (HMP) on 2 August 2011 and were combined into two Bowtie indexes. The metagenomic sequence reads were aligned to reference genomes using Bowtie[29] with the following parameters: -n 2 -l 35 -e 300 -best -p 8 -chunkmbs 1024 -X 600 -tryhard. Mapping results from the two indexes were merged by selecting the alignment with fewest mismatches and a minimum of 90% identities; if a read was aligned to a reference genome with the same number of mismatches, each genome was assigned 1/2 read. The relative abundance of each genome was calculated by summing the number of reads aligned to that genome divided by the total number of reads and scaled by the genome size. In each subject, the relative abundance was scaled to sum to one. The taxonomic rank for every genome (species, genus and phyla) was downloaded from NCBI taxonomy. The relative abundance of taxonomical ranks was calculated by summing the relative abundance of all its members.

***De novo* assembly and gene calling.** High-quality reads were used for *de novo* assembly with Velvet[30] into contigs of at least 500-bp length using a k-mer length of 39 coverage cut-off of 3. k-mer length was tuned to maximize the N50 value. Reads from each subject were assembled separately; unassembled reads were then used in a global final assembly to also identify rare genes. Genes were predicted on the contigs with MetaGeneMark[31]. A non-redundant gene catalogue was constructed with CD-HIT[32] using a sequence indentity cut-off of 0.95, with a minimum coverage cut-off of 0.9 for the shorter sequences. This catalogue contained 5,997,383 microbial genes (Supplementary Table 20) and was merged with the MetaHIT gene catalogue[15] by adding genes that are unique to our study; the combined gene catalogue was used to align reads. To assess the abundance of genes, reads were aligned to the gene catalogue with Bowtie[29] using parameters: -n 2 -l 35 -e 300 -best -p 8 -chunkmbs 1024 -X 600 -tryhard.

There were 4,778,619 genes unique to our catalogue, which could depend on the fact that we used CD-HIT for clustering whereas BLAT was used in the MetaHIT study[15], although the same criteria of 95% sequence identity and 90% coverage on the shorter sequence were used. Alternatively, this could be owing to the younger age of the MetaHIT cohort (52 ± 11 years (mean ± s.d.) versus 70 ± 1 years for our cohort; $P < 0.001$, Student's t-test), as it is known that the faecal microbiota of adults $>$65 years is different from that of younger adults[6–8]. We also tested the hypothesis that the increased number of genes in our catalogue could be due to chimaeric or misassembled reads, and tested this hypothesis by applying the method for assembly validation and quality control described previously[33]. This analysis showed that the high number of genes with limited overlap to the MetaHIT catalogue is not due to chimaeric or misassembled reads.

Metagenomic clusters. Genes were clustered based on their profile across samples with the assumption that genes from the same genome should have a similar abundance in one subject. Clustering was done by calculating the correlation distance (1 − correlation coefficient) and clustering with the Markov cluster algorithm implemented in the MCL software[34]. We considered only genes that are shared among 10 or more subjects and calculated the correlation coefficient across subjects, creating edges between genes with values above 0.85. The network was divided into clusters by the MCL software using the suggested values for inflation parameters of 1.4, 2 and 6. Clustering was marginally affected by this change, which suggests a robust clustering. Inflation of 1.4 was chosen for further analysis. Cluster abundance was calculated by summing the relative abundance of all genes in a cluster. To validate clustering, clusters were taxonomically annotated by blasting each gene in a cluster to NCBI non-redundant database with blastp using 1×10^{-5} as E-value cut-off. MGCs taxonomical origin (LCA) was determined by blasting the genes in each cluster against the NCBI non-redundant catalogue and requiring that at least 50% of the genes had a best hit to the same phylogenetic group.

Functional annotation. All genes in our catalogue were translated to amino acid sequences and aligned to the KEGG database version 59 using USEARCH[10] ($E < 1 \times 10^{-5}$). Each protein was assigned a KEGG orthologue based on the best hit gene in the KEGG database. Using this approach, 30% of the genes could be assigned a KEGG orthologue and 5,971 unique KEGG orthologues were found. The abundance of a KEGG orthologue was calculated by summing the abundance of genes annotated to a feature.

Statistical analysis. All statistical analyses were made in the R software[35]. Differential abundance of species and MGC was tested by Wilcoxon rank sum test. Correlations between serum biomarkers and species or MGCs were tested with Spearman's correlation. When multiple hypotheses were considered simultaneously, P values were adjusted to control the false discovery rate with the method described previously[36]. Only species with a maximum relative abundance in any subject above 10^{-5} was considered in the analyses.

The random forest model has been shown to be a suitable model for exploiting non-normal and dependent data such as metagenomic data[37]. Random forest models were trained using the random forest package in R to identify T2D status in a test set of the NGT and T2D subjects and using the profiles of species and MGCs. The performance of the predictive model was evaluated with a tenfold cross-validation approach and measured as cross-validation error and AUC. Variable importance by mean decrease in accuracy was calculated for the random forest models using the full set of species or MGCs. By ranking the variables by importance, smaller models were constructed containing only the most important variables. The random forest model trained on NGT and T2D subjects was used to classify IGT subjects as NGT or T2D using the profiles of species and MGCs in the random forest package in R with default parameters and 10,000 trees.

For the functional analysis using KEGG orthologues, Wilcoxon rank sum test was used to test differential abundance between groups, and P values were corrected for multiple testing with the method described previously[36]. The KEGG grouping of orthologues into pathways was used as input to the reporter feature algorithm[22] and calculating reporter pathways in which there are differential abundant KEGG orthologues. This algorithm takes as inputs the adjusted P values and fold changes for each KEGG orthologue in a comparison together with the annotation of KEGG orthologues into pathways from the KEGG database. Each pathway is then scored based on the contributing P values of KEGG orthologues and direction by fold changes to calculate a global P value for each pathway.

27. 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 (1998).

28. Bingley, P. J., Bonifacio, E. & Mueller, P. W. Diabetes Antibody Standardization Program: first assay proficiency evaluation. *Diabetes* **52,** 1128–1136 (2003).

29. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).
30. Zerbino, D. R. & Birney, E. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* **18,** 821–829 (2008).
31. Zhu, W., Lomsadze, A. & Borodovsky, M. *Ab initio* gene identification in metagenomic sequences. *Nucleic Acids Res.* **38,** e132 (2010).
32. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22,** 1658–1659 (2006).
33. Hess, M. *et al.* Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science* **331,** 463–467 (2011).
34. Dongen, S. v. *Graph Clustering by Flow Simulation*. PhD thesis, Univ. Utrecht (2000).
35. R Development Core Team. *R: A Language and Environment for Statistical Computing* http://www.R-project.org (R Foundation for Statistical Computing, 2012).
36. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate — a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B* **57,** 289–300 (1995).
37. Knights, D., Costello, E. K. & Knight, R. Supervised classification of human microbiota. *FEMS Microbiol. Rev.* **35,** 343–359 (2011).