Clinical detection of human probiotics and human pathogenic bacteria by using a novel high-throughput platform based on next generation sequencing

Chih-Min Chiu1†, Feng-Mao Lin1†, Tzu-Hao Chang2, Wei-Chih Huang1, Chao Liang1, Ting Yang1, Wei-Yun Wu1, Tzu-Ling Yang1, Shun-Long Weng1,4,5,6* and Hsien-Da Huang1,3*

Abstract
Background: The human body plays host to a vast array of bacteria, found in oral cavities, skin, gastrointestinal tract and the vagina. Some bacteria are harmful while others are beneficial to the host. Despite the availability of many methods to identify bacteria, most of them are only applicable to specific and cultivable bacteria and are also tedious. Based on high throughput sequencing technology, this work derives 16S rRNA sequences of bacteria and analyzes probiotics and pathogens species.

Results: We constructed a database that recorded the species of probiotics and pathogens from literature, along with a modified Smith-Waterman algorithm for assigning the taxonomy of the sequenced 16S rRNA sequences. We also constructed a bacteria disease risk model for seven diseases based on 98 samples. Applicability of the proposed platform is demonstrated by collecting the microbiome in human gut of 13 samples.

Conclusions: The proposed platform provides a relatively easy means of identifying a certain amount of bacteria and their species (including uncultivable pathogens) for clinical microbiology applications. That is, detecting how probiotics and pathogens inhabit humans and how affect their health can significantly contribute to develop a diagnosis and treatment method.

Background
High throughput sequencing can analyze a large amount of sequences, enabling sequencing of 16S rRNA to identify complex bacteria species of pathogens and probiotic bacteria. Many naturally occurring bacteria form a complex population in the environment. The human body plays host to a vast array of bacteria, found in oral cavities, skin, gastrointestinal tract and the vagina. Some bacteria are harmful while others are beneficial to the host.

A pathogen is a microorganism that causes disease in its host. For example, bacterial pathogen include Corynebacterium diphtheria (causes diphtheria), Listeria monocytogenes (causes food poisons), and Legionella pneumophila (causes Legionnaires’ disease). Probiotics, another microorganism, benefit the host and has received considerable attention in recent years. A FAO report in 2001 [1] cited the advantages of probiotics as increasing immunity [2,3], reducing gastrointestinal discomfort [4,5], and protecting the flora within urogenital tract [6]. As is well known, probiotics can ameliorate symptoms of diseases [7] and reduce the risk of suffering from diseases [8,9].

Despite the availability of many approaches to identify probiotics and pathogens, most of them are only applicable to specific and cultivable bacteria but time consuming. For instance, conventional methods detect growth of cultured bacteria in approximately two days, or an additional five days to obtain no-growth culture results [10], which is laborious. Besides, some bacteria cannot be cultured [11], subsequently increasing the difficulty of specifying pathogenic bacteria. Moreover, it is hard to
determine whether an infection is caused by one or more bacteria types.

16S rRNA sequences, capable of identifying bacteria on a molecular level, can detect uncultivable bacteria [12]. Use of 16S rRNA sequencing can overcome some problems of conventional culture method [13]. Although 16S rRNA sequencing is a more effective means of identifying bacteria than conventional culture method, 16S rRNA sequencing takes a considerable amount of time in amplifying DNA sequences [14]. Sanger sequencing known as “first-generation” or “conventional” sequencing has been used for DNA sequencing for almost two decades. Next generation sequencing (NGS) can analyze large-scale sequences quicker, enable massively parallel analysis, reduce reagent costs and the size of sample components, and perform high throughput [15]. Thus NGS is more efficient than the Sanger method, which generates one read per sample. In addition, NGS of 16S rRNA more easily identify cultivable or uncultivable bacteria [12].

Because of the improvement of sequencing technology and Bioinformatics approaches, the accuracy in distinguishing bacteria with those methods has been increased. Based on high throughput sequencing technology, this work identifies 16S rRNA sequences of bacteria and analyzes bacteria species. High-throughput sequencing can sequence a large number of 16S rRNA sequence more efficiently; with high-throughput sequencing, researchers can acquire information to identify pathogens and probiotic bacteria [16-18].

Results
Platform application: gut probiotics and pathogens detection
The read statistics of quality filtering and taxonomy assignment are demonstrated in Table 1. Figure 1A illustrated the percentage of probiotics detected by the proposed platform. Table 2 listed the quantities (matched sequenced reads) of probiotics identified in the samples in the case study. The top three identified probiotics in 12 samples are Lactococcus salivarius, Streptococcus thermophilus, and Bifidobacterium longum. Figure 1B and Table 3 listed the proportion and quantities of pathogens, of which top three pathogens are Escherichia coli, Salmonella enteric, and Haemophilus influenza.

Table 4 listed the results of disease risk evaluations. It showed that three diseases of two samples (B031 and B034) had similar distributions in the control group. The three diseases are obesity, colorectal cancer, and constipation. Sample B031 had reached the significance level with P-value 0.0333 and 0.0121 < 0.05 of distribution in constipation and colorectal cancer respectively compared to 98 samples control group using binomial test. Sample B034 had reached the significance level with P-value 0.00257 and 0.0121 < 0.05 of distribution in obesity and colorectal cancer. Evaluated by the association of bacterial risk markers and disease, the results suggested that these two samples had higher risk than 98 samples control group in constipation, colorectal cancer, and obesity. Their enterotypes of gut probiotics and pathogens may be one of risk factors which would cause disease.

Reproducibility and accuracy evaluation of proposed platform
Two replicated experiments of four samples were performed to estimate the reproducibility of the proposed platform. The results of repeated experiments were consistent. The similarity between two repeated experiments were calculated by using UniFrac [19]. Results of each

| ID   | Raw reads | QC   | Bacteria identified | Probiotics | Pathogens |
|------|-----------|------|---------------------|------------|-----------|
| B011 | 125420    | 117451 | 93.65%              | 90952      | 77.44%    | 60  0.07% | 3509 3.86% |
| B012 | 132240    | 120134 | 90.85%              | 94679      | 78.81%    | 3457 3.65% | 20109 21.24% |
| B013 | 151876    | 142585 | 93.88%              | 99025      | 69.45%    | 3452 3.49% | 21341 21.55% |
| B014 | 136419    | 126794 | 94.18%              | 95377      | 75.23%    | 611  0.64% | 6665 6.99%  |
| B016 | 135457    | 126507 | 93.39%              | 89407      | 70.67%    | 49   0.05% | 20870 23.34% |
| B017 | 141682    | 131968 | 93.14%              | 89465      | 67.79%    | 1064 1.19% | 8944 10.00% |
| B018 | 111228    | 102382 | 92.05%              | 56981      | 55.66%    | 910  1.60% | 11630 20.41% |
| B019 | 128532    | 120719 | 93.92%              | 76877      | 63.68%    | 305  0.40% | 2775 3.61%  |
| B020 | 126441    | 121957 | 94.95%              | 89618      | 73.48%    | 123  0.14% | 3673 4.10%  |
| B031 | 140291    | 132311 | 93.88%              | 97962      | 74.04%    | 2129 2.17% | 5194 5.30%  |
| B033 | 142462    | 134554 | 94.45%              | 80548      | 59.86%    | 229  0.28% | 2725 3.38%  |
| B034 | 148854    | 140059 | 94.09%              | 106050     | 75.72%    | 9857 9.29% | 15436 14.56% |
| Total| 1621752   | 1517411| 93.54%              | 1066941    | 70.31%    | 22246 2.09%| 122871 11.52% |

http://www.jclinbioinformatics.com/content/4/1/1
Figure 1 Relative abundance of probiotics and pathogenic bacteria from human gut of all samples. (A) The percentage of probiotics was identified in the samples. (B) The proportion of pathogenic bacteria was identified in the samples in the case study.

Table 2 The quantities (matched sequenced reads) of probiotics identified in the samples in the case study

| Probiotics            | B011 | B012 | B013 | B014 | B016 | B017 | B018 | B019 | B020 | B031 | B033 | B034 |
|-----------------------|------|------|------|------|------|------|------|------|------|------|------|------|
| Lactobacillus salivarius |     |      |      |      |      |      |      |      |      |      |      | 2724 |
| Streptococcus thermophilus |    |      |      |      |      |      |      |      |      |      |      | 4221 |
| Bifidobacterium longum |     |      |      |      |      |      |      |      |      |      |      | 2094 |
| Bifidobacterium adolescentis | 375 | 242  | 2724 | 113  | 53   | 9    | 13   | 5    | 55   | 238  | 4221 |       |
| Bifidobacterium bifidum | 104 | 6120| 1092 | 1092 | 1092 | 1092 | 1092 | 1092 | 1092 | 1092 | 1092 |       |
| Bifidobacterium breve | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   |       |
| Lactobacillus brevis | 11  | 11  | 11  | 11  | 11  | 11  | 11  | 11  | 11  | 11  | 11  |       |
| Lactobacillus casei | 13  | 13  | 13  | 13  | 13  | 13  | 13  | 13  | 13  | 13  | 13  |       |
| Lactobacillus fermentum | 34  | 34  | 34  | 34  | 34  | 34  | 34  | 34  | 34  | 34  | 34  |       |
| Lactobacillus gasseri | 79  | 79  | 79  | 79  | 79  | 79  | 79  | 79  | 79  | 79  | 79  |       |
| Lactobacillus johnsonii | 7   | 7   | 7   | 7   | 7   | 7   | 7   | 7   | 7   | 7   | 7   |       |
| Lactobacillus paracasei | 4   | 4   | 4   | 4   | 4   | 4   | 4   | 4   | 4   | 4   | 4   |       |
| Lactobacillus plantarum | 3   | 3   | 3   | 3   | 3   | 3   | 3   | 3   | 3   | 3   | 3   |       |
| Lactobacillus reuteri | 2   | 2   | 2   | 2   | 2   | 2   | 2   | 2   | 2   | 2   | 2   |       |
| Lactobacillus rhamnosus | 6792| 6792| 6792| 6792| 6792| 6792| 6792| 6792| 6792| 6792| 6792| 6792 |
| Lactococcus lactis | 2   | 2   | 2   | 2   | 2   | 2   | 2   | 2   | 2   | 2   | 2   |       |
| Streptococcus thermophilus | 5762| 5762| 5762| 5762| 5762| 5762| 5762| 5762| 5762| 5762| 5762|       |

For each species, if the number of reads is 0 for all samples, that species was not shown.

*The leading three probiotics are Lactobacillus salivarius, Streptococcus thermophilus and Bifidobacterium longum.
sample pair (replicate 1 and 2) closely resemble each other. The similarity of UniFrac distance of each sample pair is higher than 0.96 (0.9617 for B014, 0.9872 for B018, 0.9914 for B020, 0.9722 for B033). This implies that the analysis results are reproducible.

Next, accuracy of the platform is evaluated by adding *Lactobacillus reuteri* to a stool sample (B050). Sample B050 contains 24,408 assigned taxons, and *Lactobacillus* has no detected count. Whether the counts of this species in positive control sample (B050S_L) are elevated must be determined. Analysis results indicate that 27,113 taxons are detected in sample B050S_L. In fact, the detected counts of *Lactobacillus reuteri* in sample B050S_L are 1,430, and the percentage of *Lactobacillus* markedly increases from 0% to 5%.

In short, our platform is accurate and reproducible in terms of detecting the quantities of bacterial species of the proposed platform. The results evaluate the accuracy and feasibility of proposed platform in order to identify probiotics and pathogens. While requiring only about one day for detection, not limited in identifying certain bacteria, the proposed platform can detect and quantify multiple bacteria simultaneously.

**Discussion**

Because of the constraint of costs and technical limitations, 16S rRNA sequences obtained in most databases are partial sequences. Many studies thus assign taxonomy by using partial 16S rRNA sequences. In our probiotics and pathogens 16S rRNA sequence database, 17,964 sequences are collected from NCBI nucleotide database, NCBI 16S microbial rRNA database, Green-genes database, and SILVA. Our probiotics and pathogens 16S rRNA database contain less than 39% of 16S rRNA sequences which are longer than 1400 bps. Only 9% of the sequences are close to full length.

This work extracts the V4 region from full length 16S rRNA of microbiome in the human gut as a platform application. Some 16S rRNA variable regions are more dependable than other regions in assigning taxonomy like V3 and V4 [20,21]; in addition, some 16S rRNA variable regions are much conserved. The proportion

---

**Table 3 The quantities (matched sequenced reads) of pathogens identified in the samples in the case study**

| Pathogens                  | B011 | B012 | B013 | B014 | B016 | B017 | B018 | B019 | B020 | B031 | B033 | B034 |
|----------------------------|------|------|------|------|------|------|------|------|------|------|------|------|
| *Bordetella pertussis*     | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 2    |
| *Bacillus abortus*         | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| *Bacillus melitensis*      | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| *Campylobacter jejuni*     | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 11   | 0    | 0    | 0    | 40   |
| *Clostridium botulinum*    | 0    | 38   | 5048 | 4    | 5    | 2    | 1    | 361  | 153  | 1211 | 115  | 59   |
| *Clostridium difficile*    | 0    | 0    | 1    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 1    |
| *Clostridium perfringens*  | 0    | 1    | 2    | 0    | 0    | 0    | 0    | 3    | 10   | 24   | 12   | 93   |
| *Corynebacterium diphtheriae* | 0   | 1    | 0    | 1    | 0    | 0    | 0    | 0    | 1    | 0    | 0    | 3    |
| *Enterococcus faecalis*    | 57   | 13   | 1    | 4    | 8    | 4    | 0    | 20   | 5    | 19   | 6    | 38   |
| *Enterococcus faecium*     | 41   | 8    | 2    | 6    | 5    | 2    | 1    | 22   | 3    | 13   | 1    | 32   |
| *Escherichia coli*         | 1744 | 8560 | 7900 | 3637 | 10651 | 4404 | 5691 | 1424 | 1733 | 210  | 165  | 4483 |
| *Haemophilus influenzae*   | 2    | 1771 | 2    | 1055 | 8    | 49   | 1    | 171  | 15   | 1802 | 2322 | 4502 |
| *Neisseria meningitidis*   | 0    | 2    | 0    | 3    | 1    | 0    | 1    | 1    | 1    | 1    | 1    | 12   |
| *Pseudomonas aeruginosa*   | 1    | 6    | 6    | 4    | 2    | 2    | 3    | 1    | 3    | 0    | 0    | 3    |
| *Salmonella enterica*      | 1570 | 9291 | 7978 | 1849 | 9864 | 4209 | 5726 | 622  | 1658 | 303  | 44   | 4495 |
| *Shigella sonnei*          | 41   | 243  | 239  | 32   | 308  | 122  | 192  | 8    | 41   | 1    | 1    | 98   |
| *Staphylococcus aureus*    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 1    | 0    | 0    | 0    | 1    |
| *Staphylococcus epidermidis* | 0   | 0    | 0    | 0    | 0    | 0    | 0    | 1    | 1    | 0    | 0    | 2    |
| *Streptococcus agalactiae* | 69   | 3    | 0    | 5    | 0    | 1    | 0    | 0    | 3    | 6    | 5    | 92   |
| *Streptococcus pneumoniae* | 46   | 26   | 9    | 16   | 1    | 36   | 3    | 46   | 25   | 272  | 5    | 154  |
| *Streptococcus pyogenes*   | 7    | 76   | 149  | 14   | 10   | 112  | 9    | 94   | 23   | 417  | 45   | 1428 |
| *Vibrio cholerae*          | 0    | 3    | 0    | 0    | 0    | 1    | 1    | 0    | 1    | 0    | 1    | 7    |
| *Yersinia pestis*          | 0    | 0    | 1    | 27   | 2    | 0    | 0    | 1    | 0    | 906  | 0    | 5    |
| *Zoonotic agents*         | 3509 | 20109 | 21341 | 6665 | 20870 | 8944 | 11630 | 2775 | 3673 | 5194 | 2725 | 15436 |

For each species, if the number of reads is 0 for all samples, that species was not shown.

*The leading three pathogens are *Escherichia coli*, *Salmonella enterica* and *Haemophilus influenzae*. *
# Table 4 The result of disease risk evaluations of 12 samples

| Disease              | B011  | B012  | B013  | B014  | B016  | B017  | B018  | B019  | B020  | B031  | B033  | B034  |
|----------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| Constipation         | 2.67E-01 | 2.67E-01 | 2.67E-01 | 1.00E + 00 | 1.00E + 00 | 2.67E-01 | 1.00E + 00 | 1.00E + 00 | 2.67E-01 | 3.34E-02 | 2.67E-01 | 1.00E + 00 |
| Obesity              | 1.34E-01 | 1.34E-01 | 1.00E + 00 | 1.34E-01 | 1.00E + 00 | 1.34E-01 | 1.34E-01 | 1.00E + 00 | 1.00E + 00 | 1.34E-01 | 1.00E + 00 | 2.57E-03 |
| IBS                  | 3.33E-01 | 7.06E-01 | 1.00E + 00 | 3.33E-01 | 1.00E + 00 | 3.33E-01 | 3.33E-01 | 1.00E + 00 | 1.10E-01 | 1.10E-01 | 7.06E-01 | 3.33E-01 |
| Ulcerative colitis   | 9.30E-02 | 4.15E-01 | 1.00E + 00 | 4.15E-01 | 1.00E + 00 | 9.30E-02 | 4.15E-01 | 1.00E + 00 | 9.30E-02 | 1.00E + 00 | 1.00E + 00 | 1.00E + 00 |
| Colorectal cancer    | 4.88E-01 | 2.59E-01 | 9.35E-01 | 7.47E-01 | 7.47E-01 | 7.47E-01 | 7.47E-01 | 2.59E-01 | 4.88E-01 | 4.88E-01 | 7.47E-01 | 1.22E-02 |
| Atopic dermatitis    | 1.83E-01 | 1.00E + 00 | 1.00E + 00 | 1.00E + 00 | 1.00E + 00 | 1.83E-01 | 1.83E-01 | 1.00E + 00 | 1.00E + 00 | 1.00E + 00 | 1.00E + 00 | 1.00E + 00 |
| Allergic rhinitis    | 1.89E-01 | 1.00E + 00 | 1.00E + 00 | 1.00E + 00 | 1.00E + 00 | 1.89E-01 | 1.89E-01 | 1.89E-01 | 1.89E-01 | 1.89E-01 | 1.89E-01 | 1.89E-01 |

The bold numbers represent two samples had reached significance level with P-value less than 0.05 of distribution in three diseases compared to 98 sample control group using evaluation model.
and diversity of probiotics and pathogens may be made diverse by using different 16S rRNA variable regions. The proposed platform is also applicable to other 16S rRNA variable regions for taxonomy assignment. Importantly, a more appropriate region than others must be selected to produce an outcome that is close to full length 16S rRNA sequence.

This work further attempt is to collect common probiotics and pathogens from the literature. Although it may be incomplete, recent advances in sequencing technology make it possible to identify and define an increasing number of bacteria, implying an obvious increase in the number of identified probiotics and pathogens in the future. Efforts are underway in our laboratory to update the list of used probiotics and pathogens.

Previous studies [22-24] identified pathogen or probiotic bacteria by using antibody, 16S rRNA gene microarrays, fluorescence in situ hybridization (FISH), and proteomic methods. In this work, the proposed platform can detect various pathogens and probiotics based on 16S rRNA (rDNA) sequences of bacteria using NGS and Bioinformatics method. An average of 126,451 reads was acquired per sample in this work. It is doubt that the sequencing depth is enough to detect a small amount of probiotics and pathogens. Although increasing the coverage of sequencing can advance the sensitivity of detecting probiotics and pathogens, the sequencing cost will increase. It is important to work out an appropriate coverage of sequencing for detecting probiotics and pathogens.

The results of disease risk evaluations revealed that most of 12 samples did not have resembled distributions of bacteria markers with control group. Only two samples had reached the significance level of distributions. The reason for the phenomenon may be the overlapped bacteria markers between diseases. 28 markers are used in colorectal cancer, and 17 markers are used in irritable bowel syndrome. Six markers are overlapped. For sample B031, the significant distributions in colorectal cancer were partly contributed to the significance in irritable bowel syndrome owing to the overlapped markers. Similarly, two overlapped markers for sample B034 were in colorectal cancer and obesity. In this kind of speculation, the influence of colorectal cancer to irritable bowel syndrome would be six (overlapped markers of CC and IBS) over seventeen (markers of IBS), and the influence of colorectal cancer to obesity would be two (overlapped markers of CC and obesity) over nine (markers of obesity). In addition, the influence of colorectal cancer to constipation and ulcerative colitis would be one over six and two over ten, respectively.

In addition to that some bacteria markers in species level are belong to the marker of genus level and species level, both genus marker and species markers may have associated with affecting the distributions mutually. Continually, collecting more markers and evaluating the distributions with markers in the same level are required for constructing a global prediction model in Taiwanese.

Conclusions
This work constructed a bacterial disease risk evaluation model for seven diseases and developed a novel platform by using NGS and Bioinformatics approach. Compared with the traditional bacteria culture method, our proposed platform can reduce experiment time. Besides, the proportion of probiotics and pathogens (including uncultivable pathogens) in the human body can be detected rapidly with 16 s RNA database of probiotics and pathogens. Furthermore, the proposed platform provides further insight into the cause of disease based on the relation of probiotics, pathogens, and disease. For instance, the type of antibiotics can be adjusted if the pathogens of disease are identified from infected patients. In addition, the proposed platform allows researchers to determine whether the intake of probiotics impacts the human body [25-29]. In the future, this preliminary study will be continuously extended for more bacterial disease markers. For more comprehensive applications, this work will also collect bacteria from other parts of human body as control group data. In fact, a detective method of how the probiotics and pathogens inhabit human can provide new insight for human health. It could improve diagnosis and treatment method.

Methods
Figure 2 illustrates the bioinformatics system flow of the proposed platform, which includes analysis pipeline of NGS. The Figure 2 contains four parts: sequence quality filtering, construction of bacteria sequence database, taxonomy assignment, and disease risk model evaluation. The detailed components in the proposed platform are described below.

Sample collection
In this study, stool samples of 98 Taiwan volunteers were gathered. The samples were collected by Sigma-transwab (Medical Wire) into a tube with Liquid Amies Transport Medium, and stored at 4°C until processing.

DNA extraction
In the case study, fresh faeces were obtained from participants. DNA was extracted directly on stool samples by using a QIAamp DNA Stool Mini Kit (Qiagen). A swab was vortexed vigorously and incubated at room temperature for 1 min. The sample was then transferred to microcentrifuge tubes containing 560 μl Buffer ASL, vortexed, and incubated at 37°C for 30 min. In addition, the suspension was incubated at 95°C for 15 min,
vortexed, and centrifuged at 14,000 rpm for 1 min into pellet stool particles. Extraction was performed following the protocol of the QIAamp DNA Stool Mini Kit. The DNA was eluted with 50 μl Buffer AE, and centrifuged at 14,000 rpm for 1 min. Moreover, the DNA extract was stored at −20°C until further analysis. Finally, DNA extraction was performed, depending on the sample collected.

Library construction and sequencing for V4 region of 16S ribosomal DNA
The PCR primers, F515 (5′-GTGCCAGCMGCCGCGGTAA-3′) and R806 (5′-GGACTACHVGGGTWTCTAAT-3′), were designed to amplify the V4 domain of bacterial 16S ribosomal DNA as described previously [30]. PCR amplification was performed in a 50 μl reaction volume containing 25 μl 2X Taq Master Mix (Thermo Scientific), 0.2 μM of each forward and reverse primer, and 20 ng DNA template. The reaction conditions consisted of an initial 95°C for 5 min, followed by 30 cycles of 95°C for 30 sec, 54°C for 1 min, and 72°C for 1 min, as well as a final extension of 72°C for 5 min. Next, amplified products were checked by 2% agarose gel electrophoresis and ethidium bromide staining. Amplicons were purified using the AMPure XP PCR Purification Kit (Agencourt), and quantified using Qubit dsDNA HS Assay Kit (Qubit) on Qubit 2.0 Fluorometer (Qubit)—all according to respective manufacturer instructions. For V4 library preparation, Illumina adapters were attached to the amplicons using the Illumina TruSeq DNA Sample Preparation v2 Kit. Purified libraries were applied for cluster generation and sequencing on the MiSeq system. The raw sequence files are available for download at http://clinic.mbc.nctu.edu.tw/.

16S rRNA (rDNA) sequence data quality filtering
The raw fastq files obtained by Illumina sequencing machine were quality-filtered using the FASTX-Toolkit. The paired-end 150 bp reads were performed using the minimum acceptable phred quality score of 20, as well as the 70% of bases that must exceed 20 phred quality score. Sequence shorter than 100 nucleotides would be omitted after quality trimming from reads tail. Notably, reads containing ambiguous characters were discarded.

Construction of probiotics and pathogens database
The list of probiotics and pathogens were obtained from literatures or the claims of official departments. Additional file 1: Table S1 lists species of probiotics which were adapted from both literatures [7,9] and the claims of official departments, such as Taiwan Food and Drug Administration [31] and Health Canada [32]. 99 bacterial pathogens were collected from literature [25,26,33-42] and Taiwan Food and Drug Administration [31] (Additional file 1: Table S2).

The 16S rRNA sequences of probiotics and pathogens used for taxonomy mapping were retrieved from the NCBI nucleotide database, NCBI 16S microbial rRNA database, Greengenes database [43] and SILVA [44].
Table 5 Disease-related biomarkers of seven diseases

| Disease            | Marker                  | Correlation | Lower bound | Upper bound | Case | Control | Pubmed ID   |
|--------------------|-------------------------|-------------|-------------|-------------|------|---------|-------------|
| Constipation       | *Escherichia coli*      | -           | 2.86E-03    | 1.52E-01    | 35   | 35      | 20039451    |
|                    | *Roseburia*             | -           | 1.41E-03    | 4.61E-02    | 14   | 12      | 22315951    |
|                    | *Lactobacillus*         | -           | 6.10E-05    | 9.45E-03    | 14   | 12      | 22315951    |
|                    | *Bifidobacterium*       | -           | 5.39E-05    | 1.73E-02    | 14   | 12      | 22315951    |
|                    | *Enterobacteriaceae*    | +           | 1.00E-02    | 4.26E-01    | 14   | 12      | 22315951    |
|                    | *Ruminococcus bromii*  | +           | 1.16E-05    | 4.98E-03    | 8    | 15      | 20014457    |
| Obesity            | *Prevotella*            | -           | 2.46E-03    | 5.36E-01    | 23   | 13      | 20876719    |
|                    | *Bifidobacterium*       | -           | 5.39E-05    | 1.73E-02    | 33   | 30      | 19498350    |
|                    | *Lachnospiraceae*       | -           | 3.11E-03    | 6.74E-02    | 3    | 3       | 19164560    |
|                    | *Verrucomicrobiae*      | -           | 1.43E-05    | 1.78E-02    | 3    | 3       | 19164560    |
|                    | *Akkermansia*           | -           | 1.43E-05    | 1.78E-02    | 3    | 3       | 19164560    |
|                    | *Faecalibacterium prausnitzii* | +   | 7.70E-04    | 2.15E-02    | 15   | 13      | 19849869    |
|                    | *Lactobacillus*         | +           | 6.10E-05    | 9.45E-03    | 20   | 20      | 19774074    |
|                    | *Coriobacteriaceae*     | +           | 3.26E-05    | 4.72E-03    | 3    | 3       | 19164560    |
|                    | *Erysipelotrichaceae*   | +           | 1.35E-04    | 6.64E-03    | 3    | 3       | 19164560    |
| Ulcerative colitis | *Bacteroides uniformis* | -           | 7.63E-04    | 5.44E-02    | 13   | 22      | 21073731    |
|                    | *Bacteroides vulgatus*  | -           | 1.55E-03    | 4.21E-02    | 13   | 22      | 21073731    |
|                    | *Parabacteroides distasonis* | -  | 2.22E-05    | 1.68E-03    | 13   | 22      | 21073731    |
|                    | *Faecalibacterium prausnitzii* | -  | 7.70E-04    | 2.15E-02    | 13   | 27      | 19235886    |
|                    | *Firmicutes*            | -           | 9.18E-02    | 4.50E-01    | 13   | 27      | 19235886    |
|                    | *Clostridium*           | -           | 2.48E-03    | 6.03E-02    | 31   | 30      | 21253779    |
|                    | *Clostridium leptum*    | -           | 9.65E-06    | 1.05E-03    | 13   | 27      | 19235886    |
|                    | *Bifidobacterium*       | -           | 5.39E-05    | 1.73E-02    | 13   | 27      | 19235886    |
|                    | *Bacteroides ovatus*    | -           | 2.04E-04    | 1.81E-02    | 13   | 27      | 21073731    |
|                    | *Escherichia coli*      | +           | 2.86E-03    | 1.52E-01    | 9    | 9       | 16954244    |
| Atopic dermatitis  | *Lactobacillus*         | -           | 6.10E-05    | 9.45E-03    | 68   | 256     | 17604093    |
|                    | *Bifidobacteriales*     | -           | 8.09E-05    | 1.84E-02    | 7    | 27      | 20623646    |
|                    | *Bacteroides*           | +           | 6.56E-02    | 6.37E-01    | 68   | 256     | 17604093    |
|                    | *Clostridium perfringens* | +         | 0.00E+00    | 1.06E-04    | 15   | 15      | 21963389    |
| Colorectal cancer  | *Bacteroides uniformis* | -           | 7.63E-04    | 5.44E-02    | 46   | 56      | 21850056    |
|                    | *Roseburia*             | -           | 1.41E-03    | 4.61E-02    | 46   | 56      | 21850056    |
|                    | *Fusobacterium*         | -           | 3.32E-05    | 2.64E-02    | 50   | 38      | 7574628     |
|                    | *Eubacterium*           | -           | 1.36E-03    | 7.92E-02    | 46   | 56      | 21850056    |
|                    | *Caproccoccus*          | -           | 1.91E-05    | 2.89E-03    | 21   | 23      | 20470058    |
|                    | *Collinsella aerofaciens* | -         | 2.39E-05    | 2.09E-03    | 50   | 38      | 7574628     |
|                    | *Alistipes*             | -           | 4.07E-04    | 2.60E-02    | 46   | 56      | 21850056    |
|                    | *Sutterellaceae*        | -           | 9.39E-04    | 4.85E-02    | 46   | 56      | 21850056    |
|                    | *Escherichia*           | +           | 3.05E-03    | 1.85E-01    | 46   | 56      | 21850056    |
|                    | *Shigella*              | +           | 1.51E-03    | 8.84E-02    | 46   | 56      | 21850056    |
|                    | *Bacteroides fragilis*  | +           | 7.22E-06    | 1.92E-02    | 46   | 56      | 21850056    |
|                    | *Porphyromonas*         | +           | 0.00E+00    | 1.59E-05    | 46   | 56      | 21850056    |
|                    | *Faecalibacterium prausnitzii* | +  | 7.70E-04    | 2.15E-02    | 50   | 38      | 7574628     |
|                    | *Ruminococcus albus*    | +           | 0.00E+00    | 4.95E-04    | 50   | 38      | 7574628     |
|                    | *Streptococcus*         | +           | 1.12E-04    | 6.83E-03    | 46   | 56      | 21850056    |

Chiu et al. Journal of Clinical Bioinformatics 2014, 4:1
http://www.jclinbioinformatics.com/content/4/1/1
Following sequence data collection, we assemble partial sequences which used the same species classification and removed redundant sequences. Additionally, we also removed the unique sequence from only one research support with 3% similarity which shared the same species classification with other sequence.

**Taxonomy mapping**

To generate taxonomy assignments, the proposed platform invoked a modified Smith-Waterman algorithm from miRExpress [45], which can compare pairs of sequences in parallel, for mapping reads to taxons. miRExpress was designed for identifying the best similarity between sequencing reads and miRNA precursor sequences. In our model, it was modified for identifying multiple hits of 16S rRNA sequence mapping results with similarity threshold 0.97. In order to reduce the storage space of output, the SAM format [46] was used to replace the original miRExpress output format for storing alignment results. Furthermore, two kinds of output format were designed. One format

---

**Table 5 Disease-related biomarkers of seven diseases** *(Continued)*

| Genus          | Fold change | P-value          | Percent | Amount | Biomarker |
|----------------|-------------|------------------|---------|--------|-----------|
| Blautia hansenii | +           | 0.00E+00        | 6.77E-05| 50     | 38        |
| Enterococcus   | +           | 0.00E+00        | 1.19E-04| 46     | 56        |
| Bifidobacterium angulatum | + | 0.00E+00        | 2.60E-05| 50     | 38        |
| Blautia producta | +          | 0.00E+00        | 1.10E-04| 50     | 38        |
| Ruminococcus gnavus | + | 1.04E-05      | 2.64E-03| 50     | 38        |
| Eubacterium eligens | +    | 7.09E-05        | 2.05E-02| 50     | 38        |
| Eubacterium rectale | +        | 8.13E-05        | 1.34E-02| 50     | 38        |
| Bacteroides stercolis | +      | 4.87E-05      | 9.42E-02| 50     | 38        |
| Enterobacteriales | +       | 1.00E-02        | 4.26E-01| 10     | 10        |
| Erysipelotrichaceae | +    | 1.35E-04        | 6.64E-03| 50     | 38        |
| Dorea           | +           | 5.67E-05        | 6.08E-03| 21     | 23        |
| Bifidobacterium longum | +      | 1.56E-05      | 3.60E-03| 50     | 38        |
| Faecalibacterium | +           | 1.66E-03        | 6.79E-02| 21     | 23        |

**Irritable bowel syndrome**

| Genus          | Fold change | P-value          | Percent | Amount | Biomarker |
|----------------|-------------|------------------|---------|--------|-----------|
| Bacteroides uniformis | -           | 7.63E-04        | 5.44E-02| 11     | 22        |
| Bacteroides vulgatus | -           | 1.55E-03        | 4.21E-02| 11     | 22        |
| Parabacteroides distasonis | -  | 2.22E-05      | 1.68E-03| 11     | 22        |
| Faecalibacterium praunstizii | - | 7.70E-04      | 2.15E-02| 23     | 23        |
| Bacteroidetes | -           | 2.87E-01        | 7.95E-01| 62     | 46        |
| Bifidobacterium | -           | 5.39E-05        | 1.73E-02| 62     | 46        |
| Bacteroides ovatus | -           | 2.04E-04        | 1.81E-02| 11     | 22        |
| Faecalibacterium | -           | 1.66E-03        | 6.79E-02| 62     | 46        |
| Escherichia coli | +           | 2.86E-03        | 1.52E-01| 14     | 18        |
| Haemophilus | +           | 1.02E-05        | 1.69E-03| 22     | 22        |
| Fusobacterium | +           | 3.32E-05        | 2.64E-02| 23     | 23        |
| Gammaproteobacteria | +    | 1.75E-02      | 4.69E-01| 22     | 22        |
| Ruminococcus | +           | 1.22E-03        | 4.08E-02| 62     | 46        |
| Enterococcus | +           | 0.00E+00        | 1.19E-04| 23     | 23        |
| Veillonella | +           | 1.12E-05        | 7.82E-03| 26     | 26        |
| Lactobacillaceae | +         | 6.10E-05        | 9.45E-03| 23     | 23        |
| Dorea           | +           | 5.67E-05        | 6.08E-03| 62     | 46        |
| Lactobacillus | -           | 6.10E-05        | 9.45E-03| 12     | 12        |
| Bifidobacterium | -           | 5.39E-05        | 1.73E-02| 67     | 20        |
| Bacteroides fragilis | +       | 7.22E-06      | 1.92E-02| 22     | 22        |
| Faecalibacterium praunstizii | + | 7.70E-04      | 2.15E-02| 22     | 22        |

**Allergic rhinitis**

| Genus          | Fold change | P-value          | Percent | Amount | Biomarker |
|----------------|-------------|------------------|---------|--------|-----------|
| Lactobacillus | -           | 6.10E-05        | 9.45E-03| 12     | 12        |
| Bifidobacterium | -           | 5.39E-05        | 1.73E-02| 67     | 20        |
| Bacteroides fragilis | +       | 7.22E-06      | 1.92E-02| 22     | 22        |
| Faecalibacterium praunstizii | + | 7.70E-04      | 2.15E-02| 22     | 22        |

The associations between bacterium and disease are majorly collected from case–control studies which the quantities of bacterium are obtained from deep sequencing data. The proportion of 78 bacteria from control group was applied as risk markers (constipation: 6, obesity: 9, IBS: 17, UC: 10, CC: 28, AD: 4, AR: 4) to predict disease risk to seven diseases in this study.
records whole mapped sequencing reads based on taxons. The other one records which taxons could be assigned based on sequencing reads. These two kinds of output could support the important information for assigning sequencing reads to suitable taxon. miRExpress was originally designed for dealing with single-end sequencing data. Therefore, the additional program was added for processing paired-end sequencing data. In this part, both end sequencing reads need to be assigned to the same taxon. If paired-end sequencing reads were mapped to different taxons, this paired sequence would be dropped. The probiotics and pathogens 16S rRNA sequence from our database were built in FASTA format. Following quality filtering, all paired-end sequences were aligned to the probiotics and pathogens database with whole read aligned from one end to the other end. Reads were then truncated with an identity lower than 97%, according to previous research in order to achieve a better compromise between sequences from PCR sequencing errors and taxonomic relatedness [27].

The construction of Bacterial disease risk evaluation model (BDREM)
To study the associations between bacteria and diseases, we collected related information from literatures. We concerned bacteria that are associated with seven diseases: constipation [28,47,48], obesity [29,49-52], irritable bowel syndrome (IBS) [28,53-58], ulcerative colitis (UC) [53,59-61], colon cancer (CC) [62-64], Atopic Dermatitis (AD) and Allergic rhinitis (AR), were collected positive correlation and negative correlation data, and the individual risk of disease was evaluated.

The association data were majorly collected from case–control studies which the quantities of bacteria were obtained from NGS data, and few well-known bacteria validated by multiple studies through cultural experiments were also included. We further eliminated some conflicted data with both positive and negative correlation between bacteria and disease in different studies.

Health Asians stool samples of 98 Taiwan volunteers were gathered. Following deep sequencing and sequencing data processing, the proportion of 78 bacteria from control group was applied as risk markers (constipation: 6, obesity: 9, IBS: 17, UC: 10, CC: 28, AD: 4, AR: 4) to predict disease risk to seven diseases in this study (Table 5).

The mathematical formula of BDREM in this study was developed as the following steps. Let \( \lambda \) be a \( N \times S \) matrix, where \( N \) is the number of markers selected in the prediction model of constipation and \( S \) is the number of health subjects in 7 prediction models. \( T_i \) was...
defined as one of the two notches of median for each row of $\lambda$ [65]. $T_i$ is a threshold to distinguish $\lambda_{ij}$ from normal proportion level to abnormal (fail to success in one trail of binominal distribution). Smaller notch was selected to $T_i$ when each marker was recorded as a negative association to the disease, and a success trail was identified when $\lambda_{ij}$ is smaller than $T_i$. On the opposite, larger notch was selected when association was positive, and a success trail was identified when $\lambda_{ij}$ is larger than $T_i$.

$$
T_i = \begin{cases} 
\text{Medain of} \ \{\lambda_{i1}, \lambda_{i2}, \ldots, \lambda_{iS}\} + \frac{1.58 \times IQR}{\sqrt{S}}, & \text{positive association} \\
\text{Medain of} \ \{\lambda_{j1}, \lambda_{j2}, \ldots, \lambda_{jS}\} - \frac{1.58 \times IQR}{\sqrt{S}}, & \text{negative association}
\end{cases}
$$

Let $P_j$ be the probability of successful trails in the $j$th column of $\lambda$. The meaning of $P_j$ is the personal probability that abnormal proportion level happened.

$$P_j = \frac{\# \text{success trails in the } j \text{th column of } \lambda}{N}$$

Let $P_p$ be the mean of $P_j$. It represents how frequent the abnormal proportion level happened to all $P_j$ in average, regarded as the hypothesized probability of success in each $P_j$.

$$P_p = \frac{\sum_{j=1}^{S} P_j}{N}$$

Assume $P_j$ obey a binomial distribution, and let $P_i$ be the hypothesized probability (0.05051 for constipation, 0.07239 for obesity, 0.06952 for IBS, 0.05227 for UC, 0.09280 for CC, 0.04924 for AD, 0.05114 for AR). A binominal test was used to $P_j$ and $P_p$. Alpha = 0.05 was choose to judge if a subject is significantly different from the others in $\lambda$.

Figure 3 illustrated an example for evaluating the risk of obesity of B034 and B031. The model used lower and upper bound proportion bound of 9 markers from 98 control samples to define risk markers of these two samples following by using binominal test. Four markers of B034 exceed the lower bound and upper bound of obesity. The binominal test P-Value of B034 is 2.572e-03 < 0.05. Since P-Value < = hypothesized probability 0.07239, this case is significantly associated (significantly) with disease than random chance. There are two markers of B031 exceed lower bound of obesity. The P-Value of B031 is 0.1344 > 0.05, the case is no more associated with disease than random chance. As the results, we can assume that B034 had higher probability to cause Obesity.

### Endnote

6http://hannonlab.cshl.edu/fastx_toolkit/index.html.

### Additional file

#### Additional file 1: The list of probiotics and pathogens were obtained from literatures or the claims of official departments: Table S1. The reference list of probiotics. Table S2. The reference list of pathogens.

#### Competing interests

The authors declare no competing interests.

#### Authors’ contributions

HDH conceived and supervised the study. CMC were responsible for the design, computational analyses, implementation of the system, and drafting the manuscript. FML, THC, CL, TY, TLY, SLW, WCH and WYW were in charge of manuscript revision and data update. All authors read and approved the final manuscript.

#### Acknowledgements

The authors would like to thank the National Science Council of the Republic of China for financially supporting this research under Contract No. NSC 101-2311-B-009-001 and NSC 100-2627-B-009-002. This work was supported in part by the UST-UCSD International Center of Excellence in Advanced Bio-engineering sponsored by the Taiwan National Science Council I-RiCE Program under Grant Number: NSC 101-2911-I-009-101 and NSC 102-2911-I-009-101, and Veterans General Hospitals and University System of Taiwan (VGHUST) Joint Research Program under Grant Number: VGHUST101-G5-1-1. This work was also partially supported by MOE ATU.

#### Author details

1. Institute of Bioinformatics and Systems Biology, National Chiao Tung University, Hsin-Chu 300, Taiwan. 2. Graduate Institute of Biomedical Informatics, Taipei Medical University, Taipei, Taiwan. 3. Department of Biologic Science and Technology, National Chiao Tung University, Hsin-Chu 300, Taiwan. 4. Department of Obstetrics and Gynecology, Hsinchu Mackay Memorial Hospital, Hsinchu, Taiwan. 5. Mackay Medicine, Nursing and Management College, Taipei, Taiwan. 6. Department of Medicine, Mackay Medical College, New Taipei City, Taiwan.

Received: 31 May 2013 Accepted: 13 December 2013
Published: 13 January 2014

#### References

1. FAO/WHO: Health and nutritional properties of probiotics in food including powdered milk with live lactic acid bacteria. In: Book Health and nutritional properties of probiotics in food including powdered milk with live lactic acid bacteria. http://www.who.int/foodsafety/publications/fs_management/en/probiotics.pdf City, 2001.
2. Kalionomaki M, Salminen S, Arvilommi H, Kero P, Koskinen P, Isolauri E: Probiotics: in vivo adherence properties of Lactobacillus casei strain Shirota on Shiga toxin-producing Escherichia coli. Int J Food Microbiol 2001, 73:444S–450S.
3. Gopal PK, Prasad J, Smart J, Gill HS: In vitro adherence properties of lactobacillus rhamnosus DR20 and bifidobacterium lactis DR10 strains and their antagonistic activity against an enterotoxigenic Escherichia coli. Int J Food Microbiol 2001, 67:207–216.
4. Ogawa M, Shimizu K, Nomoto K, Takahashi M, Watanuki M, Tanaka R, Tanaka T, Hamabata T, Yamashita S, Takeda Y: Protective effect of Lactobacillus casei strain Shirota on Shiga toxin-producing Escherichia coli O157:H7 infection in infant rabbits. Infect Immun 2001, 69:1101–1108.
5. Reid G: Probiotic agents to protect the urogenital tract against infection. Am J Clin Nutr 2001, 73:437S–443S.
6. Wan YM, Zhu YQ, Xia B, Luo J: Treating TNBS-induced colitis in rats with probiotics. Turk J Gastroenterol 2011, 22:486–493.
8. Taipale T, Penihakinnen K, Isolauri E, Larsen C, Brodbeck E, Alanen P, Jokela J, Soderling E: Bifidobacterium animalis subsp. lactis BB-12 in reducing the risk of infections in infancy. Br J Nutr 2011, 105:409–416.

9. Wickens K, Black PN, Stanley TV, Mitchell E, Fitzharis P, Tannock GW, Purdie G, Crane J: A differential effect of 2 probiotics in the prevention of eczema and atopy: a double-blind, randomized, placebo-controlled trial. J Allergy Clin Immunol 2008, 122:788–794.

10. Jordan JA, Durso MB: Real-time polymerase chain reaction for detecting bacterial DNA directly from blood of neonates being evaluated for sepsis. J Mol Diagn 2005, 7:575–581.

11. Wilson M: Bacteriology of Humans: An Ecological Perspective. Hoboken: Wiley-Blackwell; 2008.

12. Goldenberger D, Kurazi A, Vogt F, Zbinden R, Altwegg M: Molecular diagnosis of bacterial endocarditis by broad-range PCR amplification and direct sequencing. J Clin Microbiol 1997, 35:2733–2739.

13. Woo PC, Lau SK, Woo GK, Ngan AH, Hui WT, Yuen KY: A differential effect of antibiotic growth promoters on intestinal microbiota in food animals: a novel model for studying the relationship between gut microbiota and human obesity? Front Microbiol 2011, 2:53.

14. Hegerle N, Guiski N: Epidemiology of whooping cough and typing of Bordetella pertussis. Future Microbiol 2013, 8(11):1391–1403.

15. Lekholm UN, Boyle SM, Sirananthan N: Brucella: a pathogen without classic virulence genes. Vet Microbiol 2008, 129:1–14.

16. Zhan P, Luo L, Qian Q, Shen XK, Qiu LY, Yu, Song Y: Chlamydia pneumoniae infection and lung cancer risk: a meta-analysis. Eur J Cancer Prev 2011, 20:743–747.

17. Pankhill J, Wren BW, Mungall K, Ketey JH, Churcher C, Basham D, Chillingworth T, Davies RM, Feltwell T, Holroyd S, et al: The genome sequence of the food-borne pathogen campylobacter jejuni reveals hypervariable sequences. Nature 2000, 403:665–668.

18. Guidoboni M, Ferrier AI, Porzoni M, Dogliolo C, Pelosi F: Infectious agents in mucosa-associated lymphoid tissue-type lymphomas: pathogenic role and therapeutic perspectives. Clin Lymphoma Myeloma 2006, 6:289–300.

19. Halley RA, Changpe PC, Fisher BD: Uppington's Illustrated Reviews: Microbiology. 2nd edition. Philadelphia: Lippincott Williams & Wilkins; 2007.

20. Fredlund H, Falk L, Junstrand M, Unemo M: Molecular genetic methods for diagnosis and characterisation of Chlamydia trachomatis and Neisseria gonorrhoeae: impact on epidemiological surveillance and interventions. APMIS 2004, 112:717–784.

21. Berdichevski T, Barsbick i, Bar-Meir S, Ben-Horin S: Pseudomonas aeruginosa in a patient with flare-up of inflammatory bowel disease (IBD): is it old or new? Curr Infect Dis Rev 2010, 12(Suppl 2):E131–E137.

22. Setterfield BA, Stewart AF, Lew CS, Pickett DO, Cohen MN, Moore EA, Luedtke PF, O'Neill KL, Robison RA: A quadruplex real-time PCR assay for rapid detection and differentiation of the Clostridium botulinum toxin genes A, B, E and F. J Med Microbiol 2010, 59:65–66.

23. Brook I: The role of anaerobic bacteria in cutaneous and soft tissue abscesses and infected cysts. Anoraoite 2007, 13:171–177.

24. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL: Greengenes, a chimera-checked 16S RNA gene database and workbench compatible with ARB. Appl Environ Microbiol 2006, 72:5269–5272.

25. quart C, Prune A, Eymaz P, Gerken J, Schwerte T, Yarza P, Peplies J, Goddke FO: The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 2013, 41:D590–D596.

26. Wang WC, Lin FM, Chang WC, Lin KY, Huang HD, Lin NS: miRExpress: analyzing high-throughput sequencing data for profiling microRNA expression. BMC Bioinformatics 2009, 10:328.

27. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S: The sequence alignment/map format and SAMtools. Bioinformatics 2009, 25:2078–2079.

28. Chmielewska A, Szajewska H: Systematic review of randomised controlled trials: probiotics for functional constipation. World J Gastroenterol 2010, 16:659–75.

29. Chassard C, Dagoigny M, Scott KP, Crouzet L, Delhomme C, Marquet P, Martin JC, Pickering G, Ardid D, Eschaler A, et al: Functional dysbiosis within the gut microbiota of patients with constipated-irritable bowel syndrome. Aliment Pharmacol Ther 2012, 35:282–383.

30. Schwartz A, Tanas D, Schafer K, Beijer S, Bos NA, Donus C, Hardt PD: Microbiota and SCFA in lean and overweight healthy subjects. Obesity (Silver Spring) 2010, 18:190–195.

31. Balabanoglu R, George G, Kabeerdoss J, Hepkett J, Chandragnasekan AM, Ramakrishna BS: Quantitative differences in intestinal Faecalibacterium prausnitzii in obese Indian children. Br J Nutr 2010, 103:335–338.

32. Million M, Maraninchi M, Hemy N, Armoogum F, Richet H, Carrier P, Valero R, Raccah D, Valettes B, Raoult D: Obesity-associated gut
microbiota is enriched in lactobacillus reuteri and depleted in Bifidobacterium animalis and methanobrevibacter smithii. Int J Obes (Lond) 2012, 36:817–825.

52. Armougom F, Henry M, Valettes B, Raccah D, Raoult D: Monitoring bacterial community of human gut microbiota reveals an increase in lactobacillus in obese patients and methanogens in anorexic patients. PloS one 2009, 4:e7125.

53. Noor SQ, Ridgway K, Scovell L, Kemsley EK, Lund EK, Jamieson C, Johnson IT, Narbad A: Ulcerative colitis and irritable bowel patients exhibit distinct abnormalities of the gut microbiota. BMC Gastroenterol 2010, 10:134.

54. Carroll IM, Ringel-Kulka T, Siddle JP, Ringel Y: Alterations in composition and diversity of the intestinal microbiota in patients with diarrhea-predominant irritable bowel syndrome. Neurogastroenterol Motil 2012, 24:e521–530. e248.

55. Duboc H, Rainteau D, Rajca S, Humbert L, Farabos D, Maubert M, Grondin V, Jouet P, Bouhassira D, Seksik P, et al: Increase in fecal primary bile acids and dysbiosis in patients with diarrhea-predominant irritable bowel syndrome. Neurogastroenterol Motil 2012, 24:e513–520. e246-517.

56. Kerckhoffs AP, Ben-Amor K, Samsom M, van der Rest ME, de Vogel J, Knol J, Akkermans LM: Molecular analysis of faecal and duodenal samples reveals significantly higher prevalence and numbers of Pseudomonas aeruginosa in irritable bowel syndrome. J Med Microbiol 2011, 60:236–245.

57. Saulnier DM, Riehle K, Mistretta TA, Diaz MA, Mandal D, Raza S, Weidler EM, Qin X, Coarfa C, Milosavljevic A, et al: Gastrointestinal microbiome signatures of pediatric patients with irritable bowel syndrome. J Gastroenterol 2011, 46:479–486.

58. Rajilic-Stojanovic M, Biagi E, Heilig HG, Kajander K, Kekkonen RA, Tims S, de Vos WM: Global and deep molecular analysis of microbiota signatures in fecal samples from patients with irritable bowel syndrome. Gastroenterology 2011, 141:1792–1801.

59. Sokol H, Seksik P, Furet JP, Firmesse O, Nion-Larmurier I, Beauperey L, Cosnes J, Corthier G, Marteau P, Dore J: Low counts of faecalibacterium prausnitzii in colitis microbiota. Inflamm Bowel Dis 2009, 15:1183–1189.

60. Andoh A, Imaeda H, Aomatsu T, Inatomi O, Bamba S, Sasaki M, Saito Y, Tsujikawa T, Fujimura Y: Comparison of the fecal microbiota profiles between ulcerative colitis and Crohn’s disease using terminal restriction fragment length polymorphism analysis. J Gastroenterol 2011, 46:479–486.

61. Sokol H, Lepage P, Seksik P, Dore J, Marteau P: Temperature gradient gel electrophoresis of fecal 16S rRNA reveals active Escherichia coli in the microbiota of patients with ulcerative colitis. J Clin Microbiol 2006, 44:3172–3177.

62. Wang T, Cai G, Qiu Y, Fei N, Zhang M, Pang X, Jia W, Cai S, Zhao L: Structural segregation of gut microbiota between colorectal cancer patients and healthy volunteers. ISME J 2012, 6:225–229.

63. Moore WE, Moore LH: Intestinal flora of populations that have a high risk of colon cancer. Appl Environ Microbiol 1995, 61:3202–3207.

64. Shen XJ, Rawls JF, Randall T, Burcal L, Mpande CN, Jenkins N, Jovov B, Abdo J, Sandler RS, Keku TO: Molecular characterization of mucosal adherent bacteria and associations with colorectal adenomas. Gut Microbes 2010, 1:138–147.

65. McGill R, Tukey JW, Larsen WA: Variations of box plots. Am Stat 1978, 32:12–16.