BMC
Microbiology

RESEARCH ARTICLE

Open Access

A multi-omic analysis of an Enterococcus faecium mutant reveals specific genetic mutations and dramatic changes in mRNA and protein expression

De Chang[1], Yuanfang Zhu[2], Li An[1], Jinwen Liu[2], Longxiang Su[1], Yinghua Guo[1], Zhenhong Chen[1], Yajuan Wang[1], Li Wang[1], Junfeng Wang[1], Tianzhi Li[1], Xiangqun Fang[1], Chengxiang Fang[2], Ruifu Yang[2,3*] and Changting Liu[1*]

Abstract

Background: For a long time, _Enterococcus faecium_ was considered a harmless commensal of the mammalian gastrointestinal (GI) tract and was used as a probiotic in fermented foods. In recent decades, _E. faecium_ has been recognised as an opportunistic pathogen that causes diseases such as neonatal meningitis, urinary tract infections, bacteremia, bacterial endocarditis and diverticulitis. _E. faecium_ could be taken into space with astronauts and exposed to the space environment. Thus, it is necessary to observe the phenotypic and molecular changes of _E. faecium_ after spaceflight.

Results: An _E. faecium_ mutant with biochemical features that are different from those of the wild-type strain was obtained from subculture after flight on the SHENZHOU-8 spacecraft. To understand the underlying mechanism causing these changes, the whole genomes of both the mutant and the WT strains were sequenced using Illumina technology. The genomic comparison revealed that _dprA_, a recombination-mediator gene, and _arpU_, a gene associated with cell wall growth, were mutated. Comparative transcriptomic and proteomic analyses showed that differentially expressed genes or proteins were involved with replication, recombination, repair, cell wall biogenesis, glycometabolism, lipid metabolism, amino acid metabolism, predicted general function and energy production/conversion.

Conclusion: This study analysed the comprehensive genomic, transcriptomic and proteomic changes of an _E. faecium_ mutant from subcultures that were loaded on the SHENZHOU-8 spacecraft. The implications of these gene mutations and expression changes and their underlying mechanisms should be investigated in the future. We hope that the current exploration of multiple "-omics" analyses of this _E. faecium_ mutant will provide clues for future studies on this opportunistic pathogen.

Keywords: _E. faecium_, Genome, Transcriptome, Proteome, Multi-omics

Background

In the past, _E. faecium_ was considered to be a harmless commensal of the mammalian GI tract and was used as a probiotic in fermented foods [1,2]. In recent decades, _E. faecium_ has been recognised as an opportunistic pathogen that causes diseases such as neonatal meningitis, urinary tract infections, bacteremia, bacterial endocarditis and diverticulitis [3-7]. Therefore, _E. faecium_ can penetrate and survive in many environments in the human body, which could potentially lead to unpredictable consequences.

Due to revolutionary advances in high-throughput DNA sequencing technologies [8] and computer-based genetic analyses, genome decoding and transcriptome sequencing (RNA-seq) [9,10] analyses are rapid and available at low costs. Moreover, the development of mass spectrometry-based proteomic analysis provides a simple and convenient approach to identify and quantify thousands of proteins in a single experiment [11,12]. By employing these high-throughput technologies, the mechanisms underlying the systematic changes of a mutant and wild-type microbe could be revealed. Here we employed multi-

* Correspondence: ruifuyang@gmail.com; liuchangt@foxmail.com
[2]BGI-Shenzhen,, Shenzhen, People's Republic of China
[1]Nanlou Respiratory Diseases Department, Chinese PLA General Hospital, Beijing 100853, China
Full list of author information is available at the end of the article

omic technologies, including genomic, transcriptomic and proteomic analysis of a mutant strain of *E. faecium* and the corresponding wild-type strain to understand the complex mechanisms behind the mutations resulting in altered biochemical metabolic features.

Methods

Acquisition of the mutant

The *E. faecium* strain that was loaded in the SHENZHO U-8 spacecraft as a stab culture was obtained from the Chinese General Microbiological Culture Collection Center (CGMCC) as CGMCC 1.2136. After spaceflight from Nov. 1st to 17th, 2011, the *E. faecium* sample was struck out and grown on solid agar with nutrients. Then, 108 separate colonies were picked randomly and screened using the 96 GEN III MicroPlate™ (Biolog, USA). The ground strain LCT-EF90 was used as the control. With the exception of spaceflight, all other culture conditions were identical between the two groups. The majority of selected subcultures showed no differences in the biochemical assays except for strain LCT-EF258. Compared with the control strain, a variety of the biochemical features of LCT-EF258 had changed after a 17-day flight in space. Based on the Biolog colour changes, strain LCT-EF258 had differences in utilisation patterns of N-acetyl-D-galactosamine, L-rhamnose, myo-inositol, L-serine, L-galactonic acid, D-gluconic acid, glucuronamide, p-hydroxy- phenylacetic acid, D-lactic acid, citric acid, L-malic acid and γ-amino-butryric acid relative to the control strain LCT-EF90 (Table 1). Despite isolation of this mutant, we could not determine if the underlying mutations were caused by the spaceflight environment. However, the mutant's tremendous metabolic pattern changes still drew our interest to uncover possible genomic, transcriptomic and proteomic differences and to further understand the mechanisms underlying these differences.

DNA, RNA and protein preparation

Both the mutant and the control strains were grown in Luria-Bertani (LB) medium at 37°C; genomic DNA was prepared by conventional phenol-chloroform extraction methods; RNAs were exacted using TIANGEN RNAprep pure Kit (Beijing, China) according to the manufacturer's instructions. Protein was extracted and quantified and was subsequently analysed by SDS-polyacrylamide gel electrophoretogram. After digestion with trypsin, the samples were labelled using the iTRAQ reagents (Applied Biosystems), which fractionates the proteins using strong cationic exchange (SCX) chromatography (Shimadzu). Each fraction was separated using a splitless nanoACQuity (Waters) system coupled to the Triple TOF 5600 System (AB SCIEX, Concord, ON).

Table 1 Phenotypic characteristics of the mutant (LCT-EF258) and the control strain (LCT-EF90) used in this study

Features	LCT-EF90	LCT-EF258
N-acetyl-D-galactosamine	-	+/−
L-rhamnose	-	+/−
Myo-inositol	-	+/−
L-serine	+/−	-
L-galactonic	-	+/−
D-gluconic acid	+/−	-
Glucuronamide	+/−	-
p-hydroxy- phenylacetic acid	+	-
D-lactic acid	-	+/−
Citric acid	+/−	-
L-malic acid	-	+
γ-amino-butryric acid	-	+

Note: " + " represents a significantly positive reaction; "+/−" represents a slightly positive reaction; "-" represents a negative reaction.

Genome sequencing and annotation

Sequencing and filtering

Using genomic DNA from the two samples, we constructed short (500 bp) and large (6 kb) random sequencing libraries and selected 90-bp read lengths for both libraries. Raw data were generated from the Illumina Hiseq2000 next-generation sequencing (NGS) platform with Illumina 1.5 format encoding a Phred quality score from 2 to 62 using ASCII 66 to 126. The raw data were then filtered through four steps, including removing reads with 5 bp of Ns' base numbers, removing reads with 20 bp of low quality (≤Q20) base numbers, removing adapter contamination, and removing duplication reads. Finally, a total of 55 million base pairs of reads were generated to reach a depth of ~190-fold of total genome coverage.

Repetitive sequences analysis

We searched the genome for tandem repeats using Tandem Repeats Finder [13] and Repbase [14] (composed of many transposable elements) to identify the interspersed repeats. Transposable elements in the genome assembly were identified both at the DNA and protein level. For identification of transposable elements at the DNA level, RepeatMasker [15] was applied using a custom library comprising a combination of Repbase. At the protein level, RepeatProteinMask, which is updated software in the RepeatMasker package, was used to perform RM-BlastX against the transposable elements protein database.

ncRNA sequences analysis

The tRNA genes were predicted by tRNAscan [16]. Aligning the rRNA template sequences from animals using BlastN with an E-value of 1e-5 identified the rRNA

fragments. The miRNA and snRNA genes were predicted by INFERNAL software [17] against the Rfam database [18].

Gene functional annotation
To ensure the biological meaning, we chose the highest quality alignment result to annotate the genes. We used BLAST to accomplish functional annotation in combination with different databases. We provided BLAST results in m8 format and produced the annotation results by alignment with selected databases.

Nucleotide sequence accession number
The whole-genome sequences of the wild-type and mutant *E. faecium* strains in this study have been deposited at DDBJ/EMBL/GenBank under the accession numbers ANAJ00000000 and ANAI00000000, respectively.

Comparative genomic analysis
SNPs calling
Raw SNPs were identified using software MUMmer (Version 3.22) [19] and SOAPaligner (Version 2.21). In all, raw SNPs were filtered by the following criteria: SNPs with quality scores < 20, SNPs covered by < 10 paired-end reads, SNPs within 5 bp on the edge of reads, and SNPs within 5 bp of two or more existing mutations. Finally, SNPs in repetitive regions found using the "Repetitive sequences analysis" method were also filtered.

Small size InDel variants calling
First, InDels (insertions and deletions) with lengths of less than 10 bp were extracted from the gap extension alignment between the genome assembly and the reference using LASTZ (Version 1.01.50). Second, we removed the unreliable InDels containing N base within 50 bp upstream and downstream, and we removed InDels with more than two mismatches within a total of 20 bp upstream and downstream. Finally, the candidate InDels were verified by comparing sample reads to the surrounding region of the InDels (100 bp each side) with the reference sequence by using BWA (Version 0.5.8) [20].

Synteny analysis
The LCT-EF258 target sequences were ordered according to the reference sequence based on MUMmer. Then, the X and Y axes of the two-dimensional synteny graphs and the upper and following axes of linear syntenic graphs were constructed after the same proportion of size reduction in the length of both sequences. The protein set P1 of the target sequence was aligned with the protein set P2 of the reference sequence using BLASTP (e-value < = 1e-5, identity > = 85%, and the best hit of each protein was selected). Finally, the results with the best-hit value

were reserved and the average of two consistent values was obtained.

Transcriptome sequencing and comparison
Sequencing and filtering
Total RNAs were purified using TRIzol (Invitrogen) and rRNA was removed. Then, cDNA synthesis was performed with random hexamers and Superscript II reverse transcriptase (Invitrogen). Meanwhile, double-stranded cDNAs were purified with a Qiaquick PCR purification kit (Qiagen) and sheared with a nebuliser (Invitrogen) to ~200 bp fragments. After end repair and poly (A) addition, the cDNAs were ligated to Illumina N-acetyl-D-galactosamine (pair end) adapter oligo mix and suitable fragments were selected as templates by gel purification. Next, the libraries were PCR amplified and were sequenced using the Illumina Hiseq 2000 platform and the paired-end sequencing module.

The filtration consisted of three steps: removing reads with 1 bp of Ns' base numbers, removing reads with 40 bp of low quality (≤Q20) base numbers, and removing adapter contamination. Additionally, reads mapped to the reference (LCT-EF90) rRNA sequences were removed. All gene expression data generated in this study have been deposited under accession numbers SRR922447 and SRR922448 (https://trace.ddbj.nig.ac.jp/DRASearch/).

Gene expression value statistics
The gene coverage was evaluated by mapping clean reads to the reference genes using SOAPaligner software, and the gene expression value was calculated by the RPKM (Reads Per kb per Million reads) formula based on the method described in Ali et al. [21]. The RPKM method was able to eliminate the influence of gene length and sequencing discrepancy on the gene expression calculation. Therefore, the calculated gene expression could be directly used for comparing the gene expression among difference samples.

Differential gene expression analysis
To control error rate and identify true differentially expressed genes (DEGs), the p-value was rectified using the FDR (False Discovery Rate) control method [22]. Both the FDR value and the RPKM ratio in different samples were calculated. Finally, genes with an RPKM ratio ≥ 2 and a FDR ≤ 0.001 between different samples were defined as DEGs. Different DEGs were enriched and clustered according to the GO and KEGG functions.

Proteomic study
Quantitative proteomics were performed using iTRAQ technology coupled with 2D-nanoLC-nano-ESI-MS/MS to examine the difference of protein profiles [23]. After identification by the TripleTOF 5600 System, data acquisition

was performed with a TripleTOF 5600 System (AB SCIEX, Concord, ON) fitted with a Nanospray III source (AB SCIEX, Concord, ON) with a pulled quartz tip as the emitter (New Objectives, Woburn, MA). Data analysis, including protein identification and relative quantification, were performed with the ProteinPilotTM software 4.0.8085 using the Paragon Algorithm version 4.0.0.0 as the search engine. Each MS/MS spectrum was searched against the genome annotation database (5263 protein sequences), and the search parameters allowed for Cys. The local FDR was set to 5%, and all identified proteins were grouped by the ProGroup algorithm (ABI) to minimise redundancy. Proteins were identified based on at least one peptide with a percent confidence above 95%. Some of the identified peptides were excluded according to the following conditions: (i) Peptides with low ID confidence (<15%) were excluded. (ii) Peptide peaks corresponding to the ITRAQ labels were not observed. (iii) Shared MS/MS spectra, due to either identical peptide sequences in more than one protein or when more than one peptide was fragmented simultaneously, were excluded. (iv) Any peptide ratio in which the S/N (signal-to-noise ratio) is too low was excluded. Several quantitative estimates provided for each protein by the Protein Pilot were utilised, including the fold change ratios of differential expression between labelled protein extracts and the P value, which represents the probability that the observed ratio is different to 1 by chance. All experiments were performed in three replicates, and the differentially expression proteins (DEPs) were selected if they appeared at least twice and the fold change was larger than 1.2 with a p-value less than 0.05. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository with the dataset identifier PXD000326.

Bioinformatics analysis
Gene ontology and GO enrichment analysis
GO (Gene Ontology) enrichment analysis provided all GO terms that were significantly enriched in a list of DEGs, and the DEGs were filtered corresponding to specific biological functions. We first mapped all DEGs to GO terms in the database, calculating gene numbers for every term, and then used the hypergeometric test to find significantly enriched GO terms based on GO::TermFinder [24]. Here, a strict algorithm was developed for the analysis:

$$P = 1 - \sum_{i=0}^{m-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

where N was the number of all genes with GO annotation; n was the number of DEGs in N; M was the number of all genes that were annotated to certain GO terms; m was the

number of DEGs in M. The calculated p-value required a corrected p-value ≤ 0.05 as a threshold by Bonferroni correction.

Pathway analysis and pathway enrichment analysis
Gene interactions play key roles in many biological functions. Pathway enrichment of DEGs was analysed by the KEGG pathway [25]. This analysis identified significantly enriched metabolic pathways in DEGs when compared with the genome background. The same analysis utilized in the GO enrichment was used for the pathway enrichment analysis. Here, N was the number of all genes with KEGG annotation, n was the number of DEGs in N, M was the number of all genes annotated to specific pathways, and m was the number of DEGs in M.

COG function analysis
Cluster of Orthologous Groups of proteins (COG) is the database for gene/protein orthologous classification (http://www.ncbi.nlm.nih.gov/COG/). Every gene/protein in a COG is supposed to be derived from a single gene/protein ancestor. Orthologs are gene/proteins derived from different species of one vertical family and have the same functions as the ancestor. Paralogs are proteins derived from gene expression and may have new, related functions. We compared identified proteins with the COG database to predict the gene or proteins' function.

Results
Genomic sequencing, assembly and annotation
Genomic DNA from both samples was sequenced using a whole-genome shotgun sequencing (WGS) approach on the Illumina Hiseq2000 system. The short (500 bp) and large (6 kb) random sequencing libraries were constructed, and the mean read length was 90 bp for both libraries. A total of 55 million base pairs of reads were generated to reach a depth of ~190-fold genome coverage (see Methods for details). The genomes were assembled using SOAPdenovo (Version 1.05) [26], which resulted in the final high quality genomic assemblies.
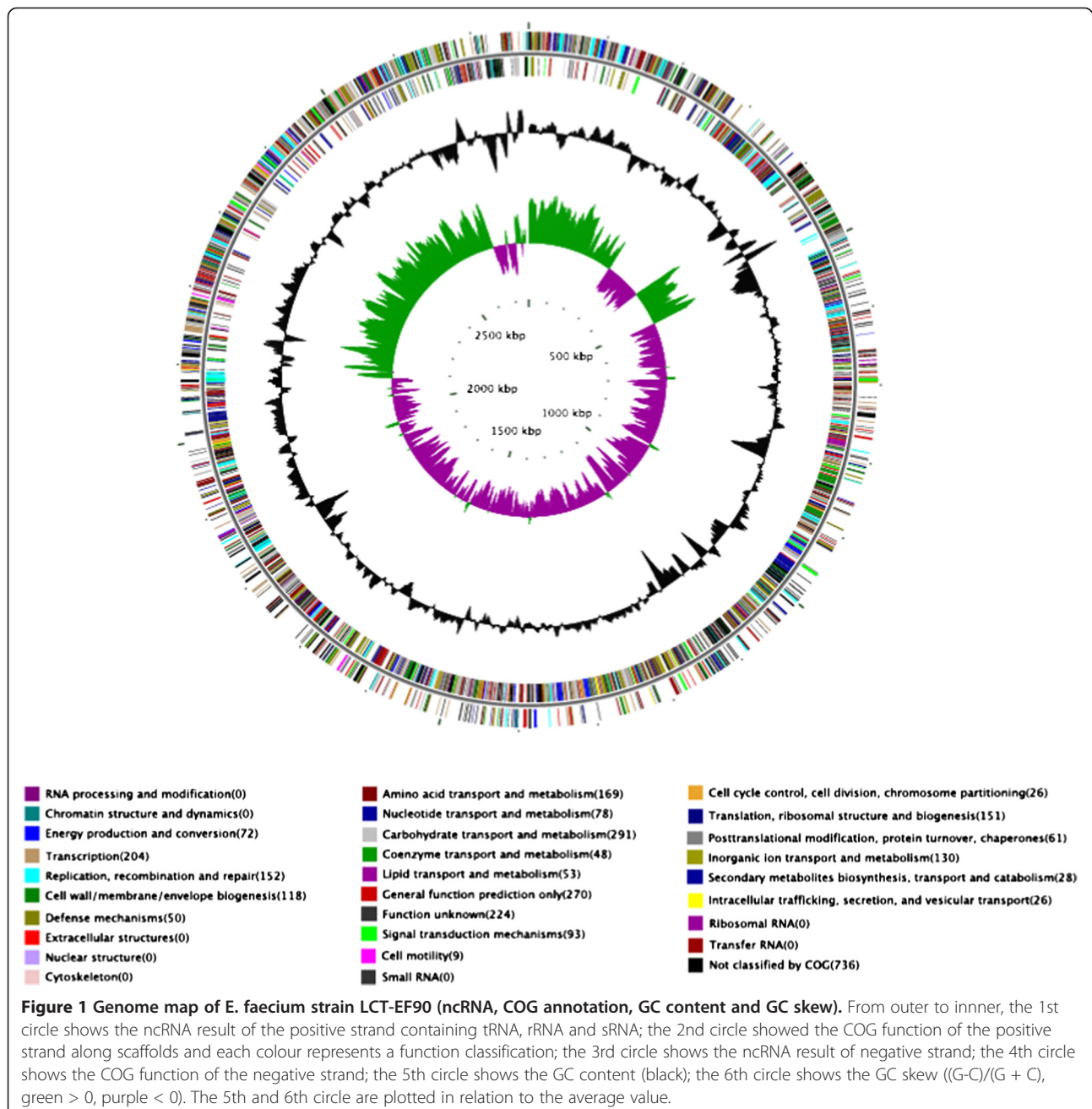
Before the comparative genomics analysis, gene models and their associated functions for strain LCT-EF90 were determined using different databases. First, we used Glimmer software [27] for gene prediction and identified 2,777 genes with a total length of 2,394,186 bp, which consisted of 86.31% of the genome. In addition, 13,090 bp of the transposon sequences and 4,787 bp of the tandem repeat sequences were identified, which consisted of 0.47% and 0.17% of genome, respectively (Additional file 1: Table S1). We identified 37 tRNA fragments with a total length of 2,807 bp and 2 snRNA (small nuclear RNA) genes with a total length of 367 bp (see Methods for details). We annotated all of the genes against the popular functional databases, including 59.60% of the genes into the GO

database (Additional file 1: Figure S1) [28], 73.50% of the genes into COG (Additional file 1: Figure S2) [29], 66.69% of the genes into KEGG (Additional file 1: Figure S3) [25], 97.34% of the genes into the NR database, 69.07% genes into SwissProt [30] and 97.34% of the genes into TrEMBL [31] (see Methods for details). Moreover, 321 genes were identified in the CAZY (Carbohydrate-Active enzymes) database [32], 210 genes in the PHI-base (Pathogen - Host Interaction) database [33], 6 genes in DBETH (a Database of Bacterial Exotoxins for Human) [34] and 387 genes in

VFDB (Virulence Factors Database) [35]. In addition, our analysis predicted genome islands, prophages and CRISPRs (Clustered Regularly Interspaced Short Palindromic Repeats), but no CRISPRs have been found. The genome map of E. faecium strain LCT-EF90 was shown in Figure 1.

Comparative genomic analysis

We used LCT-EF90 as the reference strain and detected variations, including SNPs, InDels and structure variations

Figure 1 Genome map of E. faecium strain LCT-EF90 (ncRNA, COG annotation, GC content and GC skew). From outer to innner, the 1st circle shows the ncRNA result of the positive strand containing tRNA, rRNA and sRNA; the 2nd circle showed the COG function of the positive strand along scaffolds and each colour represents a function classification; the 3rd circle shows the ncRNA result of negative strand; the 4th circle shows the COG function of the negative strand; the 5th circle shows the GC content (black); the 6th circle shows the GC skew ((G-C)/(G + C), green > 0, purple < 0). The 5th and 6th circle are plotted in relation to the average value.

(SVs) between LCT-EF258 and LCT-EF90 (Figure 2). For SNP identification, the query sequence was aligned with the reference sequence using MUMmer software (Version 3.22) [36] (see Methods for details). The raw variation sites were identified and then filtered with strict standards to detect potential SNP sites. Finally, 1 SNP for *E. faecium* LCT-EF258 was detected and was located in the functional gene LCT-EF90GL001983 (Additional file 1: Table S2). The SNP mutation in LCT-EF90GL001983 was a non-synonymous substitution in *dprA*, a gene encoding a DNA processing protein based on KEGG pathway analysis, and may play an important role in phenotypic variation.

To detect more variations, we used the LASTZ (Version 1.01.50) tool to identify InDels less than or equal to 10 bp (see Methods for details). After a series of filtering conditions, we have found 8 InDels between LCT-EF90 and LCT-EF258 (Additional file 1: Table S3), including 7 InDels in intergenic regions and only one in a coding region. The coding region InDel was identified in LCT-EF90GL000008, which is annotated as an *arpU* family gene related to transcriptional regulators in the NR database (Additional file 1: Table S4) but not in VFDB (Virulence Factors Database). While small size InDels were found in sample LCT-EF258, we were also interested in large scale structural variations. We aligned the two samples with a reference at the nucleic acid level (see Methods for details) but did not identify any large scale SVs. The probable reason may be that the generation time was so short that the variations did not have enough time to accumulate.

Transcriptomic analysis

Using gene difference expression analysis, 2,679 genes between LCT-EF90 and LCT-EF258 were detected. After filtering conditions of FDR ≤ 0.001 and RPKM Ratio ≥ 2, 1,159 genes remained. Both up-regulated and down-regulated genes were identified in this analysis. Approximately 123 genes were up-regulated, and 1,036 genes were down-regulated between LCT-EF90 and LCT-EF258 (Figure 3A). We found that the down-regulated genes significantly outnumbered up-regulated genes, suggesting that gene expression and metabolism were inhibited in LCT-EF258.

Different DEGs were enriched and clustered according to GO, COG and KEGG analyses. For COG, the up-regulated and down-regulated genes were summed and were compared with unchanged genes. The most change was annotated into the translation, ribosomal structure and biogenesis function classes (Figure 3B). For gene ontology, the DEGs that showed statistical significance (P-value ≤0.05) were the component, function and process ontologies. For LCT-EF90 and LCT-EF258, seven categories, including 601 DEGs (identical DEGs may fall into different categories), were shown to be meaningful (Figure 3C). For the KEGG functional cluster, there were eleven categories, including 283 DEGs, between LCT-EF90 and LCT-EF258. Most of the genes were annotated into three categories: purine metabolism, pyrimidine metabolism and ribosome (Figure 3D).

Comparative proteomic analysis

Using Protein Pilot software, 1188 proteins that appeared at least twice in three replicates were identified [37]. Relatively quantitative analysis shows that 213 DEPs were
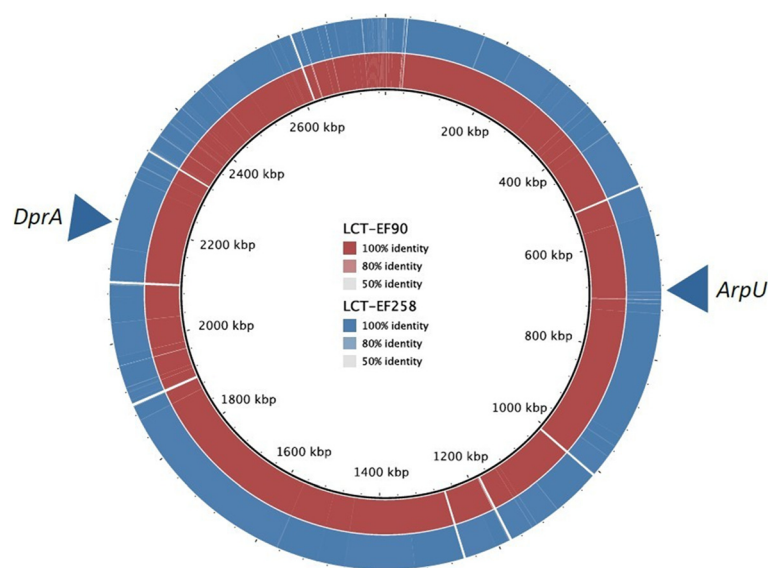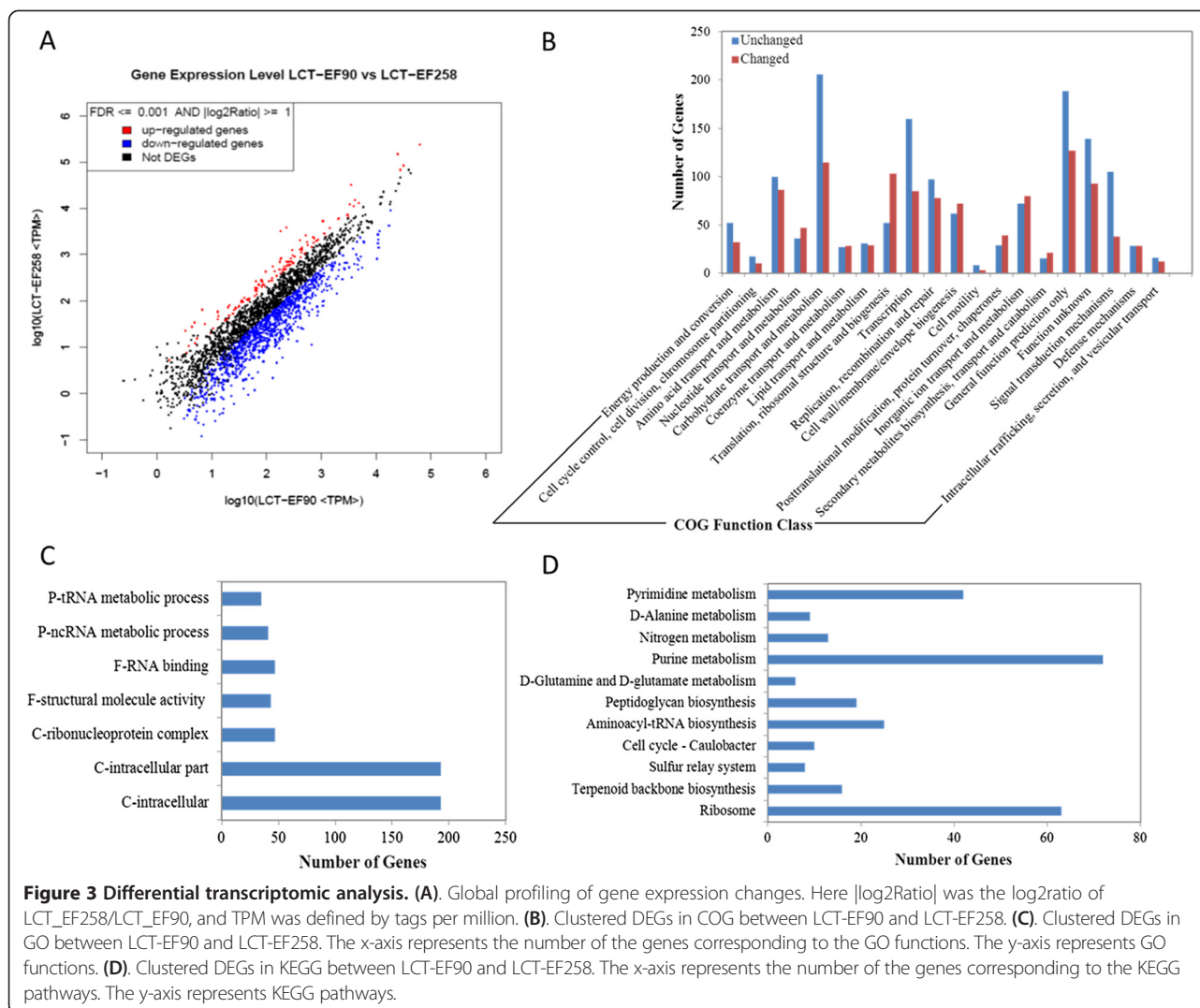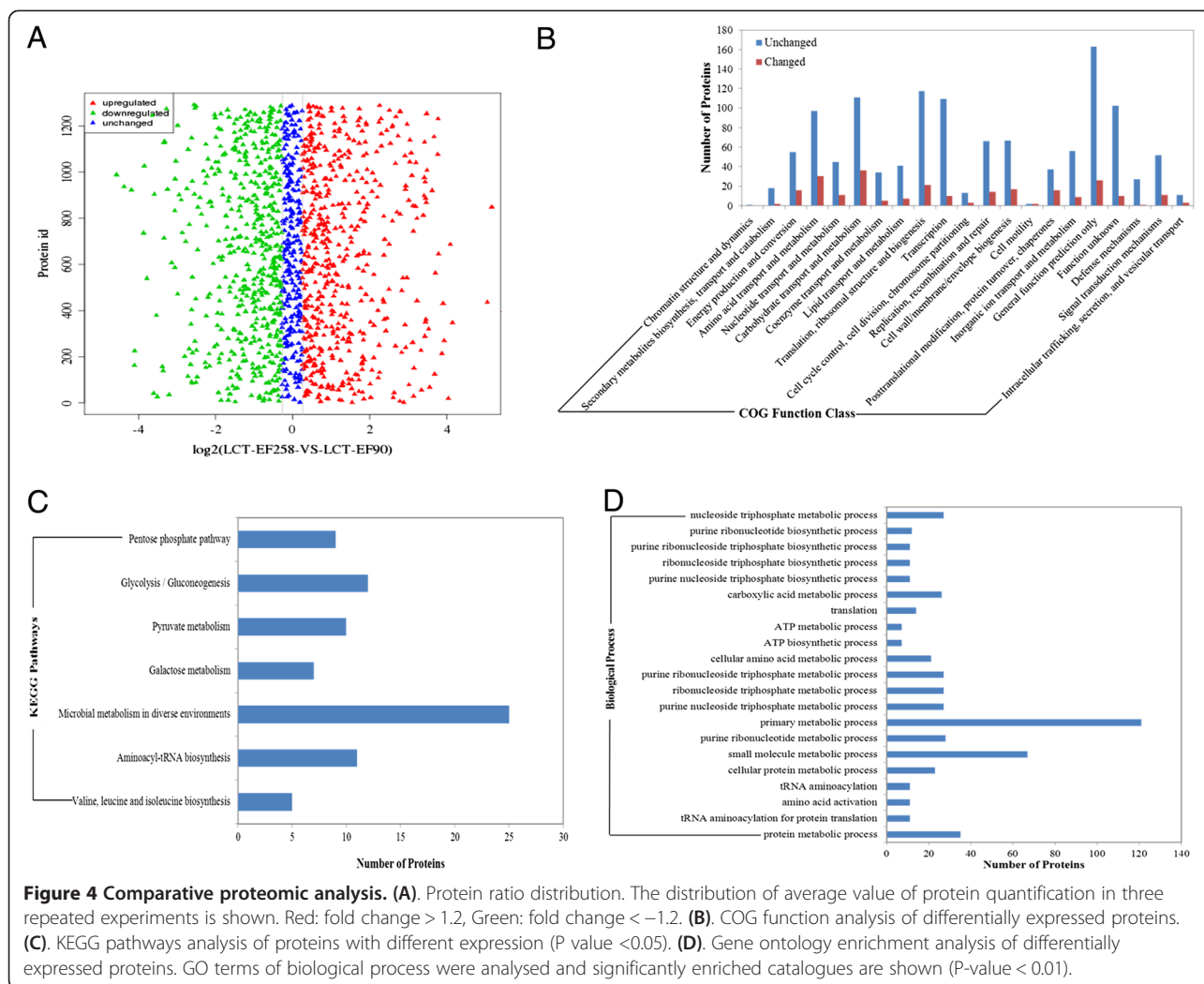

Figure 2 Comparative genomic analysis. We used BRIG software to achieve alignment results of three genomes. The gray circle is LCT-EF90, and blue circle is LCT-EF258. There are some white regions in two circles, which are the gaps in genomes. The triangles indicate the general positions of the mutations with SNPs and InDels, which were annotated into genes *dprA* and *arpU*.

Figure 3 Differential transcriptomic analysis. (A). Global profiling of gene expression changes. Here |log2Ratio| was the log2ratio of LCT_EF258/LCT_EF90, and TPM was defined by tags per million. **(B)**. Clustered DEGs in COG between LCT-EF90 and LCT-EF258. **(C)**. Clustered DEGs in GO between LCT-EF90 and LCT-EF258. The x-axis represents the number of the genes corresponding to the GO functions. The y-axis represents GO functions. **(D)**. Clustered DEGs in KEGG between LCT-EF90 and LCT-EF258. The x-axis represents the number of the genes corresponding to the KEGG pathways. The y-axis represents KEGG pathways.

identified, including 116 down-regulated proteins and 97 up-regulated proteins (Figure 4A). Subsequently, DEPs were classified according to COG function category. It is clear that the expression of proteins involved in functions such as energy production, metabolism, transcription, translation, posttranslational modification, DNA recombination and repair, cell wall biogenesis and signal transduction mechanisms changed the most (Figure 4B). The enrichment and cluster of DEPs were performed according to Gene Ontology and KEGG Pathways functional analysis. The metabolic and biosynthetic biological processes were found to be different in the mutant (Figure 4C). As to KEGG functions affected in the mutant, significant difference was found in the following pathways: valine, leucine and isoleucine biosynthesis; aminoacyl-tRNA biosynthesis; pyruvate metabolism; galactose metabolism; glycolysis; pentose phosphate pathway; and microbial metabolism in diverse environments (Figure 4D).

Integration of transcriptomic and proteomic analysis

Most previous studies suggest a weak correlation between mRNA expression and protein expression, which may be due to post-transcriptional regulation of protein synthesis, post-translational modification or experimental errors [38-40]. However, according to the central dogma of molecular genetics, genetic information is transmitted from DNA to message RNAs that are subsequently translated to proteins [41,42]. Thus, we integrated the DEFs and DEPs to identify the overlapping genes that are expressed differently in both the transcriptome and the proteome. One-hundred and two genes were selected (Figure 5A), and those genes with either up-regulated or down-regulated expression at both the mRNA and protein levels were subjected to bioinformatic analysis. The Gene Ontology study indicated that biological processes such as metabolic processes, catabolic processes, biosynthetic processes and translation may be affected in the mutant
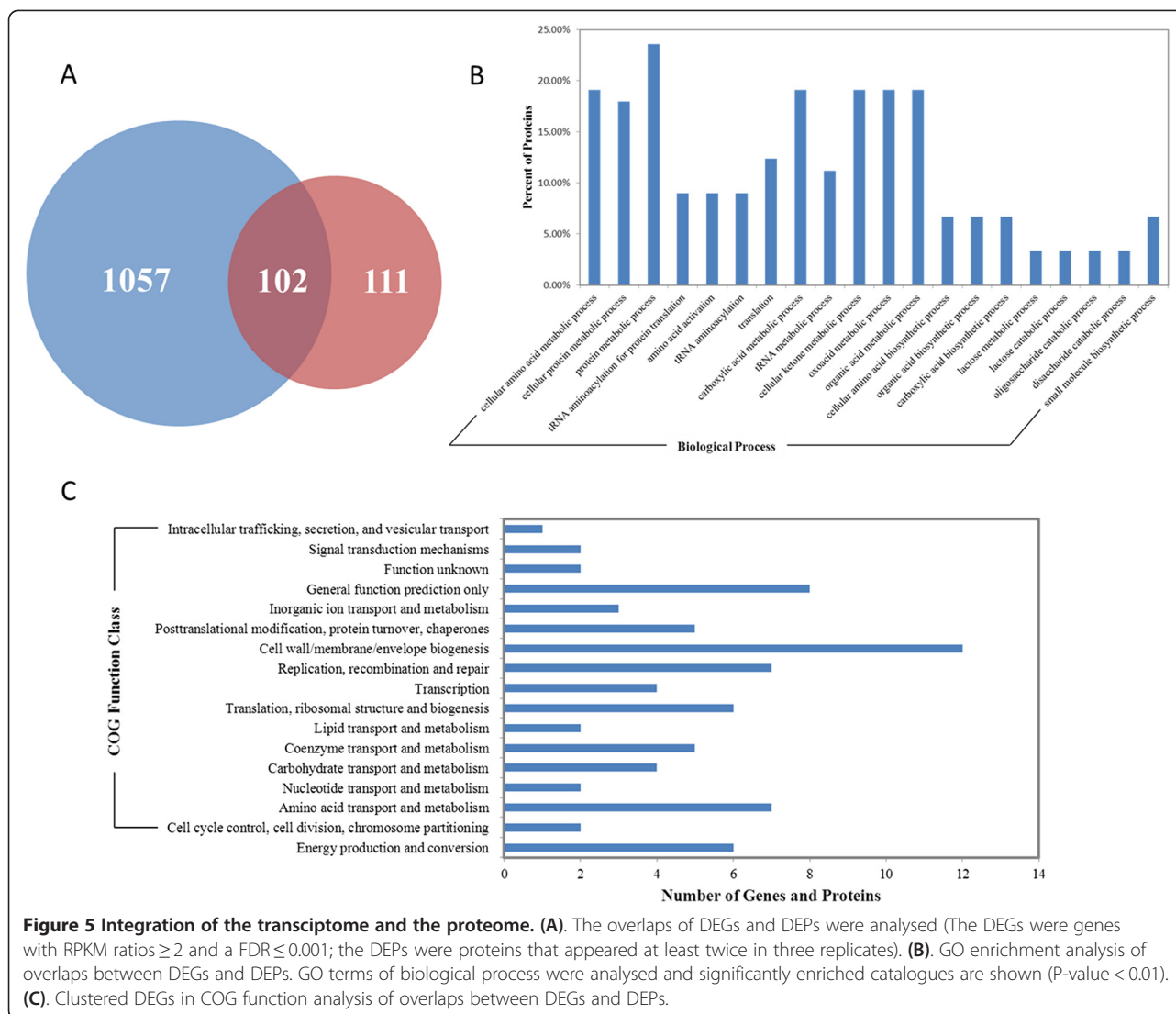
Figure 4 Comparative proteomic analysis. (A). Protein ratio distribution. The distribution of average value of protein quantification in three repeated experiments is shown. Red: fold change > 1.2, Green: fold change < −1.2. **(B)**. COG function analysis of differentially expressed proteins. **(C)**. KEGG pathways analysis of proteins with different expression (P value <0.05). **(D)**. Gene ontology enrichment analysis of differentially expressed proteins. GO terms of biological process were analysed and significantly enriched catalogues are shown (P-value < 0.01).

strain (Figure 5B). Functional classification according to COG function category indicates that, except for the general function prediction catalogue and the amino acid transport and metabolism catalogue, the genes with the greatest change in expression are classified into the cell wall/membrane/envelope biogenesis and replication catalogue and the recombination and repair catalogue (Figure 5C). Interestingly, the genetic comparison revealed that gene mutations were identified in *dprA* and *arpU*. The former gene was described as a competence gene involved in the protection of incoming DNA, and the latter gene was a transcriptional regulator that plays a role in cell wall growth and division [43].

Discussion

E. faecium is a part of the normal flora in human and animal intestines and is a ubiquitous opportunistic nosocomial pathogen. *E. faecium* was isolated from spacecraft-associated environments for the first time in 2009 [44]. Immune system suppression may make crew members

susceptible to *E. faecium* during spaceflight. Furthermore, the virulence of *E. faecium* may be enhanced during spaceflight. There is no comprehensive genetic information currently available for *E. faecium* after spaceflight, which makes it difficult to study the pathogenicity of the organism after exposure to this unique environment. We originally planned to research the impact of spaceflight environments on bacteria using *E. faecium* as a model. However, because the subculture may also produce unknown mutations, we cannot exclusively determine that the mutations identified after spaceflight were caused by the spaceflight environment. However, we did not obtain any mutants from the ground control strain subcultures. We were still interested in revealing the possible mechanisms of the mutant compared to the control strain using multiple 'omics' analysis. This study presents the whole genome, transcriptome and proteome of a mutant *E. faecium* strain. Our results show that 2,777 genes were predicted, and two point mutations were identified and were located in *dprA* and a transcriptional regulator
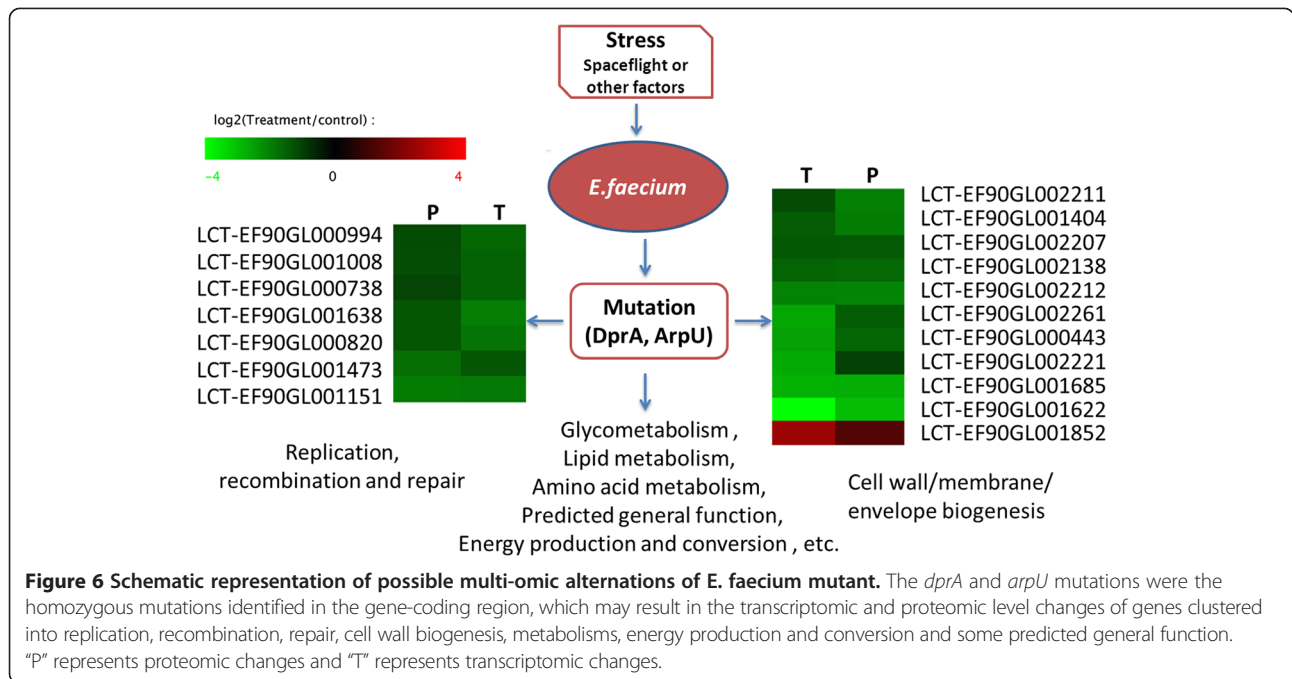
Figure 5 Integration of the transciptome and the proteome. (A). The overlaps of DEGs and DEPs were analysed (The DEGs were genes with RPKM ratios ≥ 2 and a FDR ≤ 0.001; the DEPs were proteins that appeared at least twice in three replicates). **(B)**. GO enrichment analysis of overlaps between DEGs and DEPs. GO terms of biological process were analysed and significantly enriched catalogues are shown (P-value < 0.01). **(C)**. Clustered DEGs in COG function analysis of overlaps between DEGs and DEPs.

(ArpU family). DprA was described as a member of a recombination-mediator protein family, which is required for natural transformation relating to horizontal gene transfer in bacteria [45-48]. ArpU was reported to control the muramidase-2 export, which plays an important role in cell wall growth and division. Mutation of *arpU* may lead to serious metabolic effects [43]. The transcriptome and proteome analysis suggests that the differentially expressed genes and proteins are mainly distributed in pathways involved in glycometabolism, lipid metabolism, amino acid metabolism, predicted general function, energy production and conversion, replication, recombination and repair, cell wall, membrane biogenesis, etc. Among these changes, the two main altered functional classifications were the replication, recombination and repair catalogue and the cell wall and membrane biogenesis catalogue, which are in accordance with the predicted functions of the mutated genes. Expression changes of genes in

the replication, recombination and repair catalogue may be caused by a stress-induced *dprA* mutation. The *arpU* mutation may affect the expression of members attributed to cell wall and membrane biogenesis (Figure 6). All of these changes at the molecular level may be caused by a stimulus during space flight. Because spacecraft are designed to provide an internal environment suitable for human life (reducing harmful conditions, such as high vacuum, extreme temperatures, orbital debris and intense solar radiation), *E. faecium* was placed in the cabin of the SHENZHOU-8 spacecraft to determine how microgravity as an external stimulus influences this bacterium.

Conclusion

This study was the first to perform comprehensive genomic, transcriptomic and proteomic analysis of an *E. faecium* mutant, an opportunistic pathogen often present in the GI tract of space inhabitants. We identified

Figure 6 Schematic representation of possible multi-omic alternations of E. faecium mutant. The *dprA* and *arpU* mutations were the homozygous mutations identified in the gene-coding region, which may result in the transcriptomic and proteomic level changes of genes clustered into replication, recombination, repair, cell wall biogenesis, metabolisms, energy production and conversion and some predicted general function. "P" represents proteomic changes and "T" represents transcriptomic changes.

dprA and *arpU* mutations, which affect genes and proteins with different expressions clustered into glycometabolism, lipid metabolism, amino acid metabolism, predicted general function, energy production, DNA recombination and cell wall biogenesis, etc. We hope that the current exploration of multiple "-omics" analyses of the *E. faecium* mutant could aid future studies of this opportunistic pathogen and determine the effects of the space environment on bacteria. However, the biochemical metabolism of bacteria is so complex that the biological meanings underlying the changes of *E. faecium* in this study is not fully understood. The implications of these gene mutations and expressions, and the mechanisms between the changes of biological features and the underlying molecular changes, should be investigated in the future. Moreover, the high cost of loading biological samples onto spacecraft and the difficult setting limits this type of exploration.

Additional file

Additional file 1: Tables S1, S2, S3, S4. Shows the repeat sequences statistics, SNP, Indels between LCT-EF258 and LCT-EF90, and annotation of InDels respectively. Supplementary figure represent function annotation in GO, GOG and KEGG database.

Competing interests

The authors declare that there are no competing interests.

Author's contributions

All authors proposed and designed the study. DC performed the approach and analyzed the results. All authors contributed to the writing of the manuscript. All authors read and approved the final manuscript.

Author details

[1]Nanlou Respiratory Diseases Department, Chinese PLA General Hospital, Beijing 100853, China. [2]BGI-Shenzhen, Shenzhen, People's Republic of China. [3]State Key Laboratory of Pathogen and Biosecurity, Beijing Institute of Microbiology and Epidemiology, Beijing 100071, China.

References

1. Franz CM, Stiles ME, Schleifer KH, Holzapfel WH: **Enterococci in foods– a conundrum for food safety.** *Int J Food Microbiol* 2003, **88**(2–3):105–122.
2. Lund B, Edlund C: **Probiotic Enterococcus faecium strain is a possible recipient of the vanA gene cluster.** *Clin Infect Dis* 2001, **32**(9):1384–1385.
3. Knoll BM, Hellmann M, Kotton CN: **Vancomycin-resistant Enterococcus faecium meningitis in adults: case series and review of the literature.** *Scand J Infect Dis* 2013, **45**(2):131–139.
4. Simjee S, White DG, McDermott PF, Wagner DD, Zervos MJ, Donabedian SM, English LL, Hayes JR, Walker RD: **Characterization of Tn1546 in vancomycin-resistant Enterococcus faecium isolated from canine urinary tract infections: evidence of gene exchange between human and animal enterococci.** *J Clin Microbiol* 2002, **40**(12):4659–4665.
5. Polidori M, Nuccorini A, Tascini C, Gemignani G, Iapoce R, Leonildi A, Tagliaferri E, Menichetti F: **Vancomycin-resistant Enterococcus faecium (VRE) bacteremia in infective endocarditis successfully treated with combination daptomycin and tigecycline.** *J Chemother* 2011, **23**(4):240–241.
6. Arias CA, Mendes RE, Stilwell MG, Jones RN, Murray BE: **Unmet needs and prospects for oritavancin in the management of vancomycin-resistant enterococcal infections.** *Clin Infect Dis* 2012, **54**(Suppl 3):S233–S238.
7. Olofsson MB, Pornull KJ, Karnell A, Telander B, Svenungsson B: **Fecal carriage of vancomycin- and ampicillin-resistant Enterococci observed in Swedish adult patients with diarrhea but not among healthy subjects.** *Scand J Infect Dis* 2001, **33**(9):659–662.

8. Shendure J, Ji H: **Next-generation DNA sequencing.** *Nat Biotechnol* 2008, **26**(10):1135–1145.
9. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics.** *Nat Rev Genet* 2009, **10**(1):57–63.
10. Lohse M, Bolger AM, Nagel A, Fernie AR, Lunn JE, Stitt M, Usadel B: **RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics.** *Nucleic Acids Res* 2012, **40**(Web Server issue):W622–W627.
11. Nanjo Y, Skultety L, Uvackova L, Klubicova K, Hajduch M, Komatsu S: **Mass spectrometry-based analysis of proteomic changes in the root tips of flooded soybean seedlings.** *J Proteome Res* 2012, **11**(1):372–385.
12. Tomazella GG, Risberg K, Mylvaganam H, Lindemann PC, Thiede B, de Souza GA, Wiker HG: **Proteomic analysis of a multi-resistant clinical Escherichia coli isolate of unknown genomic background.** *J Proteomics* 2012, **75**(6):1830–1837.
13. Benson G: **Tandem repeats finder: a program to analyze DNA sequences.** *Nucleic Acids Res* 1999, **27**(2):573–580.
14. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J: **Repbase update, a database of eukaryotic repetitive elements.** *Cytogenet Genome Res* 2005, **110**(1–4):462–467.
15. Chen N: *Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences, Current Protocols in Bioinformatics/Editoral Board, Andreas D Baxevanis [et al.]* 2004, **Chapter 4**:Unit 4 10. ; 2004.
16. Lowe TM, Eddy SR: **tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence.** *Nucleic Acids Res* 1997, **25**(5):955–964.
17. Nawrocki EP, Eddy SR: **Query-dependent banding (QDB) for faster RNA similarity searches.** *PLoS Comput Biol* 2007, **3**(3):e56.
18. Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR: **Rfam: an RNA family database.** *Nucleic Acids Res* 2003, **31**(1):439–441.
19. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL: **Versatile and open software for comparing large genomes.** *Genome Biol* 2004, **5**(2):R12.
20. Li H, Durbin R: **Fast and accurate short read alignment with burrows-wheeler transform.** *Bioinformatics* 2009, **25**(14):1754–1760.
21. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq.** *Nat Methods* 2008, **5**(7):621–628.
22. Audic S, Claverie JM: **The significance of digital gene expression profiles.** *Genome Res* 1997, **7**(10):986–995.
23. Unwin RD, Griffiths JR, Whetton AD: **Simultaneous analysis of relative protein expression levels across multiple samples using iTRAQ isobaric tags with 2D nano LC-MS/MS.** *Nat Protoc* 2010, **5**(9):1574–1582.
24. Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, Sherlock G: **GO: TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes.** *Bioinformatics* 2004, **20**(18):3710–3715.
25. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M: **KEGG for representation and analysis of molecular networks involving diseases and drugs.** *Nucleic Acids Res* 2010, **38**(Database issue):D355–D360.
26. Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, et al: **De novo assembly of human genomes with massively parallel short read sequencing.** *Genome Res* 2010, **20**(2):265–272.
27. Delcher AL, Bratke KA, Powers EC, Salzberg SL: **Identifying bacterial genes and endosymbiont DNA with Glimmer.** *Bioinformatics* 2007, **23**(6):673–679.
28. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25**(1):25–29.
29. Tatusov RL, Galperin MY, Natale DA, Koonin EV: **The COG database: a tool for genome-scale analysis of protein functions and evolution.** *Nucleic Acids Res* 2000, **28**(1):33–36.
30. Bairoch A, Apweiler R: **The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999.** *Nucleic Acids Res* 1999, **27**(1):49–54.
31. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, et al: **The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003.** *Nucleic Acids Res* 2003, **31**(1):365–370.
32. Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B: **The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics.** *Nucleic Acids Res* 2009, **37**(Database issue):D233–D238.
33. Winnenburg R, Baldwin TK, Urban M, Rawlings C, Kohler J, Hammond-Kosack KE: **PHI-base: a new database for pathogen host interactions.** *Nucleic Acids Res* 2006, **34**(Database issue):D459–D464.
34. Chakraborty A, Ghosh S, Chowdhary G, Maulik U, Chakrabarti S: **DBETH: a database of bacterial exotoxins for human.** *Nucleic Acids Res* 2012, **40**(Database issue):D615–D620.
35. Chen L, Yang J, Yu J, Yao Z, Sun L, Shen Y, Jin Q: **VFDB: a reference database for bacterial virulence factors.** *Nucleic Acids Res* 2005, **33**(Database issue):D325–D328.
36. Delcher AL, Salzberg SL, Phillippy AM: *Using MUMmer to Identify Similar Regions in Large Sequence Sets, Current Protocols in Bioinformatics/Editoral Board, Andreas D Baxevanis [et al.]* 2003, **Chapter 10**:Unit 10 13; 2003.
37. Lemeer S, Hahne H, Pachl F, Kuster B: **Software tools for MS-based quantitative proteomics: a brief overview.** *Methods Mol Biol* 2012, **893**:489–499.
38. Greenbaum D, Jansen R, Gerstein M: **Analysis of mRNA expression and protein abundance data: an approach for the comparison of the enrichment of features in the cellular population of proteins and transcripts.** *Bioinformatics* 2002, **18**(4):585–596.
39. Zhang W, Culley DE, Scholten JC, Hogan M, Vitiritti L, Brockman FJ: **Global transcriptomic analysis of Desulfovibrio vulgaris on different electron donors.** *Antonie Van Leeuwenhoek* 2006, **89**(2):221–237.
40. Nie L, Wu G, Culley DE, Scholten JC, Zhang W: **Integrative analysis of transcriptomic and proteomic data: challenges, solutions and applications.** *Crit Rev Biotechnol* 2007, **27**(2):63–75.
41. Crick F: **Central dogma of molecular biology.** *Nature* 1970, **227**(5258):561–563.
42. Gygi SP, Rochon Y, Franza BR, Aebersold R: **Correlation between protein and mRNA abundance in yeast.** *Mol Cell Biol* 1999, **19**(3):1720–1730.
43. Lleo MM, Fontana R, Solioz M: **Identification of a gene (arpU) controlling muramidase-2 export in Enterococcus hirae.** *J Bacteriol* 1995, **177**(20):5912–5917.
44. Stieglmeier M, Wirth R, Kminek G, Moissl-Eichinger C: **Cultivation of anaerobic and facultatively anaerobic bacteria from spacecraft-associated clean rooms.** *Appl Environ Microbiol* 2009, **75**(11):3484–3491.
45. Zhang XS, Blaser MJ: **DprB facilitates inter- and intragenomic recombination in Helicobacter pylori.** *J Bacteriol* 2012, **194**(15):3891–3903.
46. Tadesse S, Graumann PL: **DprA/Smf protein localizes at the DNA uptake machinery in competent Bacillus subtilis cells.** *BMC Microbiol* 2007, **7**:105.
47. Mortier-Barriere I, Velten M, Dupaigne P, Mirouze N, Pietrement O, McGovern S, Fichant G, Martin B, Noirot P, Le Cam E, et al: **A key presynaptic role in transformation for a widespread bacterial protein: DprA conveys incoming ssDNA to RecA.** *Cell* 2007, **130**(5):824–836.
48. Yadav T, Carrasco B, Myers AR, George NP, Keck JL, Alonso JC: **Genetic recombination in Bacillus subtilis: a division of labor between two single-strand DNA-binding proteins.** *Nucleic Acids Res* 2012, **40**(12):5546–5559.