Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Genome re-sequencing and reannotation of the Escherichia coli ER2566 strain and transcriptome sequencing under overexpression conditions

Lizhi Zhou
Xiamen University
Kaihang Wang
Xiamen University
Tingting Chen
Xiamen University
Yue Ma
Xiamen University
Yang Huang
Xiamen University
Jiajia Li
Xiamen University
Liqin Liu
Xiamen University
Yuqian Li
Xiamen University
Zhibo Kong
Xiamen University
Qingbing Zheng
Xiamen University
Yingbin Wang
Xiamen University
Ying Gu
Xiamen University
Hai Yu
Xiamen University
Shaowei Li (shaowei@xmu.edu.cn)
Xiamen University https://orcid.org/0000-0002-3374-1038
Ningshao Xia
Xiamen University

Research article

Keywords: Escherichia coli ER2566, genome reannotation, transcriptome sequencing, engineer bacteria

Posted Date: March 25th, 2020

DOI: https://doi.org/10.21203/rs.2.21231/v2

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Version of Record: A version of this preprint was published on June 16th, 2020. See the published version at https://doi.org/10.1186/s12864-020-06818-1.

Abstract

Background The Escherichia coli ER2566 strain (NC_CP014268.2) was developed as a BL21 (DE3) derivative strain and has been widely used in recombinant protein expression. However, like many other current RefSeq annotations, the annotation of the ER2566 strain is incomplete, with missing gene names and miscellaneous RNAs, as well as uncorrected annotations of some pseudogenes. Here, we performed a systematic reannotation of the ER2566 genome by combining multiple annotation tools with manual revision to provide a comprehensive understanding of the E. coli ER2566 strain, and used high-throughput sequencing to explore how the strain adapts under external pressure. Results The reannotation included noteworthy corrections to all protein-coding genes, led to the exclusion of 120 hypothetical genes or pseudogenes, and resulted in the addition of 65 coding sequences and 230 miscellaneous noncoding RNAs and 2 tRNAs. In addition, we further manually examined all 194 pseudogenes in the Ref-seq annotation and directly identified 144 (74%) as coding genes. The remaining pseudogenes without explicit function were removed. We then used whole-genome sequencing and high-throughput RNA sequencing to assess mutational adaptations under consecutive subculture or overexpression burden. Whereas no mutations were detected in response to consecutive subculture, overexpression of the human papillomavirus 16 type capsid led to the identification of a mutation (position 1,094,824 within the 3' non-coding region) positioned 19-bp away from the lac I gene in the transcribed RNA, which was not detected at the genomic level by Sanger sequencing. Conclusion The ER2566 strain is used by both the general scientific community and the biotechnology industry. Reannotation of the E.coli ER2566 strain not only improved the RefSeq data but uncovered a key site that might involve in the transcription and translation of genes encoding the lactose operon repressor. We propose that our pipeline might offer a universal method for the reannotation of other bacterial genomes with high speed and accuracy. This study may facilitate a better understanding of gene function for the ER2566 strain under external burden and provide more clues to engineer bacteria for biotechnological applications.

Background

The *Escherichia coli* expression system is one of the most well-characterized classical expression systems for recombinant protein expression in biological science. *E. coli* offers clear advantages over other expression systems, including a clear genetic background, fast breeding, low cost, high cell density cultures, and high protein expression levels [1-3]. Indeed, more than 60% of recombinant proteins and nearly 30% of approved recombinant therapeutic proteins are produced using *E. coli* expression [4]. The *E.coli* ER2566 strain is a common laboratory tool that takes advantage of the expression and growth properties of the B line strain [5]. In 2016, the first complete genome of ER2566 (C2566, NC_CP014268.2) competent cells was sequenced by New England Biolabs and deposited into GenBank [6], with automatic annotation by the NCBI prokaryotic genome annotation pipeline (PGAP) [7].

Along with the rapid development of biological laboratory techniques, there has been a significant advance in sequencing technologies. However, this has not been matched by the production of better sequences; rather, sequencing advancements have led to the deposition of an increased number of

"draft" bacterial genomes into public databases [8], which tend to be incomplete and fragmented. In addition, the massive amounts of genomic data generated by next-generation sequencing platforms has also increased the probability of errors in genome annotations, since most genomes are annotated automatically and not subjected to any manual review. High-quality annotations of bacterial genomes are critical to understanding biological processes and enhancing these genomes has become a major task in the post-genomic era. Therefore, the reannotation of previously published genomes with manual reviewing is necessary to improve databases and supply accurate information [9]. Although numerous tools have been used to identify relevant genes, gene prediction is still imperfect. In addition, some genuine genes are missed by gene finder tools because the algorithms are directed to maintain a balance between specificity and sensitivity to avoid false-positive predictions. Thus, the use of multiple *ab initio* gene finders along with blast searching will help to identify genes correctly and lead to more accurate annotations [10, 11].

Next-generation sequencing (NGS) is a cost-effective tool for the study of gene function and experimental bacterial evolution. Indeed, NGS technology was used by Luhachack and colleagues to successfully identify the function of the transcription factor YcjW as a regulator of the complex interaction between carbohydrate metabolism and H2S production in bacteria [12], and whole-genome re-sequencing in *E.coli* by Herring et al. led to the identification of mutations that conveyed a selective growth advantage during adaptation to a glycerol-based growth medium [13]. Even though the ER2566 engineering strain is widely used for the preparation of virus-like particles [14], and bacterial surface display [15], among other functions, exploring bacterial adaptive evolution of the ER2566 strain under various external pressures through NGS remains difficult and inconclusive because of limitations in the annotations of the strain [13].

In this study, we employed a series of automated annotations and combined this with manual inspections with high-throughput analyses to reannotate the ER2566 genome. As a result of our analysis, we further propose a universal reannotation pipeline for other bacterial genomes that can be undertaken with high speed and accuracy. Our reannotation eliminated 120 hypothetical genes and pseudogenes from the RefSeq annotation, and now includes an additional 65 coding sequences with definitive gene names and functions. The number of miscellaneous noncoding RNAs was also increased from 15 to 245. Subsequently, we applied whole-genome sequencing and RNA-sequencing to assess for mutational adaptations that occur following continuous subculture pressure or overexpression burden. We detected a mutation located within the 3' non-coding region (1,094,824 position) 19-bp away from the *lac*I gene at the RNA level, which may be involve in the transcription and translation of genes encoding the lactose operon repressor. Our reannotation and sequencing results will provide a better understanding of some of the biological processes of the ER2566 strain, and may offer insight into future biotechnological applications in bacterial engineering.

Results & Discussions

The process of genome reannotation combined with detailed manual reviewing encompasses the re-identification and labeling of characteristic features of a sequenced genome, and is a process that has been performed extensively for numerous organisms, including bacteria [16, 17]. Unlike the human genome, which is about 1.3% protein coding, 90% of the bacterial genome codes for proteins, with only short intergenic stretches[18]. Precise genomic annotation is thus fundamental to the further interpretation of the biochemical and physiological characteristics of organisms, to provide detailed information on protein coding sequences, pseudogenes, non-coding RNAs, repeat sequences and various other genomic data [19]. In this study, we reannotated the genome of the *E.coli* strain ER2566 through a reannotation pipeline, as illustrated in Figure 1, with high speed and accuracy.

We employed a series of automated annotation tools combined with manual inspection to reannotate the ER2566 genome. In our pipeline, the automation part using Prokka combined with others gene finders (GLIMMER, Zcurve and GeneMark) could finish a complete bacterial genome annotation in about 30 min. Compared with some online tools, this pipeline showed higher speed and accuracy. For example, NCBI provides a Prokaryotic Genomes Automatic Annotation Pipeline service via email, with a turn-around time of several days. RAST is another web server for annotating bacterial and archaeal genomes that provides results within one day. Some local stand-alone annotation tools, such as RATT[20], Rapid Annotation Transfer Tool, can transfer annotations from a high-quality reference to a new genome on the basis of conserved synteny. However, due to the limitation of its algorithm, RATT cannot effectively identify pseudogenes, indels, etc. Besides, numerous automated tools have been developed for genome annotation, including Mypro[21], MAKER[22], BlastLKOLA[23] and so on. To avoid false-positive predictions, the algorithms of these annotation tools are designed to balance specificity and sensitivity of their results. In contrast, combination of multiple ab initio gene finders combined with blast searching and manual inspection will help to confirm identified genes and generate more accurate annotations.

Improvement in coding sequences (CDSs)

For the systematic reannotation of the CDSs, the prediction and identification of coding genes occurred in two stages. In the first stage, Prodigal software was used to predict a total of 4,180 CDSs on the complete ER2566 genome deposited in GenBank (accession number NC_CP014268.2). Using sequence alignment to the Swiss-Prot database [24] by blastp [25], with a threshold e-value of $<10^{-6}$, all CDSs were artificially annotated to provide accurate information regarding the sequences and functions of the enrolled proteins. A total of 4,044 (97%) of the 4,180 CDSs were annotated as protein-coding genes, with the remaining CDSs (136 CDSs; 3%) marked as hypothetical genes, with no registration in the Swiss-prot database. To improve upon this prediction, three other well-established gene finders—GLIMMER, Zcurve and GeneMark—were independently used, identifying 4,231, 4,287, and 4,213 CDSs, respectively. These putative CDS sets were subsequently filtered by blastn against the first Prodigal-predicted CDS set. Overall, an additional 428 CDSs were found: 194 CDSs were identified using GLIMMER, 123 using GeneMark, and 201 using ZCURVE. These additional 428 CDSs were then searched against the Swiss-Prot database by blastp with a stricter threshold e-value $<10^{-10}$, coverage >80%, and an identity >70%. This filtered out 402 of these additional CDSs, resulting in an additional 43 genes. This led to a total of

4,087 protein-coding CDSs (4,044 + 43) included in the reannotation of the ER2566 genome, along with 136 CDSs for hypothetical genes.

Due to trimming or splitting, genes with real function can often be incorrectly assigned as pseudogenes through protein homology alignment. In our reannotation, we manually reviewed 194 pseudogenes from the RefSeq database annotation. In total, 144 of the 194 pseudogenes were directly identified as coding genes and are now found in the reannotated list, while the remaining pseudogenes without any function were removed. These newly identified protein-coding genes included 34 mobile genetic elements, a common type of genetic change considered to be important in evolution. For instance, a genomic positive strand region (3,296,328 - 3,297,025 bp), previously annotated as a pseudogene without function, was identified to be two genes, insA and insB, which are homologues of the insertion element protein, IS1, and related to DNA binding and transposase activity (Fig. 2a) [26]. In addition, two annotated pseudogenes (C2566-RS05300 and C2566_RS05310) and one hypothetical protein (C2566_RS22600) in the RefSeq annotation (range 1,088,980-1,094,720) were reannotated as three new genes (*lac*Z1, *lac*Z2, and ECBD_2906, respectively), and one related pseudogene, C2566_RS05305, was removed; these changes are consistent with previous results [5] (Fig. 2b). These three new genes are flanked by the lactose permease gene *lac*Y upstream and the lactose operon repressor gene *lac*I downstream, both of which are essential to the lac operon system. This reannotation had uncovered genes related to the transcription and translation of genes encoding the lactose operon repressor in ER2566 strain.

By comparing with RefSeq annotation, a total of 120 protein-coding CDSs were removed, as they were identified as either hypothetical proteins or pseudogenes, with most of them having no assigned function (Additional Files S1). Meanwhile, 65 new CDSs were added (Additional Files S2). The complete reannotation list can be found in Additional File S3. ER2566 strain is a BL21/K-12 hybrid strain, where about 6% sequence of K-12 strain replaces about 7% sequence in BL21(DE3) genome. The genome alignment of BL21 and ER2566 demonstrates a high degree of consistency [27] (Fig 3). The recent version of BL21(DE3) annotation contains 4,197 CDSs, in which 3873 (92.3%) CDSs were identified in ER2566 annotation as identical gene symbols or alias as well. The unidentical 7.7% CDSs annotated in BL21 genome as compared to ER2566 is comprised of ~7% sequence corresponding to the hybrid part in ER2566 and other CDSs that does not have an official gene name (Additional File S4). Overall, we determined 4,223 protein-coding genes for the ER2566 genome, including 136 CDSs labeled as hypothetical proteins and 3873 CDSs identical to BL21(DE3) that account for about 99% of total 3,903 CDSs (4,197*93%) equivalent to BL21(DE3) CDSs within ER2566 genome. This reannotation effectively eliminated the possibility of false interpretations introduced by the original annotation and provides a more integral view of the regulatory networks in ER2566 strain (Table 1).

Integrated proteogenomics search database (iPtgxDB) is widely used to provide protein expression evidence and could confirm the validity of the annotation, which was used to identify the short protein-coding genes that have numerous functions[28]. Thus, the combination of transcriptome data and reannotation results was used to generate an integrated proteomics database [29], which provided an

important optimal basis for genome-scale regulatory or metabolic predictions and comprehensive exploration on the genome information and underlying genes' functions (Additional File S8).

Miscellaneous RNA improvement

Miscellaneous RNA, such as transfer RNA (tRNA), ribosomal RNA, and other non-coding RNAs, play pivotal biological roles in cellular activity. To date, about 119 RNA molecules in the E.coli ER2566 strain have been identified, including 85 tRNAs, 22 rRNAs, and 12 ncRNAs[6]. However, almost all of these ncRNAs are missing from the original RefSeq annotation. We used Aragorn 1.2, Barrnap 0.9, and infernal 1.1 ncRNA finders independently to predict genes coding for tRNAs, rRNAs, and non-coding RNA. Compared with the auto-annotation, 2 tRNAs and 230 ncRNAs are new, with most having functions in translation, DNA replication, and expression regulation (Additional Files S5). In addition, most of the ncRNA functions have been verified experimentally previously. For instance, about 94 of the nucleoid-associated ncRNA molecules play key functions in DNA-RNA interactions [30]. Meanwhile, the fragments per kilobase of transcript per million mapped reads (FPKM) was used to quantify the transcription level of the newly added ncRNA and to analyze the transcriptomics data of ER2566 under different induction conditions (Additional File S6). Quantitative analysis indicates that 85% (208/245) of newly added ncRNA are detectable, while the other 15% (37/245) ncRNA are undetectable possibly due to some specific functions requiring certain conditions. In summary, our reannotation introduced 230 new ncRNAs and 2 new tRNAs, with an overall tally of genes encoding for 87 tRNAs, 22 rRNAs and 245 ncRNAs.

Variant calling of whole-genome re-sequencing under consecutive subculture

Genomic variations in bacterial species usually reflect an evolutionary response that occurs under various external—usually unfavorable—environmental stressors. Thus, we next performed variant calling to identify any nucleotide-level differences (i.e., single nucleotide polymorphisms (SNPs), insertions and deletions (indels), and/or structural variations) in the ER2566 strain. There are two approaches for variant calling: by mapping reads against the reference genome directly or by assembling a *de novo* genome to compare against a reference genome. In most cases, mapping reads produces a better resolution for SNPs and indels than genome assembly, whereas the latter is optimal for identifying structural variants and regions with high divergence. Here, we used both methods to interrogate the ER2566 genome (Fig. 4).

First, we re-sequenced the whole genome of the ER2566 strain grown in our laboratory under consecutive subculturing. Two different-sized insert libraries (500 bp, 2000 bp) were built, and a total of 10.0 million paired-end reads of 90 bp in length were generated using an Illumina HiSeq. The raw reads were mapped to the C2566 reference genome (NC_CP014268.2) with a good coverage depth (>100-times). No SNP or indel was detected. A pipeline optimized for longer assembly was designed to accomplish the re-sequencing of our ER2566 strain. Various de novo assembly softwares were used to construct a confident and long (4,469,460-bp) scaffold, and assembly results for each step were assessed by alignment of final sequences back to the reference genome (Fig S1). One inversion was found (Fig S1c), which turned out to be a genome assembly issue caused by high repetition region and was corrected by subsequent Sanger sequencing. The technical difference between short-reads sequencing and single

molecule long-reads sequencing may result in the generation of inverted region. Nevertheless, our pipeline by the combination of reads mapping, de novo assembly and Sanger confirmation generate an intact ER2566 genome in our practice. The resequencing of ER2566 also suggest that the continuous subculture of E. coli ER2566 strain in our lab did not cause mutation in the genomic sequence.

Mutation detections by RNA-sequencing under overexpression

RNA-Seq is widely used in quantitative gene expression studies for the identification of non-annotated transcripts and polymorphisms, and for RNA editing in transcribed regions. Thus, to identify any variations in the ER2566 genome due to overexpression pressure, we used RNA-seq to analyze the transcriptomes of the ER2566 strain growing at 37°C without plasmids (B37, three replicates) or overexpressing human papillomavirus 16 type capsid protein L1 via plasmid-based inducible expression (Y37, three replicates) (Fig. 5a).From a total of 81.3 million 125-bp paired-end reads, 78.6 million reads (96.74%) were mapped to the reference genome in B37 (control) samples. Yet, for the Y37 (overexpressed) samples, among the 76.0 million reads, only 24.0 million (31.50%) reads mapped to the reference genome, which was significantly lower than that for the B37 samples. The cause of lower mapping rates in Y37 samples was due to the large number of mRNAs transcribed by the engineered plasmids which is not related to genome sequence (Table 2).

The variant detection analysis using BactSNP revealed one mutation in the Y37 samples (1,094,824 position; Fig. 5b), located in the non-coding region 19-bp downstream from the lacI gene. Interestingly, lacI is the highest transcribed gene in the Y37 samples by comparing the FPKM for all genes (Additional Files S7). The analysis of transcriptome data indicated a mutation of C to T substitution at position 109,824 in the three replicates of Y37 samples, as confirmed by nearly 100% mutation rate in all observed reads (910,914 out 911,345 reads). Furthermore, the mutation was found in three replicates of B37 samples as well, with a mutation rate up to 85% (655 out 773 reads). Surprisingly, such mutation could not be detected in the bacterial genome by Sanger sequencing (Fig. 5c). The discrepancy between transcribe RNA and genome may arise from the modification during RNA transcription instead of sequencing error. The presumption is supported by the mutation position being located in a high-efficiency RNA methylation site, which is often accompany with spontaneous deamination of 5-methylcytosine and consequently producing thymine in aqueous solution [31].

Conclusions

Here, we employed a series of automated annotation tools along with manual inspection to reannotate the ER2566 genome. The major updates include the noteworthy correction of all protein-coding genes, the exclusion of 120 CDSs from the Refseq annotation, and the addition of 65 new CDSs with definitive name or putative function. Moreover, there is an increase in the number of miscellaneous RNA from 15 to 245. These new additions will help to provide a more informative profile of the ER2566 genome and provide a better base for exploring the molecular mechanisms of stress in response to changes in the bacterial cellular milieu. Nevertheless, this reannotation still has further room for improvement, with the

continuing advancement of the algorithm and the accumulation of next-generation sequence data. We also carried out whole-genome sequencing and RNA-seq to detect variant calling under different conditions of external pressure, and detected one mutation within the non-coding region of the *lac*I gene. However, this mutation was not detected at the genomic level by Sanger sequencing, which may indicate that this is a RNA modification related to the biological strain of overexpression pressure in ER2566.

The ER2566 strain is used widely within the scientific community, and our reannotation not only improved the characterization of the strain but uncovered a key site that might involve in the transcription and translation of genes encoding the lactose operon repressor. Our reannotation pipeline with high speed and accuracy could thus be extrapolated for the reannotation of other bacterial genomes to provide a better understanding of gene function under external burden and provide more clues to engineer bacteria for biotechnological applications.

Methods

Bacterial strains, plasmids, and culture conditions

The *E.coli* B Strain *ER2566* was purchased from New England Biolabs (NEB). Cells were grown at 37°C with sharking at 180 rpm in Luria-Bertani (LB) broth (5 g/L NaCl, 10 g/L tryptone, 5 g/L yeast extract, pH 7.0) under the pressure of continual passaging.

The gene encoding human papillomavirus 16 type L1 was cloned into the pTO-T7 expression vector. *E.coli ER2566* with the pTO-T7-HPV-16L1 vector was cultured in 250 ml LB broth containing 20 μg/ml kanamycin at 37°C with shaking at 180 rpm. Upon reaching an OD_{600} of 1.0, 5 ml of culture was transferred into a flask containing 500 ml LB and 20 μg/ml kanamycin, and incubated at 37°C with shaking at 180 rpm. At OD_{600} of 0.6, the culture was induced with a final concentration of 0.1 mM/L isopropyl-β-d-thiogalactopyranoside (IPTG) and incubated for 7 h at 37°C with shaking at 180 rpm.

DNA extraction and sequencing

Genomic DNA was extracted using a cetyltrimethylammonium bromide (CTAB)-based protocol[32]. Total genomic DNA concentration and quality were determined using a NanoDrop2000 Spectrophotometer (Thermo Fisher Scientific). DNA libraries for Illumina sequencing were constructed for each accession number according to the manufacturer's specifications. After DNA library construction, sequencing was performed by a commercial service (Novogene, Beijing, China) on an Illumina HiSeq platform with 90-bp read lengths. Finally, 10.7 million raw reads were obtained for subsequent analyses.

RNA extraction and sequencing

Cells were harvested by centrifugation at 7000 rpm for 10 min at room temperature and total RNA was extracted using the MasterPure RNA Purification Kit, according to the manufacturer's protocol (Lucigen). DNase was added to reduce the chance of genomic DNA contamination. Total RNA was extracted in 50 μl

RNase-free DEPC-treated water. RNA concentration was measured using an RNA Assay Kit in a Qubit 2.0 Fluorometer. A total of 3 μg RNA per sample was used as the input material for RNA sample preparations. Libraries of RNA-seq template were constructed using the NEBNext Ultra RNA Library Prep Kit for Illumina (NEB, USA), following the manufacturer's recommendations. Sequencing was performed on an Illumina HiSeq 2500 platform and 125-bp paired-end reads were generated.

Genome reannotation of the ER2566 strain

The ER2566 strain (NC_CP014268.2) whole genome was downloaded from the NCBI Reference Sequence Database as an input file for Prokka, a widely used annotation software that annotates a bacterial genome in about 10 min on a laptop [33]. To attain a rich and reliable annotation, we coordinated a suite of existing tools, including Prodigal 2.6 [34], Aragorn 1.2 [35], Barrnap 0.9 (https://github.com/tseemann/barrnap), Infernal 1.1 [36] and MinCED 0.2.0 [37] for our reannotation pipeline to predict, respectively, coding sequences (CDS), ribosomal RNA genes (rRNA), transfer RNA genes (tRNA), non-coding RNA genes (ncNRA) and clustered regularly interspaced short palindromic repeats (CRISPRs). Subsequently, various gene finders (GLIMMER 2.03 [38], GeneMark [39] and ZCURVE [40]) were used to further confirm the coding sequences.

Variant calling of whole-genome sequencing by reads mapping

The quality of the raw reads were determined using FastQC [41], and appropriately truncated and filtered using fastp [42] to remove low-quality bases and Illumina adapter contamination with default parameters. The clean reads were then mapped against the C2566 reference genome using bwa-mem [43] with standard settings, and sorted by location as bam files. Bam files were converted to sam files using Samtools [44]. BactSNP was used to remove duplications and to detect variant calling [45].

Variant calling of whole-genome sequencing by genome assembly

Draft contigs were created using SPAdes [46], with optimized parameters. The assembled draft contigs and sequencing libraries were used as input into Redundans [47] and scaffold genome assembly was performed with the recommended parameters. This resulted in fewer fragments, longer sequences and fewer gaps, as compared with using the input contigs. Scaffolds were then further improved using ABACAS [48], which rapidly aligned, ordered and orientated the scaffolds based on the following user-provided references: perl abacas.pl –r reference.fa –q scaffolds.fa –p nucmer. Finally, the assemblies were used to identify the mutants or indels against the reference genome through the program Harvest [49].

Abbreviations

BLAST: Basic local alignment search tool

CDS: Coding sequence

IPTG: Isopropyl-β-d-thiogalactopyranoside

NGS: Next-generation sequencing

ORF: Open reading frame

PGAP: Prokaryotic genome annotation pipeline

RNA-seq: RNA Sequencing

SNP: Single nucleotide polymorphisms

WGS: Whole-genome sequencing

FPKM: The Fragments per kilobase of transcript per million mapped reads

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The genomic sequence of the E.coli B strain ER2566 was downloaded in FASTA format from the NCBI-Microbial Genome Database (NZ_CP014268.2) and was annotated by Prokka using multiple annotation tools and manual review. Whole-genome re-sequencing and RNA-seq data in this study were respectively deposited in the NIH Sequence Read Archive (www.ncbi.nlm.nih.gov/sra/). The raw reads of whole-genome re-sequencing for different insert sizes:ER2566_500-SRR10828732 (https://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR10828732), ER2566_2000-SRR10828731 (https://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR10828731), RNA-Seq raw data:BB372-SRR10828730(https://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR10828730), BB372-SRR10828729(https://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR10828729), BB373-SRR108287828(https://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR10828728), YY371-SRR108287827(https://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR10828727), YY372-SRR108287826(https://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR10828726), YY373-SRR108287825(https://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR10828725).

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Chinese government: the National Natural Science Foundation of China (grant no. U1705283, 31670935, 81971932), the Natural Science Foundation of Fujian Province (Grant 2017J07005), and the New drug Invention project (grant no.2018ZX09738-008). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors' Contributions

Y.G., H.Y. and S.L. conceived the study. L.Z., K.W., T.C. and Y.M. performed DNA-library preparation, whole-genome and RNA-seqing by Illumina. L.Z., K.W., J.L., L.L., Y.L. and H.Y. performed SNP analysis, RNA-seq analysis and reannotation. L.Z., K.W., Z.K., Q.Z., Y.W., Y.G., H.Y., S.L. and N.X. analyzed data. L.Z., Y.G., H.Y. and S.L. wrote the manuscript. All authors have read and approved the final manuscript.

Acknowledgements

We thank Mrs. Rebecca Jackson for polishing and revising of the final manuscript.

Authors' information

Affiliations

State Key Laboratory of Molecular Vaccinology and Molecular Diagnostics, School of Public Health, Xiamen University, Xiamen, Fujian, 361102, People's of Republic of China.

Lizhi zhou, Zhibo Kong, Yingbin Wang, Qingbing Zheng, Hai Yu, Ying Gu, Shaowei Li, Ningshao Xia

National Institute of Diagnostics and Vaccine Development in Infectious Disease, School of Life Sciences, Xiamen University, Xiamen, Fujian, 361102, People's of Republic of China.

Kaihang Wang, Tingting Chen, Yue Ma, Yang Huang, Jiajia Li, Liqin Liu, Yuqian Li, Ying Gu, Shaowei Li, Ningshao Xia

Corresponding authors

Correspondence to Ying Gu or Hai Yu or Shaowei Li

References

1. Sezonov G, Joseleau-Petit D, D'Ari R: **Escherichia coli physiology in Luria-Bertani broth**. *J Bacteriol* 2007, **189**(23):8746-8749.
2. Shiloach J, Fass R: **Growing E-coli to high cell density - A historical perspective on method development**. *Biotechnol Adv* 2005, **23**(5):345-357.

3. Rosano GL, Ceccarelli EA: **Recombinant protein expression in Escherichia coli: advances and challenges**. *Front Microbiol* 2014, **5**.
4. Correa A, Oppezzo P: **Overcoming the solubility problem in E. coli: available approaches for recombinant protein production**. *Methods Mol Biol* 2015, **1258**:27-44.
5. Fomenkov A, Sun Z, Dila DK, Anton BP, Roberts RJ, Raleigh EA: **EcoBLMcrX, a classical modification-dependent restriction enzyme in Escherichia coli B: Characterization in vivo and in vitro with a new approach to cleavage site determination**. *PLoS One* 2017, **12**(6):e0179853.
6. Anton BP, Fomenkov A, Raleigh EA, Berkmen M: **Complete Genome Sequence of the Engineered Escherichia coli SHuffle Strains and Their Wild-Type Parents**. *Genome Announc* 2016, **4**(2).
7. Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, Lomsadze A, Pruitt KD, Borodovsky M, Ostell J: **NCBI prokaryotic genome annotation pipeline**. *Nucleic Acids Res* 2016, **44**(14):6614-6624.
8. Goodwin S, McPherson JD, McCombie WR: **Coming of age: ten years of next-generation sequencing technologies**. *Nat Rev Genet* 2016, **17**(6):333-351.
9. Ouzounis CA, Karp PD: **The past, present and future of genome-wide re-annotation**. *Genome Biol* 2002, **3**(2):COMMENT2001.
10. Luo CW, Hu GQ, Zhu HQ: **Genome reannotation of Escherichia coli CFT073 with new insights into virulence**. *Bmc Genomics* 2009, **10**.
11. Warren AS, Archuleta J, Feng WC, Setubal JC: **Missing genes in the annotation of prokaryotic genomes**. *Bmc Bioinformatics* 2010, **11**.
12. Luhachack L, Rasouly A, Shamovsky I, Nudler E: **Transcription factor YcjW controls the emergency H2S production in E. coli**. *Nat Commun* 2019, **10**(1):2868.
13. Herring CD, Raghunathan A, Honisch C, Patel T, Applebee MK, Joyce AR, Albert TJ, Blattner FR, van den Boom D, Cantor CR *et al*: **Comparative genome sequencing of Escherichia coli allows observation of bacterial evolution on a laboratory timescale**. *Nat Genet* 2006, **38**(12):1406-1412.
14. Zhang X, Wei M, Pan H, Lin Z, Wang K, Weng Z, Zhu Y, Xin L, Zhang J, Li S *et al*: **Robust manufacturing and comprehensive characterization of recombinant hepatitis E virus-like particles in Hecolin((R))**. *Vaccine* 2014, **32**(32):4039-4050.
15. Chen TT, Wang KH, Chi X, Zhou LZ, Li JJ, Liu LQ, Zheng QB, Wang YB, Yu H, Gu Y *et al*: **Construction of a bacterial surface display system based on outer membrane protein F**. *Microb Cell Fact* 2019, **18**.
16. Chowdhary N, Selvaraj A, KrishnaKumaar L, Kumar GR: **Genome Wide Re-Annotation of Caldicellulosiruptor saccharolyticus with New Insights into Genes Involved in Biomass Degradation and Hydrogen Production**. *Plos One* 2015, **10**(7).
17. Slager J, Aprianto R, Veening JW: **Deep genome annotation of the opportunistic human pathogen Streptococcus pneumoniae D39**. *Nucleic Acids Research* 2018, **46**(19):9971-9989.
18. Salzberg SL: **Next-generation genome annotation: we still struggle to get it right**. *Genome Biol* 2019, **20**(1):92.

19. Armengaud J: **Reannotation of Genomes by Means of Proteomics Data**. *Proteomics in Biology, Pt A* 2017, **585**:201-216.
20. Otto TD, Dillon GP, Degrave WS, Berriman M: **RATT: Rapid Annotation Transfer Tool**. *Nucleic Acids Research* 2011, **39**(9).
21. Liao YC, Lin HH, Sabharwal A, Haase EM, Scannapieco FA: **MyPro: A seamless pipeline for automated prokaryotic genome assembly and annotation**. *J Microbiol Methods* 2015, **113**:72-74.
22. Campbell MS, Holt C, Moore B, Yandell M: **Genome Annotation and Curation Using MAKER and MAKER-P**. *Curr Protoc Bioinformatics* 2014, **48**:4 11 11-39.
23. Kanehisa M, Sato Y, Morishima K: **BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences**. *J Mol Biol* 2016, **428**(4):726-731.
24. Bairoch A, Apweiler R: **The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000**. *Nucleic Acids Research* 2000, **28**(1):45-48.
25. McGinnis S, Madden TL: **BLAST: at the core of a powerful and diverse set of sequence analysis tools**. *Nucleic Acids Research* 2004, **32**:W20-W25.
26. Badia J, Ibanez E, Sabate M, Baldoma L, Aguilar J: **A rare 920-kilobase chromosomal inversion mediated by IS1 transposition causes constitutive expression of the yiaK-S operon for carbohydrate utilization in Escherichia coli**. *J Biol Chem* 1998, **273**(14):8376-8381.
27. Kim S, Jeong H, Kim EY, Kim JF, Lee SY, Yoon SH: **Genomic and transcriptomic landscape of Escherichia coli BL21(DE3)**. *Nucleic Acids Res* 2017, **45**(9):5285-5293.
28. Storz G, Wolf YI, Ramamurthi KS: **Small Proteins Can No Longer Be Ignored**. *Annu Rev Biochem* 2014, **83**:753-+.
29. Omasits U, Varadarajan AR, Schmid M, Goetze S, Melidis D, Bourqui M, Nikolayeva O, Quebatte M, Patrignani A, Dehio C *et al*: **An integrative strategy to identify the entire protein coding potential of prokaryotic genomes by proteogenomics**. *Genome Res* 2017, **27**(12):2083-2095.
30. Qian Z, Zhurkin VB, Adhya S: **DNA-RNA interactions are critical for chromosome condensation in Escherichia coli**. *P Natl Acad Sci USA* 2017, **114**(46):12225-12230.
31. Nabel CS, Manning SA, Kohli RM: **The curious chemical biology of cytosine: deamination, methylation, and oxidation as modulators of genomic potential**. *ACS Chem Biol* 2012, **7**(1):20-30.
32. Sharma RC, Murphy AJ, DeWald MG, Schimke RT: **A rapid procedure for isolation of RNA-free genomic DNA from mammalian cells**. *Biotechniques* 1993, **14**(2):176-178.
33. Seemann T: **Prokka: rapid prokaryotic genome annotation**. *Bioinformatics* 2014, **30**(14):2068-2069.
34. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ: **Prodigal: prokaryotic gene recognition and translation initiation site identification**. *Bmc Bioinformatics* 2010, **11**.
35. Laslett D, Canback B: **ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences**. *Nucleic Acids Research* 2004, **32**(1):11-16.
36. Nawrocki EP, Kolbe DL, Eddy SR: **Infernal 1.0: inference of RNA alignments**. *Bioinformatics* 2009, **25**(10):1335-1337.

37. Bland C, Ramsey TL, Sabree F, Lowe M, Brown K, Kyrpides NC, Hugenholtz P: **CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats**. *Bmc Bioinformatics* 2007, **8**.
38. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL: **Improved microbial gene identification with GLIMMER**. *Nucleic Acids Research* 1999, **27**(23):4636-4641.
39. Lukashin AV, Borodovsky M: **GeneMark.hmm: new solutions for gene finding**. *Nucleic Acids Research* 1998, **26**(4):1107-1115.
40. Guo FB, Zhang CT: **ZCURVE_V: a new self-training system for recognizing protein-coding genes in viral and phage genomes**. *Bmc Bioinformatics* 2006, **7**.
41. Brown J, Pirrung M, McCue LA: **FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool**. *Bioinformatics* 2017, **33**(19):3137-3139.
42. Chen SF, Zhou YQ, Chen YR, Gu J: **fastp: an ultra-fast all-in-one FASTQ preprocessor**. *Bioinformatics* 2018, **34**(17):884-890.
43. Houtgast EJ, Sima VM, Bertels K, Al-Ars Z: **Hardware acceleration of BWA-MEM genomic short read mapping for longer read lengths**. *Comput Biol Chem* 2018, **75**:54-64.
44. 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**(16):2078-2079.
45. Brouard JS, Schenkel F, Marete A, Bissonnette N: **The GATK joint genotyping workflow is appropriate for calling variants in RNA-seq experiments**. *J Anim Sci Biotechnol* 2019, **10**:44.
46. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD *et al*: **SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing**. *J Comput Biol* 2012, **19**(5):455-477.
47. Pryszcz LP, Gabaldon T: **Redundans: an assembly pipeline for highly heterozygous genomes**. *Nucleic Acids Res* 2016, **44**(12):e113.
48. Assefa S, Keane TM, Otto TD, Newbold C, Berriman M: **ABACAS: algorithm-based automatic contiguation of assembled sequences**. *Bioinformatics* 2009, **25**(15):1968-1969.
49. Treangen TJ, Ondov BD, Koren S, Phillippy AM: **The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes**. *Genome Biol* 2014, **15**(11):524.

Tables

Table1: Overview of the differences between the original annotation, the reannotation and BL21(DE3) annotation

	Original annotation (NZ_CP014268.2)	Reannotation	BL21(DE3)
Genome length(bp)	4,478,958		4,558,953
Plasmids		None	
G+C%	50.81%		50.83%
Genes(total)	4,628	4,577	4370
Protein_coding genes	4,258	4,223	4197
Pseudogenes	248	0	70
tRNAs	85	87	85
rRNAs	22	22	22
Miscellaneous RNAs[a]	15	245	66
Backbone genes	4,380(4,258 protein-coding genes,85 tRNA genes,22rRNAs and 15 misc RNAs)	4,577(4,223 protein-coding genes,87 tRNA genes,22rRNAs and 245 misc RNAs)	4,370(4,197 protein-coding genes,85 tRNA genes,22rRNAs and 66 misc RNAs)

a: The concept of miscellaneous RNA includes ncRNA ,tmRNA and all other ncRNAs.

Table 2: Statistical analysis of RNA-Seq data.

Group	Run	Raw sequences reads	Unidentified Reads	Human papillomavirus type 16 L1 Reads	Escherichia coli Reads	Mapping reads
B37	1	11,390,697	125,298 (1.1%)	0%	11,265,400 (98.9%)	11,144,771(97.84%)
	2	11,381,924	159,347 (1.4%)	0%	11,222,577 (98.6%)	11,114,562(97.65%)
	3	14,942,498	298,850 (2%)	0%	14,643,648 (98%)	14,625,644(97.88%)
Y37	1	14,146,289	2,263,406 (16%)	6,365,830 (45%)	5,517,053 (39%)	4,448,202(31.44%)
	2	12,607,439	2,143,264 (17%)	5,673,348 (45%)	4,790,827 (38%)	3,678,800(29.88%)
	3	11,260,708	1,689,106 (15%)	5,067,319 (45%)	4,504,283 (40%)	3,586,189(31.85%)

Additional Files

Additional file S1: TableS1. The list of the ruled-out genes in the reannotation.

Additional file S2: TableS2. The list of newly added protein-coding in the reannotation.

Additional file S3: TableS3. The list of complete CDSs in the reannotation.

Additional file S4: TableS4. Comparison with BL21(DE3) and ER2566 annotation.

Additional file S5: TableS5. The list of newly added miscellaneous ncRNA genes in the reannotation.

Additional file S6: Table S6. The transcription level of the newly added ncRNA.

Additional file S7: TableS7. The top 50 the highly expressed genes.

Additional file S8: An integrated proteogenomics search database of ER2566 strain.

FigureS1. Pairwise alignment and visualization.

Figures

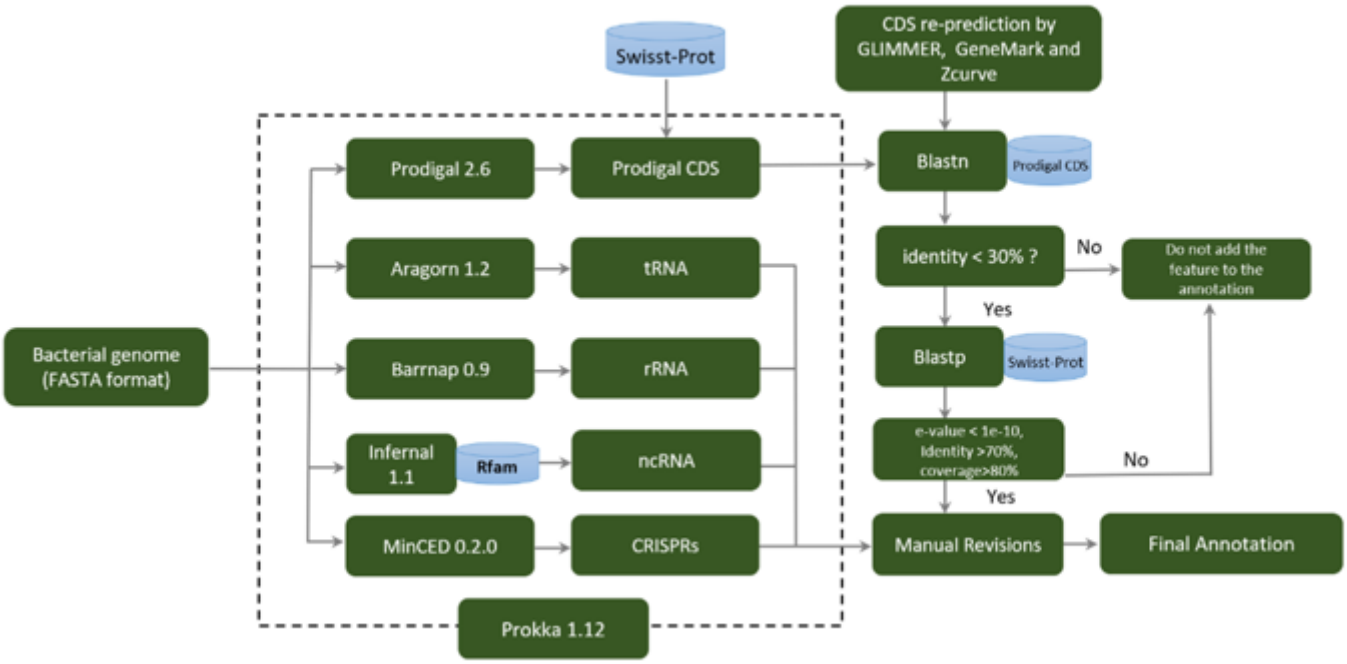

Figure 1

Flowchart depicting the pipeline and methods used for bacterial genome reannotation of the E.coli strain ER2566.

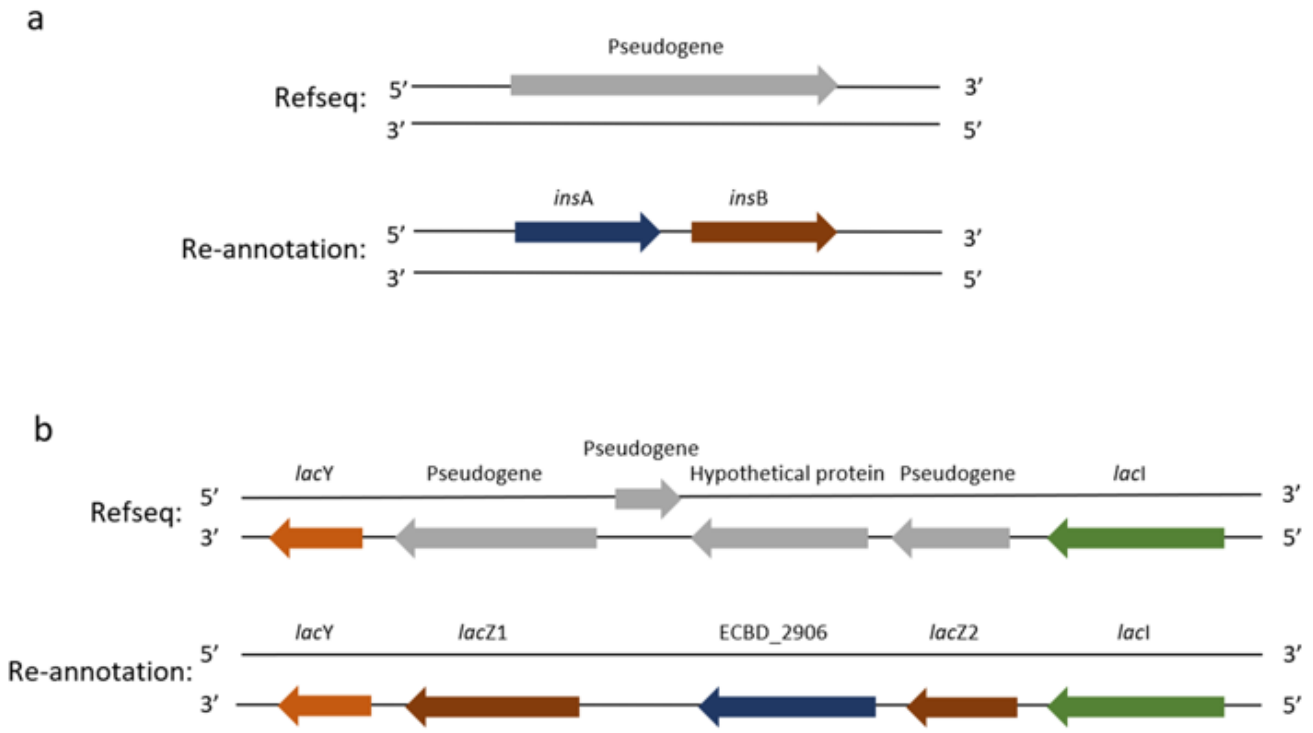

Figure 2

Examples of the differences between the original RefSeq annotation and our reannotation. a) In the reannotation, one pseudogene (RS16270) was identified as two genes, insA and insB, which show strong homology to the insertion element protein, IS1. B) In the reannotation, two pseudogenes were re-identified as two genes (lacZ1 and lacZ2), whereas the hypothetical protein was reannotated and shown to be highly homologous with the DNA-directed RNA polymerase gene ECBD_2906 from E.coli strain BL21-DE3.

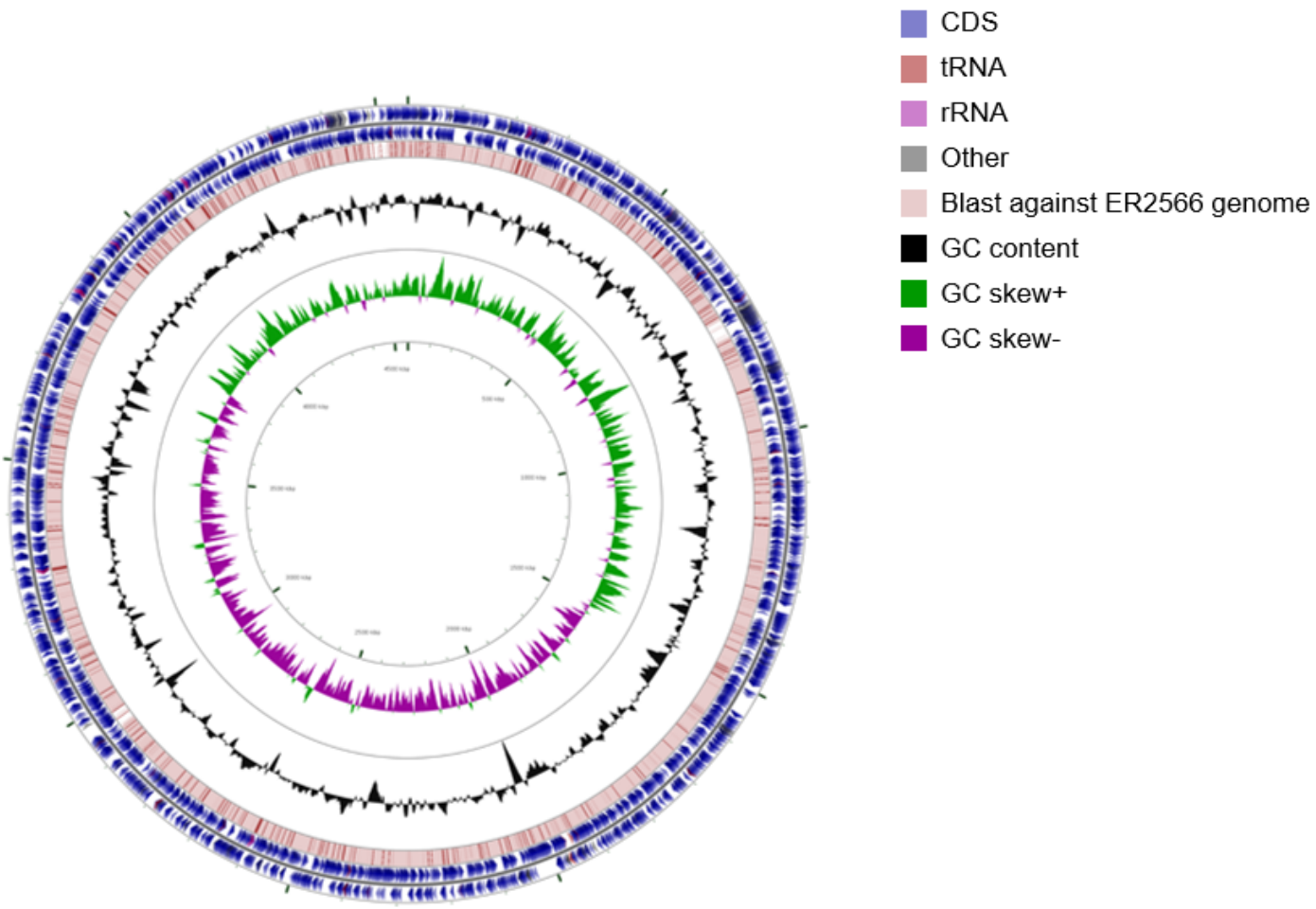

Figure 3

Comparison between BL21(DE3) genome and ER2566 genome. Viewing from outside to inside rings, the outermost two rings, respectively representing plus-strand and minus-strand, show features extracted from the BL21(DE3) genome GenBank file (GenBank: CP001509.3); the next ring shows the positions of BLAST hits between the BL21(DE3) genome and the ER2566 genome detected by blastn. The height of each line in the third ring showing BLAST results is proportional to the percent identity of the hit, and overlapping hits renders as darker lines. The next two rings show GC content and GC skew.

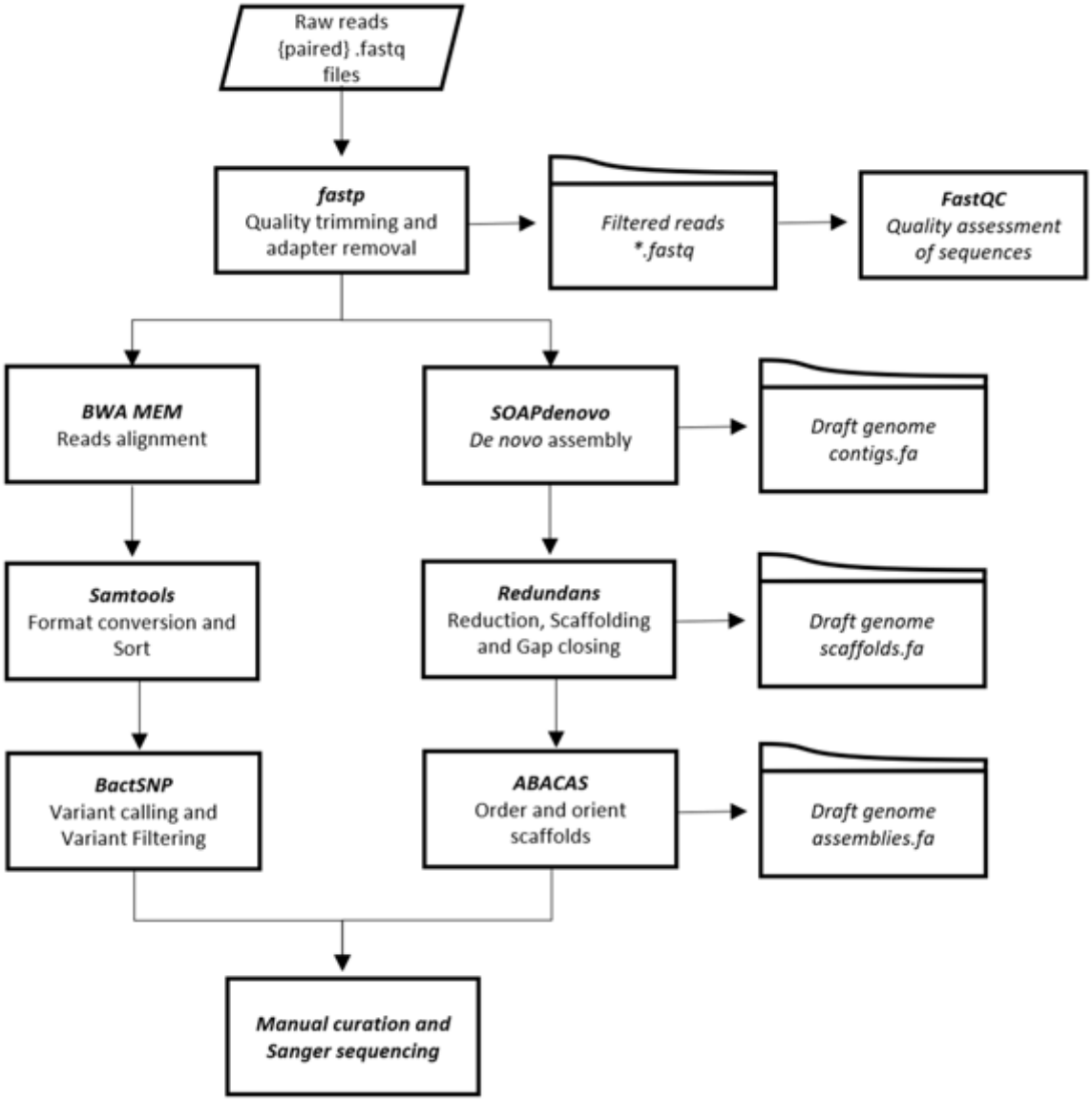

Figure 4

Flow-chart of variant calling, combining reads mapping and de novo assembly.

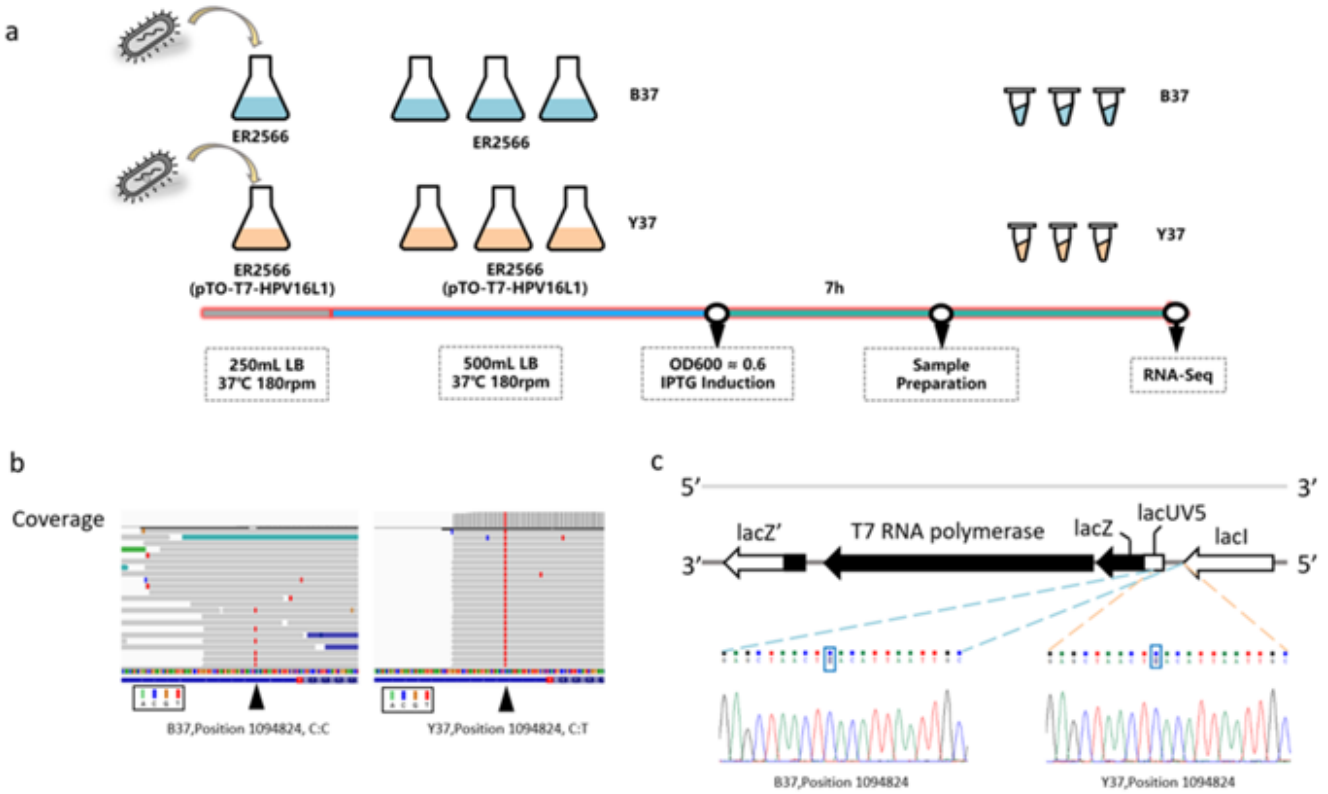

Figure 5

RNA-seq for variant calling under pressure from overexpression. a) The experimental design. Each group (B37, without plasmid; Y37, with pTO-T7 plasmid overexpressed) had three biological replicates. b) Visualization of BAM files of the B37 (left panel) and Y37 (right panel) in the Integrative Genomics Viewer. Based on the reannotation, one mutant was identified at position 10,948,24C>T, located within the 3’ non-coding region of the transcription factor gene lacI. c. Mutation detected by Sanger sequencing of the B37 and Y37 genomic samples.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AdditionalfileS2.xlsx
- FigureS1.jpg
- AdditionalfileS4.xlsx
- AdditionalfileS1.xlsx
- AdditionalfileS7.xlsx
- AdditionalfileS6.xlsx

- AdditionalfileS8.zip
- AdditionalfileS3.xlsx
- AdditionalfileS5.xlsx