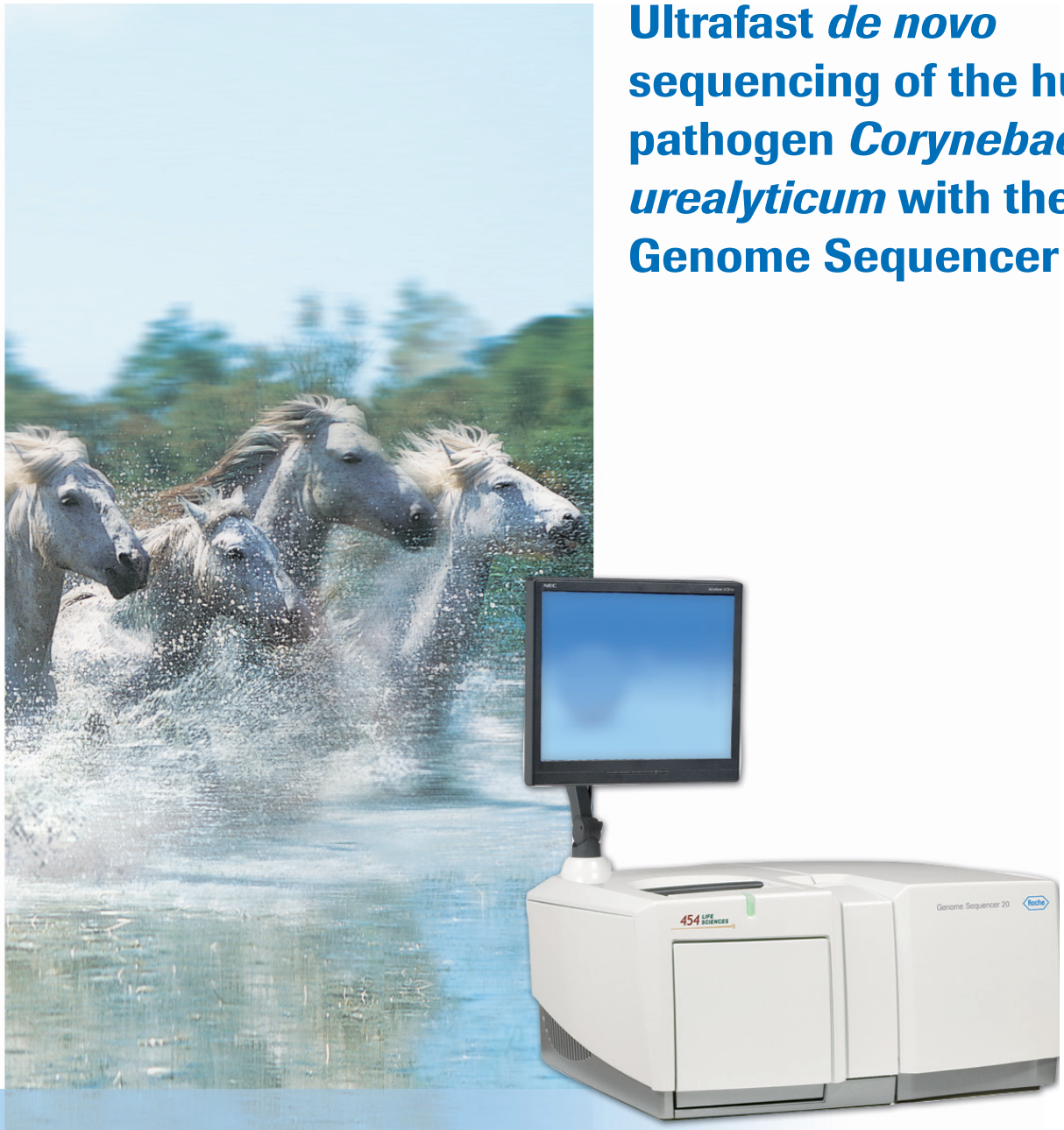
Roche

Diagnostics

Genome Sequencer System

Application Note No. 2/July 2006



Ultrafast *de novo* sequencing of the human pathogen *Corynebacterium urealyticum* with the Genome Sequencer System

www.roche-applied-science.com

454 LIFE SCIENCES

Ultrafast *de novo* sequencing of the human pathogen *Corynebacterium urealyticum* with the Genome Sequencer System

Andreas Tauch[1]*, Eva Trost[1], Thomas Bekel[2], Alexander Goesmann[2], Ulrike Ludewig[3] and Alfred Pühler[1]
[1] Institut für Genomforschung, CeBiTec, Universität Bielefeld, Bielefeld, Germany
[2] Bioinformatics Resource Facility, CeBiTec, Universität Bielefeld, Bielefeld, Germany
[3] Roche Applied Science, Penzberg, Germany
* Corresponding author: Andreas.Tauch@Genetik.Uni-Bielefeld.de

Introduction

A microbial genome sequence provides a wealth of data and specific information that cannot be obtained by other experimental approaches. A genome sequence can be regarded as the starting point for detailed bioinformatics analysis, metabolic reconstruction and systematic functional examination of all of the identified genes. In particular, genome sequencing has revealed important information on the genes and deduced proteins of human pathogens, including putative virulence factors and potential new drug targets. The most widely used strategy for the sequencing of a microbial genome is the direct shotgun approach with the application of Sanger technology, along with a small-insert clone library of the organism of interest [1]. An emerging alternative technology in genome sequencing by the direct shotgun approach is provided by the Genome Sequencer System. The special design of this revolutionary technology relies on small DNA fragments and adaptor sequences, rather than small-insert clone libraries, and highly parallel sequencing in picoliter-scale volumes. This results in an ultrafast sequencing process [2]. This sequencing strategy is therefore ideally suited for the rapid determination of genome sequences of hitherto uncharacterized human pathogens.

Materials and Methods

Genome sequencing and assembly

Materials

Equipment:
Genome Sequencer 20 Instrument (Software Version 1.0.52)

Reagents:
Sample Preparation: GS DNA Library Preparation Kit (5 µg sample DNA was processed in the standard DNA Library Preparation procedure); GS emPCR Kit I (Shotgun)
Sequencing: GS 20 Sequencing Kit; GS PicoTiterPlate Kit

A detailed list of all necessary equipment and reagents is provided in the Genome Sequencer 20 User's Manuals and Guides.

Methods

Preparation of genomic DNA
C. urealyticum DSM7109, originally isolated from a bladder stone was grown in brain heart broth supplemented with 1% (vol/vol) Tween 80. Genomic DNA was isolated from 8×10^9 cells and purified by an iterative phenol-chloroform extraction method [4]. This procedure yielded high-purity genomic DNA with a concentration of 1175 ng/µL and an A_{280}/A_{260} ratio of 1.89.

For full details on the preparation procedure and sequencing, please refer to the GS DNA Library Preparation User's Manual, the GS emPCR Kit User's Manual and the Genome Sequencer 20 Operator's Manual.

In this Application Note we describe the de novo sequencing of the human pathogen *Corynebacterium urealyticum* utilizing the Genome Sequencer System. *C. urealyticum* is frequently isolated from urine research samples of catheterized intensive care patients and causes urinary tract infections that are significantly associated with stone formation [3]. Moreover, *C. urealyticum* is multiply resistant to clinically relevant antibiotics and is often only susceptible to glycopeptides. The new sequencing approach provided, for the first time, detailed information on the genome of *C. urealyticum*. In combination with elaborated bioinformatics tools, the annotated genome sequence revealed valuable insights into the gene inventory of this emerging nosocomial pathogen.

Procedure

Bioinformatics

Post-Run Analysis: Flow-Data was assembled using the Assembly Software (Newbler Assembler) of the Genome Sequencer 20 Software Version 1.0.52 as described in the GS 20 Data Processing Software Manual.

Genome annotation with GenDB and SAMS

After *de novo* assembly of the *C. urealyticum* genome, the obtained sequence contigs were filtered according to their size. Subsequently, 69 contigs with a minimal length of 501 bp were chained together into a pseudochromosome in which the contigs were separated by 12-mer linkers (CTAGCTAGCATG) containing stop codons in all six reading frames. This pseudochromosome was automatically annotated by applying the GenDB platform [5] and a standard analysis pipeline: After gene prediction with a combined approach using GLIMMER and CRITICA, an automatic function prediction (METANOR) with a combination of standard bioinformatics tools, like BLAST, HMMer and InterPro, was performed for each identified protein-coding sequence. This approach led to consistent gene annotations, assigning gene names, gene products, EC numbers, functional protein categories (COGs), and other attributes.

Small sequence contigs (≤ 500 bp) were analyzed by using the Sequence Analysis and Management System (SAMS) which is based on GenDB. SAMS was originally designed and implemented for quality control of sequence data obtained during the high-throughput phase of genome sequencing projects. Similar to the analysis of potential coding regions predicted on a bacterial genome sequence, individual (small) sequences can be processed and annotated by using a bioinformatics pipeline analogous to that described above. Accordingly, the application of SAMS results in consistantly annotated small sequence contigs.

Results and Discussion

Sequencing of the *C. urealyticum* genome with the Genome Sequencer System yielded 657,410 sequence reads that were finally used for *de novo* genome assembly. By applying a contig length cut-off of 500 bp, a total of 2,294,755 bases were assembled into 69 contigs with an average contig size of 33,257 bp. Of these assembled bases, 2,291,059 (99.8%) have been determined with PHRED 40 quality, meaning that the accuracy of the base call is at least 99.99% at the respective position. The largest assembled contig has a size of 175,964 bp. In addition, 154 small contigs (≤ 500 bp) with an average length of 144 bp have been assembled, producing a total of 22,211 bp.

The contiguous sequences of the *C. urealyticum* genome were uploaded into SAMS and GenDB for rapid annotation of the sequence data [5]. Small contigs were analyzed with a set of bioinformatics tools included in SAMS, and the resulting observations were stored in the SAMS database to enable

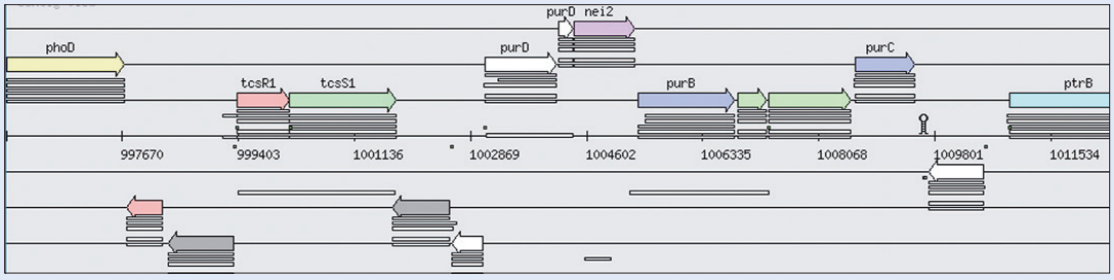
all the assembled sequence contigs to be evaluated. Although many of the small contigs revealed no significant hits in relation to public database entries, this automated analysis identified 57 partial protein-coding regions of which 25 (44%) apparently code for transposases of mobile genetic elements. The 69 large contigs were used for a precise gene prediction by the combined action of the software tools GLIMMER and CRITICA which are integrated in the GenDB platform, resulting in the prediction of 2027 protein-coding regions. Relevant data derived from this genome annotation are summarized in Table 1.

Feature	Value
No. of assembled bases	2,294,755
No. of assembled contigs (> 500 bp)	69
Mean G+C content of DNA	64.4%
No. of ribosomal RNAs	5
No. of transfer RNAs	46
No. of protein-coding sequences	2027 (100%)
No. of proteins homologous with proteins of *C. jeikeium*	1589 (78.4%)
No. of proteins non-homologous with proteins of *C. jeikeium*	438 (21.6%)
Coding density	90.2%
Mean size of coding sequences	1036 bp
Mean size of intergenic regions	136.7 bp

Table 1: General features of the *Corynebacterium urealyticum* genome.

To assess the quality of the *C. urealyticum* genome sequence by an additional bioinformatics approach, we utilized the automatic frameshift prediction performed by the CRITICA tool during the annotation process (Figure 1).

Among the 121 candidate regions predicted by CRITICA only seven turned out to contain a frameshift when considering the highly similar gene arrangement in the genome of the taxonomically closely related pathogen *Corynebacterium jeikeium* [6].



▷ **Figure 1:** Annotation of the *C. urealyticum* genome sequence with the GenDB platform. An excerpt of the genome annotation visualized by the GenDB system is shown. Arrows indicate the identified protein-coding regions. An apparent frameshift in the *purD* gene region associated with a homopolymer stretch of length seven has been detected automatically.

Interestingly, six potential frameshifts are apparently located in sequence regions that contain homopolymer stretches of length six and seven, respectively. These frameshifts might be due to the special detection of nucleotide incorporation during the sequencing process that is based on the release of pyrophosphate and the generation of photons. The number of incorporated nucleotides is thus indicated by the signal intensity which is, in principle, linear to at least homopolymers of length eight [2]. The low number of frameshifts that are apparently associated with homopolymers in the present study indicates the very high quality of the established contig sequences of the *C. urealyticum* genome, especially when considering the presence of 451 homopolymer stretches of length 6 to 13 within

the 69 contigs. Subsequently, functional annotation of the predicted coding regions was performed with the METANOR tool in such a way that orthology information from the *C. jeikeium* genome [6] was used to generate functional protein assignments. This bioinformatics approach allowed us to detect those proteins that are homologous in both corynebacteria and those that are only encoded by the *C. urealyticum* genome. The functional classification of the deduced proteins into Clusters of Orthologous Groups of proteins (COGs) [7] is shown in Figure 2A.

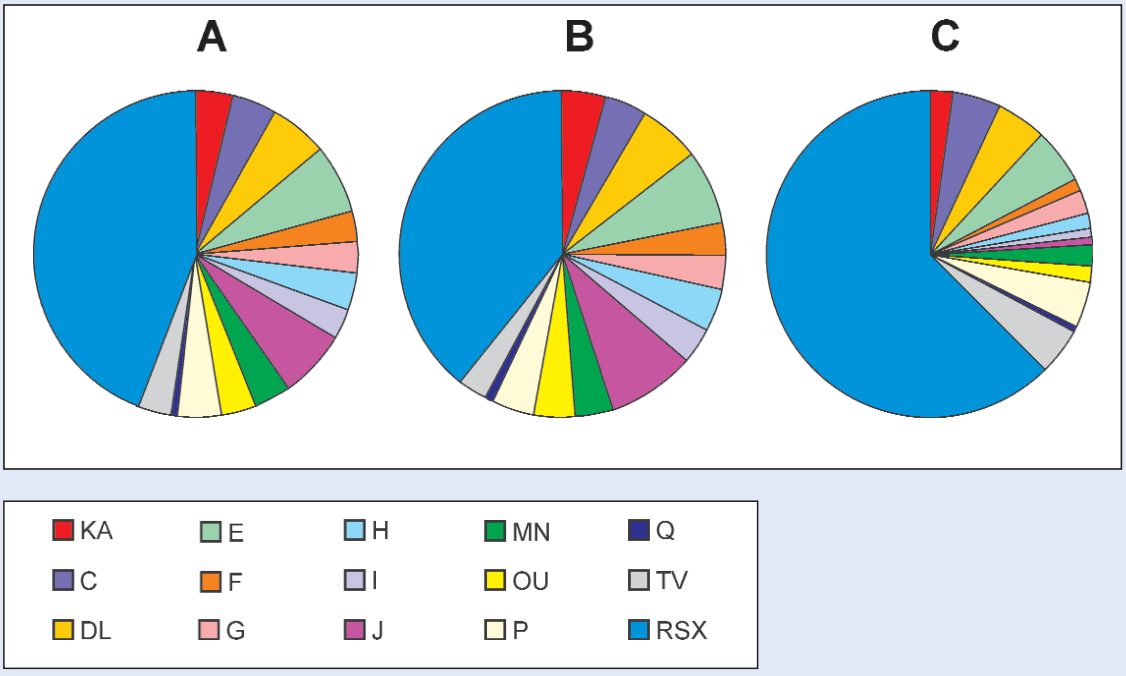


Figure 2: Functional classification of the predicted proteins of *C. urealyticum*. Distribution into COG classes of the 2027 predicted proteins of *C. urealyticum* (A) and of the subset of proteins sharing homology (1589; B) or lacking homology (438; C) with proteins of *C. jeikeium*. Functional categories are abbreviated as follows: KA, transcription and RNA modification; C, energy production; DL, cell cycle control, replication, recombination and repair; E, amino acid metabolism; F, nucleotide metabolism; G, carbohydrate metabolism; H, coenzyme metabolism; I, lipid metabolism; J, translation; MN, cell wall/membrane biogenesis and motility; OU, posttranslational modification and secretion; P, inorganic ion metabolism; Q, secondary metabolites catabolism; TV signal transduction and defense mechanisms; RSX, general function prediction only, function unknown, not in COGs.

Approximately three-quarters of the predicted proteins (1589) revealed significant similarity to proteins that are encoded by the *C. jeikeium* genome and are apparently shared between both corynebacteria (Figure 2B), whereas 438 proteins were non-homologous between both species (Figure 2C). Among the set of proteins lacking homology with proteins from *C. jeikeium* we identified the components of a microbial urease and accessory proteins which could, together, constitute a functional urease machinery involved in the utilization of urea as nitrogen source and in the concomitant splitting of urea. This enzymatic reaction could be responsible for alkalinization of the human urine, which, in turn could cause damage of epithelial cells of the urinary tract along with struvite stone formation. The respective proteins can therefore be regarded as prominent virulence factors of *C. urealyticum*.

Conclusions

This study clearly demonstrates that the Genome Sequencer System provides a powerful technology to determine high-quality *de novo* sequences of bacterial genomes without prior DNA cloning. In conjunction with elaborated bioinformatics platforms for genome annotation, this new sequencing approach enables researchers rapidly to decipher and analyze the gene inventory of previously uncharacterized human pathogens. It may soon be possible to gain important insights into the lifestyle and pathophysiology of *C. urealyticum* and the molecular mechanisms of multidrug resistance from in-depth analyses of the annotated genome sequence.

References

[1] Frangeul L (1999) Microbiology 145: 2625-2634
[2] Margulies M (2005) Nature 437: 376-380
[3] Funke G (1997) Clin Microbiol Rev 10: 125-159
[4] Tauch A (1995) Plasmid 33: 168-179
[5] Meyer F (2003) Nucleic Acids Res 31: 2187-2195
[6] Tauch A (2005) J Bacteriol 187: 4671-4682
[7] Tatusov RL (1997) Science 278: 631-637

For more information, visit
www.genome-sequencing.com

NOTICE TO PURCHASER
RESTRICTION ON USE: Purchaser is only authorized to use the Genome Sequencer 20 Instrument with PicoTiterPlate devices supplied by 454 Life Sciences Corporation and in conformity with the procedures contained in the Operator's Manual.

Trademarks
PICOTITERPLATE is a trademark of 454 Life Sciences Corporation, Branford, CT, USA.

04918312001(1)0706

454 LIFE SCIENCES

Roche Diagnostics

Roche Diagnostics GmbH
Roche Applied Science
68298 Mannheim
Germany
www.roche-applied-science.com