A Draft de Novo Genome Assembly
of the PCB-Degrader Pseudomonas
Pseudoalcaligenes KF707

CALABRESE C (1), SIMONE D (2), CALABRESE F M (2), PIREDDA R (2),
CAPPELLETTI M (3), TRISCARI BARBERI T (3), FEDI S (3),
ATTIMONELLI M (2), ZANNONI D (3), TURNER R J (4)

(1) Department of Gynecological, Obstetrics and Pediatric Sciences, Unit of Medical Genetics,
Policlinico S. Orsola-Malpighi, University of Bologna, Italy
(2) Department of Biochemistry and Molecular Biology ‘E. Quagliariello’
University of Bari, Italy
(3) Department of Experimental Evolutionary Biology,
Unit of Applied and General Microbiology, University of Bologna, Italy
(4) Department of Biological Sciences & Biofilm Research Group
University of Calgary, Calgary, Alberta, Canada

Motivation
The study of Pseudomonas pseudoalcaligenes KF707 has provided new interesting ins-
sights on the use of bacterial strains for bioremediation(1). Indeed, KF707 is a Gram-
negative aerobic bacterium able to degrade polychlorinated biphenyls (PCBs)(2) and to
tolerate several toxic metal(loid)s(1). Both chemotaxis and biofilm formation have been
recently analysed and they have been described to have an effect on these degradation
properties. According to this, the goal of the project is the whole sequencing and the
de novo complete assembly of P. pseudoalcaligenes KF707 genome in order to obtain
information about the genetic basis for these peculiar physiological aspects.

Methods
The whole P. pseudoalcaligenes genome was sequenced with the 454 shotgun se-
quencing, performed on the GS-FLX instrument, using 1/4 of the plate. Preliminary as-
sessments of the length distribution and the sequencing quality were made. Software
systems of Newbler (v. 1.51, Roche 454 Life Sciences) and Abyss(4) (v 1.2.4) were used
in parallel to assemble the 454 reads in contigs. Statistical estimation of per-contig
genomic copy number (g.c.n.) was performed on 655 (out of 729) Newbler contigs,
with a length 500bp and a mean per-contig read depth of at least 5.6x. For this aim a
pipeline implemented in R was used, based on a kernel density estimation of the per-
contig median read depths distribution(5). Ab initio annotation of protein coding genes
was performed on both Newbler and Abyss contig sets, using RAST(6) (Rapid Annotation
using Subsystems Technology), which implements a propagation strategy based on
manually curated subsystems and subsystem-based protein families that automatically
guarantees a high degree of assignment consistency.

**Results**

From the sequencing 213,206 reads were obtained (7.7Mb), with a length ranging from 60 to 540 nt and a modal value of ~370 nt. Newbler assembler produced 729 contigs ranging in length from 500bp to 51,361bp (mean = 8303bp, total bases assembled = 6,053,515bp) with a N50* of 14148. Abyss assembler produced 4120 contigs ranging in length from 100bp to 13633bp (mean = 814, total bases assembled = 5,960,640) with a N50 of 2449. A high-quality assessment of 5694 gene functions and an initial metabolic reconstruction were obtained by using RAST. Each g. c. n. was calculated by dividing per-contig median read depth with the upper limit (= 7.4x) of the Confidence Interval (CI) of the tallest peak of the kernel density; contigs with g. c. n > 2 were considered regions repeated in the genome. On these bases, three sets of contigs were identified: a) 639 Newbler contigs (562bp to 51,361bp; mean length= 9,188bp) non-repeated in the P. pseudoalcaligenes genome; b) 8 contigs repeated from three to seven times and c) 8 contigs repeated less than three times. The g.c.n. predicted on set (c) of contigs were not as robust, because both their CIs overlap with some other CI central values of non-repeated contigs and RAST did not find any prediction of transposase or ribosomal subunits within their sequence, as it did for the other set of 8 contigs. Real time PCR will be used to validate this in silico estimation. Further steps of the project will be focused on contigs gaps filling by comparing Newbler and Abyss assemblies to obtain longer contigs: our preliminary data suggest that the two contig sets share about 5,76 Mb. Contigs obtained by merging the two sets will be used again for g.c.n. predictions and protein coding genes annotations and for the optical mapping strategy, to perform the final scaffolding. Any short physical gap in the genomic map will be solved by using PCR amplification with primers designed on neighbouring scaffold boundaries.

**References**


Contact email
dome.simone@gmail.com