Supplementary Material: Large Scale Read Classification for Next Generation Sequencing

James M. Hogan[1*] and Timothy Peut[1]
[1] School of EECS, Queensland University of Technology, Brisbane, Australia.
j.hogan@qut.edu.au, tim@timpeut.com

Abstract

Next Generation Sequencing (NGS) has revolutionised molecular biology, resulting in an explosion of data sets and a pressing need for rapid identification as a prelude to annotation and further analysis. NGS data consists of a substantial number of short sequence reads, given context through downstream assembly and annotation, a process requiring reads consistent with the assumed species or species group. Highly accurate results have been obtained for restricted sets using SVM classifiers, but such methods are difficult to parallelise and success depends on significant attention to feature selection. This work examines the problem at very large scale, using a mix of synthetic and real data with a view to determining the overall structure of the problem and the effectiveness of parallel ensembles of simpler classifiers (principally random forests) in addressing the challenges of large scale genomics.

1 Introduction

This document provides a list of sequences used in the study *Large Scale Read Classification for Next Generation Sequencing*, [Proceedings of the 14th International Conference on Computational Science, Cairns, 2014]. The list provided is that of the base or reference sequences during our analysis, these being origin sequences for many simulated reads using ART (see the main paper). Full details of the list are provided over the page.

Please see the main paper for any further technical details.

* Author to whom all correspondence should be addressed.

Reference ID	Organism	Num Generated (for 130 projects)	Class
NC_002927.3	Bordetella bronchiseptica RB50 chromosome, complete genome	2	Negative
NC_009495.1	Clostridium botulinum A str. ATCC 3502 chromosome, complete genome	2	Negative
NC_022121.1	Chlamydia trachomatis strain J/31-98, complete genome	2	Negative
NC_014121.1	Enterobacter cloacae subsp. cloacae ATCC 13047 chromosome, complete genome	2	Negative
NC_004431.1	Escherichia coli CFT073 chromosome, complete genome	11 / 65	Negative / Positive
NC_012973.1	Helicobacter pylori B38 chromosome, complete genome	2	Negative
NZ_CM001376.1	Jonquetella anthropi DSM 22815 chromosome, whole genome shotgun sequence	2	Negative
NC_000962.3	Mycobacterium tuberculosis H37Rv complete genome	2	Negative
NC_002946.2	Neisseria gonorrhoeae FA 1090 chromosome, complete genome	2	Negative
NC_003116.1	Neisseria meningitidis Z2491 chromosome, complete genome	2	Negative
NC_008463.1	Pseudomonas aeruginosa UCBPP-PA14 chromosome, complete genome	2	Negative
NC_002952.2	Staphylococcus aureus subsp. aureus MRSA252 chromosome, complete genome	65 / 11	Positive / Negative
NC_012121.1	Staphylococcus carnosus subsp. carnosus TM300 chromosome, complete genome	4	Negative
NC_007168.1	Staphylococcus haemolyticus JCSC1435 chromosome, complete genome	4	Negative
NZ_GL545260.1	Staphylococcus hominis subsp. hominis C80 genomic scaffold supercont1.9, whole genome shotgun sequence	4	Negative
NC_004350.2	Streptococcus mutans UA159 chromosome, complete genome	4	Negative
NC_008533.1	Streptococcus pneumoniae D39 chromosome, complete genome	4	Negative
NC_004070.1	Streptococcus pyogenes MGAS315 chromosome, complete genome	4	Negative
NC_007350.1	Staphylococcus saprophyticus subsp. saprophyticus ATCC 15305, complete genome	4	Negative
NC_020164.1	Staphylococcus warneri SG1, complete genome	4	Negative
NC_005810.1	Yersinia pestis biovar Microtus str. 91001 chromosome, complete genome	2	Negative