GIs/non-GIs dataset from 118 bacterial genomes

Starting from 675 complete bacterial genome sequences, Langille et al. used stringent but potentially flexible criteria with distance cutoffs to select query genomes containing a sufficient number of suitably related species or strains for GI analysis [1]. They identified some regions conserved across all genomes as negative dataset and constructed standard dataset to investigate the accuracy of several sequence composition-based GI prediction tools. There exist 771 positive GIs and 3770 negative GIs (non-GIs) whose lengths varied from 8 kb to 31 kb. Given that these GIs and non-GIs were obtained from 118 genomes, representative species genomes from the domains of *Bacteria* and *Archaea*. Thus, this dataset should not be overly biased.