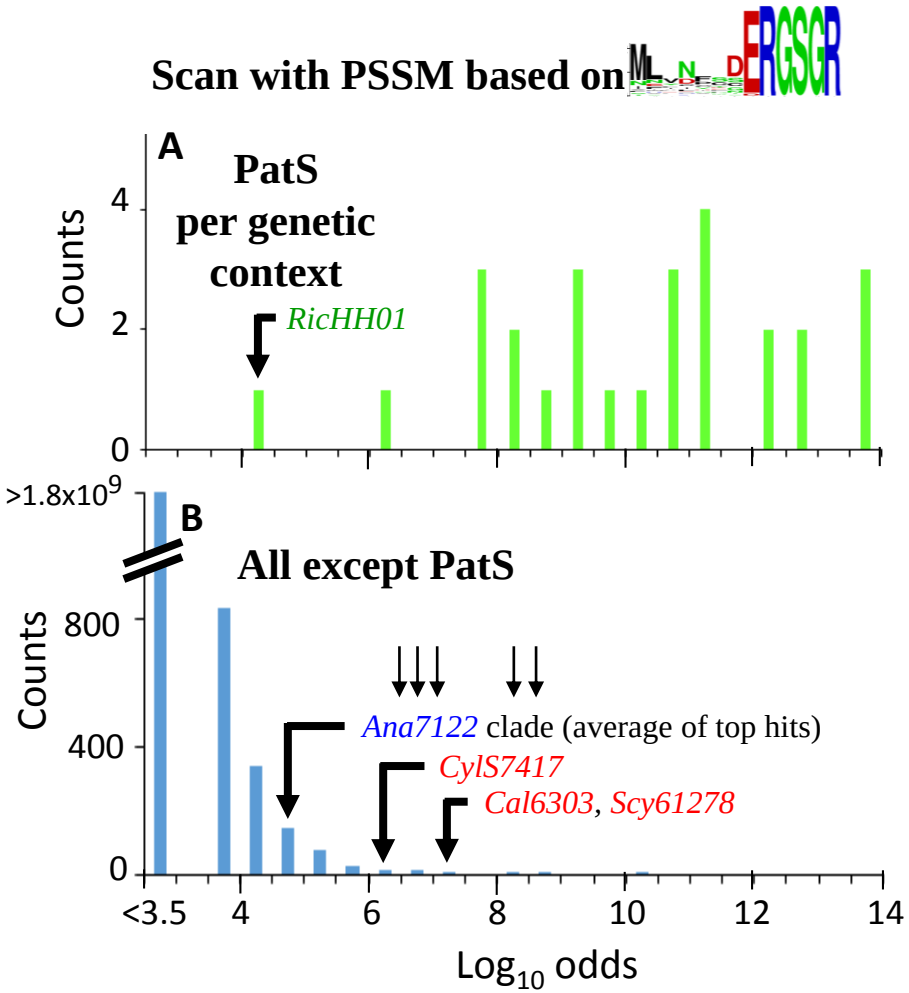
Supporting Information Figure S5: Scan of genomes with PatS- and PatX-based PSSMs

Figure S5 A-B: Scan of genomes of heterocyst-forming strains with PatS-based PSSM



Position-specific scoring matrices (PSSMs) were constructed as described in Methods, from aligned portions of PatS and PatX proteins (graphical representation from Fig. 8), using all proteins from the target organisms as the source of overall amino acid frequencies. The PSSMs were applied to the genomes of all heterocyst-forming (**A-F**) or filamentous (**G-J**) cyanobacteria, translated in all six reading frames. Scores represent a ratio between the likelihoods that a given amino acid sequence would be produced from (a) the frequencies of the aligned PatS or PatX sequences and (b) the frequencies of amino acids in all proteins produced by the target organism. The ratio is expressed as a base-10 logarithm. Scores for PatS (**A**) and PatX (**C,E,G,I**) were computed using PSSMs derived from training sets excluding the sequence from the target organisms and related organisms, to simulate the process of discovering the PatS and PatX sequences if they were not already known. When these sequences are not excluded, scores for PatS and PatX sequences are one to two orders of magnitude higher. Scores for sequences not already identified as PatS and PatX by genetic context used (**B**, **D**, **F, H**, **J**) were computed using the entire PatS and PatX training sets. Scores were filtered to exclude those below 3.5 (**A**, **B**, **E-J**) or 1.5 (**C** and **D**). Also indicated (**B**, **D**, and **F**) are the scores for candidate PatS or PatX proteins not determined by genetic context (in red), plus those of alternative PatS or PatX proteins (small arrows). See Supporting Tables 8 – 10 for underlying data.

Figure S5 C-F: Scan of heterocyst-forming strains with PatX-based PSSMs

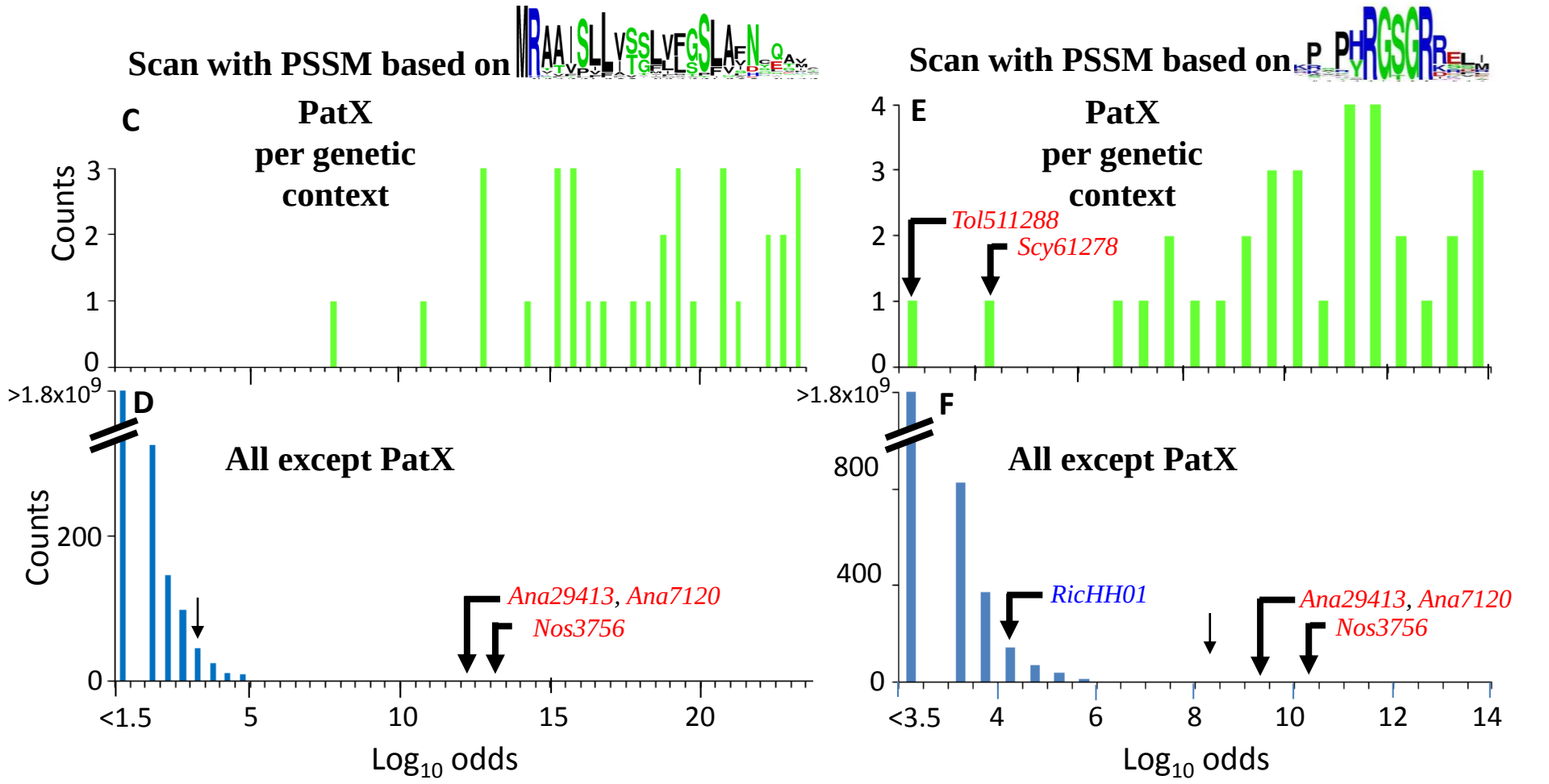


Figure S5 G-J: Scan of filamentous strains with PatX-based PSSMs