



**Figure S4.** Distribution of sequence identities and size of the sequence sets studied. A. Training sets. B. Test sets. C. *Shewanella* sequence sets.