Protein Sequences from 100 Phylogenetically Diverse Genomes
N=313,139

Calculate Pairwise Sequence Identity (BLASTP, e10^-10, 80% coverage)

Homology Clustering N=23,336 (MCL)

Identification of Widely Distributed SFams N=720

Family Size >= 250

Diverse Representative Sequence Selection

Multiple Sequence Alignment (MUSCLE)

Construction of Widely Distributed SFam HMMs (HMMBUILD) N=720

Family Size < 250

New Genomes N=1,894

Extract Protein Sequence N=7,165,275

Screen for SFam Homologs (HMMSCAN, e10^-5, 80% coverage) N=5,624,246

Add Homologs to SFams N=1,541,029

Calculate Pairwise Sequence Identity (BLASTP, e10^-5, 80% coverage) N=5,624,246

Homology Clustering (MCL) N=344,921

Family Size >= 250

Diverse Representative Sequence Selection

Multiple Sequence Alignment (MUSCLE)

Construction SFam HMMs (HMMBUILD) N=344,921

Family Size < 250