1. Predict Protein Sequences from genomes or metagenomes (SEED and GenDB developers)

2. Annotations and Subsystem Characterizations (SEED Annotators)

3. Compare all predicted proteins with all other predicted proteins (precalculated in SEED)

4. Find predicted proteins in subsystems or predicted proteins with homologs (E < 10^{-20}) in subsystems

5. Classify protein, homolog, domain of homolog, and subsystems for each predicted protein in Sargasso, AMD, and SEED-nr databases

6. Choose 10,000 samples with replacement from each database. Calculate the difference for each subsystem. Repeat 10,000 times and report the median differences

7. Choose 20,000 samples at random from both databases to generate two mixed populations and calculate the difference. Repeat 10,000 times and obtain the range of differences

8. Compare the median of #6 with the range from #7 for a given confidence interval to determine whether it is significant