Whole genome sequencing of 24 strains

Identification of 47,997 coding sequences

Coding sequences separated into 3100 clusters based on 70% identity of 70% of length

Clusters separated into 4536 subclusters based on 95% identity of 95% of length

Longest sequence in each subcluster is chosen for probe design

~3 probes created for each core gene and ~13 probes for each distributed gene (31307 total)

Chips produced by Nimblegen with all probes in duplicate