

QC Filtering Flowchart

Start
3088 genomes
with GO data

Completeness >90
NaN→0: 1
Lost: 259
Remaining: 2829



Contamination <5
NaN→100: 353
Lost: 543
Remaining: 2286



Genes >0
Lost: 0
Remaining: 2286



Valid Environment
(not null/empty/"Unclassified")
Lost: 77
Remaining: 2209

High Quality Dataset
2209 genomes
11 environments