Figure: Stacked area chart showing count of sequences across data curation stages.

Legend: **Dataset** — zuber, archiveII, Multi-databases, RNAstralign, bpRNA, Ribonanza

Left-axis labels (initial stage):
- Total (N = 951,465)
- Ribonanza (N = 806,573)
- bpRNA (N = 102,318)
- RNAstralign (N = 37,149)
- archiveII (N = 3,975)
- zuber (N = 1,450)

Right-axis labels (final stage):
- Total (N = 158,599)
- Ribonanza (N = 108,412)
- bpRNA90 (N = 24,833)
- RNAstralign (N = 24,213)
- Multi-databases (N = 3,392)
- archiveII (N = 977)
- zuber (N = 164)

Y-axis: Count of sequences (100 to 1M, log scale)

X-axis (Data curation stage): Fetched from online repository, Reads > 500 and S/N > 1, AUROC, BLAST, Duplicates, Non-ACGU characters, Sequences below 10nt, No structure, Redundancy across databases