Step 1a: Download nucleotide database (nt.fa)

NCBI

nt.fa

> gi | 407466711
AGCTTTTCATTCTGACTGCAACGG...

> gi | 209395693
TTTAAAGAGACCGGCGATTCTAGT...

> gi | 9629367
AGCGAAAGCAGGTCAATTATATTC...
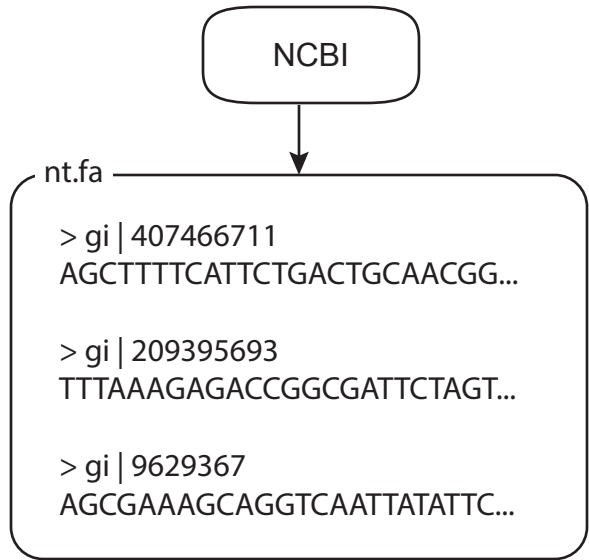
Step 1b: Obtain organism, taxon ID, and gi relationships

PathoDB   -or-   NCBI

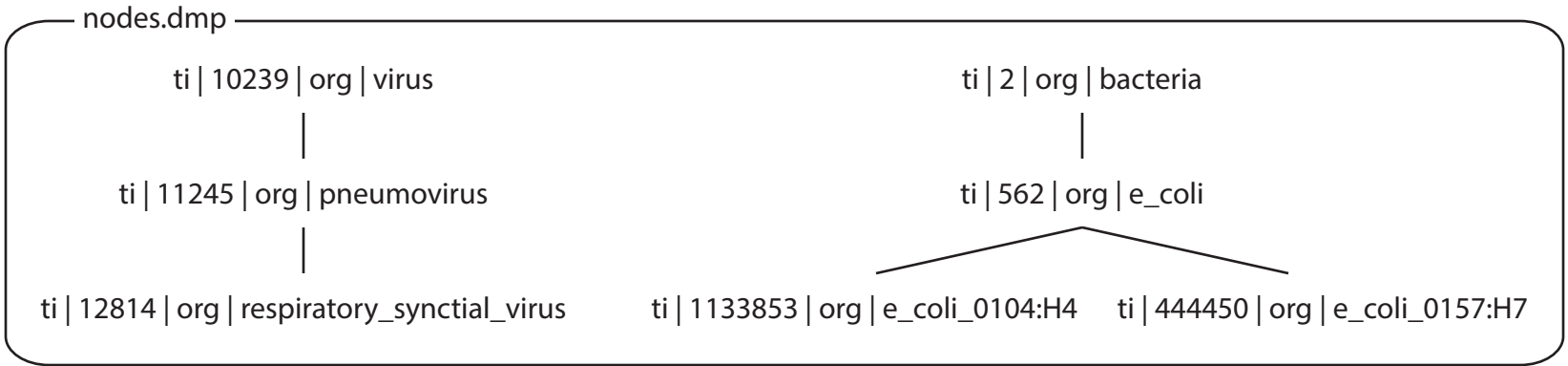| org	ti	gi
E. coli 0104:H4	1133853	407466711
E. coli 0157:H7	444450	209395693
E. coli 0157:H7	444450	209157093
Respiratory syncytial virus	12814	9629367
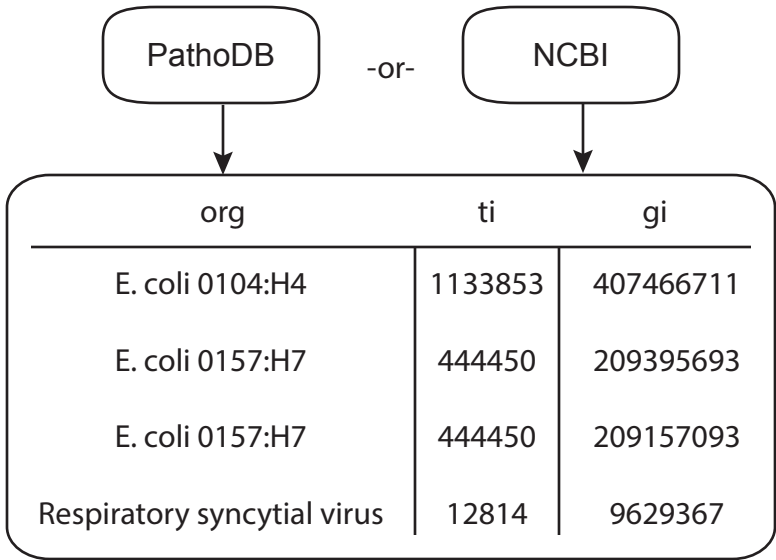
Step 2: Append taxon ID and organism to all sequence headers (nt_ti.fa)

nt_ti.fa

> ti | 1133853 | org | e_coli_0104:H4 | gi | 407466711
AGCTTTTCATTCTGACTGCAACGG...

> ti | 444450 | org | e_coli_0157:H7 | gi | 209395693
TTTAAAGAGACCGGCGATTCTAGT...

> ti | 12814 | org | respiratory_syncytial_virus | gi | 9629367
AGCGAAAGCAGGTCAATTATATTC...
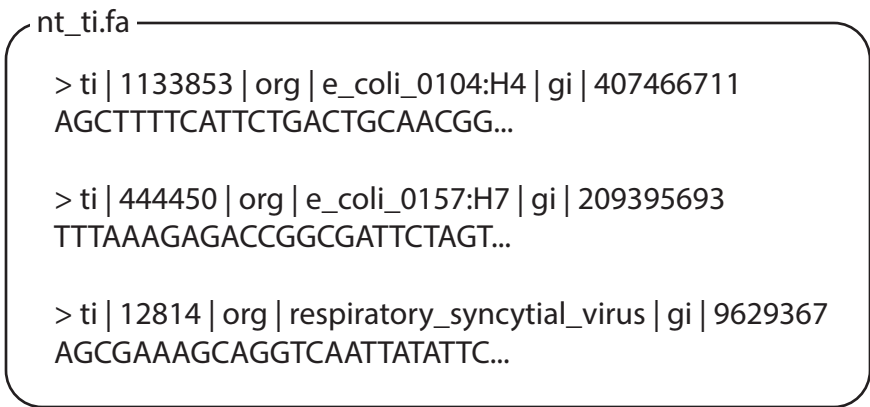
Step 3: Download NCBI taxonomy tree (nodes.dmp)

nodes.dmp

ti | 10239 | org | virus
|
ti | 11245 | org | pneumovirus
|
ti | 12814 | org | respiratory_synctial_virus

ti | 2 | org | bacteria
|
ti | 562 | org | e_coli
ti | 1133853 | org | e_coli_0104:H4   ti | 444450 | org | e_coli_0157:H7

Step 4: User input taxon IDs

10239 (all viruses)

113385 (E. coli O104:H4)
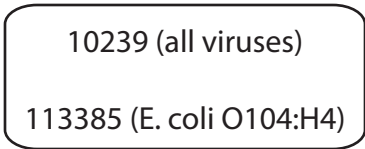
Step 5: Extract target library

Target.fa

> ti | 1133853 | org | e_coli_0104:H4 | gi | 407466711
AGCTTTTCATTCTGACTGCAACGG...

> ti | 12814 | org | respiratory_syncytial_virus | gi | 9629367
AGCGAAAGCAGGTCAATTATATTC...