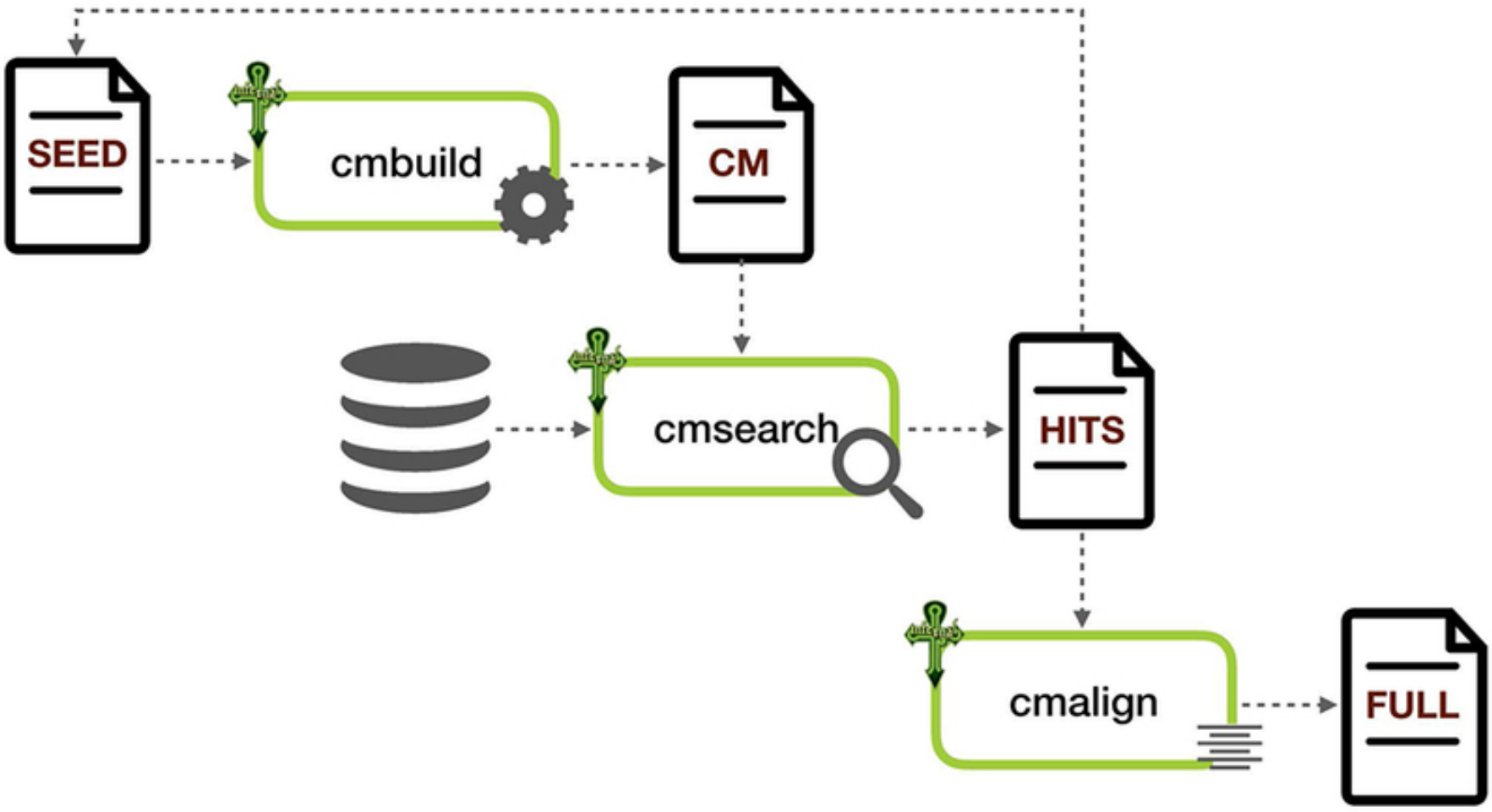
RNAcentral uses Rfam to annotate sequences

- 16.5 million sequences
- 50% are from Rfam
- 40% have Rfam hits
- 2% are potential sources of new Rfam families

Rfam: RNA families database (3016 families)

- Each family is represented by:
  - representative SEED alignment annotated with secondary-structure
  - covariance model (CM) built from the SEED
  - hits in Rfamseq database above GA threshold (FULL)



Rfamseq: switch from subset of ENA to genome-centric database

- about 8000 reference genomes
- reduces redundancy
- more scalable

Rfamseq: switch from subset of ENA to genome-centric database

- about 8000 reference genomes
- reduces redundancy
- more scalable

Rfamseq: genome-centric database means less flexible SEEDs

- previous requirement: SEED sequences must be in Rfamseq
- new approach allows any GenBank or RNAcentral sequence
- verification of sequences utilizes GenBank and RNAcentral API

Future directions for Rfam

- improve Rfam families based on crystal structures
- synchronize with mirBase
- use model reference coordinates to annotate important features