Figure 1: K-mer homology is an alignment-free and reference-independent method for cross-species analyses

A

Rhinolophus sinicus

Mus musculus

Homo sapiens

Time (MYA)

~100 80 60 40 20 0

Lung

Single-cell RNA-seq

B

Uniformly subsample biochemical synonyms from sequencing data

Identify interesting biochemical synonyms

Query database for definitions of interesting biochemical synonyms

Subsample amino acid k-mers from each transcriptome

Find enriched k-mers

Query for functional annotation

Transcriptomes
Species A

Species B

leaftea

Protein-coding

Input reads

Noncoding

Categorize on biochemical properties

Uniformly subsample k-mers

Reduced amino acid alphabet *k*-mers

— Nucleotide
∿ Protein
∿ Categorized protein

Build cell type database tree

Query unannotated species' cells against reference

Cell type enriched differential *k*-mers

Interesting k-mer   *Putative homologs not identifiable by traditional methods*

*Not in genome assembly
genome

In reference genome
genome

Query protein seq of reads with mystery k-mers for annotation

*Not in annotated gene
gene

In annotated gene
gene

*Unknown homolog

Known homolog