Supplemental Figure 3
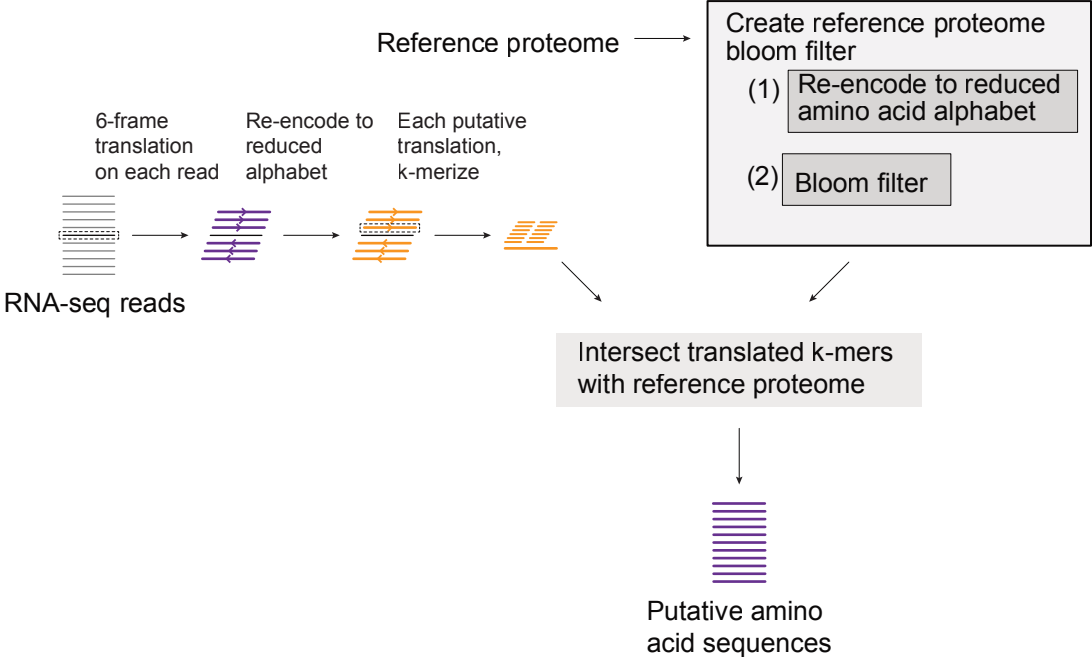
A 1. Extract cell barcodes with sufficient nUMI from single-cell bam via `bam2fasta`

2. Predict amino acid sequence from RNA-seq reads via `khtools extract_coding`

Reference proteome ⟶ Create reference proteome bloom filter
(1) Re-encode to reduced amino acid alphabet
(2) Bloom filter

6-frame translation on each read → Re-encode to reduced alphabet → Each putative translation, k-merize

RNA-seq reads

Intersect translated k-mers with reference proteome

Putative amino acid sequences

3. Randomly subsample amino acid k-mers via MinHash algorithm in `sourmash sketch`

(b) `sourmash sketch`   MinHash protein k-mers

(c) `sourmash compare`   Compute cell-cell similarity

(d) `sourmash index`   Build sequence bloom tree (SBT)

(e) `sourmash knn`   Build nearest neighbor graph

(f) `sourmash umap`   Build UMAP off of knn

B .bam + .bed
(if input is bam, convert to fastq)

1. `samtools fastq` → FASTQs

2. `fastp` → Adapter, poly-A, trimmed

3. `khtools extract_coding` → Predicted amino acid sequence

UNIPROT/SwissProt manually curated eukaryotic protein sequences

4. `diamond blastp` →
XP_008705901.1 PREDICTED: phosph
XP_026370896.1 phospholipase A2
XP_011803952.1 PREDICTED: phosph
NP_001177150.1 cytochrome c oxid
XP_003419308.1 cytochrome c oxid

Putatitive homologs

NCBI RefSeq non-redundant protein

C

Non-coding / Coding / Stop codon in all translation frames / All translations shorter than peptide k-mer size + 1 / Read too short / Low complexity peptide — dot plots by species (Human, Chimpanzee, Bonobo, Gorilla, Orangutan, Macaque, Mouse, Opossum, Platypus, Chicken); x-axis Percentage.