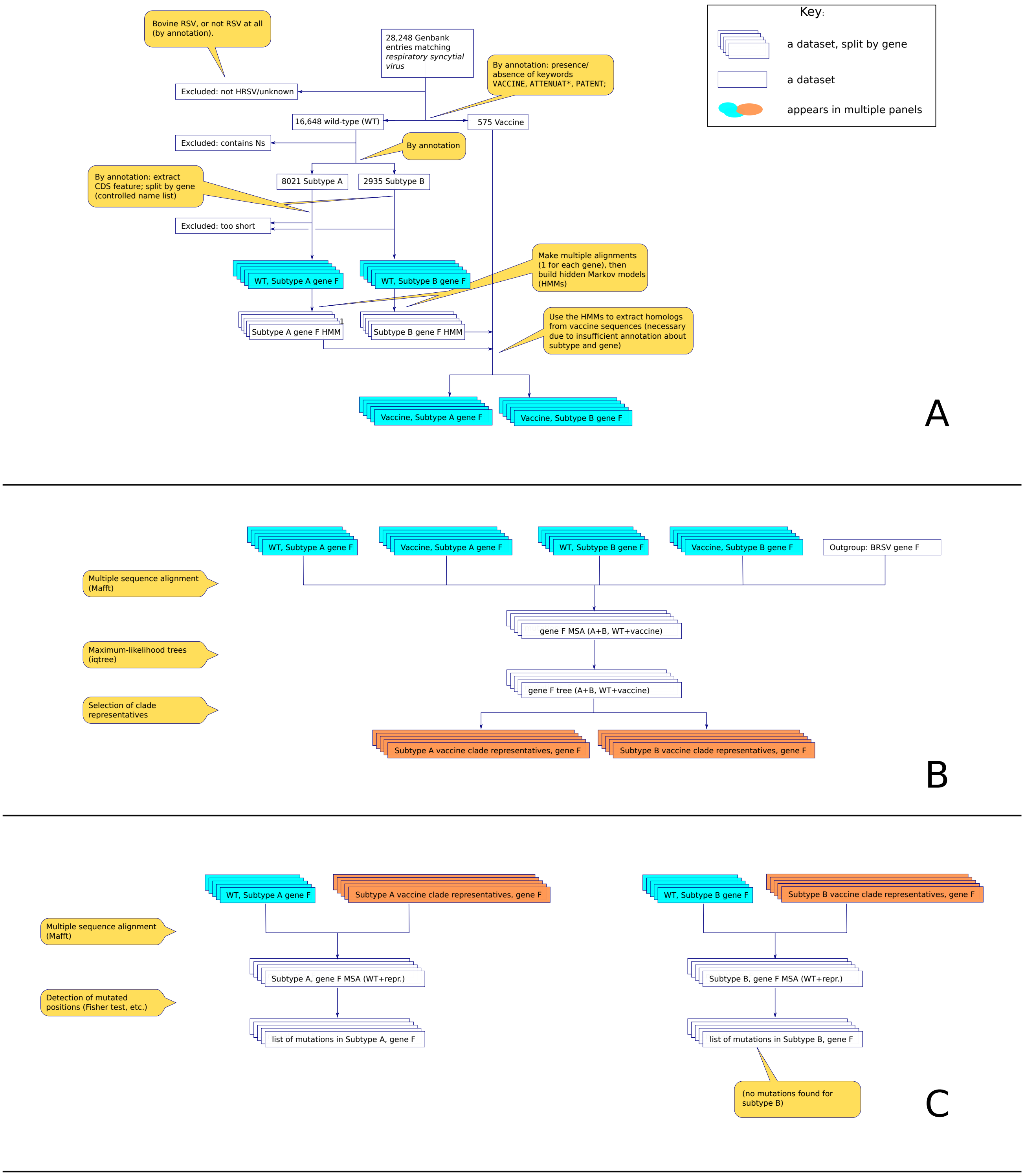
Key:
a dataset, split by gene
a dataset
appears in multiple panels
Bovine RSV, or not RSV at all (by annotation).
28,248 Genbank entries matching *respiratory syncytial virus*
Excluded: not HRSV/unknown
By annotation: presence/ absence of keywords VACCINE, ATTENUAT*, PATENT;
16,648 wild-type (WT)
575 Vaccine
Excluded: contains Ns
By annotation
By annotation: extract CDS feature; split by gene (controlled name list)
8021 Subtype A
2935 Subtype B
Excluded: too short
WT, Subtype A gene F
WT, Subtype B gene F
Make multiple alignments (1 for each gene), then build hidden Markov models (HMMs)
Subtype A gene F HMM
Subtype B gene F HMM
Use the HMMs to extract homologs from vaccine sequences (necessary due to insufficient annotation about subtype and gene)
Vaccine, Subtype A gene F
Vaccine, Subtype B gene F
A
WT, Subtype A gene F
Vaccine, Subtype A gene F
WT, Subtype B gene F
Vaccine, Subtype B gene F
Outgroup: BRSV gene F
Multiple sequence alignment (Mafft)
gene F MSA (A+B, WT+vaccine)
Maximum-likelihood trees (iqtree)
gene F tree (A+B, WT+vaccine)
Selection of clade representatives
Subtype A vaccine clade representatives, gene F
Subtype B vaccine clade representatives, gene F
B
WT, Subtype A gene F
Subtype A vaccine clade representatives, gene F
WT, Subtype B gene F
Subtype B vaccine clade representatives, gene F
Multiple sequence alignment (Mafft)
Subtype A, gene F MSA (WT+repr.)
Subtype B, gene F MSA (WT+repr.)
Detection of mutated positions (Fisher test, etc.)
list of mutations in Subtype A, gene F
list of mutations in Subtype B, gene F
(no mutations found for subtype B)
C

Data Flow:

A: Extracting and sorting sequences by gene, subtype and origin (WT/vaccine). Input is the set of GenBank entries pertaining to RSV. Outputs are sets of sequences, one each for each gene/subtype/origin combination (cyan). Classification is by annotation (WT) or by sequence similarity to WT (vaccines).
B: Building phylogenies and extracting clade representatives. Inputs are the sequence sets from step **A**, outputs are sets of sequences representative of each clade (orange). See text for rationale.
C: Finding mutated positions. Inputs are results from steps **A** and **B**, outputs are lists of positions which have significantly different major/minor allele frequencies among vaccines by comparison with WT sequences. Significance is measured by Fisher's exact test.