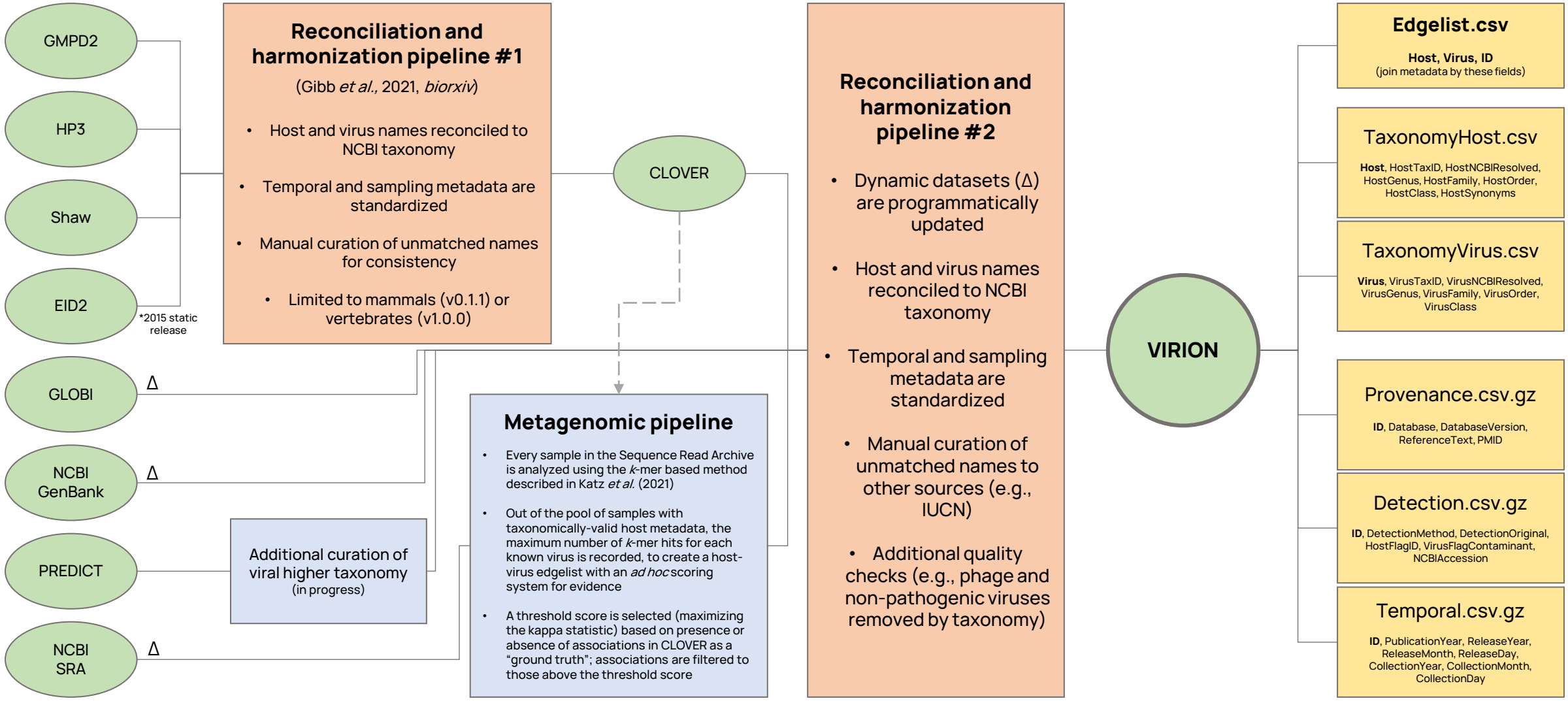
GMPD2
HP3
Shaw
EID2
*2015 static release
GLOBI
Δ
NCBI GenBank
Δ
PREDICT
NCBI SRA
Δ
Reconciliation and harmonization pipeline #1
(Gibb et al., 2021, biorxiv)
• Host and virus names reconciled to NCBI taxonomy
• Temporal and sampling metadata are standardized
• Manual curation of unmatched names for consistency
• Limited to mammals (v0.1.1) or vertebrates (v1.0.0)
Additional curation of viral higher taxonomy
(in progress)
CLOVER
Metagenomic pipeline
• Every sample in the Sequence Read Archive is analyzed using the k-mer based method described in Katz et al. (2021)
• Out of the pool of samples with taxonomically-valid host metadata, the maximum number of k-mer hits for each known virus is recorded, to create a host-virus edgelist with an ad hoc scoring system for evidence
• A threshold score is selected (maximizing the kappa statistic) based on presence or absence of associations in CLOVER as a "ground truth"; associations are filtered to those above the threshold score
Reconciliation and harmonization pipeline #2
• Dynamic datasets (Δ) are programmatically updated
• Host and virus names reconciled to NCBI taxonomy
• Temporal and sampling metadata are standardized
• Manual curation of unmatched names to other sources (e.g., IUCN)
• Additional quality checks (e.g., phage and non-pathogenic viruses removed by taxonomy)
VIRION
Edgelist.csv
Host, Virus, ID
(join metadata by these fields)
TaxonomyHost.csv
Host, HostTaxID, HostNCBIResolved, HostGenus, HostFamily, HostOrder, HostClass, HostSynonyms
TaxonomyVirus.csv
Virus, VirusTaxID, VirusNCBIResolved, VirusGenus, VirusFamily, VirusOrder, VirusClass
Provenance.csv.gz
ID, Database, DatabaseVersion, ReferenceText, PMID
Detection.csv.gz
ID, DetectionMethod, DetectionOriginal, HostFlagID, VirusFlagContaminant, NCBIAccession
Temporal.csv.gz
ID, PublicationYear, ReleaseYear, ReleaseMonth, ReleaseDay, CollectionYear, CollectionMonth, CollectionDay