The evolving VectorBase gene build: mixing automated and manual approaches when annotating vector genomes

Daniel Lawson

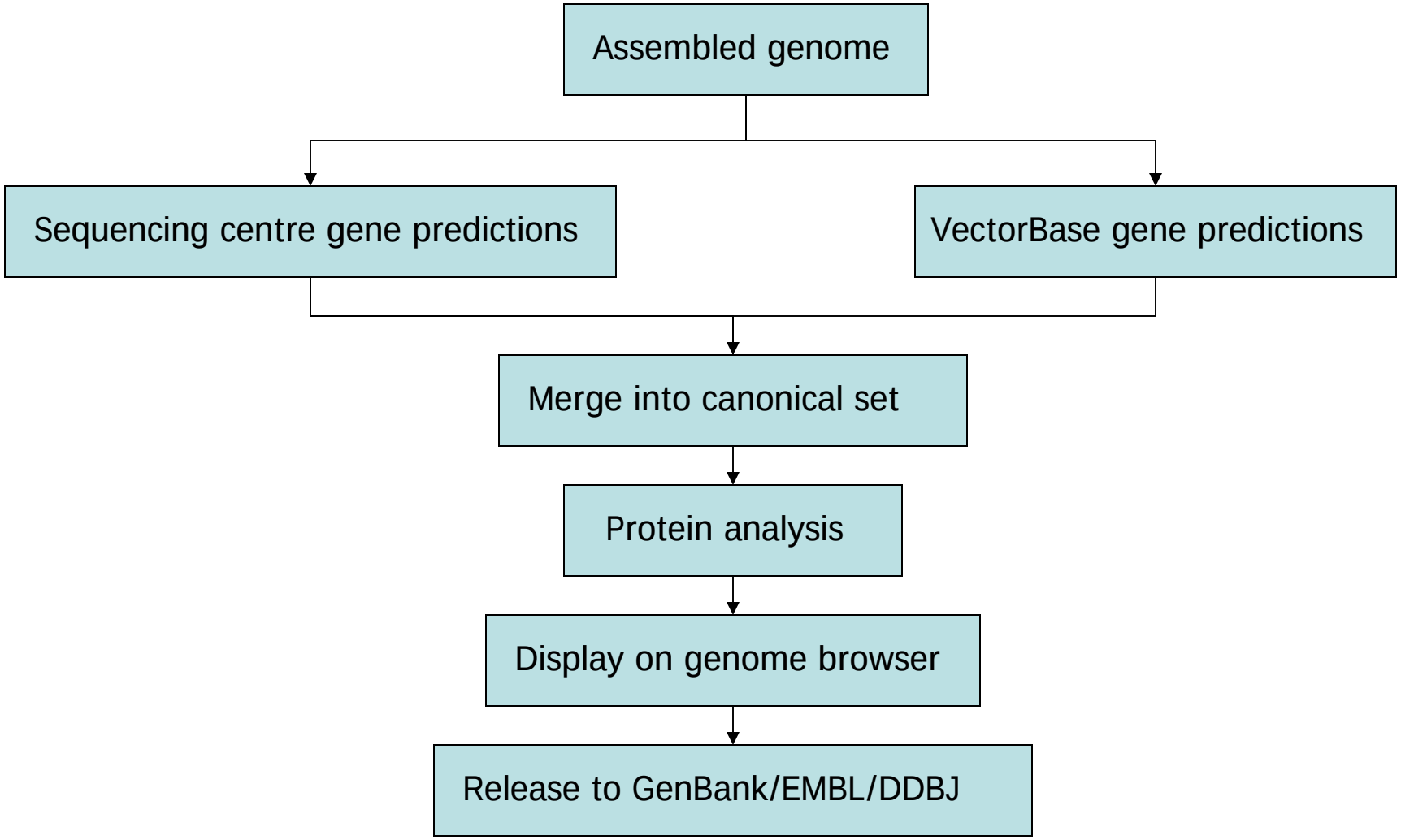
VectorBase-EBI, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton UK

Topics

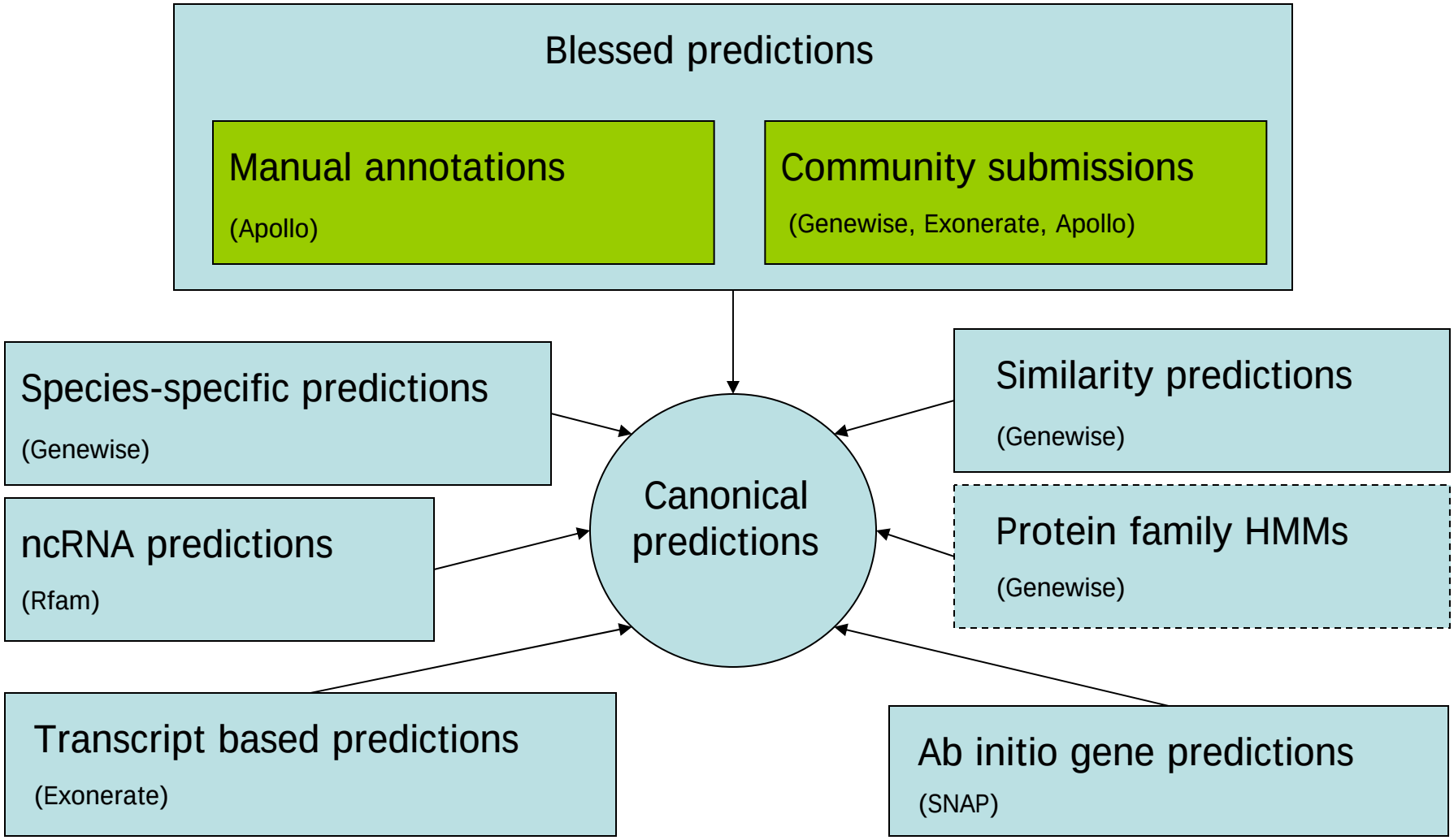
- ★ VectorBase species
- Generic gene prediction system (new genomes)
- ★ VectorBase gene prediction system (new developments)
- Influence of manual annotation
- ★ Visualizing manual annotation
- ★ Keeping information up to date

Anopheles gambiae PEST *Aedes aegypti*	Annotated
Culex pipiens quinquefasciatus	Assembly
Ixodes scapularis	Sequencing
Anopheles gambiae M & S form *Pediculus humanus*	Sequencing
Lutzomyia longipalpis *Glossina morsitans morsitans* *Phlebotomus papatasi* *Rhodnius prolixus*	Initiated

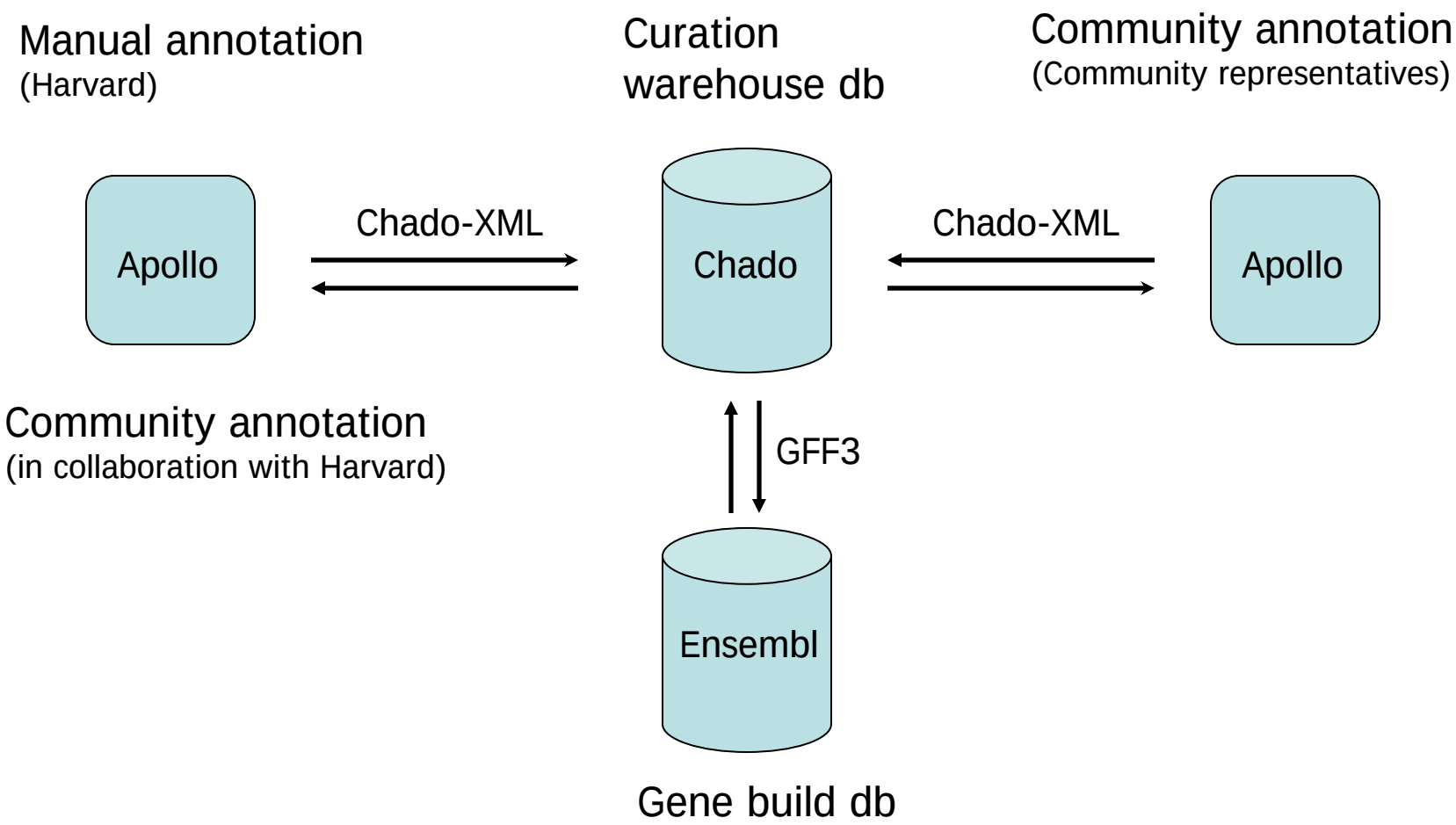
Annotation of new genomes



VectorBase gene prediction pipeline



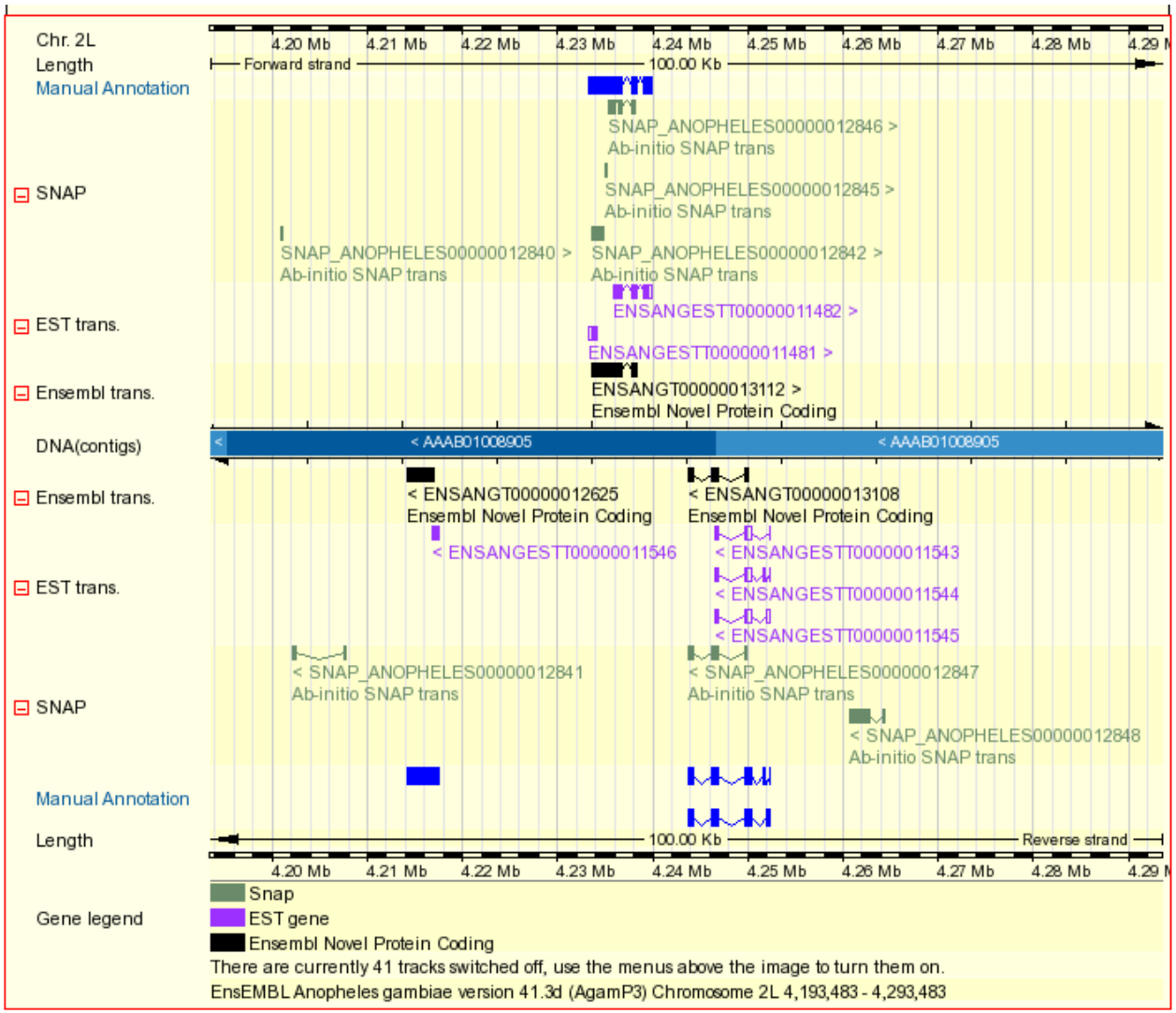
VectorBase curation database pipeline for manual/community annotation



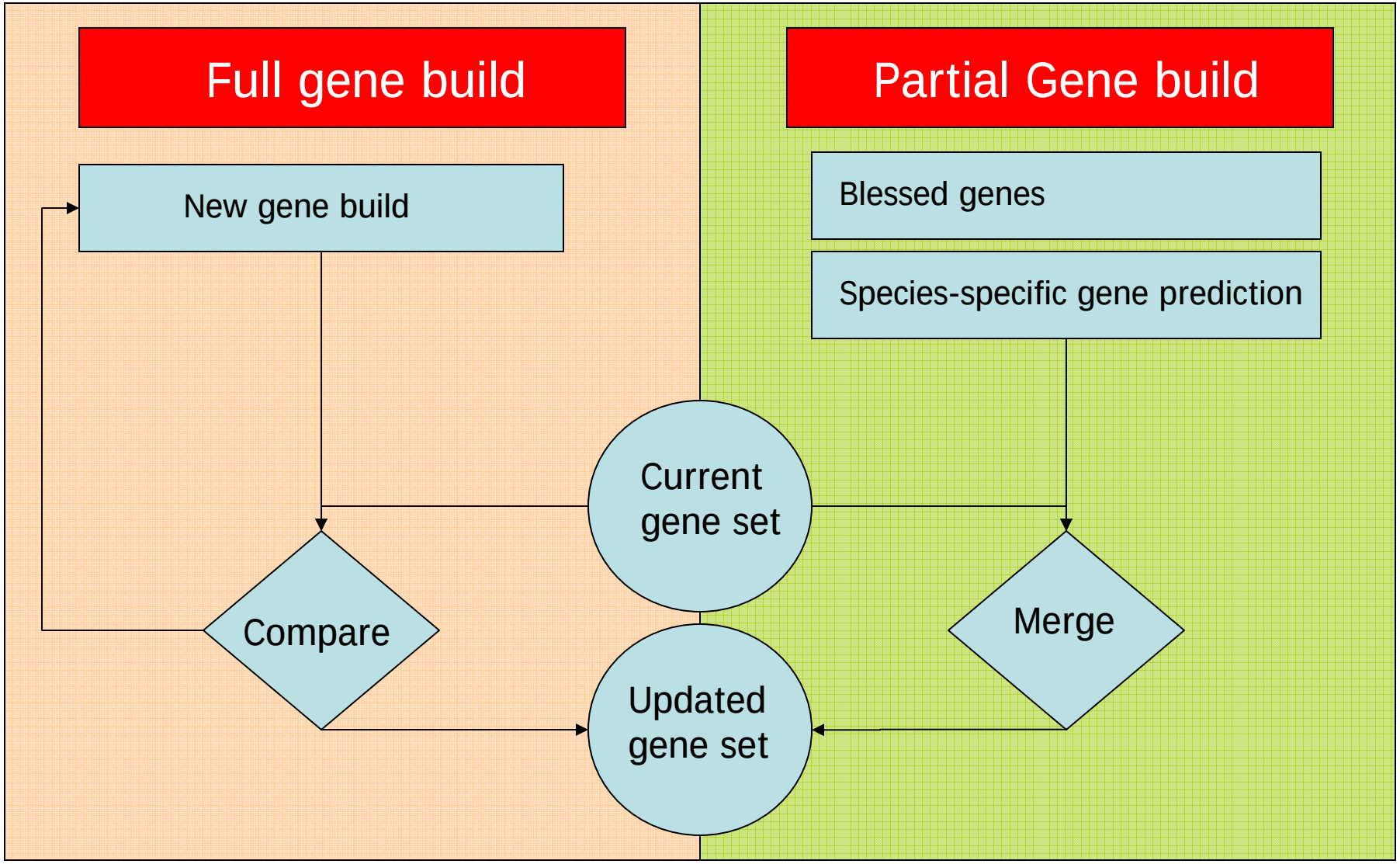
Manual annotation progress

	Protein-coding gene No.	VectorBase manual	Community submission
Anopheles gambiae			
AgamP3.3	13,277	261 (2.0 %)	667 (5.0 %)
current		2474 (18.6 %)	667* (5.0 %)
Aedes aegypti			
AaegL1.1	15,419	0 (0.0 %)	0 (0.0 %)
current		0 (0.0 %)	341 (2.2 %)

Manual annotation visualisation



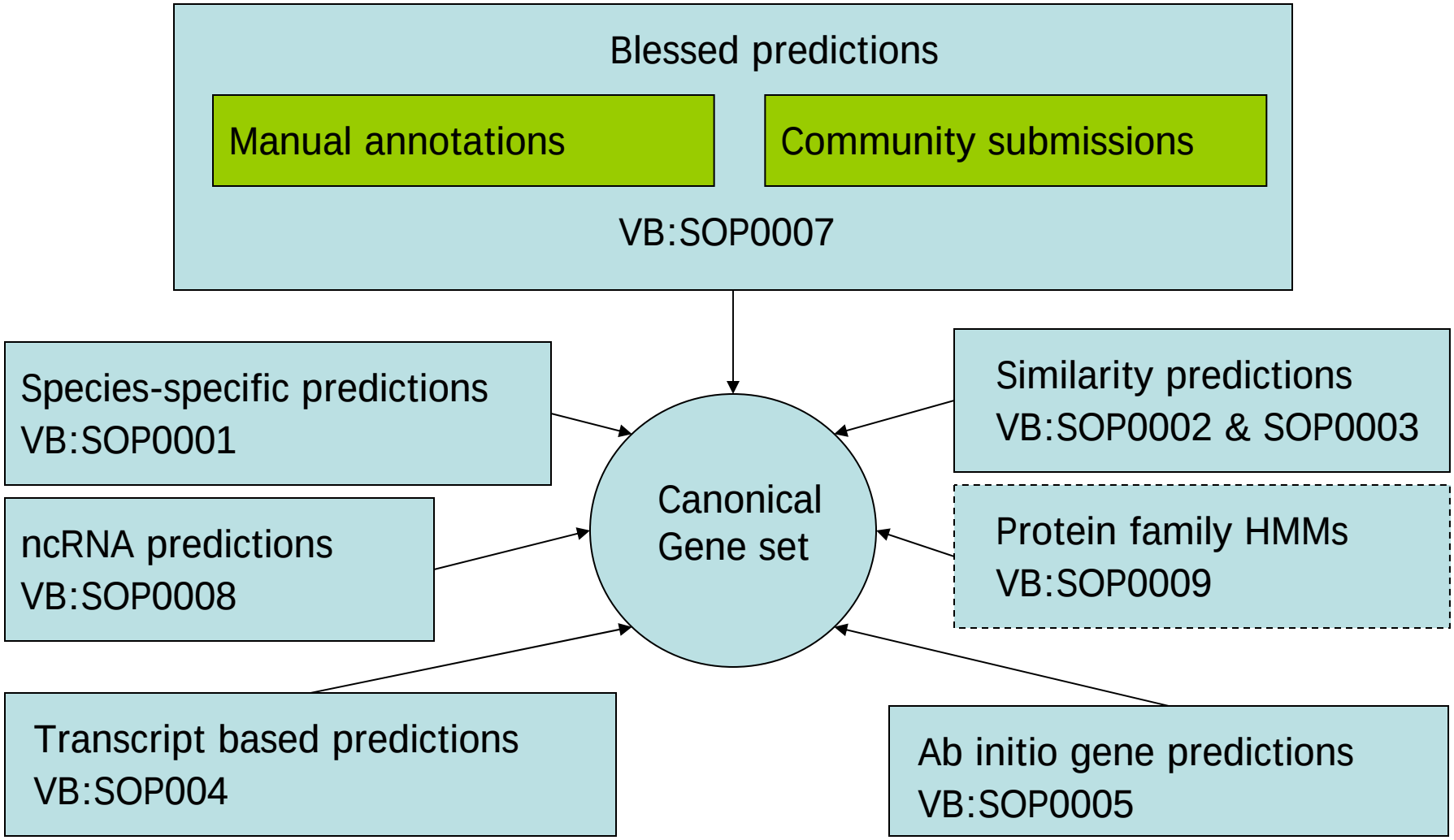
Overview of proposed re-annotation system



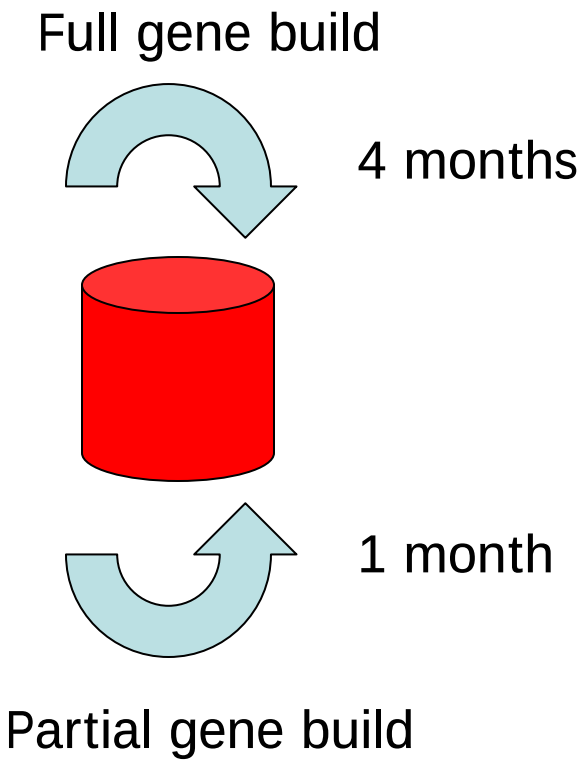
Comparing new gene builds with the old one

- Use of manual annotation for validation of automated gene build improvements
- Simple statistics (CDS length, intron size, CDS matching TE's)
- BRC annotation metrics
  - Supporting evidence for a gene prediction (citation, expression, orthology)
  - Attachment of Standard Operating Procedures (SOPs)

VectorBase gene prediction pipeline (SOP)



Gene build schedules



Triggers for re-annotation

- Temporal
- Data
  - New data for species
  - New genomes
  - Re-annotated genomes

3 wise annotators