Annotation of genomes at VectorBase

Daniel Lawson

Updated by James Allen & Gloria I. Giraldo-Calderón

February 2017

VectorBase

Bioinformatics Resource for Invertebrate Vectors of Human Pathogens

Genome annotation - the goal!

- Defining important features of the genome sequence
- Labelling/describing features of the genome
- ‘Adding value’ to the genome sequence

- Annotation is an ongoing process
- Annotation is almost always incomplete

- Complete set of gene predictions (protein-coding and ncRNA)
- Short description of the putative function for each prediction
- Species/Group dependant catalogue of other data types

Annotation from a genome project perspective

- Initial ‘first pass’ annotation prior to publication
- Subsequent annotation is a collaboration with the community
- Focused on protein-coding genes
- ‘Best guess’ predictions
- Little emphasis on transposons or pseudogenes
- Predicting gene loci is more important than getting 100% accuracy for gene structures
- Predicting accurate gene structures is more important than granularity of molecular functions

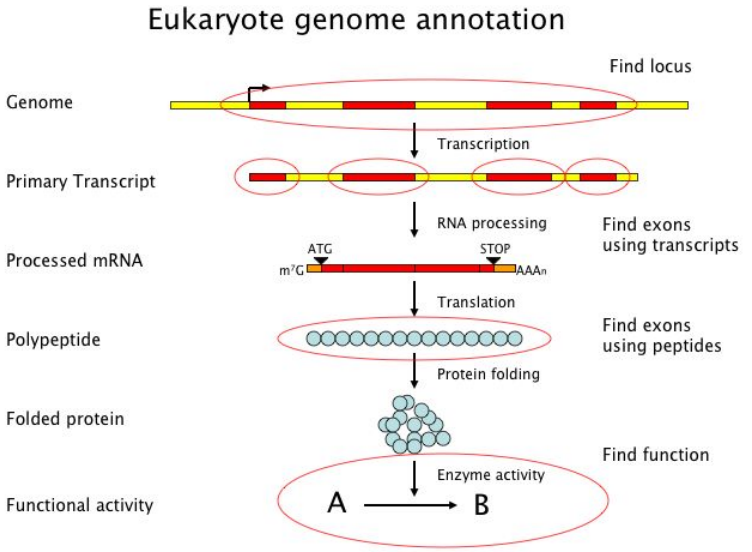
When to annotate?

- Genome assembly is ‘complete’
- Assembly passes some set of rudimentary QA/QC procedures
- Assess completeness & accuracy
- Does this assembly fulfil the original requirements (gene sets, synteny)
- Ancillary supporting data for gene prediction (RNAseq) are available

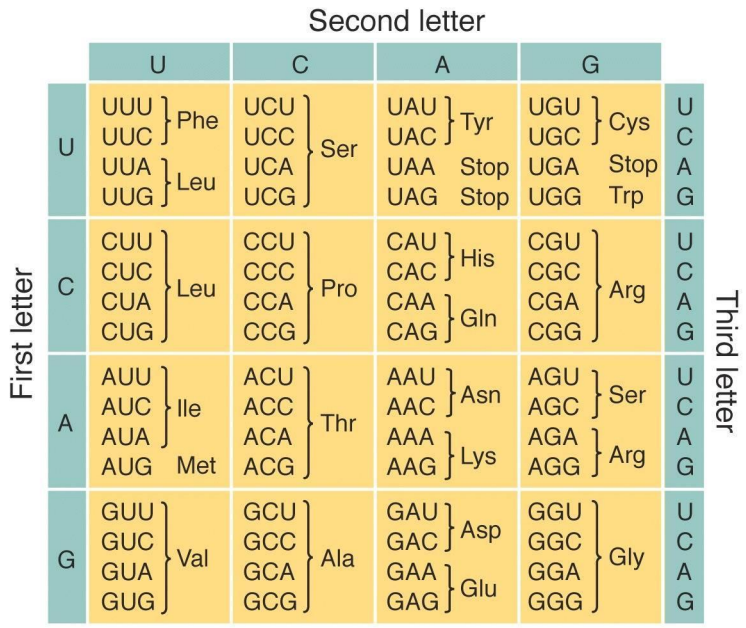
- The more partners/participants, the more important this is
- ‘Freeze’ assembly - do not meddle with this no matter how tempting it is
- Agree nomenclature for contigs/scaffolds, gene predictions
- Accept as much help as you can find (community involvement)
- Be pedantic, be boring, be thorough (or as much as you can be)

Genome annotation

- First-pass genome annotation is almost always based on “automatic” computational approaches
- *ab initio*
- Similarity based
  - Transcript (ESTs, RNAseq)
  - Protein (nr protein database)



Automatic annotation strategies



ab initio

```
Score =  349 bits (176), Expect = 3e-94
Identities = 176/176 (100%)
Strand = Plus / Minus

Query: 1       tacactacagttagaatgctgatgctgcactatagcaaacacggcgagtgcatactccag 60
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 3196614 tacactacagttagaatgctgatgctgcactatagcaaacacggcgagtgcatactccag 3196555

Query: 61      gaaattggtgcagcatttcgcggagagcatgcgacagatctgctactcatctgtgatggc 120
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 3196554 gaaattggtgcagcatttcgcggagagcatgcgacagatctgctactcatctgtgatggc 3196495

Query: 121     aaggagactgtgcgagcacacaagttggtactggcggctgccagtccactcatacg 176
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 3196494 aaggagactgtgcgagcacacaagttggtactggcggctgccagtccactcatacg 3196439

Score =  446 bits (225), Expect = e-123
Identities = 225/225 (100%)
Strand = Plus / Minus

Query: 176     gaatgattttggaagagactccgatgctggagggcgaaaccaccgtttacttcccggatg 235
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 3196347 gaatgattttggaagagactccgatgctggagggcgaaaccaccgtttacttcccggatg 3196288

Query: 236     tgcaggtgtgttacttccggctgctgctcgacttcctgtactccgggcaagtgtacgtgc 295
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 3196287 tgcaggtgtgttacttccggctgctgctcgacttcctgtactccgggcaagtgtacgtgc 3196228

Query: 296     ccgcaaacgaggtgcaccacctgcaagatctcttagcgttactacaaattaagcccagca 355
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 3196227 ccgcaaacgaggtgcaccacctgcaagatctcttagcgttactacaaattaagcccagca 3196168

Query: 356     tctggaaaaactccgattgctccaacgacagtggtaagtggtggt 400
                |||||||||||||||||||||||||||||||||||||||||||
Sbjct: 3196167 tctggaaaaactccgattgctccaacgacagtggtaagtggtggt 3196123
```

similarity

ab initio gene predictions

- Use compositional features of the DNA sequence to define coding segments (essentially exons)
  - ORFs
  - Coding bias
  - Splice site consensus sequences
  - Start and Stop codons

- Each feature is assigned a log likelihood score
- Use dynamic programming to find the highest scoring path
- Need to be trained using a known set of coding sequences

- Examples: Genefinder, Augustus, Glimmer, SNAP, fgenesh

Similarity gene predictions

- Use known coding sequences to define coding regions
- Transcriptome sequences (Sanger, 454, Illumina, SOLiD)
- Peptide sequences (taxonomically restricted)
- Needs to handle fuzzy alignment (especially around splice junctions)
- Needs to attempt to find start and stop codons

- e.g. Genewise, exonerate, gsnap, cufflinks

RNAseq based transcript reconstruction

Aim: Gene prediction using high-throughput transcriptome data a.k.a ‘RNAseq’

Overview

- Alternative method for generating transcript-based gene predictions.
- Uses Illumina or 454 reads as well as traditional Sanger sequenced ESTs
- Relatively short read lengths makes intron-exon junction prediction hard countered by the very high volume of data generated (millions of reads)
- Pipeline uses existing short-read algorithms for gene prediction:
- tophat, cufflinks, scripture, **trinity**

Potential problems

- Data sets require significant filtering and pre-analysis QC
- Mis-calling of homopolymer runs in 454 data leads to data noise and mis-prediction of splice sites
- Large data sets include many inappropriate splicing events (intron read through, NMD targets etc.)
- Alignment issues of data ‘noise’, especially from cufflinks

Summary: Effective at finding UTR regions and validating/improving predictions which is vital for making sense of sequence based measures of gene expression

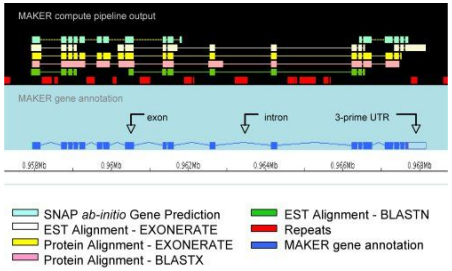
MAKER annotation with RNAseq and reference proteomes

Aims

- Gene prediction pipeline for the masses
- Used for a number of arthropod genome projects
- Touted as the default pipeline for many more (part of the GMOD toolkit)

Overview

- *ab-initio* gene predictions from SNAP, Augustus & FGENESH
- Final gene models from MAKER
- Similarity alignments from both EXONERATE and BLAST
- Repeats from RepeatFinder & RepeatMasker
- Additional data sets integrated via GFF3 files (RNA-Seq)
- Uses MPI for parallelization over a compute farm
- Optimization for long scaffolds



Summary

- Iterative runs give acceptable reference gene sets
- Used for Heliconius, Glossina, sandflies and the first tranche of the 16 Anophelines
- Used by others for Strigamia, Manduca, published ant genomes

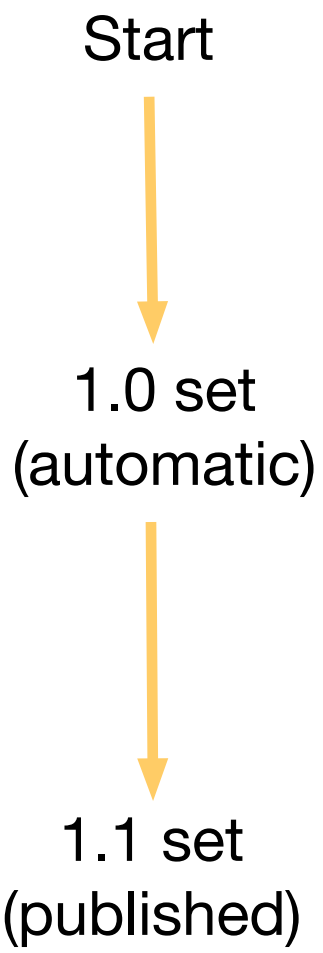
MAKER
Annotate this!

Current VectorBase annotation pipeline

- MAKER based automatic annotation
- Includes SNAP training and *ab initio*
- RNAseq based transcript similarity prediction
- Taxonomically constrained peptide similarity prediction
- 3+ rounds of prediction refinement

- Community annotation phase
- Capture gene structure changes
- Metadata associated with locus (symbol, description, citation)

- Submission to INSDC, propagation to UniProt
- Presentation through VectorBase & Ensembl Genomes



Functional annotation - Protein domains

- Protein domains have a number of definitions based on their size, folding and function/evolution.
- Domains are a part of protein structure description
- Domains with a similar structure are likely to be related evolutionarily and have a similar function
- We can use this to infer function (& structure) for an unknown protein be comparison to known proteins
- The tool of choice here is a Hidden Markov Model (HMM)

Protein Domain databases

- InterPro
- UniProt - protein database
- Prosite - database of regular expressions
- Pfam - profile HMMs
- PRINTS - conserved protein signatures
- Prodom - collection of multiple sequence alignments
- SMART - HMMs
- TIGRfams - HMMs
- PIRSF
- Superfamily
- Gene3D
- Panther - HMMs

Functional annotation - Other features

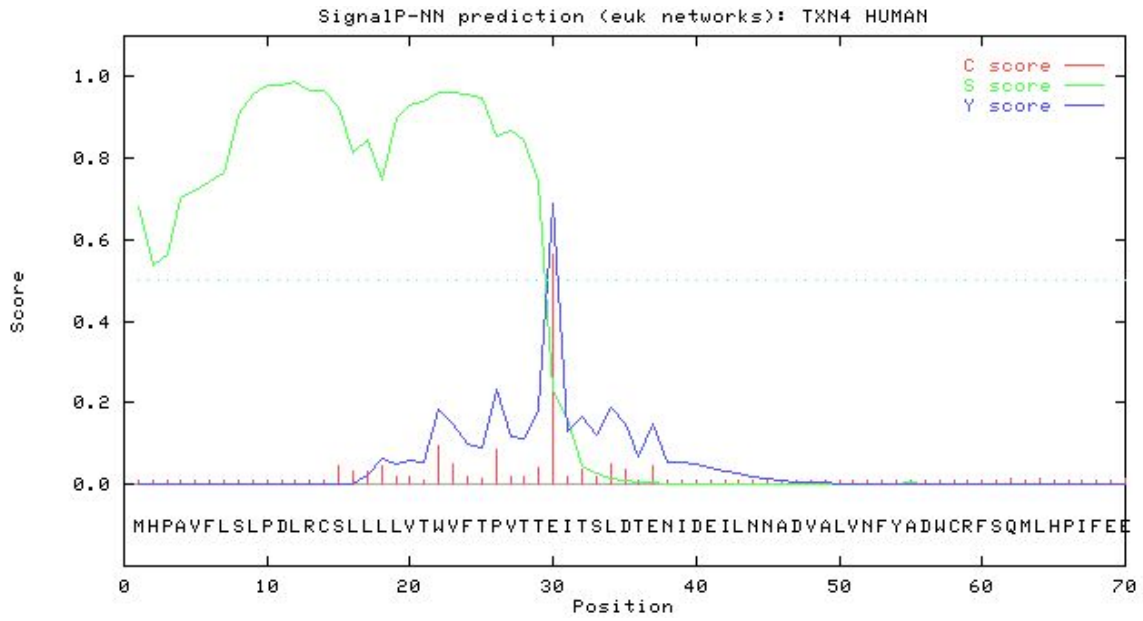
- Other features which can be determined
- Signal peptides
- Transmembrane domains
- Low complexity regions
- Various binding sites, glycosylation sites etc.

- See http://expasy.org/tools/ for a good list of possible prediction algorithms

Signal peptides

- Short peptide sequence found at the N-terminus of a pre-protein which mark the peptide for transport across one or more membranes

e.g. SignalP



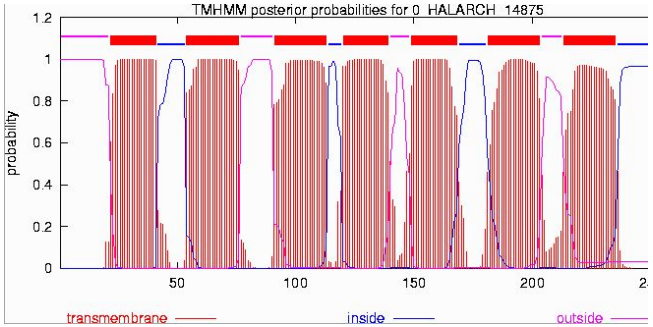
Transmembrane domains

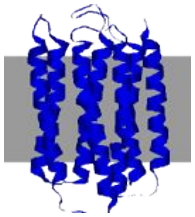
- Simple hydrophobic regions which sit inside a membrane
- Transmembrane domains anchor proteins in a membrane and can orient other domains in the protein correctly

Examples: Receptors, transporters, ion channels

- Identified based on the protein composition using a simple sliding window algorithm or an HMM

e.g. Tmpred, TMHMM





Ontologies

- Use of ontologies to annotate gene products

- Gene Ontology (GO)
  - Cellular component
  - Molecular function
  - Biological process

- Sequence Ontology (SO)

- GO terms mapped via interproscan and curated interpro2go file
- Assigned at lowest level of evidence (Inferred from Electronic Annotation IEA)

RNA gene annotations

- Most recent update, February 2017 release
- Data sources: miRBase, tRNAscan-SE, Rfam
- All data sources are available as alignment tracks in the genome browser
- Models can be used in Web Apollo for manual gene annotation

Projecting gene descriptions

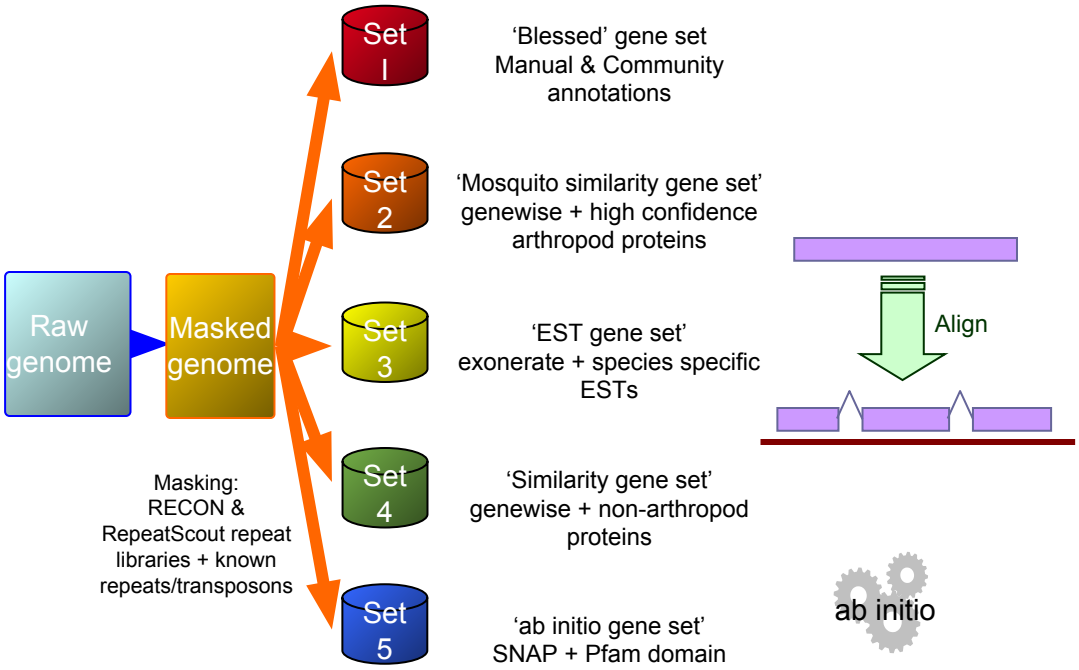
- Some of the species in VectorBase have been annotated more extensively than others, and it is useful to propagate gene descriptions to closely related species. Gene descriptions (but not gene names) are propagated based on orthology.

- Descriptions are projected from a gene to its ortholog if the pair share >30% amino acid sequence identity, and their alignment covers >66% of both genes' lengths.

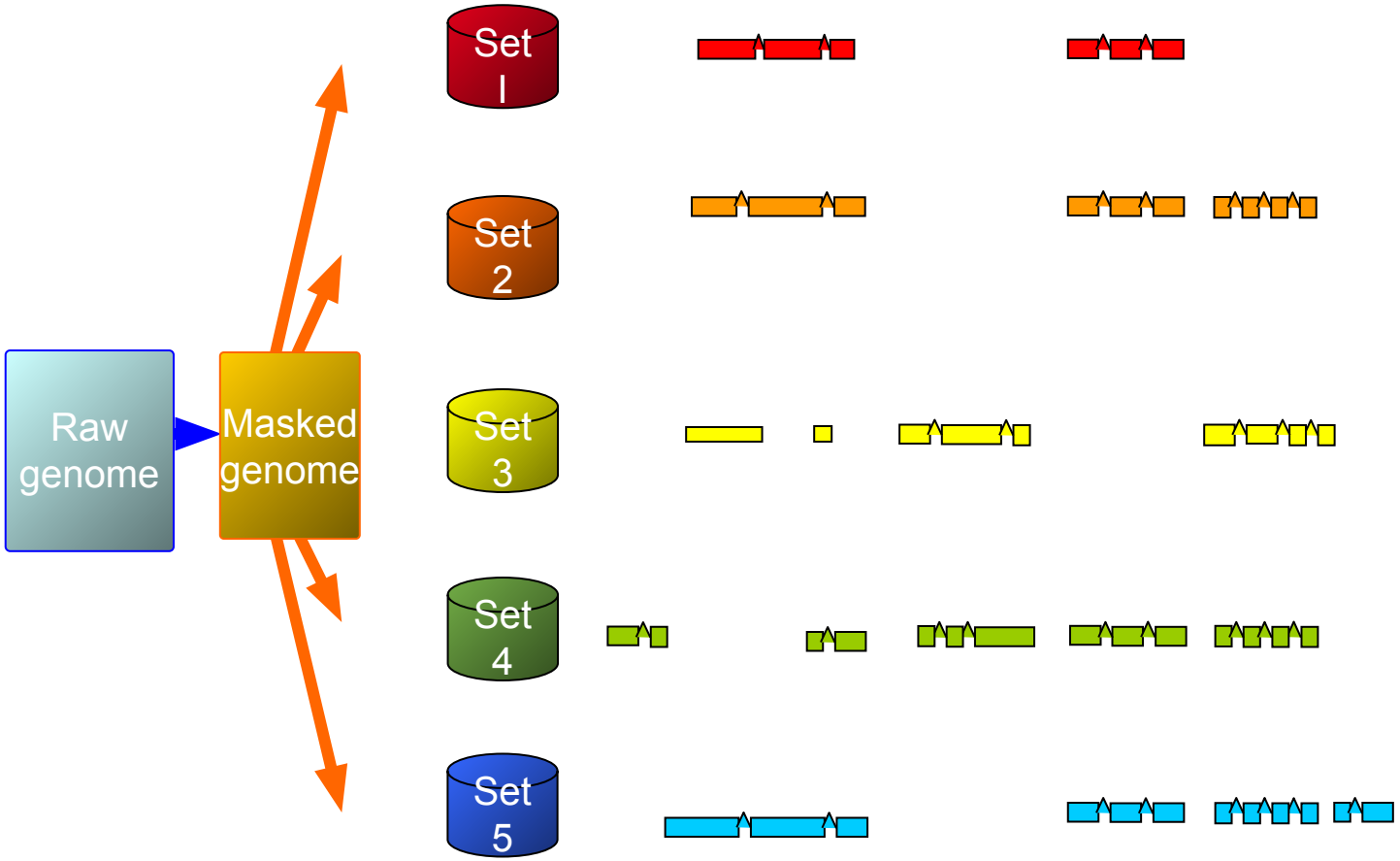
- Descriptions are propagated between the following species:
  - *Aedes aegypti* to *Aedes albopictus*
  - *Anopheles gambiae* to the other Anophelines
  - *Glossina morsitans* to the other *Glossinidae*, *Musca domestica*, *Stomoxys calcitrans*
  - *Drosophila melanogaster* to *Glossinidae*, *Musca domestica*, *Stomoxys calcitrans*

Previous (pre-2013) VectorBase genome annotation overview

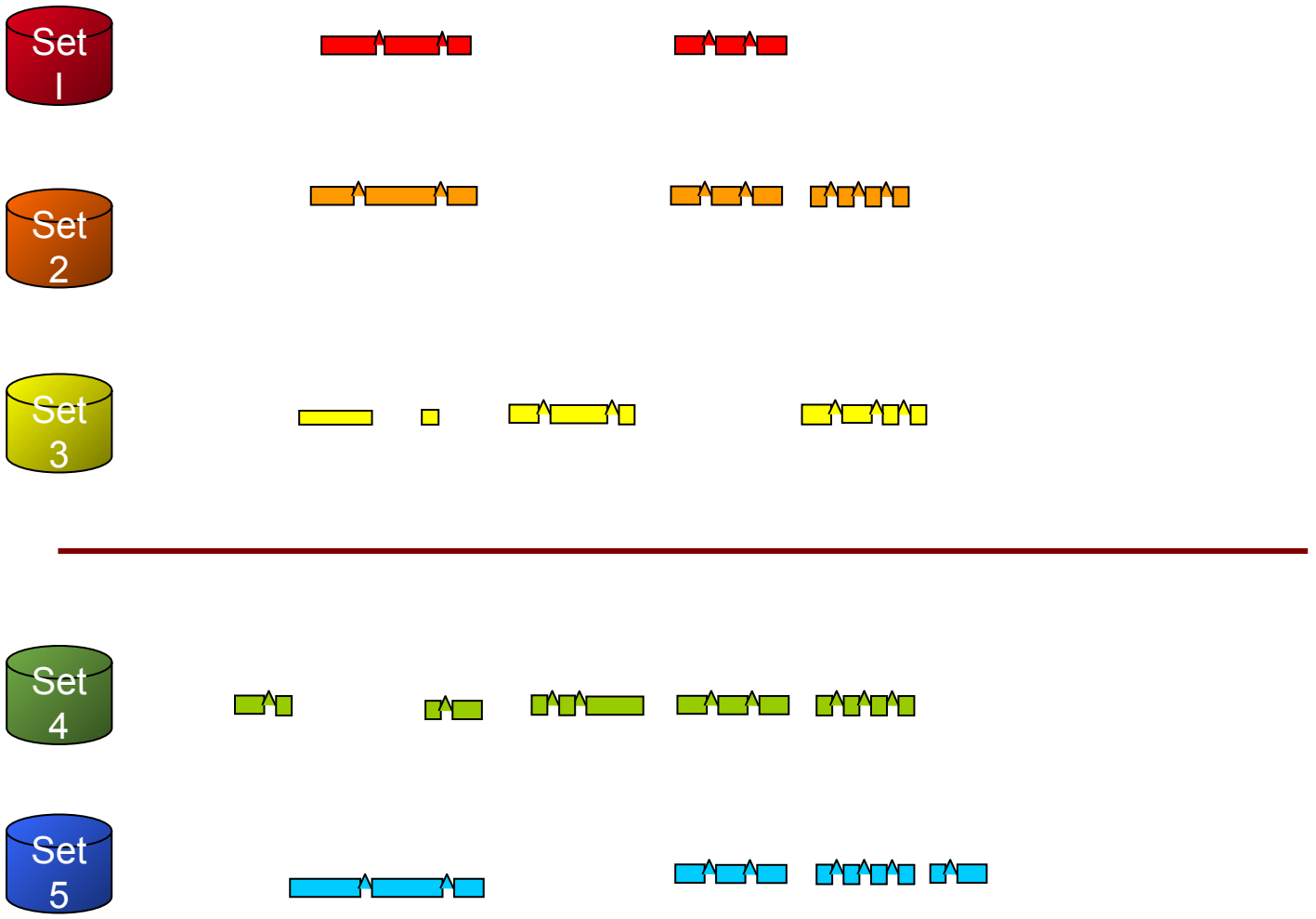
- VectorBase annotation pipeline based on Ensembl (used for many vertebrate genomes)
- (Relative) lack of evidence for predicting genes in comparison to vertebrates
- ‘Gap filling’ approach to aggregating the various prediction sets into the final canonical set



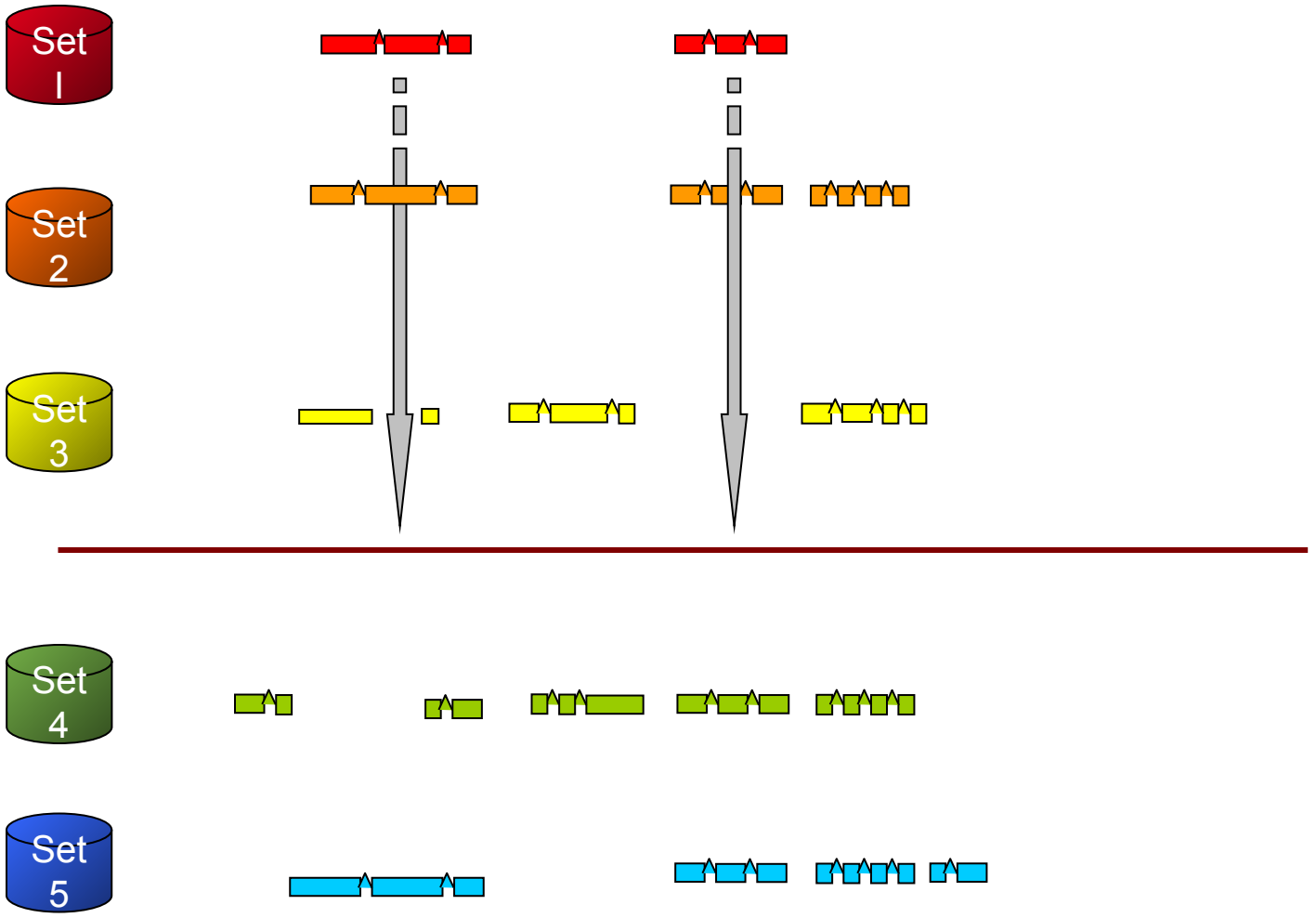
Make multiple sets of gene predictions



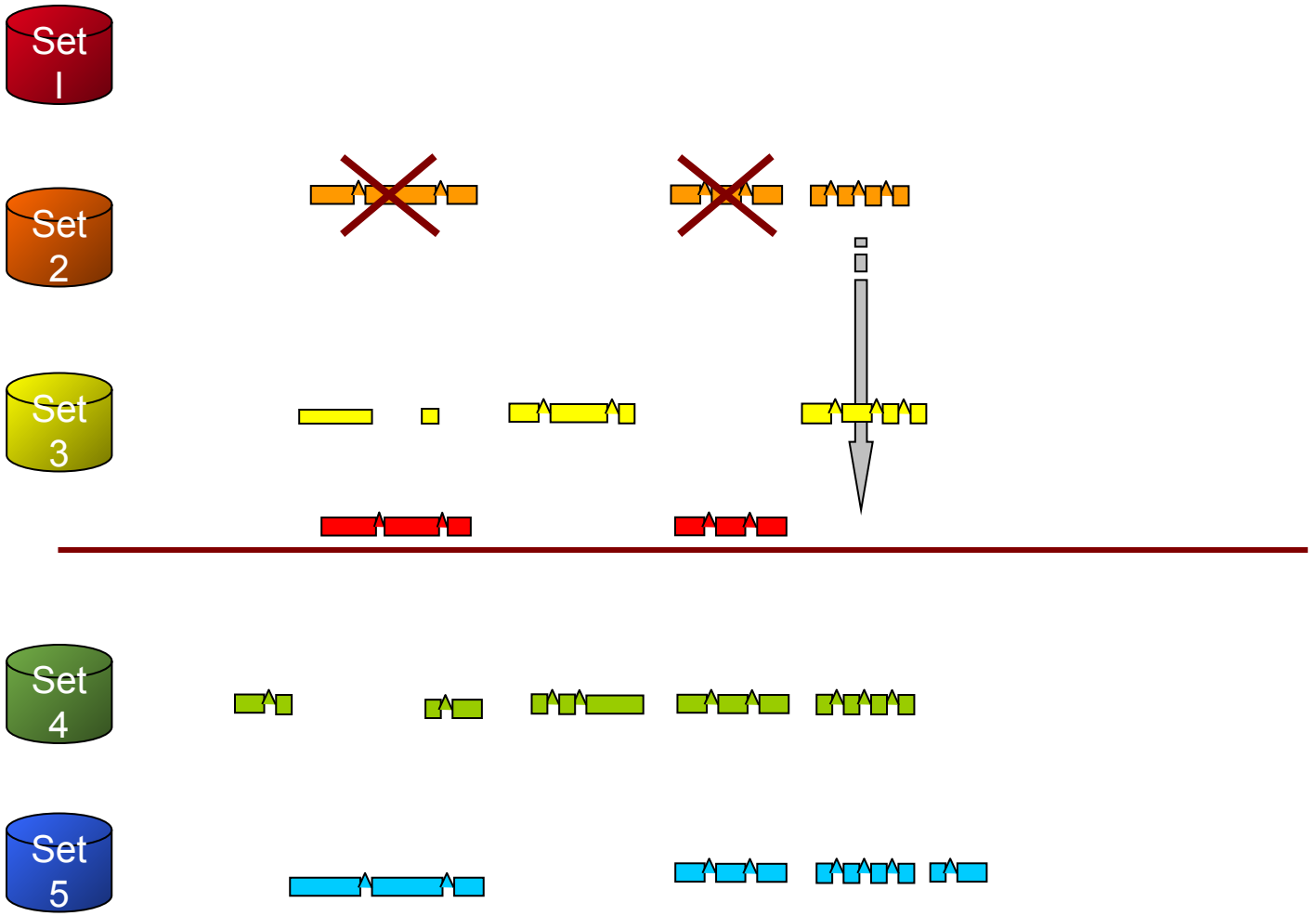
Confirm highest confidence gene set



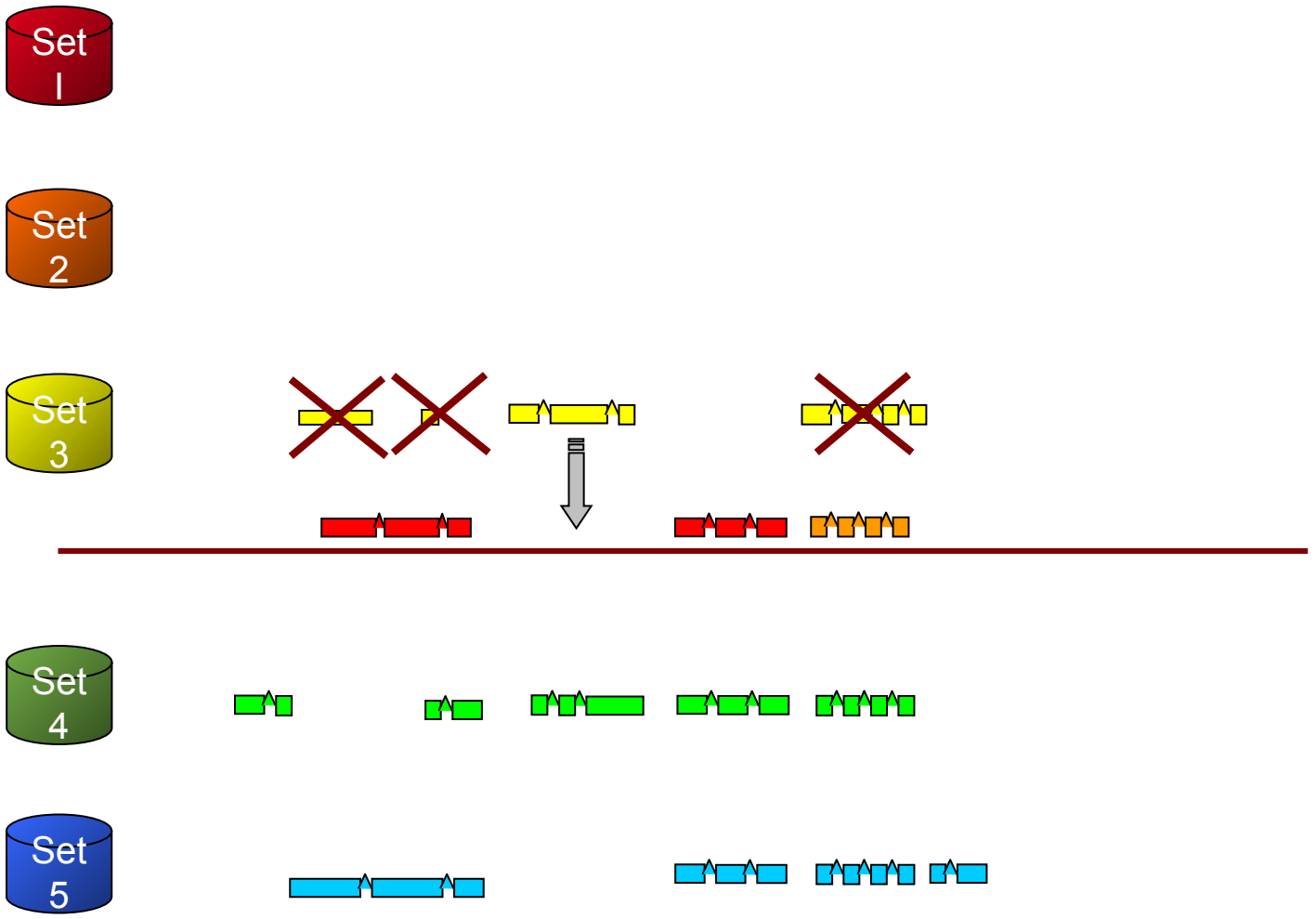
Confirm highest confidence gene set



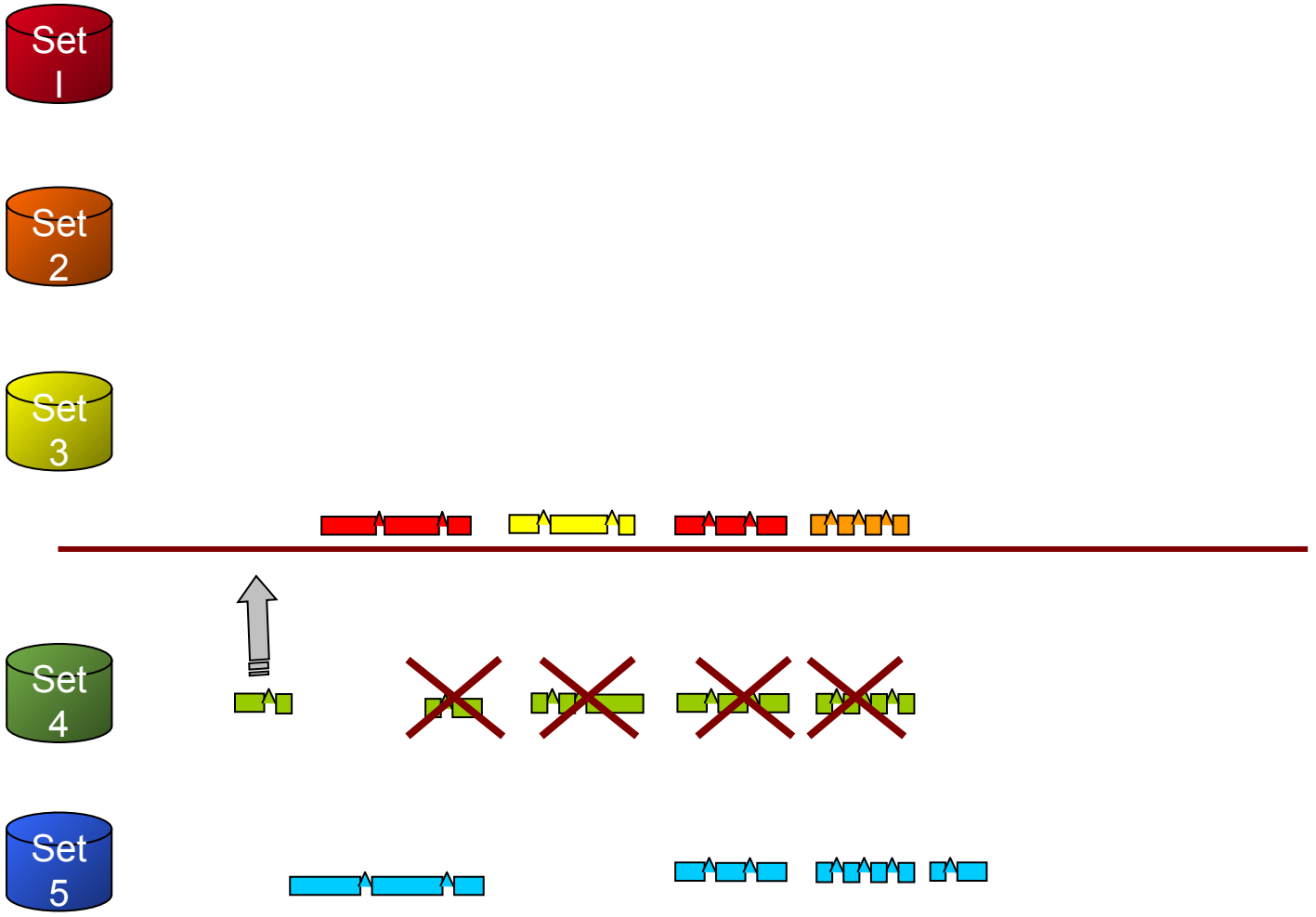
‘Gap fill’ with next highest confidence gene set



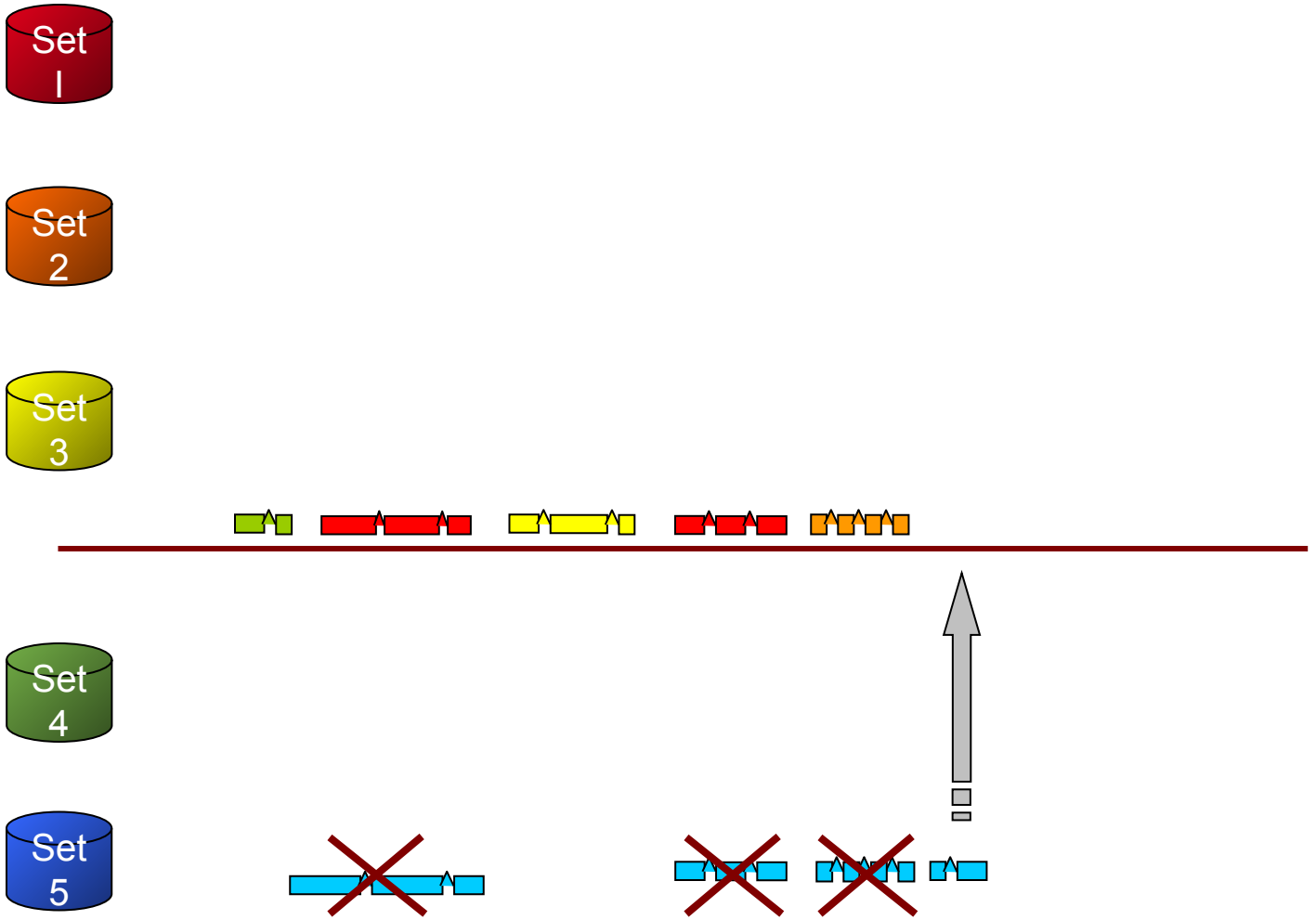
Repeat ‘Gap fill’ with next highest confidence gene set



Repeat ‘Gap fill’ with next highest confidence gene set



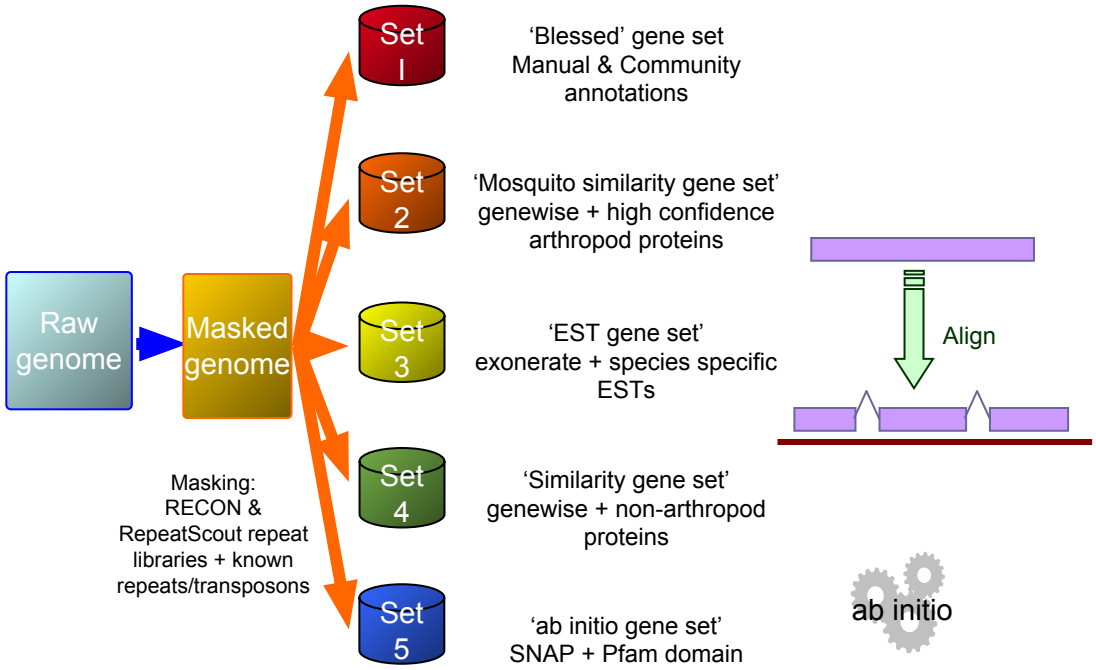
Repeat ‘Gap fill’ with next highest confidence gene set



Until all gene sets have been merged into a single canonical set

Previous (pre-2013) VectorBase genome annotation overview

- Gene prediction sets are subjective based on available evidence and annotators experience
- Significant time is spent in ‘Gap filling’, usually requiring a number of runs with subsequent quality assessment
- Difficult to parallelise (not necessarily in terms of the compute)
- Average 3-6 months per genome



How to search for more information or help?

E-mail us at
info@vectorbase.org