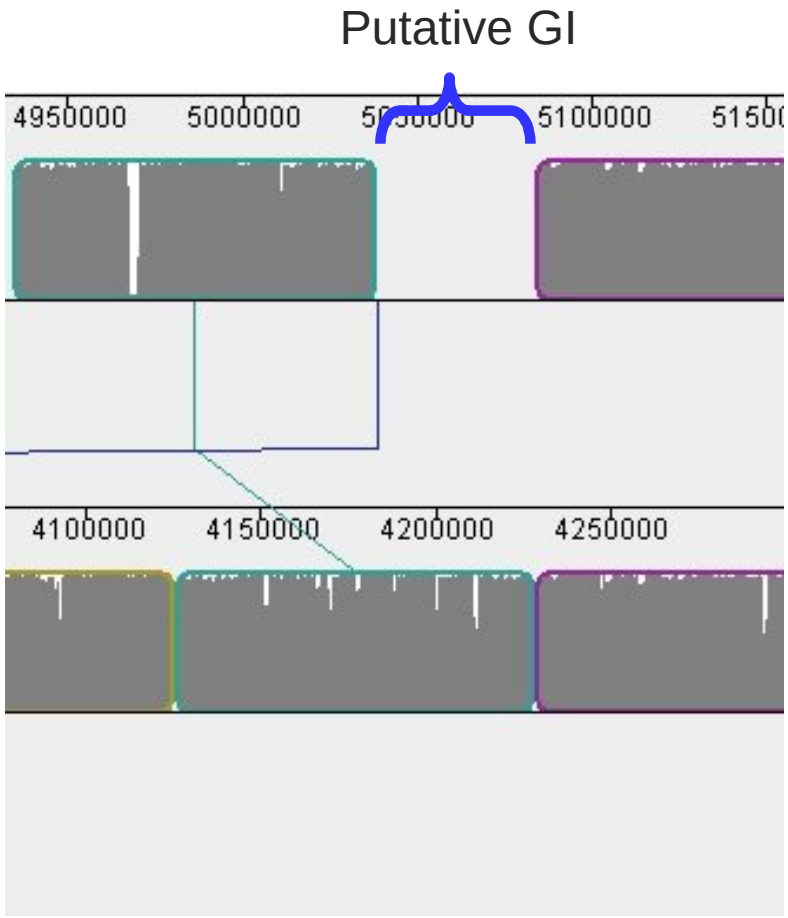
Overview of IslandPick pipeline and the generation of GI datasets

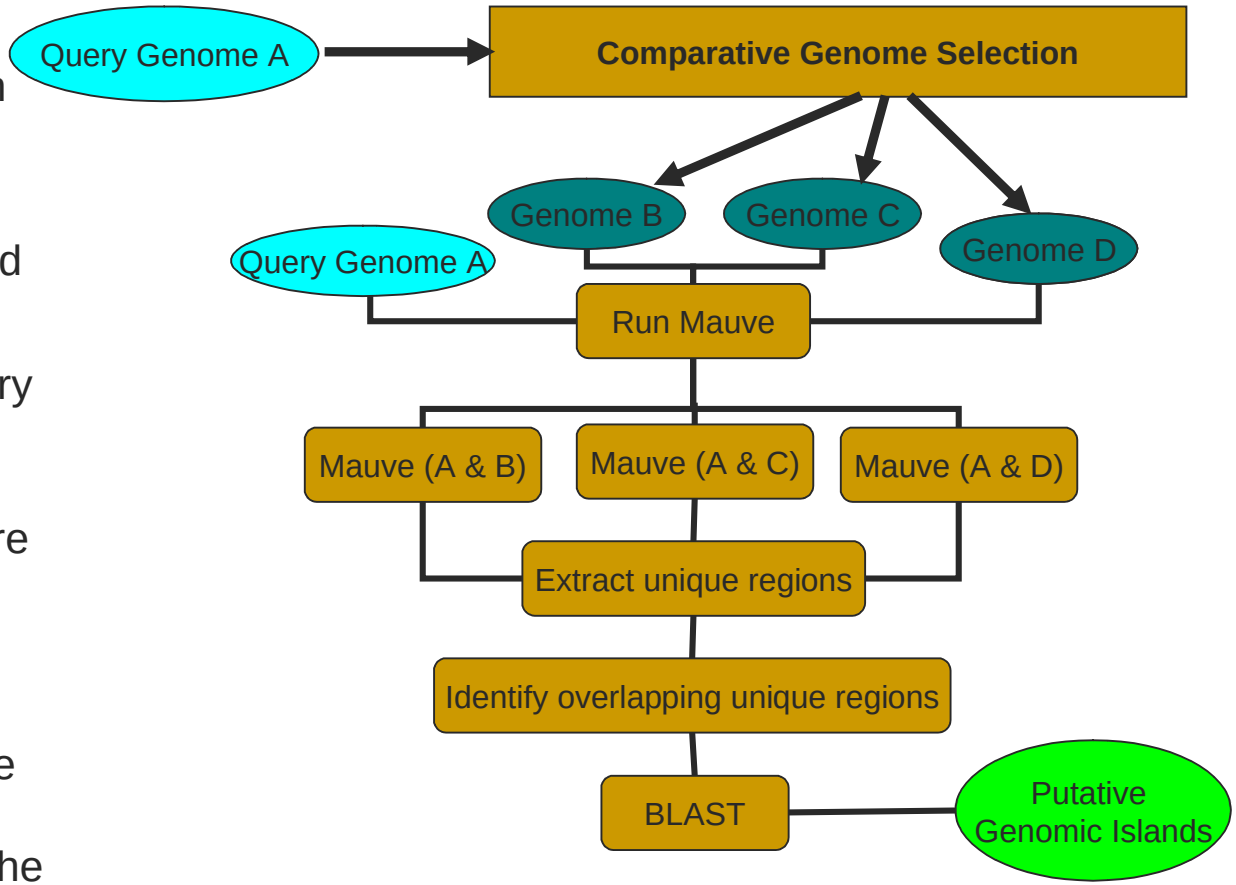
Predicting GIs using comparative genomics

- By using whole genome alignments we can identify regions that are present in one genome but not in another
- Large regions that are not aligned are probably unique to that genome
- Regions that are unique when compared against many closely related genomes can be considered putative GIs
- The IslandPick pipeline uses these principles to identify GIs



Outline of IslandPick pipeline

1. Given a query genome, IslandPick picks suitable genomes for comparison
2. Pair-wise whole genome alignments are performed
3. Large regions in the query genome that can not be aligned to any of the comparative genomes are identified
4. Blast is used as a secondary filter to ensure the regions are not genome duplications in the query genome



Comparative Genome Selection

- For any given query genome how do we pick genomes for comparison to produce a robust dataset of GIs consistently?

Comparative Genome Selection

- First, we build an all against all distance matrix for all sequence genomes using CVTree (Qi, 2004, PMID: 15215347)
- We then use various cutoffs to select genomes that are suitable for comparison with the query genome such that only highly probable GIs are predicted (see later slides)
- Query genomes that do not have suitable comparative genomes are not used for prediction of GIs

Comparative genome selection cutoffs

- The three obvious cutoffs that are needed when choosing genomes are:
  - Minimum Distance Cutoff
    - Eliminates the use of genomes that have not diverged enough (very closely related strains)
  - Max Distance Cutoff
    - Eliminates the use of genomes that have diverged too much (noise)
  - Min Number Genomes
    - Eliminates the use of too few comparative genomes
- Next, we introduce a cutoff that controls the prediction of more recent or ancient GI insertions

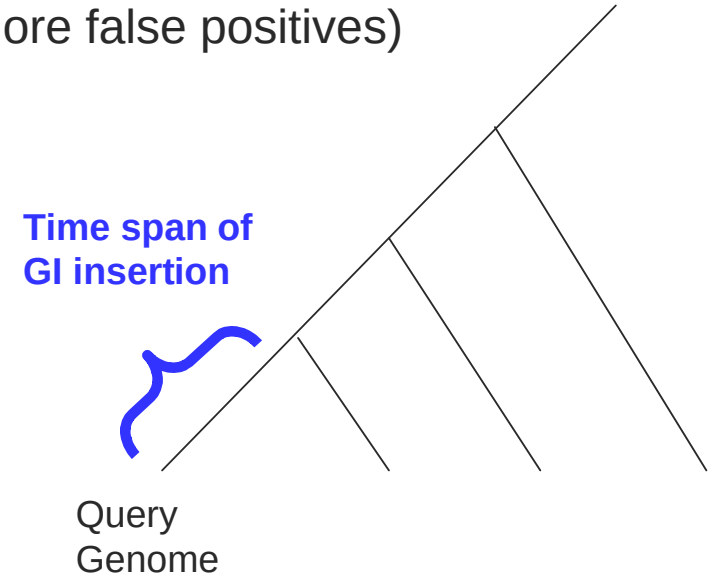
Predicting Similar Aged GIs

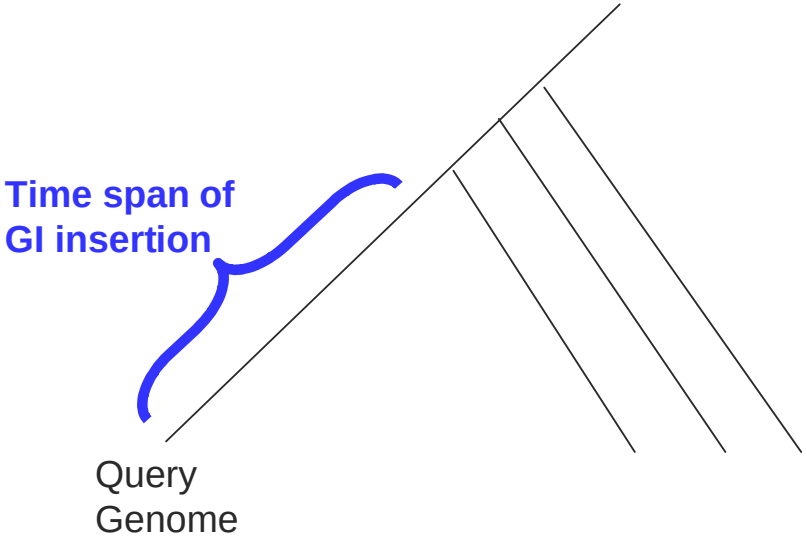
-Detecting time of GI insertion depends on the closest relative used in the comparison

-By selecting a more distantly related genome as the closest relative to the query genome, we can detect the insertion of older GIs

-We introduce a cutoff that controls the selection of the closest genome called the “Single Close Genome Cutoff”

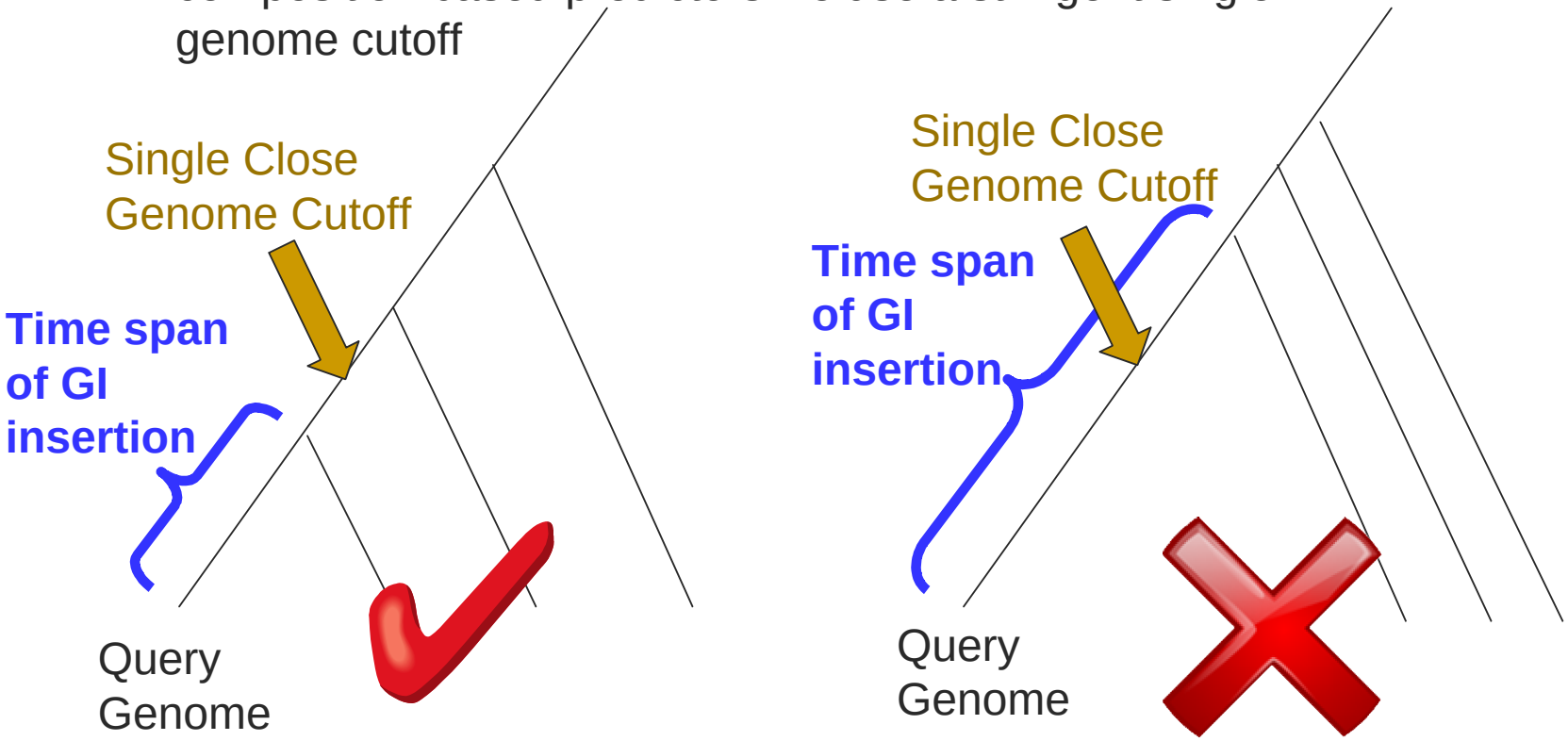
-Increasing this cutoff allows the prediction of more ancient GIs (increasing recall), whereas decreasing this cutoff increases precision (because predicting old GIs will allow more false positives)





Single Close Genome Cutoff

- To reduce the prediction of false positives in our GI dataset and allow for improved accuracy calculations of sequence composition based predictors we use a stringent single genome cutoff

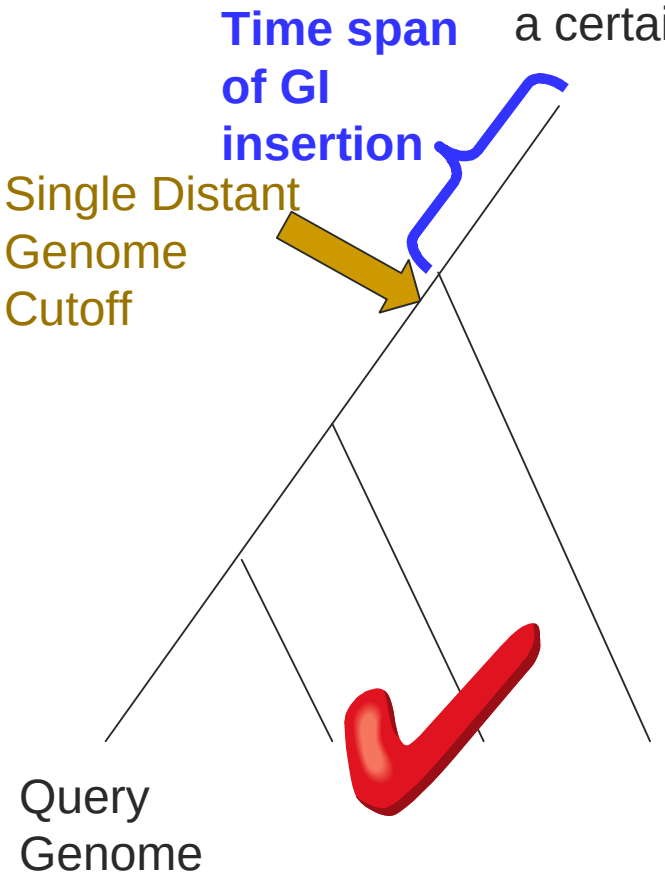


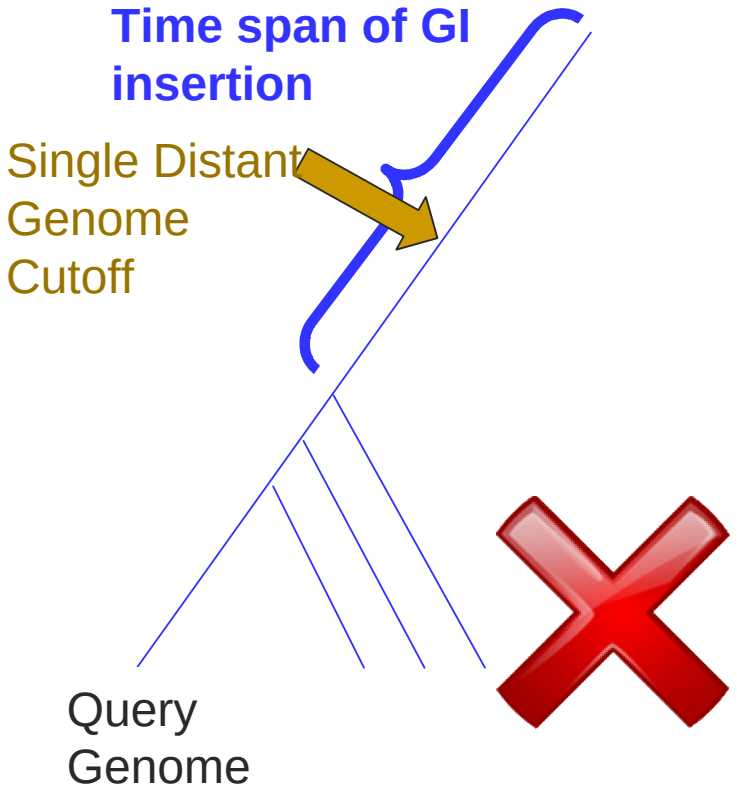
Negative Dataset Creation

- To generate a dataset of non-GIs we identify regions that are conserved across multiple species using whole genome alignments
- The same comparative genome selection is used with an additional cutoff that ensures that recent insertions are not mistaken as conserved regions
- This “Single Distant Genome Cutoff” ensures that at least one genome is phylogenetically distant from the query genome
- Increasing this cutoff decreases false positives (ie. GIs) in the negative dataset, whereas decreasing this cutoff produces a larger negative dataset

Predicting good non-GIs

-To reduce the prediction of very ancient GIs within the negative dataset we require that at least 1 genome is at least a certain distance away from the query genome





Evaluating GI Predictors

-By generating robust positive and negative GI datasets using a method that does not rely on sequence composition bias, we can better evaluate the accuracy of several sequence composition based GI predictors