Advanced genomics analysis

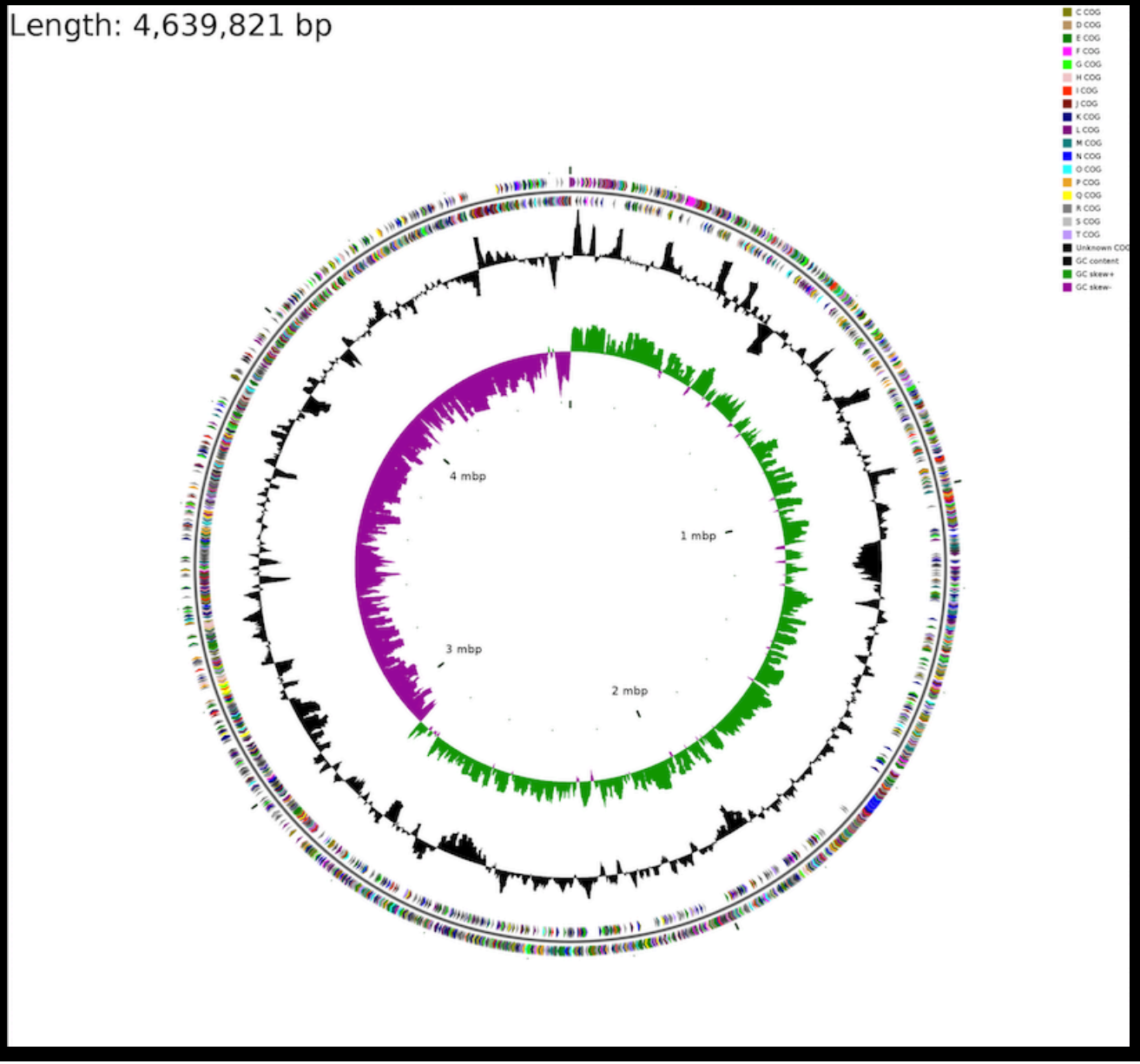
- Calculate GC content, GC skew and visualize the genomic feature in a circular way
- Deduce ortholog relationship between two species and multiple species
- Find the syntenic regions according to the gene order
- Construct the gene families

GC content and GC skew

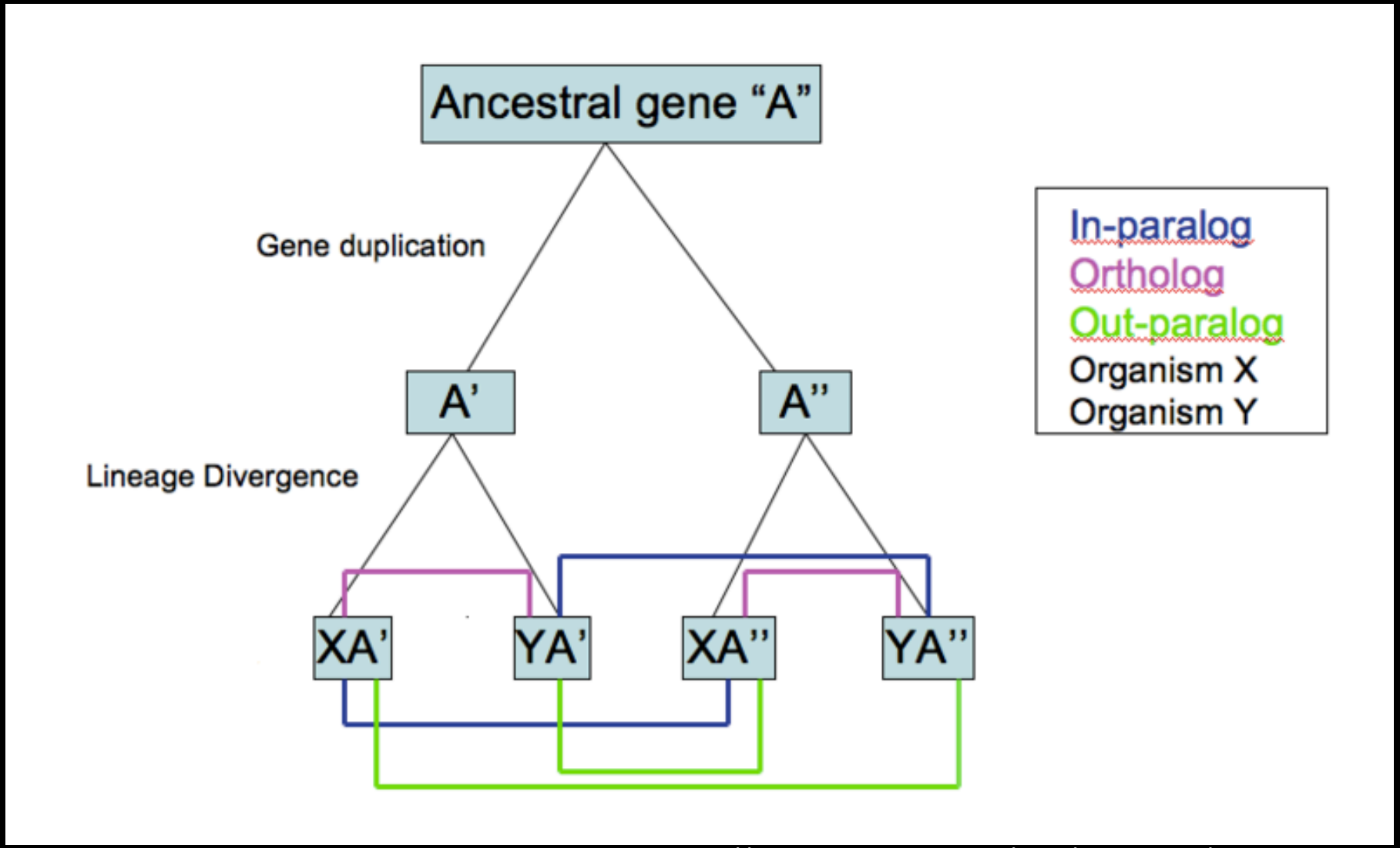
- GC content
  - Percentage of GC bases in a given stretch of DNA
  - High GC more stable than low GC
  - Isochore (GC biased large fragment) contain many coding genes
  - Related with mutation bias, recombination, etc
- GC skew
  - (G - C)/(G + C)
  - Uneven distribution in prokaryotic genome
  - Related to the leading and lagging strand in DNA replication

https://www.yourgenome.org/facts/what-is-dna-replication

GC content estimation and circular genome visualization

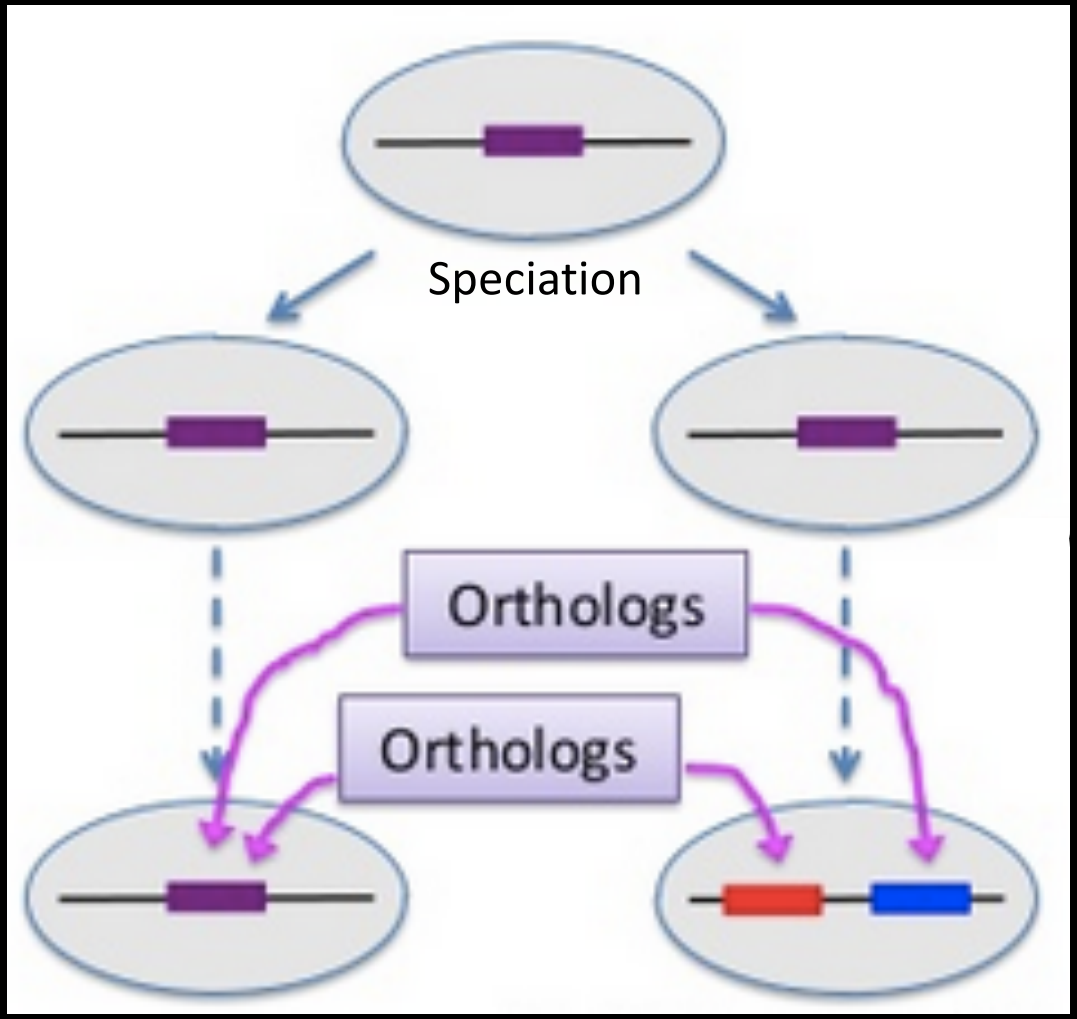


Orthologs and paralogs



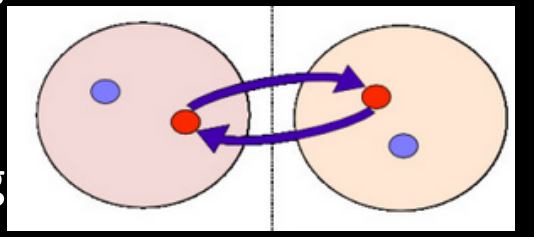
https://genomevolution.org/wiki/index.php/In-Paralog

Ortholog is not transitive

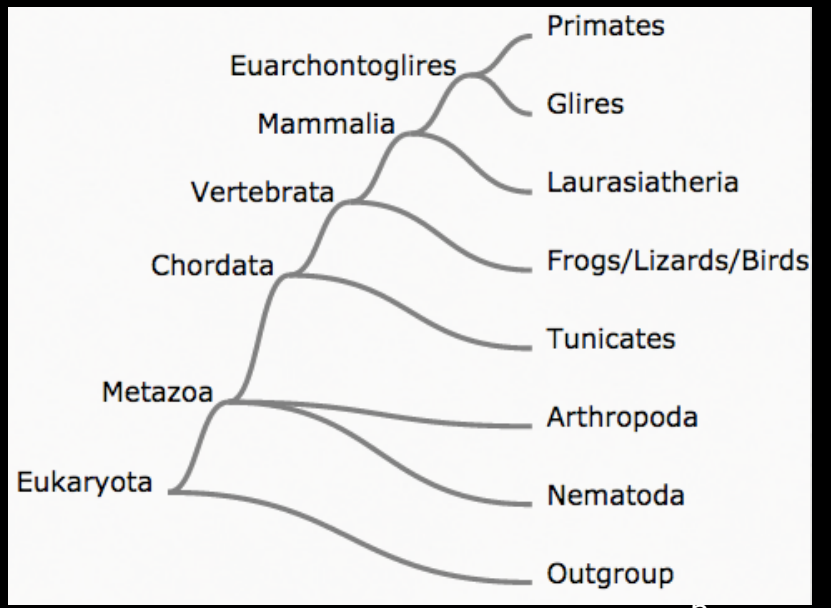


Identify ortholog relationships between two gene sets

- Commonly used methodology/software
  - Homologous search based methods
    - Reciprocal best hits (RBH), most widely used methods
    - Inparanoid, another popular softwares
      Include 1-to-many and many-to-many relationships
    - TribeMCL or orthoMCL,Markov chain based clustering
      Include 1-to-many and many-to-many relationships



  - Phylogeny based methods
    - MetaPhors
    - TreeFam
    - EnsemblCompara
    - PhylomeDB
    - …



Treefam

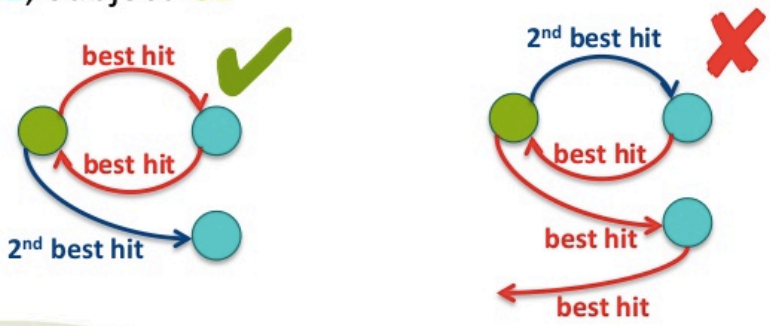
- Never do
  - Directly blast search to nr, let alone particular function subsets

Ortholog identification based on reciprocal best hits (RBH)

Reciprocal Best BLAST Hits

The James Hutton Institute

- S1, S2 are the gene sequence sets from two organisms
- BLASTP:
  - Query=S1, Subject=S2
  - Query=S2, Subject=S1

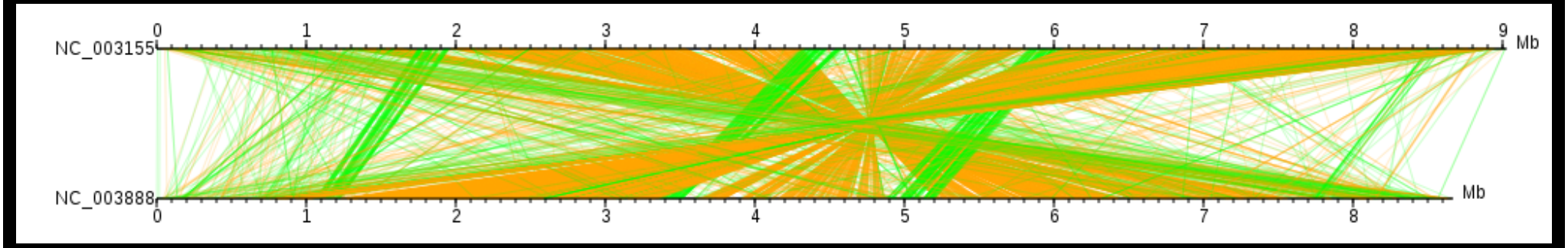


- Optionally filter BLAST hits (e.g. on %identity and %coverage)
- Find all pairs of sequences {G_{S1n}, G_{S2n}} in S1, S2 where G_{S1n} is the best BLAST match to G_{S2n} and G_{S2n} is the best BLAST match to G_{S1n}.

Leighton Pritchard, http://www.slideshare.net/

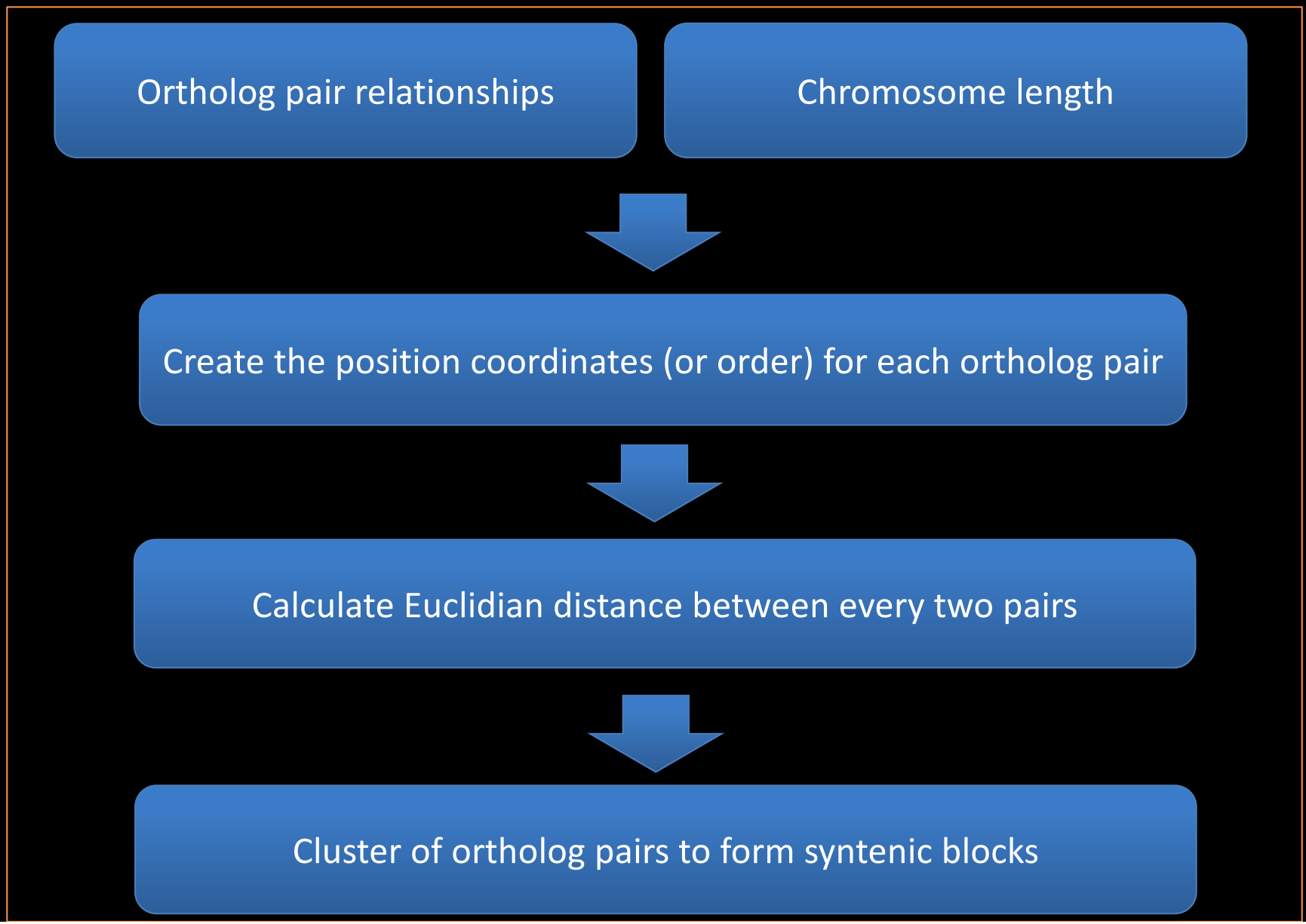
Find the syntenic regions according to gene order

- Synteny
  - collinear relationship between two genomic regions
  - Indication of ortholog relationship for flexible genomes



- Methodology
  - Whole genome alignment based
    - Mummer based
  - **Gene cluster based**

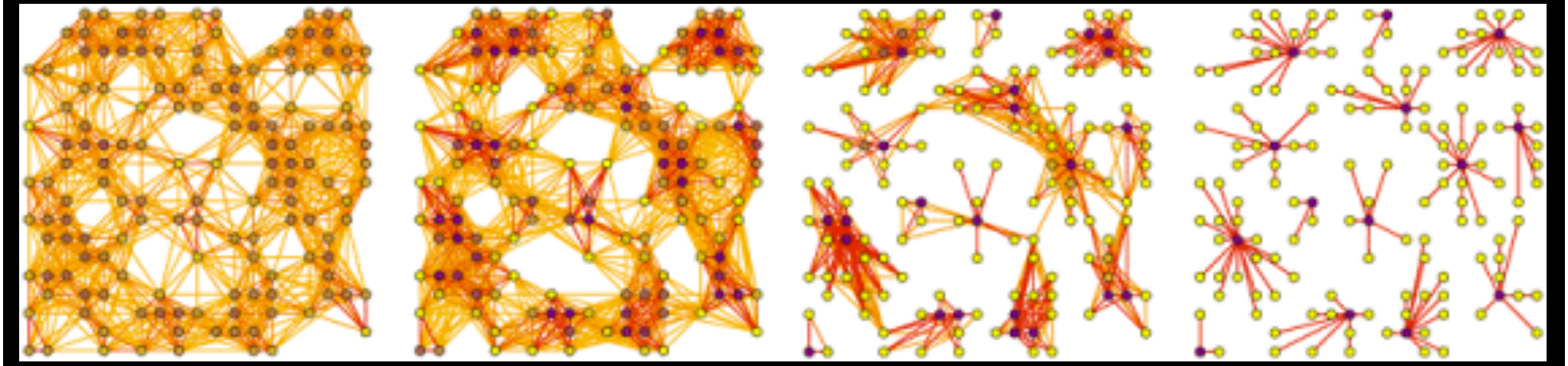
Pipeline for cluster-based synteny



Identify gene families

- Hierarchical clustering based on paired wised relationships

- Markov chain clustering



http://micans.org/mcl/

One parameter tuned: Inflation, the limitation of intra-cluster random walks.

Practice

- Estimate of GC content and GC skew in a genome and visualize it
- Visualize a circular genome regarding GC content, GC skew and COG functional annotation
- Use reciprocal best hit method to find ortholog relationships between gene sets from to bacterial species
- Visualize the identity between the orthologs
- Detect the synteny between two chromosomes
- Cluster gene families using Markov cluster