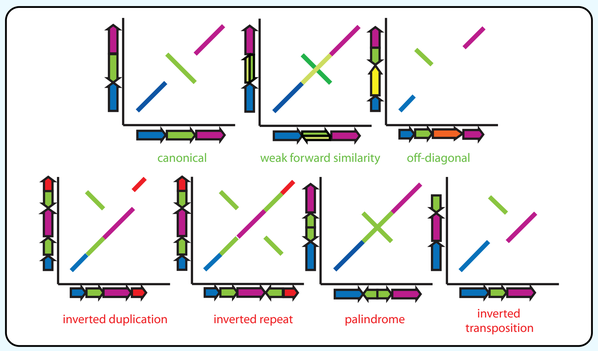
Micro-inversions in mammalian evolution

Mark J. Chaisson[1,4], Benjamin J. Raphael[2], Pavel A. Pevzner[3]
[1]UCSD Bioinformaics Program, [2]Computer Science Department, Brown University
[3]Computer Science Department, UCSD,, [4]mchaisso@bioinf.ucsd.edu

Abstract
The availability of comparative targeted genomic sequences allows for micro-inversion based phylogenetic reconstruction. We propose a new approach for identifying micro-inversions across different species, and show that micro-inversions provide a new source of low homoplasy evolutionary characters. These characters may be used as "certificates" to verify different branches in a phylogenetic tree turning the challenging problem of phylogeny reconstruction into a relatively simple algorithmic problem. Our approach allows us to show that there are around 400 micro-inversions between human and chimpanzee and to construct phylogeny on 38 mammalian species. We estimate that there exists thousands of micro-inversions in genomes of mammals from comparative sequencing projects, an untapped source of phylogenetic characters.

Motivation
- Sequence based phylogenetic reconstruction is subject to both homoplasy (independent mutation to the same state) and model dependent parameters (Bayesian prior, and substitution model).
- Multiple-genome rearrangement based phylogeny is a difficult problem to solve.
- Non-overlapping inversions represent rearrangements where the ancestral gene order may be reconstructed in polynomial time.

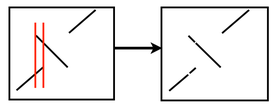
Detecting Micro-inversions
While micro-inversions represent powerful evolutionary characters, their detection is far from being simple. A naïve approach is to detect reverse-strand local alignments between orthologous sequences. However, reverse-strand local alignments may also be caused by palindromes and inverted repeats, ubiquitous genomic features that do not reflect any variations in the genomic architecture between two genomes, i.e., they may be detected within a single genome without a need to align to another genome. Reverse strand alignments may also be detected in inverted transpositions and more complex interleaving rearrangement events. Diagrammatic dot plots of sequences that generate such alignments are shown below.

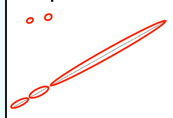


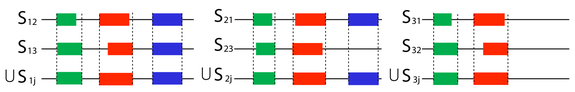
The *InvChecker* Method
Given sequences for a set of species, we would like to find all *inversion loci*: orthologous sequences that are involved in inversions. We developed a method, *InvChecker*, to find inversions using pairwise BLASTZ alignments of repeat-masked sequences, and GRIMM-Synteny, and assign orthology based on transitive relationships defined by the alignments. The inversion discovery and validation pipeline is shown below.

Step 1. Alignment

- Align all pairs of sequences using BLASTZ.
Each ungapped diagonal is a block.

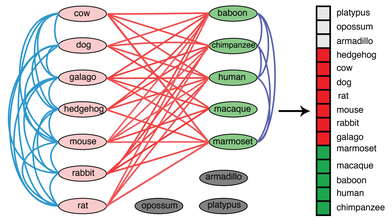
Step 2. Filter overlapping alignments

- Fragment overlapping blocks.
- Remove low-scoring blocks.

Step 3. Locate inverted segments.

- Detect synteny blocks using GRIMM-Synteny (blocks shown in red circles).
- For each synteny block, enumerate its composition BLASTZ blocks as 1... n in their order in one (reference) genome. Output $b_1 \ldots b_n$, the order of the blocks in the query genome, using $-b_i$ to denote a reversed block.
- Microinversions are negative blocks $-b_i$ where $|-b_i| = i$.

Step 4. Define inversion boundaries


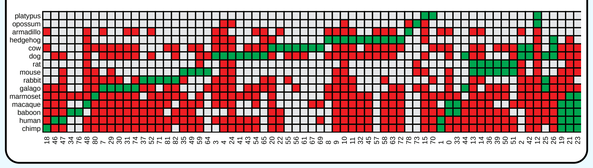
- Once all inverted blocks are discovered in pairwise alignments, use the union of coordinates of inversions for each species against all others to define the boundaries of inversions.

Step 5, Find divergent loci.
Locating divergent inversion loci: a "phylogenetic trek"
- Motivation: if an inversion locus is detected in a human/mouse, but not human/rat whole genome alignments, locate the locus in rat using the inverted mouse sequence.
- Constrain search in rat using orthologous positions defined by the net (main diagonal) of the mouse/rat alignment.
- Resulted in a 58% increase on the number of loci found in the ENCODE CFTR Region.

Step 6. Determine loci orientations


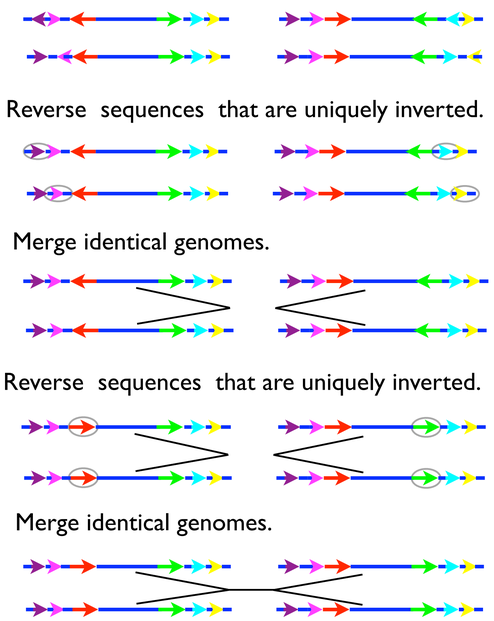
Assigning inversion orientation: inversion graph
- Red edges are in opposite orientation, blue same.
- Compare the alignment scores in forward and reverse directions to determine orientation.
- Consistent loci have a bipartite inversion graph on red edges.
- Discard inconsistent loci.

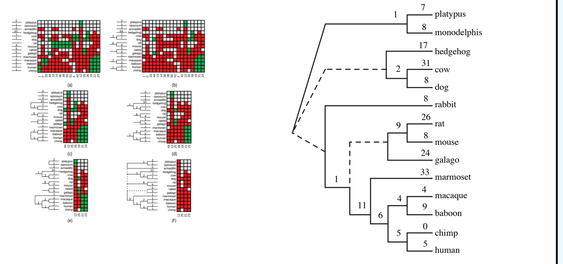
Example: Greater CFTR Region
InvChecker was ran on the greater CFTR region, a 1.7 Mb gene rich region from the human genome and its orthologous sequence from 14 other species sequenced for the ENCODE project. The result is shown below where each column is an inversion locus, red and green cells are in opposite orientation, and gray are of unknown (deleted or divergent) orientation.



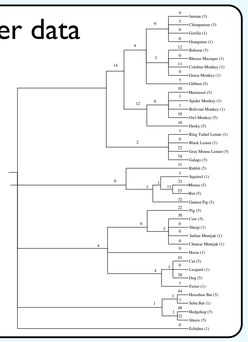
Detecting conflicting characters as *four gamete violations*
- The *four-gamete condition*: given an *m* species by *n* character matrix, each character ∈ {0,1} no two columns (characters) may have the rows { [0 0], [01], [10] [11]}.
- Use Maximum Conflict Removal to remove violating characters.

Constructing phylogeny with micro-inversions
- Two phase iterative technique:
  - Inversion-phase: find inversions that bring a species closer to all other species, and reverse them.
  - Merging-phase: merge all species with identical genomic architectures.
  - Repeat until there is only one species.

Example of phylogenetic reconstruction on four sequences.


Reconstruction on ENCODE CFTR region


Reconstruction on a larger data set
We constructed the phylogeny of 38 species using sequences from the NISC Comparative Vertebrate Sequencing Project. Between 1.7 and 6 Mb of sequence were available for each species.



Conclusion
The use of micro-inversions for phylogenetic reconstruction is simple and viable. The detection of ancient micro-inversions to resolve early branches in mammalian evolution represents a hurdle that will likely be surmountable with future whole-genome sequences.