Reviews in
Computational Biology

Multiple Whole Genome Alignments

Christophe Dessimoz

May 23rd, 2011

ETH Zürich inf | Informatik
Computer Science

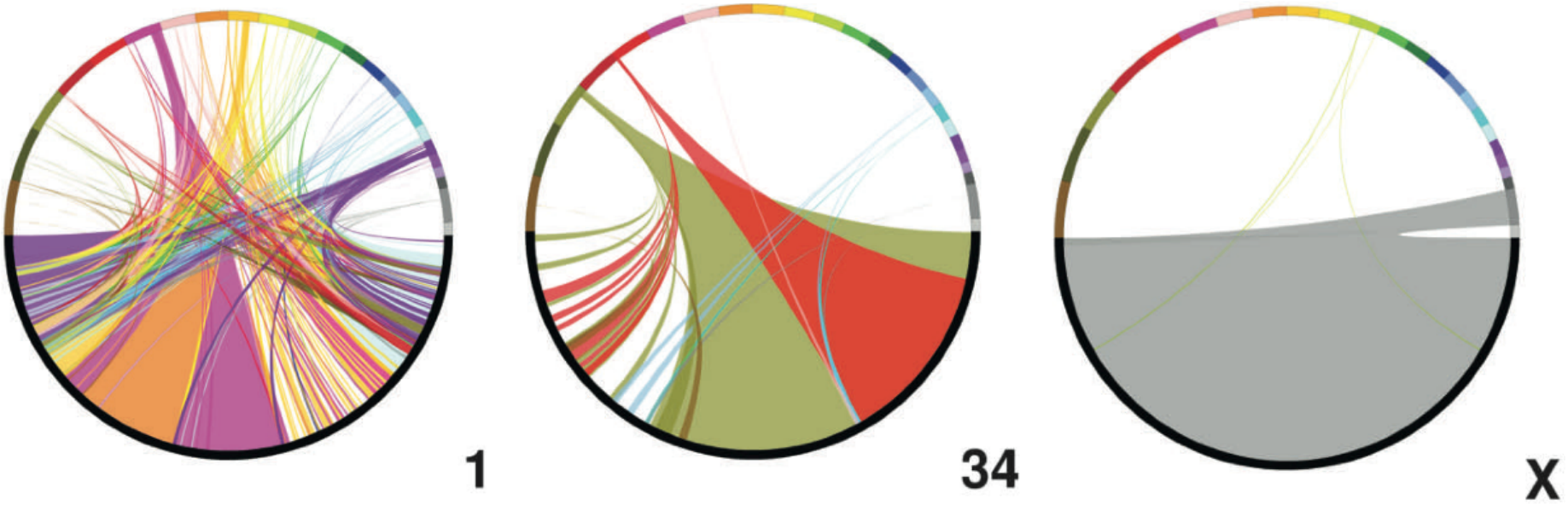
Outline

- **Background/Motivation**
  - Genome evolution
  - Previous reviews
- **Objectives of whole genome alignment**
- **Recent developments**
- **Benchmarking**

Genome Evolution

- “**Mutations**”
  - Inversions
  - (Retro-)Transpositions
  - Losses
  - Lateral gene transfer
  - Duplications (short segment, whole genome)
  - ….
- “**Branching**”
  - Speciation events

Human vs. Dog



Krzywinski et al. Circos: an information aesthetic for comparative genomics. Genome Research (2009) vol. 19 (9) pp. 1639-45

Previous Reviews

BRIEFINGS IN BIOINFORMATICS VOL 6. NO 1. 6–22. MARCH 2005

The many faces of sequence alignment

Serafim Batzoglou

- General survey of alignment at protein and genome levels

Human Molecular Genetics, 2006, Vol. 15, Review Issue 1

Evolution at the nucleotide level: the problem of multiple whole-genome alignment

Colin N. Dewey[1,*] and Lior Pachter[2]

- Brief survey of objectives
- Survey of methods

Computation and Analysis of Genomic Multi-Sequence Alignments

Mathieu Blanchette

Annual Review of Genomics and Human Genetics
2007, 8:193-213

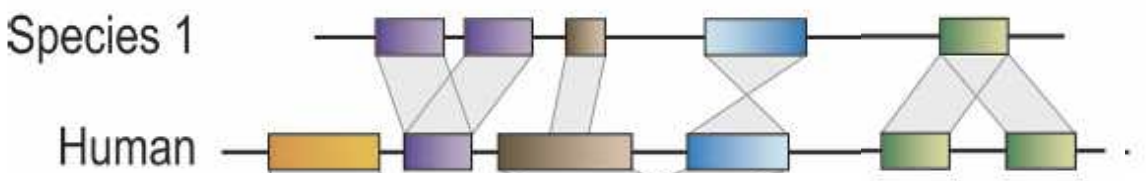
- Brief survey of objectives
- Survey of methods
- Benchmarking
- Applications

In this review

- Discuss the (sometimes conflicting) objectives of genome alignments.
- Illustrate these objectives with recent methodological developments.
- Show that inconsistent objectives complicates validation/comparison of methods.

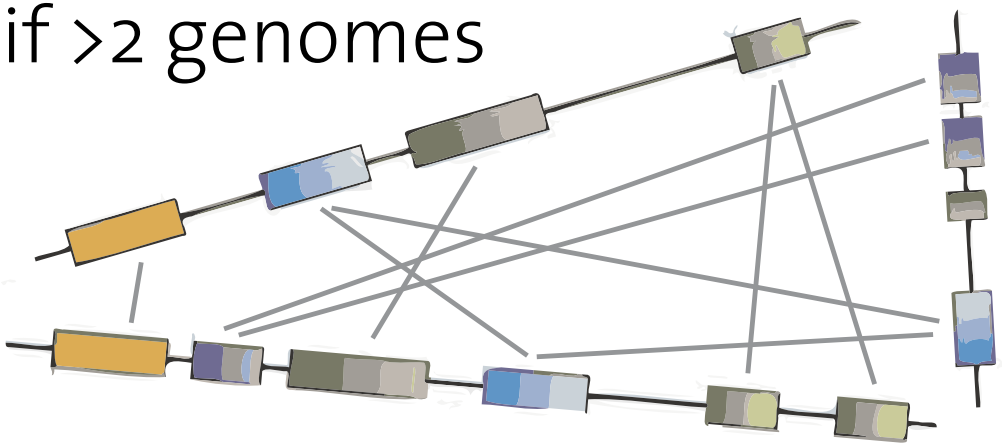
Alignment objectives

- In the absence of duplication, usually **homologous characters**
- Else, not so obvious, because no 1:1 relation between genomes:



Margulies et al. Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. Genome Research (2007) vol. 17 (6) pp. 760-74

- Especially if >2 genomes

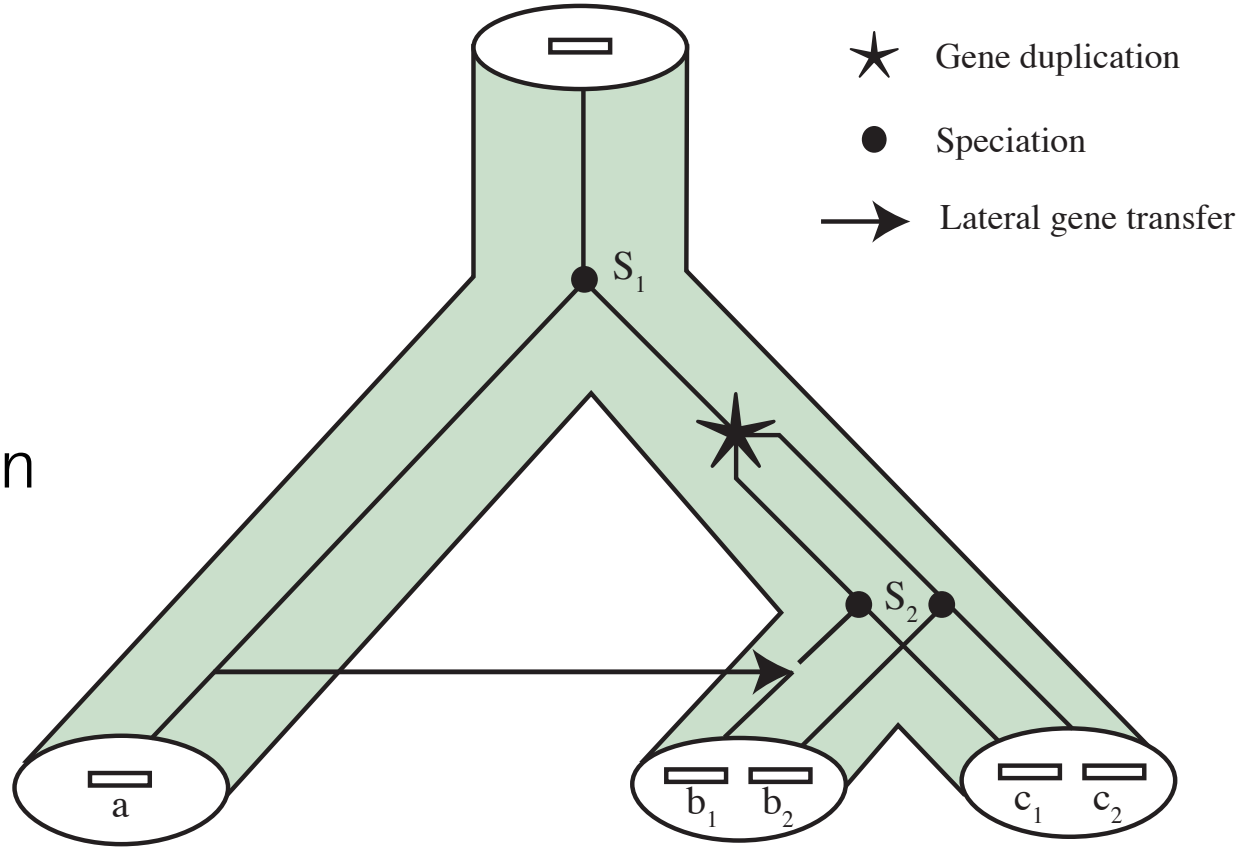


Refining Homology

Two genes (or characters) are *homologs* if they have a common ancestor.

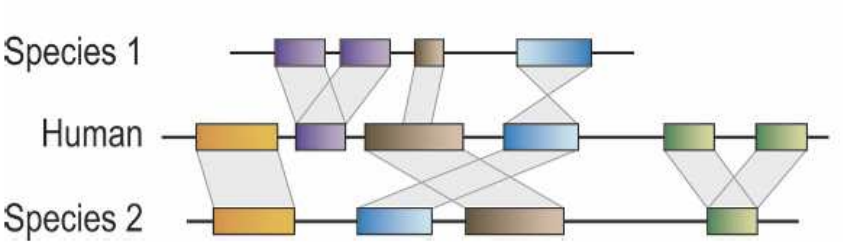
Main Subtypes

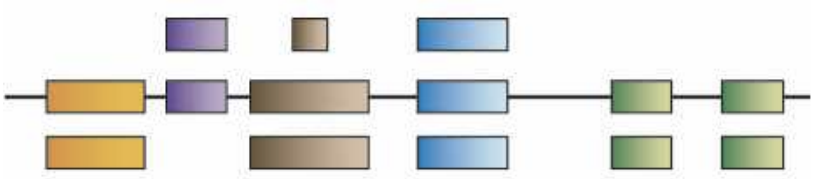
- *Orthologs*: through speciation
- *Paralogs*: through duplication
- *Xenologs*: through lateral transfer



Alignment Objectives

Align orthologs to a reference genome: define a reference genome; map each of its character to at most 1 orthologous character in each "target" species





e.g. Ensembl, Vista

Align 1-to-1 orthologs

e.g. Mauve

Align 1-to-1 positional orthologs ("monotopoorthologs")

e.g. Mercator/MAVID

Identify a more general structure ("threaded blocksets", graphs)

e.g. ABA, TBA, Enredo

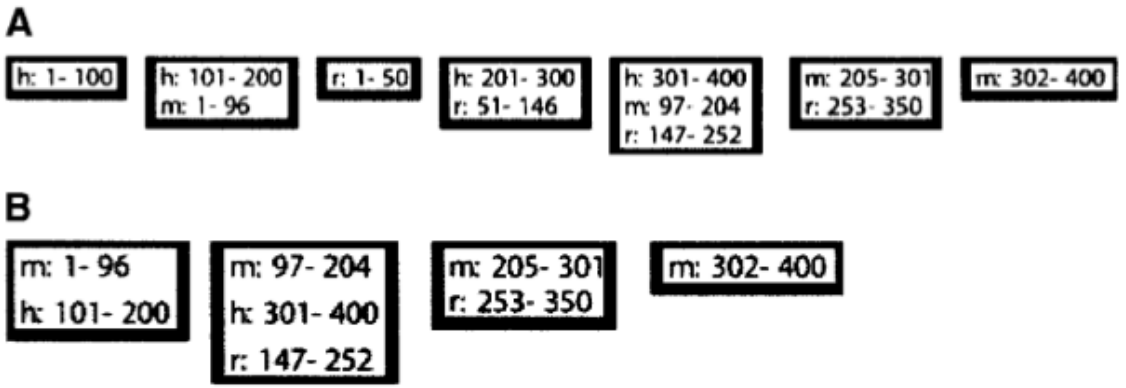
Do something else (unspecified)

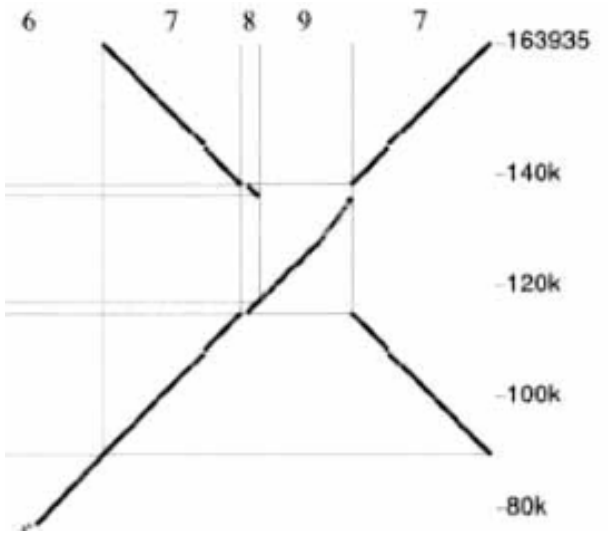
e.g. VPG Ancestral Align, M-GCAT

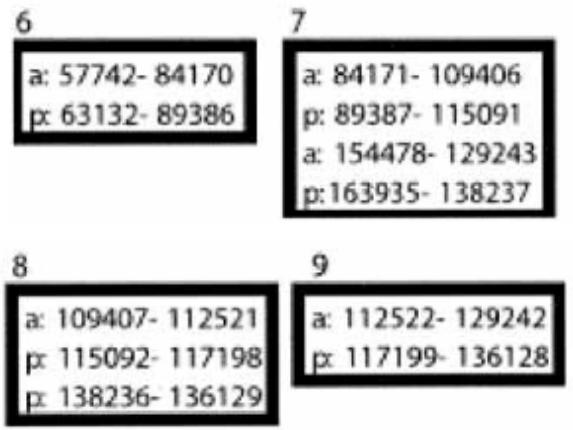
Margulies et al. Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. Genome Research (2007) vol. 17 (6) pp. 760-74

Threaded Blocks

h:human
m: mouse
r: rat



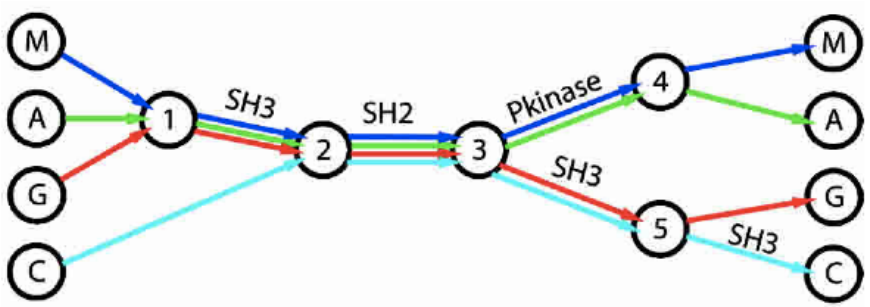




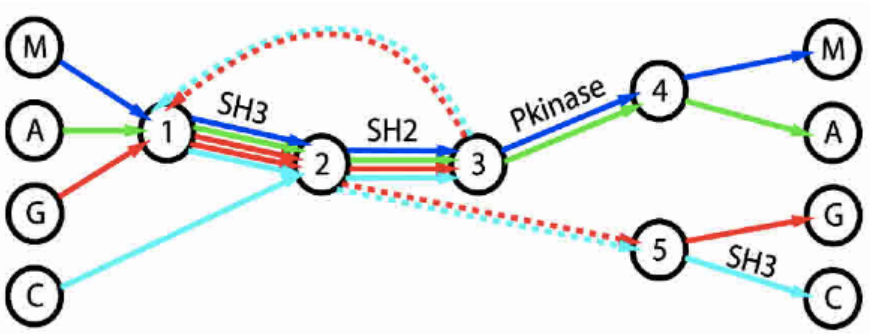
Blanchette et al. Aligning multiple genomic sequences with the threaded blockset aligner. Genome Research (2004) vol. 14 (4) pp. 708-15

A-Bruijn Graph

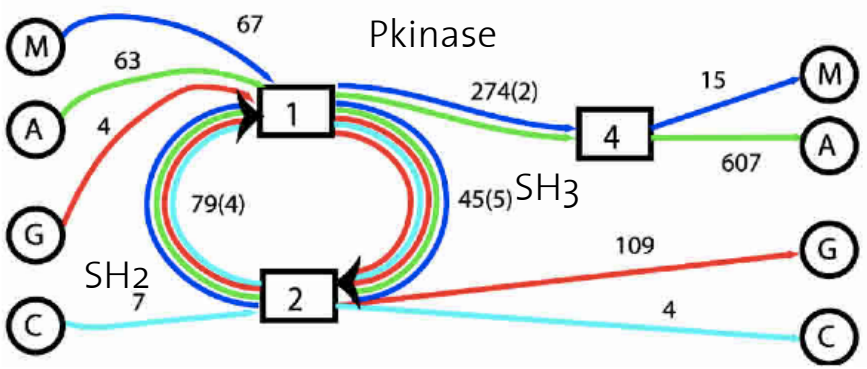
Partial Order Alignment



A-Bruijn Graph (simplified)

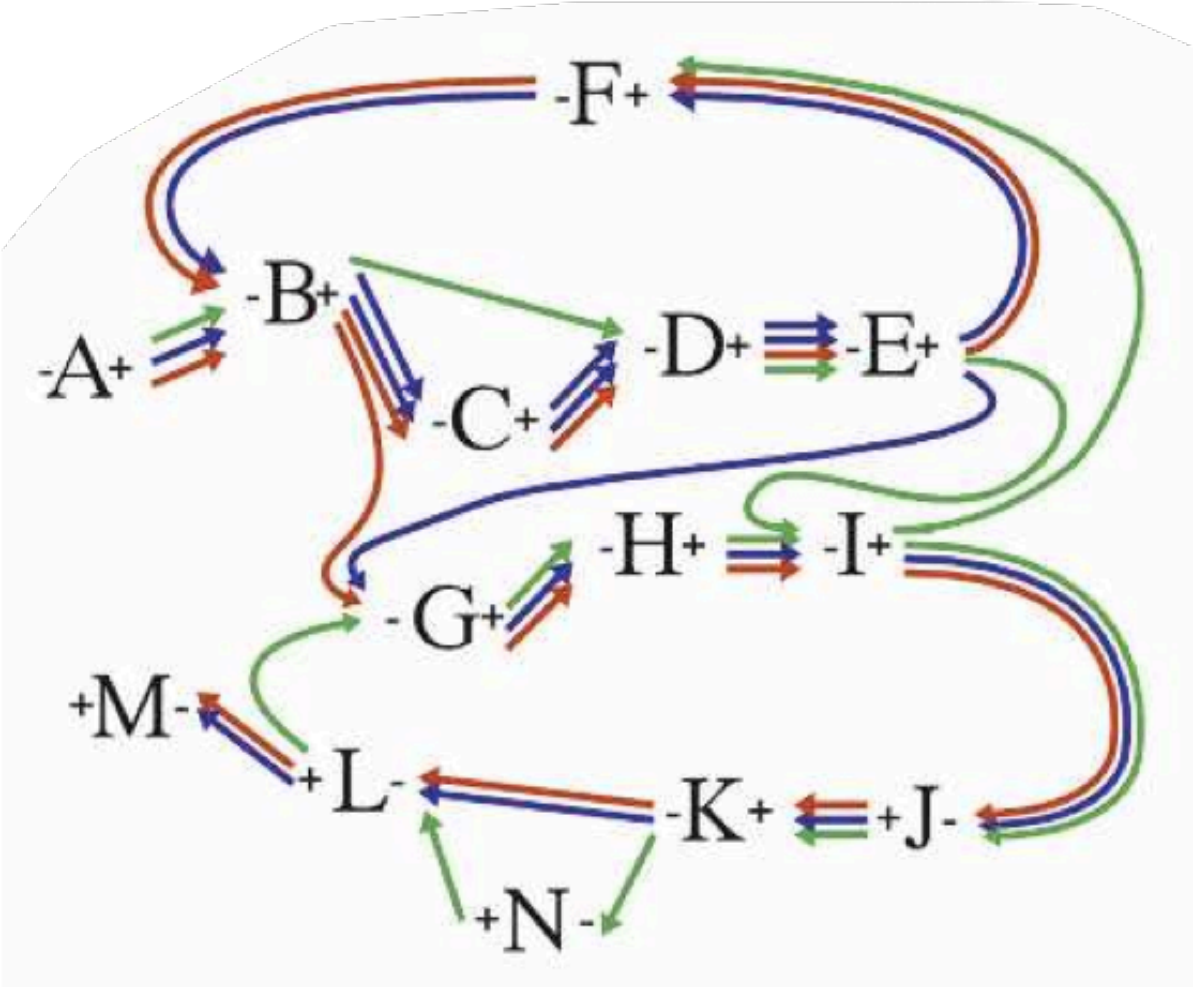


A-Bruijn Graph



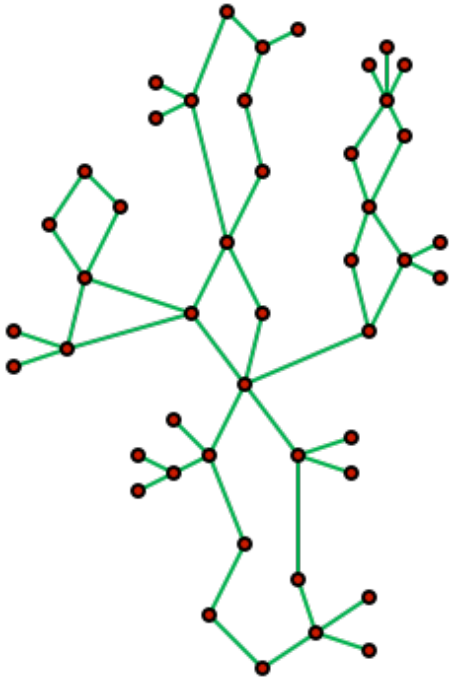
Raphael et al. A novel method for multiple alignment of sequences with repeated and shuffled elements. Genome Research (2004) vol. 14 (11) pp. 2336-46

Enredo Graph

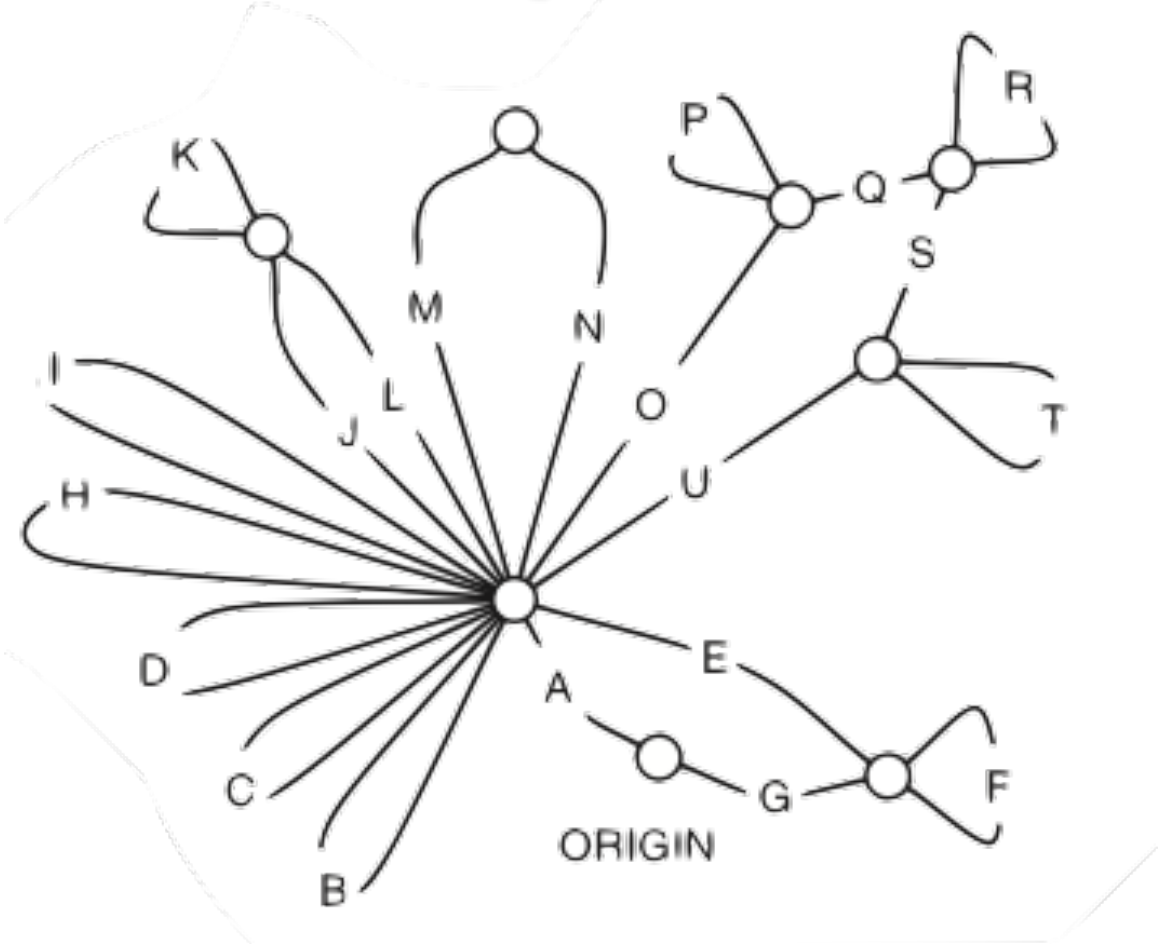


Paten et al. Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. Genome Research (2008) vol. 18 (11) pp. 1814-28

Cactus Graphs



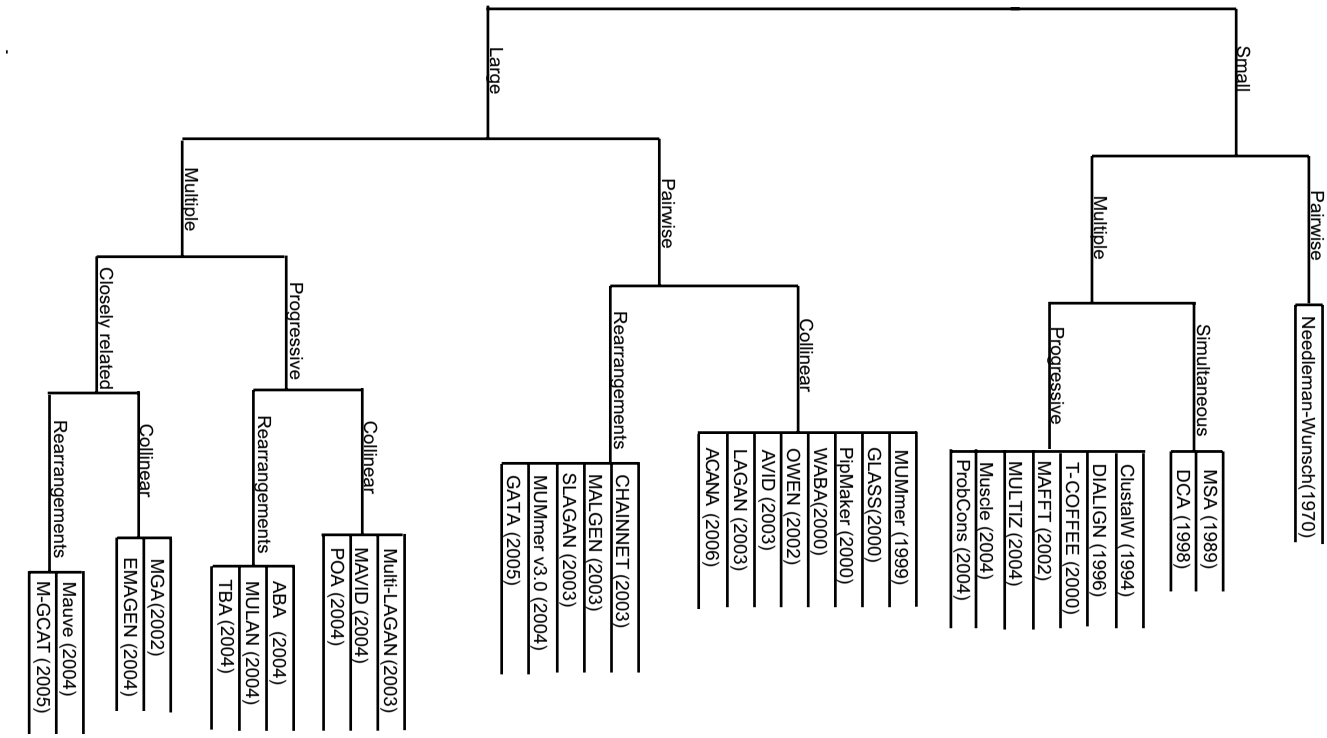
wikipedia



Paten et al. Cactus graphs for genome comparisons. Journal of computational biology (2011) vol. 18 (3) pp. 469-81

Typical Approach

- **Multiple whole-genome alignment is in most formulations NP-hard. Cannot be solved exactly.**
- **Typical (heuristic) approach:**
  1. Identify “anchors” (a.k.a.“seeds”, conserved short sequences) across genomes
  2. Use anchors to map co-linear homologous segments across genomes
  3. Align co-linear segments (efficiently)



Treangen and Messeguer. M-GCAT: interactively and efficiently constructing large-scale multiple genome comparison frameworks in closely related species. BMC Bioinformatics (2006) vol. 7 pp. 433

Some Tools Published Lately

BIOINFORMATICS ORIGINAL PAPER
Vol. 27 no. 3 2011, pages 334–342
doi:10.1093/bioinformatics/btq665
Sequence analysis
Advance Access publication December 9, 2010

Mugsy: fast multiple alignment of closely related whole genomes

Samuel V. Angiuoli[1,2,*] and Steven L. Salzberg[1]

Enredo and Pecan: Genome-wide mammalian consistency-based multiple alignment with paralogs

Benedict Paten,[1,3,4] Javier Herrero,[2,3] Kathryn Beal,[2] Stephen Fitzgerald,[2] and Ewan Birney[2,4]

[1]Center for Biomolecular Science and Engineering, University of California, Santa Cruz, California 95064, USA; [2]EMBL European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom

Pairwise whole-genome alignment involves the creation of a homology map, capable of performing a near complete transformation of one genome into another. For multiple genomes this problem is generalized to finding a set of consistent homology maps for converting each genome in the set of aligned genomes into any of the others. The problem can be divided into two principal stages. First, the partitioning of the input genomes into a set of colinear segments, a process which essentially deals with the complex processes of rearrangement. Second, the generation of a base pair level alignment map for each colinear segment. We have developed a new genome-wide segmentation program, Enredo, which produces colinear segments from extant genomes handling rearrangements, including duplications. We have then applied the new alignment program Pecan, which makes the consistency alignment methodology practical at a large scale, to create a new set of genome-wide mammalian alignments. We test both Enredo and Pecan using novel and existing assessment analyses that incorporate both real biological data and simulations, and show that both independently and in combination they outperform existing programs. Alignments from our pipeline are publicly available within the Ensembl genome browser.

[Supplemental material is available online at www.genome.org. Enredo and Pecan are freely available at http://www.ebi.ac.uk/~jherrero/downloads/enredo/ and http://www.ebi.ac.uk/~bjp/pecan/, respectively.]

Uncertainty in homology inferences: Assessing and improving genomic sequence alignment

Gerton Lunter,[1,3] A
Alexandre Caldeira

[1]MRC Functional Genetics
Oxford OX1 3QX, United K
Oxford, OX1 2TG, United K

Sequence alignment
particular, the stat
phylogenetic inferen
and simulation study
aligned bases are in
each leading to syste
alignment quality; however, these improvements are modest compared with the remaining alignment errors, even with exact knowledge of the evolutionary model, emphasizing the need for statistical approaches to account for uncertainty. We develop a new algorithm, Marginalized Posterior Decoding (MPD), which explicitly accounts for uncertainties, is less biased and more accurate than other algorithms we consider, and reduces the proportion of misaligned bases by a third compared with the
first nonheuristic algorithm for DNA sequence al
Needleman–Wunsch algorithm. Despite this, considerab
conclude that a probabilistic treatment is essential, both
uncertainty. This is becoming increasingly relevant wit
DNA, whose study relies heavily on alignments. Aligr
drawing conclusions from alignments. Software and al
http://genserv.anat.ox.ac.uk/grape/.

[Supplemental material is available online at www.genon

OPEN ACCESS Freely available online
PLoS one

progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement

Aaron E. Darling[1]*[¤], Bob Mau[2], Nicole T. Perna[3]

Contents lists available at ScienceDirect
Genomics
journal homepage: www.elsevier.com/locate/ygeno
ELSEVIER

NcDNAlign: Plausible multiple alignments of non-protein-coding genomic sequences

Dominic Rose [a], Jana Hertel [a], Kristin Reiche [a,d], Peter F. Stadler [a,d,b,c], Jörg Hackermüller [d,*]

[a] Bioinformatics Group, Department of Computer Science, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany
[b] Department of Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria
[c] Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA

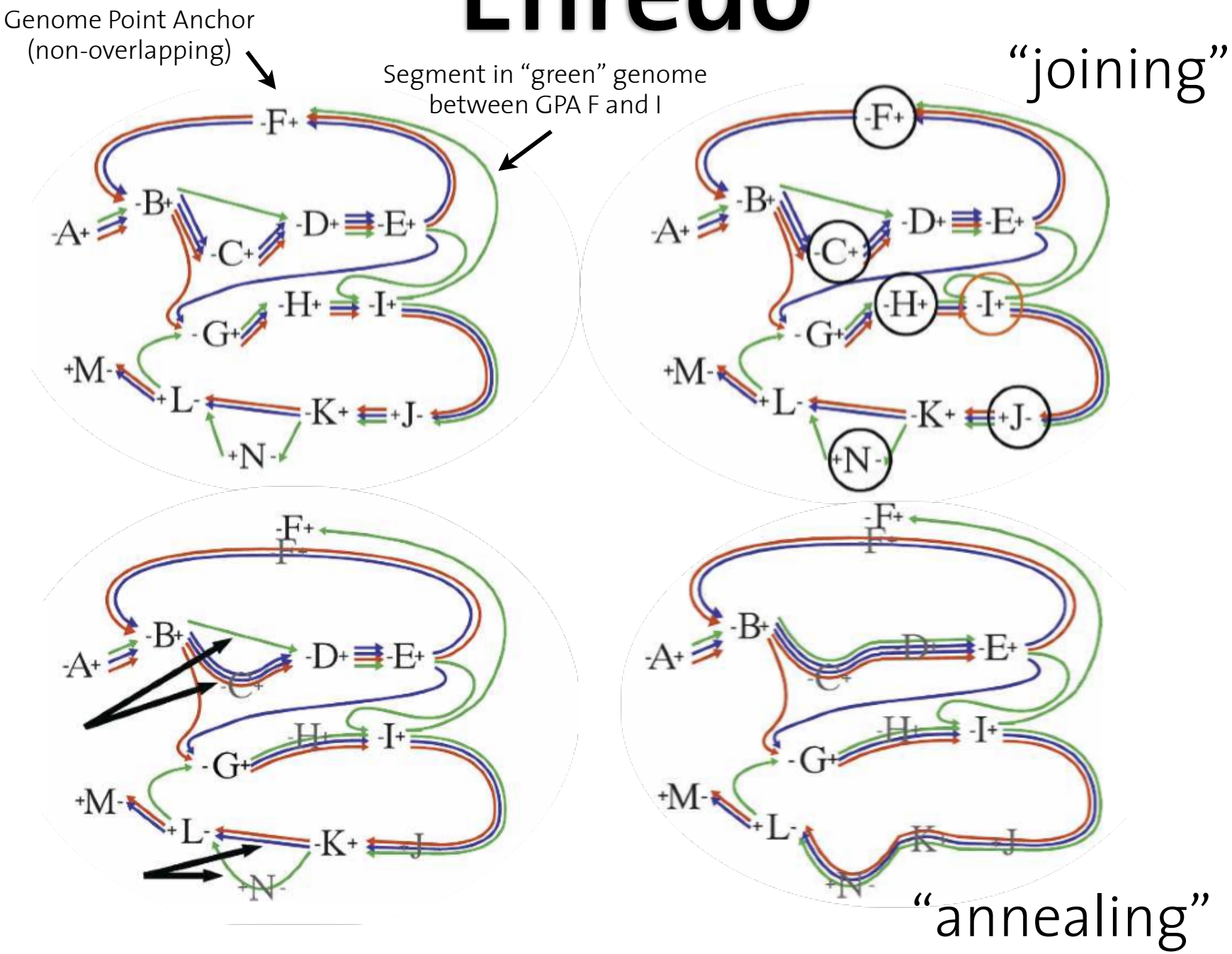
Multiple whole-genome alignments without a reference organism

Inna Dubchak,[1,2] Alexander Poliakov,[1] Andrey Kislyuk,[3] and Michael Brudno[4,5]

[1]Genome Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA; [2]DOE Joint Genome Institutes, Walnut Creek, California 94598, USA; [3]Department of Computer Science, Georgia Institute of Technology, Atlanta, Georgia 30332, USA; [4]Department of Computer Science, Banting and Best Department of Medical Research, and Centre for Analysis of Genome Evolution and Function, University of Toronto, Toronto, Ontario M5R 3G4, Canada

Multiple sequence alignments have become one of the most commonly used resources in genomics research. Most algorithms for multiple alignment of whole genomes rely either on a reference genome, against which all of the other sequences are laid out, or require a one-to-one mapping between the nucleotides of the genomes, preventing the alignment of recently duplicated regions. Both approaches have drawbacks for whole-genome comparisons. In this paper we present a novel symmetric alignment algorithm. The resulting alignments not only represent all of the genomes equally well, but also include all relevant duplications that occurred since the divergence from the last common ancestor. Our algorithm, implemented as a part of the VISTA Genome Pipeline (VGP), was used to align seven vertebrate and six *Drosophila* genomes. The resulting whole-genome alignments demonstrate a higher sensitivity and specificity than the pairwise alignments previously available through the VGP and have higher exon alignment accuracy than comparable public whole-genome alignments. Of the multiple alignment methods tested, ours performed the best at aligning genes from multigene families—perhaps the most challenging test for whole-genome alignments. Our whole-genome multiple alignments are available through the VISTA Browser at http://genome.lbl.gov/vista/index.shtml.

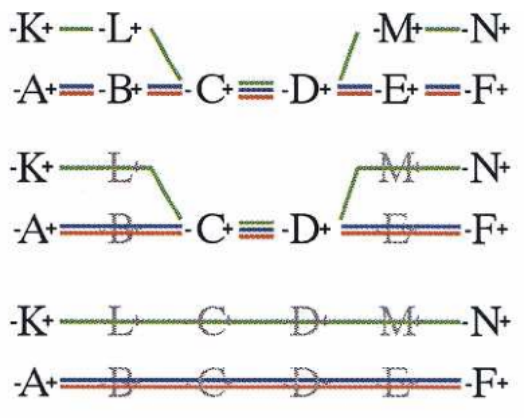
Enredo



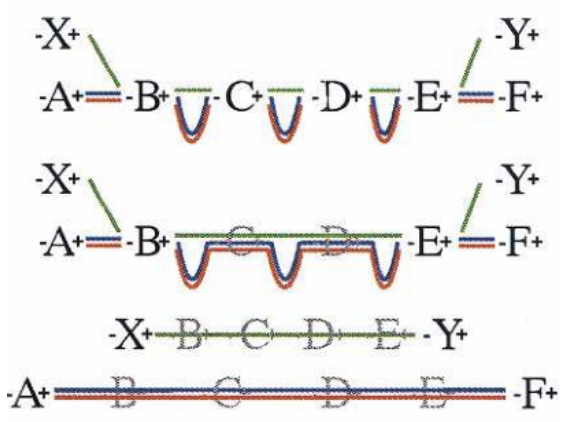
Paten et al. Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. Genome Research (2008) vol. 18 (11) pp. 1814-28

Enredo

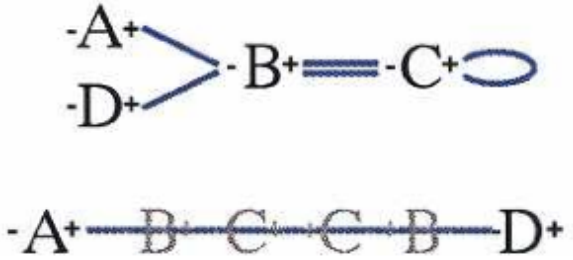
Heuristics on Graph:



Limits disruptive effect of transposable elements and repeats



Limits disruptive effect of retrotransposed genes



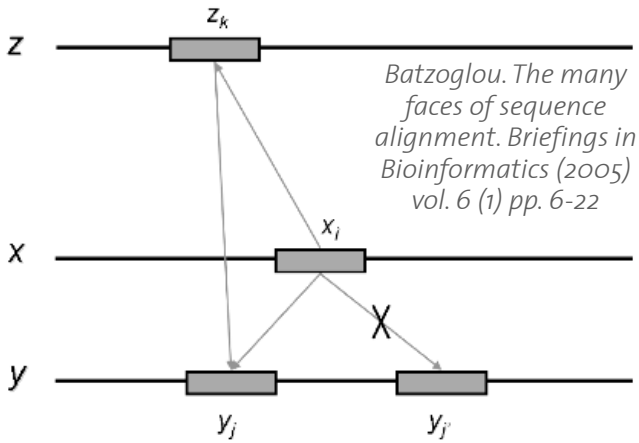
Limits disruptive effect of short tandem repeats

Paten et al. Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. Genome Research (2008) vol. 18 (11) pp. 1814-28

Pecan

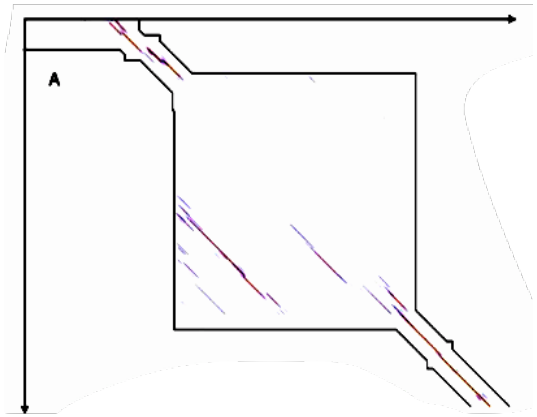
Consistency-based alignment of long co-linear segments

$$P'(x_i \Diamond y_j|\theta) = \frac{1}{n-1} \cdot \left(P(x_i \Diamond y_j|\theta) + \sum_{z \in \mathcal{L}-\{x,y\}} \sum_k P(x_i \Diamond z_k|\theta) \cdot P(y_j \Diamond z_k|\theta) \right)$$

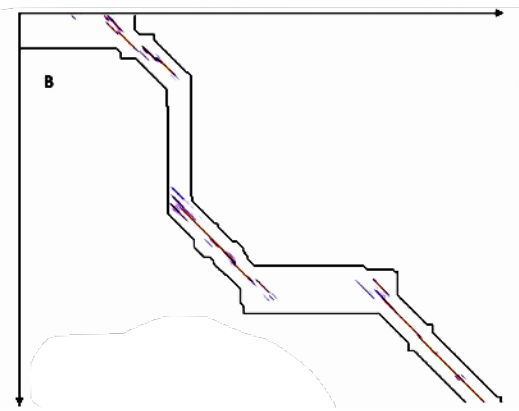


Batzoglou. The many faces of sequence alignment. Briefings in Bioinformatics (2005) vol. 6 (1) pp. 6-22

Consistency

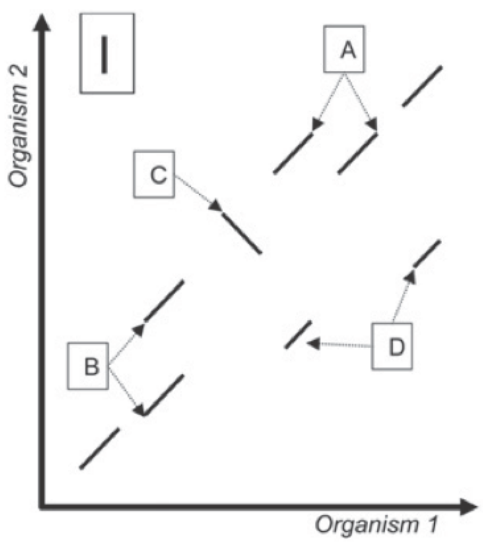


Constrained Alignment

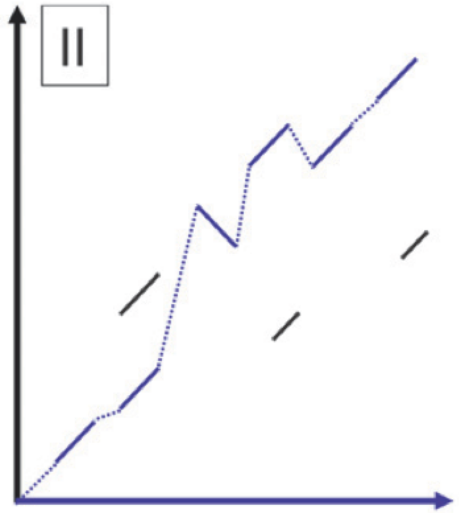


Transitive Anchoring

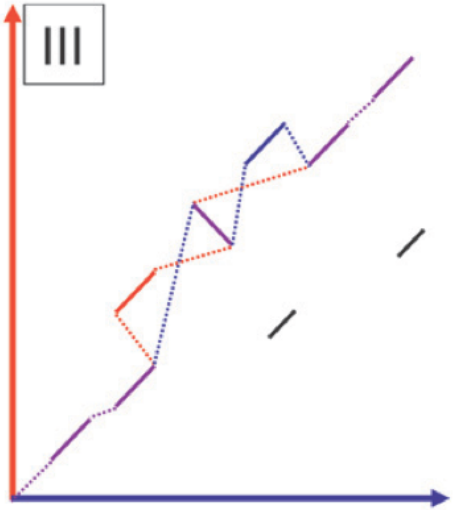
Vista “Ancestral Align”



CHAOS
local alignments
(akin to *BLAST*)



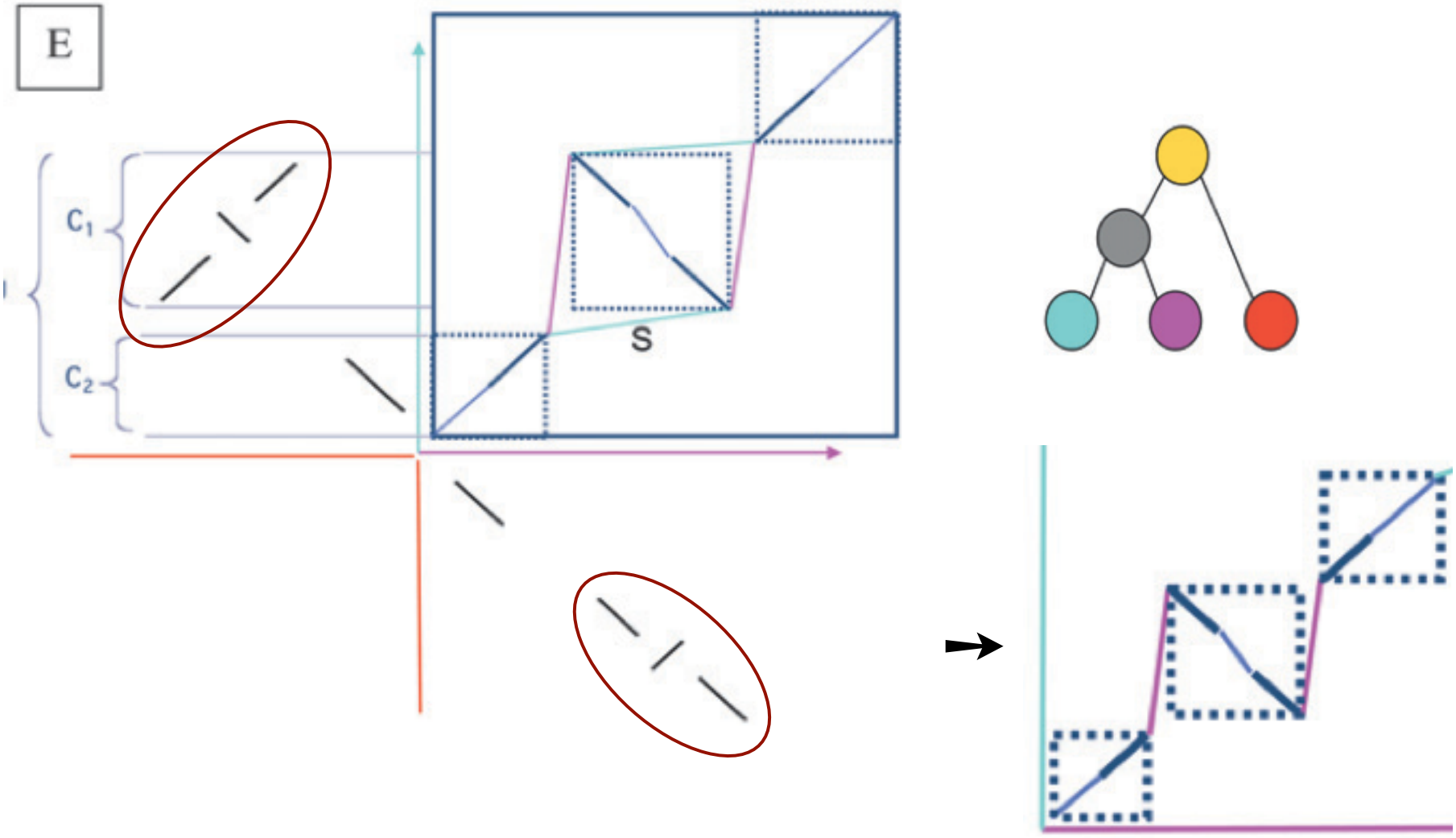
Shuffle LAGAN
highest scoring
1-monotonic map
w.r.t. 1st genome (blue)



Shuffle LAGAN
highest scoring
1-monotonic map
w.r.t. 2nd genome (red)

Dubchak et al. Multiple whole-genome alignments without a reference organism. Genome Research (2009) vol. 19 (4) pp. 682-9

Vista “Ancestral Align”



Benchmarking

Simulation

Surrogate measures

e.g. coverage: e.g. % of total length aligned

conservation of exon structure (synteny, strand, order)

different substitution patterns @3rd codon (in exons)

"known" features of specific dataset (e.g. ALU seqs only in primates, ancestral repeats in all mammals, few rearrangements in X chromosome)

Statistical tests, information criteria

e.g. null hypothesis: sequences are unrelated

If for all bipartitions, Ho is rejected, conclude that the sequences are homologous

Benchmarking

Simulation

Towards realistic benchmarks for multiple alignments of non-coding sequences

Jaebum Kim[1], Saurabh Sinha[1,2*]

Progressive Mauve: Multiple alignment of genomes with gene flux and rearrangement

Aaron E. Darling[1,2,3] Bob Mau[4]

Nicole T. Perna[5]

Surrogate measures

Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome

Elliott H. Margulies, Gregory M. Cooper, George Asimenos, et al.

Genome Res. 2007 17: 760-774

Research Article

Assessing the Quality of Whole Genome Alignments in Bacteria

Firas Swidan and Ron Shamir

Enredo and Pecan: Genome-wide mammalian consistency-based multiple alignment with paralogs

Benedict Paten, Javier Herrero, Kathryn Beal, et al.

Uncertainty in homology inferences: Assessing and improving genomic sequence alignment

Gerton Lunter,[1,3] Andrea Rocco,[2] Naila Mimouni,[2] Andreas Heger,[1] Alexandre Caldeira,[2] and Jotun Hein[2]

NATURE BIOTECHNOLOGY VOLUME 28 NUMBER 6 JUNE 2010

Comparative assessment of methods for aligning multiple genome sequences

Xiaoyu Chen & Martin Tompa

Multiple whole-genome alignments without a reference organism

Inna Dubchak, Alexander Poliakov, Andrey Kislyuk, et al.

Statistical tests, information criteria

Method

Open Access

Measuring the accuracy of genome-size multiple alignments

Amol Prakash*† and Martin Tompa*

NATURE BIOTECHNOLOGY VOLUME 28 NUMBER 6 JUNE 2010

Comparative assessment of methods for aligning multiple genome sequences

Xiaoyu Chen & Martin Tompa

Challenges

- **Simulation:** genome evolution is hard to model "realistically". Strongly dependent on model assumptions.
- **Surrogate measures:** indirect, only test some aspects. Surrogates must be reliable.
- **Statistical tests/criteria:** merely rejecting null hypothesis may not be so informative. Difficult to rank methods. Dependent on model assumptions.
- **The methods often differ in their (implicit or explicit) objectives**
  → comparison difficult or misleading

Examples

Software

Open Access

M-GCAT: interactively and efficiently constructing large-scale multiple genome comparison frameworks in closely related species

Todd J Treangen* and Xavier Messeguer

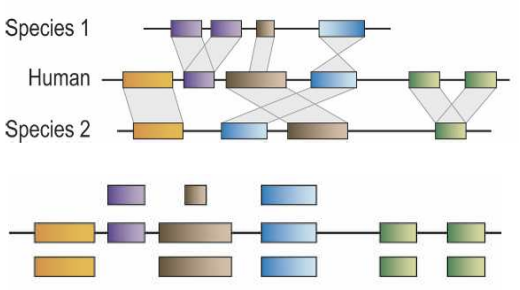
"We use the COG [41] identifier to determine if two or more genes from distinct genomes are orthologous."

Detecting non-orthology in the COGs database and other approaches grouping orthologs using genome-specific best hits

Christophe Dessimoz*, Brigitte Boeckmann[1], Alexander C. J. Roth and Gaston H. Gonnet

"Our results show that a very significant fraction of the COG groups include non-orthologs: using conservative parameters, the algorithm detects non-orthology in a third of all COG groups."

Apples vs. Oranges?



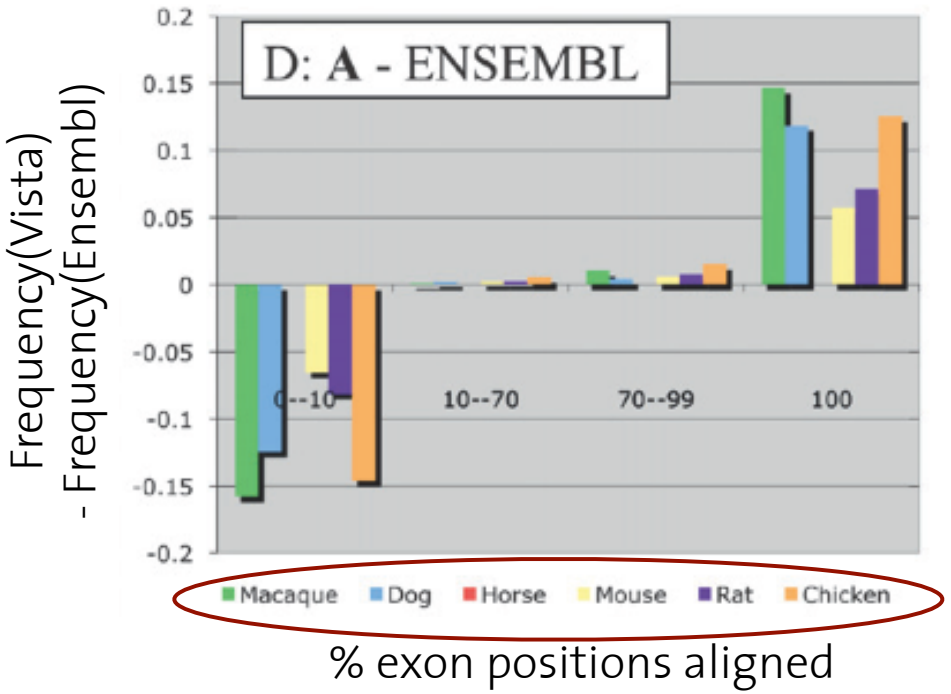


Table 2. A comparison of the alignments at the UCSC Genome Browser, Ensembl, and our alignments (VISTA) based on Inparanoid gene clusters

(Human-Mouse)

	VISTA		UCSC/MultiZ		Ensembl/Enredo	
	Genes	Exons	Genes	Exons	Genes	Exons
Aligned to gene/exon (of 13,780 genes, 141,244 exons)	13,444 97.6%	134,446 95.2%	13,207 95.8%	133,498 94.5%	11,592 84.1%	113,971 80.7%
Of these, aligned to orthologs only	12,978 94.2%	133,264 94.4%	13,170 95.6%	133,363 94.4%	11,567 83.9%	113,897 80.6%
Of these, aligned to orthologs and paralogs	417 3.0%	943 0.7%	19 0.1%	7 0%	11 0.1%	2 0%
Of these, aligned to paralogs only	49 0.4%	239 0.2%	18 0.1%	128 0.1%	14 0.1%	72 0.1%
Aligned to any ortholog, many–many clusters (of 182 genes, 862 exons)	128 70.3%	549 63.7%	112 61.5%	475 55.1%	38 20.9%	162 18.8%
Of these, aligned to all orthologs	39 21.4%	126 14.6%	4 2.2%	2 0.2%	1 0.5%	1 0.1%
Aligned to any ortholog, one–many clusters (of 305 genes, 2500 exons)	242 79.3%	2131 85.2%	226 74.1%	1909 76.4%	133 43.6%	1153 46.1%
Of these, aligned to all orthologs	97 31.8%	655 26.2%	7 2.3%	36 1.4%	7 2.3%	46 1.8%

Dubchak et al. Multiple whole-genome alignments without a reference organism. Genome Research (2009) vol. 19 (4) pp. 682-9

Conversely....

Table 2. Broad coverage statistics of the three segmentation methods

Method	No. of blocks[a]	N50 on human[b]	Percent of human bases[c]	Percent of full human genes[d]	Percent of partial genes[e]	Percent of genes covered[f]
Mercator	4436	832,834	44.46	59.0	25.9	85
MULTIZ	16,74,1834	35,679	75.70	37.4	60.1	97.5
Enredo	29,323	237,998	84.47	80.6	9.4	90

[a]The total number of blocks in the segmentation.
[b]The weighted median (N50) of segment lengths, using the human as the reference.
[c]The percentage of bases in the human genome covered by the segmentation.
[d]The percentage of human genes fully contained within the segmentation.
[e]The percentage of genes partially contained within the segmentation.
[f]The rounded sum of the previous two columns.

Paten et al. Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. Genome Research (2008) vol. 18 (11) pp. 1814-28

Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome

Elliott H. Margulies, Gregory M. Cooper, George Asimenos, et al.

Genome Res. 2007 17: 760-774

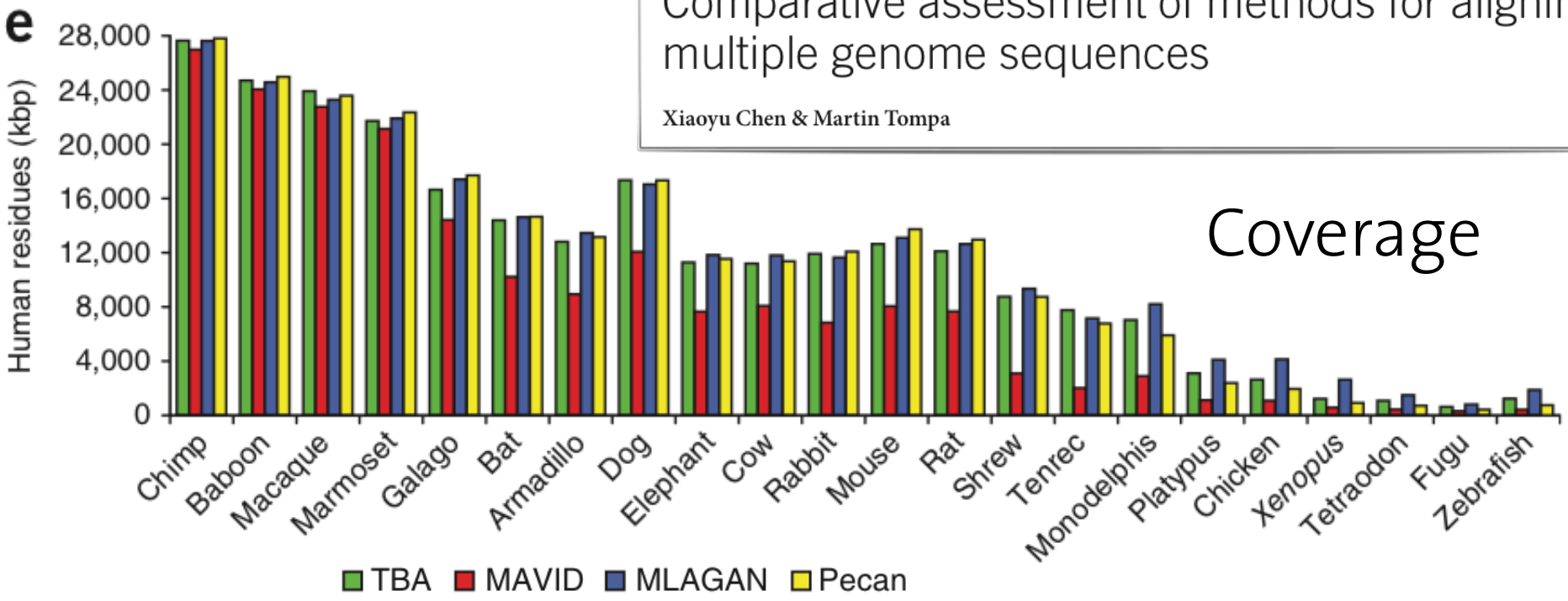
Explaining alignment discrepancies

We observed substantial differences between the four alignments; determining the sources for these differences is difficult, but a few conclusions can be drawn. For example, MAVID's lower coverage estimates likely result from the strict one-to-one orthology requirement, which eliminates human-specific duplications. The discrepancy in coverage between MAVID and the other aligners that is due to this restriction can be upper-bounded by

NATURE BIOTECHNOLOGY VOLUME 28 NUMBER 6 JUNE 2010

Comparative assessment of methods for aligning multiple genome sequences

Xiaoyu Chen & Martin Tompa



Coverage

Alignment accuracy

Wherever alignments do not agree, which alignment, if any, is correct? This is difficult to assess because the true alignment (the one that aligns all and only orthologous residues) is inherently unknown.

ONLINE METHODS

Comparison percentages. The ENCODE alignments are human-centric and represented in the coordinates of the human sequence. We therefore use the human sequence as our reference when measuring the level of agreement of the alignments. We compare, for each coordinate h in the human sequence and

Conclusion

- Multiple whole genome alignment is an active area of research, fueled by sequencing revolution.
- Whole genome aligners can differ considerably in the objectives they pursue.
- This complicates benchmarking (an already difficult problem)