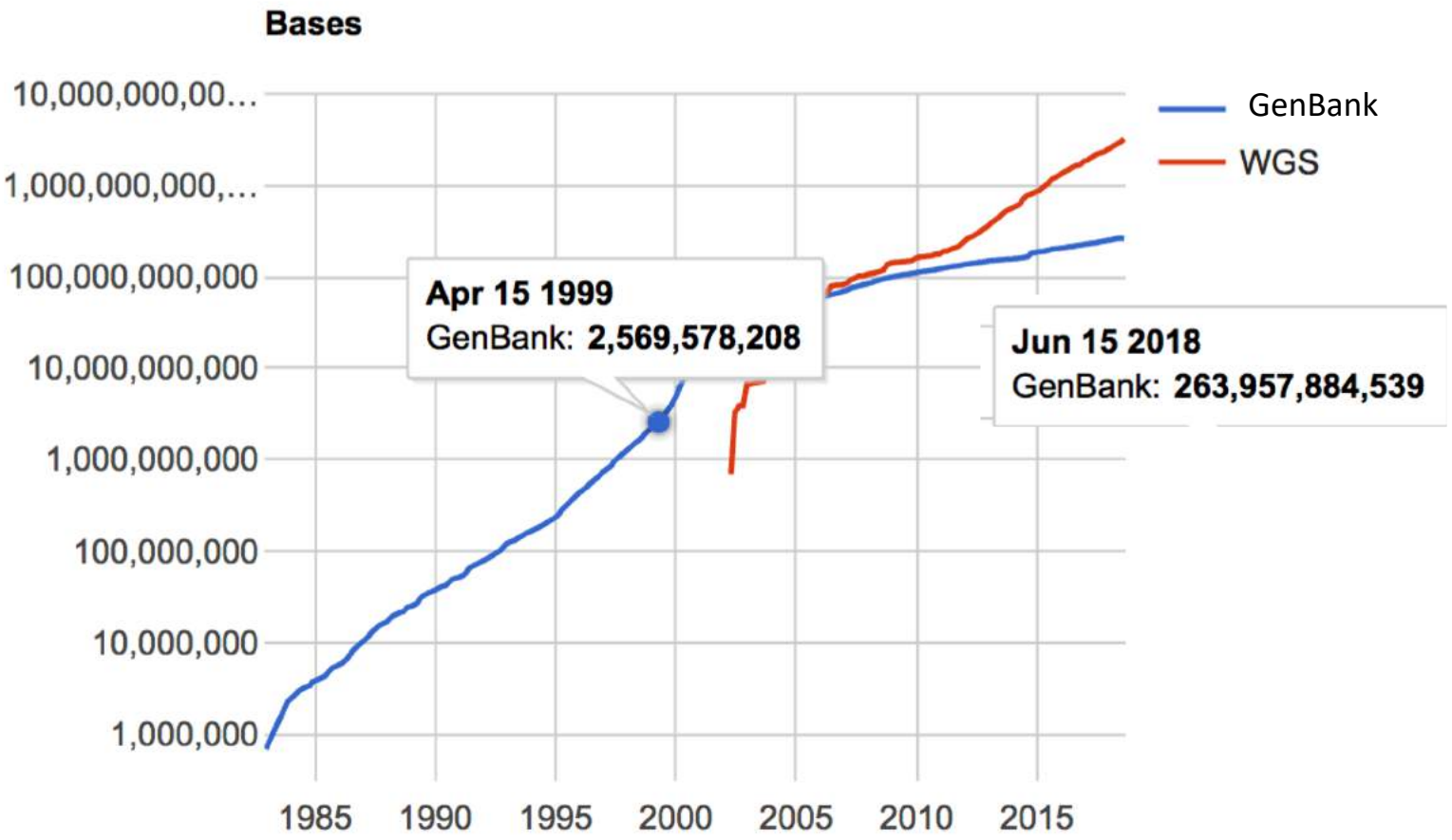
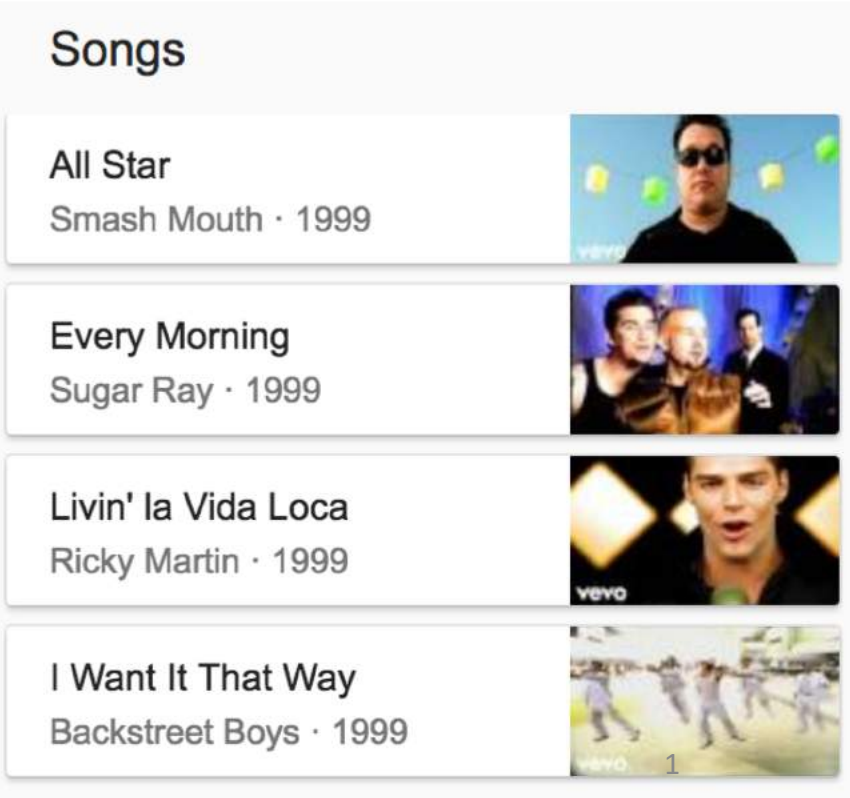
The MUMmer Saga

Charlotte Darby

10/11/18

Genomics was nascent in 1999





https://www.ncbi.nlm.nih.gov/genbank/statistics/ Accessed 4 October 2018

MUMmer was developed to produce “a very detailed and inclusive base-to-base mapping between two sequences”

Nucleic Acids Research, 1999, Vol. 27, No. 11 **2369–2376**

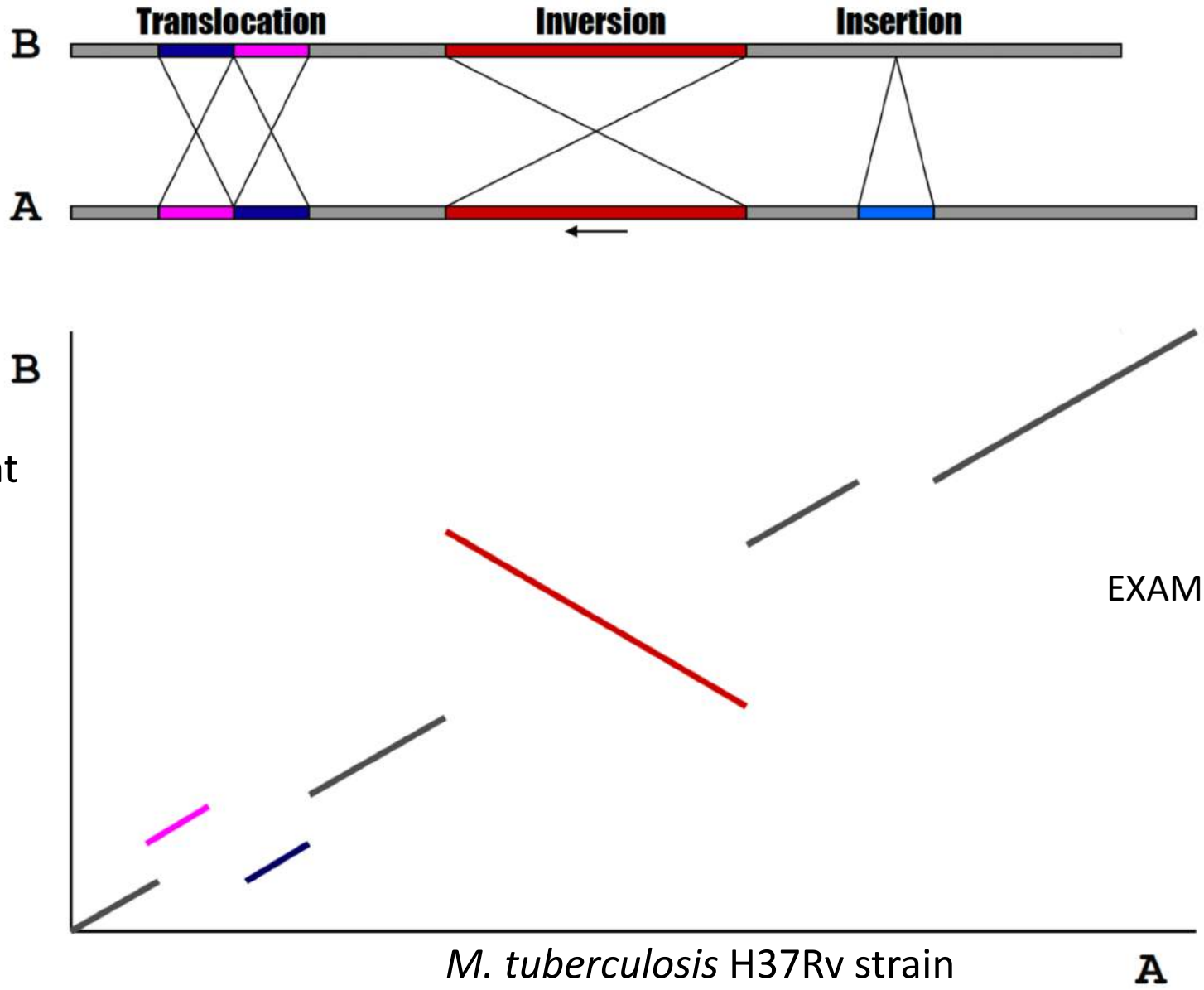
Alignment of whole genomes

Arthur L. Delcher[1,2], Simon Kasif[3], Robert D. Fleischmann[4], Jeremy Peterson[4], Owen White[4] and Steven L. Salzberg[4,*]

[1]Department of Computer Science, Loyola College in Maryland, Baltimore, MD 21210, USA, [2]Celera Genomics, 45 West Gude Drive, Rockville, MD 20850, USA, [3]Department of Electrical Engineering and Computer Science, University of Illinois, Chicago, IL 60607, USA and [4]The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA

Received February 2, 1999; Revised and Accepted April 14, 1999

A dotplot is used to visualize a pairwise alignment



New assembly of a different *M. tuberculosis* strain

EXAMPLE: not real data

Figure from Mike Schatz,
Applied Comparative Genomics 2018

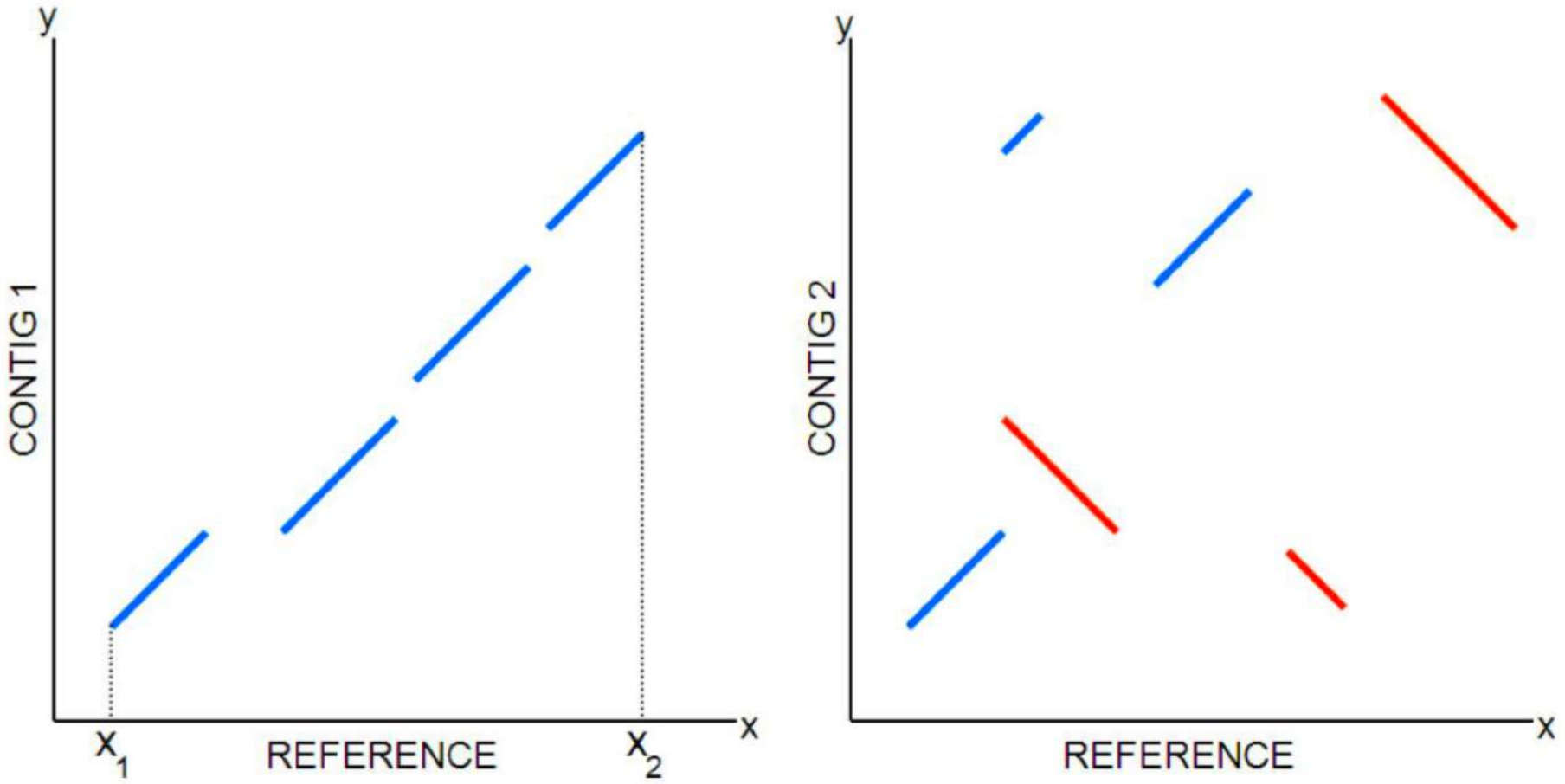
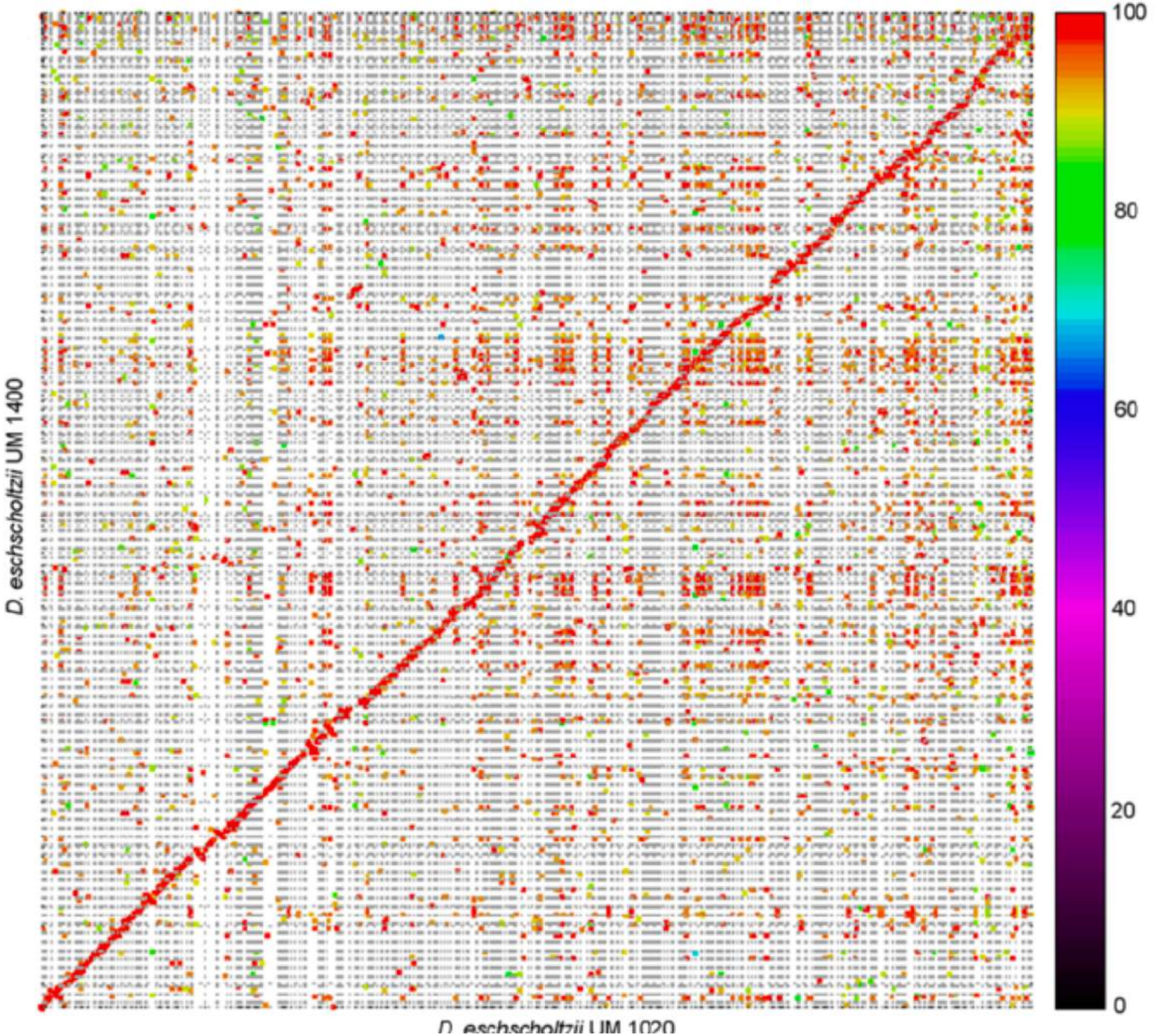


Figure 1
The dot plot of the alignment of two different contigs to a reference sequence. The dot plot of two different contigs as they are aligned to a reference genome. Contig 1 (left) strongly aligns to the region between x_1 and x_2, but Contig 2 (right) does not align as a whole.

González and Liao, Clustering exact matches of pairwise sequence alignments by weighted linear regression, BMC Bioinformatics 2008



"Dot-plot representing whole genome comparison between UM 1400 and UM 1020 isolates. The comparison was performed using MUMmer 3.23 (NUCmer). Dots closest to the diagonal line represent co-linearity between the two genomes. The dots are color-coded as indicative of percent sequence identity. The red dots or diagonal line depicts the high sequence identity between both genomes."

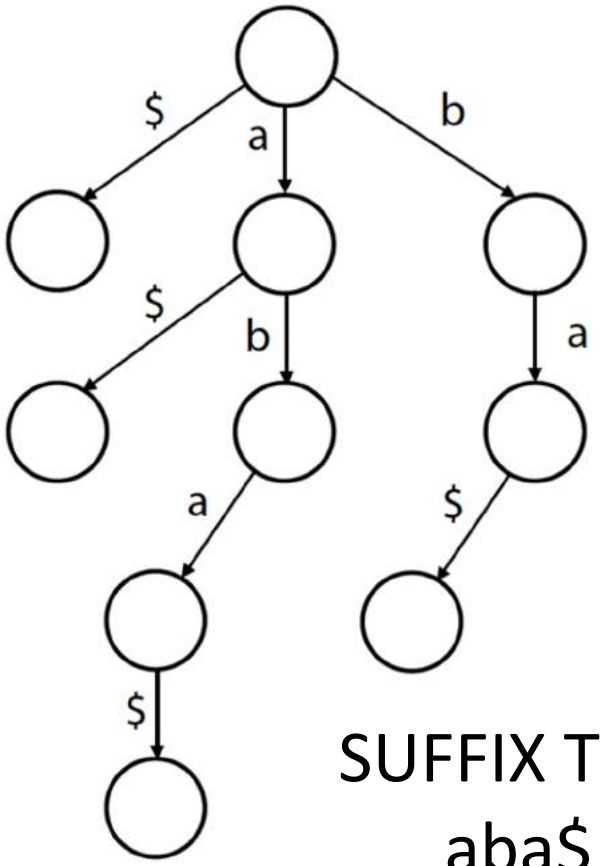
Chan et al., Genome analysis of *Daldinia eschscholtzii* strains UM 1400 and UM 1020, wood-decaying fungi isolated from human hosts, BMC Genomics 2015

Genome size ~ 35 Mb
UM 1400: 104 scaffolds > 1000bp
UM 1020: 598 scaffolds > 1000 bp

MUMmer's pairwise alignment algorithm

1. Find all maximal unique matching subsequences
   - MUM: "subsequence that occurs exactly once in A and once in B and is not contained in any longer such sequence"
   - Assume that MUMs are part of the desired global alignment
   - Build **suffix tree** of both genomes together and scan it (single pass)
2. Sort MUMs (á la longest increasing subsequence)
3. Refine locally
   - SNP, insertion, highly polymorphic, repeat
4. Report

Recall the suffix trie data structure: every path from root to leaf corresponds to a suffix in the string



SUFFIX TRIE

aba$

Making the suffix trie into a suffix tree:

- Collapse non-branching paths
- Label edges with the offsets in the original string of the letters they stand for
- Label nodes with the offset in the original string the suffix comes from

The $ is a terminator character that is not in the string

Figure from Ben Langmead

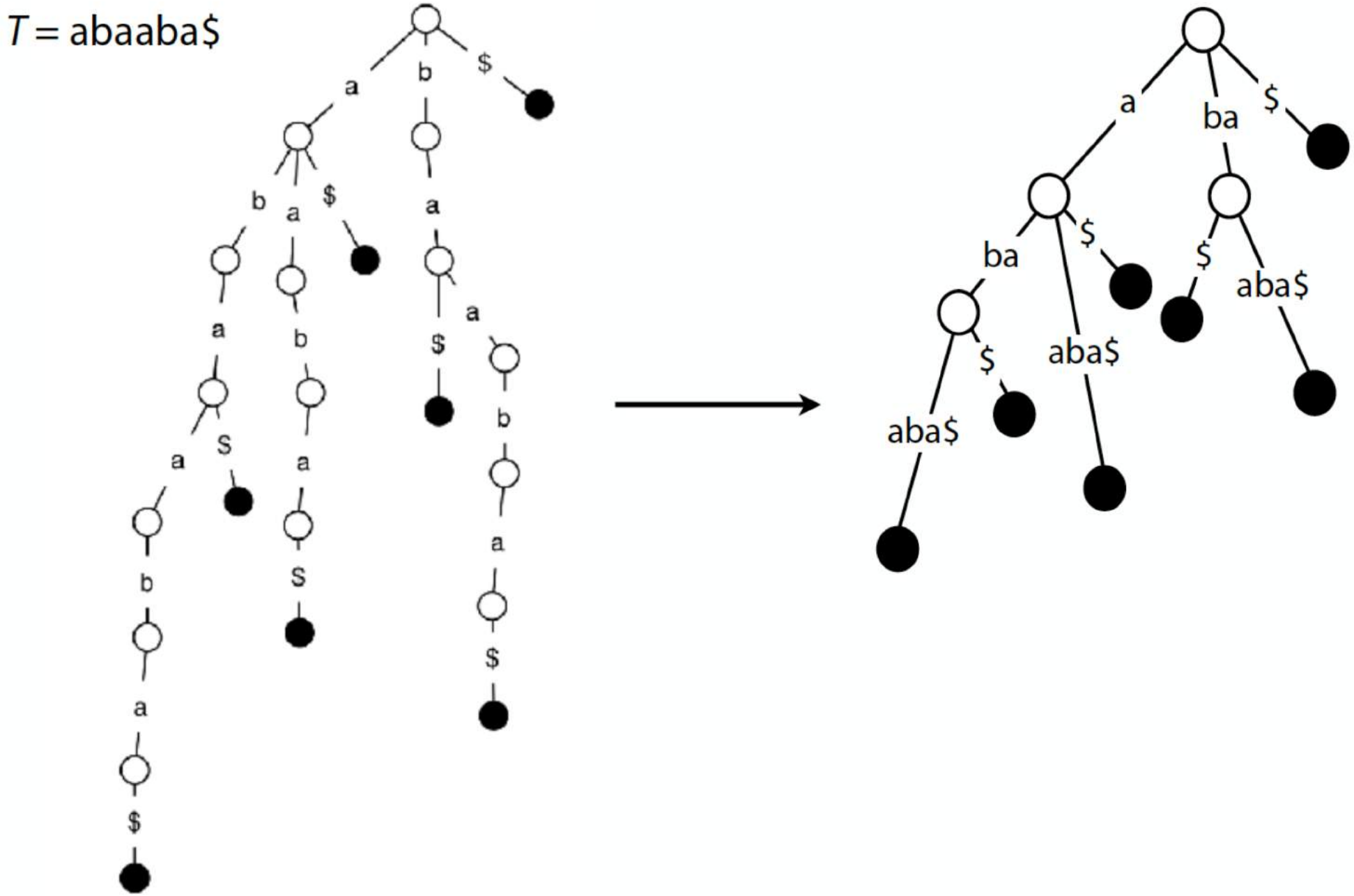


Figure from Ben Langmead

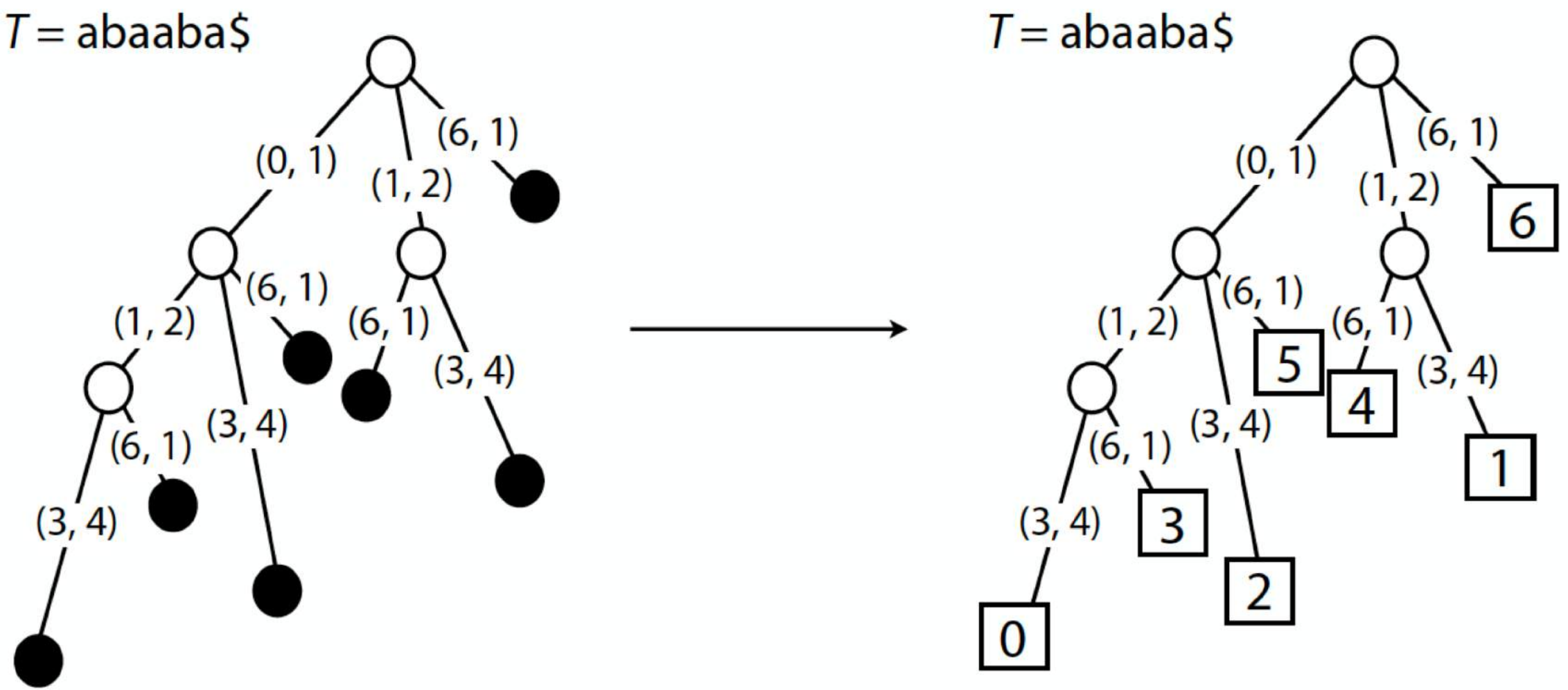


Figure from Ben Langmead

Structure of the suffix tree built from both strings reveals MUMs

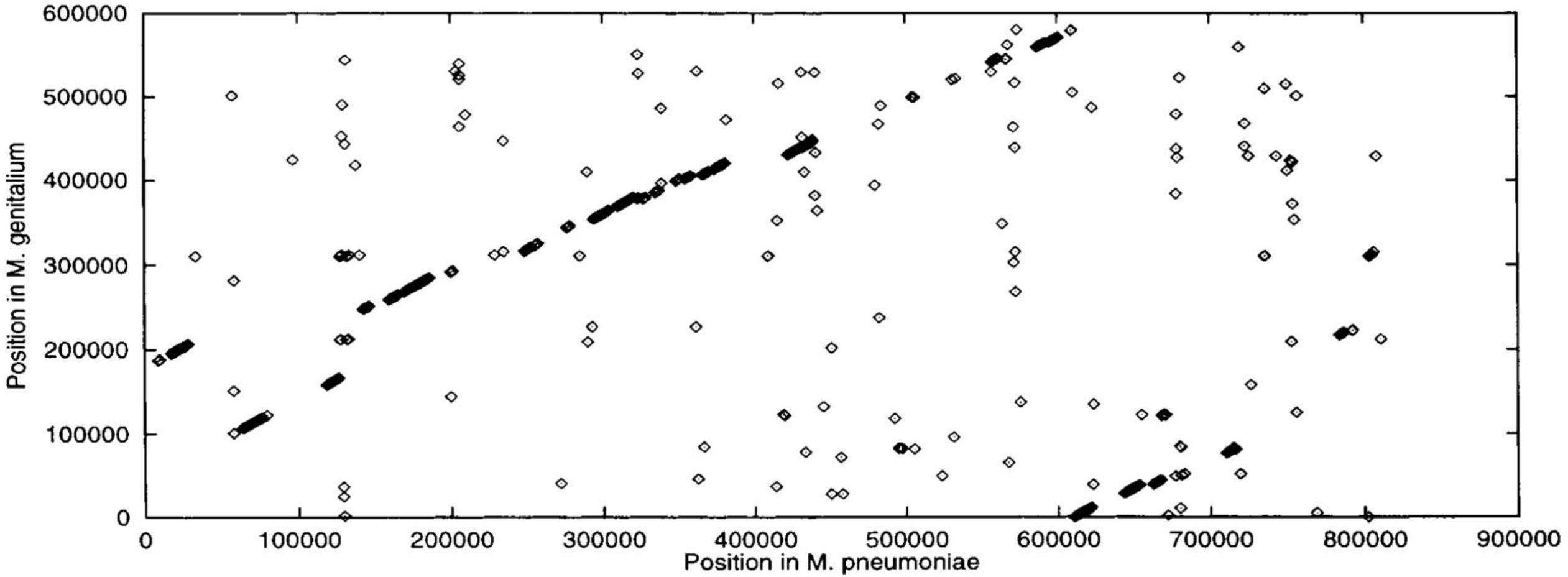
“It is not hard to see that every unique matching sequence is represented by an internal node with exactly two child nodes, such that the child nodes are leaf nodes from different genomes. The unique matches that are maximal can be identified by mismatches at their ends. (MUMmer as actually implemented determines whether a match is maximal based on pointers used to construct the suffix tree.) Thus, in a single scan of the suffix tree, all MUMs can be identified.”

MUMmer1 Applications

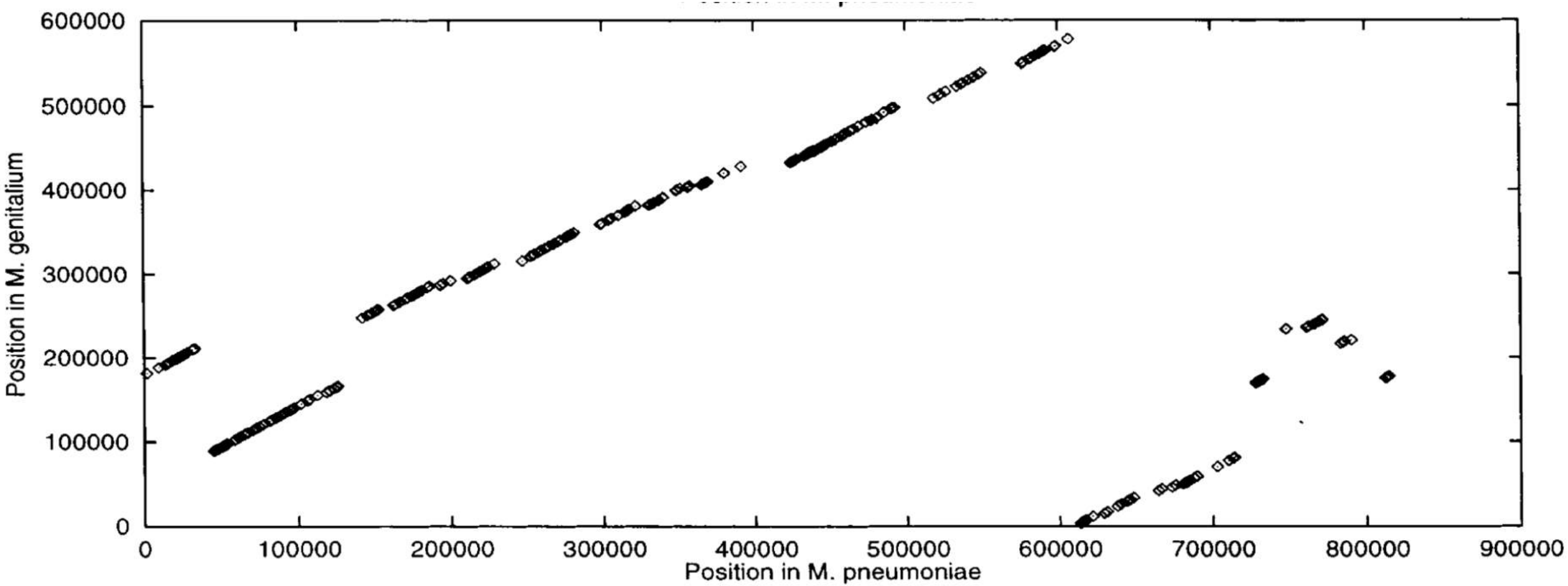
- Two *M. tuberculosis* strains (55 sec)
  - SNPs
  - Insertions
- *M. genitalium* and *M. pneumoniae* (122.5 sec)
  - 20-30bp MUMs “strung together”
- 200 kbp syntenic region of human and mouse (29 sec)
  - 15bp MUM

966 Citations (Google Scholar)

1000bp chunk all-versus-all alignments with the FASTA program



MUMs between *M. genitalium* and *M. pneumoniae*



Sequences are getting larger: need an upgrade

- *Arabidopsis thaliana* chromosomes
- Many pairs of prokaryotic genomes
- Human genome assembly – want to align different versions of the assemblies of human chromosomes ... Too memory-intensive

MUMmer2 Advances

- Proteins
- Build suffix tree for A and ”streaming” search the suffix tree for B
  - Matches will be unique in A but not necessarily in B
- Maximal exact matches (not necessarily unique in A and B)
- Report multiple alignments by clustering small alignments instead of reporting just one large alignment
- One-third the memory and time

MUMmer2 Components

- NUCmer: align contigs to reference (or other set of contigs)
  - Order and orient based on closely related reference
  - Compare versions of assembly
- PROmer: comparative genomics
  - Similarity between all pairs of protein coding regions in 2 sets

- Both result in chained local exact matches -> inexact matches

815 Citations (Google Scholar)

MUMmer2 Applications

- Bacteria undergo “chromosome-scale inversions” around the origin of replication
  - More bacteria genomes are being sequenced – can compare all pairs
- Recent whole genome duplication in *Arabidopsis thaliana*
  - Can do chromosome-scale alignments 17-29Mbp
- Human genome ancient duplications: “over 70% of human chromosome 14 appears to be an ancient duplication of part of chromosome 2”
  - Proteins show this while DNA does not

Delcher, Salzberg and Phillippy. *Current Protocols in Bioinformatics* 2003

Another upgrade: MUMmer3 Advances

- Open-source
- GUI features
- Better suffix tree - ¼ less memory
- “well-structured and well-documented software library”
- Less “hacky” multi-contig query/reference alignment
- NUCmer and PROmer as fast as MUMmer

Time and space improvements enable human genome scale analysis

Time and space requirements of MUMmer 2.1 and 3.0 when computing all exact matches of length 20 or longer for different pairs of sequences

Reference genome		Query genome		MUMmer 2.1		MUMmer 3.0	
Species	Size (Mbp)	Species	Size (Mbp)	Time (sec)	Space (MB)	Time (sec)	Space (MB)
E. coli K12	4.6	*E. coli* O157:H7	5.5	18	102	17	77
A. fumigatus	28.0	*A. nidulans*	30.1	128	578	120	459
Saccharomyces cerevisiae	13.0	*Schizosaccharomyces pombe*	13.8	51	261	47	204
D. melanogaster (chromosome 2L)	22.2	*D. pseudoobscura* (all chromosomes)	150.0	546	684	520	485
Homo sapiens (chromosome 21)	44.7	*Mus musculus* (chromosome 16)	99.2	-	-	430	759

Timings were done on a Linux-based computer with a 2.4 GHz Pentium processor. The human-mouse comparison was run only with MUMmer 3.0. Mbp, millions of base pairs; MB, megabytes. A suffix tree is constructed only for the reference genome.

Pair of human genome assemblies: 101.5 hours, <4GB memory, 11 hours longest chromosome

2756 Citations (Google Scholar)

PROmer reveals synteny in two *Aspergillus*

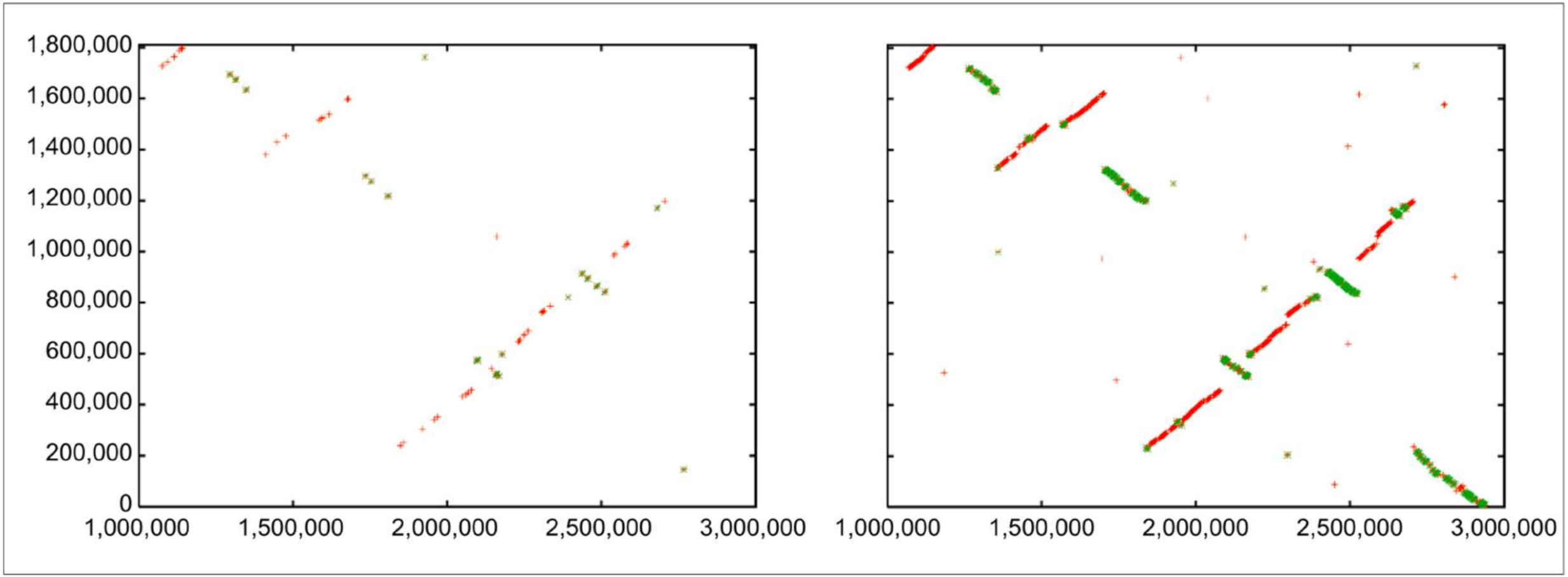
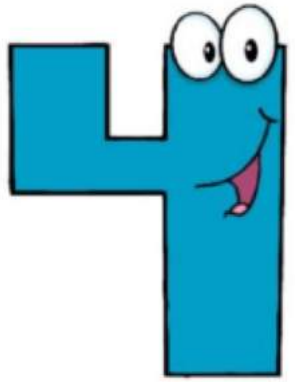


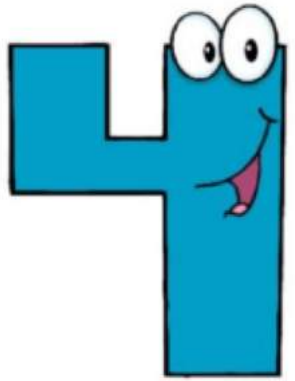
Figure 1
Dot-plot alignments of a 2.9 Mbp chromosome of *A. fumigatus* (*x*-axis) to a 2.1 Mbp scaffold of *A. nidulans* (*y*-axis). Left: nucleotide-based alignment with Nucmer. Right: amino-acid-based alignment with Promer. Aligned segments are represented as dots or lines, up to 3,000 bp long in the Nucmer alignment and up to 9,500 bp in the Promer alignment. These alignments were generated by the mummerplot script and the Unix program gnuplot.

MUMmer4 Advances



- 48-bit essaMEM suffix array (141 Tbp length limit instead of 500 Mb/4 Gb)
- SAM file output
- Parallelism
- Library
- Precompute and save SA

MUMmer4 Applications



		Arabidopsis	Tardigrade	Human/Chimp
nucmer3	Wall time (min)	17.5	19.6	fail
	CPU time (min)	17.1	19.2	fail
	Memory (GB)	2.1	2.3	fail
nucmer4	Wall time (min)	3.7	4.0	207
	CPU time (min)	22	26	2897
	Memory (GB)	4.6	4.9	66

Align reads to a reference
Less sensitive, faster except for bowtie2, more memory with respect to short reads

Using MUMmer as a short-read aligner: MUMmerGPU

BMC Bioinformatics

Home About Articles Submission Guidelines

Software | Open Access

High-throughput sequence alignment using Graphics Processing Units

Michael C Schatz †, Cole Trapnell †, Arthur L Delcher and Amitabh Varshney

†Contributed equally

BMC Bioinformatics 2007 **8**:474

https://doi.org/10.1186/1471-2105-8-474

Received: 20 August 2007 | **Accepted:** 10 December 2007 | **Published:** 10 December 2007

See also MUMmerGPU 2.0: Trapnell and Schatz, *Parallel Computing* 2009

Core data structure became more space- efficient

- Suffix tree bytes per base pair ~= memory requirement
- 1: 37
- 2: 20
- 3: 15.5
- 4: 13-14 (suffix array) can be batched

Two strains of *Y. pestis* have high nucleotide identity but substantial genome rearrangements

- KIM vs. CO92
- This species has a highly repetitive genome already

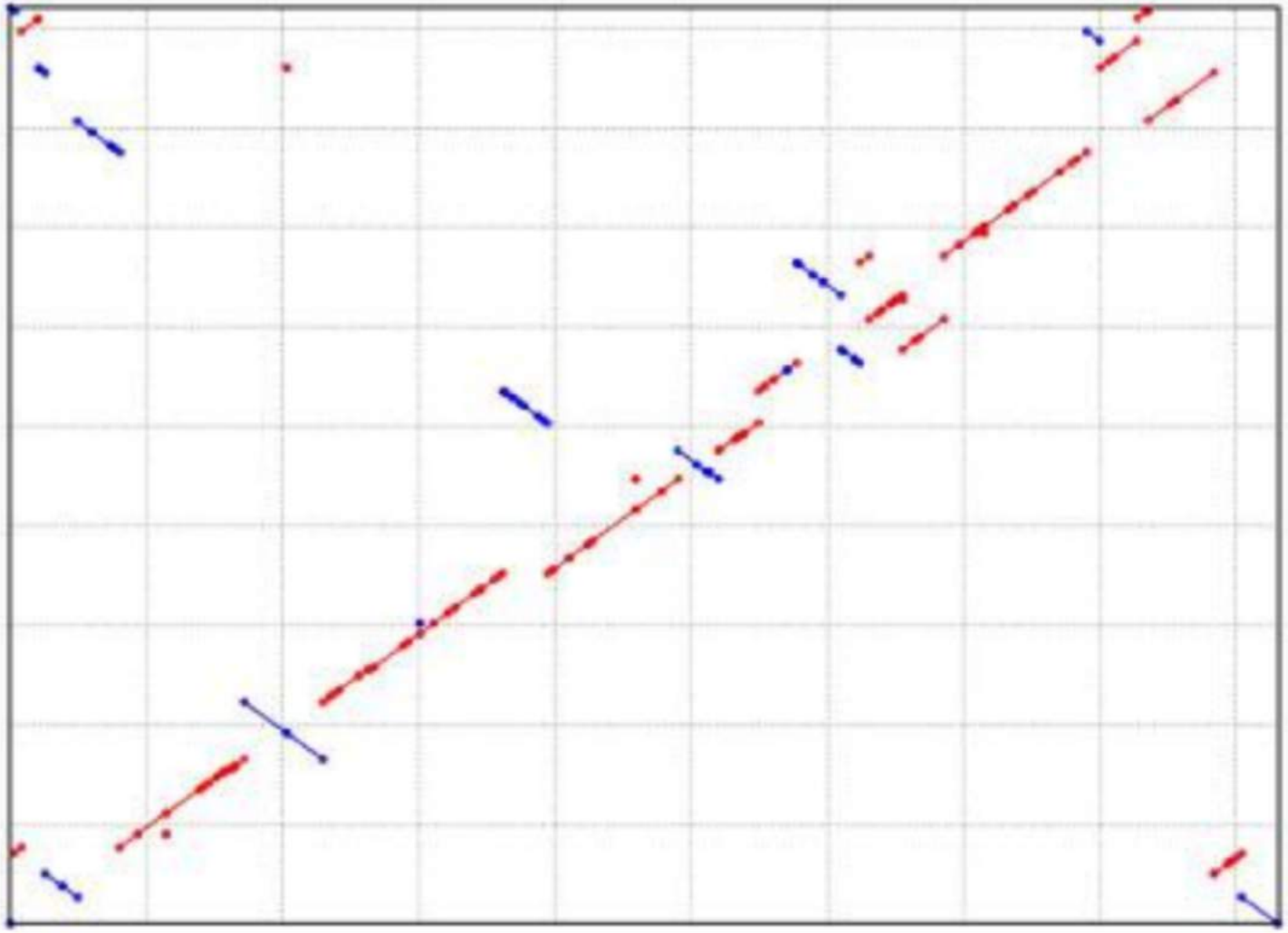
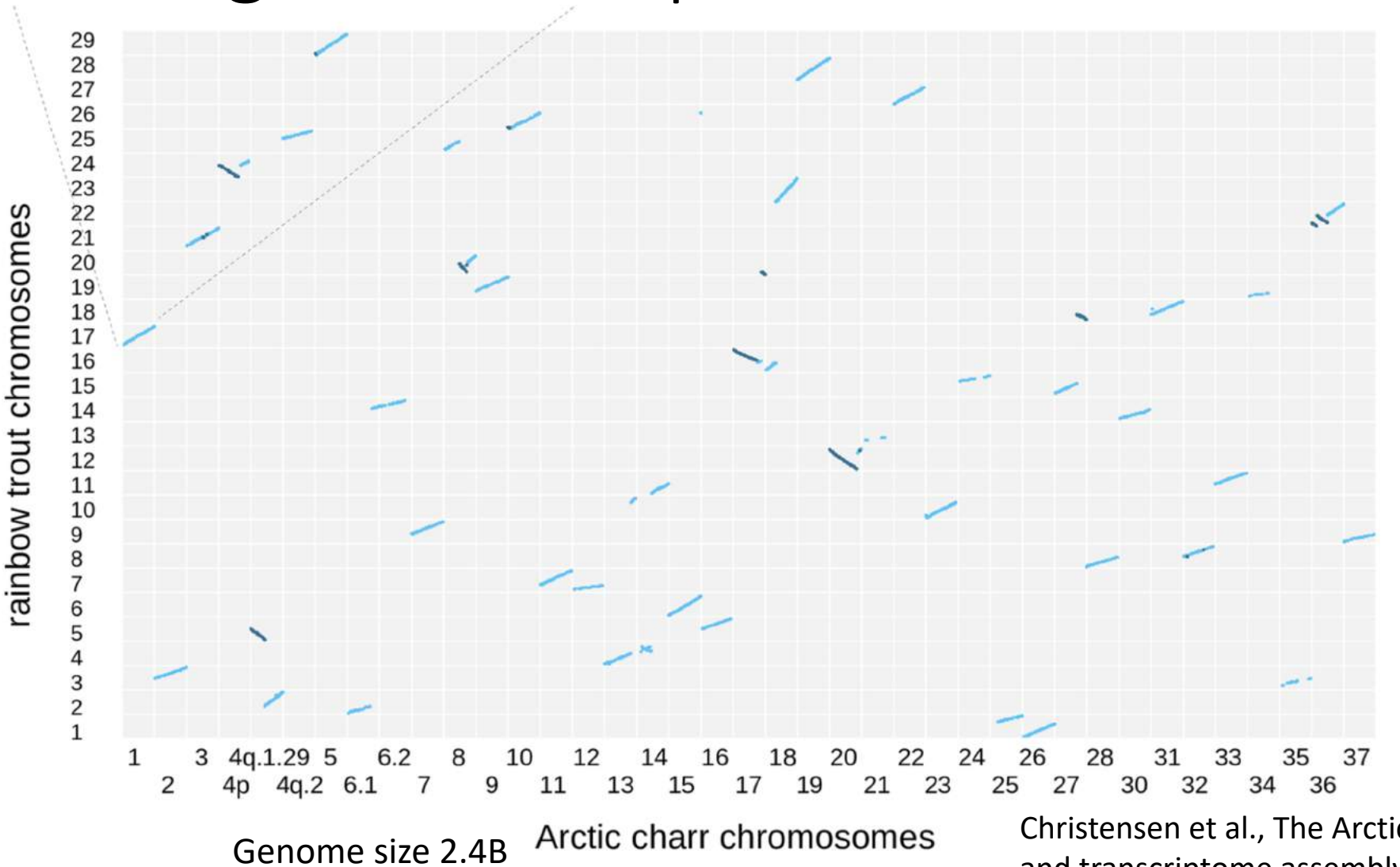


Image from Mike Schatz

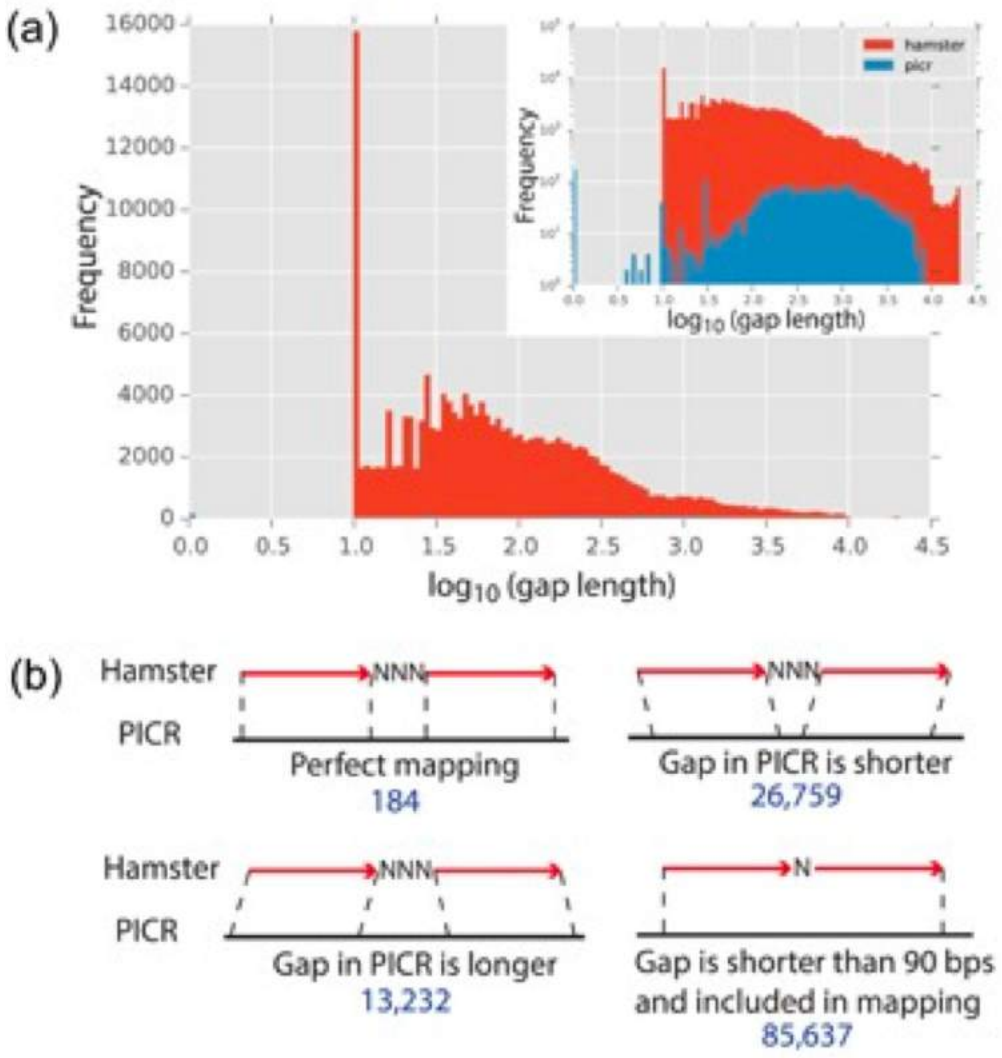
Arctic charr genome: alignments of contigs to the trout genome helped order and orient the contigs



Genome size 2.4B

Christensen et al., The Arctic charr (*Salvelinus alpinus*) genome and transcriptome assembly, PLOS One 2018

Chinese hamster genome: filling gaps in the old assembly using WGA to the new assembly



Rupp et al., A reference genome of the Chinese hamster based on a hybrid assembly strategy, Biotechnol Bioeng 2018

"By aligning the RefSeq assembly to PICR using MUMmer3.0 we identified the missing sequence for 125,812 (76%) of the RefSeq gaps (Figure 3b)"

https://www.agefotostock.com/age/en/Stock-Images/Rights-Managed/RDC-ad-208816

Salmonella enteria serovar classification using SNP distance calculated from NUCmer alignment

- Calculate genetic distance and determine a cutoff for “IN” versus “OUT” for various serovars
- One method: number of SNPs called with NUCmer
- Other methods: MLST, k-mer based distances (Jaccard, MASH)

Pettengill et al., Real-Time Pathogen Detection in the Era of Whole-Genome Sequencing and Big Data: Comparison of k-mer and Site-Based Methods for Inferring the Genetic Distances among Tens of Thousands of *Salmonella* Samples, PLOS One, 2016