CoreDetector: A flexible and efficient program for core-genome alignment of evolutionary diverse genomes

Mario Fruzangohar[1*], Paula Moolhuijzen[2], Nicolette Bakaj[1], Julian Taylor[1]

[1]The Biometry Hub, School of Agriculture, Food and Wine, University of Adelaide, Australia
[2]Centre for Crop and Disease Management, School of Molecular and Life Sciences, Curtin University, Bentley, WA 6102, Australia

[*]Corresponding author

Table of Contents

Supplementary methods Notes

Extracting query segments from pair-wise MAF entries

Post LCB identification, for each entry in the pair-wise MAF file, one FASTA entry contig is generated as below: sequence of contig is simply the query sequence of the MAF entry where gaps (character "-") are removed and the contig name is constructed as below:

$$query\_contig\_name! \, query\_start\_pos! \, query\_end\_pos$$

Start and end positions of the query are added to the query contig name separated by the '!' character. These indices help the final backtracking algorithm. The generated FASTA file is used to align against the next subject genome FASTA file

CoreDetector Backtracking algorithm

All *N - 1* generated pair-wise MAF files are processed recursively (starting from the last MAF file through to the first one). For each HSP entry of the last MAF file, a backtracking algorithm extracts each subject sequence part and finally builds an MSA entry containing *N* sequences. MSA entries are written as a final output MAF file or all concatenated into a final output FASTA file containing *N* contigs.

Given *N* genomes results in *N - 1* pairwise MAF file alignments, each pairwise alignment is accessed through an array defined as *MAF[1 .. N - 1]*. Foreach pairwise MAF alignment, let msa^K be a variable list of *K* sequences that represent the multiple sequence alignment profile of the

K sequences, and let the function *getHomolog (hsp, query, maf)* query part of the input *hsp* in the input pairwise alignment maf file to return the subject *hsp* equivalent.

$For\ each\ entry\ of\ MAF[N-1]\ as\ 'hsp'\ \ begin$

$Append\ query\ sequence\ of\ 'hsp'\ to\ msaN[1];$

$For\ i\ \ from\ N-1\ to\ 1\ begin$

$Append\ subject\ sequence\ of\ getHomolog('hsp'.query, MAF[i])\ to\ msaN[N+1-i];$

$Adjust\ alignment\ gaps\ in\ msaN;$

$Trim\ last\ 2\ numbers\ separated\ by\ !\ \ from\ 'hsp'.query\_name$

$End\ For;$

$End\ For;$

The adjustment of alignment gaps occurs using an algorithm that consists of two processes. First process is *gap surveyor* that essentially prepares a map of gap positions and their numbers. The second process is *gap merger* that merges the gap survey of 2 sequences. The algorithm starts with preparing a gap survey for the query sequence as a Java Map object. When a subject profile is added to the end of alignment profile of msa^K, the gaps survey is merged with each previous sequence's gap survey profile in msa^K and as a result of merging, all sequences in the profile and their gap surveys are updated.

Complexity of Backtracking algorithm

Order of the algorithm is a function of number of genomes (N), level of divergence between genomes (D) and length of genomes (L). The first loop in the algorithm has an order equal to number of HSPs in the final pair-wise alignment, that itself is a function of N, L and D and can be shown with $h(N, L, D)$. Then the order of the algorithm is $O(h(h + N^2))$.

Selecting homologous sequences

The pairwise aligner is run in low and high sensitivity modes and the results of the two runs are merged into a single MAF file. In low sensitivity mode, more similar sequences are captured and in high sensitivity mode more divergent sequences are detected. Adjacent hits from the two runs are then merged into larger hits.

As pairwise alignments can report multiple hits (HSPs) due to paralogue sequences (duplication) or by chance, CoreDetector can use the chromosome number (if available) and relative position of an HSP in that chromosome to select the best hit that likely reflects the true orthologs. HSPs shorter than a user selected threshold (default 500bp) is excluded from the results to minimize the chance of a false positive orthologs.

Experiment 1) Analysis of 27 fungal pathogen (Ptr) genomes

Table S1.*Pyrenophora tritici-repentis* (Ptr) isolate genomes

ASSEMBLY ACCESSION	STRAIN	SEQUENCING TECHONOLGY	LENGTH	ASSEMBLY LEVEL
GCA_003171515.3	M4	PacBio	40738834	Chromosome
GCA_003171545.1	ARCrossB10v1	Illumina	33619417	Scaffold
GCA_003231325.2	Ptr134	Illumina	40622625	Chromosome
GCA_003231345.1	Ptr5213	Illumina	34183844	Scaffold
GCA_003231355.1	Ptr11137	Illumina	33921061	Scaffold
GCA_003231365.1	Ptr239	Illumina	34483177	Scaffold
GCA_003231415.2	DW5	PacBio	40616955	Chromosome
GCA_003231425.2	86-124	PacBio	40927694	Chromosome
GCA_008692205.1	V0001	PacBio	40136298	Contig
GCA_018492725.1	AR CrossB10v2	PacBio	39876747	Contig
GCA_022544795.1	Alg130	Illumina	34824145	Scaffold

GCA_022544805.1	Alg215	Illumina	34570617	Scaffold
GCA_022578365.1	T205	Illumina	34133756	Scaffold
GCA_022578395.1	T199	Illumina	34281657	Scaffold
GCA_022788405.1	EW4-4	Illumina	34368509	Scaffold
GCA_022788415.1	SN001A	Illumina	34152043	Scaffold
GCA_022788425.1	SN002B	Illumina	35159469	Scaffold
GCA_022788435.1	EW7m1	Illumina	34228317	Scaffold
GCA_022788445.1	SN001C	Illumina	34292940	Scaffold
GCA_022788505.1	EW306-2-1	Illumina	34540381	Scaffold
GCA_022788515.1	CC142	Illumina	34345883	Scaffold
GCA_022813025.1	Biotrigo9-1	PacBio	42003943	Chromosome
GCA_022813065.1	L13-192	PacBio	36964309	Chromosome
GCA_022837075.1	Ptr90-2	Pacbio	39319516	Chromosome
GCA_000149985.1	Pt-1C-BFP	Sanger	37840464	Scaffold
	*DW7	Illumina		Scaffold
	*SD20	Illumina		Scaffold

*Assembly originally sourced from DOI: 10.1186/s12864-018-4680-3 published in 2018.

CoreDetector command:

$$/usr/bin/time\ -v\ "pipeline\_Minimap.sh\ ptr\_genomes.txt\ coredector\_ptr27\ 10\ 16"$$

Mugsy command:

$$/usr/bin/time\ -v\ mugsy\ --directory\ /data/mugsy\ -prefix\ ptr\ /data/*.fna\ /data/*.genome.fa$$

Parsnp command:

$$/usr/bin/time\ -v\ parsnp\ -r\ !\ -d\ ptr\ -p\ 16\ -P\ 61000000000\ --verbose\ \&>\ ptr.log$$

Phylonium command:

$$/usr/bin/time\ -v\ phylonium\ -b\ 100\ -t\ 16\ -v\ ptr/*\ \&>\ phylonium\_ptr.txt$$

Table S2. The comparative matrix of fungal pathogen rooted phylogenetic tree metrics generated from CoreDetector, Parsnp and Phylonium results.

Distance Metric	Triplets	RFCluster0.5	Matching Pair	Nodal Splitted	Matching Cluster	MAST	CopheneticL2

CoreDetector vs Parsnp	448	**9**	94	44.5	52	**7**	37.61
CoreDetector vs Phylonium	667	13	109	55.24	68	10	42.94
Parsnp vs Phylonium	**370**	11	**64**	**36.08**	**38**	9	**20.02**

Experiment 2) Analysis of 12 Drosophila genomes

Table S3 Drosophila genus genomes

Assembly Accession	***Organism Name***	***Length***	***Assembly Level***
GCA_000001215.4	Drosophila melanogaster	143706478	Chromosome
GCA_003285735.2	Drosophila virilis	189443829	Contig
GCA_003286085.2	Drosophila persimilis	195512972	Contig
GCA_003286155.2	Drosophila erecta	146538397	Contig
GCA_004382195.2	Drosophila sechellia	153084571	Chromosome
GCA_009870125.2	Drosophila pseudoobscura	163266851	Chromosome
GCA_016746365.2	Drosophila yakuba	147883098	Chromosome
GCA_016746395.2	Drosophila simulans	131663590	Chromosome
GCA_017639315.2	Drosophila ananassae	213817545	Chromosome
GCA_018153295.1	Drosophila grimshawi	191382978	Contig
GCA_018153725.1	Drosophila mojavensis	163170721	Contig
GCA_018902025.2	Drosophila willistoni	246985538	Chromosome

CoreDetector command:

$$./pipeline\_Minimap.sh\ \ /path/to/genomes.txt\ \ /output/folder\ \ 40\ \ 32$$

Parsnp command:

$$time\ ./parsnp\ \ -r\ !\ \ -d\ /dir/To/fasta/files\ \ -p\ 32\ \ -c\ \ -P\ 1500000000\ \ -o\ /output/dir$$

Phylonium commands:

1. Run Phylonium

$$time\ ./phylonium\ mel.fasta\ \ sec.fasta\ sim.fasta\ yak.fasta\ ere.fasta \setminus ana.fasta\ pse.fasta\ per.fasta\ wil.fasta\ vir.fasta\ moj.fasta\ gri.fasta > phylip.mat$$

2. Install mattools from https://github.com/EvolBioInf/mattools

$$./mat\ nj\ phylip.mat$$

ProgressiveCactus commands:

1. Make sure you update TOIL directory to a directory that has enough space:

$$export\ TOIL\_WORKDIR = /path/to/large/space$$

$$time\ ./cactus\ /Path/to/cactus\_output\ /path/to/genomes.txt\ \ flymsa.hal$$

2. Run halStats to generate phylogenetics tree from hal file:

$$halStats\ \ flymsa.hal$$

3. Convert hal to maf format:

$$./hal2maf\ \ flymsa.hal\ cactusflymsa.maf$$

4. Analysis of maf file and the extraction of the core sequence:

$$java\ -jar\ MFbio.jar\ --task\ analyzecactus\ --srcdir\ /cactusflymsa.maf\ --destdir\ /cactus\_out/alignmen\_core.maf\ --file1\ \ /cactus\_out/alignment\_Core.fa$$

SibeliaZ commands:

Using 32 cores and 256GB of RAM

1. Run Sibeliaz

$$time\ -v\ sibeliaz\ -t\ 32\ \ mel.fasta\ \ sec.fasta\ sim.fasta\ yak.fasta\ ere.fasta \setminus ana.fasta\ pse.fasta\ per.fasta\ wil.fasta\ vir.fasta\ moj.fasta\ gri.fasta$$

2. Analysis of maf file and the extraction of the core sequence:

$$java\ -jar\ MFbio.jar\ --task\ analyzesibeliaz\ --srcdir\ /sibeliaz\_out/alignment.maf\ --destdir\ /sibeliaz\_out/alignmen\_core.maf\ --file1\ \ /sibeliaz\_out/alignment\_Core.fa$$

Table S4. The comparative phylogenetic tree distance metrics of fly unrooted ground truth tree compared to Neighbor Joining unrooted phylogenetic trees generated from the results obtained from CoreDetector, Phylonium, ProgressiveCactus and SibeliaZ.

Distance Metric	Quartet	Path Difference	Matching Triplets	Matching Split	UMAST	GeoUnrooted	RF Weighted(0.5)	RF(0.5)
CoreDetector	0.0	0.00	4	0	0	80.99	131.64	0
Phylonium	0.0	0.00	10	0	0	81.26	132.06	0
ProgressiveCactus	0.0	0.00	11	0	0	80.98	131.63	0
SibeliaZ	21.0	5.65	43	4	1	81.17	131.92	1

Experiment 3) Analysis of 34 Rodent genomes

Table S5 Rodent order genomes

ASSEMBLY ACCESSION	ORGANISM NAME	LENGTH	ASSEMBLY LEVEL
GCA_004027535.1	Acomys cahirinus	2306070819	Scaffold
GCA_004027875.1	Aplodontia rufa	3005535537	Scaffold
GCA_001984765.1	Castor canadensis	2518306565	Scaffold
GCA_004027575.1	Cricetomys gambianus	2397721602	Scaffold
GCA_000223135.1	Cricetulus griseus	2399770464	Scaffold
GCA_000151885.2	Dipodomys ordii	2236368823	Scaffold
GCA_004024685.1	Dipodomys stephensi	2346418196	Scaffold
GCA_001685075.1	Ellobius lutescens	2353188398	Scaffold
GCA_001685095.1	Ellobius talpinus	2265966243	Scaffold
GCA_004027185.1	Glis glis	2462087207	Scaffold
GCA_004027655.1	Graphiurus murinus	2815014629	Scaffold
GCA_016881025.1	Ictidomys tridecemlineatus	2478949113	Chromosome
GCA_020740685.1	Jaculus jaculus	2863848715	Chromosome
GCA_001458135.2	Marmota marmota marmota	2506852125	Scaffold
GCA_002204375.1	Meriones unguiculatus	2523107715	Scaffold
GCA_017639785.1	Mesocricetus auratus	2457062007	Scaffold
GCA_000317375.1	Microtus ochrogaster	2287340943	Chromosome

GCA_900094665.2	Mus caroli	2553112587	Chromosome
GCA_000001635.9	Mus musculus	2728206152	Chromosome
GCA_900095145.2	Mus pahari	2475012951	Chromosome
GCA_921997135.2	Mus spretus	2546527799	Chromosome
GCA_004027005.1	Muscardinus avellanarius	2527147110	Scaffold
GCA_000622305.1	Nannospalax galili	3061408210	Scaffold
GCA_004026605.1	Ondatra zibethicus	2562752769	Scaffold
GCA_903995425.1	Onychomys torridus	2468394440	Chromosome
*GCA_004027895.1	Orientallactaga bullata	3093575781	Scaffold
GCA_023159225.1	Perognathus longimembris pacificus	2212099196	Chromosome
GCA_003704035.3	Peromyscus maniculatus bairdii	2512423440	Chromosome
GCA_907164565.1	Psammomys obesus	2364980841	Scaffold
GCA_015227675.2	Rattus norvegicus	2647899415	Chromosome
GCA_004025045.1	Sigmodon hispidus	2730600022	Scaffold
GCA_002406435.1	Spermophilus dauricus	3106271744	Scaffold
GCA_004024805.1	Xerus inauris	2601418404	Scaffold
GCA_004024765.1	Zapus hudsonius	2611189839	Scaffold

- Note Orientallactaga bullata (Gobi jerboa)= Allactaga_bullata

CoreDetector command:

1. Run CoreDetector

$$time\ ./pipeline.sh\ 34genomes\_noindex.txt\ Minimap\_2\ 40\ 32$$

Parsnp (Harvest) command:

Resources: We allocated a machine with 16 OCPU and 256GB RAM.

1. Run Parsnp

$$time\ ./parsnp\ -r\ !\ -d\ /path/to/Rodent/fasta/files\ -p\ 32\ -c\ -P\ 1500000000\ -o\ /path/to/output$$

Mugsy command:

Resources: We allocated a machine with 16 OCPU and 256GB RAM.

2. Run Mugsy

$$time\ ./mugsy\ --directory\ /path/to/output\ --prefix\ Rodentgenomes$$

$Acomys\_cahirinus.fna\ Allactaga\_bullata.fna\ Aplodontia\_rufa.fna$
$\backslash Castor\_canadensis.fna\ Cricetomys\_gambianus.fna\ Cricetulus\_griseus.fna$
$\backslash Dipodomys\_ordii.fna\ Dipodomys\_stephensi.fna\ Ellobius\_lutescens\_v1.fna$
$\backslash Ellobius\_talpinus.fna\ GliGli.fna\ Graphiurus\_murinus.fna$
$\backslash Ictidomys\_tridecemlineatus.fna\ Jaculus\_jaculus.fna\ Marmota\_marmota.fna$
$\backslash Meriones\_unguiculatus.fna\ Mesocricetus\_auratus.fna\ Microtus\_ochrogaster.fna$
$\backslash Mus\_caroli.fna\ Mus\_pahari.fna\ Mus\_spretus.fna\ Muscardinus\_avellanarius\_v1.fna$
$\backslash Nannospalax\_galili.fna\ Ondatra\_zibethicus.fna\ Onychomys\_torridus.fna$
$\backslash Perognathus\_longimembris.fna\ Peromyscus\_maniculatus.fna\ Psammomys\_obesus.fna$
$\backslash Rattus\_norvegicus.fna\ Sigmodon\_hispidus.fna\ Spermophilus\_dauricus.fna$
$\backslash Xerus\_inauris.fna\ Zapus\_hudsonius.fna\ mus\_musculus\_v39.fna$

ProgressiveCactus commands:

Resources: We allocated a machine with 32 OCPU and 512GB RAM.

1. Make sure you update TOIL directory to a directory that has enough space:
   $export\ TOIL\_WORKDIR = /mnt/NFS/mario/temp/$

2. Run ProgressiveCactus
   $time\ cactus\ /path/to/output\ \ /path/to/34genomes.txt\ /rodent\_msa.hal$

Run Skmer command:
Using 32 Cores and 128GB of RAM

$$time\ skmer\ reference\ /references/Rodent\ -l\ /skmer\_out\ -t\ -o\ JCtree\ -p\ 32$$

Table S6. Comparative phylogenetic tree distance metrics of unrooted rodent ground truth tree compared to unrooted trees generated from CoreDetector, Phylonium and ProgressiveCactus and SibeliaZ results.

Distance Metric	Quartet	Path Difference	Matching Triplets	Matching Split	Umast	GeoUnrooted	RF Weighted(0.5)	RF(0.5)
CoreDetector	672	18.43	471	12	2	150.17	456.16	3
Skmer	2240	26.38	648	20	4	150.15	456.03	5