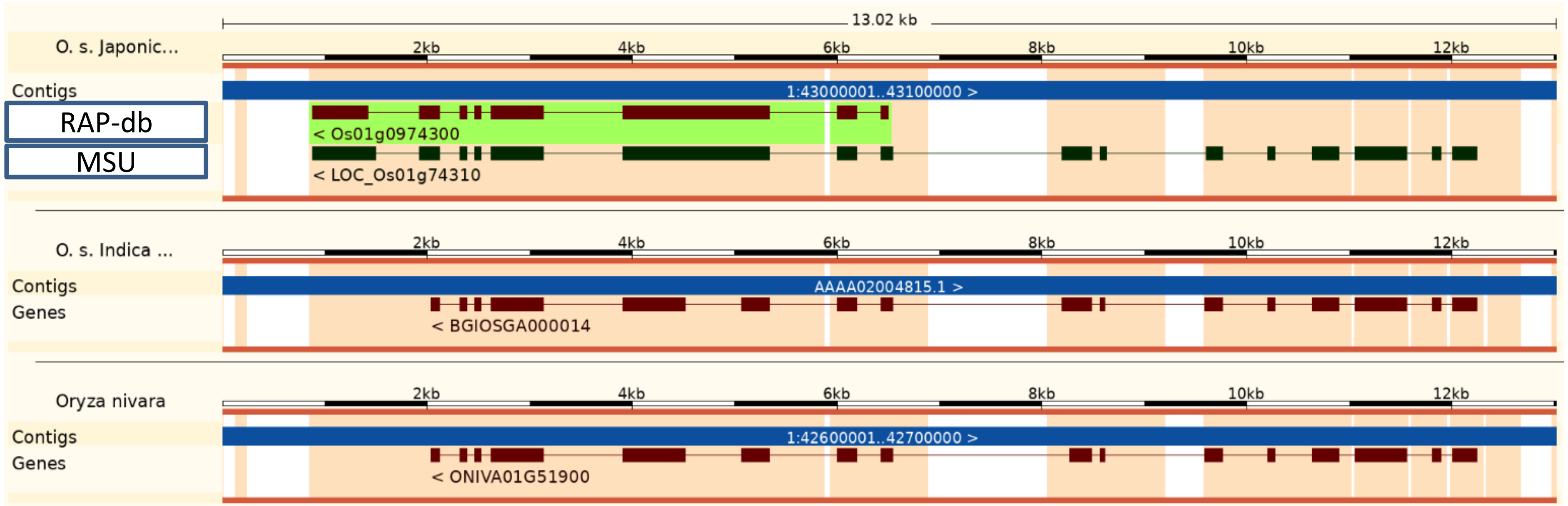
Building Pangene Sets from Plant Genome Alignments Confirms Presence-Absence Variation

Bruno Contreras-Moreira, Shradha Saraf, Guy Naamati,
Sandeep S. Amberkar, Paul Flicek, Andrew R. Jones,
Sarah Dyer

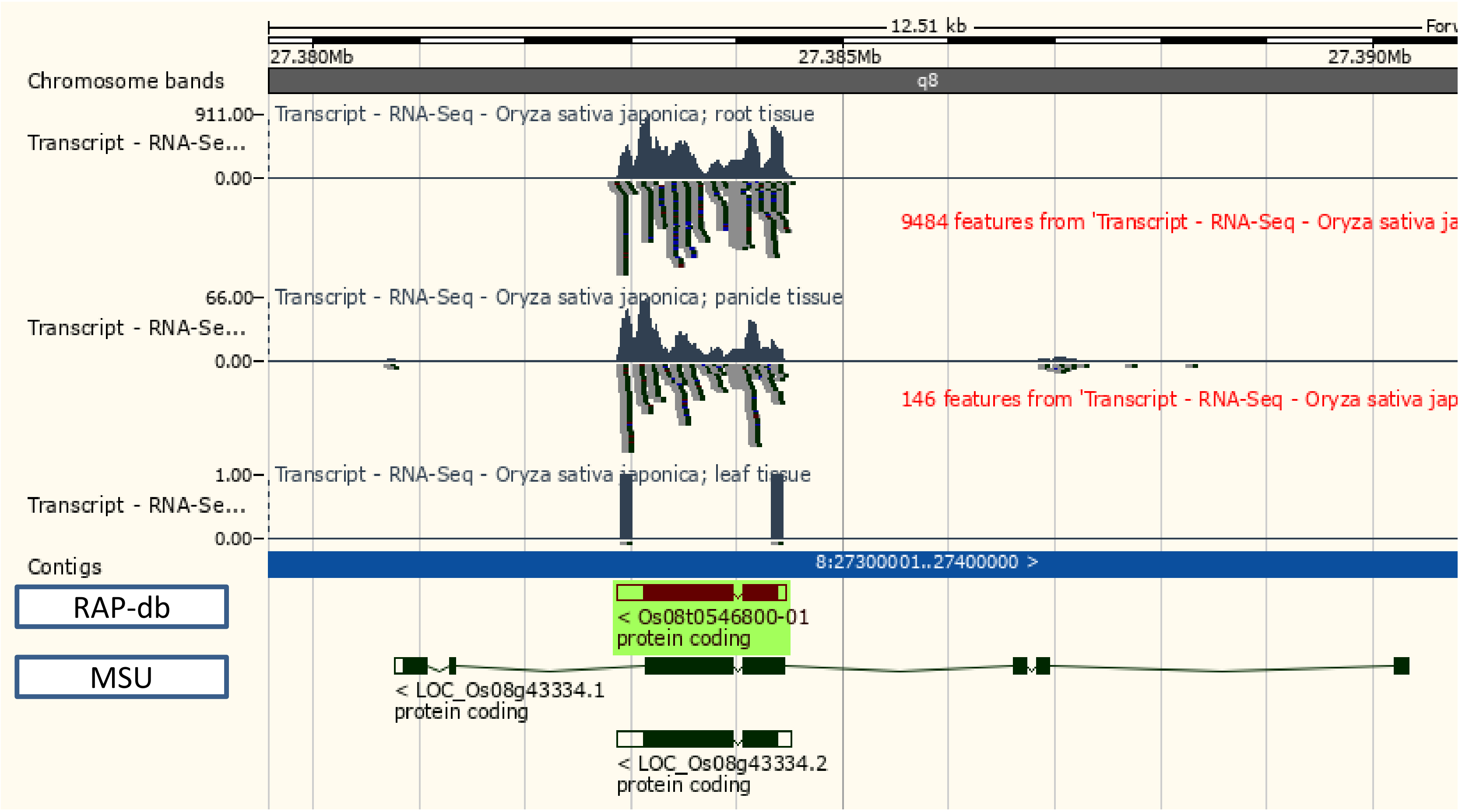
European Molecular Biology Laboratory-EBI
Estación Experimental de Aula Dei-CSIC

Problem: unconsistent gene annotation

	Gene models MSU (2011)	Gene models RAP-db	Protein products SwissProt
Oryza sativa	55.801	37.859	4.168



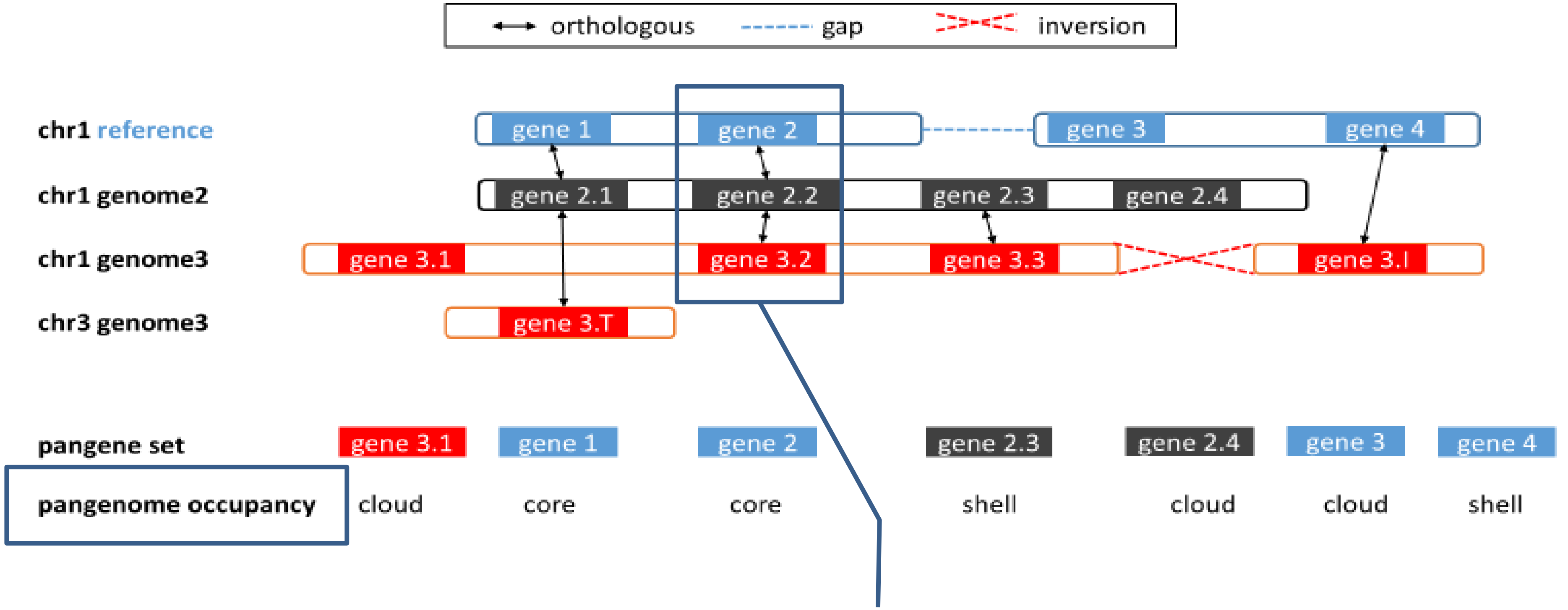
Problem: multiple isoforms, are they all valid and relevant?



9484 features from 'Transcript - RNA-Seq - Oryza sativa ja

146 features from 'Transcript - RNA-Seq - Oryza sativa jap

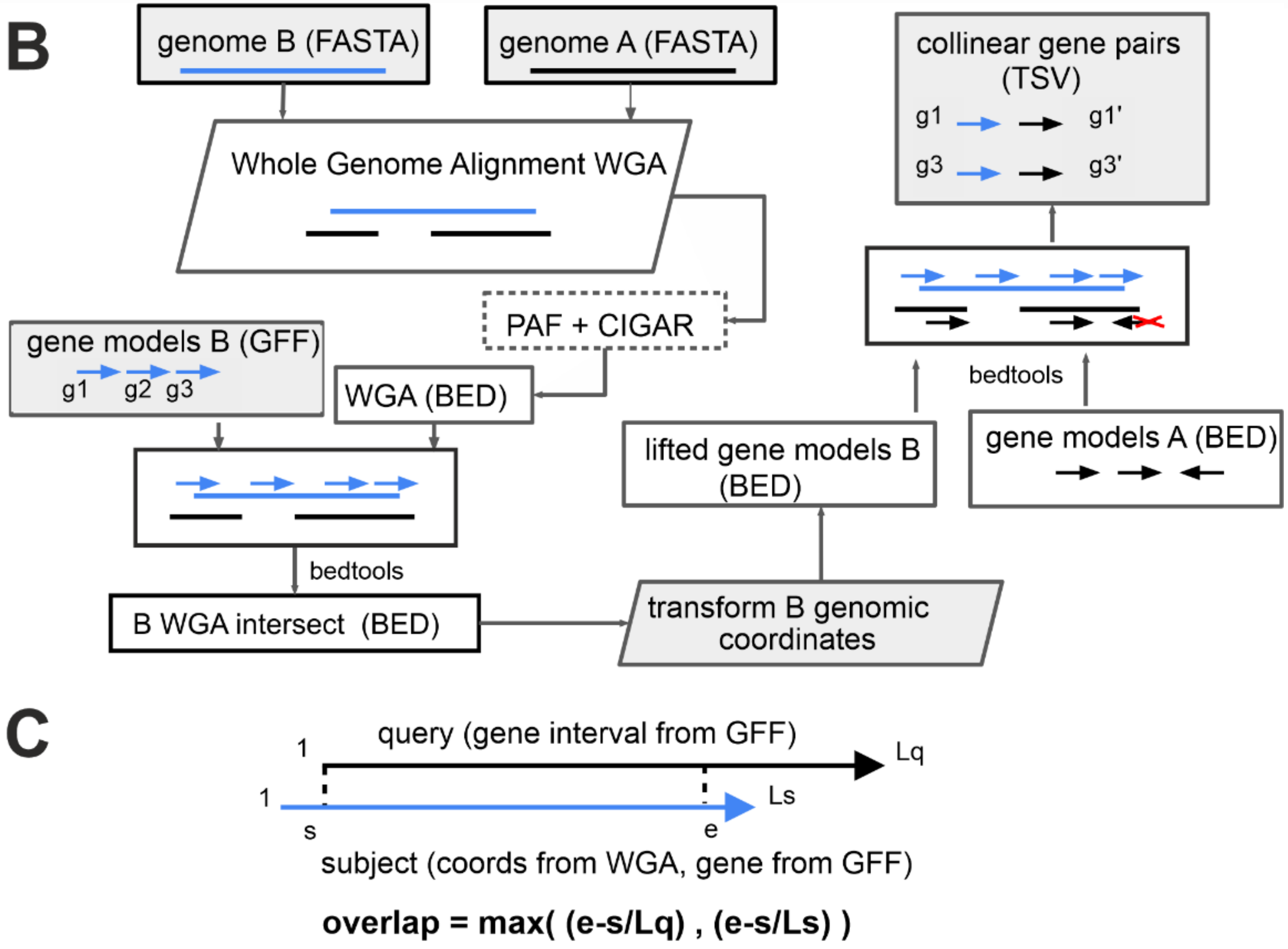
Goal: definition and nomenclature of pangenes



pangene = cluster of overlapping genomic
regions containing gene models

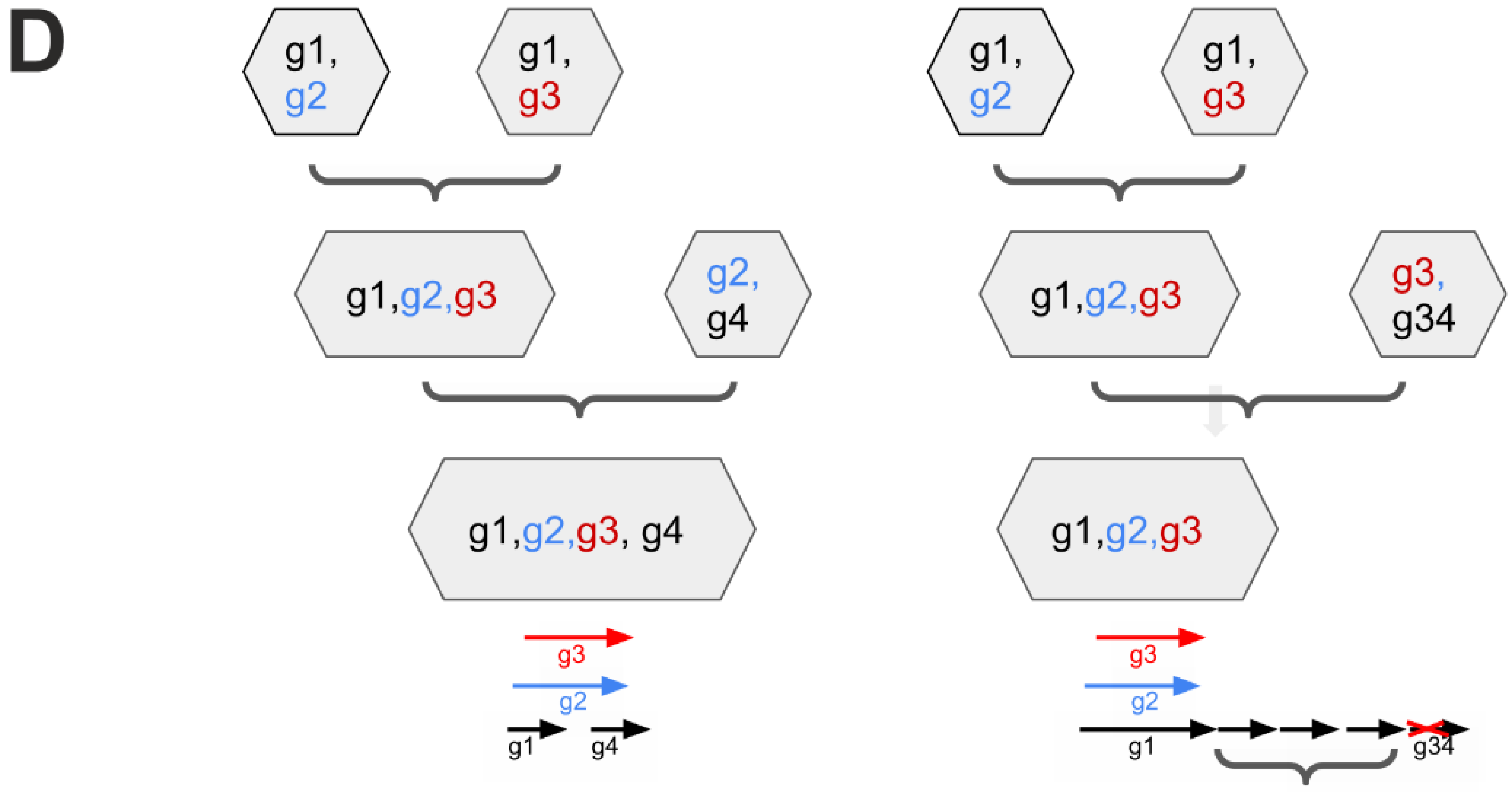
https://github.com/Ensembl/plant-scripts

A prototype to produce clusters of pangenes


B
genome B (FASTA)
genome A (FASTA)
Whole Genome Alignment WGA
PAF + CIGAR
gene models B (GFF)
g1
g2
g3
WGA (BED)
bedtools
B WGA intersect (BED)
transform B genomic coordinates
lifted gene models B (BED)
collinear gene pairs (TSV)
g1
g1'
g3
g3'
bedtools
gene models A (BED)
C
query (gene interval from GFF)
1
Lq
1
Ls
s
e
subject (coords from WGA, gene from GFF)
overlap = max((e-s/Lq) , (e-s/Ls))

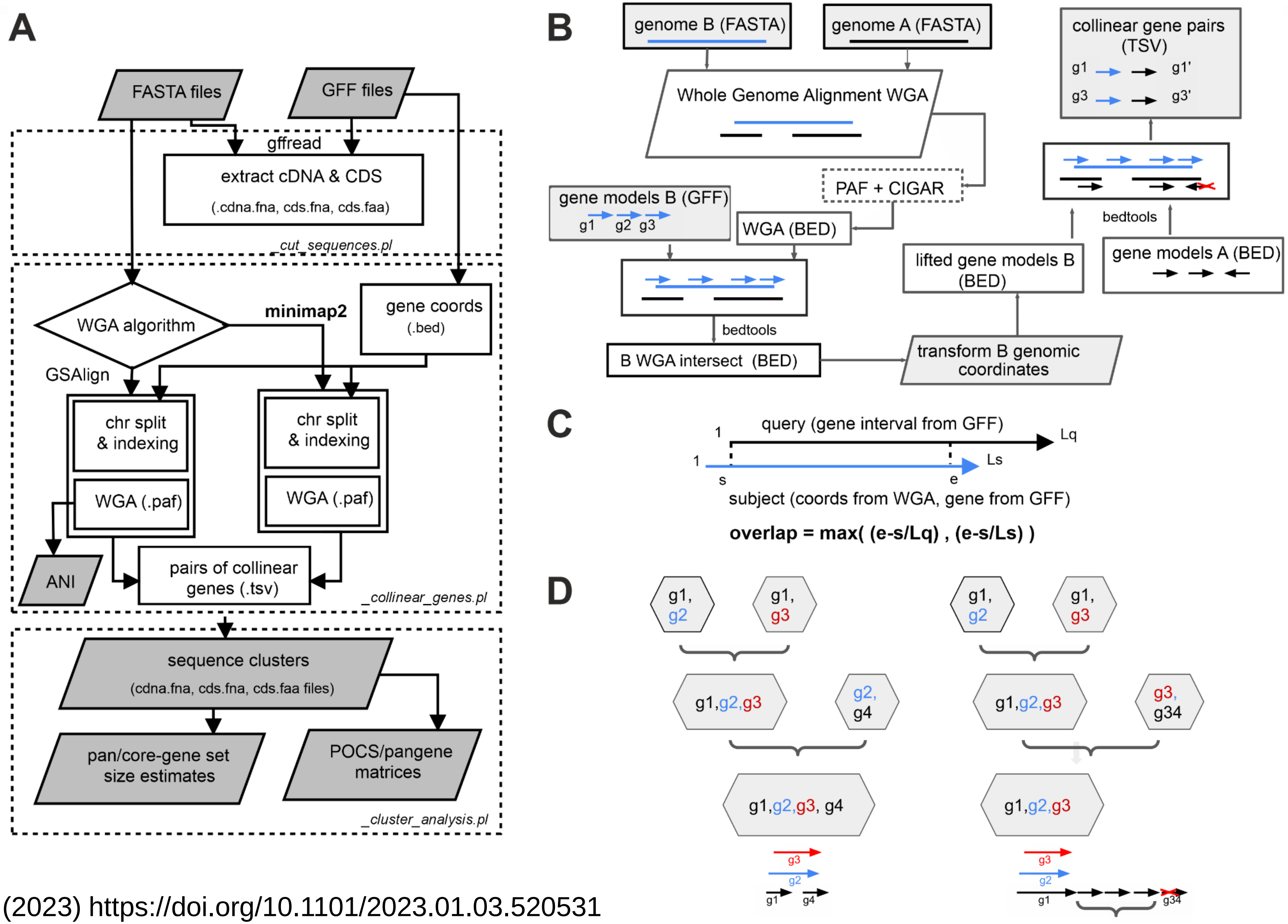
Contreras-Moreira et al (2023) https://doi.org/10.1101/2023.01.03.520531

A prototype to produce clusters of pangenes



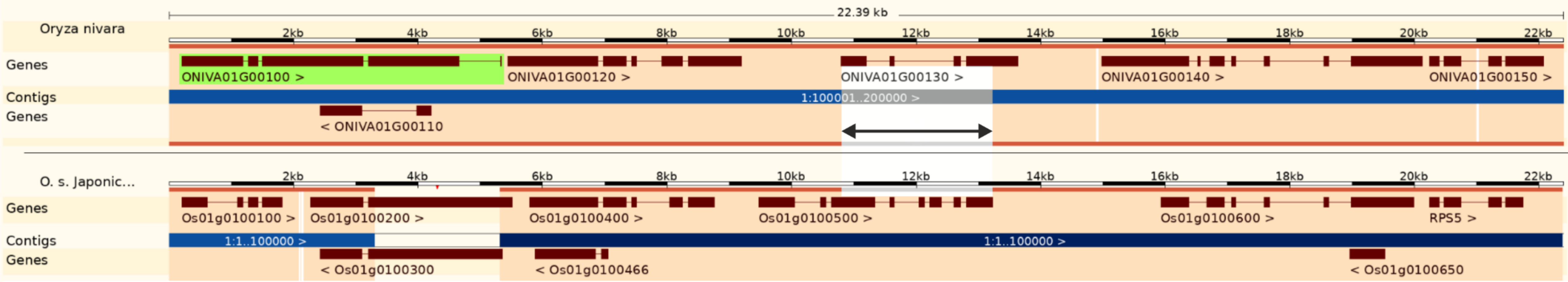
Contreras-Moreira et al (2023) https://doi.org/10.1101/2023.01.03.520531

A prototype to produce clusters of pangenes



Contreras-Moreira et al (2023) https://doi.org/10.1101/2023.01.03.520531

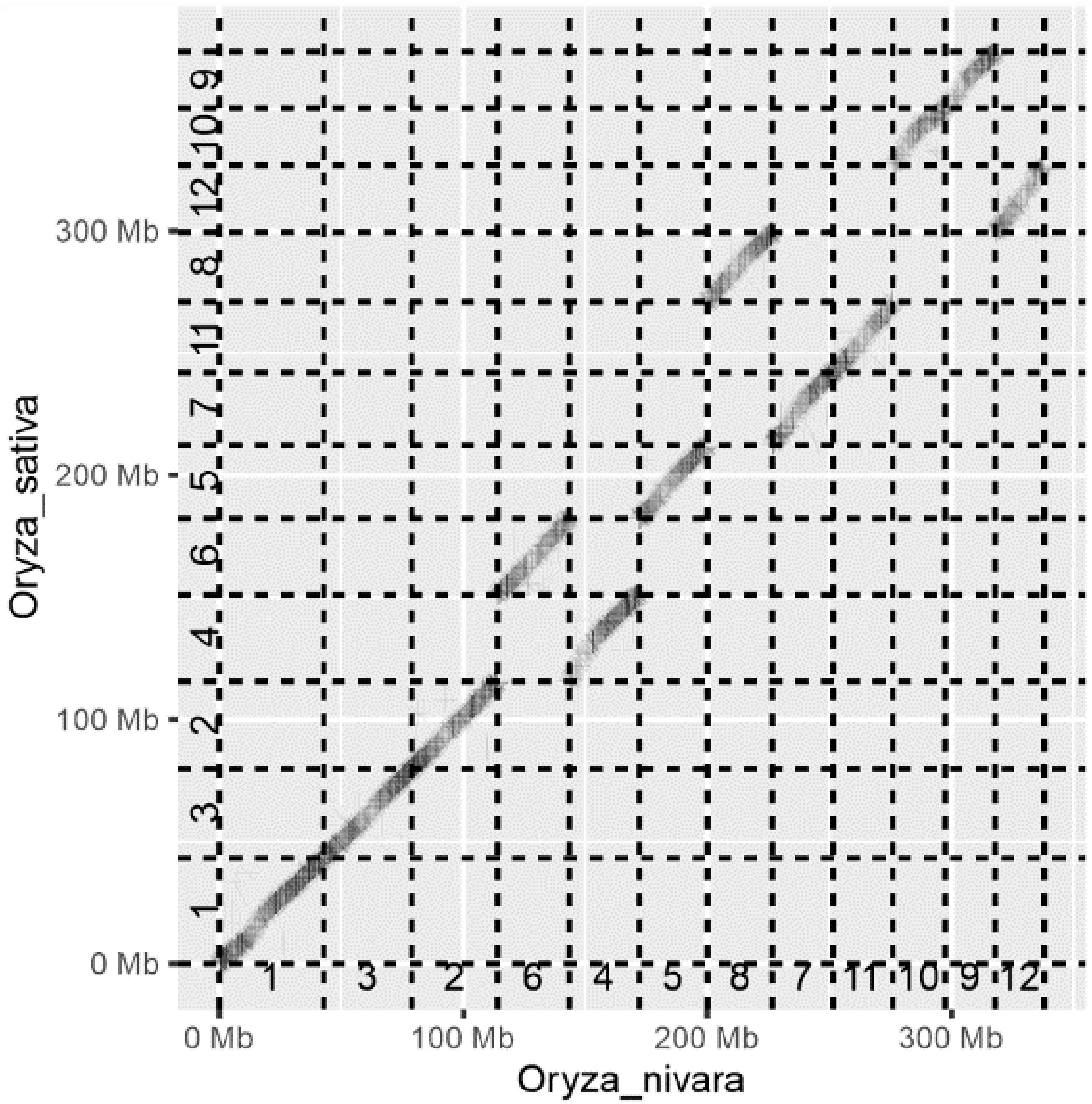
Region in chr1 of Oryza nivara (top) and Oryza sativa (bottom) in Ensembl

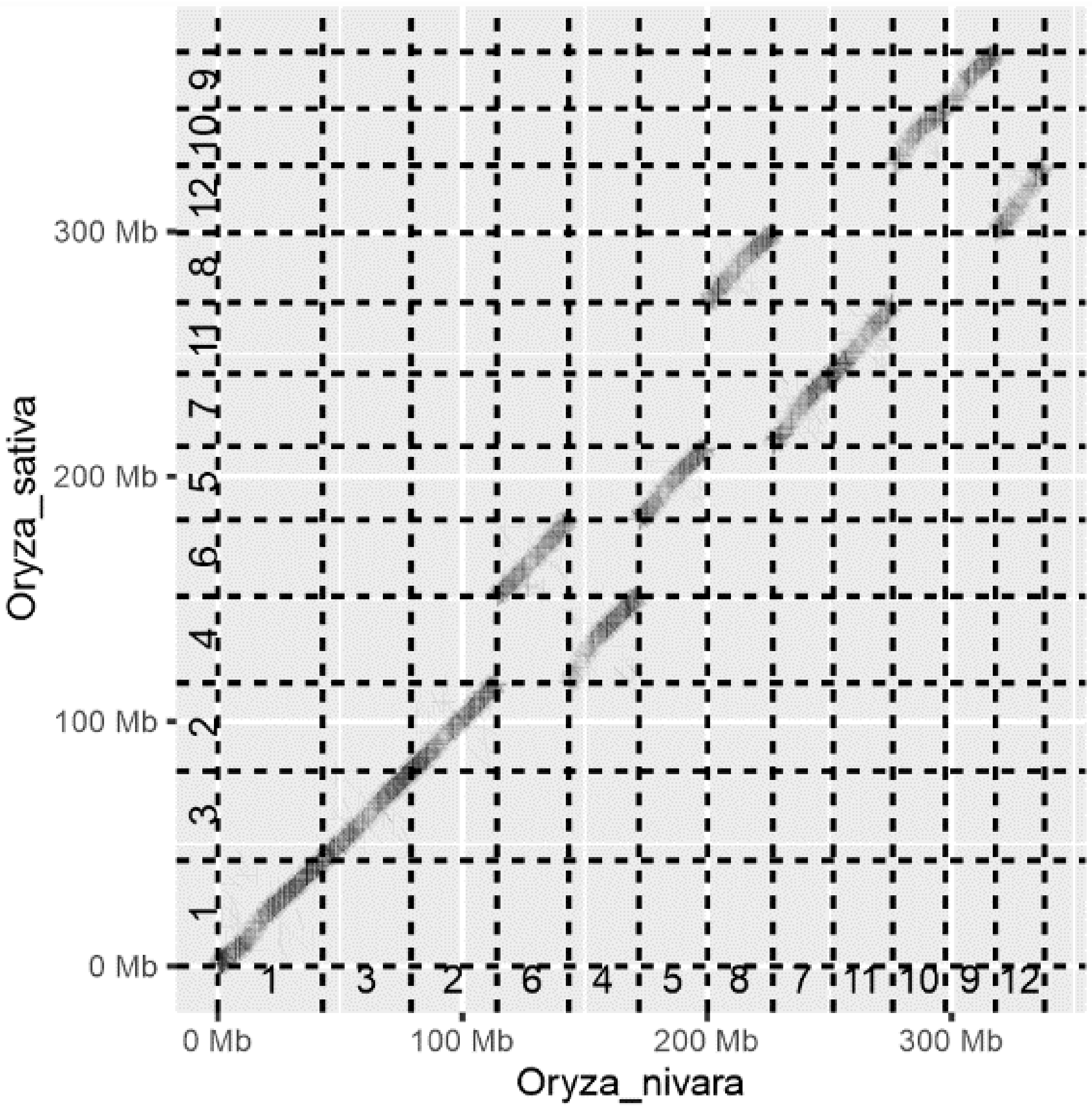


Oryza *nivara*	ONIVA01G00100	ONIVA01G00110	ONIVA01G00120	[gDNA segment]	ONIVA01G00130	ONIVA01G00140	[gDNA segment]	ONIVA01G00150
minimap2 overlap (bp)	7839, 3065	1787	2963	1169	2428	4063	564	1471
GSAlign overlap (bp)	7827, 6546		2956	1173	2415	4052	562	1501
Oryza sativa Japonica	Os01g0100100 Os01g0100200	Os01g0100300	Os01g0100400	Os01g0100466	Os01g0100500	Os01g0100600	Os01g0100650	RPS5

Contreras-Moreira et al (2023) https://doi.org/10.1101/2023.01.03.520531

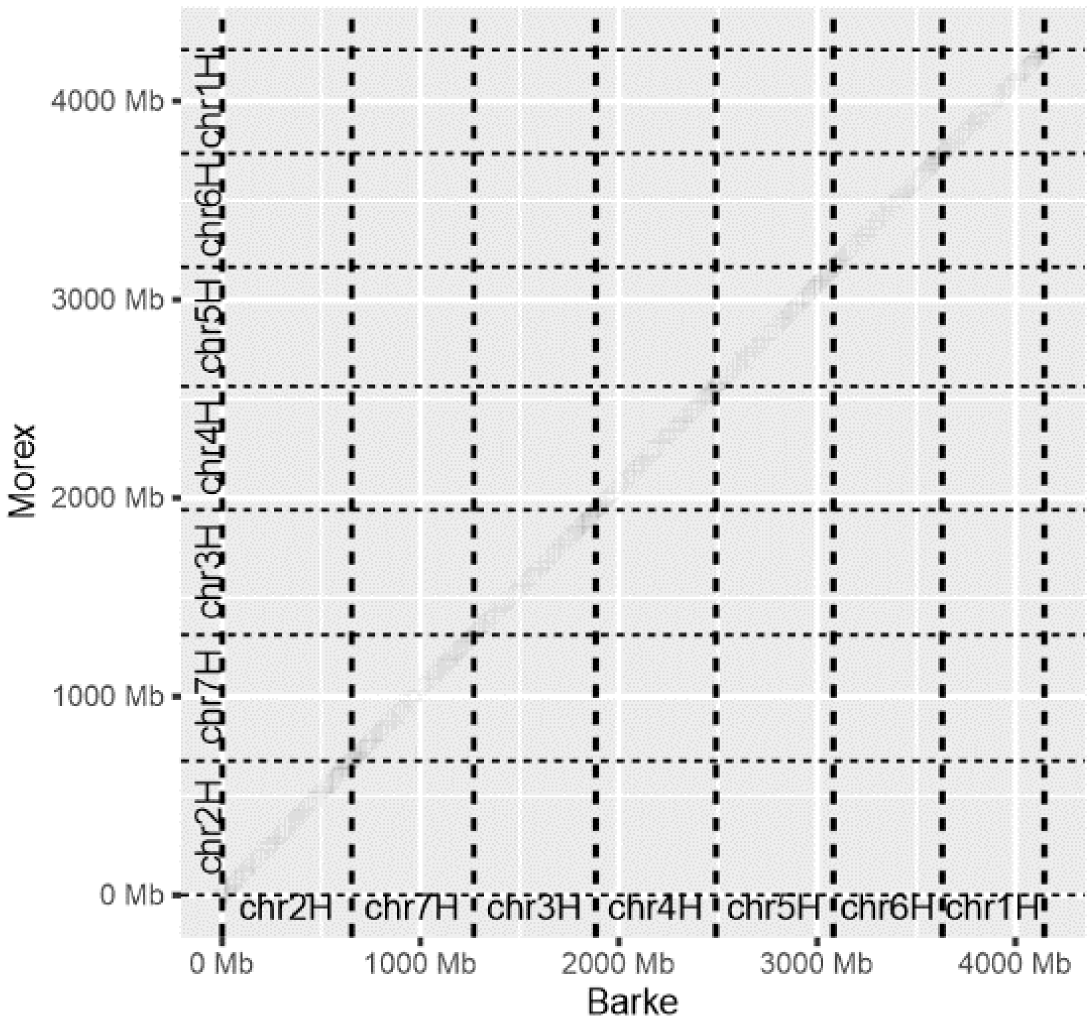
Pangenes are collinear in Whole Genome Alignments (rice)

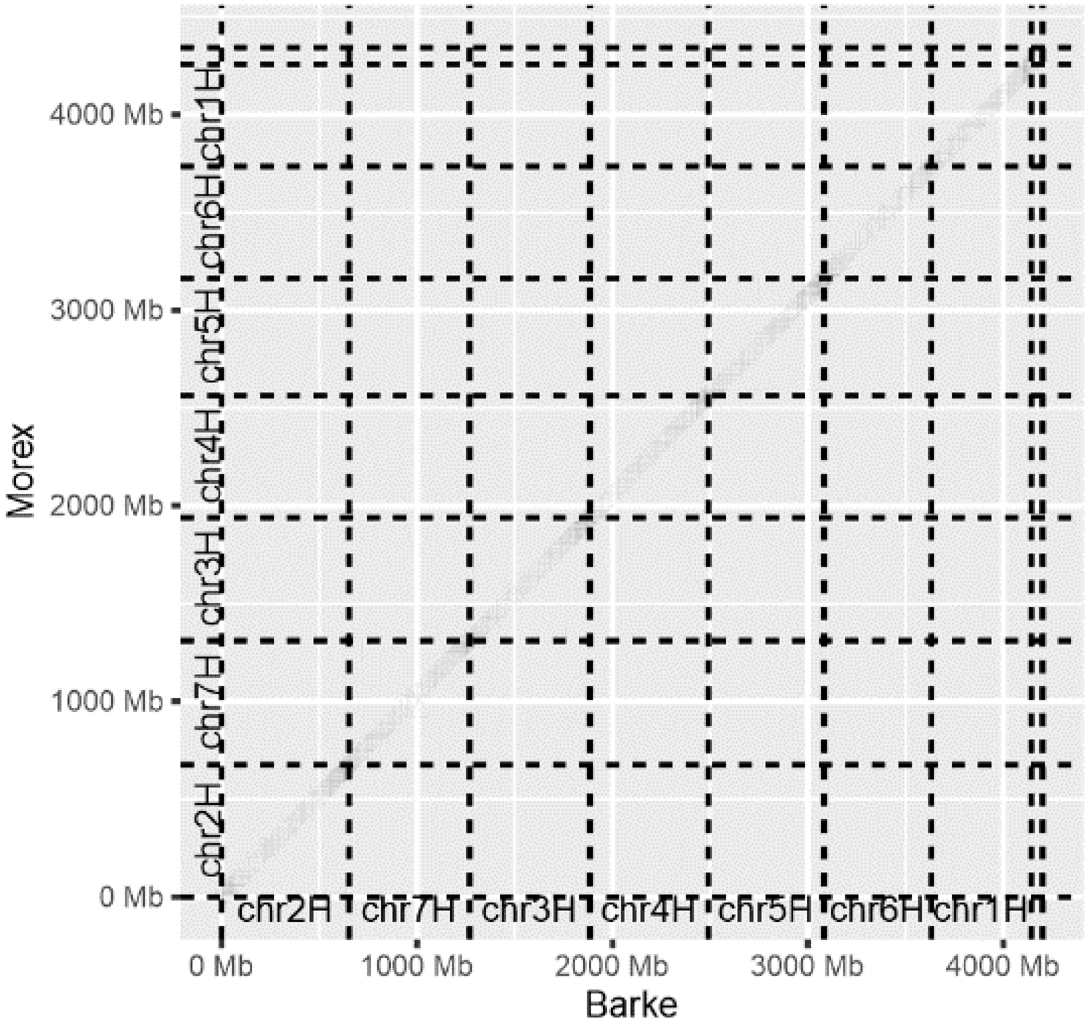




Contreras-Moreira et al (2023) https://doi.org/10.1101/2023.01.03.520531

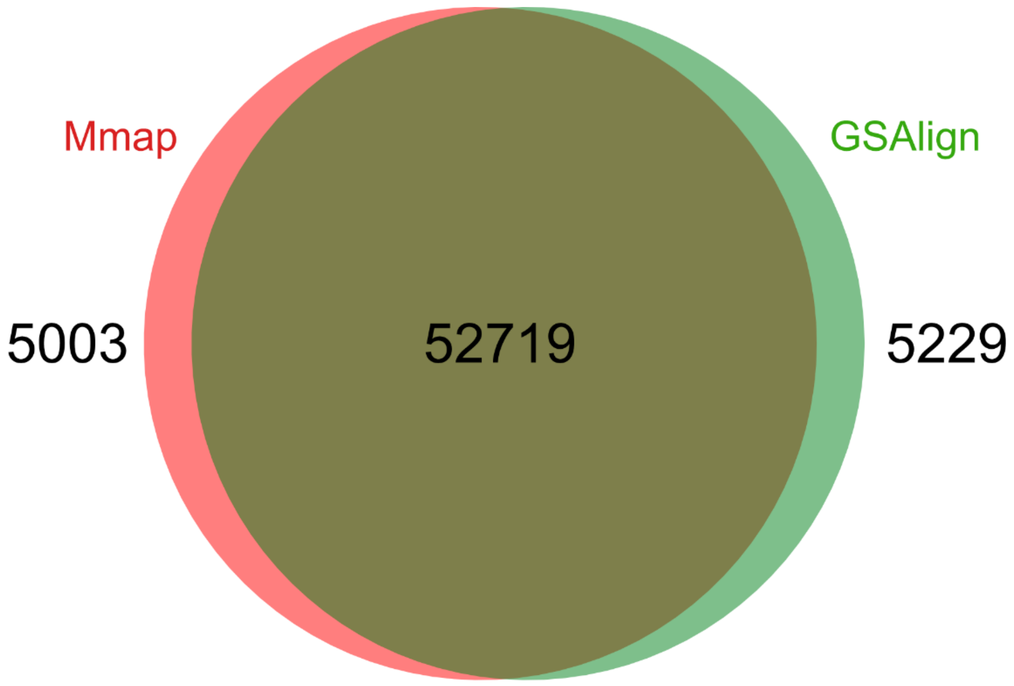
Pangenes are collinear in Whole Genome Alignments (barley)



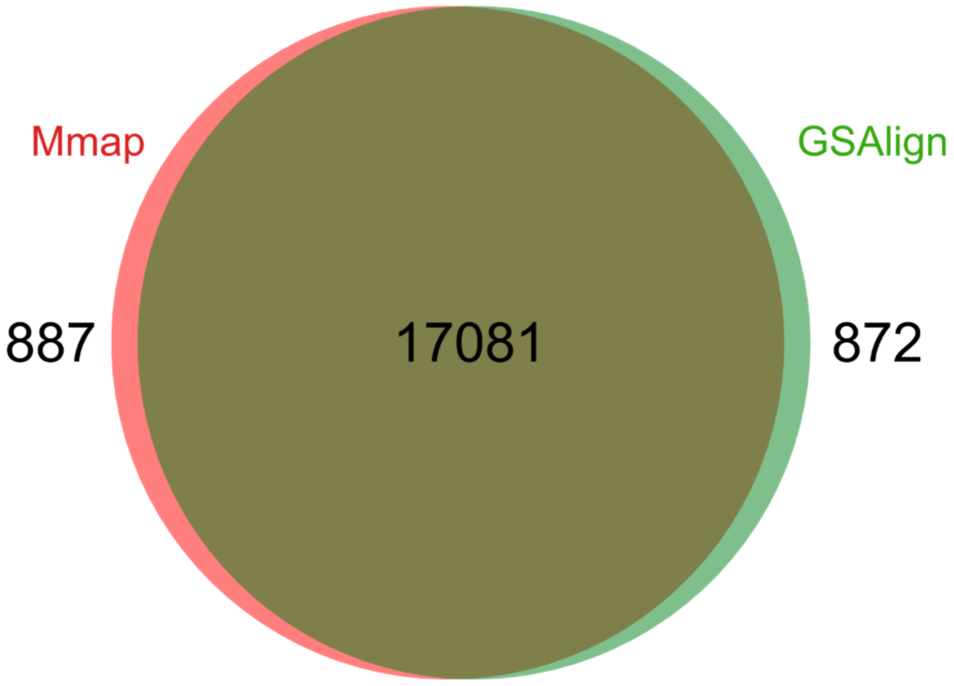


Contreras-Moreira et al (2023) https://doi.org/10.1101/2023.01.03.520531

Minimap2 and GSAlign produce similar pangenes (rice3)



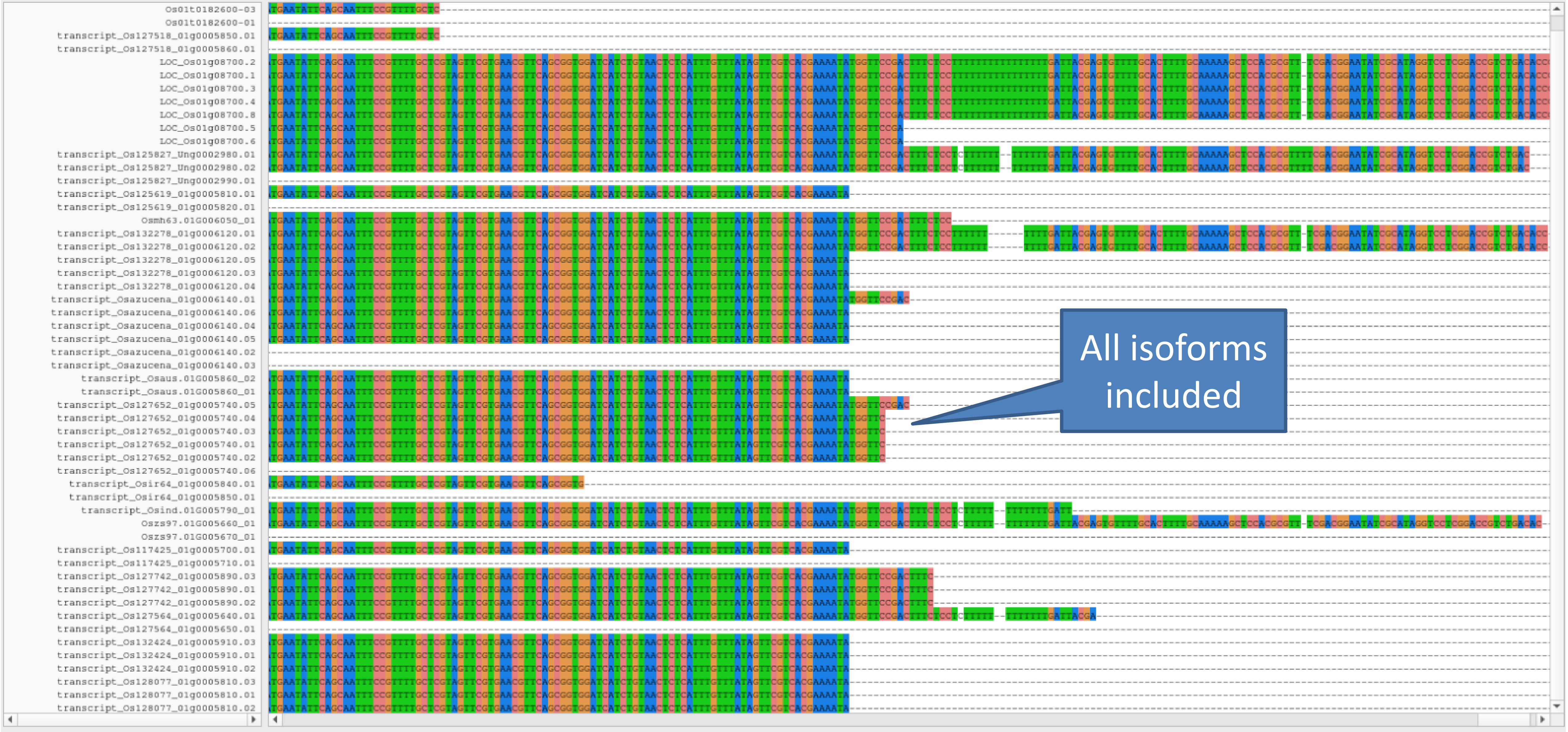
all clusters



core clusters

Contreras-Moreira et al (2023) https://doi.org/10.1101/2023.01.03.520531

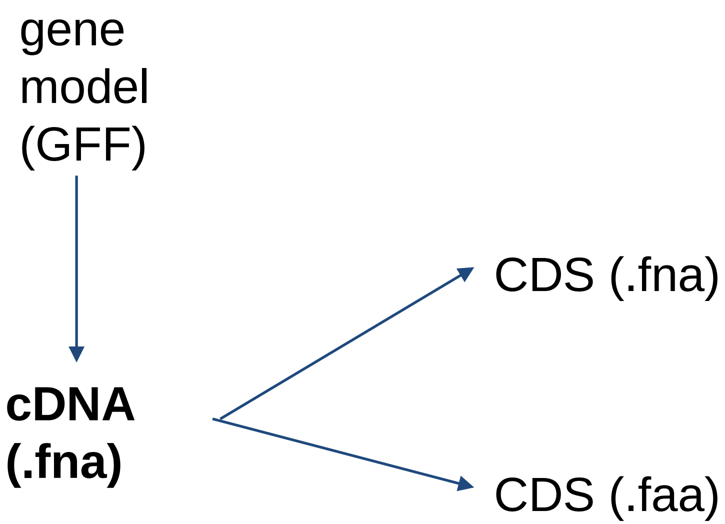
How do pangene clusters look like? GIGANTEA Os01g0182600, 3’



```
Os01t0182600-03
Os01t0182600-01
transcript_Os127518_01g0005850.01
transcript_Os127518_01g0005860.01
LOC_Os01g08700.2
LOC_Os01g08700.1
LOC_Os01g08700.3
LOC_Os01g08700.4
LOC_Os01g08700.8
LOC_Os01g08700.5
LOC_Os01g08700.6
transcript_Os125827_Ung0002980.01
transcript_Os125827_Ung0002980.02
transcript_Os125827_Ung0002990.01
transcript_Os125619_01g0005810.01
transcript_Os125619_01g0005820.01
Osmh63.01G006050_01
transcript_Os132278_01g0006120.01
transcript_Os132278_01g0006120.02
transcript_Os132278_01g0006120.05
transcript_Os132278_01g0006120.03
transcript_Os132278_01g0006120.04
transcript_Osazucena_01g0006140.01
transcript_Osazucena_01g0006140.06
transcript_Osazucena_01g0006140.04
transcript_Osazucena_01g0006140.05
transcript_Osazucena_01g0006140.02
transcript_Osazucena_01g0006140.03
transcript_Osaus.01G005860_02
transcript_Osaus.01G005860_01
transcript_Os127652_01g0005740.05
transcript_Os127652_01g0005740.04
transcript_Os127652_01g0005740.03
transcript_Os127652_01g0005740.01
transcript_Os127652_01g0005740.02
transcript_Os127652_01g0005740.06
transcript_Osir64_01g0005840.01
transcript_Osir64_01g0005850.01
transcript_Osind.01G005790_01
Oszs97.01G005660_01
Oszs97.01G005670_01
transcript_Os117425_01g0005700.01
transcript_Os117425_01g0005710.01
transcript_Os127742_01g0005890.03
transcript_Os127742_01g0005890.01
transcript_Os127742_01g0005890.02
transcript_Os127564_01g0005640.01
transcript_Os127564_01g0005650.01
transcript_Os132424_01g0005910.03
transcript_Os132424_01g0005910.01
transcript_Os132424_01g0005910.02
transcript_Os128077_01g0005810.03
transcript_Os128077_01g0005810.01
transcript_Os128077_01g0005810.02
```

All isoforms included

Different versions of pangene clusters: cDNA, CDS nucl, CDS pep



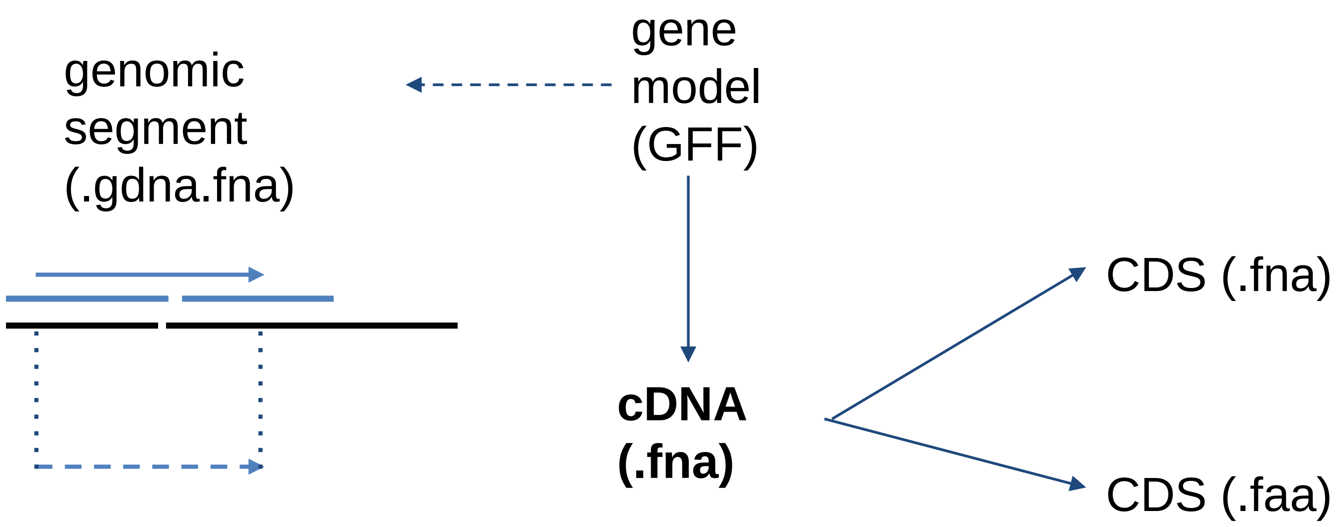
gene model (GFF)

cDNA (.fna)

CDS (.fna)

CDS (.faa)

Different versions of pangene clusters: cDNA, CDS nucl, CDS pep + **gDNA**



genomic segment (.gdna.fna)

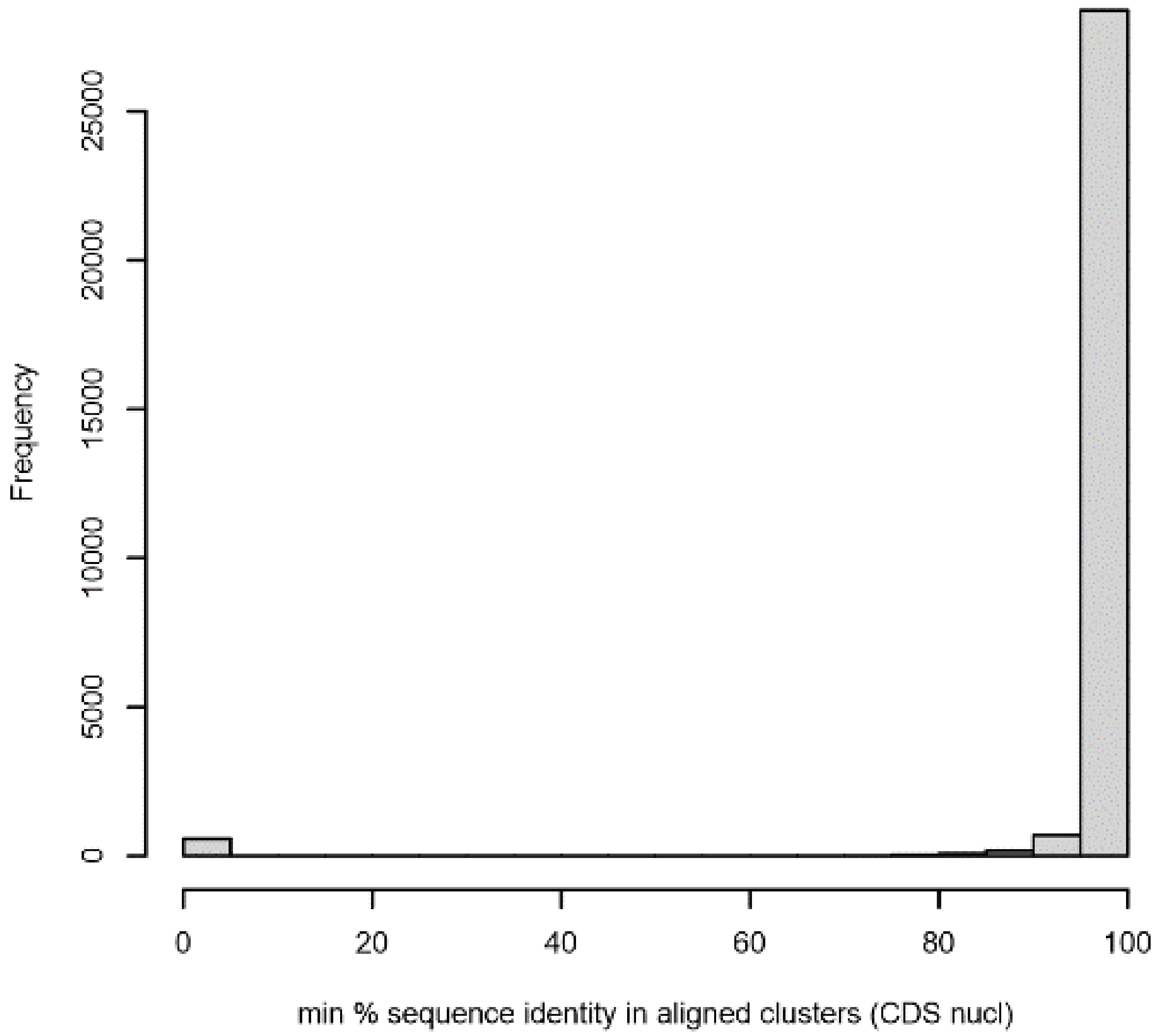
gene model (GFF)

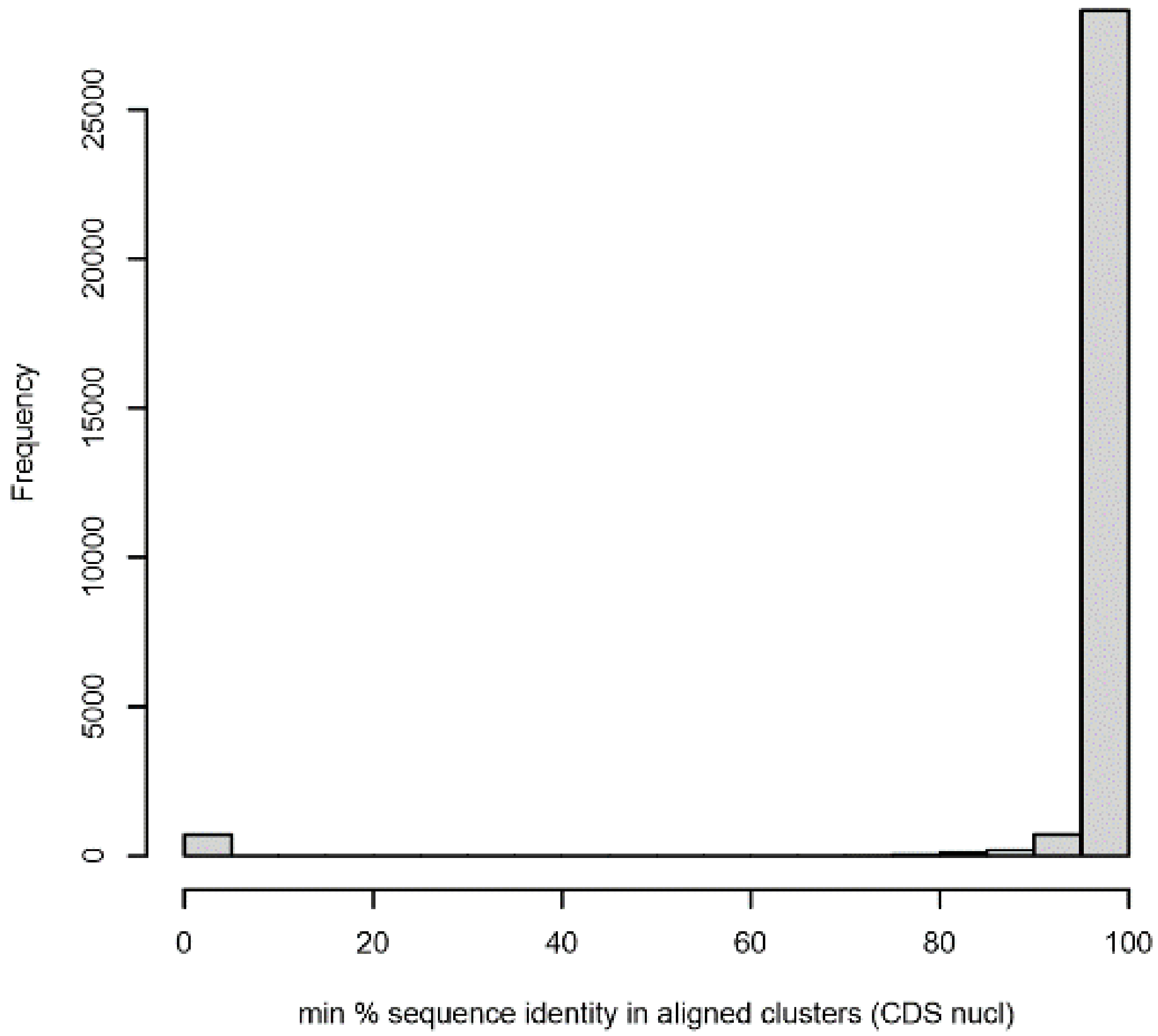
cDNA (.fna)

CDS (.fna)

CDS (.faa)

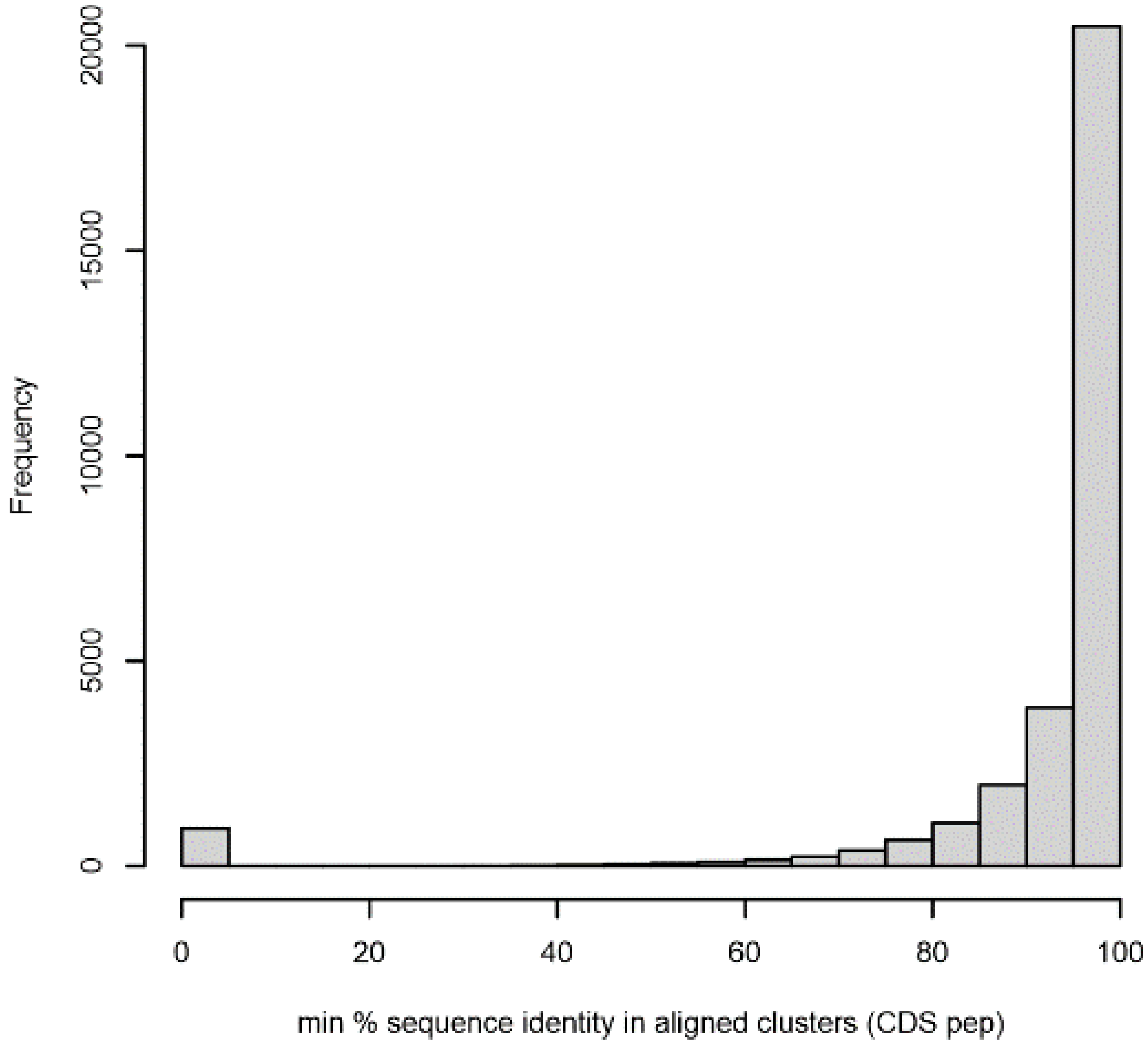
Minimap2 & GSAlign make pangene clusters with high sequence identity (nucl)

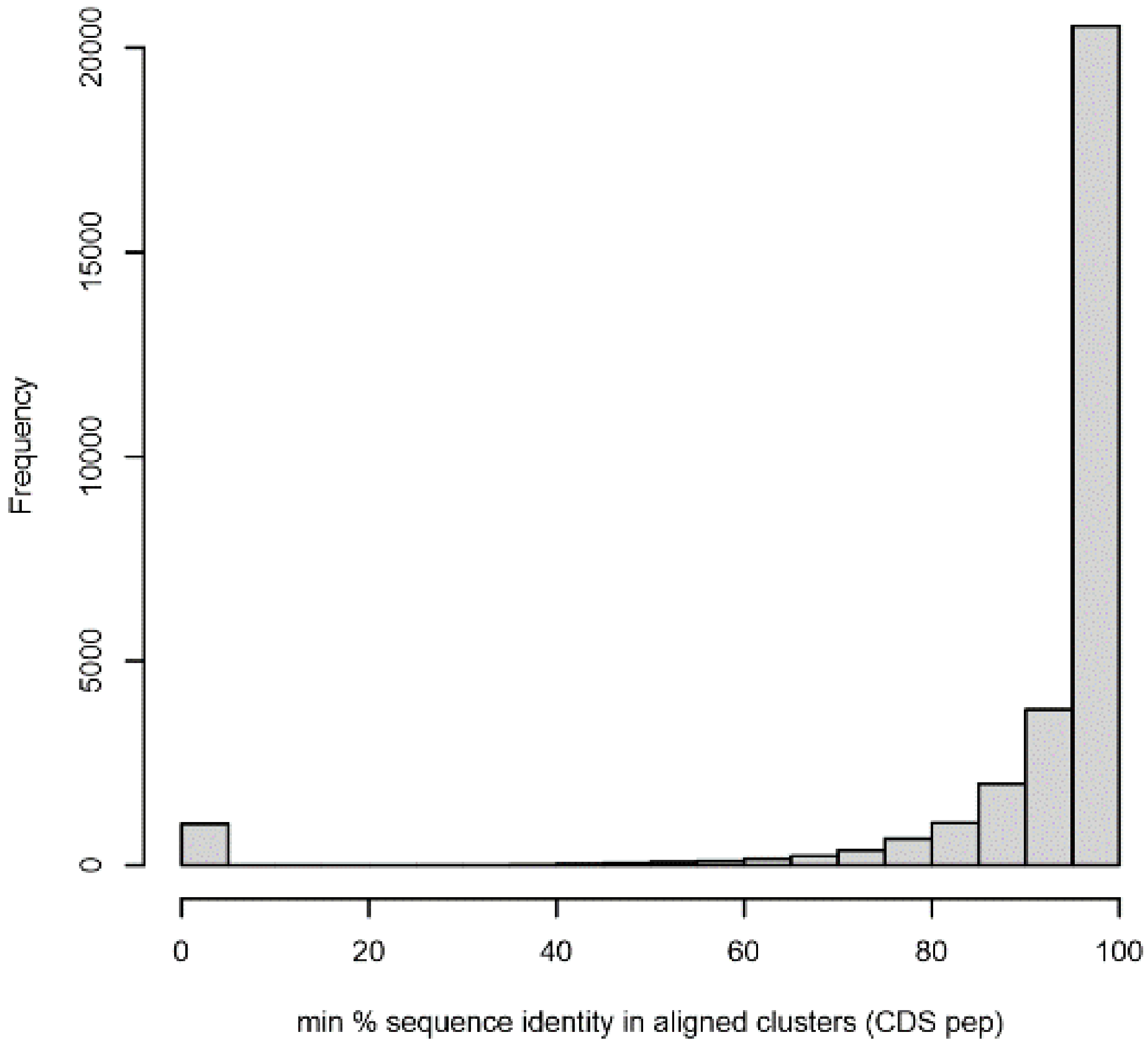




Contreras-Moreira et al (2023) https://doi.org/10.1101/2023.01.03.520531

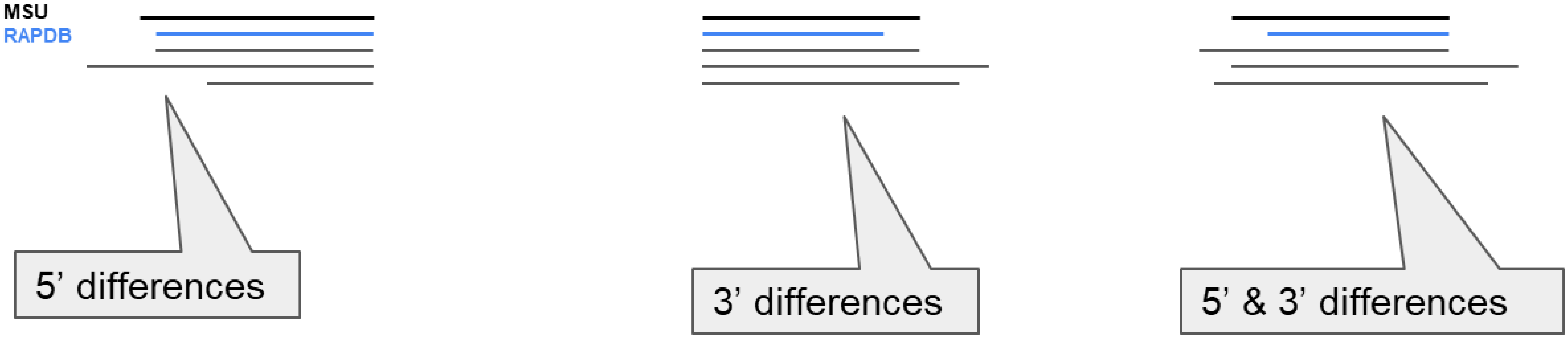
Minimap2 & GSAlign make pangene clusters with high sequence identity (pep)





Contreras-Moreira et al (2023) https://doi.org/10.1101/2023.01.03.520531

Apps: pangene clusters provide evidence for gene model annotation (rice)



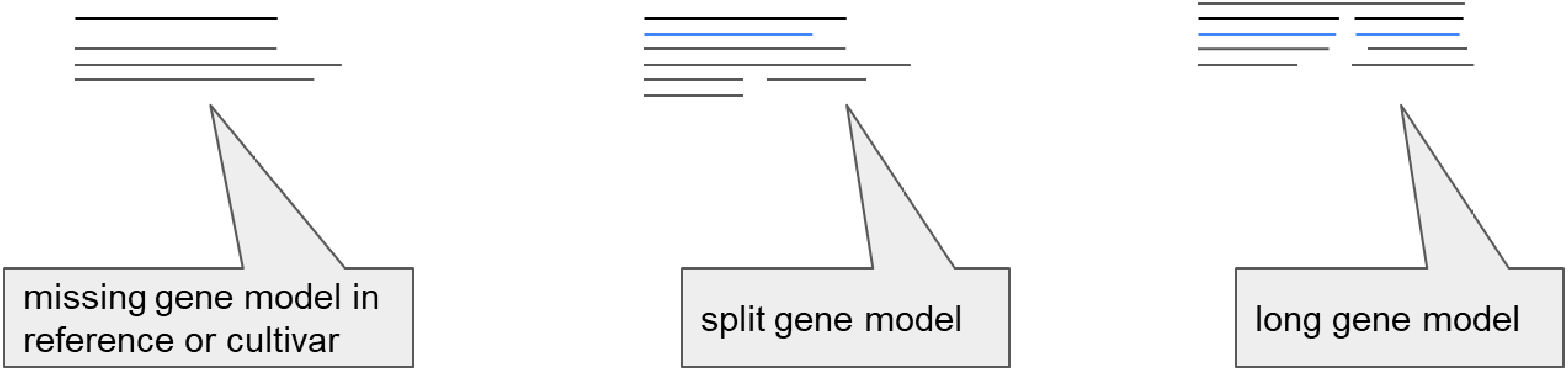
5’ differences

3’ differences

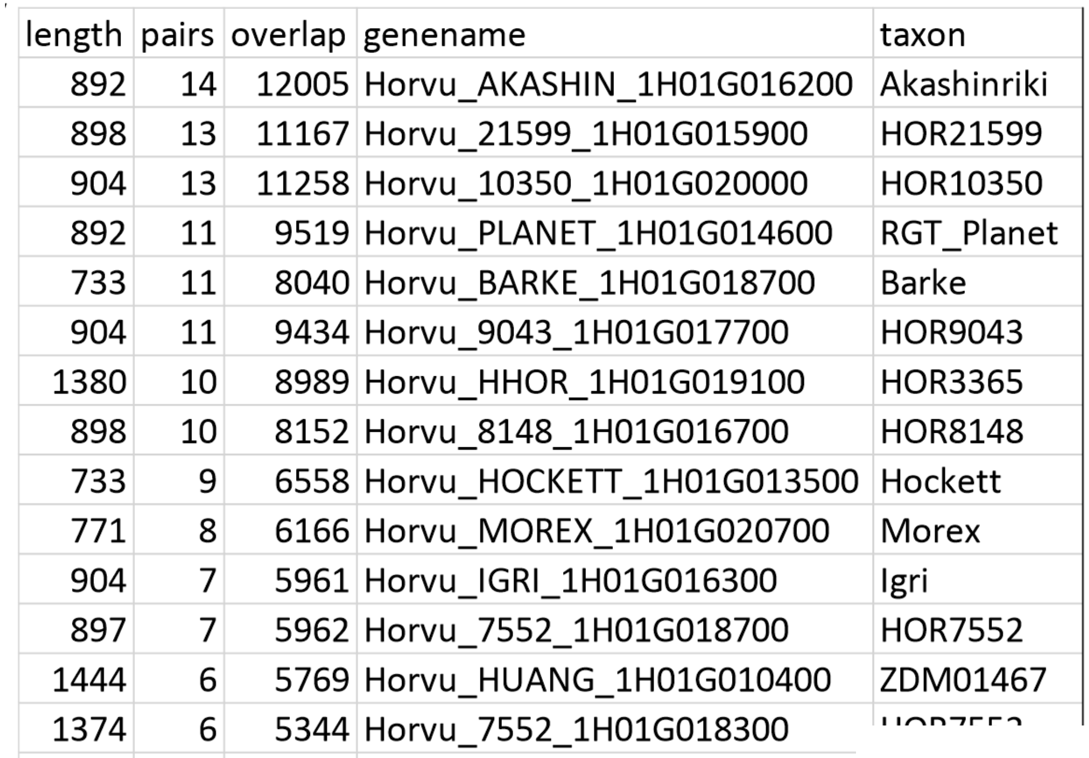
5’ & 3’ differences

benchmark set	isoforms	in pangenes	match cluster mode
Oryza sativa genes curated in RAP-db	3895	2937	1700 (57.8%)
Oryza proteins curated in SwisProt	5685	3876	1650 (42.5%)

Apps: pangene clusters provide evidence to fix annotation errors (rice)

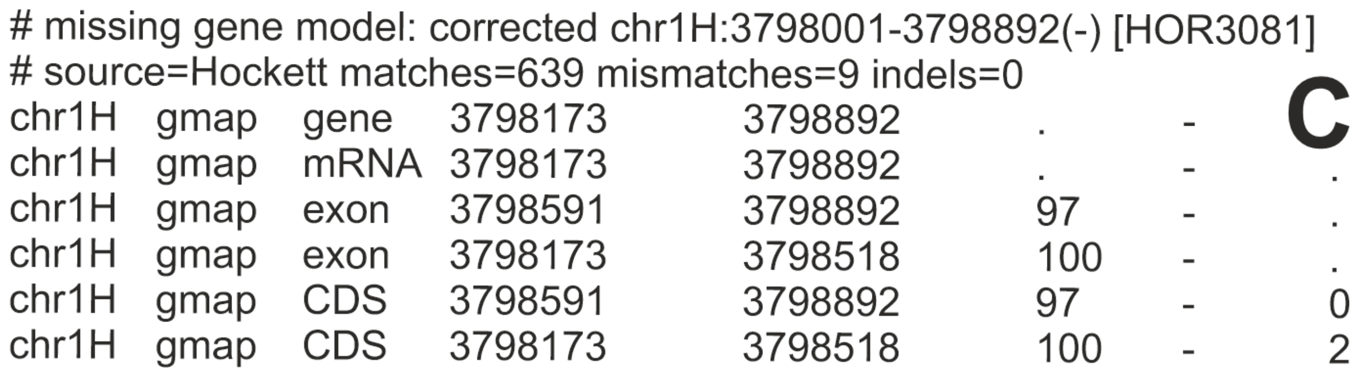


Example: validating PAV in barley



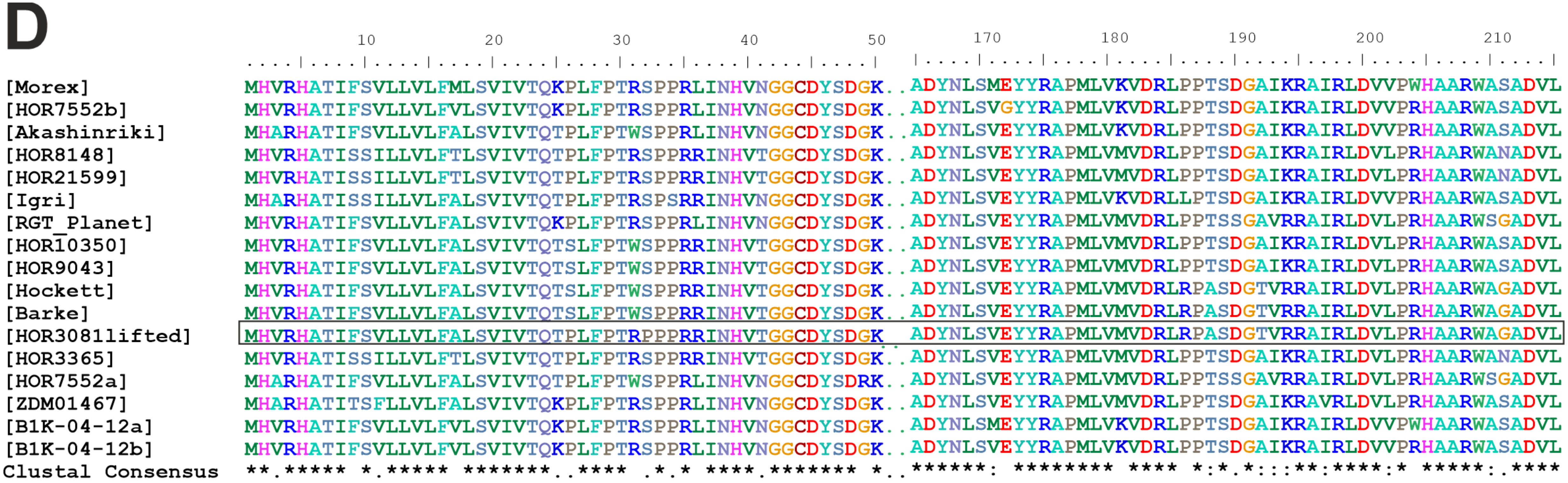
length	pairs	overlap	genename	taxon
892	14	12005	Horvu_AKASHIN_1H01G016200	Akashinriki
898	13	11167	Horvu_21599_1H01G015900	HOR21599
904	13	11258	Horvu_10350_1H01G020000	HOR10350
892	11	9519	Horvu_PLANET_1H01G014600	RGT_Planet
733	11	8040	Horvu_BARKE_1H01G018700	Barke
904	11	9434	Horvu_9043_1H01G017700	HOR9043
1380	10	8989	Horvu_HHOR_1H01G019100	HOR3365
898	10	8152	Horvu_8148_1H01G016700	HOR8148
733	9	6558	Horvu_HOCKETT_1H01G013500	Hockett
771	8	6166	Horvu_MOREX_1H01G020700	Morex
904	7	5961	Horvu_IGRI_1H01G016300	Igri
897	7	5962	Horvu_7552_1H01G018700	HOR7552
1444	6	5769	Horvu_HUANG_1H01G010400	ZDM01467
1374	6	5344	Horvu_7552_1H01G018300	HOR7552
1380	2	2247	Horvu_FT11_1H01G019600	
1380	2	1661	Horvu_FT11_1H01G019300	

GMAP lift-over
Genomic segment



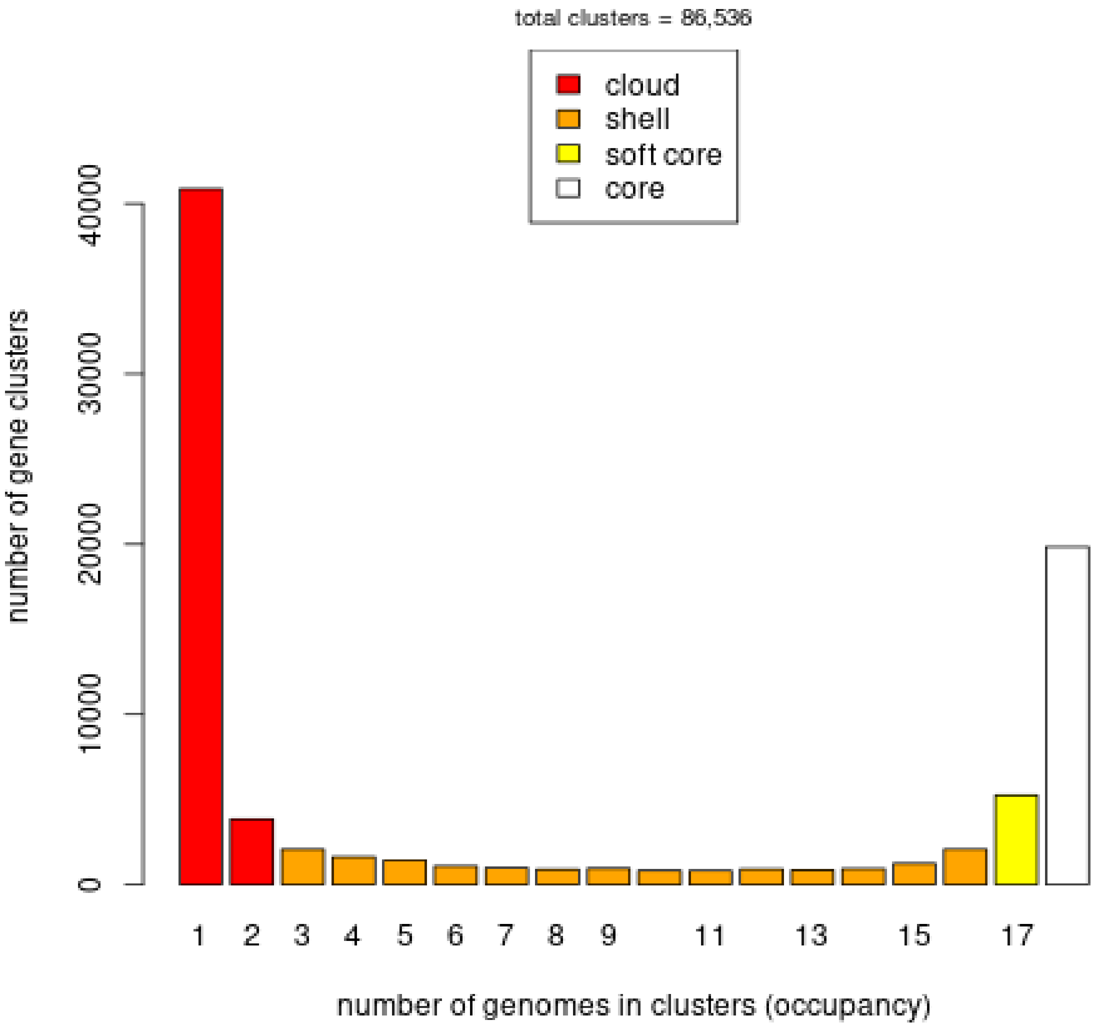
```
# missing gene model: corrected chr1H:3798001-3798892(-) [HOR3081]
# source=Hockett matches=639 mismatches=9 indels=0
chr1H  gmap  gene  3798173  3798892  .    -
chr1H  gmap  mRNA  3798173  3798892  .    -   .
chr1H  gmap  exon  3798591  3798892  97   -   .
chr1H  gmap  exon  3798173  3798518  100  -   .
chr1H  gmap  CDS   3798591  3798892  97   -   0
chr1H  gmap  CDS   3798173  3798518  100  -   2
```

Example: validating PAV in barley



```
D
                                10        20        30        40        50      170       180       190       200       210
                       ....|....|....|....|....|....|....|....|....|....|  |....|....|....|....|....|....|....|....|....|....|
[Morex]                MHVRHATIFSVLLVLFMLSVIVTQKPLFPTRSPPRLINHVNGGCDYSDGK..ADYNLSMEYYRAPMLVKVDRLPPTSDGAIKRAIRLDVVPWHAARWASADVL
[HOR7552b]             MHVRHATIFSVLLVLFVLSVIVTQKPLFPTRSPPRLINHVNGGCDYSDGK..ADYNLSVGYYRAPMLVKVDRLPPTSDGAIKRAIRLDVVPRHAARWASADVL
[Akashinriki]          MHARHATIFSVLLVLFALSVIVTQTPLFPTWSPPRLINHVNGGCDYSDGK..ADYNLSVEYYRAPMLVKVDRLPPTSDGAIKRAIRLDVVPRHAARWASADVL
[HOR8148]              MHVRHATISSILLVLFTLSVIVTQTPLFPTRSPPRRINHVTGGCDYSDGK..ADYNLSVEYYRAPMLVMVDRLPPTSDGAIKRAIRLDVLPRHAARWANADVL
[HOR21599]             MHVRHATISSILLVLFTLSVIVTQTPLFPTRSPPRRINHVTGGCDYSDGK..ADYNLSVEYYRAPMLVMVDRLPPTSDGAIKRAIRLDVLPRHAARWANADVL
[Igri]                 MHARHATISSILLVLFALSVIVTQTPLFPTRSPSRRINHVNGGCDYSDGK..ADYNLSVEYYRAPMLVKVDRLLPTSDGAIKRAIRLDVVPRHAARWASADVL
[RGT_Planet]           MHVRHATIFSVLLVLFALSVIVTQKPLFPTRSPPRLINHVNGGCDYSDGK..ADYNLSVEYYRAPMLVMVDRLPPTSSGAVRRAIRLDVLPRHAARWSGADVL
[HOR10350]             MHVRHATIFSVLLVLFALSVIVTQTSLFPTWSPPRRINHVTGGCDYSDGK..ADYNLSVEYYRAPMLVMVDRLPPTSDGAIKRAIRLDVLPRHAARWASADVL
[HOR9043]              MHVRHATIFSVLLVLFALSVIVTQTSLFPTWSPPRRINHVTGGCDYSDGK..ADYNLSVEYYRAPMLVMVDRLPPTSDGAIKRAIRLDVLPRHAARWASADVL
[Hockett]              MHVRHATIFSVLLVLFALSVIVTQTSLFPTWSPPRRINHVTGGCDYSDGK..ADYNLSVEYYRAPMLVMVDRLRPASDGTVRRAIRLDVLPRHAARWAGADVL
[Barke]                MHVRHATIFSVLLVLFALSVIVTQTSLFPTWSPPRRINHVTGGCDYSDGK..ADYNLSVEYYRAPMLVMVDRLRPASDGTVRRAIRLDVLPRHAARWAGADVL
[HOR3081lifted]        MHVRHATIFSVLLVLFALSVIVTQTPLFPTRPPPRRINHVTGGCDYSDGK  ADYNLSVEYYRAPMLVMVDRLRPASDGTVRRAIRLDVLPRHAARWAGADVL
[HOR3365]              MHVRHATISSILLVLFTLSVIVTQTPLFPTRSPPRRINHVTGGCDYSDGK..ADYNLSVEYYRAPMLVMVDRLPPTSDGAIKRAIRLDVLPRHAARWANADVL
[HOR7552a]             MHARHATIFSVLLVLFALSVIVTQTPLFPTWSPPRLINHVNGGCDYSDRK..ADYNLSVEYYRAPMLVMVDRLPPTSSGAVRRAIRLDVLPRHAARWSGADVL
[ZDM01467]             MHARHATITSFLLVLFALSVIVTQKPLFPTRSPPRLINHVNGGCDYSDGK..ADYNLSVEYYRAPMLVMVDRLPPTSDGAIKRAVRLDVLPRHAARWASADVL
[B1K-04-12a]           MHVRHATIFSVLLVLFVLSVIVTQKPLFPTRSPPRLINHVNGGCDYSDGK..ADYNLSMEYYRAPMLVKVDRLPPTSDGAIKRAIRLDVVPWHAARWASADVL
[B1K-04-12b]           MHVRHATIFSVLLVLFVLSVIVTQKPLFPTRSPPRLINHVNGGCDYSDGK..ADYNLSVEYYRAPMLVKVDRLPPTSDGAIKRAIRLDVVPRHAARWASADVL
Clustal Consensus      **.***** *.***** *******..**** .*.* ****.******* *..******: ********* **** *:*.*:::**:****:* *****:.****
```

Apps: pangene clusters capture pangenome dynamics (rice)



	% BUSCO complete
softcore (all isoforms)	**96.9**
oryza_sativa_RAPDB.cds	85.3
oryza_sativa_MSU.cds	94.7
oryza_sativa_arc.cds	94.7
oryza_sativa_azucena.cds	95.3
oryza_sativa_chaomeo.cds	95.4
oryza_sativa_gobolsailbalam.cds	94.4
oryza_sativa_ir64.cds	94.8
oryza_sativa_ketannangka.cds	95.0
oryza_sativa_khaoyaiguang.cds	95.2
oryza_sativa_larhamugad.cds	94.1
oryza_sativa_lima.cds	94.5
oryza_sativa_liuxu.cds	94.7
oryza_sativa_natelboro.cds	93.4
oryza_indica.cds	95.1
oryza_sativa_ZS97.cds	94.5
oryza_sativa_n22.cds	95.3
oryza_sativa_mh63.cds	94.5

TO BE DONE

- PanOryza project: define rice pangenes that will be curated in UniProt in collaboration with the rice community

- Create rules to name pangenes:
  - stable and consistent
  - support future addition of new annotation sets

- For you: try it out at https://github.com/Ensembl/plant-scripts

Building Pangene Sets from Plant Genome Alignments Confirms Presence-Absence Variation

Bruno Contreras-Moreira, Shradha Saraf, Guy Naamati, Sandeep S. Amberkar, Paul Flicek, Andrew R. Jones, Sarah Dyer

@BrunoContrerasM

EXTRA SLIDES

	max RAM (GB)	WGA N50 (Kbp)	% genes blocks3+	total clusters	(soft) core clusters	% BUSCO complete
ACK2	4.5	6.1	34.3	43,951	15,768	74.3
rice3	1.4	[27.4, 29]	[74.4, 78.4]	62,915	18,681	83.9
chr1wheat10 (-H)	64.5	[80.8, 142.4]	[38.8, 54.2]	30,173	7,872	
barley20 (-H)	46.3	[43.6, 75.7]	[24.9, 35.4]	180,667	15,674 (23,888) {34,074}	61.3 (82.1) {95.7}

Contreras-Moreira et al (2023) https://doi.org/10.1101/2023.01.03.520531

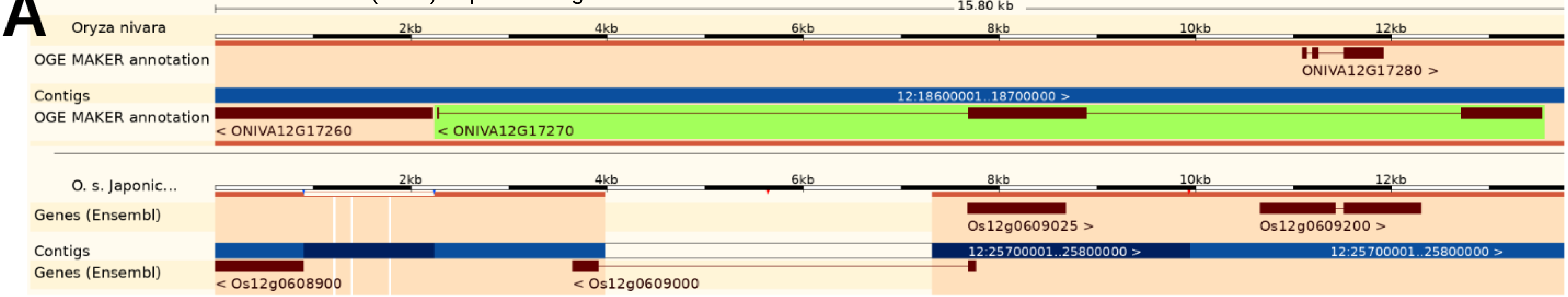
	max RAM (GB)	WGA N50 (Kbp)	% genes blocks3+	total clusters	(soft) core clusters	% BUSCO complete	% ANI
ACK2	4.5	4.3	23.4	43,340	16,432	74.7	84.7
rice3	3.3	[15.2, 16.9]	[51.6, 56.6]	62,844	18,626	84.2	[96.4, 97.6]
chr1wheat10	83.4	[40.9, 72.1]	[20.2, 34.2]	30,135	7,723		[98.9, 99.4]
barley20	113.1	[17.1, 34.3]	[10.5, 15.9]	173,984	13,934 (21,171)	56.8 (76.2)	[96.9, 99.3]
barley20 -H	110.1	[16.6, 32.9]	[10.5, 15.9]	188,289	13,957		[96.8, 99.3]

Contreras-Moreira et al (2023) https://doi.org/10.1101/2023.01.03.520531

	dataset	core clusters	multiple copies	shell clusters	gDNA segments	match Compara	share InterPro domains
Compara orthogroups	ACK2	20,192	161				[18,259]
minimap2 clusters	ACK2	15,768	490			14,044	[14,371]
GSAlign clusters	ACK2	16,432	446			14,145	[14,790]
Compara orthogroups	rice3	13,020	219	6,386			16,766 [11,571]
minimap2 clusters	rice3	20,419	3,022	9,503	5,593	17,317	22,796 [17,232]
GSAlign clusters	rice3	20,224	2,831	9.863	6,173	17,103	22,818 [16,957]

Contreras-Moreira et al (2023) https://doi.org/10.1101/2023.01.03.520531

Contreras-Moreira et al (2023) https://doi.org/10.1101/2023.01.03.520531



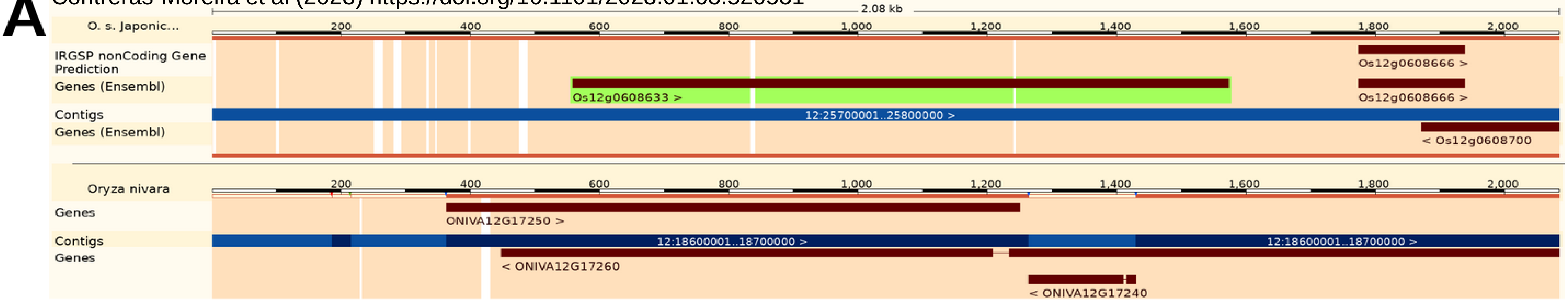
B

>transcript:ONIVA12G17270.1 gene:ONIVA12G17270 12:18626386-18637647(-) [Oryza_nivara]
MQLLFFSFLFLLLARETSAVAADGCSRRCGGLVVPYPFGFSGSCPIMLSCNVDGGSNSTAALILQGNDAT
TTDRSYTVVDGSFNSTASTFTVSVPPSCNRTVSDARRWLSGANYGVSSRTGLFLRGCRNATSSDCSVPVE
TMLRTTRCSGGGGNETASSSLTCIASLSPATPAERGLGGLFAQWEKVEEPRCENLLTSVYGDTREGVFSL
EFAAAEMRWWVNGSCGGGVDDLGRCAANATCIPMQTPSGNWGHRCECLPWMAGDGFAAGEGCYAGKRRRM
RVVEFATAGSVAFLLCLALSVWCLLRRRQWRRNNAKLTVKMARKHLPKDARFFRGKPIEDELELEAAGPR
RFHYGELAAATANFSDDRRLGSGGFGSVYRGFLNGGDVAVKRVAETSRQGWKEFVAEVRIISRLRHRNLV
PLVGWCHDGGDELLLVYELMPNGSLDAHIHSSGNVLPWPARYEVVLGVGAALMYLHHEAEQRVVHRDIKP
SNVMLDASFSARLGDFGLARLIDDGRRSRTTGIAGTMGYIDAECFLLAGRASVESDVYSFGVVLLEVACG
RRPAVVINGGEDAIHLTQWVWDTHGGAAGGGILDAADTRLNGEFDVAEMERVLAVGLWCAHPDRGLRPSI
RQAVSVLRFEAPLPSLPVR**MPVATYGPPVSTASAPTSNDTSAGR**LHP

>transcript:Os12t0609000-00 gene:Os12g0609000 12:25722111-25722886(-) [Oryza_sativa]
MPVATYGPPVSTASAPTSNDTSAGRDSATRTVKSEDPLPPRLYARQGQLDSHLFPLAFIEPPFVEHLACM
PIELAIAICLALHLVRRRAPPACACHPLAVLAFLSPWRPSA

Contreras-Moreira et al (2023) https://doi.org/10.1101/2023.01.03.520531

A



B

>transcript:Os12t0608633-00 gene:Os12g0608633 12:25710726-25711730(+) [Oryza_sativa]
CGGTGGCAGGAACGTCGCCACCGGCATCCTCGCCGGGAGGCTCGGCGGCGGCGCCTCACCCCTCAGCACGCTGATGGCCTGCCTGATCACCGGCCTCACGCTCCGGTCAGGGTGCGCGCACCAGAGCCCGACAACCATG
ACACGCTCCATCTCGCCACCGTCGAACTCGCCGGTGAGCCGCCGGTCAGCAGCGTCAAGAATCCTCCCATTGCCGTACAAATCCCAGACCCATTGAGCAAGGTGGATCCGGTCTTCGTCTACCTCGGATTGGTGGTCCG
CCATTATCGGTCGCCGGCCACAGGCGATCTCAAGGAGGAGGACGCCAAAGCTGTACGCGTCGGACTCGGCGTTGGCCCTGCCGGTGATCATGCACTCCGGGTCCATGTACCCCATCGTGCCGGCGAGCACCGTGGTGTG
TGGTCCCCGGCCGTGGTCGACAAGCCTGGCGAGGCCGAAGTCGCCGAGCTTAGCGTTGAAGGCGGCGTCGAGCATGATGTTGCTTGGCTTGATGTCGCGGTGCACCACGCACTGCTCCCACTCCTCGTGCAGGTACAGC
AGAGCCGAGCCGATCCCGAGCACGATCTCGTGCCTGAGTGGCCATGGGAGCACGCCGGCGCTGGCTTTGTAGAGGTGGGTGTCGAGGCTGGCGTTGGGCATGAGCTCGTAGACGAGGAGCTCGCCACCGCCGTGGCACC
AGCCGATGAGCTGCACCAGGTTGCGGTGGCGGAGGCGGCTGATGATCCGCACCTCCGACGCGTACTCCTTCCTCCCCTGCTTCGAGCTCTTGGACACCCTTTTAATGGCGACGTCAAGGTTCAACTCTTTCAAGAATCC
TCTGTACACTGATCCGAACCCGCCTTCCCCGAGCTTGTGCTCGTCGGAGAAGTCGTCGGTGGCGATGGCGAGCTCGCCAAAGCGGAATCTCTTCGGCCCTGTCCCCTTCTCGAAGTCGTCTTCCATGGCCGTCTCATCA
TCGAAGATGCCCCCCTCTTCCATCTCCTGCTC

>transcript:ONIVA12G17250.1 gene:ONIVA12G17250 12:18611207-18616986(+) [Oryza_nivara]
ATGGCGGCCACCCTCCTCTCCCGGGCAGCCGCCGCCGCCGCCGTCTCCCTCCGCGGCGCCCGCTCGCACCACATATTGCCCTCCTCCCTCCCCAAGGAAACCCTCCTTCCTCCTCTCCTCCTCCTCATCCTCACCACCA
CCACCGCCTCGCTCCTCGCGGTCGGGGGAAGAGTCGGATGGGCGCGCGCCGCCGAGGAGGGCGCCGGGTTCGGGTGCAGGGCGTCGGTGCCGGCGGCGCCGGGGGGCGTCGGTTCGTTCGGGATCGCGGCGCGCTGCAA
CGCCACCTCGTCGTCGGCCGTTTCGGAGGCTACCAACGCGCTGCCGAGGACGGAGCCCGTGGTTTCGGCCGAGTGGCTGCACGCCAACCTCAAGGACCCCGATGTGAAGGTACTTGATGCCTCTTGGTATATGCCTGCC
GAACAAAGGAATCCTCTTCAGGAGTACCAGGTGGCTCATATTCCTGGGGCGCTATTCTTTGATGTTGATGGAATATCAGACAGAACATCAAGCTTGCCGCACATGCTGCCATCTGAAAAGGCATTTTCTGCTGCTGTAT
CTTCGCTTGGAATATACAACAAAGATGGGATAGTAGTTTATGATGGAAAGGGACTATTTAGCGCTGCTCGTGTTTGGTGGATGTTTCGTGTTTTCGGACATGATAAAGTTTGGGTGTTGGATGGTGGTTTGCCCCAATG
GCGTGCTTCTGGGTATGATGTTGAATCAAGTGCCTCTAGTGATGCCATCTTAAAAGCCAGTGCTGCTCGTGAAGCAATTGAGAAAGTTTATCAGGGGCAGTTGGTTGGTCCCTCCACATTTGAAGCAAAGTTGCAGCCT
CATCTTATTTGGAATCTTGACCAGGTAAAAGAGAACATTGATGCCAAGACACATCAACTTATAGATGCTCGAGGAAAGCCTAGATTTGATGGTGCAGTTCCAGAGCCACGGAAAGGAATAAGAAGTGGGCATGTGCCTG
GGAGCAAATGTGTTCCTTTCCCTCAGTTGCTTGACAGTTCGCAAAAGCTATTACCTCCAGAAGAGCTCCGTAAACGATTTGAACAAGAAGGTATATCGCTTGATCAACCCCTGGTGACCTCATGTGGTACTGGTGTGAC
AGCATGTATATTAGCTCTGGGCCTCCACCGCCTCGGCAAAACCGATGTTCCTGTATATGATGGATCATGGACTGAATGGGGAGCCCATCCTGACACTCCAGTTGCCACTGCTGCTTAATTAGTACCAGTTACAATCTTT
TGGAGGACTCTTGATTATACTTCCCCTGCACCGTGAATGGTCCTATGGAAGGAAGAAAGCAAAGGAGATTGGAGAGGTCGGCTGGGCGTAGGTTGGGGGGAAGGGACGCCATTAGACTGGCCTTTTCGTTGTAAATGGT
TTTGAAAATAGACATACACTGTGAATTATTATGAGCTGTTGATCTACCTTTCTGAAGTCCCTTTTATATGATGGCCTGGGGCAATAATAAATGAGAACACTAATTAATACGCTGAAAGTGTCAATACCTGTATCCTTTT
TTGTTTGTGCCATTGTGTTCTTGTGTTATGCAACTAATGTTCATCATATTACCGGTCCAAAATCAGGTATGATGCTCCTGTAAACAATGTGTCAAAATCAATTCATTTCTGCAGTTAATTAATCTGATAAAAAACGTCA
TTGCTCAGGTGTCATTTCAGCAAGGAAGATGTCTCGGTCGTCCTGGACGACCGAGTAGTGTCTGTGCTGGCGCTGCTGCTTCCTGTCGCAACTGAAGATGATGTGTGGTTGAAAGCATCAAC**CGGCGGCAGGAACGTCG**
CCACCGGCATCCTCGCCGGGAGGCTCGGCGGCGGCCTCGCCCCTCAGCACGCCGATGGCCTGCCTGATCGTCGGCCGCAGGCTCCGGTCAGGGTGCGCGCACCAGAGCCCGACGACCATGACGGCCTCCATCTCCCC
GCCGTCGAACTCGCCGGTGAGCCGCCGGTCAGCGGCGTCAAGAATCCTCCCACTGCCGTACAAATTCCAGACCCACTGAGCAATGTGGATCCGGTCTTCGTCTACCTCGGATTGGTGGTGGTGGTCCGCCATTATTGGT
CGCCGGCCGCAGGCGATCTCAAGGAGGACGACGCCGAAGCTGTACACGTCGGACTCGGCGTTGGCCCTACCGGTGATCATGCACTCCGGGTCCATGTACCCCATCGTGCCGGCGAGCACCGTGGTGTGTGAGCCCCGGC
CATGGTCGACGAGCCTGGCGAGGCCGAAGTCGCCGAGCTTGGCGTTGAAGGCGGCGTCGAGCATGATGTTGCTCGGCTTGATGTCGCGGTGCACCACGCACTGCTCCCACTCCTCGTGCAGGTACAGCAGCGCCGAGCC
GATCCCGAGCACGATCTCGTGCCTGAGTGGCCACGGGAGCACGCCGGCGTTGGCGCTGTAGAGGTGGGTGTCGAGGCTGGCGTTGGGCATGAGCTCGTAGACGAGGAGGAGCTCGCCG