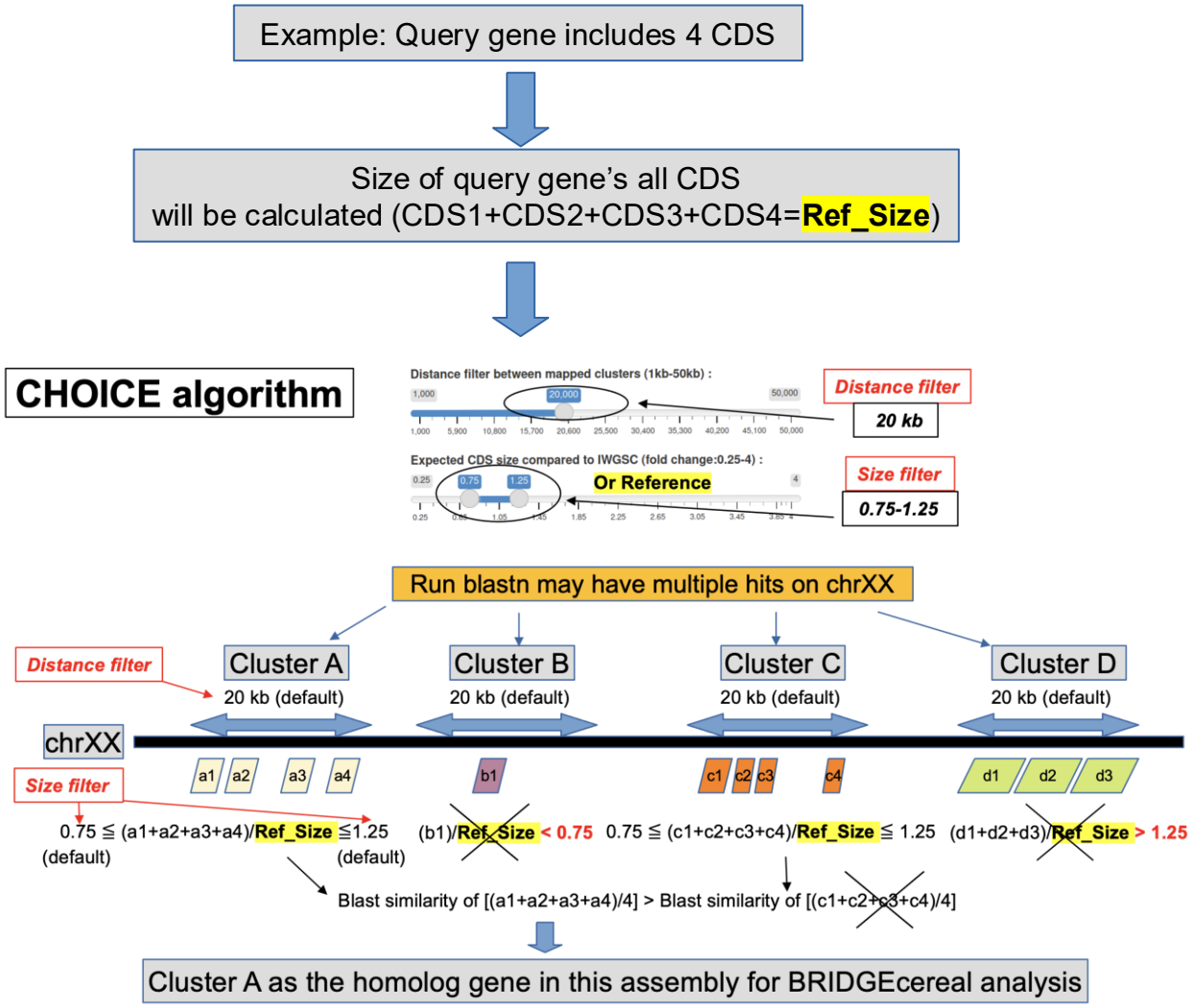
Two default CHOICE parameters explained:

Two parameters are applied as distance and size filters in the CHOICE algorithm to select homologs in pan-genomes. For more information and demo code, please visit (https://github.com/xianranli/CHOICE_CLIPS).



Submit a CDS:

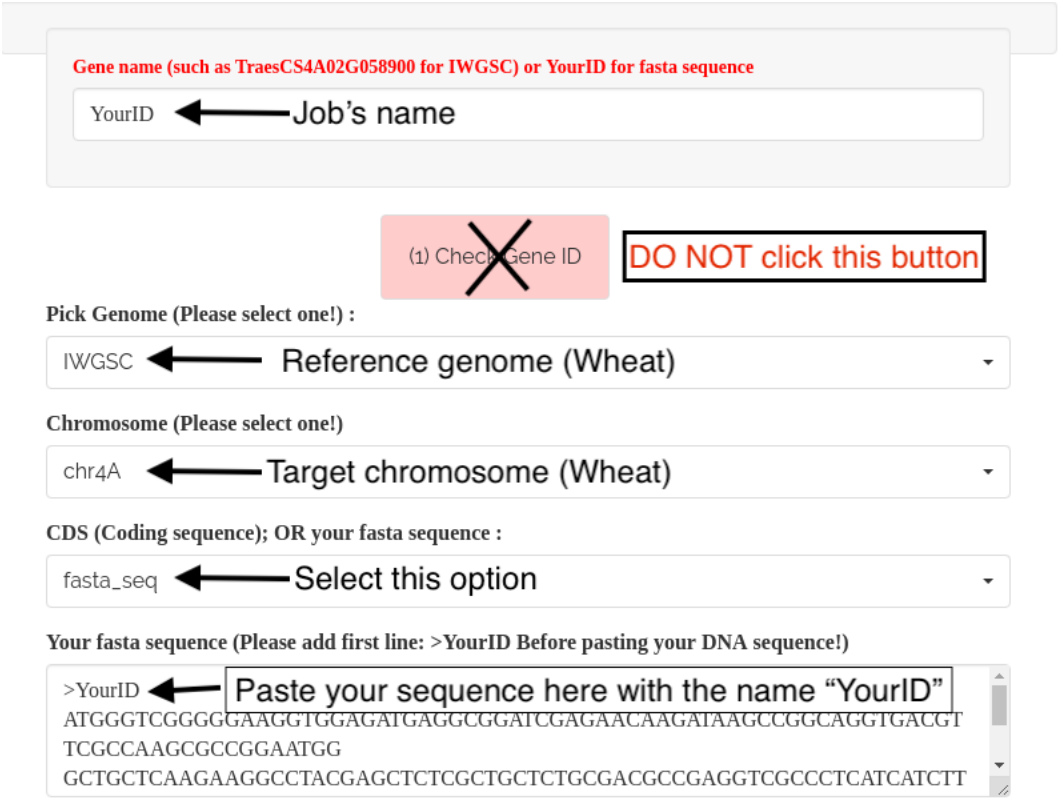
To blast a coding sequence against pan-genomes

1). Input your sequence ID name, such as YourID, in the Gene name box.

DO NOT click Button "(1) Check Gene ID"!

2). Manually select the Reference and Chromosome, select “fasta_seq” in the “CDS” box, then paste the sequence in the “Your fasta sequence” box. For your fasta sequence, the first ID line should be >YourID, the same name used in the Gene name box.

3). Click “(2) Submit” button to start the process. The query sequence together with other pan-genomes will be plotted in Panel 1.



Upload a contig or a chromosome:

To compare a sequenced chromosome (or large DNA fragment/contig) with pan-genomes on specific gene.

For a contig or a large fragment containing the gene of interest, change the first ID line of the fasta file as the corresponding chromosome. For example, a sequenced wheat contig/fragment contains a gene, which is known located at chromosome4A in the default reference genome IWGSC, the first fasta line should be as >chr4A.

```
>chr4A
ATGCATGC
```

For other crops (chromosome 4): Maize (>chr4); Sorghum (>chr4); Rice (>chr4); Barley (>chr4H).

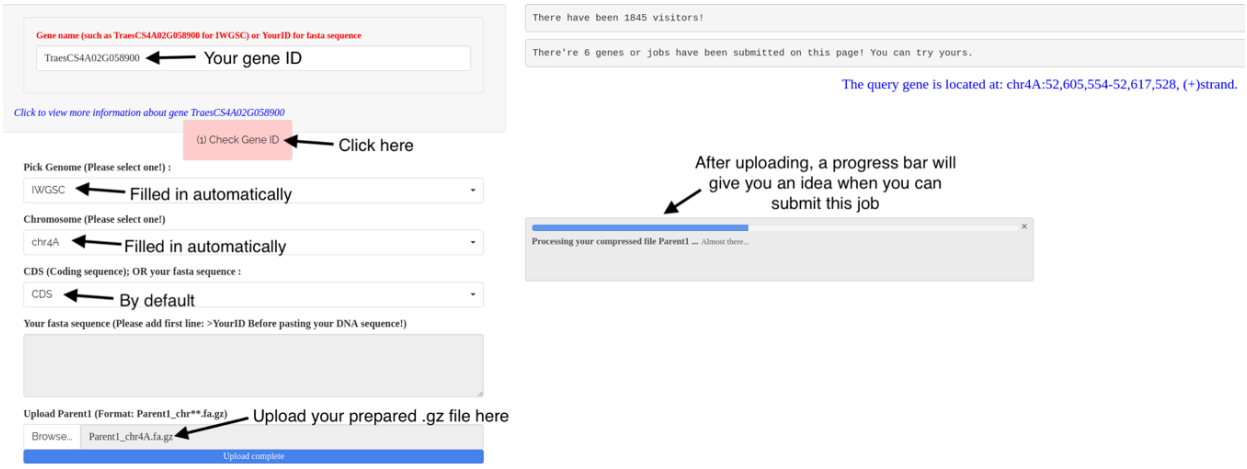
Rename this FASTA file as Parent1_chr4A.fa (the other as Parent2_chr4A.fa)

Then Run bgzip to compress the FASTA file. bgzip can be downloaded at (http://www.htslib.org/download/)

```
bgzip Parent1_chr4A.fa
```

1). Input the gene model ID in the Gene name box, then click "(1) Check Gene ID" button to automatically fill the Boxes for Reference and Chromosome.

2). To upload the compressed Parent1_chr4A.fa.gz file (max allowed file size = 300 MB), click "Browse" button to select and upload the formatted file (Parent1_chr4A.fa.gz).



3). When the progress bar (blue) showing "Upload complete", then click "(2) Submit (large file)" button (yellow color) to start the process. The "Parent1" genome together with other pan-genomes will be plotted in Panel 1.

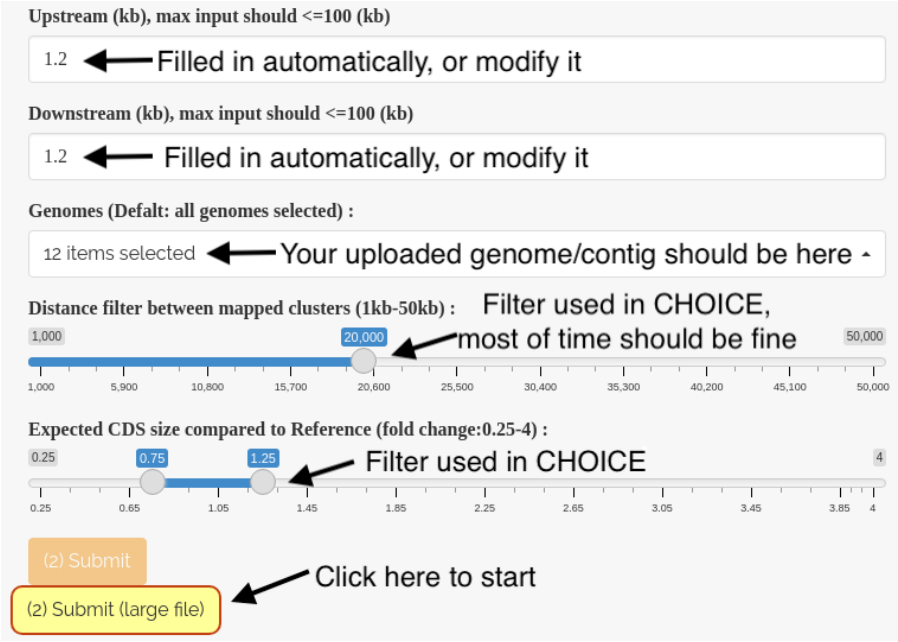
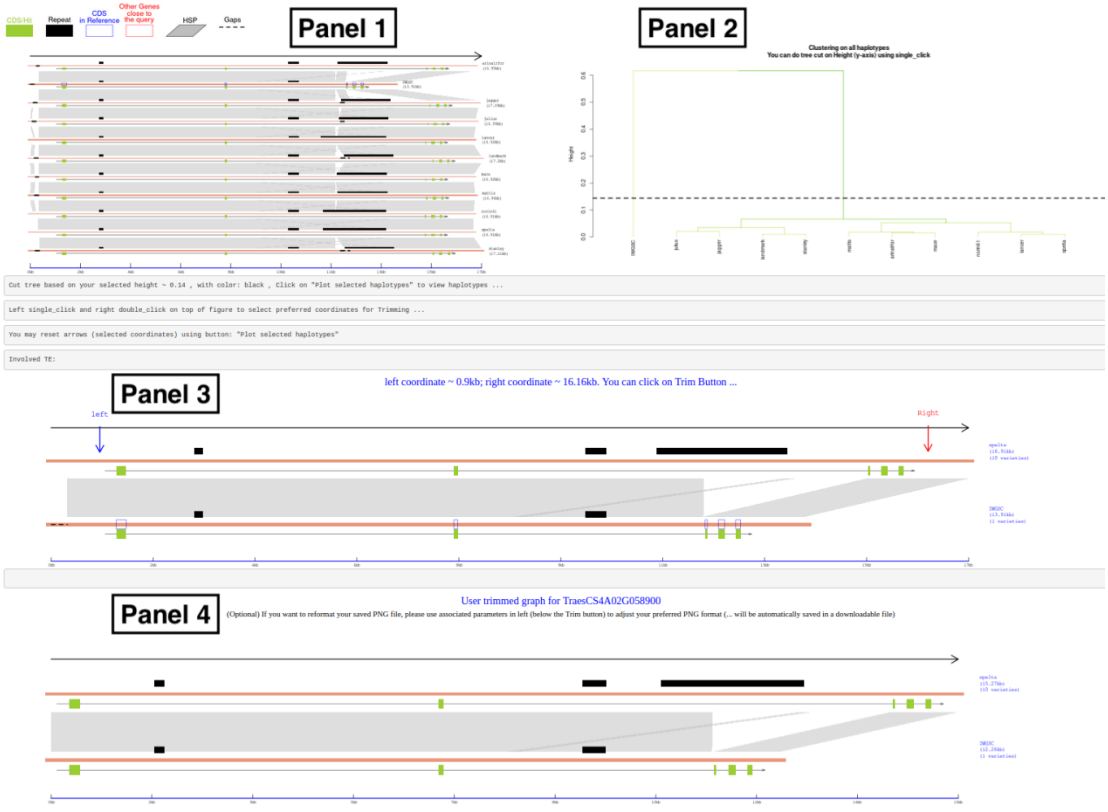


Figure layout:

An example from wheat:



1). Panel 1 (Output of Button (2)): Graphs of genes across all pan-genomes.

Outcome of CHOICE algorithm.

2). Panel 2 (Output of Button (3)): Tree/cluster.

Panel 2 is derived from Panel 1.

Outcome of CLIPS algorithm.

To interact with this tree, cut the tree on y-axis with a single click.

3). Panel 3 (Output of Button (4)): Graph of representative haplotypes.

Panel 3 is derived from Panel 2.

To trim the graph, single click at left (top of Panel 3), and double click at right (top of Panel 3).

4). Panel 4 (Output of Button (5)): Graph of trimmed figure.

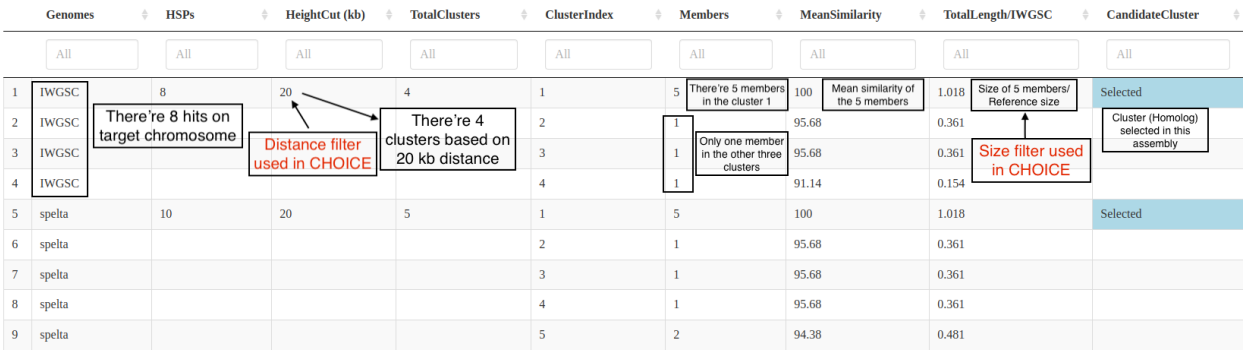
Panel 4 is derived from Panel 3.

Output tables:

1), Varieties in each group based on the tree-cut. This table is used for plotting Panel 3 and Panel 4.

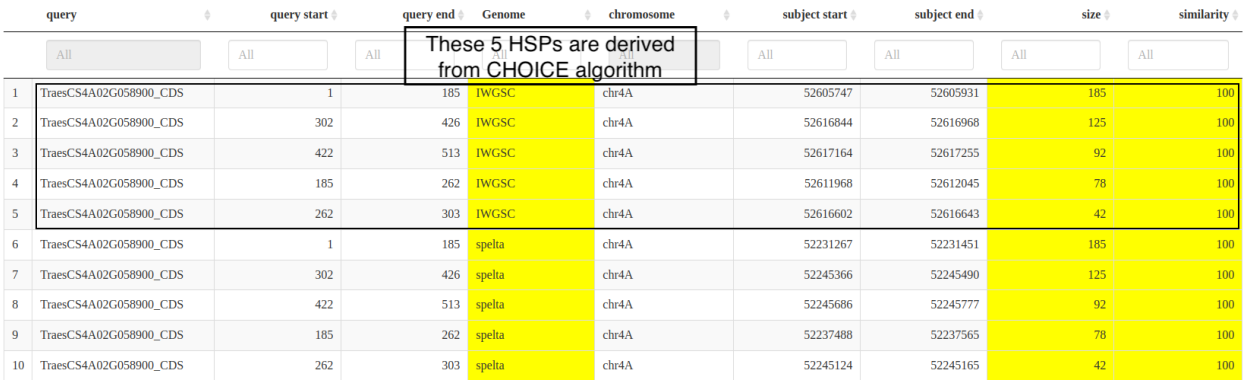
	Variety groups	Representative	Members	Varieties
	All	All	All	All
1	1	mattis	10	arinalrfor, jagger, julius, lancer, landmark, mace, mattis, norin61, spelta, stanley
2	2	IWGSC	1	IWGSC

2), CHOICE output about how to select homologs in assemblies.



	Genomes	HSPs	HeightCut (kb)	TotalClusters	ClusterIndex	Members	MeanSimilarity	TotalLength/IWGSC	CandidateCluster
	All	All	All	All	All	All	All	All	All
1	IWGSC	8	20	4	1	5	100	1.018	Selected
2	IWGSC				2	1	95.68	0.361	
3	IWGSC				3	1	95.68	0.361	
4	IWGSC				4	1	91.14	0.154	
5	spelta	10	20	5	1	5	100	1.018	Selected
6	spelta				2	1	95.68	0.361	
7	spelta				3	1	95.68	0.361	
8	spelta				4	1	95.68	0.361	
9	spelta				5	2	94.38	0.481	

3), Blast results used for plotting. This table is used for plotting Panel 1 to Panel 4.



	query	query start	query end	Genome	chromosome	subject start	subject end	size	similarity
	All	All	All	All	All	All	All	All	All
1	TraesCS4A02G058900_CDS	1	185	IWGSC	chr4A	52605747	52605931	185	100
2	TraesCS4A02G058900_CDS	302	426	IWGSC	chr4A	52616844	52616968	125	100
3	TraesCS4A02G058900_CDS	422	513	IWGSC	chr4A	52617164	52617255	92	100
4	TraesCS4A02G058900_CDS	185	262	IWGSC	chr4A	52611968	52612045	78	100
5	TraesCS4A02G058900_CDS	262	303	IWGSC	chr4A	52616602	52616643	42	100
6	TraesCS4A02G058900_CDS	1	185	spelta	chr4A	52231267	52231451	185	100
7	TraesCS4A02G058900_CDS	302	426	spelta	chr4A	52245366	52245490	125	100
8	TraesCS4A02G058900_CDS	422	513	spelta	chr4A	52245686	52245777	92	100
9	TraesCS4A02G058900_CDS	185	262	spelta	chr4A	52237488	52237565	78	100
10	TraesCS4A02G058900_CDS	262	303	spelta	chr4A	52245124	52245165	42	100

Downloadable files:

All files, except the GeneID.png, are text files, which can be opened and viewed by any text editor, such as Notepad.

Output files in the compressed .zip	Details
GeneID.png	Final trimmed figure (Panel 4) in .png format (600 dpi).
GeneID_Blast_Original	Blast output (with –outfmt 6) using the CDSs search against the target chromosome of each assembly. (Blast of query gene's CDS to all assemblies)

GeneID_Haplotype_syn	Output from CHOICE. (Derived from GeneID_Blast_Original file) HSPs selected as homolog in each assembly.
GeneID_ref_CDS-Haplotype_out_m8	Blast output of CHOICE selected coding sequences in all assemblies. Used for plotting the HSPs of CDS in each assembly (green box)
GeneID_repMask2	Repeats found in selected segments. Used for plotting repeats (black box).
GeneID_Haplotype_N_Gaps	Gaps found in selected segments. Used for plotting gaps (dashed line).
GeneID_Haplotype-Self_out_m8	Blast output of genomic DNA in pan-genomes. (Pairwise) Used for plotting HSPs (grey polygon).
GeneID_CDS.fa	Query gene's coding sequence.
GeneID_Haplotype.fa	Genomic sequences extracted from all assemblies in selected segments.
GeneID_User_Selected.fa (optional)	Output of DNA sequences based on selected haplotypes and trimmed coordinates.