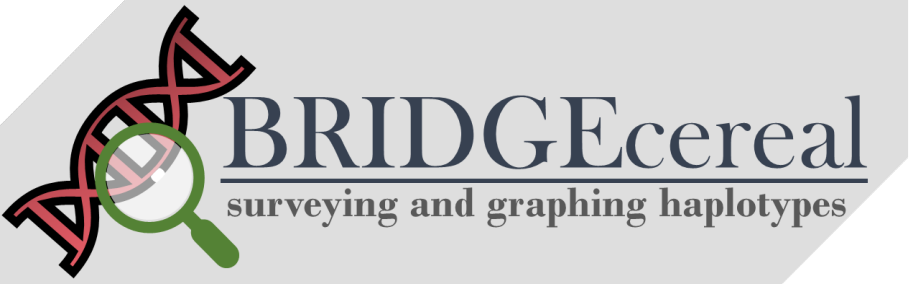


BRIDGEcereal streamlines unsupervised learning to graph indel-based haplotype from pan-genome*

Xianran Li

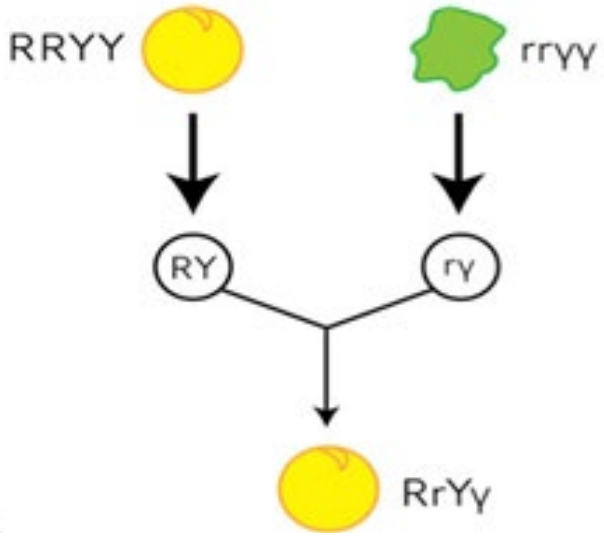
USDA-ARS Wheat Health, Genetics, and Quality

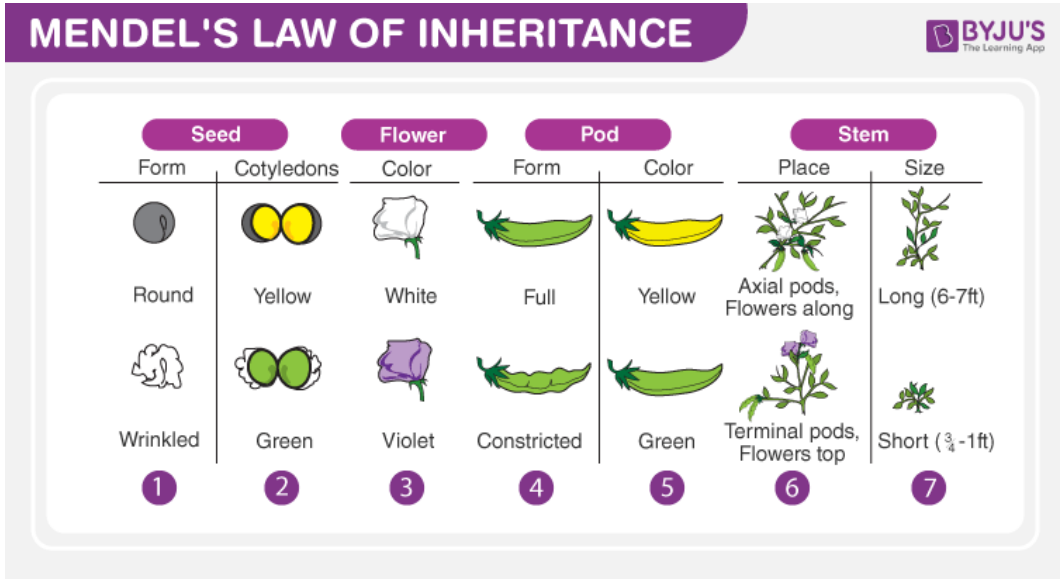
Pullman, WA

https://compbiolab.org/

*How to find large indel polymorphisms for your favorite genes from available pan-genome

Genetics: phenotype & casual DNA polymorphisms





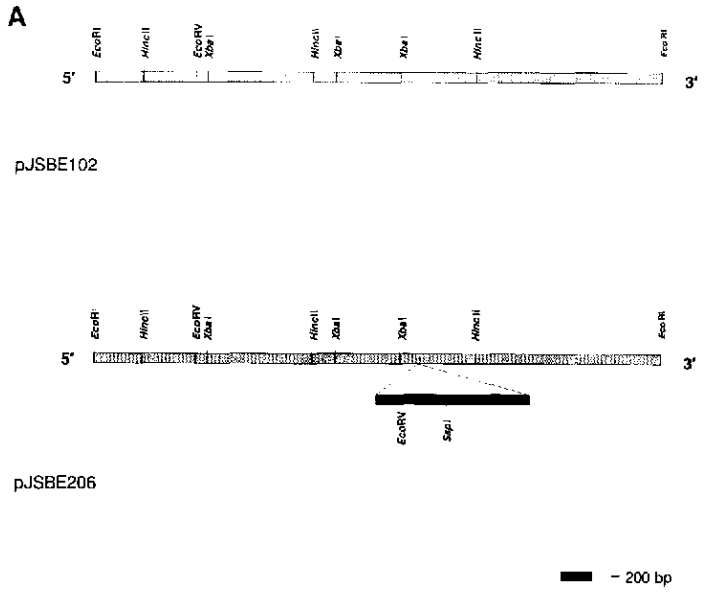
Cell Supports open access

The first cloned Mendel's gene segregates a large insertion (0.8-kb)!

ARTICLE | VOLUME 60, ISSUE 1, P115-122, JANUARY 12, 1990 Download Full Issue

The wrinkled-seed character of pea described by Mendel is caused by a transposon-like insertion in a gene encoding starch-branching enzyme

Madan K. Bhattacharyya • Alison M. Smith • T.H.Noel Ellis • Cliff Hedley • Cathie Martin



Two decades of cereal genome sequencing: from single reference to pan-genome

Drafted 2002
Finished 2005
Drafted 2009
Drafted 2009
Drafted 2012
Drafted 2014
Improved 2017
Improved 2017
Improved 2018
Improved 2018
Pan-genome 2020
Pan-genome 2020
Pan-genome 2021
Pan-genome 2021

Rice
Maize
Wheat
Barley
Sorghum

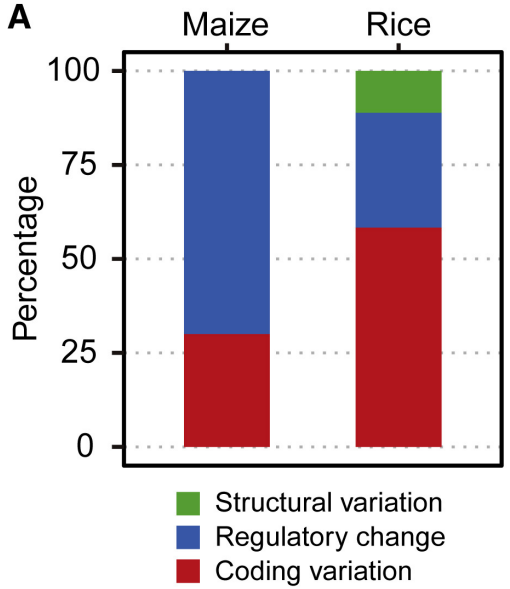
REVIEW

How the pan-genome is changing crop genomics and improvement!!!

Rafael Della Coletta[1], Yinjie Qiu[1], Shujun Ou[2], Matthew B. Hufford[2*] and Candice N. Hirsch[1*]

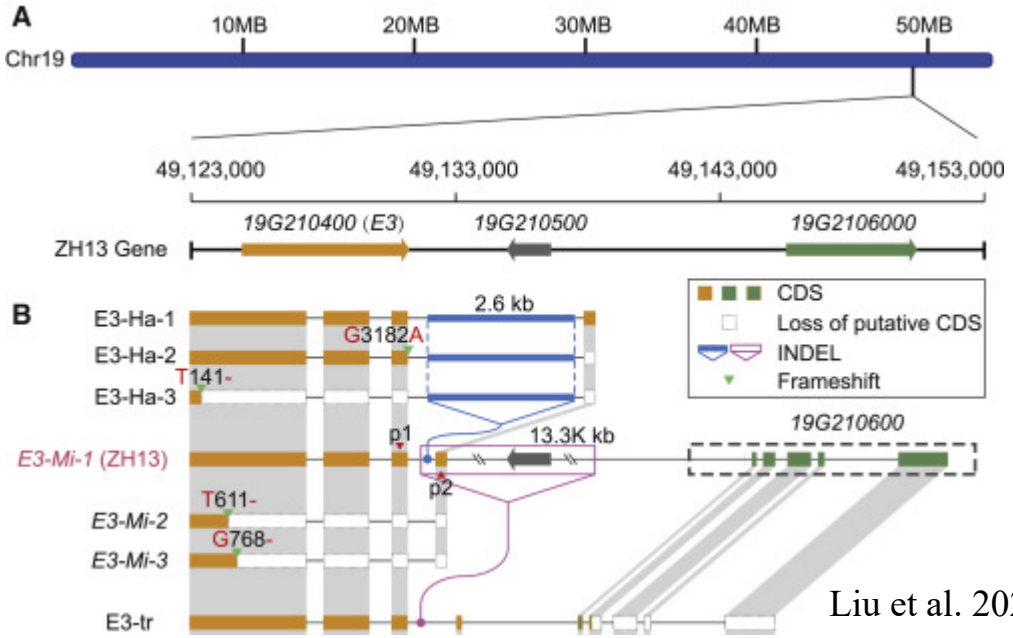
Advancing QTL mapping and GWAS using pan-genomes

Half of identified casual polymorphisms are large indels.



Chen et al. 2021

Haplotypes of the *E3* gene because of indels



Liu et al. 2020

How the pan-genome is changing crop genomics and improvement?

is

How to leverage pan-genomes for your genes?

ChatGPT

Li how to identify potential indels for a gene from pan-genomes

To identify potential indels for a gene from pan-genomes, you can follow these general steps:

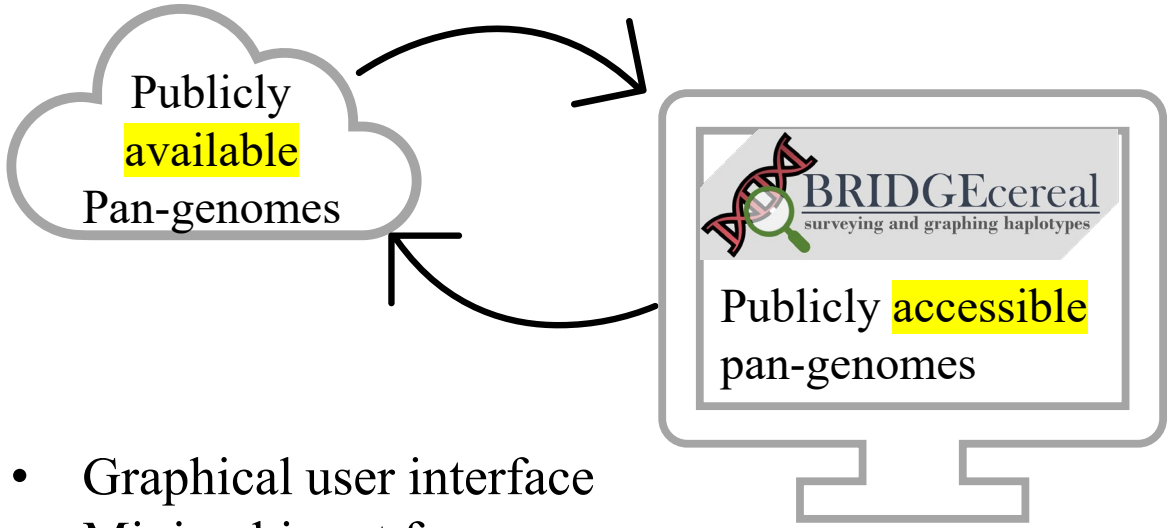
1. Obtain the pan-genome data: This can be obtained from various sources, such as public databases or through sequencing and assembling the genomes of multiple individuals or strains of the same species.
2. Identify the gene of interest: Use bioinformatics tools to identify the gene of interest in the pan-genome data. This can be done by comparing the gene sequence to the pan-genome data using alignment tools such as BLAST or MUMmer.
3. Compare the gene sequence across the pan-genomes: Compare the gene sequence across the pan-genomes to identify any insertions or deletions (indels). You can use alignment tools such as MAFFT or Clustal Omega to align the gene sequences across the pan-genomes.
4. Identify potential indels: Look for gaps or differences in the alignment that indicate potential indels. These can be large or small and can be insertions or deletions relative to the reference sequence.
5. Validate potential indels: Once you have identified potential indels, you can validate them using various methods such as PCR, Sanger sequencing, or high-throughput sequencing techniques.

It's important to note that the analysis of pan-genomes can be complex and requires specialized expertise in genomics and bioinformatics. Additionally, the accuracy of the

BRIDGEcereal: new bioinformatic bridge for identifying large indels

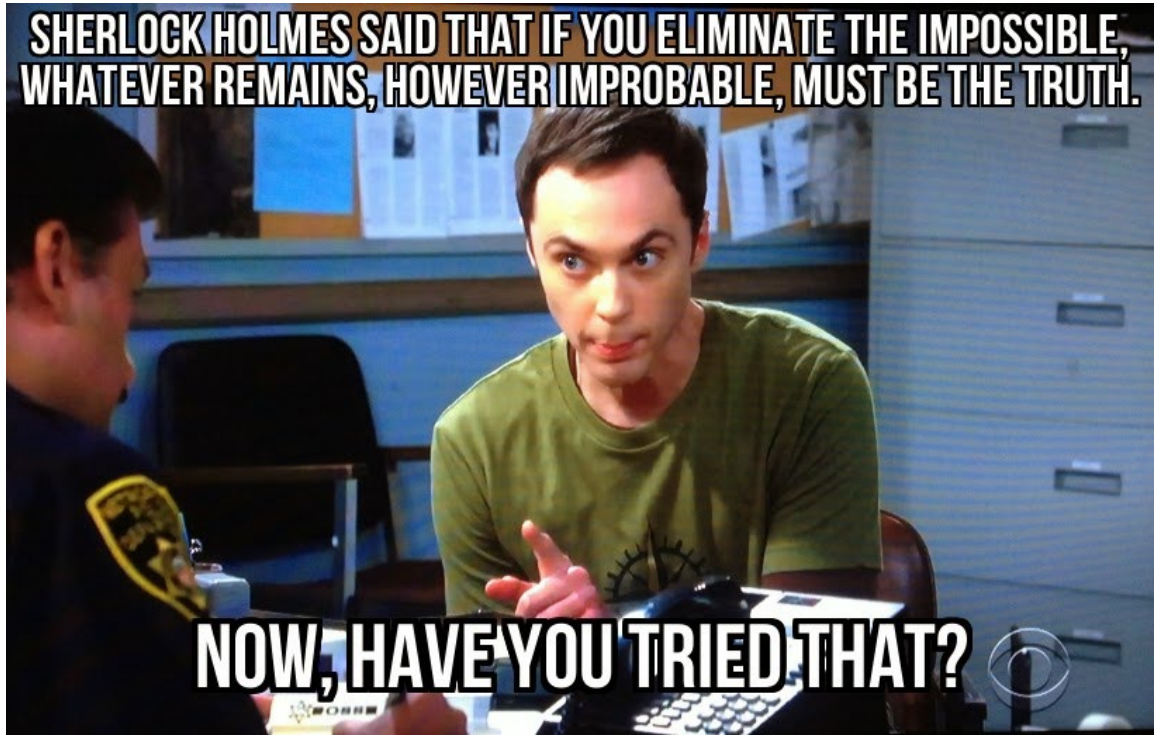
Bosen Zhang

Laura Tibbs-Cortes

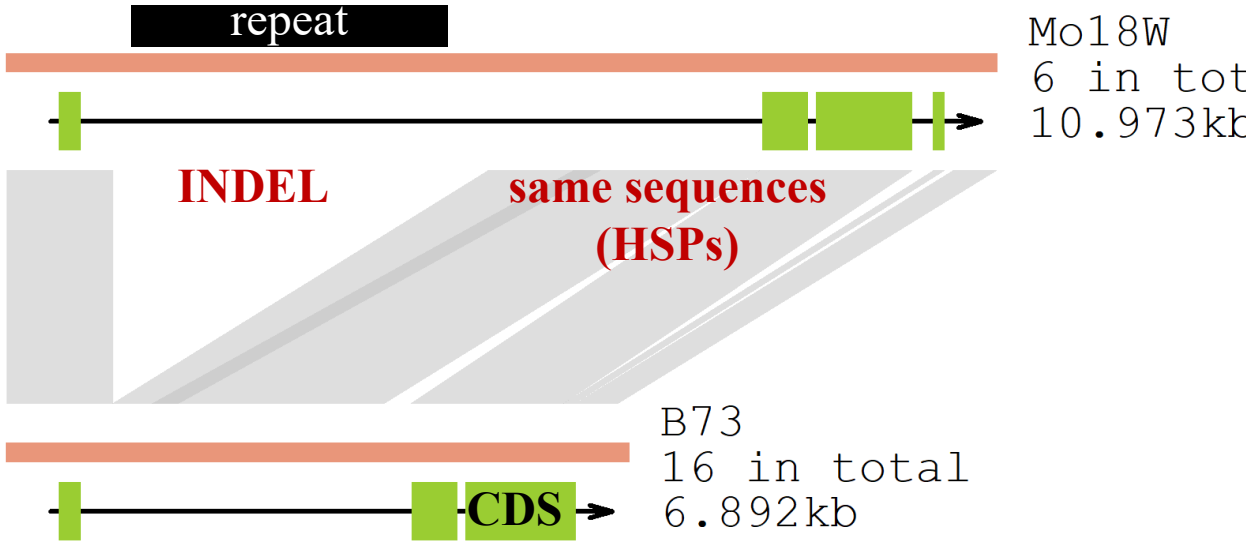


- Graphical user interface
- Minimal input for users
- Adjustable parameters

Blueprint of BRIDGEcereal



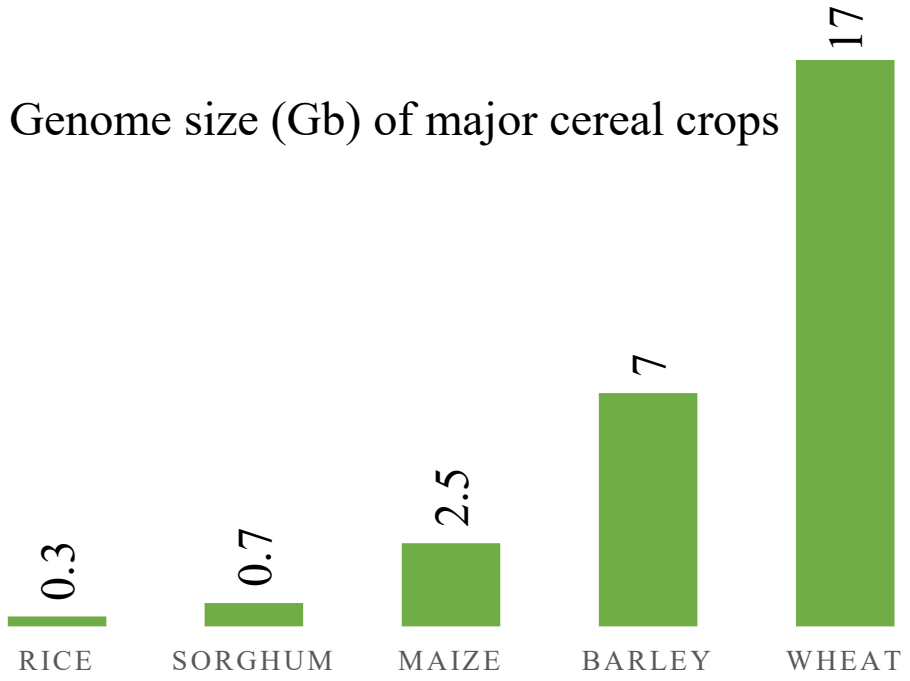
"If you eliminate same sequences, whatever remains, however improbable, must be INDELs"



Same sequences = HSP in BLAST

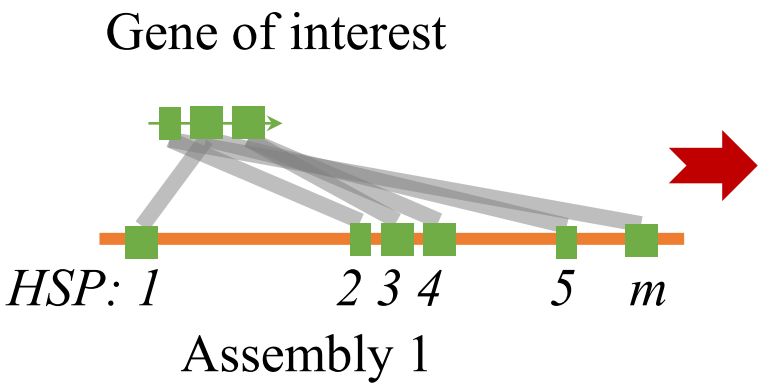
HSP: High-scoring Segment Pairs

Challenge 1: where is the homolog of a gene of interest in each assembly?



Gene annotation is typically only available for the first reference.

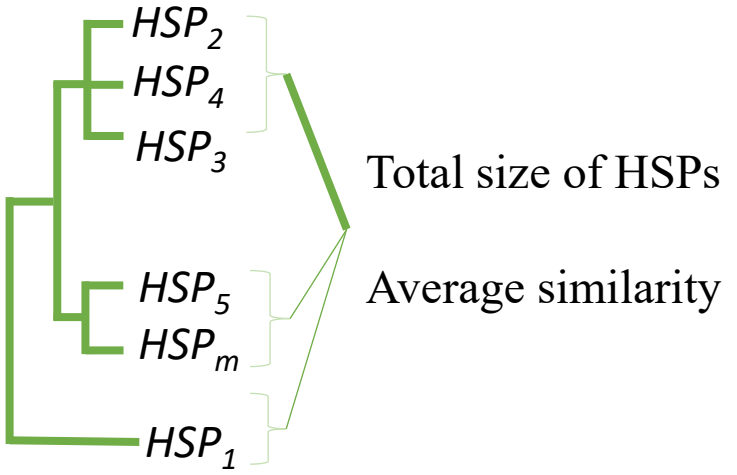
CHOICE to identify and extract segments harboring the ortholog from each assembly



Distance matrix

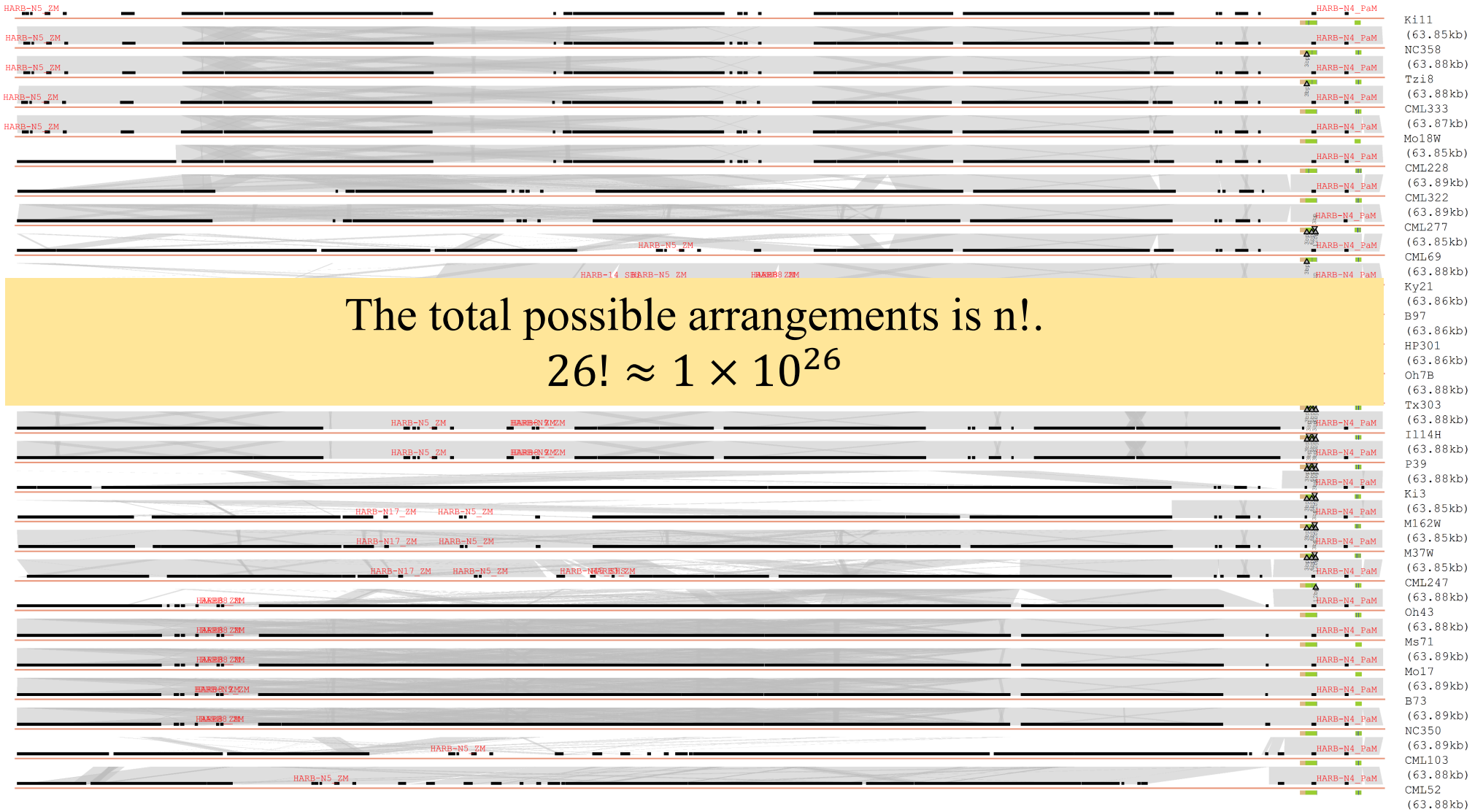
	HSP_1	HSP_2	...	HSP_m
HSP_1	0	$\lvert s_1 - s_2 \rvert$		$\lvert s_1 - s_m \rvert$
HSP_2	$\lvert s_2 - s_1 \rvert$	0		$\lvert s_2 - s_m \rvert$
...				
HSP_m	$\lvert s_m - s_1 \rvert$	$\lvert s_m - s_2 \rvert$		0

s_m: start coordinate of the m^{th} *HSP*



Clustering HSPs for Ortholog Identification
via Coordinates and Equivalence

Challenge 2: Complex and indiscernible pattern from a large number assemblies

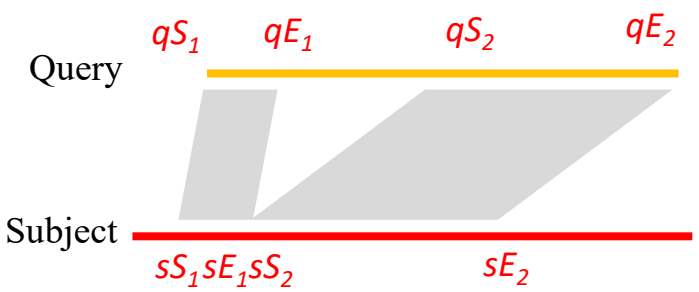


The total possible arrangements is n!.

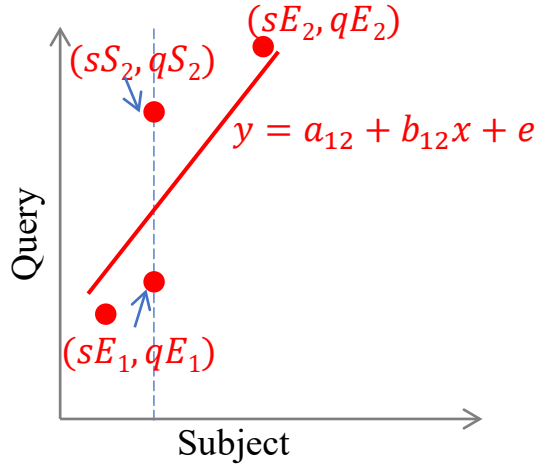
$26! \approx 1 \times 10^{26}$

26 assemblies

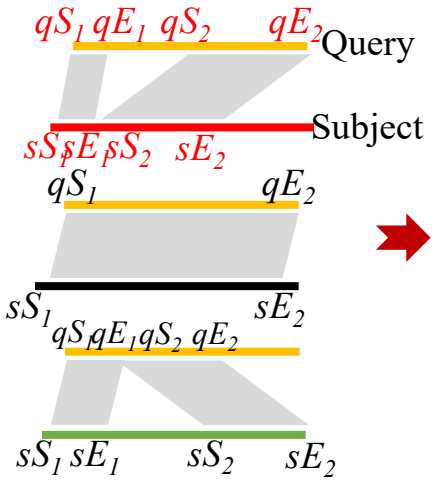
CLIPS: Clustering via Large Indel Permuted Slopes

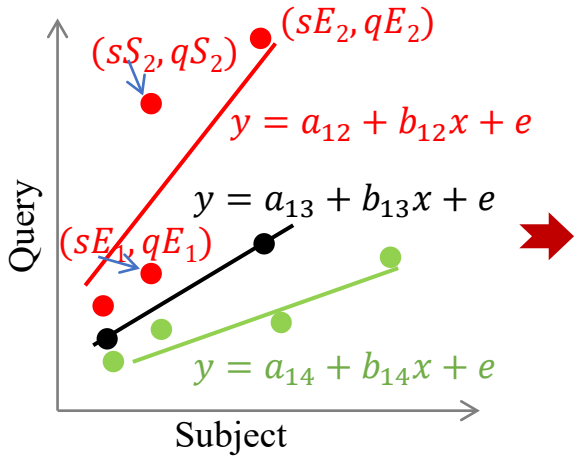


Query	$y = (qS_1, qE_1, qS_2, qE_2 \ldots qS_n, qE_n)$
Subject	$x = (sS_1, sE_1, sS_2, sE_2 \ldots sS_n, sE_n)$



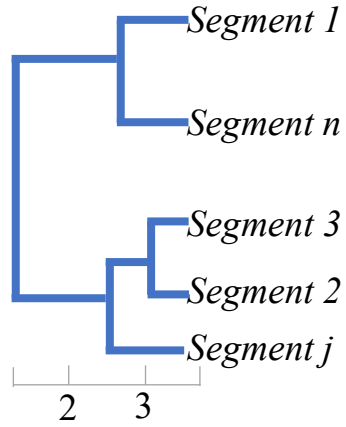
CLIPS: Clustering via Large Indel Permuted Slopes



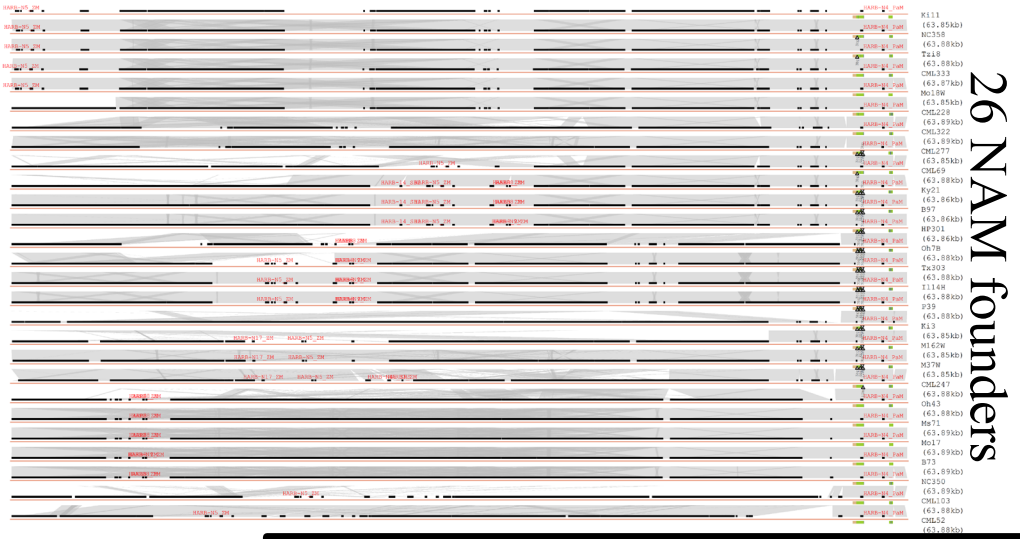


Slope matrix

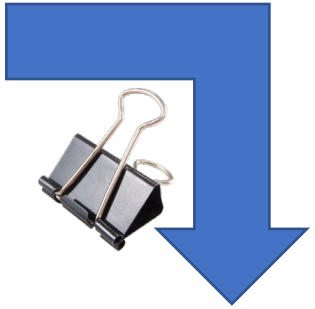
	Segment 1	Segment 2	…	Segment n
Segment 1	1	b_{12}		b_{1n}
Segment 2	b_{21}	1		b_{2n}
…				
Segment n	b_{n1}	b_{n2}		1



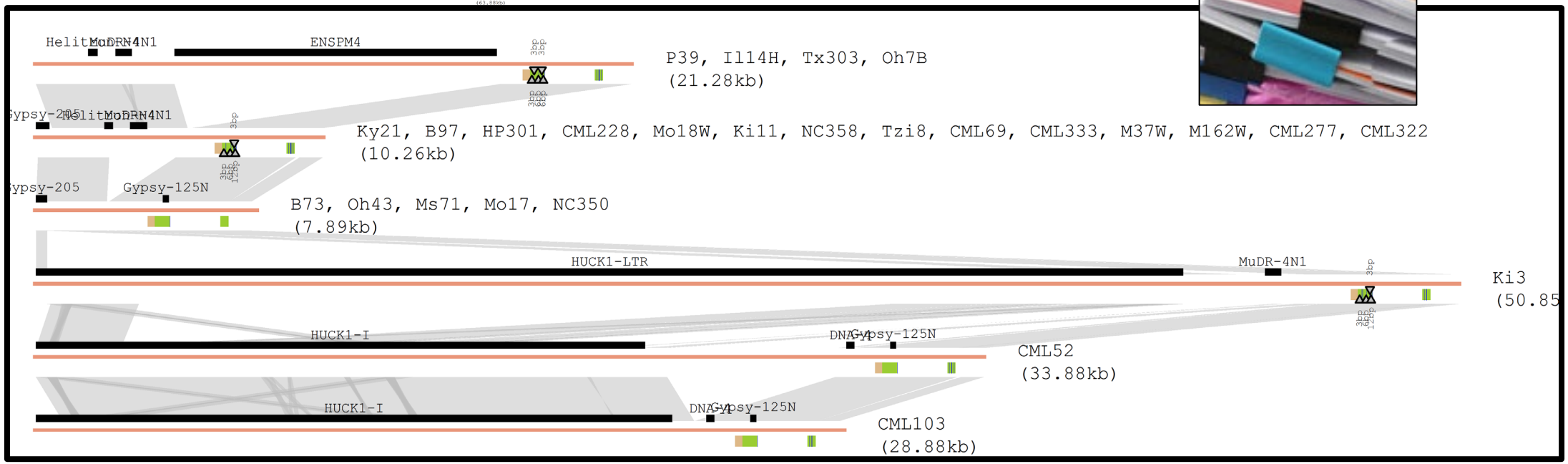
CLIPS clamps a complex gene pile into a concise and legible haplotype graph



26 NAM founders

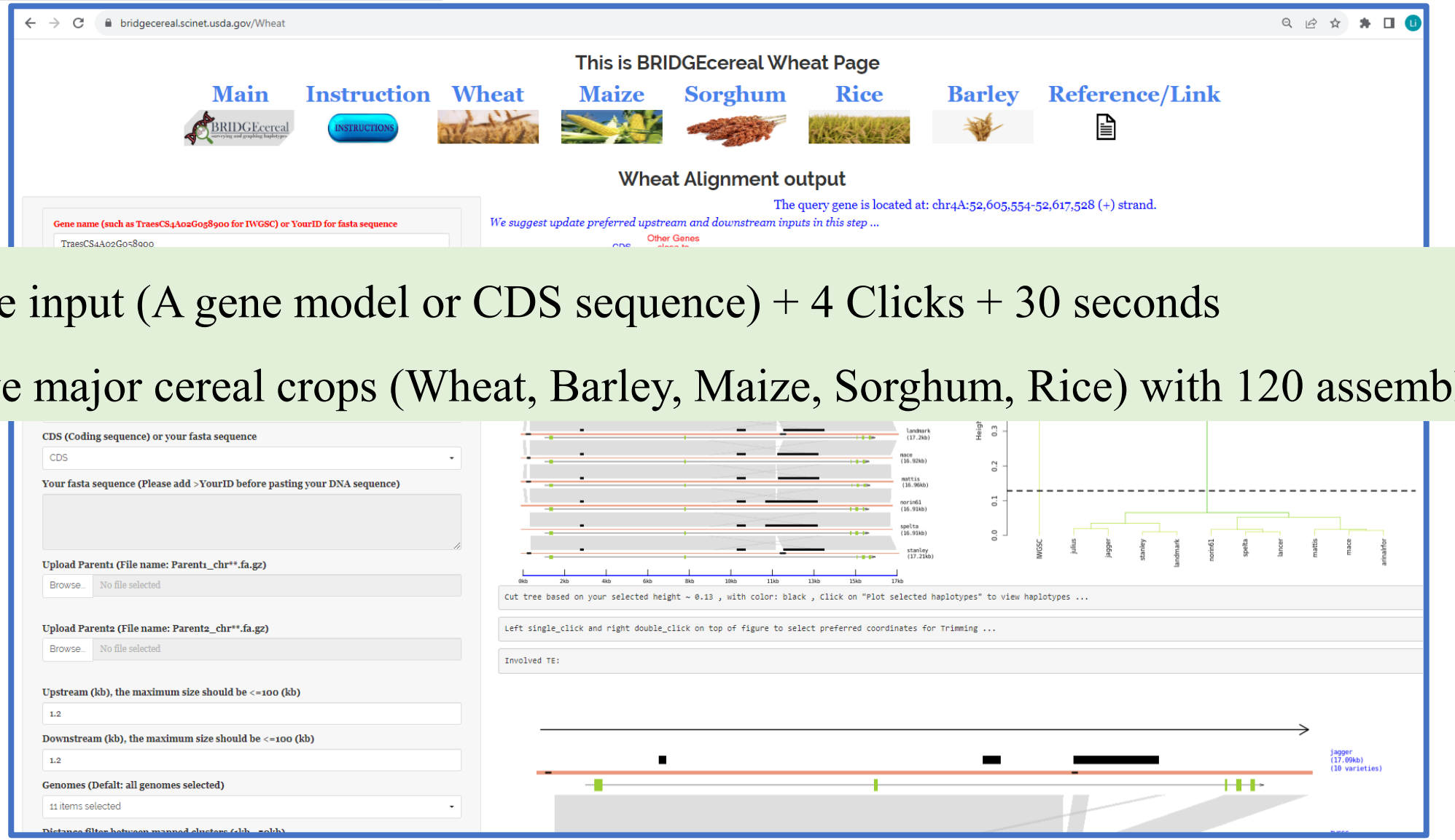


6 haplotypes



BRIDGEcereal
surveying and graphing haplotypes

https://bridgecereal.scinet.usda.gov/



- One input (A gene model or CDS sequence) + 4 Clicks + 30 seconds
- Five major cereal crops (Wheat, Barley, Maize, Sorghum, Rice) with 120 assemblies

Demonstration on BRIDGEcereal facilitating gene identification and characterization

Population	Gene	Prior information		BRIDGEcereal results
		Gene	Causal site	
Chinese Spring × Paragon RIL	*Rc-D1*	√	√	New recessive alleles
	B1	√	×	A 13-kb deletion 6-kb upstream
	Hooded	×	×	3-kb deletion removing 3 exons as a potential causal site

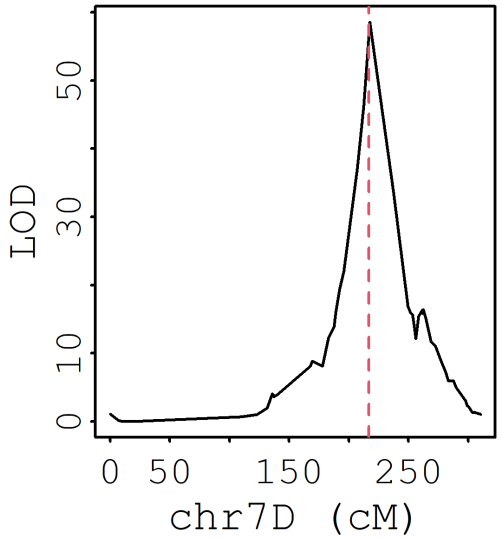
Wheat Chinese Spring × Paragon RIL population:

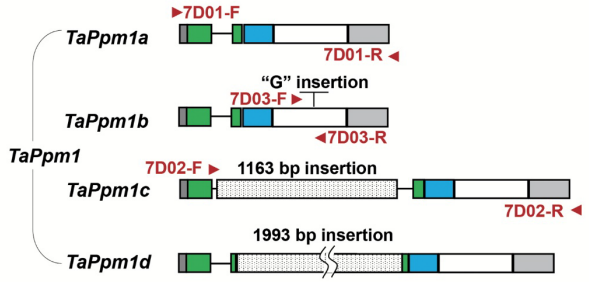
- Chinese Spring is the reference genome.
- Paragon was assembled at the scaffold level.

New recessive alleles of the well characterized wheat *Rc-D1* gene

Prior information	
Gene	Causal site
√	√

Trait → QTL

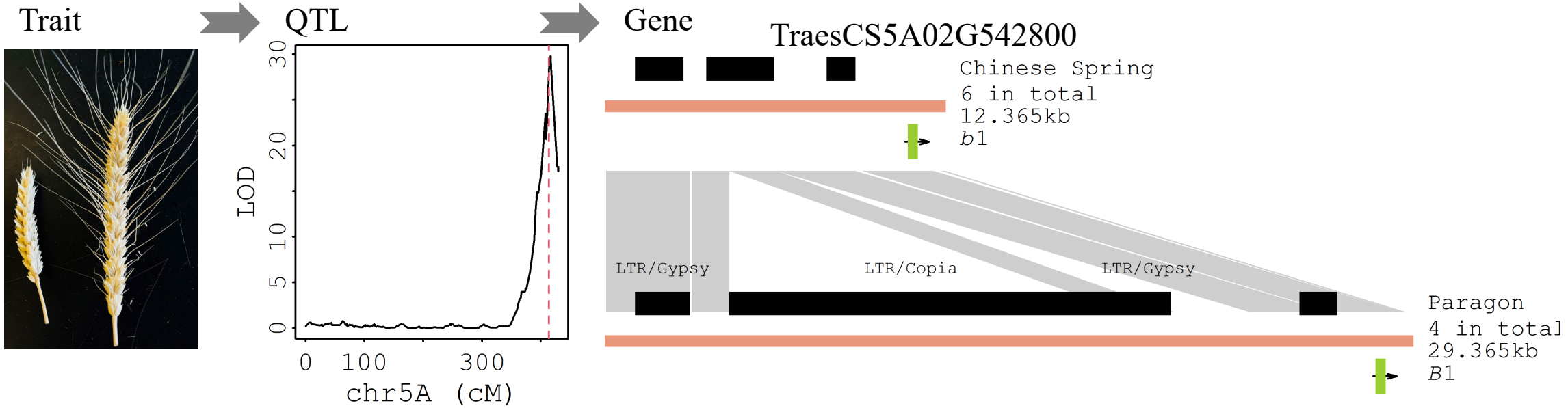




Jiang et al. 2018

Two large indels (13-kb + 4-kb) in 6-kb upstream as potential causal site for *B1*

Prior information	
Gene	Causal site
√	×



- In 2020, three papers reported TraesCS5A02G542800 as the underlying gene with the consensus of an unknown casual polymorphism outside of gene altering its expression.

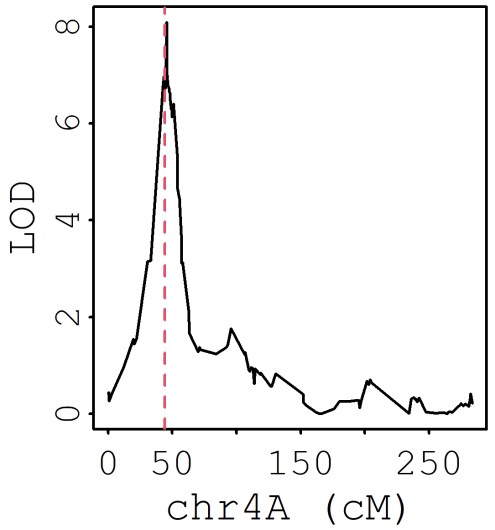
TaDL is less likely as candidate for the classic wheat *Hooded* QTL

Prior information	
Gene	Causal site
×	×

Trait

Chinese Spring × Paragon

QTL



TraesCS4A02G058800 (*TaDL*)

```
TraesPAR_scaffold_035072_01G00      MQSMDLVSPSEHLCYVRCTYCNTVLAVGVPCKRLMDTVTVKCGHCNNLSF
TraesCS4A02G058800                  MQSMDLVSPSEHLCYVRCTYCNTVLAVGVPCKRLMDTVTVKCGHCNNLSF
                                    **************************************************

TraesPAR_scaffold_035072_01G00      LSPRPPPMVQPLSPNDHHHPMGPFQGWTDCRRNQPLPPLASPTSSDASPR
TraesCS4A02G058800                  LSPRPPPMVQPLSPNDHHHPMGPFQGCTDCRRNQPLPPLASPTSSDASPR
                                    ************************** ********************

TraesPAR_scaffold_035072_01G00      APFVVKPPEKKHRLPSAYNRFMREEIQRIKAAKPDIPHREAFSMAAKNWA
TraesCS4A02G058800                  APFVVKPPEKKHRLPSAYNRFMREEIQRIKAAKPDIPHREAFSMAAKNWA
                                    **************************************************

TraesPAR_scaffold_035072_01G00      KCDPRCSTTVSASNSAPEPRIVVPGPQLQERATEQVVESFDIFKQMERSA
TraesCS4A02G058800                  KCDPRCSSTVSASNSAPEPRIVVPGPQLQERATEQVVESFDIFKQMERSA
                                    *******:******************************************
```

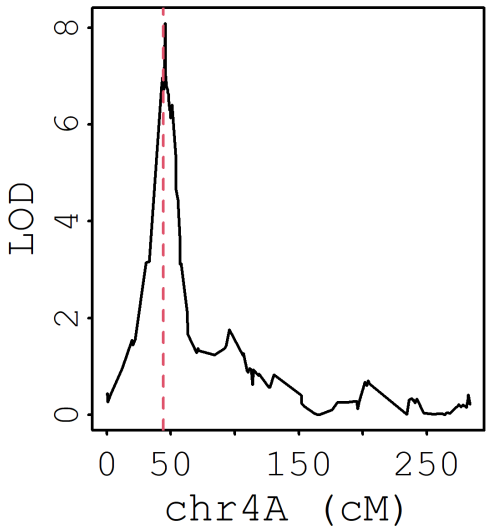
Two parental lines have not obvious causal functional polymorphisms

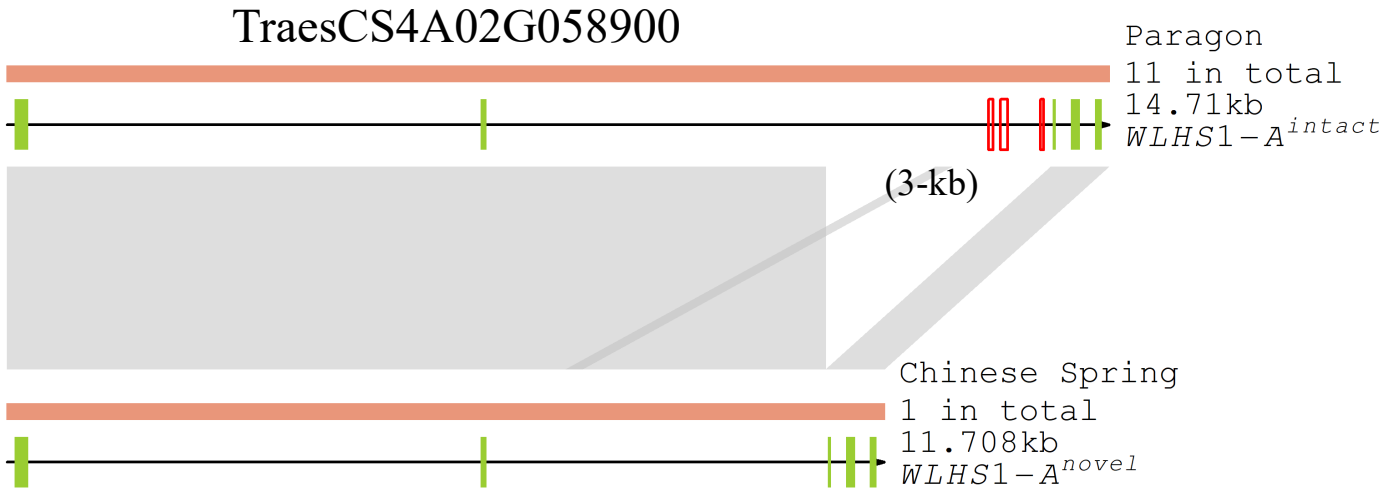
A *MADS-box* gene more likely underlying the classic wheat *Hooded* QTL

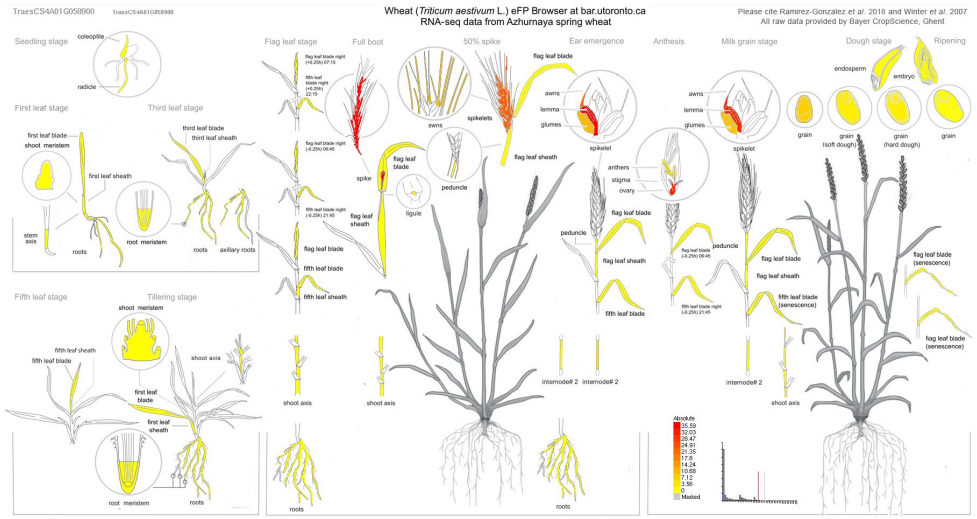
Prior information	
Gene	Causal site
×	×

Trait → QTL → Gene

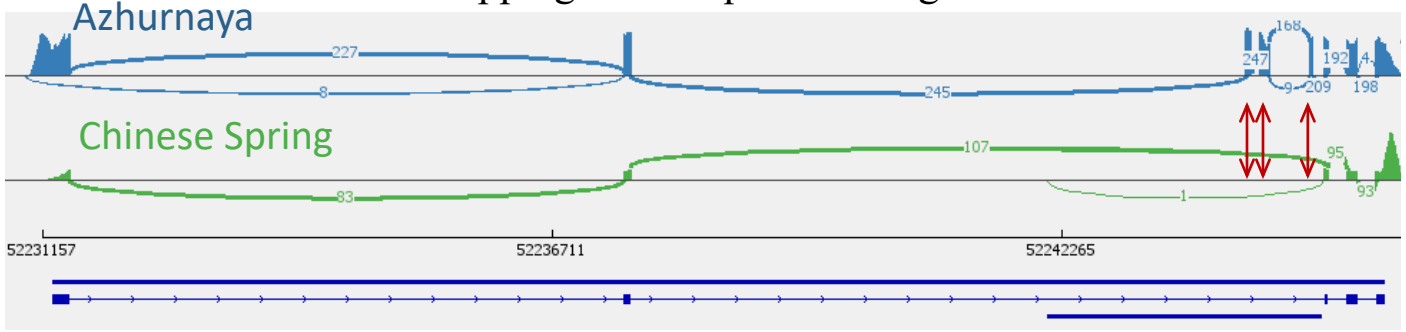
Chinese Spring × Paragon







Mapping RNAseq reads to a genome without the deletion



BRIDGEcereal
surveying and graphing haplotypes

https://bridgecereal.scinet.usda.gov/

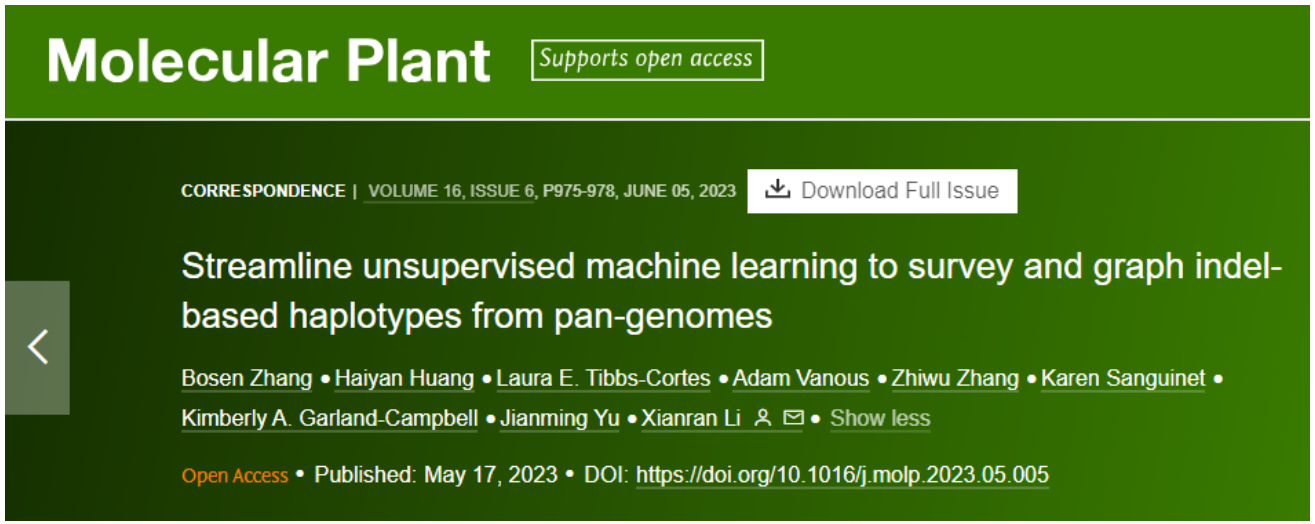
① Search “BRIDGEcereal” :

Google

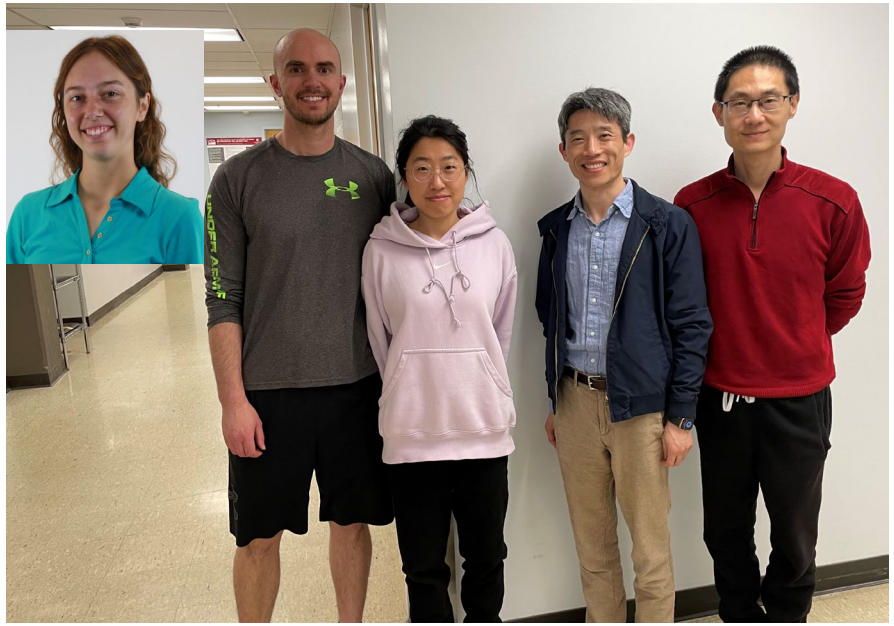
Bing

yahoo!

②



Acknowledgements



USDA Agricultural Research Service

Compbio Lab

Bosen Zhang

Laura Tibbs-Cortes

Ryan Benke

Linqian Han

Haiyan Huang

Adam Vanous

Kimberly Garland-Campbell

Iowa State University
Department of Agronomy

Jianming Yu

Washington State University

Zhiwu Zhang

Karen Sanguinet

Germplasm Resources Unit
..... a national capability supported by the BBSRC at the John Innes Centre

Chinese Spring × Paragon RILs

USDA Agricultural Research Service
U.S. DEPARTMENT OF AGRICULTURE

SCINet
Scientific Computing Initiative

Iowa State University
CROP BIOENGINEERING CENTER