HIGH-FIDELITY LONG-READ SEQUENCING ENABLES RAPID DETECTION OF STRUCTURAL AND COPY NUMBER VARIANTS: A CASE STUDY IN SOFT WINTER WHEAT

DANIELA MILLER, PHD CANDIDATE, NORTH CAROLINA STATE UNIVERSITY

BROWN-GUEDIRA LAB, NCSU, USDA-ARS EASTERN REGIONAL SMALL GRAINS GENOTYPING LABORATORY (ERSGGL)

HULSE-KEMP LAB, NCSU, USDA-ARS GENOMICS AND BIOINFORMATICS RESEARCH UNIT (GBRU)

SATURDAY 14 JANUARY 2023 **#PAG30**

INTERNATIONAL WHEAT GENOME SEQUENCING CONSORTIUM (IWGSC) WORKSHOP

Small Grains Genotyping

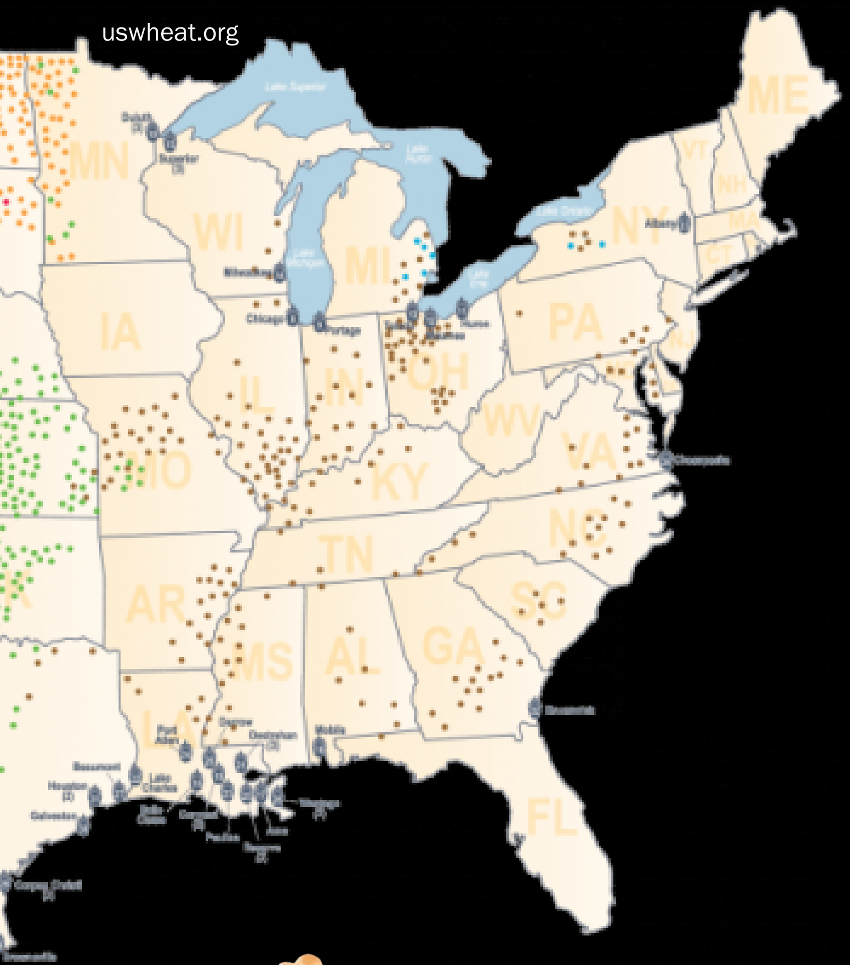
OUTLINE

1. Introduction
2. HiFi Soft Winter Wheat Genome Assemblies
3. Copy Number Variant (CNV) Detection: *VERNALIZATION-A1* Gene
4. Structural Variant (SV) Detection: 5B/5G Introgression

INTRODUCTION

SOFT WINTER WHEAT GENOME ASSEMBLY

SOFT WINTER WHEAT GERMPLASM IS NOT YET REPRESENTED IN CURRENT WHEAT ASSEMBLIES

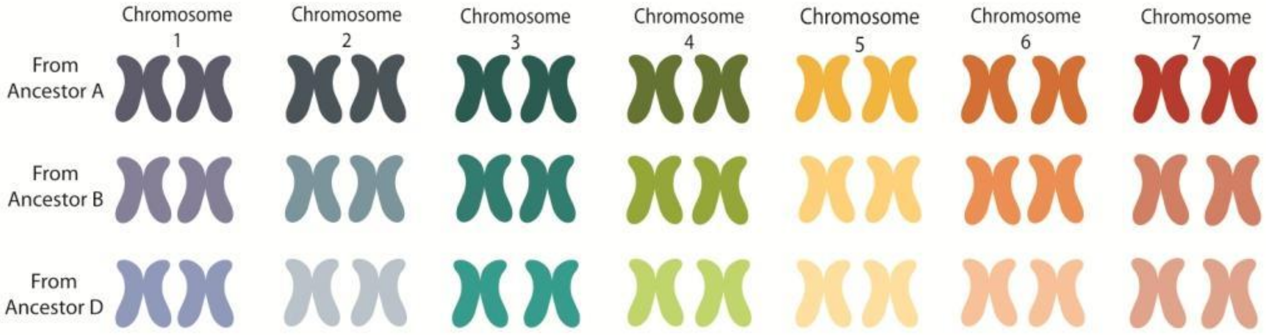


uswheat.org

- Of the existing wheat assemblies, only ‘Jagger’ is a US winter wheat cultivar.
- Soft Winter Wheat (SWW) is the most common market class in eastern US.
  - Soft wheat is used for crackers & cookies.
  - Winter wheat is sown in autumn and harvested in spring.
- Unique regional germplasm:
  1. ‘AGS2000’ is representative of SE US regional SWW germplasm, is well-adapted to warmer climates, and has stem rust (Ug99) resistance (see poster #46883).
  2. ‘Hilliard’ is a broadly-adapted SWW cultivar with notable Fusarium head blight (FHB) resistance.

WHEAT GENOME ASSEMBLY REMAINS CHALLENGING AMIDST NEXT GENERATION SEQUENCING (NGS) BOON

The 15 gigabase hexaploid wheat genome (2n = 6x = 42, AABBDD) is 80% repetitive with large complex repeat structures.



coloradowheat.org

SHORT-READ SEQUENCING (NGS)

- Relatively low error rates
- Repeat sequences longer than read lengths (i.e. > 600 bp) cannot be resolved
- Minor errors still cause mis-assembly between highly homo(eo)logous regions

LONG-READ SEQUENCING (ONT, PacBio CLR)

- Long reads can span large repeats
- High error rates hamper assembly process

HIGH-FIDELITY (HIFI) LONG-READ SEQUENCING

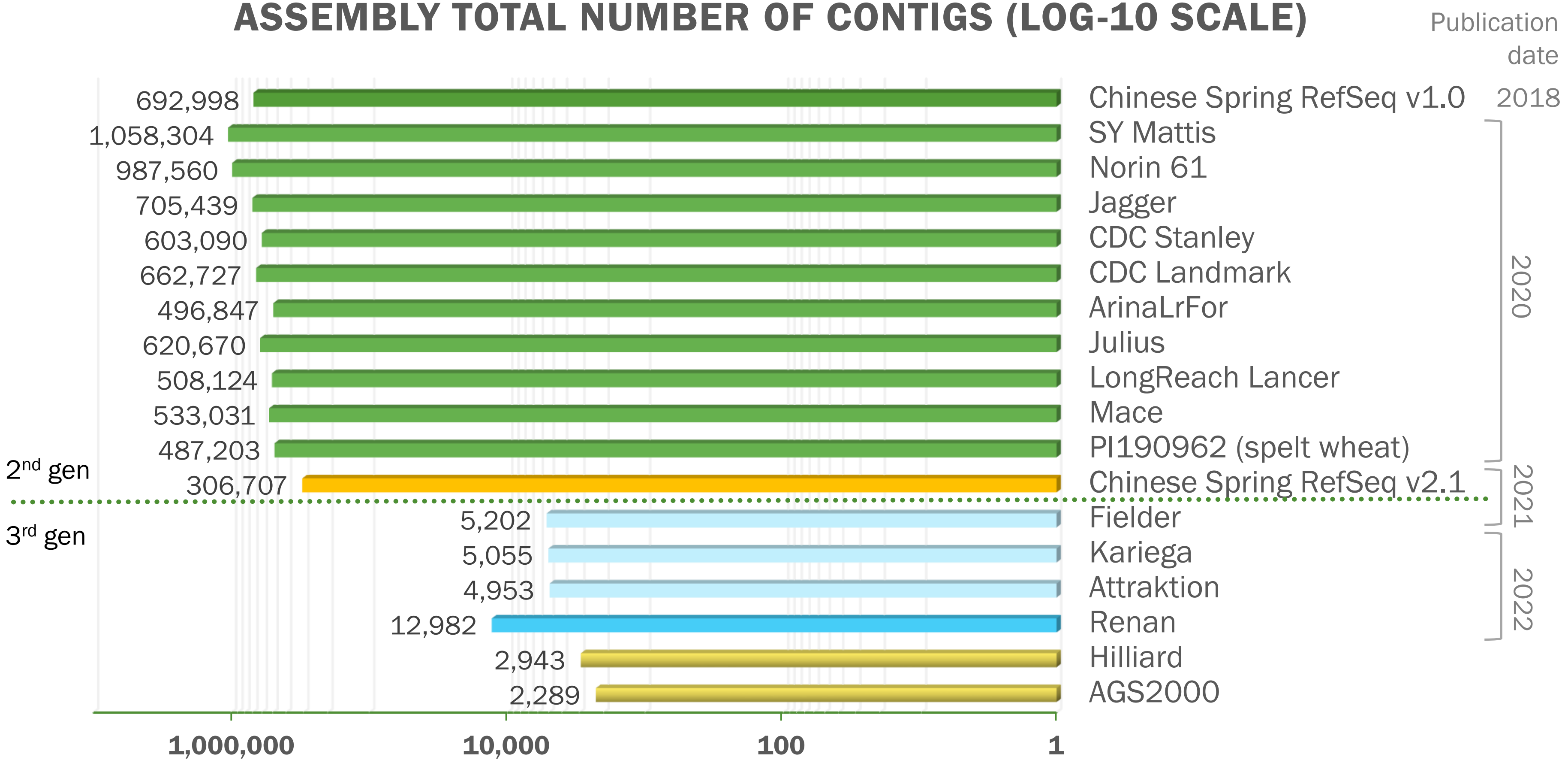
- PacBio Circular Consensus Sequencing (CCS) resolves errors by sequencing in multiple passes
- HiFi reads: consensus reads >=Q20

HIFI SOFT WINTER WHEAT GENOME ASSEMBLIES

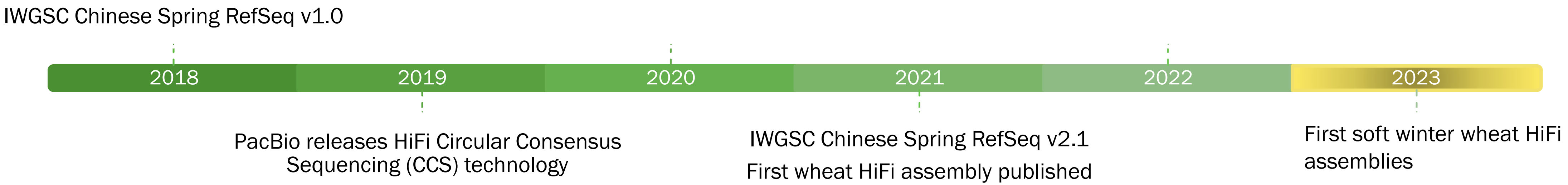
CULTIVARS AGS2000 & HILLIARD

WHEAT HIFI ASSEMBLIES HAVE ORDERS OF MAGNITUDE GREATER CONTIGUITY

ASSEMBLY TOTAL NUMBER OF CONTIGS (LOG-10 SCALE)



- 🟩 Short reads +
- 🟨 Current RefSeq
- 🟦 HiFi long-read assemblies
- 🟦 ONT long-read assembly
- 🟨 Presented herein



IWGSC Chinese Spring RefSeq v1.0

PacBio releases HiFi Circular Consensus Sequencing (CCS) technology

IWGSC Chinese Spring RefSeq v2.1

First wheat HiFi assembly published

First soft winter wheat HiFi assemblies

GENOME ASSEMBLY STATISTICS FOR SWW CULTIVARS ‘AGS2000’ AND ‘HILLIARD’

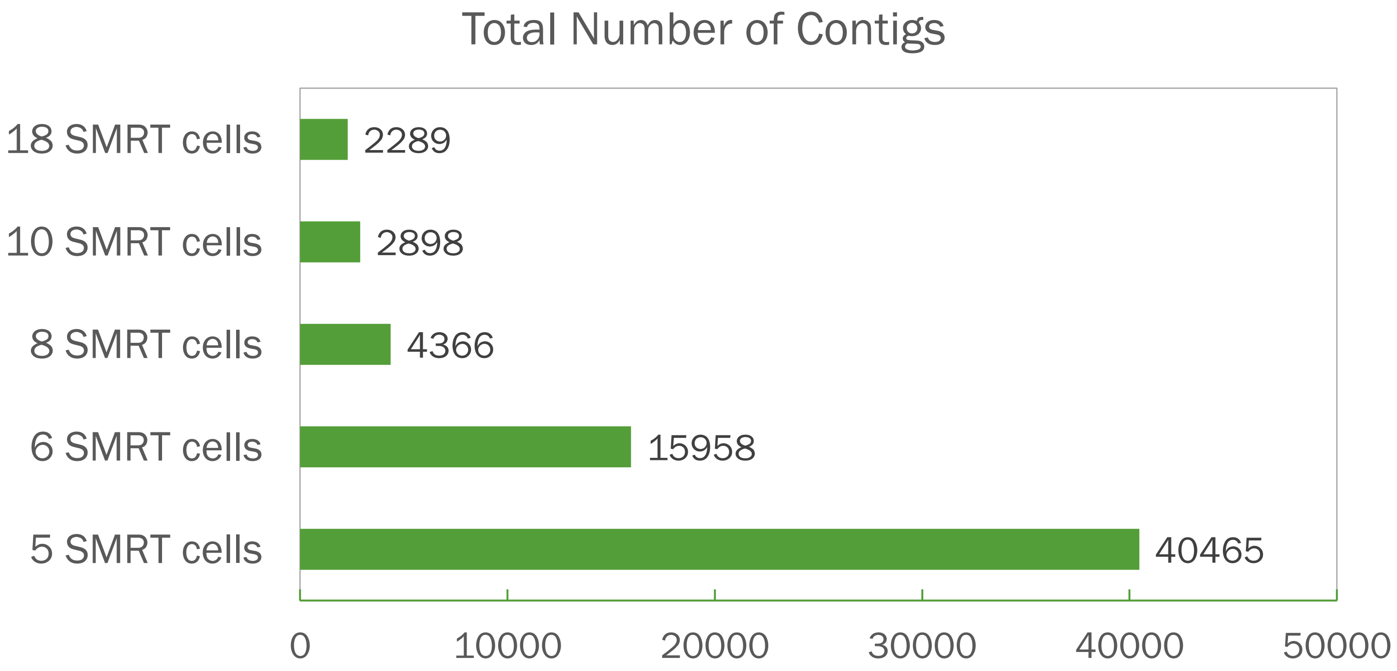
- Scaffolding with `RagTag` using reference genome ‘Attraktion’ Kale et al 2022.
ENA accession PRJEB48529.
- 10 SMRT cells yields a high quality assembly.
- Yet, significant improvements in contig N50 and L50 can still be gained with increased sequencing depth.

Sheron Simpson, USDA-ARS GBRU
Cal Youngblood, MSU

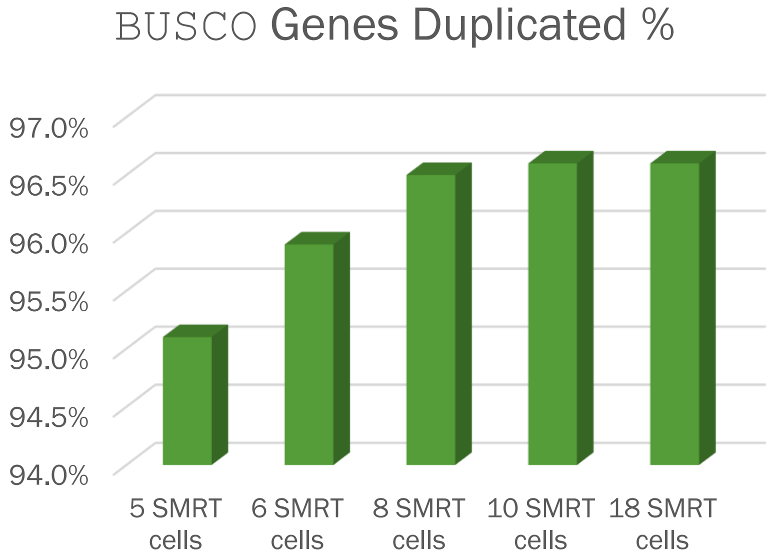
Sample	AGS2000	Hilliard
# SMRT Cells	**18**	**10**
PACBIO CCS HIFI DATA		
Raw Total Yield (Tb)	9.42	4.92
Input Coverage (X)	35.3	19.0
`RagTag` SCAFFOLD ASSEMBLY		
Scaffold Total Size (Gb)	14.642	14.616
# Pseudomolecules	21	21
`HifiASM` CONTIG ASSEMBLY		
Contig #	2289	2943
Contig N50 (Mb)	63.44	23.14
Contig L50 (# contigs)	56	161
Contig % in >50 Kb	99.74%	99.82%
Max Contig Length (Mb)	262.20	172.19

WHEAT HIFIASM ASSEMBLY DOWNSAMPLING

FROM ‘AGS2000’



- HifiASM did not assemble with <5 SMRT cells
- 8 - 18 SMRT cells covered comparable gene space in BUSCO analysis



Sample	AGS2000	Hilliard
# SMRT Cells	18	10
BUSCO v5.2.2		
Gene set	poales_odb10	
Complete %	99.4%	99.4%
Single %	2.8%	2.7%
Duplicated %	96.6%	96.7%
Fragmented %	0.0%	0.0%
Missing %	0.6%	0.6%

COPY NUMBER VARIANT (CNV) DETECTION:

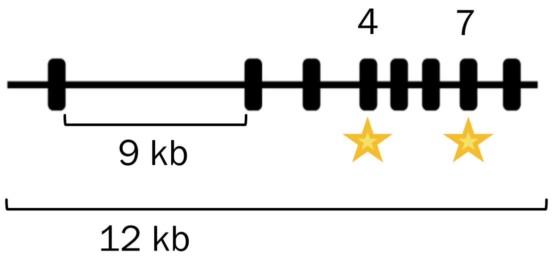
VERNALIZATION-A1 GENE

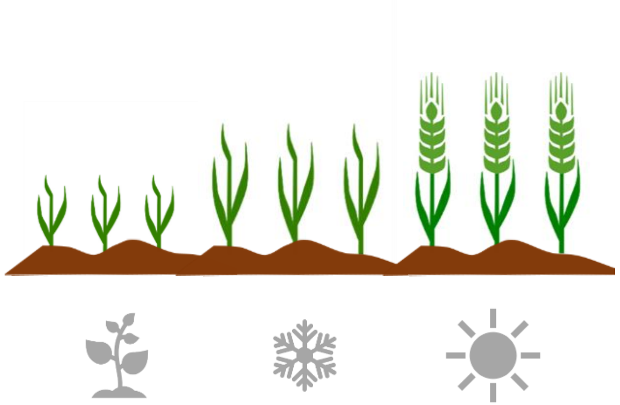
VERNALIZATION-A1 (*VRN-A1*) GENE

Vernalization is a response to prolonged cold exposure required for initiation of flowering in winter wheat and other plants.

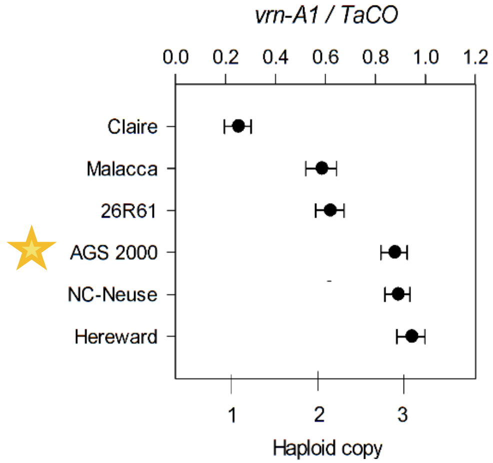
- The large (12 kb) *VRN-A1* gene is a central regulator of flowering in wheat.
- Known ‘tandem’ copy number variation exists in *VRN-A1* (see table).
- **Winter wheats** most commonly have **3 *vrn-A1* copies**.
- Increased copies of *vrn-A1* are associated with longer vernalization requirement.
- SNP in exons 4 and 7 have been associated with functional outcomes and correlated with copy number.
- The structure of the multi-copy *vrn-A1* region **remains elusive, often collapsed** in assembly.

VRN-A1



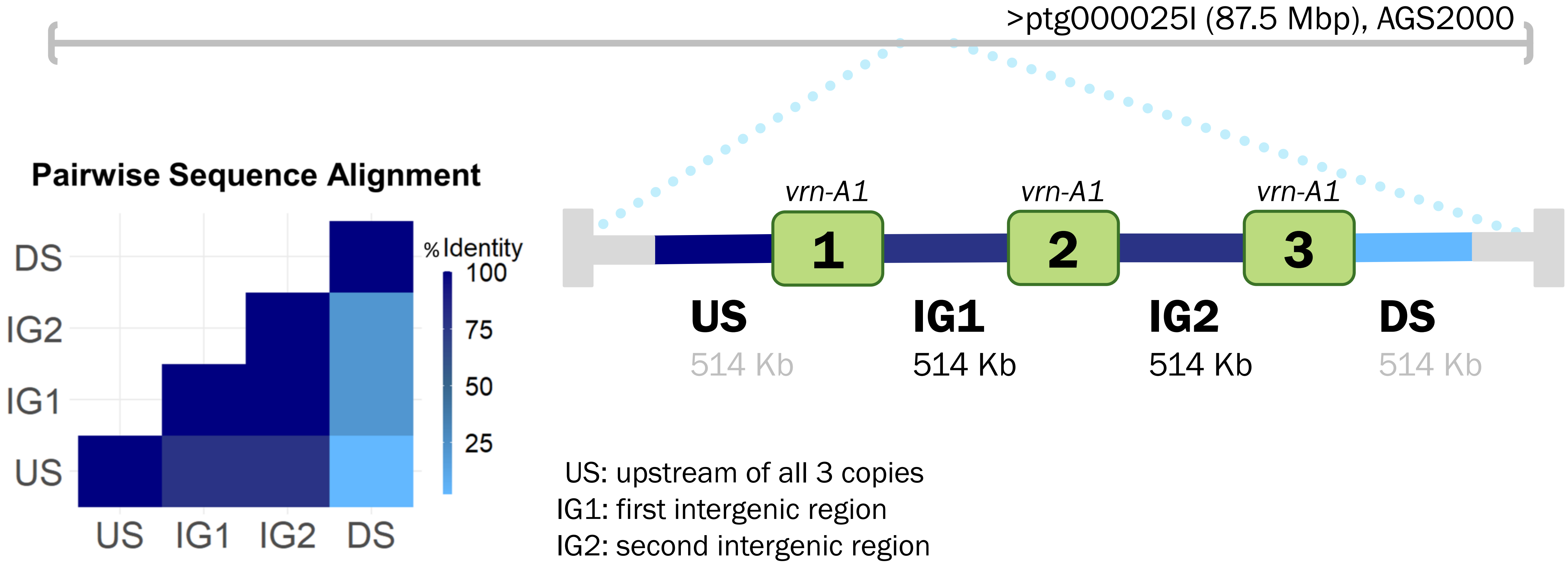


Copy Number Variation in *vrn-A1*



Guedira et al. 2016 PlosONE

ALL 3 COPIES OF *VRN-A1* ASSEMBLED IN A SINGLE 87.5 MB CONTIG REVEALING LARGE (514 KB) INTERGENIC REGIONS



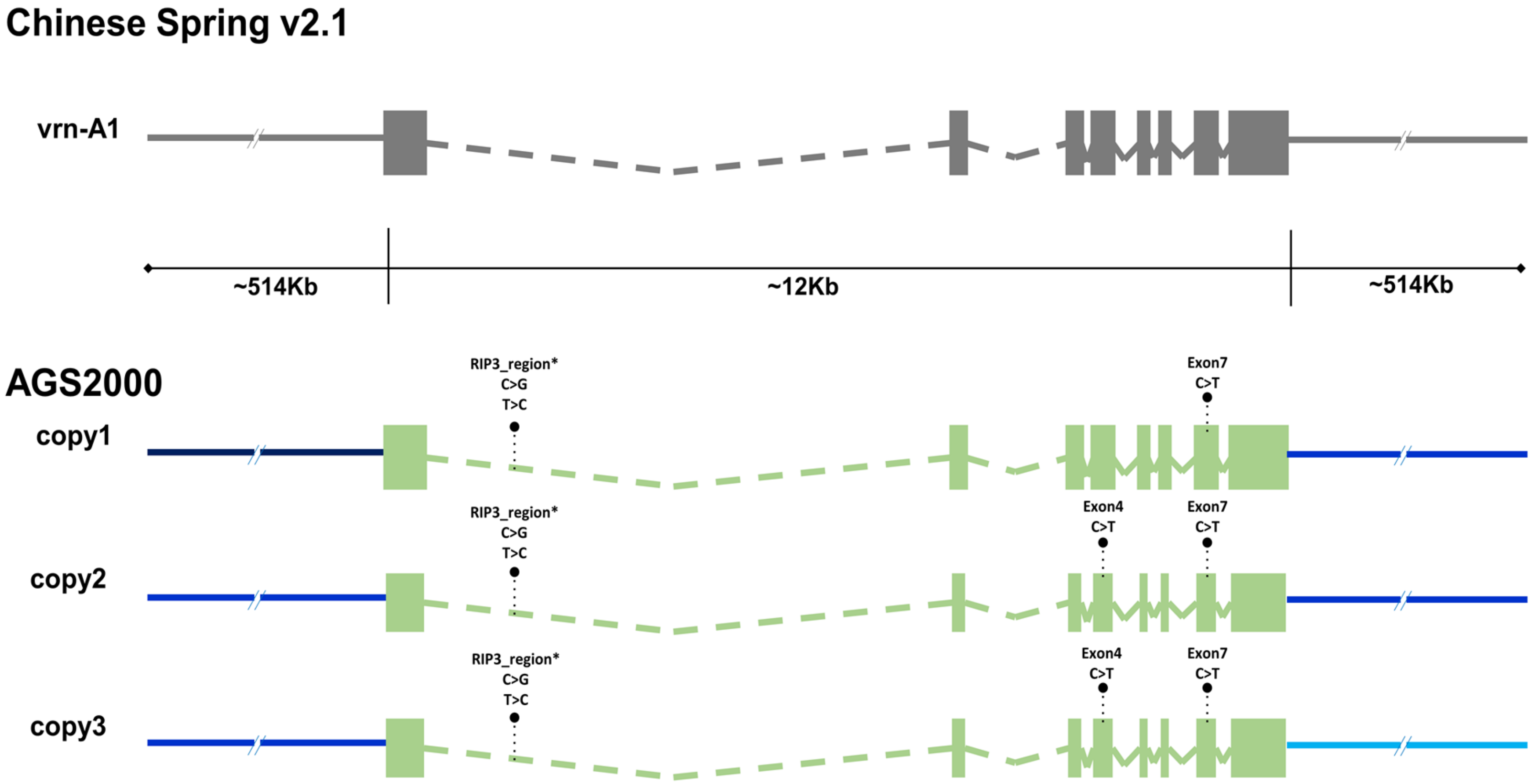
US: upstream of all 3 copies
IG1: first intergenic region
IG2: second intergenic region
DS: downstream of all 3 copies

COMPLETE ASSEMBLY OF VRN-A1 REGION ENABLES VARIANT CALLING AMONG 3 TANDEM COPIES IN 'AGS2000'

Luis Rivera-Burgos, NCSU

Functional SNP in the 3 *vrn-A1* copies mapped against reference Chinese Spring:

- Intron 1: 2 SNP (C>G;T>C) in GRP3 binding region in all 3 copies
- Exon 4: 1 SNP (C>T) in **two** of the **three** copies
- Exon 7: 1 SNP (C>T) in all 3 copies



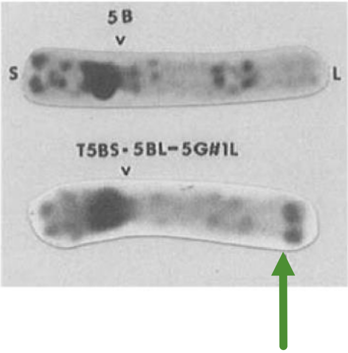
STRUCTURAL VARIANT (SV) DETECTION:

CHROMOSOME 5B/5G INTROGRESSION

CHROMOSOME 5B/5G INTROGRESSION
FROM T. TIMOPHEEVII

Triticum timopheevii
KSU Wheat Genetics Resource Center

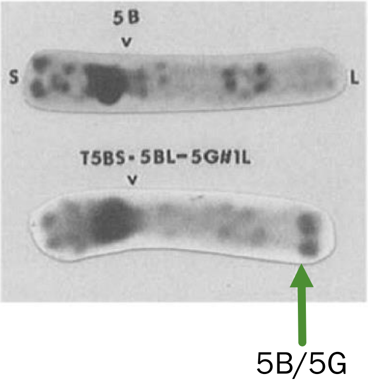
- *Triticum timopheevii* subsp. *timopheevii* (2n = 4x = 28, A^tA^tGG) is a cultivated tetraploid wheat relative native to Iran, Iraq, and Turkey.
- Genome is partially-homologous with *T. aestivum* (2n = 6x = 42, AABBDD)
- Introduced the ***Lr18*** leaf rust (*Puccinia triticina*) resistance gene into wheat germplasm via introgression on long arm of chromosome 5B.
  - ***Lr18*** introgression is present in cultivar ‘Hilliard’, not in ‘AGS2000.’
  - ***Lr18*** present in 37% of elite lines in 2022 SWW regional nurseries.
- Exact genomic position and extent of the introgression remains unknown.

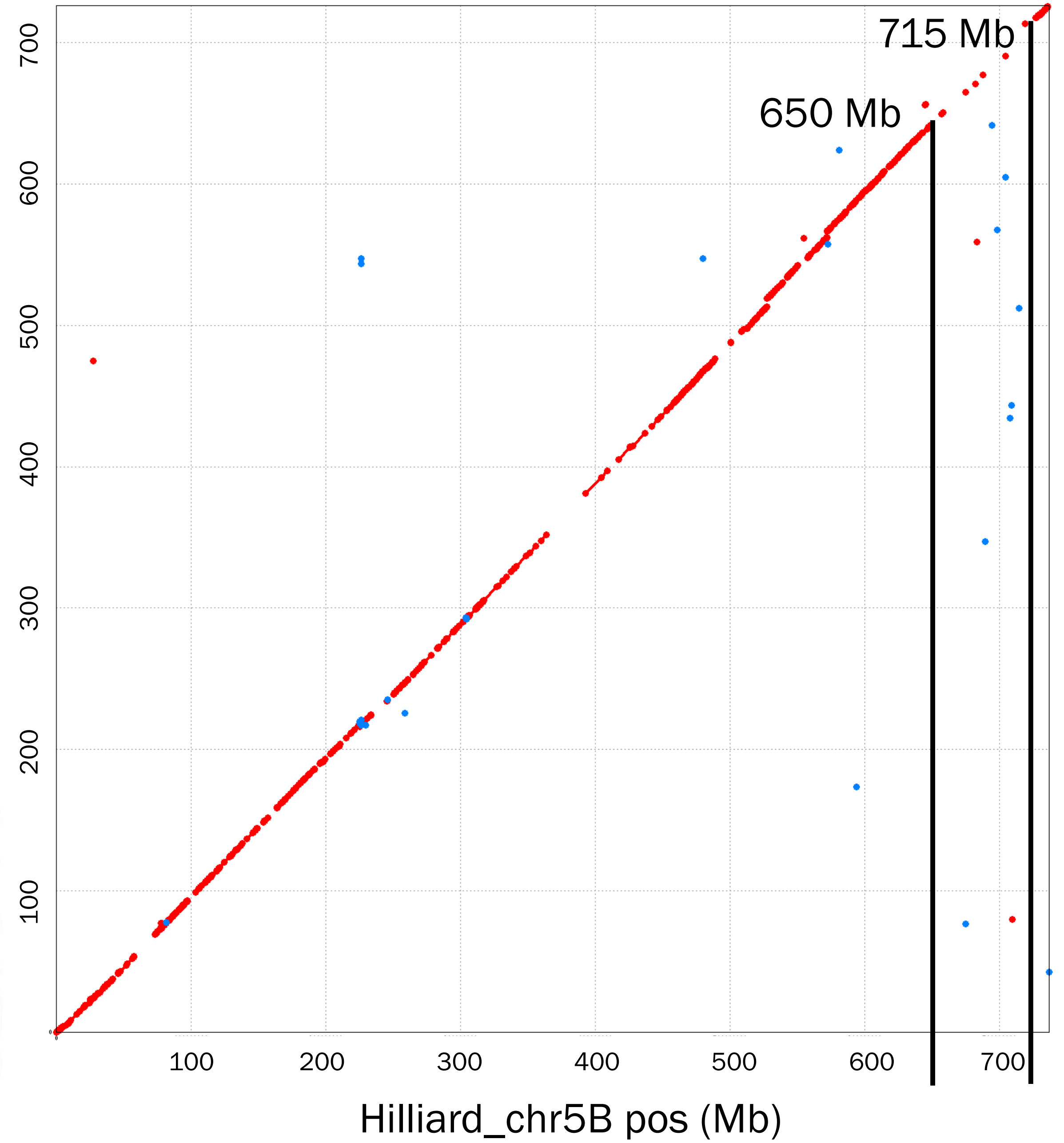


C-band staining of chrom 5B
Arrow shows 5G introgression from *T. timopheevii*.
Friebe et al. *1996 Euphytica* **91:** 59-87.

5B/5G INTROGRESSION DELINEATION

- MUMMER plot of alignment between chromosomes 5B from ‘Hilliard’ and ‘AGS2000’ assemblies
- Alignments filtered to 99% identity revealed sequence divergence between positions: 650 Mb – 715 Mb
  - *This divergent region matches the 5B/5G introgression.*
- Interestingly, the terminal sequence remains conserved.

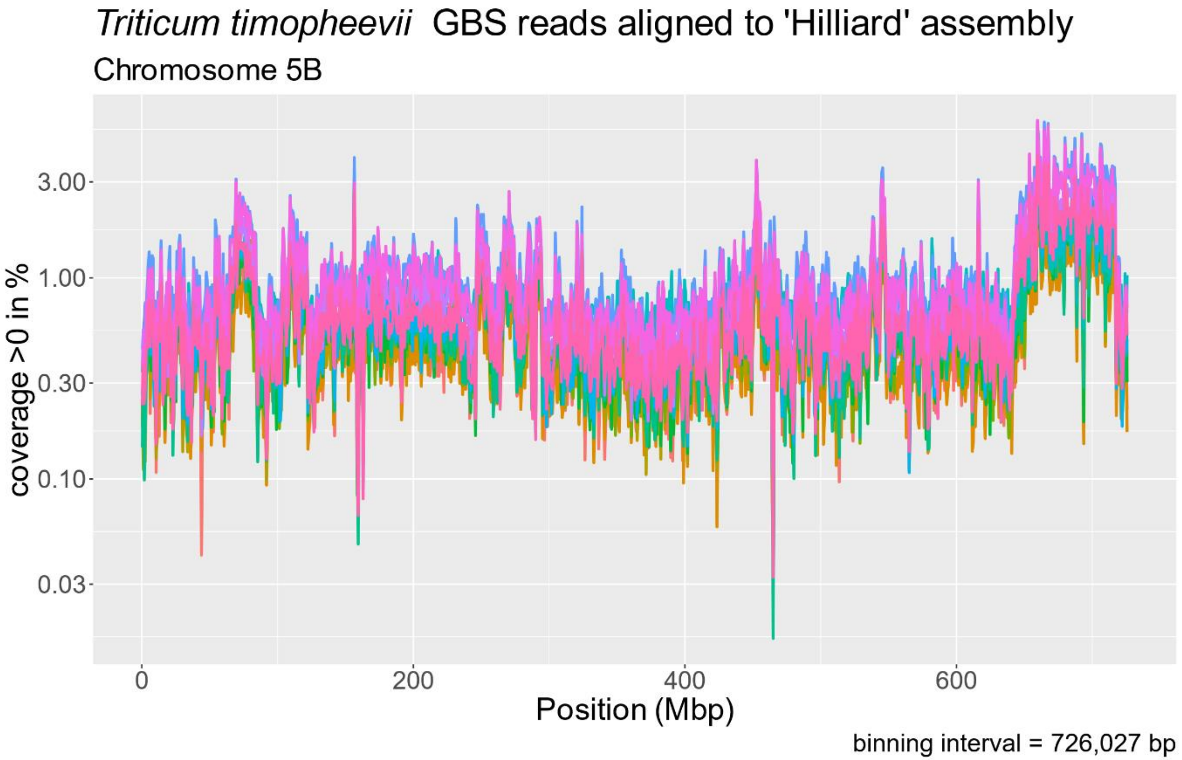




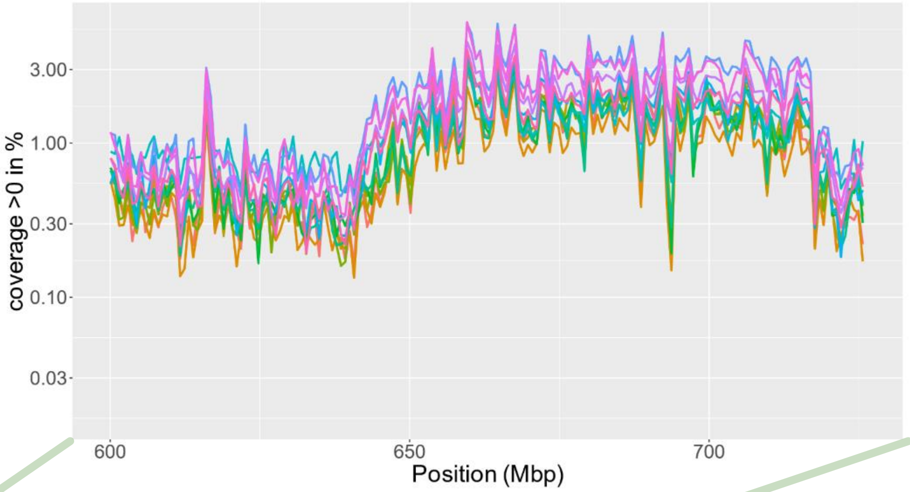
T. TIMOPHEEVII READ MAPPING
COVERAGE ANALYSIS CONFIRMS 5B/5G INTROGRESSION OF APPROX. 65 MB

~65 Mb
Introgression

	Estimated Start Position (Mb)	Estimated End Position (Mb)
mean (n=12)	652.2	715.7
Minimum	647.3	714.8
maximum	653.8	716.9


Triticum timopheevii GBS reads aligned to 'Hilliard' assembly

Chromosome 5B



Colors represent read mappings from 12 different *T. Timopheevii* accessions genotyped by GBS in Hyun et al. 2020. NCBI Project PRJNA601245.

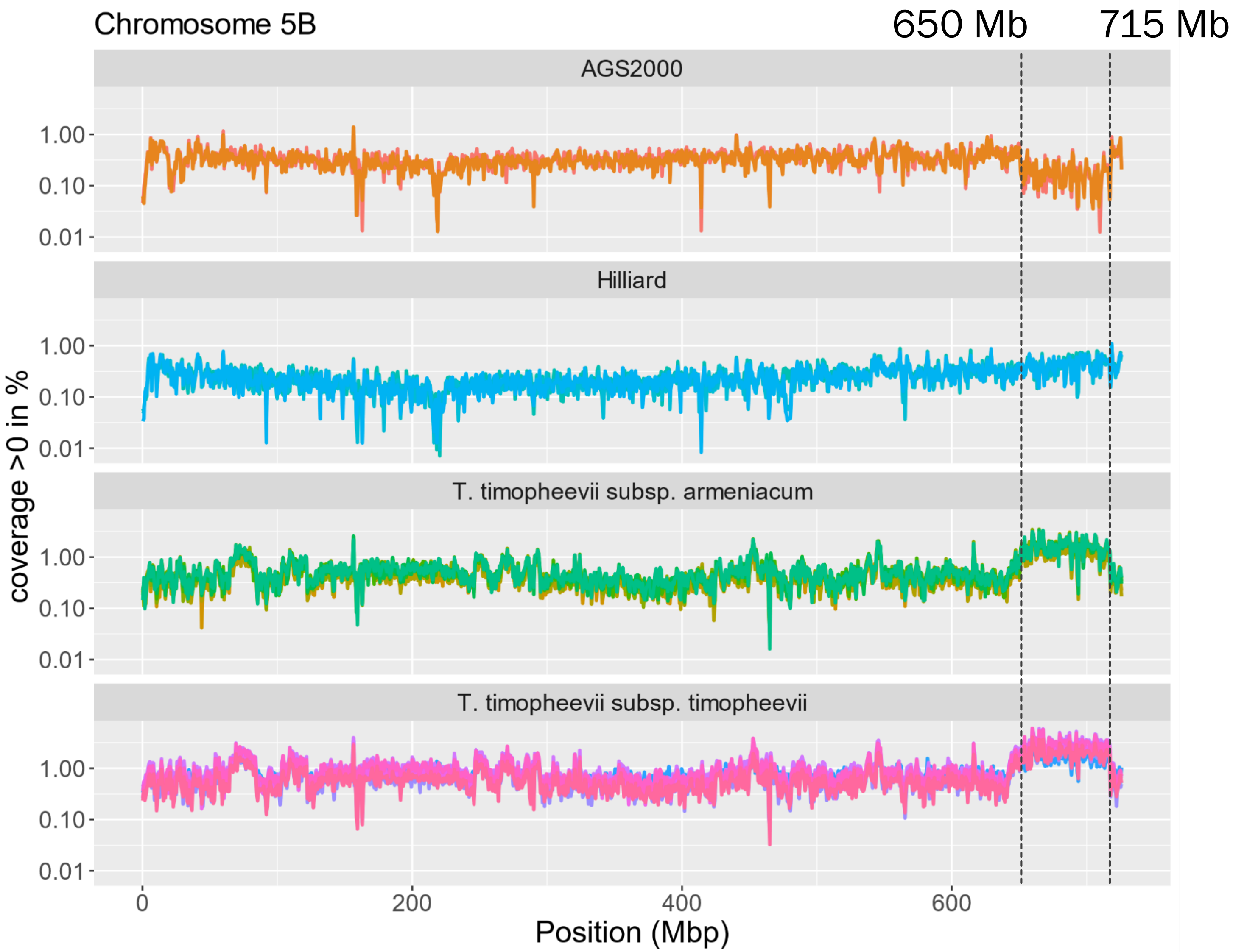
COVERAGE ANALYSIS CONFIRMS 5B/5G INTROGRESSION OF APPROX. 65 MB

- Mapping genotyping-by-sequencing (GBS) short reads to the ‘Hilliard’ assembly
- 5B/5G introgression location 650 Mb – 715 Mb supported by:
  1. Drop in ‘AGS2000’ reads mapped
  2. Increase in *T. timopheevii* reads mapped from both subspecies *timopheevii* and *armeniacum*

Matthew Willman, NCSU

GBS reads aligned to 'Hilliard' assembly

Chromosome 5B



binning interval = 726,027 bp

ACKNOWLEDGEMENTS

Brown-Guedira Lab USDA-ARS Eastern Regional Small Grains Genotyping Lab (ERSGGL), NCSU

Luis Rivera-Burgos

Matthew Willman

Kim Howell

Jared Smith

Joy Horovitz

Nico Lara

Noah DeWitt

Mohammed Guedira

Jo Wood, Wellcome Sanger Institute

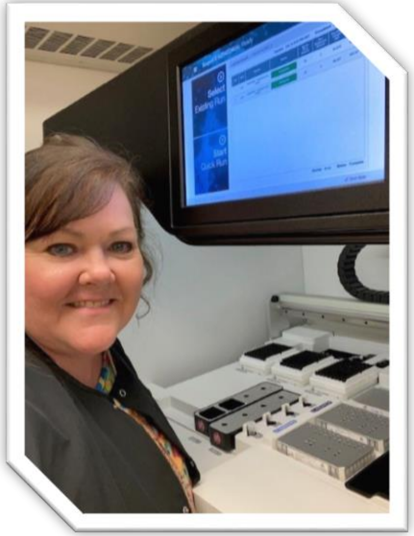
USDA-ARS **Genomics and Bioinformatics Research Unit** (GBRU)

Sheron Simpson ►

Cal Youngblood, Mississippi State University

Amanda Hulse-Kemp

Brian Scheffler



Hulse-Kemp Lab USDA-ARS GBRU, NCSU

Emily Delorean

Keo Corak

Ashley Schoonmaker

Cassie Newman

Grant Billings

Heather Manching

Jeremy Winders

WHEAT CAP
Coordinated Agricultural Project

NC STATE UNIVERSITY

USDA

Small Grains Genotyping