CoreDetector: scalable, chromosome-agnostic core-genome alignment for structural-variation discovery in polyploid wheat

Mario Fruzangohar[1*], Wasin Pipattungsakul[1], Julian Taylor[1*]

[1]Biometry Hub, Adelaide University, South Australia | *mario.fruzangohar@adelaide.edu.au

AAGI Analytics for the Australian Grains Industry

ABSTRACT

The Australian grains industry has a strong focus on improving industry relevant traits of polyploid crops such as wheat and canola. In pre-breeding this can often involve an initial alignment of a set of population genomes to be used in downstream analytics of crop improvement pipelines. Generating these alignments is computationally demanding, and current chromosome-by-chromosome workflows miss interchromosomal events such as introgressions and translocations. To solve this we enhanced our core-genome alignment tool, CoreDetector, to handle polyploid genomes and built a toolkit that converts nucleotide-level core alignments into genome-wide profiles of synteny, copy-number variation (CNV), unique insertions and gaps, with no reference genome or predefined chromosome relationships. Across 16 hexaploid bread wheat assemblies, CoreDetector produced an 11.02 Gbp core alignment in 15 hours using 32 cores and 100 GB RAM, where PGGB and Minigraph-Cactus failed. The toolkit recovered known wheat rearrangements and revealed new structural variants, insertions and CNV loci at agronomically important genes. These discoveries could be exploited to provide impactful outcomes for Australian grain growers.

1. Introduction

- Pangenomes capture diversity beyond a single linear reference, but whole-genome alignment of large, repeat-rich polyploids such as hexaploid wheat strains current graph-based tools.
- Splitting genomes by chromosome needs chromosome-level assemblies and predefined correspondence. This approach misses introgressions and translocations.
- **Aim:** Generate a scalable, chromosome-agnostic core-genome alignment and structural-variation analysis for polyploid pangenomes.

2. Methods

- 16 hexaploid wheat assemblies (Ensembl Plants, soft-masked).
- Phylogeny-guided progressive intersection: closest genomes are aligned first with Minimap2, keeping only shared sequence (Fig. 1a).
- Chromosome-agnostic: same-named chromosomes are searched first, then all others.
- Toolkit: gaps/PAV, unique insertions (Fig. 1b), CNVs, BED export and synteny plots.
- Benchmarked against PGGB and Minigraph-Cactus.

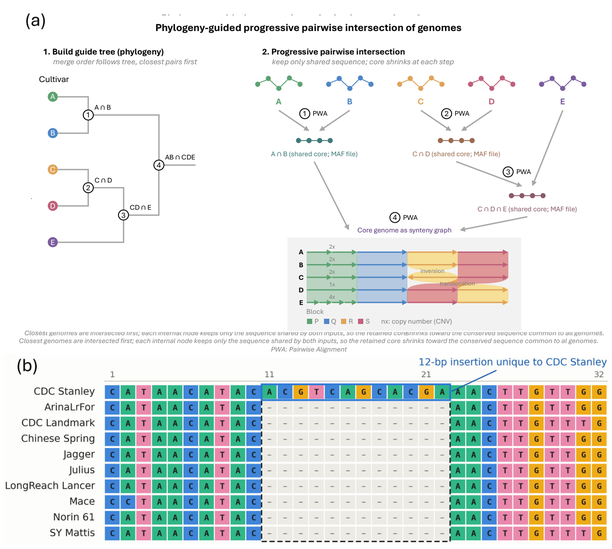


Fig. 1 (a) CoreDetector progressive core alignment; (b) unique insertion in CDC Stanley.

3. Results: scalability and core genome

Tool	Genomes	CPU / RAM	Outcome	Time (h)
CoreDetector	**16**	**32 / 100 GB**	**Completed**	**15**
Minigraph-Cactus	16	48 / 360 GB	Out of memory	<1
PGGB	16	48 / 360 GB	Unfinished (1 week)	>168
Minigraph-Cactus	Chr 1A	48 / 360 GB	Completed	41
PGGB	Chr 1A	48 / 360 GB	Crashed	45

Table 1. Runtime and resources on the 16-genome panel and on chromosome 1A alone.

- **Chromosome 1A vs Minigraph-Cactus:** 492 vs 312 Mbp aligned (82% vs 52%); block N50 15.3 vs 3.7 kbp; longest block 164 vs 10 kbp.
- **Reference-free insertions:** CoreDetector reports cultivar-specific insertions for all 16 cultivars (largest 86 kbp, CDC Stanley); Minigraph-Cactus reports only CS-specific ones.
- **Whole panel:** 11.02 Gbp core (77% of genome), 2.4 M conserved blocks, N50 36.4 kbp.

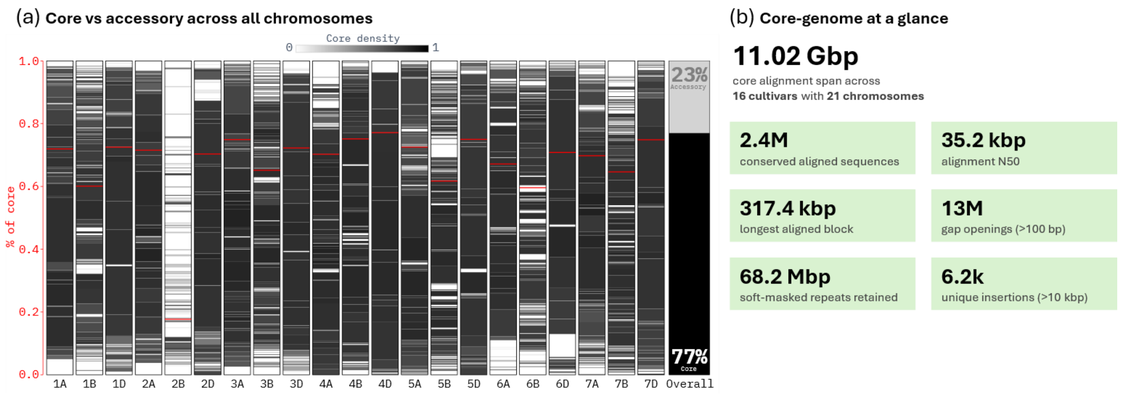


Fig. 2 Core vs accessory proportion per chromosome, and whole-genome core statistics.

4. Results: chromosome-scale synteny

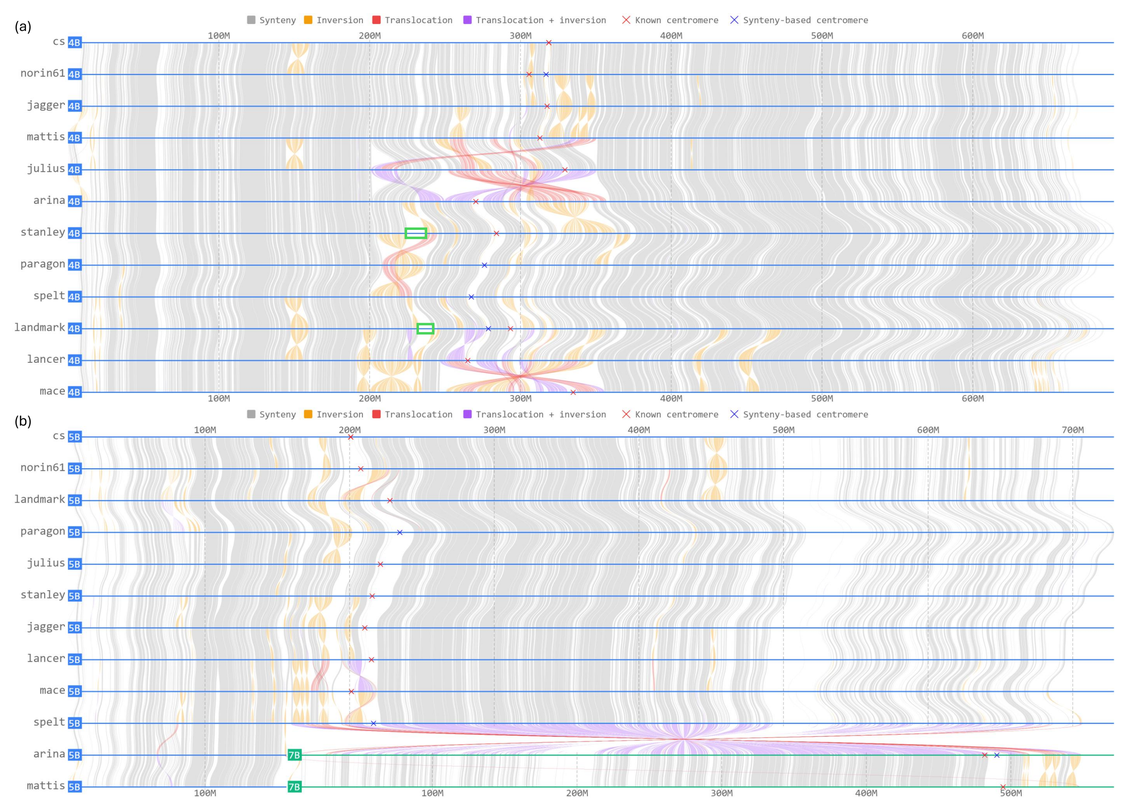


Fig. 5 Synteny across 12 chromosome-level cultivars (translocations red, inversions yellow). (a) 4B: complex pericentric rearrangements and expansions in Stanley and Landmark (green boxes). (b) 5B: known 5B–7B recombination in ArinaLrFor and SY Mattis.

- **Centromeres from synteny:** low-synteny gaps matched ChIP-seq centromeres in 193 of 210 chromosome–cultivar pairs; positions are inferred for Paragon and Spelt.

5. Genome features and discussion

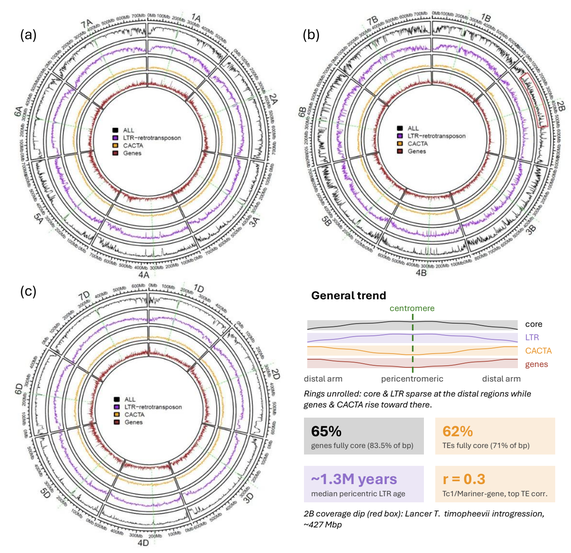


Fig. 4 Core (black), LTR (purple), CACTA (orange) and gene (brown) coverage in 1 Mbp windows; green dashes = centromeres.

- Core is densest pericentrically and drops sharply at centromeres.
- 2B dip: ~427 Mbp *T. timopheevii* introgression in LongReach Lancer.
- 6,218 unique insertions >10 kbp; the largest (104 kbp, Stanley 1D) carries ω-gliadin and NBS-LRR genes and is absent from Wheat Panache.
- 83.5% of CS genic space lies in the core.
- **CNVs:** 113,711 events >1 kbp across 15 cultivars; Paragon and Spelt carry the most. 29 defence-related genes vary in copy number.
- **Transposons:** genes with extra copies are enriched for Tc1/mariner (DTT, $p < 0.001$), suggesting cut-and-paste-driven duplication.

6. Take home and industry impact

- The CoreDetector algorithm generates a reference-free, chromosome-scale structural alignment of wheat on modest hardware.
- It comes with a rich toolkit of visualization and information extraction functionality for investigating structural variation.
- Resulting genome-alignments can be used in downstream analytics to understand important associations to industry relevant traits.
- These can be also used by breeders to exploit genomic features such as copy number variation and amino acid changes that are specific to a cultivar.

7. References

Fruzangohar M, et al. CoreDetector: core-genome alignment of evolutionary diverse genomes. Bioinformatics 2023;39:btad628.

Walkowiak S, et al. Multiple wheat genomes reveal global variation in modern breeding. Nature 2020;588:277–283.

Hickey G, et al. Pangenome graph construction with Minigraph-Cactus. Nat. Biotechnol 2024;42:663–673.

Garrison E, et al. Building pangenome graphs. Nat Methods 2024;21:2008–2012.

Bayer PE, et al. Wheat Panache. Plant Genome 2022;15:e20221.

GitHub: https://github.com/mfruzan/CoreDetector

GRDC Grains Research & Development Corporation | Adelaide University | Curtin University | The University of Queensland Australia

The support of the Grains Research and Development Corporation (GRDC) provided through the Analytics for the Australian Grains Industry (AAGI) project is acknowledged.