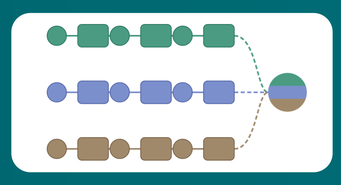
Centre for Crop and Disease Management

Curtin Biometry and Agricultural Data Analytics

Curtin University

PanAnnoGraph: from gene presence/absence to functional gene states in pangenome graphs



Kristina K Gagalova[1], Mitchell Bestry[2], Teng Li[2], David Edwards[2], Mark Derbyshire[1]

[1]Centre for Crop and Disease Management, Curtin University | [2]Applied Bioinformatics group, University of Western Australia

ABSTRACT

Pangenome graphs hold every assembly of a species, yet graph annotation still stops at present or absent: a copy with a premature stop or frameshift counts as present, hiding the function of alleles that matter for breeding. We compared tools for projecting annotations onto graphs and built **PanAnnoGraph**, which calls exon-level gene presence/absence from Minigraph-Cactus walks. A new sequence-consequence module combines variant-centric and gene-centric tracks to give a **functional state for every gene in every haplotype**.

1 Annotation projection on graphs

Tools place annotations in graph space, but none yet reports a gene-level state for every haplotype **and what each variant does to the protein**.

	Graph space	Reports all haplotypes	Gene PAV	Exon level	Codon effect	Visual view
vg rna transcripts projected onto every haplotype	✓	✓	·	·	·	·
vg annotate GFF → GAF in graph space	✓	·	·	·	·	·
odgi inject GFF/BED embedded as paths	✓	·	·	·	·	○
GrAnnot2 (GraSuite) transfer to all genomes · PAV matrix	✓	✓	✓	◐	·	·
SaVanache linearised pangenome viewer	✓	✓	·	·	·	✓
GraphPAV-Cactus *Teng Li, unpublished* walks or HAL liftover · whole-gene coverage	✓	✓	✓	·	·	·
PanAnnoGraph v0.1 walks, exon coverage	✓	✓	✓	✓	·	·
PanAnnoGraph v1.0 this work: Track A + Track B	✓	✓	✓	✓	★	·

● full ◐ partial ○ via other tools · not covered ★ new module (in development)

2 Minigraph-Cactus graph: what we extract

a Input assemblies

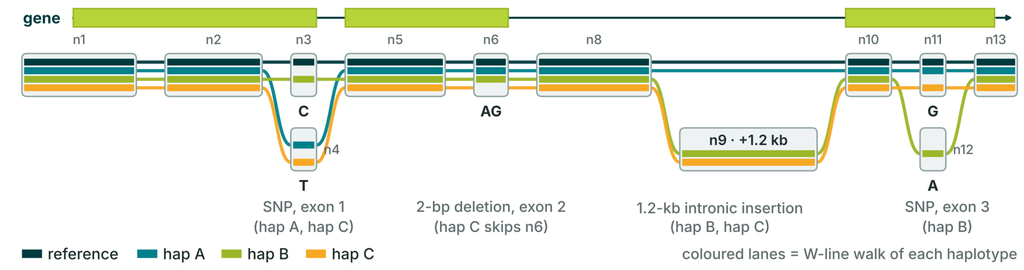
reference
hap A
hap B
hap C

minigraph SV backbone + Cactus base-level alignment

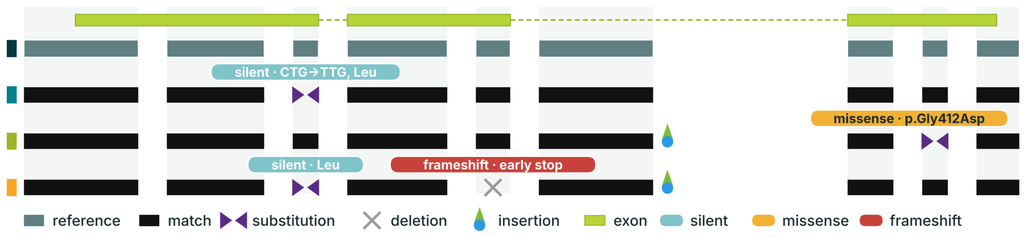
```
S n3   C                                  GFA 1.1
S n4   T
L n2 + n3 + 0M
L n2 + n4 + 0M
W ref    0 A01 >n1>n2>n3>n5…
W hapA   1 A01 >n1>n2>n4>n5…
W hapC   1 A01 >n1>n2>n4>n5>n8>n9…
```

S segment: node + sequence · L link: edge between nodes · W walk: one haplotype's path

b Minigraph-Cactus variation graph: nodes, bubbles (snarls) and walks



c Linearised view relative to the reference + coding consequence



d What PanAnnoGraph extracts

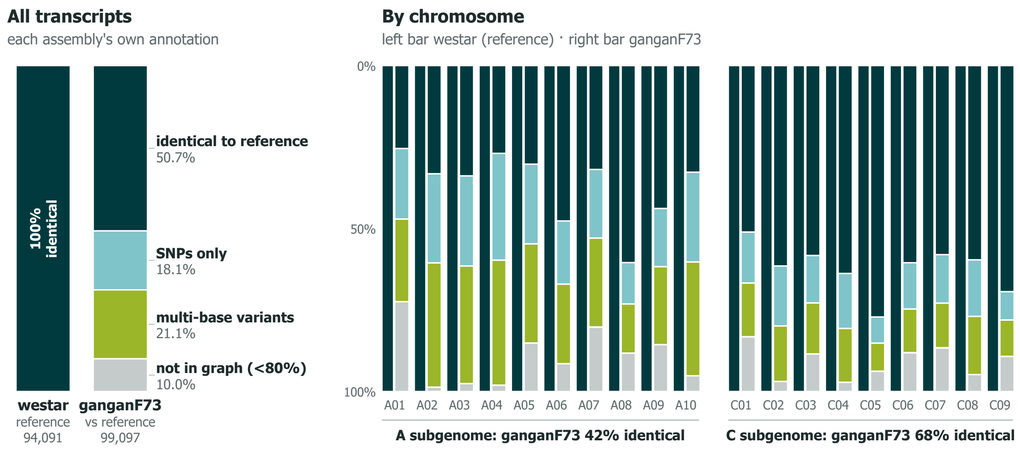
Exon × walk node coverage
→ gene presence/absence per haplotype

Snarls → vg deconstruct → VCF
→ Track A: variant consequences

Gene loci shared across walks
→ Track B: haplotype CDS & protein

3 PanAnnoGraph preliminary results in canola

Comparing annotations against the Westar reference in a canola pangenome graph showed that **44%** of present **ganganF73** genes differ from Westar in exon sequence.



% identical among transcripts present in the graph · 28 transcripts partly outside the graph not visible · deletions relative to Westar are not detected

4 A sequence-consequence layer

For each SNP and indel in a gene: is it **non-synonymous**, does it add a **stop codon** or **frameshift**, or change a **splice site**? No graph tool reports this yet; we propose two complementary tracks, and their union gives one state per gene copy.

Graph VCF — vg deconstruct, phased
Reference GFF3 — + per-assembly GFFs
Assembly FASTAs — + PanAnnoGraph gene loci

TRACK A · variant-centric — *What does each variant do?*

1. **Normalise graph VCF** — pop nested bubbles, realign alleles — `vcfbub` `vcfwave` `bcftools norm`
2. **Restrict to gene space** — CDS + UTR ± 50 bp splice window — `bcftools view -R`
3. **Haplotype-aware consequences** — phase-aware: compound codons, rescued frameshifts — `bcftools csq -p a` `SnpEff`
4. **Per variant × haplotype** — missense · stop-gain · frameshift · splice · start/stop-lost

Exon-level PAV from PanAnnoGraph v0.1 — present · partial · absent

TRACK B · gene-centric — *What does each gene copy look like?*

1. **Pair each copy with the reference** — shared locus from reference-projected models — `gffcompare`
2. **Extract or rescue the copy** — own model → CDS; no model → align ref protein — `gffread` `miniprot`
3. **Compare with reference** — codon alignment · stops · GT-AG splice sites — `MAFFT` `needle`
4. **Per gene × haplotype** — % protein retained · syn / non-syn

Integrate → one functional state per gene × haplotype

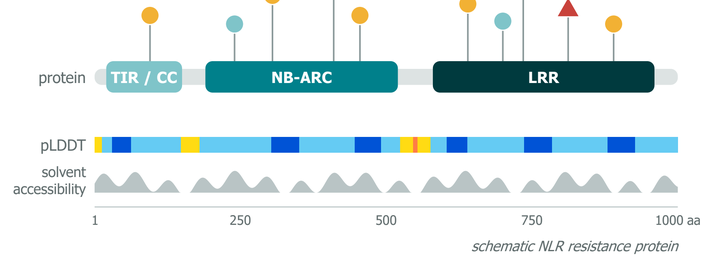
precedence (all or most severe):
- absent
- partial
- loss-of-function
- missense / in-frame
- synonymous
- identical

gene × haplotype state matrix (schematic)

outputs
- `state_matrix.tsv`
- `variant_consequences.tsv`
- **protein haplotypes per gene**
- **% protein retained per copy**

5 Longer term plan: variants in 3D

Mapping genetic variants onto predicted protein structures helps assess their potential structural and functional impact.



3D structure prediction with AlphaFold, ESMFold or Boltz

per residue
- Prediction confidence (pLDDT)
- Predict solvent accessibility
- Predicted functional domains - Pfam
- distance to pocket (P2Rank)

6 Why it matters for plant pre-breeding

Pangenome diversity → Gene loci and copies → Functional variation → Breeding insights

Allele mining
Identify donor lines with intact or favourable gene variants.

Natural knockouts
Discover naturally disrupted genes to study gene function and trait effects.

Diagnostic markers
Identify functional variants for targeted marker development.

Targeted introgression
Characterise donor haplotypes and flag linked disruptive variants (linkage drag).

Disease resistance
Compare resistance gene copies and integrity across germplasm.

Candidate gene discovery
Prioritise trait-associated genes using predicted variant effects.

References

Song J-M et al. (2020) Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of Brassica napus. Nat Plants 6:34–45.

Hickey G et al. (2024) Pangenome graph construction from genome alignments with Minigraph-Cactus. Nat Biotechnol 42:663–673.

Garrison E et al. (2018) Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nat Biotechnol 36:875–879.

Guarracino A et al. (2022) ODGI: understanding pangenome graphs. Bioinformatics 38:3319–3326.

Ravel S et al. (2026) GraTools, a user-friendly tool for exploring and manipulating pangenome variation graphs. BMC Bioinformatics.

AAGI Analytics for the Australian Grains Industry

GRDC Grains Research & Development Corporation

Adelaide University · Curtin University · The University of Queensland Australia · applied bioinformatics group · The University of Western Australia

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