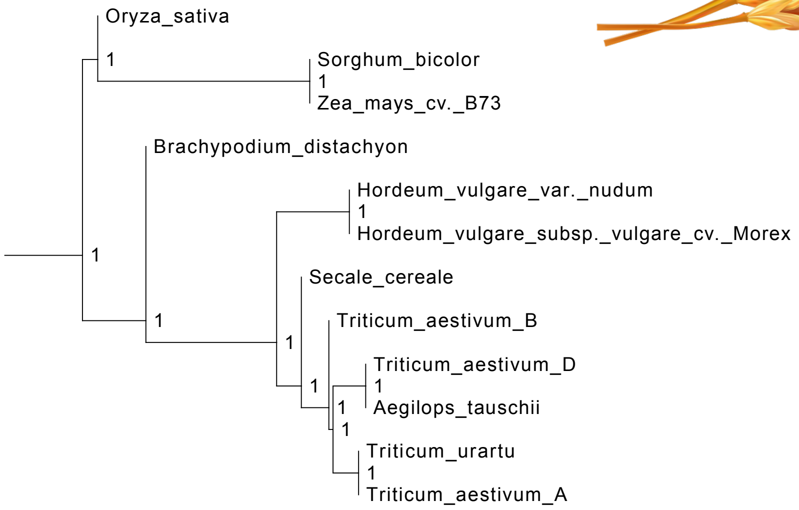
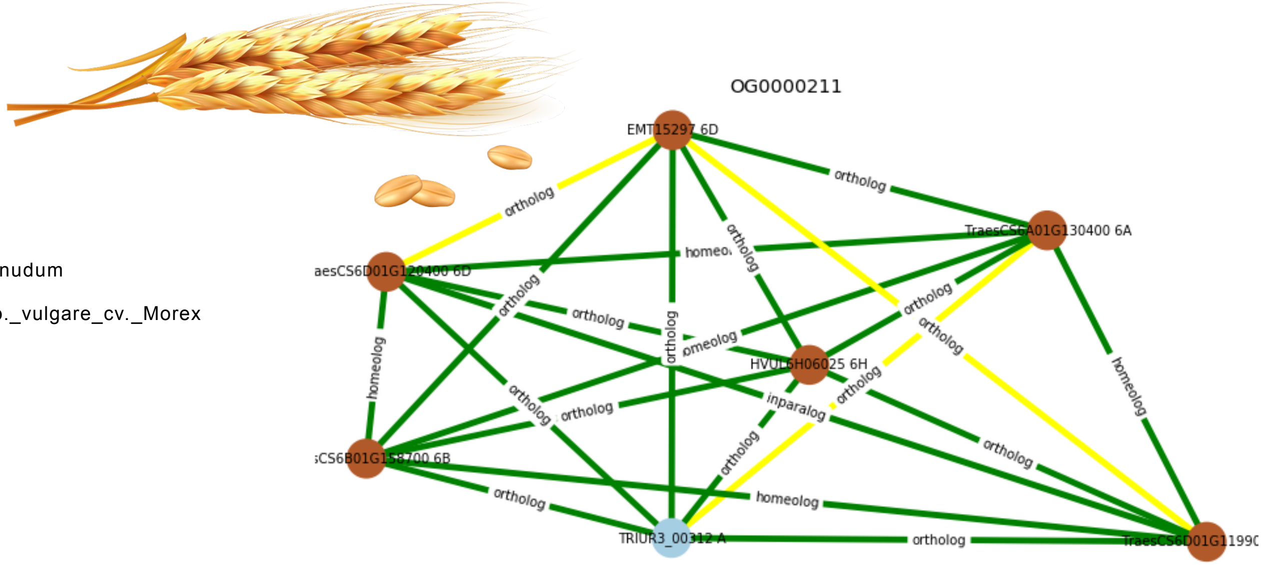
HelmholtzZentrum münchen
German Research Center for Environmental Health

PGSB Plant Genome and Systems Biology

Phylogenomics framework used to study gene copy number variations (GCNVs) in bread wheat

Daniel Lang|daniel.lang@helmholtz-muenchen.de





deNBI GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE

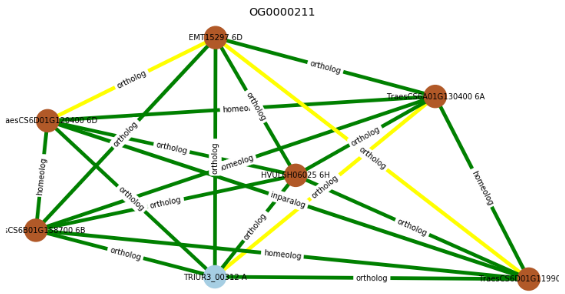
HELMHOLTZ ASSOCIATION

Phylogenomics/annotation resources published with the **bread wheat V1 genome papers**

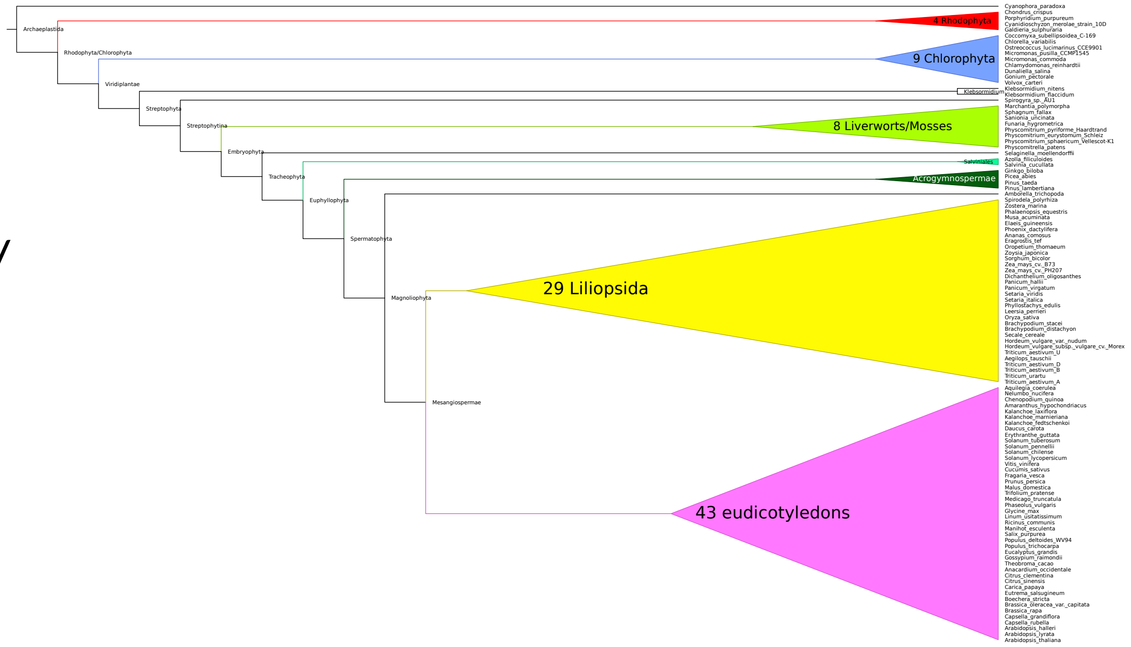
The tragic ripples of an epic fraud p. 634
Insect pest profits from maize defenses pp. 642 & 694
Photoredox activation of methane pp. 647 & 668
Science
$15
17 AUGUST 2018
sciencemag.org
AAAS
ROAD MAP FOR
WHEAT
Ordered sequence will speed research pp. 635, 661, & 662

- Resources:
  1. classified homeologs
  2. orthology-based GOA|POA|TOA
  3. wheat Transcription Associated Proteins (TAPs=TF+TR)
  4. wheat gene families including orthologs from other Viridiplantae
  5. gene families with significant GCNVs
- all downloadable → links also in main paper supplement
  - https://wheat-urgi.versailles.inra.fr/Seq-Repository/Annotations
  - http://dx.doi.org/10.5447/IPK/2018/5
  - http://doi.org/10.1126/science.aar7191
- relevant resources are highlighted on the slides: **resource X**

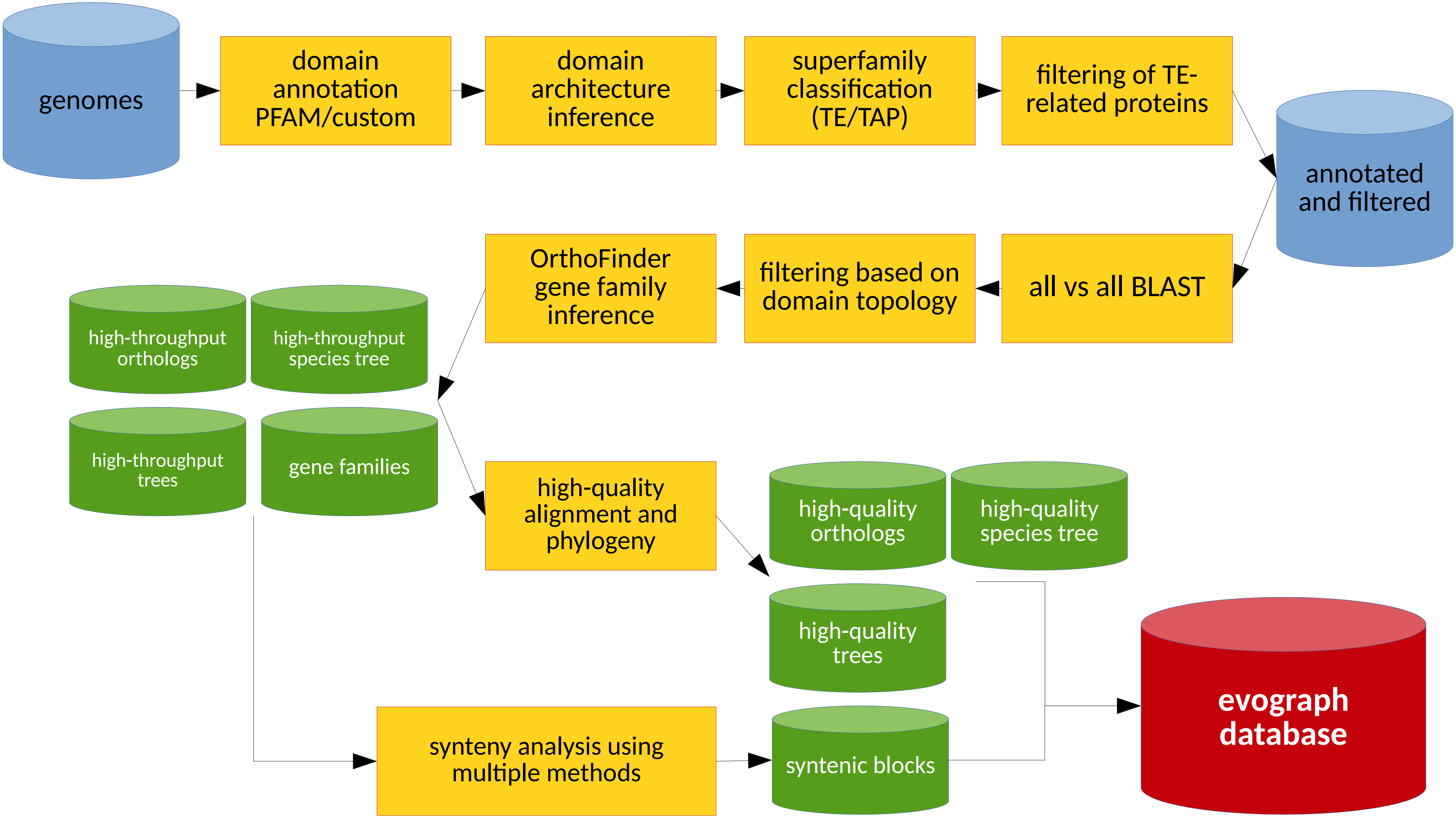
evograph
A graph of all plant gene families



- scalable, iterative phylogenomics workflow: *evograph*
  - targeting Plantae diversity
  - combine multiple iterations
  - broad taxon sampling ↔ family/ pangenome
- specific run for the bread wheat analysis
  - focussing on grasses → Triticeae
  - with representative Viridiplantae as outgroups

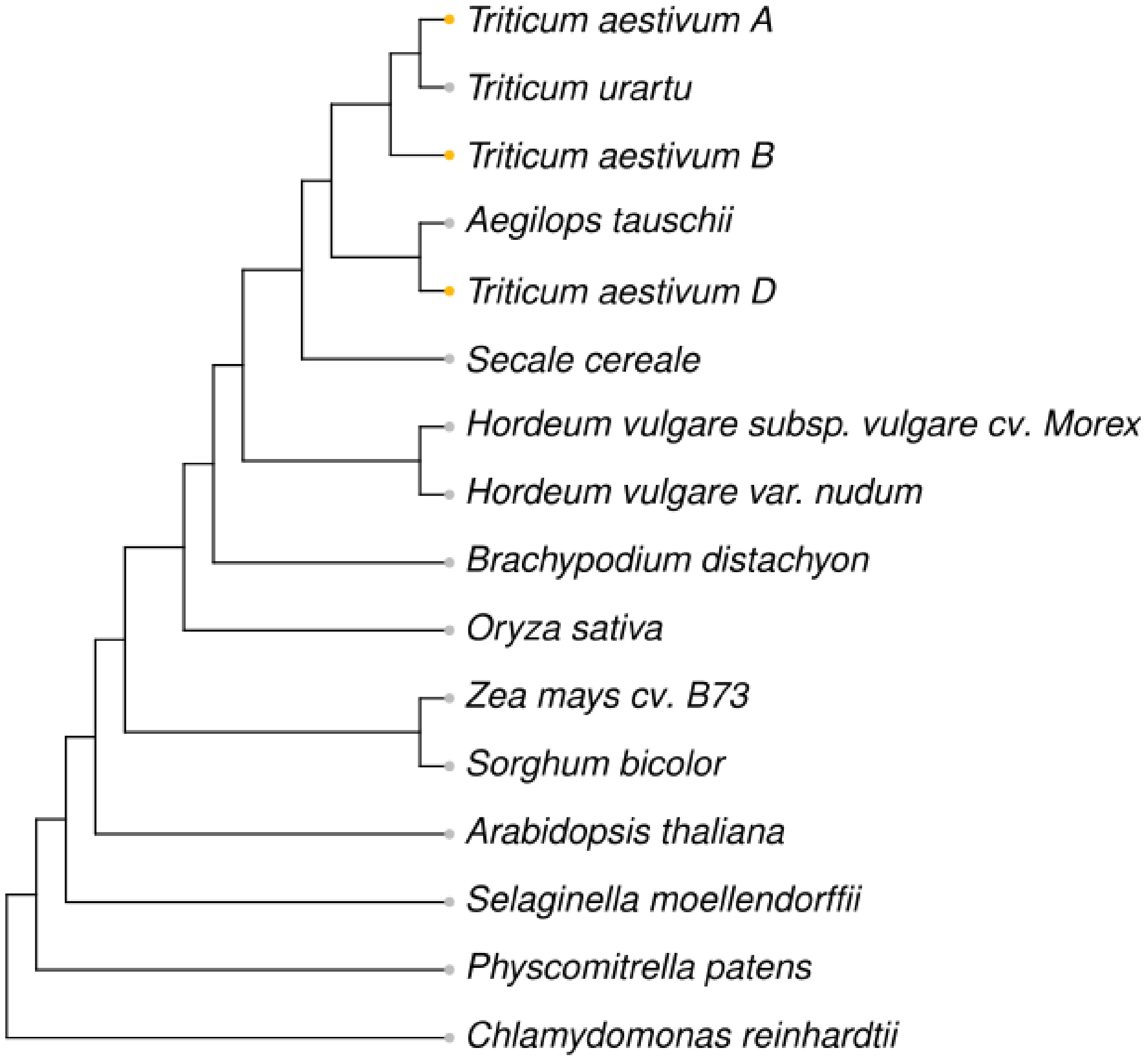


evograph
iterative workflow


genomes
domain annotation PFAM/custom
domain architecture inference
superfamily classification (TE/TAP)
filtering of TE-related proteins
annotated and filtered
all vs all BLAST
filtering based on domain topology
OrthoFinder gene family inference
high-throughput orthologs
high-throughput species tree
high-throughput trees
gene families
high-quality alignment and phylogeny
high-quality orthologs
high-quality species tree
high-quality trees
synteny analysis using multiple methods
syntenic blocks
evograph database

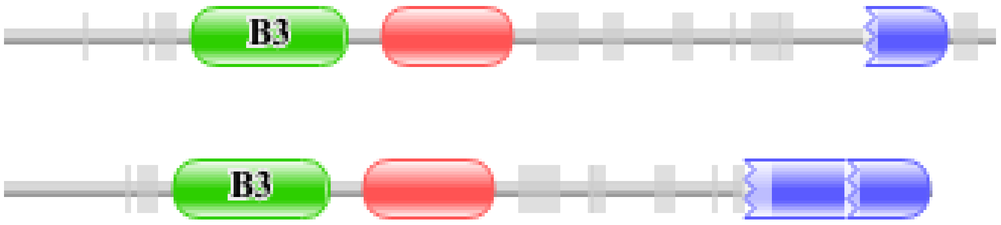
Bread wheat phylogenomics study

- 12 grass genomes
  - 9 grass species
  - A/B/D subgenomes as individual taxa
  - 2 barley varieties
- 4 Viridiplantae outgroups
- using OrthoFinder on filtered BLAST links
  - DendroBLAST phylogenies
- custom species tree
  - based on >12k subfamily trees [ASTRAL]
- dated chronogram of grasses used for comparative analysis
  - dated based on previous molecular clock estimates and fossil dates



Protein **domain architectures** define gene families

- e.g. these represent 2 distinct families of ARFs:



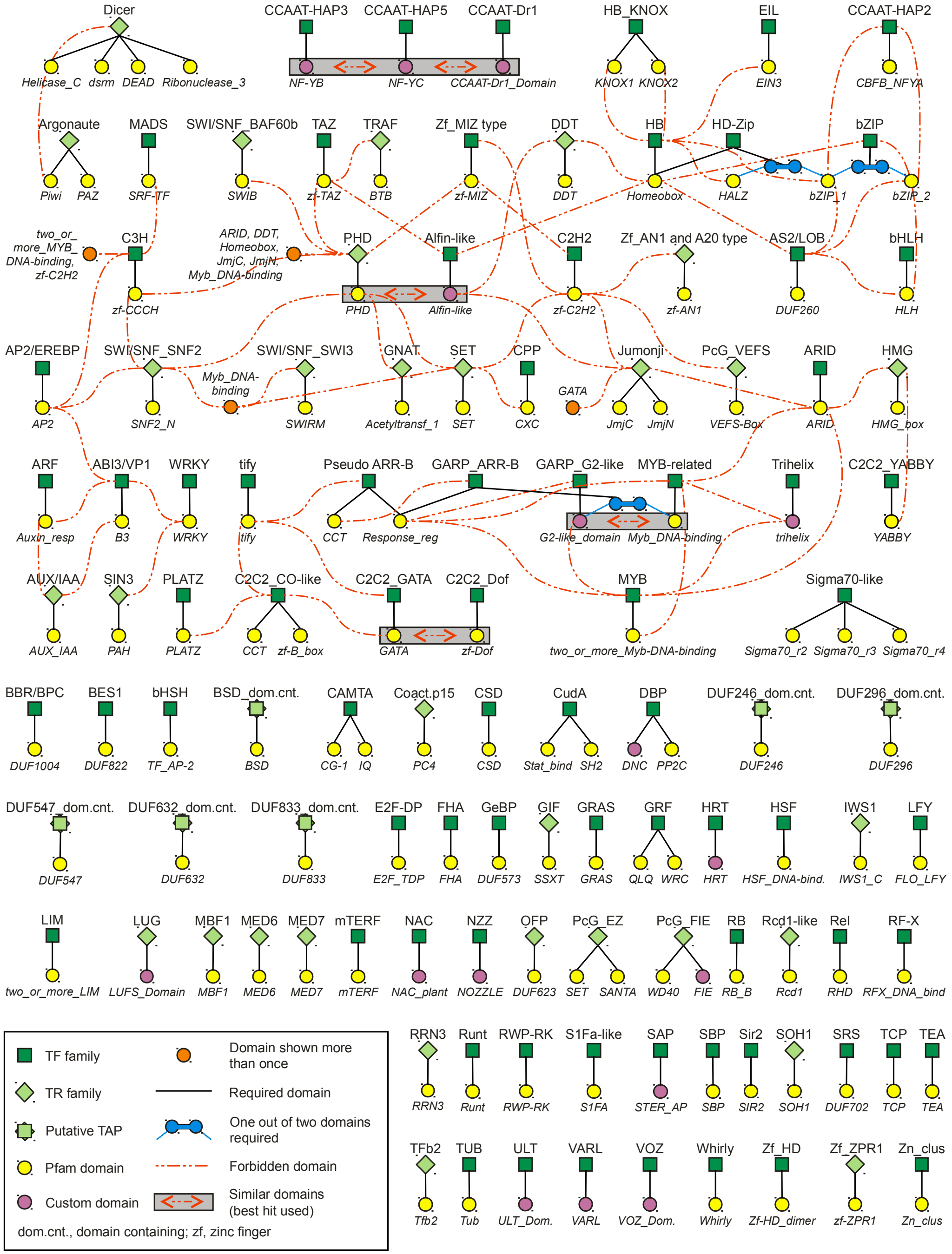
- subfamily = comprising only orthologs and inparalogs
- gene family here:
  - 1-few subfamilies ideally traced to one ancestor gene in LCA of Viridiplantae
- inference of domain architectures
  - PFAM and custom HMMs
  - filtering of all vs. all BLAST links used as OrthoFinder inputs
    - hits must cover at least domain architecture
  - domain presence/abscence rules:
    - TE classification
    - TF superfamily

resource iv

Using protein domain presence/absence to classify proteins e.g. **genome wide annotation of wheat Transcription Associated Proteins (TAPs) = TFs + TRs**

resource iii

genome	genes	TAP super families
A	3025	104
B	3211	103
D	3163	103
U	158	45
TAPs	**9557**	



Lang, D., B. Weiche, G. Timmerhaus, S. Richardt, D.M. Riaño-Pachón, L.G.G. Corrêa, R. Reski, B. Mueller-Roeber, S.A. Rensing (2010): Genome-wide phylogenetic comparative analysis of plant transcriptional regulation: a timeline of loss, gain, expansion and correlation with complexity Genome Biology and Evolution 2, 488-503.

Homolog relationships / **ortholog and homeolog** inference

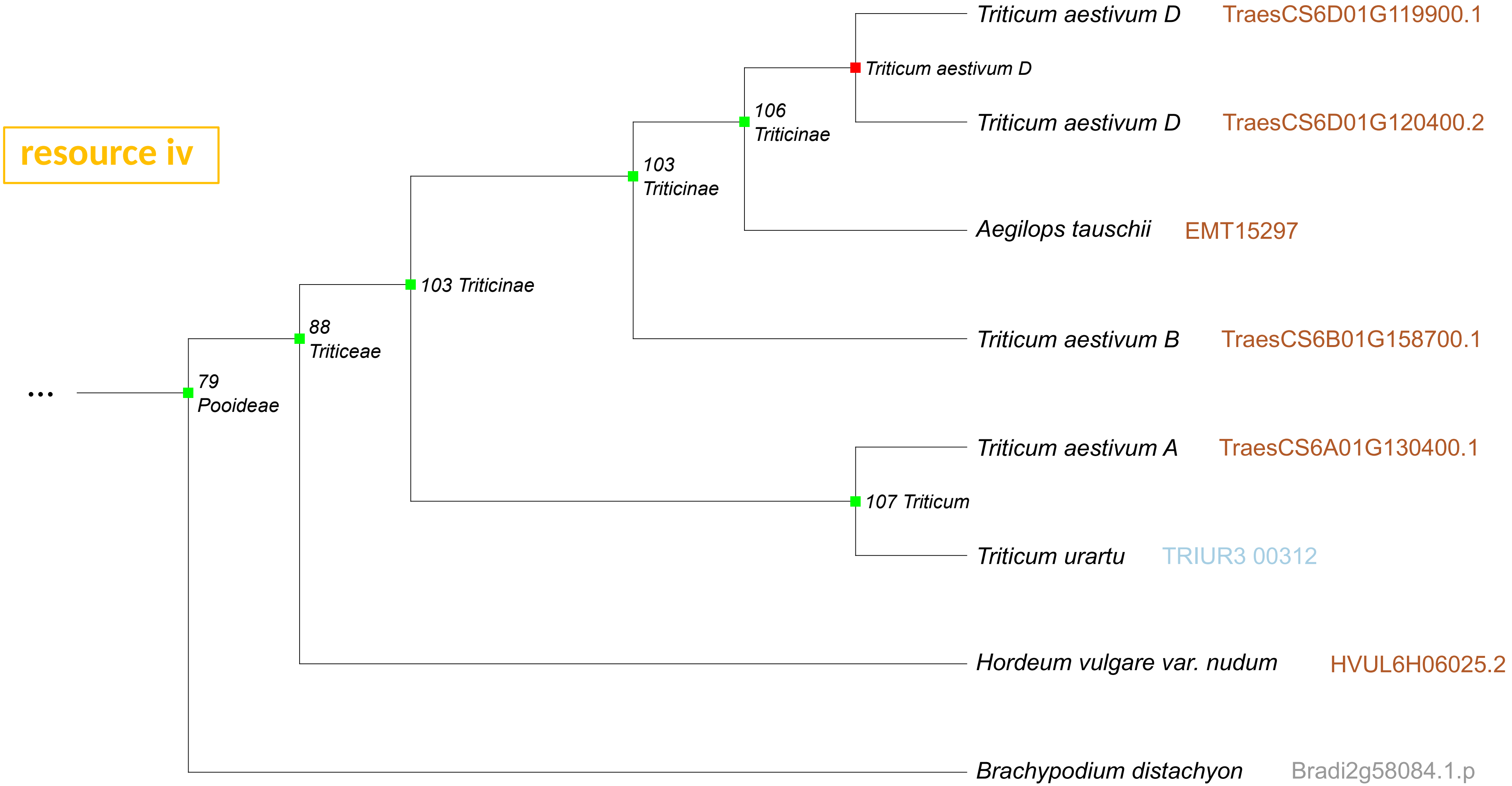
- reconciliation of gene trees with the species using Species Overlap algorithm
  - infer **speciation** and **duplication** events → **ortholog**, **inparalog** and **outparalog** relationships
- wheat subgenomes treated as independent taxa
  - subgenome orthologs → **homeologs**
- use colinearity, chromosomal ancestry and position to assess translocations/transduplications

class	pairs
homeolog	123,588
inparalog	1,797,522
ortholog	2,555,680
outparalog	59,046,632

resource i

Example of an annotated **reconciled gene tree** of a Pooideae-specific subfamily

resource iv


Triticum aestivum D TraesCS6D01G119900.1
Triticum aestivum D
106 Triticinae
Triticum aestivum D TraesCS6D01G120400.2
103 Triticinae
Aegilops tauschii EMT15297
103 Triticinae
88 Triticeae
Triticum aestivum B TraesCS6B01G158700.1
...
79 Pooideae
Triticum aestivum A TraesCS6A01G130400.1
107 Triticum
Triticum urartu TRIUR3 00312
Hordeum vulgare var. nudum HVUL6H06025.2
Brachypodium distachyon Bradi2g58084.1.p

Orthology-based **ontology annotation** for wheat V1

- concept: transfer of ontology term annotation from orthologs
  - rich data from Gene Ontology consortium, TAIR, Gramene/Planteome
- Gene Ontology (**GO**), Plant Ontology (**PO**), Trait Ontology (**TO**)
- evidence quality:
  - transfer: IEA =IEA, all other: ISO (inferred by sequence orthology)
  - pfam2GO/domain architecture: ISM (inferred from sequence model)
- sequence similarity-based annotation using AHRD (only IEA):
  - 120,185 genes
  - 1,820 GO terms
- orthology-based yields more and more specific annotations:

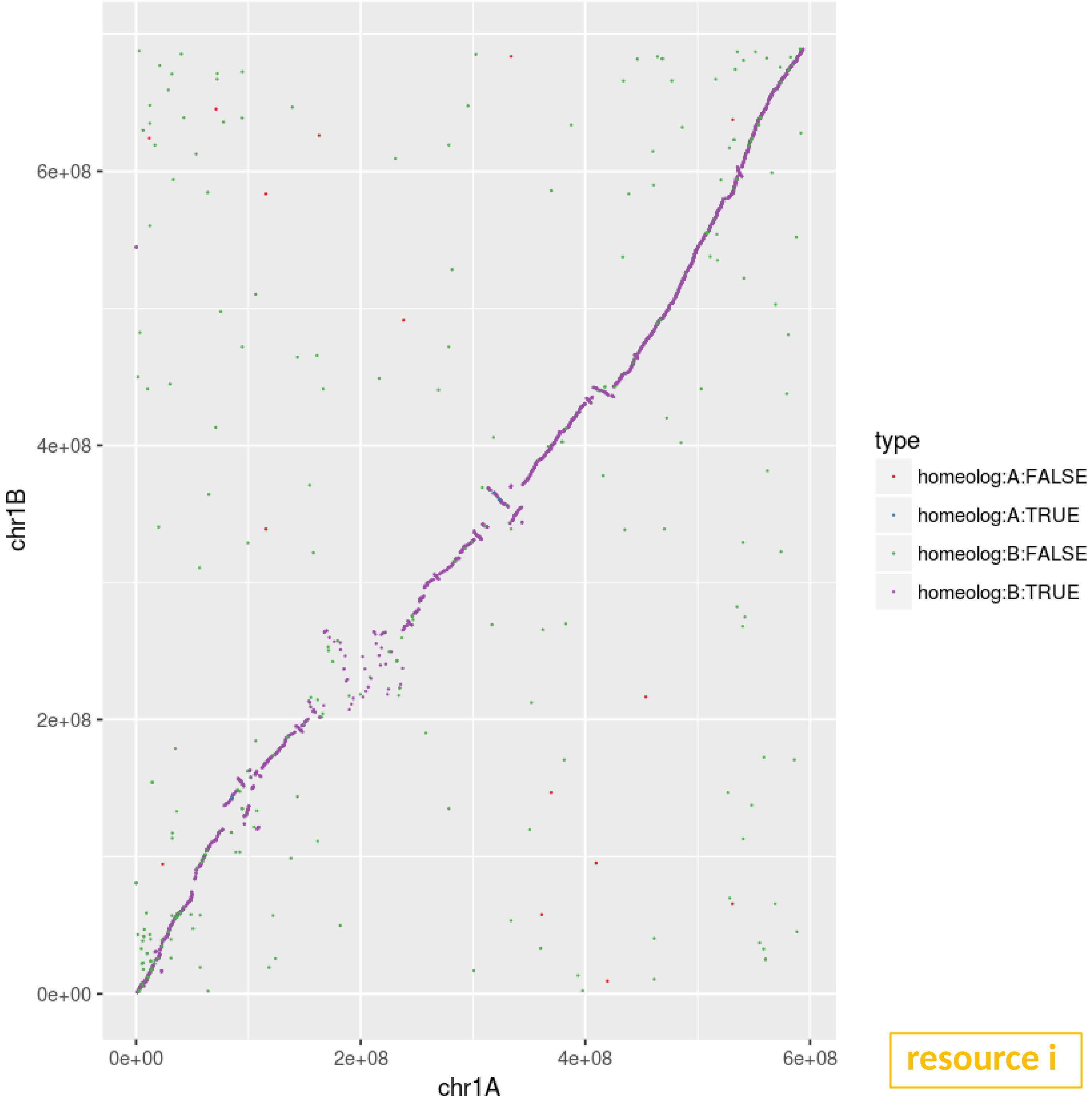
ontology	nassoc	nfamilies	nseq	nterms
PO	3,622,724	12,631	79,530	446
GO	1,560,102	15,983	113,815	**7,408**
TO	8,173	227	1,060	279

evidence_code	count
ISO	**4,989,419**
IEA	177,154
ISM	155,326

resource ii

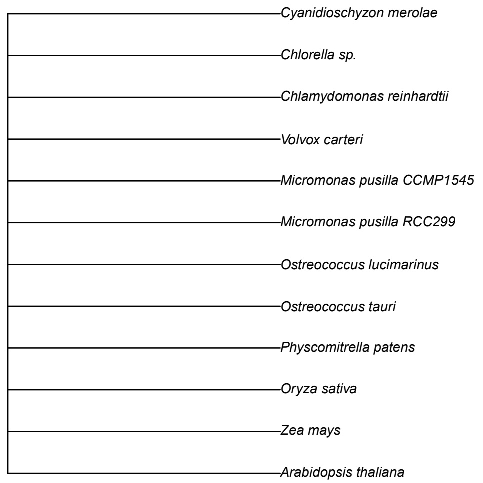
Genome-scale resolution and **classification of homeologs:**

syntenic (TRUE) ↔ translocated/-duplicated (FALSE)



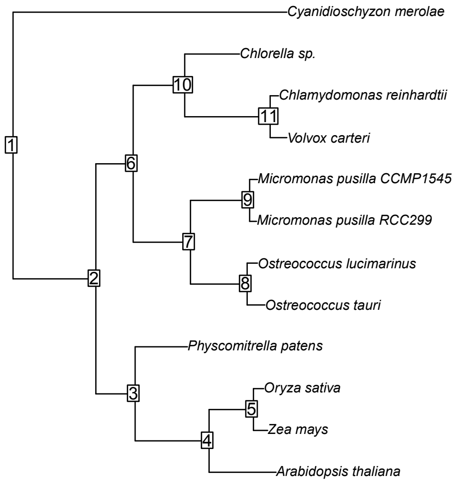
resource i

Any **comparative analysis** among species is affected by phylogenetic relationships



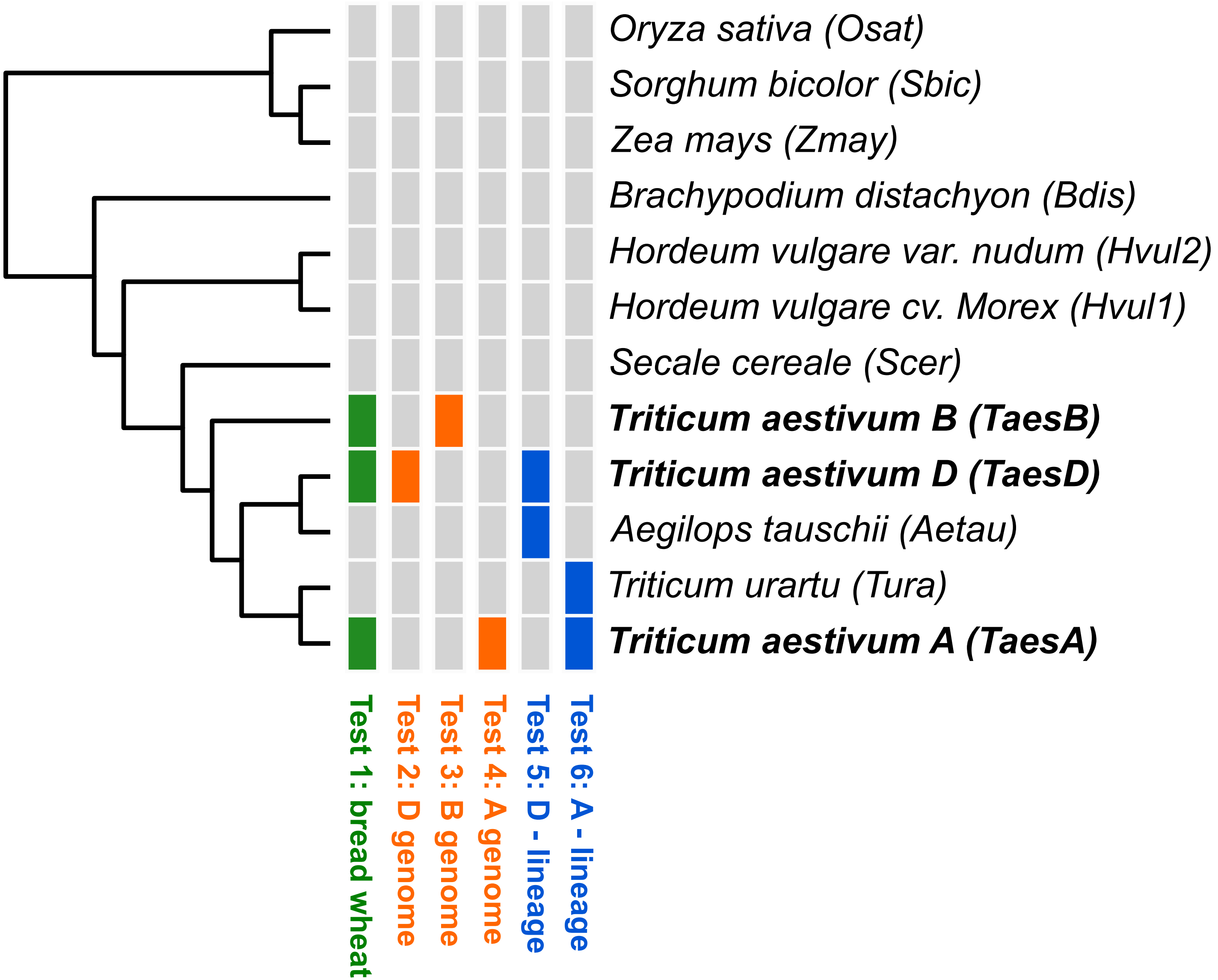
✗

- correct for phylogenetic dependence with **phylogenetic comparative (PC) methods**
  - developed for morphological traits
  - e.g. Felsenstein's PIC 1985



✓

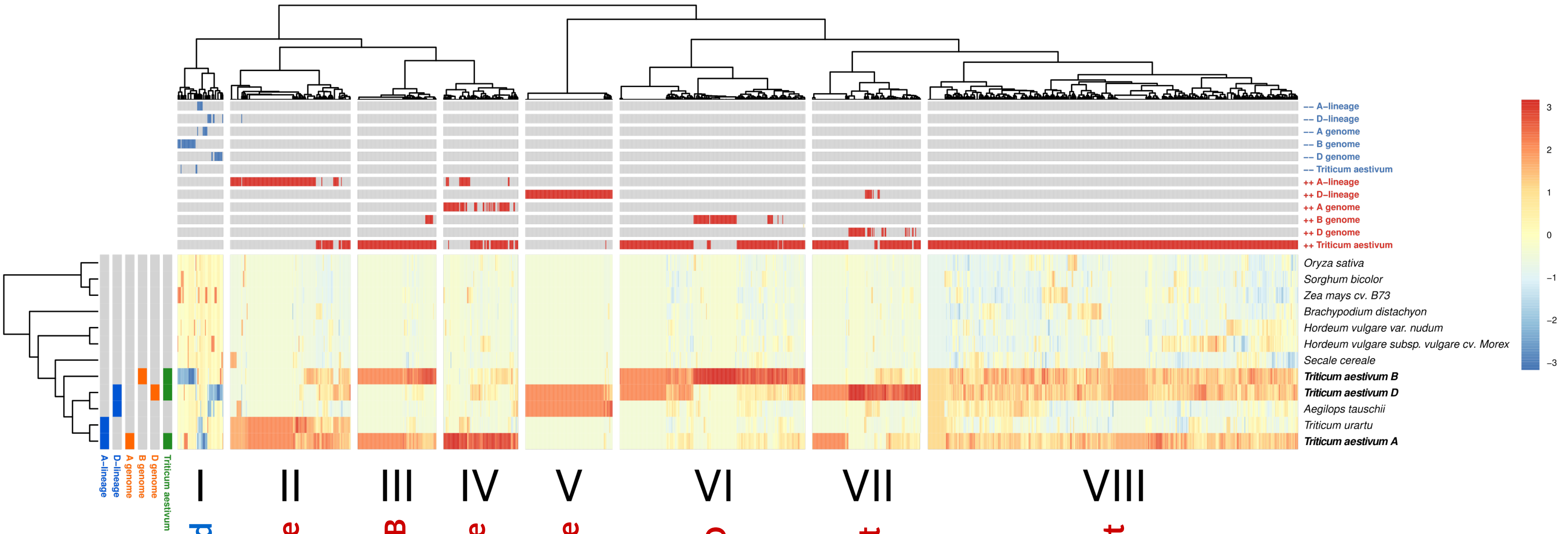
Phylogenetic one-way ANOVA testing for **Gene Copy Number Variations (GCNVs)** in these sets



Oryza sativa (Osat)
Sorghum bicolor (Sbic)
Zea mays (Zmay)
Brachypodium distachyon (Bdis)
Hordeum vulgare var. nudum (Hvul2)
Hordeum vulgare cv. Morex (Hvul1)
Secale cereale (Scer)
Triticum aestivum B (TaesB)
Triticum aestivum D (TaesD)
Aegilops tauschii (Aetau)
Triticum urartu (Tura)
Triticum aestivum A (TaesA)
Test 1: bread wheat
Test 2: D genome
Test 3: B genome
Test 4: A genome
Test 5: D - lineage
Test 6: A - lineage

resource v

GCNVs in gene families shaping the traits that make bread wheat



I -- contracted

II ++ A-lineage

III ++ A & B

IV ++ A genome

V ++ D lineage

VI ++ B & D

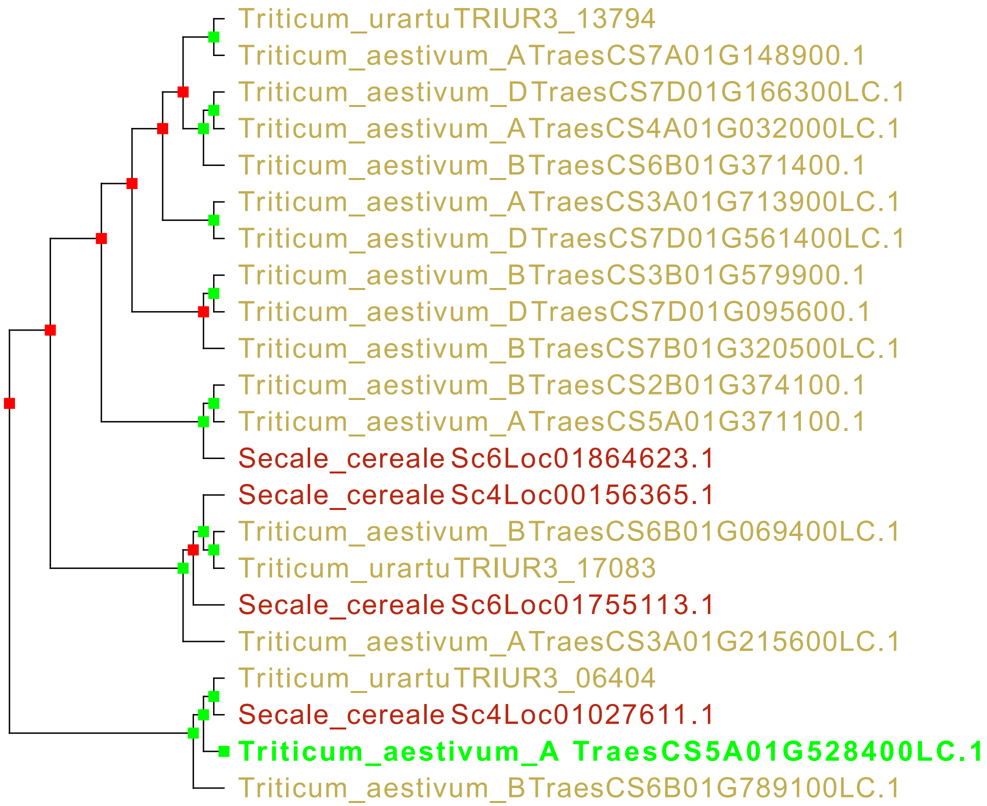
VII ++ D & bread wheat

VIII ++ bread wheat

resource v

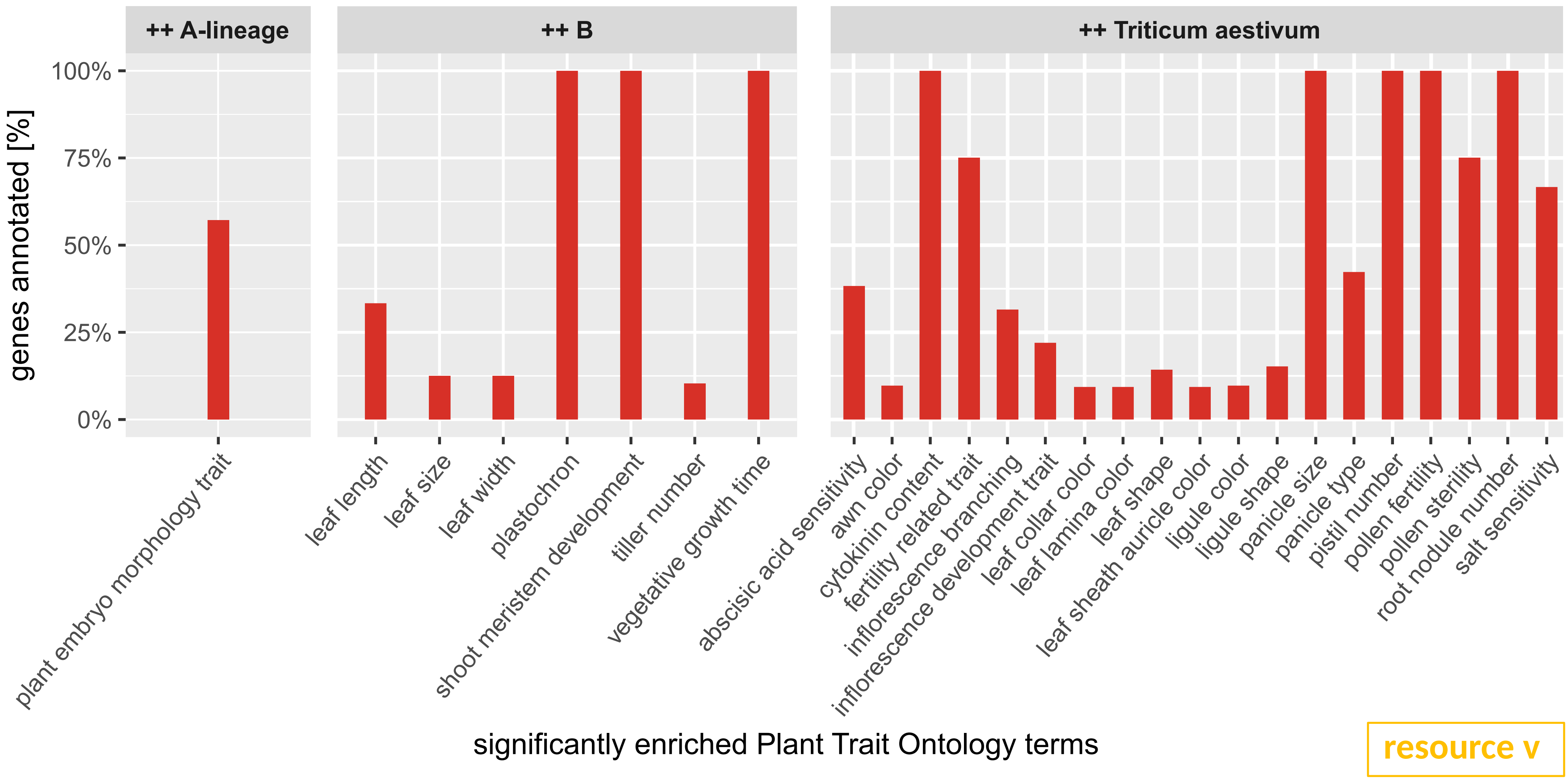
Wheat **GCNVs**: rich resource to find novel/interesting **traits<->genes**!

- **all expert-curated gene families** are in expansions e.g. prolamins, NLRs, HMW glutenin…
- Overlap with mapped **QTLs**:
  - IWGSC/Salse collection: 22, families, 25 genes, 23 marker, 62 QTLs
  - comprise at least 3 connections to literature:
    - FAR1-like → Earliness
    - (u)DENN → senescence
    - Sulfate_transp::STAS → patent: enhanced traits in wheat
  - QTLs mapped to orthologs in other species → Plant **T**rait **O**ntology (**TO**)


Triticum_urartu TRIUR3_13794
Triticum_aestivum_A TraesCS7A01G148900.1
Triticum_aestivum_D TraesCS7D01G166300LC.1
Triticum_aestivum_A TraesCS4A01G032000LC.1
Triticum_aestivum_B TraesCS6B01G371400.1
Triticum_aestivum_A TraesCS3A01G713900LC.1
Triticum_aestivum_D TraesCS7D01G561400LC.1
Triticum_aestivum_B TraesCS3B01G579900.1
Triticum_aestivum_D TraesCS7D01G095600.1
Triticum_aestivum_B TraesCS7B01G320500LC.1
Triticum_aestivum_B TraesCS2B01G374100.1
Triticum_aestivum_A TraesCS5A01G371100.1
Secale_cereale Sc6Loc01864623.1
Secale_cereale Sc4Loc00156365.1
Triticum_aestivum_B TraesCS6B01G069400LC.1
Triticum_urartu TRIUR3_17083
Secale_cereale Sc6Loc01755113.1
Triticum_aestivum_A TraesCS3A01G215600LC.1
Triticum_urartu TRIUR3_06404
Secale_cereale Sc4Loc01027611.1
Triticum_aestivum_A TraesCS5A01G528400LC.1
Triticum_aestivum_B TraesCS6B01G789100LC.1

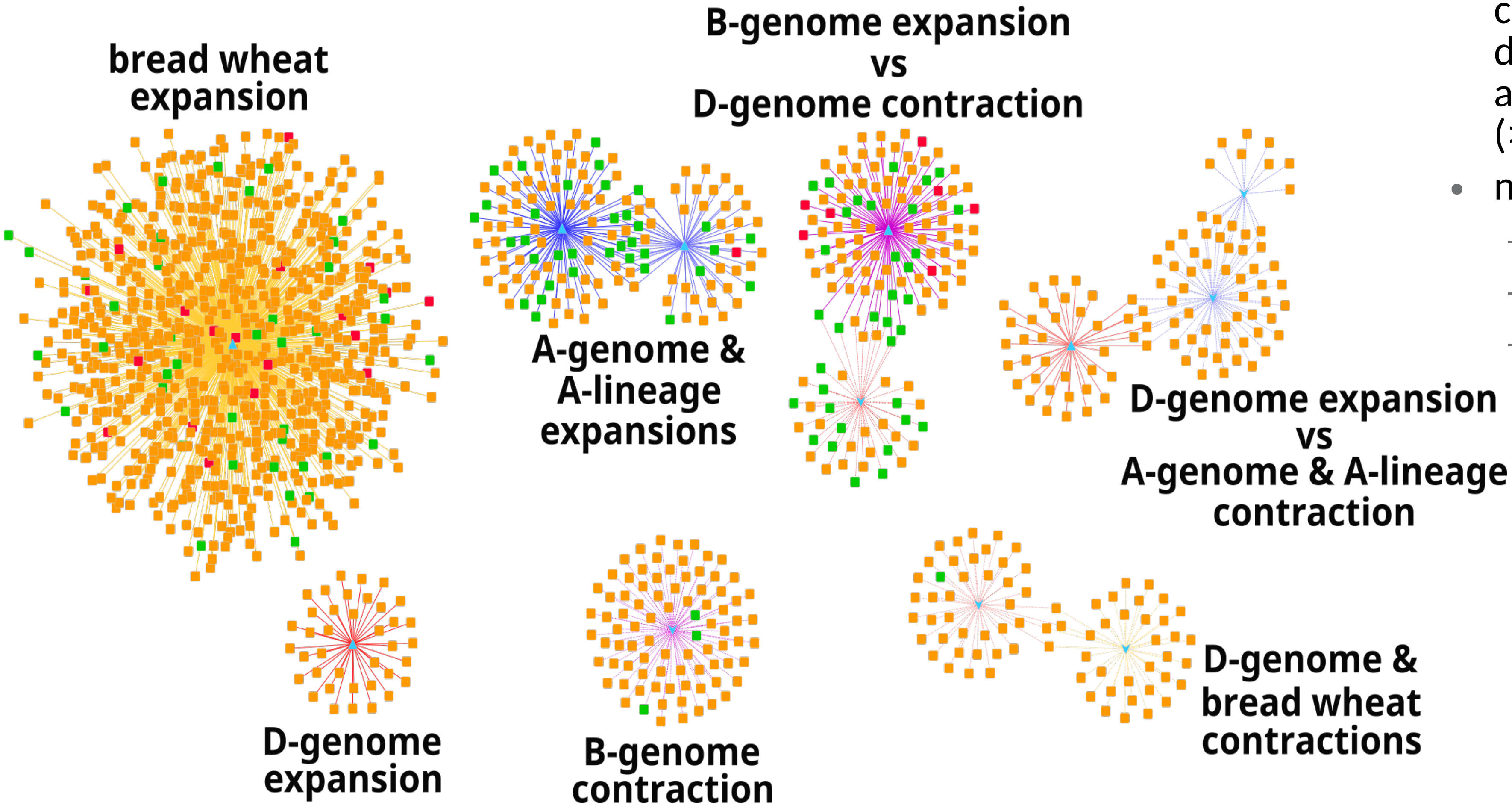
resource v

The **expanded families** are enriched with previously mapped, orthologous **plant traits** (**TO**)



resource v

Clustering of functional networks of enriched ontology terms: **GCNVs** affected broad range of **functions** hinting at subgenome(-lineage)-specific adaptations or subgenomic gene flow



bread wheat expansion

B-genome expansion vs D-genome contraction

A-genome & A-lineage expansions

D-genome expansion vs A-genome & A-lineage contraction

D-genome expansion

B-genome contraction

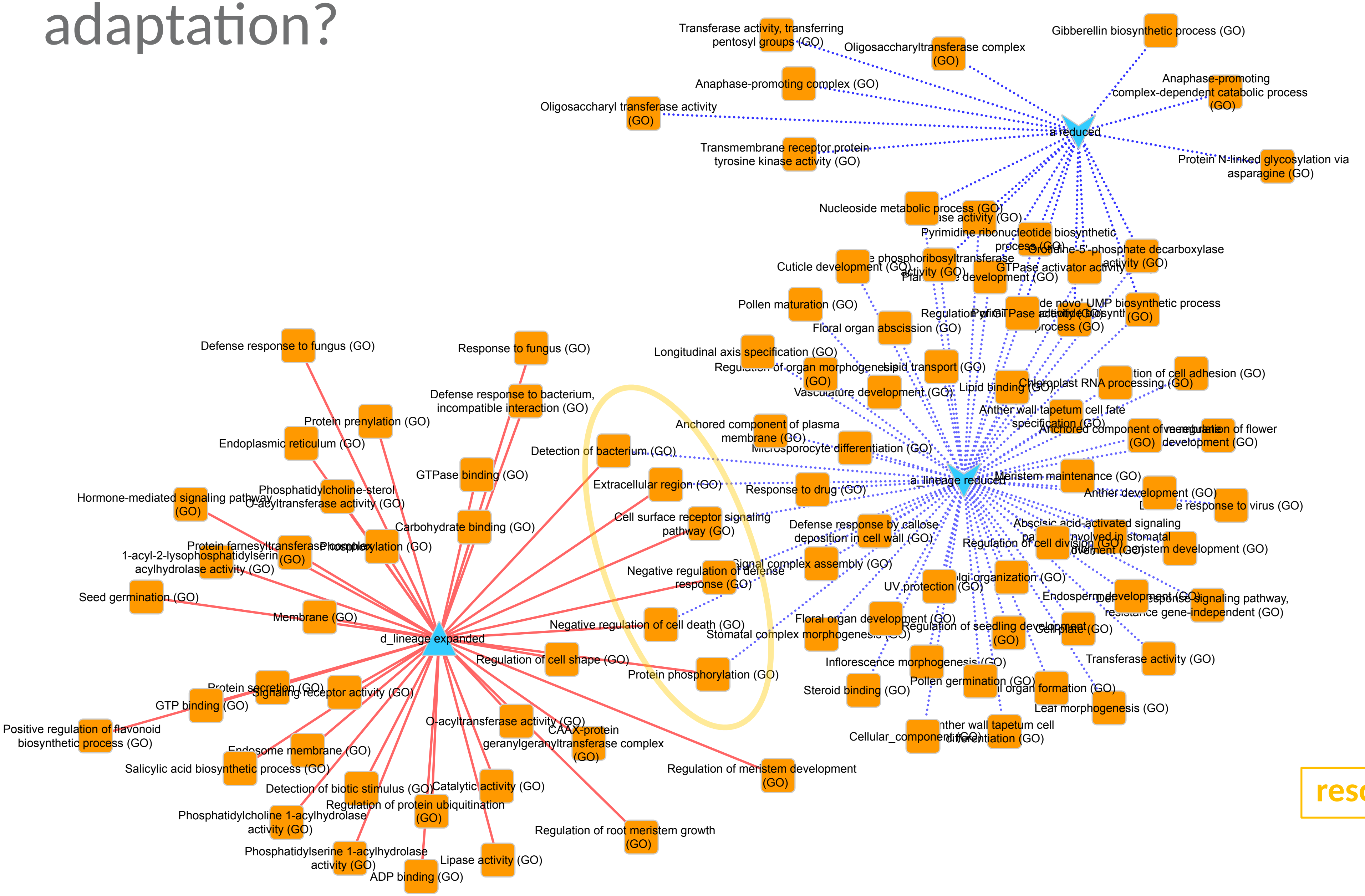
D-genome & bread wheat contractions

- based on FDR<0.1 sets for gene families consistent domain architectures (>50%)
- node colors:
  - GO orange
  - PO green
  - TO red

resource v

GCNVs: Defense responses expanded in D-lineage and reduced in A-lineage

Subgenomic gene flow or lineage-specific adaptation?



Transferase activity, transferring pentosyl groups (GO)
Oligosaccharyltransferase complex (GO)
Gibberellin biosynthetic process (GO)
Anaphase-promoting complex (GO)
Anaphase-promoting complex-dependent catabolic process (GO)
Oligosaccharyl transferase activity (GO)
a reduced
Transmembrane receptor protein tyrosine kinase activity (GO)
Protein N-linked glycosylation via asparagine (GO)
Nucleoside metabolic process (GO)
ase activity (GO)
Pyrimidine ribonucleotide biosynthetic process (GO)
Orotidine-5'-phosphate decarboxylase activity (GO)
Cuticle development (GO)
e phosphoribosyltransferase activity (GO)
GTPase activator activity
development (GO)
Pollen maturation (GO)
'de novo' UMP biosynthetic process (GO)
Regulation of GTPase activity (GO)
process (GO)
Floral organ abscission (GO)
Longitudinal axis specification (GO)
Regulation of organ morphogenesis (GO)
Lipid transport (GO)
Vasculature development (GO)
Lipid binding (GO)
Chloroplast RNA processing (GO)
tion of cell adhesion (GO)
Anther wall tapetum cell fate specification (GO)
Anchored component of plasma membrane (GO)
Anchored component of membrane (GO)
Regulation of flower development (GO)
Microsporocyte differentiation (GO)
Meristem maintenance (GO)
a_lineage reduced
Response to drug (GO)
Anther development (GO)
e response to virus (GO)
Defense response by callose deposition in cell wall (GO)
Abscisic acid-activated signaling pathway involved in stomatal movement (GO)
Regulation of cell division (GO)
Meristem development (GO)
Signal complex assembly (GO)
lgi organization (GO)
UV protection (GO)
Endosperm development (GO)
Defense response signaling pathway, resistance gene-independent (GO)
Floral organ development (GO)
Stomatal complex morphogenesis (GO)
Regulation of seedling development (GO)
Cell plate (GO)
Inflorescence morphogenesis (GO)
Transferase activity (GO)
Steroid binding (GO)
Pollen germination (GO)
l organ formation (GO)
Leaf morphogenesis (GO)
Cellular_component (GO)
nther wall tapetum cell differentiation (GO)
Defense response to fungus (GO)
Response to fungus (GO)
Defense response to bacterium, incompatible interaction (GO)
Protein prenylation (GO)
Endoplasmic reticulum (GO)
Detection of bacterium (GO)
GTPase binding (GO)
Extracellular region (GO)
Hormone-mediated signaling pathway (GO)
Phosphatidylcholine-sterol O-acyltransferase activity (GO)
Carbohydrate binding (GO)
Cell surface receptor signaling pathway (GO)
Protein farnesyltransferase complex
Phosphorylation (GO)
1-acyl-2-lysophosphatidylserine acylhydrolase activity (GO)
Negative regulation of defense response (GO)
Seed germination (GO)
Membrane (GO)
Negative regulation of cell death (GO)
d_lineage expanded
Regulation of cell shape (GO)
Protein phosphorylation (GO)
Protein secretion (GO)
Signaling receptor activity (GO)
GTP binding (GO)
Positive regulation of flavonoid biosynthetic process (GO)
O-acyltransferase activity (GO)
CAAX-protein geranylgeranyltransferase complex (GO)
Endosome membrane (GO)
Salicylic acid biosynthetic process (GO)
Regulation of meristem development (GO)
Detection of biotic stimulus (GO)
Catalytic activity (GO)
Regulation of protein ubiquitination (GO)
Phosphatidylcholine 1-acylhydrolase activity (GO)
Regulation of root meristem growth (GO)
Phosphatidylserine 1-acylhydrolase activity (GO)
Lipase activity (GO)
ADP binding (GO)

resource v

Phylogenomics/annotation resources published with the **bread wheat V1 genome papers**

The tragic ripples of an epic fraud p. 636
Insect pest profits from maize defenses pp. 642 & 694
Photoredox activation of methane pp. 647 & 668
Science
$15
17 AUGUST 2018
sciencemag.org
AAAS
ROAD MAP FOR
WHEAT
Ordered sequence will speed research pp. 635, 661, & 662

- Resources:
  1. classified homeologs
  2. orthology-based GOA|POA|TOA
  3. wheat Transcription Associated Proteins (TAPs=TF+TR)
  4. wheat gene families including orthologs from other Viridiplantae
  5. gene families with significant GCNVs
- all downloadable → links also in main paper supplement
  - https://wheat-urgi.versailles.inra.fr/Seq-Repository/Annotations
  - http://dx.doi.org/10.5447/IPK/2018/5
  - http://doi.org/10.1126/science.aar7191
- relevant resources were highlighted on the slides: **resource X**

Thank you!

- everyone @PGSB
- IWGSC cooperators
  - especially the comparative and gene family working groups
- for your attention :-)