THE UNIVERSITY OF WESTERN AUSTRALIA

Centre for Applied Bioinformatics

Pangenomics and machine learning for disease resistance

Dave Edwards

Director, Centre for Applied Bioinformatics
University of Western Australia

Dave.Edwards@uwa.edu.au

THE UNIVERSITY OF WESTERN AUSTRALIA

2023

Centre for Applied Bioinformatics

- Illumina
- Good, cheap and short

- PacBio Sequel
- Long HiFi reads

- Oxford nanopore
- Long reads

Sequence genomes

THE UNIVERSITY OF WESTERN AUSTRALIA

Centre for Applied Bioinformatics

The genome of the mesopolyploid crop species *Brassica rapa*

A reference genome for pea provides insight into legume genome evolution.

A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome

Studying the genetic diversity of yam bean using a new draft genome assembly.

Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome

The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes

Assembly of the non-heading pak choi genome and comparison with the genomes of heading Chinese cabbage and the oilseed yellow sarson.

Dispersion and domestication shaped the genome of bread wheat

Assembly and comparison of two closely related *Brassica napus* genomes.

Shifting the limits in wheat research and breeding using a fully annotated reference genome.

Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement

The genome of a southern hemisphere seagrass species (*Zostera muelleri*)[1]

Sequencing and assembly of low copy and genic regions of isolated *Triticum aestivum* chromosome arm 7DS

A comprehensive draft genome sequence for lupin (*Lupinus angustifolius*), an emerging health food: Insights into plant-microbe interactions and legume evolution.

The improved assembly of 7DL chromosome provides insight into the structure and evolution of bread wheat.

Draft genome assembly and transcriptome dataset for European turnip (*Brassica rapa* L. ssp. rapifera), ECD4 carrying clubroot resistance

The genome sequence of the Antarctic bullhead notothen reveals evolutionary adaptations to a cold environment

THE UNIVERSITY OF WESTERN AUSTRALIA

Pangenomes are the new reference

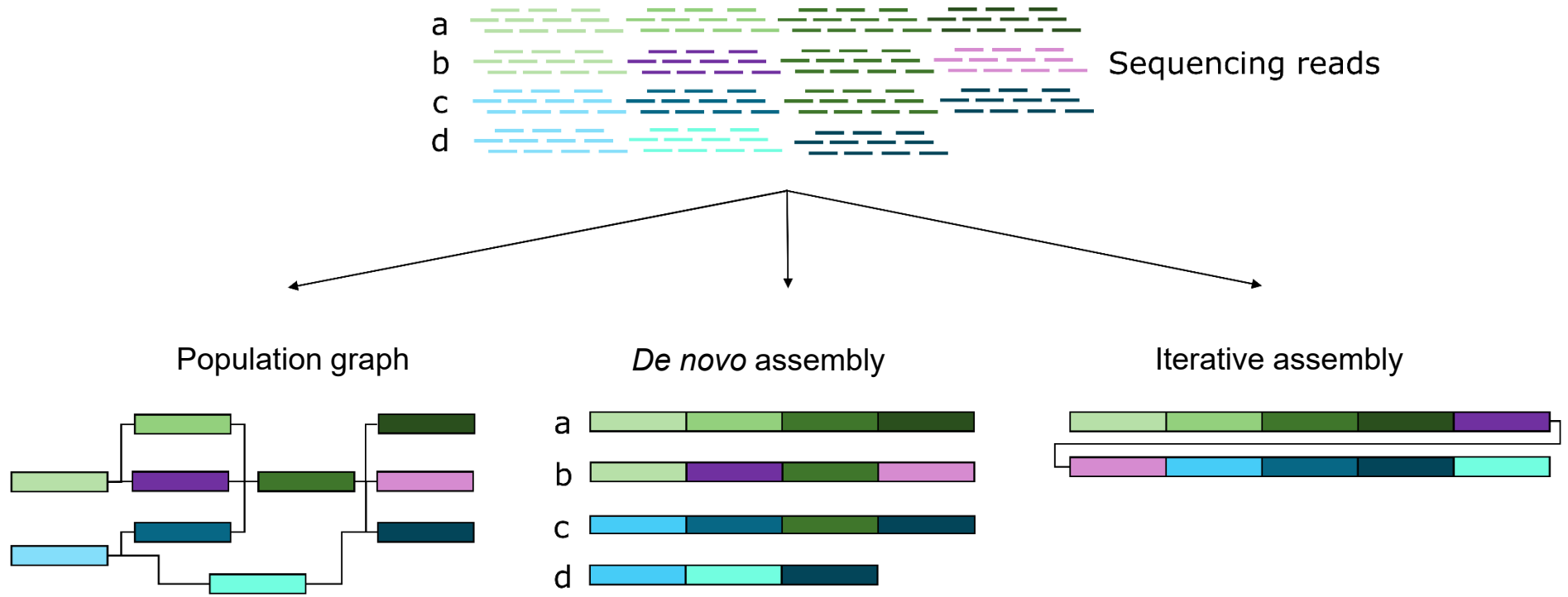
Centre for Applied Bioinformatics

- A single reference genome does not represent the diversity of a species
- PAV genes are responsible for important agronomic traits
- Pangenomes capture heritability missing in single genomes
- Need to know gene content for genome editing

THE UNIVERSITY OF WESTERN AUSTRALIA

Building a pangenome

Centre for Applied Bioinformatics



THE UNIVERSITY OF WESTERN AUSTRALIA

Brassica oleracea pangenome

Centre for Applied Bioinformatics

ARTICLE

Received 22 Aug 2016 | Accepted 28 Sep 2016 | Published 11 Nov 2016

DOI: 10.1038/ncomms13390 OPEN

The pangenome of an agronomically important crop plant *Brassica oleracea*

Agnieszka A. Golicz[1], Philipp E. Bayer[2], Guy C. Barker[3], Patrick P. Edger[4], HyeRan Kim[5], Paula A. Martinez[1], Chon Kit Kenneth Chan[2], Anita Severn-Ellis[2], W. Richard McCombie[6], Isobel A.P. Parkin[7], Andrew H. Paterson[8], J. Chris Pires[9], Andrew G. Sharpe[10], Haibao Tang[11], Graham R. Teakle[3], Christopher D. Town[12], Jacqueline Batley[2] & David Edwards[2]

Diverse morphotypes

- Cabbage (2)
- Cauliflower (2)
- Broccoli
- Kale
- Brussels sprout
- Rapid cycler TO1000
- *B. macrocarpa*

THE UNIVERSITY OF WESTERN AUSTRALIA

Brassica oleracea pangenome

Centre for Applied Bioinformatics

Previous reference genomes:

B. oleracea TO1000	54,458 genes, 488 Mbp
B. oleracea var. capitata	45,758 genes, 535 Mbp

Pangenome	61,379 genes, 587 Mbp,

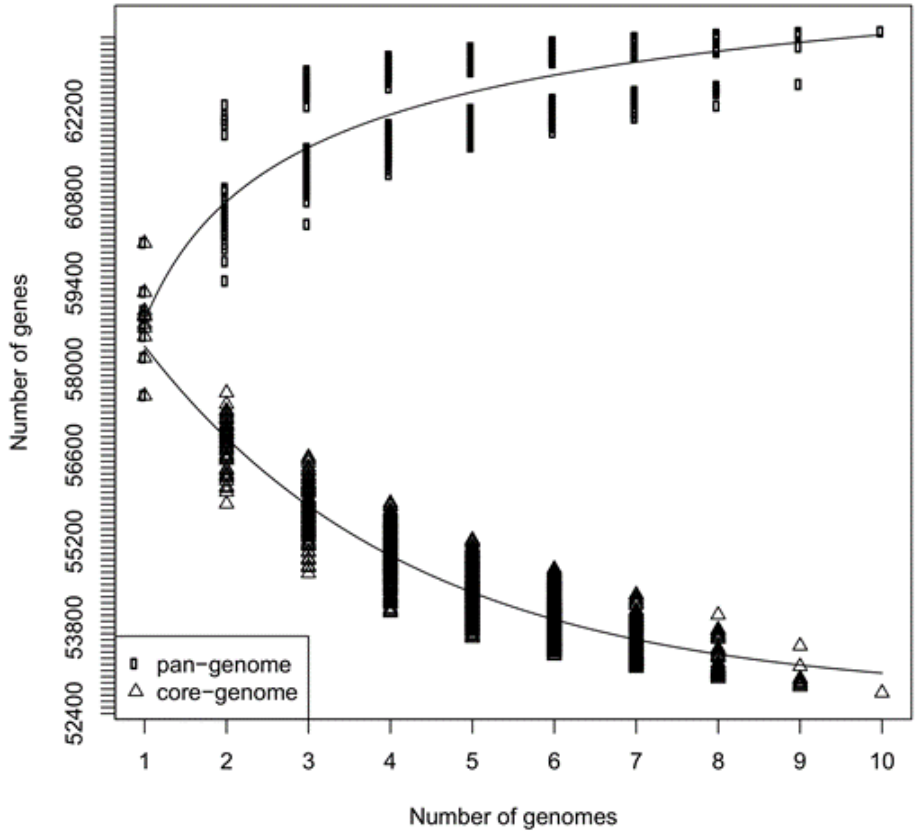
18.7% of genes are variable

THE UNIVERSITY OF WESTERN AUSTRALIA

Brassica oleracea pangenome

Centre for Applied Bioinformatics

Adding more genomes captures more genes and defined the core and variable genome



Brassica oleracea

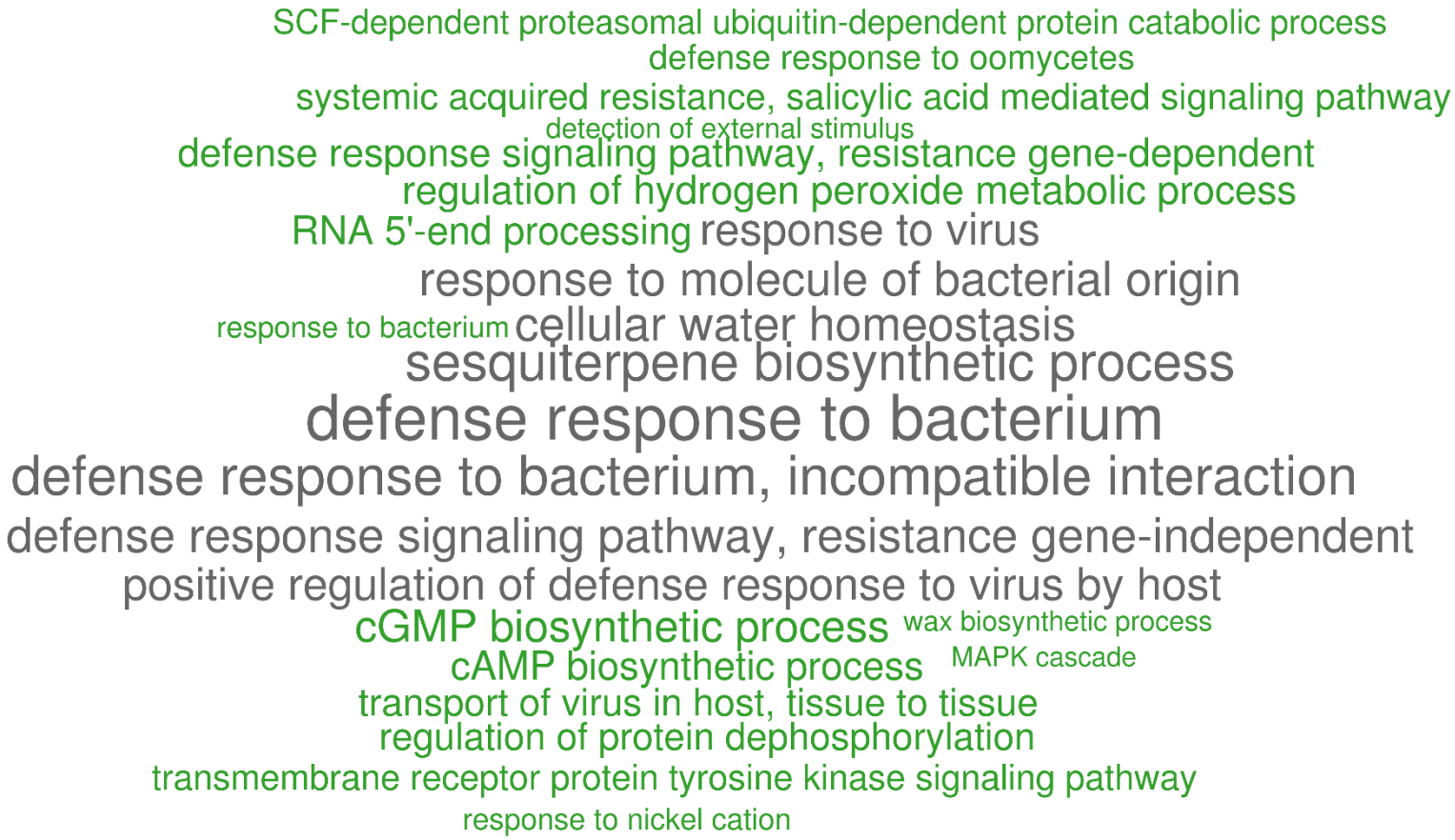
Predicted pangenome of 61,198 ± 394 genes (35,462 ± 250 gene families)

Predicted core genome size of 49,676 ± 96 genes (28,489 ± 51 gene families)

THE UNIVERSITY OF WESTERN AUSTRALIA

Brassica oleracea pangenome

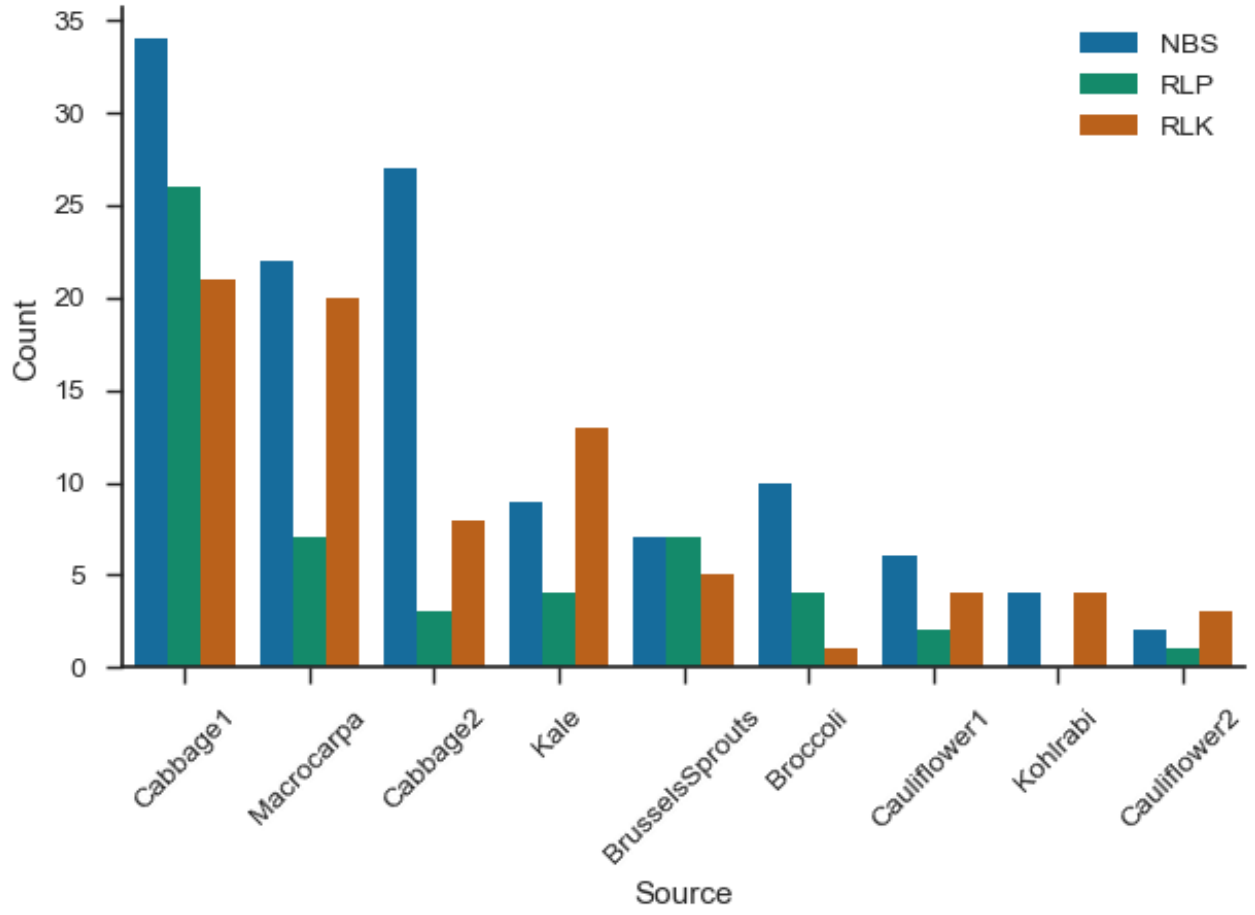
Centre for Applied Bioinformatics



THE UNIVERSITY OF WESTERN AUSTRALIA

Disease resistance genes

Centre for Applied Bioinformatics



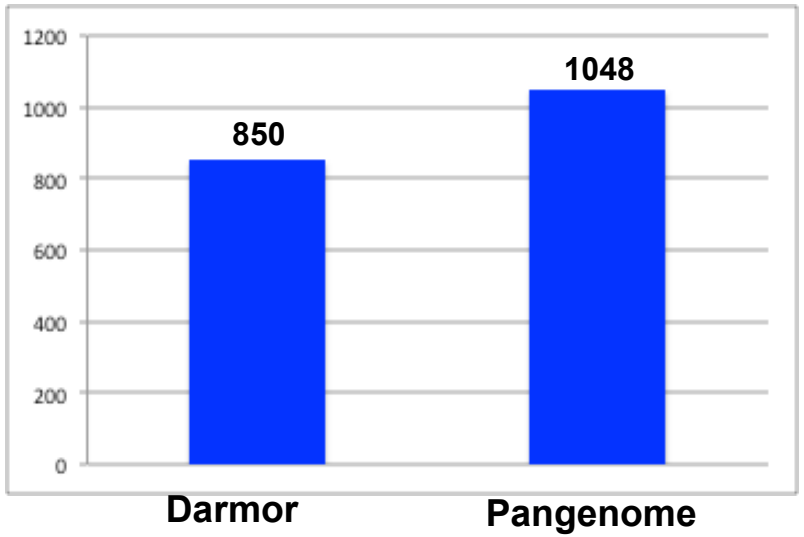
THE UNIVERSITY OF WESTERN AUSTRALIA

Brassica napus

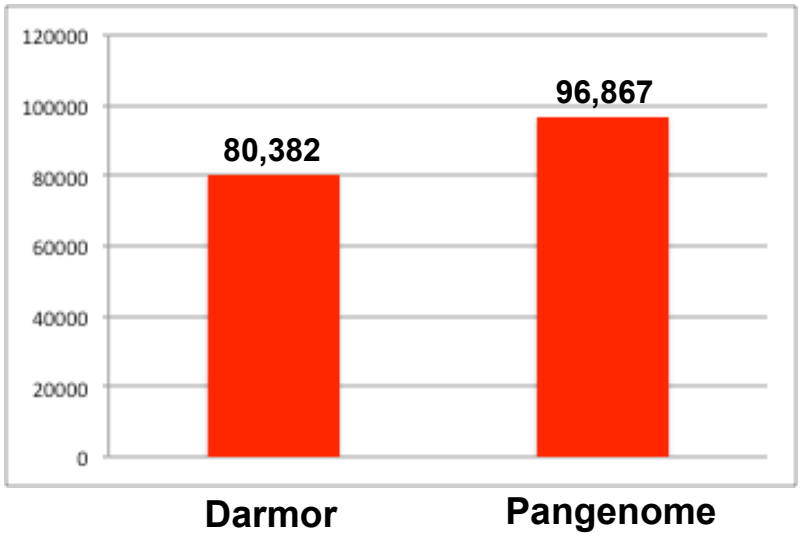
Centre for Applied Bioinformatics

- 33 *B. napus*
- 20 resynthesised *B. napus*

Assembly size (Mb)



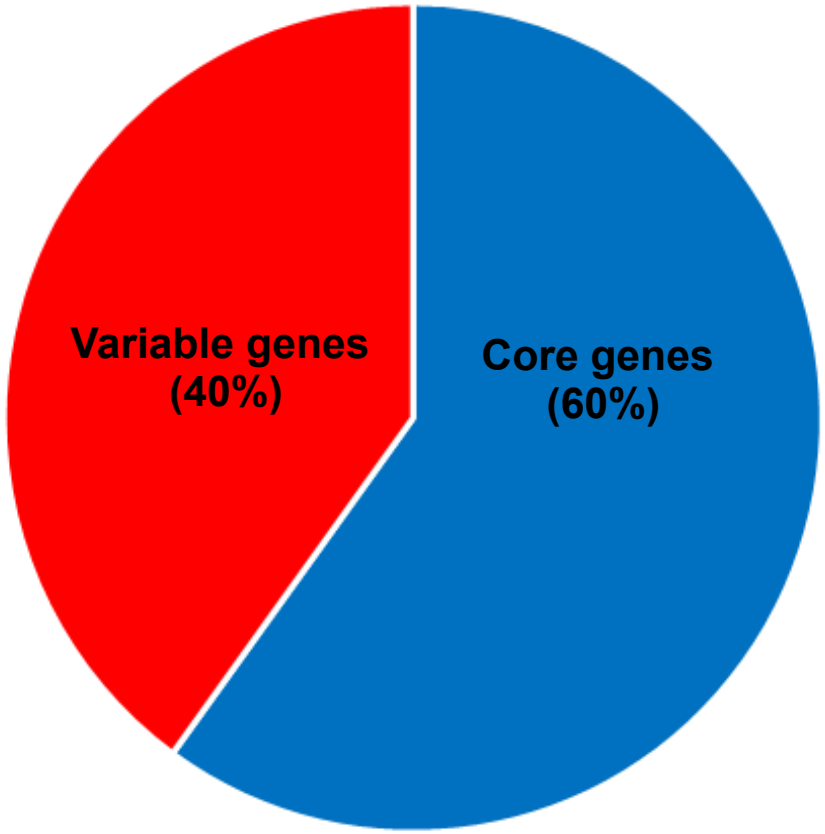
No. of genes



THE UNIVERSITY OF WESTERN AUSTRALIA

Brassica napus

Centre for Applied Bioinformatics

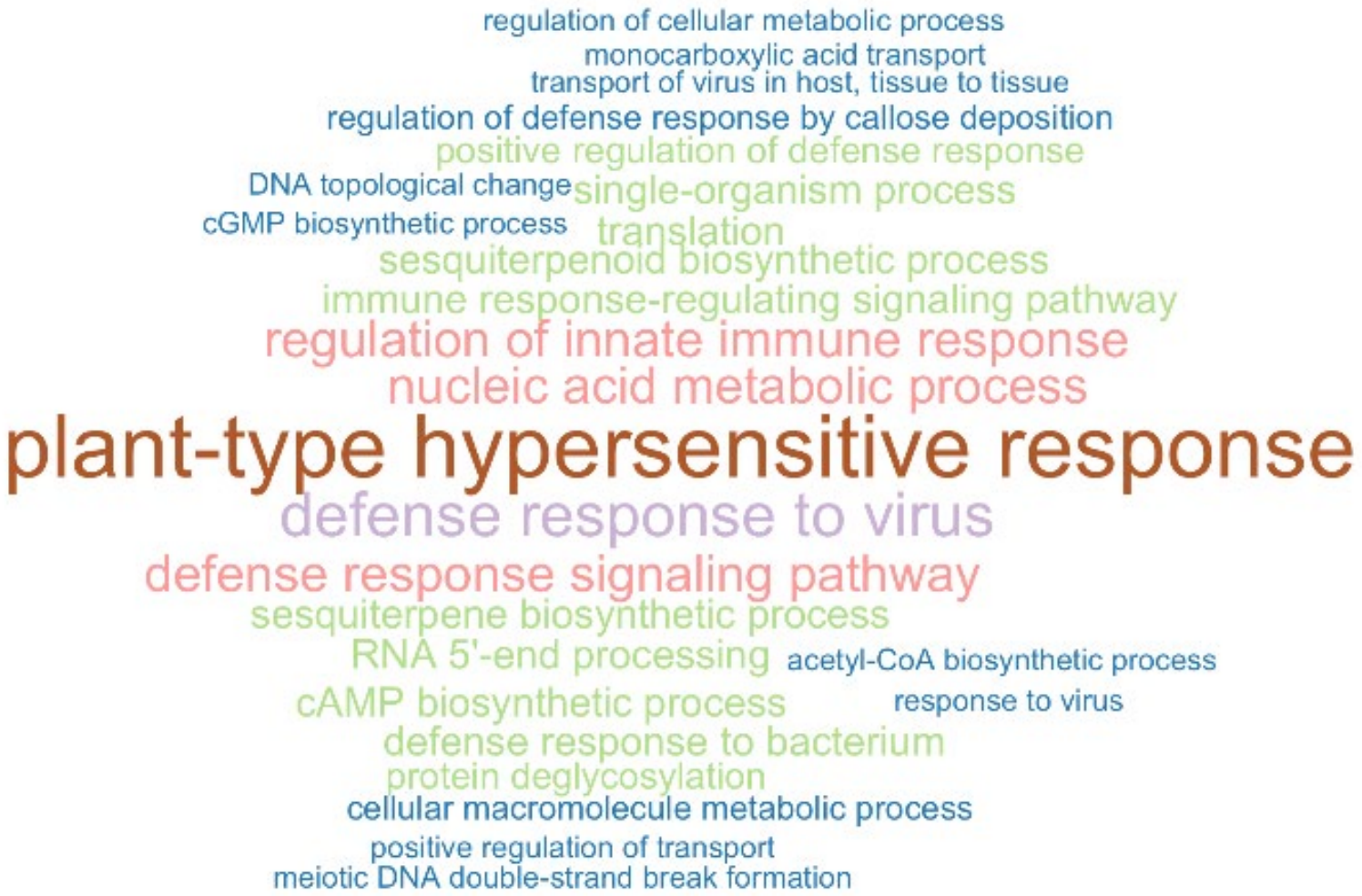


- Variable genes are shorter and have fewer exons than core genes
- Number of variable genes indicates how much diversity is present

THE UNIVERSITY OF WESTERN AUSTRALIA

Brassica napus

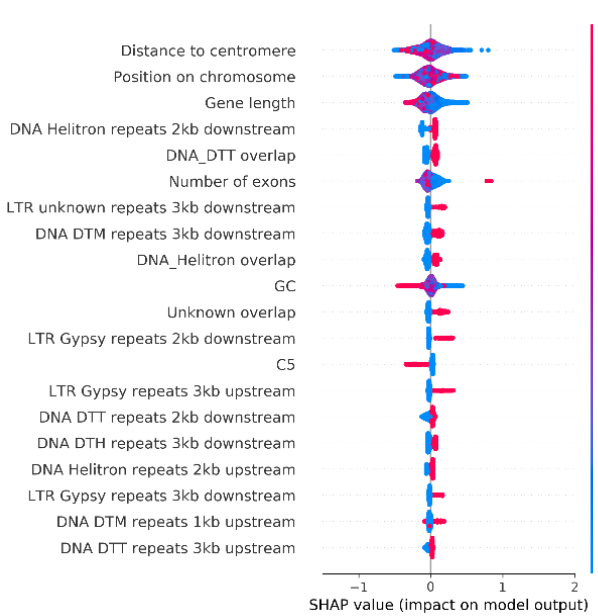
Centre for Applied Bioinformatics



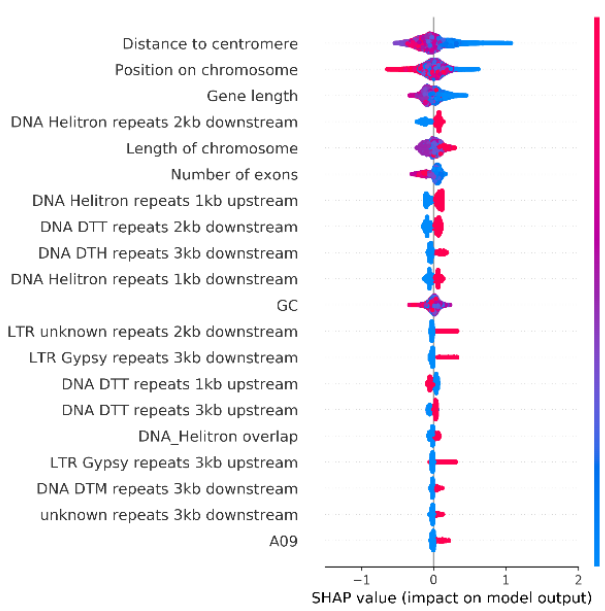
THE UNIVERSITY OF WESTERN AUSTRALIA

Mechanism of PAV

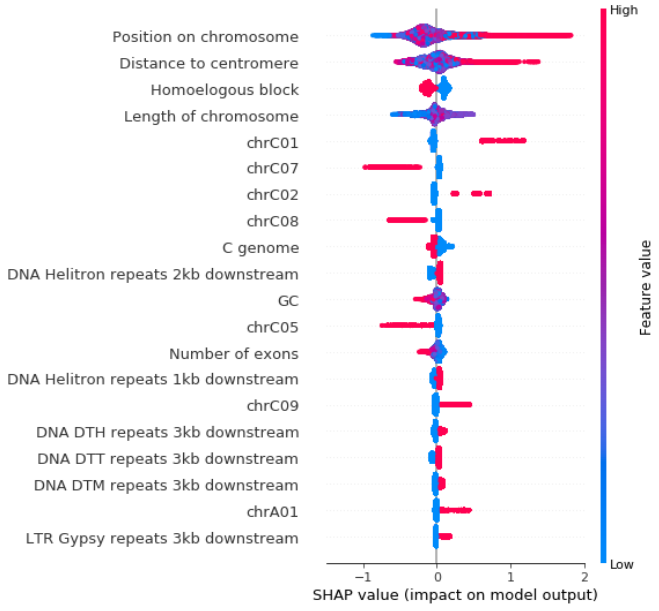
Centre for Applied Bioinformatics



B. oleracea



B. rapa



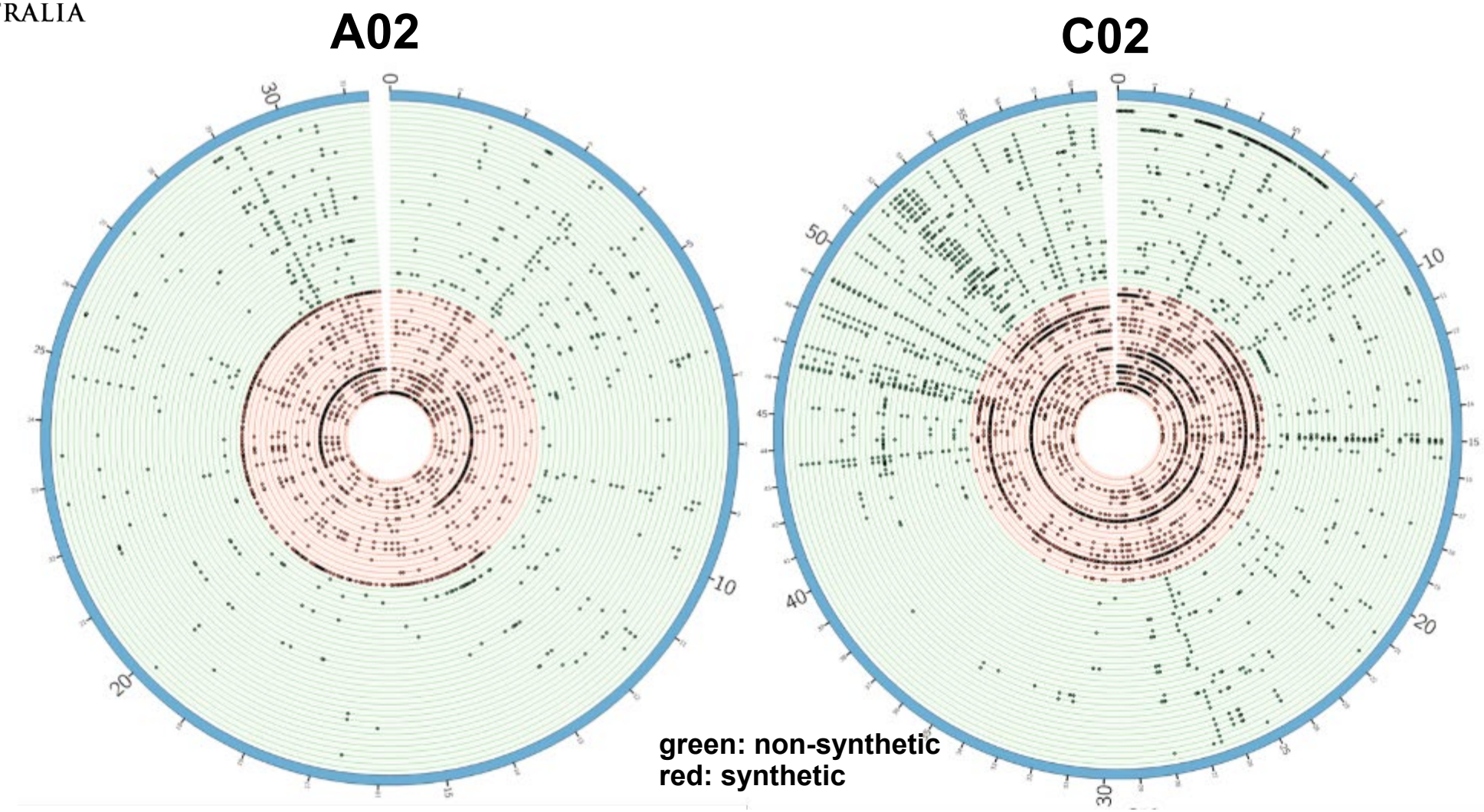
B. napus

Bayer et al. (2021) Modelling of gene loss propensity in the pangenomes of three Brassica species suggests different mechanisms between polyploids and diploids. Plant Biotechnology Journal. 19 (12): 2488-2500

THE UNIVERSITY OF WESTERN AUSTRALIA

Brassica napus

Centre for Applied Bioinformatics



THE UNIVERSITY OF WESTERN AUSTRALIA

Bayer PE, Petereit J, Durant E, Monat C, Rouard M, Hu H, Chapman B, Li C, Cheng S, Batley J, Edwards D. (2022) Wheat Panache: A pangenome graph database representing presence–absence variation across sixteen bread wheat genomes. The Plant Genome.

Golicz et al. (2020) Pangenomics comes of age: From bacteria to plant and animal applications. Trends in Genetics 36(2): 132-145

Zanini et al. (2021) Pangenomics in crop improvement – from coding SVs to finding regulatory variants with pangenome graphs. Plant Genome. 15 (1): e20177

Bayer et al. (2021) Modelling of gene loss propensity in the pangenomes of three Brassica species suggests different mechanisms between polyploids and diploids. Plant Biotechnology Journal. Plant Biotechnology Journal. 19 (12): 2488-2500

Varshney et al. (2021) A chickpea genetic variation map based on the sequencing of 3,366 genomes. Nature. 599: 622–627 https://doi.org/10.1038/s41586-021-04066-1

Hurgobin et al. (2018) Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid Brassica napus. Plant Biotechnology Journal. 16 (7), 1265-1274

Bayer et al. (2022) Sequencing the USDA core soybean collection reveals gene loss during domestication and breeding. The Plant Genome. 15: e20109

Khan et al. (2020) Super-pangenome by integrating the wild side of a species for accelerated crop improvement. Trends in Plant Science. 25 (2): 148-158

Dolatabadian A, Bayer P, Tirnaz S, Hurgobin B, Edwards D, Batley J. (2020) Characterisation of disease resistance genes in the Brassica napus pangenome reveals significant structural variation. Plant Biotechnology Journal. 18 (4): 969-982

Ruperao et al. (2021) Sorghum pan-genome explores the functional utility for genomic- assisted breeding to accelerate the genetic gain. Frontiers in Plant Science. 12:963

Zhao et al. (2020) Trait associations in the pangenome of pigeon pea (Cajanus cajan) Plant Biotechnology Journal. 18: 1946-1954

Golicz et al. (2016) The pangenome of an agronomically important crop Brassica oleracea. Nature Communications 7:13390

Yu et al. (2019) Insight into the evolution and functional characteristics of the pan-genome assembly from sesame landraces and modern cultivars. Plant Biotechnology Journal. 17 (5): 881-892

Rijzaani H, Bayer PE, Rouard M, Doležel J, Batley J, Edwards D. (2022) The pangenome of banana highlights differences between genera and genomes. Plant Genome. 15: e20100

Bayer et al. (2017) Assembly and comparison of two closely related Brassica napus genomes. Plant Biotechnology Journal. 15 (12):1602-1610

Bayer et al. (2020) Plant pan-genomes are the new reference. Nature Plants. 6 (8): 1-7

Bayer et al. (2021) The application of pangenomics and machine learning in genomic selection. Plant Genome. e20112.

Wang et al. (2021). The chicken pan-genome reveals gene content variation and a regulatory region deletion in IGF2BP1 affecting body size. Molecular Biology and Evolution. 38 (11): 5066–5081

Golicz A, Batley J and Edwards D. (2016) Towards plant pangenomics. Plant Biotechnology Journal. 14 (4):1099-105

Montenegro JDM, Golicz AA, Bayer PE, Hurgobin B, Lee HT, Chan CKK, Visendi P, Lai K, Doležel J, Batley J, Edwards D. (2017) The pangenome of modern hexaploid bread wheat. Plant Journal. 90 (5): 1007-1013

Danilevicz et al. (2020) Plant Pangenomics: Approaches, Applications and Advancements. Current Opinion in Plant Biology. 54: 15-25

Hurgobin H and Edwards D. (2017) SNP discovery using a pangenome: has the single reference approach become obsolete? Biology 6 (1): E21

Tay Fernandez et al. (2022) Expanding gene-editing potential in crop improvement with pangenomes. International Journal of Molecular Sciences 23(4): 2276.

Hu H, Scheben A, Verpaalen B, Tirnaz S, Bayer PE, Hodel R, Batley J, Soltis D, Soltis P, Edwards D. (2022) Amborella gene presence/absence variation is associated with abiotic stress responses that may contribute to environmental adaptation. New Phytologist. 233 (4): 1548-1555

Tay Fernandez et al. (2022) Pangenomes as a resource to accelerate breeding of under-utilised crop species. International Journal of Molecular Sciences 23 (5): 2671

Centre for Applied Bioinformatics

Graph pangenome visualisation

THE UNIVERSITY OF WESTERN AUSTRALIA

Centre for Applied Bioinformatics

PLANT GENOMICS

Graph pangenomes find missing heritability

The use of association studies to identify candidate genes for complex biological traits in plants has been challenging due to a reliance on single reference genomes, leading to missing heritability. Graphical pangenomes and the identification of causal variants help overcome this and provide an important advance for crop breeding.

David Edwards and Jacqueline Batley

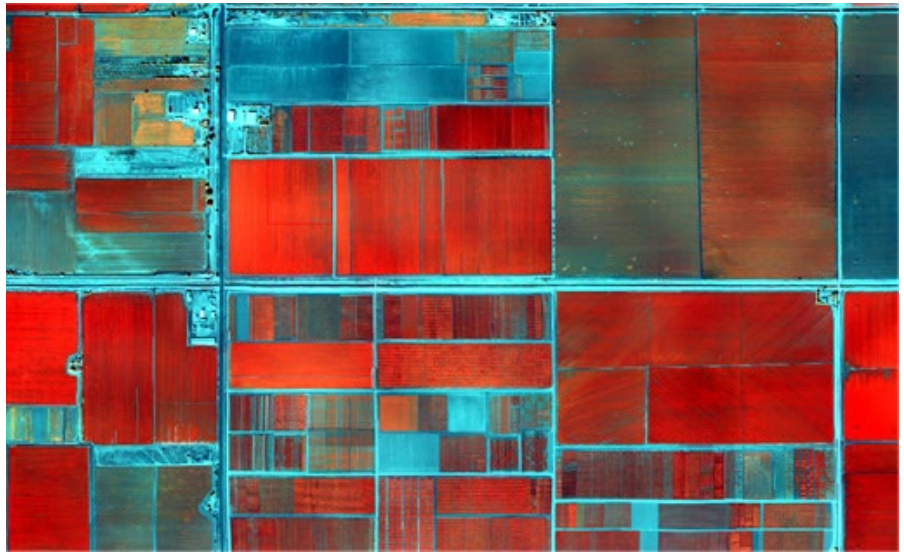
Edwards and Batley (2022) Nature Genetics 54: 919-920

THE UNIVERSITY OF WESTERN AUSTRALIA

Link phenotypes to genome variation

Centre for Applied Bioinformatics

Next generation phenotyping



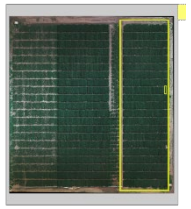
THE UNIVERSITY OF WESTERN AUSTRALIA

The need for machine learning

Centre for Applied Bioinformatics

Types of Inputs

Image data

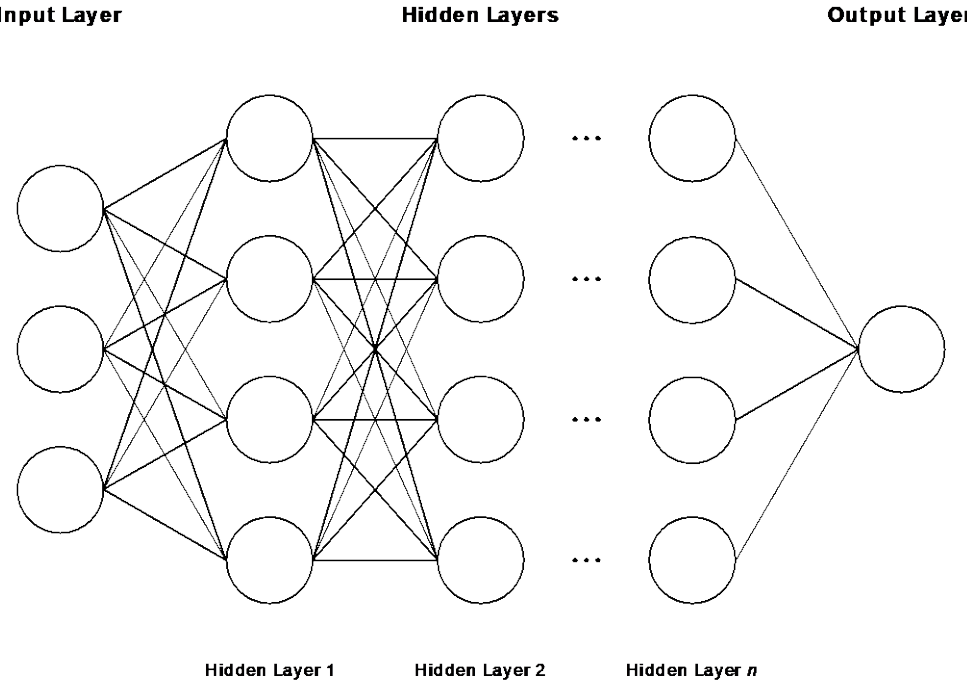


Text data

```
ACGTCACGTACTAG
CATGCATCGTAGCT
GTACGTACGTAGCT
```

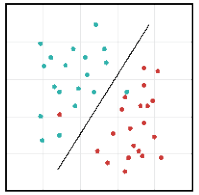
Tabular data

Contig	FDR	Control LB	Treatment MG
ptsG	5.10e-10	0.0	-3.96
setA	6.85e-8	0.0	4.24
sucB	3.10e-6	0.0	2.21
sucD	3.10e-6	0.0	2.46
deoC	3.10e-6	0.0	2.53

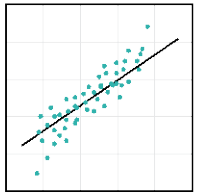


Types of Outputs

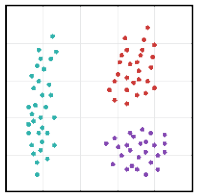
Classi cation



Regression



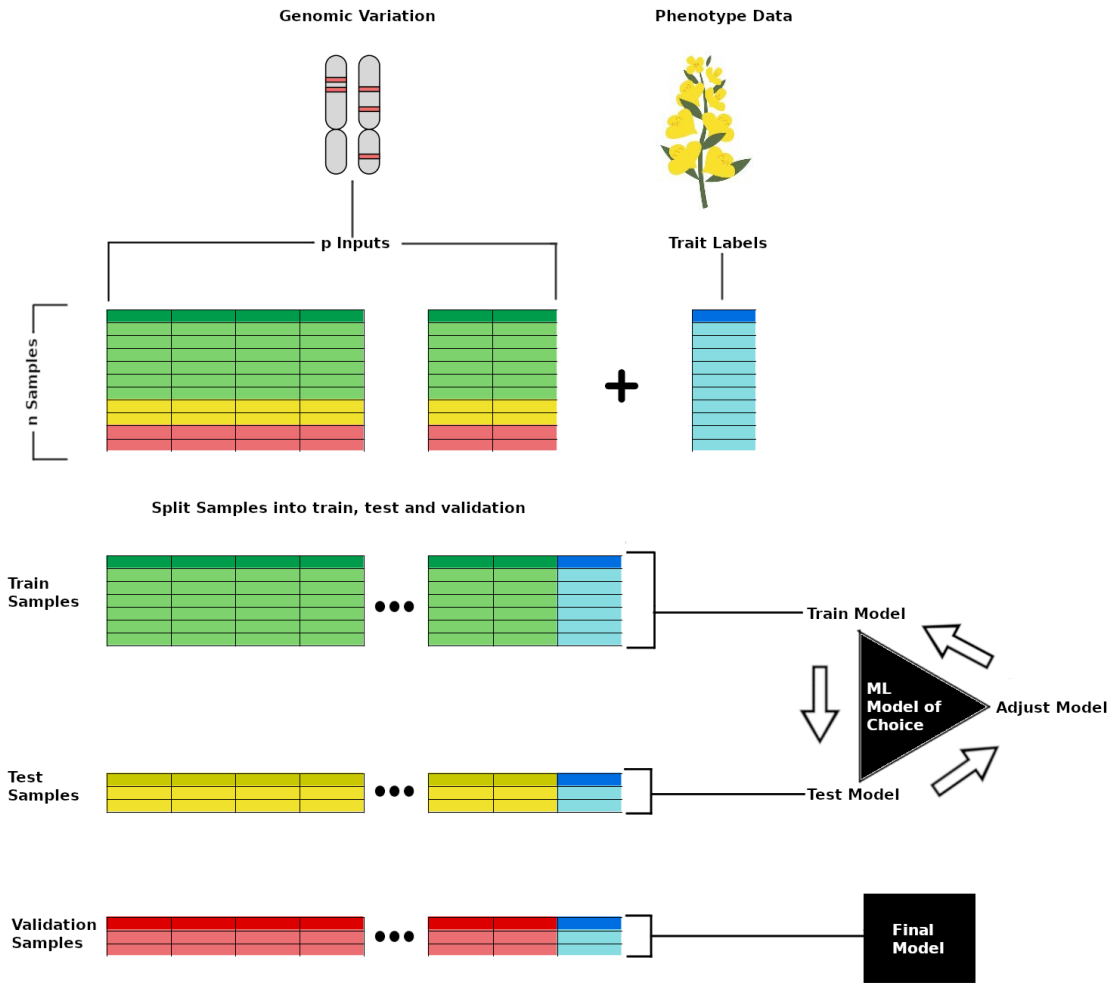
Clustering



THE UNIVERSITY OF WESTERN AUSTRALIA

Centre for Applied Bioinformatics

The need for machine learning



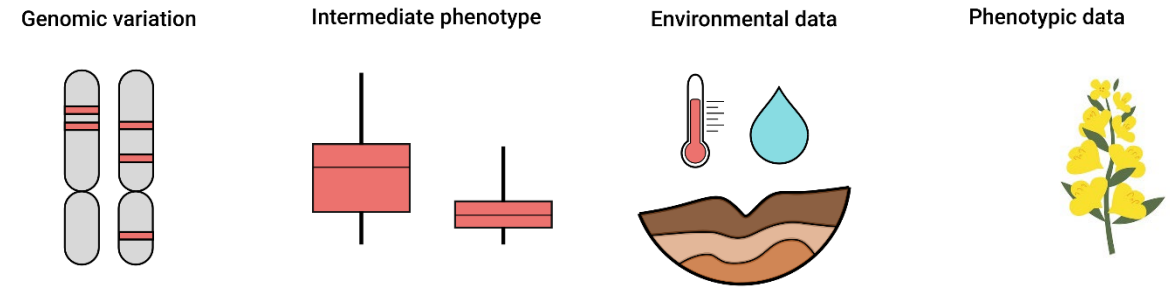
Danilevicz MF, Anderson R, Gill M, Batley J, Bennamoun M, Bayer PE, Edwards D. (2022) Plant genotype to phenotype prediction using machine learning. Frontiers in Genetics. 13: 822173

THE UNIVERSITY OF WESTERN AUSTRALIA

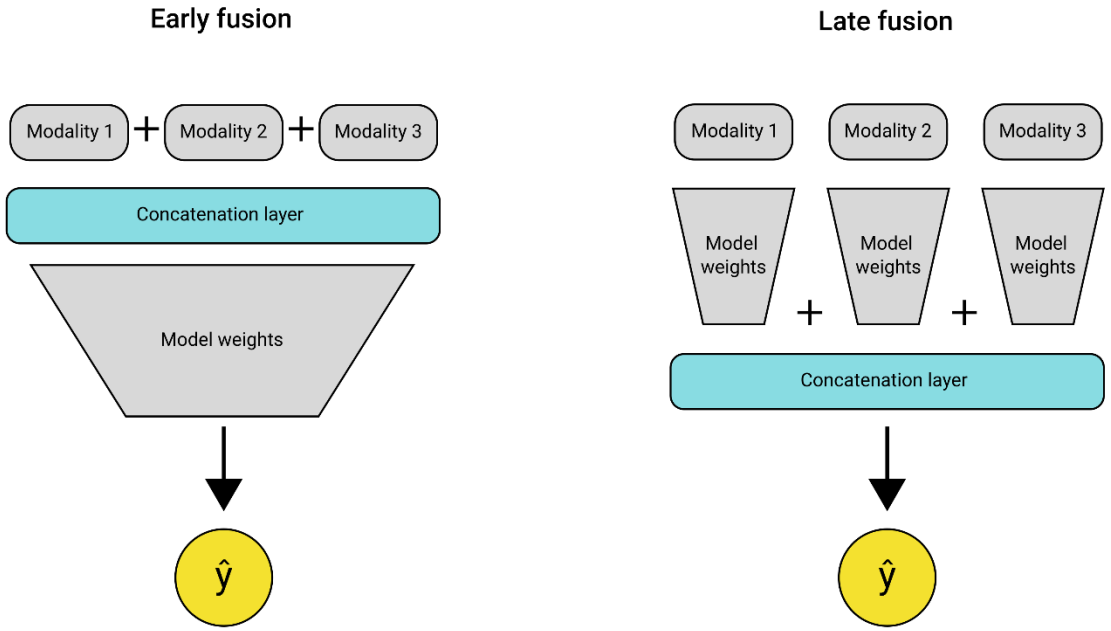
The need for machine learning

Centre for Applied Bioinformatics

Data modalities



Multimodal deep learning



THE UNIVERSITY OF WESTERN AUSTRALIA

Trait prediction from SNP data

Centre for Applied Bioinformatics

5.5 million SNPs, 700-1000 individuals per trait

Evaluation Metric	Trait	Learning Algorithm			
		XGB†	RF†	CNN†	DNN†
Accuracy	Flower Colour	**96.79%**	95.51%	87.82%	81.41%
	Pod Colour	**84.52%**	76.13%	70.32%	70.32%
	Pubescence Density	**91.56%**	81.17%	83.77%	80.52%
	Seed Coat Colour	**89.68%**	85.16%	84.52%	83.87%
RMSE % of Mean	Seed Oil Percentage	11.14%	**10.67%**	13.44%	10.74%
	Seed Protein Percentage	6.41%	**6.32%**	6.33%	6.91%
	Seed Weight	19.03%	21.63%	**17.77%**	18.91%

Identified genomic loci overlap with GWAS

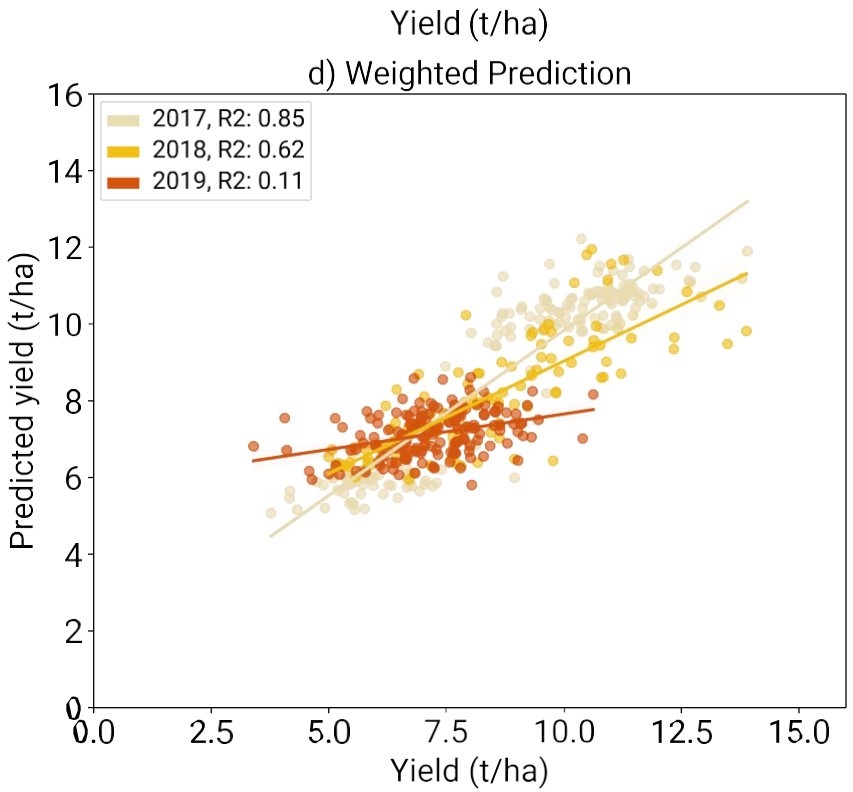
Machine Learning outperforms Deep Learning

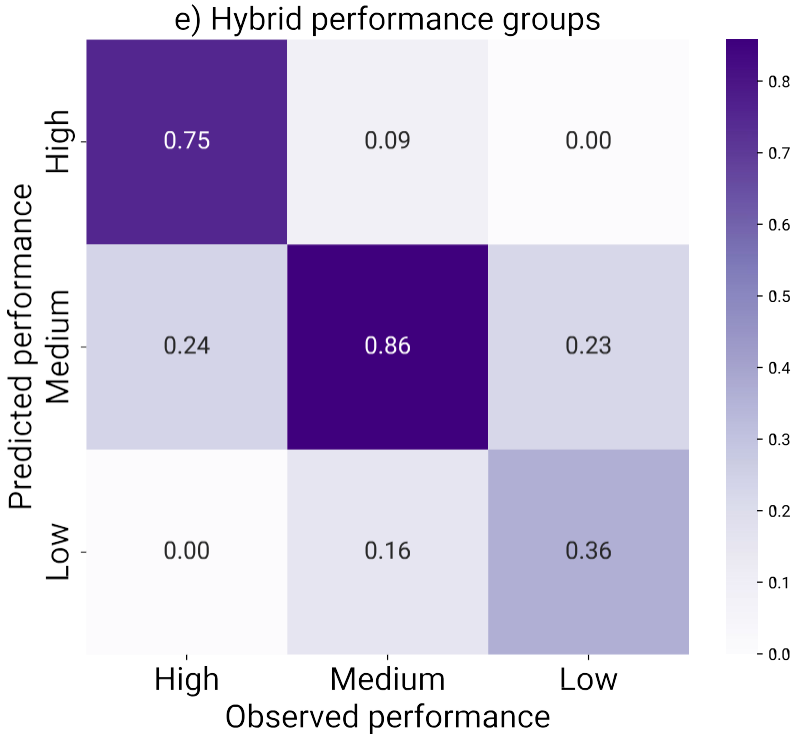
Gill M, Anderson R, Hu H, Bennamoun M, Petereit J, Valliyodan B, Nguyen HT, Batley J, Bayer PE, Edwards D. (2022) Machine learning models outperform deep learning models, provide interpretation and facilitate feature selection for soybean trait prediction. BMC Plant Biology 2, 180

THE UNIVERSITY OF WESTERN AUSTRALIA

Yield prediction from image data

Centre for Applied Bioinformatics





Danilevicz MF, Bayer PE, Boussaid F, Bennamoun M, Edwards D. (2021) Maize yield prediction at an early developmental stage using multispectral images and genotype data for preliminary hybrid selection in the field. Remote Sensing. 13 (19): 3976.

THE UNIVERSITY OF WESTERN AUSTRALIA

R gene prediction in canola

Centre for Applied Bioinformatics

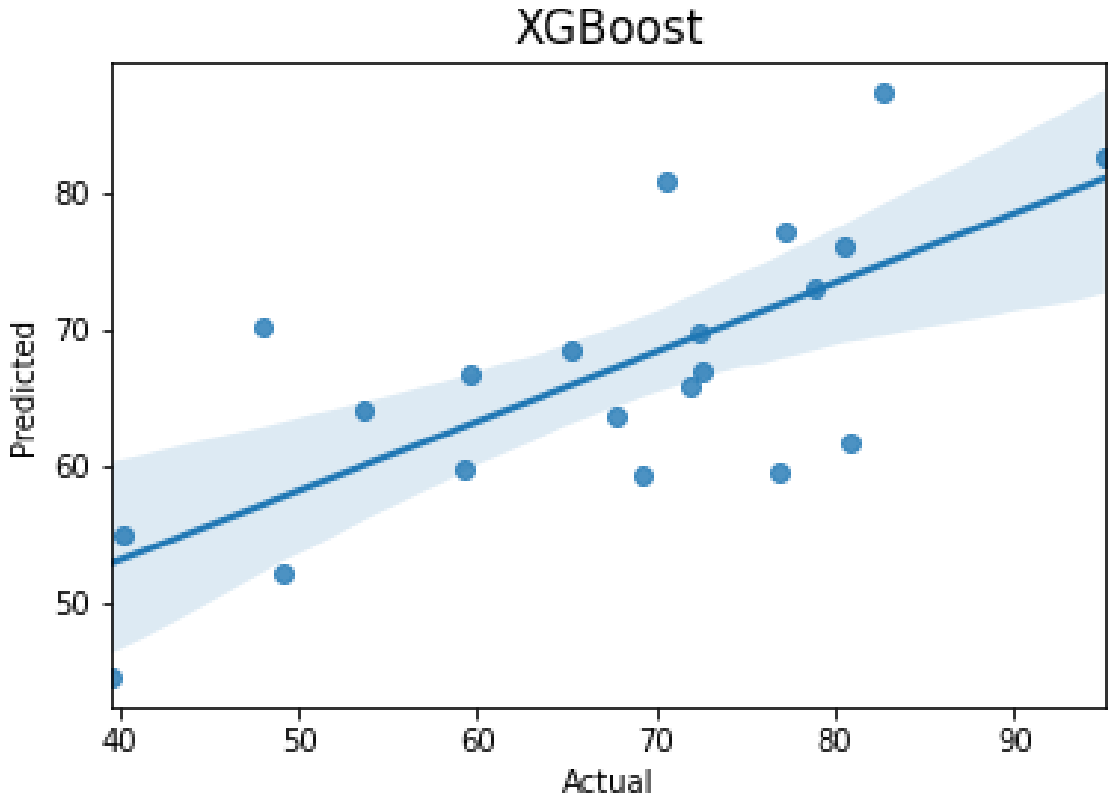
Class	Correct prediction
Rlm1	1951 (89.0%)
Rlm2	2160 (98.5%)
Rlm3	2042 (93.2%)
Rlm4	1977 (90.2%)
Rlm6	2160 (98.5%)
Rlm7	2181 (99.5%)
Rlm9	2031 (92.7%)
RlmS	2179 (99.4%)
LepR1	2151 (98.1%)
LepR2	2178 (99.4%)
LepR3	2178 (99.4%)

Based on phenotypic assays from isolate panel using XGBoost model

THE UNIVERSITY OF WESTERN AUSTRALIA

Genotype based quantitative blackleg resistance in canola

Centre for Applied Bioinformatics



Small dataset, multi location multi year phenotypes
Loci important for prediction overlap with those identified using GWAS

THE UNIVERSITY OF WESTERN AUSTRALIA

Next steps

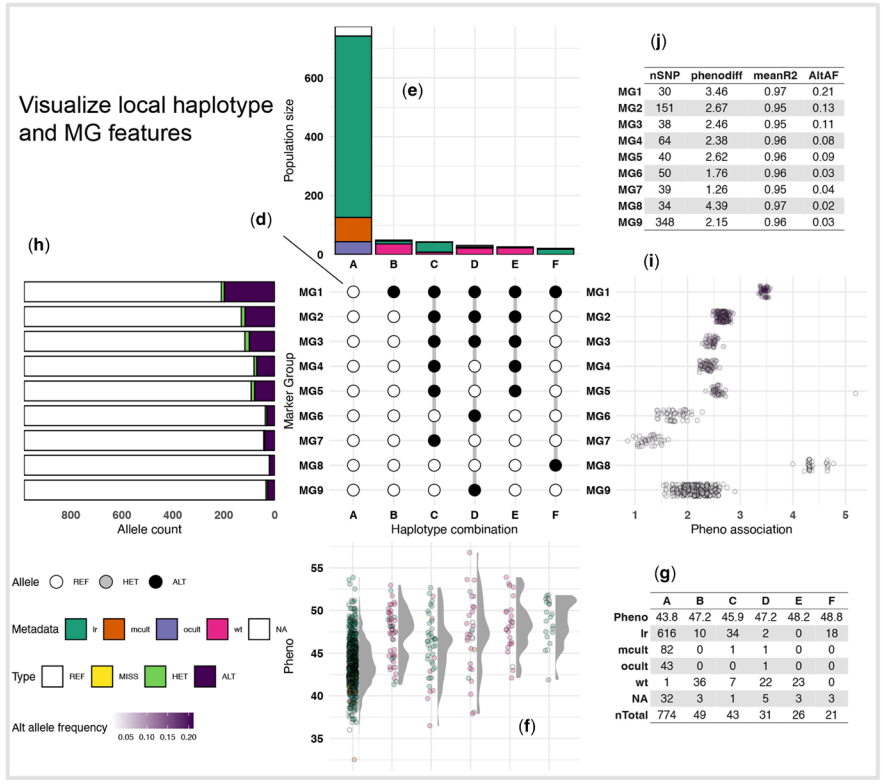
Centre for Applied Bioinformatics

Cluster SNPs into Marker Groups (MGs)

Haplotype individuals by their MG allelles

Optimize haplotyping with clustering tree

Visualize local haplotype and MG features



(j)

	nSNP	phenodiff	meanR2	AltAF
MG1	30	3.46	0.97	0.21
MG2	151	2.67	0.95	0.13
MG3	38	2.46	0.95	0.11
MG4	64	2.38	0.96	0.08
MG5	40	2.62	0.96	0.09
MG6	50	1.76	0.96	0.03
MG7	39	1.26	0.95	0.04
MG8	34	4.39	0.97	0.02
MG9	348	2.15	0.96	0.03

(g)

	A	B	C	D	E	F
Pheno	43.8	47.2	45.9	47.2	48.2	48.8
lr	616	10	34	2	0	18
mcult	82	0	1	1	0	0
ocult	43	0	0	1	0	0
wt	1	36	7	22	23	0
NA	32	3	1	5	3	3
nTotal	774	49	43	31	26	21

Moving from SNPs to haplotypes

Genetics and population analysis

crosshap: R package for local haplotype visualization for trait association analysis

Jacob I. Marsh [1,2], Jakob Petereit [1,2], Brady A. Johnston [3], Philipp E. Bayer [1,2], Cassandria G. Tay Fernandez[1,2], Hawlader A. Al-Mamun[1,2], Jacqueline Batley [1,2], David Edwards [1,2,*]

THE UNIVERSITY OF WESTERN AUSTRALIA

Next steps

Centre for Applied Bioinformatics

Build large language models based on publications (ChatGPT for crops)

Integrate data from knowledge graphs

Develop a comprehensive AI platform for crops and disease

THE UNIVERSITY OF WESTERN AUSTRALIA

Thanks

Centre for Applied Bioinformatics

Thanks to my team past and present

Thanks to all my collaborators

Thanks for my funders

Australian Government

Australian Research Council

THE UNIVERSITY OF WESTERN AUSTRALIA