Applying Networks of Prior Knowledge to Improve SNP Selection for Genomic Prediction and Polygenic Risk Scores

Thomas Crow, Prof Mark Cooper, Dr David Kainer

(With thanks to Prof Ricardo Mancera and the CMS Biomolecular Modelling Group)

t.crow@student.uq.edu.au

@scicrow.bsky.social

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

pawsey

ARC CENTRE OF EXCELLENCE FOR
PLANT SUCCESS
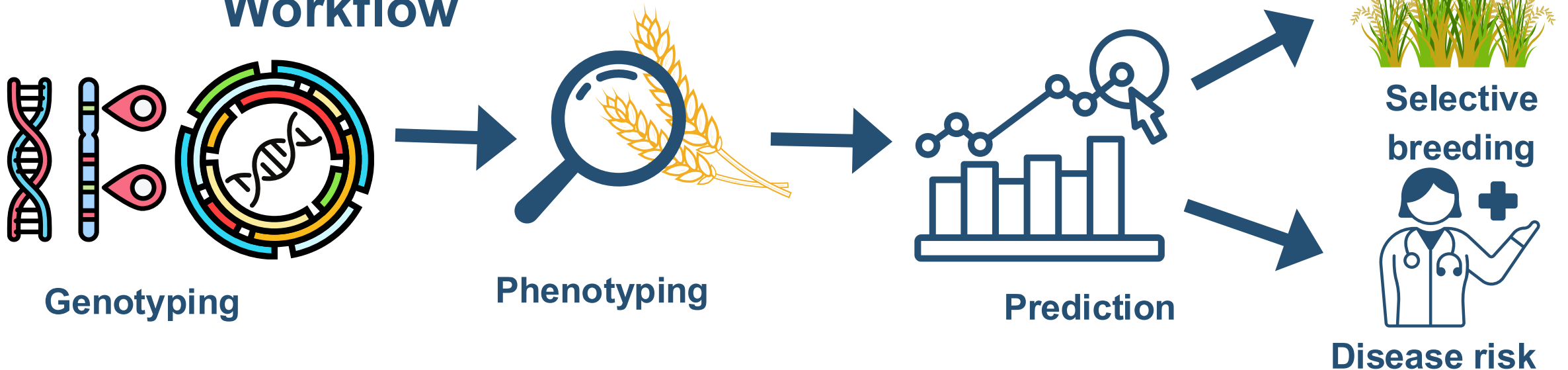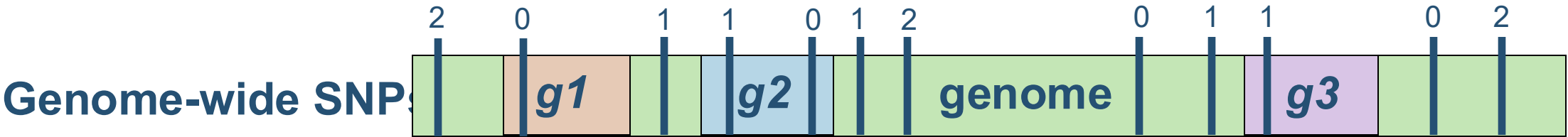IN NATURE AND AGRICULTURE

Curtin University

The Problem

Single nucleotide polymorphisms (SNPs) can influence organism traits.

But which SNPs do we use? Too much or too little SNP information can decrease accuracy.
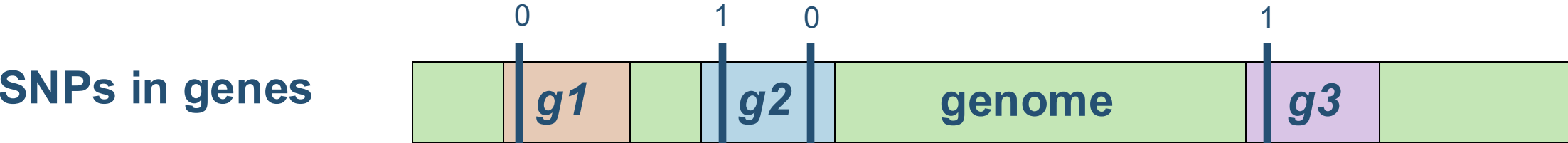
Genomic Prediction Workflow

Genotyping

Phenotyping

Prediction

Selective breeding

Disease risk

SNP Selection Methods

Genome-wide SNPs

Assumes infinitesimal contribution of all SNPs across entire genome. Biologically naive, dependent on large populations for accuracy (1).

SNPs in genes

Limit SNPs to genic regions. Allows for smaller samples but depends on annotated reference genome and misses regulatory effects (2).

SNPs from GWAS
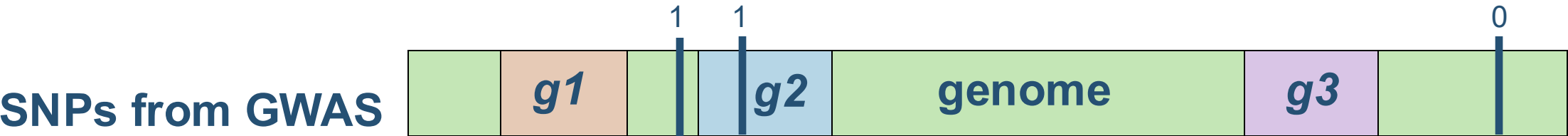
Biologically-informed SNP selection that can predict with small populations. Misses many additive effects and emergent effects from SNP interactions (3).
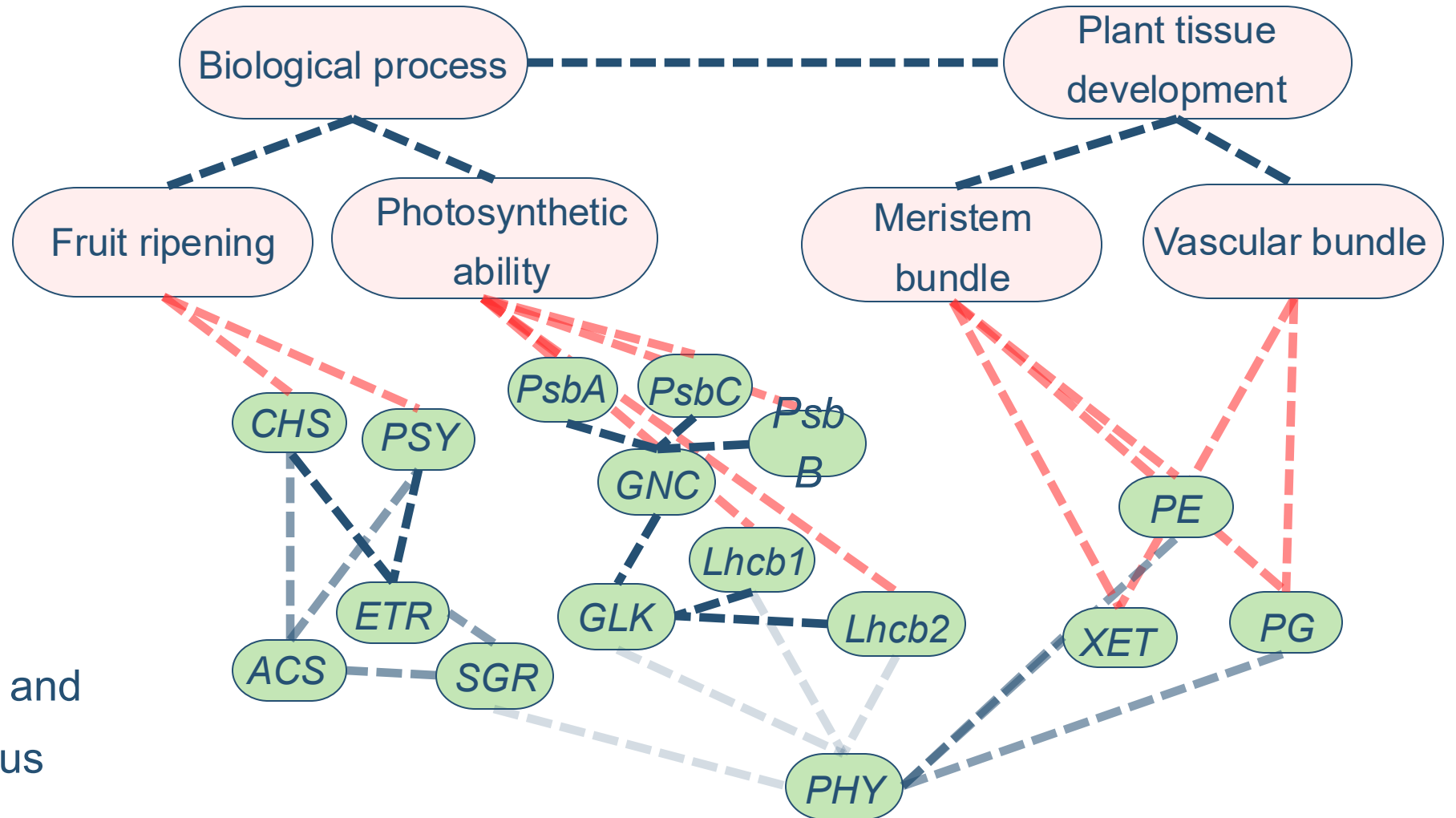
Our approach: building knowledge networks

1. Build a trait-to-trait network using ontologies (4)

2. Build a gene-to-gene network from existing knowledge

3. Link a trait-to-gene network with edges (relationships)

4. Build graph complexity and incorporate a heterogenous graph neural network

Network-guided SNP selection

Random walk with restart (RWR) explores graph connections by randomly travelling across edges. These require parallelized workflows.

Random walkers may find indirect gene-trait relationships. We can then filter for SNPs within direct and indirect genes.

1. Start at a seed node (point 0)

2. At each step, randomly select a direction.

RESTART

3. Each step may randomly result in a restart. Returning to seed node to explore other paths.