Genomic Selection

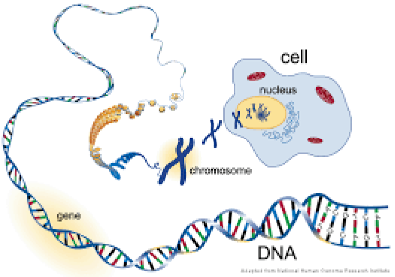
John McEwan

RPBC technical meeting Day 1 10th May

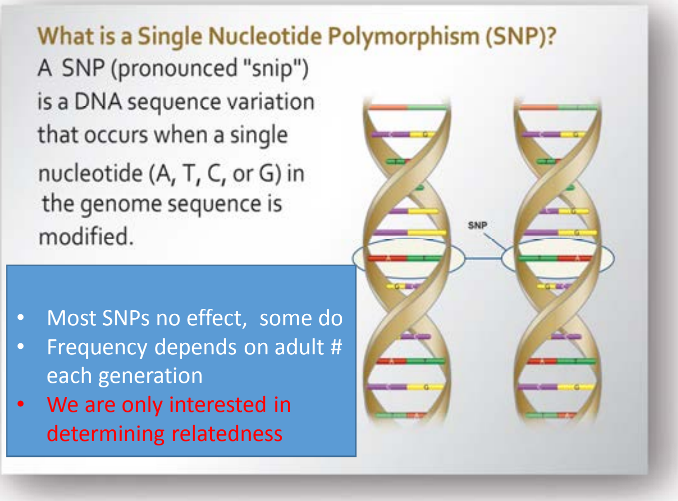
Distinction Hotel Rotorua

What is a SNP?

https://youtu.be/tJjXpiWKMyA



- Each cell has a nucleus
- Contain chromosomes (Pinus n=12)
- Diploids (eg Pinus) have 2 copies
- Each chromosome has 1 copy of DNA
- Each DNA copy has two strands



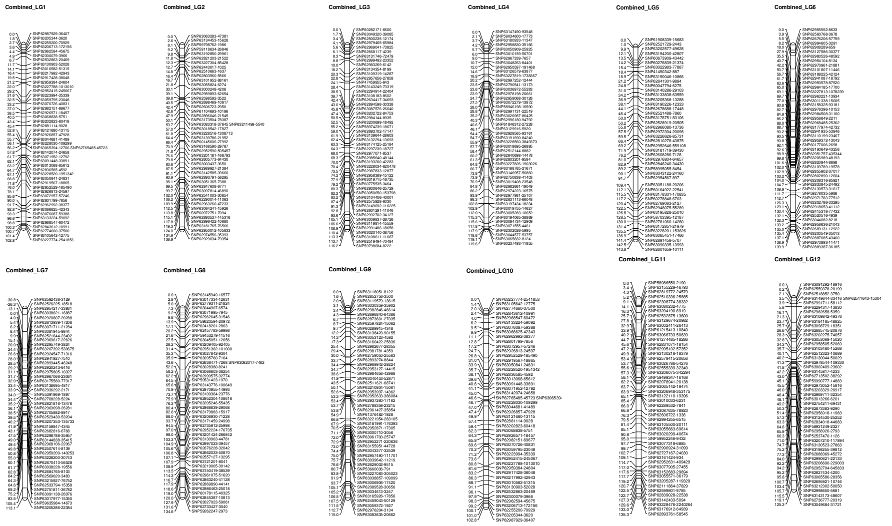
What is a Single Nucleotide Polymorphism (SNP)?

A SNP (pronounced "snip") is a DNA sequence variation that occurs when a single nucleotide (A, T, C, or G) in the genome sequence is modified.

- Most SNPs no effect, some do
- Frequency depends on adult # each generation
- We are only interested in determining relatedness

Exome Capture-derived SNP Consensus Linkage Map of *Pinus radiata* D.Don November 2016

Phil Wilcox



"length" 14.35 Morgans

Why?

- We can select trees to produce seed based on their own recorded traits
- But more accurate to also use information from relatives:
  = breeding values (predicts the heritable component)
  - E.g. progeny testing (backwards selection)
  - Can use pedigrees i.e. information all relatives (for forward & backward selection)
  - Assists greatly when want to rank individuals across environments
- Genomics provides more accurate information on relationships between individuals
  - More accurate breeding values
  - In particular makes better use of more distant relatives data
  - Can discriminate between full sibs even without measurements
  - Can be done at an early age
  - Captures unrecorded ancestral relationships and identifies recording mistakes
  - Valuable when pedigrees are shallow and subject to variable recording
  - Valuable when short genome length (recombination versus map)
- *Genomics now widely used in many plant and animal improvement industries*

Why?

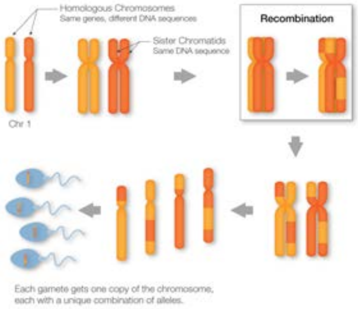
- When combined with improved breeding scheme designs results in faster genetic progress
- Need to have clear breeding objective of what traits need improvement….
- Caveats:
  - Need a low cost genotyping system
  - Need to genotype enough individuals to validate the system
    - Best if all breeding candidates and progeny are genotyped
  - Need to integrate genotyping into existing breeding scheme to accelerate gains
  - Increased selection accuracy and decrease generation length

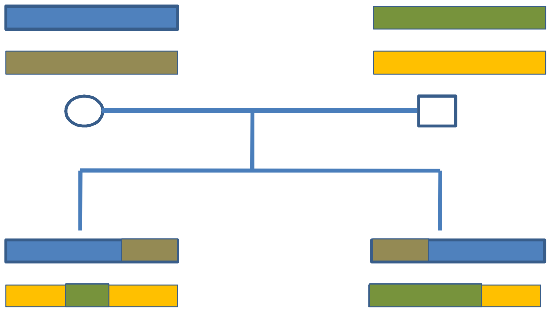
Genomic Selection

- Historically breeding values have been determined by measuring the individual, although in some cases it has to be predicted from relatives (pedigrees). currently use “BLUP” to efficiently combine data for breeding values.
- Increasingly (molecular) breeding values or **MBVs** are being estimated directly from an individuals genotypes.
- How is this done and what is its scientific underpinning?
- The field is still rapidly developing, but it rests on two phenomena:
  - **Mendelian sampling**… the sampling of chromosomal segments from parents
  - **Linkage disequilibrium** … the tendency for variants physically close together to be inherited as a unit.

Mendelian sampling

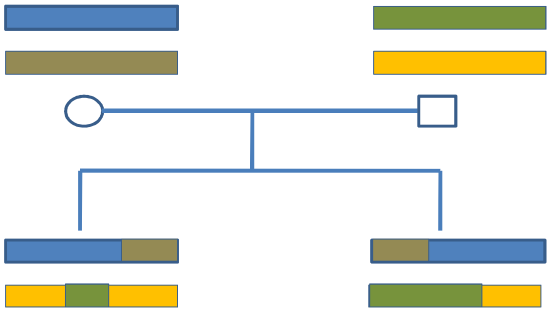
- 2 full sibs may inherit quite different grandparental segments
- Based on pedigree each should share half their DNA
- However ~50% of the genetic variation in offspring is due to Mendelian sampling!


Homologous Chromosomes
Same genes, different DNA sequences
Sister Chromatids
Same DNA sequence
Recombination
Chr 1
Each gamete gets one copy of the chromosome,
each with a unique combination of alleles.



Mendelian sampling

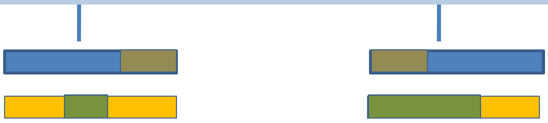
- 2 full sibs may inherit quite different grandparental segments
- Based on pedigree each should share half their DNA
- However ~50% of the genetic variation in offspring is due to Mendelian sampling!



Mendelian sampling

	mother	father	sib 1	sib 2
mother	1			
father	0	1		
sib 1	0.5	0.5	1	
sib 2	0.5	0.5	0.5	1

their DNA



- However ~50% of the genetic variation in offspring is due to Mendelian sampling!

Mendelian sampling

	mother	father	sib 1	sib 2
mother	1			
father	0	1		
sib 1	0.5	0.5	1	
sib 2	0.5	0.5	0.5	1

their DNA

So how big are the differences?

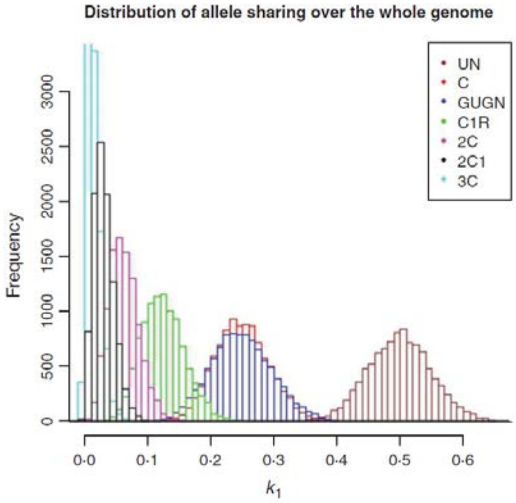


Fig. 5. Distribution of actual genome sharing ($\breve{k}_1$) for samples of 'human' genomes for different degrees of pedigree relationship of descendants of full sibs (as Fig. 2)

- Humans following diagram
- Pinus ~2X the variability
- Its big especially as a proportion of more distant relatives

GBLUP: what is a GRM?

Parentage derived or A relationship matrix

	mother	father	sib 1	sib 2
mother	1			
father	0	1		
sib 1	0.5	0.5	1	
sib 2	0.5	0.5	0.5	1

DNA derived relationship or G relationship matrix

	mother	father	sib1	sib2
mother	1			
father	0.1	1.1		
sib1	0.5	0.5	1	
sib2	0.5	0.5	0.43	1

Its just BETTER!

Linkage disequilibrium (LD)

The closer the variants are the less likely they are to be disturbed by recombination

It also depends on effective population size (Ne) with lower effective population sizes having more linkage disequilibrium over longer distances.

Many agricultural species $\mathbf{N_e}$ is markedly smaller than the physical numbers suggest. Often there may be millions of individuals but $\mathbf{N_e}$ may only number in the low hundreds.

You have an apex breeding population and a multiplication tier or tiers so the effective breeding population can be hundreds or thousands of times smaller.

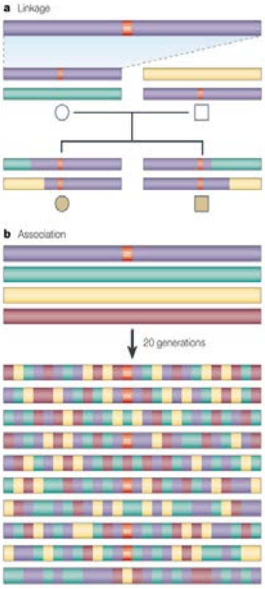
The effective population size is also typically closer to the number of males used per generation and with AI (artificial insemination) that may be very small, because the effective population size is proportional to the “harmonic” mean of the male and females used.

$$1/N_e \approx 1/4N_m + 1/4N_f$$

Look at this carefully its just mendelian sampling carried out over more generations!

Pedigrees become increasingly worse in percentage terms to determine shared segments as individuals are more distantly related

LD methods attempt to track segments … but it’s the same!



Called GBLUP

Pedigree relationship matrix is replaced by genomic relationship

Based on shared DNA variants from SNP chip genotyping

Most commonly used method in industry

- Identifies problems in traditional evaluation
  - Preferential feeding dams, wrong genetic groups
  - Incorrect pedigrees … wrong samples

Assumes infinite number of loci, infinitely small, additive effects….. Yeah right

Active development: maths & understanding genome architecture

Called GBLUP

Its ~~life~~ BLUP Jim but not as we know it

Molecular breeding values: SNP-BLUP

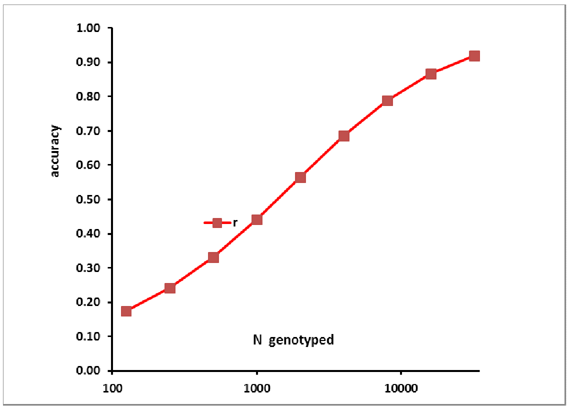
- An alternative to GBLUP is to simply track the alleles of closely spaced markers as proxies for ancestry of that segment of DNA
- Assumes an additive model i.e. 0, 1 or 2 alleles
- Each segment has an estimated effect on trait

SNP-BLUP mBVs cont

- For instance say you have AA, AB and BB for a marker 1 and a coefficient of +0.5cm for dbh and the average count of B allelesin the discovery set is 1
- The difference in the allele frequency from the mean B allele frequency is AA =-1 AB=0 and BB=1.
- So a BB individual would contribute 1*0.5 =+0.5kg and an AA individual would be -0.5.
- This is repeated for marker 2 and marker 3 etc … only the average count of B alleles and the coefficients change.
- Calculating mBVs. Once you have the genotypes and coefficients this is simple
- MBV = Σ ((count of B alleles for marker N –mean alleles marker N) * coefficient marker N)
- **It works out this is the same as GBLUP under "normal" conditions**

What about the accuracy of the mBV?

- This aspect is currently under significant development:
- In GBLUP calculated the same as for BLUP based on relationships
- The marker spacing also has to be dense enough. This is usually estimated as the number of independent chromosome segments = **$2N_eL$** where **L** is the length of the genome in Morgans (14.35 in Pinus)
- The effective population size **N_e** of the elites <200 so you need at least 2*200*14.35 = 5,740 markers
- The accuracy degrades the further (number of generations) away you are from the training set. This has to do with the breakdown of the linkage disequilibrium over generations (see graph for numbers needed after >10 generations)



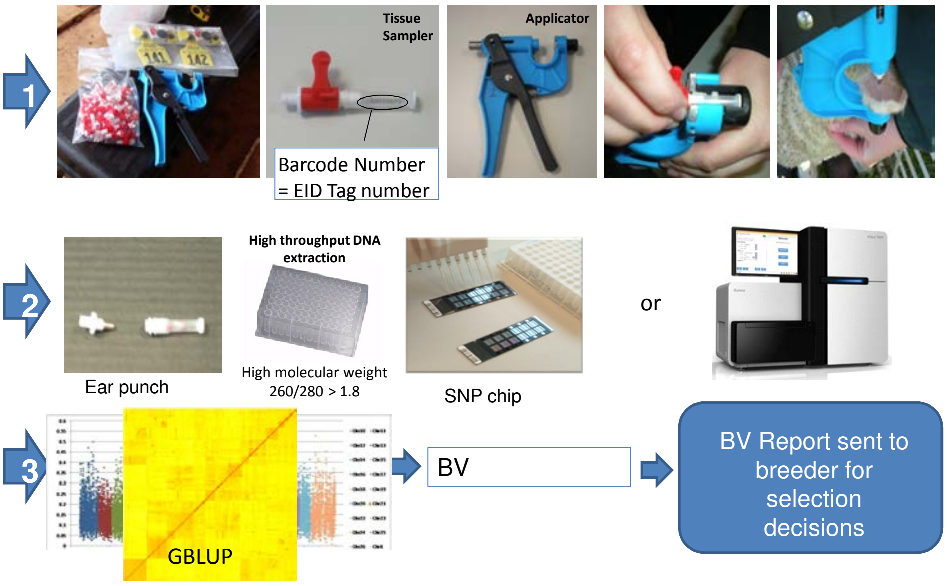
Summary: GBLUP vs SNP-BLUP

- 2 phenomena:
  - Linkage disequilibrium
  - Mendelian sampling
- Like looking on different sides of a coin
  - You will need both concepts
- Most use GBLUP, but as n > SNP move to SNP-BLUP
- Still under active development (across breed evaluation)

RPBC Radiata Pine Breeding Co

From tissue sample to breeding values

RPBC Radiata Pine Breeding Co

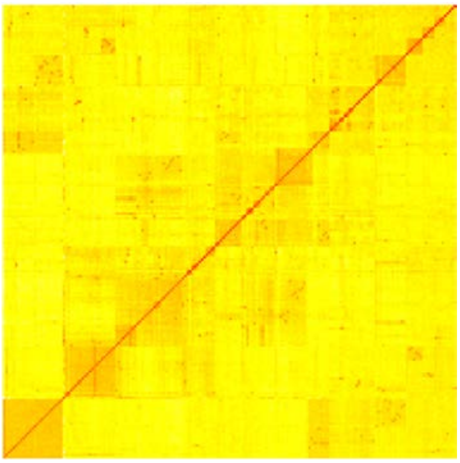


GRM uses

- Parentage
- Inbreeding
- strain prediction
  - & strain composition
- Co-ancestry

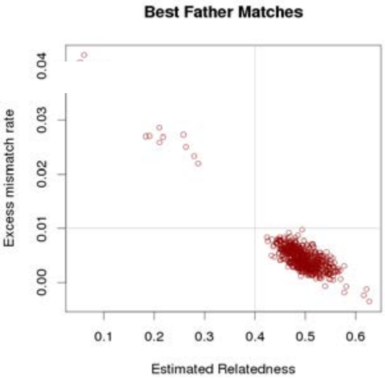
- Estimate h^2 without pedigrees
- Calculate mBVs

- Estimate genetic diversity
  - E.g. PCA plots via GRM



RPBC
Radiata Pine Breeding Co

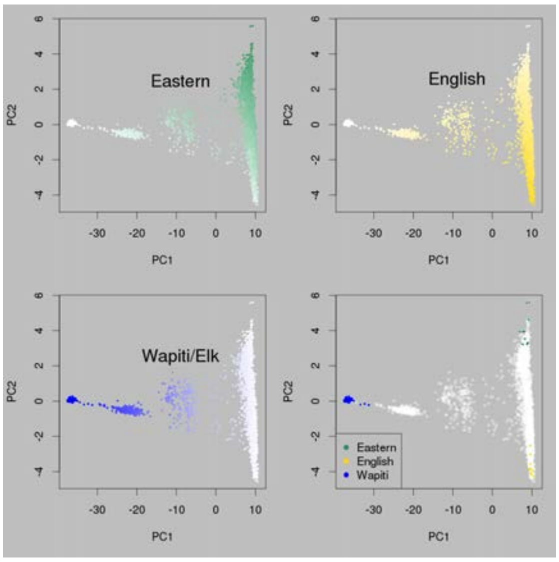
Best father matches (Red deer analysis)



76,926 SNPs, mean depth 3.3

RPBC
Radiata Pine Breeding Co

Cervus Elaphus (red deer)



Heritability & GBV estimation: dairy goats

Foote herd: ssBLUP analysis outputs

Trait	No. of records	Heritability (± s.e.)	Repeatability (± s.e.)
Protein, kg/day	49689	0.32 ± 0.02	0.54 ± 0.01
Volume, litres/day	55337	0.30 ± 0.02	0.53 ± 0.01
$\log_{10}$SCC	47575	0.20 ± 0.02	0.44 ± 0.01
Live weight, kg	8126	0.62 ± 0.03	0.80 ± 0.01

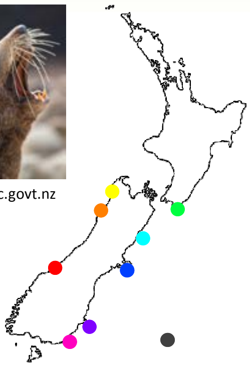
Meredith Dairy: GBLUP analysis outputs

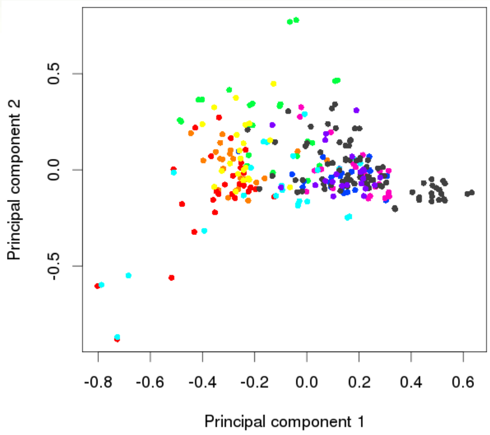
Trait	No. of records	Heritability (± s.e.)	Repeatability (± s.e.)
Fat%	7524	0.21 ± 0.02	0.22 ± 0.02
Protein%	7544	0.32 ± 0.02	0.39 ± 0.02
290-day volume, litres	8016	0.27 ± 0.02	0.48 ± 0.02
$\log_{10}$SCC	7552	0.10 ± 0.02	0.30 ± 0.02
Joining weight, kg	7081	0.30 ± 0.02	0.56 ± 0.01

Meredith: no pedigrees GBLUP gBVs within 3 months
mBV accuracies for milk vol est ~0.74
56K SNPs, mean depth 1.9 reads (Wheeler et al. WCGALP 2018)

Population genetics: Fur seal (*kekeno*)

www.doc.govt.nz





Will Stovall, Neil Gemmell,
Kim Rutherford

Other uses

- Linkage analysis (for genome assembly and mapping)
- Linkage dis-equilbrium analysis (for genome assembly and estimation of effective population size)

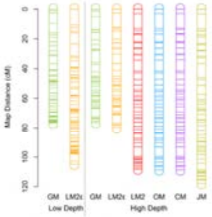
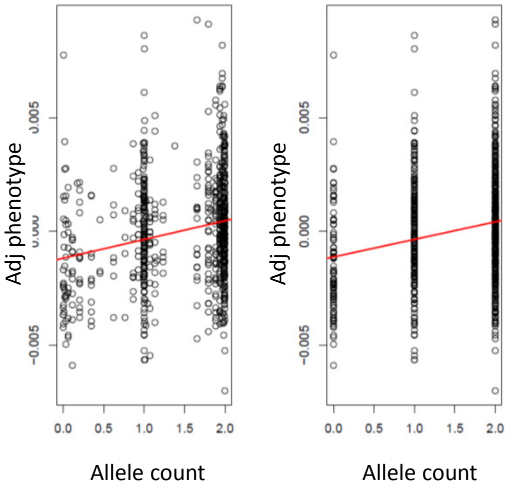
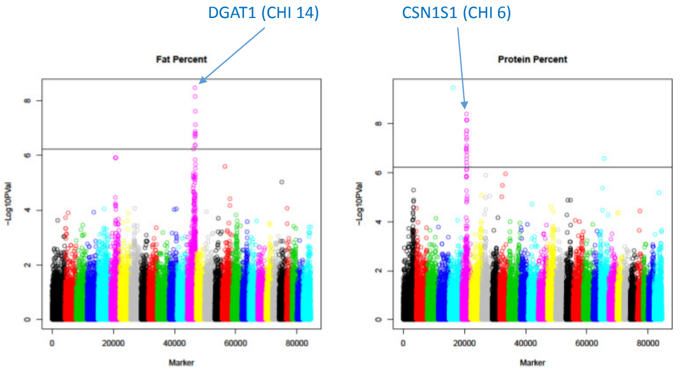


Figure 5 Subset of linkage maps for chromosome 11 of mānuka computed using the various software packages. Low Depth refers to the maps produced using SNPs with a mean read depth below 6, while High Depth refers to maps produced using SNPs with less than 20% missing data after setting genotypes with a read depth below 20 to missing. Map distances are in centimorgans and were computed using the Haldane mapping function.





The concept

- Perfect genotype = 0,1,2 and regression
- Probabilistic genotype =0 to 2 and regression

RPBC Genomic Selection Project

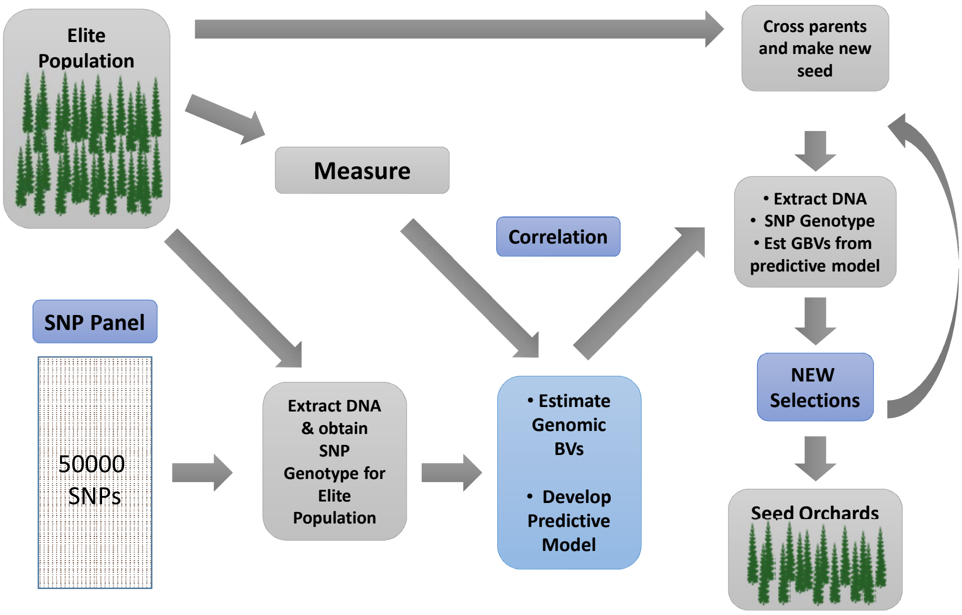
Goal:

- **Implement GS and forward selection** in radiata pine improvement

How:

- Develop **genome-wide marker panel** (50K SNPs)
- **Establish** cloned **populations** to **implement GS and forward selection**

Genomic Selection – How does it work?



Elite Population

Measure

Cross parents and make new seed

SNP Panel

50000 SNPs

Extract DNA & obtain SNP Genotype for Elite Population

Correlation

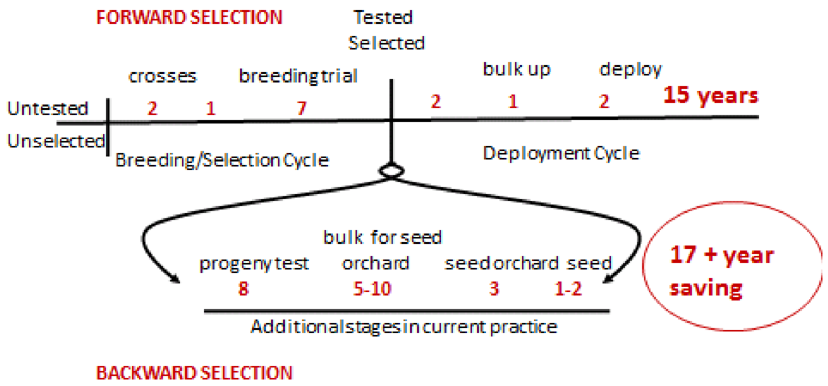
- Estimate Genomic BVs
- Develop Predictive Model

- Extract DNA
- SNP Genotype
- Est GBVs from predictive model

NEW Selections

Seed Orchards

Breeding and Deployment Timelines



FORWARD SELECTION

Tested Selected

crosses

breeding trial

bulk up

deploy

Untested Unselected

2 1 7

2 1 2

15 years

Breeding/Selection Cycle

Deployment Cycle

bulk for seed orchard

progeny test

seed orchard seed

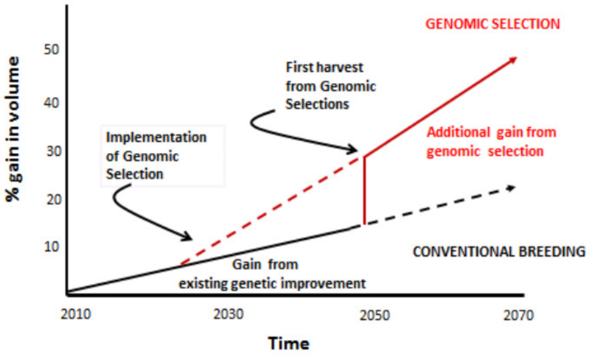
8 5-10 3 1-2

Additional stages in current practice

17 + year saving

BACKWARD SELECTION

What was proposed



"On-going exploitation of GS, commencing with GS/ forward selections of the 2013 elites in 2017/18, will add further pulses of up to 10% gain at intervals of 8 to 10 years, representing a doubling of genetic gain per unit time over conventional breeding. This will add further significant financial value over time."

"At a national level, the economic impact of Genomic Selection becomes very significant. Assuming the current average recoverable volume 450m³/ha, 70% uptake, plantings of 40,000 hectares per annum, and log price at harvest of $90, the increased for a 30% increase in volume would be $425 million. Financial gains from other traits could increase the additional value at harvest to >$500 million. An additional 10% gain arising from the next tranche of genetic improvement should add a further $250-300 million at time of harvest. All gains are cumulative and ongoing."

RPBC
Radiata Pine Breeding Co

Summary

- SNP = DNA variant
- Genomic selection replaces a relationship derived from pedigree with that estimated from DNA
- Genomics provides more accurate information on relationships between individuals
  - More accurate breeding values
  - In particular makes better use of more distant relatives data
  - Can discriminate between full sibs even without measurements
  - Can be done at an early age
  - Captures unrecorded ancestral relationships and identifies recording mistakes
  - Valuable when pedigrees are shallow and subject to variable recording

- Major objective create low cost genotyping system: complete ~$35/sample measures 50K SNPs
- "Research" molecular BVs available for Pinus
- Validation of benefits using molecular BV accuracies ongoing
- Breeding scheme & deployment to maximise genomic selection benefits
  - Underway expected to be finalised September
- Strategic workshop on overall programme and results November 2019

RPBC
Radiata Pine Breeding Co

Thanks

- Natalie Graham, Heidi Dungey, Emily Telfer, Yongjun Li, Jaroslav Klápště
- Brian Cullis and Alison Smith
- Phil Wilcox
- Drs Apiolaza and Evison
- John Hay, John Butcher, Dr Kirst and many others