WEDNESDAY, AUGUST 27, 2014

Keynote Lecture

Opportunities for genomics in forestry

Bruce Tier
University of New England, Australia

It is now possible to test individuals of many species for hundreds of thousands of loci. This can be done directly or by imputing large numbers of loci from smaller, representative subsets. Tests for hundreds of thousands of single-nucleotide-polymorphisms (SNPs) on a single chip are available as are subsets. Also genotype s can be obtained by direct sequencing. Indeed, the cost of obtaining the genome sequence for new species has also become affordable. As a consequence considerable use of genomic information has been made by breeders of many species with tens of thousands of individuals being tested. The way that this type of information has been used varies with the industry, and depends upon the end product which range from inbred lines in plants through to livestock where cross breeding is necessarily ongoing. The problem for forestry breeders lies between these extremes, with goals in the long term, long times between making selection decisions and obtaining results and a variety of problems common to both plant and animal breeders. This paper examines how forestry breeding enterprises can best take advantage of the flood of genomic data that is becoming available. From the search for genes, through to genomic selection and the design of breeding programs for improvement and deployment, the use of genomic information is discussed with the goal of making better use of the available phenothypes to breed trees that are more productive in any environment.

Opportunities for genomic information in Tree Breeding

Bruce Tier
Animal Genetics and Breeding Unit
University of New England
Armidale, NSW, Australia

Outline

- Background
- Genomic information
- Genetic evaluation
- Genomic selection
- Results from livestock industries
- Tree breeding and genomics
- A future

Background

- Quantitative genetics
  - Inference
  - "You can't look up an animal's or a tree's genes"
- Genomic information and quantifying effects
  - P >> N
- WCGALP Vancouver (last week)
- Breeder's equation:

$$R = ih^2\sigma_p/L = ih\sigma_a/L$$

GENOMIC INFORMATION

Genomic information

- Diploid species
- Chromosomes
- Inheritance
- We can observe (parts of) one's genome
- Uses of Genomic Information:
  - Status at one or more loci
  - Sharing across individuals

Genomic information

- Single and multiple loci
  - AFLPs, RFLPs, SSR (µSats), SNP (all polymorphic), indels, transpositions ….
  - Sequence data
- "Identify causal mutations"
  - Transcriptome
  - µRNAs, cis and trans acting factors
  - Interactions between loci, GxE …

Genomic information

- Tests are usually for genotypes
  - Two chromosomes, one from each gamete
- Requires Phasing/Haplotyping:
  - Identify parental gametes
    - Numerous methods, very reliable particularly with large family data

Sequence data

- Relatively cheap now (cf human genome)
- Provides order of bases/loci on chromosomes
  - Genetic map
- Genotype by sequencing also useful for small regions

Genomic information

- Some phenotypic variation results from genetic variation
  - Can ignore invariant regions (most of the genome)
- Markers can track large chunks of chromosomes
  - Interior parts between markers can be imputed
    - Haplotypes are required
  - Recombination relatively rare occurrence given the length of the genome
    - LD in trees

Genotyping Paradigm

- (Obtain sequence)
- Genotype large numbers of individuals with dense markers/sequence
  - Impute their haplotypes
- Genotype others with cheaper, less dense methods and impute the dense part

  ➔ **all** individuals with **dense** genotypes

Imputation

- Parental Haplotypes:

AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB

- Progeny genotype:

A?????A???A????A???A???A?????B??

A?????B???A????B???A???B?????B??

Imputation

- Progeny haplotype:

AAAAAAAAAAAAAAAAAAAAAAAA?????BBB

A?????B???A????B???A???B?????B??

- or:

AAAAAAAAAAAAAAAAAAAA???BBBBBBBBB

A?????B???A????B???A???B?????B??

Cost of genotyping (Cattle)

- $20: 150 SNPs – pedigree inference
- $50: 10000 SNPs – low density
- $80: 50000 SNPs – medium density
- $200: 750,000 SNPs – high density
- $1000: full sequence
- Prices depend upon volume
- Can impute from 10000 ➔ Sequence
  - In stages, reasonably reliable

Sequencing Strategy

- 25x coverage is thorough for one individual
  - All sequence genotyped
- 0.2x coverage is cheaper for the whole group
  - Inference and pedigree provides the rest
- Choice depends upon purpose

Genomic information

- Do we want to find genes?
- Do we want to breed 'better' trees?
- Does knowing the first one help the second?
  - Yes, if genetic disorder or,
  - If it explains a significant amount of genetic variation
    - Unlikely if selection has been successful!!
      - DGAT1, Calpastatin

Variation in the genome

- Mammalian genome ~3Gb
  - Genes: ~20000
  - Gene products: 200,000+
  - Variants multimillions
- We can know the genes,
  - And even what some of them do
  - But estimating the size of their effects is problematic

Tree genomes

- Conifer Genome ~30Gb
  - 10x the mammalian genome
  - Gene products: ?
  - Variants: ?
  - Massive numbers of repeated sequences
- Eucalypyt genome
  - Smaller than mammalian genome

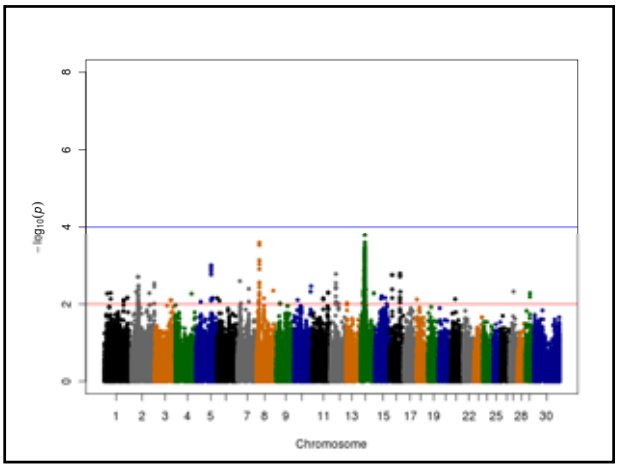
Fundamental propositions

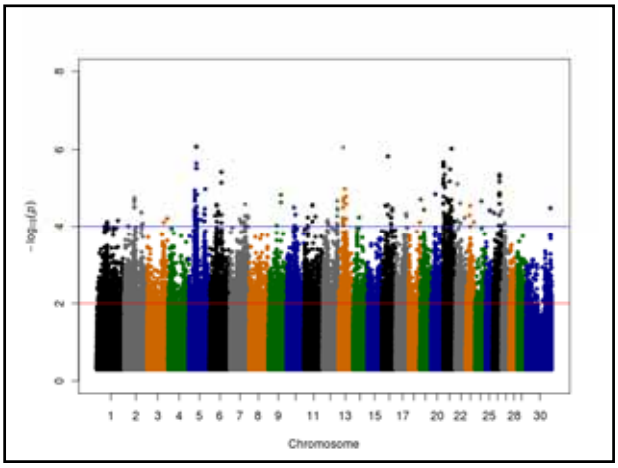
QTL (causal mutations):

- in *complete* LD with markers
- act consistently
- explain significant proportion of genetic variance

Finding quantitative genes

- Looking for QTL of large effect amongst a lot of other variation
  - the wide cross
- Genome Wide Association Study (GWAS)
  - **y*** = **q**g + **e**,
- Fit number of alleles for one locus as a covariate, plot -$\log_{10}$(p) values, as a Manhattan plot:





Finding genes: quantitative

- GWAS: multiple testing: expect to find 5% at p<0.05
  - Multiple testing problems, False discovery rate
- Testing loci individually ignores other loci
- GWAS explains multiples of the genetic variance
- Estimating the thousands of effects well
  - requires a lot of data, analysed jointly

GENETIC EVALUATION

Genetic evaluation

- Phenotype = Genotype + Environment

In linear model land:

- **y = Xb + Zu + e**
  - Phenotypes (**y**) are a function of fixed effects (**b**) and random effects (**u**) and a residual (**e**). **X** and **Z** are incidence matrices assigning observations to effects.
- Expected means and variances of **u** and **e** must also be specified.

$y = Xb + Zu + e$

- **b** could contain trial, block and replicate, location, treatment, propagation effects;
- **u** would normally contain any genetic effects (breeding values, GCA, SCA, epigenetic, maternal effects), GxE interactions, as well as,
  - other random effects, that arise from shared histories, such as spatial (row and column) and competition effects.

$y = Xb + Zu + e$

- A very general model
- Designed to accommodate 'unbalanced' data
  - Why not exploit it?
- Best Linear Unbiased Prediction (BLUP)

Estimated Breeding Values

- Covariances amongst breeding values are commonly assumed to be $\mathbf{A}\sigma_a^2$
  - **A** is the Numerator relationship matrix.
  - Covariances weight observations according to perceived sharing of genetic material
- Dominance effects can be included in the model with **D** – the Dominance relationship matrix.

Comparing individuals

- At the phenotype level individuals can only be sensibly compared with others that have been treated alike.
- Estimated breeding values (in **u**) can be compared, across space and time, on the genetic scale.
  - For individuals in the **same analysis**

GENOMIC SELECTION

Genomic Selection

- Predict individual's genetic merit given their genotype
  - Develop prediction equation using training set
  - Use prediction equation to generate genomic breeding values for prediction set
    - Variety of methods

Genomic Selection Methods

- All methods based on a model like:
  - $\mathbf{y = Xb + (Z_1u) + Z_1Qg + e}$, where **Q** is a matrix of genotypes and **g** is a vector of effects of one allele at that locus.
  - We might want to know **g**, but **g** can hold tens (or hundreds) of thousands of effects!
  - Generally use SNPs
- Methods vary in how Var(**g**) is treated

Methods for genomic selection

- SNPBLUP $\text{Var}(\mathbf{g})=I\sigma_g^2$
- Bayesian "alphabet" to determine marker effects, all MCMC method
  - A: all SNP in the model
  - B: some SNP in the model
  - C_π: some SNP in the model, π is the proportion
  - R: mixture model, all SNP in the model variety of distributions with different variances (One very small ~ zero)

Methods for genomic selection

- GBLUP
  - Use genotype data to build genomic relationship matrix (**G**)
  - Equivalent to SNPBLUP: $\mathbf{g} = \lambda \mathbf{M G^{-1} u}$
- SNP can be weighted: WGBLUP
  - Equivalent to Bayes alphabet methods

Genomic Selection

- Methods give similar results for any data set
- Bayesian methods slightly more successful when there are large QTL
- Variety of results across species
  - Results are best when effective population size is small and there are plenty of data
- All models imply that **g** effects are estimable
  - Variation at the individual (not marker) level

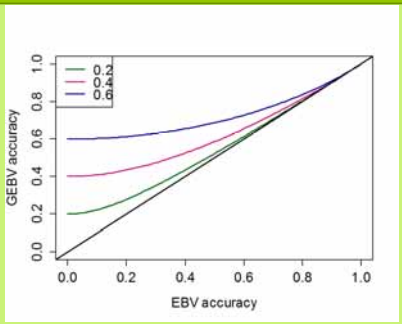
Use of Pooled DNA

- Individuals are phenotyped
- Grouped on phenotype ranks
- DNA pooled (in equal amounts)
- Fewer genotype tests required
- Similar models and methods
- Early results suggest similar accuracy

"Blending"

- Genomic predictions (GBVs) for young animals are combined with their EBVs from routine evaluations
- Availability of useful phenotypes early in life affects the value of GBVs
  - Blending (= selection index) of EBV with GBV
- Heritability of GBV = 1
  - Only deals with part of the genetic variation

Value of genomic prediction



Genomic Selection in Livestock

- Quantity and quality of data
  - Phenotypes vs accurate EBVs
- Breeds in the population
- Family size (AI)
- Effective population size
- Numbers of families
- Relationship between training and prediction sets

GS in Livestock

	Dairy	Beef	Sheep	Pigs
Breeds	1 major, few minor	20+	100+	Few major
Data recording	Intensive for production	Sparse	Sparse	Intensive in nucleus
Accuracy of training data	High	Low	Low	low
Family Size	Very large	1-10000+	1-1000	10-1000
Relationship to training set	Very high	Modest	Modest	Good

GS in Livestock

	Dairy	Beef	Sheep	Pigs
Availability of early records (ie before selection)	Nil	Some	Some	Some
Results	0.4-0.8	0.2-0.5	0.2-0.6	0.4-0.7
Usefulness	High Young bulls receive highly accurate EBVs as calves cf progeny test requiring 4 years	Moderate GS accuracies are low, correlated phenotypes provide partial results		In between

GS in livestock

- Results variable depend on
  - quality and quantity of data in training population
  - relationship of predictees to training population
- Value declines with
  - Genetic distance from the training set
    - E.g. Dairy Industry: Young bulls can be one (sons) or two (grandsons) generations from training set
  - Availability of other information

Genomic Selection – Single Step

- Single step is where all individuals are combined in the one evaluation - where training and validation is simultaneous
  - Genotyped or not
  - Phenotyped or not
  - BLUP Equations must be modified
- All individuals get the benefit of the genotypes
  - Essential to be 'Best'

Relationship matrices

- Specifies how genes are shared
- **A** uses the pedigree
  - Based on identity by descent (IBD)
  - Founders unrelated, pedigree traces IBD
- Genomic relationship matrices (**G**)
  - Based on identity by state (IBS) – founders 'related'
  - **G** based on pedigree and LD is better

Story unfolding

Genomic Selection – Single Step

- Big problems with the joint analysis of populations derived from multiple origins
  - Genetic groups
- Active area of research
  - **G** based on pedigree and LD better
- Haplotype models to come
  - LD captured more accurately
  - More tractable computationally

TREE BREEDING

Tree Breeding

- Wild population – many ancestors still alive!
- Arboretum where crosses are made
  - (Breeding Program)
- Seed harvested and germinated
- Planted in designed trials across range of environments
  - "Progeny testing"
- Phenotypes measured

Tree Breeding

- Breeding values estimated for trees in arboretum and trials (all data)
- Trees selected for both arboretum and orchards for seed production
- Large number of breeding objectives
  - Different and changing environments
  - Different traits and combinations of traits

Tree Breeding

- Modest number of trees selected for crossing
  - Relatively few parents
- Lots of progeny in trials
- Data recorded (8-10 years after crossing)
  - Correlated with Objective traits
- Parents and families evaluated very accurately
- Over time many trials and many families
- Significant logistic problems

Tree Breeding

- In the end we have
  - A large number of trees recorded
  - Relatively small number of parents
    - All with a large number of progeny, and consequently highly accurate EBVs.
    - Each trial tree has its own record
  - Large quantity of data
  - Plenty of scope for selection

Opportunities

- Based on breeder's equation
  - Accuracy (h)
  - Selection intensity (i)
  - Genetic variation (σ_a)
  - Generation interval (L)

Opportunities: Accuracy

- Better defined relationships
  - Pedigree corrected (OP, polymix …)
  - Genomic data defines remote relationships better
- Better models (**G**, single step)
- Better family estimates
- Potentially more data
  - Production stands may provide useful phenotypes
  - Genotypes are the link back to arboretum

Opportunities: Selection Intensity

- Better breeding program design:
  - More parents, fewer progeny per family
- Early screening possible with GBVs for
  - rarely expressed traits e.g. rare diseases
  - expensive to measure traits e.g. Kraft Pulp Yield
  - traits expressed late in rotation
- Individuals to be measured can be selected on genotype and breeding value

Opportunities: Genetic Variation

- Characterise wild populations
- Select from the wild
  - could be useful individuals in the forest
- Manage variation in the arboretum using the genomic data

Opportunities: Earlier selection

- Reduced generation interval
- Genomic prediction equations (or the single step method) provide much more accurate EBVs at very young ages
- Many individuals can be screened before planting in trials
  - And grafted into arboretum and orchards before phenotyping

Conclusion

Where to?

- Obtain moderately dense markers for your species (even sequence?)
- Completely genotype arboretum
  - Develop genomic prediction system based on single step or haplotype model
- Use all trial data in one analysis
- Select some trial individuals for genotyping
  - Graft best into arboretum (as juveniles)

Where to?

- Additional selection steps within and across families before entering orchards and/or trials
  - Need to retain selection data to avoid bias
  - Use cheap sparse chips and impute
- (Sample wild relatives?)
- Essential to keep phenotyping!!!
  - Selectively use pooled DNA
- Watch animal breeding literature as methods evolve

Where to?

- Do not overemphasise the genomic part of the genetic variance
  - It will limit progress in the long run
- Exploit the power of the evaluation with genotypes by increasing the number of families and reducing their size

Where to?

- Do not overemphasise the genomic part of the genetic variance
  - It will limit progress in the long run
- Exploit the power of the evaluation with genotypes by increasing the number of families and reducing their size
- Did I say keep phenotyping?

Where to?

- Do not overemphasise the genomic part of the genetic variance
  - It will limit progress in the long run
- Exploit the power of the evaluation with genotypes by increasing the number of families and reducing their size
- Did I say keep phenotyping?
- Keep phenotyping, I can't say it enough


Thank you