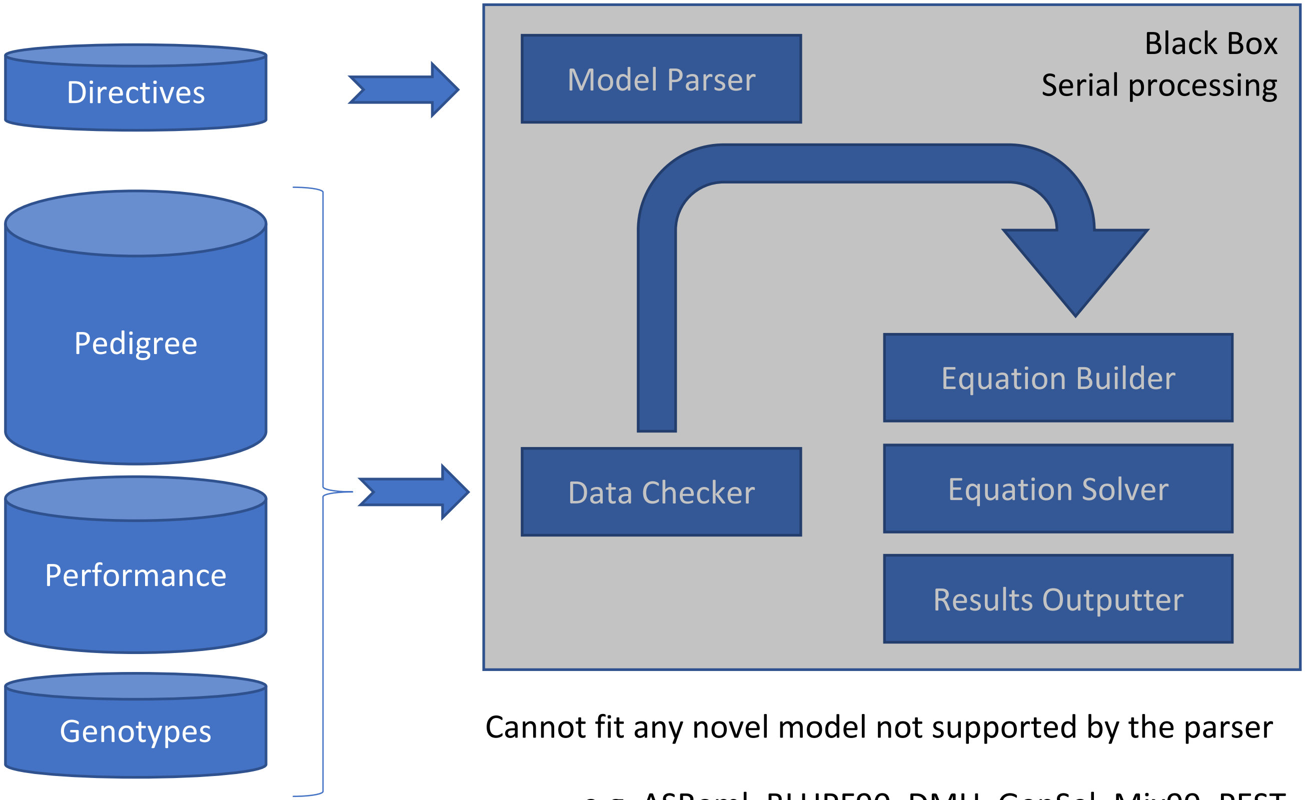
An introduction to BOLT Software for genetic and genomic evaluations

Dorian Garrick (Dorian@ThetaSolutionsLLC.com)

Daniel Garrick & Bruce Golden

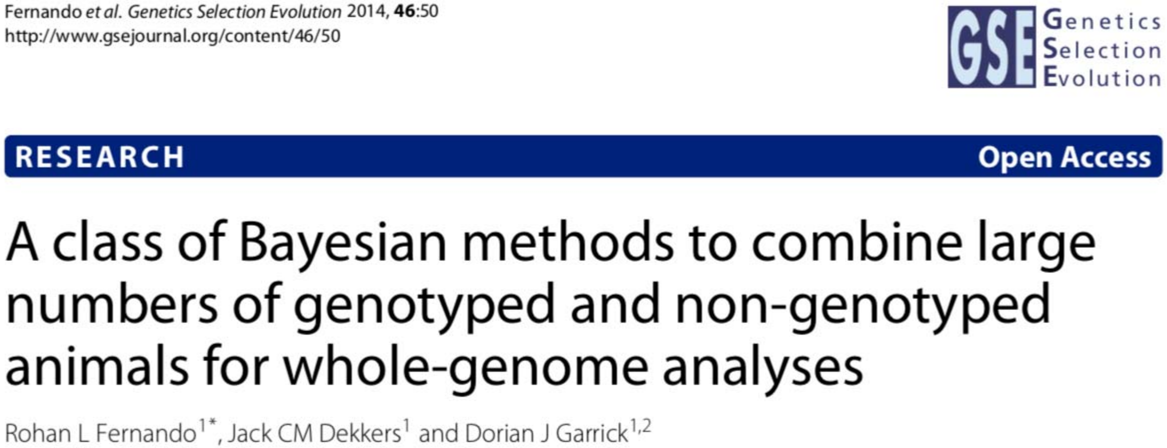
ThetaSolutions

Typical Genetic Evaluation Software



Cannot fit any novel model not supported by the parser

e.g. ASReml, BLUPF90, DMU, GenSel, Mix99, PEST, etc



Fernando *et al. Genetics Selection Evolution* 2014, **46**:50
http://www.gsejournal.org/content/46/50

Genetics Selection Evolution

RESEARCH **Open Access**

A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses

Rohan L Fernando[1*], Jack CM Dekkers[1] and Dorian J Garrick[1,2]

$$\begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\end{bmatrix} = \begin{bmatrix}\mathbf{X}_1\\ \mathbf{X}_2\end{bmatrix}\boldsymbol{\beta} + \begin{bmatrix}\mathbf{Z}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{Z}_2\end{bmatrix}\begin{bmatrix}\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{T}_2\boldsymbol{\alpha} + \boldsymbol{\epsilon}\\ \mathbf{T}_2\boldsymbol{\alpha}\end{bmatrix} + \mathbf{e}$$

$$\begin{bmatrix}\mathbf{X}^{*\prime}\mathbf{X}^{*} & \mathbf{X}^{*\prime}\mathbf{W} & \mathbf{X}_1^{*\prime}\mathbf{Z}_1\\ \mathbf{W}'\mathbf{X}^{*} & \mathbf{W}'\mathbf{W} + \mathbf{I}\frac{\sigma_e^2}{\sigma_\alpha^2} & \mathbf{W}_1'\mathbf{Z}_1\\ \mathbf{Z}_1'\mathbf{X}_1^{*} & \mathbf{Z}_1'\mathbf{W}_1 & \mathbf{Z}_1'\mathbf{Z}_1 + \mathbf{A}^{11}\frac{\sigma_e^2}{\sigma_g^2}\end{bmatrix}\begin{bmatrix}\hat{\boldsymbol{\beta}}^{*}\\ \hat{\boldsymbol{\alpha}}\\ \hat{\boldsymbol{\epsilon}}\end{bmatrix} = \begin{bmatrix}\mathbf{X}'^{*}\mathbf{y}\\ \mathbf{W}'\mathbf{y}\\ \mathbf{Z}_1'\mathbf{y}_1\end{bmatrix}$$

$$\mathbf{W} = \begin{bmatrix}\mathbf{W}_1\\ \mathbf{W}_2\end{bmatrix} = \begin{bmatrix}\mathbf{Z}_1\hat{\mathbf{M}}_1\\ \mathbf{Z}_2\mathbf{M}_2\end{bmatrix}$$

Application for Federal Funding

2013 USDA-AFRI Funding Application

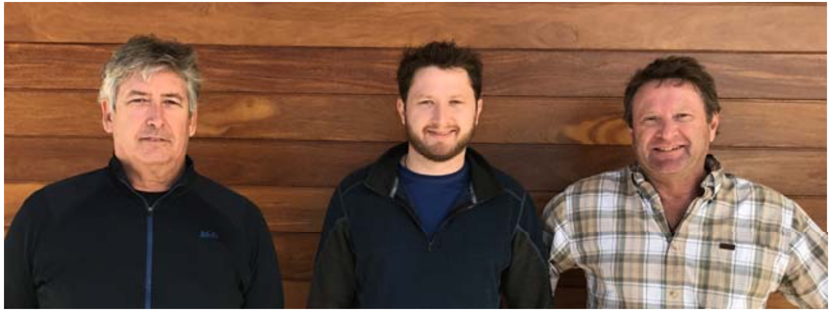
To develop, program, and evaluate the single step marker effect Bayesian regression approach

We regret that the program is ***not*** *able to provide funding*

"The panel would have been more convinced if results using real multi-trait data from any of the commodity groups were presented showing that the method described can achieve the claimed results."

ThetaSolutions

- Formed in 2014 to develop and license a suite of tools (BOLT) for genetic and genomic evaluation



Expansion in 2017

We leverage technology for computer gaming

Computer gaming/animation is >$100 billion per year industry!!

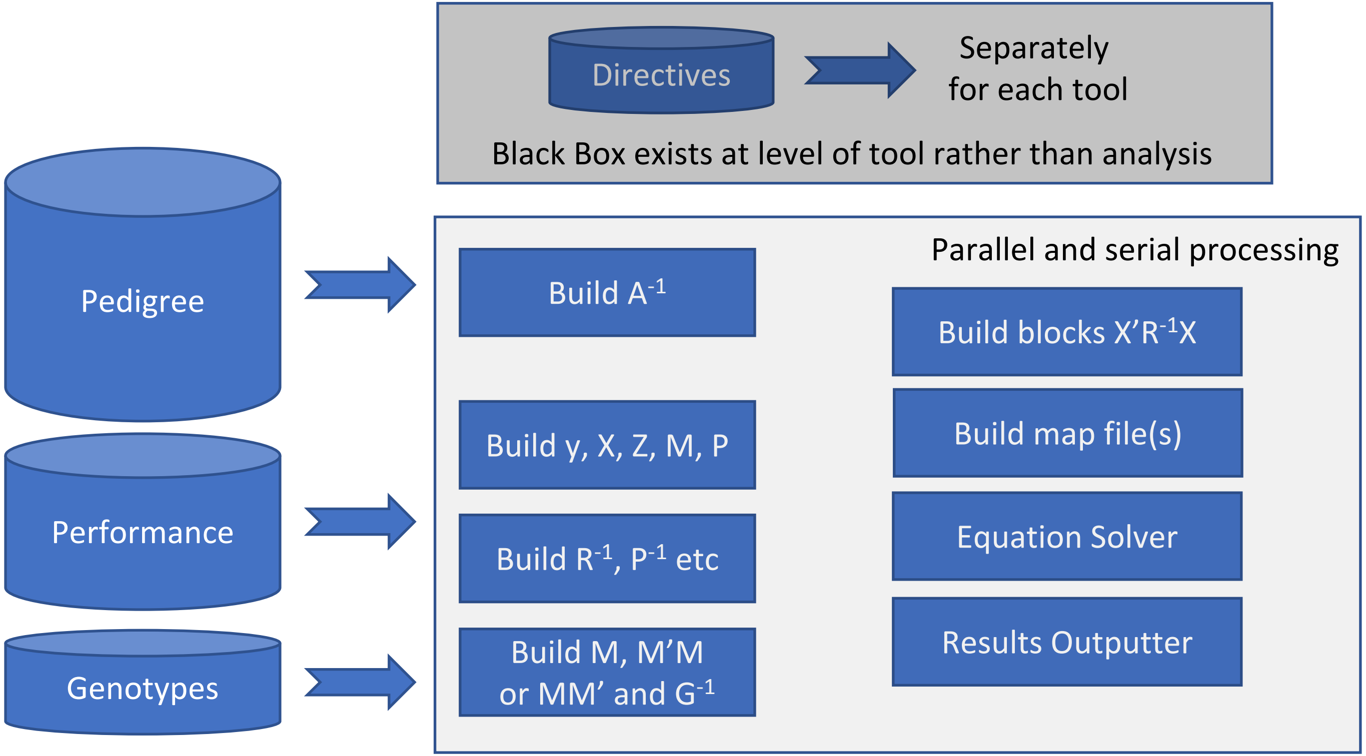
18,688 nodes each with 16 cores and each with NVIDIA GPU

10x more powerful than predecessor Jaguar but uses same space and power

2012 Cray supercomputer at Oak Ridge National Laboratories – fastest in US, 2nd fastest in world
After 6 years, it has now slipped to 5th fastest computer in the world

Cost-effective gaming computers with multiple Titan graphic cards (GPU)

BOLT Evaluation



Directives → Separately for each tool

Black Box exists at level of tool rather than analysis

Pedigree

Performance

Genotypes

Parallel and serial processing

Build A^{-1}

Build y, X, Z, M, P

Build R^{-1}, P^{-1} etc

Build M, M'M or MM' and G^{-1}

Build blocks $X'R^{-1}X$

Build map file(s)

Equation Solver

Results Outputter

Can fit any model you know how to assemble

Single Trait PBLUP

$$y = Xb + Zu + e$$

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + \lambda A^{-1} \end{bmatrix} \begin{bmatrix} b \\ u \end{bmatrix} \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

Given text files (e.g. comma or space delimited)
alphanumeric pedigree (animal, sire, dam)
performance file with phenotype, covariates, class identifiers

BOLT approach

- Construct map files for lhs and rhs

$$\begin{array}{ll} MAP & \\ XpX & XpZ \\ ZpX & ZpZ + Ainv * 3.5 \end{array} \qquad \begin{array}{l} MAP \\ Xpy \\ Zpy \end{array}$$

- Solver like pcgmgpu, pcgbigd, or pcgsbig
  - Provides all the solutions
- Gibbs sampler like ssGibbs
  - Provides MCMC samples, solutions and PEV
    - e.g 6 hours for AHA BW, WW, PWG super-hybrid model with

Multi-trait PBLUP

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} + \begin{bmatrix} Z_1 & 0 \\ 0 & Z_2 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}$$

$$\begin{bmatrix} X_1'R^{11}X_1 & X_1'R^{12}X_2 & X_1'R^{11}Z_1 & X_1'R^{12}Z_2 \\ . & X_2'R^{22}X_2 & X_2'R^{12}Z_1 & X_2'R^{22}Z_2 \\ . & . & Z_1'R^{11}Z_1 + g^{11}A^{-1} & Z_1'R^{12}Z_2 + g^{12}A^{-1} \\ symmetric & . & . & Z_2'R^{22}Z_2 + g^{22}A^{-1} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ u_1 \\ u_2 \end{bmatrix}$$

$$= \begin{bmatrix} X_1'R^{11}y_1 + X_1'R^{12}y_2 \\ X_2'R^{12}y_1 + X_2'R^{22}y_2 \\ Z_1'R^{11}y_1 + Z_1'R^{12}y_2 \\ Z_2'R^{12}y_1 + Z_2'R^{22}y_2 \end{bmatrix}$$

Trivial Extensions

- Introducing EBV from an external analysis
  - Just add relevant blocks computed using other BOLT tools to the relevant blocks of the MME
- Using all kinds of variance-covariance structures
  - relationship matrix for honey bees!
  - relationship matrix allowing selfing (some plants)
  - random regression or autoregressive models
  - heterogeneous genetic and/or residual variances

MME for Single-step “Hybrid” model

$$M_n = A_{ng} A_{gg}^{-1} M_g$$

$$\begin{bmatrix} X'X & X'ZM & X_n'Z_n & X'Z \\ & M'Z'ZM + I\frac{\sigma_e^2}{\sigma_\alpha^2} & M_n'Z_n'Z_n & M'Z'Z \\ & & Z_n'Z_n + A^{nn}\frac{\sigma_e^2}{c\sigma_a^2} & Z_n'Z_n \\ & sym & & Z'Z + A^{-1}\frac{\sigma_e^2}{(1-c)\sigma_a^2} \end{bmatrix} \begin{bmatrix} b \\ \alpha \\ \varepsilon_n \\ \phi \end{bmatrix} = \begin{bmatrix} X'y \\ M'Z'y \\ Z_n'y_n \\ Z'y \end{bmatrix}$$

M_n on off-diagonal can be a big matrix – but is not explicitly required

When few animals genotyped it still might have large M_n

Computations can be simplified by avoiding explicit use of

Great when g>>n

MME for Single-step “Super Hybrid” model

$$\begin{bmatrix} X'X & X_g'Z_gM_g & X_n'Z_n & X'Z \\ & M_g'Z_g'Z_gM_g + I\frac{\sigma_e^2}{\sigma_a^2} + M_n'A^{nn}M_n\frac{\sigma_e^2}{c\sigma_a^2} & M_g'A^{gn}\frac{\sigma_e^2}{c\sigma_a^2} & M_g'Z_g'Z_g \\ & & Z_n'Z_n + A^{nn}\frac{\sigma_e^2}{c\sigma_a^2} & Z_n'Z_n \\ & sym & & Z'Z + A^{-1}\frac{\sigma_e^2}{(1-c)\sigma_a^2} \end{bmatrix} \begin{bmatrix} b \\ \alpha \\ u_n \\ \phi \end{bmatrix} = \begin{bmatrix} X'y \\ M_g'Z_g'y_g \\ Z_n'y_n \\ Z'y \end{bmatrix}$$

Only M_g (and not M_n) appears in the off-diagonal

Practical for multiple-trait national evaluation with large g

Great when g<<n

30 mins to build M_n and its product matrices for 13 million IGS *multibreed* animals with 100,000 genotyped

15 mins for all other matrix products **1.5 hours** for this model for 1,000 PCG iterations

Summary

- BOLT is a set of proprietary command line unix tools
  - Available through an annual licence fee that includes training and consulting
- Empowers an organization to develop efficient turnkey evaluations
  - Without needing to compile or learn a low-level programming language
- We constantly improving algorithms & implementation (monthly updates)
- Leverage NVIDIA hardware developments (eg Titan V vs Titan X vs Black)
  - About 2x faster realized each generation
  - Increased GPU ram offers even greater advances
- We enable end-to-end automated pipelines to drive faster genetic gain