UNIVERSITY OF GEORGIA

Os desafios do uso de *Big Data* no melhoramento genético animal

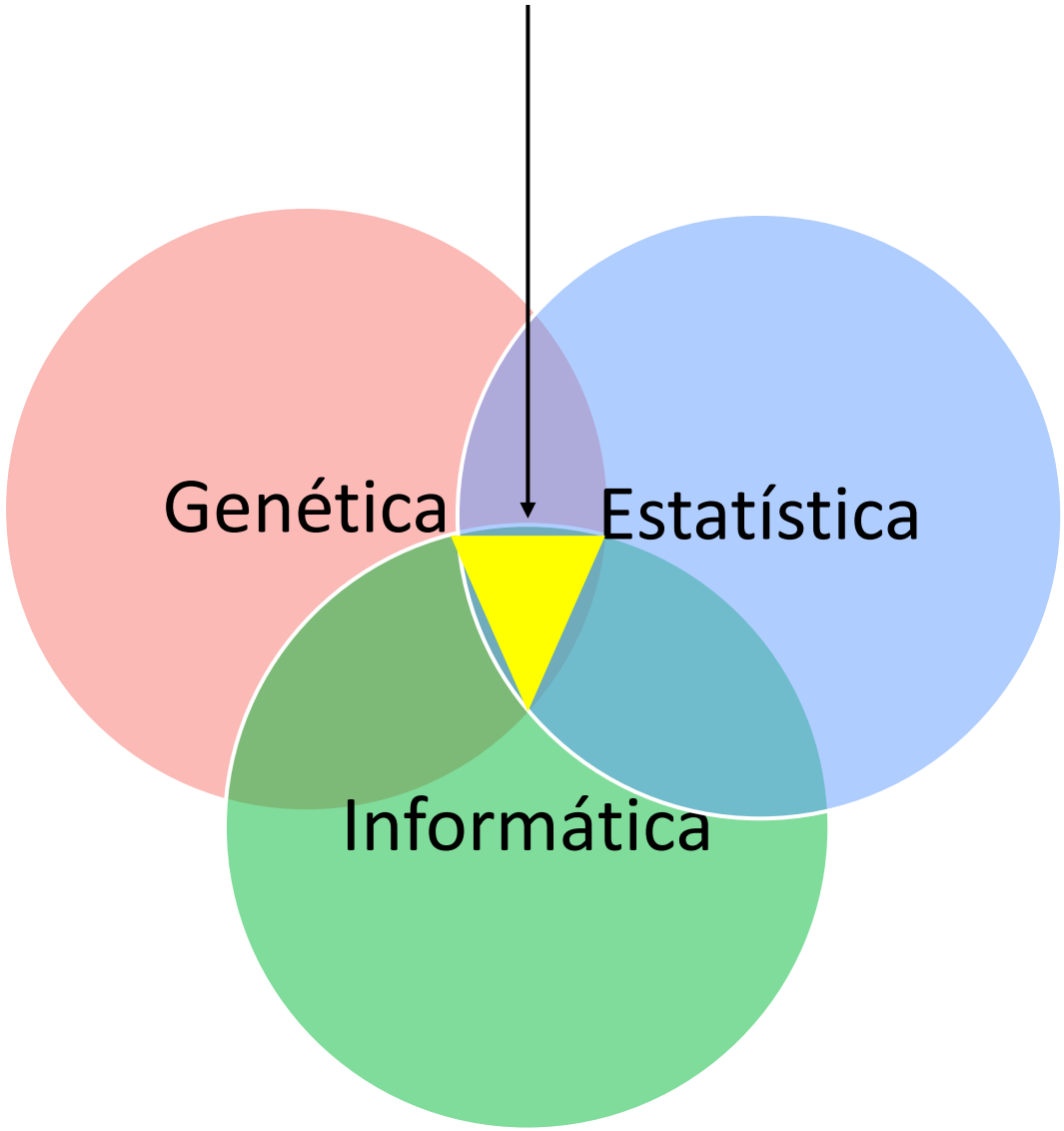
Daniela Lourenco, PhD

I Encontro de Modelagem Estatística – DES – UEM
05/12/2017

+Daniela
Postdoc - UGA:
Melhoramento Genético Animal
Professora e Pesquisadora - UGA
United States
Athens, GA
North Atlantic Ocean
Gulf of California
Mexico
Gulf of Mexico
Cuba
Dominican Republic
Puerto Rico
Guatemala
Caribbean Sea
7.500 Km
Nicaragua
Costa Rica
Panama
Venezuela
Guyana
French Guiana
Suriname
Colombia
Ecuador
Peru
Brazil
Bolivia
Paraguay
Astorga - PR
Chile
Uruguay
Graduação - UEM:
Ciências Biológicas
Mestrado e Doutorado - UEM:
Melhoramento Genético Animal
Google

Grupo de Melhoramento Genético Animal

Melhoramento Genético Animal



Melhoramento genético animal

- Selecionar os melhores animais para serem pais das próximas gerações
- Objetivo geral: melhorar/aumentar a produção
- Objetivo global: produzir alimentos para bilhões de pessoas

Como faz?

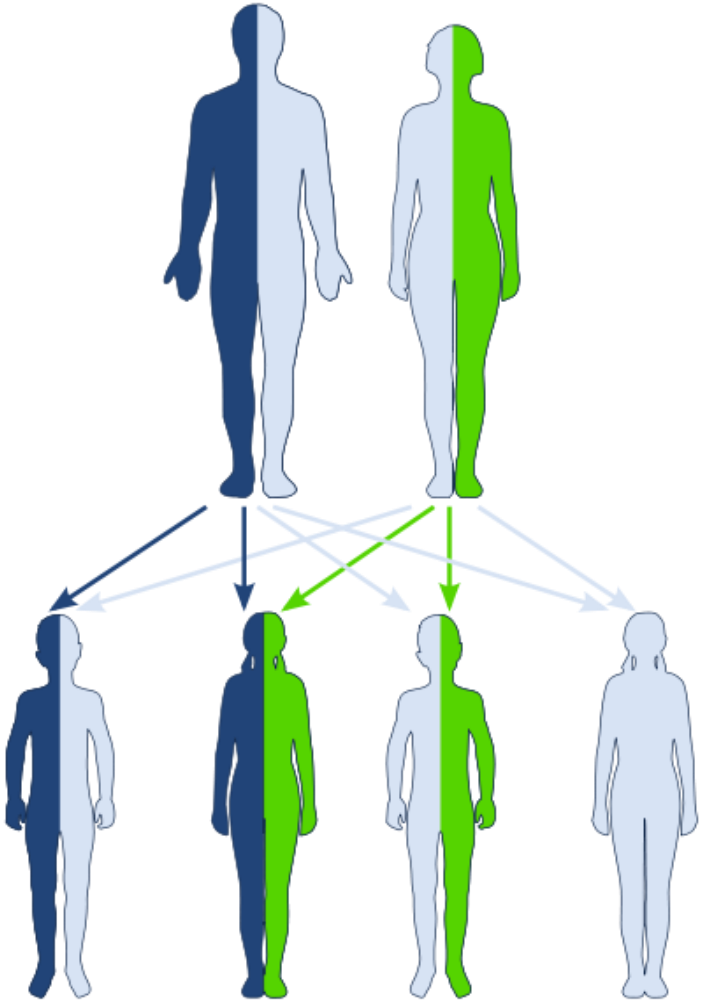
- Mensurar o fenótipo para a característica desejada

$$P = G + E$$

$$\Downarrow$$

Porção herdada

Herança de genes



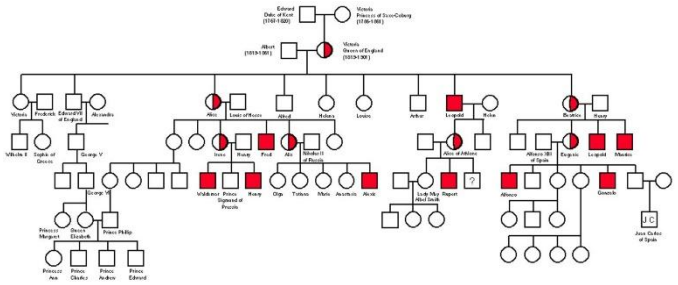
https://commons.wikimedia.org/wiki/File:ABO_system_codominance.svg

- Genes determinam as características
- Conjunto de genes = genoma
- Genoma = manual de instruções
  - Criar
  - Funcionar
  - Reparar

Valor genético

- Transformar o potencial genético do animais em números

+

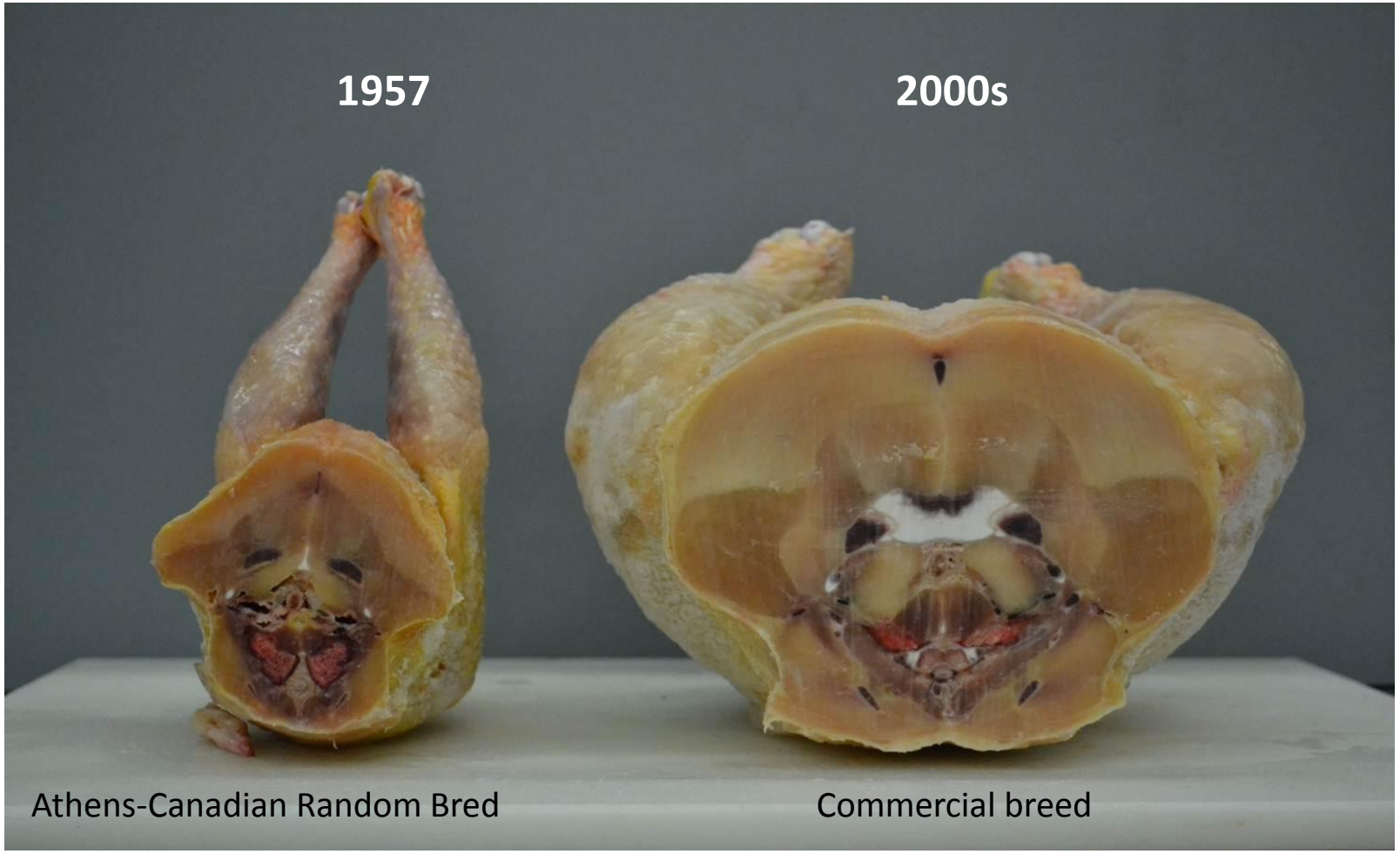


http://www.emersonkent.com/source_documents/victoria_hemophilia.htm

=

Valor genético = EBV

Selecionando animais baseado no valor genético



Courtesy of Dr. Nick Dale, Poultry Science, UGA

Como estimar o valor genético?

$$y = \mathbf{X}\beta + \mathbf{Z}u + e$$

$$u \sim N(0, \mathrm{A}\sigma_u^2)$$

$$e \sim N(0, \mathrm{R}\sigma_e^2)$$

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}+\mathbf{A}^{-1}\lambda \end{bmatrix} \begin{bmatrix} \beta \\ u \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathrm{y} \\ \mathbf{Z}'\mathrm{y} \end{bmatrix}$$

estrutura de covariância

$$\lambda = \frac{\sigma_e^2}{\sigma_u^2}$$

BLUP
(Henderson, 1950)

Matriz de parentesco entre os animais (**A**)

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

- **A** é fácil de ser construída para poucos animais
  - Proporção de genes compartilhados entre parentes
- **A** deve ser invertida
- Custo computacional para inversão de matrizes = O($\boldsymbol{n}^3$)

“Nossa habilidade de gerar dados em melhoramento genético animal está ultrapassando nossa habilidade de analisá-los”
Arthur Gilmour

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

1963

Reprinted From
STATISTICAL GENETICS AND PLANT BREEDING
NAS–NRC 982 1963

Selection Index and Expected Genetic Advance

C. R. HENDERSON
Department of Animal Husbandry, Cornell University, Ithaca, New York

Sem recursos tecnológicos para inverter a matriz **A** in 1963

BLUP MME ficou inativo por um tempo

1976

BIOMETRICS *32*, 69–83
March, 1976

A SIMPLE METHOD FOR COMPUTING THE INVERSE OF A NUMERATOR RELATIONSHIP MATRIX USED IN PREDICTION OF BREEDING VALUES

C. R. HENDERSON

Department of Animal Science, Cornell University, Ithaca, New York 14853, U.S.A.

1988

Research-Article *J. Dairy Sci.*

Additive Genetic Model with Groups and Relationships

R.L. Quaas [1]

Como construir a $\mathbf{A}^{-1}$

- **A** é densa
- $\mathbf{A}^{-1}$ é esparsa

- Recursão para $\mathbf{A}^{-1}$

$$u_i \mid u_1, u_2, .., u_{i-1}, u_{i+1}, .., u_n = u_i \mid u_{si}, u_{di}$$

$$u_i = 0.5(u_{s_i} + u_{d_i}) + \varphi_i \qquad \mathbf{u} = \mathbf{P}\mathbf{u} + \mathbf{\Phi} \qquad \mathbf{A}^{-1} = (\mathbf{I} - \mathbf{P})'\mathbf{M}^{-1}(\mathbf{I} - \mathbf{P})$$

Henderson (1976); Quaas (1988)

- Custo computacional alto se **P** é densa
- P –> 2 elementos $\neq$ 0 por linha

Como construir a $\mathbf{A}^{-1}$

$$\mathbf{A}^{-1} = (\mathbf{I} - \mathbf{P})'\mathbf{M}^{-1}(\mathbf{I} - \mathbf{P})$$

```
ImP=(/1., -.5, -.5/)
Minv=(/2., 4/3., 1., 0./)
ainv=0.0

  do
   read(2,*,iostat=io) animal, sire, dam,par_stat
   if (io /= 0) exit
   p(1)=animal
   p(2)=sire
   p(3)=dam
    do i=0,nanimal
      do k=1,3
        do l=1,3
          ainv(k,l)=ainv(k,l)+ImP(k)*ImP(l)*Minv(par_stat)
        enddo
      enddo
    enddo
```

Exemplos de banco de dados (AB&G - UGA)

- US Holsteins
  - 70 milhões de animais no pedigree
  - 42 milhões usados nas análises
  - $\mathbf{A}^{-1}$ = 42M x 42M
  - 32 milhões de animais com fenótipo
  - Modelos multi-característica
  - Vários efeitos aleatórios
  - Características contínuas e categóricas

- American Angus
  - 10 milhões de animais no pedigree
  - 10 milhões usados nas análises
  - $\mathbf{A}^{-1}$ = 10M x 10M
  - 8.7 milhões de animais com fenótipo
  - Modelos multi-característica
  - Vários efeitos aleatórios
  - Características contínuas e categóricas

Número de equações

- American Angus
  - Peso 1, 205, 365 dias

$$y = \mathbf{X}\beta + \mathbf{Z}_1 u + \mathbf{Z}_2 m + \mathbf{W} mpe + e$$

- $\beta = 640.985 * 3$
- $u = 10.173.507 * 3$
- $m = 10.173.507 * 3$
- $mpe = 10.173.507 * 3$

✓ Número de equações = 93.484.518

Número de equações

- **US Holsteins**
  - Leite, gordura, proteína aos 305 dias

$$y = \mathbf{X}\beta + \mathbf{K}hs + \mathbf{Z}u + \mathbf{W}pe + e$$

- $\beta = (388.6799+273+1+1)*3$
- $hs = 7.311.252*3$
- $u = 29.651.923*3$
- $pe = 21.527.040*3$

✓ Número de equações = 187.131.867

Número de equações

- **US Holsteins**
  - 18 características de tipo racial

$$y = \mathbf{X}\beta + \mathbf{K}hs + \mathbf{Z}u + \mathbf{W}pe + e$$

- $\beta = (392.069+704+32) * 18$
- $hs = 3.387.024 * 18$
- $u = 9.730.972 * 18$
- $pe = 6.550.442 * 18$

✓ Número de equações = 361.102.374

Número de equações

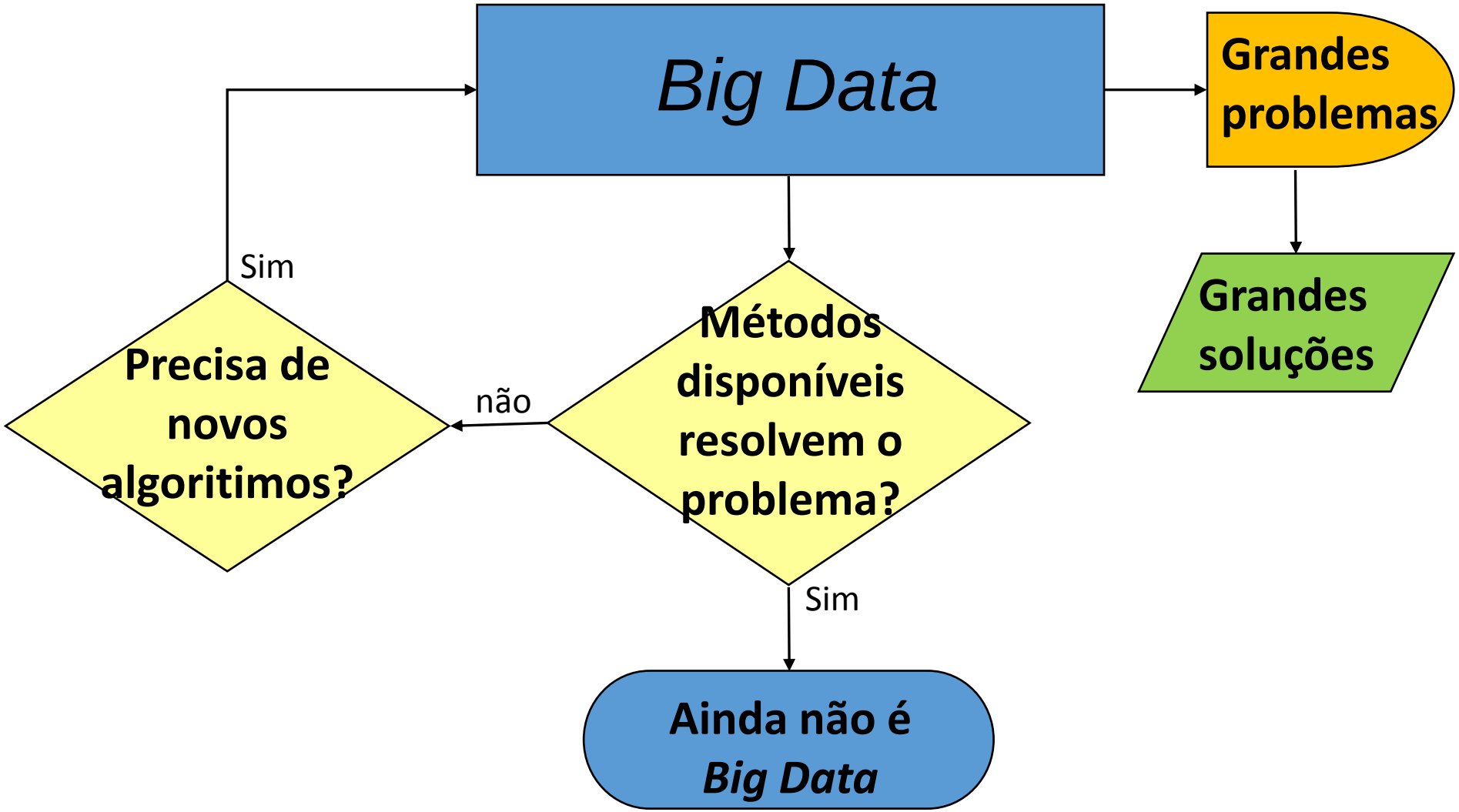
- **Canadian Holsteins**
  - Leite, gordura, proteína no dia do teste *3 partos = 9 características

$$y = \mathbf{X}_1\beta_1 + \sum_{j=0}^{4}\mathbf{X}_{2j}\,\beta_{2_j} + \sum_{j=0}^{4}\mathbf{K}_j\,hy_j + \sum_{j=0}^{4}\mathbf{Z}_j\,u_j + \sum_{j=0}^{4}\mathbf{W}_j\,pe_j + E$$

10 classes de variância residual cada característica

- $\beta_1 = 136.017 * 9; \qquad \beta_2 = 96 * 9 * 5$
- $hy = 15.984 * 9 * 5$
- $u = 198.555 * 9 * 5$
- $pe = 118.404 * 9 * 5$

✓ Número de equações = 430.167.393


Big Data
Grandes problemas
Grandes soluções
Métodos disponíveis resolvem o problema?
não
Precisa de novos algoritimos?
Sim
Sim
Ainda não é Big Data

Big Data – AB&G UGA software

nce.ads.uga.edu/wiki/doku.php?id=application_programs

Apps computing R

[[application_programs]]

Trace: • readme.renumf90 • readme.blupf90 • start • application_programs

The programs support mixed models with multiple-correlated effects, multiple animal models and dominance.

Most of the programs now support being called with the parameter file as an argument, e.g.

- BLUPF90 - BLUP in memory
- REMLF90 - accelerated EM REML
- QXPAK - joint analysis of QTL and polygenic effects (M. Perez-Enciso) QxPak web page
- AIREMLF90 - Average Information REML with several options including EM-REML and heterogeneous residual variances (S. Tsuruta)
- CBLUP90 - solutions for bivariate linear-threshold models
- CBLUP90THR - as above but with thresholds computed and many linear traits (B. Auvray)
- CBLUP90REML - as above but with quasi REML (B. Auvray)
- GIBBSF90 - simple block implementation of Gibbs sampling
- GIBBS1F90 - as above but faster for creating mixed model equations only once
- GIBBS2F90 - as above but with joint sampling of correlated effects
- GIBBS3F90 - as above with support for heterogeneous residual variances
- GIBBS2CEN - support censored data (J. Arango)
- GIBBS2RECUR - support recursive models (J. P. Sanchez)
- POSTGIBBSF90 - statistics and graphics for post-Gibbs analysis (S. Tsuruta)
- THRGIBBSF90 - Gibbs sampling for any combination of categorical and linear traits (D. Lee)
- THRGIBBS1F90 - as above but simplified with several options (S. Tsuruta)
- RENUMF90 - a renumbering program that also can check pedigrees and assign unknown parent groups; supports large data sets
- INBUPGF90 - a program to calculate inbreeding coefficients with incomplete pedigree (I. Aguilar)
- SEEKPARENTF90 - a program to verify paternity and parent discovery using SNP markers (I. Aguilar)
- PREDICTF90 - a program to calculate adjusted y, y_hat, and residuals (I. Aguilar)
- PREDF90 - a program to predict direct genomic value (DGV) for animals based on genotypes and SNP solution

Available by request

- MRF90 - Method R program suitable for very large data sets; contact T. Druet.
- COXF90 – Bayesian Cox model - contact J. P. Sanchez (JuanPablo.Sanchez@irta.cat)
- BLUPF90HYP – BLUPF90 with hypothesis testing (F and Chi2 tests) - contact J. P. Sanchez as above

Available only under research agreement

- BLUP90IOD2 - BLUP by iteration on data with PCG for very large models (S. Tsuruta)
- CBLUP90IOD - BLUP by iteration on data for threshold-linear models
- ACCF90 - approximation of accuracies for breeding values
- BLUP90MBE - BLUP by iteration on data with support for very large models for multi-breed evaluations
- BLUP90ADJ - BLUP data preadjustment tool

Included in application programs

- PREGSF90 – genomic preprocessor that combines genomic and pedigree relationships (I. Aguilar)
- POSTGSF90 – genomic postprocessor that extracts SNP solutions after genomic evaluations (single step, GBLUP) (I. Aguilar)

FORTRAN

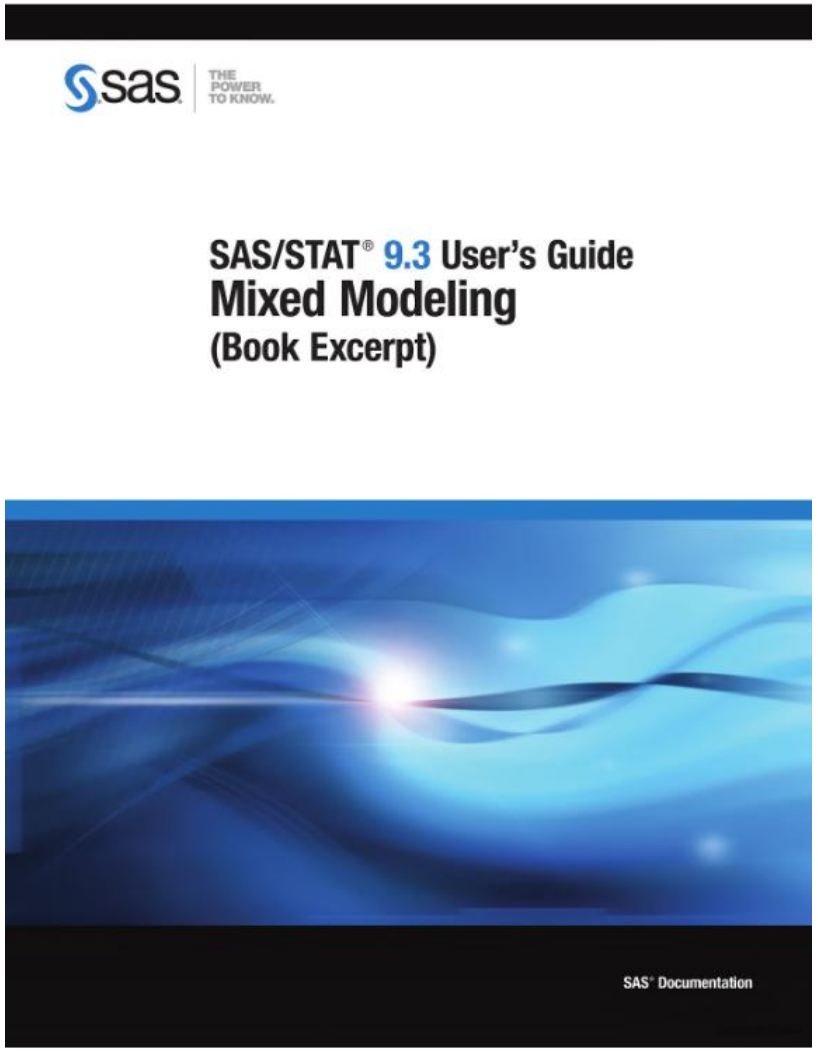
Computação numérica e científica

Como resolver grandes sistemas de equações?

- **Métodos finitos**
  - Ax = b
  - LUx=b or LL'x=b or LDL'x=b
  - Soluções com alta precisão
  - Custo O($\boldsymbol{n^3}$)
  - Usados para obter PEV
  - Matrix esparsa
    - FSPACK (Misztal et al., 1993)
    - YAMS (Masuda et al., 2015)

- **Métodos iterativos**
  - Gauss-Seidel
  - Gradiente conjugado precondicionado (PCG)
    - $M^{-1}Ax = M^{-1}b$
  - PCG + Iteração nos dados (Tsuruta et al., 2001)
  - Coeficientes do LHS gerados a cada iteração
  - Simples operações com vetores e escalares
  - Baixo custo computacional

Resolvendo um grande número de equações



Overview: HPMIXED Procedure

The HPMIXED procedure uses a number of specializ ques to
models with variance component structure. The HPMIX desig
estimation problems involving a large number of fixed e dom e
number of observations.

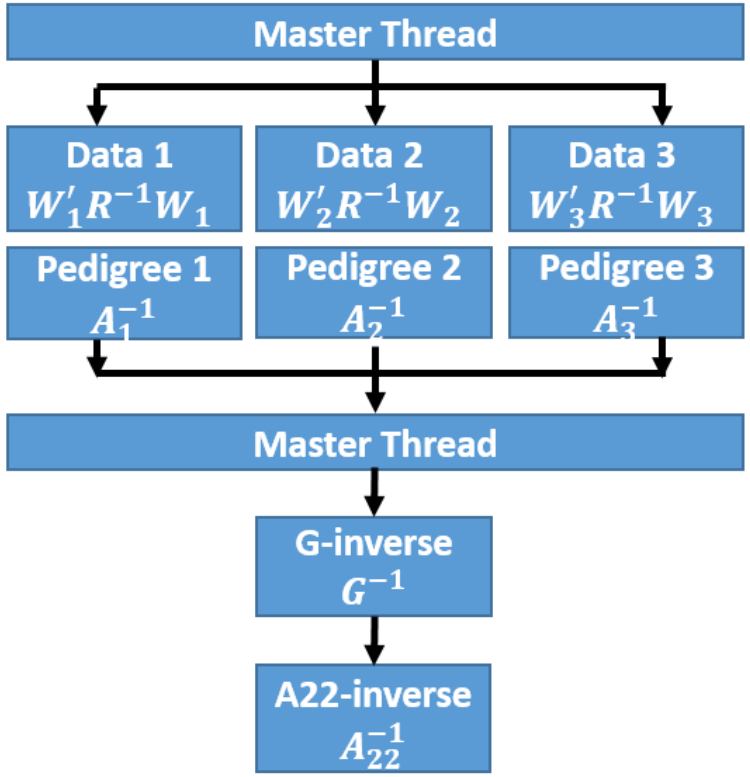
Conjugate Gradient Algorithm and Iteration-on-Data Technology

Solving the mixed model equations is a critical component of linear mixed model analysis. The two main components of the preconditioned conjugate gradient (PCCG) algorithm are preconditioning and matrix-vector product computing (Shewchuk 1994). The algorithm is guaranteed to converge to the solution within n_e iterations, where n_e is equal to the number of distinct eigenvalues of the mixed model equations. This simple yet powerful algorithm can be easily implemented with an iteration-on-data (IOD) technique (Tsuruta, Misztal, and Stranden 2001) that can yield significant savings of memory resources.

The combination of the PCCG algorithm and iteration on data makes it possible to efficiently compute best linear unbiased predictors (BLUPs) for the random effects in mixed models with large mixed model equations.

Como resolver grandes sistemas de equações?

- **Métodos iterativos**
  - PCG + Iteração nos dados (Tsuruta et al., 2001)
  - PCG + Iteração nos dados + OpenMP (Yutaka Masuda)



Tempo de computação

PCG (bloco-diagonal) + Iteração nos dados + OpenMP

- American Angus
  - 3 características

✓ Número de equações $= 93.484.518$

✓ Tempo $= 5{:}30\text{h}$

- US Holsteins
  - 18 características

✓ Número de equações $= 361.102.374$

✓ Tempo $= 10\text{h}$

Tempo de computação

PCG (bloco-diagonal) + Iteração nos dados + OpenMP

- Canadian Holsteins
  - 9 características MRA

✓ $\text{Número de equações} = 430.167.393$

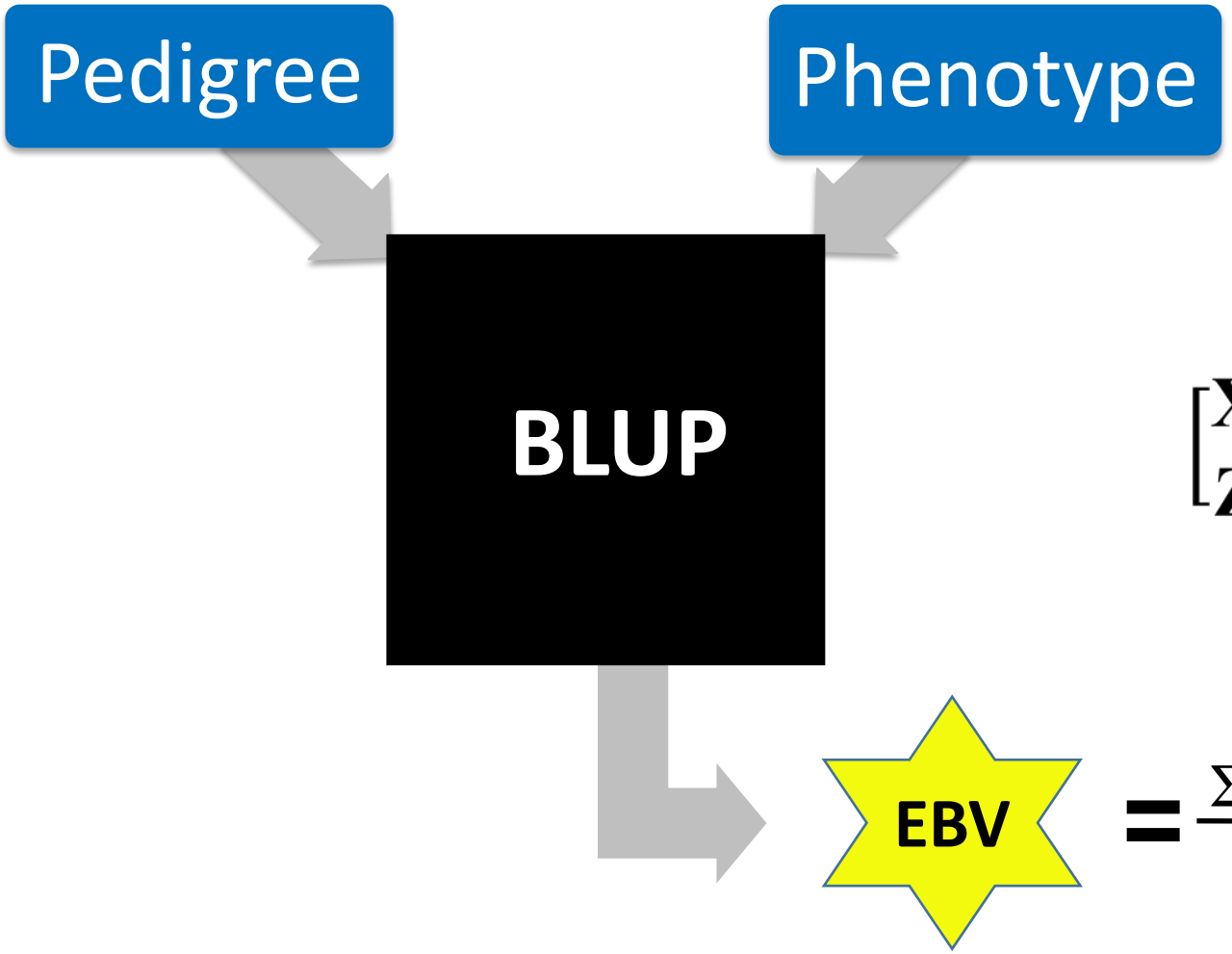
✓ $\text{Tempo} = 4.3 \text{ dias}$

Resolver grandes sistemas de equações

✓

Proximo desafio…

Genômica – um novo desafio



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

$$= \frac{\Sigma \text{ efeitos dos genes}}{2}$$

Genômica – um novo desafio

E se conseguíssemos alguma informação a respeito dos genes?

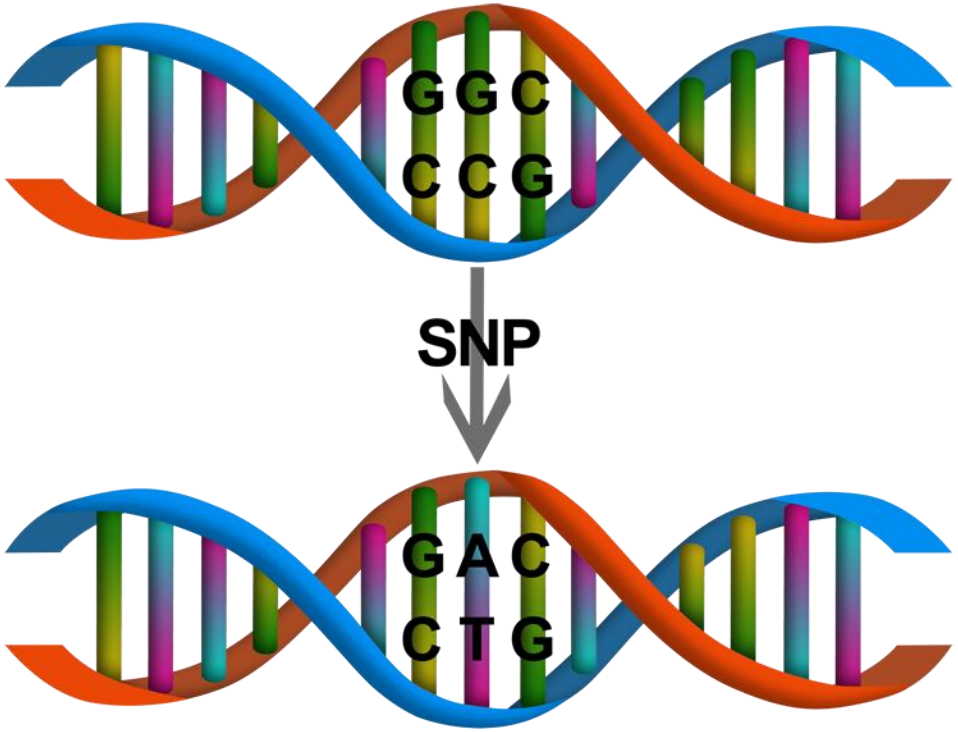
Conseguiríamos estimar EBV de forma mais acurada?

Real proporção de genes X esperança da proporção de genes

Melhor estrutura de covariância

Qual tipo de informação Genômica?

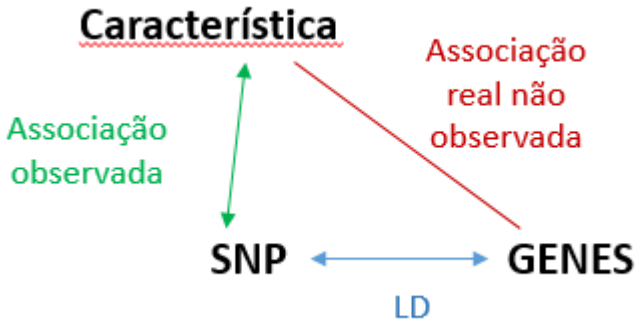
Indivíduo 1



Indivíduo 2

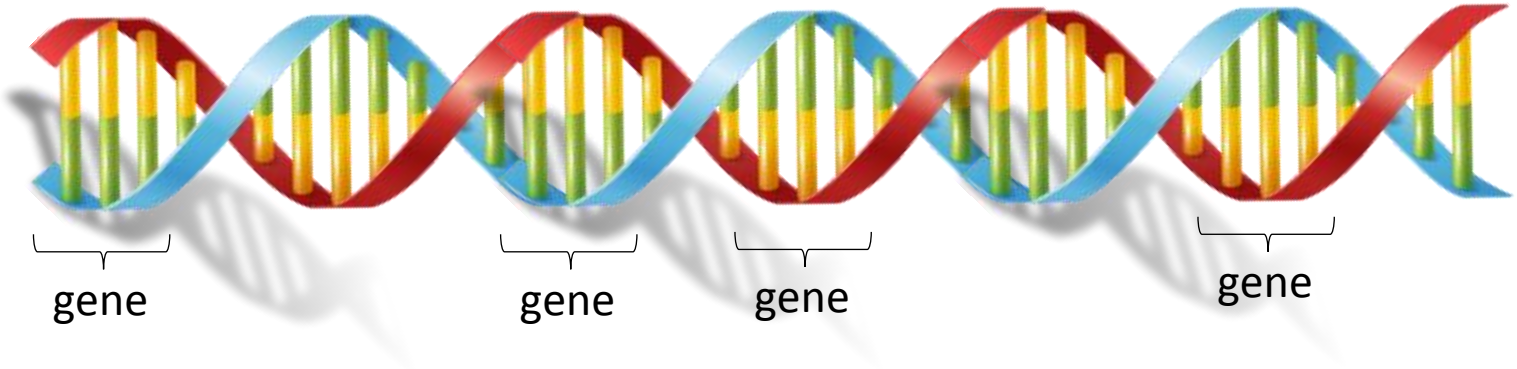
http://www.thinnergene.com/about-thinnergene/genetics-101/

- **SNP = Erros no DNA**
- Reparados
- Transmitidos
- Alteram a performance
- Efeito benéfico
- Efeito deletério



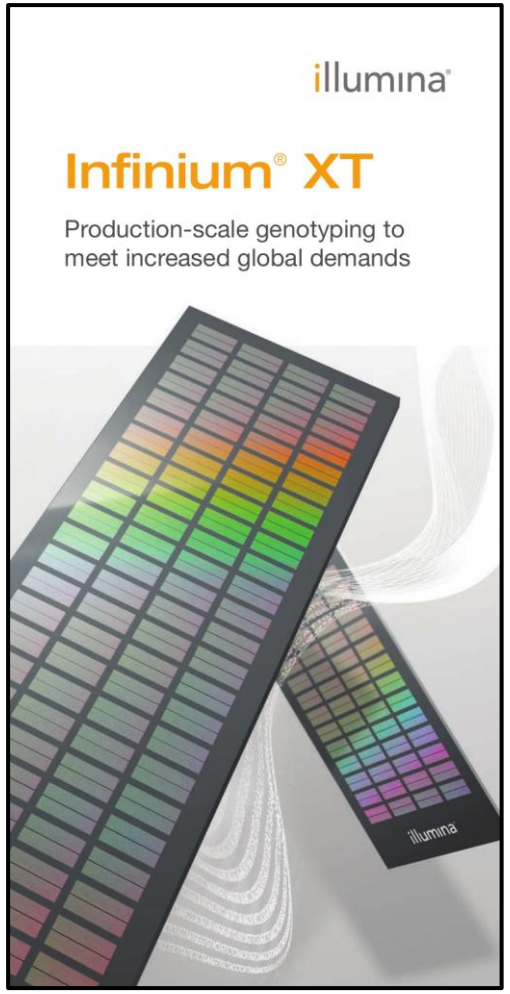
Qual tipo de informação Genômica?

- Características de interesse são poligênicas

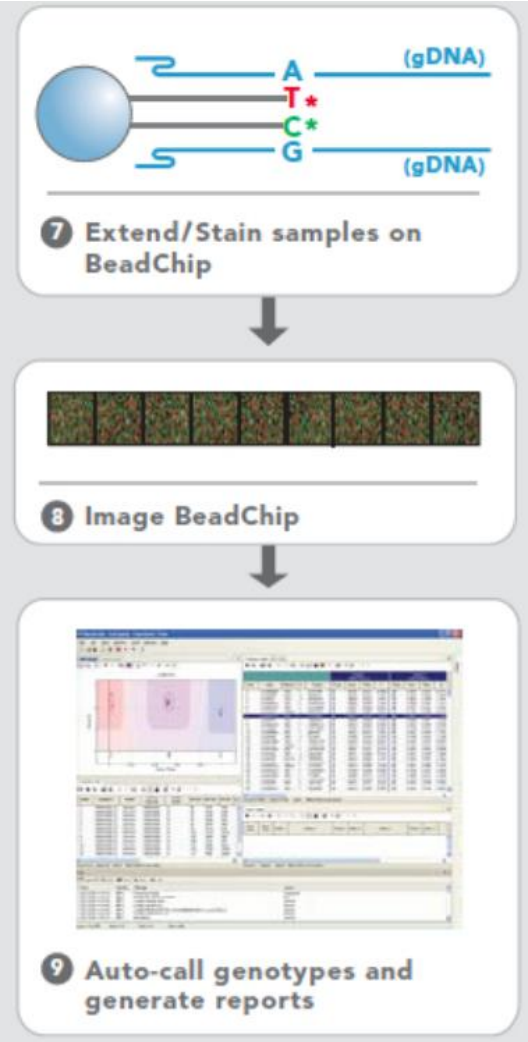


Milhares de genes → Milhares de SNP

Qual tipo de informação Genômica?



https://www.illumina.com/products



- 50.000 SNP por animal
  - 2009: $300
  - 2017: $25 - $50

Qual o tamanho dos bancos de dados?

```
USA_287506 022101021221111100021111 ->
USA_287507 220000111021212011021021 ->
USA_287508 121100122121221111020110 ->
USA_287509 220000222020222022020020 ->
USA_287510 110101211111121121020010 ->
USA_287511 121001010121102000022022 ->
USA_287513 220000122121221111020110 ->
USA_287514 121001121120111012011020 ->
USA_287515 011002010121101101121111 ->
USA_287516 121001211021212111121010 ->
USA_287518 121001010121102000022022 ->
USA_287519 220000222020222022020020 ->
USA_287520 121001121120111012011020 ->
USA_287521 220000222020222022020020 ->
USA_287522 110101211111121121020010 ->
USA_287523 022101021221111100021111 ->
USA_287524 022101021221111100021111 ->
USA_287525 121100122121221111020110 ->
USA_287526 011002010121101101121111 ->
USA_287527 220000111021212011020020 ->
USA_287528 121001010121102000022022 ->
USA_287529 220000222020222022020020 ->
USA_287530 012111121220111111021122 ->
USA_287531 112101020111102101012011 ->
```

BB = 0

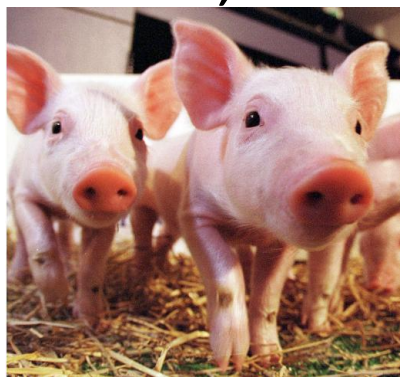
Bb = 1

bb = 2

NA = 5

15Gb

250,000



http://sesenfarm.com/raising-pigs/

19Gb

400,000

http://www.angus.org/AGI/default.aspx

10Gb

182,600

https://www.usjersey.com/AJCA-NAJ-JMS/AJCA/AnimalIdentificationServices/HerdRegister.aspx

107Gb

1,900,000

http://www.holsteinusa.net/programs_services/backgrounds.html

Como usar a informação genômica?

USA_287506	022101021221111100021111 ->
USA_287507	220000111021212011021021 ->
USA_287508	121100122121221111020110 ->
USA_287509	220000222020222022020020 ->
USA_287510	110101211111121121020010 ->
USA_287511	121001010121102000022022 ->
USA_287513	220000122121221111020110 ->
USA_287514	121001121120111012011020 ->
USA_287515	011002010121101101121111 ->
USA_287516	121001211021212111121010 ->
USA_287518	121001010121102000022022 ->
USA_287519	220000222020222022020020 ->
USA_287520	121001121120111012011020 ->
USA_287521	220000222020222022020020 ->
USA_287522	110101211111121121020010 ->
USA_287523	022101021221111100021111 ->
USA_287524	022101021221111100021111 ->
USA_287525	121100122121221111020110 ->
USA_287526	011002010121101101121111 ->
USA_287527	220000111021212011020020 ->
USA_287528	121001010121102000022022 ->
USA_287529	220000222020222022020020 ->
USA_287530	012111121220111111021122 ->
USA_287531	112101020111102101012011 ->

M

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

BLUP

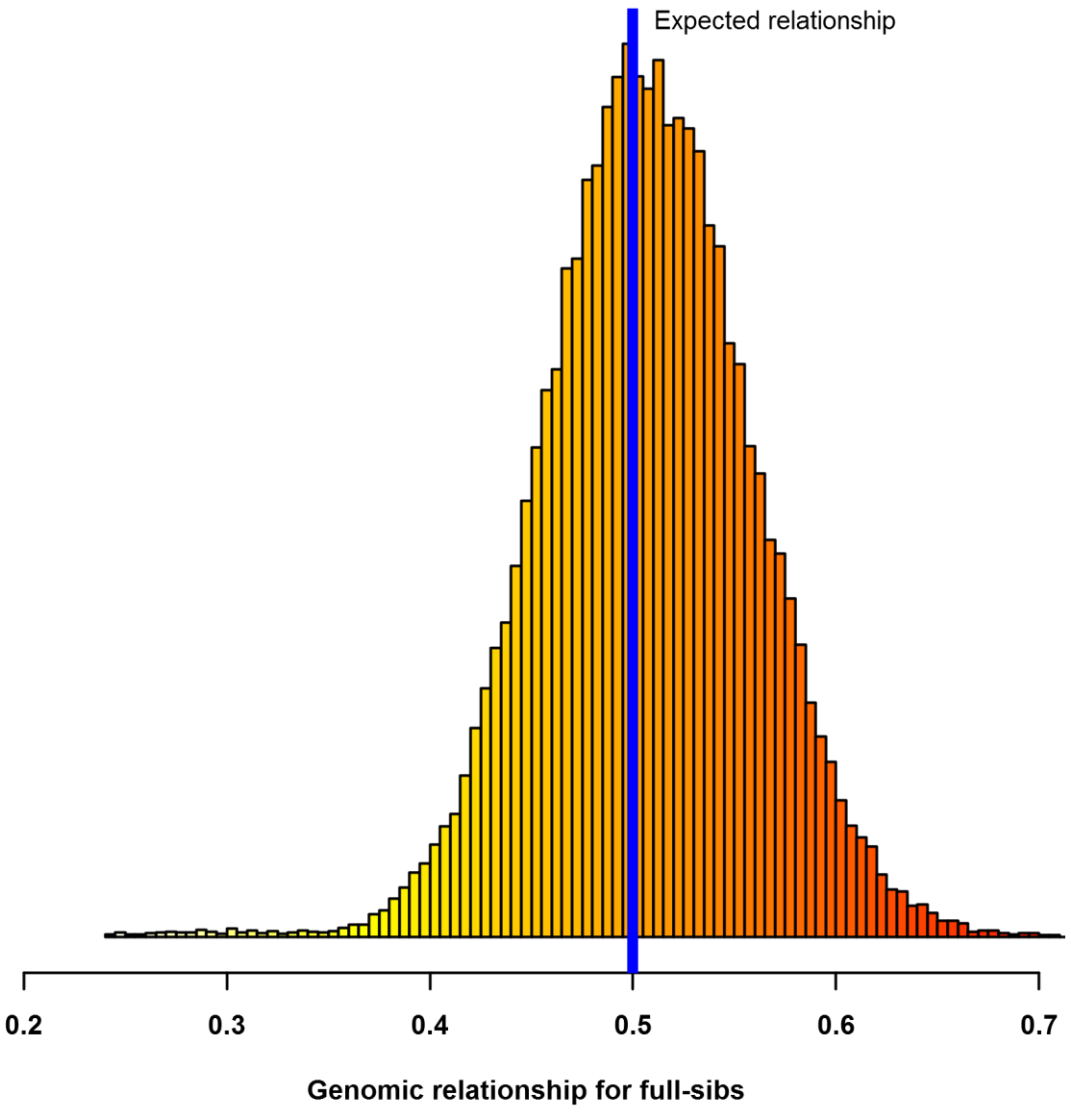
$$\mathbf{G} = \mathbf{MM'}/\mathbf{k}$$

Proporção de SNP compartilhados entre TODOS

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z}+\mathbf{G}^{-1}\lambda \end{bmatrix} \begin{bmatrix} \beta \\ u \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{bmatrix}$$

GBLUP

Comparando **A** e **G**



Lourenco et al. (2015a)

Problemas?

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}+\mathbf{G}^{-1}\lambda \end{bmatrix} \begin{bmatrix} \beta \\ u \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}$$

GBLUP

- Apenas animais genotipados no modelo
- Angus
  - 10M de animais na **A**
  - 400mil na **G**

- Solução 1
  - GBLUP para animais genotipados
  - BLUP para não genotipados
  - Múltiplos passos para combinar as informações

Solução 2

- Combinar as estruturas de covariância
  - Única matriz com as informações de **A** e **G**

Misztal et al., 2009

Legarra et al., 2009

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}$$

Joint distribution of EBV for ungenotyped and genotyped

$$p(\mathbf{u}_1, \mathbf{u}_2) = p(\mathbf{u}_2)\, p(\mathbf{u}_1 | \mathbf{u}_2)$$

$$p(\mathbf{u}_2) = N(\mathbf{0}, \mathbf{G}) \quad \Longrightarrow \quad Var(\mathbf{u}_2) = \mathbf{G}$$

$$p(\mathbf{u}_1 | \mathbf{u}_2) = N\left(\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{u}_2, \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\right)$$

Solução 2

- Combinar as estruturas de covariância
  - Única matriz com as informações de **A** e **G**

Misztal et al., 2009

Legarra et al., 2009

Variância da predição de genótipos para animais não genotipados

$$Var\begin{pmatrix}\mathbf{u}_1\\ \mathbf{u}_2\end{pmatrix}=\mathbf{H}=\begin{bmatrix}\mathbf{H}_{11} & \mathbf{H}_{12}\\ \mathbf{H}_{21} & \mathbf{H}_{22}\end{bmatrix}=$$

$$\begin{bmatrix}\overbrace{\mathbf{A}_{11}-\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}}+\overbrace{\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}} & \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G}\\ \mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{G}\end{bmatrix}$$

Erro de predição

A predição «gera» uma covariância

G construída com base nos SNP

Solução 2 - Single-step GBLUP

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z}+\mathbf{H}^{-1}\lambda \end{bmatrix} \begin{bmatrix} \beta \\ u \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{bmatrix}$$

ssGBLUP

UGA

Misztal et al., 2009

Legarra et al., 2009

Aguilar et al., 2010

- **H** é complicada

$$\mathbf{H} = \mathbf{A} + \begin{bmatrix} \mathbf{A}_{12}\mathbf{A}_{22}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{I} \\ \mathbf{I} \end{bmatrix} [\mathbf{G} - \mathbf{A}_{22}] [\mathbf{I} \quad \mathbf{I}] \begin{bmatrix} \mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}$$

- $\mathbf{H}^{-1}$ é simples

$$\mathrm{H}^{-1} = \mathrm{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \mathrm{G}^{-1} - \mathrm{A}_{22}^{-1} \end{bmatrix}$$

Problemas?

$$H^{-1} = A^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & G^{-1} - A_{22}^{-1} \end{bmatrix}$$

- Recursão para **A**$^{-1}$

$$u_i = 0.5(u_{s_i} + u_{d_i}) + \varphi_i \qquad \mathbf{u} = \mathbf{P}\mathbf{u} + \mathbf{\Phi}$$

- $\mathbf{A}_{22}^{-1}$

$$\mathbf{A}_{22}^{-1} q = \mathbf{A}^{22} q - (\mathbf{A}^{12})'[(\mathbf{A}^{11})^{-1}[\mathbf{A}^{12} q]]$$

- Recursão para **G**$^{-1}$
  - Dividir os animais genotipados em *core* and *non-core*

Misztal et al. (2014)
Lourenco et al. (2015b)

$$u_i \mid u_1, u_2, \ldots, u_{i-1} = \sum_{j = \text{core}} p_{ij} u_j + \varepsilon_i \qquad \mathbf{u}_n = \mathbf{P}_{nc}\mathbf{u}_c + \mathbf{\Phi}_n$$

APY – algoritimo para *core* e *non-core*

USA_287506
USA_287507
USA_287508
USA_287509
USA_287510
USA_287511
USA_287513
USA_287514
USA_287515
USA_287516
USA_287518
USA_287519
USA_287520
USA_287521
USA_287522
USA_287523
USA_287524
USA_287525
USA_287526
USA_287527
USA_287528
USA_287529
USA_287530

core

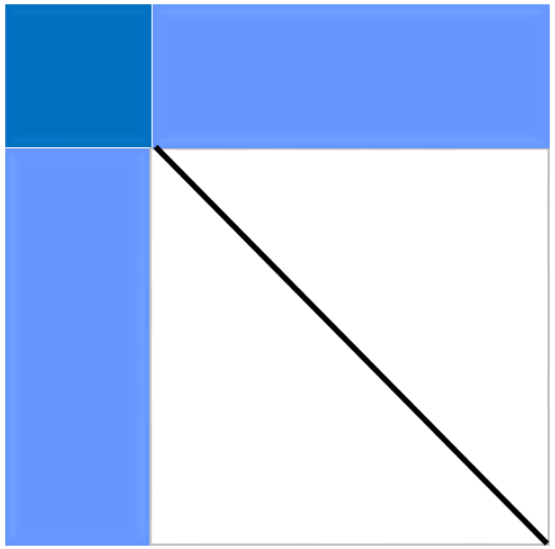
non-core

$$\mathbf{G}_{\mathrm{APY}}^{-1} = \begin{bmatrix} \mathbf{G}_{cc}^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} -\mathbf{G}_{cc}^{-1}\mathbf{G}_{cn} \\ \mathrm{I} \end{bmatrix} \mathbf{M}_{nn}^{-1} \begin{bmatrix} -\mathbf{G}_{nc}\mathbf{G}_{cc}^{-1} & \mathrm{I} \end{bmatrix}$$

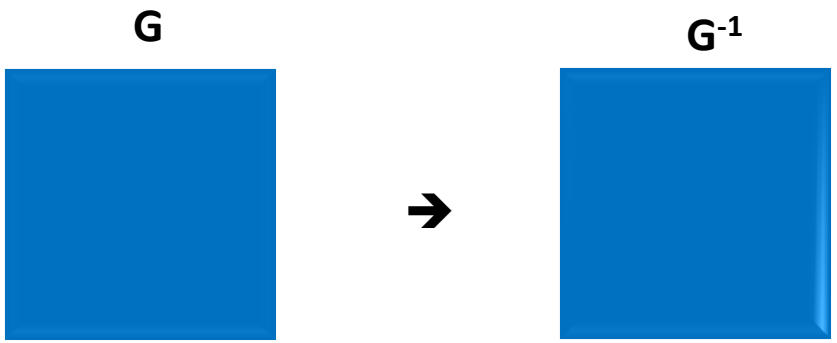
$$\mathrm{M}_{nn} = \mathrm{diag}\{\mathrm{g}_{ii} - \mathrm{g}_{ic}\mathbf{G}_{cc}^{-1}\,\mathrm{g}_{ci}\}$$

$$\mathbf{G} = \begin{bmatrix} \mathbf{G}_{\mathrm{cc}} & \mathbf{G}_{\mathrm{cn}} \\ \mathbf{G}_{\mathrm{nc}} & \mathbf{G}_{\mathrm{nn}} \end{bmatrix}$$

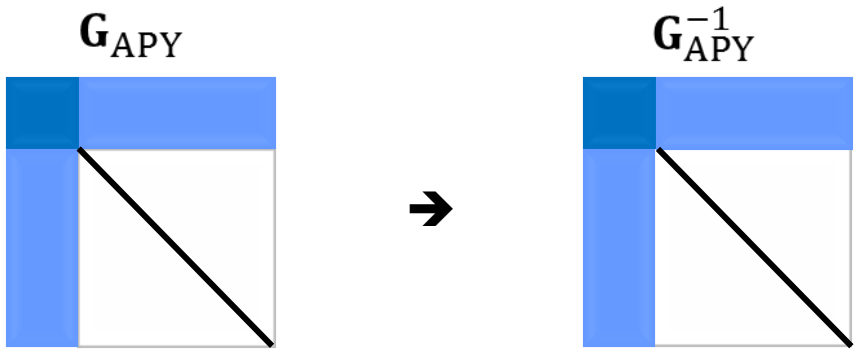
$\mathbf{G}_{\mathrm{APY}}^{-1}$



APY – algoritimo para *core* e *non-core*



- O($\boldsymbol{n}^3$) cost and $\boldsymbol{n}^2$ memory



- $\mathbf{G}_{\mathrm{APY}}^{-1}$ é esparsa
- Computação mais eficiente
- *core* = O($\boldsymbol{n}^3$) cost and $\boldsymbol{n}^2$ memory
- *Non-core* = O($\boldsymbol{n}$) cost and $\boldsymbol{n}$ memory

APY – Número de Animais *core*

- A dimensionalidade da **G** é limitada
  - Depende do tamanho da população
    - Número de SNP independents
    - Número de animais genotipados

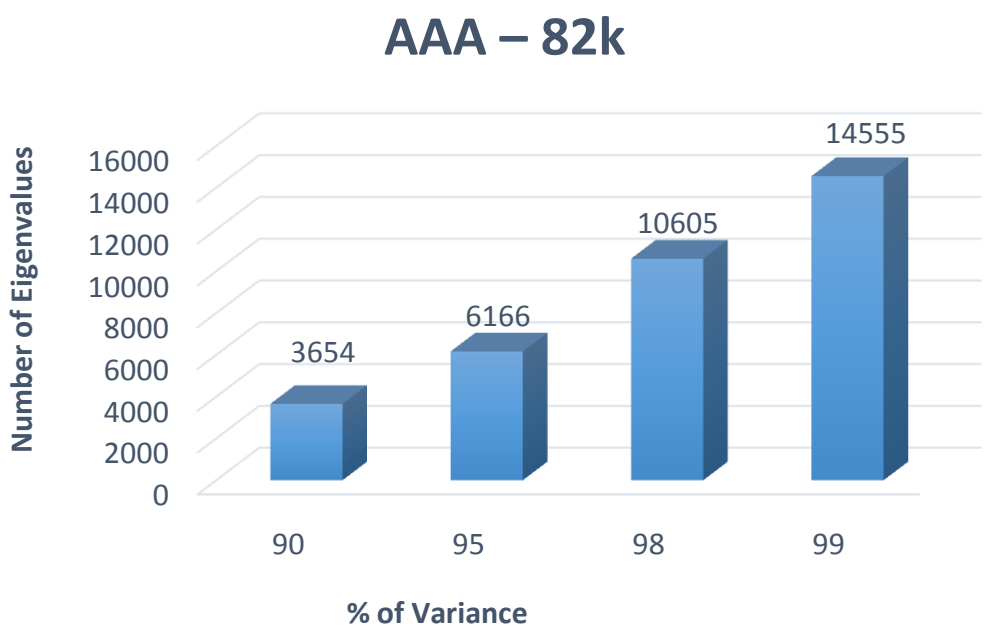
The Dimensionality of Genomic Information and Its Effect on Genomic Prediction

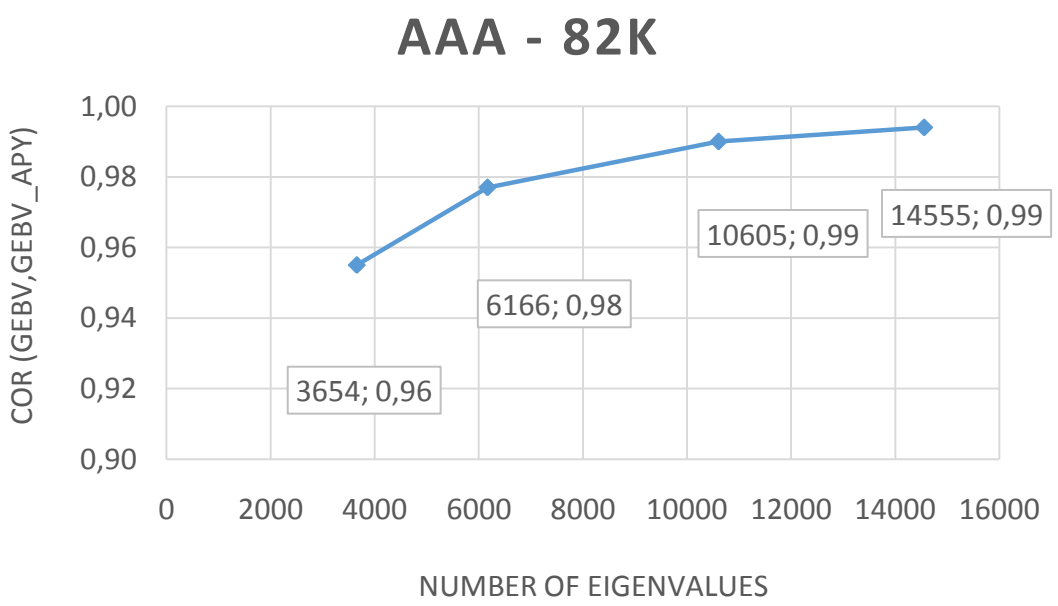
Ivan Pocrnic,*[,1] Daniela A. L. Lourenco,* Yutaka Masuda,* Andres Legarra,† and Ignacy Misztal*
*Department of Animal and Dairy Science, University of Georgia, Athens, Georgia 30602, and †Institut National de la Recherche Agronomique, Génétique, Physiologie et Systèmes d'Elevage, F-31326 Castanet-Tolosan, France

core = número de autovalores que explicam 98% da variância da **G**

APY – Número de Animais *core*

- SVD de **Z** vs. Decomposição de autovalor de **G**

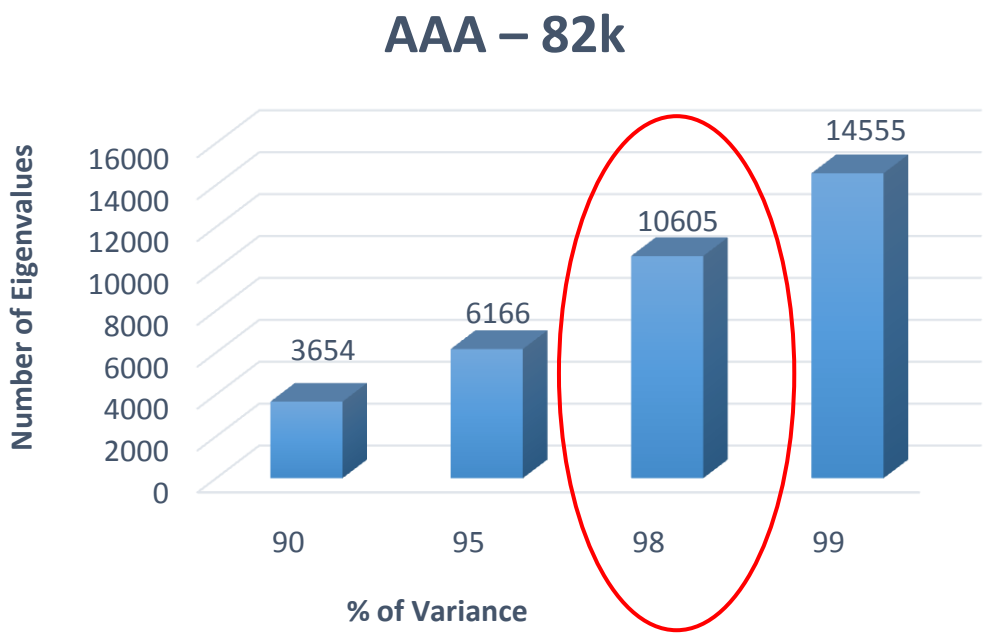




Lourenco et al. (2015b)

APY – Número de Animais *core*

- Tempo de computação e memória: $\mathbf{G}^{-1}$ vs. $\mathbf{G}_{\mathrm{APY}}^{-1}$
- 10.000 *core* e 72.000 *non-core*

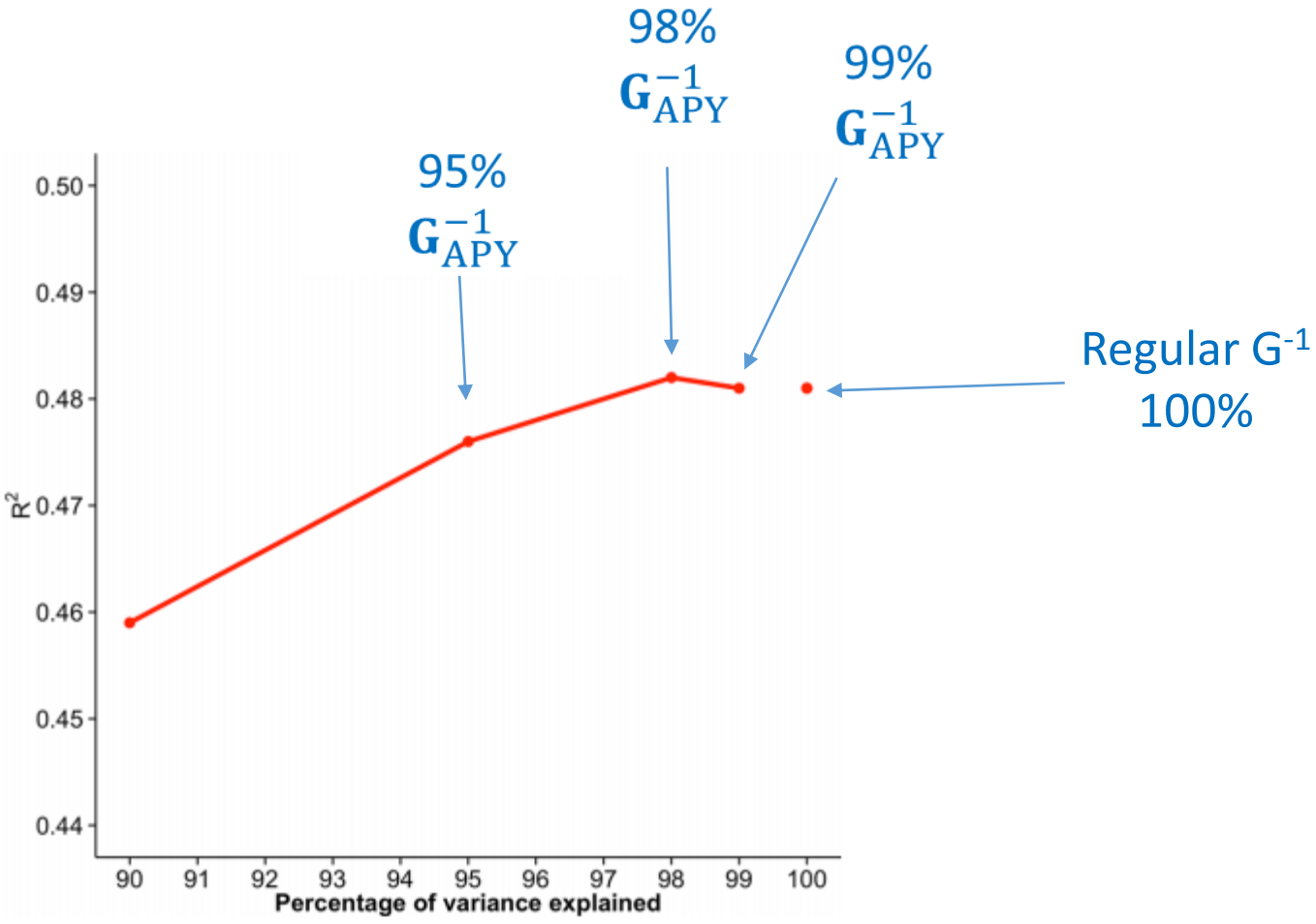


- Regular
  - 213’
  - 230Gb
- APY
  - 28’
  - 16Gb

Lourenco et al. (2015b)

CPU 2.90GHz 16 cores 256 Gb

APY – Por que usar 98%?



Pocrnic et al. (2016)

Com e sem informação genômica

PCG (bloco-diagonal) + Iteração nos dados + OpenMP

- American Angus
- Peso 1, 205, 365 dias
- Número de equações = 93.484.518

A^{-1}

5.5 horas

H^{-1}

core = 15k

non-core = 320k

34 horas

CPU 3.0GHz 24 cores 1 Tb

ssGBLUP APY

- American Angus

  - 400k animais genotipados
  - 19k *core*
  - 18 cacacterísticas
  - ~ 68Gb RAM
  - ~ 1 hour (G^{-1} APY)

Julho/2017

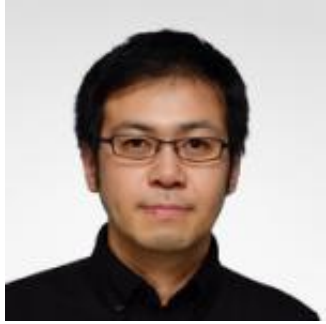
- US Holsteins

  - >250k animais genotipados
  - 7 características

Novembro/2016

O maior sistema de equações com ssGBLUP APY

- US Holsteins
  - 42M pedigree
  - 32M phenotypes
  - Leite, gordura, proteína
  - 760k animais genotipados
  - 18k *core*
  - ~ 74Gb RAM
  - Conv_crit = 10^{-15}



Masuda et al. (2017)

Tempo = 41 horas

6 cores

É possível usar 2M de animais genotipados?

$\mathbf{G}^{-1}$ vs. $\mathbf{G}_{\mathrm{APY}}^{-1}$

29 Tb

208 Gb

Our “dodo1”

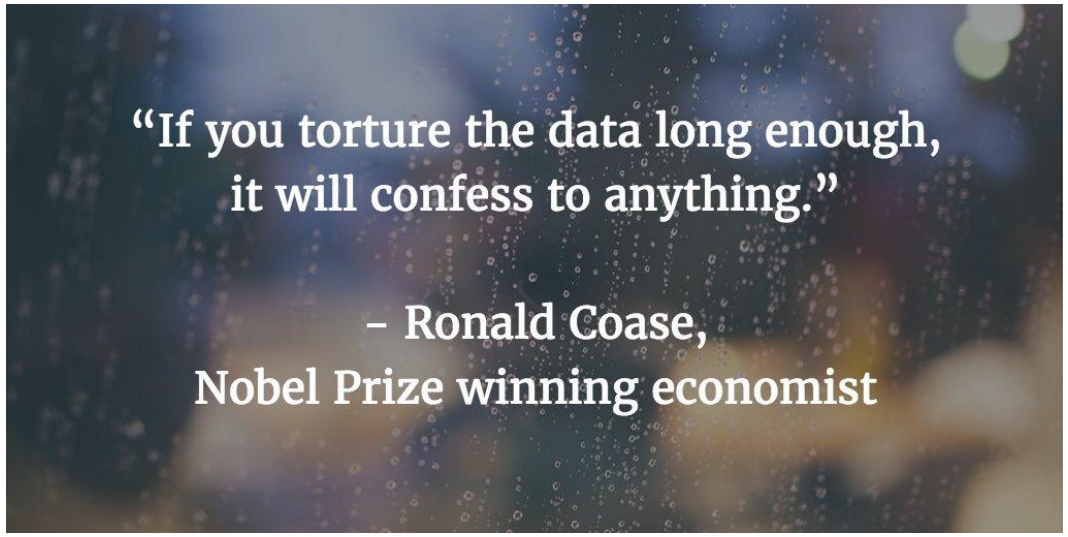
CPU 3.0GHz 24 cores 1 Tb

Próximo Desafio

Milhares de genes

Milhares de SNP (50k)

Dados de sequenciamento
3.000 bovinos



Milhões de SNP (32M)

EBV com maior accurácia?

Big data em Melhoramento Genético Animal

- *Big data* é um desafio desde ~ 1970 e está se tornando maior
- Métodos e Algoritimos estão acompanhando
  - Problemas resolvidos com sucesso
  - Melhoristas = codebreakers
- Poder computacional está aumentando

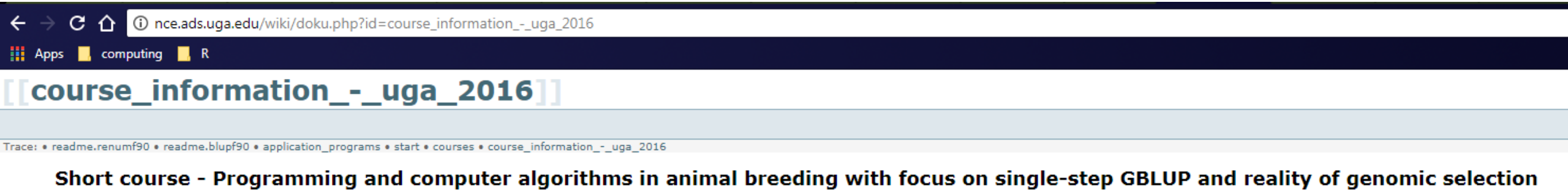
✓ Pessoas capacitadas para resolver problemas

✓ Conhecimento estatístico

✓ Habilidade de programação

Grupo de Melhoramento Genético Animal - UGA

http://nce.ads.uga.edu/



Short course - Programming and computer algorithms in animal breeding with focus on single-step GBLUP and reality of genomic selection

- 3 Semanas
- 7/5/2018 a 25/5/2018

Agradecimentos

USDA NIFA

ANGUS THE BUSINESS BREED
AMERICAN ANGUS ASSOCIATION
3201 Frederick Avenue • St. Joseph, MO 64506 • (816) 383-5100 • Fax (816) 233-9703 • E-mail: angus@angus.org

AGI

zoetis

PIC

Cobb

U.S. Registered Holsteins
FOR MAXIMUM PROFIT
HOLSTEIN ASSOCIATION USA, INC.

Smithfield

UNIVERSITY OF GEORGIA