Centre for Tropical Livestock Genetics and Health

Single Step Genome-wide association study (ssGWAS)

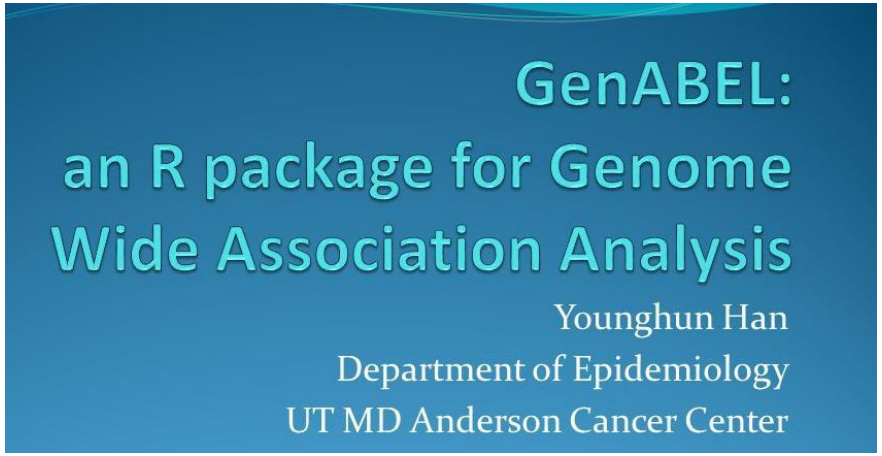
Getinet M. Tarekegn

February 2026

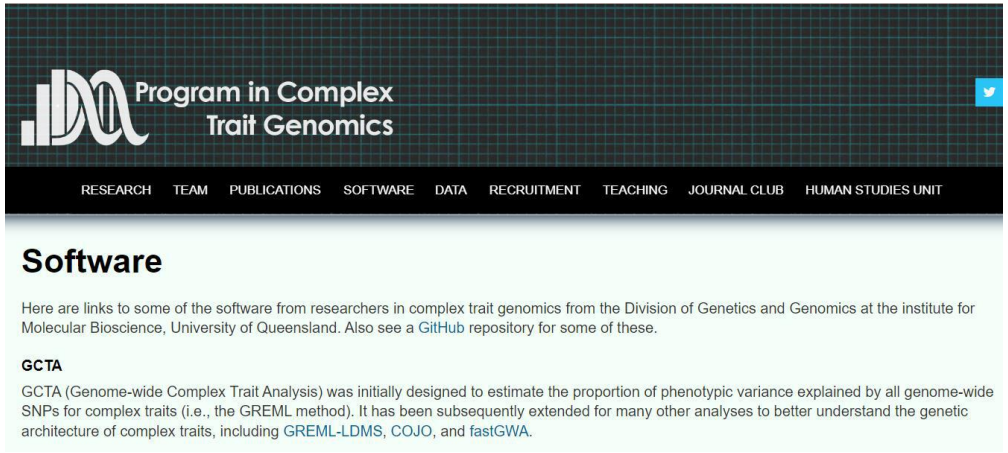
Options to implement GWAS

Centre for Tropical Livestock Genetics and Health

1)



2)



3)

genetics-statistics/
GEMMA

Genome-wide Efficient Mixed Model Association

4)

GWAS using the ssGBLUP

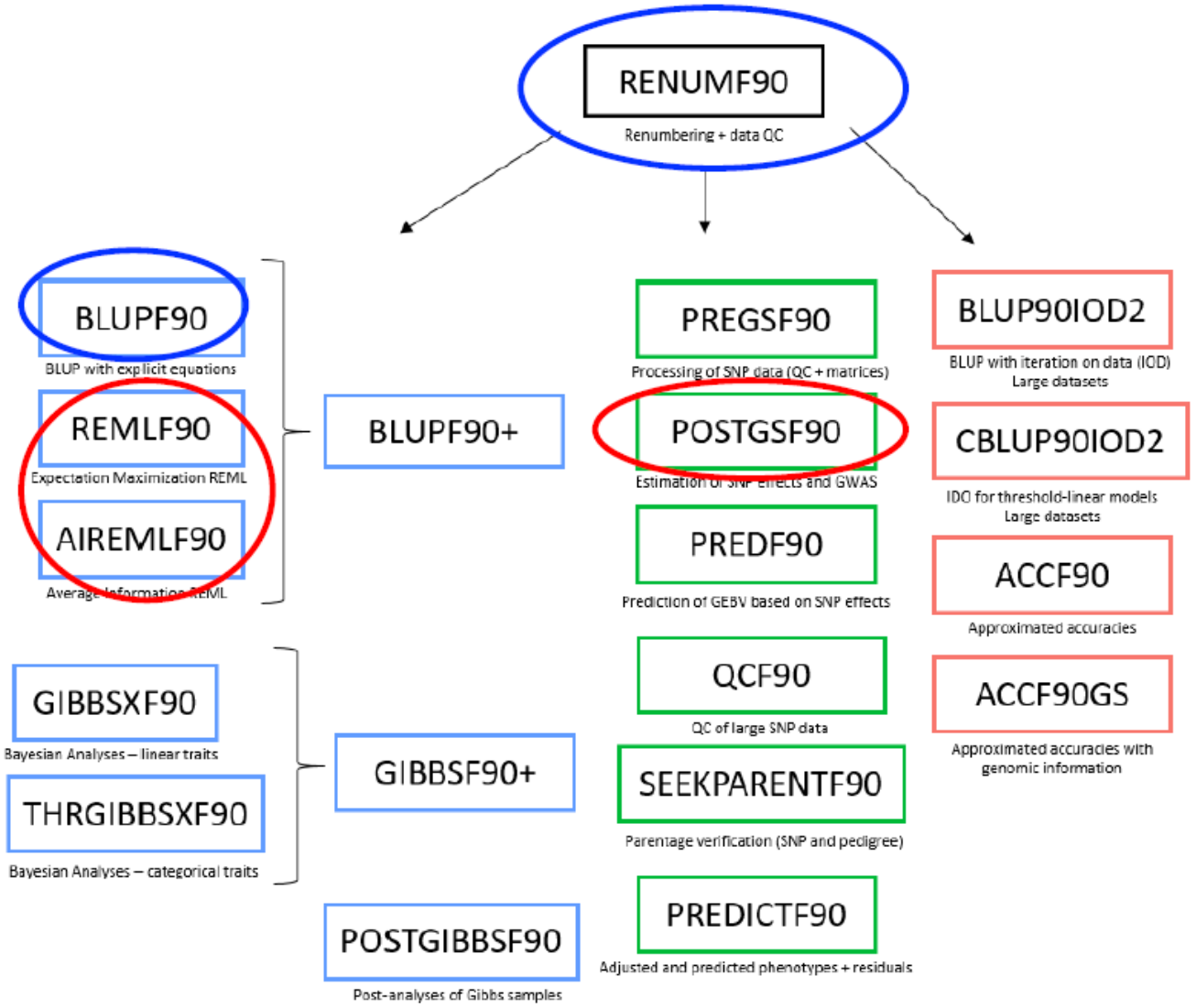
BLUPF90 programs

Centre for Tropical Livestock Genetics and Health

- It is a family of programs for mixed-model computations focusing on animal breeding applications.
- The programs can: **estimate variances** using several methods,
- Calculate BLUP for very large data sets,
- Use SNP information for improved accuracy of breeding values and genome-wide association studies (GWAS).
- The programs have been designed with 3 goals in mind:
  1. Flexibility to support a large set of models found in animal breeding applications.
  2. Simplicity of softwares to minimize errors and facilitate modifications.
  3. Efficiency at the algorithmic level.

- Aside from being used in hundreds of studies, the programs are utilized for commercial genetic evaluation in **dairy, beef, pigs, broiler chicken, and fish** major companies/institutions/associations in the US and beyond.
- The programs are written in Fortran90/95 and originated as exercises for a class taught by Ignacy Misztal at the University of Georgia.
- Over time, they have been upgraded and enhanced by many contributors. Details on programming and computing algorithms are available @ Misztal (1999).

BLUPF90 programs … cont'd



How ssGBLUP does?

- The inverse of H combining the additive and genomic relationship matrices, A and G, respectively, can be calculated as:

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

where:

A− 1 = inverse of relationship matrix based on pedigree information

G− 1 = inverse of genomic relationship matrix (VanRaden et al., 2008)

A− 1 22 = inverse of pedigree-based relationship matrix for genotyped animals

How ssGBLUP does?

Centre for Tropical Livestock Genetics and Health

- Assume **a** is a vector for breeding values for genotyped animals, and **u** is a vector for SNP marker effects, and one is related to another with the following equation.

  a=Zu (Wang et al., 2012)

  - **a** = Vector of **breeding values** (additive genetic effects).
  - **Z** = Incidence matrix linking animals to their **SNPs**.
  - **u** = Vector of SNP effects in genomic prediction.

The variance is: $\text{var}(\mathbf{a}) = \mathbf{G}\sigma_a^2$ and $\text{var}(\mathbf{Zu}) = \mathbf{ZDZ}'\sigma_u^2$; where **D** is a diagonal matrix of weights accounting for variances of SNP markers. This matrix is usually assumed to be **I** in the regular ssGBLUP.

a=Zu

For single animal: $a_i = \sum_{j=1}^{m} Z_{ij}u_j$

(Where m = number of markers.)

Matrix Form: $\begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} Z_{11} & Z_{12} & \cdots & Z_{1m} \\ Z_{21} & Z_{22} & \cdots & Z_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{n1} & Z_{n2} & \cdots & Z_{nm} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_m \end{pmatrix}$

Above 2 variances are identical, so we can derive

$$\mathbf{G} = \mathbf{ZDZ}'\frac{\sigma_u^2}{\sigma_a^2} = \mathbf{ZDZ}'\lambda$$ where **λ=σ²u/σ²a**.

According to the definition of **G**, the variance ratio can be: $\lambda = \frac{\sigma_u^2}{\sigma_a^2} = \frac{1}{2\sum_j p_j(1-p_j)}$.

CGIAR

- The prediction of breeding value $\hat{\mathbf{a}}$ is calculated with ssGBLUP
- The prediction of SNP effects $\hat{\mathbf{u}}$ is also calculated with the best prediction

$$\begin{aligned}\hat{\mathbf{u}} &= \mathrm{cov}(\mathbf{u}, \mathbf{a}')[\mathrm{var}(\mathbf{a})]^{-1}\hat{\mathbf{a}} \\ &= \lambda \mathbf{D}\mathbf{Z}'\mathbf{G}^{-1}\hat{\mathbf{a}} \\ &= \mathbf{D}\mathbf{Z}'(\mathbf{Z}\mathbf{D}\mathbf{Z}')^{-1}\hat{\mathbf{a}}\end{aligned}$$

- With the prediction of an SNP effect, we can **give weights to markers based on SNP solutions**.
- The current default weight is as in Wang et al. (2012): $w_i = 2p_i q_i \hat{u}_i^2$.

How the programs work?

Centre for Tropical Livestock Genetics and Health

RENUMF90

- RENUMF90 is a renumbering program to create input (data, pedigree, and parameter) files for BLUPF90 programs and provide basic statistics.
- RENUMF90-specific parameter file should be prepared as follows.
  - The file consists of pairs of **keyword (**is always capital**)** and the corresponding **value(s)**.
  - The following keywords are mandatory and must appear in the following order: **DATAFILE**, **TRAITS**, **FIELDS_PASSED TO OUTPUT**, **WEIGHT(S)**, **RESIDUAL_VARIANCE** and **EFFECT**. If you don't actually need **FIELDS_PASSED TO OUTPUT** and **WEIGHT(S)**, simply leave an empty line.
  - The remaining keywords are optional

```
base ~ (0.045s)
~/bin/renumf90 --show-template
# parameter file for renumf90
DATAFILE

TRAITS

FIELDS_PASSED TO OUTPUT

WEIGHT(S)

RESIDUAL_VARIANCE

EFFECT

#RANDOM
#
#OPTIONAL
#
#FILE
#
#FILE_POS
#
#SNP_FILE
#
```

Variance components estimation (ai/remlf90)

Centre for Tropical Livestock Genetics and Health

Options to estimate variance components:

REMLF90:

- **It** uses expectation maximization (EM) REML.
- It is the most reliable algorithm for most problems but can take hundreds of rounds of iterations.
- **REMLF90** was found to have problems converging with random regression models. In this case, using starting variances that are too large than too small usually helps.
- EM does **not calculate standard errors** for the estimates.

AIREMLF90:

- It uses Average Information (AI) REML.
- It usually converges much faster but sometimes doesn't converge.
- Very slow convergence usually indicates that **the model is over parameterized**, and **there is insufficient information to estimate some variances**.
- AI REML calculates **standard errors** for the estimates.

Variance components … cont'd

Centre for Tropical Livestock Genetics and Health

- **OPTION missing -99999**
- Specifies missing observations (default 0).
- This is only for data, not pedigree (always 0 for missing pedigrees). There is no missing covariable, so 0 is treated as a level.
- **OPTION SNP_file snp :-** Specifies the SNP file name **snp** to use genotype data.
- **OPTION use_yams:-** Run the program with YAMS (modified FSPAK). The computing time can be dramatically improved.

RENUMF90 Output files

Centre for Tropical Livestock Genetics and Health

RENUMF90 generates several files:

- *renf90.par*: parameter template file for BLUPF90 and other application programs
- *renf90.tables*: table relating the original code and the renumbered code
- *renf90.dat*: data file for BLUPF90
- *renaddxx.ped*: pedigree file for BLUPF90; *xx* is an integer number that indicates the position of animal effect among all model effects in renf90.par. This file will be created only if **RANDOM animal** is specified.
- *SNPfile_XrefID*: cross-reference file for genomic analysis, which contains renumbered ID and original ID; *SNPfile* is the original SNP marker file. This file will be created only if **SNP_FILE** is specified.
- *renf90.inb*: inbreeding coefficients. This file will be created only if **INBREEDING pedigree** is specified.
- *renf90.fields*: has detailed information about the data fields.

Output pedigree file

Centre for Tropical Livestock Genetics and Health

- The additive pedigree file built by RENUMF90 is renaddxx.ped.

Check your ped.outputFile

Structure of the pedigree file:

1. animal number (from 1)
2. parent 1 number or unknown parent group number for parent 1
3. parent 2 number or unknown parent group number for parent 2
4. 3 minus number of known parents (this column is replaced by inbreeding code if **INBREEDING** is specified or by default in RENUMF90 ≥ v1.157)
5. known or estimated year of birth (0 if not provided)
6. number of known parents (for genotyped animals, if any: 10 + number of known parents)
7. number of records
8. number of progenies as parent 1
9. number of progenies as parent 2
10. original animal id

Example

Input file - data

```
aa 1 10
aa 2 12
bb 1 11
cc 1 12
cc 2 14
dd 2 13
ee 2 14
```

Pedigree file - ped

```
aa ff ee 2004
bb hh gg 2004
cc hh ii 2004
dd ff 0 2004
ee ff 0 2002
ff 0 0 2002
gg ff 0 2002
```

Output pedigree file - renadd02.ped

```
Animal, sire, dam, inbreeding code (3-
#unknown parents if no-inbreeding), birth
year, #known parents, #records, #progeny of
sire, # progeny of dam, original animal ID
1 6 11 1333 2002 1 1 0 1 ee
2 8 7 2000 2004 2 1 0 0 bb
7 6 11 1333 2002 1 0 0 1 gg
3 6 12 1333 2004 1 1 0 0 dd
9 11 11 1000 2002 0 0 0 1 ii
4 6 1 2000 2004 2 2 0 0 aa
6 11 11 1000 2002 0 0 4 0 ff
```

BLUPF90

- **OPTION use_yams**
  - Runs the program with YAMS (modified FSPAK).
- **OPTION saveA22Inverse:** Saves **A**22−1 in "A22i".
- **OPTION saveGInverse:** Saves **G**-1 in "Gi"; to use a genomic relationship matrix **G**, the file needs to contain **G**-1
- **OPTION saveA22: -** Saves **A**22 in "A22".
- **OPTION saveG**:- saves G in "G"
- **OPTION SNP_file snp:-** Specifies the SNP file name snp to use genotype data.
- **OPTION snp_p_value**
  - Computes the elements of the inverse of the Mixed Model Equations that are needed for exact GWAS with p-values using postGSf90.
  - This requires quite a lot of memory and time.

```
freqdata.count
sum2pq
A22
G
A22i
Gi
Check_Diagonal_GimA22i
solutions
xx_ija
```

POSTGSF90

Centre for Tropical Livestock Genetics and Health

Basic options

- The program calculates SNP effects using the ssGBLUP framework (**Wang et al., 2012**).
- It needs **OPTION map_file** to assign SNP to their location for Manhattan plots, so chromosomes are visualized in different colors.

The following options for **POSTGSF90** (ssGWAS) are available:

- **OPTION Manhattan_plot:-** Plots the Manhattan plot (SNP effects) for each trait and correlated effects using **GNUPLOT**.
- **OPTION windows_variance n:** Calculates the variance explained by n adjacent SNPs.
- **Hint:** When this option is used, the sum of variance explained by n adjacent SNPs (column 8 of snp_sol or column 3 of chrsnpvar) is not 100%. This is because moving variance is used. If windows size is 20, the proportion of variance assigned to SNP 1 is calculated from SNP 1 to 20, for SNP 2 it goes from 2 to 21, for SNP 3 it goes from 3 to 22, and so forth. A file called windows_variance has variance that sums to 100% in column 9.

POSTGSF90 … cont'd

Centre for Tropical Livestock Genetics and Health

- **OPTION readGInverse <file> :-** Reads $\mathbf{G}-1$ from "Gi" by default, or from a user-supplied **file**. See the caution below.
- **OPTION readA22Inverse <file> :-** Reads $\mathbf{A}22-1$ from "A22i" by default, or from a user-supplied **file**. See the caution below.

Caution:

- With the options **readGInverse** and **readA22Inverse**, the program applies τ to the loaded $\mathbf{G}-1$ and ω to the loaded $\mathbf{A}22-1$ regardless of whether the matrices have been already scaled with τ or ω. In other words, the loaded matrix could be scaled twice if the user used τ or ω both in saving and reading the matrix. Be careful to use the scaling factors combined with the input/output options.
- **Hint:** OPTION TauOmega was needed when inbreeding was not considered for $\mathbf{A}-1$. Because inbreeding is now considered for $\mathbf{A}-1$, we recommend not using this option anymore.

POSTGSF90 … cont'd

- **OPTION windows_variance_type n:** Sets windows type for variances calculations
  1. moving windows
  2. exclusive windows
- **OPTION which_weight x :** Generates a weight variable to construct a weighted genomic relationship matrix **G=ZDZ′**
- **OPTION snp_p_value:-** Computes p-values for GWAS from elements of the inverse of the Mixed Model Equations previously obtained from blupf90. This requires quite a lot of memory and time. See **Aguilar et al. (2019)** for more details.
- **OPTION snp_var:-** Creates a file with prediction error covariance (PEC) for SNP to be used in **PREDF90** to compute reliability for indirect predictions. This option works when OPTION snp_p_value is used in BLUPF90+.

Output files for POSTGSF90

Centre for Tropical Livestock Genetics and Health

- "snp_sol" contains solutions of SNP and weights

1. trait
2. effect
3. SNP
4. Chromosome
5. Position
6. SNP solution
7. weight (can be used as the weight to calculate the weighted **G** matrix)
8. variance explained by n adjacent SNP (if **OPTION windows_variance** is used)
9. variance of the SNP solution (used to compute the p-value if **OPTION snp_p_value** is used)

- "chrsnp" contains data to create the plot by GNUPLOT

1. trait
2. effect
3. values of SNP effects to use in Manhattan plots, i.e., (abs(SNP_i)/var(SNP))
4. SNP
5. Chromosome
6. Position

- "chrsnpvar" contains data to create plot by GNUPLOT

1. trait
2. effect
3. variance explained by n adjacent SNP
4. SNP
5. Chromosome
6. Position

Output files for POSTGSF90 …

Centre for Tropical Livestock Genetics and Health

"windows_segment" contains information of windows segments used to get variance explained

1. label
2. window size (number of SNP)
3. Start SNP number for the window
4. End SNP number for the window
5. identification of window: (ChrNumber)'_'(startPositionMBP)
6. Start (ChrNumber)'_'(Position) for the window
7. End (ChrNumber)'_'(Position) for the window

"windows_variance" contains variance explained for the biggest non-overlapping windows segments

1. trait
2. effect
3. Start SNP number or SNP name for the window
4. End SNP number or SNP name for the window
5. window size (number of SNP)
6. Start (ChrNumber)'_'(Position) for the window
7. End (ChrNumber)'_'(Position) for the window
8. identification of window: (ChrNumber)'_'(startPositionMBP)
9. variance explained by n adjacents SNP

"snp_pred" contains allele frequencies + SNP effects

Output files for POSTGSF90 …

Centre for Tropical Livestock Genetics and Health

Graphic control files:

- Several files are created to generate graphics using either GNUPLOT or R.
- File names rules
- "Sft1e2.R". The first letter indicates "S" for solutions of SNP, "V" for variance explained, and "P" for p-values.
- "t1e2" indicates that the file is for the trait 1 and the effect 2. 45

Filename extension

- xxx.gnuplot => GNUPLOT
- xxx.R => R programs
- xxx.pdf => image
- xxx.png => image
- xxx.tif => image

Output files for POSTGSF90 …

Centre for Tropical Livestock Genetics and Health

Graphic control files:

- Several files are created to generate graphics using either GNUPLOT or R.
- File names rules
- "Sft1e2.R". The first letter indicates "S" for solutions of SNP, "V" for variance explained, and "P" for p-values.
- "t1e2" indicates that the file is for the trait 1 and the effect 2. 45

Filename extension

- xxx.gnuplot => GNUPLOT
- xxx.R => R programs
- xxx.pdf => image
- xxx.png => image
- xxx.tif => image

Step to prepare parameter file for ssGWAS

Centre for Tropical Livestock Genetics and Health

- Prepare the parameter file for renumf90; you may name it ***renum.par***
- Run renumf90 as: ./renumf90 *renum.par*
- Run either ai/remlf90 as it is without changing anything on its parameterFile, i.e. on *renf90.par*
- Once the variance components are obtained here, copy the *renf90.par* to *blupf90.par* and use the latter parameterFile for BLUPF90.
- Open *blupf90.par* and change the following:
  - Replace the variance component estimates by the new estimates generated while running ai/remlf90
  - OPTION saveGInverse
  - OPTION saveA22Inverse
  - OPTION map_file gwasmap.txt
  - OPTION use_yams
  - OPTION snp_p_value
  - OPTION saveA22
  - OPTION saveG
- Then, copy *blup.par* to *postGSF90.par*
- Change/add the ff options:
  - OPTION readGInverse
  - OPTION readA22Inverse
  - OPTION use_yams
  - OPTION snp_p_value
  - OPTION weightedG wt
  - OPTION windows_variance 20
  - OPTION Manhattan_plot
  - OPTION map_file gwasmap.txt

CTLGH Funders

Bill & Melinda Gates foundation

UKaid from the British people

UKRI Biotechnology and Biological Sciences Research Council