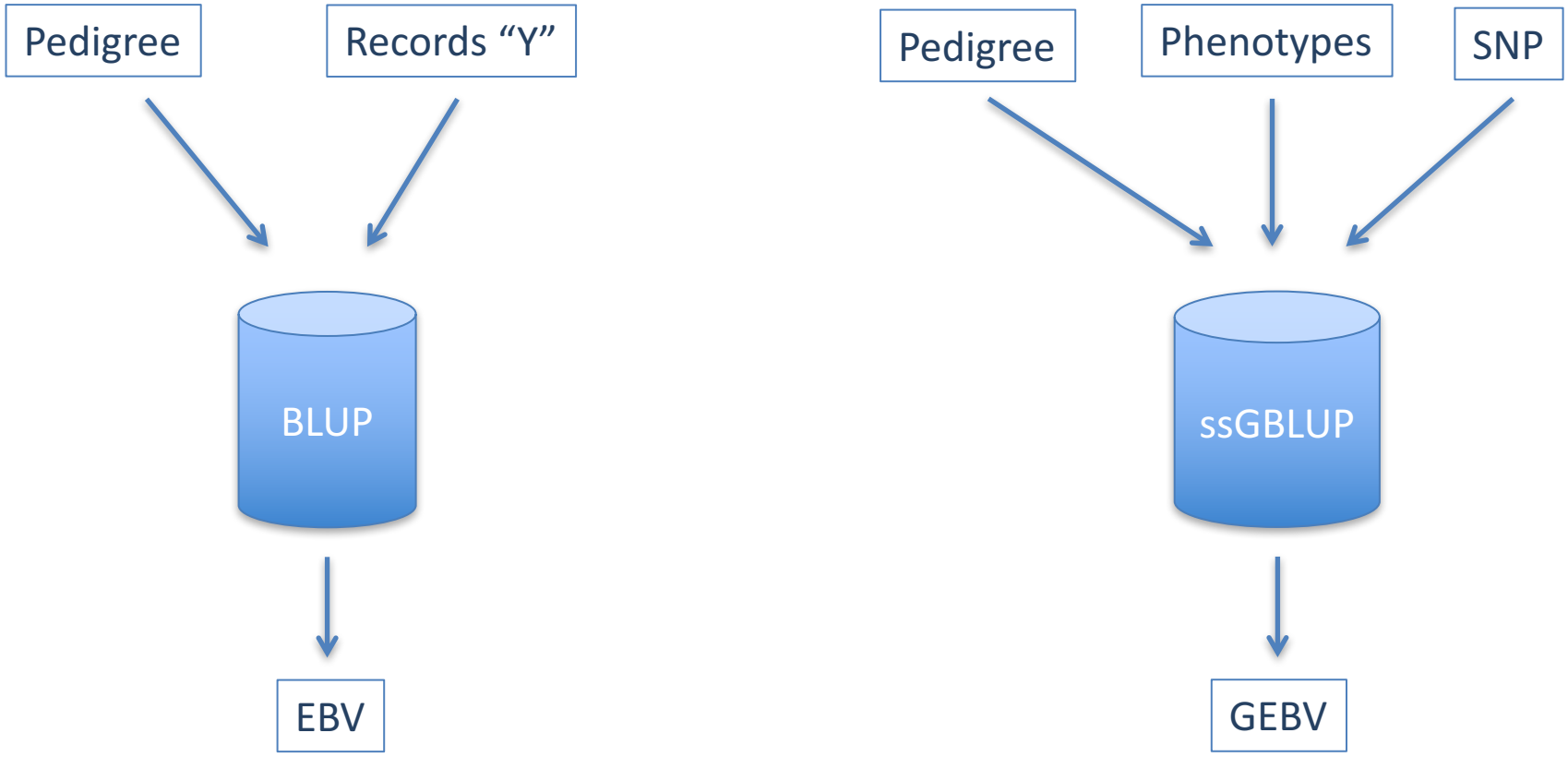
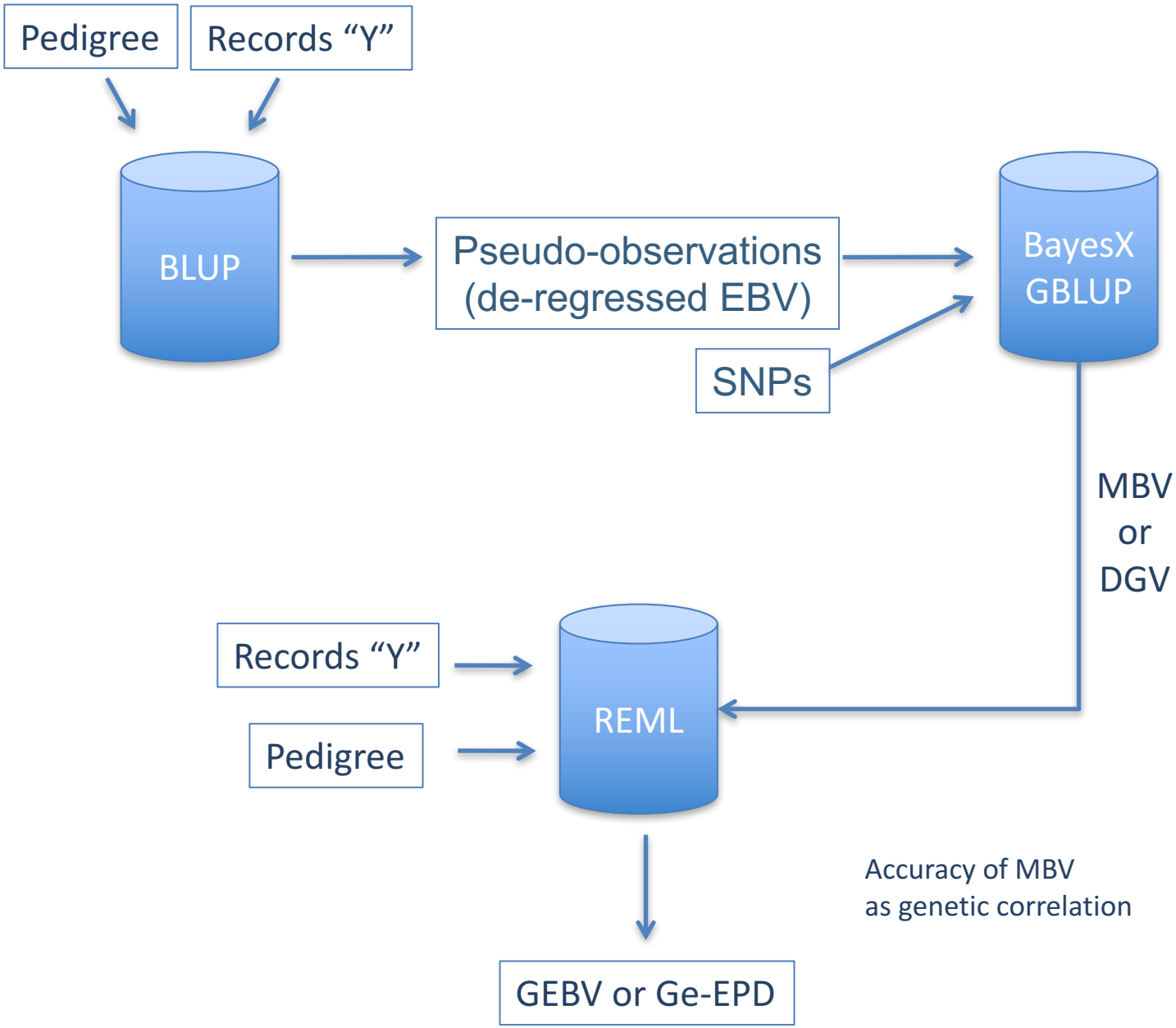
Creation of genomic relationship matrices with preGSf90 and Forming Single-step mixed model equation

BLUP vs. ssGBLUP



multistep vs. ssGBLUP



Extra matrices required for single-step

- Inverses

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

PREGSF90

  - Pedigree relationships between genotyped animals
  - Genomic relationships

Parameters file

RENUMF90
renum.par

```
DATAFILE
phenotypes.txt
TRAITS
3
FIELDS_PASSED TO OUTPUT

WEIGHT(S)

RESIDUAL_VARIANCE
 0.9038
EFFECT
1 cross alpha  # mu
EFFECT
2 cross alpha  # animal
RANDOM
animal
FILE
pedigree
SNP_FILE
marker.geno.clean
(CO)VARIANCES
    0.9951E-01
```

BLUPF90
renf90.par

```
DATAFILE
 renf90.dat
NUMBER_OF_TRAITS
           1
NUMBER_OF_EFFECTS
           2
OBSERVATION(S)
    1
WEIGHT(S)

EFFECTS: POSITIONS_IN_DATAFILE NUMBE
  2          1 cross
  3      15800 cross
RANDOM_RESIDUAL VALUES
  0.90380
 RANDOM_GROUP
     2
 RANDOM_TYPE
 add_animal
 FILE
renadd02.ped
(CO)VARIANCES
  0.99510E-01
OPTION SNP_file marker.geno.clean
```

BLUPF90 programs using Genomics

- Genomic programs
  - controled by adding OPTIONS commands to the parameter file
  - `OPTION SNP_file` *`marker.geno.clean`*
  - Read 2 files:
    - `marker.geno.clean`
    - `marker.geno.clean_XrefID`

SNP file & Cross Reference Id

SNP File First col: Identification, could be alphanumeric
Second col: SNP markers {codes: 0,1,2 and 5 for missing}

```
80   2110101100201201101101011011111
8014 2111010151110112022111011151111
516  2110010120225202112021012110211
181  2111011111220112055020002010101
```

Renumber ID

Cross Reference ID

```
1732 80
8474 8014
406 516
9441 181
```

Pedigree File (from RENUMF90)

```
1732 11010 10584 1 3 12  1 0 0 80
8474  8691  9908 1 3 12  1 0 0 8014
 406  8691  9825 1 3 12  1 0 2 516
9441  8691  8829 1 3 12  1 0 0 181
```

Original ID

Genomic Relationship Matrix - G

- G = ZZ'/k
  - Z = centered matrix for SNP marker
  - Dimension Z= n*p
  - n animals,
  - p markers

Genotype Codes
0 – Homozygous
1 – Heterozygous
2 – Homozygous

5 – No Call (Missing)

Data file with SNP marker

```
80    211010110020120110110101101111112111112101 00
8014  211101015111011202211101115111121011122101 00
516   211001012022520211202101211021112022121111 01
181   211101111122011205020002010102221221111110 0
```

HOWTO: Creating Genomic Matrix

- Read SNP marker information => M

$$\begin{bmatrix} 2 & 1 & 2 & .. \\ 0 & 1 & 0 & .. \\ .. & .. & .. & .. \end{bmatrix}$$

- Get ‘means’ to center
  - Calculate allele frequency from observed genotypes (p_i)
  - p_i= sum($SNPcode_i$)/2n

- Centered matrix Z = M-2P

$$\begin{bmatrix} p_1 & p_2 & p_3 & p_n \\ p_1 & p_2 & p_3 & p_n \\ p_1 & p_2 & p_3 & p_n \\ & & \ldots & \end{bmatrix}$$

Genomic Matrix default options

- G* = ZZ'/k as in VanRaden, 2008
- With:
  - Z center using current allele frequencies
  - k = 2 sum (p * (1-p))

- G = G*0.95 + A_{22}*0.05 (to invert)

- Tunning of G (see Vitezica et al., 2011)
  - Adjust G to have mean of diagonals and off-diagonals equal to A_{22}

Genomic Matrix Options

- OPTION whichfreq x
  - 0: read from file *freqdata* or other specified
  - 1: 0.5
  - 2: current calculated from genotypes (default)
- OPTION FreqFile *file*
  - Reads allele frequencies from a file
- OPTION maxsnps x
  - Set the maximum length of string for reading marker data from file => BovineHD chip

Options for Blending G and A

- OPTION AlphaBeta alpha beta
  - $G = alpha*G^r + beta*A_{22}$
- OPTION tunedG
  - 0: no adjustment
  - 1: mean(diag(G))=1, mean(offdiag(G))=0
  - 2: mean(diag(G))=mean(diag(A)), mean(offdiag(G))=mean(offdiag(A)) (default)
  - 3: mean(G)=mean(A)
  - 4: Use Fst adjustment Powell et al. (2010) & Vitezica et al. (2011)

Creating a ‘raw’ genomic matrix ‘GBLUP’

- Tricks:
- Use dummy pedigree

  1 0 0
  2 0 0
  …
- Change blending parameters
  - OPTION AlphaBeta 0.99 0.00
  - OPTION GamaDelta 0.01 0.00
- No adjustment for compatibility with A
  - OPTION tunedG 0

$$G = 0.99*G + 0.01*I$$

Creating a ‘raw’ genomic matrix ‘GBLUP’

- Change blending parameters
  - OPTION AlphaBeta 0.99 0.00
  - OPTION GamaDelta 0.01 0.00

$$G = \mathit{Alpha}\ G^{r} + \mathit{Beta}\ A22 + \mathit{Gamma}\ I + \mathit{Delta}$$

Storing and Reading Matrices

- Matrices that can be stored:
  - A22, inv(A22), G, inv(G), GmA22, inv(GmA22), inv(H)

- All matrices are stored in same format:
  - upper triangular
  - By default in binary format
  - But to store in text (Ascii) format:
    - Use: `OPTION saveAscii`

- Values
  - i j val
  - i & j refers to the row number in the genotype file !!!!!
  - Renumber ID could be obtained from the XrefID file

Storing and Reading Matrices

To save our ‘raw’ genomic matrix:

- OPTION saveG [all]
  - If the optional *all* is present all intermediate G matrices will be saved!!!

or it inverse

- OPTION saveGInverse
  - Only the final matrix G, after blending, scaling, etc. is inverted !!!

- Look in wiki for keywords for other matrices

Storing with Original IDs

- Some matrices could be stored in text files with the original IDs extracted from *renaddxx.ped* created by the RENUMF90 program (col #10)

- For example:
  - OPTION saveGOrig
  - OPTION saveDiagGOrig
  - OPTION saveHinvOrig

- Values
  - origID_i, origID_j, val

OUTPUT

- Only GimA22i , other requested matrices, and some reports are stored
- Main log is printout to the screen!!!
- Use the command `tee` to save in a log file
- This will allow to save and see the messages from the program

- `echo renf90.par | preGSf90 | tee pregs.log`

Printout: Same heading as other programs

```
Options read from parameter file:

 * SNP file: marker.geno.clean
 * SNP Xref file: marker.geno.clean_XrefID
 * Matrix in Ascii format(default=binary)

 *-------------------------------------------------------------*
 *                  Genomic Library: Version 1.110             *
 *                                                             *
 *  Modified relationship matrix (H) created for effect:   2   *
 *-------------------------------------------------------------*

 Read 18600 animals from pedigree file: "renadd02.ped"
 Number of Genotyped Animals: 1500
```

All options entered in the parameter file should be here!
IF not check if keywords are correct (upper and lower case)

Check number of animals and individuals with genotypes

Printout

```
Creating A22
   Extracting subset of: 4634 pedigrees from: 18600 elapsed time:     0.0019
   Calculating A22 Matrix by Colleau ...elapsed time   1.250464

Reading SNP file
   Column position in file for the first marker: 7
   Format to read SNP file: (6x,400000i1)
   Number of SNPs: 3000
   Number of Genotyped animals: 1500
   Reading SNP file elapsed time: .41

Statistics of alleles frequencies in the
   N:            3000
   Mean:        0.500
   Min:         0.101
   Max:         0.898
   Var:         0.016
```

Information from genotype file. The format is detected from the first line!!!

All genotypes should start in the same column!!!

Number of SNP is also determined by the first line!!!

Looking at stored matrices

- Avoid open with text editors, huge files!!!
- For example:
- 1500 genotyped individuals => 1,125,750 rows
- Inspection could be done by Unix commands:
  - `head G` => first 10 lines
  - `tail G` => last 10 lines
  - `less G` => scroll document by line/page
  - `wc -l G` => count number of lines
    good for checks with the number of genotypes (n) = (n*(n+1)/2)

head G

```
1 1 .999382118619
1 2 .355052761478
2 2 1.014521277458
1 3 -.048184197960
2 3 -.057513012886
3 3 .976558921904
1 4 -.101734083083
2 4 -.007644724611
3 4 .196757165096
4 4 1.018165021903
```

PreGSf90 inside BLUPF90?

- Almost all programs from package support creation of genomic relationship matrices, Hinv, etc.
- OPTION SNP_file xxxx

- Why preGSF90 ?
  - Same genomic relationship matrix for several models, traits, etc. Just do it once and store.
  - Uses optimized subroutines for efficient matrix multiplications, inversion and with support for parallel processing

Creating a subset of relationship matrix (A_{22})

- Create a relationship matrix for only genotyped animals (~ thousands)
- Full pedigree (~millions)
- Trace only ancestors of genotyped
- Colleau's algorithm to create A_{22}

Tabular method vs. Colleau algorithm

- Testing
  - 6,500 genotyped Holsteins
  - 57,000 pedigrees

	Tabular*	Colleau method
CPU Time	311 s	45 s
Memory	12.1GB	322MB

* Gmatrix.f90 (VanRaden, 2009)