Université de Toulouse

Data simulation (including genomics) QMSim software

Zulma G. Vitezica

zulma.vitezica@inra.fr

GenPhySE

Génétique, Physiologie et Systèmes d'Elevage

INRA

SCIENCE & IMPACT

QMSim: why to use it ?

✓It was design to simulate large-scale genotyping data in multiple and complex livestock pedigrees

✓A wide variety of genome architectures from infinitesimal model to single-locus model

✓It is a user-friendly tool for simulating data

✓Computationally efficient in terms of both time and memory

QMSim†: where to find it ?

†Sargolzaei & Schenkel (2009), Bioinformatics 25:680-681.

The code is written in C++ language

Executable files are freely available for Windows and Linux and Mac at:

http://www.aps.uoguelph.ca/~msargol/qmsim/

How the simulation is carried out ?

In 2 steps:

✓*First step:* A **historical population** is simulated

–in order to create initial LD and

–to establish mutation-drift equilibrium

–expansion and contraction of the population

✓*Second step:* One or multiple **recent population structures** are generated

Parameter file

- ✓It must be in ASCII format
- ✓It consists of **five** main sections
- ✓The order of commands within each section is not important
- ✓All commands end with a semicolon
- ✓No semicolon → error message and program exits.

```
/*********************************
 **      Global parameters      **
 *********************************/
title = "Example 1 - 10k SNP panel"
...;

/*********************************
 **    Historical population    **
 *********************************/
begin_hp;
       ....;
end_hp;

/*********************************
 **         Populations         **
 *********************************/
begin_pop = "p1";
       ....;
end_pop;

/*********************************
 **           Genome            **
 *********************************/
begin_genome;
      ....;
end_genome;

/*********************************
 **        Output options       **
 *********************************/
begin_output;
       ....;
end_output;
```

1. Global parameters section

```
/*************************************
 **      Global parameters          **
 *************************************/
title = "Example 1 - 10k SNP panel";
seed = "seed.txt";
```

An arbitrary title

The random number generator (RNG*) requires a seed file.

If it is not specified → RNG will be seeded from the system clock

For each run the initial seed numbers will be backed up in output folder

Seed + Number of threads for parallel processing → This allows to generate the same simulated data !

Parameter file: ex01.prm

Output folder: r_ex01/

Output

```
.-----------------------------------.
| Example 1 - 10k SNP panel         |
`-----------------------------------'
```

Initial seed is backed up in [r_ex01/seed].
parameter file is backed up in [r_ex01/ex01.prm].

* Mersenne Twister algorithm (Matsumoto & Nishimura, 1998)

1. Global parameters section

```
/*********************************
 **      Global parameters      **
 *********************************/
title = "Example 1 - 10k SNP panel";
nrep  = 1;        //Number of replicates
h2    = 0.2;      //Heritability
qtlh2 = 0.2;      //QTL heritability
phvar = 1.0;      //Phenotypic variance
```

Range: 0 - 10,000

QTL effect is simulated

Overall heritability (Polygenic + QTL)

```
title = "Example 8
nrep  = 1;
h2    = 0.2;
qtlh2 = 0.0;
phvar = 1.0;
```

Only polygenic effect is simulated

```
title = "Example 11
nrep  = 1;
h2    = 0.2;
qtlh2 = 0.05;
phvar = 1.0;
```

Both, polygenic and QTL effects are simulated

1. Global parameters section

```
/*********************************
 **      Global parameters      **
 *********************************/
title = "Example 1 - 10k SNP panel";
nrep  = 1;        //Number of replicates
h2    = 0.2;      //Heritability
qtlh2 = 0.2;      //QTL heritability
phvar = 1.0;      //Phenotypic variance
no_male_rec;      // No record for males
```

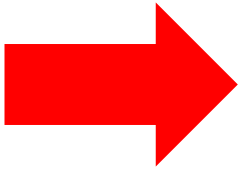
A sex limited trait like milk yield

When males do not have records, but selection or culling are based on

- **Phenotypes** → Males will be randomly selected or culled
- **EBVs**

Parameter file

- ✓ It consists of **five** main sections



```
/*********************************
 **      Global parameters      **
 *********************************/
title = "Example 1 - 10k SNP panel"
...;

/*********************************
 **    Historical population    **
 *********************************/
begin_hp;
       ....;
end_hp;

/*********************************
 **         Populations         **
 *********************************/
begin_pop = "p1";
       ....;
end_pop;

/*********************************
 **           Genome            **
 *********************************/
begin_genome;
      ....;
end_genome;

/*********************************
 **        Output options       **
 *********************************/
begin_output;
       ....;
end_output;
```

2. Historical population section

```
/**********************************
 **   Historical population    **
 **********************************
begin_hp;
   hg_size = 420 [0]              //
             420 [200];
   nmlhg   = 20;                  //
end_hp;
```

- To create initial LD
- Evolutionary forces: mutation and drift (no selection, no migration)
- Random mating: union of gametes randomly sampled from the male and female gametic pools
- Discrete generations
- Only a single historical population

2. Historical population section

Historical generation sizes

```
/*********************************
 **    Historical population    **
 *********************************
begin_hp;
   hg_size = 420 [0]
             420 [200];
   nmlhg   = 20;                  //
end_hp;
```

Constant size of 420

hg_size = **v1** [**v2**]

v1 the historical generation **size**
Range: 2 – 100,000

v2 the historical generation **number**
Range: 0 – 150,000

2. Historical population section

Historical **bottleneck** or **expansion** can be simulated

```
/*********************************
 **   Historical population   **
 *********************************/
begin_hp;
   hg_size = 2000 [0]
             200 [1000];
   nmlhg   = 40;
end_hp;
```

Gradual decrease in size from 2000 to 200

```
/*********************************
 **   Historical population   **
 *********************************/
begin_hp;
   hg_size = 100 [0]
             100 [950]
            3000 [1000];
   nmlhg   = 200;
end_hp;
```

Expansion in the last historical generation from 100 to 3000

2. Historical population section

```
/*********************************
 **   Historical population   **
 *********************************/
begin_hp;
   hg_size = 2000 [0]
             200 [1000];
   nmfhg   = 40;
end_hp;
```

Number of males

Default : equal number of males and females

nm**f**hg → **first** historical generation

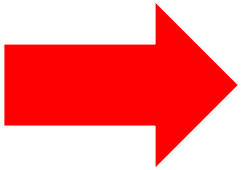
Sex ratio will be constant across historical generations. It can be changed in the last generation

nm**l**hg → **last** historical generation

```
/*********************************
 **   Historical population   **
 *********************************/
begin_hp;
   hg_size = 2000 [0]
             200 [1000];
   nmlhg   = 40;
end_hp;
```

Parameter file

- ✓ It consists of **five** main sections



```
/*********************************
 **      Global parameters      **
 *********************************/
title = "Example 1 - 10k SNP panel"
...;

/*********************************
 **    Historical population    **
 *********************************/
begin_hp;
        ....;
end_hp;

/*********************************
 **         Populations         **
 *********************************/
begin_pop = "p1";
        ....;
end_pop;

/*********************************
 **           Genome            **
 *********************************/
begin_genome;
       ....;
end_genome;

/*********************************
 **        Output options       **
 *********************************/
begin_output;
        ....;
end_output;
```

3. Population section

```
/*********************************
 **          Populations          **
 *********************************/
begin_pop = "p1";
        ....;
end_pop;

begin_pop = "p2";
        ....;
end_pop;
```

One or multiple recent populations

For the ***first defined recent population*** (i.e. p1), founders must come from **the last historical generation**

For ***subsequent populations*** (i.e. p2), founders can be chosen from one or more (up to 10) **previously defined populations** (i.e. p1)

Multiple recent populations can be analyzed (*joint_pop in Hist pop section*)
separately (one pedigree for each population) or
jointly (by creating one pedigree for all populations) for inbreeding and EBV

3. Population section

Choosing founders for a population

```
/*********************************
 **          Populations          **
 *********************************/
begin_pop = "line1";
   begin_founder;
      male   [n = 20,  pop = "hp", select = tbv /h];
      female [n = 400, pop = "hp", select = tbv /h];
   end founder;
```

Parameters for the founders

Number of male/female to be selected

It indicates from which population the base animals must be selected

Type of selection

select: rnd (default), phen, tbv and ebv
/l : to select low values
/h : to select high values

hp: historical population (last historical generation)

Choosing founders for a population for F1 design

```
/*********************************
 **          Populations          **
 *********************************/
begin_pop = "line1";
   begin_founder;
      male   [n = 20,  pop = "hp", select = tbv /h];
      female [n = 400, pop = "hp", select = tbv /h];
   end_founder;
   ng  = 20;          //Number of generations
end_pop;

begin_pop = "line2";
   begin_founder;
      male   [n = 20,  pop = "hp", select = tbv /l];
      female [n = 400, pop = "hp", select = tbv /l];
   end_founder;
   ng  = 20;          //Number of generations
end_pop;

//Cross between line1 and line 2 to generate F2
begin_pop = "cross";
   begin_founder;
      male   [n = 20, pop = "line1", gen = 20];
      female [n = 400, pop = "line2", gen = 20];
   end_founder;
   ng  = 2;           //Number of generations
```

Crossing between populations/lines is allowed

3. Population section

Matting design

```
/*********************************
 **       Populations       **
 *********************************/
begin_pop = "p1";
  begin_founder;
    male   [n = 4500, pop = "hp"];
    female [n = 48000, pop = "hp"];
  end_founder;
  ls  = 1;              //Litter size
  pmp = 0.5;            //Proportion of male progeny
  ng  = 10;             //Number of generations
  md  = minf;           //Mating design - control of inbreeding
  sr  = 0.4;            //Replacement ratio for sires
  dr  = 0.2;            //Replacement ratio for dams
  sd  = ebv /h;         //Selection design
  cd  = ebv /l;         //Culling design
  ebv_est = blup;
```

rnd : default
p_assort : similarity
minf : inbreeding is minimized in the next generation

Assortative mating base on **phen**, **ebv** or **tbv**

```
ng   = 10;              //Number of generations
md   = p assort/ebv;    //Mating design
```

3. Population section

Replacement

```
/*********************************
 **          Populations          **
 *********************************/
begin pop = "p1";
  _founder;
  le    [n =   50, pop = "hp"];
  male [n = 2500, pop = "hp"];
  rounder;
    = 1 2 [0.2];        //Litter size
  p = 0.5;              //Proportion of male progeny
  ng = 10;              //Number of generations
  md = p_assort/ebv;    //Mating design
  sr = 0.4;             //Replacement ratio for sires
  dr = 0.2;             //Replacement ratio for dams
  sd = ebv /h;          //Selection design
  cd = phen/l;          //Culling design
  ebv_est = blup;
```

Overlapping generations

sr : 1, discrete generations (default)

sr : 40% of sires will be replaced in all generations

sr : 0.4 [1] 0.5 [5]
40% of sires will be culled for generation 1 to 5, and 50% from generation 5 to last generation

3. Population section

Selection and culling designs

```
/*************************************
 **          Populations            **
 *************************************/
begin_pop = "p1";
   begin_founder;
      male   [n =                    hp"];
      female [n = 2                  hp"];
   end_founder;
   ls  = 1 2 [0.2];                  //Litter size
   pmp = 0.5;                        //Proportion of male progeny
   ng  = 10;                         //Number of generations
   md  = p_assort/ bv;               //Mating design
   sr  = 0.4;                        //Replacement ratio for sires
   dr  = 0.2;                        //Replacement ratio for dams
   sd  = ebv /h;                     //Selection design
   cd  = phen/l;                     //Culling design
   ebv_est = blup;
```

rnd, phen, tbv ebv and **age** (only for culling)

/l or **/h** to select low or high values

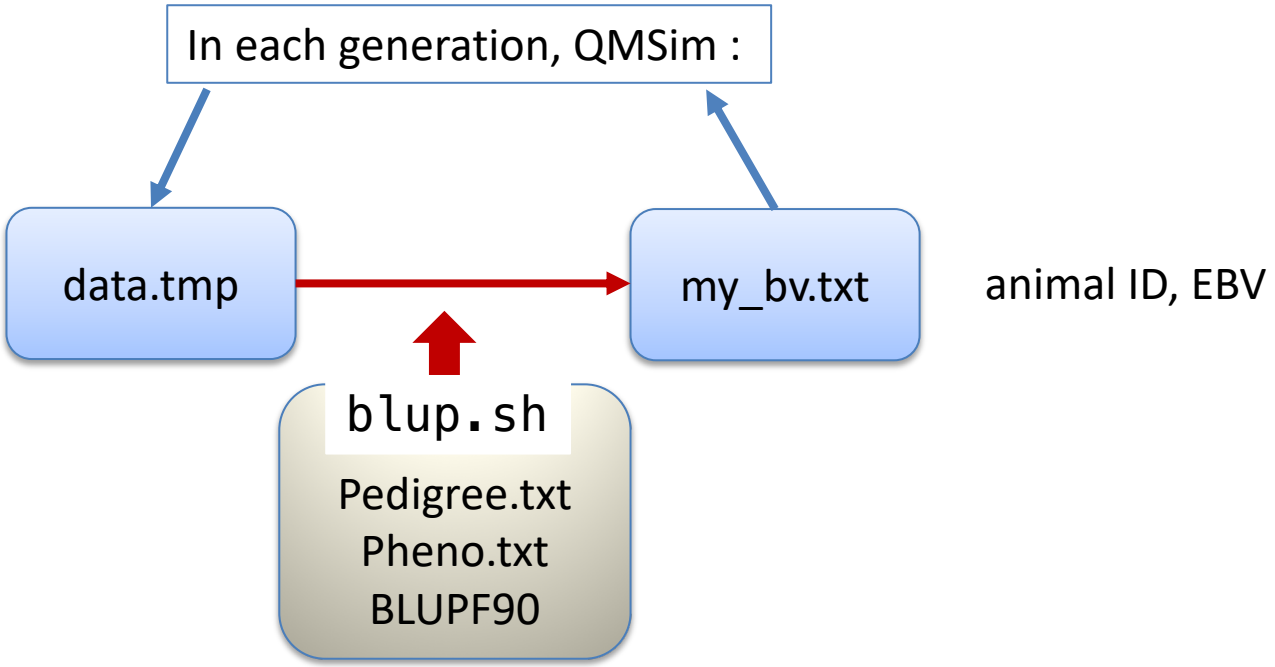
Breeding value estimation method

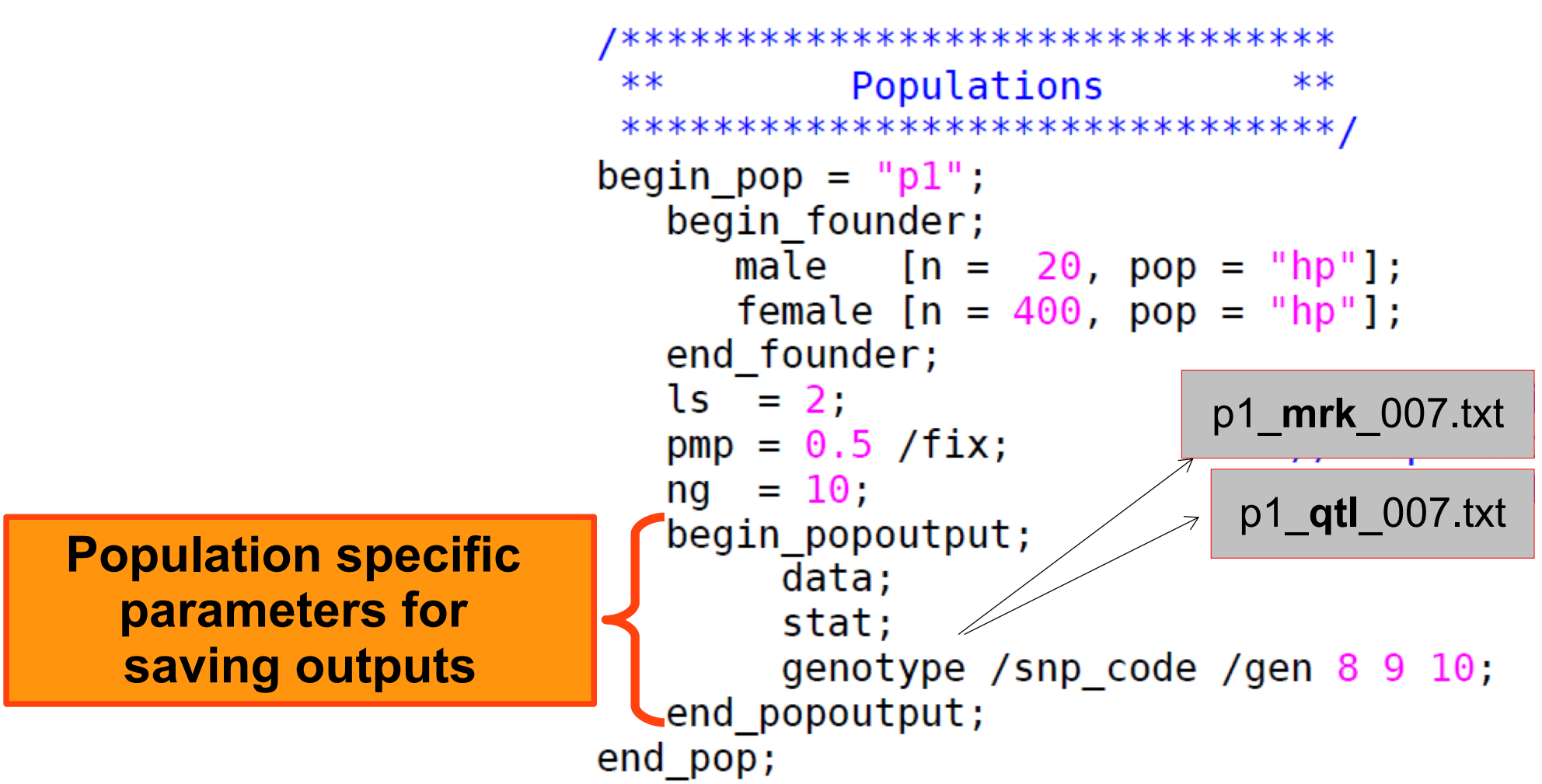
3. Population section

Selection and culling designs

```
pmp = 0.5;              //Proportion of male progeny
ng  = 10;               //Number of generations
md  = p_assort/ebv;     //Mating design
sr  = 0.4;              //Replacement ratio for sires
dr  = 0.2;              //Replacement ratio for dams
sd  = ebv /h;           //Selection design
cd  = phen/l;           //Culling design
ebv_est
```

```
ebv_est = external_bv "./blup.sh";
```





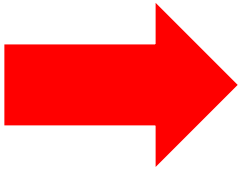
data: save individual's data except their genopype
(*File name:* 'population name'_**data**_'replicate number'.txt

stat: save brief statistic on simulated data

genotype: save genotype data

Parameter file

- ✓ It consists of **five** main sections



```
/*********************************
 **      Global parameters      **
 *********************************/
title = "Example 1 - 10k SNP panel"
...;

/*********************************
 **    Historical population    **
 *********************************/
begin_hp;
        ....;
end_hp;

/*********************************
 **         Populations         **
 *********************************/
begin_pop = "p1";
        ....;
end_pop;

/*********************************
 **           Genome            **
 *********************************/
begin_genome;
       ....;
end_genome;

/*********************************
 **        Output options       **
 *********************************/
begin_output;
        ....;
end_output;
```

4. Genome section

Marker information

Example – 30k SNP panel

```
/*********************************
 **        Genome          **
 *********************************/
begin_genome;
  begin_chr = 30;
    chrlen = 100;          //Chromo
    nmloci = 3000;          //Number of markers
    mpos   = rnd;          //Marker positions
    nma    = all 2;        //Number of marker alleles
    maf    = eql;          //Marker allele frequencies
    nqloci = 50;           //Number of QTL was 50
    qpos   = rnd;          //QTL positions
    nqa    = all 2;        //Number of QTL alleles
    qaf    = eql;          //QTL allele frequencies
    qae    = rndg 0.4;     //QTL allele effects
  end_chr;
  mmutr    = 2.5e-5 /recurrent; //Marker mutation
rate
  qmutr    = 2.5e-5;          //QTL mutation rate
  r_mpos_g;                 //
positions across genome
  r_qpos_g;                 // F
across genome
end_genome;
```

Number of chromosomes: 30
chrlen : range 1-5,000 cM

Position is sampled from uniform distribution

All marker loci will have 2 alleles

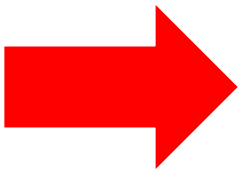
In the first historical generation, then drift and mutation

It's sampled from gamma distribution with shape 0.4

Other possibilities :
- Missing marker/QTL genotypes
- Genotyping errors can be simulated (marker/QTL)

Parameter file

- ✓ It consists of **five** main sections



```
/*********************************
 **      Global parameters      **
 *********************************/
title = "Example 1 - 10k SNP panel"
...;

/*********************************
 **    Historical population    **
 *********************************/
begin_hp;
        ....;
end_hp;

/*********************************
 **         Populations         **
 *********************************/
begin_pop = "p1";
        ....;
end_pop;

/*********************************
 **           Genome            **
 *********************************/
begin_genome;
       ....;
end_genome;

/*********************************
 **        Output options       **
 *********************************/
begin_output;
        ....;
end_output;
```

5. Output section

```
/*********************************
 **        Output options       **
 *********************************/
begin_output;
   linkage_map;
   hp_stat;
end_output;
```

Marker and QTL linkage map (GWAS)

Save brief statistics on historical population

```
/*********************************
 **        Output options       **
 *********************************/
begin_output;
   linkage_map;
   allele_effect;
end_output;
```

Save allele effects

```
 .-----------.
|  Example 1  |
 `-----------'
ID      Chr   Allele:Effect ...
-----------------------------------------
Q1        1     1: 0.066403   2:-0.001068
Q2        1     1:-0.050405   2: 0.031267
Q3        1     1:-0.006917   2: 0.009631
Q4        1     1:-0.000543   2: 0.000171
```

QMSim outputs

```
/*********************************
 **          Populations          **
 *********************************/
begin_pop = "p1";
   begin_founder;
      male   [n =  20, pop = "hp"];
      female [n = 400, pop = "hp"];
   end_founder;
   ls  = 2;
   pmp = 0.5 /fix;
   ng  = 10;
   begin_popoutput;
         data;
         stat;
         genotype /snp_code /gen 8 9 10;
   end_popoutput;
end_pop;
```

p1_**data**_001.txt

Progeny	Sire	Dam	Sex	G	NMPrg	NFPrg	F	Homo	Phen	Res	Polygene	QTL	Final_EBV
28795	15853	20301	F	10	0	0	0.035156	0.741889	+2.384481	+0.551262	+0.000000	+1.833219	+1.614136
28796	15853	10632	F	10	0	0	0.005859	0.699111	+2.150498	+0.802941	+0.000000	+1.347557	+1.558238
28797	15853	10844	F	10	0	0	0.010742	0.709444	+0.639003	-1.011156	+0.000000	+1.650158	+1.177224
28798	15853	21272	F	10	0	0	0.005249	0.782111	+2.628546	+0.842557	+0.000000	+1.785989	+1.681527
28799	15853	13409	M	10	0	0	0.006348	0.737889	---	+0.000000	+0.000000	+1.055616	+1.348642
28800	15853	13208	M	10	0	0	0.001953	0.715889	---	+0.000000	+0.000000	+0.424422	+1.368958

p1_**stat**_001.txt

```
begin_popoutput;
    data;
    stat;
    genotype /gen 8 9 10;
end_popoutput;
end_pop;
```

```
 .-----------.
|  Example 1  |
 `-----------'
------------------- Inbreeding -------------------
                      Inbred           All
Gen.         No.      Mean     SD      Mean     SD
0              0    0.0000 0.0000    0.0000 0.0000
1              0    0.0000 0.0000    0.0000 0.0000

------------------ Homozygosity ------------------
Gen.                    Mean                  SD
0                 0.68254159          0.01207245
1                 0.68200626          0.01103250

------------------- Phenotype --------------------
Gen.                    Mean                  SD
0                 0.08440969          1.01093563
1                 0.04504056          1.02152016

---------------------- QTL -----------------------
Gen.                    Mean                  SD
0                 0.04889285          0.56092140
1                -0.00533798          0.55392545
```

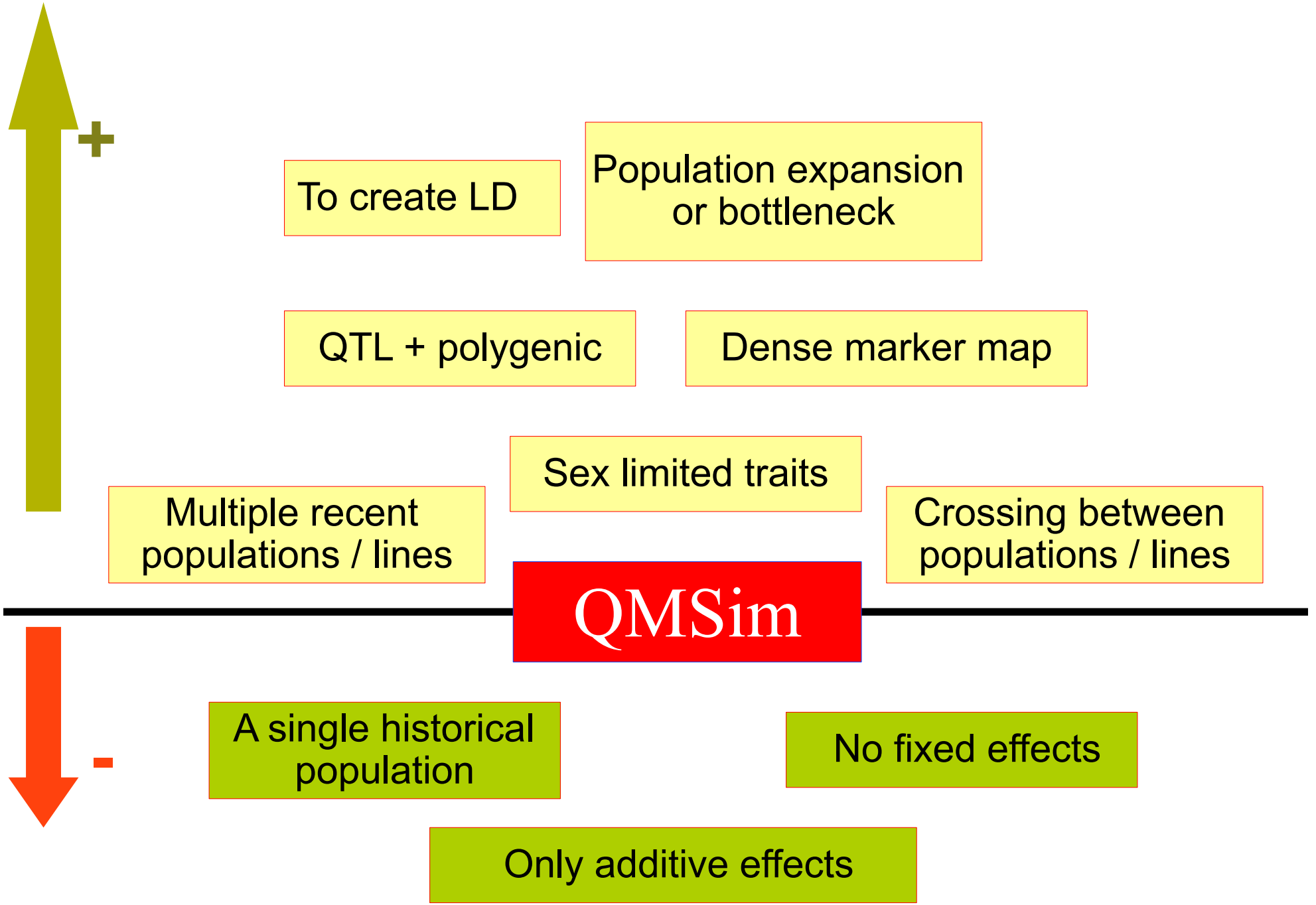
```
------------------------------------------ Brief structure ------------------------------------------
Gen.      Progeny    Male%     Male Selected   Female Selected     Sire   Culled      Dam   Culled
0             420 0.047619       20        0      400        0        0        0        0        0
1             400 0.500000      200        8      200       80       20        8      400       80
2             400 0.500000      200        8      200       80       20        8      400       80
3             400 0.500000      200        8      200       80       20        8      400       80
4             400 0.500000      200        8      200       80       20        8      400       80
5             400 0.500000      200        0      200        0       20        0      400        0
Overall      2420 0.421488     1020       32     1400      320      100       32     2000      320
```

p1_**mrk**_001.txt

```
begin_popoutput;
    data;
    stat;
    genotype /snp_code /gen 4 5;
end_popoutput;
end_pop;
```

```
 .-----------.
|  Example 1  |
 `-----------'
Progeny Sire    Dam     Genotypes (0 = a1,a1; 2 = a2,a2; 3 = a1,a2; 4 = a2,a1; 5 = missing; The first
allele is paternal and the second allele is maternal) ...
1621    1408    365     0 2 2 2 2 2 2 0 2 2 2 2 4 4 3 0 3 3 3 0 0 0 2 0 3 2 2 4 0 2 0 0 4 3 2 3 3 3 3
0 2 2 0 0 3 4 3 4 4 0 2 0 4 0 0 3 2 2 4 2 0 4 2 3 2 3 0 0 4 0 4 2 2 3 0 4 2 0 3 0 0 4 3 3 0 2 0 2 2 0
2 2 2 0 0 0 0 0 2 0 0 2 0 0 2 0 4 2 2 3 4 0 4 2 2 0 0 0 0 4 4 2 2 3 4 0 0 2 3 2 4 0 3 4 0 2 2 4 0 0 2
0 2 0 3 0 0 3 2 2 0 0 0 4 0 4 2 2 2 2 0 3 2 2 2 0 4 2 2 2 0 2 2 0 2 0 0 2 2 4 3 2 0 2 0 0 2 0 2 2 0 2
2 2 2 2 1623    1408    1096    3 4 4 4 2 2 4 3 4 2 4 4 2 2 0 0 0 0 3 0 0 0 2 0 3 2 2 4 0 2 0 0 4 3 2
3 3 3 3 0 2 2 0 0 3 4 3 4 4 0 2 0 4 0 0 3 2 2 4 2 0 4 2 3 2 3 0 0 4 0 4 2 2 3 0 4 2 0 3 0 0 4 0 3 3 4
0 4 2 0 4 4 2 3 3 3 3 0 2 0 0 2 0 0 2 0 2 2 2 0 2 0 4 2 2 0 0 0 0 2 2 2 2 0 2 0 0 2 0 4 4 0 3 4 0 2 2
4 0 3 2 3 2 0 0 0 0 0 2 4 3 0 3 2 3 4 4 4 2 2 3 0 2 2 2 0 2 2 2 2 0 2 2 0 2 0 0 2 2 2 0 2 0 2 0 0 2 0
4 2 3 4 2 2 2 2 0 0 2 2 0 0 0 2 2 2 0 2 0 3 4 0 4 4 2 0 2 2 2 2 0 0 2 0 2 2 2 3 4 2 3 4 2 4 4 4 3 4 2
3 2 0 0 2 0 0 0 0
```

```
 .-----------.
|  Example 1  |
 `-----------'
Progeny Sire    Dam     Genotypes (paternal allele, maternal allele) ...
1621    1139    285     2 2 1 1 2 1 1 1 2 2 1 2 1 1 2 1 1 1 1 2 1 2 1 1 1 1 2 1 1 2 2 1 1 1 1 1 1 2
2 1 1 1 1 2 2 2 2 2 2 2 2 2 1 1 2 2 2 1 1 1 1 2 1 2 1 1 2 1 1 1 2 1 1 2 1 1 2 1 2 1 1 2 1 2 1 2 1 1 2
2 2 2 1 2 1 2 2 1 2 1 1 2 2 2 2 2 2 2 1 1 1 2 1 1 2 2 2 1 2 1 2 1 1 2 2 2 1 1 2 1 2 2 2 2 2 2 2 1 1 1
1 2 2 2 2 2 1 1 1 1 1 2 1 2 2 2 1 2 1 2 2 1 1 1 2 1 2 2 1 1 2 1 1 2 2 1 2 2 1 2 1 1 2 2 2 1 1 1 1 1 2
2 2 2 1 2 1 1 1 2 1 1 1 1 2 2 2 2 2 1 2 1 1 1 1 2 1 2 2 2 2 1 1 1 1 2 1 2 2 2 1 2 2 2 1 2 1 2 1 2 2 2
1 2 1 1 1 2 1 2 2 1 2 2 2 1 1 1 1 1 1 1 1 1 2 1 1 2 1 2 2 2 2 1 1 1 1 2 1 2 1 1 1 1 1 2 1 2 1 1 2 1 2
1 1 2 1 1 1 1 1 1 1 1 1 1 2 1 2 1 2 1 2 1 2 1 1 2 1 1 1 1 1 2 2 2 2 2 2 2 2 1 2 2 1 1 2 2 1 1 1 2 1 1
2 2 1 2 2 1 2 1 2 1 2 1 2 2 1 1 2 2 2 1 1 2 2 1 1 2 2 2 2 2 2 2 2 2 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1
2 2 2 2 2 2 2 2
```


+
To create LD
Population expansion or bottleneck
QTL + polygenic
Dense marker map
Sex limited traits
Multiple recent populations / lines
Crossing between populations / lines
QMSim
-
A single historical population
No fixed effects
Only additive effects

Some advices

Simulation MUST be a mirror of real life, as much as possible

Heritability according to the trait

Number of markers is fixed in first historical generation, then drift and mutation

- Check the number of informative markers in recent population
- Final number of segregating markers should be ~ 50K

Number of QTL on each chromosome > 200

- Our traits are complex, most of them polygenic

Some advices

Number of chromosomes !!!

Wallaby: 10 chromosomes

Chickens: macro and micro-chromosomes (chromosomes should be defined separately with different sizes)

Most livestock animal populations have overlapping generations

- Thank you for your attention!