Short course on

Methods and Tool for Genomic Predictions and GWAS in Breeding Programs

Mehdi Sargolzaei
Select Sires Inc.
University of Guelph

Daniela Lourenco
University of Georgia

20 -24 February 2023, University of New England

1

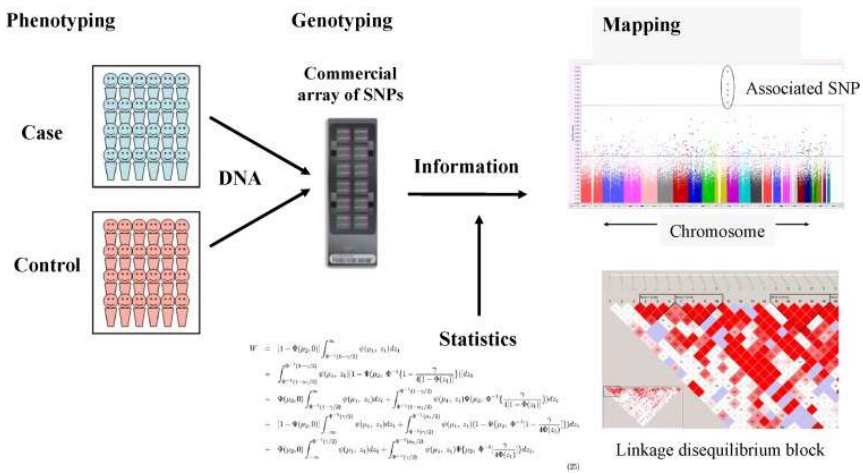
Genomics Revolution – Human Genome Project

- Precision/Personalized medicine was the reason behind the 13 years effort and $2.7 billion spent on the Human Genome Project (Completed in April 2003)
- A human genome can now be mapped in just a few hours for less than $1,000
- A complete human genome contains three billion base pairs of DNA
- Genomes of all humans are extremely similar. There are minor DNA sequence variations in each individual (between 1% and 3%) that makes them unique.
- The human genome includes approximately 20,000 different genes that encode proteins.

2

2

Genomics Revolution – Gold Rush (GWAS)



Genomics Inform. 2012 Dec; 10(4): 220–225

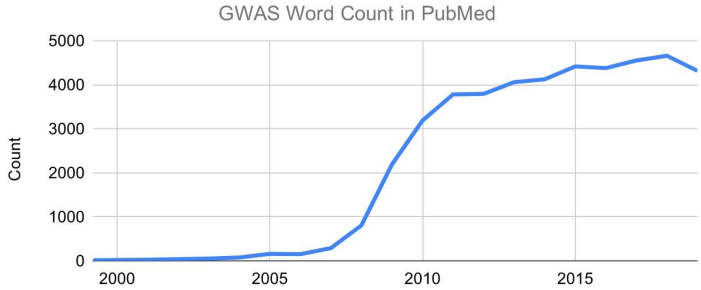
3

3

Genomics Revolution – Gold Rush (GWAS)

First successful GWAS was carried out in 2002 (Nat Genet. 2002;32:650–654)

The next publication was 3 years later in 2005

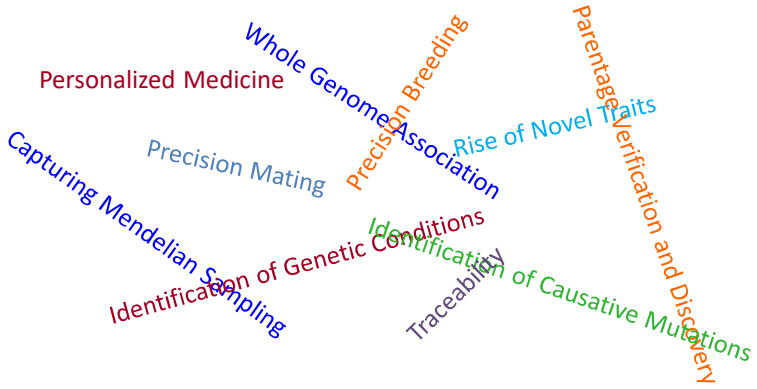


Source: https://bioinformaticshome.com/tools/gwas/gwas.html)

4

4

Genomics Revolution



What's changed?

- Our ability to sequence the genome
- Our ability to analyze the large genomic data

5

5

Introduction to Genomic

Human:

- Understand, diagnose, monitor, treat, predict and prevent diseases
- Development of medicine (e.g., Vaccines)

Livestock:

- Genetic selection to increase production and profitability
- Efficiency
- Resiliency
- Sustainability
- Food security
- Welfare

6

6

Introduction to Genomic

By 2020, the impact of genetics on medicine will be even more widespread. The pharmacogenomics approach for predicting drug responsiveness will be standard practice for quite a number of disorders and drugs. New gene-based "designer drugs" will be introduced to the market for diabetes mellitus, hypertension, mental illness, and many other conditions. Improved diagnosis and treatment of cancer will likely be the most advanced of the clinical consequences of genetics, since a vast amount of molecular information already has been collected about the genetic basis of malignancy. By 2020, it is likely that every tumor will have a precise molecular fingerprint determined, cataloging the genes that have gone awry, and therapy will be individually targeted to that fingerprint.

Implications of the Human Genome Project for Medical Science
Francis S. Collins, MD, PhD; Victor A. McKusick, MD
JAMA. 2001;285(5):540-544. doi:10.1001/jama.285.5.540

7

7

Introduction to Genomic

Source: Hildebrand Factory chocolate, Germany

8

8

Introduction to Genomic

1st UK child to receive gene therapy for fatal genetic disorder is now 'happy and healthy'

By Nicoletta Lanese published 3 days ago

A baby with a rare inherited disorder became the first child in the U.K. to receive a new gene therapy for the condition.

Comments (0)

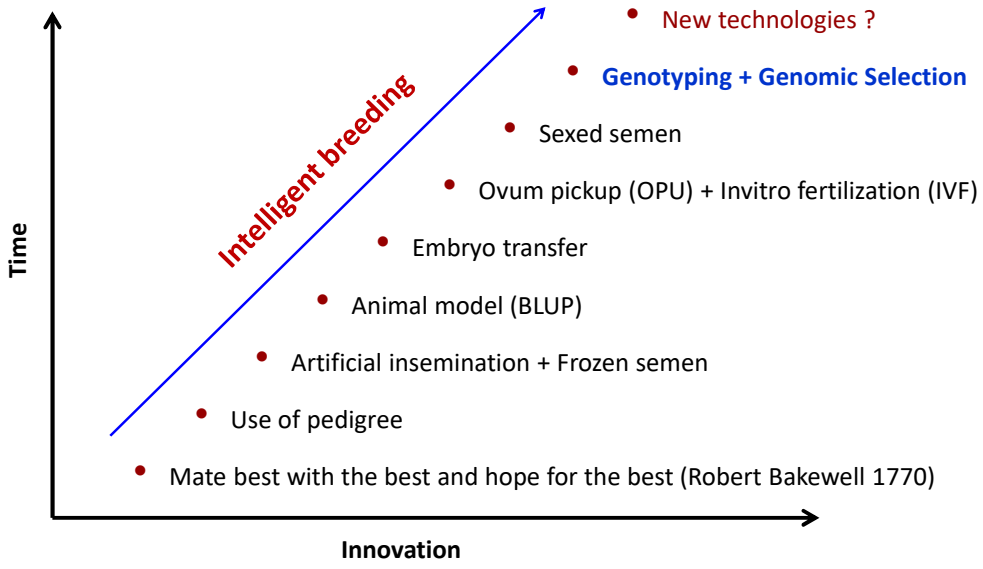
Feb 17, 2023

https://www.livescience.com/1st-uk-child-to-receive-gene-therapy-for-fatal-genetic-disorder-is-now-happy-and-healthy

9

9

Timeline of Major Events in Dairy Cattle Breeding

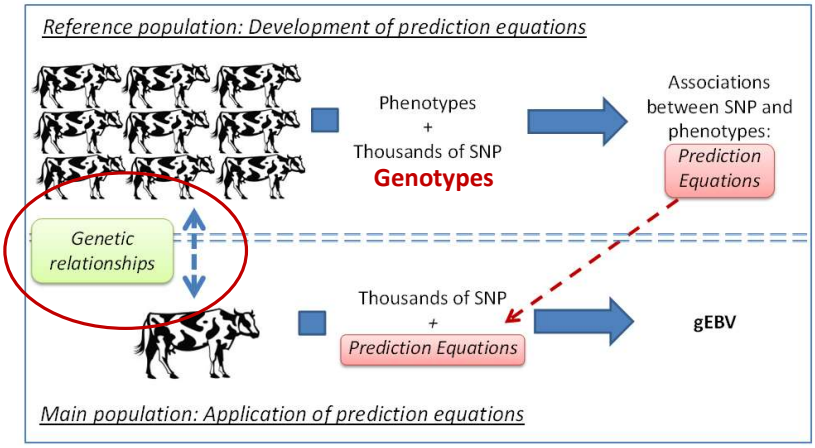


10

10

Introduction to Genomic Selection

Investment in continuous data collection is the key



Ref: Kor Oldenbroek and Liesbeth van der Waaij, 2015.

11

11

Introduction to Genomic Selection

Genetic evaluation:

- Variance components
- Genetic similarity between individuals

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z} + \mathbf{A}^{-1}\dfrac{\sigma_e^2}{\sigma_a^2} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{bmatrix}$$

G H

12

12

Introduction to Genomic Selection

From theory to application – An example:

- Theory of Individual Cow Model (Animal Model) was first proposed by Henderson in mid 1960's
- The national application of Animal Model in US began in 1989

Almost 20 years gap!

Computer Power

13

13

Introduction to Genomic Selection

From theory to application:

- Van Arendonk et al., · 1994; (Genetics: 137: 319-329)
- Nejati-Javaremi et al., 1997 (JAS 75: 1738-1745)
- Meuwissen et al., 2001 (Genetics 157: 1819-1829)
- VanRaden, 2008 (JDS 91: 4414-4423)

The national application of Genomic Selection in US began in 2008

- Dense marker data
- Computer power

The same story for Single Step BLUP

14

14

Genetic Marker

Genetic marker can help track the inheritance of a gene and it can be anything like DNA sequence or a phenotypic characteristic like polledness.

Examples of DNA markers:

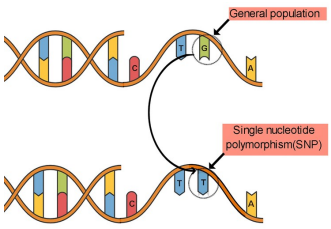
- Single Nucleotide Polymorphism (SNPs)
- Microsatellites
- Indels
- Restriction Fragment Length Polymorphisms (RFLPs)
- Variable Number of Tandem Repeats (VNTRs)
- Copy Number Variants (CNVs)

A DNA marker may or may not have a function

15

15

Genetic marker - Single Nucleotide Polymorphism



Source: DOI:10.1007/s11356-022-19981-7

An SNP is a genomic variant at a single base position in the DNA with at least 1% frequency in the population.

- Bi-allelic
- Most common type of DNA marker
- Uniformly distributed across the genome
- Lower information content at single locus
- Lower cost per marker
- SNPs make up about 90% of human genetic variation

16

16

Genotyping Technologies

- SNP Microarrays
  - Illumina's Infinium Beadchip assay
  - Affymetrix GeneChip Array
- TaqMan SNP assay
- MassArray SNP
- DNA Sequencing (NGS)

https://biocertica.com/blogs/genetics/what-are-other-types-of-genotyping-technologies

17

17

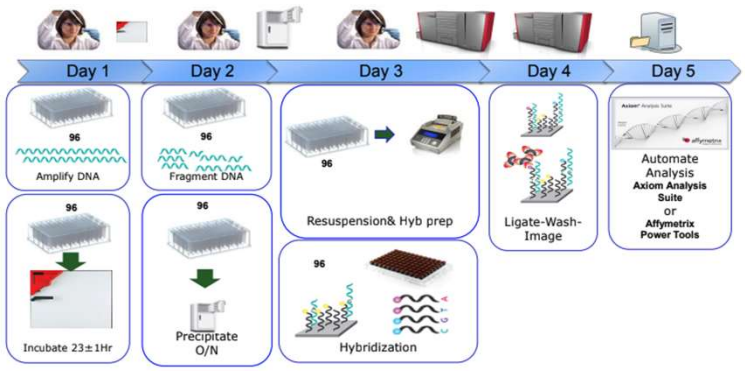
Genotyping Technologies – Important Factors

	Illumina Infinium microarray
• Call rate	>99%
• Accuracy of genotype call	>99.9%
• Reproducibility	>99.9%

18

18

Genotyping Technologies

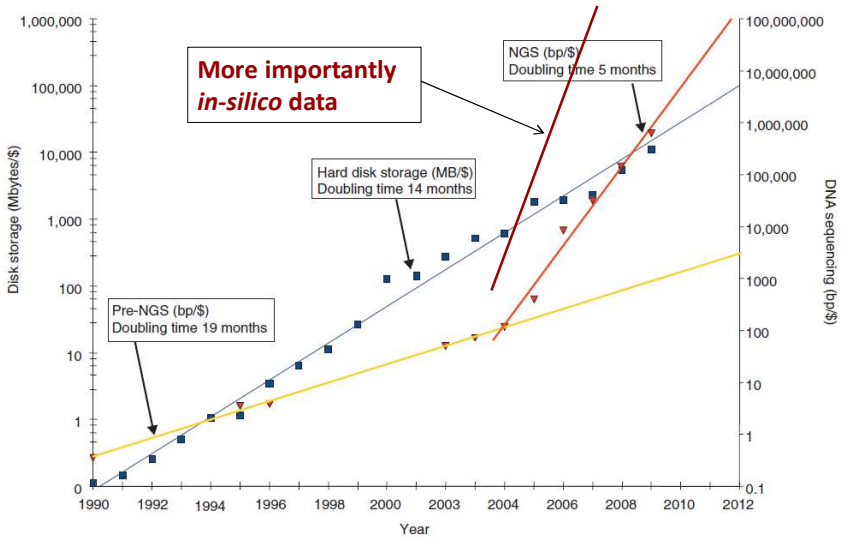


Source: https://biocertica.com/blogs/genetics/how-do-we-perform-genotyping

19

19

Challenges with Genomic Data – Big Data?



Ref: Stein 2010

20

20

Challenges with Genomic Data

Mardis *Genome Medicine* 2010, **2**:84
http://genomemedicine.com/content/2/11/84

Genome Medicine

MUSINGS

The $1,000 genome, the $100,000 analysis?

Elaine R Mardis*

Having recently attended the Personal Genomes meeting at Cold Spring Harbor Laboratories (I was an organizer this year), I was struck by the number of talks that described the use of whole-genome sequencing and analysis to reveal the genetic basis of disease in patients. These patients included a child with irritable bowel disease, a child with severe combined immunodeficiency, required for it to occur. I therefore offer the following as food for thought.

One source of difficulty in using resequencing approaches for diagnosis centers on the need to improve the quality and completeness of the human reference genome. In terms of quality, it is clear that the clone-based methods used to map, assign a minimal tiling path,

21

21

Genomic Data

Two main providers of microarray SNP chip and genotype call software are Illumina and Affymetrix.
The genotype files come in “Tall” or “Wide” format:

```
[Header]
GSGT Version  2.0.4
Processing Date            2023-01-01 12:00 AM
Content                    GGP100k.bpm
Num SNPs      95256
Total SNPs    95256
Num Samples   20
Total Samples 20
[Data]
SNP Name                    Sample ID  Allele1-Forward  Allele2-Forward  Allele1-Top  Allele2-Top  Allele1-AB  Allele2-AB  GC Score
10-104012831-C-G-rs442869917   Sample1    C    C    C    C    A    A    0.5420
10-15108992-A-G-rs384947169    Sample1    A    A    A    A    A    A    0.9090
10-15835936-G-A-rs209130723    Sample1    G    G    G    G    B    B    0.3396
10-26681293-G-A-rs453101503    Sample1    G    G    G    G    B    B    0.6591
10-26948606-C-T-rs384792959    Sample1    C    C    G    G    B    B    0.3390
10-27008241-A-C-rs42918694     Sample1    C    C    C    C    B    B    0.7581
10-27895449-A-G-rs451556029    Sample1    A    A    A    A    A    A    0.1042
10-37505397-T-A-rs135642375    Sample1    A    A    T    T    B    B    0.6645
10-37505419-T-C-rs136559242    Sample1    C    C    G    G    B    B    0.7314
10-46144755-G-A-rs135125777    Sample1    A    G    A    G    A    B    0.8879
10-47509723-A-T-rs467796086    Sample1    A    A    A    A    A    A    0.5256
10-49904259-G-A-rs471723345    Sample1    G    G    G    G    B    B    0.7448
10-6988001-T-C-rs211553144     Sample1    T    T    A    A    A    A    0.4207
10-81024106-T-G-rs448413483    Sample1    T    T    A    A    A    A    0.5362
```

22

22

Genomic Data

AB “Wide” format

Index	SNP Name	Sample1.Gtype	Sample2.Gtype	Sample3.Gtype	Sample4.Gtype	Sample5.Gtype		
1	ARS-BFGL-BAC-10919	AA	AA	AA	AB	AA	AA
2	ARS-BFGL-BAC-10975	AB	AA	AA	AA	AB	AB
3	ARS-BFGL-BAC-11000	AB	AA	AA	AB	AA	AA
4	ARS-BFGL-BAC-11003	AB	AA	AB	AA	AA	AA
5	ARS-BFGL-BAC-11025	AB	BB	BB	AA	BB	BB
6	ARS-BFGL-BAC-11044	AB	AB	AB	AB	BB	BB
7	ARS-BFGL-BAC-11193	AB	BB	AB	AB	AB	AA
8	ARS-BFGL-BAC-11215	AB	BB	AB	AA	AB	BB
9	ARS-BFGL-BAC-11218	AB	AB	AB	BB	BB	BB
10	ARS-BFGL-BAC-11276	BB	BB	BB	AB	AB	BB
11	ARS-BFGL-BAC-11283	AA	AA	AA	AA	AB	AB
12	ARS-BFGL-BAC-11513	AB	AA	AB	AB	AA	AA
13	ARS-BFGL-BAC-11612	BB	BB	AB	AB	AB	BB
14	ARS-BFGL-BAC-11657	BB	BB	BB	AB	BB	BB
15	ARS-BFGL-BAC-11666	AB	BB	AB	AB	BB	AB

Genomic Data

Why is AB allele coding preferred?

Designation of strand and allele is usually not consistent across platform, organization and assemblies

AB coding is a simple method that ensures uniformity of genotype calls

Illumina® SNP Genotyping

TECHNICAL NOTE

“TOP/BOT” Strand and “A/B” Allele

A guide to Illumina’s method for determining Strand and Allele for the GoldenGate® and Infinium™ Assays.

INTRODUCTION

To address DNA strand designation and orientation for both human and non-human species, Illumina has devel-

consistently designate the same SNP orientation and allele calls even if public SNP databases and genome assemblies change. This will enable researchers world-

https://www.illumina.com/documents/products/technotes/technote_topbot.pdf

Genomic Data – Format Conversion

The most common format to store genotypes:

- VCF (Variant Call Format)
- BED – PLINK
- **Plain text file**

VCF:

- Very flexible → Phased/unphased SNP, Indel
- Complex and not compressed
- Big file size → Slow to read

BED

- Specific to PLINK
- Binary
- Simple
- Slightly compressed

25

25

Genomic Data – Allele Coding for Additive Gene Action

Gene Content:

Gene Content is the number of copies of a reference allele for an SNP

- In the absent of mutation, its h2 is 1

Gene Content can be calculated on A B allele

2 = AA
1 = AB
0 = BB

Centered Coding:

1 = AA
0 = AB
-1 = BB

Parameter estimates (Same)
Marker effect (Same)
Breeding vales (Same)
GRM (Different)
Reliability (Different)

GSE: 2011, 43:25

Distances between allele codes within a marker is the same for both methods.

26

26

Genomic Data

Example of genotype file:

ID	genotypes
SAMPLE_123	00210215022102011...
SAMPLE_124	01201012212201111...
SAMPLE_125	11101202201220110...
SAMPLE_126	22102110021102101...
SAMPLE_127	10120050110010200...
SAMPLE_128	02222201052101111...
SAMPLE_129	11202210021102122...
SAMPLE_130	00021150120011201...
SAMPLE_131	21102020022010252...

.
.
.

QMSim / BLUPF90

ID	Chip	genotypes
SAMPLE_123	1	00210215022102011...
SAMPLE_124	1	01201012212201111...
SAMPLE_125	1	11101202201220110...
SAMPLE_126	1	22102110021102101...
SAMPLE_127	1	10120050110010200...
SAMPLE_128	1	02222201052101111...
SAMPLE_129	1	11202210021102122...
SAMPLE_130	2	00021150120011201...
SAMPLE_131	2	21102020022010252...

.
.
.

FImpute

27

27

Genomic Data

Map file:

SNPID	Chr	Pos
rs100	1	115
rs220	1	1567
rs272	1	2369
rs343	1	4034
rs423	1	8921
rs487	1	10561
rs499	1	11834
rs542	1	12956
rs589	1	14283

.
.
.

QMSim / BLUPF90

SNPID	Chr	Pos	Chip_HD	Chip_LD
rs100	1	115	1	0
rs220	1	1567	2	1
rs272	1	2369	3	0
rs343	1	4034	4	0
rs423	1	8921	5	2
rs487	1	10561	6	0
rs499	1	11834	7	3
rs542	1	12956	8	0
rs589	1	14283	9	4

.
.
.

FImpute

28

Genomic Data

Pedigree file:

ID	Sire	Dam	Gender
SAMPLE_123	Sire_A	Dam_F	M
SAMPLE_124	Sire_B	Dam_J	F
SAMPLE_125	Sire_D	Dam_B	M
SAMPLE_126	Sire_B	Dam_O	F
SAMPLE_127	Sire_H	Dam_I	M
SAMPLE_128	Sire_K	Dam_Q	M
SAMPLE_129	Sire_A	Dam_S	M
SAMPLE_130	Sire_H	Dam_V	M
SAMPLE_131	Sire_M	Dam_A	F

.
.
.

29

Statistics of Genomic Data & Quality Check

Most common:

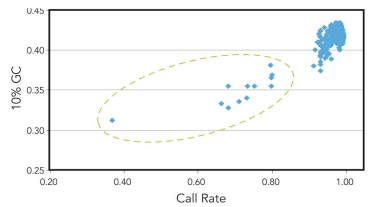
- Call Rate
- Minor Allele Frequency (MAF)
- Heterozygosity Rate
- Duplicate Samples/SNP
- Mendelian Errors
  - Parentage Test
  - Gender Conflicts
- Hardy-Weinberg Equilibrium
- Linkage Disequilibrium and Identification of Misplaced SNPs

30

QC: Call Rate

- Low call rate (either per SNP over samples or per sample over SNPs) is normally associated with low genotype quality

Figure 2: Scatter Plot of 10% GC Score Compared to Call Rates of a Sample Set



Poorly performing samples are obvious outliers from most of the samples when 10% GC Score is plotted against sample call rate (green circle).

Source: Illumina Genotyping Technical Note

- For Illumina and Affymetrix microarray genotypes, samples/SNPs with call rate bellow 95% should be excluded

31

QC: Minor Allele Frequency (MAF)

- If p is frequency of allele 1 and q frequency of allele 2 → MAF = min(p , q)
- If MAF is zero, the SNP is not segregating and has no information
- If MAF is very low, it may cause problem both numerically and statistically
- MAF filter level depends on power to detect associations (e.g. sample size)
- Usually, MAF filter is set to 1% but for large sample size it could be smaller

32

QC: Heterozygosity Rate

- Higher than expected heterozygosity is an indication of sample contamination and low genotyping quality
- Samples with 3 SD deviation from mean heterozygosity of population should be removed
- Excess of heterozygosity can also be used to removed low quality genotype (Difference between expected and observed)

abs(Observed Hetero rate -2pq) > 0.15

(Wiggans 2011)

33

QC: Duplicate Samples/SNP

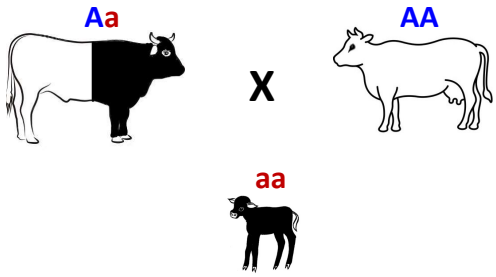
- Challenge: identification of duplicates in large data set
- A simple but not optimized solution:
  - Sort SNP based on allele frequency
  - Use partial search
  - Do parallel processing

```
T=200; //for 50K panel
Err=0;
for(i=0;i<n;++i)
{
   for(j=0;j<n;j+=500)
   {
      for(k=j;k<n && k<(j+500);++k)
      {
          if(g1!=g2) ++Err;
      }
      if(Err>T) break;
   }
   //if Err <=T duplicate detected
}
```

34

QC: Mendelian Errors

- Mendelian error is a genotype inconsistency between progeny and parent
  - Pedigree error
  - Sample mix-up
  - Poor genotype quality



35

QC: Hardy-Weinberg Equilibrium

Under HWE allele frequencies will stay the same across generations

- HWE assumption:
  - No mutation
  - random mating
  - no gene flow
  - infinite population size
  - no selection

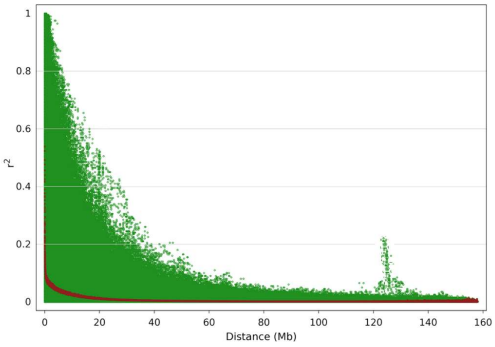
Genotype frequency under HWE: p^2, $2pq$, q^2

HWE test in livestock under intense selection may not be a definite indicator of poor genotype quality except for excess of heterozygosity

36

QC: Identification of Misplaced SNPs Using LD

- LD is non-random association of alleles at two loci
- SNPs located close to each other show high allelic correlation (most of the time inherited together)
- SNPs far apart or on different chromosomes show low allelic correlation



37

Simulation

38

Approaches to Solve the Problems

- **Field experiment**
  - Realistic but expensive and sometimes infeasible
- **Mathematical modeling**
  - Very useful but very simplified model
- **Simulation**
  - Flexible, complex model and works when math does not work

39

39

Simulation

- A simulation is the imitation of the operation of a real-world process or system over time (Banks et al. 2001)

40

40

Simulation Basics

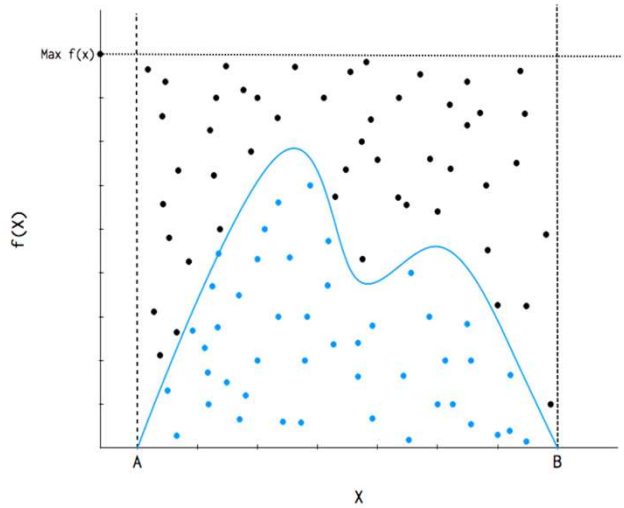
- Priors/Parameters (collected over time on historical data)
- We understand how the system works, but generating data/samples by modeling the system is very complex
- Simulation uses priors/parameters to generate each single force in the system and try to evolve the system over time (it may need long processing time)
- Simulation is used in almost all fields!

41

41

Simple Example of Stochastic Process Application

- Integration

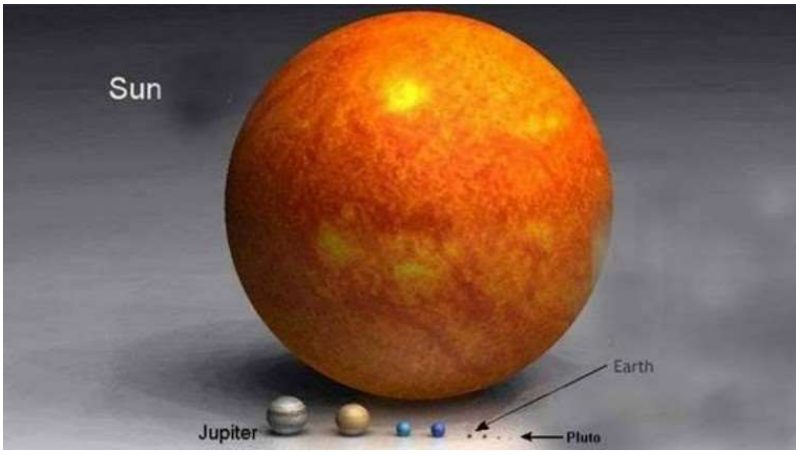


42

42

Another Example

- How many earth planets can fit inside the sun? **~1,300,000**
  Known: Sun radius is 109 times of earth radius



43

43

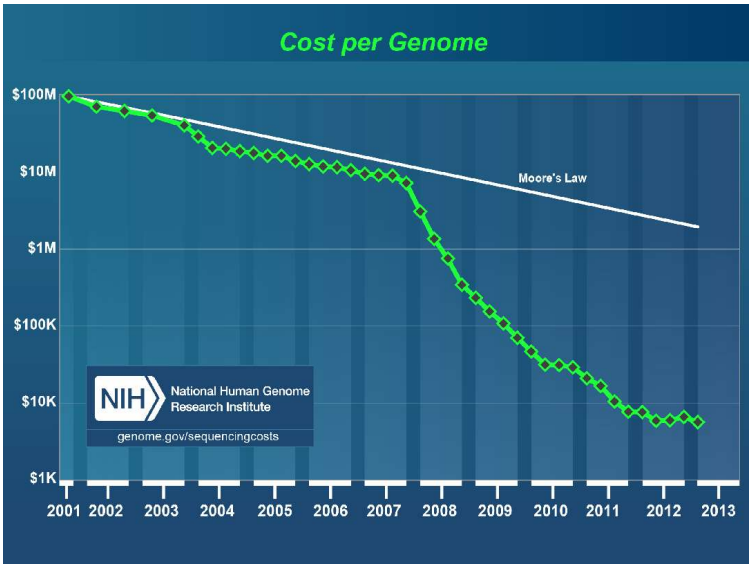
Why Genome Simulation

- Genomic/Sequence data is not available to many researchers
- Extremely cost effectives
- Generate tuned data for method comparison
- Simulation provides more detailed results
- Genome is very complex so deterministic modeling is difficult

44

44

Sequencing Cost



45

45

Population simulation

- Environmental effects
- Genetic effects
- Interactions

- Micro-evolution (within population)
  - Continuous
  - Over long period of time
- Macro-evolution (e.g. origin and extinction of populations)
  - Can be slow or fast

46

46

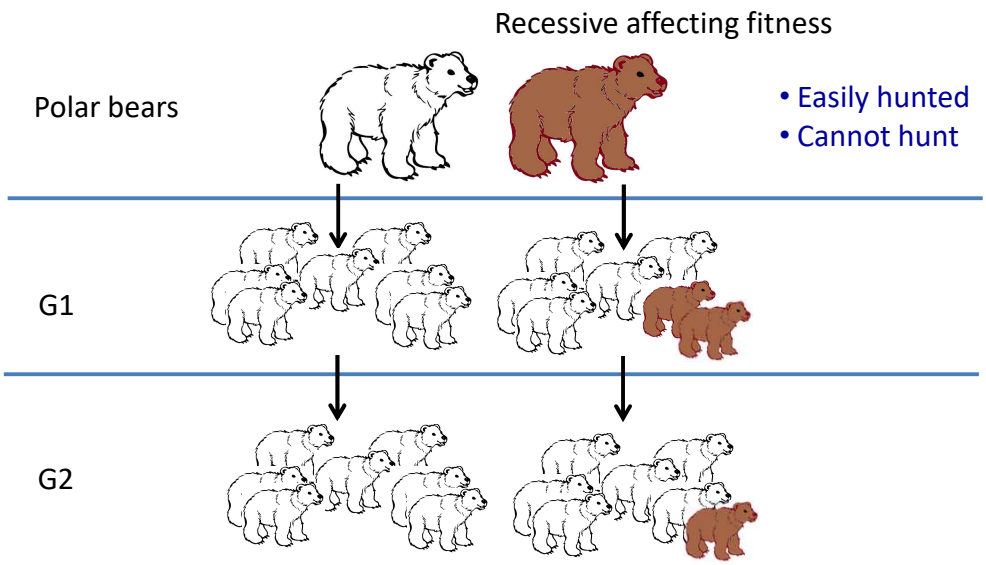
Evolution

- The genotype of an individual is static. But the fitness of individuals plays critical role in evolution
- Evolutionary changes happen in the population over generations when one generation turns to another
- Evolutionary forces change frequency of alleles in favor of adaptation to the environment

47

47

Evolutionary Forces



48

48

Evolutionary Forces

- Mutation
- Drift
- Migration (i.e. gene flow)
- Natural selection
- …

Evolution is a result of changes in allele frequencies

- Better adaptation to the environment (fitness)
- Higher rate of reproduction or more contribution to the gene pool of the next generations

49

49

Types of Simulation

- **Deterministic**
  - It has no random elements
  - Entire relation between input and output is known
  - One replicate is needed
  - Fast
  - Difficult to program
  - Complex systems are difficult to model
- **Stochastic**
  - It has at least one random element
  - It is usually not solved analytically
  - It is based on probability theory
  - Generally, large number of replicates are needed
  - Time consuming
  - Easy to program

50

50

Main Types of Population Simulation

- **Backward simulation** (coalescent)
- **Forward simulation** (gene dropping)

Backward method takes a set of sequences sampled today and work backwards in time to reconstruct their common ancestral sequence while forward simulation works generation by generation forward in time.

51

51

Forward *vs* Backward

Most of classical population genetics considers the future of a population given a starting point, the coalescent considers the present, while taking the past into account.

One can combine the two:

- Simulate the present population structure using backward simulation
- Simulate future generation using forward simulation

52

52

Pros and Cons

- Forward simulation can keep track of the complete ancestral information
- Forward simulation is much easier to take into account evolutionary forces like selection, migration, …
- Forward simulation can simulate more realistic samples
- Backward simulation is computationally more efficient than forward simulation
- Forward simulation is much more flexible

53

53

Hardy-Weinberg Equilibrium

- No drift (i.e., very large population size)
- No mutation
- No migration
- No selection
- Random mating
- Sex ratio of 1

Idealised population (no evolution)

54

54

How to Detect Evolution?

- If evolution is not occurring, then we should observe HWE or no change in allele frequencies over time
- Magnitude of change (+ or -) in allele frequency over generations points to intensity of evolution
  - A good example might be comparison of change in allele frequencies related to intelligence between species (human is still evolving!)

55

55

The Challenges

- We need priors/parameters for simulation (e.g., distributions, h^2, Vp)
- Estimation of past allele frequencies – Most of the time no historical data is available
- Past population size is not known accurately
- Genealogy

56

56

Some Solutions

- Simplify the model
- Calculated population size from current linkage disequilibrium information
- Start the simulation with equal allele frequencies or non-segregating alleles

57

57

Steps

- Simulate founders/base population
- Simulate recent generations from the founders

58

58

Simulating Polygenic Effect

- For base population

$$a \sim N(0, \sigma_a^2)$$

$$a_i = \sigma_a * NRnd$$

$$NRnd \sim N(0, 1)$$

- For recent generations

$$a_i = \frac{(a_{sire} + a_{dam})}{2} + MS_i$$

$$MS_i = \sigma_a * NRnd * \sqrt{0.5 - 0.25(F_{sire} + F_{dam})}$$

59

59

Simulating Residual Effect

- For all individuals

$$e \sim N(0, \sigma_e^2)$$

$$e_i = \sigma_e * NRnd$$

$$NRnd \sim N(0, 1)$$

60

60

Simulating Markers and QTLs

- For base population (historical)

  Priors/parameters

  - Allele frequency distribution
  - Number of alleles
  - Genetic architecture (Additive, polygenic, major genes?)
  - Demographic events in the past (domestication, migration, bottleneck, …)
  - Past effective population size

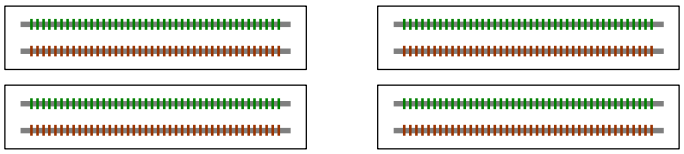
61

61

Simulating Markers and QTLs

Founder generation

Haplotypes



- Simulate large enough number of generations to reach mutation-drift equilibrium
- Simulate known demographic event
- No selection

62

62

Simulating QTL Effects

- QTL effects are simulated in the last historical generation

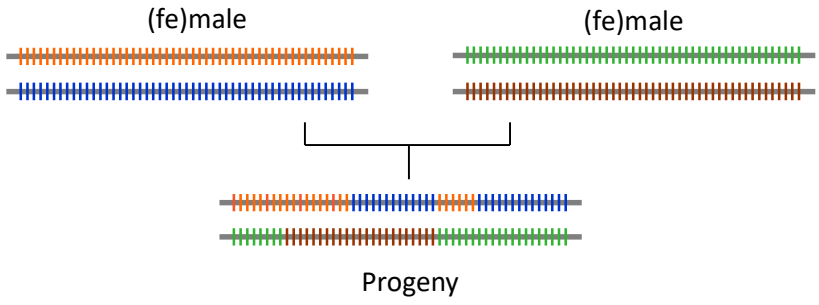
QTL allelic effects are first sampled based on the specified distribution (i.e., gamma, normal or uniform distribution) and then are scaled such that the sum of QTL variances in the last historical generation equals the input QTL variance.

63

63

Simulating Markers and QTLs

- For progeny in historical population
  - Sample two individuals regardless of gender (random union of gametes)
  - Sample the number of crossovers based on binomial distribution
  - Place the crossovers randomly
  - Start with one haplotype randomly and walk through the haplotypes
  - Simulate mutation

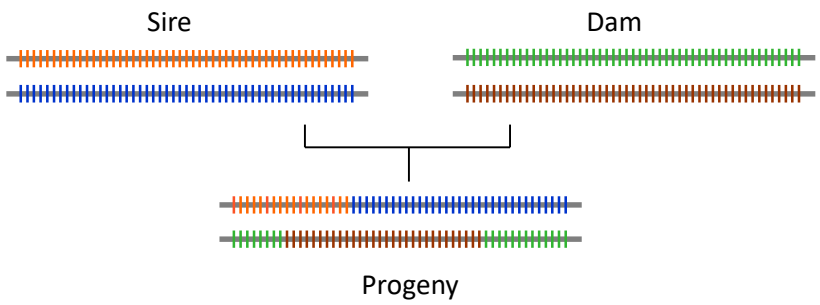


64

64

Simulating Markers and QTLs

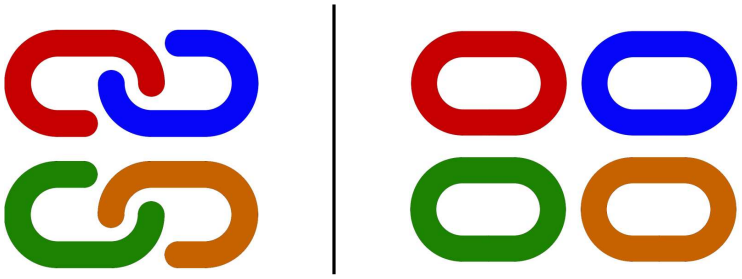
- For progeny after historical generations
  - Select a male and female based on desired mating design
  - Sample the number of crossovers based on binomial distribution
  - Place the crossovers randomly
  - Start with one haplotype randomly and walk through the haplotypes



65

65

Linkage Disequilibrium



66

Linkage Disequilibrium

- Non-random association of alleles at two loci because of physical linkage or co-selection
- If we know genotypes/alleles at one locus how well this information can help to predict unobserved genotypes/alleles at neighboring locus

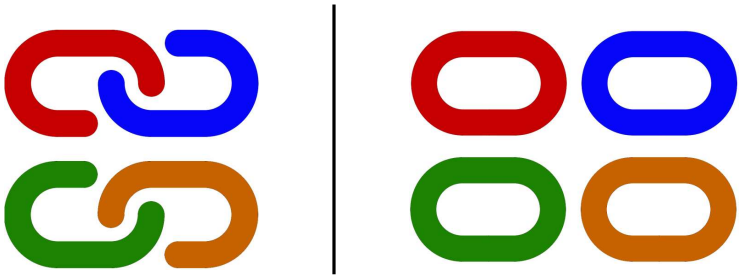
67

67

Linkage Disequilibrium – Cont'd

Linked
on the same chromosome

Unlinked
On different chromosomes



68

68

Linkage Disequilibrium – Cont'd

Observed frequencies

		Locus 2	
		B = ■	b = □
		$f_B = 0.5$	$f_b = 0.5$
Locus 1	A = ●		
	$f_A = 0.5$	● ■ $f_{AB} = 0.25$	● □ $f_{Ab} = 0.25$
	a = ○		
	$f_a = 0.5$	○ ■ $f_{aB} = 0.25$	○ □ $f_{ab} = 0.25$

Linkage equilibrium

		Locus 2	
		B = ■	b = □
		$f_B = 0.5$	$f_b = 0.5$
Locus 1	A = ●		
	$f_A = 0.5$	● ■ $f_{AB} = 0.35$	● □ $f_{Ab} = 0.15$
	a = ○		
	$f_a = 0.5$	○ ■ $f_{aB} = 0.15$	○ □ $f_{ab} = 0.35$

Linkage disequilibrium

69

69

Linkage Disequilibrium – Cont'd

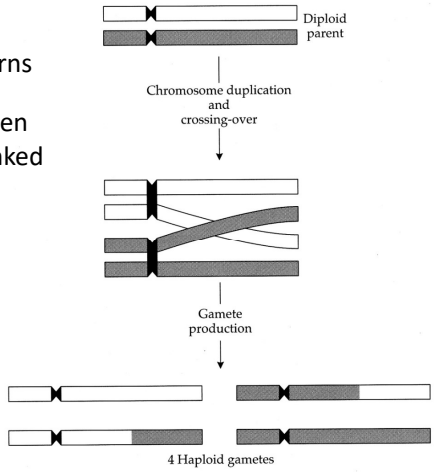
- When genes are linked, statistical dependence exists
- Linked genes tend to be inherited together
- This tendency declines as distance increases
- The decline is mainly the results of crossing-over

70

70

Crossing-over

- Cross-over happens when generation turns
- There is less chance of cross-over between close variants and therefore they stay linked together for longer period of time



Source: Lynch and Walsh, 1998

71

71

Causes of LD – Departure of Gamete Frequencies From Expectation

- Drift (small effective population size)
- Selection & assortative matings
- Mutation
- Migration
- Crossing

Main causes in livestock populations

72

72

How is LD Measured?

Deviation of gamete frequencies from expectation

When haplotypes are reconstructed/known

$$D = f_{AB} - f_A f_B$$

When haplotypes are unknown

$$D = \frac{n}{n-1}\left[\frac{4n_{AABB} + 2(n_{AABb} + n_{AaBB}) + n_{AaBb}}{2n} - 2f_A f_B\right]$$

where n is the number of individuals and n_{AABB}, n_{AaBB}, n_{AABb} and n_{AaBb} are the numbers of individuals for each genotype combination (Lynch and Walsh, 1998).

D can be positive or negative depending if A and B are in coupling or repulsion phase

73

73

How is LD Measured? – Cont'd

For bi-allelic markers / SNP

$$r^2 = \frac{D^2}{f_A f_a f_B f_b}$$

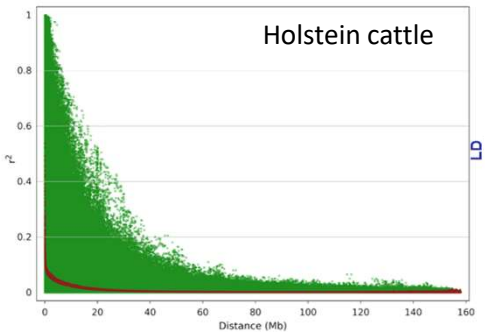
For multi-allelic markers D' or X^2 are usually used.

74

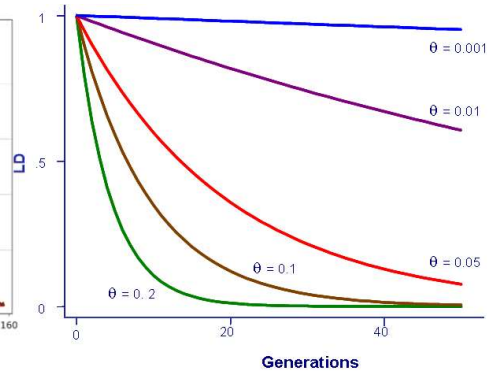
74

Decay of Linkage Disequilibrium

Decay of LD
as a function of distance



Decay of LD
as a function of generation



Persistency of LD over generations is a function of recombination or distance between the SNPs

75

75

LD *vs* Linkage

LD is non-random association of alleles at two loci in population

Linkage could be considered as LD within family

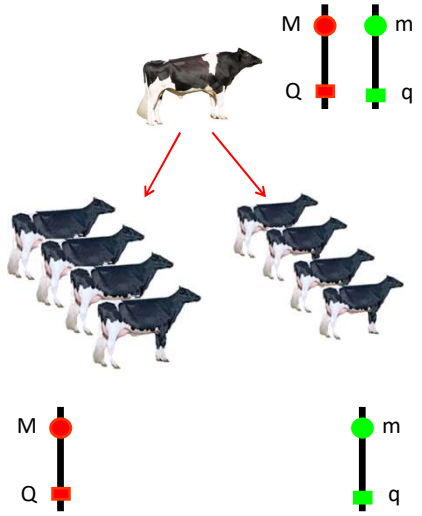
Decay of LD over generations is a function of distance between the markers

76

76

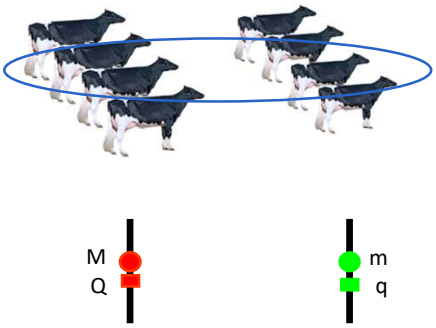
LD *vs* Linkage– Cont'd

Linkage



Linkage disequilibrium

- Random sample



77

77

Linkage Disequilibrium and Genetic Relationship

Extent of linkage disequilibrium:

Human ↓ Dairy cattle ↑

Relationship between individuals:

Human ↓ Dairy cattle ↑

Genetic drift
Selection

78

78

Linkage Disequilibrium and Genetic Relationship – Cont'd

- Close relatives share long haplotypes
- Distant relatives share short haplotypes

↓

Length of shared haplotypes between two individuals stores information about the genetic relationship

↓

Age of relationship (Meuwissen et al. 2014)

79

79

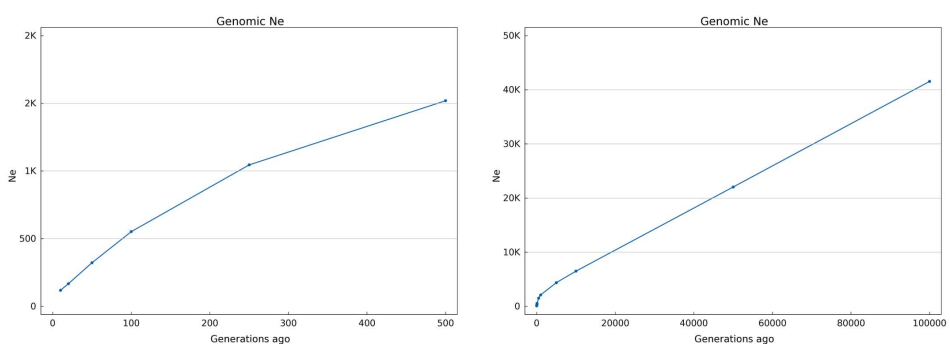
LD and Estimation of Past Ne

$$E(r^2) = \frac{1}{1 + 4Ne * c}$$

Ref: (Sved, 1971); c is distance in Morgan

$$\text{Age of Ne} = \frac{1}{2c}$$

Ref: (Hayes et al., 2003)



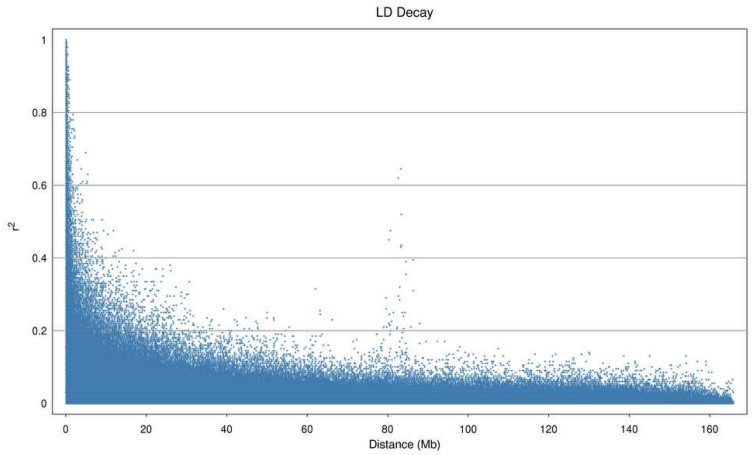
Holstein cattle

80

80

Finding Misplaced SNPs

A misplaced SNP shows low level of LD with nearby SNP but strong LD with distant SNPs



81

81

Available Resources

- Genetic simulation resources: https://popmodels.cancercontrol.cancer.gov/gsr/packages/
- List of generic simulation software: https://bioinformaticsonline.com/pages/view/8265/list-of-generic-simulation-softwaretoolsresource-with-brief-description-and-homepage

82

82

QMSim

83

83

QMSim – QTL & Marker Simulator

Designed for simulating:

- livestock populations
- Large scale genomic data
- Family based data (complex pedigree)
- Multiple populations
- Evolutionary forces
- Computational efficiency

84

84

Where to Get It?

https://animalbiosciences.uoguelph.ca/~msargol/qmsim/

QTL and Marker Simulator

Overview

Linkage disequilibrium (LD) and linkage analyses have been used extensively to identify quantitative trait loci (QTL) in human and livestock. Owing to the recent developments in sequencing and genotyping technologies, very dense marker maps are now available. Simulation is a very useful tool for assessing and validating new methods for genomic studies at low cost. During the last few decades, simulation has played a major role in answering a wide variety of questions in genomics. Several software have been developed for simulating genomes especially in human research. However most of the developed software tools do not provide functionality required for many of the applications in livestock.

QMSim was developed to simulate large scale genomic data in livestock populations. QMSim is a family based simulator, which can also take into account predefined evolutionary features, such as LD, mutation, bottlenecks and expansions. The simulation is basically carried out in two steps: In the first step, a historical population is simulated to establish mutation-drift equilibrium and, in the second step, recent population structures are generated, which can be complex. QMSim allows for a wide range of parameters to be incorporated in the simulation models in order to produce appropriate simulated data.

Features

- Simulates historical generations to establish mutation-drift equilibrium and create linkage disequilibrium.
- Recombination is appropriately modeled. Interference is allowed.
- Multiple chromosomes, each with different or similar density of markers and QTL maps, can be generated.
- Very dense marker map and also sequence data can be simulated.
- Missing genotypes and genotyping errors can be simulated.
- Markers can be either SNP or microsatellites.
- Males and females can have different genome length in cM.
- Unbalanced sex ratio is also allowed in historical populations.

85

85

How it Works?

Step 1

Historical population
- Random mating
- No selection
- Mutation
- Drift
- Bottleneck/expansion events

- Creating desired extent of LD
- Mutation-drift equilibrium

Step 2

Recent population(s)

Generation 1
↓
Generation 2
↓
Generation n

- Creating desired population structure
- Long range LD
- No mutation

86

86

Parameter file structure

- Global parameters
- Historical population parameters
- Recent population(s) parameters
- Genome parameters
- Output options

87

87

How to write a parameter file

Basic rules:

- Each commands end with a ;
- Each section begins with “begin” and ends with “end” command
- Comments
  //
  /* … */ multi-line

88

88

Global parameters

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

Whole genetic variation is explained
by simulated QTLs
Polygene h2 = h2 – qtlh2

89

89

Global parameters

```
/******************************
 **     Global parameters     **
 ******************************/
title = "Example 1 - 10k SNP panel";
nrep  = 1;                    //Number of replicates
h2    = 0.2;                  //Total heritability
qtlh2 = 0.1;                  //QTL heritability
phvar = 1.0;                  //Phenotypic variance
```

50% o genetic variation explained by simulated QTLs and 50% by polygenes

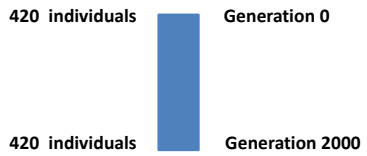
90

90

Historical population

```
/******************************
 **   Historical population   **
 ******************************/
begin_hp;
   hg_size = 420 [0] 420 [2000];   //Size of the historical generations
   nmlhg   = 20;                   //Number of males in the last generation
end_hp;
```

Population size [generation number]



91

91

Historical population

```
/******************************
 **   Historical population   **
 ******************************/
begin_hp;
   hg_size = 2000 [0] 150 [2000];  //Size of the historical generations
end_hp;
```

Represents a livestock population

92

92

Historical population

```
/******************************
 **   Historical population   **
 ******************************/
begin_hp;
   hg_size = 1000 [0] 3000 [4000];  //Size of the historical generations
end_hp;
```

Similar to the human population

93

93

Recent population(s)

```
/******************************
 **     Recent population     **
 ******************************/
begin_pop = "p1";             //Population name
 begin_founder;
      male   [n =  20, pop = "hp"];
      female [n = 400, pop = "hp"];
   end_founder;
   ls  = 2;                   //Litter size
   pmp = 0.5 /fix;            //Proportion of male progeny
   ng  = 10;                  //Number of generations
   md  = rnd;                 //Mating design
   sd  = tbv;                 //Selection design
   cd  = age;                 //Culling design
   begin_popoutput;
        data;
        stat;
        genotype /gen 8 9 10;
   end_popoutput;
end_pop;
```

94

94

Genome parameters

```
/******************************
 **          Genome           **
 ******************************/
begin_genome;
   begin_chr = 30;            //Number of chromosomes
      chrlen = 100;           //Chromosome length
      nmloci = 333;           //Number of markers
      mpos   = rnd;           //Marker positions
      nma    = all 2;         //Number of marker alleles
      maf    = eql;           //Marker allele frequencies
      nqloci = 25;            //Number of QTL
      qpos   = rnd;           //QTL positions
      nqa    = rnd 2 3 4;     //Number of QTL alleles
      qaf    = eql;           //QTL allele frequencies
      qae    = rndg 0.4;      //QTL allele effects
   end_chr;
   interference = 25;
end_genome;
```

95

95

Output options

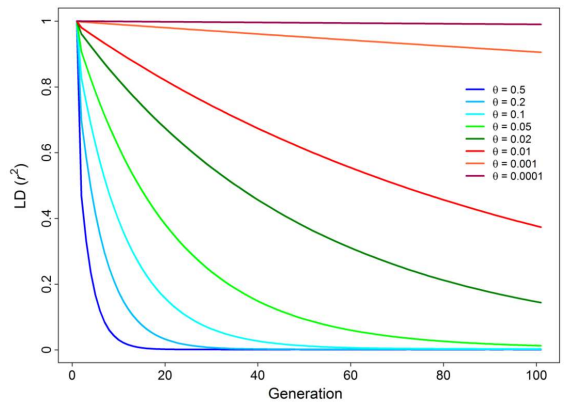
```
/******************************
 **      Output options       **
 ******************************/
begin_output;
   output_folder="output_s1"; //output folder name
   hp_stat;       //Save brief statistics on historical population
   linkage_map;   //Report linkage map (centiMorgan)
   allele_effect; //Save allele effect
end_output;
```

96

96

Decay of linkage disequilibrium

In an ideal population linkage disequilibrium decays exponentially over generations for long distances but not for short distances



Observed decay of linkage disequilibrium (LD) between adjacent marker pairs for different recombination rates (θ) in a simulated data set

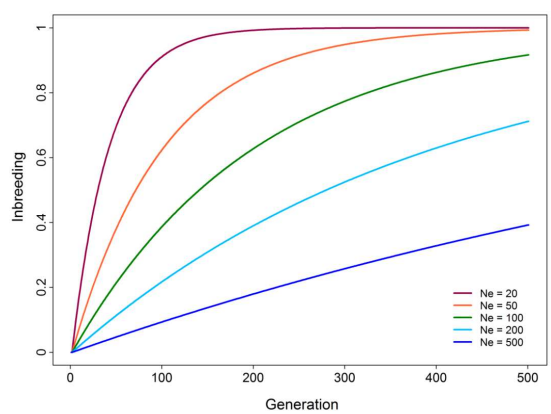
97

97

Inbreeding

Mean inbreeding in population:

$$F_t = 1 - (1 - \frac{1}{2Ne})^t$$



Observed average inbreeding coefficients over generations for different effective population sizes (*Ne*).

98

98

Mutation-drift equilibrium

- Mutation generates new variation
- Drift removes variation

- Mutation rates
- Effective population size (Ne)

Mutation:

- Infinite-allele mutation model
- Finite-allele mutation model

99

99

Mutation-drift equilibrium

At equilibrium

$$F_t = F_{t-1} = F_{t-2} \cdots = F$$

- Infinite-allele mutation model:

$$F = \frac{1}{1 + 4Neu}$$

- Finite-allele mutation model:

$$F = \frac{1 + \frac{4Neu}{k-1}}{1 + \frac{4Neuk}{k-1}}$$

k = the number of possible alleles

100

100

Mutation-drift equilibrium

Allele frequencies at equilibrium:

4Neu > 1 Normally distributed
4Neu = 1 Uniform
4Neu < 1 U-shape distribution

Ref: Wright, 1931

101

101

Mutation-drift equilibrium

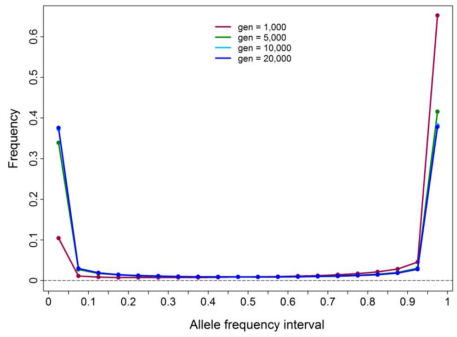
Parameters for different scenarios of mutation-drift equilibrium.

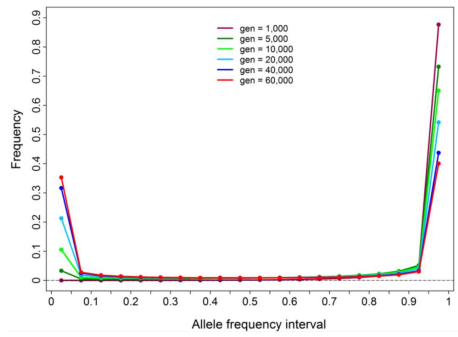
Parameters	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5	Scenario 6
Ne	100	1,000	1,000	500	1,000	2,000
u	2.5e-4	2.5e-5	2.5e-5	2.5e-4	2.5e-4	2.5e-4
4*Neu*	0.1	0.1	0.1	0.5	1	2
No. of SNP	10,000	10,000	10,000	10,000	10,000	10,000
Initial allele freq.	fixed	fixed	0.5	fixed	fixed	fixed
No. of gen.	1,000	1,000	1,000	1,000	1,000	1,000
	5,000	5,000	2,000	5,000	5,000	5,000
	10,000	10,000	5,000	10,000	10,000	10,000
	20,000	20,000	10,000	20,000	20,000	20,000
		40,000				
		60,000				
No. of replicates	100	100	100	100	100	100

102

102

Mutation-drift equilibrium





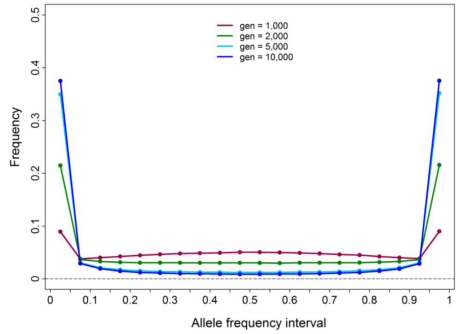
Scenario 1

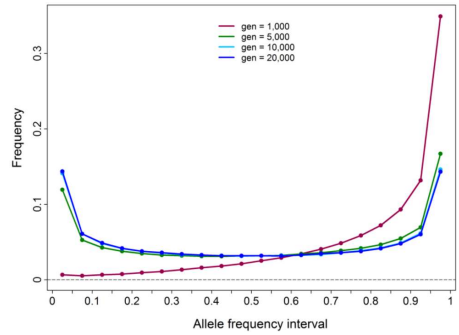
Scenario 2

103

103

Mutation-drift equilibrium





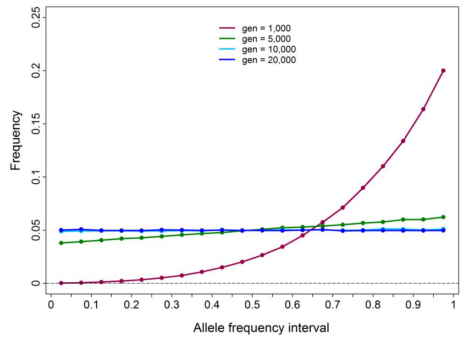
Scenario 3

Scenario 4

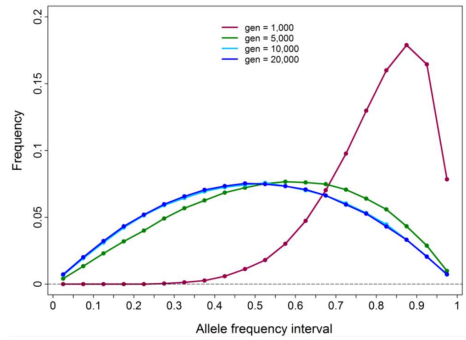
104

104

Mutation-drift equilibrium



Scenario 5



Scenario 6

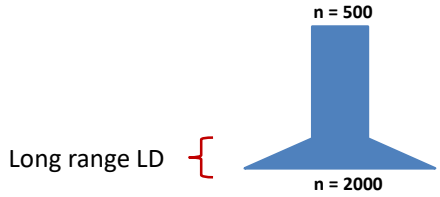
105

105

Mutation-drift equilibrium

Issue 1:
When large population size is considered, large number of generations is needed to establish mutation-drift equilibrium

Solution:
Simulate small number of individuals over smaller number of generations and then expand the population gradually to achieve desired population size.



106

106

Mutation-drift equilibrium

Issue 2:
Too many loci are going to fixation!

Solution:
Simulate larger number of loci and then select those that are segregating to get the desired density

```
/******************************
 **          Genome           **
 ******************************/
begin_genome;
   begin_chr = 10;
   ⋮
   end_chr;
   select_seg_loci /maft 0.01 /nmrk 50000 /nqtl 500;
genomeend_genome;
```

107

107

Monitoring the equilibrium

```
/******************************
 **      Output options       **
 ******************************/
begin_output;
   monitor_hp_homo /freq 100;
end_output;
```

---

hp_homo_mrk file:

```
Gen    Mean homozygosity
0     0.360734
100   0.398209
200   0.433752
300   0.468322
400   0.499805


3600  0.886277
3700  0.888996
3800  0.890092
3900  0.891668
4000  0.892113
```

108

108

Selecting founders for recent populations

```
/******************************
 **     Recent population     **
 ******************************/
begin_pop = "p1";
 begin_founder;
      male   [n =  20, pop = "hp"];
      female [n = 400, pop = "hp"];
   end_founder;
   ⋮
end_pop;

begin_pop = "p2";
 begin_founder;
      male   [n =  20, pop = "p1", gen = 10, select = tbv /h];
      female [n = 400, pop = "p1", gen = 10, select = tbv /h];
   end_founder;
   ⋮
end_pop;
```

From the last historical generation

Should be defined before “p2”

109

109

Selecting founders for recent populations

```
begin_pop = "p3";
 begin_founder;
      male   [n =  10, pop = "p1", gen = 10, select = tbv /h];
      male   [n =  10, pop = "p2", gen = 10, select = tbv /h];
      female [n = 200, pop = "p1", gen = 10, select = tbv /h];
      female [n = 200, pop = "p2", gen = 10, select = tbv /h];
   end_founder;
   ⋮
end_pop;
```

110

110

Mating & selection in recent populations

```
/******************************
 **     Recent population     **
 ******************************/
begin_pop = "p1";
 begin_founder;
      male   [n =  20, pop = "hp"];
      female [n = 400, pop = "hp"];
   end_founder;
   ⋮
   md  = rnd;                 //Mating design
   sd  = tbv;                 //Selection design
   cd  = age;                 //Culling design
end_pop;
```

Mating design (md) can be:
rnd = random; rnd_ug = random union of gametes;
minf = minimizing inbreeding; maxf = maximizing inbreeding;
p_assort = positive assortative; n_assort = negative assortative; (/phen, /ebv, /tbv)

111

111

Mating & selection in recent populations

```
/******************************
 **     Recent population     **
 ******************************/
begin_pop = "p1";
 begin_founder;
      male   [n =  20, pop = "hp"];
      female [n = 400, pop = "hp"];
   end_founder;
   ⋮
   md  = rnd;                 //Mating design
   sd  = tbv;                 //Selection design
   cd  = age;                 //Culling design
end_pop;
```

Selection design (sd) can be:
rnd = random
phen = based on phenotype
tbv = based on true breeding value
ebv = based on estimated breeding value

/l or /h can be used to select low or high values

112

112

Mating & selection in recent populations

```
/******************************
 **     Recent population     **
 ******************************/
begin_pop = "p1";
 begin_founder;
      male   [n =  20, pop = "hp"];
      female [n = 400, pop = "hp"];
   end_founder;
   ⋮
   md  = rnd;                 //Mating design
   sd  = tbv;                 //Selection design
   cd  = age;                 //Culling design
end_pop;
```

Culling design (cd) can be:

rnd = random
phen = based on phenotype
tbv = based on true breeding value
ebv = based on estimated breeding value
age = based on age

/l or /h can be used to select low or high values

113

113

Estimation of breeding values

```
/******************************
 **     Recent population     **
 ******************************/
begin_pop = "p1";
 begin_founder;
      male   [n =  20, pop = "hp"];
      female [n = 400, pop = "hp"];
   end_founder;
   ⋮
   md  = rnd;                 //Mating design
   sd  = ebv;                 //Selection design
   cd  = age;                 //Culling design
   ebv_est = blup;
end_pop;
```

EBV estimation method can be:

blup = best linear unbiased prediction
approx = approximation based on sibs information
accur = approximation based on user defined accuracy
external_bv = user is responsible to estimate breeding values in each generation

114

114

External estimation of breeding values

ebv_est = external_bv "external_solver";

Each generation

1) QMSim creates “data.tmp”
2) QMSim runs “external_solver”

   2.1) external_solver reads “data.tmp”, estimates breeding values and writes the results in “my_bv.txt”

3) QMSim reads “my_bv.txt” and progresses to the next generation

115

115

Outputs

ld_decay file

```
No. marker pairs
--------------------
Bin\Gen.          0         1
[0,.05)        5497      5560
[.05,.1)       5541      5479
[.1,.2)       10666     10896
[.2,.3)       10716     10533
[.3,.4)       10478     10427
[.4,.5)       10569     10401
[.5,.6)       10417     10437
[.6,.7)       10363     10372
[.7,.8)       10271     10204
[.8,.9)       10230     10195
[.9,1)        10177     10052
[1,2)        100300    100636
[2,3)         99853     99905
[3,4)         98099     97956
[4,5)         96133     95345
```

Distance in cM

```
Mean R2(SD)
--------------------
Bin\Gen.                0                1
[0,.05)   0.3709(0.3994) 0.3543(0.3961)
[.05,.1)  0.3308(0.3758) 0.3100(0.3666)
[.1,.2)   0.2862(0.3446) 0.2641(0.3339)
[.2,.3)   0.2408(0.3071) 0.2264(0.3011)
[.3,.4)   0.2255(0.2930) 0.2099(0.2834)
[.4,.5)   0.2063(0.2771) 0.1914(0.2677)
[.5,.6)   0.1916(0.2641) 0.1793(0.2555)
[.6,.7)   0.1860(0.2555) 0.1695(0.2436)
[.7,.8)   0.1729(0.2401) 0.1568(0.2274)
[.8,.9)   0.1674(0.2319) 0.1497(0.2202)
[.9,1)    0.1604(0.2267) 0.1447(0.2139)
[1,2)     0.1384(0.2037) 0.1263(0.1934)
[2,3)     0.1123(0.1697) 0.1030(0.1621)
[3,4)     0.0961(0.1481) 0.0873(0.1404)
[4,5)     0.0872(0.1370) 0.0793(0.1300)
```

116

116

Outputs

mrk file

```
ID     Genotypes (paternal allele, maternal allele) ...
35521   2 2 1 1 1 1 1 1 2 1 2 2 2 2 1 2 2 2 2 2 2 2 2 2 1 1 1 2...
35522   2 2 1 1 1 1 1 1 2 1 2 2 2 2 1 2 2 2 2 2 2 2 2 2 1 1 1 2...
35523   2 2 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 2 2...
35524   2 2 1 1 1 1 1 1 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 1 1 1 2...
35525   2 2 1 2 1 1 1 1 2 1 2 2 2 2 1 2 2 2 2 2 2 2 2 2 1 1 1 2...
35526   2 2 1 1 1 1 1 1 2 1 2 2 2 2 1 2 2 2 2 2 2 2 2 2 1 1 1 2...
35527   2 2 1 1 1 1 1 1 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 1 1 1 2...
35528   2 2 1 1 1 1 1 1 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 1 1 1 2...
35529   2 2 1 1 1 1 1 1 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 1 1 1 2...
.
.
.
```

117

117

Outputs

mrk file (in SNP genotype code)

```
ID     Genotypes (0 = a1,a1; 2 = a2,a2; 3 =
35521  2222224222022440323224222332424222202 2...
35522  2222422222022220020222222002222222202 2...
35523  2222224222022440323224222332424222202 2...
35524  2222224222022440323224222332424222202 2...
35525  2222222222022440323224223324222222022...
35526  2222222222022440323224223324222222022...
35527  2222222222022440323224223324242222022...
35528  2222222222022220020222222002222222022...
35529  2222222222022440323224223324242222022...
35530  2222222222022440323224223324242222022...
35531  2222222222022220323224223324242222022...
35532  2222224222022440323224223324242222022...
35533  2222422222022220020222222002222222022...
.
.
.
```

118

118

Outputs

linkage map file

```
ID          Chr    Position
M1           1     0.07522
M2           1     0.14365
M3           1     0.37741
M4           1     0.38784
M5           1     0.41830
M6           1     0.54802
M7           1     0.71100
M8           1     0.78760
M9           1     0.82219
M10          1     0.86985
M11          1     0.92948
M12          1     1.02755
M13          1     1.06030
M14          1     1.07831
.
.
.
```

119

119

Outputs

freq file

```
ID        Gen Chr  Allele:Freq ...
M1          1   1    2:1.000000
M2          1   1    1:0.970000   2:0.030000
M3          1   1    1:1.000000
M4          1   1    1:1.000000
M5          1   1    1:0.688000   2:0.312000
M6          1   1    1:0.069000   2:0.931000
M7          1   1    2:1.000000
M8          1   1    1:0.100000   2:0.900000
M9          1   1    2:1.000000
M10         1   1    2:1.000000
M11         1   1    2:1.000000
.
.
.
```

120

120

Outputs

qtl effect file

```
ID      Chr   Allele:Effect ...
Q1        1     1:-0.000059   2: 0.000174
Q2        1     1:-0.000131   2: 0.015543
Q3        1     1: 0.000894   2:-0.004144
Q4        1     1: 0.000004   2:-0.000004
Q5        1     1:-0.002999   2: 0.007286   3: 0.006781
Q6        1     1: 0.000039   2:-0.000245
Q7        1     1:-0.000104   2: 0.000891
Q8        1     1:-0.000024   2: 0.004732
Q9        1     1: 0.001907   2:-0.002520
Q10       1     1: 0.000703   2:-0.000330
Q11       1     1:-0.000248   2: 0.008515
Q12       1     1:-0.000040   2: 0.003418
Q13       1     1:-0.000007   2: 0.000335
Q14       1     1: 0.000079   2:-0.004041
Q15       1     1: 0.000138   2:-0.000046
Q16       1     1:-0.003273   2:-0.011271   3: 0.007570
.
.
.
```

121

121

Data backup and transfer

- All you need is to backup the initial seed and the parameter file
- Use the same parameter file and initial seed with the “seed” command to generate identical output

122

122

QMSim main limitations

- One historical population
- Single trait
- No dominance and epistatic effects
- And many more !!!

123

123