

발간등록번호

11-1543000-000790-01

개체 유전체 육종가 추정 시스템(GBLUP) 및 개량체계 개발

(Development of Individual Genomic Breeding Value
Estimation and Genetic Improvement System)

한경대학교 산학협력단

농림축산식품부

제 출 문

농림축산식품부 장관 귀하

이 보고서를 “개체 유전체 육종가 추정 시스템(GBLUP) 및 개량체계 개발” 과제의 보고서로 제출합니다.

2015 년 1 월 27 일

주관연구기관명 : 한경대학교

주관연구책임자 : 이 득 환

세부연구책임자 : 이 득 환

연 구 원 : 이 학 교

연 구 원 : 도 창 회

연 구 원 : 이 준 호

연 구 원 : 조 충 일

연 구 원 : 박 종 은

연 구 원 : 김 병 우

연 구 원 : 손 지 현

연 구 원 : 박 성 식

연 구 원 : 진 석 훈

연 구 원 : 나 호 철

연 구 원 : 강 현 중

연 구 원 : 김 자 영

연 구 원 : Daniel Vasco

협동연구기관명 : 경상대학교

협동연구책임자 : 이 정 규

연 구 원 : 임 현 태

연 구 원 : 김 열 중

연 구 원 : 손 난 주

연 구 원 : 문 슬 기

연 구 원 : 김 희 성

연 구 원 : 신 준 호

연 구 원 : 전 진 태

연 구 원 : 이 길 주

연 구 원 : 박 경 옥

연 구 원 : 김 범 모

연 구 원 : 이 재 봉

연 구 원 : 정 은 지

연 구 원 : 류 제 선

연 구 원 : 이 한 준

연 구 원 : 장 명 석

협동연구기관명 : (주)젠닥스

협동연구책임자 : 유 병 호

연 구 원 : 이 용 욱

연 구 원 : 최 규 명

연 구 원 : 차 현 욱

연 구 원 : 이 제 현

연 구 원 : 박 순 욱

연 구 원 : 장 진 원

연 구 원 : 최 미 란

연 구 원 : 김 정 근

연 구 원 : 배 현 경

연 구 원 : 박 종 관

연 구 원 : 정 우 영

연 구 원 : 신 지 섭

연 구 원 : 김 대 중

연 구 원 : 이 성 수

연 구 원 : 김 덕 임

연 구 원 : 김 진 호

연 구 원 : 이 학 필

연 구 원 : 노 승 희

요 약 문

I. 제 목 : 개체 유전체 육종가 추정 시스템(GBLUP) 및 개량체계 개발

II. 연구개발의 목적 및 필요성

- 가축의 유전체 관련 연구들을 활용하여 산업적으로 사용되는 사례가 매우 적었으나 최근 들어 선진국을 중심으로 유전체 선발이라는 새로운 기술이 도입되어 경제형질의 개량에 유전체 정보를 이용하는 사례가 있는 시점에 국내 종축에 대한 유전체 정보 활용 개량방안에 대한 규명 및 산업화가 절실히 필요한 시점임
- 어린가축에 대하여 유전체 정보를 활용한 선발방법의 개발로써 보다 정확한 방법에 의한 조기 선발 및 선발종축의 활용으로 세대간격을 단축함으로써 경제적 소요비용의 절감 및 개량의 효율을 증가시키기 위함
- 한편, 단일염기다형(SNP) 마커 등을 통해 집단의 유전적 구성이 지속적으로 변화하기 때문에 유전체 정보를 활용한 유전능력 추정(GBLUP)과 개별 SNP 마커별 추정치가 요구되는데, 이러한 추정치는 집단이 다르게 되면 정확성이 현저히 떨어지게 되므로 우리나라의 한우, 젓소, 돼지 등에 적용하기 위해서는 우리나라의 종축집단에 대하여 마커분석 및 능력검정을 지속적으로 실시할 필요성이 있으며 한우, 젓소 및 종돈 등의 신속한 개량을 위해서는 GBLUP체계 구축을 위한 연구가 필요함

III. 연구개발 내용 및 범위

1. 연구개발의 내용

- 가. 한우 및 종돈 종축의 유전체 마커정보와 혈통 정보를 동시에 수집하여 육종가 추정
- 나. 유전체 정보 활용 종축(한우 및 종돈)의 유전능력 평가방법 개발(참조집단)
- 다. 후보 종축들의 유전능력 조기 예측방법 및 예측치의 정확도 추정(적용집단)
- 라. 유전체 정보 활용 한우 및 종돈 개량체계 개발

2. 연구개발의 범위

- 가. 한우의 종모우 및 검정송아지들에 대한 유전체 정보 및 도체성적 조사(1,500두/5년)
- 나. 종돈(Landrace순종돈)의 모돈 및 웅돈들에 대한 유전체 정보 및 번식성적 조사(1,000두/5년)
- 다. 유전체 정보 활용 육종가 평가방법 개발 및 평가방법에 따른 육종가 추정치의 정확도 모의실험
- 라. 한우의 수집된 유전체 정보 활용 도체형질에 대한 육종가 추정 및 기존 육종가 추정방법과 비교
- 마. 종돈의 수집된 유전체 정보 활용 번식형질에 대한 육종가 추정 및 기존 육종가 추정방법과 비교
- 바. 한우의 수집된 유전체 정보 활용 어린 후보축의 도체형질에 대한 육종가 추정방법 개발
- 사. 종돈의 수집된 유전체 정보 활용 어린 후보축의 번식형질에 대한 육종가 추정방법 개발
- 아. 한우의 수집된 유전체 정보 활용 종축선발체계 개발
- 자. 종돈의 수집된 유전체 정보 활용 종축선발체계 개발

IV. 연구개발결과

1. 모의실험을 통한 유전체 정보 활용 육종가 평가방법 개발 및 정확도 비교

- 유전체 추정육종가(GEBV)에 영향을 미치는 요인은 양적형질좌위(QTL) 밀도, 유전력, 표지인자밀도, 참조집단의 크기 등을 고려할 수 있으며, 각각의 조건에 따라 총 16개 조합에 따른 모의 유전체 자료를 각 20회 반복 생성하여 베이지안과 GBLUP 방법에 따른 GEBV의 정확도에 대한 모의실험을 실시한 결과,
- 각각의 조건에 따라 정확도 추정치의 차이를 보였으나, 여러 조건 중 형질에 대한 조건(양적형질좌위 밀도, 유전력)은 분석자가 바꿀 수 있는 내용이 아니므로 표지인자의 밀도와 참조집단의 크기가 조절 가능한 조건이 되는데 특히 참조집단의 크기가 400두일 때와 2,000두일 때 약 13%의 육종가 추정치에 대한 정확도의 차이를 보였다.
- 따라서, 유전체 정보를 이용한 육종가 추정은 지속적으로 참조집단의 유전체 정보를 누적하는 것이 더 높은 정확도를 얻기 위해 중요하다 할 수 있다.
- 또한 다량의 표지인자를 이용한 유전체 육종가의 추정은 기존의 혈통자료를 이용하는 BLUP 방법보다 향상된 정확도를 보였으며, 특히 표현형이 없이 유전체 정보만을 가진 개체에서도 높은 정확도를 보여 조기선발을 정확하게 시행할 수 있다는 가능성을 확인할 수 있었다.

2. 한우 유전체 정보 수집(Reference Population) 및 도체형질에 대한 유전체 추정육종가의 정확도 추정

- 5년간 한우개량사업소 생산 후대검정우 및 후보씨수소 총 **1,887두**에 대하여 채혈을 실시하였으며, 이들 축군에 대한 도체성적과 혈통자료를 수집하여 연구에 이용하였다.
- 유전체 전장의 SNP 정보는 Illumina BovineSNP50 v1, v2 Kit를 활용 959두의 시험축군에 대하여 수집하였으며, Illumina Bovine HD Beadchip Kit 활용 720두에 대하여 수집하여 총 **1,679두(목표:1,500두)**에 대한 유전체전장의 단일염기다형(SNP)를 분석하였다.

년도	SNP 분석			DNA Analysis Kit
	후보종모우	후대검정우	계	
2010	35	253	288	BovineSNP50 V1 Beadchip
2011	39	297	336	BovineHD Beadchip(700K)
2012	37	347	384	BovineHD Beadchip(700K)
2013	35	301	336	BovineSNP50 V2 Beadchip
2014	36	299	335	BovineSNP50 V2 Beadchip
계	182	1,497	1,679	

- 본 연구에서는 50K와 700K로 상이한 유전체 데이터를 수집하였기 때문에 아래 두 자료 변환 방법에 따라 이상치에 대한 품질평가 및 육종가 추정을 실시하였다.
- ① Imputation 프로그램을 이용하여 50K를 700K로 변환시키는 방법으로 29개 상업색체 내의

유효 SNP 735,239개 중에서, 시험축군에서 다형성이 없는 마커 45,722개, Minor allele frequency가 5%이하인 마커 110,057개, 결측률이 10% 이상인 마커 5,702개를 제외하고 573,758개의 마커를 이용하여 ssGBLUP 방법을 통해 도체형질에 대한 육종가를 추정하였으며 육종가의 정확도를 계산한 결과, 개체 자신의 표현형이 없는 집단에서 유전체정보를 활용한 육종가 추정치는 유전체 정보를 활용하지 않았을 때 보다 약 4-6%의 추정치에 대한 정확도의 향상이 있었다.

- ② 50K와 700K 간 중첩되는 SNP를 선별하고 품질평가에 의한 유효 마커 30,748개의 유전체 정보를 활용하여 육종가를 추정한 결과, 700K로 유전자형을 Imputation하여 얻은 결과와 동일하게 BLUP방법에 의해 추정한 육종가보다 5-6%의 육종가추정치의 정확도가 향상되었다.
- 또한 50K 유전체 정보에서 연관불평형의 값을 이용하여 Tag SNP을 선별하여 저밀도 (8K) SNP를 구성하고 유전체 육종가 추정을 통하여, 유전체선발에서 저밀도 chip의 활용 가능성에 대해 알아본 결과, 표현형정보가 없는 집단에서 등심단면적 0.50, 등지방두께 0.52, 근내지방도 0.53의 유전체 추정육종가의 정확도가 있는 것으로 추정되어 약 8-10%의 향상효과가 있는 것으로 추정되었다.
- 따라서 가축개량을 위한 한우의 육종가 추정 시 유전체자료의 활용은 육종가의 정확도 향상에 도움을 줄 수 있을 것이라 판단되며, 표현형을 가지는 개체의 육종가 정확도가 70% 수준임을 감안할 때, 향후 유전체 선발의 정확도가 70% 이상이 되도록 지속적으로 참조집단의 유전체 정보를 추가할 필요가 있음을 시사한다.

3. 유전체 정보 활용 한우의 육종체계 및 산업화 방안

- 한우의 유전체 정보의 활용은 크게 씨수소 선발과 우량 암소의 조기선발에 활용될 수 있을 것이라 판단된다.
- 씨수소 선발을 위한 검정체계에서 당대검정 대상 송아지 선발 시 유전체 전장의 선별된 SNP(8K) 정보를 활용하여 육질관련 형질들에 대한 육종가를 기초로 1차 선발(1,000두/2000두)함으로써 선발강도 및 선발의 정확도를 향상하면 연간 약 640억원의 추가 이익이 발생할 수 있는 것으로 추정되었다. 또한 후대검정 송아지 및 8K SNP정보를 갖는 송아지들에 대한 표현형 자료를 수집함으로써 참조집단의 크기를 향상하여 유전체 정보를 활용한 선발의 정확도를 향상해야할 것으로 판단된다. 향후 10,000두의 참조집단정보를 확보하면 SNP정보에 의한 육종가 추정치의 정확도를 70%까지 향상시킬 수 있을 것으로 사료되었다.
- 암소집단에서의 유전체정보 활용은 암송아지의 조기선발 및 친자확인 등 복합적으로 적용할 수 있을 것이라 판단되며, 한우 10만두/년 이상을 8K chip을 통하여 분석할 경우 소요비용은 두당 80,000원 이하로 분석이 가능하여, 현행 지역단위 한우암소개량사업의 초음파도체형질의 육종가 및 Microsatelite marker를 활용한 친자감별을 대체하여 더 정확한 선발지표로서 사용이 가능할 것으로 예측되며 암소집단에 적용 시 연간 약 270억원의 개량효과가 있을 것으로 추정되었다.

4. 종돈 유전체 정보 수집 및 번식형질에 대한 유전체 육종가 정확도의 추정

- 5년간 Landrace 종돈 1,448두에 대하여 채혈을 실시하였으며, 이들 축군에 대한 번식성적과 혈통 자료를 수집하여 연구에 이용하였다.
- SNP를 통한 유전체 정보는 Illumina PorcineSNP60 v1를 이용하여 680두, v2로 381두의 시험축군에 대하여 수집하였으며, 총 1,061두(목표:1,000두)에 대한 유전체전당의 단일염기다형(SNP)를 분석하였다.

년도	DNA sample				SNP 분석			비고
	모돈	용돈	자돈	계	모돈	용돈	계	
2010	245	21		266	219	21	240	PorcineSNP60 v1 Beadchip
2011	207	29		236	224	29	253	PorcineSNP60 v1 Beadchip
2012	200	22		222	172	15	187	PorcineSNP60 v1 Beadchip
2013	144	9	70	223	144	0	144	PorcineSNP60 v2 Beadchip
2014	341	161		501	214	26	237	PorcineSNP60 v2 Beadchip
계	1,137	242	70	1,448	973	91	1,061	

- 2종의 60K chip에 대해 품질평가를 통해 1차적으로 불필요한 마커를 제거하였다. 62,621개의 SNP 마커(PorcineSNP60 V1) 및 64,232개 SNP 마커(PorcineSNP60 V2)중에서 서로 중첩되는 61,177개의 마커를 선별하였다. 이후 상기 선별된 마커 중 염색체 정보 또는 위치(거리) 정보가 없는 마커 12,577개를 제거하였고, 수소의 경우 X 염색체를 반수체 형태로 보유하기 때문에 유전평가의 바이어스 발생을 줄이기 위해 X 염색체 내 마커 1,320개를 제거하였다. 또한, 각 마커별 2개의 대립유전자에 대하여 빈도를 조사하고 소수 유전자 빈도가 5% 미만인 13,847개 마커를 제거한 후 총 33,433개의 SNP 마커를 선별하였다.
- 이후 현행 돼지의 육종가 추정방법인 BLUP방법과 ssGBLUP 방법을 통해 번식형질에 대한 육종가를 추정하여 육종가의 정확도를 계산한 결과 개체 자신의 표현형이 없는 후보 집단(Candidate population)에서 임신기간 0.57, 생존 산자수 0.56 총산자수(미라제외) 0.55 총산자수(미라포함) 0.55의 육종가 정확도가 추정으로 약 10-11%의 정확도가 향상되었다.
- 또한 60K 유전체 정보에서 연관불평형의 값을 이용하여 Tag SNP를 선별하여 저밀도(6K) SNP를 구성하고, 이들 정보를 이용하여 60K chip으로 Imputation하였을 때 97%의 정확도로 추정되었으며, 6K chip을 이용한 유전평가에서 60K chip과 유사한 정확도의 수치를 보였다.
- 따라서 가축개량을 위한 돼지의 육종가 추정시 유전체자료의 활용은 육종가의 정확도 향상에 도움을 줄 수 있을 것이라 판단되며, 표현형을 가지는 개체의 육종가 정확도가 70% 수준임을 감안할 때, 향후 유전체 선별의 정확도가 70% 이상이 되도록 지속적으로 참조집단의 유전체 정보를 추가할 필요가 있다고 사료되었다.

5. 유전체 정보 활용 종돈의 육종체계 및 산업화 방안

- 돼지의 유전체 정보의 활용방안은 육종가추정치의 정확도 향상(11%이상)으로 인한 종돈의 개량효율과, 그것으로 인해 발생하는 경제적 이윤에 대한 측면으로 유전체선발 및 저밀도(6K) chip의 활용방안을 제시하였다.
- 돼지의 경우 산자수 등 번식형질의 유전력이 매우 낮으며, 선발간격이 길고, 근교퇴화에 민감하다는 요건을 가지고 있어 개량이 매우 어렵다. 또한 개체검정이 불가능하기 때문에 선조 및 방계 친척의 혈연정보에 의한 유전평가에 의존하고 있기 때문에 평가의 정확도가 비교적 낮다. 따라서 유전체 육종가의 추정치를 선발에 적용시키면 육종가 추정치의 정확도 향상(0.45→0.56)으로 인한 유전적 개량량 증가효과를 얻을 수 있을 것으로 보여 지며, 이는 기존방법에 비해 약 295%의 개량효율이 있는 것으로 추정되었다. 이를 산업적 측면에 따른 경제적 가치로 계산해 보았을 때, 연간 약 100억원의 추가이익의 효과가 있을 것으로 보인다.
- 현재 SNP chip의 가격이 지속적으로 낮아지고 있으며, Imputation 기술이 발달함에 따라 6K SNP chip을 두당 55천원 정도에 이용이 가능할 것으로 보인다. 또한 유전체선발의 적용은 혈통 자료를 이용하여 육종가를 추정할 경우 동복개체의 육종가가 모두 같은 값으로 추정되는 문제를 해결 할 수 있다. 따라서 GGP 종돈장의 경우 자돈 단계에서 6K chip을 이용한 유전체 선발을 적용하는 것이 효율적일 것으로 판단된다.

V. 연구성과 및 성과활용 계획

- 한우의 SNP자료 확보(50K이상 1,679두) - 유전체 정보 활용 종축선발을 위한 정보로 활용계획(향후 타 사업과 연결하여 자료수집을 통한 선발의 정확도 70%이상 목표)
- 한우의 SNP 정보를 활용, 연관불평형(LD)에 근거한 Selected SNP(약 8K)정보를 활용하여 육종가 추정시 높은 추정의 정확도(MS=0.53)를 얻음. Selected SNP을 이용한 chip제작과 이를 이용하여 한우 암송아지의 유전체 육종가 추정 후 농가 송아지 선발에 활용(현재 초음파 분석사업을 대체) - 기술이전 후 산업화 실시
- 또한 Selected SNP Chip을 활용하여 당대검정 대상우(출생 신생송아지)들의 육종가를 추정하고 1차 선발을 실시함으로써 선발효율의 증대 - 기술이전 후 산업화 실시(예: 농협중앙회 등)
- 종돈의 SNP자료 확보(60K이상 1,061두) - 유전체 정보 활용 종축선발을 위한 정보로 활용계획(향후 선발의 정확도 70%이상 목표)
- 종돈의 SNP정보를 활용, 연관불평형(LD)에 근거한 Selected SNP(약 6K)정보를 활용하여 육종가 추정시 높은 추정의 정확도(생존산자수=0.55)를 얻음. Selected SNP을 이용한 chip제작과 이를 이용하여 검정대상 자돈의 유전체 육종가 추정 후 1단계 검정대상돈 선발로 선발의 정확도 및 선발강도 향상으로 개량효율 증대 - 기술이전 후 산업화 실시(예: 선진한마을)

SUMMARY

I. Title

Development of individual genomic breeding value estimation and genetic improvement system

II. Objective and Necessity of Research and Development

- There are few cases applying genomic researches in livestock industry. Recently, new technology called 'Genomic Selection' is introduced in development of economic traits by advanced countries. Therefore, the improvement of the domestic breeding livestock using genomic information is highly required.
- The objective of this research is development of selection strategy for the young animals using genomic information to select the sire in early lifetime of livestock candidates. The final goal is to decrease the generation interval and economical expense for growing and to increase efficiency of breeding improvement
- In addition, applying genomic selection using SNP marker, the genetic composition of population is consistently changed by generation. Changing of reference population such as other breed, the accuracy of genomic selection will be decreased. Moreover, gradual update of genomic data and genetic evaluation is required to apply our source of livestock. Finally, GBLUP is essential research area for rapid improvement of Hanwoo, dairy cattle and pig.

III. Contents and Range of Research and Development

1. Contents

- The estimation of breeding value using genomic and pedigree information in Korean beef cattle (Hanwoo) and landrace pig
- Development of genetic evaluation using genomic information in Korean beef cattle (Hanwoo) and landrace pig (reference population)
- Estimation of accuracy of candidate sire in early life time (candidate population)
- Development of genetic improvement system applying genomic information in Korean beef cattle (Hanwoo) and landrace pig

2. Ranges

- Investigation of the genomic and carcass traits' information in Korean beef cattle (Hanwoo) (1,500 animals/ 5 years)
- Investigation of the genomic and reproduction traits' information in Korean landrace pig (1,000 animals/ 5 years)
- Development of genetic evaluation system using genomic information and simulation

study for accuracy of genomic selection

- Comparison of genomic estimated breeding value of carcass traits using genomic information in Korean beef cattle (Hanwoo)
- Comparison of genomic estimated breeding value of reproduction traits using genomic information in Korean landrace pig
- Development of genomic evaluation for young candidate sire of carcass traits using genomic information in Korean beef cattle (Hanwoo)
- Development of genomic evaluation for young candidate sire of reproduction traits using genomic information in Korean landrace pig
- Development of selection system of sire using genomic information in Korean beef cattle (Hanwoo)
- Development of selection system of sire using genomic information in Korean landrace pig

IV. Results

1. Comparison of accuracy of genomic selection using simulated data

- The components to affect genomic estimated breeding value (GEBV) are density of QTL, heritability, marker density, size of reference population and etc. Here, we applied multiple methods of genomic selection to simulated data with 16 combination of above component with 20 replication.
- As a result of study, density of marker and size of reference population were suggested as essential component to affect the accuracy of genomic selection. Especially, 13% difference of accuracy of genomic selection between 400 and 2000 of reference population was reported.
- Therefore, this result suggested that consistent accumulation of genomic information for reference population is most important factor to increase the accuracy of genomic selection.
- The GBLUP using numerous marker information showed higher accuracy than traditional BLUP using pedigree information. Especially, the accuracy of genomic prediction in young animal without phenotype but genotype will facilitate the selection of early lifetime.

2. Retainment of the genomic and accuracy of genomic selection for carcass traits' information in Korean beef cattle (Hanwoo)

- A total 1,887 animals including Hanwoo steers and candidate sire for five years were used in research of genomic selection. The blood sample, carcass trait and pedigree information of these animal was retained.
- The SNP genotyping using Illumina BovineSNP50 v1 and v2 for 959 animals

and Illumina Bovine HD Beadchip for 720 animals was performed, thereby total 1,679 animals (1,500 animal planned).

Year	Numer of genotyped animal			DNA Analysis Kit
	Bulls	Steers	Total	
2010	35	253	288	BovineSNP50 V1 Beadchip
2011	39	297	336	BovineHD Beadchip(700K)
2012	37	347	384	BovineHD Beadchip(700K)
2013	35	301	336	BovineSNP50 V2 Beadchip
2014	36	299	335	BovineSNP50 V2 Beadchip
Total	182	1,497	1,679	

○ The quality control and outlier checking for 50K and 700K SNP chip were performed for GEBV calculation.

① The 50K SNPs were imputed to 700K including 735,239 SNPs. Then, quality control (QC) was performed with exclusion threshold which was monomorphic, minor allele frequency < 5% and missing rate > 10%. After QC, 573,758 SNPs were remained. The accuracy of GEBV of carcass traits using ssGBLUP was increased 4-6% than using traditional BLUP for unphenotyped animals.

② Common 30,748 SNPs between 50K and 700K chip after QC, the accuracy of GEBV of carcass traits using ssGBLUP was increased 5-6% than using traditional BLUP for unphenotyped animals.

○ To validate the efficiency of low density chip, 8K SNPs using tag SNPs based on LD is used to estimate the accuracy of genomic selection. the accuracies of GEBV of carcass traits using ssGBLUP were 0.50, 0.52 and 0.53 for eye muscle area, backfat thickness and marbling score, respectively.

○ Therefore, GEBV using genomic selection for genetic improvement can help to increase the accuracy of selection. To improve the accuracy of genomic selection at 70% which is accuracy of traditional EBV with phenotype, consistent update of genomic information is required.

3. Breeding system and industrial application applying genomic selection in Hanwoo

○ Genomic selection using genomic information could be applied in early selection for sire and cow .

○ In sire selection, first selection (1,000/2,000 animals) based on GEBV of carcass traits will have the additional profit of 6,400 million per year. And additional update of genomic and phenotypic information for calf as 10,000 animal will increase the accuracy of GEBV as 70%.

○ In dam population, genomic selection can be applied to early selection in heifer calf

and paternity test. Using 8K SNP chip in hanwoo, over 100,000 animal/year would be analyzed in 80,000 won/animal. Moreover, in current genetic improvement system, GEBV and genomic information could replace the ultrasound trait and progeny test, respectively. And application of this system will have genetic improvement of the 2,700 million won per year.

4. Retainment of the genomic and accuracy of genomic selection for reproduction traits' information in Korean Landrace

- A total 1,448 animals including Landrace sire for five years were used in research of genomic selection. The blood sample, reproduction trait and pedigree information of these animal was retained.
- The SNP genotyping using Illumina PorcineSNP50 v1 and v2 for 680 and 381 animals was performed, thereby total 1,061 animals (1,000 animal planned).

Year	DNA sample				SNP Gentyed sample			DNA Analysis Kit
	Sows	Boar	Piglet	Total	Sows	Boar	Total	
2010	245	21		266	219	21	240	PorcineSNP60 v1 Beadchip
2011	207	29		236	224	29	253	PorcineSNP60 v1 Beadchip
2012	200	22		222	172	15	187	PorcineSNP60 v1 Beadchip
2013	144	9	70	223	144	0	144	PorcineSNP60 v2 Beadchip
2014	341	161		501	214	26	237	PorcineSNP60 v2 Beadchip
계	1,137	242	70	1,448	973	91	1,061	

- The quality control and outlier checking for 60K v1 and v2 SNP chip were performed for GEBV calculation.
- Using common 33,433 SNPs between 50K v1 and v2 chip after QC, the accuracy of GEBV of carcass traits using ssGBLUP was increased 5-6% than using traditional BLUP for unphenotyped animals.
- To validate the efficiency of low density chip, 6K SNPs using tag SNPs based on LD is used to estimate the accuracy of genomic selection. the accuracies of GEBV of reproduction traits using ssGBLUP were 0.57, 0.56, 0.55 and 0.55 for gestational length(GL), total number of born (TNB), total number of born alive excluding mummy (NBA1) and number of born alive excluding mummy (NBA1), respectively. The range of increase of accuracy of GEBV than EBV was 10-11%.
- Moreover, correlation of GEBV between 60K and 6K SNP chip was 98% and the level of accuracy among chips was similar.
- Therefore, GEBV using genomic selection for genetic improvement can help to increase the accuracy of selection. To improve the accuracy of genomic selection at 70% which is accuracy of traditional EBV with phenotype, consistent update of genomic information is required.

5. Breeding system and industrial application applying genomic selection in pig
- Genomic selection increase the efficiency of genetic improvement resulted from increased accuracy of GEBV(>11%) and can be applied to breeding industry by low density (6K) SNP chip.
 - In reproduction trait, genetic improvement is very hard, which is caused by low heritability, long selection interval, sensitive inbreeding depression. And accuracy of genetic evaluation is low since genetic evaluation using information of relative and ancestor. Therefore, increased accuracy of selection (0.45→0.56) will make increased genetic with 259% of breeding efficiency comparing current system. In economic aspect, application of genomic selection will have additional profit of 10 billion won/year.
 - Decreasing the price of SNP chip and developing imputation technology, 6K SNP chip will be available 55,000 won/animal. And genomic selection can solve the problem of same EBV in equal uterine offsprings. Therefore, genomic selection using 6K SNP chip can be applied to piglet in GGP.

V. Research Outcome and Utilization Plan

- Annual update of SNP data of Hanwoo to achieve the 70% accuracy of genomic selection (current: 50K with 1679 animals)
- Application of selected SNP chip (8K) based on LD for estimation of GEBV (accuracy of MS = 0.53) and usage for the selection system of cow in local farm (substitution to current ultrasound method) - technology transfer and industrial application
- Using selected SNP chip apply to estimate GEBV for calf and resulted in improvement of selection efficiency at 1st stage selection. - technology transfer and industrial application (eg. Nonghyup)
- Annual update of SNP data of pig to achieve the 70% accuracy of genomic selection (current: 60K with 1061 animals)
- Application of selected SNP chip (6K) based on LD for estimation of GEBV (accuracy of TNB = 0.55), customization of selected SNP chip and improvement of selection efficiency at 1st stage selection for piglet by increase in selection accuracy and intensity - technology transfer and industrial application (eg. Sunjin Co., Ltd.)

CONTENTS

Chapter 1 Summary of Research Project

Clause 1 Necessity of Research

Clause 2 Objective, Contents, and Ranges of Research

Chapter 2 Outlook of Technology Development

Clause 1 International Status of Technology Development

1. Background of Genomic Selection
2. Livestock Improvement using Genomic Information
3. International Application of Genomic Selection Technique

Clause 2 Domestic Status of Technology Development

1. Domestic Application of Genomic Selection Technique by livestock species

Chapter 3 Contents and Results of Research

Clause 1 Comparison of Accuracy and Method of Genomic Selection using Simulation Analysis

1. Comparison of accuracy of genomic selection for methods and influencing factors using simulation data
2. Comparison of accuracy of genomic selection in simulated populations for Hanwoo

Clause 2 Application of genomic information for genetic improvement in Hanwoo

I. Material and Methods

1. Genomic, phenotypic carcass and pedigree data collection of candidate sire and offsprings in Hanwoo
 - 가. Blood sampling and extraction of genomic DNA
 - 나. SNP genotyping
 - 다. Genomic data collection
 - 라. Phenotypic data
 - 마. Pedigree data
2. Estimation method of EBV and GEBV in Hanwoo
 - 가. Genomic data quality control
 - 나. Inspection of phenotype and pedigree data
 - 다. Construction of genomic relationship matrix

- 라. Estimation of genetic parameter and breeding value, accuracy comparison
- 3. Accuracy comparison of genomic selection between high (HD) and low density (LD) SNP chip
 - 가. Imputation to generate high density SNP data
 - 나. Generation of low density SNP data using tag SNP

II. Results and Discussion

- 4. Estimation method of EBV and GEBV in Hanwoo
 - 가. Statistical analysis of pedigree and phenotype data, environmental factor analysis
 - 나. Estimation of genetic parameter in carcass trait
 - 다. SNP quality control
 - 라. Quality control using genomic relationship matrix
 - 마. Comparison of accuracy of GEBV according genomic data
- 5. Accuracy comparison of genomic selection between high (HD) and low density (LD) SNP chip
 - 가. GEBV analysis using high density SNP data
 - 나. GEBV analysis using low density SNP data

Clause 3 Application of genomic information for genetic improvement in breeding stock of Korean Ladrace pig

I. Material and Methods

- 1. Genomic data collection and genotyping analysis in Korean Ladrace pig
 - 가. Blood sampling and extraction of genomic DNA
 - 나. Genomic data collection
 - 다. Phenotypic data
 - 라. Pedigree data
- 2. Estimation method of EBV and GEBV in pig
 - 가. Genomic data quality control
 - 나. Inspection of phenotype and pedigree data
 - 다. Construction of genomic relationship matrix
 - 라. Estimation of genetic parameter and breeding value, accuracy comparison
- 3. Selection of low density SNP and comparison of accuracy of GEBV

II. Results and Discussion

- 4. Estimation method of EBV and GEBV in pig
 - 가. Statistical analysis of pedigree and phenotype data, environmental factor analysis

- 나. Estimation of genetic parameter in reproduction trait
- 다. SNP quality control
- 라. Quality control using genomic relationship matrix
- 마. Comparison of accuracy of GEBV according genomic data
- 5. Accuracy comparison of genomic selection and low density (LD) SNP chip
 - 가. SNP selection for low density (6K) SNP chip
 - 나. Accuracy comparison of GEBV between 6K and 60K SNP chip

Clause 4 Economic effect and industrial application of genomic selection in Hanwoo and pig

1. Economic effect of genomic selection in Hanwoo
 - 가. Economic analysis of selection in proven bulls
 - 나. Economic value of selection in proven bulls using genomic information
 - 다. Economic value of selection in cow using genomic informatio
2. Economic effect and industrial application of genomic selection in pig
3. Application of GEBV in young animal using low density SNP effect

Chapter 4 Level of Achievement and Contribution to Related Area

Chapter 5 Research Outcomes and Plan to Utilization

Chapter 6 International Science and Technology Information during Research Progress

Chapter 7 Research Facilities and Equipments

Chapter 8 References

목 차

개체 유전체 육종가 추정 시스템(GBLUP) 및 개량체계 개발

제 1 장 연구개발과제의 개요

- 제 1 절 연구의 필요성 및 목표
- 제 2 절 연구 목적, 내용 및 범위

제 2 장 국내외 기술개발 현황

- 제 1 절 해외 기술개발 현황
- 제 2 절 국내 기술개발 현황

제 3 장 연구개발수행 내용 및 결과

제 1 절 모의자료 분석을 통한 유전체선발방법의 정확도 및 방법론 검증

- 1. 유전체 선발기법 및 영향요인에 따른 정확도 비교
- 2. 한우집단의 유전체 육종가 추정 모의시험

제 2 절 한우 종축에 대한 유전체 정보 활용 분석기법 및 개량방안 연구

I. 재료 및 방법

- 1. 한우 후보종모우와 후대검정 대상우의 유전체, 도체형질 및 혈통 데이터 수집
 - 가. 유전체 자료 형성을 위한 한우 시험축 채혈 및 gDNA 추출
 - 나. SNP chip 분석 및 Genotype 확보
 - 다. 유전체자료 수집
 - 라. 표현형자료
 - 마. 혈통자료
- 2. 한우의 기존(EBV) 및 유전체(GEBV) 육종가 추정 방법
 - 가. 유전체 데이터 품질평가
 - 나. 표현형 및 혈통 데이터 점검
 - 다. 유전체 혈연계수 행렬 구성
 - 라. 유전모수, 육종가 추정 및 정확도 비교
- 3. 고밀도(HD) 및 저밀도(LD) chip에 대한 유전체 육종가 정확도 비교
 - 가. 대치(Imputation)을 이용한 고밀도 유전체 데이터 생성
 - 나. Tag SNP을 이용한 저밀도 유전체 데이터 생성

II. 결과 및 고찰

4. 유전체 정보를 이용한 육종가(GEBV) 추정 및 기존 육종가(EBV)와 비교
가. 한우 혈통 및 검정자료에 대한 기초통계분석 및 환경요인 분석
나. 한우 도체형질에 대한 유전모수 추정
다. SNP마커의 품질평가
라. 유전체 혈연계수 행렬을 통한 품질평가
마. 유전체 정보 유무에 따른 추정 육종가 정확도 추정 및 비교
5. 고밀도(HD) 및 저밀도(LD)칩에 대한 유전체 육종가 정확도 비교
가. HD(700K) chip을 이용한 유전체 육종가 분석
나. 저밀도(8K) chip을 이용한 유전체 육종가 분석

제 3 절 돼지 종축에 대한 유전체 정보 활용 분석기법 및 개량방안 연구

I. 재료 및 방법

1. 유전체(SNP) 정보 및 검정자료 수집
가. 채혈 및 genomic DNA 추출
나. 유전체자료 수집
다. 표현형자료
라. 혈통자료
2. 돼지의 기존(EBV) 및 유전체(GEBV) 육종가 추정 방법
가. 유전체 데이터 품질평가
나. 표현형 및 혈통 데이터 검정
다. 유전체 혈연계수 행렬 구성
라. 유전모수, 육종가 추정 및 정확도 비교
3. 저밀도(6K) chip 선별 방법 및 유전체 육종가 정확도 비교

II. 결과 및 고찰

4. 유전체 육종가(GEBV) 및 혈통 육종가(EBV) 비교
가. 돼지 혈통 및 검정자료에 대한 기초통계분석 및 환경요인 분석
나. 돼지 번식형질에 대한 유전모수 추정
다. SNP마커의 품질평가
라. 유전체 혈연계수 행렬
마. 유전체 정보 유무에 따른 추정 육종가 정확도 추정 및 비교
5. 저밀도(6K) chip 선별 및 유전체 육종가 정확도 비교
가. 저밀도칩 분석을 위한 SNP의 선별
나. 저밀도(6K) chip과 60K chip 간의 유전체 육종가 정확도 비교

제 4절 한우 및 돼지 종축에 대한 유전체 정보 활용 산업화 방안

1. 한우 유전체 선발에 대한 경제적 효과
 - 가. 한우 보증씨수소 선발을 통한 경제 가치 분석
 - 나. 한우 유전체 정보 활용 보증씨수소 선발을 통한 경제 가치 분석
 - 다. 한우 유전체 정보 활용 암소 선발을 통한 경제 가치 분석
2. 돼지 유전체 선발에 대한 경제적 효과 및 산업화 방안
3. 추정된 마커효과를 이용한 어린 후보축의 저밀도 SNP 기반 유전체 추정 육종가 활용방안

제 4 장 목표달성도 및 관련분야에의 기여도

제 5 장 연구개발 성과 및 성과활용 계획

제 6 장 연구개발과정에서 수집한 해외과학기술정보

제 7 장 연구시설·장비 현황

제 8 장 참고문헌

제 1 장 연구개발과제의 개요

제 1 절 연구개발의 필요성 및 목표

1. 연구개발의 필요성

가. 산업적 측면

- (1) 축산물에 대한 해외 시장 개방 이후 국내 축산산업은 경쟁력을 갖추기 위하여 생산의 효율성 증진에 대한 필요성을 인식하고 있으며, 이에 따라 한시적이지 않고 그 효과가 미래까지 지속되는 중축산업의 중요도가 증가
- (2) 유전체에 대하여 현재까지 많은 연구들이 시행되었음에도 불구하고 아직까지는 유전체 연구들을 활용하여 산업적으로 사용되는 사례가 매우 적었으나 최근 들어 선진국을 중심으로 유전체 선발이라는 새로운 기술이 도입되어 경제적 형질 개량에 유전체 정보를 이용하는 사례가 있음

나. 학술적 측면

- (1) 현재 일반적으로 활용하는 유전능력분석(BLUP) 시스템은 혈통기록에 의존하고 있어 능력검정을 받지 않은 어린 가축에 대한 유전능력 추정치의 정확도는 매우 낮은 한계가 있음(예: 한우의 도축형질 - 40% 전후, 돼지의 산자수 - 35% 전후)
- (2) 또한, 능력기록이 없는 개체는 단순히 부와 모의 평균능력을 이용하기 때문에 감수분열 단계에서 발생하는 멘델리안 샘플링(mendelian sampling) 효과를 계산하지 못하고 동일한 부모에서 태어난 개체는 동일한 유전능력으로 추정하게 되기 때문에 선발한 개체들 간의 혈연계수가 높아지게 되어 유전적 다양성을 유지하는데 저해요인으로 작용하고 있음
- (3) 유럽 및 미국에서는 개체가 가지고 있는 모든 염색체에 골고루 흩어져 있는 수만~수십만 개의 단일염기다형마커(Single Nucleotide Polymorphism; SNP Marker)를 동시에 분석하여(micro-array) 나타난 마커 유전자형과 능력검정을 통하여 수집한 능력정보를 동시에 분석하는 유전체유전능력평가기법(genomic BLUP; GBLUP)을 2000년대 초부터 개발하기 시작하여 현재 젓소, 돼지 등의 축종에서 실용화하고 있는 상태임
 - (가) 일부 학자들은 2008년 이후 유전체 선발의 실용화가 시작되면서 유전체 추정 육종가(Genomic Estimated Breeding value; GEBV)이론이 밝혀진 이후에 동물육종의 혁명이라고까지 함
 - (나) 미국 및 캐나다는 젓소에 대한 50K SNP chip을 활용하여 유전체정보를 수집하고 기존의 능력검정 선발체계에 통합·운영함으로써 이전보다 향상된 선발의 정확도로써 보증씨수소를 선발하고 있음(2009년부터 선발에 활용)
- (4) 축산선진국에서 연구개발한 방법은 전통적인 선발방법인 경제적 능력과 혈연관계에 대한 유전평가방법(BLUP)에 다량의 유전체 정보를 추가하여 평가하는 방법으로 기존방법이 능력기록이 없는 어린 가축에 대하여 단순 혈통 자료에 의존하여 멘델리안 샘플링을 고려하지 못하는 문제를 해결함은 물론 유전체 정보만으로도 10여두 이상의 자손의 능력을 검정한 것과 비슷한 정확도를 가지게 되므로 어린 가축을 조기에 선발하여 이를 능력검정 함

으로써 우수 가축 선발의 선발가능성과 선발의 정확도를 획기적으로 개선할 수 있음

- (5) 한편, SNP마커정보를 활용한 선발을 실시할 때, 집단의 유전적 구성이 지속적으로 변화하거나 집단이 다르게 되면 정확성이 현저히 떨어지게 되므로 우리나라의 한우, 젓소, 돼지 등에 적용하기 위해서는 우리나라 종축집단의 마커분석, 능력검정을 지속적으로 실시할 필요성이 있기 때문에 한우, 젓소 및 종돈 등의 신속한 개량을 위해서는 GBLUP체계 구축을 위한 연구에 적극적인 투자가 필요함

제 2 절 연구목적, 내용 및 범위

3. 연구개발의 목적

- 가. 한우 참조집단의 유전체 마커정보와 능력검정 정보를 활용하여 생산송아지의 유전능력을 조기에 예측(GEBV, 정확도 70%이상)
- 나. 송아지의 도체형질들에 대한 GEBV에 근거하여 당대검정 대상우를 선발하여 개량의 효율성 향상
- 다. 유전체 정보를 활용한 한우의 국가단위 육종 개량체계의 재설정으로 한우개량의 효율성 향상
- 라. 종돈의 유전체 정보와 능력검정정보를 활용하여 후보돈의 유전능력을 조기에 예측
- 마. 후보돈의 조기 선발을 통한 개량효율성 향상
- 바. GBLUP을 활용한 국산 순종돈의 개발 및 유전능력 향상으로 외화 절감 및 양돈 산업의 경쟁력 강화
- 사. GBLUP을 활용한 국산 종돈 개량체계 제시

4. 연구개발의 내용 및 범위

- 가. 모의시험을 통한 유전체 선발의 효율성 검증 및 검정체계의 효율성 검증
- 나. 한우 후보종모우와 후대검정우의 유전체 50K SNP Markers정보 수집
- 다. 한우 후대검정우의 SNP Marker정보를 이용한 육질에 대한 GBLUP추정과 혈통정보를 이용한 BLUP추정치의 비교
- 라. GBLUP 추정치의 정확도 검증
- 마. 대량의 SNP Marker들에 대한 유전적 효과 추정
- 바. 종모우 및 이들의 자축에 대한 SNP Marker에 대한 유전적 효과 추정치를 활용한 자축의 GBLUP추정
- 사. 자축의 GBLUP추정치에 의한 당대검정 대상우 선발 및 당대검정 실시
- 아. 당대검정 성적, 혈통정보, 유전체 정보를 활용한 GBLUP과 BLUP추정치 비교로 후보종모우 선발 정확도 검증
- 자. 후대검정을 통한 보증종모우 선발 정확도와 GBLUP에 의한 보증종모우 선발 정확도 검증
- 차. 돼지 GGP 종돈의 유전체 정보 및 검정성적(번식형질) 수집
- 카. 검정돈의 GBLUP과 BLUP간의 육종가 추정치 비교 및 GBLUP의 효율성 검증
- 타. GBLUP에 의한 종돈 선발체계 효율성 검증 및 신규 종돈 개량체계 제시

제 2 장 국내외 기술개발 현황

제 1 절 해외 기술개발 현황

1. 유전체 선발기술 배경

- 가. 유전체 선발은 Meuwissen 등(2001)에 의해 처음으로 제안되었고 Simulation으로 조밀한 간격의 Marker Mapping을 통해 개체의 유전능력을 예측하였다. 마커 간 간격이 조밀할 수록 마커와 양적형질유전자좌위(QTL)간 연관불평형이 높게 되며, 이 때 인접한 마커를 통해 QTL의 효과를 간접적으로 예측하는 방법을 이용할 수 있다고 보고함.
- 나. Gengler 등(2007)이 유전체 정보가 없는 개체에 대한 유전체 정보를 예측할 수 있는 알고리즘을 제안하였고, VanRaden 등(2008)이 유전체 혈연계수행렬 계산 알고리즘 및 유전체 육종가 계산방법을 기술하였다. Misztal 등(2009)이 기존혈통정보와 유전체정보 결합 알고리즘을 발표하였고 현재는 Liu, Goddard 등(2014)이 SSSmodel 모델을 개발하여 Single-setp genomic model 내에서 SNP 효과를 직접 추정 할 수 있는 방법을 제시하였고 독일, 캐나다 등에서 국가단위 유전평가에 적용시키고 있음.

2. 유전체 정보를 이용한 가축 개량기술

- 가. 덴마크 듀록 : 덴마크의 듀록 품종에서 일당증체량 및 사료 요구율 형질에 대하여 기존의 혈통정보와 검정정보를 이용하여 추정한 BLUP방법과 유전체 정보를 갖는 개체에 대해서만 추정한 GBLUP, 그리고 모든 정보(유전체 정보 보유 및 미 보유 개체 포함)를 하나의 모델에 적합하여 분석한 Single step GBLUP방법에 대한 정확도 및 회귀분석을 실시한 결과, 유전체 정보를 이용하여 분석한 것이 혈통정보만 이용한 것에 비하여 정확도가 높다고 보고하였으며, Single step GBLUP을 이용하였을 때 유전체 정보를 갖지 않는 개체에 대해 BLUP추정방법에 비해 정확도를 더 향상시킬 수 있다고 보고함 (Christensen 등, 2012).

3. 유전체 선발기술의 해외 적용현황

- 가. 캐나다 돼지 개량 프로그램(Canadian Centre for Swine Improvement) : 60K SNP Chip을 이용하여 유전체 정보를 수집하고 검정데이터와 유전체정보를 모두 갖고 있는 개체의 집단(참조집단)에서 각 SNP의 효과를 추정하고 이를 검정정보 없이 유전체정보만 갖는 개체의 집단(검정집단)에 적용하여 기존 BLUP방법과 유전체 선발방법 간 비교하였다. 그 결과, 등지방두께에서 유전체 정보 없이 육종가를 추정하였을 때 0.35의 정확도를 보였으며, 유전체 정보를 이용하였을 때 0.39의 정확도로 정확도가 4%증가하였을 때 유전적 개량량이 11%증가의 효과가 있다고 보고하였으며, 총산자수는 0.55와 0.66으로 유전적 개량량이 20%증가하였다고 보고하였음. 초음파 등심단면적의 경우 유전체 정보를 이용함으로써 67%의 유전적 개량량 향상효과가 있다고 보고함 (CCSI, 2013).

- 나. PIC 중돈회사는 후보돈에 대하여 300 - 400 개의 마커와 부모의 60K SNP Chip을 통해 Imputation하는 기술을 보유하고 있으며 이러한 방법은 유전적 개량량을 가속화 시킬 수 있다고 보고하였다. 총산자수에 대해 부모의 60K SNP chip정보만 가지고 후보돈에 대해 유전체 선발을 하였을 때 11%의 정확도가 향상된 반면, 후보돈 자신의 60K SNP chip정보를 포함하여 분석하였을 때 68%의 정확도를 갖는 것으로 보고함 (PIC, 2013).
- 다. Patrick Stratz등(2014)은 유전체 선발 비용을 줄이기 위한 연구를 실시하였는데, 이들의 연구결과에 따르면 독일의 Pietrain수돼지 895두에 대하여 저밀도의 SNP(384 SNP)을 Marker간의 거리를 균등하게 선정한 경우와 연관불평형(LD)의 크기에 기준하여 선정한 경우를 비교한 결과, 전자의 경우가 후자의 경우 보다 유전체 육종가추정치의 정확도가 높았다고 하였음.
- 라. 홀스타인의 경우 전 세계에 걸쳐 유전자원을 공유하고 있기 때문에 국가 간 유전능력 평가를 국제유전능력평가기구(INTERBULL)의 참여국에 한하여 실시하고 있는데 2010년 이후로 유전체 정보를 포함하여 분석하자는 분위기가 조성이 되었으며 유전체 선발에 대한 검증단계를 거쳐 현재 일부 참여국에 대해서 국제유전체유전능력평가(GMACE)를 실시하고 있다. 또한 CDN(Canadian Dairy Network)에서는 유전체정보를 활용하여 유전능력평가 결과를 농가에 제공하고 있으며, 미국(USDA), 독일(VIT)등의 국가에서도 유전체 분석결과를 서비스하고 있음.

제 2 절 국내 기술개발 현황

국내 일부 QTL관련 Marker규명에 대한 연구는 상당히 많이 진행된 반면에 국내 유전체 정보를 활용한 선발기술에 대한 연구는 아직 미약한 수준이며 Lee와 Vasco(2011)의 연구결과 등 일부 연구보고가 있음. 각 축종별 연구수행결과를 요약하면 아래와 같음.

1. 축종별 유전체 선발기술의 연구현황

가. 한우

- (1) 이등(2014a)은 한우의 도체형질과 관련이 있는 10개의 DNA 마커를 규명하였는데, 이러한 DNA 마커로 마블링 점수 정확도를 예측할 수는 없지만 마블링 점수의 정확도를 예측하는데 활용 할 수 있다는 연구를 보고하였다.
- (2) 50K 고밀도 칩을 활용하여 황갈색 한우와 칙소간의 유전적 차이를 비교 연구한 결과가 있음 (이등, 2014b).

나. 젃소

- (3) 국내 젃소의 경우 2013년 2월 이전까지 유전체선발에 관한 연구가 전무한 실정이었다가 2013년 2월부터 젃소유전체선발을 위한 참조집단축군 조성을 위한 시료확보 및 유전체자료 수집에 관한 연구가 수행 중에 있으며 3년간 약 2,000여두의 참조 축군을 조성할 예정에 있음.
- (4) 가상의 국내 젃소유전체 데이터를 생성하여 통계분석방법, 참조집단의 크기 및 유전력의 크기에 따라 유전평가의 정확도 변화에 대한 연구가 진행되었고 모의실험을 실시한 결과로써 통계분석방법 간 육종가 추정치의 정확도는 표현형정보, 혈통정보 및 유전체 정보를 모두 활용하여 개체의 유전능력을 평가하는 GBLUP이 유전체 정보 없이 분석하는 BLUP방법에 비해 높게 나타났으며, 특히 자신의 표현형 정보가 존재하지 않는 검정집단 축군에서 통계분석방법 간 육종가 추정치의 정확도에 대한 편차가 크게 나타났다. 검정형질의 유전력에 대한 크기별 육종가 정확도를 비교한 결과, BLUP과 GBLUP에서 모두 유전력이 높을수록 육종가 추정치의 정확도가 높게 나타났으며, 또한 참조집단크기별 육종가 추정치의 정확도는 크게 영향을 받는 것으로 나타났다는 연구결과가 있음(조등, 2013).

다. 돼지

- (1) 이 등(2013)은 폐쇄 종돈집단의 랜드레이스 순종돈 448두에 대하여 Porcine SNP 60K chip을 이용하여 수집된 유전체 정보를 활용하여 개체 간 혈연계수행렬을 5가지 방법(G05, GMF, GOF, GOF* 및 GN)으로 추정하여 비교한 결과, GOF 방법에 의하여 계산된 혈연계수가 기존의 혈통정보를 이용한 혈연계수와 가장 작은 편차를 갖는 것으로 추정되었고 GMF방법(소수대립유전자빈도 활용방법)에서 가장 큰 편차를 보이는 것으로 보고하였다. GOF방법에 의한 혈연계수를 이용하여 추정된 유전체육종가가 기존의 혈통정보를 이용하여 추정된 육종가와 가장 높은 상관성을 갖는 것으로 보고하였다.
- (2) 조 등(2013)은 국립축산과학원에서 사육중인 제주재래돼지와 랜드레이스의 상호 교차 교

배를 통해 생산된 417두의 F2집단을 대상으로 성장형질을 측정하고, 개체별 혈액으로부터 Illumina Porcine Snp 60K BeadChip을 이용하여 유전자형 분석을 실시하여 성장형질과 유전체전장의 단일염기다형의 유전자형 간에 연관성을 알아보았으며. 체중에 영향을 미치는 고정효과에 대해 일반선형모형을 설정하여 사전 보정을 실시하였고, 여기서 얻어진 잔차값을 이용하여 표지인자와 월령별 체중간의 연관성 분석을 실시한 결과, 연관성이 강한($p < 10^{-6}$) 표지인자가 생시(BWB), 3주령(BW3), 10주령(BW10) 및 20주령(BW20) 체중에서 각각 657개, 846개, 49개, 122개가 존재하는 것으로 보고하였다. 특히 생시 및 3주령 체중에 영향을 미치는 표지인자는 특정 염색체(SSC9)에서 고도의 통계적 유의차($p < 10^{-15}$)를 나타내는 유전자 좌위가 있는 것으로 추정되었음.

제 3 장 연구개발수행 내용 및 결과

제 1 절 모의자료 분석을 통한 유전체선발방법의 정확도 및 방법론 검증

본 연구의 초기 단계인 1차년도에 경우 한우 및 돼지 시험 축군에 대한 DNA 샘플 및 유전체 데이터가 수집되지 않았기 때문에 7회 세미나를 통한 논문 리뷰 및 유전체 선발의 타당성을 검증하고자 모의실험 연구를 실시하였다.



그림1. 동물유전체 육종연구 세미나 결과 인쇄본

1. 유전체 선발기법 및 영향요인에 따른 정확도 비교

본 연구는 표지인자 수(밀도), 유전력, 형질에 관여하는 양적형질좌위 수(밀도), 참조집단의 크기를 서로 다르게 하여 모의 자료를 형성하고 각 모의자료를 베이지안 방법을 이용한 8가지 유전체 육종가 추정방법과 기존의 BLUP 방법을 적용하여 유전체 선발의 정확도에 영향을 미치는 요인을 확인하고 각 방법론별 정확도 차이를 비교함으로써 조건에 따른 최적 방법론을 제시하여 국내에 유전체 선발방법을 도입하여 활용하고자 실시하였다.

유전체 육종가는 다양한 베이지안 이론을 이용하여 분석하였으며 (BayesA, BayesB1 (Meuwissen, 2001), BayesB, BayesBfast, BayesC, BayesCpi, BayesCscale, GBLUPgibbs(Fernando, 2010)), 분석형질 및 유전체 정보량에 따른 가정은 그림 2과 같다.

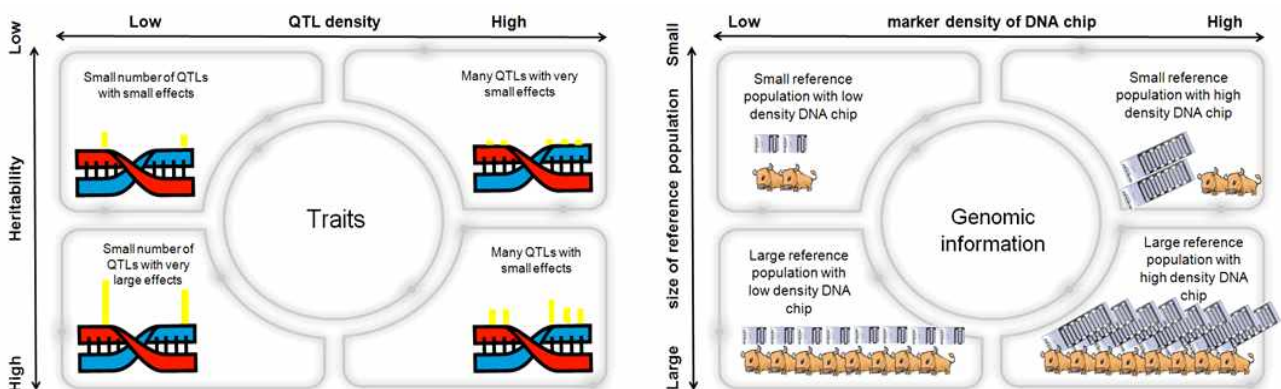


그림 2. 유전체 정보 및 표현형의 유전력에 따른 조건표

가. 재료 및 방법

(1) 유전체 모의자료 생성

분석에 사용되는 표지인자 수, 형질에 관여하는 양적형질좌위 수, 유전력 그리고 유전체 정보와 표현형 자료를 가지고 있는 개체 수를 다르게 한 16조합의 모의 유전체 자료를 생성하였고, 하여 육종가를 추정하고 참육종가와 상관을 통해 유전평가의 정확도를 계산하였다.

모의실험에 대한 신뢰도를 높이기 위하여 각 조합별로 방법론별 20회 반복실험을 실시하였으며, 이 때 생성되는 반복 모의 자료는 각 조합의 조건에 따른 임의 생성자료로 서로 조건은 같으나 형질발현에 영향을 주는 양적 형질좌위의 위치와 크기는 서로 다르게 생성하였다.

표 1. 양적형질유전자좌위수(Q), 유전력(H), 표지인자수(M) 및 참조집단크기(P)의 가정 및 모의자료

모의자료	양적형질좌위수	유전력	표지인자수	참조집단크기
	(밀도)		(밀도)	
	Q	H	M	P
	형질의 차이		유전체 정보의 차이	
Q1H1M1P1			600	400
Q1H1M1P5		0.1	(0.5cM)	2,000
Q1H1M5P1			3,000	400
Q1H1M5P5	60		(0.1cM)	2,000
Q1H5M1P1	(5cM)		600	400
Q1H5M1P5			(0.5cM)	2,000
Q1H5M5P1		0.5	3,000	400
Q1H5M5P5	(0.1cM)		2,000	
Q5H1M1P1			600	400
Q5H1M1P5		0.1	(0.5cM)	2,000
Q5H1M5P1			3,000	400
Q5H1M5P5	300		(0.1cM)	2,000
Q5H5M1P1	(1cM)		600	400
Q5H5M1P5			(0.5 cM)	2,000
Q5H5M5P1		0.5	3,000	400
Q5H5M5P5	(0.1cM)		2,000	

본 모의실험자료의 형성에서는 3쌍의 염색체만 존재한다고 가정하였으며, 각 염색체는 서로 그 길이를 다르게 하여 110cM, 100cM, 90cM으로 정의하였고, 표지인자는 균일하게 위치시켰으며, QTL은 임의로 위치시켰다. 초기의 유효집단의 수는 100두로 고정하였으며, 수컷과 암컷은 1:1의 비율로 임의 생성하였다.

Marker는 2.5×10^{-3} , QTL는 2.5×10^{-5} 의 확률로 돌연변이를 발생시켰고, 교차는 Haldane's 계산에 의해 동일 염색체 내에 두 개의 Marker간의 교차가 일어날 확률로 발생시켜 Hardy-Weinberg 평형집단($N_e = 100$)으로서 1000세대까지 성장시켰다.

표지인자의 경우 돌연변이 및 교차가 발생하면 다형성이 나타나게 되며, QTL의 경우 돌연변이가 발생하면 감마분포(gamma distribution : scale=1.66, shape=0.4)에서 임의값을

발생시켜 양수와 음수 각각 0.5의 확률로 QTL효과를 생성하였다.

1001-1002세대에서는 육종가에 임의 잔차값을 더해서 표현형자료를 생성하였으며, 잔차값은 $N(0, \sigma_e^2)$ 로부터 생성되었다. 여기서 σ_e^2 는 유전력이 h^2 일 경우 $\sigma_e^2 = \frac{1-h^2\sigma_g^2}{h^2}$ 으로 값이 얻어졌으며, 유전체자료와 같이 생성하였다. 1001-1002세대는 표현형자료와 유전체자료를 생성하였으며, 1003세대(progeny1)와 1004세대(progeny2)는 유전체자료만 생성하였고 1001~1004세대의 유전체자료 및 표현형자료를 활용하여 개체별 표지인자효과 및 육종가를 추정하였다. 모의집단생성을 위한 모수의 가정은 표 2에 제시하였다.

표 2. 모의실험에서 유전체 자료 생성을 위한 가정 및 모수

	모수(Parameter)
염색체 수	3
염색체 길이	90cM 100cM 110cM
초기 유효집단 크기	100두
분포 분할 방식	표지인자 - 균등분할 QTL - 임의분할
표지인자의 돌연변이율	2.5×10^{-3}
QTL의 돌연변이율	2.5×10^{-5}
돌연변이 효과의 분포가정	감마분포(Scale=1.66,Shape=0.4)
재조합율(Haldane's map function)	$r = (1 - e^{-2.0})/2$
1-1000세대까지 세대당 집단크기	100두
1001-1004세대까지 세대당 집단크기	400두, 2000두
교배방법	임의교배
총 세대수	1004세대
표현형과 유전체 데이터 생성	1001 - 1002세대
유전체 데이터만 생성	1003 - 1004세대

(2) 유전체 육종가 추정방법

(가) 베이저안 방법을 이용한 유전체 육종가 추정 방법론별 차이는 표3와 같으며(Fernando, 2010), BayesA, GBLUP방법의 경우 모든 표지인자에 대하여 효과가 존재하는 것으로 가정하고 BayesB, BayesB1, BayesBfast, BayesC, BayesCpi, BayesCscale 방법은 일부의 표지인자만이 효과를 갖는다고(fraction=1- π) 가정 하였다.

(나) BayesA, BayesB, BayesB1 및 BayesBfast 방법은 표지인자별로 분산을 추정하며, GBLUP과 BayesC 계열 방법은 전체 유전체 분산을 유용표지인자수로 나누어 사용한다. 따라서 GBLUP방법은 모든 표지인자를 유용표지인자로 가정하므로 전체 분산을 표지인자수로 나누어 모든 표지인자에서 매우 작고 같은 크기의 분산을 가정하였다.

표 3. 베이저안 추론에서 유전체 육종가 추정방법별 가정 및 분석방법 비교

표지인자의 분포가정

분산	모든 표지인자의 분산존재(pi=0)	일부의 표지인자에서만 분산존재(1-π)
Random-Individual variance σ_j^2	$\sigma_j^2 \sim \nu_\alpha S_{\nu_\alpha}^2 \chi_{\nu_\alpha}^{-2}$ Bayes A(B0) $(\alpha_j \sigma_j^2) \sim N(0, \sigma_j^2)$	$(\sigma_j^2 \nu_\alpha, S_\alpha^2) \sim \nu_\alpha S_\alpha^2 x_{\nu_\alpha}^{-2}$ BayesB, B1(known π) $(\alpha_j \pi, \sigma_j^2) \begin{cases} \sim N(0, \sigma_j^2) & \text{Pr}(1 - \pi), \\ = 0 & \text{Pr}(\pi) \end{cases}$ MH sampling BayesBfast(known π) sample $\delta_j = 1$ with probability 0.5 MH sampling
Random-Constant variance σ_α^2	$\sigma_\alpha^2 \sim \nu_\alpha S_{\nu_\alpha}^2 x_{\nu_\alpha}^{-2}$ Bayes C0=GBLUP	$(\sigma_\alpha^2 \nu_\alpha, S_\alpha^2) \sim \nu_\alpha S_\alpha^2 x_{\nu_\alpha}^{-2}$ BayesC(known π) $(\alpha_j \pi, \sigma_\alpha^2) \begin{cases} \sim N(0, \sigma_\alpha^2) & \text{Pr}(1 - \pi), \\ = 0 & \text{Pr}(\pi) \end{cases}$ sample $\delta_j = 1$ with probability 0.5
Random-Constant var (estimated π) σ_α^2		BayesCpi(estimated π) $(\alpha_j \pi, \sigma_\alpha^2) \begin{cases} \sim N(0, \sigma_\alpha^2) & \text{Pr}(1 - \pi), \\ = 0 & \text{Pr}(\pi) \end{cases}$ sample $\delta_j = 1$ with probability 0.5 sample π with beta distribution
Random-Constant var (estimated π, scale) σ_α^2		BayesCscale(estimated π) $(\alpha_j \pi, \sigma_\alpha^2) \begin{cases} \sim N(0, \sigma_\alpha^2) & \text{Pr}(1 - \pi), \\ = 0 & \text{Pr}(\pi) \end{cases}$ sample $\delta_j = 1$ with probability 0.5 sample π with beta distribution sample S_α^2 with gamma distribution

* B0 : BayesB with MH sampling=0, C0 : BayesC with π=0

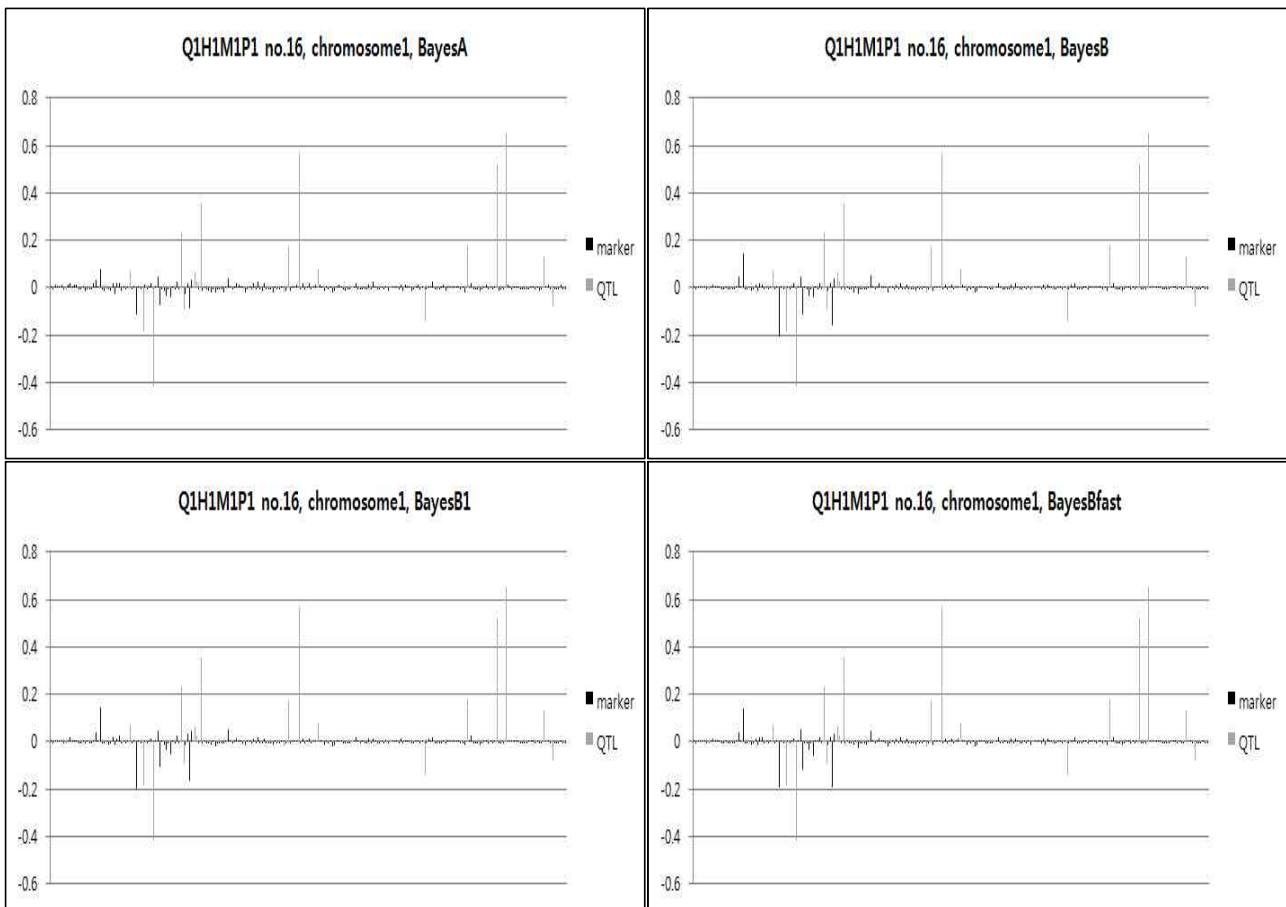
(다) BayesCpi, BayesCscale방법은 유전체 내에 존재하는 유용표지인자(effective marker)와 비유용표지인자(non-effective marker)의 비율의 계산에 사용되는 π를 추정하여 이용하며, BayesCscale 방법은 표지인자의 효과를 샘플링하기 위한 표지인자 효과분산의 scale parameter를 추정하였다.

(라) 본 연구에서 π는 0.9 scale parameter를 위한 이형집합성은 0.5를 가정하였으며 방법에 따라 π와 scale parameter는 추정하여 유전체 육종가를 계산(BayesCpi, BayesCscale)하였으며, Gibbs sampling 기법을 이용하여 각 표지인자별 효과를 추정하고 이를 개체별 유전자형에 곱한 값들의 합으로 육종가 계산하였다.

나. 결과 및 고찰

(1) 통계분석방법 및 가정에 따른 표지인자 효과

그림 8는 각 방법론별 표지인자 효과를 추정한 것으로, 양적형질 좌위 밀도(수)를 5cM(22개, 염색체1번의 경우)으로, 유전력을 0.1로, 표지인자의 수(밀도)를 0.5cM(220개, 염색체1번의 경우)으로, 참조집단의 크기를 400두로 가정한 모의실험에서 BayesA, BayesB, BayesB1, BayesBfast, BayesC, BayesCpi, BayesCscale, GBLUP방법으로 분석한 결과를 나타내었다. GBLUP을 제외한 나머지 방법에서 모두 비슷한 좌위에 효과가 있는 것으로 추정되었으나 그 크기는 서로 차이를 보이는 것을 확인할 수 있다. GBLUP방법의 경우 전체 유전분산을 표지인자의 수로 나누어 사용하므로 그 분산의 크기가 작기 때문에 각 표지인자별 효과가 다른 방법에 비하여 상대적으로 작게 나타나는 것을 확인할 수 있었다. 또한 형질의 유전적 특성인 양적형질좌위수(밀도)와 유전력, 유전체 정보량인 표지인자수(밀도)와 참조집단의 크기가 표지인자 효과 구명에 미치는 영향을 확인하기 위하여 각 요인별 가정을 다르게 한 모의실험에서 BayesCpi 방법을 이용하여 각 표지인자들의 효과를 추정한 결과, 같은 방법의 표지인자 효과 추정이라 할지라도 유전체 정보의 조건에 따라 그 형태가 다르게 나타남을 확인할 수 있었다(그림 4). 즉 양적형질유전자좌위수가 많고 유전력이 높으며 표지인자의 수 및 참조집단이 클수록 표지인자의 모수 추정치가 향상됨을 알 수 있었다.



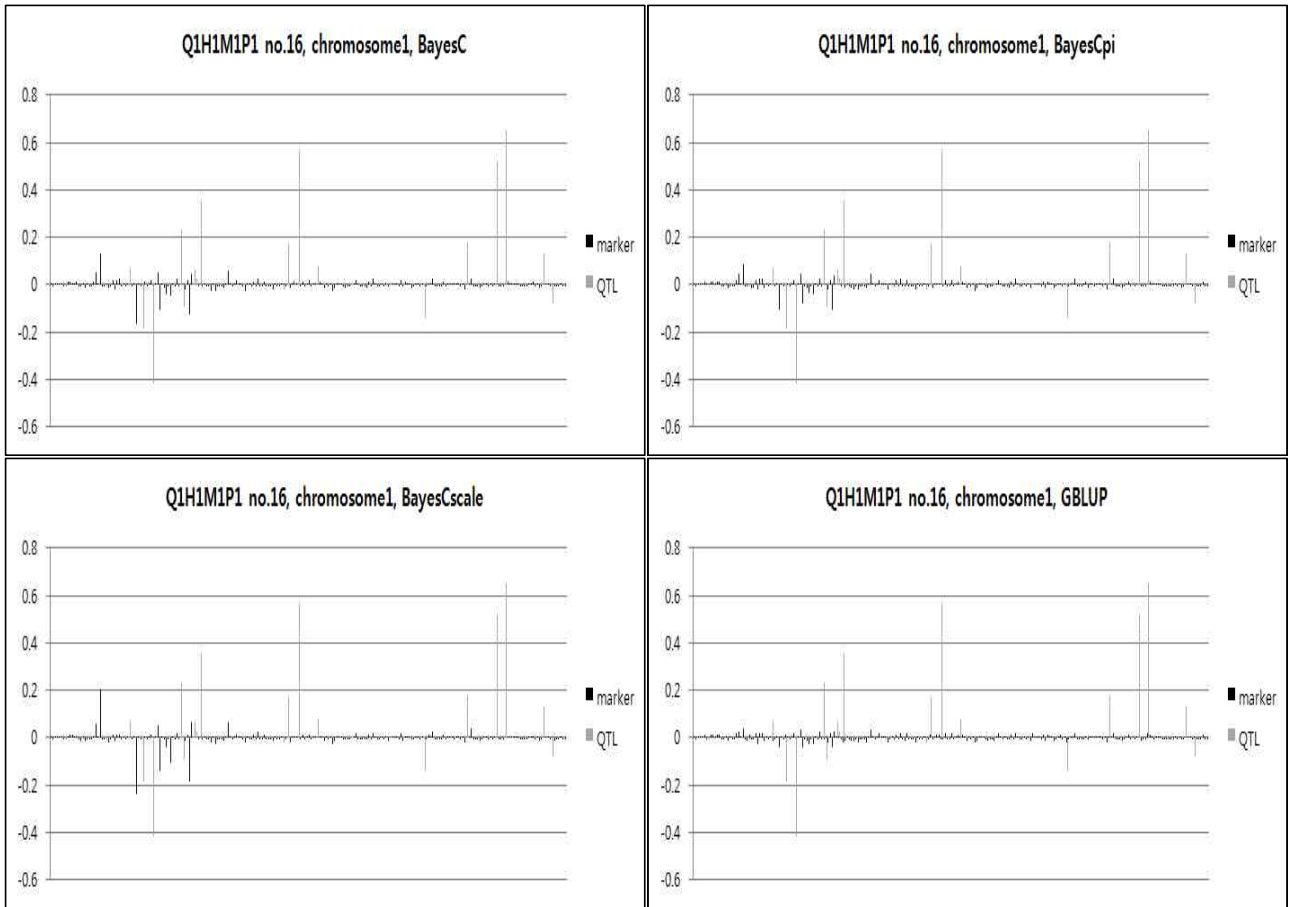


그림 3. 통계분석 방법별 양적형질좌위의 유전효과 추정치 비교(가정: Q1H1M1P1)

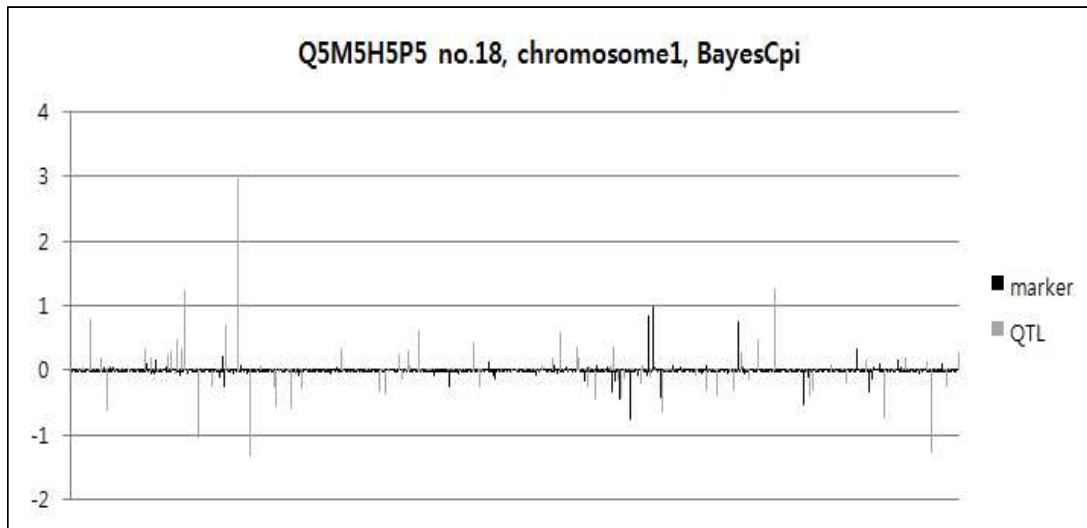


그림 4. 특정 모의실험조건에서 BayesCpi 방법을 이용한 표지인자 효과 추정 예(QTL 밀도=1cM, 유전력=0.5, Marker밀도=0.1cM, 집단크기=2000)

(3) 모의시험 가정별 정확도 비교

각 조건별 모의 유전체 자료를 BayesA, BayesB, BayesB1, BayesBfast, BayesC, BayesCpi, BayesCscale, GBLUP 방법을 이용하여 각 표지인자의 효과를 추정하고 이를 각 개체별로 모든 표지인자의 유전자형에 곱하고 이를 합하여 육종가(GEBV, genomic estimated breeding value)를 추정한 뒤 알고 있는 참육종가와 상관 계산을 통하여 정확도를 계산하고 각 모의자료의 조건별로 비교하였다.

(가) 양적형질좌위수(밀도)

동일한 조건하에서 형질에 관여하는 양적형질좌위의 밀도(수)가 5cM(60개)에서 1cM(300개)으로 증가하였을 경우 특이적으로 유전력이 낮은 경우 정확도가 감소하는 모양을 나타내었으며(그림 5). 유전력이 높을 경우 오히려 정확도가 증가하는 것을 확인할 수 있었다(그림 6).

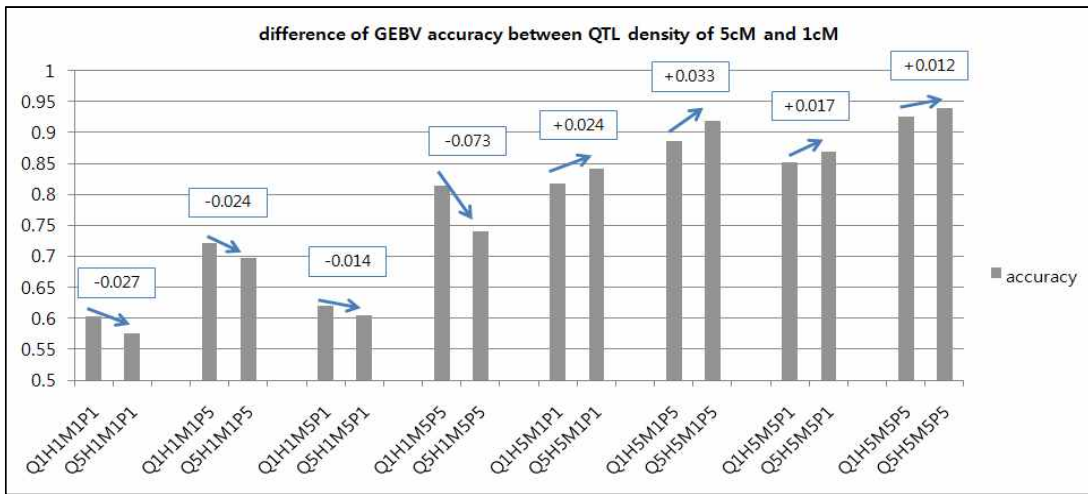


그림 5. 양적형질 좌위의 밀도가 5cM에서 1cM로 변화할 경우의 조건별 정확도 변화 추이 (Q1, Q5 : 양적형질좌위의 분포 간격을 각각 5cM 단위와 1cM 단위로 가정(임의적 간격이나 평균이 각각 0.5, 0.1cM 임), H1, H5 : 유전력을 각각 0.1과 0.5를 가정, M1, M5 : 표지인자의 분포 간격을 각각 0.5cM 단위와 0.1cM 단위로 가정(등간격 배치), P1, P5 : 참조집단의 규모를 각각 400두, 2000두 가정)

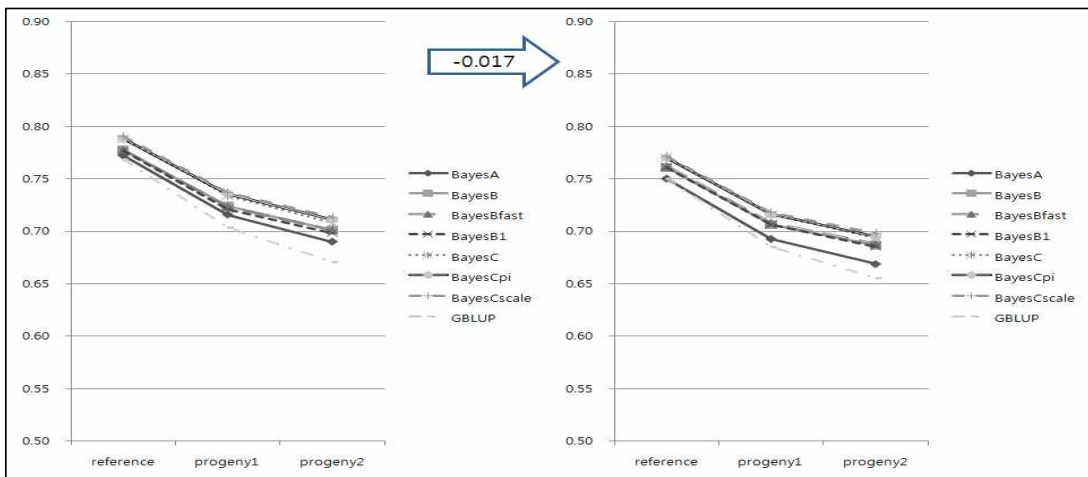


그림 6. 양적형질 좌위의 밀도가 5cM에서 1cM로 변화할 경우에서 추정된 유전체 육종가추정치 의 정확도에 대한 추정방법별 비교(참조집단 및 2대 후손 자료)

(나) 유전력

동일한 조건하에서 형질의 유전력이 0.1에서 0.5로 증가하였을 경우 전반적으로 정확도가 크게 증가하는 것을 확인할 수 있다. 증가의 상승폭은 형질발현에 관여하는 양적형질좌위 밀도(수)가 높을수록 크게 나타났으며, 참조집단의 크기가 클 경우 증가량은 상대적으로 참조집단이 작을 때보다 작은 것으로 나타나 유전력과 참조집단의 크기는 정확도의 상승에 공통적인 영향을 미치는 것으로 사료된다.

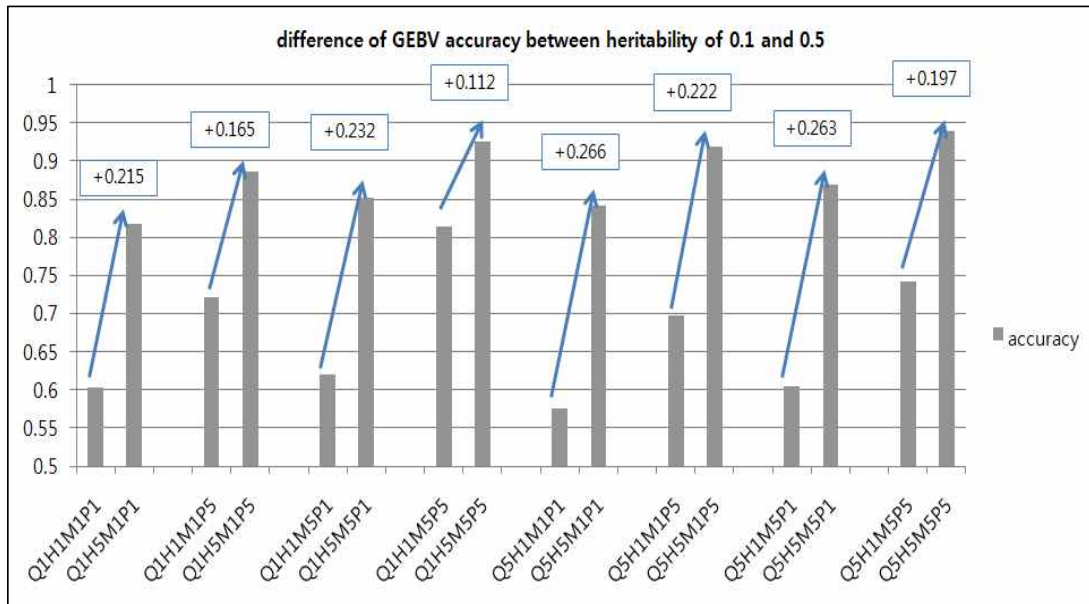


그림 7. 유전력이 0.1에서 0.5로 변화할 경우의 조건별 유전체육종가 추정치의 정확도 변화 추이

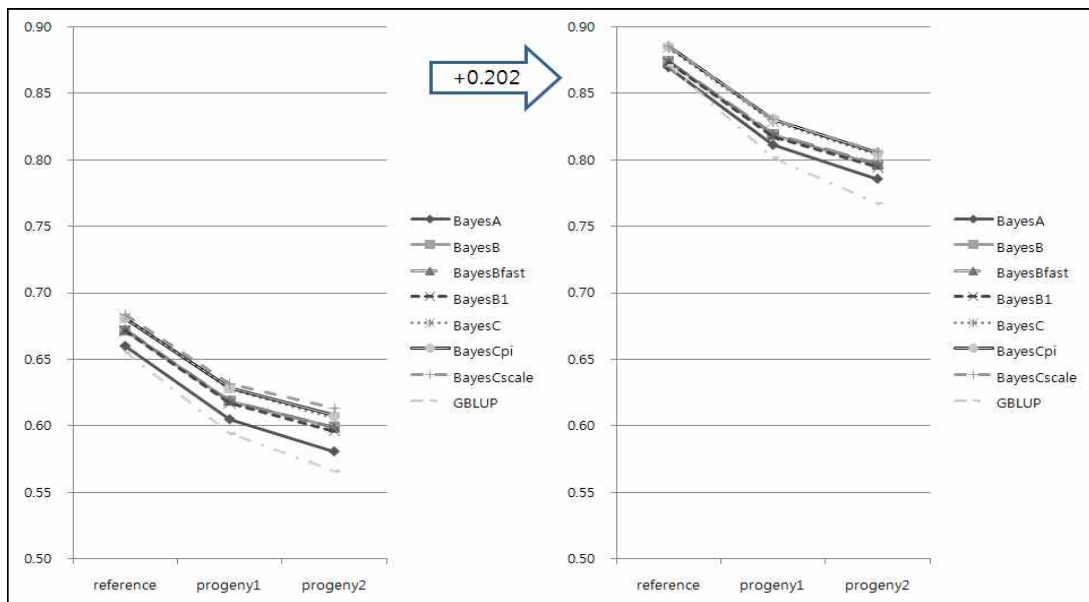


그림 8. 유전력이 0.1에서 0.5로 변화할 경우에서 추정된 유전체 육종가 추정치의 정확도에 대한 통계분석방법별 비교(참조집단 및 2대 후손 자료)

(다) 표지인자수(밀도)

동일한 조건하에서 표지인자의 밀도(수)가 0.5cM에서 0.1cM으로 증가하였을 경우 전반적으로 정확도가 증가하는 것을 확인할 수 있으며, 특이적으로 GBLUP방법의 정확도가 크게 향상되었다.

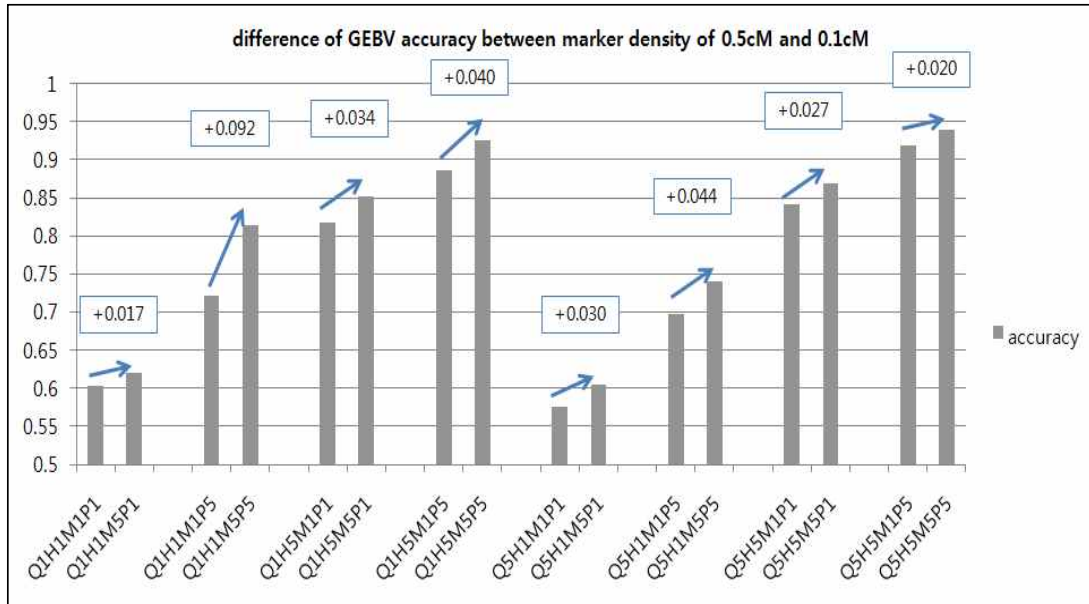


그림 9. 표지인자의 밀도에 따른 모의시험별 유전체 육종가 추정치의 정확도 비교

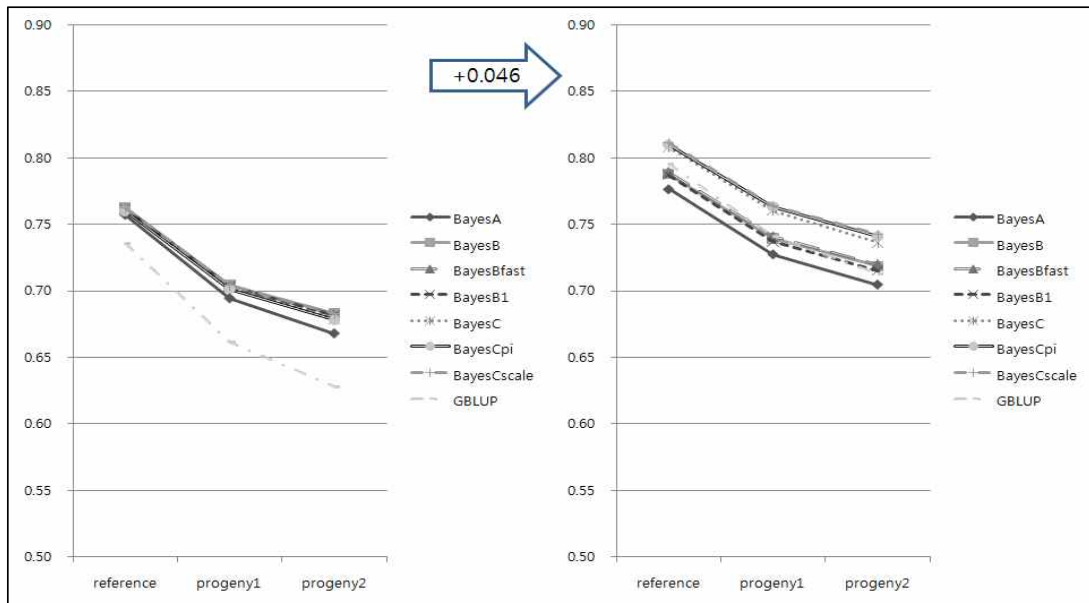


그림 10. 표지인자의 밀도가 0.5cM에서 0.1cM로 변화할 경우에서 추정된 유전체 육종가 추정치의 정확도에 대한 추정방법별 비교(참조집단 및 2대 후손 자료)

(라) 참조집단 크기

참조집단의 크기가 400두인 자료에서 표지인자 효과를 추정하여 육종가를 계산하는 경우와 2,000두 자료로 육종가를 계산하는 경우 전반적으로 정확도가 크게 증가하는 것을 확인할 수 있다.

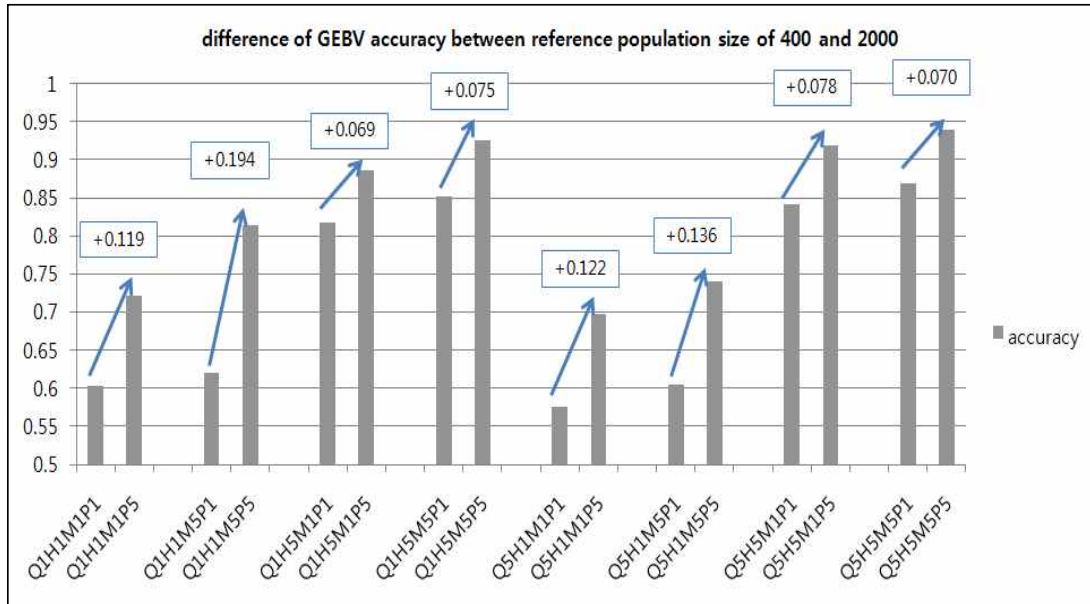


그림 11. 참조집단 크기에 따른 모의시험별 유전체 육종가 추정치의 정확도 비교

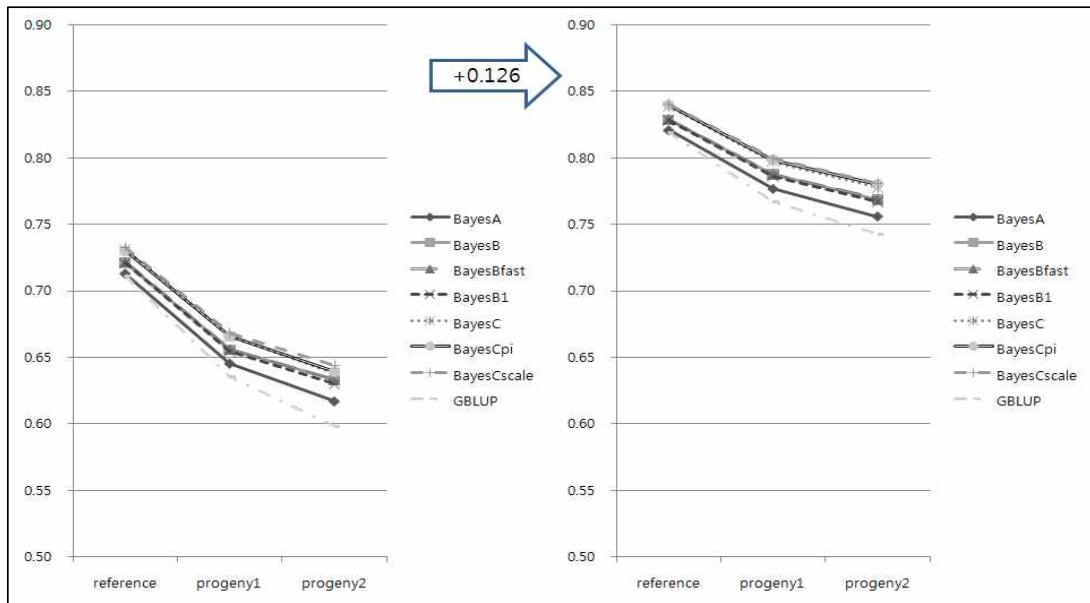


그림 12. 참조집단 크기에 따른 통계분석방법별 유전체 육종가 추정치의 정확도 비교(참조집단 및 2대 후손 자료)

(마) 통계분석방법

기존의 육종가 추정 방법과 비교하여 얼마나 정확도가 향상하였는가를 판단하고 가장 정확한 유전체 육종가 추정 방법을 제시하며 연산 효율에 따른 최적 방법론을 제시하기 위하여 각 방법론별 비교를 실시하였다.

유전체 정보를 이용하여 육종가를 추정할 경우 정확도가 기존의 BLUP방법을 이용하였을 때보다 매우 높게 나타난 것을 확인할 수 있으며, 특히 표현형자료가 없는 후대에서 그 차이가 확실하게 나타나고 있다. 베이지안 방법을 이용한 유전체 육종가의 정확도는 BayesC 계열(BayesC, BayesCpi, BayesCscale)에서 가장 높게 나타났고 BayesB 계열(BayesB, BayesBfast, BayesB1), BayesA, GBLUP순으로 나타났다.

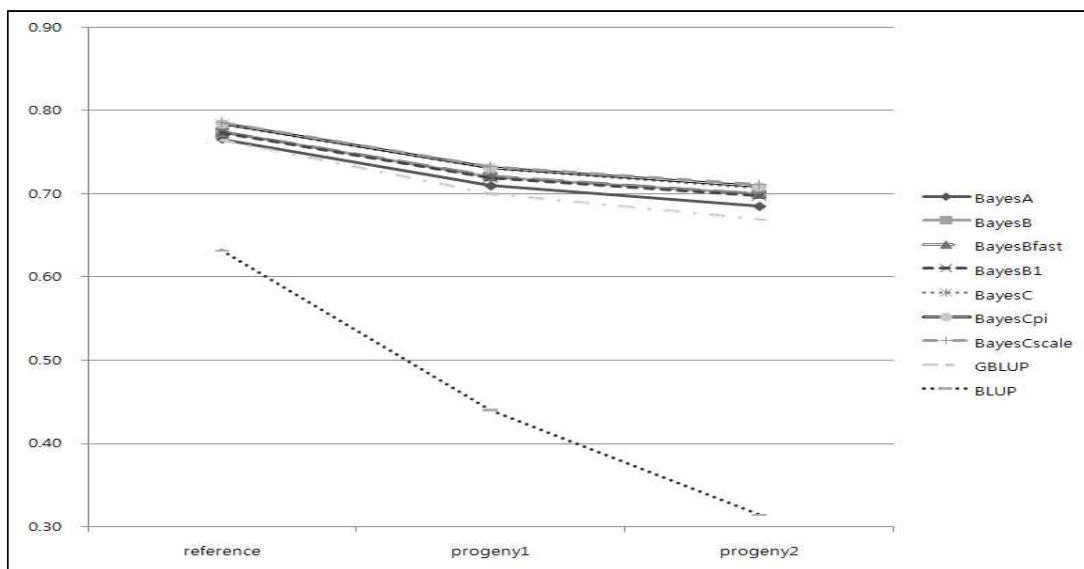


그림 13. 혈통정보를 이용하는 BLUP방법과 유전체 정보를 이용하는 통계분석방법간의 육종가 추정치의 정확도 비교(참조집단 및 2대 후손)

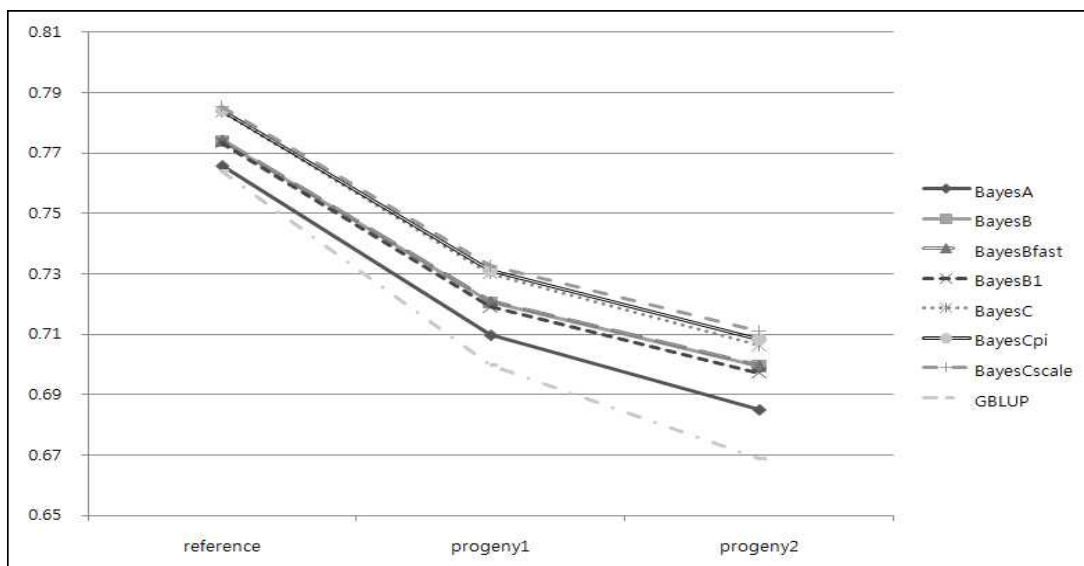


그림 14. 통계분석방법별 유전체육종가 추정치의 정확도 비교

(바) 요인별 육종가 추정치의 비교

모의시험별 육종가 추정치의 정확도에 대한 유의적인 차이가 있는지를 검정하기 위하여 SAS GLM을 이용한 duncan 검정을 실시하였으며, 결과는 표 4와 같다.

표 4. 모의시험조건별 육종가 추정치의 정확도에 대한 유의성 검정

요인		참조집단	1대 후대	2대 후대
		평균	평균	평균
QTL density	Q1 (5cM)	0.765 ^{NS}	0.693 ^{NS}	0.657 ^{NS}
	Q5 (1cM)	0.757 ^{NS}	0.688 ^{NS}	0.654 ^{NS}
heritability	H1 (0.1)	0.652 ^b	0.586 ^b	0.556 ^b
	H5 (0.5)	0.870 ^a	0.795 ^a	0.755 ^a
marker density	M1 (0.5)	0.742 ^b	0.667 ^b	0.632 ^b
	M5 (0.1)	0.780 ^a	0.714 ^a	0.679 ^a
reference population size	P1 (400)	0.710 ^b	0.631 ^b	0.594 ^b
	P5 (2,000)	0.811 ^a	0.750 ^a	0.718 ^a
methods	BLUP	0.635 ^c	0.442 ^d	0.316 ^d
	GBLUP	0.765 ^b	0.701 ^c	0.670 ^c
	BayesA	0.767 ^{ab}	0.711 ^{bc}	0.686 ^{bc}
	BayesB1	0.774 ^{ab}	0.720 ^{ab}	0.698 ^{ab}
	BayesBfast	0.775 ^{ab}	0.722 ^{ab}	0.701 ^{ab}
	BayesB	0.775 ^{ab}	0.722 ^{ab}	0.701 ^{ab}
	BayesC	0.785 ^{ab}	0.731 ^a	0.708 ^a
	BayesCpi	0.785 ^{ab}	0.732 ^a	0.710 ^a
	BayesCscale	0.787 ^a	0.734 ^a	0.712 ^a

NS : No significance, Different alphabets on superscript of same column are significant difference (p<0.001)

2. 한우집단의 유전체육종가 추정 모의시험

본 연구에서는 현행 SNP chip과 유사한 형태의 모의자료를 생성하여, 이를 이용한 유전체 육종가 예측 연구를 수행하였다. 모의시험방법으로써 SNP자료생성 후 유전자형에 대한 pruning 이전 및 이후와 유전체 육종가 추정을 위한 반응변수를 표현형 및 BLUP방법으로 추정된 EBV로 설정하여 유전체육종가 추정치의 정확도를 비교하였다. 모의 자료 생성에 있어서는 QMSim (Sargolzaei, M, & Schenkel, F. S.,2009, Bioinformatics)을 사용하였고, 유전체 육종가(GEBV)의 추정은 **rrBLUP package** (Endelman, J. B., 2011, The Plant Genome)를 사용하였다.

가. 재료 및 방법

(1) 모의자료 생성

(가) 모의 자료 생성을 위한 유전모수

본 모의실험에서는 Bovine 50K SNP pannel과 유사하게 유전체 데이터를 구성하였으며, 표현형 형질과 QTL에 대한 유전모수 및 단일염기다형 Marker수에 대해 표 5와 같이 가정하였다.

표 5. 모의 자료 생성에 이용한 모수

항목	설정값
h^2	0.3
QTL variance	0.3
Residual variance	0.7
Phenotypic variance	1
No. chromosomes	29
No. QTL in total	725
No. markers in total	53,998

모의실험을 통해 생성된 자료에 대한 세대별 기초 통계치는 다음 표 6와 같다.

표 6. 표현형, 참육종가 (TBV), 잔차, 추정육종가에 대한 기초 통계량

Gen.	Phenotype		TBV (QTL)		Residual		EBV	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
0	-0.069	0.968	0.000	0.548	-0.069	0.821	-0.021	0.290
1	-0.052	1.019	-0.055	0.522	0.003	0.860	-0.056	0.365
2	-0.141	0.976	-0.163	0.525	0.022	0.861	-0.149	0.372
3	-0.011	0.989	-0.075	0.576	0.064	0.821	-0.037	0.413
4	0.089	0.971	0.031	0.504	0.059	0.852	0.059	0.322
5	0.018	1.050	-0.001	0.501	0.019	0.909	0.017	0.329
6	0.001	0.998	-0.085	0.513	0.086	0.836	-0.009	0.334
7	-0.143	0.929	-0.129	0.495	-0.014	0.805	-0.116	0.275
8	0.127	1.020	0.064	0.481	0.063	0.856	0.098	0.310
9	0.111	0.997	0.089	0.493	0.021	0.825	0.103	0.323
10	0.150	1.000	0.158	0.532	-0.008	0.852	0.160	0.320
Overall	0.011	0.998	-0.016	0.524	0.026	0.846	0.005	0.332

표현형과 혈통정보를 이용한 추정육종가(EBV)와 참육종가(TBV)간의 상관계수(r) 즉 추정육종가의 정확도는 표 7과 같으며 모의시험에서 생성된 전체 집단에서 정확도는 0.62, 유전자형 정보를 지닌 집단(8, 9, 10 세대; genotyped animal)의 정확도는 0.60으로 추정되었다.

표 7. 표현형과 혈통정보를 이용한 추정육종가(EBV)와 참육종가(TBV)간의 상관계수(r)

Gen.	r(EBV,TBV)	r(EBV,TBV)		r(EBV,TBV)	
		_Male	_Female	_Sire	_Dam
0	0.530	0.810	0.519	0.000	0.000
1	0.655	0.642	0.669	0.922	0.611
2	0.659	0.648	0.669	0.919	0.691
3	0.737	0.738	0.736	0.958	0.689
4	0.615	0.632	0.598	0.764	0.765
5	0.623	0.641	0.609	0.968	0.609
6	0.656	0.653	0.662	0.746	0.633
7	0.574	0.585	0.572	0.751	0.705
8	0.550	0.577	0.519	0.463	0.621
9	0.605	0.659	0.545	0.827	0.598
10	0.649	0.637	0.655	0.858	0.623
¹⁾ GA	0.601	0.624	0.573	0.716	0.614
Total	0.623	0.657	0.614	0.743	0.595

¹⁾GA : genotyped animal

(2) 모의자료 전처리

생성된 모의 자료의 전처리과정을 통하여, 결측치가 적은 SNP만을 선택하여 분석에 이용하였다. 전처리 기준은 다음(표 8)과 같으며, 최종적으로 53,253 개의 SNP을 유전체육종가 추정을 위한 분석에 이용하였다.

표 8. 모의자료 DNA샘플로부터 유전자형 결측률 5%이상, MAF(Minor allele frequency)<10%, Hardy-Weinberg 평형 검정 ($p > 0.00001$)미만 요인에 대한 품질평가(QC) 내역

Category	Number of Marker
Genotyped SNP marker	53,998
Hardy-Weinberg 평형 검정 ($p > 0.00001$)	11
Minor allele frequency < 10 %	734
Percentage of missing genotype < 5 %	0
Selected SNP markers	53,253

(3) 연관불평형(LD)을 이용한 SNP pruning

많은 양의 SNP중에서 서로 연관평형(Linkage Equilibrium)을 이루는 SNP만을 선별하는 pruning과정을 통해 low density marker를 이용한 GEBV의 추정을 통하여 정확도를 비교하고자 하였다. pruning 기준으로는 분산 팽창 지수 (variance inflation factor(VIF))를 이용하였으며, window size 50, moving size 5 그리고 VIF 임계점은 2로 설정하였다. 여기서 $VIF = 1/(1-R^2)$ 이며 R^2 는 다중상관계수이다 (즉, $R^2 = 0.50$). Pruning 과정을 통하여, 4,768 개의 SNP subset을 통하여, pruning 전후의 추정 정확도를 5 fold 교차 검증을 통하여 비교하였다.

(4) 표현형 변화에 따른 유전체추정 육종가(GEBV)의 정확도 비교

표현형을 이용하여 GEBV 추정하는 것이 가장 일반적인 경우이나, 현실적으로 표현형 자료가 충분치 않은 경우 다양한 추정치를 GEBV추정을 위한 반응변수로 사용하게 되며, 이러한 반응변수로는 DYD, de-regressed proofs, national EBVs 또는 MACE evaluations 등을 실제 International Bull Evaluation (Interbull)에 사용하고 있다(<http://interbull2.slu.se/www/v1/>). 본 실험에서는 모의 자료로부터 추정된 EBV를 반응변수로 사용하였을 때, 유전체추정육종가의 정확도 차이를 비교하고자 하였다.

나. 결과

pruning 전후, EBV/표현형에 따른 네 가지의 경우에 대한 5-fold 교차검증의 결과는 다음과 같다.

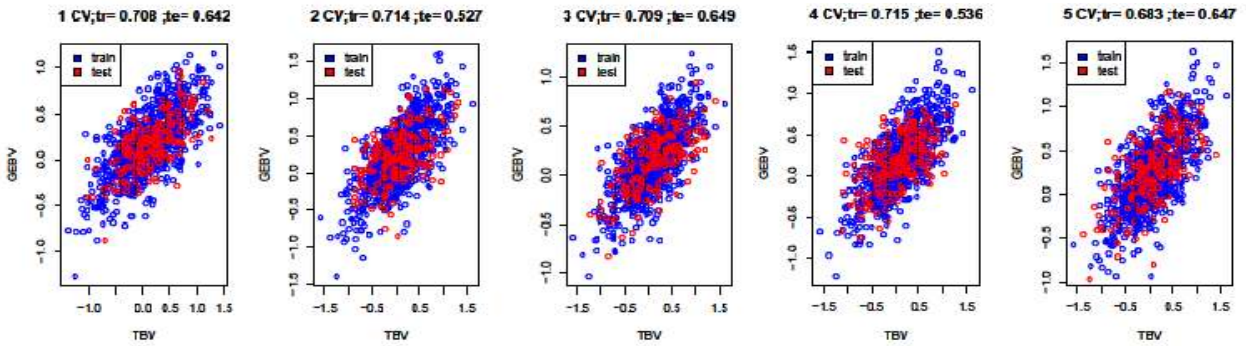


그림15. pruning 후의 표현형 정보를 이용한 GEBV 정확도

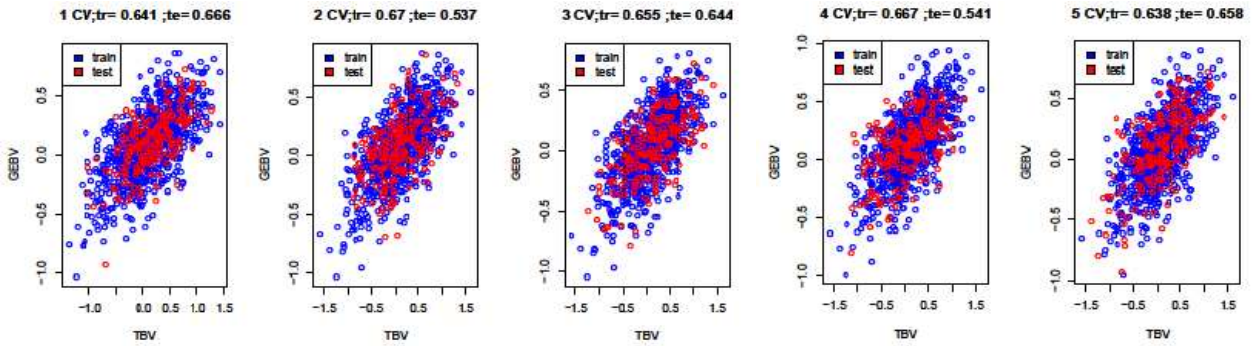


그림16. pruning 후의 EBV 정보를 이용하여 추정한 GEBV의 5-fold 교차 검증

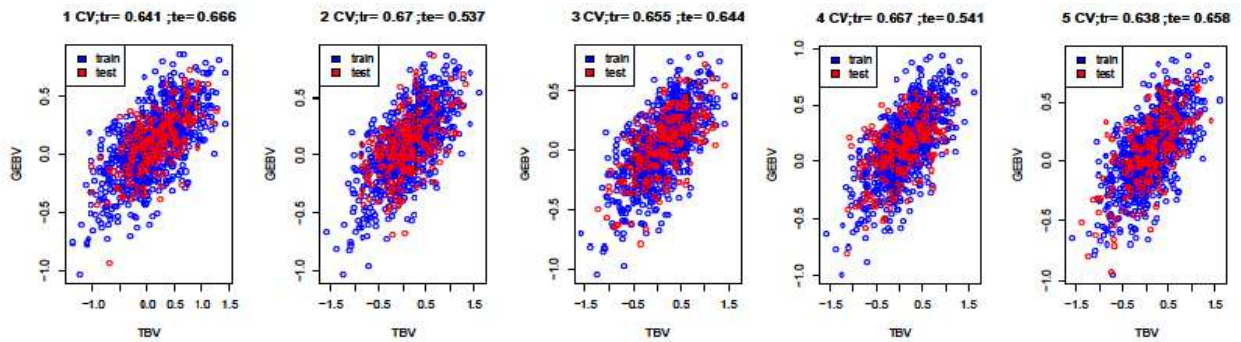


그림17. pruning 전의 표현형 정보를 이용한 GEBV 5-fold 교차 검증

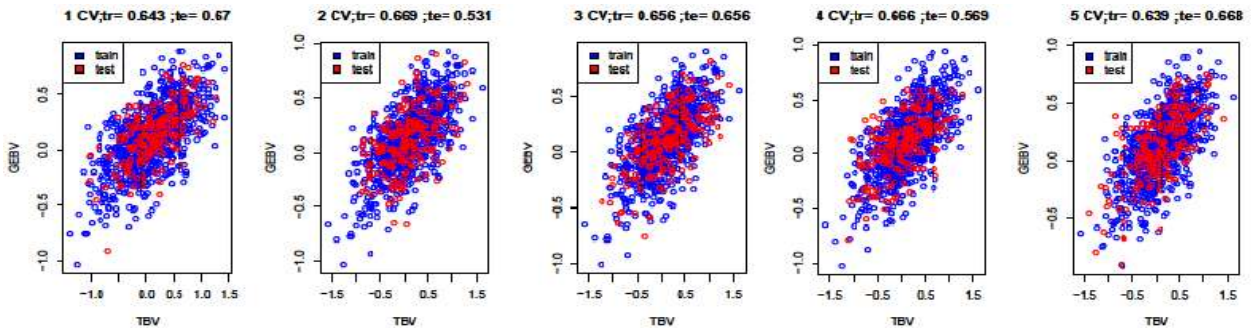


그림18. pruning 전의 EBV 정보를 이용한 GEBV 5-fold 교차 검증

전체적으로 네 가지 경우 TBV 대비 모두 60-62%의 정확도를 표현형정보가 없는 집단(candidate)에서 나타내었으며 이상 네 가지 경우에 대한 GEBV 추정결과의 정확도를 표9에 나타냈다.

표 9. pruning 전후 및 표현형/EBV를 반응변수로 하였을 때의 정확도 비교

Pruning		Pruned				Unpruned			
y		Phenotype		EBV		Phenotype		EBV	
Replicate	Reference	Candidate	Reference	Candidate	Reference	Candidate	Reference	Candidate	
CV1	0.71	0.64	0.64	0.67	0.72	0.66	0.64	0.67	
CV2	0.71	0.53	0.67	0.54	0.72	0.55	0.67	0.53	
CV3	0.71	0.65	0.66	0.64	0.72	0.67	0.66	0.66	
CV4	0.71	0.54	0.67	0.54	0.72	0.59	0.67	0.57	
CV5	0.68	0.65	0.64	0.66	0.69	0.66	0.64	0.67	
Mean	0.71	0.60	0.65	0.61	0.71	0.62	0.65	0.62	
SD	0.01	0.06	0.01	0.06	0.01	0.06	0.01	0.06	

Candidate집단의 유전체 육종가 추정치는 pruning 자료에 비하여 unpruned 자료에서 1-2%정확도 상승효과를 보였으며, 반응변수에 따른 육종가추정치의 정확도를 살펴보면, 표현형 자료를 사용하였을 경우가 EBV를 사용하였을 때보다 Reference집단에서 뚜렷한 정확도 상승(6%)을 보였으나 Candidate집단에서는 차이를 보이지 않았다. 따라서, 적정 수준의 LD pruning을 통한 효율적인 SNP pannel의 가능성이 있음을 알 수 있다.

제 2 절 한우 종축에 대한 유전체 정보 활용 분석

I. 재료 및 방법

1. 한우 후보종모우와 후대검정 대상우의 유전체, 도체형질 및 혈통 데이터 수집

가축의 유전적 개량량을 증대시키기 위해서는 선발의 강도, 유전표준편차, 정확도를 높이고 세대간격을 단축해야 가능하다. 이 중 현행 개량체계를 유지하면서 유전체 선발을 통해 유전적 개량량을 증대시킬 수 있는 방안은 유전평가의 정확도를 향상시키는 것이다. 유전체 선발의 정확도는 형질의 유전력, 참조집단크기, QTL수, 유효집단크기, SNP밀도, 분석방법 등 다양한 요인에 따라 영향을 받는데 이 중 현실적으로 인위적 변경가능하며 정확도에 영향을 많이 미치는 요인이 참조집단의 크기라고 판단된다. 참조집단은 일반적으로 자신의 표현형데이터와 유전체정보를 모두 갖는 개체의 집단을 지칭하며, 본 연구에서는 후대검정우가 참조집단이 된다. 연간 350두 내외로 생산되는 후대검정우에 대한 참조집단 크기를 증가시키기 위하여 5년에 걸쳐 본 연구가 수행되었다.

가. 유전체 자료 형성을 위한 한우 시험축 채혈 및 gDNA 추출

(1) 유전체 자료 형성을 위한 시험축 채혈

SNP chip 분석을 위한 시료는 농협중앙회 한우개량사업소에서 보유하고 있는 한우 후보종모우 및 이들의 후대검정우를 이용하였으며, 경정맥으로부터 약 5ml를 항응고제가 들어있는 튜브에 채혈하였다. 구제역으로 인하여 일부 혈액을 구할 수 없는 경우가 발생하여 이러한 개체에 대해서는 정액에서 genomic DNA(gDNA)를 추출하였다. 각 시료에 대한 연도별 채취 현황은 표 10에 나타내었다.

표 10. 연도별 후보종모우 및 후대검정우의 시료 채취 현황

연차	채혈일자	대상	내역		전체
			후보종모우	후대검정우	
1차년도	2010년 02월	46차	18	204	222
		47차	17	173	190
2차년도	2011년 06월	48차	21	187	208
		49차	18	169	187
3차년도	2011년 12월	50차	18	184	202
		51차	19	186	205
4차년도	2013년 1월	52차	17	187	204
		53차	12	114	126
5차년도	2013년 3월	64차당대	44	299	343
전체			184	1,703	1,887

(2) gDNA 추출 및 준비

gDNA 추출은 혈액 2ml로부터 modified Isopropanol법으로 추출하였으며, 200ng/ul이상의 고농도, 고순도의 DNA를 일정하게 준비하였다. 추출된 gDNA는 그림 19과 같이 전기영동을 실시하여 추출결과를 확인하였으며, 분광광도계를 이용하여 농도와 OD260/280, OD260/230을 측정하여 품질을 평가하였다(표 11).

표 11. 한우 1,887두에 대한 gDNA 추출 결과(일부)

Name	Conc(ng/μL)	OD260/280	OD260/230	OD260	OD280	OD230
501	667.55	1.91	2.12	13.443	7.077	6.394
502	320.69	1.91	2.09	6.44	3.378	3.096
503	348.8	1.9	1.99	7.035	3.722	3.56
504	264.97	1.92	1.98	5.33	2.788	2.713
505	344.77	1.9	2.02	6.991	3.725	3.51
506	436.65	1.89	2.12	8.824	4.701	4.214
507	301.71	1.9	2.04	6.082	3.227	3.009
508	330.55	1.88	1.95	6.645	3.55	3.427
509	449.51	1.9	2.13	9.037	4.782	4.276
510	404.11	1.9	2.02	8.209	4.38	4.135
511	584.58	1.91	2.14	11.712	6.134	5.48
512	308.09	1.91	2.06	6.136	3.199	2.963
513	472.29	1.89	2.05	10.036	5.58	5.189
514	431.25	1.9	2.13	8.676	4.588	4.099
.
.
.

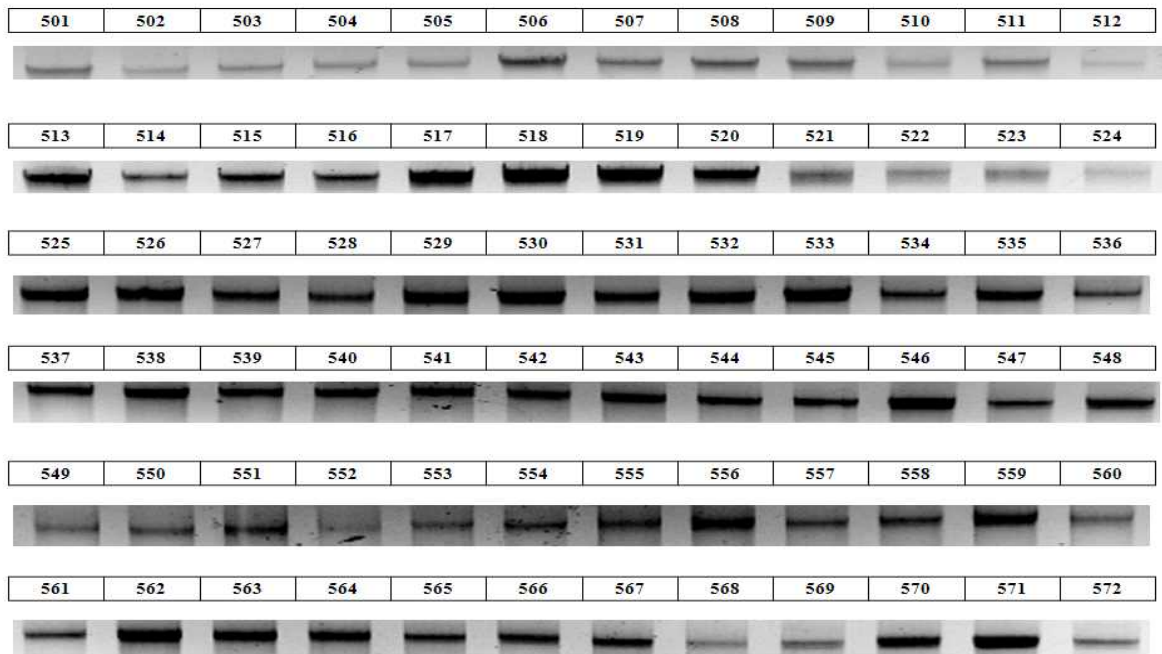


그림 19. 전기영동(gel electrophoresis)사진.(일부)

한우 sample 1,887두에 대하여 Chip assay에 사용하기 적절한 50~100 ng/ul로 희석 하였으며, OD260/280의 값이 1.8 미만인 두수에 대하여 Purification를 실시하여 Chip assay를 준비하였다.

나. SNP chip 분석 및 Genotype 확보

한우 후대검정에 공시되어 표현형자료가 수집된 개체들을 대상으로 생후 약 6개월령에 혈액을 채취하여 DNA를 추출한 다음 추출된 DNA 표본을 대상으로 약 50,000개의 단일염기다형(Single nucleotide polymorphism; SNP)을 갖는 Illumina BovineSNP50 Beadchip(V1, V2)과 Bovine HD chip (약 777,000개)을 활용하여 개체별 대용량 유전자형(유전체) 분석을 실시하였다.



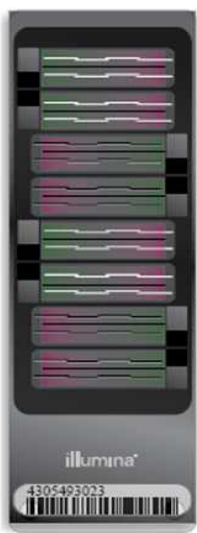
BovineSNP50 Beadchip	BovineSNP50 V2 Beadchip	BovineHD Beadchip
		
54,001SNP	54,609SNP	777,962SNP

그림 20. 다양한 Bovine Chip에 대한 비교

SNP Chip 분석을 위한 실험절차는 그림 21와 같으며, 3일간 진행되는 실험의 각 과정별 분석 내용은 그림 21와 같다.

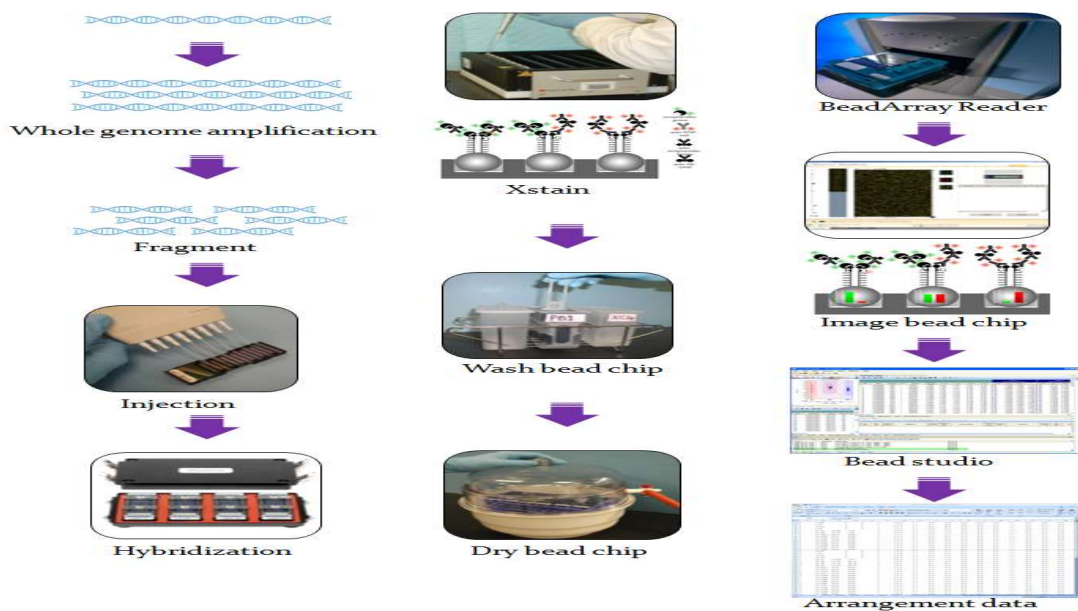


그림 21. SNP chip 분석과정 모식도(출처, Illumina USA)

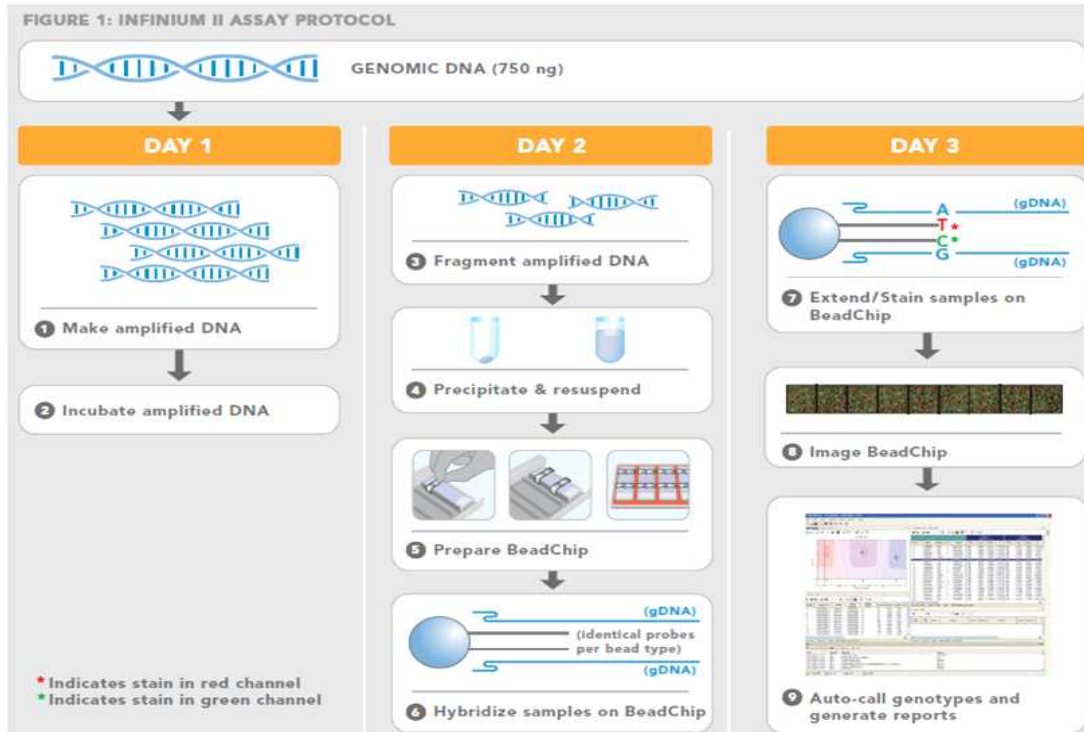


그림 22. SNP chip 분석과정별 내용

Chip assay 과정에서 발생할 수 있는 오류를 최소화하기 위하여 그림 23와 같이 실험날짜 및 실험자, 기계 정보 등의 기본정보 및 사용된 시약의 Lot no, Chip number 정보와 chip 위에 올려진 sample 위치를 기록 하면서 실험을 수행하였다.

그림 23. Lab Tracking Form

SNP chip으로부터 유전자형 분석하는 방법은 실험이 완료된 Chip을 scan기기를 이용하여 이미지 파일과 idat 파일이 생성하며, 이미지 파일의 경우는 sample별 green색과 red색 등, 두 가지 구별된

색상으로 유전자형을 판독하였다.

SNP Genotyping의 경우 Genome-Studio 프로그램을 이용하여 그림 24과 같이 감광 분석된 chip data를 염기로 구성된 유전자형으로 전환하였다.

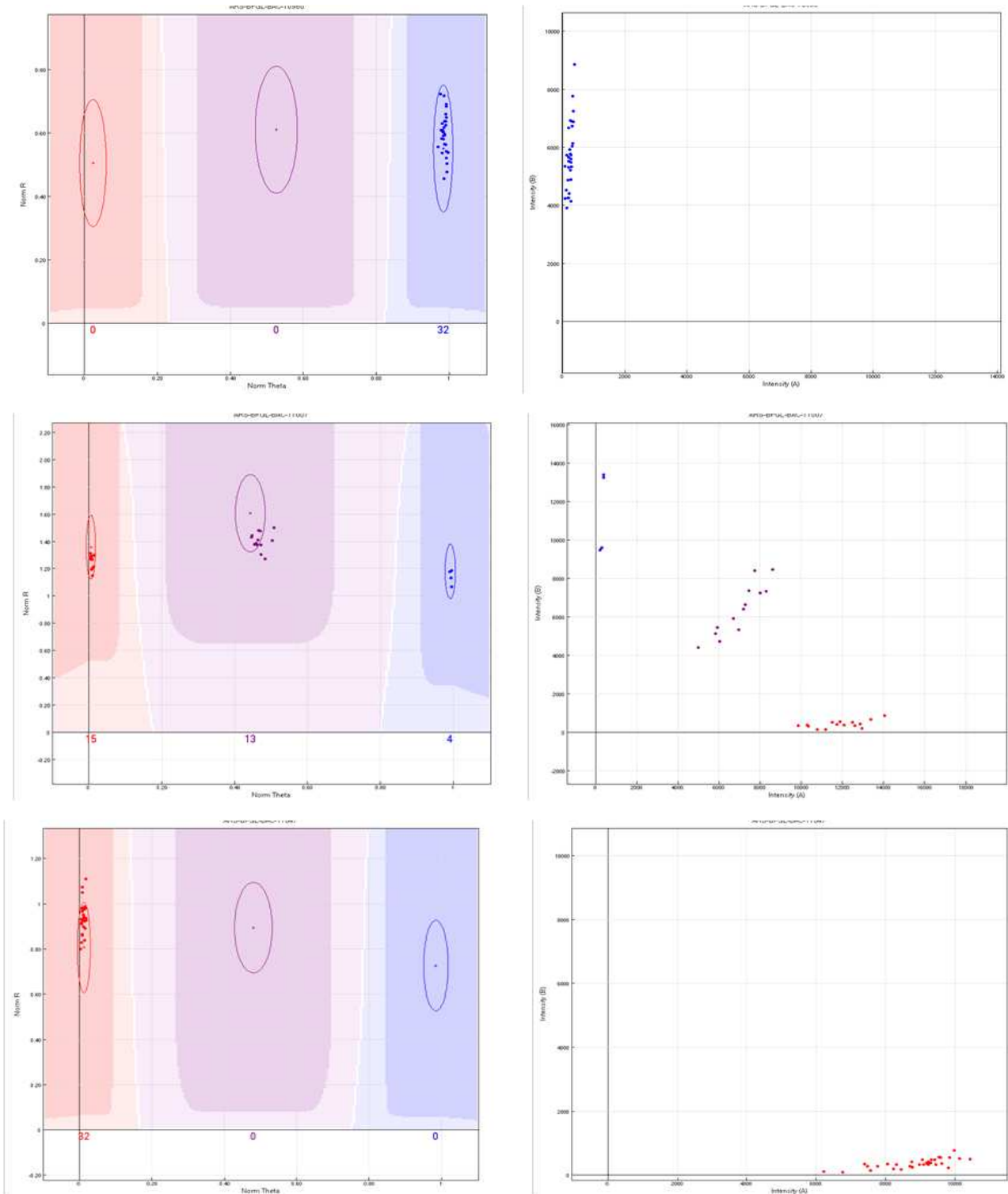


그림 24. Genome-Studio 프로그램을 이용한 SNP의 genotyping (BB Homotype 일 경우(상), AB Heterotype 일 경우(중), AA Homotype 일 경우(하))

다. 유전체자료 수집

유전자형으로 전환된 유전체자료의 결과는 표 12와 같으며, 5년간 한우종축 1,679두의 SNP 분석 정보는 표 13에 나타냈다. 연구 초기인 1차년도에는 고밀도 칩인 HD의 분석비용이 많이 들어가기 때문에 고밀도칩의 선택보다는 유전체 분석두수를 늘리기 위하여 50K칩을 사용하였다. 그러나 이후 HD칩의 비용이 감소함으로 인해 HD칩의 비용이 이전의 50K칩 비용과 거의 비슷하게 되어 2 - 3차년도 연구과제에서 HD칩을 수집하여 연구하였다. 4 - 5차년 연구는 2단계 연구로써 유전체 선발을 검증하는 단계의 연구를 수행하고자 하였다.

표 12. BovineHD SNP chip 분석결과(일부)

Index	Name	Chr	Position	시료번호				
				104021	104022	104060	104106	104107
1	ARS-BFGL-BAC-10172	14	6371334	AG	AG	AA	AG	AG
2	ARS-BFGL-BAC-1020	14	7928189	AG	AG	AA	AG	AA
3	ARS-BFGL-BAC-10245	14	31819743	GG	GG	GG	GG	GG
4	ARS-BFGL-BAC-10345	14	6133529	AA	AA	AA	AA	AA
5	ARS-BFGL-BAC-10365	14	27005721	CC	CC	CC	CC	CC
6	ARS-BFGL-BAC-10591	14	17544926	AG	AA	AG	AG	AA
7	ARS-BFGL-BAC-10793	14	29259114	CG	GG	GG	CG	CG
8	ARS-BFGL-BAC-10867	14	34639444	CC	CG	CG	CG	CG
9	ARS-BFGL-BAC-10951	10	17911906	AA	AA	AA	AA	AA
10	ARS-BFGL-BAC-10952	10	18882288	AA	AA	AA	AA	AA
11	ARS-BFGL-BAC-10960	10	20609250	GG	GG	GG	GG	GG
12	ARS-BFGL-BAC-10972	10	20792754	GG	GG	GG	GG	GG
13	ARS-BFGL-BAC-10975	10	21225382	AA	AA	AA	AA	AA
14	ARS-BFGL-BAC-11000	10	79252023	AA	AA	AA	AA	AA
15	ARS-BFGL-BAC-11003	10	80410977	AA	AA	AA	AA	AA
16	ARS-BFGL-BAC-11007	10	80783719	AA	AA	AA	AA	AG
17	ARS-BFGL-BAC-11025	10	84516867	AC	CC	AC	CC	AC

표 13. 연도별 SNP Chip 분석정보

년도	SNP 분석			DNA Analysis Kit
	후보종모우	후대검정우	계	
2010	35	253	288	BovineSNP50 V1 Beadchip
2011	39	297	336	BovineHD Beadchip
2012	37	347	384	BovineHD Beadchip
2013	35	301	336	BovineSNP50 V2 Beadchip
2014	36	299	335	BovineSNP50 V2 Beadchip
계	182	1,497	1,679	

라. 표현형자료

우리나라의 한우품종은 고유품종으로써 외부로부터의 유전자원이 도입되지 않는 폐쇄된 집단이며, 한우 수소의 개량은 농협중앙회 한우개량사업소를 통해 국가에서 실시하고 있다. 본 분석에 이용된 표현형자료는 23차부터 54차까지 농협중앙회 한우개량사업소에서 수행하고 있는 한우 후대검정에 공시하여 도축된 후대검정우 5,641두로부터 수집된 도축 시 등심단면적(EMA), 등지방두께(BF) 및 근내지방도(MS)에 대한 도축성적을 이용하였다.

한우에 대한 도축 성적은 소를 도축한 후 2등 분할된 왼쪽 반도체의 마지막등뼈(흉추)와 제 1허리뼈(요추) 사이를 절개한 후 배최장근의 등심쪽 절개면에 대하여 등급을 판정하게 되며 형질별 측정방법은 다음과 같다.

(1) 등지방두께

등지방두께는 등급판정부위에서 배최장근단면의 오른쪽면을 따라 복부 쪽으로 3분의 2들어간 지점의 등지방을 mm단위로 측정하며 등지방두께가 1mm이하인 경우에는 1mm로 간주한다.

(2) 등심단면적

등심단면적은 등급판정부위에서 가로, 세로가 1cm단위로 표시된 면적자를 이용하여 배최장근의 단면적을 cm^2 단위로 측정하며 배최장근 주위의 배다열근, 두반극근, 배반극근은 제외한다.

(3) 근내지방도

근내지방도는 배최장근 단면적 측정부위에서 지방침착도를 기준표(1 - 9 단계)에 따라 등급을 구분하며, 육색, 지방색, 조직감, 성숙도의 상태에 따라 근내지방도의 등급이 하향될 수 있다.

마. 혈통자료

한우의 유전능력 평가를 위해 표현형 정보를 갖는 개체들의 선조 혈통을 추적하여 총 23,296개의 혈통정보를 수집하여 이용하였다.

2. 한우의 기존(EBV) 및 유전체(GEBV) 육종가 추정 방법

가. 유전체 데이터 품질평가

유전체 데이터의 품질평가는 크게 두 개로 개체와 SNP마커에 대한 품질평가로 진행하였다.

(1) 개체 품질평가

- (가) Genome studio를 통해 생성된 유전자형 데이터(A, T, G, C, Missing)를 이용하여 결측률이 10%이상인 개체,
- (나) 혈통정보를 이용하여 구성한 혈연계수행렬(NRM)과 유전체정보를 이용하여 구성한 혈연계수행렬(GRM)간 비교를 통하여 오류를 발생시키는 개체,
- (다) 유전체 정보를 활용하여 멘델리안 불일치 오류를 발생시키는 개체를 확인한 후 제거 또는 수정을 실시하였다.

(2) SNP마커 품질평가

- (가) Genome studio를 통해 생성된 유전자형 데이터(A, T, G, C, Missing)를 이용하여 결측률이 10%이상인 마커,
- (나) 모두 동형인 마커 (다형성이 없는 마커),
- (다) Minor allele frequency(MAF)가 5%이하인 마커,
- (라) Hardy-Weinberg equilibrium test에서 p-value가 10^{-6} 이하인 마커,
- (마) 다른 마커와 상당히 높은 상관을 갖는 마커에 대하여 품질평가 실시하였다.

나. 표현형 및 혈통 데이터 점검

(1) 표현형 데이터 점검

각 형질별 기초통계량 분석 및 분포도 그래프를 이용하여 이상치 자료를 제거하였는데 여기서 유효자료를 결정하기 위한 유효범위는 아래와 같다.

- (가) 등심단면적 : 40 - 123cm²
- (나) 등지방두께 : 1 - 35mm
- (다) 근내지방도 : 1 - 9점

(2) 혈통 데이터 점검

혈통의 오류를 규명하기 위하여 아래와 같은 사항에 대하여 점검을 실시하였다.

- (가) 등록번호 규칙 오류 (9자리)
- (나) 성별 오류
- (다) 세대순 오류
- (라) 개체와 부 또는 개체와 모의 등록번호가 동일한 오류

다. 유전체 혈연계수 행렬 구성

(1) 혈통정보를 이용한 혈연계수 행렬

수집된 집단 전체의 혈통정보를 Tabular method(Emik and Terrill, 1949)를 이용하여 혈연계수 행렬을 구성하고 이 중 유전체 정보를 형성한 개체들의 혈연계수를 따로 출력하여 유전체 관계 행렬과 비교할 수 있도록 하였다.

(2) 유전체 정보를 이용한 혈연계수 행렬

유전체 혈연계수 행렬은 전장의 SNP정보를 이용하여 다음과 같이 실시하였다.

개체의 수를 n , 좌위(locus)의 수를 m 이라고 하면, 행의 크기는 n 이고 열의 크기는 m 인 M 의 행렬이 된다. 추정 목적에 따라 $MM'(n*n)$ 또는 $M'M(m*m)$ 의 형태로 사용할 수 있다. M 행렬의 각 원소는 소수 대립유전자(Miner allele)의 동형접합체에서 -1 , 다수 대립유전자(Major allele)의 동형접합체의 경우 1 , 그리고 이형접합체는 0 로 구성된 행렬이다. MM' 행렬을 구성하면 대각선의 값은 각 개체에 대한 동형접합체의 수이며, 비대각선의 값은 두 개체 간 공동으로 소유하는 대립유전자의 수이다. 이에 반해서 $M'M$ 의 대각선은 각 좌위에 대한 동형접합체의 수, 비대각선은 서로 다른 좌위에서 공동으로 소유하는 대립유전자의 수이다. i 번째 좌위에 두 번째 대립유전자의 빈도를 p_i 라고 하고, P 행렬의 i 번째 열을 $2(p_i - 0.5)$ 로 정의하면 P 는 0 을 평균으로 갖는 기대유전자형에 대한 빈도 값이 된다. 관측된 M 의 대립유전자 행렬에서 기대대립 유전자형에 대한 값으로 나타나는 P 에 대한 행렬의 값을 빼 주게 되면 대립유전자 효과의 평균이 0 이 되는 (Centered value) Z 행렬을 구성할 수 있다. 위에서 계산된 Z 행렬을 이용하여 아래와 같은 수식에 의하여 GRM 행렬을 구성할 수 있다(VanRaden, 2007, VanRaden, 2008).

$$P_{ij} = 2(p_i - 0.5)$$

$$Z = M - P$$

$$G = \frac{ZZ'}{2\sum p_i(1-p_i)}$$

라. 유전모수, 육종가 추정 및 정확도 비교

여러 형질을 동시에 고려한 다형질 평가모형을 아래와 같이 설정하여 유전 및 환경 (공)분산성분과 유전모수를 REML방법으로 추정하였다.

$$y = Xb + Za + e$$

여기서 y 는 도체형질(등심단면적, 등지방두께, 근내지방도)에 대한 관측치 벡터이고, b 는 고정효과에 대한 벡터이며 a 는 해당개체의 육종가이며, e 는 잔차효과이다.

상기 모형에서 동물의 유전효과에 대한 다음의 2가지 가정을 설정하여 분석모형을 설정하였다.

가정1: $a \sim MVN(0, G \otimes V)$

가정2:

여기서 G 는 혈통정보에 의한 혈연계수행렬

$$H = \begin{bmatrix} A_{12}A_{22}^{-1}G^*A_{22}^{-1}A_{21} & A_{12}A_{22}^{-1}G^* \\ G^*A_{22}^{-1}A_{21} & G^* \end{bmatrix}, A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, G^* = G - A_{22}$$

여기서 A_{11} 은 유전체 정보를 갖지 않은 개체 간의 혈통정보에 의한 혈연관계행렬, A_{22} 는 유전체 정보를 갖는 개체들의 혈통에 의한 혈연정보행렬, G 는 유전체 혈연관계행렬로 정의된 것이다.

상기의 가정1과 가정2에서 잔차의 분포는 $e \sim MVN(0, H \otimes R)$ 으로 가정하였으며 형질간의 상관은 없는 것으로 가정하였다.

또한 상기 가정 2에서 개체들의 육종가를 유전체 정보를 갖지 않은 개체(a_1)과 유전체 정보를 갖는 개체(a_2)로 구분하여 분포특성에 대한 가정은 다음과 같다.

$$\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \sim MVN\left(0, \begin{bmatrix} A_{11} + A_{12}A_{22}^{-1}(G - A_{22})A_{22}^{-1}A_{21} & A_{12}A_{22}^{-1}G \\ GA_{22}^{-1}A_{21} & G \end{bmatrix} \otimes V\right)$$

상기의 가정1에서 임의효과에 대한 분산성분 추정은 기존의 육종학적 관행에 의하였으며 가정2의 경우 다음과 같이 정규방정식을 설정하였다.

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + H^{-1} \otimes V^{-1} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} X'R^{-1}y \\ Z'R^{-1}y \end{bmatrix}$$

여기서 ,

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

유전평가의 신뢰도(r^2)는 예측분산오차(Prediction error variance; PEV)의 값을 이용하여 계산하였으며, 수식은 아래와 같다.

$$r^2 = 1 - \left(\frac{PEV}{\sigma_a^2} \right)$$

3. 고밀도(HD) 및 저밀도(LD) chip에 대한 유전체 육종가 정확도 비교

한우에서 수집된 Illumina Bovine 50K SNP chip과 Bovine HD SNP chip의 유전체 정보를 이용하여 1) 50K정보를 700K(HD)정보로 확대하는 Imputation연구를 수행하였고, 이 때 Imputation의 정확도 및 이를 활용한 유전체 선발의 정확도를 분석함으로써 기존 50K chip과의 정확도 차이를 비교하였으며, 2) 50K 유전체 정보에서 연관불평형의 값을 이용하여 Tag SNP을 선별하여 저밀도(8K) SNP 데이터를 구성하였다. 또한 유전체 육종가 계산을 통하여, 8K, 50K, 700K간 정확도 비교를 통해 유전체 선발에서 저밀도 칩(8K)의 활용 가능성에 대해 알아보았는데 이에 대한 모식도는 그림 25과 같다.

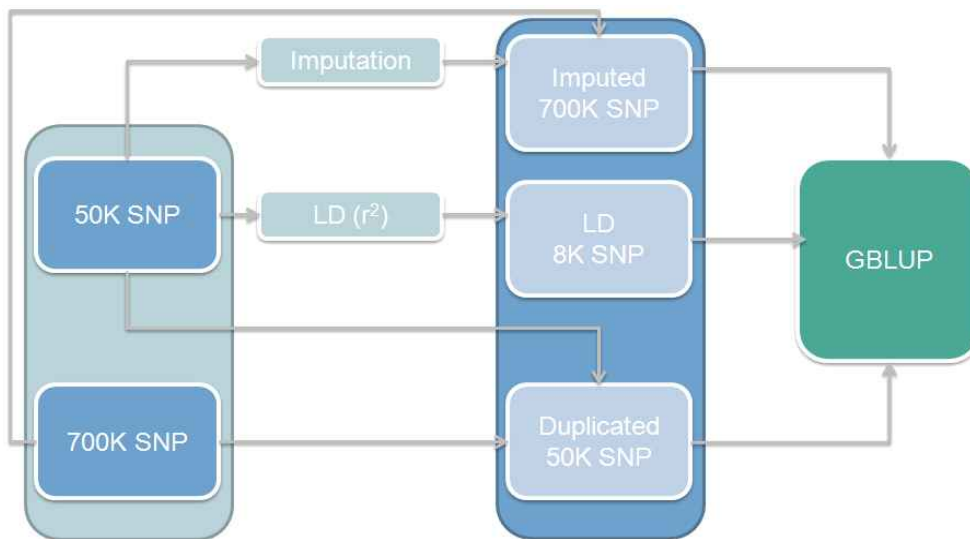


그림 25. 서로 다른 SNP정보를 이용한 유전체 육종가 비교 프로세스

가. 대치(Imputation)을 이용한 고밀도 유전체 데이터 생성

초기 Imputation의 용도는 유전자형이 결정되지 않은 즉, 결측된 유전자형에 대하여 통계적으로 추정함으로써 대치하는데 활용되었다. 그러나 최근 분자생물학기술에 다양한 분야의 기술이 접목됨에 따라 대용량의 유전정보를 작은 유전체 칩 안에 담을 수 있게 되었으며, 분석기간 또한 단축됨으로써 대용량의 유전체 정보를 얻기에 용이하게 되었다. 반면 기존의 분석한 저밀도칩으로 분석된 개체들의 유전체 정보를 활용하기 어려워짐에 따라 이들의 정보를 집단유전학적 개념에서 연관불평형 및 가계혈통 등의 정보를 이용하여 Imputation하는 연구가 다수 수행되었으며, 비록 참조집단의 크기에 영향 받지만 Imputation의 정확도는 분석방법의 향상에 따라 상당히 높은 것으로 보고되고 있다.

한우의 HD(700K) 정보 갖는 664두의 참조집단 축군을 이용하여 50K칩으로 분석된 942두의 검정축군을 대상으로 FImpute program(Sargolzaei 등. 2011)으로 Imputation하였다. Imputation의 정확도를 알아보기로 HD칩으로 분석된 664두 중 임의로 50두를 선별하여 검정축군으로 설정하였으며, 이들의 참값과 추정된 Imputation간 일치율을 통해 정확도(Accuracy)를 계산하였는데 이에 대

한 수식은 아래와 같다.

$$ACC_{\text{impute}}(\%) = \frac{N_{\text{imputed} = \text{observed}}}{N_{\text{total}}} \times 100$$

나. Tag SNP을 이용한 저밀도 유전체 데이터 생성

Tag SNP은 Genome상에 haplotype상의 유사한 SNP간 높은 연결성(LD)에 의해 그룹을 나타내는 것 중 대표적인 SNP를 지칭한다. 이들 Tag SNP은 특정 그룹에 대한 유전적 변이를 대표하기 때문에 표현형과의 연관성 분석에서 모든 SNP정보를 대치하여 대표의 Tag SNP을 통해 분석이 가능하기 때문에 이를 통해 분석시간 및 SNP분석 비용 감소가 가능한 장점이 있기 때문에 50K SNP정보에서 이웃된 SNP간 분석된 연관불평형의 값을 바탕으로 연관성이 높은 SNP에 대해 제외해 나감으로써 Tag SNP을 선별하였다.

(1) 연관불평형 추정

일반적으로 연관불평형의 크기는 서로 다른 두 대립유전자가 서로 연관되어 유전되는 정도를 나타내는 값으로 D 를 표준화시킨 D' (Lewontin, 1964) 또는 r^2 (Hill, 1968)로 추정할 수 있다. 그러나 D' 을 통한 연관불평형 추정은 집단 크기가 작거나 대립유전자빈도가 적을 경우 과대추정(Overestimate)될 수 있기 때문에(Hayes, 2008; McRae 등. 2002) r^2 를 이용하여 연관불평형을 추정하였다.

동일 염색체에 존재하는 두 개의 대립유전자가 존재하는(biallelic) 서로 다른 두 좌위 A, B의 연관불평형 크기(r^2)는 다음과 같이 계산하였다.

$$r^2 = \frac{D^2}{P(A_1) \times P(A_2) \times P(B_1) \times P(B_2)}$$

여기서, $p(A_1)$ 과 $p(A_2)$ 및 $p(B_1)$ 과 $p(B_2)$ 는 집단 내 A 또는 B 두 좌위의 각 대립유전자 빈도를 나타내며, D 는 다음과 같다.

$$D = p(A_1B_1) \times p(A_2B_2) - p(A_1B_2) \times p(A_2B_1)$$

여기서, $p(A_1B_1)$ 과 $p(A_2B_2)$ 및 $p(A_1B_2)$ 과 $p(A_2B_1)$ 는 집단 내 A 및 B 두 좌위의 각 대립유전자로 구성된 일배체형(Haplotype)의 빈도를 나타낸다.

두 좌위가 각각 동형이거나 두 좌위 중 하나의 유전자형이 동형일 경우 일배체형의 빈도를 계산할 수 있으나, 두 좌위 모두 이형일 경우(double heterozygotes)는 DNA chip의 분석 결과로는 상인상태(A_1B_1/A_2B_2), 상반상태(A_1B_2/A_2B_1)를 구분하지 못하기 때문에 EM algorithm (Expectation Maximization Algorithm)를 이용하여 A_1B_1/A_2B_2 와 A_1B_2/A_2B_1 의 조건부 확률을 계산하여 변화량이 일정량 이하(10^{-5})감소할 때까지 반복 연산하여 연관불평형의 값을 추정하였다 (Excoffier and Slatkin, 1995).

II. 결과 및 고찰

4. 유전체 정보를 이용한 육종가(GEBV) 추정 및 기존 육종가(EBV)와 비교

가. 한우 혈통 및 검정자료에 대한 기초통계분석 및 환경요인 분석

표 6에는 표현형정보를 통해 추적된 세대별 두수 및 근친된 개체수에 대하여 표기하였다. 23,296두 중 3,116두가 근친이 되어 있었으며 이들에 대한 평균 근친도는 1.94%으로 추정되었다.

표 14. 한우의 세대별, 성별, 근친된 개체 수

Ancestor	Sex				No. of inbreeding
	Male	Female	Steer	Total	
14	1	1	0	2	0
13	2	2	0	4	0
12	1	3	0	4	0
11	14	26	0	40	0
10	33	71	0	104	0
9	51	153	0	204	1
8	80	335	0	415	3
7	84	598	0	682	3
6	96	1,065	0	1,161	11
5	93	1,921	0	2,014	33
4	121	3,186	0	3,307	64
3	156	4,610	0	4,766	200
2	301	4,791	0	5,092	591
1	0	0	5,501	5,501	2,210
Total	1,033	16,762	5,501	23,296	3,116

검정자료가 수집된 5,641두에 대하여 혈통정보를 추적한 결과 3대 혈통 정보를 모두 아는 개체가 5,586두로 약 99.0%였고, 3대의 혈통정보 중 모를 모르는 개체는 1두, 외조부-외조모를 모르는 개체가 55두로 전체적으로 혈통등록 및 기록이 매우 우수하게 나타났다 (표 15, 표 16).

표 15. 도체검정이 수집된 개체의 알려진 부모 정보

분류	빈도	비율	누적빈도	누적비율
3대 혈통 정보를 모두 아는 경우 (* , **, ****)	5,586	98.99	5,586	98.99
모와 외조부,외조모를 모르는 경우 (* , *- , **--)	1	0.02	5,587	99.01
부와 조부,조모를 모르는 경우 (* , -* , --**)	1	0.02	5,642	99.03
외조부, 외조모를 모르는 경우 (* , **, **--)	55	0.97	5,697	100.00

(* , **, **** : 개체, 부-모, 조부-조모-외조부-외조모)

표 16. 검정차수별 혈통 정보

검정차수	3대 혈통 완벽 (* , **, ****)	모와 외조부, 외조모를 모르는 경우 (* , *- , **--)	부와 조부, 조모를 모르는 경우 (* , -* , --**)	외조부, 외조모 모름 (* , **, **--)	계
23	106	13			119
24	181	7			188
25	166	10			176
26	164	8			172
27	79	2			81
28	143			1	144
29	116				116
30	144				144
31	100				100
32	158	2			160
33	146	1			147
34	143				143
35	146				146
36	148				148
37	140				140
38	161				161
39	183				183
40	184				184
41	173	3			176
42	176				176
43	199	4			203
44	191	1			192
45	191				191
46	204				204
47	172				172
48	186	1			187
49	168				168
50	199		1		200
51	197				197

52	186		186
53	113		113
54	202		202
-	420	2	422

(* , ** , **** : 개체, 부-모, 조부-조모-외조부-외조모)

표 17 에 도체자료가 수집된 5,641두에 대하여 혈통 추석한 결과에 대하여 나타났다. 그 결과 표현형이 수집된 개체들은 총 597두의 종모우와 5,323두의 종빈우로부터 태어난 자손으로 나타났다. 한우 종모우별 낭우에 대한 두수를 보면 최소 1두에서 최대 36두까지 자손을 생산하였으며, 평균 종모우당 약 9.45두의 자손을 갖는 것으로 나타났다. 또한 종빈우는 대부분 1두의 자손을 가졌으며, 최대 4두까지 자손을 갖는 종빈우가 1두 존재하는 것으로 나타났다.

표 17. 도체자료가 수집된 개체에 대한 종축 당 자손수에 대한 빈도표 (표현형데이터 참고)

No. of offspring per sire	Count	No. of offspring per dam	Count
1	9	1	5,023
2	10	2	283
3	21	3	16
4	34	4	1
5	28		
6	32		
7	54		
8	68		
9	70		
10	63		
11	55		
12	39		
13	30		
14	30		
15	13		
16	13		
17	4		
18	5		
19	2		
20	1		
21	5		
22	1		
23	1		
24	2		
26	2		
27	2		
28	1		
29	1		
36	1		
Number of sire	597	Number of dam	5,323

한우 24 - 54차 후대검정우 5,641두의 도체형질에 대한 기초통계량은 표 18에 나타났다. 한우 후대검정우의 도축월령은 22개월에서 26개월사이에 도축되었으며 평균 24개월령에 도축되는 것으로 나타났고, 등심단면적, 등지방두께, 근내지방도의 평균값은 각각 $78.7 \pm 9.2 \text{cm}^2$, $8.6 \pm 3.7 \text{mm}$, $3.2 \pm 1.6 \text{score}$ 로 분석되었다.

표 18. 한우 도체형질에 대한 기초통계량

Traits	No.	Mean	S.D.	Min.	Max.
EMA	5,214	78.7	9.2	40	123
BF	5,219	8.6	3.7	1	35
MS	3,383	3.2	1.6	1	9

개체의 유전능력 분석을 위하여 도체형질에 영향을 미치는 환경요인을 알아보려고 분산분석을 실시하였으며 각 형질에 영향을 미치는 요인별 자승평균 및 유의수준에 대하여 표 19에 나타났다. 한우 후대검정 대상우는 모두 수소이기 때문에 성별효과는 고정되어 있었으며, 동기우군 효과로 검정차수에 대한 효과가 유의수준 1% 이내에서 영향을 미치는 것으로 나타났으며, 도축 시 개체 나이에 대한 공변이 효과도 등지방두께를 제외하고 1% 유의수준에서 도체형질에 영향을 미치는 것으로 분석되었다.

표 19. 한우에서 도체형질에 대한 요인별 분산분석표

Source	DF	EMA	BF	MS
Batch ¹⁾	369	482.0 **	51.8 **	37.0 **
Age ²⁾	1	449.5 **	0.2 ^{NS}	20.0 **

¹⁾Batch : 검정차수-검정지-도축일; ²⁾Age : 도축월령; NS : $p > 0.05$; ** : $p < 0.01$;

나. 한우 도체형질에 대한 유전모수 추정

한우의 유전능력 평가 모형 설정을 위하여 분산분석을 실시한 결과 동기우군효과(검정차수-검정지-도축일)과 도축시 나이가 영향하는 것으로 나타나 유전분석모형에 고정효과로 포함시켰다. 유전체 정보의 유무에 따라 또는 유전체 정보 밀도에 따라 비교하는 연구에서 유전모수의 값이 조건에 따라 변경될 경우 육종가의 수치 및 육종가 정확도에 영향을 미치기 때문에 유전모수의 추정은 유전체 정보 없이 표현형과 혈통정보를 기반으로 REML방법에 의하여 추정하였으며, 그 결과에 대하여 표 20에 나타냈다. 본 연구에서 추정된 유전분산 결과를 살펴보면, 등심단면적에서 27.18, 등지방두께 5.50, 근내지방도 1.44로 추정되었으며 잔차분산의 경우 위와 동일한 형질에 대하여 각각 34.97, 5.80, 1.01로 나타났다. 또한 유전력을 살펴보면 등심단면적에서 0.44, 등지방두께에서 0.49, 근내지방도에서 0.59로 추정되었다. 이는 현행 국가단위 후보 및 보증씨수소 선발에서 사용하고 있는 도체형질에 대한 유전력 (등심단면적 0.41, 등지방두께 0.40, 근내지방도 0.50)과 다소 차이를 보이기는 했으나, 2010년에 평가된 결과로 고려하여 볼 때 (국립축산과학원, 2014) 충분히 변화할 수 있는 차이로 판단된다. 따라서 여기서 추정된 유전모수의 값을 이용하여 개체의 육종가 평가에 활용하였다.

표 20. 한우의 도체형질에 대한 유전모수 추정

	REML		
	EMA	BF	MS
Genetic (co)variances			
EMA	27.18	-2.20	1.86
BF	-2.20	5.50	0.03
MS	1.86	0.03	1.44
Residual (co)variances			
EMA	34.97	2.13	0.68
BF	2.13	5.80	0.41
MS	0.68	0.41	1.01
Heritability	0.44	0.49	0.59

다. SNP마커의 품질평가

본 연구에서 Illumina BovineSNP50 V1, V2로 959두의 시험축군에 대하여 유전체 데이터를 수집하였으며, 720두는 Illumina Bovine HD Beadchip을 이용하여 유전정보 데이터를 수집하였다. 우선 개체단위에서 유전자형의 결측률이 10%이상인 개체가 73두 존재하여 제거하고 1,606두에 대하여 SNP단위의 품질평가를 실시하였다. 한우의 유전체 데이터는 서로 다른 세 종류의 SNP chip로 chip간 중복되는 SNP에 대하여 조사하였으며 그 결과는 표 21에 나타났다. 50K Version 1(V1)의 경우 54,001개의 SNP가 존재하였으며, Version 2(V2)는 54,609개, HD는 777,962개의 SNP가 존재하였다. 50K V1과 50K V2간 중복되는 SNP의 수를 확인해 본 결과 48,186개의 SNP를 공유하고 있었으며, 이 중 염색체 정보 및 거리정보가 누락된 마커(1,479개)와 성염색체 마커(692개)를 제외하고 46,015개의 마커가 1차적으로 선정되었다. 46,015개의 SNP와 HD간 겹치는 SNP를 조사하여 본 결과 41,878개가 있었으며, 이 중 다형성이 없는 마커, Minor allele frequency 가 5%이하인 마커, 결측률이 10%이상인 마커, high correlation된 마커, Hardy-Weinberg equilibrium 검정의 p-value가 10^{-6} 보다 적은 마커 15,267개를 제외하고 남은 30,748개의 유전체 정보가 다음 분석에서 사용되었다.

표 21. 한우 50K SNP의 유전자형 자료 및 전처리

자료 및 전처리	Number of SNP
BovineSNP50 beadchip version1	54,001
BovineSNP50 beadchip version2	54,609
중복 SNP(version1, version2)	48,186
염색체 정보 및 거리정보 결측 ^①	1,479
성염색체(X, Y) ^②	692
품질평가 (MAF, HWE 등) ^③	15,267
유용 SNP*	30,748

※ 유용 SNP : ①-②-③

라. 유전체 혈연계수 행렬을 통한 품질평가

한우의 성염색체를 제외한 나머지 상염색체 내 30,748개 좌위의 유전자형 정보를 이용한 개체간 혈연계수의 추정치(Genomic relationship matrix; GRM)와 혈통정보를 기반으로 추정된 혈연계수값(Numerical relationship matrix; NRM)을 이용하여 이들 간 산포도 및 회귀식을 추정하였다(그림 26). 그림 26에 제시된 바와 같이 부와 자식간 혈연관계(NRM)는 $0.5 \pm \alpha$ 으로 나타나는 반

면에 유전체 정보를 이용한 혈연계수 행렬의 값은 약 40% - 60%로 NRM값을 중심으로 정규분포 형태의 변이가 존재함을 알 수 있는데, 이는 유전체 정보를 이용하면 혈연계수에 대한 기댓값이 아닌 멘델리안 샘플링의 효과까지 고려하여 실제 개체들 간의 혈연관계의 크기를 계산하기 때문에 혈통정보에서 동일한 혈연계수의 크기를 갖는 개체 간에도 변이가 발생하고 있음을 알 수 있었다.

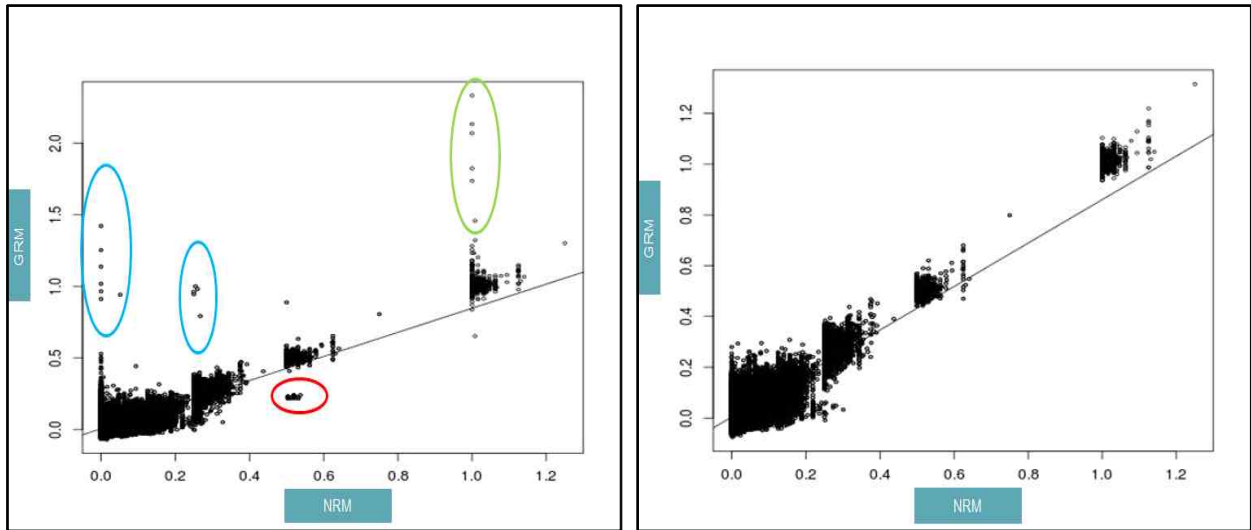


그림 26. 혈통정보를 이용한 혈연계수행렬(NRM)과 유전체 정보를 활용한 혈연계수행렬 (GRM)간 비교 (왼쪽 그래프 : 혈통오류 수정 전, 오른쪽 그래프 : 혈통오류 수정 후)

또한 유전체 정보를 이용하여 혈통정보 기반의 혈연계수 추정치와 비교를 통하여 잘못된 혈통정보를 인지할 수도 있는데, 예를 들면 그림 26의 왼쪽 그래프에서, NRM이 0.2이하이면서 GRM이 약 0.5값으로 나타난 개체 조합이 다수 나타났다. 이는 실제 혈연관계가 없다고 혈통상에 기록되어 있으나 실제로는 부-자손간 개체의 조합으로 추측할 수도 있으며, 이와 반대 경우로 혈통정보에서는 부-자손으로 기록된 개체 조합이 유전정보를 이용하였을 때는 혈연관계가 약 0.1이하로 미약한 것으로 나타났는데 이는 실제 전형매 또는 부모-자손간 관계가 아님에도 불구하고 혈통기록상 $0.5 \pm a$ 의 혈연관계로 예측된 경우도 있었다. 또한 혈통기록상에 혈연관계가 없다고 나타났으나 유전체정보를 통한 결과에서는 $1.0 \pm a$ 의 값을 갖는 것으로 보아 이는 동일한 개체 또는 오류의 데이터 일 것으로 추정되었다. 또한 혈통정보에서는 반형매로 추정되었는데 유전체 상에서는 전형매 이상의 혈연관계 크기를 갖거나 반대의 경우(유전체에서 반형매, 혈통에서 전형매), 개체 자신에서 1보다 높게 나타나는 개체가 존재하였다. 위의 다양한 이상치를 근거로 오류를 발생시키는 개체 18두를 찾아냈으며, 이 중 혈통오류를 발생시키는 개체가 13두 존재하였고 동일개체가 5두 존재하였다. 혈통오류를 발생시키는 13두 중 11두에 대해서는 친부를 확인하여 혈통기록을 변경하였으며, 나머지 혈통오류를 발생시키는 2두에 대해서는 혈통기록 제거를 하였으며, 동일개체 5두에 대해 유전체 데이터를 제거하고 남은 1,601두와 이들이 갖는 30,748개의 유전체 데이터를 이용하여 유전체 유전분석을 실시하였다 (표 22).

표 22. 유전자형의 결측률이 10%이상인 개체 및 혈통정보 오류개체에 대한 품질평가(QC)

Category	Number of animal
Genotyped animals	1,679
Percentage of missing genotype in animal < 10 %	73
Pedigree information error (동일개체)	5
Selected animal	1,601

마. 유전체 정보 유무에 따른 추정육종가 정확도 추정 및 비교

(1) BLUP 및 ssGBLUP별 육종가의 정확도 추정 및 육종가 상관 비교

유전체 정보 없이 혈통 및 표현형 기반으로 개체의 육종가를 추정하였으며(BLUP), 혈통 및 표현형정보에 추가적으로 유전체 정보를 포함하여 육종가를 추정하였다 (ssGBLUP). 추정 육종가의 정확도는 자신의 표현형 정보 유무, 표현형 정보를 갖는 개체와 혈연관계 크기 등의 요인에 따라 영향을 받기 때문에 자신의 표현형 데이터를 보유하고 있는 개체 집단(Reference Population), 자신의 표현형 데이터 없이 반형매 자손의 기록을 갖는 개체 집단(Sire), 자신의 표현형 데이터가 없고 자식의 표현형 데이터가 없는 개체 집단(Candidate)로 구분하여 표 23에 나타냈다. 전체 1,601두 중 Sire 집단이 140두, Reference 집단이 1,180두, Candidate 집단이 281두로 분석되었으며, BLUP의 육종가추정치의 정확도에 대한 평균은 Sire집단에서 등심단면적의 경우 0.75, 등지방두께는 0.76, 근내지방도는 0.78로 추정되었고, Reference 집단에서 등심단면적 0.70, 등지방두께 0.72, 근내지방도 0.77로 나타났으며, Candidate집단에서 등심단면적 0.42, 등지방두께 0.43, 근내지방도 0.43로 분석되었다. 50K 유전체 정보를 이용하였을 때 Sire집단 및 Reference집단의 육종가추정치의 정확도에 대한 평균은 BLUP방법에 비하여 0 - 2%로 증가하는 것으로 분석되었으며, 유전체 선발에서 가장 강점인 자신의 표현형 데이터가 없는 Candidate집단에서는 BLUP방법에 비해 **5 - 6% 육종가 정확도가 향상**되는 것으로 분석되었다.

표 23. 도체 형질별 BLUP과 ssGBLUP 육종가 정확도 비교

	BLUP without genotypic data			GBLUP with 50K SNP chip (ssGBLUP)		
	Sire	Reference	Candidate	Sire	Reference	Candidate
N	140	1,180	281	140	1,180	281
EMA	0.75	0.70	0.42	0.76	0.71	0.47
BF	0.76	0.72	0.43	0.77	0.73	0.48
MS	0.78	0.77	0.43	0.8	0.77	0.49

(2) 검정(표현형)자료를 얻기 전과 후의 유전체육종가 추정치간 상관 비교

유전체 선발의 강점은 앞에서 나타난 육종가 정확도와 같이 표현형 데이터를 얻기 이전에 효과가 큰 것으로 알려져 있으며 본 결과 역시도 선행연구들과 유사한 패턴을 나타냈다. 유전체 정보를 한우의 보증씨수소가 선발되기 이전인 후보씨수소 선발 또는 당대검정우 매입시기에 활용한다고 가정하였을 때, 후보씨수소에 대하여 표현형 자료가 수집되기 이전과 수집된 이후 육종가 추정치 비교 연구하였는데, 이에 대한 수행을 위하여 53차 후대검정우 108두의 표현형 정보를 수집되기 이전과 수집된 후 각각에 대하여 개체의 육종가를 계산하고 이를 두 개의 육종가를 비교하여 그림 27과 표 24에 나타냈다.

그림 27에 왼쪽은 유전체 정보 없이 혈통과 표현형정보만을 이용하여 표현형 정보가 있을 때(y축)와 표현형 정보가 없을 때(x축) 108두의 후대검정우에 대한 각 형질별 육종가추정치에 대한 산포도를 제시하였다. 그 결과 등심단면적(EMA)에서 회귀계수 추정치는 0.86로 나타났으며, 상관은 0.37으로 추정되었다. 반면 동일한 형질에 대하여 유전체 정보를 포함하여 표현형 정보가 있을 때와 표현형 정보가 없을 때에 대하여 회귀계수를 추정한 결과 0.99로 매우 높은 계수값을 갖는 것으로 추정되었으며, 상관계수값은 0.52로 추정되었다. 등지방두께(BF)에서 유전체 정보 없이 육종가(EBV)를 비교하였을 때 회귀계수가 0.74로 추정된 반면에, 유전체 정보를 포함하여 추정된 육종가(GEBV)를 비교하였을 때는 0.91로 상대적으로 높게 추정되었다. 또한 상관계수 추정치는 0.44에서 0.67로 유전체 정보를 사용함으로써 0.23의 상관관계가 상승하는 결과를 얻었다. 근내지방도(MS)에서 유전체 정보 없이 육종가(EBV)를 비교하였을 때 회귀계수값은 0.72로 나타났으며, 유전체 정보를 포함한 유전체육종가(GEBV)를 비교하였을 때는 1.00로 거의 완벽한 회귀관계를 갖는 것으로 추정 되었다(그림 27).

유전체정보를 포함하여 표현형을 얻기 이전에 추정한 육종가와 표현형을 얻은 후에 추정한 육종가 간 상관성이 높다는 것은 조기에 개체를 선발하는 것이 가능하다는 것을 의미하는데, 이는 앞서 언급한 당대검정우 매입 및 후보씨수소 선발에 활용할 수 있는 가능성을 충분히 보여준 결과라고 판단되며, 차후 참조집단의 크기가 지금보다 더욱 더 누적될 경우 유전체 선발에 대한 파급효과는 더 커질 것으로 사료된다. 현행 후보씨수소 선발 시 혈통지수에 입각한 근내지방도 형질이 고려되고 있는데 이 시기에 유전체 정보를 활용할 경우 지금보다 더 정확하게 개체를 선발해 나갈 수 있을 것으로 판단된다.

표 24. 표현형 정보 얻기 이전과 이후의 육종가추정치에 대한 상관

Trait	$cor(EBV_{wp}, EBV_{wop})^{1)}$	$cor(GEBV_{wp}, GEBV_{wop})^{2)}$	Diff. $(GEBV_{cor}, EBV_{cor})^{3)}$
EMA	0.37	0.52	0.15
BF	0.44	0.67	0.23
MS	0.34	0.61	0.27

¹⁾ $cor(EBV_{wp}, EBV_{wop})$: 표현형정보가 없을 때와 표현형정보가 있을 때 육종가(EBV)의 상관

²⁾ $cor(GEBV_{wp}, GEBV_{wop})$: 표현형정보가 없을 때와 표현형정보가 있을 때 유전체 육종가(GEBV)의 상관

³⁾ Diff. $(GEBV_{cor}, EBV_{cor})$: 유전체 육종가(GEBV)와 육종가(EBV)의 정확도 차이

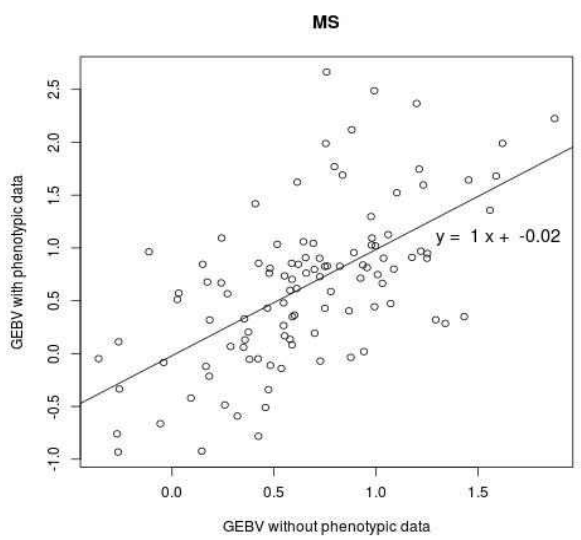
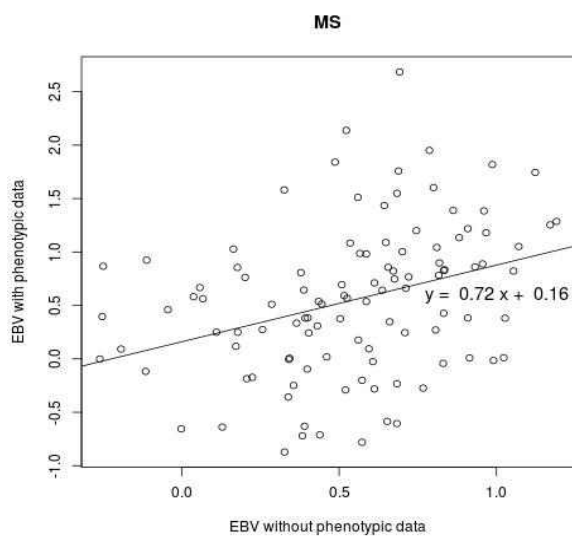
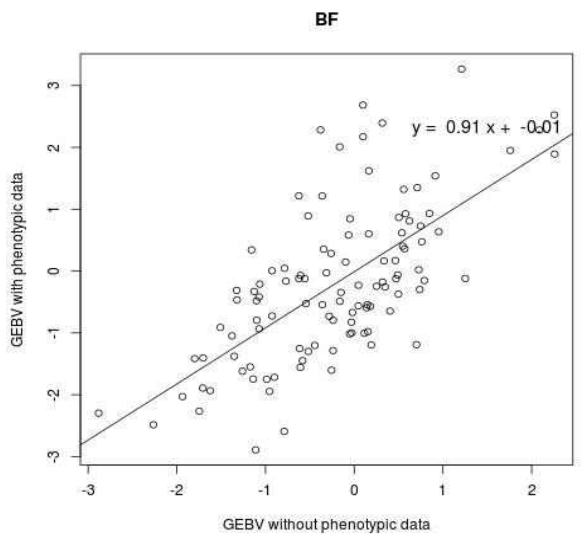
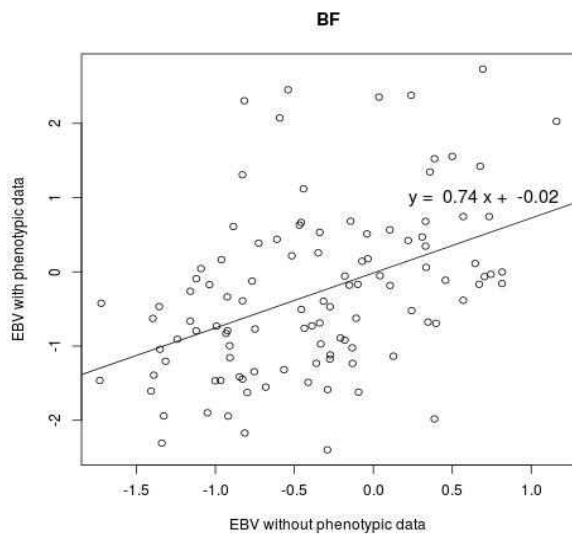
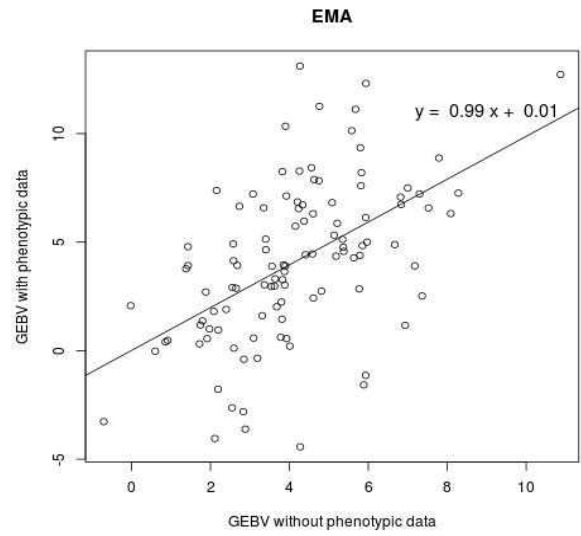
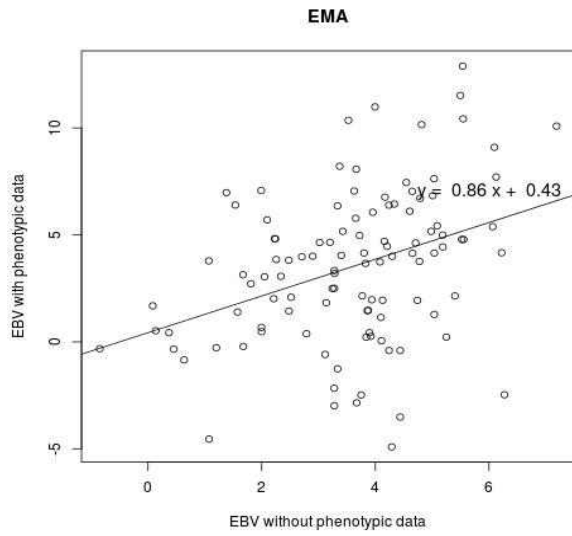


그림 27. 표현형 정보 얻기 이전과 이후의 유전체 육종가 상관 Graph (BLUP : 왼쪽, ssGBLUP : 오른쪽)

5. 고밀도(HD) 및 저밀도(LD)칩에 대한 유전체 육종가 정확도 비교

가. HD(700K) chip을 이용한 유전체 육종가 분석

(1) 품질평가

유전체 정보를 수집한 개체들 중 혈통 및 유전자형 결측률이 높은 개체를 제외한 나머지 1,601두에 대하여 50K 정보를 갖는 개체는 887두이었고 HD정보를 갖는 개체 714두이었는데, Imputation실시 이전에 분석에 유효하지 않은 마커를 제외하기 위하여 품질평가를 실시하였다.

HD chip은 전체 777,762개의 SNP정보를 담고 있으며, 이 중 성염색체를 제외한 29개 상염색체 내 735,239개의 SNP정보가 분석에 이용되었다. 이 중 다형성이 없는 마커(45,722개), Minor allele frequency 가 5%이하인 마커(113,936개), 결측률이 10%이상인 마커(5,702개), high correlation된 마커(203,985개), Hardy-Weinberg equilibrium 검정의 p-value가 10^{-6} 보다 적은 마커(24개)를 제외하고 365,870개의 유전체 정보가 Imputation분석 및 유전체 육종가 분석에 이용되었는데, 그림 28에 50K와 HD chip에 염색체별 유전정보에 대해서 제시하였다.

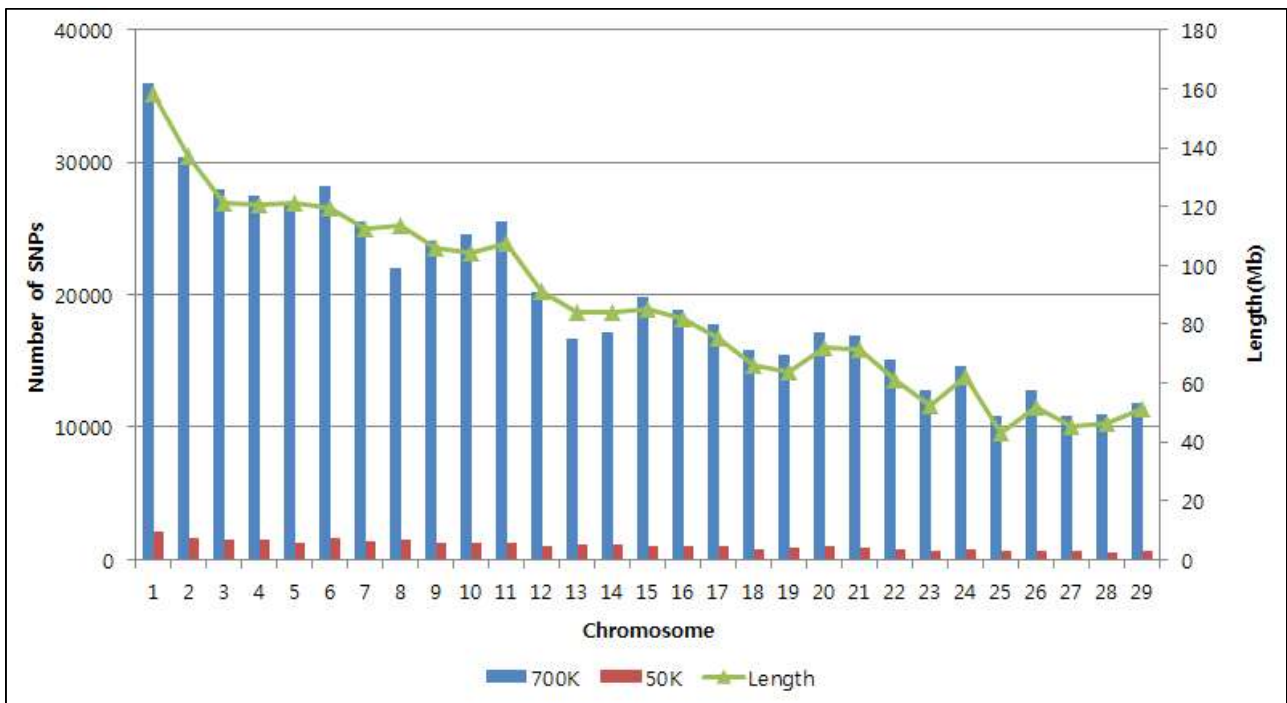


그림 28. 한우 50K 및 HD(700K) chip의 염색체별 마커수 및 염색체 길이

(2) SNP Genotype Imputation 정확도

Imputation에서 고밀도의 칩(HD)을 갖는 개체의 집단이 참조집단이며 참조집단 크기는 Imputation의 정확도에 영향을 미친다. HD chip을 갖는 714두 중 임의로 50두를 선정하여 50K SNP chip에서 얻은 정보와 중첩되는 SNP정보만을 선정하여 후, Imputation한 후에 실제 HD SNP 정보와 비교하는 방식으로 Imputation의 정확도를 추정하였는데, 그 결과는 표25에 나타났다. 표 25에 제시된 바와 같이 임의로 선별된 50두에 대한 Imputation 정확도는 0.925 - 0.992로 분포하였으며, 평균 0.964의 정확도로 상당히 높은 정확도로 Imputation됨을 볼 수 있었다.

표25. 50K chip에서 HD chip으로 Imputation시 정확도

	N	MEAN	SD	MIN	MAX
Accuracy	50	0.964	0.017	0.925	0.992

(3) Imputation정보(HD chip)를 이용한 유전체 육종가 정확도 비교

유전체 육종가 정확도를 자신의 표현형 데이터를 보유하고 있는 개체 집단(Reference population), 자신의 표현형 데이터 없이 반형매 자손의 기록을 갖는 개체 집단(Sire), 자신의 표현형 데이터가 없고 자식의 표현형 데이터가 없는 개체 집단(Candidate)로 구분하여 표 26에 나타냈다. Imputation된 유전체 육종가 추정치의 정확도는 Sire집단에서 등심단면적 0.75, 등지방두께 0.77, 근내지방도 0.79로 나타났으며, Reference집단에서 등심단면적 0.71, 등지방두께 0.73, 근내지방도 0.77로 나타났으며, Candidate집단에서 등심단면적 0.46, 등지방두께 0.47, 근내지방도 0.48로 분석되었다. 50K 유전체 정보를 이용한 결과와 비교하여 보았을 때, Sire, Reference, Candidate집단 모두 정확도의 큰 차이가 없는 것으로 분석되었다.

표 26. 50K와 Imputed SNP chip (HD)간 유전체 육종가 정확도 비교

	ssGBLUP with 50K SNP chip			ssGBLUP with Imputed HD SNP(700k)		
	Sire	Reference	Candidate	Sire	Reference	Candidate
N	140	1180	281	140	1180	281
EMA	0.76	0.71	0.47	0.75	0.71	0.46
BF	0.77	0.73	0.48	0.77	0.73	0.47
MS	0.80	0.77	0.49	0.79	0.77	0.48

나. 저밀도(8K) chip을 이용한 유전체 육종가 분석

(1) LD chip 분석을 위한 SNP 선별

저밀도 SNP chip을 개발하는 방법에는 크게 세 가지 방법이 있을 수 있겠다. 첫째는 경제형질에 영향을 주는 유효 SNP를 찾아내고 이들 정보를 SNP chip안에 담음으로써 SNP의 수를 축소시키는 방법이 있을 것이고, 둘째는 유전적 거리정보를 이용하여 균등하게 SNP정보를 추출하는 방안이 있을 수 있으며, 셋째는 유전적 불평형의 크기에 따라 SNP정보를 추출하는 방법을 들 수 있다. 첫 번째 방법은 특정 경제형질에 의존되어 SNP chip이 개발될 수 있어 다른 형질에는 집 목하기 어려운 단점이 있기 때문에 대부분의 경제형질의 경우 다수의 유전자가 소수의 효과를 나타내기에 전체적인 분산을 설명하는데 어려움이 있는 단점이 있으며, 둘째 방법은 유전적 거리 정보를 기반으로 어느 형질에서도 적용할 수 있고, 차후 필요시 Imputation에도 활용할 수 있는 장점이 있으나 이것은 생물학적으로 비과학적인 단점이 있다. 따라서 본 연구는 Imputation 및 다수의 형질에 적용할 수 있는 특징의 연관불평형(LD) 추정치를 이용한 SNP 정보의 선별을 실시하였다.

저밀도 SNP chip 개발을 위한 SNP 선별을 위하여 50K 내 이웃하는 SNP마커 간 연관불평형

(LD)분석을 실시하였으며, 그래프는 그림 29와 같다. 그림 29에서는 염색체 1번부터 29번까지를 순차적으로 나타냈으며 x축은 SNP마커간 거리를, y축은 연관불평형(r^2)값으로 표현하였다. 그림에서 보는 바와 같이 전반적으로 가까운 거리에 있는 SNP간 연관불평형의 크기가 컸으며, SNP마커간 거리가 멀어지면 멀어질수록 연관불평형의 값이 적은 것으로 나타났다.

두 개의 SNP마커간 연관불평형의 크기가 크다는 것은 다음세대로 넘어갈 때 재조합이나 돌연변이 등의 요인 없이 같이 붙어서 다음 세대로 전달될 확률이 높음을 의미하며, 이들은 물리적으로 두 개의 SNP마커이나 하나처럼 효과를 나타낸다고 말할 수 있다. 따라서 이러한 점에 착안하여, SNP간 연관불평형이 높은 조합을 찾아내서 제거해나가는 순서로 진행함으로써 50K SNP 정보 중 7,801개의 SNP를 선별하였다.

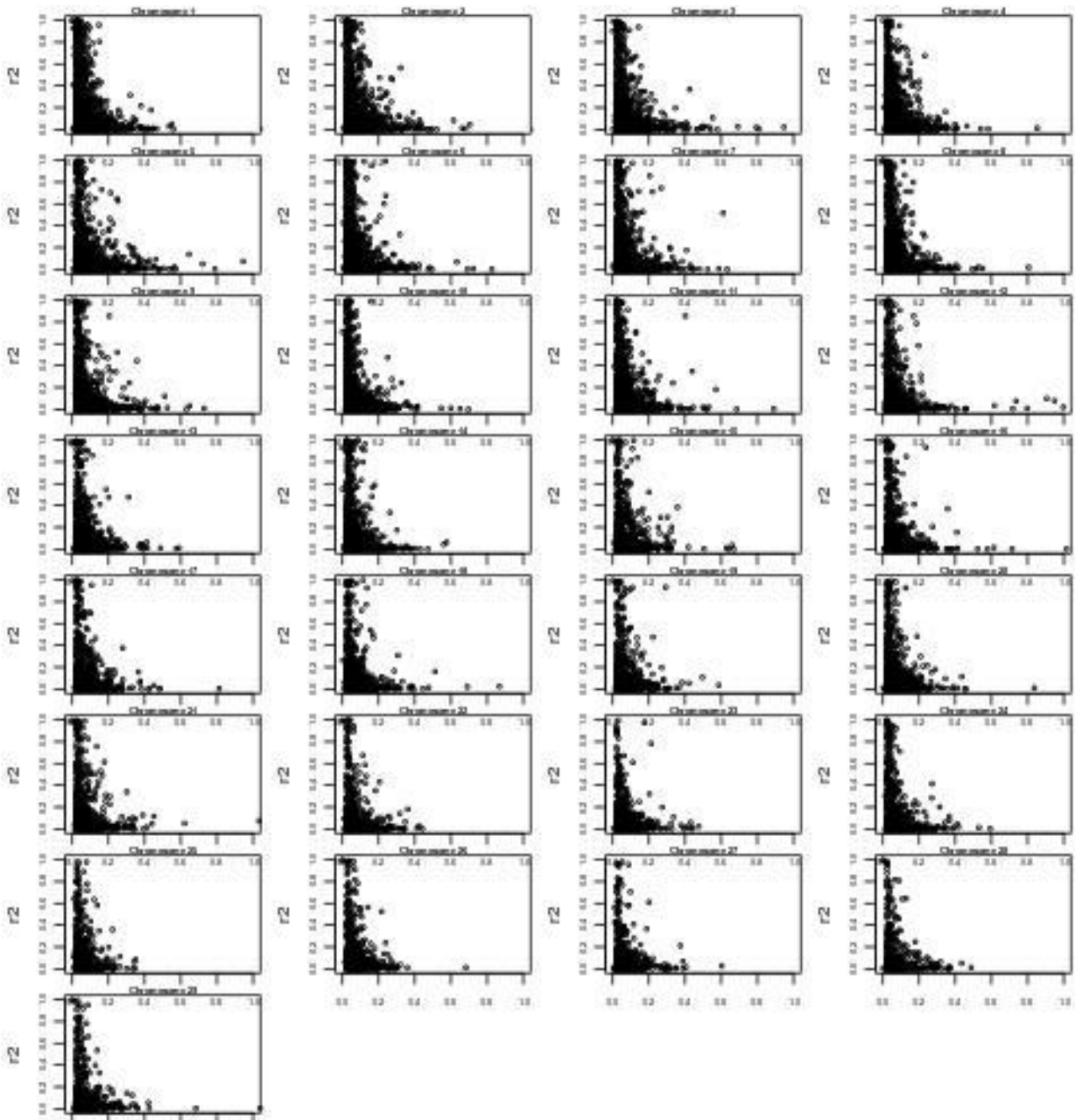


그림 29. 50K 정보를 이용한 염색체별 이웃하는 SNP간 연관불평형 분석

7,801개 축소된 SNP에 대한 연관불평형 그래프는 그림 30에 나타냈다. 그림 30은 그림 29와 다르게 거리가 가까운 SNP마커간 연관불평형의 크기가 많이 축소됨을 볼 수 있다. 다만 거리가 멀어지면 서도 연관불평형이 큰 일부의 SNP마커 조합이 보이게 되는데, 이는 50K SNP 정보를 이용한 LD 추정방법에서도 추정되었지만 이웃하는 SNP마커만 그래프에 표기하여 제시되지 않은 부분이다. 이러한 부분까지도 제거해 나가면 더 작은 수를 갖는 저밀도 칩을 개발할 수 있겠으나, SNP의 숫자가 너무 줄어들면 이 또한 차후 유전체 유전분석 정확도에 영향을 미칠 것으로 판단되어 제거하지 않았다.

또한 표 27에 염색체 1번에서 50K와 8K(저밀도) SNP chip간 유전적 거리별 연관불평형 크기의 평균에 대해서 나타냈다. 50K chip의 경우 거리가 0.05 Mb(Mega base pairs)이하에서 0.217의 평균 연관불평형 값을, 0.05 ~ 0.1Mb에서는 0.119, 0.1 ~ 0.2 Mb에서는 0.053로 분석된 반면 8K chip의 경우 동일한 거리에서 각각 0.041, 0.072, 0.056로 SNP 마커간 연관불평형이 50K chip에 비해 상대적으로 많이 낮게 나타났음을 볼 수 있다. 따라서 저밀도 SNP chip 개발을 위한 SNP선별은 연관불평형의 크기에 의해서 상관 높은 SNP를 균등하게 잘 선별했다고 판단된다.

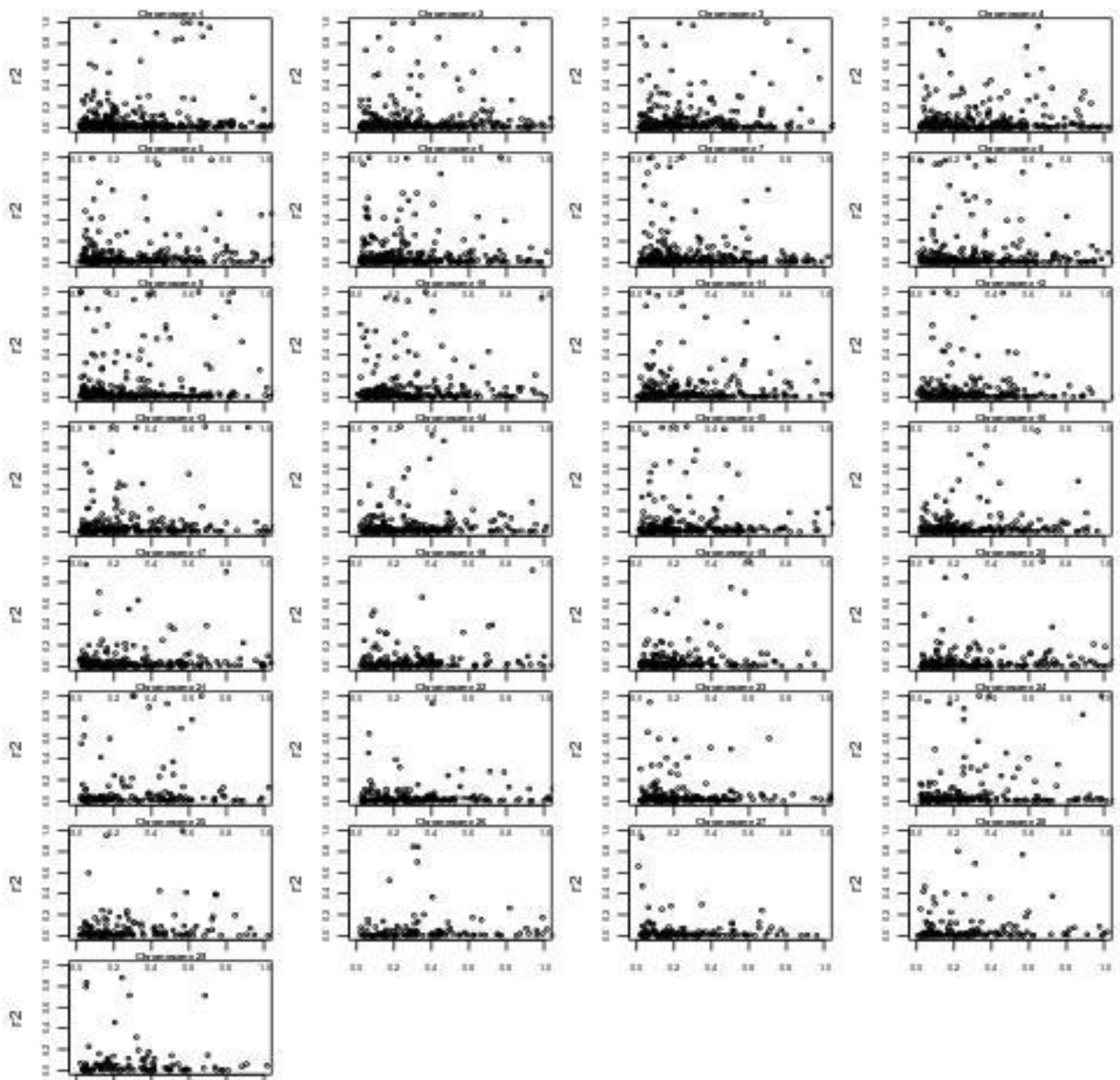


그림 30. 저밀도칩을 이용한 염색체별 이웃하는 SNP간 연관불평형 분석

표 27. 1번 염색체에서 유전 거리별 연관불평형(LD) 평균 값 비교

Distance(Mbp)	Mean of linkage disequilibrium(r^2)	
	50K SNP chip	8K SNP chip
0.00 ~ 0.05	0.217	0.041
0.05 ~ 0.10	0.119	0.072
0.10 ~ 0.20	0.053	0.056
0.20 ~ 0.30	0.038	0.059
0.30 ~ 0.40	0.036	0.051
0.40 ~ 0.50	0.015	0.049

(2) 품질평가

54,001개의 SNP 마커(BovineSNP50 V1) 및 54,609개 SNP 마커(BovineSNP50 V2)를 갖는 2종의 Illumina Bovine 50K chip에서 수집된 886개체의 50K SNP정보 중 서로 중복되는 48,186개의 마커를 선별 한 후 분석에 유효하지 않은 마커를 제외하기 위한 품질 평가를 실시하였다. 상기 선별된 마커 중 염색체 정보 또는 위치(거리) 정보가 없는 마커 1,479개를 제거하였고, 수소의 경우 X 염색체를 반수체 형태로 보유하기 때문에 유전 평가의 바이어스(bias) 발생을 줄이기 위해 X염색체 내 마커 692개를 제거하였다. 또한 Minor allele frequency가 5% 이하인 마커 및 다형성이 없는 마커 12,731개의 마커를 제거하였으며, 상기과정을 통해 총 33,284개의 SNP 마커를 선별하였다.

(3) 저밀도(8K) chip을 이용한 유전체 육종가 정확도 비교

유전체 육종가추정치의 정확도를 자신의 표현형 데이터를 보유하고 있는 개체 집단(Reference population), 자신의 표현형 데이터 없이 반형매 자손의 기록을 갖는 개체 집단(Sire), 자신의 표현형 데이터가 없고 자식의 표현형 데이터가 없는 개체 집단(Candidate)로 구분하여 표 28에 나타냈다. 저밀도(8K) chip으로 추정된 유전체 육종가 정확도는 Sire집단에서 등심단면적 0.77, 등지방두께 0.78, 근내지방도 0.80로 나타났으며, Reference 집단에서 등심단면적 0.72, 등지방두께 0.74, 근내지방도 0.78로 나타났으며, Candidate 집단에서 등심단면적 0.50, 등지방두께 0.52, 근내지방도 0.53로 분석되었다. 50K 유전체 정보를 이용한 결과와 비교하여 보았을 때, Sire, Reference 집단에서의 정확도는 큰 차이 없이 유사하게 나타났으나, 표현형 정보가 없는 Candidate집단에서 3 - 4%의 정확도 향상 효과를 나타냈다.

표 28. 50K와 저밀도(8K) chip 간 유전체 육종가추정치의 정확도 비교

	ssGBLUP with 50K SNP chip			ssGBLUP with LD chip (8K)		
	Sire	Reference	Candidate	Sire	Reference	Candidate
N	140	1180	281	140	1180	281
EMA	0.76	0.71	0.47	0.77	0.72	0.50
BF	0.77	0.73	0.48	0.78	0.74	0.52
MS	0.80	0.77	0.49	0.80	0.78	0.53

제 3 절 돼지 종축에 대한 유전체 정보 활용 분석기법 및 개량방안 연구

I. 재료 및 방법

1. 유전체(SNP) 정보 및 검정자료 수집

종돈의 현행 개량체계를 개선하고 개량효율성을 극대화하기 위해서 DNA상의 유전체 정보를 활용한 선발을 통해 유전평가의 정확도를 향상시키고 선발강도를 높임으로써 개량의 효율을 높이는 개선대책이 필요하므로, 이를 위한 연구를 수행하였다. 유전체 선발의 정확도는 형질의 유전력, 참조집단크기, QTL수, 유효집단크기, SNP밀도, 분석방법 등 다양한 요인에 따라 영향을 받는데 이 중 현실적으로 실현 가능하며 정확도에 영향을 많이 미치는 요인이 참조집단의 크기라고 판단된다. 참조집단은 일반적으로 자신의 표현형자료와 유전체정보를 모두 갖는 개체의 집단을 말한다. 본 연구에 개량의 목적이 되는 형질은 연령이 일정기간 경과되어 조사되며 한쪽 성에만 발현되고 후대검정이 지난한 모계종돈의 번식형질인 산자수의 개량이 대체적으로 유전력이 낮고 개량의 효율성이 낮기 때문에 유전체 정보를 이용한 선발방법을 규명하는데 대상형질로 정의하였다. 이러한 형질에 대한 종돈에서의 자료 수집은 선발된 모돈에서만 조사되기 때문에 모돈이 참조집단이 될 수 있고 웅돈의 경우에는 높은 혈연관계를 갖는 개체가 다수 존재할 수 있기 때문에 추가적으로 참조집단에 포함하였다. 또한 이러한 참조집단의 크기를 증가시키기 위하여 연간 약 200두 내외의 유전체자료 및 검정자료를 동일 돈군을 대상으로 5년간 수집하는 연구가 수행되었다.

가. 채혈 및 genomic DNA 추출

(1) 혈액 시료채취

SNP chip 분석을 위한 시료는 단양 GGP의 Landrace 종축를 이용하였으며, 경정맥으로부터 약 5ml를 항응고제(EDTA)가 들어있는 튜브에 채혈하였다. (그림 31) 각 시료에 대한 연도별 채취 현황은 표 29에 나타내었다.

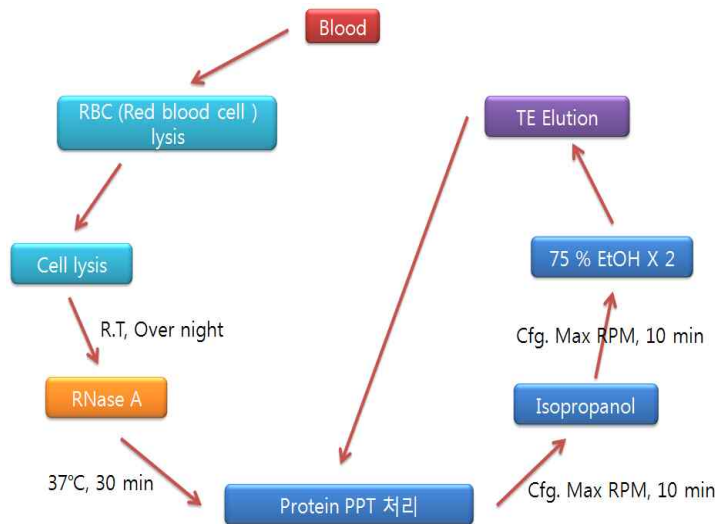


그림 31. 전혈로부터의 genomic DNA의 추출 과정

표 29. 연도별 후보종모우 및 후대검정우의 시료 채취 현황

	응돈	모돈	합계
2010년(1차년도)	21	245	266
2011년(2차년도)	29	221	250
2012년(3차년도)	22	200	222
2013년(4차년도)	9	155	164
2014년(5차년도)	161	341	502
총 합계	242	1,162	1,404

(2) SNP Chip 분석

Chip 분석을 위한 실험절차는 Illumina사에서 제시한 SNP chip 분석과정에 따라 진행하였다. Chip assay 과정에서 발생할 수 있는 오류를 최소화하기 위하여 실험날짜 및 실험자, 기계정보 등의 기본정보 및 사용된 시약의 Lot no, Chip number 정보와 chip 위에 올려진 sample 위치까지 기록 하면서 실험을 수행하였다. Chip 분석을 위한 실험절차는 3일간 진행되는 실험의 각 과정별 분석 내용은 그림 32와 같다.

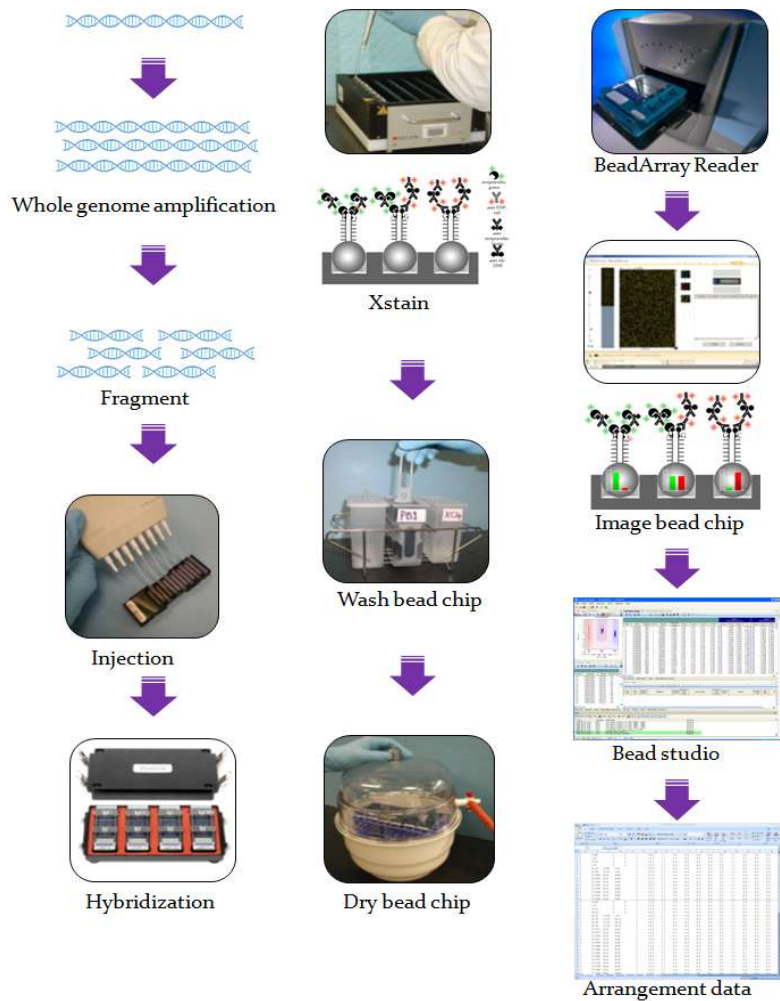
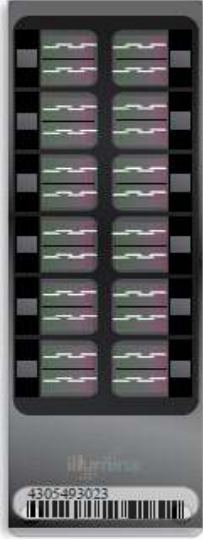


그림 32. SNP chip 분석과정 모식도(출처, Illumina, USA).

분석에 이용된 Chip은 약 60,000개의 단일염기다형(Single nucleotide polymorphism; SNP)을 갖는 Illumina PorcineSNP60 Beadchip(V1, V2)를 활용하여 개체별 대용량 유전자형(유전체) 분석을 실시하였다.

표 30. 버전에 따른 PorcineSNP60 Beadchip에 대한 비교

Chip version	PorcineSNP60 V1 Beadchip	PorcineSNP60 V2 Beadchip
No. of SNP	62,163	64,232
Chip Image		

(3) SNP chip으로부터 유전자형 스캔

분석이 완료된 Chip을 Scan 하게 되면 다음과 그림 33과 같은 이미지 파일과 idat 파일이 생성되며, 이미지 파일의 경우는 시료 표본별 green색과 red색의 두 가지 구별된 색상으로 유전자형을 판독하였다.

SNP Genotyping의 경우 GenomeStudio 프로그램을 이용하여 그림 34와 같이 감광 분석된 chip data를 염기로 구성된 유전자형으로 전환한다.



그림 33. Chip 9347037037의 스캔 사진

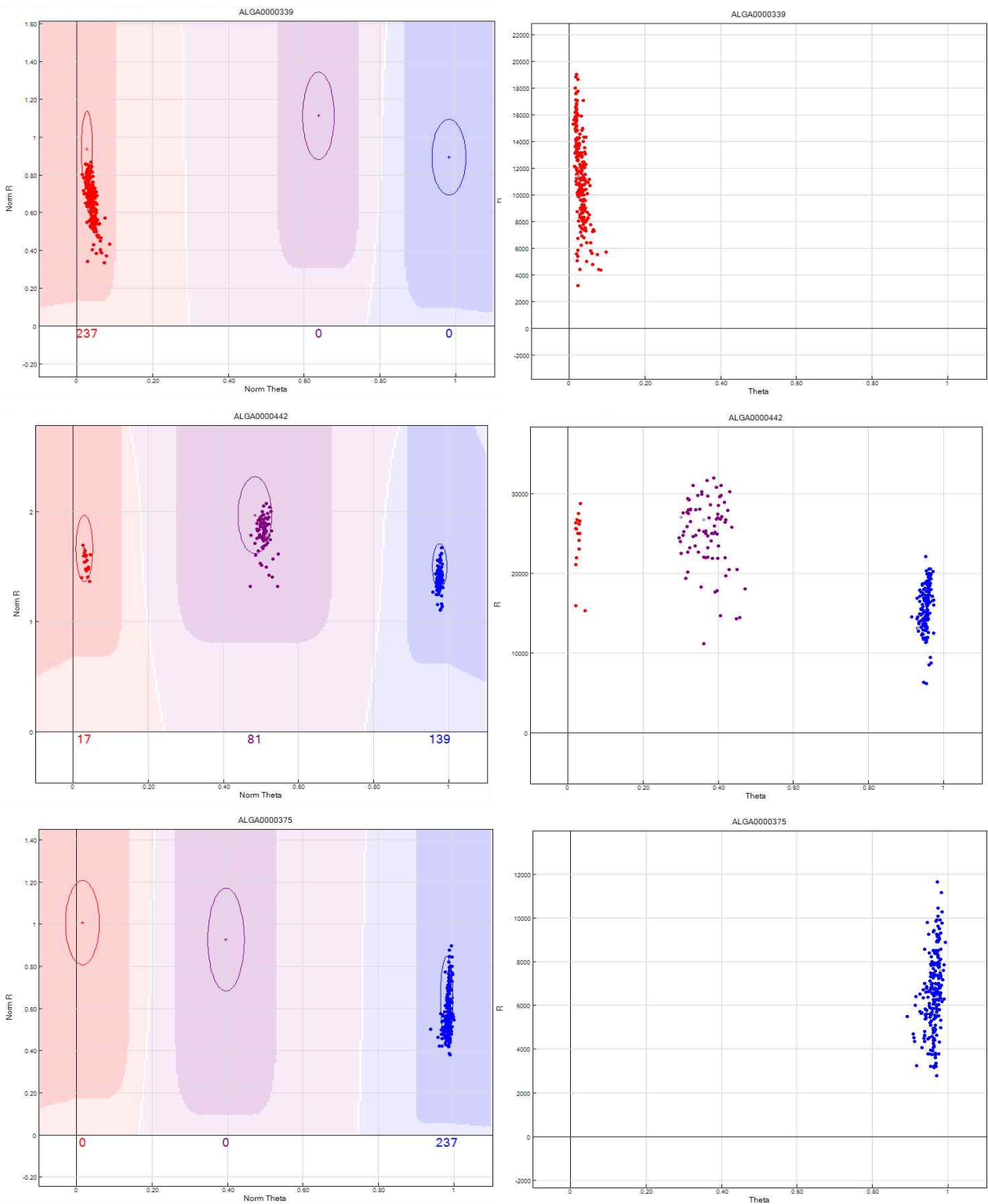


그림 34. GenomeStudio 프로그램을 이용한 SNP의 genotyping (AA Homotype 일 경우(상), AB Heterotype 일 경우(중), AA Homotype 일 경우(하))

나. 유전체자료 수집

유전자형으로 전환된 유전체자료의 결과는 표 31와 같으며, 5년간 돼지 Landrace 집단의 종축 1,061두의 SNP분석 정보는 표 32에 나타내었다.

표 31. PocineSNP60 beadchip 분석결과(일부)

Index	Name	Chr	Position	시료번호				
				LL09066048	LL08131059	LL09147068	LL09070019	LL06184064
1	ALGA0000009	1	52297	AG	GG	GG	GG	GG
2	ALGA0000014	1	79763	TC	CC	CC	CC	CC
3	ALGA0000021	1	209568	AC	AA	AA	AA	AA
4	ALGA0000022	1	292758	AG	AA	AA	AA	AA
5	ALGA0000046	1	747831	AG	GG	GG	GG	GG
6	ALGA0063147	11	53081518	TC	TC	CC	TC	CC
7	ALGA0063150	11	53093586	CC	CC	TC	CC	CC
8	ALGA0063183	11	53340428	TG	GG	TG	GG	TG
9	ALGA0063194	11	53364705	AG	GG	AG	GG	AG
10	ALGA0063203	11	53441272	AC	AA	AC	AA	AC
11	INRA0045306	14	86227451	TG	TG	GG	TT	GG
12	INRA0045347	14	87300576	TC	CC	TC	TC	TC
13	INRA0045367	14	88065435	TT	TT	TT	TT	TT
14	INRA0045371	14	88169829	CC	CC	CC	CC	CC
15	INRA0045383	14	89081607	TG	TT	TG	TT	TG
16	INRA0055464	18	11780172	TC	TC	TC	CC	CC

표 32. 연도별 SNP Chip 분석정보

년도	DNA sample				SNP 분석			비고
	모든	응돈	자돈	계	모든	응돈	계	
2010	245	21		266	219	21	240	PorcineSNP60 v1 Beadchip
2011	207	29		236	224	29	253	PorcineSNP60 v1 Beadchip
2012	200	22		222	172	15	187	PorcineSNP60 v1 Beadchip
2013	144	9	70	223	144	0	144	PorcineSNP60 v2 Beadchip
2014	341	161		501	214	26	237	PorcineSNP60 v2 Beadchip
계	1,137	242	70	1,448	973	91	1,061	

다. 표현형자료

대상 GGP 집단은 1997년 모계 순종돈(Landrace) 을 입식하였고 1998년 1월부터 순종돈 생산 및 생산된 순종 자돈을 대상으로 검정 및 선발과 선발 종돈을 활용한 교배 등, 폐쇄번식을 통하여 개량사업을 현재까지 추진하여 왔다. 산자수의 유전력은 약 0.12-0.15으로 매우 낮으며, 옹돈의 경우 개체 자신의 표현형이 없고 모돈의 경우 산차기록이 누적되어야 하므로 개량이 매우 어려운 형질 중 하나이지만 경제적으로 매우 중요한 가치를 지닌 형질이다. 이러한 산자수에 관련된 표지인자 효과를 추정하고 유전체 육종가를 추정하기 위하여 2,247두의 9,182복 산자수 표현형을 수집하였고 형질별 측정방법은 다음과 같다.

(1) 임신기간(GL)

모돈의 분만일로부터 해당 산차의 최종 교배일까지의 기간으로 정의하였으며 유산한 모돈의 임신기간은 분석에서 제외하였다.

(2) 총산자수1(TNB1) : 출생시 생존 + 사산

총산자수1은 복당 분만자돈에서 생존한 개체수와 사산 개체수를 합친 것으로 자궁 내에서 사망하여 미라가 된 개체를 제외한 산자수를 의미한다. 사산 개체는 분만 또는 임신 후기에 사망한 개체로 분만직전까지 생존산자일 가능성이 높으므로 개체의 가능 번식능력으로 판단할 수 있다.

(3) 총산자수2(TNB2) : 출생시 생존 + 사산 + 미라

총산자수2는 복당 분만자돈에서 생존한 개체수와 사산 및 미라 개체수를 합친 것으로 배란된 난자와 주입된 정자가 수정되어 자궁에 착상된 개체 수로 판단할 수 있으며, 개체의 배란 능력과 연관되었다고 할 수 있다.

(4) 생존산자수(NBA) : 출생시 생존산자수

생존산자수는 복당 분만자돈에서 실제적으로 생존한 개체 수만 측정한 것으로 산자수 중 경제성에 가장 밀접하게 연관되어 있다.

라. 혈통자료

돼지의 혈통 및 유전체 육종가 추정, 혈통오류 확인을 위하여 약 14대로 구성된 82,643개의 혈통정보를 수집하여 이용하였다.

2. 돼지의 기존(EBV) 및 유전체(GEBV) 육종가 추정 방법

가. 유전체 데이터 품질평가

유전체 데이터의 품질평가는 크게 두 개로 개체와 SNP마커에 대한 품질평가로 진행하였다.

(1) 개체 품질평가

- (가) Genome studio를 통해 생성된 유전자형 데이터(A, T, G, C, Missing)를 이용하여 결측률이 10%이상인 개체,
- (나) 혈통정보를 이용하여 구성한 혈연계수행렬과 유전체정보(NRM)를 이용하여 구성한 혈연계수행렬(GRM)간 비교를 통하여 오류를 발생시키는 개체,
- (다) 유전체 정보를 활용하여 멘델리안 불일치 오류를 발생시키는 개체를 확인한 후 제거 또는 수정을 실시하였다.

(2) SNP마커 품질평가

- (가) Genome studio를 통해 생성된 유전자형 데이터(A, T, G, C, Missing)를 이용하여 결측률이 10%이상인 마커,
- (나) 모두 동형인 마커 (다형성이 없는 마커),
- (다) Minor allele frequency(MAF)가 5%이하인 마커,
- (라) Hardy-Weinberg equilibrium test에서 p value가 10^{-6} 이하인 마커,
- (마) 다른 마커와 상당히 높은 상관을 갖는 마커에 대하여 품질평가 실시하였다.

나. 표현형 및 혈통 데이터 점검

(1) 표현형 데이터 점검

각 형질별 기초통계량 분석 및 분포도 그래프를 이용하여 이상치 제거를 실시하였다. 이상치를 제거하기 위한 각 형질별 유효자료의 범위는 아래와 같이 설정하여 이상치 자료를 제거하였다.

- (가) 임신기간 : 102 - 125일
- (나) 총산자수1 : 5 - 22두
- (다) 총산자수2 : 5 - 22두
- (라) 생존산자수 : 5 - 22두

(2) 혈통 데이터 점검

혈통정보에 대한 오류를 검정하기 위하여 아래의 사항에 대한 점검을 실시하였다.

- (가) 등록번호 규칙 오류 (9자리)
- (나) 성별 오류
- (다) 세대순 오류
- (라) 개체와 부 또는 개체와 모의 등록번호가 동일한 오류

다. 유전체 혈연계수 행렬 구성

(1) 혈통정보를 이용한 혈연계수 행렬

수집된 집단 전체의 혈통정보를 Tabular method(Emik and Terrill, 1949)를 이용하여 혈연계수 행렬을 구성하고 이 중 유전체 정보를 형성한 개체들의 혈연계수를 따로 출력하여 유전체 관계 행렬과 비교할 수 있도록 하였다.

(2) 유전체 정보를 이용한 혈연계수 행렬

유전체 혈연계수 행렬은 전장의 SNP정보를 이용하여 다음과 같이 실시하였다.

개체의 수를 n , 좌위(locus)의 수를 m 이라고 하면, 행의 크기는 n 이고 열의 크기는 m 인 M 의 행렬이 된다. 추정 목적에 따라 $MM'(n*n)$ 또는 $M'M(m*m)$ 의 형태로 사용할 수 있다. M 행렬의 각 원소는 소수 대립유전자(Miner allele)의 동형접합체에서 -1 , 다수 대립유전자(Major allele)의 동형접합체의 경우 1 , 그리고 이형접합체는 0 로 구성된 행렬이다. MM' 행렬을 구성하면 대각선의 값은 각 개체에 대한 동형접합체의 수이며, 비대각선의 값은 두 개체 간 공동으로 소유하는 대립유전자의 수이다. 이에 반해서 $M'M$ 의 대각선은 각 좌위에 대한 동형접합체의 수, 비대각선은 서로 다른 좌위에서 공동으로 소유하는 대립유전자의 수이다. i 번째 좌위에 두 번째 대립유전자의 빈도를 p_i 라고 하고, P 행렬의 i 번째 열을 $2(p_i - 0.5)$ 로 정의하면 P 는 0 을 평균으로 갖는 기대유전자형에 대한 빈도값이 된다. 관측된 M 의 대립유전자 행렬에서 기대대립 유전자형에 대한 값으로 나타나는 P 에 대한 행렬의 값을 빼 주게 되면 대립유전자 효과의 평균이 0 이 되는 (Centered value) Z 행렬을 구성할 수 있다. 위에서 계산된 Z 행렬을 이용하여 아래와 같은 수식에 의하여 GRM 행렬을 구성할 수 있다(VanRaden, 2007, VanRaden, 2008).

$$P = 2(p_i - 0.5) \dots \dots \dots \textcircled{1}$$

$$Z = M - P \dots \dots \dots \textcircled{2}$$

$$G = \frac{ZZ'}{2\sum p_i(1-p_i)} \dots \dots \dots \textcircled{3}$$

(가) G05

G05 방법은 수식 ②에서와 같이 M 에서 P 를 빼는 방식이 적용된다. P 행렬은 아래와 같이 i 번째 표지인자의 유전자 빈도와 기준 빈도 0.5 의 편차값에 2 를 곱해준 것이며, 이러한 경우 대립유전자효과의 평균값(mean value of the allele effect)이 0 으로 표준화된다(VanRaden, 2008). G05 방법으로 생성되는 유전체 관계행렬(G)은 ZZ' 행렬을 유전체 관계행렬의 척도(scale)인

$2\sum_{i=1}^m p_i(1-p_i)$ 로 나누게 된다.

$$P = 2(p_i - 0.5)$$

(나) GMF

GMF 방법은 G05방법과 매우 유사하나 수식①의 P 행렬을 구성할 때 기준빈도를 0.5 대신 모든 표지인자의 소수대립유전자빈도평균(\overline{MAF})을 이용하는 것으로 P 행렬은 다음과 같다.

$$P = 2(p_i - \overline{MAF})$$

(다) GOF

GOF 방법은 G05방법 및 GMF방법과는 다르게 유전체 관계행렬의 척도(scale)인 $2\sum_{i=1}^m p_i(1-p_i)$ 을 표지인자별로 세분화하여 관측 유전자 빈도(Observed Gene frequency)를 고려한 방법이며 다음과 같다.

$$GOF = ZDZ', \quad Z = M - P, \quad P = 2(p_i - 0.5),$$

여기서 D 행렬은 대각행렬이며 이 행렬의 i번째 원소 D_{ii} 는 아래와 같다.

$$D_{ii} = \frac{1}{m[2p_i(1-p_i)]}$$

(라) GOF*

GOF* 방법은 원리적으로 GOF방법과 유사하나 유전자 빈도의 분산을 위하여 베타함수를 적용한 것이며 다음과 같다.

$$GOF^* = \frac{ZZ'}{\left[(p_0 - q_0)^2 + \left(\frac{\sum_{i=1}^m p_i(1-p_i)}{m} \right) \left(\frac{\alpha + \beta + 2}{\alpha + \beta} \right) \right] m}, \quad Z = M - P, \quad P = 2(p_i - 0.5)$$

여기서 α 와 β 는 Beta process $B(\alpha, \beta)$ 의 두 모수이며 아래와 같이 추정하였다.

$$\hat{\alpha} = \bar{p} \left(\frac{\bar{p}(1-\bar{p})}{m} - 1 \right), \quad \hat{\beta} = (1-\bar{p}) \left(\frac{\bar{p}(1-\bar{p})}{m} - 1 \right)$$

여기서 \bar{p} 는 평균 대립유전자 빈도이며 m은 표지인자수이며, p_0, q_0 는 다음과 같다.

$$E(p) = \frac{\alpha}{(\alpha + \beta)} = p_0, \quad E(q) = \frac{\beta}{(\alpha + \beta)} = q_0$$

(마) GN

GN 방법은 근교되지 않은 혈통 혈연계수행렬과 유전체 관계행렬을 유사하게 만들기 위하여 대각행렬값을 1에 가깝게 표준화한 것이며, 다음과 같다.

$$GN = \frac{ZZprime}{\text{trace}[ZZ']/n}, Z = M - P, P = 2(p_i - 0.5)$$

여기서 n 은 유전체 정보를 소유한 개체수이며 **trace**는 대각행렬의 합이다.

각 유전체 관계행렬은 개체 간의 유사성이 기대치 보다 크게 나타나는 경우에 비정칙행렬 (Singular Matrix)이어서 역행렬 계산이 불가능하므로 혈통정보를 이용한 혈연계수행렬(A)에 가중치($w=0.95$)를 부여하여 다음과 같이 혈연계수행렬을 변경하였다.(Aguilar 등, 2011)

$$G = wG + (1 - w)A, G = G05, GMF, GOF, GOF*, GN$$

각 유전체 관계행렬과 혈통을 이용한 혈연관계행렬을 구성하기 위하여 관련 프로그램을 gFortran 언어를 이용하여 직접 개발하였으며, 혈통정보를 이용한 혈연계수 A는 Henderson(1976)이 제안한 순환참조방법(recursive method)를 적용한 작표 방법(tabular method)으로, 각 유전체 관계행렬(G)은 Aguilar(2011)의 ORIG 방법에 Forni(2011)가 제시한 5가지 방법을 적용하여 개발하였다.

라. 유전모수, 육종가 추정 및 정확도 비교

여러 형질을 동시에 고려한 다형질 평가모형을 아래와 같이 설정하여 유전 및 환경 (공)분산성분과 유전모수를 REML방법으로 추정하였다.

$$y = Xb + Za + e$$

여기서 y 는 도체형질(등심단면적, 등지방두께, 근내지방도)에 대한 관측치 벡터이고, b 는 고정효과에 대한 벡터이며 a 는 해당개체의 육종가이며, e 는 잔차효과이다.

상기 모형에서 동물의 유전효과에 대한 다음의 2가지 가정을 설정하여 분석모형을 설정하였다.

가정1: $a \sim MVN(0, G \otimes V)$

가정2: $a \sim MVN(0, H \otimes V)$

여기서 G 는 혈통정보에 의한 혈연계수행렬

$$H = \begin{bmatrix} A_{12}A_{22}^{-1}G^*A_{22}^{-1}A_{21} & A_{12}A_{22}^{-1}G^* \\ G^*A_{22}^{-1}A_{21} & G^* \end{bmatrix}, A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, G^* = G - A_{22}$$

여기서 A_{11} 은 유전체 정보를 갖지 않은 개체 간의 혈통정보에 의한 혈연관계행렬, A_{22} 는 유전체 정보를 갖는 개체들의 혈통에 의한 혈연정보행렬, G 는 유전체 혈연관계행렬로 정의된 것이다.

상기의 가정1과 가정2에서 잔차의 분포는 $e \sim MVN(0, H \otimes R)$ 으로 가정하였으며 형질간의 상관은 없는 것으로 가정하였다.

또한 상기 가정 2에서 개체들의 육종가를 유전체 정보를 갖지 않은 개체(a_1)과 유전체 정보를 갖는 개체(a_2)로 구분하여 분포특성에 대한 가정은 다음과 같다.

$$\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \sim MVN\left(0, \begin{bmatrix} A_{11} + A_{12}A_{22}^{-1}(G - A_{22})A_{22}^{-1}A_{21} & A_{12}A_{22}^{-1}G \\ GA_{22}^{-1}A_{12} & G \end{bmatrix} \otimes V\right)$$

상기의 가정1에서 임의효과에 대한 분산성분 추정은 기존의 육종학적 관행에 의하였으며 가정2의 경우 다음과 같이 정규방정식을 설정하였다.

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + H^{-1} \otimes V^{-1} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} X'R^{-1}y \\ Z'R^{-1}y \end{bmatrix}$$

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

유전평가의 신뢰도(r^2)는 예측분산오차(Prediction error variance; PEV)의 값을 이용하여 계산하였으며, 수식은 아래와 같다.

$$r^2 = 1 - \left(\frac{PEV}{\sigma_a^2} \right)$$

3. 저밀도(6K) chip 선별 방법 및 유전체 육종가 정확도 비교

돼지 Landrace 품종에서 수집된 Illumina PorcineSNP60 beadchip과 60K 유전체 정보에서 연관불평형의 값을 이용하여 Tag SNP을 선별하여 저밀도(6K) SNP 데이터를 구성하고 유전체 육종가 계산을 통하여, 6K와 60K간 정확도 비교를 통해 유전체 선발에서 6K chip의 활용 가능성에 대해 알아보고자 실시하였으며 자세한 모식도는 그림35에 나타났다.

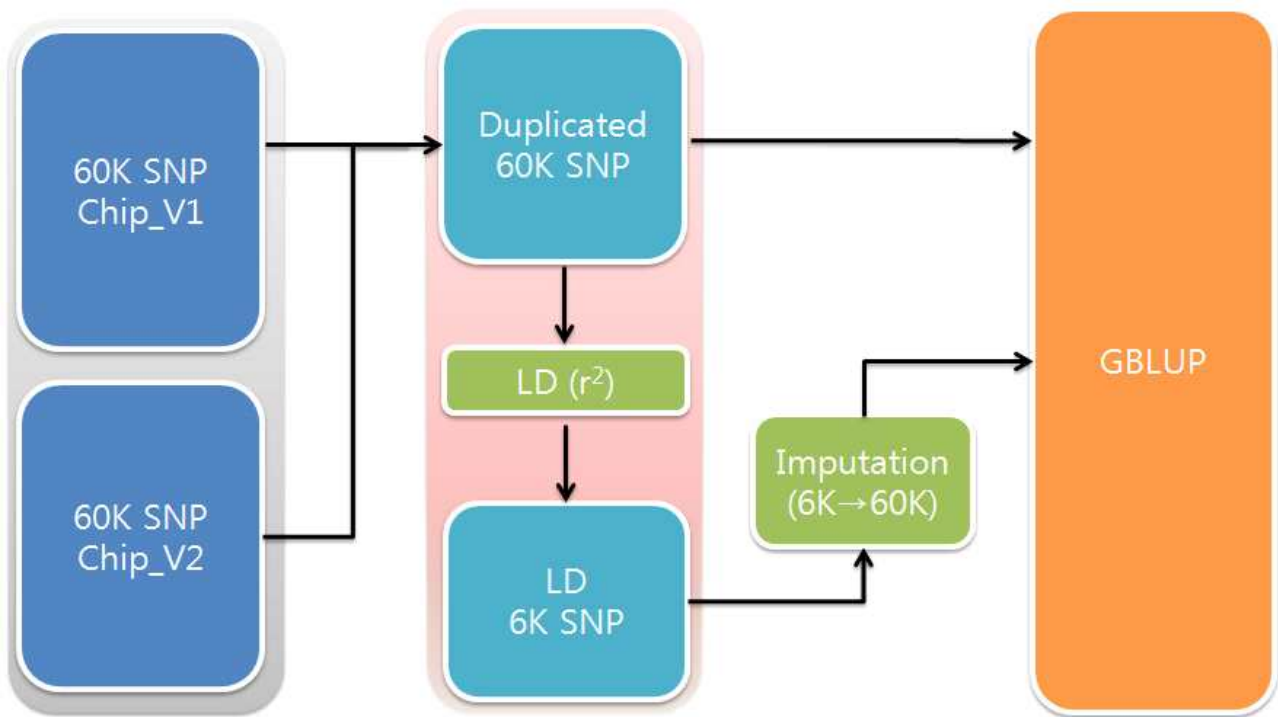


그림35. 돼지의 유전체 육종가 추정과정

가. Tag SNP을 이용한 저밀도(6K) 유전체 데이터 생성

Tag SNP은 게놈상에 haplotype상 유사한 SNP간 높은 연결성(LD)에 의해 그룹을 나타내는 것 중 대표적인 SNP를 지칭하며, 이들 Tag SNP은 특정 그룹에 대한 유전적 변이를 대표하기에 표현형과 연관성 분석에서 모든 SNP사용 대신에 대표의 Tag SNP을 통해 분석이 가능하며 이를 통해 분석시간 및 SNP분석 비용 감소가 가능하다. 60K SNP에서 이웃된 SNP간 분석된 연관불평형의 값을 바탕으로 연관성이 높은 SNP에 대해 제외해 나감으로써 Tag SNP을 선별하였다.

(1) 연관불평형 추정

일반적으로 연관불평형의 크기는 서로 다른 두 대립유전자가 서로 연관되어 유전되는 정도를 나타내는 값으로 D를 표준화시킨 D' (Lewontin, 1964) 또는 r^2 (Hill, 1968)로 추정할 수 있다. 그러나 D'을 통한 연관불평형 추정은 집단 크기가 작거나 대립유전자빈도가 적을 경우 과대추정(Overestimate)될 수 있기 때문에(Hayes, 2008; McRae 등. 2002) r^2 를 이용하여 연관불평형을 추정하였다.

동일 염색체에 존재하는 두 개의 대립유전자가 존재하는(biallelic) 서로 다른 두 좌위 A, B의 연관불평형 크기(r^2)는 다음과 같이 계산하였다.

$$r^2 = \frac{D^2}{P(A_1) \times P(A_2) \times P(B_1) \times P(B_2)}$$

여기서, $p(A_1)$ 과 $p(A_2)$ 및 $p(B_1)$ 과 $p(B_2)$ 는 집단 내 A 또는 B 두 좌위의 각 대립유전자 빈도를 나타내며, D는 다음과 같다.

$$D = p(A_1B_1) \times p(A_2B_2) - p(A_1B_2) \times p(A_2B_1)$$

여기서, $p(A_1B_1)$ 과 $p(A_2B_2)$ 및 $p(A_1B_2)$ 과 $p(A_2B_1)$ 는 집단 내 A 및 B 두 좌위의 각 대립유전자로 구성된 일배체형(Haplotype)의 빈도를 나타낸다.

두 좌위가 각각 동형이거나 두 좌위 중 하나의 유전자형이 동형일 경우 일배체형의 빈도를 계산할 수 있으나, 두 좌위 모두 이형일 경우(double heterozygotes)는 DNA chip의 분석 결과로는 상인상태(A_1B_1/A_2B_2), 상반상태(A_1B_2/A_2B_1)를 구분하지 못하기 때문에 EM algorithm (Expectation Maximization Algorithm)를 이용하여 A_1B_1/A_2B_2 와 A_1B_2/A_2B_1 의 조건부 확률을 계산하여 변화량이 일정량 이하(10^{-5})감소할 때까지 반복 연산하여 연관불평형의 값을 추정하였다 (Excoffier and Slatkin, 1995).

II. 결과 및 고찰

4. 유전체 육종가(GEBV) 및 혈통 육종가(EBV) 비교

가. 돼지 혈통 및 검정자료에 대한 기초통계분석 및 환경요인 분석

혈통자료 분석을 통하여 계산된 출생년도별 근친도는 표 33, 34와 같다.

표 33. Landrace품종(모돈)의 출생년도별 근친도에 대한 기초통계량(단위 %)

출생년도	두수	평균값	표준편차	최소값	최대값
1998	159	1.55	1.69	0	7.72
1999	116	1.63	2.11	0	13.11
2000	123	2.82	2.27	0	9.41
2001	142	3.41	2.27	0	9.26
2002	126	3.88	1.81	0.24	9.74
2003	126	4.56	1.60	0.24	9.31
2004	108	5.10	1.42	2.36	9.84
2005	114	6.33	2.39	2.65	16.38
2006	105	6.75	2.36	3.54	17.25
2007	143	7.30	2.37	4.40	20.87
2008	128	6.71	1.17	5.28	11.53
2009	141	7.58	2.31	5.06	18.63
2010	172	7.72	1.26	5.54	13.12
2011	158	8.30	1.37	6.04	13.38
합계	1,861				

표 34. Landrace품종(웅돈)의 출생년도별 근친도에 대한 기초통계량(단위 %)

출생년도	두수	평균값	표준편차	최소값	최대값
1998	8	0.96	0.74	0.12	2.08
1999	17	0.72	0.92	0	2.87
2000	12	2.73	1.78	0	6.12
2001	15	2.71	2.22	0	6.50
2002	11	4.00	1.24	2.02	5.91
2003	19	4.14	1.76	0.44	7.39
2004	22	5.59	1.77	1.98	9.84
2005	12	5.34	1.14	3.71	6.91
2006	22	5.58	0.96	4.00	7.69
2007	19	7.87	3.55	4.67	20.87
2008	16	7.73	2.11	5.43	12.19
2009	22	7.12	1.01	5.74	9.23
2010	23	7.79	1.54	5.89	13.52
2011	22	8.65	1.70	6.39	12.74
2012	6	9.18	1.66	7.18	11.15
합계	246				

오랜 기간 폐쇄된 집단에서 교배되어 연도별로 근교계수가 증가가고 있으며, 전체 집단 의 근 교계수는 약 0.08이상으로 상당히 높은 근친도를 보이고 있었으며 개체들간의 비교적 높은 혈연 관계를 갖는 개체들로 구성되어 있음을 알 수 있었다.

GGP 랜드레이스 순종돈 집단의 산자수 및 산차의 기초통계량은 표 35에 나타내었다. 모돈의 임신기간의 평균은 114.17로 나타났으며, 생존산자수 평균은 10.67, 미이라를 제외한 총산자수 평 균은 11.35, 미이라를 제외한 총산자수는 11.67로 나타났다.

표 35. 모돈의 번식형질에 대한 기술통계량

Trait	Mean	SD	Min	Max
GL	114.17	1.28	110	119
NBA	10.67	2.54	5	20
TNB1	11.35	2.70	5	21
TNB2	11.67	2.80	5	22

개체의 유전능력 분석을 위하여 도체형질에 영향을 환경요인을 알아보고자 분산분석을 실시 하였으며 각 형질에 영향을 요인별 자승평균 및 유의수준에 대하여 표 36에 나타냈다

표 36. 번식형질에 영향을 환경요인에 대한 분산분석

effect	Landrace				
	df	GL	NBA	TNB1	TNB2
PAR	10	13.69**	120.51**	145.23**	170.54**
YS	204	5.57**	9.50**	9.46**	10.42**
SS	309	2.88**	7.88**	9.31**	9.71**
Error	7,099	1.19	6.03	6.84	7.29

PAR : 산차, YS : 분만년도-월, SS : 교배용돈

** : P<0.01, * : P<0.05 , NS : non significant

분석결과 번식형질에 영향을 요인들로서 모돈의 산차, 교배차수, 교배용돈 등이 영향을 미 치는 것으로 나타났다. 모든 요인에서 고도의 유의적인 차이(p<0.01)를 보이는 것으로 나타났다.

나. 돼지 번식형질에 대한 유전모수 추정

앞서 돼지의 유전능력 평가 모형 설정을 위하여 분산분석을 실시한 결과 산차, 교배차수, 교배용돈 등이 영향을 미치는 것으로 나타나 유전분석모형에 고정효과(산차, 교배차수) 및 임의효과(교배용돈)로 포함시켰다. 유전체 정보의 유무에 따라 또는 유전체 정보 밀도에 따라 비교하는 연구에서 유전모수의 값이 조건에 따라 변경될 경우 육종가의 수치 및 육종가 정확도에 영향을 미치기 때문에 유전모수의 추정은 유전체 정보 없이 표현형과 혈통정보를 기반으로 REML방법에 의하여 추정하였으며, 그 결과를 표 37에 나타냈다.

육종가 추정방법은 BLUP 추정방법으로 실시하였으며 유전모수 추정에 고려된 모형식은 다음과 같다.

$$y_{ijkl} = \mu + Par_i + YM_j + SS_k + PE_{ijkl} + a_{ijkl} + e_{ijkl}$$

여기서 y_{ijkl} 는 i 번째 산차에 j 번째 분만년도-월에 대한 번식형질 관측치, μ 는 전체평균, Par_i 은 i 번째 산차의 고정효과(수준수: 11), YM_j 는 분만년도-월의 고정효과(수준수: 207), SS_k 는 최종 교배용돈에 대한 임의효과 $\sim N(0, I\sigma_{ss}^2)$, PE_{ijkl} 은 영구환경 임의효과 $\sim N(0, I\sigma_{pe}^2)$, a_{ijkl} 는 개체의 상가적 임의유전효과 $\sim N(0, A\sigma_a^2)$, 그리고 e_{ijkl} 는 개체의 임의오차효과 $\sim N(0, I\sigma_e^2)$ 로 가정하였다.

표 37. 돼지의 번식형질에 대한 따른 유전모수 추정결과

	GL	NBA	TNB1	TNB2
Genetic (co)variances				
GL	0.207	-0.096	-0.118	-0.112
NBA		0.928	0.930	0.958
TNB1			1.003	1.038
TNB2	symm.			1.087
PE (co)variances				
GL	0.183	-0.049	-0.054	-0.054
NBA		0.442	0.505	0.513
TNB1			0.653	0.668
TNB2	symm.			0.700
SS (co)variances				
GL	0.059	-0.020	-0.029	-0.029
NBA		0.084	0.083	0.075
TNB1			0.095	0.089
TNB2	symm.			0.088
Residual (co)variances				
GL	0.884	-0.348	-0.408	-0.406
NBA		6.075	5.841	5.660
TNB1			6.582	6.439
TNB2	symm.			6.851
Heritability	0.155	0.123	0.120	0.125
Repeatability	0.293	0.283	0.282	0.276

TNB1:생존산자수+사산, TNB2:생존산자수+사산+미라, NBA:생존산자수, GL:임신기간, PE:영구환경효과, SS:교배용돈

다. SNP마커의 품질평가

본 연구에서 Illumina PorcineSNP60 V1 beadchip으로 1-3차년도 까지 680두의 시험축군에 대하여 유전정보 데이터를 수집하였으며, 4-5차년도에는 Illumina PorcineSNP60 v2 beadchip을 이용하여 381두의 유전정보 데이터를 수집하여 총 1,061두에 대하여 SNP단위의 품질평가를 실시하였다. 60K chip의 v1의 경우 62,163개의 SNP가 존재하였으며, v2는 64,232개의 SNP가 존재하였다. v1과 v2간 중복되는 SNP 수를 확인해 본 결과 61,177개의 SNP를 공유하고 있었으며, 이 중 염색체 정보 및 거리정보가 결측된 12,577개와 성염색체 내에 존재하는 1,320개의 마커를 제외하고 47,280개의 유전체 정보가 다음 분석에서 사용되었다.

표 38. 랜드레이스 종돈의 60K SNP의 유전자형 자료 및 전처리

자료 및 전처리	Number of SNP
PorcineSNP60 beadchip version1	62,163
PorcineSNP60 beadchip version2	64,232
중복 SNP(version1, version2) ^①	61,177
염색체 정보 및 거리정보 결측 ^②	12,577
성염색체(X, Y) ^③	1,320
유용SNP*	47,280

※ 유용SNP: ①-②-③

라. 유전체 혈연계수 행렬

(1) 방법에 따른 유전체 혈연계수 행렬(1-3차년도)

25,325개의 혈통자료를 이용하여 혈연관계행렬(numerical relationship matrix; NRM)을 구성한 뒤 유전체 정보를 소유한 680두의 조합별 혈연계수를 상삼각행렬 형태로 추출하였다. 또한 이상 표지인자 및 성염색체상의 표지인자를 제거한 680두의 40,651개 SNP 표지인자를 이용하여 5가지 유전체 관계행렬(genomic relationship matrix; GRM)을 구성하고 이를 혈통을 이용한 혈연계수와 비교하였다.

개체별 유전체 정보의 유효성을 판정하기 위하여 혈통혈연계수 대비 유전체 혈연계수를 도화한(ploting) 결과 그림 36의 위쪽 그림에서 표기번호와 같이 혈통정보와 유전체 정보가 어긋나는 개체들이 발견되었다. 부모-자손관계 또는 전형매일 경우 혈통혈연계수는 근친을 제외하였을 경우 이론적으로 $0.5 \pm \alpha$ 로 나타나며, 선조-2대 후손 또는 반형매일 경우 $0.25 \pm \alpha$ 로 나타나게 된다. 혈통-유전체의 부합(mismatch)으로 판단되는 개체들은 혈통혈연계수는 높게 나타나나 유전체 혈연계수는 $0 \pm \alpha$ 로 나타나는 경우와 반대로 혈통혈연계수는 0로 나타나나 유전체 혈연계수는 높게

나타나는 경우가 많이 발견된다. 이러한 개체들은 표 39의 하단과 같이 부 개체에서 CC유전자형이나 자손개체가 TT유전자형으로 나타나는 멘델리안 오류(Mendelian error)들이 다수 발견되며, 채혈, DNA추출, SNPchip 분석, chip 스캔 및 유전자형 구명으로 연결되는 일련의 분석 단계에서의 표기 오류(labeling error)에 의하여 발생된 것으로 사료된다. 혈통-유전체의 부합(mismatch) 검증 결과 총 680두의 유전체 형성 개체 중 20두가 혈통정보와 유전체 정보가 서로 부합되는 것으로 나타나 제거하였으며, 이러한 결과로 선형회귀분석의 결정계수 값이 0.709에서 0.739으로 향상되었다(그림 36).

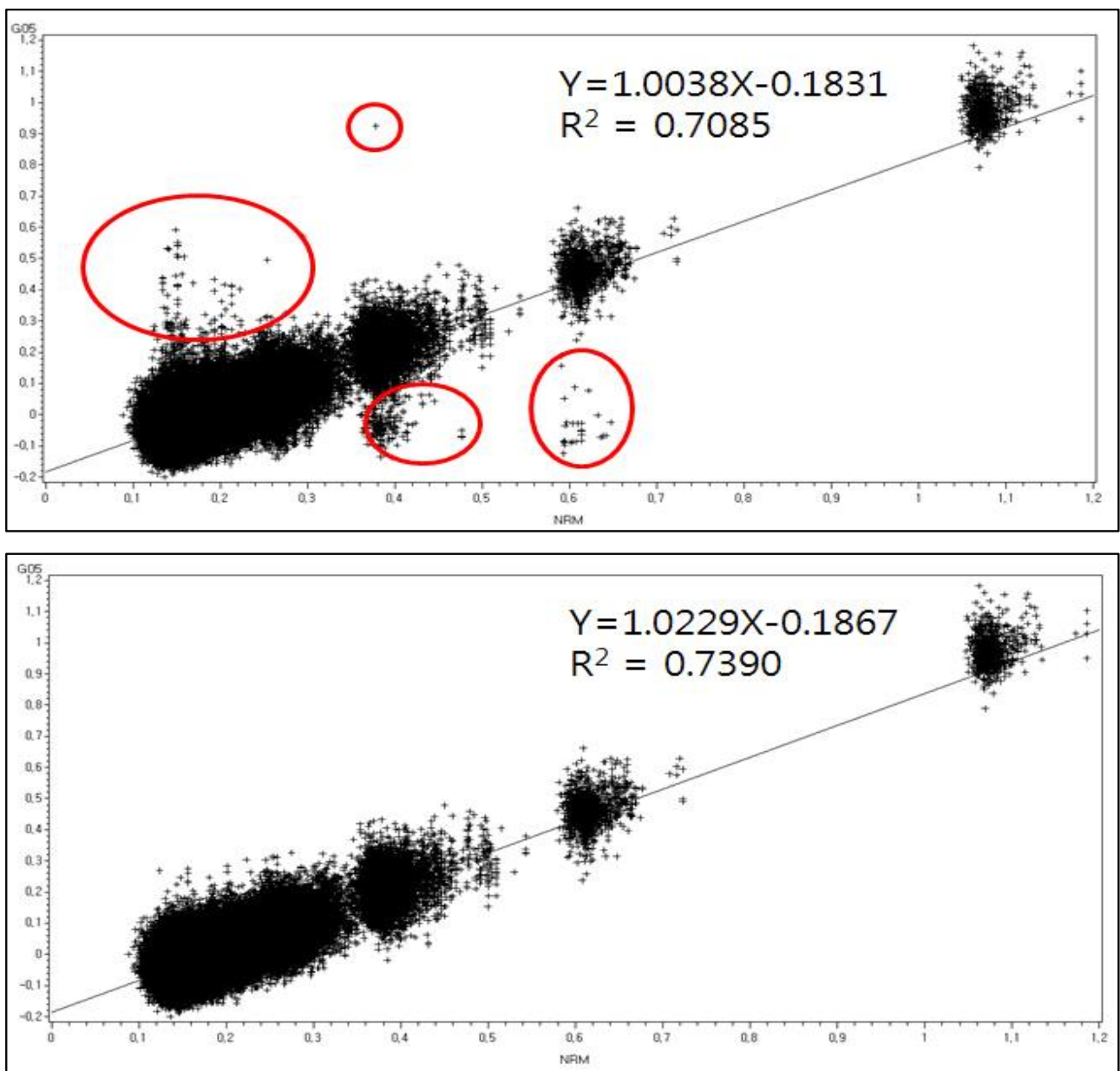


그림 36. 혈통을 이용한 혈연계수와 유전체 정보를 이용한 혈연계수의 차이가 발생하는 개체를 이상치로 판정하여 제거하기 전(위) 후(아래)의 혈연계수 plot

표 39. 이상치로 판단된 개체의 유전자형 검토(정상가계(상), 이상치가 포함된 가계(하))

No. of SNP	정상 가계									
	부	자손1			자손2		자손3		자손4	
	유전자형	유전자형	멘델오류	유전자형	멘델오류	유전자형	멘델오류	유전자형	멘델오류	
1	CC	CC	0	CC	0	CC	0	CC	0	
2	TC	TC	0	TC	0	TC	0	TC	0	
3	AG	GG	0	AG	0	GG	0	GG	0	
4	CC	TC	0	CC	0	CC	0	TC	0	
5	GG	GG	0	GG	0	TG	0	GG	0	
6	CC	CC	0	CC	0	TC	0	CC	0	
7	GG	AG	0	GG	0	GG	0	AG	0	
8	AA	AA	0	AA	0	AA	0	AA	0	
9	GG	AG	0	GG	0	AG	0	AG	0	
10	TT	TC	0	TT	0	TT	0	TC	0	
11	CC	TC	0	CC	0	TC	0	TC	0	
12	AG	GG	0	AG	0	GG	0	GG	0	
13	AG	GG	0	AG	0	AG	0	AG	0	
14	CC	CC	0	CC	0	CC	0	CC	0	
15	CC	TC	0	CC	0	CC	0	CC	0	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
40,706	TT	TC	0	TT	0	TT	0	TT	0	
Total			27		2		3		2	
혈통 유전체 정보 부합 개체가 포함되어 있는 가계(자손6)										
No. of SNP	부	자손5			자손6		자손7		자손8	
	유전자형	유전자형	멘델오류	유전자형	멘델오류	유전자형	멘델오류	유전자형	멘델오류	
1	CC	CC	0	CC	0	CC	0	CC	0	
2	CC	CC	0	TT	1	TC	0	TC	0	
3	GG	GG	0	AG	0	AG	0	GG	0	
4	CC	CC	0	TC	0	CC	0	CC	0	
5	GG	GG	0	GG	0	GG	0	TG	0	
6	TT	TT	0	TC	0	TC	0	TT	0	
7	AA	AA	0	AG	0	AG	0	AG	0	
8	AA	AA	0	AG	0	AA	0	AA	0	
9	GG	GG	0	AA	1	GG	0	AG	0	
10	CC	CC	0	TC	0	TC	0	TC	0	
11	CC	CC	0	TC	0	CC	0	TC	0	
12	GG	GG	0	AG	0	AG	0	GG	0	
13	GG	GG	0	GG	0	AG	0	AG	0	
14	TC	TT	0	CC	0	TC	0	TC	0	
15	CC	CC	0	TT	1	CC	0	CC	0	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
40,706	CC	CC	0	CC	0	CC	0	TC	0	
Total			5		3,258		3		6	

이상치를 제거한 660두의 혈통 및 유전체 정보를 이용한 혈연계수 행렬의 대각원소 및 비대각 원소에 대한 기초 통계량은 표 40과 같다.

표 40. 각 혈연계수 행렬의 혈연계수 기초 통계

	대각원소			
	Mean	Minimum	Maximum	Variance
A	1.077	1.048	1.186	0.000
G05	0.979	0.790	1.183	0.003
GMF	2.756	2.302	2.973	0.004
GOF	0.980	0.815	1.332	0.005
GOF*	0.516	0.416	0.624	0.001
GN	1.000	0.806	1.208	0.003
	비대각 원소			
	Mean	Minimum	Maximum	Variance
A	0.181	0.088	0.723	0.004
G05	-0.001	-0.199	0.661	0.005
GMF	1.775	1.440	2.473	0.008
GOF	-0.001	-0.171	0.740	0.005
GOF*	-0.001	-0.105	0.349	0.002
GN	-0.002	-0.203	0.676	0.006

정보를 수집한 GGP 집단은 1997년 Landrace 순종돈을 입식하고 현재까지 외부의 유전자원의 도입 없이 폐쇄 번식을 통하여 개량사업을 추진해왔기 때문에 혈통정보를 이용한 혈연계수에서 근교계수가 높은 것으로 나타났으나 유전체 정보를 이용한 혈연계수에서는 이러한 근교계수가 잘 나타나지 않은 것을 확인할 수 있다. 그림 36의 위쪽 그림에서와 같이 혈통혈연계수에 대한 유전체 혈연계수의 회귀직선 기울기는 1로 나타났으나 절편이 -0.18로 혈통혈연계수와 유전체 혈연계수가 잘 일치하지만 혈통정보를 이용한 혈연계수는 근교계수만큼 중심이 이동한 것을 확인할 수 있다. 이러한 차이의 원인은 Lee 등(2011)의 모의자료 분석결과에서 표지인자 밀도가 증가함에 따라 유전체 근교계수가 낮게 나타나는 현상 및 폐쇄집단에서 유전적 부동(genetic drift)이 발생한 유전변이들이 유전체 정보의 유의성 검사과정(QC)에서 제거되어 유전체 관계행렬 구성 시 반영되지 않기 때문인 것으로 예상되나 근본적 원인을 명확하게 구명하기 위해서는 추가적인 연구가 필요하다고 사료된다.

유전체 관계행렬 방법론별 차이를 보면 GMF방법에서는 대각행렬 원소의 평균이 높으며, 반대로 GOF*방법에서는 낮은 것으로 나타났다. 나머지 G05, GOF, GN방법의 대각행렬 원소 평균은 혈통을 이용한 혈연계수와 유사하게 나타났으며, 특히 GN방법의 경우 대각원소의 평균이 1로 표준화되어 있는 것을 확인할 수 있다. 혈연계수행렬의 대각원소 표준편차는 혈통을 이용한 혈연계수인 NRM에서 가장 낮게 나타났으며, 유전체 정보를 이용한 혈연계수에서는 GOF*를 제외한 나머지 방법론에서 높게 나타났다. 비대각원소에 대한 비교 결과 NRM과 GMF를 제외한 나머지 방법론에서 평균이 0으로 나타났으며, NRM은 그림36의 선형회귀의 절편값과 유사하게 평균이 0.18정도로 나타났고, GMF는 대각행렬의 비교결과와 유사하게 다른 방법론보다 혈연계수 평균치가 높게 나타났다. 각 유전체 혈연계수의 위와 같은 특징은 같은 Illumina Porcine SNP60 beadchip을 사용한 Forni(2011)의 결과와 매우 유사하게 나타났으며, 한우 종축집단을 분석한 이(2012)의 결과와는 GOF*에서 차이를 보였다. GMF 방법에 의한 혈연계수가 다른 방법론에 비하여 크게 나타나는 이유는 P 행렬을 구성할 때 기준빈도를 0.5 대신 0.5보다 작은 소수대립유전자빈도평균(MAF)을 이용하기 때문인 것으로 사료된다. 또한 GOF*의 대각행렬원소의 평균과 전체 혈연계수의 분산이 작게 계산된 것은 G05, GMF방법등에서 이

용되는 유전체 관계행렬의 척도(scale)인 $2\sum_{j=1}^m p_j(1-p_j)$ 대신에 베타함수를 이용한 유전자빈도분산을 이용하므로 분포 값이 커진 이유로 사료된다.

표 40과 같이 유전체 혈연계수의 표준편차(Variance)가 크게 나타난 것은 이 등(2010)이 언급한 멘델리안 샘플링(mendelian sampling)이 고려된 것으로, 확률적 개념의 혈통혈연계수는 혈연관계에 따라 일률적인 혈연계수를 가지지만 유전체 혈연계수는 실제적 유전변이의 유사도가 고려되어 같은 혈연관계에서도 유전자 구성의 유사도가 멘델리안 샘플링 크기에 따라 변이를 나타낸 것이라 할 수 있다. 그림 37는 혈통 혈연계수와 5가지 유전체 혈연계수의 편차값을 계산하여 빈도분포를 확인한 결과로 모든 유전체 혈연계수계산 방법론에서 정규분포형태의 멘델리안샘플링을 나타내었다. GMF방법을 제외한 나머지 방법에서 혈통혈연계수와 유전체 혈연계수의 편차인 0.18을 중심으로 정규분포 형태를 나타내고 있으며, GMF의 경우 -1.59로 분포의 중심이 이동한 것을 확인할 수 있다. 혈통 혈연계수와 각 방법론별 유전체 혈연계수 편차값의 표준편차는 G05, GMF, GOF, GOF*, GN방법에서 각각 0.046, 0.065, 0.043, 0.045, 0.47로 나타나 GMF의 변이가 가장 크고, GOF의 혈연계수가 혈통혈연계수와 가장 유사한 것을 확인할 수 있었다. 이러한 멘델리안 샘플링이 고려된 유전체 관계행렬은 혈연관계의 정밀도를 향상시키며, 이를 육종가 추정에 이용할 경우 추정의 정확도 향상에 기여할 수 있다고 사료된다.

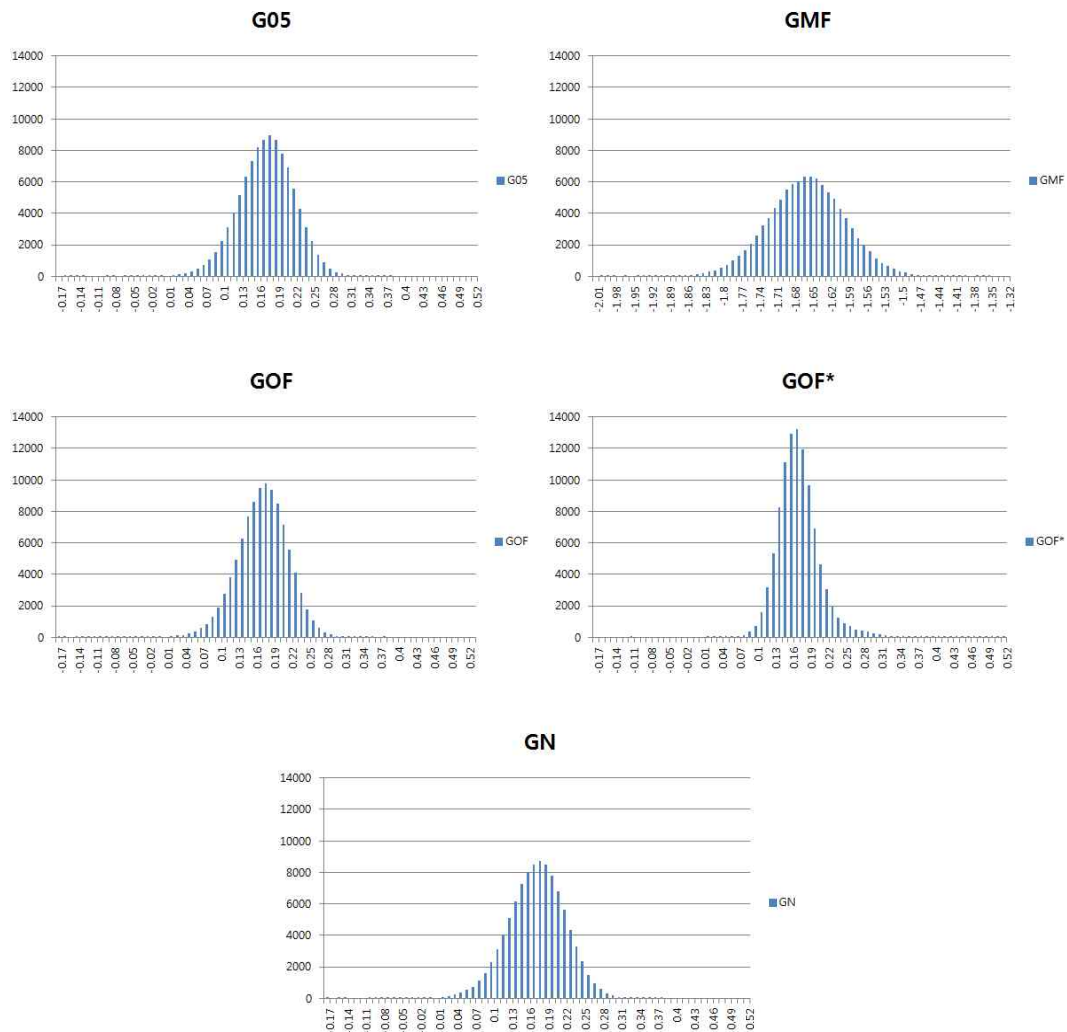


그림 37. 각 유전체 관계행렬 구성 방법론별 멘델리안 샘플링 분포 모양

표 41. 각 혈연계수 행렬의 상관분석결과

	A	G05	GMF	GOF	GOF*	GN
A	1					
G05	0.860	1				
GMF	0.800	0.861	1			
GOF	0.871	0.995	0.856	1		
GOF*	0.860	1.000	0.861	0.995	1	
GN	0.860	1.000	0.861	0.995	1.000	1

표 41는 혈통 및 유전체 정보를 이용하여 계산한 혈연계수의 상관분석을 실시한 자료이며, 혈통 혈연계수와 유전체 혈연계수의 편차에 대한 표준편차가 가장 작았던 GOF방법에 의한 혈연계수가 혈통을 이용한 혈연계수와 가장 높은 상관을 나타내었다. 혈통 혈연계수와 유전체 혈연계수의 상관은 GMF방법에서 가장 낮게 나타났으며, 나머지 유전체 관계행렬 구성방법과도 상대적으로 낮게 나타났다. 5가지 유전체 관계행렬구성 방법론들 중에는 G05, GOF, GOF*, GN방법론간의 혈연계수에서 상관이 높게 나타났으며, 특히 G05, GOF* 및 GN 방법론은 모두 같은 P 행렬을 이용하여 100%의 상관을 보였다.

(2) 유전체 혈연계수 행렬(1-5차년도)

위의 '(1)방법에 따른 유전체 혈연계수 행렬(1-3차년도)'을 통해 제거된 20두를 제외한 660두의 PorcineSNP60 V1 beadchip 유전정보데이터와 4-5차년도 연구를 통해 수집된 381두의 V2 beadchip 유전정보데이터를 합하여 총 1,041두에 대한 혈통기반(NRM) 및 유전체활용(GRM) 혈연계수 추정치를 이용하여 이들 간 산포도 및 회귀식을 추정하였다

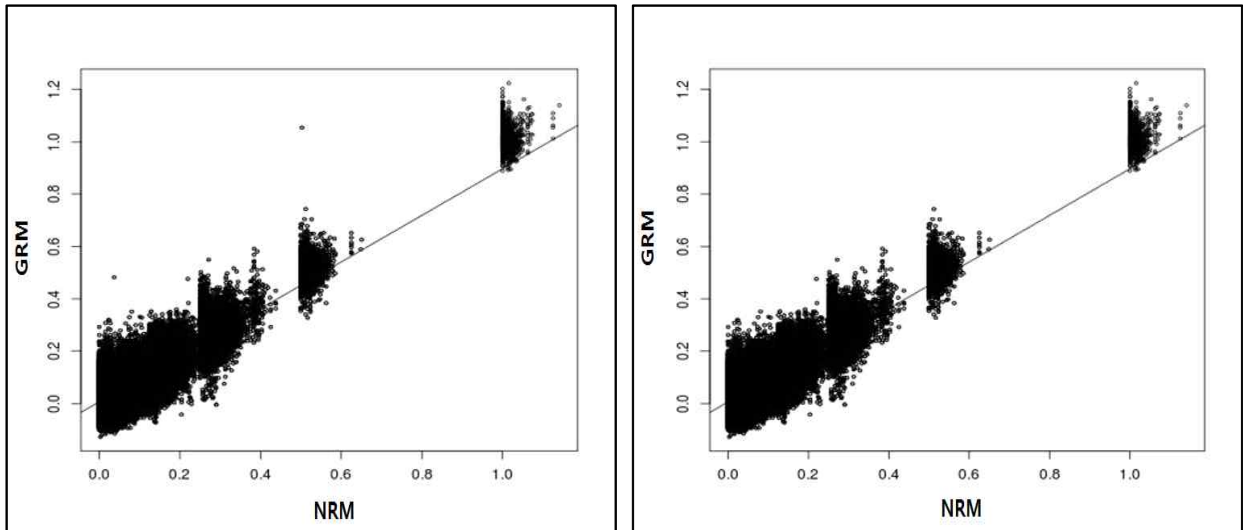


그림 38. 혈통을 이용한 혈연계수(NRM)와 유전체 정보를 이용한 혈연계수(GRM)의 차이가 발생하는 개체를 이상치로 판정하여 제거하기 전(위) 후(아래)의 혈연계수 plot

또한 혈연계수 추정치의 비교를 통하여 잘못된 혈통정보를 가진 개체 3두를 확인하였다. 확인결과 3두의 경우 2014년에 새롭게 도입된 수입정액의 기록으로서 기존 집단과의 연결성이 없어 분석에서 제외하였으며 이에 대한 정보는 표 42에 나타내었다.

표 42. 유전자형의 결측률이 10%이상인 개체 및 혈통정보 오류개체에 대한 품질평가(QC)

Category	Number of animal
Genotyped animals	1,041
Percentage of missing genotype in animal < 10 %	3
Pedigree information error (수입정액)	3
Selected animal	1,035

마. 유전체 정보 유무에 따른 추정육종가 정확도 추정 및 비교

(1) BLUP 및 ssGBLUP별 육종가의 정확도 추정 및 육종가 상관 비교

유전체 정보 없이 혈통 및 표현형 기반으로 개체의 육종가를 추정하였으며(BLUP), 혈통 및 표현형정보에 추가적으로 유전체 정보를 포함하여 육종가를 추정하였다 (ssGBLUP). 추정육종가의 정확도는 자신의 표현형 정보 유무, 표현형 정보를 갖는 개체와 혈연관계 크기 등의 요인에 따라 영향받기 때문에 자신의 표현형 데이터를 보유하고 있는 개체 집단(Reference Population), 자신의 표현형 데이터 없이 반형매 자손의 기록을 갖는 개체 집단(DAM, SIRE), 자신의 표현형 데이터가 없고 자식의 표현형 데이터가 없는 개체 집단(Candidate)로 구분하였으며, 전체 1,035두 중 DAM 집단 및 sire 집단이 각각 139두 및 85두, Reference 집단이 660두, Candidate 집단이 151두로 분석되었고, BLUP의 추정육종가의 정확도는 표 43에 나타내었다. 60K 유전체 정보를 이용하였을 때 DAM 및 SIRE집단 및 Reference집단의 육종가 정확도는 BLUP방법에 비하여 5 - 8%로 증가하는 것으로 분석되었으며, 유전체 선발에서 가장 강점인 자신의 표현형 데이터가 없는 Candidate집단에서는 BLUP방법에 비해 **10 - 11% 육종가 정확도가 향상**되는 것으로 분석되었다.

표43. 번식 형질별 BLUP과 ssGBLUP 육종가 정확도 비교

	BLUP without genotypic data				GBLUP with 60K SNP chip (ssGBLUP)			
	DAM	SIRE	Reference Population	candidate	DAM	SIRE	Reference Population	candidate
N	139	85	660	151	139	85	660	151
GL	0.66	0.62	0.65	0.46	0.72	0.70	0.71	0.57
NBA	0.65	0.62	0.64	0.45	0.71	0.69	0.70	0.56
TNB1	0.65	0.61	0.63	0.45	0.70	0.69	0.69	0.55
TNB2	0.65	0.61	0.63	0.45	0.70	0.69	0.69	0.55

GL : 임신기간, NBA : 생존산자수, TNB1: 생존산자수+사산 , TNB2 : 생존산자수+사산+미라,
DAM : 교배모돈, Sire : 교배웅돈

(2) 검정(표현형)자료를 얻기 전과 후의 유전체육종가 추정치간 상관 비교

또한 분석에 이용된 1,035두의 대상우 중 임의로 220두의 모돈을 선발하여 검정데이터를 얻기 이전에 분석한 유전체 육종가와 검정데이터를 획득한 이후 분석한 유전체 추정육종가간 상관에 대하여 추정하였다.

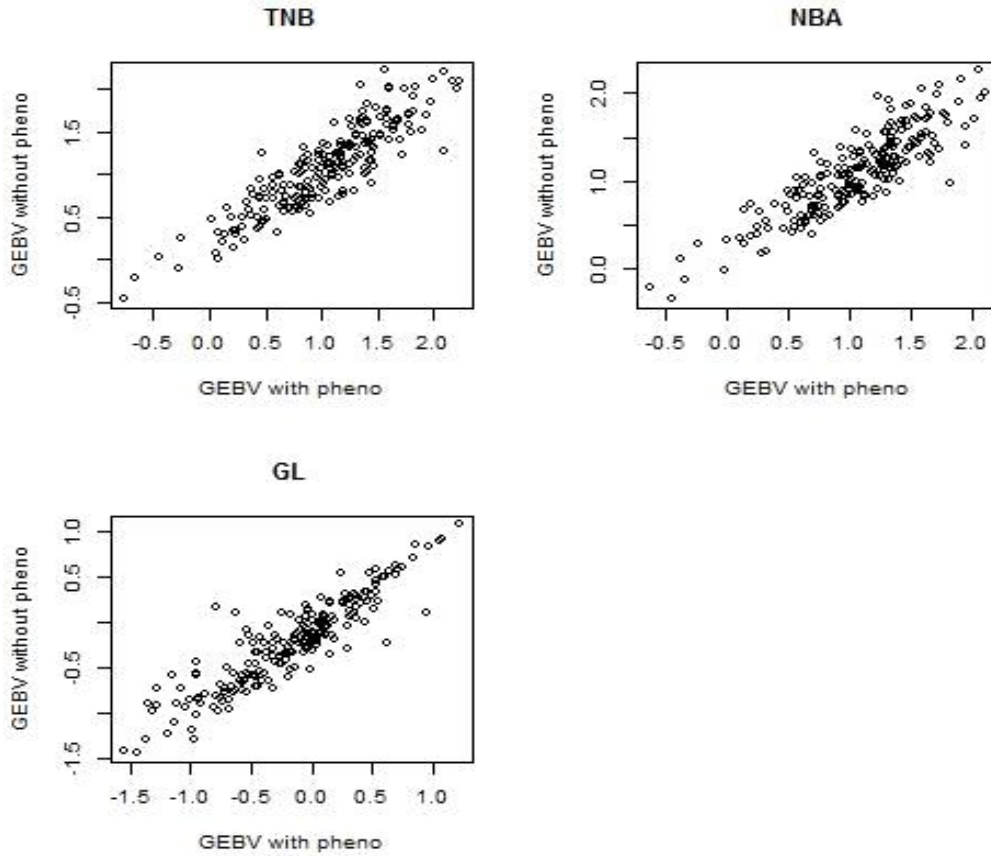


그림 39. 표현형 정보 얻기 이전과 이후의 유전체 육종가추정치간 상관 Graph

그 결과 검정성적을 얻기 전과 후간 유전상관이 총산자수(TNB1)에서 0.89, 생존산자수(NBA)에서 0.87, 임신기간(GL)에서 0.91로 매우 높게 나타났다.

표 44. 표현형 정보 얻기 이전과 이후의 유전체 육종가추정치에 대한 상관

형질	TNB1	NBA	GL
상관추정치	0.89	0.87	0.91

6. 저밀도(6K) chip 선별 및 유전체 육종가 정확도 비교

가. 저밀도칩 분석을 위한 SNP의 선별

저밀도 SNP chip 개발을 위한 SNP 선별을 위하여 60K 내 이웃하는 SNP마커 간 연관불평형 (LD)분석을 실시하였으며, 그래프는 그림 40와 같으며 염색체 1번부터 18번까지를 순차적으로 나타냈으며 x축은 SNP마커간 거리를, y축은 연관불평형(r^2)값으로 표현하였다. 그림에서 보는 바와 같이 전반적으로 가까운 거리에 있는 SNP간 연관불평형의 크기가 컸으며, SNP마커간 거리가 멀어지면 멀어질수록 연관불평형의 값이 적은 것으로 나타났다. SNP간 연관불평형이 높은 조합을 찾아내서 제거해나가는 순서로 진행함으로써 60K SNP 데이터를 5,802개로 축소하게 되었다.

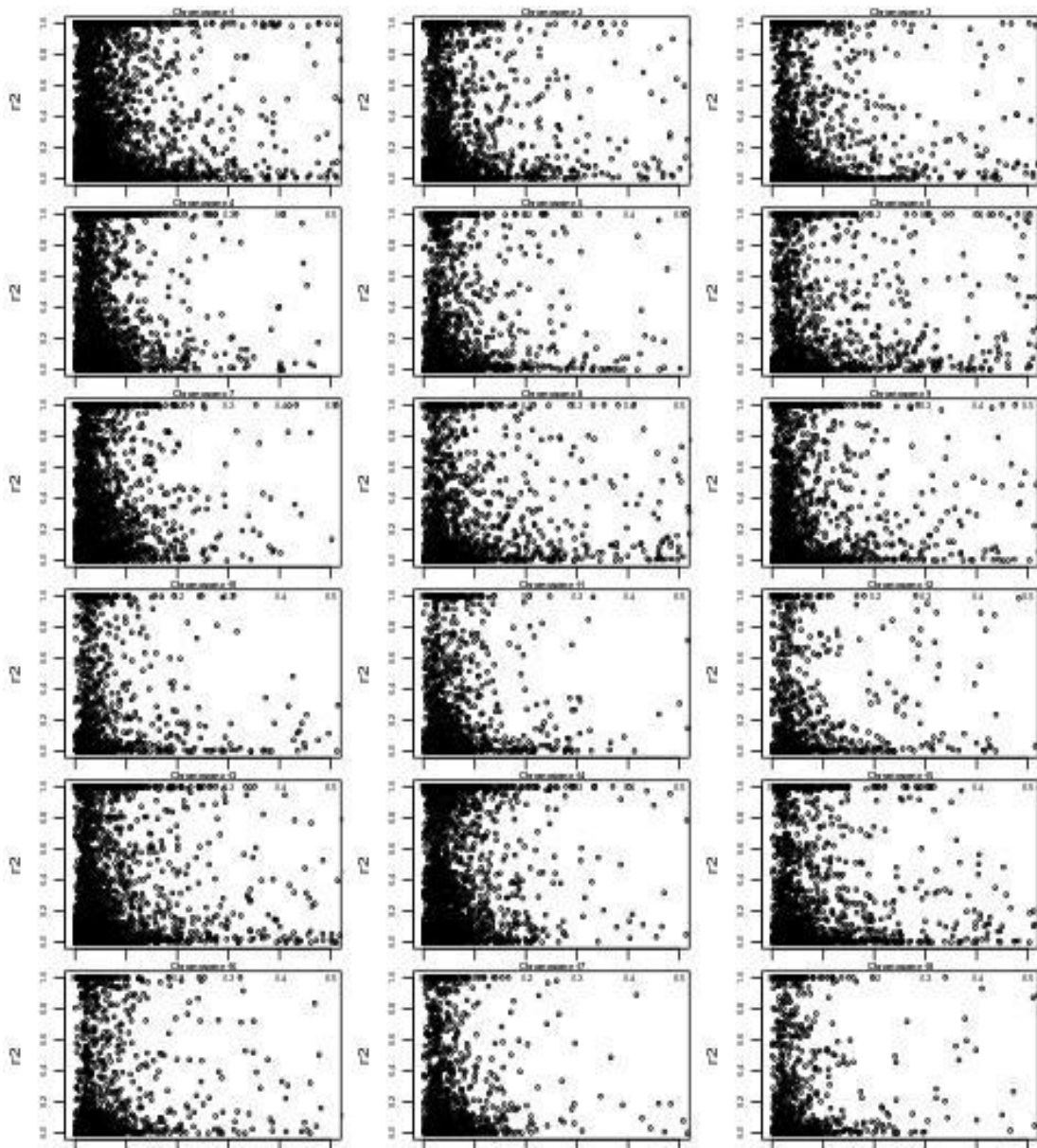


그림 40. 60K 정보를 이용한 염색체별 이웃하는 SNP간 연관불평형 분석

5,802개 축소된 SNP에 대한 연관불평형 그래프는 그림 41에 나타냈다. 그림 41은 그림 40와 다르게 거리가 가까운 SNP마커간 연관불평형의 크기가 많이 축소됨을 볼 수 있다. 다만 거리가 멀어지면 서도 연관불평형이 큰 일부의 SNP마커 조합이 보이게 되는데, 이는 60K에서도 실제로 존재했던 부분이었는데 이웃하는 SNP마커만 그래프에 표기하다보니 나타나지 않은 부분이다. 이러한 부분까지도 제거해 나가면 더 작은 수를 갖는 저밀도 칩을 개발할 수 있겠으나, SNP의 숫자가 너무 줄어들면 이 또한 차후 유전체 유전분석 정확도에 영향을 미칠 것으로 판단되어 제거하지 않았다.

또한 표 45에 염색체 1번에서 60K와 6K(저밀도) SNP chip간 유전적 거리별 연관불평형 크기의 평균에 대해서 나타냈다. 60K chip의 경우 거리가 0.05 Mb(Mega base pairs)이하에서 0.346의 평균 연관불평형 값을, 0.05 ~ 0.10Mb에서는 0.345, 0.10 ~ 0.20 Mb에서는 0.312로 분석된 반면 6K chip의 경우 동일한 거리에서 각각 0.204, 0.215, 0.164로 SNP 마커간 연관불평형이 60K chip에 비해 상대적으로 많이 낮게 나타났음을 볼 수 있다. 따라서 저밀도 SNP 칩 개발을 위한 SNP선별은 연관불평형의 크기에 의해서 상관 높은 SNP를 균등하게 잘 선별했다고 판단된다.

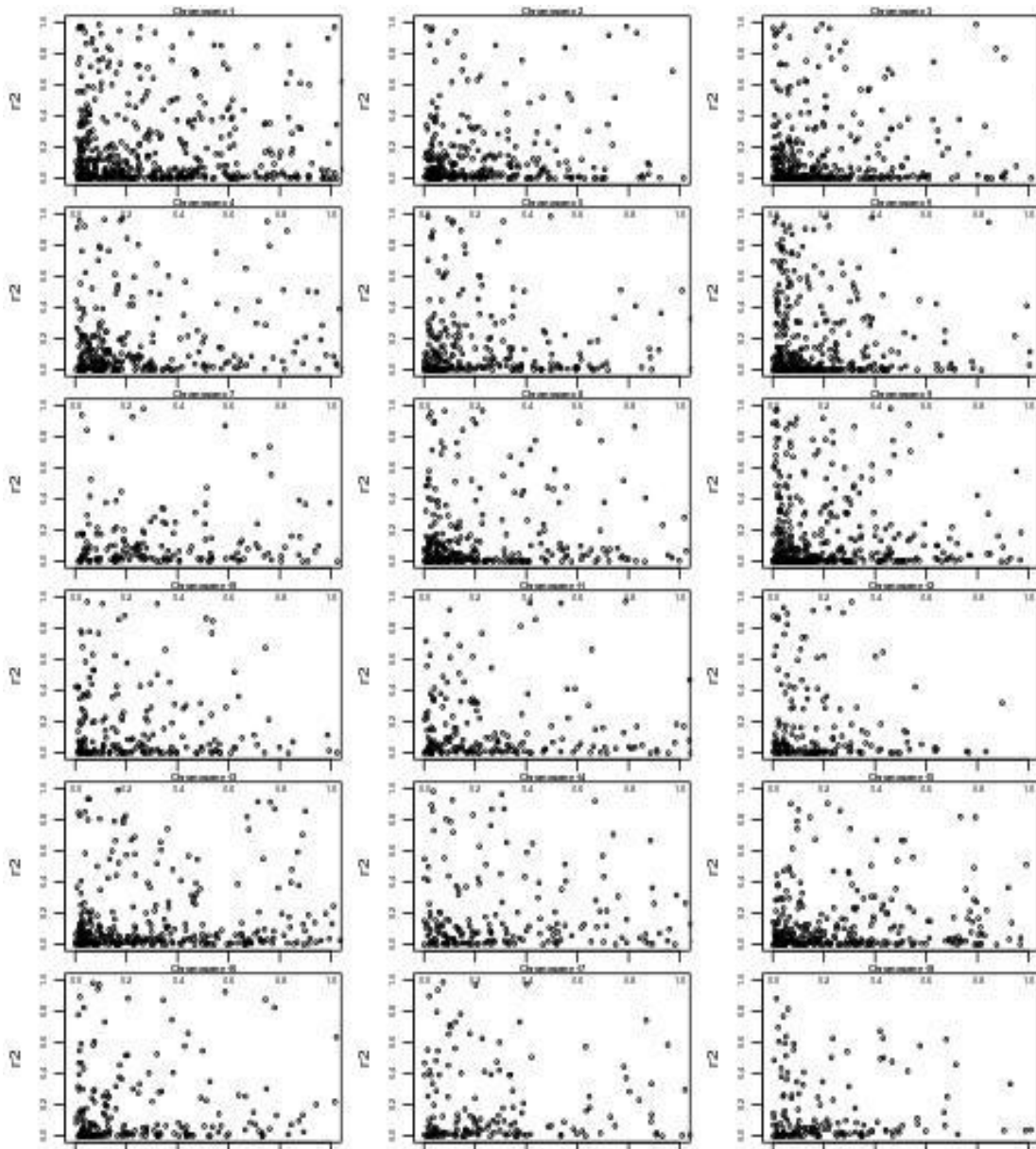


그림 41. 저밀도(6K) chip을 이용한 염색체별 이웃하는 SNP간 연관불평형 분석

표 45. 1번 염색체에서 유전 거리별 연관불평형 평균 값 비교

Distance	Means of Linkage Disequilibrium(r^2)	
	60K SNP chip	6K SNP chip
0.00 ~ 0.05	0.346	0.204
0.05 ~ 0.10	0.345	0.215
0.10 ~ 0.20	0.312	0.164
0.20 ~ 0.30	0.247	0.206
0.30 ~ 0.40	0.384	0.139
0.40 ~ 0.50	0.405	0.235

나. 저밀도(6K) chip과 60K chip 간의 유전체육종가 정확도 비교

(1) SNP Genotype Imputation 정확도

Imputation에서 60K chip 정보를 갖는 개체의 집단이 참조집단이며 참조집단 크기는 Imputation의 정확도에 영향을 미친다. 60K chip 정보를 갖는 1,041두 중 임의로 각각 50%(521두), 70%(729두), 90%(937두)를 선정하여 참조집단으로 하고, 그 외에 개체에 대해선 6K chip에서 얻은 정보와 중첩되는 SNP정보만을 선정하여 다시 60K로 Imputation한 후에 실제 60K chip 정보를 (검정축) 통해 추정육종가와 비교하는 방식으로 Imputation의 정확도를 추정하였다. 표 46에 제시된 바와 같이 임의로 선별된 검정축 104두(10%)에 대한 Imputation 정확도는 평균 0.977로 상당히 높은 정확도로 Imputation됨을 볼 수 있었다.

표 46. 랜드레이스 돼지집단에서 (1,041두) 저밀도(6K) 및 60K density chip 보유두수별 Imputation 정확도에 대한 모의시험 평균 및 표준편차(100회 반복시험)

Proportion of animals with 60K SNP	No. of animals with 6K SNP	No. of animals with 60K SNP	Accuracy of imputation	
			MEAN	SD
0.50	520	521	0.97	0.001
0.70	312	729	0.97	0.001
0.90	104	937	0.98	0.002

(2) 6K chip을 이용한 유전체 육종가 정확도 비교

유전체 육종가 정확도를 자신의 표현형 데이터를 보유하고 있는 개체 집단(Reference population), 자신의 표현형 데이터 없이 반형매 자손의 기록을 갖는 개체 집단(DAM, SIRE), 자신의 표현형 데이터가 없고 자식의 표현형 데이터가 없는 개체 집단(Candidate)로 구분하여 표 47에 나타냈다. 60K 유전체 정보를 이용한 결과와 비교하여 보았을 때, 모든 집단에서의 정확도는 큰 차이 없이(1%이내) 유사하게 나타났다.

표47. 번식 형질별 BLUP과 ssGBLUP 육종가 정확도 비교

	GBLUP with 60K SNP chip (ssGBLUP)				GBLUP with 6K SNP chip (ssGBLUP)			
	DAM	SIRE	Reference Population	candidate	DAM	SIRE	Reference Population	candidate
N	139	85	660	151	139	85	660	151
GL	0.72	0.70	0.71	0.57	0.71	0.69	0.70	0.56
NBA	0.71	0.69	0.70	0.56	0.70	0.69	0.69	0.55
TNB1	0.70	0.69	0.69	0.55	0.70	0.68	0.69	0.55
TNB2	0.70	0.69	0.69	0.55	0.70	0.68	0.69	0.55

GL : 임신기간, NBA : 생존산자수, TNB1: 생존산자수+사산 , TNB2 : 생존산자수+사산+미라,

DAM : 교배모돈, Sire : 교배용돈

제 4 절 한우 및 돼지 종축에 대한 유전체 정보 활용 산업화 방안

1. 한우 유전체 선발에 대한 경제적 효과

가. 한우 보증씨수소 선발을 통한 경제 가치 분석

국내 한우에서 대차대조표와 같은 지표를 통하여 한우생산의 손익지표에 대한 자료 및 경영진단에 대한 자료가 많지 않기 때문에 도축가격에 영향을 미치는 요인에 대한 정확한 분석이 어렵다. 따라서 본 연구에서는 ‘생산비, 자산비 등의 현장에 투입되는 개체 생산비는 모든 개체에서 동일하다’라는 가정 하에 도체가격에 영향을 미치는 도체 형질의 환경요인에 대하여 보정한 후 다중회귀 분석을 통해 12개월령 체중(당대 씨수소 선발형질) 및 도체형질별 도체가격에 기여도를 계산하였다.

(1) 도체가격에 대한 요인별 기여도 분석

농협중앙회 한우개량사업소 36 - 48차 후대검정우 1,757두의 12개월 체중, 도체성적 및 도체단가 자료를 이용하였다. 우선 도체단가 형성에 영향을 미치는 환경요인에 대해 알아보하고자 분산 실시하였다. 그 결과 후대검정 차수와 도축시 개체 나이가 영향을 미치는 것으로 나타났다. 도체가격에 영향을 미치는 요인에 대한 분석을 위하여 1차 사전보정작업의 일환으로 후대검정 차수와 도축시 개체나이를 독립변수로 구성하고 종속변수로 도체가격으로 관계식을 구성한 후 아래와 같이 일반선형모형식(Generalized linear model)을 통해 잔차값을 계산하였다.

$$y = \mu + \text{batch} + \beta \text{age} + e$$

도체형질에 영향을 미치는 환경요인에 대하여 1차 사전보정된 잔차값을 이용하여 도체가격을 종속변수로 나머지 4개의 도체형질 및 12개월 체중을 독립변수로 구성하여 아래와 같이 다중회귀분석분석을 수행하였다.

$$y = b_1 \text{CWD} + b_2 \text{EMA} + b_3 \text{BF} + b_4 \text{MS} + b_5 \text{W12} + b_6$$

여기서 y 는 종속변수로써 도체가격에 대한 잔차값이며, $b_1 - b_5$ 는 각각 도체중, 등심단면적, 등지방두께, 근내지방도, 12개월 체중에 대한 회귀계수값이며, b_6 은 절편값이다.

그 결과 다중회귀모형식의 결정계수는 0.85로 12개월 체중, 등심단면적, 등지방두께, 근내지방도, 도체중이 도체가격을 형성에 크게 영향을 미치는 것으로 분석되었다. 또한 각 도체형질에 대한 편회귀값은 도체중이 14,097원, 등심단면적 2,791원, 등지방두께 -35,015원, 근내지방도 300,252원, 12개월 체중 1,466원, 절편값이 0로 분석되었다. 이는 해당형질 외 다른 도체형질이 고정되었다는 가정하에 단위당 증가량을 설명한다고 할 수 있다 (예: 근내지방도가 1점 증가하면 도체가격이 300,252원 증가한다).

(2) 유전적 개량량을 이용한 현행 한우 선발에 대한 경제 가치 분석 (수소)

유전적 개량량 수식에 경제적 가치를 적용하여 선발을 통해 얻을 수 있는 유전효과에 대한 경제 가치 분석이 가능하다. 현재 농협중앙회 한우개량사업소에서 실시하고 있는 후대검정체계를 보면 그림 42와 같이 우량 씨수소와 우량 암소로부터 700두의 수송아지를 생산하여 최종적으로 보증 씨수소 20여두를 생산하는 시스템이다. 한우 당대검정에서 700두의 수송아지 중 55두(7.9%)로 선발한다. 이를 각 형질의 측정단위를 없애고 표준화된 선발차의 값으로 나타내는 선발강도로 변환시키면 1.87이다. 또한 여기서 선발된 후보씨수소(55두)에 대하여 후대검정을 통해 약 20두(36.4%)의 보증 씨수소를 선발하게 되며, 이를 선발강도로 변환하면 1.03이다.

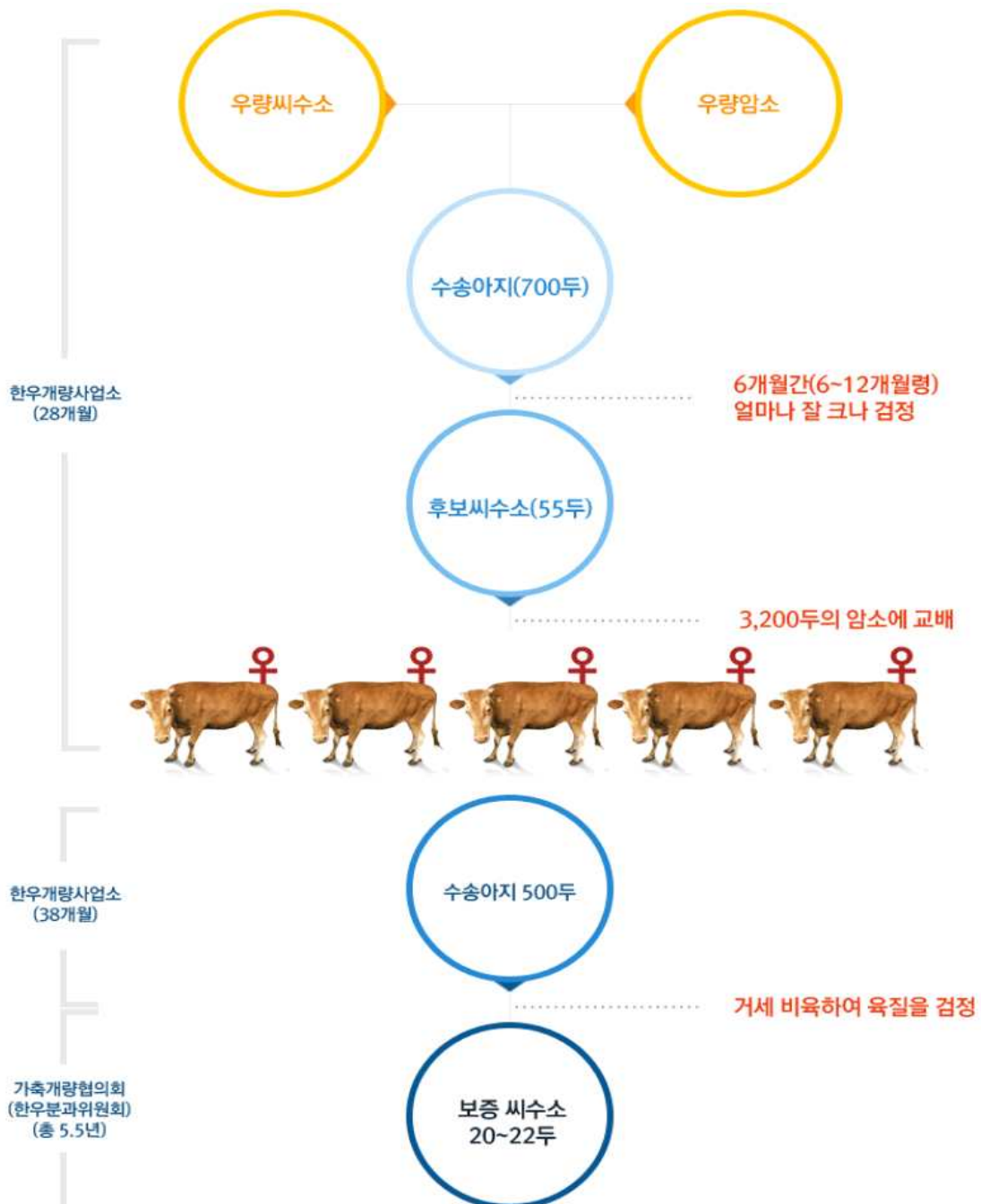


그림 42. 현행 한우 후대검정을 통한 씨수소 선발 체계도(출처:한우개량사업소)

가축의 유전적 개량량은 선발강도(i), 육종가의 표준편차(σ_A)와 육종가와 표현형간의 상관계수(r_{AP})로써 나타낼 수 있다.

$$\Delta G = i\sigma_A r_{AP}$$

현행 한우 개량의 주요 경제형질은 12개월령 체중(당대검정), 근내지방도(당대검정, 후대검정), 등지방 두께(후대검정), 등심단면적(후대검정) 네 가지 형질이다. 또한 당·후대검정의 선발지수에 대한 가중치가 후보씨수소 선발에서 12개월령 체중, 근내지방도에서 각각 1의 가중치를 두고 있으며, 보증 씨수소 선발에서 등심단면적, 등지방두께, 근내지방도에 대해 각각 1, -1, 6의 가중치를 통해 선발하고 있다. 따라서 당·후대검정을 분리하여 유전적 개량량 분석을 실시하였다.

본 연구에서 12개월령 체중에 대한 유전평가는 실시하지 않았기에 정확도 및 유전표준편차에 대한 정보는 한우 유전능력 평가보고서(국립축산과학원, 2014) 및 당대 유전평가기록을 참고하였다. 이외 나머지 도축형질에 대한 정확도 및 유전표준편차는 본 연구에서 추정된 값을 활용하여 유전적 개량량을 계산하였다(표 48). 그 결과, 한우의 당대검정을 통해 세대당 156,493원/두의 개량효과가 있었으며, 후대검정을 통해 302,029원/두로 한우개량을 통해(전체) 세대당 458,522원/두의 개량효과를 얻고 있다.

표 48 . 현행 한우의 당·후대검정 시스템을 통한 유전적 개량량 분석표

구분	기존(현행)					전체
	당대검정		후대검정			
	W12	MS	EMA	BF	MS	
정확도	0.57	0.43	0.75	0.76	0.78	
유전표준편차	14.89	1.20	5.21	2.34	1.20	
선발강도	1.87	1.87	1.03	1.03	1.03	
도체가격	1,466	300,252	2,791	-35,015	300,252	
유전적개량량	23,267	289,719	11,233	-64,139	289,467	
유전적개량량 (선발지수)	당대 : 156,493원		후대: 302,029원			458,522원

나. 한우 유전체 정보 활용 보증씨수소 선발을 통한 경제 가치 분석

(1) 현행 한우 보증씨수소 선발 체계 활용 유전체 선발을 통한 경제 가치

현행 한우 보증씨수소 선발 체계에서 유전체 정보를 활용하였을 때 얻을 수 있는 경제 가치에 대하여 분석을 실시하였으며, 유전체 밀도에 따라 저밀도(8K) chip, 50K chip, 700K(Imputation) chip으로 구분하여 비교하였다.

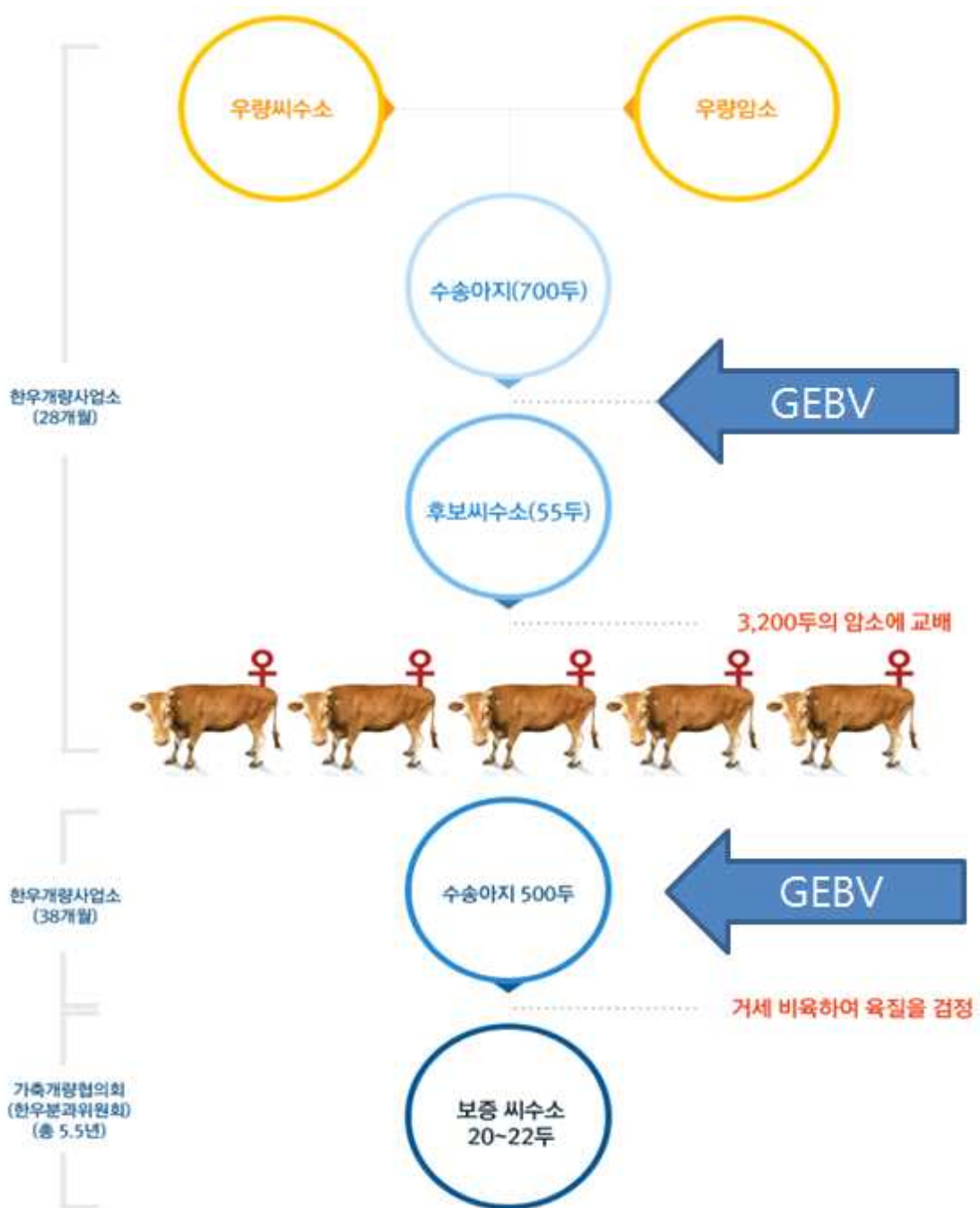


그림 43. 한우 유전체 정보 활용 씨수소 선발 체계도

그림 43와 같이 현행 후대검정 체계에서 후보씨수소 선발과 보증씨수소 선발 단계에서 유전체 정보를 포함하여 유전평가 결과를 적용하였을 때, 유전체 분석 칩의 종류에 따라 유전적 개량량에 대한 변화에 대한 결과를 표49, 50, 51에 나타냈다. 현행 한우 당·후대검정체계에 8K chip을 이용하였을 때, 당대검정에서의 근내지방도 정확도가 0.43에서 0.53으로 증가하였으며, 후대검정에서는 각 도체형질별 0.02의 정확도가 향상되었다. 이들 정보를 활용하여 당·후대 선발에서의 유전적 개량량 변화를 보면, 한우 당대검정을 통해 세대당 190,181원/두, 후대검정을 통해 309,782원/두로 분석되었으며, 한우 당·후대선발을 통한 유전개량을 통해 총 세대당 499,964원/두의 개량 효과를 얻을 수 있을 것으로 기대되었다.

현행 한우의 당·후대검정체계에 50K chip을 이용할 경우, 한우 당·후대선발을 통한 유전개량을 통해 세대당 486,323원/두의 개량효과를 얻을 수 있을 것으로 분석되었고, 700K chip 정보를 이용하였을 경우 세대당 479,218원/두의 유전적 개량의 가치를 창출할 수 있을 것으로 추정되었다.

한우 당·후대검정체계에서 유전체 정보를 활용하였을 때 유전체 정보의 밀도에 따라서 차이를 보였으나 유전체 정보를 활용하지 않은 기준은 방법에 비하여 최소 세대당 20,696원/두의 효과를 기대해 볼 수 있을 것으로 분석되었으며, 세 가지의 서로 다른 밀도 칩 중 가장 저밀도인 8K chip의 경우 50K나 700K와 견주어 비교하여도 전혀 손색이 없고 오히려 상대적으로 고밀도인 이들 칩에 비해 유전적 개량량을 향상할 수 있을 것으로 분석되어 차후 대용량 칩 분석비용을 절감하면서 더 많은 참조집단을 구성하는데 활용될 수 있을 것으로 사료된다.

표 49. 한우의 당·후대검정을 통한 유전적 개량량 (8K chip 이용)

구분	유전체 정보 활용 (8K chip)					전체
	당대검정			후대검정		
	W12	MS	EMA	BF	MS	
정확도	0.57	0.53	0.77	0.78	0.80	
유전표준편차	14.89	1.20	5.21	2.34	1.20	
선발강도	1.87	1.87	1.03	1.03	1.03	
도체가격	1,466	300,252	2,791	-35,015	300,252	
유전적개량량	23,267	357,096	11,533	-65,827	296,889	
유전적개량량 (선발지수)	당대 : 190,181원			후대: 309,782원		499,964원

표 50. 한우의 당·후대검정을 통한 유전적 개량량 (50K chip 이용)

구분	유전체 정보 활용 (50K)					전체
	당대검정		후대검정			
	W12	MS	EMA	BF	MS	
정확도	0.57	0.53	0.77	0.78	0.80	
유전표준편차	14.89	1.20	5.21	2.34	1.20	
선발강도	1.87	1.87	1.03	1.03	1.03	
도체가격	1,466	300,252	2,791	-35,015	300,252	
유전적개량량	23,267	330,145	11,383	-64,983	296,889	
유전적개량량 (선발지수)	당대 : 176,706원		후대: 309,617원			486,323원

표 51. 한우의 당·후대검정을 통한 유전적 개량량 (700K chip 이용)

구분	유전체 정보 활용 (700K chip)					전체
	당대검정		후대검정			
	W12	MS	EMA	BF	MS	
정확도	0.57	0.53	0.77	0.78	0.80	
유전표준편차	14.89	1.20	5.21	2.34	1.20	
선발강도	1.87	1.87	1.03	1.03	1.03	
도체가격	1,466	300,252	2,791	-35,015	300,252	
유전적개량량	23,267	323,407	11,233	-64,983	293,178	
유전적개량량 (선발지수)	당대 : 173,337원		후대: 305,881원			479,218원

(2) 한우 당·후대 동일 기준 유전체 선발을 통한 경제 가치

유전체 선발의 강점은 표현형이 수집되지 않은 개체, 즉 한우의 선발체계에서 보면 당대검정시기인데 현행 시스템에서는 당대검정에서 근내지방도 평가에서만 유전체 선발의 강점이 발휘될 수 있다. 만약 당대검정에서 후대검정과 동일한 도체형질 그리고 가중치로 선발한다면 유전체 선발의 강점이 더 극대화 될 것으로 판단된다. 따라서 당·후대검정체계를 동일하게 구성하여 현행 방법과 유전체 선발 방법 간 비교를 실시하여 표 52, 53, 54, 55에 나타냈다.

표 52 . 한우의 동일한 당·후대검정 시스템을 통한 유전적 개량량 (유전체 정보 없이 선발)

구분	유전체 정보 없이 선발						전체
	당대검정			후대검정			
	EMA	BF	MS	EMA	BF	MS	
정확도	0.42	0.43	0.43	0.75	0.76	0.78	
유전표준편차	5.21	2.34	1.20	5.21	2.34	1.20	
선발강도	1.87	1.87	1.87	1.03	1.03	1.03	
도체가격	2,791	-35,015	300,252	2,791	-35,015	300,252	
유전적개량량	11,421	-65,884	289,719	11,233	-64,139	289,467	
유전적개량량 (선발지수)	당대 : 302,603원			후대 : 302,029원			604,632원

표 53 . 한우의 동일한 당·후대검정 시스템을 통한 유전적 개량량 (8K chip 이용)

구분	유전체 정보 활용 (8K chip)						전체
	당대검정			후대검정			
	EMA	BF	MS	EMA	BF	MS	
정확도	0.50	0.52	0.53	0.77	0.78	0.80	
유전표준편차	5.21	2.34	1.20	5.21	2.34	1.20	
선발강도	1.87	1.87	1.87	1.03	1.03	1.03	
도체가격	2,791	-35,015	300,252	2,791	-35,015	300,252	
유전적개량량	13,596	-79,674	357,096	11,533	-65,827	296,889	
유전적개량량 (선발지수)	당대 : 372,641원			후대 : 309,782원			682,423원

표 54 . 한우의 동일한 당·후대검정 시스템을 통한 유전적 개량량 (50K chip 이용)

구분	유전체 정보 활용 (50K)						전체
	당대검정			후대검정			
	EMA	BF	MS	EMA	BF	MS	
정확도	0.47	0.48	0.49	0.76	0.77	0.80	
유전표준편차	5.21	2.34	1.20	5.21	2.34	1.20	
선발강도	1.87	1.87	1.87	1.03	1.03	1.03	
도체가격	2,791	-35,015	300,252	2,791	-35,015	300,252	
유전적개량량	12,780	-73,545	330,145	11,383	-64,983	296,889	
유전적개량량 (선발지수)	당대 : 344,533원			후대 : 309,617원			654,149원

표 55. 한우의 동일한 당·후대검정 시스템을 통한 유전적 개량량 (700K chip 이용)

구분	유전체 정보 활용 (700K)						전체
	당대검정			후대검정			
	EMA	BF	MS	EMA	BF	MS	
정확도	0.46	0.47	0.48	0.75	0.77	0.79	
유전표준편차	5.21	2.34	1.2	5.21	2.34	1.2	
선발강도	1.87	1.87	1.87	1.03	1.03	1.03	
도체가격	2,791	-35,015	300,252	2,791	-35,015	300,252	
유전적개량량	12,508	-72,013	323,407	11,233	-64,983	293,178	
유전적개량량 (선발지수)	당대:337,494원			후대:305,881원			643,375원

그 결과 유전체 정보 없이 한우 당·후대검정 시스템에 적용하였을 때, 한우의 당대검정을 통해 세대당 302,603원/두의 개량효과가 있었으며, 후대검정을 통해 302,029원/두로 한우개량을 통해 (전체) 세대당 604,632원/두의 개량효과를 얻을 수 있다고 분석되었으며, 현재 실시하고 있는 한우 후대검정체계에 비해 유전체 정보를 이용한 한우 후대검정체계에서 약 32% 향상된 유전적 개량량을 나타낼 것으로 추정되었다. 동일한 조건에서 저밀도 유전체 정보를 활용하여 유전체 선발을 실시할 경우 세대당 682,423원/두, 50K 정보를 활용할 경우 세대당 654,149원/두, 700K 정보를 활용할 경우 세대당 643,375원/두의 가치를 획득할 수 있을 것으로 분석되었다.

현행 한우 보증씨수소 선발은 육종농가에서 혈통, 질병 및 외모심사를 통해 개체 선발 단계를 시작으로 당대검정 그리고 후대검정을 거쳐 진행된다. 이 중 가장 선발강도가 높은 단계는 육종농가대상우 약 10,000두에서 800여두의 당대검정 개체를 선발하는 과정이다. 육종농가에서 개체를 선발하는 단계가 유전체 선발을 통해 가장 높은 효율을 얻을 수 있는 단계로 판단된다. 다만 선발의 대상이 되는 개체의 수가 많기 때문에 저렴한 비용을 통해 유전체 데이터를 수집할 수 있는 방법 및 시스템이 필요하다. 현행 유전체 분석비용에 대한 단가는 분석두수, 칩밀도 등에 따라 차이가 있다. 본 연구 결과에 의하면 8K chip 활용한 유전평가 결과가 50K 또는 700K와 견주어 손색이 없기 때문에 8K chip을 이용하면 유전체 분석의 비용을 보다 저렴한 비용으로 수집이 가능할 것이고, 이것이 한우개량 시스템에 접목이 된다면 8K chip의 분석비용을 두당 약 5 ~ 10만원로 가능할 것이다. 이는 육종농가개체 선발단계에서 유전체 정보를 이용하고자 할 때 500백만원 ~ 1000백만원의 비용이 추가되게 된다. 한우개량사업에서 육종농가개체선발 단계에서부터 유전체 선발을 실시한 것인지에 대해서는 차후 유전체 분석 비용 대비 얻을 수 있는 경제성에 대한 타당성 연구가 추가적으로 필요할 것으로 판단되며, 현재 참조집단 축군을 이용하여 얻을 수 있는 득이 적다면 우선 당·후대검정시스템에서 유전체 선발을 적용하면서 참조집단의 축군을 증가한 후 적용하는 방안도 고려해 보아야 할 것이다.

(3) 한우 유전체 정보를 활용한 당·후대검정체계 개선(안)

상기의 검정체계별 유전적 개량량에 따른 경제적 효율성을 비교하면(그림 44) 세대당 약 223.9천원/두의 개량효율이 증가함을 알 수 있다. 만일 이러한 개량효율이 선발된 종모우의 정액으로 전국에 공급되고 이들의 자손이 연간 1,000천두가 생산되고 세대간격을 3.5년으로 가정한다면 약 640억원(223,900백만원/3.5년)의 경제적 효율이 증가함을 추정할 수 있다.

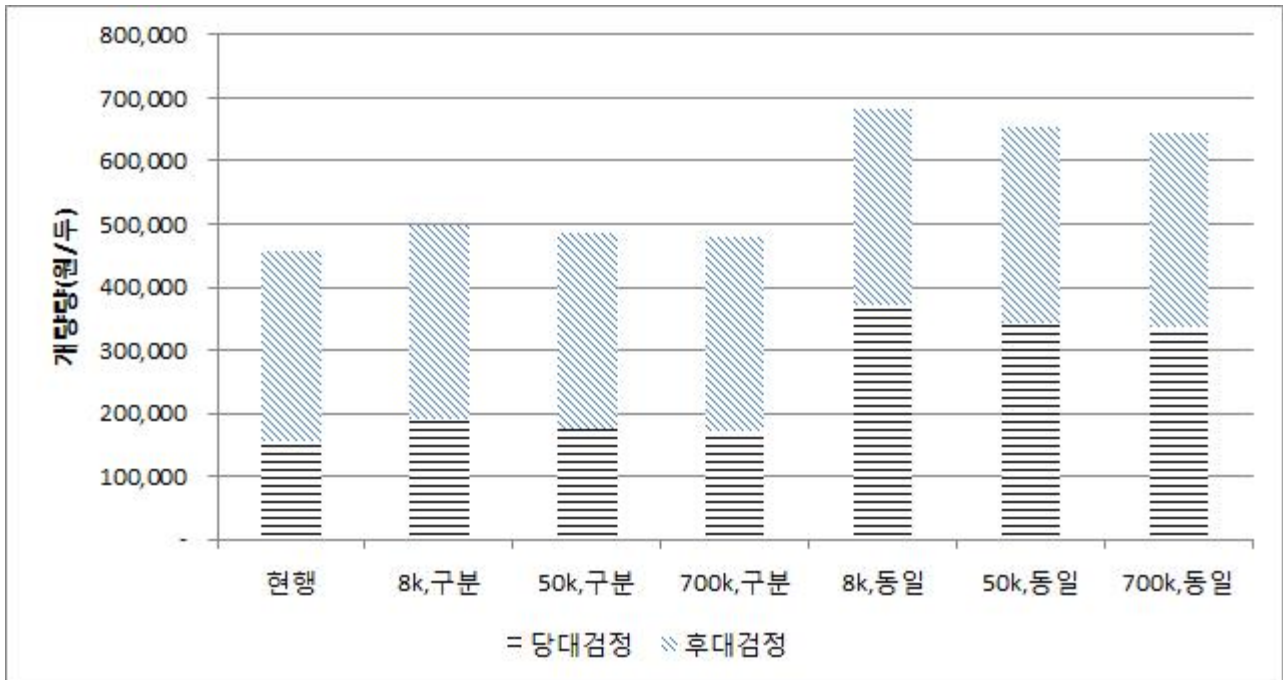


그림 44. 검정체계 및 선발방법에 따른 두당, 세대당 개량량 추정치

주) 8K,50K,700K: SNP자료 수집량,

구분: 현행의 후보종모우 및 보증종모우 선발시 적용되는 선발지수식

동일: 후보종모우 및 보증종모우 선발시 적용되는 선발지수식을 보증종모우 선발시 선발지수식으로 동일하게 적용할 때

개량 효율성을 증가하기 위한 검정체계(연간 계획)의 개선안은 다음과 같이 요약하여 제안할 수 있다.

- ① 1등급 씨수소(약 40두)와 전국 최우수 종빈우(약 9,000두 선발)를 활용하여 우량 수송아지 2,500두 생산
- ② 생산된 수송아지 중 2000두를 1차 선발하여 LD기반 SNP(8K)분석(3-6개월령)
- ③ SNP자료에 기반하여 도축형질에 대한 GEBV를 추정하고 추정치에 기반하여 1,000두의 당대검정 대상우 선발
- ④ GEBV에 근거하여 탈락된 송아지는 6개월령에 거세 비육 후 30개월령까지 비육 도축하여 육질 자료 수집 - 수집된 자료는 유전체 분석용 참조집단으로 활용
- ⑤ 당대검정 대상 선발 송아지는 12개월령까지 당대검정후 12개월령 체중에 근거한 후보종모우 선발(약 60두)
- ⑥ 선발된 후보씨수소는 후대검정 실시

- ⑦ 후보씨수소 및 후대검정 송아지는 50K SNP분석 후 도체성적을 포함한 참조집단으로 활용
- ⑧ 보증종모우 선발은 SNP정보를 활용한 육종가 추정에 근거하여 선발

참조집단의 크기가 10,000두에 도달할 때까지 본 사업을 적용한다.

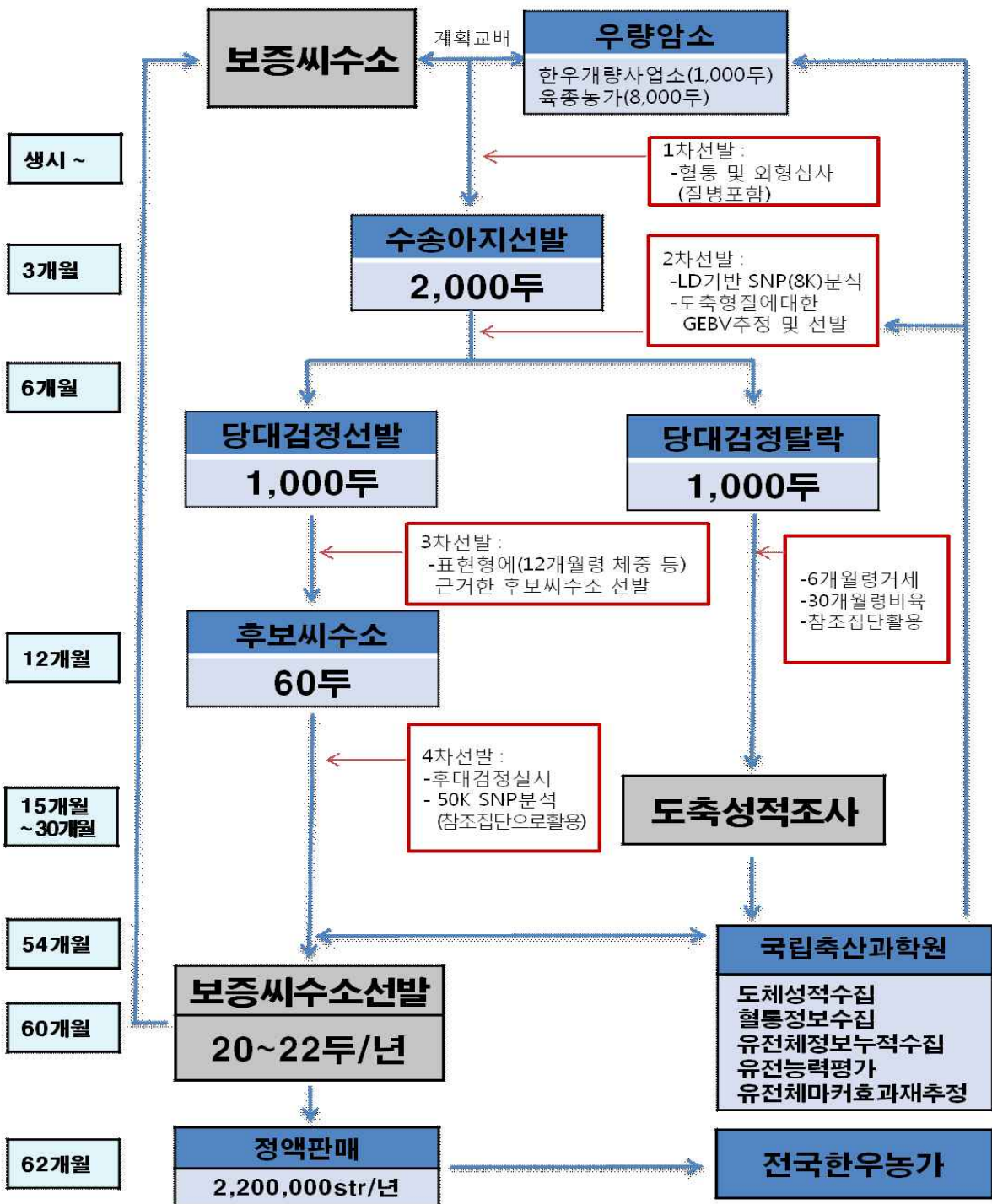


그림 45. 한우 유전체 정보를 활용한 당·후대검정체계 개선(모식도)

다. 한우 유전체 정보 활용 암소 선발을 통한 경제 가치 분석

한우 수소의 개량은 국가단위에서 진행되고 있으며, 한우 암소의 개량은 농가의 몫이다. 현재 농가단위에서 개체를 선발하는 체계적인 시스템은 없는 상태이며, 한우암소검정사업을 통해 초음파 또는 혈통정보를 이용한 개체의 유전평가로 암소축에 대하여 농가 스스로 선발 및 도태를 할 수 있도록 사업을 진행하고 있으나 농가 활용이 미진한 실정이다. 본 연구에서는 한우 암소에서 초음파정보를 이용한 선발방법과 유전체정보를 이용한 선발방법간 비교를 통해 얻을 수 있는 유전적 개량량에 대한 경제 수익을 비교하고자 한다.

(1) 초음파 활용 암소선발

초음파는 한우에서 개체의 도체형질(등심단면적, 등지방두께, 근내지방도)을 예측하는데 하나의 도구로 사용되고 있다. 초음파는 개체를 도축하지 않고도 능력을 미리 예측함으로써 상당히 유용한 방법으로 활용되고 있으나 이것이 실제 관측치가 아닌 초음파를 통한 추정치이므로 표현형 측정의 정확도가 떨어지며 관측하는데 시간이 소요되기에 초음파 정보를 이용한 유전분석결과 보다는 초음파 표현형정보 기반에 개체 선발을 대부분의 농가에서 수행하고 있을 것으로 판단된다. 앞서 연구한 결과 및 가정(Assumption)을 통하여 초음파를 통한 유전적 개량량에 대하여 추정하였다(표 56). 암소에서 도체형질에 대한 측정의 정확도는 70%, 초음파 정보를 이용하여 전체 암소 중 70%선발한다고 가정하였다.

등심단면적의 정확도는 0.28, 등지방두께 0.29, 근내지방도 0.32로 추정되었으며, 앞서 분석된 형질별 유전표준편차, 도체가격을 적용한 결과 등심단면적의 경우 세대당 0.729cm^2 유전적 개량량의 효과를 얻을 수 있을 것으로 추정되었으며 등지방두께는 0.339mm, 근내지방도는 0.192점의 유전적 개량효과가 있는 것으로 분석되었다. 이를 도체가격의 가중치를 주어 경제가치로 환산한 결과 등심단면적, 등지방두께, 근내지방도에서 각각 2,036원, -11,881원, 57,648원/두/세대의 효과를 얻을 수 있을 것으로 추정되었다.

표 56. 한우 암소 초음파 정보 활용 유전적 개량량 예측

구분	초음파 정보활용 암소선발		
	EMA	BF	MS
정확도	0.28	0.29	0.32
유전표준편차	5.21	2.34	1.2
선발강도	0.5	0.5	0.5
도체가격	2,791	-35,015	300,252
유전적개량량	0.729	0.339	0.192
유전적개량량(원)	2,036	-11,881	57,648

(2) 유전체정보 활용 암소선발

초음파기록 대신에 유전체 정보를 암소 선발에 활용시 기대할 수 있는 유전적 개량량에 대해서 표 57에 나타냈다.

암소의 선발강도를 제외한 나머지에 대한 정보는 앞서 분석한 결과를 이용하였다. 그 결과 등심 단면적의 경우 세대당 1.303cm² 유전적 개량량의 효과를 얻을 수 있을 것으로 추정되었으며 등지방두께는 0.608mm, 근내지방도는 0.318점의 유전적 개량효과가 있는 것으로 분석되었다. 이를 도체가격의 가중치를 주어 경제가치로 환산한 결과 등심단면적, 등지방두께, 근내지방도에서 각각 3,635원, -21,303원, 95,480원/두/세대의 효과를 얻을 수 있을 것으로 추정되었다.

한우 암소 선발에 초음파정보 대신 유전체 정보를 이용하여 분석하였을 때 등심단면적, 등지방두께, 근내지방도에서 각각 1,599원, -9,422원, 37,832원/두/세대 증가될 것으로 추정되었다. 한우 암소의 세대간격을 3년, 가임암소의 수를 80만이라고 가정한다면 유전체 정보를 한우 암소 개량에 사용하게 됨으로써 연간 얻을 수 있는 기대 차익은 등심단면적 426백만원, 등지방두께 2,512백만원, 근내지방도 10,088백만원로 추정되었다. 참고로 여기서 나타낸 결과는 유전적 개량량만을 서로 비교한 것으로 좀 더 정확한 분석을 위해서 자료수집에 소요되는 총 경비를 고려해야 할 것이며, 차후 이들에 대한 경제성 분석 연구 역시 필요할 것이다.

표 57. 한우 암소 초음파 정보 활용 유전적 개량량 예측

구분	유전체 정보활용(8K) 암소선발		
	EMA	BF	MS
정확도	0.50	0.52	0.53
유전표준편차	5.21	2.34	1.2
선발강도	0.5	0.5	0.5
도체가격	2,791	-35,015	300,252
유전적개량량	1.303	0.608	0.318
유전적개량량(원)	3,635	-21,303	95,480

(3) 한우 유전체 정보를 이용한 암소개량 산업화 방안

상기에서 언급한 바와 같이 초음파 정보를 이용한 암소의 유전평가와 이에 기초한 우량암소 선발은 사업수행에 있어서 크게 2가지 문제점이 있다. 첫째는 초음파 측정의 정확도가 초음파 측정자, 판독자등에 따라 크게 다를 수 있으며 유전평가의 활용을 위한 시기적 적절성을 구현하는데 어려움이 있다는 것이다. 즉 초음파를 측정하고 측정된 자료를 판독, 자료화 및 혈연관계가 있는 개체들의 정보를 수집하여 유전평가 하는데 상당기간이 소요되기 때문에 농가에 정보제공 및 활용이 제한적이라는 것이다. 둘째, 초음파 측정에 소요되는 비용이 두당 약 30천원으로 농가에 제공하는 정보에 비교하여 높게 형성되어 있으므로 농가의 적용이 지난한 것이 현실이다.

유전체 SNP(8K)정보를 활용한 산업화 방안은 초음파 정보를 활용한 방안에 비교하여 육종가 추정치의 정확도가 높을 뿐만 아니라 자료수집이 쉽고 유전분석에 소요되는 시간이 짧으며 농가의

정보제공이 초음파측정 결과에서 보다 많은 장점이 있다. 예를 들면 육종가 추정치 뿐만 아니라 친자감정이 가능하며 또한 친자가 오류가 있을 경우에 친자를 찾을 수도 있으며 향후 질병과 같은 유전자분석이 이루어 졌을 때, 이에 대한 정보도 제공할 수 있는 많은 가능성이 있는 사업이다.

추진절차는 다음과 같이 요약할 수 있다.

- ① 송아지 생산으로 이표 장착 시 혈통부 정리(등록) 및 채혈
- ② 혈액으로부터 SNP(8K) 정보 생산
- ③ SNP(8K)정보에 의한 친자확인(친자찾기), 질병확인, 육질관련 형질에 대한 육종가 예측
- ④ 농가에 정보제공 및 송아지 선발
- ⑤ 선발송아지 활용 암소개신 및 이들 암소로부터 후대축 생산
- ⑥ 생산된 후대축의 비육도축으로 육질 판정결과 수집
- ⑦ 육질 결과에 의한 SNP정보 활용 육종가 평가 결과 검증



그림 46. 한우 유전체 정보를 이용한 암소개량 산업화 방안

2. 돼지 유전체 선발에 대한 경제적 효과 및 산업화 방안

현재까지 누적된 참조집단(reference population; 유전체 정보와 표현형을 가지는 개체 수) 규모는 산자수에 대하여(이상치 제외) 1,061두로 적으며, 따라서 유전체 선발의 개량 효과를 현 시점에서 명확하게 예측하기 어렵다. 그러나 작은 참조집단임에도 불구하고 산자수에 대한 유전체 육종가는 표현형이 없는 개체에 대하여 기존의 BLUP추정방법 보다 약 11% 상향되는 것으로 추정되었다. 이러한 정확도 상승이 유전체 선발 기술을 현 종돈 개량체계에 적용하였을 경우의 어떠한 효과를 유발하는지 알아보기 위하여 유전적 개량량을 예측해 보았다.

백색계인 Landrace 품종은 모계 종축으로 사용되기 때문에 산자수는 매우 중요한 경제형질이라 할 수 있다. 그러나 산자수는 유전력이 매우 낮고 옹돈의 경우 표현형이 생성되지 않는다. 또한 모든 경우 분만을 해야 표현형이 생기므로 선발 간격이 길고, 근교퇴화에 민감하다는 요건을 가지고 있어 개량이 매우 어렵다. 따라서 산자수는 현 개량 체계에서 그림 47와 같이 종돈 선발시에 자신의 표현형이 없어 부모 육종가 평균을 이용하기 때문에 매우 낮은 정확도의 육종가 추정치 정보를 이용하고 있다.

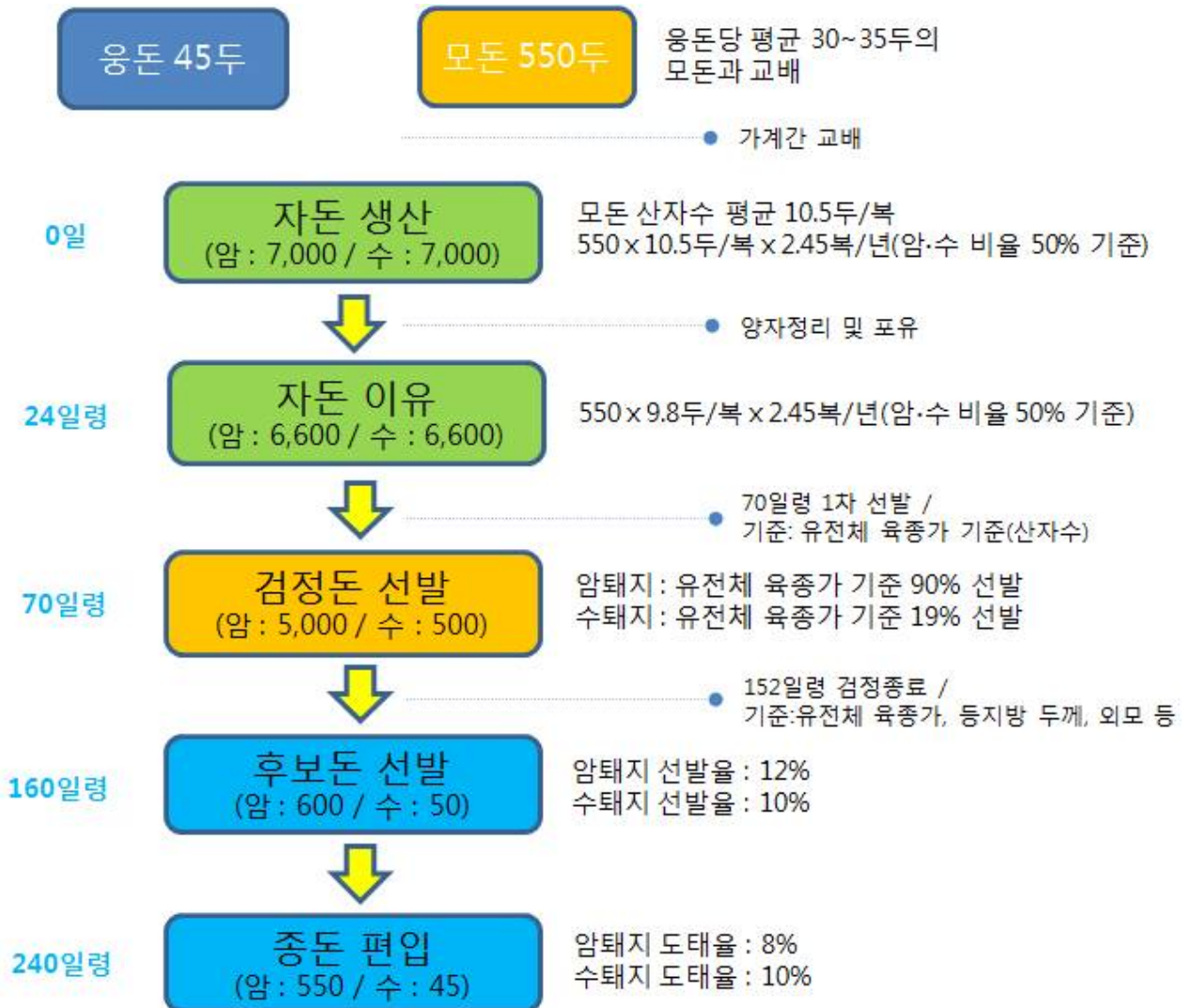


그림 47. GGP의 현재 검정 및 개량체계

현 체계에서의 연간 개량량을 아래와 같은 공식으로 추산하였다.

$$\Delta G = i\sigma_A r_{AP}$$

현행 산자수에 대한 선발은 검정돈으로부터 모든 선발시 혈통정보에 근거하여 이용되며 검정돈(암돼지) 5,000두중 모든 550두(최근들어 모든 교체율이 100%임), 검정돈(숫돼지) 500두중 45두를 선발하여 선발강도가 평균 1.77으로 선발을 실시하고 있다. 생존산자수를 기준으로 할 때에 유전표준편차는 0.9이고 검정돈의 경우 산자수가 없기 때문에 혈통정보를 이용한 육종가 추정의 정확도는 0.45으로 추정되었다. 또한 현행 체계에서 후보돈 선발에는 산자수 이외에 등지방 두께, 90kg 도달일령, 외모 등을 고려하는데 각 기준별 가중치를 1:1:1:1로 가정하고 세대간격을 1년으로 가정한다면 연간 개량량은 약 0.19두가 된다. 아래그림은 실제 종돈장에서 장기간 개량한 종돈의 개량량을 연도별로 계산한 값이다. 그림 48에서도 알 수 있듯이 실제 개량량은 연간 약 0.1두에 불과하였는데 이는 그동안 연간 모든 교체율이 약 60%에 불과하였고 시기에 따라 선발 지표가 다소 상이하였기 때문으로 판단된다.

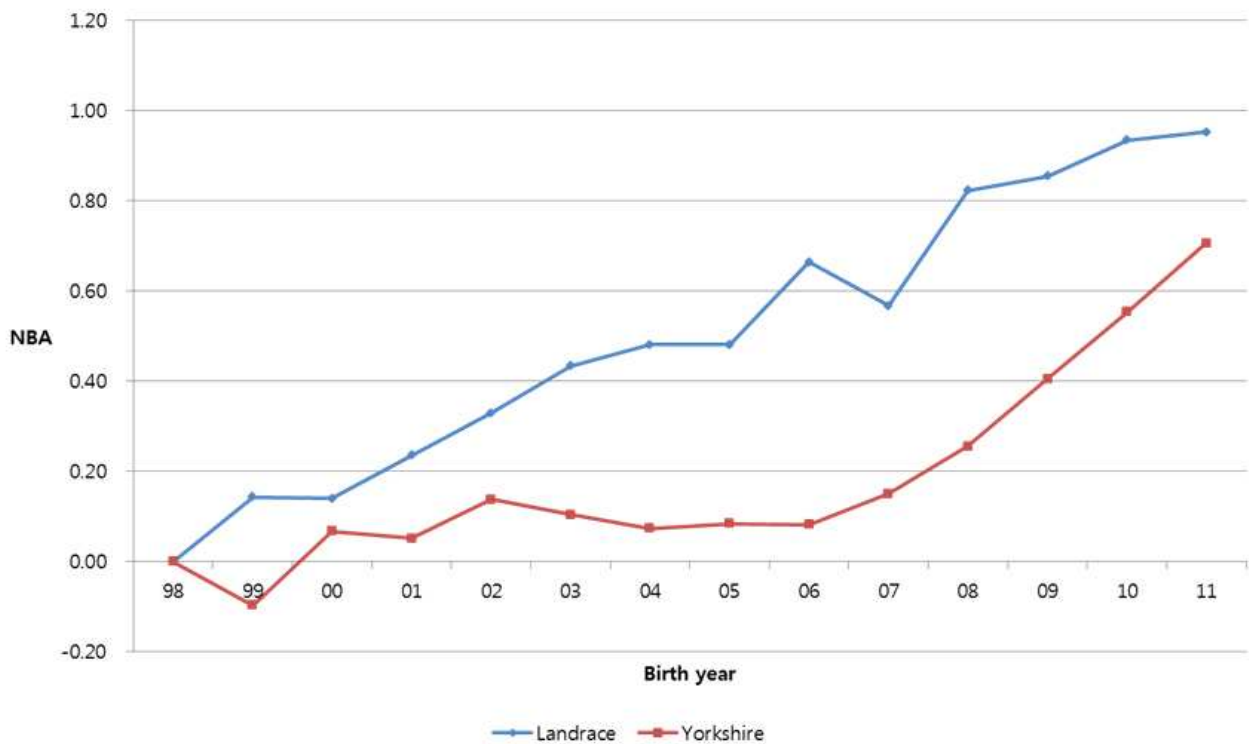


그림48. 생존산자수에 대한 품종별/년도별 개량량 추정치

표 58. 개량효율에 대한 현행의 종돈선발체계와 유전체 정보 활용시 종돈선발 체계 비교

항목	현재		개선(안)	
	산출내역	추정치	산출내역	추정치
유전표준편차	유전분산(0.928두 ²)의 제곱근	0.963	유전분산(0.928두 ²)의 제곱근	0.963
선발강도	모돈: 5000두 검정하여 550두 선발 (선발강도=1.76) 웅돈: 500두 검정하여 45두 선발 (선발강도=1.78)	1.77	모돈: 5390두 SNP 유전체 정보에 의한 육종가에 근거하여 550두 선발(선발강도=1.76) 웅돈: 2695두 유전체 SNP정보에 의한 육종가에 근거하여 45두 선발(선발강도=2.60)	2.18
정확도	연구결과에 의함	0.45	연구결과에 의함	0.56
세대간격	100%의 replacement rate가정시	1년	100%의 replacement rate가정시	1년
단일 형질에 대한 유전적 개량량	연간개량량 = 정확도×유전표준편차×선발강도	0.51두	연간개량량 = 정확도×유전표준편차×선발강도	1.18두
선발에 고려된 형질	4개 형질(NBA:D90kg:BF:Conform)을 각각 동일비중으로 선발을 가정하여 연간 개량량을 1/4함	0.19두	4개 형질(NBA:D90kg:BF:Conform)을 각각 동일비중으로 선발을 가정하여 연간 개량량을 1/4함	0.295두
실용돈기준	10두(GP용모돈)×20두(MSY/모돈)×0.19두×550두×450천원/두	94억원	10두(GP용모돈)×20두(MSY/모돈)×0.295두×550두×450천원/두	146억원

반면에 유전체 정보를 이용한다면 연간개량량은 약 0.295두에 해당하며 이는 현재 연간개량량 0.1두에 비하여 295%의 개량효율이 있는 것으로 평가된다. 또한 이에 대한 경제적 가치는 연간 약 146억원으로 현행보다 약 50억원의 추가이득의 효과가 있으며, 해당 집단의 실제 연간개량량인 0.1두의 경제적가치인 49억과 비교했을 때 약 100억원의 추가이득 효과가 있을 것으로 보인다. 반면에 유전체 정보를 얻기 위한 SNP분석에 소요되는 비용은 SNP분석비용은 두당 10만원 산정한다면 약 8억원의 소요비용이 추가 요구된다.



그림 49. GGP의 유전체 선발 적용 개량체계

현재 SNPchip 가격이 지속적으로 낮아지고 있으며, Imputation 기술이 발달함에 따라 현재 6K SNP개당 55천원 정도의 저밀도 chip을 이용하는 방법이 제안되고 있다. GGP의 경우 생산된 자돈들의 혈연관계가 잘 구성되어 있으나 동복자료가 많아 혈통정보를 이용할 경우 모두 같은 육종가로 추정되는 문제가 존재하기 때문에 멘델리안 샘플링이 고려되는 유전체 선발 방법을 도입하여 자돈 단계에서부터의 저밀도 chip을 이용한 유전체 선발을 적용하는 것이 가장 효율적일 것으로 사료된다.

3. 추정된 마커효과를 이용한 어린 후보축의 저밀도 SNP 기반 유전체 추정육종가 활용방안

가. 참조집단을 이용한 마커효과의 (u vector) 추정

기존의 구성된 참조집단으로부터 n개의 tag SNP을 이용한 저밀도 SNP 마커셋의 선별이 가능하며, ssBLUP을 이용하여 이들 마커에 대한 효과(u)를 추정한다. 이들 마커의 효과는 SNP유전자형에 대한 회귀 계수이며, 이를 통하여 후보군 개체의 유전체 추정육종가(GEBV)를 얻을 수 있다. 뿐만 아니라, 매년 선발된 후대검정우에 대해서는 유전자형 분석 자료를 업데이트하여 추정육종가의 정확도 향상에 투자한다.

나. 후보축의 저밀도 SNP 유전자형 (G matrix) 분석

참조집단으로부터 선정된 저밀도 SNP 마커셋의 칩을 제작하여, 후보축의 유전자형 분석을 수행한다. 이는 회귀계수의 계획행렬(G)에 해당하는 유전자형 행렬을 얻기 위한 과정이다. 각 SNP는 {aa, Aa, AA}의 유전자형에 따라, {-1, 0, 1} 또는 {0, 1, 2}로 코딩한다.

다. 어린 후보축의 유전체 추정 육종가 산출

위에서 얻은 G와 u를 곱하여 최종적으로 개체의 GEBV를 얻을 수 있다.

이를 풀어서 나타내면, 하나의 개체를 예로 들어 G를 구성하는 각 유전자형을 각각 $g_1, g_2, g_3 \dots g_n$ 이라 할 때, 각 마커의 효과를 각각 $u_1, u_2, u_3 \dots u_n$ 으로 놓으면, 개체의 유전체 추정 육종가는 다음과 같다.

$$GEBV = Gu = g_1u_1 + g_2u_2 + g_3u_3 + \dots + g_nu_n$$

이에 대한 예시를 아래 표 59과 같이 정리하면 다음과 같다.

표 59. 한우 근내지방도 형질의 유전체 추정육종가(GEBV) 추정(예시)

Index	Chr	Position(bp)	SNP_effect(u)	Major_allele	Genotype (G)						
					Animal_1	Animal_2	...	Animal_n			
SNP1	1	113,641	0.011994	A	AA	1	AT	0	...	TT	-1
SNP2	1	1,247,320	0.008446	G	AG	0	AG	0	...	GG	1
SNP3	1	1,791,640	0.001023	T	CC	-1	TT	1	...	TT	1
SNP4	1	1,936,698	-0.002504	T	TT	1	TG	0	...	TG	0
SNP5	1	2,019,348	0.001288	C	CA	0	CC	1	...	AA	-1
SNP6	1	2,223,442	0.004370	G	GG	1	GC	0	...	CC	-1
.
.
.
SNP6822	29	43,480,021	0.001668	G	AG	0	AG	0	...	GG	1
SNP6823	29	43,850,360	0.010366	C	CA	0	CC	1	...	AA	-1
GEBV*						0.013	0.013	...		-0.017	

*GEBV= SNP1_genotype X effect + SNP2_genotype X effect + ... +SNPn_genotype X effect

라. 유전체 추정 육종가 기반 개체 선발

이렇게 추정된 GEBV를 이용하여, 후보축의 개체선발의 지표로 사용한다. 선발에 탈락된 개체는 거세후 비육군으로 분류하여 활용하며, 표현형 데이터 수집에 활용한다. 마지막으로 GEBV 정확도 향상을 위하여 최종 선발된 개체는 저밀도가 아닌 기존의 상용화 chip (50-60K) 유전자형 분석을 수행한다.

제 4 장 목표달성도 및 관련분야에의 기여도

1. 연구개발 목표의 달성도

구분	연구개발의 목표	달성도(%)
1차년도 (2010)	○ 유전체 자료를 활용한 선발 효율성에 대한 모의시험으로 개량 효율성 검증	100
	○ 한우 및 돼지의 종축에 대한 유전체 및 검정자료의 분석	100
	○ 한우 후보종축의 부모와 자손의 유전체 정보에 대한 다형성 분석	100
	○ 돼지 GGP 부모의 유전체 정보에 대한 다형성 분석	100
2차년도 (2011)	○ 유전체 모의시험 정보를 이용한 육종가 계산 방법들의 정확도 비교 추가 연구(BayesC BayesC π 등) 및 프로그램 개발	100
	○ 한우 및 돼지의 종축에 대한 유전체 및 검정자료의 분석	100
	○ 한우 후보종축의 부모와 자손의 유전체 정보에 대한 다형성 분석	100
	○ 돼지 GGP 생산 검정돈의 유전체 정보에 대한 다형성 분석	100
3차년도 (2012)	○ 유전체 육종가 계산 알고리즘 추가 연구 및 프로그램 개발	100
	○ 개발된 프로그램을 이용하여 각 경제형질에 대한 종축의 유전체 육종가 추정 및 유용 표지인자의 탐색	100
	○ 한우 후보종축의 부모와 자손의 유전체 정보에 대한 다형성 분석	100
	○ 돼지 GGP 부모의 유전체 정보에 대한 다형성 분석	100
4차년도 (2013)	○ 한우 후보종축 및 자손의 유전체 유전능력 추정	100
	○ 돼지 모돈 및 후보돈의 유전체 유전능력 추정	100
	○ 한우 당대검정우의 사전선발을 위한 육종가 정확도 교차분석 검증	100
	○ 돼지 검정돈의 사전선발을 위한 육종가 정확도 교차분석 검증	100
	○ 한우 후보종축 및 자손의 DNA추출 및 유전체 정보 다형성 분석	100
	○ 돼지 GGP 부모 및 자돈의 DNA추출 및 유전체 정보 다형성 분석	100
5차년도 (2014)	○ 한우의 유전체 정보를 이용한 종축선발의 효율성 검증	100
	○ 유전체 정보를 이용한 한우의 검정방법 제시	100
	○ 돼지 검정돈의 사전선발을 위한 검정체계 제시	100
	○ 우수종돈 선발을 위한 유전체 정보 활용 유전능력 평가방법 제시	100
	○ 한우 후보종모우 및 당대검정 대상우의 유전체 정보 수집 및 분석	100
	○ 돼지의 유전체 정보가 미확보된 GGP 부모 및 생산 자돈들에 대한 유전체 정보 수집 및 분석	100
최종	○ 한우 및 돼지의 종축에 대한 DNA 분리 및 SNP 자료 수집 분석 ○ 한우 및 돼지의 유전체 선발의 정확도 검증 ○ 한우의 유전체 정보 활용 검정체계 제시로 산업화 방안 제시 ○ 돼지의 유전체 정보 활용 종돈 선발체계 제시	100

2. 정량적 연구개발 성과
가. 논문성과

계재 연도	논문명	저자			학술지명	Vol.(No.)	국내외 구분	SCI구분
		주저자	교신저자	공동저자				
2013	Genome-wide association analysis identifies quantitative trait loci for growth in a Landrace purebred population	정은지, 박희복	김병우 임현태	이재봉, 유채경, 김범모, 김혜인	Animal Genetics	45(3)442-444	국외	SCI
2013	종돈의 산육형질에 미치는 산차 및 환경요인의 효과추정	차대협	임현태	선두원, 이재봉, 이정규	농업생명과학연구	47(4)91-101	국내	비SCI
2013	유전체 관계행렬 구성에 따른 Landrace 순종돈의 육종가 비교	이준호	이득환	조광현, 조충일, 박경도	한국동물자원과학회지	55(3)165-171	국내	비SCI
2013	종돈의 번식형질에 미치는 산차 및 환경요인의 효과 추정	차대협	이정규	선두원, 김현권, 이재봉, 임현태, 김병우	농업생명과학연구	47(4)103-113	국내	비SCI
2012	돼지의 번식형질과 산육형질에 대한 유전모수 추정	조충일	이득환	안진국, 이준호	한국동물자원과학회지	54(1)9-14	국내	비SCI
2012	한우에서 전장의 유전체 정보를 활용한 연관불평형 및 유효집단크기 추정 에 관한 연구	조충일	이득환	이준호	한국생명과학회지	22(3)366-372	국내	비SCI
2011	Predicting the Accuracy of Breeding Values Using High Density Genome Scans	이득환	이득환	Daniel A. Vasco	Asian-Australasian Journal of Animal Sciences	24(2)162-172	국외	SCI
2011	가축 유전체정보 활용 종축 유전능력 평가 연구-표지인자 효과 추정 모의 실험	조충일	이득환		한국동물자원과학회지	53(1)1-6	국내	비SCI
2010	한우의 유전체 표지인자 활용 개체 혈연관계 추정	이득환	이득환	조충일, 김내수	한국동물자원과학회지	52(5)357-366	국내	비SCI

나. 특허 및 프로그램 등록 성과

등록연도	등록명	등록인	등록국	등록번호
2011	한우유전체 데이터 활용 집단의 다양성 분석 프로그램	한경대학교 산학협력단	컴퓨터프로그램 보호위원회	2011-01-123-002956
2012	유전체 육종가 추정 및 친자부정 탐색을 위한 유전체 관계행렬 생성	한경대학교 산학협력단	한국저작권위원회	C-2012-021177
2012	대용량 유전체 정보를 이용한 연관불평형 분석 및 유효집단크기 추정	한경대학교 산학협력단	한국저작권위원회	C-2012-021174
2012	5종류의 유전체 관계행렬 구성 및 혈연계수 출력	한경대학교 산학협력단	한국저작권위원회	C-2012-021176
2012	DNACHIP결과 자료의 이상치 제거 및 fastPHASE프로그램 구동을 위한 자료 구성	한경대학교 산학협력단	한국저작권위원회	C-2012-021172
2012	Single(싱글) Marker(마커) Regression(리그레션)을 이용한 표지인자의 표현형 연관성분석	한경대학교 산학협력단	한국저작권위원회	C-2012-021175
2012	fastPHASE프로그램 결과자료를 이용한 집단유전학적 통계 계산 및 allelic code 생성	한경대학교 산학협력단	한국저작권위원회	C-2012-021173
2014	한우의 단일염기다형을 분석하기 위한 유전자 조성물	한경대학교 산학협력단	특허청	10-2014-0159234

다. 학술발표

게재연도	논문명	발표자	학술회의명	Vol.(No.)	발표일시	장소	국내외구분
2010	폐쇄 종돈집단의 번식형질에 대한 장기간 선발 효과 추정	박성식,이득환	한국동물자원과학회	Proceedings vol. II/52p	2010-06-25	진주산업대학교	국내
2010	유전체정보 활용 개체 유전능력 평가의 정확도에 대한 모의실험	조충일,이득환,이준호	한국동물자원과학회	Proceedings vol. II/52p	2010-06-25	진주산업대학교	국내
2011	Bayes Cpi 방법을 이용한 Landrace 종축의 유전체 육종가 추정	이준호,이득환,조충일,박성식	한국동물자원과학회	Proceedings vol. II/89p	2011-06-24	부산대학교 밀양캠퍼스	국내
2011	모의실험을 통한 유전체 선발의 정확도 연구	이준호,이득환,조충일,박성식	한국동물자원과학회	Proceedings vol. II/88p	2011-06-24	부산대학교 밀양캠퍼스	국내
2011	한우에서 전장의 유전체 정보를 활용한 연관불평형 및 유효집단크기 추정에 관한 연구	조충일,이득환,이준호	한국동물자원과학회	Proceedings vol. II/65p	2011-06-23	부산대학교 밀양캠퍼스	국내

계재연도	논문명	발표자	학술회의명	Vol.(No.)	발표일시	장소	국내외구분
2012	유전체 선발, 연관불평형 및 양적형질좌위 탐색을 위한 고밀도 유전체 정보자료 처리 프로그램 개발	이준호,이득환, 조충일	한국동물자원과학회	Proceedings vol2/129p	2012-06-29	대전/충남대학교	국내
2012	유전체 관계행렬을 이용한 랜드레이스 종축의 유전체 육종가 추정	이준호,이득환	한국동물자원과학회	Proceedings vol2/129p	2012-06-29	대전/충남대학교	국내
2012	폐쇄돈군 선발에 따른 산자수 유전변이 효과 추정	이득환,손지현, 박주완	한국동물자원과학회	Proceedings vol2/128p	2012-06-29	대전/충남대학교	국내
2013	Comparison of linkage disequilibrium and effective population size for Landrace and native pig breeds in Korea	박종은,이득환, 이준호,손지현	WCAP2013	WCAP2013-1-01-029	2013-10-16	중국/베이징	국제
2013	Genome-wide association analysis identifies quantitative trait loci for growth in Landrace purebred population	정은지,임현태, 박희복,이재봉, 유채경,김범모, 김혜인,이정규	한국동물자원과학회	Proceedings vol. II/129p	2013-06-27	제주한화리조트	국내
2013	특정 폐쇄종돈 집단의 번식형질 및 산육형질에 대한 유전적 개량량 추정	박성식,이득환, 박주완,진석훈, 손지현	한국동물자원과학회	Proceedings vol. II/113p	2013-06-28	제주한화리조트	국내
2013	특정 폐쇄종돈 집단의 번식형질 및 산육형질에 대한 유전적 개량량 추정	박성식,이득환, 박주완,이규철, 손지현	한국동물자원과학회	Proceedings vol. II/148p	2013-06-28	제주한화리조트	국내
2013	특정 폐쇄종돈집단의 산자수에 대한 유전체 정보와 혈연정보를 활용한 육종가 추정치 비교	손지현,이득환, 이준호,박종은, 양희근	한국동물자원과학회	Proceedings vol. II/32p	2013-06-27	제주한화리조트	국내
2013	한우 도체형질에 대한 유전체 육종가 정확도 추정	조충일,이득환, 최연호,박병호, 윤종택,이학교	한국동물자원과학회	Proceedings vol. II/147p	2013-06-28	제주한화리조트	국내
2014	Estimation of linkage disequilibrium and effective population size using whole genome single nucleotide polymorphisms in Korean native pig X Landrace	박종은,이준호, 손지현,이득환	10th WCGALP	10th WCGALP / 57p	2014-08-18	캐나다/밴쿠버	국제
2014	Accuracy of genomic prediction in a simulated pig populations	손지현,박종은, 강현중,김자영, 이득환	10th WCGALP	10th WCGALP / 63p	2014-08-18	캐나다/밴쿠버	국제

라. 연구성과 활용

구분		기술실시(이전)	정책자료	언론홍보	기타홍보	교육지도	기타
활용건수	목표	2	2	1			
	달성	1 ⁽¹⁾		1 ⁽²⁾	1 ⁽³⁾	8 ⁽⁴⁻¹¹⁾	15 ⁽¹²⁾

- (1) ⁽¹⁾기술이전실시 : 소유권이전(특허:한우의 단일염기 다형을 분석하기 위한 유전자 조성물)
- (2) ⁽²⁾축산신문, 분자유전학적 정보를 이용한 가축개량방법 (2014년10월17일)
- (3) ⁽³⁾동물자원과학회, 학술발표수상(포스터최우수상) : 특정 폐쇄 종돈집단의 번식형질 및 산육형질들에 대한 유전적 개량량 추정
- (4) ⁽⁴⁾제 1회 유전체 육종연구 워크샵 : 유전체 선발의 기본 개념 및 분석 방법 (2010년2월18일)
- ⁽⁵⁾제 2회 유전체 육종연구 워크샵 : 유전체선발 분석 알고리즘 학습1 (2010년3월25일)
- ⁽⁶⁾제 3회 유전체 육종연구 워크샵 : 유전체선발 분석 알고리즘 학습2 (2010년4월29일)
- ⁽⁷⁾제 4회 유전체 육종연구 워크샵 : 유전체선발 분석 알고리즘 학습3 (2010년5월27일)
- ⁽⁸⁾제 5회 유전체 육종연구 워크샵 : 유전체선발 분석 알고리즘 학습4 (2010년6월23일)
- ⁽⁹⁾제 6회 유전체 육종연구 워크샵 : 유전체선발 분석 알고리즘 학습4 (2010년7월29일)
- ⁽¹⁰⁾제 7회 유전체 육종연구 워크샵 : 유전체선발 분석 알고리즘 학습 및 관련 프로그램 개발 연구(2010년10월8일)
- ⁽¹¹⁾유전체 연구 세미나:동물 유전체 정보를 활용한 가축개량 산업화 방안 (2011년05월03일~04일)
- (5) ⁽¹²⁾(주)선진한마을 유전체선발 및 종돈평가 기술지도 12회 (2009년 ~ 2014년)

마. 연구인력 활용/ 양성 성과

- (1) 박사학위 배출 : 3명
- (2) 석사학위 배출 : 4명
- (3) 학사학위 배출 : 8명

바. 장·단기 연수지원 성과

- (1) 단기국외 : 2명 2주(이준호, 조충일, Univ. of Iowa, 2011)
- (2) 단기국외 : 2명 2주(이득환, 조충일, Univ. of Georgia, 2012)

3. 연구개발 결과의 관련분야에의 기여도

가. 기술적인 측면

- 한우의 SNP자료 확보(50K이상 1,679두) - 유전체 정보 활용 종축선발을 위한 정보로 활용계획(향후 선발의 정확도 70%이상 목표)
- 한우의 SNP 정보를 활용, 연관불평형(LD)에 근거한 selected SNP(약 8K)정보를 활용하여 육종가 추정시 높은 추정의 정확도(MS=0.53) 향상
- 종돈의 SNP자료 확보(60K이상 1061두) - 유전체 정보 활용 종축선발을 위한 정보로 활용계획(향후 선발의 정확도 70%이상 목표)
- 종돈의 SNP정보를 활용, 연관불평형(LD)에 근거한 selected SNP(약 6K)정보를 활용하여 육종가 추정시 높은 추정의 정확도(생존산자수=0.55) 향상

- 한우 및 돼지의 개체별 유전능력평가 정확도(신뢰도) 향상으로 인한 유전적 개량량 상승 효과 및 DNA 수준의 조기 선발체계 구축

나. 산업적인 측면

- Selected SNP을 이용한 chip제작과 이를 이용하여 한우 암송아지의 유전체 육종가 추정 후 농가 송아지 선발에 활용(현재 초음파 분석사업을 대체)하여 한우개량사업의 효율성 강화
- 또한 Selected SNP Chip을 활용하여 당대검정 대상우(출생 신생송아지)들의 육종가를 추정하고 1차 선발을 실시함으로써 선발효율의 증대 - 기술이전 후 산업화 실시(예: 농협중앙회 등)
- Selected SNP을 이용한 chip제작과 이를 이용하여 검정대상 자돈의 유전체 육종가 추정 후 1단계 검정대상돈 선발로 선발의 정확도 및 선발강도 향상으로 개량효율 증대 및 양돈 농가의 소득증대
- 증대된 개량효율을 가진 우수한 종돈의 국산화로 수입종돈에 의한 해외 의존도 절감

제 5 장 연구개발 성과 및 성과활용 계획

1. 개발기술의 실용화 및 산업화 계획

- (1) 한우의 SNP자료 축적 - 향후 유전체 선발의 정확도가 70%이상 이 되도록 향상
- (2) 한우의 저밀도 chip(약 8K)제작 및 한우 암소에서 출생한 암송아지의 유전체 육종가 추정 후 농가 송아지 선발에 활용(현재 초음파 분석사업을 대체) - 기술이전 후 산업화 실시
- (3) 한우의 저밀도 Chip(약 8K)제작 및 한우 암소에서 출생한 수송아지의 유전체 육종가 추정 후 당대검정용 송아지 선발에 활용 - 기술이전 후 산업화 실시
- (4) 종돈의 SNP자료 축적 - 향후 유전체 선발의 정확도가 70%이상 이 되도록 향상
- (5) 종돈의 저밀도 Chip(약 6K)제작 및 모계종돈(모돈)에서 출생한 자돈의 유전체 육종가 추정 (이유시) 후 검정자돈 선발에 활용 - 기술이전 후 산업화 실시

2. 교육·지도·홍보 등 기술확산 계획 등

- (1) 유전체 정보 활용 지역단위 한우 암소개량을 위한 송아지 선발방법 정책사업 제안
- (2) 한우사업단 암소검정사업 실시 농가를 대상으로 유전체 선발 활용방법 기술지도 계획

3. 특허 및 논문 등 지식재산권 확보 계획

- (1) 한우의 단일염기다형을 분석하기 위한 유전자 조성물(한우의 저밀도 8K chip) - 2015년이 내 특허등록(2014.11월 특허 출원)
- (2) 돼지의 단일염기다형을 분석하기 위한 유전자 조성물(돼지의 저밀도 6K chip) - 2015년 특허 출원계획(마커선별 완료)
- (3) 유효집단 추정 및 유전체 선발의 저밀도 마커 이용시 정확도 비교 - 논문발표계획(2015년)

4. 추가연구, 타연구에 활용 계획 등

- (1) Golden Seed 연구과제에 연관하여 국산 종돈 개발에 활용
- (2) 진흥청 차세대바이오그린과제(동물유전체사업단)에 지속추진으로 한우 및 종돈의 추가 SNP자료 확보 및 연구수행

제 6 장 연구개발과정에서 수집한 해외과학기술정보

1. Iowa state University 방문 기술 정보 수집(조충일, 이준호)

- 유전체선발 기술 이론 및 Bayesian 방법을 활용한 유전체선발

Iowa state University 방문 기술 정보 수집내역

contents

1. Introduction, background, and Mathematical Foundation of Quantitative Genetics
2. introduction - Genomic selection in Livestock
3. LD_linkage : Genomic selection why does it work?
4. The problem of prediction genetic merit
5. genomic selection- Shrinkage of marker effects
6. Bayesian Methods in Genome Association Studies
7. Reliabilities
8. Genomic selection using Low-density SNPs
9. Response and inbreeding from Genomic Selection
10. R-reference card

↳(PDF 파일 Link)

2. UGA(University fo Georgia) 기술정보 수집내역(이득환, 조충일)

- GBLUP 및 ssGBLUP의 이론 및 활용방법 수집(프로그램 코드 등 수집)
- ssGBLUP 개발팀의 Dr. Shogo Tsuruta 국내초청

1

Genetic analysis using application (BGF90) programs in animal breeding

Shogo Tsuruta
Department of Animal and Dairy Science
College of Agricultural and Environmental Sciences
The University of Georgia
Athens, GA 30602 USA
Email: shogo@uga.edu
February 2013

Contents

0. Fundamental Statistics
1. Analysis of Variance
2. BLUP & BLUE
3. ML and REML
4. Bayesian (MCMC)
5. Practical modeling
6. Genomic selection
7. Application (BGF90) programs for BLUP and variance component estimation
8. References

└(PDF 파일 Link)

3. 제10회 WCGALP 참가 기술정보 수집

(10th World Congress on Genetics Applied to Livestock Production, in CANADA)

- 동물육종에서의 최신 연구동향 수집(Presentation(PDF파일) and Oral(동영상파일)별도첨부)

제 7 장 연구시설·장비 현황

- 해당사항 없음

제 8 장 참고문헌

- BROWNING, Brian L. BEAGLE 3.3. 2010.
- Badke, M. Y., Bates, R. O., Ernst, C. W., Schwab, C., and Steibel J. 2012. Estimation of linkage disequilibrium in four US pig breeds. *BMC Genomics* 2012, 13:24
- Christensen, O. F., P. Madsen, B. Nielsen, T. Ostersen and G. Su. 2012. Single-step methods for genomic evaluation in pigs. *Animal*. 6(10):1565–1571.
- CCSI. 2013. <http://www.swineimprovement.com/>
- Druet, T., Georges. M., DAG-PHASE. 2010.
- Du, F. X., Clutter, A. C. and Lohuis, M. M. 2007. Characterizing Linkage Disequilibrium in Pig Populations. *Int.J.Biol.Sci.* 3(3):166–178
- Emik, L. O. and Terrill, C. E. 1949. Systemic procedures for calculating inbreeding coefficients. *J. Hered.* 40:51–55
- Endelman, J. B. 2011. Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Gen.* 4:250–255
- Excoffier, L. and Slatkin, M. 1995. Maximum-likelihood-estimation of molecular haplotype frequencies in a diploid population. *Molecular Biology and Evolution*, 12(5):921 - 927
- Findhap. VanRaden, P. M. 2011
- Fernando, R.L., Bayesian Methods in Genome Association Studies. Technical report. Iowa State Univ. 2010.
- Goddard, M. 2009. Genomic selection : prediction of accuracy and maximisation of long term response. *Genetica*, 136:245–257
- GENSEL. Fernando R., Garrick D. J. 2008. (First Edition) (Third Edition 2009)
- Henderson, C. R. 1975. Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 31(2):423–447
- Hayes, B. J., Bowman, P. J., Chamberlain, A. C., Verbyla, K. and Goddard, M. E. 2008. Accuracy of genomic breeding values in multi-breed dairy cattle populations. *Genetic Selection. Evolution.* 41–51.
- Hayes, B. J., Visscher, P. E., McPartlan, H. and Goddard, M. E. 2003. A novel multi-locus measure of linkage disequilibrium and its use to estimate past effective population size. *Genome Research*, 13:635–643.
- Hill, W. G. and Robertson, A. 1968. Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics*, 38:226–231.
- Miształ, I., A. Legarra, I. Aguilar 2009. A relationship matrix including full pedigree and genomic information. *Journal of Dairy Science*, 92(9):4656–4663
- INTERBULL. 2014. <http://interbull2.slu.se/www/v1/>.
- Illumina. 2014. <http://www.illumina.com/>
- Lee, S. H., Cho, Y. M., Lim, D., Kim, H. C., Choi, B. H., Park, H. S., Kim, O. H., Kim, S., Kim, T. H., Yoon, D. and Hong, S. K. 2011. Linkage Disequilibrium and Effective

Population Size in Hanwoo Korean Cattle. *Asian-Aust. J. Anim. Sci.*, 24(12):1660–1665.

- Lee, D. H. and Daniel A. V. 2011. Predicting the Accuracy of Breeding Values Using High Density Genome Scans . *Asian-Australasian J. Animal Sci.* 24(2)062–172
- Legarra, A., and I. Misztal. 2008. Computing strategies in genomewide selection. *J. Dairy Sci.*, 91(1):360–366.
- Lewontin, R. C. 1964. The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics*, 49(1):49–67.
- Luo, Z. W. 1998. Detecting Linkage Disequilibrium Between a Polymorphic Marker Locus and a Trait Locus in Natural Populations. *Heredity*, 80(2):198–208
- McKay, S. D., Schnabel, R. D., Murdoch, B. M., Matukumalli, L. K., Aerts, J., Coppieters, W., Crews, D., Dias Neto, E., Gill, C. A., Gao, C., Mannen, H., Stothard, P., Wang, Z., Van Tassell, C. P., Williams, J. L., Taylor, J. F. and Moore, S. S. 2007. Whole genome linkage disequilibrium maps in cattle. *BMC Genetics*, 8–74
- McRae, A. F., McEwan, J. C., Dodds, K. G., Wilson, T., Crawford, A. M. and Slate, J. 2002. Linkage disequilibrium in domestic sheep. *Genetics*, 160(3):1113–1122.
- Meuwissen, T. H., Hayes, B. J. and Goddard, M. E. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4):1819–1829.
- REMLF90. 2001. <http://nce.ads.uga.edu/~ignacy/newprograms.html>.
- Mujibi, F. D. N., Nkrumah, J. D., Durunna, O. N., Stothard, P., Mah, J., Wang, Z., ... & Moore, S. S. 2011. Accuracy of genomic breeding values for residual feed intake in crossbred beef cattle. *Journal of animal science*, 89(11): 3353–3361.
- Gengler, N., S. Abras, C. Verkenne, S. Vanderick, M. Szydlowski, R. Renaville. 2007. Accuracy of Prediction of Gene Content in Large Animal Populations and its Use for Candidate Gene Detection and Genetic Evaluation. *Journal of Dairy Science* 91(4):1652–1659
- Resende Jr, M. F., Munoz, P., Resende, M. D. V., Garrick, D. J., Fernando, R. L., Davis, J. M., Jokela, E. J., Martin, T. A., Peter, G. F. and Kirst, M. 2012. Accuracy of Genomic Selection Methods in a Standard Data Set of Loblolly Pine. *Genetics*, 190(4):1503–0510.
- PIC. 2013. <http://www.picaustralia.com.au/>
- Qanbari, S., Pimentel, E. C. G., Tetens, J., Thaller, G., Lichtner, P., Sharifi, A. R. and Simianer, H. 2010. The pattern of linkage disequilibrium in German Holstein cattle. *Animal Genetics*, 41(4):346–356.
- Sargolzaei, M., Schenkel, F. S., Jansen, G. B. and Schaeffer, L. R. 2008. Extent of Linkage Disequilibrium in Holstein Cattle in North America. *Journal of Dairy Science*, 91(5):2106–2117.
- Sargolzaei, M., Schenkel, F. S., 2009. QMSim : a large-scale genome simulator for livestock. *Bioinformatics* 25(5):680–681
- Sargolzaei, M., Chesnais, J. P. & Schenkel, F. S. (2011). FImpute - An efficient imputation algorithm for dairy cattle populations. *Journal of Animal Science*. 89 (E-Suppl. 1)
- Sved, J. A. 1971. Linkage disequilibrium and homozygosity of chromosome segments in finite population. *Theoretical Population Biology*, 2(2):125–141.

- Stratz, P., Wellmann, R., Bennewitz, J. 2014. Strategies to implement genomic selection in pig breeding using very low marker density. Proceeding, 10th World Congress of Genetics Applied to Livestock Production
- Thevenon, S., Dayo, G. K., Sylla, S., Sidibe, I., Berthier, D., Legros, H., Boichard, D., Eggen, A. and Gautier, M. 2007. The extent of linkage disequilibrium in a large cattle population of western Africa and its consequences for association studies. *Animal Genetics*, 38(3):277-286
- Toosi, A., Fernando, R. L. and Dekkers, J. C. M. 2010. Genomic selection in admixed and crossbred populations. *Journal of Animal Science*, 88(1):32-46.
- VnaRaden, P. M. 2007. Genomic measures of relationship and inbreeding. *Interbull Bull.* 37:33-36.
- VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. *Journal of Dairy Science*, 91(11):4414-4423.
- VanRaden, P. M., Van Tassell, C. P., Wiggans, G. R., Sonstegard, T. S., Schnabel, R. D., Taylor, J. F., Schenkel, F. S. 2009. Invited review: Reliability of genomic predictions for North American Holstein bulls. *Journal of Dairy Science*, 92(1):16-24.
- Liu, M., M. E. Goddard, F. Reinhardt, and Reent, R. 2014. A single-step genomic model with direct estimation of marker effects. *J Dairy Sci.* 97(9):5833-5850.
- 국립축산과학원, 2014, 유전능력평가보고서32호
- 조충일, 최태정. 2013. 젓소 유전체 선발 효율성에 관한 모의실험 연구. *Journal of Agriculture & Life Science* 47(6):145-153.
- 조광현, 최태정, 조충일, 민홍립, 박상출, 박경도, 오재돈, 이학교, 이준호. 2014. 젓소 종모우 선발의 신뢰도 향상을 위한 MACE proofs의 활용. *Journal of Agriculture & Life Science* 48(3):166-172
- 조충일, 이준호, 박병호, 이득환. 2013. 제주재래돼지와 랜드레이스간 F2 교잡종 집단에서 성장형질과 전장의 유전체 정보간 연관성 분석. *Journal of Agriculture & Life Science* 47(2):75-84
- 이준호, 조광현, 조충일, 박경도, 이득환. 2013. 유전체 관계행렬 구성에 따른 Landrace 순종돈의 육종가 비교. *Journal of animal science and technology* 55(3):165-171
- 이승환, 조용민, 윤두학. 2014a. 한우 도체형질 연관 DNA marker 현장 검증연구. *Journal of Agriculture & Life Science* 48(2):133-141
- 이재봉, 서보영, 황정혜, 민찬식, 오기철, 임현태. 2014b. Rsb 분석법을 활용한 최소의 선발신호 탐색 연구. *Journal of Agriculture & Life Science* 48(4):219-229

<부록> 유전체 자료분석을 위한 프로그램 개발

1. 유전체 자료 전처리 프로그램

가. 기본 포맷

각 표지인자의 유전자형이 행으로, 개체별 유전자형이 열로 구성된 유전체 정보의 기본 포맷은 아래와 같으며, 염색체 번호, 위치, 표지인자 이름을 기본으로 정렬한다.

기본 포맷형성

염색체 번호, 위치, 표지인자 이름으로 정렬

염색체 번호 변환

Name	Chr	Position	Name	Chr	Position
ARS-BFGL-BAC-11602	0	0	ARS-BFGL-BAC-11602	99	0
ARS-BFGL-NGS-100091	X	95462301	ARS-BFGL-NGS-100091	31	95462301
ARS-BFGL-NGS-106614	Y	0	ARS-BFGL-NGS-106614	32	0
BovineHD3200000002	MT	190	BovineHD3200000002	33	190

Name	Chr	Position	7524405020_R01C01.Top Alleles	7524405020_R02C01.Top Alleles
Name	Chr	Position	225183380	225249147

Name	Chr	Position	225183380	225249147	225183294	225229450
BovineHD0100000005	1	16947	CC	CC	CC	CC
BovineHD0100000015	1	36337	AG	AA	AG	GG
BovineHD0100000024	1	67130	AA	AA	AA	AA
BovineHD0100000026	1	78655	AG	AA	AG	GG
BovineHD0100000027	1	83412	GG	GG	GG	GG
BovineHD0100046367	1	89725	AA	AA	AA	AA
BovineHD0100000030	1	100260	ZZ	ZZ	ZZ	AA
BovineHD0100000033	1	111645	AC	CC	AC	AA
BovineHD0100000035	1	120183	AC	AA	AC	CC
BovineHD0100000037	1	127885	AA	AA	AA	AA
Hapmap43437-BTA-101873	1	135098	AG	GG	AG	AA
BovineHD0100000038	1	137548	GG	GG	GG	GG
BovineHD0100000039	1	146011	AG	AA	AG	GG
BovineHD0100000040	1	147231	AG	AA	AG	GG
BovineHD0100000042	1	149772	AC	AA	AC	CC
BovineHD0100000043	1	151060	AG	GG	AG	AA
BovineHD0100000044	1	152374	AG	AA	AG	GG
BovineHD0100000046	1	155938	AG	AA	AG	GG
BovineHD0100000047	1	156978	AA	AA	AA	AA

Chip ID -> 등록 번호 변환

결측 유전자형 통일,

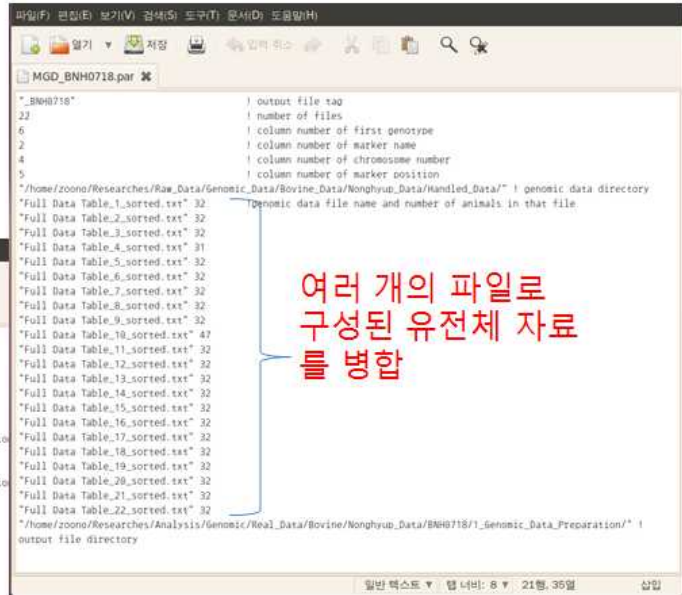
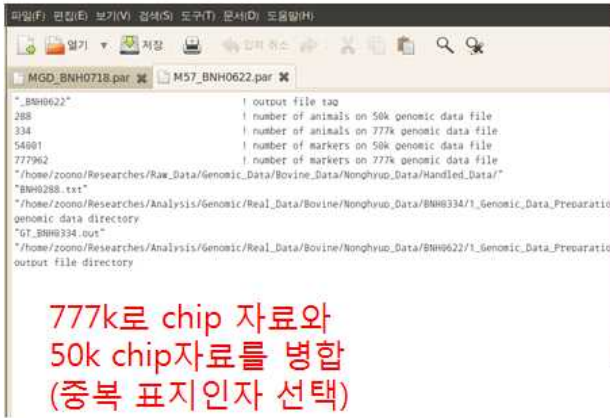
유전체 자료를 염색체 번호, 위치, 이름을 기준으로 하여 정렬 (excel 또는 awk, sort 이용)

나. 유전체 자료 병합, 700K • 50K 전환 병합 프로그램

소의 700K chip의 유전자형 같은 고밀도 유전체 정보는 그 정보량이 크기 때문에 여러 개의 파일로 분리되어 있는 경우가 많다. 각각의 분리된 파일을 정확하게 병합하고 이러한 병합시에 발생할 수 있는 오류를 점검하기 위한 프로그램을 개발하였다. 서로 밀도가 다른 50K, 700K 유전체 정보를 하나로 병합시키는 프로그램도 개발하였다. 50K, 700K chip에서 중복된 표지인자 자료만 추출하여 병합시키며, 중복 표지인자의 속성정보(염색체번호, 위치)가 다를 경우 이를 출력한다.

기본 포맷형성

유전체 자료 병합,
777k->50k 전환 병합



자료 병합에서 발생할 수 있는
오류 방지를 위하여 유전체 자료
병합 프로그램 개발

다. 이상치 제거, 기초통계 및 결측자료 대치 준비 프로그램
모든 표지인자에 대한 유전자 빈도를 계산한 뒤 이를 이용하여 Minor allele frequency 결측률,
Hardy_Weinberg disequilibrium chi_square를 계산하고 parameter로 입력한 기준치를 초과하는
표지인자 정보 및 개체 정보를 제거한다.

이상치 점검

표지인자 결측률, MAF,
HWD, 개체별 결측률

Marker Name	MarkerNO	SQSM	FOMGT	FONIG	FOHE	C50HWT	chromosome	position
ARS-BFGL-N65-19289	3	OL_F_H 0.0088876 0.0085441 0.0088859	239.811	1	305793			
ARS-BFGL-BAC-31497	16	OL_HWT 0.0108401 0.0250990 0.0363014	95.202	1	905632			
ARS-BFGL-BAC-36895	15	OL_F_H 0.0101626 0.0068446 0.0068446	360.233	1	1078256			
ARS-BFGL-BAC-31703	17	OL_MAF 0.0101626 0.0055021 0.0130048	0.063	1	1153822			
ARS-BFGL-N65-87636	17	OL_F_H 0.0142276 0.0051546 0.0034364	643.328	1	1683944			
Haomaz24070-BTA-123581	28	OL_HWT 0.0067751 0.0129604 0.0177353	137.994	1	1625471			
ARS-BFGL-BAC-754	29	OL_F_H 0.0108401 0.0065068 0.0102748	61.570	1	1712261			
ARS-BFGL-BAC-16253	38	OL_HWT 0.0033875 0.0380693 0.0611829	39.067	1	1933465			
ARS-BFGL-BAC-11152	40	OL_F_H 0.0088876 0.0085441 0.0102529	228.067	1	2013659			
ARS-BFGL-N65-75879	68	OL_F_H 0.0081301 0.0081967 0.0081967	359.975	1	3322254			
ARS-BFGL-N65-107481	69	OL_HWT 0.0676732 0.0108932 0.0116195	292.379	1	3493059			
ARS-BFGL-N65-109014	74	OL_HWT 0.0108401 0.0236301 0.0363014	66.422	1	3495838			
BTB-00003162	81	OL_MGR 0.1226287 0.3953668 0.4883889	0.028	1	3846214			
BTB-00002874	83	OL_MAF 0.0176152 0.0062069 0.0118345	16.157	1	3999035			
BTB-01160548	93	OL_M_H 0.0558266 0.4251969 0.0944882	330.586	1	4626587			
BTB-01160486	94	OL_F_H 0.0135501 0.0082418 0.0109890	156.445	1	4648383			
Haomao55613-rs29014454	117	OL_HWT 0.0866434 0.1104522 0.1245367	180.945	1	5803078			
BTB-01560572	118	OL_HWT 0.0054201 0.0102188 0.0136240	156.445	1	5821018			
BTB-01560539	119	OL_F_H 0.0094851 0.0075239 0.0123119	45.809	1	5849779			
BTB-01489336	122	OL_HWT 0.0142276 0.0127148 0.0213058	33.339	1	5997487			
BTA-44248-no-rs	129	OL_HWT 0.0108401 0.0126712 0.0138137	336.238	1	6450803			
UA-IFASA-8265	130	OL_HWT 0.0067751 0.0159602 0.0173353	75.091	1	6492449			
BTA-44295-no-rs	131	OL_F_H 0.0162062 0.0099062 0.0130854	166.099	1	6526757			
BTA-44295-no-rs	137	OL_HWT 0.0063057 0.0190565 0.0207531	78.954	1	6940697			
BTA-44295-no-rs	141	OL_F_H 0.0054201 0.0081744 0.0136240	37.486	1	7162655			
BTA-44295-no-rs	146	OL_HWT 0.0149051 0.0574278 0.0942228	24.444	1	8148041			
ARS-BFGL-N65-41048	160	OL_F_H 0.0128226 0.0065202 0.0075498	253.664	1	8266939			
BTA-52266-no-rs	172	OL_HWT 0.0514965 0.3287014 0.3214286	103.203	1	8484167			
Haomao49509-BTA-17127	187	OL_M_H 0.1938894 0.4611779 0.2292191	345.766	1	10494823			
Haomao68823-rs29019235	189	OL_F_H 0.0047425 0.0098707 0.0129340	168.119	1	10587853			
ARS-BFGL-BAC-6567	191	OL_HWT 0.0101626 0.0184805 0.0314853	25.499	1	10717144			
ARS-BFGL-BAC-6567	195	OL_F_H 0.0094851 0.0075239 0.0123119	104.393	1	11109134			

유효한 표지인자가 제거되지 않도록 재점검

표지인자 이상치 제거 후 개체별 유전체 정보의 결측률을 계산하여 기준치를 초과하는 개체의 유전체 정보를 제거한다.

이상치 점검

표지인자 결측률, MAF, HWD, 개체별 결측률

제거된 개체별 유전체 자료와 제거사유 점검

Animal ID	Animal NO.	NO. Missing	Missing rate
221960935	288	12182	0.3510663
221960939	289	9616	0.2771181
221998274	290	9471	0.2729395
221960937	291	9439	0.2720173
221960152	292	11544	0.3326801
221998056	293	13686	0.3944992
221990450	294	10438	0.3088069
221990452	295	10279	0.2962248
221990448	296	10823	0.3119020
221990449	297	11102	0.3199424
221990453	298	14076	0.4056484
221965017	299	8801	0.2536311
221977165	388	18120	0.5221902
223009114	389	21764	0.6272046
223004478	390	19293	0.5559942
223004479	391	22512	0.6487608
223005738	398	8946	0.2578098
223004480	434	22090	0.6365904
223230603	514	16113	0.4643516
223182261	515	18569	0.5351297
223226071	516	14459	0.4166859
223182286	517	16517	0.4759942
223220753	518	15039	0.4334006
223220729	519	11389	0.3282132
223182157	520	19039	0.5486743
223233375	551	12751	0.3074640

이상치 점검

Missing genotype imputation 준비

이상치를 제거한 이후 나머지 결측치의 대치 fastPHASE 또는 BEAGLE 프로그램 이용
"?"로 되어있는 유전자형이 대치됨
BEAGLE format

fastPHASE format

BEAGLE format

이후 fastPHASE 또는 BEAGLE 프로그램 입력 포맷으로 변환하여 염색체별 파일을 형성한다.

Chip 분석 결과 기초통계

이상치 발생 사유별 빈도, 염색체별 집단유전 통계, 유전체 자료 밀도

이상치 발생 원인 별 빈도 분석, 보유 자료의 유효성 파악

```

*****Basic statistics*****
Number of markers on chromosome X          938
Number of markers on chromosome Y          1
Number of markers with Unknown marker location(position) 428
Total number of markers                    44899
Number of markers on autosomal chromosomes 43548
Number of selected(useful) markers         34700
Number of outlier markers                  10199
Number of animals with missing over       0.25% 26
Total missing rate(Number of Missing genotypes/Number of genotypes(NO.animal*NO.markers)) 0.0185548197

*** total ***
Number of all missing markers              0
Number of all homo_genotype markers       4
Number of all hetero_genotype markers     0
Number of all homo_genotype markers and missing over 10.00% 0
Number of all hetero_genotype markers and missing over 10.00% 0
Number of markers with no missing        0
Number of markers with missing over       10.00% 516
Number of markers with minor allele frequency less than 0.01 5117
Number of markers with HWchisquare over 23.93 8928
Number of markers with missing over 10.00% and HWchisquare over 23.93 253
Number of markers with minor allele frequency less than 0.01 and HWchisquare over 23.93 4508
Number of markers with missing over 10.00% and minor allele frequency less than 0.01 24
Number of markers with missing over 10.00%, minor allele frequency less than 0.01 and HWchisquare over 23.93 21
    
```

이상치 발생 원인별 빈도 결과를 출력하여 보유 유전체 자료의 유효성 현황을 파악한다.

Chip 분석 결과 기초통계

이상치 발생 사유별 빈도, 염색체별 집단유전 통계, 유전체 자료 밀도

Minor allele frequency 분포, 표지인자사이의 거리 빈도(이상치 제거 전,후)

```

***** frequency of MAF *****
number of total markers with Minor allel frequency
  minor allele frequency < 0.1 : 14886
  0.1 <= minor allele frequency < 0.2 : 7517
  0.2 <= minor allele frequency < 0.4 : 7037
  0.3 <= minor allele frequency < 0.4 : 6972
  0.4 <= minor allele frequency : 7128

number of selected markers with Minor allel frequency(except for unusual data)
  minor allele frequency < 0.1 : 6659
  0.1 <= minor allele frequency < 0.2 : 7338
  0.2 <= minor allele frequency < 0.4 : 6878
  0.3 <= minor allele frequency < 0.4 : 6827
  0.4 <= minor allele frequency : 6998

***** frequency of distance *****
number of total markers with distance between adjacent markers
  distance between adjacent markers < 1 : 53
  1 <= distance between adjacent markers < 10 : 176
  10 <= distance between adjacent markers < 100 : 38142
  100 <= distance between adjacent markers < 1000 : 5130
  1000 <= distance between adjacent markers : 10

number of selected markers with distance between adjacent markers(except for unusual data)
  distance between adjacent markers < 1 : 52
  1 <= distance between adjacent markers < 10 : 127
  10 <= distance between adjacent markers < 100 : 27539
  100 <= distance between adjacent markers < 1000 : 6939
  1000 <= distance between adjacent markers : 14
    
```

이상치 제거 전 • 후의 MAF와 인접 표지인자사이의 거리 빈도를 출력하여 비교할 수 있게 한다.

Chip 분석 결과 기초통계

이상치 발생 사유별 빈도, 염색체별 집단유전 통계, 유전체 자료 밀도

(이상치 제거 전)

- 염색체별 표지인자수,
- 평균거리,
- 평균 MAF,
- F-statistics
- HW_chi값
- 평균 PIC
- 평균 Observed heterozygosity
- 평균 expected heterozygosity

Chromosome NO.	NOM	AODAMB	AOMAF	AO18C	AOCOSHW	AOPIC	AOHE	AOEH
chromosome 1	2925	53.6	0.20665	0.07406	42.88392	0.22466	0.27562	0.27894
chromosome 2	2354	57.6	0.20485	0.07474	42.69677	0.22155	0.27234	0.27541
chromosome 3	2062	58.3	0.20960	0.06598	37.59238	0.22812	0.28887	0.28341
chromosome 4	2172	55.0	0.20843	0.07711	46.34416	0.22292	0.27560	0.27783
chromosome 5	1848	65.1	0.20211	0.07856	46.67067	0.21872	0.26885	0.27172
chromosome 6	2221	53.1	0.21610	0.06394	35.94174	0.23183	0.28689	0.28889
chromosome 7	1944	57.3	0.20496	0.07570	42.41660	0.22687	0.27092	0.27472
chromosome 8	1998	56.1	0.20501	0.07130	42.23496	0.22365	0.27454	0.27750
chromosome 9	1776	59.0	0.20007	0.07451	43.16001	0.21800	0.26713	0.27006
chromosome 10	1789	57.3	0.20486	0.07028	41.80390	0.22181	0.27328	0.27547
chromosome 11	1858	57.2	0.20835	0.07677	44.95185	0.22313	0.27551	0.27812
chromosome 12	1468	61.6	0.20282	0.08123	46.61858	0.22065	0.27004	0.27403
chromosome 13	1498	55.5	0.20504	0.06716	37.77251	0.22230	0.27387	0.27612
chromosome 14	1569	52.5	0.21354	0.06793	42.26511	0.23039	0.28584	0.28692
chromosome 15	1412	59.4	0.20345	0.07514	43.27602	0.22155	0.27218	0.27521
chromosome 16	1397	57.7	0.19946	0.08524	51.17747	0.21584	0.26490	0.26799
chromosome 17	1304	57.0	0.20256	0.07135	40.52315	0.22264	0.27209	0.27586
chromosome 18	1086	59.8	0.21012	0.07377	42.86182	0.22647	0.27938	0.28205
chromosome 19	1120	56.3	0.21190	0.06997	42.35153	0.23022	0.28368	0.28622
chromosome 20	1299	54.6	0.20995	0.07547	42.50047	0.22619	0.27742	0.28136
chromosome 21	1187	59.5	0.21168	0.08027	49.43174	0.22787	0.28076	0.28397
chromosome 22	1088	55.8	0.19834	0.08230	50.25576	0.21610	0.26452	0.26766
chromosome 23	885	58.6	0.22059	0.05938	33.72194	0.23637	0.29159	0.29470
chromosome 24	1097	56.2	0.20983	0.06835	39.66836	0.22718	0.27999	0.28255
chromosome 25	811	52.2	0.22294	0.07318	45.85807	0.23322	0.29051	0.29235
chromosome 26	907	56.4	0.20668	0.07928	48.61783	0.22409	0.27559	0.27858
chromosome 27	805	55.9	0.21005	0.07333	43.83945	0.22522	0.27808	0.28057
chromosome 28	788	58.1	0.20914	0.07030	40.14561	0.22618	0.27821	0.28132
chromosome 29	872	58.6	0.20604	0.06650	37.92678	0.22268	0.27507	0.27687
total	43540	57.0	0.20727	0.07328	42.85614	0.22412	0.27580	0.27865

이상치 제거 전의 유전체 구성에 관한 기초 통계를 출력한다.(이상치 제거 전후 비교용)

Chip 분석 결과 기초통계

이상치 발생 사유별 빈도, 염색체별 집단유전 통계, 유전체 자료 밀도

(이상치 제거 후)

- 염색체별 표지인자수, 제거사유별 빈도
- 평균거리,
- 평균 MAF, 평균 F-statistics
- 평균 HW_chi값, 평균 PIC
- 평균 Observed heterozygosity
- 평균 expected heterozygosity

Chromosome NO.	NOM	NOSM	NOOM	NOOM_M	NOOM_F	NOOM_H	NOOM_M_H	NOOM_F_H	NOOM_M_F	NOOM_MH	NOOM_AH	NOOM_AH	AODAMB	AOMAF	AO18C	AOCOSHW	AOPIC	AOHE	AOEH
chromosome 1	2925	2345	588	24	310	528	12	270	1	1	0	0	66.9	0.25034	0.00996	2.09212	0.26950	0.33493	0.33591
chromosome 2	2354	1874	480	18	289	433	7	253	2	2	0	0	72.3	0.24962	0.01050	2.15333	0.26758	0.33286	0.33385
chromosome 3	2062	1679	383	26	197	344	9	175	2	2	0	0	71.7	0.24870	0.00892	2.06242	0.26863	0.33434	0.33485
chromosome 4	2172	1724	448	10	251	417	4	226	1	1	0	0	69.2	0.25612	0.00877	2.15548	0.27195	0.33897	0.33918
chromosome 5	1848	1447	401	28	242	363	13	219	2	2	0	0	83.2	0.24903	0.00767	1.99984	0.26672	0.33271	0.33267
chromosome 6	2221	1825	396	21	212	359	11	185	0	0	0	0	64.0	0.25666	0.00924	2.13534	0.27254	0.33996	0.34071
chromosome 7	1944	1538	414	28	228	365	14	192	1	0	0	0	73.0	0.25159	0.01129	2.08935	0.26834	0.33354	0.33506
chromosome 8	1998	1593	405	21	215	371	11	191	1	1	0	0	70.5	0.24859	0.00637	1.87906	0.26863	0.33479	0.33461
chromosome 9	1776	1403	373	23	201	335	12	173	2	1	0	0	74.4	0.24464	0.00838	2.12943	0.26421	0.32815	0.32841
chromosome 10	1789	1434	355	21	201	321	8	180	0	0	0	0	71.4	0.24733	0.00712	2.04291	0.26565	0.33117	0.33190
chromosome 11	1858	1458	400	19	233	360	5	207	0	0	0	0	73.1	0.25701	0.00820	2.07263	0.27265	0.34084	0.34116
chromosome 12	1468	1149	319	14	167	291	8	145	0	0	0	0	78.7	0.25131	0.01166	2.40197	0.27040	0.33602	0.33733
chromosome 13	1498	1206	292	15	159	269	10	141	2	2	0	0	69.1	0.24795	0.00716	2.08624	0.26640	0.33234	0.33214
chromosome 14	1569	1269	300	15	175	276	8	150	3	3	0	0	65.1	0.25814	0.00407	1.90773	0.27596	0.34508	0.34490
chromosome 15	1412	1107	305	9	173	278	5	150	0	0	0	0	76.0	0.24914	0.00771	2.09546	0.26872	0.33514	0.33505
chromosome 16	1397	1076	321	20	196	288	11	172	1	1	0	0	75.1	0.25015	0.00838	2.11079	0.26747	0.33360	0.33358
chromosome 17	1304	1055	249	14	132	225	4	118	0	0	0	0	70.5	0.24113	0.01119	2.07910	0.26344	0.32579	0.32726
chromosome 18	1086	860	226	10	120	204	3	105	0	0	0	0	75.2	0.25752	0.00900	2.17817	0.27432	0.34258	0.34315
chromosome 19	1120	899	221	11	104	198	4	88	0	0	0	0	70.2	0.25491	0.00655	1.79163	0.27469	0.34284	0.34277
chromosome 20	1299	1040	259	16	136	240	7	126	0	0	0	0	68.1	0.25376	0.01133	2.12463	0.27088	0.33676	0.33824
chromosome 21	1187	938	249	13	139	225	5	123	0	0	0	0	75.5	0.26074	0.00851	1.90775	0.27749	0.34648	0.34734
chromosome 22	1088	853	235	10	125	211	4	107	0	0	0	0	71.3	0.24449	0.00923	2.15130	0.26369	0.32745	0.32791
chromosome 23	885	731	154	12	75	137	6	64	1	1	0	0	70.8	0.25606	0.00754	1.97052	0.27378	0.34104	0.34217
chromosome 24	1097	875	222	12	119	197	5	101	0	0	0	0	70.6	0.25290	0.00765	1.84760	0.27210	0.33968	0.33948
chromosome 25	811	640	171	15	103	147	2	91	2	1	0	0	66.1	0.27210	0.00562	1.79755	0.28265	0.35546	0.35544
chromosome 26	907	713	194	10	101	170	1	86	1	1	0	0	71.0	0.25219	0.00887	2.07835	0.27183	0.33847	0.33899
chromosome 27	805	643	162	4	90	151	2	81	0	0	0	0	78.1	0.25601	0.00665	1.88124	0.27228	0.34038	0.34038
chromosome 28	788	632	156	14	82	149	9	80	0	0	0	0	72.6	0.25136	0.00582	1.95711	0.26959	0.33715	0.33642
chromosome 29	872	702	170	7	108	150	2	93	1	1	0	0	72.3	0.24947	0.00856	2.33886	0.26753	0.33352	0.33370
total	43540	34700	8840	460	4883	8002	202	4300	23	20	0	0	71.6	0.25185	0.00856	2.06359	0.26986	0.33634	0.33676

이상치 제거 이후의 통계를 출력하며, 각 이상치 제거 사유별 빈도를 염색체별로 출력한다.

다. 프로그램 코드

(1)DNACHIP결과 자료의 이상치 제거 및 fastPHASE프로그램 구동을 위한 자료 구성

- DNACHIP 분석결과인 대용량의 SNP 표지인자 유전자형 자료 중 이상치 제거
- 표지인자 결측률, minor allele frequency, H_W test chi_square값, 개체 결측률의 이상치 판정 기준을 입력받아 기준에 부합되는 자료 제거
- 전체 및 선발된 표지인자 자료의 기초통계량 계산 및 출력
- 결측자료 대치 프로그램인 fastPHASE연동을 위한 자료 format 변경

```
! Name : Outlier Elimination, Basic Statistics of Genomic Raw Data & Preparation for fastPHASE
! Developed by : Joonho Lee - PhD working in Hankyong National Univ.(Korea) zoonolee@gmail.com,
+82-10-3408-2895
! Version : Linux 4.0 (2012-03-16)
! Contents :      1. Data reading(animal ID, SNP marker name, chromosome number, marker position,
phenotype(name, records))
!
!                  2. Calculation of gene & genotype frequency
!                  3. Unusual data elimination
!                  4. Statistics of data & population genetics
!                  5. Transformation of genotype data for fastPHASE
program OEGRDPF
integer::io,i,j,k,l,m,n,o,p=0,q=0,r,iz,i1,i2,SMNBC=1 liostate & do_loop
integer::NOAC=0,XCN=19,YCN=20,UCN=99,ANO=0,EXCN=999 !number of autosomal chromosomes & chromosome
number for X, Y, Unknown
integer::BC(5)
integer::GC(11)
integer::NOA,NOZ=0,NOZBC=0,NANO=0,NOMOAC=0,NOM,NOMAG=0,NOMIG=0,NOMIH=0,NOHE=0,NOMAH=0,
NOMG=0,NOMOXC=0,NOMOYC=0,NOMOUC=0,TNOMG=0 !counting
integer::NOAMM=0,NOAMMBC=0,NONMM=0,NODM=0,NODMBC=0,NODA=0,NODMBOFOMGT=0,NODMBOAH
=0,NODMBOFOMIG=0,NODMBOCSOHWT=0,NODMBOAHE=0 !counting
integer::ACNONMM=0,ACNOAMM=0,ACNODMBOFOMGT=0,ACNODMBOAH=0,ACNODMBOFOMIG=0,ACNOD
MBOCSOHWT=0,ACNODMBOAHE=0 !counting
integer::UCNONMM=0,UCNOAMM=0,UCNODMBOFOMGT=0,UCNODMBOAH=0,UCNODMBOFOMIG=0,UCNOD
MBOCSOHWT=0,UCNODMBOAHE=0 !counting
integer::SCNONMM=0,SCNOAMM=0,SCNODMBOFOMGT=0,SCNODMBOAH=0,SCNODMBOFOMIG=0,SCNODM
BOCSOHWT=0,SCNODMBOAHE=0 !counting
integer::NODMBO_M_H=0,NODMBO_F_H=0,NODMBO_M_F=0,NODMBO_MHF=0,NODMBO_HOM=0,NODMBO_
HEM=0 !counting
integer::ACNODMBO_M_H=0,ACNODMBO_F_H=0,ACNODMBO_M_F=0,ACNODMBO_MHF=0,ACNODMBO_HO
M=0,ACNODMBO_HEM=0 !counting
integer::UCNODMBO_M_H=0,UCNODMBO_F_H=0,UCNODMBO_M_F=0,UCNODMBO_MHF=0,UCNODMBO_HO
M=0,UCNODMBO_HEM=0 !counting
integer::SCNODMBO_M_H=0,SCNODMBO_F_H=0,SCNODMBO_M_F=0,SCNODMBO_MHF=0,SCNODMBO_HOM
=0,SCNODMBO_HEM=0 !counting
integer::NODMBOFOMGTBC=0,NODMBOAHBC=0,NODMBOFOMIGBC=0,NODMBOCSOHWTBC=0,NODMBOA
HEBC=0 !counting
integer::NODMBO_M_HBC=0,NODMBO_F_HBC=0,NODMBO_M_FBC=0,NODMBO_MHFBC=0,NODMBO_HOMBC
=0,NODMBO_HEMBC=0 !counting
integer::NOSMWMAF1=0,NOSMWMAF2=0,NOSMWMAF3=0,NOSMWMAF4=0,NOSMWMAF5=0,NOTMWMAF1
=0,NOTMWMAF2=0,NOTMWMAF3=0,NOTMWMAF4=0,NOTMWMAF5=0
integer::NODBAM1=0,NODBAM2=0,NODBAM3=0,NODBAM4=0,NODBAM5=0,NODBASM1=0,NODBASM2=0,NO
DBASM3=0,NODBASM4=0,NODBASM5=0
integer::CMPBC=0,EXMPBC=0
integer::CMPBC_S=0,EXMPBC_S=0
```

```

integer*8::SODBAMBC=0,SOTDBAM=0,SODBAMBC_S=0,SOTDBAM_S=0
integer::NOMBC=0,MDG(4) ! for statistics
integer,allocatable::IOM(:,:),GN(:,:),NOMGBA(:),NOSMBC(:) !data input & missing data output
real::GF(5),GTF(11),FOMAH,FOMIH,FOHE,FOMAG,FOMIG,SOGF,SOGTF,FOMG,FOMGT,IOL(4),SOPIC=0,AOPI
C=0 !frequencies
real::DBAMBC=0,DBAMBC_S=0
real::ENOMIH=0,ENOHE=0,ENOMAH=0,CSOHW=0,EH=0,PIC=0,IBC=0,TMGR=0,AODBAMBC_S=0,AODBAMBC
=0,AOTDBAM_S=0,AOTDBAM=0 !statistics for population genetics
real::SOMAFBC_S=0,SOIBCBC_S=0,SOCISOHW=0,SOPICBC_S=0,SOHEBC_S=0,SOEHBC_S=0
real::AOMAFBC_S=0,AOIBCBC_S=0,AOCISOHW=0,AOPICBC_S=0,AOHEBC_S=0,AOEHBC_S=0
real::SOTMAF_S=0,SOTIBC_S=0,SOTCSOHW=0,SOTPIC_S=0,SOTHE_S=0,SOTEH_S=0 !statistics for
population genetics by chromosome
real::AOTMAF_S=0,AOTIBC_S=0,AOTCSOHW=0,AOTPIC_S=0,AOTHE_S=0,AOTEH_S=0 !statistics for
population genetics by chromosome
real::SOMAFBC=0,SOIBCBC=0,SOCISOHW=0,SOPICBC=0,SOHEBC=0,SOEHBC=0,SOTMAF=0,SOTIBC=0,SO
TCSOHW=0,SOTPIC=0,SOTHE=0,SOTEH=0 !statistics for population genetics by chromosome
real::AOMAFBC=0,AOIBCBC=0,AOCISOHW=0,AOPICBC=0,AOHEBC=0,AOEHBC=0,AOTMAF=0,AOTIBC=0,
AOTCSOHW=0,AOTPIC=0,AOTHE=0,AOTEH=0 !statistics for population genetics by chromosome
character(len=50)::MN, D1, D2, D3
character(len=1)::BA(5)=('A','T','G','C','Z'),MABA,MIBA,MG
character(len=2)::GT(11)=('AA','AT','AG','AC','TT','TG','TC','GG','GC','CC','ZZ'),MAH,MIH,HE,CHNO
character(len=200)::GDD,ODD,IDD,PDD
character(len=30)::GDFN,WF
character(len=15)::MWF(5)
character(len=6)::SOSM="NORMAL"
character(len=8)::OFT="_DEF0000"
character(len=230)::GDFNIP,ODFNIP,IDFNIP,PDFNIP,PFIP
character(len=20),allocatable::ID(:)
character(len=2),allocatable::GTS(:)
character(len=1),allocatable::GTDFFP(:,:)
!reading parameters
print
print *, "*****"
print *, "*****"
print *, "OEGRDPF_Lv4 : Outlier Elimination of Genomic Raw Data & Preparation for fastPHASE Linux
version 4.0"
print *, "Developed by Joonho Lee : zoonolee@gmail.com, +82-10-3408-2895"
print *, "Data format : (1st row) Markername Chromosome Position LL09066048 LL08131059
LL09147068 ... (Animal IDs)"
print *, "Data format : (2nd row) ALGA0000009 1 52297 AG GG GG . . .
(Genotypes)"
print *, "Input data must be sorted by chromosome No.(1st) and Marker position(2nd)"
print *, "Sex chromosome and markers of unknown chromosome must be numbered more then the biggest
chromosome number"
print *, "Length of Output file tag must be 8 characters"
print *, "Please enter the name of parameter file including directory(ex) ./parGRDH"
print
print *, "*****"
print *, "*****"
read *, PFIP
PFIP=adjustl(PFIP)
open (99, file=PFIP)
read (99,*) OFT !Output file tag
read (99,*) NOA

```

```

read (99,*) GDD
read (99,*) GDFN
read (99,*) ODD
read (99,*) IDD
read (99,*) PDD
read (99,*) GT(11)
read (99,*) IOL(:)
read (99,*) NOAC
read (99,*) XCN, YCN, UCN
read (99,*) DXCN, DYCEN, DUCN
read (99,*) MDG(:)
print *, "number of animals", NOA
print *, "directory of genomic data file", GDD
print *, "name of genomic data file?", GDFN
print *, "directory of output file", ODD
print *, "missing genotype", GT(11)
print *, "range of outliers"
print *, "  marker-missing% : ", IOL(1)
print *, "  minor allele frequency : ", IOL(2)
print *, "  HW-chisquare(df=1) : ", IOL(3)
print *, "  animal-missing% : ", IOL(4)
print *, "number of autosomal chromosomes : ", NOAC
print *, "X chromosome No. , Y chromosome No. Unknown chromosome No. ", XCN, YCN, UCN
GDFN=adjustl(GDFN)
GDD=adjustr(GDD)
GDFNIP=GDD//GDFN
GDFNIP=adjustl(GDFNIP)
print *, "input genomic data file is ",GDFNIP
print *, "Output file for Statistics will be created in",ODD
print *, "Output file for Imputation will be created in",IDD
print *, "Parameter file for Reformation will be created in",PDD
ODD=adjustr(ODD)
IDD=adjustr(IDD)
PDD=adjustr(PDD)
close(99)
open(1,file=GDFNIP,status='unknown')
ODFNIP=ODD//"Deleted_Markers"//OFT//".out"
ODFNIP=adjustl(ODFNIP)
open(11,file=ODFNIP,status='unknown')
write(11,fmt='(a50,6a10,2a15)')
Name","MarkerNO","SOSM","FOMGT","FOMIG","FOHE","CSOHW","chromosome","position"
ODFNIP=ODD//"Selected_Markers"//OFT//".out"
ODFNIP=adjustl(ODFNIP)
open(12,file=ODFNIP,status='unknown')
write(12,fmt='(a50,a10,2a15)') "Marker Name","MarkerNO","chromosome","position"
ODFNIP=ODD//"Deleted_Animals"//OFT//".out"
ODFNIP=adjustl(ODFNIP)
open(13,file=ODFNIP,status='unknown')
write(13,fmt='(a20,a10,2a15)') "Animal ID      ","Animal NO.,"NO.Missing","Missing rate"
ODFNIP=ODD//"Selected_Animals"//OFT//".out"
ODFNIP=adjustl(ODFNIP)
open(14,file=ODFNIP,status='unknown')
write(14,fmt='(a20,2a10)') "Animal ID      ","Animal NO.,"Animal NNO."
IDFNIP=IDD//"run_fastPHASE"//OFT//".run"

```

"Marker

```

IDFNIP=adjustl(IDFNIP)
open(10,file=IDFNIP,status='unknown')
write(10,*) "cd "//adjustl(IDD)
do i=1,NOAC
  iz=ichar('0')
  i1=floor(i/10.)+iz
  i2=mod(i,10)+iz
  CHNO(1:1)=achar(i1)
  CHNO(2:2)=achar(i2)
  IDFNIP=IDD//"To_fastPHASE"//CHNO(1:2)//OFT//".inp"
  IDFNIP=adjustl(IDFNIP)
  write(10,*)      "/fastPHASE_Linux      -T10      -S20      -H-4      -o"//CHNO(1:2)//OFT//
./To_fastPHASE"//CHNO(1:2)//OFT//".inp"
  open(i+20,file=IDFNIP,status='unknown')
enddo
ODFNIP=ODD//"Basic_Statistics"//OFT//".out"
ODFNIP=adjustl(ODFNIP)
open(16,file=ODFNIP,status='unknown')
ODFNIP=ODD//"By_Chromosome_S"//OFT//".out"
ODFNIP=adjustl(ODFNIP)
open(17,file=ODFNIP,status='unknown')
write(17,fmt='(a14,19a10)')
NO.", "NOM", "NOSM", "NODM", "NODM_M", "NODM_F", "NODM_H", "NODM_M_H", "NODM_F_H", &
&"NODM_M_F", "NODM_MHF", "NODM_AHO", "NODM_AHE", "AODAMB", "AOMAF", "AOIBC", "AOC SOHWT", "A
OPIC", "AOHE", "AOEH"
ODFNIP=ODD//"By_Chromosome_A"//OFT//".out"
ODFNIP=adjustl(ODFNIP)
open(20,file=ODFNIP,status='unknown')
write(20,fmt='(a14,8a10)')
NO.", "NOM", "AODAMB", "AOMAF", "AOIBC", "AOC SOHWT", "AOPIC", "AOHE", "AOEH"
PDFNIP=PDD//"PIP"//OFT//".par"
PDFNIP=adjustl(PDFNIP)
open(18,file=PDFNIP,status='unknown')
ODFNIP=ODD//"Whole_Results"//OFT//".out"
ODFNIP=adjustl(ODFNIP)
open(19,file=ODFNIP,status='unknown')
write(19,fmt='(a50,a5,a15,3a10,3a5,3a6,3a10,2a5,2a6,6a10)')
Name", "CHno", "position", "SOSM", "FOMGT", "NOMGT", "MAH", "HE", &
"MIH", "NOMAH", "NOHE", "NOMIH", "FOMAH", "FOHE", "FOMIH", "MABA", "MIBA", "NOMAG", "NOMIG", "FOMAG
", "FOMIG", "EH", "CSOHWT", "PIC", "IBC"
NOM=0
do
  read(1,*,iostat=io)
  if(io/=0) exit
  NOM=NOM+1
enddo
rewind(unit=1)
NOM=NOM-1
allocate(IOM(NOM,2),GN(NOM,NOA),GTDFFP(NOAC*2,NOM))
allocate(ID(NOAC),GTS(NOAC),NOMGBA(NOAC),NOSMBC(NOAC))
NOSMBC=0
IOM=0
GN=0
NOMGBA=0

```



```

NODA=0
GTDFFP=' '
read(1,*)D1,D2,D3,ID
MG=GT(11)(1:1)
NODM=0
do i=1,NOM
  GTS=' '
  GF=0.0
  GTF=0.0
  FOMGT=0.0; FOMG=0.0
  read(1,*,iostat=io) MN, IOM(i,:), GTS
  if(io/=0) exit
  BC=0; GC=0; nMI=0; nMA=0; NOMIH=0; NOHE=0; NOMAH=0; NOMG=0
  do j=1,NOA
    if(GTS(j).EQ.GT(11)) GTS(j)="???"
    do k=1,2
      select case (GTS(j)(k:k))
        case("A")
          BC(1)=BC(1)+1
        case("T")
          BC(2)=BC(2)+1
        case("G")
          BC(3)=BC(3)+1
        case("C")
          BC(4)=BC(4)+1
        case("??")
          BC(5)=BC(5)+1
      end select
    enddo
    select case (GTS(j))
      case("AA")
        GC(1)=GC(1)+1
      case("AT")
        GC(2)=GC(2)+1
      case("TA")
        GC(2)=GC(2)+1
      case("AG")
        GC(3)=GC(3)+1
      case("GA")
        GC(3)=GC(3)+1
      case("AC")
        GC(4)=GC(4)+1
      case("CA")
        GC(4)=GC(4)+1
      case("TT")
        GC(5)=GC(5)+1
      case("TG")
        GC(6)=GC(6)+1
      case("GT")
        GC(6)=GC(6)+1
      case("TC")
        GC(7)=GC(7)+1
      case("CT")
        GC(7)=GC(7)+1
    endselect
  enddo
enddo

```

```

case("GG")
  GC(8)=GC(8)+1
case("GC")
  GC(9)=GC(9)+1
case("CG")
  GC(9)=GC(9)+1
case("CC")
  GC(10)=GC(10)+1
case("??")
  GC(11)=GC(11)+1
  TNOMG=TNOMG+1
end select
enddo
GF=real(BC)/real(sum(BC(:4)))
GTF=real(GC)/real(sum(GC(:10)))
FOMG=real(BC(5))/real(NOAA*2)
FOMGT=real(GC(11))/(real(NOAA))
FOMIH=0.0; FOHE=0.0; FOMAH=0.0; FOMIG=0.0; FOMAG=0.0;
MIH='?';MAH='?';HE='?';MIBA='?';MABA='?';SOSM='NORMAL'
do k=1,4
  if(GF(k).NE.0 .AND. GF(k)<0.5) then
    MIBA=BA(k)
    FOMIG=GF(k)
    NOMIG=BC(k)
  elseif(GF(k).NE.0 .AND. GF(k)>0.5) then
    MABA=BA(k)
    FOMAG=GF(k)
    NOMAG=BC(k)
  elseif(GF(k)==0.5) then
    if(sum(GF(:k)).eq.0.5) then
      NOMIG=BC(k)
      MIBA=BA(k)
      FOMIG=GF(k)
    else
      MABA=BA(k)
      FOMAG=GF(k)
      NOMAG=BC(k)
    endif
  endif
enddo
MIH=MIBA//MIBA
MAH=MABA//MABA
HE=MABA//MIBA
if (HE=='TA') then
  HE="AT"
elseif(HE=='GA') then
  HE="AG"
elseif(HE=='CA') then
  HE="AC"
elseif(HE=='GT') then
  HE="TG"
elseif(HE=='CT') then
  HE="TC"
elseif(HE=='CG') then

```

```

HE="GC"
endif
do k=1,10
  if(GT(k)==MIH) then
    FOMIH=GTF(k)
    NOMIH=GC(k)
  elseif(GT(k)==MAH) then
    FOMAH=GTF(k)
    NOMAH=GC(k)
  elseif(GT(k)==HE) then
    FOHE=GTF(k)
    NOHE=GC(k)
  endif
enddo
if(EXCN.ne.XCN.and.EXCN.ne.YCN.and.EXCN.ne.UCN.and.EXCN.ne.999.and.IOM(i,1).ne.EXCN) then
  AOMAFBC_S=SOMAFBC_S/real(NOSMBC(IOM(i-1,1)))
  AOIBCBC_S=SOIBCBC_S/real(NOSMBC(IOM(i-1,1)))
  AOC SOHWTBC_S=SOC SOHWTBC_S/real(NOSMBC(IOM(i-1,1)))
  AOPICBC_S=SOPICBC_S/real(NOSMBC(IOM(i-1,1)))
  AOHEBC_S=SOHEBC_S/real(NOSMBC(IOM(i-1,1)))
  AOEHBC_S=SOEHBC_S/real(NOSMBC(IOM(i-1,1)))
  AODBAMBC_S=SODBAMBC_S/real(NOSMBC(IOM(i-1,1))-1)
  AOMAFBC=SOMAFBC/real(NOMBC)
  AOIBCBC=SOIBCBC/real(NOMBC-NOZBC)
  AOC SOHWTBC=SOC SOHWTBC/real(NOMBC-NOZBC)
  AOPICBC=SOPICBC/real(NOMBC)
  AOHEBC=SOHEBC/real(NOMBC)
  AOEHBC=SOEHBC/real(NOMBC)
  AODBAMBC=SODBAMBC/real(NOMBC-1)
  write(19,fmt='(a1)' " "
  write(17,fmt='(a10,i4,12i10,f10.1,6f10.5)')
"chromosome",EXCN,NOMBC,NOSMBC(IOM(i-1,1)),NODMBC,NODMBOFOMGTBC,NODMBOFOMIGBC,&
NODMBOC SOHWTBC,NODMBO_M_HBC,NODMBO_F_HBC,NODMBO_M_FBC,NODMBO_MHFBC,NODMBO_H
OM,NODMBO_HEM,AODBAMBC_S,AOMAFBC_S,AOIBCBC_S,&
  AOC SOHWTBC_S,AOPICBC_S,AOHEBC_S,AOEHBC_S
  write(20,fmt='(a10,i4,i10,f10.1,6f10.5)')
"chromosome",EXCN,NOMBC,AODBAMBC,AOMAFBC,AOIBCBC,AOC SOHWTBC,AOPICBC,AOHEBC,AOEHBC
NODMBOC SOHWTBC=0;NODMBOFOMGTBC=0;NODMBOFOMIGBC=0;NODMBO_M_HBC=0;NODMBO_F_HB
C=0;NODMBO_M_FBC=0;NODMBO_MHFBC=0;NOZBC=0
  NODMBO_HOM=0;NODMBO_HEM=0
SODBAMBC=0;SOMAFBC=0;SOIBCBC=0;SOC SOHWTBC=0;SOPICBC=0;SOHEBC=0;SOEHBC=0;NOMBC=0;NO
DMBC=0
SODBAMBC_S=0;SOMAFBC_S=0;SOIBCBC_S=0;SOC SOHWTBC_S=0;SOPICBC_S=0;SOHEBC_S=0;SOEHBC_S=
0
end if
ENOMIH=(FOMIG**2)*real((NOMIH+NOHE+NOMAH))
ENOHE=(2*FOMIG*FOMAG)*real((NOMIH+NOHE+NOMAH))
ENOMAH=(FOMAG**2)*real((NOMIH+NOHE+NOMAH))
CSOHW=(real(NOMIH)-ENOMIH)**2/ENOMIH+((real(NOHE)-ENOHE)**2/ENOHE)+((real(NOMAH)-ENOM

```

```

AH)**2/ENOMAH)
  EH=1-(FOMIG**2+FOMAG**2)
  PIC=1-(FOMIG**2+FOMAG**2)-2*((FOMIG**2)*(FOMAG**2))
  IBC=1-(real(NOHE)/ENOHE)
  if(ENOMIH.eq.0.or.ENOMAH.eq.0.or.ENOHE.eq.0) then;
NOZ=NOZ+1;NOZBC=NOZBC+1;CSOHWT=0;IBC=0;endif;
  if(CSOHWT>IOL(3)) then;
NODMBOCSOHWT=NODMBOCSOHWT+1;SOSM="OL_HWT";NODMBOCSOHWTBC=NODMBOCSOHWTBC+1
;endif;
  if(FOMIG<IOL(2)) then;
NODMBOFOMIG=NODMBOFOMIG+1;SOSM="OL_MAF";NODMBOFOMIGBC=NODMBOFOMIGBC+1;endif;
  if(FOMIG<IOL(2).and.CSOHWT>IOL(3)) then;
NODMBO_F_H=NODMBO_F_H+1;SOSM="OL_F_H";NODMBO_F_HBC=NODMBO_F_HBC+1;endif;
  if(FOMGT>IOL(1)) then;
NODMBOFOMGT=NODMBOFOMGT+1;SOSM="OL_MGR";NODMBOFOMGTBC=NODMBOFOMGTBC+1;endif;
  if(FOMGT>IOL(1).and.CSOHWT>IOL(3)) then;
NODMBO_M_H=NODMBO_M_H+1;SOSM="OL_M_H";NODMBO_M_HBC=NODMBO_M_HBC+1;endif;
  if(FOMGT>IOL(1).and.FOMIG<IOL(2)) then;
NODMBO_M_F=NODMBO_M_F+1;SOSM="OL_M_F";NODMBO_M_FBC=NODMBO_M_FBC+1;endif;
  if(FOMGT>IOL(1).and.FOMIG<IOL(2).and.CSOHWT>IOL(3)) then;
NODMBO_MHF=NODMBO_MHF+1;SOSM="OL_MHF";NODMBO_MHFBC=NODMBO_MHFBC+1;endif;
  if(FOMGT.eq.1) then; NOAMM=NOAMM+1;SOSM="OL_AMG";NOAMMBC=NOAMMBC+1;endif;
  if(FOMAH.eq.1) then;
NODMBOAH=NODMBOAH+1;SOSM="OL_AHO";NODMBOAHBC=NODMBOAHBC+1;endif;
  if(FOHE.eq.1) then;
NODMBOAHE=NODMBOAHE+1;SOSM="OL_AHE";NODMBOAHEBC=NODMBOAHEBC+1;endif;
  if(FOMAH.eq.1 .and. FOMGT>IOL(1)) then;
NODMBO_HOM=NODMBO_HOM+1;SOSM="OL_HOM";NODMBO_HOMBC=NODMBO_HOMBC+1;endif;
  if(FOHE.eq.1 .and. FOMGT>IOL(1)) then;
NODMBO_HEM=NODMBO_HEM+1;SOSM="OL_HEM";NODMBO_HEMBC=NODMBOA_HEMBC+1;endif;
  if(FOMGT.eq.0) NONMM=NONMM+1
  if(IOM(i,1).ne.XCN.and.IOM(i,1).ne.YCN.and.IOM(i,1).ne.UCN) then
    if(CSOHWT>IOL(3)) ACNODMBOCSOHWT=ACNODMBOCSOHWT+1
    if(FOMIG<IOL(2)) ACNODMBOFOMIG=ACNODMBOFOMIG+1
    if(FOMGT>IOL(1)) ACNODMBOFOMGT=ACNODMBOFOMGT+1
    if(FOMIG<IOL(2).and.CSOHWT>IOL(3)) ACNODMBO_F_H=ACNODMBO_F_H+1
    if(FOMGT>IOL(1).and.CSOHWT>IOL(3)) ACNODMBO_M_H=ACNODMBO_M_H+1
    if(FOMGT>IOL(1).and.FOMIG<IOL(2)) ACNODMBO_M_F=ACNODMBO_M_F+1
    if(FOMGT>IOL(1).and.FOMIG<IOL(2).and.CSOHWT>IOL(3)) ACNODMBO_MHF=ACNODMBO_MHF+1
    if(FOMGT.eq.1) ACNOAMM=ACNOAMM+1
    if(FOMGT.eq.0) ACNONMM=ACNONMM+1
    if(FOMAH.eq.1) ACNODMBOAH=ACNODMBOAH+1
    if(FOHE.eq.1) ACNODMBOAHE=ACNODMBOAHE+1
    if(FOMAH.eq.1.and.FOMGT>IOL(1)) ACNODMBO_HOM=ACNODMBO_HOM+1
    if(FOHE.eq.1.and.FOMGT>IOL(1)) ACNODMBO_HEM=ACNODMBO_HEM+1
  elseif(IOM(i,1).eq.UCN) then
    if(CSOHWT>IOL(3)) UCNODMBOCSOHWT=UCNODMBOCSOHWT+1
    if(FOMIG<IOL(2)) UCNODMBOFOMIG=UCNODMBOFOMIG+1
    if(FOMGT>IOL(1)) UCNODMBOFOMGT=UCNODMBOFOMGT+1
    if(FOMIG<IOL(2).and.CSOHWT>IOL(3)) UCNODMBO_F_H=UCNODMBO_F_H+1
    if(FOMGT>IOL(1).and.CSOHWT>IOL(3)) UCNODMBO_M_H=UCNODMBO_M_H+1
    if(FOMGT>IOL(1).and.FOMIG<IOL(2)) UCNODMBO_M_F=UCNODMBO_M_F+1
    if(FOMGT>IOL(1).and.FOMIG<IOL(2).and.CSOHWT>IOL(3)) UCNODMBO_MHF=UCNODMBO_MHF+1
    if(FOMGT.eq.1) UCNOAMM=UCNOAMM+1

```

```

if(FOMGT.eq.0) UCNONMM=UCNONMM+1
if(FOMAH.eq.1) UCNODMBOAH=UCNODMBOAH+1
if(FOHE.eq.1) UCNODMBOAHE=UCNODMBOAHE+1
if(FOMAH.eq.1.and.FOMGT>IOL(1)) UCNODMBO_HOM=UCNODMBO_HOM+1
if(FOHE.eq.1.and.FOMGT>IOL(1)) UCNODMBO_HEM=UCNODMBO_HEM+1
elseif(IOM(i,1).eq.XCN.or.IOM(i,1).eq.YCN) then
if(CSOHWT>IOL(3)) SCNODMBOCSOHWT=SCNODMBOCSOHWT+1
if(FOMIG<IOL(2)) SCNODMBOFOMIG=SCNODMBOFOMIG+1
if(FOMGT>IOL(1)) SCNODMBOFOMGT=SCNODMBOFOMGT+1
if(FOMIG<IOL(2).and.CSOHWT>IOL(3)) SCNODMBO_F_H=SCNODMBO_F_H+1
if(FOMGT>IOL(1).and.CSOHWT>IOL(3)) SCNODMBO_M_H=SCNODMBO_M_H+1
if(FOMGT>IOL(1).and.FOMIG<IOL(2)) SCNODMBO_M_F=SCNODMBO_M_F+1
if(FOMGT>IOL(1).and.FOMIG<IOL(2).and.CSOHWT>IOL(3)) SCNODMBO_MHF=SCNODMBO_MHF+1
if(FOMGT.eq.1) SCNOAMM=SCNOAMM+1
if(FOMGT.eq.0) SCNONMM=SCNONMM+1
if(FOMAH.eq.1) SCNODMBOAH=SCNODMBOAH+1
if(FOHE.eq.1) SCNODMBOAHE=SCNODMBOAHE+1
if(FOMAH.eq.1.and.FOMGT>IOL(1)) SCNODMBO_HOM=SCNODMBO_HOM+1
if(FOHE.eq.1.and.FOMGT>IOL(1)) SCNODMBO_HEM=SCNODMBO_HEM+1
endif
if(IOM(i,1) .eq. XCN) then: NOMOXC=NOMOXC+1;SOSM="OL_XCH";endif;
if(IOM(i,1) .eq. YCN) then: NOMOYC=NOMOYC+1;SOSM="OL_YCH";endif;
if(IOM(i,1) .eq. UCN) then: NOMOUC=NOMOUC+1;SOSM="OL_UCH";endif;
write(19,fmt='(a50,i5,i15,a10,f10.6,i10,3a5,3i6,3f10.6,2a5,2i6,3f10.6,f10.3,2f10.6)')
MN,IOM(i,1),IOM(i,2),SOSM,FOMGT,GC(11),&

```

&MAH,HE,MIH,NOMAH,NOHE,NOMIH,FOMAH,FOHE,FOMIH,MABA,MIBA,NOMAG,NOMIG,FOMAG,FOMIG,E
H,CSOHWT,PIC,IBC

```

if(IOM(i,1).ne.DXCN.and.IOM(i,1).ne.DYCN.and.IOM(i,1).ne.DUCN) then
NOMOAC=NOMOAC+1
NOMBC=NOMBC+1
if(FOMIG .LT. 0.1) then
NOTMWMAF1=NOTMWMAF1+1
elseif(FOMIG .GE. 0.1 .and. FOMIG .LT. 0.2) then
NOTMWMAF2=NOTMWMAF2+1
elseif(FOMIG .GE. 0.2 .and. FOMIG .LT. 0.3) then
NOTMWMAF3=NOTMWMAF3+1
elseif(FOMIG .GE. 0.3 .and. FOMIG .LT. 0.4) then
NOTMWMAF4=NOTMWMAF4+1
elseif(FOMIG .GE. 0.4 ) then
NOTMWMAF5=NOTMWMAF5+1
endif
CMPBC=IOM(i,2)
if(NOMBC.ne.1) then
DBAMBC=real(CMPBC-EXMPBC)/1000.0
SODBAMBC=SODBAMBC+DBAMBC
SOTDBAM=SOTDBAM+DBAMBC
if(DBAMBC .LT. MDG(1)) then
NODBAM1=NODBAM1+1
elseif(DBAMBC .GE. MDG(1) .and. DBAMBC .LT. MDG(2)) then
NODBAM2=NODBAM2+1
elseif(DBAMBC .GE. MDG(2) .and. DBAMBC .LT. MDG(3)) then
NODBAM3=NODBAM3+1
elseif(DBAMBC .GE. MDG(3) .and. DBAMBC .LT. MDG(4)) then

```

```

        NODBAM4=NODBAM4+1
        elseif(DBAMBC .GE. MDG(4) ) then
            NODBAM5=NODBAM5+1
        endif
    endif
    EXMPBC=COMPBC

SOMAFBC=SOMAFBC+FOMIG;SOIBCBC=SOIBCBC+IBC;SOCSOHWTBC=SOC SOHWTBC+CSOHWT;SOPICBC
=SOPICBC+PIC;SOHEBC=SOHEBC+FOHE;SOEHBC=SOEHBC+EH

SOTMAF=SOTMAF+FOMIG;SOTIBC=SOTIBC+IBC;SOTCSOHWT=SOTCSOHWT+CSOHWT;SOTPIC=SOTPIC
+PIC;SOTHE=SOTHE+FOHE;SOTEH=SOTEH+EH
    endif

if(FOMGT>IOL(1).or.FOMIG<IOL(2).or.CSOHWT>IOL(3).or.FOHE.EQ.1.or.IOM(i,1).eq.DXCN.or.IOM(i,1).eq.DYC
N&
    &.or.IOM(i,1).eq.DUCN) then
        NODM=NODM+1
        NODMBC=NODMBC+1
        write(11,fmt='(a50,i10,a10,3f10.7,f10.3,2i15)') MN,i,SOSM,FOMGT,FOMIG,FOHE,CSOHWT,IOM(i,1),IOM(i,2)
    else

SOMAFBC_S=SOMAFBC_S+FOMIG;SOIBCBC_S=SOIBCBC_S+IBC;SOC SOHWTBC_S=SOC SOHWTBC_S+CSO
HWT;
        SOPICBC_S=SOPICBC_S+PIC;SOHEBC_S=SOHEBC_S+FOHE;SOEHBC_S=SOEHBC_S+EH

SOTMAF_S=SOTMAF_S+FOMIG;SOTIBC_S=SOTIBC_S+IBC;SOTCSOHWT_S=SOTCSOHWT_S+CSOHWT;
SOTPIC_S=SOTPIC_S+PIC;SOTHE_S=SOTHE_S+FOHE;SOTEH_S=SOTEH_S+EH
        NOSMBC(IOM(i,1))=NOSMBC(IOM(i,1))+1
        CMPBC_S=IOM(i,2)
        if(NOSMBC(IOM(i,1)).ne.1) then;
            DBAMBC_S=real(CMPBC_S-EXMPBC_S)/1000.0
            SODBAMBC_S=SODBAMBC_S+DBAMBC_S
            SOTDBAM_S=SOTDBAM_S+DBAMBC_S
            if(DBAMBC_S .LT. MDG(1)) then
                NODBASM1=NODBASM1+1
            elseif(DBAMBC_S .GE. MDG(1) .and. DBAMBC_S .LT. MDG(2)) then
                NODBASM2=NODBASM2+1
            elseif(DBAMBC_S .GE. MDG(2) .and. DBAMBC_S .LT. MDG(3)) then
                NODBASM3=NODBASM3+1
            elseif(DBAMBC_S .GE. MDG(3) .and. DBAMBC_S .LT. MDG(4)) then
                NODBASM4=NODBASM4+1
            elseif(DBAMBC_S .GE. MDG(4) ) then
                NODBASM5=NODBASM5+1
            endif
        endif
        EXMPBC_S=COMPBC_S
        p=p+1
        write(12,fmt='(a50,i10,2i15)') MN,i,IOM(i,1),IOM(i,2)
        do l=1,NOA
            GTDFFP(l*2-1,p)=GTS(l)(1:1)
            GTDFFP(l*2,p)=GTS(l)(2:2)
            if(GTS(l).eq.'??') NOMGBA(l)=NOMGBA(l)+1
        enddo
    enddo

```

```

if(FOMIG .LT. 0.1) then
  NOSMWMAF1=NOSMWMAF1+1
elseif(FOMIG .GE. 0.1 .and. FOMIG .LT. 0.2) then
  NOSMWMAF2=NOSMWMAF2+1
elseif(FOMIG .GE. 0.2 .and. FOMIG .LT. 0.3) then
  NOSMWMAF3=NOSMWMAF3+1
elseif(FOMIG .GE. 0.3 .and. FOMIG .LT. 0.4) then
  NOSMWMAF4=NOSMWMAF4+1
elseif(FOMIG .GE. 0.4 ) then
  NOSMWMAF5=NOSMWMAF5+1
endif
endif
EXCN=IOM(i,1)
enddo
AOTMAF_S=SOTMAF_S/real(NOM-NODM)
AOTIBC_S=SOTIBC_S/real(NOM-NODM)
AOTCSOHWT_S=SOTCSOHWT_S/real(NOM-NODM)
AOTPIC_S=SOTPIC_S/real(NOM-NODM)
AOTHE_S=SOTHE_S/real(NOM-NODM)
AOTEH_S=SOTEH_S/real(NOM-NODM)
AOTDBAM_S=SOTDBAM_S/real(NOM-NODM-NOAC)
AOTMAF=SOTMAF/real(NOMOAC)
AOTIBC=SOTIBC/real(NOMOAC-NOZ)
AOTCSOHWT=SOTCSOHWT/real(NOMOAC-NOZ)
AOTPIC=SOTPIC/real(NOMOAC)
AOTHE=SOTHE/real(NOMOAC)
AOTEH=SOTEH/real(NOMOAC)
AOTDBAM=SOTDBAM/real(NOMOAC-NOAC)
write(17,fmt='(a14,i10,f10.1,6f10.5)')
",NOMOAC,NOM-NODM,NOMOAC-(NOM-NODM),ACNODMBOFOMGT,ACNODMBOFOMIG,&
ACNODMBOCSOHWT,ACNODMBO_M_H,ACNODMBO_F_H,ACNODMBO_M_F,ACNODMBO_MHF,ACNODMBO
O_HOM,ACNODMBO_HEM,AOTDBAM_S,AOTMAF_S,&
&AOTIBC_S,AOTCSOHWT_S,AOTPIC_S,AOTHE_S,AOTEH_S
write(20,fmt='(a14,i10,f10.1,6f10.5)')
",NOMOAC,AOTDBAM,AOTMAF,AOTIBC,AOTCSOHWT,AOTPIC,AOTHE,AOTEH
print *, "Done with the outlier marker elimination : ",NODM,"markers were deleted"
do m=1,NOA
  if((real(NOMGBA(m))/real((NOM-NODM))).GT.IOL(4)) then
    NODA=NODA+1
    write(13,fmt='(a20,i10,i15,f15.7)') ID(m),m,NOMGBA(m),real(NOMGBA(m))/real((NOM-NODM))
  else
    NANO=NANO+1
    write(14,fmt='(a20,i10)') ID(m),m,NANO
  endif
enddo
if(NOANODA.lt.10) then
  write(MWF(3),fmt='(a4)') '(a1)'
elseif(NOANODA.ge.10.and.NOANODA.lt.100) then
  write(MWF(3),fmt='(a4)') '(a2)'
elseif(NOANODA.ge.100.and.NOANODA.lt.1000) then
  write(MWF(3),fmt='(a4)') '(a3)'
elseif(NOANODA.ge.1000.and.NOANODA.lt.10000) then
  write(MWF(3),fmt='(a4)') '(a4)'
elseif(NOANODA.ge.10000.and.NOANODA.lt.100000) then

```

```

write(MWF(3),fmt='(a4)' '(a5)'
endif
write(MWF(1),fmt='(i10)' NOA-NODA
MWF(1)=adjustl(MWF(1))
do o=1,NOAC
write(o+20,fmt=MWF(3)) MWF(1)
if(NOSMBC(o).lt.10) then
write(MWF(3),fmt='(a4)' '(a1)'
elseif(NOSMBC(o).ge.10.and.NOSMBC(o).lt.100) then
write(MWF(3),fmt='(a4)' '(a2)'
elseif(NOSMBC(o).ge.100.and.NOSMBC(o).lt.1000) then
write(MWF(3),fmt='(a4)' '(a3)'
elseif(NOSMBC(o).ge.1000.and.NOSMBC(o).lt.10000) then
write(MWF(3),fmt='(a4)' '(a4)'
elseif(NOSMBC(o).ge.10000.and.NOSMBC(o).lt.100000) then
write(MWF(3),fmt='(a4)' '(a5)'
elseif(NOSMBC(o).ge.100000.and.NOSMBC(o).lt.1000000) then
write(MWF(3),fmt='(a4)' '(a6)'
endif
write(MWF(2),fmt='(i10)' NOSMBC(o)
MWF(2)=adjustl(MWF(2))
write(o+20,fmt=MWF(3)) MWF(2)
write(MWF(5),fmt='(a1,i10,a3)' '( ',NOSMBC(o),'a1)'
MWF(5)=trim(MWF(5))
MWF(5)=adjustl(MWF(5))
print *, "Number of markers on chromosome No.",o,NOSMBC(o),MWF(5)
ANO=0
do m=1,NOA
if((real(NOMGBA(m))/real((NOM-NODM))).LE.IOL(4)) then
ANO=ANO+1
write(MWF(4),fmt='(i10)' ANO
MWF(4)=adjustl(MWF(4))
if(ANO.lt.10) then
write(MWF(3),fmt='(a10)' '(a4,1x,a1)'
elseif(ANO.ge.10.and.ANO.lt.100) then
write(MWF(3),fmt='(a10)' '(a4,1x,a2)'
elseif(ANO.ge.100.and.ANO.lt.1000) then
write(MWF(3),fmt='(a10)' '(a4,1x,a3)'
elseif(ANO.ge.1000.and.ANO.lt.10000) then
write(MWF(3),fmt='(a10)' '(a4,1x,a4)'
elseif(ANO.ge.10000.and.ANO.lt.100000) then
write(MWF(3),fmt='(a10)' '(a4,1x,a5)'
endif
write(o+20,fmt=MWF(3)) "# id",MWF(4)
write(o+20,fmt=MWF(5)) GTDFFP(m*2-1,SMNBC:(SMNBC+NOSMBC(o)-1))
write(o+20,fmt=MWF(5)) GTDFFP(m*2,SMNBC:(SMNBC+NOSMBC(o)-1))
endif
enddo
SMNBC=SMNBC+NOSMBC(o)
close(o+20)
enddo
print *, "Done with the outlier animal elimination : ",NODA,"animals were deleted"
write(16,fmt='(a120)' "*****Raw data & parameter description*****"
write(16,fmt='(a120,i15)' "number of animals", NOA

```



```

write(16,fmt='(a120,a230)') "directory of genomic data file", GDD
write(16,fmt='(a120,a230)') "name of genomic data file", GDFN
write(16,fmt='(a120,a230)') "directory of output file", ODD
write(16,fmt='(a120,a15)') "missing genotype", GT(11)
write(16,fmt='(a120,i15)') "range of outliers"
write(16,fmt='(a120,f15.3)') " marker-missing% :", IOL(1)
write(16,fmt='(a120,f15.3)') " minor allele frequency :", IOL(2)
write(16,fmt='(a120,f15.3)') " HW-chisquare(df=1) :", IOL(3)
write(16,fmt='(a120,f15.3)') " animal-missing% :", IOL(4)
write(16,fmt='(a120,i15)') "number of autosomal chromosomes :", NOAC
write(16,fmt='(a120,3i5)') "X chromosome No. , Y chromosome No. Unknown chromosome No. ", XCN,
YCN, UCN
write(16,fmt='(a120)') " "
write(16,fmt='(a120)') "*****Basic statistics*****"
write(16,fmt='(a120,i15)') "Number of markers on chromosome X",NOMOXC
write(16,fmt='(a120,i15)') "Number of markers on chromosome Y",NOMOYC
write(16,fmt='(a120,i15)') "Number of markers with Unknown marker location(position)",NOMOUC
write(16,fmt='(a120,i15)') "Total number of markers", NOM
write(16,fmt='(a120,i15)') "Number of markers on atuosomal chromosomes", NOMOAC
write(16,fmt='(a120,i15)') "Number of selected(useful) markers", NOM-NODM
write(16,fmt='(a120,i15)') "Number of outlier markers", NODM
write(16,fmt='(a109,f10.2,a,i15)') "Number of animals with missing over ",IOL(4),"%", NODA
TMGR=real(TNOMG)/real(NOM*NOA)
write(16,fmt='(a120,f15.10)') "Total missing rate(Number of Missing genotypes/Number of
genotypes(NO.animal*NO.markers))" , TMGR
write(16,fmt='(a120)') " "
write(16,fmt='(a120)') "*** total ***"
write(16,fmt='(a120,i15)') "Number of all missing markers", NOAMM
write(16,fmt='(a120,i15)') "Number of all homo_genotype markers", NODMBOAH
write(16,fmt='(a120,i15)') "Number of all hetero_genotype markers", NODMBOAHE
write(16,fmt='(a109,f10.2,a,i15)') "Number of all homo_genotype markers and missing over ",
IOL(1)*100,"%", NODMBO_HOM
write(16,fmt='(a109,f10.2,a,i15)') "Number of all hetero_genotype markers and missing over ",
IOL(1)*100,"%", NODMBO_HEM
write(16,fmt='(a120,i15)') "Number of markers with no missing", NONMM
write(16,fmt='(a109,f10.2,a,i15)') "Number of markers with missing over",IOL(1)*100,"%",
NODMBOFOMGT
write(16,fmt='(a110,f10.2,i15)') "Number of markers with minor allele frequency less than",IOL(2),
NODMBOFOMIG
write(16,fmt='(a110,f10.2,i15)') "Number of markers with HWchisquare over",IOL(3), NODMBOCSOHW
write(16,fmt='(a78,f10.2,a22,f10.2,i15)') "Number of markers with missing over",IOL(1)*100,"% and
HWchisquare over",IOL(3),&
NODMBO_M_H
write(16,fmt='(a79,f10.2,a21,f10.2,i15)') "Number of markers with minor allele frequency less than",IOL(2),&
" and HWchisquare over",IOL(3),NODMBO_F_H
write(16,fmt='(a62,f10.2,a38,f10.2,i15)') "Number of markers with missing over ",IOL(1)*100,&
"% and minor allele frequency less than" ,IOL(2),NODMBO_M_F
write(16,fmt='(a35,f9.2,a35,f10.2,a21,f10.2,i15)') "Number of markers with missing over",IOL(1)*100,&
"% , minor allele frequency less than" ,IOL(2)," and HWchisquare over",IOL(3), NODMBO_MHF
write(16,fmt='(a120)') " "
write(16,fmt='(a120)') "*** on autosomal chromosome ***"
write(16,fmt='(a120,i15)') "Number of all missing markers", ACNOAMM
write(16,fmt='(a120,i15)') "Number of all homo_genotype markers", ACNODMBOAH
write(16,fmt='(a120,i15)') "Number of all hetero_genotype markers", ACNODMBOAHE

```

```

write(16,fmt='(a109,f10.2,a1,i15)') "Number of all homo_genotype markers and missing over ",
IOL(1)*100,"%", ACNODMBO_HOM
write(16,fmt='(a109,f10.2,a1,i15)') "Number of all hetero_genotype markers and missing over ",
IOL(1)*100,"%", ACNODMBO_HEM
write(16,fmt='(a120,i15)') "Number of markers with no missing", ACNONMM
write(16,fmt='(a109,f10.2,a1,i15)') "Number of markers with missing over",IOL(1)*100,"%",
ACNODMBOFOMGT
write(16,fmt='(a110,f10.2,i15)') "Number of markers with minor allele frequency less than",IOL(2),
ACNODMBOFOMIG
write(16,fmt='(a110,f10.2,i15)') "Number of markers with HWchisquare over",IOL(3), ACNODMBOCISOHWT
write(16,fmt='(a78,f10.2,a22,f10.2,i15)') "Number of markers with missing over",IOL(1)*100,"% and
HWchisquare over",IOL(3),&
ACNODMBO_M_H
write(16,fmt='(a79,f10.2,a21,f10.2,i15)') "Number of markers with minor allele frequency less than",IOL(2),&
" and HWchisquare over",IOL(3),ACNODMBO_F_H
write(16,fmt='(a62,f10.2,a38,f10.2,i15)') "Number of markers with missing over ",IOL(1)*100,&
"% and minor allele frequency less than" ,IOL(2),ACNODMBO_M_F
write(16,fmt='(a35,f9.2,a35,f10.2,a21,f10.2,i15)') "Number of markers with missing over",IOL(1)*100,&
"% , minor allele frequency less than" ,IOL(2)," and HWchisquare over",IOL(3), ACNODMBO_MHF
write(16,fmt='(a120)') " "
write(16,fmt='(a120)') "*** on unknown chromosome ***"
write(16,fmt='(a120,i15)') "Number of all missing markers", UCNOAMM
write(16,fmt='(a120,i15)') "Number of all homo_genotype markers", UCNODMBOAH
write(16,fmt='(a120,i15)') "Number of all hetero_genotype markers", UCNODMBOAHE
write(16,fmt='(a109,f10.2,a1,i15)') "Number of all homo_genotype markers and missing over ",
IOL(1)*100,"%", UCNODMBO_HOM
write(16,fmt='(a109,f10.2,a1,i15)') "Number of all hetero_genotype markers and missing over ",
IOL(1)*100,"%", UCNODMBO_HEM
write(16,fmt='(a120,i15)') "Number of markers with no missing", UCNONMM
write(16,fmt='(a109,f10.2,a1,i15)') "Number of markers with missing over",IOL(1)*100,"%",
UCNODMBOFOMGT
write(16,fmt='(a110,f10.2,i15)') "Number of markers with minor allele frequency less than",IOL(2),
UCNODMBOFOMIG
write(16,fmt='(a110,f10.2,i15)') "Number of markers with HWchisquare over",IOL(3), UCNODMBOCISOHWT
write(16,fmt='(a78,f10.2,a22,f10.2,i15)') "Number of markers with missing over",IOL(1)*100,"% and
HWchisquare over",IOL(3),&
UCNODMBO_M_H
write(16,fmt='(a79,f10.2,a21,f10.2,i15)') "Number of markers with minor allele frequency less than",IOL(2),&
" and HWchisquare over",IOL(3),UCNODMBO_F_H
write(16,fmt='(a62,f10.2,a38,f10.2,i15)') "Number of markers with missing over ",IOL(1)*100,&
"% and minor allele frequency less than" ,IOL(2),UCNODMBO_M_F
write(16,fmt='(a35,f9.2,a35,f10.2,a21,f10.2,i15)') "Number of markers with missing over",IOL(1)*100,&
"% , minor allele frequency less than" ,IOL(2)," and HWchisquare over",IOL(3), UCNODMBO_MHF
write(16,fmt='(a120)') " "
write(16,fmt='(a120)') "*** on sex chromosome ***"
write(16,fmt='(a120,i15)') "Number of all missing markers", SCNOAMM
write(16,fmt='(a120,i15)') "Number of all homo_genotype markers", SCNODMBOAH
write(16,fmt='(a120,i15)') "Number of all hetero_genotype markers", SCNODMBOAHE
write(16,fmt='(a109,f10.2,a1,i15)') "Number of all homo_genotype markers and missing over ",
IOL(1)*100,"%", SCNODMBO_HOM
write(16,fmt='(a109,f10.2,a1,i15)') "Number of all hetero_genotype markers and missing over ",
IOL(1)*100,"%", SCNODMBO_HEM
write(16,fmt='(a120,i15)') "Number of markers with no missing", SCNONMM
write(16,fmt='(a109,f10.2,a1,i15)') "Number of markers with missing over",IOL(1)*100,"%",

```

SCNODMBOFOMGT

write(16,fmt='(a110,f10.2,i15)') "Number of markers with minor allele frequency less than",IOL(2),
SCNODMBOFOMIG

write(16,fmt='(a110,f10.2,i15)') "Number of markers with HWchisquare over",IOL(3), SCNODMBOCSOHW
write(16,fmt='(a78,f10.2,a22,f10.2,i15)') "Number of markers with missing over",IOL(1)*100,"% and
HWchisquare over",IOL(3),&

SCNODMBO_M_H

write(16,fmt='(a79,f10.2,a21,f10.2,i15)') "Number of markers with minor allele frequency less than",IOL(2),&
" and HWchisquare over",IOL(3),SCNODMBO_F_H

write(16,fmt='(a62,f10.2,a38,f10.2,i15)') "Number of markers with missing over ",IOL(1)*100,&
"% and minor allele frequency less than" ,IOL(2),SCNODMBO_M_F

write(16,fmt='(a35,f9.2,a35,f10.2,a21,f10.2,i15)') "Number of markers with missing over",IOL(1)*100,&
"% , minor allele frequency less than" ,IOL(2)," and HWchisquare over",IOL(3), SCNODMBO_MHF

write(16,fmt='(a120)') " "

write(16,fmt='(a120)') "***** frequency of MAF *****"

write(16,fmt='(a120)') "number of total markers with Minor allele frequency"

write(16,fmt='(a120,i15)') " minor allele frequency < 0.1 : ", NOTMWMAF1

write(16,fmt='(a120,i15)') "0.1 <= minor allele frequency < 0.2 : ", NOTMWMAF2

write(16,fmt='(a120,i15)') "0.2 <= minor allele frequency < 0.4 : ", NOTMWMAF3

write(16,fmt='(a120,i15)') "0.3 <= minor allele frequency < 0.4 : ", NOTMWMAF4

write(16,fmt='(a120,i15)') "0.4 <= minor allele frequency : ", NOTMWMAF5

write(16,fmt='(a120)') " "

write(16,fmt='(a120)') "number of selected markers with Minor allele frequency(except for unusual data)"

write(16,fmt='(a120,i15)') " minor allele frequency < 0.1 : ", NOSMWMAF1

write(16,fmt='(a120,i15)') "0.1 <= minor allele frequency < 0.2 : ", NOSMWMAF2

write(16,fmt='(a120,i15)') "0.2 <= minor allele frequency < 0.4 : ", NOSMWMAF3

write(16,fmt='(a120,i15)') "0.3 <= minor allele frequency < 0.4 : ", NOSMWMAF4

write(16,fmt='(a120,i15)') "0.4 <= minor allele frequency : ", NOSMWMAF5

write(16,fmt='(a120)') " "

write(16,fmt='(a120)') "***** frequency of distance *****"

write(16,fmt='(a120)') "number of total markers with distance between adjacent markers"

write(16,fmt='(50x,a55,2i15)') " distance between adjacent markers < ",MDG(1), NODBAM1

write(16,fmt='(50x,i15,a40,2i15)') MDG(1),"<= distance between adjacent markers < ",MDG(2), NODBAM2

write(16,fmt='(50x,i15,a40,2i15)') MDG(2),"<= distance between adjacent markers < ",MDG(3), NODBAM3

write(16,fmt='(50x,i15,a40,2i15)') MDG(3),"<= distance between adjacent markers < ",MDG(4), NODBAM4

write(16,fmt='(50x,i15,a55,i15)') MDG(4),"<= distance between adjacent markers "

NODBAM5

write(16,fmt='(a120)') " "

write(16,fmt='(a120)') "number of selected markers with distance between adjacent markers(except for
unusual data)"

write(16,fmt='(50x,a55,2i15)') " distance between adjacent markers < ",MDG(1), NODBASM1

write(16,fmt='(50x,i15,a40,2i15)') MDG(1),"<= distance between adjacent markers < ",MDG(2), NODBASM2

write(16,fmt='(50x,i15,a40,2i15)') MDG(2),"<= distance between adjacent markers < ",MDG(3), NODBASM3

write(16,fmt='(50x,i15,a40,2i15)') MDG(3),"<= distance between adjacent markers < ",MDG(4), NODBASM4

write(16,fmt='(50x,i15,a55,i15)') MDG(4),"<= distance between adjacent markers "

NODBASM5

write(18,fmt='(a10)') ""//OFT//""

write(18,*) ANO

IDD=""//adjustl(IDD)

write(18,*) IDD//""

ODD=""//adjustl(ODD)

write(18,*) ODD//""

write(18,fmt='(a18)') ""//Selected_Markers//""

PDD=""//adjustl(PDD)

```
write(18,*) PDD//""
write(18,fmt='(4f10.5)') IOL(:)
write(18,fmt='(4i15)') MDG(:)
write(18,*) NOAC
do r=1,NOAC
  write(18,*) NOSMBC(r)
enddo
print*, "Done with Outlier Elimination"
print*, "Run fastPHASE ex) ./fastPHASE_Linux -T10 -S20 -H-4 -o01 ./To_fastPHASE01//OFT//".inp"
print*, "Parameter file for 3_Post_Imputation is ",PDD
print*, "Parameter file name for 3_Post_Imputation is ",PIP//OFT//".par"
end program OEGRDPF
```


유전자형전환 및 통계

결측 자료 대치 이후의 통계 산출 및 각 표지인자별 유전자형의 변환 (0, 1, 2 또는 -1, 0, 1)

유전자 빈도 0, 1, 2 변환 유전자형 유전자형 빈도

The screenshots show the following data:

- GT_mach1:** Shows a list of markers with two alleles (0, 1) and their corresponding genotype values.
- GT_maching_012_BNH1476.04:** Shows a list of markers with three alleles (0, 1, 2) and their corresponding genotype values.
- GT_mach:** Shows a list of markers with two alleles (0, 1) and their corresponding genotype values.
- G_maching_01_BNH1476.04 & Genefre:** Shows a list of markers with three alleles (0, 1, 2) and their corresponding genotype values.

수량화된 유전자형을 이후에 다시 염기 조합 형태로 전환할 수 있도록 이를 출력하고 유전자 빈도 및 유전자형 빈도를 출력한다.(유전체 관계행렬 구성에 사용됨)

유전자형전환 및 통계

결측 자료 대치 이후의 통계 산출 및 각 표지인자별 유전자형의 변환 (0, 1, 2 또는 -1, 0, 1)

MARKER NO.	MAF	IBC	CSOHW	PIC	O_HE	E_HE	DBAM
1	0.3421596	-0.0587406	5.0169721	0.3488450	0.4766162	0.4501728	0.0000000
2	0.0687758	0.0227917	0.7552990	0.1198876	0.1251719	0.1280913	132.8419952
3	0.0144429	0.0819792	9.7717371	0.0280634	0.0261348	0.0284686	125.3079987
4	0.2517194	0.0396927	2.2980025	0.3057570	0.3617606	0.3767135	77.8300018
5	0.0577716	-0.0107751	0.1688138	0.1029421	0.1100413	0.1088682	45.3260002
6	0.0646492	0.0787392	9.0145626	0.1136262	0.1114168	0.1209394	54.9360008
7	0.1464924	-0.0066166	0.0636563	0.2187985	0.2517194	0.2508647	274.1539917
8	0.3618729	-0.0285100	1.1818371	0.3549542	0.4745530	0.4613985	38.4010010
9	0.0718707	0.1184614	26.4041252	0.1245114	0.1176066	0.1334106	45.7220001
10	0.0643054	0.0341472	1.6954064	0.1130994	0.1162311	0.1203403	21.2240009
11	0.0471114	0.0731218	7.7742376	0.0857532	0.0832187	0.0897838	23.7450008
12	0.4466988	-0.0226253	0.7443104	0.3721429	0.5055020	0.4943100	34.9179993
13	0.4126548	-0.0272208	1.0773799	0.3672544	0.4979367	0.4847416	104.9179993
14	0.0618982	0.0287716	1.2036276	0.1093902	0.1127923	0.1161337	74.9599991
15	0.3535076	0.0189503	0.5221487	0.3526189	0.4484182	0.4570799	19.9260006
16	0.4769601	-0.0324537	1.5314077	0.3744686	0.5151307	0.4989383	24.0640003
17	0.4848693	-0.0380895	2.1094849	0.3747789	0.5185695	0.4995421	30.1970005
18	0.4803989	-0.0318471	1.4747020	0.3746155	0.5151307	0.4992316	47.9539986
19	0.2843879	0.0047502	0.0328086	0.3241890	0.4050804	0.4070228	47.6279984
20	0.1767538	-0.0067385	0.0660248	0.2486763	0.2929849	0.2910237	76.5520020
21	0.1231087	0.0125095	0.2275358	0.1925982	0.2132050	0.2159059	76.7740021
22	0.2558459	0.0246556	0.8838877	0.3082818	0.3713893	0.3807776	69.5510025
23	0.4019945	-0.0256567	0.9566721	0.3652105	0.4931224	0.4807899	152.9510040
24	0.3442288	0.0021829	0.0069284	0.3495558	0.4504814	0.4514670	24.3339996
25	0.3349381	-0.0258548	0.9127347	0.3462699	0.4566712	0.4455091	22.6110001
26	0.0213205	0.0778975	8.6426010	0.0408612	0.0305144	0.0417320	22.0600006
27	0.0615543	0.0058446	0.0496666	0.1088571	0.1148556	0.1155307	23.0690002
28	0.3259973	-0.0204167	0.6060913	0.3428897	0.4484182	0.4394461	22.7460003
29	0.3122421	-0.0088326	0.1134329	0.3372614	0.4332875	0.4294940	44.7130013
30	0.3033012	-0.0080695	0.1169793	0.3333157	0.4264099	0.4226192	25.6439991
31	0.3232462	0.0033774	0.0165862	0.3418060	0.4360385	0.4375162	62.1459999
32	0.1915406	0.0428857	2.6741619	0.2617469	0.2964237	0.3097056	65.4080011
33	0.0116919	-0.0118301	0.2034926	0.0228433	0.0233838	0.0231104	40.2470016
34	0.4030261	0.0423829	2.6118393	0.3654193	0.4607978	0.4811922	39.2770004

결측치 대치 이후 표지인자별 MAF, F-statistics, HW_chi_square, PIC, Observed Heterozygosity Expected Heterozygosity Distance between adjacent markers

유효성 판정이 완료된 자료의 통계를 출력한다.

다. 프로그램 코드

(1) fastPHASE프로그램 결과자료를 이용한 집단유전학적 통계 계산 및 allelic code생성

- 결측자료 대치 프로그램인 fastPHASE결과자료를 정리
- fastPHASE프로그램 구동 이후 유전체 자료의 이상치 발생확인
- 염기형태(A, T, G, C)의 유전자형 자료를 allelic code 형태로 변환
- 전체 및 염색체별 표지인자 자료의 기초통계량 계산 및 출력
- 결측자료 대치 프로그램인 fastPHASE연동을 위한 자료 format 변경

```
! Name : Post Imputation & Basic Statistics Analysis
! Developed by : Joonho Lee - PhD working in Hankyong National Univ.(Korea) zoonlee@gmail.com,
+82-10-3408-2895
! Version : Linux 4.0 (2012-03-21)
! Contents :
!           1. Converting Imputation Results -> 0,1,2
!           2. Calculation of gene & genotype frequency
!           4. Statistics of data & population genetics
program RIRnBSA
integer::i,j,k,lz,i1,i2,p,q,r,s liostate & do_loop
integer::NOAC=0,SMNBC=1,NOA=0,ANO=0,NOM=0,MNO=0 !number of chromosomes, animals and markers &
chromosome, animal and marekr number
integer::BC(4)
integer::GC(10)
integer::CMPBC=0,EXMPBC=0,MDG(4)
integer::NODBAM1=0,NODBAM2=0,NODBAM3=0,NODBAM4=0,NODBAM5=0
integer::NOMWMAF1=0,NOMWMAF2=0,NOMWMAF3=0,NOMWMAF4=0,NOMWMAF5=0,CMP=0,EXMP=0,CN=0
,EXCN=999, SODBAM=0,NOMBCN=0 ! for statistics
integer::NODMBOFOMGT=0,NODMBOAH=0,NODMBOFOMIG=0,NODMBOCSOHW=0,NODMBONP=0,NODMB
OAHE=0,EXCNO=0
integer,allocatable::G012DS(:,),CG101DS(:,),NOSMBC(:,),IOM(:,) !data input
real::GF(4),GTF(10),FOMAH,FOMIH,FOHE,FOMAG,FOMIG,IOL(4) !frequencies
real::HWD=0,CSOHW=0,EH=0,PIC=0,IBC=0 !statistics for population genetics
real::SOMAFBC=0,SOIBCBC=0,SOCISOHW=0,SOPICBC=0,SOHEBC=0,SOEHBC=0,SOTMAF=0,SOTIBC=0,SO
TCSOHW=0,SOTPIC=0,SOTHE=0,SOTEH=0 !statistics for population genetics by chromosome
real::SSOMAFBC=0,SSOIBCBC=0,SSOCISOHW=0,SSOPICBC=0,SSOHEBC=0,SSOEHBC=0,SSOTMAF=0,SSO
TIBC=0,SSOTCSOHW=0,SSOTPIC=0,SSOTHE=0 !statistics for population genetics by chromosome
real::SSOTEH=0 !statistics for population genetics by chromosome
real::AOMAFBC=0,AOIBCBC=0,AOCISOHW=0,AOPICBC=0,AOHEBC=0,AOEHBC=0,AOTMAF=0,AOTIBC=0,
AOTCSOHW=0,AOTPIC=0,AOTHE=0,AOTEH=0 !statistics for population genetics by chromosome
real::SDOMAFBC=0,SDOIBCBC=0,SDOCISOHW=0,SDOPICBC=0,SDOHEBC=0,SDOEHBC=0
real::SDOTMAF=0,SDOTIBC=0,SDOTCSOHW=0,SDOTPIC=0,SDOTHE=0,SDOTEH=0 !statistics for population
genetics by chromosome
real::MINOMAFBC=0,MINOIBCBC=0,MINOCISOHW=0,MINOPICBC=0,MINOHEBC=0,MINOEHBC=0
real::MINOTMAF=0,MINOTIBC=0,MINOTCSOHW=0,MINOTPIC=0,MINOTHE=0,MINOTEH=0 !statistics for
population genetics by chromosome
real::MAXOMAFBC=0,MAXOIBCBC=0,MAXOCISOHW=0,MAXOPICBC=0,MAXOHEBC=0,MAXOEHBC=0
real::MAXOTMAF=0,MAXOTIBC=0,MAXOTCSOHW=0,MAXOTPIC=0,MAXOTHE=0,MAXOTEH=0 !statistics
for population genetics by chromosome
real(10)::SODBAMBC=0,SOTDBAM=0,SSODBAMBC=0,SSOTDBAM=0,MINODBAMBC=0,MAXODBAMBC=0,MI
NOTDBAM=0,MAXOTDBAM=0
real(10)::DBAMBC=0,AODBAMBC=0,SDODBAMBC=0,AOTDBAM=0,SDOTDBAM=0
```

```

character(len=1)::BA(4)=(/'A','T','G','C'/),MABA,MIBA
character(len=2)::GT(10)=(/'AA','AT','AG','AC','TT','TG','TC','GG','GC','CC'/),MAH,MIH,HE,CHNO
character(len=8)::OFT,NOAF
character(len=12)::WF1,WF2,WF3
character(len=24)::IDFN='01_DEF0000_genotypes.out'
character(len=300)::Dummy
character(len=220)::IDD,ODD,GDD
character(len=16)::GDFN
character(len=15)::MWF
character(len=250)::IDFNIP,GDFNIP,ODFNIP,PF
character(len=1),allocatable::GDS(:,:)
character(len=2),allocatable::GTDS(:,:)
character(len=50),allocatable::MN(:)
!reading parameters
print
*,
"*****"
*****"
print *, "RIRnBSA_Lv3 : Reformation of Imputation Results & Basic Statistics Analysis Linux version 3.0"
print *, "Developed by Joonho Lee : zoonolee@gmail.com, +82-10-3408-2895"
print *, "Data form is the results of fastPHASE"
print *, "Please enter the name of parameter file including path(ex) ./parGRDH"
print
*,
"*****"
*****"
read *, PF
PF=adjustl(PF)
open (99, file=PF)
read (99,*) OFT
read (99,*) NOA
rewind(unit=99)
read (99,*) OFT
print *, "file tag", OFT
read (99,*) NOAF
print *, "number of animals", NOA
read (99,*) IDD
print *, "directory of imputed genomic data file ", IDD
read (99,*) GDD
print *, "directory of previous data file ", GDD
read (99,*) GDFN
print *, "name of selected marker information file ", GDFN
read (99,*) ODD
print *, "directory of output file", ODD
read (99,fmt='(4f10.5)') IOL(:)
print *, "Outliers : MAF<",IOL(2),"HW-chisquare>",IOL(3)
read (99,fmt='(4i15)') MDG(:)
read (99,*) NOAC
print *, "number of autosomal chromosomes : ", NOAC
allocate(NOSMBC(NOAC))
do i=1,NOAC
  read (99,*) NOSMBC(i)
  NOM=NOM+NOSMBC(i)
enddo
close(99)

```



```

print *, "number of markers", NOM
do p=1,NOAC
  print *, "Number of markers on chromosome No.",p," : ", NOSMBC(p)
enddo
IDD=adjustr(IDD)
allocate(GDS(NOA*2,NOM))
allocate(GTDS(NOM,NOA))
allocate(G012DS(NOM,NOA))
allocate(CG101DS(NOM,NOA))
allocate(IOM(NOM,3))
allocate(MN(NOM))
WF1='          ';WF2='          ';WF3='          ';
write(WF1,fmt='(a12)') '('//adjustr(NOAF)//'i1)'
write(WF2,fmt='(a12)') '('//adjustr(NOAF)//'a3)'
write(WF3,fmt='(a12)') '('//adjustr(NOAF)//'i3)'
print *, WF1
ODD=adjustr(ODD)
GDD=adjustr(GDD)
ODFNIP=ODD//'"Genotype012"//OFT//".out"'
ODFNIP=adjustl(ODFNIP)
open(98,file=ODFNIP,status='unknown')
ODFNIP=ODD//'"Genefrequency"//OFT//".out"'
ODFNIP=adjustl(ODFNIP)
open(97,file=ODFNIP,status='unknown')
write(97,fmt='(2a15)') "0","1"
ODFNIP=ODD//'"Genotypefrequency"//OFT//".out"'
ODFNIP=adjustl(ODFNIP)
open(96,file=ODFNIP,status='unknown')
write(96,fmt='(3a15)') "0","1","2"
ODFNIP=ODD//'"GT_maching_012"//OFT//".out"'
ODFNIP=adjustl(ODFNIP)
open(95,file=ODFNIP,status='unknown')
write(95,fmt='(3a3)') "0","1","2"
ODFNIP=ODD//'"G_maching_01"//OFT//".out"'
ODFNIP=adjustl(ODFNIP)
open(94,file=ODFNIP,status='unknown')
write(94,fmt='(2a3)') "0","1"
ODFNIP=ODD//'"Pop_Gen_Stat"//OFT//".out"'
ODFNIP=adjustl(ODFNIP)
open(93,file=ODFNIP,status='unknown')
write(93,fmt='(8a15)') "MARKER NO.,"MAF","IBC","CSOHWT","PIC","O_HE","E_HE","DBAM"
ODFNIP=ODD//'"Pop_Gen_Stat_BC"//OFT//".out"'
ODFNIP=adjustl(ODFNIP)
open(92,file=ODFNIP,status='unknown')
w r i t e ( 9 2 , f m t = ' ( a 5 , 2 8 a 1 5 ) ' )
"CHNO","AMAF","SDMAF","MINMAF","MAXMAF","AIBC","SDIBC","MINIBC","MAXIBC","ACSOHWT","SDCS
OHWT"&
&,"MINCSOHWT","MAXCSOHWT","APIC","SDPIC","MINPIC","MAXPIC","AHE","SDHE","MINHE","MAXHE","
AEH","SDEH","MINEH","MAXEH"&
&,"ADBAM","SDDBAM","MINDBAM","MAXDBAM"
ODFNIP=ODD//'"Genotype"//OFT//".out"'
ODFNIP=adjustl(ODFNIP)
open(91,file=ODFNIP,status='unknown')

```

```

ODFNIP=ODD//CenteredGT_101//OFT//.out"
ODFNIP=adjustl(ODFNIP)
open(90,file=ODFNIP,status='unknown')
GDFNIP=GDD//GDFN//OFT//.out"
GDFNIP=adjustl(GDFNIP)
open(89,file=GDFNIP,status='old')
ODFNIP=ODD//Statistics//OFT//.out"
ODFNIP=adjustl(ODFNIP)
open(88,file=ODFNIP,status='unknown')
do q=1, NOAC
  iz=ichar('0')
  i1=floor(q/10.)+iz
  i2=mod(q,10)+iz
  CHNO(1:1)=achar(i1)
  CHNO(2:2)=achar(i2)
  IDFN(3:10)=OFT(1:8)
  IDFN(1:2)=CHNO(1:2)
  IDFNIP=IDD//IDFN
  IDFNIP=adjustl(IDFNIP)
  open(q+20,file=IDFNIP,status='old')
  print*,"reading",IDFNIP
  do r=1, 21
    read(q+20,fmt='(a300)',iostat=io) Dummy
    if(io/=0) exit
  enddo
  do s=1, NOA
    read(q+20,fmt='(a300)',iostat=io) Dummy
    read(q+20,*,iostat=io) GDS(s*2-1,SMNBC:(SMNBC+NOSMBC(q)-1))
    read(q+20,*,iostat=io) GDS(s*2,SMNBC:(SMNBC+NOSMBC(q)-1))
  enddo
  print *,"Reading imputed genotype from",SMNBC,"th marker to ",(SMNBC+NOSMBC(q)-1),"th marker"
  SMNBC=SMNBC+NOSMBC(q)
  print *,"Reading imputed genotype file",q,"finished"
close(q+20)
enddo
SOTMAF=0;SOTIBC=0;SOTCSOHWT=0;SOTPIC=0;SOTHE=0;SOTEH=0
SSOTMAF=0;SSOTIBC=0;SSOTCSOHWT=0;SSOTPIC=0;SSOTHE=0;SSOTEH=0
AOTMAF=0;AOTIBC=0;AOTCSOHWT=0;AOTPIC=0;AOTHE=0;AOTEH=0
SDOTMAF=0;SDOTIBC=0;SDOTCSOHWT=0;SDOTPIC=0;SDOTHE=0;SDOTEH=0
MINOTMAF=0;MINOTIBC=0;MINOTCSOHWT=0;MINOTPIC=0;MINOTHE=0;MINOTEH=0
MAXOTMAF=0;MAXOTIBC=0;MAXOCSOHTWT=0;MAXOTPIC=0;MAXTOHE=0;MAXOTEH=0
read(89,fmt='(a300)',iostat=io) Dummy
print *, Dummy
do i=1, NOAC
SOMAFBC=0;SOIBCBC=0;SOCSOHWTBC=0;SOPICBC=0;SOHEBC=0;SOEHBC=0
SSOMAFBC=0;SSOIBCBC=0;SSOCSOHWTBC=0;SSOPICBC=0;SSOHEBC=0;SSOEHBC=0
AOMAFBC=0;AOIBCBC=0;AOCSOHWTBC=0;AOPICBC=0;AOHEBC=0;AOEHBC=0
SDOMAFBC=0;SDOIBCBC=0;SDOCSOHWTBC=0;SDOPICBC=0;SDOHEBC=0;SDOEHBC=0
MINOMAFBC=0;MINOIBCBC=0;MINOCSOHWTBC=0;MINOPICBC=0;MINOHEBC=0;MINOEHBC=0
MAXOMAFBC=0;MAXOIBCBC=0;MAXOCSOHWTBC=0;MAXOPICBC=0;MAXOHEBC=0;MAXOEHBC=0
SODBAMBC=0;SSODBAMBC=0;AODBAMBC=0;SDODBAMBC=0;MINODBAMBC=0;MAXODBAMBC=0;
DBAMBC=0;CMPBC=0;EXMPBC=0
do j=1, NOSMBC(i)

```

```

MNO=MNO+1
read(89,fmt='(a50,i10,2i15)') MN(MNO),IOM(MNO,1),IOM(MNO,2),IOM(MNO,3)
GF=0.0;GTF=0.0;BC=0;GC=0;
do k=1, NOA
  GTDS(MNO,k)=GDS(k*2-1,MNO)//GDS(k*2,MNO)
  select case (GDS(k*2-1,MNO))
    case("A")
      BC(1)=BC(1)+1
    case("T")
      BC(2)=BC(2)+1
    case("G")
      BC(3)=BC(3)+1
    case("C")
      BC(4)=BC(4)+1
  end select
  select case (GDS(k*2,MNO))
    case("A")
      BC(1)=BC(1)+1
    case("T")
      BC(2)=BC(2)+1
    case("G")
      BC(3)=BC(3)+1
    case("C")
      BC(4)=BC(4)+1
  end select
  select case (GTDS(MNO,k))
    case("AA")
      GC(1)=GC(1)+1
    case("AT")
      GC(2)=GC(2)+1
    case("TA")
      GC(2)=GC(2)+1
    case("AG")
      GC(3)=GC(3)+1
    case("GA")
      GC(3)=GC(3)+1
    case("AC")
      GC(4)=GC(4)+1
    case("CA")
      GC(4)=GC(4)+1
    case("TT")
      GC(5)=GC(5)+1
    case("TG")
      GC(6)=GC(6)+1
    case("GT")
      GC(6)=GC(6)+1
    case("TC")
      GC(7)=GC(7)+1
    case("CT")
      GC(7)=GC(7)+1
    case("GG")
      GC(8)=GC(8)+1
    case("GC")

```

```

        GC(9)=GC(9)+1
    case("CG")
        GC(9)=GC(9)+1
    case("CC")
        GC(10)=GC(10)+1
    end select
enddo
GF=real(BC)/real(sum(BC(1:4)))
GTF=real(GC)/real(sum(GC(1:10)))
FOMIH=0.0; FOHE=0.0; FOMAH=0.0; FOMIG=0.0; FOMAG=0.0;
MIH='??';MAH='??';HE='??';MIBA='?';MABA='?'
do l=1,4
    if(GF(l).NE.0 .AND. GF(l)<0.5) then
        MIBA=BA(l)
        FOMIG=GF(l)
        NOMIG=BC(l)
    elseif(GF(l).NE.0 .AND. GF(l)>0.5) then
        MABA=BA(l)
        FOMAG=GF(l)
        NOMAG=BC(l)
    elseif(GF(l)==0.5) then
        if(sum(GF(:l)).eq.0.5) then
            NOMIG=BC(l)
            MIBA=BA(l)
            FOMIG=GF(l)
        else
            MABA=BA(l)
            FOMAG=GF(l)
            NOMAG=BC(l)
        endif
    endif
enddo
MIH=MIBA//MIBA
MAH=MABA//MABA
HE=MABA//MIBA
if (HE=='TA') then
    HE="AT"
elseif(HE=='GA') then
    HE="AG"
elseif(HE=='CA') then
    HE="AC"
elseif(HE=='GT') then
    HE="TG"
elseif(HE=='CT') then
    HE="TC"
elseif(HE=='CG') then
    HE="GC"
endif
do l=1,10
    if(GT(l)==MIH) then
        FOMIH=GTF(l)
        NOMIH=GC(l)
    elseif(GT(l)==MAH) then

```

```

FOMAH=GTF(I)
NOMAH=GC(I)
elseif(GT(I)==HE) then
FOHE=GTF(I)
NOHE=GC(I)
endif
enddo
do I=1,NOA
if(GTDS(MNO,I)==MIH) then
G012DS(MNO,I)=0
CG101DS(MNO,I)=-1
elseif(GTDS(MNO,I)==MAH) then
G012DS(MNO,I)=2
CG101DS(MNO,I)=1
else
G012DS(MNO,I)=1
CG101DS(MNO,I)=0
endif
enddo
write(98,fmt=WF1) G012DS(MNO,1:NOA)
write(90,fmt=WF3) CG101DS(MNO,1:NOA)
write(97,fmt='(2F15.10)') FOMIG,FOMAG
write(96,fmt='(3F15.10)') FOMIH,FOHE,FOMAH
write(95,fmt='(3a3)') MIH,HE,MAH
write(94,fmt='(2a3)') MIBA,MABA
write(91,fmt=WF2) GTDS(MNO,1:NOA)
ENOMIH=(FOMIG**2)*real((NOMIH+NOHE+NOMAH))
ENOHE=(2*FOMIG*FOMAG)*real((NOMIH+NOHE+NOMAH))
ENOMAH=(FOMAG**2)*real((NOMIH+NOHE+NOMAH))

```

CSOHWT=((real(NOMIH)-ENOMIH)**2/ENOMIH)+((real(NOHE)-ENOHE)**2/ENOHE)+((real(NOMAH)-ENOMAH)**2/ENOMAH)

```

EH=1-(FOMIG**2+FOMAG**2)
PIC=1-(FOMIG**2+FOMAG**2)-2*((FOMIG**2)*(FOMAG**2))
IBC=1-(real(NOHE)/ENOHE)
if(FOMIG<IOL(2).or.CSOHWT>IOL(3).or.FOHE.eq.1) print *, "Marker",MNO,"looks
unusual",FOMIG,FOHE,CSOHWT
if(FOMIG .LT. 0.1) then
NOSMWMAF1=NOSMWMAF1+1
elseif(FOMIG .GE. 0.1 .and. FOMIG .LT. 0.2) then
NOSMWMAF2=NOSMWMAF2+1
elseif(FOMIG .GE. 0.2 .and. FOMIG .LT. 0.3) then
NOSMWMAF3=NOSMWMAF3+1
elseif(FOMIG .GE. 0.3 .and. FOMIG .LT. 0.4) then
NOSMWMAF4=NOSMWMAF4+1
elseif(FOMIG .GE. 0.4 ) then
NOSMWMAF5=NOSMWMAF5+1
endif
CMPBC=IOM(MNO,3)
if(j.ne.1) then
DBAMBC=real(CMPBC-EXMPBC)/1000
SODBAMBC=SODBAMBC+DBAMBC
SSODBAMBC=SSODBAMBC+DBAMBC**2

```

```

if(DBAMBC .LT. MDG(1)) then
  NODBAM1=NODBAM1+1
elseif(DBAMBC .GE. MDG(1) .and. DBAMBC .LT. MDG(2)) then
  NODBAM2=NODBAM2+1
elseif(DBAMBC .GE. MDG(2) .and. DBAMBC .LT. MDG(3)) then
  NODBAM3=NODBAM3+1
elseif(DBAMBC .GE. MDG(3) .and. DBAMBC .LT. MDG(4)) then
  NODBAM4=NODBAM4+1
elseif(DBAMBC .GE. MDG(4) ) then
  NODBAM5=NODBAM5+1
endif
if(j.eq.2) then;
  MINODBAMBC=DBAMBC
  MAXODBAMBC=DBAMBC
endif
MINODBAMBC=MIN(MINODBAMBC,DBAMBC)
MAXODBAMBC=MAX(MAXODBAMBC,DBAMBC)
endif
EXMPBC=CMPC
if(j.eq.1) then;

```

```

MINOMAFBC=FOMIG;MINOIBCBC=IBC;MINOCSOHWBTBC=CSOHWBTBC;MINOPICBC=PIC;MINOHEBC=FOHE;MINOEHBC=EH

```

```

MAXOMAFBC=FOMIG;MAXOIBCBC=IBC;MAXOCSOHWBTBC=CSOHWBTBC;MAXOPICBC=PIC;MAXOHEBC=FOHE;MAXOEHBC=EH
endif

```

```

MINOMAFBC=MIN(MINOMAFBC,FOMIG);MINOIBCBC=MIN(MINOIBCBC,IBC);MINOCSOHWBTBC=MIN(MINOCSOHWBTBC,CSOHWBTBC);

```

```

MINOPICBC=MIN(MINOPICBC,PIC);MINOHEBC=MIN(MINOHEBC,FOHE);MINOEHBC=MIN(MINOEHBC,EH)

```

```

MAXOMAFBC=MAX(MAXOMAFBC,FOMIG);MAXOIBCBC=MAX(MAXOIBCBC,IBC);MAXOCSOHWBTBC=MAX(MAXOCSOHWBTBC,CSOHWBTBC);

```

```

MAXOPICBC=MAX(MAXOPICBC,PIC);MAXOHEBC=MAX(MAXOHEBC,FOHE);MAXOEHBC=MAX(MAXOEHBC,EH)

```

```

SOMAFBC=SOMAFBC+FOMIG
SOIBCBC=SOIBCBC+IBC
SOCOHWBTBC=SOCOHWBTBC+CSOHWBTBC
SOPICBC=SOPICBC+PIC
SOHEBC=SOHEBC+FOHE
SOEHBC=SOEHBC+EH
SSOMAFBC=SSOMAFBC+FOMIG**2
SSOIBCBC=SSOIBCBC+IBC**2
SSOCOHWBTBC=SSOCOHWBTBC+CSOHWBTBC**2
SSOPICBC=SSOPICBC+PIC**2
SSOHEBC=SSOHEBC+FOHE**2
SSOEHBC=SSOEHBC+EH**2

```

```

write(93,fmt='(i15,7F15.7)') MNO,FOMIG,IBC,CSOHWBTBC,PIC,FOHE,EH,DBAMBC

```

```

enddo

```

```

AOMAFBC=SOMAFBC/real(NOSMBC(i))

```

```

AOIBCBC=SOIBCBC/real(NOSMBC(i))
AOC SOHWTBC=SOCSOHWTBC/real(NOSMBC(i))
AOPICBC=SOPICBC/real(NOSMBC(i))
AOHEBC=SOHEBC/real(NOSMBC(i))
AOEHBC=SOEHBC/real(NOSMBC(i))
AODBAMBC=SODBAMBC/real(NOSMBC(i)-1)
SDOMAFBC=(SSOMAFBC-SOMAFBC**2/real(NOSMBC(i)))/real(NOSMBC(i))
SDOIBCBC=(SSOIBCBC-SOIBCBC**2/real(NOSMBC(i)))/real(NOSMBC(i))
SDOCSOHWTBC=(SSOCSOHWTBC-SOCSOHWTBC**2/real(NOSMBC(i)))/real(NOSMBC(i))
SDOPICBC=(SSOPICBC-SOPICBC**2/real(NOSMBC(i)))/real(NOSMBC(i))
SDOHEBC=(SSOHEBC-SOHEBC**2/real(NOSMBC(i)))/real(NOSMBC(i))
SDOEHBC=(SSOEHBC-SOEHBC**2/real(NOSMBC(i)))/real(NOSMBC(i))
SDODBAMBC=(SSODBAMBC-SODBAMBC**2/real(NOSMBC(i)-1))/real(NOSMBC(i)-1)
w r i t e ( 9 2 , f m t = ' ( i 5 , 2 8 f 1 5 . 5 ) ' )
i,AOMAFBC,SDOMAFBC,MINOMAFBC,MAXOMAFBC,AOIBCBC,SDOIBCBC,MINOIBCBC,MAXOIBCBC,&
AOC SOHWTBC,SDOCSOHWTBC,MINOCSOHWTBC,MAXOCSOHWTBC,AOPICBC,SDOPICBC,MINOPICBC,M
AXOPICBC,AOHEBC,SDOHEBC,&
MINOHEBC,MAXOHEBC,AOEHBC,SDOEHBC,MINOEHBC,MAXOEHBC,AODBAMBC,SDODBAMBC,MINODBA
MBC,MAXODBAMBC
SOTMAF=SOTMAF+SOMAFBC
SOTIBC=SOTIBC+SOIBCBC
SOTCSOHWT=SOTCSOHWT+SOCSOHWTBC
SOTPIC=SOTPIC+SOPICBC
SOTHE=SOTHE+SOHEBC
SOTEH=SOTEH+SOEHBC
SOTDBAM=SOTDBAM+SODBAMBC
SSOTMAF=SSOTMAF+SSOMAFBC
SSOTIBC=SSOTIBC+SSOIBCBC
SSOTCSOHWT=SSOTCSOHWT+SSOCSOHWTBC
SSOTPIC=SSOTPIC+SSOPICBC
SSOTHE=SSOTHE+SSOHEBC
SSOTEH=SSOTEH+SSOEHBC
SSOTDBAM=SSOTDBAM+SSODBAMBC
if(i.eq.1) then;
MINOTMAF=MINOMAFBC;MINOTIBC=MINOIBCBC;MINOTCSOHWT=MINOCSOHWTBC;MINOTPIC=MINOP
ICBC;MINOTHE=MINOHEBC;MINOTEH=MINOEHBC
MAXOTMAF=MAXOMAFBC;MAXOTIBC=MAXOIBCBC;MAXOTCSOHWT=MAXOCSOHWTBC;MAXOTPIC=
MAXOPICBC;MAXOTHE=MAXOHEBC;MAXOTEH=MAXOEHBC
MINOTDBAM=MINODBAMBC;MAXOTDBAM=MAXODBAMBC;
endif
MINOTMAF=MIN(MINOMAFBC,MINOTMAF);MINOTIBC=MIN(MINOIBCBC,MINOTIBC);MINOTCSOHWT=M
IN(MINOCSOHWTBC,MINOTCSOHWT);
MINOTPIC=MIN(MINOPICBC,MINOTPIC);MINOTHE=MIN(MINOHEBC,MINOTHE);MINOTEH=MIN(MINOEH
BC,MINOTEH)
MAXOTMAF=MAX(MAXOMAFBC,MAXOTMAF);MAXOTIBC=MAX(MAXOIBCBC,MAXOTIBC);MAXOTCSO
HWT=MAX(MAXOCSOHWTBC,MAXOTCSOHWT);

```

```

MAXOTPIC=MAX(MAXOPICBC,MAXOTPIC);MAXOTHE=MAX(MAXOHEBC,MAXOTHE);MAXOTEH=MAX(
MAXOEHBC,MAXOTEH)
  MINOTDBAM=MIN(MINODBAMBC,MINOTDBAM);MAXOTDBAM=MAX(MAXODBAMBC,MAXOTDBAM);
  print *,"Reading marker information on",IOM(MNO,2),"th chromosome"
  print *,"Converting imputed genotype file and analyzing statistics of chromosome",i,"finished"
enddo
AOTMAF=SOTMAF/real(NOM)
AOTIBC=SOTIBC/real(NOM)
AOTCSOHWI=SOTCSOHWI/real(NOM)
AOTPIC=SOTPIC/real(NOM)
AOTHE=SOTHE/real(NOM)
AOTEH=SOTEH/real(NOM)
AOTDBAM=SOTDBAM/real(NOM-NOAC)
SDOTMAF=(SSOTMAF-SOTMAF**2/real(NOM))/real(NOM)
SDOTIBC=(SSOTIBC-SOTIBC**2/real(NOM))/real(NOM)
SDOTCSOHWI=(SSOTCSOHWI-SOTCSOHWI**2/real(NOM))/real(NOM)
SDOTPIC=(SSOTPIC-SOTPIC**2/real(NOM))/real(NOM)
SDOTHE=(SSOTHE-SOTHE**2/real(NOM))/real(NOM)
SDOTEH=(SSOTEH-SOTEH**2/real(NOM))/real(NOM)
SDOTDBAM=(SSOTDBAM-SOTDBAM**2/real(NOM-NOAC))/real(NOM-NOAC)
write(92,fmt='(a5,28f15.5)')
  "Tot",AOTMAF,SDOTMAF,MINOTMAF,MAXOTMAF,AOTIBC,SDOTIBC,MINOTIBC,MAXOTIBC,&
AOTCSOHWI,SDOTCSOHWI,MINOTCSOHWI,MAXOTCSOHWI,AOTPIC,SDOTPIC,MINOTPIC,MAXOTPIC,
AOTHE,SDOTHE,MINOTHE,MAXOTHE,&
&AOTEH,SDOTEH,MINOTEH,MAXOTEH,AOTDBAM,SDOTDBAM,MINOTDBAM,MAXOTDBAM
  write(88,fmt='(a120)') "***** frequency of MAF *****"
  write(88,fmt='(a120)') "number of markers with Minor allele frequency"
  write(88,fmt='(a120,i15)') "      minor allele frequency < 0.1 : ", NOSMWMAF1
  write(88,fmt='(a120,i15)') "0.1 <= minor allele frequency < 0.2 : ", NOSMWMAF2
  write(88,fmt='(a120,i15)') "0.2 <= minor allele frequency < 0.4 : ", NOSMWMAF3
  write(88,fmt='(a120,i15)') "0.3 <= minor allele frequency < 0.4 : ", NOSMWMAF4
  write(88,fmt='(a120,i15)') "0.4 <= minor allele frequency      : ", NOSMWMAF5
  write(88,fmt='(a120)') " "
  write(88,fmt='(a120)') "***** frequency of distance *****"
  write(88,fmt='(a120)') "number of total markers with distance between adjacent markers"
  write(88,fmt='(50x,a55,2i15)') "      distance between adjacent markers < ",MDG(1), NODBAM1
  write(88,fmt='(50x,i15,a40,2i15)') MDG(1),"<= distance between adjacent markers <",MDG(2), NODBAM2
  write(88,fmt='(50x,i15,a40,2i15)') MDG(2),"<= distance between adjacent markers <",MDG(3), NODBAM3
  write(88,fmt='(50x,i15,a40,2i15)') MDG(3),"<= distance between adjacent markers <",MDG(4), NODBAM4
  write(88,fmt='(50x,i15,a55,i15)') MDG(4),"<= distance between adjacent markers      ",
NODBAM5
end program RIRnBSA

```


(2) 대용량 유전체 정보를 이용한 연관불평형 분석 및 유효집단크기 추정

- 대용량 유전체 정보를 이용한 연관불평형 분석
- 동일 염색체에 존재는 표지인자 쌍의 haplotype빈도 추정
- 연관불평형 분석 자료를 이용한 유효집단크기 추정

```
! Name : Extent of Linkage Disequilibrium
! Developed by : Joonho Lee - PhD working in Hankyong National Univ.(Korea) zoonolee@gmail.com,
+82-10-3408-2895
! Version : Linux 1.0 (2012-03-26)
! Contents :
!           1. Calculation of Haplotype frequency
!           2. Estimation of Haplotype frequency using EM algorithm
!           3. Calculation of Linkage Disequilibrium
!           4. Estimation of Effective Population Size
program EOLD
integer::CNO,MNO1,MNO2,NEJ,iter,DummyI,iz,i1,i2,NONEG !iostate & do_loop
integer::NOA=0,ANO=0,NOM=0,MNO=0,NOHTBC=0,CBPTCM=1000000 !number of animals and markers &
animal and marekr number
integer::nMIHMIH,nMIHHE,nMIHMAH,nHEMIH,nHEHE,nHEMAH,nMAHMIH,nMAHHE,nMAHMAH
integer,allocatable::G0I2DS(:,:),SEMNBC(:,:),MP(:,:),NOSMBC(:,:),NEGA(:,:),CNEG(:) !data storage
real::D=0,rsquare=0,CONV=0,DLFLDD=0,HG=0,DFNEG=0,EPS=0
real::nMIMI,nMIMA,nMAMI,nMAMA
real::MIMIif,MIMAf,MAMIif,MAMAf,newMIMIif,newMIMAf,newMAMIif,newMAMAf,prMIMIMAMA,prMIMAMAM
I,diff=0.0
real(10)::SLDBC=0,ALDBC=0,SALDBC=0,AALDBC=0,STLD=0,ATLD=0,STALD=0,ATALD=0
real,allocatable::GF(:,:),NEG(:,:),SNEG(:,:),ANEG(:) !output data
character(len=2)::CHNO
character(len=50)::MN
character(len=8)::OFT,NOAF
character(len=16)::WF1
character(len=300)::Dummy
character(len=220)::MDD,ODD,GDD
character(len=30)::GDFN1,GDFN2,MDFN
character(len=250)::MDFNIP,GDFNIP,ODFNIP,PF
character(len=20),allocatable::ID(:)
!reading parameters
print
print
"*****"
*****"
print *, "EOLD_Lv1 : Extent of Linkage Disequilibrium Linux version 1.0"
print *, "Developed by Joonho Lee : zoonolee@gmail.com, +82-10-3408-2895"
print *, "Data form is the results of Post_fastPHASE and of converted genotype 0,1,2"
print *, "Please enter the name of parameter file including path(ex) ./parGRDH"
print
"*****"
*****"
read *, PF
PF=adjustl(PF)
open (99, file=PF)
read (99,*) OFT
print *, "file tag ", OFT
```

```

read (99,*) NOA
rewind(unit=99)
read (99,*) OFT
read (99,*) NOAF
print *, "number of animals", NOA
read (99,*) NOM
print *, "number of markers", NOM
read (99,*) GDD
print *, "directory of Post_Imputation output file ", GDD
read (99,*) GDFN1
print *, "name of allelic coding data file(example : Genotype012_OUT_TAG.out) ", GDFN1
read (99,*) GDFN2
print *, "name of allele frequency data file(example : Genefrequency_OUT_TAG.out) ", GDFN2
read (99,*) MDD
print *, "directory of Selected markers file ", MDD
read (99,*) MDFN
print *, "name of Selected markers file(example : Selected_markers_OUT_TAG.out) ", MDFN
read (99,*) ODD
print *, "directory of output file ", ODD
read (99,*) CONV
print *, "convergence of haplotype frequency estimation ", CONV
read (99,*) CBPTCM
print *, "1 centi morgan(cM) to base pair(bp) (example 1000000) ", CBPTCM
read (99,*) DLFLDD
print *, "distance limit for LD decay (unit : cM) ", DLFLDD
read (99,*) NOAC
print *, "number of autosomal chromosomes : ", NOAC
allocate(SEMNBC(NOAC,2))
allocate(NOSMBC(NOAC))
allocate(G012DS(NOM,NOA))
allocate(MP(NOM))
allocate(GF(NOM,2))
ODD=adjustr(ODD)
GDD=adjustr(GDD)
MDD=adjustr(MDD)
MNO=0
do CNO=1,NOAC
  read (99,*) NOSMBC(CNO)
  SEMNBC(CNO,1)=MNO+1
  MNO=MNO+NOSMBC(CNO)
  SEMNBC(CNO,2)=MNO
  print *, "Marker number on chromosome No.",CNO," from ", SEMNBC(CNO,1)," to ", SEMNBC(CNO,2)
  print *, "Number of markers on chromosome No.",CNO," : ", NOSMBC(CNO)
  iz=ichar('0')
  i1=floor(CNO/10.)+iz
  i2=mod(CNO,10)+iz
  CHNO(1:1)=achar(i1)
  CHNO(2:2)=achar(i2)
  print *, "Name of LD analysis result of chromosome No.",CNO," : ", "LD_CH_"//CHNO(1:2)//OFT//".out"
  ODFNIP=ODD// "LD_CH_"//CHNO(1:2)//OFT//".out"

```

```

ODFNIP=adjustl(ODFNIP)
open(CNO+20,file=ODFNIP,status='unknown')
enddo
read(99,*) NONEG
print *, "number of groups for NE estimation : ", NONEG
allocate(NEG(NONEG,2))
allocate(NEGA(NONEG,2))
allocate(CNEG(NONEG))
allocate(SNEG(NONEG))
allocate(ANEG(NONEG))
do i=1,NONEG
  read(99,*) NEGA(i,:)
  NEG(i,1)=1/real(NEGA(i,1)*2)*100;NEG(i,2)=1/real(NEGA(i,2)*2)*100
  print *, "effective population size from ",NEGA(i,1)," to ",NEGA(i,2)," generations ago "
  print *, "using average LD of distance between ",NEG(i,1)," and ",NEG(i,2)
enddo
close(99)
WF1='
write(WF1,fmt='(a16)') '(//adjustr(NOAF)//"i1    )'
print *, "reading format for Selected markers file is : ",WF1
GDFNIP=GDD//GDFN1
GDFNIP=adjustl(GDFNIP)
open(98,file=GDFNIP,status='old')
GDFNIP=GDD//GDFN2
GDFNIP=adjustl(GDFNIP)
open(97,file=GDFNIP,status='old')
MDFNIP=MDD//MDFN
MDFNIP=adjustl(MDFNIP)
open(96,file=MDFNIP,status='old')
ODFNIP=ODD//LD_Decay//OFT//".out"
ODFNIP=adjustl(ODFNIP)
open(95,file=ODFNIP,status='unknown')
ODFNIP=ODD//LD_STAT_BC//OFT//".out"
ODFNIP=adjustl(ODFNIP)
open(94,file=ODFNIP,status='unknown')
ODFNIP=ODD//NE_estimation//OFT//".out"
ODFNIP=adjustl(ODFNIP)
open(93,file=ODFNIP,status='unknown')
ODFNIP=ODD//LD_Adjacent//OFT//".out"
ODFNIP=adjustl(ODFNIP)
open(92,file=ODFNIP,status='unknown')
read(97,*) DummyI,DummyI
read(96,*)
do MNO=1, NOM
  read(98,fmt=WF1) G012DS(MNO,:)
  read(97,*) GF(MNO,:)
  read(96,*) MN,DummyI,DummyI,MP(MNO)
enddo
do CNO=1, NOAC
NOHTBC=0

```

```

write(92,*) "chromosome No. ", CNO
SLDBC=0;ALDBC=0;SALDBC=0;AALDBC=0
do MNO1=SEMNBC(CNO,1), SEMNBC(CNO,2)
do MNO2=MNO1+1,SEMNBC(CNO,2)
HG=real(MP(MNO2)-MP(MNO1))/CBPTCM
NOHTBC=NOHTBC+1

nMIHMIH=0;nMIHHE=0;nMIHMAH=0;nHEMIH=0;nHEHE=0;nHEMAH=0;nMAHMIH=0;nMAHHE=0;nMAHMAH=
0
do ANO=1,NOA
select case (G012DS(MNO1,ANO))
case(0)
select case (G012DS(MNO2,ANO))
case(0)
nMIHMIH=nMIHMIH+1
case(1)
nMIHHE=nMIHHE+1
case(2)
nMIHMAH=nMIHMAH+1
end select
case(1)
select case (G012DS(MNO2,ANO))
case(0)
nHEMIH=nHEMIH+1
case(1)
nHEHE=nHEHE+1
case(2)
nHEMAH=nHEMAH+1
end select
case(2)
select case (G012DS(MNO2,ANO))
case(0)
nMAHMIH=nMAHMIH+1
case(1)
nMAHHE=nMAHHE+1
case(2)
nMAHMAH=nMAHMAH+1
end select
end select
enddo
nMIMI=2*nMIHMIH+nMIHHE+nHEMIH+real(nHEHE)/2
nMIMA=nMIHHE+2*nMIHMAH+real(nHEHE)/2+nHEMAH
nMAMI=nMAHHE+2*nMAHMIH+real(nHEHE)/2+nHEMIH
nMAMA=2*nMAHMAH+nMAHHE+nHEMAH+real(nHEHE)/2
do iter=1,50
MIMI f=nMIMI/(nMIMI+nMIMA+nMAMI+nMAMA)
MIMA f=nMIMA/(nMIMI+nMIMA+nMAMI+nMAMA)
MAMI f=nMAMI/(nMIMI+nMIMA+nMAMI+nMAMA)
MAMA f=nMAMA/(nMIMI+nMIMA+nMAMI+nMAMA)
prMIMIMAMA=(MIMI f*MAMA f)/((MIMI f*MAMA f)+(MIMA f*MAMI f))

```

```

prMIMAMAMI=(MIMAf*MAMIf)/(MIMIf*MAMAf)+(MIMAf*MAMIf)
nMIMI=2*nMIHMIH+nMIHHE+nHEHE*prMIMIMAMA)
nMIMA=nMIHHE+2*nMIHMAH+(nHEHE*prMIMAMAMI)+nHEMAH
nMAMI=nMAHHE+2*nMAHMIH+(nHEHE*prMIMAMAMI)+nHEMIH
nMAMA=2*nMAHMAH+nMAHHE+nHEMAH+(nHEHE*prMIMIMAMA)
newMIMIf=nMIMI/(nMIMI+nMIMA+nMAMI+nMAMA)
newMIMAf=nMIMA/(nMIMI+nMIMA+nMAMI+nMAMA)
newMAMIf=nMAMI/(nMIMI+nMIMA+nMAMI+nMAMA)
newMAMAf=nMAMA/(nMIMI+nMIMA+nMAMI+nMAMA)
DIFF=2*abs(newMIMIf-MIMIf)
if( DIFF < CONV ) exit
enddo
D=(newMIMIf*newMAMAf)-(newMIMAf*newMAMIf)
rsquare=(D**2)/(GF(MNO1,1)*GF(MNO1,2)*GF(MNO2,1)*GF(MNO2,2))
SLDBC=SLDBC+rsquare
write(CNO+20,fmt='(2i15,f15.10,f15.12)') MNO1,MNO2,HG,rsquare
if(HG.le.DLFLDD) write(95,fmt='(2f15.10)') HG, rsquare
if(MNO2-MNO1.eq.1) then
  SALDBC=SALDBC+rsquare
  write(92,fmt='(2f15.10)') HG, rsquare
endif
do NEJ=1, NONEG
  if(HG.gt.NEG(NEJ,2).and.HG.gt.NEG(NEJ,2)) then
    CNEG(NEJ)=CNEG(NEJ)+1
    SNEG(NEJ)=SNEG(NEJ) + rsquare
  endif
enddo
enddo
enddo
print *, "haplotype frequency calculation of chromosome", CNO, "is finished"
print *, "haplotype number on chromosome", CNO, "is", NOHTBC
close(CNO+20)
ALDBC=SLDBC/real(NOHTBC);AALDBC=SALDBC/real(NOSMBC(CNO)+1)
STLD=STLD+SLDBC;STALD=STALD+SALDBC
NOTHT=NOTHT+NOHTBC
write(94,fmt='(a12,i5,i10,2f15.10)') 'chromosome',CNO,NOHTBC,ALDBC,AALDBC
enddo
ATLD=STLD/real(NOTHT);ATALD=STALD/real(NOM-NOAC)
write(94,fmt='(a27, 2f15.10)') 'total', ATLD, ATALD
write(93,*) 'Estimated effective population size'
do NEJ=1, NONEG
  ANEG(NEJ)=SNEG(NEJ)/CNEG(NEJ)
  EPS=(1/((NEG(NEJ,1)+NEG(NEJ,2))/200))*((1/ANEG(NEJ))-1.0)
  write(93,fmt='(i4,a3,i4,10x,f20.10)') NEGA(NEJ,1)," ~ ",NEGA(NEJ,2),EPS
enddo
end program EOLD

```

(3) Single Marker Regression을 이용한 표지인자의 표현형 연관성분석

- Single Marker Regression을 이용한 표지인자의 표현형 연관성분석
- 표지인자의 표현형연관 유의성에 대한 F검정 및 유의성 출력

```
! Name : Single Marker Regression
! Developed by : Joonho Lee - PhD working in Hankyong National Univ.(Korea) zoonolee@gmail.com,
+82-10-3408-2895
! Version : Linux 1.0 (2012-04-09)
! Contents :          1. Estimation of marker effect
!                  2. Test of significant using F distribution
program SMR
integer::ANO=0,NOM=0,MNO=0,NOR=0,RNO=0,NOTR=0,TRNO=0,FENO=0,NOFE=0,Dummy,PCHNO=1,MAXP=0,S
G !numbers and counting variables
integer,allocatable::TG012DS(:,),G012DS(:,),NOA(:,),FEDS(:,),IOM(:,),NOMWSG(:,),OMN(:,),NOSM(:)      !data
storage
real::JtJ,JtX,XtJ,XtX,JtY,XtY,YtY,det,mu_hat,g_hat,F_val,tmp,P_val,log10P_val,CLOP
real,allocatable::PRDS(:)
character(len=3),allocatable::NAOTR(:)
character(len=8),allocatable::NOAF(:)
character(len=50),allocatable::MN(:)
character(len=8)::OFT,NOMF,NOFEF,NOTRF
character(len=4)::MAXPF
character(len=16)::WF1,WF2,WF3,WF4,WF5
character(len=220)::MDD,ODD,GDD1,GDD2,PDD
character(len=30)::GDFN1,GDFN2,PDFN,MDFN
character(len=250)::MDFNIP,GDFNIP1,GDFNIP2,ODFNIP,PDFNIP,PF
!reading parameters
print
print
"*****"
*****"
print *, "MGP_Lv1 : Extent of Linkage Disequilibrium Linux version 1.0"
print *, "Developed by Joonho Lee : zoonolee@gmail.com, +82-10-3408-2895"
print *, "Genotype data form is the results of Post_fastPHASE and of converted genotype 0,1,2"
print *, "Maximum number of trait is 20"
print *, "Phenotype data form is animalID, Fixedeffect...,Trait..."
print *, "Fixed effect must be numbered as integer"
print *, "Trait name must be consisted of 3 character ex) EMA,BFT,..."
print *, "Please enter the name of parameter file including path(ex) ./parGRDH"
print
print
"*****"
*****"
read *, PF
PF=adjustl(PF)
open (99, file=PF)
read (99,*) OFT
print *, "file tag ", OFT
read (99,*) NOM
read (99,*) MAXP
```

```

read (99,*) CLOP
read (99,*) NOTR
allocate(MN(NOM))
allocate(OMN(NOM))
allocate(IOM(NOM,2))
rewind(unit=99)
read (99,*) OFT
read (99,*) NOMF
read (99,*) MAXPF
read (99,*) CLOP
read (99,*) NOTRF
print *, "number of markers", NOM
print *, "number of traits", NOTR
allocate(NAOTR(NOTR))
allocate(NOANOTR(NOTR))
allocate(NOAF(NOTR))
allocate(NOSM(NOTR))
allocate(NOMWSG(MAXP+1,NOTR))
read (99,*) NAOTR(:)
print *, "name of traits and number of records"
do TRNO=1,NOTR
  read (99,*) NOA(TRNO)
  backspace(unit=99)
  read (99,*) NOAF(TRNO)
  print *, NAOTR(TRNO),NOA(TRNO),NOAF(TRNO)
enddo
read (99,*) GDD1
print *, "directory of matched genotype and phenotype data file ", GDD1
read (99,*) GDD2
print *, "directory of Selected_Markers file ", GDD2
read (99,*) GDFN2
print *, "file name of Selected_Markers ", GDFN2
read (99,*) ODD
print *, "directory of output file ", ODD
close(99)
ODD=adjustr(ODD)
GDD1=adjustr(GDD1)
GDD2=adjustr(GDD2)
GDFN2=adjustl(GDFN2)
GDFNIP2=GDD2//GDFN2
GDFNIP2=adjustl(GDFNIP2)
open(98,file=GDFNIP2,status='old')
read(98,*)
do MNO=1,NOM
  read(98,*) MN(MNO),OMN(MNO),IOM(MNO,1),IOM(MNO,2)
enddo
WF1='          ';WF2='          ';WF3='          ';WF4='          '
write(WF1,fmt='(a16)') '(//adjustr(NOMF)//i1      )'

```

```

write(WF2,fmt='(a16)') '(a9,//adjustr(NOTRF)//a15)'
write(WF3,fmt='(a16)') '(i9,//adjustr(NOTRF)//i15)'
write(WF4,fmt='(a16)') '(a9,//adjustr(NOTRF)//i15)'
print *, "reading format for matched genotype file is : ",WF1
print *, "writing format for number of signigicant markers file is : ",WF2,WF3,WF4
ODFNIP=ODD//SM_frequency//OFT//.out"
ODFNIP=adjustl(ODFNIP)
open(97,file=ODFNIP,status='unknown')
write(97,fmt=WF2) "-log10_P", NAOTR(:)
ODFNIP=ODD//SM_numbers//OFT//.out"
ODFNIP=adjustl(ODFNIP)
open(96,file=ODFNIP,status='unknown')
NOMWSG=0
NOSM=0
do TRNO=1, NOTR
  WF5='
  write(WF5,fmt='(a16)') '(//adjustr(NOAF(TRNO))//i1 )'
  GDFNIP1=GDD1//GT//NAOTR(TRNO)//OFT//.out"
  GDFNIP1=adjustl(GDFNIP1)
  open(TRNO+10,file=GDFNIP1,status='old')
  GDFNIP1=GDD1//PR//NAOTR(TRNO)//OFT//.out"
  GDFNIP1=adjustl(GDFNIP1)
  open(TRNO+30,file=GDFNIP1,status='old')
  ODFNIP=ODD//ME//NAOTR(TRNO)//OFT//.out"
  ODFNIP=adjustl(ODFNIP)
  open(TRNO+70,file=ODFNIP,status='unknown')
  write(TRNO+70,fmt='(a50,7a15)')"Marker
  Name","Chromosome","Position","Mu_hat","Me_hat","F_val","P_val","-log10P_val"
  allocate(TG012DS(NOAF(TRNO),NOM))
  allocate(G012DS(NOM,NOAF(TRNO)))
  allocate(PRDS(NOAF(TRNO)))
  do ANO=1,NOAF(TRNO)
    read(TRNO+10,fmt=WF1) TG012DS(ANO,:)
    read(TRNO+30,*) PRDS(ANO)
  enddo
  G012DS=transpose(TG012DS)
  close(TRNO+10)
  close(TRNO+30)
  ODFNIP=ODD//RG//NAOTR(TRNO)//OFT//.out"
  ODFNIP=adjustl(ODFNIP)
  open(TRNO+10,file=ODFNIP,status='unknown')
  ODFNIP=ODD//RM//NAOTR(TRNO)//OFT//.out"
  ODFNIP=adjustl(ODFNIP)
  open(TRNO+30,file=ODFNIP,status='unknown')
  JtJ=NOAF(TRNO)
! print *, "JtJ",JtJ
  JtY=sum(PRDS)
! print *, "JtY",JtY

```



```

YtY=sum(PRDS**2)
! print *, "YtY",YtY
do MNO=1,NOM
  JtX=sum(TG012DS(:,MNO))
  XtJ=JtX
!   print *, "XtJ and JtX",XtJ, JtX
  XtX=sum(TG012DS(:,MNO)**2)
!   print *, "XtX",XtX
  XtY=dot_product(TG012DS(:,MNO),PRDS)
!   print *, "XtY",XtY
  det=1/(JtJ*XtX-XtJ**2)
!   print *, "det",det
  mu_hat=(XtX*JtY-JtX*XtY)*det
  g_hat=(JtJ*XtY-XtJ*JtY)*det
  F_val=(NOA(TRNO)-1)*abs(g_hat)*XtY/(YtY-abs(g_hat)*XtY)
  tmp=(NOA(TRNO)-1)/((NOA(TRNO)-1)+1*F_val)
  P_val=BETAI((NOA(TRNO)-1)/2.,1/2.,tmp)
  log10P_val=log10(P_val)*(-1.0)
  if(P_val.lt.CLOP) then
    NOSM(TRNO)=NOSM(TRNO)+1
    write(TRNO+30,fmt='(a50,i10,2i15,f15.10)') MN(MNO),OMN(MNO),IOM(MNO,1),IOM(MNO,2),P_val
    write(TRNO+10,fmt=WF5) G012DS(MNO,:)
  endif
  do SG=1, MAXP
    if(log10P_val.ge.(SG-1).and.log10P_val.lt.SG) NOMWSG(SG,TRNO)=NOMWSG(SG,TRNO)+1
  enddo
  if(log10P_val.ge.MAXP) NOMWSG(MAXP+1,TRNO)=NOMWSG(MAXP+1,TRNO)+1
  if(PCHNO.ne.IOM(MNO,1)) then
    print *, "Association test of trait-",NAOTR(TRNO)," using markers on chromosome",PCHNO,"is
finished"
  endif
  PCHNO=IOM(MNO,1)
  write(TRNO+70,fmt='(a50,2i15,5f15.8)') MN(MNO),IOM(MNO,1),IOM(MNO,2),mu_hat, g_hat, F_val, P_val,
log10P_val
!   print '(5a15)', "mu_hat", "g_hat", "F_val", "P_val", "log10P_val"
!   print *, mu_hat, g_hat, F_val, P_val, log10P_val
  enddo
  deallocate(TG012DS,G012DS,PRDS)
enddo
print *, "Association test of trait-",NAOTR(NOTR)," using markers on chromosome",PCHNO,"is finished"
do SG=1, MAXP
  write(97,fmt=WF3) SG, NOMWSG(SG,:)
enddo
write(97,fmt=WF4) "over"//MAXPF,NOMWSG(MAXP+1,:)
write(97,fmt=WF4) "recordsNo",NOA(:)
write(97,fmt=WF4) "No.SM",NOSM(:)
write(96,fmt='(2a20)') "Trait","Significant markers"
do TRNO=1, NOTR

```

```

write(96,fmt='(a20,i20)') NAOTR(TRNO),NOSM(TRNO)
enddo
end program SMR
!*****
!Returns the value ln(Gamma(xx)) for xx>0. Full accuracy is obtained
!for xx > 1. For 0<xx<1, the reflection formula can be used first:
!!   Gamma(1-z) = pi/Gamma(z)/sin(pi*z) = pi*z/Gamma(1+z)/sin(pi*z)
!*****
Function GAMMLN(xx)
real*8 cof(6),stp,half,one,fpf,x,tmp,ser !internal arithmetic in double precision
data cof,stp/76.18009173D0,-86.50532033D0,24.01409822D0,      &
      -1.231739516D0,.120858003D-2,-.536382D-5,2.50662827465D0/
data half,one,fpf/0.5d0,1.d0,5.5d0/
  x=xx-one
  tmp=x+fpf
  tmp=(x+half)*log(tmp)-tmp
  ser=one
  do j=1,6
    x=x+one
    ser=ser+cof(j)/x
  end do
  GAMMLN=tmp+log(stp*ser)
  return
End
!*****
!Returns the incomplete beta function Ix(a,b)
!*****
Function BETAI(a,b,x)
  if(x.lt.0..or.x.gt.1.) print *, ' Bad argument x in function BetaI.'
  if(x.eq.0..or.x.eq.1.) then
    bt=0.
  else
    tmp=gammln(a+b)-gammln(a)-gammln(b)+a*log(x)+b*log(1.-x)
    bt=exp(gammln(a+b)-gammln(a)-gammln(b) &
      +a*log(x)+b*log(1.-x))
  end if
  if (x.lt.(a+1.)/(a+b+2.)) then !use continued fraction directly
    BETAI=bt*BETACF(a,b,x)/a
    return
  else
    BETAI=1.-bt*BETACF(b,a,1.-x)/b !use continued fraction after
    return !making the symmetry transformation
  end if
End
!*****
! Continued fraction for incomplete beta function (used by BETAI)
!*****
Function BETACF(a,b,x)

```

```

Parameter(MAXIT=100,EPS=3e-7,FPMIN=1e-30)
qab=a+b; qap=a+1.0; qam=a-1.0
c=1.0
d=1.0-qab*x/qap
if (abs(d) < FPMIN) d=FPMIN
d=1.0/d; h=d
do m=1, MAXIT
  m2=2*m
  aa=m*(b-m)*x/((qam+m2)*(a+m2))
  d=1.0+aa*d
  if (abs(d) < FPMIN) d=FPMIN
  c=1.0+aa/c
  if (abs(c) < FPMIN) c=FPMIN
  d=1.0/d
  h=h*d*c
  aa=-(a+m)*(qab+m)*x/((qap+m2)*(a+m2))
  d=1.0+aa*d
  if (abs(d) < FPMIN) d=FPMIN
  c=1.0+aa/c
  if (abs(c) < FPMIN) c=FPMIN
  d=1.0/d
  del=d*c
  h=h*del
  if (abs(del-1.0) < EPS) goto 10
end do !m loop
10 if (m>MAXIT) then
  print *, ' a or b too big, or MAXIT too small.'
else
!   print *, ' Number of iterations: ', m
end if
BETACF=h
return
END

```

3. 혈연계수 계산 프로그램

가. 혈통 이용 혈연계수 계산 프로그램

혈연계수 계산

혈통, 유전체 정보를 이용하여 혈연계수 행렬 구성

혈통정보를 이용한 혈연계수

11193	11193	1.0024414063	223919342	223919342
11193	11199	0.8371993759	223919342	223936702
11193	11108	0.2512207031	223919342	224051383
11193	11206	0.0566406250	223919342	223936700
11193	11063	0.0354083906	223919342	223977275
11193	11042	0.0366821289	223919342	224037130
11193	11105	0.0488281250	223919342	224051499
11193	11136	0.0300292969	223919342	223916193
11193	11101	0.0263671875	223919342	224051497
11193	11153	0.0244140625	223919342	223896791
11193	11203	0.0420532227	223919342	223935986
11193	11190	0.8327148438	223919342	223941199
11193	8943	0.0473632813	223919342	223036890
11193	9142	0.0089765625	223919342	223141795
11193	9241	0.0357666616	223919342	223891962
11193	9161	0.025078125	223919342	223876501
11193	9913	0.0430988293	223919342	223011736
11193	9205	0.5024414063	223919342	223096626
11193	6007	0.0405273438	223919342	223198123
11193	9201	0.0507812500	223919342	223096628
11193	8535	0.1348876953	223919342	223198150
11193	9175	0.0439453125	223919342	223061773
11193	8527	0.0504150391	223919342	223198160
11193	11770	0.0261230469	223919342	224211962
11193	10829	0.0325622559	223919342	224125668
11193	11779	0.0279235840	223919342	224253999
11193	11749	0.0301818848	223919342	224242584
11193	11820	0.0279235840	223919342	224259290
11193	11732	0.0366210938	223919342	224239987
11193	10705	0.047448340	223919342	224171935

나. 유전체 정보 이용 혈연계수 계산 프로그램

혈연계수 계산

혈통, 유전체 정보를 이용하여 혈연계수 행렬 구성

유전체 정보를 이용한 혈연계수

1	1	0.9962559342	223919342	223919342
1	2	-0.0064443224	223919342	223936702
1	3	0.2312972546	223919342	224051383
1	4	0.0599610992	223919342	223936700
1	5	-0.0309693199	223919342	223977275
1	6	-0.0079109762	223919342	224037130
1	7	0.0043734135	223919342	224051499
1	8	-0.0141996955	223919342	223916193
1	9	-0.0106075155	223919342	224051497
1	10	-0.0090846767	223919342	223896791
1	11	-0.0161186811	223919342	223935986
1	12	-0.0187407937	223919342	223941199
1	13	0.0012445223	223919342	223036890
1	14	-0.0315277405	223919342	223141795
1	15	-0.0323233855	223919342	223891962
1	16	0.0386150815	223919342	223876501
1	17	-0.0090984311	223919342	223011736
1	18	0.4049960697	223919342	223096626
1	19	-0.0065277643	223919342	223198123
1	20	-0.0058182143	223919342	223096628
1	21	0.0971846953	223919342	223198150
1	22	-0.0058097515	223919342	223061773
1	23	-0.0333520100	223919342	223198160
1	24	-0.0216616699	223919342	224211962
1	25	0.0201679002	223919342	224125668
1	26	0.0355000058	223919342	224253999
1	27	0.0211225897	223919342	224242584
1	28	0.0075268778	223919342	224259290
1	29	0.0074934070	223919342	224239987
1	30	0.0161227075	223919342	224171935

다. 프로그램 코드

(1) 5종류의 유전체 관계행렬 구성 및 혈연계수 출력

- 유전체 정보를 이용한 유전체 관계행렬 구성
- 5종의 유전체 관계행렬 알고리즘을 적용한 유전체 관계행렬 구성
- 각 개체 조합별 유전체 혈연계수 출력 및 근교계수 통계량 출력

```
! Name : Making Genomic Relationship Matrix
! Developed by : Joonho Lee - PhD working in Hankyong National Univ.(Korea) zoonolee@gmail.com,
+82-10-3408-2895
! Version : Linux 4.0 (2012-06-15)
! Contents :
!           1. Ordering Animal information for Genomic Relationship Matrix
!           2. Making Genomic Relationship Matrix
!           3. Inversion of Genomic Relationship Matrix
!           4. Statistics of Relationship

program MGRM
integer::i,j,l,mk,an,ErrorF liostate & do_loop
integer::NOA=0,ANO=0,NOM=0,MNO=0 !number of animals and markers & animal and marekr number
integer,allocatable::MT(:,:) !data storage
real(10)::GF1=0,GF2=0,SEH=0,SG05=0,SGMF=0,AOMAF=0,SGF1,SSGF1,AGF1,VGF1,ALPHA,BETA,P0,Q0,KGOFS,
TRG05 !gnen frequencies & expected heterozygosity
real(10)::SOIBCG05=0,SSOIBCG05=0,AOIBCG05=0,SDOIBCG05=0,MINOIBCG05=0,MAXOIBCG05=0 !statistics for
population genetics
real(10)::SOIBCGMF=0,SSOIBCGMF=0,AOIBCGMF=0,SDOIBCGMF=0,MINOIBCGMF=0,MAXOIBCGMF=0
!statistics for population genetics
real(10)::SOIBCGOF=0,SSOIBCGOF=0,AOIBCGOF=0,SDOIBCGOF=0,MINOIBCGOF=0,MAXOIBCGOF=0 !statistics
for population genetics
real(10)::SOIBCGOFS=0,SSOIBCGOFS=0,AOIBCGOFS=0,SDOIBCGOFS=0,MINOIBCGOFS=0,MAXOIBCGOFS=0
!statistics for population genetics
real(10)::SOIBCGN=0,SSOIBCGN=0,AOIBCGN=0,SDOIBCGN=0,MINOIBCGN=0,MAXOIBCGN=0 !statistics for
population genetics
real,allocatable::G05(:,:),IG05(:,:),IBCG05(:,:),W05(:,:),WMF(:,:) !output data
real,allocatable::GMF(:,:),IGMF(:,:),IBCGMF(:)
real,allocatable::GOF(:,:),IGOF(:,:),IBCGOF(:)
real,allocatable::GOFS(:,:),IGOFS(:,:),IBCGOFS(:)
real,allocatable::GN(:,:),IGN(:,:),IBCGN(:)
real,allocatable::k(:),DII(:)
character(len=8)::OFT,NOAF
character(len=16)::WF1,WF2,WF3
character(len=300)::Dummy
character(len=220)::ADD,ODD,GDD
character(len=30)::GDFN1,GDFN2,ADFN
character(len=250)::ADFNIP,GDFNIP,ODFNIP,PF
character(len=20),allocatable::ID(:)
!reading parameters
print
*****
*****"
```

```

print *, "MGRM_Lv2 : Making Genomic Relationship Matrix & Basic Statistics Analysis Linux version 2.0"
print *, "Developed by Joonho Lee : zoonolee@gmail.com, +82-10-3408-2895"
print *, "Data form is the results of Post_fastPHASE and of converted genotype 0,1,2"
print *, "Please enter the name of parameter file including path(ex) ./parGRDH"
print
"*****"
*****"
read *, PF
PF=adjustl(PF)
open (99, file=PF)
read (99,*) OFT
print *, "file tag ", OFT
read (99,*) NOA
rewind(unit=99)
read (99,*) OFT
read (99,*) NOAF
print *, "number of animals", NOA
read (99,*) NOM
print *, "number of markers", NOM
read (99,*) GDD
print *, "directory of Post_Imputation output file ", GDD
read (99,*) GDFN1
print *, "name of allelic coding data file(example : Genotype012_OUT_TAG.out) ", GDFN1
read (99,*) GDFN2
print *, "name of allele frequency data file(example : Genefrequency_OUT_TAG.out) ", GDFN2
read (99,*) ADD
print *, "directory of Selected animals file ", ADD
read (99,*) ADFN
print *, "name of Selected animals file(example : Selected_Animals_OUT_TAG.out) ", ADFN
read (99,*) ODD
print *, "directory of output file ", ODD
close(99)
allocate(ID(NOAA))
allocate(MT(NOM,NOAA))
allocate(G05(NOAA,NOAA))
allocate(IG05(NOAA,NOAA))
allocate(IBC05(NOAA))
allocate(W05(3,NOM))
allocate(k(NOAA))
allocate(GMF(NOAA,NOAA))
allocate(IGMF(NOAA,NOAA))
allocate(IBC05GMF(NOAA))
allocate(WMF(3,NOM))
allocate(GOF(NOAA,NOAA))
allocate(IGOF(NOAA,NOAA))
allocate(IBC05GOF(NOAA))
allocate(DII(NOM))
allocate(GOFS(NOAA,NOAA))

```

```

allocate(IGOFS(NO,NOA))
allocate(IBCGOFS(NO,NOA))
allocate(GN(NO,NOA))
allocate(IGN(NO,NOA))
allocate(IBCIGN(NO,NOA))
G05=0;IG05=0
GMF=0;IGMF=0
GOF=0;IGOF=0;
GOFS=0;IGOFS=0;
GN=0;IGN=0
WF1='          ';WF2='          ';WF3='          ';
write(WF1,fmt='(a16)') '(/adjustr(NOAF)//f15.11)'
write(WF2,fmt='(a16)') '(/adjustr(NOAF)//f22.15)'
write(WF3,fmt='(a16)') '(/adjustr(NOAF)//i1      )'
print *, WF1,WF2,WF3
ODD=adjustr(ODD)
GDD=adjustr(GDD)
ADD=adjustr(ADD)
GDFNIP=GDD//GDFN1
GDFNIP=adjustl(GDFNIP)
open(98,file=GDFNIP,status='old')
GDFNIP=GDD//GDFN2
GDFNIP=adjustl(GDFNIP)
open(97,file=GDFNIP,status='old')
ADFNIP=ADD//ADFN
ADFNIP=adjustl(ADFNIP)
open(96,file=ADFNIP,status='old')
ODFNIP=ODD//G05//OFT//".out"
ODFNIP=adjustl(ODFNIP)
open(95,file=ODFNIP,status='unknown')
ODFNIP=ODD//IG05//OFT//".out"
ODFNIP=adjustl(ODFNIP)
open(94,file=ODFNIP,status='unknown')
ODFNIP=ODD//IBC05//OFT//".out"
ODFNIP=adjustl(ODFNIP)
open(93,file=ODFNIP,status='unknown')
write(93,fmt='(2a20)') "Animal ID","Inbreeding coeff"
ODFNIP=ODD//G05RC_selected//OFT//".out"
ODFNIP=adjustl(ODFNIP)
open(92,file=ODFNIP,status='unknown')
ODFNIP=ODD//GMF//OFT//".out"
ODFNIP=adjustl(ODFNIP)
open(91,file=ODFNIP,status='unknown')
ODFNIP=ODD//IGMF//OFT//".out"
ODFNIP=adjustl(ODFNIP)
open(90,file=ODFNIP,status='unknown')
ODFNIP=ODD//IBC05GMF//OFT//".out"
ODFNIP=adjustl(ODFNIP)

```

```

open(89,file=ODFNIP,status='unknown')
write(89,fmt='(2a20)') "Animal ID","Inbreeding coeff"
ODFNIP=ODD/"GMFRC_selected"/OFT/"out"
ODFNIP=adjustl(ODFNIP)
open(88,file=ODFNIP,status='unknown')
ODFNIP=ODD/"GOF"/OFT/"out"
ODFNIP=adjustl(ODFNIP)
open(87,file=ODFNIP,status='unknown')
ODFNIP=ODD/"IGOF"/OFT/"out"
ODFNIP=adjustl(ODFNIP)
open(86,file=ODFNIP,status='unknown')
ODFNIP=ODD/"IBCGOF"/OFT/"out"
ODFNIP=adjustl(ODFNIP)
open(85,file=ODFNIP,status='unknown')
write(85,fmt='(2a20)') "Animal ID","Inbreeding coeff"
ODFNIP=ODD/"GOFRC_selected"/OFT/"out"
ODFNIP=adjustl(ODFNIP)
open(84,file=ODFNIP,status='unknown')
ODFNIP=ODD/"GOFs"/OFT/"out"
ODFNIP=adjustl(ODFNIP)
open(83,file=ODFNIP,status='unknown')
ODFNIP=ODD/"IGOFs"/OFT/"out"
ODFNIP=adjustl(ODFNIP)
open(82,file=ODFNIP,status='unknown')
ODFNIP=ODD/"IBCGOFs"/OFT/"out"
ODFNIP=adjustl(ODFNIP)
open(81,file=ODFNIP,status='unknown')
write(81,fmt='(2a20)') "Animal ID","Inbreeding coeff"
ODFNIP=ODD/"GOFsRC_selected"/OFT/"out"
ODFNIP=adjustl(ODFNIP)
open(80,file=ODFNIP,status='unknown')
ODFNIP=ODD/"GN"/OFT/"out"
ODFNIP=adjustl(ODFNIP)
open(79,file=ODFNIP,status='unknown')
ODFNIP=ODD/"IGN"/OFT/"out"
ODFNIP=adjustl(ODFNIP)
open(78,file=ODFNIP,status='unknown')
ODFNIP=ODD/"IBCGN"/OFT/"out"
ODFNIP=adjustl(ODFNIP)
open(77,file=ODFNIP,status='unknown')
write(77,fmt='(2a20)') "Animal ID","Inbreeding coeff"
ODFNIP=ODD/"GNRC_selected"/OFT/"out"
ODFNIP=adjustl(ODFNIP)
open(76,file=ODFNIP,status='unknown')
read(97,*) GF1,GF2
SGF1=0;SSGF1=0
do mk=1, NOM
  read(98,fmt=WF3) MT(mk,:)

```



```

read(97,*) GF1,GF2
AOMAF=AOMAF+GF1
SEH=SEH+(GF1*GF2*2)
DII(mk)=1.0/(NOM*GF1*GF2*2)
W05(1,mk)=GF2*(-2)
W05(2,mk)=1.0-GF2*2
W05(3,mk)=2.0-GF2*2
SGF1=SGF1+GF1
! SGF2=SGF2+GF2
SSGF1=SSGF1+GF1**2
! SSGF2=SSGF2+GF2**2
enddo
AGF1=SGF1/NOM
!AGF2=SGF2/NOM
VGF1=(SSGF1-(SGF1**2/NOM))/NOM
!VGF2=(SSGF2-(SGF2**2/NOM))/NOM
ALPHA=AGF1*(AGF1*(1-AGF1)/VGF1-1)
BETA=(1-AGF1)*(AGF1*(1-AGF1)/VGF1-1)
AOMAF=AOMAF/NOM
Q0=ALPHA/(ALPHA+BETA)
P0=BETA/(ALPHA+BETA)
KGOFS=NOM*((P0-Q0)**2+(SEH/NOM)*((ALPHA+BETA+2)/(ALPHA+BETA)))
close(98)
close(97)
k=SEH
do mk=1, NOM
  WMF(1,mk)=AOMAF*2-GF1*2-1.0
  WMF(2,mk)=AOMAF*2-GF1*2
  WMF(3,mk)=AOMAF*2-GF1*2+1.0
enddo
print *, "reading genotype is finished"
TRG05=0
read(96,*)
do i=1,NOA
  read(96,*) ID(i)
  do j=1,NOA
    SG05=0
    SGMF=0
    SGOF=0
    do l=1,NOM
      SG05=SG05+W05(MT(l,i)+1,l)*W05(MT(l,j)+1,l)
      SGMF=SGMF+WMF(MT(l,i)+1,l)*WMF(MT(l,j)+1,l)
      SGOF=SGOF+W05(MT(l,i)+1,l)*W05(MT(l,j)+1,l)*DII(l)
    enddo
    IF(i.EQ.j) TRG05=TRG05+SG05
    G05(i,j)=SG05/sqrt(k(i)*k(j))
    G05(j,i)=G05(i,j)
    GMF(i,j)=SGMF/sqrt(k(i)*k(j))
  enddo
enddo

```

```

GMF(j,i)=GMF(i,j)
GOF(i,j)=SGOF
GOF(j,i)=GOF(i,j)
GOF5(i,j)=SG05/KGOF5
GOF5(j,i)=GOF5(i,j)
GN(i,j)=SG05
GN(j,i)=GN(i,j)
enddo
enddo
GN=GN/(TRG05/NOA)
close(96)
print *, "making genomic relationship matrix is finished "
CALL FindInv(G05, IG05, NOA, ErrorF)
if(ErrorF.lt.0) print *, "error status during inversion"
print *, "making inverse genomic relationship matrix is finished -G05"
CALL FindInv(GMF, IGMF, NOA, ErrorF)
if(ErrorF.lt.0) print *, "error status during inversion"
print *, "making inverse genomic relationship matrix is finished -GMF"
CALL FindInv(GOF, IGOF, NOA, ErrorF)
if(ErrorF.lt.0) print *, "error status during inversion"
print *, "making inverse genomic relationship matrix is finished -GOF"
CALL FindInv(GOFS, IGOFS, NOA, ErrorF)
if(ErrorF.lt.0) print *, "error status during inversion"
print *, "making inverse genomic relationship matrix is finished -GOFS"
CALL FindInv(GN, IGN, NOA, ErrorF)
if(ErrorF.lt.0) print *, "error status during inversion"
print *, "making inverse genomic relationship matrix is finished -GN"
do j=1, NOA
  IBCG05(j)=G05(j,j)-1.0
  SSOIBCG05=SSOIBCG05+IBCG05(j)**2
  write(93,fmt='(a20,f20.10)') ID(j),IBCG05(j)
  write(95,fmt=WF1) G05(j,:)
  write(94,fmt=WF2) IG05(j,:)
  do i=1,NOA
    if(i.ge.j) write(92,fmt='(2i10,f15.10,2a22)') j,i,G05(j,i),ID(j),ID(i)
  enddo
enddo
AOIBCG05=sum(IBCG05)/NOA;SDIBCG05=(SSOIBCG05-(sum(IBCG05)**2/NOA))/NOA
MINOIBCG05=minval(IBCG05);MAXOIBCG05=maxval(IBCG05)
write(93,fmt='(4a20)') "AOIBC","SDIBC","MINOIBC","MAXOIBC"
write(93,fmt='(4f20.10)') AOIBCG05,SDIBCG05,MINOIBCG05,MAXOIBCG05
do j=1, NOA
  IBCGMF(j)=GMF(j,j)-1.0
  SSOIBCGMF=SSOIBCGMF+IBCGMF(j)**2
  write(89,fmt='(a20,f20.10)') ID(j),IBCGMF(j)
  write(91,fmt=WF1) GMF(j,:)
  write(90,fmt=WF2) IGMF(j,:)
  do i=1,NOA

```

```

    if(i.ge.j) write(88,fmt='(2i10,f15.10,2a22)') j,i,GMF(j,i),ID(j),ID(i)
  enddo
enddo
AOIBCGMF=sum(IBCGMF)/NOA;SDIBCGMF=(SSOIBCGMF-(sum(IBCGMF)**2/NOA))/NOA
MINOIBCGMF=minval(IBCGMF);MAXOIBCGMF=maxval(IBCGMF)
write(89,fmt='(4a20)') "AOIBC","SDIBC","MINOIBC","MAXOIBC"
write(89,fmt='(4f20.10)') AOIBCGMF,SDIBCGMF,MINOIBCGMF,MAXOIBCGMF
do j=1, NOA
  IBCGOF(j)=GOF(j,j)-1.0
  SSOIBCGOF=SSOIBCGOF+IBCGOF(j)**2
  write(85,fmt='(a20,f20.10)') ID(j),IBCGOF(j)
  write(87,fmt=WF1) GOF(j,:)
  write(86,fmt=WF2) IGOF(j,:)
do i=1,NOA
  if(i.ge.j) write(84,fmt='(2i10,f15.10,2a22)') j,i,GOF(j,i),ID(j),ID(i)
  enddo
enddo
AOIBCGOF=sum(IBCGOF)/NOA;SDIBCGOF=(SSOIBCGOF-(sum(IBCGOF)**2/NOA))/NOA
MINOIBCGOF=minval(IBCGOF);MAXOIBC=maxval(IBCGOF)
write(85,fmt='(4a20)') "AOIBC","SDIBC","MINOIBC","MAXOIBC"
write(85,fmt='(4f20.10)') AOIBCGOF,SDIBCGOF,MINOIBCGOF,MAXOIBCGOF
do j=1, NOA
  IBCGOFs(j)=GOFs(j,j)-1.0
  SSOIBCGOFs=SSOIBCGOFs+IBCGOFS(j)**2
  write(81,fmt='(a20,f20.10)') ID(j),IBCGOFS(j)
  write(83,fmt=WF1) GOFs(j,:)
  write(82,fmt=WF2) IGOFS(j,:)
do i=1,NOA
  if(i.ge.j) write(80,fmt='(2i10,f15.10,2a22)') j,i,GOFs(j,i),ID(j),ID(i)
  enddo
enddo
AOIBCGOFs=sum(IBCGOFS)/NOA;SDIBCGOFs=(SSOIBCGOFs-(sum(IBCGOFS)**2/NOA))/NOA
MINOIBCGOFs=minval(IBCGOFS);MAXOIBCGOFs=maxval(IBCGOFS)
write(81,fmt='(4a20)') "AOIBC","SDIBC","MINOIBC","MAXOIBC"
write(81,fmt='(4f20.10)') AOIBCGOFs,SDIBCGOFs,MINOIBCGOFs,MAXOIBCGOFs
do j=1, NOA
  IBCGN(j)=GN(j,j)-1.0
  SSOIBCGN=SSOIBCGN+IBCGN(j)**2
  write(77,fmt='(a20,f20.10)') ID(j),IBCGN(j)
  write(79,fmt=WF1) GN(j,:)
  write(78,fmt=WF2) IGN(j,:)
do i=1,NOA
  if(i.ge.j) write(76,fmt='(2i10,f15.10,2a22)') j,i,GN(j,i),ID(j),ID(i)
  enddo
enddo
AOIBCGN=sum(IBCGN)/NOA;SDIBCGN=(SSOIBCGN-(sum(IBCGN)**2/NOA))/NOA
MINOIBCGN=minval(IBCGN);MAXOIBCGN=maxval(IBCGN)
write(77,fmt='(4a20)') "AOIBC","SDIBC","MINOIBC","MAXOIBC"

```

```

write(77,fmt='(4f20.10)') AOIBCGN,SDIBCGN,MINOIBCGN,MAXOIBCGN
print *, "estimating inbreeding coefficient is finished "
end program MGRM
!Subroutine comes here
!Subroutine to find the inverse of a square matrix
!Author : Louisda16th a.k.a Ashwith J. Rego
!Reference : Algorithm has been well explained in:
!http://math.uww.edu/~mcfarlat/inverse.htm
!http://www.tutor.ms.unimelb.edu.au/matrix/matrix\_inverse.html
SUBROUTINE FINDInv(matrix, inverse, n, errorflag)
    IMPLICIT NONE
    !Declarations
    INTEGER, INTENT(IN) :: n
    INTEGER, INTENT(OUT) :: errorflag !Return error status. -1 for error, 0 for normal
    REAL, INTENT(IN), DIMENSION(n,n) :: matrix !Input matrix
    REAL, INTENT(OUT), DIMENSION(n,n) :: inverse !Inverted matrix
    LOGICAL :: FLAG = .TRUE.
    INTEGER :: i, j, k, l
    REAL :: m
    REAL, DIMENSION(n,2*n) :: augmatrix !augmented matrix
    !Augment input matrix with an identity matrix
    DO i = 1, n
        DO j = 1, 2*n
            IF (j <= n ) THEN
                augmatrix(i,j) = matrix(i,j)
            ELSE IF ((i+n) == j) THEN
                augmatrix(i,j) = 1
            Else
                augmatrix(i,j) = 0
            ENDIF
        END DO
    END DO
    !Reduce augmented matrix to upper traingular form
    DO k =1, n-1
        IF (augmatrix(k,k) == 0) THEN
            FLAG = .FALSE.
            DO i = k+1, n
                IF (augmatrix(i,k) /= 0) THEN
                    DO j = 1,2*n
                        augmatrix(k,j) = augmatrix(k,j)+augmatrix(i,j)
                    END DO
                    FLAG = .TRUE.
                    EXIT
                ENDIF
            END IF
            IF (FLAG .EQV. .FALSE.) THEN
                PRINT*, "Matrix is non - invertible"
                inverse = 0
                errorflag = -1
            END IF
        END IF
    END DO

```

```

                                return
                                ENDIF
                                END DO
                                ENDIF
                                DO j = k+1, n
                                    m = augmatrix(j,k)/augmatrix(k,k)
                                    DO i = k, 2*n
                                        augmatrix(j,i) = augmatrix(j,i) - m*augmatrix(k,i)
                                    END DO
                                END DO
                                END DO
                                END DO
                                !Test for invertibility
                                DO i = 1, n
                                    IF (augmatrix(i,i) == 0) THEN
                                        PRINT*, "Matrix is non - invertible"
                                        inverse = 0
                                        errorflag = -1
                                        return
                                    ENDIF
                                END DO
                                !Make diagonal elements as 1
                                DO i = 1, n
                                    m = augmatrix(i,i)
                                    DO j = i, (2 * n)
                                        augmatrix(i,j) = (augmatrix(i,j) / m)
                                    END DO
                                END DO
                                !Reduced right side half of augmented matrix to identity matrix
                                DO k = n-1, 1, -1
                                    DO i = 1, k
                                        m = augmatrix(i,k+1)
                                        DO j = k, (2*n)
                                            augmatrix(i,j) = augmatrix(i,j) -augmatrix(k+1,j) * m
                                        END DO
                                    END DO
                                END DO
                                !store answer
                                DO i = 1, n
                                    DO j = 1, n
                                        inverse(i,j) = augmatrix(i,j+n)
                                    END DO
                                END DO
                                errorflag = 0
                                END SUBROUTINE FINDinv

```

(2) 한우의 유전체 정보 활용 유전체 혈연관계행렬 추정프로그램

- 대용량의 SNP 표지인자 정보를 이용하여 개체 간 혈연적 유사도를 추정가능
- 개체 간 추정된 유전체 혈연계수를 이용하여 한우 친자판정에 활용 가능

```
program G_matrix

use typedef

implicit none

integer :: i, j, k, io, ii, jj, NUL
integer :: chro, nmk, tnmk, n_nmk
integer, parameter :: nid=1232, nchro=18
character(len=128) :: filename1, filename2
character(len=10000) :: ex
integer :: unitF
character(len=1)::BASE(4)=('A','T','G','C')
type GENOV
  character(1) :: INFO0
  character(11):: INFO1
  character(len=1) :: HAP(2,nid)
  integer :: GENOT(nid)
  integer :: temp(4)
  real :: FREQ(4)
  character(len=1) :: MaMi(2)
end type GENOV
type(GENOV),allocatable:: GENO(:)
integer, allocatable :: GENOTYPE(,:;)
real, allocatable :: g_matrix(,:;)

do chro = 1, nchro
  filename1=iofname3('gzip -d exam.Mark01.phased.gz',chro)
  CALL system(filename1)
  filename2 = filename1(9:26)
  nmk=N_recf(filename2)
  nmk = nmk-1
  tnmk = nmk + tnmk
enddo

allocate(g_matrix(nid,nid), GENOTYPE(nid,tnmk))

tnmk = 0; n_nmk = 0; g_matrix = 0.0;
GENOTYPE = 0;
do chro=1, nchro
  filename1 = iofname3('gzip -d exam.Mark01.phased.gz',chro)
  filename2 = filename1(9:26)
  nmk=N_recf(filename2)
  nmk = nmk-1
  unitF=fopen(filename2)
  allocate(GENO(nmk))
```

```

read(unitF, *)
print *, "chromosome=", chro, "nmk=", nmk

do i=1, nmk
  read(unitF,*iostat=io) GENO(i)%INFO0, GENO(i)%INFO1, &
    ((GENO(i)%HAP(j, k),j=1,2),k=1,nid)
  if(io .ne. 0) exit
enddo

do i = 1, nmk
  GENO(i)%temp = 0;
  do j= 1, nid
    do k = 1, 2
      select case(GENO(i)%HAP(k,j))
        case("A")
          GENO(i)%temp(1) = GENO(i)%temp(1) + 1
        case("T")
          GENO(i)%temp(2) = GENO(i)%temp(2) + 1
        case("G")
          GENO(i)%temp(3) = GENO(i)%temp(3) + 1
        case("C")
          GENO(i)%temp(4) = GENO(i)%temp(4) + 1
      end select
    enddo
  enddo
  NUL = 0;
  do jj = 1, 4
    GENO(i)%FREQ(jj) = real(GENO(i)%temp(jj)) / real(nid * 2.0)

    if(GENO(i)%FREQ(jj) > 0.5) then
      GENO(i)%MaMi(1) = BASE(jj)
    else if(GENO(i)%FREQ(jj) .eq. 0.5) then
      NUL = NUL + 1
      GENO(i)%MaMi(NUL) = BASE(jj)
    else if(GENO(i)%FREQ(jj) .ne. 0.0) then
      GENO(i)%MaMi(2) = BASE(jj)
    endif
  enddo

enddo

do i = 1, nmk
  n_nmk = n_nmk + 1
  do j = 1, nid
    if(GENO(i)%HAP(1,j) .ne. GENO(i)%HAP(2,j)) then
      GENO(i)%GENOT(j) = 1
    else if(GENO(i)%HAP(1,j) .eq. GENO(i)%HAP(2,j)) then
      if(GENO(i)%HAP(1,j) .eq. GENO(i)%MaMi(2)) then
        GENO(i)%GENOT(j) = 0
      else if(GENO(i)%HAP(1,j) .eq. GENO(i)%MaMi(1)) then
        GENO(i)%GENOT(j) = 2
      endif
    endif
  enddo
enddo

```

```

        else
            GENO(i)%GENOT(j) = 5 ! This is error code
        endif
    endif
    GENOTYPE(j,n_nmk) = GENO(i)%GENOT(j)
enddo
enddo

deallocate(GENO)
close(unitF)

enddo

open(50, file='./genotype.out')

do i = 1, nid
    write(50, fmt='(50000(i2))') GENOTYPE(i, :)
enddo

g_matrix=GOF(GENOTYPE)

open(51, file='./GOF.out', status='unknown')

do i = 1, nid
    do j = i, nid
        write(51, fmt='(2i5, 50000f8.4)') i, j, g_matrix(i,j) ! row, col, value
    enddo
enddo

contains

function GOF(GENOTYPE) result(g_matrix)

integer :: i, j, nid, nmk
integer :: GENOTYPE(:, :)
real :: pq2
real, allocatable :: p_matrix(:, :), pi(:), p(:), z_matrix(:, :)
real, allocatable :: tz_matrix(:, :), ztz_matrix(:, :), g_matrix(:, :)

nid = size(GENOTYPE,1)
nmk = size(GENOTYPE,2)

allocate(p_matrix(nid,nmk), pi(nmk), p(nmk), z_matrix(nid,nmk))
allocate(tz_matrix(nmk,nid), ztz_matrix(nid,nid), g_matrix(nid,nid))

pi = 0.0; p = 0.0; p_matrix = 0.0; z_matrix = 0.0;
pq2 = 0.0; tz_matrix = 0.0; ztz_matrix = 0.0;

GENOTYPE = GENOTYPE - 1
do i=1, nmk
    pi(i) = real(sum(GENOTYPE(:,i)) + nid) / real(nid*2)

```



```

enddo

do i=1, nmk
  p(i)=(pi(i) - 0.5)*2.0
  p_matrix(:,i) = p(i)
enddo

z_matrix=GENOTYPE-p_matrix

do i=1, nmk
  pq2 = (2.0 * pi(i) * (1.0-pi(i))) + pq2
enddo

tz_matrix = transpose_m(nid,nmk,z_matrix)
ztz_matrix=matmul(z_matrix,tz_matrix)

g_matrix = (ztz_matrix / pq2)

end function

function iofname3(nfile,iv) result(fname)
  character(len=29)::nfile
  integer::iv,i1,i2,iz
  character(len=29)::fname
  character(len=2)::ch1
  i1=0; i2=0; iz=0;
  iz=ichar('0')
  i1=floor(iv/10.)+iz
  i2=mod(iv,10)+iz
  ch1(1:1)=achar(i1)
  ch1(2:2)=achar(i2)
  fname(1:17)=nfile(1:17)
  fname(18:19)=ch1(1:2)
  fname(20:29)=nfile(20:29)
  fname=trim(fname)
end function

function transpose_m(nrow, ncol, matrix) result(t_matrix)
  integer :: n, i, nrow, ncol
  real :: matrix(nrow, ncol), t_matrix(ncol, nrow)
  t_matrix=0.0;
  do i=1, nrow
    t_matrix(:,i) = matrix(i,:)
  enddo
end function

end program

```

4. 유전체 육종가 추정 및 비교 프로그램

가. 시뮬레이션 프로그램

```
program simgen4

! This program is to simulate haplotype data by generation and
! calculate linkage disequilibrium by map distance and generation
! programed by Deukhwan Lee(May 1, 2009)

use prob; use kinds
implicit none
real,parameter::tolerance=1.0e-8,lengthU=1000,infinity=lengthU*1000
integer,parameter::slength=9
integer,allocatable::PNTHAP(:,:),PROHAP(:,:),PED(:,:),popgen(:)
integer,allocatable::ChromL(:),nmarker(:),nqtl(:),nloci(:),LOCATION(:),LOCATE(:),genotype(:)
integer::PNT(2),RY(5),cross(100)
real,allocatable::EFFECT(:),PNTQTL(:,:),PROQTL(:,:),ALFREQ(:),TALFREQ(:),Chr(:),BV(:)
real(r8),allocatable:: varG(:,:),varE(:)
integer::popsize,nbaseP,ngen,startgen,lastgen,igen,anim,nanim,maxchrL,nsir,ndam,nupro,tanimal
integer::maxmarker,maxqtl,maxloci,mdist,scatter1,scatter2,wild,nchr,ichr,fromgen,togen,maxM,maxQ
integer::pnsir,pndam,pnupro,egen
integer::locus1,locus2,next,id,io,ir,ic,icount,gamete,locus,mlocus,which,ncross,LDopt
logical,allocatable::CHKMKR(:),CHKQTL(:)
real::rand,mutateM,mutateQ,sumco,Mdis,Qdis,val,heri,phe,bvt,r2
real(r8)::scal,shap
integer ::kkk
!integer::startld, interld
character(len=50)::parfile
character(len=slength)::genfile

write(*,'(a)',advance='no')' name of parameter file?'
read '(a)',parfile
parfile=adjustl(parfile)
open(66,file=parfile)

read(66,*)
read(66,*)nbaseP
read(66,*)popsize
read(66,*)nsir,ndam,nupro
read(66,*)pnsir,pndam,pnupro
read(66,*)ngen
read(66,*)egen
allocate(popgen(0:ngen))
read(66,*)Mdis,Qdis      ! 1cM distance(1,000,000bp) unit
read(66,*)scatter1,scatter2 ! 1cM distance(1,000,000bp) unit
read(66,*)scal,shap     ! 1cM distance(1,000,000bp) unit
read(66,*)maxM,maxQ     ! maximum number of alleles on Marker and QTL locus
```

```

read(66,*)mutateM,mutateQ
mutateQ=mutateQ*Qdis/Mdis
read(66,*)fromgen,togen          ! mutation time(from and to)
read(66,*)LDopt                  ! LD calculate(1=calculate, others= no calculate)
read(66,*)startgen,lastgen
!read(66,*)startld, interld
read(66,*)heri
read(66,*)nchr                   ! number of chromosomes
allocate(Chr(nchr),ChromL(nchr),nmarker(nchr),nqtl(nchr),nloci(nchr))
do ichr=1,nchr
  read(66,*)Chr(ichr)            ! 1cM distance(1,000,000bp) unit
enddo
do ichr=1,nchr
  ChromL(ichr)=Chr(ichr)*lengthU
  nmarker(ichr)=Chr(ichr)/Mdis
  nqtl(ichr)=Chr(ichr)/Qdis
  nloci(ichr)=nmarker(ichr)+nqtl(ichr)
enddo
maxchrL=maxval(ChromL)
maxmarker=maxval(nmarker)
maxqtl=maxval(nqtl)
maxloci=maxval(nloci)

allocate(LOCATE(maxchrL),LOCATION(maxloci),CHKMKR(maxloci),CHKQTL(maxloci),EFFECT(maxloci),&
        ALFREQ(maxloci),TALFREQ(maxloci),genotype(maxmarker))

call initseed()

call genped(popgen)
print*, 'finished generating pedigree, generation & total animal=',ngen,popgen(ngen)

tanimal=popgen(ngen)
allocate(BV(tanimal,nchr),varG(0:ngen,nchr),varE(0:ngen))
BV=0.0; varG=0.d0; varE=0.d0

open(19, file="./locus_var",status="unknown")          !!!!!!!!!!!!!!!!!!!!!!!
open(21, file="./QTL_effect",status="unknown")        !!!!!!!!!!!!!!!!!!!!!!!
open(29, file="./tlocus_var",status="unknown")        !!!!!!!!!!!!!!!!!!!!!!!
do ichr=1,nchr
  call set_seed(ChromL(ichr))
  open(1,file='./simped.dat',status='unknown')
  genfile=iofname('HAPLOTYPE',ichr)
  open(2,file=genfile,status='unknown',recl=5000)
  genfile=iofname('GENOTYPE',ichr)
  open(3,file=genfile,status='unknown',recl=5000)
  genfile=iofname('LDCALC',ichr)
  open(4,file=genfile,status='unknown')
  call posit(ChromL(ichr),nmarker(ic

```

나. 베이저안을 이용한 SNP효과 추정 프로그램

```
! Bayesian program
! programed by Chungil CHO

program Bayesian
  use prob; use kinds
  implicit none

integer, parameter :: no_chro=29, numiter=10000
integer :: allele(10000)
integer, allocatable :: x(:,:)
real(kind=8), allocatable :: y(:), truea(:)
character(10) :: filename0
integer :: i, j, k, s, io
integer :: nmarker, nmk, nid
character(5000) :: aaa

real(kind=8) :: mean2pq, vara, scalea, xb
real(kind=8), allocatable :: b(:), meanb(:), var(:), ycorr(:)

integer :: iter
real (kind=8) :: vare, rhs
real (r8):: v, invlhs, mean
real (kind=8), allocatable :: xb1(:)
integer :: locus
real (kind=8):: txy, txx, lhs
real (kind=8)::summ, accuracy
real (kind=8), allocatable :: ahat(:)
call random_seed
call initseed()

nid=0
open(2, file='./pheno')
do
  read(2, *, iostat=io)
  if(io .ne. 0) exit
  nid=nid+1
enddo
print *, nid
close(2)

nmk=0;
do s=1, no_chro
  filename0=iofname1('geno',s)
  open(1, file=filename0)
  read(1, *) aaa
  nmarker=len_trim(aaa)
```

```

nmk=nmarker+nmk
aaa=' '
print *, 'chromosome=', s, 'nmk=', nmarker
close(1)
enddo
print *, 'nid=', nid, 'nmk=', nmk
nmk=nmk+1
allocate(x(nid, nmk), y(nid), truea(nid))
open(2, file='./pheno')
do i=1, nid
  read(2, *) y(i)
enddo
x=1
do s=1, no_chro
  filename0=iofname1('geno',s)
  open(1, file=filename0, iostat=io)
  do i=1, nid
    read(1, fmt='(10000i1)',iostat=io) allele
    x(i, (nmarker*(s-1)+2):(nmarker*(s-1)+nmarker+1))= allele
  enddo
  close(1)
enddo
mean2pq=0.5; vara= 1.0/20.0;
scalea =0.5*vara / (nmk*mean2pq)
allocate(b(nmk), meanb(nmk), var(nmk), ycorr(nid))
b=0.0; meanb=0.0; var=0.0; ycorr=0.0
b(1)=sum(y)/real(nid)
do i=1, nid
  xb=0.0
  do j=1, nmk
    xb=x(i,j)*b(j)+xb
  enddo
  ycorr(i)=y(i)-xb
enddo
allocate(xb1(nid), ahat(nid))
vare=0.0; v=1.0
do iter=1, numiter
  do i=1, nid
    vare=(ycorr(i)*ycorr(i))+vare
  enddo
  vare=vare/gen_chisq(v,(nid+3))
  do i=1, nid
    xb1(i)=x(i,1)*b(1)
  enddo
  do i=1, nid
    ycorr(i)=ycorr(i)+xb1(i)
  enddo
  rhs=sum(ycorr)/real(vare)

```

```

invlhs=1.0/(nid/vare)
mean=rhs*invlhs
b(1)=gen_normal_scalar(mean, invlhs)
xb1=0.0
do i=1, nid
  xb1(i)=x(i,1)*b(1)
enddo
do i=1, nid
  ycorr(i)=ycorr(i)-xb1(i)
enddo
meanb(1)=meanb(1)+b(1)
do locus=2, nmk
  var(locus)=(scalea*4.0+b(locus)*b(locus)) / gen_chisq(v,5) ! 4+1
enddo
do locus=2, nmk
  xb1=0.0
  do i=1, nid
    xb1(i)=x(i,locus)*b(locus)
  enddo
  do i=1, nid
    ycorr(i)=ycorr(i)+xb1(i)
  enddo

  txy=0.0
  do i=1, nid
    txy=x(i, locus)*ycorr(i)+txy
  enddo
  rhs=txy / vare

  txx=0.0
  do i=1, nid
    txx=x(i, locus)*x(i,locus)+txx
  enddo
  lhs=(txx / vare) + (1.0 / var(locus))

  invlhs=1.0 / lhs
  mean=invlhs*rhs
  b(locus)= gen_normal_scalar(mean, invlhs)
  xb1=0.0
  do i=1, nid
    xb1(i)=x(i,locus)*b(locus)
  enddo
  do i=1, nid
    ycorr(i)=ycorr(i)-xb1(i)
  enddo
  meanb(locus) = meanb(locus) +b(locus)
enddo
if(mod(iter,100) .eq. 0) then

```

```

                print *, "iteration=", iter
                print *, "b(locus)= ", b(1:3)
            endif
        enddo

do i=1, nmk
    meanb(i) = meanb(i) / real(numiter)
enddo

do i=1, nid
    summ=0.0
    do j=1, nmk
        summ=x(i,j)*meanb(j)+summ
    enddo
    ahat(i)=summ
enddo

open(3, file='./truea', iostat=io)
do i=1, nid
    read(3, *) truea(i)
enddo

accuracy=corr(truea, ahat)
print *, "corr=", accuracy

open (51, file='./ahat.out')
open (52, file='./meanb.out')

do i=1, nid
    write (51, fmt='(f10.6)') ahat(i)
enddo
do i=1, nmk
    write (52, fmt='(f10.6)') meanb(i)
enddo

contains

function iofname1(nfile,iv) result(fname)
    character(len=*)::nfile
    integer::iv,i1,i2,iz
    character(len=6)::fname
    character(len=2)::ch1
    i1=0; i2=0; iz=0;
    iz=ichar('0')
    i1=floor(iv/10.)+iz

```

```

i2=mod(iv,10)+iz
ch1(1:1)=achar(i1)
ch1(2:2)=achar(i2)
fname(1:4)=nfile(1:4)
fname(5:6)=ch1(1:2)
fname=trim(fname)
end function

function corr(truea,ahat) result(accuracy)
real(kind=8)::truea(:),ahat(:),accuracy
real(kind=8)::sm(2),ss(2,2)
integer::nr,ir
nr=size(truea,1)
sm=0.0; ss=0.0; accuracy=0.0
do ir=1,nr
    sm(1)=sm(1)+truea(ir)
    sm(2)=sm(2)+ahat(ir)
    ss(1,1)=ss(1,1)+truea(ir)*truea(ir)
    ss(1,2)=ss(1,2)+truea(ir)*ahat(ir)
    ss(2,2)=ss(2,2)+ahat(ir)*ahat(ir)
enddo
accuracy=(ss(1,2)-sm(1)*sm(2)/nr)/sqrt((ss(1,1)-sm(1)*sm(1)/nr)*(ss(2,2)-sm(2)*sm(2)/nr))
end function

function average(x) result(z)
real::x(:),z,sm
integer::nr,ir
nr=size(x,1)
sm=0.0; z=0.0
do ir=1,nr
    sm=sm+x(ir)
enddo
z=sm/nr
end function

end program

```


다. BLUP 및 GBLUP 육종가 추정 및 비교 프로그램(R)

```
! BLUP and GBLUP
! programed by Chungil CHO

library("MASS")
desgn <- function(v) {
  if(is.numeric(v)){
    va=v
    mrow = length(va)
    mcol = max(va)
  }
  if(is.character(v)){
    vf = factor(v)
    va = as.numeric(vf)
    mrow = length(va)
    mcol = length(levels(vf))
  }
  X = matrix(data=c(0), nrow=mrow, ncol=mcol)
  for(i in 1:mrow){
    ic = va[i]
    X[i,ic] = 1
  }
  return(X)
}
ainv = function(pedi){
  n = nrow(pedi)
  Ainv = matrix(c(0), nrow = n, ncol = n)
  for ( i in 1:n) {
    animal = pedi[i,1]
    sire   = pedi[i,2]
    dam    = pedi[i,3]
    if (sire == 0 & dam == 0) {
      alpha = 1
      Ainv[animal,animal] = alpha + Ainv[animal,animal]
    }
    else if (sire != 0 & dam == 0) {
      alpha = 4/3
      Ainv[animal,animal] = alpha + Ainv[animal,animal]
      Ainv[animal,sire]   = -alpha/2 + Ainv[animal,sire]
      Ainv[sire,animal]   = -alpha/2 + Ainv[sire,animal]
      Ainv[sire,sire]     = alpha/4 + Ainv[sire,sire]
    }
    else if (sire == 0 & dam != 0) {
      alpha = 4/3
      Ainv[animal,animal] = alpha + Ainv[animal,animal]
      Ainv[animal,dam]    = -alpha/2 + Ainv[animal,dam]
      Ainv[dam,animal]    = -alpha/2 + Ainv[dam,animal]
```

```

    Ainv[dam,dam]      = alpha/4 + Ainv[dam,dam]
  }
  else {
    alpha = 2
    Ainv[animal,animal] = alpha + Ainv[animal,animal]
    Ainv[animal,sire]   = -alpha/2 + Ainv[animal,sire]
    Ainv[sire,animal]   = -alpha/2 + Ainv[sire,animal]
    Ainv[animal,dam]    = -alpha/2 + Ainv[animal,dam]
    Ainv[dam,animal]    = -alpha/2 + Ainv[dam,animal]
    Ainv[sire,sire]     = alpha/4 + Ainv[sire,sire]
    Ainv[sire,dam]      = alpha/4 + Ainv[sire,dam]
    Ainv[dam,sire]      = alpha/4 + Ainv[dam,sire]
    Ainv[dam,dam]      = alpha/4 + Ainv[dam,dam]
  }
}
return(Ainv)
}
gma=function(gped){
  library(MASS)
  library(GeneticsPed)
  M1=gped
  M= M1[,2:ncol(M1)]-1
  p1=round((apply(M,2,sum)+nrow(M))/(nrow(M)*2),3)
  p=2*(p1-.5)
  P = matrix(p,byrow=T,nrow=nrow(M),ncol=ncol(M))
  Z = as.matrix(M-P)
  b=1-p1
  c=p1*b
  d=2*(sum(c))
  ZZt = Z %*% t(Z)
  G = (ZZt/d)
  return(G)
}
a_matrix=function(ped){
  NRM=matrix(0,nrow(ped),nrow(ped))

  for(i in 1:nrow(ped)){
    a=ped[i,1]
    s=ped[i,2]
    d=ped[i,3]
    if(s > 0 && d > 0){
      for(j in 1:(a-1)){
        NRM[j,a] = 0.5*(NRM[j,s]+ NRM[j,d])
        NRM[a,j] = NRM[j,a]
      }
      NRM[a,a] = 1 + 0.5 * NRM[s,d]
    } else{
      if (s+d > 0 && s*d == 0){

```

```

    for(j in 1:(a-1)){
      NRM[j,a] = 0.5 * NRM[j,s+d]
      NRM[a,j] = NRM[j,a]
    }
  }
  NRM[a,a] = 1.0
}
}
return(NRM)
}
data=read.table("data.dat")
ped=read.table("data.ped")
sigma_a = 0.1
sigma_e = 0.9
alpha = sigma_e / sigma_a
X = desgn(data[,2])
Z = desgn(data[,1])
y = data[,3]
y=as.matrix(y)
ai = ainv(ped)
LHS = rbind(
  cbind(t(X) %*% X, t(X) %*% Z),
  cbind(t(Z) %*% X, t(Z) %*% Z + ai * alpha)
)
RHS = rbind(t(X) %*% y, t(Z) %*% y)
gi_LHS = ginv(LHS)
sol = gi_LHS %*% RHS
gped=read.table("marker.geno")
gped=gped[,c(-1)]
gm = gma(gped)
gi = ginv(gm)
LHS = rbind(
  cbind(t(X) %*% X, t(X) %*% Z),
  cbind(t(Z) %*% X, t(Z) %*% Z + gi * alpha)
)
RHS = rbind(t(X) %*% y, t(Z) %*% y)
gi_LHS = ginv(LHS)
sol1 = gi_LHS %*% RHS
tbv=read

```


주 의

1. 이 보고서는 농림축산식품부에서 시행한 첨단생산기술개발사업의 연구보고서입니다.
2. 이 보고서 내용을 발표할 때에는 반드시 농림축산식품부에서 시행한 첨단생산기술사업사업의 연구결과임을 밝혀야 합니다.
3. 국가과학기술 기밀유지에 필요한 내용은 대외적으로 발표 또는 공개하여서는 아니 됩니다.