ISSN (Print) 1225-9918
ISSN (Online) 2287-3406
Journal of Life Science **2014** Vol. 24. No. 11. 1258~1267 DOI : http://dx.doi.org/10.5352/JLS.2014.24.11.1258

- Review -

Development of an Economic-trait Genetic Marker by Applying Next-generation Sequencing Technologies in a Whole Genome

Jeong-An Gim and Heui-Soo Kim*

Department of Biological Sciences, College of Natural Sciences, Pusan National University, Busan 609-735, Korea

Received September 17, 2014 / Revised October 13, 2014 / Accepted November 5, 2014

Developing economic traits with a high growth rate, robustness, and disease resistance in livestock is an important challenge. RFLP and AFLP are the classical methods used to develop economic traits. Whole-genome-based economic traits have recently been detected with the advent of next-generation sequencing (NGS) technologies. However, NGS technologies are rather costly for use in studies, and RNA-seq, RAD-Seq, RRL, MSG, and GBS have been used to overcome the issue of high costs. In this study, recent NGS-based studies were reviewed, particularly those that focused on minimum costs and maximum effects. Then, we presented further prospects on how to apply for selection of high economic-trait livestock.

Key words : Genetic marker, NGS technologies, restriction enzyme, whole genome

서 론

유전자 마커란 인간을 비롯한 각 생명체의 표현형질을 나타내는 유전적 특성을 말한다[16, 18] 지금까지 인간에 있어서 정상과 질병 샘플을 비교한 연구가 많이 진행되었으며[49, 50], 가축에 있어서는 각 개체군에서 우수한 경제형질을 가진 샘플과 그렇지 않은 샘플을 비교한 연구가 많이 진행되었다[17, 25]. 이러한 표현형질을 나타내는 유전자 마커는 생명체의 게놈 상에 존재하고 있으나, 전장게놈의 경우 매우 방대하기에 기본적으로 형질과 관련되어 있다고 알려진 중요한 유전자 위주로 연구되고 있었다. 이전에는 RFLP (Restriction Fragment Length Polymorphism), AFLP (Amplified Fragment Length Polymorphism), microsatellite와 같은 기법을 사용해서 개체군 간 SNP (Single Nucleotide Polymorphism) 및 CNV (Copy Number Variation)와 같은 다양한 유전자 마커를 발굴하고 있었으나[13, 68], 이러한 기법들은 상대적으로 적은 정보와 많은 노동력 및 연구비용을 필요로 한다. 현재, NGS 기법의 대두로 인하여 시퀀스 정보와 유전형질이 많은 양의 데이터로 발굴되고 있는 실정이다. 이러한 NGS 데이터가 많이 쏟아져 나옴에 따라 기능유전체 분야의 연구가 가속화되고 있고, 인간 유전체에서 ENCODE (Encyclopedia of DNA Elements) 프로젝트가 주목받고 있다[15]. 인간 뿐만 아니라 가축 등 경제동물 및 식물에서도 유전체의 기능을 밝히기 위한 많은 NGS 데이터를 생성해 내고 있으며, QTL (quantitative trait locus)등을 비롯한 많은 게놈상의 영역에 대한 기능이 밝혀지고 있다[8, 45, 52, 59]. 하지만, 이와 같은 데이터를 생성 시 한 샘플 당 많은 실험 비용이 필요하고, 유전자 마커 발굴에 불필요한 부분까지 분석하는 경우가 많이 일어난다[41]. 때라서, 실험 비용을 줄이기 위하여 전체 샘플에 제한효소를 처리 한 후 특정 서열의 어댑터를 붙인 후, 어댑터 근처의 서열을 NGS 기법으로 분석한다든지, 혹은 전장게놈이 아닌 샘플의 전체 전사체를 분석하는 기법이 많이 활용되고 있다[6, 20, 48]. 이러한 기법으로 전장게놈에서의 SNP 및 구조적 변이체를 쉽고 빠르게 발굴할 수 있다. 고전적 마커 개발에 사용되었던 제한효소가 NGS 기법에 있어서 다시 주목받고 있는 점이 흥미롭다. 본 논문에서는 이렇게 적은 분석 비용으로 높은 효과를 얻을 수 있는 NGS 기법에 대한 소개 및 적용 가능성을 탐색해 보고자 한다. 이는 많은 수의 다양한 질병 환자 샘플 및 다양한 가축 종 샘플에서 우수한 유전자 마커를 찾는 데 도움을 줄 것으로 생각된다.

본 론

우수한 경제형질을 발굴 및 개체 선별에 활용하기 위한 노력으로써, 기존의 생어 시퀀싱 기법을 뛰어넘는 NGS 기법이 활용되고 있다. 이러한 NGS 기법은 다양한 플랫폼 하에서 상당히 많은 양의 서열 정보를 생성해 낼 수 있는 특징을 가지고 있다. 먼저 본 논문에서는 기본적인 NGS에 대한 소개 및, NGS 데이터를 만들어 낼 수 있는 플랫폼에 대하여 알아보았다. 그 다음으로, NGS 기법을 응용한 방법들인 RNA-seq, RAD-Seq, RRL, MSG, 그리고 GBS 기법에 대하여 알아보았다 (Table 1). 주로 샘플에 제한효소를 처리하거나, 혹은 여러 개

***Corresponding author**
Tel : +82-51-510-2259, Fax : +82-51-581-2962
E-mail : khs307@pusan.ac.kr

Table 1. The list of NGS-based analysis for confirming genetic polymorphisms

Experimental type	Study aims	Genome size	Reference genome	Population	Samples per population (or experimental set)	Size of recognition site	References
RNA-Seq	Confirming transcripts, or isoforms. Assessing gene expression level.	Big size genome is available.	*De novo* sequencing is available.	Medium	Medium	Not applicable	[9, 12, 48]
RAD-Seq	Ecological, or phylogeographic study in many samples.	Big size genome is available.	Not prerequisite.	Many	Small	Big	[6, 7, 21]
RRL	Polymorphism analysis, SNP analysis, DNA methylation.	Big size genome is available.	It is better to have a reference genome.	Medium	Medium	Depend on genome size	[30, 32, 42, 60]
MSG	QTL mapping, phenotype analysis, or making genetic map	Small size is better.	It is better to have a reference genome.	Small	Many	Small	[5, 19, 67]
GBS	Polymorphism analysis, detecting selective markers.	Big size genome is available.	It is better to have a reference genome.	Medium	Medium	Small, but methyl-sensitive	[10, 20]

의 샘플을 시퀀싱 이전에 통합하는 방법으로 개체군에 대한 샘플 분석을 진행한 연구에 대하여 소개하였다. 또한 이러한 기법들이 어떠한 연구에 어떻게 활용될 수 있는지에 대해 알아보고자 한다.

NGS 기법 소개

유전정보를 갖고 있는 DNA의 서열을 확인하기 위해 시퀀싱 기술이 필요한 바, 이에 1977년 프레드릭 생어와 그 동료들에 의해 최초로 사슬종결법을 통한 phi X174의 게놈이 해독되었다[53]. NGS 기법이 등장하기 이전인 2000년대 중반까지, 생어의 시퀀싱 기법을 기반으로 한 서열분석법이 주로 사용되었다. 저비용으로 대량의 데이터를 생산하고자 하는 연구자들의 요구에 따라, 다양한 플랫폼을 기반으로 한 NGS 기술이 대두하였다. 이러한 NGS 기법의 기본적 개념은 다음과 같다.

플랫폼마다 약간의 차이는 있지만, 기본적으로 NGS 분석을 위해 샘플에 대한 제한효소 처리를 통하여 큰 사이즈의 DNA 서열을 짧은 조각으로 만드는 작업이 가장 먼저 수행된다(Fig. 1A). 그 다음으로 각 짧은 DNA 조각의 양 끝에 어댑터를 ligation 한 후, 두 가닥을 한 가닥으로 만든다(Fig. 1B). 어댑터와 상보적인 서열인 프라이머를 붙이고, 프라이머의 3′ 말단에서 상보적으로 나머지 서열이 합성될 때 하나의 뉴클레오타이드 당 세 가지의 신호가 나타난다(Fig. 1C). 이는 첫 번째로 두 개의 phosphate, 두 번째로 하나의 수소 이온, 마지막으로 네 가지의 서로 다른 색으로 각각 표시된 염기이다. 하나의 뉴클레오타이드가 합성될 때 마다 NGS 기계가 위의 신호를 읽어 들여 서열을 확인한다. 이렇게 읽어들일 수 있는 짧은 DNA 조각의 길이는 플랫폼마다 다르지만 짧게는 100 bp 부터 길게는 400 bp에 이른다. 이렇게 읽어들인 짧은 DNA 조각의 서열들을 read라고 부르고, read를 모아서 하나의 서열을 만드는 과정을 assembly라고 한다. 이 때, 짧은 DNA 조각인 read를 모아 전체 게놈을 맞추는 작업을 bottom-up 시퀀싱 전략이라고 한다[56]. Read가 모여서 하나의 연속된 DNA 조각을 만들 수 있는데, 이를 contig라고 일컫고 연속적이라는 의미인 contiguous로부터 유래하였다. Contig가 모여서 scaf-

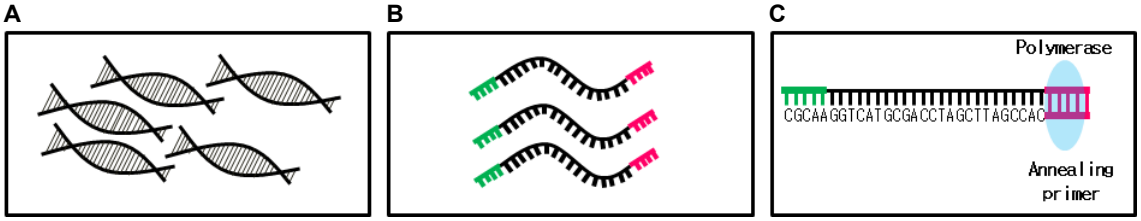

Fig. 1. The method applied by the NGS machine to amplify DNA fragment for sequencing. Whole genome is separated into small fragment (A), then ligated to both of adaptors at the end (B). Primers complementary to adaptor are annealed, and bases are synthesized. In this time, two phosphates, one hydrogen ion, and one fluorescence-labeled base are indicated as signal, and the machine reads these signals (C).

fold를 이루고, 이러한 scaffold가 모여서 한 샘플에 대한 전체 게놈이 완성된다. 이렇게 하여 얻어진 assembly 된 read를 참조서열에 붙여서 비교하는 과정을 mapping이라고 한다[54]. Mapping을 통하여 연구하고자 하는 샘플 서열에서의 SNP, CNV 및 INDEL을 확인할 수 있고, assembly 과정에서 시퀀싱되지 않은 부분인 gap을 줄일 수 있다. 이 때 하나의 샘플을 많이 읽을수록 많은 read를 얻게 되고, 그 만큼 서열의 정확도는 높일 수 있으나 비용도 그에 따라 증가하게 된다. 시퀀싱 과정에서 하나의 서열에 대해 읽어들인 read 수를 depth, 또는 coverage라고 한다. 이러한 depth를 구하는 공식은 아래 수식과 같다.

(The number of reads) × (The average read length) / (The length of genome) = Depth [35]

한편, 참조서열이 밝혀지지 않은 새로운 종의 서열을 NGS 기법을 통하여 확인하는 방법으로써, *de novo* assembly 시퀀싱 기법이 있다[58]. De novo 시퀀싱의 경우 많은 depth를 필요로 하고, 이에 많은 시간과 노력이 필요한 기법이다. 시간과 노력을 줄임과 동시에 *de novo* 시퀀싱의 장점을 살리기 위한 방법으로써, 아래에 설명할 RNA-Seq과 같은 전사체 분석 및, RAD-Seq과 같은 제한효소자리 인접 영역 분석에 *de novo* 시퀀싱이 많이 활용되고 있다[7, 21, 58].

이러한 시퀀싱의 목적은 각 생명체의 표현형질에 대한 유전특성을 설명할 수 있는 유전자 마커를 발굴하기 위한 것이다. 개체의 질병, 성장률, 강건성, 항병성, 약물 반응성과 같은 다양한 특성에 대한 유전적 요인을 게놈 단계에서 전반적으로 탐색하는 연구 방법을 Genome-wide association study (GWAS)라고 하는데[46, 47], NGS 기법이 대두함에 따라 이를 활용한 다양한 연구 결과가 발표되고 있다[11, 27, 28, 57, 65]. GWAS 기법을 이용하여 실험군과 대조군 각 두 그룹의 게놈에서, SNP 및 CNV와 같은 표현형질에 영향을 줄 수 있는 연구들이 많이 진행되고 있다. 이 때 특정 유전변이가 대조군과 비교하여 실험군에서 통계적으로 유의하게 나타날 경우, 이를 “associated” 라고 표현한다. 이렇게 발굴된 associated traits는 개체의 표현형에 영향을 줄 수 있는 유전자 마커가 될 수 있다. 특히 NGS 기법의 발달로 인하여 전장게놈 서열이 다양한 실험군에서 분석 가능하게 됨으로써, GWAS는 유전체학 분야에서 유전자 마커를 발굴하는 매우 유용한 기법으로써 많이 활용되고 있다.

하지만, 전장게놈 시퀀싱은 한 샘플 당 많은 비용이 들어갈 뿐만 아니라, 표현형을 나타내는 영역은 전체 게놈 영역 중 매우 일부분에 불과하다. 이에 전체 게놈이 아닌 게놈의 특정 부분만을 NGS 기법으로 시퀀싱 하는 기법인 RNA-seq, RAD-Seq, RRL, MSG, 및 GBS 등을 활용함으로써 표현형질을 발굴하는 있는 연구들이 많이 진행되고 있다.

NGS 플랫폼 소개

NGS 기술의 경우, 플랫폼의 선정에 따라서 실험 비용 및 산출 결과 해석이 달라지는 특성을 갖고 있기에, 실험 설계 단계에 있어서 어떠한 플랫폼을 사용하여 원하는 결과를 도출할 것인지에 대한 고찰이 필수적이다. 생어법을 기반으로 한 시퀀싱 플랫폼은 모세관 전기영동(capillary electrophoresis)을 활용한 자동화 기기가 1990년대 보급되었고, 이에 인간 게놈 프로젝트를 가능하게 하였다. 이러한 기기로는 대표적으로 Applied Biosystems 사의 ABI 3730xl 등을 들 수가 있으며, 현재에도 짧은 서열에 대한 시퀀싱 용도로 많이 활용되고 있다. 2000년대 중반 이후 생어 시퀀싱 기법을 뛰어넘는 새로운 염기 서열 분석법이 여러 회사에서 다양한 플랫폼을 토대로 제시되고 있고, 아래에 소개하고자 한다.

먼저, 2000년대 중반에 대두된 NGS 플랫폼은 기본적으로 전체 게놈을 제한효소를 이용하여 짧은 단편으로 만들고, 이들 DNA 단편에 대한 증폭을 하고 시퀀싱 과정에서 떨어져 나오는 형광 표지자 등을 탐지할 수 있는 소자를 통해 염기를 읽어 들인다. Roche 사에서 개발된 454 기종의 경우 반응할 샘플 내에 oil 성분과 한 쪽 PCR 프라이머가 표면에 붙어 있는 microbead를 넣어 준 후, emulsion을 유도하여 단일 가닥 DNA 단편들이 emulsion 내에서 PCR 반응이 일어나도록 하는 emulsion PCR 과정을 거친다. 이러한 과정 후, 각 bead에는 복제된 DNA 단편들이 존재하게 되고, 이들이 각각의 well에 들어간 후 시퀀싱 반응에 들어가게 된다. 이 때 생성되는 phosphate의 신호 강도를 확인하여 결과를 수치화함으로써 시퀀싱이 이루어진다. 이를 pyrosequencing이라고 하며, 상대적으로 길이가 긴 read를 얻을 수 있기에 assembly 및 mapping 과정이 수월하다는 장점이 있으나(600-1,000 bp), 정확도가 떨어진다는 단점을 갖고 있다[39-41]. 이러한 454 기종을 활용한 연구 중 대표적인 하나가, DNA 이중나선 구조를 발견한 제임스 왓슨의 게놈을 4개월 반 만에 해독한 것이다[64].

Illumina 사가 개발한 HiSeq 2000 기종의 경우 Roche의 454에 비해 짧은 read가 산출된다는 단점이 있으나(100bp), 상대적으로 저렴한 비용으로 많은 양의 데이터 생산이 가능한 특징을 갖고 있다. 반응할 샘플을 단편화시킨 후, 어댑터를 ligation 시킨 후 이를 FlowCell이라는 슬라이드에 흘려 준다. 이 때 FlowCell에 고정되어 있는 어댑터 서열과 상보적인 프라이머가 상호 결합하게 된다. 이 상태에서 PCR 반응이 진행되면 주변의 FlowCell에 있는 프라이머에 다른 쪽의 어댑터 서열이 붙으면서, DNA 단편이 알파벳 “U”의 형태로 구부러진다. PCR 반응을 통하여 특정 서열의 DNA 단편이 증폭되게 되는데 이를 cluster라고 하며, 이러한 과정을 bridge amplification이라고 한다[33, 40]. Cluster를 주형으로 하여 네 가지 염기에 대한 형광값을 시퀀싱 과정에서 읽어들인 후, 각각의 cluster 유래 서열을 assembly 하여 샘플에 대한 서열을 해독한다.

2010년대 현재 대두되고 있는 최신 NGS 플랫폼의 경우,

시퀀싱 전 PCR 증폭 과정을 생략하고DNA 단편을 바로 시퀀싱함으로써 시간과 노력을 최소화할 수 있는 장점을 갖고 있다. 이를 single molecule, real time (SMRT) 기법이라고 하고 [23, 24], Pacific BioScience 사가 개발한 PacBio RS 및 Life technologies 사의 Ion Proton과 같은 장비들에 의해 적용되고 있다. 이러한 플랫폼의 경우 반도체 소자에 위치한 작은 well에서 DNA 단편에 염기가 합성 될 때 신호를 직접 읽어 들임으로써 시퀀싱을 수행한다[62]. 앞으로도 다양한 회사에서 새로운 시퀀싱 플랫폼이 개발될 것으로 기대되는 바, NGS 플랫폼에 대한 정확한 이해를 통하여 다양한 샘플에서 보다 더 저렴한 비용과 작은 노력으로 최대의 결과를 도출해야 할 것이다.

RNA-Sequencing (RNA-Seq)

RNA-Seq법은 NGS 기법을 활용하여 전사체로부터 얻어진 cDNA (complementary DNA)를 시퀀싱 하는 기술이다. 시퀀싱 기법 및 과정은 전장게놈 시퀀싱 방법이랑 같다. 먼저 NGS 기법을 사용하여 cDNA로부터 시퀀싱을 수행하며, 시퀀싱 된 단편들은 참조서열과 비교 과정을 거친 후 참조서열 내의 적절한 위치에 시퀀싱 단편을 붙임으로써 RNA-Seq 과정이 수행된다. 이 때 SOAP [37]나 MAQ [36] 등의 생물정보학적 도구를 이용한다. 본 실험 기법은 RNA로부터 유래한 cDNA로 실험하기에 RNA를 추출할 수 있는 어떠한 샘플이면 실험 가능하다. 게다가, 시퀀싱 된 염기 서열 정보를 읽을 수 있을 뿐만 아니라, 각 전사체의 발현량 역시 읽어들인 수에 의해 추론 가능하다는 큰 장점을 가지고 있다. 또한 RNA-Seq은 전장게놈에 비해 훨씬 짧은 전사체 영역만 시퀀싱하기에, 상대적으로 저비용이라는 장점 또한 갖고 있다. 현재까지 인간 세포주 [12], 경주마 운동 전후[48], 소의 우유[9] 등과 같은 샘플에서 RNA-Seq이 적용된 사례가 보고되었다. 이러한 사례를 바탕으로, RNA-Seq의 경우 SNP 분석[12] 및 전사체의 변이체[48]를 발굴해 내는 데 사용될 수 있다는 것을 시사할 수 있다. 이러한 전사체 및 전사체의 변이체를 발굴하는 것은, 전사체로 인한 표현형의 변화를 확인하는 연구에서 매우 필수적인 과정이다. RNA-Seq의 경우 지금까지 개발된 방법들에 비해 전체 전사체의 발현량 및 발현 정도를 가장 정확하고 빠르게 분석할 수 있는 방법이라고 할 수 있다[63]. 지금까지 분석된 종의 전장게놈 길이는 사람과 비슷하거나 약간 짧은 수준이기에, 사람 정도의 상대적으로 긴 전장게놈도 쉽게 분석할 수 있다.

한편, RNA-Seq의 경우 단점도 발견되는데, 전사체의 발현량을 정확히 추론하기 위한 정량화에 대한 연구가 상대적으로 많이 진행되지 않은 실정이다[14]. 또한, 정확한 참조서열이 공개되지 않은 종에서는 RNA-Seq 수행이 곤란하다는 점을 들 수 있다. 하지만, 인간을 비롯한 대부분의 경제동물에서 참조서열이 속속 공개되고 있고, *de novo* 전사체 분석이 이루어짐으로써[34] 위의 단점은 완화되고 있는 실정이다. RNA-Seq의 경우 기존의 마이크로어레이 실험을 대체하는 매우 강력한 기법으로써, 다양한 샘플에서 많은 양의 데이터를 만들어 내고 있는 기법이다[63]. 앞으로 보다 더 정확한 전사체의 동정과 발현량의 추론 그리고 더 많은 전사체의 변이체 발굴과 더불어서, 보다 더 쉽고 빠르게 분석해 주는 생물정보학적 도구가 더 출현할 것으로 기대한다.

Restriction site-associated DNA sequencing (RAD-Seq)

RAD-Seq 기법은 제한효소로 각 DNA 샘플을 단편화시킨 후 NGS 기법을 활용하여 읽어들인 서열데이터를 분석하는 기법이다. RAD 마커의 경우 본래 마이크로어레이 실험에서 처음 적용되었으며[43], 이후 NGS 기술과 더불어 큰가시고기(stickleback)의 SNP를 통한 유전적 다형성을 연구하는 데 적용되었다[6]. RAD-Seq의 경우 최소 300 ng의 정제된 DNA가 필요하고, 여기에 제한효소를 처리 한 후, 샘플을 표지할 수 있는 각 샘플 특이적인 1차 어댑터를 붙이는 과정을 거치게 된다. 1차 어댑터는 Illumina Genome Analyzer 기반의 시퀀싱을 위한 프라이머 서열과, 각 샘플 별 특이적인 4-5 bp로 이루어져 있는 바코드 서열로 구성되어 있다. 샘플 별 구분을 용이하게 하기 위해, 바코드 서열은 최소 두 개 이상의 뉴클레오타이드를 샘플 별로 다르게 구성할 것이 요구된다[6]. 1차 어댑터가 달린 서열은 연구 목적에 따라 샘플을 섞고, DNA 단편의 사이즈 선택(300-700 bp)을 수행한다. 또한, NGS 분석을 위한 DNA 전처리 과정인 다듬기(shear) 과정을 소니케이터를 사용하여 거친다. 다듬기 과정을 거친 DNA 단편에 1차 어댑터와 구별되는 2차 어댑터를 붙인 다음 PCR을 통해 각 단편을 증폭해 내고 참조서열과 비교하여 특성을 확인한다 (Fig. 2).

본 논문에서 소개한 기법 중 상대적으로 많은 비용과 실험 단계가 필요하지만, 바코드 서열이 각 개체별 특이성을 높여준다는 장점을 갖고 있다. 이러한 RAD-Seq 기법은, 큰가시고기의 개체군 분화와 선택 연구에 적용된 바 있다[29]. 또한 non-model organism인 식충식물 기생 모기(*Wyeomyia smithii*)의 계통지리학(phylogeography)적 분석에 적용하여기후 변화에 따른 유전적 변화를 밝혀 내었다[21]. 마찬가지로 배추좀나방(*Plutella xylostella*)의 유전자 지도 작성에 있어서, 비교유전체학의 관점에서 RAD-Seq이 적용된 바 있으며[7] 열대어의 계통분류 및 적응방산(adaptive radiation) 연구에도 본 RAD-Seq 기법이 사용되었다[61]. RAD-Seq 기법의 경우 참조서열이 없거나 정확하지 않는 non-model organism의 개체군 연구에 강점을 보이고 있다[7, 21]. 이에 흥미롭게도, RAD-Seq의 경우 상대적으로 생태학적, 진화학적 및 생물지리학 등의 분야에서 많이 사용되는 것을 알 수 있다. 아래 소개 할 기법들 (RRL, MSG 및 GBS) 역시 RAD-Seq을 기반으로 하여 발전한 기법이자, 샘플 및 실험 조건에 따라 RAD-Seq을 적절히 변형한 기법들이라고 할 수 있다. 앞으로 보다 더 쉬우면서도 저비용으로 다양한 샘플에서 NGS 분석을 수행하게 될 RAD-Seq

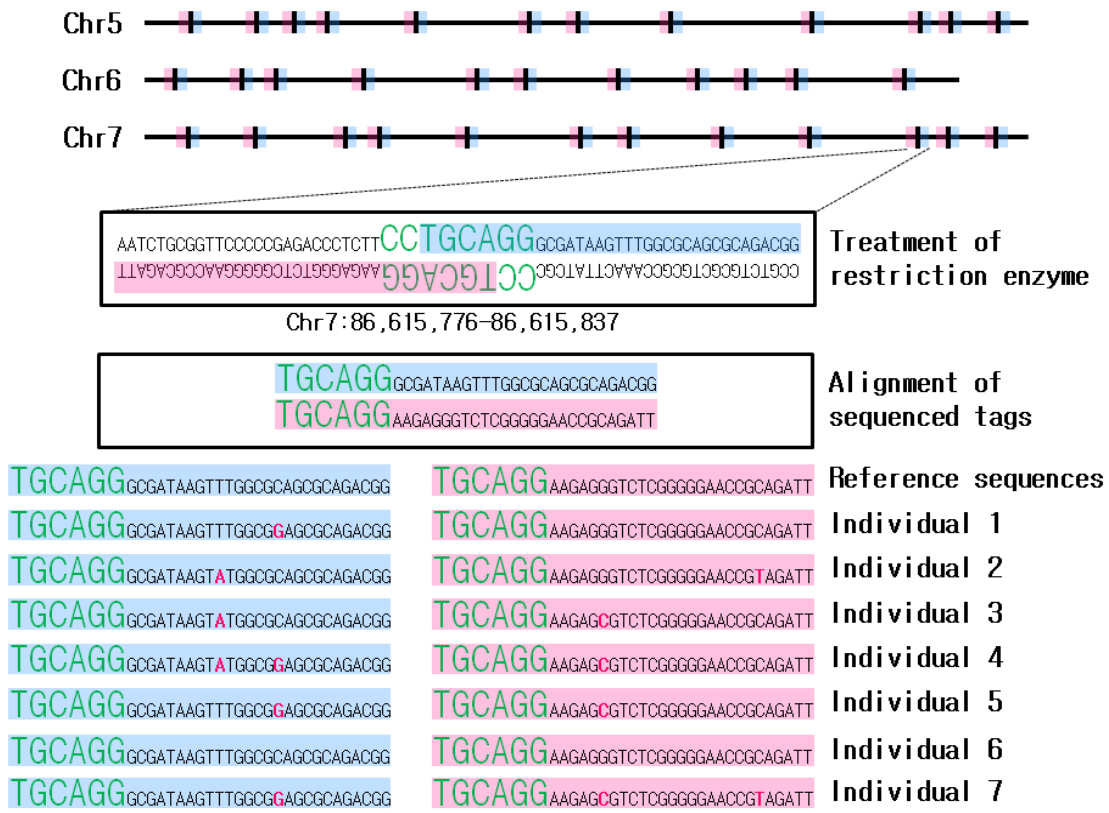

Fig. 2. The scheme of methods for NGS-based marker discovery. After treatment of restriction enzyme to whole-genome, each DNA fragments were aligned (or ligated adaptor sequences). Then, DNA fragments were aligned, and analyzed their differences among individual in population. When *Sbf*I is used, the restriction enzyme site is indicated as large letters in horse genome.

의 발전 방향이 기대된다.

Reduced-representation library (RRL)

위 단락에서 소개한 RAD-Seq과 마찬가지로 RRL 기법은 각 샘플로부터 유래한 genomic DNA를 대상으로 제한효소 처리를 하여, 100-200 bp 정도의 단편을 회수 한 후 어댑터를 달아 NGS 기법으로 시퀀싱 하는 방법이다. 어댑터가 달린 단편 양 끝을 읽어들인 후 참조서열과 비교하여 각 샘플 간 다형성을 확인한다. 현재까지 소[60], 돼지[4, 51], 및 칠면조[32] 등과 같은 경제동물 뿐만 아니라, 콩[30] 및 옥수수[26] 등과 같은 식물에서 RRL를 이용한 다형성 분석이 이루어졌다. 사람에 있어서는 총 18 개의 지역별 개체군에서 선정한 개인의 DNA에서 RRL가 인간의 진화에 대한 연구를 개체군 유전학의 관점에서 적용된 바 있다[38]. 이 방법은 최초로 인간 게놈 SNP를 분석하는 데 사용되었으며[2], 실제로 개체군 내에서의 형질과 관련된 SNP를 발굴하는 데 많은 연구가 진행되고 있다. 그리고 RRL를 통해 얻어진 SNP 데이터는 전장게놈 재 시퀀싱 프로젝트에 등록할 수 있다[44]. 상대적으로 저비용 및 간단한 실험 결과로써 많은 데이터를 얻을 수 있고, 상대적으로 큰 전장게놈에도 적용할 수 있다는 장점이 있으나, 연구 목표에 적절한 마커를 제공할 수 있는 제한효소의 선정이 중요하다.

또한 RRL의 장점을 응용한 reduced-representation bisulfite sequencing (RRBS) 기법이 소개되고 있다[42]. DNA 메틸레이션 분석에 있어서 bisulfite를 처리한 후 시퀀싱 했을 때, 메틸화가 된 시토신은 시토신 그대로 표시되는 반면, 메틸화가 되지 않은 시토신은 티민으로 읽히게 된다. 이에 DNA 메틸화 양상에 따라 SNP가 생기게 되는데, RRL의 장점을 살린 RRBS를 통하여 DNA 메틸레이션 양상을 저비용 및 고효율의 데이터로 확인할 수 있다.

Multiplexed shotgun genotyping (MSG)

상대적으로 보다 간소화된 실험기법과 저비용에 대한 요구로, Multiplexed shotgun genotyping (MSG) 기법이 개발되었다[5]. MSG의 경우 위에서 소개한 RAD-Seq을 기반으로 하는 연구이고, 어댑터에 6 bp의 바코드 서열을 붙인다는 점에서 비슷한 기법이나 보다 단순화된 실험 프로토콜을 가지고 있다. 특히, 적은 양의 DNA로도 실험이 가능하여, 많은 양의 DNA를 얻기 힘든 샘플인 초파리에서 최초로 MSG가 사용되었다. 특히 MSG의 강점 중의 하나가 MSG가 적용된 각 개체군 샘플에서 Hidden Markov Model (HMM)을 사용하여 각 개체

군의 조상을 추측할 수가 있다는 점이다. HMM이란 실제 관측이 불가능한 결과를 비슷한 데이터들을 바탕으로 예측해 내는 통계적 기법이다. HMM을 사용하여 개체군 시퀀싱 후 조상 단계의 유전형 분석에 적용된 바 있으며[66], MSG에서 역시 적용되었다[5]. 교잡된 벼의 QTL을 분석하기 위해 MSG가 적용되었고 새로운 여섯 개의 QTL을 새로 동정한 사례가 있다[19]. 또한, 갈색등근바리(*Epinephelus coioides*)에서의 유전자 연관 지도를 만드는 데에 적용된 바 있고, 차후 QTL 분석 등 경제형질 관련 마커를 발굴하는 데 많은 역할을 할 것으로 예상하고 있다[67]. 이와 같이 MSG 기법은 기존 데이터를 바탕으로 실제 관측이 곤란한 데이터까지 추정할 수 있는 HMM과 같은 외삽법(Imputation)을 적용하기 쉽다는 점에서 가장 큰 장점이 있다고 할 수 있다. 물론 이러한 HMM을 적용하기 위해서 마커의 수와 각 개체 수를 많이 해야 한다는 전제 조건이 있고[5], 상대적으로 약간 까다로운 실험 설계 및 지나치게 큰 크기의 전장게놈 분석이 힘들다는 단점이 있으나, 잘 활용한다면 위에 설명한 바와 같이 많은 장점을 가져다 주는 실험 기법이다.

Genotyping by sequencing (GBS)

RAD-Seq과 비슷한 기법으로 Genotyping by sequencing (GBS) 기법이 도입되었다. RAD-Seq과 비교하여 더 작은 양(100 ng)의 DNA가 필요하며, 다듬기 과정 및 DNA 단편의 사이즈 선택 과정이 생략된다는 강점이 있다. 제한 효소 처리 후, 바코드 서열이 포함된 어댑터와 공통서열이 포함된 어댑터 두 가지를 동시에 섞고 DNA 단편에 ligation 된다. 이 때 모든 어댑터와 연결된 DNA 단편이 시퀀스 되지 않는데, 이는 1 kb 이상의 긴 서열이나, 한 DNA 단편에 바코드 서열만 포함되었다든지 또는 공통서열만 포함될 경우 시퀀싱 되지 않는다. 나머지 과정은 RAD-Seq과 비슷하거나 같으며, 상대적으로 저비용으로 SNP와 같은 많은 마커들을 발굴해 낼 수 있다. GBS는 2011년에 기존의 RAD-Seq을 단순화한 새로운 실험 기법으로 소개되었으며, 옥수수와 보리에서 GBS 마커가 분석된 바 있다[20]. 특히 본 연구의 분석에서 저자들은 메틸레이션-sensitive한 제한효소인 *Ape*KI를 사용할 것을 제안하고 있다. *Ape*KI의 경우 GCWGC로 이루어진 8 bp의 제한효소 절단길이를 가지고 있으며, 3′ 말단의 사이토신이 메틸화되어 있을 경우 작동하지 못하는 특성을 갖고 있다. 이동성 유전인자들(transposable elements, TEs)을 비롯한 게놈 상의 반복서열들(repetitive sequences, REs)의 경우 일반적으로 유전체 불안정성(genomic instability)를 유도하기에, 숙주세포(host cell)는 RE에 대한 메틸레이션을 유도함으로써 RE에 의한 유전체 불안정성을 막는 기작을 보여 준다[10]. 이에 상대적으로 RE에 메틸레이션이 많이 되어 있기에, RE가 아닌 게놈 상의 다른 영역을 제한효소로 자를 확률을 높여주는 효과를 제공한다. 이러한 GBS 기법은 콩의 총 8가지 유전자형 분석에서도 적용되었는데, 이 실험에서는 복잡도를 감소시키기 위해 선택적인 프라이머를 사용하였다. 이에 SNP가 40% 정도 증가하였고 시퀀스 해독 배수(depth)는 두 배 정도 늘어남을 확인할 수 있었다[55]. 이러한 GBS의 경우 RAD-Seq에 비해 저렴하면서도 고품질의 데이터를 제공해 주기에, 앞으로 발전 및 적용 가능성이 기대되는 기법이라고 할 수 있다.

연구 대상종 선정 및 연구 목적에 맞는 해상도 결정

NGS 데이터의 경우 방대한 데이터를 제공하는 반면, 명확한 실험 계획 및 분석 없이는 의미 없는 데이터를 만들어 내는 경우가 많다. 이에 저비용으로 NGS 데이터 분석을 하여 최대한의 많은 효과를 얻기 위해서는, 어떠한 목적으로 어떠한 실험 대상 종을 선정해서 얼마나 많은 샘플에 대해 적용할 것인지의 물음에 대한 답이 필요하다.

NGS 데이터의 몇 가지 특징 중 하나가 참조서열이 있으면 분석하기 쉽다는 것이다. 참조서열이 있는 종을 분석하였을 경우, 그 종에서 NGS를 통하여 만들어진 서열 단편들을 쉽게 참조서열에 정렬할 수 있을 뿐만 아니라, 유전체 상의 어떠한 위치에 위치하는지를 쉽게 알 수 있다. 또한, 제한효소를 사용하는 NGS 실험에 있어서 참조서열의 길이를 알 수 있다면, 제한효소 절단길이에 따라 적절한 제한 효소를 사용하여 예상되는 마커 site의 개수를 아래 수식과 같이 예측할 수 있다. 이에, 아래의 수식을 이용하여 참조서열의 길이에 따른 적절한 제한효소를 선정하는 것이 중요해진다[16].

$(0.25)^n \times$ (reference genome size) = expected site number [22]

NGS 데이터의 몇 가지 특징 중 다른 하나는, 시퀀스 해독 배수가 높으면 정확한 서열을 얻을 수 있으나 비용은 증가한다는 것이다. 이에, 연구 목표에 맞는 시퀀스의 해독 배수를 정해 적당량의 데이터를 얻어 내는 것이 중요하다. 낮은 시퀀스의 해독 배수는 저비용으로 쉽게 분석할 수 있는 장점이 있으나, 생물학적 의미를 가진 데이터로서의 가치가 떨어질 수 있다. 반면, 높은 시퀀스의 해독 배수는 충분한 데이터를 제공해 줄 수 있으나, 높은 비용을 필요로 하고 생물정보학적 분석에 많은 시간 및 노력이 소모된다는 단점이 있다.

개체군 수와 개체군 당 개체의 수

개체군 연구에 있어서 각 개체 간 변이를 찾아내고, 변이에 따른 우수 유전형질 및 특이적 유전형질을 찾아내는 것은 중요하다. 이러한 개체군 연구의 세부 목적으로는 크게 우수 유전형질에 대한 인공선택의 결과 연구, 계통지리학적 연구, QTL 분석을 통한 유전자 지도 작성 및 특정 과정 이후 생명체의 유전자 발현이 급격하게 많이 바뀌는 경우 등과 같이 나누어 볼 수 있다.

먼저 인공선택의 결과 연구에서, 가축화된 동물에 있어서 게놈상의 특정 영역에 대해 F_{ST} (Wright's fixation index) 등의

통계적 수식을 통해 개체군유전학의 관점에서 접근하는 연구가 있다. 가축화된 종은 인간에 의해서 인위적으로 특정 형질을 가진 개체 위주로 교배되는 인공선택을 받기에, 특정 영역에 대한 상대적으로 빠른 진화율이 확인되고, 이는 개체군에 있어서 selective sweep 등의 현상으로 나타난다. 이러한 인공선택의 결과를 파악하는 것은 특정 형질에 대한 마커를 확인할 수 있을 뿐만 아니라, 개체군에서의 형질다형성 연구에 좋은 연구 테마를 제시해 줄 수 있다. 실제로 돼지의 네 가지 품종 및 멧돼지에서 각 품종 당 최소 23개체 최대 36개체에 대하여 RRL을 실시하였고 SNP 분석을 실시한 사례가 있다. 총 세 종류의 제한효소(*Alu*I, *Hae*III 그리고 *Msp*I)를 사용한 후 NGS 기법으로 시퀀싱을 수행하였다[51]. 후속 연구로써, 가축화된 품종은 비슷한 선택압을 받았고 전체 게놈 중 약 7%가 선택 받은 영역임을 밝혀 내었다[3]. 뿐만 아니라, 가축화된 품종과 멧돼지를 비교했을 때, 성장률과 관련된 유전자들의 선택 양상이 가축화된 품종 내에서 더 많이 관련성을 보인다는 것을 밝혀 내었다. 앞으로 돼지뿐만 아니라 소, 말, 개 등의 경제동물의 개체군에서도 선택압을 받은 영역을 찾아낸다면 추후 우수한 형질을 가진 마커를 쉽게 찾을 수 있고, 보존할 수 있을 것이다.

그 다음 세계적 추세에 비추어 볼 때 대한민국에서는 상대적으로 덜 활발히 이루어지는 연구인 계통지리학의 개체군유전학적 연구에도 본 논문에서 소개한 NGS 기법이 어떻게 적용될 것인지 고찰해보고자 한다. 지역별 및 시기별 동물상 및 식물상을 고찰하기 위해서 각 샘플의 개체군을 분석하게 된다. 이에 본 실험은 앞에서 소개한 인공선택 영역을 발굴하는 실험에 비해 상대적으로 높은 개체군 수와, 상대적으로 낮은 개체군 내 개체 수를 갖는다는 특징이 있다[16]. 위에서 논의한 RAD-Seq을 통하여 *W. smithii*의 계통지리학적 분석을 한 연구가 대표적이다. 실험 대상종인 모기는 참조서열이 없고 야외에 있는 개체군에서 수행했기에 이들의 조상에 대한 유전형을 알 수 없는 상태였다. 이에 조상에 대한 유전형을 외삽법을 통해 알아내었고, 각각의 SNP가 분석되어 21개의 개체군에 대한 연관관계가 확인되었다. 이러한 실험에서는 각 마커 사이의 밀도가 높을 필요성이 없으므로 제한효소의 절단길이가 긴 것으로 하는 것이 바람직하며, 실제로 8 bp의 제한효소 절단길이를 가지는 *Sbf*I를 사용하였다[21]. 따라서 RAD-Seq 뿐만 아니라 기존 데이터를 갖고 비슷한 실험 대상의 데이터를 예측하는 HMM 등의 외삽법을 활용할 수 있는 MSG 및 GBS와 같은 방법을 적용해 볼 수 있을 것이다.

이러한 NGS 기법은 유전자 지도를 만드는 데 사용되기도 한다. 이러한 연구의 경우 전장게놈 시퀀싱이 가장 이상적이긴 하나, 분석 비용 및 시간 문제로 본 논문에서 소개한 기법을 사용하여 유전자 지도를 만들게 된다. 유전자 지도 작성의 특성 상, 게놈 사이즈가 적을수록 분석이 용이해지고 참조서열의 의존성이 높다는 점을 들 수 있다. 이에 상대적으로 전장게놈 사이즈가 작은 벼(389 Mb)[31]와 초파리(120 Mb)[1] 에서 QTL 분석을 위해 MSG를 사용한 사례가 있다[5, 19]. 이러한 기능 유전체 분석의 데이터 정확도를 높이기 위해서 각 마커 사이의 밀도는 상대적으로 높아야 할 필요성이 있다. 따라서 제한효소는 벼 분석에서는 *Mse*I, 초파리 분석에서는 *Taq*I 제한효소가 사용되었는데 이는 각각 4 bp의 제한효소 절단길이를 가진다. 두 실험 모두 단일 개체군에서 벼 분석의 경우 781개, 그리고 초파리 분석의 경우 96개의 개체에서 실험되었다. 이렇게 나온 실험 결과를 바탕으로, 여러 가지 이유로 실험용 샘플을 구하지 못하는 비슷한 연구에 HMM 등의 외삽법을 통하여 예측할 수 있다는 장점 또한 가지고 있다[5].

RNA-Seq의 경우 전장 전사체에 대한 분석을 통하여 유전자 발현, SNP 발굴 및 전사체에 대한 변이체들을 발굴해 내는데 적용될 수 있다. 기존 유전자 발현을 대량으로 확인할 수 있었던 microarray에 비해, 샘플의 모든 전사체를 확인할 수 있는 장점을 갖고 있다. 이에 운동능력이 좋은 경주마 3두, 그리고 낮은 경주마 3두 총 6두에서 운동 전후 혈액과 골격근 각각 총 24 샘플에서 RNA-Seq이 수행된 실험이 있다. 운동 전후로 발현이 달라진 유전자들 및 우수마와 열등마 특이적 발현 유전자들이 분석되었고, 서로 발현이 달라지는 유전자의 변이체들을 발굴하였다[48]. 상대적으로 저렴한 비용으로 많은 결과를 얻을 수 있고 후속 연구에 파급력이 높은 데이터를 제시해 줄 수 있다는 점에서, 앞으로 발전 가능성이 기대되는 기법이라고 할 수 있다.

정리하자면, 실험 디자인에 있어서 어떠한 연구를 할 것인지, 그리고 연구를 통해 어떠한 데이터를 얻을 것인지에 대해 결정하는 것이 필요하다. 그 다음으로 연구 대상 종의 전장게놈 크기에 따라 적절한 제한효소 절단길이를 가지는 제한효소를 선정해야 하고, 참조서열을 구할 수 있는지에 따라 DNA 단편을 얼마만큼의 시퀀스 해독 배수로 읽을 것인가를 결정해야 한다. 또한, 개체군 수와 개체군 내 개체 수가 얼마나 되는지에 따라 적절한 방법을 선정해야 한다.

결 론

NGS 기술의 대두 이후 지난 10여년 간 NGS를 활용한 새로운 실험 기법이 급속히 발전하고 있다. 이러한 발전 이유 중 하나는 최소한의 비용으로 최대한의 효과를 얻기 위한 과학자들의 노력에 있었다고 볼 수 있다. 최근 암 생물학, 동물육종학에서 분자생태학에 이르기까지 거의 생명과학과 생명공학의 모든 분야에서 NGS를 활용한 유전체 특성 및 형질 특이적 영역을 찾으려는 노력이 계속되어 왔고 앞으로 그럴 것이다. 본 논문에서 최근까지 연구된 NGS 기법을 설명하고, 실제 실험에 어떻게 활용되었는지에 대해 설명하였다. 본 논문에서 소개된 다양한 NGS 기법에 대해 상세한 이해를 기초로 하여, 실험하고자 하는 샘플의 특성과 NGS 기술의 특성을 잘 파악

하여 실험을 잘 수행한다면, 저비용으로 가치 있고 많은 양의 데이터를 얻을 수 있을 것으로 기대된다.

감사의 글

본 논문은 농촌진흥청 공동연구사업(과제번호: PJ009254)의 지원에 의해 이루어진 것임.

References

1. Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A. and Galle, R. F., et al. 2000. The genome sequence of Drosophila melanogaster. *Science* **287**, 2185-2195.
2. Altshuler, D., Pollara, V. J., Cowles, C. R., Van Etten, W. J., Baldwin, J., Linton, L. and Lander, E. S. 2000. An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407**, 513-516.
3. Amaral, A. J., Ferretti, L., Megens, H. J., Crooijmans, R. P., Nie, H., Ramos-Onsins, S. E., Perez-Enciso, M., Schook, L. B. and Groenen, M. A. 2011. Genome-wide footprints of pig domestication and selection revealed through massive parallel sequencing of pooled DNA. *PLoS One* **6**, e14782.
4. Amaral, A. J., Megens, H. J., Kerstens, H. H., Heuven, H. C., Dibbits, B., Crooijmans, R. P., den Dunnen, J. T. and Groenen, M. A. 2009. Application of massive parallel sequencing to whole genome SNP discovery in the porcine genome. *BMC Genomics* **10**, 374.
5. Andolfatto, P., Davison, D., Erezyilmaz, D., Hu, T. T., Mast, J., Sunayama-Morita, T. and Stern, D. L. 2011. Multiplexed shotgun genotyping for rapid and efficient genetic mapping. *Genome Res* **21**, 610-617.
6. Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., Selker, E. U., Cresko, W. A. and Johnson, E. A. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* **3**, e3376.
7. Baxter, S. W., Davey, J. W., Johnston, J. S., Shelton, A. M., Heckel, D. G., Jiggins, C. D. and Blaxter, M. L. 2011. Linkage mapping and comparative genomics using next-generation RAD sequencing of a non-model organism. *PLoS One* **6**, e19315.
8. Bonneau, J., Taylor, J., Parent, B., Bennett, D., Reynolds, M., Feuillet, C., Langridge, P. and Mather, D. 2013. Multi-environment analysis and improved mapping of a yield-related QTL on chromosome 3B of wheat. *Theor Appl Genet* **126**, 747-761.
9. Cánovas, A., Rincon, G., Islas-Trejo, A., Wickramasinghe, S. and Medrano, J. F. 2010. SNP discovery in the bovine milk transcriptome using RNA-Seq technology. *Mamm Genome* **21**, 592-598.
10. Carnell, A. N. and Goodman, J. I. 2003. The long (LINEs) and the short (SINEs) of it: altered methylation as a precursor to toxicity. *Toxicol Sci* **75**, 229-235.
11. Chen, W., Gao, Y., Xie, W., Gong, L., Lu, K., Wang, W., Li, Y., Liu, X., Zhang, H. and Dong, H., et al. 2014. Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism. *Nat Genet* **46**, 714-721.
12. Chepelev, I., Wei, G., Tang, Q. and Zhao, K. 2009. Detection of single nucleotide variations in expressed exons of the human genome using RNA-Seq. *Nucleic Acids Res* **37**, e106-e106.
13. Cho, Y., McCouch, S., Kuiper, M., Kang, M. R., Pot, J., Groenen, J. and Eun, M. 1998. Integrated map of AFLP, SSLP and RFLP markers using a recombinant inbred population of rice (Oryza sativa L.). *Theor Appl Genet* **97**, 370-380.
14. Christodoulou, D. C., Gorham, J. M., Herman, D. S. and Seidman, J. 2011. Construction of normalized RNA seq libraries for next generation sequencing using the crab duplex specific nuclease. *Curr Protoc Mol Biol* **12**.
15. Consortium, E. P. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74.
16. Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M. and Blaxter, M. L. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat Rev Genet* **12**, 499-510.
17. Dekkers, J. C. 2004. Commercial application of marker- and gene-assisted selection in livestock: strategies and lessons. *J Anim Sci* **82 E-Suppl**, E313-328.
18. Desta, Z. A. and Ortiz, R. 2014. Genomic selection: genome-wide prediction in plant improvement. *Trends Plant Sci* **19**, 592-601.
19. Duan, M., Sun, Z., Shu, L., Tan, Y., Yu, D., Sun, X., Liu, R., Li, Y., Gong, S. and Yuan, D. 2013. Genetic analysis of an elite super-hybrid rice parent using high-density SNP markers. *Rice* **6**, 1-15.
20. Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S. and Mitchell, S. E. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* **6**, e19379.
21. Emerson, K. J., Merz, C. R., Catchen, J. M., Hohenlohe, P. A., Cresko, W. A., Bradshaw, W. E. and Holzapfel, C. M. 2010. Resolving postglacial phylogeography using high-throughput sequencing. *Proc Natl Acad Sci USA* **107**, 16196-16200.
22. Etter, P. D., Bassham, S., Hohenlohe, P. A., Johnson, E. A. and Cresko, W. A. 2011. SNP discovery and genotyping for evolutionary genetics using RAD sequencing. *Methods Mol Biol* **772**, 157-178.
23. Fang, G., Munera, D., Friedman, D. I., Mandlik, A., Chao, M. C., Banerjee, O., Feng, Z., Losic, B., Mahajan, M. C. and Jabado, O. J., et al. 2012. Genome-wide mapping of methylated adenine residues in pathogenic Escherichia coli using single-molecule real-time sequencing. *Nat Biotechnol* **30**, 1232-1239.
24. Flusberg, B. A., Webster, D. R., Lee, J. H., Travers, K. J., Olivares, E. C., Clark, T. A., Korlach, J. and Turner, S. W. 2010. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat Methods* **7**, 461-465.

25. Gomez-Raya, L., Olsen, H. G., Lingaas, F., Klungland, H., Vage, D. I., Olsaker, I., Talle, S. B., Aasland, M. and Lien, S. 2002. The use of genetic markers to measure genomic response to selection in livestock. *Genetics* **162**, 1381-1388.
26. Gore, M. A., Chia, J. M., Elshire, R. J., Sun, Q., Ersoz, E. S., Hurwitz, B. L., Peiffer, J. A., McMullen, M. D., Grills, G. S. and Ross-Ibarra, J., et al. 2009. A first-generation haplotype map of maize. *Science* **326**, 1115-1117.
27. Grossi, D. D., Buzanskas, M. E., Grupioni, N. V., de Paz, C. C., Regitano, L. C., de Alencar, M. M., Schenkel, F. S. and Munari, D. P. 2014. Effect of IGF1, GH, and PIT1 markers on the genetic parameters of growth and reproduction traits in Canchim cattle. *Mol Biol Rep* [Epub ahead of print].
28. Gurung, S., Mamidi, S., Bonman, J. M., Xiong, M., Brown-Guedira, G. and Adhikari, T. B. 2014. Genome-wide association study reveals novel quantitative trait Loci associated with resistance to multiple leaf spot diseases of spring wheat. *PLoS One* **9**, e108179.
29. Hohenlohe, P. A., Bassham, S., Etter, P. D., Stiffler, N., Johnson, E. A. and Cresko, W. A. 2010. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genet* **6**, e1000862.
30. Hyten, D. L., Cannon, S. B., Song, Q., Weeks, N., Fickus, E. W., Shoemaker, R. C., Specht, J. E., Farmer, A. D., May, G. D. and Cregan, P. B. 2010. High-throughput SNP discovery through deep resequencing of a reduced representation library to anchor and orient scaffolds in the soybean whole genome sequence. *BMC Genomics* **11**, 38.
31. International Rice Genome Sequencing Project. 2005. The map-based sequence of the rice genome. *Nature* **436**, 793-800.
32. Kerstens, H. H., Crooijmans, R. P., Veenendaal, A., Dibbits, B. W., Chin, A. W. T. F., den Dunnen, J. T. and Groenen, M. A. 2009. Large scale single nucleotide polymorphism discovery in unsequenced genomes using second generation high throughput sequencing technology: applied to turkey. *BMC Genomics* **10**, 479.
33. Kozarewa, I. and Turner, D. J. 2011. Amplification-free library preparation for paired-end Illumina sequencing. *Methods Mol Biol* **733**, 257-266.
34. Kumar, S. and Blaxter, M. L. 2010. Comparing de novo assemblers for 454 transcriptome data. *BMC genomics* **11**, 571.
35. Lander, E. S. and Waterman, M. S. 1988. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* **2**, 231-239.
36. Li, H., Ruan, J. and Durbin, R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18**, 1851-1858.
37. Li, R., Li, Y., Kristiansen, K. and Wang, J. 2008. SOAP: short oligonucleotide alignment program. *Bioinformatics* **24**, 713-714.
38. Luca, F., Hudson, R. R., Witonsky, D. B. and Di Rienzo, A. 2011. A reduced representation approach to population genetic analyses and applications to human evolution. *Genome Res* **21**, 1087-1098.
39. Luo, C., Tsementzi, D., Kyrpides, N., Read, T. and Konstantinidis, K. T. 2012. Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. *PLoS One* **7**, e30087.
40. Mardis, E. R. 2008. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet* **9**, 387-402.
41. Mardis, E. R. 2013. Next-generation sequencing platforms. *Annu Rev Anal Chem* **6**, 287-303.
42. Meissner, A., Gnirke, A., Bell, G. W., Ramsahoye, B., Lander, E. S. and Jaenisch, R. 2005. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res* **33**, 5868-5877.
43. Miller, M. R., Dunham, J. P., Amores, A., Cresko, W. A. and Johnson, E. A. 2007. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Res* **17**, 240-248.
44. Nielsen, R., Paul, J. S., Albrechtsen, A. and Song, Y. S. 2011. Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet* **12**, 443-451.
45. Onteru, S. K., Fan, B., Nikkila, M. T., Garrick, D. J., Stalder, K. J. and Rothschild, M. F. 2011. Whole-genome association analyses for lifetime reproductive traits in the pig. *J Anim Sci* **89**, 988-995.
46. Ozaki, K. and Tanaka, T. 2006. Genome-wide association study to identify single-nucleotide polymorphisms conferring risk of myocardial infarction. *Methods Mol Med* **128**, 173-180.
47. Ozaki, K. and Tanaka, T. 2005. Genome-wide association study to identify SNPs conferring risk of myocardial infarction and their functional analyses. *Cell Mol Life Sci* **62**, 1804-1813.
48. Park, K. D., Park, J., Ko, J., Kim, B. C., Kim, H. S., Ahn, K., Do, K. T., Choi, H., Kim, H. M. and Song, S., et al. 2012. Whole transcriptome analyses of six thoroughbred horses before and after exercise using RNA-Seq. *BMC Genomics* **13**, 473.
49. Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y. and Shen, D., et al. 2012. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490**, 55-60.
50. Qin, Q., Xu, Y., He, T., Qin, C. and Xu, J. 2012. Normal and disease-related biological functions of Twist1 and underlying molecular mechanisms. *Cell Res* **22**, 90-106.
51. Ramos, A. M., Crooijmans, R. P., Affara, N. A., Amaral, A. J., Archibald, A. L., Beever, J. E., Bendixen, C., Churcher, C., Clark, R. and Dehais, P., et al. 2009. Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. *PLoS One* **4**, e6524.
52. Rothammer, S., Seichter, D., Forster, M. and Medugorac, I. 2013. A genome-wide scan for signatures of differential artificial selection in ten cattle breeds. *BMC Genomics* **14**, 908.
53. Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, C. A., Hutchison, C. A., Slocombe, P. M. and Smith, M. 1977. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* **265**, 687-695.
54. Shendure, J. and Ji, H. 2008. Next-generation DNA sequencing. *Nat Biotechnol* **26**, 1135-1145.

55. Sonah, H., Bastien, M., Iquira, E., Tardivel, A., Legare, G., Boyle, B., Normandeau, E., Laroche, J., Larose, S. and Jean, M., et al. 2013. An improved genotyping by sequencing (GBS) approach offering increased versatility and efficiency of SNP discovery and genotyping. *PLoS One* **8**, e54603.
56. Staden, R. 1979. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Res* **6**, 2601-2610.
57. Strillacci, M., Frigo, E., Schiavini, F., Samore, A., Canavesi, F., Vevey, M., Cozzi, M., Soller, M., Lipkin, E. and Bagnato, A. 2014. Genome-wide association study for somatic cell score in Valdostana Red Pied cattle breed using pooled DNA. *BMC Genet* **15**, 106.
58. Surget-Groba, Y. and Montoya-Burgos, J. I. 2010. Optimization of de novo transcriptome assembly from next-generation sequencing data. *Genome Res* **20**, 1432-1440.
59. Swamy, B. P. and Kumar, A. 2013. Genomics-based precision breeding approaches to improve drought tolerance in rice. *Biotechnol Adv* **31**, 1308-1318.
60. Van Tassell, C. P., Smith, T. P., Matukumalli, L. K., Taylor, J. F., Schnabel, R. D., Lawley, C. T., Haudenschild, C. D., Moore, S. S., Warren, W. C. and Sonstegard, T. S. 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nat Methods* **5**, 247-252.
61. Wagner, C. E., Keller, I., Wittwer, S., Selz, O. M., Mwaiko, S., Greuter, L., Sivasundar, A. and Seehausen, O. 2013. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. *Mol Ecol* **22**, 787-798.
62. Wang, Y., Wen, Z., Shen, J., Cheng, W., Li, J., Qin, X., Ma, D. and Shi, Y. 2014. Comparison of the performance of Ion Torrent chips in noninvasive prenatal trisomy detection. *J Hum Genet* **59**, 393-396.
63. Wang, Z., Gerstein, M. and Snyder, M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**, 57-63.
64. Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y. J., Makhijani, V. and Roth, G. T., et al. 2008. The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872-876.
65. Witt, S. H., Kleindienst, N., Frank, J., Treutlein, J., Muhleisen, T., Degenhardt, F., Jungkunz, M., Krumm, B., Cichon, S. and Tadic, A., et al. 2014. Analysis of genome-wide significant bipolar disorder genes in borderline personality disorder. *Psychiatr Genet* **24**, 262-265.
66. Xie, W., Feng, Q., Yu, H., Huang, X., Zhao, Q., Xing, Y., Yu, S., Han, B. and Zhang, Q. 2010. Parent-independent genotyping for constructing an ultrahigh-density linkage map based on population sequencing. *Proc Natl Acad Sci USA* **107**, 10578-10583.
67. You, X., Shu, L., Li, S., Chen, J., Luo, J., Lu, J., Mu, Q., Bai, J., Xia, Q. and Chen, Q., et al. 2013. Construction of high-density genetic linkage maps for orange-spotted grouper Epinephelus coioides using multiplexed shotgun genotyping. *BMC Genet* **14**, 113.
68. Yu, H., Xie, W., Wang, J., Xing, Y., Xu, C., Li, X., Xiao, J. and Zhang, Q. 2011. Gains in QTL detection using an ultra-high density SNP map based on population sequencing relative to traditional RFLP/SSR markers. *PLoS One* **6**, e17595.

초록 : NGS 기법을 활용한 전장게놈에서의 경제형질 관련 유전자 마커 발굴

김정안 · 김희수*
(부산대학교 자연과학대학 생명과학과)

가축의 고 성장률, 강건성, 질병 저항성과 같은 경제적 형질을 발굴하는 것은 매우 중요한 과제이다. 이에 경제적 형질을 발굴하기 위한 방법으로 전통적으로 RFLP, AFLP와 같은 방법이 대두되었으며, 최근 NGS 기법이 발달함에 따라 이러한 경제적 형질을 전장게놈의 수준에서 발굴하려는 노력이 계속되고 있다. 하지만, NGS 기법의 경우 상대적으로 많은 연구 비용이 필요한 실정이다. 이를 극복하기 위한 노력으로써 RNA-seq, RAD-Seq, RRL, MSG, GBS 등과 같은 기법이 활용되고 있다. 본 논문에서는 NGS 기법을 기반으로 한 최근 연구 동향을 확인하고자 하며, 특히 최소의 연구 비용으로 최대의 효과를 낼 수 있는 연구 방법을 소개하는 데 초점을 맞추었다. 또한 이러한 연구 방법이 우수한 경제형질을 가진 가축을 선정하는 데 어떻게 적용될 수 있는지에 대해 토의하였다.