creative commons
COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

석사학위논문

차세대염기서열분석법을 통한 돌연변이감귤 아라온주, 제대온주의 유전체분석

제주대학교 대학원

바이오소재공학과

정 명 언

2020년 2월

제주대학교 중앙도서관
JEJU NATIONAL UNIVERSITY LIBRARY

차세대염기서열분석법을 통한 돌연변이감귤 아라온주, 제대온주의 유전체분석

지도교수 김 인 중

정 명 언

이 논문을 바이오소재공학 석사학위 논문으로 제출함

2019년 12월

정명언 바이오소재공학 석사학위 논문을 인준함

심사위원장 이효연

위 원 김인중

위 원 정용석

제주대학교 대학원

2019년 12월

제주대학교 중앙도서관
JEJU NATIONAL UNIVERSITY LIBRARY

Variant Analysis of Citrus Mutant Lines, Ara unshiu and Jedae unshiu, using Next Generation Sequencing

Myungun Jung
(Supervised by professor In-Jung Kim)

A thesis submitted in partial fulfillment of the requirement
for the degree of Master of Science

2019. 12

Department of Biomaterials Science and Technology
GRADUATE SCHOOL
JEJU NATIONAL UNIVERSITY

제주대학교 중앙도서관
JEJU NATIONAL UNIVERSITY LIBRARY

목 차

제주대학교 중앙도서관
JEJU NATIONAL UNIVERSITY LIBRARY

List of tables

제주대학교 중앙도서관
JEJU NATIONAL UNIVERSITY LIBRARY

List of figures

제주대학교 중앙도서관
JEJU NATIONAL UNIVERSITY LIBRARY

Abstract

Recently, the development of next generation sequencing method (NGS) and cost reduction, researches using genome analysis have been actively conducted, and whole genome re-sequencing that aligns the leads of the sample to the previously coded reference gene. The cost is very low, unlike De novo assembly, which uses the overlap information of leads that sequence new samples without standard genome. In this study, we used the reference genes of the already registered C. unshiu Marc. Cv. Miyagawa-wase(Shimizu T, etal. 2017) and used the control groups Citrus unshiu and mutant Ara-unshiu, Jedae-unshiu overall genom analysis was conducted. In this experiment, the reason of Citrus unshiu already registered on NCBI was analyzed is that the nucleotide sequence may have naturally mutated depending on factors such as environmental effects. As a result, the SNPs were changed by 2,355,501 compared to the reference sequence registered with NCBI, and the insertions had 98,321 deletions with a difference of about 87,696, showing a considerable difference from the reference gene registered with NCBI. Through this, we are performing the registration as a new reference gene. Through these results, this study analyzed SNP and Indel of Ara-unshiu and Jedae-unshiu compared with Citrus unshiu. Sequence of filtered fragment after quality adjustment of FASTQ file reads obtained using Whole genome re-seq instrument using FastQC, PICARD tools after alignment using BWA Samtools Remove duplicates that occur during the sequencing process using, extract SNPs and InDels using Samtools, and analyze and summarize the occurrence region, frequency, and frequency of each DNA polymorphism using the SNPeff tool. And annotation is performed to classify the functions of each SNPs and Indels so that the skins are not

제주대학교 중앙도서관
JEJU NATIONAL UNIVERSITY LIBRARY

smooth, and the valleys are vertically sloping from the nipple and have an uneven appearance. And the content of citrus flavonoids is increased compared to the fruits of C grown under the same conditions, and the contents of each of the components of Hesperidin and Narirutin are also increased compared to the contents of each of the fruits of Gungcheonsaeng grown under the same conditions. J gene related to (Inoue T1, Yoshinaga A., et al. 2015, Susoma Jannat, Md Yousof Ali., Et al. 2016), flowering time is the same as C, but fruit matures later than C, We have developed a molecular marker using ANP and Indel to identify genes for A, which is characterized by a 1 to 3 month delay in pigmentation time compared to C (Fujii H, Ohta S., et. al. (2016).

제주대학교 중앙도서관
JEJU NATIONAL UNIVERSITY LIBRARY

Ⅰ. Introduction

종자관련 산업은 '품종'을 매개로하는 대표적인 지식재산 산업으로 기존 품종에 비해 산업적인 측면에서 우수한 품종을 개발했을 때 오랜 기간 동안 독점적인 권리 획득기회로 인한 고부가가치 유망산업이다. 95년도에 발효된 TRIPs(agreement on Trade-Related Aspects of Intellectual Property Rights) 의 효력으로 WTO 에 가입된 나라는 식물 품종을 특허법 등으로 보호하도록 하는 의무가 부과되었고 이로 인하여 자연스럽게 우리나라 또한 종자산업법 및 식물신품종 보호법 등 이에 관련된 여러 제도들을 정비하였다.(2018 이길우,장인호) 또한 유전자원으로 의약품 및 화장품 같은 연구개발 및 상품화에 따른 이익을 낸다면 해당 유전자원 이용자가 유전자원 제공자와 이익을 공유해야 하는 나고야 의정서가 2010년 생물다양성협약 제 10차 당사국총회에서 채택되었고 우리나라에서 2014년 10월 12일에 발효되었으며 이후 2018년 8월 18일 "유전자원 접근 및 이익공유에 관한 법률"이 전면 시행, 2017년 8월 17일 나고야 의정서 비준서를 유엔 사무국에 기탁해 17일부터 당사국 지위를 받게 됨에 따라 유전자원에 대한 로열티 규정 및 지급에 대한 분쟁의 심화가 예상되었다. 이에 본 연구실에서는 제주특별자치도 농업기술원으로부터 공여 받은 궁천조생(C.unshiu Marc.cv.Miyagawa-wase)의 가지의 눈에 제주대학교 원자력과학기술연구소의 코발트-60(^{60}CO) 방사선 조사시설(선원 C-188,Nordion International Ltd.,Canada)을 이용하여 방사선 선원으로부터 0.68 m의 거리에서 40Gy 조사량으로 22시간 동안 감마선 조사를 통해 돌연변이를 유도한 제대온주와 ^{60}Co의 감마선을 80 Gray 선량으로 조사한 아라온주를 개발하였다. 현재는 대비품종인 궁천조생보다 과실의 성숙시기가 늦고 당도가 높으며 신맛이 적으며 과실의 성숙이 늦어 일반 감귤의 출하시기인 11월 하순부터 12월 중순까지의 출하 집중시기를 피할 수 있어 경제적 가치가 높은 아라온주(Ara-unshiu)와, 마찬가지로 가지 눈에 감마선으로 조사한 후 접목하여 무성번식 시킨후 개화, 착과, 당산도 등이 궁천조생과 매우 닮았으나, 과실의 과피가 매끄럽지 못하고, 플라보노이드 성분(Eom HJ, Lee D., et al. 201

제주대학교 중앙도서관
JEJU NATIONAL UNIVERSITY LIBRARY

6)의 함량이 변화된 새로운 변종 감귤 식물 제대온주(Jedae-unshiu)를 개발하여 특허등록을 마친 상태이다. 최근에는 차세대 염기서열 분석기술의 발전으로 여러 식물체의 유전체 정보가 해독되고 있으며 이를 활용한 마커들의 개발이 활발히 진행 중 이고 이중 본 실험실에서 돌연변이 품종을 만드는데 사용된 궁천조생(C. unshiu Marc.cv.Miyagawa-wase)의 경우 별도의 reference gene 없이 유전체 초안지도를 작성한 De novo assembly가 2017년에 수행된바가 있다(Tokurou shimizu et al., 2017).본 실험실 에서는 신품종 보호를 위하여 작물의 환경적 요인에 크게 영향을 받지 않고 DNA 염기서열의 차이를 기반으로 하는 분자마커 개발이 필요성을 느껴 분석이 끝난 NGS 유전체 데이터를 활용하여 마커를 개발 중이다. 이번 연구에서는 궁천조생, 아라온주, 제대온주 에 대한 Whole Genome Re-sequencing을 수행하였으며 본 실험실에서 제공한 각 시료들로 Sample Preparation을 하여 샘플QC 후 각 분석항목에 맞추어 Library를 제작하고(Saunders HE. 2019) Macrogen 사의 Illumina sequencer을 분석에 이용(Yohe , Thyagarajan 2017,Cheng Y, Jiang S., et al. 2019)하였다.re-sequencing의 경우 염기 서열 해독 시 시퀀싱 기술의 핵심인 리드들의 overlap 정보를 이용한 전체시퀀스를 밝혀내는 De novo assembly 와 달리 이미 해독이 된 reference gene에 샘플의 리드들을 alignment(mapping)하는 것으로, 어떤 유전자가 얼마나 많이 발현되었는지 볼 수 있으며(Jennings LJ, Arcila ME., et al. 2017) 본 연구와 같이 돌연변이 품종과 기존 샘플과의 유전 변이 연구에 주로 이용되며 이러한 Resequencing을 통하여 산출해 낼 수 있는 유전변이중 가장 대표적인 것으로는 SNP(single nucleotide polymorphism)이다. 이렇게 filter 된 SNP 는 돌연변이 형질과 Indel 사이의 분석에 활용될 수 있다. 본 연구에서 진행한 Whole genome resequencing 의 경우 실험실에서 제공한 Genomic DNA 를 이용하여 Sample QC 후 각 분석항목에 맞추어 Library를 제작 후 Raw Data를 산출하였다.

제주대학교 중앙도서관
JEJU NATIONAL UNIVERSITY LIBRARY

Ⅱ. Method & Materials

1. Plant material

각 식물체의 re-sequencing 분석을 위해 서귀포 아열대농업연구소에서 시료를 채취한 뒤 액체질소를 이용해 -80℃에 샘플을 보관하였다.

2. Genomic DNA extraction

Macrogen 사의 sequencing 분석에 필요한 Genomic DNA 추출은 GeneALL사의Exgene GeneALL Plant SV mini,250p kit를이용하였다. 첫번째로deep freezer에보관된 시료를 -196도의 액체질소로 냉각 하여 미세한 분말로 빠르고 완벽하게 파쇄한 후 최대 100mg(wet)의 샘플을2.0ml microcentrifuge tube에 넣는다. 그 다음400ul의CTAB extraction buffer를 넣어주면 버퍼 속의 EDTA가 Chelate 작용을 하여 DNA가 식물 조직으로부터 분리된다. 식물조직이 파쇄 되면 DNase가 세포 내 에서 빠져 나와 DNA를 파괴할 수 있으므로 Chloroform 등의 단백질 비활성화 물질의 처리과정을 거쳐 모든 enzyme 활동을 정지시킨다. 또한 DNA는 염(salt)의 존재 하에 70% EtOH에서 pellet을 형성한다. 이 과정에서 13000rpm에서 원심 분리를 수행하면 엉긴 DNA는 가라앉고 다른 여러 이물질이 상층액 에 남아 있게 된다. 그 후 70ul의 Tris-EDTA buffer에 DNA를 녹여 얻을 수 있다. 이렇게 얻어진 Genomic DNA는 NANO Drop을 통하여 NGS 분석 시 필요한 DNA양을 정량하였다.

제주대학교 중앙도서관
JEJU NATIONAL UNIVERSITY LIBRARY

3. Quality control & trimming process

FASTQ 파일은 DNA 의 염기서열을 나타내는 텍스트를 기반으로 한 염기 데이터 형식이다. NGS 처리 후 해독한 시퀀스를 FASTQ 파일로 저장을 하는데 이를 Raw data 라 부른다. 보통 각 리드의 염기서열에 관한 정보를 처리하는 단계인 Raw data 를 생성 후 참조 염기서열에 read 의 alignment(Reinert K, Langmead B., et al. 2015)후 모든 read 에 대한 처리결과를 통합하는 variant calling 으로 과정이 3 단계가 나뉜다(Muzzey D, Evans EA., et al. 2015). 이 단계의 결과물들은 각각 FASTQ, SAM/BAM, VCF 파일로 산출된다(Fig 1.)(Roy S, Coldren, et al. 2017) 본 연구에서 사용된 Illumina sequencer 의 read type 은 Paired-end reads 이다. 양끝에서 각각 염기서열을 해독 후 리드 당 같은 크기의 FASTQ 파일을 한 쌍 생성하는데 이것을 Paired-end reads 라고 부른다(H.P.J.Buermans, J.T.den Dunnen 2014). 그리고 해독한 염기의 정확도는 Phred score 로 나타내는데(quality score) NGS 데이터의 QC(Quality control)을 하는데 있어서 다양한 툴들이 있다(Katta MA, Khan AW, et al.2015). FastQC 는 biases 나 어떤 problems 가 있는지 확인하기 위한 절차다(El Allali A, Arshad M., et al. 2019). 본 연구에서는 FastQC (V0.11.5)를 사용하였으며 다운로드한(http://www.bioinformatics.babraham.ac.uk/projects/fastqc) 파일을 압축 해제 후 Java runtime environment 를 설치한 뒤 run-fastqc 파일을 실행한다. 그럼 reading 이 되는데 분석결과에 따라 각각 html 로 저장이 되며 정상이면 초록색, 조금 비정상이면 오렌지색, 비정상이면 빨간색으로 표기 된다.(Figure 2.) QC 에는 일반적으로 phred scoring scheme 를 이용하는데 계산법은 $Q\,Phred = -10\ \log 10\, p_{\varepsilon}$ 로 정의되며, 여기서 p_{ε} 는 error rate 이다. $Q\,phred$ 는 일반적으로 0 에서 41 사이의 수치를 나타내며 프레드 수치가 클수록 염기의 정확도는 높고, 당연히 오류율은 그 반대이다. 예를 들어, $Q\,Phred$=20 은 오류율 p_{ε} =1%에 해당하며, 이것은 염기를 잘못 해독할 확률이 1%임을 의미한다. 그리고 Sequencing file 의 adaptor sequence 자르기와 base quality 조절 즉 정확도가 낮은 염기를 FASTQ data 에서 제거하는 작업을 trimming 이라 하는데 이때 사용된 software 는 Trimmomatic (v0.36)이다. 리눅스 지원이 되면 설치 후 vi /etc/bashrcalias trimmomatic='java

제주대학교 중앙도서관
JEJU NATIONAL UNIVERSITY LIBRARY

-jar /파일경로/trimmomatic-0.36.jar' #제일 아래에 쓰고 ESC+:wqsource /etc/bashrc # 동기화 trimmomatic # 실행여부 확인 후 trimmomatic PE -phred33 cell10_1.fastq.gz cell10_2.fastq.gz -baseout cell10_adaptcut ILLUMINACLIP:/run/media/admin/f0bb15f3-2214-43d2-8f4f-895adde38d60/resource/ref_and_tools/tools/trimmomatic/Trimmomatic-0.36/adapters/TruSeq3-PE.fa:2:30:10 SLIDINGWINDOW:4:15 MINLEN:36 LEADING:3 TRAILING:3 으로 실행하면 불완전한 adapter sequences 들을 NGS data 에서 제거할 수 있어 data analysis 의 정확도를 높여준다. 결과적으로 Table 1. 에 나와 있듯이 아라온주의 Q20 값은 99.01 Q30 은 96.64 로 정확도가 매우 높으며 제대온주 또한 Q20 값이 98.94, Q30 값은 96.43 으로 높으며 대조구인 궁천조생(Satsuma)또한 Q20 값이 99.03, Q30 값이 96.69 로 정확도가 매우 높은 것으로 확인되었다.

제주대학교 중앙도서관
JEJU NATIONAL UNIVERSITY LIBRARY

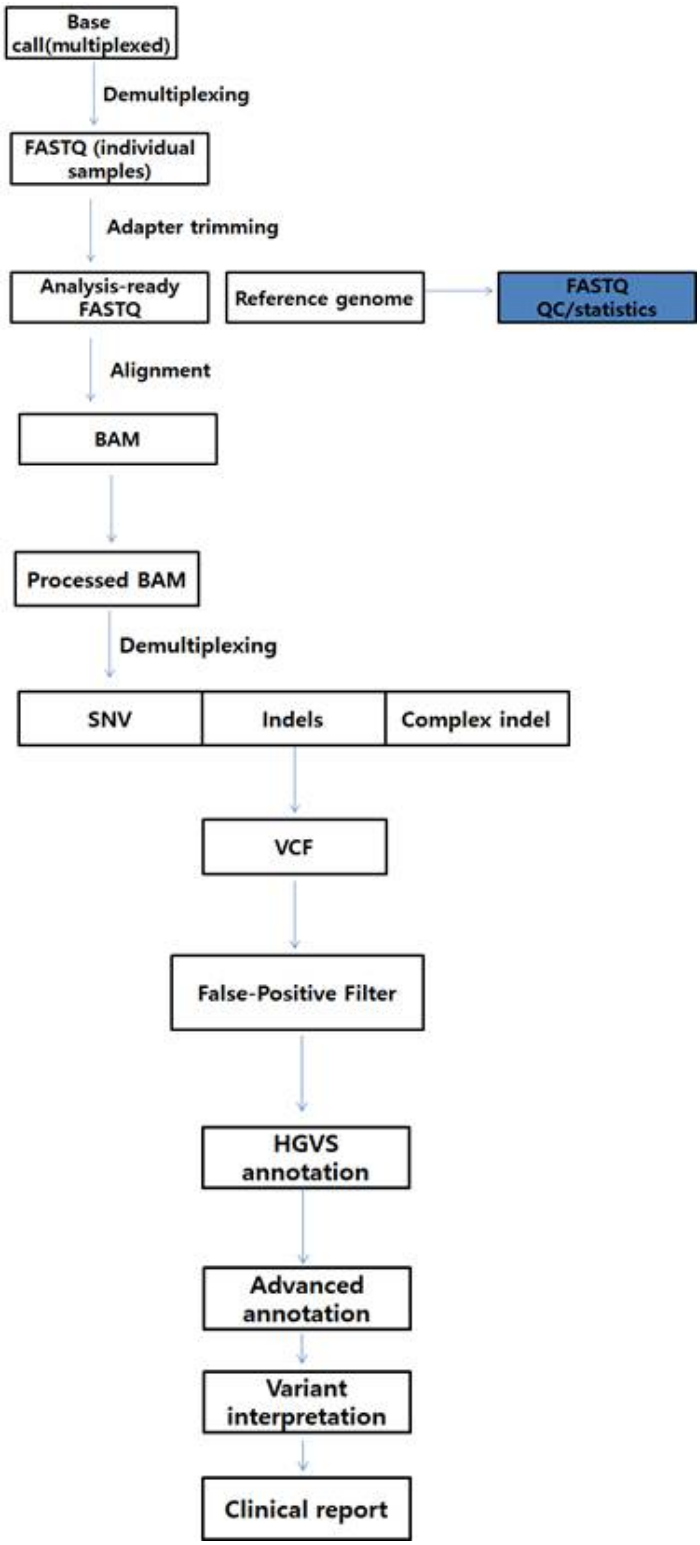

Fig 1.Variant calling pipeline.

제주대학교 중앙도서관
JEJU NATIONAL UNIVERSITY LIBRARY

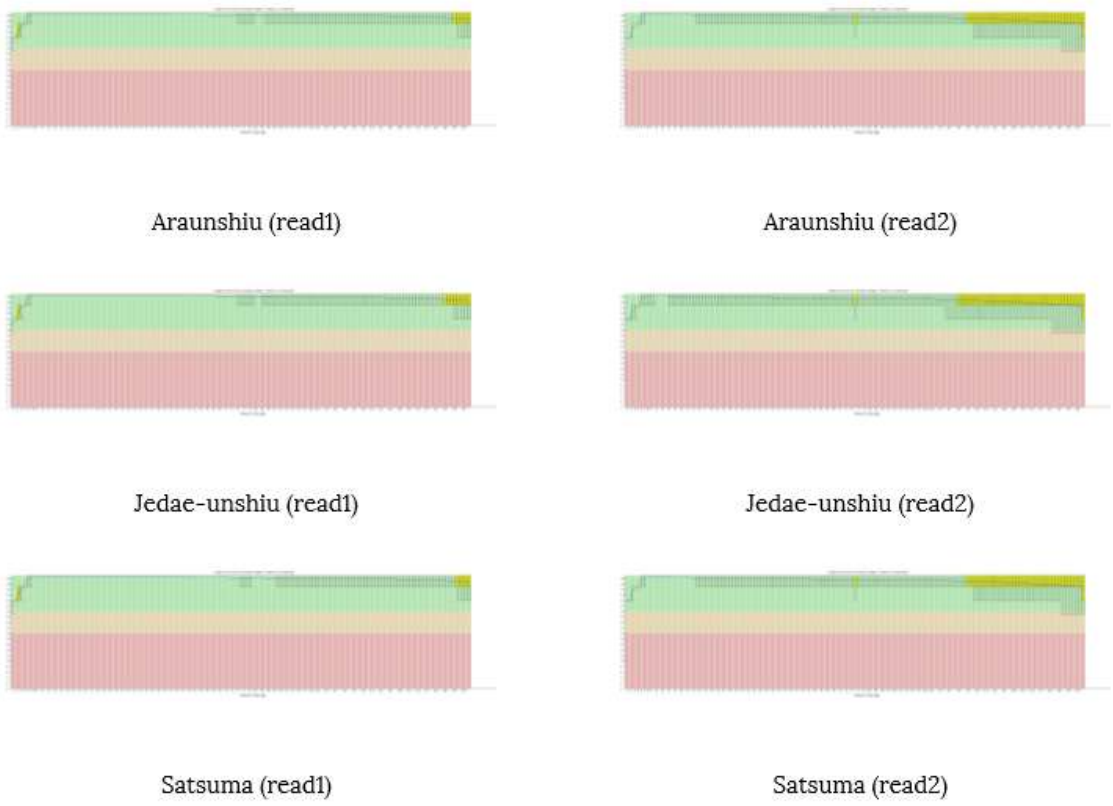

Figure 2. Read quality at each cycle of samples

· The x-axis and y-axis are respectively the number of cycles, and phred quality score
· Phred quality score of 20 means 99% accuracy and reads with quality score over 20 are g enerally accepted as good quality reads.
· Yellow box : Interquartile range (25-75%) of phred score at each cycle.
· Red line : Median phred score at each cycle.
· Blue line : Average phred score at each cycle.
· Upper & Lower whiskers : Point of 10% and 90%.
· Green background : Good quality.
· Orange background : Acceptable quality.
· Red background : Bad quality.

제주대학교 중앙도서관
JEJU NATIONAL UNIVERSITY LIBRARY

Table 1. Raw data and Filtered data Statistics

1) Raw data

Library name	Total read bases (bp)	Total reads	GC(%)	Q20(%)	Q30(%)
Araunshiu	14,231,492,696	94,248,296	38.07	95.41	90.76
Jedae-unshiu	13,143,904,626	87,045,726	37.90	95.13	90.21
Satsuma	13,180,410,688	87,287,488	37.91	95.55	91.00

2) Filtered data

Library name	Total read bases (bp)	Total reads	GC(%)	Q20(%)	Q30(%)
Araunshiu	12,256,490,060	86,959,256	37.17	99.01	96.64
Jedae-unshiu	11,203,039,998	79,817,640	36.99	98.94	96.43
Satsuma	11,402,822,432	80,736,000	37.08	99.03	96.69

· Library name : Sample name.
· Total read bases : Total number of bases sequenced.
· Total reads : Total number of reads. In Illumina paired-end sequencing, this value refers to t he sum of read1 and read2.
· GC(%) : GC content.
· Q20(%) : Ratio of bases that have phred quality score of over 20.
· Q30(%) : Ratio of bases that have phred quality score of over 30.

제주대학교 중앙도서관
JEJU NATIONAL UNIVERSITY LIBRARY

Ⅲ. Results and disscussion

1. Mapping Data Statistics analysis

서열정렬은 생성된 리드의 염기서열을 reference genome과 비교하여 해당 DNA fragment 유전체상 원래 위치를 추정하는 과정이다. 이 정렬에서 가장 큰 문제점은 바로 read length가 짧고 sequence 자체에 반복영역이 있어 Reference sequence의 여러 곳에서 동시정렬위험이 있다. 이에 현재 NGS를 이용한 short reads 로 reference genome에 mapping을 할 때 처리량과 정확도가 매우 높은 alignment software로 BWA(Burrows-wheeler aligner-http://bio-bwa.sourceforge.net/)가 쓰인다. 여기서 말하는 정확도란 리드의 정확한 위치를 찾아내는 능력을 말하며 이 tool은 BWA aln 과 BWA samse/sampe 로 나뉘어 지는데(http://bio-bwa.sourceforge.net/bwa.shtml) BWA aln의 경우 리피런스 시퀀스에 대해 BWT(Burrows-wheeler transform)알고리즘을 이용해 색인을 만든 뒤 만들어진 색인에 input으로 이용되는 reads 와 reference를 비교하여 suffix array를 만드는 과정이다. 그리고 BWA aln 과정에서는 suffix array에 대한 정보만 있기 때문에 여기서 single read 인지 paired read 인지에 따라 BWA samse/sampe를 통해 앞에서 만들어진 suffix array data로 reference 상의 position을 잡아준고 SAM이라는 형식의 파일을 만들게 된다. 분석 방법은 BWA를 설치한뒤 export PATH 명령어로 환경변수를 지정후(export PATH=$PATH:/path/to/bwa-0.5.9)index 파일을 만들고 mapping한다. 전반적인 pre-processing 과정을 거치면 변이 검출 단계까지 간다. 다음 자료(Table 2.)는 pre-processing 의 mapping(alignment)단계에서 BWA-MEM 프로그램을 사용했으며 인풋 파일의 경우 FASTQ 파일이고 아웃 파일의 경우 SAM/BAM 파일이다(Roser LG, Agüero F,. et al. 2019). 이후 sequencing duplication(서열 중복)의 제거를 하는데 PCR 과정에서 발생하는 중복 리드를 표시하거나 제거하는 과정이다. 이후 재정렬(Realignment) 과정을 거치는데 이는 정렬된 리드에서 일치하지 않는 염기가 최소화 되도록 부분적으로

제주대학교 중앙도서관
JEJU NATIONAL UNIVERSITY LIBRARY

재 정렬 하는 과정이며 정렬후 보정하는 과정에서 서열의 삽입, 결손이 빈번하게 발생하는 위치에 대해 정리된 데이터베이스를 활용하여 BAM 파일을 재구성 할 수 있다. SAM은 Text file로 저장되어 있어 바로 열람이 가능하며 BAM binary 형식이기 때문에 바로 열람이 불가능하다. SAM 과 BAM 파일은 모두 염기서열을 저장하며 같은 정보를 갖고 있지만 BAM 의 경우 reference sequence names, length들이 포함되어 있다.

Table 2. Mapped data Statistics

Library name	Ref.Length	Mapped Sites (>=1x)	Total Reads	Mapped Reads	Mapped Bases	Mean Depth
Araunshiu	331,457,320	313,032,762 (94.44%)	86,959,256	78,322,051 (90.07%)	10,772,956,109	32.50
Jedae-unshiu	331,457,320	313,094,620 (94.46%)	79,817,640	72,305,675 (90.59%)	9,986,592,608	30.13
Satsuma	331,457,320	313,011,634 (94.43%)	80,736,000	73,302,553 (90.79%)	10,056,329,767	30.34

· Library name : Sample name.
· Ref.Length : Length of reference genome.
· Mapped Sites : Length of mapped site.
· Total Reads : Number of total read.
· Mapped Reads : Number of reads mapped to the reference.
· Mapped Bases : Number of bases in reads mapped to the reference.
· Mean Depth : Average alignment depth.

제주대학교 중앙도서관
JEJU NATIONAL UNIVERSITY LIBRARY

2. Variants count analysis

본 연구에서 quality 조절 후 필터링 된 각 단편의 염기서열은 BWA Samtools를 사용하여 정렬 화 시킨 후 PICARD tools를 사용하여 sequencing 과정중 발생하는 duplicate를 제거, Samtools 를 사용하여 SNPs 와 InDels를 추출, 분석하였다(Liu Y, Loewer M., et al. 2016). 더 자세히 설명해 보자면 Variant calling 이전 BAM/SAM 파일의 preprocessing으로 정렬된 결과를 염색체 별로 분류하는 정렬 후속과정을 SAMtools(http://samtools.sourceforge.net/)과 Picard로 처리하였다.Variant calling이란 염기 정확도의 보정을 거친 read들의 BAM/SAM 파일을 통합후 SNP/Indel로 확률을 표현하며 Bayesian method로 계산한다(Yuan Ji, Yanxun Xu,, et al. 2011). 사용된 소프트웨어로는 SAMTools 가 있다. SAMtools의 경우 SAMtools를 이용하여 추출할 수 있는 염기서열변이 정보는 단일염기서열변이(SNV: Single Nucleotide Variation)와 짧은 삽입/결실 (Short InDel) 정보등을 산출해내는데 사용하였다. 연구 진행과정은 samtools 프로그램을 이용하여 reference sequence 의 FASTA 파일로부터 reference.fa.fai를 생성하였다. 리눅스 환경에서 SAMtools 사용시 samtools view [in.bam] > [out.sam]standard out으로 나오므로 > 로 받아 BAM 파일을 SAM 파일형식으로 변환 하며 SAM 파일을 BAM 파일형식으로 변환 할 경우 samtools view -Sb [in.sam] > [out.bam]standard out으로 산출 > 로 받아준다.(Manual page from samtools-1.9released on 18 July 2018) 또한 Picard 의 경우 JAVA를 기반으로 하며 파일다운(http://broadinstitute.github.io/picard/) 후 java jvm-args -jar picard.jar PicardToolName OPTION1=value1 OPTION2=value2 와 같이 software를 실행하였다. Fig 3에서 보는바와 같이 아라온주의 경우 SNPs 는 2,367095, Insertions 100,144, Deletions는 89,043 개이며 제대온주는 SNPs 2,355,926 ,Insertions 99,002, Deletions 88,194개이고 궁천조생의 경우 SNPs 2,355,501 ,Insertions 98,321 , Deletions 87,696개로 NCBI Reference gene 대비 큰 차이를 보였다. 여기서 같은 종인 궁천조생의 경우를 보았을 때 NCBI reference gene과 비교하면 SNPs 및 Indel 값이 차이 가 커서 대조구를 본 실험실이 보유한 궁천조생을 Reference gene으로 잡고 SNP, Indel 분석 한 결과(Fig 4.) 아라온주는 SNP 347,818 Insertions 34,958

제주대학교 중앙도서관
JEJU NATIONAL UNIVERSITY LIBRARY

Deletions 34,595 의 변이가 있었으며 제대온주의 경우 SNP 342,789 Insertions 34,476 Deletions 34,345의 차이를 보여주었다(Table 3.). 이 후 gene-ID로 어떤 유전자들이 발현되었는지 알아보기위해 gene ID를 보유한 SNP, Indel만 추출하였고(Table 4.), 확실한 분석을 위하여 homology를 띤 SNP, Indel을 따로 분석하였으며(Table 5), 이후 각 SNP, Indel에서 아라온주와 제대온주가 서로 다른sequence만 추출하여 분석정확도를 높이기 위해 read depth가 30이상인 데이터값만 추출하여 아라온주와(Fig 4.) 제대온주(Fig 5.) 각각의 Annotation을 진행하였다(Iacoangeli A1, Al Khleifat A., et al.2019).

제주대학교 중앙도서관
JEJU NATIONAL UNIVERSITY LIBRARY

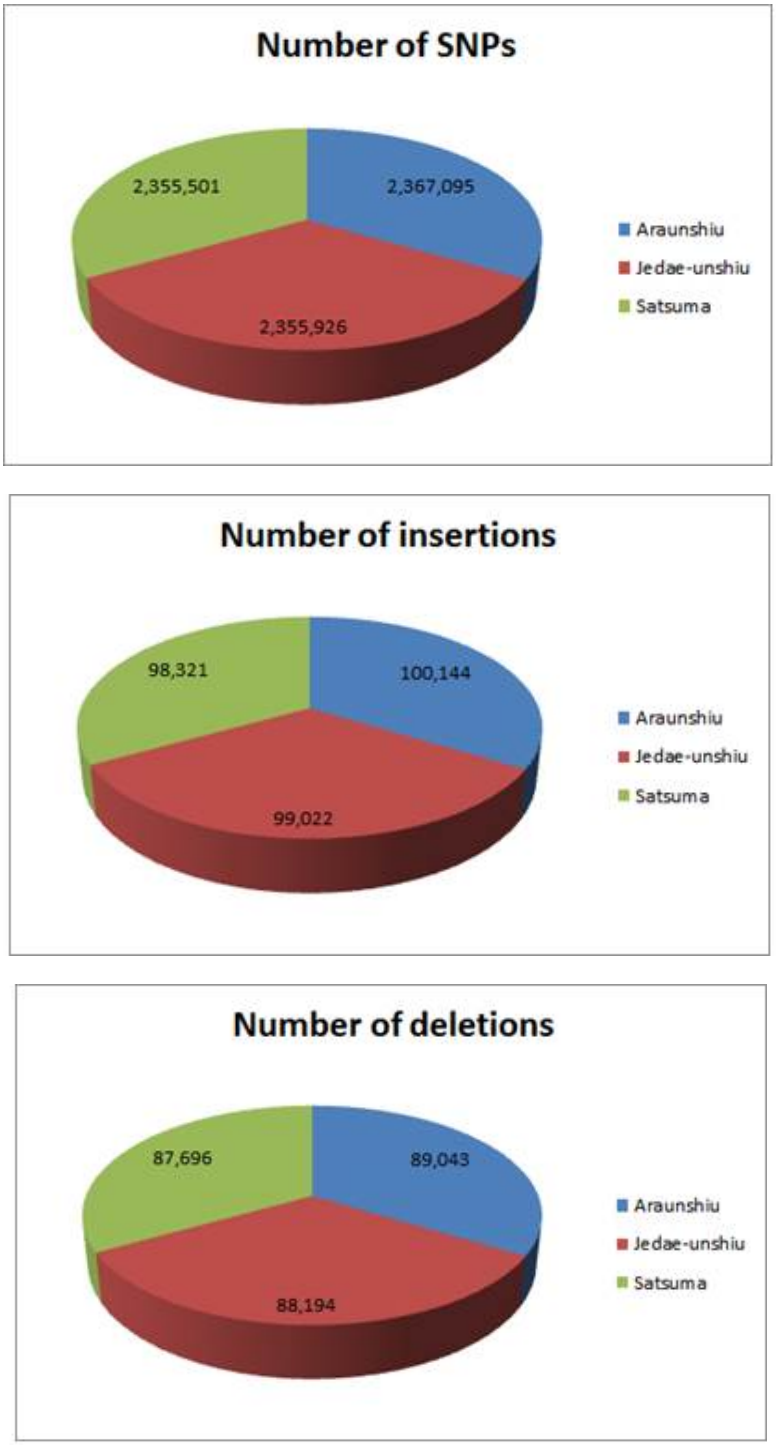

Fig 3. Compared analysis with reference Satsuma SNPs, Indel

제주대학교 중앙도서관
JEJU NATIONAL UNIVERSITY LIBRARY

#Chromosome	Pos					Satsuma					Annotation			
		Alt	Hom/Het	Read Depth	Alt Depth	Alt	Hom/Het	Read Depth	Alt Depth	Gene Name	Gene ID	Start	End	
BDQV01000008.1	651616	aAGaa	hom	30	21	aaa	hom			CUMW_040990	gene4100	649691	659633	DEFINITION LOB domain-containing protein 27 [Citrus clementina]. ACCESSION XP_006439670 VERSION XP_006439670.2 DBLINK BioProject: PRJNA232045 96.92%
BDQV01000078.1	143119	ATTCTTT	hom	47	47	ATTT	hom			CUMW_139820	gene13988	132855	147128	DEFINITION LOW QUALITY PROTEIN: pentatricopeptide repeat-containing protein At4g04790, mitochondrial [Citrus clementina]. ACCESSION XP_006452171 VERSION XP_006452171.2 DBLINK BioProject: PRJNA232045 90.45%
BDQV01000045.1	1024158	ATTTTTT	hom	32	31	ATTTTT	hom			CUMW_112700	gene11274	1020466	1024318	DEFINITION 1-acyl-sn-glycerol-3-phosphate acyltransferase 1, chloroplastic [Citrus clementina]. ACCESSION XP_006453607 VERSION XP_006453607.1 DBLINK BioProject: PRJNA232045 99.14%
BDQV01000038.1	188997	attttttt	hom	35	23	attttttttt	hom			CUMW_103310	gene10333	167945	211069	DEFINITION cytochrome P450 71A1 [Citrus clementina]. ACCESSION XP_006439950 VERSION XP_006439950.1 DBLINK BioProject: PRJNA232045 99.43%
BDQV01000007.1	1904003	attttttttt	hom	33	22	attttttttt	hom			CUMW_038990	gene3900	1902003	1906385	hypothetical protein
BDQV01000033.1	1097657	C	hom	357	351	A	hom			CUMW_097910	gene9793	1097206	1097724	DEFINITION RNA polymerase alpha subunit (chloroplast) [Citrus polytrifolia]. ACCESSION YP_009722803 VERSION YP_009722803.1 DBLINK BioProject: PRJNA595267 96.27%
BDQV01000008.1	865721	caaaaaaa	hom	33	27	caaaaaaaa	hom			CUMW_041310	gene4132	859484	865924	DEFINITION ABC transporter A family member 7-like [Citrus sinensis]. ACCESSION XP_006476636 VERSION XP_006476636.1 DBLINK BioProject: PRJNA225998 99.37%
BDQV01000047.1	329954	caaaaaaaa	hom	36	24	caaaaaaaa	hom			CUMW_113610	gene11365	325725	333833	DEFINITION calponin homology domain-containing protein DDB_G0272472 [Citrus sinensis]. ACCESSION XP_006486946 VERSION XP_006486946.1 DBLINK BioProject: PRJNA225998 85.13%(query cover : 74%)
BDQV01000090.1	514262	cCTttttttt	hom	31	21	cttttttttt	hom			CUMW_148300	gene14837	511405	514873	DEFINITION DNA-directed RNA polymerases II and V subunit 8A [Citrus clementina]. ACCESSION XP_006442881 VERSION XP_006442881.1 DBLINK BioProject: PRJNA232045 100%
BDQV01000048.1	693661	gttttttttt	hom	36	28	gttttttttt	hom			CUMW_115350	gene11539	693487	696046	DEFINITION protein EMBRYO DEFECTIVE 1674 [Citrus clementina]. ACCESSION XP_006443109 VERSION XP_006443109.1
BDQV01000015.1	257865	t	hom	32	28	taaaa	hom			CUMW_062690	gene6270	254867	265514	DEFINITION E3 ubiquitin-protein ligase BRE1-like 1 [Citrus sinensis]. ACCESSION XP_006464461 VERSION XP_006464461.1
BDQV01000097.1	193265	aTGttttttt	hom	30	19	attttttttt	hom			CUMW_151550	gene15163	191113	196568	DEFINITION protein EARLY-RESPONSIVE TO DEHYDRATION 7, chloroplastic [Citrus sinensis]. ACCESSION XP_006466976 VERSION XP_006466976.1 DBLINK BioProject: PRJNA225998 92.96%
BDQV01000103.1	579510	attttttt	hom	32	25	attttttt	hom			CUMW_155910	gene15599	577299	580190	DEFINITION O-acyltransferase WSD1 [Citrus clementina]. ACCESSION XP_006446860 VERSION XP_006446860.2 DBLINK BioProject: PRJNA232045 99.10%
BDQV01000230.1	277693	cttttttttt	hom	34	31	cttttttttt	hom			CUMW_200520	gene20062	277353	277788	hypothetical protein
BDQV01000114.1	499780	Ta	hom	166	160	Taaa	hom			CUMW_162140	gene16222	494777	501597	DEFINITION subtilisin-like protease SBT5.3 [Citrus clementina]. ACCESSION XP_024047172 VERSION XP_024047172.1
BDQV01000529.1	2527	aTc	hom	1023	1023					CUMW_244880	gene24500	1341	2734	DEFINITION acetyl-CoA carboxylase carboxyltransferase beta subunit (plastid) [Citrus reticulata]. ACCESSION YP_009364765 VERSION YP_009364765.1 DBLINK BioProject: PRJNA387851 61%(38.97%)
BDQV01000992.1	8937	aTTTTTttt	hom	30	17					CUMW_267320	gene26745	115	10438	DEFINITION putative disease resistance protein RGA4 [Citrus clementina]. ACCESSION XP_024043067 VERSION XP_024043067.1 DBLINK BioProject: PRJNA232045 98.37%(70%)
BDQV01000638.1	96357	CAT	hom	30	30					CUMW_252690	gene25281	91272	96954	hypothetical protein
BDQV01003266.1	771	G	hom	982	906					CUMW_282570	gene28271	713	1805	hypothetical protein

Fig 4. Araunshiu annotation

제주대학교 중앙도서관
JEJU NATIONAL UNIVERSITY LIBRARY

#Chromosome	Pos					Satsuma					Annotation			
		Alt	Hom/Het	Read Depth	Alt Depth	Alt	Hom/Het	Read Depth	Alt Depth	Gene Name	Gene ID	Start	End	
BDQV01000002.1	2055431	A	hom	5	5	AC	hom			CUMW_010310	gene103(	2055214	2056878	DEFINITION cysteine-rich receptor-like protein kinase 10 [Citrus sinensis]. ACCESSION XP_024950645 VERSION XP_024950645.1 DBLINK BioProject: PRJNA225998 98.41%
BDQV01000012.1	2024315	a	hom	35	35	aac	hom			CUMW_056330, CUMW_056320	gene5633	2024042(	2024547(	DEFINITION importin subunit alpha [Citrus clementina]. ACCESSION XP_006444447 VERSION XP_006444447.2 DBLINK BioProject: PRJNA232045 100%
BDQV01000077.1	137816	AACGCCC	hom	38	38	AA	hom			CUMW_138920	gene1389	127516	137977	DEFINITION probable serine/threonine-protein kinase PBL16 [Citrus clementina]. ACCESSION XP_006434330 VERSION XP_006434330.1
BDQV01000034.1	315149	C	hom	39	37	G	hom			CUMW_098100	gene9812	314246	315757	DEFINITION ammonium transporter 1 member 1 [Citrus clementina]. ACCESSION XP_006420410 VERSION XP_006420410.1 DBLINK BioProject: PRJNA232045 100%(75%)
BDQV01000040.1	45826	C	hom	160	151	G	hom			CUMW_105950	gene1059	45545	46102	DEFINITION NADH-plastoquinone oxidoreductase subunit 7 (plastid) [Citrus reticulata]. ACCESSION YP_009364813 VERSION YP_009364813.1 DBLINK BioProject: PRJNA387851 60.22%(96%)
BDQV01000023.1	64552	ctttttttttt	hom	30	19	ctttttttttt	hom			CUMW_080900	gene8091	61628	65092	DEFINITION ribosome-binding factor PSRP1, chloroplastic [Citrus sinensis]. ACCESSION XP_006493202 VERSION XP_006493202.1 DBLINK BioProject: PRJNA225998 98.12%
														DEFINITION ammonium transporter 1 member 1 [Citrus clementina].
BDQV01000034.1	315151	G	hom	39	37	C	hom			CUMW_098100	gene9812	314246	315757	DEFINITION ammonium transporter 1 member 1 [Citrus clementina]. ACCESSION XP_006420410 VERSION XP_006420410.1 DBLINK BioProject: PRJNA232045 100%(75%)
BDQV01000090.1	549730	gTGTTTtt	hom	30	19	gttttttttt	hom			CUMW_148360	gene1484	549102	549872	DEFINITION probable E3 ubiquitin-protein ligase RNF217 [Citrus clementina]. ACCESSION XP_024045330 VERSION XP_024045330.1 DBLINK BioProject: PRJNA232045 98.08%(65%)
BDQV01000031.1	561803	taaaaaaa	hom	30	22	taaaaaaaa	hom			CUMW_093970	gene9399	557059	563116	DEFINITION hippocampus abundant transcript-like protein 1 [Citrus sinensis]. ACCESSION XP_006486681 VERSION XP_006486681.1 DBLINK BioProject: PRJNA225998 95.07%(95%)
BDQV01000016.1	170264	taaaaaaaa	hom	30	24	taaaaaaaa	hom			CUMW_065660	gene6567	167423	171851	hypothetical protein
BDQV01000001.1	3783474	tTAAaaaa	hom	30	19	taaaaaaaa	hom			CUMW_005340	gene533	3782897	3785999	DEFINITION peptidyl-prolyl cis-trans isomerase CYP18-2 [Citrus clementina]. ACCESSION XP_006430139 VERSION XP_006430139.1 DBLINK BioProject: PRJNA232045 100%
BDQV01000092.1	19565	aTttttttt	hom	35	27					CUMW_148800	gene1488	16869	21556	DEFINITION protein SAR DEFICIENT 1 [Citrus clementina]. ACCESSION XP_006442616 VERSION XP_006442616.1 DBLINK BioProject: PRJNA232045 99.58%
BDQV01000367.1	127116	cCTttttttt	hom	62	50					CUMW_227780	gene2279	126077	129268	hypothetical protein
BDQV01000275.1	294406	cTCTttttttt	hom	48	48					CUMW_211380	gene2114	293436	297166	hypothetical protein
BDQV01000247.1	280134	CTTTTTTT	hom	58	46					CUMW_205120	gene2052	279198	280554	DEFINITION 40S ribosomal protein S8 [Citrus clementina]. ACCESSION XP_006431931 VERSION XP_006431931.1 DBLINK BioProject: PRJNA232045 100%
BDQV01000163.1	6034	gTTttttttt	hom	48	42					CUMW_180000	gene1800	4765	6578	DEFINITION feruloyl CoA ortho-hydroxylase 1 [Citrus clementina]. ACCESSION XP_006445647 VERSION XP_006445647.1 DBLINK BioProject: PRJNA232045 64%(74.3%)
BDQV01000114.1	499780	Tac	hom	139	131					CUMW_162140	gene1622	494777	501597	DEFINITION subtilisin-like protease SBT5.3 [Citrus clementina]. ACCESSION XP_024047172 VERSION XP_024047172.1 DBLINK BioProject: PRJNA232045 95%(75.53%)
BDQV01000328.1	245897	tTGtt	hom	51	37					CUMW_221900	gene2220	245127	246455	hypothetical protein
BDQV01000525.1	35444	CCGG	hom	33	31					CUMW_244500	gene2446	34503	35543	DEFINITION geraniol 8-hydroxylase [Citrus clementina]. ACCESSION XP_006423715 VERSION XP_006423715.1 DBLINK BioProject: PRJNA232045 91.36%
BDQV01000495.1	57104	T	hom	153	137					CUMW_242010	gene2421	56934	57365	hypothetical protein

Fig 5. Jedae-unshiu annotation

제주대학교 중앙도서관
JEJU NATIONAL UNIVERSITY LIBRARY

Table 3. Compared analysis with Satsuma SNPs, Indel

	SNP	Insertion	Deletion
Ara-unshiu	347,818	34,958	34,595
Jedae-unshiu	342,789	34,476	34,345

Table 4. Compared analysis with Satsuma SNPs, Indel(Gene-ID)

	SNP	Insertion	Deletion
Ara-unshiu	39,600	6,582	6,610
Jedae-unshiu	38,384	6,528	6,597

Table 5. Compared analysis with Satsuma SNPs, Indel(Homozygous)

	SNP	Insertion	Deletion
Ara-unshiu	19,926	3,300	3,914
Jedae-unshiu	19,698	3,337	3,983

제주대학교 중앙도서관
JEJU NATIONAL UNIVERSITY LIBRARY

3. Transition and Transversion Information analysis

생물종에서 발생되는 염기서열변이의 종류에 따라 purine(A, G) 의 변이 혹은 pyrimidine(C, T)의 변이를 transition이라 한다. DNA substitution mutation은 DNA sequence의 length 변화가 없이 염기의 조성이 변하는 것을 뜻한다. transversion의 경우 purine과 pyrimidine의 변이를 뜻한다. 생명체는 진화과정 등에서 유전적 변이를 거치게 된다. 모든 종은 서로 다른 transition, transversion 수치를 보이는데 이를 나타내는 지표로써 Ts/Tv ratio가 있다(Yohe S, Thyagarajan B. 2017).purine의 경우 pyrmidine과 유사하나 고리구조를 하나 더 가지고 있다는 특징이 있다. 이 말인 즉, transition에 반해 transversion의 경우 분자의 구조적 변화가 더 심하다고 볼 수 있다(Zhang X, Chen X,, et al.2018). transversion은 경우의 수로 따지면 8가지이며 transition보다 2배 더 많지만 이러한 구조적 특성 때문에 실제 발현 빈도수는 상대적으로 적다. 이러한 특성은 Wobble 현상이라 하여 염기서열의 치환이 아미노산 서열은 일으키지 않는 silent substitution 되기 쉽다(S. Joakim Näsvall, Peng Chen,. et al. 2007). 한 예로 cytosine -> thymine 그리고 guanine -> adenine이 더 빈번하게 일어나는데 이는 그들의 구조와 깊은 관련이 있다. Cytosine의 경우 methyltransferase에 의해 5-methylcytosine으로 구조변환이 되며 deamination에 의해 amine기를 일게 되면 thymine의 형태를 띠게 된다. 이러한 cytosine과 guanine이 많은 영역을 CpG라 칭한다. transition에 의한 silent substitution의 경우 변이가 세대를 거듭하여도 보존되는 경향이 강하며 대표적인 예가 single nucleotide polymorphysm(SNP)이다(Korn, J.M., et al. 2008).일반적인 생물체에서 Ts/Tv는 2에 가까우며 단백질영역에서는 3까지도 관찰된다고 한다(Guo C1, McDowell IC., et al. 2017). 특히 CpG영역의 경우 후성유전체학적(epigenetic)으로 생명현상에 중요한 영향을 미치는 것으로 알려져 있다(Mahamdallie S, Ruark E, et al.2018). Ts/Tv 의 software로 SnpEFF (http://snpeff.sourceforge.net/download.html)를 사용하였다(Cingolani P1, Platts A, et al. 2012). 유전적 변이에 대한 annotation을 수행하는 software이며 주로 염기서열의 변화에 따른 아미노산 서열의 변화를 예측하는 목적으로 사용되고 input file 로 SNP, Indel등의 유전정보를 담고있는 VCF(variant call format) 파일

제주대학교 중앙도서관
JEJU NATIONAL UNIVERSITY LIBRARY

을 output으로 한다(Jiang Y, Wu C. et al. 2019). 사용 방법은 간단하다. SnpEFF에 VCF(variant call format)를 넣어주면 프로그램이 annotation을 진행해준다. 이 프로그램역시 리눅스나 맥환경에서 사용가능하며 $ wget http://sourceforge.net/projects/snpeff/files/snpEff_latest_core.zip 라는 커맨드로 다운이 가능하다. 다운 후 zip 파일을 압축해제 후 SnpEFF에서 제공해주는 데이터 베이스를 $ java -jar snpEff.jar download <Database filename>명령을 내려 다운받고 실행하면 Data 라는 항목이 생기고, 그속에서도 각 DB별 항목이 생긴다. 그뒤 4.2 Command line 명령어 $ java -Xmx4g -jar snpEff.jar -v GRCh37.75 examples/test.chr22.vcf > test.chr22.ann.vcf 를 실행하면 자동으로 SnpEFF가 DB에 관한 유무를 체크후 자동으로 다운로드실행이 된다. 그 후 Annotation Type을 분석하였다(Table 6.).아라 온주의 transition 값은(Fig 6.) 846,417, transversion 은 1,520,678 이고 궁천조생의 경우 transition 값은 842,964 , transversion 은 1,512,537 이며 제대온주의 transition 값은 843,480 , transversion 은 1,512,446 으로 각 샘플들의 Tv의 경우 Ts 에 비해 대략 2배정도 로 산출되었으며 분석샘플 궁천조생과 대비하여 아라온주의 Ts값은 3,453이며 제대의 Ts 값은 516으로 아라온주 에서 큰 차이를 보이고 있으며 제대온주 또한 분명한 차이를 보이고 있다.

제주대학교 중앙도서관
JEJU NATIONAL UNIVERSITY LIBRARY

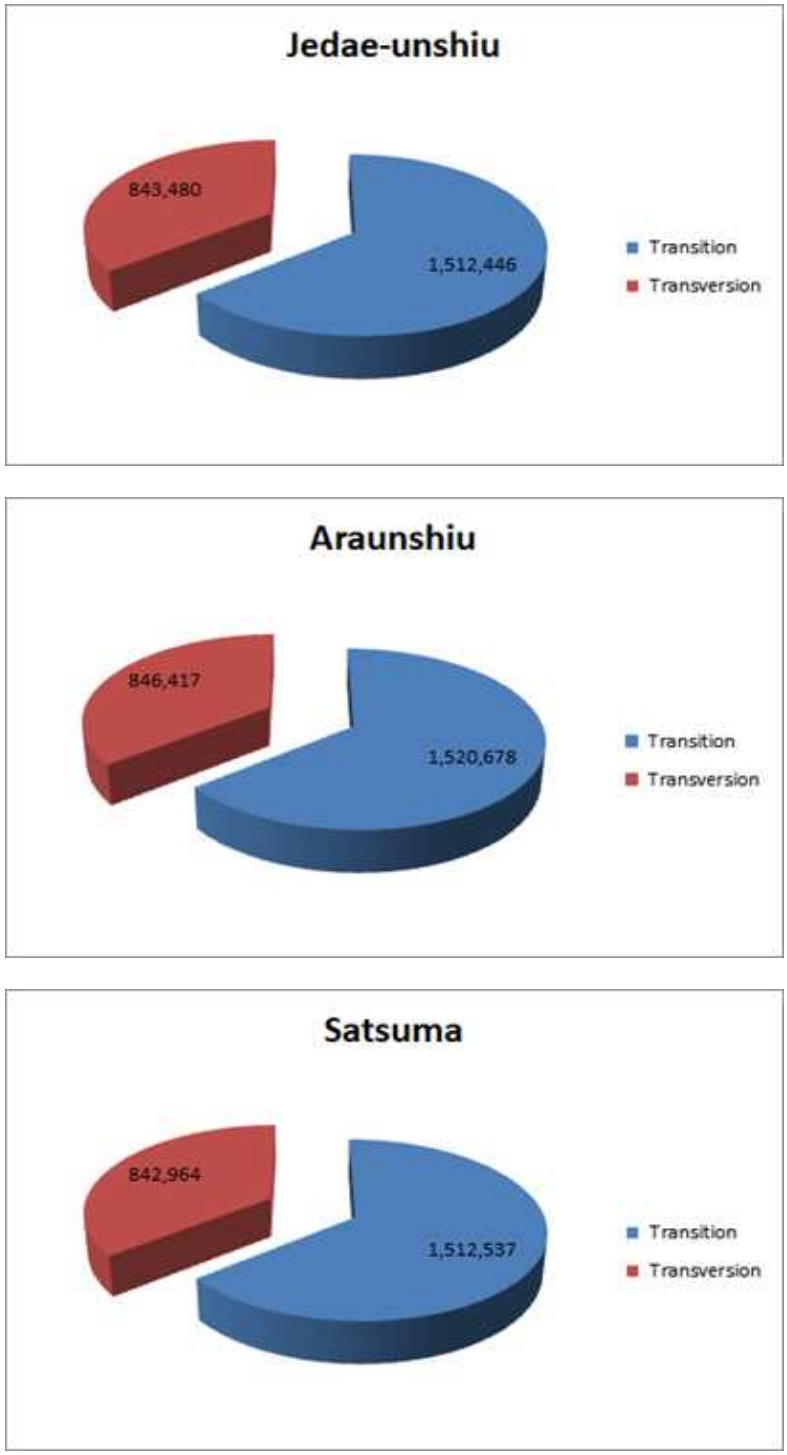

Fig 6. Ts/Tv ratio

제주대학교 중앙도서관
JEJU NATIONAL UNIVERSITY LIBRARY

Table 6. Annotation type count & information

Library name	Type of annotation	Count	Ratio
Araunshiu	intron_variant	298,414	52.09%
	missense_variant	104,174	18.18%
	synonymous_variant	75,828	13.24%
	3_prime_UTR_variant	35,402	6.18%
	5_prime_UTR_variant	18,418	3.21%
	splice_region_variant & intron_variant	9,326	1.63%
	upstream_gene_variant	8,324	1.45%
	downstream_gene_variant	6,424	1.12%
	frameshift_variant	3,037	0.53%
	5_prime_UTR_premature_start_codon_gain_variant	2,669	0.47%
Jedae-unshiu	intron_variant	298,323	52.18%
	missense_variant	103,704	18.14%
	synonymous_variant	75,600	13.22%
	3_prime_UTR_variant	35,394	6.19%
	5_prime_UTR_variant	18,264	3.19%
	splice_region_variant & intron_variant	9,318	1.63%
	upstream_gene_variant	8,323	1.46%
	downstream_gene_variant	6,380	1.12%
	frameshift_variant	2,969	0.52%
	5_prime_UTR_premature_start_codon_gain_variant	2,645	0.46%
Satsuma	intron_variant	298,310	52.17%
	missense_variant	103,818	18.16%
	synonymous_variant	75,623	13.23%
	3_prime_UTR_variant	35,297	6.17%
	5_prime_UTR_variant	18,303	3.2%
	splice_region_variant & intron_variant	9,298	1.63%
	upstream_gene_variant	8,266	1.45%
	downstream_gene_variant	6,358	1.11%
	frameshift_variant	2,994	0.52%
	5_prime_UTR_premature_start_codon_gain_variant	2,651	0.46%

제주대학교 중앙도서관
JEJU NATIONAL UNIVERSITY LIBRARY

Type of annotation	Description	Impact
coding_sequence_variant	The variant hits a CDS.	MODIFIER
chromosome	A large part (over 1% or 1,000,000 bases) of the chromosome was deleted.	HIGH
duplication	Duplication of a large chromoome segment (over 1% or 1,000,000 bases).	HIGH
inversion	Inversion of a large chromoome segment (over 1% or 1,000,000 bases).	HIGH
coding_sequence_variant	One or many codons are changed.	LOW
inframe_insertion	One or many codons are inserted (e.g.: An insert multiple of three in a codon boundary).	MODERATE
disruptive_inframe_insertion	One codon is changed and one or many codons are inserted (e.g.: An insert of size multiple of three, not at codon boundary).	MODERATE
inframe_deletion	One or many codons are deleted (e.g.: A deletion multiple of three at codon boundary).	MODERATE
disruptive_inframe_deletion	One codon is changed and one or more codons are deleted (e.g.: A deletion of size multiple of three, not at codon boundary).	MODERATE
downstream_gene_variant	Downstream of a gene (default length: 5K bases).	MODIFIER
exon_variant	The variant hits an exon (from a non-coding transcript) or a retained intron.	MODIFIER
exon_loss_variant	A deletion removes the whole exon.	HIGH
exon_loss_variant	Deletion affecting part of an exon.	HIGH
duplication	Duplication of an exon.	HIGH
duplication	Duplication affecting part of an exon.	HIGH
inversion	Inversion of an exon.	HIGH
inversion	Inversion affecting part of an exon.	HIGH
frameshift_variant	Insertion or deletion causes a frame shift (e.g.: An indel size is not multple of 3).	HIGH
gene_variant	The variant hits a gene.	MODIFIER
feature_ablation	Deletion of a gene.	HIGH
duplication	Duplication of a gene.	MODERATE
gene_fusion	Fusion of two genes.	HIGH
gene_fusion	Fusion of one gene and an intergenic region.	HIGH
bidirectional_gene_fusion	Fusion of two genes in opposite directions.	HIGH

제주대학교 중앙도서관
JEJU NATIONAL UNIVERSITY LIBRARY

rearranged_at_DNA_level	Rearrengment affecting one or more genes.	HIGH
intergenic_region	The variant is in an intergenic region.	MODIFIER
conserved_intergenic_variant	The variant is in a highly conserved intergenic region.	MODIFIER
intragenic_variant	The variant hits a gene, but no transcripts within the gene.	MODIFIER
intron_variant	Variant hits and intron. Technically, hits no exon in the transcript.	MODIFIER
conserved_intron_variant	The variant is in a highly conserved intronic region.	MODIFIER
miRNA	Variant affects an miRNA.	MODIFIER
missense_variant	Variant causes a codon that produces a different amino acid (e.g.: Tgg/Cgg, W/R).	MODERATE
initiator_codon_variant	Variant causes start codon to be mutated into another start codon (the new codon produces a different AA). (e.g.: Atg/Ctg, M/L (ATG and CTG can be START codons))	LOW
stop_retained_variant	Variant causes stop codon to be mutated into another stop codon (the new codon produces a different AA). (e.g.: Atg/Ctg, M/L (ATG and CTG can be START codons))	LOW
protein_protein_contact	Protein-Protein interacion loci.	HIGH
structural_interaction_variant	Within protein interacion loci (e.g. two AA that are in contact within the same protein, prossibly helping structural conformation).	HIGH
rare_amino_acid_variant	The variant hits a rare amino acid thus is likely to produce protein loss of function..	HIGH
splice_acceptor_variant	The variant hits a splice acceptor site (defined as two bases before exon start, except for the first exon).	HIGH
splice_donor_variant	The variant hits a Splice donor site (defined as two bases after coding exon end, except for the last exon).	HIGH
splice_region_variant	A sequence variant in which a change has occurred within the region of the splice site, either within 1-3 bases of the exon or 3-8 bases of the intron.	LOW
splice_region_variant	A variant affective putative (Lariat) branch point, located in the intron.	LOW
splice_region_variant	A variant affective putative (Lariat) branch point from U12 splicing machinery, located in the intron.	MODERATE

제주대학교 중앙도서관
JEJU NATIONAL UNIVERSITY LIBRARY

stop_lost	Variant causes stop codon to be mutated into a non-stop codon (e.g.: Tga/Cga, */R).	HIGH
5_prime_UTR_premature start_codon_gain_variant	A variant in 5'UTR region produces a three base sequence that can be a START codon.	LOW
start_lost	Variant causes start codon to be mutated into a non-start codon (e.g.: aTg/aGg, M/R).	HIGH
stop_gained	Variant causes a STOP codon (e.g.: Cag/Tag, Q/*).	HIGH
synonymous_variant	Variant causes a codon that produces the same amino acid (e.g.: Ttg/Ctg, L/L).	LOW
start_retained	Variant causes start codon to be mutated into another start codon (e.g.: Ttg/Ctg, L/L (TTG and CTG can be START codons)).	LOW
stop_retained_variant	Variant causes stop codon to be mutated into another stop codon (e.g.: taA/taG, */*).	LOW
transcript_variant	The variant hits a transcript.	MODIFIER
feature_ablation	Deletion of a transcript.	HIGH
regulatory_region_variant	The variant hits a known regulatory feature (non-coding).	MODIFIER
upstream_gene_variant	Upstream of a gene (default length: 5K bases).	MODIFIER
3_prime_UTR_variant	Variant hits 3'UTR region.	MODIFIER
3_prime_UTR_truncation + exon_loss	The variant deletes an exon which is in the 3'UTR of the transcript.	MODERATE
5_prime_UTR_variant	Variant hits 5'UTR region.	MODIFIER
5_prime_UTR_truncation + exon_loss_variant	The variant deletes an exon which is in the 5'UTR of the transcript.	MODERATE

· Type of annotation : Sequence ontology which allows to standardize terminology used for assessing sequence changes and impact.
· Description : Detailed description of the effect (annotation).
· Impact : Effects are categorized by 'impact': {High, Moderate, Low,Modifier}. These are pre-defined categories to help users find more significant variants.
  - HIGH : The variant is assumed to have high(disruptive)impact on the protein, probably causing protein truncation, loss of function or triggering nonsense mediated decay.
  - MODERATE : A non-disruptive variant that might change protein effectiveness.
  - LOW : Assumed to be mostly harmless or unlikely to change protein behavior.
  - MODIFIER : Usually non-coding variants or variants affecting non-coding genes, where predictions are difficult or there is no evidence of impact.

제주대학교 중앙도서관
JEJU NATIONAL UNIVERSITY LIBRARY

Ⅳ. Conclusion

아라온주의 경우 대조구인 궁천조생에 비해 과실의 성숙시기가 1에서 3개월 가까이 늦어지는 특징을 가지고 있으며 제대온주의 경우 과형 및 플라보노이드 함량에 대한 변화를 가지고 있다. 이번 NGS 분석후 아라온주와 제대온주에서만 발견된 SNP,Indel을 Annotation을 통해 돌연변이 감귤의 특징들을 파악하였고 돌연변이원으로 추정되는 유전자 시퀀스를 바탕으로 molecular marker를 제작할 계획이다. 궁천 조생의 경우 NGS 분석 결과 NCBI상에 등록되어있는 Reference gene과 큰 차이를 보이는바 본 실험결과로 나온 시퀀스를 NCBI에 제주 Satsuma로 올릴 예정이며 아라온주의 경우 NGS 결과를 통하여 보유중인 궁천조생과 비교분석해보았을 때 그 결과가 부정확한heterozygous mutant type을 제외한 유전자형이 Homozygous 인 경우 SNP 차이가 커 이를 활용하여 다양한 마커가 나올 것이라 기대가 되며 제대온주 또한 아라온주 보다는 SNPs 및 Indel 수가 적지만 명확히 차이를 보이는바 SNP 및 Indel 로 molecular marker를 제작할 경우 그 확률이 높을 것으로 기대된다. SNP 마커의 경우 Resequencing된 개체를 reference genome과 염기서열 비교를 통해 genome-wide SNP를 찾아내 목적형질 관련 유전자에서 개체 간의 SNP를 비교하고, 개체 간의 차이를 보이는 SNP를 분자마커로 개발하여 특정 품종 구분용 마커 및 중요형질 및 목적형질 식별 마커로 쓰인다. 국내 분자마커의 경우 본 실험실에서 개발한 아라온주,제대온주는 포함하지 않기 때문에 분자마커 개발을 통하여 신품종인 아라온주 및 제대온주 에 대한 품종보호가 가능하다.

제주대학교 중앙도서관
JEJU NATIONAL UNIVERSITY LIBRARY

Ⅴ. Reference

이길우·장인호 (2018). “우리나라 품종 보호의 국가적 차원에서 효율적 관리 방안” (KISTEP issue weekly 통권 제254호)

Cingolani P1, Platts A, et al.(2012) "A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3."Fly (Austin).6(2):80-92.

Cheng Y, Jiang S., et al.(2019) "Whole-Genome Re-Sequencing of Corylus heterophylla Blank-Nut Mutants Reveals Sequence Variations in Genes Associated With Embryo Abortion." Front Plant Sci. 10:1465.

Eom HJ, Lee D., et al.(2016) "Flavonoids and a Limonoid from the Fruits of Citrus unshiu and Their Biological Activity."J Agric Food Chem. 64(38):7171-8.

El Allali A, Arshad M., et al.(2019) “MZPAQ: a FASTQ data compression tool.”Source Code Biol Med. 14:3.

Fujii H, Ohta S., et al. (2016) "Parental diagnosis of satsuma mandarin (Citrus unshiu Marc.) revealed by nuclear and cytoplasmic markers." Breed Sci. 66(5):683-691.

Guo C1, McDowell IC., et al.(2017) “Transversions have larger regulatory effects than transitions.” BMC Genomics. 18(1):394.

제주대학교 중앙도서관
JEJU NATIONAL UNIVERSITY LIBRARY

H.P.J.Buermans, J.T.den Dunnen (2014) "Next generation sequencing technology: Advances and applications" Biochim Biophys Acta. 1842(10):1932-1941.

Inoue T1, Yoshinaga A., et al.(2015) "In situ detection and identification of hesperidin crystals in satsuma mandarin (Citrus unshiu) peel cells." Phytochem Anal. 26(2):105-10.

Jennings LJ, Arcila ME., et al.(2017) "Guidelines for Validation of Next-Generation Sequencing - Based Oncology Panels" 19(3):341-365.

Jiang Y, Wu C., et al.(2019) "GTX.Digest.VCF: an online NGS data interpretation system based on intelligent gene ranking and large-scale text mining." BMC Med Genomics. 12(Suppl 8):193.

Korn, J.M., et al.(2008). "Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat Genet." 40: 1253 - 1260

Katta MA, Khan AW,. et al.(2015) "NGS-QCbox and Raspberry for Parallel, Automated and Rapid Quality Control Analysis of Large-Scale Next Generation Sequencing (Illumina) Data." PLoS One. 10(10):e0139868.

Levy SE, Myers RM (2016). "Advancements in Next-Generation Sequencing." Annu Rev Genomics Hum Genet. 17:95-115.

Liu Y, Loewer M., et al.(2016) "SNVSniffer: an integrated caller for germline and somatic single-nucleotide and indel mutations." BMC Syst Biol. 10 Suppl 2:47.

제주대학교 중앙도서관
JEJU NATIONAL UNIVERSITY LIBRARY

Muzzey D, Evans EA., et al.(2015) “Understanding the Basics of NGS: From Mechanism to Variant Calling.” Curr Genet Med Rep. (4):158-165.

Mahamdallie S, Ruark E, et al.(2018) “The ICR639 CPG NGS validation series: A resource to assess analytical sensitivity of cancer predisposition gene testing.” Wellcome Open Res. 3:68.

Reinert K, Langmead B., et al.(2015) “Alignment of Next-Generation Sequencing Reads.” Annu Rev Genomics Hum Genet. 16:133-51.

Roy S, Coldren, et al.(2017). "Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines: A Joint Recommendation of the Association for Molecular Pathology and the College of American Pathologists." J Mol Diagn. 20(1):4-27.

Roser LG, Agüero F,, et al.(2019) “FastqCleaner: an interactive Bioconductor application for quality-control, filtering and trimming of FASTQ files.” BMC Bioinformatics. 20(1):361.

S. Joakim Näsvall, Peng Chen,. et al.(2007) “The wobble hypothesis revisited: Uridine-5-oxyacetic acid is critical for reading of G-ending codons” RNA. 13(12):2151-64.

Susoma Jannat, Md Yousof Ali., et al.(2016) “Protective Effects of Sweet Orange, Unshiu Mikan, and Mini Tomato Juice Powders on t-BHP-Induced Oxidative Stress in HepG2 Cells” Preventive Nutrition and Food Science 21(3): 208-220.

제주대학교 중앙도서관
JEJU NATIONAL UNIVERSITY LIBRARY

Shimizu T, Tanizawa Y., et al. (2017) "Draft Sequencing of the Heterozygous Diploid Genome of Satsuma (Citrus unshiu Marc.) Using a Hybrid Assembly Approach." Front Genet. 8:180. doi: 10.3389

Sohn JI, Nam JW.(2018) "The present and future of de novo whole-genome assembly." Brief Bioinform. 19(1):23-40.

Saunders HE. (2019) "Liquid Biopsy Quality Control and the Importance of Plasma Quality, Sample Preparation, and Library Input for Next Generation Sequencing Analysis." J Biomol Tech. 30(Suppl):S26.

Yuan Ji, Yanxun Xu,, et al.(2011) "BM-Map: Bayesian Mapping of Multireads for Next-Generation Sequencing Data" Biometrics. 67(4): 1215 - 1224.

Yohe S, Thyagarajan B. (2017) "Review of Clinical Next-Generation Sequencing." Arch Pathol Lab Med. 141(11):1544-1557.

Yohe , Thyagarajan (2017). "Review of Clinical Next-Generation Sequencing." Arch Pathol Lab Med. 141(11):1544-1557

Zhang X, Chen X,, et al.(2018) "Cataloging Plant Genome Structural Variations." Curr Issues Mol Biol. 27:181-194.

Zhang X, Liu B., et al.(2019) "Whole Genome Re-sequencing Reveals Natural Variation and Adaptive Evolution of Phytophthora sojae." Front Microbiol. 10:2792.

Zhao Y, Wang K., et al.(2019) "A high-throughput SNP discovery strategy for RNA-seq data." BMC Genomics. 20(1):160.

제주대학교 중앙도서관
JEJU NATIONAL UNIVERSITY LIBRARY