생산 절차서

(한국인 단염기 다형성 데이터)

문 서 번 호	GRC-MP-011
제 정 일 자	2010.06.01
개 정 일 자	2022.12.06
개 정 번 호	11

결재	구 분	작 성	검 토	승 인
	직 위	연구원	수 석	센터장
	성 명	최재필	이성훈	김정은
	서 명			
	일 자			

표준게놈데이터센터

충청북도 청주시 흥덕구 오송읍
오송생명1로 194-41 기업연구관II 604호

전화 : 043-235-8687 팩스 : 043-235-8688

개정이력

개정번호	제/개정일	주요 제/개정 내용	승인자
00	2010.06.01	• 초기제정	이성훈
01	2010.08.13	• Axiom, NGS 분석내용 추가	이성훈
02	2013.10.13	• 변이체 데이터 생산에 대한 내용으로 전반적 개정	신영아
	2014.01.29	• 오기 교정, 용어 통일 • 전문위원회 및 기술위원회 평가에 따른 부분적 개정	허혜진
03	2014.10.27	• 오기 교정, 목차 수정 • 전문위원회 및 기술위원회 평가에 따른 부분적 개정	박종화
04	2015.09.02	• 주소 변경 • 타액 샘플 및 HiSeq™ 4000 해독기 사용 추가 • 데이터 관리 절차 및 양식 제정	박종화
05	2016.06.17	• 센터명 변경 • 데이터 보급 절차 제정	박종화
06	2017.06.05	• 센터 영문 약칭 변경 • HiSeq X™ Ten 해독기 사용 추가	박종화
07	2017.10.26	• 한국인 질병군 변이체 데이터 생산 절차 추가 • 삽입결실, 게놈배수, 게놈구조 분석 절차 추가	박종화
08	2019.12.10	• 게놈구조 변이 분석 절차 삭제 • 통계 검정 방법 추가	박종화
09	2020.10.01	• MGISEQ-T7 해독기 사용 추가 • 명칭 변경에 따라 TGI를 KOGIC으로 수정 • '10.4.1 단염기 유전형 결정' 항목 내용 추가 • 단염기 유전형 일관성 검사 절차 추가	박종화
10	2021.12.08	• 변이체 분석에 참조서열을 hg19에서 hg38로 변경에 따른 생산 절차 개정 • '10. 변이체 데이터 생산' 항목에 순서도 추가 • '11. 불확도' 내용을 현황에 맞게 수정 • '12. 소급성 확보' 항목 추가 • '13. 데이터 형식' 개정에 따른 내용 변경	박종화
11	2022.12.06	• '11. 불확도' 내용을 현황에 맞게 수정 • '12. 소급성 확보' 내용 수정	박종화
12	2023.09.20	• 측정량 정의, 목적을 분리하여 표시, 기대활용분야 추가 • '23년 내부심사에 따라 KS ID 삭제 • GRC-MP-09 이전에 사용된 데이터 생산 순서도 삭제 • '12 소급성 부분 명확히 명시함	김정은

< 목 차 >

개정이력

목차

한국인 변이체 데이터 생산 절차 모식도

1. 측정량의 정의

2. 기관생명윤리심의위원회(IRB) 통과

3. 참여자 및 시료 확보

4. 참여자의 기초임상정보

5. 해독데이터 생산에 필요한 장비

6. DNA 확보 방법

7. 시료의 정성 및 정량 분석 방법

8. 해독 데이터 생산

9. 공개된 해독 데이터 수집

10. 변이체 데이터 생산

11. 소급성 확보

12. 불확도

13. 데이터 형식

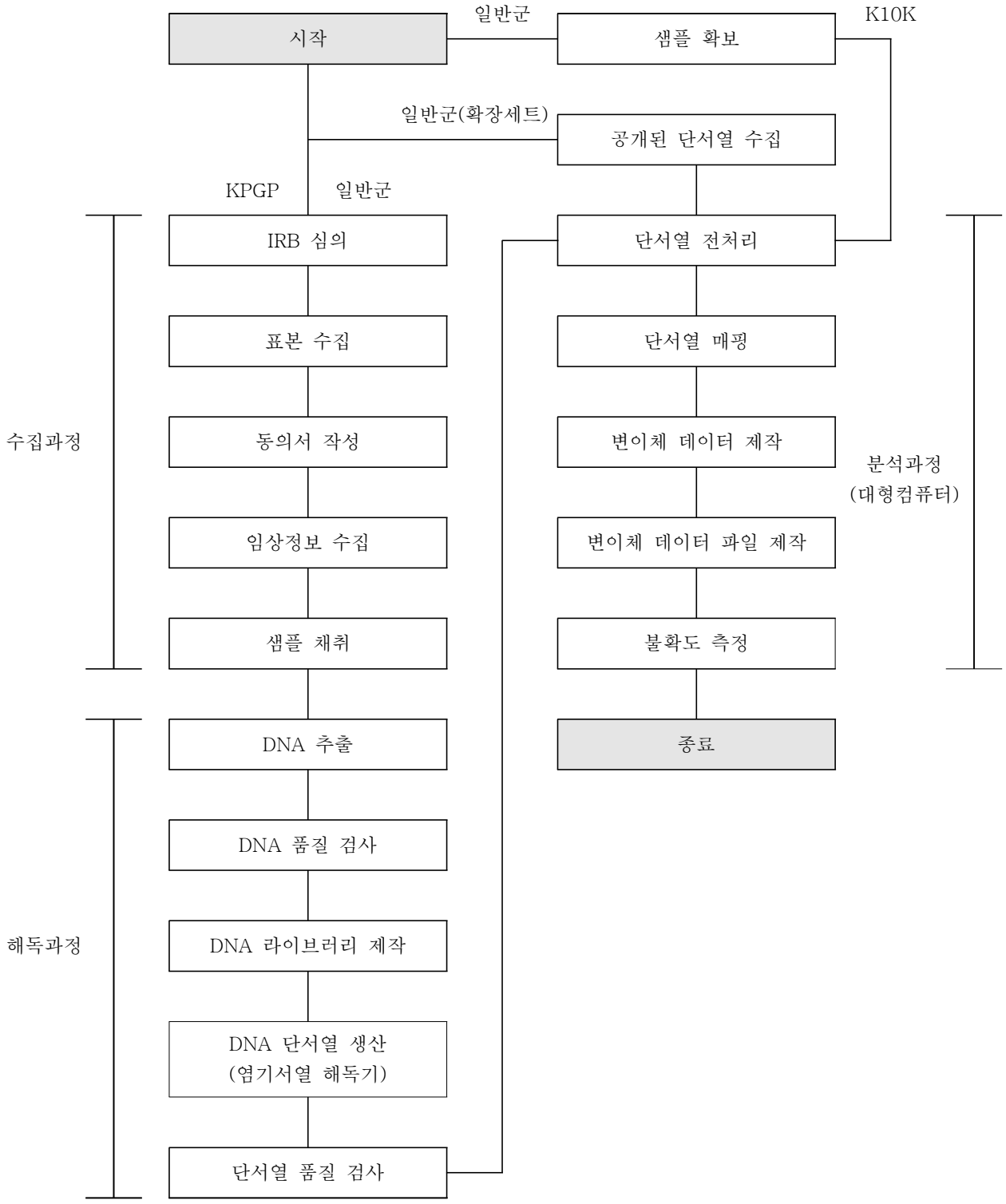
14. 데이터 관리

15. 데이터 보급

16. 참고문헌

17. 양식

한국인 변이체 데이터 생산 절차 모식도



1. 측정량의 정의

○ 측정량

한국인 변이체 참조표준 데이터는 한국인 일반군 및 질병군 집단의 게놈 상에 존재하는 단염기 변이(SNV: Single Nucleotide Variant)정보로 구성한다. 단염기 변이에 대한 측정량은 한국인 일반군 및 질병군 집단에서의 **게놈 상의 단염기 변이 및 그 변이의 빈도**로 한다(표 1). 단염기 변이는 염기를 표현하는 A,T,G,C로 표기하며 변이의 빈도는 0-1 사이의 실수로 표기한다.

측정량	표기
단염기 변이	A,T,G,C 염기
단염기 변이의 빈도	0-1 실수 [0.1234]

표 1. 변이체 참조표준 데이터의 측정량

SNV란 한 개의 염기가 다른 것으로 개개인의 모든 차이를 만드는 원인이며, 개인과 개인 간의 DNA에 존재하는 한 염기쌍(single base-pair variation)의 차이로 DNA sequence 다형성 중에서 가장 많이 존재하는 형태이다. 인간의 경우 대략 1,000 bp에 1개의 SNV가 존재한다고 알려져 있으며, 인종간의 차이를 고려한다면, 이보다 더 높은 변이를 가질 것으로 추정하고 있다. 여러 인종의 같은 게놈위치에서 SNV가 발굴되더라도 인구집단 내에서의 대립유전자(allele)의 빈도(frequency)는 다르게 나타나며, 이것은 인종이 가지는 고유의 특정 형질(trait)과 관련이 있다[1, 2]. 빈도수의 차이는 인종 간 뿐 아니라, 질환자 집단과 정상인 집단에서도 차이를 보이기 때문에 질병과 직접 또는 간접적으로 연관(association)이 있다. 그러므로, 특정 인구집단 내에서의 대립유전자 빈도(allele frequency)를 계산하는 것은 SNV 발굴과 함께 인구집단의 특징, 질환 마커 발굴 등에 매우 중요하다. 최근에는 1000게놈 프로젝트(1000 Genomes Project) [3]를 통해 다양한 인종에 대해 전장게놈시퀀싱(whole genome sequencing) 결과가 공개되고 있어, 인종별 변이체 발굴과 대립유전자 빈도가 계산 가능해졌다[3].

○ 한국인 단염기 다형성 참조표준 목적

다양한 인종과 한국인의 변이체 데이터와 함께 공개된 다양한 질병데이터를 데이터베이스화 한다면 인종 및 질병 특이적인 변이를 규명할 수 있다. 하지만, 이와 같은 국제 컨소시엄을 통한 다양한 인종의 전장게놈서열(whole genome sequence)데이터가 생산되었지만 각 샘플별로 보면 데이터 생산량이 전체 게놈의 약 2~6배수 정도이므로 빈도수가 낮은 변이에 대해서는 정확도가 떨어지고 있는 실정이다. 이러한 이유로 질병관련 마커 개발은 SNP array 기반의 전장게놈연관분석(GWAS: Genome-Wide Association Study)의 결과로부터 확보된 것을 최근까지 주로 활용하고 있다. GWAS를 통해 발굴된 마커는 국립인간게놈연구소(NHGRI: National Human Genome

Research Institute) GWAS catalog를 통해 현재까지 총 11,555개 마커가 공개되고 있고, 대부분이 질병에 직접적으로 영향을 주지 않는 간접적(indirect) 마커이다. 이론적인 모델에 따르면, 'Linkage Disequilibrium(LD) block' 단위에서 질환연관성이 있는 유전형(genotype)을 분석할 경우, 특정 질병과 연관된 단염기 변이 근처에 마커를 해독(fine mapping) 할 수 있게 됨으로써 질병과 연관된 직접적(direct) 유전자를 발견하는 연구를 수행할 수 있으며, 이를 위해서는 차세대 염기서열 해독(NGS: Next Generation Sequencing) 기법을 이용한 민족별 전장게놈(whole genome)에 대한 데이터가 요구된다. 그러므로 한국인의 단염기 변이 빈도(allele frequency)의 표준을 분석하여 이를 기준자료로 활용하는 것이 참조표준의 생산 목적이다.

○ 기대 활용 분야

최근 질병의 원인 유전변이를 발굴하기 위한 전장게놈연관분석이 활발히 진행되면서 사람들의 단염기 변이 다형성(SNP: Single Nucleotide Polymorphism)에 대한 관심이 높아지고 있다. 차세대 시퀀싱 분석 (NGS: Next Generation sequencing) 기술의 발달로 짧은 시간과 저비용으로 전장게놈해독(whole genome sequencing) 및 분석이 가능해지면서 NGS 기반의 GWAS 연구가 활발해졌다. 전 세계적으로 생산된 인간게놈해독 데이터 (예, 1000 genome project)는 다양한 인종의 기원 및 특성을 밝히는데 활용 될 수 있다. 변이체 데이터의 크기가 커지면서, 인종/민족별 단염기 변이의 차이에의한 질병 원인, 약물 반응성 등의 민족별 차이가 명확해 지고 있다. 이러한 결과는 한국인 단염기 다형성 데이터를 기반으로 한 바이오 산업의 활성을 기대할 수 있다.

영국, 미국, 핀란드 등의 바이오 빅데이터 선진국에서는 NGS 기반의 바이오 빅데이터를 구축함으로써 질병의 원인 분석 및 대응에 활용 중이다. 최근 우리나라도 "국가 통합 바이오 빅데이터 구축사업"을 통해 100만명의 한국인 유전체 연구를 수행을 위한 예비타당성조사가 통과됨으로써, 한국형 맞춤 의학(precision medicine)을 준비하고 있기 때문에 한국인 정상인에 대한 단염기 변이 다형성의 활용도는 더 높아지게 되었다.

2. 기관생명윤리심의위원회(IRB) 통과

임상시험을 하는 병원에서 연구계획서 또는 변경계획서, 피보험자로부터 서면동의를 얻기 위해 사용하는 방법이나 제공되는 정보를 검토하고 지속적으로 이를 확인함으로써 임상시험에 참여하는 피험자의 권리, 안전, 복지를 보호하기 위해 시험기관 내에 독립적으로 설치하는 상설위원회를 기관생명윤리심의위원회(IRB: Institutional Review Board)라 한다. 임상시험의 윤리성을 보장하기 위한 가장 기본적이고 필수적인 기구이며, 병원의 의사나 교수 이외에도 기관에 소속되지 않은 제 3자인 종교인, 철학자, 변호사 등도 참여하여 임상시험 계획서의 윤리적 타당성을 심의한다.

기관생명윤리위원회는 윤리적, 사회적으로 심각한 영향을 미칠 수 있는 생명과학기술의 연구, 개발 또는 이용하는 기관으로서 보건복지부령이 정하는 기관(생명윤리법 제9조 각호)에 설치하도록 하여 총괄적인 형태로 설치기관을 규정함으로서 일반법으로서의 기능을 유지할 수 있도록 하였다.

기관생명윤리위원회는 배아연구기관, 유전자은행, 유전자 치료기관 및 보건복지부령에 정하는 기관의 경우에 설치되며(생명윤리법 제9조 제1항 각호), 생명과학연구계획서의 윤리적, 과학적 타당성, 동의의 적법성, 개인 정보의 보호 대책 및 각호의 기관에서 행하는 생명과학기술의 연구, 개발 또는 이용에 관한 사항(생명윤리법 제9조 제2항 각호)을 심의한다. 각호의 기관의 장은 당해 기관에서 행하여지는 생명과학기술의 연구, 개발 또는 이용으로 인하여 생명윤리 또는 안전에 중대한 위해가 발생하거나 발생할 우려가 있는 경우에는 지체 없이 기관위원회를 소집하여 이를 심의하도록 하고, 그 결과를 보건복지부장관에게 보고하여야 한다(생명윤리법 제9조 제3).

3. 참여자 및 시료 확보

한국인게놈프로젝트(KPGP: Korean Personal Genome Project)는 한국인 표준게놈 데이터 구축을 위해 진행되는 게놈해독 프로젝트로, 2006년 세계 최초로 시작된 개인게놈프로젝트(PGP: Personal Genome Project)와 공동연구로 추진되고 있다. KPGP의 참여자는 100% 자발적 참여이며, 참여 절차는 다음과 같다. 1) 참가 동의서 및 유전자 연구 동의서를 작성한다. 2) 생물학에 대한 일정 수준의 교육을 이수하였는지 확인하고, 그렇지 않은 경우에는 소정의 평가를 실시한다. 3) 임상설문지를 작성한다.

한국인만명게놈프로젝트(K10K; Korea 10k Genome Project)는 한국인 1 만 명을 대상으로 게놈을 해독하고 분석하여 오믹스 빅데이터를 구축하고자 하는 프로젝트로써, 이를 통한 게놈 기반 산업의 발전을 목적으로 하고 있다. 울산광역시, 울산과학기술원, 울산대학교병원 주도로 2015년 시작되었으며, KPGP와 마찬가지로 PGP와도 협력하고 있다.

참조표준 등급의 변이체 데이터 생산에 사용할 일반군 표본은 KPGP 또는 K10K의 참여자로 게놈정보 공개에 동의한 자를 대상으로 한다. 표본은 비혈연 관계의 한국인으로, 암 또는 희귀 유전질환의 경력이 없는 자를 대상으로 하여, 지역, 성별, 나이에 상관없이 무작위로 선발한다. 단, 미성년자는 제외한다.

KPGP는 본 생산절차에 따라 시료 확보, 게놈 해독, 데이터 분석을 수행한다. K10K는 자체 생산절차에 따라 시료 확보 및 게놈 해독을 수행한다. 이 외의 프로젝트로부터는 해독 데이터를 수집하여 사용한다. 한국인 변이체 데이터 제작을 위한 샘플로는 「평가 절차서 및 세부평가 기준서」의 평가 기준에 부합하는 샘플을 선별하여 확보하며, KPGP 및 K10K 프로젝트 이 외의 프

로젝트로부터 해독 데이터를 수집하는 경우에는 데이터 분석에 대한 평가 기준 만을 적용한다. 데이터 분석은 본 생산절차에 따라 수행한다. 표준게놈데이터센터는 각 프로젝트로부터 해독 데이터나 분석 데이터를 제공받아 추가 분석 작업을 수행하여 한국인 변이체 데이터를 제작한다.

4. 참여자의 기초임상정보

참여자의 기초 임상정보는 대부분이 설문에 의해 확보되며, 성별, 샘플 수집 당시의 나이, 지역 정보를 빼고는 모두 비공개로 한다. 기초 임상정보는 한국인 게놈에 대한 관련 정보를 확보하기 위한 최소 정보를 포함하고 있다. 참여자로부터 확보된 대부분의 시료는 고유 아이디를 부여하여 익명화를 통해 공개된다.

순번	항목	설명	공개여부
1	ID (identification)	피 측정자의 고유번호	공개
2	Sex	성별 (1: 남자, 2: 여자)	공개
3	Age	나이	공개

표 2. 한국인 참여자의 기초임상정보

5. 해독데이터 생산에 필요한 장비

샘플을 채취하여 해독데이터를 생산하기까지는 각 용도별로 아래 표에 제시된 것과 같거나 대체 가능한 장비가 필요하다.

기자재/시설/장비명	제조사	용 도
Infinite F200	Tecan	샘플 품질관리
NanoDrop	Thermo Fisher Scientific	
Epoch Microplate Spectrophotometer	BioTek	
Qubit Fluorometric Quantification	Thermo Fisher Scientific	
Electrophoresis Power Supplies Horizontal Electrophoresis Systems	Bio-Rad	
Covaris S2	Covaris	라이브러리 제작 시 DNA절단
Mastercycler® nexus	Eppendorf	라이브러리 제작
T100 Thermal Cycler	Bio-Rad	
AllInOneCycler™	Bioneer	
LightCycler® 480II	Roche	라이브러리 품질관리
QuantStudio 6 QuantStudio 7 pro	Applied Biosystems	
Agilent 2100 Bioanalyzer	Agilent	
4200 TapeStation system		
Qubit Fluorometric Quantification	Thermo Fisher Scientific	
HiSeq™ 2000, HiSeq™ 2500 HiSeq™ 4000, HiSeq X™ Ten	Illumina	염기해독
BGISEQ-500, MGISEQ-T7	MGI	

표 3. 게놈해독에 필요한 장비

게놈 해독기 별 성능은 다음의 표와 같다.

해독기	Max. Output	Max. Reads Per Run	Max. Read Length
HiSeq™ 2000	200 Gb	2 billion	2 × 100 bp
HiSeq™ 2500	1000 Gb	4 billion	2 × 150 bp
HiSeq™ 4000	1500 Gb	5 billion	2 × 150 bp
HISeq™ X Ten	1800 Gb	6 billion	2 × 150 bp
BGISEQ-500	520 Gb	2.6 billion	2 × 100 bp
MGISEQ-T7	1.5 Tb	5 billion	2 × 150 bp

* HiSeq™ 2500은 High-Output Mode를 기준으로 함.

표 4. 게놈 해독기 별 성능

6. DNA 확보 방법

Germline 변이 동정을 위하여, 혈액 또는 타액의 DNA를 채취하여 분석에 이용하고, 샘플 채취방법, DNA 추출방법 등을 기록한다. 샘플 채취방법은 현재 다양한 종류의 튜브가 시판되고 있으므로, 튜브의 타입, 사이즈, 방법 등을 명시하여야 한다. DNA 추출 방법은 샘플로부터 DNA를 추출하는 Protocol을 각 단계별로 사용되는 시약과 양, 처리방법과 시간 등을 구체적으로 명시하도록 한다. 특히 DNA 추출을 위해 상용화된 Kit를 사용할 경우 제조사 및 해당 Kit에 대한 정보를 명시하도록 한다.

DNA 추출을 위한 샘플 중 혈액 채취 시 용량은 5 ml 이상 채혈하도록 하며 EDTA tube 사용을 권장한다.

예)

혈액채취 방법
BD Vacutainer / 13 mm * 75 mm / EDTA 튜브 사용
DNA 추출 방법 (QIAamp DNA Blood Mini Kit)

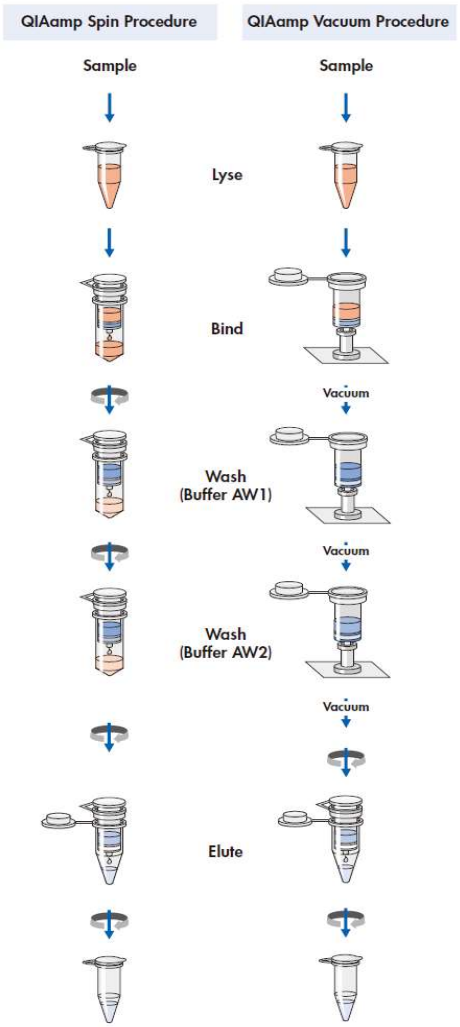


그림 1. DNA 추출 절차

QIAamp DNA Blood Mini Kit은 whole blood, plasma, serum, buffy coat, bone marrow, other body fluids, lymphocytes, cultured cells, tissue에서 genomic, viral, mitochondrial DNA를 포함하는 모든 DNA를 추출할 수 있다. 약 50 Kb 크기까지 분리가 가능하며 대부분은 20~30 kb 사이즈로 잘려진 DNA들이 분리된다. DNA 분리 과정은 아래와 같다.

1) 1.5 mL tube에 proteinase K를 40 µL와 RNase A 8 µL을 넣고 샘플(혈액, 타액)을 400 µL 넣는다.
2) 1)에 lysis buffer 인 AL buffer 400 µL를 넣고 15초간 섞어준다.

3) 56 ℃에서 10분간 반응하고 spin down 한다.
4) 100% 에탄올 400 µL를 넣고 잘 섞어준 다음 QIAamp Mini spin column에 옮긴다.
5) 뚜껑을 닫고 13,500 rpm에서 1분간 원심분리한다.
6) column을 새 collection tube에 옮기고 600 µL Buffer AW1을 넣고 13,500 rpm에서 1분간 원심분리한다.
7) column을 새 collection tube에 옮기고 600 µL Buffer AW2를 넣고 13,500 rpm에서 3분간 원심분리한다.
8) 폐액을 버리고 13,500 rpm에서 1분간 원심분리한다.
9) column을 새 1.5 µL tube에 옮기고 Buffer AE 혹은 멸균된 증류수를 40 µL 넣는다.
10) 상온에서 5분간 incubation한 후 13,500 rpm에서 1분간 원심분리한다.

7. 시료의 정성 및 정량 분석 방법

DNA를 확보하고, 정성 및 정량분석에서 정한 기준치 이상일 경우에만 전장게놈해독(whole genome sequencing)을 해야 분석 결과의 오류를 최대한 줄일 수 있다.

7.1 정성분석

시료의 정성분석으로는 흡광도(260/280 ratio, 260/230 ratio) 측정과 전기패턴을 기준으로 한다. 흡광도 측정은 기기를 사용하며, 기기 제조사의 가이드라인에 따라 진행하고, 흡광도 260/280 ratio가 1.7~2.0, 260/230 ratio가 1.5 이상 되어야 한다. 전기영동은 0.7% agarose gel에 30 ng 시료와 Lambda DNA/HindIII 마커를 함께 로딩한 후, 특정 시스템을 사용하여 이미지를 확인하여, 전기영동 결과 다음과 같이 23 Kb 위치의 밴드가 뚜렷하게 보이는 시료를 선택한다(그림 2).

그림2. 정도 관리에 통과된 시료의 전기영동 사진

7.2 정량분석

시료의 정량분석으로는 형광농도를 측정한다. 형광농도 측정은 double-strand DNA의 농도만을 측정하는 방법으로, 상용화된 kit을 사용하여 제조사의 가이드라인에 따라 진행하며, 형광농도 50 ng/μL 이상, 총량 1 μg 이상 되어야 한다.

8. 해독 데이터 생산

차세대 시퀀싱 (NGS: Next Generation Sequencing) 기술을 이용하여 한국인의 전장게놈(whole genome)에 대한 해독데이터를 생산한다. 장비의 제조업체 별로 제공되는 절차에 따라 해독데이터를 생산하도록 하며, 본 생산 절차서에는 Illumina社와 MGI社의 해독 절차를 명기하였다.

8.1 차세대 염기서열 해독기

차세대 염기서열 해독기는 전장게놈(whole genome)의 영역을 염기해독 함으로써 게놈 전체 영역의 변이체를 찾을 수 있는 장비로서, 0.2~10 Kb까지의 다양한 DNA insert를 가진 paired-end 리드에 대하여 염기서열해독이 가능하다.

그림 3. Illumina HiSeq2000 장비

Performance Parameter	High Output Mode
Read length (bp)	2 × 100
Yield (Gb)	~600
Run time	~11 days
Bases > Q30	> 80%
Reads passing filter	> 90%
Number of flow cells	2
Lanes/flow cell	8
Cluster generation	cBot

표 5. Illumina HiSeq2000 성능에 대한 사양표

그림 4 DNBSEQ-T7 장비

8.2 서열 해독을 위한 DNA 라이브러리 제작 (Ilumina 해독장비)

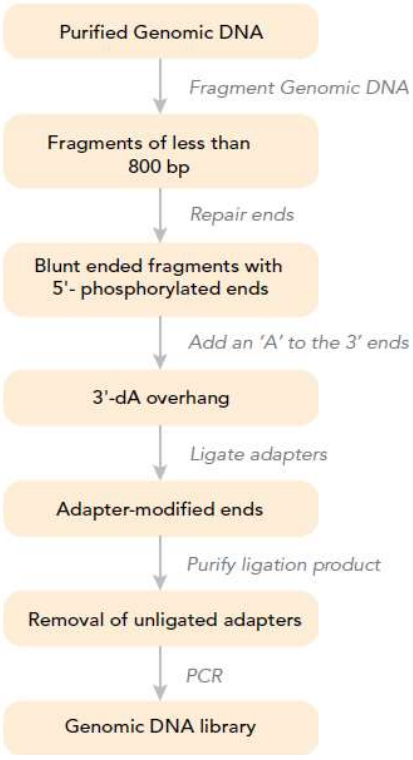


그림 5. 라이브러리 제작 절차

Resequencing을 위한 라이브러리 제작을 위해서 정제된 genomic DNA 1~5 μg을 Covaris를 이용하여 단편화시킨 뒤, BioAnalyzer 2100으로 단편화 크기를 확인한다. 크기를 확인한 후 TruSeq DNA Sample preparation kit을 이용하여 라이브러리를 만든다. 라이브러리는 end repair를 수행하고, AMPure bead로 정제한다. End repair된 DNA의 3'말단에 A를 붙이고 DNA ligation 시스템을 이용하여 시퀀싱 어댑터를 DNA에 ligation 시킨다. 2% agarose gel, 1X TAE, 120 V로 전기영동한 뒤 목표 크기의 밴드를 잘라내어 QIAGEN Minelute gel Extraction Kit로 정제한다. Adaptor-ligated DNA는 PCR에 의하여 증폭된 후, 정제되어 라이브러리의 품질검사를 수행한다. 품질검사를 통과한 라이브러리는 qPCR을 통해 정확한 몰농도가 측정되고, 시퀀싱에 적절한 클러스터를 합성할 수 있는 목표 몰농도로 희석되어 염기해독에 사용된다.

8.3 게놈 염기해독 (Ilumina 해독장비)

라이브러리 품질검사를 통과한 라이브러리는 클러스터 합성을 위하여 cBot기기를 이용하여 클러스터를 합성한다. 합성이 종료되면, 염기해독시약인 SBS kit를 원하는 염기해독 싸이클에 맞는 양만큼 준비한다. 그런 다음 각각의 염기해독 시약은 염기서열 해독기 (HiSeq 2000, HiSeq

2500, 또는 HiSeq 4000) 내의 정해진 위치에 로딩한다. 클러스터가 합성된 flow cell을 기기 내에 위치시킨 후, reagent delivery check를 수행하여 Fluidic system에 문제가 없음을 확인한다. Fluidic system 정상일 경우, 염기해독을 진행한다. Read 1 해독이 종료되면, Read 2를 위한 시약을 장착하고 염기서열 해독을 계속 진행한다. 생산된 염기해독 데이터는 Illumina사의 pipeline 프로그램인 CASAVA 1.8.2를 이용하여 품질검사를 수행한다.

8.4 서열 해독을 위한 DNA 라이브러리 제작 (MGI 해독장비)

DNA를 단편을 100 bp~1000 bp로 조각화 하고, Bead를 사용해 Paired-End 100 또는 Paired-End 150 라이브러리 제작이 가능하도록 300 bp~500 bp로 size selection을 진행한다. Agilent 2100 Bioanalyzer로 선택된 gDNA의 size를 확인하고 Qubit dsDNA HS Assay kit로 농도를 측정한 후 선택된 DNA 단편을 repair하여 blunt end로 만들고, 3' 말단에 dATP를 붙여준다. 그리고 dTTP tailed 어댑터를 DNA 단편 끝에 연결하여 flow cell에 hybridization 될 수 있도록 처리한다. Adapter와 ligation된 product를 PCR을 이용해 증폭시킨다. 최종 PCR 산물의 크기를 Agilent 2100 Bioanalyzer 이용하여 확인하고, Qubit dsDNA HS Assay kit으로 농도 측정 후 1 pmol PCR product에 해당되는 Mass (ng)를 계산한다. 1pmol PCR product를 heat-denature 시켜 생겨난 single-strand molecule를 DNA ligase로 연결시키고 남아있는 linear molecule은 exonuclease로 digest 시켜준다.

8.5 게놈 염기해독 (MGI 해독장비)

DNA nano ball 제작을 위해 40 fmol의 single-strand circular DNA library를 primer hybridization 시키고, DNB Enzyme (Phi29)을 이용한 RCA (Rolling Circle Amplification)를 15분간 수행한 후, Qubit ssDNA Assay kit으로 library 농도를 측정한다. Pooling할 샘플들의 concentration reciprocal의 합계, 평균값 및 400/평균값인 Parameter B를 계산하여, 각 샘플들의 pooling volume를 확정한다. DNB Rapid Reagent kit에 있는 DNB Load Buffer I과 IIII를 이용해 DNB loading Mix를 준비한다. MGIDL-7에 flow cell 과 DNB loader의 바코드를 인식시키서 장착하고, 준비된 DNB loading Mix를 DNB tube hole에 넣은 후, Flow cell loading을 시작한다. Sequencing Cartridge와 Washing cartridge를 DNBSEQ-T7RS에 장착하고 DNB loading이 완료된 flow을 인식시켜 sequencing을 진행한다.

9. 공개된 해독 데이터의 수집

질병군 표본은 논문을 통하여 공개된 해독 데이터로부터 확보한다. 「평가 절차서 및 세부평가

기준서」의 평가기준에 만족하는 게놈 해독 데이터를 대상으로 하여, 해당 표본의 SNP 칩 데이터가 함께 공개된 표본을 수집한다. 확보한 표본에 대해서는 출처 논문에 대한 정보를 기록한다.

10. 변이체 데이터 생산

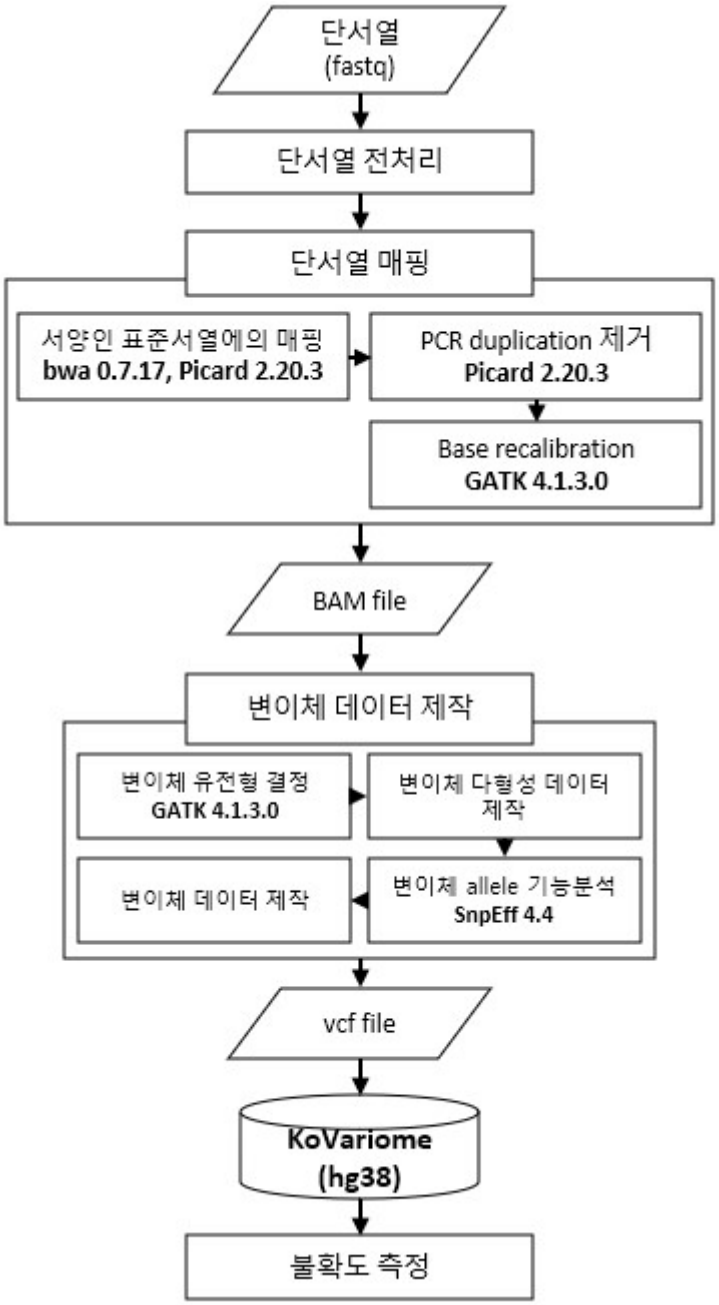


그림 6 변이체 데이터 생산 신버전(hg38) 순서도

변이체 데이터 생산을 위해서는 크게, 해독기계에서 생산된 단서열(short read)의 매핑(mapping), 변이체 발굴, 변이체 annotation의 세 가지 과정이 필요하다. 이 과정은 한국인 일반군과 질병군 집단 모두에게 적용된다. 인간게놈참조서열 hg38 버전이 적용된 개정된 생산절차서(GRC-MP-010)부터는 이전의 생산절차서(GRC-MP-009)와 다르게 단염기와 삽입결실을 통합한 변이체 데이터를 생산하는 절차를 따른다.

10.1 단서열 전처리

생산된 단서열은 그 단서열을 생산한 flow cell의 lane과 index 별로 분류하여 각각의 파일로 저장한다. 단서열의 ID는 {PAIR_ID}#{INDEX}/{FIRST_OR_SECOND}의 형식으로 변경한다. 단서열 ID의 예시는 다음과 같다.

```
@FCB0971ABXX:4:1101:1730:2096#CAGATCAT/1
```

해독기계에서 생산된 단서열로부터 정확한 변이정보를 얻기 위해, 해독된 단서열 중 quality가 떨어지는 read를 제거한다. 이를 위해 Sickle 프로그램 (https://github.com/najoshi/sickle)을 이용한다. 이 프로그램은 각 단서열의 평균 Q score를 계산하여, 이 값이 사용자가 선택한 quality 기준값 미만(Q score < 30)이면 단서열을 끝에서부터 시작하여 하나씩 자르고(trimming) 나머지 염기로 다시 Q score를 계산하여 평균값이 기준 quality 이상이 나올 때까지 이 작업을 반복한다. Trimming은 3’ 말단에 대하여 수행한다. 최종적으로 얻어진 단서열의 길이가 사용자가 선택한 기준값(50 bp) 보다 작을 경우, 해당 단서열을 제거하고 그렇지 않을 경우 해당 단서열의 필터링을 종료한다.

10.2 단서열 매핑

필터링 과정을 거친 단서열을 정확하게 분석하기 위해 서양인 표준 서열에의 매핑(mapping), PCR duplication 제거, base recalibration의 단계를 거친다. 이를 위해 Burrows-Wheeler Aligner (BWA), Picard 및 Genome Analysis Toolkit (GATK) 툴을 사용한다[4].

10.2.1 서양인 표준 서열에의 매핑

한국인 샘플에 대해 샘플 당 전장게놈(whole genome)의 약 20배수 이상에 해당하는 단서열을 서양인 표준 서열(human reference sequence, NCBI build GRCh38, UCSC build hg38)에 효과적으로 매핑하기 위해 Burrows-Wheeler Aligner (BWA) 0.7.17 [5] 버전의 프로그램을 이용

한다. 이때 사용하는 명령어와 옵션은 아래와 같다.

```
bwa mem -M -R "@RGWtID:LANE_Name_LibWtSM:ID_SamplesWtPL:ILLUMINAWtLB:LANE_Name_LibWt" -t [number of thread] [ref.fa] [sample1_R1.trimed.fq.gz] [sample1_R2.trimed.fq.gz]
```

분석의 편의성 및 호환성을 위해, Picard 2.20.3 [6] 버전의 프로그램을 이용하여 매핑 결과 파일인 SAM 포맷 파일을 정렬함과 동시에 BAM포맷으로 변환한다. 이때 사용하는 명령어와 옵션은 아래와 같다.

```
java -XX:ParallelGCThreads=4 -Xmx8g -jar picard.jar SortSam INPUT=[sample1.sam]
OUTPUT=[sample1.sort.bam] VALIDATION_STRINGENCY=LENIENT
SORT_ORDER=coordinate CREATE_INDEX=true TMP_DIR=[/Temp/]
```

10.2.2 PCR duplication 제거

정확한 변이체 발굴을 위해 Picard 2.20.3 버전의 프로그램을 이용하여 PCR duplicate reads를 제거한다. 이 후 생성한 BAM의 index를 생성한다. 이때 사용하는 명령어와 옵션은 아래와 같다. 본 단계는 각각 작성된 BAM 파일을 한 개의 BAM 파일로 통합한 후 수행한다.

```
java -XX:ParallelGCThreads=4 -Xmx16g -jar picard.jar MarkDuplicates INPUT=[sample1.sort.bam]
OUTPUT=[sample1.sort.dedup.bam]                    METRICS_FILE=[marked_DupMetrics.txt]
VALIDATION_STRINGENCY=LENIENT TMP_DIR=[/Temp/]

java -XX:ParallelGCThreads=4 -Xmx16g -jar picard.jar BuildBamIndex
INPUT=[sample1.sort.dedup.bam]
```

10.2.3 Base recalibration

GATK 4.1.3.0 프로그램에서 제공하는 BaseRecalibrator 옵션을 이용하여, 단서열의 성질과 관련하여 사용자가 지정한 다양한 covariate들을 바탕으로 recalibration 테이블을 만들고, 이에 따라 ApplyBQSR 옵션을 사용하여 염기의 quality score를 재보정(recalibration)한다. 사용자가 지정할 수 있는 covariate 들로는 read group, 기존에 보고된 quality score, machine cycle 및 nucleotide context 값들이 있다. 재보정한 결과로 생성한 BAM파일에 대하여 index도 생성한다. 이 단계에서 사용하는 명령어와 옵션은 아래와 같다.

```
gatk --java-options "-XX:ParallelGCThreads=4 -Xmx16g" BaseRecalibrator -R [ref.fa] -I
[sample1.sort.dedup.bam] --known-sites [dbsnp_versionNo.hg38.vcf] -O [sample1recal.table]

gatk --java-options "-XX:ParallelGCThreads=4 -Xmx16g" ApplyBQSR -R [ref.fa] -I
[sample1.sort.dedup.bam] --bqsr-recal-file [sample1.recal.table] -O [sample1.final.bam]

java -XX:ParallelGCThreads=4 -Xmx16g -jar picard.jar BuildBamIndex INPUT=[sample1.final.bam]
```

10.3 변이체 데이터 제작

변이체 데이터는 1) 변이체 유전형 결정, 2) 변이체 다형성 데이터 제작, 3) 변이체 allele 기능 분석, 4) 변이체 데이터 제작의 총 4가지 단계를 거쳐 제작한다. 이 과정에서는 GATK 4.1.3.0 프로그램과 SnpEff-4.3 프로그램을 사용한다. 단염기 변이체 데이터는 참조표준 등급 부여의 대상이다.

10.3.1 변이체 유전형 결정

한국인 샘플 각각에서 GATK 4.1.3.0 프로그램의 HaplotypeCaller 옵션을 사용하여 DNA 위치별 유전형(genotype)을 결정한다.

```
gatk -java-options "-XX:ParallelGCThreads=4 -Xmx32g" HaplotypeCaller -R [ref.fa] -I
[sample1.final.bam] -L chr[no] --genotyping-mode DISCOVERY -stand-call-conf 30 --dbsnp
[dbsnp_versionNo.hg38.vcf] -O [sample1.Chr[no].variant.vcf]
GATK --java-options "-XX:ParallelGCThreads=4 -Xmx32g" HaplotypeCaller -R [ref.fa] -I
[sample1.final.bam] -L chr[no] --genotyping-mode DISCOVERY -stand-call-conf 30 -ERC GVCF --dbsnp
[dbsnp_versionNo.hg38.vcf] -O [sample1.Chr[no].raw.g.vcf]
```

각 유전형에 대한 effective coverage는 다음과 같이 계산한다[7].

$$C(j) = \sum_{i=1}^{depth(j)} (1 - 10^{-m_{ij}/10}) \times (1 - 10^{-q_{ij}/10})$$

m_{ij} 와 q_{ij} 는 i 번째 단서열의 j 번째 위치에서의 각각 mapping quality score와 base-call quality score를 말한다.

10.3.2 변이체 다형성 데이터 제작

10.3.1 단계에서 생산된 데이터를 대상으로, read depth coverage와 effective coverage가 2 이상인 allele로부터 단염기 다형성(SNP) 데이터를 생산한다. Allele의 빈도수는 다음의 수식과 같이 계산한다.

$F_t = N_t / N$, where $t = \{A, T, G, C\}$

F_t는 특정 allele의 빈도수를 의미하고, N_t는 특정 allele의 개수, N은 전체 allele의 개수를 의미한다.

10.3.3 변이체 allele 기능 분석

SNP의 allele 중 인간표준게놈지도의 염기와 다른 allele에 대해서는 annotation을 수행한다. SNP annotation은 SnpEff 4.3 프로그램을 이용하여 다음과 같이 수행한다.

```
java -jar snpEff.jar eff -chr chr1 -c [snpEff.config] -csvStats [chr1.target.csv] -s
[chr1.target.html] hg38 [chr1.target.vcf] > [chr1.snpEff.vcf]
```

10.3.4 변이체 데이터 제작

위 과정을 거쳐 생산된 데이터로부터 변이체 데이터 파일을 작성한다. 이 과정은 in-house 프로그램을 사용한다. SnpEff 프로그램이 제공하는 각 변이 별 기능 구분 정보는 한국인 변이체 데이터에서 정의한 기능 구분으로 변환한다. 변환은 다음의 표에 맞추어 수행한다.

SnpEff의 기능 구분	한국인 변이체 데이터의 기능 구분	설명
INTERGENIC	INTERGENIC	유전자 간 영역에 변이가 존재하는 경우
UPSTREAM	UPSTREAM	유전자의 upstream 영역에 변이가 존재하는 경우
DOWNSTREAM	DOWNSTREAM	유전자의 downstream 영역에 변이가 존재하는 경우
INTRON	INTRONIC	유전자의 인트론 영역에 변이가 존재하는 경우
SPLICE_SITE_DONOR	SPLICE-SITE	유전자의 splice donor site 상에 변이가 존재하는 경우
SPLICE_SITE_ACCEPTOR		유전자의 splice acceptor site 상에 변이가 존재하는 경우
EXON	EXONIC	유전자의 엑손 영역에 변이가 존재하는 경우
UTR_5_PRIME	5'UTR	유전자의 5'UTR 영역에 변이가 존재하는 경우
START_GAINED		
UTR_3_PRIME	3'UTR	유전자의 3'UTR 영역에 변이가 존재하는 경우
SYNONYMOUS_CODING	SYNONYMOUS	유전자의 CDS 영역에 존재하는 변이가 존재하며, 그 변이가 단백질 서열에 변이를 일으키지 않는 경우
SYNONYMOUS_START		
SYNONYMOUS_STOP		
NON_SYNONYMOUS_START	NON-SYNONYMOUS	유전자의 CDS 영역에 존재하는 변이가 존재하며, 그 변이가 단백질 서열에 변이를 일으키는 경우
NON_SYNONYMOUS_CODING		
START_LOST		
STOP_GAINED		
STOP_LOST		
CODON_DELETION	CODON_CHANGE	유전자의 CDS 영역에 변이가 존재하며, 단백질 서열상의 Amino Acid가 추가되거나 없어지는 경우
CODON_INSERTION		
CODON_CHANGE_PLUS_CODON_INSERTION		유전자의 CDS 영역에 변이가 존재하며, 단백질 서열상의 Amino Acid가 변화되고 추가되거나 없어지는 경우
CODON_CHANGE_PLUS_CODON_DELETION		
FRAME_SHIFT	FRAME_SHIFT	유전자의 CDS 영역에 변이가 존재하며, 변이 이후의 단백질 서열이 완전히 바뀌는 경우
START_LOST		
STOP_GAINED		
STOP_LOST		

한국인 변이체 데이터에는 간편한 표시를 위하여 각 SNP에 대해 id를 부여한다. Allele이 한 개인 경우에는 id를 할당하지 않는다. 동일한 게놈 상 위치에는 그 allele의 구성이 다른, 두 가지 이상의 SNP가 존재할 수 있으며, 이들에 대해서는 서로 별도의 SNP id를 부여한다.

ks1

11. 소급성 확보

○ 참조표준 기술평가기준 가이드(NCSRD-2022-01, 제정본) 7.2.1에 따라, 우선적으로 VIM의 Metrological traceability (측정소급성) 정의와 ILAC-P10 문서에 있는 측정결과의 측정소급성에 관한 국제시험기관 인정협의체(ILAC) 방침에 따라 소급성을 확보할 수 있도록 한다.

○ 소급성 확보가 현실적으로 불가능한 경우, 참조표준 기술평가기준 가이드 7.2.2에 따라 7.2.2.3을 참고하여, 측정장비의 Specification을 제시하여 측정결과의 신뢰성을 보장하기에 충분함을 확인 및 제시하고, 실제 측정을 하는 기간에 측정하고자 하는 데이터 생산 과정에 문제가 없었음을 증명할 수 있는 정도관리(또는 자체점검) 결과를 제시한다 [8-13]. 기기실 온도, 습도 점검, 시료 보관, 장비점검에 대해서는 국내 공인기관(유전자검사기관, DTC 인증기관 등)의 관리 기준을 따른다. [1)]

○ 정도관리 결과로는 DNA 흡광도, 단서열 품질, 단변이 비율 자료를 준비하여 제시한다.

○ 유전체 분석 과정에서 사용되는 사람 유전체 표준 서열은 세계적으로 공신력을 갖춘 국제 집합체인 Genome Reference Consortium에서 제공하는 인간게놈서열(GRCh38) [14, 15]을 측정결과가 신뢰할 수 있는 기준에서 비롯되었다는 근거 중 하나로 제시하도록 한다.

○ 데이터가 간접생산 방식인 경우 수집에 이용한 문서, 데이터베이스 등이 충분히 신뢰할 수 있는 것임을 입증하기 위해, 제목/이름, 저자 또는 저널명 및 출판사(기관)등 정보를 확인한다. 논문의 경우, 기본적으로 peer-reviewed 논문이어야 한다.

12. 불확도

12.1 단염기 변이의 불확도

게놈서열은 측정이 아니라 분류를 기반으로 하기 때문에, 측정불확도 표현 지침(Guide to the Expression of Uncertainty in Measurement, GUM)에 따른 불확도 산출에 다소 어려움이 있다. 단염기 변이의 경우에는 실험절차, 장비, 생물정보학적 분석 방법 등에 의해 불확도가 발생할 수 있다. 이에 대하여 각 절차별 오류확률을 종합한 최종적으로 결정된 변이의 오류확률값을 이용하여 측정불확도를 정의한다 [16, 17].

1) 단, 국외에서 생산된 데이터, 다양한 연구실에서 생산된 데이터에 대해서는 데이터의 정도 관리 결과를 제시한다.

최종적으로 결정된 변이의 Phred-scaled quality score(*QUAL*)값은 아래와 같이 정의할 수 있다.

$$QUAL = -10\log_{10}Pr$$

*Pr : 오류 확률(error-probability)

위의 정의에 따라 *QUAL*값을 역산하여 구한 오류 확률은 변이를 구하는 중간 절차의 quality score들을 종합한 일종의 합성 불확도라고 볼 수 있다.
이 값은 결정된 변이(SNP)의 오류율에 관한 측도이며 불확도를 구하기 위한 확률로 사용한다 (NIST Technical Note 1990 참조 [18]). 이때 불확도는 엔트로피 값을 의미한다. 위에서 정의한 확률에 따라 측정된 변이의 해당 위치에서 *QUAL* 값이 q인 경우, 측정불확도(최대엔트로피) H는 아래와 같다.

$$H = -(1-10^{-\frac{q}{10}})\ln(1-10^{-\frac{q}{10}}) - \sum_{k=1}^{3}\frac{10^{-\frac{q}{10}}}{3}\ln(\frac{10^{-\frac{q}{10}}}{3})$$

정리하면

$$H = -(1-10^{-\frac{q}{10}})\ln(1-10^{-\frac{q}{10}}) - 10^{-\frac{q}{10}}(\ln(10^{-\frac{q}{10}} - \ln(3))$$

12.2 단염기 변이 빈도의 불확도

단염기 변이 빈도는 Minor allele frequency(MAF)로서 두번째로 많은 allele의 비율을 의미한다. 따라서 0.5보다 항상 작은 값을 가진다. MAF는 the second most common allele의 비율 p를 추정하는 것이므로, 이 경우 모비율의 추정(Estimatin a Proportion)의 이론에 따라 명확하게 불확도를 정의 할 수 있다 [19].

n개의 샘플에서 측정된 MAF를 $\hat{p}$라고 했을 때, 표준불확도는 표본비율의 표준편차로 아래와 같이 정의한다.

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

이 경우 자유도는 n-1, 포함인자는 z-value 또는 t-value (1.96 for 95% 신뢰수준)이다. 모비율의 추정 이론에 따르므로 표본수가 늘어날 수록 불확도는 작아지게 된다. 또한

$$\hat{p}^2 - \left(z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)^2 = \hat{p}^2 - z^2\frac{\hat{p}(1-\hat{p})}{n} = \frac{\hat{p}}{n}(n\hat{p} - z^2(1-\hat{p}))$$

이므로,

$$z^2 \leq \frac{n\hat{p}}{1-\hat{p}}$$

이면

$$\hat{p} \geq z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

이다. 이 경우 확장불확도에 따른 값의 범위에 음의 영역이 포함될 수 없다. 그러나 $\hat{p} \leq 0.5$이므로, $1-\hat{p} \geq 0.5$ 이고 따라서 $\frac{\hat{p}}{1-\hat{p}} \leq 1$. 즉, 아래와 같다.

$$z^2 \leq \frac{n\hat{p}}{1-\hat{p}} \leq n$$

이때, 95% 신뢰수준 가정시 $z = 1.96$이므로, $z^2 = 3.84$이고 샘플수(n)가 4이상일 경우 확장불확도에 따른 값의 범위에 음의 영역이 포함될 수 없다. t-value를 사용하는 경우, $n = 7$인 경우 자유도 6이고, 이때 $t^2 = 2.447^2 = 5.99 < n$으로서 확장불확도에 따른 값의 범위에 음의 영역이 포함 될 수 없음($n = 6$인 경우 음의 영역 포함 가능성 있음). 따라서 확장불확도의 정확한 정의를 위해 z-value를 사용할 경우 샘플 수 4(2명)이상, t-value를 사용할 경우 샘플 수 7(4명) 이상의 조건을 만족해야 한다.

13. 데이터 형식

13.1 데이터 생산자 정보

순번	분 류	항 목	설 명
1	데이터 생산자 정 보	provider id	데이터를 생산한 곳의 등록번호
2		protocol id	유전자 결정방법 실험 시 사용한 방법의 id
3		contact name	담당자 이름
4		contact email	담당자 이메일 주소
5		contact phone number	담당자 전화번호
6		institution	데이터를 생산한 곳의 이름
7		address	데이터를 생산한 곳의 주소
8		postal code	우편번호

13.2 참조 데이터 정보

순번	분 류	항 목	설 명
1	참조 데이터 정보	genome build	UCSC의 genome build 번호 [hg38]
2		dbSNP build	NCBI의 dbSNP build 번호 [dbSNP155]

13.3 표본 정보

순번	분 류	항 목	설 명
1	표본 정보	sample id	표본의 개체 등록번호
2		country code	국적 (한국의 경우 KR)
3		ethnicity	민족 (순수 한국인의 경우 KR)
4		gender	성별 [1=male(남성), 2=female(여성)]
5		age	나이
6	해독 정보	sequencer	해독기 기종

13.4 변이체 다형성 정보

순번	항목		설명
1	#CHROM		변이가 존재하는 염색체 번호 [chr1]
2	POS		변이의 염색체 상의 위치
3	ID		dbSNP에서 제공되는 rs id, 한국인 변이로 등록된 ks id
4	REF		인간 게놈 참조서열의 allele
5	ALT		한국인 샘플에서의 allele
6	QUAL		ALT에 다형성이 있을 확률
7	FILTER		변이의 Quality가 낮을 경우 [LowQual]로 표기
8	INFO	AC	Allele count
		AF	Allele frequency
		AN	Total number of alleles in called genotypes
		BaseQRankSum	Z-score from Willcoxon rank sum test of Alt Vs. Ref base qualities
		DB	dbSNP Membership
		DP	Approximate read depth
		ExcessHet	Phred-scaled p-value for exact test of excess heterozygosity
		FS	Phred-scaled p-value using Fisher's exact test to detect strand bias
		MLEAC	Maximum likelihood expectation (MLE) for the allele counts
		MLEAF	Maximum likelihood expectation (MLE) for the allele frequency
		MQ	RMS Mapping Quality
		MQRankSum	Z-score From Wilcoxon rank sum test of Alt vs. Ref read mapping qualities
		QD	Variant Confidence/Quality by Depth
		ReadPosRankSum	Z-score from Wilcoxon rank sum test of Alt vs. Ref read position bias
		SOR	Symmetric Odds Ratio of 2x2 contingency table to detect strand bias

		ANN[1)]	Functional annotations
		MU	Minor allele frequency uncertainty
9	FORMAT	GT	Genotype
		AD	Allelic depths for the ref and alt alleles in the order listed
		DP	Approximate read depth
		GQ	Genotype Quality
		PL	Normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specification
		GU	Genotype uncertainty

[1)] Format : 'Allele | Annotation | Annotation_Impact | Gene_Name | Gene_ID | Feature_Type | Feature_ID | Transcript_BioType | Rank | HGVS.c | HGVS.p | cDNA.pos / cDNA.length | CDS.pos / CDS.length | AA.pos / AA.length | Distance | ERRORS / WARNINGS / INFO'

14. 데이터 관리

표본 수집 내역, 샘플 해독 내역, 분석 내역, 변이체 데이터 작성 내역에 대해서는 문서로 기록을 남긴다. 게놈 해독 데이터와 분석 데이터는 분석 서버의 지정된 폴더에 둔다. 각 분석 단계별로 지정된 폴더에 분석 결과파일을 두고, 이에 대한 설명을 기록한 README 파일을 작성한다. README 파일에는 분석 입력파일, 분석방법 등을 기록한다. 샘플 목록, 참조 데이터, 게놈 해독 데이터, 매핑 데이터, 최종 변이체 데이터는 데이터 업데이트 시 지정된 백업서버에 백업한다.

15. 데이터 보급

생산된 변이체 데이터는 국가참조표준센터에 전달하고, KOGIC FTP와 데이터센터의 웹사이트를 통하여 공개한다. 변이체 데이터의 생산에 이용된 게놈 해독 데이터는 KOGIC FTP, KOBIC FTP, NCBI SRA를 통해 공개한다.

데이터 보급 사이트명	웹 주소
국가참조표준센터 웹사이트	http://www.srd.re.kr
표준게놈데이터센터 웹사이트	http://variome.kr
KOGIC FTP	ftp://biodisk.org
KOBIC FTP	ftp://ftp.kobic.re.kr
NCBI SRA	http://www.ncbi.nlm.nih.gov

16. 참고문헌

1. Yoshiura, K. *et al*. A SNP in the ABCC11 gene is the determinant of human earwax type. *Nat. Genet.* **38**, 324-330 **(**2006)
2. Fujimoto, A. *et al*. A scan for genetic determinants of human hair morphology: EDAR is associated with Asian hair thickness. *Hum. Mol. Genet.* **17**, 835-843 (2008)
3. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56-65 (2012)
4. Liu, X. *et al*. Variant Callers for Next-Generation Sequencing Data: A Comparison Study. *PLoS One* **8**, e75619 (2013)
5. Li H. and Durbin R. Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics 25:1754-60 (2009)
6. http://broadinstitute.github.io/picard
7. Gronau, I. *et al.* Bayesian inference of ancient human demography from individual genome sequences. *Nat. Genet.* **43**, 1031-1034 (2011)
8. Krusche, P., et al., Best practices for benchmarking germline small-variant calls in human genomes. Nat Biotechnol, 2019. 37(5): p. 555-560.
9. Zook, J.M., et al., An open resource for accurately benchmarking small variant and reference calls. Nat Biotechnol,2019. 37(5): p. 561-566.
10. Mattocks, C.J., et al., A standardized framework for the validation and verification of clinical molecular genetic tests. EurJ Hum Genet,2010. 18(12): p. 1276-88.
11. Gargis, A.S., et al., Assuring the quality of next-generation sequencing in clinical laboratory practice. Nat Biotechnol, 2012. 30(11): p. 1033-6.
12. Roy, S., et al., 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, 2018. 20(1): p. 4-27.
13. Schrijver, I., et al., Opportunities and challenges associated with clinical diagnostic genome sequencing: a report of the Association for Molecular Pathology. J Mol Diagn,2012. 14(6): p. 525-40
14. Church, D.M. et al. Extending reference assembly models. *Genome Biol* **16**, 13 (2015)
15. Zheng-Bradley, X. et al. Alignment of 1000 Genomes Project reads to reference assembly GRCh38. *Gigascience* **6**, 1-8 (2017)
16. Ewing B; Hillier L; Wendl MC; Green P. Base-calling of automated sequencer traces using phred. I Accuracy assessment. Genome Res 8, (3) 175-185 (1998).
17. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error

probabilities. *Genome Res* **8**, 186-94 (1998).

18. Antonio Possolo Simple guide for evaluating and expressing the uncertainty of NIST measurement results. NIST Technical Note 1900. http://dx.doi.org/10.6028/NIST.TN.1900
19. Walpole R. E., Myers R. H., Myers S. L., Ye K. Probability and Statistics for Engineers and Scientists 9th edition PEARSOM, ISBN 978-0-321-62911-1.1

17. 양식

GRC-MP-CF-003 한국인 일반군 단염기 표본 측정정보
GRC-MP-AF-004 한국인 일반군 단염기 표본 측정결과
GRC-MP-RF-001 변이체 데이터 보급 내역
GRC-MP-DF-001 변이체 데이터 제공 내역

* 양식 안에 기재된 붉은 글씨 내용은 예시를 표현함

GRC-MP-CF-003

한국인 일반군 단염기 표본 측정정보

번호	표본 ID	등록 년도	국가 코드	지역	성별	연령 (yr)	샘플 종류	DNA 추출량(㎍)	DNA 흡광도		해독기
									260/280 ratio	260/230 ratio	
1	KPGP-00001	2014	KR	대전	여	68	혈액	342	1.8	.	HiSeq2000
2	KPGP-00002	2014	KR	대전	남	74	혈액	196	1.75	.	HiSeq2000
3	KPGP-00006	2014	KR	경기도	남	45	혈액	318	1.84	.	HiSeq2000
4	KPGP-00032	2014	KR	경기도	남	35	혈액	333	1.74	.	HiSeq2000
5	KPGP-00033	2014	KR	경기도	여	31	혈액	504	1.8	.	HiSeq2000
6	KPGP-00039	2014	KR	경기도	여	31	혈액	303.8	1.8	2.28	HiSeq2500
7	KPGP-00056	2014	KR	경기도	남	30	혈액	311	1.79	.	HiSeq2500
8	KPGP-00086	2014	KR	대전	남	33	혈액	116.62	1.82	2.62	HiSeq2500
9	KPGP-00088	2014	KR	경기도	여	20	혈액	273	1.73	.	HiSeq2000
10	KPGP-00090	2014	KR	경기도	남	23	혈액	234	1.86	.	HiSeq2000
11	KPGP-00117	2014	KR	대전	남	28	혈액	369	1.83	2.37	HiSeq2000
12	KPGP-00120	2014	KR	경기도	남	26	혈액	43.15	1.81	2.11	HiSeq2000
13	KPGP-00121	2014	KR	경상북도	여	24	혈액	454	1.8	2.26	HiSeq2000
14	KPGP-00122	2014	KR	서울	여	24	혈액	11.75	1.73	1.58	HiSeq2000
15	KPGP-00124	2014	KR	부산	남	27	혈액	349	1.79	2.4	HiSeq2000
16	KPGP-00125	2014	KR	전라북도	남	28	혈액	31.6	1.71	2.43	HiSeq2000
17	KPGP-00127	2014	KR	강원도	남	31	혈액	291	1.8	2.39	HiSeq2000
18	KPGP-00128	2014	KR	인천	여	22	혈액	146	1.84	2.55	HiSeq2000
19	KPGP-00129	2014	KR	서울	남	46	혈액	343	1.81	2.36	HiSeq2000

- 등록년도: 기반 프로젝트로부터 표본 및 샘플 정보를 제공받는 년도. 샘플 채취년도를 의미하지 않음.
- 지역: 기반 프로젝트 참여 당시의 주소지
- 연령: 샘플을 채취할 당시의 표본 연령
- 이전 기록 대비 교정된 부분은 밑줄 표시함
- 이전 기록 대비 새로 추가된 샘플은 굵은 글씨로 기록함

교 정 내 역

<2017년 12월 14일>

- 부적합 샘플 삭제 : KPGP-00317 (매핑률 미달)

최종기록일: 2015년 8월 19일
최종기록자: 김 창 수 (서명)

GRC-MP-AF-004

한국인 일반군 단염기 표본 측정결과

번호	표본 ID	Depth (x)	Q Score				Read 길이				Mapping Rate (%)	SNV 비율 (He/Ho)	Ts/Tv 비율
			최소		평균		최소		평균				
			R1	R2	R1	R2	R1	R2	R1	R2			
1	KPGP-00001	28.20	23	23	38	37	50	50	89	88	99.64%	1.44	1.97
2	KPGP-00002	28.38	23	23	37	37	50	50	89	88	99.77%	1.43	1.98
3	KPGP-00003	25.76	23	23	36	35	50	50	88	87	99.52%	1.43	1.98
4	KPGP-00004	30.13	24	23	37	36	50	50	88	86	99.48%	1.43	1.97
5	KPGP-00005	29.25	24	23	37	36	50	50	88	87	99.51%	1.44	1.97
6	KPGP-00006	29.80	24	23	37	37	50	50	99	99	99.67%	1.45	1.99
7	KPGP-00007	31.74	25	23	37	37	50	50	99	99	99.24%	1.47	1.96
8	KPGP-00008	29.26	25	23	37	37	50	50	99	99	99.25%	1.46	1.97
9	KPGP-00009	27.36	24	22	37	36	50	50	88	88	99.18%	1.39	1.98
10	KPGP-00010	26.38	22	23	36	36	50	50	88	88	99.25%	1.46	1.98
11	KPGP-00011	30.82	24	23	36	36	50	50	87	87	99.61%	1.40	1.98
12	KPGP-00012	31.74	24	23	36	36	50	50	88	87	99.61%	1.43	1.98
13	KPGP-00013	31.87	24	23	36	36	50	50	87	87	99.63%	1.45	1.98
14	KPGP-00014	26.29	23	23	36	36	50	50	86	85	99.36%	1.39	1.99
15	KPGP-00015	31.07	24	23	36	36	50	50	87	87	99.62%	1.40	1.98
16	KPGP-00016	32.20	24	23	36	36	50	50	88	87	99.55%	1.40	1.99
17	KPGP-00017	32.31	24	23	36	36	50	50	88	87	99.50%	1.42	1.98
18	KPGP-00018	31.76	23	23	36	36	50	50	87	87	99.62%	1.41	1.99
19	KPGP-00019	29.23	24	23	36	36	50	50	87	87	99.59%	1.40	1.99

- C: Clean up
- 이전 기록 대비 교정된 부분은 밑줄 표시함
- 이전 기록 대비 새로 추가된 샘플은 굵은 글씨로 기록함

교 정 내 역

<2019년 12월 10일>

- 공개 가능 여부 열을 추가함

최종기록일: 2020년 8월 19일

최종기록자: 김 창 수 (서명)

GRC-MP-RF-001

변이체 데이터 보급 내역

변이체 데이터 명	데이터 종류	UNIST KOGIC FTP	NCBI SRA	KOBIC FTP	OpenKPGP 웹사이트	데이터센터 웹사이트	담당자
VariomeData.0.5	샘플정보	-	2017.03.03	-	2017.02.23	-	김창수
	단서열	2017.02.23	2017.03.03	2017.03.22	-	-	
	변이체	2016.11.09	-	-	-	-	

- 이전 기록 대비 변경된 부분은 밑줄 표시함
- 이전 기록 대비 새로 추가된 샘플은 굵은 글씨로 기록함

최종기록일: 2017년 5월 15일

최종기록자: 김 창 수 (서명)

GRC-MP-DF-001

변이체 데이터 제공 내역

일자	데이터 종류	요청자 소속	요청자	제공자
2018.11.27	VariomeData.0.6	대한민국 임시정부	이봉창	안창호
2019.03.19	KPGP 30 명	.	윤봉길	안창호
2019.07.11	KPGP WGS	.	안중근	안창호