한국인 변이체 데이터 평가 절차서

문 서 번 호	GRC-EP-008
제 정 일 자	2010.06.01
개 정 일 자	2019.12.10
개 정 번 호	08

결재	구 분	작 성	검 토	승 인
	직 위	선 임	책 임	센터장
	성 명	조성웅	김정은	박종화
	서 명			
	일 자			

표준게놈데이터센터

충청북도 청주시 흥덕구 오송읍
오송생명1로 194-41 기업연구관II 6층

전화 : 043-235-8687 팩스 : 043-235-8688

개정이력

개정번호	제/개정일	주요 제/개정 내용	승인자
00	2010.06.01	• 초기 제정	이성훈
01	2013.10.13	• 변이체 데이터 생산에 대한 내용으로 전반적 개정	신영아
02	2014.01.21	• 용어 교정 • 전문위원회 및 기술위원회 평가에 따른 부분적 개정	허혜진
03	2014.09.22	• 오기 교정 • 전문위원회 및 기술위원회 평가에 따른 부분적 개정 • 양식 및 예시 추가	박종화
04	2015.09.02	• 주소 변경	박종화
05	2016.06.17	• 센터명 변경	박종화
06	2017.06.05	• 센터 영문 약칭 변경 • Ts/Tv ratio 평가 기준 추가	박종화
07	2017.10.26	• 내용 간소화에 따른 전반적 개정	박종화
08	2019.10.26	• 불확도 추정에 대한 세부평가 기준 변경	박종화

< 목 차 >

1. 목적

1.1 한국인 게놈서열

생명체의 게놈(genome) 서열(sequence)은 생명 현상의 가장 기본이 되는 유전자 및 유전자의 조절에 관한 일차적인 기본 정보를 제공하기 때문에 이를 밝히는 것은 생명 현상을 이해하는 데 있어서 아주 중요하다. 최근에는 게놈 해독(genome sequencing) 기술의 발달에 힘입어 게놈 서열 결정 속도가 가속화되고 있다. 하지만, 전 세계의 서로 다른 실험실에서 생산된 게놈 서열 데이터들은 표준화의 미비로 인하여 데이터를 서로 교환 혹은 공유하거나, 정교하게 상호 비교 분석하는 데 있어서 어려움이 따른다. 본 문서는 한국인 변이체 참조표준 데이터 제작을 위한 정량화된 품질 평가 기준을 마련함으로써 재현성 확보는 물론 높은 품질 수준을 갖춘 참조표준 데이터의 생산이 가능하도록 하는 것을 그 목적으로 한다.

1.2 한국인 변이체 데이터

최근 질병의 원인 유전변이를 발굴하기 위한 전장유전체연관분석(GWAS: Genome-Wide Association Study)이 활발히 진행되면서 사람들의 단일 염기 다형성(SNP: Single Nucleotide Polymorphism)에 대한 관심이 높아지고 있고, 맞춤 의학(personalized medicine)이 관심대상이 되면서 한국에서도 많은 연구가 진행되고 있다. 하지만, GWAS 연구는 인구집단에서 대립유전자(allele frequency)의 빈도수가 높은 SNP를 대상으로 제작된 SNP array를 이용하여 수행하기 때문에, 질병과 관련된 간접적(indirect)인 마커만을 발굴하는 한계점을 가지고 있다. 이러한 한계점을 극복하고, 질병에 직접적(direct)으로 영향을 주는 마커를 발굴하기 위해서는 차세대 시퀀싱 분석 (NGS: Next Generation sequencing)기술에 의한 게놈해독(genome sequencing)이 필요하다. 최근에는 짧은 시간과 저비용으로 게놈 해독 및 분석이 가능해지면서, 1000 Genomes Project와 같은 국제컨소시엄을 비롯한 여러 연구그룹에서 다양한 인종에 대한 전장게놈해독(whole genome sequencing) 분석 결과가 발표되고 있다. 전 세계적으로 생산된 인간게놈해독 데이터는 다양한 인종 특성, 질병의 발생 기작 연구 등에 직접 활용할 수 있어, 맞춤의학의 시대로 가는데 많은 기여를 할 것으로 예상된다. 한국인에 대해서도 NGS 기술 적용하여 정확도가 높은 전장게놈 데이터를 분석하여 변이체 데이터를 생산하여, 발굴된 변이에 대해 빈도(frequency)를 계산하고, 이에 대해 재현가능하며 정량화 할 수 있는 참조표준 데이터를 생산, 평가한다.

2. 용어 정의

2.1 게놈

생명체의 몸은 수많은 세포로 이루어져있고, 세포는 모두 핵을 가지고 있으며, 핵 안의 염색체(chromosome)라는 물질에는 생명체의 유전정보를 저장하는 DNA (deoxyribonucleic acid) 가 있다. 이 DNA는 생명체가 다음 세대로 자신의 유전정보를 전달 할 수 있는 유전전달 물질로써,

한 쌍의 염색체를 부모로부터 하나씩 받게 된다. 인간은 22쌍의 상동염색체(autosomal chromosome), 1쌍의 성염색체(sex chromosome)로 이루어진 총 23쌍을 가진다. 전체 염색체를 구성하는 DNA를 게놈(genome)이라고 한다.

2.2 게놈 염기서열 데이터

유전체 데이터는 보통 염색체를 구성하고 있는 DNA에 저장되어 있고, DNA는 아데닌(A), 구아닌(G), 시토신(C), 티민(T)과 같은 특정 염기를 갖는 뉴클레오티드(nucleotide) 사슬로 구성되어 있기 때문에, 유전체 염기서열의 순서가 곧 유전체 정보를 대변한다고 볼 수 있다. 따라서 게놈 염기서열 데이터는 이러한 게놈 정보가 저장되어 있는 DNA의 염기서열 순서를 일컫는다.

2.3 유전자(gene)

게놈(genome)중 특정 기능을 암호화 할 수 있는 영역(region)을 유전자라고 한다. 이 유전자를 생명체의 여러 형질 발현과 직접적인 연관이 있으며, 인간의 경우 인종의 특징, 질병의 특징, 개인 간의 특징을 유전자를 통해 예측을 할 수 있다.

2.4 유전자 주석(annotation)

인간게놈(human genome)에서 유전자의 위치(locus), 기능(function) 등의 정보를 분석한다.

2.5 유전자 기능 데이터

유전자 예측 데이터는 DNA 상에서 유전자의 정확한 위치를 나타낼 뿐 그 유전자의 자세한 기능에 관련된 데이터는 따로 추가해 주어야 한다. 즉, 그 유전자가 세포의 어느 위치에서 발현되어 어떤 생화학 작용을 일으키고, 어떤 생명현상에 관여하는지 자세한 기능을 알아내어 이 정보를 추가해야 한다. 하지만, 이러한 기능은 이미 밝혀져 있는 경우도 있지만, 아직 그 정보가 부족한 경우도 있고, 단순히 서열 유사성을 통해 유추만 가능한 경우도 있고, 아예 기능을 알 수 없는 경우도 존재하기 때문에 예측된 유전자에 기능을 부여할 때는 예측된 기능은 물론 예측 방법 및 신뢰도에 대한 설명도 필요하다. 이와 같은 유전자가 어떤 생물학적 기능에 대한 데이터를 유전자 기능(functional annotation) 데이터라고 한다.

2.6 염기서열 평균 중첩도(average depth)

전장게놈서열을 해독하여, 게놈 서열을 확보할 경우, 전체서열의 수십 배수 이상의 서열을 반복적으로 생산하여 서열의 정확도 높여야만, 오류가 상대적으로 적은 서열 생산은 물론 변이체 데이터를 생산 할 수 있다.

2.7 단일 염기 변이 (SNV: Single Nucleotide Variant)

모든 생물은 같은 종이라도 게놈(genome)서열의 아주 적은 부분이 차이가 날 수 있으며, 인간

의 경우 게놈(genome) 서열(sequence)이 각 개인 간에 99.9% 일치하지만 0.1%의 차이를 가진다. 염기서열 차이 중 90%는 특정위치(locus)에서 한 염기서열(base pair)가 다른 염기로 치환이 되었으며, 이러한 locus의 특징을 단일염기변이(SNV: Single Nucleotide Variant)라고 한다. 즉, 인간은 99.9% 유전자가 일치하지만 0.1%의 게놈서열의 차이를 가져 인종 또는 개인의 특이적인 형질을 보이게 되는 것이다. SNV는 인종에 따라 차이가 나타나므로 유전적 거리를 계산하는데도 이용할 수 있고, 인종의 머리색, 키, 눈색 등의 특징도 많은 부분 설명이 가능하다.

2.8 변이체(variome)

변이체란 단일 염기 다형성, 구조적변이 등의 데이터 집합체로서 정의된다.

2.9 대립유전자 빈도(allele frequency)

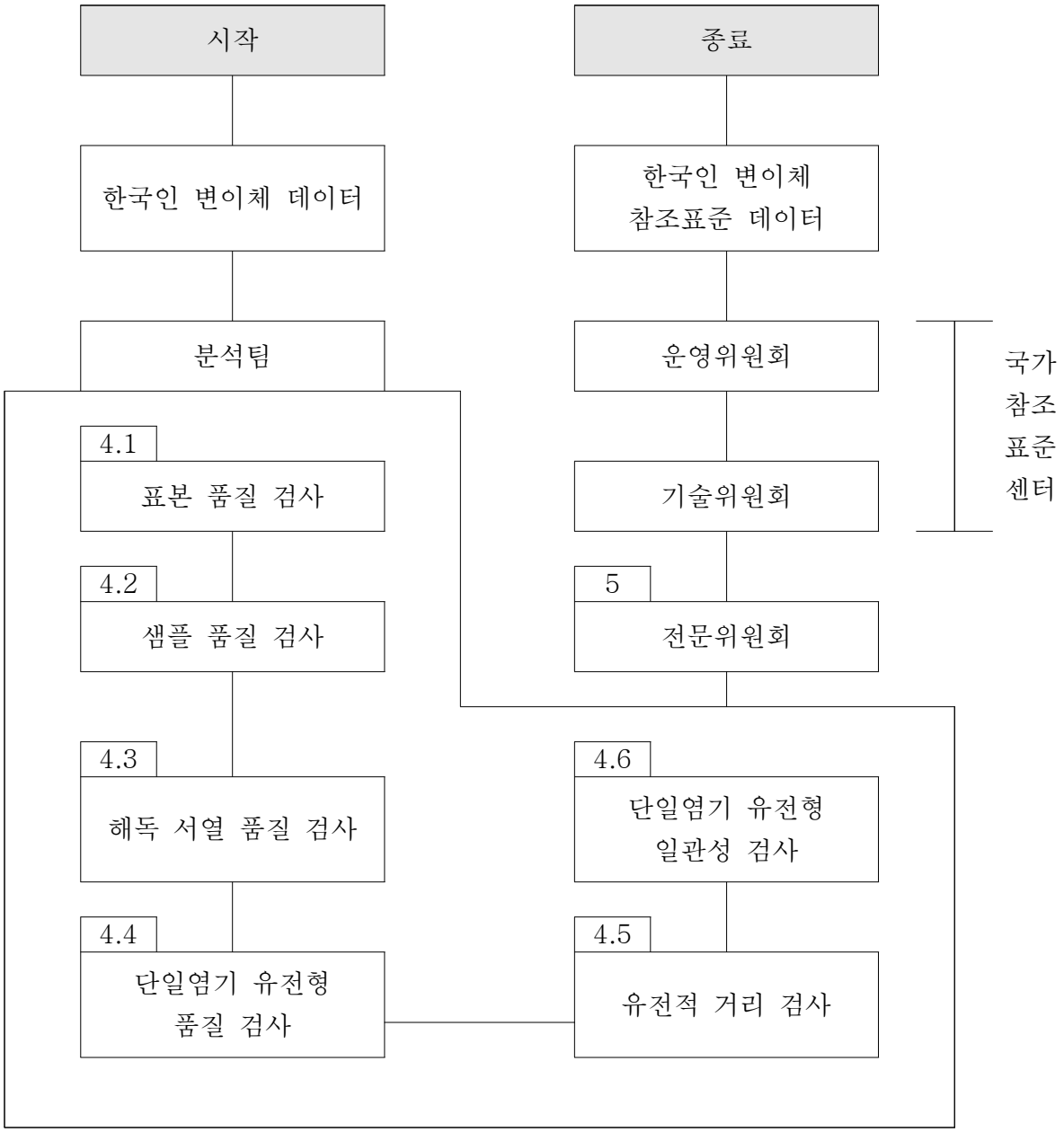
대부분의 생물종은 이배체(haploid)의 염색체(chromosome)를 가지고 있으며, 이러한 염색체를 상동염색체(autosomal chromosome)라 한다. 상동염색체의 DNA서열은 한 쌍을 이루며, 특정 위치(locus)에서의 DNA의 쌍을 유전형(genotype)이라고 하고, 이를 구성하는 DNA 서열 각각을 대립유전자(allele)라고 한다. 특정 인구 집단 내에서 유전형(genotype)의 구성을 알아보기 위해서 게놈상의 유전변이가 일어난 위치(locus)에서 대립유전자의 구성을 알아보고, 대립유전자 빈도수(frequency)를 구할 수 있다.

2.10 질병 연관성 연구(Disease Association study)

연관성(association)은 게놈(genome)의 특정 위치(locus)에 위치한 유전형(genotype)이 표현형(phenotype)과 함께 발생함을 나타내는 통계적 수치이며, 어떤 특정 집단(인종집단, 질환자 집단)에서 유전형(genotype)의 대립유전자 빈도(allele frequency)에 따라 계산된다. 연관성 분석에 주로 사용되는 통계적 방법은 여러 환경변수를 함께 고려하여, 질환/대조군 분석에서는 로지스틱 회기분석(logistic regression), 특정 형질이 연속된 수치로 표현이 되는 경우에는 선형회귀 분석(linear regression)을 이용한다.

3. 평가 절차도

3.1 변이체 데이터 평가 절차도



4. 평가 기준

4.1 표본 품질 검사

표본 품질 검사는 「임상설문지」와 「변이체 데이터 표본 수집 내역」을 대상으로 한다. 한국인 일반인 변이체 데이터의 경우에는 다음의 조건을 만족하는지 확인한다.

1) 희귀 유전 질환이 없는 표본
2) 암 발병 이력이 없는 표본
3) 연령, 성별, 주소지 정보가 확보되어 있는 표본
4) 비혈연 관계의 표본

한국인 질병군 변이체 데이터의 경우에는 다음의 조건이 만족하는지 확인한다.

1) 출처 문헌 정보가 확보되어 있는 표본
2) 비혈연 관계의 표본

4.2 샘플 품질 검사

샘플 품질 검사는 「변이체 데이터 샘플 해독 내역」을 대상으로 하며, 다음의 조건을 만족하는지 확인한다.

1) 혈액 또는 타액
2) DNA의 흡광도 260/280 ratio가 1.7 ~ 2.0
3) DNA의 흡광도 260/230 ratio가 1.5 이상

4.3 해독서열 품질 검사

해독 서열 품질 검사는 「변이체 데이터 샘플 해독 내역」과 「변이체 데이터 분석 내역」을 대상으로 하며, 다음의 조건을 만족하는지 확인한다.

1) 단서열은 paired-end로 해독
2) 필터링 후 단서열의 평균 Q score 20 이상
3) 필터링 후 단서열의 길이 50 bp 이상
4) 필터링 후 단서열의 양이 인간 게놈 서열 대비 20 배수 이상 [1, 2]
5) 필터링 후 단서열의 인간 게놈 서열에 대한 매핑률이 90% 이상

4.4 단일염기 유전형 품질 검사

단일염기 유전형 품질 검사는 「변이체 데이터 분석 내역」을 대상으로 하며, 다음의 조건을 만족하는지 확인한다.

1) Homozygous SNV 대비 heterozygous SNV 개수의 비율이 1.3 이상 [1]
2) Heterozygous SNV의 Tv (transversion) 대비 Ts (transition) 비율이 1.5 이상

4.5 유전적 거리 검사

표본이 서로 비혈연 관계인지는 표본 간의 유전적 거리를 계산하여 비교함으로써 게놈 수준에서 재확인하도록 한다. 유전적 거리 계산에는 GATK Lite 2.3.9 프로그램의 UnifiedGenotyper를 이용하여 variant site를 대상으로 발굴한 단일염기 유전형 정보를 사용한다. 두 샘플 간의 유전적 거리(π)는 다음과 같이 계산한다.

$$\pi = \frac{\sum_{k=1}^{s} d_k}{s}, \quad d = \frac{n(\{a_i b_j | a_i b_j \in P, a_i \neq b_j\})}{n(P)}, \quad P = \{a_1 b_1, a_1 b_2, ..., a_2 b_1, a_2 b_2, ..., a_i b_j\}$$

s는 두 샘플 간에 유전형이 서로 다르게 결정된 게놈 상 단일염기 위치의 총 개수를 뜻하고, a_i는 첫 번째 샘플의 한 개 유전형을 구성하는 각 allele을 뜻하며, b_j는 두 번째 샘플의 한 개 유전형을 구성하는 각 allele을 뜻한다. d 값의 경우, 예를 들면, 두 샘플의 유전형이 각각 A/A와 A/A이면 0, A/A와 T/T이면 1, A/A와 A/T이거나, A/T와 A/T이면 0.5가 된다. 유전형이 결정되지 않은 위치에 대해서는 reference homozygous 유전형으로 간주한다.

유전적 거리 검사는 상기의 방법으로 계산된 유전적 거리를 대상으로, 다음의 조건을 만족하는지 확인한다.

1) 다른 표본과의 유전적 거리가 모두 8.4×10^{-4} 이상

4.6 단일염기 유전형 일관성 검사

참조표준 등급부여의 대상이 되는 단일염기 변이체 데이터에 대해서는 일부 샘플에 대하여, 본 생산절차에서 기록된 방법이 아닌 다른 방법으로 대립유전자 빈도수(allele frequency)를 측정하여, 결과가 일관성이 있는지 확인한다. Genome-wide SNP chip으로 일관성이 확인된 경우, 각 샘플 당 일치율이 98% 이상이어야 한다[3, 4]. SNP의 genotype 재현 확인을 위한 일치율 계산은 아래와 같다.

		장비 1에서 얻은 유전형			
		aa	ab	bb	Others
장비 2에서 얻은 유전형	aa	A	B	C	D
	ab	E	F	G	H
	bb	I	J	L	M
	Others	N	O	P	Q
일치율(%) = (A+ F+ L) / (A+ B+ C+ E+ F+ G+ I+ J+ L) * 100					

일치율 측정 시 Hardy-Weinberg disequilibrium, Low call rate, minor allele frequency, filtering 기준을 적용했을 때, 포함되지 않는 marker는 제거한다.

5. 전문위원회

전문위원회는 평가 절차도에서 조건을 만족시키지 못하는 데이터들을 개별적으로 확인하여 데이터의 문제점을 재차 확인하고, 평가 프로세스의 문제점을 점검하고, 데이터의 특성상 완전히 획일화하여 평가할 수 없는 예외적 부분을 개별적으로 평가하여 반영한다.

6. 참고문헌

1. Y.S. Ju, Y.J. Yoo, J.I. Kim, & J.S. Seo **The first Irish genome and ways of improving sequence accuracy.** *Genome Biol.* 11, 132 (2010).
2. William Brockman, et al. **Quality scores and SNP detection in sequencing-by-synthesis systems.** *Genome Res.* 18.5 (2008): 763-770.
3. Weixin Wang, Zhi Wei, Tak-Wah Lam, and Junwen Wang. **Next generation sequencing has lower sequence coverage and poorer SNP-detection capability in the regulatory regions.** *Sci Rep.* 1 (2011).
4. Kimberly Pelak, Kevin V. Shianna, Dongliang Ge, Jessica M. Maia, Mingfu Zhu, Jason P. Smith, Elizabeth T. Cirulli et al. **The characterization of twenty sequenced human genomes.** *PLoS genet.* 6, no. 9 (2010): e1001111.