(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2020년02월17일
(11) 등록번호 10-2078200
(24) 등록일자 2020년02월11일

(51) 국제특허분류(Int. Cl.)
G16B 20/00 (2019.01) *G16B 25/00* (2019.01)
G16B 40/00 (2019.01) *G16B 50/00* (2019.01)
(52) CPC특허분류
G16B 20/00 (2019.02)
G16B 25/00 (2019.02)
(21) 출원번호 10-2018-0166476(분할)
(22) 출원일자 2018년12월20일
심사청구일자 2018년12월20일
(65) 공개번호 10-2019-0000342
(43) 공개일자 2019년01월02일
(62) 원출원 특허 10-2016-0172053
원출원일자 2016년12월15일
심사청구일자 2016년12월15일
(56) 선행기술조사문헌
Y. Diao 외 1인, "Building Highly-Optimized, Low-Latency pipelines for Genomic Data Analysis", 7th Biennial Conference on Innovative Data Systems Research (CIDR' 15), 2015.01. 1부.*
Y. Liu, "Towards Elastic High Performance Distributed Storage Systems in the Cloud", Licentiate Thesis, KTH Royal Institute of Technology Stockholm, Sweden, 2015. 1부.*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
(주)신테카바이오
대전광역시 유성구 테크노2로 187, 비동 512호(용산동, 미건테크노월드2차)
(72) 발명자
정종선
대전광역시 유성구 엑스포로 448, 201동 1602호(전민동, 엑스포아파트)
이선호
서울특별시 관악구 인헌15길 6-11, 101호(봉천동, 봉일타운)
(뒷면에 계속)
(74) 대리인
권형석

전체 청구항 수 : 총 2 항 심사관 : 성경아

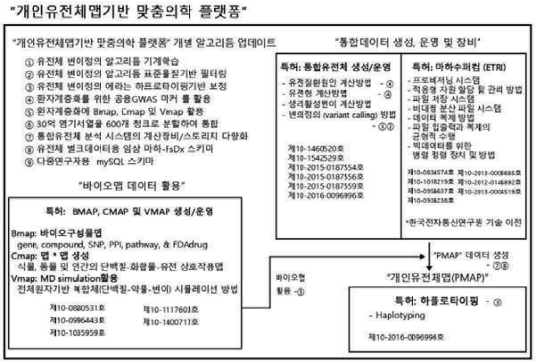
(54) 발명의 명칭 **개인 유전체 맵 기반 맞춤의학 분석 플랫폼 및 이를 이용한 분석 방법**

(57) 요 약

본 발명은 상용화된 "개인유전체맵기반 맞춤의학 플랫폼" 기반 정밀의학을 실현하기 위한 요구사항을 개선하기 위해 안출된 것으로, 본 발명은 개인 유전체의 변이 정확도(precision) 및 재현성 (recall or reproducibility)을 높이는 방법 및 유전형기반 환자계층화를 수행하기 위하여 공용 GWAS(genome wide association study)마커

(뒷면에 계속)

대 표 도 - 도1



를 사용하는 방법, 및 한국전자통신연구원이 개발한 마하수퍼컴과 연동에서의 문제를 개선 및 향상을 위한 방법을 위하여 고안된 것이다. 그리고 유전체 빅데이터 및 수퍼컴환경에서 다중 사용자에게 용이하도록 제공될 수 있도록 하는 리포팅 데이터베이스 분석 모듈이 포함된 맞춤의학 및 진단기기로 사용을 위한 개인유전체맵기반 맞춤의학 플랫폼을 상용화를 위한 개선점에 대한 연구결과이다. 본 기술이 상용화되면 국가 간의 유전체데이터기반 맞춤의학이 대부분 아카데미아에 머물러 있고 부분적으로만 가능한 현실에서 첫 전체 영역에서 상용이 가능한 맞춤의학 플랫폼이 제공될 수 있다.

(52) CPC특허분류
G16B 40/00 (2019.02)
G16B 50/00 (2019.02)

(72) 발명자
가소정
서울특별시 관악구 낙성대역6길 34, 101호(봉천동)
홍종희
서울특별시 동대문구 고산자로 534, 108동 2203호
(제기동, 한신아파트)
조양래
충청북도 청주시 흥덕구 강내면 태성탑연로
391-28, 101동 805호

명 세 서

청구범위

청구항 1

수퍼컴 기반 통합유전체맵 및 유전체데이터를 분석하기 위하여,

컴퓨터속도를 기반하여 이종 계산 시스템 및 데이터 I/O를 고려한 레벨의 서로 다른 병렬 분산 스토리지가 구비되고;

통합유전체DB에서 다중 사용자가 사용할 수 있도록, 선행계산결과 및 통계 툴이 구비되며;

외부 통계 툴의 입력파일을 생성하여, 다중 사용자가 사용할 수 있도록 mySQL 데이터베이스가 구축되며:

상기 통합유전체DB는,

수퍼컴(pc-cluster)을 기반으로 운용되어 유전체 염기서열의 대비를 통해 유전변이를 분석하도록, 인간 유전체 30억 염기를 규정된 크기의 구분한 청크들을 포함하여 구성되고;

상기 청크는,

서로 다른 복수의 대상자들로부터 추출된 유전체 염기서열이 서로 대비될 수 있도록 동일 순서와 크기로 구분되어, 상기 대상자들의 순서에 따라 정렬된 2차원 매트릭스 형태로 구성되되:

상기 청크에 포함되는 염기서열의 크기는, 상기 대상자들의 수에 따라 증감되어, 상기 청크의 데이터 크기가 10 내지 20 Gbyte가 되도록 설정됨을 특징으로 하는 개인 유전체 맵 기반 맞춤의학 분석 플랫폼.

청구항 2

제 1 항에 있어서,

상기 병렬분산 스토리지는,

(a) I/O의 한계를 설정하는 단계와;

(b) 다중 노드(다중 cpu세트)에서 계산한 데이터 분석결과의 일관성, 재현성 및 한계를 확인하는 단계를 수행함에 의해 설정됨을 특징으로 하는 개인 유전체 맵 기반 맞춤의학 분석 플랫폼.

청구항 3

삭제

발명의 설명

기 술 분 야

[0001] 본 발명은 게놈 프로젝트에 의해 구축된 다수 전장 (엑솜, 혹은 표적) 유전체 DB와 입력된 개인 전장 유전체 정보를 비교하여 개인 유전체로부터 유전정보를 분석하여 제공하는 수퍼컴퓨팅 시스템을 활용한 개인 유전체 맵 기반 맞춤의학 분석 플랫폼에 대한 발명이다.

배 경 기 술

[0003] 현재 IT 시장의 추세는 구글(Google), 페이스북(fasebook), 아마존(amazon), 클라우드컴퓨팅 및 유비쿼터스(Ubiquitous) 순으로 변화하고 있고, 이와 동시에 바이오 메디컬, 생물정보 및 유전체 영역도 바이오 구글, 시스템 바이오, 개인별 맞춤의학 그리고 정밀의학 (precision medicine) 순으로 새로운 트랜드에 맞춰 바뀌어 가고 있다. 특히 포스트 인간게놈프로젝트는 차세대 시퀀싱 기술이 급격하게 발전하여 개인별 맞춤의학을 현실화

하기 위한 노력이 활발히 진행되고 있다.

[0004] 현재 차세대 시퀀싱 기술은 인간 1명 (x30)의 전장유전체를 시퀀싱(해독)하고 분석하는데 약 1주일 정도 소요가 되는 것으로 알려져 있다. 그리고 현재 전 세계에 차세대 시퀀서가 100,000여 대가 공급된 것으로 보고되었고, 제3세대 시퀀서 (Ion Torrent: 2.5세대, Pacific BioScience의 제3세대)의 주요 개발회사들에게 많은 자금이 투자된 것으로 보고되었다.

[0005] 그 이외에 전 세계적으로는 해당분야는 모든 사업 중에서도 가장 빠르게 발전 및 개발이 되는 분야이다. 이러한, 추세대로 진행이 되면 향후 2~3년 후에는 1명의 전장 유전체 시퀀싱 및 분석이 약 $1,000이하로 낮아질 것으로 예상된다. 위의 차세대기술기반의 가장 활용성이 높고 바로 실용화되는 기술은 임상유전체(clinical genomics), 약물유전체학(pharmaco - genomics) 및 중개 임상 (translational medicine)있다, 그리고 최근에 이러한 임상유전체가 의학유전체(medical genomics)로 변신이 되고 있고, 이러한 의학유전체는 환자계층화(patient stratification)기술과 더불어 미국 대통령이 언급한 정밀의학 (precision medicine)이라는 새로운 학문 및 신 조어를 만들어 내게 되었다.

[0006] 이와 같은, 유전체 변이 관련 정보는 매년 증가하고 있으며, 본 발명은 검증 데이터의 확장에 의해 분석 정확도 영역이 지속적으로 확대될 것이다.

[0007] 한편, 본 출원인은 언급된 유전자 분석 분야의 기술적 요구사항을 개선하기 위해 지속적인 기술의 개발을 수행하고 있다.

[0008] 이와 같은 노력의 결과, 정밀의학 (precision medicine)을 위한, 바이오 빅데이터와 관련된, 임상관련 정보, 단백체 및 유전체 정보, 그리고 이들의 분석 속도를 향상시키기 위한 분석 시스템 구축, 등을 위한 방법을 개발하였고, 특히, 분석속도를 위한 GPU(graphic process unit) 기반의 분석시스템을 개발하였고(특허등록: 10-0996443), 데이터의 비교 속도를 향상시키기 위한 기법인 RVR(records virtual rack)분석 툴의 특징은 파일을 기반으로는 정보 검색 방법(특허등록: 10-0880531, 특허등록: 10-1035959, 및 특허등록: 10-1117603)을 개발하였다.

[0009] 또한, RVR 및 GPU(graphic process unit)에 기반하여 단백체에 적용시킨 (특허등록: 10-1400717), 변이의 정의(variant calling) 및 대조군과 개인 유전체 사이의 희귀변이 정도를 효율적으로 판단하기 위하여 대립유전자깊이기반 ADISCAN 분석 툴을 개발하였다 (특허등록: 10-1460520, 10-1542529, 및 10-2014-0020738).

[0010] 그리고 유전체정보를 효율적으로 관리를 하기 위한 통합유전체 DB 생성, 질병원인을 위한 변이발굴 및 환자계층화를 위한 유전형 계산 방법 (특허등록: 10-2015-0187554, 10-2015-0187556, 및 10-2015-0187559) 및 유전체정보에서 휴먼하플로 타이핑을 계산하는 방법 (특허출원: 10-2016-0096996)을 개발하였다.

[0011] 또한, 통합유전체 DB 같은 빅데이터를 위한 스토리지(storage) 운용에 특화된 미들웨어(middleware)는 한국전자통신연구원(ETRI)에서 개발한 병렬분산 환경에서 동시에 수천개의 유전체 벌크 데이터 분석이 가능하게 만든 마하수퍼컴퓨팅 시스템 (특허등록 10-1460520, 10-1010219, 10-0956637, 10-093623, 10-2013-0005685, 10-2012-0146892 및 10-2013-0004519)이 개발되었다.

[0012] 본 출원인은 한국전자통신연구원으로부터 마하시스템을 제공받아 임상환경에 적용을 위한 바이오 빅데이터를 활용한 최적화 환경을 갖추고, 정밀의학 구현을 위한 통합유전체분석 시스템과 연동된 국내 첫 수퍼컴퓨팅 시스템을 개발하였다.

[0013] 특히, 마하-Fs (유전체와 같은 버크데이터용 초고속 I/O를 위한 스토리지 시스템)는 일반 클라우드컴퓨팅 환경에 맞추어 졌지만, 본 출원인은 재현성 및 정밀성 그리고 시스템의 한계를 명확하게 정의하여, 임상환경 즉 병원에서 진단용으로 사용가능한 마하-FsDx를 개발하였다.

선행기술문헌

특허문헌

[0015] (특허문헌 0001) (001) 대한민국 등록특허 제10-0880531호

(특허문헌 0002) (002) 대한민국 등록특허 제10-0996443호

(특허문헌 0003) (003) 대한민국 등록특허 제10-1035959호

(특허문헌 0004) (004) 대한민국 등록특허 제10-1117603호

(특허문헌 0005) (005) 대한민국 등록특허 제10-1400717호

(특허문헌 0006) (006) 대한민국 등록특허 제10-1460520호

(특허문헌 0007) (007) 대한민국 등록특허 제10-1542529호

(특허문헌 0008) (008) 대한민국 특허출원 제10-2015-0187554호

(특허문헌 0009) (009) 대한민국 특허출원 제10-2015-0187556호

(특허문헌 0010) (010) 대한민국 특허출원 제10-2015-0187559호

(특허문헌 0011) (011) 대한민국 특허출원 제10-2016-0096996호

(특허문헌 0012) (012) 대한민국 등록특허 제10-0834574호

(특허문헌 0013) (013) 대한민국 등록특허 제10-1010219호

(특허문헌 0014) (014) 대한민국 등록특허 제10-0956637호

(특허문헌 0015) (015) 대한민국 등록특허 제10-0936238호

(특허문헌 0016) (016) 대한민국 특허출원 제10-2013-0005685호

(특허문헌 0017) (017) 대한민국 특허출원 제10-2012-0146892호

(특허문헌 0018) (018) 대한민국 특허출원 제10-2013-0004519호

발명의 내용

해결하려는 과제

[0016] 본 발명은 상기와 같은 상용화된 개인유전체맵기반 정밀의학을 실현하기 위한 요구사항을 개선하기 위해 안출된 것으로, 본 발명은 개인 유전체의 변이 검출 속도 및 효율을 향상시킬 수 있는 통합유전체 DB를 향상하고, 또한, 환자계층화의 효율을 높이기 위하여 공용 GWAS(genome wide association study)마커 기반 환자계층화를 향상하기 위한 방법 및, 통합유전체DB의 구성물인 변이(variant)를 위한 알고리듬 방법의 정확도(precision) 및 재현성 (recall or reproducibility)을 높이고 정확한 변이를 정의하는 과제등과 같은 통합유전체DB 와 마하수퍼컴과 연동에서의 문제를 개선 및 향상을 위한 방법을 위하여 고안된 것이다.

[0017] 또한, 본 발명은 검출된 변이정보를 빅데이터 및 수퍼컴환경에서 다중 사용자에게 용이하도록 제공될 수 있도록 하는 리포팅 데이터베이스 분석 모듈이 포함된 통합유전체 DB 및 마하수퍼컴퓨팅 기반 맞춤의학 및 진단기기로 사용을 위한 개인 유전체맵 기반 맞춤의학 플랫폼을 제공하기 위한 것이다.

과제의 해결 수단

[0019] 상기한 바와 같은 목적을 달성하기 위한 본 발명의 특징에 따르면, 본 발명은 검사자의 통합유전체DB, 바이오메디컬 정보, 단백체 데이터베이스 및 이들을 분석하기 위한 시스템 관련 기술들로 구성된 분석 플랫폼에 있어서, 분석 효율을 향상시키기 위한 특징들을 갖는다.

[0020] 또한, 본 발명은 본 발명에 의한 통합유전체 데이터베이스를 운영하기 위하여는 대한민국 등록특허 제10-0834574호, 제10-1010219호, 제10-0956637호, 제10-0936238호 및 대한민국 특허출원 제10-2013-0005685호, 제10-2012-0146892호, 제10-2013-0004519호에 개시된 마하수퍼컴시스템과 연동하여 하나의 진단기기용 마하-FsDx의 사용자환경, 기능 및 운영 효율을 높이기 위하여 다음과 같은 특징을 갖는다.

[0021] 개인 유전체 분석에서 상업화된 맞춤의학 플랫폼기반 서비스를 제공하기 위하여는 변이정의(variant calling), 환자계층화 및 통합유전체 시스템에 대한 최적화가 필요하다. 이를 위해 본 발명은 변이정의(variant calling), 환자계층화 및 통합유전체 시스템 구축에 대하여 기술적 특징 1 내지 특징 9를 적용하였다.

[0022] 먼저, 본 발명은 변이정의(variant calling)와 관련하여 3개의 기술적 특징이 적용된다.

[0023] 상기 변이정의(variant calling)는 분석결과의 기준이 되는 것으로, 바이어스(bias, 편향성) 없이 표준화가 되어야 한다.

[0024] 그러나 현재까지 알려진 유전체 변이정의 모델은 베이시안모델(Bayesian model)을 기반으로 개발된 GATK, STRELKA, EBcall, MuTect 및 SomaticSniper가 있고, 확률적 예측 및 그래프 모델 (Fisher's Exact Test and string graphs) 방식을 기반으로 개발된 jointSNVMix, VarScan 및 Fermi 등이 있다. 또한, 지식 및 통계기반 모델(Prior knowledge and statistical models)을 기반으로 개발된 GNUmap, GATK, SOAPsnp, SAMTools, 및 SNVer 등이 있다.

[0025] 그러나 위의 3가지 다른 모델은 모델 자체에 바이어스(bias)가 포함되어 들어가 있기 때문에 편향적인 결과를 도출한다.

[0026] 즉, 베이시안모델(Bayesian model)로 개발된 툴 (GATK, STRELKA, EBcall, MuTect 및 SomaticSniper)은 선행지식 기반으로 최적화(fitting)가 되었기 때문에 선행지식 기반으로 알려진 변이는 예측을 잘하지만 신규변이는 예측이 잘 안되는 경우도 있다.

[0027] 그리고 확률적 예측 및 그래프모델기반 알고리듬 (jointSNVMix, VarScan 및Fermi)들의 경우, 일반통계에서는 정규분포에서 특정 변이분포가 얼마나 벗어났는지를 계산 한다. 그러나, 실제 변이는 이러한 정규성을 가지고 있지 않고, 특정유전자, 특정질병, 특정 환경 요인에 예민할 수 있다.

[0028] 한편, 지식 및 통계 모델 (Prior knowledge and statistical models)기반 툴(GNUmap, GATK, SOAPsnp, SAMTools, 및 SNVer)들은 베이시안모델과 다른 방식의 선행지식 기반 다양한 통계를 사용한다.

[0029] 따라서, 본 발명은 변이정의에 편향성(bias)을 배제하기 위하여, ADISCAN을 적용한다.

[0030] 상기 ADISCAN의 기본 개념은 모든 인간의 유전자는 배수체(엄마의 유전자, 아빠의 유전자)를 가지고 있고, 이를 조각으로 만들고 다시 검출을 했을 때, 염기조각의 서로 다른 대립유전자의 비율이 50:50 확률로 검출을 한다는 가정하에, 검출된 대립유전자의 비율이 50:50이면 hetero 변이로 정의(homozygote)하고, 100:0이면 homo로 정의(reference homo만 수집됨)하며, 0:100이면 alternative homo로 정의(alternative homo만 수집됨)을 하는 방식을 말한다.

[0031] 그리고 이를 편향성이 없는 ratio (50/50 = 1, 0/100 혹은 100/0 = 0)로 표시할 수 있고, 이러한 모든 수치는 tangent 함수 및 다양한 다른 함수로 표현을 할 수 있다(단, tangent 함수가 성질 특성을 표현하는데 적합함).

[0032] 이와 같은, ADISCAN의 기본 개념은 출원인의 특허등록 제10-1400717호 및 제10-1542529호에 개시한 바 있다.

[0033] 이에 본 발명은, 상기 ADISCAN을 이용하여, 변이정의를 수행함에 있어, 다음과 같은 기술적 특징을 추가 한정하였다.

[0034] 즉, 본 발명은, 상기 ADISCAN을 이용하여, 변이정의를 수행함에 있어, 대립유전자깊이(depth)에 따른 변이 민감도 점수의 최고점 및 최저점을 0 내지 1 단위로 표준화하고, 변이정의(variant calling)가 불가능한 영역(Twilight Zone)의 범위를 규정하였으며, 이를 상수 파라미터(constant parameters)화 하였다(특징 1)

[0035] 그리고 본 발명은 상기 ADISCAN을 이용하여, 변이정의를 수행함에 있어, 표준물질(NA12878)을 지정하고, 상기 표준물질을 정답기반으로 하여 기계학습을 수행하여 전술한 상수파라미터를 확정하고, 위양성(false positive)의 대상이 되는 모든 변이를 표준물질(NA12878)과 비교하여 오류를 검출한다(특징 2).

[0036] 또한, 본 발명은 상기 ADISCAN을 이용하여, 변이정의를 수행함에 있어, 인간 유전체의 변이정의(variant calling)의 정밀도를 향상시키기 위하여, 하플로타입 기반의 보정을 수행한다(특징 3).

[0037] 다음으로, 본 발명은 환자계층화와 관련하여, 2개의 기술적 특징이 적용된다.

[0038] 환자계층화(patient stratification)는 현실적 임상에 사용 가능하고, 국제환경에서 표준화가 필요하다.

[0039] 본 발명은 통합유전체데이터 기반 하플로타입(haplotype) 염기서열을 이용하여 유전적 변이를 통해 환자계층화를 수행하고, 이와 같은 환자계층화의 기본적 원리는 출원인의 선행 특허 특허문헌 8, 9, 10에 개시된 바 있다.

[0040] 이와 같은 환자계층화 방법은, 유전자 및 다중유전자 단위에서 하플로타입(gene haplotype) 기반 유전자단위 및 다중유전자들의 군집화하므로, 집단유전체 정보가 충분이 있는 경우에는 활용성이 높으나, 인종에 기인한 국제적인 데이터 확보가 어려운 경우에는 활용성이 낮은 문제점이 있었다.

[0041] 즉, 국제컨소시엄(international consortium)기반으로 알려진 공용 GWAS (genomewide association study) 마커를 기반으로 환자계층화를 유전자단위에서 계산하면 효율이 향상되므로, 이러한 공용 GWAS 마커기반 유전자단위 방식의 환자계층화가 요구된다.

[0042] 그러나 공용 GWAS 마커는 희박한 (sparse) 분포에 기인하여, 즉, 많은 마커가 유전자와 유전자 중간에 혹은 인트론(intron)에 위치하여, 단백질 기능성과 연결을 못 시키는 문제점이 있다.

[0043] 현재 GWAS 마커는 GWAS catalog (https://www.ebi.ac.uk/gwas/, 유럽생물정보기관, EBI)에서 만든 공용 마커로, DB는 수천 내지 수십만 명을 사용하여 계산하는 질병연관성 연구는 국제컨소시엄의 결과물이고, 환자계층화연구의 중요한 자산이 된다.

[0044] 다만, 50만 개 변이가 30억 개 염기를 대변해야 하기 때문에 변이 한 개의 염기(혹은 변이)를 중심으로 앞과 뒤에 6,000 base pair씩 12,000 bp에 1개의 염기가 있는 방식으로 사용이 되었다.

[0045] 이 역시 대부분이 비-기능성 변이인 것이 문제점이다.

[0046] 이와 같은 문제를 해결하기 위해서, 본 발명은 연관 불평형(LD, linkage disequilibrium)기반으로 유전자 단위의 환자 계층화를 수행하는 기술(특징 4)이 적용된다.

[0047] 또한, 본 발명은 공용 마커의 희박성을 보완하기 위해 바이오데이터 간 연결을 통해 연관성을 확보하는 기술(특징 5)가 적용된다.

[0048] 마지막으로 본 발명은 통합유전체 시스템 구성과 관련하여, 4개의 기술적 특징이 적용된다.

[0049] 본 발명에 적용되는 통합유전체 시스템의 기본적 구성은 대한민국 등록특허 제10-1460520호 및 대한민국 특허출원 제10-2015-0187554호에 개시된 바 있다.

[0050] 그러나 개시된 바와 같은 통합유전체 시스템은 pc-cluster(서버 급 노드)를 100 - 1,000대를 연결하여 사용하는 수퍼컴퓨팅 환경에서, 많은 CPU 기반으로 동시에 다수의 인간 유전체데이터를 계산하고 통합하여야 한다. 이와 같은 환경에서, 합리적인 가격으로 pc-cluster를 구성을 하기 위하여, 표준화된 전장유전체 분석을 수행하기 위한 분석 데이터의 분할 구조가 필요하다.

[0051] 본 발명은 이와 같은 표준화된 분석 데이터 분할 구조를 제공하기 위해, 30억개의 전장유전체를 규정된 개수의 청크 (chunks)로 분할하여 노드(node: multiple CPU)당 최적화된 메모리 크기를 할당한다(특징 6).

[0052] 또한, 본 발명은 전장유전체 분석의 효율성을 향상시킨 유전체분석용 수퍼컴을 구성하기 위하여, 메모리기반 CPU, GPGPU기반 GPU 및 I/O성능(1,000대의 동시계산을 수행 가능), 메모리-케시-서버 및 병렬분산형 스토리지에 대한 표준화된 구성(특징 7)을 제공한다.

[0053] 아울러, 본 발명은 이와 같은 하드웨어를 임상 진단기기로 적용하기 위해, 한계 및 재현성 실험결과를 제공(특징 8)한다.

[0054] 그리고 마지막으로 본 발명은 개인유전체맵기반 맞춤의학 플랫폼과 같은 대형 시스템은 일반 연구 개발자가 사용할 수 있도록 하기 위하여, 선행계산 및 공용툴을 장착한 표준화된 데이터베이스를 제공(특징 9)한다.

발명의 효과

[0056] 위에서 살핀 바와 같은 본 발명에 의한 개인 유전체 맵 기반 맞춤의학 분석 플랫폼을 이용하면, 맞춤의학용 플랫폼의 상용화가 가능해지고, 유전체 변이정의(variant calling)에 대한 정밀한 수행이 가능해지며, 환자계층화의 결과는 아카데미아 수준으로부터 상용화 수준으로 향상될 수 있으며, 통합유전체DB 시스템은 반자동화 환경에서 자동화가 가능한 효과가 있다.

도면의 간단한 설명

[0058] 도 1은 본 발명에 의한 개인 유전체 맵 기반 맞춤의학 분석 플랫폼 및 이를 이용한 분석 방법에 있어, 핵심 기술구성을 개시한 개요도.

도 2는 본 발명에 의한 유전체 변이정의 알고리듬(ADISCAN)의 기계학습 개념을 도시한 개념도.

도 3은 본 발명에 의한 표준물질기반 유전체 변이정의 알고리듬 필터링 및 하프로타이핑 기반 오류보정 개념을 도시한 개념도.

도 4는 본 발명에 의한 개인 유전체 맵 기반 맞춤의학 분석 플랫폼 및 이를 이용한 분석 방법에 적용되는 하플로타이핑의 기본 개념을 도시한 예시도.

도 5는 본 발명에 의한 환자 계층화를 위한 유전체 기술 히스토리와 개념을 정리한 개념도.

도 6는 본 발명에 의한 통합유전체 DB에서 환자계층화를 위한 공용 GWAS 마커 활용 및 Bio-map이 활용된 예를 도시한 예시도.

도 7는 본 발명에 의한 GWAS 마커 기반 환자계층화 생성 방법을 도시한 예시도.

도 8는 본 발명에 의한 GWAS 마커 기반 환자계층화 생성예를 도시한 예시도.

도 9는 본 발명에 적용되는 PC-cluster에서 운영을 위한 인간 유전체 30억 염기에 대한 분할된 청크를 도시한 개념도.

도 10은 본 발명에 의한 유전체데이터 분석을 위한 계산장비(a1-a6), 스토리지(b1-b3)구성 및 마하-FsDx의 구성을 도시한 개념도.

도 11a 내지 도 11f는 본 발명에 의해 마하-Fs를 마하-FsDx 화하여 안정성, 한계성 및 재현성에 대하여 평가한 결과를 도시한 예시도.

도 12a 및 도 12b는 한계성에 대하여, MAHA-FsDx 구성 규모 대비 최대치 성능을 검증한 예시도.

도 13은 본 발명에 의한 통합유전체 DB 플랫폼기반 다중연구자 데이터베이스의 구성을 도시한 개념도

도 14은 본 발명에 의한 통합유전체 DB의 대립유전자깊이, 지노타입 및 하플로타입 DB의 구성을 도시한 개념도.

도 15는 본 발명에 의한 각각의 청크 단위 메트릭스에서 SNV, INDEL 및 CNV의 검출 원리를 도시한 개념도.

도 16은 본 발명에 의한 통합유전체 DB에서 유전자 및 다중유전자기반 유전형 계산을 통한 환자계층화를 도시한 개념도.

발명을 실시하기 위한 구체적인 내용

[0059] 이하에서는 첨부된 도면을 참조하여 본 발명의 구체적인 실시 예에 의한 개인 유전체 맵 기반 맞춤의학 분석 플랫폼을 상세히 살펴보기로 한다.

[0060] 본 발명의 각 특징을 설명하기에 앞서 먼저, 본 발명이 적용되는 유전자 분석 서비스의 구성을 간단히 살펴보기로 한다. 유전자 분석 서비스는 병원 등의 개인 유전자 수집 기관으로부터 혈액 등의 샘플을 수집하여, 해당 샘플을 DNA 시퀀싱 회사에 시퀀싱을 의뢰하게 된다.

[0061] 그리고 상기 DNA 시퀀싱 회사는 수집된 샘플로부터 DNA(NGS, next generration sequencing) 해독을 수행한다. 물론, 최근에는 기술적 발전에 따라 다양한 방법에 의해 DNA sequencing을 생성할 수 있으므로, 상기 DNA 해독 생성 방법은 DNA 시퀀싱 회사의 기술 수준에 따라 다양한 방법에 의해 수행될 수 있다. 이와 같이 생성된 DNA 해독은 본 발명과 같은 시스템을 통해 개인 유전체에 포함된 유전적 정보가 분석되고, 분석된 분석정보는 병원 등의 진단기관 또는 수요자에게 전달된다.

[0062] 물론, 상기 DNA 시퀀싱 회사로부터 DNA 벌크(bulk, dummy) 데이터가 제공되는 경우, 이로부터 고집적 인덱싱 파일로 형성하여 빅데이터인 유전체 염기서열을 분석할 수도 있다.

[0063] 즉, 본 발명은 DNA 해독 정보로부터 개인 유전체에 포함된 유전적 정보를 분석하는 분석 및 진단 시스템에 관한 것으로, 이하에서 본 발명에 의한 통합유전체 빅데이터를 활용한 개인 유전체 맵 기반 맞춤의학 분석 플랫폼 및 이를 이용한 분석 방법에 있어, 기술적 특징 1 내지 9를 중심으로, 본 발명의 기술 구성을 상세히 설명하기로 한다.

[0064] 도 1은 본 발명에 의한 개인 유전체 맵 기반 맞춤의학 분석 플랫폼 및 이를 이용한 분석 방법에 있어, 핵심 기술구성을 개시한 개요도이다. 이에 도시된 바와 같이, 본 발명은, 개인 유전체 분석에서 상업화된 맞춤의학 플랫폼 기반 서비스를 제공하기 위하여 변이정의(variant calling), 환자계층화 및 통합유전체 시스템 구축에 대하여 기술적 특징 1 내지 특징 9가 적용된다.

[0065] 도 2는 본 발명에 의한 유전체 변이정의 알고리듬(ADISCAN)의 기계학습 개념을 도시한 개념도로 본 발명의 기술적 특징 1의 개념을 도시하고 있다.

[0066] 도 2에 도시된 바와 같은, 본 발명의 특징 1은 ADISCAN을 이용하여, 변이정의를 수행함에 있어, 대립유전자깊이(depth)에 따른 변이 민감도 점수의 최고점 및 최저점을 0 내지 1 단위로 표준화하고, 변이정의(variant calling)가 불가능한 영역(Twilight Zone)의 범위를 규정하였으며, 이를 상수 파라미터(constant parameters)화 한 것이다.

[0067] 즉, 기계학습을 위한 score function을 적용하고, 상기 score function은 3가지 상수 파라미터(a, b, 및 c) 가 적용된 tangent 함수를 이용하여 유전체의 변이정의(variant calling)를 적용하되, 상수 파라미터(a)는 대립유전자깊이(depth)에 따른 변이 민감도, 상수 파라미터(b)는 점수의 최고점 및 최저점을 0 - 1단위로 표시하며, 상수 파라미터(c)는 변이정의(variant calling)가 불가능한 영역(Twilight Zone)의 범위를 정한다.

[0068] 즉, 본 발명에 의한 ADISCAN을 위한 score function은

[0069] Score function S(i)= tan(D-0.5)× a - log b (max(b,min(depth)))+c 에 의해 산출된다.

[0070] 위 Score function 산출식의 의미는, 대립유전자깊이 비율(ratio)(=D), 대립유전자 깊이 중에 첫 번째 대립유전자(reference depth) 깊이 및 두 번째 대립유전자 깊이(alternative depth) 중 작은 것과 b(상수) 중에 큰 것을 log의 밑 b(상수)으로 하는 로그값, tan(tangent 함수) 값의 가중치 a(상수) 그리고, 0에서 1의 scale을 위한 c(상수)로 유전자 변이정의 값을 표현할 수 있다는 것이다.

[0071] 이와 같이, 상수 a, b 및 c를 알면 바로 변이를 homo, hetero 또는 alternative homo로 편향성이 없는 판정을 할 수 있다.

[0072] 이를 위해 알려진 정답에 상수 a, b 및 c 형식으로 기계학습을 수행하여 상수를 계산하였고, 상수 파라미터(a)는 대립유전자깊이(depth)에 따른 변이민감도, 상수파라미터(b)는 점수의 최고점 및 최저점을 0 - 1단위로 표시하였으며, 상수파라미터(c)는 변이정의(variant calling)가 불가능한 영역(Twilight Zone)의 범위를 최적화를 한다.

[0073] 이에 따라 본 발명의 특징 1에 따르면, 범용성을 위한 알고리듬의 재현성 및 정확성이 일관성이 유지될 수 있는 효과가 있다.

[0074] 한편, 본 발명의 특징 2는 상기 Score function의 상수 파라미터를 표준물질(NA12878)에서 제공한 데이터를 활용하여 기계학습을 수행하여 변수를 확정하는 것이다.

[0075] 도 3은 본 발명에 의한 표준물질기반 유전체 변이정의 알고리듬 필터링 및 하프로타이핑 기반 오류보정 개념을 도시한 개념도.

[0076] 도 3에 도시된 바와 같이, 본 발명의 특징 2는 유전체의 변이정의 (variant calling)에 있어, 추가적으로 표준물질(NA12878)과 비교하여 오류를 수정한다.

[0077] 인간유전체는 진화가 되는 과정에서 수없이 많은 반복되는 서열을 가지게 되었고, 반복되는 염기조각이 여러 다른 유전자에 흩어져서 정렬이 된다.

[0078] 이때, 레퍼런스 게놈(Reference genome)에 정렬된 염기조각을 BAM(binary alignment map)이라 부르고, 상기 BAM기반으로 변이정의(variant calling)을 하게 되면 모두 위양성(false positive)의 대상이 된다.

[0079] 따라서, 이와 같이 중복해서 정렬되는 염기조각을 가지는 유전자를 미리 필터링을 하고, 보정 하면 인간 유전체의 변이정의(variant calling)를 더욱 정밀하게 할 수 있다.

[0080] 이때, 상기 표준물질(NA12878)은 전 세계 연구자들이 가장 많이 인용하고 비교용으로 사용하는 유전체정보 및 검체 정보를 말하는 것으로, 미국표준연구원의 지원으로 저스틴이 편집한(Justin et. al., Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls,Nature Biotechnology 32, 246251, 2014) nature biotech 논문에서 언급된 바 있다.

[0081] 여기서, 상기 표준물질은 한 사람의 셀라인(cell line) 검체에 대하여 영구적으로 생존하는 셀라인 (cell line)을 만들고, 다양한 플랫폼으로 한 사람에 존재하는 모든 변이를 검출하여, 표준적으로 사용할 수 있고, 공유하는 검체를 말한다.

[0082] 상기 NA12878은 다양한 차세대 시퀀싱 플랫폼(4개 이상)을 사용하여 정답 변이가 약 350만 개를 생성이 되어있다.

[0083] 한편, 변이를 검출하는 툴을 적용하기 전 단계를 BAM(binary alignment map)이라고 한다. 그리고 이러한 BAM파일이 주어지면, ADISCAN 변이검출 툴을 사용하여 변이를 검출을 하는데, 변이를 정답과 비교하면 틀린 답을 선별할 수 있다.

[0084] 여기서, 틀리게 나오는 이유는 변이 검출 툴 자체가 틀린 것이 아니고, BAM파일이 잘못 만들어진 데 기인한다. 따라서, NA12878의 정답 변이 350만개는 BAM파일의 에러(Error)를 보정하는 용으로 사용 가능하다.

[0085] 구체적으로 상기 BAM파일의 에러보정은 아래 세단계의 작업과정에 의해 수행된다.

[0086] 첫째로, 상기 NA12878의 정답 변이를 활용하여, 가장 표준적인 유전체 분석파이프라인을 사용하여 계산을 수행한다. 가장 표준적인 유전체분석 파이프라인 1-8 단계가 있고, 본 발명에 의한 변이 분석 역시 아래와 같은 8단계에 의해 수행된다.

[0087] 1) Trimming, 2) Mapping, 3) Sort BAMs, 4) Merge BAMs, 5) Remove Duplication, 6) Realign InDEL, 7) Recalibrate Scores, 8) Variant Calling

[0088] 이러한 파이프라인을 사용하고, 표준분석 툴(BWA-MEM)을사용하여 새로 변이를 검출을 위한 기초데이터인 BAM(Binary alignment map)을 생성한다.

[0089] 둘째로 변이 검출 툴인 ADISCAN을 사용하여 변이를 검출하고, 변이가 누적되어 잘못 산출되는 유전자변이를 특정유전자에 누적 합을 계산한다.

[0090] 누적 합이 1개 이상 나오는 유전자를 선별하고, 1개 이상 나오는 유전자에 대하여 원인분석을 한다. 즉, 원인이 염기조각(read)이 중복(duplication)적으로 정렬되는 reference 자체의 중복성이 있는지, 혹은 특정, exon 또는 intron에 중복이 있는지 확인한다.

[0091] 셋째로 중복성이 밝혀진 유전자구조(exon, intron, utr 또는 intergenic)영역에 대하여 하플로타이핑(haplotyping)을 하기 위한 1000게놈프로젝트(haplotype 5008명)를 사용하여 하플로타이핑을 수행한다.

[0092] 한편, 도 4에는 본 발명에 의한 개인 유전체 맵 기반 맞춤의학 분석 플랫폼 및 이를 이용한 분석 방법에 적용되는 하플로타이핑의 기본 개념이 도시되어 있다.

[0093] 상기 하플로타이핑 방법에 대한 기본 개념은 특허출원: 제10-2016-0096996호에 개시된 바 있고, 이를 다시 요약하면, 도 4에 도시된 바와 같이, (A)염기 서열조각모음(fastq)을 (B)인간표준유전체(GRCH38)에 정렬하고, 정렬된 서열(fastq)을 다시 추출하여, (C) 기구성된 다른 하플로타입 데이터베이스 (IMGT/HLA: HLA유전자의 경우에 특수하게 제작된 여러 종류의 하플로타입을 포함하는 데이터베이스)에 정렬을 한다. 그리고 (D) 연속적으로 위양성(false positive)를 제거하면서 랭크를 정하고, (E)최종으로 한 사람의 아빠/엄마의 각 하플로타입 2개를 분리하는 과정을 통해 수행한다. 그리고 (E) 제 (D) 단계의 결과물로 생긴 2가닥의 하플로타입이 나오면, 그것을 합쳐서 배수체(genotype)로 만들어 변이를 보고하는 과정을 통해 수행된다.

[0094] 이때, 본 발명에 의한 하플로타이핑은 IMGT/HLA 데이터베이스 대신에 1000게놈프로젝트(haplotype)의 5008 명의 하플로타입 데이터베이스를 사용한다.

[0095] 한편, 본 발명의 특징 3은 전장유전체의 변이를 계산하는 방법 및 시퀀싱(해독)된 염기서열조각(read)이 중복된 영역에 정렬되어 발생되는 오류(에러)(variant calling error)를 보정하는 방법, 즉 본원 발명 특징 2의 결과를 보정하는 하플로타이핑 방법을 적용시킨 새로운 효율적인 방식이다.

[0096] 본 발명에 적용되는 하플로타이핑은 1,000게놈 프로젝트에서 생성된 26개 인종기반 HaplotypeDB 및 다른 확립된 하플로타입 DB가 포함된다. 따라서, 하플로타이핑 방식에 의해 하플로타이핑을 수행하고, 변이정의(variant calling)을 수행할 수 있다. 상기 하플로타이핑에 대한 기본 개념은 다시 설명하기로 한다.

[0097] 이때, 레퍼런스 유전체(reference genome)에 정렬된 염기조각(read)를 추출하고(fastq), 추출한 염기조각을 위에서 언급한 1,000게놈의 하플로타입 DB에 에셈블러(bwa-mem, 등)을 통하여 정렬하며, 하플로타이핑 방법을 사용하여 2개의 대립유전자(alleles)를 선별하고, 도 3에 도시된 두 개의 allele A, B를 합쳐, 합쳐진 결과를 기반으로 변이정의(variant calling)을 수행한다.

[0098] 이하에서는 환자계층화와 관련된 본원 발명의 특징 4 및 특징 5를 설명하기로 한다.

[0099] 도 5는 본 발명에 의한 환자 계층화를 위한 유전체 기술 히스토리와 개념을 정리한 개념도이다.

[0100] 이에 도시된 바와 같이, 환자계층화는 인간유전체가 완성이 된 2002년 이래, SNP칩기반 질병연관성 기술이 활발히 진행이 되었고, 특히, GWAS기반 마커 발굴 국제 컨소시엄이 활발했었고, 이어서 eQTL관련 국제컨소시엄도 계속 생겨났다.

[0101] 그리고 이의 결과로서 GWAS_Catalog 약 40,000개 GWAS 마커 및 eQTL의 약 600만개 마커가 공개되었고, 미국 NCBI에서 희귀질환, 암질환, 및 일반질환 관련 유전자 변이마커를 Clinvar (Clinical variant)용으로 데이터베이스화를 시작했고, 현재 약 100,000개의 희귀질환 마커가 공개되었다.

[0102] 또한, 2007년 전후에 일루미나 및 프로톤 (PGM-ION) NGS기반 실용화가 된 시퀀싱 장비가 출시되었고, 새로운 변이 혹은 암-변이 기반 진단용 임상유전체가 등장했고, 국제암유전체컨소시엄 및 1,000게놈프로젝트 등에서 각, 2,100쌍의 암환자 전장유전체, 2600명의 정상인 유전체를 공개하기에 이르렀다. 그리고, 신테카바이오가 세계에서 처음으로 두 개의 빅데이터를 30억 염기 * 6,800개 전장유전체를 하나의 대형 메트릭스(matrix)화를 수행을 하였다. 이것을 통합유전체DB라고 명명을 하고 있다.

[0103] 도 5에 도시된 바와 같이, 전장유전체와 SNP칩데이터의 마커간의 거리를 비교해보면, SNP칩데이터의 각 마커는 전장유전체의 12,000당 하나씩 존재한다. 즉, 12,000/2 * 500,000마커 ~= 30억 염기가 된다. 따라서, SNP칩에서 찾은 GWAS마커에 기능정보를 연결하려면 LD (Linkage disequilibrium)에 의하여 두 개의 로커스가 서로 함께 유전이 된다는 사실을 기반으로 기능정보를 가진 eQTL마커, BAV (bioactive variant) 및 Clinvar 마커를 연결하는 시도가 필요하다. 또한, 변이의 질병연관성(penetrance) 정보는 임상에서 아주 중요한 질병을 연결하는 척도인데, 이러한 penetrance의 정보 중에서 희귀하지만 낮은 penetrance정보를 주는 영역에 대한 역할이 거의 알려지지 않고 있다. 그리고 이러한 정보는 환자계층화에 희귀변이 혹은 개인별 특정단백질의 활성도에 영향을 줄 것으로 판단을 하고 있다.

[0104] 도 6는 본 발명에 의한 통합유전체 DB에서 환자계층화를 위한 공용 GWAS 마커 활용 및 Bio-map이 활용된 예를 도시한 예시도이고, 도 14은 본 발명에 의한 통합유전체 DB의 대립유전자깊이, 지노타입 및 하플로타입 DB의 구성을 도시한 개념도이며, 도 16은 본 발명에 의한 통합유전체 DB에서 유전자 및 다중유전자기반 유전형 계산을 통한 환자계층화를 도시한 개념도이다.

[0105] 도 6에 도시된 바와 같이, 본 발명에 의한 개인 유전체 맵 기반 맞춤의학 분석 플랫폼 및 이를 이용한 분석 방법의 특징 4는 연관불평형을 기반으로 공용마커를 활용하여 환자계층화를 수행한다.

[0106] 즉, 본 발명에 의한 환자계층화는 도 14 및 도 16에 도시된 바와 같이, 유전자단위 및 다중유전자에서 대조군과 개인 유전체 사이의 일반변이 (SNP: single nucleotide polymorphism) 및 희귀변이 (rare variant)마커를 검출하고, 이들을 기반으로 유전형을 계산하여 환자계층화(patient stratification)를 수행한다.

[0107] 이와 같은 환자계층화의 기본 개념은 대한민국 특허출원 제10-2015-0187554호, 제10-2015-0187556호 및 제10-2015-0187559에 개시된 바와 같다.

[0108] 그러나 이와 같은 환자 계층화 방식(유전자 및 다중유전자 단위에서 하프로타입(gene haplotype) 기반 유전자단위 및 다중유전자들의 군집화를 하는 방법에 의한 유전형 계산법)은 집단정보가 충분이 있는 경우에는 활용성이 높으나, 국제적인 데이터 확보가 어려운 경우에는 효율성이 떨어지게 된다.

[0109] 즉, 국제컨소시엄(international consortium)기반으로 알려진 공용 GWAS(genome wide association study) 마커기반 환자계층화를 유전자단위에서 계산하면 효율이 좋아지므로, 이러한 공용 GWAS 마커기반 유전자단위 방식의 환자계층화가 요구된다.

[0110] 그러나 공용 GWAS 마커들은 변이가 넓은 영역에 희박한 (sparse) 분포로 되어 있는 것이 대부분이므로, 마커가 특정 질병에 대한 기능 설명되지 못하는 경향이 있다.

[0111] 따라서, 본 발명은 연관 불평형(LD, linkage disequilibrium)에 따른 지수 R^2을 이용하여 기준값(0.7)을 기준(R^2>0.7)으로 비기능적 변이를 기능변이와 연결을 시켜 공용(GWAS 등)마커기반의 유전자단위 환자계층화를 수행한다.

[0112] 이때, 상기 연관 불평형 지수는, 아래 산술식에 의해 산출될 수 있다.

$$D' = \frac{|D|}{D_{max}}$$

$$where \quad D_{max} = \min(p_A p_b, p_a p_B) \; if \; D > 0;$$

$$D_{max} = \min(p_A p_B, p_a p_b) \; if \; D < 0$$

$$r^2 = \frac{D^2}{p_A p_a p_B p_b}$$

[0113]

[0114] 이때, r^2 즉, 연관불평형(D~=LD) 지수 계산은 두 개의 단일염기 다형성 간에 존재하는 연관관계의 강도를 계산한다.

[0115] 각 단일염기 다형성에서 관측되는 대립유전자를 이용하여 계산한 하플로타입 빈도와 무작위로 나타날 하플로타입 빈도의 차이를 이용하여 LD 지수를 계산한다.

[0116] LD 지수로는 현재 D' 이 가장 많이 이용되고 있으며, 일반적으로 |D'| > 0.8 인 경우 두 단일염기 다형성 간에 강한 연관관계가 있다고 판단한다.

[0117] 또한, r^2가 0.7보다 크면 두 단일염기 다형성 간에 강한 연관관계가 있다고 판단한다.

[0118] 여기서, Pa는 첫 번째 locus 에 A 대립유전자, Pa는 첫 번째 locus 에 a 대립유전자, PB는 두 번째 locus에 B 대립유전자 및 Pb는 두 번째 locus에 b 대립유전자를 가지는 하플로타입의 빈도이다.

[0119] 한편, 도 7에는 본 발명에 의한 GWAS 마커 기반 환자계층화 생성 방법이 예시도로 도시되어 있다.

[0120] 이에 도시된 바와 같이, 본 발명에서는 환자계층화에서 GWAS마커 유전형은 major 및 minor allele가 있고, 이중에서 GWAS마커와 LD가 형성이 되는 LOF 마커와 eQTL마커는 같은 allele로 되어 있어야 한다.

[0121] 그리고 GWAS마커는 기능성 정보를 가지고 있지 않고, LD로 연결이 되는 마커는 기능성 정보를 항상 필요로 한다.

[0122] 따라서, 환자계층화 생성 규칙을 정의할 수 있고, Genetic variation과 expression의 관계는 GWAS마커, eQTL, LOF, 및 발현(expression)되는 유전자의 간단한 관계도를 보여줄 수 있다.

[0123] 도 8에는 본 발명에 의한 GWAS 마커 기반 환자계층화 생성예가 도시되어 있다.

[0124] 도 8에 도시된 바와 같이, GWAS마커 (밤색)가 선별이 되고, GWAS마커의 minor allele를 가지고 있으면서 동시에 eQTL마커 (빨강색)가 minor allele인 경우가 선별이 되고, 다음으로 GWAS마커(밤색)의 minor allele를 가지고 있고 LOF 마커(초록)를 가지고 있는 것을 선별한다.

[0125] 그리고 질병계층화-1 ~ 질병계층화-N을 선별하고, 비-질병 계층화 군을 분류할 수 있다.

[0126] 또한, 나중에 실제 약물을 투입했을 때 약물에 반응하는 환자 (responder)가 특정 질병계층화-i에 포함이 되면, 약물과 특정 환자계층화 마커와의 관련성이 있다고 할 수 있다.

[0127] 그리고 환자계층화 마커와 기능연관성 정보와의 분자생물학 영역의 메커니즘을 규명하는 작업을 할 수 있다. 여기서, eQTL마커는 발현하는 mRNA 코딩 유전자를 알 수 있고, LOF (loss of function)는 GWAS마커의 기능을 직접적으로 설명 하게 된다.

[0128] 한편, 전술한 바와 같이, 공용 GWAS 마커와 eQTL 및 유전자변이(gene variants)는 R^2 >0.7과 같은 LD기반으로 기능정보를 연결하지만, eQTL 및 유전자변이(gene variants)가 기능정보를 주지 못할 수 있다.

[0129] 이는 12,000 bp 당 하나씩 존재하는 GWAS 마커가 너무 희박한 (sparse) 분포로 생성된 마커이기 때문에, 여전히 기능성과 연결을 못 시킬 수 있기 때문이다.

[0130] 이러한 문제점을 해결하기 위해 본 발명은 바이오맵(bmap, cmap 및 vmap, 특허등록 제10-0880531호, 제10-0996443호, 제10-1035959호, 제10-1117603호 및 제10-1400717호)에서 사용한 바이오데이터 인덱싱(rvr), 바이오구성물을 상동성기반으로 생성된 맵(bmap), 맵과 맵을 연결기키는 방법(cmap) 및 단백질-약물-변이 복합체의

시뮬레이션(vmap)을 통해 연관성을 확보하여 희박한 (sparse) 분포로 만들어진 공용 GWAS 마커의 기능성을 설명하는 기술(특징 5)이 적용된다.

[0131] 즉, 본 발명은 바이오맵기반 바이오정보의 활용방법은 특징 4와 같이, 맞춤의학 플랫폼의 통합유전체DB에서 GWAS마커를 전술한 바와 같은 연관불평형(LD)에 의하여 eQTL마커와 연결이 지어지고, eQTL에 의하여 발현되는 유전자가 질병 메카니즘 관련 기능성이 설명되지 않는 경우, 특징 5와 같이, 바이오맵(bmap), 맵x맵(cmap) 및 복합체(변이-약물-단백질, vmap)의 정보를 통하여 기능성을 설명하고, 이와 같은 기능을 RVR 기술을 기반으로 신속하게 네트워킹 및 검색하도록 한다.

[0132] 여기서, rvr, bmap, cmap,및 vmap에 대한 개념은 특허등록 제10-0880531호, 제10-0996443호, 제10-1035959호, 제10-1117603호 및 제10-1400717호에 개시되어 있으나, 이를 간단히 정리하면 다음과 같다.

[0133] 첫 번째로 rvr(records virtual rack)은 단일 데이터 검색을 위한 파일 생성 방법 및 단일 데이터 파일의 검색 방법으로, 단일 파일 검색을 위한 rat(record allocation table)파일을 이용한다. rvr은 컴퓨터공학에서 사용하는 인덱싱 기법(inverted indexing)을 바이오데이터에 적용하도록 변형한 것이다.

[0134] 요점은 다량의 바이오 데이터들의 특성은 벌크데이터 혹은 빅데이터 이므로, 파일의 주소정보를 외부에 기록을 한 파일을 생성(rat: record allocation table)하면, batch 처리를 비롯해서 대용량 바이오데이터를 핸들링 하기가 용이해진다.

[0135] 본 방법을 사용하여, 염기서열, 단백질서열, 다중정렬파일, text정보, 단백질 3D정보, 변이정보, 이미지정보, 오디오정보, 등, 수많은 정형, 비정형의 다양한 종류의 바이오관련 빅데이터 자료를 처리할 수 있다.

[0136] 두 번째로 bmap(bio map)은 군집 및 백본 데이터베이스 기반 바이오 메디컬 통합 정보 검색 방법으로, 대한민국 특허등록 제10-1035959호에 개시된 바 있다.

[0137] 상기 bmap은 사용자가 알고자 하는 유전자의 주석(annotation)을 공용(public) DB에서 해당 유전자(gene)와 관련된 모든 정보를 상동성(Homology) 기반으로 수집 및 rvr 기반의 인덱싱을 사용하여 바이오구글 개념의 검색을 하는 방법을 말한다.

[0138] 공용 데이터의 컨텐츠는 Gene expression, Pathway, Disease, Nucleotide, Protein, Regulation, Interaction, Metabolite, SNP & mRNA SNPChip data, Drug and ProteinChemical, 및 Interaction 관련 DB들을 중심으로 유전자맵 결과를 만들 수 있다.

[0139] 같은 방식으로 약물, 변이, 상호작용 및 FDA 약물 맵도 만들 수 있다.

[0140] 세 번째로 cmap(cell map)은 상호 연계 가능한 다중 맵 생성을 통한 바이오메디컬 기능연관정보 제공 시스템에 대한 내용으로, 대한민국 특허등록 제10-1117603호에 개시되어 있다.

[0141] 상기 cmap는 바이오맵에 사용하는 keyword 혹은 ID를 바이오맵에 연결을 시키는 방법 및 바이오맵과 바이오맵을 연결시키는 방법에 관한 것이다.

[0142] 그리고 상기 cmap의 특징은 아래 3가지로 구분될 수 있다.

[0143] 1) 쿼리(Query)는 유전자 ID, 화합물 ID, 질병 ID 및 중요한 keyword 등 또는 이들의 조합이 될 수 있다.

[0144] 상동성(유사성) Query는 Query와 상동성(유사성)을 주는 모든 컨텐츠를 의미하고, 소문자 query로 사용한다.

[0145] 데이터베이스(DB)는 레코드들을 포함하는 집단을 의미하고 레코드는 하나의 어브젝트의 단위이고 어브젝트는 각각의 논문, 바이오 원시정보(예 : 염기서열, 아미노산서열, 분자의 구성물인 원자의 좌표, 화합물 좌표,화학기호, 아미노산, 등)를 포함한다.

[0146] 2)쿼리 Q를 가지고 데이터베이스 DB를 검색한 결과를 Q x DB로 표시하고, 쿼리와 상동성 쿼리모음을 (q1, q2,..qN)으로 표현한다. 따라서 Q 및 상동성 q를 가지고 검색된 한 셋트의 자료(q1.db1, ... qN.db1, ... q1.dbN ...qN.dbN')이고 이것을 맵(Map)이라 정의한다.

[0147] 3)모듬맵/모듬맵은 맵(M1)과 맵(M2)에서 사용한 방식과 같은 방식으로 표현이 가능하다. 즉, 질병관련 유전자 모음(동일 유전자 기능성 모음, 유사한 기능 약물의 모듬, 질병 저항성 유전자 모음, 혹은 유전체, 등)과 연관된 유전자를 모아둔 맵들의 모음과 단일맵의 비교는 M1 x M2으로 표현한다.

[0148] 네 번째로 vmap(virtual map)은 전체원자기반 고분자 복합체의 시뮬레이션 방법에 관한 것으로, 대한민국 특허

등록 제10-1400717호에 개시된 바 있다.

[0149] 상기 vmap은 단백질-약물-변이의 복합체를 시뮬레이션을 통하여 자유에너지의 차이를 계산을 하고, 변이로 인한 약물의 표적에 대한저항성을 계산할 수 있고, 단백질 및 약물을 반복적으로 수행하는 경우, 특정 원자/아미노산을 고정하고, 스트레스를 많이 주면, 에너지에 가중치 개념의 증폭을 주는 효과가 있으며, 에너지기반 검출을 쉽게 할 수 있다.

[0150] 특히, 백신의 경우 본 방법에서는 여러 가지 타입의 MHC(면역 수용체 단백질) 구조를 템플릿으로 만들어 둔 후, 면역 기능에 관련된 여러 유전자 서열의 패턴 분석을 통해 얻어진 아미노산 서열들을 MHC내의 알려진 펩타이드 결합부위에 치환시켜서 펩타이드를 고정시킨 후 시뮬레이션을 진행한 후 각 펩타이드의 에너지 변화를 비교 분석하는 방법으로 MHC와 안정하게 결합하는 펩타이드를 GPGPU기반 스크리닝을 통해 예측해 낼 수 있다.

[0151] 이하에서는 본 발명에 의한 통합유전체 시스템에 관련된 본 발명의 기술적 특징을 설명하기로 한다.

[0152] 도 9는 본 발명에 적용되는 PC-cluster에서 운영을 위한 인간 유전체 30억 염기에 대한 분할된 청크를 도시한 개념도이고, 도 14은 본 발명에 의한 통합유전체 DB의 대립유전자깊이, 지노타입 및 하플로타입 DB의 구성을 도시한 개념도이다.

[0153] 도 9에 도시된 바와 같이, 본 발명의 특징 6은 전장유전체분석 및 통합유전체DB에 대한 것으로, 그 기본 구성은 대한민국 등록 특허 제10-1460520호 및 대한민국 특허출원 제10-2015-0187554호에 개시된 바 있다.

[0154] 그러나 현실적으로 pc-cluster(다중 서버급 노드)를 100 ~1,000대 연결하여 사용하는 수퍼컴퓨팅 환경)에서, 많은 CPU기반으로 동시에 많은 수의 인간 유전체데이터를 프로세싱하고 통합을 하면서도, 합리적인 비용으로 pc-cluster를 구성을 하려면, 노드(node: multiple CPU)당 처리 용량을 규정하는 것이 바람직하다.

[0155] 이에 본 발명은 인간 유전체 분석에 따라 약 64GB의 메모리를 분산 할당하는 것이 바람직하다.

[0156] 따라서, 본 발명은 이러한 pc-cluster환경에서 인간 전장유전체(30억 염기 base pair * N명)를 통합을 하기 위하여, 전장유전체를 일정한 크기 (약 50,000,000 염기 base pair * N명 = 1개 청크)로 분할하여 600개의 청크(chunks)로 만들고 운영하도록 하는 기술(특징 6)을 적용하여, PC-cluster기반 분석의 효율을 향상시킨다.

[0157] 이는 1개의 청크에 10,000명 - 20,000명의 통합된 샘플을 다중정렬된 환경으로 만들려면 약 10 - 20 GB의 메모리가 필요하므로, 이와 같이 셋팅 된 분석 플랫폼의 경제성이 현저히 상승하기 때문이다.

[0158] 한편, 도 10은 본 발명에 의한 유전체데이터 분석을 위한 계산장비(a1-a6), 스토리지(b1-b3)구성 및 마하-FsDx의 구성을 도시한 개념도이다.

[0159] 도 10에 도시된 바와 같이, 본 발명의 특징 7은 유전체 벌크데이터 운용을 위한 수퍼컴퓨터 환경의 최적화, 데이터 I/O 속도에 따른 파이프라인에서 하드웨어 및 네트워크카드 속도 테스트 및 배치 그리고 동시에 수백개 및 수천개의 계산을 할 때 사용할 수 있는 미들웨어급의 병렬분산 스토리지 환경(예, 클라우드컴퓨팅)의 구성에 관한 것이다.

[0160] 이때, 유전체분석용 수퍼컴퓨터를 구성하려면, 특성상 메모리기반 초고속 계산흐름(a1)(그러나 비교적 작은 I/O), 초고속 계산 후에 대량의 데이터(추가 I/O성능 + 추가 초고속 메모리-케시-서버 + 대형 스토리지)를 필요로 하는 경우(a2), 초고속계산을 수행하는 GPGPU의 초고속 계산 후에 대량의 데이터(추가 I/O성능 + 추가 초고속 메모리-케시-서버 + 대형 스토리지)를 위한 흐름(a3), 수천 대의 일반 계산용 서버에서 수천 개의 계산을 동시에 수행하는 흐름(a4) 및 중간형 속도를 위한 흐름(a5) 그리고 데이터 백업을 하는 흐름(a6)이 필요하다.

[0161] 그리고 스토리지 관점에서도 메모리 디스크 드라이버(b1), 병렬분산형 초고속 대용량 I/O용 스토리지(b2) 그리고 일반 스토리지(b3) 등 추가적으로 I/O성능을 내기위한 초고속 CPU 및 스토리지가 제공되어야 한다.

[0162] 한편, 본 발명의 특징 8은 이러한 대형장비 (마하-fsdx)를 임상에 활용하려면, 재현성 및 일관성이 보장이 되어야 하고, 이에 따라 시스템 한계에 대하여 명확한 정의 한정된 것을 말한다. 따라서, 현재의 마하-Fs를 마하-FsDx화 (진단장비 및 의료기기)를 위한 임상환경의 작업이 필요하다.

[0163] 다음은 평가내용의 실시예를 보여준다.

[0164] 도 11a 내지 도 11f는 본 발명에 의해 마하-Fs를 마하-FsDx 화하여 안정성, 한계성 및 재현성에 대하여 평가한 결과를 나타내고 있다.

[0165] 여기서, 검증 테스트 환경은,

[0166] 1. MAHA-FsDx Master Node : MDS(CPU L5520 4Core x2ea/64GB Memory/128GB Disk)

[0167] 2. MAHA-FsDx Data Node : DS1~5 (CPU E5630 2.53GHz x1EA/24GB Memory/2TB x5ea)

[0168] 3. MAHA-FsDx가용량 : Physical 50TB / usable 25TB (MAHA Replica 2 option (안정성))

[0169] 4. Compute Node : CPU Xeon E5-2630 v2 2.60GHz / 64GB Memory / RAID 11TB

[0170] 5. 샘플 데이터 : CHA-NS161103-0001_R1.fastq.gz CHA-NS161103-0001_R2.fastq.gz Genome TS(Target Sequence) 샘플 (420MB x2ea) 이고;

[0172] 테스트 방법은,

[0173] 1. 재현성 테스트 : Local Disk /awork01와 /maha4에서 동일하게 테스트

[0174] 2. 안정성 테스트 :5대 MAHA DS 노드 중 한대를 장애 유발 기존 데이터 이슈 상태 여부

[0175] 3. 한계성 테스트 :Compute Node 당 2Jobs 으로 테스트.

[0176] A. no caching : 30 / 40 / 60 jobs 결과

[0177] B. caching : 60 / 200 / 240 / 360 jobs 결과

[0178] C. no caching 대용량 1.4PB : 392 / 582 / 776 / 784 Jobs 결과 (추가 포함)

[0179] 4. 사용된 유전체 분석 파이프라인 및 원시 데이터

[0180] 이에 제시된 바와 같이, 본 발명에 의한 특징 7 및 8이 적용된 마하-FsDx가 마하-Fs에 비하여 재현성, 안정성 및 한계성에서 우수함을 알 수 있다.

[0182] 도 12a 및 도 12b는 한계성에 대하여, MAHA-FsDx 구성 규모 대비 최대치 성능을 검증한 예이다.

[0183] 이때, 일반 서버에 5대 구성, 각 서버는 2TB HDD 5개로 MAHA-FsDx로 구성되고, 일반 서버에 5대에 메모리 케쉬(SDRAM chching) 추가 구성하였으며, 일반 고용량 서버 10대를 구성하였고, 각 서버 6TB HDD 24개로 MAHA-FsDx로 구성하였다.

[0184] 이들 실험 결과로부터 MAHA-FsDx가 한계성이 현저히 향상되었음을 확인할 수 있다.

[0185] 도 13은 본 발명에 의한 통합유전체 DB 플랫폼기반 다중연구자 데이터베이스의 구성을 도시한 개념도이고, 도 15는 본 발명에 의한 각각의 청크 단위 메트릭스에서 SNV, INDEL 및 CNV의 검출 원리를 도시한 개념도이다.

[0186] 도 13에 도시된 바와 같이, 본 발명의 특징 9는 개인유전체맵기반 맞춤의학 플랫폼과 같은 대형 시스템에 다수의 연구개발자가 접근이 가능한 형태의 선행계산이 된 데이터 및 툴을 통합하여 분석을 할 수 있는 미들웨어 상의 표준화된 데이터베이스(mySQL 및 mongoDB)기반 사용자환경을 제공하는 것이다.

[0187] 본 발명은 통합유전체 DB에서 제공하는 다양한 툴 및 공용 툴을 사용하여 다중 연구자들이 필요한 선행 데이터를 계산 및 결과데이터 추출을 한다. 특히, SNV, INDEL 및 CNV의 기능정보, 특히 LOF (loss of function)를 polyphen 및 SNPeffecter를 사용하여 계산을 한 결과가 저장이 되고, GWAS마커, LD계산정보, eQTL, MAF, phenotype, 등 다양한 공용 통계분석 툴을 장착하여, 유전자단위 기반 환자계층화 및 독성마커 계산을 수행한다.

[0188] 도 14은 통합유전체 DB를 5,000명의 각 유전체 30억개의 염기서열을 청크단위 (5천 만개)단위로 염기서열을 나누어고 5,000명의 메트릭스 형태로 만든 대립유전자, 지노타입, 및 하플로타입 DB에 대한 예시이다.

[0189] 그리고 도 15는 각각 나누어진, 각 청크 단위에서 본 발명의 특징 9에 따라 연구개발자가 접근 가능한 형태의 선행계산이 된 데이터 (SNV, INDEL, 및 CNV)를 추출하는 계략도이다.

[0190] 또한, 도 16은 본 발명에 의한 통합유전체 DB에서 유전자 및 다중유전자기반 유전형 계산을 통한 환자계층화를 보여준다.

[0191] 본 발명의 권리는 위에서 설명된 실시 예에 한정되지 않고 청구범위에 기재된 바에 의해 정의되며, 본 발명의 분야에서 통상의 지식을 가진 자가 청구범위에 기재된 권리범위 내에서 다양한 변형과 개작을 할 수 있다는 것

은 자명하다.

[0193] 이하에서는 본 발명의 구현을 위한 기반기술 중 일부에 대한 개요를 간단히 정리하여 설명한다.

[0195] 대한민국 특허출원 제10-2015-0187554호(특허문헌 8), 대한민국 특허출원 제10-2015-0187556호(특허문헌 9) 및 대한민국 특허출원 제10-2015-0187559호(특허문헌 10)의 기술요지

[0196] 특허문헌 8, 9 및 10에 의한 질병원인 발굴 시스템은 분석데이터 입력부(100), 검색제어부(200), 결과 리포트 제공부(300), HaploScan DB(400), ADISCAN DB(500), IDA DB(600), 생리활성 DB(700) 및 레퍼런스 DB(800)를 포함하여 구성된다.

[0197] 상기 분석데이터 입력부(100)는 개인 유전체 정보를 입력받는 부분으로, DNA sequencing 데이터를 입력받는다.

[0198] 그리고 상기 검색제어부(200)는 입력된 DNA sequencing으로부터 각 유전자의 유전형, 표현형에 대한 유전형, 희귀변이, 질병변이 및 생리활성변이를 검출하는 부분으로, 이를 위해 상기 검색제어부(200)는 HaploScan엔진(210), ADISCAN 엔진(220), IDA 검색엔진(230) 및 생리활성변이 검색엔진(240)을 포함하여 구성된다.

[0199] 상기 HaploScan 엔진(210)은 상기 분석데이터(입력된 DNA Sequencying)을 후술할 HaploScan DB(400)에 저장된 Haplo MAP(414, 424)과 대비하여 유전형을 판별하는 역할을 수행한다.

[0200] 상기 HaploScan DB(400)의 구조 및 상기 HaploScan 엔진(210)의 검색 방식은 이후 다시 상세히 설명하기로 한다.

[0201] 그리고 상기 ADISCAN 엔진(220)은 입력된 분석데이터에 포함된 각 염기에 대하여 ADISCAN DB(500)과 ADISCAN 방식으로 대비하여, 집단대조군 대비 희귀성을 산출하는 역할을 수행한다.

[0202] 또한, 상기 IDA 검색엔진(230)은 이미 알려진 유전자 관련 질병변이를 검출하는 것으로, 알려진 질병변이가 저장된 IDA DB(600)와 분석데이터를 비교하여 질병변이를 검출한다.

[0203] 그리고 상기 생리활성변이 검색엔진(240)은, 단백질 대사관련 유전 변이를 검출하는 것으로, 크게 단백질-약물, 단백질-DNA 및 단백질-단백질 결합에 관여하는 아미노산에 대한 유전변이 여부를 판별한다.

[0204] 이때, 상기 생리활성변이 검색엔진(240)은 BAV DB(700)와 분석데이터를 비교하여 상기 분석 데이터 중 상기 BAV DB(700)에 저장된 단백질 결합 관련한 아미노산에 대응하는 염기들의 변이 여부를 판별하게 된다.

[0205] 한편, 상기 검색제어부(200)는 HaploScan 엔진(210) 및 ADISCAN 엔진(220)에 의해 판별된 유전형과 각 염기의 유의성(희귀성)을 진단자(또는 사용자)가 가시적으로 용이하게 파악할 수 있도록 맨하탄 플롯 및 방사형 변이 유의성 차트를 이용하여 결과리포트를 생성한다.

[0206] 그리고 생성된 상기 결과리포트는 결과리포트제공부(300)를 통해 사용자에게 제공된다.

[0207] 이하에서는 특허문헌 8, 9 및 10에 의한 질병원인 발굴 시스템의 데이터베이스 구조를 설명하기로 한다.

[0208] 특허문헌 8, 9 및 10에 의한 질병원인 발굴 시스템은 크게 HaploScan DB(400)와 ADISCAN DB(500), IDA DB(600), BAV DB(700) 그리고 Reference DB(800)를 포함하여 구성된다.

[0209] 상기 HaploScan DB(400)는 도 3에 도시된 바와 같이, 분석 대상인 개인 유전체 정보로부터 유전형을 산출하기 위해 대조군 유전자의 유전형을 정리한 DB로, 상기 HaploScan DB(400)는 도 2에 도시된 바와 같이, 단일유전자 정보데이터베이스(410)와, 다중유전자정보 데이터베이스(420)를 포함하여 구성된다.

[0210] 그리고 상기 단일유전자정보 데이터베이스(410)는 단일유전자에 대한 유전형들을 저장한 데이터 베이스로, 단일유전자 Haplo 맵(414)과 단일유전자 하플로 프리컨시 정보(412)를 포함하여 구성된다.

[0211] 한편, 도 6에 도시된 바와 같이, 상기 단일유전자 Haplo 맵(414)은 전체 대조군의 동일 유전자에 대하여, 변이 분포를 점유 비율 별로 구분(군집)하여 저장한 것으로, 각 유전자를 활용한 세계 26개 인종의 반수체(haplotype)계산 및 특정 형질의 빈도 및 각 서브-인종의 빈도를 계산하여 정리한 것이다.

[0212] 그리고 상기 단일유전자 하플로 프리컨시 정보(412)는 상기 각각의 변이에 대한 정보를 저장한 것이다. 이때, 상기 단일유전자 하플로 프리컨시 정보(412)는 변이정보를 직접 저장한 데이터일 수도 있고, 후술할 Reperence DB(800)에 저장된 정보를 위치를 표시하는 식별인자로 구성될 수도 있다. 즉, 상기 단일유전자 하플로 프리컨시 정보(412)는 인간의 39,000개 유전자와 5 천명의 세계인종에서의 각 유전자에서 빈도 및 다양한 질병연관 주석 정보를 제공한다.

[0213] 또한, 상기 다중유전자정보 데이터베이스(420)는 다중유전자에 대한 변이 분포 및 정보를 제공하기 위한 데이터베이스로, 다중유전자 Haplo 맵(424)과 다중유전자 하플로 프리컨시 정보(422)를 포함하여 구성된다.

[0214] 이때, 상기 다중유전자 Haplo 맵(424)은 다중유전자에 의해 표현형이 특정되는 유전 특성에 있어, 각 표현형 별로 전체 대조군의 관련 염기에 대한 변이 분포를 점유 비율 별로 군집화하여 저장한 것으로, 표현형(phenotype)의 원인 변이를 활용한 세계 26개 인종의 반수체(haplotype)계산 및 특정 형질의 빈도 및 각 서브-인종의 빈도를 계산하여 정리한 것이다.

[0215] 그리고 상기 다중유전자 하플로 프리컨시 정보(422)는 상기 각각의 변이에 대한 정보를 저장한 것이다. 이때, 상기 다중유전자 하플로 프리컨시 정보(422) 역시 변이정보를 직접 저장한 데이터일 수도 있고, 후술할 Reperence DB(800)에 저장된 정보를 위치를 표시하는 식별인자로 구성될 수도 있다.

[0216] 즉, 상기 다중유전자 하플로 프리컨시 정보(422)는 인간의 39,000개 유전자와 5천명의 세계인종에서의 표현형(phenotype) 연관 유전자 셋트 들의 빈도 및 다양한 질병연관 주석정보를 제공한다.

[0217] HaploScan DB(400)의 X축은 30억 염기서열이고, 상기 염기서열에서 유전자는 39,000개가 있다. 이의 스키마에서 특정 유전자(i)에서 변이가 N(개) 발견이 되었다면, 상기 변이를 Y축: 5,000명에서 haplotype 및 genotype 모두를 사용하여 군집화를 할 수 있고, 군집화가 된 형태가 HaploMap이된다.

[0218] 이때, 각 군집은 각 유전형을 의미하는데 이들의 내용을 살펴보면, 첫 번째 GP*47*0 는 그 유전형이 세계인에서 47%를 차지하고, 세계인의 평균과 비교해서 0 bit 다르고(동일하고), 두 번째 유전형 GP*25*1은 세계인에서 25%를 차지함을 의미하며, 세계인의 평균과 비교해서 1 bit 다르다는 것을 의미한다.

[0219] 또한, 다중유전기반 HaploMap도 동일한 방식에 의해 분류 및 구분된다.

[0220] 상기 ADISCAN DB(500)는 대조군 집단의 유전체 정보를 저장한 DB로, 구체적으로 집단유전체는 글로벌 게놈프로젝트 수행에 의해 공지된 유전체 정보가 활용될 수 있다.

[0221] 한편, 상기 ADISCAN DB(500)는 대조군 집단의 전장 유전체 정보를 저장하되, 인종 등의 유전형의 군을 형성하는 구분기준에 따라 구분되어 저장될 수 있다.

[0222] 이때, 상기 인종별 구분은 5개 대분류의 구분일 수도 있고, 26개 소분류의 구분일 수도 있는데, 이는 인종별 유전특성을 반영하여 변이 유전자 여부를 판별/검출하기 위함이다.

[0223] 그리고 상기 IDA DB(600)는 이미 알려진 질병과 이에 관련된 유전 변이가 저장되는 곳으로, 다양한 질병 별로 각 질병에 관련된 유전자 변이 정보 및 이들 변이 정보를 뒷받침하는 문헌 정보가 정리되어 저장된다.

[0224] 또한, BAV DB(700)에는 다양한 단백질의 바인딩 위치의 아미노산 형태를 결정하는 유전자 정보가 저장된다.

[0225] 구체적으로는, 단백질-약물, 단백질-DNA 및 단백질-단백질 간의 바인딩에 있어, 이들 결합에 영향을 미치는 아미노산과 해당 아미노산에 영향을 미치는 유전자 정보가 저장된다.

[0226] 이에 따라, 특정 대사물의 바인딩을 관장하는 아미노산에 대한 염기들에 변이가 다수 발생한 경우, 해당 분석 데이터의 피검사자는 해당 대사물에 대하여 정상적인 체내 처리가 어려워질 가능성이 높아지게 된다.

[0227] 즉, 상기 BAV DB(700)에는 알려진 질병변이를 포함하여 단백질의 약물 결합 위치, Promoter 위치 및 결합상태의 단백질 활성이 예측되는 변이들이 저장된다.

[0228] 상기 BAV DB(700)는 생리활성관련 유전자 정보를 저장하는 데이터 베이스로, 유전자와 약물, 대사물 및 음식물에 대한 저항성 및 감수성 관련정보가 저장된다. 이때, 상기 BAV DB(700) 또한, 공신력이 확보된 공지된 데이터를 연계하여 구축할 수 있고, 예를 들어, 약물은행에 공지된 6,000 여 개의 약물정보(상호작용 단백질과 바인딩 영역 정보 등), 대사물 은행에 공지된 12,000 여 개의 대사물 정보(상호작용 단백질과 바인딩 영역 정보 등) 및 DMET(drug metabolizing enzyme and transporter gene)에 있는 200여 개의 유전자의 약물 대사관련 변이 위치에 대한 정보를 활용할 수 있다.

[0229] 한편, 상기 레퍼런스 DB(800)는 알려진 유전체의 변이에 대한 정보를 저장하는 DB로, 문헌정보 뿐만 아니라 공개된 정보 데이터베이스와 연계되어 구축될 수 있다.

[0230] 예를 들어, PheWAS-GWAS(Genome wide association study) data 및 eMERGE (Electronic Medical Records and Genomics) data가 레퍼런스 DB에 적용될 수 있다.

[0231] 한편, 도시되지는 않았으나, 상기 검색제어부(200)가 임상정보 기반의 질병원인 예측 결과를 도출하기 위해 유전적 특성과 함께 고려되어야할 피검사 대상자의 환경적 소인 정보가 저장되는 임상정보 DB를 더 포함하여 구성될 수도 있다.

[0232] 이때, 상기 임상정보 DB는 개인의 환경적 요인 결과물 데이터와 집단 평균 및 기준정보가 저장된다.

[0233] 그리고 상기 개인의 환경적 요인 결과물 데이터는 개인의 종합검진 데이터 등의 임상정보 데이터일 수 있고, 상기 집단 평균 및 기준정보는 질병관리본부가 제공하는 지역사회 코호트 연구 결과를 활용할 수 있다.

[0234] 이하에서는 특허문헌 8, 9 및 10에 의한 개인 전장 유전체를 이용한 유전정보 분석 방법을 상세히 살펴보기로 한다.

[0235] 먼저, 특허문헌 8, 9 및 10에 의한 개인 전장 유전체를 이용한 유전정보 분석 방법은 먼저, 분석데이터 입력부가 분석 대상이 되는 분석 데이터(DNA Sequencing)을 수신받는 것으로부터 시작된다(S100).

[0236] 이때, 상기 분석 데이터가 DNA 조각들로 구성된 Dumy 형태로 제공될 수도 있는데, 이 경우 본 발명은 도 15에 도시된 바와 같이, 제공된 Dumy 데이터에 고집적 인덱싱을 통해 RVR 파일 형태로 DNA sequencing 을 생성하여 저장한다.

[0237] 이후, 본 발명에 의한 개인 전장 유전체를 이용한 유전정보 분석 방법은 분석 대상에 따라 크게 4가지 분석을 수행한다.

[0238] 즉, 특허문헌 8, 9 및 10에 의한 개인 전장 유전체를 이용한 유전정보 분석은 1) 유전형 판별(S200), 2) 희귀변이 산출(S300), 3) 질병변이 산출(S400) 및 생리활성변이 산출(S500)의 4가지 분석을 수행하는 바, 이하에서는 각각에 대하여 상세히 살펴보기로 한다.

[0240] [유전형 판별]

[0241] 상기 HaploScan 엔진(210)은 상기 DNA Sequencying을 HaploScan DB(400)에 저장된 Haplo Frequency(412) 및 MAP(414)과 대비하여 단일 유전자 및 표현형에 대하여 유전형이 속하는 군집 및 이에 대한 정보를 검출한다.

[0242] 구체적으로 상기 HaploScan 엔진(210)은 상기 DNA sequencying의 i번째 유전자에 대하여 상기 단일유전자 Haplo Frequency(412)의 i번째 유전자 정보와 대비하여(S211), 분석 대상인 개인 유전체의 i번째 유전자가 단일유전자 Haplo MAP(414)에 분류된 단일유전자 분류중 어느 군집에 포함되는지 여부를 판별한다(S213, S215).

[0243] 이후, 상기 HaploScan 엔진(210)은 i=1 부터 마지막까지(약 i=39,000) 반복하여 분석데이터의 전체 유전자에 대한 유전형을 판별한다(S217, S219).

[0244] 또한, 상기 HaploScan 엔진(210)은 상기 DNA sequencying을 상기 다중유전자 Haplo Frequency(422)와 대비하여(S221), 각 표현형에 대한 분석 대상 유전체의 다수 유전자의 조합이 다중유전자 Haplo MAP(424)에 분류된 다중유전자 조합의 분류중 어느 군집에 포함되는지 여부를 판별한다(S223, S225).

[0245] 이에서도 역시, 상기 HaploScan 엔진(210)은 다중유전자정보 데이터베이스(420)에 저장된 모든 표현형에 대하여 반복하여 분석데이터의 유전형을 판별한다(S227, S229).

[0246] 이와 같은 HaploScaning 과정을 통해 분석 대상 유전체에 포함된 단일 유전자 변이 및 다중 유전자 변이에 따른 유전형을 정의할 수 있다.

[0248] [희귀변이 산출]

[0249] 희귀변이는 극히 이례적인 특정 유전 변이에 의해 유발되는 염기 변이로, 일반적으로 희귀질병과 관련된 경우가 많은 것으로, 특정 염기에 대한 변이 유무 또는 차이를 검출하여, 희귀질병 발병 가능성 등을 판단할 수 있다.

[0250] 이를 위해 본 발명은 먼저, 도 14에 도시된 바와 같이, ADISCAN 엔진(220)이 대조군을 선별한다(S310).

[0251] 이때 상기 대조군이란, 해당 변이에 대한 희귀성을 판단하게 될 대조 집단으로, 특정 인종을 한정하거나 특정 국가를 대상으로 한정할 수도 있다.

[0252] 이후, 상기 ADISCAN 엔진(200)은 특정 로커스의 염기에 대하여 대조군 DB의 염기와 ADISCAN 방식으로 변이지수를 산출하고, 이와 같은 과정을 전체 유전체에 대하여(n=1 부터 n=약 30억) 수행한다(S320, S330, S340).

[0253] 이에 따라 전체 염기서열에 대하여 염기들의 희귀성을 산출한다(S350).

[0254] 한편, 상기 희귀변이 산출을 위한 ADISCAN(allelic depth and imbalance scanning)이란 정상과 이상 유전자의 차이를 주는 마커들을 스크리닝하는 기법으로, 대립유전자깊이곱탄젠트차이, 대립유전자제곱승차이, 대립유전자절대값차이, 기하학적대립유전자차이, 통계적대립유전자차이 또는 대립유전자불균형비율에 따라 판단된다.

[0256] [질병변이 산출]

[0257] 상기 질병변이 검출은 IDA 검색엔진(230)이 분석데이터를 IDA DB(600)의 변이정보와 비교하여, 해당 질병의 위험도를 판단하게 된다(S410).

[0258] 이와 같은 방법으로, 상기 IDA DB에 포함된 모든 질병에 대하여 상기 분석데이터를 검토한 후(S420), 유의미한 변이관련 질병들을 산출하게 된다(S430).

[0260] [생리활성변이 산출]

[0261] 상기 생리활성변이 검출은 생리활성변이 검색엔진(240)이 BAV DB(생리활성변이 DB)를 검색하여(S510), 단백질의 결합에 관여하는 아미노산에 정보를 검출한다(S520).

[0262] 이때, 상기 단백질 결합은 단백질-약물, 단백질-DNA 및 단백질-단백질의 결합을 포함하고, 상기 아미노산 정보에는 상기 아미노산에 관련된 염기의 정보가 포함된다.

[0263] 이후, 상기 생리활성변이 검색엔진(240)은 상기 아미노산 정보에 포함된 염기와 분석데이터를 대비하여 분석 데이터 상에 변이가 발생 된 아미노산 및 이에 관련된 대사물 정보 등을 검출한다(S530, S540).

[0264] 그리고 상기 생리활성변이 검색엔진(240)은 전체 아미노산에 대하여 변이 검출을 반복수행하고, 검출된 정보를 통합하여 생리활성변이정보를 산출한다(S550, S560).

[0266] *이후 상기 검색제어부(200)는 판별 또는 산출된 유전형, 희귀변이, 질병변이 및 생리활성변이를 통합하여, 사용자에게 제공될 결과리포트를 생성한다(S600).

[0267] 이때, 상기 검색제어부(200)는 피검사자의 임상정보가 제공된 경우 이를 바탕으로 임상정보 기반 질병원인을 산출하여 제공할 수 있다.

[0268] 구체적으로, 질병의 원인을 예측하려면 현 상태의 환경적인 요인 결과물(종합검진데이터 및 임상정보)을 포함하는 PHR (personal health records)이 필요하다. 특히, 환경적인 요인에서 집단의 평균 및 기준정보가 필요하게 된다(본 발명에서 상기 기준정보는 질병관리본부에서 제공하는 제2기 지역사회 코호트 연구결과를 활용). 여기서, 이러한 환경적인 요인의 결과물과 유전형과 연계를 지은 것을 PHR-trait 이라고 부른다.

[0269] 질병원인 관계도(Ⅱ) 검출식은, logistic regression분석 방법을 활용한 것으로, 변수 χ는 전술한 바와 같이 산출된 유전형, 희귀변이, 질병변이 및 생리활성변이에 따라 결정되는 값이고, 변수 β는 상기 PHR로부터 결정되는 값이다.

[0270] 즉, 상기 질병원인 관계도는 Gene, Disease 혹은 Drug의 유전형 (group or cluster of genotypes) vs. PHR (BMI, AGE, SEX, 등)의 연관성을 계산할 수 있게 된다.

[0271] 따라서, 현재의 임상상태 (clinical condition: normal, disease, or phenotype)와 39,000유전자에서 계산한 Gene, Disease, Drug유전형과의 연관성을 계산하여 전체유전자기반 질병원인을 계산한다.

[0272] 한편, 특허문헌 8, 9 및 10에 의한 질병원인 발굴 시스템은 산출된 유전자 변이정보로부터 리포팅 데이터를 생성한다.

[0273] 이때 산출되는 결과 리포트는, 산출물에 따라 각각 다소 차이는 있으나, 기본적으로 변이 유전자에 대한 가시화를 위해 매하탄 플롯 및 방사형 변이 차트를 활용한다.

[0274] 상기 맨하탄 플롯(Manhattan plot)은 39,000 개의 유전자에 대하여, 알려진 모든 SNP의 non-sym 변이들을 기준으로 게놈프로젝트의 표준 유전자를 유전형에 따라 분류하여 누적된 값을 점(point)으로 가시화 한 그래프를 의미한다.

[0275] 이에 분석 대상 유전체의 유전자를 표시하면, 대조군 대비 분석 대상 유전자의 변이 특이성을 용이하게 인식할 수 있다.

[0276] 이와 같은 맨하탄 플롯(Manhattan plot)은 변이 로커스를 손쉽게 파악할 수 있을 뿐만 아니라, 변이 정도도 용

이하게 파악할 수 있다.

[0277] 한편, 상기 맨하탄 플롯에 의해 표시된 유의성 변이들은 변이 정도 및 유전적 특성에 따라 방사형 변이 차트로 표시될 수 있다.

[0278] 이때, 상기 분석 대상 유전체의 유전적 변이 정도와 대조군 평균을 함께 표시하여, 분석 대상 유전체의 변이정도를 가시적으로 명확하게 표시할 수 있을 뿐만 아니라, 유전적 특성 정보를 추가적으로 포함시켜 결과리포트를 생성할 수도 있다.

[0279] 전술한 바와 같은 방법으로 생성된 상기 결과리포트는 결과리포트 제공부를 통해 제공된다.

[0281] 대한민국 특허출원 제10-2016-0096996호(특허문헌 11)의 기술요지

[0282] 특허문헌 11에 의한 하플로타이핑 시스템 및 방법은 인간의 전장 유전체에 대한 하플로타이핑에 적용될 수도 있고, 특정 영역의 SNP에 대하여 적용될 수 있다.

[0283] 여기서 특정 영역이라 함은 특정 기능 수행에 관련된 유전자(또는 유전자들의 조합) 영역을 의미하는 것으로, 대표적으로는, 인간의 면역체계조절기능을 담당하는 인간 백혈구 항원 유전자(HLA gene) 영역, 약물대사관련 기능을 담당하는 유전자(DMET gene) 영역, 면역세포 발현에 관련된 유전자(KIR gene) 영역 및 혈액 특성에 관련된 유전자(ABO gene) 영역 등이 될 수 있다.

[0284] 따라서, 특허문헌 11은 인간 전장 유전체에 대한 하플로타이핑뿐만 아니라, HLA 타이핑, DMET 타이핑, KIR 타이핑 및 ABO 타이핑 등 특정 영역에 대한 하플로타이핑에도 적용될 수 있다.

[0285] 여기서, DMET (Drug Metabolizing Enzymes and Transporters)은 약물의 흡수(absorption)와 처리(disposition), 약물작용에 관여하는 단백질 효소(enzymes)와 전달자(transporters)들을 일컫는다.

[0286] 예를 들면, cytochrome p450 enzyme family (CYPs), uptake transporters, efflux transporters 등이 이에 속하고, 한 혈족(family) 안에 여러 개의 유전자가 있으며, 이들의 유전자 서열은 서로 비슷하면서도 다형성(polymorphism)을 갖는다.

[0287] 개인간 DMET 유전자 서열 차이는 약물반응, 부작용, 질병민감성 등에 영향을 미칠 뿐만아니라 적절한 약물선택의 기준이 될 수 있기 때문에 최근 약물유전학(pharmacogenetics)에서 주목받는 연구분야이다.

[0288] 그리고 KIR (Killer-cell Immunoglobulin-like Receptors)은 Natural killer (NK) cell이나 T cell 과 같은 특정 면역세포의 표면에 발현되는 단백질이다.

[0289] KIR은 다른 세포의 표면에 있는 major histocompatibililty (MHC, 주조직적합성) class I 과 상호작용함으로써 NK cell과 T cell의 세포를 죽이는 능력을 조절한다.

[0290] 따라서, KIR의 이러한 기능은 감염, 자가면역질환, 암 등에 대한 민감성과 반응성향과 관련이 있다.

[0291] 그리고 상기 KIR은 매우 다양(polymorphic)하여 유전자 서열이 개인마다 차이가 크며, 개인마다 가지고 있는 유전자 양이나 종류가 다르다.

[0292] 한편, ABO(blood type)는 ABO 혈액형과 수혈관계를 따지는 데 주요한 역할을 하는 유전자로, 크로모좀 9q34에 위치해 있으며 전통적인 혈청기법으로는 3개의 대립유전자(allele)(A, B, O types)을 구분할 수 있다.

[0293] A, B, O 각각의 대립유전자(allele)에도 세부그룹(subgroup)이 존재하며 드물게 같은 혈액형이라 하더라도 세부그룹(subgroup)간에 수혈이 불가능한 문제가 생기기도 한다.

[0294] 이하에서는, 설명의 구체성을 확보하기 위해, HLA 타이핑을 대표적인 실시예로 설명하기로 한다.

[0295] 주요 조직 적합성 복합체 분야(The major histocompatibility complex regions)는 휴먼게놈(human genome) 중에서 가장 복잡한 영역 중 하나이고 인간의 면역체계 조절 기능(the regulation of the immune system)을 책임지고 있다. 그 중 인간의 백혈구 항원(the Human leukocyte antigens, HLAs) 유전자는 6번 염색체(chromosome)의 약 3Mbp stretch에 존재하고 병원균(pathogen)을 억제하고 제거하는 적응형 면역 반응(adaptive immune response)에 큰 역할을 담당한다.

[0296] 임상 관점에서는 장기이식을 할 때 기증자(donor)와 수증자(recipient) 간의 HLA 유전자가 유사할 경우 거부반응(rejection)의 위험을 줄일 수 있다. 따라서 정확한 HLA 타이핑(typing)은 매우 중요한 문제이다.

[0297] 그러나 HLA 유전자(genes)의 높은 다형성(highly polymorphic), 연관불균형(linkage disequilibrium) 및 유전자간 서열 유사성(sequence similarity) 때문에 정확한 HLA 타이핑은 매우 어렵다.

[0298] 예를 들면 엑손(exons) 2-4 of HLA-A gene in class I에 대해 IMGT/HLA 데이터베이스(database)에 보고된 대립유전자(alleles)는 수 천 개가 존재하고, HLA-A, B 및 C genes 간의 대립유전자(alleles)들은 매우 유사하다.

[0299] 낮은 해상도(2-digits)에 의해 같은 항원 펩티드(antigen peptide)일지라도 아미노산(amino acid)의 차이로 인해 동종 반응(allogeneic response)을 유발할 수 있기 때문에 아미노산 수준(amino acid level)의 고 해상도(4-digits)까지 HLA 타이핑(ydping)이 필요하다.

[0300] 고해상도(High resolution) HLA 타이핑(typing)의 기존 방법은 특정 올리고 뉴클레오티드 시퀀스(SSO)에 의한 PCR 법(polymerase chain reaction by sequence specific oligonucleotide)과 SBT(sequence-based typing)법이 있지만 이와 같은 방법은 작업인력의 노동력에 의존하여 처리되어, 낮은 처리량(low-throughput)과 고비용이 문제시된다.

[0301] 한편, TAS(Targeted amplicon sequencing) 접근법은 PCR법에 비해 상대적으로 높은 처리량(high-throughput)을 나타내므로, 저렴한 비용으로, 수백 bases의 long reads를 생성하여 높은 정확성을 가지는 HLA 타이핑이 가능하다.

[0302] 그러나 효율성과 비용 때문에 최근 생성되고 있는 대다수의 데이터는 genome-wide sequence, whole genome sequence (WGS) 또는 whole exome sequence (WES)이고, 이와 같은 데이터는 long reads가 아닌 short sequence reads (~101bp)를 가진다. 따라서 이와 같은 short sequence reads 이용하여 TAS 접근법과 같은(또는 그 이상) 정확도와 경제성을 갖춘 HLA 타이핑에 대한 필요성이 대두되고 있다.

[0303] 즉, 특허문헌 11에 의한 HLA 타이핑은 short read를 이용하여, 정확성 및 효율성이 확보된 HLA 타이핑을 제공하기 위한 것이다.

[0304] Short sequence reads를 이용한 HLA typing 방법은 크게 두 분류로 나뉜다.

[0305] 하나는 short reads들을 조합(assemble)하여 긴 콘티그(contigs)를 생성하여 전체 HLA type을 결정하는 것이고, 다른 하나는 알려진 대립 유전자 시퀀스(allele sequences)를 레퍼런스(reference)로 하여 short sequence reads 들을 정렬(align)한 후 정렬된 정보로 실제 대립 유전자(alleles)를 결정하는 방법이다.

[0306] 조합(Assembly)에 기반한 방법은 short reads를 사용할 경우 phasing issue로 인한 대립유전자의 부정합(false positive allele) 판정 문제를 해결하기 어렵고, 요구되는 시간도 길어지게 된다.

[0307] 한편, 얼라인먼트(Alignment)에 기반한 방법은 HLA 유전자 영역의 높은 다형성(high polymorphic)으로 인해 알려진 대립유전자(alleles)들이 매우 유사하기 때문에 실제 대립유전자(alleles)를 결정하는 것이 쉽지 않다.

[0308] 이러한 문제점에도 불구하고 연구자들의 많은 관심 속에 조합에 기반한 방법으로는 HLAreporter가 소개되었고, 얼라인먼트에 기반한 방법으로는 PHLAT 등이 소개되었으며, 최근에 발표된 HLAreporter과 PHLAT은 이전 HLA 타이핑에 비하여 정확한 HLA 타이핑 결과를 나타낸다.

[0309] 특허문헌 11에 의한 HLA 타이핑은 genome-wide short sequencing data에 대해 매우 정확한 HLA 타이핑을 수행하는 것으로, 이하에서는 이를 HLAscan이라 칭한다.

[0310] PHLAT등의 종래기술에서는 정렬된 리드(aligned read)와 유전자 깊이(depth coverage)로 대립유전자 후보군(candidate alleles)을 선별하였으나, 본 발명에 의한 HLAscan은 대립유전자(alleles)에 정렬(align)된 리드의 분포도(read distribution)를 이용한다.

[0311] 또한, 특허문헌 11은 phase issue로 선택된 대립유전자의 부정합(false positive alleles)을 제거하기 위한 알고리즘(이하 '고유리드 연산 알고리즘'이라 한다)이 적용된다.

[0312] 특허문헌 11에 의한 HLAscan을 이용하여, 시험한 결과, 11개의 1000 genome samples, 51개의 HapMap samples, 자체 5개의 samples 에 대하여 종래기술에 비하여 정확성이 매우 향상된 결과를 보였다.

[0314] 이하에서는 특허문헌 11에 의한 HLAscan의 구체적인 구성을 설명하기로 한다.

[0315] 특허문헌 11에 의한 HLAscan은,

[0316] 1) 대립유전자 후보군(Candidate alleles)을 선별함에 있어, 정렬된 리드의 분포도를 고려한 스코어 기능

(score function considering the distribution of aligned reads)을 제공하고;

[0317] 2) phase issue로 생성된 대립유전자의 부정합(false positive alleles)을 검출하기 위한 고유리드 연산 알고리즘을 제공한다.

[0319] 특허문헌 11에 의한 HLAscan은 기본적으로 정렬기반의 접근법(Alignment-based approach)을 기반으로 한 것으로, 도 2에 도시된 바와 같이, 크게 두 단계로 구분된다.

[0320] 제1단계(Tier1)는 NGS 장비로부터 생성된 원시 시퀀스 리드(raw sequence reads)를 전장 유전체 레퍼런스(whole genome reference)에 정렬(alignment)하여 이진정렬맵(binary alignment/map, BAM)을 생성한 후 HLA 유전자 영역(genes region)에 해당되는 시퀀스 리드(sequence reads)들을 선별하는 과정이다.

[0321] 제2단계(Tier2)는 먼저, IMGT/HLA database에 존재하는 모든 대립유전자(alleles)에 각각 그 시퀀스 리드(sequence reads)들을 각각 정렬(alignment)한다.

[0322] HLA-A 를 예로 들면 IMGT/HLA database에 HLA-A gene의 알려진 대립유전자(alleles)는 3182개가 존재하고, HLAscan은 이들 대립유전자(alleles)들을 레퍼런스로 하여 수집된 시퀀스 리드(sequence reads)를 각각 정렬(alignment)한다.

[0323] 그리고 정렬(alignment)된 정보를 이용하여 최종 대립유전자(alleles)를 결정한다.

[0324] 이때, 특허문헌 11은 정렬된 정보로부터 최종 대립유전자를 결정함에 있어, 후보군 대립유전자를 선별하기 위하여 정렬된 리드 분포도를 고려한 스코어기능을 제공하고; 최종 대립유전자의 선별함에 있어, phase issue를 해결하기 위해 고유리드 연산 알고리즘을 제공한다.

[0325] 이하에서는, 특허문헌 11에 의한 HLAscan에서 제공하하는 정렬된 리드 분포도를 고려한 스코어기능 및 고유리드 연산 알고리즘의 구체적인 내용을 설명하기로 한다.

[0327] **정렬된 리드 분포도를 고려한 스코어 기능(score function considering the distribution of aligned reads)**

[0328] HLAscan은 IMGT/HLA database에 저장된 수천(약 8,000)개의 대립유전자 모델(alleles from) 중에 진정 대립유전자(true alleles)를 선택하기 위해 정렬된 리드의 분포도를 고려한 스코어 기능(score function considering the distribution of aligned reads)을 사용하여 허위 대립유전자(false alleles)를 제거한다.

[0329] 이 과정에서 제거되지 않고 남은 대립유전자(alleles)는 후보 대립유전자(candidate allele)라고 한다.

[0330] 이때 상기 정렬된 리드의 분포도를 고려한 스코어 기능(score function considering the distribution of aligned reads)은 정렬된 리드의 레퍼런스 상의 분포도가 균일하게 분산되지 않은 경우, 해당 레퍼런스의 대립유전자를 허위 대립유전자로 판정하는 것을 말한다.

[0331] 예를 들어, 레퍼런스 시퀀스(Reference sequence, ref)의 position s_i 내지 e_i에 정렬(alignment)된 read_i (1≤i≤n)가 주어졌다고 가정하면, 이때, 'read_i는 ref의 position (s_i+e_i)/2 에 정렬(alignment)되었다'라고 정의된다. 그리고 리드(Read)가 정렬(align)되어 있지 않은 reference의 연속 포지션(consecutive positions)들을 noread_j (1≤j≤m)라 한다.

[0332] 이 경우, score function은,

[0333] $$score = \sum_{j=1}^{m} \left(\frac{|noreadj|}{c} \right)^4$$

[0334] (c is a constant)

[0335] 에 의해 산출될 수 있다.

[0336] 이때, 산출된 스코어가 기준치보다 크게 산출된 레퍼런스를 허위 대립유전자에 대한 레퍼런스로 판정하여, 해당 대립유전자를 후보에서 제외시킬 수 있다.

[0338] **고유리드 연산 알고리즘**

[0339] 특허문헌 11에 의한 고유리드 연산 알고리즘은 1) 후보 대립유전자(Candidate alleles)가 다수 존재할 경우 불

합치 후보 대립유전자(false positive candidate alleles)를 제거하는 알고리즘과, 2) 3개 이하의 후보 대립유전자(candidate alleles)가 존재할 경우 phase issue로 선택된 후보 대립유전자(candidate alleles)를 검출하여 제거하는 알고리즘을 포함한다.

[0340] 또한, 특허문헌 11에 의한 고유리드 연산 알고리즘은 전술한 바와 같은 판단결과를 바탕으로 최종 대립유전자가 동형접합체(homozygous)의 대립유전자 인지 이형접합체(heterozygous)의 대립유전자인지 여부를 판별할 수 있다.

[0342] 예를 들어, 타이핑할 유전자(gene)로부터 시퀀스 리드(sequence reads)를 수집하였고 시퀀싱에 오류(sequencing error)가 없다고 가정한다.

[0343] 이때, t 개의 candidate allele_i (1≤i≤t) 중 서로 다른 시퀀스를 갖는 두 개의 리드 A, B(two reads A and B which have different sequence)가 서로 다른 allele_p 및 allele_q (1≤p,q≤t)의 position x to y 에 각각 100% 매치 되어 맵핑(mapping with 100% match)되고, 다른 영역에는 맵핑(mapping)되지 않았을 때, 해당 검사체의 실제 유전자는 리드 A의 시퀀스(sequence)를 포함한 한 가닥 그리고 리드B의 시퀀스(sequence)를 포함한 한 가닥을 가진 이형집합체(heterozygous)이다.

[0344] 따라서 리드 A 및 리드 B 중 어떤 것도 mapping with 100% match 되지 않는 후보 대립유전자(candidate alleles)는 불합치 대립유전자(false positive allele)이므로 제거한다.

[0346] 또한, 특허문헌 11에 의한 고유리드 연산 알고리즘은 3개 이하의 후보 대립유전자(candidate alleles)가 존재할 때, 각각의 후보 대립유전자(candidate allele)에 대하여 오직 자신에게만 정렬(aligned)되어 있는 시퀀스 리드(sequence reads)의 개수를 카운트(count) 하여 그 순으로 제1후보 대립유전자(the first candidate allele) 및 제2후보 대립유전자(the second candidate allele)를 선정한다. 그리고 선택된 두 개의 후보 대립유전자(candidate alleles)에 대해 같은 과정을 반복한다.

[0347] 만약 두 후보 대립유전자(candidate alleles) 모두 고유 정렬 리드(unique aligned reads)를 가지고 있을 때 2개의 대립유전자(alleles)를 최종결과물로 산출한다. 이 경우 해당 대립유전자는 이형접합체의 대립유전자임을 의미한다.

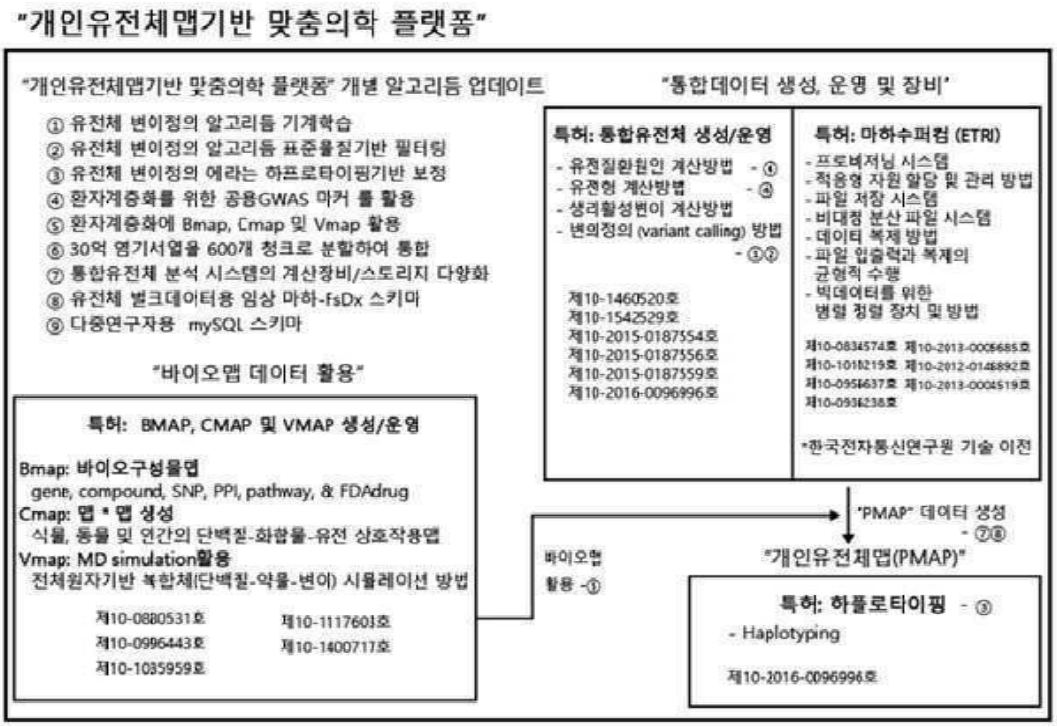
[0348] 그리고 하나의 후보 대립유전자(candidate allele)만 고유 정렬 리드(unique aligned reads)를 가지고 있을 때(하나의 allele에 aligned reads가 다른 allele의 모든 aligned reads를 포함한 경우), 고유 정렬 리드를 가진 대립유전자만을 최종 결과로 출력한다. 이 경우 해당 대립유전자는 동형접합체의 대립유전자임을 나타낸다.

산업상 이용가능성

[0350] 본 발명은 최근 가속화되어 발전하고 있는 맞춤의학 플랫폼의 핵심기술로, 본 발명에 의하면, 현재 부분적으로 독성 여부의 위험도 정도를 검출하는 수준을 넘어, 개인유전체 맵 기반 맞춤의학 플랫폼을 통해 표준화되고, 재현 가능하며 자동화된 기술을 적용하여 상용화된 맞춤의학용 플랫폼이 가능해지고, 이에 따라 인류의 의학적 발전에 크게 기여할 수 있다.

도면

도면1


"개인유전체맵기반 맞춤의학 플랫폼"
"개인유전체맵기반 맞춤의학 플랫폼" 개별 알고리듬 업데이트
① 유전체 변이정의 알고리듬 기계학습
② 유전체 변이정의 알고리듬 표준물질기반 필터링
③ 유전체 변이정의 에러는 하프로타이핑기반 보정
④ 환자계층화를 위한 공용GWAS 마커 를 활용
⑤ 환자계층화에 Bmap, Cmap 및 Vmap 활용
⑥ 30억 염기서열을 600개 청크로 분할하여 통합
⑦ 통합유전체 분석 시스템의 계산장비/스토리지 다양화
⑧ 유전체 빌크데이터용 임상 마하-FsDx 스키마
⑨ 다중연구자용 mySQL 스키마
"통합데이터 생성, 운영 및 장비"
특허: 통합유전체 생성/운영
- 유전질환원인 계산방법 - ①
- 유전형 계산방법 - ④
- 생리활성변이 계산방법
- 변의정의 (variant calling) 방법 - ①②
제10-1460520호
제10-1542529호
제10-2015-0187554호
제10-2015-0187556호
제10-2015-0187559호
제10-2016-0096996호
특허: 마하수퍼컴 (ETRI)
- 프로세저닝 시스템
- 적응형 자원 할당 및 관리 방법
- 파일 저장 시스템
- 비대칭 분산 파일 시스템
- 데이터 복제 방법
- 파일 입출력과 복제의 균형적 수행
- 빅데이터를 위한 병렬 정렬 장치 및 방법
제10-0834574호 제10-2013-0005685호
제10-1018219호 제10-2012-0148892호
제10-0956637호 제10-2013-0004519호
제10-0936238호
*한국전자통신연구원 기술 이전
"바이오맵 데이터 활용"
특허: BMAP, CMAP 및 VMAP 생성/운영
Bmap: 바이오구성물맵
gene, compound, SNP, PPI, pathway, & FDAdrug
Cmap: 맵 * 맵 생성
식물, 동물 및 인간의 단백질-화합물-유전 상호작용맵
Vmap: MD simulation활용
전체원자기반 복합체(단백질-약물-변이) 시뮬레이션 방법
제10-0880531호
제10-0996443호
제10-1035959호
제10-1117601호
제10-1400717호
바이오맵 활용 -①
"PMAP" 데이터 생성 - ⑦⑧
"개인유전체맵(PMAP)"
특허: 하플로타이핑 - ③
- Haplotyping
제10-2016-0096996호

도면2

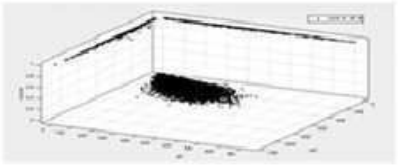
"ADISCAN 기계학습"

① • Score function $S(i) = tan(D - 0.5) \times a - log_b(\max(b, \min(depth))) + c$
- a : non-determined (중간점) 범위를 결정 (그래프상으로는 기울기)
- b : depth에 따른 변이 민감도 조절
- c : homo(최고점), hetero(최저점) 범위를 조절 (그래프상으로는 x축 평행이동)

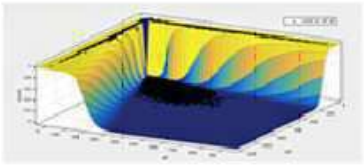
② • Machine Learning을 통한 변이 검색 함수 최적화
- 표준물질(NA12878) 변이정보를 통한 최적화

③ • Given the depth information of position i for a sample
- Score function
  - $Adiscore = \frac{1}{1+e^{-S(i)}}$
  - $S(i) = tan(D - 0.5) \times 49.11 + log_4(\max(4, \min(DP_{ref(i)}, DP_{att'(i)}))) - 13.72$,
    Where D $= 1 - R(i)$, a, b, c are constants
  - $R(i) = \frac{DP_{att(i)}+1}{DP_{ref(i)}+1}$ if $DP_{ref(i)} \geq DP_{att(i)}$
    $\frac{DP_{att(i)}+1}{DP_{ref(i)}+1}$ else

④



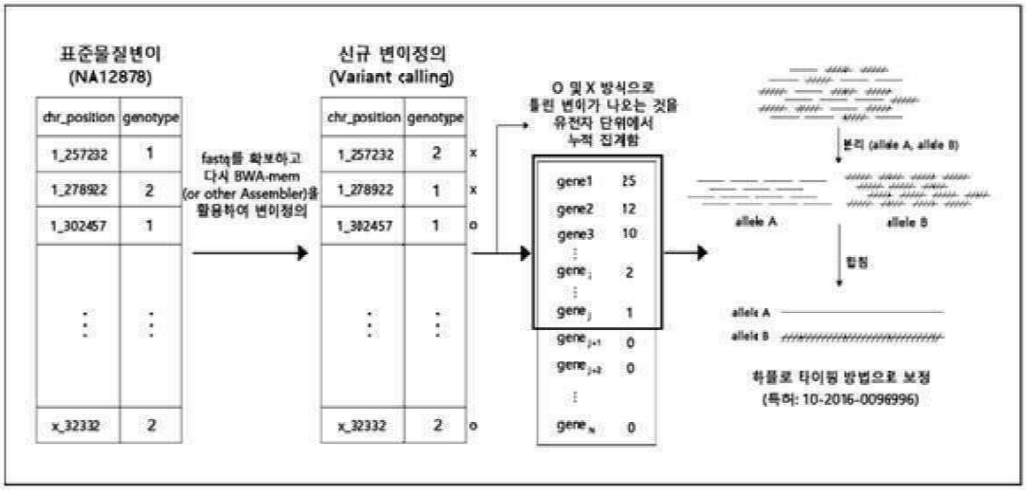
(a) 학습된 방식으로 분리한 결과



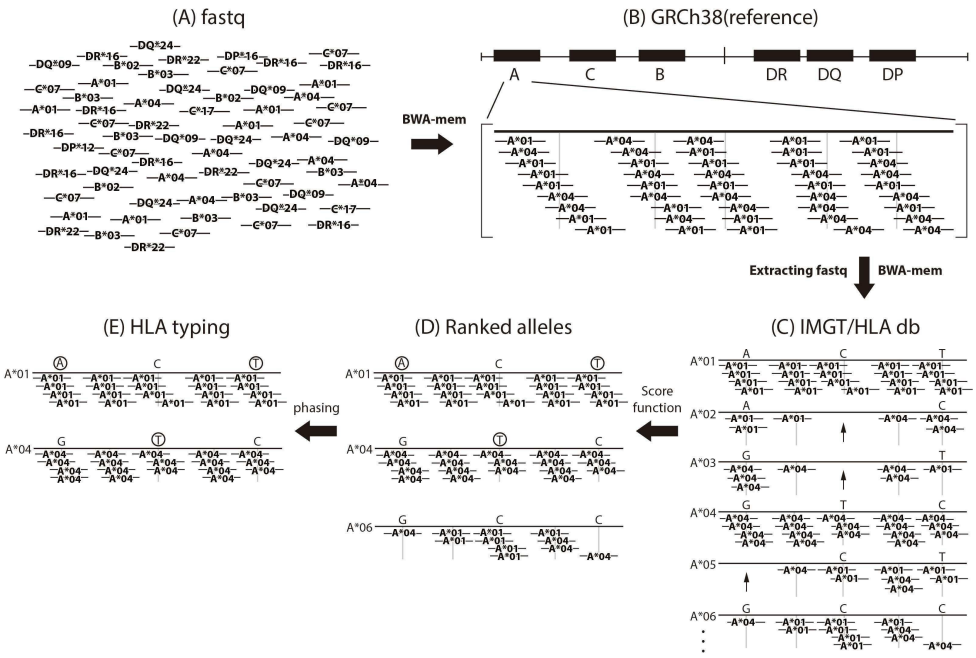
(b) 및 3D 분포 plot

도면3

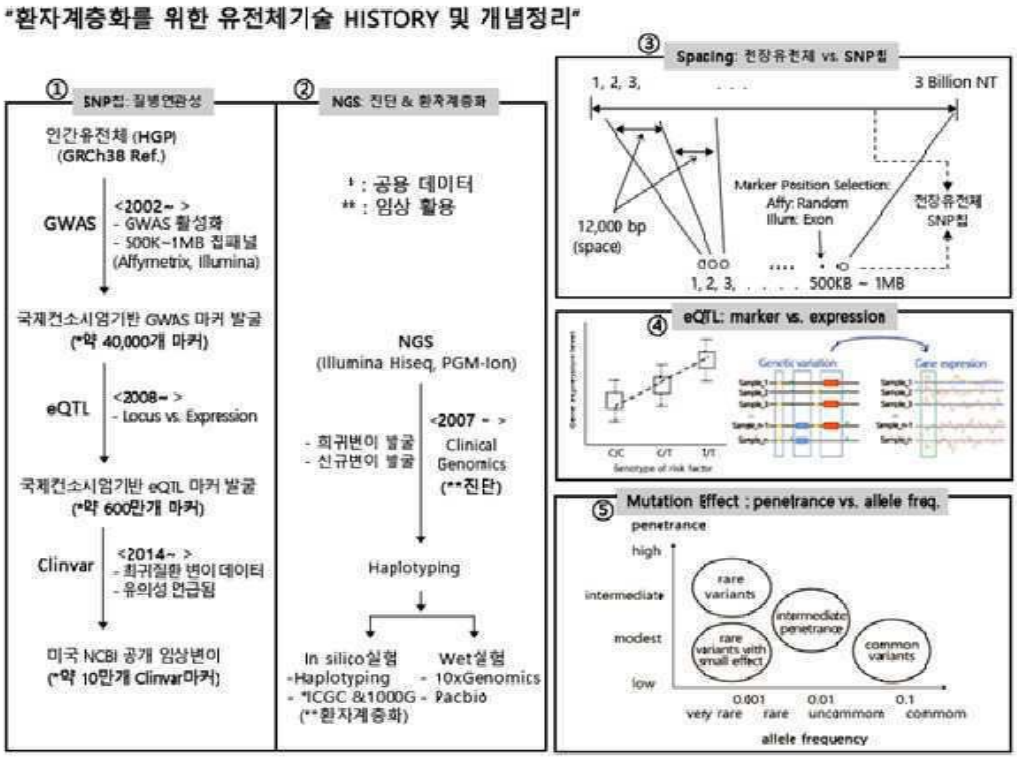
"표준물질 및 하플로타이핑기반 게놈에셈블리 에라 수정 방법"



도면4

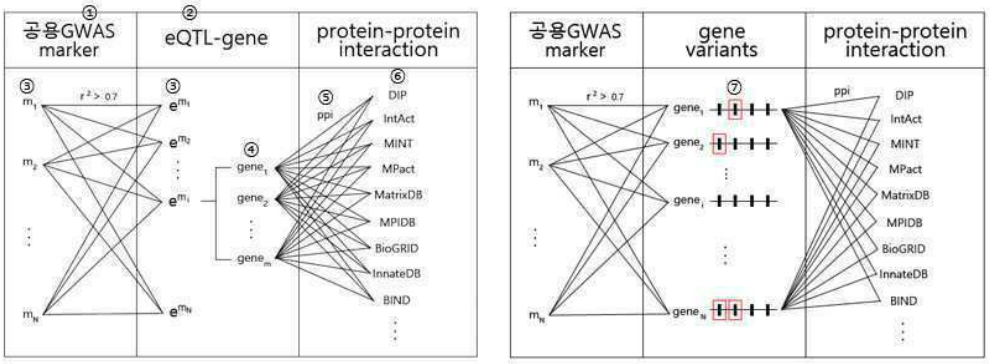

(A) fastq
(B) GRCh38(reference)
A
C
B
DR
DQ
DP
BWA-mem
Extracting fastq
BWA-mem
(E) HLA typing
(D) Ranked alleles
(C) IMGT/HLA db
phasing
Score function

도면5


"환자계층화를 위한 유전체기술 HISTORY 및 개념정리"
① SNP칩: 질병연관성
인간유전체 (HGP)
(GRCh38 Ref.)
GWAS
<2002~ >
- GWAS 활성화
- 500K~1MB 집패널
(Affymetrix, Illumina)
국제컨소시엄기반 GWAS 마커 발굴
(*약 40,000개 마커)
eQTL
<2008~ >
- Locus vs. Expression
국제컨소시엄기반 eQTL 마커 발굴
(*약 600만개 마커)
Clinvar
<2014~ >
- 희귀질환 변이 데이터
- 유의성 연급됨
미국 NCBI 공개 임상변이
(*약 10만개 Clinvar마커)
② NGS: 진단 & 환자계층화
* : 공용 데이터
** : 임상 활용
NGS
(Illumina Hiseq, PGM-Ion)
- 희귀변이 발굴
- 신규변이 발굴
<2007 ~ >
Clinical Genomics
(**진단)
Haplotyping
In silico실험
-Haplotyping
- *ICGC &1000G
(**환자계층화)
Wet실험
- 10xGenomics
- Pacbio
③ Spacing: 전장유전체 vs. SNP칩
1, 2, 3, ...
3 Billion NT
Marker Position Selection:
Affy: Random
Illum: Exon
전장유전체
SNP칩
12,000 bp
(space)
1, 2, 3, ... 500KB ~ 1MB
④ eQTL: marker vs. expression
Genetic variation
Gene expression
⑤ Mutation Effect : penetrance vs. allele freq.
penetrance
high
intermediate
modest
low
rare variants
intermediate penetrance
rare variants with small effect
common variants
0.001
very rare
0.01
rare
0.1
uncommon
common
allele frequency

도면6

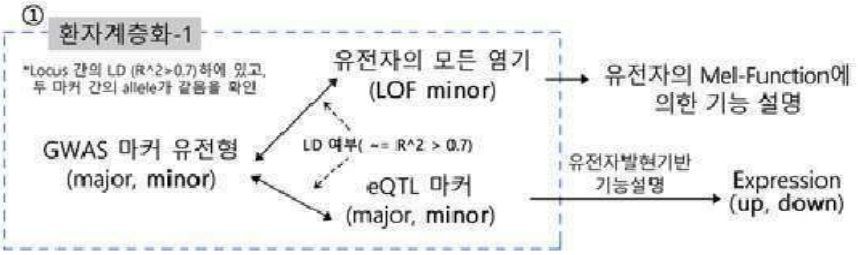
"GWAS 마커를 활용한 환자계층화"



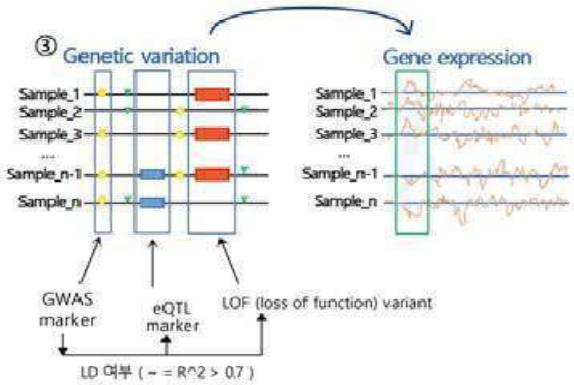
① GWAS : Gemone wide Association Study
② eQTL : expression Quantitative Trait Loci
③ m, e^{m1} : 마커를 위한 인덱스
④ gene : 마커에 기인한 발현된 유전자(mRNA)
⑤ ppi :공용 protein-protein interaction 데이터베이스
⑥ ppiDB : 공용 단백질 상호작용 DB 및 바이오맵(bmap, cmap & vmap)
⑦ LOF; 변이(SNP, INDEL, CNV, Duplication 등)를 polyphen, SNP effector 등의 툴을 돌려서 기능성 정보 LOF(Loss of Function)를 계산함. 빨강색 Box는 LOF임.
⑧ Sampling : ③의 마커에서는 minor allele를 사용하고, ⑦에서도 minor allele를 사용함

도면7

"GWAS마커 기반 환자계층화 생성 오버뷰"

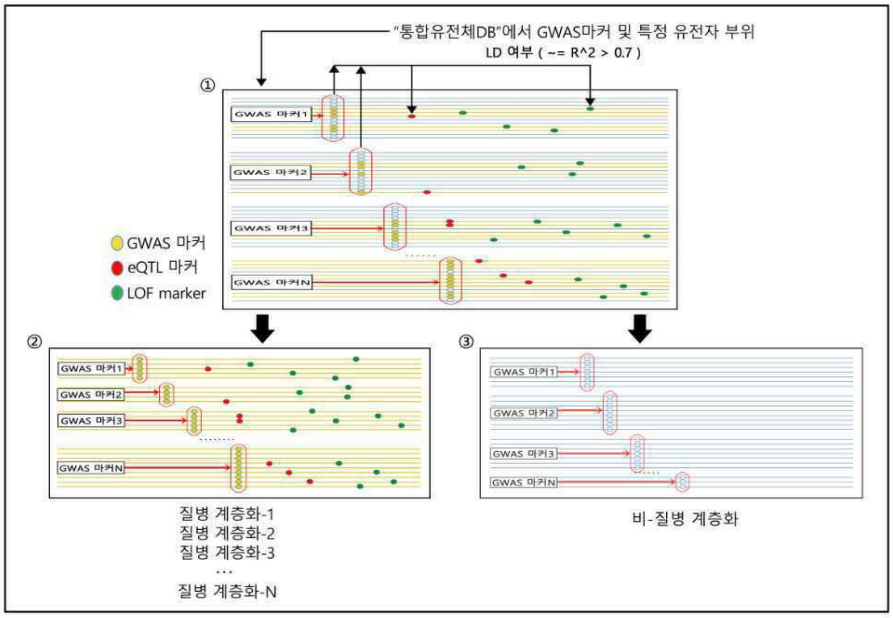


② 환자계층화 생성 규칙
- GWAS의 allele type은 특정한 경우를 제외하고, minor allele를 쓰는 것을 원칙으로 한다. *특정한 경우는 연구결과가 major를 써야 하는 경우로 한정 함.
- LOF는 GWAS의 allele type과 같은 type을 사용함. LOF자체가 기능을 설명 함.
- eQTL 마커는 GWAS의 allele type과 같은 것을 사용함.
- eQTL 마커에 기인한 Expression의 up혹은 down의 모든 유전자기반 기능설명을 함.



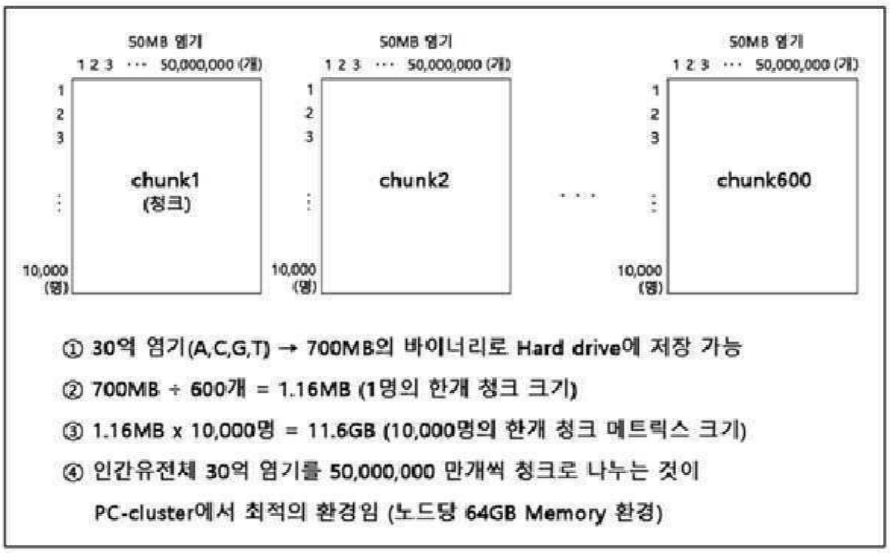
도면8

"GWAS마커 기반 환자계층화 실시 예문"

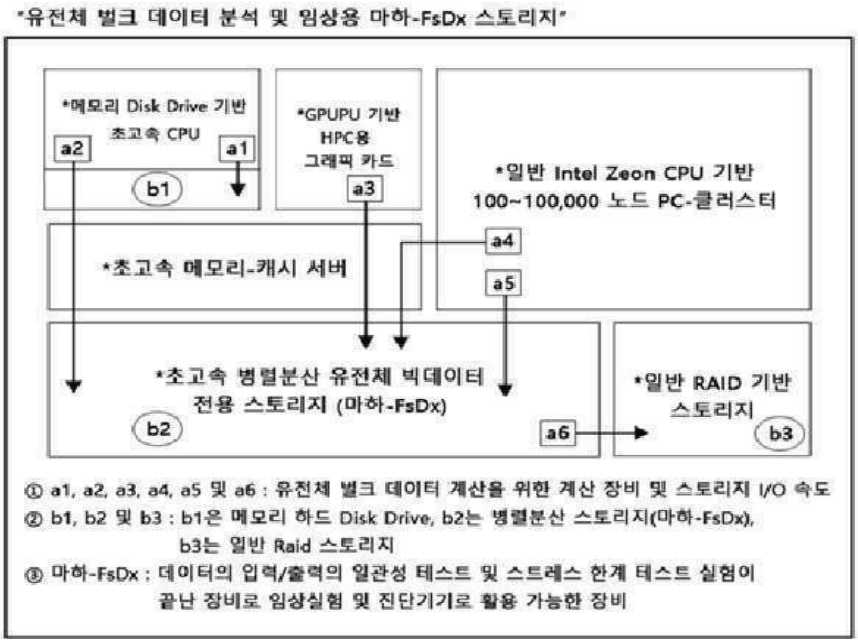


도면9

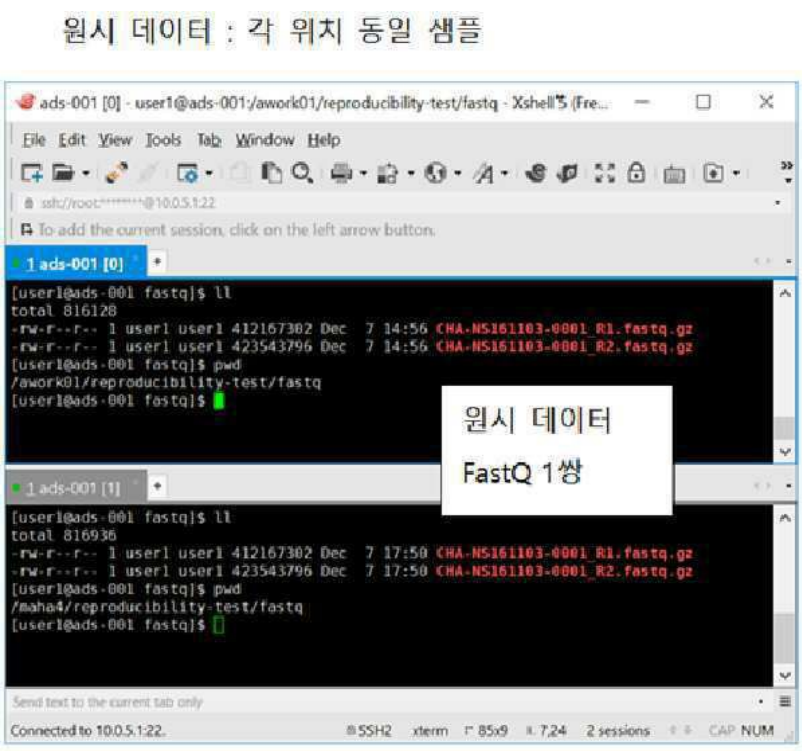
"통합유전체 데이터 관리 방법"



도면10

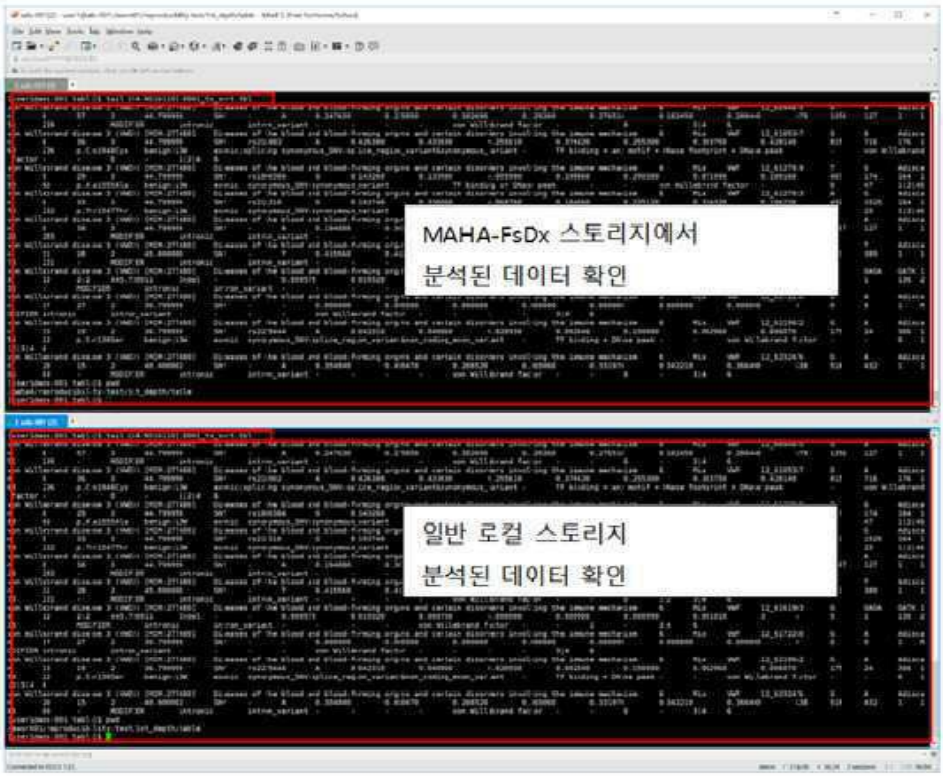

"유전체 벌크 데이터 분석 및 임상용 마하-FsDx 스토리지"
*메모리 Disk Drive 기반
초고속 CPU
a2
a1
b1
*GPUPU 기반
HPC용
그래픽 카드
a3
*일반 Intel Zeon CPU 기반
100~100,000 노드 PC-클러스터
a4
a5
*초고속 메모리-캐시 서버
*초고속 병렬분산 유전체 빅데이터
전용 스토리지 (마하-FsDx)
b2
a6
*일반 RAID 기반
스토리지
b3
① a1, a2, a3, a4, a5 및 a6 : 유전체 벌크 데이터 계산을 위한 계산 장비 및 스토리지 I/O 속도
② b1, b2 및 b3 : b1은 메모리 하드 Disk Drive, b2는 병렬분산 스토리지(마하-FsDx),
b3는 일반 Raid 스토리지
③ 마하-FsDx : 데이터의 입력/출력의 일관성 테스트 및 스트레스 한계 테스트 실험이
끝난 장비로 임상실험 및 진단기기로 활용 가능한 장비

도면11a


원시 데이터 : 각 위치 동일 샘플
ads-001 [0] - user1@ads-001:/awork01/reproducibility-test/fastq - Xshell 5 (Fre...
File Edit View Tools Tab Window Help
To add the current session, click on the left arrow button.
1 ads-001 [0]
[user1@ads-001 fastq]$ ll
total 816128
-rw-r--r-- 1 user1 user1 412167302 Dec 7 14:56 CHA-NS161103-0001_R1.fastq.gz
-rw-r--r-- 1 user1 user1 423543796 Dec 7 14:56 CHA-NS161103-0001_R2.fastq.gz
[user1@ads-001 fastq]$ pwd
/awork01/reproducibility-test/fastq
[user1@ads-001 fastq]$
원시 데이터
FastQ 1쌍
1 ads-001 [1]
[user1@ads-001 fastq]$ ll
total 816936
-rw-r--r-- 1 user1 user1 412167302 Dec 7 17:50 CHA-NS161103-0001_R1.fastq.gz
-rw-r--r-- 1 user1 user1 423543796 Dec 7 17:50 CHA-NS161103-0001_R2.fastq.gz
[user1@ads-001 fastq]$ pwd
/maha4/reproducibility-test/fastq
[user1@ads-001 fastq]$
Send text to the current tab only
Connected to 10.0.5.1:22.
SSH2 xterm 85x9 7,24 2 sessions CAP NUM

도면11b

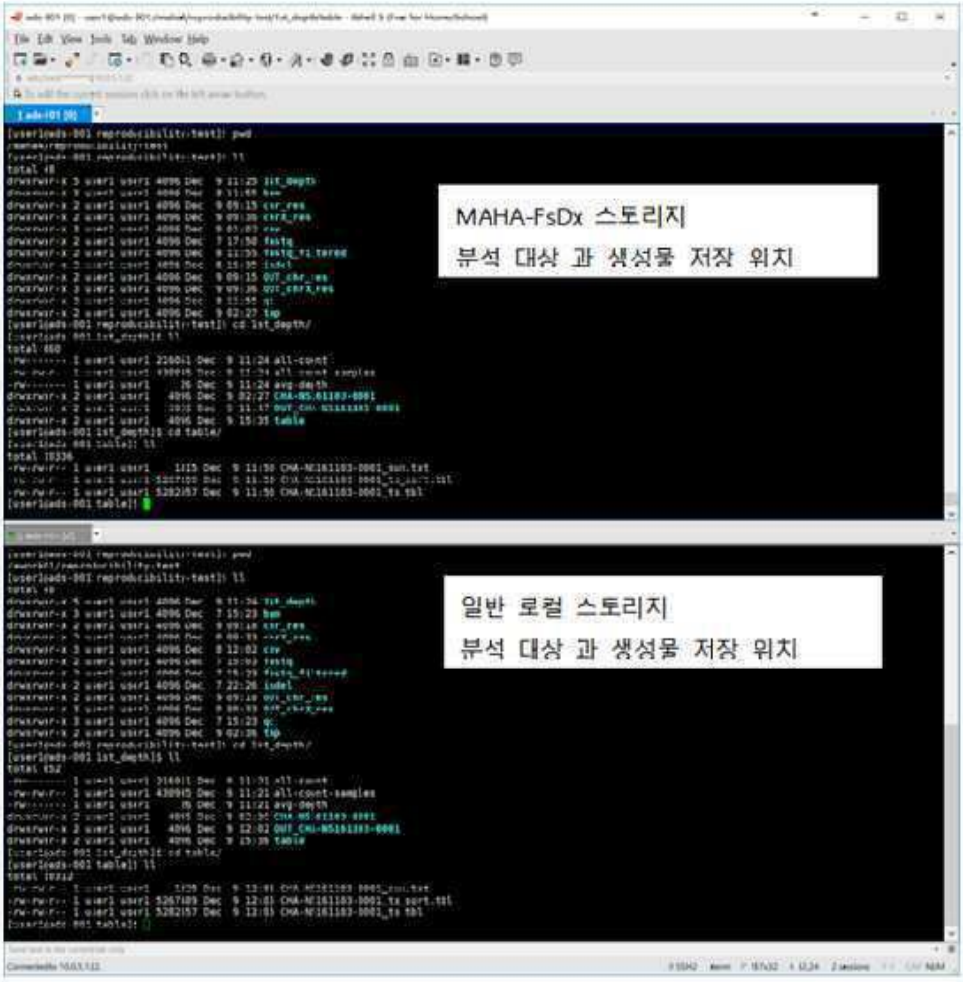
1. 재현성검증 : 로컬 디스크 와 MAHA-FsDx의 스토리지 의 동일 분석 후 데이터 재현 검증



도면11c

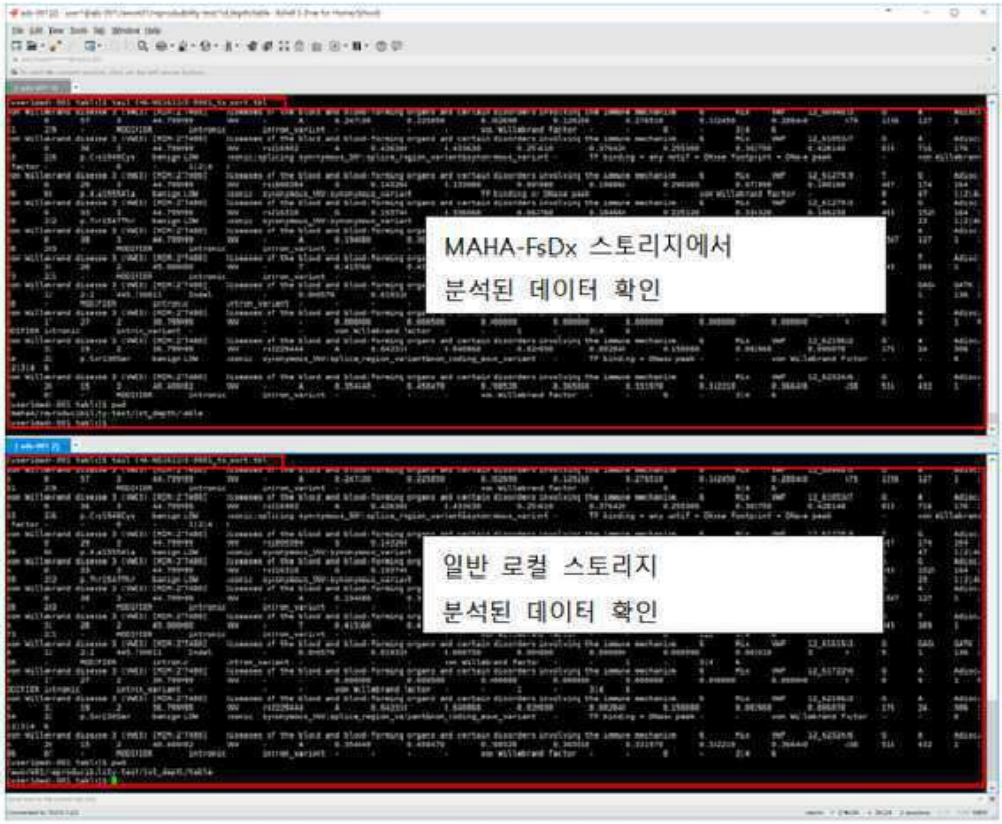
2. 안정성검증:분석 -> 결과 확인 -> 장애 유발 ->이상유무 확인 -> 복구 -> 이상유무 확인

A. MAHA-FsDx 저장 데이터 확인



도면11d

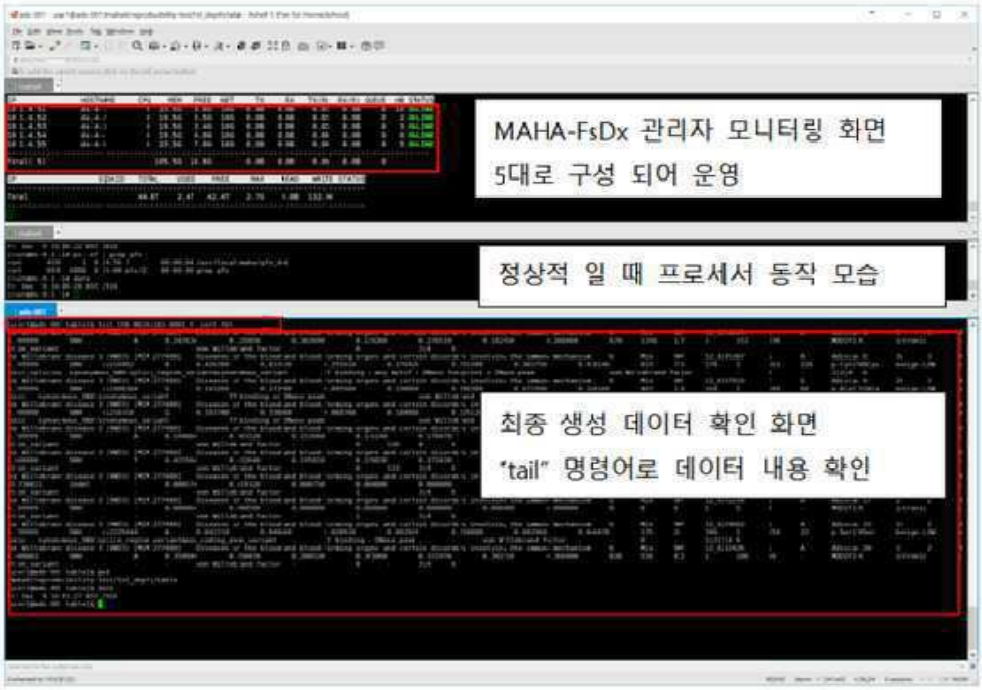
B. 데이터 분석 후 각 스토리지 생성 데이터 확인



도면11e

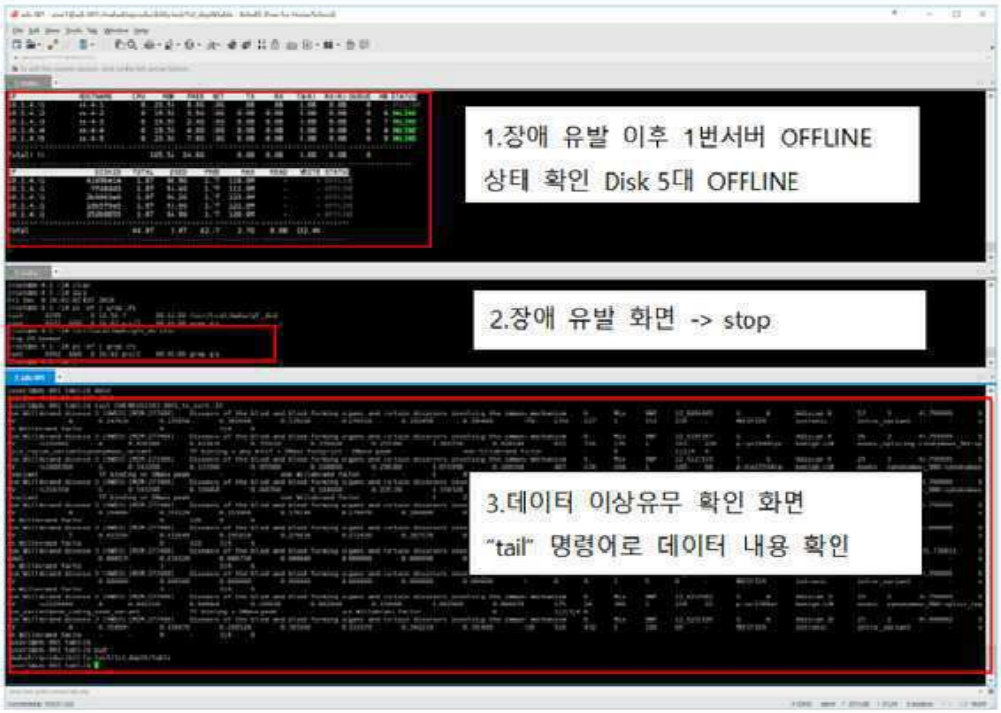
C. MAHA-FsDx 구성 DS 노드 한대 장애 유발

i. 장애 전 화면 캡쳐 -> MDS 에서 모니터링 화면

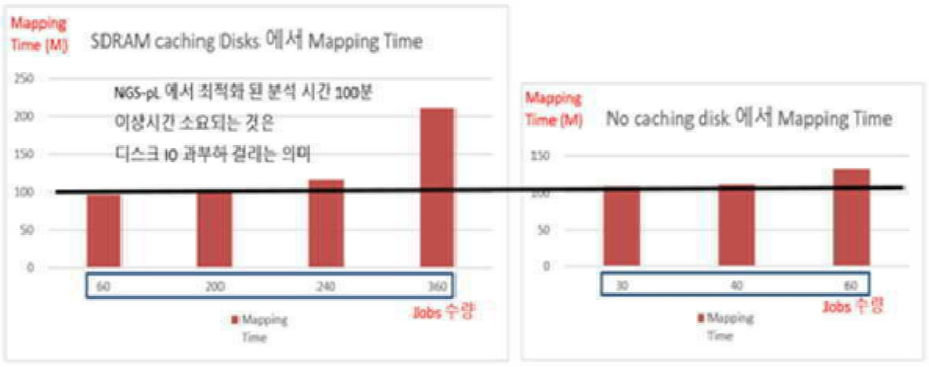


도면11f

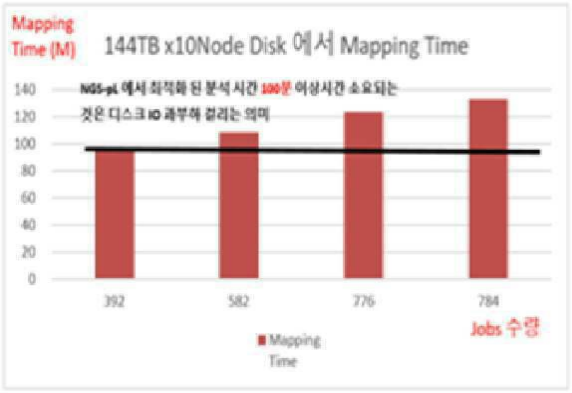
ii. 장애 유발 후 대상 서버에서 데이터 결과물 명령어로 확인 이상유무 체크 ->검증



도면12a

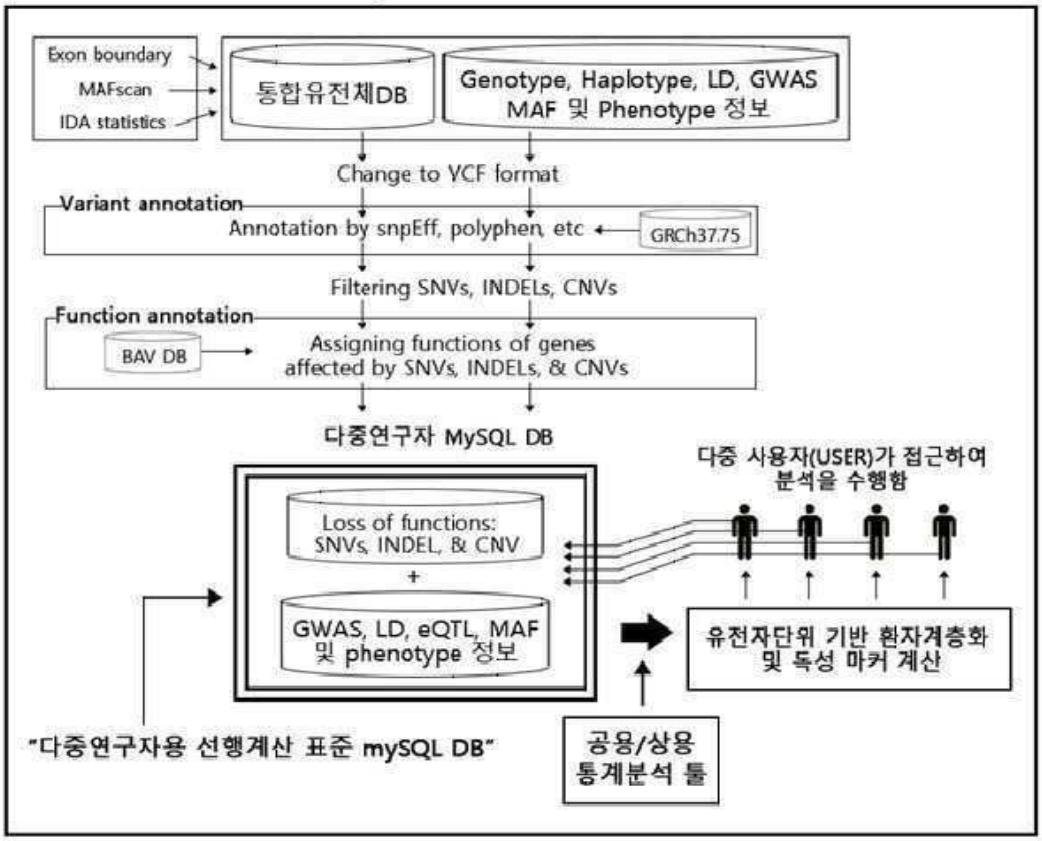

Mapping Time (M)
SDRAM caching Disks 에서 Mapping Time
NGS-pL 에서 최적화 된 분석 시간 100분
이상시간 소요되는 것은
디스크 IO 과부하 걸리는 의미
250
200
150
100
50
0
60
200
240
360
Jobs 수량
Mapping Time
Mapping Time (M)
No caching disk 에서 Mapping Time
150
100
50
0
30
40
60
Jobs 수량
Mapping Time

도면12b

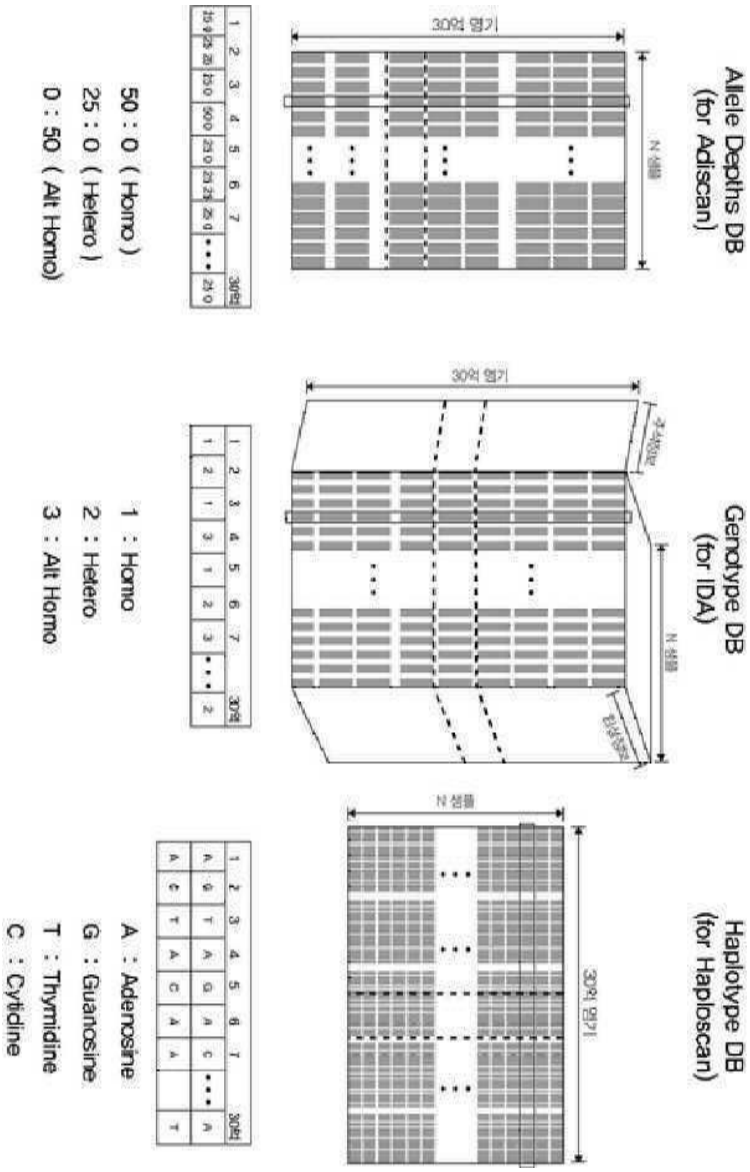

Mapping Time (M)
144TB x10Node Disk 에서 Mapping Time
NGS-pL 에서 최적화 된 분석 시간 100분 이상시간 소요되는
것은 디스크 IO 과부하 걸리는 의미
140
120
100
80
60
40
20
0
392
582
776
784
Jobs 수량
Mapping Time

도면13

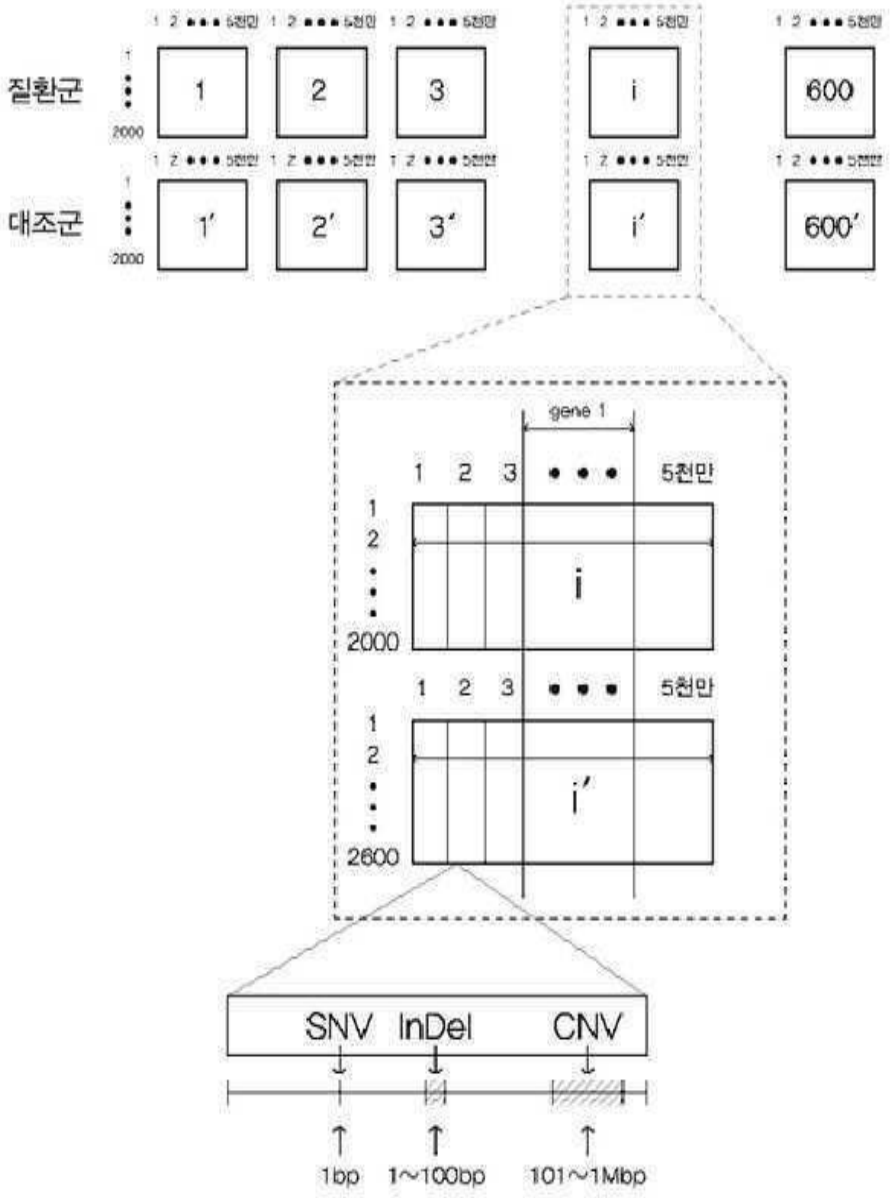
"다중연구자용 선행계산 표준 mySQL DB 계략도"



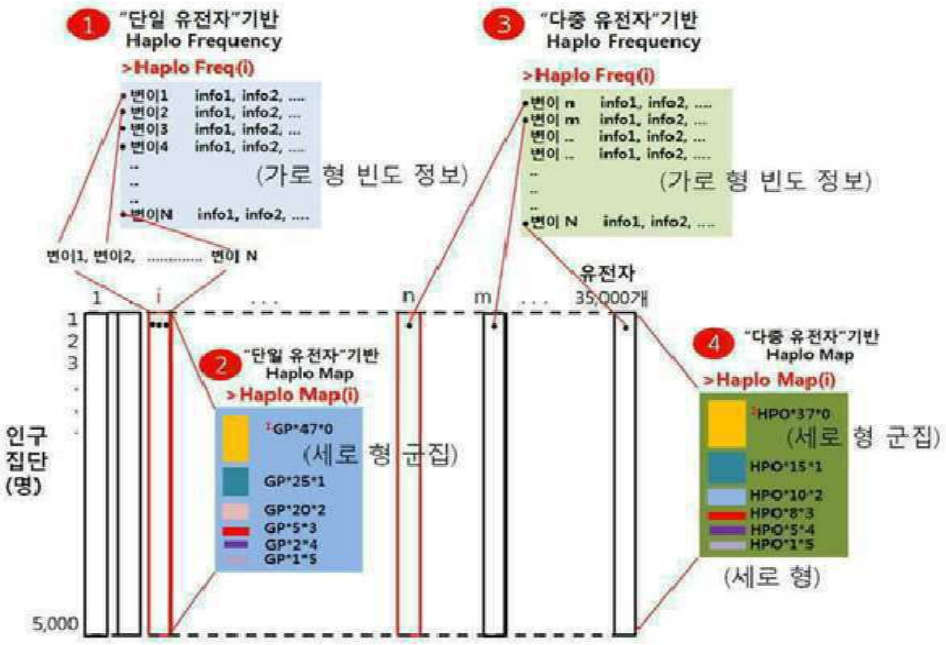
도면14


Allele Depths DB
(for Adiscan)
30억 염기
N 샘플
50 : 0 (Homo)
25 : 0 (Hetero)
0 : 50 (Alt Homo)
Genotype DB
(for IDA)
30억 염기
N 샘플
1 : Homo
2 : Hetero
3 : Alt Homo
Haplotype DB
(for Haploscan)
N 샘플
30억 염기
A : Adenosine
G : Guanosine
T : Thymidine
C : Cytidine

도면15


질환군
대조군
1
2
3
i
600
1'
2'
3'
i'
600'
2000
gene 1
5천만
2600
SNV
InDel
CNV
1bp
1~100bp
101~1Mbp

도면16


1 "단일 유전자"기반 Haplo Frequency
>Haplo Freq(i)
변이1 info1, info2, ...
변이2 info1, info2, ...
변이3 info1, info2, ...
변이4 info1, info2, ...
변이N info1, info2, ...
(가로 형 빈도 정보)
3 "다중 유전자"기반 Haplo Frequency
>Haplo Freq(i)
변이 n info1, info2, ...
변이 m info1, info2, ...
(가로 형 빈도 정보)
변이1, 변이2, ... 변이 N
유전자 35,000개
1 i ... n m ... 35,000
인구 집단 (명)
1
2
3
5,000
2 "단일 유전자"기반 Haplo Map
>Haplo Map(i)
GP*47*0
GP*25*1
GP*20*2
GP*5*3
GP*2*4
GP*1*5
(세로 형 군집)
4 "다중 유전자"기반 Haplo Map
>Haplo Map(i)
HPO*37*0
HPO*15*1
HPO*10*2
HPO*8*3
HPO*5*4
HPO*1*5
(세로 형 군집)
(세로 형)