이학박사학위논문

개인별 맞춤 약물유전체 : 생물학적 기능 요소와 개인유전체 변이를 기반으로한 약물학적 접근

Personal pharmacogenomics : pharmacological approach based on biological functional element and genomic variant

2016년 2월

서울대학교 대학원

자연과학대학 생물정보협동과정

이 수 연

논 문 초 록

1. 국문요약(국문초록)

요약(국문초록)

개인별 맞춤 약물 처방을 위한 약물 유전체(Pharmacogenomics) 연구는 각 개인별 약물효능(efficacy), 필요용량(dose requirements), 이상반응(adverse events)등의 약물학적 반응을 조절하는 유전적 변이 정보를 고려하여 약물을 처방함으로써 약물부작용(ADR;Adverse Drug Reaction)을 방지하고 치료효과를 극대화하는 데 그 목적이 있다. 최근까지는 특정 약물반응성에 차이를 보이는 실험군과 대조군(case-control)을 모집하고 모집된 환자들의 변이 정보를 전장유전체분석연구(GWAS; Genome Wide Association Studies) 기법을 사용하여 해당 약물의 반응성을 조절하는 10개 미만의 변이를 찾아내는 인구기반의 관찰연구(Population-based observational studies)가 주를 이루었으며, 미국 식약청(FDA; Food and Drug Administration)에서는 약물 처방 시에 이러한 연구결과들을 고려하여 처방하도록 권고하고 있다.

이러한 성공적인 연구결과들과 실제 적용사례에도 불구하고 인구기반의 관찰연구는 표본형성에 필요한 막대한 비용, 인구통계학적 요인/조건(Demographic condition)에 영향을 받는 연구결과, 그리고 희귀변이(Rare variant) 혹은 개인변이(Private variant)가 연구결과에 포함될 수 없다는 점 등의 많은 한계점을 드러냈다.

2013년 우리는 이러한 문제점을 극복하고자 개인의 유전적 변이 정보를 바탕으로 유전자, 약물 점수를 계산하여 개인에게 위험한 약물의 순위를 제공하는 PharmSafe 알고리즘을 개발하고 PharmGKB를 사용하여 알고리즘의 성능을 평가하였다. 하지만 암호영역(Coding region)의 변이만 사용하였다는 점과 생물학적 지식(Biological knowledge)은 단백질-단백질 상호작용(PPI;Protein-protein interaction)만 적용했다는 점을 한계점으로 제시하였다.

본 논문에서는 변이로 인하여 단백질의 기능을 변화시키는 RNA 편집 싸이트를 검출 할 수 있는 도구를 제작하고, 비암호영역(non-coding region)의 변이 정보를 사용함은 물론, RNA 편집 싸이트 등과 같이 변이에 의해 조절되는 유전자 정보를 비롯한 7가지 생물학, 약리학, 통계학적 지식요소를 가중치로 사용하여 개선된 PharmSafe 알고리즘을 개발하였으며 1092명의 개인유전체 데이터를 2503명(1000 Genome Project Phase 3)으로 확장함으로써 더 많은 개인의 유전체 데이터에서의 PharmSafe 알고리즘의 재현성 또한 입증하였다. 그 결과 7가지 지식요소를 반영한 개선된 PharmSafe 알고리즘에서는 에서는 약물학적 유전자 종류 중 약물을 분해하는 효소를 가중치로 사용한 알고리즘에서 AUC 0.5857~0.6502, (0.6224±0.222;최솟값~최댓값, 평균±표준편차)로 가장 높은 평과 결과를 얻었으며 약물 군별 평가에서는 혈압강하(Antihypertensives; n=22)군에서AUC 0.6234~0.8896(0.7340±0.0539;최솟값~최댓값, 평균±표준편차)로 가장 높은 평가 결과를 얻었다. 개선된PharmSafe 알고리즘은 환자들의 유전적 변이 정보를 바탕으로 의학적 의사결정 지원시스템(CDSS;Clinical Decision Support System)에서 각 환자별로 위험요소가 적은 약물을 알맞은 농도로 처방받는데 매우 유용하게 쓰일 것이다.

주요어 : 개인유전체, 약물유전체,약물학,맞춤의학, RNA 편집 싸이트

학 번 : 201030127

목 차

1 장 개인별 맞춤 약물유전체 : 개인유전체 변이를 기반으로한 약물학적 접근

소 개

미국 오바마(Barack Obama) 대통령은 2015년 1월 신년국정연설 당시, 2016년부터 미화 215 million달러(2150억원)를 개인맞춤의학(personalized medical treatments)을 위한 연구에 투자하겠다고 발표했다. 환자 개개인의 특성이 아닌 환자군의 평균적 특성을 고려한 기존 의학치료(medical treatments)의 만병통치약(one-size-fits-all-approach) 적인 접근 방법은 특정 개인에게는 유용한 치료이나 그렇지 못한 개인이 다수 발생하는 결과를 초래 했다고 지적하며, 이러한 문제점을 극복하기 위해 이번 대규모 투자를 통해 지놈(genome)을 이용한 생물-의학(bio-medical)연구, 특히 개인의 유전체 정보를 바탕으로한 개인맞춤형 약물처방 연구분야를 활성화할 것이며 나아가 이러한 시도가 앞으로 시행될 개인맞춤의학의 핵심이 될 것이라고 언급했다. 이처럼 현재 맞춤의학에 대한 연구 패러다임은 크게 변화하고 있으며 지놈을 이용한 맞춤의학의 연구적, 경제적 중요성이 크게 대두되고 있다.

맞춤의학(personalized medicine)이라는 컨셉은 1960대부터 사용되었고, 용어는 1999년부터 처음 사용되었다[1]. 이 용어는 개인을 위한 맞춤 의료서비스(healthcare)를 할 수 있는 의학적 모델을 제시하는 것으로 정의되었지만 보다 실질적인 연구 측면에서 중요한 의미는 개인의 약물유전체(pharmacogenomics)정보를 바탕으로 한 개인별 맞춤 약물처방을 위한 의학적 모델을 연구하고 제시하는

것이었다. [2]. 약물유전체를 바탕으로 한 맞춤의학 연구의 목적은 약물부작용(ADR;Adverse Drug Reaction)을 최소화하고 치료효과를 극대화(maximize therapeutic benefit)하며 비용 절감을 통한 경제적 효과를 얻는 것이다[2].

미국에서는 매년 2,000,000(6.7%) 이상의 입원 환자가 약물부작용으로 인한 심각한 손상을 겪고 있으며, 이 중 100,000(50%)의 환자는 치명적인 손상을 겪은 것으로 보고되었다[3]. 또한 주요사망원인 중 약물부작용이 4~6 순위를 차지할 정도로 빈도가 높고 심각한 문제이며[3, 4], 그로 인한 경제적 손실 또한 미국 $137-177 billion ,독일 €434 million, 영국GBP£2 billion으로 매우 크다 [5-7]. 제약 산업계에서도 약물부작용은 주요 부담으로 작용하고 있다. 미국에서는 실제로 약물부작용으로 인해 1990년부터 2012년까지 43개의 약물이 허가취소(withdraw)되었고[8], 캐나다 보건복지부 발표에 따르면 새로 승인된 약물의 50%에서 심각한 약물부작용사례가 보고되었고 이 중 95%가 시장 출시 이후에 발견되었다고 한다 [5].

이러한 약물부작용의 발생을 줄이는데 가장 좋은 대안으로 떠오른 분야는 각 개인별 약물효능(efficacy), 필요용량(dose requirements), 이상반응(adverse events)등의 약물학적 반응을 조절하는 유전적 변이를 비롯한 유전학적 요소를 찾는 약물유전체 연구이다 [2]. 개인 지놈에서 나타나는 변이들의 약 2-90%가 약물반응성 차이를 설명한다고 알려져 있다 [9, 10]. 이러한 약동학(Pharmacokinetics), 약력학 (Pharmacodynamics)과 연관된 변이들은 약물 수송체(drug transporter), 인간주조직적합성항원(HLAs;Human-Leukocyte Antigens), 약물 대사효소(drug metabolizing enzyme) 유전자에 존재하며 이 유전적 변이들이 약물투여량(drug dose), 약물의 혈장농도(drug plasma levels)등에 영향을 미쳐 약물반응성을 조절하여 임상적으로 예후가 좋은 치료효과를 나타내기도 하고 심각한 약물

부작용을 초래하기도 한다 [2, 11, 12]. 이렇게 유전적 변이 혹은 유전자발현으로 인해 약물의 반응성이 조절받는 대표적인 약물로는 트라스투주맙(trastuzumab)과 아바카비어 (abacavir)가 있다. 트라스투주맙(Trastuzumb)은 유전자발현으로 인해 약물반응성이 조절받는 대표적인 약물로 HER2(ERBB2) 양성 전이성 유방암 환자(HER2-positive breast cancer)의 항암요법에 사용된다. HER2가 과발현(over expression)된 후기(late-stage) HER2(ERBB2) 양성 전이성 유방암 환자에게 트라스투주맙을 투여할 경우 그렇지 않은 환자보다 중간생존시간(median survival time)이 20.3개월 대 25.1개월로 높게 나타났다. [13, 14]. 아바카비어(Abacavir)는 유전적 변이에 의해 약물의 반응성이 조절받는 대표적인 약물로 인간 면역 결핍 바이러스 감염(HIV type 1) 환자 중 인간주조직적합성복합체 유전자의 5701번째 위치에 변이(Human Leukocyte Antigen-B*5701 allele) 를 가지고 있는 환자의 48-61%가 아바카비어(abacavir)에 과민반응(hypersensitivity)을 보였다 [15]. 이러한 연구결과를 바탕으로 FDA(Food and Drug Administration)에서는 아바카비어(abacavir)의 라벨에 해당 변이에 대한 문구를 표시하도록 권고하였으며 [16] 비슷한 사례의 약물 100개 이상에 대해서도 유전변이정보를 라벨에 표기하도록 권고하였다 [17]. 1950년 근육이완제인 염화석시닐콜린(suxamethonium chloride)과 N-acetyltransferase 효소에 의해 대사되는 약물의 반응에 영향을 미치는 변이를 찾아내는 연구를 시작으로 약물반응성에 영향을 미치는 변이를 찾는 연구가 시작되었고[18] 1990년대 SNP array 기술을 거쳐 2010년 차세대 시퀀싱(NGS;Next-generation sequencing) 기법이 도입되면서 폭팔적으로 증가하게 되었다.

미화 약 30억 달러의 비용과 10년의 시간을 들여 시행된 인간게놈프로젝트(HGP;Human Genome Project)에 의해 2백만개에

달하는 인간의 유전변이가 밝혀졌고 그로부터 10년 후 차세대 시퀀싱(NGS;Next-generation sequencing)기법의 발달로 약 18,837달러의 비용과 열흘의 시간으로 한 개인의 전체 유전체를 해독할 수 있게 되었다. 그 후 차세대시퀀싱(NGS;Next-generation sequencig)기술은 다양하게 발전하여 2015년 4월 4,211달러의 비용으로 한 개인의 전체 유전체를 해독하게 되었다 [19]. 차세대 시퀀싱(NGS;Next-generation sequencing)기술의 발달에 따른 가파른 비용 감소로 인해 대량의 개인유전체를 분석하기 위한 HapMap Consortium, 1000 Genomes project, TCGA등의 컨소시움들이 만들어지고 데이터가 생성되었다. 2008년에 시작된 1000 genome project는 전장유전체 시퀀싱(Whole genome sequencing)을 이용해 14개의 인구집단(populations)에 속하는 1,092명의 개인유전체 데이터를 생성하였고 그 선행연구(pilot study)가 2010년에 완성되고 공개되었다. 그 결과 38,000,000개의 단일염기다형성변이 (SNPs;Single nucleotide polymorphisms), 1,400,000개의 짧은삽입(short insertions)과 결손(deletions), 그리고 14,000개의 큰 결손(larger deletions)을 밝혀냈으며 특히 한 개인당 평균 250-300개의 기능상실변이(loss-of-function variant)를 가진다는 것이 보고되었다 [20]. 현재까지 1000 genome project는 27개의 인구집단으로부터 추출한 2,503명의 개인유전체 데이터를 공개했다. 이러한 대규모 개인 유전체 데이터들은 약물유전체 연구에도 사용되었고 Hapmap 데이터로부터 얻은 본래의 변이 정보를 재검토(evaluate)하여 약물 관련 유전자(Pharmacogene), 변이(Pharmacovariant) 정보를 담은 데이터 베이스들이 새로 출시되었다. 특히 2009년 4월 약물유전학적 지식베이스(Knowledge base)인 PharmGKB는 35개의 HapMap CEU 샘플과 26개의 HapMap YRI 샘플로부터 38개의 약물유전학적 후보 유전자(Pharmacogenetic candidate genes)를 선정하고 해당 프로젝트를 Very Important

Pharmacogenes(VIP)으로 명명하였으며, 약물유전학과 관련된 일배체형(Haplotypes), 스플라이싱변이(Splicing variants), 그리고 해당 변이들이 포함된 유전자들의 정보등을 제공하였다 [21].

하지만 1000 Genome project를 비롯한 대량의 개인 유전체 데이터가 개인의 표현형(Phenotype)정보를 포함하고 있지 않기 때문에 이러한 데이터들을 약물유전체 연구에 적용함에 있어 큰 제약점이 되었고, 이로 인하여 개인의 유전체 정보는 물론 약물학적 표현형정보를 가지고 있는 데이터의 필요성이 크게 제시 되었다. 이로인해 Pharmacogenomics Research Network(PGRN)등의 컨소시움이 생성되었다.

약물유전체를 바탕으로 한 약물부작용(ADR) 예측 연구는 대부분 약물 반응성의 차이를 보이는 실험군과 대조군(case-control)을 모집하고 이 환자들의 혈액을 채취하여 마이크로어레이(microarray)또는 차세대시퀀싱(NGS)기법을 이용하여 개인의 유전체 서열을 해독한 후 이 데이터를 전장유전체분석연구(GWAS;Genome-wide association study)를 통해 약물 반응성을 조절하는 하나 또는 다수의 변이를 찾는 것이 대표적이다. 전장유전체분석연구(GWAS;Genome-wide association study)는 2005년 “Common disease, common variant” 라는 가설 [22] 을 바탕으로 소개되었으며 질병(Disease), 약물반응성(Drug response)등 인간에게서 나타나는 의학적 표현형(clinical phenotype)을 유발하는 원인이 되는 단일염기다형성변이(SNPs)를 검출해내는 강력한 도구로 널리 쓰여왔다. 특히 차세대시퀀싱기술(NGS)의 발달로 전장유전체분석연구(GWAS)가 폭팔적으로 증가하였으며, 그 결과 2015년 현재 638개의 연구가 GWAS catalog에 등록되었으며 9450개의 단일염기다형성변이(SNP)마커가 등록되어 있다 [23]. 전장유전체분석연구(GWAS)의 장점은 첫째 일반적으로 혈통, 가계도를 사용하는 기존 유전학적 연구에 비해 환자집단을 모으기가 수월하다는

점이다. 두번째로는 기존에 사용하던 연관성연구들(Linkage studies)에 비해 작은 유전효과(genetic effects)를 검출해 내는데 높은 통계적 힘(statistical power)을 가진다. 왜냐하면 기존에 사용하던 연관비평형(LD;linkage disequilibrium)방법은 10kb(kilobases)에서부터 몇 Mb(megabases) 범위의 유전체안의 표현형을 유발하는 변이를 검출 할 수 있는데 반해 전장유전체분석연구(GWAS)는 국소적으로 미세한 범위에 해당하는 변이의 검출이 가능하기 때문이다 [24]. 약물유전체(Pharmacogenomics)를 대상으로 하는 전장유전체분석연구(GWAS)에서 약물반응성에 크게 영향을 미치는 변이를 검출하기 위해서는 충분한 표본 크기(sample size), 치료계획(treatment protocol), 복용량(dosage), 자발적 부작용 보고여부, 인종정보(Ethnicity)등을 포함하는 환자특성 (patient features)을 만족하는 표본 형성과 모집된 환자 표본집단의 특성을 고려한 실험디자인등의 조건이 만족될 때 좋은 결과를 얻을 수 있다 [25].

2011년 7월총 48개의 약물반응성을 조절하는 변이를 검출한 전장유전체분석연구(GWAS)가 NHGRI GWAS Catalog에 등록되었으며 이중 대표적인 연구는 항혈소판제인 와파린(Warfarin), 클로피도그렐(Clopidogrel), 타목시펜(Tamoxifen)에 관한 연구이다. 특히 VKORC1과 CYP2C9 변이들이 와파린(Warfarin)의 복용량을 조절한다는 것은 전장유전체분석연구(GWAS)기법을 사용해 약물반응성에 영향을 미치는 변이를 밝혀낸 가장 대표적인 사례라고 할 수 있다. [26]. 유사한 인구통계학적 요인/조건(Demographic condition)의 CYP2C9*1/*1 유전형을 가지는 환자 49명에서의 안전한 약물효과를 위한 일일 평균 와파린용량은 7.9mg이었으나, CYP2C9*1/*3 유전형을 가지는 환자 10명에서 요구되는 일일 평균

와파린용량은 2.2mg으로 유전형에 따라 요구되는 일일 평균 와파린용량의 차이가 있다고 보고되었다 [27]. 또한 기존 학술문헌을 이용한 메타분석(Meta-analysis)에 의하면 CYP2C9의 유전형은 와파린용량 가변성(Warfarin dose variability)의 약 12%를 설명하며, VKCOR1 유전형은 약 25%를 설명 할 수 있다고 한다 [28]. 이러한 연구결과를 바탕으로 미국 FDA(Food and Drug Administration)에서는 초기 와파린용량(Warfarin dose) 결정에 있어 CYP2C9과 VKCOR1의 유전형 검사를 권장하고있다. 구체적인 예로는 항혈소판제인 클로피도그렐(Clopidogrel)을 복용한 환자 중에서 CYP2C19*17 유전형을 가지는 환자는 활성을 나타내는 클로피도그렐 대사체가 과다하게 변환되기 때문에 출혈(bleeding)의 위험성이 높다는 보고가 있으며 [29], 타목시펜(Tamoxifen)을 CYP2D6*4 동형접합성 유전형(homozygous)을 가지는 유방암 환자에게 투약하는 경우 무재발 생존비율(Disease-free survial)이 낮게 나타난다고 하는 보고가 있다 [30].

이러한 좋은연구결과들과 광범위한 사용에도 불구하고 최근 약물유전체를 대상으로 한 전장유전체분석연구(Pharmacogenomic GWAS)에 대한 많은 한계점이 제기되고 있다. 전장유전체분석연구(GWAS)에서 중요한 요소 중 하나는 표본크기(sample size)이다. 약물유전체를 대상으로 한 전장유전체분석연구(Pharmacogenomic GWAS)에서는 이 점이 중요한 잠재된 제한점으로 작용한다. 그 이유는 일반적으로 약물을 복용한 500명 중 1명의 비율로 심각한 약물 부작용(Drug adverse reaction)이 발생한다고 알려져 있다 [31]. 일반적으로 전장유전체분석연구(GWAS)에서 사용하는 실험군과 대조군 연구는 당뇨병과 같이 흔히 나타나는 질병(common disease)에 걸린 사람들을 대상으로 표본을 구성하는 것에 비해 약물유전체를 대상으로 한 전장유전체분석연구(Pharmacogenomic GWAS)는 특정 질병에 걸린

사람들 중 해당 약물을 복용했을 때 심각한 약물부작용을 겪은 사람들을 모집해야 하기 때문에 표본형성에 큰 어려움이 있다. 예를 들어 높은 콜레스테롤 수치의 환자나 고혈압 환자는 모집하기가 비교적 수월하나 높은 콜레스테롤 수치를 가진 환자 중 스타틴(Statins)에 반응하지 않는 환자나 고혈압 환자 중 베타차단제(beta blocker)에 반응성을 보이지 않는 환자군은 그 수가 확연히 작기 때문에 일정한 수치만큼 환자들을 모집하기가 상대적으로 매우 어렵다 [32]. 이러한 이유로 약물유전체를 대상으로 한 전장유전체분석연구(Pharmacogenomic GWAS)의 표본 크기는 작을 수 밖에 없고 이로 인하여 통계적 힘(Statistical power)이 작아지기 때문에 정확하게 약물반응성을 조절하는 변이를 검출하는 것이 어려우며 또한 이를 검증(Validation)하고 재현(Replication)하는 것 또한 어렵게 된다. 뿐만 아니라 현재 약물유전체를 대상으로 한 전장유전체분석연구(Pharmacogenomic GWAS)를 위해 모집된 표본의 인종은 유럽인이거나 유럽계 미국인이 대부분이고 아시아인 이나 아프리카인을 대상으로 한 연구는 매우 드물다. 따라서 기존연구에서 검출된 대부분의 변이들은 해당 민족 이외의 다른 민족이나 인종에게 적용하는것이 위험하다. 이는 개인에서도 동일한 문제점으로 작용한다. 1000 Genome Project에서 밝혀졌듯이 한 개인당 평균 250-300개의 기능상실변이(loss-of-function variant)가 존재하는데 이를 같은 결과로 해석하는 것은 아주 큰 문제가 있다. 약물유전체를 대상으로 한 전장유전체분석연구(Pharmacogenomic GWAS)의 또 다른 한계점은 1000 Genome Project 데이터를 비롯한 대규모 차세대 시퀀싱(NGS) 데이터에서 대량으로 나타난 대립유전자형빈도(MAF;Minor allele frequency) 5% 미만인 희귀변이(Rare variant)와 한 개인에게서만 나타나는 변이(Private variant)를 검출해내기 어렵다는 점이다. 실제로 여러 희귀변이(Rare variant)들이 알츠하이머병(Alzheimer's disease)을 비롯한 여러 질병을 유발하는 원인이 되는 변이(Causal

variant)라는 것은 물론, 항암제인 메토트렉세이트(Methotrexate)를 비롯한 여러 약물들의 약물반응성을 조절하는 원인변이라는 것이 밝혀져있다 [33-35]. 약물반응성을 조절하는 원인이 되는 희귀변이(Rare variant) 혹은 개인변이(Private variant)를 검출하기 위해서는 그에 맞는 큰 표본집단이 있어야 하는데 이는 현실적으로 한계가 있기 때문에 전장유전체분석연구(GWAS)를 이용하게 되면 통계적인 힘(Statistical power)이 떨어질 수 밖에 없으며 [25] 검출된 희귀변이의 신뢰성(Reliability) 또한 떨어지게 된다. 앞서 언급한 수많은 전장유전체분석연구(GWAS)의 한계점에도 불구하고 기존의 거의 모든 약물유전체 연구들이 전장유전체분석연구(GWAS)를 사용해 약물반응성을 조절하는 변이를 찾는 연구를 시행해 왔다.

2013년 우리는 인구기반의 관찰연구(population-based observational studies)인 전장유전체분석연구(GWAS)의 단점을 극복하고 개인의 희귀변이(Rare variant)와 개인변이(Private variant)를 포함한 유전적 변이정보를 바탕으로 각 개인별 위험한 약물부작용(ADR)을 예측하여 임상에서 개인마다 안전한 약물을 선택하여 처방할 때 사용될 수 있는 정보를 제공하는 개인 약물유전학(“personal pharmacogenomics”)의 개념을 제시하고 이를 실현하는 알고리즘인 PharmSafe를 개발하였다(reference). PharmSafe는 개인의 유전체 서열을 입력값으로 하고 약동학적(PK;pharmacokinetics), 약력학적(PD;pharmacodynamics) 으로 영향을 받는 유전자와 연결된 모든 약물에 대하여 유해한 정도를 점수로 제공한다. 낮은 개인의 PharmSafe 약물 점수(personalized PharmSafe score)는 개인이 해당 약물을 복용시 부작용(ADR)이 발생할 수 있는 확률이 증가한다는 것을 의미한다. 따라서 낮은 점수를 가진 약물을 개인이 복용할 때에는 복용량(dosage)을 조절하거나 복용하지 않는 것을 권고한다. PharmSafe 알고리즘은 크게 3단계로 구성되어 있다. 첫번째는

입력값으로 받은 유전체 서열에 나타난 변이들에 SIFT(Sorting Intolerant From Tolerant)점수 [36] 를 사용하여 변이로 인한 아미노산 치환이 단백질 기능에 영향을 미치는 정도를 변이점수(Variant score)로 나타낸다. 두번째는 각 변이별 점수를 해당 변이들이 포함되는 유전자 별로 기하평균(Geometric mean)을 사용하여 요약(Summarize)하고 이를 유전자점수(Gene score)라고 명명하였다. 세번째는 DrugBank [37], KEGG drug [38] 등에 포함된 약물과 유전자 사이의 관계를 이용하여 유전자점수(Gene Score)와 같이 기하평균(Geometric mean)을 사용하여 약물점수(Drug score)를 계산한다. 앞서 언급한 바와 같이 변이점수(Variant score), 유전자점수(Gene score), 약물점수(Drug score) 모두 낮을 수록 유해함을 나타낸다. 앞선 연구에서는 입력값으로는 1092명의 개인 유전체를 포함한 1000 Genome Project phase 1 데이터를 사용했고 정답값(Gold standard)으로는 대표적인 약물유전학 지식베이스인 PharmGKB [39] 의 정보를 이용하여 AUC(Area under curve)를 계산함으로써 PharmSafe 알고리즘의 효과를 입증하였다. PharmSafe의 알고리즘 성능을 평가하기 위하여 두가지 방법을 채택하였다. 첫번째는 인종정보없이 평가한 비인종평가 (Ethnicity-non-specific validation), 두번째는 1000 Genome Project 데이터가 포함하고 있는 4가지 인종정보(AFR;African, AMR;American, ASN;Asian, EUR;European)를 사용한 인종별평가(Ethnicity-specific validation)이다. 497개의 약물에 대하여 PharmSafe 알고리즘을 계산한 결과, 인종별 평가(Ethnicity-specific validation)에서는 0.662±0.081 (평균±표준편차,0.637~0.742), 비인종별 평가(Ethnicity-non-specific validation)에서는 0.633±0.038(평균±표준편차, 0.622~0.642)의 AUC값을 얻었다. 인종별 평가, 비인종별 평가를 비교했을 때 인종별 평가에서 비인종별 평가에 비해 우월한 AUC

점수를 보여 PharmSafe의 효과를 확실하게 증명하였다. 하지만 앞선 연구에서 입력값으로 모두 암호영역(Coding region)의 변이만을 사용했다는 점과 생물학적 지식(Biological knowledge)은 단백질-단백질 상호작용 (PPI;Protein-protein interaction)만 적용했다는 점을 한계점으로 제시하였다.

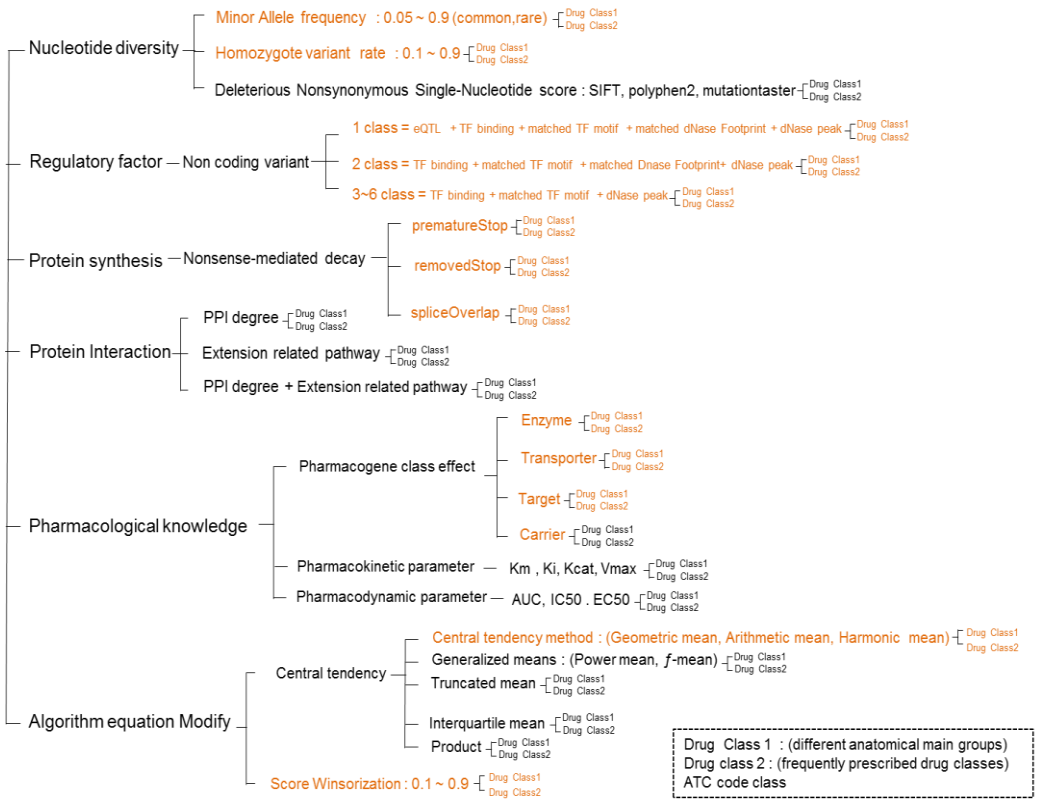


그림 0 생물학적, 약리학적, 통계학적 요소

따라서 본 연구에서는 앞선 연구의 한계점을 극복하고자 비암호영역(Non-coding region)의 변이 정보를 비롯한 8개의 생물학적(변이빈도;Variant frequency, 동형접합체 변이 비율;Homozygote variant rate, 대립유전자형빈도;Minor allele frequency, 넌센스-매개 전사체 붕괴;Nonsense-mediated decay), 약리학적(약물학적유전자; Pharmacogene class), 통계학적

지식(중심경향성;Central tendency method, 변이점수윈저화; Variant score winsorization)을 사용함은 물론, 해당 지식 최상의 조합을 찾아 알고리즘의 성능을 향상 시키고자 하였다(그림 0). 또한 1092명의 개인유전체 데이터를 2503명(1000 Genome Project Phase 3)으로 확장함으로써 더 많은 개인의 유전체 데이터에서의 PharmSafe 알고리즘의 재현성 또한 입증하고자 한다.

방 법

개인별 지놈 및 약물 데이터

입력값으로 사용한 개인별 지놈 데이터는 1000 Genome 프로젝트 [40] (2015년 6월 기준)로 부터 다운로드 받았다. 데이터는 JPT(Japanese in Tokyo, Japan) 104명, BEB(Bengali from Bangladesh) 86명을 포함한 전체 2504명으로 구성되어 있으며 작게는 26개의 부분 인종군(Subpopulation), 크게는 AFR(African), AMR(Admixed American), EAS(East Asian), SAS(South Asian), EUR(European)의 5개 인종군(Super population)으로 구분되어 있다(보충 표 1, 보충 그림 1) [20, 41].

KEGG drug [38] 와 Drug bank 4.0 [37] 로 부터 약물정보(drug information), 와 약물관련 유전자인 표적(Target), 수송체(transporter), 효소(enzyme), 수송기구(carrier)에 대한 정보를 수집하였다(자세한 추출 과정은 Pharmsafe1의 *"Drugs, drug-related genes, and drug-gene association"* 참조). 약물 분류군에 대한 정보는 ATC(Anatomical Therapeutic Chemical Classification System)와 자주 처방되는 약물분류(15 most frequently prescribed drug classes)를 사용하였다. ATC는 WHOCC [42] 로부터 다운로드 받았으며 이 중 14 해부학적 주요 그룹(Anatomical main groups)에 대한 정보를 추출하여 사용하였다. 자주 처방되는 약물 분류는 National Center for Health Statistics [43] 로 부터 다운로드 받아 사용하였다(보충 표 2 와 3).

생물학적 지식 정보 데이터

알고리즘 향상을 위해 비암호영역의 변이(Noncoding variant)등을 포함한 7가지 생물학, 약리학, 통계학적 지식정보를 각 지식베이스로부터 다운로드 받거나 혹은 데이터로부터 추출하여 사용하였다. 입력값으로 받은 유전체 서열에 포함된 변이로 인한 아미노산 치환이 단백질 기능에 영향을 미치는 정도를 점수로 나타내기 위해 SIFT [44] 를 사용하였다 [36]. 변이빈도(Variant frequency), 동형접합체변이 비율(Homozygote variant rate), 대립유전자형빈도(Minor allele frequency)는 1000 Genome 데이터로부터 추출하였다. 수집한 약물 데이터로부터 497개의 약물, 4226개의 약물-유전자 연관정보(drug-gene relations)를 추출하였으며 약물 관련 유전자에 대해서는 표적(Target) 440개, 수송체(transporter) 54개, 효소(enzyme) 74개, 수송기구(carrier) 10개로 총 545개를 추출하였다. 조기 번역정지(Prematurestop), 정지코돈삭제(removed stop), 이어맞추기위치변이(splice overlap)를 포함하는 난센스-매개 전사체 붕괴(NMD; Nonsense-Mediated mRNA Decay)를 유발하는 변이 정보를 Variant Annotation Tool [45] 을 사용하여 얻은 후 SIFT 점수가 있는 개인별 1000 Genome 데이터에 매핑하여 사용하였다 [46]. 디엔에이가수분해효소 과민반응 위치(DNase hypersensitivity site), 전사인자의 결합부위(binding sites of transcription factors), 촉진제 위치(promoter regions)등을 포함하는 전사 조절기전(regulation transcription)에 영향을 미치면서 유전자간부위(intergenic region)에 속하는 비암호영역의 변이(Noncoding variant)정보를 사용하기 위하여 전사조절 영향력에 대해 7단계로 분류, 제공하는 Regulomedb[47]로부터 19,493개의 유전자와 관련된 26,561,892개의 변이 그리고 99,845,325개의 유전자와 변이의 연관정보(gene-variant relation)를 다운받았다. 다운받은 변이 정보를 SIFT 점수가 있는 개인별 1000 Genome

데이터에 매핑하여 사용하였다. [48].

생물학적 지식을 사용한 변이, 유전자, 약물점수

개인의 유전체 서열에 나타난 변이들로부터 각 변이가 속한 유전자 그리고 해당 유전자가 영향을 미치는 약물의 유해한 정도를 정량화 하기 위해 우리는 기존 PharmSafe 논문을 통해 변이점수(variant score), 유전자점수(gene score), 약물점수(drug score)라고 명명한 세 단계의 점수를 고안해내고 해당 점수를 사용하여 개인의 유해한 약물순위를 예측하는 PharmSafe 알고리즘을 개발하고 그 유용성을 증명하였다 [49]. 본 논문에서는 앞서 발표한 PharmSafe 알고리즘의 성능을 향상시키고자 비암호영역(Non-coding region)의 변이 정보를 비롯한 8개의 생물학적(변이빈도;Variant frequency, 동형접합체 변이 비율;Homozygote variant rate, 대립유전자형빈도;Minor allele frequency, 넌센스-매개 전사체 붕괴;Nonsense-mediated decay), 약리학적(약물학적유전자;Pharmacogene class), 통계학적 지식(중심경향성;Central tendency method, 변이점수윈저화; Variant score winsorization)요소들을 사용하였으며 해당 지식 정보들을 입력값으로 사용하여 최상의 조합을 찾아 알고리즘의 성능을 향상 시키고자 하였다. 앞으로 열거될 모든 수식에서 S_{vi} , S_{gj} , S_{dk} 는 각각 변이 점수, 유전자 점수, 약물 점수를 의미하며 모든 수식에 사용된 기호는 보충 표 4을 참조하면 된다.

변이, 유전자, 약물점수

변이, 유전자, 약물점수는 앞서 발표한 Baik et al의 Pharmsafe

알고리즘을 사용하였다[50]. Pharmsafe의 알고리즘은 다음과 같다. 첫번째로 변이점수(Variant score)는 입력값으로 받은 개인의 유전체 서열에서 나타난 변이 정보를 담은 파일인 VCF(Variant Call Format)에 포함된 변이로 인한 아미노산 치환이 단백질 기능에 영향을 미치는 정도를 점수로 나타내기 위해 SIFT를 사용하였다[36]. 계산에 사용된 변이는 비동의성 암호영역 변이(non-synonymous coding variant)로써 변이 i 의 변이점수를 S_{Vi} 로 정의하고 SIFT 알고리즘의 점수를 사용하였다. 수식은 아래와 같다.

$$S_{Vi} = \mathrm{SIFT}(V_i)$$

변이점수는 0 부터 1 사이의 범위를 가지며 점수가 낮을수록 해당 변이로 인한 아미노산 치환이 단백질기능에 유해한(deleteriousness)영향을 미치는 것으로 해석된다. 유전자점수는 유전자 j 에 속하는 비동의성 암호영역 변이들의 변이점수를 기하평균을 이용하여 합산하며 이를 S_{g_j} 로 정의하였다. G_j 는 유전자 j 에 속하면서 변이점수를 가지는 동의성 암호영역 변이들의 집합을 의미한다. 수식은 아래와 같다.

$$s_{g_j} = \begin{cases} 1 & if\ |G_j| = 0 \\ \left(\prod_{v_i \in G_j} s_{v_i} \right)^{1/|G_j|} & if\ |G_j| > 0 \end{cases}$$

약물점수는 약물 k 의 약물반응성에 영향을 미친다고 알려진 약동학(PK;Pharmacokinetics), 약력학(PD;Pharmacodynamics)적 유전자들의 유전자점수를 기하평균을 이용하여 합산한 점수이며 이를 S_{dk} 라고 정의하고 수식은 아래와 같다.

$$s_{d_k} = \left(\prod_{g_j \in D_k} s_{g_j} \right)^{1/|D_k|}$$

D_k 는 약물 k 와 약동학, 약력학적 연관성이 있으면서 유전자 점수를 가지는 유전자들의 집합을 의미한다. 유전자점수와 약물점수 또한 0 부터 1 사이의 범위를 가지며 유전자점수는 낮을수록 해당 유전자가 약동학, 약력학적 기능을 수행함에 있어 유해한 영향을 미쳐 이로 인하여 약물반응성이 영향을 받는다고 해석되고 약물점수 역시 낮을수록 개인이 해당 약물을 복용했을 때 부작용를 비롯한 유해한 약물반응성을 보일 것이라고 해석된다. 이러한 약물점수는 복용할 약물 선택시 중요한 기준으로 사용될 것이다.

중심 경향 방법(central tendency method)

모집단으로부터 얻어진 자료를 살펴보면 특정값으로 몰리는 현상을 보이는데 이를 중심경향(central tendency)이라고 하고 해당 특정값을 중심경향값(central tendency value)이라고 한다. 이러한 중심경향을 나타내는 값은 평균(Mean), 중앙값(Median), 최빈값(Mode)등이 대표적이다. 특히 평균이 가장 많이 쓰이는데 평균값을 구하는 방법은 산술평균(Arithmetic mean), 기하평균(Geometric mean), 조화평균(Harmonic mean)등 7가지가 있다[51]. 기존의 PharmSafe 알고리즘에서는 유전자점수와 약물점수 합산시 기하평균(Geometric mean)을 사용하였으나 본 연구에서는 변이점수는 앞선 방법과 동일하게 계산하고 유전자점수와 약물점수는 중심 경향 방법에 대표적인 산술평균(Arithmetic mean), 조화평균(Harmonic mean) 그리고 곱(Product)를 사용하여 계산하고 그 결과를 기하평균(Geometric

mean)과 비교하였으며 각각의 방법에 따른 수식은 아래와 같다.

산술평균의 유전자점수 및 약물점수 수식

$$s_{g_j} = \begin{cases} 1 & if\ |G_j| = 0 \\ \left(\frac{1}{n} \sum_{v_i \in G_j} s_{v_i} \right) & if\ |G_j| > 0 \end{cases} \quad , \quad s_{d_k} = \left(\frac{1}{n} \sum_{g_j \in D_k} s_{g_j} \right)$$

조화평균의 유전자점수 및 약물점수 수식

$$s_{g_j} = \begin{cases} 1 & if\ |G_j| = 0 \\ \left(\frac{n}{\frac{1}{n} \sum_{v_i \in G_j} s_{v_i}} \right) & if\ |G_j| > 0 \end{cases} \quad , \quad s_{d_k} = \left(\frac{n}{\frac{1}{n} \sum_{g_j \in D_k} \frac{1}{s_{g_j}}} \right)$$

곱의 유전자점수 및 약물점수 수식

$$s_{g_j} = \begin{cases} 1 & if\ |G_j| = 0 \\ \left(\prod_{v_i \in G_j} s_{v_i} \right) & if\ |G_j| > 0 \end{cases} \quad , \quad s_{d_k} = \left(\prod_{g_j \in D_k} s_{g_j} \right)$$

약물학적 유전자 종류(Pharmacogene type)

약물의 반응성은 약물의 흡수, 분포, 생체내 변화 및 배설을 포함하는 약동학(PK;pharmacokinetics)적 작용기전과 생체에 대한 약물의 생리학적, 생화학적 작용기전을 나타내는 약력학(PD;Pharmacodynamics)적 작용기전에 따라 달라진다. 약물의 표적(Target)이 되는 유전자가 약동학적 작용기전을 조절하며 약력학적

작용기전은 약물수송체(transporter), 약물분해효소(enzyme), 약물수송기구(carrier) 유전자에 의해 영향을 받는다. 이러한 약물학적 유전자 종류에 따라 약물의 반응성이 달라 질 수 있음을 이용하여 기존 PharmSafe 알고리즘에 각각의 약물학적 유전자 종류에 속하는 변이에 가중치를 달리하여 적용하였다. 표적, 수송체, 효소, 수송기구 별로 특정 변이 i 가 유전자 j 의 영역에 포함되면 해당 변이점수를 제곱하여 가중치점수(Weight score)를 계산하였으며 이를 가중치 변이점수(Weighted variant score)로 명명하고 W_{v_i} 로 정의하였다. 수식은 아래 명시하였다. 약물학적 유전자 종류별 계산에서 만약 변이 i 가 해당 약물학적 유전자 종류의 유전자 영역에 포함되지 않거나 어떠한 약물학적 유전자의 영역에도 포함되지 않았다면 가중치를 가하지 않은 기존의 변이점수를 그대로 사용하였다. V_{PGT} 는 약물학적 유전자의 영역 안에 속하는 변이점수를 가진 변이들의 집합이다. 변이점수 및 가중치 변이점수의 수식은 아래와 같다.

$$S_{v_i} = SIFT\left(v_i\right)$$

$$W_{v_i} = \begin{cases} \left(S_{v_i}\right)^2 & if\ S_{v_i} \in V_{PGT} \\ S_{v_i} & else \end{cases},$$

$$V_{PGT} = \left\{ v_i \mid v_i \in Target / Transporter / enzyme / carrier\ gene\ region \right\}$$

유전자점수와 약물점수는 가중치 변이점수를 입력값으로 받아 위와 같이 기하평균을 사용하여 계산하였으며 수식은 아래와 같다.

$$s_{g_j} = \begin{cases} 1 & if\ \left|G_j\right| = 0 \\ \left(\prod_{v_i \in G_j} w_{v_i} \right)^{1/\left|G_j\right|} & if\ \left|G_j\right| > 0 \end{cases}, \quad s_{d_k} = \left(\prod_{g_j \in D_k} s_{g_j} \right)^{1/\left|D_k\right|}$$

변이 점수 윈저화(Variant score winsorization)

윈저화(Winsorization)는 이상치(Outlier)에 의해 왜곡되는 영향을 줄이기 위해 최극단에 속하는 값들을 특정 값으로 변환(Transformation)하여 사용하는 통계적인 기법이다[52]. PharmSafe 알고리즘에서는 0 부터 1 까지의 범위를 갖는 SIFT 점수를 변이점수로 사용하고 있으며, SIFT에서는 0.05이하의 점수를 가지는 변이에 대해 유해하다고(deleterious) 정의 하고 있다. 하지만 실제로 0.05 이상의 점수를 가지는 변이들에 대해서는 유해한 정도에 대한 언급을 하지 않고 있다[44]. 예를들어, SIFT점수 0은 매우 유해하고 1은 전혀 유해하지 않다고 정의 되어있기 때문에 0값을 가지는 변이와 1값을 가지는 변이 사이의 유해한 정도의 차이는 크다고 할 수 있지만 0.7값을 가지는 변이와 0.8값을 가지는 변이 사이의 유해한 정도의 차이는 가늠하기가 상당히 어렵다. 따라서 점수 구간 사이의 유해한 정도 차이가 거의 없어 예측 결과에 잡음(noise)을 유발하는 점수 구간의 절단점(cut-off point)을 찾아 윈저화(Winsorization) 방법을 통해 예측 결과의 잡음을 제거하고자 하였다. 이를 위해 0 부터 1 까지의 점수를 0.1, 0.2, … , 0.9 와 같이 0.1 간격으로 10개의 윈도우로 나누고 이를 SR 이라고 명명하였다. 각 윈도우별로 절단점(cut-off point) 이상의 값들은 변이점수 1로 변환하였다. 예를들어 0.7 윈도우에서는 0.7 이상의 변이점수들은 1로 변환하였다. 변이점수와 윈저화 변이점수 수식은 아래와 같다.

$$S_{v_i} = SIFT\left(v_i\right)$$

$$WS_{v_i} = \begin{cases} S_{v_i} & if\ S_{v_i} \le SR_l \\ 1 & else \end{cases}, \qquad SR = \{0.1 \sim 0.9\}$$

유전자점수는 윈저화 변이점수를 입력값으로 받아 위와 같이 기하평균을 사용하여 계산하였으며 약물점수는 앞서 계산된 유전자점수를 사용하여 계산하였다. 수식은 아래와 같다.

$$s_{g_j} = \begin{cases} 1 & if\ |G_j| = 0 \\ \left(\prod_{v_i \in G_j} WS_{v_i} \right)^{1/|G_j|} & if\ |G_j| > 0 \end{cases}, \quad s_{d_k} = \left(\prod_{g_j \in D_k} s_{g_j} \right)^{1/|D_k|}$$

낮은 대립형질 빈도(Minor Allele Frequency)

인구집단에서 낮은 대립형질 빈도(MAF)가 5%이하인 변이들에 대하여 Hapmap project를 통해 희귀변이(Rare variant)로 정의되었고, 그 후 1000 Genome project를 통해 한 개인당 평균 250-300개의 기능상실변이(loss-of-function variant)를 가지며 그 중 10~20개가 희귀변이라는 것이 밝혀졌다. [20]. 또한 이러한 희귀변이들이 알츠하이머, 파킨슨병을 비롯한 많은 질병을 유발하는 원인변이라는 것이 밝혀지고 이러한 연구들이 잃어버린 유전 가능성(missing heritability)의 많은 부분을 설명한다고 보고되었다[53]. 하지만 희귀변이가 약물의 반응성에 영향을 미친다는 연구는 이제 시작단계에 있으며 아직 희귀변이가 특정 약물의 반응성을 조절한다는 연구결과는 발표되지 않았다(2015년 8월 기준). 우리는 희귀변이가 약물 반응성에 미치는 영향을 반영하고자 기존의 Pharmsafe 알고리즘에 낮은 대립형질 빈도(MAF)정보를 반영하였다. 1000 Genome 데이터로부터 낮은 대립형질 빈도(MAF)정보를 추출하고 0부터 0.01까지의 빈도범위에 대하여 0.001의 간격을 적용하여 10개의 윈도우로 나누어 이를 ***MAFR***이라고 명명했다. 각 윈도우별로 해당 윈도우 이하의 낮은

대립형질 빈도값을 가지는 변이들에 대하여 기존의 변이점수를 제곱하여 가중치 변이점수를 계산하였으며 수식은 아래와 같다. V_{MAFR} 는 l 번째 $MAFR$에 해당하는 낮은 대립형질 빈도값을 가지는 변이들의 집합이다.

$$S_{v_i} = SIFT(v_i)$$

$$W_{v_i} = \begin{cases} (S_{v_i})^2 & if\ MAF_{v_i} \in V_{MAFR_m} \\ S_{v_i} & else \end{cases}, \quad \begin{aligned} MAFR &= \{0.001 \sim 0.009, 0.01 \le MAF_{v_i}\} \\ V_{MAFR} &= \{v_i \mid \mathrm{MAF}_{v_i} \in MAFR_m\} \end{aligned}$$

앞서 만든 가중치 유전자 점수를 사용해 유전자 점수를 계산하고 이를 이용해 약물 점수를 계산하였으며 수식은 아래와 같다.

$$s_{g_j} = \begin{cases} 1 & if\ |G_j| = 0 \\ \left(\prod_{v_i \in G_j} w_{v_i} \right)^{1/|G_j|} & if\ |G_j| > 0 \end{cases}, \quad s_{d_k} = \left(\prod_{g_j \in D_k} s_{g_j} \right)^{1/|D_k|}$$

동형접합변이 비율(Homozygote mutation rate)

대립유전자(allele)의 변이(mutation)에 따른 유전형(genotype)은 크게 동형접합야생형(homozygous wild type), 이형접합변이형(heterozygous mutant type), 동형접합변이형(homozygous mutant type)의 3가지 종류로 나뉠 수 있다. 동형접합야생형은 대립유전자 양쪽 모두 변이가 없는 경우를 의미하고 이형접합변이형은 한쪽에만 변이가 있는 경우를, 동형접합변이형은 양쪽 모두 변이가 있는 경우를 의미한다[54]. 일반적으로 희귀질환에서는 질병을 유발하는 열성돌연변이(recessive mutation)가 동형접합변이형인 경우 높은 질병발생률을 보인다고

알려져 있다. 약물유전체 연구에서도 동형접합변이로 인해 약물반응성이 달라지는 연구결과들이 보고되어 있으며, 대표적으로는 Ser9 유전자의 이형접합변이형 및 동형접합변이형이 동형접합야생형에 비하여 도파민 D3 수용체의 선택적 리간드인 GR99841의 결합능력을 증가시킨다는 것이 그 예이다[55]. 또한 클로피도그렐을 복용한 사람중 CYP2C19*17 동형접합변이형을 가진 사람이 그렇지 않은 사람에 비하여 혈소판 응집이 과도하게 일어났음이 보고되었다[56]. 이러한 연구결과를 바탕으로 우리는 전체 인구에서 동형접합변이의 비율이 높은 변이일수록 약물반응성에 미치는 영향이 클 것이라고 가정하고 1000 Genome 데이터로부터 동형접합변이의 비율을 추출하고 이를 Pharmsafe 알고리즘에 반영하여 계산하였다. 0.1부터 0.9까지의 동형접합변이 비율을 0.1의 간격을 적용하여 9개의 윈도우로 나누고 이를 ***HVFR*** 이라고 명명했다. 각 윈도우별로 해당 윈도우에 속하는 동형접합변이 비율을 가지는 변이들에 대하여 기존의 변이점수를 제곱하여 가중치 변이점수를 계산하였으며 수식은 아래와 같다. HV_{HVFR} 는 l 번째 ***HVFR*** 에 해당하는 동형접합변이비율을 가지는 변이들의 집합이다.

$$S_{v_i} = SIFT(v_i)$$

$$w_{v_i} = \begin{cases} (S_{v_i})^2 & if\ HVF_{v_i} \in HV_{HVFR_m} \\ S_{v_i} & else \end{cases} \qquad \begin{aligned} HVFR &= \{0.1 \sim 0.9\} \\ HV_{HVFR} &= \{v_i \mid \mathrm{HVF}_{v_i} \in HVFR_m\} \end{aligned}$$

앞서 만든 가중치 유전자 점수를 사용해 유전자 점수를 계산하고 이를 이용해 약물 점수를 계산하였으며 수식은 아래와 같다.

$$s_{g_j} = \begin{cases} 1 & if\ |G_j| = 0 \\ \left(\prod_{v_i \in G_j} w_{v_i} \right)^{1/|G_j|} & if\ |G_j| > 0 \end{cases}, \quad s_{d_k} = \left(\prod_{g_j \in D_k} s_{g_j} \right)^{1/|D_k|}$$

mRNA 안정성 조절 기전 (Nonsense-mediated mRNA decay; NMD)

난센스-매개 전사체 붕괴(NMD;Nonsense-mediated mRNA decay)는 유전자 발현 조절 기전 중 mRNA의 질적 조절(quality control)에 관여하는 가장 대표적인 기작이다[57]. 난센스-매개 전사체 붕괴는 유전자 내에 조기종결코돈(premature stop codons)이 포함되어 mRNA가 정상길이의 단백질보다 짧은 단백질을 생성하게 되고 유전자가 본래의 기능을 상실하고 해롭게(deleterious) 또는 이롭게(gain-of-function) 변형되거나 비정상 유전자에서 생산된 물질이 정상 유전자에서 생산된 물질과 결합하여 정상 유전자의 기능마저 비정상으로 만드는 우성-음성(Dominant negative)효과를 나타내는 기작이다[58]. 최근에는 조기 번역정지(Prematurestop), 정지코돈삭제(removed stop), 이어맞추기위치변이(splice overlap) 등의 기작 또한 난센스-매개 전사체 붕괴의 원인이 됨이 밝혀졌다[59]. 본래 난센스-매개 전사체 붕괴는 세포의 정상적인 기능을 위해 비정상적인 유전자를 제거하는 기전이지만 특정 유전적인 변이에 의해 비정상적으로 발생한 난센스-매개 전사체 붕괴로 인해 암 또는 유전질환등의 다양한 질병이 발생한다. 대표적인 예로 베타글로빈(β-globin) 유전자의 상위영역(upstream)에 존재하는 변이들로 인해 헤모글로빈의 합성이 감소되어 발생하는 혈액 질환인 베타 탈라세미아가

있다[60]. 우리는 변이에 의해 발생되는 난센스-매개 전사체 붕괴의 생물학적 영향을 PharmSafe 알고리즘에 반영하기 위해 Variant Annotation Tool(VAT)을 사용하여 난센스-매개 전사체 붕괴에 관여한다고 알려진 변이들의 정보를 추출했다[46]. 특정 변이 i 가 조기번역정지(Premature stop), 정지코돈삭제(removed stop), 이어맞추기위치변이(splice overlap) 중 하나라도 관여하는 변이라면 해당 변이의 변이점수를 제곱하고 그렇지 않으면 기존의 변이점수를 그대로 사용하여 가중치 변이점수(W_{v_i})를 계산하였다. V_{NMD} 는 난센스-매개 전사체 붕괴에 관여하며 변이점수를 가진 변이들의 집합이다. 변이점수 및 가중치 변이점수 수식은 아래와 같다.

$$S_{v_i} = SIFT\left(v_i\right)$$

$$w_{v_i} = \begin{cases} \left(S_{v_i}\right)^2 & if\ v_i \in V_{NMD_m} \\ S_{v_i} & else \end{cases}$$

$$NMD = \left\{ \text{Pr } emature\ Stop, \text{Remove Stop, Spicingover, PrematureStrop\& Removestop, PrematureStop\& RemoveStop\& Splicingover} \right\}$$

$$V_{NMD} = \left\{ v_i \mid \text{v}_i \in premature / removestop / splicingover\ \text{var}\, iant \right\}$$

앞서 만든 가중치 유전자 점수를 사용해 유전자 점수를 계산하고 이를 이용해 약물 점수를 계산하였으며 수식은 아래와 같다.

$$s_{g_j} = \begin{cases} 1 & if\ \left|G_j\right| = 0 \\ \left(\prod_{v_i \in G_j} w_{v_i} \right)^{1/|G_j|} & if\ \left|G_j\right| > 0 \end{cases} \quad , \quad s_{d_k} = \left(\prod_{g_j \in D_k} s_{g_j} \right)^{1/|D_k|}$$

비암호 변이를 포함한 유전자 기능 조절 변이 (Regulatory variants including noncoding region variants)

기존 PharmSafe 논문에서 가장 큰 한계점은 변이점수로 사용한 SIFT가 암호영역의 변이(coding variant)만 포함하고 있어 유전자간 영역(IGR;Intergenic region), 비발현부위(intron)등을 포함한 비암호영역의 변이(non-coding region)들을 반영 할 수 없었다는 것이었다. 이러한 한계점을 극복하고자 이번 논문에서는 비번역 변이들의 정보를 PharmSafe 알고리즘에 반영했다. 비암호영역, 암호영역변이들 중 유전자의 기능을 조절하는 변이들의 정보를 RegulomeDB[48]에서 다운로드 하였다. RegulomeDB는 특정 변이가 디엔에이가수분해효소 과민반응 위치(DNase hypersensitivity site), 전사인자의 결합부위(binding sites of transcription factors), 촉진제 위치(promoter regions)등을 포함하는 전사 조절기전(regulation transcription)에 영향을 미치는 전사조절 영향력에 대해 1a부터 6까지의 14 단계로 나타냈다. 1a 로 갈수록 변이 i 가 여러가지의 조절기작에 관여하는 것으로 정의하였다. 다운로드 받은 데이터는 각 변이의 염색체상의 위치정보, 해당 변이에 의해 전사조절이 되는 유전자의 목록, 그리고 앞서 언급한 RegulomeDB의 전사조절영향력 단계정보로 구성되어 있었다. 우리는 여러 유전자의 전사조절에 영향을 미치는 변이가 유해(deleterious)하다고 가정하고 RegulomeDB로부터 얻은 1부터 6까지 총 6단계별로 해당 정보를 활용하여 PharmSafe 알고리즘을 다음과 같이 변형하여 계산하고 검증하였다. 유전자 기능 조절변이 RV_i에 의해 조절기작이 영향을 받는 유전자 G_j를 RG_j 라고 정의하고 RG_j 의 조절 기작에 관여하는 변이의 갯수를 세어

가중치점수인 WS_{RG_j} 를 계산했다. 기존의 유전자점수 대신 WS_{RG_j} 를 가중치로 하여 가중기하평균(weighted geometric mean)을 통해 가중 유전자점수 WS_{g_j} 를 계산하고 이를 이용하여 약물점수를 계산하였으며 수식은 아래와 같다.

$$RC = \{class1 \sim class6, sum\,of\ class\} \quad , \quad RV_i = \{v_i \mid v_i \in RC_m \,\&\, v_i \in RG_j\}$$

$$S_{v_i} = SIFT(v_i) \quad , \quad WS_{RG_j} = \#RV_i$$

$$ws_{g_j} = \begin{cases} 1 & if\,|G_j| = 0 \\ \left(\prod_{v_i \in RG_j} v_i^{WS_{G_j}}\right)^{1 / \sum_{n=1}^{n} WS_{G_j}} & if\,|G_j| > 0 \end{cases} \quad , \quad s_{d_k} = \left(\prod_{g_j \in D_k} s_{g_j}\right)^{1/|D_k|}$$

각 요소별 최적의 조건 선정 및 조합 실험(Combination test)

앞서 6개의 생물학적(동형접합체 변이 비율;Homozygote variant rate, 대립유전자형빈도;Minor allele frequency, 넌센스-매개 전사체 붕괴;Nonsense-mediated decay), 약리학적(약물학적유전자;Pharmacogene class), 통계학적 지식(중심경향성;Central tendency method, 변이점수윈저화; Variant score winsorization) 요소들을 가중치로 사용하여 계산한 결과를 바탕으로 각 요소별 가장 높은 평가점수(AUC)를 가지는 7가지 조건을 선별하였다. 선별방법은 기하평균의 인종평가 AUC값(0.6076, SEA),

비인종평가 AUC값(0.6271, SNA)을 기준으로 두고 각 요소 E 에 속한 l 번째 조건 E_{C_l} 의 인종평가 AUC값($EA_{E_{C_l}}$), 비인종평가 AUC값($NA_{E_{C_l}}$)을 아래 수식과 같이 인종, 비인종평가 별로 각각의 기준 AUC값에서 요소 E 에 속한 l 번째 조건 E_{C_l} 의 AUC값을 뺀 $EAD_{E_{C_l}}$, $NAD_{E_{C_l}}$ 를 계산한 후 두 값의 평균을 취해 $MAD_{E_{C_l}}$ 을 계산하였다.

$$EAD_{E_{C_l}} = SEA - EA_{E_{c_i}},\ NAD_{E_{C_l}} = SNA - NA_{E_{c_i}}$$

중심경향값을 제외한 6가지 요소별로 가장 높은 $MAD_{E_{C_l}}$ 를 가지는 조건 한가지를 선택하여 총 6가지 요소별 조건을 선정하였다(보충표 7). 선택된 요소별 조건은 변이점수 윈저화(SW)는 0.2, 낮은 대립형질 빈도(MAF)는 0.007, 약물학적 유전자 종류(PGT)에서는 효소, 동형접합변이 비율(HR)은 0.7, mRNA 안정성 조절 기전(NMD)에서는 조기 종결코돈, 유전자 기능 조절변이(RV)에서는 4단계를 선택하였고 이렇게 선정된 6가지 요소별 조건들의 모든 58가지 조합에 대하여 PharmSafe 알고리즘을 적용하여 계산하였다. 선택된 요소별 조건들의 집합을 SE_C 라 정의하고 m번째 조합에 속하는 변이 i 를 $CB_{m_{v_i}}$, 이 변이들의 집합을 CB_{m_v} 로 정의했다. 변이 i 가 $CB_{m_{v_i}}$ 인 경우 변이점수에 제곱을 하고 그렇지 않으면 기존의 변이점수를 그대로 사용하여 가중치 변이점수(W_{v_i})를 계산하였다. 각 변이 별로 하나 이상의 $CB_{m_{v_i}}$ 에 속하는 경우 변이 i 의 가중치 변이점수(W_{v_i})를 곱하여 조합 가중치점수(CBW_{v_i})를 계산하여 유전자점수의 입력값으로 사용하였으며 수식은 아래와 같다.

$$S_{v_i} = SIFT(v_i)$$

$$W_{v_i} = \begin{cases} (S_{v_i})^2 & if\ v_i \in CB_{m_{v_i}} \\ S_{v_i} & else \end{cases},$$

$$SE_C = \{SW(0.2), MAF(0.007), PGT(enzyme), HR(0.7), NMD(PS\ \&\ \mathrm{RV}), RV(4class)\}$$
$$CB_{m_v} = \{v_i \mid v_i \in {}_{SE_C} C_r\}$$
$$r = \{1,2,3,4,5,6\}$$

$$CBW_{v_i} = \tilde{\bigcirc}_{v_i \hat{1} C_{v_i}} W_{v_i}$$

앞서 만든 조합 가중치 유전자 점수를 사용해 유전자 점수를 계산하고 이를 이용해 약물 점수를 계산하였으며 수식은 아래와 같다.

$$s_{g_j} = \begin{cases} 1 & if\ |G_j| = 0 \\ \left(\prod_{v_i \in G_j} CBW_{v_i}\right)^{1/|G_j|} & if\ |G_j| > 0 \end{cases}, \quad s_{d_k} = \left(\prod_{g_j \in D_k} s_{g_j}\right)^{1/|D_k|}$$

PharmGKB를 이용한 Pharmsafe 알고리즘의 성능 평가

앞서 발표한 PharmSafe 논문[50]과 같은 방법으로 본 논문에서도 PharmGKB를 사용하여 생물학적 지식 정보를 가중치로 사용한 개선된 PharmSafe 알고리즘의 성능평가를 시행하였다. PharmGKB는 전장유전체분석연구(GWAS) 결과 중 약물학적 지식과 관련된 원인변이와 약물 관계정보를 휴먼 큐레이션(Human curation)을 통해 검증하여 데이터베이스화 한 지식베이스이다. 2015년 1월 23일 PharmGKB [39]로부터 약물 392개, 변이 1176개를 포함한 3248개의

약물반응성을 조절하는 원인변이와 약물 관계정보를 다운로드 받았다. 이 중 앞서 선정한 497개의 약물에 해당하는 약물 290개, 변이 840개를 포함하는 1807개(55.63%)의 원인변이와 약물 관계정보를 추출하였다. 추출된 1807개의 변이가 속하는 유전자는 471개였으며 이는 표적(Target) 225, 수송체(transporter) 18, 효소(enzyme) 20, 수송기구(carrier) 7, 비약물학적 유전자 201개로 구성되어져 있었다(보충그림 6). 인종정보를 이용한 검증을 위해 PharmGKB로부터 흑인 혹은 아프리카계 미국인 관련 관계정보 83개, 아시아인 관련 관계정보 329개, 백인 관련 관계정보 647개, 기타 인종관련 6개, 인종정보가 없는 관계정보 1997개의 원인변이와 약물 관계정보를 받아 이 중 앞서 선정한 497개의 약물에 해당하는 흑인 혹은 아프리카계 미국인 관련 관계정보 58개(69.87%), 아시아인 관련 관계정보 329개(63.88%), 백인 관련 관계정보 451개(69.70%), 기타 인종관련 관계정보 5개(83.33%), 인종정보가 없는 관계정보 1128개(56.48%)의 약물, 원인변이 그리고 인종의 관계정보를 추출하였으며 PharmGKB에 대해서는 아프리카(Black or African American), 아시아(Asian), 미국 그리고 유럽(White)으로 인종을 재분류하고 1000 Genome 데이터에 대해서는 아프리카(AFR;African), 유럽(EUR;European), 아시아(EAS;East Asian, SAS;South Asian), 미국(AMR;Ad Mixed American)으로 분류하여 사용했다(보충표5). 위의 정보를 1000 Genome 데이터에 매핑한 결과, 한 개인당 평균 647.5±73.14(평균±표준편차,474~891)개의 약물과 원인변이 관계를 가지고 있었다(보충 그림4).

PharmSafe 알고리즘의 결과인 약물점수를 바탕으로 각 개인 별 위험약물 순위를 매겨 PharmGKB와 대조하였을 때 위험순위가 높은 약물들이 실제 환자군을 대상으로 한 실험에서 밝혀진 PharmGKB의 위험 약물 목록과 일치되고 위험하지 않은 약물들은 일치되지 않는다면

PharmSafe 알고리즘이 개인의 변이정보를 사용하여 해당 개인에게 위험한 약물과 그렇지 않은 약물을 구별하는 능력이 높다는 것을 증명할 수 있다는 가정을 바탕으로 아래와 같이 두 가지 데이터를 구성하여 민감도(sensitivity)와 특이도(specificity)를 계산하고 검증하였다. 사용된 첫번째 데이터는 497개의 약물정보를 사용하여 계산한 개인별 PharmSafe 약물점수를 낮은 점수순으로 순위(rank)를 매겨 구성한 개인별 Pharmsafe 위험 약물 목록이다. 두번째로는 앞서 설명한 바와 같이 PharmGKB로부터 추출한 497개의 약물과 관련된 원인변이와 약물 관계정보를 1000 Genome 데이터에 각 개인별로 매핑하여 PharmGKB와 일치된 원인변이와 약물 그리고 인종 관계정보를 구성하고 이를 Gold Standard(GS) 데이터로 사용하였다. 각 개인별 위험약물 순위목록을 임계값(threshold)을 낮춰가며 Gold Standard 데이터와 대조하여 Gold Standard의 약물이 임계값 범위 안의 개인별 위험약물 목록과 일치하면 참값(True)으로, 불일치하면 거짓값(False)으로 구분하고 이를 이용하여 민감도(sensitivity)와 특이도(specificity)를 아래의 수식으로 계산하였다.

$$sensitivity = \frac{|D_L \cap GS|}{|GS|}\ ,\ specificity = 1 - \frac{|D_L - GS|}{|D - GS|}$$

D 는 497개의 약물들의 집합을 의미하고, D_L 은 임계값 L 범위안의 약물들의 집합을 의미한다. 우리는 앞서 추출한 인종정보를 반영(Ethnic specific)하거나 혹은 반영하지 않는(Non-ethnic specific) 방법으로 알고리즘 평가점수인 AUC를 계산하여 성능을 검증하였다. 인종정보를 사용한 검증의 경우는 앞선 방법과 같이 임계값 범위 안의 개인의 위험약물 목록과 인종정보를 Gold Standard의 약물과 인종정보와 비교하여 약물과 인종 모두 일치하면 참값(True)으로, 불일치하면

거짓값(False)으로 판별하였으며 해당 과정을 각 4가지 인종별로 진행하였다.

결 과

생물학적, 통계학적 , 약물학적 7가지 요소별 변이, 유전자, 약물 평균 점수 분포

중심경항방법, 약물학적 유전자 종류, 변이점수 원저화, 낮은 대립형질 빈도, 동형접합변이 빈도, mRNA 안정성 조절 기전, 유전자 기능 조절변이 7가지로 구성되고 각각 4~10가지 조건이 포함된 가중치 요소들을 사용하여 2504명의 1000 지놈 개인 유전체 데이터에 매핑하고 각 요소별 가중치 알고리즘에 적용하여 계산한 변이점수, 유전자 점수, 약물 점수를 2504명 각 사람별로 평균점수를 계산하여 각 요소에 포함되어 있는 조건별로 평균 개인 점수 분포를 계산하였다 (그림 1). 중심경항방법을 비롯한 7가지 가중치 요소에 포함된 4~10 가지 조건을 적용하여 계산한 각 요소의 조건별 변이점수의 개인평균은 0.591~0.853(0.652±0.048, 평균±표준편차), 50509 개의 유전자 점수의 개인평균은 0.172~0.521(0.328±0.062, 평균±표준편차), 497 개의 약물 점수의 개인평균은 0.244~0.825(0.652±0.069, 평균±표준편차) 였다. 동형접합변이 빈도가 낮을수록 유전자 점수가 낮은 것을 확인할 수 있었다. 이는 동형접합변이이면서 빈도가 낮은 희귀 동형접합 변이들 중 non-synonymous 변이가 다수 포함되어 있음을 의미하며 이는 희귀 동형접합변이가 아미노산 서열 변형을 잘 일으키고 이 작용으로 인하여 유전자의 기능 상실되어 약물 반응성에 영향을 미친다고 해석할 수 있다. 중심경향법에서 곱이 가장 낮은 유전자 점수와 약물 점수를 보였지만 곱의 특성상 유전자의 길이가 길어 변이 수가 많거나 영향을 미치는 유전자가 많이 연구되어 밝혀진 약물의 경우는 유전자나 약물의 위험도와 상관없이 유전자, 약물 점수가 낮아지기 때문에 표준화(normalization)이 되는 기하평균을 다음

계산에서는 표준 계산법으로 이용하였다.

7가지 요소별 조건을 이용한 인종, 비인종 별 Pharmsafe 알고리즘 평가

7가지 요소는 각각 중심경향방법 4(산술평균, 기하평균,조화평균, 곱), 약물학적 유전자 종류 4(표적,수송체,효소,수송기구), 변이점수 윈저화 9(0.1~0.9), 낮은 대립형질 빈도 10(0~0.1≦MAF), 동형접합변이 빈도 9(0.1~0.9), mRNA 안정성 조절 기전 5(조기종결코돈, 종결 코돈 제거, splice-overlap, 조기종결코돈 + 종결코돈제거,조기종결코동 + 종결코돈제거 + splice-overlap), 유전자 기능 조절변이 7(1~total class)가지 조건으로 구성되어 있다. 각 요소에 포함된 54가지 조건을 가중치로 사용하여 변이(n=977842), 유전자(n=50509), 약물(n=497) 점수를 1000 Genome 에 포함된 2504명 개인별로 계산하고 PhrmsGKB로부터 다운받은 유전자-변이-약물 연관과 이에 해당하는 인종정보를 사용하여 만든 GS(Gold-standard)와 함께 각 인종, 비인종 각각 조건별로 AUC를 계산하여 Pharmsafe 알고리즘의 성능을 평가 비교 하였다. 계산결과 AUC는 인종평가 0.5633 ~ 0.6436,0.6093±0.011(최솟값~최댓값,평균±표준편차), 비 인종 평가 0.5857 ~ 0.6502,0.6229±0.011(최솟값~최댓값,평균±표준편차)로 나타나 인종 평가에 비해 비 인종평가가 높은 AUC를 나타냈다(그림 2, 보충 표 6). 이렇게 비 인종 평가가 낮에 나타난 이유는 평가 기준으로 사용된 PharmGKB의 1699 개의 유전자-변이-약물 연관 중 인종 정보가 포함된 유전자-변이-약물 연관은 45.96%(781)개로 절반도 되지 않은 것이 원인이라고 예상된다. 이는 PharmGKB가 GWAS 연구의 결과를 논문들로부터 추출하여 만든 지식정보 베이스 이기 때문에 적은 수의 유전자-변이-약물 연관 정보 및 인종정보를 포함 할

수 밖에 없다. 그 이유는 GWAS연구 특성상 특정 약물 복용여부를 바탕으로 모집한 비교 대조군(case-control) 개인 유전체 서열 데이터을 대량으로 모집해야 하는데 이는 현실적으로 불가능해 많은 수의 유전자-변이-약물 연관을 밝혀 낼 수 없기 때문이다. 따라서 Pharmsafe 알고리즘의 개인별 위험한 약물을 판단하는 변별력이 떨어진다기 보다는 검증에 사용된 적은 수의 유전자-변이-약물 연관 정보가 문제가 된다. 앞으로 더 많은 인종정보가 포함된 유전자-변이-약물 연관 데이터가 생성되면 인종평가의 AUC는 물론 전체 AUC 값 또한 상승될 것이라고 예상된다. 가장 높은 AUC를 보인 조건은 인종 평가에서는 약물학적 유전자 종류의 효소(0.6436) 이었으며 비인종 평가에서는 중심경향법의 곱(0.6502) 이었으며 가장 낮은 AUC를 보인 조건은 인종 평가에서는 중심경향법의 산술평균(0.5633), 비인종 평가에서는 유전자 기능 조절변이 total class(0.5857)이었다(그림 2, 보충 표 6).

기하평균, 산술평균, 조화평균, 곱 4가지를 포함한 중심경향법의 수식 각각을 적용해 2054명의 1000 Genome 개인유전체 데이터를 사용해 Pharmsafe 알고리즘을 변형하여 계산한 AUC 결과는 인종 0.5633 ~ 0.6163, 0.5964±0.033(최솟값~최댓값,평균±표준편차), 비인종0.5935~0.6502, 0.6269±0.021(최솟값~최댓값,평균±표준편차)값을 나타냈다. 기하평균과 조화평균, 곱이 산술평균에 비해 에서 높은 AUC 값을 보였으나 각 조건의 인종간의 표준편차가 각각 인종평가 0.017, 0.048, 0.021으로 기하평균에서 가장 작은 표준편차를 보여 인종에 상관없이 높은 AUC를 나타냈다. 특히 아프리카 인종에서의 AUC는 조화평균, 곱에서 현저히 낮아졌다. 이는 조화평균이나 곱은 유전자의 길이가 길어 변이의 수가 많거나 특정 약물에 관여하는 유전자의 수가 많으면 점수가 내려가는 특성을 반영한 것으로 표준화가 되지 않아 안정된 AUC값의 표준편차가

큰 것 으로 해석할 수 있다. 이러한 이유로 다음에 하는 모든 실험에서 기준 수식으로 기하평균을 사용하였다. (그림 2, 보충 그림 S4 A).

약물학적 유전자 종류 4가지(표적, 수송체, 효소, 수송기구)에 속하는 유전자에 포함되는 변이 점수를 제곱하는 방법으로 가중 변이점수로 만들어 Pharmsafe알고리즘으로 약물 점수를 계산하고 평가한 AUC는 인종 0.5928 ~ 0.6248, 0.6055± 0.022(최솟값~최댓값,평균±표준편차) 비 인종 0.5928 ~ 0.6248, 0.6270± 0.018(최솟값~최댓값,평균±표준편차) 이었다. 인종, 비인종 모두 효소에서 가장 높은 AUC를 나타냈고 표적은 가장 낮은 AUC를 나타냈다. 이는 표적 유전자는 대부분 하나의 약물에만 영향을 미치지만 효소의 경우는 한 개의 유전자가 다수의 약물에 영향을 미친다. 따라서 같은 개수의 유전자에 변이가 생겨 기능을 상실한다 해도 표적은 한 개 혹은 두개의 약물의 영향력에 영향력을 미미치는 것 비해 효소는 다수의 약물에 영향을 미치기 때문이라고 예상된다(그림 2, 보충 그림 4 B).

잡음을 줄이거나 제거하기 위한 방법중 하나로 통계학에서 쓰이는 윈저화는 잡음을 유발하는 절단점을 찾는 것을 목표로 한다. Pharmsafe 계산시 잡음을 제거해 개인에게 위험한 약물을 구분하는 변별력을 높이기 위해 변이 점수 윈저화를 시행하였다. 0 부터 1 사이의 점수를 0.1, 0.2 … 0.9로 각각 10개의 윈도우로 나누고 각 윈도우 별로 절단점값(cut off point value) 이상의 값들은 변이 점수 1로 변환하였다. 예를 들어 0.7 윈도에서는 0.7 이상의 변이 점수들은 1로 변환하는 변이 점수 윈저화를 실시하고 Pharmsafe 알고리즘으로 약물 점수를 계산하였다. 그 결과 AUC는 인종 0.6074 ~ 0.6167, 0.6046± 0.016(최솟값~최댓값,평균±표준편차), 비인종 0.6219 ~ 0.6364, 0.6278± 0.002(최솟값~최댓값,평균±표준편차) 이었다. 인종에서는 0.3 단계가 가장 높은 AUC 0.6167 를 나타냈고 비 인종 단계에서는 0.2 단계가 가장 높은 AUC 0.6364 를 나타냈다(그림 2, 보충 그림 4

C).

낮은 대립형질 빈도는 1000 Genome에서 나타난 변이들의 낮은 대립형질 빈도 데이터를 이용하여 0.001 ~ 0.001 이상 의 10단계로 낮은 대립형질 빈도를 나누고 단계별로 m 번째 단계에 포함되는 변이 점수를 제곱하여 가중치 변이 점수를 만들어 Pharmsafe 알고리즘을 사용하여 약물 점수를 계산하였다. 그 결과 AUC는 인종 0.6008 ~ 0.6078, 0.6018± 0.016(최솟값~최댓값,평균±표준편차) , 비인종 0.6121 ~ 0..6272, 0.6260± 0.003(최솟값~최댓값,평균±표준편차) 이었다. 인종평가 에서는 0.007이 AUC 0.6078로 가장 높았으며 비인종 평가 에서는 0.005로 0.6272로 가장 높았으나 0.01 이상 단계를 제외한 모든 단계에서 비슷한 AUC 를 나타냈다 (그림 2, 보충 그림 4 D). 낮은대립형질빈도가 5%(0.05) 미만인 희귀변이(rare variant)들이Alzheimer's disease 를 비롯한 여러 질병을 유발하는 원인 변이라는 것은 물론 항암제methotrexate 를 비롯한 약물들의 약물반응성을 조절하는 원인 변이라는것이 밝혀져 있다 [33, 34] 하지만 5% 라는 기준은 Hapmap 프로젝트에서 제시한 기준으로 일부 다른 논문에서는 1% 이하를 희귀변이라고 주장한다. 또한 희귀변이가 질병에 영향을 미친다는 연구는 다수 발표되어 있으나 약물반응성에 대하여 희귀변이가 영향을 미친다는 연구는 항암제 몇종에 대한 연구 뿐이다. 따라서 약물 반응성을 조절하는 희귀변이의 절단점값은 아직 모호하다. 하지만 이번 결과로 미루어 0.01 이하의 낮은 대립형질 빈도를 가지는 변이들이 약물작용에 미치는 영향이 0.01 이상의 낮은대립형질을 가지는 변이에 비해 크다는 것으로 해석할 수 있음으로 약물반응성에 영향을 미치는 희귀변이의 절단점값은 1% 이하 라고 해석할 수 있다.

1000 Genome 에 속한 2504명의 데이터를 기준으로 변이당 동형접합변이 빈도를 계산한 후 0.1 ~0.9의 9단계로 빈도를 구분하여

개인에게서 나타난 동형접합변이가 단계별로 m 번째 단계에 속하면 변이 점수를 제곱하여 가중치 변이 점수를 만들고 이를 이용해 Pharmsafe 알고리즘으로 계산하여 약물 점수를 계산하고 평가하였다. 그결과 인종평가 에서는 0.5995 ~ 0.6073, 0.6022± 0.015(최솟값~최댓값,평균±표준편차), 비인종평가 에서는 0.6186 ~ 0.6276, 0.6268± 0.002(최솟값~최댓값,평균±표준편차) 으로 비인종평가의 0.7 단계에서만 AUC 0.6276 기준 AUC 0.6271 보다 높게 나타났다(그림림 2, 보충 그림 4 E).

mRNA 안정성 조절 기전은 조기 종결코돈(premature stop codons), 종결 코돈 제거(removed stop codons), splice-overlap 으로 총 3가지 기작으로 알려져 있다. VAT를 사용하여 종결코톤종결 등 3가지 기작을 일으키는 변이 정보를 1000 Genome 2504명에서 한번이라도 나타난 변이 목록에 매핑하였다. 3가지 기작을 조기종결코돈, 종결 코돈 제거, splice-overlap, 조기종결코돈 + 종결코돈제거,조기종결코돈 + 종결코돈제거 + splice-overlap 로 구분하여 총 5단계로 나누고 단계마다 해당 단계에 개인별로 해당 단계의 변이 정보에 해당하는 변이를 변이 점수를 제곱하여 가중치 변이 점수로 계산하고 Pharmsafe 알고리즘에 적용하여 개인당 약물별 점수를 계산하고 이를 평가하였다. 그 결과 인종평가 0.6074 ~ 0.6163, 0.6096± 0.018(최솟값~최댓값,평균±표준편차), 비인종평가 0.6271 ~ 0.6301, 0.6286± 0.002(최솟값~최댓값,평균±표준편차)이었고 인종 ,비인종 평가 모두에서 조기종결코돈이 가장 높은 AUC를 나타냈다(그림 2, 보충 그림 4 F). 종결코돈제거는 단백질의 합성시 함성을 중지하는 코돈이 상실되어 생기는 기작으로 단백질의 기능을 하는 부분의 아미노산이 대부분 상실되지 않는다. 또한 splicing-overlap 의 경우에도 단백질 합성시 일정부분의 RNA 서열이 상실되는 경우로 단백질의 기능을 하는 부분의 아미노산이 합성 될 수 도 있고 그렇지

않을 수 도 있어 실제로 단백질의 기능에 미치는 영향력이 크기 않아 많은 연구가 이루어지지 않은 기작이다. 따라서 두가지 기작으로 인하여 생성된 단백질로 인하여 약물의 반응성에 미치게되는 영향력이 크지 않거나 없을 수도 있다. 하지만 조기종결코돈 같은 경우는 아예 단백질 합성시 종결코돈이 실제보다 앞서 나타나 단백질 합성 자체가 극히 일부분만 되는 경우로 단백질의 기능 자체가 상실되는 경우가 많다고 알려져 있어 mRNA 안정성 조절 기전의 가장 대표적인 기전으로 알려져 있다. 따라서 이번 실험결과가 이러한 생물학적 기전을 반영한 결과라고 할 수 있다.

RegulomeDB 는 각 변이가 eQTL, TF binding, matched TF motif, matched DNase Footprint, DNase peak 등의 유전자의 기능을 조절하는 기작 에 관여하는 정도를 1a ~ 6의 14 단계의 점수로 나타내 제공한다. 데이터는 26,561,892 변이와 해당 변이가 기능에 영향을 주는 19,493 개의 유전자로 구성되어 있으며 이중 1.2%(n = 301,551)가 엑손 영역에 있는 변이이고 98.8%(n =26,260,341)가 인트론 및 비번역 부분에 존재하는 변이로 구성되어 있다. 기존 Pharsmafe 논문에서는 SIFT 점수가 존재하는 액손 영역의 변이만 계산영역에 넣었다. 따라서 비번역 영역의 변이는 알고리즘에 적용할 수 없어 큰 한계점으로 지적되었다. 따라서 이번 실험에서 26,260,341개의 비번역 변이를 알고리즘 계산에 추가함을써 코딩영역 변이 뿐 아니라 비번역 영역의 변이도 추가함으로써 지놈 전체 영역에서 발생하는 변이를 모두 Pharmsafe 알고리즘에 적용하여 한계점을 극복하였다. 또한 유전자 점수 계산시 기존 알고리즘 에서는 유전자 j 영역에 속하는 변이들의 점수를 기하평균을 내어 사용했다. 하지만 이번 실험에서는 변이 i 가 기능을 조절하는 유전자 j 에 해당하는 변이들의 수를 세어 유전자 j 에 가중치를 주어 가중치 유전자 점수로 사용해 DNA 영역을 벗어나 전사 단계에서 일어나는 유전자에 대한 영향력 까지 고려하여

알고리즘을 개선하였다. RegulomeDB 데이터를 다운로드 받아 1 단계부터 전 단계를 합친 전체 단계(total)등 총 7단계로 나누고 1000 Genome 2504명 유전체 데이터에 개인별 변이 목록 중 단계별로 유전자 기능 조절 단계 m에 해당하면 해당 변이 점수를 제곱하여 가중치 변이 점수를 만들어 Pharmsafe 알고리즘을 계산하여 개인별로 약물점수를 계산하고 이를 평가하였다. 그 결과 인종평가 0.5788 ~ 0.6258, 0.6037± 0.024(최솟값~최댓값,평균±표준편차), 비인종평가 0.5857 ~ 0.6189, 0.6027± 0.015(최솟값~최댓값,평균±표준편차)의 AUC 결과가 도출되었다. 인종평가에서는 4 단계가 0.6258로 가장 높은 AUC를 나타냈으며, 비인종 평가는 기준 AUC 보다 모두 낮은 AUC를 보였다(그림 2, 보충 그림 4 G).

각 요소들을 가중치로 반영해 계산한 Pharmsafe 알고리즘들이 각 약물 분류 정보군에서의 작용을 알아보기 위하여 약물 분류정보 ATC 에서 추출한 14 가지 군 그리고 WHOCC 에서 추출한 15 가지 가장 자주 처방 받은 약물 군 총 29 가지 약물 군 별로 54 가지 조건을 가중치로 적용하여 AUC 를 계산, 비교하였다. 7 개의 모든군에서 인종 비인종 모두에서 B(혈액 및 혈액 형성 기관) 0.5289 ~ 0.7727±0.0469, C(심장 혈관 시스템) 0.5076 ~ 0.6994±0.0418, G(비뇨 생식계 그리고 성 호르몬) 0.3805 ~ 0.8731±0.0685, G03(성 호르몬과 생식 시스템의 조절기) 0.7396 ~ 0.8619±0.025, C03(이뇨제) 0.4071 ~ 0.7928±0.0876, N06A(항 우울제) 0.4942 ~ 0.7364±0.0556 (최저값 ~ 최고값±표준편차)가 공통적으로 높은 AUC 를 나타냈다. 특이적으로 mRNA 안정성 조절 기전 인종평가 에서 M(근골격계)가 0.7719~0.8322±0.0256 로 높게 나타났다(그림 3, 보충 그림 4). 주로 혈관계, 비뇨생식계 그리고 성호르몬 관련 약물군들이 높은 AUC 를 나타내는 것을 확인하였다.

요소별 조건 중 최적의 조합을 사용한 가중 Pharmsafe 알고리즘 성능 평가

앞서 6개의 생물학적(Homozygote variant rate, Minor allele frequency, Nonsense-mediated decay), 약리학적(Pharmacogene type), 통계학적(SIFT Score filtering) 요소들을 가중치로 사용하여 가중 pharmsafe 알고리즘을 평가한 AUC를 기준으로 6가지 요소별 조건(변이 점수 윈저화(SW) 0.2, 낮은 대립형질 빈도(MAF) 0.007, 약물학적 유전자 종류(PGT) 효소, 동형접합변이 빈도(HR) 0.7, mRNA 안정성 조절 기전(NMD) 조기 종결코돈, 유전자 기능 조절 변이(RV) 4단계)을 선정하였다(방법 각 요소별 최적의 조건 선정 및 조합 실험 참조). 선택된 요소별 조건의 모든 56가지 조합을 입력값으로 하여 가중 Pharmsafe 알고리즘을 사용하여 1000 Genome 2504명의 개인별 약물점수를 계산하고 이를 검증하였다. 인종평가에서는 평균적으로 전체 0.6068±0.222(평균±표준편차)였고, 비인종평가 에서는 평균적으로 전체 0.6095±0.02(평균±표준편차)였다. 인종평가 비인종평가 모두에서 가장 높은 AUC 를 보였던 조합은 변이 점수윈저화 & 약물학적 유전자 타입으로 전체 AUC 0.6235, 0.6426 이었다. 이는 기준 AUC보다 0.0159, 0.015 상승하였다(그림 4, 보충 그림 8). 각 조합의 AUC 중 상위 10% 안에 든 조합 6개의 조합을 살펴보면 인종평가 에서는 약물학적 유전자 타입이 6회, 변이 점수 윈저화 3회, 낮은대립형질빈도 와 mRNA안정기전이 각각 2회씩 포함되어 있었고 비인종평가 에서는 약물학적 유전자 타입이 6회, 변이 점수 윈저화 4회, 낮은대립형질빈도 와 mRNA안정기전이 각각 2회 그리고 동형접합변이빈도가 1회씩 포함되어 있었다. 이 결과로 미루어보아 알고리즘 계산시 절단점값(0.2)을 찾고 절단점값 이상에서 나타나는 잡음을 제거하는 점수 윈저화가 필수적이라는 것을 알 수 있다. 또한

약물 유전자의 여러가지 종류 중 한개의 유전자가 많은 약물에 영향을 미치는 효소가 약물반응성에 큰 영향력을 미치고 있음을 알 수 있다.

고찰

약물부작용(ADR)의 원인이 되는 변이를 밝히는 연구는 꾸준히 진행되어 왔으며 응고인자 활성을 억제하는 항응고제인 와파린(warfarin), 간질치료제로 쓰이는 카바마제핀(carbamazepine), 진통제로서 감기약의 성분이 되는 코데인(Codeine)등에 관련된 연구가 그 대표적인 예이다. 코데인의 경우, CYP2D6*1xN/*2xN/*17xN/*35xN 유전형을 가진 환자에서 CYP2D6 유전자가 초고속대사자(UMs;Ultrarapid Metabolizers) 표현형을 띄게 되고 코데인의 대사가 정상 유전형을 가진 경우보다 훨씬 빠르게 되어 독성을 일으킨다고 보고되었다.[61, 62] 또한 VKORC1과 CYP2C9 유전자에 −1639G>A , CYP2C9*2 , CYP2C9*3 유전형을 가진 환자의 경우 일반 환자에서보다 와파린을 37% 적게 복용해야하며 그렇지 않은 경우 과도한 항응고 반응으로 인한 약물부작용이 일어날 수 있다고 보고되었다.[63, 64].

앞선 예시와 같이 유전적 변이에 의해 약물부작용이 발생할 수 있다는 사실이 그동안 다수의 연구결과로 밝혀져 이러한 원인 변이를 찾는 것이 약물부작용 연구에 있어 매우 중요하다고 알려져 있음에도 불구하고 현재 알려진 약물부작용의 원인 변이는 상당히 적다. 이러한 현상의 원인은 현재 연구 방법의 한계점에서 기인한다. 현재까지 약물부작용의 원인 변이를 밝히는 연구는 대부분 하나의 약물을 대상으로 부작용을 일으키는 실험군과 대조군(case−control)을 표본추출하여 전장유전체분석연구(GWAS)기법을 통해 주로 5개 미만의 원인 변이를 밝히는 인구기반의 관찰연구(population−based observational studies)였다. 인구기반의 관찰연구는 연구대상이 되는 표본을 추출하기 위해 소요되는 시간이 길고 그에 따른 큰 비용이 요구되는데 반해 상대적으로 적은 수의 원인 변이를 결과로 얻게 되어 현실적으로 많은 수의 연구를 수행하기 어렵다는 단점을 가지고 있다. 또한 인구기반의

관찰연구 특성상 추출된 표본의 인구통계학적 요인/조건(Demographic condition)에 의해 영향을 크게 받게 되는데 대부분의 연구대상 인종이 백인(Caucasian)이기 때문에 아시아나 아프리카 인종에 대해 동일한 연구결과를 적용하는데 어려움이 있었다. 이러한 문제점들으로 인해 개인의 유전적 정보를 활용하여 개인별 맞춤 약물처방을 제공하는 맞춤의료(Personalized medicine)는 실현되기 어려웠다.

이러한 문제점을 극복하고자 우리는 2013년 개인의 유전체정보를 사용하여 각 개인별 위험한 약물 의 순위를 제공하는 PharmSafe 알고리즘을 개발하고 이를 약물과 해당약물부작용의 원인변이 정보를 지식베이스화한 PharmGKB를 이용하여 알고리즘을 평가해 알고리즘의 성능을 입증하였다 [49]. PharmSafe 알고리즘은 각 개인의 유전체 서열로 부터 변이정보를 추출하여 변이, 유전자, 약물점수를 계산하기 때문에 앞서 제시된 인구기반의 관찰연구의 비용적인 문제와 인구통계학적 요인에 의존적인 결과등의 한계점들을 극복하였다. 하지만 암호영역의 변이에 대해서만 적용이 가능하고 여러가지 생물학, 약리학, 통계학적 요소들을 충분하게 반영하지 못한다는 한계점을 가지고 있었다. 이러한 점들을 극복하고자 본 논문에서는 생물학, 약리학, 통계학적 지식요소 7가지(중심경향방법, 약물학적 유전자 종류, 변이점수 원저화, 낮은 대립형질 빈도, 동형접합변이 빈도, mRNA 안정성 조절 기전, 유전자 기능 조절변이)에 속하는 54가지 조건을 반영하여 PharmSafe 알고리즘을 계산하고 평가하였다. 54가지 조건을 반영하여 알고리즘을 계산, 평가한 결과 전체 인종을 대상으로한 평가(Ethnicity-specific validation)에서는 AUC 0.5633~0.6436, 0.6057±0.012 (최솟값~최댓값, 평균±표준편차)로 가장 높은 AUC값을 얻은 조건은 약물학적 유전자 종류 중 약물분해효소에 대한 가중치 계산이었다. 비인종 평가(Ethnicity-non-specific validation)에서는 AUC 0.5857~0.6502, 0.6224±0.222(최솟값~최댓값,평균±표준편차)으로

가장 높은 AUC값을 얻은 조건은 중심경향방법의 요인에 속하는 곱이었다(그림 2). 약동학(PK;Pharmacokinetics)에서는 약물의 생체 작용을 설명하는 흡수, 분포, 대사, 배설(ADME;Absortion, Distribution, Metabolism, Excretion) 중 대사작용이 약물의 농도 조절과 부작용의 많은 부분을 설명한다고 제시하고 있으며, 이러한 주장은 와파린을 비롯한 많은 약물들의 부작용의 원인변이가 CYP 유전자군에서 발견되는 것으로 증명되고 있다[64]. 알고리즘의 검정을 위해 사용한 PharmGKB의 1807개의 변이 약물 연관관계 데이터에서 효소가 전체 471개중 20개(4.24%) 인 결과를 반영하여도 효소가 약물반응성에 큰 영향을 미친다는 것을 알 수 있다. 이러한 연구결과들과 약동학적 이론으로 미루어 보아 약물반응성에 가장 큰 영향을 미치는 것이 약물대사효소 유전자라는 것을 알 수 있다. 인종대상 평과(Ethnicity-specific validaton)결과에서 약물대사효소 유전자에 대해 가중치를 사용했을때 가장 좋은 평가결과를 얻은 것 또한 같은 맥락으로 생각될 수 있다. 또한 하나의 약물대사효소 유전자에 변이가 발생하는 경우 약물표적 유전자의 경우와 달리 다수의 약물대사에 영향을 미치기 때문에 약물대사효소 유전자가 약물반응성에 중요한 역할을 한다고 할 수 있다.

이렇게 개선된 PharmSafe 알고리즘이 실제 약물학적 메커니즘을 반영하여 각 개인별 위험한 약물 순위를 도출하고 있으며 이 결과가 증명됨을 알 수 있다. 54가지 가중치 조건의 조합실험에서도 점수윈저화와 약물학적 유전자 타입 조합이 AUC값 0.6235, 0.6426으로 인종, 비인종 평가 모두에서 가장 좋은 결과를 보였다(그림 3). 이 결과 역시 약물학적 유전자 종류에 따라 약동학적인 작용이 영향을 받으며 특히 약물대사효소가 중요한 역할을 한다는 것을 알 수 있다. 전체적으로 약물학적 유전자 종류를 가중치 요소로 사용한 4가지 조건 실험을 제외하고 모두 비인종평가가 인종평가보다 높은 AUC를

나타냈다. 그 원인은 평가기준으로 사용한 PharmGKB데이터의 특성 때문으로 예상된다. 실제 PharmGKB로부터 다운로드 받은 데이터에는 3,248개의 약물과 원인변이 관계정보가 포함되어 있었으며 이 중 인종정보가 포함된 약물과 원인변이 관계정보는 781개(45.96%)로 절반도 되지 않았다. 앞으로 약물, 원인변이 그리고 인종의 관계정보가 많아지면 PharmSafe 알고리즘의 평가점수 또한 상승될 것으로 예상된다. 54가지 조건을 반영하여 계산한 모든 평가결과에서 아프리카 인종의 AUC 편차가 큰 이유는 PharmGKB에 속해 있는 인종정보가 포함된 약물과 원인변이 관계정보 781개 중 아프리카 인종에 관련된 관계정보는 112개(13.28%)로 백인에 관련된 관계정보가 413개(48.99%)인 것에 비해 월등하게 낮기 때문이다. 이 또한 앞서 언급한 바와 같이 약물, 원인변이 그리고 인종의 관계정보가 많아지면 전체적인 AUC 값은 상승하고 편차는 낮아질 것으로 생각된다. 약물 분류 정보군를 가중치로 이용한 실험에서는 B(혈액 및 혈액 형성기관) 0.5289~0.7727±0.0469, C(심장혈관 시스템) 0.5076~0.6994±0.0418, G(비뇨 생식계 그리고 성 호르몬) 0.3805~0.8731±0.0685, G03(성 호르몬과 생식 시스템의 조절기) 0.7396~0.8619±0.025, C03(이뇨제) 0.4071~0.7928±0.0876, N06A(항 우울제) 0.4942~0.7364±0.0556 (최솟값~최댓값±표준편차) 약물 군에 대한 결과가 공통적으로 AUC값이 높았다. 혈액 및 혈액 형성기관에 속하는 대표적인 약물은 클로피도그렐(clopidogrel), 와파린(warfarin)등이며 앞서 언급한 바와 같이 이들 약물은 부작용의 원인이 되는 변이에 대한 연구가 잘 알려져 있으며 이를 고려한 실제 임상사례들 또한 많다. 다리페나신(Darifenacin)은 과민성방광 치료에 사용되는 약물로CYP2D6 유전자에 의해 대사가 조절되는 것으로 알려있으며 이는 FDA drug lable에 기재되어 있다 [65]. 심장혈관 시스템 군에 속하는 약물 중 혈압강하제인 아테놀(atenolol)은

GALNT2 유전자에 NC_000001.10:g.230294916C>T(rs2144300) 변이가 존재하면 혈압과 고밀도지단백 콜레스테롤(HDL-C) 수치를 낮춘다고 보고되어 있으며 HDL 콜레스테롤 수치가 낮으면 일반적으로 심장질환에 위험하다고 알려져 있다[66].

본 논문에서는 앞선 PharmSafe 알고리즘의 가장 큰 한계점인 암호영역의 변이만 고려된다는 한계점을 극복하고 비암호영역의 변이정보를 가중치 정보로 사용했다. 하지만 비암호영역에 해당하는 전체 변이를 사용한 것이 아니라 그 중 26,260,341개의 비암호영역 변이만을 사용했기 때문에 여전히 한계점을 가지고 있다. 이전 논문에 이어 이번 논문에서도 알고리즘의 성능평가를 위해 PharmGKB를 사용했기 때문에 여전히 실제 특정 약물에 대한 부작용을 겪은 환자집단에서 평가하지 못한 한계점 또한 가지고 있다. 이러한 한계점에도 불구하고 본 논문을 통해 소개한 개선된 PharmSafe 알고리즘은 기존 알고리즘에 생물학, 약리학, 통계학적 지식요소를 기반으로 가중치를 반영, 기존 알고리즘 보다 높은 평가수치(AUC)를 보였다. 이 결과는 암호영역의 변이 존재유무만을 반영하여 위험한 약물 순위를 도출하는 것보다 생물학, 약리학, 통계학적 요소를 추가적으로 반영하는 것이 개인에게 위험한 약물을 예측하는데 훨씬 효과적임을 나타낸다. 현대 약물유전학의 지향점은 개인 유전체 서열을 기반으로 개인의 약물반응성을 예측하고 이를 약물 처방시 적용하여 개인 맞춤 약물 처방을 하는 것이다. 우리는 각 개인의 유전변이정보를 반영하여 위험한 약물 순위를 제공하는 PharmSafe 알고리즘이 이러한 점에서 현대 약물유전학의 지향점에 가장 부합하는 알고리즘이라고 생각한다. 앞으로 우리는 인구통계학적 요인/조건(Demographic condition)을 추가적으로 반영하여 각 인종적, 지역성, 환경적 특성을 고려할 수 있도록 PharmSafe 알고리즘을 개선하고자 한다. PharmSafe 알고리즘은 개인에게 위험한 약물 순위를 제공하기 때문에 의학적

의사결정 지원시스템(CDSS;Clinical Decision Support System)에서 사용되면 환자가 자신의 유전체 서열을 바탕으로 환자 자신에게 위험요소가 없는 약물을 알맞은 농도로 처방받는데 매우 유용하게 쓰일 수 있을 것이다. 의사 또한 해당 환자의 유전적 특성을 쉽게 파악하여 맞춤 처방을 할 수 있게 될 것이다. 이는 약물부작용으로 인한 심각한 피해를 감소시킬 뿐아니라 치료비용 또한 획기적으로 줄여줄 것이다. 또한 약물오남용에 의한 부작용이나 비용 역시 감소시킬 것으로 기대한다.

결과 그림

그림 1. 1000 genome 에 포함된 2504명의 개인별 지놈데이터의 각 가중치 요소별 변이, 유전자 그리고 약물 점수 분포

1000 Genome 2504명의 개인 유전체 데이터에 속한 변이, 50509개의 유전자, 497개의 약물정보를 이용한 각 요소별 변이, 유전자, 약물 점수의 개인별 점수 평균 분포(A) 중심경항방법 (B)약물학적 유전자 종류 (C) mRNA 안정성 조절 기전(PS : 조기 종결코돈(Premature Stop codons), RS : 종결 코돈 제거(Removed Stop codons), SO : Splice-Overlap) (D) 낮은 대립형질 빈도 (E) 변이점수 윈저화 (F) 동형접합변이 빈도 (G) 유전자 기능 조절변이(1~6,total : 1 class ~total class)

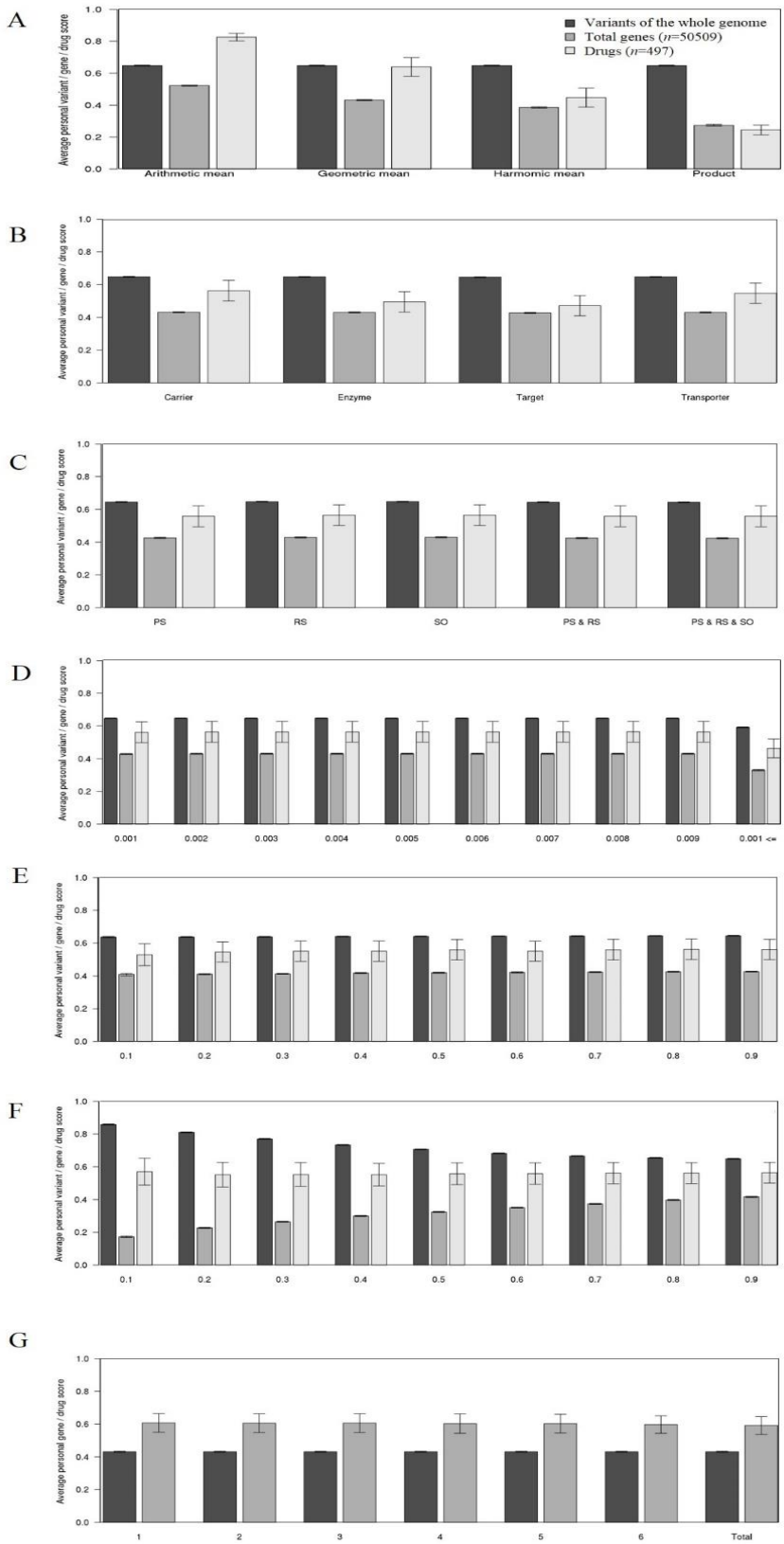
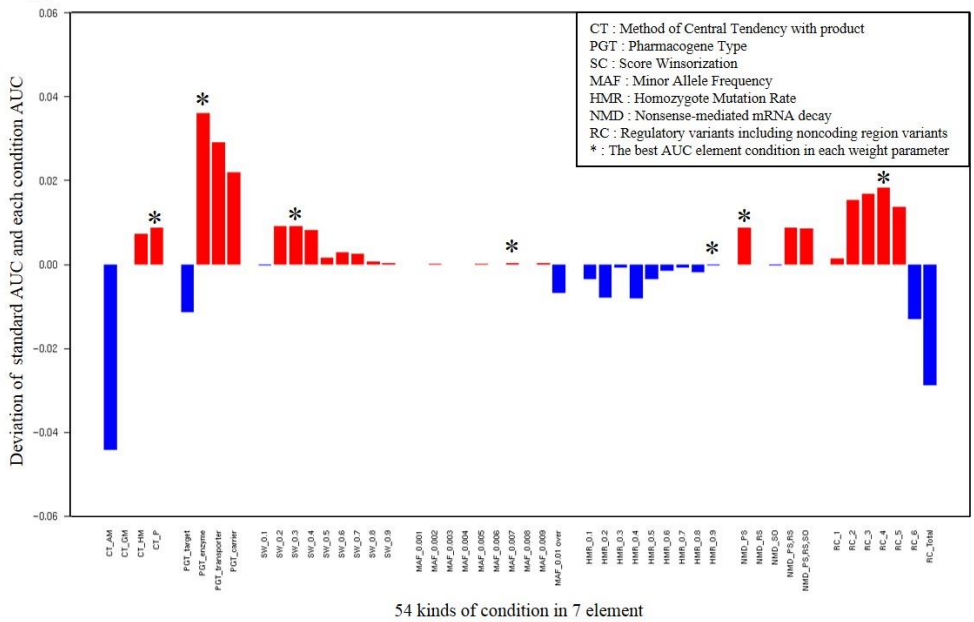

A
Average personal variant / gene / drug score
Variants of the whole genome
Total genes (n=50509)
Drugs (n=497)
Arithmetic mean
Geometric mean
Harmonic mean
Product
B
Carrier
Enzyme
Target
Transporter
C
PS
RS
SO
PS & RS
PS & RS & SO
D
0.001
0.002
0.003
0.004
0.005
0.006
0.007
0.008
0.009
0.001 <=
E
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
F
G
Average personal gene / drug score
1
2
3
4
5
6
Total

그림 2. 7가지 각 요소에 포함된 48가지 조건 별 인종, 비인종 Pharmsafe 알고리즘 검증 결과 비교

1000 Genome 에 속한 2504명의 개인 유전체 데이터를 각 7가지 요소에 포함된 48가지 조건을 가중치로 사용하여 계산한 497 약물별 점수를 PharmGKB로 다운받은 유전자-변이-약물 연관과 인종정보를 사용하여 만든 GS를 사용하여 AUC를 계산하여 검증했다. 54가지 조건별 (A)인종 검증 (B)비인종 검증. 각 막대그래프의 값은 기하평균의 전체 AUC(인종 : 0.6076, 비인종 : 0.6271)값을 기준 AUC 값으로 정하고 이 값에서 각 조건별 AUC값을 뺀 편차이며 적색은 기준 AUC보다 각 조건의 AUC가 상승했음을 의미하고 청색은 기준 AUC보다 각 조건은 AUC값이 하락했음을 의미한다. * 표시는 각 요소별 가장 높은 AUC값을 가진 조건을 의미한다.

A


CT : Method of Central Tendency with product
PGT : Pharmacogene Type
SC : Score Winsorization
MAF : Minor Allele Frequency
HMR : Homozygote Mutation Rate
NMD : Nonsense-mediated mRNA decay
RC : Regulatory variants including noncoding region variants
* : The best AUC element condition in each weight parameter
Deviation of standard AUC and each condition AUC
0.06
0.04
0.02
0.00
-0.02
-0.04
-0.06
CT_AM
CT_GM
CT_HM
CT_P
PGT_target
PGT_enzyme
PGT_transporter
PGT_carrier
SW_0.1
SW_0.2
SW_0.3
SW_0.4
SW_0.5
SW_0.6
SW_0.7
SW_0.8
SW_0.9
MAF_0.001
MAF_0.002
MAF_0.003
MAF_0.004
MAF_0.005
MAF_0.006
MAF_0.007
MAF_0.008
MAF_0.009
MAF_0.01 over
HMR_0.1
HMR_0.2
HMR_0.3
HMR_0.4
HMR_0.5
HMR_0.6
HMR_0.7
HMR_0.8
HMR_0.9
NMD_FS
NMD_RS
NMD_SO
NMD_FS,RS
NMD_FS,RS,SO
RC_1
RC_2
RC_3
RC_4
RC_5
RC_6
RC_Total
54 kinds of condition in 7 element

B

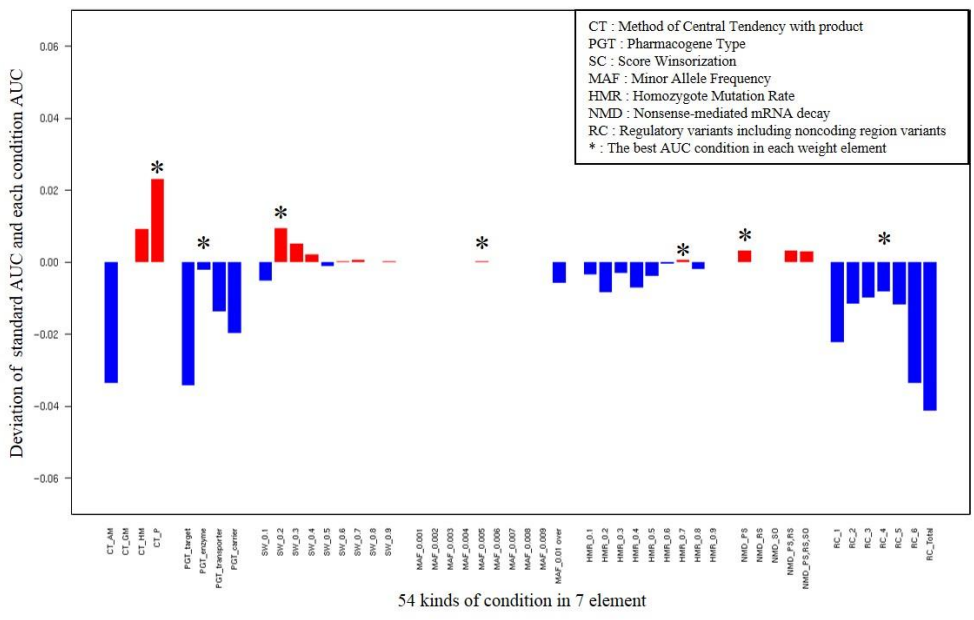
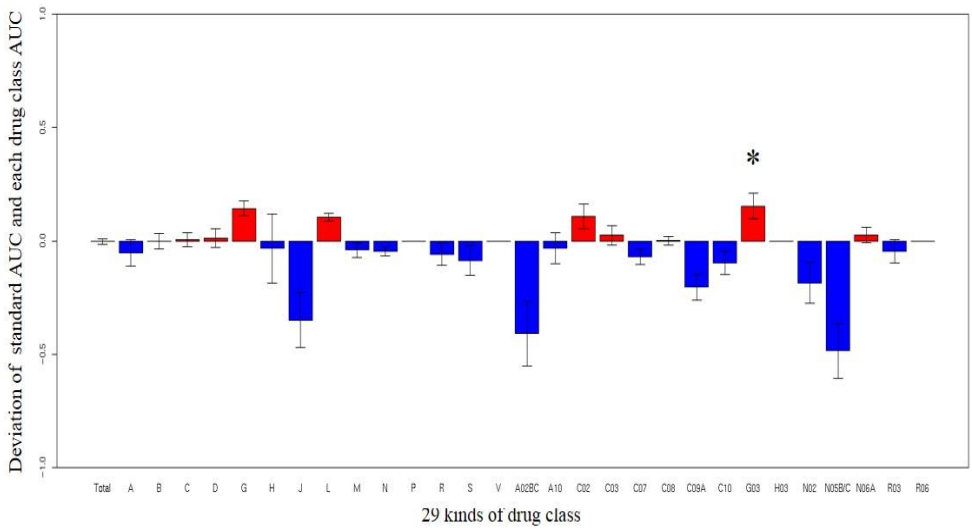

CT : Method of Central Tendency with product
PGT : Pharmacogene Type
SC : Score Winsorization
MAF : Minor Allele Frequency
HMR : Homozygote Mutation Rate
NMD : Nonsense-mediated mRNA decay
RC : Regulatory variants including noncoding region variants
* : The best AUC condition in each weight element
Deviation of standard AUC and each condition AUC
0.06
0.04
0.02
0.00
-0.02
-0.04
-0.06
CT_AM
CT_GM
CT_HM
CT_P
PGT_target
PGT_enzyme
PGT_transporter
PGT_carrier
SW_0.1
SW_0.2
SW_0.3
SW_0.4
SW_0.5
SW_0.6
SW_0.7
SW_0.8
SW_0.9
MAF_0.001
MAF_0.002
MAF_0.003
MAF_0.004
MAF_0.005
MAF_0.006
MAF_0.007
MAF_0.008
MAF_0.009
MAF_0.01 over
HMR_0.1
HMR_0.2
HMR_0.3
HMR_0.4
HMR_0.5
HMR_0.6
HMR_0.7
HMR_0.8
HMR_0.9
NMD_FS
NMD_RS
NMD_SO
NMD_FS,RS
NMD_FS,RS,SO
RC_1
RC_2
RC_3
RC_4
RC_5
RC_6
RC_Total
54 kinds of condition in 7 element

그림 3. 7가지 각 요소에 포함된 48가지 조건 별 29 약물분류군별 인종, 비인종 Pharmsafe 알고리즘 검증 결과 비교

1000 Genome 에 속한 2504명의 개인 유전체 데이터를 각 7가지 요소에 포함된 48가지 조건을 가중치로 사용하여 계산한 497 약물별 점수를 PharmGKB로 다운받은 유전자-변이-약물 연관과 인종정보를 사용하여 만든 GS를 사용하여 29가지 약물분류군별로 AUC를 계산하여 검증했다. 29가지 약물분류군별 (A)인종 검증 (B)비인종 검증. 각 막대그래프의 값은 기하평균의 전체 AUC(인종 : 0.6076, 비인종 : 0.6271)값을 기준 AUC 값으로 정하고 이 값에서 각 조건별 AUC값을 뺀 편차이며 적색은 기준 AUC보다 각 조건의 AUC가 상승했음을 의미하고 청색은 기준 AUC보다 각 조건은 AUC값이 하락했음을 의미한다. * 표시는 각 요소별 가장 높은 AUC값을 가진 조건을 의미한다.

A


Deviation of standard AUC and each drug class AUC
1.0
0.5
0.0
−0.5
−1.0
*
Total A B C D G H J L M N P R S V A02BC A10 C02 C03 C07 C08 C09A C10 G03 H03 N02 N05B/C N06A R03 R06
29 kinds of drug class

B

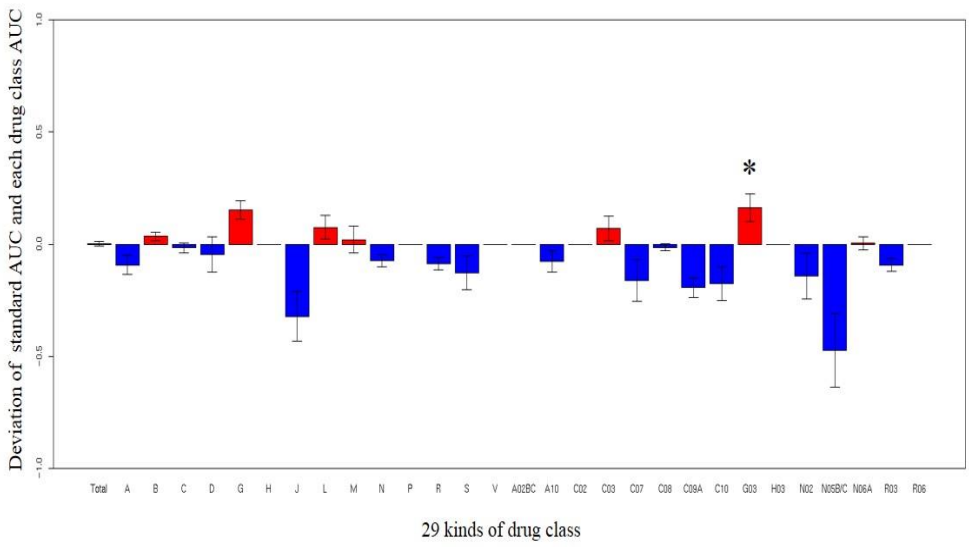
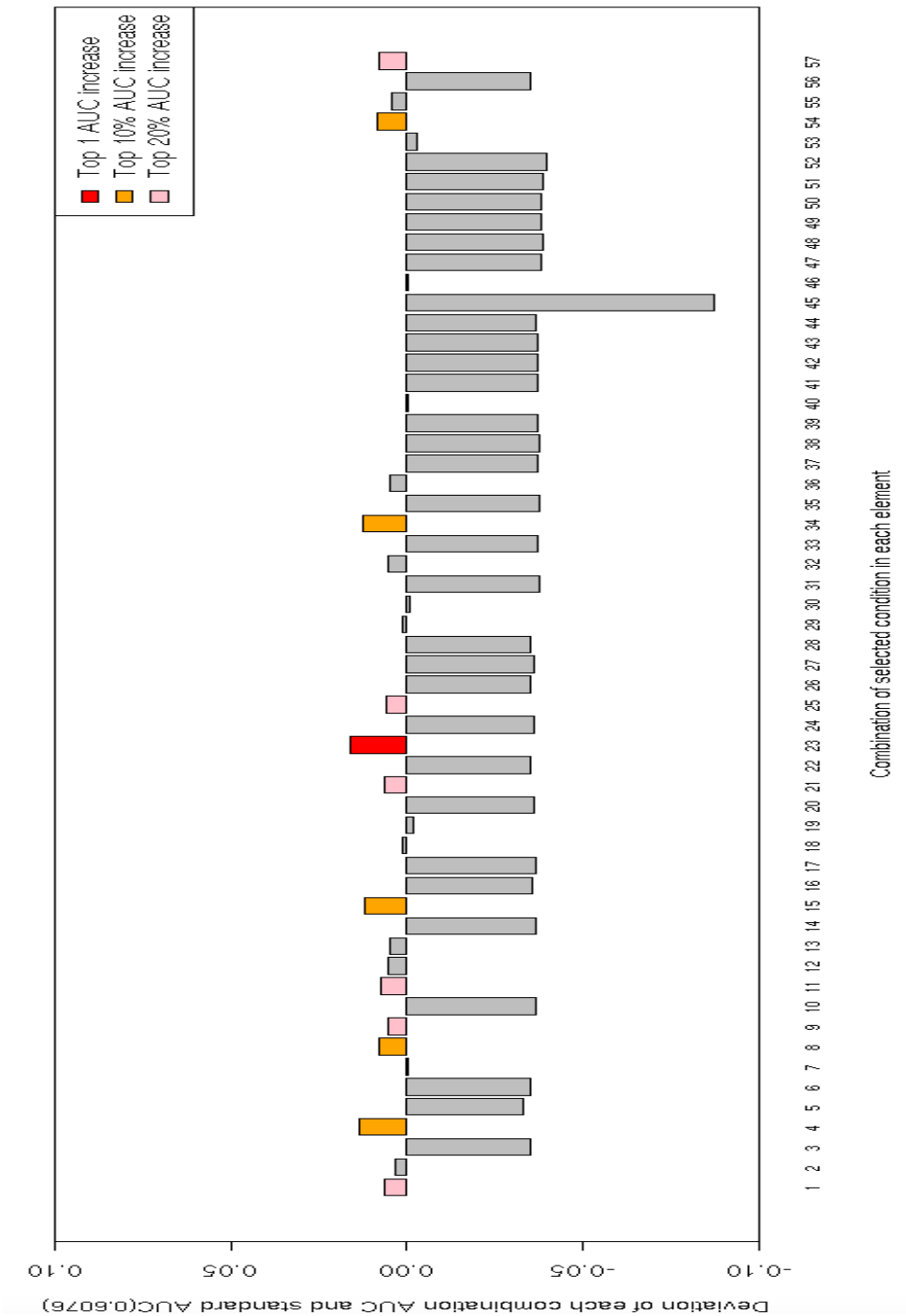

Deviation of standard AUC and each drug class AUC
1.0
0.5
0.0
−0.5
−1.0
*
Total A B C D G H J L M N P R S V A02BC A10 C02 C03 C07 C08 C09A C10 G03 H03 N02 N05B/C N06A R03 R06
29 kinds of drug class

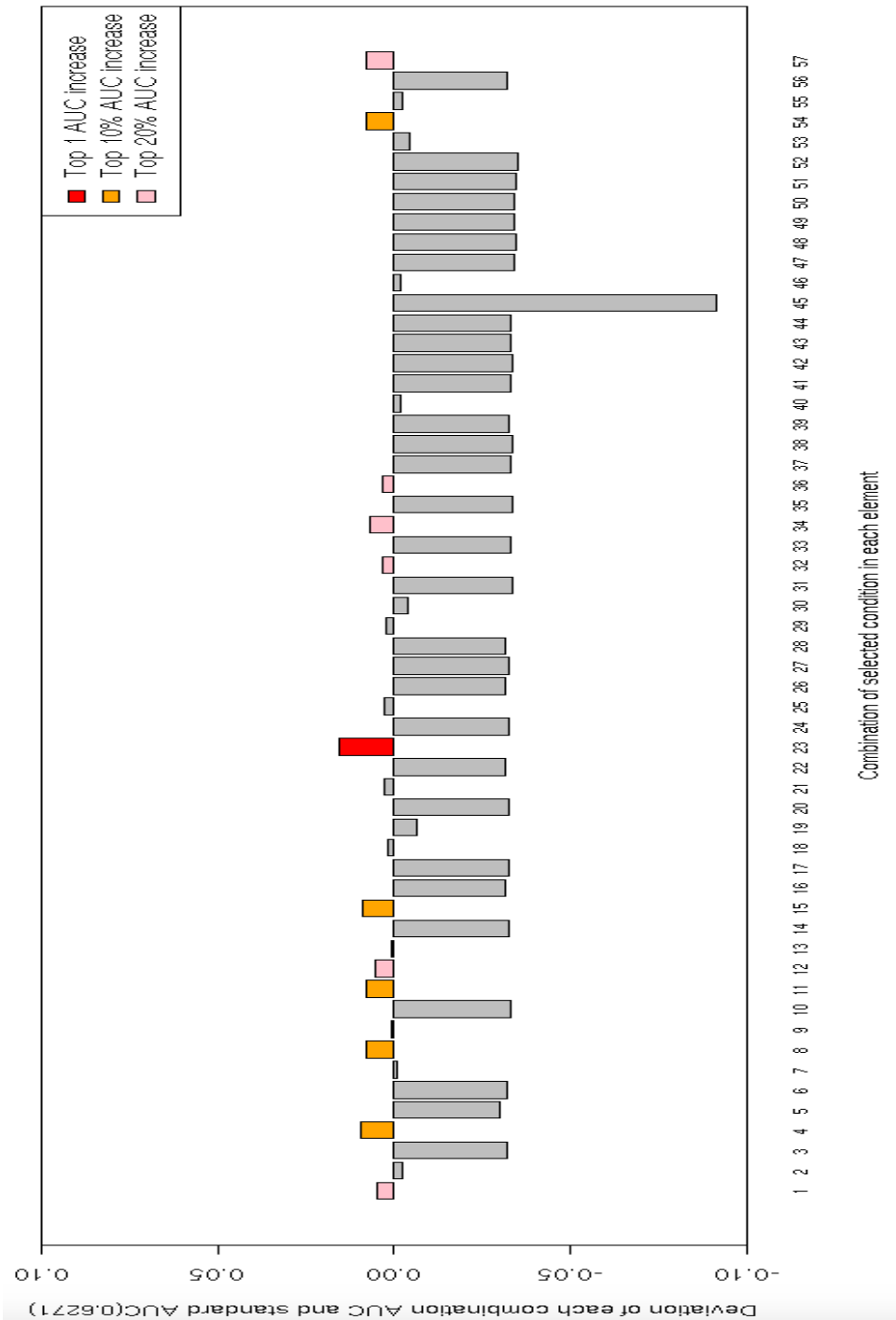
그림 4. 7가지 각 요소에 포함된 48가지의 54조합별 인종, 비인종 Pharmsafe 알고리즘 검증 결과 비교각 7가지 요소에 포함된 48가지 조건중 선택된 6가지 조건(변이 점수 윈저화(SW) 0.2, 낮은 대립형질 빈도(MAF) 0.007, 약물학적 유전자 종류(PGT) 효소, 동형접합변이 빈도(HR) 0.7, mRNA 안정성 조절 기전(NMD) 조기 종결코돈, 유전자 기능 조절 변이(RV))에 대한 모든 57가지 조합을 가중치로 사용하여 2504명 개인별로 Pharmsafe 알고리즘으로 계산하여 약물점수를 도출하고 이를 평가하여 각 조합별로 AUC를 계산하였다. (A) 인종평가, (B)비인종 평가이다. 그래프에서 적색은 가장 높은 AUC를 보인 조합이고 주황색은 상위 10%에 속한 AUC 가진 조합, 분홍색은 상위 20%에 속하는 AUC를 가진 조합이다. 세로축은 인종, 비인종별 기준 AUC(0.6076,0.6271)에서 각 조합의 AUC를 뺀 편차이며 가로축은 각 조합의 순번이다

.0(SW,MAF),1(SW,HR),2(SW,PGT),3(SW,NMD),4(SW,RV),5(MAF,HR),6(MAF,PGT),7(MAF,NMD),8(MAF,RV),9(HR,PGT),10(HR,NMD),11(HR,RV),12(PGT,NMD),13(PGT,RV),14(NMD,RV),15(SW,MAF,HR),16(SW,MAF,PGT),17(SW,MAF,NMD),18(SW,MAF,RV),19(SW,HR,PGT),20(SW,HR,NMD),21(SW,HR,RV),22(SW,PGT,NMD),23(SW,PGT,RV),24(SW,NMD,RV),25(MAF,HR,PGT),26(MAF,HR,NMD),27(MAF,HR,RV),28(MAF,PGT,NMD),29(MAF,PGT,RV),30(MAF,NMD,RV),31(HR,PGT,NMD),32(HR,PGT,RV),33(HR,NMD,RV),34(PGT,NMD,RV),35(SW,MAF,HR,PGT),36(SW,MAF,HR,NMD),37(SW,MAF,HR,RV),38(SW,MAF,PGT,NMD),39(SW,MAF,PGT,RV),40(SW,MAF,NMD,RV),41(SW,HR,PGT,NMD),42(SW,HR,PGT,RV),43(SW,HR,NMD,RV),44(SW,PGT,NMD,RV),45(MAF,HR,PGT,NMD),46(MAF,HR,PGT,RV),47(MAF,HR,NMD,RV),48(MAF,PGT,NMD,RV),49(HR,PGT,NMD,RV),50(SW,MAF,HR,PGT,NMD),51(SW,MAF,HR,PGT,RV),52(SW,MAF,HR,NMD,RV),53(SW,MAF,PGT,NMD,RV),54(SW,HR,PGT,NMD,RV),55(MAF,HR,PGT,NMD,RV),56(SW,MAF,HR,PGT,NMD,RV)

A. 인종평가


Top 1 AUC increase
Top 10% AUC increase
Top 20% AUC increase
Deviation of each combination AUC and standard AUC(0.6076)
0.10
0.05
0.00
-0.05
-0.10
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57
Combination of selected condition in each element

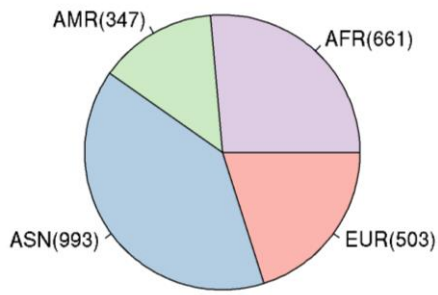
B. 비인종평가


Top 1 AUC increase
Top 10% AUC increase
Top 20% AUC increase
Deviation of each combination AUC and standard AUC(0.6271)
-0.10
-0.05
0.00
0.05
0.10
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57
Combination of selected condition in each element

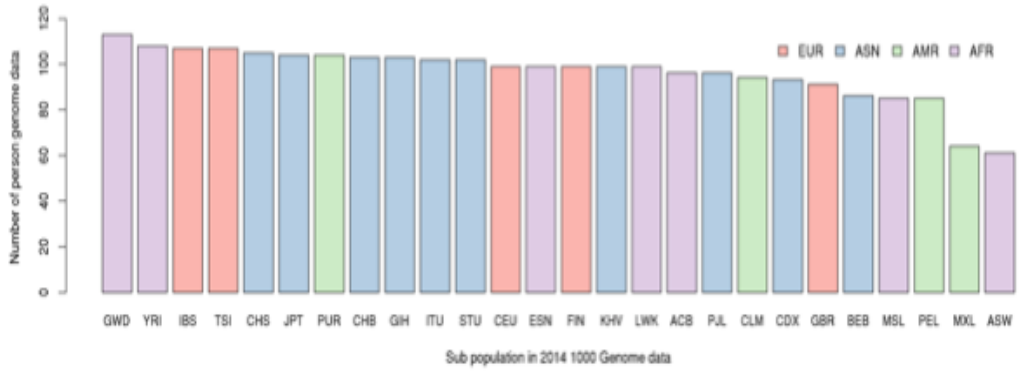
보충 그림

보충 그림 1. 2014년도 1000 Genome 데이터의 2504명의 개인유전체의 하위 , 상위 인구집단(sub & super population) 구성. 2014년도 1000 지놈 데이터의 26 하위 그리고 4 상위 인구집단 구성정보 (2015 6). A. 상위 인구집단 4개에 따른 2504명의 분포(AFR(African), EUR(European),ASN(EAS;East Asian),ASN(SAS;South Asian), AMR(Ad Mixed American)). B. 하위 인구집단 26개에 따른 2504명의 분포(AFR(YRI(Yoruba in Ibadan, Nigeria),LWK(Luhya in Webuye, Kenya),GWD(Gambian in Western Divisions in the Gambia),MSL(Mende in Sierra Leone),ESN(Esan in Nigeria),ASW(Americans of African Ancestry in SW USA),ACB(African Caribbeans in Barbados)),EUR(CEU(Utah Residents (CEPH) with Northern andWestern European Ancestry),TSI(Toscani in Italia), FIN(Finnish in Finland),GBR(British in England and Scotland),IBS(Iberian Population in Spain)), ASN(EAS,SAS;CHB(Han Chinese in Bejing, China),JPT(apanese in Tokyo, Japan),CHS(Southern Han Chinese),CDX(Chinese Dai in Xishuangbanna, China),KHV(Kinh in Ho Chi Minh City, Vietnam), GIH(Gujarati Indian from Houston, Texas),PJL(Punjabi from Lahore, Pakistan),BEB(Bengali from Bangladesh), STU(Sri Lankan Tamil from the UK), ITU(Indian Telugu from the UK)),AMR(MXL(Mexican Ancestry from Los Angeles USA),PUR(Puerto Ricans from Puerto Rico),CLM(Colombians from Medellin, Colombia),PEL(Peruvians from Lima, Peru)).

A.

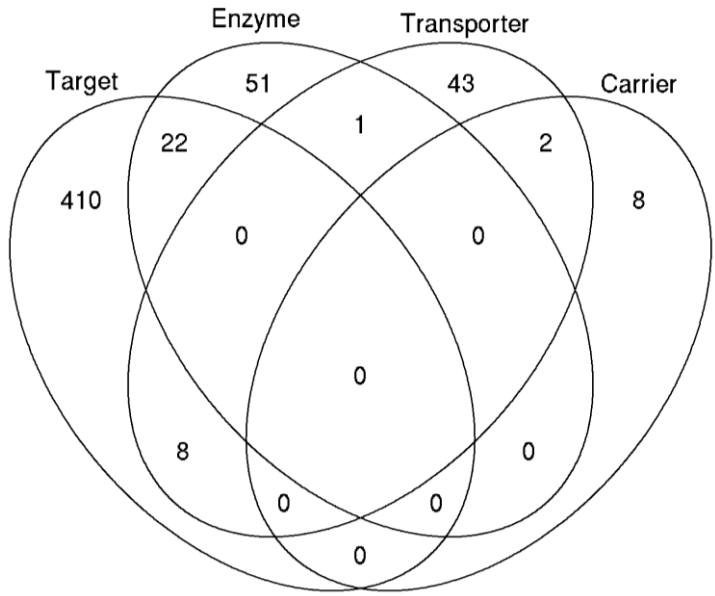

AMR(347)
AFR(661)
ASN(993)
EUR(503)

B

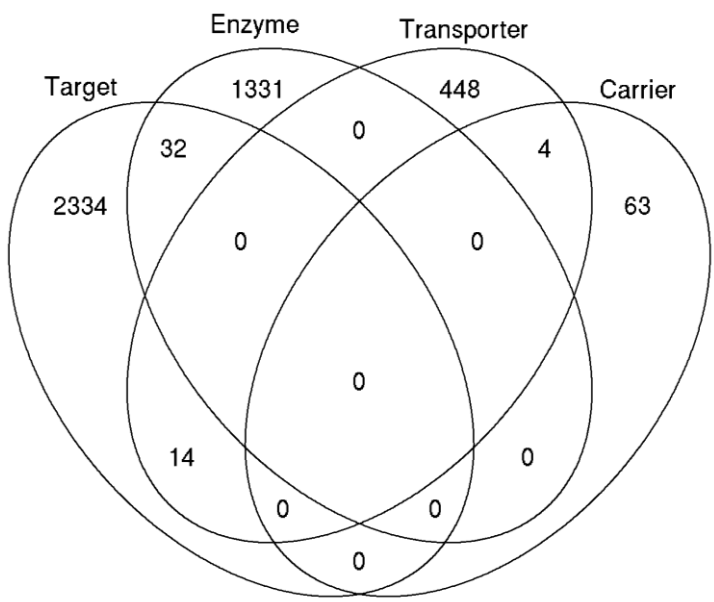

EUR
ASN
AMR
AFR
Number of person genome data
0
20
40
60
80
100
120
GWD
YRI
IBS
TSI
CHS
JPT
PUR
CHB
GIH
ITU
STU
CEU
ESN
FIN
KHV
LWK
ACB
PJL
CLM
CDX
GBR
BEB
MSL
PEL
MXL
ASW
Sub population in 2014 1000 Genome data

보충 그림 2. Pharmsafe 알고리즘에 사용된 약물학적 유전자 군별 약물, 약물-유전자 연관 개수. A. Pharmsafe 알고리즘에 사용된 약물 497 의 약물학적 유전자 군별 분포. B. Pharmsafe 알고리즘에 사용된 약물-유전자 연관 4,426 개의 분포.

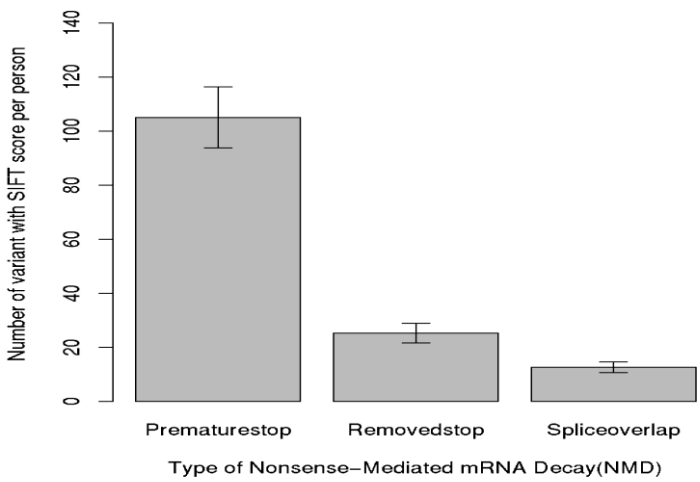
A


Enzyme
Transporter
Target
Carrier
51
43
22
1
2
410
8
0
0
0
8
0
0
0
0

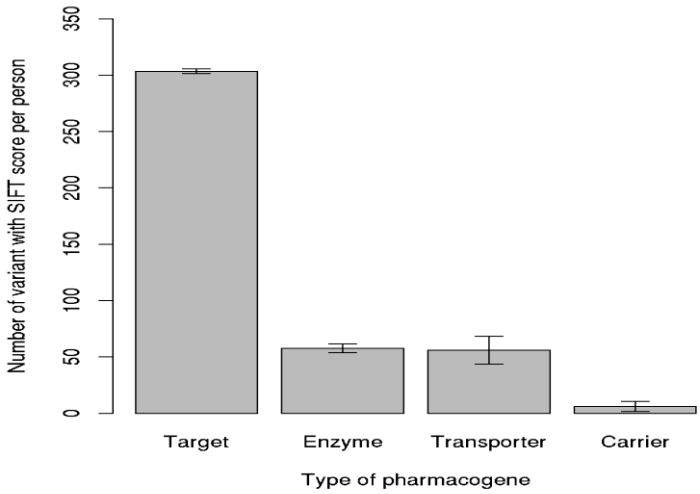
B


Enzyme
Transporter
Target
Carrier
1331
448
32
0
4
2334
63
0
0
0
14
0
0
0
0

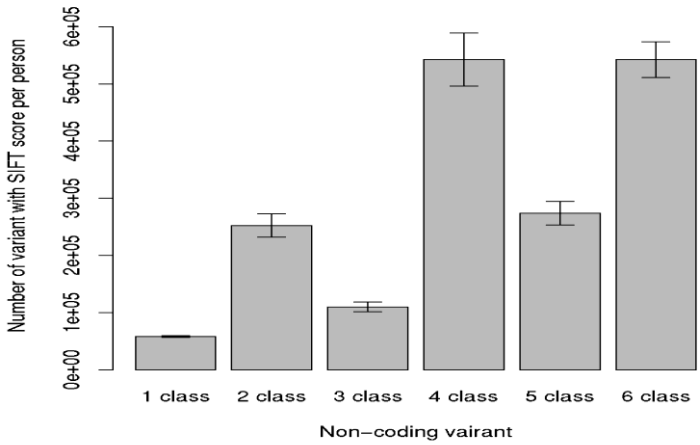
보충 그림 3. 각 생물학적 지식요소별 1000 지놈 데이터에 속하는 변이 개수.



NMD type	No. of variants (mean ± SD)
Prematurestop	105.05±11.23(75~146)
Removedstop	25.19±3.61(16~39)
Spliceoverlap	12.54±1.98(6~20)
Total	142.79 ±13.82(109~194)

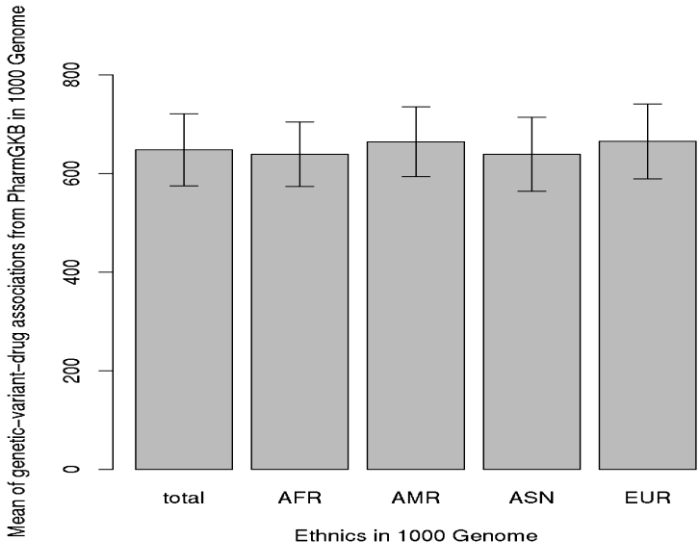


Pharmacogene type	No. of variants (mean ± SD)
Target	303.4±2.21(292~322)
Enayme	57.6±3.91(52~67)
Transporter	56.0±12.40(50~62)
Carrier	6.0±4.47(3~8)
Total	423±12.44(408~436)



Non−coding variant classes	No. of variants (mean ± SD)
1	58,104.42±1490.867 (53,261~6,2725)
2	252,297.80±20,426.453 (224,145~299,733)
3	109,733.28±8,462.391 (98,148~128,443)
4	542,564.47±46,231.893 (481,555~643,982)
5	273,761.18±20,501.018 (246,910~317,964)
6	542,329.02±31,127.380 (495,426~610,183)
Total	1,778,784±126,353.8 (1,600,470~2,056,404)

보충 그림 4. PharmsGKB 와 1000 지놈에 공통적으로 속하는 유전적 변이-약물(genetic-variant-drug associations;GVDA).



Ethnic*	No. of GVDA (mean ± SD)
AFR	638.7±65.3 (505~844)
AMR	663.9±70.7(529~891)
ASN	639.0±75.0(474~868)
EUR	664.7±75.7(493~869)
Total	647.5±73.14(474~891)

*PharmGKB ethnic information AFR(Black or African American), AMR and EUR(White), ASN(Asian). 1000 Genome ethnic information AFR(African), EUR(European), ASN(EAS;East Asian, SAS;South Asian), AMR(Ad Mixed American)

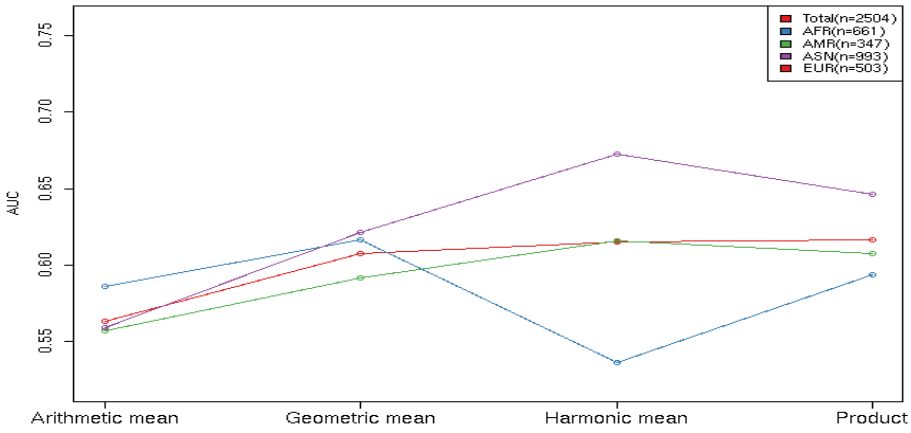
보충 그림 5. 497 전체 약물집합과 각 29가지 약물 군별로 7가지 각 요소에 포함된 54 가지 조건별로 가중치를 반영하여 Pharmsafe 알고리즘을 인종, 비인종으로 평가 .

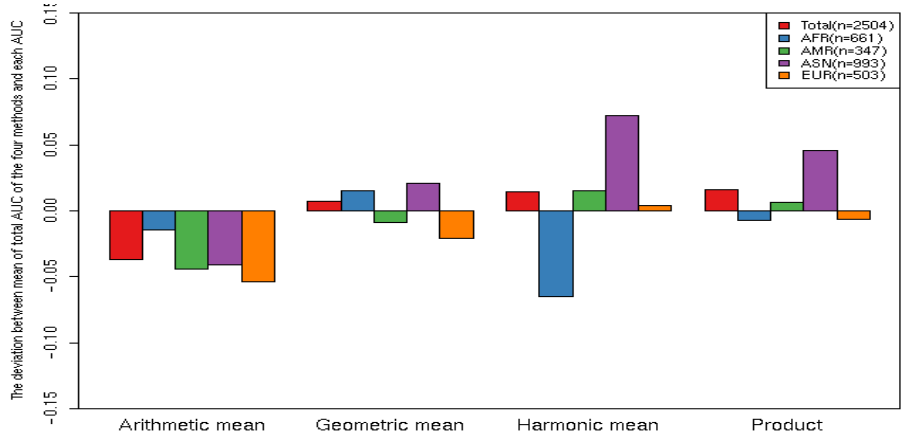
왼쪽 꺾은선 그래프는 각 조건별 AUC를 4가지 인종과 모든 인종을 합한 전체인종 별로 표현한 그래프이다. 오른쪽 그래프는 기준 AUC(기하평균의 AUC)에서 각 조건의 AUC를 뺀 편차이다(단 중심경향법은 기준 AUC를 각 방법의 전체(total) AUC의 평균을 사용하였다). 각 그래프의 색상은 적색은 전체, 청색은 아프리카, 초록색은 미국, 보라색은 아시아, 주황색은 유럽을 나타낸다. 표는 전체 인종 그리고 각 인종별 AUC를 기술해 놓았다. Heatmap은 인종 비 인종별로 표시하였으며 각 셀의 값은 기준 AUC(기하평균의 AUC)에서 각 조건의 AUC를 뺀 편차이며, 보라색은 편차의 상승을 의미하고 청색은 편차의 하강을 의미한다. 각 셀안에 기입된 숫자는 각 실험별 AUC값을 의미한다. 왼쪽 회색, 노란색, 연한 회색은 ATC 에서 추출한 해부학적 그룹 14 가지([A] Alimentary tract and metabolism,[B] Blood and blood forming organs,[C] Cardiovascular system,[D] Dermatologicals,[G] Genito urinary system and sex hormones, systemic [H] hormonal preparations and excl. sex hormones and insulins,[J] Antiinfectives for systemic use,[L] Antineoplastic and immunomodulating agents, [M] musculo-skeletal system,[N] Nervous system,[P] antiparasitic products, insecticides and repellents,[R] Respiratory system,[S] Sensory organs,[V] Various),HOCC에서 추출한 15가지 가장 자주 처방받은 약물 군([A02BC] Proton pump inhibitors,[A10] Drugs used in diabetes, [C02] Antihypertensives,[C03] Diuretics,[C07] Beta blocking agents,[C09A] ACE inhibitors plain,[C08] Calcium channel blockers,[C10] Lipid modifying agents, [G03] Sex hormones and modulators of the genital system,[H03] Thyroid therapy,[N02] Analgesics,[N05B,N05C] Anxiolytics and hypnotics/sedatives, [N06A] Antidepressants, [R03] Drugs for obstructive airway diseases,[R06] Antihistamines for systemic use) 그리고 약물관련 유전자, 비 약물관련 유전자를 의미한다. 각 heatmap 그래프에서 공백은 해당

약물군의 속하는 약물에 대한 변이가 해당 인종에서 존재하지 않아 계산이 되지 않은 것을 의미한다.

A. 중심경향방법(central tendency method)

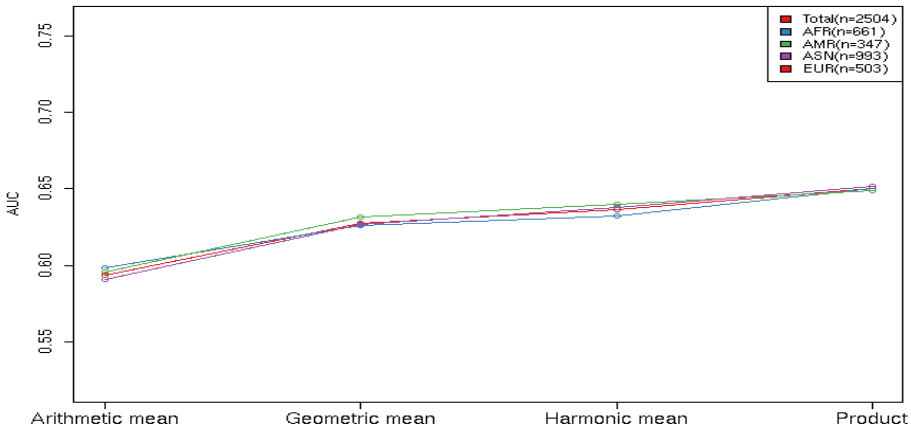
인종평가

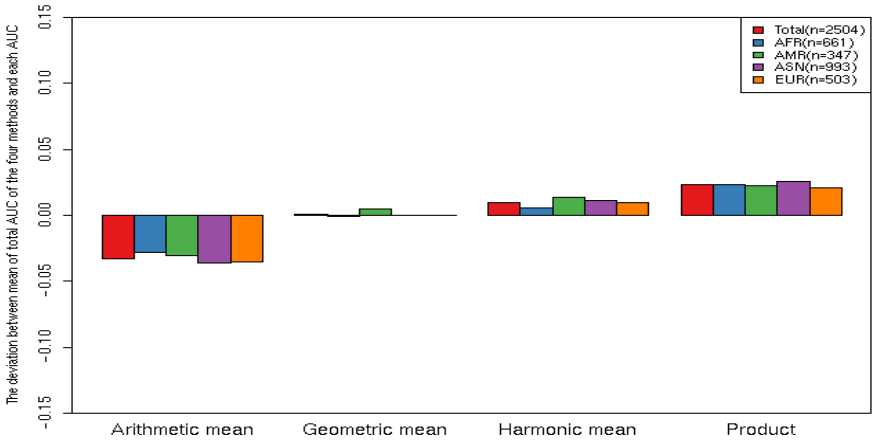




	Arithmetic mean	Geometric mean	Harmonic mean	Product
Total(*n*=2504)	0.5633	0.6076	0.6149	0.6163
AFR(*n*=661)	0.5858	0.6161	0.5358	0.5932
AMR(*n*=347)	0.5565	0.5918	0.6156	0.6071
ASN(*n*=993)	0.5592	0.6214	0.6726	0.6459
EUR(*n*=503)	0.5467	0.5799	0.6046	0.5944

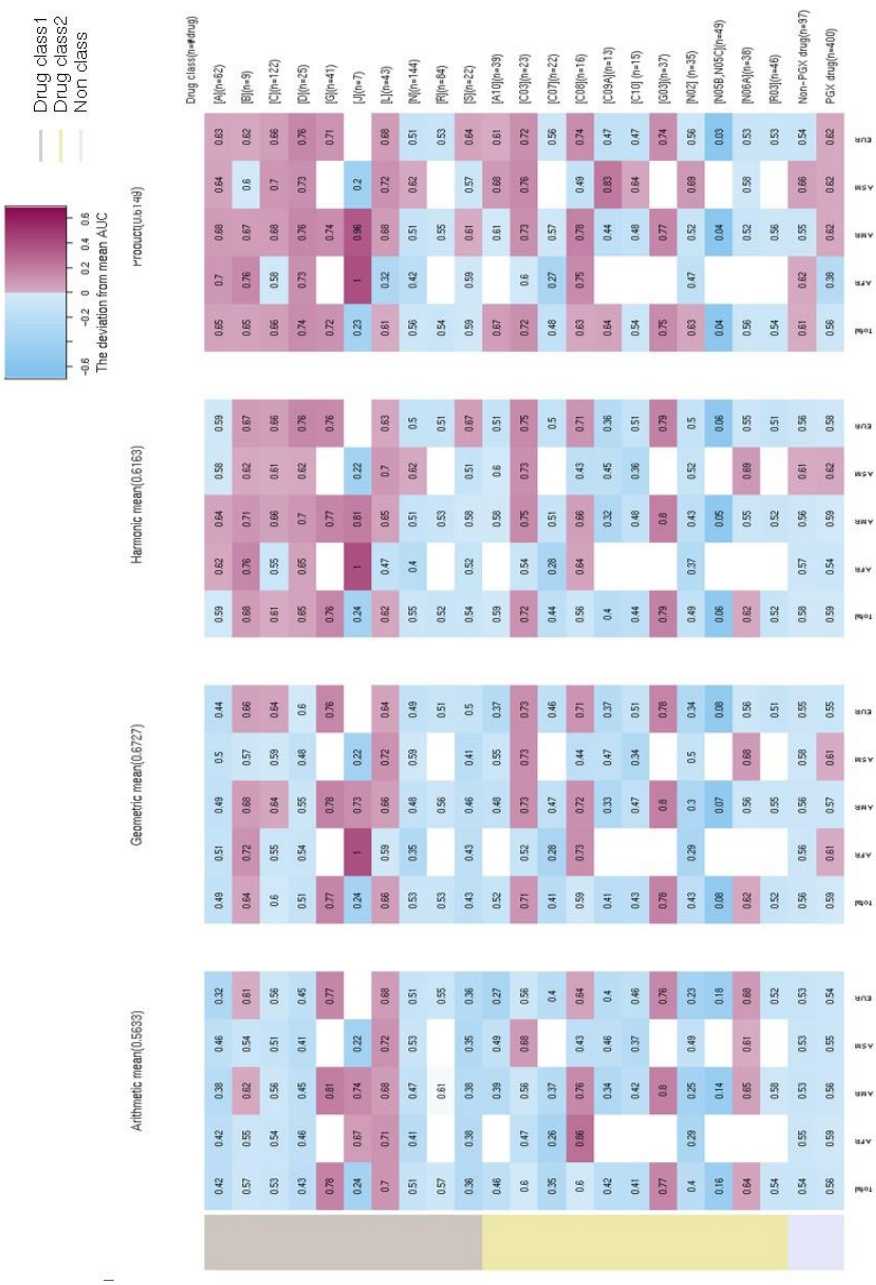
비인종평가



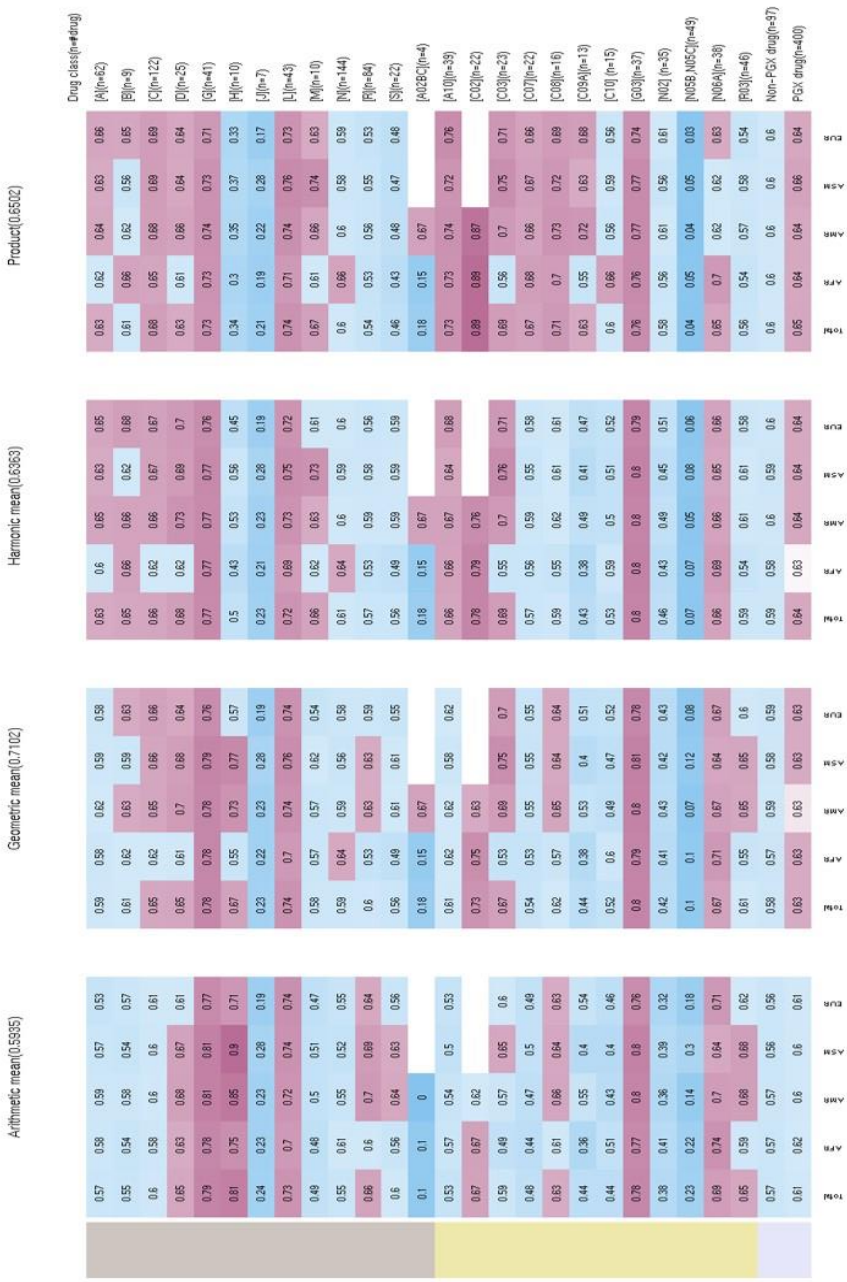


	Arithmetic mean	Geometric mean	Harmonic mean	Product
Total(*n*=2504)	0.5935	0.6271	0.6363	0.6502
AFR(*n*=661)	0.5983	0.6261	0.6325	0.6502
AMR(*n*=347)	0.5958	0.6314	0.6402	0.6491
ASN(*n*=993)	0.5908	0.6266	0.6376	0.652
EUR(*n*=503)	0.5912	0.6264	0.636	0.6472

인종평가

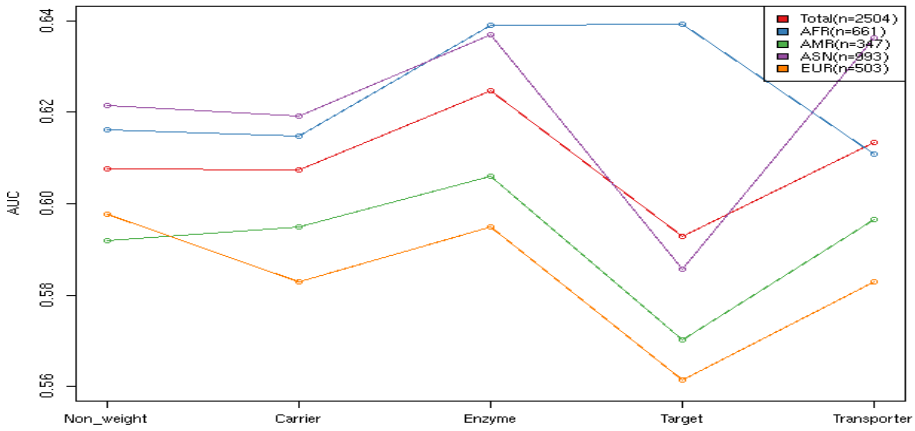

Drug class1
Drug class2
Non class
The deviation from mean AUC
−0.6 −0.2 0 0.2 0.4 0.6
Arithmetic mean(0.5633)
Geometric mean(0.6727)
Harmonic mean(0.6163)
Product(0.149)
Drug class(n=#drug)
[A](n=62)
[B](n=9)
[C](n=122)
[D](n=25)
[G](n=41)
[J](n=7)
[L](n=43)
[N](n=144)
[R](n=64)
[S](n=22)
[A10](n=39)
[C03](n=23)
[C07](n=22)
[C08](n=16)
[C09A](n=13)
[C10] (n=15)
[G03](n=37)
[N02] (n=35)
[N05B,N05C](n=49)
[N06A](n=38)
[R03](n=46)
Non−PGX drug(n=97)
PGX drug(n=400)
Total
AFR
AMR
ASN
EUR

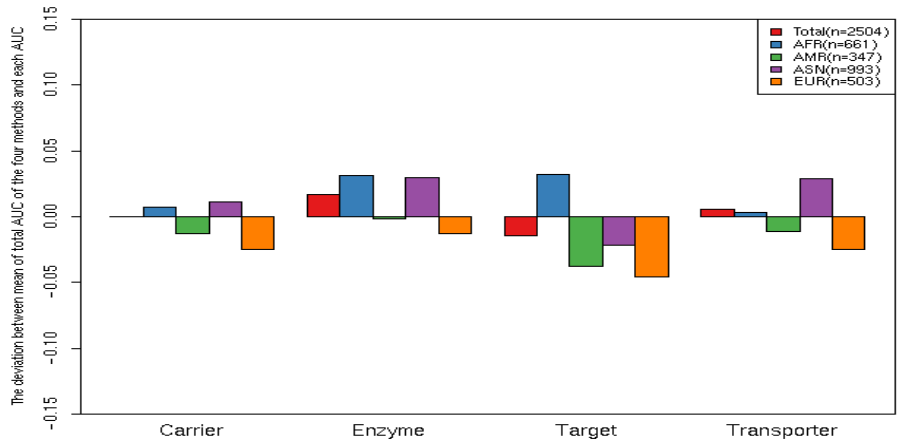
비인종평가


Arithmetic mean(0.5935)
Geometric mean(0.7102)
Harmonic mean(0.6363)
Product(0.6502)
Drug class(n=#drug)
[A](n=62)
[B](n=9)
[C](n=122)
[D](n=25)
[G](n=41)
[H](n=10)
[J](n=7)
[L](n=43)
[M](n=10)
[N](n=144)
[R](n=84)
[S](n=22)
[A02BC](n=4)
[A10](n=39)
[C02](n=22)
[C03](n=23)
[C07](n=22)
[C08](n=16)
[C09A](n=13)
[C10] (n=15)
[G03](n=37)
[N02] (n=35)
[N05B ,N05C](n=49)
[N06A](n=38)
[R03](n=46)
Non-PGX drug(n=97)
PGX drug(n=400)
total
AFR
AMR
ASN
EUR

B. 약물학적 유전자 종류(Pharmacogene type)

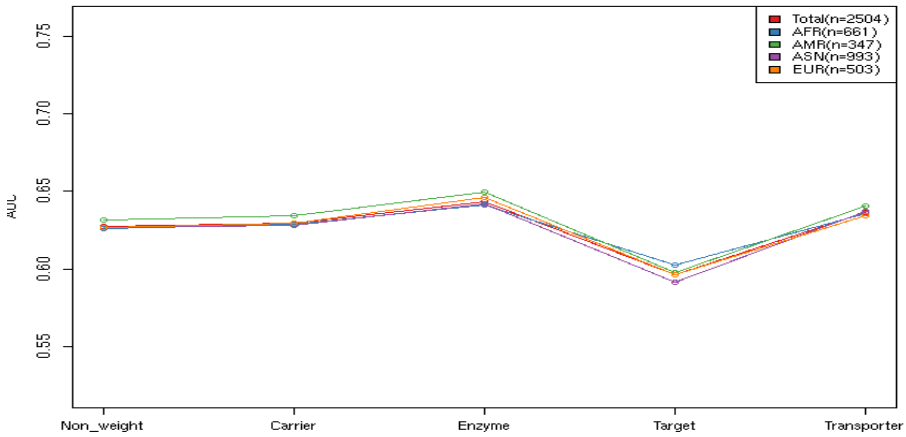
인종평가

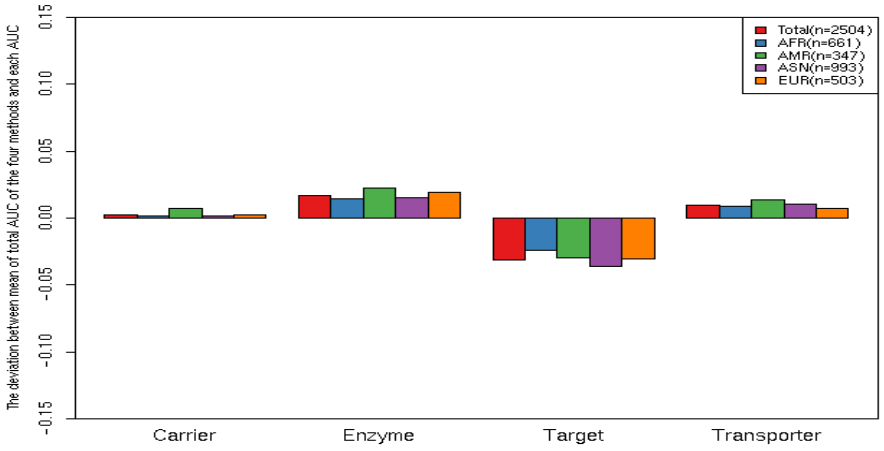




	Non-weight	Carrier	Enzyme	Target	Transporter
Total(*n*=2504)	0.6076	0.6074	0.6248	0.5928	0.6134
AFR(*n*=661)	0.6161	0.6149	0.639	0.6393	0.6109
AMR(*n*=347)	0.5918	0.5948	0.6059	0.5701	0.5965
ASN(*n*=993)	0.6214	0.6191	0.637	0.5857	0.6363
EUR(*n*=503)	0.5799	0.583	0.5949	0.5615	0.583

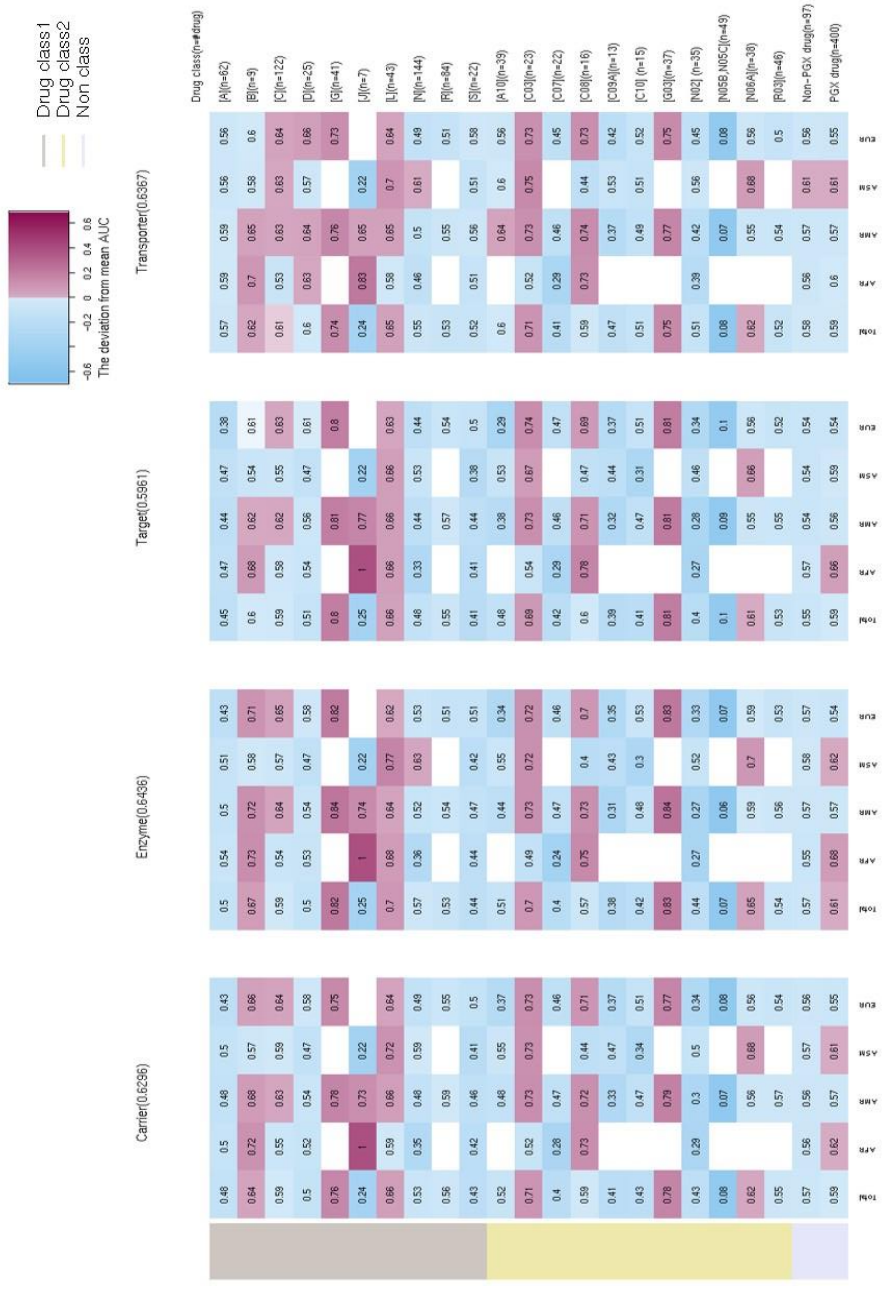
비인종평가



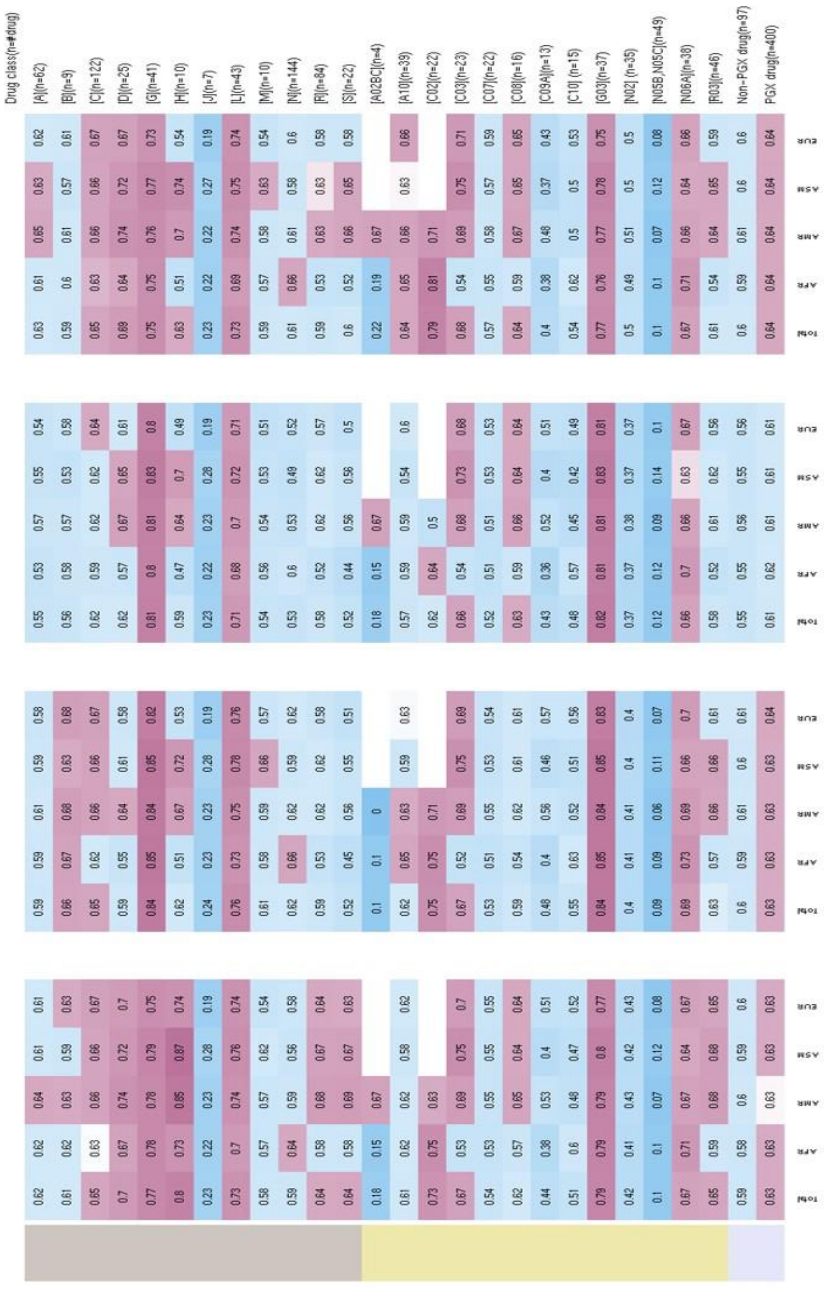


	Non-weight	Carrier	Enzyme	Target	Transporter
Total(*n*=2504)	0.6271	0.6296	0.6436	0.5961	0.6367
AFR(*n*=661)	0.6261	0.6289	0.6411	0.6026	0.6357
AMR(*n*=347)	0.6314	0.6343	0.6493	0.5974	0.6409
ASN(*n*=993)	0.6266	0.6284	0.6422	0.5911	0.6372
EUR(*n*=503)	0.6264	0.6298	0.6459	0.5964	0.6343

인종평가

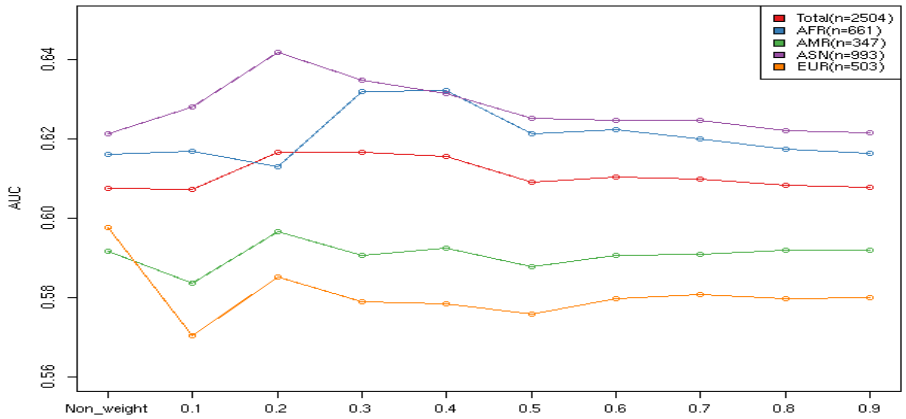

Drug class1
Drug class2
Non class
The deviation from mean AUC
−0.6
−0.2
0
0.2
0.4
0.6
Carrier(0.6296)
Enzyme(0.6436)
Target(0.5961)
Transporter(0.6367)
Drug class(n=#drug)
[A](n=62)
[B](n=9)
[C](n=122)
[D](n=25)
[G](n=41)
[J](n=7)
[L](n=43)
[N](n=144)
[R](n=84)
[S](n=22)
[A10](n=39)
[C03](n=23)
[C07](n=22)
[C08](n=16)
[C09A](n=13)
[C10] (n=15)
[G03](n=37)
[N02] (n=35)
[N05B,N05C](n=49)
[N06A](n=38)
[R03](n=46)
Non−PGX drug(n=37)
PGX drug(n=400)
total
AFR
AMR
ASN
EUR

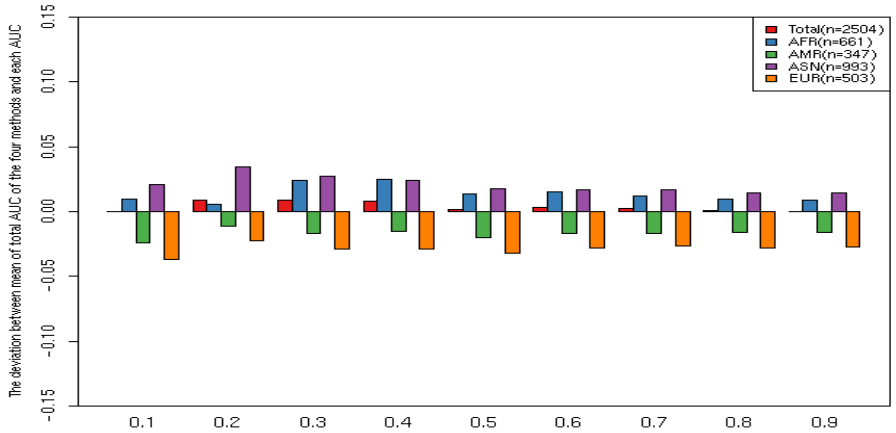
비인종평가


Carrier(0.6296)
Enzyme(0.6436)
Target(0.5961)
Transporter(0.6367)
Drug class(n=##drug)
[A](n=62)
[B](n=9)
[C](n=122)
[D](n=25)
[G](n=41)
[H](n=10)
[J](n=7)
[L](n=43)
[M](n=10)
[N](n=144)
[R](n=84)
[S](n=22)
[A02BC](n=4)
[A10](n=39)
[C02](n=22)
[C03](n=23)
[C07](n=22)
[C08](n=16)
[C09A](n=13)
[C10] (n=15)
[G03](n=37)
[N02] (n=35)
[N05B,N05C](n=49)
[N06A](n=38)
[R03](n=46)
Non-PGX drug(n=97)
PGX drug(n=400)
total
AFR
AMR
ASN
EUR

C. 변이 점수 윈저화(Variant score winsorization)

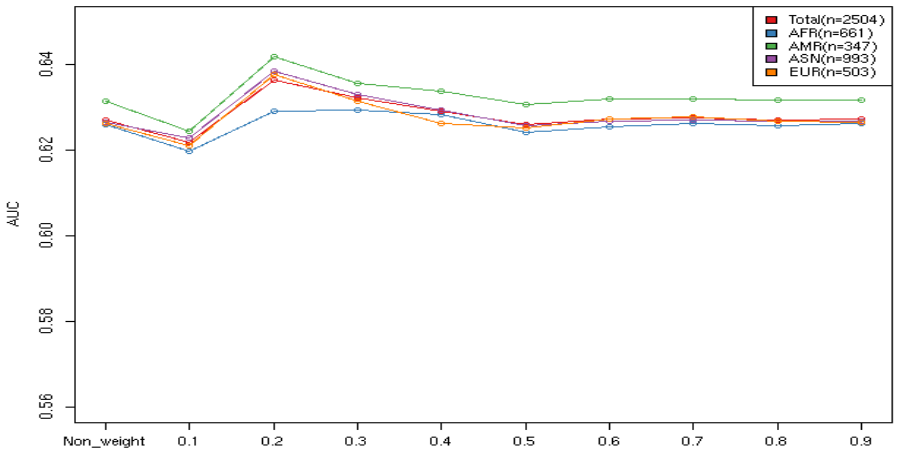
인종평가

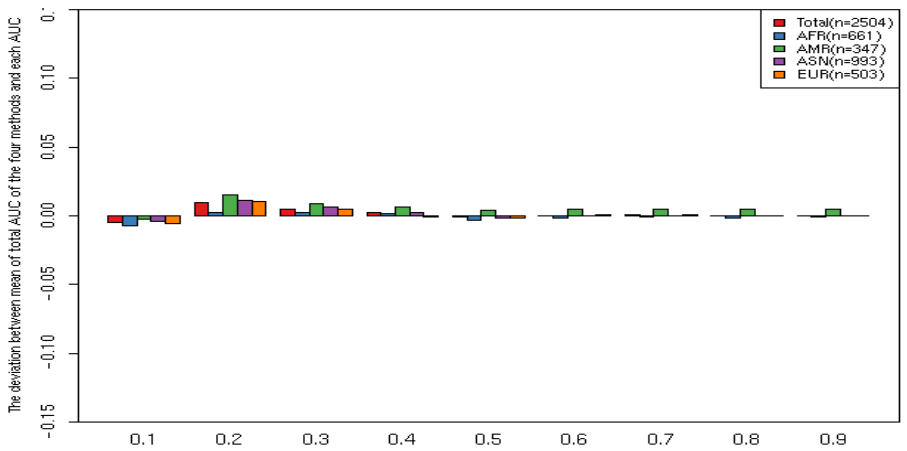




	Non-weight	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Total(*n*=2504)	0.6076	0.6074	0.6166	0.6167	0.6157	0.6092	0.6104	0.61	0.6083	0.6078
AFR(*n*=661)	0.6161	0.6169	0.613	0.632	0.6324	0.6214	0.6225	0.62	0.6175	0.6164
AMR(*n*=347)	0.5918	0.5837	0.5966	0.5906	0.5926	0.5879	0.5907	0.591	0.5919	0.5919
ASN(*n*=993)	0.6214	0.6282	0.6419	0.6348	0.6314	0.6254	0.6247	0.6248	0.6222	0.6217
EUR(*n*=503)	0.5799	0.5704	0.5853	0.5789	0.5785	0.5758	0.5797	0.5809	0.5798	0.58

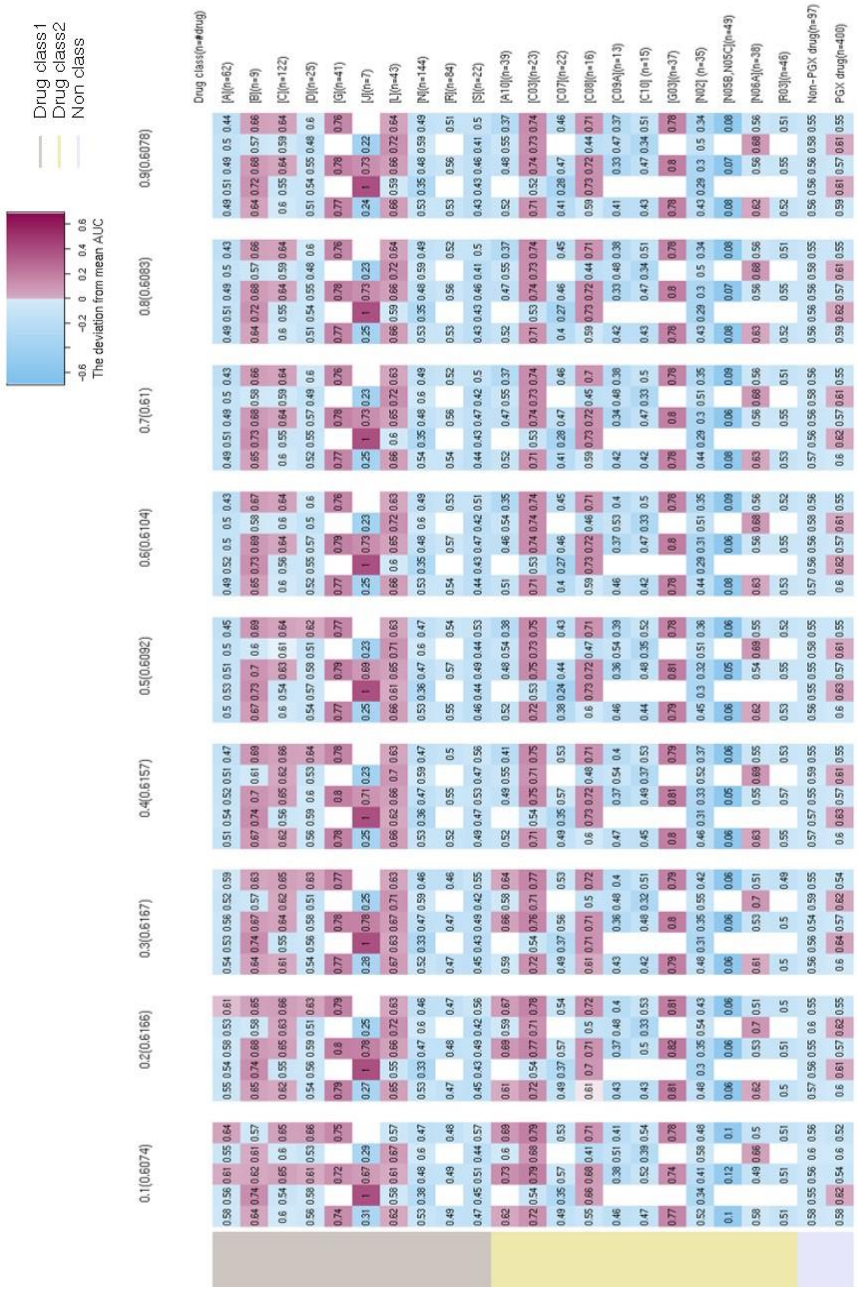
비인종평가



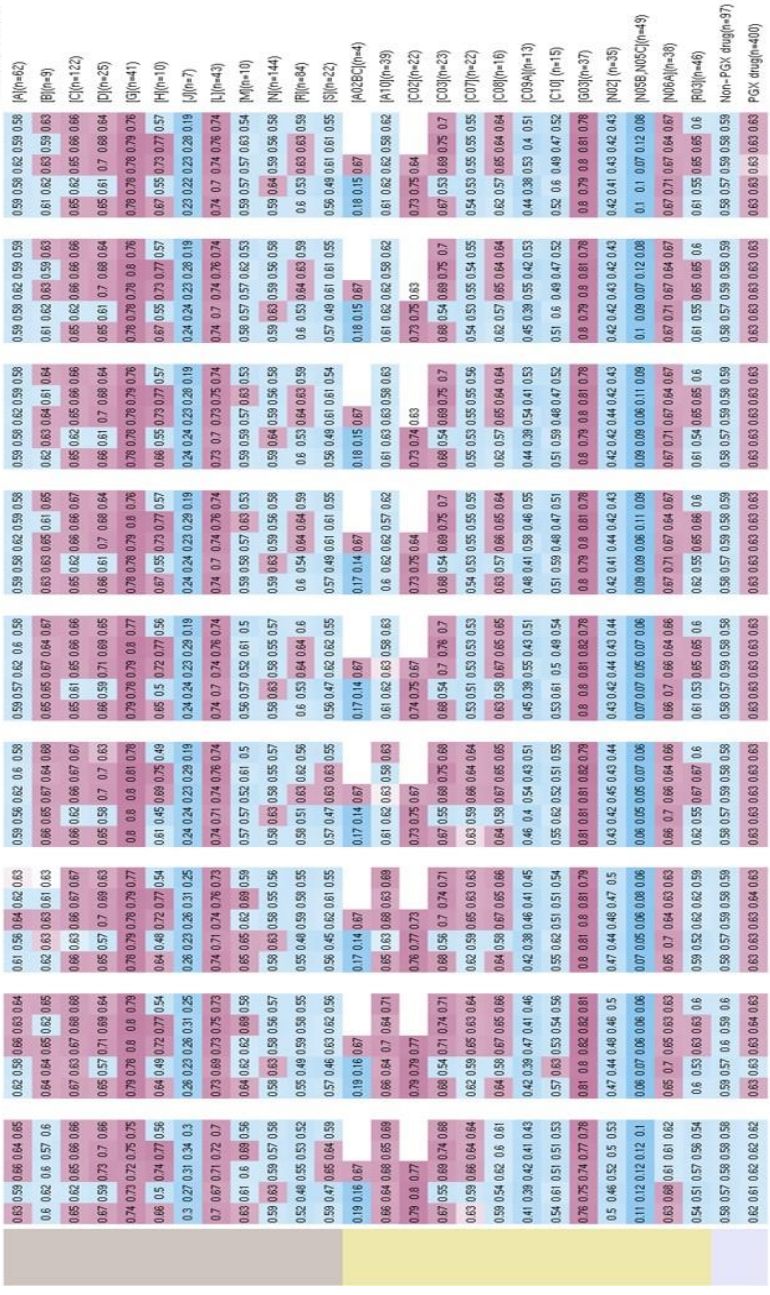


	Non-weight	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Total(n=2504)	0.6271	0.6219	0.6364	0.6322	0.6291	0.6259	0.6273	0.6277	0.6271	0.6273
AFR(n=661)	0.6261	0.6198	0.6292	0.6295	0.6283	0.6241	0.6254	0.6263	0.6257	0.6262
AMR(n=347)	0.6314	0.6244	0.6419	0.6356	0.6338	0.6307	0.632	0.632	0.6318	0.6317
ASN(n=993)	0.6266	0.6228	0.6385	0.6331	0.6293	0.6257	0.6268	0.6271	0.6267	0.6268
EUR(n=503)	0.6264	0.6211	0.6377	0.6315	0.6264	0.6253	0.6274	0.6278	0.6267	0.6266

인종평가

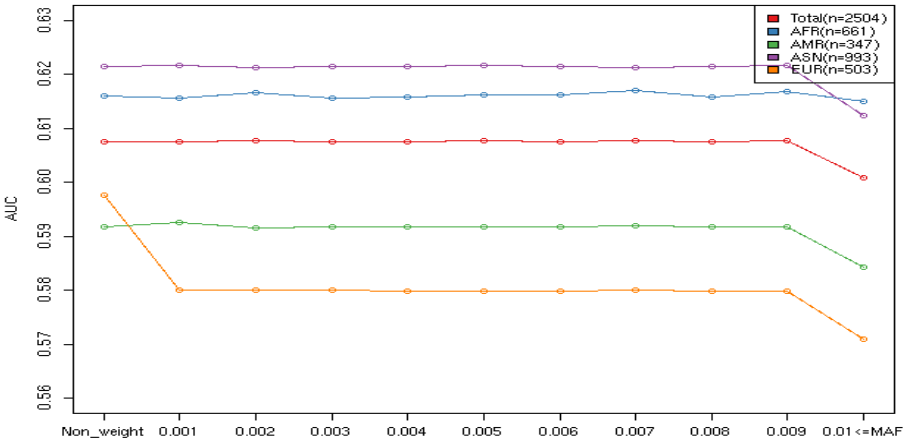

Drug class1
Drug class2
Non class
The deviation from mean AUC
−0.6 −0.2 0 0.2 0.4 0.6
Drug class(n=#drug)
[A](n=62)
[B](n=9)
[C](n=122)
[D](n=25)
[G](n=41)
[J](n=7)
[L](n=43)
[N](n=144)
[R](n=84)
[S](n=22)
[A10](n=39)
[C03](n=23)
[C07](n=22)
[C08](n=16)
[C09A](n=13)
[C10](n=15)
[G03](n=37)
[N02](n=35)
[N05B,N05C](n=49)
[N06A](n=38)
[R03](n=46)
Non-PGx drug(n=97)
PGx drug(n=400)
0.1(0.6074)
0.2(0.6166)
0.3(0.6167)
0.4(0.6157)
0.5(0.6092)
0.6(0.6104)
0.7(0.61)
0.8(0.6083)
0.9(0.6078)

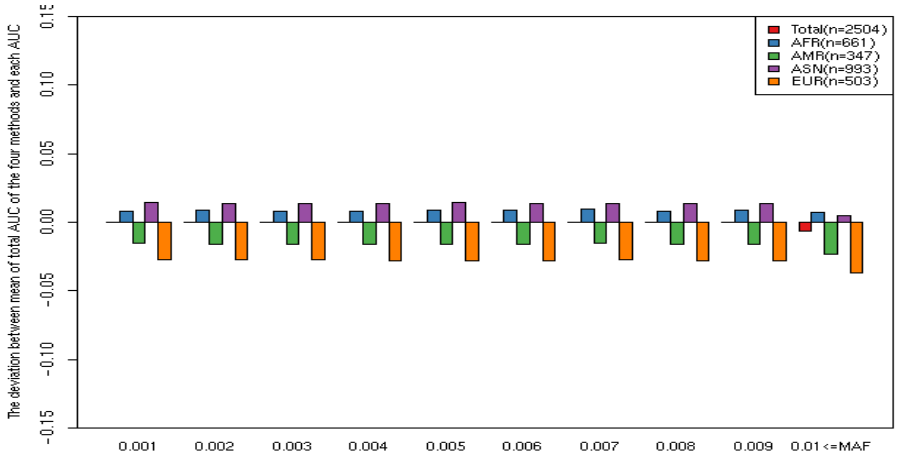
비인종평가


Drug class(n=#drug)
0.1(0.6219)
0.2(0.6364)
0.3(0.6322)
0.4(0.6291)
0.5(0.6259)
0.6(0.6273)
0.7(0.6277)
0.8(0.6271)
0.9(0.6273)
[A](n=62)
[B](n=9)
[C](n=122)
[D](n=25)
[G](n=41)
[H](n=10)
[J](n=7)
[L](n=43)
[M](n=10)
[N](n=144)
[R](n=84)
[S](n=22)
[A02BC](n=4)
[A10](n=39)
[C02](n=22)
[C03](n=23)
[C07](n=22)
[C08](n=16)
[C09A](n=13)
[C10] (n=15)
[G03](n=37)
[N02] (n=35)
[N05B,N05C](n=49)
[N06A](n=30)
[R03](n=46)
Non-PGX drug(n=97)
PGX drug(n=400)

D. 낮은 대립형질 빈도(Minor Allele Frequency)

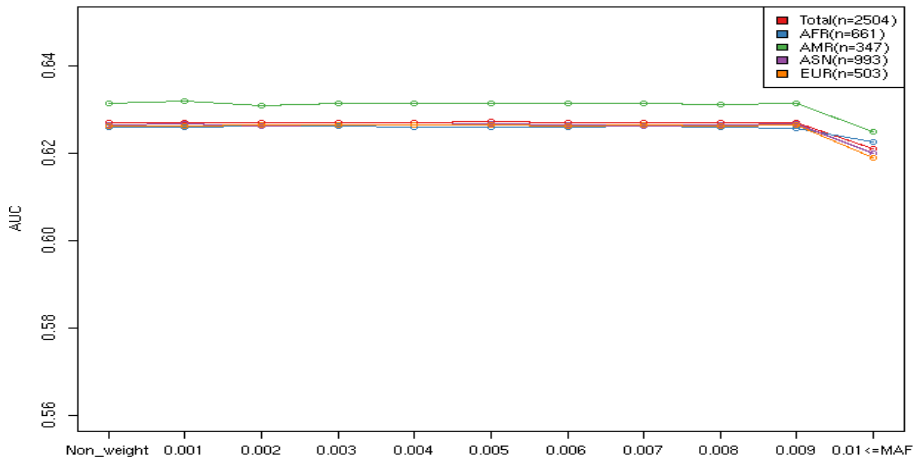
인종평가

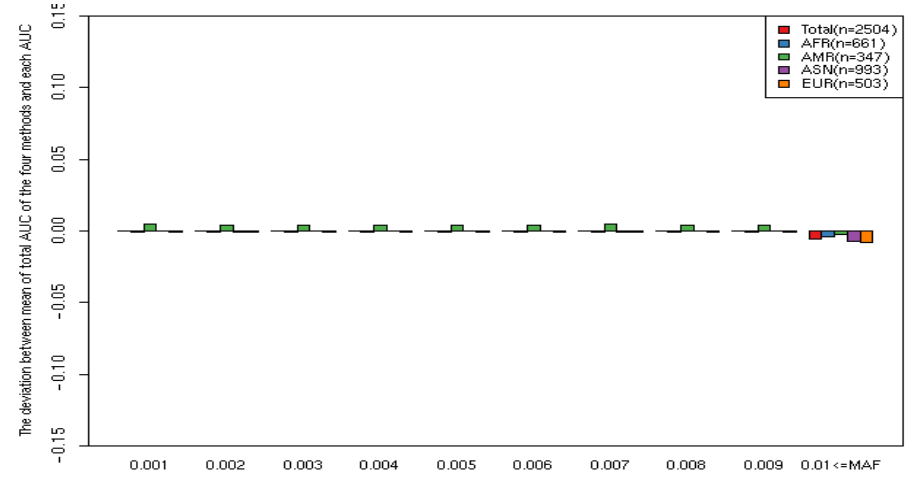




	Non-weight	0.001	0.002	0.003	0.004	0.005	0.006	0.007	0.008	0.009	0.01<=MAF
Total(*n*=2504)	0.6076	0.6076	0.6077	0.6075	0.6075	0.6077	0.6076	0.6078	0.6075	0.6078	0.6008
AFR(*n*=661)	0.6161	0.6156	0.6167	0.6157	0.6159	0.6162	0.6162	0.6171	0.6159	0.6168	0.6151
AMR(*n*=347)	0.5918	0.5926	0.5915	0.5918	0.5918	0.5918	0.5918	0.592	0.5917	0.5918	0.5843
ASN(*n*=993)	0.6214	0.6217	0.6213	0.6215	0.6215	0.6217	0.6214	0.6213	0.6215	0.6216	0.6123
EUR(*n*=503)	0.5799	0.58	0.5801	0.58	0.5799	0.5799	0.5799	0.5801	0.5799	0.5799	0.5709

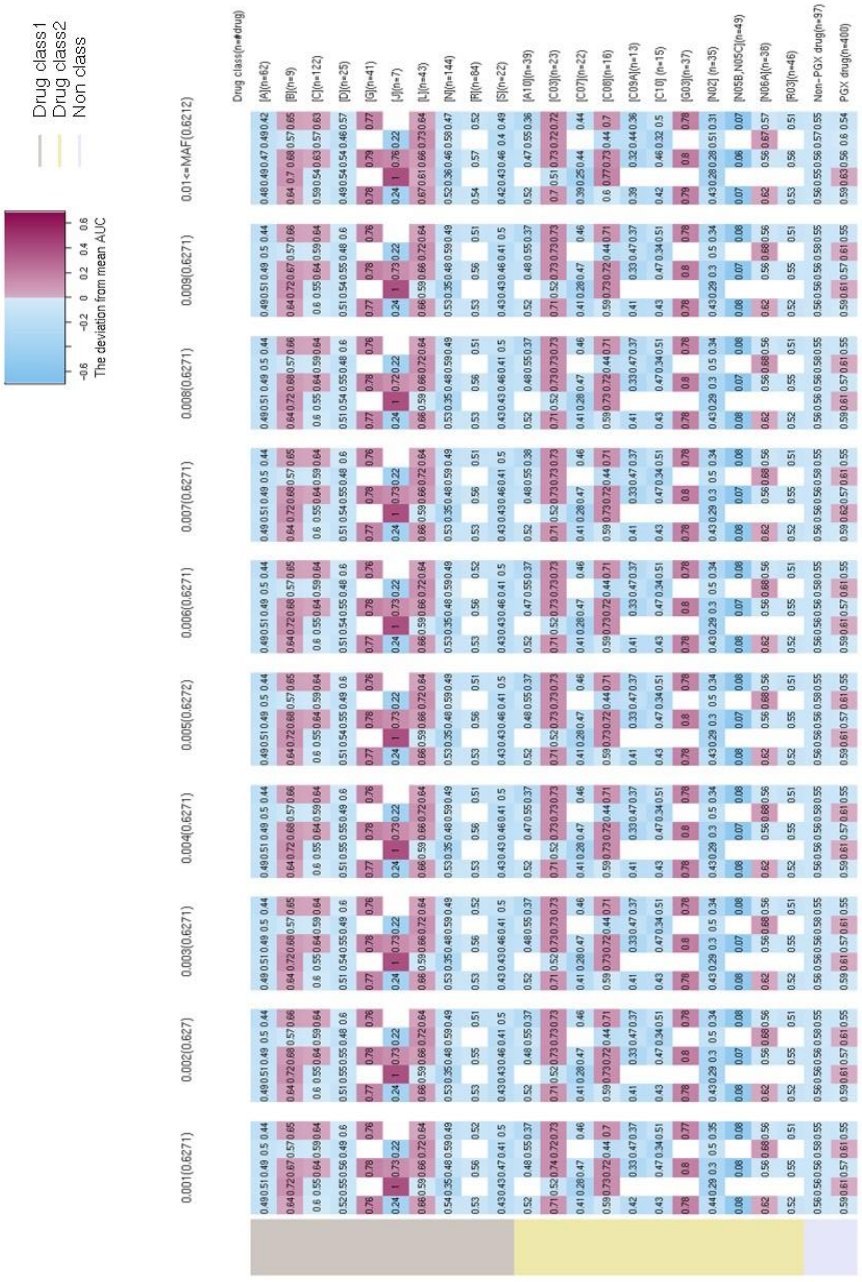
비인종평가



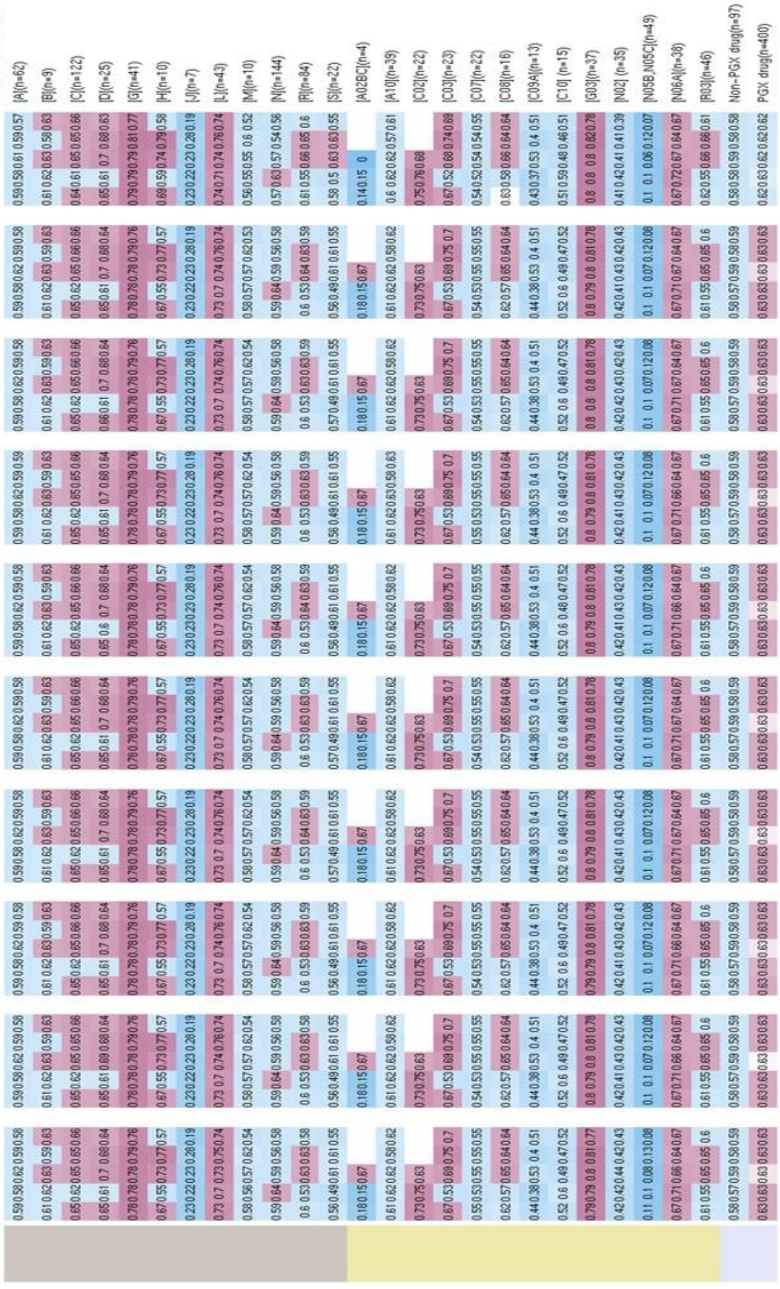


	Non-weight	0.001	0.002	0.003	0.004	0.005	0.006	0.007	0.008	0.009	0.01<=MAF
Total(*n*=2504)	0.6271	0.6271	0.627	0.6271	0.6271	0.6272	0.6271	0.6271	0.6271	0.6271	0.6212
AFR(*n*=661)	0.6261	0.626	0.6262	0.6262	0.626	0.6261	0.626	0.6262	0.6261	0.6258	0.6226
AMR(*n*=347)	0.6314	0.6319	0.631	0.6314	0.6314	0.6314	0.6314	0.6316	0.6313	0.6314	0.6249
ASN(*n*=993)	0.6266	0.6267	0.6264	0.6266	0.6266	0.6268	0.6266	0.6264	0.6266	0.6267	0.6201
EUR(*n*=503)	0.6264	0.6262	0.6265	0.6265	0.6265	0.6265	0.6264	0.6265	0.6264	0.6265	0.619

인종평가

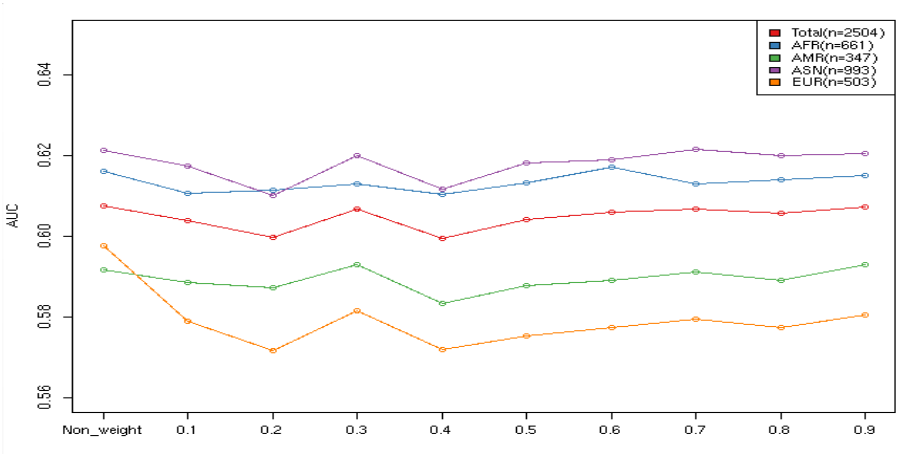

Drug class1
Drug class2
Non class
−0.6 −0.2 0 0.2 0.4 0.6
The deviation from mean AUC
Drug class(n=#drug)
0.001(0.6271)
0.002(0.627)
0.003(0.6271)
0.004(0.6271)
0.005(0.6272)
0.006(0.6271)
0.007(0.6271)
0.008(0.6271)
0.009(0.6271)
0.01<=MAF(0.6212)
[A](n=82)
[B](n=9)
[C](n=122)
[D](n=25)
[G](n=41)
[J](n=7)
[L](n=43)
[N](n=144)
[R](n=84)
[S](n=22)
[A10](n=39)
[C03](n=23)
[C07](n=22)
[C08](n=16)
[C09A](n=13)
[C10] (n=15)
[G03](n=37)
[N02] (n=35)
[N05B,N05C](n=49)
[N06A](n=38)
[P03](n=46)
Non-PGX drug(n=97)
PGX drug(n=400)

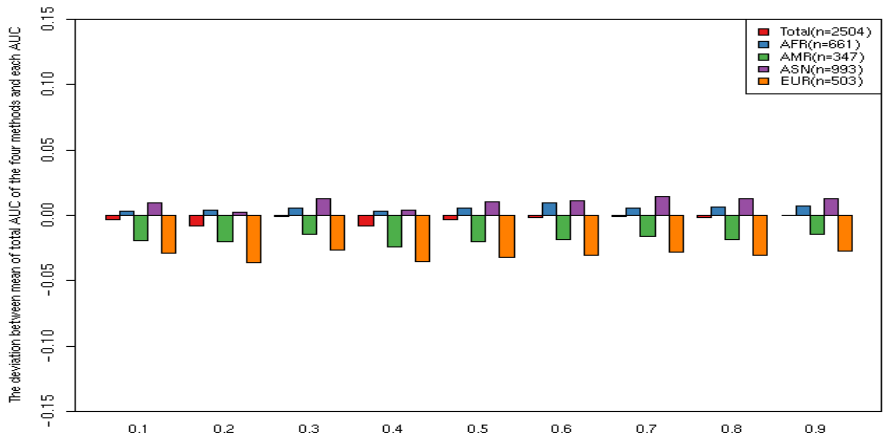
비인종평가


0.001(0.6271)
0.002(0.627)
0.003(0.6271)
0.004(0.6271)
0.005(0.6272)
0.006(0.6271)
0.007(0.6271)
0.008(0.6271)
0.009(0.6271)
0.01<=MAF(0.6212)
Drug class(n=#drug)
[A](n=62)
[B](n=9)
[C](n=122)
[D](n=25)
[G](n=41)
[H](n=10)
[J](n=7)
[L](n=43)
[M](n=10)
[N](n=144)
[R](n=84)
[S](n=22)
[A02BC](n=4)
[A10](n=39)
[C02](n=22)
[C03](n=23)
[C07](n=22)
[C08](n=16)
[C09A](n=13)
[C10] (n=15)
[G03](n=37)
[N02] (n=35)
[N05B,N05C](n=49)
[N06A](n=38)
[R03](n=46)
Non-PGx drug(n=97)
PGx drug(n=400)

E. 동형접합변이 비율 (Homozygote mutation rate)

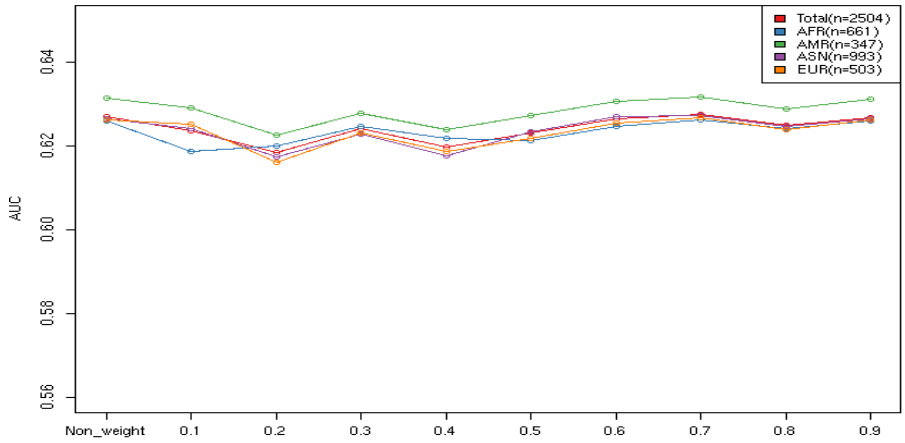
인종평가

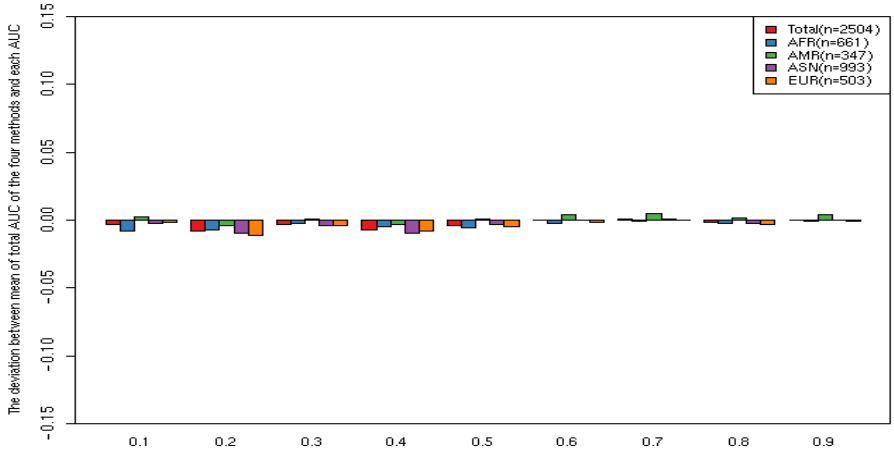




	Non-weight	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Total(*n*=2504)	0.6076	0.604	0.5997	0.6067	0.5995	0.6041	0.6061	0.6067	0.6057	0.6073
AFR(*n*=661)	0.6161	0.6108	0.6115	0.613	0.6104	0.6132	0.6172	0.6131	0.6142	0.6151
AMR(*n*=347)	0.5918	0.5885	0.5873	0.593	0.5835	0.5879	0.5891	0.5912	0.5891	0.5931
ASN(*n*=993)	0.6214	0.6176	0.6103	0.6201	0.6118	0.6182	0.6191	0.6217	0.6202	0.6207
EUR(*n*=503)	0.5799	0.5789	0.5717	0.5815	0.572	0.5754	0.5775	0.5795	0.5774	0.5806

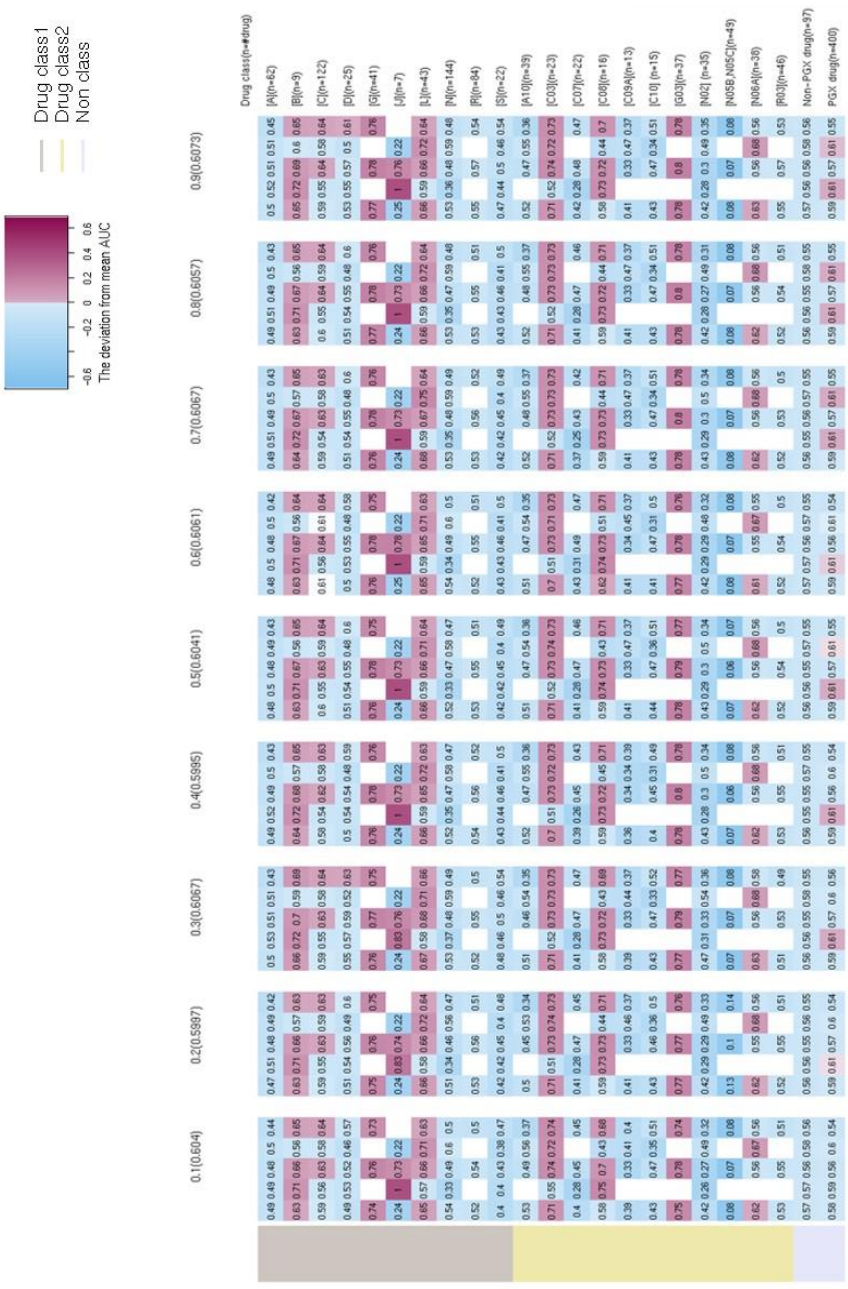
비인종평가



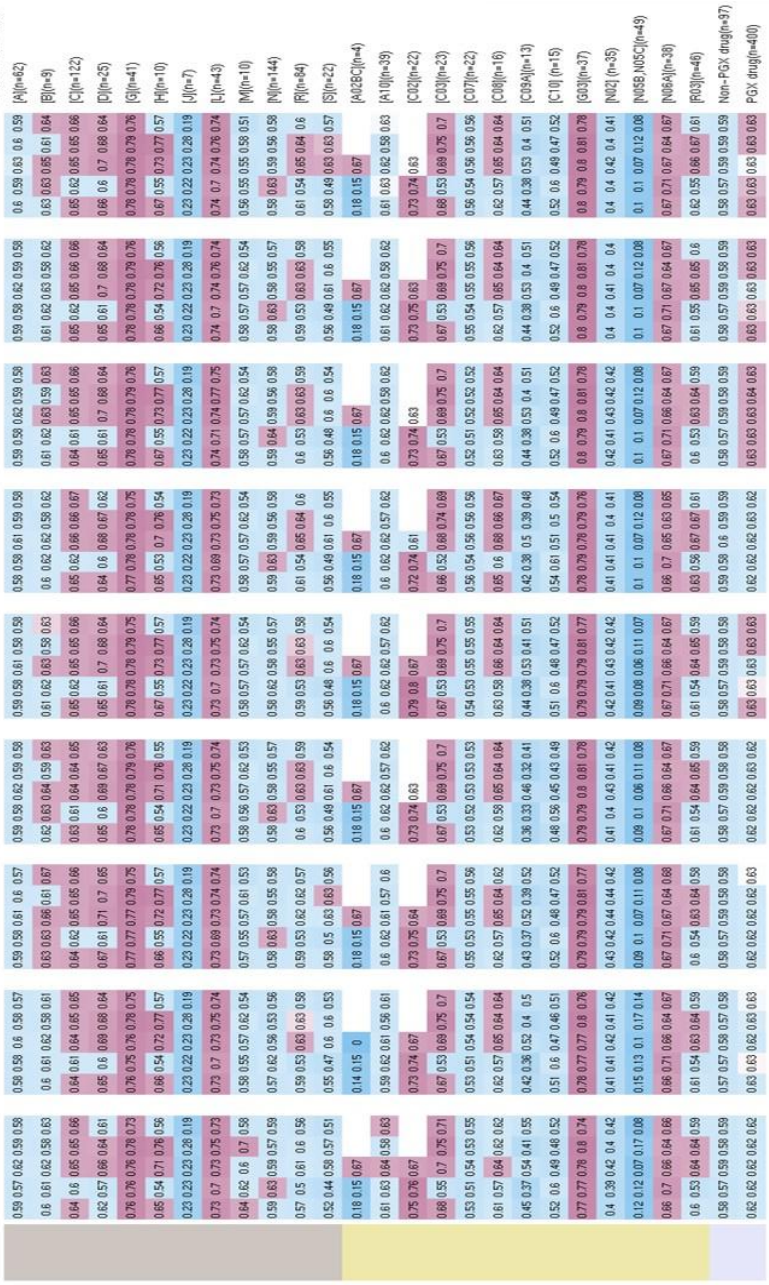


	Non-weight	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Total(n=2504)	0.6271	0.6236	0.6186	0.6241	0.6199	0.6231	0.6266	0.6276	0.625	0.6269
AFR(n=661)	0.6261	0.6187	0.6201	0.6248	0.6219	0.6214	0.6247	0.6262	0.6243	0.6259
AMR(n=347)	0.6314	0.6291	0.6226	0.6279	0.624	0.6274	0.6307	0.6317	0.6289	0.6311
ASN(n=993)	0.6266	0.6242	0.6174	0.6228	0.6177	0.6234	0.627	0.6274	0.6247	0.6266
EUR(n=503)	0.6264	0.6253	0.6161	0.6231	0.6188	0.6218	0.6254	0.6268	0.6239	0.6263

인종평가

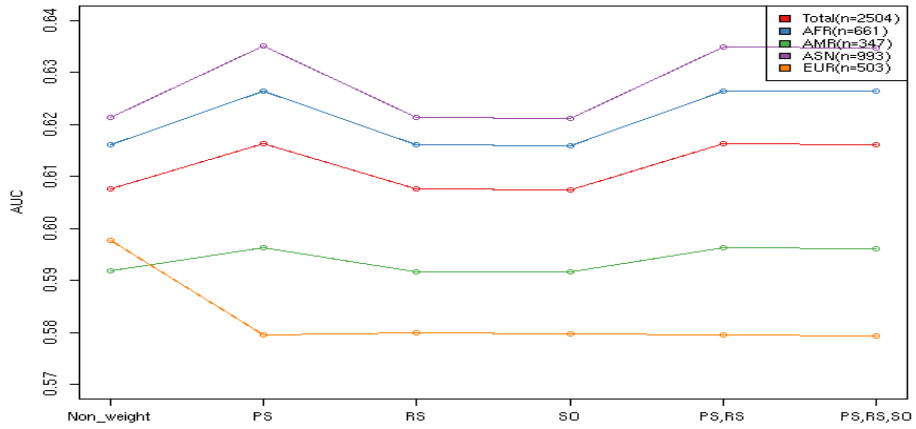

Drug class1
Drug class2
Non class
−0.6 −0.2 0 0.2 0.4 0.6
The deviation from mean AUC
0.1(0.604)
0.2(0.5997)
0.3(0.6067)
0.4(0.5985)
0.5(0.6041)
0.6(0.6061)
0.7(0.6067)
0.8(0.6057)
0.9(0.6073)
Drug class(n=#drug)
[A](n=62)
[B](n=9)
[C](n=122)
[D](n=25)
[G](n=41)
[J](n=7)
[L](n=43)
[N](n=144)
[R](n=64)
[S](n=22)
[A10](n=39)
[C03](n=23)
[C07](n=22)
[C08](n=16)
[C09A](n=13)
[C10] (n=15)
[G03](n=37)
[M02] (n=35)
[N05B,N05C](n=49)
[N06A](n=39)
[R03](n=46)
Non-PGX drug(n=97)
PGX drug(n=400)

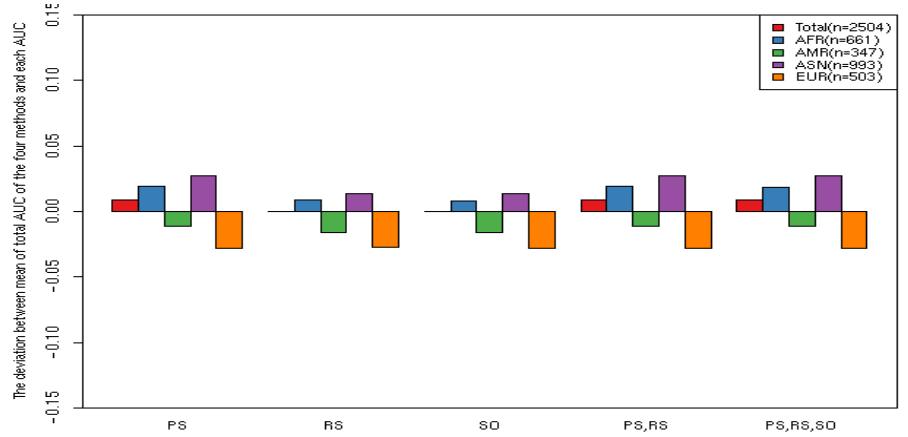
비인종평가


0.1(0.6236)
0.2(0.6186)
0.3(0.6241)
0.4(0.6199)
0.5(0.6231)
0.6(0.6266)
0.7(0.6276)
0.8(0.625)
0.9(0.6269)
Drug class(n=#drug)
[A](n=62)
[B](n=9)
[C](n=122)
[D](n=25)
[G](n=41)
[H](n=10)
[J](n=7)
[L](n=43)
[M](n=10)
[N](n=144)
[R](n=84)
[S](n=22)
[A02BC](n=4)
[A10](n=39)
[C02](n=22)
[C03](n=23)
[C07](n=22)
[C08](n=16)
[C09A](n=13)
[C10] (n=15)
[G03](n=37)
[N02] (n=35)
[N05B,N05C](n=49)
[N06A](n=39)
[R03](n=46)
Non-PGx drug(n=97)
PGx drug(n=400)

F. mRNA 안정성 조절 기전(Nonsense-mediated mRNA decay)

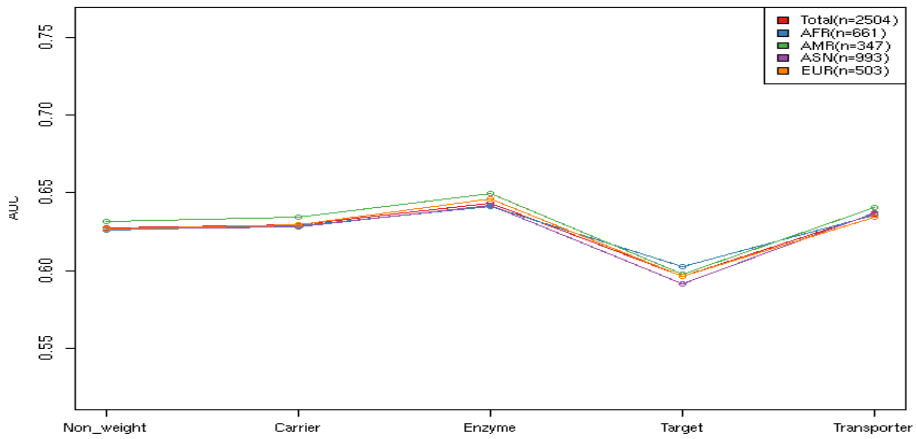
인종평가

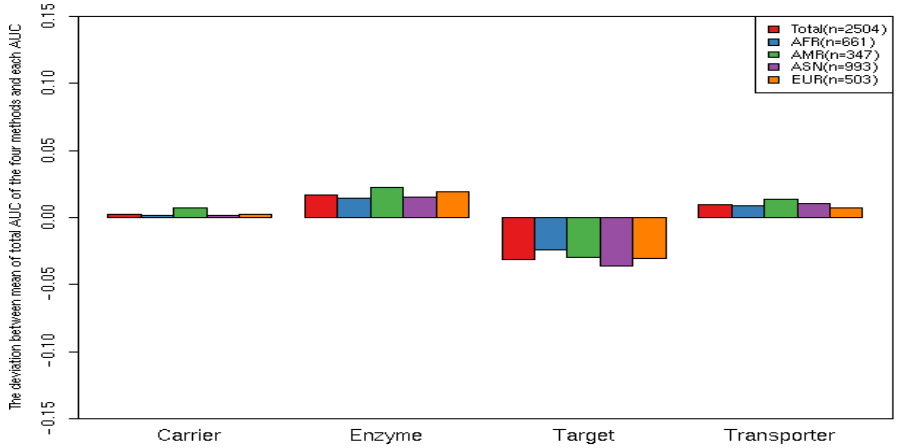




	Non-weight	PS	RS	SO	PS,RS	PS,RS,SO
Total(*n*=2504)	0.6076	0.6163	0.6076	0.6074	0.6163	0.6161
AFR(*n*=661)	0.6161	0.6265	0.6161	0.616	0.6265	0.6264
AMR(*n*=347)	0.5918	0.5964	0.5917	0.5916	0.5963	0.5961
ASN(*n*=993)	0.6214	0.6351	0.6214	0.6212	0.635	0.6348
EUR(*n*=503)	0.5799	0.5795	0.58	0.5798	0.5795	0.5794

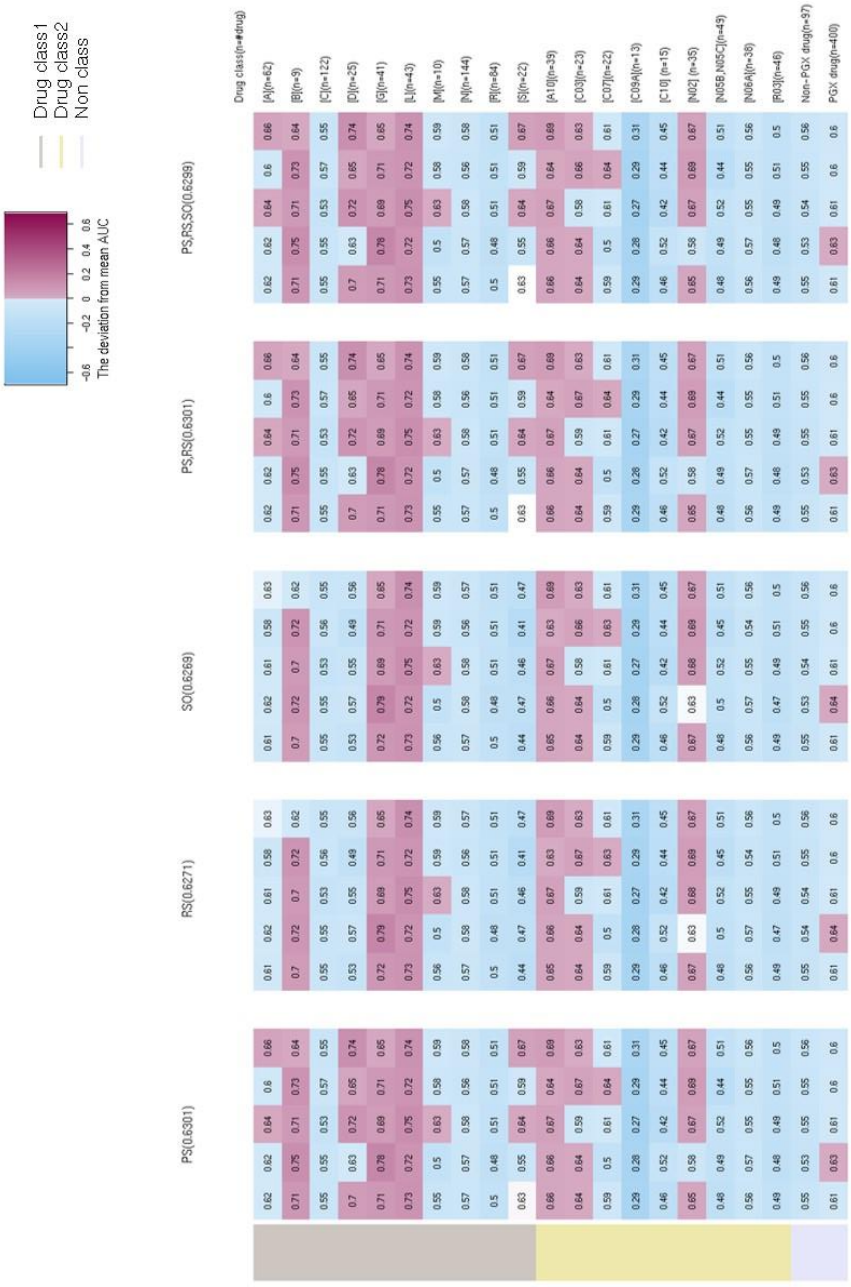
비인종평가



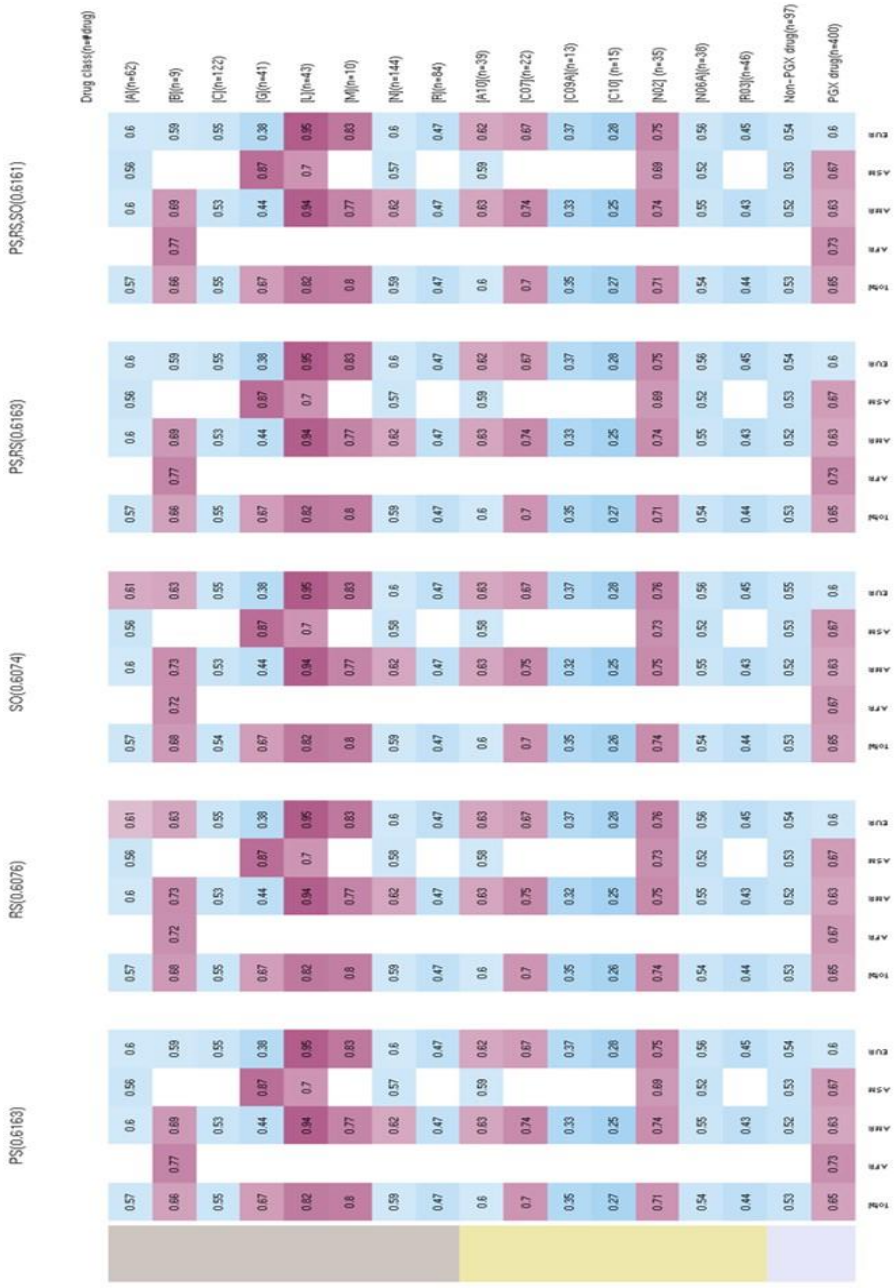


	Non-weight	PS	RS	SO	PS,RS	PS,RS,SO
Total(*n*=2504)	0.6271	0.6301	0.6271	0.6269	0.6301	0.6299
AFR(*n*=661)	0.6261	0.6271	0.6261	0.6259	0.6271	0.627
AMR(*n*=347)	0.6314	0.6329	0.6313	0.6312	0.6328	0.6326
ASN(*n*=993)	0.6266	0.6322	0.6265	0.6263	0.6322	0.6319
EUR(*n*=503)	0.6264	0.6281	0.6265	0.6263	0.6281	0.628

인종평가

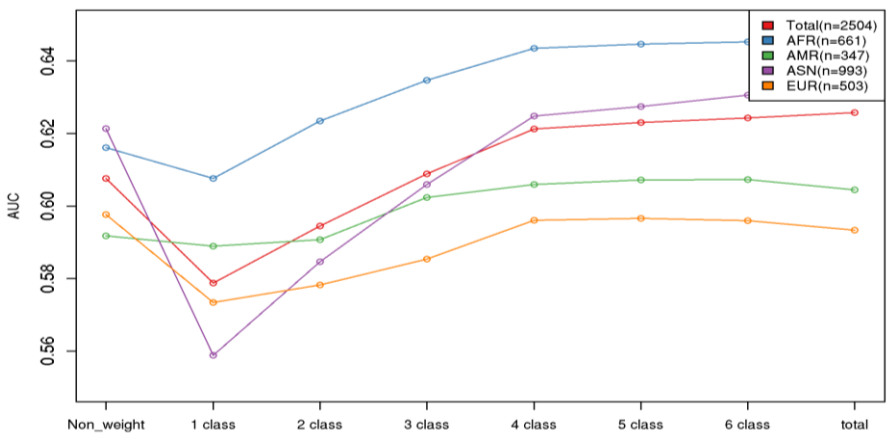

Drug class1
Drug class2
Non class
-0.6 -0.2 0 0.2 0.4 0.6
The deviation from mean AUC
PS(0.6301)
PS(0.6271)
SO(0.6269)
PS,PS(0.6301)
PS,RS,SO(0.6239)
Drug class(n=#drug)
[A](n=62)
[B](n=9)
[C](n=122)
[D](n=25)
[G](n=41)
[L](n=43)
[M](n=10)
[N](n=144)
[R](n=84)
[S](n=22)
[A10](n=39)
[C03](n=23)
[C07](n=22)
[C09A](n=13)
[C10](n=15)
[N02] (n=35)
[N05B,N05C](n=49)
[N06A](n=36)
[R03](n=46)
Non-PGX drug(n=97)
PGX drug(n=100)
TWN AFR AMR ASN EUR

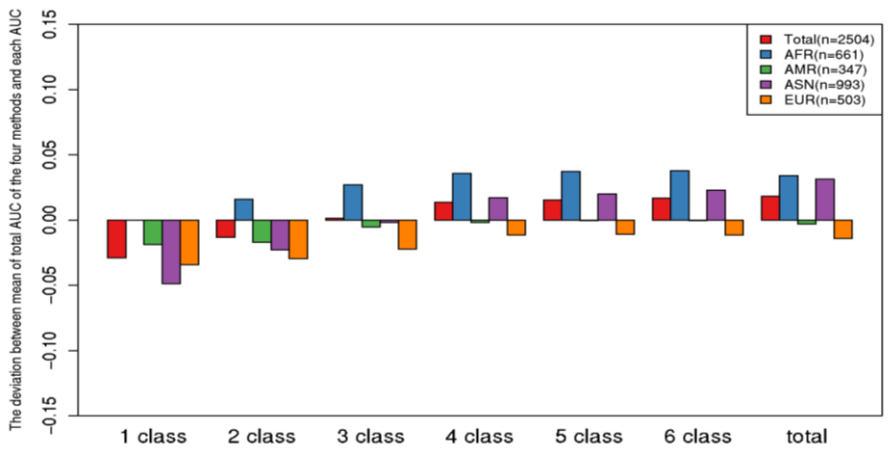
비인종평가


PS;RS;SO(0.6161)
PS;RS(0.6163)
SO(0.6074)
RS(0.6076)
PS(0.6163)
Drug class(n=#drug)
[A](n=62)
[B](n=9)
[C](n=122)
[G](n=41)
[L](n=43)
[M](n=10)
[N](n=144)
[R](n=84)
[A10](n=39)
[C07](n=22)
[C09A](n=13)
[C10] (n=15)
[N02] (n=35)
[N06A](n=38)
[R03](n=46)
Non-PGX drug(n=97)
PGX drug(n=400)
EUR
ASN
AMR
AFR
Total

G. 비번역 변이를 포함한 유전자 기능 조절 변이 (Regulatory variants including noncoding region variants)

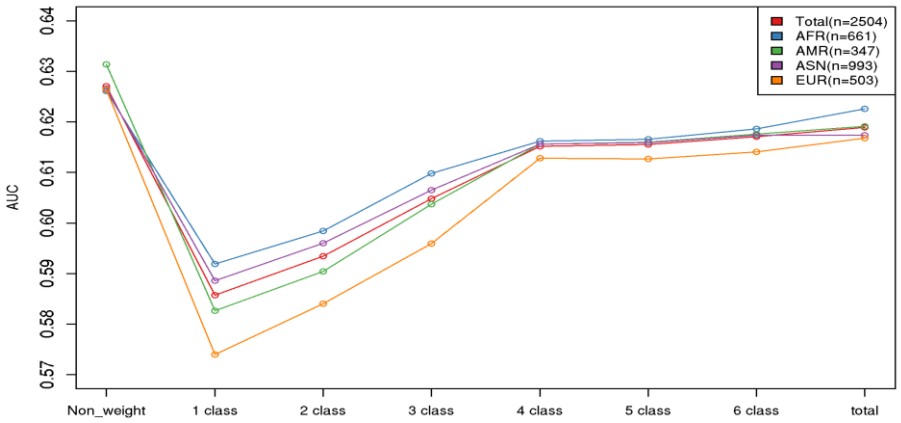
인종평가

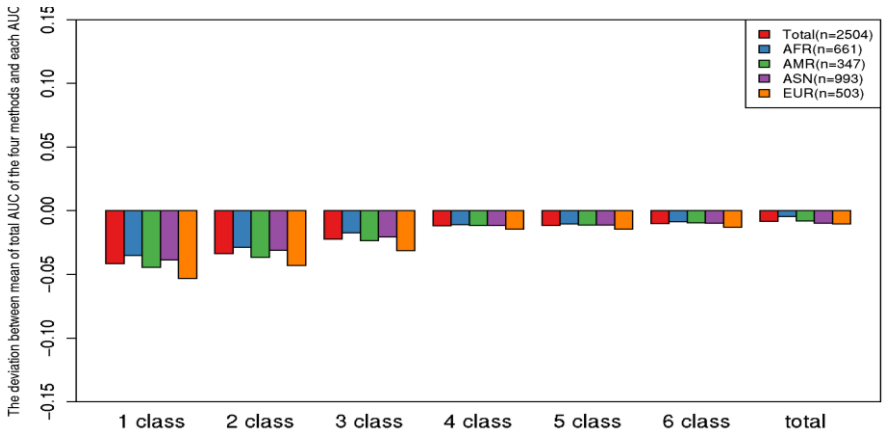




	Non−weight	1 class	2 class	3 class	4 class	5 class	6 class	total
Total(*n*=2504)	0.6076	0.6089	0.623	0.6243	0.6258	0.6213	0.5945	0.5788
AFR(*n*=661)	0.6161	0.6347	0.6447	0.6453	0.6416	0.6435	0.6235	0.6076
AMR(*n*=347)	0.5918	0.6024	0.6072	0.6073	0.6045	0.6059	0.5907	0.589
ASN(*n*=993)	0.6214	0.6059	0.6275	0.6306	0.6391	0.6248	0.5847	0.5588
EUR(*n*=503)	0.5799	0.5854	0.5966	0.596	0.5934	0.5961	0.5782	0.5734

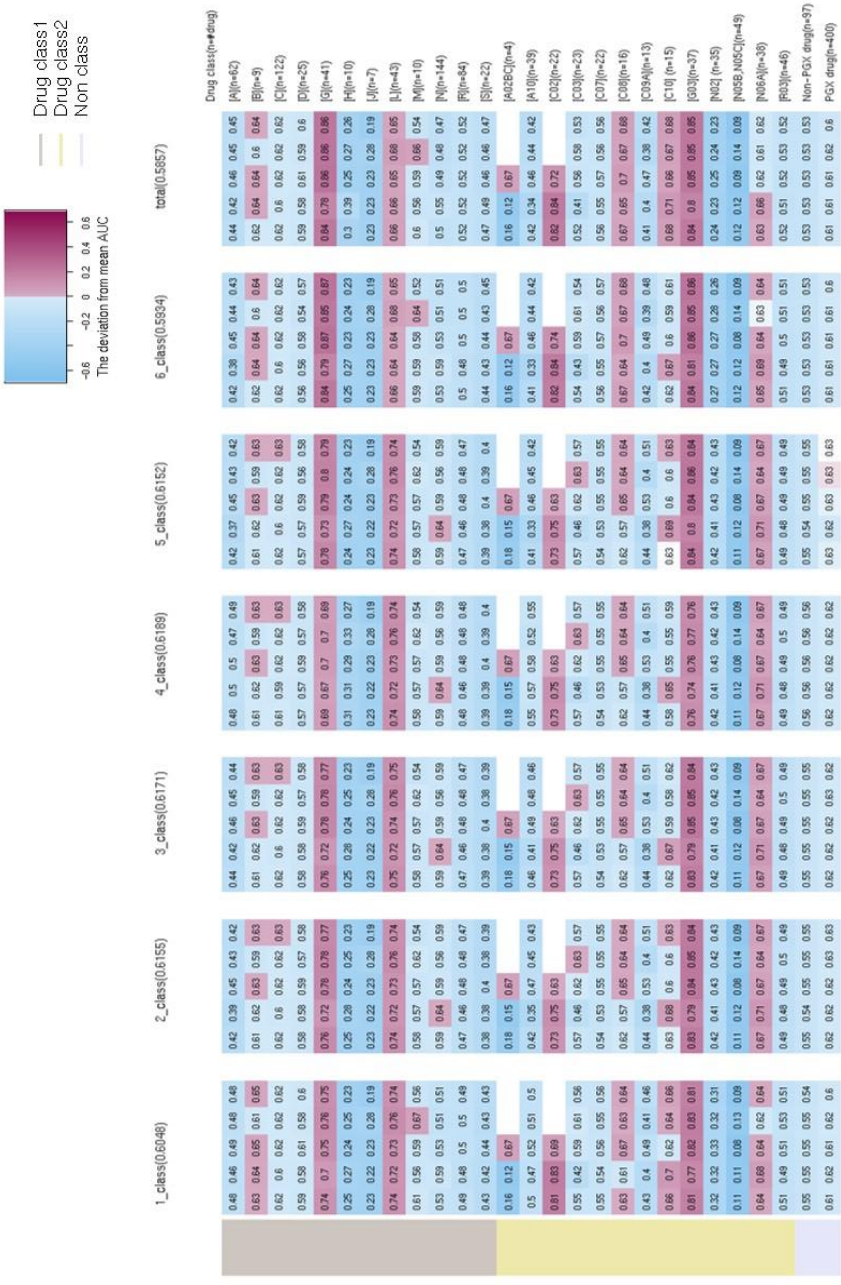
비인종평가



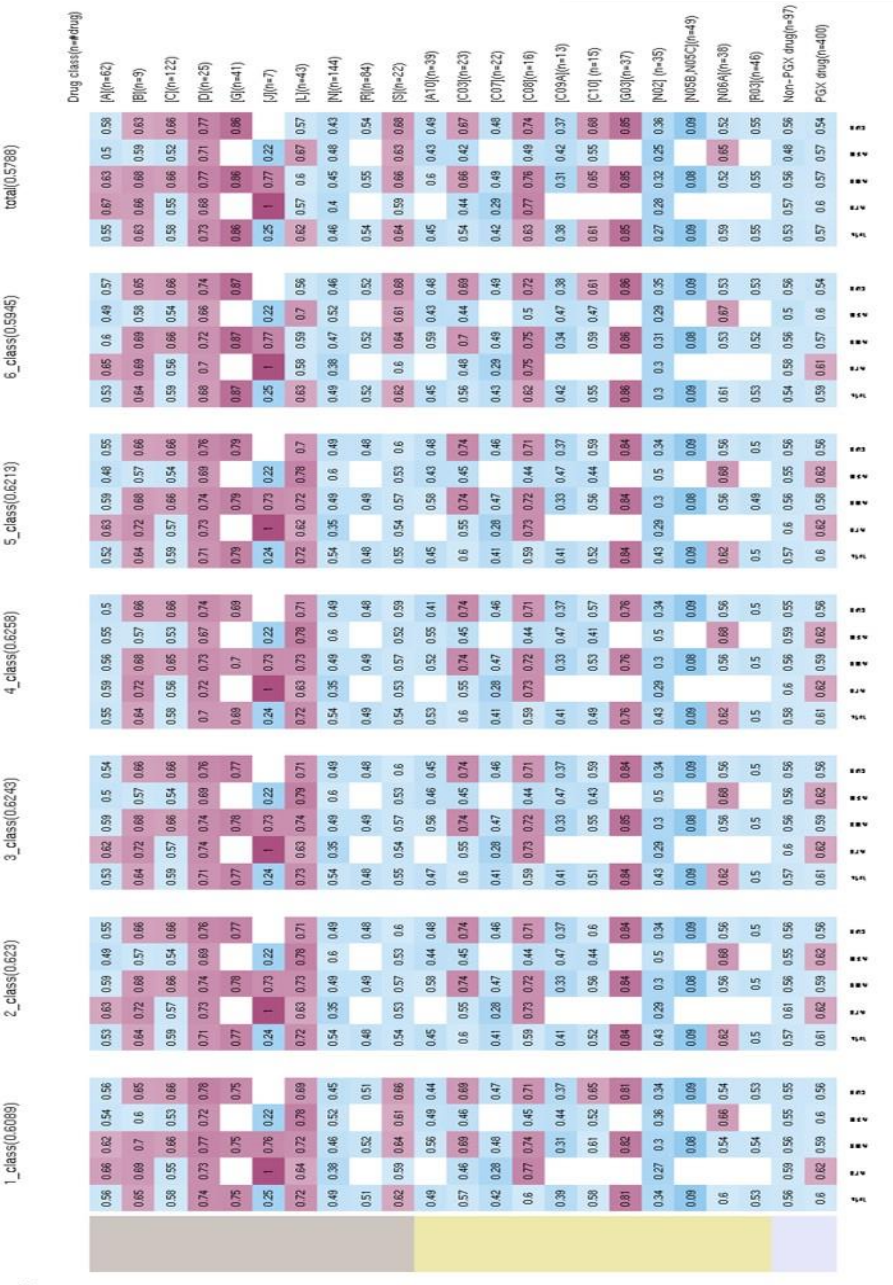


	Non−weight	1 class	2 class	3 class	4 class	5 class	6 class	total
Total(*n*=2504)	0.6271	0.6048	0.6155	0.6171	0.6189	0.6152	0.5934	0.5857
AFR(*n*=661)	0.6261	0.6098	0.6166	0.6186	0.6226	0.6162	0.5984	0.5919
AMR(*n*=347)	0.6314	0.6037	0.616	0.6176	0.6191	0.6156	0.5904	0.5826
ASN(*n*=993)	0.6266	0.6065	0.6159	0.6174	0.6174	0.6156	0.596	0.5886
EUR(*n*=503)	0.6264	0.5959	0.6127	0.6141	0.6168	0.6128	0.584	0.574

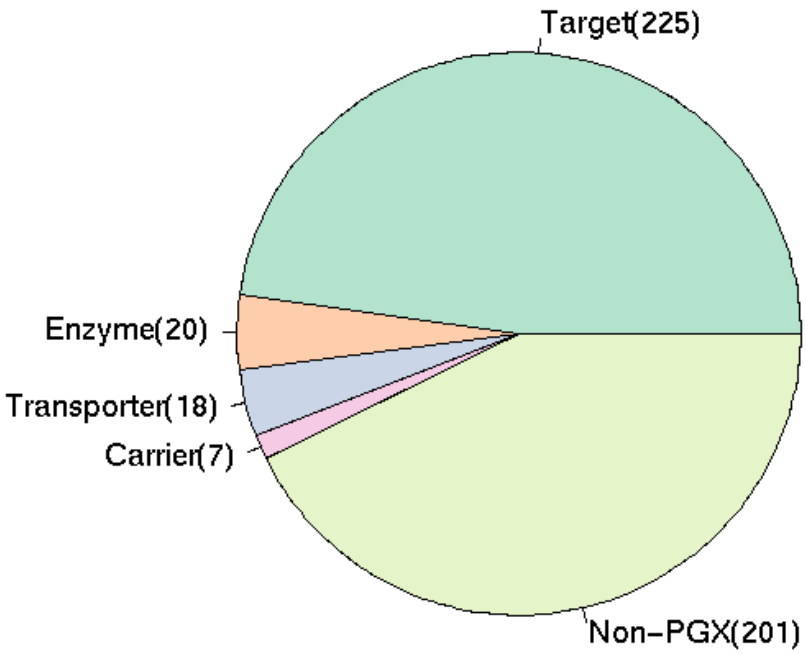
인종평가


Drug class1
Drug class2
Non class
-0.6 -0.2 0 0.2 0.4 0.6
The deviation from mean AUC
1_class(0.6048)
2_class(0.6155)
3_class(0.6171)
4_class(0.6189)
5_class(0.6152)
6_class(0.5934)
total(0.5857)
Drug class(n=#drug)
[A](n=62)
[B](n=9)
[C](n=122)
[D](n=25)
[G](n=41)
[H](n=10)
[J](n=7)
[L](n=43)
[M](n=10)
[N](n=144)
[R](n=84)
[S](n=22)
[A02BC](n=4)
[A10](n=39)
[C02](n=22)
[C03](n=23)
[C07](n=22)
[C09](n=16)
[C09A](n=13)
[C10] (n=15)
[G03](n=37)
[N02] (n=35)
[N05B,N05C](n=49)
[N06A](n=36)
[R03](n=46)
Non-PGX drug(n=97)
PGX drug(n=400)

비인종평가


Drug class(n=#drug)
[A](n=62)
[B](n=9)
[C](n=122)
[D](n=25)
[G](n=41)
[J](n=7)
[L](n=43)
[N](n=144)
[R](n=84)
[S](n=22)
[A10](n=39)
[C03](n=23)
[C07](n=22)
[C08](n=16)
[C09A](n=13)
[C10] (n=15)
[G03](n=37)
[N02] (n=35)
[N05B, N05C](n=49)
[N06A](n=36)
[R03](n=46)
Non-PGX drug(n=97)
PGX drug(n=400)
total(0.5788)
6_class(0.5945)
5_class(0.6213)
4_class(0.6258)
3_class(0.6243)
2_class(0.623)
1_class(0.6089)

보충 그림 6. PharmGKB로부터 추출한 1807개의 변이-약물 연관관계에 포함된 변이가 속한 유전자 471개의 약물학적 유전자 카테고리별 분포.



보충 표

보충 표 1. 2014 1000 지놈 데이터의 인종별 개인유전체 갯수

Sub population	Super population	No. of personal genome data
GWD	AFR	113
YRI	AFR	108
ESN	AFR	99
LWK	AFR	99
ACB	AFR	96
MSL	AFR	85
ASW	AFR	61
PUR	AMR	104
CLM	AMR	94
PEL	AMR	85
MXL	AMR	64
CHS	ASN	105
JPT	ASN	104
CHB	ASN	103
GIH	ASN	103
ITU	ASN	102
STU	ASN	102
KHV	ASN	99
PJL	ASN	96

CDX	ASN	93
BEB	ASN	86
IBS	EUR	107
TSI	EUR	107
CEU	EUR	99
FIN	EUR	99
GBR	EUR	91
	Total	
	2504	

26 2014 1000 지놈 데이터의 하위 26개, 상위 4개 인종군 정보. (2015 6). 상위 인종군 : AFR(African), EUR(European),ASN(EAS;East Asian),ASN(SAS;South Asian), AMR(Ad Mixed American). 하위 인종군 : AFR(YRI(Yoruba in Ibadan, Nigeria),LWK(Luhya in Webuye, Kenya),GWD(Gambian in Western Divisions in the Gambia),MSL(Mende in Sierra Leone),ESN(Esan in Nigeria),ASW(Americans of African Ancestry in SW USA),ACB(African Caribbeans in Barbados)),EUR(CEU(Utah Residents (CEPH) with Northern andWestern European Ancestry),TSI(Toscani in Italia), FIN(Finnish in Finland),GBR(British in England and Scotland),IBS(Iberian Population in Spain)), ASN(EAS,SAS;CHB(Han Chinese in Bejing, China),JPT(apanese in Tokyo, Japan),CHS(Southern Han Chinese),CDX(Chinese Dai in Xishuangbanna, China),KHV(Kinh in Ho Chi Minh City, Vietnam), GIH(Gujarati Indian from Houston, Texas),PJL(Punjabi from Lahore, Pakistan),BEB(Bengali from Bangladesh), STU(Sri Lankan Tamil from the UK), ITU(Indian Telugu from the UK)),AMR(MXL(Mexican Ancestry from Los Angeles USA),PUR(Puerto Ricans from Puerto Rico),CLM(Colombians from Medellin, Colombia),PEL(Peruvians from Lima, Peru)).

보충 표 2. ATC 약물 분류군별 PharmGKB로 부터 추출한 약물조절변이 갯수

Drug class	ATC code	No. of drugs	PharmGKB*
Alimentary tract and metabolism	A	62	77
Blood and blood forming organs	B	9	99
Cardiovascular system	C	122	291
Dermatologicals	D	25	56
Genito urinary system and sex hormones	G	41	9
Systemic hormonal preparations, excl. sex hormones and insulins	H	10	11
antiinfectives for systemic use	J	7	5
Antineoplastic and immunomo- dulating agents	L	43	294
Musculo-skeletal system	M	10	7
Nervous system	N	144	504
Antiparasitic products, insecticides and repellents	P	2	0
Respiratory system	R	84	58
Sensory organs	S	22	63
Various	V	3	1
Sub-total		493	1475
Others(unclassified)		4	205

* Number of existing genetic-variant-drug associations in PharmGKB 2015 version(2015, 1)

보충 표 3. ATC 자주 처방받은 약물 분류군별 PharmGKB로 부터 추출한 약물조절변이 갯수

Drug class	ATC code	No. of drugs	PharmGKB*
Proton pump inhibitors	A02BC	4	12
Drugs used in diabetes	A10	39	26
Antihypertensives	C02	22	6
Diuretics	C03	23	33
Beta blocking agents	C07	22	50
Calcium channel blockers	C08	16	28
ACE inhibitors, plain	C09A	13	35
Lipid modifying agents	C10	15	125
Sex hormones and modulators of the genital system	G03	37	9
Thyroid therapy	H03	5	1
Analgesics	N02	35	131
Anxiolytics and hypnotics and sedatives	N05B/ N05	49	1
Antidepressants	N06A	38	265
Drugs for obstructive airway diseases	R03	46	40
Antihistamines for systemic use	R06	35	1
Sub-total		395	763
Others(unclassified)		102	917

* Number of existing genetic-variant-drug associations in PharmGKB 2015 version(2015, 1)

보충 표 4. Pharmsafe & 가중치 Pharmsafe 알고리즘에 쓰인 기호.

Symbol	Definition
v_i	Variant i
g_j	Gene j
S_{v_i}	SIFT score of variant i
S_{g_j}	Damaged score of gene(gene score) j
S_{d_k}	Damaged score of drug(drug score) k
G_j	Set of variant with SIFT score in gene j
D_k	Set of gene related with drug k
WS_{v_i}	Winsorized variant score of variant i
W_{v_i}	Weight score of variant i
W_{g_j}	Weight score of gene j
V_{PGT}	Set of variants in PharmacoGene j region
SR	SIFT score filter Range
MAF_{v_i}	Minor Allele Frequency of variant i in 1000 genome (n=2504)
$MAFR$	Minor Allele Frequency Range in in 1000 genome(n=2504)
V_{MAFR}	Set of variants with SIFT score in $MAFR_l$
HVF_{v_i}	Homozygote Variant Frequency of variant i in 1000 genome (n=2504)

$HVFR$	Homozygote Variant Frequency Range in 1000 genome(m=2504)
HV_{HVFR}	Set of Homozygote Variants with SIFT score in $HVFR_l$
V_{NMD}	Set of NMD variants with SIFT score
RG_j	Gene j is regulated by *variant i*
RS_{v_i}	Regulation score of variant *i*
RV	Variants with the ability to regulated the gene
RC	Regulation variant Classs
RV_i	Regulation variant *i*
WS_{RG_j}	Weight score of regulated gene *j*
$EAD_{E_{c\ l}}$	Ethic testing AUC diviation of condition *l* in E element
$NAD_{E_{c\ l}}$	Non-Ethic testing AUC diviation of condition *l* in E element
SEA	Standard Ethnic AUC
SNA	Standard Non-ethnic AUC
$EA_{E_{cl}}$	Ethnic AUC of condition *l* in element e
$NA_{E_{cl}}$	Non-ethnic AUC of condition *l* in element e
$CB_{m_{v_i}}$	Variant in *i* Conbination m

* pharmacogene type consist of target,transporter,enzyme,carrier.

보충 표 5. PharmGKB, 497 약물, 1000 지놈에 공통적으로 속하는 유전적 변이-약물 연관 갯수

Element	Pharm GKB*	497 drugs (%)+	1000 Genome ∩ 497 drugs(%)+
Genetic-variant-drug associations	3248	1807(55.63)	1100(33.86)
Black or African American associations	83	58(69.87)	40(48.19)
Asian associations	515	329(63.88)	224(43.49)
White associations	647	451(69.70)	303(46.83)
Hispanic or Latino associations	6	5(83.33)	3(50)
Others(unclassified)	1997	1128(56.48)	652(32.64)
Drugs	391	290(74.18)	251(64.19)
Variants	1175	840(71.48)	522(44.42)

* Number of existing genetic-variant-drug associations in PharmGKB 2015 version(2015, 1)

+ Percentage of genetic-variant-drug associations in 497 drugs or 497 drugs and 1000 Genome from ParmGKB

보충 표 6. 7가지 요소에 속하는 54가지 조건별 인종/비인종 평가 지수.

Weight parameter	Element of each parameter	Ethnic AUC	Non-ethnic AUC
Central Tendency method	Arithmetic mean	0.5633	0.5935
	Geometric mean	0.6076	0.6271
	Harmonic mean	0.6149	0.6363
	product	0.6163	0.6502
Pharmacogene Type	Target	0.5961	0.5928
	Enzyme	0.6436	0.6248
	Transporter	0.6367	0.6134
	Carrier	0.6296	0.6074
Score Winsorization	0.1	0.6074	0.6219
	0.2	0.6166	0.6364
	0.3	0.6167	0.6322
	0.4	0.6157	0.6291
	0.5	0.6092	0.6259
	0.6	0.6104	0.6273
	0.7	0.61	0.6277
	0.8	0.6083	0.6271
	0.9	0.6078	0.6273
Minor Allele Frequency	0.001	0.6076	0.6271
	0.002	0.6077	0.627
	0.003	0.6075	0.6271
	0.004	0.6075	0.6271
	0.005	0.6077	0.6272

	0.006	0.6076	0.6271
	0.007	0.6078	0.6271
	0.008	0.6075	0.6271
	0.009	0.6078	0.6271
	0.01 over	0.6008	0.6212
Homozygote mutation rate	0.1	0.604	0.6236
	0.2	0.5997	0.6186
	0.3	0.6067	0.6241
	0.4	0.5995	0.6199
	0.5	0.6041	0.6231
	0.6	0.6061	0.6266
	0.7	0.6067	0.6276
	0.8	0.6057	0.625
	0.9	0.6073	0.6269
Nonsense-mediated mRNA decay	Premature Stop codons	0.6163	0.6301
	Removed Stop codons	0.6076	0.6271
	Splice-Overlap	0.6074	0.6269
	Premature Stop codons, Removed Stop codons	0.6163	0.6301
	Premature Stop codons, Removed Stop codons, Splice-Overlap	0.6161	0.6299
Regulatory variants	1	0.6089	0.6048

including noncoding	2	0.623	0.6155
region variants	3	0.6243	0.6171
	4	0.6258	0.6189
	5	0.6213	0.6152
	6	0.5945	0.5934
	Total	0.5788	0.5857

보충 표 7. 각 요소의 조건별 인종, 비인종 AUC 및 편차 및 편차 평균.

SW : 변이 점수 원저화, MAF : 낮은 대립형질 빈도 , PGT : 약물학적 유전자 종류, HR : 동형접합변이 빈도,NMD : 7 mRNA 안정성 조절 기전, RC : 유전자 기능 조절 변이

요소	조건	인종평가 전체 AUC	비인종평가 전체 AUC	인종평가 편차	비인종평가 편차	평균편차
SW	0.1	0.6074	0.6219	−0.0002	−0.0052	−0.0027
	0.2	0.6166	0.6364	0.009	0.0093	0.00915
	0.3	0.6167	0.6322	0.0091	0.0051	0.0071
	0.4	0.6157	0.6291	0.0081	0.002	0.00505
	0.5	0.6092	0.6259	0.0016	−0.0012	0.0002
	0.6	0.6104	0.6273	0.0028	0.0002	0.0015
	0.7	0.61	0.6277	0.0024	0.0006	0.0015
	0.8	0.6083	0.6271	0.0007	0	0.00035
	0.9	0.6078	0.6273	0.0002	0.0002	0.0002
MAF	0.001	0.6076	0.6271	0	0	0
	0.002	0.6077	0.627	1E−04	−1E−04	0
	0.003	0.6075	0.6271	−1E−04	0	−5E−05
	0.004	0.6075	0.6271	−1E−04	0	−5E−05
	0.005	0.6077	0.6272	1E−04	1E−04	1E−04
	0.006	0.6076	0.6271	0	0	0
	0.007	0.6078	0.6271	0.0002	0	1E−04
	0.008	0.6075	0.6271	−1E−04	0	−5E−05
	0.009	0.6078	0.6271	0.0002	0	1E−04
	0.01 ≤MAF	0.6008	0.6212	−0.0068	−0.0059	−0.00635
PGT	표적	0.5961	0.5928	−0.0115	−0.0343	−0.0229
	효소	0.6436	0.6248	0.036	−0.0023	0.01685

	수송체	0.6367	0.6134	0.0291	−0.0137	0.0077
	수송기구	0.6296	0.6074	0.022	−0.0197	0.00115
요소	조건	인종평가 전체 AUC	비인종평가 전체 AUC	인종평가 편차	비인종평가 편차	평균편차
HR	0.1	0.604	0.6236	−0.0036	−0.0035	−0.00355
	0.2	0.5997	0.6186	−0.0079	−0.0085	−0.0082
	0.3	0.6067	0.6241	−0.0009	−0.003	−0.00195
	0.4	0.5995	0.6199	−0.0081	−0.0072	−0.00765
	0.5	0.6041	0.6231	−0.0035	−0.004	−0.00375
	0.6	0.6061	0.6266	−0.0015	−0.0005	−0.001
	0.7	0.6067	0.6276	−0.0009	0.0005	−0.0002
	0.8	0.6057	0.625	−0.0019	−0.0021	−0.002
	0.9	0.6073	0.6269	−0.0003	−0.0002	−0.00025
NMD	PS	0.6163	0.6301	0.0087	0.003	0.00585
	RS	0.6076	0.6271	0	0	0
	SO	0.6074	0.6269	−0.0002	−0.0002	−0.0002
	PS,RS	0.6163	0.6301	0.0087	0.003	0.00585
	PS,RS,SO	0.6161	0.6299	0.0085	0.0028	0.00565
RC	1	0.6089	0.6048	0.0013	−0.0223	−0.0105
	2	0.623	0.6155	0.0154	−0.0116	0.0019
	3	0.6243	0.6171	0.0167	−0.01	0.00335
	4	0.6258	0.6189	0.0182	−0.0082	0.005
	5	0.6213	0.6152	0.0137	−0.0119	0.0009
	6	0.5945	0.5934	−0.0131	−0.0337	−0.0234
	Total	0.5788	0.5857	−0.0288	−0.0414	−0.0351

보충 표 8. 각 요소의 선택 된 6가지 조건의 56가지 조합과 Pharmsafe 알고리즘을 사용해 계산한 개인별 약물 점수를 평가한 인종, 비인종 AUC. (A)인종평가 , (B) 비인종평가 이다. 각 약자는 다음과 같다. SW : 변이 점수 윈저화, MAF : 낮은 대립형질 빈도 , PGT : 약물학적 유전자 종류, HR : 동형접합변이 빈도,NMD : 7 mRNA 안정성 조절 기전, RC : 유전자 기능 조절 변이.

A. 인종평가

순번	조합 종류	전체	아프리카	미국	아시아	유럽
0	SF,MAF	0.614	0.632	0.5961	0.6243	0.5825
1	SF,HR	0.613	0.6293	0.5954	0.6241	0.5818
2	SF,PGX	0.6235	0.6443	0.6042	0.6323	0.5923
3	SF,LOF	0.62	0.6401	0.5987	0.6334	0.5819
4	SF,NC	0.5203	0.4703	0.5549	0.5156	0.5706
5	MAF,HR	0.6047	0.63	0.5865	0.6098	0.5738
6	MAF,PGX	0.6162	0.6431	0.5974	0.6203	0.5858
7	MAF,LOF	0.6117	0.6409	0.5898	0.6192	0.5736
8	MAF,NC	0.5722	0.4464	0.669	0.5891	0.6356
9	HR,PGX	0.641	0.5969	0.6202	0.5853	0.5852
10	HR,LOF	0.6109	0.6385	0.5892	0.6194	0.573
11	HR,NC	0.5722	0.4465	0.6691	0.5892	0.6356
12	PGX,LOF	0.6213	0.6508	0.5994	0.628	0.5847
13	PGX,NC	0.5743	0.4478	0.671	0.593	0.6355
14	LOF,NC	0.5722	0.4466	0.669	0.5891	0.6358
15	SF,MAF,HR	0.6074	0.6353	0.5891	0.6113	0.5757
16	SF,MAF,PGX	0.6156	0.6443	0.5968	0.6188	0.5846
17	SF,MAF,LOF	0.6131	0.6448	0.5916	0.6186	0.5756

18	SF,MAF,NC	0.5709	0.4462	0.6681	0.5861	0.6362
19	SF,HR,PGX	0.6149	0.6425	0.5963	0.6186	0.5841
20	SF,HR,LOF	0.6124	0.6428	0.591	0.6185	0.575
21	SF,HR,NC	0.571	0.4462	0.6682	0.5862	0.6362
22	SF,PGX,LOF	0.6198	0.6511	0.5983	0.6248	0.5838
23	SF,PGX,NC	0.5721	0.4472	0.6693	0.5886	0.6353
24	SF,LOF,NC	0.571	0.4464	0.6681	0.5862	0.6364
25	MAF,HR,PGX	0.6089	0.6408	0.5909	0.6092	0.579
26	MAF,HR,LOF	0.6059	0.6424	0.5845	0.6078	0.5692
27	MAF,HR,NC	0.5712	0.4442	0.6684	0.5889	0.6349
28	MAF,PGX,LOF	0.6138	0.6495	0.5927	0.6154	0.5785
29	MAF,PGX,NC	0.5726	0.445	0.6695	0.5912	0.635
30	MAF,LOF,NC	0.5713	0.4443	0.6682	0.5888	0.635
31	HR,PGX,LOF	0.6133	0.6479	0.5923	0.6153	0.5781
32	HR,PGX,NC	0.5726	0.445	0.6696	0.5912	0.635
33	HR,LOF,NC	0.5713	0.4444	0.6683	0.5888	0.635
34	PGX,LOF,NC	0.5726	0.4451	0.6694	0.5911	0.6352
35	SF,MAF,HR,PGX	0.6087	0.6408	0.5906	0.6091	0.5782
36	SF,MAF,HR,LOF	0.6068	0.6432	0.5859	0.6081	0.5705
37	SF,MAF,HR,NC	0.5701	0.444	0.6676	0.5862	0.6353
38	SF,MAF,PGX,LOF	0.613	0.6488	0.5921	0.6143	0.5779
39	SF,MAF,PGX,NC	0.5706	0.4445	0.6676	0.588	0.6335
40	SF,MAF,LOF,NC	0.5701	0.4441	0.6675	0.5861	0.6355
41	SF,HR,PGX,LOF	0.6125	0.6473	0.5917	0.6142	0.5775
42	SF,HR,PGX,NC	0.5706	0.4445	0.6677	0.588	0.6335
43	SF,HR,LOF,NC	0.5701	0.4441	0.6675	0.5861	0.6355
44	SF,PGX,LOF,NC	0.5706	0.4447	0.6676	0.5879	0.6337

45	MAF,HR,PGX,LOF	0.6074	0.6454	0.5872	0.6064	0.5733
46	MAF,HR,PGX,NC	0.5706	0.4442	0.6677	0.5893	0.6316
47	MAF,HR,LOF,NC	0.5702	0.4435	0.6669	0.5885	0.6325
48	MAF,PGX,LOF,NC	0.5706	0.4443	0.6675	0.5892	0.6318
49	HR,PGX,LOF,NC	0.5707	0.4443	0.6676	0.5892	0.6318
50	SF,MAF,HR,PGX,LOF	0.6071	0.6446	0.587	0.6064	0.573
51	SF,MAF,HR,PGX,NC	0.5694	0.4439	0.6665	0.5871	0.6308
52	SF,MAF,HR,LOF,NC	0.569	0.4435	0.6659	0.5858	0.6325
53	SF,MAF,PGX,LOF,NC	0.5694	0.4441	0.6664	0.587	0.6309
54	SF,HR,PGX,LOF,NC	0.5694	0.4441	0.6664	0.587	0.6309
55	MAF,HR,PGX,LOF,NC	0.5687	0.4444	0.666	0.5858	0.6298
56	SF,MAF,HR,PGX,LOF,NC	0.568	0.4439	0.6652	0.5849	0.6292

B. 비인종평가

순번	조합 종류	전체	아프리카	미국	아시아	유럽
0	SF,MAF	0.6321	0.6308	0.6367	0.6317	0.6315
1	SF,HR	0.6322	0.6307	0.6365	0.6321	0.6313
2	SF,PGX	0.6426	0.6406	0.6477	0.6414	0.644
3	SF,LOF	0.6341	0.6315	0.6373	0.6357	0.6322
4	SF,NC	0.5356	0.5323	0.5391	0.5368	0.5353
5	MAF,HR	0.6228	0.6252	0.6265	0.6205	0.6217
6	MAF,PGX	0.635	0.6362	0.6398	0.6321	0.636
7	MAF,LOF	0.6247	0.626	0.6271	0.6242	0.6222
8	MAF,NC	0.5951	0.582	0.5978	0.6049	0.591
9	HR,PGX	0.636	0.6396	0.6324	0.6358	0.6155
10	HR,LOF	0.6249	0.626	0.6271	0.6248	0.6224

11	HR,NC	0.5951	0.582	0.5979	0.6049	0.591
12	PGX,LOF	0.6365	0.6365	0.64	0.6354	0.636
13	PGX,NC	0.5972	0.5844	0.6001	0.6071	0.5928
14	LOF,NC	0.5951	0.5821	0.5978	0.605	0.591
15	SF,MAF,HR	0.6262	0.6278	0.6297	0.6244	0.625
16	SF,MAF,PGX	0.6351	0.636	0.6392	0.6326	0.6357
17	SF,MAF,LOF	0.6276	0.6285	0.6302	0.6273	0.6254
18	SF,MAF,NC	0.5942	0.5812	0.597	0.6039	0.5904
19	SF,HR,PGX	0.635	0.6358	0.639	0.6328	0.6354
20	SF,HR,LOF	0.6277	0.6284	0.63	0.6276	0.6253
21	SF,HR,NC	0.5943	0.5813	0.5971	0.6039	0.5904
22	SF,PGX,LOF	0.6362	0.6364	0.6393	0.6352	0.6357
23	SF,PGX,NC	0.5955	0.5826	0.5985	0.605	0.5914
24	SF,LOF,NC	0.5943	0.5813	0.597	0.604	0.5904
25	MAF,HR,PGX	0.6288	0.6318	0.6326	0.6254	0.629
26	MAF,HR,LOF	0.6207	0.6239	0.6226	0.6192	0.6181
27	MAF,HR,NC	0.5945	0.5804	0.5975	0.6047	0.5907
28	MAF,PGX,LOF	0.63	0.6324	0.6328	0.6279	0.6291
29	MAF,PGX,NC	0.5957	0.5813	0.5986	0.6062	0.5916
30	MAF,LOF,NC	0.5945	0.5804	0.5974	0.6048	0.5907
31	HR,PGX,LOF	0.63	0.6322	0.6327	0.6281	0.629
32	HR,PGX,NC	0.5957	0.5813	0.5986	0.6062	0.5916
33	HR,LOF,NC	0.5945	0.5805	0.5974	0.6048	0.5907
34	PGX,LOF,NC	0.5957	0.5814	0.5985	0.6063	0.5917
35	SF,MAF,HR,PGX	0.6292	0.6318	0.6326	0.6263	0.6292
36	SF,MAF,HR,LOF	0.623	0.6257	0.6249	0.6218	0.6205
37	SF,MAF,HR,NC	0.5936	0.5794	0.5967	0.6038	0.59

38	SF,MAF,PGX,LOF	0.6302	0.6324	0.6327	0.6284	0.6292
39	SF,MAF,PGX,NC	0.5942	0.5802	0.5974	0.6044	0.5904
40	SF,MAF,LOF,NC	0.5936	0.5794	0.5966	0.6039	0.59
41	SF,HR,PGX,LOF	0.6302	0.6322	0.6326	0.6286	0.6291
42	SF,HR,PGX,NC	0.5942	0.5802	0.5974	0.6044	0.5904
43	SF,HR,LOF,NC	0.5937	0.5794	0.5967	0.6039	0.59
44	SF,PGX,LOF,NC	0.5943	0.5802	0.5973	0.6045	0.5904
45	MAF,HR,PGX,LOF	0.625	0.6286	0.6273	0.6225	0.6237
46	MAF,HR,PGX,NC	0.5938	0.5792	0.5972	0.6049	0.5888
47	MAF,HR,LOF,NC	0.5933	0.5785	0.5965	0.6045	0.5886
48	MAF,PGX,LOF,NC	0.5939	0.5792	0.5972	0.605	0.5888
49	HR,PGX,LOF,NC	0.5939	0.5793	0.5972	0.605	0.5888
50	SF,MAF,HR,PGX,LOF	0.6255	0.6289	0.6276	0.6233	0.624
51	SF,MAF,HR,PGX,NC	0.5928	0.5783	0.5963	0.6038	0.5878
52	SF,MAF,HR,LOF,NC	0.5925	0.5776	0.5957	0.6036	0.5878
53	SF,MAF,PGX,LOF,NC	0.5929	0.5784	0.5962	0.6039	0.5879
54	SF,HR,PGX,LOF,NC	0.5929	0.5784	0.5962	0.6039	0.5879
55	MAF,HR,PGX,LOF,NC	0.5926	0.5784	0.5959	0.6038	0.587
56	SF,MAF,HR,PGX,LOF,NC	0.5921	0.5777	0.5957	0.6034	0.5866

2 장 RNA 편집 위치 검출을 위한 RNA 서열 비교 및 생물학적 주석처리 도구 개발

소 개

RNA 편집(RNA editing)은 RNA 전사 후에 한개의 뉴클레오타이드(nucleotide) 시퀀스(sequence)가 변형되는 현상을 의미하며 이는 mRNA, 비번역 RNA(ncRNA), 마이크로 RNA(miRNA) 등을 포함한 모든 전사체에서 일어난다[67]. 포유 동물에서 가장 많이 발생하는 RNA 편집유형은 A 에서 I 로 바뀌는 편집이며 이는 주로 ADAR 이라는 효소(enzyme) 에 의해 발생한다[68-70]. 최근 RNA 서열을 NGS 기법을 이용하여 초고속으로 분석하는 RNA-seq 기술이 발달함에 따라 G-A,C-U,T-C,C-A,G-C,T-A,A-등의 새로운 RNA 편집형태도 인간의 세포주에서 발견되었다[71, 72]. 대부분의 RNA 편집 혹은 RNA-DNA 서열변형 위치(RDD site)는 인트론(intron), 5 ‘번역 영역 (5’ UTR), 3 ‘ 번역 영역 (3’ UTR)(7) 그리고 Alu 서열에서 발생한다. 그렇지만 RNA 편집이 번역영역(coding region)에서 발생하게 되면 이로인해 nonsynonymous 단백질 변형, 선택적 스플라이싱(Alternative splicing), 유전자 발현변화 등을 일으켜 원래 단백질의 기능을 상실 시키거나 변형시킬 수 있다. (7-9) 또한 RNA 편집은 마이크로 RNA, 짧은 간섭 RNA(small interfering RNA, siRNA), piRNA(Piwi-interacting RNA)등의 기능에도 영향을 미칠 수 있다. 많은 RNA 편집들은 간질, 뇌허열, 우울증, 뇌종양등 다양한 인간 질병에 영향을 미친다고 보고되었다[67]. 지난 몇년동안 엄청난 양의 DNA, RNA-seq 데이터 들이 생산되고 GEO(www.ncbi.nlm.nih.gov/geo/), ENCODE (http://genome.ucsc.edu/ENCODE/) 등의 공공 저장소에 저장되고 공개되었다. 이러한 데이터들을 기반으로, 많은 RNA 편집과 RDD site 을 검출하기 위한 많은 도구들이 개발되었다. 알려지지 않은 RNA

편집 위치를 찾아내는 대표적은 도구로는 rddChecker (http://genomics.jhu.edu/software/rddChecker/) 가 있으며 이런 도구들은 동일한 시료에서 DNA, RNA-seq을 통해 얻어진 DNA, RNA 서열을 비교하여 후보 RNA 편집위치를 검출하고 이를 단일염기다형성(SNP), 알려진 RDD site 으로 필터하여 알려지지 않은 새로운 RDD 혹은 RNA 편집 위치를 찾아낸다. 하지만 이러한 도구들은 수많은 위양성(false-positive) 결과 들을 포함하고 있다. 많은 새로운 RNA 편집 위치들이 발견되었지만[73], 이 결과들이 많은 위양성 결과를 포함하고 있음이 증명되었다[74, 75]. 이러한 관점으로 보아 RNA 편집위치, RDD site 를 검출해 내는데 있어 가장 중요한 관점은 발견된 위치들중에서 진정한 RNA 편집 위치와 위양성 위치를 구별하는 기술이라고 할 수 있다.

출간된 논문으로 부터 인간 검수를 통해 추출된 RNA 편집 위치를 들을 기반으로 한 DARNED (a database of RNAediting in humans) [76], RADAR (a rigorously annotated database of A-to-I RNA editing) [77] 등의 데이터베이스들이 개발되었다. 그중 DARNED 는 가장 유명한 데이터베이스로써 약 42,000 개의 인간 RNA 편집 싸이트를 가지고 있다. RADAR 은 인간 1,343,464, 쥐 7,272 그리고 초파리 3,155 개의 RNA 편집싸이트를 내포하고 있다. 이러한 데이터베이들을 기반으로 웹기반은 RNA 편집 위치를 주석해주는(annotation) ExpEdit 같은 도구들이 개발되었다[78]. ExpEdit 는 DARNED 를 기반으로 하여 입력된 RNA-seq 데이터에 DARNED 를 매핑하여 정보를 제공하는 웹 기반의 도구로써 신뢰성 있는 데이터를 제공한다. 하지만 알려지지 않은 새로운 RNA 편집 위치를 검출해 내지는 못하는 한계점을 가지고 있다. 하지만 이러한 단점에서 불구하고 원시 RNA-seq 데이터 형태인 FASTQ, SAM (sequence alignment map), BAM (binary alignment map) 파일을 입력값으로 받아 웹에서 사용자가 손쉽게 신뢰성 높은 RNA 편집 위치를 결과로 얻고 또 검색할 수 있다는 점에서 각광받고 있다. 그렇지만 용량이 많은 원시 RNA-seq 데이터를 업로드 하는데 BAM 700MB 파일을 기준으로 28 시간이라는 효율적이지 못한 시간이 걸린다는 치명적인 단점이 있다. 파이썬 패키지로 제작된 REDtools 는 업로딩 시간을 단축하였으며 입력값이 같은 샘플에서 얻어진 DNA,

RNA-seq 데이터일 경우 알려지지 않은 새로운 RNA 편집 싸이트를 찾을 수 있음을 물론 입력값이 RNA-seq 결과 하나인 경우에는 기존에 있던 알려진 RNA 편집싸이트 정보를 매핑하여 제공할 수 있게 보완되어 만들어 졌다[79]. 또한 A-I 편집싸이트 정보를 웹기반의 도구로 제공하는 VIRGO[80]도 개발되었다. 하지만 ExpEdit, REDtools 는 제공한 RNA 편집위치에 대한 신뢰성 정도를 제공하지 않으며 VIRGO 는 A-I 에 한정하여 제공한다는 한계점이 있다.

본 연구는 RNA-seq 서열비교를 통한 RNA 편집위치 주석을 제공하는 새로운 도구(RCARE)로써 각 샘플에서만 특정적으로 나타나는 따른 RNA 편집 위치, 여러가지 문헌과 데이터 베이스를 통합한 지식베이스를 기반으로한 풍부하고 체계적인 지식 주석, 데이터를 기반으로한 신뢰성 등급(evidence level)을 각 RNA 편집위치 마다 제공한다. 또한 사용자가 같은 샘플에서 얻은 DNA, RNA-seq 데이터를 입력값으로 넣으면 알려지지 않은 새로운 RNA 편집 위치를 검출하여 제공한다. 사용자가 RNA-seq 의 variant fall format(VCF)를 입력값으로 사용자 친화적인 웹 페이지에 업로드 하면 한개의 RNA-seq 데이터에 대한 주석값 혹은 여러 RNA-seq 데이터 간의 비교값에 대하여 각 생물학적 요소별 요약 그래프를 제공한다. 또한 사용자가 원시 RNA-seq 데이터인 FASTQ,BAM,SAM 파일을 RNA-seq VCF 를 빠른시간에 자신의 데스크탑에서 변환할 수 있는 파이썬 스크립트를 제공한다. RCARE 는 http://www.snubi.org/ 에서 자유롭게 사용할 수 있습니다.

방 법

데이터 수집

RCARE 는 314,880, 6,830, 그리고 13,018 의 mRNA 편집 위치를 각각 DARNED(NCBI37/hg19)[76], 인간 ENCODE RNA-Seq 데이터[67], Bahn et al, [71] 로 부터 다운로드 받아 통합했다. 앞서 다운받은 1,379,404 와 RADRA [77] ,Li et al [73].에서 다운받은 10,115 개의 인간 RNA 편집 위치를 사용하여 참조(reference) 기반의 각 RNA 편집 위치별 신뢰성 등급을 만들었다. 각 RNA 편집 위치별 풍부한 생물학적 주석 정보를 위해 우리는 Homo_sapiens.GR-Ch37.69.gtf, Repeat-Masker database information from Ensembl (ensembl.org), UCSC (http://genome.ucsc.edu/buildGRCCh37/hg19)를 다운받아 ncRNA, Ensembl Gene (ENSG) ID, Transcript(ENST) ID, Exon ID (ENSE) 그리고 repetitive element (Alu, nonrepetitive)를 각 RNA 편집 위치 혹은 RDD 위치별로 제공 하였다. 또한 ANNOVAR (http://www.openbioinformatics.org/annovar/) [81]를 사용하여 intron, intergenic, splicing region, downstream, upstream, 3' UTR, 5' UTR, 그리고 synonymous/nonsynonymous 주석 정보를 각 RNA 편집 위치 혹은 RDD 위치별로 제공하였다(Table 1).

RNA-seq 데이터 전처리(preprocessing)

RCARE 는 FASTQ 혹은 BAM 파일로 부터 hg 19 VCF 파일로 전환시켜주는 자동 변환 유틸리티를 제공한다(그림 1A). 이 유틸리티는 TopHat [82], SAMtools[83], VCFtools[84], Tabix[85], 그리고 Bowtie2[86] 를 포함하고 있으며 이들을 자동으로 실행시켜 RNA-seq 원시 파일을 VCF 형태로 바꿔주는 파이썬 스크립트를 포함하고 있어 파일 크기가 큰 RNA-seq 원시 데이터를 업로드 하는 시간을 없앴다. TopHat 과 SAMtools 의 설정 파라미터들은 쉽게 변환

가능하도록 되어 있다. 변환 유틸리티는 2 가지 유형으로 구성되어져 있다.

전체 버전 변환 유틸리티는 tophat, 알려지지 않은 새로운 RDD 위치 혹은 RNA 편집 위치를 찾는 파이썬 스크립트와 서로 다른 RNA-seq 데이터를 비교하여 특정 샘플에서만 나타나는 RNA 편집 위치 혹은 RDD 위치를 찾는 파이썬 스크립트, 시험 가동을 위한 샘플 파일등을 비롯한 BAM 혹은 FASTQ를 VCF로 변환하는데 필요한 모든 도구들이 포함있다. 하지만 만약 유저가 tophat, samtools 갖은 도구들이 설치되어 있다면 라이트 버전을 빠르게 다운받아 사용하면 된다. 자세한 데이터 처리 과정은 그림 1 을 참조하면 된다. 만약 유저가 STAR 그리고 GATK 같은 다른 도구를 사용하고 싶다면 해당 도구를 사용해 VCF 를 생성한뒤 웹 버전에 업로드 하면 알려진 RNA 편집 위치 주석, 그리고 해당 위치들에 대한 생물학적 정보를 얻을 수 있으며 2 가지 이상의 데이터를 업로드 하면 서로 비교하여 특정 샘플에서만 나타나는 RNA 편집 위치를 검출해 낼 수 있으며 같은 시료에서 얻어진 DNA, RNA 서열 데이터를 VCF 형태로 입력하면 염색체(chromosome), 위치(position)을 비교하여 알려지지 않은 새로운 RDD 위치 혹은 RNA 편집 위치등을 검출 할 수 있다.

RNA 편집 위치들의 비교

RACE 에서는 다른 샘플에서 얻은 RNA-seq VCF 를 염색체(chromosome), 위치(position)을 기준으로비교하여 특정 샘플에서만 나타나는 기능이 알려진 RNA 편집 위치를 식별해 준다. 자세한 사항은 그림 3 에 기술했다.

웹 인터페이스 구축

RCARE 의 웹 기반의 응용프로그램은 HTML5 (Hypertext Markup Language 5), CSS3(Cascading Style Sheets 3), jQuery, and

Highcharts API(http://www.highcharts.com/)를 사용하여 구축하였다. 이 응용프로그램은 RNA-seq 데이터 처리, 비교, 주석 그리고 시각화로 구성된 4 가지 기능을 제공한다.

결 과

우리는 DARNED(NCBI37/hg19)[76], 인간 ENCODE RNA-Seq 데이터 [67],Bahn et al.[71]로 부터 321,008 개의 RNA 편집 위치를 수집하였다. 이 데이터는 30 가지 샘플 종류로 구성된 154 개의 샘플, 23 개의 논문 그리고 11,299 개의 유전자, 12 가지의 RNA 편집 종류로 포함되어 있다(그림 2). RCARE 의 생물학적 정보 주석은 (1) synonymous vs. nonsynonymous 변화, (2) splicing junction 지역 안의 포함 여부, (3) genomic features, (4) Alu 연관여부,(5) ncRNA 지역안의 포함 여부, (6) 유전자 기호(gene symbol), 다양한 ID 등의 유전자 정보(자세한정보는 첨부파일 참조) (7) 샘플 기원, (8) 참조 논문,(9) 참조 데이터를 기반으로 한 신뢰성 등급등 9 가지의 카테고리로 구성되어 있다. 또한 서로 다른 모든 샘플들을 비교하여 샘플 사이에서 교차 혹은 차이가 나는 RNA 편집 위치 마다 17 가지의 유용한 생물학적 주석을 제공한다(표 2). RNA 편집 위치를 검출하는 것도 중요하지만 검출된 데이터의 신뢰성은 무엇보다 중요하다. 검출된 RNA 편집 위치의 신뢰성을 향상시키기 위해 우리는 신뢰성 등급(Evidence levels)을 생성하였다. 신뢰성 등급은 각 RNA 편집 위치가 관련 논문이나 데이터베이스에서 언급된 횟수를 기반으로 생성하였다. 우리는 신뢰성 등급을 생성하기 위하여 첫번째로 5 가지 논문 혹은 지식베이스(DARNED, RADAR, Bahn et al.[71], Li et al.[73], Park et al.[67])를 수집하였다. 두번째로 모아진 데이터들을 염색체, 위치 정보, reference/alternative 서열정보를 사용하여 통합하고 각 RNA 편집 싸이트가 언급된 횟수를 세어 신뢰성 등급을 생성하였다. 신뢰성 등급은 A 부터 E 까지 총 5 단계로 구성되어져 있으며 A 로 갈수록 높은 등급의 신뢰를 가진다고 말할 수 있다(그림 3, 보충자료 1). 예를 들어 만약 한개의 RNA 편집 위치가 등급이 A 라면 이는 5 개의 데이터베이스 혹은 관련 논문에서 찾아 졌다고 할 수 있다. 반대로 E 등급이라면 한개의 데이터베이스 혹은 논문에서 찾아 졌다고 할 수 있다. 통합된 데이터의 85%의 RNA 편집 위치는 D 등급을 가지고 있었다. 이 신뢰성 등급은 사용자가 등급 스케일에 따라 해당 RNA 편집 위치가 위양성 인지 참인지를 결정하는데 큰 도움을 줄것이다.

웹 인터페이스는 사용자 설명서, 도구 다운로드, 분석 이렇게 3 부분으로 구성되어 있다. 사용자 설명서 영역은 변환 유틸리티, 웹을 사용한 생물학적 주석 달기, 샘플간의 비교 그리고 결과 설명으로 이루어져 있다(그림 1B). 모든 설명서 파일은 PDF 파일로 다운로드 가능하다. 도구 다운로드 영역은 변환 유틸리티를 다운로드 할 수 있따. 우리는 2 가지 버전의 변환 유틸리티를 제공하고 있다. 전체 버전은 TopHat[82], SAMtools[83], VCFtools[84], Tabix[85], Bowtie2[86], 자동 파이썬 스크립트 등 원시 RNA-seq 데이터를 VCF 로 변환하는 모든 도구를 포함하고 있으며 라이트 버전은 파이선 스크립트 만을 포함하고 있다. 분석 영역은 16 가지 유용한 생물학적 주석과 신뢰성 등급을 VCF 에 있는 모든 RNA 편집 위치마다 제공한다. 비교 부분은 서로 다른 샘플로 부터 생성된 RNA-seq VCF 파일의 모든 조합을 비교하여 특정 샘플에서만 나타나는 RNA 편집 위치를 16 가지 유용한 생물학적 주석 정보와 함께 제공한다. 결과 페이지는 genomic 기능, 유전자, ncRNAs, synonymous vs nonsynonymous 변화 , RNA editing 유형 그리고 검출된 RNA 편집위치들의 각 염색체 혹은 샘플별 분포를 요약 그래프를 제공한다.(Figure 1C) 모든 그래프 이미지는 다운로드와 인쇄가 가능하다. 결과를 시험하기 위하여 우리는 MCF-7(a breast cancer cell line), HUVEC (a human umbilicalvein endothelial cell line) 그리고 HeLa-S3 (a cervical carcinoma cell line) RNA-seq 데이터를 ENCODE (http://genome.ucsc.edu/ENCODE/)로 부터 다운로드 받아 2347,646, 1190 개의 RNA 편집 위치를 검출했다. 이 위치들 중 HUVEC 205 개, MCF-7 605 개, HeLa-S3 334 개가 A 부터 C 사이의 신뢰성 등급을 가지고 있었다. 하지만 이것은 전체 데이터중 각각 31.7%, 24.8% 그리고 28.07% 에 해당하였다. 이 결과로 미루어 보아 신뢰성 등급이 RNA 편집 위치를 검출하고 분석하는데 반드시 필요함을 알 수 있었다. 우리는 또한 MCF-7 과 HUVEC 을 비교하여 각 샘플에서만 나타나는 알려진 RNA 편집위치를 검출하였다. 그결과 MCF-7 으로 부터 특이적으로 유방암에서만 나타나는 2,080 개의 RNA 편집 위치를 검출하였다. 이러한 결과는 특정 조건에서 RNA 편집 위치를 검출하는것의 중요성을 보여준다. 우리는 37.3 Mb 의 샘플 파일로 Expedit 와 RCARE 의 실행 시간을 비교를 위해 RACRE 의

변환처리과정과 주석 분석하는 시간을 측정하였다. BAM 파일을 VCF 로 변환하는 처리과정은 3.0 GHz CPU, 2048 MB RAM 의 데스트탑 환경에서 192 초에 수행되었고 RNA 편집 위치당 생물학적 주석을 처리하는 과정은 14.19 MBps 네트워크 환경에서 7 초가 걸렸다.

고 찰

RCARE 는 신뢰할 수 있는 RNA 편집 위치에 대한 생물학적 주석처리, 비교, 그래프 시각화를 효율적이고 사용자 친화적인 웹기반의 시스템이다. RCARE 는 321,008 개의 인간 RNA 편집위치와 이에 해당하는 풍부한 생물학적 주석정보 그리고 유용한 요약 그래프를 신뢰성 등급과 함께 제공하고 있다. 게다가 이 도구는 RNA-seq 원시 데이터를 VCF 로 자동을 변환해 주는 변환 유틸리티를 파이썬 기반의 스크립트와 함께 제공하고 있다. 이 유틸리티로부터 나온 VCF 를 웹에 업로드 하면 생물학적 분석 그리고 서로 다른 샘플간의 비교 하여 생물학적 주석정보와 함께 제공 받을 수 있다. RCARE 웹 인터페이스는 쉽게 주석처리와 시각화를 할 수 있으며 그 그결과를 CSV 형태와 JPG 형태로 다운로드 받을 수 있다. 최근 2~3 년 사이에 새로운 RNA 편집 위치를 찾아내는 것이 이 분야 연구의 초점이 되어왔다[67, 71, 73]. 그러나 최근 연구동향은 확인된 RNA 편집위치의 신뢰성 판정 여부로 옮겨 갔다. 왜냐하면 발견된 RNA 편집 위치들에서 많은 위양성이 발견되었기 때문이다. 우리는 RCARE 의 신뢰성 등급이 RNA 편집 위치 연구 분야에서 그 수요가 크게 증가할 것이라고 예상한다. RCARE 는 신뢰할 수 잇는 RNA 편집 위치를 동정하는데 크게 도움이 될 것이라고 예상된다.

결과그림

그림 1 RCARE 의 RNA-seq 데이터 처리 과정 단계 및 생물학적 주석처리 결과 시각화 과정 A) RNA-seq 데이터 처리 변환 유틸리티 처리 과정(pipeline). B) 생물학적 주석 처리와 비교 분석을 위한 웹 인터페이스, C)주석 결과의 요약 그래프 예제

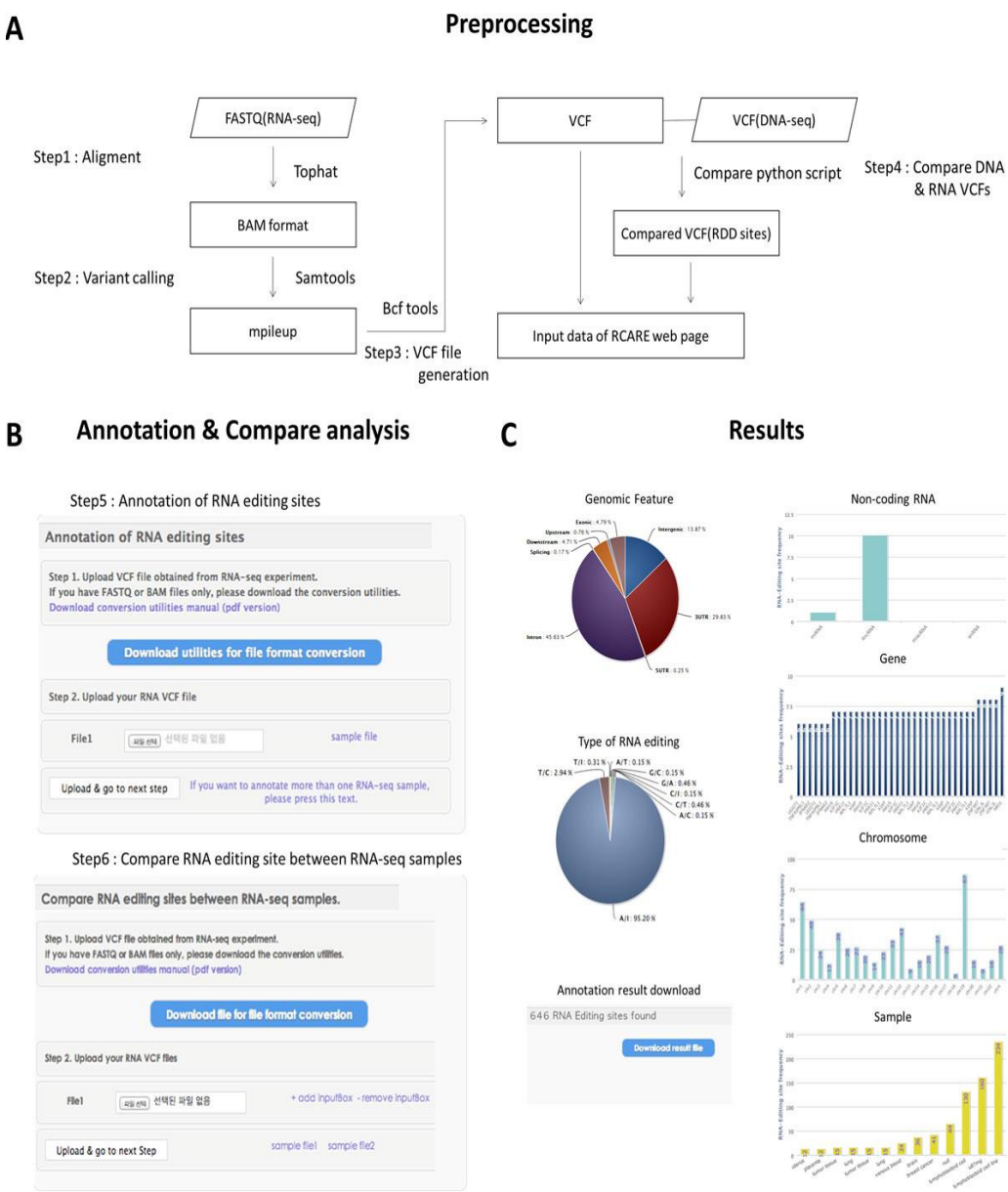
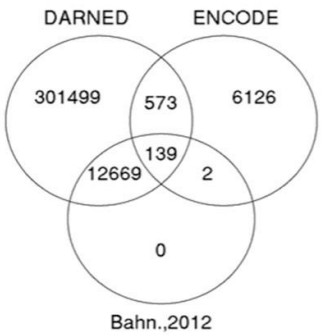
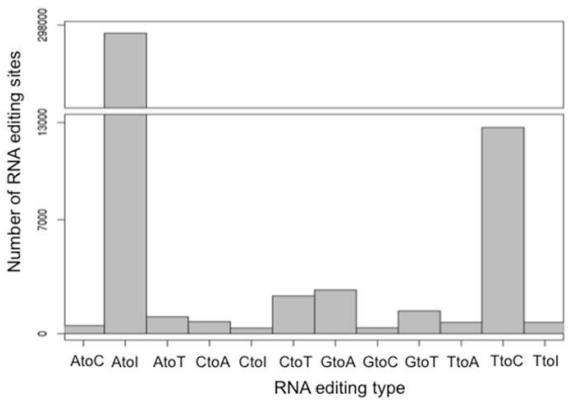


그림 2 데이터 구성. A) DARNED, ENCODE, Bahn et[71] 데이터에 포함되어 있는 RNA 편집 위치 개수를 나타내는 벤다이어그램. B) 각 샘플당 RNA 편집 위치 개수. C) 각 RNA 편집 타입별 RNA 편집 위치 개수. D)각 염색체당 포함되어 있는 RNA 편집위치 개수. E) 각 게놈 기능(genomic feature)당 포함되어 있는 RNA 편집위치 개수.

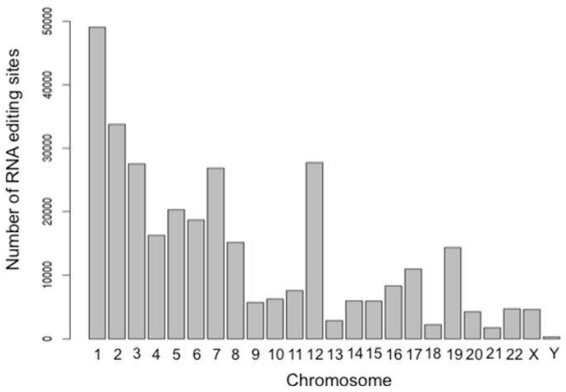
A



C



D



B

Sample type	No. RNA editing sites
Adipose	85
Adrenal	467
Blood	251,796
Blood Vessel	132
Bone	223
Brain	16,287
Breast	528
Cancer	16,811
Disease	436
Eye	736
Heart	243
Kidney	1,618
Large Intestine	375
Liver	296
Lung	1,836
Muscle	322
Pancreas	65
Prostate	1,276
Salivary Gland	196
Skin	287
Small Intestine	665
Spleen	1,586
Stem Cell	586
Stomach	227
Testis	1,364
Throat	591
Thymus	2,188
Thyroid	174
Uterus	3,295
Not Specified	25,299
Total	329,992

E

Genomic Feature	No. RNA editing sites
3'UTR	17,483
5'UTR	345
Intergenic region	38,319
Exon	3,634
Intron	255,891
Splice site	45
Upstream	1,035
Downstream	4,309
Total	321,061

결과 표

표 1 현재 RCARE 에 RNA 편집 위치들의 데이터의 출처 목록.

각 데이터베이스 별 항목은 고유한 전체 RNA 편집 싸이트 수를 의미한다.

	샘플 수	RNA 편집 위치 수	Alu 서열 수	참고문헌 수
DARNED	29	314,880	15,783	34
ENCODE	27	6,840	347	18
RADAR	30	291,901	14,318	31
Bahn et al. [71]	27	12,810	2,916	15
Li et al. [73]	1	1	0	1
Total	114	626,432	33,364	99

표 2 RCARE의 주석 결과 형식.

순번	항목	설명	참고문헌
1	Chr	Chromosome of the RNA editing site in the reference genome.	[71, 73, 76]
2	Pos	Coordinate of the RNA editing site in the reference genome.	
3	In DNA	Base of the RNA editing site in the DNA reference sequence.	
4	In RNA	Base of the RNA editing site in the RNA sequence of sample.	
5	Gene	Gene name to which the RNA editing site belongs.	
6	*Evidence level*	The *evidence level* consists of five levels (A–E), where A is highest level (e.g., if an RNA editing site had level "A," it appeared in all five of the resource databases/papers used). *Level A: The RNA editing site appeared in five resources (evidence No. 5). *Level B: The RNA editing site appeared in four resources (evidence No. 4). *Level C: The RNA editing site appeared in three resources (evidence No. 3). *Level D: The RNA editing site appeared in two resources (evidence No. 2). *Level E: The RNA editing site appeared in one resource (evidence No. 1).	[70,72,75,76] RepeatMasker
7	Strand	+ for positive strand; − for negative strand.	[70,72,75]
8	Source	This field contains information regarding the tissue source from which the RNA editing instance was obtained.	
9	PubMed ID	This field provides the reference article from which the RNA editing data was extracted.	

10	Alu	This field provides information of Alu at the RNA editing site.	RepeatMasker
11	Data reference	Reference database.	Each database or reference
12	ENSG	Ensembl Gene ID.	GTF(*Homo sapiens*, GRCH37.17) in Ensembl
13	ENST	Ensembl Transcript ID.	
14	ENSE	Ensembl Exon ID.	
15	Genomic feature	Genomic feature of the RNA editing site. *Exonic: the variant overlaps a coding exon. *Splicing: the variant is within 2 bp of a splicing junction. *ncRNA: the variant overlaps a transcript without coding annotation in the gene definition. *5’ UTR: the variant overlaps a 5' untranslated region. *3’ UTR: the variant overlaps a 3' untranslated region. *Intronic: the variant overlaps an intron. *Upstream: the variant overlaps the 1−kb region upstream of the transcription start site. *Downstream: the variant overlaps the 1−kb region downstream of the transcription end site. *Intergenic: a variant is in the intergenic region.	[81]
16	Synonymous or nonsynonymous	Synonymous or nonsynonymous substitutions at the RNA editing site.	[80]

17	Noncoding RNA	This field indicates whether the location of an RNA editing site is in ncRNA.	GTF (*Homo sapiens*, GRCH37.17) in Ensembl

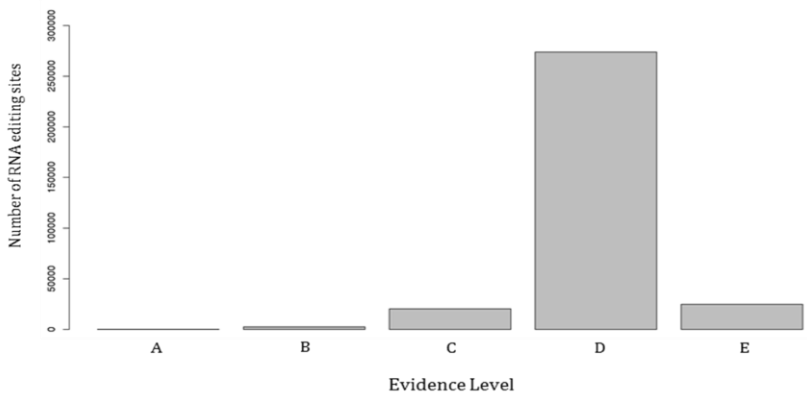
표 3 RCARE 의 신뢰성 등급 분류 체계.

분류	신뢰성 등급	등급의 구성 자료	수
A	5	DARNED, ENCODE, Bahn et al. [5], RADAR, Alu	33
B	4	DARNED, RADAR, Bahn et al. [5], Alu	2,204
		DARNED, ENCODE, RADAR, Alu	23
		DARNED, ENCODE, Bahn et al. [5], RADAR	106
C	3	DARNED, Bahn et al. [5], RADAR	7,037
		DARNED, Bahn et al. [5], Alu	679
		DARNED, ENCODE, RADAR	550
		DARNED, RADAR, Alu	12,038
		DARNED, RADAR, Li et al. [11]	1
		ENCODE, Bahn et al. [5], RADAR	1
		ENCODE, RADAR, Alu	20
D	2	DARNED, Alu	806
		DARNED, RADAR	269,547
		DARNED, Bahn et al. [5]	2,749
		ENCODE, Alu	271
		ENCODE, RADAR	341
		ENCODE, Bahn et al. [5]	1
E	1	DARNED	19,107
		ENCODE	5,494
	Total		321,008

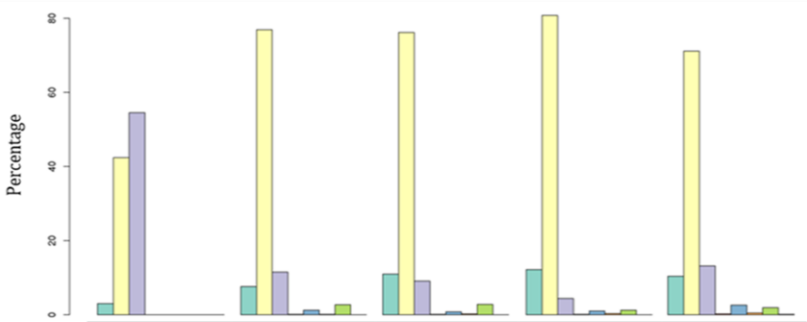
보충자료 1 각 신뢰성 등급, genomic feature 당 속하는 RNA 편집 위치 수.

A) 각 신뢰성 등급당 RNA 편집 위치 수 B) 각 genomic feature 에 속하는 신뢰성 등급당 RNA 편집 위치 수

A



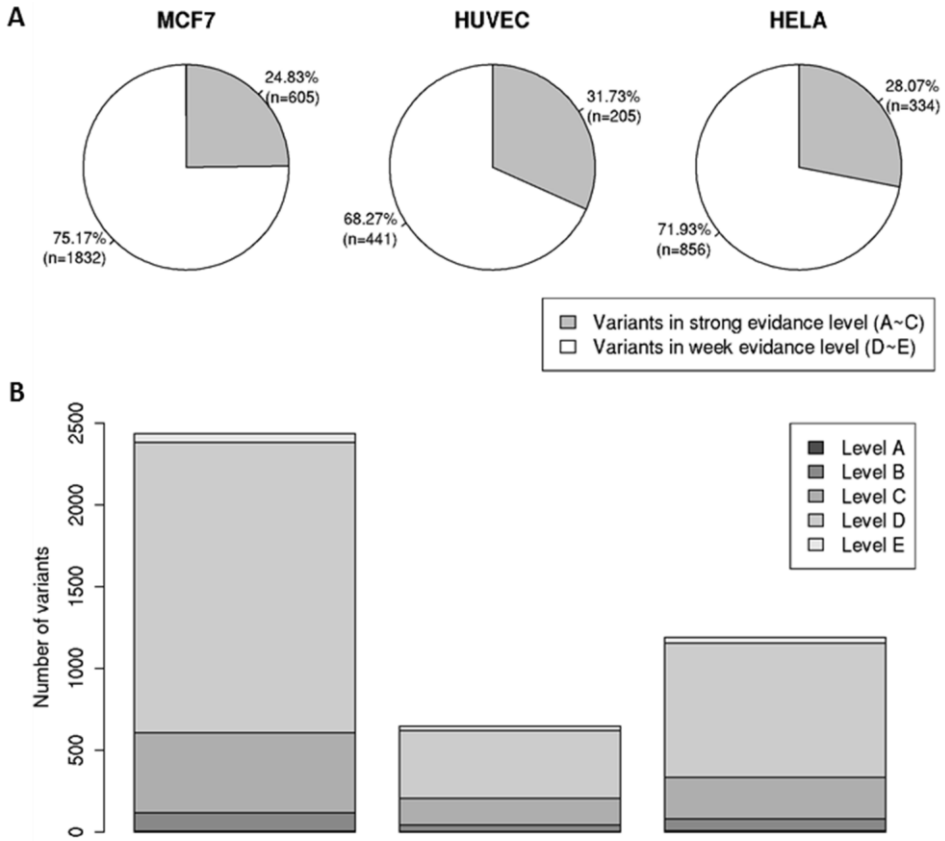
B



Evidence level	A	B	C	D	E
Intergenic region	1	177	2,237	33,344	2,560
Intronic region	14	1,794	15,486	221,100	17,497
3'UTR	18	269	1,842	12,106	3,248
5'UTR	0	2	11	280	52
Exonic region	0	27	154	2,818	635
Upstream	0	2	36	882	115
Downstream	0	62	562	3,206	479
Splice site	0	0	3	15	27
Total	33	2,333	20,331	273,751	24,613

보충자료 2 3 개의 세포주에 속하는 신뢰성 등급의 비율.

A) MCF-7 (a breast cancer cell line), HUVEC (a human umbilical vein endothelial cell line) and HeLa-S3 (a cervical carcinoma cell line) 세포주에서 검출된 RNA 편집 위치의 신뢰성 등급 A-C 와 D-E 에 속하는 편집위치 수의 비율. B)3 개의 세포주당 RNA 편집위치 수

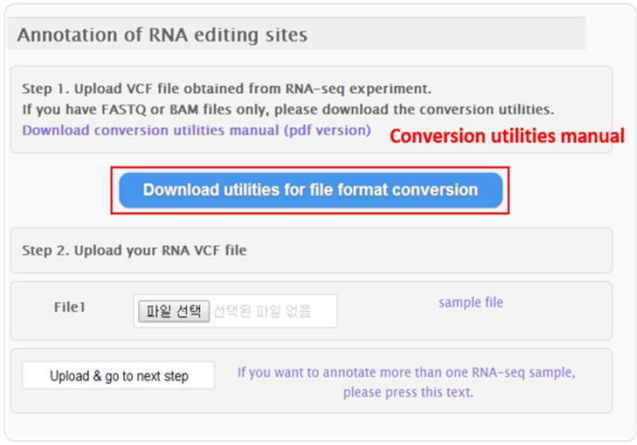


Cell line	MCF7	HUVEC	HELA
Level A	4 (0.16%)	2 (0.31%)	7 (0.59%)
Level B	112 (4.60%)	39 (6.04%)	71 (5.96%)
Level C	489 (20.07%)	164 (25.39%)	256 (21.51%)
Level D	1,779 (73.00%)	415 (64.24%)	822 (69.08%)
Level E	53 (2.17%)	26 (4.02%)	34 (2.86%)
Total	2,437 (100.00%)	646 (100.00%)	1,190 (100.00%)

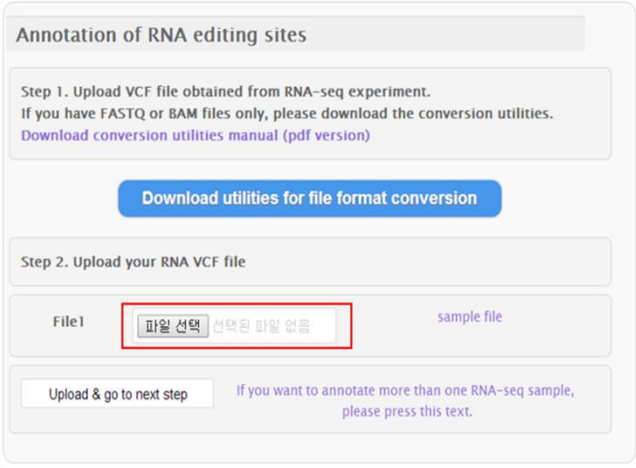
보충자료 3 RCARE 웹 인터페이스 사용자 설명서

1. Annotation section

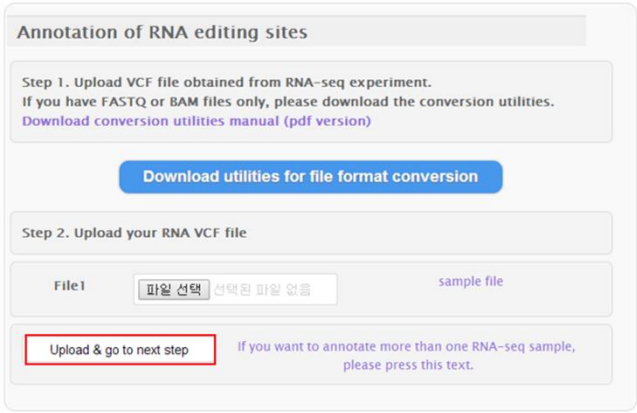
A. If you have FASTQ or BAM files only, please download the conversion utilities.



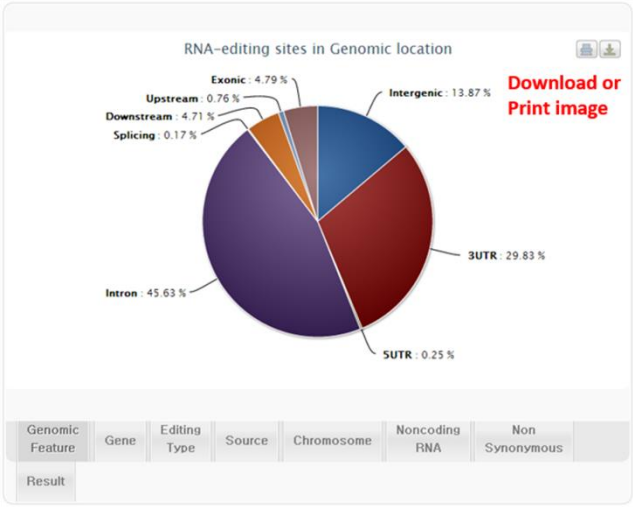
B. Upload a VCF file for RNA editing site annotations.



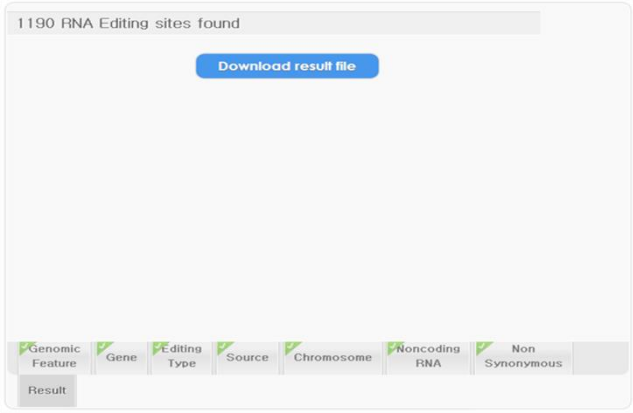
C. Press ‘Upload & to go next step’ button for RNA editing site annotations.



D. Result graphs show annotated results.

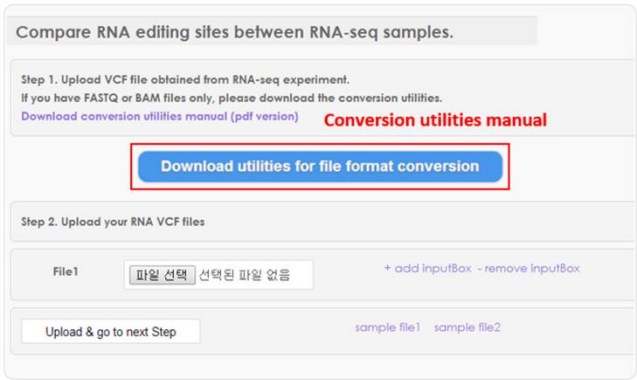


E. Press "*Download result file*" button to download result file.

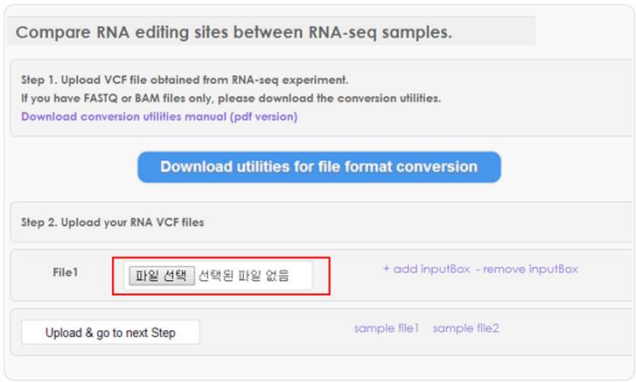


2. Compare section

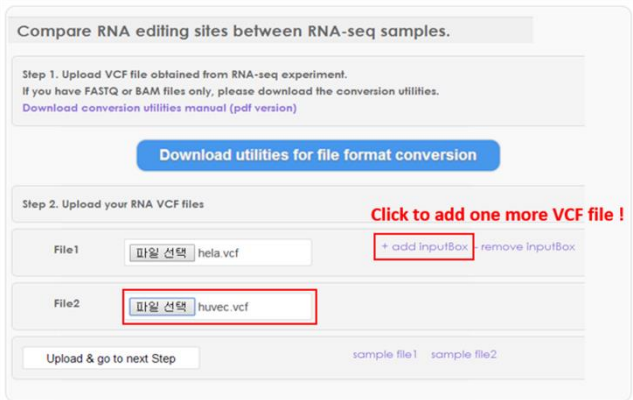
A. If you have FASTQ or BAM files only, please download the conversion utilities.



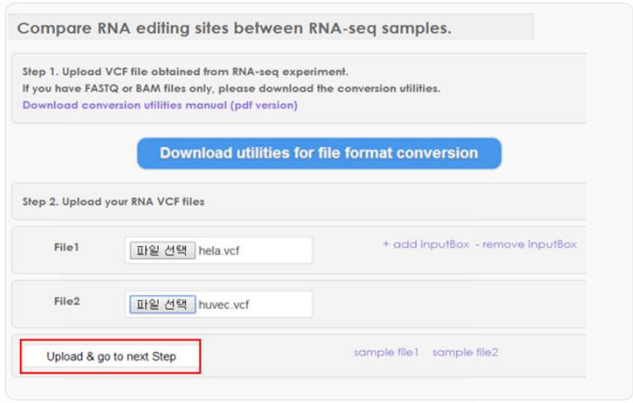
B. Upload a VCF file for annotation of RNA editing sites.



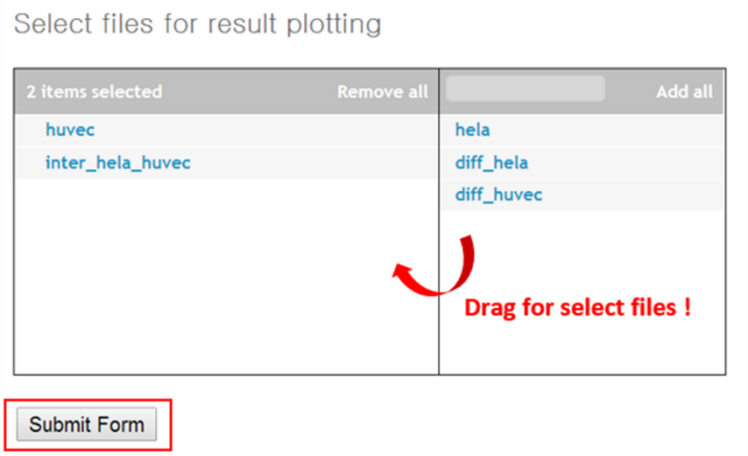
C. Add one more VCF file for comparison.



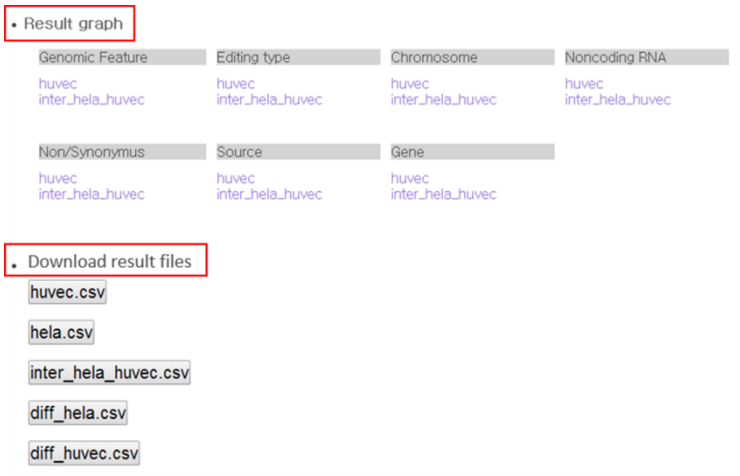
D. Press ‘Upload & to go next step’ button to annotate RNA editing sites.



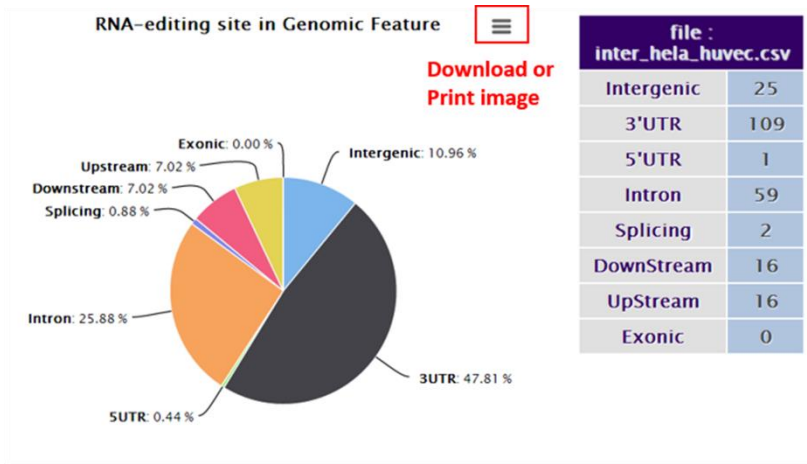
Select result files by dragging the files from the right column to the left, and press “*Submit Form*” button.



E. Select result graph to view and download result files.



F. Result graphs show annotated results.



보충자료 4 RCARE 변환 유틸리티 사용자 설명서

1. Description

RCARE convert utilities is a set of Python-based utilities for converting FASTQ and BAM (binary format for storing sequence data) into VCF (variant call format) and comparing RNA and DNA VCF files from the same sample. The package provides customized TopHat and SAMtools commands that the user can execute. RCARE convert utilities provides an autoinstallation function for the tools. This is very easy for researchers to use, even for those with no experience of RNA-Seq data analysis.

RCARE convert utilities contains TopHat (http://tophat.cbcb.umd.edu/), SAMtools (http://samtools.sourceforge.net/), Tabix (http://samtools.sourceforge.net/tabix.shtml), VCFtools (http://vcftools.sourceforge.net/), and Bowtie2 (http://bowtie-bio.sourceforge.net/bowtie2). If user presetup tools including TopHat, download light RCARE convert utilities and installation.

2. Input data format

RCARE convert utilities convert three sequence formats (FASTQ, BAM, and VCF) to VCF, which is the input format on the RCARE website.

- FASTQ format
  - ➔ FASTQ format is a text-based format for storing both a biological sequence (usually a nucleotide sequence) and its corresponding quality scores.
- BAM format
  - ➔ BAM format is a binary format for storing sequence data. (http://samtools.sourceforge.net/SAMv1.pdf).
- VCF format (variant call format)
  - ➔ VCF is a text file format (most likely stored in a compressed manner). It contains meta information lines, a header line, and

data lines, each containing information about a position in the genome.

3. Installing and testing the installation

3-1 Install quick start

- RCARE needs a presetup in the Python environment.
- Download RCARE convert utilities (4.75G) from the website.
- Unzip RCARE convert utilities.
  - ➔ Tar -xvf RCARE-pre-processing.tar.gz
- Run rcare.py for your purposes.

3-2 Test the installation

The sample BAM data contained only 21 chromosome. These data were extracted from paired-end RNA-Seq using HeLa cells in ENCODE (http://genome.ucsc.edu/ENCODE).

- Input data confirmation
  - ➔ ls ./input_data/bam/
- Test command
  - ➔ python rcare.py -ib sample.bam -fn sample_bam_test
- Result confirmation
  - ➔ ls ./result_data/vcf/sample_bam_test/

4. Synopsis and example

4.1 Input data folder consists of FASTQ, BAM, and VCF. Insert into row data in each folder.

4.2 Convert paired-end FASTQ files into VCF format

- ➔ Python rcare.py -if -p S1.fastq S2.fastq -fn fastq_test
- Result confirmation
  - ➔ ls ./input_data/vcf/fastq_test/

4.3 Convert single FASTQ file into VCF format

- ➔ python rcare.py -if -s S1.fastq -fn single_fastq_test
- Result confirmation
  - ➔ ls ./input_data/vcf/fastq_test

4.4 Convert BAM into convert to VCF format

➔ python rcare.py −ib sample.bam −fn sample_bam_test

− Result confirmation

➔ ./result_data/vcf/sample_bam_test

− Compare RNA VCF with DNA VCF file

➔ python rcare.py −c DNA.vcf RNA.vcf −fn 1_compare_test

− Result confirmation

➔ ls ./result_data/compare/1_compare_test/

4.5 Customized TopHat command running

➔ python rcare.py −tc "tophat" −fn test

4.6 Customized SAMtools command running

➔ python rcare.py −sc "samtools" −fn test\

5. RCARE convert utilities options

Option	Description
−if	Input file format: FASTQ file
−ib	Input file format: BAM file
−p	Paired end FASTQ file
−c	Compare VCF(RNA) with VCF(DNA)
−fn	Result file name
−tc	Customized TopHat commands
−sc	Customized SAMtools commands

Folder file name	Description
input_data	Insert input data
result_data	Save result data

resource	Files required for preprocessing
tools	Tools required for preprocessing
rcare.py	Batch file of convert utilities

6. Package composition

7. Light RCARE convert utilities installation

- Download light-RCARE-pre-processing.tar.gz (35.62 MB) from RCARE website.
- Download tools:
  1. TopHat: http://tophat.cbcb.umd.edu/
  2. SAMtools: http://samtools.sourceforge.net/
  3. Bowtie2: http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
  4. Tabix: http://samtools.sourceforge.net/tabix.shtml
  5. VCFtools: http://vcftools.sourceforge.net/
- All tools insert into the tools folder in the RCARE convert utilities
- If user has used previous setup tools, initialize each tool's environment settings

8. Authors

Ju Han Kim and Soo Youn Lee from SNUBI (Seoul National University Biomedical Informatics; http://www.snubi.org/)

9. References

1. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009; 25: 2078-9.

2. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 2009; 25: 1105-11.
3. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R; 1000 Genomes Project Analysis Group. The variant call format and VCFtools. Bioinformatics 2011; 27: 2156-8.
4. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature methods 2012; 9: 357-9.
5. Li H. Tabix: fast retrieval of sequence features from generic TAB-delimited files. Bioinformatics 2011; 27: 718-9.

참고문헌

1. Jain, K.K., *Personalized medicine.* Curr Opin Mol Ther, 2002. **4**(6): p. 548-58.
2. Wei, C.Y., M.T. Lee, and Y.T. Chen, *Pharmacogenomics of adverse drug reactions: implementing personalized medicine.* Hum Mol Genet, 2012. **21**(R1): p. R58-65.
3. Lazarou, J., B.H. Pomeranz, and P.N. Corey, *Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies.* JAMA, 1998. **279**(15): p. 1200-5.
4. Severino, G. and M. Del Zompo, *Adverse drug reactions: role of pharmacogenomics.* Pharmacol Res, 2004. **49**(4): p. 363-73.
5. Pirmohamed, M., et al., *Adverse drug reactions as cause of admission to hospital: prospective analysis of 18 820 patients.* BMJ, 2004. **329**(7456): p. 15-9.
6. Rottenkolber, D., et al., *Adverse drug reactions in Germany: direct costs of internal medicine hospitalizations.* Pharmacoepidemiol Drug Saf, 2011. **20**(6): p. 626-34.
7. Xie, H.F., F., *Pharmacogenomics steps toward personalized medicine.* Personalized Medicine, 2005. **2**(4): p. 325-337.
8. Need, A.C., A.G. Motulsky, and D.B. Goldstein, *Priorities and standards in pharmacogenetic research.* Nat Genet, 2005. **37**(7): p. 671-81.
9. Crews, K.R., et al., *Pharmacogenomics and individualized medicine: translating science into practice.* Clin Pharmacol Ther, 2012. **92**(4): p. 467-75.
10. Zhou, K. and E.R. Pearson, *Insights from genome-wide association studies of drug response.* Annu Rev Pharmacol Toxicol, 2013. **53**: p. 299-310.
11. Stocco, G., K.R. Crews, and W.E. Evans, *Genetic polymorphism of inosine-triphosphate-pyrophosphatase influences mercaptopurine metabolism and toxicity during treatment of acute lymphoblastic leukemia individualized for thiopurine-S-methyl-transferase status.* Expert Opin Drug Saf, 2010. **9**(1): p. 23-37.
12. Zaza, G., et al., *Pharmacogenomics: a new paradigm to personalize treatments in nephrology patients.* Clin Exp Immunol, 2010. **159**(3): p. 268-80.
13. Hudis, C.A., *Trastuzumab--mechanism of action and use in clinical practice.* N Engl J Med, 2007. **357**(1): p. 39-51.
14. Petak, I., et al., *Integrating molecular diagnostics into anticancer drug discovery.* Nat Rev Drug Discov, 2010. **9**(7): p. 523-35.
15. Eichler, H.G., et al., *Bridging the efficacy-effectiveness gap: a regulator's perspective on addressing variability of drug response.* Nat Rev Drug Discov, 2011. **10**(7): p. 495-506.

16. Wang, L., H.L. McLeod, and R.M. Weinshilboum, *Genomics and drug response.* N Engl J Med, 2011. **364**(12): p. 1144-53.
17. FDA. *Table of Pharmacogenomic Biomarkers in Drug Labeling*. May 20, 2015; Available from: http://www.fda.gov/Drugs/ScienceResearch/ResearchAreas/Pharmacogenetics/ucm083378.htm
18. Wikipedia. *Pharmacogenetics*. August 26, 2015; Available from: https://en.wikipedia.org/wiki/Pharmacogenetics.
19. Wetterstrand, K. *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)*. June 15, 2015; Available from: http://www.genome.gov/sequencingcosts/.
20. Genomes Project, C., et al., *An integrated map of genetic variation from 1,092 human genomes.* Nature, 2012. **491**(7422): p. 56-65.
21. Gamazon, E.R., et al., *A pharmacogene database enhanced by the 1000 Genomes Project.* Pharmacogenet Genomics, 2009. **19**(10): p. 829-32.
22. Altshuler, D., M.J. Daly, and E.S. Lander, *Genetic mapping in human disease.* Science, 2008. **322**(5903): p. 881-8.
23. Hindorff LA, M.J., Morales J, Junkins HA, Hall PN, Klemm AK, and Manolio TA. *A Catalog of Published Genome-Wide Association Studies*. March 23, 2015; Available from: http://www.genome.gov/gwastudies/.
24. Risch, N. and K. Merikangas, *The future of genetic studies of complex human diseases.* Science, 1996. **273**(5281): p. 1516-7.
25. Motsinger-Reif, A.A., et al., *Genome-wide association studies in pharmacogenomics: successes and lessons.* Pharmacogenet Genomics, 2013. **23**(8): p. 383-94.
26. Takeuchi, F., et al., *A genome-wide association study confirms VKORC1, CYP2C9, and CYP4F2 as principal genetic determinants of warfarin dose.* PLoS Genet, 2009. **5**(3): p. e1000433.
27. Loebstein, R., et al., *Interindividual variability in sensitivity to warfarin--Nature or nurture?* Clin Pharmacol Ther, 2001. **70**(2): p. 159-64.
28. Au, N. and A.E. Rettie, *Pharmacogenomics of 4-hydroxycoumarin anticoagulants.* Drug Metab Rev, 2008. **40**(2): p. 355-75.
29. Sibbing, D., et al., *Cytochrome 2C19*17 allelic variant, platelet aggregation, bleeding events, and stent thrombosis in clopidogrel-treated patients with coronary stent placement.* Circulation, 2010. **121**(4): p. 512-8.
30. Schroth, W., et al., *Breast cancer treatment outcome with adjuvant tamoxifen*

relative to patient CYP2D6 and CYP2C19 genotypes. J Clin Oncol, 2007. **25**(33): p. 5187-93.

31. Alomar, M.J., *Factors affecting the development of adverse drug reactions (Review article).* Saudi Pharm J, 2014. **22**(2): p. 83-94.
32. Ong, F.S., et al., *Clinical utility of pharmacogenetic biomarkers in cardiovascular therapeutics: a challenge for clinical implementation.* Pharmacogenomics, 2012. **13**(4): p. 465-75.
33. Lord, J., A.J. Lu, and C. Cruchaga, *Identification of rare variants in Alzheimer's disease.* Front Genet, 2014. **5**: p. 369.
34. Cirulli, E.T. and D.B. Goldstein, *Uncovering the roles of rare variants in common disease through whole-genome sequencing.* Nat Rev Genet, 2010. **11**(6): p. 415-25.
35. Ramsey, L.B., et al., *Rare versus common variants in pharmacogenetics: SLCO1B1 variation and methotrexate disposition.* Genome Res, 2012. **22**(1): p. 1-8.
36. Ng, P.C. and S. Henikoff, *SIFT: Predicting amino acid changes that affect protein function.* Nucleic Acids Res, 2003. **31**(13): p. 3812-4.
37. Law, V., et al., *DrugBank 4.0: shedding new light on drug metabolism.* Nucleic Acids Res, 2014. **42**(Database issue): p. D1091-7.
38. Kanehisa, M., et al., *From genomics to chemical genomics: new developments in KEGG.* Nucleic Acids Res, 2006. **34**(Database issue): p. D354-7.
39. Whirl-Carrillo M, M.E., Hebert JM, Gong L, Sangkuhl K, Thorn CF, Altman RB, Klein TE. *Pharmacogenomics Knowledge Base*. Available from: https://www.pharmgkb.org.
40. 1000Genomes. *1000 Genomes Project*. Available from: http://www.1000genomes.org/.
41. 1000Genomes. *Which populations are part of your study?* ; Available from: http://www.1000genomes.org/category/frequently-asked-questions/population.
42. WHOCC. *International language for drug utilization research*. May 5, 2015.
43. CDC, *National Center for Health Statistics.*
44. Ng, P.C. *SIFT Help*. August 2001; Available from: http://sift.jcvi.org/www/SIFT_help.html.
45. Habegger, L. *Variant Annotation Tool*. May 27, 2014; Available from: http://vat.gersteinlab.org/.
46. Habegger, L., et al., *VAT: a computational framework to functionally annotate variants in personal genomes within a cloud-computing environment.* Bioinformatics, 2012. **28**(17): p. 2267-9.
47. Boyle, A. *RegulomeDB*. 2012; Available from: http://www.regulomedb.org/.

48. Boyle, A.P., et al., *Annotation of functional variation in personal genomes using RegulomeDB.* Genome Res, 2012. **22**(9): p. 1790-7.
49. Baik SY, L.S., Park CH, Yoon JH, Kim JH, *Deleterious Coding Variant Analysis for Personalized Prevention of Adverse Drug Reactions.* 2013.
50. Su Youn Baik, S.Y.L., Chan Hee Park, Jun Hee Yoon, Ju Han Kim, *Deleterious Coding Variant Analysis for Personalized Prevention of Adverse Drug Reactions.* 2015.
51. Wikipedia. *Central Tendency*. July 21, 2015; Available from: https://en.wikipedia.org/wiki/Central_tendency.
52. Wikipedia. *Winsorising*. July 23, 2015; Available from: https://en.wikipedia.org/wiki/Winsorising.
53. Lin, Y.C., et al., *Identifying rare and common disease associated variants in genomic data using Parkinson's disease as a model.* J Biomed Sci, 2014. **21**: p. 88.
54. Wikipedia. *Mutation*. September 18, 2015; Available from: https://en.wikipedia.org/wiki/Mutation.
55. Lundstrom, K. and M.P. Turpin, *Proposed schizophrenia-related gene polymorphism: expression of the Ser9Gly mutant human dopamine D3 receptor with the Semliki Forest virus system.* Biochem Biophys Res Commun, 1996. **225**(3): p. 1068-72.
56. Li-Wan-Po, A., et al., *Pharmacogenetics of CYP2C19: functional and clinical implications of a new variant CYP2C19*17.* Br J Clin Pharmacol, 2010. **69**(3): p. 222-30.
57. Wikipedia. *Nonsense-mediated Decay*. September 20, 2015; Available from: https://en.wikipedia.org/wiki/Nonsense-mediated_decay.
58. Chang, Y.F., J.S. Imam, and M.F. Wilkinson, *The nonsense-mediated decay RNA surveillance pathway.* Annu Rev Biochem, 2007. **76**: p. 51-74.
59. Ni, J.Z., et al., *Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay.* Genes Dev, 2007. **21**(6): p. 708-18.
60. Frischmeyer, P.A. and H.C. Dietz, *Nonsense-mediated mRNA decay in health and disease.* Hum Mol Genet, 1999. **8**(10): p. 1893-900.
61. Ciszkowski, C., et al., *Codeine, ultrarapid-metabolism genotype, and postoperative death.* N Engl J Med, 2009. **361**(8): p. 827-8.
62. Gasche, Y., et al., *Codeine intoxication associated with ultrarapid CYP2D6 metabolism.* N Engl J Med, 2004. **351**(27): p. 2827-31.
63. Rieder, M.J., et al., *Effect of VKORC1 haplotypes on transcriptional regulation and*

warfarin dose. N Engl J Med, 2005. **352**(22): p. 2285-93.

64. Aithal, G.P., et al., *Association of polymorphisms in the cytochrome P450 CYP2C9 with warfarin dose requirement and risk of bleeding complications.* Lancet, 1999. **353**(9154): p. 717-9.
65. FDA. *Full Prescribing Information*. 2012; Available from: http://www.accessdata.fda.gov/drugsatfda_docs/label/2012/021513s010lbl.pdf.
66. McDonough, C.W., et al., *Atenolol induced HDL-C change in the pharmacogenomic evaluation of antihypertensive responses (PEAR) study.* PLoS One, 2013. **8**(10): p. e76984.
67. Park, E., et al., *RNA editing in the human ENCODE RNA-seq data.* Genome Res, 2012. **22**(9): p. 1626-33.
68. Kim, U., et al., *Molecular cloning of cDNA for double-stranded RNA adenosine deaminase, a candidate enzyme for nuclear RNA editing.* Proc Natl Acad Sci U S A, 1994. **91**(24): p. 11457-61.
69. Kumar, M. and G.G. Carmichael, *Nuclear antisense RNA induces extensive adenosine modifications and nuclear retention of target transcripts.* Proc Natl Acad Sci U S A, 1997. **94**(8): p. 3542-7.
70. Wagner, R.W., et al., *A double-stranded RNA unwinding activity introduces structural alterations by means of adenosine to inosine conversions in mammalian cells and Xenopus eggs.* Proc Natl Acad Sci U S A, 1989. **86**(8): p. 2647-51.
71. Bahn, J.H., et al., *Accurate identification of A-to-I RNA editing in human by transcriptome sequencing.* Genome Res, 2012. **22**(1): p. 142-50.
72. Genomes Project, C., et al., *A map of human genome variation from population-scale sequencing.* Nature, 2010. **467**(7319): p. 1061-73.
73. Li, M., et al., *Widespread RNA and DNA sequence differences in the human transcriptome.* Science, 2011. **333**(6038): p. 53-8.
74. Pickrell, J.K., Y. Gilad, and J.K. Pritchard, *Comment on "Widespread RNA and DNA sequence differences in the human transcriptome".* Science, 2012. **335**(6074): p. 1302; author reply 1302.
75. Lin, W., et al., *Comment on "Widespread RNA and DNA sequence differences in the human transcriptome".* Science, 2012. **335**(6074): p. 1302; author reply 1302.
76. Kiran, A. and P.V. Baranov, *DARNED: a DAtabase of RNa EDiting in humans.* Bioinformatics, 2010. **26**(14): p. 1772-6.
77. Ramaswami, G. and J.B. Li, *RADAR: a rigorously annotated database of A-to-I RNA editing.* Nucleic Acids Res, 2014. **42**(Database issue): p. D109-13.
78. Picardi, E., et al., *ExpEdit: a webserver to explore human RNA editing in RNA-Seq*

experiments. Bioinformatics, 2011. **27**(9): p. 1311-2.

79. Picardi, E. and G. Pesole, *REDItools: high-throughput RNA editing detection made easy.* Bioinformatics, 2013. **29**(14): p. 1813-4.
80. Distefano, R., et al., *VIRGO: visualization of A-to-I RNA editing sites in genomic sequences.* BMC Bioinformatics, 2013. **14 Suppl 7**: p. S5.
81. Wang, K., M. Li, and H. Hakonarson, *ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data.* Nucleic Acids Res, 2010. **38**(16): p. e164.
82. Trapnell, C., L. Pachter, and S.L. Salzberg, *TopHat: discovering splice junctions with RNA-Seq.* Bioinformatics, 2009. **25**(9): p. 1105-11.
83. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.
84. Danecek, P., et al., *The variant call format and VCFtools.* Bioinformatics, 2011. **27**(15): p. 2156-8.
85. Li, H., *Tabix: fast retrieval of sequence features from generic TAB-delimited files.* Bioinformatics, 2011. **27**(5): p. 718-9.
86. Langdon, W.B., *Performance of genetic programming optimised Bowtie2 on genome comparison and analytic testing (GCAT) benchmarks.* BioData Min, 2015. **8**(1): p. 1.

논 문 초 록

1. 영문요약(영문초록)

Abstract

Personal pharmacogenomics : pharmacological approach based on biological functional element and genomic variant

LEE SOO-YOUN

자연과학대학 생물정보협동과정

The Graduate School

Seoul National University

The main objective of a pharmacogenomics study, particularly those that pertain to personal drug prescription, is to prevent adverse drug reactions and harness the maximum clinical benefits by analyzing genetic variations known to control pharmacodynamic

effects such as drug efficacy, dose requirements, adverse events, etc. Until recently, population-based observational studies were the norm where genetic information of case and control groups, recruited based on altered drug reaction, is investigated to identify less than 10 variants that are critical in controlling drug reactions. The U.S. Food and Drug Administration (FDA) recommends that drugs be prescribed accordingly to the study results. Despite successful applications of these results, population-based observational studies are hindered by: immense budget issues in developing a standard set; results affected by demographic conditions; rare or private variants unable to be included in the results; etc.

In 2013, in an attempt to overcome such issues, we developed the PharmSafe algorithm, which calculates gene and drug scores based on an individual's genetic variation information and ranks drugs that are possibly hazardous. Performance of the algorithm was evaluated using PharmGKB. However, the algorithm only considers variants within the coding region and of all biological knowledge, only protein-protein interaction is applied within the algorithm.

In this paper, we developed a new and improved PharmSafe algorithm where variants from non-coding region and biological functional element including regulational factor such as RNA editings are also considered and seven biological, pharmacological, and statistical knowledge elements are used as weight parameters. In addition to the aforementioned updates, we were able to validate the reproducibility of the PharmSafe algorithm in larger genome

datasets by using 1000 Genome Project Phase 3 data, which expands the preexisting datasets to 2,503 samples.

Of the pharmacodynamic genes, the new PharmSafe algorithm was able to achieve the highest area under the curve (AUC) of 0.5857~0.6502, (0.6224±0.222; minimum ~ maximum, mean±standard deviation) in catalytic enzyme genes. In drug class evaluations, antihypertensives (n=22) had the highest AUC of 0.6234~0.8896 (0.7340±0.0539; minimum ~ maximum, mean±standard deviation).

We believe that the new PharmSafe algorithm would be a valuable tool for a clinical decision support system (CDSS) in prescribing drugs safely and efficiently at the right dosages based on an individual's genetic variation information.

Keyword: Personal Genome, Pharmacogenomics, Pharmacology, Personalized Medicine, Personalized Pharmacogenomics, RNA editing

Student ID : 201030127

이학박사학위논문

개인별 맞춤 약물유전체 : 생물학적 기능 요소와 개인유전체 변이를 기반으로한 약물학적 접근

Personal pharmacogenomics : pharmacological approach based on biological functional element and genomic variant

2016년 2월

서울대학교 대학원

자연과학대학 생물정보협동과정

이 수 연

논 문 초 록

1. 국문요약(국문초록)

요약(국문초록)

개인별 맞춤 약물 처방을 위한 약물 유전체(Pharmacogenomics) 연구는 각 개인별 약물효능(efficacy), 필요용량(dose requirements), 이상반응(adverse events)등의 약물학적 반응을 조절하는 유전적 변이 정보를 고려하여 약물을 처방함으로써 약물부작용(ADR;Adverse Drug Reaction)을 방지하고 치료효과를 극대화하는 데 그 목적이 있다. 최근까지는 특정 약물반응성에 차이를 보이는 실험군과 대조군(case-control)을 모집하고 모집된 환자들의 변이 정보를 전장유전체분석연구(GWAS; Genome Wide Association Studies) 기법을 사용하여 해당 약물의 반응성을 조절하는 10개 미만의 변이를 찾아내는 인구기반의 관찰연구(Population-based observational studies)가 주를 이루었으며, 미국 식약청(FDA; Food and Drug Administration)에서는 약물 처방 시에 이러한 연구결과들을 고려하여 처방하도록 권고하고 있다.

이러한 성공적인 연구결과들과 실제 적용사례에도 불구하고 인구기반의 관찰연구는 표본형성에 필요한 막대한 비용, 인구통계학적 요인/조건(Demographic condition)에 영향을 받는 연구결과, 그리고 희귀변이(Rare variant) 혹은 개인변이(Private variant)가 연구결과에 포함될 수 없다는 점 등의 많은 한계점을 드러냈다.

2013년 우리는 이러한 문제점을 극복하고자 개인의 유전적 변이 정보를 바탕으로 유전자, 약물 점수를 계산하여 개인에게 위험한 약물의 순위를 제공하는 PharmSafe 알고리즘을 개발하고 PharmGKB를 사용하여 알고리즘의 성능을 평가하였다. 하지만 암호영역(Coding region)의 변이만 사용하였다는 점과 생물학적 지식(Biological knowledge)은 단백질-단백질 상호작용(PPI;Protein-protein interaction)만 적용했다는 점을 한계점으로 제시하였다.

본 논문에서는 변이로 인하여 단백질의 기능을 변화시키는 RNA 편집 싸이트를 검출 할 수 있는 도구를 제작하고, 비암호영역(non-coding region)의 변이 정보를 사용함은 물론, RNA 편집 싸이트 등과 같이 변이에 의해 조절되는 유전자 정보를 비롯한 7가지 생물학, 약리학, 통계학적 지식요소를 가중치로 사용하여 개선된 PharmSafe 알고리즘을 개발하였으며 1092명의 개인유전체 데이터를 2503명(1000 Genome Project Phase 3)으로 확장함으로써 더 많은 개인의 유전체 데이터에서의 PharmSafe 알고리즘의 재현성 또한 입증하였다. 그 결과 7가지 지식요소를 반영한 개선된 PharmSafe 알고리즘에서는 에서는 약물학적 유전자 종류 중 약물을 분해하는 효소를 가중치로 사용한 알고리즘에서 AUC 0.5857~0.6502, (0.6224±0.222;최솟값~최댓값, 평균±표준편차)로 가장 높은 평과 결과를 얻었으며 약물 군별 평가에서는 혈압강하(Antihypertensives; n=22)군에서AUC 0.6234~0.8896(0.7340±0.0539;최솟값~최댓값, 평균±표준편차)로 가장 높은 평가 결과를 얻었다. 개선된PharmSafe 알고리즘은 환자들의 유전적 변이 정보를 바탕으로 의학적 의사결정 지원시스템(CDSS;Clinical Decision Support System)에서 각 환자별로 위험요소가 적은 약물을 알맞은 농도로 처방받는데 매우 유용하게 쓰일 것이다.

주요어 : 개인유전체, 약물유전체,약물학,맞춤의학, RNA 편집 싸이트

학 번 : 201030127

목 차

1 장 개인별 맞춤 약물유전체 : 개인유전체 변이를 기반으로한 약물학적 접근

소 개

미국 오바마(Barack Obama) 대통령은 2015년 1월 신년국정연설 당시, 2016년부터 미화 215 million달러(2150억원)를 개인맞춤의학(personalized medical treatments)을 위한 연구에 투자하겠다고 발표했다. 환자 개개인의 특성이 아닌 환자군의 평균적 특성을 고려한 기존 의학치료(medical treatments)의 만병통치약(one-size-fits-all-approach) 적인 접근 방법은 특정 개인에게는 유용한 치료이나 그렇지 못한 개인이 다수 발생하는 결과를 초래 했다고 지적하며, 이러한 문제점을 극복하기 위해 이번 대규모 투자를 통해 지놈(genome)을 이용한 생물-의학(bio-medical)연구, 특히 개인의 유전체 정보를 바탕으로한 개인맞춤형 약물처방 연구분야를 활성화할 것이며 나아가 이러한 시도가 앞으로 시행될 개인맞춤의학의 핵심이 될 것이라고 언급했다. 이처럼 현재 맞춤의학에 대한 연구 패러다임은 크게 변화하고 있으며 지놈을 이용한 맞춤의학의 연구적, 경제적 중요성이 크게 대두되고 있다.

맞춤의학(personalized medicine)이라는 컨셉은 1960대부터 사용되었고, 용어는 1999년부터 처음 사용되었다[1]. 이 용어는 개인을 위한 맞춤 의료서비스(healthcare)를 할 수 있는 의학적 모델을 제시하는 것으로 정의되었지만 보다 실질적인 연구 측면에서 중요한 의미는 개인의 약물유전체(pharmacogenomics)정보를 바탕으로 한 개인별 맞춤 약물처방을 위한 의학적 모델을 연구하고 제시하는

것이었다. [2]. 약물유전체를 바탕으로 한 맞춤의학 연구의 목적은 약물부작용(ADR;Adverse Drug Reaction)을 최소화하고 치료효과를 극대화(maximize therapeutic benefit)하며 비용 절감을 통한 경제적 효과를 얻는 것이다[2].

미국에서는 매년 2,000,000(6.7%) 이상의 입원 환자가 약물부작용으로 인한 심각한 손상을 겪고 있으며, 이 중 100,000(50%)의 환자는 치명적인 손상을 겪은 것으로 보고되었다[3]. 또한 주요사망원인 중 약물부작용이 4~6 순위를 차지할 정도로 빈도가 높고 심각한 문제이며[3, 4], 그로 인한 경제적 손실 또한 미국 $137-177 billion ,독일 €434 million, 영국GBP£2 billion으로 매우 크다 [5-7]. 제약 산업계에서도 약물부작용은 주요 부담으로 작용하고 있다. 미국에서는 실제로 약물부작용으로 인해 1990년부터 2012년까지 43개의 약물이 허가취소(withdraw)되었고[8], 캐나다 보건복지부 발표에 따르면 새로 승인된 약물의 50%에서 심각한 약물부작용사례가 보고되었고 이 중 95%가 시장 출시 이후에 발견되었다고 한다 [5].

이러한 약물부작용의 발생을 줄이는데 가장 좋은 대안으로 떠오른 분야는 각 개인별 약물효능(efficacy), 필요용량(dose requirements), 이상반응(adverse events)등의 약물학적 반응을 조절하는 유전적 변이를 비롯한 유전학적 요소를 찾는 약물유전체 연구이다 [2]. 개인 지놈에서 나타나는 변이들의 약 2-90%가 약물반응성 차이를 설명한다고 알려져 있다 [9, 10]. 이러한 약동학(Pharmacokinetics), 약력학 (Pharmacodynamics)과 연관된 변이들은 약물 수송체(drug transporter), 인간주조직적합성항원(HLAs;Human-Leukocyte Antigens), 약물 대사효소(drug metabolizing enzyme) 유전자에 존재하며 이 유전적 변이들이 약물투여량(drug dose), 약물의 혈장농도(drug plasma levels)등에 영향을 미쳐 약물반응성을 조절하여 임상적으로 예후가 좋은 치료효과를 나타내기도 하고 심각한 약물

부작용을 초래하기도 한다 [2, 11, 12]. 이렇게 유전적 변이 혹은 유전자발현으로 인해 약물의 반응성이 조절받는 대표적인 약물로는 트라스투주맙(trastuzumab)과 아바카비어 (abacavir)가 있다. 트라스투주맙(Trastuzumb)은 유전자발현으로 인해 약물반응성이 조절받는 대표적인 약물로 HER2(ERBB2) 양성 전이성 유방암 환자(HER2-positive breast cancer)의 항암요법에 사용된다. HER2가 과발현(over expression)된 후기(late-stage) HER2(ERBB2) 양성 전이성 유방암 환자에게 트라스투주맙을 투여할 경우 그렇지 않은 환자보다 중간생존시간(median survival time)이 20.3개월 대 25.1개월로 높게 나타났다. [13, 14]. 아바카비어(Abacavir)는 유전적 변이에 의해 약물의 반응성이 조절받는 대표적인 약물로 인간 면역 결핍 바이러스 감염(HIV type 1) 환자 중 인간주조직적합성복합체 유전자의 5701번째 위치에 변이(Human Leukocyte Antigen-B*5701 allele) 를 가지고 있는 환자의 48-61%가 아바카비어(abacavir)에 과민반응(hypersensitivity)을 보였다 [15]. 이러한 연구결과를 바탕으로 FDA(Food and Drug Administration)에서는 아바카비어(abacavir)의 라벨에 해당 변이에 대한 문구를 표시하도록 권고하였으며 [16] 비슷한 사례의 약물 100개 이상에 대해서도 유전변이정보를 라벨에 표기하도록 권고하였다 [17]. 1950년 근육이완제인 염화석시닐콜린(suxamethonium chloride)과 N-acetyltransferase 효소에 의해 대사되는 약물의 반응에 영향을 미치는 변이를 찾아내는 연구를 시작으로 약물반응성에 영향을 미치는 변이를 찾는 연구가 시작되었고[18] 1990년대 SNP array 기술을 거쳐 2010년 차세대 시퀀싱(NGS;Next-generation sequencing) 기법이 도입되면서 폭팔적으로 증가하게 되었다.

미화 약 30억 달러의 비용과 10년의 시간을 들여 시행된 인간게놈프로젝트(HGP;Human Genome Project)에 의해 2백만개에

달하는 인간의 유전변이가 밝혀졌고 그로부터 10년 후 차세대 시퀀싱(NGS;Next-generation sequencing)기법의 발달로 약 18,837달러의 비용과 열흘의 시간으로 한 개인의 전체 유전체를 해독할 수 있게 되었다. 그 후 차세대시퀀싱(NGS;Next-generation sequencig)기술은 다양하게 발전하여 2015년 4월 4,211달러의 비용으로 한 개인의 전체 유전체를 해독하게 되었다 [19]. 차세대 시퀀싱(NGS;Next-generation sequencing)기술의 발달에 따른 가파른 비용 감소로 인해 대량의 개인유전체를 분석하기 위한 HapMap Consortium, 1000 Genomes project, TCGA등의 컨소시움들이 만들어지고 데이터가 생성되었다. 2008년에 시작된 1000 genome project는 전장유전체 시퀀싱(Whole genome sequencing)을 이용해 14개의 인구집단(populations)에 속하는 1,092명의 개인유전체 데이터를 생성하였고 그 선행연구(pilot study)가 2010년에 완성되고 공개되었다. 그 결과 38,000,000개의 단일염기다형성변이 (SNPs;Single nucleotide polymorphisms), 1,400,000개의 짧은삽입(short insertions)과 결손(deletions), 그리고 14,000개의 큰 결손(larger deletions)을 밝혀냈으며 특히 한 개인당 평균 250-300개의 기능상실변이(loss-of-function variant)를 가진다는 것이 보고되었다 [20]. 현재까지 1000 genome project는 27개의 인구집단으로부터 추출한 2,503명의 개인유전체 데이터를 공개했다. 이러한 대규모 개인 유전체 데이터들은 약물유전체 연구에도 사용되었고 Hapmap 데이터로부터 얻은 본래의 변이 정보를 재검토(evaluate)하여 약물 관련 유전자(Pharmacogene), 변이(Pharmacovariant) 정보를 담은 데이터 베이스들이 새로 출시되었다. 특히 2009년 4월 약물유전학적 지식베이스(Knowledge base)인 PharmGKB는 35개의 HapMap CEU 샘플과 26개의 HapMap YRI 샘플로부터 38개의 약물유전학적 후보 유전자(Pharmacogenetic candidate genes)를 선정하고 해당 프로젝트를 Very Important

Pharmacogenes(VIP)으로 명명하였으며, 약물유전학과 관련된 일배체형(Haplotypes), 스플라이싱변이(Splicing variants), 그리고 해당 변이들이 포함된 유전자들의 정보등을 제공하였다 [21].

하지만 1000 Genome project를 비롯한 대량의 개인 유전체 데이터가 개인의 표현형(Phenotype)정보를 포함하고 있지 않기 때문에 이러한 데이터들을 약물유전체 연구에 적용함에 있어 큰 제약점이 되었고, 이로 인하여 개인의 유전체 정보는 물론 약물학적 표현형정보를 가지고 있는 데이터의 필요성이 크게 제시 되었다. 이로인해 Pharmacogenomics Research Network(PGRN)등의 컨소시움이 생성되었다.

약물유전체를 바탕으로 한 약물부작용(ADR) 예측 연구는 대부분 약물 반응성의 차이를 보이는 실험군과 대조군(case-control)을 모집하고 이 환자들의 혈액을 채취하여 마이크로어레이(microarray)또는 차세대시퀀싱(NGS)기법을 이용하여 개인의 유전체 서열을 해독한 후 이 데이터를 전장유전체분석연구(GWAS;Genome-wide association study)를 통해 약물 반응성을 조절하는 하나 또는 다수의 변이를 찾는 것이 대표적이다. 전장유전체분석연구(GWAS;Genome-wide association study)는 2005년 “Common disease, common variant” 라는 가설 [22] 을 바탕으로 소개되었으며 질병(Disease), 약물반응성(Drug response)등 인간에게서 나타나는 의학적 표현형(clinical phenotype)을 유발하는 원인이 되는 단일염기다형성변이(SNPs)를 검출해내는 강력한 도구로 널리 쓰여왔다. 특히 차세대시퀀싱기술(NGS)의 발달로 전장유전체분석연구(GWAS)가 폭팔적으로 증가하였으며, 그 결과 2015년 현재 638개의 연구가 GWAS catalog에 등록되었으며 9450개의 단일염기다형성변이(SNP)마커가 등록되어 있다 [23]. 전장유전체분석연구(GWAS)의 장점은 첫째 일반적으로 혈통, 가계도를 사용하는 기존 유전학적 연구에 비해 환자집단을 모으기가 수월하다는

점이다. 두번째로는 기존에 사용하던 연관성연구들(Linkage studies)에 비해 작은 유전효과(genetic effects)를 검출해 내는데 높은 통계적 힘(statistical power)을 가진다. 왜냐하면 기존에 사용하던 연관비평형(LD;linkage disequilibrium)방법은 10kb(kilobases)에서부터 몇 Mb(megabases) 범위의 유전체안의 표현형을 유발하는 변이를 검출 할 수 있는데 반해 전장유전체분석연구(GWAS)는 국소적으로 미세한 범위에 해당하는 변이의 검출이 가능하기 때문이다 [24]. 약물유전체(Pharmacogenomics)를 대상으로 하는 전장유전체분석연구(GWAS)에서 약물반응성에 크게 영향을 미치는 변이를 검출하기 위해서는 충분한 표본 크기(sample size), 치료계획(treatment protocol), 복용량(dosage), 자발적 부작용 보고여부, 인종정보(Ethnicity)등을 포함하는 환자특성 (patient features)을 만족하는 표본 형성과 모집된 환자 표본집단의 특성을 고려한 실험디자인등의 조건이 만족될 때 좋은 결과를 얻을 수 있다 [25].

2011년 7월총 48개의 약물반응성을 조절하는 변이를 검출한 전장유전체분석연구(GWAS)가 NHGRI GWAS Catalog에 등록되었으며 이중 대표적인 연구는 항혈소판제인 와파린(Warfarin), 클로피도그렐(Clopidogrel), 타목시펜(Tamoxifen)에 관한 연구이다. 특히 VKORC1과 CYP2C9 변이들이 와파린(Warfarin)의 복용량을 조절한다는 것은 전장유전체분석연구(GWAS)기법을 사용해 약물반응성에 영향을 미치는 변이를 밝혀낸 가장 대표적인 사례라고 할 수 있다. [26]. 유사한 인구통계학적 요인/조건(Demographic condition)의 CYP2C9*1/*1 유전형을 가지는 환자 49명에서의 안전한 약물효과를 위한 일일 평균 와파린용량은 7.9mg이었으나, CYP2C9*1/*3 유전형을 가지는 환자 10명에서 요구되는 일일 평균

와파린용량은 2.2mg으로 유전형에 따라 요구되는 일일 평균 와파린용량의 차이가 있다고 보고되었다 [27]. 또한 기존 학술문헌을 이용한 메타분석(Meta-analysis)에 의하면 CYP2C9의 유전형은 와파린용량 가변성(Warfarin dose variability)의 약 12%를 설명하며, VKCOR1 유전형은 약 25%를 설명 할 수 있다고 한다 [28]. 이러한 연구결과를 바탕으로 미국 FDA(Food and Drug Administration)에서는 초기 와파린용량(Warfarin dose) 결정에 있어 CYP2C9과 VKCOR1의 유전형 검사를 권장하고있다. 구체적인 예로는 항혈소판제인 클로피도그렐(Clopidogrel)을 복용한 환자 중에서 CYP2C19*17 유전형을 가지는 환자는 활성을 나타내는 클로피도그렐 대사체가 과다하게 변환되기 때문에 출혈(bleeding)의 위험성이 높다는 보고가 있으며 [29], 타목시펜(Tamoxifen)을 CYP2D6*4 동형접합성 유전형(homozygous)을 가지는 유방암 환자에게 투약하는 경우 무재발 생존비율(Disease-free survial)이 낮게 나타난다고 하는 보고가 있다 [30].

이러한 좋은연구결과들과 광범위한 사용에도 불구하고 최근 약물유전체를 대상으로 한 전장유전체분석연구(Pharmacogenomic GWAS)에 대한 많은 한계점이 제기되고 있다. 전장유전체분석연구(GWAS)에서 중요한 요소 중 하나는 표본크기(sample size)이다. 약물유전체를 대상으로 한 전장유전체분석연구(Pharmacogenomic GWAS)에서는 이 점이 중요한 잠재된 제한점으로 작용한다. 그 이유는 일반적으로 약물을 복용한 500명 중 1명의 비율로 심각한 약물 부작용(Drug adverse reaction)이 발생한다고 알려져 있다 [31]. 일반적으로 전장유전체분석연구(GWAS)에서 사용하는 실험군과 대조군 연구는 당뇨병과 같이 흔히 나타나는 질병(common disease)에 걸린 사람들을 대상으로 표본을 구성하는 것에 비해 약물유전체를 대상으로 한 전장유전체분석연구(Pharmacogenomic GWAS)는 특정 질병에 걸린

사람들 중 해당 약물을 복용했을 때 심각한 약물부작용을 겪은 사람들을 모집해야 하기 때문에 표본형성에 큰 어려움이 있다. 예를 들어 높은 콜레스테롤 수치의 환자나 고혈압 환자는 모집하기가 비교적 수월하나 높은 콜레스테롤 수치를 가진 환자 중 스타틴(Statins)에 반응하지 않는 환자나 고혈압 환자 중 베타차단제(beta blocker)에 반응성을 보이지 않는 환자군은 그 수가 확연히 작기 때문에 일정한 수치만큼 환자들을 모집하기가 상대적으로 매우 어렵다 [32]. 이러한 이유로 약물유전체를 대상으로 한 전장유전체분석연구(Pharmacogenomic GWAS)의 표본 크기는 작을 수 밖에 없고 이로 인하여 통계적 힘(Statistical power)이 작아지기 때문에 정확하게 약물반응성을 조절하는 변이를 검출하는 것이 어려우며 또한 이를 검증(Validation)하고 재현(Replication)하는 것 또한 어렵게 된다. 뿐만 아니라 현재 약물유전체를 대상으로 한 전장유전체분석연구(Pharmacogenomic GWAS)를 위해 모집된 표본의 인종은 유럽인이거나 유럽계 미국인이 대부분이고 아시아인 이나 아프리카인을 대상으로 한 연구는 매우 드물다. 따라서 기존연구에서 검출된 대부분의 변이들은 해당 민족 이외의 다른 민족이나 인종에게 적용하는것이 위험하다. 이는 개인에서도 동일한 문제점으로 작용한다. 1000 Genome Project에서 밝혀졌듯이 한 개인당 평균 250-300개의 기능상실변이(loss-of-function variant)가 존재하는데 이를 같은 결과로 해석하는 것은 아주 큰 문제가 있다. 약물유전체를 대상으로 한 전장유전체분석연구(Pharmacogenomic GWAS)의 또 다른 한계점은 1000 Genome Project 데이터를 비롯한 대규모 차세대 시퀀싱(NGS) 데이터에서 대량으로 나타난 대립유전자형빈도(MAF;Minor allele frequency) 5% 미만인 희귀변이(Rare variant)와 한 개인에게서만 나타나는 변이(Private variant)를 검출해내기 어렵다는 점이다. 실제로 여러 희귀변이(Rare variant)들이 알츠하이머병(Alzheimer's disease)을 비롯한 여러 질병을 유발하는 원인이 되는 변이(Causal

variant)라는 것은 물론, 항암제인 메토트렉세이트(Methotrexate)를 비롯한 여러 약물들의 약물반응성을 조절하는 원인변이라는 것이 밝혀져있다 [33-35]. 약물반응성을 조절하는 원인이 되는 희귀변이(Rare variant) 혹은 개인변이(Private variant)를 검출하기 위해서는 그에 맞는 큰 표본집단이 있어야 하는데 이는 현실적으로 한계가 있기 때문에 전장유전체분석연구(GWAS)를 이용하게 되면 통계적인 힘(Statistical power)이 떨어질 수 밖에 없으며 [25] 검출된 희귀변이의 신뢰성(Reliability) 또한 떨어지게 된다. 앞서 언급한 수많은 전장유전체분석연구(GWAS)의 한계점에도 불구하고 기존의 거의 모든 약물유전체 연구들이 전장유전체분석연구(GWAS)를 사용해 약물반응성을 조절하는 변이를 찾는 연구를 시행해 왔다.

2013년 우리는 인구기반의 관찰연구(population-based observational studies)인 전장유전체분석연구(GWAS)의 단점을 극복하고 개인의 희귀변이(Rare variant)와 개인변이(Private variant)를 포함한 유전적 변이정보를 바탕으로 각 개인별 위험한 약물부작용(ADR)을 예측하여 임상에서 개인마다 안전한 약물을 선택하여 처방할 때 사용될 수 있는 정보를 제공하는 개인 약물유전학(“personal pharmacogenomics”)의 개념을 제시하고 이를 실현하는 알고리즘인 PharmSafe를 개발하였다(reference). PharmSafe는 개인의 유전체 서열을 입력값으로 하고 약동학적(PK;pharmacokinetics), 약력학적(PD;pharmacodynamics) 으로 영향을 받는 유전자와 연결된 모든 약물에 대하여 유해한 정도를 점수로 제공한다. 낮은 개인의 PharmSafe 약물 점수(personalized PharmSafe score)는 개인이 해당 약물을 복용시 부작용(ADR)이 발생할 수 있는 확률이 증가한다는 것을 의미한다. 따라서 낮은 점수를 가진 약물을 개인이 복용할 때에는 복용량(dosage)을 조절하거나 복용하지 않는 것을 권고한다. PharmSafe 알고리즘은 크게 3단계로 구성되어 있다. 첫번째는

입력값으로 받은 유전체 서열에 나타난 변이들에 SIFT(Sorting Intolerant From Tolerant)점수 [36] 를 사용하여 변이로 인한 아미노산 치환이 단백질 기능에 영향을 미치는 정도를 변이점수(Variant score)로 나타낸다. 두번째는 각 변이별 점수를 해당 변이들이 포함되는 유전자 별로 기하평균(Geometric mean)을 사용하여 요약(Summarize)하고 이를 유전자점수(Gene score)라고 명명하였다. 세번째는 DrugBank [37], KEGG drug [38] 등에 포함된 약물과 유전자 사이의 관계를 이용하여 유전자점수(Gene Score)와 같이 기하평균(Geometric mean)을 사용하여 약물점수(Drug score)를 계산한다. 앞서 언급한 바와 같이 변이점수(Variant score), 유전자점수(Gene score), 약물점수(Drug score) 모두 낮을 수록 유해함을 나타낸다. 앞선 연구에서는 입력값으로는 1092명의 개인 유전체를 포함한 1000 Genome Project phase 1 데이터를 사용했고 정답값(Gold standard)으로는 대표적인 약물유전학 지식베이스인 PharmGKB [39] 의 정보를 이용하여 AUC(Area under curve)를 계산함으로써 PharmSafe 알고리즘의 효과를 입증하였다. PharmSafe의 알고리즘 성능을 평가하기 위하여 두가지 방법을 채택하였다. 첫번째는 인종정보없이 평가한 비인종평가 (Ethnicity-non-specific validation), 두번째는 1000 Genome Project 데이터가 포함하고 있는 4가지 인종정보(AFR;African, AMR;American, ASN;Asian, EUR;European)를 사용한 인종별평가(Ethnicity-specific validation)이다. 497개의 약물에 대하여 PharmSafe 알고리즘을 계산한 결과, 인종별 평가(Ethnicity-specific validation)에서는 0.662±0.081 (평균±표준편차,0.637~0.742), 비인종별 평가(Ethnicity-non-specific validation)에서는 0.633±0.038(평균±표준편차, 0.622~0.642)의 AUC값을 얻었다. 인종별 평가, 비인종별 평가를 비교했을 때 인종별 평가에서 비인종별 평가에 비해 우월한 AUC

점수를 보여 PharmSafe의 효과를 확실하게 증명하였다. 하지만 앞선 연구에서 입력값으로 모두 암호영역(Coding region)의 변이만을 사용했다는 점과 생물학적 지식(Biological knowledge)은 단백질-단백질 상호작용 (PPI;Protein-protein interaction)만 적용했다는 점을 한계점으로 제시하였다.

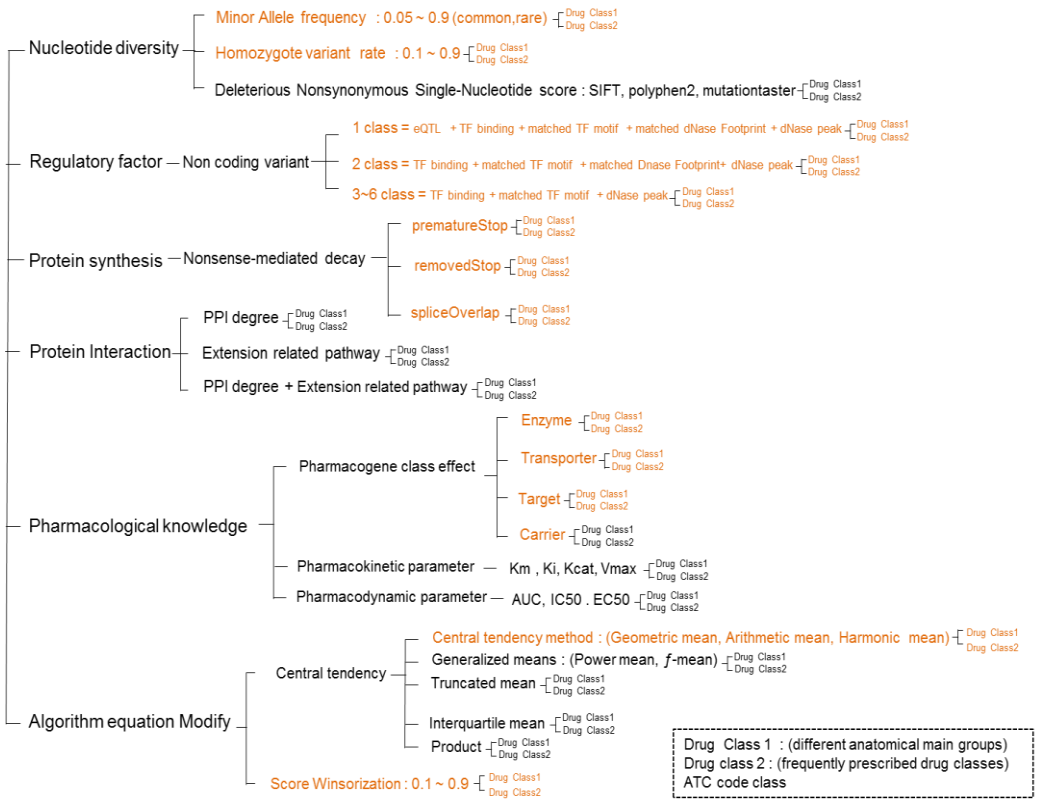


그림 0 생물학적, 약리학적, 통계학적 요소

따라서 본 연구에서는 앞선 연구의 한계점을 극복하고자 비암호영역(Non-coding region)의 변이 정보를 비롯한 8개의 생물학적(변이빈도;Variant frequency, 동형접합체 변이 비율;Homozygote variant rate, 대립유전자형빈도;Minor allele frequency, 넌센스-매개 전사체 붕괴;Nonsense-mediated decay), 약리학적(약물학적유전자; Pharmacogene class), 통계학적

지식(중심경향성;Central tendency method, 변이점수윈저화; Variant score winsorization)을 사용함은 물론, 해당 지식 최상의 조합을 찾아 알고리즘의 성능을 향상 시키고자 하였다(그림 0). 또한 1092명의 개인유전체 데이터를 2503명(1000 Genome Project Phase 3)으로 확장함으로써 더 많은 개인의 유전체 데이터에서의 PharmSafe 알고리즘의 재현성 또한 입증하고자 한다.

방 법

개인별 지놈 및 약물 데이터

입력값으로 사용한 개인별 지놈 데이터는 1000 Genome 프로젝트 [40] (2015년 6월 기준)로 부터 다운로드 받았다. 데이터는 JPT(Japanese in Tokyo, Japan) 104명, BEB(Bengali from Bangladesh) 86명을 포함한 전체 2504명으로 구성되어 있으며 작게는 26개의 부분 인종군(Subpopulation), 크게는 AFR(African), AMR(Admixed American), EAS(East Asian), SAS(South Asian), EUR(European)의 5개 인종군(Super population)으로 구분되어 있다(보충 표 1, 보충 그림 1) [20, 41].

KEGG drug [38] 와 Drug bank 4.0 [37] 로 부터 약물정보(drug information), 와 약물관련 유전자인 표적(Target), 수송체(transporter), 효소(enzyme), 수송기구(carrier)에 대한 정보를 수집하였다(자세한 추출 과정은 Pharmsafe1의 *"Drugs, drug-related genes, and drug-gene association"* 참조). 약물 분류군에 대한 정보는 ATC(Anatomical Therapeutic Chemical Classification System)와 자주 처방되는 약물분류(15 most frequently prescribed drug classes)를 사용하였다. ATC는 WHOCC [42] 로부터 다운로드 받았으며 이 중 14 해부학적 주요 그룹(Anatomical main groups)에 대한 정보를 추출하여 사용하였다. 자주 처방되는 약물 분류는 National Center for Health Statistics [43] 로 부터 다운로드 받아 사용하였다(보충 표 2 와 3).

생물학적 지식 정보 데이터

알고리즘 향상을 위해 비암호영역의 변이(Noncoding variant)등을 포함한 7가지 생물학, 약리학, 통계학적 지식정보를 각 지식베이스로부터 다운로드 받거나 혹은 데이터로부터 추출하여 사용하였다. 입력값으로 받은 유전체 서열에 포함된 변이로 인한 아미노산 치환이 단백질 기능에 영향을 미치는 정도를 점수로 나타내기 위해 SIFT [44] 를 사용하였다 [36]. 변이빈도(Variant frequency), 동형접합체변이 비율(Homozygote variant rate), 대립유전자형빈도(Minor allele frequency)는 1000 Genome 데이터로부터 추출하였다. 수집한 약물 데이터로부터 497개의 약물, 4226개의 약물-유전자 연관정보(drug-gene relations)를 추출하였으며 약물 관련 유전자에 대해서는 표적(Target) 440개, 수송체(transporter) 54개, 효소(enzyme) 74개, 수송기구(carrier) 10개로 총 545개를 추출하였다. 조기 번역정지(Prematurestop), 정지코돈삭제(removed stop), 이어맞추기위치변이(splice overlap)를 포함하는 난센스-매개 전사체 붕괴(NMD; Nonsense-Mediated mRNA Decay)를 유발하는 변이 정보를 Variant Annotation Tool [45] 을 사용하여 얻은 후 SIFT 점수가 있는 개인별 1000 Genome 데이터에 매핑하여 사용하였다 [46]. 디엔에이가수분해효소 과민반응 위치(DNase hypersensitivity site), 전사인자의 결합부위(binding sites of transcription factors), 촉진제 위치(promoter regions)등을 포함하는 전사 조절기전(regulation transcription)에 영향을 미치면서 유전자간부위(intergenic region)에 속하는 비암호영역의 변이(Noncoding variant)정보를 사용하기 위하여 전사조절 영향력에 대해 7단계로 분류, 제공하는 Regulomedb[47]로부터 19,493개의 유전자와 관련된 26,561,892개의 변이 그리고 99,845,325개의 유전자와 변이의 연관정보(gene-variant relation)를 다운받았다. 다운받은 변이 정보를 SIFT 점수가 있는 개인별 1000 Genome

데이터에 매핑하여 사용하였다. [48].

생물학적 지식을 사용한 변이, 유전자, 약물점수

개인의 유전체 서열에 나타난 변이들로부터 각 변이가 속한 유전자 그리고 해당 유전자가 영향을 미치는 약물의 유해한 정도를 정량화 하기 위해 우리는 기존 PharmSafe 논문을 통해 변이점수(variant score), 유전자점수(gene score), 약물점수(drug score)라고 명명한 세 단계의 점수를 고안해내고 해당 점수를 사용하여 개인의 유해한 약물순위를 예측하는 PharmSafe 알고리즘을 개발하고 그 유용성을 증명하였다 [49]. 본 논문에서는 앞서 발표한 PharmSafe 알고리즘의 성능을 향상시키고자 비암호영역(Non-coding region)의 변이 정보를 비롯한 8개의 생물학적(변이빈도;Variant frequency, 동형접합체 변이 비율;Homozygote variant rate, 대립유전자형빈도;Minor allele frequency, 넌센스-매개 전사체 붕괴;Nonsense-mediated decay), 약리학적(약물학적유전자;Pharmacogene class), 통계학적 지식(중심경향성;Central tendency method, 변이점수윈저화; Variant score winsorization)요소들을 사용하였으며 해당 지식 정보들을 입력값으로 사용하여 최상의 조합을 찾아 알고리즘의 성능을 향상 시키고자 하였다. 앞으로 열거될 모든 수식에서 S_{vi} , S_{gj} , S_{dk} 는 각각 변이 점수, 유전자 점수, 약물 점수를 의미하며 모든 수식에 사용된 기호는 보충 표 4을 참조하면 된다.

변이, 유전자, 약물점수

변이, 유전자, 약물점수는 앞서 발표한 Baik et al의 Pharmsafe

알고리즘을 사용하였다[50]. Pharmsafe의 알고리즘은 다음과 같다. 첫번째로 변이점수(Variant score)는 입력값으로 받은 개인의 유전체 서열에서 나타난 변이 정보를 담은 파일인 VCF(Variant Call Format)에 포함된 변이로 인한 아미노산 치환이 단백질 기능에 영향을 미치는 정도를 점수로 나타내기 위해 SIFT를 사용하였다[36]. 계산에 사용된 변이는 비동의성 암호영역 변이(non-synonymous coding variant)로써 변이 i 의 변이점수를 S_{Vi} 로 정의하고 SIFT 알고리즘의 점수를 사용하였다. 수식은 아래와 같다.

$$S_{Vi} = \mathrm{SIFT}(V_i)$$

변이점수는 0 부터 1 사이의 범위를 가지며 점수가 낮을수록 해당 변이로 인한 아미노산 치환이 단백질기능에 유해한(deleteriousness)영향을 미치는 것으로 해석된다. 유전자점수는 유전자 j 에 속하는 비동의성 암호영역 변이들의 변이점수를 기하평균을 이용하여 합산하며 이를 S_{g_j} 로 정의하였다. G_j 는 유전자 j 에 속하면서 변이점수를 가지는 동의성 암호영역 변이들의 집합을 의미한다. 수식은 아래와 같다.

$$s_{g_j} = \begin{cases} 1 & if\ |G_j| = 0 \\ \left(\prod_{v_i \in G_j} s_{v_i} \right)^{1/|G_j|} & if\ |G_j| > 0 \end{cases}$$

약물점수는 약물 k 의 약물반응성에 영향을 미친다고 알려진 약동학(PK;Pharmacokinetics), 약력학(PD;Pharmacodynamics)적 유전자들의 유전자점수를 기하평균을 이용하여 합산한 점수이며 이를 S_{dk} 라고 정의하고 수식은 아래와 같다.

$$s_{d_k} = \left(\prod_{g_j \in D_k} s_{g_j} \right)^{1/|D_k|}$$

D_k 는 약물 k 와 약동학, 약력학적 연관성이 있으면서 유전자 점수를 가지는 유전자들의 집합을 의미한다. 유전자점수와 약물점수 또한 0 부터 1 사이의 범위를 가지며 유전자점수는 낮을수록 해당 유전자가 약동학, 약력학적 기능을 수행함에 있어 유해한 영향을 미쳐 이로 인하여 약물반응성이 영향을 받는다고 해석되고 약물점수 역시 낮을수록 개인이 해당 약물을 복용했을 때 부작용를 비롯한 유해한 약물반응성을 보일 것이라고 해석된다. 이러한 약물점수는 복용할 약물 선택시 중요한 기준으로 사용될 것이다.

중심 경향 방법(central tendency method)

모집단으로부터 얻어진 자료를 살펴보면 특정값으로 몰리는 현상을 보이는데 이를 중심경향(central tendency)이라고 하고 해당 특정값을 중심경향값(central tendency value)이라고 한다. 이러한 중심경향을 나타내는 값은 평균(Mean), 중앙값(Median), 최빈값(Mode)등이 대표적이다. 특히 평균이 가장 많이 쓰이는데 평균값을 구하는 방법은 산술평균(Arithmetic mean), 기하평균(Geometric mean), 조화평균(Harmonic mean)등 7가지가 있다[51]. 기존의 PharmSafe 알고리즘에서는 유전자점수와 약물점수 합산시 기하평균(Geometric mean)을 사용하였으나 본 연구에서는 변이점수는 앞선 방법과 동일하게 계산하고 유전자점수와 약물점수는 중심 경향 방법에 대표적인 산술평균(Arithmetic mean), 조화평균(Harmonic mean) 그리고 곱(Product)를 사용하여 계산하고 그 결과를 기하평균(Geometric

mean)과 비교하였으며 각각의 방법에 따른 수식은 아래와 같다.

산술평균의 유전자점수 및 약물점수 수식

$$s_{g_j} = \begin{cases} 1 & if\ |G_j| = 0 \\ \left(\frac{1}{n} \sum_{v_i \in G_j} s_{v_i} \right) & if\ |G_j| > 0 \end{cases} \quad , \quad s_{d_k} = \left(\frac{1}{n} \sum_{g_j \in D_k} s_{g_j} \right)$$

조화평균의 유전자점수 및 약물점수 수식

$$s_{g_j} = \begin{cases} 1 & if\ |G_j| = 0 \\ \left(\frac{n}{\frac{1}{n} \sum_{v_i \in G_j} s_{v_i}} \right) & if\ |G_j| > 0 \end{cases} \quad , \quad s_{d_k} = \left(\frac{n}{\frac{1}{n} \sum_{g_j \in D_k} \frac{1}{s_{g_j}}} \right)$$

곱의 유전자점수 및 약물점수 수식

$$s_{g_j} = \begin{cases} 1 & if\ |G_j| = 0 \\ \left(\prod_{v_i \in G_j} s_{v_i} \right) & if\ |G_j| > 0 \end{cases} \quad , \quad s_{d_k} = \left(\prod_{g_j \in D_k} s_{g_j} \right)$$

약물학적 유전자 종류(Pharmacogene type)

약물의 반응성은 약물의 흡수, 분포, 생체내 변화 및 배설을 포함하는 약동학(PK;pharmacokinetics)적 작용기전과 생체에 대한 약물의 생리학적, 생화학적 작용기전을 나타내는 약력학(PD;Pharmacodynamics)적 작용기전에 따라 달라진다. 약물의 표적(Target)이 되는 유전자가 약동학적 작용기전을 조절하며 약력학적

작용기전은 약물수송체(transporter), 약물분해효소(enzyme), 약물수송기구(carrier) 유전자에 의해 영향을 받는다. 이러한 약물학적 유전자 종류에 따라 약물의 반응성이 달라 질 수 있음을 이용하여 기존 PharmSafe 알고리즘에 각각의 약물학적 유전자 종류에 속하는 변이에 가중치를 달리하여 적용하였다. 표적, 수송체, 효소, 수송기구 별로 특정 변이 i 가 유전자 j 의 영역에 포함되면 해당 변이점수를 제곱하여 가중치점수(Weight score)를 계산하였으며 이를 가중치 변이점수(Weighted variant score)로 명명하고 W_{v_i} 로 정의하였다. 수식은 아래 명시하였다. 약물학적 유전자 종류별 계산에서 만약 변이 i 가 해당 약물학적 유전자 종류의 유전자 영역에 포함되지 않거나 어떠한 약물학적 유전자의 영역에도 포함되지 않았다면 가중치를 가하지 않은 기존의 변이점수를 그대로 사용하였다. V_{PGT} 는 약물학적 유전자의 영역 안에 속하는 변이점수를 가진 변이들의 집합이다. 변이점수 및 가중치 변이점수의 수식은 아래와 같다.

$$S_{v_i} = SIFT\left(v_i\right)$$

$$W_{v_i} = \begin{cases} \left(S_{v_i}\right)^2 & if\ S_{v_i} \in V_{PGT} \\ S_{v_i} & else \end{cases},$$

$$V_{PGT} = \left\{ v_i \mid v_i \in Target / Transporter / enzyme / carrier\ gene\ region \right\}$$

유전자점수와 약물점수는 가중치 변이점수를 입력값으로 받아 위와 같이 기하평균을 사용하여 계산하였으며 수식은 아래와 같다.

$$s_{g_j} = \begin{cases} 1 & if\ \left|G_j\right| = 0 \\ \left(\prod_{v_i \in G_j} w_{v_i} \right)^{1/\left|G_j\right|} & if\ \left|G_j\right| > 0 \end{cases}, \quad s_{d_k} = \left(\prod_{g_j \in D_k} s_{g_j} \right)^{1/\left|D_k\right|}$$

변이 점수 윈저화(Variant score winsorization)

윈저화(Winsorization)는 이상치(Outlier)에 의해 왜곡되는 영향을 줄이기 위해 최극단에 속하는 값들을 특정 값으로 변환(Transformation)하여 사용하는 통계적인 기법이다[52]. PharmSafe 알고리즘에서는 0 부터 1 까지의 범위를 갖는 SIFT 점수를 변이점수로 사용하고 있으며, SIFT에서는 0.05이하의 점수를 가지는 변이에 대해 유해하다고(deleterious) 정의 하고 있다. 하지만 실제로 0.05 이상의 점수를 가지는 변이들에 대해서는 유해한 정도에 대한 언급을 하지 않고 있다[44]. 예를들어, SIFT점수 0은 매우 유해하고 1은 전혀 유해하지 않다고 정의 되어있기 때문에 0값을 가지는 변이와 1값을 가지는 변이 사이의 유해한 정도의 차이는 크다고 할 수 있지만 0.7값을 가지는 변이와 0.8값을 가지는 변이 사이의 유해한 정도의 차이는 가늠하기가 상당히 어렵다. 따라서 점수 구간 사이의 유해한 정도 차이가 거의 없어 예측 결과에 잡음(noise)을 유발하는 점수 구간의 절단점(cut-off point)을 찾아 윈저화(Winsorization) 방법을 통해 예측 결과의 잡음을 제거하고자 하였다. 이를 위해 0 부터 1 까지의 점수를 0.1, 0.2, … , 0.9 와 같이 0.1 간격으로 10개의 윈도우로 나누고 이를 SR 이라고 명명하였다. 각 윈도우별로 절단점(cut-off point) 이상의 값들은 변이점수 1로 변환하였다. 예를들어 0.7 윈도우에서는 0.7 이상의 변이점수들은 1로 변환하였다. 변이점수와 윈저화 변이점수 수식은 아래와 같다.

$$S_{v_i} = SIFT\left(v_i\right)$$

$$WS_{v_i} = \begin{cases} S_{v_i} & if\ S_{v_i} \le SR_l \\ 1 & else \end{cases}, \qquad SR = \{0.1 \sim 0.9\}$$

유전자점수는 윈저화 변이점수를 입력값으로 받아 위와 같이 기하평균을 사용하여 계산하였으며 약물점수는 앞서 계산된 유전자점수를 사용하여 계산하였다. 수식은 아래와 같다.

$$s_{g_j} = \begin{cases} 1 & if\ |G_j| = 0 \\ \left(\prod_{v_i \in G_j} WS_{v_i} \right)^{1/|G_j|} & if\ |G_j| > 0 \end{cases}, \quad s_{d_k} = \left(\prod_{g_j \in D_k} s_{g_j} \right)^{1/|D_k|}$$

낮은 대립형질 빈도(Minor Allele Frequency)

인구집단에서 낮은 대립형질 빈도(MAF)가 5%이하인 변이들에 대하여 Hapmap project를 통해 희귀변이(Rare variant)로 정의되었고, 그 후 1000 Genome project를 통해 한 개인당 평균 250-300개의 기능상실변이(loss-of-function variant)를 가지며 그 중 10~20개가 희귀변이라는 것이 밝혀졌다. [20]. 또한 이러한 희귀변이들이 알츠하이머, 파킨슨병을 비롯한 많은 질병을 유발하는 원인변이라는 것이 밝혀지고 이러한 연구들이 잃어버린 유전 가능성(missing heritability)의 많은 부분을 설명한다고 보고되었다[53]. 하지만 희귀변이가 약물의 반응성에 영향을 미친다는 연구는 이제 시작단계에 있으며 아직 희귀변이가 특정 약물의 반응성을 조절한다는 연구결과는 발표되지 않았다(2015년 8월 기준). 우리는 희귀변이가 약물 반응성에 미치는 영향을 반영하고자 기존의 Pharmsafe 알고리즘에 낮은 대립형질 빈도(MAF)정보를 반영하였다. 1000 Genome 데이터로부터 낮은 대립형질 빈도(MAF)정보를 추출하고 0부터 0.01까지의 빈도범위에 대하여 0.001의 간격을 적용하여 10개의 윈도우로 나누어 이를 ***MAFR***이라고 명명했다. 각 윈도우별로 해당 윈도우 이하의 낮은

대립형질 빈도값을 가지는 변이들에 대하여 기존의 변이점수를 제곱하여 가중치 변이점수를 계산하였으며 수식은 아래와 같다. V_{MAFR} 는 l 번째 $MAFR$에 해당하는 낮은 대립형질 빈도값을 가지는 변이들의 집합이다.

$$S_{v_i} = SIFT(v_i)$$

$$W_{v_i} = \begin{cases} (S_{v_i})^2 & if\ MAF_{v_i} \in V_{MAFR_m} \\ S_{v_i} & else \end{cases}, \quad \begin{aligned} MAFR &= \{0.001 \sim 0.009, 0.01 \le MAF_{v_i}\} \\ V_{MAFR} &= \{v_i \mid \mathrm{MAF}_{v_i} \in MAFR_m\} \end{aligned}$$

앞서 만든 가중치 유전자 점수를 사용해 유전자 점수를 계산하고 이를 이용해 약물 점수를 계산하였으며 수식은 아래와 같다.

$$s_{g_j} = \begin{cases} 1 & if\ |G_j| = 0 \\ \left(\prod_{v_i \in G_j} w_{v_i} \right)^{1/|G_j|} & if\ |G_j| > 0 \end{cases}, \quad s_{d_k} = \left(\prod_{g_j \in D_k} s_{g_j} \right)^{1/|D_k|}$$

동형접합변이 비율(Homozygote mutation rate)

대립유전자(allele)의 변이(mutation)에 따른 유전형(genotype)은 크게 동형접합야생형(homozygous wild type), 이형접합변이형(heterozygous mutant type), 동형접합변이형(homozygous mutant type)의 3가지 종류로 나뉠 수 있다. 동형접합야생형은 대립유전자 양쪽 모두 변이가 없는 경우를 의미하고 이형접합변이형은 한쪽에만 변이가 있는 경우를, 동형접합변이형은 양쪽 모두 변이가 있는 경우를 의미한다[54]. 일반적으로 희귀질환에서는 질병을 유발하는 열성돌연변이(recessive mutation)가 동형접합변이형인 경우 높은 질병발생률을 보인다고

알려져 있다. 약물유전체 연구에서도 동형접합변이로 인해 약물반응성이 달라지는 연구결과들이 보고되어 있으며, 대표적으로는 Ser9 유전자의 이형접합변이형 및 동형접합변이형이 동형접합야생형에 비하여 도파민 D3 수용체의 선택적 리간드인 GR99841의 결합능력을 증가시킨다는 것이 그 예이다[55]. 또한 클로피도그렐을 복용한 사람중 CYP2C19*17 동형접합변이형을 가진 사람이 그렇지 않은 사람에 비하여 혈소판 응집이 과도하게 일어났음이 보고되었다[56]. 이러한 연구결과를 바탕으로 우리는 전체 인구에서 동형접합변이의 비율이 높은 변이일수록 약물반응성에 미치는 영향이 클 것이라고 가정하고 1000 Genome 데이터로부터 동형접합변이의 비율을 추출하고 이를 Pharmsafe 알고리즘에 반영하여 계산하였다. 0.1부터 0.9까지의 동형접합변이 비율을 0.1의 간격을 적용하여 9개의 윈도우로 나누고 이를 ***HVFR*** 이라고 명명했다. 각 윈도우별로 해당 윈도우에 속하는 동형접합변이 비율을 가지는 변이들에 대하여 기존의 변이점수를 제곱하여 가중치 변이점수를 계산하였으며 수식은 아래와 같다. HV_{HVFR} 는 l 번째 ***HVFR*** 에 해당하는 동형접합변이비율을 가지는 변이들의 집합이다.

$$S_{v_i} = SIFT(v_i)$$

$$w_{v_i} = \begin{cases} (S_{v_i})^2 & if\ HVF_{v_i} \in HV_{HVFR_m} \\ S_{v_i} & else \end{cases} \qquad \begin{aligned} HVFR &= \{0.1 \sim 0.9\} \\ HV_{HVFR} &= \{v_i \mid \mathrm{HVF}_{v_i} \in HVFR_m\} \end{aligned}$$

앞서 만든 가중치 유전자 점수를 사용해 유전자 점수를 계산하고 이를 이용해 약물 점수를 계산하였으며 수식은 아래와 같다.

$$s_{g_j} = \begin{cases} 1 & if\ |G_j| = 0 \\ \left(\prod_{v_i \in G_j} w_{v_i} \right)^{1/|G_j|} & if\ |G_j| > 0 \end{cases}, \quad s_{d_k} = \left(\prod_{g_j \in D_k} s_{g_j} \right)^{1/|D_k|}$$

mRNA 안정성 조절 기전 (Nonsense-mediated mRNA decay; NMD)

난센스-매개 전사체 붕괴(NMD;Nonsense-mediated mRNA decay)는 유전자 발현 조절 기전 중 mRNA의 질적 조절(quality control)에 관여하는 가장 대표적인 기작이다[57]. 난센스-매개 전사체 붕괴는 유전자 내에 조기종결코돈(premature stop codons)이 포함되어 mRNA가 정상길이의 단백질보다 짧은 단백질을 생성하게 되고 유전자가 본래의 기능을 상실하고 해롭게(deleterious) 또는 이롭게(gain-of-function) 변형되거나 비정상 유전자에서 생산된 물질이 정상 유전자에서 생산된 물질과 결합하여 정상 유전자의 기능마저 비정상으로 만드는 우성-음성(Dominant negative)효과를 나타내는 기작이다[58]. 최근에는 조기 번역정지(Prematurestop), 정지코돈삭제(removed stop), 이어맞추기위치변이(splice overlap) 등의 기작 또한 난센스-매개 전사체 붕괴의 원인이 됨이 밝혀졌다[59]. 본래 난센스-매개 전사체 붕괴는 세포의 정상적인 기능을 위해 비정상적인 유전자를 제거하는 기전이지만 특정 유전적인 변이에 의해 비정상적으로 발생한 난센스-매개 전사체 붕괴로 인해 암 또는 유전질환등의 다양한 질병이 발생한다. 대표적인 예로 베타글로빈(β-globin) 유전자의 상위영역(upstream)에 존재하는 변이들로 인해 헤모글로빈의 합성이 감소되어 발생하는 혈액 질환인 베타 탈라세미아가

있다[60]. 우리는 변이에 의해 발생되는 난센스-매개 전사체 붕괴의 생물학적 영향을 PharmSafe 알고리즘에 반영하기 위해 Variant Annotation Tool(VAT)을 사용하여 난센스-매개 전사체 붕괴에 관여한다고 알려진 변이들의 정보를 추출했다[46]. 특정 변이 i 가 조기번역정지(Premature stop), 정지코돈삭제(removed stop), 이어맞추기위치변이(splice overlap) 중 하나라도 관여하는 변이라면 해당 변이의 변이점수를 제곱하고 그렇지 않으면 기존의 변이점수를 그대로 사용하여 가중치 변이점수(W_{v_i})를 계산하였다. V_{NMD} 는 난센스-매개 전사체 붕괴에 관여하며 변이점수를 가진 변이들의 집합이다. 변이점수 및 가중치 변이점수 수식은 아래와 같다.

$$S_{v_i} = SIFT\left(v_i\right)$$

$$w_{v_i} = \begin{cases} \left(S_{v_i}\right)^2 & if\ v_i \in V_{NMD_m} \\ S_{v_i} & else \end{cases}$$

$$NMD = \left\{ \text{Pr } emature\ Stop, \text{Remove Stop, Spicingover, PrematureStrop\& Removestop, PrematureStop\& RemoveStop\& Splicingover} \right\}$$

$$V_{NMD} = \left\{ v_i \mid \text{v}_i \in premature / removestop / splicingover\ \text{var}\, iant \right\}$$

앞서 만든 가중치 유전자 점수를 사용해 유전자 점수를 계산하고 이를 이용해 약물 점수를 계산하였으며 수식은 아래와 같다.

$$s_{g_j} = \begin{cases} 1 & if\ \left|G_j\right| = 0 \\ \left(\prod_{v_i \in G_j} w_{v_i} \right)^{1/|G_j|} & if\ \left|G_j\right| > 0 \end{cases} \quad , \quad s_{d_k} = \left(\prod_{g_j \in D_k} s_{g_j} \right)^{1/|D_k|}$$

비암호 변이를 포함한 유전자 기능 조절 변이 (Regulatory variants including noncoding region variants)

기존 PharmSafe 논문에서 가장 큰 한계점은 변이점수로 사용한 SIFT가 암호영역의 변이(coding variant)만 포함하고 있어 유전자간 영역(IGR;Intergenic region), 비발현부위(intron)등을 포함한 비암호영역의 변이(non-coding region)들을 반영 할 수 없었다는 것이었다. 이러한 한계점을 극복하고자 이번 논문에서는 비번역 변이들의 정보를 PharmSafe 알고리즘에 반영했다. 비암호영역, 암호영역변이들 중 유전자의 기능을 조절하는 변이들의 정보를 RegulomeDB[48]에서 다운로드 하였다. RegulomeDB는 특정 변이가 디엔에이가수분해효소 과민반응 위치(DNase hypersensitivity site), 전사인자의 결합부위(binding sites of transcription factors), 촉진제 위치(promoter regions)등을 포함하는 전사 조절기전(regulation transcription)에 영향을 미치는 전사조절 영향력에 대해 1a부터 6까지의 14 단계로 나타냈다. 1a 로 갈수록 변이 i 가 여러가지의 조절기작에 관여하는 것으로 정의하였다. 다운로드 받은 데이터는 각 변이의 염색체상의 위치정보, 해당 변이에 의해 전사조절이 되는 유전자의 목록, 그리고 앞서 언급한 RegulomeDB의 전사조절영향력 단계정보로 구성되어 있었다. 우리는 여러 유전자의 전사조절에 영향을 미치는 변이가 유해(deleterious)하다고 가정하고 RegulomeDB로부터 얻은 1부터 6까지 총 6단계별로 해당 정보를 활용하여 PharmSafe 알고리즘을 다음과 같이 변형하여 계산하고 검증하였다. 유전자 기능 조절변이 RV_i에 의해 조절기작이 영향을 받는 유전자 G_j를 RG_j 라고 정의하고 RG_j 의 조절 기작에 관여하는 변이의 갯수를 세어

가중치점수인 WS_{RG_j} 를 계산했다. 기존의 유전자점수 대신 WS_{RG_j} 를 가중치로 하여 가중기하평균(weighted geometric mean)을 통해 가중 유전자점수 WS_{g_j} 를 계산하고 이를 이용하여 약물점수를 계산하였으며 수식은 아래와 같다.

$$RC = \{class1 \sim class6, sum\,of\ class\} \quad , \quad RV_i = \{v_i \mid v_i \in RC_m \,\&\, v_i \in RG_j\}$$

$$S_{v_i} = SIFT(v_i) \quad , \quad WS_{RG_j} = \#RV_i$$

$$ws_{g_j} = \begin{cases} 1 & if\,|G_j| = 0 \\ \left(\prod_{v_i \in RG_j} v_i^{WS_{G_j}}\right)^{1 / \sum_{n=1}^{n} WS_{G_j}} & if\,|G_j| > 0 \end{cases} \quad , \quad s_{d_k} = \left(\prod_{g_j \in D_k} s_{g_j}\right)^{1/|D_k|}$$

각 요소별 최적의 조건 선정 및 조합 실험(Combination test)

앞서 6개의 생물학적(동형접합체 변이 비율;Homozygote variant rate, 대립유전자형빈도;Minor allele frequency, 넌센스-매개 전사체 붕괴;Nonsense-mediated decay), 약리학적(약물학적유전자;Pharmacogene class), 통계학적 지식(중심경향성;Central tendency method, 변이점수윈저화; Variant score winsorization) 요소들을 가중치로 사용하여 계산한 결과를 바탕으로 각 요소별 가장 높은 평가점수(AUC)를 가지는 7가지 조건을 선별하였다. 선별방법은 기하평균의 인종평가 AUC값(0.6076, SEA),

비인종평가 AUC값(0.6271, SNA)을 기준으로 두고 각 요소 E 에 속한 l 번째 조건 E_{C_l} 의 인종평가 AUC값($EA_{E_{C_l}}$), 비인종평가 AUC값($NA_{E_{C_l}}$)을 아래 수식과 같이 인종, 비인종평가 별로 각각의 기준 AUC값에서 요소 E 에 속한 l 번째 조건 E_{C_l} 의 AUC값을 뺀 $EAD_{E_{C_l}}$, $NAD_{E_{C_l}}$ 를 계산한 후 두 값의 평균을 취해 $MAD_{E_{C_l}}$ 을 계산하였다.

$$EAD_{E_{C_l}} = SEA - EA_{E_{c_i}},\ NAD_{E_{C_l}} = SNA - NA_{E_{c_i}}$$

중심경향값을 제외한 6가지 요소별로 가장 높은 $MAD_{E_{C_l}}$ 를 가지는 조건 한가지를 선택하여 총 6가지 요소별 조건을 선정하였다(보충표 7). 선택된 요소별 조건은 변이점수 윈저화(SW)는 0.2, 낮은 대립형질 빈도(MAF)는 0.007, 약물학적 유전자 종류(PGT)에서는 효소, 동형접합변이 비율(HR)은 0.7, mRNA 안정성 조절 기전(NMD)에서는 조기 종결코돈, 유전자 기능 조절변이(RV)에서는 4단계를 선택하였고 이렇게 선정된 6가지 요소별 조건들의 모든 58가지 조합에 대하여 PharmSafe 알고리즘을 적용하여 계산하였다. 선택된 요소별 조건들의 집합을 SE_C 라 정의하고 m번째 조합에 속하는 변이 i 를 $CB_{m_{v_i}}$, 이 변이들의 집합을 CB_{m_v} 로 정의했다. 변이 i 가 $CB_{m_{v_i}}$ 인 경우 변이점수에 제곱을 하고 그렇지 않으면 기존의 변이점수를 그대로 사용하여 가중치 변이점수(W_{v_i})를 계산하였다. 각 변이 별로 하나 이상의 $CB_{m_{v_i}}$ 에 속하는 경우 변이 i 의 가중치 변이점수(W_{v_i})를 곱하여 조합 가중치점수(CBW_{v_i})를 계산하여 유전자점수의 입력값으로 사용하였으며 수식은 아래와 같다.

$$S_{v_i} = SIFT(v_i)$$

$$W_{v_i} = \begin{cases} (S_{v_i})^2 & if\ v_i \in CB_{m_{v_i}} \\ S_{v_i} & else \end{cases},$$

$$SE_C = \{SW(0.2), MAF(0.007), PGT(enzyme), HR(0.7), NMD(PS\ \&\ \mathrm{RV}), RV(4class)\}$$
$$CB_{m_v} = \{v_i \mid v_i \in {}_{SE_C} C_r\}$$
$$r = \{1,2,3,4,5,6\}$$

$$CBW_{v_i} = \tilde{\bigcirc}_{v_i \hat{1} C_{v_i}} W_{v_i}$$

앞서 만든 조합 가중치 유전자 점수를 사용해 유전자 점수를 계산하고 이를 이용해 약물 점수를 계산하였으며 수식은 아래와 같다.

$$s_{g_j} = \begin{cases} 1 & if\ |G_j| = 0 \\ \left(\prod_{v_i \in G_j} CBW_{v_i}\right)^{1/|G_j|} & if\ |G_j| > 0 \end{cases}, \quad s_{d_k} = \left(\prod_{g_j \in D_k} s_{g_j}\right)^{1/|D_k|}$$

PharmGKB를 이용한 Pharmsafe 알고리즘의 성능 평가

앞서 발표한 PharmSafe 논문[50]과 같은 방법으로 본 논문에서도 PharmGKB를 사용하여 생물학적 지식 정보를 가중치로 사용한 개선된 PharmSafe 알고리즘의 성능평가를 시행하였다. PharmGKB는 전장유전체분석연구(GWAS) 결과 중 약물학적 지식과 관련된 원인변이와 약물 관계정보를 휴먼 큐레이션(Human curation)을 통해 검증하여 데이터베이스화 한 지식베이스이다. 2015년 1월 23일 PharmGKB [39]로부터 약물 392개, 변이 1176개를 포함한 3248개의

약물반응성을 조절하는 원인변이와 약물 관계정보를 다운로드 받았다. 이 중 앞서 선정한 497개의 약물에 해당하는 약물 290개, 변이 840개를 포함하는 1807개(55.63%)의 원인변이와 약물 관계정보를 추출하였다. 추출된 1807개의 변이가 속하는 유전자는 471개였으며 이는 표적(Target) 225, 수송체(transporter) 18, 효소(enzyme) 20, 수송기구(carrier) 7, 비약물학적 유전자 201개로 구성되어져 있었다(보충그림 6). 인종정보를 이용한 검증을 위해 PharmGKB로부터 흑인 혹은 아프리카계 미국인 관련 관계정보 83개, 아시아인 관련 관계정보 329개, 백인 관련 관계정보 647개, 기타 인종관련 6개, 인종정보가 없는 관계정보 1997개의 원인변이와 약물 관계정보를 받아 이 중 앞서 선정한 497개의 약물에 해당하는 흑인 혹은 아프리카계 미국인 관련 관계정보 58개(69.87%), 아시아인 관련 관계정보 329개(63.88%), 백인 관련 관계정보 451개(69.70%), 기타 인종관련 관계정보 5개(83.33%), 인종정보가 없는 관계정보 1128개(56.48%)의 약물, 원인변이 그리고 인종의 관계정보를 추출하였으며 PharmGKB에 대해서는 아프리카(Black or African American), 아시아(Asian), 미국 그리고 유럽(White)으로 인종을 재분류하고 1000 Genome 데이터에 대해서는 아프리카(AFR;African), 유럽(EUR;European), 아시아(EAS;East Asian, SAS;South Asian), 미국(AMR;Ad Mixed American)으로 분류하여 사용했다(보충표5). 위의 정보를 1000 Genome 데이터에 매핑한 결과, 한 개인당 평균 647.5±73.14(평균±표준편차,474~891)개의 약물과 원인변이 관계를 가지고 있었다(보충 그림4).

PharmSafe 알고리즘의 결과인 약물점수를 바탕으로 각 개인 별 위험약물 순위를 매겨 PharmGKB와 대조하였을 때 위험순위가 높은 약물들이 실제 환자군을 대상으로 한 실험에서 밝혀진 PharmGKB의 위험 약물 목록과 일치되고 위험하지 않은 약물들은 일치되지 않는다면

PharmSafe 알고리즘이 개인의 변이정보를 사용하여 해당 개인에게 위험한 약물과 그렇지 않은 약물을 구별하는 능력이 높다는 것을 증명할 수 있다는 가정을 바탕으로 아래와 같이 두 가지 데이터를 구성하여 민감도(sensitivity)와 특이도(specificity)를 계산하고 검증하였다. 사용된 첫번째 데이터는 497개의 약물정보를 사용하여 계산한 개인별 PharmSafe 약물점수를 낮은 점수순으로 순위(rank)를 매겨 구성한 개인별 Pharmsafe 위험 약물 목록이다. 두번째로는 앞서 설명한 바와 같이 PharmGKB로부터 추출한 497개의 약물과 관련된 원인변이와 약물 관계정보를 1000 Genome 데이터에 각 개인별로 매핑하여 PharmGKB와 일치된 원인변이와 약물 그리고 인종 관계정보를 구성하고 이를 Gold Standard(GS) 데이터로 사용하였다. 각 개인별 위험약물 순위목록을 임계값(threshold)을 낮춰가며 Gold Standard 데이터와 대조하여 Gold Standard의 약물이 임계값 범위 안의 개인별 위험약물 목록과 일치하면 참값(True)으로, 불일치하면 거짓값(False)으로 구분하고 이를 이용하여 민감도(sensitivity)와 특이도(specificity)를 아래의 수식으로 계산하였다.

$$sensitivity = \frac{|D_L \cap GS|}{|GS|}\ ,\ specificity = 1 - \frac{|D_L - GS|}{|D - GS|}$$

D 는 497개의 약물들의 집합을 의미하고, D_L 은 임계값 L 범위안의 약물들의 집합을 의미한다. 우리는 앞서 추출한 인종정보를 반영(Ethnic specific)하거나 혹은 반영하지 않는(Non-ethnic specific) 방법으로 알고리즘 평가점수인 AUC를 계산하여 성능을 검증하였다. 인종정보를 사용한 검증의 경우는 앞선 방법과 같이 임계값 범위 안의 개인의 위험약물 목록과 인종정보를 Gold Standard의 약물과 인종정보와 비교하여 약물과 인종 모두 일치하면 참값(True)으로, 불일치하면

거짓값(False)으로 판별하였으며 해당 과정을 각 4가지 인종별로 진행하였다.

결 과

생물학적, 통계학적 , 약물학적 7가지 요소별 변이, 유전자, 약물 평균 점수 분포

중심경항방법, 약물학적 유전자 종류, 변이점수 원저화, 낮은 대립형질 빈도, 동형접합변이 빈도, mRNA 안정성 조절 기전, 유전자 기능 조절변이 7가지로 구성되고 각각 4~10가지 조건이 포함된 가중치 요소들을 사용하여 2504명의 1000 지놈 개인 유전체 데이터에 매핑하고 각 요소별 가중치 알고리즘에 적용하여 계산한 변이점수, 유전자 점수, 약물 점수를 2504명 각 사람별로 평균점수를 계산하여 각 요소에 포함되어 있는 조건별로 평균 개인 점수 분포를 계산하였다 (그림 1). 중심경항방법을 비롯한 7가지 가중치 요소에 포함된 4~10 가지 조건을 적용하여 계산한 각 요소의 조건별 변이점수의 개인평균은 0.591~0.853(0.652±0.048, 평균±표준편차), 50509 개의 유전자 점수의 개인평균은 0.172~0.521(0.328±0.062, 평균±표준편차), 497 개의 약물 점수의 개인평균은 0.244~0.825(0.652±0.069, 평균±표준편차) 였다. 동형접합변이 빈도가 낮을수록 유전자 점수가 낮은 것을 확인할 수 있었다. 이는 동형접합변이이면서 빈도가 낮은 희귀 동형접합 변이들 중 non-synonymous 변이가 다수 포함되어 있음을 의미하며 이는 희귀 동형접합변이가 아미노산 서열 변형을 잘 일으키고 이 작용으로 인하여 유전자의 기능 상실되어 약물 반응성에 영향을 미친다고 해석할 수 있다. 중심경향법에서 곱이 가장 낮은 유전자 점수와 약물 점수를 보였지만 곱의 특성상 유전자의 길이가 길어 변이 수가 많거나 영향을 미치는 유전자가 많이 연구되어 밝혀진 약물의 경우는 유전자나 약물의 위험도와 상관없이 유전자, 약물 점수가 낮아지기 때문에 표준화(normalization)이 되는 기하평균을 다음

계산에서는 표준 계산법으로 이용하였다.

7가지 요소별 조건을 이용한 인종, 비인종 별 Pharmsafe 알고리즘 평가

7가지 요소는 각각 중심경향방법 4(산술평균, 기하평균,조화평균, 곱), 약물학적 유전자 종류 4(표적,수송체,효소,수송기구), 변이점수 윈저화 9(0.1~0.9), 낮은 대립형질 빈도 10(0~0.1≦MAF), 동형접합변이 빈도 9(0.1~0.9), mRNA 안정성 조절 기전 5(조기종결코돈, 종결 코돈 제거, splice-overlap, 조기종결코돈 + 종결코돈제거,조기종결코동 + 종결코돈제거 + splice-overlap), 유전자 기능 조절변이 7(1~total class)가지 조건으로 구성되어 있다. 각 요소에 포함된 54가지 조건을 가중치로 사용하여 변이(n=977842), 유전자(n=50509), 약물(n=497) 점수를 1000 Genome 에 포함된 2504명 개인별로 계산하고 PhrmsGKB로부터 다운받은 유전자-변이-약물 연관과 이에 해당하는 인종정보를 사용하여 만든 GS(Gold-standard)와 함께 각 인종, 비인종 각각 조건별로 AUC를 계산하여 Pharmsafe 알고리즘의 성능을 평가 비교 하였다. 계산결과 AUC는 인종평가 0.5633 ~ 0.6436,0.6093±0.011(최솟값~최댓값,평균±표준편차), 비 인종 평가 0.5857 ~ 0.6502,0.6229±0.011(최솟값~최댓값,평균±표준편차)로 나타나 인종 평가에 비해 비 인종평가가 높은 AUC를 나타냈다(그림 2, 보충 표 6). 이렇게 비 인종 평가가 낮에 나타난 이유는 평가 기준으로 사용된 PharmGKB의 1699 개의 유전자-변이-약물 연관 중 인종 정보가 포함된 유전자-변이-약물 연관은 45.96%(781)개로 절반도 되지 않은 것이 원인이라고 예상된다. 이는 PharmGKB가 GWAS 연구의 결과를 논문들로부터 추출하여 만든 지식정보 베이스 이기 때문에 적은 수의 유전자-변이-약물 연관 정보 및 인종정보를 포함 할

수 밖에 없다. 그 이유는 GWAS연구 특성상 특정 약물 복용여부를 바탕으로 모집한 비교 대조군(case-control) 개인 유전체 서열 데이터을 대량으로 모집해야 하는데 이는 현실적으로 불가능해 많은 수의 유전자-변이-약물 연관을 밝혀 낼 수 없기 때문이다. 따라서 Pharmsafe 알고리즘의 개인별 위험한 약물을 판단하는 변별력이 떨어진다기 보다는 검증에 사용된 적은 수의 유전자-변이-약물 연관 정보가 문제가 된다. 앞으로 더 많은 인종정보가 포함된 유전자-변이-약물 연관 데이터가 생성되면 인종평가의 AUC는 물론 전체 AUC 값 또한 상승될 것이라고 예상된다. 가장 높은 AUC를 보인 조건은 인종 평가에서는 약물학적 유전자 종류의 효소(0.6436) 이었으며 비인종 평가에서는 중심경향법의 곱(0.6502) 이었으며 가장 낮은 AUC를 보인 조건은 인종 평가에서는 중심경향법의 산술평균(0.5633), 비인종 평가에서는 유전자 기능 조절변이 total class(0.5857)이었다(그림 2, 보충 표 6).

기하평균, 산술평균, 조화평균, 곱 4가지를 포함한 중심경향법의 수식 각각을 적용해 2054명의 1000 Genome 개인유전체 데이터를 사용해 Pharmsafe 알고리즘을 변형하여 계산한 AUC 결과는 인종 0.5633 ~ 0.6163, 0.5964±0.033(최솟값~최댓값,평균±표준편차), 비인종0.5935~0.6502, 0.6269±0.021(최솟값~최댓값,평균±표준편차)값을 나타냈다. 기하평균과 조화평균, 곱이 산술평균에 비해 에서 높은 AUC 값을 보였으나 각 조건의 인종간의 표준편차가 각각 인종평가 0.017, 0.048, 0.021으로 기하평균에서 가장 작은 표준편차를 보여 인종에 상관없이 높은 AUC를 나타냈다. 특히 아프리카 인종에서의 AUC는 조화평균, 곱에서 현저히 낮아졌다. 이는 조화평균이나 곱은 유전자의 길이가 길어 변이의 수가 많거나 특정 약물에 관여하는 유전자의 수가 많으면 점수가 내려가는 특성을 반영한 것으로 표준화가 되지 않아 안정된 AUC값의 표준편차가

큰 것 으로 해석할 수 있다. 이러한 이유로 다음에 하는 모든 실험에서 기준 수식으로 기하평균을 사용하였다. (그림 2, 보충 그림 S4 A).

약물학적 유전자 종류 4가지(표적, 수송체, 효소, 수송기구)에 속하는 유전자에 포함되는 변이 점수를 제곱하는 방법으로 가중 변이점수로 만들어 Pharmsafe알고리즘으로 약물 점수를 계산하고 평가한 AUC는 인종 0.5928 ~ 0.6248, 0.6055± 0.022(최솟값~최댓값,평균±표준편차) 비 인종 0.5928 ~ 0.6248, 0.6270± 0.018(최솟값~최댓값,평균±표준편차) 이었다. 인종, 비인종 모두 효소에서 가장 높은 AUC를 나타냈고 표적은 가장 낮은 AUC를 나타냈다. 이는 표적 유전자는 대부분 하나의 약물에만 영향을 미치지만 효소의 경우는 한 개의 유전자가 다수의 약물에 영향을 미친다. 따라서 같은 개수의 유전자에 변이가 생겨 기능을 상실한다 해도 표적은 한 개 혹은 두개의 약물의 영향력에 영향력을 미미치는 것 비해 효소는 다수의 약물에 영향을 미치기 때문이라고 예상된다(그림 2, 보충 그림 4 B).

잡음을 줄이거나 제거하기 위한 방법중 하나로 통계학에서 쓰이는 윈저화는 잡음을 유발하는 절단점을 찾는 것을 목표로 한다. Pharmsafe 계산시 잡음을 제거해 개인에게 위험한 약물을 구분하는 변별력을 높이기 위해 변이 점수 윈저화를 시행하였다. 0 부터 1 사이의 점수를 0.1, 0.2 … 0.9로 각각 10개의 윈도우로 나누고 각 윈도우 별로 절단점값(cut off point value) 이상의 값들은 변이 점수 1로 변환하였다. 예를 들어 0.7 윈도에서는 0.7 이상의 변이 점수들은 1로 변환하는 변이 점수 윈저화를 실시하고 Pharmsafe 알고리즘으로 약물 점수를 계산하였다. 그 결과 AUC는 인종 0.6074 ~ 0.6167, 0.6046± 0.016(최솟값~최댓값,평균±표준편차), 비인종 0.6219 ~ 0.6364, 0.6278± 0.002(최솟값~최댓값,평균±표준편차) 이었다. 인종에서는 0.3 단계가 가장 높은 AUC 0.6167 를 나타냈고 비 인종 단계에서는 0.2 단계가 가장 높은 AUC 0.6364 를 나타냈다(그림 2, 보충 그림 4

C).

낮은 대립형질 빈도는 1000 Genome에서 나타난 변이들의 낮은 대립형질 빈도 데이터를 이용하여 0.001 ~ 0.001 이상 의 10단계로 낮은 대립형질 빈도를 나누고 단계별로 m 번째 단계에 포함되는 변이 점수를 제곱하여 가중치 변이 점수를 만들어 Pharmsafe 알고리즘을 사용하여 약물 점수를 계산하였다. 그 결과 AUC는 인종 0.6008 ~ 0.6078, 0.6018± 0.016(최솟값~최댓값,평균±표준편차) , 비인종 0.6121 ~ 0..6272, 0.6260± 0.003(최솟값~최댓값,평균±표준편차) 이었다. 인종평가 에서는 0.007이 AUC 0.6078로 가장 높았으며 비인종 평가 에서는 0.005로 0.6272로 가장 높았으나 0.01 이상 단계를 제외한 모든 단계에서 비슷한 AUC 를 나타냈다 (그림 2, 보충 그림 4 D). 낮은대립형질빈도가 5%(0.05) 미만인 희귀변이(rare variant)들이Alzheimer's disease 를 비롯한 여러 질병을 유발하는 원인 변이라는 것은 물론 항암제methotrexate 를 비롯한 약물들의 약물반응성을 조절하는 원인 변이라는것이 밝혀져 있다 [33, 34] 하지만 5% 라는 기준은 Hapmap 프로젝트에서 제시한 기준으로 일부 다른 논문에서는 1% 이하를 희귀변이라고 주장한다. 또한 희귀변이가 질병에 영향을 미친다는 연구는 다수 발표되어 있으나 약물반응성에 대하여 희귀변이가 영향을 미친다는 연구는 항암제 몇종에 대한 연구 뿐이다. 따라서 약물 반응성을 조절하는 희귀변이의 절단점값은 아직 모호하다. 하지만 이번 결과로 미루어 0.01 이하의 낮은 대립형질 빈도를 가지는 변이들이 약물작용에 미치는 영향이 0.01 이상의 낮은대립형질을 가지는 변이에 비해 크다는 것으로 해석할 수 있음으로 약물반응성에 영향을 미치는 희귀변이의 절단점값은 1% 이하 라고 해석할 수 있다.

1000 Genome 에 속한 2504명의 데이터를 기준으로 변이당 동형접합변이 빈도를 계산한 후 0.1 ~0.9의 9단계로 빈도를 구분하여

개인에게서 나타난 동형접합변이가 단계별로 m 번째 단계에 속하면 변이 점수를 제곱하여 가중치 변이 점수를 만들고 이를 이용해 Pharmsafe 알고리즘으로 계산하여 약물 점수를 계산하고 평가하였다. 그결과 인종평가 에서는 0.5995 ~ 0.6073, 0.6022± 0.015(최솟값~최댓값,평균±표준편차), 비인종평가 에서는 0.6186 ~ 0.6276, 0.6268± 0.002(최솟값~최댓값,평균±표준편차) 으로 비인종평가의 0.7 단계에서만 AUC 0.6276 기준 AUC 0.6271 보다 높게 나타났다(그림림 2, 보충 그림 4 E).

mRNA 안정성 조절 기전은 조기 종결코돈(premature stop codons), 종결 코돈 제거(removed stop codons), splice-overlap 으로 총 3가지 기작으로 알려져 있다. VAT를 사용하여 종결코톤종결 등 3가지 기작을 일으키는 변이 정보를 1000 Genome 2504명에서 한번이라도 나타난 변이 목록에 매핑하였다. 3가지 기작을 조기종결코돈, 종결 코돈 제거, splice-overlap, 조기종결코돈 + 종결코돈제거,조기종결코돈 + 종결코돈제거 + splice-overlap 로 구분하여 총 5단계로 나누고 단계마다 해당 단계에 개인별로 해당 단계의 변이 정보에 해당하는 변이를 변이 점수를 제곱하여 가중치 변이 점수로 계산하고 Pharmsafe 알고리즘에 적용하여 개인당 약물별 점수를 계산하고 이를 평가하였다. 그 결과 인종평가 0.6074 ~ 0.6163, 0.6096± 0.018(최솟값~최댓값,평균±표준편차), 비인종평가 0.6271 ~ 0.6301, 0.6286± 0.002(최솟값~최댓값,평균±표준편차)이었고 인종 ,비인종 평가 모두에서 조기종결코돈이 가장 높은 AUC를 나타냈다(그림 2, 보충 그림 4 F). 종결코돈제거는 단백질의 합성시 함성을 중지하는 코돈이 상실되어 생기는 기작으로 단백질의 기능을 하는 부분의 아미노산이 대부분 상실되지 않는다. 또한 splicing-overlap 의 경우에도 단백질 합성시 일정부분의 RNA 서열이 상실되는 경우로 단백질의 기능을 하는 부분의 아미노산이 합성 될 수 도 있고 그렇지

않을 수 도 있어 실제로 단백질의 기능에 미치는 영향력이 크기 않아 많은 연구가 이루어지지 않은 기작이다. 따라서 두가지 기작으로 인하여 생성된 단백질로 인하여 약물의 반응성에 미치게되는 영향력이 크지 않거나 없을 수도 있다. 하지만 조기종결코돈 같은 경우는 아예 단백질 합성시 종결코돈이 실제보다 앞서 나타나 단백질 합성 자체가 극히 일부분만 되는 경우로 단백질의 기능 자체가 상실되는 경우가 많다고 알려져 있어 mRNA 안정성 조절 기전의 가장 대표적인 기전으로 알려져 있다. 따라서 이번 실험결과가 이러한 생물학적 기전을 반영한 결과라고 할 수 있다.

RegulomeDB 는 각 변이가 eQTL, TF binding, matched TF motif, matched DNase Footprint, DNase peak 등의 유전자의 기능을 조절하는 기작 에 관여하는 정도를 1a ~ 6의 14 단계의 점수로 나타내 제공한다. 데이터는 26,561,892 변이와 해당 변이가 기능에 영향을 주는 19,493 개의 유전자로 구성되어 있으며 이중 1.2%(n = 301,551)가 엑손 영역에 있는 변이이고 98.8%(n =26,260,341)가 인트론 및 비번역 부분에 존재하는 변이로 구성되어 있다. 기존 Pharsmafe 논문에서는 SIFT 점수가 존재하는 액손 영역의 변이만 계산영역에 넣었다. 따라서 비번역 영역의 변이는 알고리즘에 적용할 수 없어 큰 한계점으로 지적되었다. 따라서 이번 실험에서 26,260,341개의 비번역 변이를 알고리즘 계산에 추가함을써 코딩영역 변이 뿐 아니라 비번역 영역의 변이도 추가함으로써 지놈 전체 영역에서 발생하는 변이를 모두 Pharmsafe 알고리즘에 적용하여 한계점을 극복하였다. 또한 유전자 점수 계산시 기존 알고리즘 에서는 유전자 j 영역에 속하는 변이들의 점수를 기하평균을 내어 사용했다. 하지만 이번 실험에서는 변이 i 가 기능을 조절하는 유전자 j 에 해당하는 변이들의 수를 세어 유전자 j 에 가중치를 주어 가중치 유전자 점수로 사용해 DNA 영역을 벗어나 전사 단계에서 일어나는 유전자에 대한 영향력 까지 고려하여

알고리즘을 개선하였다. RegulomeDB 데이터를 다운로드 받아 1 단계부터 전 단계를 합친 전체 단계(total)등 총 7단계로 나누고 1000 Genome 2504명 유전체 데이터에 개인별 변이 목록 중 단계별로 유전자 기능 조절 단계 m에 해당하면 해당 변이 점수를 제곱하여 가중치 변이 점수를 만들어 Pharmsafe 알고리즘을 계산하여 개인별로 약물점수를 계산하고 이를 평가하였다. 그 결과 인종평가 0.5788 ~ 0.6258, 0.6037± 0.024(최솟값~최댓값,평균±표준편차), 비인종평가 0.5857 ~ 0.6189, 0.6027± 0.015(최솟값~최댓값,평균±표준편차)의 AUC 결과가 도출되었다. 인종평가에서는 4 단계가 0.6258로 가장 높은 AUC를 나타냈으며, 비인종 평가는 기준 AUC 보다 모두 낮은 AUC를 보였다(그림 2, 보충 그림 4 G).

각 요소들을 가중치로 반영해 계산한 Pharmsafe 알고리즘들이 각 약물 분류 정보군에서의 작용을 알아보기 위하여 약물 분류정보 ATC 에서 추출한 14 가지 군 그리고 WHOCC 에서 추출한 15 가지 가장 자주 처방 받은 약물 군 총 29 가지 약물 군 별로 54 가지 조건을 가중치로 적용하여 AUC 를 계산, 비교하였다. 7 개의 모든군에서 인종 비인종 모두에서 B(혈액 및 혈액 형성 기관) 0.5289 ~ 0.7727±0.0469, C(심장 혈관 시스템) 0.5076 ~ 0.6994±0.0418, G(비뇨 생식계 그리고 성 호르몬) 0.3805 ~ 0.8731±0.0685, G03(성 호르몬과 생식 시스템의 조절기) 0.7396 ~ 0.8619±0.025, C03(이뇨제) 0.4071 ~ 0.7928±0.0876, N06A(항 우울제) 0.4942 ~ 0.7364±0.0556 (최저값 ~ 최고값±표준편차)가 공통적으로 높은 AUC 를 나타냈다. 특이적으로 mRNA 안정성 조절 기전 인종평가 에서 M(근골격계)가 0.7719~0.8322±0.0256 로 높게 나타났다(그림 3, 보충 그림 4). 주로 혈관계, 비뇨생식계 그리고 성호르몬 관련 약물군들이 높은 AUC 를 나타내는 것을 확인하였다.

요소별 조건 중 최적의 조합을 사용한 가중 Pharmsafe 알고리즘 성능 평가

앞서 6개의 생물학적(Homozygote variant rate, Minor allele frequency, Nonsense-mediated decay), 약리학적(Pharmacogene type), 통계학적(SIFT Score filtering) 요소들을 가중치로 사용하여 가중 pharmsafe 알고리즘을 평가한 AUC를 기준으로 6가지 요소별 조건(변이 점수 윈저화(SW) 0.2, 낮은 대립형질 빈도(MAF) 0.007, 약물학적 유전자 종류(PGT) 효소, 동형접합변이 빈도(HR) 0.7, mRNA 안정성 조절 기전(NMD) 조기 종결코돈, 유전자 기능 조절 변이(RV) 4단계)을 선정하였다(방법 각 요소별 최적의 조건 선정 및 조합 실험 참조). 선택된 요소별 조건의 모든 56가지 조합을 입력값으로 하여 가중 Pharmsafe 알고리즘을 사용하여 1000 Genome 2504명의 개인별 약물점수를 계산하고 이를 검증하였다. 인종평가에서는 평균적으로 전체 0.6068±0.222(평균±표준편차)였고, 비인종평가 에서는 평균적으로 전체 0.6095±0.02(평균±표준편차)였다. 인종평가 비인종평가 모두에서 가장 높은 AUC 를 보였던 조합은 변이 점수윈저화 & 약물학적 유전자 타입으로 전체 AUC 0.6235, 0.6426 이었다. 이는 기준 AUC보다 0.0159, 0.015 상승하였다(그림 4, 보충 그림 8). 각 조합의 AUC 중 상위 10% 안에 든 조합 6개의 조합을 살펴보면 인종평가 에서는 약물학적 유전자 타입이 6회, 변이 점수 윈저화 3회, 낮은대립형질빈도 와 mRNA안정기전이 각각 2회씩 포함되어 있었고 비인종평가 에서는 약물학적 유전자 타입이 6회, 변이 점수 윈저화 4회, 낮은대립형질빈도 와 mRNA안정기전이 각각 2회 그리고 동형접합변이빈도가 1회씩 포함되어 있었다. 이 결과로 미루어보아 알고리즘 계산시 절단점값(0.2)을 찾고 절단점값 이상에서 나타나는 잡음을 제거하는 점수 윈저화가 필수적이라는 것을 알 수 있다. 또한

약물 유전자의 여러가지 종류 중 한개의 유전자가 많은 약물에 영향을 미치는 효소가 약물반응성에 큰 영향력을 미치고 있음을 알 수 있다.

고찰

약물부작용(ADR)의 원인이 되는 변이를 밝히는 연구는 꾸준히 진행되어 왔으며 응고인자 활성을 억제하는 항응고제인 와파린(warfarin), 간질치료제로 쓰이는 카바마제핀(carbamazepine), 진통제로서 감기약의 성분이 되는 코데인(Codeine)등에 관련된 연구가 그 대표적인 예이다. 코데인의 경우, CYP2D6*1xN/*2xN/*17xN/*35xN 유전형을 가진 환자에서 CYP2D6 유전자가 초고속대사자(UMs;Ultrarapid Metabolizers) 표현형을 띄게 되고 코데인의 대사가 정상 유전형을 가진 경우보다 훨씬 빠르게 되어 독성을 일으킨다고 보고되었다.[61, 62] 또한 VKORC1과 CYP2C9 유전자에 −1639G>A , CYP2C9*2 , CYP2C9*3 유전형을 가진 환자의 경우 일반 환자에서보다 와파린을 37% 적게 복용해야하며 그렇지 않은 경우 과도한 항응고 반응으로 인한 약물부작용이 일어날 수 있다고 보고되었다.[63, 64].

앞선 예시와 같이 유전적 변이에 의해 약물부작용이 발생할 수 있다는 사실이 그동안 다수의 연구결과로 밝혀져 이러한 원인 변이를 찾는 것이 약물부작용 연구에 있어 매우 중요하다고 알려져 있음에도 불구하고 현재 알려진 약물부작용의 원인 변이는 상당히 적다. 이러한 현상의 원인은 현재 연구 방법의 한계점에서 기인한다. 현재까지 약물부작용의 원인 변이를 밝히는 연구는 대부분 하나의 약물을 대상으로 부작용을 일으키는 실험군과 대조군(case−control)을 표본추출하여 전장유전체분석연구(GWAS)기법을 통해 주로 5개 미만의 원인 변이를 밝히는 인구기반의 관찰연구(population−based observational studies)였다. 인구기반의 관찰연구는 연구대상이 되는 표본을 추출하기 위해 소요되는 시간이 길고 그에 따른 큰 비용이 요구되는데 반해 상대적으로 적은 수의 원인 변이를 결과로 얻게 되어 현실적으로 많은 수의 연구를 수행하기 어렵다는 단점을 가지고 있다. 또한 인구기반의

관찰연구 특성상 추출된 표본의 인구통계학적 요인/조건(Demographic condition)에 의해 영향을 크게 받게 되는데 대부분의 연구대상 인종이 백인(Caucasian)이기 때문에 아시아나 아프리카 인종에 대해 동일한 연구결과를 적용하는데 어려움이 있었다. 이러한 문제점들으로 인해 개인의 유전적 정보를 활용하여 개인별 맞춤 약물처방을 제공하는 맞춤의료(Personalized medicine)는 실현되기 어려웠다.

이러한 문제점을 극복하고자 우리는 2013년 개인의 유전체정보를 사용하여 각 개인별 위험한 약물 의 순위를 제공하는 PharmSafe 알고리즘을 개발하고 이를 약물과 해당약물부작용의 원인변이 정보를 지식베이스화한 PharmGKB를 이용하여 알고리즘을 평가해 알고리즘의 성능을 입증하였다 [49]. PharmSafe 알고리즘은 각 개인의 유전체 서열로 부터 변이정보를 추출하여 변이, 유전자, 약물점수를 계산하기 때문에 앞서 제시된 인구기반의 관찰연구의 비용적인 문제와 인구통계학적 요인에 의존적인 결과등의 한계점들을 극복하였다. 하지만 암호영역의 변이에 대해서만 적용이 가능하고 여러가지 생물학, 약리학, 통계학적 요소들을 충분하게 반영하지 못한다는 한계점을 가지고 있었다. 이러한 점들을 극복하고자 본 논문에서는 생물학, 약리학, 통계학적 지식요소 7가지(중심경향방법, 약물학적 유전자 종류, 변이점수 원저화, 낮은 대립형질 빈도, 동형접합변이 빈도, mRNA 안정성 조절 기전, 유전자 기능 조절변이)에 속하는 54가지 조건을 반영하여 PharmSafe 알고리즘을 계산하고 평가하였다. 54가지 조건을 반영하여 알고리즘을 계산, 평가한 결과 전체 인종을 대상으로한 평가(Ethnicity-specific validation)에서는 AUC 0.5633~0.6436, 0.6057±0.012 (최솟값~최댓값, 평균±표준편차)로 가장 높은 AUC값을 얻은 조건은 약물학적 유전자 종류 중 약물분해효소에 대한 가중치 계산이었다. 비인종 평가(Ethnicity-non-specific validation)에서는 AUC 0.5857~0.6502, 0.6224±0.222(최솟값~최댓값,평균±표준편차)으로

가장 높은 AUC값을 얻은 조건은 중심경향방법의 요인에 속하는 곱이었다(그림 2). 약동학(PK;Pharmacokinetics)에서는 약물의 생체 작용을 설명하는 흡수, 분포, 대사, 배설(ADME;Absortion, Distribution, Metabolism, Excretion) 중 대사작용이 약물의 농도 조절과 부작용의 많은 부분을 설명한다고 제시하고 있으며, 이러한 주장은 와파린을 비롯한 많은 약물들의 부작용의 원인변이가 CYP 유전자군에서 발견되는 것으로 증명되고 있다[64]. 알고리즘의 검정을 위해 사용한 PharmGKB의 1807개의 변이 약물 연관관계 데이터에서 효소가 전체 471개중 20개(4.24%) 인 결과를 반영하여도 효소가 약물반응성에 큰 영향을 미친다는 것을 알 수 있다. 이러한 연구결과들과 약동학적 이론으로 미루어 보아 약물반응성에 가장 큰 영향을 미치는 것이 약물대사효소 유전자라는 것을 알 수 있다. 인종대상 평과(Ethnicity-specific validaton)결과에서 약물대사효소 유전자에 대해 가중치를 사용했을때 가장 좋은 평가결과를 얻은 것 또한 같은 맥락으로 생각될 수 있다. 또한 하나의 약물대사효소 유전자에 변이가 발생하는 경우 약물표적 유전자의 경우와 달리 다수의 약물대사에 영향을 미치기 때문에 약물대사효소 유전자가 약물반응성에 중요한 역할을 한다고 할 수 있다.

이렇게 개선된 PharmSafe 알고리즘이 실제 약물학적 메커니즘을 반영하여 각 개인별 위험한 약물 순위를 도출하고 있으며 이 결과가 증명됨을 알 수 있다. 54가지 가중치 조건의 조합실험에서도 점수윈저화와 약물학적 유전자 타입 조합이 AUC값 0.6235, 0.6426으로 인종, 비인종 평가 모두에서 가장 좋은 결과를 보였다(그림 3). 이 결과 역시 약물학적 유전자 종류에 따라 약동학적인 작용이 영향을 받으며 특히 약물대사효소가 중요한 역할을 한다는 것을 알 수 있다. 전체적으로 약물학적 유전자 종류를 가중치 요소로 사용한 4가지 조건 실험을 제외하고 모두 비인종평가가 인종평가보다 높은 AUC를

나타냈다. 그 원인은 평가기준으로 사용한 PharmGKB데이터의 특성 때문으로 예상된다. 실제 PharmGKB로부터 다운로드 받은 데이터에는 3,248개의 약물과 원인변이 관계정보가 포함되어 있었으며 이 중 인종정보가 포함된 약물과 원인변이 관계정보는 781개(45.96%)로 절반도 되지 않았다. 앞으로 약물, 원인변이 그리고 인종의 관계정보가 많아지면 PharmSafe 알고리즘의 평가점수 또한 상승될 것으로 예상된다. 54가지 조건을 반영하여 계산한 모든 평가결과에서 아프리카 인종의 AUC 편차가 큰 이유는 PharmGKB에 속해 있는 인종정보가 포함된 약물과 원인변이 관계정보 781개 중 아프리카 인종에 관련된 관계정보는 112개(13.28%)로 백인에 관련된 관계정보가 413개(48.99%)인 것에 비해 월등하게 낮기 때문이다. 이 또한 앞서 언급한 바와 같이 약물, 원인변이 그리고 인종의 관계정보가 많아지면 전체적인 AUC 값은 상승하고 편차는 낮아질 것으로 생각된다. 약물 분류 정보군를 가중치로 이용한 실험에서는 B(혈액 및 혈액 형성기관) 0.5289~0.7727±0.0469, C(심장혈관 시스템) 0.5076~0.6994±0.0418, G(비뇨 생식계 그리고 성 호르몬) 0.3805~0.8731±0.0685, G03(성 호르몬과 생식 시스템의 조절기) 0.7396~0.8619±0.025, C03(이뇨제) 0.4071~0.7928±0.0876, N06A(항 우울제) 0.4942~0.7364±0.0556 (최솟값~최댓값±표준편차) 약물 군에 대한 결과가 공통적으로 AUC값이 높았다. 혈액 및 혈액 형성기관에 속하는 대표적인 약물은 클로피도그렐(clopidogrel), 와파린(warfarin)등이며 앞서 언급한 바와 같이 이들 약물은 부작용의 원인이 되는 변이에 대한 연구가 잘 알려져 있으며 이를 고려한 실제 임상사례들 또한 많다. 다리페나신(Darifenacin)은 과민성방광 치료에 사용되는 약물로CYP2D6 유전자에 의해 대사가 조절되는 것으로 알려있으며 이는 FDA drug lable에 기재되어 있다 [65]. 심장혈관 시스템 군에 속하는 약물 중 혈압강하제인 아테놀(atenolol)은

GALNT2 유전자에 NC_000001.10:g.230294916C>T(rs2144300) 변이가 존재하면 혈압과 고밀도지단백 콜레스테롤(HDL-C) 수치를 낮춘다고 보고되어 있으며 HDL 콜레스테롤 수치가 낮으면 일반적으로 심장질환에 위험하다고 알려져 있다[66].

본 논문에서는 앞선 PharmSafe 알고리즘의 가장 큰 한계점인 암호영역의 변이만 고려된다는 한계점을 극복하고 비암호영역의 변이정보를 가중치 정보로 사용했다. 하지만 비암호영역에 해당하는 전체 변이를 사용한 것이 아니라 그 중 26,260,341개의 비암호영역 변이만을 사용했기 때문에 여전히 한계점을 가지고 있다. 이전 논문에 이어 이번 논문에서도 알고리즘의 성능평가를 위해 PharmGKB를 사용했기 때문에 여전히 실제 특정 약물에 대한 부작용을 겪은 환자집단에서 평가하지 못한 한계점 또한 가지고 있다. 이러한 한계점에도 불구하고 본 논문을 통해 소개한 개선된 PharmSafe 알고리즘은 기존 알고리즘에 생물학, 약리학, 통계학적 지식요소를 기반으로 가중치를 반영, 기존 알고리즘 보다 높은 평가수치(AUC)를 보였다. 이 결과는 암호영역의 변이 존재유무만을 반영하여 위험한 약물 순위를 도출하는 것보다 생물학, 약리학, 통계학적 요소를 추가적으로 반영하는 것이 개인에게 위험한 약물을 예측하는데 훨씬 효과적임을 나타낸다. 현대 약물유전학의 지향점은 개인 유전체 서열을 기반으로 개인의 약물반응성을 예측하고 이를 약물 처방시 적용하여 개인 맞춤 약물 처방을 하는 것이다. 우리는 각 개인의 유전변이정보를 반영하여 위험한 약물 순위를 제공하는 PharmSafe 알고리즘이 이러한 점에서 현대 약물유전학의 지향점에 가장 부합하는 알고리즘이라고 생각한다. 앞으로 우리는 인구통계학적 요인/조건(Demographic condition)을 추가적으로 반영하여 각 인종적, 지역성, 환경적 특성을 고려할 수 있도록 PharmSafe 알고리즘을 개선하고자 한다. PharmSafe 알고리즘은 개인에게 위험한 약물 순위를 제공하기 때문에 의학적

의사결정 지원시스템(CDSS;Clinical Decision Support System)에서 사용되면 환자가 자신의 유전체 서열을 바탕으로 환자 자신에게 위험요소가 없는 약물을 알맞은 농도로 처방받는데 매우 유용하게 쓰일 수 있을 것이다. 의사 또한 해당 환자의 유전적 특성을 쉽게 파악하여 맞춤 처방을 할 수 있게 될 것이다. 이는 약물부작용으로 인한 심각한 피해를 감소시킬 뿐아니라 치료비용 또한 획기적으로 줄여줄 것이다. 또한 약물오남용에 의한 부작용이나 비용 역시 감소시킬 것으로 기대한다.

결과 그림

그림 1. 1000 genome 에 포함된 2504명의 개인별 지놈데이터의 각 가중치 요소별 변이, 유전자 그리고 약물 점수 분포

1000 Genome 2504명의 개인 유전체 데이터에 속한 변이, 50509개의 유전자, 497개의 약물정보를 이용한 각 요소별 변이, 유전자, 약물 점수의 개인별 점수 평균 분포(A) 중심경항방법 (B)약물학적 유전자 종류 (C) mRNA 안정성 조절 기전(PS : 조기 종결코돈(Premature Stop codons), RS : 종결 코돈 제거(Removed Stop codons), SO : Splice-Overlap) (D) 낮은 대립형질 빈도 (E) 변이점수 윈저화 (F) 동형접합변이 빈도 (G) 유전자 기능 조절변이(1~6,total : 1 class ~total class)

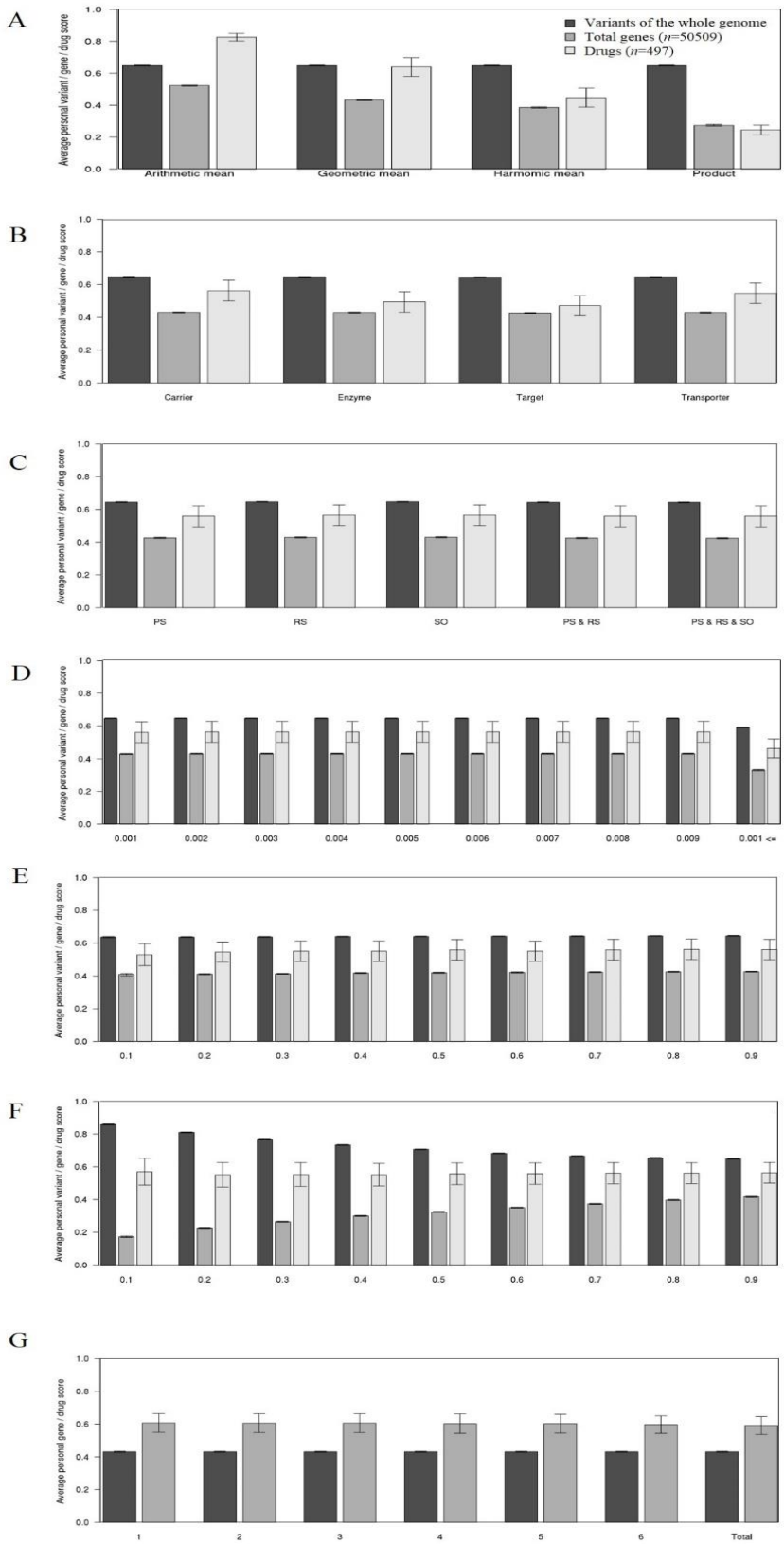
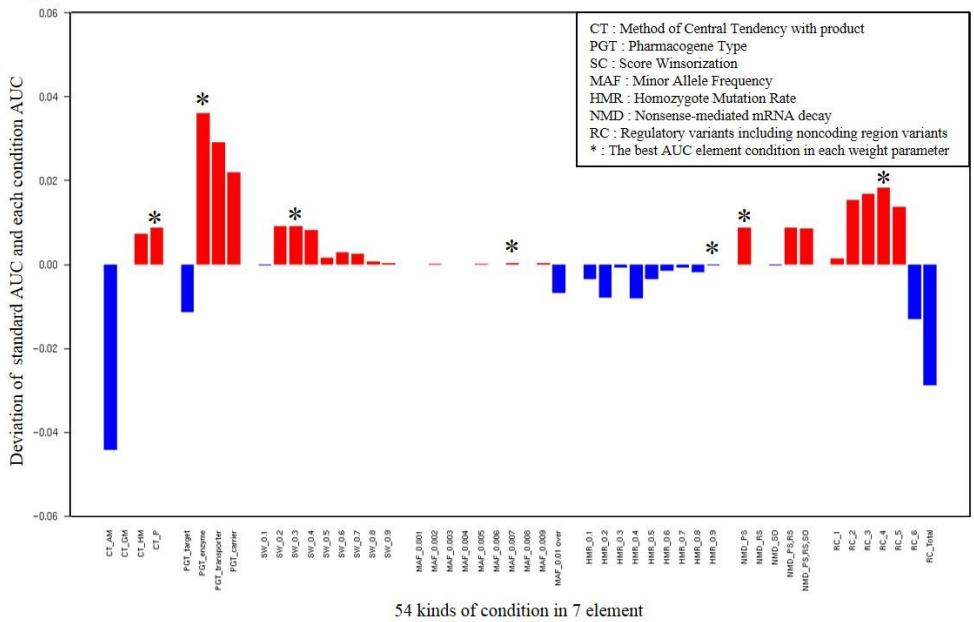

A
Average personal variant / gene / drug score
Variants of the whole genome
Total genes (n=50509)
Drugs (n=497)
Arithmetic mean
Geometric mean
Harmonic mean
Product
B
Carrier
Enzyme
Target
Transporter
C
PS
RS
SO
PS & RS
PS & RS & SO
D
0.001
0.002
0.003
0.004
0.005
0.006
0.007
0.008
0.009
0.001 <=
E
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
F
G
Average personal gene / drug score
1
2
3
4
5
6
Total

그림 2. 7가지 각 요소에 포함된 48가지 조건 별 인종, 비인종 Pharmsafe 알고리즘 검증 결과 비교

1000 Genome 에 속한 2504명의 개인 유전체 데이터를 각 7가지 요소에 포함된 48가지 조건을 가중치로 사용하여 계산한 497 약물별 점수를 PharmGKB로 다운받은 유전자-변이-약물 연관과 인종정보를 사용하여 만든 GS를 사용하여 AUC를 계산하여 검증했다. 54가지 조건별 (A)인종 검증 (B)비인종 검증. 각 막대그래프의 값은 기하평균의 전체 AUC(인종 : 0.6076, 비인종 : 0.6271)값을 기준 AUC 값으로 정하고 이 값에서 각 조건별 AUC값을 뺀 편차이며 적색은 기준 AUC보다 각 조건의 AUC가 상승했음을 의미하고 청색은 기준 AUC보다 각 조건은 AUC값이 하락했음을 의미한다. * 표시는 각 요소별 가장 높은 AUC값을 가진 조건을 의미한다.

A


CT : Method of Central Tendency with product
PGT : Pharmacogene Type
SC : Score Winsorization
MAF : Minor Allele Frequency
HMR : Homozygote Mutation Rate
NMD : Nonsense-mediated mRNA decay
RC : Regulatory variants including noncoding region variants
* : The best AUC element condition in each weight parameter
Deviation of standard AUC and each condition AUC
0.06
0.04
0.02
0.00
-0.02
-0.04
-0.06
CT_AM
CT_GM
CT_HM
CT_P
PGT_target
PGT_enzyme
PGT_transporter
PGT_carrier
SW_0.1
SW_0.2
SW_0.3
SW_0.4
SW_0.5
SW_0.6
SW_0.7
SW_0.8
SW_0.9
MAF_0.001
MAF_0.002
MAF_0.003
MAF_0.004
MAF_0.005
MAF_0.006
MAF_0.007
MAF_0.008
MAF_0.009
MAF_0.01 over
HMR_0.1
HMR_0.2
HMR_0.3
HMR_0.4
HMR_0.5
HMR_0.6
HMR_0.7
HMR_0.8
HMR_0.9
NMD_FS
NMD_RS
NMD_SO
NMD_FS,RS
NMD_FS,RS,SO
RC_1
RC_2
RC_3
RC_4
RC_5
RC_6
RC_Total
54 kinds of condition in 7 element

B

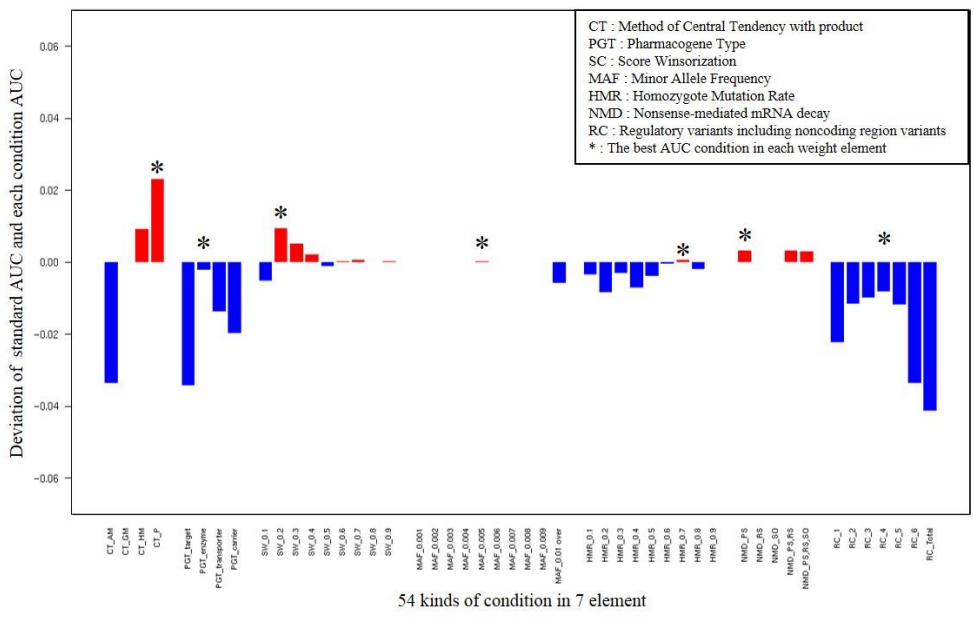
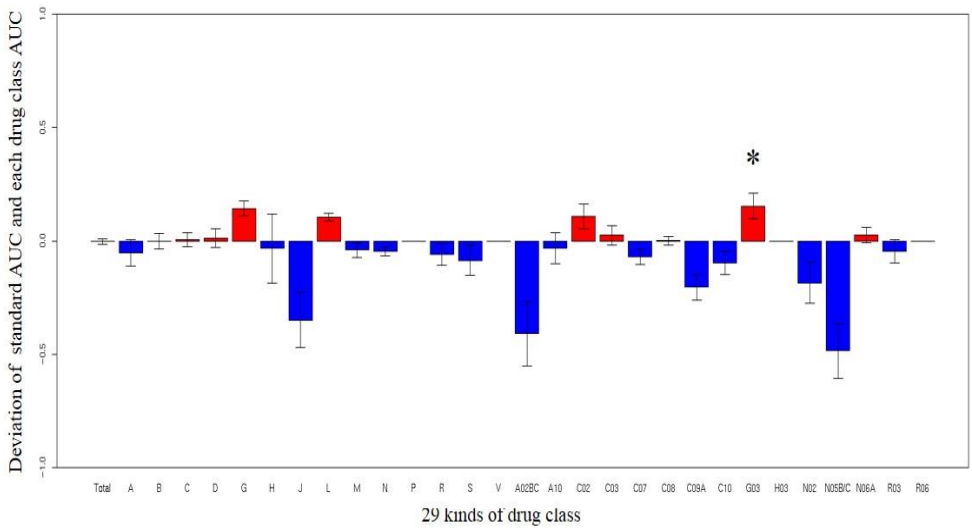

CT : Method of Central Tendency with product
PGT : Pharmacogene Type
SC : Score Winsorization
MAF : Minor Allele Frequency
HMR : Homozygote Mutation Rate
NMD : Nonsense-mediated mRNA decay
RC : Regulatory variants including noncoding region variants
* : The best AUC condition in each weight element
Deviation of standard AUC and each condition AUC
0.06
0.04
0.02
0.00
-0.02
-0.04
-0.06
CT_AM
CT_GM
CT_HM
CT_P
PGT_target
PGT_enzyme
PGT_transporter
PGT_carrier
SW_0.1
SW_0.2
SW_0.3
SW_0.4
SW_0.5
SW_0.6
SW_0.7
SW_0.8
SW_0.9
MAF_0.001
MAF_0.002
MAF_0.003
MAF_0.004
MAF_0.005
MAF_0.006
MAF_0.007
MAF_0.008
MAF_0.009
MAF_0.01 over
HMR_0.1
HMR_0.2
HMR_0.3
HMR_0.4
HMR_0.5
HMR_0.6
HMR_0.7
HMR_0.8
HMR_0.9
NMD_FS
NMD_RS
NMD_SO
NMD_FS,RS
NMD_FS,RS,SO
RC_1
RC_2
RC_3
RC_4
RC_5
RC_6
RC_Total
54 kinds of condition in 7 element

그림 3. 7가지 각 요소에 포함된 48가지 조건 별 29 약물분류군별 인종, 비인종 Pharmsafe 알고리즘 검증 결과 비교

1000 Genome 에 속한 2504명의 개인 유전체 데이터를 각 7가지 요소에 포함된 48가지 조건을 가중치로 사용하여 계산한 497 약물별 점수를 PharmGKB로 다운받은 유전자-변이-약물 연관과 인종정보를 사용하여 만든 GS를 사용하여 29가지 약물분류군별로 AUC를 계산하여 검증했다. 29가지 약물분류군별 (A)인종 검증 (B)비인종 검증. 각 막대그래프의 값은 기하평균의 전체 AUC(인종 : 0.6076, 비인종 : 0.6271)값을 기준 AUC 값으로 정하고 이 값에서 각 조건별 AUC값을 뺀 편차이며 적색은 기준 AUC보다 각 조건의 AUC가 상승했음을 의미하고 청색은 기준 AUC보다 각 조건은 AUC값이 하락했음을 의미한다. * 표시는 각 요소별 가장 높은 AUC값을 가진 조건을 의미한다.

A


Deviation of standard AUC and each drug class AUC
1.0
0.5
0.0
−0.5
−1.0
*
Total A B C D G H J L M N P R S V A02BC A10 C02 C03 C07 C08 C09A C10 G03 H03 N02 N05B/C N06A R03 R06
29 kinds of drug class

B

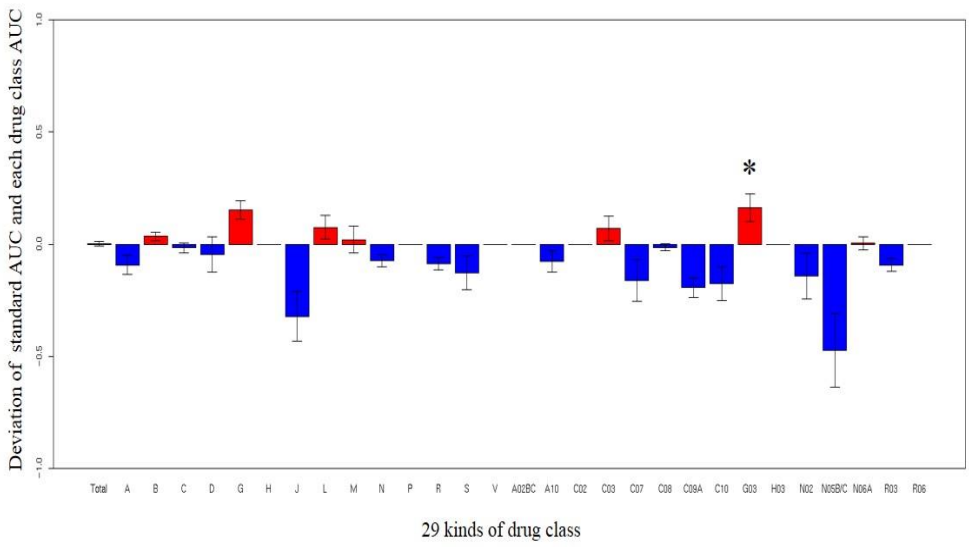
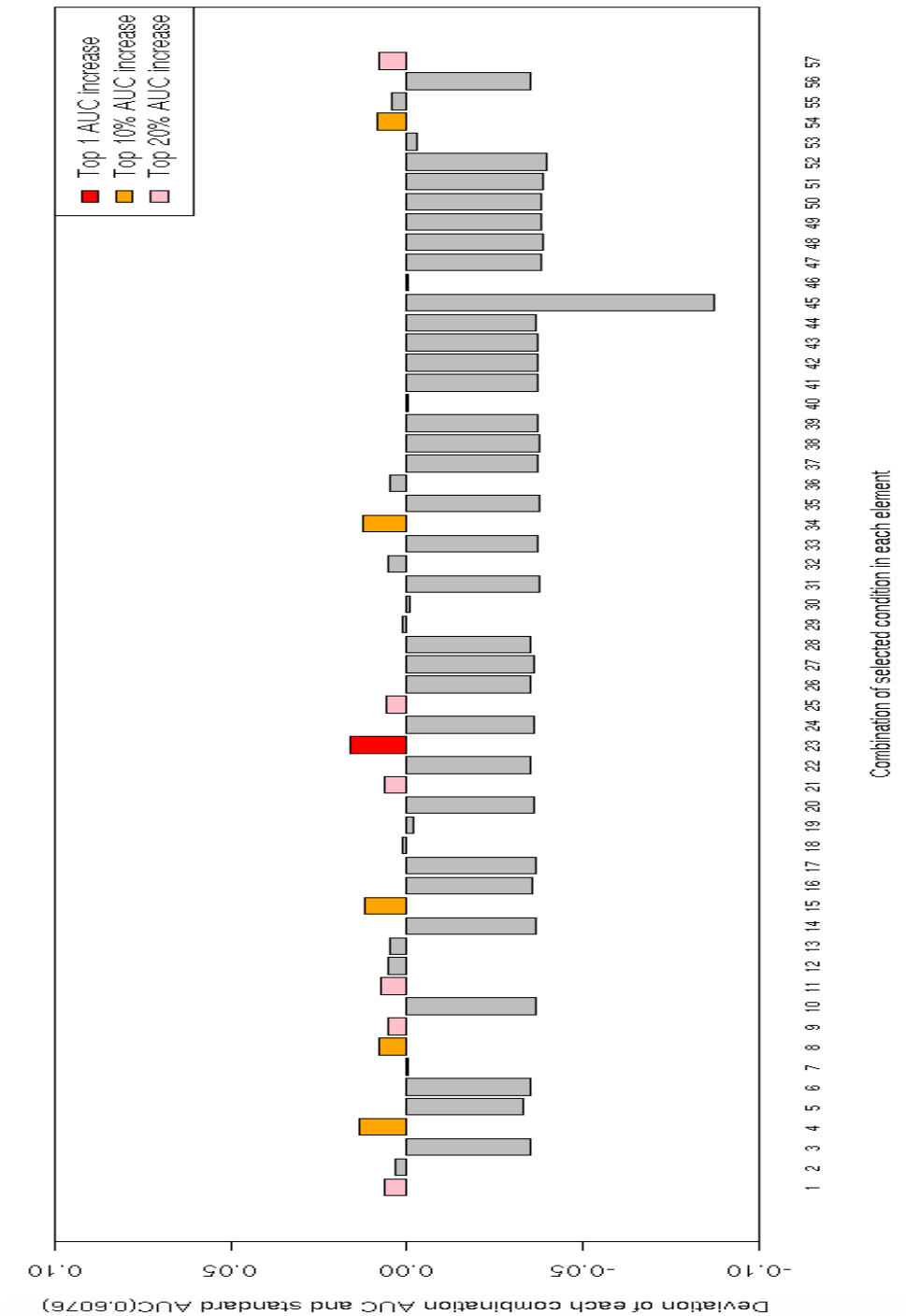

Deviation of standard AUC and each drug class AUC
1.0
0.5
0.0
−0.5
−1.0
*
Total A B C D G H J L M N P R S V A02BC A10 C02 C03 C07 C08 C09A C10 G03 H03 N02 N05B/C N06A R03 R06
29 kinds of drug class

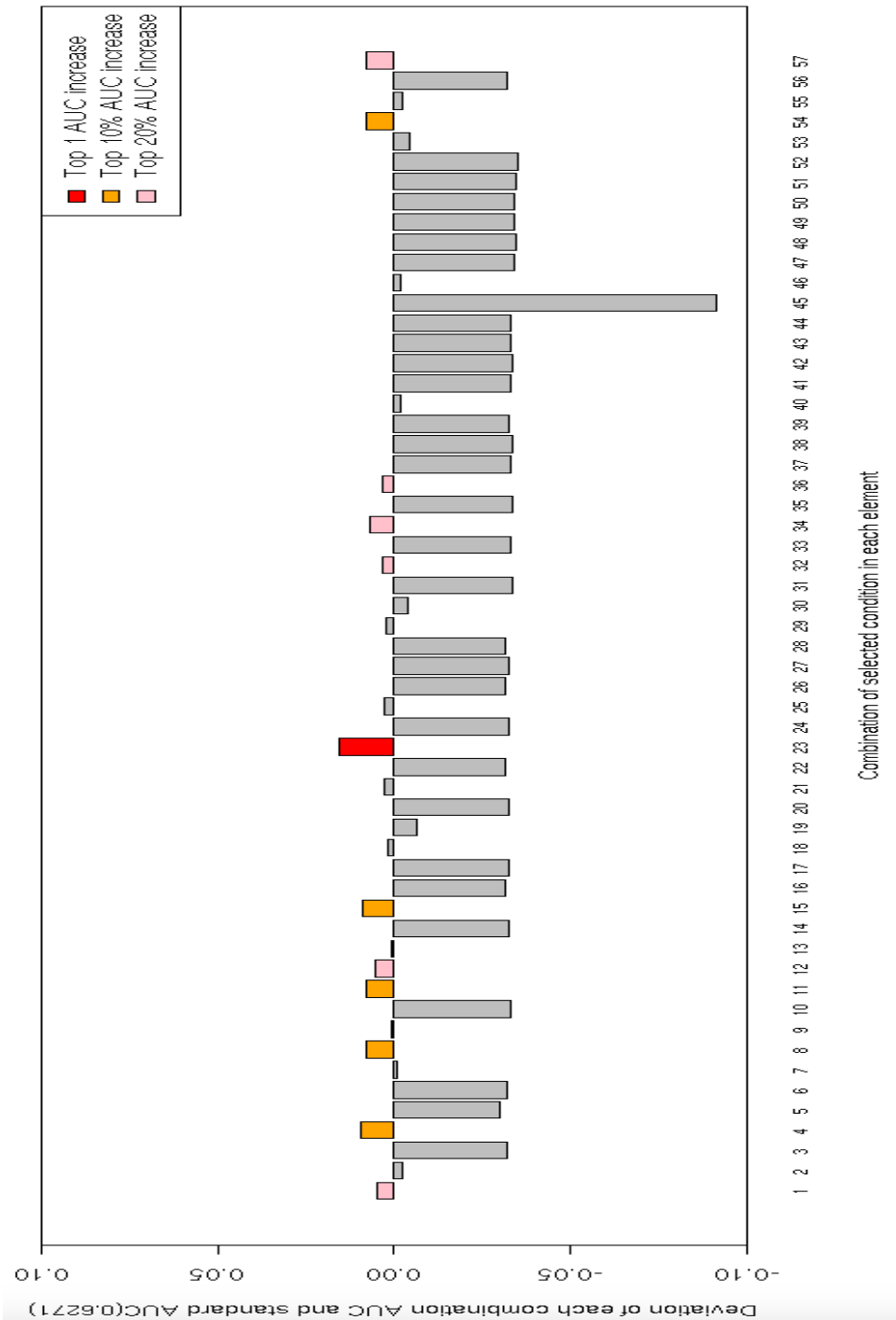
그림 4. 7가지 각 요소에 포함된 48가지의 54조합별 인종, 비인종 Pharmsafe 알고리즘 검증 결과 비교각 7가지 요소에 포함된 48가지 조건중 선택된 6가지 조건(변이 점수 윈저화(SW) 0.2, 낮은 대립형질 빈도(MAF) 0.007, 약물학적 유전자 종류(PGT) 효소, 동형접합변이 빈도(HR) 0.7, mRNA 안정성 조절 기전(NMD) 조기 종결코돈, 유전자 기능 조절 변이(RV))에 대한 모든 57가지 조합을 가중치로 사용하여 2504명 개인별로 Pharmsafe 알고리즘으로 계산하여 약물점수를 도출하고 이를 평가하여 각 조합별로 AUC를 계산하였다. (A) 인종평가, (B)비인종 평가이다. 그래프에서 적색은 가장 높은 AUC를 보인 조합이고 주황색은 상위 10%에 속한 AUC 가진 조합, 분홍색은 상위 20%에 속하는 AUC를 가진 조합이다. 세로축은 인종, 비인종별 기준 AUC(0.6076,0.6271)에서 각 조합의 AUC를 뺀 편차이며 가로축은 각 조합의 순번이다

.0(SW,MAF),1(SW,HR),2(SW,PGT),3(SW,NMD),4(SW,RV),5(MAF,HR),6(MAF,PGT),7(MAF,NMD),8(MAF,RV),9(HR,PGT),10(HR,NMD),11(HR,RV),12(PGT,NMD),13(PGT,RV),14(NMD,RV),15(SW,MAF,HR),16(SW,MAF,PGT),17(SW,MAF,NMD),18(SW,MAF,RV),19(SW,HR,PGT),20(SW,HR,NMD),21(SW,HR,RV),22(SW,PGT,NMD),23(SW,PGT,RV),24(SW,NMD,RV),25(MAF,HR,PGT),26(MAF,HR,NMD),27(MAF,HR,RV),28(MAF,PGT,NMD),29(MAF,PGT,RV),30(MAF,NMD,RV),31(HR,PGT,NMD),32(HR,PGT,RV),33(HR,NMD,RV),34(PGT,NMD,RV),35(SW,MAF,HR,PGT),36(SW,MAF,HR,NMD),37(SW,MAF,HR,RV),38(SW,MAF,PGT,NMD),39(SW,MAF,PGT,RV),40(SW,MAF,NMD,RV),41(SW,HR,PGT,NMD),42(SW,HR,PGT,RV),43(SW,HR,NMD,RV),44(SW,PGT,NMD,RV),45(MAF,HR,PGT,NMD),46(MAF,HR,PGT,RV),47(MAF,HR,NMD,RV),48(MAF,PGT,NMD,RV),49(HR,PGT,NMD,RV),50(SW,MAF,HR,PGT,NMD),51(SW,MAF,HR,PGT,RV),52(SW,MAF,HR,NMD,RV),53(SW,MAF,PGT,NMD,RV),54(SW,HR,PGT,NMD,RV),55(MAF,HR,PGT,NMD,RV),56(SW,MAF,HR,PGT,NMD,RV)

A. 인종평가


Top 1 AUC increase
Top 10% AUC increase
Top 20% AUC increase
Deviation of each combination AUC and standard AUC(0.6076)
0.10
0.05
0.00
-0.05
-0.10
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57
Combination of selected condition in each element

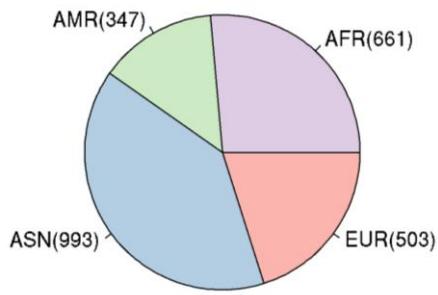
B. 비인종평가


Top 1 AUC increase
Top 10% AUC increase
Top 20% AUC increase
Deviation of each combination AUC and standard AUC(0.6271)
-0.10
-0.05
0.00
0.05
0.10
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57
Combination of selected condition in each element

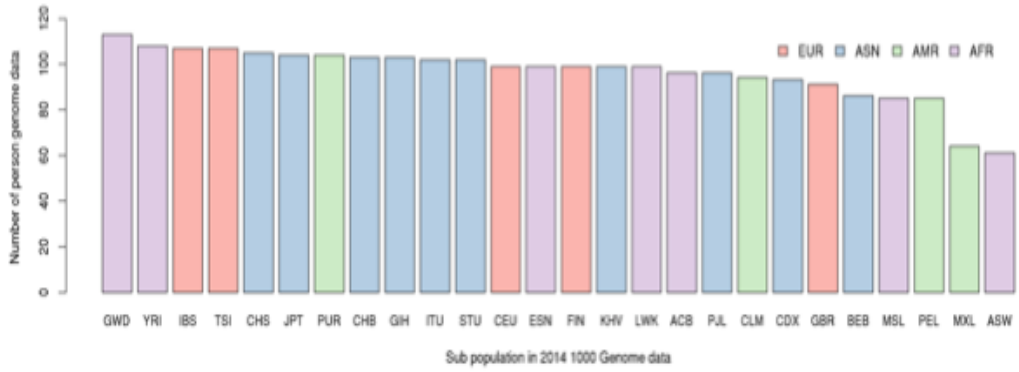
보충 그림

보충 그림 1. 2014년도 1000 Genome 데이터의 2504명의 개인유전체의 하위 , 상위 인구집단(sub & super population) 구성. 2014년도 1000 지놈 데이터의 26 하위 그리고 4 상위 인구집단 구성정보 (2015 6). A. 상위 인구집단 4개에 따른 2504명의 분포(AFR(African), EUR(European),ASN(EAS;East Asian),ASN(SAS;South Asian), AMR(Ad Mixed American)). B. 하위 인구집단 26개에 따른 2504명의 분포(AFR(YRI(Yoruba in Ibadan, Nigeria),LWK(Luhya in Webuye, Kenya),GWD(Gambian in Western Divisions in the Gambia),MSL(Mende in Sierra Leone),ESN(Esan in Nigeria),ASW(Americans of African Ancestry in SW USA),ACB(African Caribbeans in Barbados)),EUR(CEU(Utah Residents (CEPH) with Northern andWestern European Ancestry),TSI(Toscani in Italia), FIN(Finnish in Finland),GBR(British in England and Scotland),IBS(Iberian Population in Spain)), ASN(EAS,SAS;CHB(Han Chinese in Bejing, China),JPT(apanese in Tokyo, Japan),CHS(Southern Han Chinese),CDX(Chinese Dai in Xishuangbanna, China),KHV(Kinh in Ho Chi Minh City, Vietnam), GIH(Gujarati Indian from Houston, Texas),PJL(Punjabi from Lahore, Pakistan),BEB(Bengali from Bangladesh), STU(Sri Lankan Tamil from the UK), ITU(Indian Telugu from the UK)),AMR(MXL(Mexican Ancestry from Los Angeles USA),PUR(Puerto Ricans from Puerto Rico),CLM(Colombians from Medellin, Colombia),PEL(Peruvians from Lima, Peru)).

A.

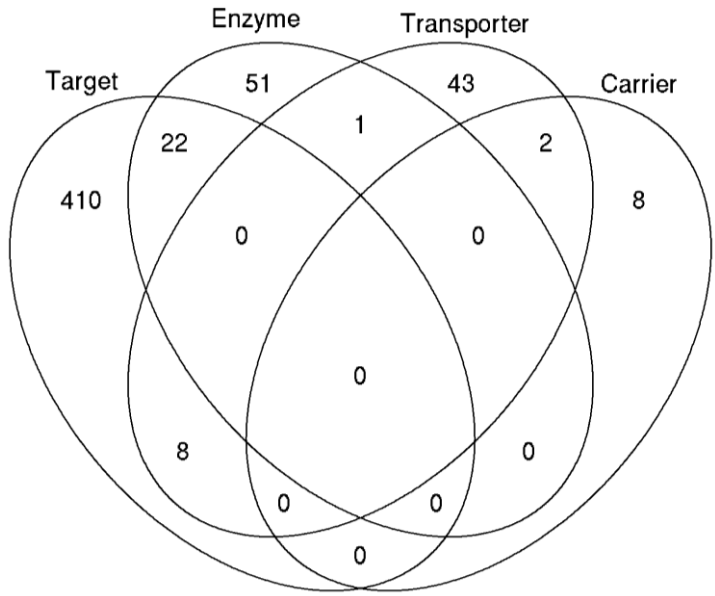

AMR(347)
AFR(661)
ASN(993)
EUR(503)

B

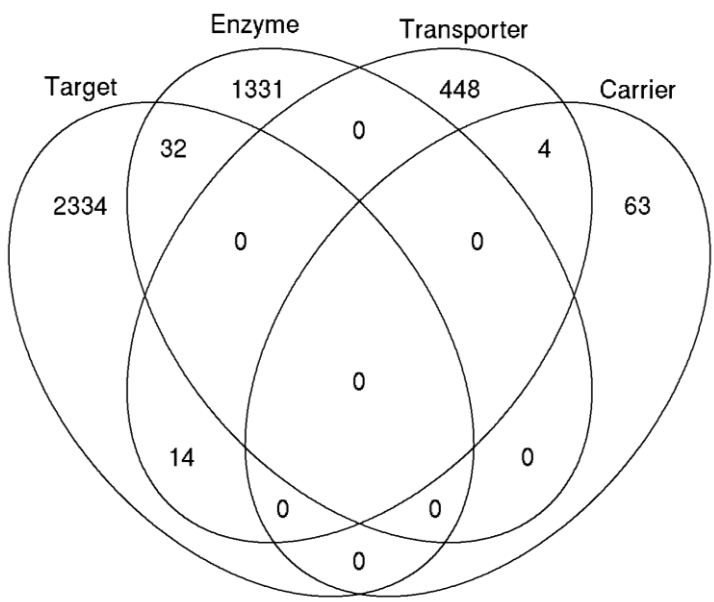

EUR
ASN
AMR
AFR
Number of person genome data
0
20
40
60
80
100
120
GWD
YRI
IBS
TSI
CHS
JPT
PUR
CHB
GIH
ITU
STU
CEU
ESN
FIN
KHV
LWK
ACB
PJL
CLM
CDX
GBR
BEB
MSL
PEL
MXL
ASW
Sub population in 2014 1000 Genome data

보충 그림 2. Pharmsafe 알고리즘에 사용된 약물학적 유전자 군별 약물, 약물-유전자 연관 개수. A. Pharmsafe 알고리즘에 사용된 약물 497 의 약물학적 유전자 군별 분포. B. Pharmsafe 알고리즘에 사용된 약물-유전자 연관 4,426 개의 분포.

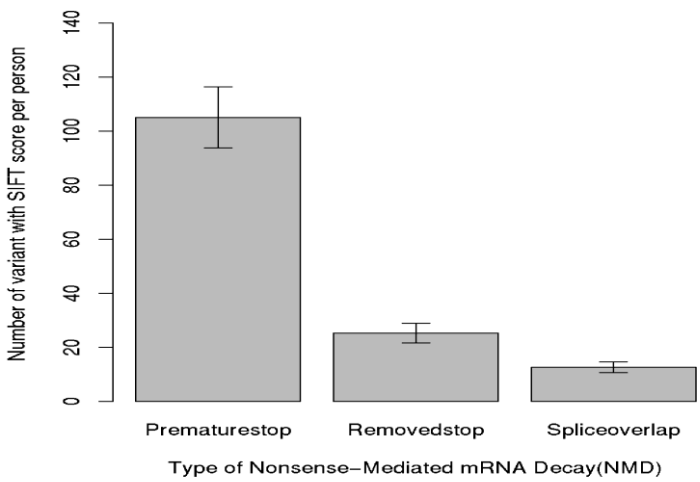
A


Enzyme
Transporter
Target
Carrier
51
43
22
1
2
410
8
0
0
0
8
0
0
0
0

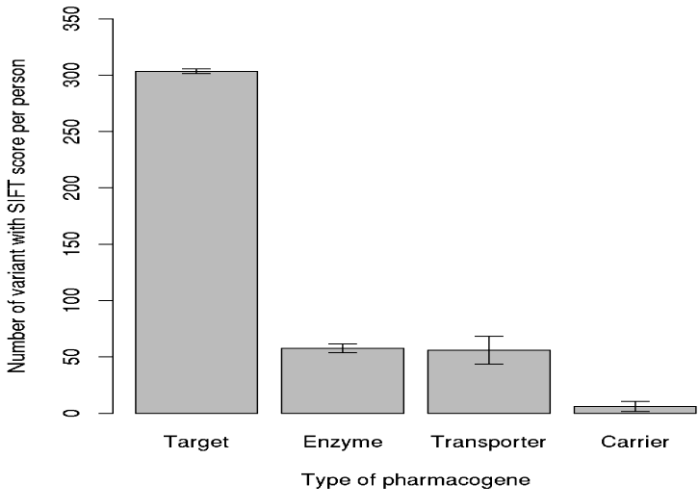
B


Enzyme
Transporter
Target
Carrier
1331
448
32
0
4
2334
63
0
0
0
14
0
0
0
0

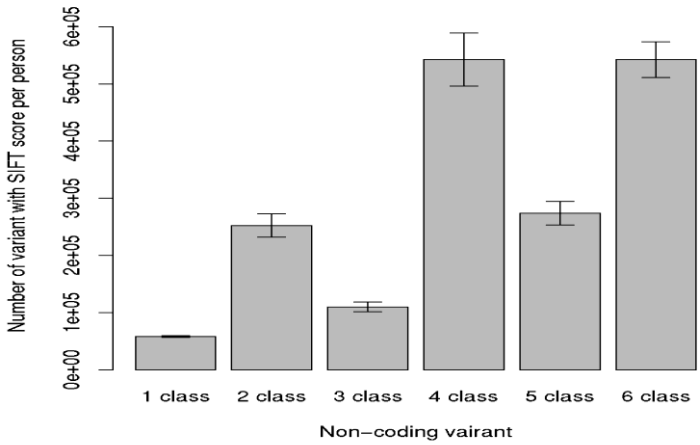
보충 그림 3. 각 생물학적 지식요소별 1000 지놈 데이터에 속하는 변이 개수.



NMD type	No. of variants (mean ± SD)
Prematurestop	105.05±11.23(75~146)
Removedstop	25.19±3.61(16~39)
Spliceoverlap	12.54±1.98(6~20)
Total	142.79 ±13.82(109~194)

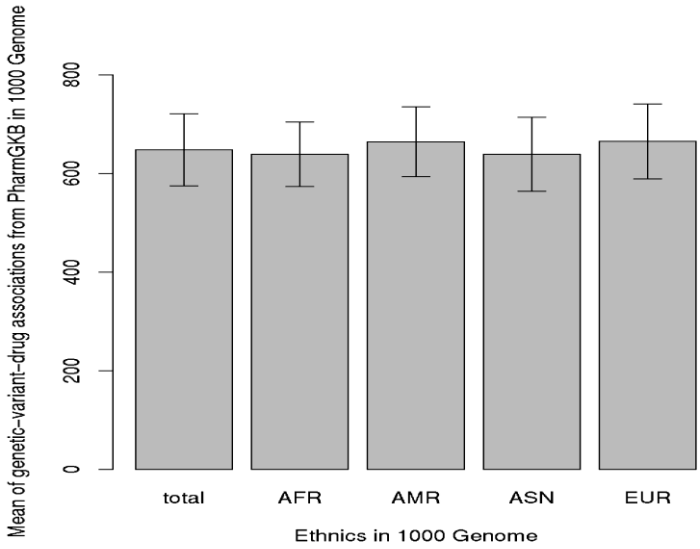


Pharmacogene type	No. of variants (mean ± SD)
Target	303.4±2.21(292~322)
Enayme	57.6±3.91(52~67)
Transporter	56.0±12.40(50~62)
Carrier	6.0±4.47(3~8)
Total	423±12.44(408~436)



Non−coding variant classes	No. of variants (mean ± SD)
1	58,104.42±1490.867 (53,261~6,2725)
2	252,297.80±20,426.453 (224,145~299,733)
3	109,733.28±8,462.391 (98,148~128,443)
4	542,564.47±46,231.893 (481,555~643,982)
5	273,761.18±20,501.018 (246,910~317,964)
6	542,329.02±31,127.380 (495,426~610,183)
Total	1,778,784±126,353.8 (1,600,470~2,056,404)

보충 그림 4. PharmsGKB 와 1000 지놈에 공통적으로 속하는 유전적 변이-약물(genetic-variant-drug associations;GVDA).



Ethnic*	No. of GVDA (mean ± SD)
AFR	638.7±65.3 (505~844)
AMR	663.9±70.7(529~891)
ASN	639.0±75.0(474~868)
EUR	664.7±75.7(493~869)
Total	647.5±73.14(474~891)

*PharmGKB ethnic information AFR(Black or African American), AMR and EUR(White), ASN(Asian). 1000 Genome ethnic information AFR(African), EUR(European), ASN(EAS;East Asian, SAS;South Asian), AMR(Ad Mixed American)

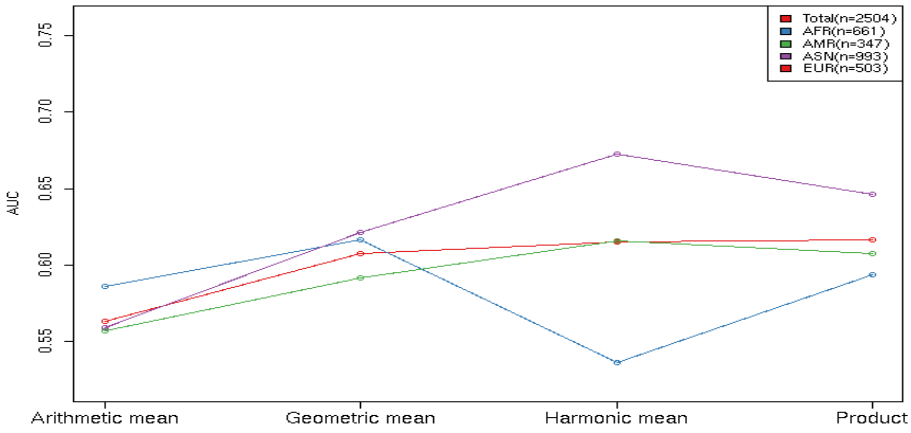
보충 그림 5. 497 전체 약물집합과 각 29가지 약물 군별로 7가지 각 요소에 포함된 54 가지 조건별로 가중치를 반영하여 Pharmsafe 알고리즘을 인종, 비인종으로 평가 .

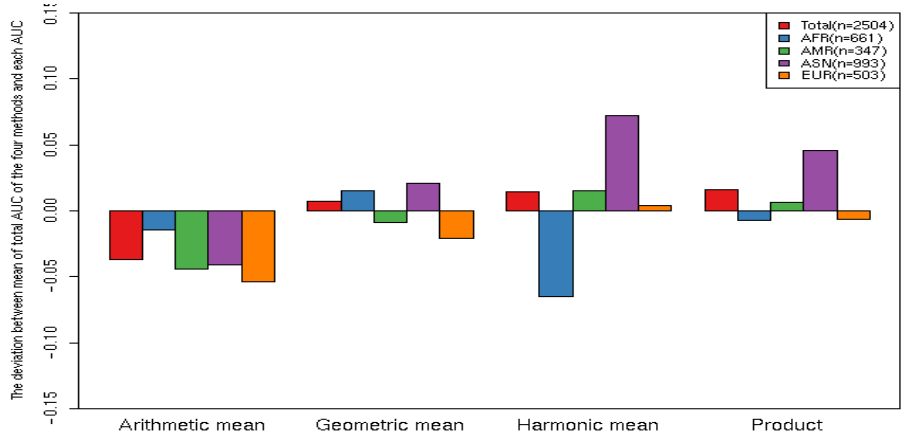
왼쪽 꺾은선 그래프는 각 조건별 AUC를 4가지 인종과 모든 인종을 합한 전체인종 별로 표현한 그래프이다. 오른쪽 그래프는 기준 AUC(기하평균의 AUC)에서 각 조건의 AUC를 뺀 편차이다(단 중심경향법은 기준 AUC를 각 방법의 전체(total) AUC의 평균을 사용하였다). 각 그래프의 색상은 적색은 전체, 청색은 아프리카, 초록색은 미국, 보라색은 아시아, 주황색은 유럽을 나타낸다. 표는 전체 인종 그리고 각 인종별 AUC를 기술해 놓았다. Heatmap은 인종 비 인종별로 표시하였으며 각 셀의 값은 기준 AUC(기하평균의 AUC)에서 각 조건의 AUC를 뺀 편차이며, 보라색은 편차의 상승을 의미하고 청색은 편차의 하강을 의미한다. 각 셀안에 기입된 숫자는 각 실험별 AUC값을 의미한다. 왼쪽 회색, 노란색, 연한 회색은 ATC 에서 추출한 해부학적 그룹 14 가지([A] Alimentary tract and metabolism,[B] Blood and blood forming organs,[C] Cardiovascular system,[D] Dermatologicals,[G] Genito urinary system and sex hormones, systemic [H] hormonal preparations and excl. sex hormones and insulins,[J] Antiinfectives for systemic use,[L] Antineoplastic and immunomodulating agents, [M] musculo-skeletal system,[N] Nervous system,[P] antiparasitic products, insecticides and repellents,[R] Respiratory system,[S] Sensory organs,[V] Various),HOCC에서 추출한 15가지 가장 자주 처방받은 약물 군([A02BC] Proton pump inhibitors,[A10] Drugs used in diabetes, [C02] Antihypertensives,[C03] Diuretics,[C07] Beta blocking agents,[C09A] ACE inhibitors plain,[C08] Calcium channel blockers,[C10] Lipid modifying agents, [G03] Sex hormones and modulators of the genital system,[H03] Thyroid therapy,[N02] Analgesics,[N05B,N05C] Anxiolytics and hypnotics/sedatives, [N06A] Antidepressants, [R03] Drugs for obstructive airway diseases,[R06] Antihistamines for systemic use) 그리고 약물관련 유전자, 비 약물관련 유전자를 의미한다. 각 heatmap 그래프에서 공백은 해당

약물군의 속하는 약물에 대한 변이가 해당 인종에서 존재하지 않아 계산이 되지 않은 것을 의미한다.

A. 중심경향방법(central tendency method)

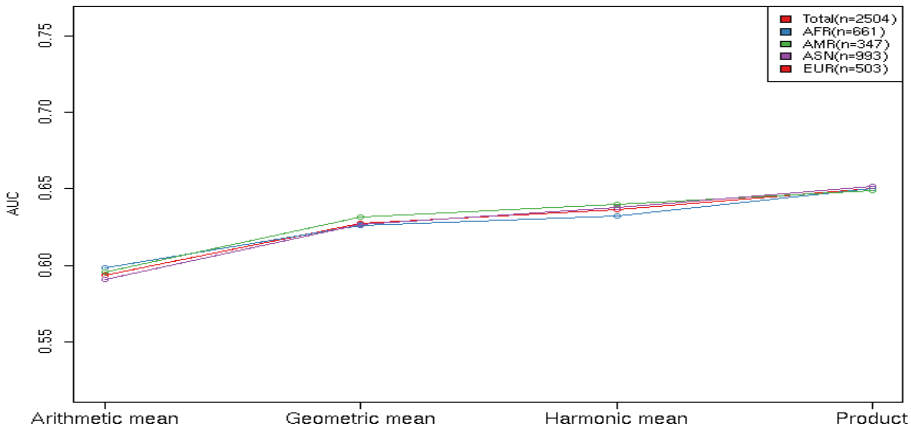
인종평가

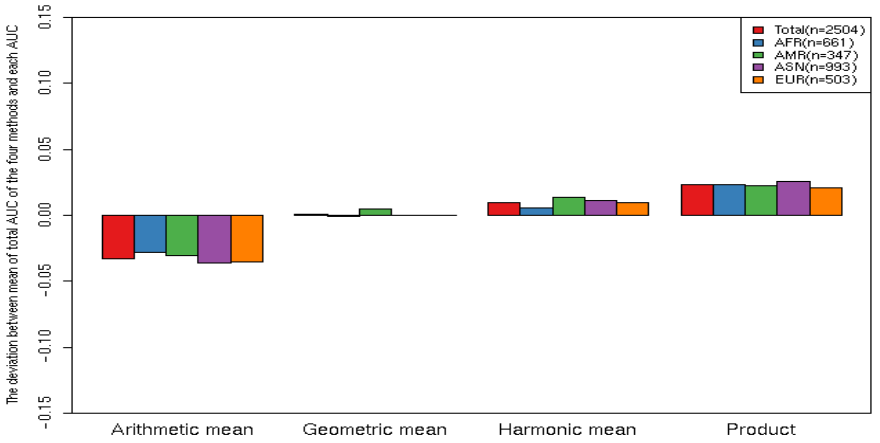




	Arithmetic mean	Geometric mean	Harmonic mean	Product
Total(*n*=2504)	0.5633	0.6076	0.6149	0.6163
AFR(*n*=661)	0.5858	0.6161	0.5358	0.5932
AMR(*n*=347)	0.5565	0.5918	0.6156	0.6071
ASN(*n*=993)	0.5592	0.6214	0.6726	0.6459
EUR(*n*=503)	0.5467	0.5799	0.6046	0.5944

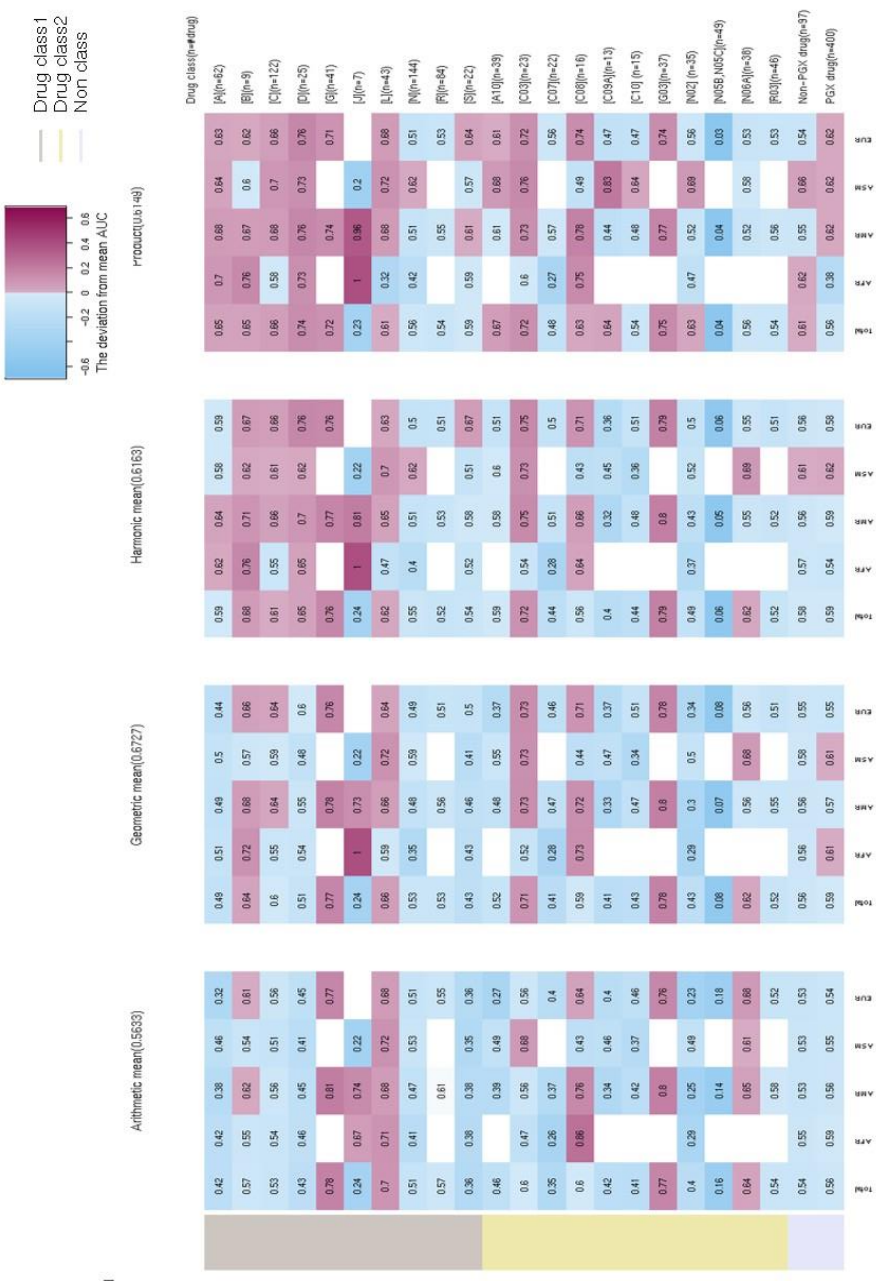
비인종평가



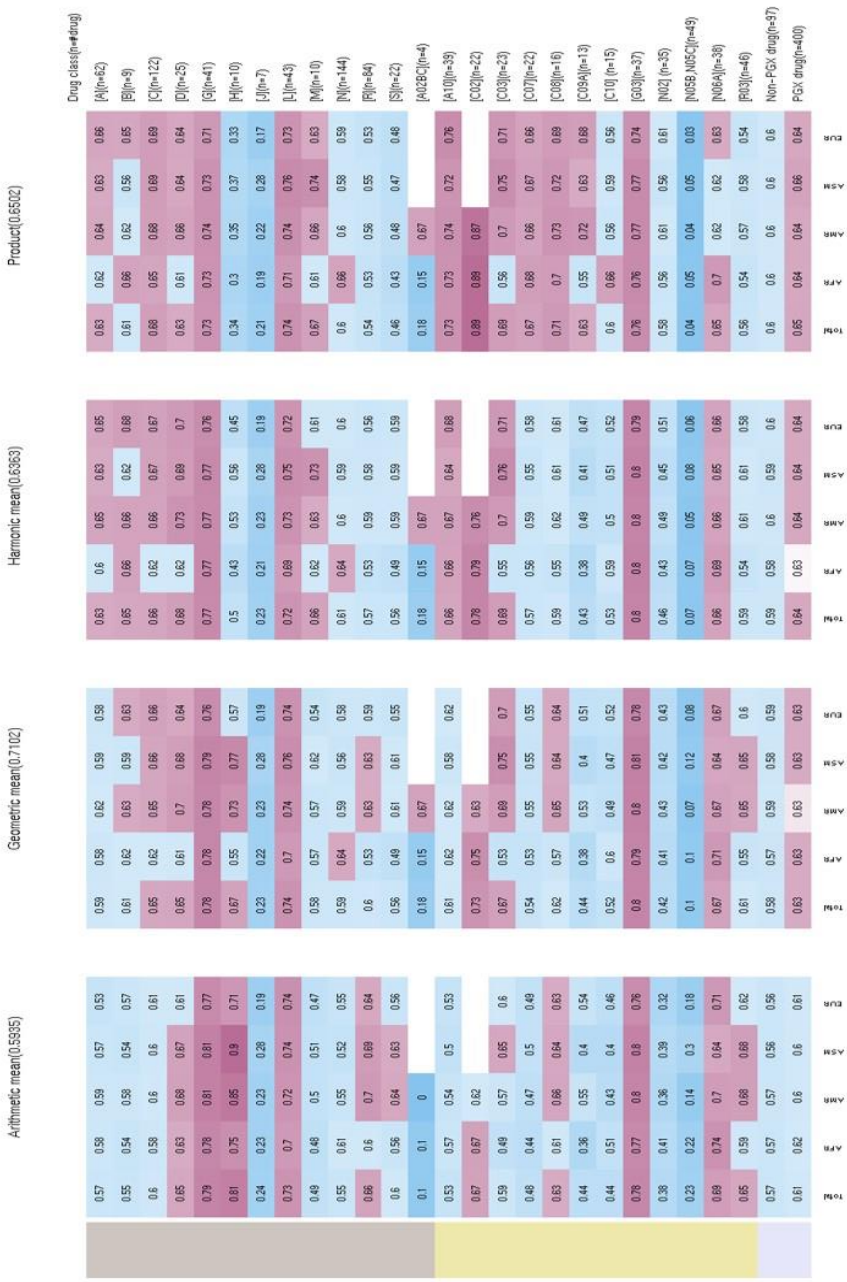


	Arithmetic mean	Geometric mean	Harmonic mean	Product
Total(*n*=2504)	0.5935	0.6271	0.6363	0.6502
AFR(*n*=661)	0.5983	0.6261	0.6325	0.6502
AMR(*n*=347)	0.5958	0.6314	0.6402	0.6491
ASN(*n*=993)	0.5908	0.6266	0.6376	0.652
EUR(*n*=503)	0.5912	0.6264	0.636	0.6472

인종평가

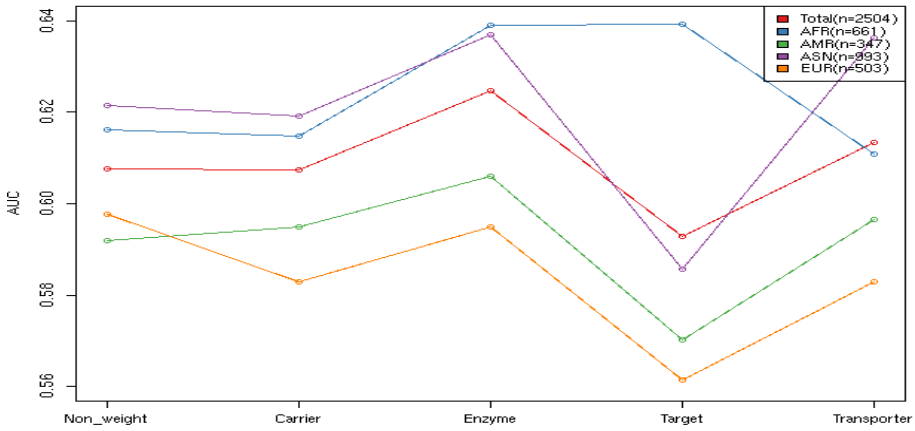

Drug class1
Drug class2
Non class
The deviation from mean AUC
−0.6 −0.2 0 0.2 0.4 0.6
Arithmetic mean(0.5633)
Geometric mean(0.6727)
Harmonic mean(0.6163)
Product(0.149)
Drug class(n=#drug)
[A](n=62)
[B](n=9)
[C](n=122)
[D](n=25)
[G](n=41)
[J](n=7)
[L](n=43)
[N](n=144)
[R](n=64)
[S](n=22)
[A10](n=39)
[C03](n=23)
[C07](n=22)
[C08](n=16)
[C09A](n=13)
[C10] (n=15)
[G03](n=37)
[N02] (n=35)
[N05B,N05C](n=49)
[N06A](n=38)
[R03](n=46)
Non−PGX drug(n=97)
PGX drug(n=400)
Total
AFR
AMR
ASN
EUR

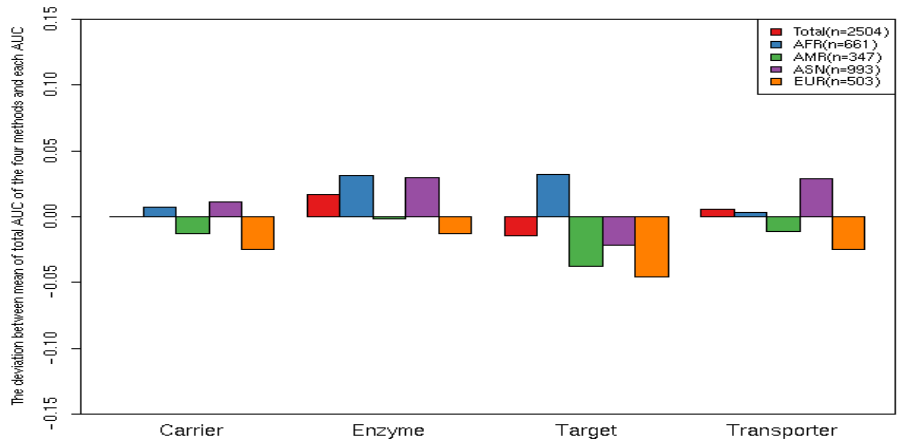
비인종평가


Arithmetic mean(0.5935)
Geometric mean(0.7102)
Harmonic mean(0.6363)
Product(0.6502)
Drug class(n=#drug)
[A](n=62)
[B](n=9)
[C](n=122)
[D](n=25)
[G](n=41)
[H](n=10)
[J](n=7)
[L](n=43)
[M](n=10)
[N](n=144)
[R](n=84)
[S](n=22)
[A02BC](n=4)
[A10](n=39)
[C02](n=22)
[C03](n=23)
[C07](n=22)
[C08](n=16)
[C09A](n=13)
[C10] (n=15)
[G03](n=37)
[N02] (n=35)
[N05B ,N05C](n=49)
[N06A](n=38)
[R03](n=46)
Non-PGX drug(n=97)
PGX drug(n=400)
total
AFR
AMR
ASN
EUR

B. 약물학적 유전자 종류(Pharmacogene type)

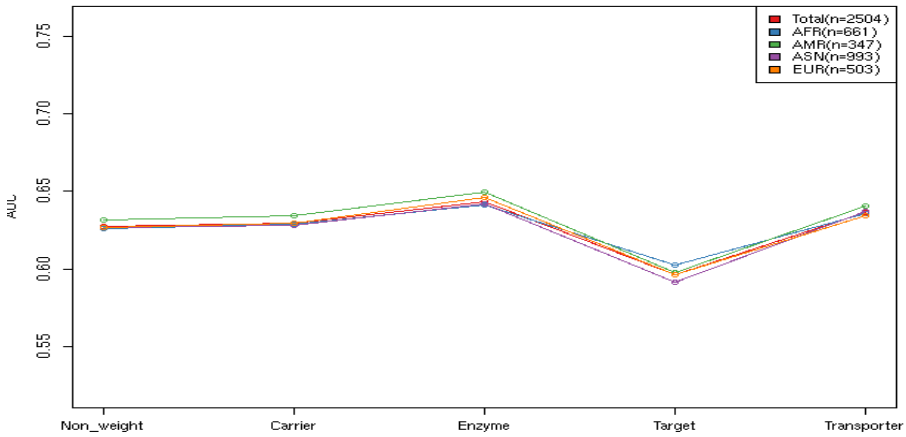
인종평가

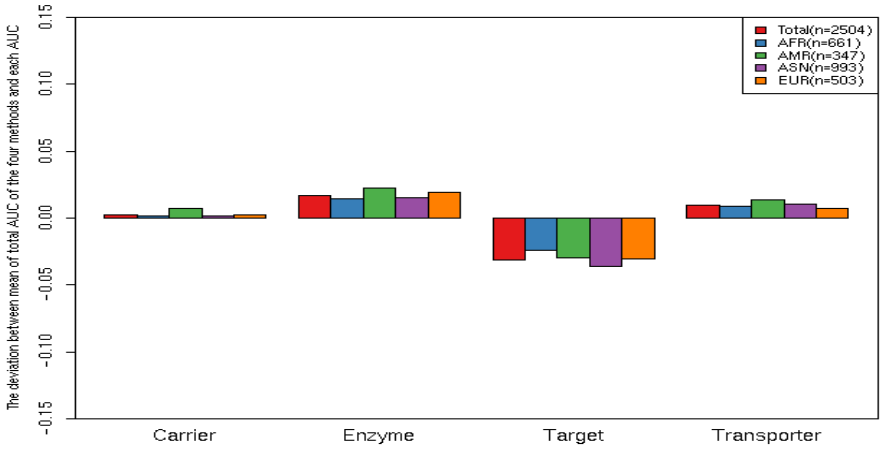




	Non-weight	Carrier	Enzyme	Target	Transporter
Total(*n*=2504)	0.6076	0.6074	0.6248	0.5928	0.6134
AFR(*n*=661)	0.6161	0.6149	0.639	0.6393	0.6109
AMR(*n*=347)	0.5918	0.5948	0.6059	0.5701	0.5965
ASN(*n*=993)	0.6214	0.6191	0.637	0.5857	0.6363
EUR(*n*=503)	0.5799	0.583	0.5949	0.5615	0.583

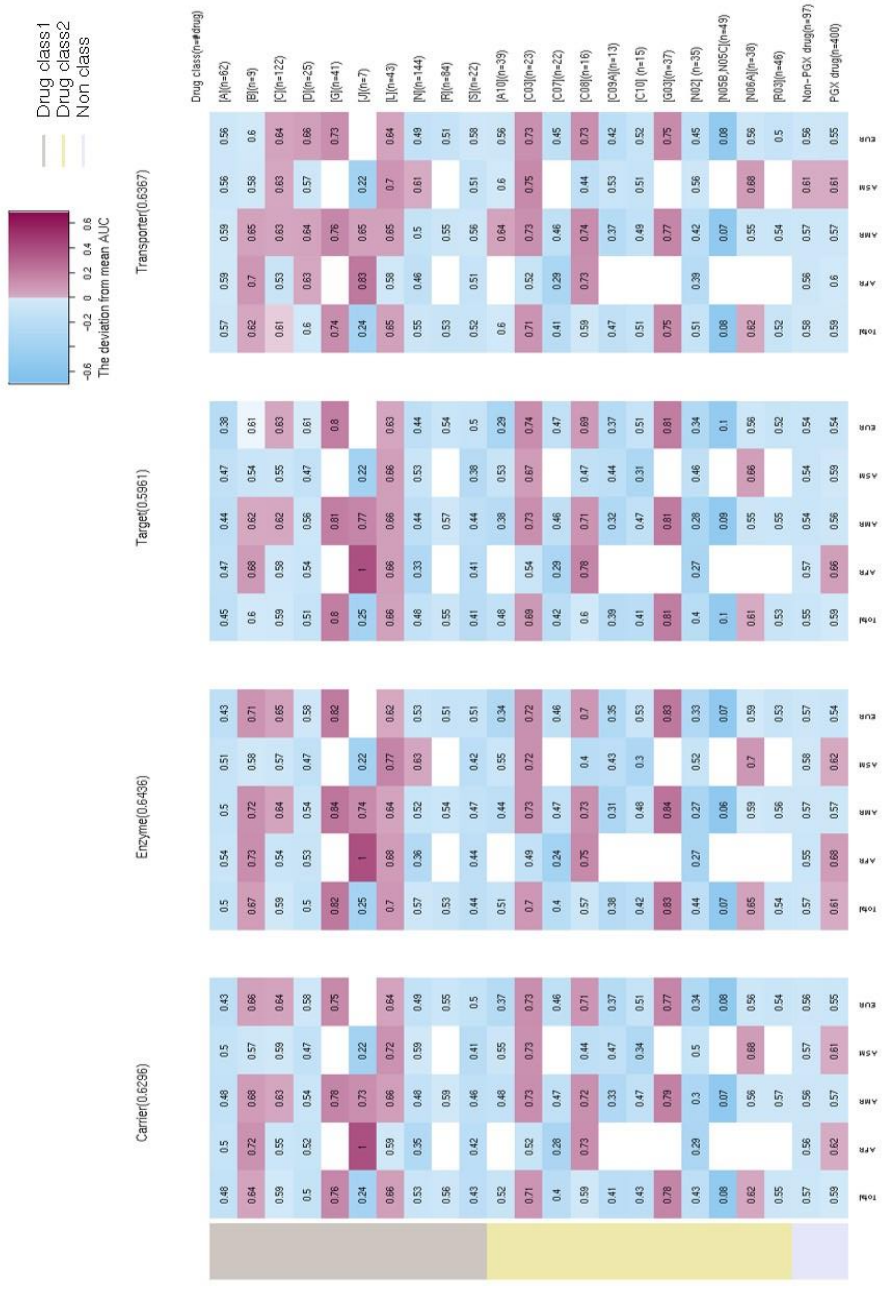
비인종평가



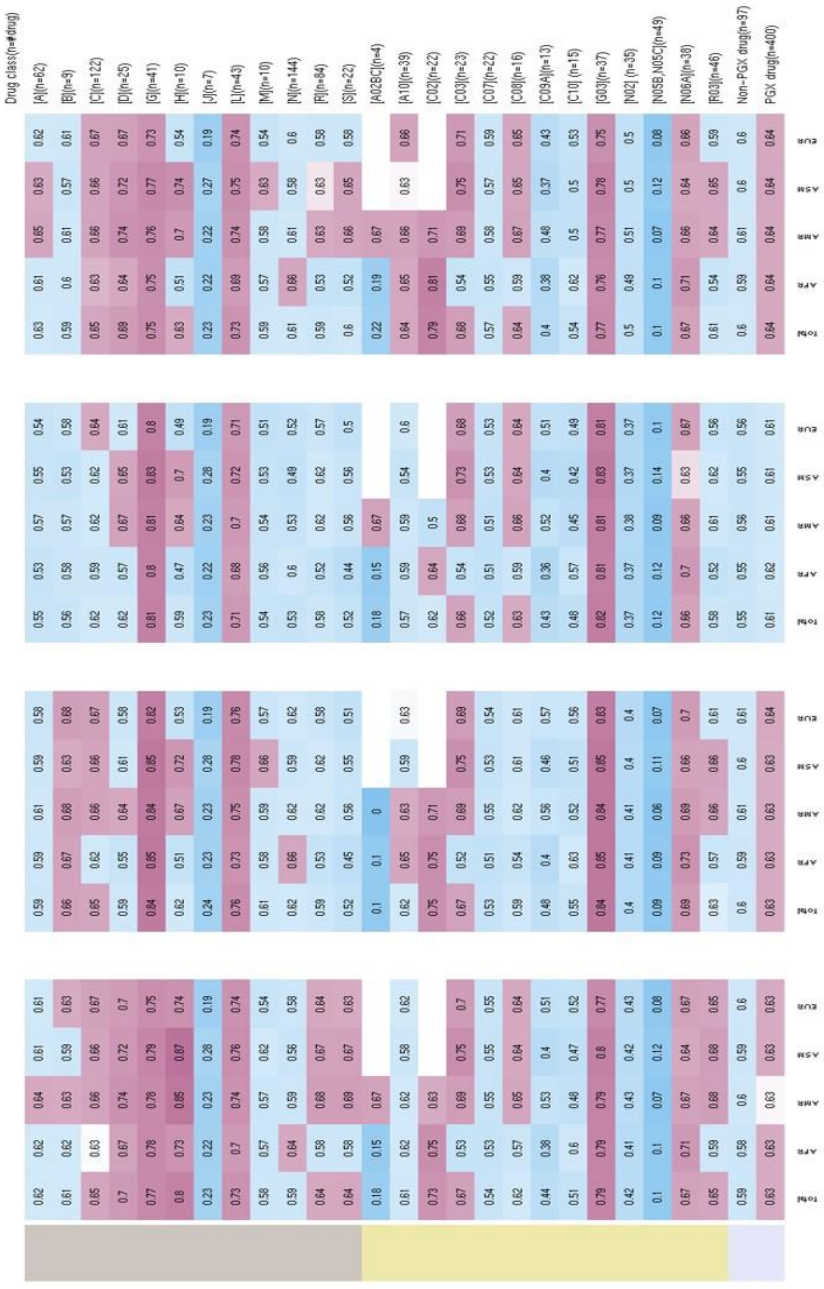


	Non-weight	Carrier	Enzyme	Target	Transporter
Total(*n*=2504)	0.6271	0.6296	0.6436	0.5961	0.6367
AFR(*n*=661)	0.6261	0.6289	0.6411	0.6026	0.6357
AMR(*n*=347)	0.6314	0.6343	0.6493	0.5974	0.6409
ASN(*n*=993)	0.6266	0.6284	0.6422	0.5911	0.6372
EUR(*n*=503)	0.6264	0.6298	0.6459	0.5964	0.6343

인종평가

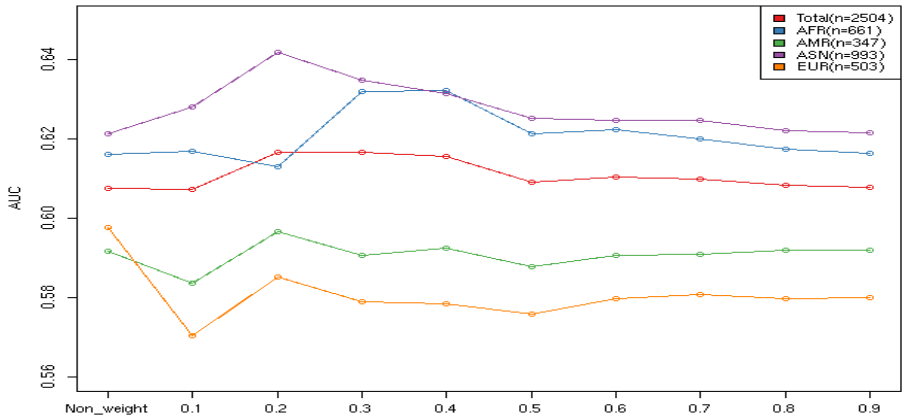

Drug class1
Drug class2
Non class
The deviation from mean AUC
−0.6
−0.2
0
0.2
0.4
0.6
Carrier(0.6296)
Enzyme(0.6436)
Target(0.5961)
Transporter(0.6367)
Drug class(n=#drug)
[A](n=62)
[B](n=9)
[C](n=122)
[D](n=25)
[G](n=41)
[J](n=7)
[L](n=43)
[N](n=144)
[R](n=84)
[S](n=22)
[A10](n=39)
[C03](n=23)
[C07](n=22)
[C08](n=16)
[C09A](n=13)
[C10] (n=15)
[G03](n=37)
[N02] (n=35)
[N05B,N05C](n=49)
[N06A](n=38)
[R03](n=46)
Non−PGX drug(n=37)
PGX drug(n=400)
total
AFR
AMR
ASN
EUR

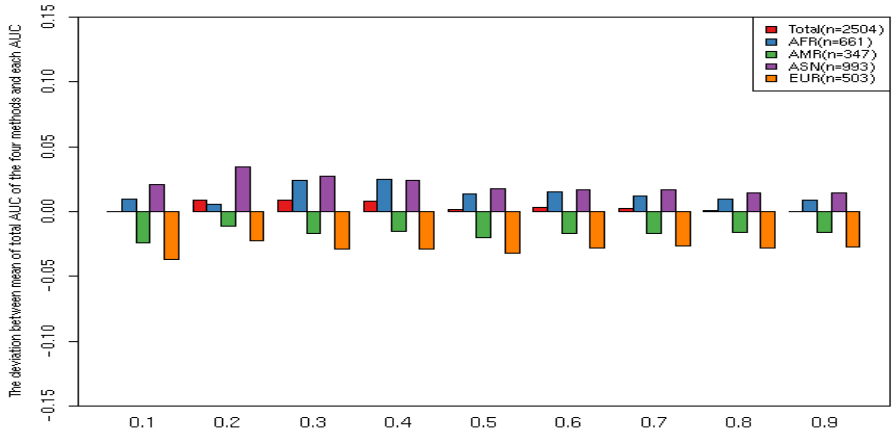
비인종평가


Carrier(0.6296)
Enzyme(0.6436)
Target(0.5961)
Transporter(0.6367)
Drug class(n=##drug)
[A](n=62)
[B](n=9)
[C](n=122)
[D](n=25)
[G](n=41)
[H](n=10)
[J](n=7)
[L](n=43)
[M](n=10)
[N](n=144)
[R](n=84)
[S](n=22)
[A02BC](n=4)
[A10](n=39)
[C02](n=22)
[C03](n=23)
[C07](n=22)
[C08](n=16)
[C09A](n=13)
[C10] (n=15)
[G03](n=37)
[N02] (n=35)
[N05B,N05C](n=49)
[N06A](n=38)
[R03](n=46)
Non-PGX drug(n=97)
PGX drug(n=400)
total
AFR
AMR
ASN
EUR

C. 변이 점수 윈저화(Variant score winsorization)

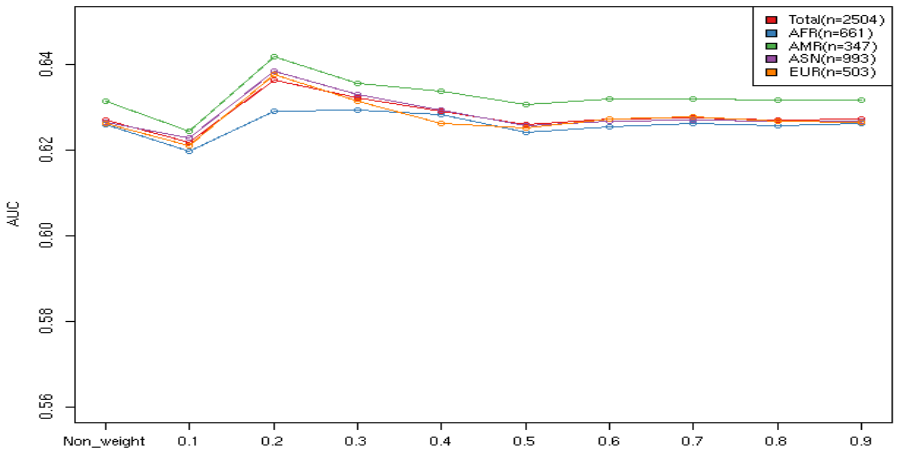
인종평가

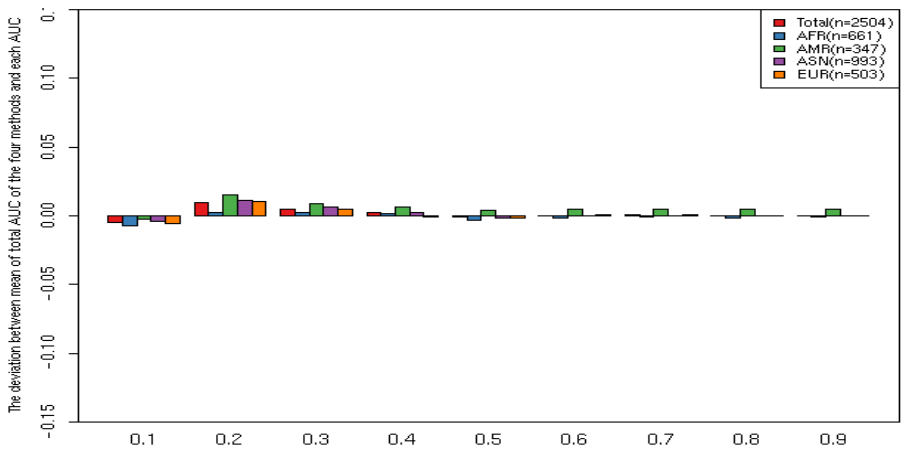




	Non-weight	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Total(*n*=2504)	0.6076	0.6074	0.6166	0.6167	0.6157	0.6092	0.6104	0.61	0.6083	0.6078
AFR(*n*=661)	0.6161	0.6169	0.613	0.632	0.6324	0.6214	0.6225	0.62	0.6175	0.6164
AMR(*n*=347)	0.5918	0.5837	0.5966	0.5906	0.5926	0.5879	0.5907	0.591	0.5919	0.5919
ASN(*n*=993)	0.6214	0.6282	0.6419	0.6348	0.6314	0.6254	0.6247	0.6248	0.6222	0.6217
EUR(*n*=503)	0.5799	0.5704	0.5853	0.5789	0.5785	0.5758	0.5797	0.5809	0.5798	0.58

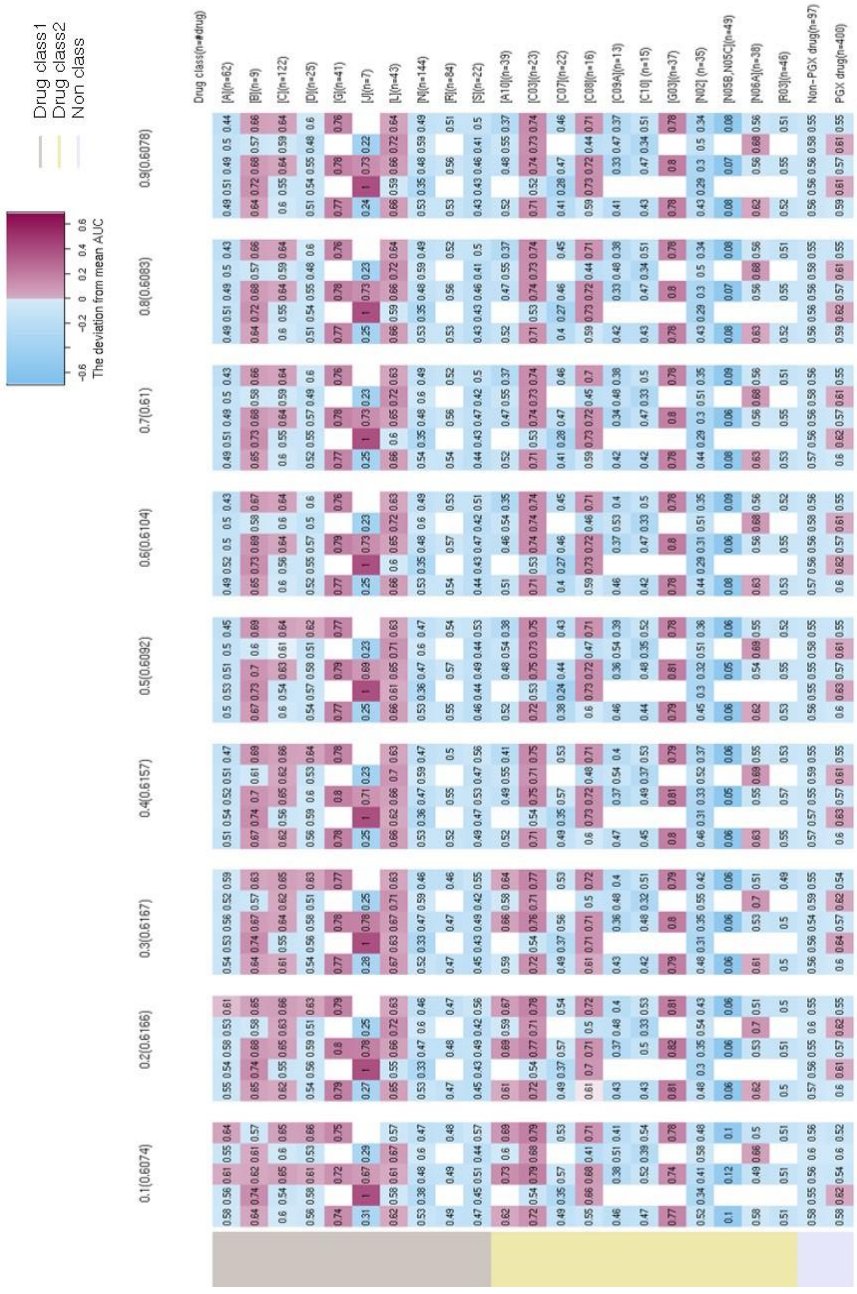
비인종평가



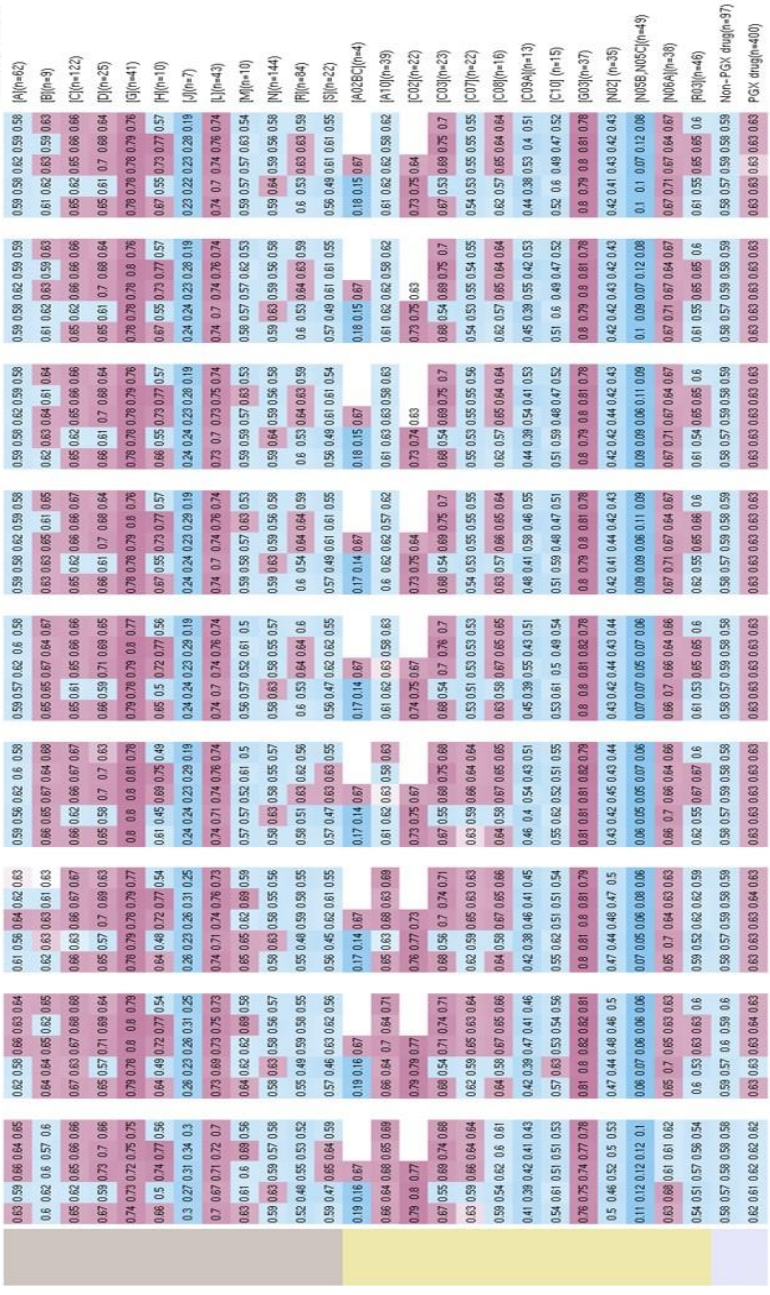


	Non-weight	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Total(n=2504)	0.6271	0.6219	0.6364	0.6322	0.6291	0.6259	0.6273	0.6277	0.6271	0.6273
AFR(n=661)	0.6261	0.6198	0.6292	0.6295	0.6283	0.6241	0.6254	0.6263	0.6257	0.6262
AMR(n=347)	0.6314	0.6244	0.6419	0.6356	0.6338	0.6307	0.632	0.632	0.6318	0.6317
ASN(n=993)	0.6266	0.6228	0.6385	0.6331	0.6293	0.6257	0.6268	0.6271	0.6267	0.6268
EUR(n=503)	0.6264	0.6211	0.6377	0.6315	0.6264	0.6253	0.6274	0.6278	0.6267	0.6266

인종평가

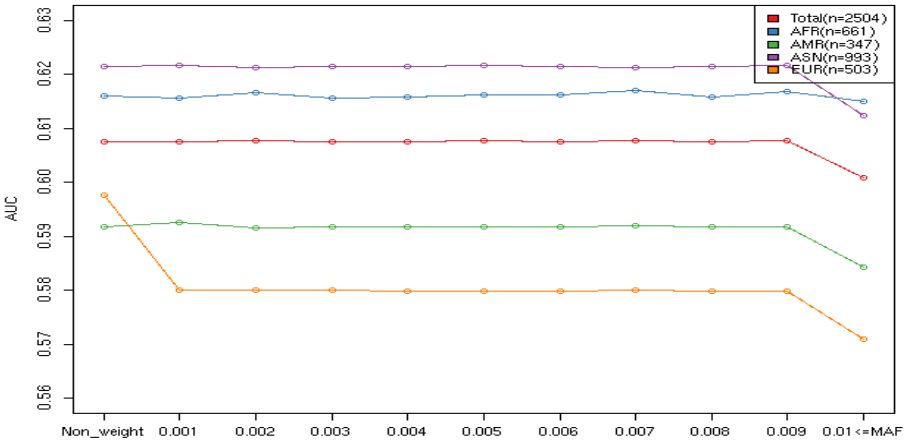

Drug class1
Drug class2
Non class
The deviation from mean AUC
−0.6 −0.2 0 0.2 0.4 0.6
Drug class(n=#drug)
[A](n=62)
[B](n=9)
[C](n=122)
[D](n=25)
[G](n=41)
[J](n=7)
[L](n=43)
[N](n=144)
[R](n=84)
[S](n=22)
[A10](n=39)
[C03](n=23)
[C07](n=22)
[C08](n=16)
[C09A](n=13)
[C10](n=15)
[G03](n=37)
[N02](n=35)
[N05B,N05C](n=49)
[N06A](n=38)
[R03](n=46)
Non-PGx drug(n=97)
PGx drug(n=400)
0.1(0.6074)
0.2(0.6166)
0.3(0.6167)
0.4(0.6157)
0.5(0.6092)
0.6(0.6104)
0.7(0.61)
0.8(0.6083)
0.9(0.6078)

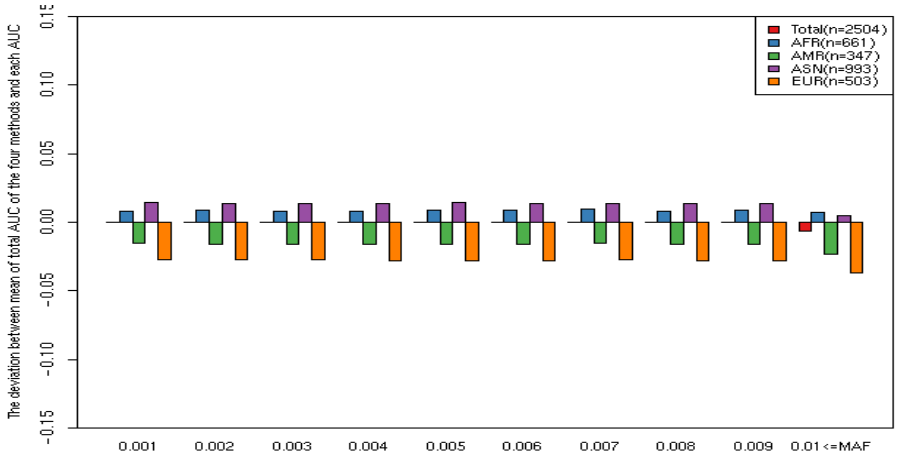
비인종평가


Drug class(n=#drug)
0.1(0.6219)
0.2(0.6364)
0.3(0.6322)
0.4(0.6291)
0.5(0.6259)
0.6(0.6273)
0.7(0.6277)
0.8(0.6271)
0.9(0.6273)
[A](n=62)
[B](n=9)
[C](n=122)
[D](n=25)
[G](n=41)
[H](n=10)
[J](n=7)
[L](n=43)
[M](n=10)
[N](n=144)
[R](n=84)
[S](n=22)
[A02BC](n=4)
[A10](n=39)
[C02](n=22)
[C03](n=23)
[C07](n=22)
[C08](n=16)
[C09A](n=13)
[C10] (n=15)
[G03](n=37)
[N02] (n=35)
[N05B,N05C](n=49)
[N06A](n=30)
[R03](n=46)
Non-PGX drug(n=97)
PGX drug(n=400)

D. 낮은 대립형질 빈도(Minor Allele Frequency)

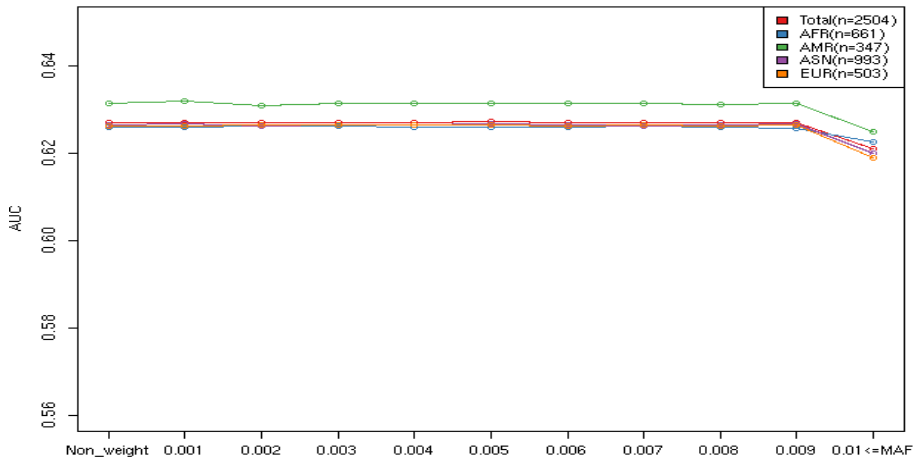
인종평가

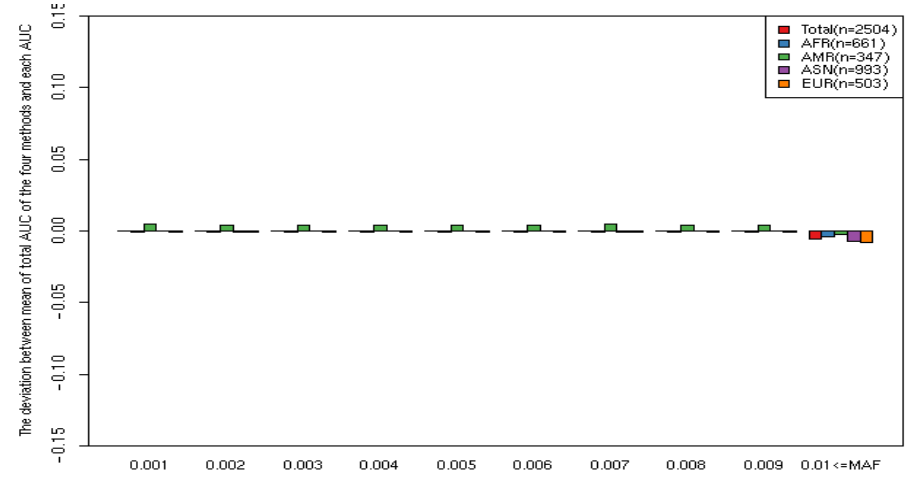




	Non-weight	0.001	0.002	0.003	0.004	0.005	0.006	0.007	0.008	0.009	0.01<=MAF
Total(*n*=2504)	0.6076	0.6076	0.6077	0.6075	0.6075	0.6077	0.6076	0.6078	0.6075	0.6078	0.6008
AFR(*n*=661)	0.6161	0.6156	0.6167	0.6157	0.6159	0.6162	0.6162	0.6171	0.6159	0.6168	0.6151
AMR(*n*=347)	0.5918	0.5926	0.5915	0.5918	0.5918	0.5918	0.5918	0.592	0.5917	0.5918	0.5843
ASN(*n*=993)	0.6214	0.6217	0.6213	0.6215	0.6215	0.6217	0.6214	0.6213	0.6215	0.6216	0.6123
EUR(*n*=503)	0.5799	0.58	0.5801	0.58	0.5799	0.5799	0.5799	0.5801	0.5799	0.5799	0.5709

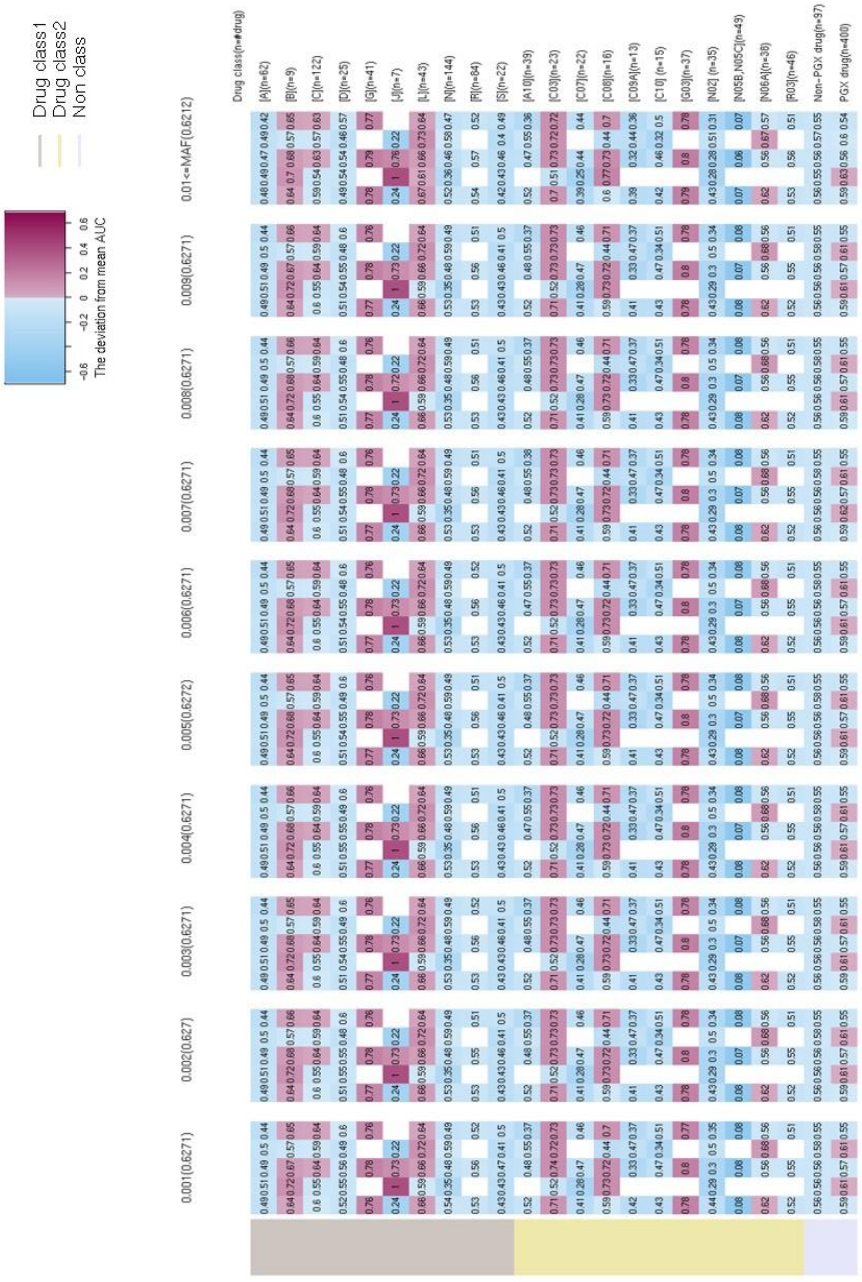
비인종평가



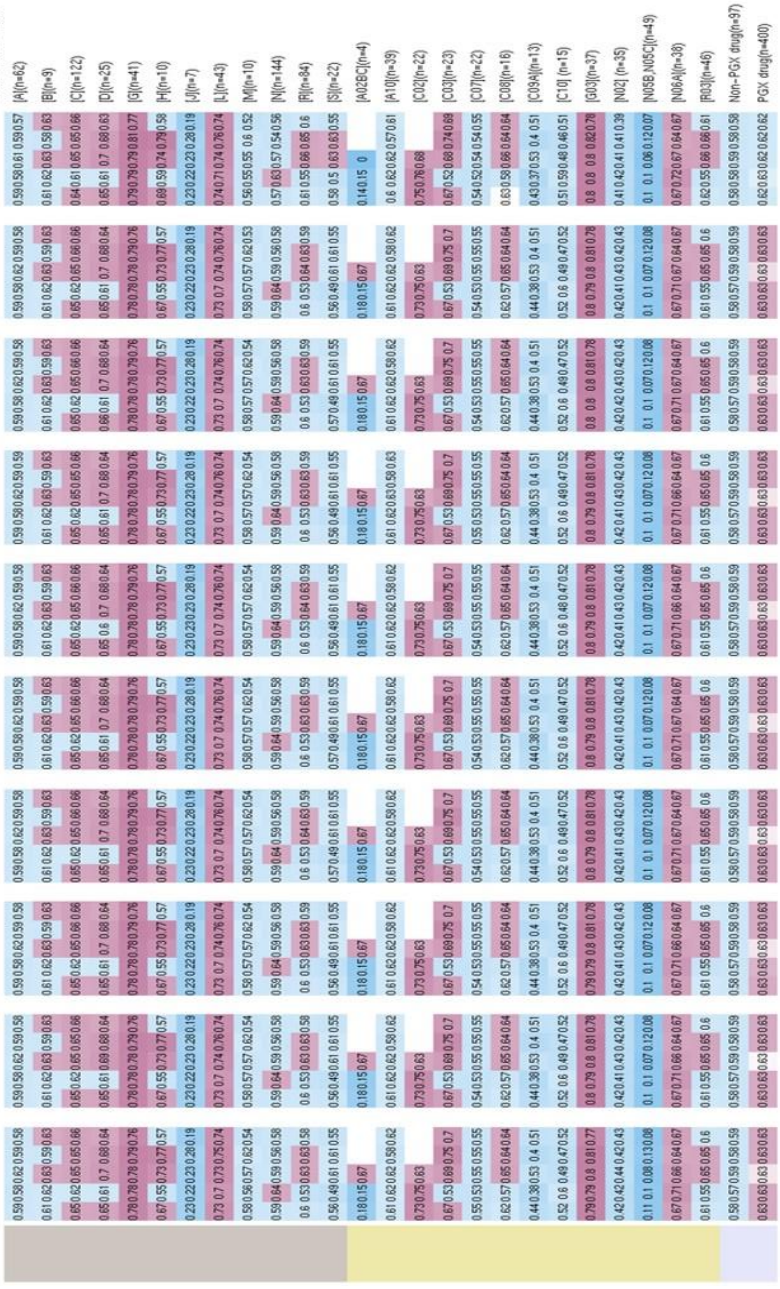


	Non-weight	0.001	0.002	0.003	0.004	0.005	0.006	0.007	0.008	0.009	0.01<=MAF
Total(*n*=2504)	0.6271	0.6271	0.627	0.6271	0.6271	0.6272	0.6271	0.6271	0.6271	0.6271	0.6212
AFR(*n*=661)	0.6261	0.626	0.6262	0.6262	0.626	0.6261	0.626	0.6262	0.6261	0.6258	0.6226
AMR(*n*=347)	0.6314	0.6319	0.631	0.6314	0.6314	0.6314	0.6314	0.6316	0.6313	0.6314	0.6249
ASN(*n*=993)	0.6266	0.6267	0.6264	0.6266	0.6266	0.6268	0.6266	0.6264	0.6266	0.6267	0.6201
EUR(*n*=503)	0.6264	0.6262	0.6265	0.6265	0.6265	0.6265	0.6264	0.6265	0.6264	0.6265	0.619

인종평가

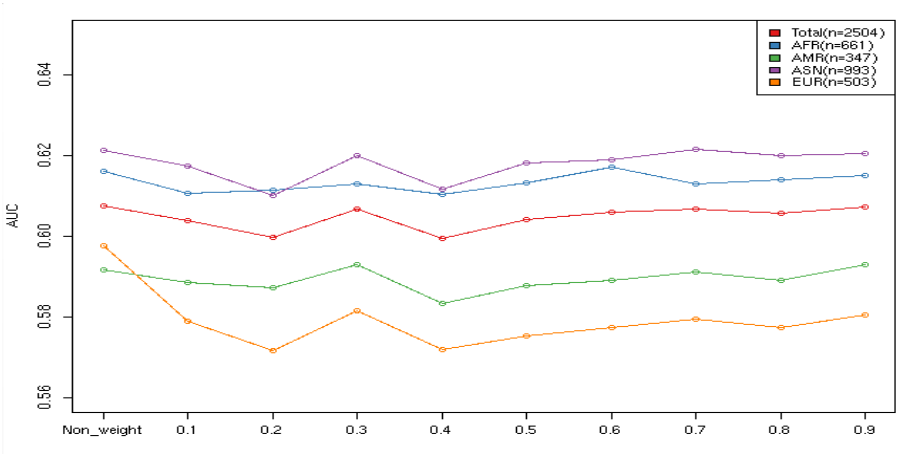

Drug class1
Drug class2
Non class
−0.6 −0.2 0 0.2 0.4 0.6
The deviation from mean AUC
Drug class(n=#drug)
0.001(0.6271)
0.002(0.627)
0.003(0.6271)
0.004(0.6271)
0.005(0.6272)
0.006(0.6271)
0.007(0.6271)
0.008(0.6271)
0.009(0.6271)
0.01<=MAF(0.6212)
[A](n=82)
[B](n=9)
[C](n=122)
[D](n=25)
[G](n=41)
[J](n=7)
[L](n=43)
[N](n=144)
[R](n=84)
[S](n=22)
[A10](n=39)
[C03](n=23)
[C07](n=22)
[C08](n=16)
[C09A](n=13)
[C10] (n=15)
[G03](n=37)
[N02] (n=35)
[N05B,N05C](n=49)
[N06A](n=38)
[P03](n=46)
Non-PGX drug(n=97)
PGX drug(n=400)

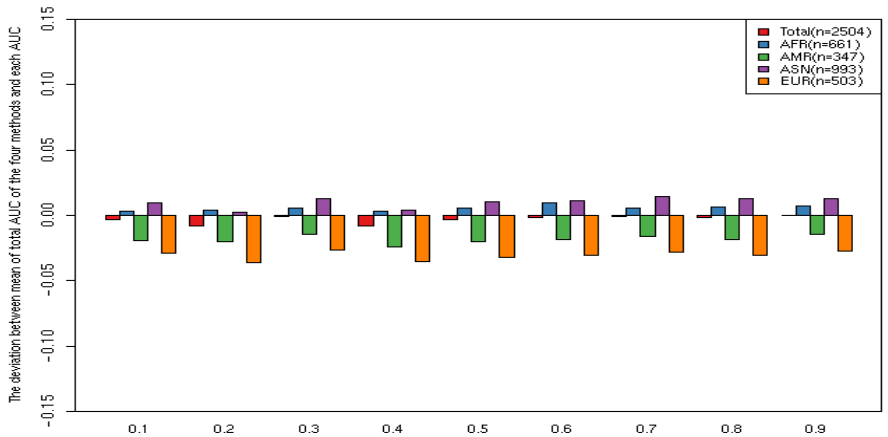
비인종평가


0.001(0.6271)
0.002(0.627)
0.003(0.6271)
0.004(0.6271)
0.005(0.6272)
0.006(0.6271)
0.007(0.6271)
0.008(0.6271)
0.009(0.6271)
0.01<=MAF(0.6212)
Drug class(n=#drug)
[A](n=62)
[B](n=9)
[C](n=122)
[D](n=25)
[G](n=41)
[H](n=10)
[J](n=7)
[L](n=43)
[M](n=10)
[N](n=144)
[R](n=84)
[S](n=22)
[A02BC](n=4)
[A10](n=39)
[C02](n=22)
[C03](n=23)
[C07](n=22)
[C08](n=16)
[C09A](n=13)
[C10] (n=15)
[G03](n=37)
[N02] (n=35)
[N05B,N05C](n=49)
[N06A](n=38)
[R03](n=46)
Non-PGx drug(n=97)
PGx drug(n=400)

E. 동형접합변이 비율 (Homozygote mutation rate)

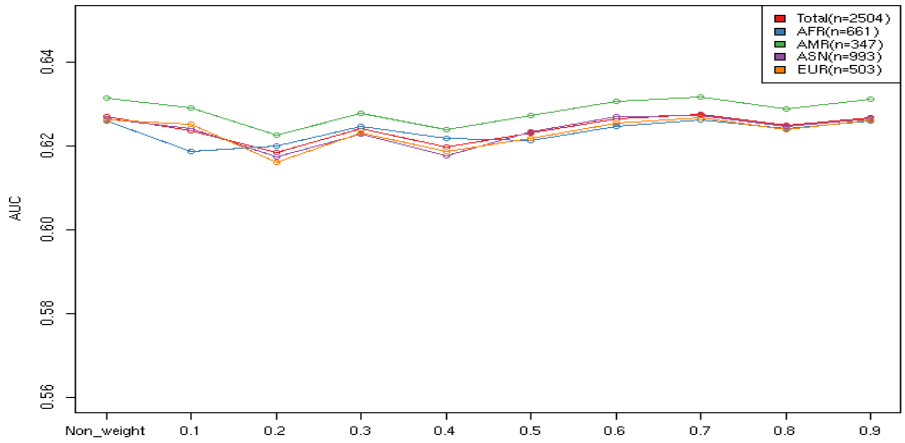
인종평가

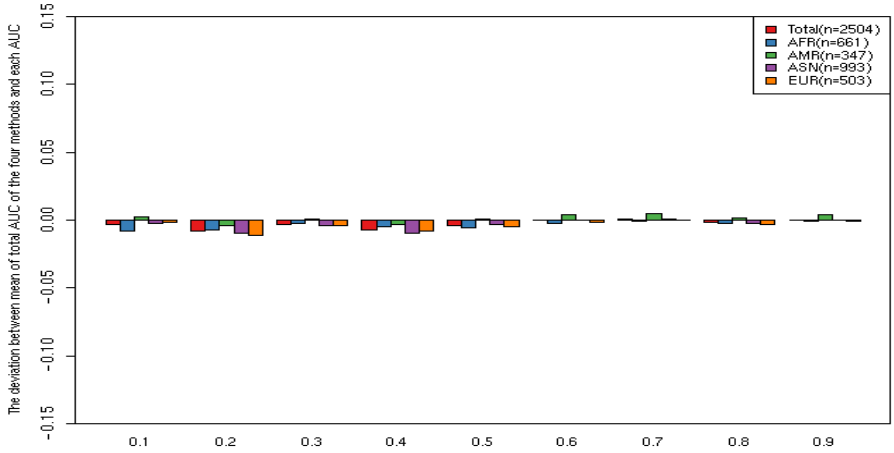




	Non-weight	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Total(*n*=2504)	0.6076	0.604	0.5997	0.6067	0.5995	0.6041	0.6061	0.6067	0.6057	0.6073
AFR(*n*=661)	0.6161	0.6108	0.6115	0.613	0.6104	0.6132	0.6172	0.6131	0.6142	0.6151
AMR(*n*=347)	0.5918	0.5885	0.5873	0.593	0.5835	0.5879	0.5891	0.5912	0.5891	0.5931
ASN(*n*=993)	0.6214	0.6176	0.6103	0.6201	0.6118	0.6182	0.6191	0.6217	0.6202	0.6207
EUR(*n*=503)	0.5799	0.5789	0.5717	0.5815	0.572	0.5754	0.5775	0.5795	0.5774	0.5806

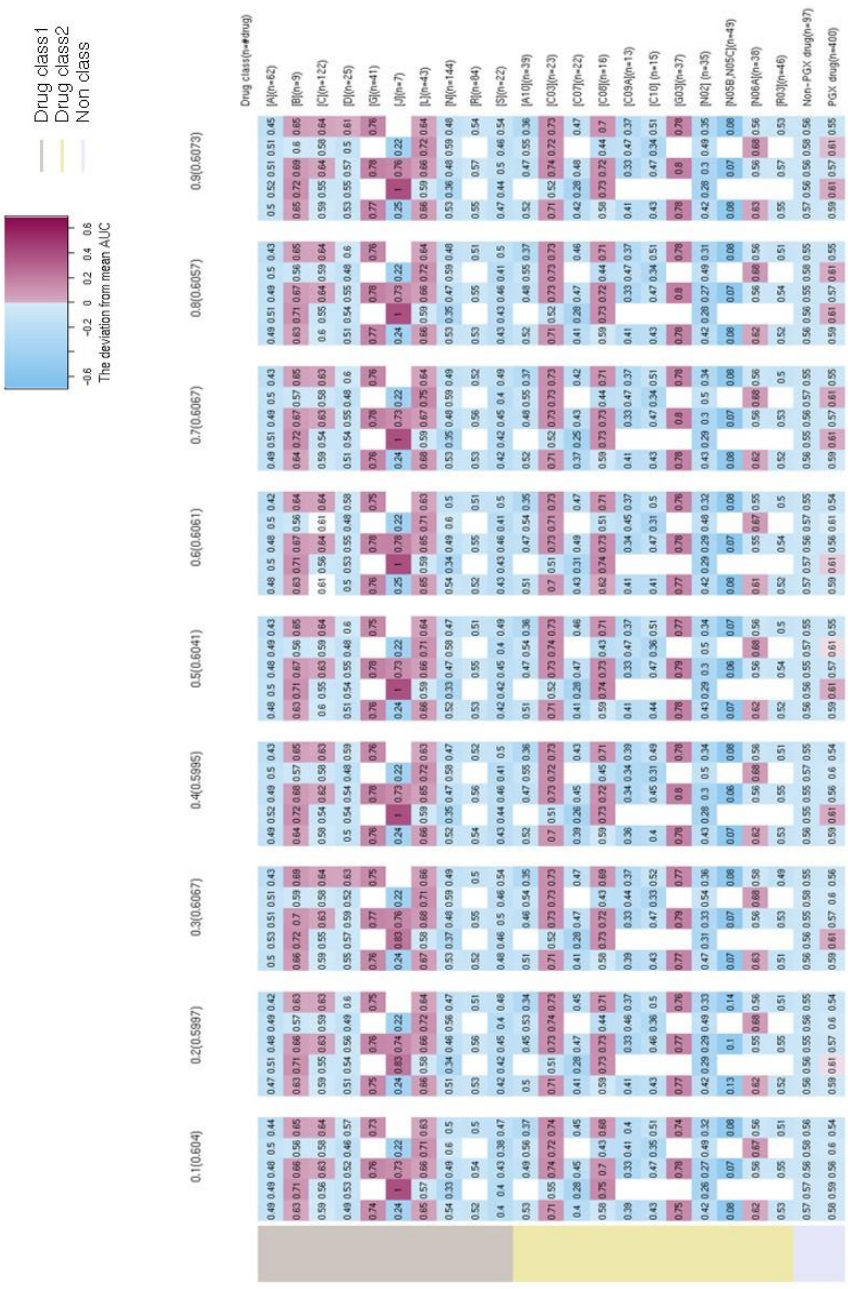
비인종평가



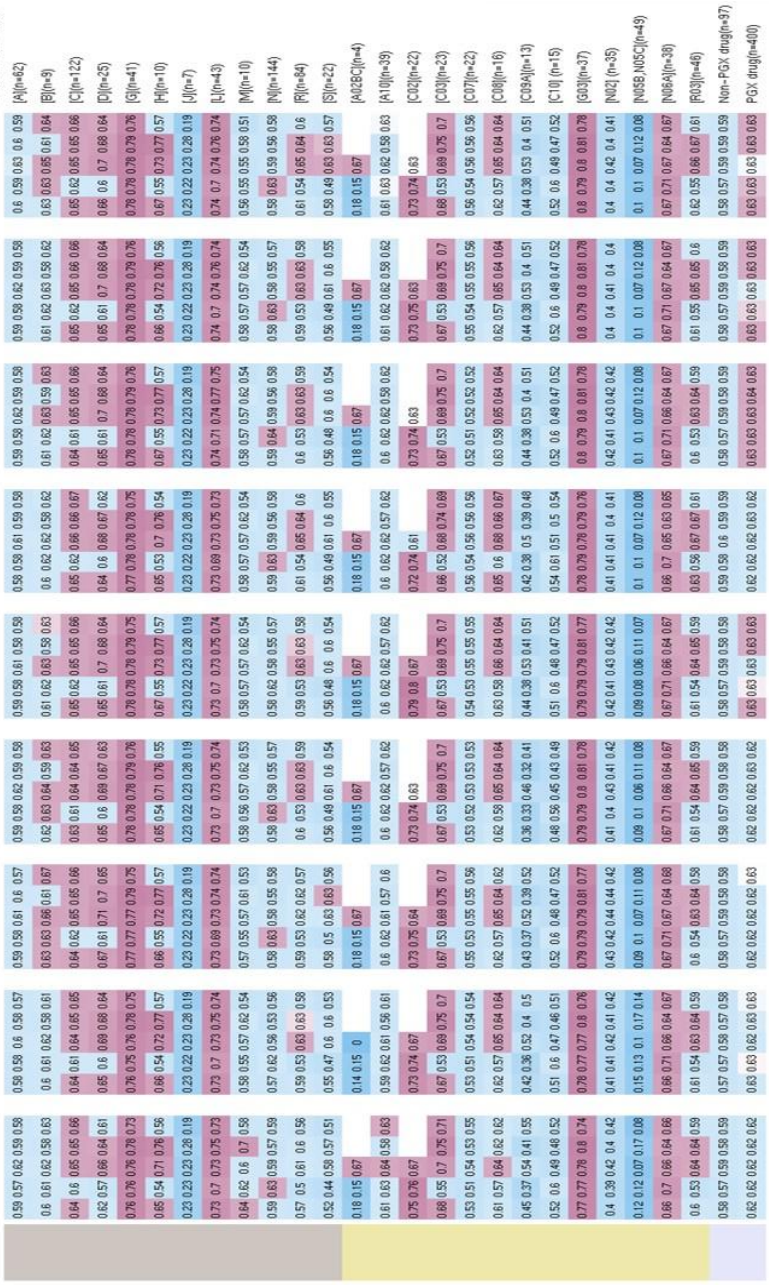


	Non-weight	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Total(n=2504)	0.6271	0.6236	0.6186	0.6241	0.6199	0.6231	0.6266	0.6276	0.625	0.6269
AFR(n=661)	0.6261	0.6187	0.6201	0.6248	0.6219	0.6214	0.6247	0.6262	0.6243	0.6259
AMR(n=347)	0.6314	0.6291	0.6226	0.6279	0.624	0.6274	0.6307	0.6317	0.6289	0.6311
ASN(n=993)	0.6266	0.6242	0.6174	0.6228	0.6177	0.6234	0.627	0.6274	0.6247	0.6266
EUR(n=503)	0.6264	0.6253	0.6161	0.6231	0.6188	0.6218	0.6254	0.6268	0.6239	0.6263

인종평가

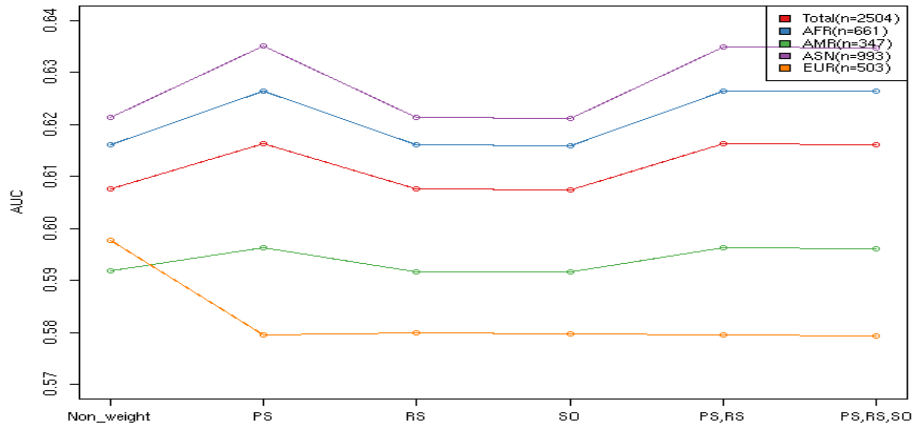

Drug class1
Drug class2
Non class
−0.6 −0.2 0 0.2 0.4 0.6
The deviation from mean AUC
0.1(0.604)
0.2(0.5997)
0.3(0.6067)
0.4(0.5985)
0.5(0.6041)
0.6(0.6061)
0.7(0.6067)
0.8(0.6057)
0.9(0.6073)
Drug class(n=#drug)
[A](n=62)
[B](n=9)
[C](n=122)
[D](n=25)
[G](n=41)
[J](n=7)
[L](n=43)
[N](n=144)
[R](n=64)
[S](n=22)
[A10](n=39)
[C03](n=23)
[C07](n=22)
[C08](n=16)
[C09A](n=13)
[C10] (n=15)
[G03](n=37)
[M02] (n=35)
[N05B,N05C](n=49)
[N06A](n=39)
[R03](n=46)
Non-PGX drug(n=97)
PGX drug(n=400)

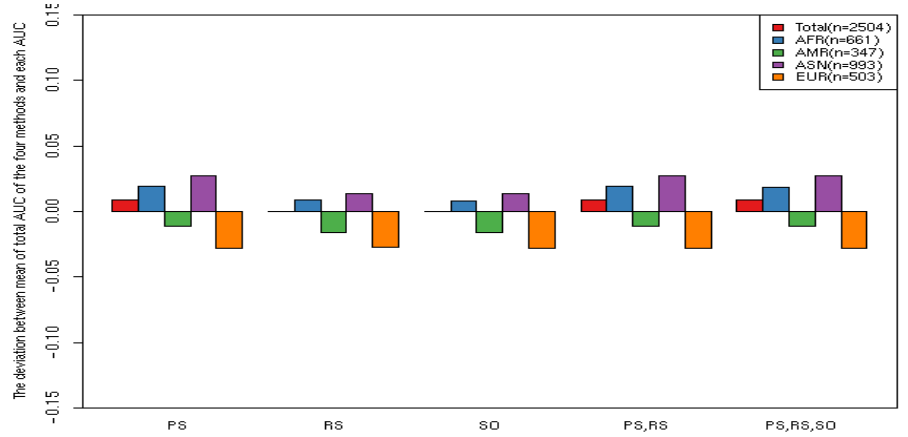
비인종평가


0.1(0.6236)
0.2(0.6186)
0.3(0.6241)
0.4(0.6199)
0.5(0.6231)
0.6(0.6266)
0.7(0.6276)
0.8(0.625)
0.9(0.6269)
Drug class(n=#drug)
[A](n=62)
[B](n=9)
[C](n=122)
[D](n=25)
[G](n=41)
[H](n=10)
[J](n=7)
[L](n=43)
[M](n=10)
[N](n=144)
[R](n=84)
[S](n=22)
[A02BC](n=4)
[A10](n=39)
[C02](n=22)
[C03](n=23)
[C07](n=22)
[C08](n=16)
[C09A](n=13)
[C10] (n=15)
[G03](n=37)
[N02] (n=35)
[N05B,N05C](n=49)
[N06A](n=39)
[R03](n=46)
Non-PGx drug(n=97)
PGx drug(n=400)

F. mRNA 안정성 조절 기전(Nonsense-mediated mRNA decay)

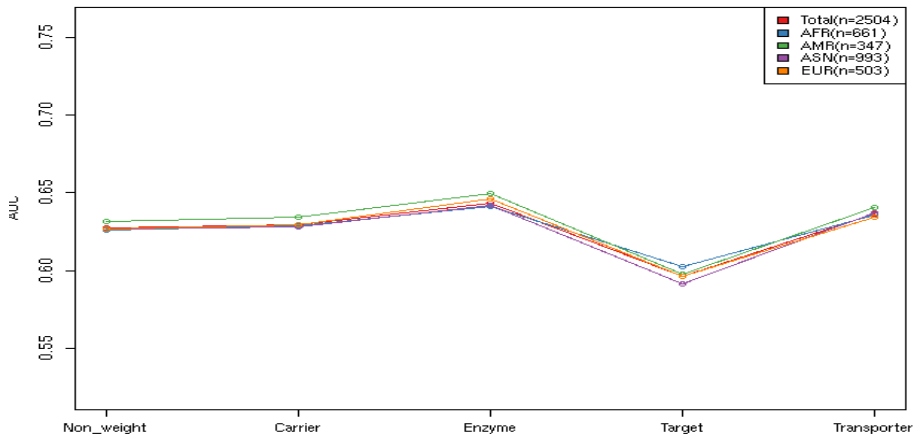
인종평가

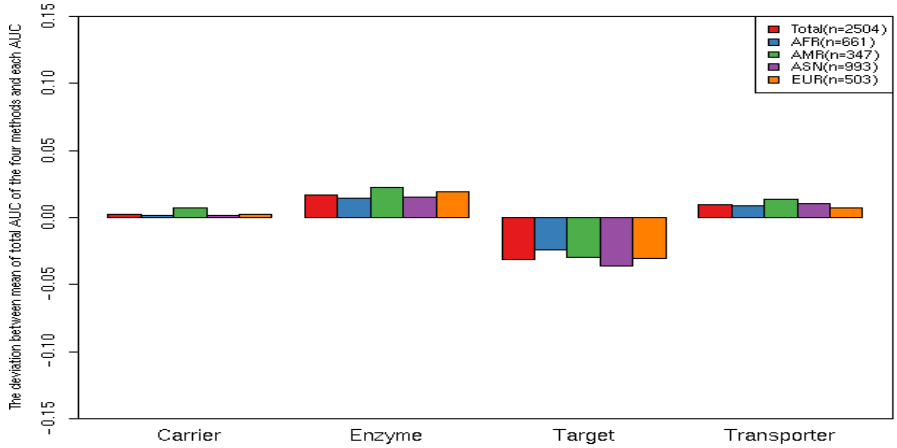




	Non-weight	PS	RS	SO	PS,RS	PS,RS,SO
Total(*n*=2504)	0.6076	0.6163	0.6076	0.6074	0.6163	0.6161
AFR(*n*=661)	0.6161	0.6265	0.6161	0.616	0.6265	0.6264
AMR(*n*=347)	0.5918	0.5964	0.5917	0.5916	0.5963	0.5961
ASN(*n*=993)	0.6214	0.6351	0.6214	0.6212	0.635	0.6348
EUR(*n*=503)	0.5799	0.5795	0.58	0.5798	0.5795	0.5794

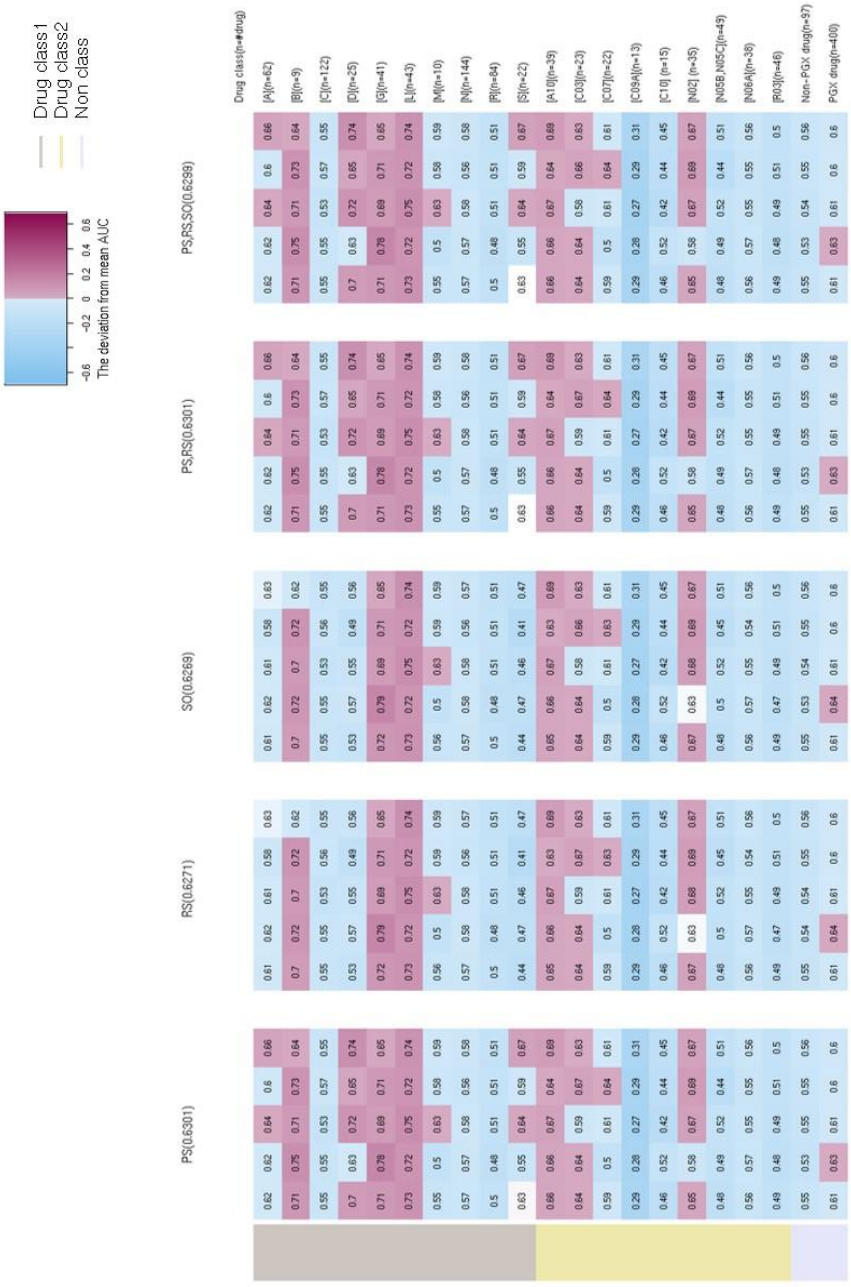
비인종평가



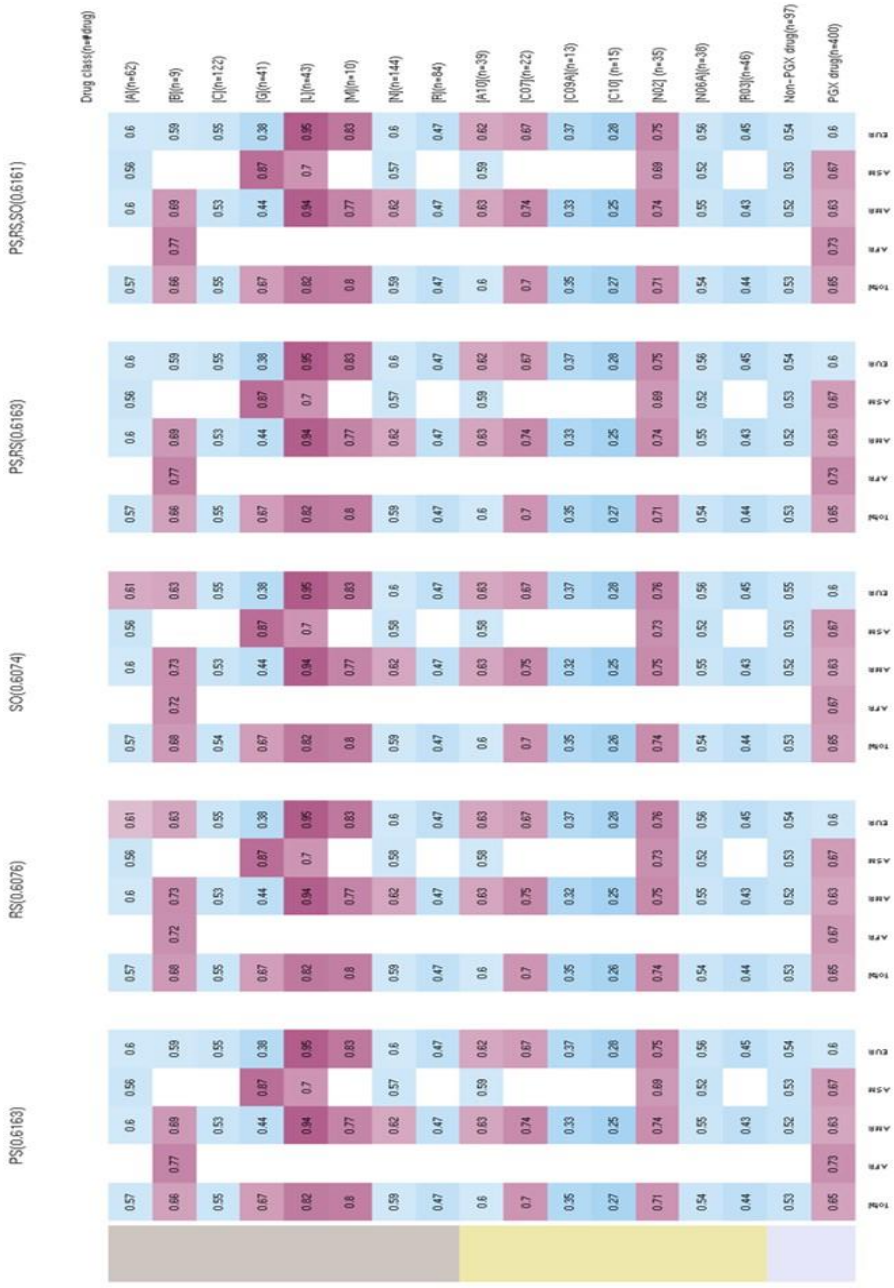


	Non-weight	PS	RS	SO	PS,RS	PS,RS,SO
Total(*n*=2504)	0.6271	0.6301	0.6271	0.6269	0.6301	0.6299
AFR(*n*=661)	0.6261	0.6271	0.6261	0.6259	0.6271	0.627
AMR(*n*=347)	0.6314	0.6329	0.6313	0.6312	0.6328	0.6326
ASN(*n*=993)	0.6266	0.6322	0.6265	0.6263	0.6322	0.6319
EUR(*n*=503)	0.6264	0.6281	0.6265	0.6263	0.6281	0.628

인종평가

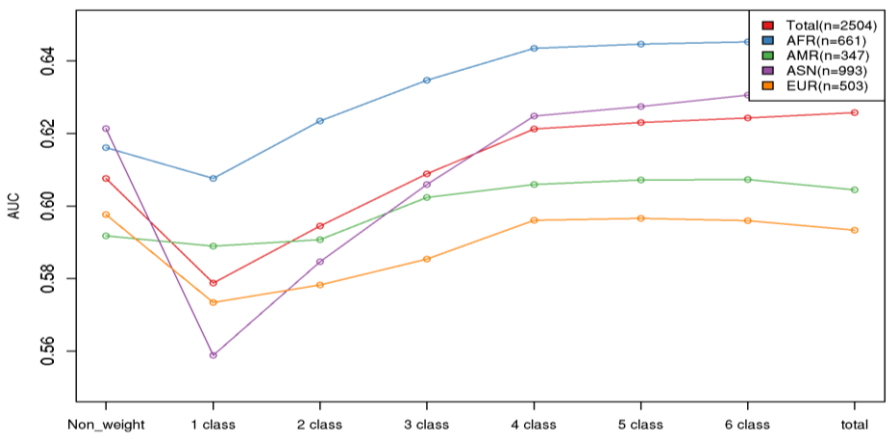

Drug class1
Drug class2
Non class
-0.6 -0.2 0 0.2 0.4 0.6
The deviation from mean AUC
PS(0.6301)
PS(0.6271)
SO(0.6269)
PS,PS(0.6301)
PS,RS,SO(0.6239)
Drug class(n=#drug)
[A](n=62)
[B](n=9)
[C](n=122)
[D](n=25)
[G](n=41)
[L](n=43)
[M](n=10)
[N](n=144)
[R](n=84)
[S](n=22)
[A10](n=39)
[C03](n=23)
[C07](n=22)
[C09A](n=13)
[C10](n=15)
[N02] (n=35)
[N05B,N05C](n=49)
[N06A](n=36)
[R03](n=46)
Non-PGX drug(n=97)
PGX drug(n=100)
TWN AFR AMR ASN EUR

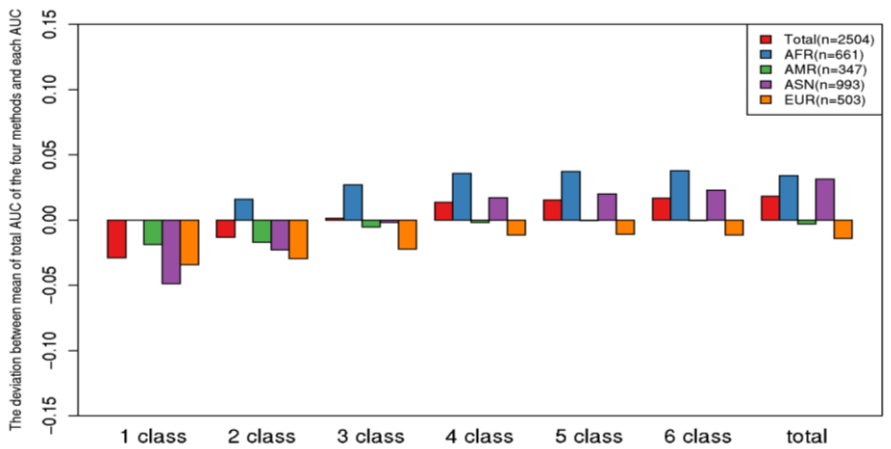
비인종평가


PS;RS;SO(0.6161)
PS;RS(0.6163)
SO(0.6074)
RS(0.6076)
PS(0.6163)
Drug class(n=#drug)
[A](n=62)
[B](n=9)
[C](n=122)
[G](n=41)
[L](n=43)
[M](n=10)
[N](n=144)
[R](n=84)
[A10](n=39)
[C07](n=22)
[C09A](n=13)
[C10] (n=15)
[N02] (n=35)
[N06A](n=38)
[R03](n=46)
Non-PGX drug(n=97)
PGX drug(n=400)
EUR
ASN
AMR
AFR
Total

G. 비번역 변이를 포함한 유전자 기능 조절 변이 (Regulatory variants including noncoding region variants)

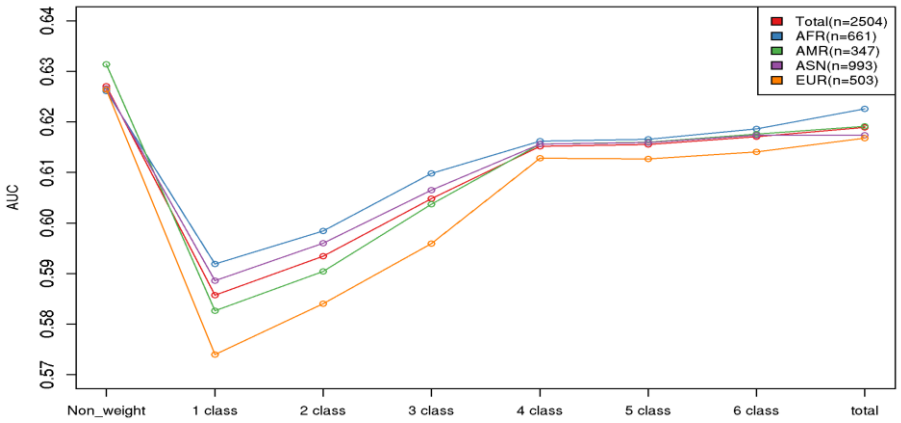
인종평가

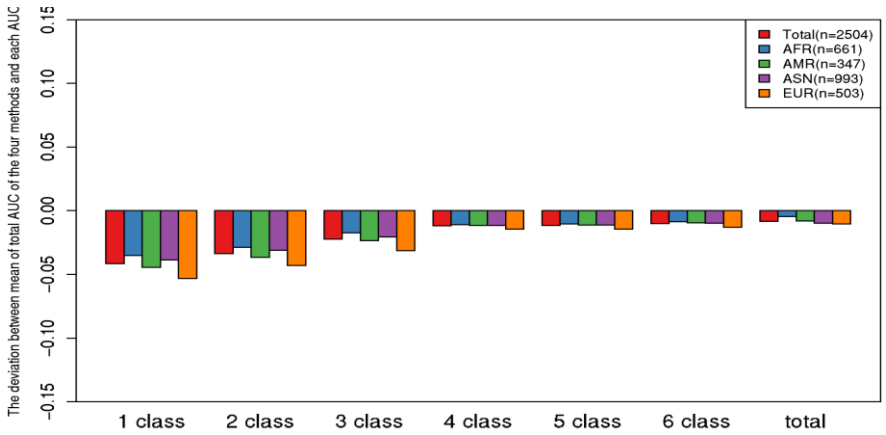




	Non−weight	1 class	2 class	3 class	4 class	5 class	6 class	total
Total(*n*=2504)	0.6076	0.6089	0.623	0.6243	0.6258	0.6213	0.5945	0.5788
AFR(*n*=661)	0.6161	0.6347	0.6447	0.6453	0.6416	0.6435	0.6235	0.6076
AMR(*n*=347)	0.5918	0.6024	0.6072	0.6073	0.6045	0.6059	0.5907	0.589
ASN(*n*=993)	0.6214	0.6059	0.6275	0.6306	0.6391	0.6248	0.5847	0.5588
EUR(*n*=503)	0.5799	0.5854	0.5966	0.596	0.5934	0.5961	0.5782	0.5734

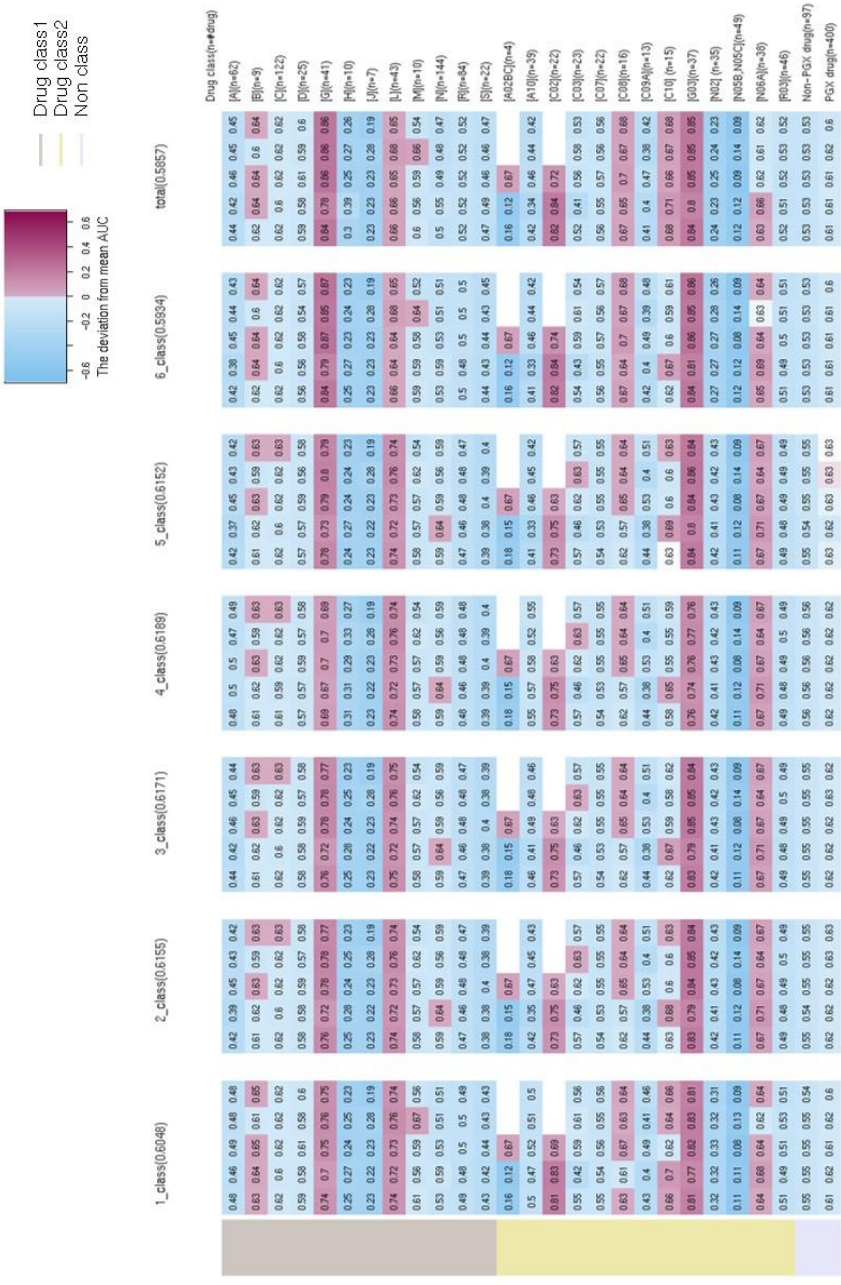
비인종평가



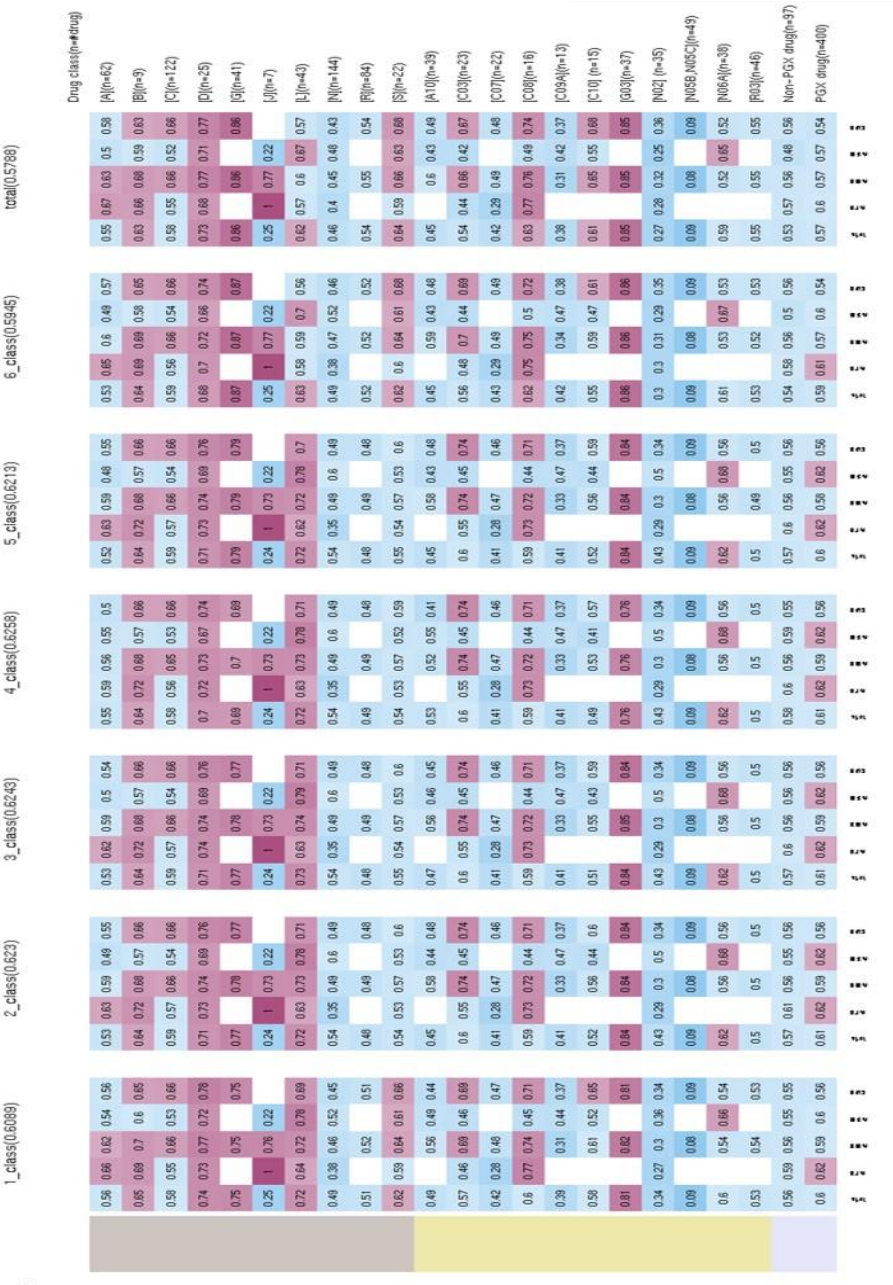


	Non−weight	1 class	2 class	3 class	4 class	5 class	6 class	total
Total(*n*=2504)	0.6271	0.6048	0.6155	0.6171	0.6189	0.6152	0.5934	0.5857
AFR(*n*=661)	0.6261	0.6098	0.6166	0.6186	0.6226	0.6162	0.5984	0.5919
AMR(*n*=347)	0.6314	0.6037	0.616	0.6176	0.6191	0.6156	0.5904	0.5826
ASN(*n*=993)	0.6266	0.6065	0.6159	0.6174	0.6174	0.6156	0.596	0.5886
EUR(*n*=503)	0.6264	0.5959	0.6127	0.6141	0.6168	0.6128	0.584	0.574

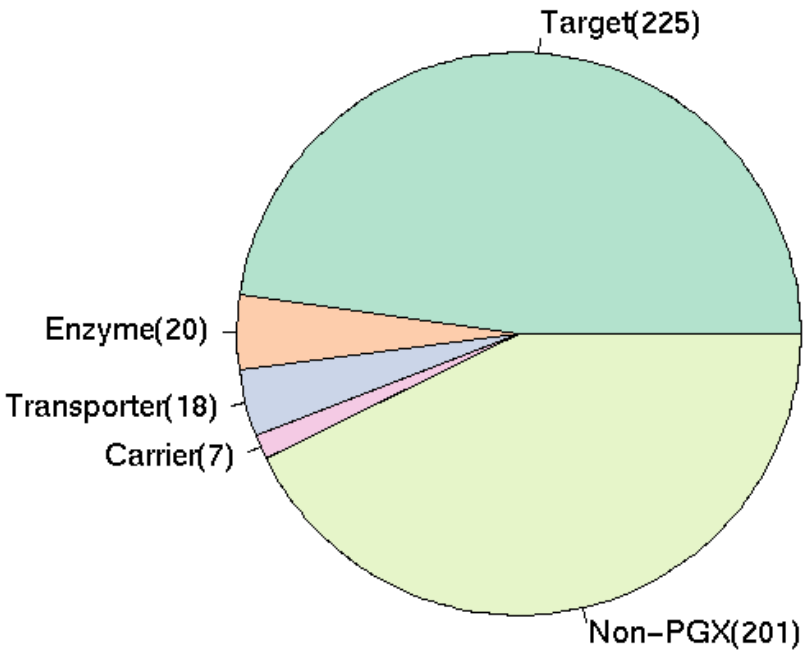
인종평가


Drug class1
Drug class2
Non class
-0.6 -0.2 0 0.2 0.4 0.6
The deviation from mean AUC
1_class(0.6048)
2_class(0.6155)
3_class(0.6171)
4_class(0.6189)
5_class(0.6152)
6_class(0.5934)
total(0.5857)
Drug class(n=#drug)
[A](n=62)
[B](n=9)
[C](n=122)
[D](n=25)
[G](n=41)
[H](n=10)
[J](n=7)
[L](n=43)
[M](n=10)
[N](n=144)
[R](n=84)
[S](n=22)
[A02BC](n=4)
[A10](n=39)
[C02](n=22)
[C03](n=23)
[C07](n=22)
[C09](n=16)
[C09A](n=13)
[C10] (n=15)
[G03](n=37)
[N02] (n=35)
[N05B,N05C](n=49)
[N06A](n=36)
[R03](n=46)
Non-PGX drug(n=97)
PGX drug(n=400)

비인종평가


Drug class(n=#drug)
[A](n=62)
[B](n=9)
[C](n=122)
[D](n=25)
[G](n=41)
[J](n=7)
[L](n=43)
[N](n=144)
[R](n=84)
[S](n=22)
[A10](n=39)
[C03](n=23)
[C07](n=22)
[C08](n=16)
[C09A](n=13)
[C10] (n=15)
[G03](n=37)
[N02] (n=35)
[N05B, N05C](n=49)
[N06A](n=36)
[R03](n=46)
Non-PGX drug(n=97)
PGX drug(n=400)
total(0.5788)
6_class(0.5945)
5_class(0.6213)
4_class(0.6258)
3_class(0.6243)
2_class(0.623)
1_class(0.6089)

보충 그림 6. PharmGKB로부터 추출한 1807개의 변이-약물 연관관계에 포함된 변이가 속한 유전자 471개의 약물학적 유전자 카테고리별 분포.



보충 표

보충 표 1. 2014 1000 지놈 데이터의 인종별 개인유전체 갯수

Sub population	Super population	No. of personal genome data
GWD	AFR	113
YRI	AFR	108
ESN	AFR	99
LWK	AFR	99
ACB	AFR	96
MSL	AFR	85
ASW	AFR	61
PUR	AMR	104
CLM	AMR	94
PEL	AMR	85
MXL	AMR	64
CHS	ASN	105
JPT	ASN	104
CHB	ASN	103
GIH	ASN	103
ITU	ASN	102
STU	ASN	102
KHV	ASN	99
PJL	ASN	96

CDX	ASN	93
BEB	ASN	86
IBS	EUR	107
TSI	EUR	107
CEU	EUR	99
FIN	EUR	99
GBR	EUR	91
	Total	
	2504	

26 2014 1000 지놈 데이터의 하위 26개, 상위 4개 인종군 정보. (2015 6). 상위 인종군 : AFR(African), EUR(European),ASN(EAS;East Asian),ASN(SAS;South Asian), AMR(Ad Mixed American). 하위 인종군 : AFR(YRI(Yoruba in Ibadan, Nigeria),LWK(Luhya in Webuye, Kenya),GWD(Gambian in Western Divisions in the Gambia),MSL(Mende in Sierra Leone),ESN(Esan in Nigeria),ASW(Americans of African Ancestry in SW USA),ACB(African Caribbeans in Barbados)),EUR(CEU(Utah Residents (CEPH) with Northern andWestern European Ancestry),TSI(Toscani in Italia), FIN(Finnish in Finland),GBR(British in England and Scotland),IBS(Iberian Population in Spain)), ASN(EAS,SAS;CHB(Han Chinese in Bejing, China),JPT(apanese in Tokyo, Japan),CHS(Southern Han Chinese),CDX(Chinese Dai in Xishuangbanna, China),KHV(Kinh in Ho Chi Minh City, Vietnam), GIH(Gujarati Indian from Houston, Texas),PJL(Punjabi from Lahore, Pakistan),BEB(Bengali from Bangladesh), STU(Sri Lankan Tamil from the UK), ITU(Indian Telugu from the UK)),AMR(MXL(Mexican Ancestry from Los Angeles USA),PUR(Puerto Ricans from Puerto Rico),CLM(Colombians from Medellin, Colombia),PEL(Peruvians from Lima, Peru)).

보충 표 2. ATC 약물 분류군별 PharmGKB로 부터 추출한 약물조절변이 갯수

Drug class	ATC code	No. of drugs	PharmGKB*
Alimentary tract and metabolism	A	62	77
Blood and blood forming organs	B	9	99
Cardiovascular system	C	122	291
Dermatologicals	D	25	56
Genito urinary system and sex hormones	G	41	9
Systemic hormonal preparations, excl. sex hormones and insulins	H	10	11
antiinfectives for systemic use	J	7	5
Antineoplastic and immunomo- dulating agents	L	43	294
Musculo-skeletal system	M	10	7
Nervous system	N	144	504
Antiparasitic products, insecticides and repellents	P	2	0
Respiratory system	R	84	58
Sensory organs	S	22	63
Various	V	3	1
Sub-total		493	1475
Others(unclassified)		4	205

* Number of existing genetic-variant-drug associations in PharmGKB 2015 version(2015, 1)

보충 표 3. ATC 자주 처방받은 약물 분류군별 PharmGKB로 부터 추출한 약물조절변이 갯수

Drug class	ATC code	No. of drugs	PharmGKB*
Proton pump inhibitors	A02BC	4	12
Drugs used in diabetes	A10	39	26
Antihypertensives	C02	22	6
Diuretics	C03	23	33
Beta blocking agents	C07	22	50
Calcium channel blockers	C08	16	28
ACE inhibitors, plain	C09A	13	35
Lipid modifying agents	C10	15	125
Sex hormones and modulators of the genital system	G03	37	9
Thyroid therapy	H03	5	1
Analgesics	N02	35	131
Anxiolytics and hypnotics and sedatives	N05B/ N05	49	1
Antidepressants	N06A	38	265
Drugs for obstructive airway diseases	R03	46	40
Antihistamines for systemic use	R06	35	1
Sub-total		395	763
Others(unclassified)		102	917

* Number of existing genetic-variant-drug associations in PharmGKB 2015 version(2015, 1)

보충 표 4. Pharmsafe & 가중치 Pharmsafe 알고리즘에 쓰인 기호.

Symbol	Definition
v_i	Variant i
g_j	Gene j
S_{v_i}	SIFT score of variant i
S_{g_j}	Damaged score of gene(gene score) j
S_{d_k}	Damaged score of drug(drug score) k
G_j	Set of variant with SIFT score in gene j
D_k	Set of gene related with drug k
WS_{v_i}	Winsorized variant score of variant i
W_{v_i}	Weight score of variant i
W_{g_j}	Weight score of gene j
V_{PGT}	Set of variants in PharmacoGene j region
SR	SIFT score filter Range
MAF_{v_i}	Minor Allele Frequency of variant i in 1000 genome (n=2504)
$MAFR$	Minor Allele Frequency Range in in 1000 genome(n=2504)
V_{MAFR}	Set of variants with SIFT score in $MAFR_l$
HVF_{v_i}	Homozygote Variant Frequency of variant i in 1000 genome (n=2504)

$HVFR$	Homozygote Variant Frequency Range in 1000 genome(m=2504)
HV_{HVFR}	Set of Homozygote Variants with SIFT score in $HVFR_l$
V_{NMD}	Set of NMD variants with SIFT score
RG_j	Gene j is regulated by *variant i*
RS_{v_i}	Regulation score of variant *i*
RV	Variants with the ability to regulated the gene
RC	Regulation variant Classs
RV_i	Regulation variant *i*
WS_{RG_j}	Weight score of regulated gene *j*
$EAD_{E_{c\ l}}$	Ethic testing AUC diviation of condition *l* in E element
$NAD_{E_{c\ l}}$	Non-Ethic testing AUC diviation of condition *l* in E element
SEA	Standard Ethnic AUC
SNA	Standard Non-ethnic AUC
$EA_{E_{cl}}$	Ethnic AUC of condition *l* in element e
$NA_{E_{cl}}$	Non-ethnic AUC of condition *l* in element e
$CB_{m_{v_i}}$	Variant in *i* Conbination m

* pharmacogene type consist of target,transporter,enzyme,carrier.

보충 표 5. PharmGKB, 497 약물, 1000 지놈에 공통적으로 속하는 유전적 변이-약물 연관 갯수

Element	Pharm GKB*	497 drugs (%)+	1000 Genome ∩ 497 drugs(%)+
Genetic-variant-drug associations	3248	1807(55.63)	1100(33.86)
Black or African American associations	83	58(69.87)	40(48.19)
Asian associations	515	329(63.88)	224(43.49)
White associations	647	451(69.70)	303(46.83)
Hispanic or Latino associations	6	5(83.33)	3(50)
Others(unclassified)	1997	1128(56.48)	652(32.64)
Drugs	391	290(74.18)	251(64.19)
Variants	1175	840(71.48)	522(44.42)

* Number of existing genetic-variant-drug associations in PharmGKB 2015 version(2015, 1)

+ Percentage of genetic-variant-drug associations in 497 drugs or 497 drugs and 1000 Genome from ParmGKB

보충 표 6. 7가지 요소에 속하는 54가지 조건별 인종/비인종 평가 지수.

Weight parameter	Element of each parameter	Ethnic AUC	Non-ethnic AUC
Central Tendency method	Arithmetic mean	0.5633	0.5935
	Geometric mean	0.6076	0.6271
	Harmonic mean	0.6149	0.6363
	product	0.6163	0.6502
Pharmacogene Type	Target	0.5961	0.5928
	Enzyme	0.6436	0.6248
	Transporter	0.6367	0.6134
	Carrier	0.6296	0.6074
Score Winsorization	0.1	0.6074	0.6219
	0.2	0.6166	0.6364
	0.3	0.6167	0.6322
	0.4	0.6157	0.6291
	0.5	0.6092	0.6259
	0.6	0.6104	0.6273
	0.7	0.61	0.6277
	0.8	0.6083	0.6271
	0.9	0.6078	0.6273
Minor Allele Frequency	0.001	0.6076	0.6271
	0.002	0.6077	0.627
	0.003	0.6075	0.6271
	0.004	0.6075	0.6271
	0.005	0.6077	0.6272

	0.006	0.6076	0.6271
	0.007	0.6078	0.6271
	0.008	0.6075	0.6271
	0.009	0.6078	0.6271
	0.01 over	0.6008	0.6212
Homozygote mutation rate	0.1	0.604	0.6236
	0.2	0.5997	0.6186
	0.3	0.6067	0.6241
	0.4	0.5995	0.6199
	0.5	0.6041	0.6231
	0.6	0.6061	0.6266
	0.7	0.6067	0.6276
	0.8	0.6057	0.625
	0.9	0.6073	0.6269
Nonsense-mediated mRNA decay	Premature Stop codons	0.6163	0.6301
	Removed Stop codons	0.6076	0.6271
	Splice-Overlap	0.6074	0.6269
	Premature Stop codons, Removed Stop codons	0.6163	0.6301
	Premature Stop codons, Removed Stop codons, Splice-Overlap	0.6161	0.6299
Regulatory variants	1	0.6089	0.6048

including noncoding	2	0.623	0.6155
region variants	3	0.6243	0.6171
	4	0.6258	0.6189
	5	0.6213	0.6152
	6	0.5945	0.5934
	Total	0.5788	0.5857

보충 표 7. 각 요소의 조건별 인종, 비인종 AUC 및 편차 및 편차 평균.

SW : 변이 점수 원저화, MAF : 낮은 대립형질 빈도 , PGT : 약물학적 유전자 종류, HR : 동형접합변이 빈도,NMD : 7 mRNA 안정성 조절 기전, RC : 유전자 기능 조절 변이

요소	조건	인종평가 전체 AUC	비인종평가 전체 AUC	인종평가 편차	비인종평가 편차	평균편차
SW	0.1	0.6074	0.6219	−0.0002	−0.0052	−0.0027
	0.2	0.6166	0.6364	0.009	0.0093	0.00915
	0.3	0.6167	0.6322	0.0091	0.0051	0.0071
	0.4	0.6157	0.6291	0.0081	0.002	0.00505
	0.5	0.6092	0.6259	0.0016	−0.0012	0.0002
	0.6	0.6104	0.6273	0.0028	0.0002	0.0015
	0.7	0.61	0.6277	0.0024	0.0006	0.0015
	0.8	0.6083	0.6271	0.0007	0	0.00035
	0.9	0.6078	0.6273	0.0002	0.0002	0.0002
MAF	0.001	0.6076	0.6271	0	0	0
	0.002	0.6077	0.627	1E−04	−1E−04	0
	0.003	0.6075	0.6271	−1E−04	0	−5E−05
	0.004	0.6075	0.6271	−1E−04	0	−5E−05
	0.005	0.6077	0.6272	1E−04	1E−04	1E−04
	0.006	0.6076	0.6271	0	0	0
	0.007	0.6078	0.6271	0.0002	0	1E−04
	0.008	0.6075	0.6271	−1E−04	0	−5E−05
	0.009	0.6078	0.6271	0.0002	0	1E−04
	0.01 ≤MAF	0.6008	0.6212	−0.0068	−0.0059	−0.00635
PGT	표적	0.5961	0.5928	−0.0115	−0.0343	−0.0229
	효소	0.6436	0.6248	0.036	−0.0023	0.01685

	수송체	0.6367	0.6134	0.0291	−0.0137	0.0077
	수송기구	0.6296	0.6074	0.022	−0.0197	0.00115
요소	조건	인종평가 전체 AUC	비인종평가 전체 AUC	인종평가 편차	비인종평가 편차	평균편차
HR	0.1	0.604	0.6236	−0.0036	−0.0035	−0.00355
	0.2	0.5997	0.6186	−0.0079	−0.0085	−0.0082
	0.3	0.6067	0.6241	−0.0009	−0.003	−0.00195
	0.4	0.5995	0.6199	−0.0081	−0.0072	−0.00765
	0.5	0.6041	0.6231	−0.0035	−0.004	−0.00375
	0.6	0.6061	0.6266	−0.0015	−0.0005	−0.001
	0.7	0.6067	0.6276	−0.0009	0.0005	−0.0002
	0.8	0.6057	0.625	−0.0019	−0.0021	−0.002
	0.9	0.6073	0.6269	−0.0003	−0.0002	−0.00025
NMD	PS	0.6163	0.6301	0.0087	0.003	0.00585
	RS	0.6076	0.6271	0	0	0
	SO	0.6074	0.6269	−0.0002	−0.0002	−0.0002
	PS,RS	0.6163	0.6301	0.0087	0.003	0.00585
	PS,RS,SO	0.6161	0.6299	0.0085	0.0028	0.00565
RC	1	0.6089	0.6048	0.0013	−0.0223	−0.0105
	2	0.623	0.6155	0.0154	−0.0116	0.0019
	3	0.6243	0.6171	0.0167	−0.01	0.00335
	4	0.6258	0.6189	0.0182	−0.0082	0.005
	5	0.6213	0.6152	0.0137	−0.0119	0.0009
	6	0.5945	0.5934	−0.0131	−0.0337	−0.0234
	Total	0.5788	0.5857	−0.0288	−0.0414	−0.0351

보충 표 8. 각 요소의 선택 된 6가지 조건의 56가지 조합과 Pharmsafe 알고리즘을 사용해 계산한 개인별 약물 점수를 평가한 인종, 비인종 AUC. (A)인종평가 , (B) 비인종평가 이다. 각 약자는 다음과 같다. SW : 변이 점수 윈저화, MAF : 낮은 대립형질 빈도 , PGT : 약물학적 유전자 종류, HR : 동형접합변이 빈도,NMD : 7 mRNA 안정성 조절 기전, RC : 유전자 기능 조절 변이.

A. 인종평가

순번	조합 종류	전체	아프리카	미국	아시아	유럽
0	SF,MAF	0.614	0.632	0.5961	0.6243	0.5825
1	SF,HR	0.613	0.6293	0.5954	0.6241	0.5818
2	SF,PGX	0.6235	0.6443	0.6042	0.6323	0.5923
3	SF,LOF	0.62	0.6401	0.5987	0.6334	0.5819
4	SF,NC	0.5203	0.4703	0.5549	0.5156	0.5706
5	MAF,HR	0.6047	0.63	0.5865	0.6098	0.5738
6	MAF,PGX	0.6162	0.6431	0.5974	0.6203	0.5858
7	MAF,LOF	0.6117	0.6409	0.5898	0.6192	0.5736
8	MAF,NC	0.5722	0.4464	0.669	0.5891	0.6356
9	HR,PGX	0.641	0.5969	0.6202	0.5853	0.5852
10	HR,LOF	0.6109	0.6385	0.5892	0.6194	0.573
11	HR,NC	0.5722	0.4465	0.6691	0.5892	0.6356
12	PGX,LOF	0.6213	0.6508	0.5994	0.628	0.5847
13	PGX,NC	0.5743	0.4478	0.671	0.593	0.6355
14	LOF,NC	0.5722	0.4466	0.669	0.5891	0.6358
15	SF,MAF,HR	0.6074	0.6353	0.5891	0.6113	0.5757
16	SF,MAF,PGX	0.6156	0.6443	0.5968	0.6188	0.5846
17	SF,MAF,LOF	0.6131	0.6448	0.5916	0.6186	0.5756

18	SF,MAF,NC	0.5709	0.4462	0.6681	0.5861	0.6362
19	SF,HR,PGX	0.6149	0.6425	0.5963	0.6186	0.5841
20	SF,HR,LOF	0.6124	0.6428	0.591	0.6185	0.575
21	SF,HR,NC	0.571	0.4462	0.6682	0.5862	0.6362
22	SF,PGX,LOF	0.6198	0.6511	0.5983	0.6248	0.5838
23	SF,PGX,NC	0.5721	0.4472	0.6693	0.5886	0.6353
24	SF,LOF,NC	0.571	0.4464	0.6681	0.5862	0.6364
25	MAF,HR,PGX	0.6089	0.6408	0.5909	0.6092	0.579
26	MAF,HR,LOF	0.6059	0.6424	0.5845	0.6078	0.5692
27	MAF,HR,NC	0.5712	0.4442	0.6684	0.5889	0.6349
28	MAF,PGX,LOF	0.6138	0.6495	0.5927	0.6154	0.5785
29	MAF,PGX,NC	0.5726	0.445	0.6695	0.5912	0.635
30	MAF,LOF,NC	0.5713	0.4443	0.6682	0.5888	0.635
31	HR,PGX,LOF	0.6133	0.6479	0.5923	0.6153	0.5781
32	HR,PGX,NC	0.5726	0.445	0.6696	0.5912	0.635
33	HR,LOF,NC	0.5713	0.4444	0.6683	0.5888	0.635
34	PGX,LOF,NC	0.5726	0.4451	0.6694	0.5911	0.6352
35	SF,MAF,HR,PGX	0.6087	0.6408	0.5906	0.6091	0.5782
36	SF,MAF,HR,LOF	0.6068	0.6432	0.5859	0.6081	0.5705
37	SF,MAF,HR,NC	0.5701	0.444	0.6676	0.5862	0.6353
38	SF,MAF,PGX,LOF	0.613	0.6488	0.5921	0.6143	0.5779
39	SF,MAF,PGX,NC	0.5706	0.4445	0.6676	0.588	0.6335
40	SF,MAF,LOF,NC	0.5701	0.4441	0.6675	0.5861	0.6355
41	SF,HR,PGX,LOF	0.6125	0.6473	0.5917	0.6142	0.5775
42	SF,HR,PGX,NC	0.5706	0.4445	0.6677	0.588	0.6335
43	SF,HR,LOF,NC	0.5701	0.4441	0.6675	0.5861	0.6355
44	SF,PGX,LOF,NC	0.5706	0.4447	0.6676	0.5879	0.6337

45	MAF,HR,PGX,LOF	0.6074	0.6454	0.5872	0.6064	0.5733
46	MAF,HR,PGX,NC	0.5706	0.4442	0.6677	0.5893	0.6316
47	MAF,HR,LOF,NC	0.5702	0.4435	0.6669	0.5885	0.6325
48	MAF,PGX,LOF,NC	0.5706	0.4443	0.6675	0.5892	0.6318
49	HR,PGX,LOF,NC	0.5707	0.4443	0.6676	0.5892	0.6318
50	SF,MAF,HR,PGX,LOF	0.6071	0.6446	0.587	0.6064	0.573
51	SF,MAF,HR,PGX,NC	0.5694	0.4439	0.6665	0.5871	0.6308
52	SF,MAF,HR,LOF,NC	0.569	0.4435	0.6659	0.5858	0.6325
53	SF,MAF,PGX,LOF,NC	0.5694	0.4441	0.6664	0.587	0.6309
54	SF,HR,PGX,LOF,NC	0.5694	0.4441	0.6664	0.587	0.6309
55	MAF,HR,PGX,LOF,NC	0.5687	0.4444	0.666	0.5858	0.6298
56	SF,MAF,HR,PGX,LOF,NC	0.568	0.4439	0.6652	0.5849	0.6292

B. 비인종평가

순번	조합 종류	전체	아프리카	미국	아시아	유럽
0	SF,MAF	0.6321	0.6308	0.6367	0.6317	0.6315
1	SF,HR	0.6322	0.6307	0.6365	0.6321	0.6313
2	SF,PGX	0.6426	0.6406	0.6477	0.6414	0.644
3	SF,LOF	0.6341	0.6315	0.6373	0.6357	0.6322
4	SF,NC	0.5356	0.5323	0.5391	0.5368	0.5353
5	MAF,HR	0.6228	0.6252	0.6265	0.6205	0.6217
6	MAF,PGX	0.635	0.6362	0.6398	0.6321	0.636
7	MAF,LOF	0.6247	0.626	0.6271	0.6242	0.6222
8	MAF,NC	0.5951	0.582	0.5978	0.6049	0.591
9	HR,PGX	0.636	0.6396	0.6324	0.6358	0.6155
10	HR,LOF	0.6249	0.626	0.6271	0.6248	0.6224

11	HR,NC	0.5951	0.582	0.5979	0.6049	0.591
12	PGX,LOF	0.6365	0.6365	0.64	0.6354	0.636
13	PGX,NC	0.5972	0.5844	0.6001	0.6071	0.5928
14	LOF,NC	0.5951	0.5821	0.5978	0.605	0.591
15	SF,MAF,HR	0.6262	0.6278	0.6297	0.6244	0.625
16	SF,MAF,PGX	0.6351	0.636	0.6392	0.6326	0.6357
17	SF,MAF,LOF	0.6276	0.6285	0.6302	0.6273	0.6254
18	SF,MAF,NC	0.5942	0.5812	0.597	0.6039	0.5904
19	SF,HR,PGX	0.635	0.6358	0.639	0.6328	0.6354
20	SF,HR,LOF	0.6277	0.6284	0.63	0.6276	0.6253
21	SF,HR,NC	0.5943	0.5813	0.5971	0.6039	0.5904
22	SF,PGX,LOF	0.6362	0.6364	0.6393	0.6352	0.6357
23	SF,PGX,NC	0.5955	0.5826	0.5985	0.605	0.5914
24	SF,LOF,NC	0.5943	0.5813	0.597	0.604	0.5904
25	MAF,HR,PGX	0.6288	0.6318	0.6326	0.6254	0.629
26	MAF,HR,LOF	0.6207	0.6239	0.6226	0.6192	0.6181
27	MAF,HR,NC	0.5945	0.5804	0.5975	0.6047	0.5907
28	MAF,PGX,LOF	0.63	0.6324	0.6328	0.6279	0.6291
29	MAF,PGX,NC	0.5957	0.5813	0.5986	0.6062	0.5916
30	MAF,LOF,NC	0.5945	0.5804	0.5974	0.6048	0.5907
31	HR,PGX,LOF	0.63	0.6322	0.6327	0.6281	0.629
32	HR,PGX,NC	0.5957	0.5813	0.5986	0.6062	0.5916
33	HR,LOF,NC	0.5945	0.5805	0.5974	0.6048	0.5907
34	PGX,LOF,NC	0.5957	0.5814	0.5985	0.6063	0.5917
35	SF,MAF,HR,PGX	0.6292	0.6318	0.6326	0.6263	0.6292
36	SF,MAF,HR,LOF	0.623	0.6257	0.6249	0.6218	0.6205
37	SF,MAF,HR,NC	0.5936	0.5794	0.5967	0.6038	0.59

38	SF,MAF,PGX,LOF	0.6302	0.6324	0.6327	0.6284	0.6292
39	SF,MAF,PGX,NC	0.5942	0.5802	0.5974	0.6044	0.5904
40	SF,MAF,LOF,NC	0.5936	0.5794	0.5966	0.6039	0.59
41	SF,HR,PGX,LOF	0.6302	0.6322	0.6326	0.6286	0.6291
42	SF,HR,PGX,NC	0.5942	0.5802	0.5974	0.6044	0.5904
43	SF,HR,LOF,NC	0.5937	0.5794	0.5967	0.6039	0.59
44	SF,PGX,LOF,NC	0.5943	0.5802	0.5973	0.6045	0.5904
45	MAF,HR,PGX,LOF	0.625	0.6286	0.6273	0.6225	0.6237
46	MAF,HR,PGX,NC	0.5938	0.5792	0.5972	0.6049	0.5888
47	MAF,HR,LOF,NC	0.5933	0.5785	0.5965	0.6045	0.5886
48	MAF,PGX,LOF,NC	0.5939	0.5792	0.5972	0.605	0.5888
49	HR,PGX,LOF,NC	0.5939	0.5793	0.5972	0.605	0.5888
50	SF,MAF,HR,PGX,LOF	0.6255	0.6289	0.6276	0.6233	0.624
51	SF,MAF,HR,PGX,NC	0.5928	0.5783	0.5963	0.6038	0.5878
52	SF,MAF,HR,LOF,NC	0.5925	0.5776	0.5957	0.6036	0.5878
53	SF,MAF,PGX,LOF,NC	0.5929	0.5784	0.5962	0.6039	0.5879
54	SF,HR,PGX,LOF,NC	0.5929	0.5784	0.5962	0.6039	0.5879
55	MAF,HR,PGX,LOF,NC	0.5926	0.5784	0.5959	0.6038	0.587
56	SF,MAF,HR,PGX,LOF,NC	0.5921	0.5777	0.5957	0.6034	0.5866

2 장 RNA 편집 위치 검출을 위한 RNA 서열 비교 및 생물학적 주석처리 도구 개발

소 개

RNA 편집(RNA editing)은 RNA 전사 후에 한개의 뉴클레오타이드(nucleotide) 시퀀스(sequence)가 변형되는 현상을 의미하며 이는 mRNA, 비번역 RNA(ncRNA), 마이크로 RNA(miRNA) 등을 포함한 모든 전사체에서 일어난다[67]. 포유 동물에서 가장 많이 발생하는 RNA 편집유형은 A 에서 I 로 바뀌는 편집이며 이는 주로 ADAR 이라는 효소(enzyme) 에 의해 발생한다[68-70]. 최근 RNA 서열을 NGS 기법을 이용하여 초고속으로 분석하는 RNA-seq 기술이 발달함에 따라 G-A,C-U,T-C,C-A,G-C,T-A,A-등의 새로운 RNA 편집형태도 인간의 세포주에서 발견되었다[71, 72]. 대부분의 RNA 편집 혹은 RNA-DNA 서열변형 위치(RDD site)는 인트론(intron), 5 ‘번역 영역 (5’ UTR), 3 ‘ 번역 영역 (3’ UTR)(7) 그리고 Alu 서열에서 발생한다. 그렇지만 RNA 편집이 번역영역(coding region)에서 발생하게 되면 이로인해 nonsynonymous 단백질 변형, 선택적 스플라이싱(Alternative splicing), 유전자 발현변화 등을 일으켜 원래 단백질의 기능을 상실 시키거나 변형시킬 수 있다. (7-9) 또한 RNA 편집은 마이크로 RNA, 짧은 간섭 RNA(small interfering RNA, siRNA), piRNA(Piwi-interacting RNA)등의 기능에도 영향을 미칠 수 있다. 많은 RNA 편집들은 간질, 뇌허열, 우울증, 뇌종양등 다양한 인간 질병에 영향을 미친다고 보고되었다[67]. 지난 몇년동안 엄청난 양의 DNA, RNA-seq 데이터 들이 생산되고 GEO(www.ncbi.nlm.nih.gov/geo/), ENCODE (http://genome.ucsc.edu/ENCODE/) 등의 공공 저장소에 저장되고 공개되었다. 이러한 데이터들을 기반으로, 많은 RNA 편집과 RDD site 을 검출하기 위한 많은 도구들이 개발되었다. 알려지지 않은 RNA

편집 위치를 찾아내는 대표적은 도구로는 rddChecker (http://genomics.jhu.edu/software/rddChecker/) 가 있으며 이런 도구들은 동일한 시료에서 DNA, RNA-seq을 통해 얻어진 DNA, RNA 서열을 비교하여 후보 RNA 편집위치를 검출하고 이를 단일염기다형성(SNP), 알려진 RDD site 으로 필터하여 알려지지 않은 새로운 RDD 혹은 RNA 편집 위치를 찾아낸다. 하지만 이러한 도구들은 수많은 위양성(false-positive) 결과 들을 포함하고 있다. 많은 새로운 RNA 편집 위치들이 발견되었지만[73], 이 결과들이 많은 위양성 결과를 포함하고 있음이 증명되었다[74, 75]. 이러한 관점으로 보아 RNA 편집위치, RDD site 를 검출해 내는데 있어 가장 중요한 관점은 발견된 위치들중에서 진정한 RNA 편집 위치와 위양성 위치를 구별하는 기술이라고 할 수 있다.

출간된 논문으로 부터 인간 검수를 통해 추출된 RNA 편집 위치를 들을 기반으로 한 DARNED (a database of RNAediting in humans) [76], RADAR (a rigorously annotated database of A-to-I RNA editing) [77] 등의 데이터베이스들이 개발되었다. 그중 DARNED 는 가장 유명한 데이터베이스로써 약 42,000 개의 인간 RNA 편집 싸이트를 가지고 있다. RADAR 은 인간 1,343,464, 쥐 7,272 그리고 초파리 3,155 개의 RNA 편집싸이트를 내포하고 있다. 이러한 데이터베이들을 기반으로 웹기반은 RNA 편집 위치를 주석해주는(annotation) ExpEdit 같은 도구들이 개발되었다[78]. ExpEdit 는 DARNED 를 기반으로 하여 입력된 RNA-seq 데이터에 DARNED 를 매핑하여 정보를 제공하는 웹 기반의 도구로써 신뢰성 있는 데이터를 제공한다. 하지만 알려지지 않은 새로운 RNA 편집 위치를 검출해 내지는 못하는 한계점을 가지고 있다. 하지만 이러한 단점에서 불구하고 원시 RNA-seq 데이터 형태인 FASTQ, SAM (sequence alignment map), BAM (binary alignment map) 파일을 입력값으로 받아 웹에서 사용자가 손쉽게 신뢰성 높은 RNA 편집 위치를 결과로 얻고 또 검색할 수 있다는 점에서 각광받고 있다. 그렇지만 용량이 많은 원시 RNA-seq 데이터를 업로드 하는데 BAM 700MB 파일을 기준으로 28 시간이라는 효율적이지 못한 시간이 걸린다는 치명적인 단점이 있다. 파이썬 패키지로 제작된 REDtools 는 업로딩 시간을 단축하였으며 입력값이 같은 샘플에서 얻어진 DNA,

RNA-seq 데이터일 경우 알려지지 않은 새로운 RNA 편집 싸이트를 찾을 수 있음을 물론 입력값이 RNA-seq 결과 하나인 경우에는 기존에 있던 알려진 RNA 편집싸이트 정보를 매핑하여 제공할 수 있게 보완되어 만들어 졌다[79]. 또한 A-I 편집싸이트 정보를 웹기반의 도구로 제공하는 VIRGO[80]도 개발되었다. 하지만 ExpEdit, REDtools 는 제공한 RNA 편집위치에 대한 신뢰성 정도를 제공하지 않으며 VIRGO 는 A-I 에 한정하여 제공한다는 한계점이 있다.

본 연구는 RNA-seq 서열비교를 통한 RNA 편집위치 주석을 제공하는 새로운 도구(RCARE)로써 각 샘플에서만 특정적으로 나타나는 따른 RNA 편집 위치, 여러가지 문헌과 데이터 베이스를 통합한 지식베이스를 기반으로한 풍부하고 체계적인 지식 주석, 데이터를 기반으로한 신뢰성 등급(evidence level)을 각 RNA 편집위치 마다 제공한다. 또한 사용자가 같은 샘플에서 얻은 DNA, RNA-seq 데이터를 입력값으로 넣으면 알려지지 않은 새로운 RNA 편집 위치를 검출하여 제공한다. 사용자가 RNA-seq 의 variant fall format(VCF)를 입력값으로 사용자 친화적인 웹 페이지에 업로드 하면 한개의 RNA-seq 데이터에 대한 주석값 혹은 여러 RNA-seq 데이터 간의 비교값에 대하여 각 생물학적 요소별 요약 그래프를 제공한다. 또한 사용자가 원시 RNA-seq 데이터인 FASTQ,BAM,SAM 파일을 RNA-seq VCF 를 빠른시간에 자신의 데스크탑에서 변환할 수 있는 파이썬 스크립트를 제공한다. RCARE 는 http://www.snubi.org/ 에서 자유롭게 사용할 수 있습니다.

방 법

데이터 수집

RCARE 는 314,880, 6,830, 그리고 13,018 의 mRNA 편집 위치를 각각 DARNED(NCBI37/hg19)[76], 인간 ENCODE RNA-Seq 데이터[67], Bahn et al, [71] 로 부터 다운로드 받아 통합했다. 앞서 다운받은 1,379,404 와 RADRA [77] ,Li et al [73].에서 다운받은 10,115 개의 인간 RNA 편집 위치를 사용하여 참조(reference) 기반의 각 RNA 편집 위치별 신뢰성 등급을 만들었다. 각 RNA 편집 위치별 풍부한 생물학적 주석 정보를 위해 우리는 Homo_sapiens.GR-Ch37.69.gtf, Repeat-Masker database information from Ensembl (ensembl.org), UCSC (http://genome.ucsc.edu/buildGRCCh37/hg19)를 다운받아 ncRNA, Ensembl Gene (ENSG) ID, Transcript(ENST) ID, Exon ID (ENSE) 그리고 repetitive element (Alu, nonrepetitive)를 각 RNA 편집 위치 혹은 RDD 위치별로 제공 하였다. 또한 ANNOVAR (http://www.openbioinformatics.org/annovar/) [81]를 사용하여 intron, intergenic, splicing region, downstream, upstream, 3' UTR, 5' UTR, 그리고 synonymous/nonsynonymous 주석 정보를 각 RNA 편집 위치 혹은 RDD 위치별로 제공하였다(Table 1).

RNA-seq 데이터 전처리(preprocessing)

RCARE 는 FASTQ 혹은 BAM 파일로 부터 hg 19 VCF 파일로 전환시켜주는 자동 변환 유틸리티를 제공한다(그림 1A). 이 유틸리티는 TopHat [82], SAMtools[83], VCFtools[84], Tabix[85], 그리고 Bowtie2[86] 를 포함하고 있으며 이들을 자동으로 실행시켜 RNA-seq 원시 파일을 VCF 형태로 바꿔주는 파이썬 스크립트를 포함하고 있어 파일 크기가 큰 RNA-seq 원시 데이터를 업로드 하는 시간을 없앴다. TopHat 과 SAMtools 의 설정 파라미터들은 쉽게 변환

가능하도록 되어 있다. 변환 유틸리티는 2 가지 유형으로 구성되어져 있다.

전체 버전 변환 유틸리티는 tophat, 알려지지 않은 새로운 RDD 위치 혹은 RNA 편집 위치를 찾는 파이썬 스크립트와 서로 다른 RNA-seq 데이터를 비교하여 특정 샘플에서만 나타나는 RNA 편집 위치 혹은 RDD 위치를 찾는 파이썬 스크립트, 시험 가동을 위한 샘플 파일등을 비롯한 BAM 혹은 FASTQ를 VCF로 변환하는데 필요한 모든 도구들이 포함있다. 하지만 만약 유저가 tophat, samtools 갖은 도구들이 설치되어 있다면 라이트 버전을 빠르게 다운받아 사용하면 된다. 자세한 데이터 처리 과정은 그림 1 을 참조하면 된다. 만약 유저가 STAR 그리고 GATK 같은 다른 도구를 사용하고 싶다면 해당 도구를 사용해 VCF 를 생성한뒤 웹 버전에 업로드 하면 알려진 RNA 편집 위치 주석, 그리고 해당 위치들에 대한 생물학적 정보를 얻을 수 있으며 2 가지 이상의 데이터를 업로드 하면 서로 비교하여 특정 샘플에서만 나타나는 RNA 편집 위치를 검출해 낼 수 있으며 같은 시료에서 얻어진 DNA, RNA 서열 데이터를 VCF 형태로 입력하면 염색체(chromosome), 위치(position)을 비교하여 알려지지 않은 새로운 RDD 위치 혹은 RNA 편집 위치등을 검출 할 수 있다.

RNA 편집 위치들의 비교

RACE 에서는 다른 샘플에서 얻은 RNA-seq VCF 를 염색체(chromosome), 위치(position)을 기준으로비교하여 특정 샘플에서만 나타나는 기능이 알려진 RNA 편집 위치를 식별해 준다. 자세한 사항은 그림 3 에 기술했다.

웹 인터페이스 구축

RCARE 의 웹 기반의 응용프로그램은 HTML5 (Hypertext Markup Language 5), CSS3(Cascading Style Sheets 3), jQuery, and

Highcharts API(http://www.highcharts.com/)를 사용하여 구축하였다. 이 응용프로그램은 RNA-seq 데이터 처리, 비교, 주석 그리고 시각화로 구성된 4 가지 기능을 제공한다.

결 과

우리는 DARNED(NCBI37/hg19)[76], 인간 ENCODE RNA-Seq 데이터 [67],Bahn et al.[71]로 부터 321,008 개의 RNA 편집 위치를 수집하였다. 이 데이터는 30 가지 샘플 종류로 구성된 154 개의 샘플, 23 개의 논문 그리고 11,299 개의 유전자, 12 가지의 RNA 편집 종류로 포함되어 있다(그림 2). RCARE 의 생물학적 정보 주석은 (1) synonymous vs. nonsynonymous 변화, (2) splicing junction 지역 안의 포함 여부, (3) genomic features, (4) Alu 연관여부,(5) ncRNA 지역안의 포함 여부, (6) 유전자 기호(gene symbol), 다양한 ID 등의 유전자 정보(자세한정보는 첨부파일 참조) (7) 샘플 기원, (8) 참조 논문,(9) 참조 데이터를 기반으로 한 신뢰성 등급등 9 가지의 카테고리로 구성되어 있다. 또한 서로 다른 모든 샘플들을 비교하여 샘플 사이에서 교차 혹은 차이가 나는 RNA 편집 위치 마다 17 가지의 유용한 생물학적 주석을 제공한다(표 2). RNA 편집 위치를 검출하는 것도 중요하지만 검출된 데이터의 신뢰성은 무엇보다 중요하다. 검출된 RNA 편집 위치의 신뢰성을 향상시키기 위해 우리는 신뢰성 등급(Evidence levels)을 생성하였다. 신뢰성 등급은 각 RNA 편집 위치가 관련 논문이나 데이터베이스에서 언급된 횟수를 기반으로 생성하였다. 우리는 신뢰성 등급을 생성하기 위하여 첫번째로 5 가지 논문 혹은 지식베이스(DARNED, RADAR, Bahn et al.[71], Li et al.[73], Park et al.[67])를 수집하였다. 두번째로 모아진 데이터들을 염색체, 위치 정보, reference/alternative 서열정보를 사용하여 통합하고 각 RNA 편집 싸이트가 언급된 횟수를 세어 신뢰성 등급을 생성하였다. 신뢰성 등급은 A 부터 E 까지 총 5 단계로 구성되어져 있으며 A 로 갈수록 높은 등급의 신뢰를 가진다고 말할 수 있다(그림 3, 보충자료 1). 예를 들어 만약 한개의 RNA 편집 위치가 등급이 A 라면 이는 5 개의 데이터베이스 혹은 관련 논문에서 찾아 졌다고 할 수 있다. 반대로 E 등급이라면 한개의 데이터베이스 혹은 논문에서 찾아 졌다고 할 수 있다. 통합된 데이터의 85%의 RNA 편집 위치는 D 등급을 가지고 있었다. 이 신뢰성 등급은 사용자가 등급 스케일에 따라 해당 RNA 편집 위치가 위양성 인지 참인지를 결정하는데 큰 도움을 줄것이다.

웹 인터페이스는 사용자 설명서, 도구 다운로드, 분석 이렇게 3 부분으로 구성되어 있다. 사용자 설명서 영역은 변환 유틸리티, 웹을 사용한 생물학적 주석 달기, 샘플간의 비교 그리고 결과 설명으로 이루어져 있다(그림 1B). 모든 설명서 파일은 PDF 파일로 다운로드 가능하다. 도구 다운로드 영역은 변환 유틸리티를 다운로드 할 수 있따. 우리는 2 가지 버전의 변환 유틸리티를 제공하고 있다. 전체 버전은 TopHat[82], SAMtools[83], VCFtools[84], Tabix[85], Bowtie2[86], 자동 파이썬 스크립트 등 원시 RNA-seq 데이터를 VCF 로 변환하는 모든 도구를 포함하고 있으며 라이트 버전은 파이선 스크립트 만을 포함하고 있다. 분석 영역은 16 가지 유용한 생물학적 주석과 신뢰성 등급을 VCF 에 있는 모든 RNA 편집 위치마다 제공한다. 비교 부분은 서로 다른 샘플로 부터 생성된 RNA-seq VCF 파일의 모든 조합을 비교하여 특정 샘플에서만 나타나는 RNA 편집 위치를 16 가지 유용한 생물학적 주석 정보와 함께 제공한다. 결과 페이지는 genomic 기능, 유전자, ncRNAs, synonymous vs nonsynonymous 변화 , RNA editing 유형 그리고 검출된 RNA 편집위치들의 각 염색체 혹은 샘플별 분포를 요약 그래프를 제공한다.(Figure 1C) 모든 그래프 이미지는 다운로드와 인쇄가 가능하다. 결과를 시험하기 위하여 우리는 MCF-7(a breast cancer cell line), HUVEC (a human umbilicalvein endothelial cell line) 그리고 HeLa-S3 (a cervical carcinoma cell line) RNA-seq 데이터를 ENCODE (http://genome.ucsc.edu/ENCODE/)로 부터 다운로드 받아 2347,646, 1190 개의 RNA 편집 위치를 검출했다. 이 위치들 중 HUVEC 205 개, MCF-7 605 개, HeLa-S3 334 개가 A 부터 C 사이의 신뢰성 등급을 가지고 있었다. 하지만 이것은 전체 데이터중 각각 31.7%, 24.8% 그리고 28.07% 에 해당하였다. 이 결과로 미루어 보아 신뢰성 등급이 RNA 편집 위치를 검출하고 분석하는데 반드시 필요함을 알 수 있었다. 우리는 또한 MCF-7 과 HUVEC 을 비교하여 각 샘플에서만 나타나는 알려진 RNA 편집위치를 검출하였다. 그결과 MCF-7 으로 부터 특이적으로 유방암에서만 나타나는 2,080 개의 RNA 편집 위치를 검출하였다. 이러한 결과는 특정 조건에서 RNA 편집 위치를 검출하는것의 중요성을 보여준다. 우리는 37.3 Mb 의 샘플 파일로 Expedit 와 RCARE 의 실행 시간을 비교를 위해 RACRE 의

변환처리과정과 주석 분석하는 시간을 측정하였다. BAM 파일을 VCF 로 변환하는 처리과정은 3.0 GHz CPU, 2048 MB RAM 의 데스트탑 환경에서 192 초에 수행되었고 RNA 편집 위치당 생물학적 주석을 처리하는 과정은 14.19 MBps 네트워크 환경에서 7 초가 걸렸다.

고 찰

RCARE 는 신뢰할 수 있는 RNA 편집 위치에 대한 생물학적 주석처리, 비교, 그래프 시각화를 효율적이고 사용자 친화적인 웹기반의 시스템이다. RCARE 는 321,008 개의 인간 RNA 편집위치와 이에 해당하는 풍부한 생물학적 주석정보 그리고 유용한 요약 그래프를 신뢰성 등급과 함께 제공하고 있다. 게다가 이 도구는 RNA-seq 원시 데이터를 VCF 로 자동을 변환해 주는 변환 유틸리티를 파이썬 기반의 스크립트와 함께 제공하고 있다. 이 유틸리티로부터 나온 VCF 를 웹에 업로드 하면 생물학적 분석 그리고 서로 다른 샘플간의 비교 하여 생물학적 주석정보와 함께 제공 받을 수 있다. RCARE 웹 인터페이스는 쉽게 주석처리와 시각화를 할 수 있으며 그 그결과를 CSV 형태와 JPG 형태로 다운로드 받을 수 있다. 최근 2~3 년 사이에 새로운 RNA 편집 위치를 찾아내는 것이 이 분야 연구의 초점이 되어왔다[67, 71, 73]. 그러나 최근 연구동향은 확인된 RNA 편집위치의 신뢰성 판정 여부로 옮겨 갔다. 왜냐하면 발견된 RNA 편집 위치들에서 많은 위양성이 발견되었기 때문이다. 우리는 RCARE 의 신뢰성 등급이 RNA 편집 위치 연구 분야에서 그 수요가 크게 증가할 것이라고 예상한다. RCARE 는 신뢰할 수 잇는 RNA 편집 위치를 동정하는데 크게 도움이 될 것이라고 예상된다.

결과그림

그림 1 RCARE 의 RNA-seq 데이터 처리 과정 단계 및 생물학적 주석처리 결과 시각화 과정 A) RNA-seq 데이터 처리 변환 유틸리티 처리 과정(pipeline). B) 생물학적 주석 처리와 비교 분석을 위한 웹 인터페이스, C)주석 결과의 요약 그래프 예제

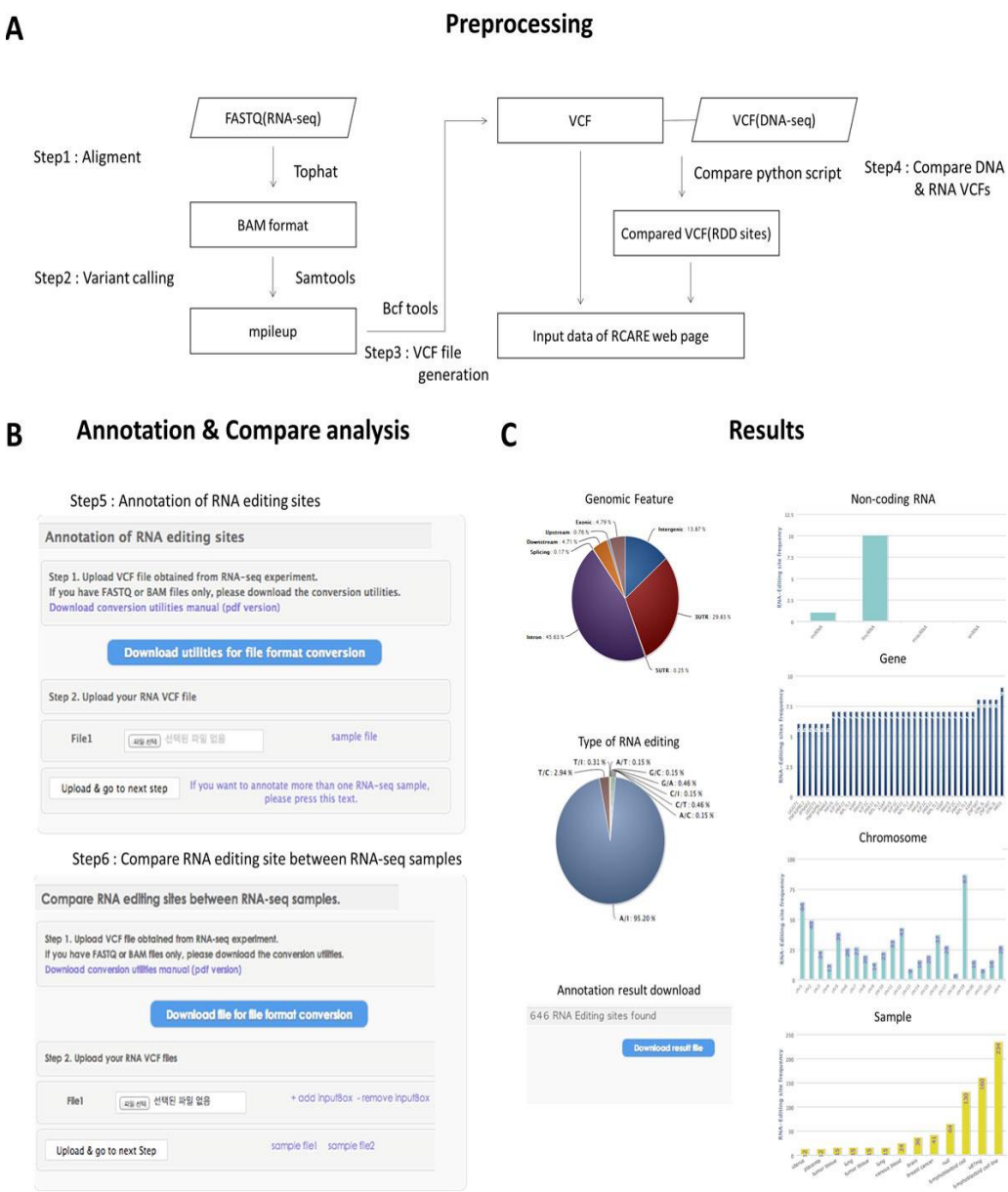
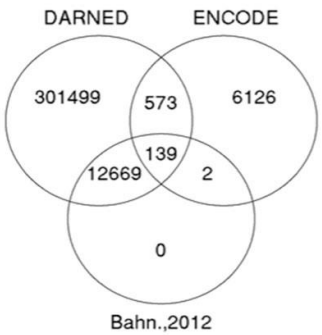
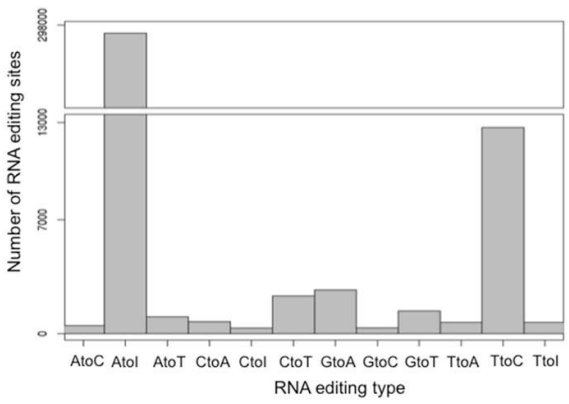


그림 2 데이터 구성. A) DARNED, ENCODE, Bahn et[71] 데이터에 포함되어 있는 RNA 편집 위치 개수를 나타내는 벤다이어그램. B) 각 샘플당 RNA 편집 위치 개수. C) 각 RNA 편집 타입별 RNA 편집 위치 개수. D)각 염색체당 포함되어 있는 RNA 편집위치 개수. E) 각 게놈 기능(genomic feature)당 포함되어 있는 RNA 편집위치 개수.

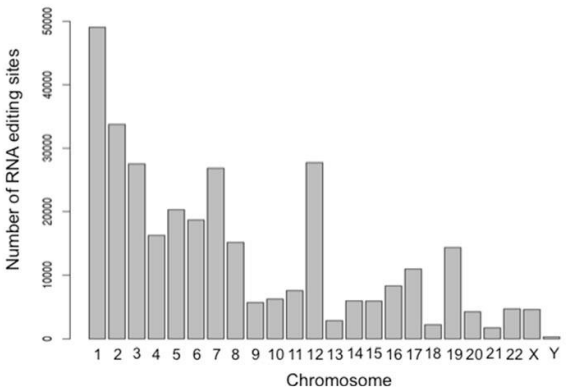
A



C



D



B

Sample type	No. RNA editing sites
Adipose	85
Adrenal	467
Blood	251,796
Blood Vessel	132
Bone	223
Brain	16,287
Breast	528
Cancer	16,811
Disease	436
Eye	736
Heart	243
Kidney	1,618
Large Intestine	375
Liver	296
Lung	1,836
Muscle	322
Pancreas	65
Prostate	1,276
Salivary Gland	196
Skin	287
Small Intestine	665
Spleen	1,586
Stem Cell	586
Stomach	227
Testis	1,364
Throat	591
Thymus	2,188
Thyroid	174
Uterus	3,295
Not Specified	25,299
Total	329,992

E

Genomic Feature	No. RNA editing sites
3'UTR	17,483
5'UTR	345
Intergenic region	38,319
Exon	3,634
Intron	255,891
Splice site	45
Upstream	1,035
Downstream	4,309
Total	321,061

결과 표

표 1 현재 RCARE 에 RNA 편집 위치들의 데이터의 출처 목록.

각 데이터베이스 별 항목은 고유한 전체 RNA 편집 싸이트 수를 의미한다.

	샘플 수	RNA 편집 위치 수	Alu 서열 수	참고문헌 수
DARNED	29	314,880	15,783	34
ENCODE	27	6,840	347	18
RADAR	30	291,901	14,318	31
Bahn et al. [71]	27	12,810	2,916	15
Li et al. [73]	1	1	0	1
Total	114	626,432	33,364	99

표 2 RCARE의 주석 결과 형식.

순번	항목	설명	참고문헌
1	Chr	Chromosome of the RNA editing site in the reference genome.	[71, 73, 76]
2	Pos	Coordinate of the RNA editing site in the reference genome.	
3	In DNA	Base of the RNA editing site in the DNA reference sequence.	
4	In RNA	Base of the RNA editing site in the RNA sequence of sample.	
5	Gene	Gene name to which the RNA editing site belongs.	
6	*Evidence level*	The *evidence level* consists of five levels (A–E), where A is highest level (e.g., if an RNA editing site had level "A," it appeared in all five of the resource databases/papers used). *Level A: The RNA editing site appeared in five resources (evidence No. 5). *Level B: The RNA editing site appeared in four resources (evidence No. 4). *Level C: The RNA editing site appeared in three resources (evidence No. 3). *Level D: The RNA editing site appeared in two resources (evidence No. 2). *Level E: The RNA editing site appeared in one resource (evidence No. 1).	[70,72,75,76] RepeatMasker
7	Strand	+ for positive strand; − for negative strand.	[70,72,75]
8	Source	This field contains information regarding the tissue source from which the RNA editing instance was obtained.	
9	PubMed ID	This field provides the reference article from which the RNA editing data was extracted.	

10	Alu	This field provides information of Alu at the RNA editing site.	RepeatMasker
11	Data reference	Reference database.	Each database or reference
12	ENSG	Ensembl Gene ID.	GTF(*Homo sapiens*, GRCH37.17) in Ensembl
13	ENST	Ensembl Transcript ID.	
14	ENSE	Ensembl Exon ID.	
15	Genomic feature	Genomic feature of the RNA editing site. *Exonic: the variant overlaps a coding exon. *Splicing: the variant is within 2 bp of a splicing junction. *ncRNA: the variant overlaps a transcript without coding annotation in the gene definition. *5’ UTR: the variant overlaps a 5' untranslated region. *3’ UTR: the variant overlaps a 3' untranslated region. *Intronic: the variant overlaps an intron. *Upstream: the variant overlaps the 1−kb region upstream of the transcription start site. *Downstream: the variant overlaps the 1−kb region downstream of the transcription end site. *Intergenic: a variant is in the intergenic region.	[81]
16	Synonymous or nonsynonymous	Synonymous or nonsynonymous substitutions at the RNA editing site.	[80]

17	Noncoding RNA	This field indicates whether the location of an RNA editing site is in ncRNA.	GTF (*Homo sapiens*, GRCH37.17) in Ensembl

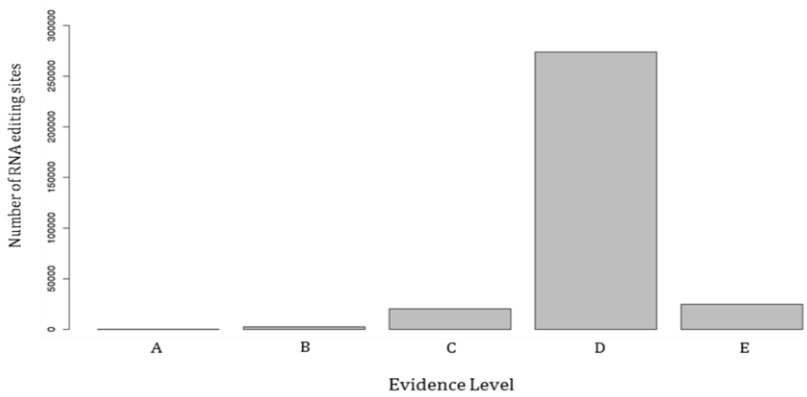
표 3 RCARE 의 신뢰성 등급 분류 체계.

분류	신뢰성 등급	등급의 구성 자료	수
A	5	DARNED, ENCODE, Bahn et al. [5], RADAR, Alu	33
B	4	DARNED, RADAR, Bahn et al. [5], Alu	2,204
		DARNED, ENCODE, RADAR, Alu	23
		DARNED, ENCODE, Bahn et al. [5], RADAR	106
C	3	DARNED, Bahn et al. [5], RADAR	7,037
		DARNED, Bahn et al. [5], Alu	679
		DARNED, ENCODE, RADAR	550
		DARNED, RADAR, Alu	12,038
		DARNED, RADAR, Li et al. [11]	1
		ENCODE, Bahn et al. [5], RADAR	1
		ENCODE, RADAR, Alu	20
D	2	DARNED, Alu	806
		DARNED, RADAR	269,547
		DARNED, Bahn et al. [5]	2,749
		ENCODE, Alu	271
		ENCODE, RADAR	341
		ENCODE, Bahn et al. [5]	1
E	1	DARNED	19,107
		ENCODE	5,494
	Total		321,008

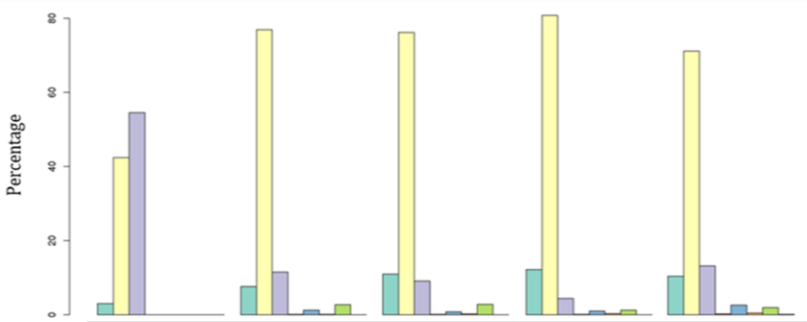
보충자료 1 각 신뢰성 등급, genomic feature 당 속하는 RNA 편집 위치 수.

A) 각 신뢰성 등급당 RNA 편집 위치 수 B) 각 genomic feature 에 속하는 신뢰성 등급당 RNA 편집 위치 수

A



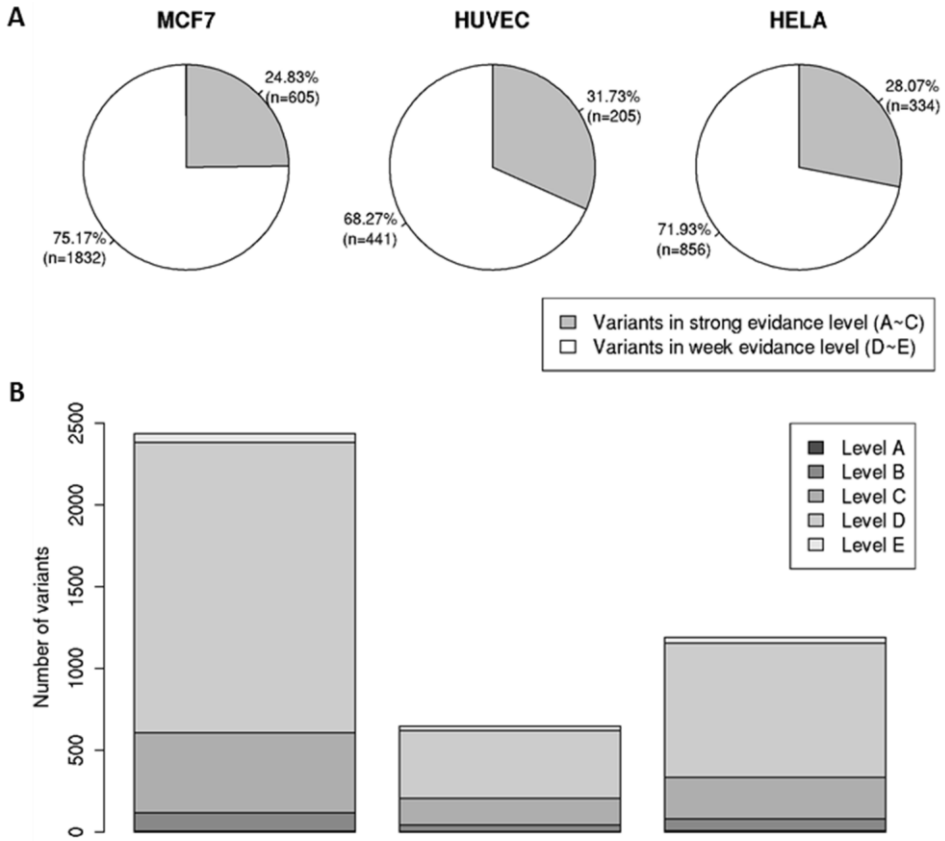
B



Evidence level	A	B	C	D	E
Intergenic region	1	177	2,237	33,344	2,560
Intronic region	14	1,794	15,486	221,100	17,497
3'UTR	18	269	1,842	12,106	3,248
5'UTR	0	2	11	280	52
Exonic region	0	27	154	2,818	635
Upstream	0	2	36	882	115
Downstream	0	62	562	3,206	479
Splice site	0	0	3	15	27
Total	33	2,333	20,331	273,751	24,613

보충자료 2 3 개의 세포주에 속하는 신뢰성 등급의 비율.

A) MCF-7 (a breast cancer cell line), HUVEC (a human umbilical vein endothelial cell line) and HeLa-S3 (a cervical carcinoma cell line) 세포주에서 검출된 RNA 편집 위치의 신뢰성 등급 A-C 와 D-E 에 속하는 편집위치 수의 비율. B)3 개의 세포주당 RNA 편집위치 수

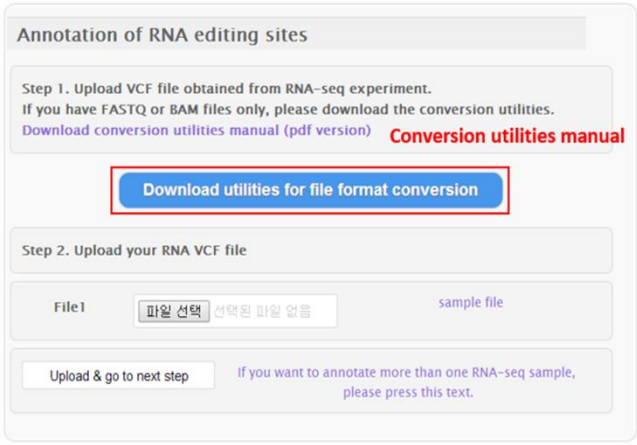


Cell line	MCF7	HUVEC	HELA
Level A	4 (0.16%)	2 (0.31%)	7 (0.59%)
Level B	112 (4.60%)	39 (6.04%)	71 (5.96%)
Level C	489 (20.07%)	164 (25.39%)	256 (21.51%)
Level D	1,779 (73.00%)	415 (64.24%)	822 (69.08%)
Level E	53 (2.17%)	26 (4.02%)	34 (2.86%)
Total	2,437 (100.00%)	646 (100.00%)	1,190 (100.00%)

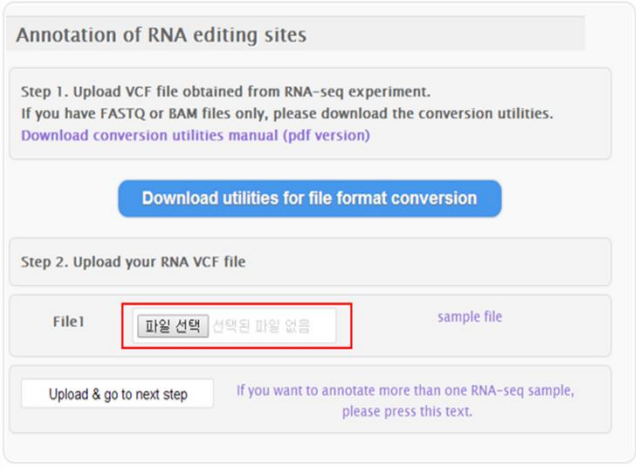
보충자료 3 RCARE 웹 인터페이스 사용자 설명서

1. Annotation section

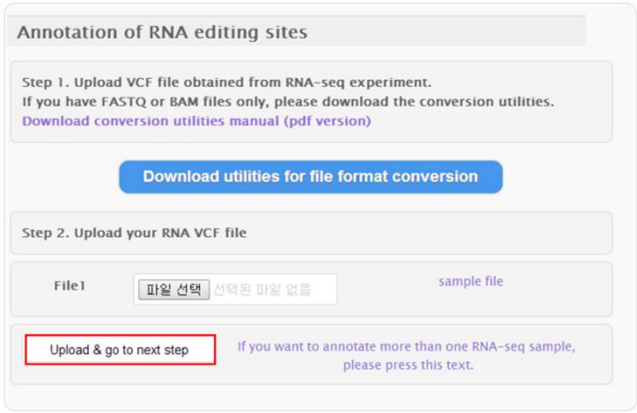
A. If you have FASTQ or BAM files only, please download the conversion utilities.



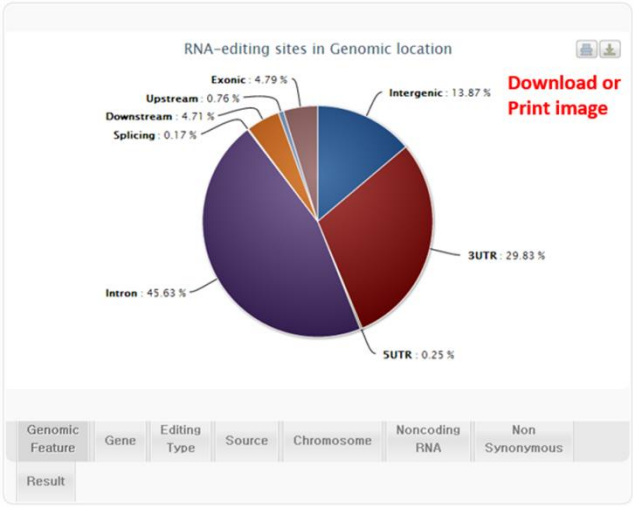
B. Upload a VCF file for RNA editing site annotations.



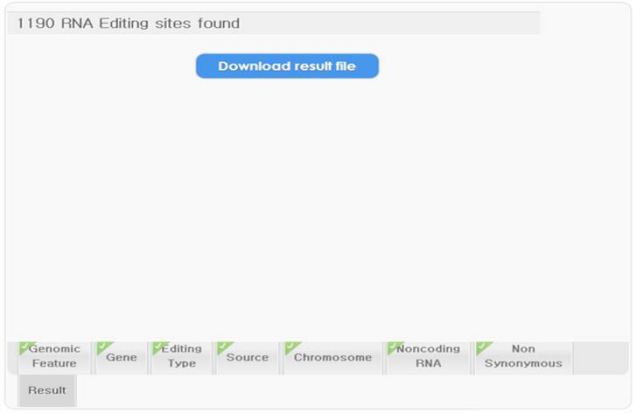
C. Press ‘Upload & to go next step’ button for RNA editing site annotations.



D. Result graphs show annotated results.

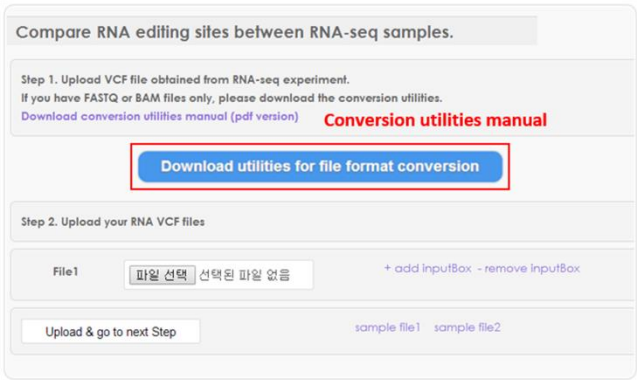


E. Press "*Download result file*" button to download result file.

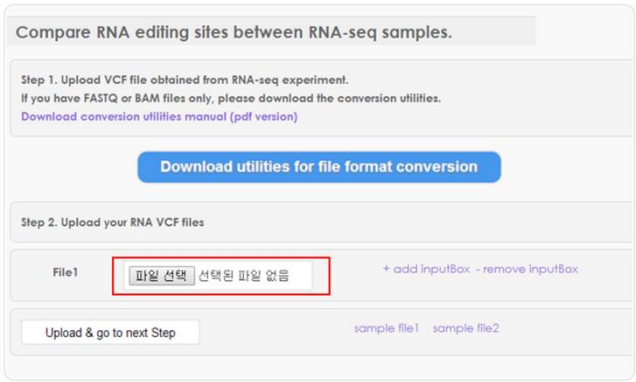


2. Compare section

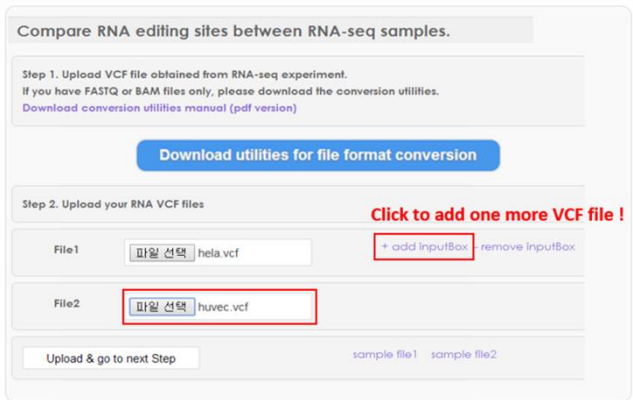
A. If you have FASTQ or BAM files only, please download the conversion utilities.



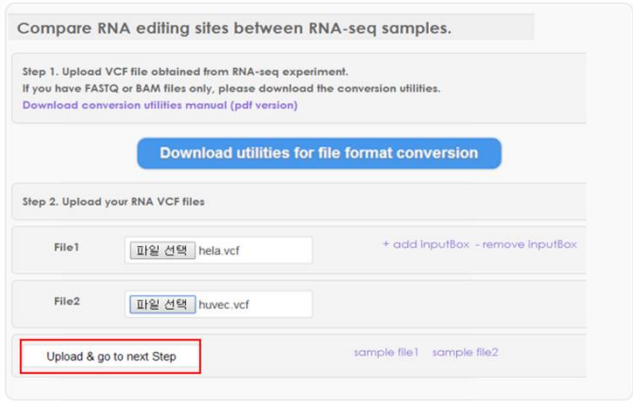
B. Upload a VCF file for annotation of RNA editing sites.



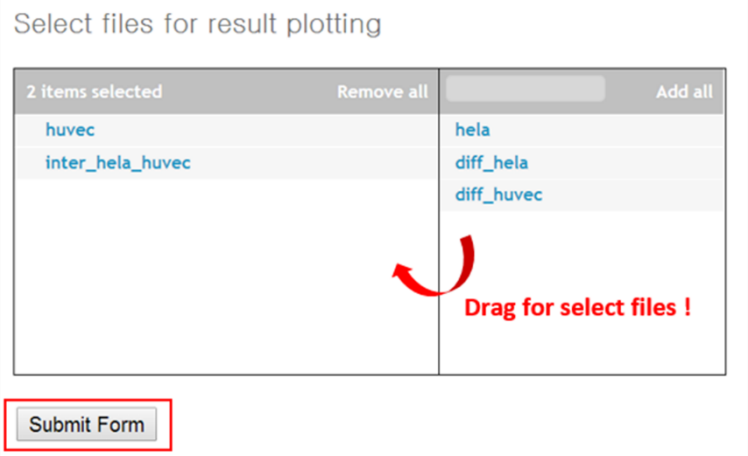
C. Add one more VCF file for comparison.



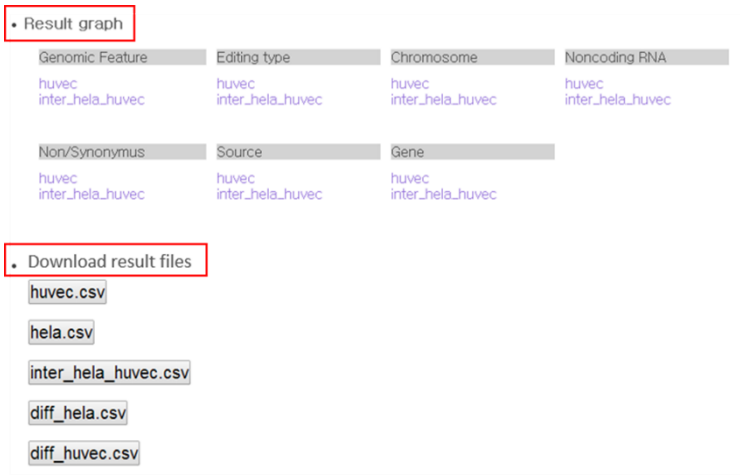
D. Press ‘Upload & to go next step’ button to annotate RNA editing sites.



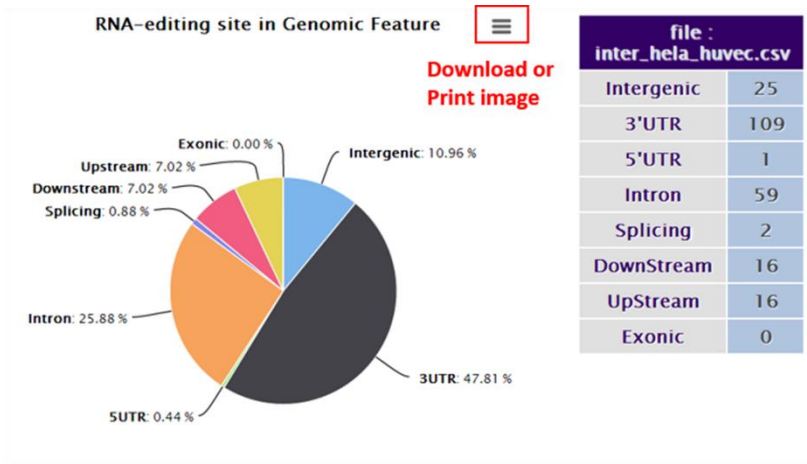
Select result files by dragging the files from the right column to the left, and press “*Submit Form*” button.



E. Select result graph to view and download result files.



F. Result graphs show annotated results.



보충자료 4 RCARE 변환 유틸리티 사용자 설명서

1. Description

RCARE convert utilities is a set of Python-based utilities for converting FASTQ and BAM (binary format for storing sequence data) into VCF (variant call format) and comparing RNA and DNA VCF files from the same sample. The package provides customized TopHat and SAMtools commands that the user can execute. RCARE convert utilities provides an autoinstallation function for the tools. This is very easy for researchers to use, even for those with no experience of RNA-Seq data analysis.

RCARE convert utilities contains TopHat (http://tophat.cbcb.umd.edu/), SAMtools (http://samtools.sourceforge.net/), Tabix (http://samtools.sourceforge.net/tabix.shtml), VCFtools (http://vcftools.sourceforge.net/), and Bowtie2 (http://bowtie-bio.sourceforge.net/bowtie2). If user presetup tools including TopHat, download light RCARE convert utilities and installation.

2. Input data format

RCARE convert utilities convert three sequence formats (FASTQ, BAM, and VCF) to VCF, which is the input format on the RCARE website.

- FASTQ format
  - ➔ FASTQ format is a text-based format for storing both a biological sequence (usually a nucleotide sequence) and its corresponding quality scores.
- BAM format
  - ➔ BAM format is a binary format for storing sequence data. (http://samtools.sourceforge.net/SAMv1.pdf).
- VCF format (variant call format)
  - ➔ VCF is a text file format (most likely stored in a compressed manner). It contains meta information lines, a header line, and

data lines, each containing information about a position in the genome.

3. Installing and testing the installation

3-1 Install quick start

- RCARE needs a presetup in the Python environment.
- Download RCARE convert utilities (4.75G) from the website.
- Unzip RCARE convert utilities.
  - ➔ Tar -xvf RCARE-pre-processing.tar.gz
- Run rcare.py for your purposes.

3-2 Test the installation

The sample BAM data contained only 21 chromosome. These data were extracted from paired-end RNA-Seq using HeLa cells in ENCODE (http://genome.ucsc.edu/ENCODE).

- Input data confirmation
  - ➔ ls ./input_data/bam/
- Test command
  - ➔ python rcare.py -ib sample.bam -fn sample_bam_test
- Result confirmation
  - ➔ ls ./result_data/vcf/sample_bam_test/

4. Synopsis and example

4.1 Input data folder consists of FASTQ, BAM, and VCF. Insert into row data in each folder.

4.2 Convert paired-end FASTQ files into VCF format

- ➔ Python rcare.py -if -p S1.fastq S2.fastq -fn fastq_test
- Result confirmation
  - ➔ ls ./input_data/vcf/fastq_test/

4.3 Convert single FASTQ file into VCF format

- ➔ python rcare.py -if -s S1.fastq -fn single_fastq_test
- Result confirmation
  - ➔ ls ./input_data/vcf/fastq_test

4.4 Convert BAM into convert to VCF format

➔ python rcare.py −ib sample.bam −fn sample_bam_test

− Result confirmation

➔ ./result_data/vcf/sample_bam_test

− Compare RNA VCF with DNA VCF file

➔ python rcare.py −c DNA.vcf RNA.vcf −fn 1_compare_test

− Result confirmation

➔ ls ./result_data/compare/1_compare_test/

4.5 Customized TopHat command running

➔ python rcare.py −tc "tophat" −fn test

4.6 Customized SAMtools command running

➔ python rcare.py −sc "samtools" −fn test\

5. RCARE convert utilities options

Option	Description
−if	Input file format: FASTQ file
−ib	Input file format: BAM file
−p	Paired end FASTQ file
−c	Compare VCF(RNA) with VCF(DNA)
−fn	Result file name
−tc	Customized TopHat commands
−sc	Customized SAMtools commands

Folder file name	Description
input_data	Insert input data
result_data	Save result data

resource	Files required for preprocessing
tools	Tools required for preprocessing
rcare.py	Batch file of convert utilities

6. Package composition

7. Light RCARE convert utilities installation

- Download light-RCARE-pre-processing.tar.gz (35.62 MB) from RCARE website.
- Download tools:
  1. TopHat: http://tophat.cbcb.umd.edu/
  2. SAMtools: http://samtools.sourceforge.net/
  3. Bowtie2: http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
  4. Tabix: http://samtools.sourceforge.net/tabix.shtml
  5. VCFtools: http://vcftools.sourceforge.net/
- All tools insert into the tools folder in the RCARE convert utilities
- If user has used previous setup tools, initialize each tool's environment settings

8. Authors

Ju Han Kim and Soo Youn Lee from SNUBI (Seoul National University Biomedical Informatics; http://www.snubi.org/)

9. References

1. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009; 25: 2078-9.

2. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 2009; 25: 1105-11.
3. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R; 1000 Genomes Project Analysis Group. The variant call format and VCFtools. Bioinformatics 2011; 27: 2156-8.
4. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature methods 2012; 9: 357-9.
5. Li H. Tabix: fast retrieval of sequence features from generic TAB-delimited files. Bioinformatics 2011; 27: 718-9.

참고문헌

1. Jain, K.K., *Personalized medicine.* Curr Opin Mol Ther, 2002. **4**(6): p. 548-58.
2. Wei, C.Y., M.T. Lee, and Y.T. Chen, *Pharmacogenomics of adverse drug reactions: implementing personalized medicine.* Hum Mol Genet, 2012. **21**(R1): p. R58-65.
3. Lazarou, J., B.H. Pomeranz, and P.N. Corey, *Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies.* JAMA, 1998. **279**(15): p. 1200-5.
4. Severino, G. and M. Del Zompo, *Adverse drug reactions: role of pharmacogenomics.* Pharmacol Res, 2004. **49**(4): p. 363-73.
5. Pirmohamed, M., et al., *Adverse drug reactions as cause of admission to hospital: prospective analysis of 18 820 patients.* BMJ, 2004. **329**(7456): p. 15-9.
6. Rottenkolber, D., et al., *Adverse drug reactions in Germany: direct costs of internal medicine hospitalizations.* Pharmacoepidemiol Drug Saf, 2011. **20**(6): p. 626-34.
7. Xie, H.F., F., *Pharmacogenomics steps toward personalized medicine.* Personalized Medicine, 2005. **2**(4): p. 325-337.
8. Need, A.C., A.G. Motulsky, and D.B. Goldstein, *Priorities and standards in pharmacogenetic research.* Nat Genet, 2005. **37**(7): p. 671-81.
9. Crews, K.R., et al., *Pharmacogenomics and individualized medicine: translating science into practice.* Clin Pharmacol Ther, 2012. **92**(4): p. 467-75.
10. Zhou, K. and E.R. Pearson, *Insights from genome-wide association studies of drug response.* Annu Rev Pharmacol Toxicol, 2013. **53**: p. 299-310.
11. Stocco, G., K.R. Crews, and W.E. Evans, *Genetic polymorphism of inosine-triphosphate-pyrophosphatase influences mercaptopurine metabolism and toxicity during treatment of acute lymphoblastic leukemia individualized for thiopurine-S-methyl-transferase status.* Expert Opin Drug Saf, 2010. **9**(1): p. 23-37.
12. Zaza, G., et al., *Pharmacogenomics: a new paradigm to personalize treatments in nephrology patients.* Clin Exp Immunol, 2010. **159**(3): p. 268-80.
13. Hudis, C.A., *Trastuzumab--mechanism of action and use in clinical practice.* N Engl J Med, 2007. **357**(1): p. 39-51.
14. Petak, I., et al., *Integrating molecular diagnostics into anticancer drug discovery.* Nat Rev Drug Discov, 2010. **9**(7): p. 523-35.
15. Eichler, H.G., et al., *Bridging the efficacy-effectiveness gap: a regulator's perspective on addressing variability of drug response.* Nat Rev Drug Discov, 2011. **10**(7): p. 495-506.

16. Wang, L., H.L. McLeod, and R.M. Weinshilboum, *Genomics and drug response.* N Engl J Med, 2011. **364**(12): p. 1144-53.
17. FDA. *Table of Pharmacogenomic Biomarkers in Drug Labeling*. May 20, 2015; Available from: http://www.fda.gov/Drugs/ScienceResearch/ResearchAreas/Pharmacogenetics/ucm083378.htm
18. Wikipedia. *Pharmacogenetics*. August 26, 2015; Available from: https://en.wikipedia.org/wiki/Pharmacogenetics.
19. Wetterstrand, K. *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)*. June 15, 2015; Available from: http://www.genome.gov/sequencingcosts/.
20. Genomes Project, C., et al., *An integrated map of genetic variation from 1,092 human genomes.* Nature, 2012. **491**(7422): p. 56-65.
21. Gamazon, E.R., et al., *A pharmacogene database enhanced by the 1000 Genomes Project.* Pharmacogenet Genomics, 2009. **19**(10): p. 829-32.
22. Altshuler, D., M.J. Daly, and E.S. Lander, *Genetic mapping in human disease.* Science, 2008. **322**(5903): p. 881-8.
23. Hindorff LA, M.J., Morales J, Junkins HA, Hall PN, Klemm AK, and Manolio TA. *A Catalog of Published Genome-Wide Association Studies*. March 23, 2015; Available from: http://www.genome.gov/gwastudies/.
24. Risch, N. and K. Merikangas, *The future of genetic studies of complex human diseases.* Science, 1996. **273**(5281): p. 1516-7.
25. Motsinger-Reif, A.A., et al., *Genome-wide association studies in pharmacogenomics: successes and lessons.* Pharmacogenet Genomics, 2013. **23**(8): p. 383-94.
26. Takeuchi, F., et al., *A genome-wide association study confirms VKORC1, CYP2C9, and CYP4F2 as principal genetic determinants of warfarin dose.* PLoS Genet, 2009. **5**(3): p. e1000433.
27. Loebstein, R., et al., *Interindividual variability in sensitivity to warfarin--Nature or nurture?* Clin Pharmacol Ther, 2001. **70**(2): p. 159-64.
28. Au, N. and A.E. Rettie, *Pharmacogenomics of 4-hydroxycoumarin anticoagulants.* Drug Metab Rev, 2008. **40**(2): p. 355-75.
29. Sibbing, D., et al., *Cytochrome 2C19*17 allelic variant, platelet aggregation, bleeding events, and stent thrombosis in clopidogrel-treated patients with coronary stent placement.* Circulation, 2010. **121**(4): p. 512-8.
30. Schroth, W., et al., *Breast cancer treatment outcome with adjuvant tamoxifen*

relative to patient CYP2D6 and CYP2C19 genotypes. J Clin Oncol, 2007. **25**(33): p. 5187-93.

31. Alomar, M.J., *Factors affecting the development of adverse drug reactions (Review article).* Saudi Pharm J, 2014. **22**(2): p. 83-94.
32. Ong, F.S., et al., *Clinical utility of pharmacogenetic biomarkers in cardiovascular therapeutics: a challenge for clinical implementation.* Pharmacogenomics, 2012. **13**(4): p. 465-75.
33. Lord, J., A.J. Lu, and C. Cruchaga, *Identification of rare variants in Alzheimer's disease.* Front Genet, 2014. **5**: p. 369.
34. Cirulli, E.T. and D.B. Goldstein, *Uncovering the roles of rare variants in common disease through whole-genome sequencing.* Nat Rev Genet, 2010. **11**(6): p. 415-25.
35. Ramsey, L.B., et al., *Rare versus common variants in pharmacogenetics: SLCO1B1 variation and methotrexate disposition.* Genome Res, 2012. **22**(1): p. 1-8.
36. Ng, P.C. and S. Henikoff, *SIFT: Predicting amino acid changes that affect protein function.* Nucleic Acids Res, 2003. **31**(13): p. 3812-4.
37. Law, V., et al., *DrugBank 4.0: shedding new light on drug metabolism.* Nucleic Acids Res, 2014. **42**(Database issue): p. D1091-7.
38. Kanehisa, M., et al., *From genomics to chemical genomics: new developments in KEGG.* Nucleic Acids Res, 2006. **34**(Database issue): p. D354-7.
39. Whirl-Carrillo M, M.E., Hebert JM, Gong L, Sangkuhl K, Thorn CF, Altman RB, Klein TE. *Pharmacogenomics Knowledge Base*. Available from: https://www.pharmgkb.org.
40. 1000Genomes. *1000 Genomes Project*. Available from: http://www.1000genomes.org/.
41. 1000Genomes. *Which populations are part of your study?* ; Available from: http://www.1000genomes.org/category/frequently-asked-questions/population.
42. WHOCC. *International language for drug utilization research*. May 5, 2015.
43. CDC, *National Center for Health Statistics.*
44. Ng, P.C. *SIFT Help*. August 2001; Available from: http://sift.jcvi.org/www/SIFT_help.html.
45. Habegger, L. *Variant Annotation Tool*. May 27, 2014; Available from: http://vat.gersteinlab.org/.
46. Habegger, L., et al., *VAT: a computational framework to functionally annotate variants in personal genomes within a cloud-computing environment.* Bioinformatics, 2012. **28**(17): p. 2267-9.
47. Boyle, A. *RegulomeDB*. 2012; Available from: http://www.regulomedb.org/.

48. Boyle, A.P., et al., *Annotation of functional variation in personal genomes using RegulomeDB.* Genome Res, 2012. **22**(9): p. 1790-7.
49. Baik SY, L.S., Park CH, Yoon JH, Kim JH, *Deleterious Coding Variant Analysis for Personalized Prevention of Adverse Drug Reactions.* 2013.
50. Su Youn Baik, S.Y.L., Chan Hee Park, Jun Hee Yoon, Ju Han Kim, *Deleterious Coding Variant Analysis for Personalized Prevention of Adverse Drug Reactions.* 2015.
51. Wikipedia. *Central Tendency*. July 21, 2015; Available from: https://en.wikipedia.org/wiki/Central_tendency.
52. Wikipedia. *Winsorising*. July 23, 2015; Available from: https://en.wikipedia.org/wiki/Winsorising.
53. Lin, Y.C., et al., *Identifying rare and common disease associated variants in genomic data using Parkinson's disease as a model.* J Biomed Sci, 2014. **21**: p. 88.
54. Wikipedia. *Mutation*. September 18, 2015; Available from: https://en.wikipedia.org/wiki/Mutation.
55. Lundstrom, K. and M.P. Turpin, *Proposed schizophrenia-related gene polymorphism: expression of the Ser9Gly mutant human dopamine D3 receptor with the Semliki Forest virus system.* Biochem Biophys Res Commun, 1996. **225**(3): p. 1068-72.
56. Li-Wan-Po, A., et al., *Pharmacogenetics of CYP2C19: functional and clinical implications of a new variant CYP2C19*17.* Br J Clin Pharmacol, 2010. **69**(3): p. 222-30.
57. Wikipedia. *Nonsense-mediated Decay*. September 20, 2015; Available from: https://en.wikipedia.org/wiki/Nonsense-mediated_decay.
58. Chang, Y.F., J.S. Imam, and M.F. Wilkinson, *The nonsense-mediated decay RNA surveillance pathway.* Annu Rev Biochem, 2007. **76**: p. 51-74.
59. Ni, J.Z., et al., *Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay.* Genes Dev, 2007. **21**(6): p. 708-18.
60. Frischmeyer, P.A. and H.C. Dietz, *Nonsense-mediated mRNA decay in health and disease.* Hum Mol Genet, 1999. **8**(10): p. 1893-900.
61. Ciszkowski, C., et al., *Codeine, ultrarapid-metabolism genotype, and postoperative death.* N Engl J Med, 2009. **361**(8): p. 827-8.
62. Gasche, Y., et al., *Codeine intoxication associated with ultrarapid CYP2D6 metabolism.* N Engl J Med, 2004. **351**(27): p. 2827-31.
63. Rieder, M.J., et al., *Effect of VKORC1 haplotypes on transcriptional regulation and*

warfarin dose. N Engl J Med, 2005. **352**(22): p. 2285-93.

64. Aithal, G.P., et al., *Association of polymorphisms in the cytochrome P450 CYP2C9 with warfarin dose requirement and risk of bleeding complications.* Lancet, 1999. **353**(9154): p. 717-9.
65. FDA. *Full Prescribing Information*. 2012; Available from: http://www.accessdata.fda.gov/drugsatfda_docs/label/2012/021513s010lbl.pdf.
66. McDonough, C.W., et al., *Atenolol induced HDL-C change in the pharmacogenomic evaluation of antihypertensive responses (PEAR) study.* PLoS One, 2013. **8**(10): p. e76984.
67. Park, E., et al., *RNA editing in the human ENCODE RNA-seq data.* Genome Res, 2012. **22**(9): p. 1626-33.
68. Kim, U., et al., *Molecular cloning of cDNA for double-stranded RNA adenosine deaminase, a candidate enzyme for nuclear RNA editing.* Proc Natl Acad Sci U S A, 1994. **91**(24): p. 11457-61.
69. Kumar, M. and G.G. Carmichael, *Nuclear antisense RNA induces extensive adenosine modifications and nuclear retention of target transcripts.* Proc Natl Acad Sci U S A, 1997. **94**(8): p. 3542-7.
70. Wagner, R.W., et al., *A double-stranded RNA unwinding activity introduces structural alterations by means of adenosine to inosine conversions in mammalian cells and Xenopus eggs.* Proc Natl Acad Sci U S A, 1989. **86**(8): p. 2647-51.
71. Bahn, J.H., et al., *Accurate identification of A-to-I RNA editing in human by transcriptome sequencing.* Genome Res, 2012. **22**(1): p. 142-50.
72. Genomes Project, C., et al., *A map of human genome variation from population-scale sequencing.* Nature, 2010. **467**(7319): p. 1061-73.
73. Li, M., et al., *Widespread RNA and DNA sequence differences in the human transcriptome.* Science, 2011. **333**(6038): p. 53-8.
74. Pickrell, J.K., Y. Gilad, and J.K. Pritchard, *Comment on "Widespread RNA and DNA sequence differences in the human transcriptome".* Science, 2012. **335**(6074): p. 1302; author reply 1302.
75. Lin, W., et al., *Comment on "Widespread RNA and DNA sequence differences in the human transcriptome".* Science, 2012. **335**(6074): p. 1302; author reply 1302.
76. Kiran, A. and P.V. Baranov, *DARNED: a DAtabase of RNa EDiting in humans.* Bioinformatics, 2010. **26**(14): p. 1772-6.
77. Ramaswami, G. and J.B. Li, *RADAR: a rigorously annotated database of A-to-I RNA editing.* Nucleic Acids Res, 2014. **42**(Database issue): p. D109-13.
78. Picardi, E., et al., *ExpEdit: a webserver to explore human RNA editing in RNA-Seq*

experiments. Bioinformatics, 2011. **27**(9): p. 1311-2.

79. Picardi, E. and G. Pesole, *REDItools: high-throughput RNA editing detection made easy.* Bioinformatics, 2013. **29**(14): p. 1813-4.
80. Distefano, R., et al., *VIRGO: visualization of A-to-I RNA editing sites in genomic sequences.* BMC Bioinformatics, 2013. **14 Suppl 7**: p. S5.
81. Wang, K., M. Li, and H. Hakonarson, *ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data.* Nucleic Acids Res, 2010. **38**(16): p. e164.
82. Trapnell, C., L. Pachter, and S.L. Salzberg, *TopHat: discovering splice junctions with RNA-Seq.* Bioinformatics, 2009. **25**(9): p. 1105-11.
83. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.
84. Danecek, P., et al., *The variant call format and VCFtools.* Bioinformatics, 2011. **27**(15): p. 2156-8.
85. Li, H., *Tabix: fast retrieval of sequence features from generic TAB-delimited files.* Bioinformatics, 2011. **27**(5): p. 718-9.
86. Langdon, W.B., *Performance of genetic programming optimised Bowtie2 on genome comparison and analytic testing (GCAT) benchmarks.* BioData Min, 2015. **8**(1): p. 1.

논 문 초 록

1. 영문요약(영문초록)

Abstract

Personal pharmacogenomics : pharmacological approach based on biological functional element and genomic variant

LEE SOO-YOUN

자연과학대학 생물정보협동과정

The Graduate School

Seoul National University

The main objective of a pharmacogenomics study, particularly those that pertain to personal drug prescription, is to prevent adverse drug reactions and harness the maximum clinical benefits by analyzing genetic variations known to control pharmacodynamic

effects such as drug efficacy, dose requirements, adverse events, etc. Until recently, population-based observational studies were the norm where genetic information of case and control groups, recruited based on altered drug reaction, is investigated to identify less than 10 variants that are critical in controlling drug reactions. The U.S. Food and Drug Administration (FDA) recommends that drugs be prescribed accordingly to the study results. Despite successful applications of these results, population-based observational studies are hindered by: immense budget issues in developing a standard set; results affected by demographic conditions; rare or private variants unable to be included in the results; etc.

In 2013, in an attempt to overcome such issues, we developed the PharmSafe algorithm, which calculates gene and drug scores based on an individual's genetic variation information and ranks drugs that are possibly hazardous. Performance of the algorithm was evaluated using PharmGKB. However, the algorithm only considers variants within the coding region and of all biological knowledge, only protein-protein interaction is applied within the algorithm.

In this paper, we developed a new and improved PharmSafe algorithm where variants from non-coding region and biological functional element including regulational factor such as RNA editings are also considered and seven biological, pharmacological, and statistical knowledge elements are used as weight parameters. In addition to the aforementioned updates, we were able to validate the reproducibility of the PharmSafe algorithm in larger genome

datasets by using 1000 Genome Project Phase 3 data, which expands the preexisting datasets to 2,503 samples.

Of the pharmacodynamic genes, the new PharmSafe algorithm was able to achieve the highest area under the curve (AUC) of 0.5857~0.6502, (0.6224±0.222; minimum ~ maximum, mean±standard deviation) in catalytic enzyme genes. In drug class evaluations, antihypertensives (n=22) had the highest AUC of 0.6234~0.8896 (0.7340±0.0539; minimum ~ maximum, mean±standard deviation).

We believe that the new PharmSafe algorithm would be a valuable tool for a clinical decision support system (CDSS) in prescribing drugs safely and efficiently at the right dosages based on an individual's genetic variation information.

Keyword: Personal Genome, Pharmacogenomics, Pharmacology, Personalized Medicine, Personalized Pharmacogenomics, RNA editing

Student ID : 201030127