



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2018-0128996
(43) 공개일자 2018년12월05일

(51) 국제특허분류(Int. Cl.)
G06F 19/12 (2011.01) C12Q 1/68 (2018.01)
G06F 19/24 (2011.01)
(52) CPC특허분류
G06F 19/12 (2013.01)
C12Q 1/6895 (2018.05)
(21) 출원번호 10-2017-0063812
(22) 출원일자 2017년05월24일
심사청구일자 2017년05월24일

(71) 출원인
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
최인걸
서울특별시 동작구 남부순환로263길 34 501호 (사당동)
민병남
서울특별시 노원구 중계로8길 21(중계동 153 house) 201호
(74) 대리인
특허법인다나

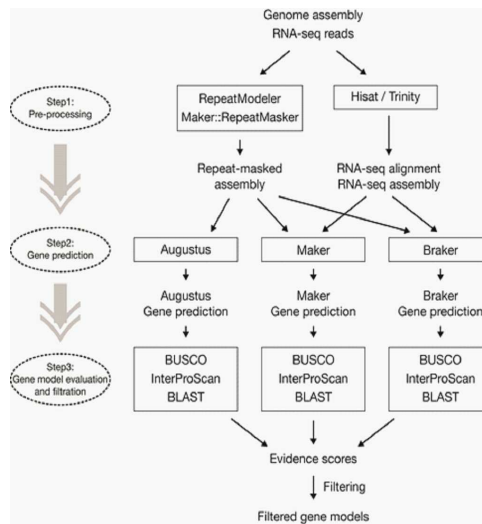
전체 청구항 수 : 총 10 항

(54) 발명의 명칭 진균류 유전체 해독을 위한 시스템 및 방법

(57) 요약

본 발명은 다양한 유전자 예측 프로그램을 통합하고 유전자 예측 모델을 평가하여 진균류 유전체를 해독하는 시스템 및 방법에 관한 것이다.

대표도 - 도1



(52) CPC특허분류

G06F 19/24 (2013.01)

이 발명을 지원한 국가연구개발사업

과제고유번호 R1415434

부처명 농촌진흥청

연구관리전문기관 농촌진흥청

연구사업명 포스트게놈 다부처유전체사업

연구과제명 [협동]버섯류 유전체 해독에 특화된 유전체조립 스캐폴딩 기술과 해독 파이프라인 개발

기 여 율 1/1

주관기관 고려대학교

연구기간 2017.01.01 ~ 2017.12.31

명세서

청구범위

청구항 1

조립된 유전체 서열 내 반복 서열을 찾아 마스킹(masking)을 하거나 mRNA 서열분석 리드(reads)를 정렬 및 조립하여 콘티그(contig)를 생성하는 전처리부;

하나 이상의 유전자 예측 프로그램을 사용하여 유전자 모델을 생성하는 유전자 예측부; 및

상기 유전자 모델을 평가하여 최종 유전자 모델을 선별하는 유전자 평가부를 포함하는, 진균류 유전체 해독 시스템.

청구항 2

제 1항에 있어서,

상기 유전자 예측 프로그램은 Augustus, Braker 및 Maker인 진균류 유전체 해독 시스템.

청구항 3

제 1항에 있어서,

유전자 모델의 평가는 BLASTp, BUSCO 및 InterProScan 으로부터 얻은 점수를 합산한 값인 증거 점수가 높은 유전자 모델을 최종 유전자 모델로 선별하는 것인, 진균류 유전체 해독 시스템.

청구항 4

(A) 조립된 유전체 서열 내 반복 서열을 찾아 마스킹(masking)하거나 mRNA 서열분석 리드들(reads)을 정렬 및 조립하여 콘티그(contig)를 생성하는 전처리 단계;

(B) 하나 이상의 유전자 예측 프로그램을 사용하여 유전자 모델을 생성하는 단계; 및

(C) 상기 유전자 모델을 평가하여 최종 유전자 모델을 선별하는 단계를 포함하는, 진균류 유전체 해독 방법.

청구항 5

제 4항에 있어서,

mRNA 서열분석 리드들을 정렬 및 조립하여 콘티그를 생성하는 것은 Hisat 및 Trinity 프로그램에 의해 수행되는 것인, 진균류 유전체 해독 방법.

청구항 6

제 4항에 있어서,

상기 유전자 예측 프로그램은 Augustus, Braker 및 Maker인 진균류 유전체 해독 방법.

청구항 7

제 4항에 있어서,

유전자 모델의 평가는 BLASTp, BUSCO 및 InterProScan 으로부터 얻은 점수를 합산한 값인 증거 점수가 높은 유전자 모델을 최종 유전자 모델로 선별하는 것인, 진균류 유전체 해독 방법.

청구항 8

제 7항에 있어서,

상기 증거 점수는 하기 방정식에 의해 계산되는 것인, 진균류 유전체 해독 방법:

$$\text{Evidence score} = (BLAST\ score * coverage) + BUSCO\ score + \sum Pfam\ scores.$$

청구항 9

제 4항 내지 제 8항 중 어느 한 항에 있어서,

상기 방법은 각 단계 또는 각 단계에서 사용되는 프로그램을 연결하는 파이프라인 프로그램을 통해 자동적으로 수행되는 것인, 진균류 유전체 해독 방법.

청구항 10

제 9항의 방법을 수행하기 위한 컴퓨터로 판독 가능한 프로그램을 기록한 기록매체.

발명의 설명

기술 분야

- [0001] 본 발명은 유전체의 유전자 모델을 예측하여 해독하는 시스템 및 방법에 관한 것이다.
- [0002] 구체적으로는 진균류 유전체에 특화된 시스템으로 진균류 유전체 및 mRNA 서열분석 데이터의 전처리, 유전자 모델 예측, 평가를 일련의 과정으로 자동수행하는 유전체 해독 파이프라인 소프트웨어 시스템에 관한 것이다.

배경 기술

- [0003] 진균류(Fungi)는 세균(Bacteria)에 비해 크고 복잡한 유전체 구조를 지니고 있으며, 2차 대사산물 생산을 위한 산업 균주로서의 활용 및 환경복원에의 다양한 응용 가능성때문에 이들 균주에 대한 유전체 분석이 활발히 진행되고 있다.
- [0004] 미국 국립생물정보센터(NCBI)에는 약 700개의 균류 유전체가 공개되어 있으며, 조인트 게놈 인스티튜트에서 제공하는 마이코코즘(<http://genome.jgi.doe.gov/programs/fungi/index.jsf>)에는 약 500여개의 균류 유전체가 공개되어 있다. 균류 유전체의 분석 및 데이터 공개의 속도는 해마다 증가하는 추세이다.
- [0005] 균류 유전체의 크기는 2.1 Mbp(Microsporidia에 속하는 기생균류)부터 215 Mbp (Uromyces viciae-fabae, 담자균류)까지 다양하며 평균 35 Mbp의 크기를 가지고 있다. 이는 대장균(4.6 Mbp)과 같은 세균류의 유전체 크기보다는 크며 인간(3 Gbp)과 같은 영장류의 유전체의 크기보다는 훨씬 작은 크기이다.
- [0006] 유전자 예측이란, 주어진 유전체 서열에서 유전자의 위치를 찾는 것이다. 원핵생물의 유전체의 경우 시작코돈과 종결코돈을 찾는 것으로 비교적 단순하지만, 진핵생물의 유전체의 경우 엑손-인트론(exon-intron)의 구조를 고려해야 하기 때문에, 단순한 알고리즘으로는 유전자 예측에 어려움이 있다.
- [0007] 진균류 유전체의 유전자 예측을 수행하는 프로그램으로는 Augustus, GeneMark, Maker, Braker 등이 알려져 있다. 이들 프로그램들은 각기 다른 알고리즘을 이용한다. 예를 들면, Augustus는 모델 유전체로부터 유전자 모델을 구축하여 유전자 예측을 위해 사용하며, Braker는 RNA 서열로부터 직접 유전자 모델을 구축하기도 한다. 각 알고리즘이 유전체 서열에 대해 종종 다른 결과를 내며, 알고리즘마다 예측정확도가 유전자마다 다르기 때문에, 보다 정확한 유전자 예측을 위해서는 이들 프로그램들을 모두 실행시킨 뒤 적절한 과정을 통해 이를 통합하는 것이 유리하다.
- [0008] 이와 같은 목적의 종래기술로는 Evidencemodeler라는 프로그램이 있다. 이 프로그램은 유전자 모델을 이용한 유전자 예측 결과와 단백질 및 전사체 서열의 정렬 결과를 통합한다. 하지만 이 프로그램은 사용자가 제공해야 하는 정보가 많기 때문에 생명정보학 기술이 충분히 숙련되지 않은 연구자가 사용하기에는 어려움이 있다.
- [0009] 또 다른 종래기술로는 위에서 언급한 Maker가 있다. Maker는 여러 프로그램을 이용하여 유전체 내 유전자 예측을 하여 모든 가능한 경우의 유전자를 예측하고, annotation edit distance(AED)를 계산하여 최종 유전자를 선별한다. 하지만 이 방법은 크기가 큰 유전자가 높은 점수를 받기에 유리하며, 유전자가 비교적 가까이 붙어있는 진균류 유전체에서는 근접한 두 유전자가 합쳐서 하나의 유전자로 나타나는 경향을 보인다. 따라서 보다 적절한

유전자들의 선별을 위해서는 개선된 알고리즘이 필요하다.

- [0010] 현존하는 많은 유전자 예측 프로그램들은 각기 다른 알고리즘을 이용하는데, 종종 다른 결과물을 주며, 현재 최적화된 예측 프로그램은 알려져 있지 않다. 따라서 정확한 유전자예측은 가능한 예측 프로그램을 모두 사용하여 예측 유전자 풀(pool)을 생성하고, 적절한 평가방법을 이용해 최적의 유전자를 결과물로 만드는 것이다.
- [0011] 이에 본 발명자들은 유전자 풀에서의 각 유전자를 평가하여 점수화하고, 최적의 유전자를 최종 선별하는 자동화된 소프트웨어 시스템 및 방법을 완성하였다.

선행기술문헌

비특허문헌

- [0012] (비특허문헌 0001) Cantarel, B.L., et al. (2008) MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res*, 18(1), 188-196.
- (비특허문헌 0002) Haas, B.J., et al. (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc*, 8(8), 1494-1512.
- (비특허문헌 0003) Kim, D., Langmead, B. and Salzberg, S.L. (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat Methods*, 12(4), 357-360.
- (비특허문헌 0004) Li, H., et al. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16), 2078-2079.
- (비특허문헌 0005) Stanke, M., et al. (2006) AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res*, 34, W435-439.
- (비특허문헌 0006) Hoff, K.J., et al. (2016) BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics*, 32(5), 767-769.
- (비특허문헌 0007) Camacho, C., et al. (2009) BLAST+: architecture and applications. *BMC bioinformatics*, 10:421.
- (비특허문헌 0008) Simao, F.A., et al. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19), 3210-3212.
- (비특허문헌 0009) Jones, P., et al. (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, 30(9), 1236-1240.

발명의 내용

해결하려는 과제

- [0013] 본 발명의 목적은 진균류 유전체 및 mRNA 서열분석 데이터의 전처리, 유전자 모델 예측, 평가를 일련의 과정으로 자동 수행하는 진균류 유전체 해독 파이프라인 소프트웨어 시스템을 제공하는 것이다.
- [0014] 본 발명의 다른 목적은 진균류 유전체 및 mRNA 서열분석 데이터의 전처리, 유전자 모델 예측, 평가를 일련의 과정을 포함하는 진균류 유전체 해독방법을 제공하는 것이다.

과제의 해결 수단

- [0016] 본 발명은 조립된 유전체 서열 내 반복 서열을 찾아 마스킹(masking)을 하거나 mRNA 서열분석 리드를 정렬 및 조립하여 콘티그(contig)를 생성하는 전처리부;
- [0017] 하나 이상의 유전자 예측 프로그램을 사용하여 유전자 모델을 생성하는 유전자 예측부;
- [0018] 상기 유전자 모델들을 평가하여 최종 유전자 모델을 선별하는 유전자 평가부를 포함하는, 진균류 유전체 해독

시스템을 제공한다.

[0020] 또한, 본 발명은,

[0021] (A) 조립된 유전체 서열 내 반복 서열을 찾아 마스킹하거나 mRNA 서열분석 리드들을 정렬 및 조립하여 콘티그를 생성하는 전처리 단계;

[0022] (B) 하나 이상의 유전자 예측 프로그램을 사용하여 유전자 모델을 생성하는 단계;

[0023] (C) 상기 유전자 모델을 평가하여 최종 유전자 모델을 선별하는 단계를 포함하는, 진균류 유전체 해독 방법을 제공한다.

[0025] 또한, 본 발명은 상기 진균류 유전체 해독 방법을 자동으로 수행하기 위한 컴퓨터로 판독 가능한 프로그램을 기록한 기록매체를 제공한다.

발명의 효과

[0026] 이 발명의 핵심은 유전자를 평가하는 방법에 있어 세 가지 방법을 동시에 이용할 수 있도록 통합시키는 알고리즘이다. 이 세 가지 방법은 Pfam 도메인 검색, BLAST를 이용한 단백질 서열 정렬, BUSCO 프로그램을 이용한 단일-카피 ortholog 검색방법이다. 다양한 서열정보를 이용하기 때문에 종래의 기술에 비교했을 때 더 높은 정확도를 기대할 수 있다.

[0027] 또한 본 발명의 방법의 각 단계 및 각 단계에서 사용되는 프로그램이 하나의 파이프라인 프로그램으로 엮여져 있기 때문에, 한 번의 프로그램 실행으로 필요한 모든 결과물을 얻을 수 있으며, 이는 다른 프로그램에서는 시도되지 않았던 내용이다. 궁극적으로 사용자가 직접해야하는 부분을 최소로 하여 유전체의 유전자 예측에 있어 연구자가 믿을만한 결과물을 손쉽게 얻을 수 있게 하는 것이 이 발명의 주된 효과이다.

도면의 간단한 설명

[0028] 도 1은 진균류 유전체 해독 파이프라인 전체 모식도이다.

도 2는 유전자 모델 평가 및 선별 모식도이다.

도 3은 유전자 모델 평가 및 최종 유전자 선별 예시로, *N. crassa* 유전체 중 일부분의 유전자 평가 과정을 도식화하였다.

도 4는 본 유전자 해독 파이프라인의 각 프로그램의 러닝 타임을 나타내었다.

발명을 실시하기 위한 구체적인 내용

[0029] 이하, 본 발명을 구체적으로 설명한다.

[0031] 본 발명은 진균류 유전체 해독 시스템을 제공한다.

[0032] 상기 시스템은 조립된 유전체 서열 내 반복 서열을 찾아 마스킹(masking)을 하거나 mRNA 서열분석 리드를 정렬 및 조립하여 콘티그(contig)를 생성하는 전처리부;

[0033] 하나 이상의 유전자 예측 프로그램을 사용하여 유전자 모델을 생성하는 유전자 예측부; 및

[0034] 상기 유전자 모델을 평가하여 최종 유전자 모델을 선별하는 유전자 평가부를 포함한다.

[0035] 일 구체예에서 상기 시스템은 전처리부; 유전자 예측부; 및 유전자 평가부로 구성될 수 있으며, 각 구성에서 사용되는 프로그램은 하나의 파이프라인으로 연결될 수 있다.

[0036] 상기 전처리부에서 사용되는 조립된 유전체는 일반적으로 알려진 유전자 조립 알고리즘을 사용하는 다양한 오픈 소스 프로그램을 이용하여 수행될 수 있다. 또 다른 구체예에서, 상기 조립된 유전체는 이하의 각 부로 구성되는 시스템을 사용하여 생성된 것일 수 있다:

- [0037] 수집부 : 수집부는 수집된 염기서열 조각(인서트)을 페이드-엔드(paired-end) 시퀀싱하여 하나의 조각 당 두 개의 리드(read) 서열을 수집 및 저장한다. 이때, 염기서열은 균류로부터 추출한 것일 수 있다.
- [0038] 리드 전처리부 : 리드 전처리부는 상기 수집부에서 수집된 리드 서열을 전처리 한다. 상기 전처리부에서는 리드 품질요약, 어댑터 서열제거, 리드 품질보정 및 k-mer 빈도 계산으로 이루어진 그룹 중에서 하나 이상을 수행할 수 있다.
- [0039] 유전체 조립부 : 유전체 조립부는 하나 이상의 유전체 조립 소프트웨어를 사용하여 하나 이상의 후보 유전체를 조립할 수 있다.
- [0040] 상기 유전체 조립부는 ALLPATHS 파이프라인부 및 Non-ALLPATHS 파이프라인부를 포함할 수 있다. 하나의 조각에서 수집된 두 개의 리드 서열의 합이 조각의 서열보다 길 경우 ALLPATHS 파이프라인부에서 조립이 수행되며, 하나의 조각에서 수집된 두 개의 리드 서열의 합이 조각의 서열보다 짧을 경우 Non-ALLPATHS 파이프라인부에서 조립이 수행될 수 있다.
- [0041] 유전체 조립 평가부 : 상기 유전체 조립부에서 조립된 후보 유전체를 평가하여 최적 유전체를 선별한다. 상기 유전체 조립 평가부에서는 우도 평가, N50값 평가, 스캐폴드 개수 평가 또는 최대 스캐폴드 길이 평가로 이루어진 그룹 중에서 하나 이상을 수행하여 후보 유전체를 평가하며, 각각의 평가 별로 순위를 부여하여 순위의 합이 가장 적은 유전체를 최적 유전체로 선별할 수 있다.
- [0042] 유전체 조립 후처리부 : 유전체 조립 평가부에서 선별된 최적 유전체를 후처리할 수 있다. 상기 후처리부에서는 외래 DNA에 의한 오염 확인, 28S rDNA 서열 확인, 미토콘드리아 서열 조립 및 웹서버 업로드로 이루어진 그룹 중에서 하나 이상을 수행할 수 있다.
- [0044] 상기 유전자 예측부는 조립된 유전체를 다양한 유전자 예측 프로그램을 사용하여 유전자를 예측하여 유전자 모델을 생성할 수 있다. 상기 유전자 예측 프로그램은 Augustus (Stanke, et al., 2006, 비특허문헌 5), Braker (Hoff, et al., 2016, 비특허문헌 6), 및 Maker (Cantarel, et al., 2008, 비특허문헌 1) 프로그램일 수 있고, 이제 제한되지 않고 알려진 유전자 예측 프로그램을 모두 사용할 수 있다. 상기 유전자 예측 프로그램들을 사용하여 얻어진 예측된 유전자 모델들을 ‘유전자 풀(pool)’이라 본 발명에서 지칭한다.
- [0045] 상기 유전자 평가부는 상기 유전자 예측부에서 생성된 유전자 모델을 평가하여 이 중에서 가장 실제에 가까운 유전자 모델을 선별한다. 상기 유전자 모델 평가부에서는 각 유전자 모델 및 유전자 모델 조합을 BLASTp (Camacho, et al., 2009, 비특허문헌 7), BUSCO(Benchmarking Universal Single-Copy Orthologs)(Simao, et al., 2015, 비특허문헌 8), 및 Pfam 도메인 검색(InterProScan, Jones, et al., 2014, 비특허문헌 9) 중 하나 이상의 정렬 프로그램을 사용하여 평가할 수 있다. 상기 프로그램을 사용하여 얻은 결과를 점수화하여 그 합산 점수가 가장 높은 유전자 모델 또는 유전자 모델 조합을 최종적으로 선별할 수 있다. 본 발명에서 상기 평가 프로그램을 이용하여 얻은 점수들을 합산한 것을 “증거 점수(Evidence score)”라고 지칭하고, 하기의 방정식에 의해 계산될 수 있다.
- $$\text{Evidence score} = (BLAST\ score * coverage) + BUSCO\ score + \sum Pfam\ scores.$$
- [0046]
- [0048] 또한, 본 발명은 진균류 해독 방법에 관한 것이다.
- [0049] 상기 방법은,
- [0050] (A) 조립된 유전체 서열 내 반복 서열을 찾아 마스킹하거나 mRNA 서열분석 리드들을 정렬 및 조립하여 콘티그를 생성하는 전처리 단계;
- [0051] (B) 하나 이상의 유전자 예측 프로그램을 사용하여 유전자 모델을 생성하는 단계;
- [0052] (C) 상기 유전자 모델들을 평가하여 최종 유전자 모델을 선별하는 단계를 포함한다.
- [0053] 이하, 첨부된 도면을 참조하여 발명을 실시하기 위한 구체적인 내용을 상세히 설명한다.

- [0055] 일 구체예에서 진균류 유전체 해독 방법은 도 1에 나타난 바와 같이,
- [0056] 상기 전처리 단계(A)에서는 RepeatModeler (<http://www.repeatmasker.org/RepeatModeler.html>)를 사용하여 조립된 유전체 서열 내 반복 서열을 찾고, Maker pipeline(Cantarel, et al., 2008, 비특허문헌 1)의 반복 부분 마스킹 기능을 수행하여 반복 서열을 마스킹할 수 있다. 또한, 전처리 단계에서 진균류로부터 얻은 mRNA-서열 리드들은 Trinity 프로그램 (Haas, et al., 2013, 비특허문헌 2), Hisat2 read aligner (Kim, et al., 2015, 비특허문헌 3) 및 Samtools format converter (SAM file to sorted BAM file) (Li, et al., 2009, 비특허문헌 4)를 사용하여 어셈블리할 수 있다.
- [0057] 상기 반복 서열의 마스킹 작업을 수행함으로써 반복 서열이 유전자 정렬 및 유전자 예측을 방해하는 것을 방지할 수 있다.
- [0058] 상기 전처리 단계를 거쳐 생성된 반복 서열이 마스킹된 유전체 및 RNA 서열분석 리드들이 조립된 콘티그는 이후 유전자 예측 단계의 프로그램에서 이용될 수 있고, 구체적으로 상기 반복 서열이 마스킹된 유전체는 Augustus (Stanke, et al., 2006, 비특허문헌 5), Braker (Hoff, et al., 2016, 비특허문헌 6), 및 Maker (Cantarel, et al., 2008, 비특허문헌 1) 프로그램에서 사용될 수 있고, 상기 RNA 서열분석 리드들이 조립된 콘티그는 Braker 및 Maker 프로그램에서 사용될 수 있다.
- [0060] 상기 유전자 예측 단계(B)에서는 Augustus, Braker 및 Maker 세 개의 유전자 예측 프로그램을 사용하여 유전자 모델을 예측하여 생성할 수 있다. 각 프로그램으로부터 예측 결과는 GFF3 및 FASTA 파일로 저장될 수 있다.
- [0061] 상기 본 발명의 Augustus 프로그램은 모델 유전체에 대해 유전자 모델을 제공하며 이 모델을 이용하여 주어진 유전체에 대해 유전자 예측을 수행한다. 즉 시작코돈부터 종결코돈까지의 엑손-인트론의 구조를 예측한다. 유전자 모델은 사용자가 지정한 *augustus_species* 파라미터로 결정될 수 있다. *softmasking* 기능이 켜져 소프트-마스킹된 조립체를 인식한다. 오버랩된 CDS 예측을 위해, 본 발명은 *singlestrand* 기능을 실행한다. 결과는 GFF3 파일로 출력되고 번역된 단백질 서열은 간단한 본 발명의 일부인 스크립트에 의해 FASTA 포맷으로 생성된다.
- [0063] 상기 유전자 예측 단계에서 Braker 프로그램은 GeneMark-ET 및 Augustus를 사용하여 비 감독 RNA 서열 기반 유전체 해독을 수행한다. 상기 GeneMark-ET는 RNA-서열 데이터를 비 감독 트레이닝에 도입하여 *ab initio* 유전자 예측 모델을 생성하는 유전자 예측 프로그램이다. 본 발명의 Braker 프로그램은 *softmasking* 기능이 켜져 소프트-마스킹된 조립체를 인식한다. Braker의 입력 파일은 전처리 단계에서 생성된 BAM 파일에서 서식이 지정된 mRNA 리드 정렬(mRNA reads alignment)이다.
- [0064] 상기 본 발명의 Maker 프로그램은 이동성이 좋고 쉽게 구성할 수 있는 유전체 주석 파이프라인으로, MAKER는 독립적으로 유전체에 주석을 달고 유전자 데이터베이스를 만드는 소규모의 진핵 생물 및 원핵 생물의 게놈 프로젝트를 가능하게 하는 프로그램이다. MAKER의 주요 기능은 반복 서열을 식별하고, EST와 단백질을 유전체에 정렬시키고, *ab initio* 유전자 예측 모델을 생성하는 것이다. 본 발명에서 Maker 프로그램은 반복적인 SNAP 유전자 모델 트레이닝을 4회 수행한다. 본 발명은 상술한 진균류 게놈의 높은 유전자 밀도로 인해 *correct_est_fusion* 기능을 사용하여 mRNA 조립에서 이웃 전사체의 융합을 수정한다. 인트론의 최대 길이는 *split_hit* 기능으로 5000bp로 설정된다. *single_exon*과 *single_length* 기능을 설정하여 50개 이상의 아미노산보다 긴 단일 엑손 유전자를 예측할 수 있다.
- [0066] 상기 유전자 모델 평가(C) 단계에서는 전 단계에서 세 가지 프로그램에 의해 생성된 모든 유전자 모델들이 BLASTp, BUSCO(Benchmarking Universal Single-Copy Orthologs) 및 InterProScan(Pfam 도메인 검색) 중 하나 이상의 정렬 프로그램을 사용하여 평가될 수 있다.
- [0067] 구체적으로 상기 유전자 예측 단계(B)에서 상기 유전자 예측 프로그램에서 의해 생성된 모든 유전자 모델들에 대하여 하나 이상의 염기쌍이 오버랩되는 유전자 모델들의 집합인 유전자 블록을 생성할 수 있다. 그 다음 상기 유전자 블록 내에서 10% 이상의 코딩 서열이 오버랩되지 않은 유전자 모델들의 모든 조합을 생성하고, 상기 유전자 모델의 조합에 대해 BLASTp, BUSCO(Benchmarking Universal Single-Copy Orthologs) 및 InterProScan(Pfam 도메인 검색) 프로그램으로 평가하여 점수로 환산할 수 있고, 상기 세 개의 프로그램으로부

터 얻은 점수를 합한 합산 점수인 증거 점수를 하기의 방정식에 의해 얻을 수 있다(도 2).

$$\text{Evidence score} = (BLAST\ score * coverage) + BUSCO\ score + \sum Pfam\ scores.$$

[0068]

[0070]

상기 BLASTp는 주어진 서열 데이터베이스로부터 유사한 서열을 검색하는 프로그램으로, 본 발명에서는 예측된 유전자 서열이 데이터베이스에 존재하는지를 검색하여 유전자 모델 평가를 위해 사용될 수 있다. 사용자는 *sister_proteome* 인수를 사용하여 계통 발생학적으로 관련된 생물체의 프로테오미를 제공한다. 본 발명에서는 사용자의 편의를 위해 *download_sister_orgs.py* 스크립트를 제공하고, 상기 스크립트를 통해서 주어진 생물체의 분류군에 대해 NCBI의 단백질 서열을 다운로드할 수 있으며 이는 BLASTp의 데이터베이스로 이용된다. 계산 시간을 줄이기 위해 본 발명은 세 가지 유전자 예측 프로그램으로부터 얻은 유전자 모델을 통합하고 동일한 유전자 모델을 제거하여 비 중복 유전자 모델을 만들 수 있다. BLAST 결과는 서열의 길이가 길수록 더 높은 bit score를 얻으므로 이를 보정하기 위해 서열 길이 커버리지를 곱하여 최종 BLAST 증거 점수를 계산한다.

[0071]

상기 BUSCO 프로그램은 모든 진균 유전체에 보존된 단일 카피 ortholog에 대해 hidden Markov 모델을 데이터베이스 형태로 제공한다. 따라서 예측 유전자가 BUSCO 데이터베이스 모델에 정렬될 경우 다른 그렇지 않은 예측 유전자에 비해 실제 유전자일 가능성이 더 높다고 판단할 수 있다. BUSCO 모델과의 서열 유사도뿐만 아니라 길이를 고려하기 위하여 BUSCO의 점수는 *hmmer* 결과에서의 “전체 서열 점수”와 길이 커버리지[$\min(\text{query length}, \text{target length}) / \max(\text{query length}, \text{target length})$]의 곱으로 계산한다.

[0072]

InterProScan (Pfam domain prediction)은 수동으로 큐레이팅된 단백질 군의 데이터베이스를 제공한다. 이 중 Pfam 도메인으로 주석 처리된 유전자 모델이 실제 유전자일 가능성이 더 높다는 가정하에 상기 프로그램을 유전자 모델 평가에 사용하였다. Pfam에 대한 점수는 InterProScan (*-f XML* 옵션)의 XML 결과에서 *hmmer3-match* 점수에 의해 직접 제공된다. 하나의 유전자 모델에서 다중 도메인의 경우, 그 점수의 합이 사용된다.

[0074]

상기 유전자 블락 내에서 서로 중복되는 유전자 모델의 조합에 대해 증거 점수가 높은 유전자 모델 조합을 최종 유전자로 선별할 수 있다(도 2).

[0075]

구체적으로, *N. crassa*의 특정 유전자 영역을 대상으로 상술한 유전자 모델을 평가하여 선별하는 과정을 도 3에 나타내었다.

[0077]

본 발명은 각 단계 또는 각 단계마다 사용되는 프로그램들을 하나의 파이프라인으로 연결하여, 유전체 전처리, 유전자 예측 및 유전자 모델 평가를 포함하는 유전체 해독을 위한 일련의 과정을 한번에 자동수행할 수 있다. 이와 같이, 각 단계에서 사용되는 프로그램을 하나의 파이프라인 프로그램으로 연결하여 진균류 유전체를 해독하는 시스템 및 방법을 수행하는 소프트웨어 프로그램을 본 발명에서 “FunGAP(Fungal Genome Annotation Pipeline)”으로 지칭한다. 즉, 본 발명에서는 진균류 유전체의 유전자 예측에 대하여 다양한 프로그램을 동시에 적용하고 이를 평가하여 최종 유전자 모델을 생성하는 자동화된 소프트웨어적 파이프라인을 제공할 수 있다. 이에 따라 각 프로그램을 독립적으로 시행할 때보다 인력을 최소한으로 할 수 있으며, 다양한 정보를 한 번의 프로그램 실행으로 얻을 수 있다.

[0079]

아울러, 본 발명은 상기 진균류 유전체 해독 방법을 자동으로 수행하기 위한 프로그램을 포함하는 컴퓨터 판독 가능한 기록매체를 제공한다.

[0081]

[실시예]

[0082]

본 발명의 진균류 유전체 해독 파이프라인(FunGAP)을 평가하기 위해, *Saccharomyces cerevisiae* (yeast), *Neurospora crassa* (Ascomycete), *Schizophyllum commune* (Basidiomycete), *Rhizopus oryzae* (Mucoromycotina), 및 *Gonapodya prolifera* (Chytridiomycota)와 같이 다섯 종의 진균 유전체를 가지고 유전자 예측 프로그램(Augustus, Braker 또는 Maker)과 본 발명을 수행하였다. 사용된 진균 유전체에 대한 기본 정보

는 하기 표 1과 같다.

표 1

Species	GenBank ID	RNA-seq reads from SRA* (# reads)	Genome size	Number of genes in reference
<i>Saccharomyces cerevisiae</i> (yeast)	GCF_000146045.2	SRX495162 (58.4 M)	12.2 Mbp	5,913
<i>Neurospora crassa</i> (Ascomycota)	GCF_000182925.2	SRX042065 (31.3 M)	41.0 Mbp	10,785
<i>Schizophyllum commune</i> (Basidiomycota)	GCF_000143185.1	SRX716964 (56.3 M)	38.5 Mbp	13,194
<i>Rhizopus oryzae</i> (Mucoromycotina)	GCA_000149305.1	SRX365613 (117.0 M)	46.1 Mbp	17,459
<i>Gonapodya prolifera</i> (Chytridiomycota)	GCA_001574975.1	SRX125310 (123.2 M)	48.8 Mbp	13,831

본 발명의 평가는 48코어 [Intel(R) Xeon(R) CPU E5-2670 v3] 및 64 GB RAM이 장착된 서버에서 수행하였다, 다중 작업의 경우 40 코어가 사용되었다.

유전자 모델 예측을 평가하기 위해 확인된 기준에는 사용자 제공 단백질 데이터베이스 및 SwissProt에 대한 BLAST hits, 모든 Pfam 도메인 및 이를 포함하는 유전자, 완전한 BUSCO 및 누락된 BUSCO, 모든 발현된 서열 태그 지원 및 높은 커버리지 (> 90 % 커버리지), 이전에 보고된 대표적인 유전자 예측 프로그램의 결과(참조)와의 비교가 포함되었다. 결과적으로 신뢰할 수 있는 점수 기능을 갖춘 여러 프로그램을 통합하면 이러한 프로그램을 별도로 사용하는 것보다 더 나은 결과를 얻는 것으로 나타났다(표 2 내지 6). 특히 3 개의 게놈 (*S. commune*, *R. oryzae* 및 *G. prolifera*)에 대한 본 발명(FunGAP)의 유전자 예측 결과는 Pfam 도메인 포함 유전자와 같이 참조 프로그램에 의한 예측보다 많은 증거가 있는 유전자를 가지고 있으며 BUSCO 평가에서 더 완전함을 확인하였다(표 4 내지 6).

위에서 설명한 수치적 평가 이외에 우리는 *S. cerevisiae* 유전자 예측과 잘 정리된 참조 프로그램 유전자 예측을 직접 비교하여 본 발명의 효과를 검증하였다. FunGAP은 참조 (5913 유전자)보다 적은 유전자 모델 (5654 개 유전자)을 예측하였다(표 2). 두 예측 모델은 5222 (90.22 %)의 동일한 유전자 구조를 가지고 432 및 691 개의 유전자가 각각 FunGAP 및 참조 유전자 모델에만 포함되었다. FunGAP에서 691 개의 주석이 달려 있지 않은 유전자를 조사한 결과 191 개의 유전자가 반복서열을 마스킹하는 전처리 단계에 의해 마스킹된 반복 영역에 위치한다는 것을 발견했다. 또한, 170 개의 유전자가 300bp 미만의 코딩 서열로 짧았고, 177 개의 유전자가 다른 시작 부위를 가지지만 동일한 엑손-인트론 구조를 가진 것을 확인하였다. 이러한 결과를 바탕으로 FunGAP가 세 가지 유전자 예측 프로그램을 안정적으로 통합한다는 것을 확인하였다.

표 2

Saccharomyces cerevisiae 게놈 유전자 예측 평가

Metrics	Augustus v3.2.1	Braker v1.8	Maker v2.31.8	FunGAP	Reference
Structural features					
- Predicted genes	6,144	5,488	5,523	5,654	5,913
- Median protein length (aa)	381	424	415	411	406
- Single-exon genes	5,416	5,191	4,881	5,275	5,585
- Short genes (<50 aa)	0	0	0	0	69

- Incomplete genes ^a	8	0	0	0	0
- Coding region (Mbp)	8.20	8.26	8.34	8.35	8.78
Evidence-supported					
- BLAST hits to protein DB ^b	5,301	5,371	5,254	5,495	5,710
- BLAST hits to SwissProt	5,327	5,404	5,292	5,534	5,904
- Pfam hits	4,571	4,630	4,535	4,706	4,899
- Pfam domains	7,250	7,310	7,368	7,395	7,728
- Repeat-related Pfam hits	7	7	8	8	100
- Complete BUSCOs	1,400	1,414	1,416	1,419	1,422
- Missing BUSCOs	7	5	1	1	0
- All EST supports	5,242	5,312	5,262	5,499	5,718
- >90% EST supports ^c	4,152	4,262	4,012	4,370	4,493
- Matches to reference	4,776	4,793	4,695	5,222	5,913

^{a)}Existence of internal stop, gap start, short intron, and internal gap

^{b)}The protein database is a set of protein sequences from relatives.

^{c)} $(Matched\ length / gene\ length) * 100$

표 3

Neurospora crassa 게놈 유전자 예측 평가

Metrics	Augustus v3.2.1	Braker v1.8	Maker v2.31.8	FunGAP	Reference
Structural features					
- Predicted genes	12,211	9,009	10,241	10,130	10,785
- Median protein length (aa)	386	441	414	408	430
- Single-exon genes	2,765	2,322	2,430	2,444	2,293
- Short genes (<50 aa)	0	1	40	40	6
- Incomplete genes ^a	7	11	0	0	0
- Coding region (Mbp)	13.36	14.05	15.01	15.03	16.90
Evidence-supported					
- BLAST hits to protein DB ^b	7,930	8,724	9,427	9,768	10,527
- BLAST hits to SwissProt	4,960	5,518	5,393	5,535	6,313
- Pfam hits	5,537	6,151	6,070	6,244	7,100
- Pfam domains	8,813	9,555	9,577	9,705	11,041
- Repeat-related Pfam hits	17	16	18	18	18
- Complete BUSCOs	1,419	1,429	1,418	1,427	1,433
- Missing BUSCOs	1	2	2	1	0
- All EST supports	6,439	6,997	7,169	7,370	8,390
- >90% EST supports ^c	3,815	4,618	4,270	4,592	5,488
- Matches to reference	5,276	6,503	5,789	6,642	10,785

^{a)}Existence of internal stop, gap start, short intron, and internal gap

^{b)}The protein database is a set of protein sequences from relatives.

[0096] ^{c)} (Matched length / gene length) * 100

표 4

[0097] *Schizophyllum commune* 게놈 유전자 예측 평가

Metrics	Augustus v3.2.1	Braker v1.8	Maker v2.31.8	FunGAP	Reference
Structural features					
- Predicted genes	22,428	13,718	13,115	15,337	13,194
- Median protein length (aa)	373	402	442	379	369
- Single-exon genes	4,921	1,097	1,893	2,172	1,469
- Short genes (<50 aa)	0	0	44	40	8
- Incomplete genes ^a	20	215	0	0	0
- Coding region (Mbp)	20.37	19.10	21.07	21.41	17.72
Evidence-supported					
- BLAST hits to protein DB ^b	9,551	12,989	11,050	13,643	13,194
- BLAST hits to SwissProt	4,504	6,467	5,300	6,496	6,278
- Pfam hits	5,587	7,744	6,510	7,863	7,446
- Pfam domains	9,427	11,508	11,025	11,797	10,922
- Repeat-related Pfam hits	80	81	113	107	42
- Complete BUSCOs	1,326	1,354	1,332	1,354	1,240
- Missing BUSCOs	14	9	10	9	19
- All EST supports	10,412	12,127	10,724	13,388	11,723
- >90% EST supports ^c	3,948	8,901	4,965	8,410	7,613
- Matches to reference	1,395	3,448	1,896	3,542	13,194

[0098] ^{a)} Existence of internal stop, gap start, short intron, and internal gap

[0099] ^{b)} The protein database is a set of protein sequences from relatives.

[0100] ^{c)} (Matched length / gene length) * 100

표 5

[0101] *Rhizopus oryzae* 게놈 유전자 예측 평가

Metrics	Augustus v3.2.1	Braker v1.8	Maker v2.31.8	FunGAP	Reference
Structural features					
- Predicted genes	13,146	13,218	15,971	15,548	17,459
- Median protein length (aa)	331.5	345	280	296.0	257
- Single-exon genes	2,313	1,858	1,909	2103	3,923
- Short genes (<50 aa)	0	0	20	18	384
- Incomplete genes ^a	17	33	0	0	0
- Coding region (Mbp)	15.33	16.25	16.77	17.49	17.97
Evidence-supported					
- BLAST hits to protein DB ^b	11,291	12,706	12,439	13,852	17,450
- BLAST hits to SwissProt	7,909	8,757	7,936	8,838	9,034
- Pfam hits	8,469	9,443	8,618	9,596	10,322
- Pfam domains	14,439	15,292	14,863	15,671	16,241

- Repeat-related Pfam hits	65	74	161	159	1,292
- Complete BUSCOs	1,232	1,268	1,280	1,283	981
- Missing BUSCOs	45	32	33	33	119
- All EST supports	9,301	10,337	9,951	10,965	12,859
- >90% EST supports ^c	5,475	6,856	5,355	6,731	6,376
- Matches to reference	3,734	3,826	3,536	4,462	17,459

^{a)} Existence of internal stop, gap start, short intron, and internal gap

^{b)} The protein database is a set of protein sequences from relatives.

^{c)} $(Matched\ length / gene\ length) * 100$

표 6

Gonapodya prolifera 게놈 유전자 예측 평가

Metrics	Augustus v3.2.1	Braker v1.8	Maker v2.31.8	FunGAP	Reference
Structural features					
- Predicted genes	38,867	15,502	2,1551	21,784	13,831
- Median protein length (aa)	238	394	314	308	334
- Single-exon genes	12,512	1,464	4,996	4,686	1,728
- Short genes (<50 aa)	6	4	15	14	2
- Incomplete genes ^a	185	41	0	0	0
- Coding region (Mbp)	26.16	22.12	25.69	26.49	17.15
Evidence-supported					
- BLAST hits to protein DB ^b	14,756	14,357	14,110	17,319	13,831
- BLAST hits to SwissProt	5,781	8,316	6,992	8,776	7,036
- Pfam hits	7,254	9,725	8,482	10,476	8,172
- Pfam domains	11,988	15,362	14,906	16,546	12,581
- Repeat-related Pfam hits	37	31	52	68	51
- Complete BUSCOs	694	1,157	1,082	1,166	1,058
- Missing BUSCOs	322	123	147	116	150
- All EST supports	15,608	13,249	13,000	15,988	11,876
- >90% EST supports ^c	2,300	8,082	3,813	7,579	7,511
- Matches to reference	554	4,770	2,180	4,980	13,831

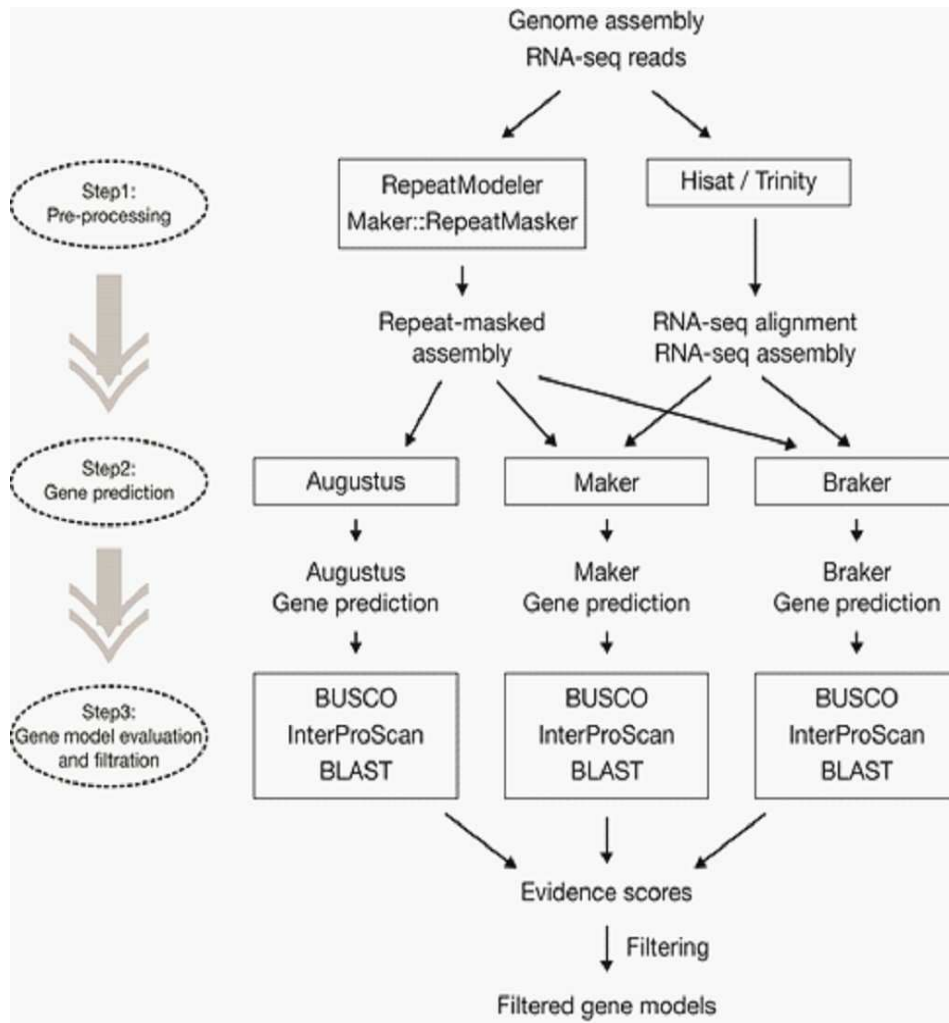
^{a)} Existence of internal stop, gap start, short intron, and internal gap

^{b)} The protein database is a set of protein sequences from relatives.

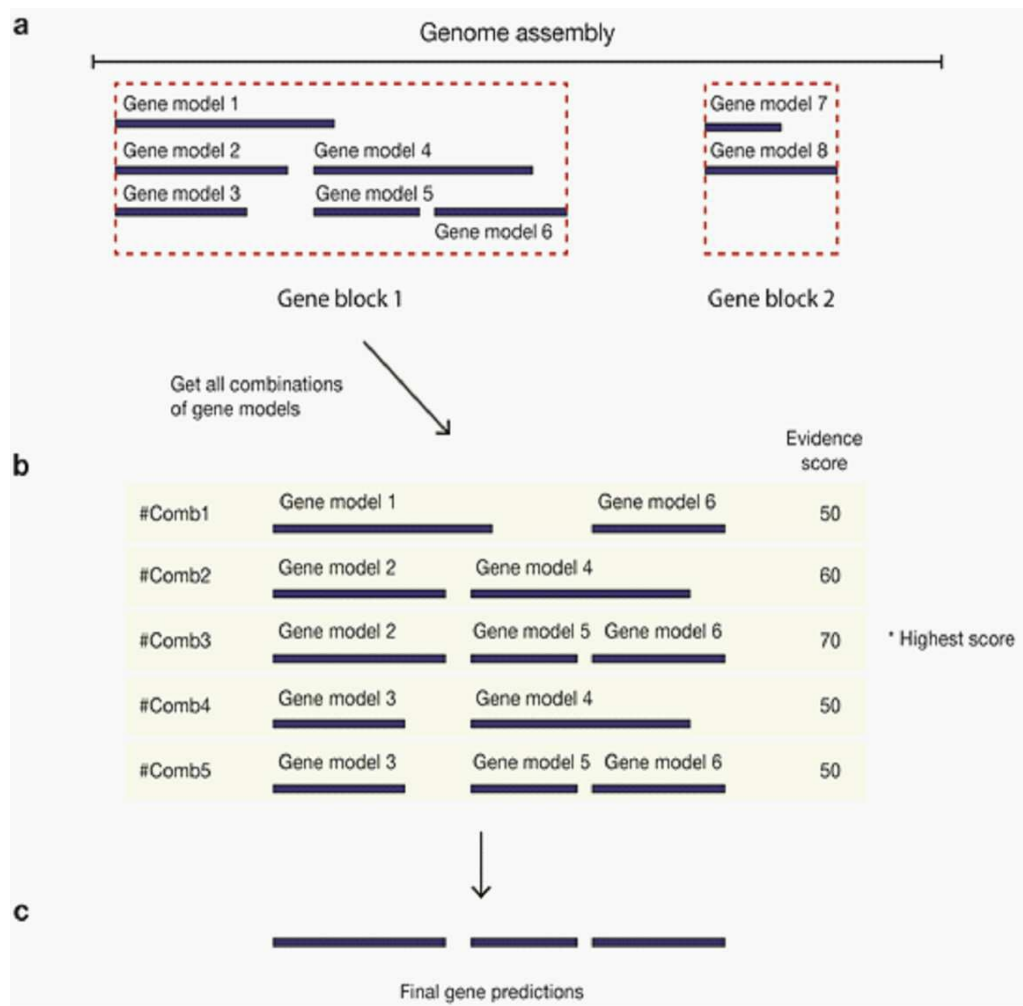
^{c)} $(Matched\ length / gene\ length) * 100$

도면

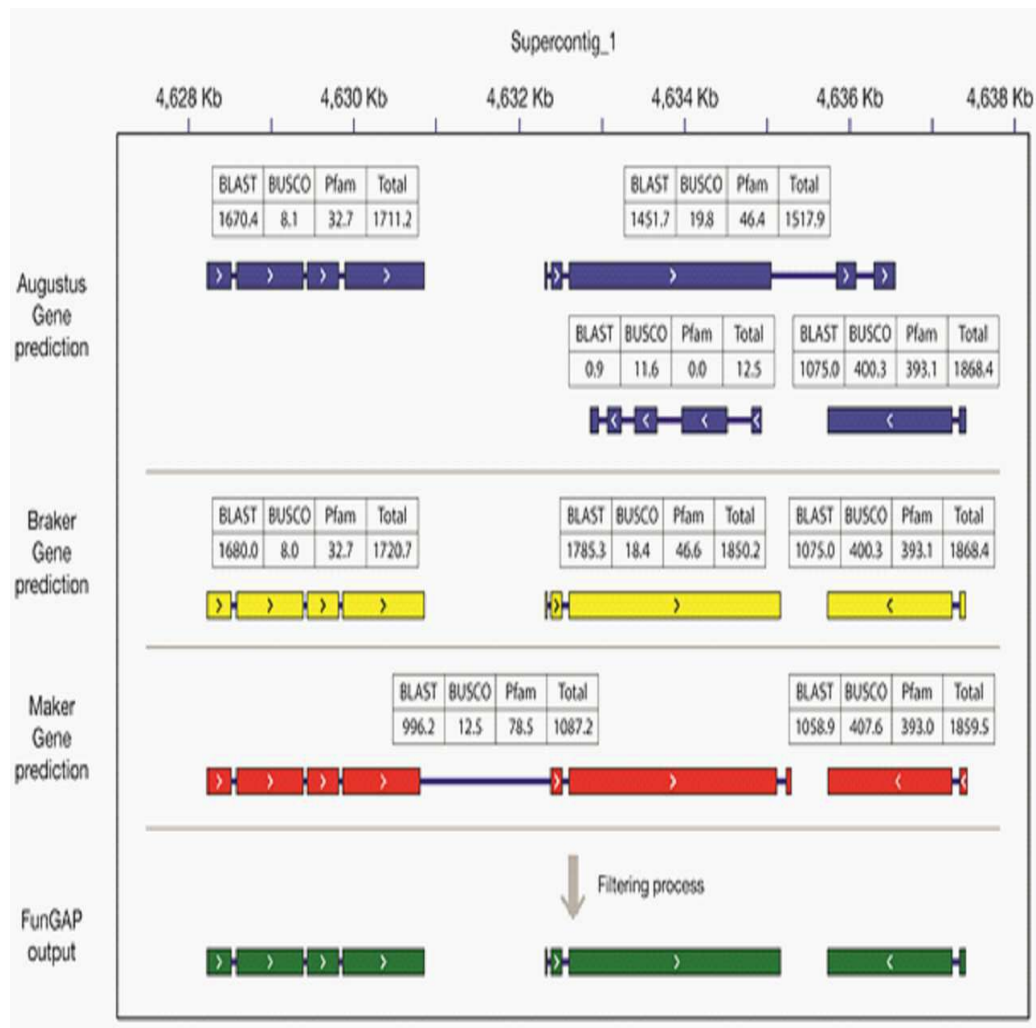
도면1



도면2



도면3



도면4

