Database Lab.

한림대학교

- 지도교수 : 윤 지 희
- 연구분야 : Database, Data Mining, BioInformatics
- 연구실 : 공학관 1314호

1 Introduction

Next Generation Sequencing (NGS) 기반의 유전체 분석 기술

- 유전체 시퀀싱 기술의 개발과 대용량 유전체 데이터의 축적
- 효율적인 NGS 데이터 분석 알고리즘 및 툴 개발

Platform Features

Feature	HiSeq2500 - Highoutput	HiSeq2500 – Rapid mode	MiSeq	PacBio RSII
Number of reads	150-180M/lane	100-150M/lane	12-15M (v2) 20-25M (v3)	50-80K/SMRT cell
Read length	2 x 100 bp	2 x 150 bp	2 x 300 bp (v3)	~ 10-20 kb
Yield per lane (PF data)	up to 35 Gb	up to 45Gb	up to 15 Gb	up to 0.4 Gb
Instrument Time	~12-14 days	~2 days	~2 days	~2 hours
Pricing per Gb	$59 (PE100)	$53 (PE150)	$108 (PE300)	$697

Simple storage for 1 million individuals (30X, WGS) will require
= 100,000,000Gb
= 100,000Tb (terabytes)
= 100Pb (petabytes)

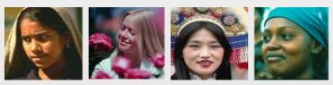
2 Introduction

유전체 구조 변이

- 유전체 구조 변이는 질병에 있어서 직접적인 원인이 되거나 위험인자로 작용
- 유전체 구조 변이
  - 서열 정보에 존재하는 인종/개인간의 차이
  - 삽입(insertion), 삭제(deletion), 전이(inversion), 단일 염기 다형성 (Single Nucleotide Polymorphism, SNP) 등이 존재

GATTTAGATCGCGATAGAG
GATTTAGATCTCGATAGAG

➢ Single nucleotide polymorphisms (SNPs)



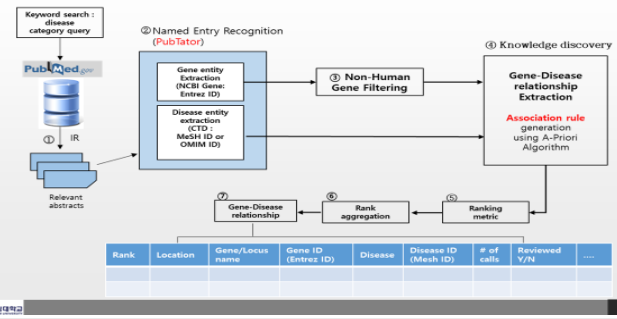
3 연구실 진행 과제

텍스트 마이닝 기반의 질병-유전자 지식 베이스 개발

- 질병-유전자-변이 관련 지식
  - 매우 빠른 속도로 생성 되어 연구 논문 및 특허 등 지속적으로 저장 축적 되고 있음 (10,000 papers per year)
  - 대규모의 관련지식을 추출/정제/가공 하기 위해서는 많은 전문가의 노력과 시간이 필요함
- 본 연구에서는 이와 같은 전문가의 큐레이션 작업 부담을 경감 시켜 데이터베이스 구축 시간을 크게 단축시킬 수 있는 텍스트 마이닝 기법을 제안하여 질병-유전자 관련 지식 베이스를 자동 생성
- 이와 같이 생성된 질병-유전자 지식 베이스는 Genome browser(PGB)에 업로드 되어 적극 활용됨
- 사례연구로써 8개의 퇴행성 뇌질환을 대상으로 하는 지식 베이스 개발 및 성능 평가

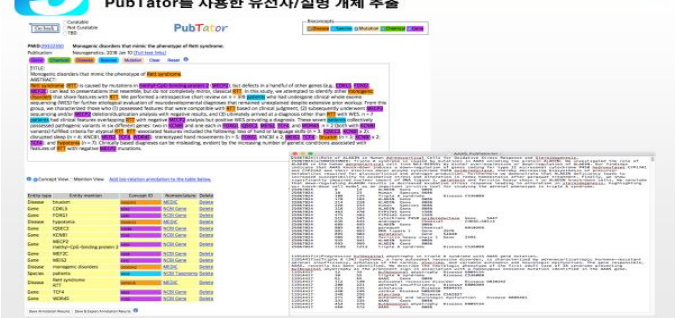
4 연구실 진행 과제

유전자-질병 연관성 추출을 위한 텍스트 마이닝 파이프라인



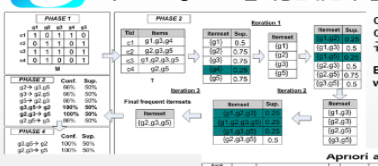
5 연구실 진행 과제

PubTator를 사용한 유전자/질병 개체 추출



6 연구실 진행 과제

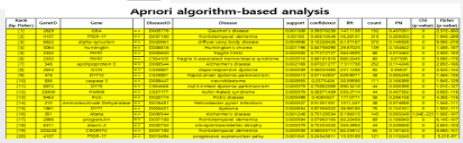
Apriori Algorithm을 이용한 유전자-질병 연관 리스트 추출



어느 두개의(gene, disease) ITEM 집합이 자주 발생하는지를 나타내는 일련의 규칙을 생성

Example of the A-Priori algorithm with support set to 0.25

Apriori algorithm-based analysis

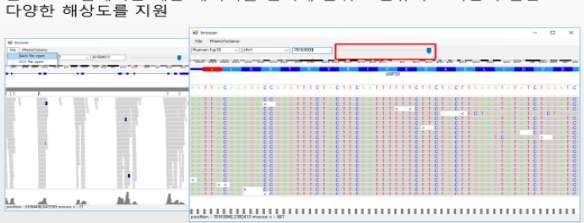


Association rule generation from frequent item sets ranking: Fisher's exact test (p value)

7 연구실 진행 과제

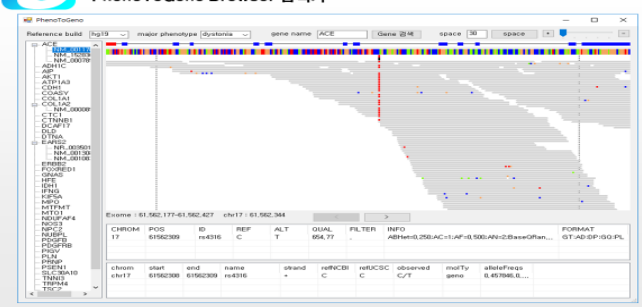
다해상도 분석 기능

- 일반적으로 SNP과 INDEL 같은 경우, 변이의 크기는 작은데 비해 정렬 대상인 레퍼런스 서열은 매우 크므로 다양한 해상도 범위에서의 변이 영역 판독을 지원해 주어야함
- 본 프로그램에서는 게놈 데이터를 염색체 단위로 분류하고 다음과 같은 다양한 해상도를 지원



8 연구실 진행 과제

PhenoToGene Browser 탐색기



- 연구내용
  - 다해상도 분석 가능한 시퀀싱 분석 브라우저(PhenoToGene Browser) 개발
  - 시퀀스 검색을 통한 유전자 변이 검색
  - RNA-Seq 데이터 분석 툴 제작
  - 유전자-질병 연관성 추출을 위한 텍스트 마이닝 기법 연구