최종보고서
[기관고유연구사업]

<table>
<tr><td>과제고유번호</td><td>1310190</td><td>연 구 분 야
(코 드)</td><td></td><td>지 원
프 로 그 램</td><td>창의
일반연구과제</td><td>공개가능여부
(공개, 비공개)</td><td>공개</td></tr>
<tr><td>연 구 사 업 명</td><td colspan="7">국 립 암 센 터 기 관 고 유 연 구 사 업</td></tr>
<tr><td>연 구 과 제 명</td><td colspan="7">국립암센터의 데이터 교환 표준 설계와 클라우드 컴퓨팅 기반의
데이터 분석/공유 시스템 설계 및 구현</td></tr>
<tr><td>과 제 책 임 자</td><td>성명</td><td>홍동완</td><td>소속</td><td colspan="2">종양면역학연구과</td><td>직위</td><td>선임연구원</td></tr>
<tr><td rowspan="4">세 부 과 제</td><td rowspan="2">구분</td><td colspan="3" rowspan="2">과제명</td><td colspan="3">과제책임자</td></tr>
<tr><td>성명</td><td>소속(직위)</td><td>전공</td></tr>
<tr><td>(1세부)</td><td colspan="3"></td><td></td><td></td><td></td></tr>
<tr><td>(2세부)</td><td colspan="3"></td><td></td><td></td><td></td></tr>
<tr><td>(3세부)</td><td colspan="3"></td><td></td><td></td><td></td></tr>
<tr><td rowspan="2">총 연 구 기 간</td><td rowspan="2">2013년1월
~
2015년12월
(총 3년)</td><td>해당단계
참 여
연구원 수</td><td>총: 2 명
내부: 2 명
외부: 명</td><td>해당단계
연 구 개 발 비</td><td colspan="3">연구비:63,000천원
민간: 천원
계: 천원</td></tr>
<tr><td>총 연구기간
참 여
연구원 수</td><td>총: 6 명
내부: 6 명
외부: 명</td><td>총 연구개발비</td><td colspan="3">연구비:183,000천원
민간: 천원
계: 천원</td></tr>
<tr><td rowspan="6">연 구 기 간 및
연 구 비
(단 위 :천 원)</td><td>구분</td><td colspan="2">연구기간</td><td>계</td><td>국립암센터</td><td colspan="3">기업부담금</td></tr>
<tr><td></td><td></td><td></td><td></td><td>소계</td><td>현금</td><td>현물</td></tr>
<tr><td>계</td><td colspan="2">2013.1~2015.12</td><td>183,000</td><td>183,000</td><td></td><td></td><td></td></tr>
<tr><td>제1차</td><td colspan="2">2013.1~2013.12</td><td>50,000</td><td>50,000</td><td></td><td></td><td></td></tr>
<tr><td>제2차</td><td colspan="2">2014.1~2014.12</td><td>70,000</td><td>70,000</td><td></td><td></td><td></td></tr>
<tr><td>제3차</td><td colspan="2">2015.1~2015.12</td><td>63,000</td><td>63,000</td><td></td><td></td><td></td></tr>
<tr><td>참여기업</td><td colspan="7">참여기업명 :</td></tr>
<tr><td>국제공동연구</td><td colspan="3">상대국명:</td><td colspan="4">상대국 연구기관명:</td></tr>
<tr><td>위 탁 연 구</td><td colspan="3">연구기관명:</td><td colspan="4">연구책임자:</td></tr>
<tr><td colspan="8">요약(연구개발성과를 중심으로 개조식으로 작성하되, 500자 이내로 작성합니다)</td></tr>
</table>

2015 년 10 월 28 일

과제책임자 : 홍 동 완 (인)

국 립 암 센 터 원 장 귀 하

< 국문 요약문 >

<table>
<tr><td>연구의
목적 및 내용</td><td>
<최종목표>
- 국립암센터의 다양한 암종의 오믹스 데이터 특성을 고려하여 정확하게 분석할 수 있는 도구를 클라우드 컴퓨팅 플랫폼에서 구축하여 국, 내외 암 연구자들에게 정확한 분석의 기회를 제공

⑴ 암종의 특성을 고려한 오믹스 데이터의 정확한 분석 알고리즘 개발
⑵ 클라우드 환경에서 운용 가능한 암종의 오믹스 데이터 분석 시스템의 구현
⑶ 암종의 변이체 특성을 고려한 데이터 교환 표준의 개발

⑴ 연구 내용
- 암종의 오믹스 데이터를 효율적으로 정확히 분석할 수 있는 알고리즘 개발 및 구현
- Hadoop 클라우드 컴퓨팅 환경에서 분석 시스템을 구현하여 국립암센터의 암 연구자에게 암종의 오믹스 데이터 분석 기회 제공

⑵ 연구 방법
- 암종의 오믹스 데이터의 암 특이성을 고려한 정확한 분석 알고리즘 개발
· 암종의 전장 유전체 시퀀싱, 트랜스크립톰 시퀀싱, 엑솜 시퀀싱 데이터의 특성에 따라 변이체를 정확하게 탐지할 수 있는 분석법 개발
- Hadoop 시스템을 이용한 클라우드 컴퓨팅 환경에서의 오믹스 데이터 분석 알고리즘 구현
· Hadoop 시스템을 설치, 스토리지 클러스터를 구현한 클라우드 컴퓨팅 테스트 환경에서 운용 가능한 암 특이적 분석 알고리즘을 구현
- Map/Reduce 기능을 이용한 효율적인 병렬 프로세싱의 구현
· Hadoop 시스템이 설치된 컴퓨터 환경에서 Map/Reduce를 이용하여 병렬 프로세싱을 지원하는 효율적인 분석 시스템을 개발
</td></tr>
<tr><td>연구개발성과</td><td>
<정량적 성과>
<table>
<tr><th>구분</th><th>달성치/목표치[1)]</th><th>달성도(%)</th></tr>
<tr><td>SCI 논문 편수</td><td>1</td><td>2/100</td></tr>
<tr><td>IF 합</td><td>4</td><td>61.792</td></tr>
<tr><td>기타 성과</td><td></td><td>특허 2편</td></tr>
</table>

<정성적 성과>

- 암종의 오믹스 데이터 분석 시 정확한 somatic point mutation 탐지 방안 제공
- 국립암센터의 암유전체 데이터가 임상에 적용될 수 있도록 하는 정확한 분석법, 데이터 표준 구조 및 플랫폼 제공
- Genome/Transcriptome 의 특성을 동시에 고려한 변이체 탐지법 개발
- Tumor suppressor inactivation의 원인 규명
</td></tr>
</table>

연구개발성과의 활용계획 (기대효과)	- 본 연구 과제 수행 기간 중 개발된 시스템을 활용하여 국립암센터의 Clinical Trials에 정확한 변이체 탐지에 적용 - 3년차 연구과제에서 발견한 이상 스플라이싱을 발생하는 유전자 변이를 임상 환자의 진단에 적용				
중심어 (5개 이내)	클라우드 컴퓨팅 시스템	하둡 파일 시스템	차세대 염기서열	암종의 오믹스 데이터 분석	데이터 교환 표준

〈 SUMMARY 〉

<table>
<tr><td>Purpose&
Contents</td><td><Final goals>

According to the genomic context as well as cancer types, this project develop tools on the cloud computing platform, it can analyze cancer omics data precisely. And, the developed tools are opened to the nationwide researchers worked on cancer research institute or hospitals.

⑴ Development of omics data analysis algorithms according to cancer types
⑵ Implementation of omics data analysis system worked on cloud computing environment
⑶ Development of data exchange standard considering the characteristics of cancer specific variants

<Contents>

- Design and implementation of an analysis method or system which can analyze cancer-specific omics data precisely and efficiently.
- Providing the analysis system developed on a hadoop based platform to researchers of National Cancer Center.

<Methods>

- Development of an precise algorithm which can analyze the cancer-specific omics data such as whole genome sequencing, whole exome sequencing and transcriptome sequencing.
- Development of parallel computational processing and clustered storage on cloud computing environment.</td></tr>
<tr><td>Results</td><td><table>
<tr><td></td><td>Ach./Exp.</td><td>Ratio(%)</td></tr>
<tr><td># of SCI papers</td><td>2/1</td><td>100</td></tr>
<tr><td>Impact Factors</td><td>61.792/4</td><td>100</td></tr>
<tr><td>ETC</td><td>2 Patents</td><td></td></tr>
</table></td></tr>
<tr><td>Expected
Contribution</td><td>- The developed methods in first year of this project could be participated to an analysis pipeline of clinical trials of National Cancer Center.
- After last base exon mutations, which occurred to an abnormal splicing including intron retention or exon skipping, were validated to cancer patients, there are expected to apply to cancer diagnosis parts of National Cancer Center.</td></tr>
<tr><td>Keywords</td><td>Cloud Computing System</td><td>Hadoop File System</td><td>Next Generation Sequencing</td><td>Cancer-specific omics data analysis</td><td>Data exchange standard</td></tr>
</table>

〈 목 차 〉

1. 연구개발과제의 개요

1-1. 연구개발 목적

- 국립암센터의 다양한 암종의 오믹스 데이터 특성을 고려하여 정확하게 분석할 수 있는 도구를 클라우드 컴퓨팅 플랫폼에서 구축하여 국, 내외 암 연구자들에게 정확한 분석의 기회를 제공

⑴ 암종의 특성을 고려한 오믹스 데이터의 정확한 분석 알고리즘 개발
⑵ 클라우드 환경에서 운용 가능한 암종의 오믹스 데이터 분석 시스템의 구현
⑶ 암종의 변이체 특성을 고려한 데이터 교환 표준의 개발

<당해연도목표>

⑴ 암종의 특성을 고려한 오믹스 데이터의 분석 알고리즘 개발
☞ 전장 유전체 시퀀싱, 트랜스크립톰 시퀀싱, 엑솜 시퀀싱에 대한 효율적이고 정확한 분석법 개발 및 구현

⑵ HDFS (Hadoop File System) 환경의 테스트 베드 구축
☞ 데스크탑 컴퓨터의 클러스터링을 통한 테스트베드 구축, 클라우드 컴퓨팅의 실험환경 제공

⑶ Map/Reduce 기법을 적용한 암종의 오믹스 데이터의 서열 정렬 및 변이체 탐지 시스템 구축
☞ 클라우드 컴퓨팅 환경에서 병렬 프로세싱이 가능한 오믹스 분석 시스템의 구현

1-2. 연구개발의 필요성

⑴ 대용량 데이터의 관리

① TCGA (The Cancer Genome Atlas) 뿐 아니라 ICGC (International Cancer Genome Consortium) 에서 생산한 데이터는 수십 페타에 달하는 것으로 실시간에 가까운 분석 및 데이터 조작 (manipulation) 이 필요함.

② 암종의 오믹스 데이터를 분석하기 위한 고성능 컴퓨팅 시스템 및 대용량 스토리지 클러스터 구축에 대한 부담이 예상되어 클라우드 컴퓨팅 환경으로 개발 플랫폼 전향

⑵ 국립암센터의 암종의 오믹스 데이터의 정확한 분석

① 국립암센터의 암 연구 전문가들은 암종의 오믹스 데이터를 통한 양질의 연구 결과를 도출하기 위한 노력 중

② 국립암센터의 암 연구 전문가들의 시퀀싱 수요가 증가할 것으로 예상되어 암종의 특이성을 고려한 정확한 분석법 개발이 필요

③ 향후 국립암센터에서 우수한 연구 실적을 도출하기 위하여 다양한 분석법을 적용한 계속적인 분석 실험과 암 연구자간의 커뮤니케이션이 지속되어야 함. 점진적으로 컴퓨팅 리소스 추가 또한 이루어져야 함

1-3. 연구개발 범위

⑴ 클라우드 컴퓨팅 환경에 운용 가능한 암종의 오믹스 데이터 분석 시스템/플랫폼 구축

① 본 연구의 1차년 과제에서는 분석 시스템의 기본 플랫폼으로 클라우드 컴퓨팅 시스템 환경을 구축. 2, 3차년 계속 과제에서의 개발 플랫폼으로의 확장 뿐 아니라 향후 본 과제에서 수행/개발된 분석 알고리즘의 테스트 환경의 역할을 함

② 본 연구 과제에서 구축한 클라우드 컴퓨팅 시스템은 고성능 컴퓨팅 환경에서 운용 가능한 분석 소프트웨어 등이 운용 가능하기 때문에 “암종의 유전체 데이터 분석 시스템 구축 및 운영” 등의 연구과제에서 활용할 수 있는 분석 파이프라인을 선 구축 테스트 가능

⑵ 암종의 특성을 고려한 오믹스 데이터의 정확한 분석 알고리즘 개발

① 본 연구과제의 책임자는 국립암센터의 암종의 오믹스 데이터를 분석하기 위하여 국립암센터에 최적화된 파이프라인을 구축
구축된 고성능 컴퓨팅 시스템과 분석 시스템을 기반으로 국립암센터의 연구자들에게 분석 업무를 지원하였으며, 정확도 높은 분석 기법 개발 및 지원이 계속 요구되었음

② Somatic point mutation 탐지 과정에서 발생할 수 있는 문제점을 발견하고 이를 해결할 수 있는 시스템을 개발하여 공개적으로 오픈하였음

⑶ 암종의 오믹스 데이터 수집

① 국립암센터의 오믹스 데이터 및 TCGA에서 공개한 암종의 오믹스 데이터 수집
(데이터 수집은 개발 도구의 목적에 맞는 암종의 오믹스 순서로 진행했으며,
본 과제의 개발 목적과 국립암센터 연구자의 요구에 따라 진행)

② 수집된 데이터를 분석 알고리즘의 성능 개선을 위한 테스트 데이터로 활용

③ 공개된 오믹스 데이터를 암종 별로 분석 후 통합 연구 결과에서 새로운 의미 도출 시도

⑷ 암종의 오믹스 데이터 분석 알고리즘/방법 개발 및 개선

① Genomic aberrant mutation 중 alternative splicing에 대해서 집중 연구를 진행함. 오믹스 통합 분석법을 개선하여 tumor suppressor inactivation의 기능/원인을 밝히는 alternative skipping 발견. 개발된 방법으로 암종의 오믹스 데이터에 적용하여 확인함. 그림 1과 같이 수집한 암종의 오믹스 데이터 중 WES로부터 획득한 somatic point mutation과 RNA-Seq으로부터 exon의 마지막 position (5’ or 3’)에서 point mutation이 있고, abnormal splicing을 보이는 패턴을 정리하여 intron retention이 tumor suppressor inactivation의 원인임을 in-silico하게 규명

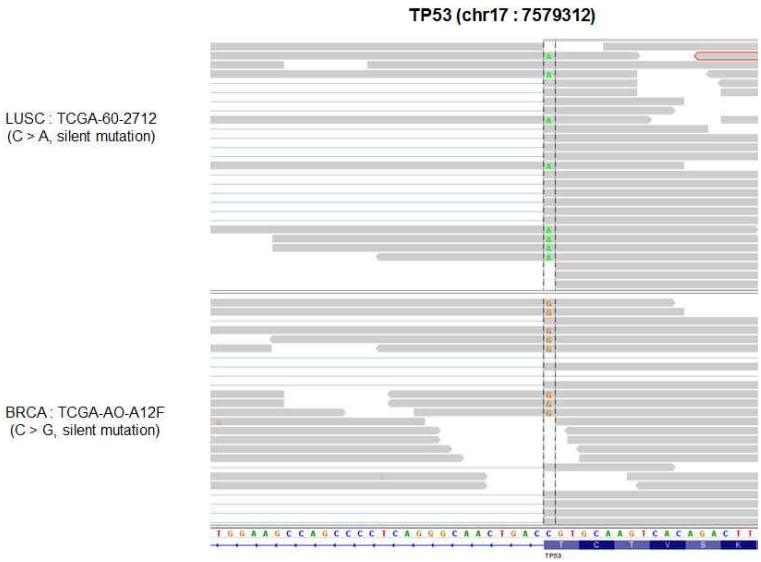


그림 1. Last Base Exon Mutation에서의 abnormal splicing 사례

⑸ 암종의 변이체 특성을 고려한 데이터 교환 표준 개발 사례

① 실제 운영을 위한 시스템을 위해 The Jackson Lab.의 Charles Lee 교수팀과의 협력 연구를 통하여 The Jackson lab.의 genomic center에서 운용하는 ELIMS 시스템의 모듈과 내부 스키마 구조를 표준화할 수 있는 형태로 수입. 그림 2, 3은 운용 가능한 내부 스키마를 보임.

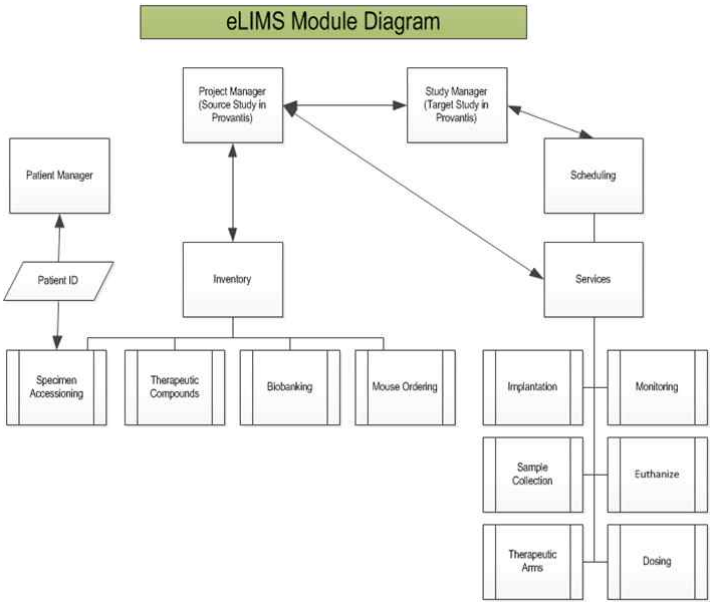


그림 2. ELIMS 시스템의 모듈 구조도

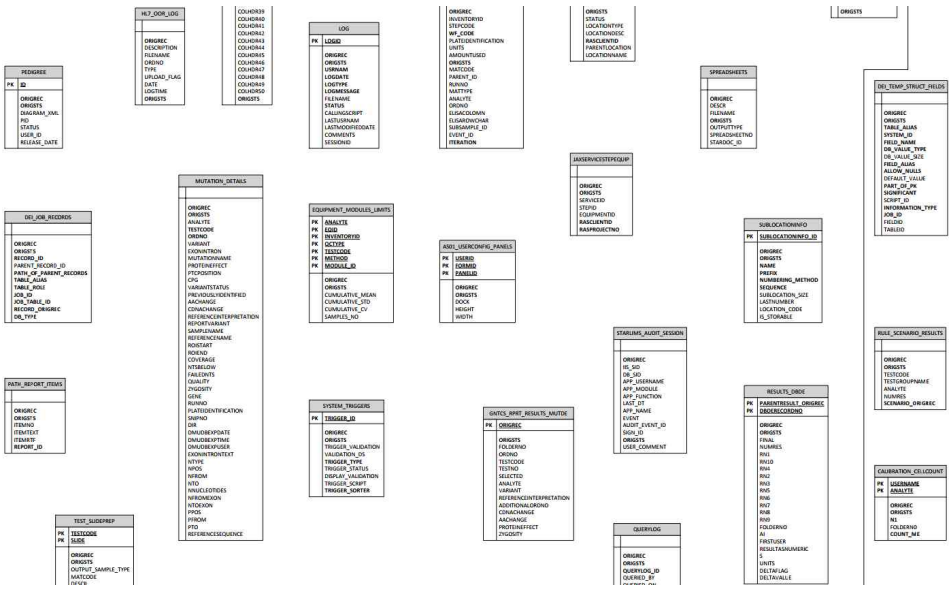


그림 3. Logical schema diagram

2. 국내외 기술개발 현황

⑴ 2009년 시작된 TCGA (The Cancer Genome Atlas) 프로젝트는 5년여에 거쳐 수행되어 암과 관련한 변이를 천만여개를 발견하였으며, 발암 유전자 500여개 등을 선정 암 환자의 진단 및 치료의 근간이 되었음 (http://cancergenome.nih.gov)

⑵ 2014년 12월 TCGA의 종료 후 NCI는 난소암, 결장직장암, 폐 선암종에 대해 강도 높은 시퀀싱을 진행하여 그 결과를 바탕으로 향후 NGS 시퀀싱의 추가 여부를 결정할 것임

⑶ 대용량 데이터의 복잡성에 대한 분석 및 관리의 어려움은 국, 내외 모두 어려움을 겪고 있는 상황임

⑷ 현재 NCI는 통합 데이터베이스를 구축할 계획이며, 그 데이터 용량이 약 20페타 바이트로 예상하고 있어 대용량 데이터베이스를 구축, 운용할 수 있는 기술이 필요함

3. 연구수행 내용 및 결과

3-1. 연구 수행 내용

- 본 연구과제에서 개선/개발한 분석 방법으로 국립암센터 연구자의 오믹스 데이터나 TCGA 데이터의 분석 후 그 결과를 연구자에게 직접 반환. 과제 종료 후에도 지속적인 연구, 협의를 통해 신규 연구를 유도하고, 연구 분야를 확장할 필요가 있음.

(1) 암종의 오믹스 데이터 분석을 위한 클라우드 컴퓨팅 환경 구축

① HDFS 설치 후 Map/Reduce 기능을 지원하는 클라우드 컴퓨팅 환경 구축

② 클라우드 컴퓨팅 환경에서 NGS 분석 파이프라인 (특히 RNA-Seq 데이터를 기본) 설치, 운영

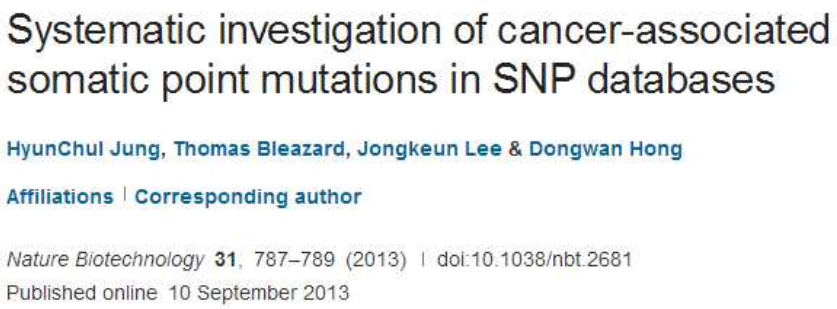
(2) 암종의 특성을 고려한 오믹스 데이터의 정확한 분석 알고리즘 개발

① 암종의 오믹스 데이터에서 somatic point mutation 탐지 과정에서 발생할 수 있는 문제점 발견

* Nature, Nature Genetics, PNAS 등 유수 저널에서 출판된 49편의 암종의 오믹스 데이터 분석 논문 조사

* 이들 논문 중 Nature Genetics, PNAS 등에 출판한 논문 중 전립선암, 방광암을 시퀀싱한 논문 중 오류가 있는 논문을 발견하였고, public database를 활용하여 필터링을 하여 somatic point mutation을 탐지하는 과정에서 문제가 발생할 수 있음을 보고

* 시제품으로 위의 문제점을 해결할 수 있는 도구 (CSTAR:Cancer genome Sequencing Tool to Acquire Reliable somatic point mutations; http://cstar-ncc.org)를 개발하여 그림 4와 같이 본 연구과제 1차년도에 **Nature biotechnology (IF: 32.44) 저널에 출판**



Systematic investigation of cancer-associated somatic point mutations in SNP databases

HyunChul Jung, Thomas Bleazard, Jongkeun Lee & Dongwan Hong

Affiliations | Corresponding author

Nature Biotechnology **31**, 787–789 (2013) | doi:10.1038/nbt.2681

Published online 10 September 2013

그림 4. 본 연구과제에 개발한 시스템이 공개된 출판논문 [1].

⑶ 비정상 스플라이싱 중 Intron retention이 암 억제 유전자의 비활성화를 시키는 원인임을 밝힘. 특히 intron retention을 발생시키는 유전자 변이를 찾아내었는데, 그 특징으로 엑손의 가장 끝자리에 위치하고 있었음. 그림 5는 본 3차년 연구과제에서 개발한 분석 방법의 결과를 **Nature Genetics (IF: 29.352) 에 출판**함 [2].

NATURE GENETICS | ANALYSIS

Intron retention is a widespread mechanism of tumor-suppressor inactivation

Hyunchul Jung, Donghoon Lee, Jongkeun Lee, Donghyun Park, Yeon Jeong Kim, Woong-Yang Park, Dongwan Hong, Peter J Park & Eunjung Lee

Affiliations | **Contributions** | **Corresponding authors**

Nature Genetics (2015) | doi:10.1038/ng.3414
Received 30 November 2014 | Accepted 08 September 2015 | Published online 05 October 2015

Citation | Reprints | Rights & permissions | Article metrics

Abstract

Abstract • References • Author information • Supplementary information

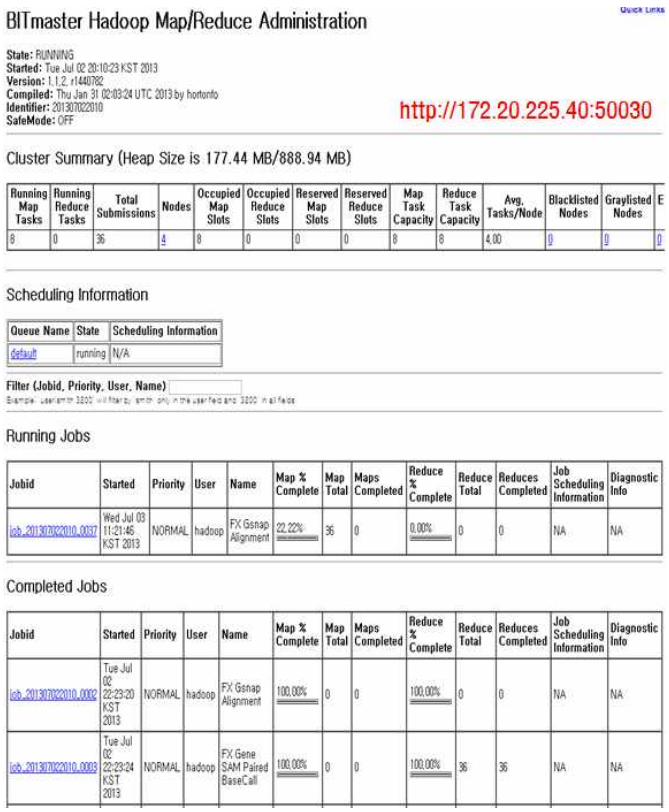
A substantial fraction of disease-causing mutations are pathogenic through aberrant splicing. Although genome profiling studies have identified somatic single-nucleotide variants (SNVs) in cancer, the extent to which these variants trigger abnormal splicing has not been systematically examined. Here we analyzed RNA sequencing and exome data from 1,812 patients with cancer and identified ~900 somatic exonic SNVs that disrupt splicing. At least 163 SNVs, including 31 synonymous ones, were shown to cause intron retention or exon skipping in an allele-specific manner, with ~70% of the SNVs occurring on the last base of exons. Notably, SNVs causing intron retention were enriched in tumor suppressors, and 97% of these SNVs generated a premature termination codon, leading to loss of function through nonsense-mediated decay or truncated protein. We also characterized the genomic features predictive of such splicing defects. Overall, this work demonstrates that intron retention is a common mechanism of tumor-suppressor inactivation.

그림 5. 본 연구과제에 발견한 변이를 공개한 출판논문

3-2. 연구 수행 결과

(1) 클라우드 컴퓨팅 환경에 운용 가능한 암종의 오믹스 데이터 분석 시스템/플랫폼 구축 결과

① 그림 6은 구축된 로컬 클라우드 환경에서 암종의 오믹스 데이터 분석 중 가장 많은 컴퓨팅 리소스와 타임 코스트가 필요한 short read (특히 1차년도에는 RNA-Seq 데이터의 분석을 주로 함)의 서열 정렬 과정을 클라우드 컴퓨팅 시스템에서 모니터링하고 있는 과정임



BITmaster Hadoop Map/Reduce Administration

Quick Links

State: RUNNING
Started: Tue Jul 02 20:10:23 KST 2013
Version: 1.1.2, r1440782
Compiled: Thu Jan 31 02:03:24 UTC 2013 by hortonfo
Identifier: 201307022010
SafeMode: OFF

http://172.20.225.40:50030

Cluster Summary (Heap Size is 177.44 MB/888.94 MB)

Running Map Tasks	Running Reduce Tasks	Total Submissions	Nodes	Occupied Map Slots	Occupied Reduce Slots	Reserved Map Slots	Reserved Reduce Slots	Map Task Capacity	Reduce Task Capacity	Avg. Tasks/Node	Blacklisted Nodes	Graylisted Nodes	E
8	0	36	4	8	0	0	0	8	8	4.00	0	0	0

Scheduling Information

Queue Name	State	Scheduling Information
default	running	N/A

Filter (Jobid, Priority, User, Name)

Running Jobs

Jobid	Started	Priority	User	Name	Map % Complete	Map Total	Maps Completed	Reduce % Complete	Reduce Total	Reduces Completed	Job Scheduling Information	Diagnostic Info
job_201307022010_0002	Wed Jul 03 11:21:45 KST 2013	NORMAL	hadoop	FX Gsnap Alignment	22.22%	36	0	0.00%	0	0	NA	NA

Completed Jobs

Jobid	Started	Priority	User	Name	Map % Complete	Map Total	Maps Completed	Reduce % Complete	Reduce Total	Reduces Completed	Job Scheduling Information	Diagnostic Info
job_201307022010_0002	Tue Jul 02 22:23:20 KST 2013	NORMAL	hadoop	FX Gsnap Alignment	100.00%	0	0	100.00%	0	0	NA	NA
job_201307022010_0003	Tue Jul 02 22:23:24 KST 2013	NORMAL	hadoop	FX Gene SAM Paired BaseCall	100.00%	0	0	100.00%	36	36	NA	NA

그림 6. 클라우드 컴퓨팅 시스템에서 헤드 노드의
Job 프로세싱의 모니터링 과정

⑵ 암종의 특성을 고려한 오믹스 데이터 분석법의 정확성을 높이기 위한 연구 수행 결과

① 분석법의 정확성을 높이기 위한 사례로 1차년도에 somatic point mutation의 탐지 시 필터링 과정에서 중요한 point mutation이 제거될 수 있는 문제점을 해결하는 사이트를 그림 7과 같이 개발하여 공개 [1].

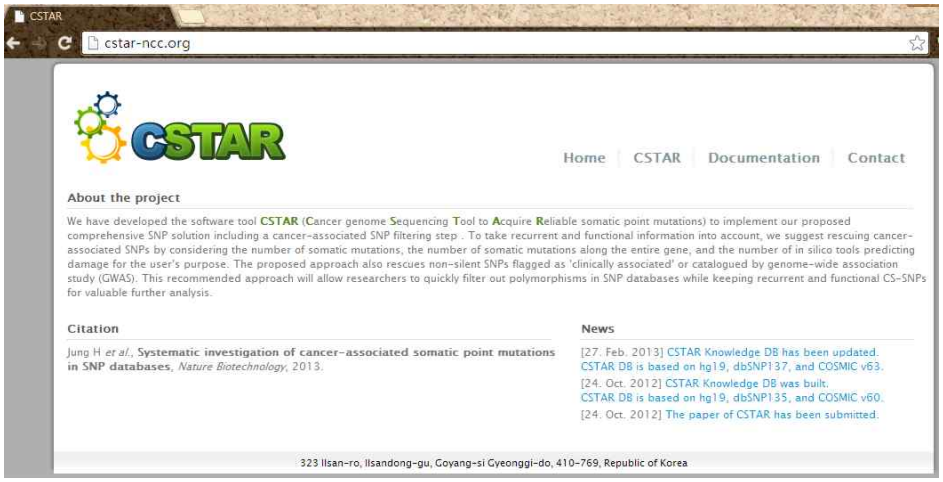


그림 7. 1차년도 개발 시스템의 사용자 인터페이스

② 1차년도 과제에서 개발한 시스템은 그림 8의 사용자 인터페이스를 이용하여 시스템을 활용할 수 있음. VCF (Variant Call Format), MAF (Mutation Annotation Format), Tab-delimited 파일을 입력으로 받아들일 수 있으며, 입력 데이터는 그림 9와 같이 Chromosome 번호와 genomic position의 필수 정보(mandatory information)와 sample name 등의 optional 정보 등으로 구성될 수 있다.

그림 8. 사용자 입력을 통한 point mutation 입력 화면

```
##fileformat=VCFv4.1
##source=TCGACRC
##reference=GRCh37
##fileDate=20130301
##INFO=<ID=GENE,Number=1,Type=String,Description="Gene name">
##INFO=<ID=STRAND,Number=1,Type=String,Description="Gene strand">
##INFO=<ID=CDS,Number=1,Type=String,Description="CDS annotation">
##INFO=<ID=AA,Number=1,Type=String,Description="Peptide annotation">
#CHROM  POS ID  REF ALT QUAL  FILTER  INFO
6 148854014 0833-10 C CT  . . GENE=SASH1;STRAND=+;CDS=c.C1646T;AA=p.P549L
4 126373784 0833-10 C CG  . . GENE=FAT4;STRAND=+;CDS=c.C11613G;AA=p.C3871W
18  29121188  0995-10 G A . . GENE=DSG2;STRAND=+;CDS=c.G1912A;AA=p.G638R
17  7577539 0995-10 G A . . GENE=TP53;STRAND=+;CDS=c.C742T;AA=p.R248W
8 19813359  0833-10 C AC  . . GENE=LPL;STRAND=+;CDS=c.C783A;AA=p.D261E
11  57185284  0833-10 C CT  . . GENE=SLC43A3;STRAND=+;CDS=c.G608A;AA=p.R203H
11  68825058  0833-10 G AG  . . GENE=TPCN2;STRAND=+;CDS=c.G442A;AA=p.G148R
4 88989191  0833-10 G AG  . . GENE=PKD2;STRAND=+;CDS=c.G2500A;AA=p.V834I
9 140510176 0833-10 G AG  . . GENE=C9orf37;STRAND=+;CDS=c.C476T;AA=p.A159V
17  10354747  0833-10 C CT  . . GENE=MYH4;STRAND=+;CDS=c.G3761A;AA=p.R1254H
11  4935994 0831-10 C AC  . . GENE=OR51G2;STRAND=+;CDS=c.G900T;AA=p.K300N
1 70505051  0831-10 G AG  . . GENE=LRRC7;STRAND=+;CDS=c.G3430A;AA=p.G1144S
17  7577121 0831-10 G AG  . . GENE=TP53;STRAND=+;CDS=c.C817T;AA=p.R273C
7 44746970  0833-10 A AG  . . GENE=OGDH;STRAND=+;CDS=c.A2779G;AA=p.I927V
20  37650519  0833-10 G AG  . . GENE=DHX35;STRAND=+;CDS=c.G1441A;AA=p.A481T
17  40998087  0833-10 C CT  . . GENE=AOC2;STRAND=+;CDS=c.C1444T;AA=p.R482W
```

그림 9. VCF 형태의 입력 파일 예

③ 그림 10은 개발 시스템에서의 실행 결과 화면으로 Cancer-associated SNP 리스트를 보임. A: 샘플 이름, B: clinically associated SNP, C: Disease susceptibility loci, D, E and E: # from preprocessed COSMIC database를 나태내고 있으며, SIFT, PolyPhen2, MutationAssessor, Phylop 등의 functional expectation 도구로부터의 결과를 함께 보임

Result-1 : Cancer-associated SNP List

A Sample Name	Gene Name	dbSNP ID	Chrom	Position	B Flagged as "clinically associated"	C Susceptibility Loci	Ref. allele	Var. allele	D # of tumor samples supporting mutation	E # of tumor samples suppoting mutation by type of variant allele	F # of mutations in gene	G # of tools predicting damage	SIFT	PolyPhen2	Mutation Assessor	Phylop
0995-10	TP53	rs121912651	17	7577539	YES		G	A	488	476	11473	3	DAMAGING	PROBABLY DAMAGING	MEDIUM	NON CONSERVED
0831-10	TP53	rs121913343	17	7577121	YES		G	A	478	456	11473	3	DAMAGING	POSSIBLY DAMAGING	MEDIUM	NON CONSERVED
0831-10	TP53	rs121913343	17	7577121	YES		G	C	478	9	11473	3	DAMAGING	PROBABLY DAMAGING	MEDIUM	NON CONSERVED
0831-10	TP53	rs28934575	17	7577548	YES		C	T	352	291	11473	3	DAMAGING	BENIGN	MEDIUM	CONSERVED
0831-10	TP53	rs28934575	17	7577548	YES		C	A	352	51	11473	3	DAMAGING	BENIGN	MEDIUM	CONSERVED
0831-10	TP53	rs28934575	17	7577548	YES		C	G	352	10	11473	4	DAMAGING	PROBABLY DAMAGING	MEDIUM	CONSERVED
0995-10	BRAF	rs113488022	7	140453136	YES		A	T	16500	16476	16898	4	DAMAGING	PROBABLY DAMAGING	LOW	CONSERVED
0995-10	BRAF	rs113488022	7	140453136	YES		A	C	16500	10	16898	4	DAMAGING	POSSIBLY DAMAGING	MEDIUM	CONSERVED
0900-09	KRAS	rs121913529	12	25398284	YES		C	T	15910	8649	25216	4	DAMAGING *WARNING! LOW CONFIDENCE.	POSSIBLY DAMAGING	MEDIUM	CONSERVED
0900-09	KRAS	rs121913529	12	25398284	YES		C	G	15910	1377	25216	4	DAMAGING *WARNING! LOW CONFIDENCE.	PROBABLY DAMAGING	MEDIUM	CONSERVED
0900-09	KRAS	rs121913529	12	25398284	YES		C	A	15910	5884	25216	4	DAMAGING *WARNING! LOW CONFIDENCE.	POSSIBLY DAMAGING	MEDIUM	CONSERVED
0900-09	KRAS	rs121913529	12	25398284	YES		C	T	15910	8649	25216	4	DAMAGING *WARNING! LOW CONFIDENCE.	POSSIBLY DAMAGING	MEDIUM	CONSERVED
0900-09	KRAS	rs121913529	12	25398284	YES		C	G	15910	1377	25216	4	DAMAGING *WARNING! LOW CONFIDENCE.	PROBABLY DAMAGING	MEDIUM	CONSERVED
0900-09	KRAS	rs121913529	12	25398284	YES		C	A	15910	5884	25216	4	DAMAGING *WARNING! LOW CONFIDENCE.	POSSIBLY DAMAGING	MEDIUM	CONSERVED
0995-10	PIK3CA	rs121913274	3	178936092	YES		A	G	168	74	4160	3	TOLERATED	PROBABLY DAMAGING	MEDIUM	CONSERVED

그림 10. 실행 결과 화면

⑶ 암종의 오믹스 데이터 특성을 고려한 통합 분석 결과

◎ 1,892명 6개 암종의 환자에서 진행한 Whole Exome Sequencing 에서 somatic point mutations 을 call한 결과와 RNA-Seq 으로부터의 splicing 상황을 통합 분석한 결과임.

◎ Exonic mutations 중 가장 끝자리에 있는 mutation (Last Base Exon Mutation; LBEM)의 경우 RNA-Seq 데이터를 확인한 결과 read가 cliping 되지 않음을 확인. 이들이 abnormal splicing의 원인이 되는지를 확인하고자 함

◎ 확인 결과 LBEM의 경우 Intron retention 등의 abnormal splicing을 일으키는 것을 확인하였고, 이 현상은 Tumor suppressor genes에서 발견됨

① 그림 11의 a, b, c, d에서 보이는 유전자 TP53, ARID1A, CDH1, TPR 등은 대표적인 암 억제 유전자로 유방암 환자의 유전자 TP53에서 Exon 4의 5’ 의 가장 끝자리 (Chr17:7579312)에 Wild type ‘C’에서 Alternate ‘G’로 C->G mutation이 있는 것을 확인할 수 있음

변이가 있는 read 들에 대해서 정상 스플라이싱이 일어나지 않고 인트론이 유지되는 abnormal splicing이 있음을 확인

Exon 4와 Exon 5 사이에 존재하는 인트론에는 “AGT”의 조기 종결 코돈 (PTC: Pre Termination Codon)이 있어 유전자 발현이 되지 않는 상황이 발생

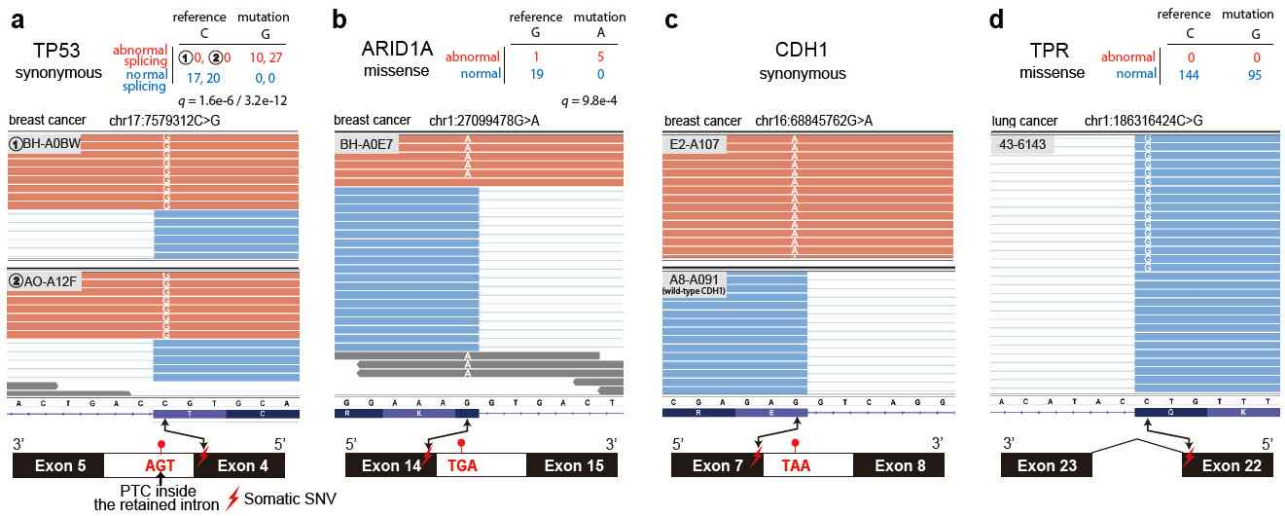


그림 11. LBEM이 비정상 스플라이싱을 일으키는 사례

② Exon의 각 자리에서 발생하는 point mutation이 비정상 스플라이싱을 일으키는 비율을 조사 (그림 12)

LBEMs이 비정상 스플라이싱을 가장 많이 일으키는 변이임을 확인

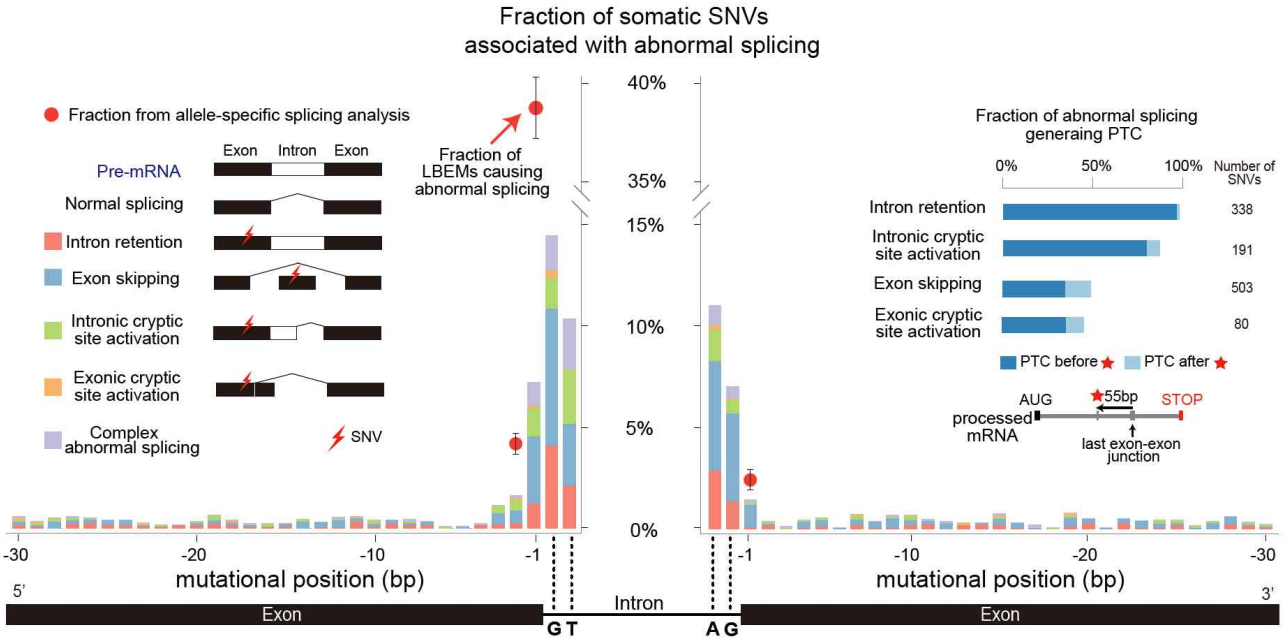


그림 12. Exon position 별 abnormal splicing 을 가진 SNV의 비율

③ Oncogene set과 Tumor suppressor gene set에서 통계적 유의성을 확인한 결과 Tumor suppressor gene에서 Intron retention의 유의도가 높음을 확인 (그림 13)

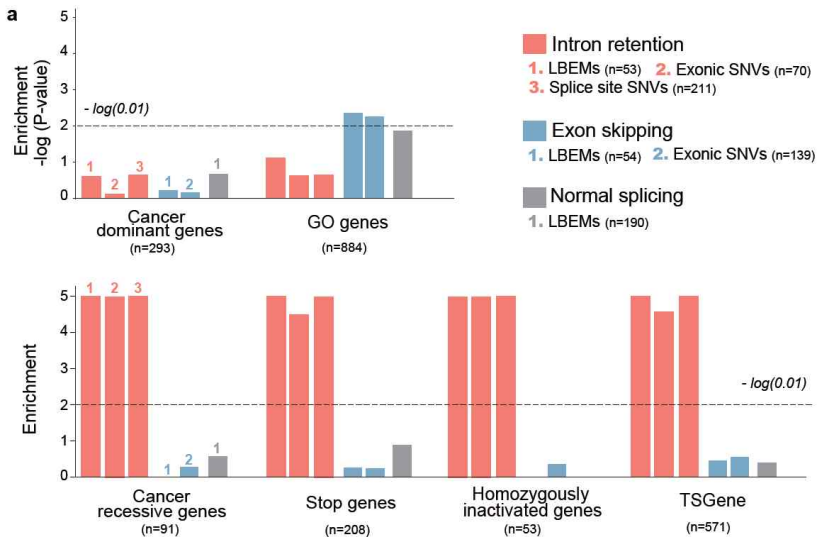


그림 13. 6개의 cancer 관련 gene sets에서의 LBEM과 intron retention 의 사례

④ 연구 결과를 “Intron retention is a widespread mechanism of tumor suppressors inactivation”의 제목으로 **Nature Genetics (IF: 29.352)에 출판**함 [2].

⑷ 데이터 추가 분석 지원 및 논문 출판 사례

※ 본 연구과제의 수행기간 동안의 논문 게재 성과 (출판 논문 I.F.: 97.44, 회색 배경: 사사논문)

논문명	저자 (저자구분)	저널명(I.F.)	Year; Vol(No):Page	구분	지원과제번호
Intron retention is a widespread mechanism of tumor suppressor inactivation [2]	홍 동 완 (교신저자)	Nature Genetics (29.648)	Online published	국외 SCI	1310190, 1410675
ALDH Inhibition Combined with Phenformin Reverses Tumor Growth by Energy Depletion in NSCLC	홍 동 완 (공동)	Cancer Cell (27.238)	submit	국외 SCI	1410675
Clinical implications of NRG1 fusion in invasive mucinous adenocarcinoma of the lung	홍 동 완 (공동)	The Pharmacogenomics Journal (5.513)	submit	국외 SCI	없음
HNF4α is a therapeutic target that links AMPK to WNT signaling in early-stage gastric cancer [3]	홍 동 완 (공동)	Gut (13.319)	2014	국외 SCI	없음
Transglutaminase 2 inhibitor Abrogated Renal Cell Carcinoma in Xenograft Model [4]	홍 동 완 (공동)	Cancer Research and Clinical Oncology (3.009)	2014; 140(5):757-67	국외 SCI	없음
Genomic profile analysis of diffuse-type gastric cancers [5]	홍 동 완 (공동)	Genome Biology (10.465)	2014; 15(4):R55	국외 SCI	없음
PATHOME: an algorithm for accurately detecting differentially expressed subpathways [6]	홍 동 완 (공동)	Oncogene (8.559)	2014; 33(41):4941-51	국외 SCI	없음
Systematic investigation of cancer-associated somatic point mutations in SNP databases [1]	홍 동 완 (교신)	Nature Biotechnology (32.44)	2013; 31(9):787-89	국외 SCI	1310190

① “HNF4α is a therapeutic target that links AMPK to WNT signaling in early-stage gastric cancer”, Gut (IF: 13.319), 2014. [3].

: 이기종의 sequencing instruments에서 생성한 시퀀싱 데이터의 분석 (그림 14)

Dataset	Ethnic Group	Comparison Group	Measurement Platform	Gene No.
TCGA (*22*)	Caucasian	29 paired tumor and non-tumor	Illumina Hiseq	27,608
Cho et al. (*21*)	Korean	65 tumor *vs.* 19 non-tumor	Illumina Human WG-6 v3.0	25,235
Kim et al. (*20*)	Korean	24 tumor *vs.* 6 non-tumor	SOLiD Single-read RNA-seq	18,890
This study	Caucasian	22 paired tumor and non-tumor	SOLiD Paired-end RNA-seq	15,987

그림 14. 시퀀싱 분석 데이터

② “Transglutaminase 2 inhibitor Abrogated Renal Cell Carcinoma in Xenograft Model”, Journal of Cancer Research and Clinical Oncology (IF: 3.009), 140(5):757-767 [4]

: Adult kidney, clear renal carcinoma patients의 expression data 분석 지원 (그림 15)

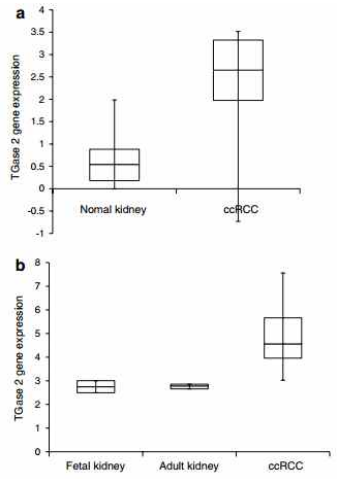


Fig. 1 Expression changes for transglutaminase 2 gene. **a** Differences in expression of transglutaminase (TGase) 2 gene in 23 normal kidney tissue samples and 23 clear cell renal cell carcinoma (ccRCC) samples (on a log2 *scale*). The statistical significance of the difference between normal kidneys and ccRCCs had a p value of 7.83e−10. **b** Distribution of the expression between normal kidney group and 26 clear cell renal cell carcinoma samples (on a log2 *scale*). The statistical significance of the difference between normal kidneys and ccRCCs had a p value of 1.71e−9

그림 15. ccRCC에서의 유전자 발현 양상

③ PATHOME: an algorithm for accurately detecting differentially expressed subpathways: Oncogene (8.559), 2014 [6].

: 위암 데이터의 transcriptome-wide expression data 처리 (그림 16)

Table 1. Summary of gastric cancer-related transcriptome-wide expression data sets used in this study

Data set	*Ethnic group*	*Comparison group*	*Profiling platform*
GSE13861[14]	Korean	65 Tumor vs 19 non-tumor	Illumina Human WG-6 v3.0
GSE15081[15]	Japanese	18 Relapse vs 38 relapse-free	Human Oligo Chip 30K
GSE36968[16]	Korean	24 Tumor vs 6 non-tumor	SOLiD Single-read RNA-seq
GSE27342[17]	Chinese	80 Tumor vs 80 non-tumor	Affymetrix Human Exon 1.0 ST array

그림 16. 암종의 오믹스 데이터에서의 발현 분석

④ “Clinical implications of NRG1 fusion in invasive mucinous adenocarcinoma of the lung”, The Pharmacogenomics Journal, submitted.

: Genome과 Transcriptome의 특성을 동시에 고려한 gene fusion 탐지 기법으로 NRG1-SLC3A2 gene fusion 탐지 (그림 17)

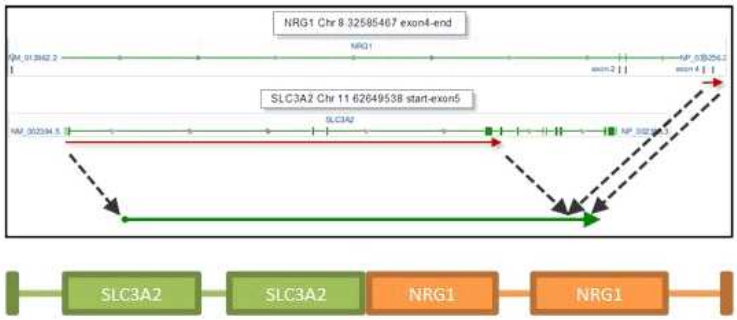


그림 17. NRG1-SLC3A2 gene fusion

⑤ “ALDH Inhibition Combined with Phenformin Reverses Tumor Growth by Energy Depletion in NSCLC”, Cancer Cell (IF: 27.238), submit.

: Omics 데이터 수집과 hierarchical gene clustering을 통하여 폐암에서의 oncometabolic target을 발굴하였고, 발굴된 물질의 약효에 대한 검증을 진행하여 논문 투고. (그림 18)

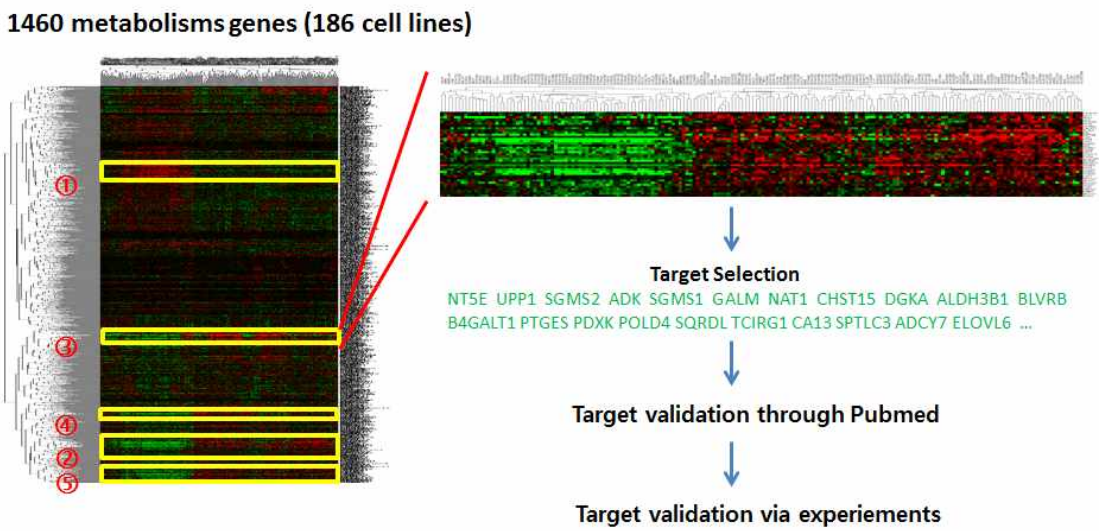


그림 18. 폐암종의 유전자 발현 양에 따른 대사 관련 유전자들의 분류와 검증 과정

⑥ 암종의 오믹스 데이터 표준 포멧 개발

● 암종의 오믹스 데이터로부터 추출한 변이체 데이터 등은 1차년도에 VCF, MAF 등의 genomic data format을 표현 범위를 정하고 XML/RDF/XML Schema 구조로 형태로 확장.
암종의 변이체 데이터는 거의 표준에 가까운 데이터를 규정/정의할 수 있으나

이를 임상에 적용하기 위해서는 탐지된 방대한 양의 변이체 중 임상 치료에 필요한 데이터를 선별하는 과정을 향후 추가 연구로 진행해야 함.

- 임상의와 유기적/지속적인 협의를 계속하고, The Jackson Lab. 의 Charles Lee 교수팀과의 협력 연구를 통하여 실제 운용 가능한 암종의 오믹스 데이터의 일부를 표준 포멧으로 개발하고 실제 운용을 가능하게 함.

4. 목표달성도 및 관련분야 기여도

4-1. 목표달성도

목 표	달성도(%)	내 용
암종의 오믹스 데이터 분석 알고리즘 개발	100%	- HDFS (Hadoop File System) 환경의 테스트 베드 구축 - TCGA 등의 공개된 암종의 시퀀싱 데이터 수집 - 암종의 변이체 특성을 고려한 필터링 시스템 개발 - 암종의 오믹스 데이터 분석 방법을 개선하여 Tumor suppressor inactivation의 원인이 되는 abnormal splicing pattern 발견 - 분석 방법 개발/개선
암종의 변이체 특성을 고려한 데이터 교환 표준 개발	100%	- 국립암센터의 오믹스 또는 변이체 데이터에 대한 데이터 표준 정의 및 확장 Patient, Genomic, Mouse schema의 interoperability의 특성을 고려한 모델링 - The Jackson Lab.의 Dr. Charles Lee와의 협업을 통해 ELIMS 시스템 포팅에 관한 연구

평가의 착안점	자 체 평 가
암종의 오믹스 데이터 분석 알고리즘 개발	- 공개된 폐암 종 등의 TCGA 오믹스 데이터 수집 진행함 - 암종의 오믹스 데이터 분석 방법을 개선하여 Tumor suppressor inactivation의 원인이 되는 abnormal splicing pattern 발견함 - RNA-Seq 데이터와 WES 데이터의 통합 분석 방법 개발 및 개선을 완료함 - Nature Genetics 에 논문 출판함
암종의 변이체 특성을 고려한 데이터 교환 표준 개발	- The Jackson Lab.의 Charles Lee 교수와의 협업을 통하여 표준 시스템 통합 개발에 관한 협력 연구를 진행함

4-2. 관련분야 기여도

(1) 정확한 분석 방법의 활용

① 본 연구과제에서는 암과 관련한 맞춤 의학에서 요구되는 정확한 분석의 필요성을 보였으며, 유전체 관련 분석 서비스를 위해 개발한 국내 시스템들이 분석 서비스에 해결 방안을 제시할 것으로 기대.

- 2011년 9월 8일 KT는 '한국인 게놈 프로젝트'에서 게놈 연구 재단과 테라젠 이텍스가 유전자 정보 분석에 클라우드 컴퓨팅을 활용해 초고속으로 분석 결과를 제공 (그림 19)

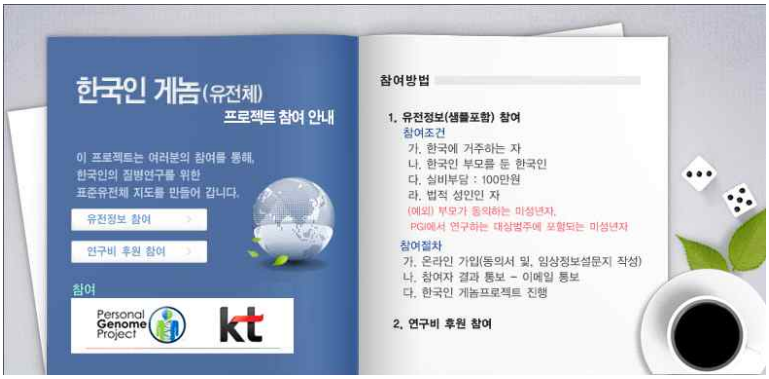


그림 19. KT와 테라젠이텍스의 클라우드 컴퓨팅을 활용한 게놈 분석

- 2011년 8월 23일 삼성 SDS는 클라우드 컴퓨팅 환경에서 NGS 데이터 분석 시스템 베타 테스팅 발표. 2011년 9월 1일부터 RNA-Seq 분석 서비스 시작. 삼성 SDS 개발팀의 자문 요청에 따라 본 기관고유연구과제의 연구 책임자는 실제 서비스 시 문제가 될 수 있는 부분에 대한 정리와 시스템 보완 방법 제공 (그림 20)

그림 20. 삼성 SDS의 자체 구축 클라우드 컴퓨팅 시스템에서의 RNA-Seq 분석 시스템

5. 연구결과의 활용계획

⑴ 클라우드 컴퓨팅 환경을 구현한 멀티 코어 프로세싱 가능 암종의 오믹스 시스템 분석 플랫폼

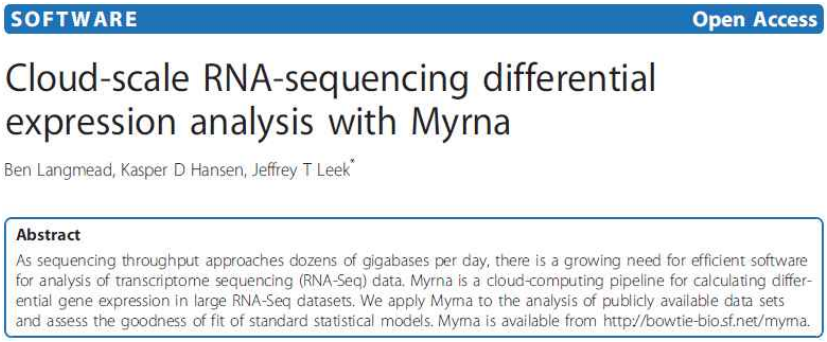
⑵ Somatic point mutation 탐지 과정 중 필터링의 문제를 해결할 수 있는 도구 CSTAR (Cancer genome Sequencing Tool to Acquire Reliable somatic point mutations; http://cstar-ncc.org) [1] 웹 사이트를 국립암센터의 Clinical Trials에 활용 적용

⑶ 인트론 유지의 이상 스플라이싱을 발생시키는 유전자 변이의 임상학적 검증을 통하여 암환자의 진단에 활용 [2]

6. 연구과정에서 수집한 해외과학기술정보

(1) 국외의 연구 사례

① 표 1에서 보였듯이 클라우드 컴퓨팅 환경에서의 오믹스 데이터 분석 시스템이 등장하고 있음. 그림 21은 죤스홉킨스의 Biostatistics과에서 연구 개발한 Myrna의 출판 논문임 (Genome Biology 2010, 11:R83). 본 기관고유연구사업의 연구 책임자는 논문 접수와 함께 2010년 Myrna의 사용 후 평가 요청을 받았음. 설치 및 활용이 극히 어렵고 single 리드만을 처리할 수 단점이 있다는 평가를 보냄



SOFTWARE Open Access

Cloud-scale RNA-sequencing differential expression analysis with Myrna

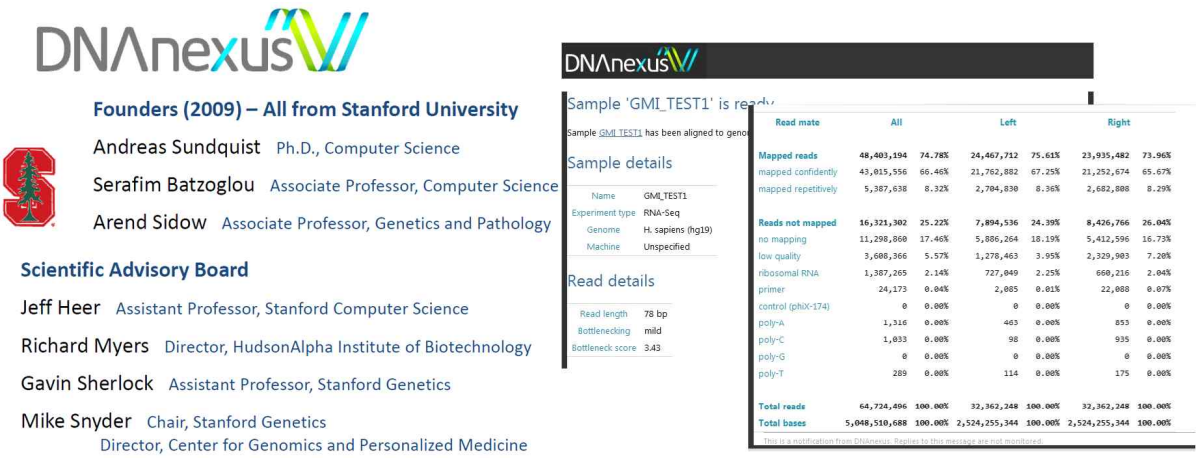
Ben Langmead, Kasper D Hansen, Jeffrey T Leek*

Abstract

As sequencing throughput approaches dozens of gigabases per day, there is a growing need for efficient software for analysis of transcriptome sequencing (RNA-Seq) data. Myrna is a cloud-computing pipeline for calculating differential gene expression in large RNA-Seq datasets. We apply Myrna to the analysis of publicly available data sets and assess the goodness of fit of standard statistical models. Myrna is available from http://bowtie-bio.sf.net/myrna.

그림 21. RNA-Seq 분석 기법을 클라우드 컴퓨팅 시스템 환경에서 구현한 예

② 그림 22의 DNAnexus는 대표적인 오믹스 데이터 분석 상업 시스템임. 스탠포드 대학의 게놈 분석을 하던 연구진들이 클라우드 컴퓨팅 환경에서 분석 시스템 개발 본격적인 서비스 시도. DNAnexus 개발팀으로부터 본 기관고유연구사업의 연구 책임자는 DNAnexus 시스템을 사용해 보고 평가해 달라는 요청을 받음. 대용량의 인프라 구축 없이 분석을 할 수 있으나 논문 출판 수준의 연구 결과를 도출하기 위해서는 분석의 정확성을 상당히 높여야 함. 현재도 분석의 정확도가 더 높아져야 함



DNAnexus

Founders (2009) – All from Stanford University

Andreas Sundquist Ph.D., Computer Science

Serafim Batzoglou Associate Professor, Computer Science

Arend Sidow Associate Professor, Genetics and Pathology

Scientific Advisory Board

Jeff Heer Assistant Professor, Stanford Computer Science

Richard Myers Director, HudsonAlpha Institute of Biotechnology

Gavin Sherlock Assistant Professor, Stanford Genetics

Mike Snyder Chair, Stanford Genetics

Director, Center for Genomics and Personalized Medicine

DNAnexus

Sample 'GMI_TEST1' is ready

Sample GMI_TEST1 has been aligned to genome

Sample details

Name	GMI_TEST1
Experiment type	RNA-Seq
Genome	H. sapiens (hg19)
Machine	Unspecified

Read details

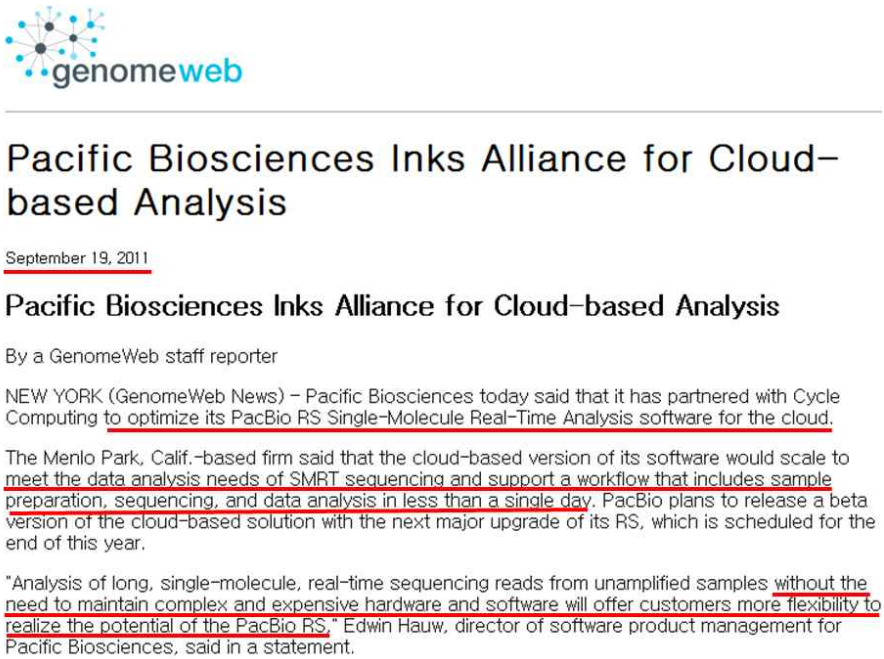
Read length	78 bp
Bottlenecking	mild
Bottleneck score	3.43

Read mate	All		Left		Right	
Mapped reads	48,403,194	74.78%	24,467,712	75.61%	23,935,482	73.96%
mapped confidently	43,015,556	66.46%	21,762,882	67.25%	21,252,674	65.67%
mapped repetitively	5,387,638	8.32%	2,704,830	8.36%	2,682,808	8.29%
Reads not mapped	16,321,302	25.22%	7,894,536	24.39%	8,426,766	26.04%
no mapping	11,298,860	17.46%	5,886,264	18.19%	5,412,596	16.73%
low quality	3,608,366	5.57%	1,278,463	3.95%	2,329,903	7.20%
ribosomal RNA	1,387,265	2.14%	727,049	2.25%	660,216	2.04%
primer	24,173	0.04%	2,085	0.01%	22,088	0.07%
control (phiX-174)	0	0.00%	0	0.00%	0	0.00%
poly-A	1,316	0.00%	463	0.00%	853	0.00%
poly-C	1,033	0.00%	98	0.00%	935	0.00%
poly-G	0	0.00%	0	0.00%	0	0.00%
poly-T	289	0.00%	114	0.00%	175	0.00%
Total reads	64,724,496	100.00%	32,362,248	100.00%	32,362,248	100.00%
Total bases	5,048,510,688	100.00%	2,524,255,344	100.00%	2,524,255,344	100.00%

This is a notification from DNAnexus. Replies to this message are not monitored.

그림 22. DNAnexus를 활용한 RNA-Seq 데이터의 분석 예

③ 3세대 시퀀싱 도구로 평가되고 있는 SMRT (Single Molecule Real Time)의 기술을 보유하고 있는 Pacific Biosciences는 UC San Diego에 위치, 샘플을 보내주면 시퀀싱 후 결과를 보내주는 형태의 서비스를 제공. 분석 결과를 클라우드 컴퓨팅 기반으로 클라이언트에게 제공해 주는 시스템 채택 (그림 23)



genomeweb

Pacific Biosciences Inks Alliance for Cloud-based Analysis

September 19, 2011

Pacific Biosciences Inks Alliance for Cloud-based Analysis

By a GenomeWeb staff reporter

NEW YORK (GenomeWeb News) – Pacific Biosciences today said that it has partnered with Cycle Computing to optimize its PacBio RS Single-Molecule Real-Time Analysis software for the cloud.

The Menlo Park, Calif.-based firm said that the cloud-based version of its software would scale to meet the data analysis needs of SMRT sequencing and support a workflow that includes sample preparation, sequencing, and data analysis in less than a single day. PacBio plans to release a beta version of the cloud-based solution with the next major upgrade of its RS, which is scheduled for the end of this year.

"Analysis of long, single-molecule, real-time sequencing reads from unamplified samples without the need to maintain complex and expensive hardware and software will offer customers more flexibility to realize the potential of the PacBio RS," Edwin Hauw, director of software product management for Pacific Biosciences, said in a statement.

그림 23. 3세대 시퀀싱 데이터의 클라우드 컴퓨팅 적용 예

7. 연구개발과제의 대표적 연구실적

번호	구분(논문/특허/기타)	논문명/특허명/기타	소속 기관명	역할	논문 게재지/특허 등록국가	Impact Factor	논문게재일/특허등록일	사사 여부 (단독사사 또는 중복사사)	특기사항 (SCI여부/인용횟수 등)
1	논문	Intron retention is a widespread mechanism of tumor suppressor inactivation	국립암센터	책임저자	미국	29.352	2015.10.05	중복사사	Nature Genetics (SCI)
2	논문	HNF4α is a therapeutic target that links AMPK to WNT signalling in early-stage gastric cancer	국립암센터	공동저자	미국	13.319	2014.11.19		Gut (SCI)
3	논문	PATHOME: an algorithm for accurately detecting differentially expressed subpathways	국립암센터	공동저자	미국	8.559	2014.10.09		Oncogene (SCI)
4	논문	Transglutaminase 2 inhibitor abrogates renal cell carcinoma in xenograft models	국립암센터	공동저자	미국	3.009	2014.05.10		Journal of Cancer Research and clinical oncology (SCI)
5	논문	Genomic profile analysis of diffuse-type gastric cancers	국립암센터	공동저자	미국	10.645	2014.04.01		Genome Biology(SCIE)
6	논문	Shape-based retrieval of CNV regions in read coverage data [8]	국립암센터	공동저자	미국	0.655	2014		International Journal of data mining and bioinformatics (SCIE)
7	논문	Systematic investigation of cancer-associated somatic point mutations in SNP databases	국립암센터	공동저자	미국	32.438	2013.09.13	중복사사	Nature Biotechnology(SCI)
8	논문	TIARA genome database: update 2013 [9]	국립암센터	제1저자	영국	4.2	2013.05.28		Database (SCI)
9	특허	유전체 단위 반복변이를 검출하는 장치 및 방법	국립암센터		대한민국		2014.01.14		
10	특허	고시폴 및 펜포르민을 유효성분으로 함유하는 암 치료용 약제학적 조성물	국립암센터		대한민국		2014.02.27		

8. 참여연구원 현황

번호	소속기관명	직위	생년월일	전공 및 학위		연구담당 분야
	성명	과학 기술인등록 번호	성별	취득 년도	학위 (전공)	과제참여 기간
	국립암센터 홍동완					

9. 기타사항

기타사항 없음

10. 참고문헌

1. Jung, H., Bleazard, T., Lee, J. & Hong, D. Systematic investigation of cancer-associated somatic point mutations in SNP databases. *Nat Biotechnol* **31**, 787-9 (2013).
2. Jung, H. *et al.* Intron retention is a widespread mechanism of tumor-suppressor inactivation. *Nat Genet* (2015).
3. Chang, H.R. *et al.* HNF4alpha is a therapeutic target that links AMPK to WNT signalling in early-stage gastric cancer. *Gut* (2014).
4. Ku, B.M. *et al.* Transglutaminase 2 inhibitor abrogates renal cell carcinoma in xenograft models. *J Cancer Res Clin Oncol* **140**, 757-67 (2014).
5. Lee, Y.S. *et al.* Genomic profile analysis of diffuse-type gastric cancers. *Genome Biol* **15**, R55 (2014).
6. Nam, S. *et al.* PATHOME: an algorithm for accurately detecting differentially expressed subpathways. *Oncogene* **33**, 4941-51 (2014).
7. Langmead, B., Hansen, K.D. & Leek, J.T. Cloud-scale RNA-sequencing differential expression analysis with Myrna. *Genome Biol* **11**, R83 (2010).
8. Hong, S. *et al.* Shape-based retrieval of CNV regions in read coverage data. *Int J Data Min Bioinform* **9**, 254-76 (2014).
9. Hong, D. *et al.* TIARA genome database: update 2013. *Database (Oxford)* **2013**, bat003 (2013).

<별첨작성 양식>

[별첨]

자체평가의견서

1. 과제현황

		과제번호		1310190	
사업구분	기관고유연구사업				
연구분야	창의 일반 연구과제			과제구분	단위
사 업 명	기관고유연구사업				주관
총괄과제	국립암센터의 데이터 교환 표준 설계와 클라우드 컴퓨팅 기반의 데이터 분석/공유 시스템 설계			총괄책임자	홍동완
과 제 명	국립암센터의 데이터 교환 표준 설계와 클라우드 컴퓨팅 기반의 데이터 분석/공유 시스템 설계			과제유형	기초
연구기관	국립암센터			연구책임자	홍동완
연구기간 연 구 비 (천원)	연차	기간	연구비	민간	계
	1차년도	2013.1.1.-2013.12.31.	50,000		
	2차년도	2014.1.1.-2014.12.31.	70,000		
	3차년도	2015.1.1.-2015.12.31.	63,000		
	계	2013.1.1.-2015.12.31	183,000		
참여기업					
상 대 국		상대국연구기관			

※ 총 연구기간이 5차년도 이상인 경우 셀을 추가하여 작성 요망

2. 평가일 : 2015.10.28

3. 평가자(과제책임자) : 홍 동 완

소속	직위	성명
종양면역학연구과	선임연구원	홍 동 완

4. 평가자(과제책임자) 확인 :

본인은 평가대상 과제에 대한 연구결과에 대하여 객관적으로 기술하였으며, 공정하게 평가하였음을 확약하며, 본 자료가 전문가 및 전문기관 평가 시에 기초자료로 활용되기를 바랍니다.

확 약	홍 동 완

Ⅰ. 연구개발실적

1. 연구개발결과의 우수성/창의성

■ 등급 : (아주우수, 우수, 보통, 미흡, 불량)

(아주우수)

- 최근 Precision Medicine에서 중요하게 여겨지고 있는 정확한 변이 탐지 과정에서 필터링 과정 중 일어날 수 있는 오류를 제거할 수 있음
- 본 연구에서 발견한 LBEM (Last Base Exon Mutation)은 abnormal splicing을 일으키고 tumor suppressor 유전자의 비활성을 야기함

2. 연구개발결과의 파급효과

■ 등급 : (아주우수, 우수, 보통, 미흡, 불량)

(아주우수)

- 본 연구에서 개발하여 공개한 시스템 (http://cstar-ncc.org)은 NGS 로부터 생산한 암 유전체 데이터의 분석에 적극 활용될 수 있어, 최근 각 대형 병원에서 실현을 위해 노력하고 있는 clinical trials에 적용되는 것을 기대할 수 있다.

3. 연구개발결과에 대한 활용가능성

■ 등급 : (아주우수, 우수, 보통, 미흡, 불량)

(아주우수)

- 본 연구에서 발굴한 LBEM 은 missense, synonymous 변이체로 기존 관심을 갖던 변이가 아닌 것으로 향후 암환자에서의 예후를 검증하는데 적용할 수 있음.

4. 연구개발 수행노력의 성실도

■ 등급 : (아주우수, 우수, 보통, 미흡, 불량)

(아주우수)

- 본 연구는 암 유전체 빅데이터를 수집하여 대용량의 자료를 분석한 것으로 수십 테라바이트 이상의 자료의 다운로드, 데이터의 오류 보정, 분석 등의 작업에 2013년, 2014년 평균 주 7일, 하루 평균 20 시간의 연구 시간을 투자하였으며, 개발 시스템을 공개한 후 현재까지 유지 보수를 진행하고 있음.

5. 공개발표된 연구개발성과(논문, 지적소유권, 발표회 개최 등)

■ 등급 : (아주우수, 우수, 보통, 미흡, 불량)

(아주우수)

본 연구의 3차년 연구 기간 중 Nature Biotechnology, Nature Genetics 등의 대표 논문에 논문을 출판하였음.

Ⅱ. 연구목표 달성도

세부연구목표 (연구계획서상의 목표)	비중 (%)	달성도 (%)	자체평가
클라우드 환경에서 운용 가능한 암종의 오믹스 데이터 분석 시스템	30	30	하둡 환경에서의 시스템 개발
암종의 특성을 고려한 오믹스 데이터의 정확한 분석 알고리즘	30	30	**개발 시스템을 Nature Biotechnology에 출판**
암종의 특성을 고려한 진단 타겟 발굴	30	30	비정상 스플라이싱으로 인한 Tumor suppressor의 비활성화 변이를 **Nature Genetics에 출판**
암종의 변이체 특성을 고려한 데이터 교환 형식	10	10	Jax Lab.에서 운영 중인 ELIMS 시스템 수입/Customizing
합계	100점	100	

Ⅲ. 종합의견

1. 연구개발결과에 대한 종합의견

본 연구과제는 최근 Precision Medicine에 필수적인 NGS 데이터의 정확한 분석법을 제시하였으며, 새로운 타겟 발굴을 성공적으로 수행하였고, 로컬의 클라우드 컴퓨팅 시스템의 개발 환경에서 과제를 수행하여 확장성에서도 용이한 분석 시스템을 개발하였음

2. 평가시 고려할 사항 또는 요구사항

본 연구과제에서는 TCGA로부터 공개된 데이터를 활용하여 분석을 진행하여 **Nature Biotechnology (2013), Nature Genetics (2015) 등의 세계 최고 권위의 journal에 논문을 책임저자로 투고하는 성과를 보였음**.

3. 연구결과의 활용방안 및 향후조치에 대한 의견

1년차 연구과제에서 개발한 시스템은 향후 NGS 분석 파이프라인에 적용가능할 것이며, 3년차 연구과제의 성과물인 암 억제 유전자를 비활성화시키는 유전자 변이는 실제 환자에서 진단 타겟으로 검증할 것을 기대함

Ⅳ. 보안성 검토

o 연구책임자의 보안성 검토의견, 연구기관 자체의 보안성 검토결과를 기재함

※ 보안성이 필요하다고 판단되는 경우 작성함.

1. 연구책임자의 의견

2. 연구기관 자체의 검토결과
