Gexp : Genemarker Expert
머신러닝 기반
멀티 클래스 분석
바이오 마커 탐지 소프트웨어

센서쟁이 팀 - 김예지, 한채은, 강서연, 이선우
발표자 - 한채은

목차

Cancer BioMarker Detection

01 What

Biomarker란?

02 Why?

왜 Gexp를
사용해야 하는가?

03 How?

Data

04 분석 시나리오

구현 예시

05 기대 효과

Biomarker
시장의 성장

What?

질병을 진단하는 센서, **Biomarker** 란?

What?

Cancer BioMarker Detection

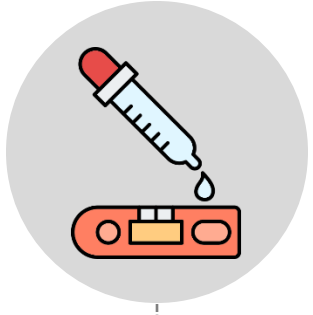
질병을 진단하는 센서, Biomarker

Biomarker : 몸 안의 변화
(병적 상태, 약물 치료 반응성, 질병의 진행)를
알아 낼 수 있는 생물학적 지표

What?

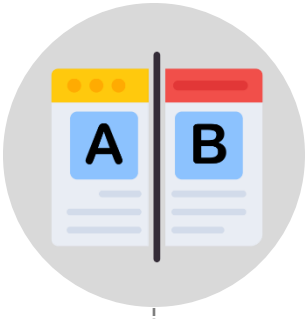
Cancer BioMarker Detection

Biomarker 탐지의 사용



진단 바이오 마커

병의 유무를 진단



예상 바이오 마커

특정 약물에 대한
반응군과 비반응군을 구별

독성 바이오 마커

특정약물에 대한 부작용을
나타낸 그룹 찾아냄

대리표지자 바이오 마커

약물 치료효과를 모니터링

예후 바이오 마커

질병의 예후를 알려줌

What?

Cancer BioMarker Detection

유전자 Biomarker란?

RNA 코로나 진단 키트

- 진단에 사용되는 유전자 바이오 마커

  : 엔벨로프(env), 뉴클레오 캡시드(N), 스파이크(S), ORF1 코딩 유전자의 조합

Oncotype DX : 유방암 예후진단

- 21개의 바이오 마커 유전자로 유방암 재발 스코어 산출
- 세계 주요 가이드 라인에 포함된 키트

Why?

왜 **Gexp**를 사용해야 하는가?

Why?

Cancer BioMarker Detection

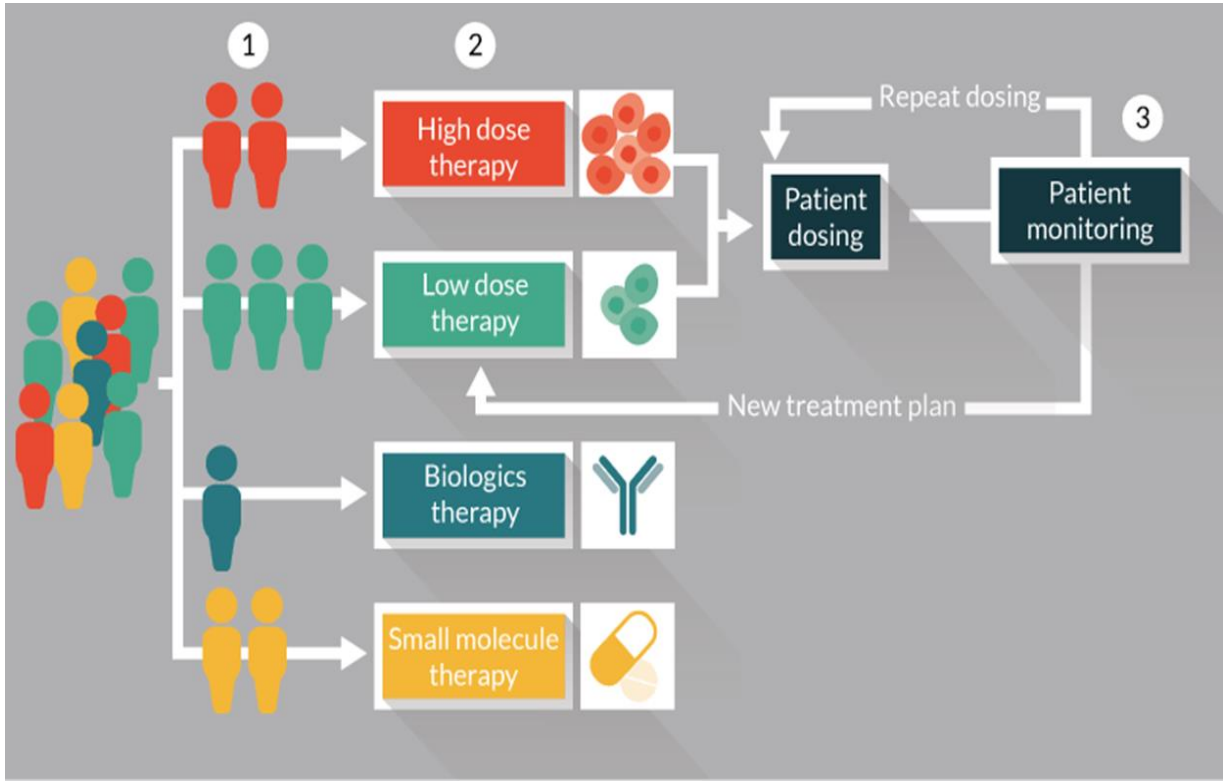
왜 gexp일까?

❖ 다중 분류(Multiclass)의 필요성

현재 → 질병 유무 판단

미래 → 질병 세부 분류 판단

맞춤의료 치료

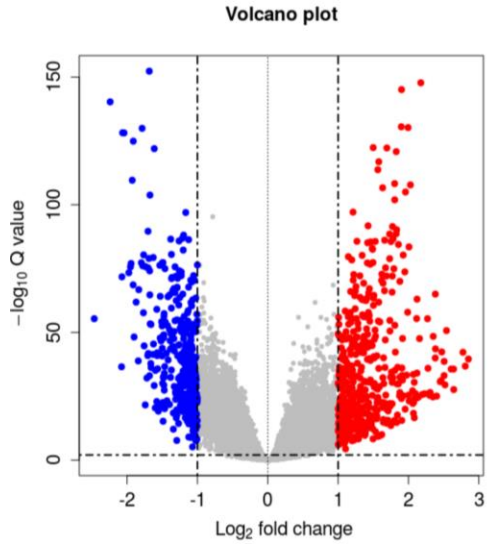


Why?

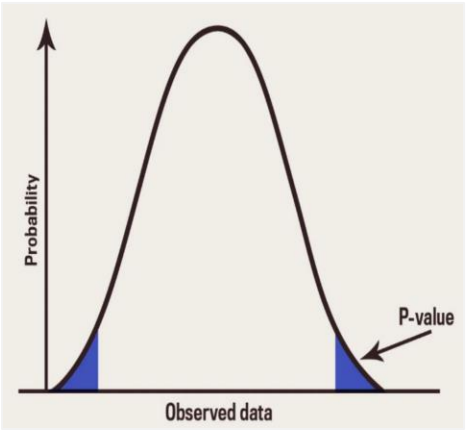
Cancer BioMarker Detection

왜 **gexp**일까?

❖ **비선형성(non-linearity) 고려 필요성**

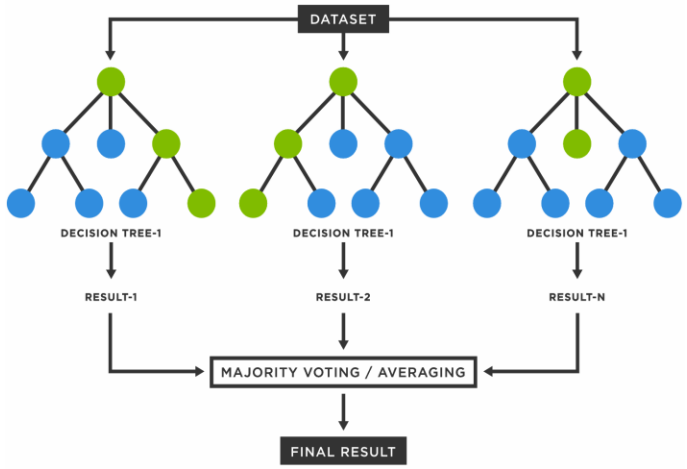


Deseq2



다중검정보정(p-value)

유전자 독립적 분석, 이진 분류(binary class)



유전자 간 관계 고려,
Multi class 분석 가능한

Tree ensemble 기반 머신 러닝 사용

Why?

Cancer BioMarker Detection

왜 gexp일까?

❖ 사용성을 높인 구현 구조

1. 편의성

- 각 기능을 함수 모듈로 구현

  1) load_labeled_data

  2) biomarker_rank

  3) plot_stepwise_accuracy

  4) desecribe_genelist

  5) plot_heatmap

2. 유연성

- 함수 옵션을 통한 커스터마이징 지원

예) 데이터 라벨링에서 37종의 암 선택 로드

```
cancer_list : list , len(list) <= 37
  ["ACC", "BLCA", "BRCA", "CESE", "CHOL", "COAD", "COADREAD", "DLBC", "ESCA",
"GBM", "GBMLGG", "HNSC", "KICH", "KIPAN", "KIRC", "KIRP", "LAML", "LGG", "LIHC",
"LUAD", "LUSC", "MESO", "OV", "PAAD", "PCPG", "PRAD", "READ", "SARC", "SKCM",
"STAD", "STES", "TGCT", "THCA", "THYM", "UCEC", "UCS", "UVM"]
  원하는 TCGA data 암 종류를 리스트 형식의 Argument로 입력
```

예) 5가지 method 별 중요도 유전자 선택

```
Model : {'RF, 'Ada', 'EXtra', 'DT', 'XGB'}
    sklearn.ensemble.RandomForestClassifier
    sklearn.ensemble.AdaBoostClassifier
    sklearn.ensemble.ExtraTreesClassifier
    sklearn.tree.DecisionTreeClassifier
    xgboost.XGBClassifier
```

예) 다양한 정확도 메트릭스 선택 가능

```
accuracy_metric : list , default = ['accuracy']
  ['f1', 'accuracy', 'precision', 'recall', 'roc', 'aic', 'bic']
  성능 평가의 지표 (sklearn.metrics)
    multi_class = True
      ['f1', 'accuracy', 'precision', 'recall', 'roc'], average='macro' 사용
    multi_class = None
      ['f1', 'accuracy', 'precision', 'recall', 'roc', 'aic', 'bic'], average='binary'을 사용
```

How?
Gexp 프로세스 및 구현

01. TCGA 데이터
02. Gexp 구현

TCGA Data

Cancer BioMarker Detection

TCGA data

FIREBROWSE

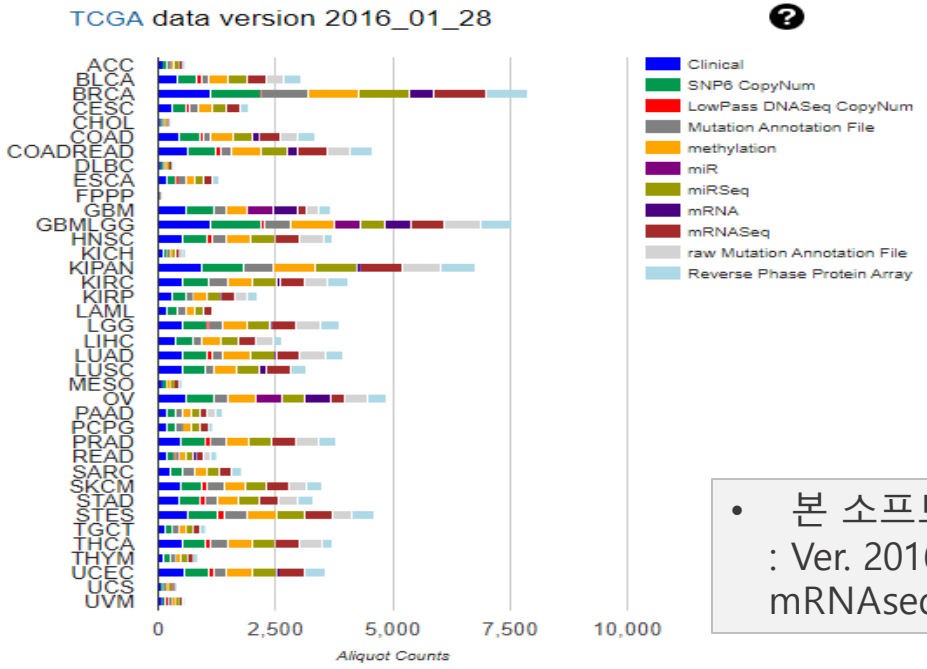
Search analysis results

HOME | BROAD GDAC | WEB API | FAQ | SAMPLES REPORT | AWG RESULTS | OLD RUNS | TUTORIAL | RELEASE NOTES | CONTACT

View Expression Profile | Enter gene name | Enter cohort abbrev | View Analysis Profile

SELECT COHORT

- Clinical Analyses
- CopyNumber Analyses
- Correlations Analyses
- miR Analyses
- miRseq Analyses
- mRNA Analyses
- mRNAseq Analyses
- Mutation Analyses
- Pathway Analyses
- RPPA Analyses



- 본 소프트웨어의 사용 data
  : Ver. 2016/01/28
  mRNAseq - gene_normalized

TCGA Dataset

Cancer BioMarker Detection

TCGA

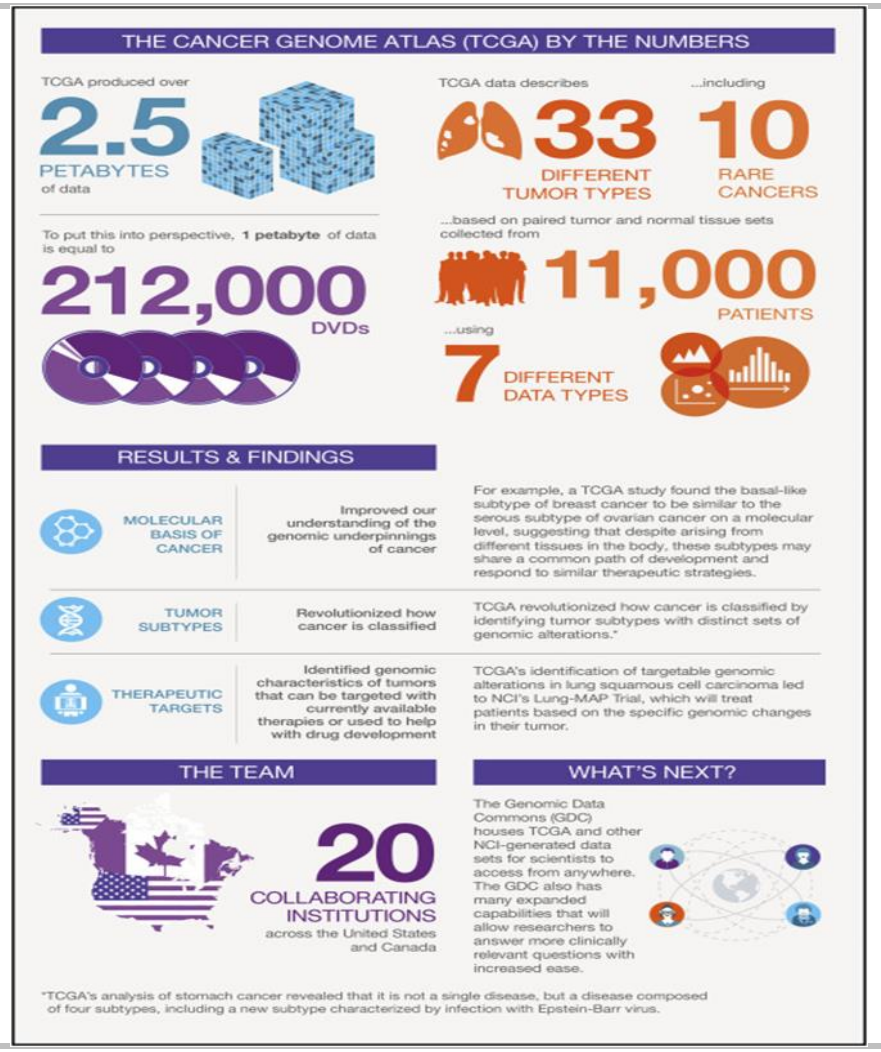
: The Cancer Genome Atlas

Since 2006 (NCI & National Human Genome Research Institute 주관)

- 미국 암 환자의 유전체 빅데이터 아카이브 프로젝트
- 전 유전체 종류 중 RNA기반 발현량 데이터 사용
  (임상(clinical), DNA, RNA, 단백질, Methylation, microRNA...)

- 33 종류 암 / 11,000명 환자

20,000개의 암 유전 분석을 하기 위해
발현 중요도가 높은 유전자를 선택하는 것이 핵심!!



TCGA – BRCA subtype

❖ 유방암 환자의 아종 라벨 데이터

- 유방암의 4개 아종 분류
  (LumA, LumB, Her2, Basal)

2019 | 해외 | 공개 | English

TCGA Breast cancer selection dataset

TCGA Breast cancer selection dataset
RICHARD TJÖRNHAMMAR;

MS and TCGA Breast cancer selection dataset

dataon의 TCGA Breast cancer selection dataset 사용
> complete_data_gen_tcga.zip
> NIHMS393293-supplement-2.xls

U
PAM50 mRNA
Basal-like
Basal-like
Basal-like
Basal-like
Basal-like
Basal-like
HER2-enriched
HER2-enriched
HER2-enriched
HER2-enriched
HER2-enriched
HER2-enriched
Luminal A
Luminal A
Luminal A
Luminal A
Luminal A
Luminal A
Luminal B
Luminal B
Luminal B
Luminal B
Luminal B
Luminal B

BRCApatients_type.csv

Hybridization REF	Tumor	Subtype
TCGA-A8-A07B-01A-1...	BRCA	LumA
TCGA-A8-A08B-01A-1...	BRCA	Her2
TCGA-A8-A08P-01A-1...	BRCA	LumB
TCGA-A8-A094-01A-1...	BRCA	Her2
TCGA-A8-A09T-01A-1...	BRCA	LumA
TCGA-AO-A03O-01A-...	BRCA	LumB
TCGA-BH-A0DZ-01A-...	BRCA	LumA
TCGA-A2-A04P-01A-3...	BRCA	Basal
TCGA-A2-A04Q-01A-...	BRCA	Basal
TCGA-A2-A04T-01A-2...	BRCA	Basal

국가연구데이터플랫폼 : DataON (kisti.re.kr)

How?
Gexp 프로세스 및 구현

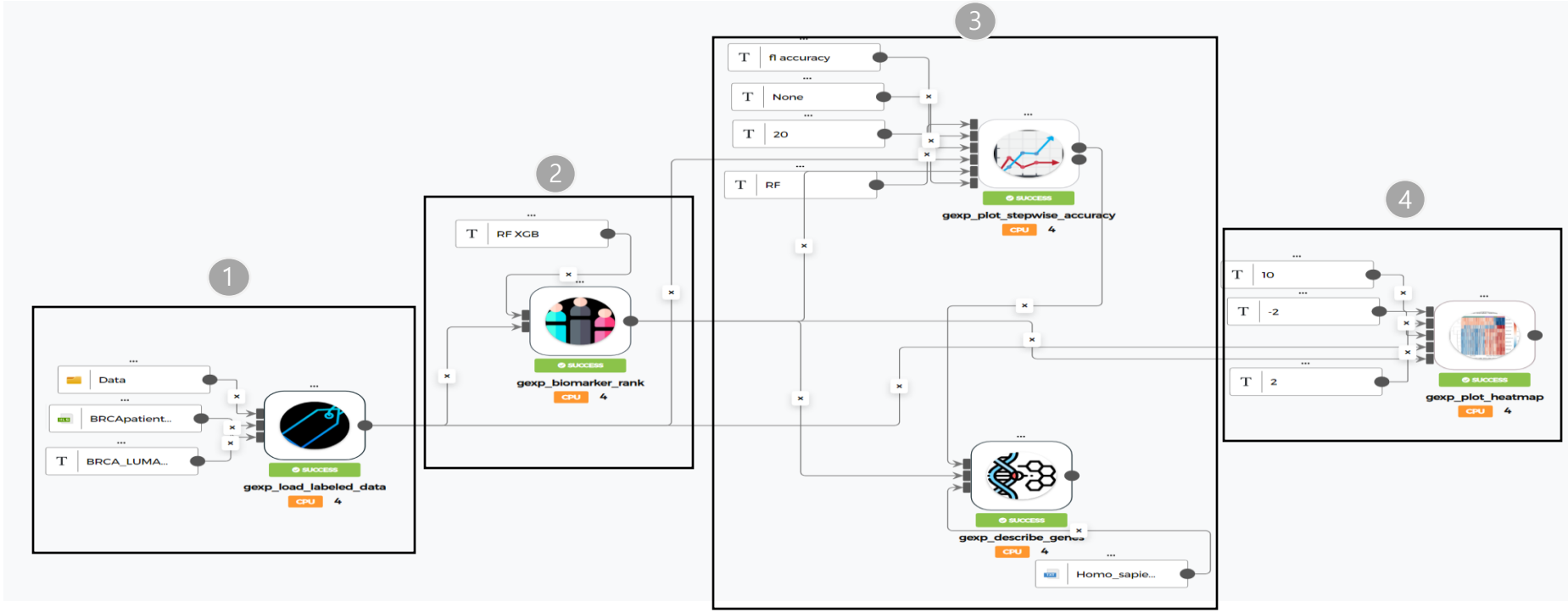
01. TCGA 데이터
02. Gexp 구현

Gexp 실행 환경

Cancer BioMarker Detection

Dataon Workflow 환경

1 TGCA 데이터 로드 및 라벨링 → 2 Gene marker 탐지 → 3 Step-wise accuracy → 4 Biomarker 선택 및 시각화



지금 바로! dataon CANVAS Beta

Gexp 실행 환경

Cancer BioMarker Detection

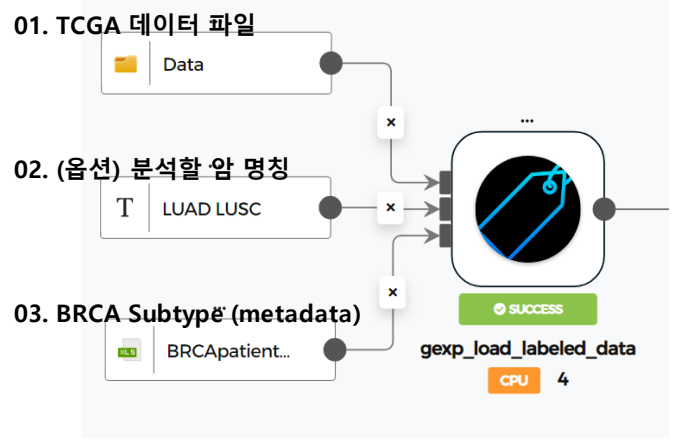
TGCA 데이터 로드 및 라벨링 → Gene marker 탐지 → Step-wise accuracy → Biomarker 선택 및 시각화

1. load_labeled_data

사용자가 선택한 암에 대하여 해당 데이터를 정상 환자를 제거한 암 환자의 유전체 데이터만 불러와 암 이름을 라벨링 한 하나의 프레임으로 만든다.

입력 예시

LUAD.rnaseqv2__illuminahiseq_rnaseqv2__unc_edu__Level_3__RSEM_genes_normalized__data.data.txt
LUSC.rnaseqv2__illuminahiseq_rnaseqv2__unc_edu__Level_3__RSEM_genes_normalized__data.data.txt



⇒

출력 예시
- cancer_data.csv

ZZEF1|23140	ZZZ3|26009	psiTPTE22|387590	tAKR|389932		Target
1989.3074	611.7619	757.3737	1.2014		LUAD
1424.1557	426.7888	5.4951	0.0		LUAD
1520.5742	477.9904	186.1244	0.4785		LUAD
1582.9518	509.7419	318.4249	0.0		LUAD
265.3543	738.5827	7.0866	0.0		LUSC
702.6641	960.9895	11.4177	0.4757		LUSC
589.6485	757.5263	5.9321	18.3894		LUSC
807.1795	1008.7179	48.2051	0.0		LUSC

Gexp 실행 환경

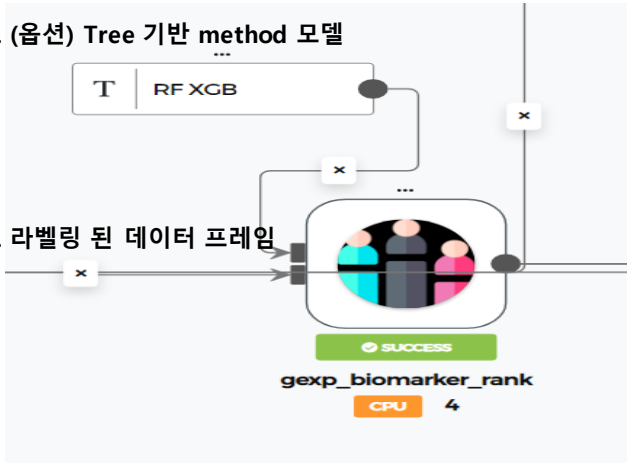
Cancer BioMarker Detection

TGCA 데이터 로드 및 라벨링 → **Gene marker 탐지** → Step-wise accuracy → Biomarker 선택 및 시각화

2. biomarker_rank

트리 기반 5가지 method을 이용해 cancer 별, subtype 별로
중요 유전자(biomarker)를 한눈에 순위화한 랭킹 프레임을 얻는다.

입력 예시



사용 가능 method 종류

- RandomForestClassifier
- XGBoostClassifier
- AdaBoostClassifier
- DecisionTreeClassifier
- ExtraTreesClassifier

출력 예시
- rank.csv

Rank

		RF	XGB
1	KRT5|3852	1	2
2	PVRL1|5818	2	150
3	KRT31|3881	3	152
4	C3orf21|152002	4	152
5	ARHGEF38|54848	5	152
6	TMEM40|55287	6	152
7	SFTA2|389376	7	152
8	CAPN8|388743	8	152
9	YEATS2|55689	9	152
10	LASS3|204219	10	4

Gexp 실행 환경

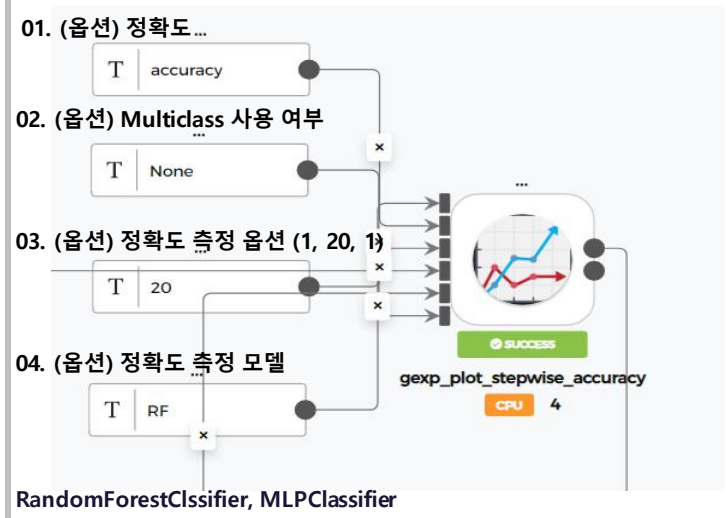
Cancer BioMarker Detection

TGCA 데이터 로드 및 라벨링 → Gene marker 탐지 → **Step-wise accuracy** → Biomarker 선택 및 시각화

3. plot_stepwise_accuracy

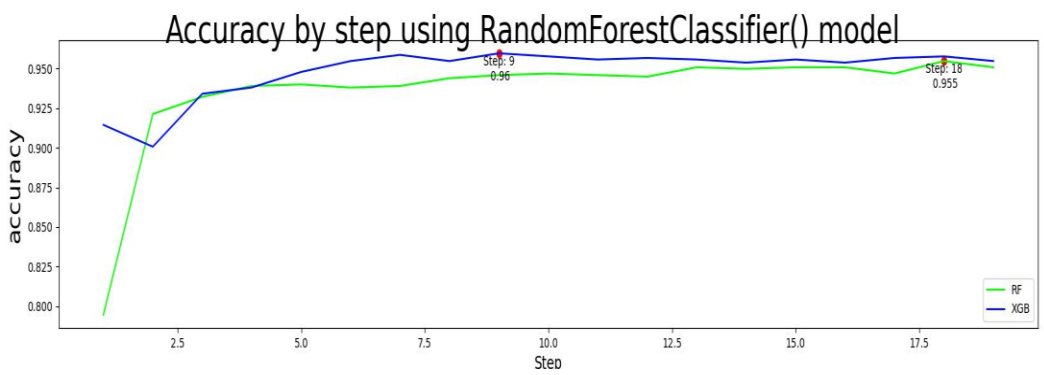
step_num의 값에 따라서 단계별로 **상위 N개 유전자**를 바이오 마커로 선택했을 때, 정확도 성능을 method 별로 비교해 시각화 한다.

입력 예시



출력 예시

- accuracy.jpeg, score.csv



	RF_accuracy	XGB_accuracy
1	0.7945281560900221	0.9145223606684052
2	0.921365787694388	0.9007002801120448
3	0.9321887375639912	0.9341591809137448
4	0.9390756302521008	0.9381097266492805
5	0.9400511928909495	0.9479329662899645
6	0.9380759200231819	0.9548005409060176
7	0.9390708007340868	0.9587269390514827
8	0.9439727615184005	0.9547957113880035
9	0.9459480343861684	0.9597073312083454
10	0.9469332560610452	0.9577513764126341
11	0.9459480343861684	0.9557761035448662
12	0.9449773012653339	0.9567613252197431
13	0.9508693132425385	0.9557857625808943

Biomarker 선택 수

Gexp 실행 환경

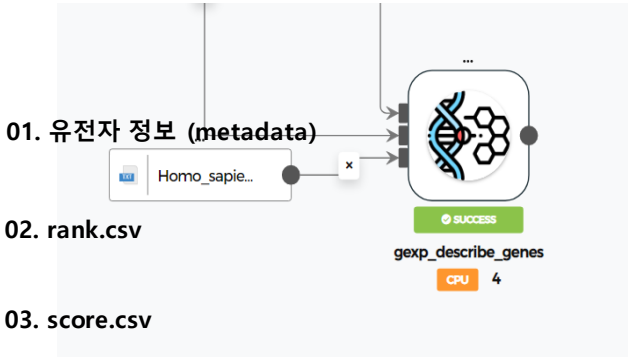
Cancer BioMarker Detection

TGCA 데이터 로드 및 라벨링 → Gene marker 탐지 → **Step-wise accuracy** → Biomarker 선택 및 시각화

4. desecribe_genelist

가장 성능이 좋은 method 방법으로 상위 유전자로 선택된 바이오 마커로
유전자 개념을 쉽게 이해 가능한 gene cards 연결 데이터 프레임 출력

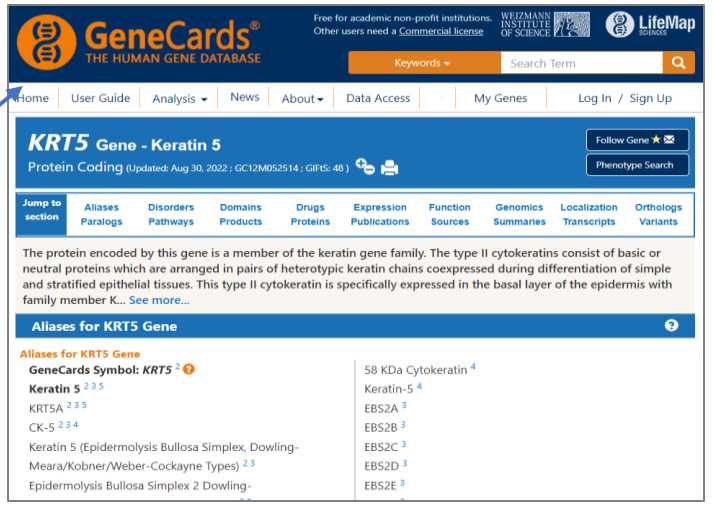
입력 예시



출력 예시
- **gene_description.csv**

유전자ID 대표 타입 위치 유전자 설명 유전자 정보 링크

	GeneID	Symbol	type_of_gene	map_location	description	links
0	54848	ARHGEF38	protein-coding	4q24	Rho guanine nucleotide exchange factor 38	https://www.genecards.org/cgi-bin/carddisp.pl?gene=54848
1	408	ARRB1	protein-coding	11q13.4	arrestin beta 1	https://www.genecards.org/cgi-bin/carddisp.pl?gene=408
2	53637	S1PR5	protein-coding	19p13.2	sphingosine-1-phosphate receptor 5	https://www.genecards.org/cgi-bin/carddisp.pl?gene=53637
3	121391	KRT74	protein-coding	12q13.13	keratin 74	https://www.genecards.org/cgi-bin/carddisp.pl?gene=121391
4	646	BNC1	protein-coding	15q25.2	basonuclin 1	https://www.genecards.org/cgi-bin/carddisp.pl?gene=646
5	339967	TMPRSS11A	protein-coding	4q13.2	transmembrane serine protease 11A	https://www.genecards.org/cgi-bin/carddisp.pl?gene=339967
6	348825	TPRXL	pseudo	3p25.1	tetrapeptide repeat homeobox like (pseudogene)	https://www.genecards.org/cgi-bin/carddisp.pl?gene=348825
7	4680	CEACAM6	protein-coding	19q13.2	CEA cell adhesion molecule 6	https://www.genecards.org/cgi-bin/carddisp.pl?gene=4680
8	221	ALDH3B1	protein-coding	11q13.2	aldehyde dehydrogenase 3 family member B1	https://www.genecards.org/cgi-bin/carddisp.pl?gene=221



Gexp 실행 환경

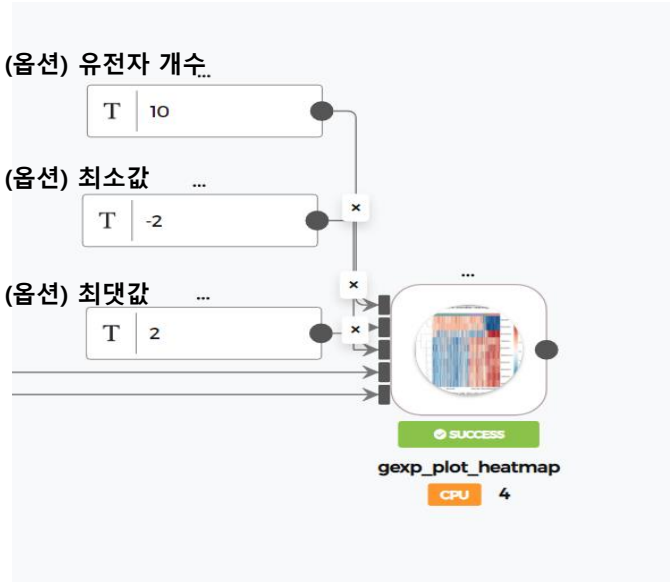
Cancer BioMarker Detection

TGCA 데이터 로드 및 라벨링 → Gene marker 탐지 → Step-wise accuracy → **Biomarker 선택 및 시각화**

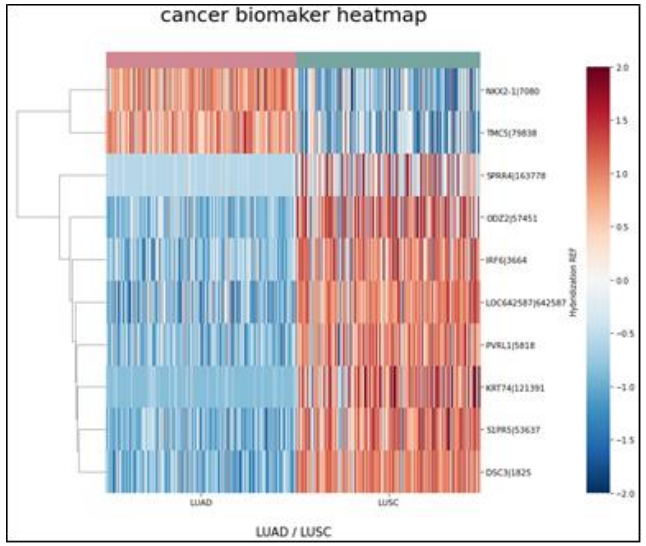
5. plot_heatmap

선택한 바이오 마커 유전자의 유전자 발현량을 시각화 합니다.

입력 예시



출력 예시
- heatmap.jpeg



선택된 바이오 마커 유전자

암 서브타입

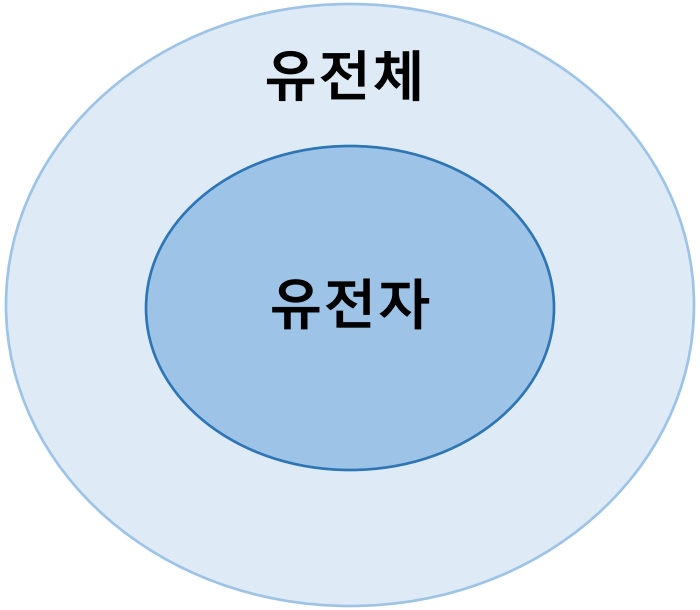
기대효과

Biomarker 시장의 동향과 기대효과

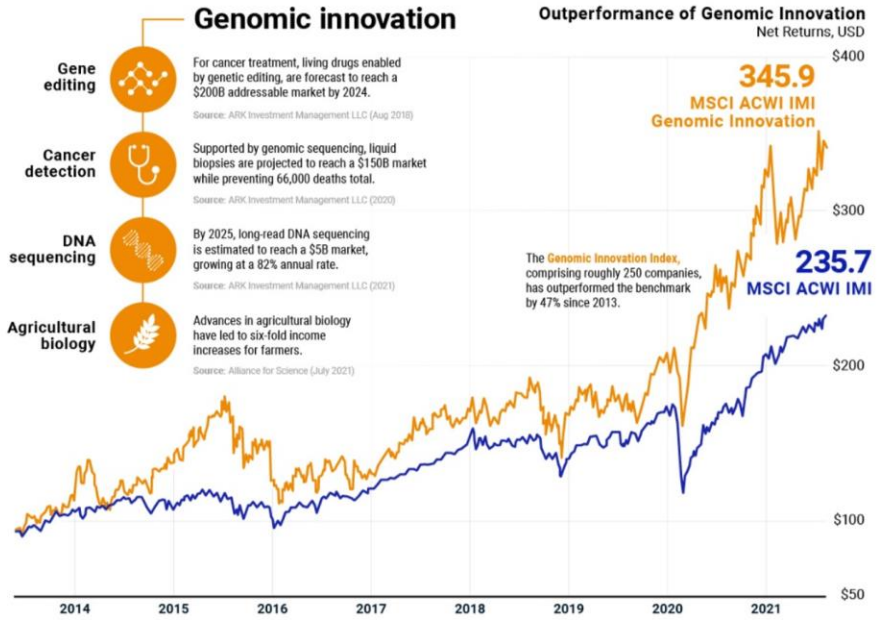
기대 효과

Cancer BioMarker Detection

❖ 차세대 유전자 분석 NGS(Next Generation Sequencing)



세포 내 분자 수준으로 더 정밀한 관측 가능!



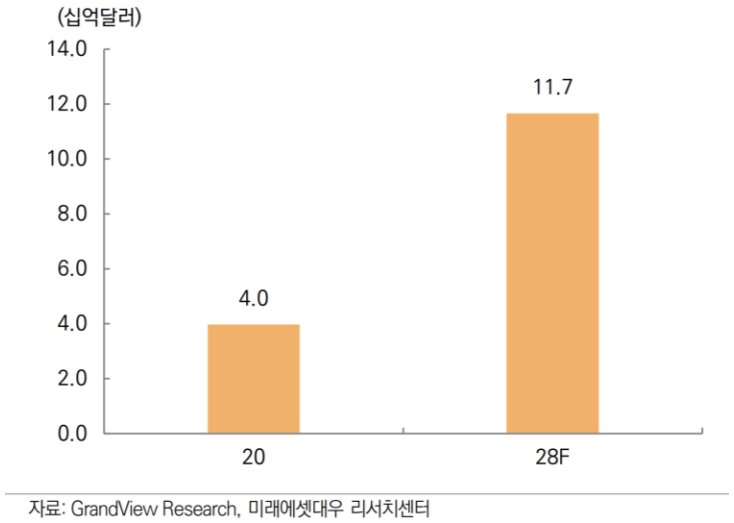
Source : MSCI(Aug 2021) Why Genomics is Poised for Growth

유전체 빅데이터, 연구는
기하급수적으로 증가

기대 효과

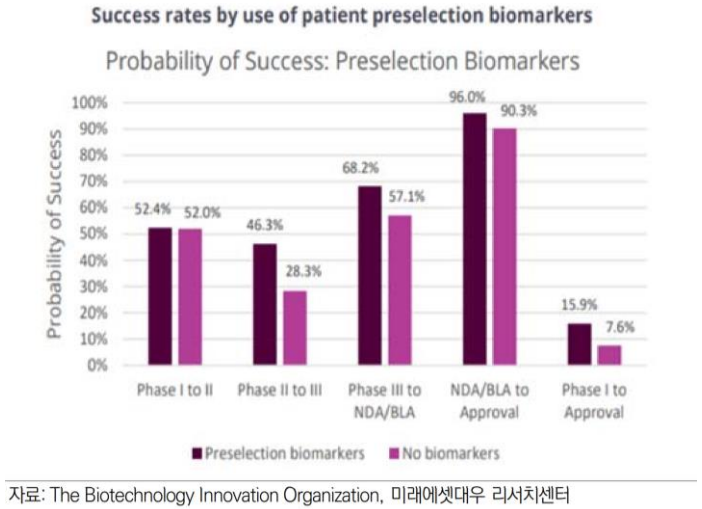
Cancer BioMarker Detection

Biomarker를 사용한 발전 전망



자료: GrandView Research, 미래에셋대우 리서치센터

유전체 분석시장 규모 추이



자료: The Biotechnology Innovation Organization, 미래에셋대우 리서치센터

바이오마커 활용 시 임상시험 성공률

Biomarker의 전망 증가, 활용도 증가


Q
A

감사합니다.