KSBi-BIML 2024

Bioinformatics & Machine Learning(BIML) Workshop for Life and Medical Scientists

생명정보학 & 머신러닝 워크샵 (온라인)

Introduction to ConnectivityMap

전민지 _ 고려대학교

KSBI KOREAN SOCIETY FOR BIOINFORMATICS | 한국생명정보학회

본 강의 자료는 한국생명정보학회가 주관하는 BIML 2024 워크샵 온라인 수업을 목적으로 제작된 것으로 해당 목적 이외의 다른 용도로 사용할 수 없음을 분명하게 알립니다.

이를 다른 사람과 공유하거나 복제, 배포, 전송할 수 없으며 만약 이러한 사항을 위반할 경우 발생하는 **모든 법적 책임은 전적으로 불법 행위자 본인에게 있음을 경고**합니다.

KSBi-BIML 2024

Bioinformatics & Machine Learning(BIML)
Workshop for Life and Medical Scientists

안녕하십니까?

한국생명정보학회가 개최하는 동계 교육 워크샵인 BIML-2024에 여러분을 초대합니다. 생명정보학 분야의 연구자들에게 최신 동향의 데이터 분석기술을 이론과 실습을 겸비해 전달하고자 도입한 전문 교육 프로그램인 BIML 워크샵은 2015년에 시작하여 올해로 벌써 10년 차를 맞이하게 되었습니다. BIML 워크샵은 국내 생명정보학 분야의 최초이자 최고 수준의 교육프로그램으로 크게 인공지능과 생명정보분석 두 개의 분야로 구성되어 있습니다. 올해 인공지능 분야에서는 최근 생명정보 분석에서도 응용이 확대되고 있는 다양한 인공지능 기반 자료모델링 기법들에 대한 현장 강의가 진행될 예정이며, 관련하여 심층학습을 이용한 단백질구조예측, 유전체분석, 신약개발에 대한 이론과 실습 강의가 함께 제공될 예정입니다. 또한 단일세포오믹스, 공간오믹스, 메타오믹스, 그리고 롱리드염기서열 자료 분석에 대한 현장 강의는 많은 연구자의 연구 수월성 확보에 큰 도움을 줄 것으로 기대하고 있습니다.

올해 BIML의 가장 큰 변화는 최근 연구 수요가 급증하고 있는 의료정보자료 분석에 대한 현장 강의를 추가하였다는 것입니다. 특히 의료정보자료 분석을 많이 수행하시는 의과학자 및 의료정보 연구자들께서 본 강좌를 통해 많은 도움을 받으실 수 있기를 기대하고 있습니다. 또한 다양한 생명정보학 분야에 대한 온라인 강좌 프로그램도 점차 증가하고 있는 생명정보 분석기술의 다양화에 발맞추기 위해 작년과 비교해 5강좌 이상을 신규로 추가했습니다. 올해는 무료 강좌 5개를 포함하여 35개 이상의 온라인 강좌가 개설되어 제공되며, 연구 주제에 따른 연관된 강좌 추천 및 강연료 할인 프로그램도 제공되며, 온라인을 통한 Q&A 세션도 마련될 예정입니다. BIML-2024는 국내 주요 연구중심 대학의 전임 교원이자 각 분야 최고 전문가들의 강의로 구성되었기에 해당 분야의 기초부터 최신 연구 동향까지 포함하는 수준 높은 내용의 강의가 될 것이라 확신합니다.

BIML-2024을 준비하기까지 너무나 많은 수고를 해주신 운영위원회의 정성원, 우현구, 백대현, 김태민, 김준일, 김상우, 장혜식, 박종은 교수님과 KOBIC 이병욱 박사님께 커다란 감사를 드립니다. 마지막으로 부족한 시간에도 불구하고 강의 부탁을 흔쾌히 허락하시고 훌륭한 현장 강의와 온라인 강의를 준비하시는데 노고를 아끼지 않으신 모든 강사분들께 깊은 감사를 드립니다.

2024년 2월

한국생명정보학회장 이 인 석

강의개요

Introduction to ConnectivityMap

ConnectivityMap 플랫폼은 3만 가지의 약물에 대한 300만 개 유전자 발현 프로파일을 포함하고 있어, 연구자들이 약물과 질병, 유전자 간의 복잡한 상호작용을 이해할 수 있게 도와주고 있다. 연구자들은 ConnectivityMap을 사용하여 기존 약물이 새로운 질병에 대해 어떤 효과를 보일 수 있는지 예측할 수 있으며, 이는 약물 개발 과정을 가속화하고 비용을 절감하는 데 큰 도움이 되고있다.

ConnectivityMap 튜토리얼 강의를 통해 이러한 대규모 데이터셋을 탐색하고 분석하는 방법을 배우게 되며, 실제 사례 연구를 통해 ConnectivityMap이 어떻게 실제 연구에 적용될 수 있는지 배운다. 또한 이론적 지식뿐만 아니라 실제적인 적용 능력을 함양하는 것을 목표로 한다.

강의는 다음의 내용을 포함한다:

- ConnectivityMap 데이터의 이해
- ConnectivityMap 데이터의 활용
- ConnectivityMap 데이터의 응용

* 교육생준비물: 노트북 (메모리 8GB 이상, 디스크 여유공간 30GB 이상)

* 강의 난이도: 초급

* 강의: 전민지 교수 (고려대학교 의과대학)

Curriculum Vitae

Speaker Name: Minji Jeon, Ph.D.

▶ **Personal Info**

Name	Minji Jeon
Title	Assistant Professor
Affiliation	Korea University

▶ **Contact Information**

Address	161, Jeongneung-ro, Seongbuk-gu, Seoul, 02708
Email	mjjeon@korea.ac.kr
Phone Number	010-2354-7084

Research Interest

AI-driven drug discovery, machine learning, bioinformatics

Educational Experience

2012	B.S. in Computer Science, Korea University, Korea
2014	M.S. in Interdisciplinary Graduate Program in Bioinformatics, Korea University, Korea
2018	Ph.D. in Computer Science, Korea University, Korea

Professional Experience

2018-2019	Research Professor, Korea University, Korea
2020-2022	Postdoctoral Fellow, Icahn School of Medicine at Mount Sinai, USA
2022-	Assistant Professor, Korea University, Korea

Selected Publications (5 maximum)

1. Zhaoping Xiong†, **Minji Jeon**†, Robert J Allaway†, Jaewoo Kang, Donghyeon Park, Jinhyuk Lee, Hwisang Jeon, Miyoung Ko, Hualiang Jiang, Minyue Zheng, Aik Choon Tan, Xindi Guo, The Multi-Targeting Drug DREAM Challenge Community, Kristen K Dang, Alex Tropsha, Chana Hecht, Tirtha K. Das, Heather A. Carlson, Ruben Abagyan, Justin Guinney, Avner Schlessinger*, Ross Cagan* "Crowdsourced identification of multi-target kinase inhibitors for RET- and TAU-based disease: the Multi-Targeting Drug DREAM Challenge" PLoS computational biology 17.9 (2021): e1009302.
2. **Minji Jeon**†, Kathleen M. Jagodnik†, Eryk Kropiwnicki, Daniel J. Stein, Avi Ma'ayan* "Prioritizing Pain-Associated Targets with Machine Learning" Biochemistry 60.18 (2021): 1430-1446.
3. **Minji Jeon**, Donghyeon Park, Jinhyuk Lee, Hwisang Jeon, Miyoung Ko, Sunkyu Kim, Yonghwa Choi, Aik-Choon Tan, Jaewoo Kang* "ReSimNet: Drug Response Similarity Prediction using Siamewe Neural Networks" Bioinformatics 35.24 (2019): 5249-5256.
4. Michael Patrick Menden, Dennis Wang, Yuanfang Guan, Michael Mason, Bence Szalai, Krishna C Bulusu, Thomas Yu, Jaewoo Kang, **Minji Jeon**, Russ Wolfinger, Tin Nguyen, Mikhail Zaslavskiy, AstraZeneca-Sanger Drug Combination DREAM Consortium, In Sock Jang, Zara Ghazoui, Mehmet Eren Ahsen, Robert Vogel, Elias Chaibub Neto, Thea Norman, Eric KY Tang, Mathew J Garnett, Giovanni Di Veroli, Stephen Fawell, Gustavo Stolovitzky, Justin Guinney, Jonathan R Dry, Julio Saez-Rodriguez*, "Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen" Nature Communications, 10.1 (2019): 2674.
5. **Minji Jeon**, Sunkyu Kim, Sungjoon Park, Heewon Lee, Jaewoo Kang* "In silico drug combination discovery for personalized cancer therapy" BMC systems biology, 2018, 12.2: 16.

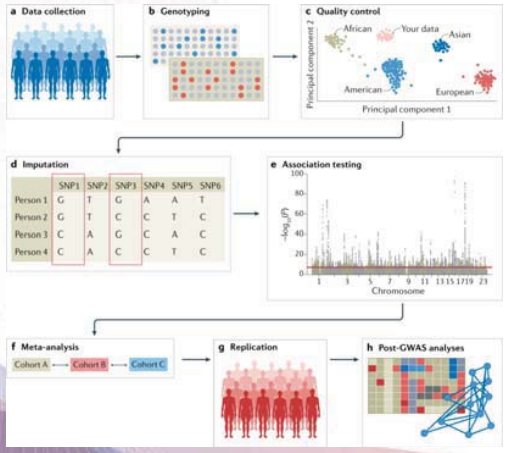
KSBi-BIML 2024

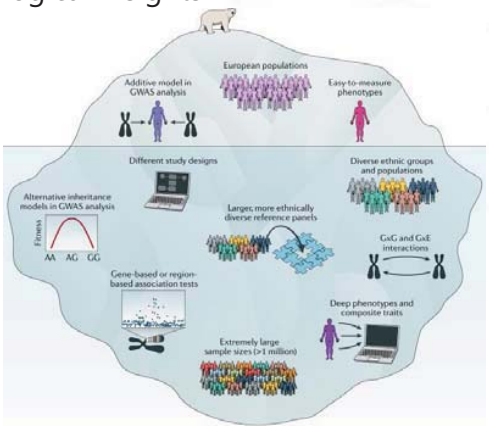
Introduction to ConnectivityMap

고려대학교 의과대학 전민지

Genome-based Disease Research

- GWAS (Genome Wide Association Study)
  - To identify genomic variants that are statistically associated with a risk for a disease or a particular trait
  - Limitations: association with disease is generally not sufficient to establish causality or to provide mechanistic and circuit-level biological insights



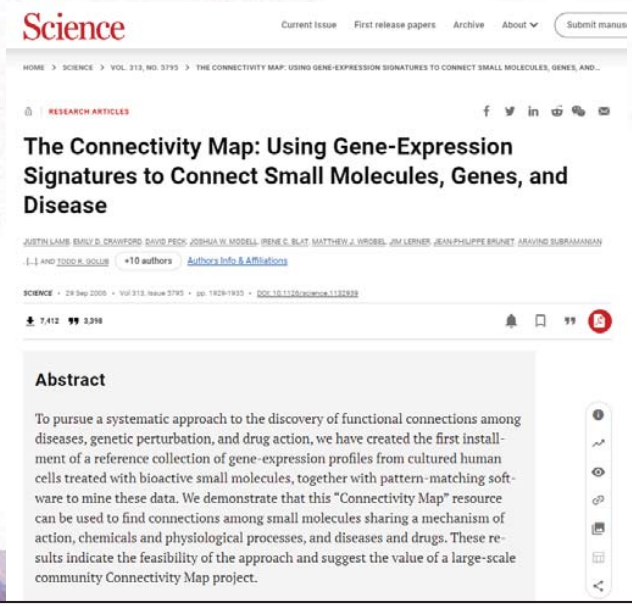


Uffelmann, Emil, et al. *Nature Reviews Methods Primers* (2021)
Tam, Vivian, et al. *Nature Reviews Genetics* (2019)

2

ConnectivityMap Concept

- ConnectivityMap: Linking disease, therapeutics and cell physiology



Science

The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease

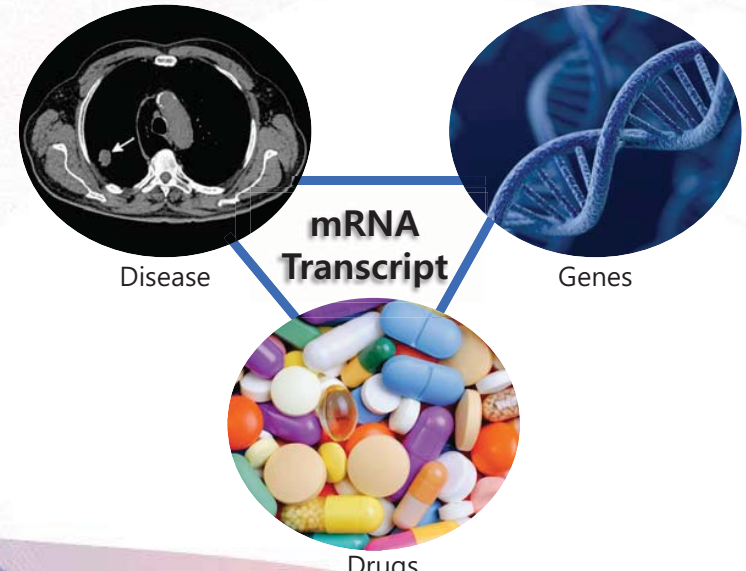
Abstract

To pursue a systematic approach to the discovery of functional connections among diseases, genetic perturbation, and drug action, we have created the first installment of a reference collection of gene-expression profiles from cultured human cells treated with bioactive small molecules, together with pattern-matching software to mine these data. We demonstrate that this "Connectivity Map" resource can be used to find connections among small molecules sharing a mechanism of action, chemicals and physiological processes, and diseases and drugs. These results indicate the feasibility of the approach and suggest the value of a large-scale community Connectivity Map project.

Lamb, Justin, et al. *science* 2006 3

ConnectivityMap Concept

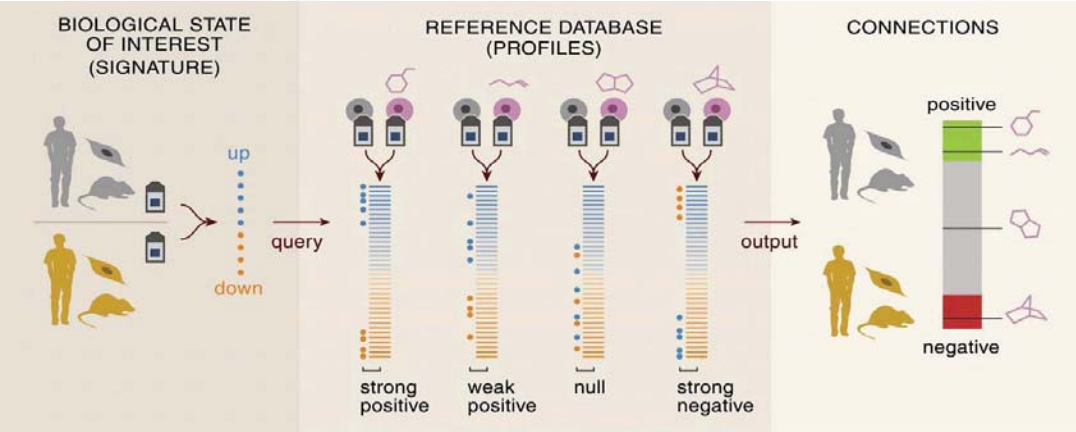
- Linking disease, therapeutics and cell physiology



4

ConnectivityMap Concept

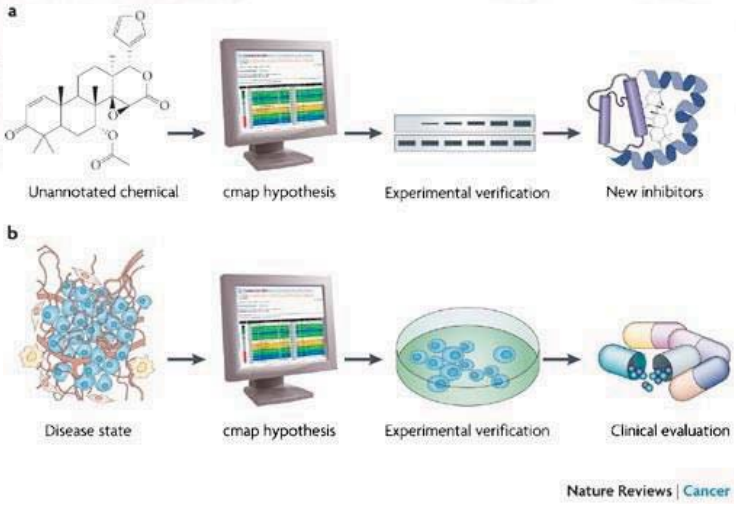
- Gene expression data could be used for the functional annotation of small molecules and genes



5

Applications of the ConnectivityMap

- The Connectivity Map is a tool for the bench researcher



Lamb, Justin. et al. *Nature reviews cancer* (2007)

6

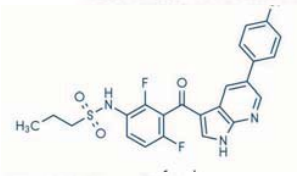
Definitions

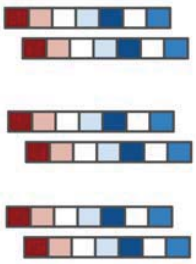
- Perturbation: an alteration of the function of a biological system, induced by external or internal mechanisms.
- Perturbagens: perturbing agents that are screened in an assay (e.g., small molecules, shRNA etc)
- Gene Signatures: differential expression of genes between two conditions, a control condition and a perturbation condition

7

ConnectivityMap v1

- Perturb cells and measure cellular responses





164 small molecules
FDA-approved drugs
nondrug bioactive tool compounds

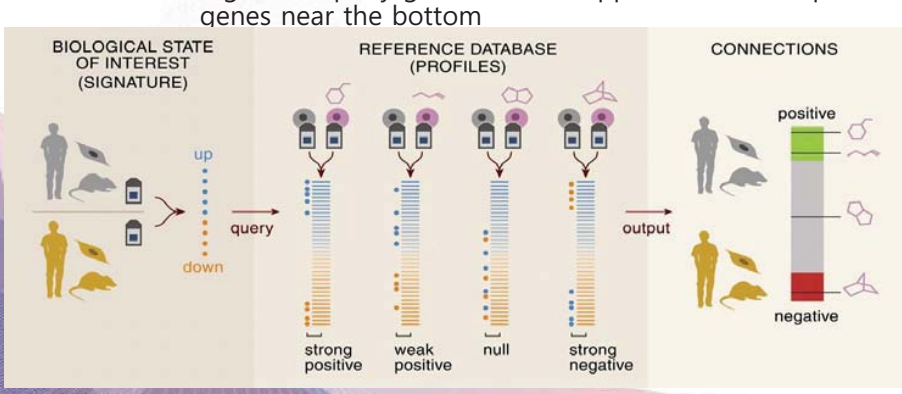
4 cell lines
MCF7 (breast cancer)
PC3 (prostate cancer)
HL60 (leukemia)
SKMEL5 (melanoma)

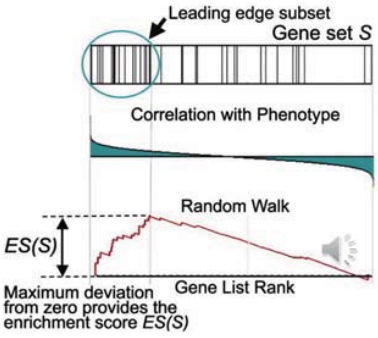
564 microarray profiles
10 uM concentration
6 or 12 hours
with controls
-> 453 gene signatures

8

ConnectivityMap v1

- Query
  - Input: gene signature (query signature)
  - Search: rank-based pattern matching strategy based on Kolmogorov-Smirnov statistics
    - The genes on the array are rank-ordered according to their differential expression relative to the control
    - The query signature is then compared to each rank-ordered list to determine whether up-regulated query genes tend to appear near the top of the list and down-regulated query genes near the bottom





9

Results: HDAC inhibitors

- Query gene signatures
  - 13 signatures downloaded from T24 (breast carcinoma) cells treated with Trichostatin A)
- Results
  - New molecules: HC toxin, valproic ac

Fig. 2. HDAC Inhibitors. **(A)** HDAC inhibitors are highly ranked with an external HDAC inhibitor signature. The "bar-view" is constructed from 453 horizontal lines, each representing an individual treatment instance, ordered by their corresponding connectivity scores with the Glaser *et al.* (*14*) signature (+1, top; −1, bottom). All valproic acid (n = 18), trichostatin A (n = 12), vorinostat (n = 2), and HC toxin (n = 1) instances in the data set are colored in black. Colors applied to the remaining instances reflect the sign of their scores (green, positive; gray, null; red, negative). The rank, name [instance id], concentration, cell line, and connectivity score for each of the selected HDAC inhibitor instances is shown. Unabridged results from this query are provided as Result S1. **(B)** Chemical structures.

rank	perturbagen	dose	cell	score
1	vorinostat [1000]	10 μM	MCF7	1
2	trichostatin A [873]	1 μM	MCF7	0.969
3	trichostatin A [992]	100 nM	MCF7	0.931
4	trichostatin A [1050]	100 nM	MCF7	0.929
5	vorinostat [1058]	10 μM	MCF7	0.917
6	trichostatin A [981]	1 μM	MCF7	0.915
7	HC toxin [909]	100 nM	MCF7	0.914
8	trichostatin A [1112]	100 nM	MCF7	0.908
9	trichostatin A [1072]	1 μM	MCF7	0.906
10	trichostatin A [1014]	1 μM	MCF7	0.893
11	trichostatin A [332]	100 nM	MCF7	0.882
12	trichostatin A [331]	100 nM	MCF7	0.846
13	trichostatin A [448]	100 nM	PC3	0.788
14	valproic acid [345]	10 mM	MCF7	0.743
15	valproic acid [23]	1 mM	MCF7	0.735
16	valproic acid [1047]	1 mM	MCF7	0.733
17	trichostatin A [413]	100 nM	ssMCF7	0.725
18	valproic acid [410]	10 mM	HL60	0.725
19	valproic acid [458]	1 mM	PC3	0.680
33	valproic acid [409]	1 mM	HL60	0.634
39	valproic acid [1020]	500 μM	MCF7	0.619
52	valproic acid [346]	2 mM	MCF7	0.582
61	valproic acid [1078]	500 μM	MCF7	0.563
71	valproic acid [629]	1 mM	SKMEL5	0.539
72	valproic acid [347]	500 μM	MCF7	0.539
73	valproic acid [989]	1 mM	MCF7	0.538
76	valproic acid [433]	1 mM	PC3	0.528
89	trichostatin A [364]	100 nM	HL60	0.507
92	valproic acid [497]	1 mM	ssMCF7	0.501
297	valproic acid [348]	50 μM	MCF7	0
388	valproic acid [994]	200 μM	MCF7	0
403	valproic acid [1002]	50 μM	MCF7	0
419	valproic acid [1060]	50 μM	MCF7	-0.537

10

Results

- anthelmintic drug parbendazole as an inducer of osteoclast differentiation (Brum et al., 2015)
- celastrol as a leptin sensitizer (Liu et al., 2015)
- compounds targeting COX2 and ADRA2A as potential diabetes treatments (Zhang et al., 2015)
- small molecules that mitigate skeletal muscular atrophy (Dyle et al., 2014) and spinal muscular atrophy (Farooq et al., 2009)
- new therapeutic hypotheses for the treatment of inflammatory bowel disease (Dudley et al., 2011) and cancer (Singh et al., 2016; Muthuswami et al., 2013; Wang et al., 2008; Schnell et al., 2015; Fortney et al., 2015; Wang et al., 2011; Churchman et al., 2015; Rosenbluth et al., 2008; Saito et al., 2009; Stockwell et al., 2012)

11

Next-Generation ConnectivityMap (CMap v2)

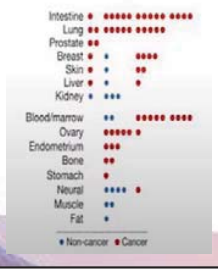
L1000 assay

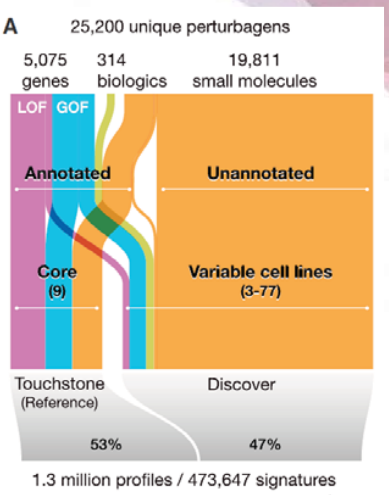
ConnectivityMap

+20,000 perturbagens

Pattern matching algorithm

+80 cell lines



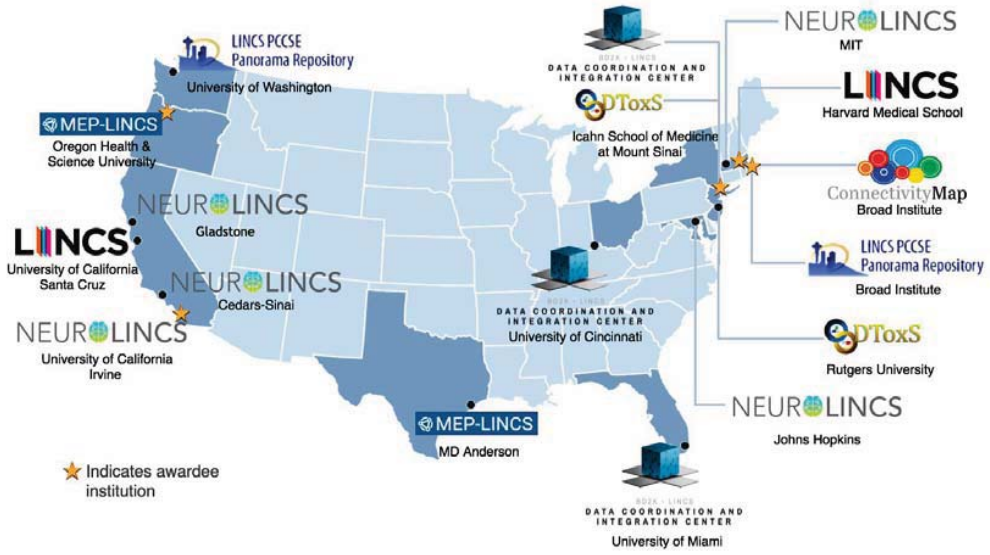


CMap version 2 with 1.3 M profiles

12

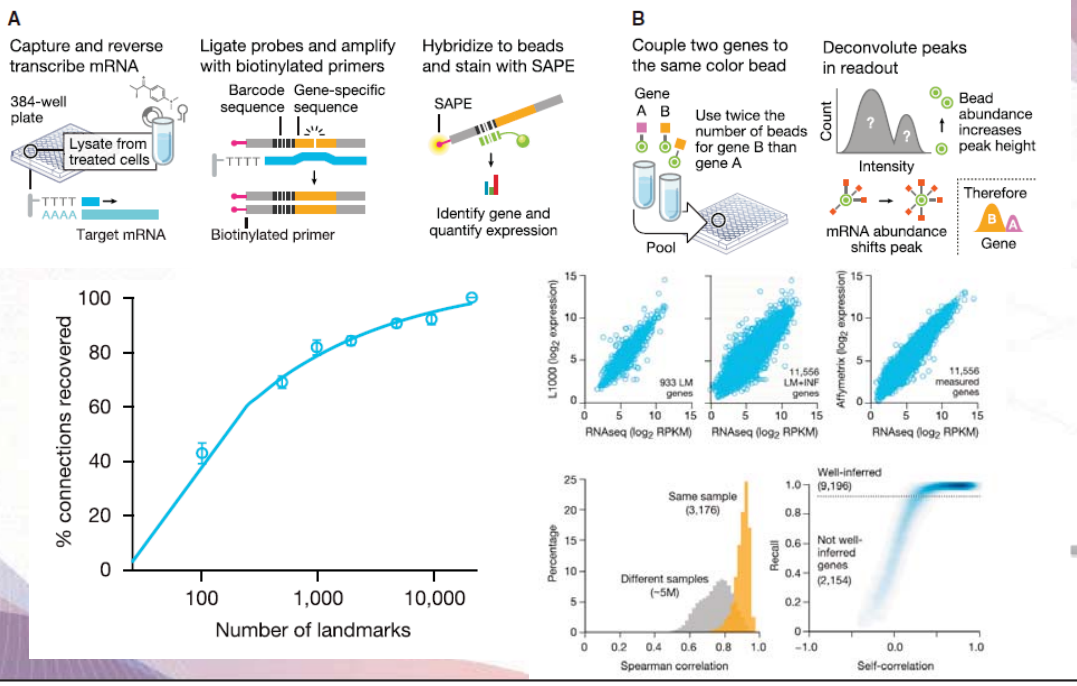
LINCS Consortium

- The Library of Integrated Network-Based Cellular Signatures (LINCS)



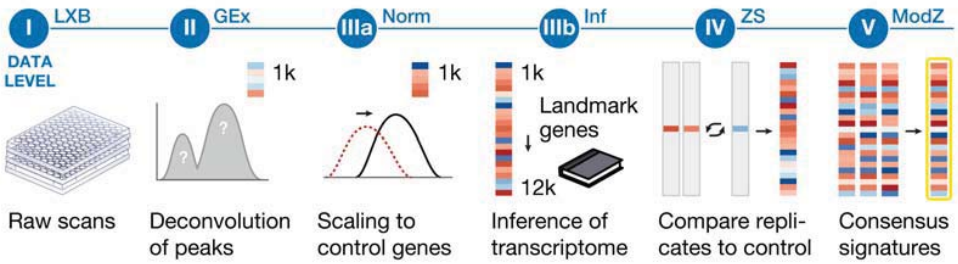
13

L1000 Assay Enables Massive Scale Up of Data

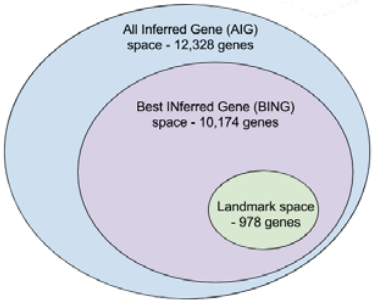


14

L1000 Assay Enables Massive Scale Up of Data

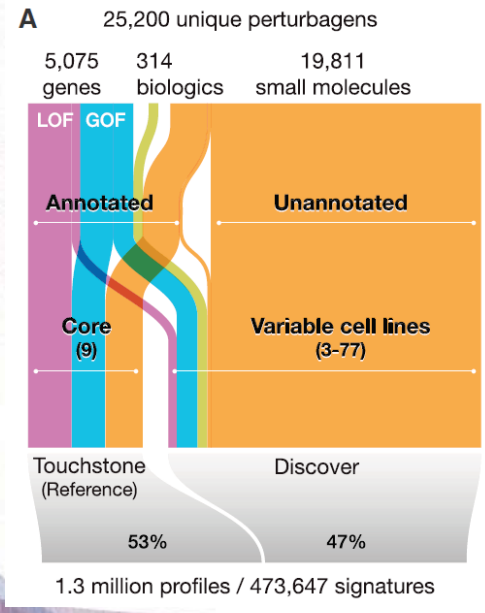


	L1000	RNA-Seq
Cost	$	$$$
Library prep?	Fast	Slow
Detection of non-abundant transcripts	No change	Needs deeper coverage
"Accuracy"	High	High
Evidence of additional benefit	?	



15

1M Profiles of CMap v2



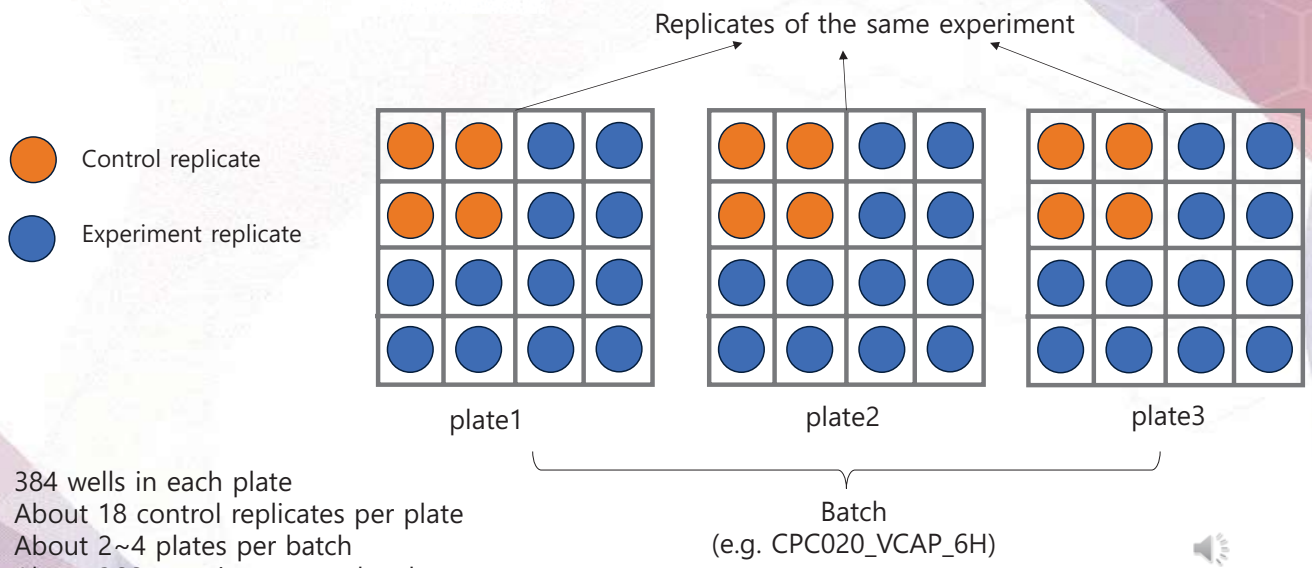
Perturbagen Type	pert_type designation in metadata files
Compound	trt_cp
Peptides and other biological agents (e.g. cytokine)	trt_lig
shRNA for loss of function (LoF) of gene	trt_sh
Consensus signature from shRNAs targeting the same gene	trt_sh.cgs
cDNA for overexpression of wild-type gene	trt_oe
cDNA for overexpression of mutated gene	trt_oe.mut
CRISPR for LLoF	trt_xpr
Controls - vehicle for compound treatment (e.g DMSO)	ctl_vehicle
Controls - vector for genetic perturbation (e.g empty vector, GFP)	ctl_vector
Controls - consensus signature from shRNAs that share a common seed sequence	trt_sh.css
Controls - consensus signature of vehicles	ctl_vehicle.cns
Controls - consensus signature of vectors	ctl_vector.cns
Controls - consensus signature of many untreated wells	ctl_untrt.cns
Controls - Untreated cells	ctl_untrt

16

1M Profiles of CMap v2

- small molecule compounds
  - ~1,300 FDA-approved drugs
  - ~5,585 bioactive tool compounds
  - 2000+ screening hits
- knocking-down genes (shRNA) or over-expressing genes
  - ~900 target/pathways of FDA-approved drugs
  - ~600 candidate disease genes
  - 500+ community nominations
- cells including primary cell lines, cancer cell lines, stem cell lines, and differentiated cell lines from different tissue types

1M Profiles of CMap v2



- 384 wells in each plate
- About 18 control replicates per plate
- About 2~4 plates per batch
- About 366 experiments per batch

Data Levels

Level 1 (LXB)	Raw fluorescent intensity (FI) values measured for every bead detected by Luminex scanners. Each 384-well plate generates 384 LXB files, where each file contains a fluorescent intensity value for each observed bead in the well.
Level 2 (GEX)	Gene expression levels for the 978 landmark genes, deconvoluted from the measured fluorescent intensity values.
Level 3 (NORM, INF)	NORM - Gene expression are **normalized** to invariant gene set curves and quantile normalized across each plate. INF- Additional values for **11,350 additional genes** not directly measured in the L10000 assay are inferred based on the normalized values for the 978 landmark genes.
Level 4 (ZS)	**Z-scores** for each gene based on Level 3 with respect to the entire plate population. This comparison of profiles to their appropriate population control generates a list of differentially expressed genes.
Level 5 (MODZ)	replicate-collapsed z-score vectors based on Level 4. Replicate collapse generates one differential expression vector, which we term a **signature**. Connectivity analyses are performed on signatures.

19

ConnectivityMap Score

- Computing similarities - Weighted Connectivity Score (WTCS)

$$w_{q,r} = \begin{cases} \left(ES_{up} - ES_{\text{down}}\right)/2, & if\ \text{sgn}\left(ES_{up}\right) \neq \text{sgn}\left(ES_{\text{down}}\right) \\ 0, & otherwise \end{cases}$$

Where ES_{up} is the enrichment of q_{up} in r and ES_{down} is the enrichment of q_{down} in r. $WTCS$ ranges between -1 and 1

- Normalization of Connectivity Scores (NCS)
  - Given a vector of WTCS values w resulting from a query, we normalize the values within each cell line (c) and perturbagen type (t) to obtain normalized connectivity scores (NCS) as

$$NCS_{c,t} = \begin{cases} w_{c,t}/\mu^{+}_{c,t} & if\ \text{sgn}\left(w_{c,t}\right) > 0 \\ w_{c,t}/\mu^{-}_{c,t} & \text{otherwise} \end{cases}$$

20

ConnectivityMap Score

- Connectivity Map Score (Tau τ)
  - by comparing each observed NCS value ncsq,r between the query q and a reference signature r to a distribution of NCS values representing the similarities between a reference compendium of queries (Qref) and r
  - Tau (τ) that ranges from −100 to +100 and represents the percentage of queries in Qref with a lower |NCS| than |ncsq,r|

$$\tau_{q,r} = \mathrm{sgn}\left(ncs_{q,r}\right) \frac{100}{N} \sum_{i=1}^{N} \left[\left|ncs_{i,r}\right| < \left|ncs_{q,r}\right|\right]$$

Results

- Discovery of MOA of Unannotated Small Molecules

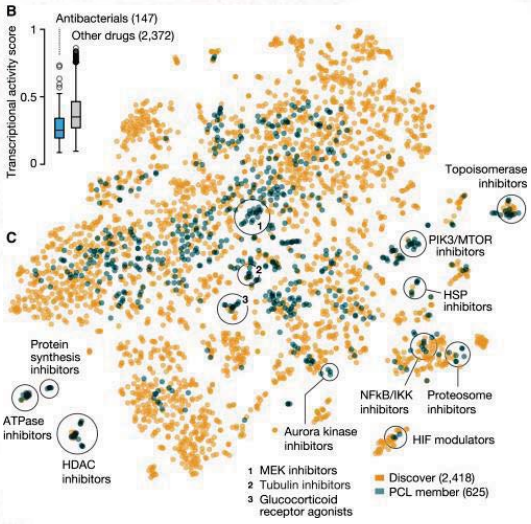


Figure 5. Characterizing Known and Unexpected Activities of Small Molecules

(A) HDAC inhibitor PCL substructure. Hierarchical clustering of pairwise connectivities of the HDAC inhibitor PCL members reveals substructure within the class. Pan-HDAC inhibitors cluster together, distinct from more isoform-selective compounds. (B) Antibacterials exhibit lower transcriptional activity than other drugs. Distributions of the maximum TAS per compound for 147 antibacterials and 2,372 known drugs in CMap-Touchstone (TS). The antibacterials' TAS distribution is significantly lower ($p < 3^{-11}$) than that of other drugs. (C) Comparison of unannotated compounds with known drugs. t-SNE projection of the signatures of 2,418 unannotated but transcriptionally active compounds (orange) with PCL members (teal). Some unannotated compounds occupy regions not covered by drugs, presenting opportunities for novel chemical development.

Results

- Discovery of a Selective CSNK1A1 inhibitor

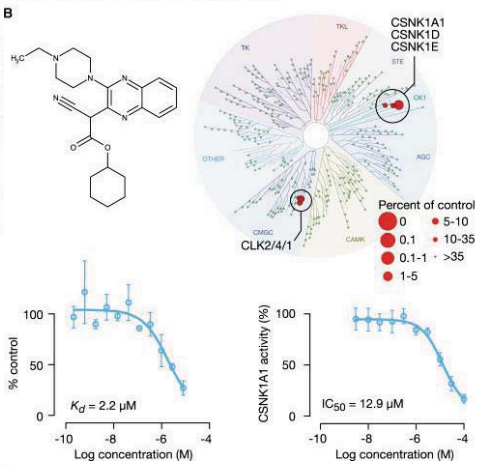


Figure 6. Kinase Inhibitor Discovery Using Reference Transcriptional Signatures
(A) Discovery of ROCK1/ROCK2 inhibitor. Top left panel: chemical structure of BRD-2751, predicted to be a ROCK inhibitor. Right: TREEspot selectivity profile of Kinomescan binding assay confirmed compound binding to ROCK1/ROCK2. Bottom left: dose response testing by Kinomescan showed ROCK1 K_D of 56 nM.
(B) Discovery of novel CSNK1A1 inhibitor. Top left panel: the chemical structure of BRD-1868. Top right: TREEspot image of Kinomescan binding assay performed with BRD-1868 at 10 μM demonstrated inhibition of 6/456 kinases tested, including CSNK1A1. Bottom left: CSNK1A1 binding by BRD-1868 confirmed by Kinomescan, with Kd 2.2 μM. Bottom right: BRD-1868 inhibits phosphorylation of peptide substrate by CSNK1A1, with IC_{50} 12.9 μM. Error bars indicate standard deviation between technical replicates.

23

Results

Table 4. An overview of the application of CMap for a number of different diseases

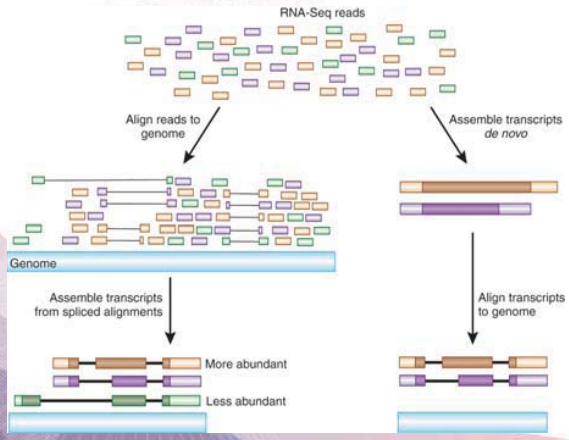
Disease	Method	Data set	Result	Drug	Reference
CNS injuries	CMap tool	Human MCF7 breast adenocarcinoma (GSE34331)	The findings show the hypothesis that inhibition of calmodulin signaling might allow neurons to alleviate substrate derived neurite growth restriction and CNS regeneration.	Calmodulin and piperazine phenothiazine (repurposed)	[54]
GBM	Pathway analysis and CMap tool	GBM data sets (GSE4290, GSE7696, GSE14805, GSE15824 and GSE16011)	Investigated antitumor drugs in GBM cell lines and identify novel drugs that can suppress GBM tumors.	Thioridazine	[55]
Gaucher disease (GD1)	Pathway analysis and CMap tool	GD1 mouse (GSE2308)	Predicted highly enriched anti-helminthic compounds for new drug action on GD1 and repurposing.	Albendazole and oxamniquine	[52]
Ovarian cancer	CMap tool	MCF7 and PC3 cell lines (GSE5258)	Found a compound as PI3K/AKT pathway inhibitor that shows the mechanism of cancer therapeutics.	Thioridazine	[56]
Stem cell leukemia (SCL)	GSEA and CMap tool	hESCs cell lines (GSE54508)	Found two HDAC inhibitors as potential inducers that can be used in treating SCL and acute megakaryoblastic leukemias.	Trichostatin A and suberoylanilide hydroxamic acid	[57]
T-cell acute lymphoblastic leukemia (T-ALL)	GSEA and CMap tool	Human and mouse T-ALL cell lines (GSE12948, GSE8416 and GSE14618)	Identified interconnecting regulatory pathways as therapeutic targets for T-ALL.	HDAC, PI3K and HSP90 inhibitors	[51]
Prostate cancer	CMap tool	Celastrol- and gedunin-treated cell lines (GSE5505 and GSE5508)	Identified target pathways of androgen receptor (AR) signaling and modulation of HSP90 MoA.	Celastrol and gedunin	[17]
Gastric cancer	Hierarchical clustering and CMap tool	Yonsei gastric cancer (GSE13861)	Predicted two possible drug candidates for gastric cancer therapy.	Vorinostat and trichostatin A	[53]
Myelomatosis	CMap tool	Human myeloma cell lines (GSE14011)	Found a drug with potential to induce suppression of cyclin D2 promoter regulation.	Pristimerin	[58]
AML	CMap tool	AML data (GSE7538)	Predicted novel treatment of human primary AML with parthenolide and transcriptional response of cells.	Celastrol	[59]

Musa, Aliyu, et al *Briefings in bioinformatics* (2018) 25

세포주 약물 반응 데이터의 활용

How to get gene signatures?

- The most fundamental problem:
  - Detecting and interpreting differences in the abundance of genes or other genomic features between experimental conditions, cell types, or disease states

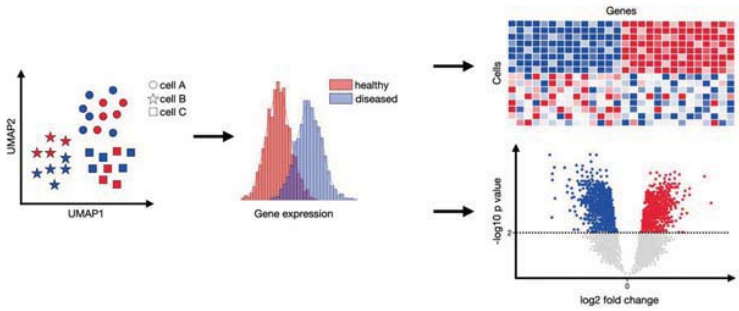


	Wild-type		Mutant	
	Mouse 1	Mouse 2	Mouse 1	Mouse 2
Gene 1	45	60	30	39
Gene 2	0	4	3	7
Gene 3	1010	800	3099	3450
…	…	…	…	…

27

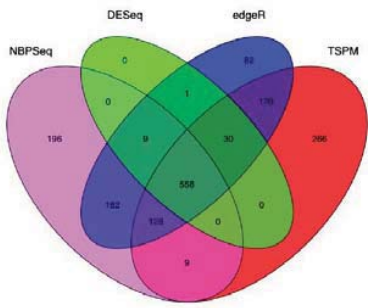
Differential expression analysis

- Differential expression analysis tests thousands of hypotheses (one test for each gene) for gene activity changes between conditions (case and control).
- Factors affecting analysis power include limited biological replicates, non-normal distribution of read counts, and measurement uncertainty for lowly expressed genes.



Differential expression analysis

- There are a number of software packages that have been developed for diff erential expression analysis of RNA-seq data
- Tools like edgeR and DESeq2 overcome these limitations using statistical m odels based on negative binomial distribution.
- There is no one method that performs optimally under all conditions

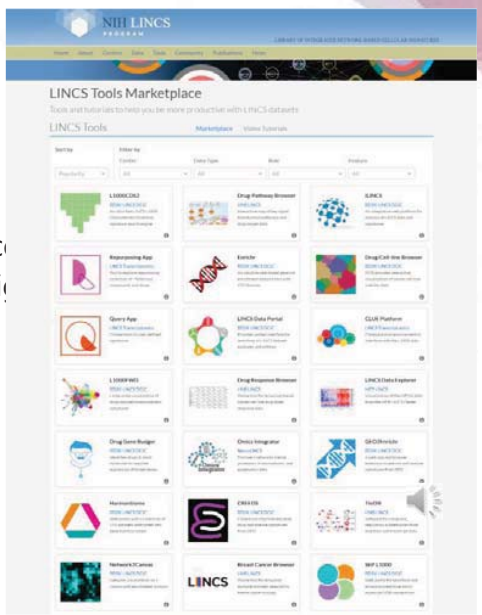


Soneson, Charlotte, and Mauro Delorenzi
BMC bioinformatics (2013)

NIH LINCS Program

- LINCS Tools
  - >50 tools developed by the consortium
  - Interactive visualizations
  - Data querying and browsing
  - Various analysis workflows
- LINCS Data
  - Transcriptomics, proteomics, epigenomics, imaging c
  - L1000 data contain >3 million samples, >1 million si
  - Programmatic access via APIs
  - User interfaces for querying and viewing signatures

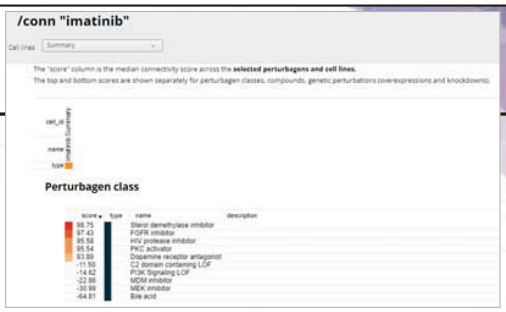
https://lincsproject.org/LINCS/tools

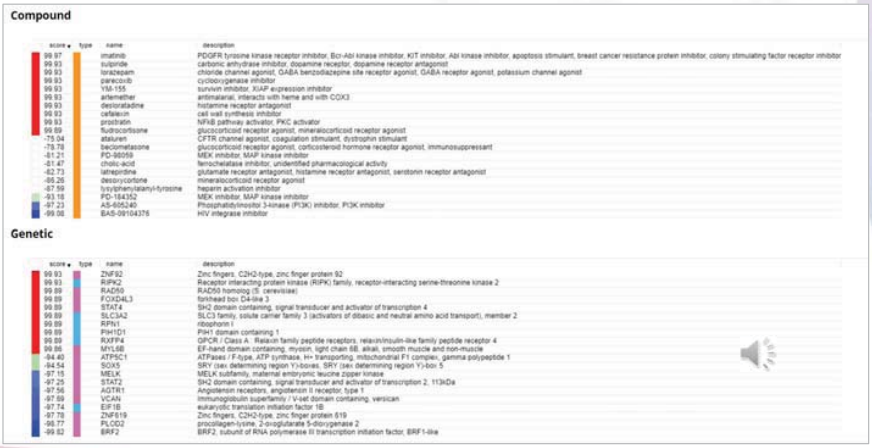


30

ConnectivityMap

- Visit clue.io
  - Single keyword search

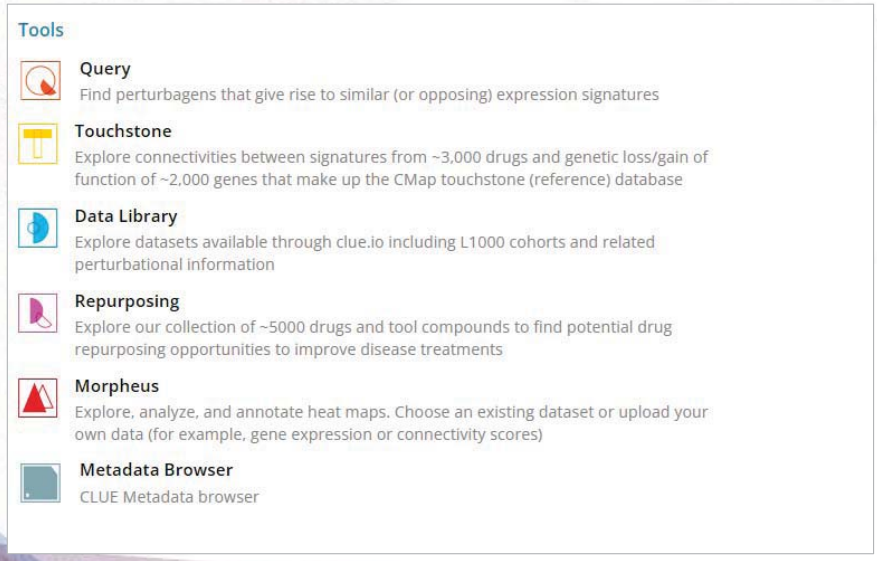




31

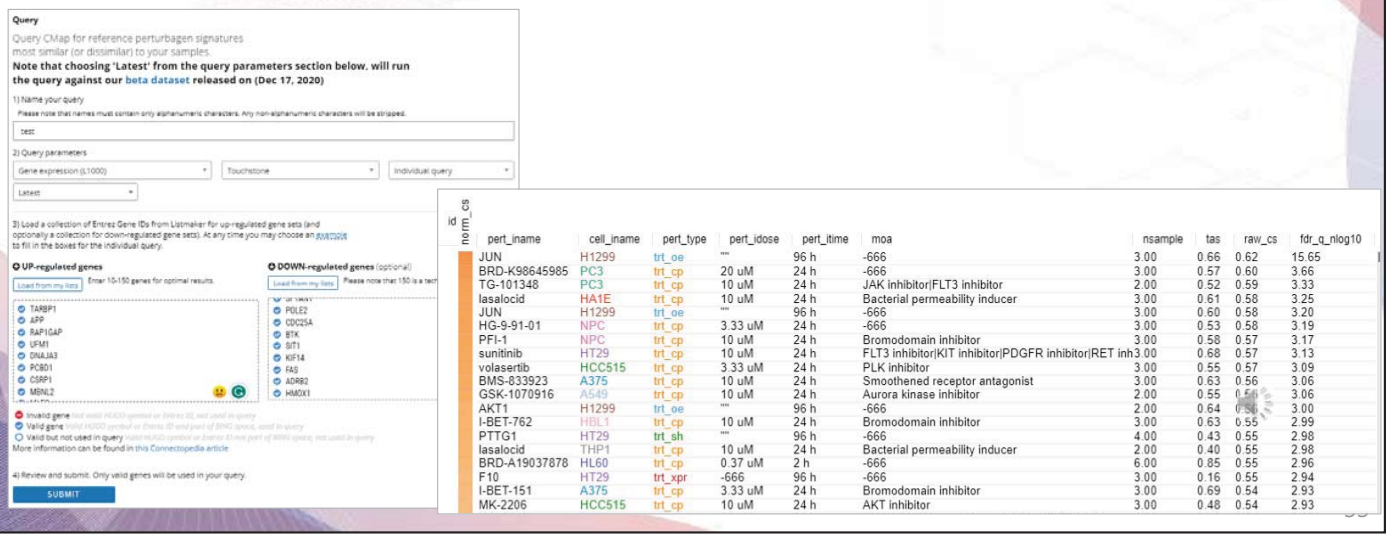
ConnectivityMap

- Search tools



ConnectivityMap

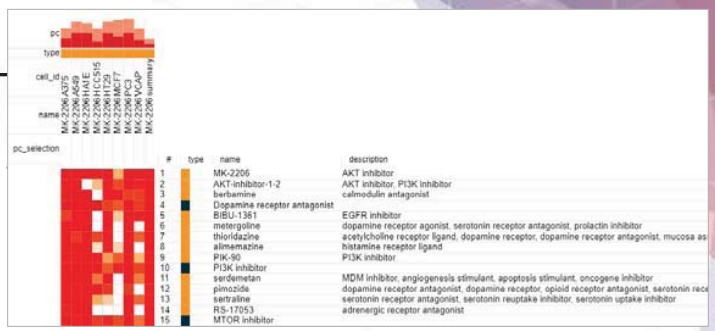
- Query signature
  - To find perturbagens that give rise to similar (or opposing) expression signatures
  - Take some time.. ~50min

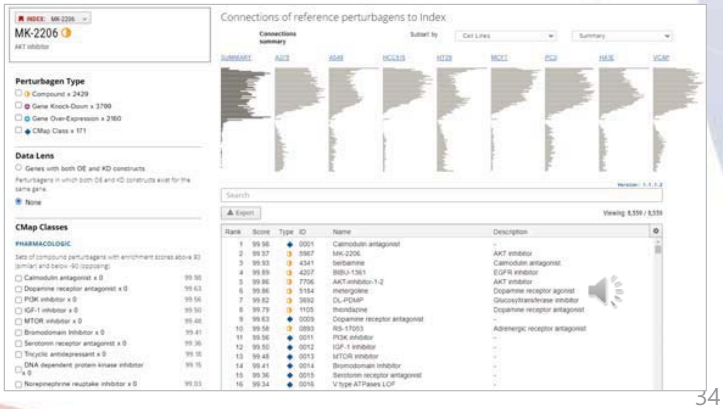


id	norm_cs	pert_iname	cell_iname	pert_type	pert_idose	pert_itime	moa	nsample	tas	raw_cs	fdr_q_nlog10
		JUN	H1299	trt_oe	""	96 h	-666	3.00	0.66	0.62	15.65
		BRD-K98645985	PC3	trt_cp	20 uM	24 h	-666	3.00	0.57	0.60	3.66
		TG-101348	PC3	trt_cp	10 uM	24 h	JAK inhibitor|FLT3 inhibitor	2.00	0.52	0.59	3.33
		lasalocid	HA1E	trt_cp	10 uM	24 h	Bacterial permeability inducer	3.00	0.61	0.58	3.25
		JUN	H1299	trt_oe	""	96 h	-666	3.00	0.60	0.58	3.20
		HG-9-91-01	NPC	trt_cp	3.33 uM	24 h	-666	3.00	0.53	0.58	3.19
		PFI-1	NPC	trt_cp	10 uM	24 h	Bromodomain inhibitor	3.00	0.58	0.57	3.17
		sunitinib	HT29	trt_cp	10 uM	24 h	FLT3 inhibitor|KIT inhibitor|PDGFR inhibitor|RET inh	3.00	0.68	0.57	3.13
		volasertib	HCC515	trt_cp	3.33 uM	24 h	PLK inhibitor	3.00	0.55	0.57	3.09
		BMS-833923	A375	trt_cp	10 uM	24 h	Smoothened receptor antagonist	3.00	0.63	0.56	3.06
		GSK-1070916	A549	trt_cp	10 uM	24 h	Aurora kinase inhibitor	2.00	0.55	0.56	3.06
		AKT1	H1299	trt_oe	""	96 h	-666	2.00	0.64	0.56	3.00
		I-BET-762	HBL1	trt_cp	10 uM	24 h	Bromodomain inhibitor	3.00	0.63	0.55	2.99
		PTTG1	HT29	trt_sh	""	96 h	-666	4.00	0.43	0.55	2.98
		lasalocid	THP1	trt_cp	10 uM	24 h	Bacterial permeability inducer	2.00	0.40	0.55	2.98
		BRD-A19037878	HL60	trt_cp	0.37 uM	2 h	-666	6.00	0.85	0.55	2.96
		F10	HT29	trt_xpr	-666	96 h	-666	3.00	0.16	0.55	2.94
		I-BET-151	A375	trt_cp	3.33 uM	24 h	Bromodomain inhibitor	3.00	0.69	0.54	2.93
		MK-2206	HCC515	trt_cp	10 uM	24 h	AKT inhibitor	3.00	0.48	0.54	2.93

ConnectivityMap

- Touchstone
  - Find connectivities between signatures function of ~2000 genes that make up
  - AKT1 inhibitor MK-2206

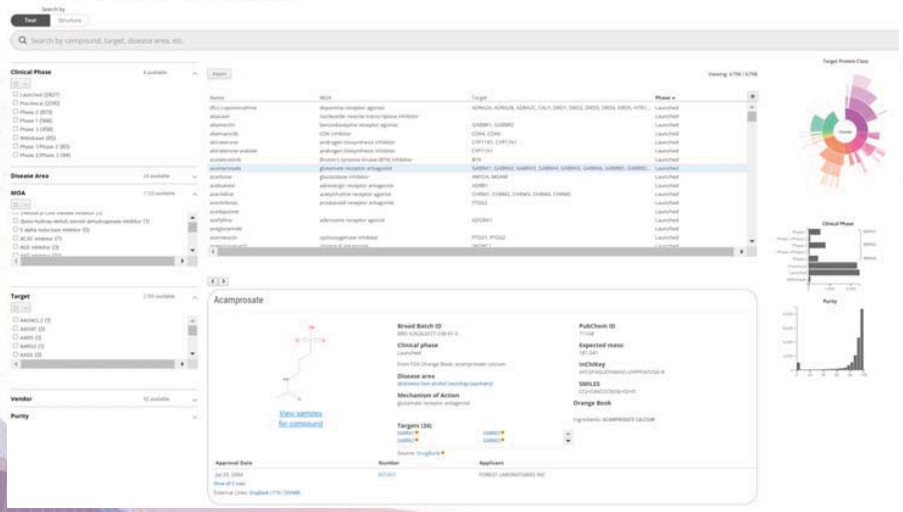




34

ConnectivityMap

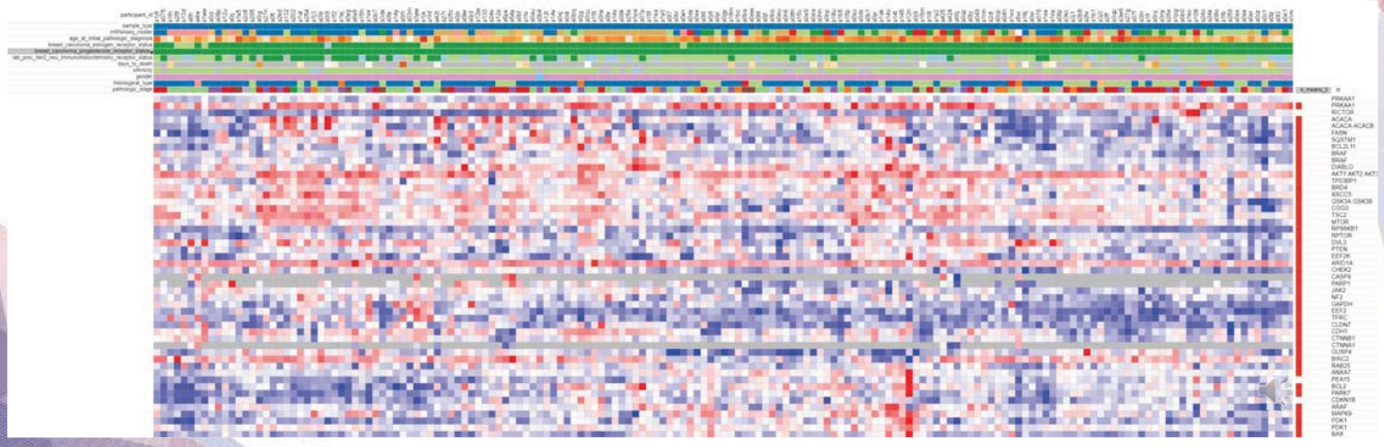
- Repurposing
  - Explore the collection of ~5000 drugs and tool compounds to find potential drug repurposing opportunities to improve disease treatments



35

ConnectivityMap

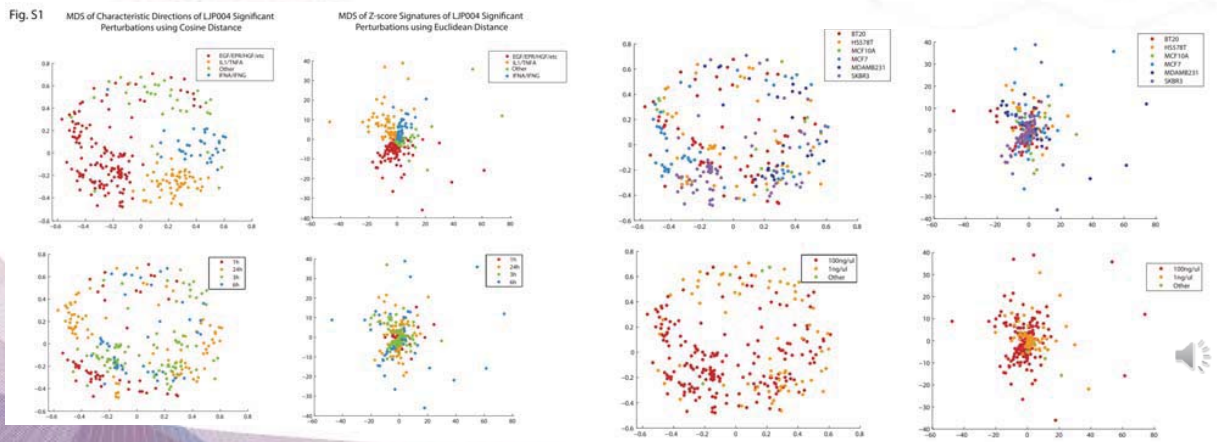
- Morpheus
  - To explore, analyze and annotate heat maps. Choose an existing dataset or upload your own data



36

L1000CDS2

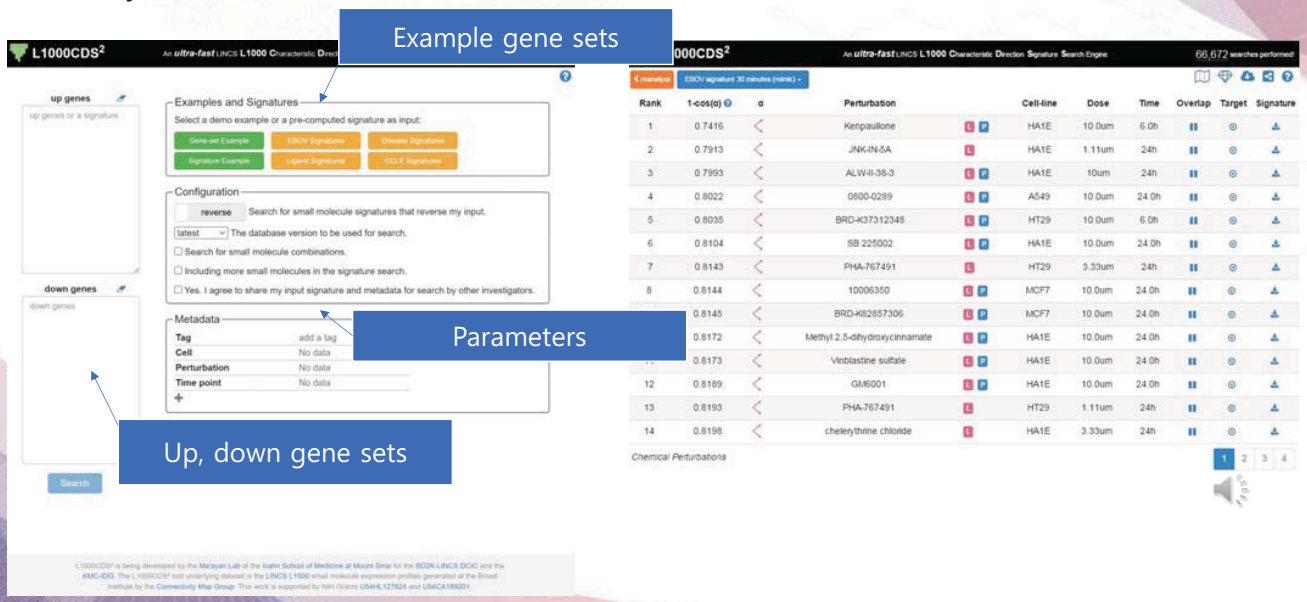
- L1000CDS2: LINCS L1000 Characteristic Direction Signature Search Engine (npj Systems Biology and Applications, 2016)
  - CD improves compared to MODZ



37

L1000CDS2

- Query and results



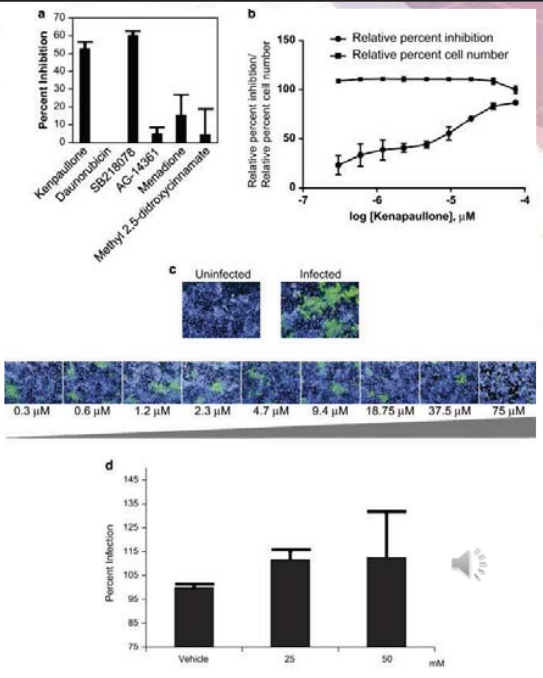
38

L1000CDS2

- Query and results
  - Input: Ebola virus signatures
  - Top candidate: Kenpaullone

Table 1. Top five predicted drugs at each time point

Drug name	Cosine distance	Batch	Cell line	Time point (h)	Concentration (μM)
30 min					
Kenpaullone	1.2584	CPC002	HA1E	6	10
0800-0289	1.1978	CPC014	A549	24	10
BRD-K37312348	1.1965	CPC016	HT29	6	10
SB 225002	1.1896	CPC001	HA1E	24	10
10006350	1.1856	CPC012	MCF7	24	10
60 min					
Kenpaullone	1.2756	CPC002	HA1E	6	10
PD 166793	1.2619	CPC001	HA1E	24	10
BAPTA-AM	1.2564	CPC001	HA1E	24	10
BRD-U74615290	1.2396	CPC014	HT29	6	10
NCGC00229596-01	1.2303	CPC008	HT29	6	10
120 min					
Kenpaullone	1.3489	CPC002	HA1E	6	10
NSC 23766	1.3478	CPC006	PC3	24	160
Rosiglitazone	1.3386	CPC006	HA1E	24	80
7-hydroxy-2, 3, 4, 5-tetrahydro-1H-[1] benzofuro[2, 3–c] azepin-1-one	1.3351	CPC007	A549	24	10
LY 364947	1.3183	CPC003	PC3	24	10



39

L1000CDS2

- Drug combinations

L1000CDS² An *ultra-fast* LINCS L1000 Characteristic Direction Signature Search Engine 66,673 searches performed!

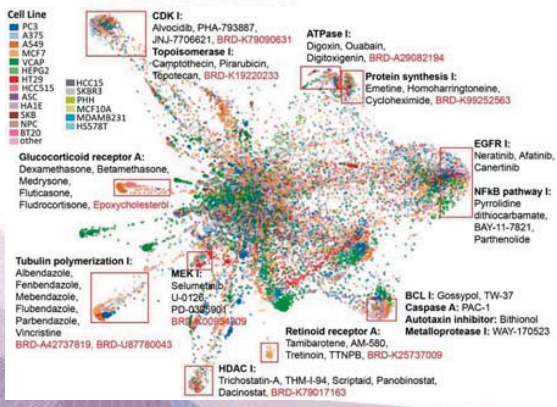
Rank	Orthogonality	Combination	
1	90°	19. GF-109203X	34. BRD-K94841585
2	89.98°	39. BRD-A68009927	47. DCC-2036
3	89.97°	35. AG14361	48. LY-2183240
4	89.97°	29. LY 2183240	38. TWS-119
5	89.96°	4. 0800-0289	32. BRD-K90382497
6	89.94°	10. Methyl 2,5-dihydroxycinnamate	19. GF-109203X
7	90.07°	19. GF-109203X	22. MLN-8054
8	90.09°	34. BRD-K94841585	45. BMS-345541
9	89.91°	17. wortmannin	19. GF-109203X
10	89.89°	25. LAWSONE	42. N9-isopropylolomoucine
11	89.88°	31. HERNIARIN	45. BMS-345541
12	89.88°	23. nilotinib	35. AG14361
13	90.17°	13. PHA-767491	30. NOCODAZOLE
14	89.82°	36. mitoxantrone	50. alvocidib

Combinations 1 2 3 4

40

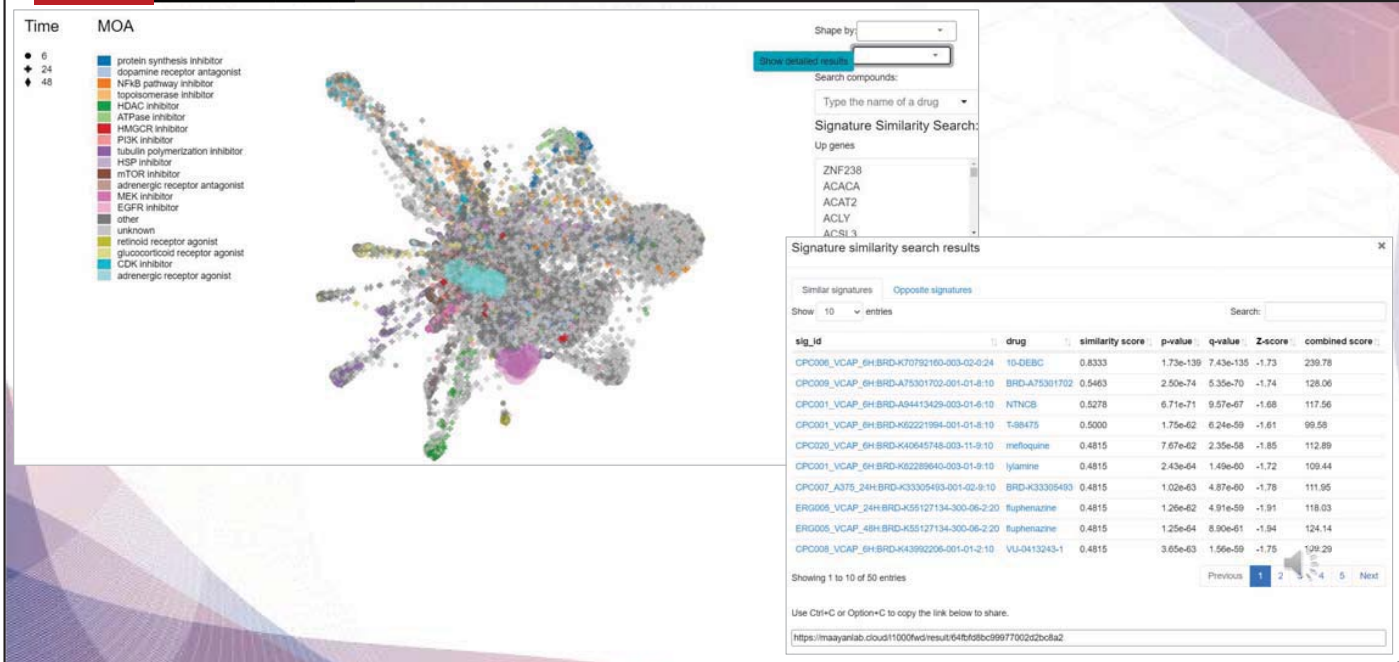
L1000FWD

- L1000 fireworks display (L1000FWD, https://maayanlab.cloud/l1000fwd/) (Bioinformatics, 2018)
  - Graph visualization of L1000 signatures
  - Similarity between signatures is computed by cosine similarity
  - 16.8K nodes and 594K edges from 68 cell lines, 3K compounds, 3 time points, 132 dosages



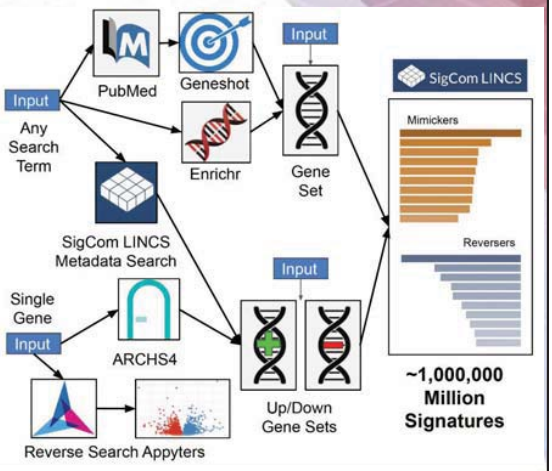
41

L1000FWD



SigCom LINCS

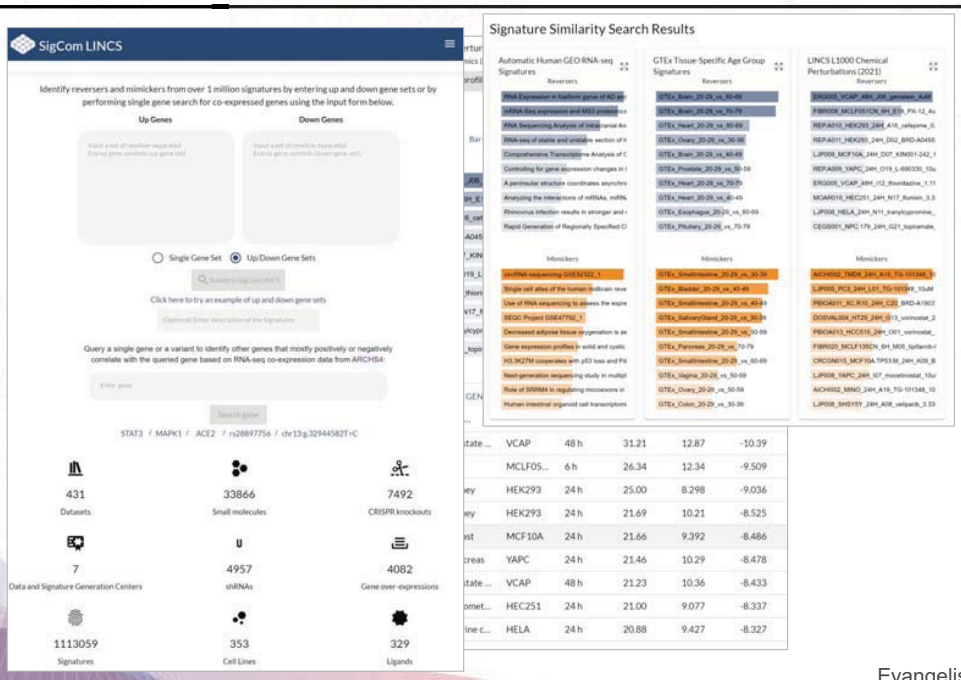
- A cloud-agonistic platform for serving, storing, and que
  - >1 million signatures from
  - LINCS
  - GTEx (Genotype-Tissue Expression)
    - Signatures of aging
  - GEO (Gene Expression Omnibus)
    - RNA-seq signatures automatically extracted from GEO
  - CREEDS (Crowd Extracted Expression of Differential Signatures)
    - Participants extracted signatures from GEO



https://maayanlab.cloud/sigcom-lincs

Evangelista, et al. *NAR* 2022

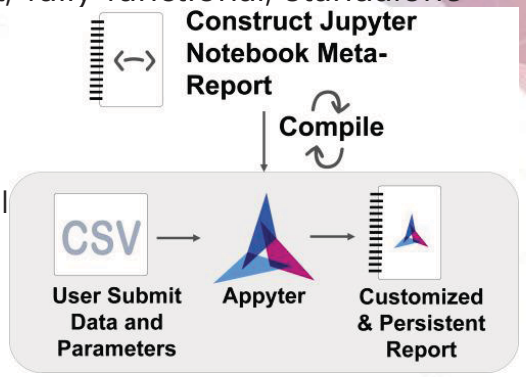
SigCom LINCS



Evangelista, et al. *NAR* 2022 44

Appyters

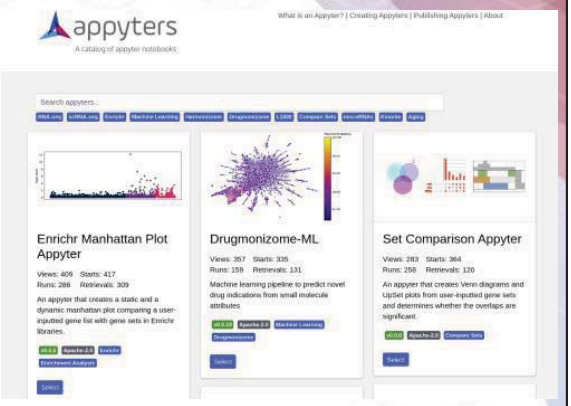
- Transform Jupyter Notebooks into lightweight, fully functional, standalone web-based bioinformatics applications
- Enable reusable workflows for
  - Customized machine learning pipelines
  - Analyzing omics data
  - Producing publishable figures and interactive visual



Clarke DJB et al. *Patterns (2021)* 45

Appyter Cataolog

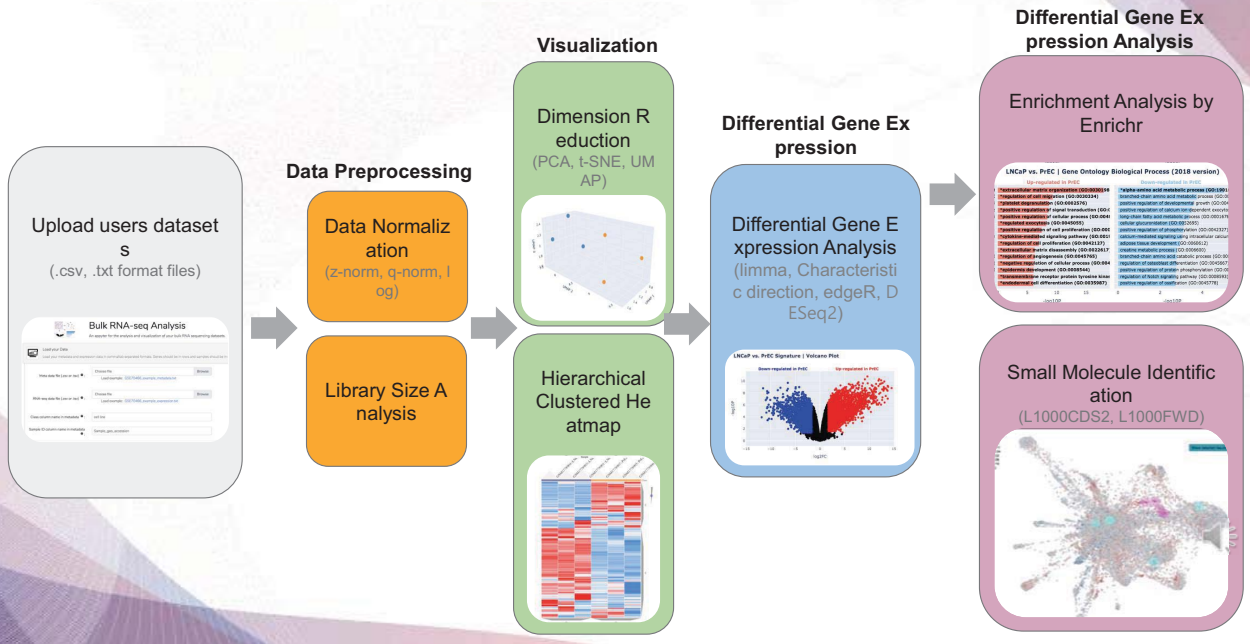
- Integrates all available Appyters (>100) into a
  - Github pull requests
  - Standardized, machine-validatable requirements
- Allows for categorization and search
- https://appyters.maayanlab.cloud



Clarke DJB et al. *Patterns (2021)* 46

URL: https://appyters.maayanlab.cloud/#/Bulk_RNA_seq

Bulk RNA-seq Analysis Appyter



Input for Bulk RNA-seq Analysis Appyter

Series GSE154613 — Query DataSets for GSE154613

Status	Public on Jul 17, 2020
Title	Modulating the transcriptional landscape of SARS-CoV-2 as an effective method for developing antiviral compounds
Organism	Homo sapiens
Experiment type	Expression profiling by high throughput sequencing
Summary	The pandemic of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has imposed a significant burden on the human population. To understand the virus and the disease it causes we sought to interfere with the transcriptional response of the infected host. Utilizing the expression pattern of SARS-CoV-2-infected cells, we identified a region in gene expression space that was unique to virus infection and inversely proportional to the transcriptional footprint of known compounds characterized in the Library of Integrated Network-based Cellular Signatures (LINCS). Here we demonstrate the successful identification of compounds that display efficacy in blocking SARS-CoV-2 replication based on their ability to counteract the virus-induced transcriptional landscape. These compounds were found to potently reduce viral load despite having no impact on viral entry or modulation of the host antiviral response in the absence of virus. RNA-Seq profiling implicated the induction of the cholesterol biosynthesis pathway as the underlying mechanism of inhibition and suggested that targeting this aspect of host biology may significantly reduce SARS-CoV-2 viral load.

URL: https://appyters.maayanlab.cloud/#/Bulk_RNA_seq

URL: https://appyters.maayanlab.cloud/#/Bulk_RNA_seq

Input for Bulk RNA-seq Analysis Appyter

Supplementary file	Size	Download	File type/resource
GSE154613_RAW.tar	7.6 Mb	(http)(custom)	TAR (of TXT)

SRA Run Selector

Raw data are available in SRA

Processed data provided as supplementary file

Custom GSE154613_RAW.tar archive:

Supplementary file	File size
GSM4675765_ACE2-A549_Amoldipine_drugonly_3.counts.genes.txt.gz	116.2 Kb
GSM4675766_ACE2-A549_Berbamine_COV2_1.counts.genes.txt.gz	117.7 Kb
GSM4675767_ACE2-A549_Berbamine_COV2_2.counts.genes.txt.gz	117.2 Kb
GSM4675768_ACE2-A549_Berbamine_COV2_3.counts.genes.txt.gz	116.8 Kb
GSM4675769_ACE2-A549-Berbamine-drugonly-1.counts.genes.txt.gz	114.4 Kb
GSM4675770_ACE2-A549-Berbamine-drugonly-2.counts.genes.txt.gz	116.3 Kb
GSM4675771_ACE2-A549-Berbamine-drugonly-3.counts.genes.txt.gz	114.5 Kb
GSM4675772_ACE2-A549-DMSO-COV2-1.counts.genes.txt.gz	116.1 Kb
GSM4675773_ACE2-A549-DMSO-COV2-2.counts.genes.txt.gz	116.3 Kb
GSM4675774_ACE2-A549-DMSO-COV2-3.counts.genes.txt.gz	116.5 Kb
GSM4675775_ACE2-A549_Loperamide_COV2_1.counts.genes.txt.gz	117.3 Kb
GSM4675776_ACE2-A549_Loperamide_COV2_2.counts.genes.txt.gz	117.2 Kb
GSM4675777_ACE2-A549_Loperamide_COV2_3.counts.genes.txt.gz	117.6 Kb
GSM4675778_ACE2-A549-Loperamide-drugonly-1.counts.genes.txt.gz	116.0 Kb
GSM4675779_ACE2-A549-Loperamide-drugonly-2.counts.genes.txt.gz	115.6 Kb
GSM4675780_ACE2-A549-Loperamide-drugonly-3.counts.genes.txt.gz	115.1 Kb
GSM4675781_ACE2-A549-Mock-1.counts.genes.txt.gz	116.7 Kb
GSM4675782_ACE2-A549-Mock-2.counts.genes.txt.gz	116.6 Kb
GSM4675783_ACE2-A549-Mock-3.counts.genes.txt.gz	115.9 Kb

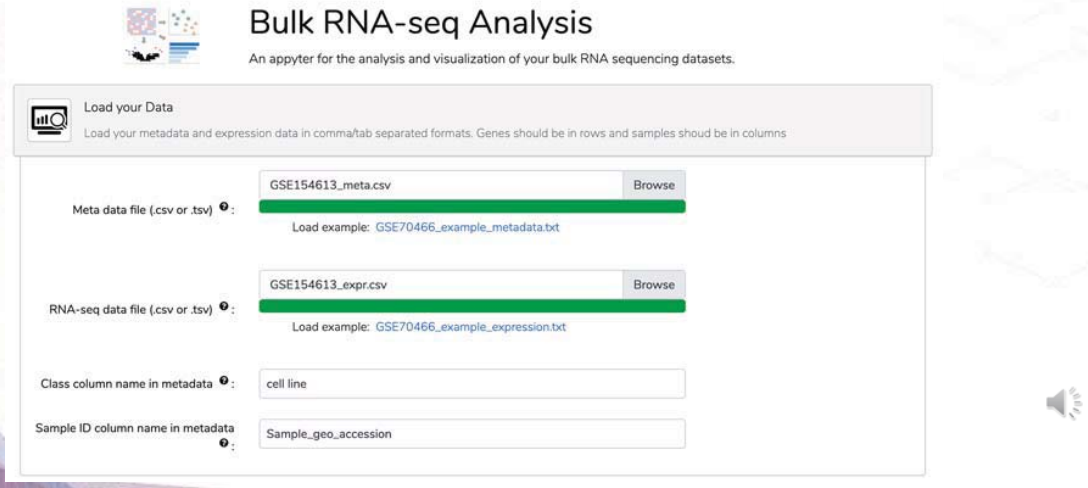
Select All | Cancel | Download | 6 file(s), 700.7 Kb

	A	B	C	D	E	F	G
1		GSM467577	GSM467576	GSM467577	GSM467577	GSM467576	GSM4675768
2	DDX11L1	0	0	1	0	0	0
3	WASH7P	23	68	15	17	98	108
4	FAM138A	0	0	0	0	0	0
5	FAM138F	0	0	0	0	0	0
6	OR4F5	0	0	0	0	0	0
7	LOC729737	1	6	5	3	13	3
8	LOC1001322	0	0	0	0	0	0
9	LOC1001320	0	0	0	0	0	0
10	LOC1001333	34	79	33	29	93	74
11	OR4F29	0	0	0	0	0	0
12	OR4F16	0	0	0	0	0	0
13	OR4F3	0	0	0	0	0	0
14	LOC1002880	12	26	12	20	43	26
15	LINC00115	2	9	3	5	13	12
16	LOC643837	169	235	166	186	249	266
17	FAM41C	0	0	0	0	0	0
18	LOC1001304	27	36	26	17	39	32
19	SAMD11	156	410	139	125	525	420
20	NOC2L	925	1457	829	833	1908	1974
21	KLHL17	66	131	45	43	152	151
22	PLEKHN1	19	34	21	14	45	49
23	C1orf170	3	9	9	6	14	19

Sample_id	Class
GSM4675774	Ctrl
GSM4675767	Case
GSM4675773	Ctrl
GSM4675772	Ctrl
GSM4675766	Case
GSM4675768	Case

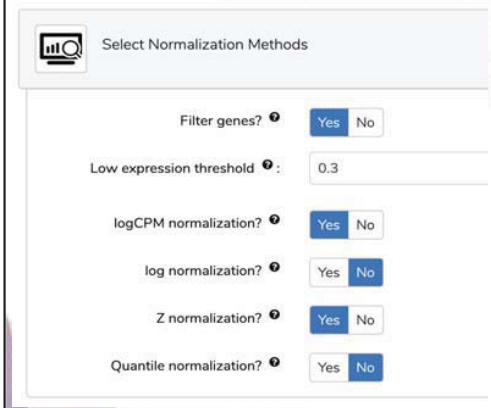
Input for Bulk RNA-seq Analysis Appyter

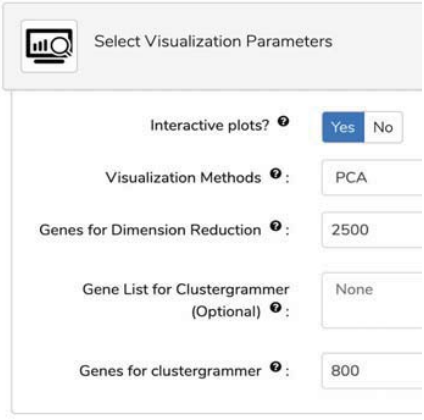
- The bulk RNA-seq analysis Appyter starts with an expression matrix of r aw read counts and metadata

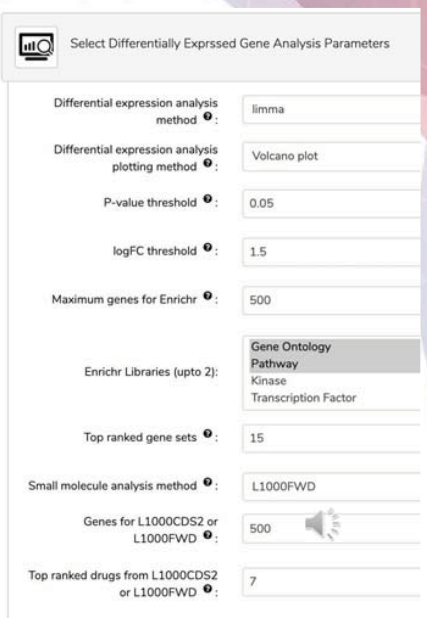


URL: https://appyters.maayanlab.cloud/#/Bulk_RNA_seq

Input for Bulk RNA-seq Analysis Appyter







URL: https://appyters.maayanlab.cloud/#/Bulk_RNA_seq

Results from Bulk RNA-seq Analysis Appyter

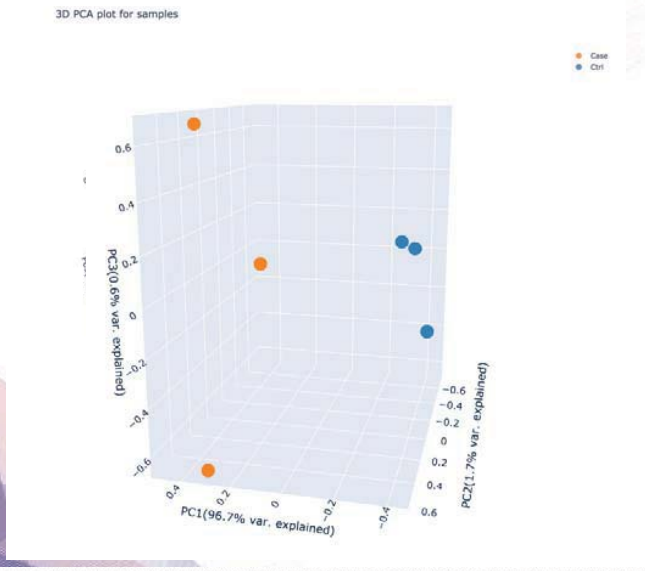


Figure 1. 3D UMAP plot for samples using 2500 genes having largest variance. The figure displays an interactive, three-dimensional scatter plot of the data. Each point represents an RNA-seq sample. Samples with similar gene expression profiles are closer in the three-dimensional space. If provided, sample groups are indicated using different colors, allowing for easier interpretation of the results.

Figure 2. Clustered heatmap plot. The figure contains an interactive heatmap displaying gene expression for each sample in the RNA-seq dataset. Every row of the heatmap represents a gene, every column represents a sample, and every cell displays normalized gene expression values. The heatmap additionally features color bars beside each column which represent prior knowledge of each sample, such as the tissue of origin or experimental treatment.

URL: https://appyters.maayanlab.cloud/#/Bulk_RNA_seq

Results from Bulk RNA-seq Analysis Appyter

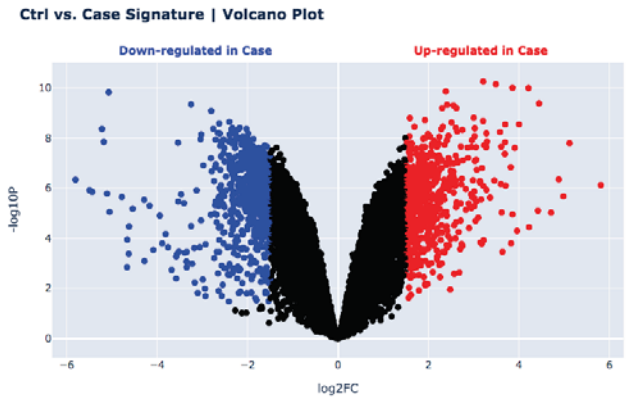


Figure 4. Volcano plot for Ctrl vs. Case. The figure contains an interactive scatter plot which displays the log2-fold changes and statistical significance of each gene calculated by performing a differential gene expression analysis. Genes with logFC > 1.0 and p-value < 0.05 in red and genes with logFC < -1.0 and p-value < 0.05 in blue. Additional information for each gene is available by hovering over it.

URL: https://appyters.maayanlab.cloud/#/Bulk_RNA_seq

Results from Bulk RNA-seq Analysis Appyter

	Signature ID	P-value	FDR	Z-score	Combined Score
1	CPC016_A549_24H:BRD-K78122587:300:01:9:10	4.665375e-25	1.997200e-20	-1.804016	43.893715
2	CPC001_VCAP_24H:BRD-K62221994:001:01:8:10	1.745531e-21	3.362437e-17	-1.638671	34.015655
3	CPC002_PC3_24H:BRD-A62336480:001:02:5:10	2.356353e-21	3.362437e-17	-1.612638	33.265106
4	CPC009_A549_6H:BRD-K90219523:003:02:	5.044622e-21	5.194117e-17	-1.743073	35.379467

Table 7. L1000FWD Results: Similar Signatures

Download CSV file: Similar Signatures for Ctrl vs. Case.csv

	Signature ID	P-value	FDR	Z-score	Combined Score
1	CPC019_VCAP_6H:BRD-K45735142:001:01:5:10	8.286689e-10	2.062470e-07	1.747248	-15.867844
2	CPC014_VCAP_24H:BRD-K05104363:001:04:3:10	2.054406e-08	3.293898e-06	1.667481	-12.818448
3	CPC014_VCAP_24H:BRD-U87780043:000:01:5:10	3.975936e-08	5.673528e-06	1.702785	-12.601566
4	CPC019_VCAP_6H:BRD-K23282736:001:01:	1.697355e-07	1.872734e-05	1.660319	-11.240738

Table 8. L1000FWD Results: Opposite Signatures

Download CSV file: Opposite Signatures for Ctrl vs. Case.csv

URL: https://appyters.maayanlab.cloud/#/Bulk_RNA_seq

MedAI Medical Science via Artificial Intelligence Lab

고려대학교 의과대학

세포주 약물 반응 데이터의 응용

Bioinformatics, 35(24), 2019, 5249–5256
doi: 10.1093/bioinformatics/btz411
Advance Access Publication Date: 22 May 2019
Original Paper

OXFORD

Data and text mining

ReSimNet: drug response similarity prediction using Siamese neural networks

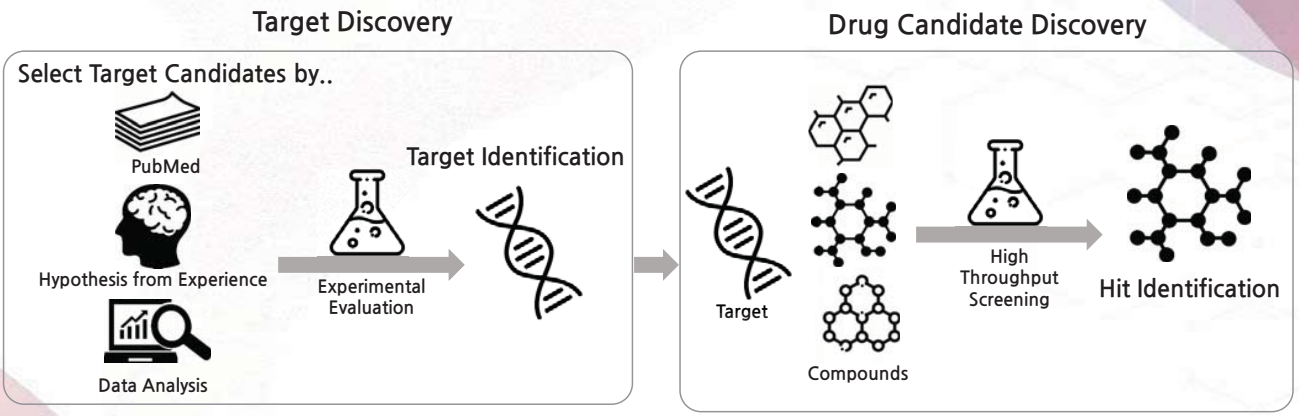
Minji Jeon[1,†], Donghyeon Park[1,†], Jinhyuk Lee[1], Hwisang Jeon[2], Miyoung Ko[1], Sunkyu Kim[1], Yonghwa Choi[1], Aik-Choon Tan[3] and Jaewoo Kang[1,2,*]

[1]Department of Computer Science and Engineering, [2]Interdisciplinary Graduate Program in Bioinformatics, Korea University, Seoul 02841, South Korea and [3]Division of Medical Oncology, Department of Medicine, Translational Bioinformatics and Cancer Systems Biology Laboratory, University of Colorado Anschutz Medical Campus, Aurora, CO 12801, USA

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Alfonso Valencia
Received on July 10, 2018; revised on April 2, 2019; editorial decision on May 11, 2019; accepted on May 16, 2019

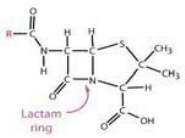
56

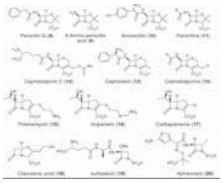
Traditional Drug Discovery Pipeline: Ligand-based Drug Discovery



Introduction

- Ligand-based drug discovery
  - To find common structures among prototype drugs that bind to a desired target
  - To design structural analogs of the prototype drugs, and evaluate them



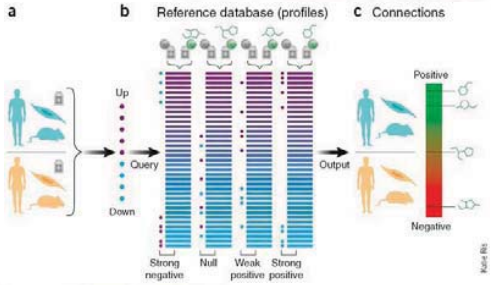


- Target discovery is essential
- Unable to work for undruggable targets
- Limited range of drug candidates: excludes drug candidates that may have similar effects despite binding to another protein

Developed a drug discovery model that can find functional analogs

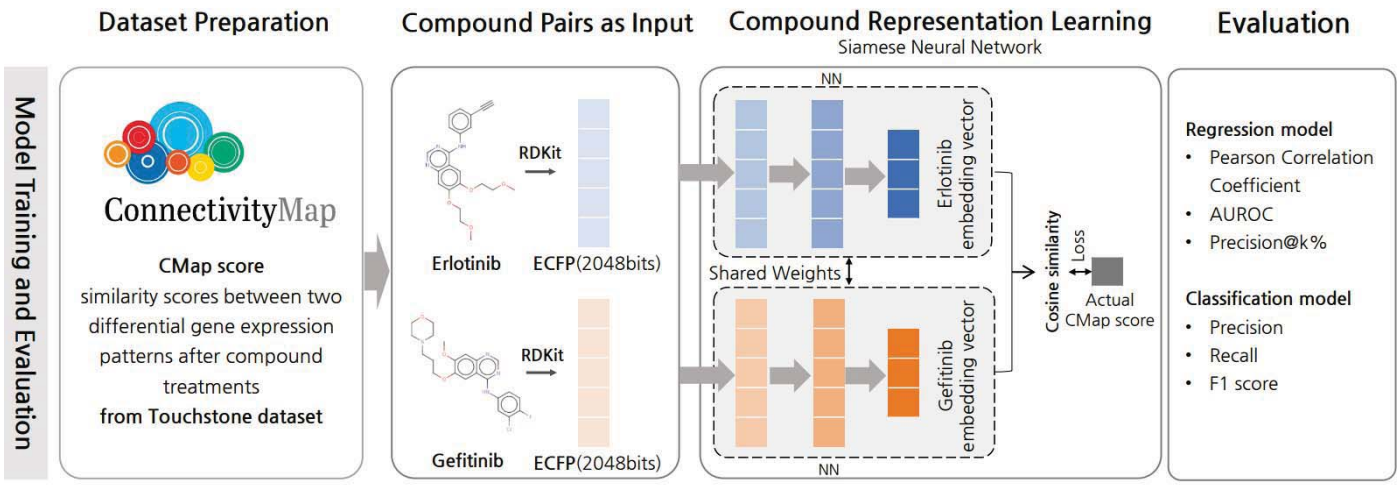
https://application.wiley-vch.de/books/sample/3527312579_c01.pdf

Gene Expression-based Drug Discovery



- ConnectivityMap (CMap) by LINCS Consortium
  - Gene expression profile-based drug repurposing platform
  - Over 3 million gene expression profiles before and after drug treatments, shRNA knockdown, CRISPR knockout, etc. and 1 million gene signatures
  - User's up- and down-regulated genes in the signature DB to find chemical compounds/shRNA/CRISPR that can mimic or reverse them
  - Provide transcriptional response-based similarity scores of compound pairs (CMap Scores)
- However, as a drug repurposing platform, it is difficult to discover novel drug candidates

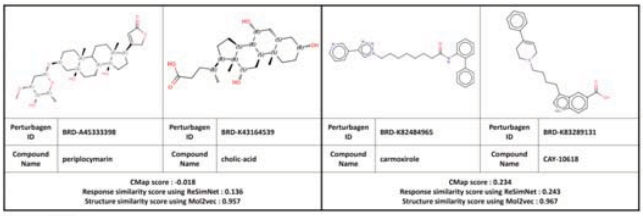
Approach



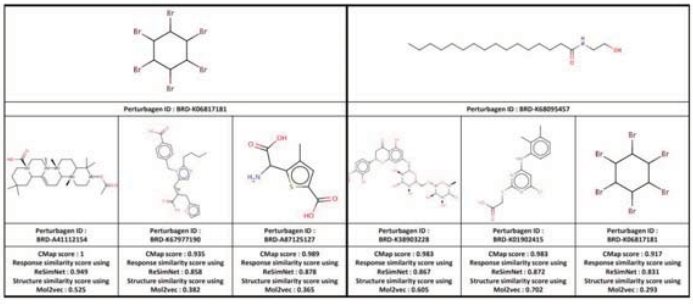
Bioinformatics, 2019

Results

▌Example pairs with similar structures but low actual CMap scores



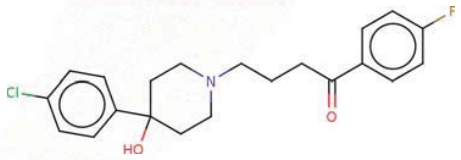
▌Example pairs with dissimilar structures but high actual CMap scores

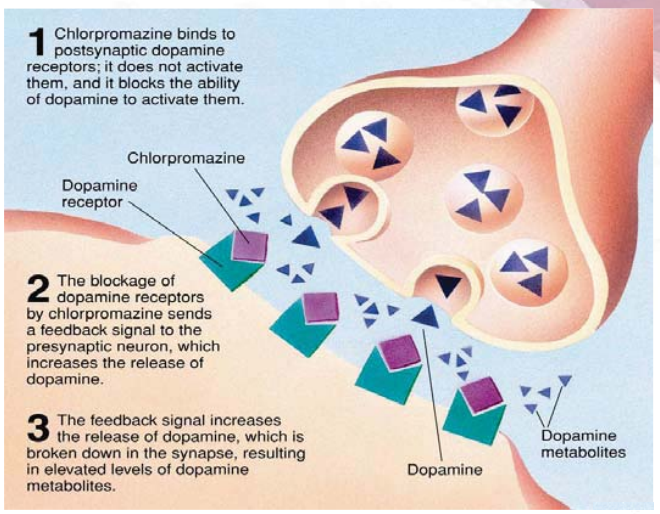


Bioinformatics, 2019

Drug discovery case study 1: Haloperidol

- Haloperidol
  - FDA-approved drug for schizophrenia
  - Dopamine receptor antagonist
  - Structure:





Drug discovery case study 1: Haloperidol

Table 2. Top drug candidates for Haloperidol from the ZINC15 database

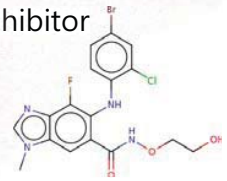
ZINC15 ID	ZINC15 name	Predicted similarity score by ReSimNet	Similarity score by ECFP	# of articles	Description
ZINC2516029	Chlorohaloperidol	0.995	0.884	6	Chlorohaloperidol targets the Dopamine D2 receptor[a]
ZINC601270	Bromperidol	0.967	0.792	66	Bromperidol is an FDA-approved drug for dementia, depression, schizophrenia, anxiety disorders and psychosomatic disorders (Yasui-Furukori *et al.*, 2002)
ZINC4214827	Amiperone	0.961	0.704	0	Amiperone targets the Dopamine D3 receptor and D3 is a potential target of Parkinson's disease and schizophrenia (Varady *et al.*, 2003)
ZINC538026	Moperone	0.955	0.792	11	Moperone is a Dopamine D2 receptor antagonist[b]
ZINC35851465	Cyantraniliprole	0.946	0.098	0	—
ZINC1481990	Budipine	0.942	0.172	1	Budipine is used in the treatment of Parkinson's disease (Klockgether *et al.*, 1993)
ZINC12494203	B-Hyodeoxycholate	0.938	0.113	0	—
ZINC3824281	Ganaxolone	0.933	0.129	1	Ganaxolone is one of neurosteroids and is used for epilepsy (Nohria and Giller, 2007)
ZINC3812988	Butorphanol	0.933	0.200	9	Butorphanol is a neuropsychiatric agent (Iyengar *et al.*, 1987)
ZINC2041178	2,3-Dibromopropanol	0.93	0.158	0	—

[a]https://pubchem.ncbi.nlm.nih.gov/compound/173712#section=ChEMBL-Target-Tree.
[b]https://www.kegg.jp/dbget-bin/www_bget? D01105.

Bioinformatics, 2019

Drug discovery case study 2: Selumetinib

- Selumetinib
  - FDA-approved cancer drug
  - MoA: MEK1/2 inhibitor
  - Structure:



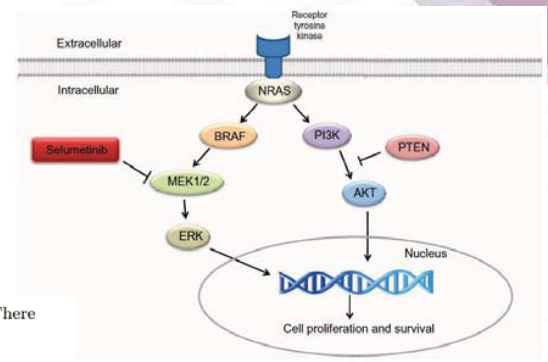


Table S 11. Top drug candidates for Selumetinib from the ZINC15 dataset. There were only two drug candidates with a score > 0.9.

ZINC15 ID	ZINC15 name	Predicted similarity score by ReSimNet	Similarity score by ECFP	# of articles	Description
ZINC38460704	binimetinib	0.993	0.841	4	Binimetinib is a MEK1/2 inhibitor [1]
ZINC43154039	buparlisib	0.937	0.084	2	Buparlisib is a PI3K inhibitor. PI3K is in a pathway parallel to the MEK pathway [2]

Bioinformatics, 2019

Jeon et al. BMC Bioinformatics (2022) 23:374
https://doi.org/10.1186/s12859-022-04895-5

BMC Bioinformatics

RESEARCH Open Access

Transforming L1000 profiles to RNA-seq-like profiles with deep learning

Minji Jeon[1,2], Zhuorui Xie[1], John E. Evangelista[1], Megan L. Wojciechowicz[1], Daniel J. B. Clarke[1] and Avi Ma'ayan[1*]

*Correspondence: avi.maayan@mssm.edu

[1] Department of Pharmacological Sciences, Mount Sinai Center for Bioinformatics, Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place, Box 1603, New York, NY 10029, USA
[2] Department of Medicine, Korea University College of Medicine, Seoul, Republic of Korea

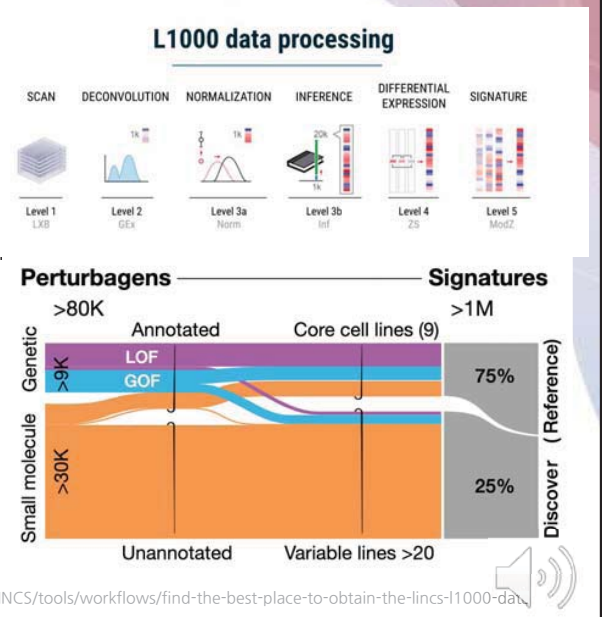
Abstract

The L1000 technology, a cost-effective high-throughput transcriptomics technology, has been applied to profile a collection of human cell lines for their gene expression response to > 30,000 chemical and genetic perturbations. In total, there are currently over 3 million available L1000 profiles. Such a dataset is invaluable for the discovery of drug and target candidates and for inferring mechanisms of action for small molecules. The L1000 assay only measures the mRNA expression of 978 landmark genes while 11,350 additional genes are computationally reliably inferred. The lack of full genome coverage limits knowledge discovery for half of the human protein coding genes, and the potential for integration with other transcriptomics profiling data. Here we present a Deep Learning two-step model that transforms L1000 profiles to RNA-seq-like profiles. The input to the model are the measured 978 landmark genes while the output is a vector of 23,614 RNA-seq-like gene expression profiles. The model first transforms the landmark genes into RNA-seq-like 978 gene profiles using a modified CycleGAN model applied to unpaired data. The transformed 978 RNA-seq-like landmark genes are then extrapolated into the full genome space with a fully connected neural network model. The two-step model achieves 0.914 Pearson's correlation coefficients and 1.167 root mean square errors when tested on a published paired L1000/RNA-seq dataset produced by the LINCS and GTEx programs. The processed RNA-seq-like profiles are made available for download, signature search, and gene centric reverse search with unique case studies.

Keywords: L1000, RNA-seq, Gene expression translation, Generative adversarial networks

Introduction – The LINCS L1000 Data

- L1000 assay
  - The L1000 assay only measures the mRNA expre
  - An additional 11,350 genes are computationally
- L1000 data
  - An expression profile is generated from a single point
  - Signatures (differentially expressed genes) are ca.. plate



https://lincsproject.org/LINCS/tools/workflows/find-the-best-place-to-obtain-the-lincs-l1000-dat..

Introduction – Limitations & Challenges

Limitations of the L1000 Data

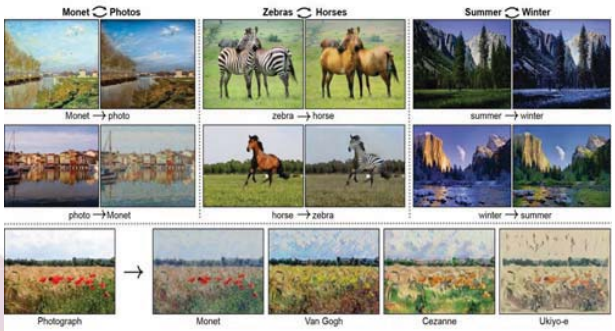
- Half of the protein-coding gene are missing from the L1000 data
  - This limits knowledge discovery about those missing genes
  - This limits the potential for integration of the CMap data with other transcriptomics pr ofiling data

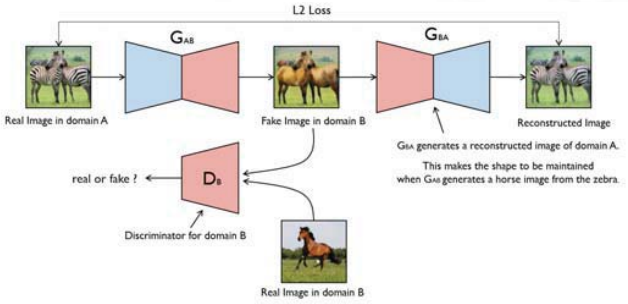
Project Challenges

- Transform L1000 profiles to RNA-seq-like profiles at the full genome scale with Deep Learning
  - How can we train a model when there are not a lot paired L1000 and RNA-seq profiles for training?
  - How can we demonstrate that predicted RNA-seq-like profiles contain knowledge?
  - What new applications the predicted RNA-seq-like profiles can provide and how we can demonstrate these applications as use cases?
  - How the trained model compares to other published models and simpler baseline models?

Methods – CycleGAN: Supervised Learning for Unpaired Data

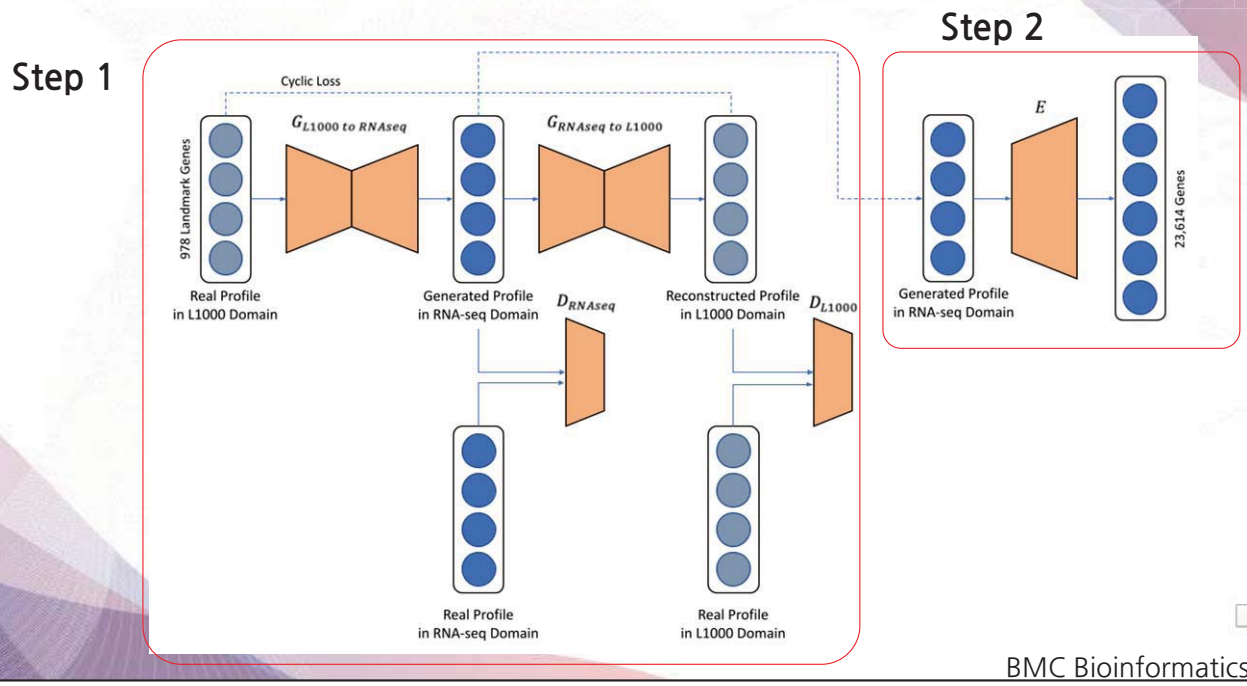
- CycleGAN (Zhu and Park et al.): An approach for learning to translate an im age from a source domain X to a target domain Y in the absence of paired samples





Zhu and Park et al., ICCV 2017

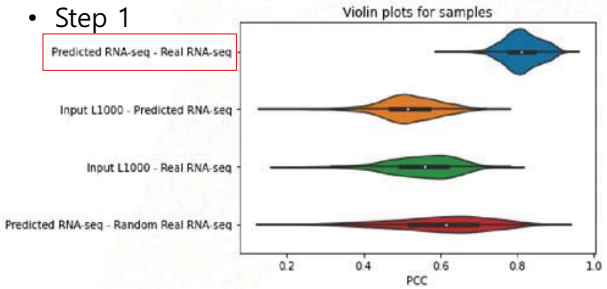
Methods – Two-Step Model for L1000 to RNA-seq-like Profiles

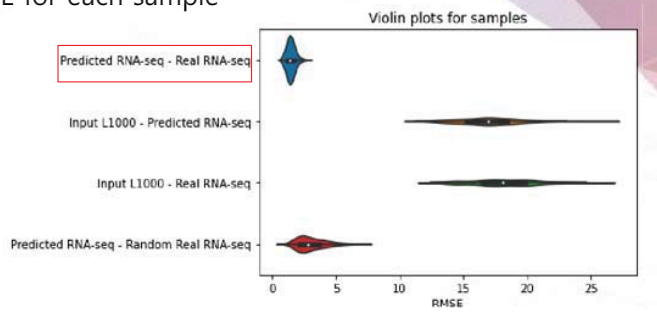


BMC Bioinformatics, 2022

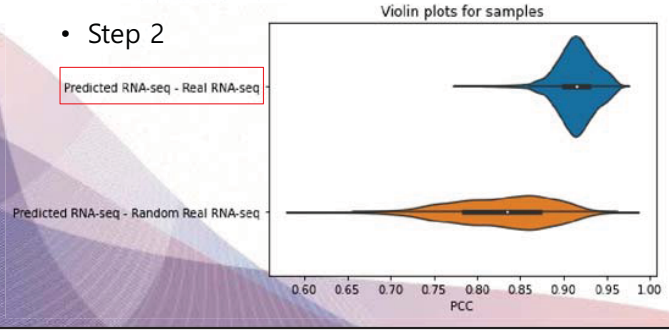
Results - GTEx Paired Samples

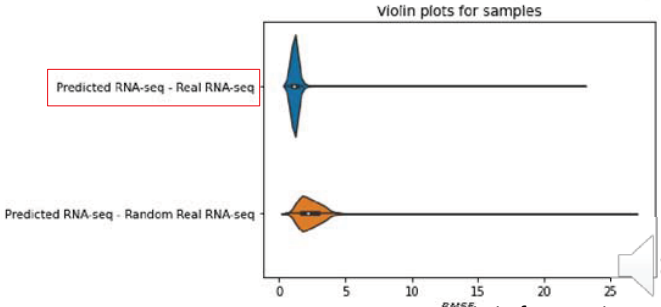
- Evaluation Metrics: The average of PCC and RMSE for each sample
- Step 1





- Step 2

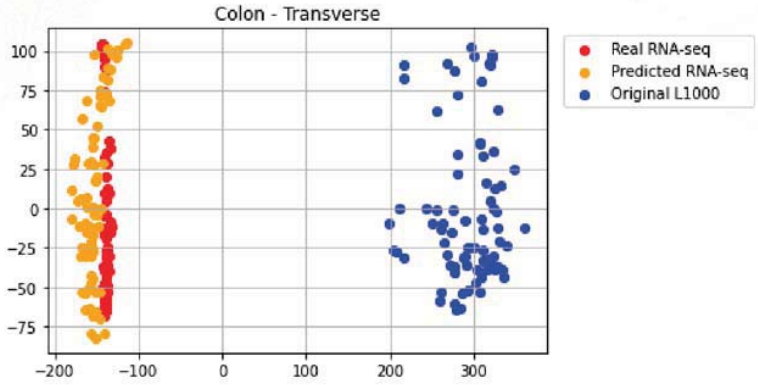




BMC Bioinformatics, 2022

Results - PCA of GTEx L1000-RNA-seq Paired Samples

- GTEx and LINCS profiled the same postmortem tissue samples using L1000 and RNA-seq
  - 2,929 samples from 53 tissues downloaded from GSE92743
- PCA plot of real RNA-seq profiles, predicted RNA-seq profiles, and original L1000 profiles fr om colon in GTEx using 11,780 common genes



BMC Bioinformatics, 2022

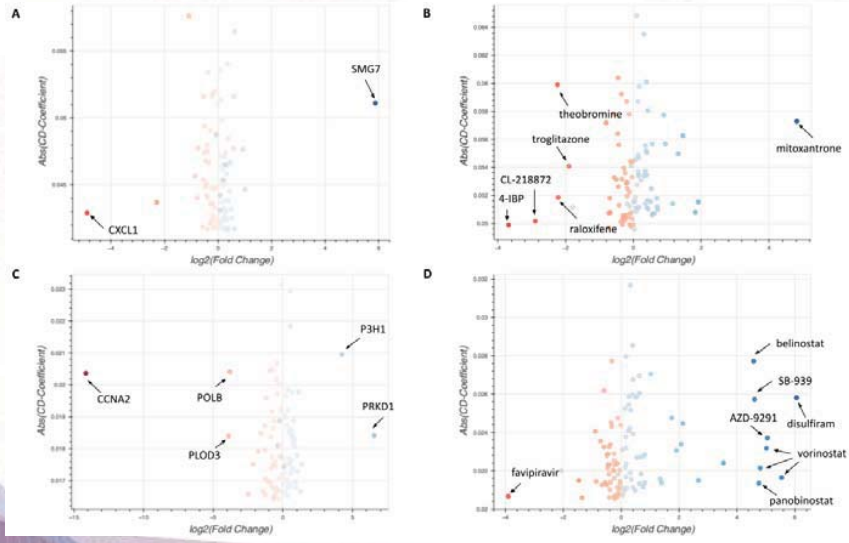
Results – Gene Centric Signature Reverse Search

KO genes

Small Molecules

SFRP2

LGI3



https://appyters.maayanlab.cloud/L1000_RNAseq_Gene_Search/

BMC Bioinformatics, 2022

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-019-13807-w OPEN

De novo generation of hit-like molecules from gene expression signatures using artificial intelligence

Oscar Méndez-Lucio[1,2*], Benoit Baillif[1], Djork-Arné Clevert[3], David Rouquié[1,5*] & Joerg Wichard[4,5*]

Finding new molecules with a desired biological activity is an extremely difficult task. In this context, artificial intelligence and generative models have been used for molecular de novo design and compound optimization. Herein, we report a generative model that bridges systems biology and molecular design, conditioning a generative adversarial network with transcriptomic data. By doing so, we can automatically design molecules that have a high probability to induce a desired transcriptomic profile. As long as the gene expression signature of the desired state is provided, this model is able to design active-like molecules for desired targets without any previous target annotation of the training compounds. Molecules designed by this model are more similar to active compounds than the ones identified by similarity of gene expression signatures. Overall, this method represents an alternative approach to bridge chemistry and biology in the long and difficult road of drug discovery.

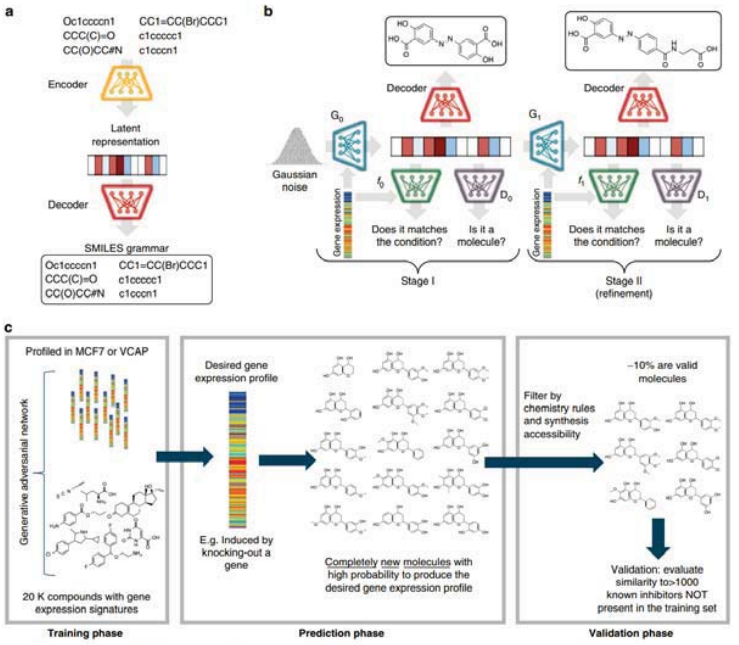


Fig. 1 Graphical representation of the models and pipeline used in the study. Molecules were encoded using a model that transforms the canonical SMILES of a molecule into a latent representation that can be later decoded into the set of grammar production rules needed to reconstruct the original SMILES (**a**). The generative adversarial network in **b** has a Stage I where the generator (G_0 in blue) takes the desired gene expression signature together with a vector of random noise to produce a molecular representation that can be decoded into SMILES using the decoder (in red). The discriminator (D_0 in purple) calculates the probability of the molecular representation to be a real molecule and the conditional network (f_0 in green) calculates the probability of the molecular representation to match the gene expression signature. In Stage II, the generator (G_1 in blue) takes as input the desired gene expression signature together with a molecular representation (e.g., the one produced by G_0) to repeat the process. The general pipeline is represented in **c** where the generative adversarial network is trained with ~20 K compounds from the L1000 dataset[25] (see Methods for details) to be able to generate compounds from a desired gene expression signature during the prediction phase.

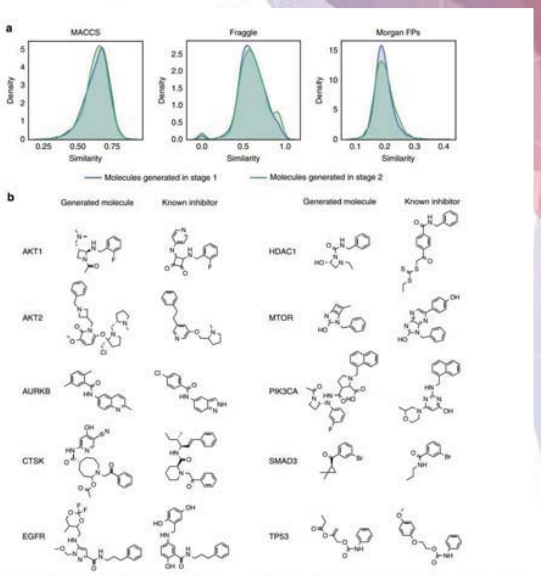


Fig. 3 Molecules generated from target knock-out gene expression signatures. a Distribution of similarity between all generated molecules and their closest active nearest neighbor using MACCS, Fraggle and Morgan Fingerprints for Stage I in blue and Stage II in green. **b** Chemical structures of some generated molecules and their closest active nearest neighbor for each of the ten different targets.

Bioinformatics, 37, 2021, i376–i382
doi: 10.1093/bioinformatics/btab275
ISMB/ECCB 2021

OXFORD

Predicting mechanism of action of novel compounds using compound structure and transcriptomic signature coembedding

Gwanghoon Jang[1], Sungjoon Park[1,*], Sanghoon Lee[1], Sunkyu Kim[1], Sejeong Park[1] and Jaewoo Kang [1,2,*]

[1]Department of Computer Science and Engineering, Korea University, Seoul, Republic of Korea and and [2]Interdisciplinary Graduate Program in Bioinformatics, Korea University, Seoul, Republic of Korea

*To whom correspondence should be addressed.

Abstract

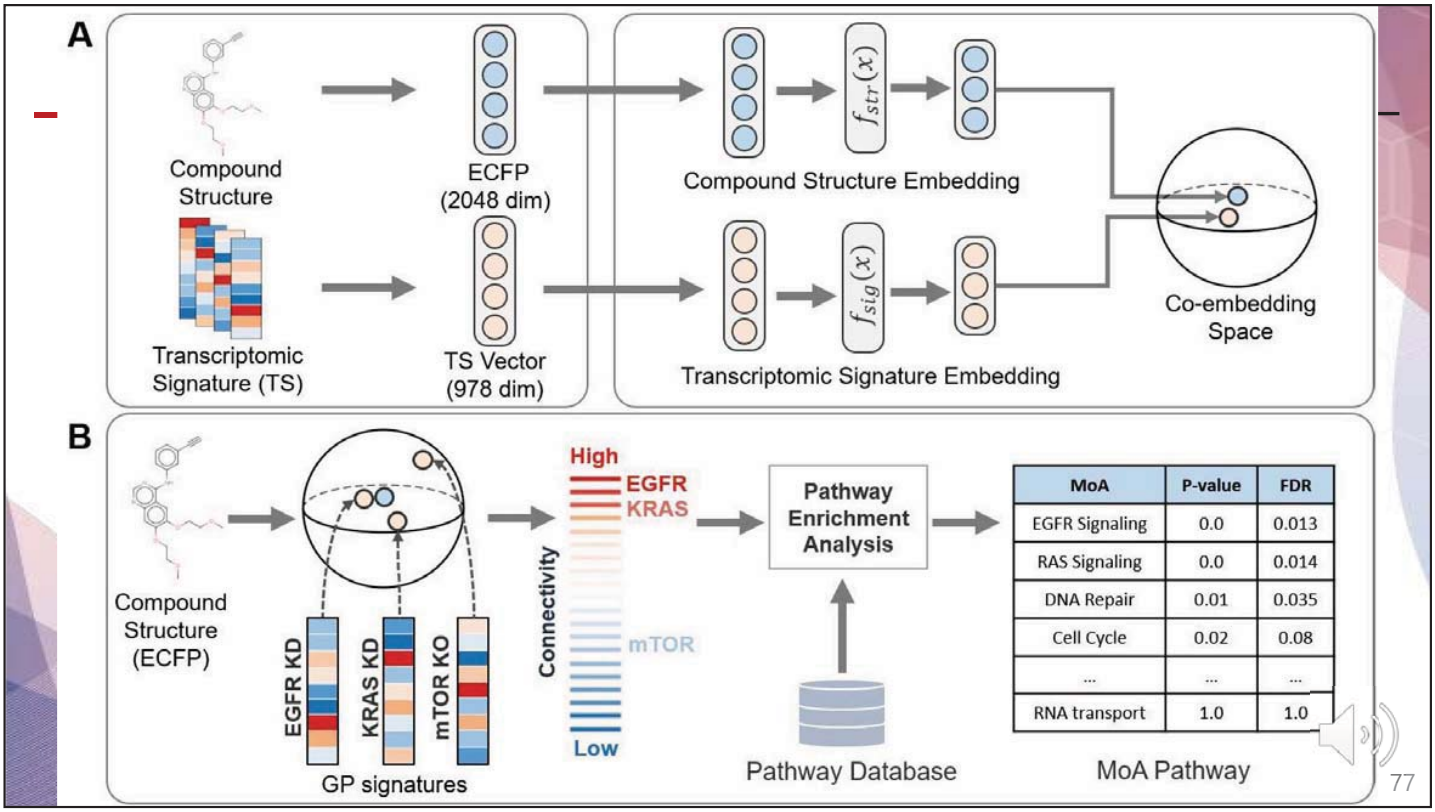
Motivation: Identifying mechanism of actions (MoA) of novel compounds is crucial in drug discovery. Careful understanding of MoA can avoid potential side effects of drug candidates. Efforts have been made to identify MoA using the transcriptomic signatures induced by compounds. However, these approaches fail to reveal MoAs in the absence of actual compound signatures.

Results: We present MoAble, which predicts MoAs without requiring compound signatures. We train a deep learning-based coembedding model to map compound signatures and compound structure into the same embedding space. The model generates low-dimensional compound signature representation from the compound structures. To predict MoAs, pathway enrichment analysis is performed based on the connectivity between embedding vectors of compounds and those of genetic perturbation. Results show that MoAble is comparable to the methods that use actual compound signatures. We demonstrate that MoAble can be used to reveal MoAs of novel compounds without measuring compound signatures with the same prediction accuracy as that with measuring them.

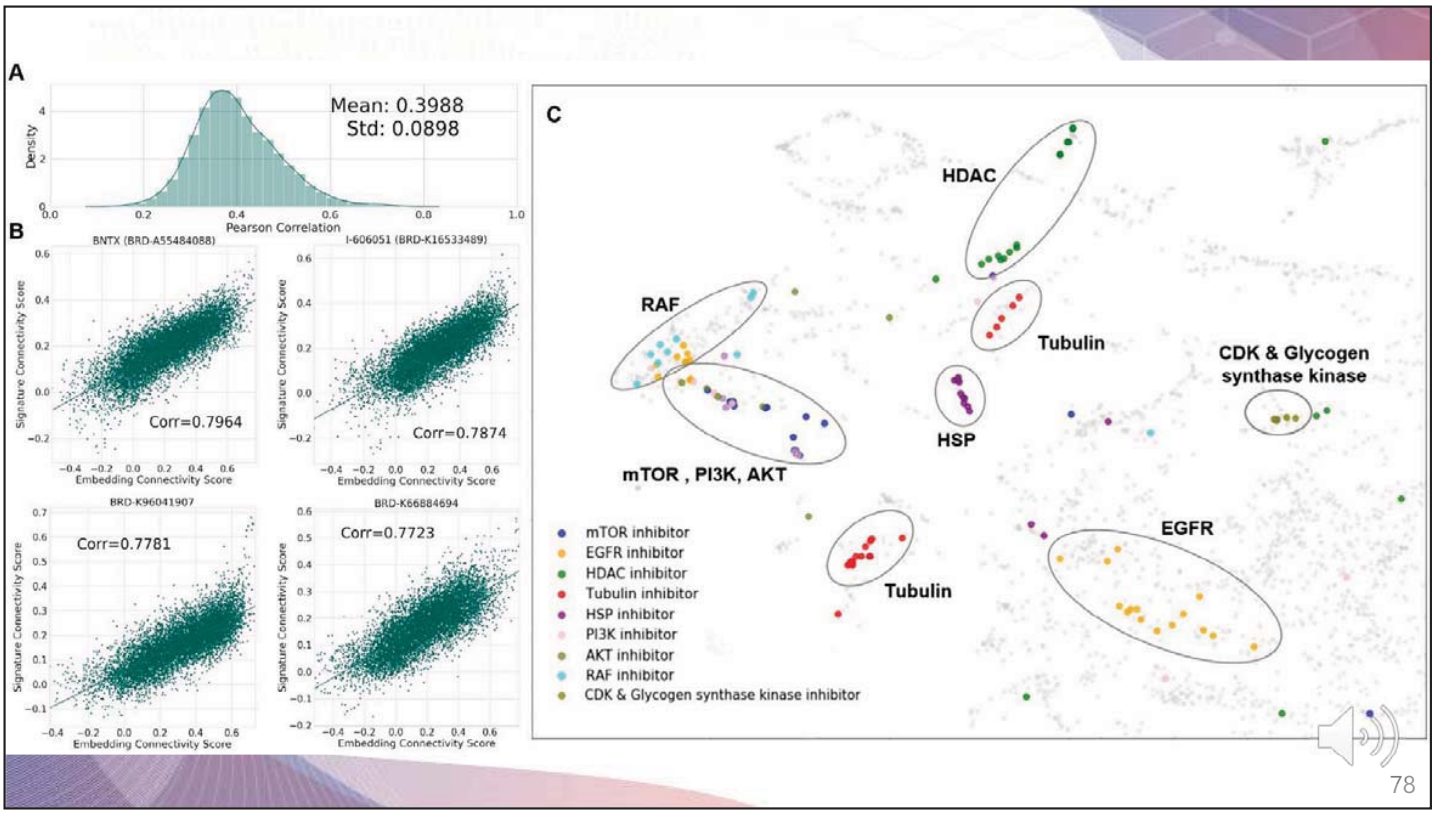
Availability and implementation: MoAble is available at https://github.com/dmis-lab/moable

Contact: : sungjoonpark@korea.ac.kr or kangj@korea.ac.kr

Supplementary information: Supplementary data are available at *Bioinformatics* online.



MoA	P-value	FDR
EGFR Signaling	0.0	0.013
RAS Signaling	0.0	0.014
DNA Repair	0.01	0.035
Cell Cycle	0.02	0.08
...	...	...
RNA transport	1.0	1.0



DeepSide: A Deep Learning Approach for Drug Side Effect Prediction

Onur Can Uner, Halil Ibrahim Kuru, R. Gokberk Cinbis, Oznur Tastan, and A. Ercument Cicek

Abstract—Drug failures due to unforeseen adverse effects at clinical trials pose health risks for the participants and lead to substantial financial losses. Side effect prediction algorithms have the potential to guide the drug design process. LINCS L1000 dataset provides a vast resource of cell line gene expression data perturbed by different drugs and creates a knowledge base for context specific features. The state-of-the-art approach that aims at using context specific information relies on only the high-quality experiments in LINCS L1000 and discards a large portion of the experiments. In this study, our goal is to boost the prediction performance by utilizing this data to its full extent. We experiment with 5 deep learning architectures. We find that a multi-modal architecture produces the best predictive performance among multi-layer perceptron-based architectures when drug chemical structure (CS), and the full set of drug perturbed gene expression profiles (GEX) are used as modalities. Overall, we observe that the CS is more informative than the GEX. A convolutional neural network-based model that uses only SMILES string representation of the drugs achieves the best results and provides 13.0% macro-AUC and 3.1% micro-AUC improvements over the state-of-the-art. We also show that the model is able to predict side effect-drug pairs that are reported in the literature but was missing in the ground truth side effect dataset. DeepSide is available at http://github.com/OnurUner/DeepSide.

Index Terms—Drug side effect prediction, deep learning, LINCS

2020 IEEE International Conference on Big Data and Smart Computing (BigComp)

A Drug-induced Liver Injury Prediction Model using Transcriptional Response Data with Graph Neural Network

Doyeong Hwang*‡, Minji Jeon*‡, Jaewoo Kang*†§
*Department of Computer Science and Engineering
†Interdisciplinary Graduate Program in Bioinformatics
Korea University
Seoul, Republic of Korea
Emails: {desertbeagle, mjjeon, kangj}@korea.ac.kr
‡Both authors contributed equally
§Corresponding author

***Abstract*—Drug-Induced Liver Injury (DILI) is a major cause of failed drug candidates in clinical trials and withdrawal of approved drugs from the market. Therefore, machine learning-based DILI prediction can be key in increasing the success rate of drug discovery because drug candidates that are predicted to potentially induce liver injury can be rejected before clinical trials. However, existing DILI prediction models mainly focus on the chemical structures of drugs. Since we cannot determine whether a drug will cause liver injury based solely on its structure, DILI prediction based on the transcriptional effect of a drug on a cell is necessary.**

Several machine learning models trained on datasets such as Liver Toxicity Knowledge Base (LTKB) [4] and Open TG-GATEs [5] were previously proposed for DILI prediction. Most of the previously proposed machine learning models are trained on drug structure information for predicting DILI [6]–[9]. However, such models used for DILI prediction do not consider genetic information or the structures and complex biological mechanisms of drugs [10]. Therefore, these models cannot predict whether a drug will cause liver injury based solely on its structure

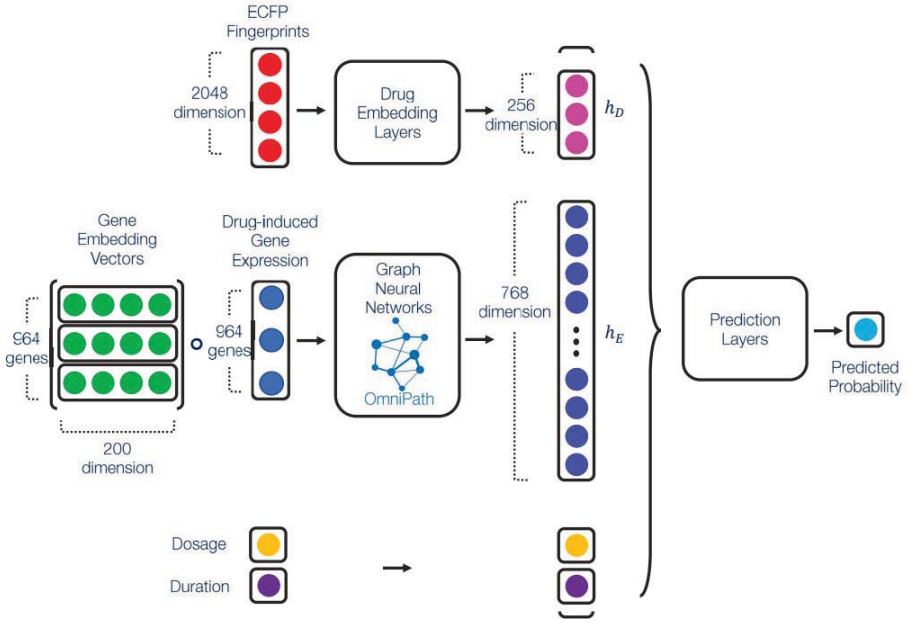


Fig. 1. The overall architecture of GLIT. The structure of a drug, the drug-induced gene expression level, and the dosage and duration of drug administration are used as inputs for GLIT. A drug embedding vector is extracted from the last layer of a neural network with three layers, and a drug-induced gene expression embedding vector is extracted from a graph neural network using a biological knowledge graph called OmniPath. The two embedding vectors are concatenated with the dosage and duration of drug administration information, and fed to the prediction layers of GLIT to predict DILI.

81

Kim and Nam *Journal of Cheminformatics* (2022) 14:9
https://doi.org/10.1186/s13321-022-00589-5

Journal of Cheminformatics

RESEARCH ARTICLE Open Access

DeSIDE-DDI: interpretable prediction of drug-drug interactions using drug-induced gene expressions

Eunyoung Kim and Hojung Nam*

Abstract

Adverse drug-drug interaction (DDI) is a major concern to polypharmacy due to its unexpected adverse side effects and must be identified at an early stage of drug discovery and development. Many computational methods have been proposed for this purpose, but most require specific types of information, or they have less concern in interpretation on underlying genes. We propose a deep learning-based framework for DDI prediction with drug-induced gene expression signatures so that the model can provide the expression level of interpretability for DDIs. The model engineers dynamic drug features using a gating mechanism that mimics the co-administration effects by imposing attention to genes. Also, each side-effect is projected into a latent space through translating embedding. As a result, the model achieved an AUC of 0.889 and an AUPR of 0.915 in unseen interaction prediction, which is competitively very accurate and outperforms other state-of-the-art methods. Furthermore, it can predict potential DDIs with new compounds not used in training. In conclusion, using drug-induced gene expression signatures followed by gating and translating embedding can increase DDI prediction accuracy while providing model interpretability. The source code is available on GitHub (https://github.com/GIST-CSBL/DeSIDE-DDI).

Keywords: Drug-drug interaction, Polypharmacy side effects, In silico prediction, Deep learning

82

NeurIPS 2022

Predicting Cellular Responses to Novel Drug Perturbations at a Single-Cell Resolution

Leon Hetzel[*1,3], **Simon Böhm**[*3], **Niki Kilbertus**[2,4],
Stephan Günnemann[2], **Mohammad Lotfollahi**[1,5], **and Fabian Theis**[1,3]

{leon.hetzel, simon.boehm, niki.kilbertus}@helmholtz-muenchen.de
s.guennemann@tum.de, {mohammad.lotfollahi, fabian.theis}@helmholtz-muenchen.de

[1]Department of Mathematics, Technical University of Munich
[2]Department of Computer Science, Technical University of Munich
[3]Helmholtz Center for Computational Health, Munich
[4]Helmholtz AI, Munich
[5] Wellcome Sanger Institute, Cambridge

Abstract

Single-cell transcriptomics enabled the study of cellular heterogeneity in response to perturbations at the resolution of individual cells. However, scaling high-throughput screens (HTSs) to measure cellular responses for many drugs remains a challenge due to technical limitations and, more importantly, the cost of such multiplexed experiments. Thus, transferring information from routinely performed bulk RNA HTS is required to enrich single-cell data meaningfully. We introduce chemCPA, a new encoder-decoder architecture to study the perturbational effects of unseen drugs. We combine the model with an architecture surgery for transfer learning and demonstrate how training on existing bulk RNA HTS datasets can improve generalisation performance. Better generalisation reduces the need for extensive and costly screens at single-cell resolution. We envision that our proposed method will facilitate more efficient experiment designs through its ability to generate in-silico hypotheses, ultimately accelerating drug discovery.

84

(1) Encoder-Decoder: (2) Attribute embeddings: (3) Adversarial classifiers:

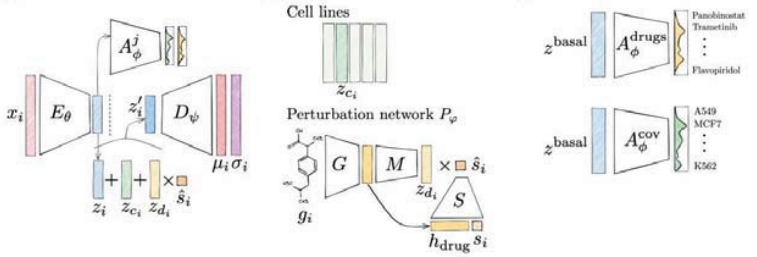


Figure 1: Architecture of chemCPA. The model consists of three parts: (1) the encoder-decoder architecure, (2) the attribute embeddings, and (3) the adversarial classifiers. The molecule encoder G can be any graph- or language-based model as long as it generates fixed-sized embeddings h_{drugs}. The MLPs S and M are trained to map the embeddings to the perturbational latent space. There, z_{d_i} is added to the basal state z_i and the covariate embedding z_{c_i}. In this work, the latter always corresponds to cell lines. The basal state $z_i = E_\theta(x_i)$ is trained to be invariant through adversarial classifiers A_ϕ^j and the decoder D_ψ gives rise to the Gaussian likelihood $\mathcal{N}(x_i \mid \mu_i, \sigma_i)$.

Table 1: Comparison of multiple models on their performance on generalisation to unseen drug-covariate combinations for dosage values of 1 μM and 10 μM.

Dose	Model	E[r^2] all	E[r^2] DEGs	Median r^2 all	Median r^2 DEGs
1 μM	Baseline	0.69	0.51	0.82	0.62
	scGen	0.73	0.59	0.77	0.68
	CPA	0.72	0.54	**0.86**	0.67
	chemCPA	0.74	0.60	**0.86**	0.66
	chemCPA pretrained	**0.77**	**0.68**	0.85	**0.76**
10 μM	Baseline	0.50	0.29	0.48	0.12
	scGen	0.62	0.47	0.66	0.49
	CPA	0.54	0.34	0.52	0.26
	chemCPA	0.71	0.58	0.77	0.64
	chemCPA pretrained	**0.76**	**0.68**	**0.82**	**0.79**

감사합니다

mjjeon@korea.ac.kr
https://medai.korea.ac.kr

2강 실습 링크1

https://colab.research.google.com/drive/1dryAvI-OyQ_XoodhcpIkzAM0syb4F5AC?usp=sharing

2강 실습 링크2

https://colab.research.google.com/drive/1YBuRAZ5TwcJ4Lq5zeUOG9j50SMSibMM8?usp=sharing

2강 실습 데이터 다운로드 링크

https://drive.google.com/file/d/1Olc2FvdGlJC2G1SwHjb5DstD3ozlB11b/view?usp=sharing

KSBi-BIML 2024

Introduction to ConnectivityMap

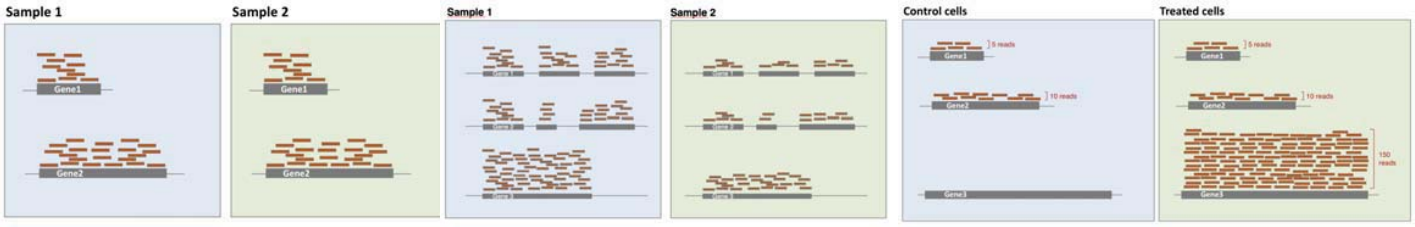
고려대학교 의과대학 전민지

DESeq2

- Models estimate parameters (mean and dispersion) to describe count data distribution accurately.
- Dispersion parameter α affects variance; edgeR and DESeq2 employ similar empirical Bayes methods to shrink α towards similar-gene dispersions, enhancing differential expression test results.
- DESeq2 workflow:
  - Normalize read counts by computing size factors, addressing differences in library sizes and library compositions.
  - Calculate dispersion estimate for each gene.
  - Plot dispersion estimates of genes against mean normalized counts, and fit a line.
  - Shrunk dispersion values of each gene towards the fitted line.
  - A Generalized Linear Model accounts for confounding variables and negative binomial distribution, fitting count data.
  - For a contrast (e.g., drug-A treated vs. untreated), differential expression test assesses l og fold change of normalized gene counts.
  - P-values are adjusted for multiple testing.

DESeq2: Normalization

- DESeq2 expects as an input a matrix of raw counts (un-normalized counts).
- These counts are supposed to reflect gene abundance (what we are interested in)
- But they are also dependent on other less interesting factors such as gene length, sequencing biases, sequencing depth or library composition

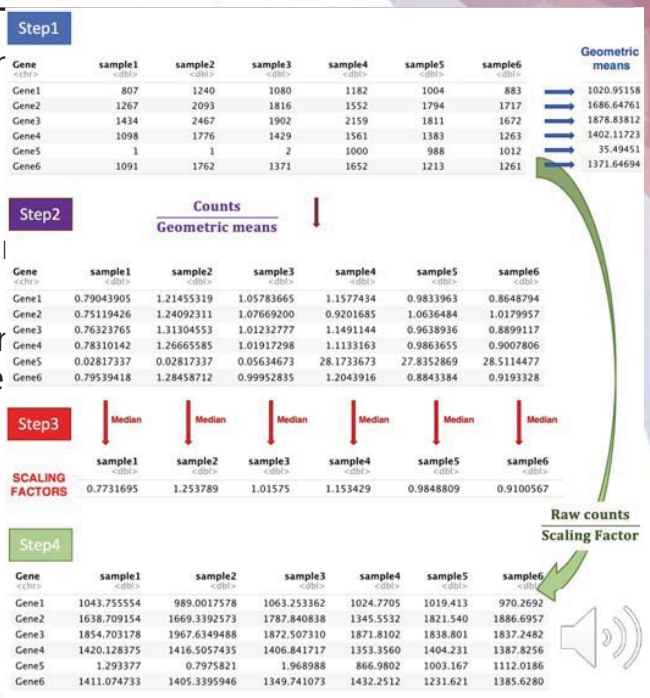


Gene length | Sequencing depth | Library composition

DESeq2: Normalization

- Step 1: DESeq2 creates a pseudo-referer wise geometric mean (for each gene).
- Step 2: For every gene in every sample, sample are calculated.
- Step 3: The median value of all ratios fo scale factor for that sample.
- Step 4: Normalized counts can be obtair values in a given sample by that sample

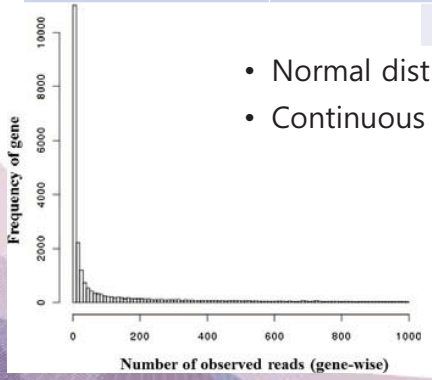
$$\hat{s}_j = \underset{i}{\text{median}} \frac{k_{ij}}{\left(\prod_{v=1}^{m} k_{iv}\right)^{1/m}}.$$



DESeq2: Count Modeling

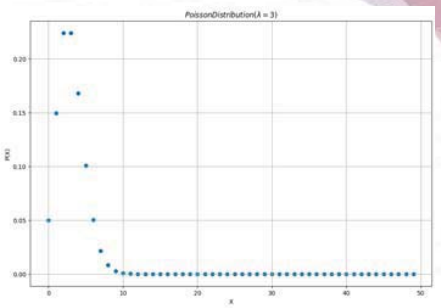
Count matrix K

	Sample1	Sample2	Samp
Gene1			
Gene2			
Gene3			



- Normal distribution (X)
- Continuous (X)

Fit Poisson distribution



- Number of cases is large
- Probability of an event happening is low
- Selecting mRNA from a large number of mRNA read
- mean = variance

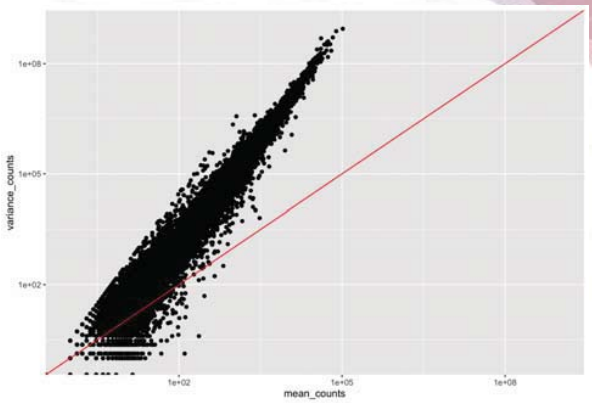
DESeq2: Count Modeling

- But RNA-seq data does not fit Poisson distribution
  - variance ≠ mean
- Negative Binomial distribution
  - variance > mean

$$K_{ij} \sim NB(mean = \mu_{ij}, dispersion = \alpha_i)$$

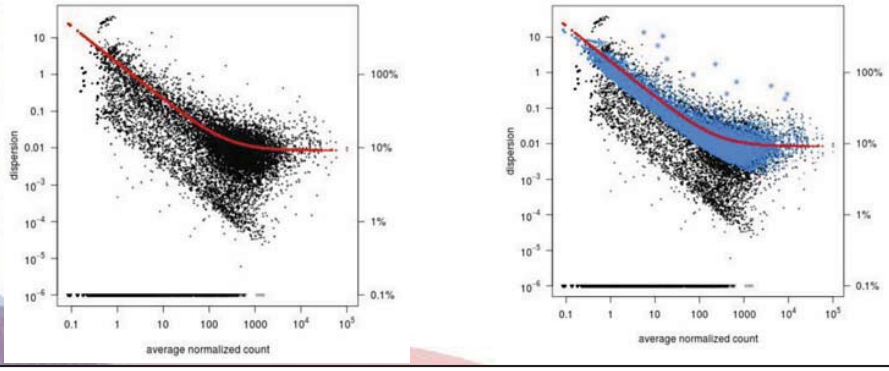
$$VAR(K_{ij}) = \mu_{ij} + \alpha_i . \mu_{ij}^2$$

K_ij = count of gene i for sample j



DESeq2: Count Modeling

- Dispersion estimation
  - Dispersion: When comparing gene expression levels between groups, it is important to know also its within-group variability
  - RNA-seq experiments typically have only few replicates
  - It is difficult to estimate within-group variability
  - Solution: pool information across genes with are expressed at similar level
    - assumes that genes with similar average expression strength have similar dispersion



DESeq2: Generalized linear model

- Generalized linear model:

$$log2(q_{ij}) = \beta_0 + \beta_1 . x_j + \epsilon$$

$$q_{ij} = \frac{\mu_{ij}}{SizeFactor_j}$$

β_0 is the log2 expression level in the reference (control samples)

β_1 is the log2FC between treated and control cells

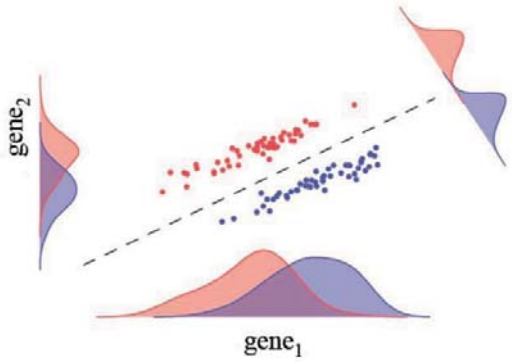
x_j = 0 if sample j is the control sample

x_j = 1 if sample j is the treated sample

8

Characteristic Direction

- Characteristic Direction (Clark et al. 2014, BMC bioinformatics)
  - Genes do not function in isolation but as part of a complex network of interactions
    - This leads to significant correlations
  - Univariate approaches can miss some structure in the data
  - Multivariate approaches are sensitive to the curse of dimensionality

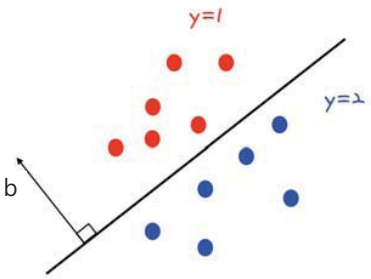


Characteristic Direction

- Linear discriminant analysis
  - Bayes rules for the classification probability
  - The contribution of each b can be interpreted as quantifying the relative contribution of each component to the total differential expression giving the significance of the corresponding gene

Normal vector = direction of characteristics in gene expression data

$$\sum_{i=1}^{p} \hat{b}_i^2 \equiv 1$$



Linear classification boundary

10