KSBi-BIML 2021

Bioinformatics & Machine Learning (BIML)
Workshop for Life Scientists

생물정보학 & 머쉰러닝 워크샵(온라인)

Single-cell Network Biology

이인석

SBi 한국생명정보학회 Korean Society for Bioinformatics
서울대학교 생물정보연구소 SNU BIOINFORMATICS INSTITUTE
Gene·2·us

Bioinformatics & Machine Learning for Life Scientists

BIML-2021

안녕하십니까?

한국생명정보학회의 동계 워크샵인 BIML-2021을 2월 15부터 2월 19일까지 개최합니다. 생명정보학 분야의 융합이론 보급과 실무역량 강화를 위해 도입한 전문 교육 프로그램인 BIML 워크샵은 2015년에 시작하였으며 올해로 7차를 맞이하게 되었습니다. 유례가 없는 코로나 대유행으로 인해 올해의 BIML 워크숍은 온라인으로 준비했습니다. 생생한 현장 강의에서만 느낄 수 있는 강의자와 수강생 사이의 상호교감을 가질 수 없다는 단점이 있지만, 온라인 강의의 여러 장점을 살려서 최근 생명정보학에서 주목받고 있는 거의 모든 분야를 망라한 강의를 준비했습니다. 또한 온라인 강의의 한계를 극복하기 위해서 실시간 Q&A 세션 또한 마련했습니다.

BIML 워크샵은 전통적으로 크게 생명정보학과 AI, 두 개의 분야로 구성되어오고 있으며 올해 역시 유사한 방식을 채택했습니다. AI 분야는 Probabilistic Modeling, Dimensionality Reduction, SVM 등과 같은 전통적인 Machine Learning부터 Deep Learning을 이용한 신약개발 및 유전체 연구까지 다양한 내용을 다루고 있습니다. 생명정보학 분야로는, Proteomics, Chemoinformatics, Single Cell Genomics, Cancer Genomics, Network Biology, 3D Epigenomics, RNA Biology, Microbiome 등 거의 모든 분야가 포함되어 있습니다. 연사들은 각 분야 최고의 전문가들이라 자부합니다.

이번 BIML-2021을 준비하기까지 너무나 많은 수고를 해주신 BIML-2021 운영위원회의 김태민 교수님, 류성호 교수님, 남진우 교수님, 백대현 교수님께 커다란 감사를 드립니다. 또한 재정적 도움을 주신, 김선 교수님 (AI-based Drug Discovery), 류성호 교수님, 남진우 교수님께 감사를 표시하고 싶습니다. 마지막으로 부족한 시간에도 불구하고 강의 부탁을 흔쾌히 허락하시고 훌륭한 강의자료를 만드는데 노력하셨을 뿐만 아니라 실시간 온라인 Q&A 세션까지 참여해 수고해 주시는 모든 연사분들께 깊이 감사드립니다.

2021년 2월

한국생명정보학회장 김동섭

강의개요

Single-cell Network Biology

최근 급속히 발전하고 있는 단일세포오믹스(single-cell omics) 기술들은 유전체연구의 패러다임을 바꾸고 있다. 특히 단일세포 수준에서 전사체 및 후성유전체의 활성 정보는 다양한 세포들이 섞여있는 조직(tissue) 및 기관(organ) 내에 존재하는 세포들의 유형별 기능과 이들 사이의 기능적 상호관계를 더 정확하게 이해할 수 있는 기회를 제공하고 있다.
단일세포오믹스는 cellular heterogeneity의 문제를 해결하였을 뿐 아니라 세포유형특이적(celltype-specific) 혹은 개인별(personal) 유전자조절네트워크(gene regulatory network or GRN)의 모델링을 가능하게 할 수 있다. 다차원 유전자 발현 데이터에 존재하는 변이(variance)를 기반으로 하는 기존 알고리즘들을 bulk RNA sequencing 정보에 적용해 GRN을 구축하기 위해서는 연구대상 샘플에 대한 많은 수의 transcriptome profiling을 수행해야 했다. 하지만 single-cell RNA sequencing (scRNA-seq)은 단일 실험에 수백-수천의 세포에 대한 transcriptome profile data를 생산하여 cell-to-cell variance를 이용한 GRN 구축이 이론적으로 가능하다. 그러므로 각 개인별 세포유형특이적인 GRN을 구축하여 보다 높은 해상도로 주어진 세포 환경에 보다 특이적인 유전자 조절 프로그램을 연구할 수 있다. 본 강좌는 scRNA-seq 데이터로부터 GRN을 구축 및 해석할 수 있는 능력을 배양하도록 도와줄 것이다.

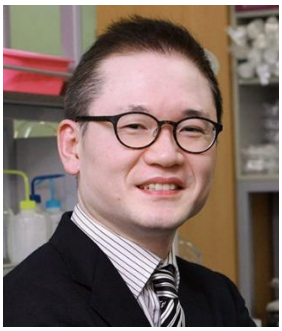
***참고강의교재:**

***교육생준비물:**

*** 강의: 이인석 교수 (연세대학교 생명공학과)**

Curriculum Vitae

Speaker Name: Insuk Lee, Ph.D.



▶Personal Info

Name Insuk Lee

Title Professor

Affiliation Yonsei University

▶Contact Information

Address 50 Yonsei-ro, Seodaemun-gu, Seoul, Korea

Email insuklee@yonsei.ac.kr

Phone Number 02-2123-5559

Research interest : Single-cell biology, Cancer immunology, Metagenomics, Human gut microbiome, Network biology, Biological data mining

Educational Experience

1993 B.S. in Biology, Hanyang University, Korea

2002 Ph.D. in Microbiology, University of Texas at Austin, USA

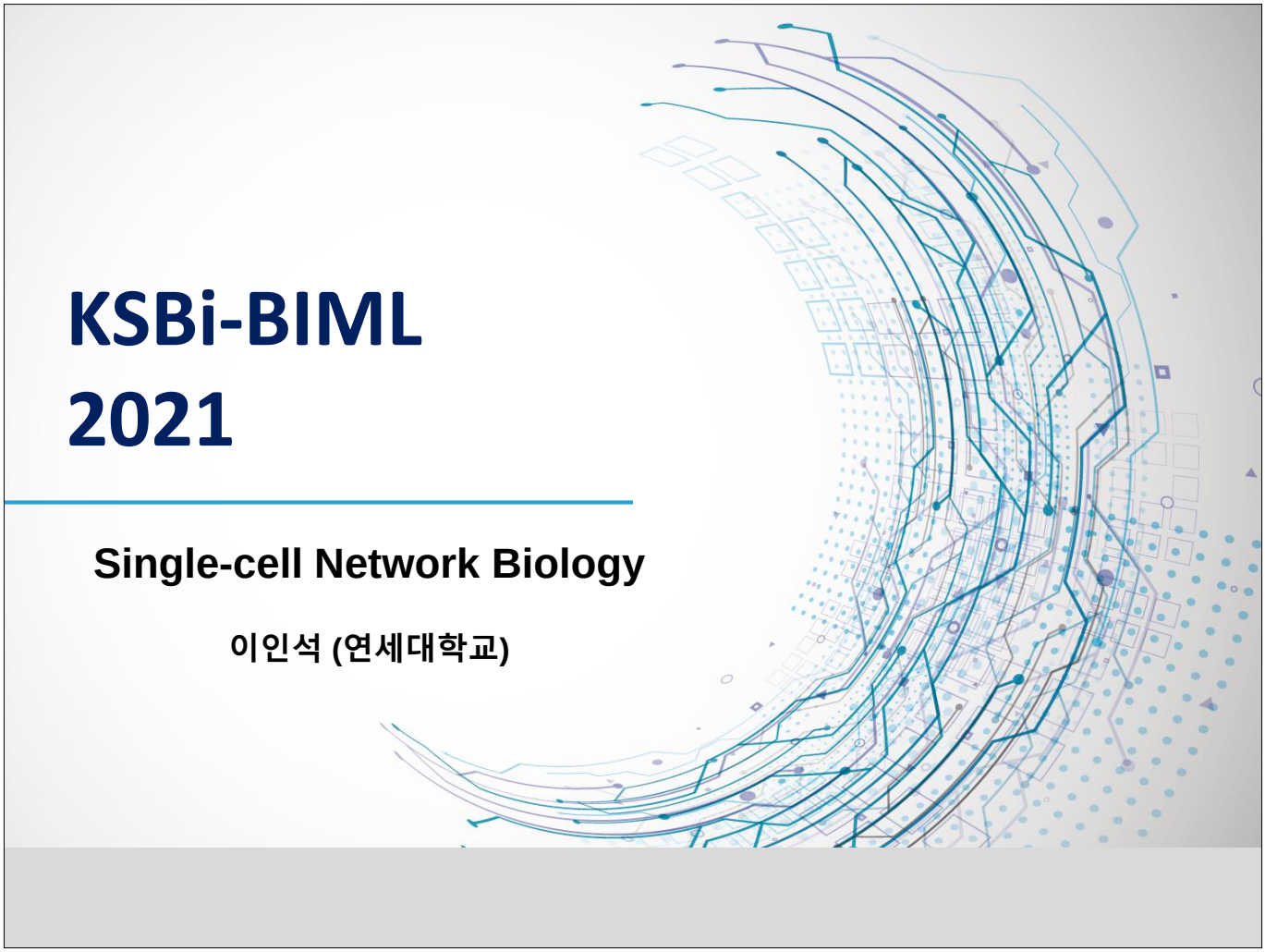
Professional Experience

2002-2008 Postdoc Fellow and Research Associate, University of Texas at Austin, USA

2008-Present Assistant/Associate/Full Professor, Yonsei University, Korea

Selected Publications (5 maximum)

1. **Junha Cha**, **Insuk Lee** Single-cell Network Biology for Resolving Cellular Heterogeneity in Human Diseases ***Experimental & Molecular Medicine*** 2020 Nov;52(11):1798-1808
2. Jimin Son, **Jae-Won Cho**, Hyo Jin Park, Jihyun Moon, Seyeon Park, Hoyoung Lee, Jeewon Lee, Ga min Kim, Su-Myeong Park, Sergio A. Lira, Andrew N. Mckenzie, Hye Young Kim, Cheol Yong Choi, Yong Taik Lim, Seong Yong Park, Hye Ryun Kim, Su-Hyung Park, Eui-Cheol Shin, **Insuk Lee** & Sang-Jun Ha, Tumor-Infiltrating Regulatory T Cell Accumulation in the Tumor Microenvironment is Mediated by IL33/ST2 Signaling ***Cancer Immunology Research*** 2020 Nov; 8(11):1393-1406
3. **Seungbyn Baek**, **Insuk Lee** Single-cell ATAC sequencing analysis: from data preprocessing to hypothesis generation ***Computational and Structural Biotechnology Journal*** 2020 June 28;18:1429-1439
4. **Kyungsoo Kim**, Seyeon Park*, Seong Yong Park, Gamin Kim, Su Myeong Park, Jae-Won Cho, Da Hee Kim, Young Min Park, Yoon Woo Koh, Hye Ryun Kim, Sang-Jun Ha** and **Insuk Lee**, Single-cell transcriptome analysis reveals TOX as a promoting factor for T cell exhaustion and a predictor for anti-PD-1 responses in human cancer ***Genome Medicine*** 2020 Feb 28;12:22
5. **Kyungsoo Kim**, Sunmo Yang, Sang-Jun Ha, **Insuk Lee**, VirtualCytometry: a webserver for evaluating immune cell differentiation using single-cell RNA sequencing data ***Bioinformatics*** 2020 Jan 15;36(2):546-551



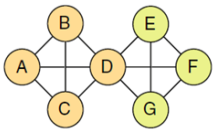
본 강의 자료는 한국생명정보학회가 주관하는 KSBi-BIML 2021 워크샵 온라인 수업을 목적으로 제작된것으로 해당 목적 이외의 다른 용도로 사용할 수 없음을 분명하게 알립니다. 수업 목적으로 배포 및 전송 받은 경우에도 이를 다른 사람과 공유하거나 복제, 배포, 전송할 수 없습니다.

만약 이러한 사항을 위반할 경우 발생하는 모든 법적 책임은 전적으로 불법 행위자 본인에게 있음을 경고합니다.

❖ Why single-cell network biology?

- The major aim of single-cell biology is **understanding cellular heterogeneity**.
- **Cell-type-specific phenotypes** are governed by activity of **cell-type-specific regulators** and target programs (i.e. **cell-type specific gene expression programs**).
- Gene expression programs are mainly **mediated by transcriptional regulatory programs** (=**transcriptional regulatory networks**).
- Therefore, **deciphering gene regulatory network (GRN) will facilitate understanding cellular heterogeneity**.
- Currently we can link only ~60% of the **disease-associated non-coding SNPs** in regulatory elements to an eQTL effect. Because many of the eQTL may have **cell-type-specific regulatory effects** (*eQTL analysis was traditionally conducted with tissue samples*).
- Therefore, **understanding mechanisms of disease genetics** needs **cell-type-specific GRN.**
- Furthermore, **single-cell gene expression data** allow to **generate GRN for individuals** (**personalized GRN**), which may facilitate implementation of **precision medicine** in the future.

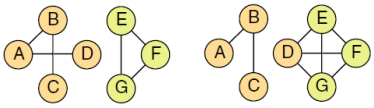
Integrated network for heterogeneous cell-types and population.



Specifying context

Single-cell data for each cell-type and individual.

Cell-type-specific and personalized GRNs



❖ Network inference from single-cell transcriptome data

➢ **Pros**

- Larger numbers of data points (>1000's cells in general) yield **higher statistical power**.
- Cell-type-specific transcriptome data contains signals **based on cell-state variation** that provides cell-type-specific pathway links **more than compositional variation**.
- Regulatory relationship can be inferred by cell-to-cell variation **within a single person (i.e. personal network)**.

➢ **Cons**

- High **noise** and **sparsity** (dropout): cause high proportion of false positive links

❖ Types of single-cell gene networks

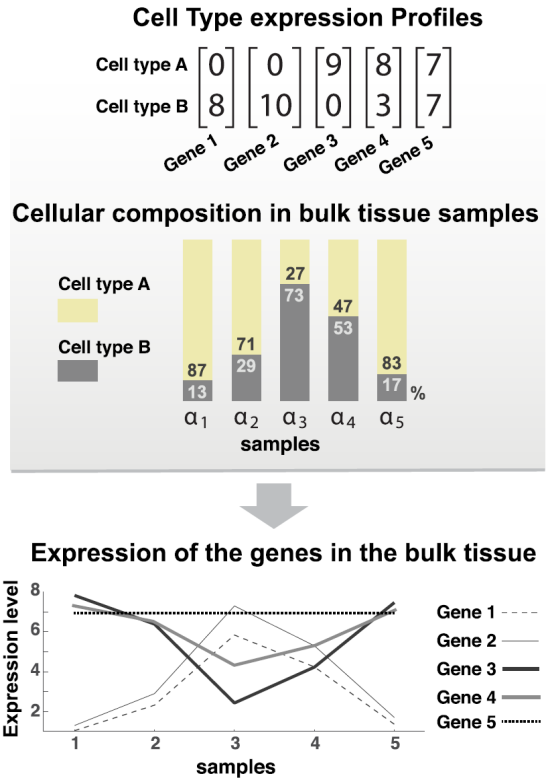
- Single-cell **Gene Regulatory Network (directional links)**
- Single-cell **Co-regulatory (or co-expression) Network (peer-to-peer links)**

❖ Approaches to single-cell network inference

- **Statistical approaches**: Correlation; Mutual information
- **Machine learning**: Random forest; Gradient boosting machine
- **Data preprocessing**: Dropout imputation; Transformation

❖ Cell-state variation vs compositional variation with bulk tissue RNA-seq data

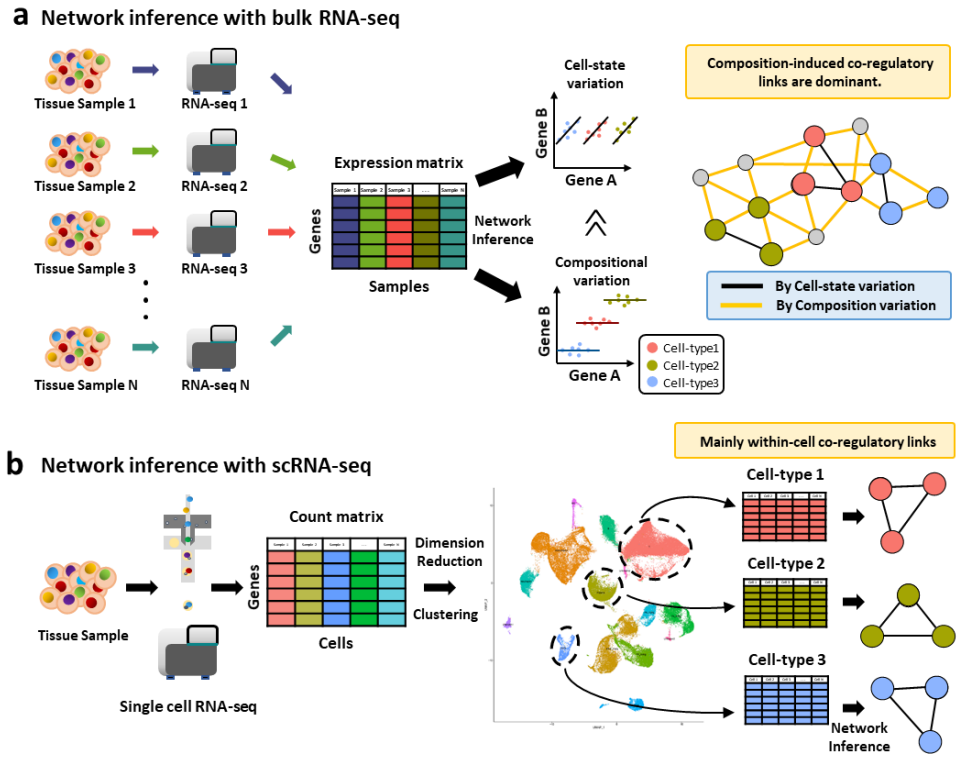
- Schematic of cellular composition effects on gene expression variance in bulk tissue.
- Top: Cell type (CT) profiles for five genes in a hypothetical tissue with two cell types. Genes 1 and 2 are marker genes for cell type B. Gene 3 is a marker gene for cell type A. Gene 4 is expressed in both cell types but at different levels, whereas Gene 5 is expressed at equal levels.
- Middle: Hypothetical cellular compositions of five bulk tissue samples. Each sample αi has the same amount of biological material but different proportions of each cell type.
- Bottom: The expected observed expression levels. Genes 1 and 2 are positively correlated and negatively correlated with Genes 3 and 4. Gene 5 is expressed at the same level in all the bulk tissue samples as it is equally expressed in all cell types.



- **Genes that have similar expression patterns across cell types will have correlated RNA levels in bulk tissue**, due to the effect of variation in cellular composition.
- **Much bulk tissue expression is explained by cellular composition variation among samples**, rather than intra-cell-type regulatory relationships.
- Dominant **cellular composition-induced co-expression** mask underlying **within-cell co-regulatory links** in bulk RNA-seq data.
- Thus, we may **need co-expression analysis** using scRNA-seq data **to map within-cell coregulatory links**.

Genome Res. 30:849 (2020)

❖ Network inference with bulk RNA-seq vs scRNA-seq



(a) In network inference with **bulk RNA-seq**, **correlation between genes by variation of cell-type composition** across tissue samples is **dominant**. Thus, network is mostly composed of cell-type composition-induced co-expression.

(b) In network inference with **scRNA-seq**, using gene-by-cell count matrix for each cell type, we can infer **networks mainly composed of within-cell co-regulatory links**.

Cha and Lee. Exp. Mol. Med. (2020)

❖ Single-cell Gene Regulatory Network from scRNA-seq data

- GRN inference from transcriptome data relies on the **assumption that regulatory information can be extracted from the expression pattern**.

➢ GRN inference methods using pseudotime ordered cells

- Most GRN modeling methods developed for scRNA-seq data requires the cells to be ordered by pseudotime in the input data.
- For example, LEAP (*Bioinformatics 33:764, 2017*) applies Pearson's correlation over temporal window of a fixed size with different time lags.
- Others are SINCERITIES (*Bioinformatics 34:258, 2018*), SCODE (*Bioinformatics 33:2314, 2017*)
- Then, network inferences will rely on the quality of pseudotime analysis of scRNA-seq data: "***robustness issue***".

➢ GRN inference methods based on Boolean models

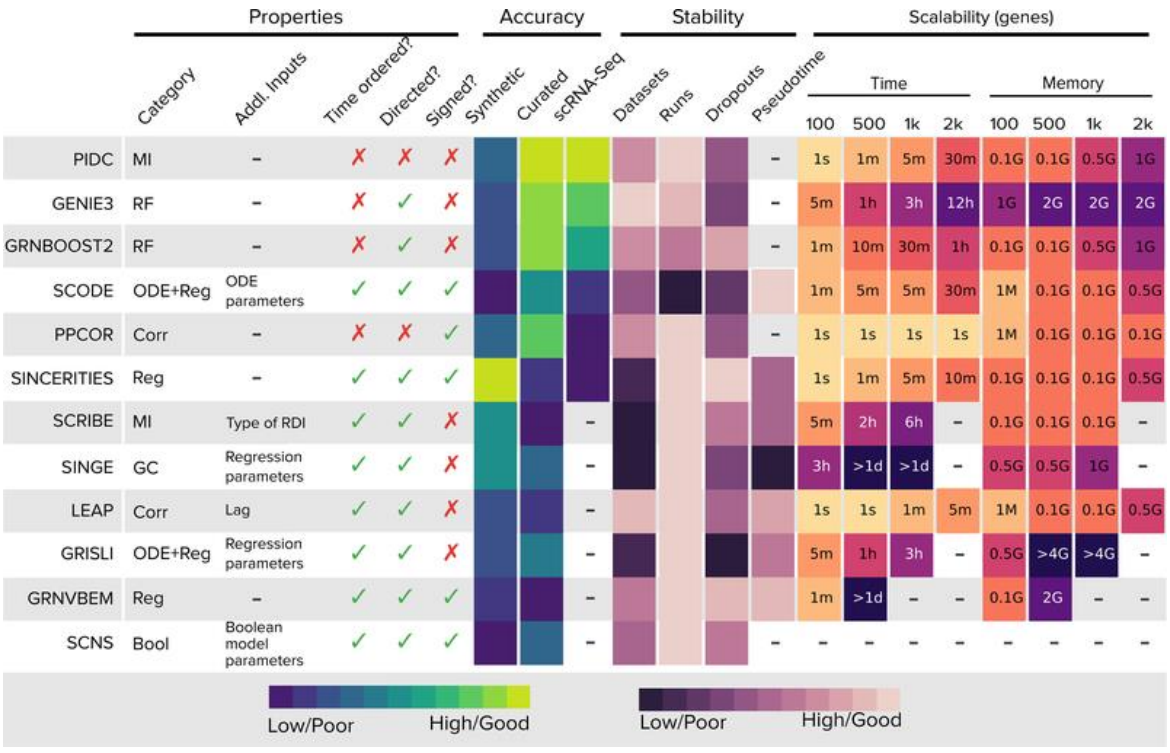
- **Boolean models** focus on **logical combination of TFs** required to transit from one state to another in dynamic process, resulting in **state-graph for key TFs involved in state changes**.
- However, it does **not provide target information** and computational demands increase rapidly with network size because of high-dimensional parameter spaces. (thus generally used for network with <100 genes): "***scalability issue***"

✓ We prefer GRN inference methods which are **robust** and **scalable** to any single-cell transcriptome data: **Partial correlation** (calculated by R package PPCOR)**, PIDC, GENIE3 and GRNBoost**

➢ Benchmarking GRN inferences from scRNA-seq data

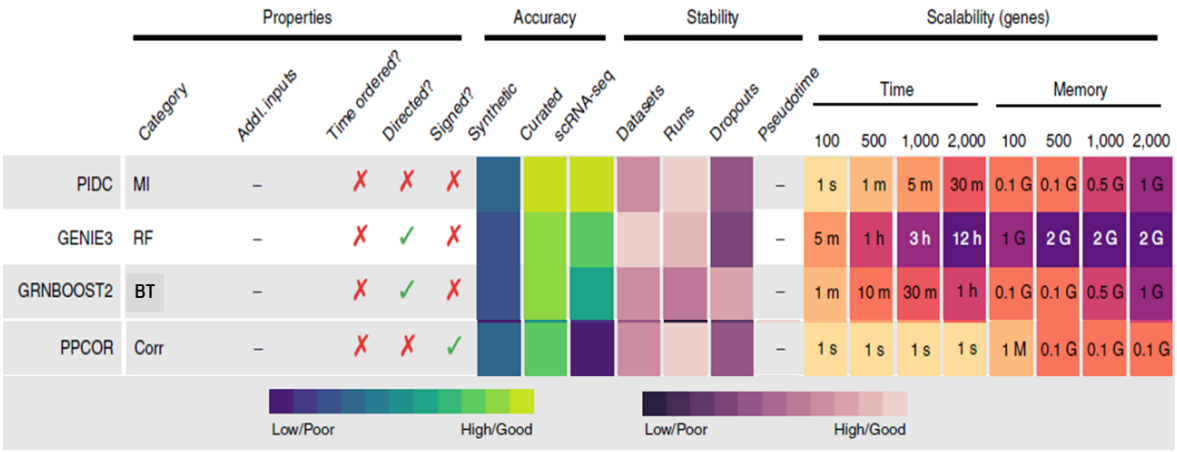
Nature Methods 17:147 (2020)

- In recent benchmarking based on scRNA-seq data from human and mouse, the GRN inference methods with no requirement for time-ordered cells were all top ranked in terms of accuracy.
- Since PIDC, GENIE3, GRNBoost2, and PPCOR do not require pseudotime-ordered cells, they are immune to any errors in pseudotime computation.



➢ **Benchmarking GRN inferences from scRNA-seq data** *Nature Methods 17:147 (2020)*

- MI (mutual information); RF (random forest); BT (boosting); Corr (correlation);
- GENIE3 (RF) and GRNBoost2 (BT) infer **directional edges** (TF → target), whereas PPCOR (Corr) and PIDC (MI) infer **unidirectional edges**.
- PPCOR is fast but shows the lowest accuracy.
- PIDC is fast and shows the highest accuracy.
- GENIE3 and GRNBoost2 show relatively high accuracy, but GENIE3 is very slow for >1000 cells.
- GENIE3 and PIDC also had better stability across multiple runs, whereas GRNBoost2 was less sensitive to the presence of dropouts.
- Since GRNBoost2 and GENIE3 have multithreaded implementations now, they are as fast as PIDC.



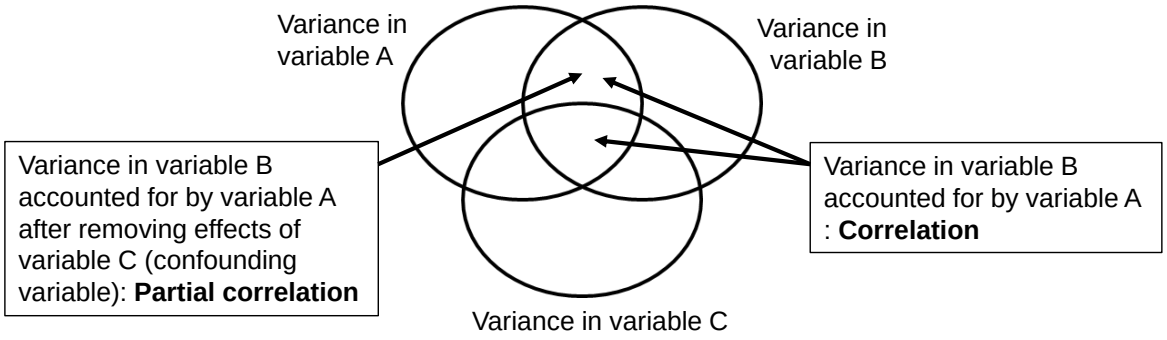
❖ **Partial Correlation**

- The **principle underlying correlation networks** is that if two genes have highly-correlated expression patterns (i.e. they are **co-expressed**), then they are **assumed to participate together in a regulatory interaction**.
- It is important to highlight that **co-expressed genes are indicative of an interaction but this is not a necessary and sufficient condition**. Partial correlation is **a measure of the relationship between two variables while controlling for the effect of other variables**.
- In complex system, **processes often interdependent**. For example, the abundance of clouds is often correlated with the amount of aerosol particles in the atmosphere.
- But both are also correlated with wind speed. Wind speed might be a "**mediating" or "confounding" variable**.
- Here we want to **test for an association two variables after controlling for the effect of one or more potentially confounding variables**.
- Correlation coefficient is adjusted for **correlations between each variable (*A*, *B*)** and **potential confounding variable *C***.

$$r_{AB.C} = \frac{r_{AB} - rACrBC}{\sqrt{1 - r^2_{AC}}\sqrt{1 - r^2_{BC}}}$$

- **Null hypothesis**: there is **no association** between the two variables **after controlling for effects of confounding variable(s).**
- Therefore the presence of an **edge between *A* and *B*** indicates that a **correlation exists between *A* and *B* regardless of which other nodes are being conditioned on**.

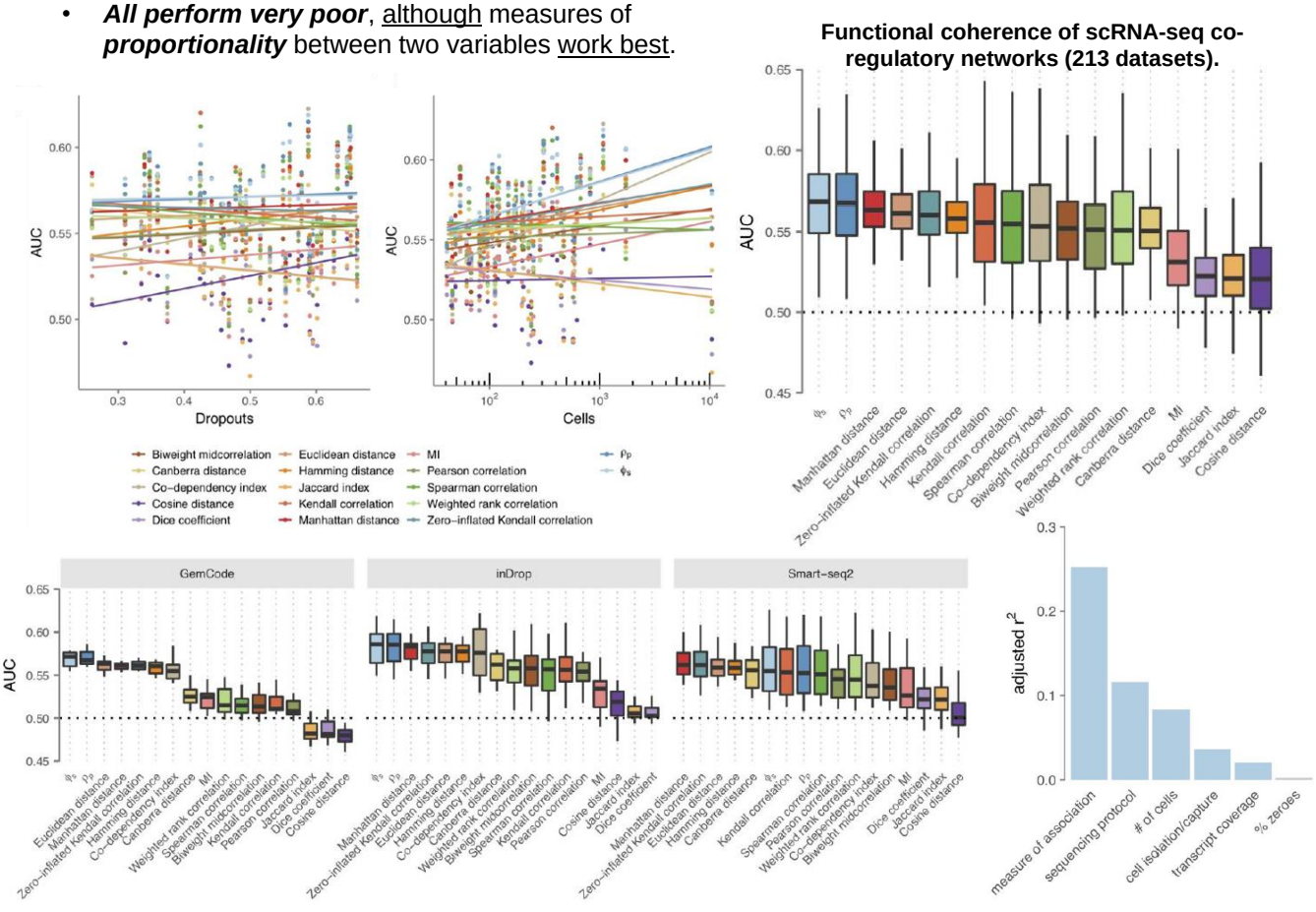
- **Venn Diagram explanation**



- Typically, **gene expression profiles from single cell data follow multimodal distribution** rather than a unimodal continuous shape. Therefore, **Pearson correlation coefficient are less suited for single cell expression data** because this metric measures a linear dependency between two variables.
- Given the non-linear nature of single cell gene expression data, nonparametric methods such as the **Spearman correlation** and **Kendall rank correlation** coefficients are more appropriate.
- It also computes a *p*-value for each correlation.
- Since these values are symmetric, this method yields an **undirected** regulatory network.
- We use the **sign of the correlation**, which is bounded between -1 and 1, to signify whether an interaction is **inhibitory (negative)** or **activating (positive).**

➢ **Benchmarking metrics for gene-gene association with scRNA-seq data** *Nat. Methods 16:381 (2019)*

- ***All perform very poor***, although measures of ***proportionality*** between two variables work best.



❖ **PIDC (Partial Information Decomposition and Context)** *Cell Syst. 2017;5(3):251–67. e3*

- **Mutual information** (**MI**) measure has advantages over correlation measure, as it can capture complex non-linear and non-monotonic dependencies.
- Calculating MI involves estimating pairwise joint probability distributions, generally requiring density estimation or data discretization, and the accuracy of these estimates depends on the sample sizes.
- The **entropy**, ***H(X)***, quantifies the uncertainty in the probability distribution, *p(x)*, of a random variable *X*. For a discrete random variable, $H(X) = -\sum_{x \in X} p(x) \log p(x)$
- MI between two random variables *X* and *Y*; *I(X;Y) = H(X) + H(Y) – H(X, Y),* where *H(X, Y)* is the **joint entropy** assuming independence of *X* and *Y.*
- For **a pair of co-regulated genes**, their observed ***H(X, Y)* is lower**, and **have higher MI**.
- Single-cell datasets are sufficiently large to allow us to accurately estimate probability distributions **between more than two variables** based on **multivariate information (MVI) theory**.
- MVI measure **improves accuracy of estimated information dependency**.
- **Partial information decomposition** (**PID**) enables to provide a meaningful measure of MVI.
- PID considers the information provided by **a set of source variables** (or genes), *S* = {***X*, *Y***}, about another **target variable, *Z***, partitioned into redundant, synergistic, and unique information.
- ***I(Z; X, Y)* = Synergy*(Z; X, Y)* + Unique$_X$*(Z; Y)* + Unique$_Y$*(Z*; *X*) + Redundancy(*Z*; *X*, *Y*)**
- **Redundant information** is the portion of information about *Z* that can be provided by either variable in *S* alone;
- **Unique information** from *X* (or *Y*) is the portion of information provided only by *X* (or only *Y*)
- **Synergistic information** is the portion of information that is only provided by knowledge of both *X* and *Y*.

➢ **Examples of MI calculation**

Below two profile pairs have same Euclidean distance. However…

1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1

For marginal entropy, P(0) = 0, P(1) = 1

H(X) = -(0*$\log_2$(0))-(1*$\log_2$(1)) = 0: low complexity of each random variable

H(Y) = H(X)

For joint entropy, P(0,0) = 0, P(0,1) = 0, P(1,0) = 0, P(1,1) = 1

H(X,Y) = -(0*$\log_2$(0))-(0*$\log_2$(0))-(0*$\log_2$(0))-(1*$\log_2$(1)) = 0: low complexity of joint incident

Thus *M(X,Y)* = 0+0-0 = 0.

1	0	1	0	1	0	1	0	1	0
1	0	1	0	1	0	1	0	1	0

For marginal entropy, P(0) = 0.5, P(1) = 0.5

H(X) = -(1/2*$\log_2$(1/2))-(1/2*$\log_2$(1/2)) = 0.5 + 0.5 = 1:

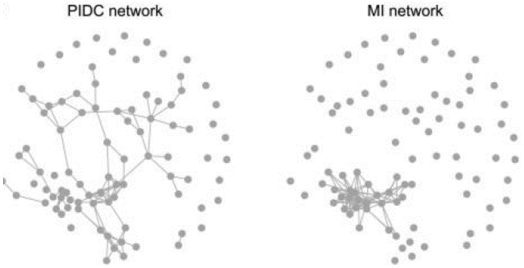
H(Y) = H(X)

For joint entropy, P(0,0) = 0.5, P(0,1) = 0, P(1,0) = 0, P(1,1) = 0.5

H(X,Y) = -(1/2*$\log_2$(1/2))-(0*$\log_2$(0))-(0*$\log_2$(0))-(1/2*$\log_2$(1/2)) = 0.5+0.5 = 1:

Thus *M(X,Y)* = 1+1-1 = 1.

- **PIDC outperforms pairwise MI-based algorithms**.
- The larger sample sizes of single-cell data are vital for PIDC-based network inference.

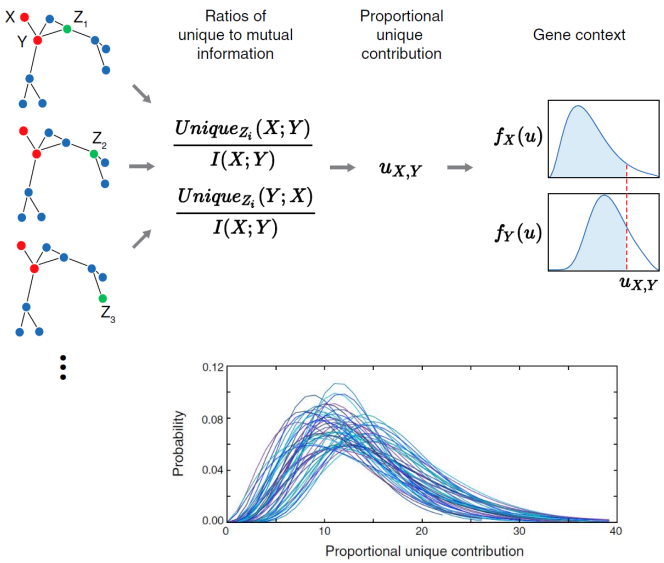


❖ **Network inference based on PID and Context**

- In a network of *n* genes, given a pair of genes X and Y, there are *n* − 2 gene triplets involving the pair. The *MI* between *X* and *Y, I(X;Y)* is unaffected by the choice of the third gene, *Z*, because MI is a pairwise measure, but $Unique_Z(X;Y)$, varies depending on *Z*.
- The ratio $Unique_Z(X;Y)/I(X;Y)$ is the proportion of *MI* that is accounted for by unique information between *X* and *Y*. ***This ratio would be higher if X and Y are connected***.
- **Proportion of unique contribution (PUC)** between two genes *X* and *Y* as the sum of this ratio calculated using every other gene *Z* in a network.
- For network inference, the redundancy and unique information contributions are first estimated for every gene triplet, then the PUC is calculated for each pair of genes in the network.
- Finding a threshold for defining an edge at this stage is problematic, because the **distributions of PUC scores differ between genes**, thus setting a global threshold for PUC scores across the whole network risks biasing the results by factors such as expression variability.
- Therefore, **confidence of an edge** that takes into **account the network context** is used.

$$c = F_X(u_{X,Y}) + F_Y(u_{X,Y})$$

- where $F_X(U)$ is the cumulative distribution function of all the PUC scores involving gene *X*.
- The resulting network is **undirected** since the proportional unique contribution is symmetric.



❖ **GENIE3 (**GEne Network Inference with Ensemble of trees) *PLoS ONE 5(9): e12776 (2010)*

- GENIE3 is a GRN inference method based on **variable selection with ensembles of regression trees**. In each of the **regression problems**, **the expression pattern of one of the genes (target gene) is predicted from the expression patterns of all the other genes (input genes)**, using tree-based ensemble methods Random Forests (RF).
- The **importance of an input gene in the prediction of the target gene expression pattern** is taken as an **indication of a putative regulatory link**.
- Putative regulatory links are then aggregated over all genes to provide a ranking of interactions from which the whole network is reconstructed.
- Tree-based ensemble methods doesn't make any assumption about the nature of gene regulation, can potentially **capture high-order conditional dependencies between expression patterns**.
- Importantly, GENIE3 produces **directed** GRNs, and naturally allows for the presence of feedback loops in the network. It is also **fast** and **scalable**.
- A network inference algorithm was defined as a procedure that **exploits a *set of gene expression vectors* to assign weights to putative regulatory links from any gene *i* to any gene *j***, with the aim of yielding large values for weights which correspond to actual regulatory interactions.
- Exploiting expression data, the identification of the regulatory genes for a given target gene is defined as **determining the subset of genes whose expression directly influences or is predictive of the expression of the target gene**.

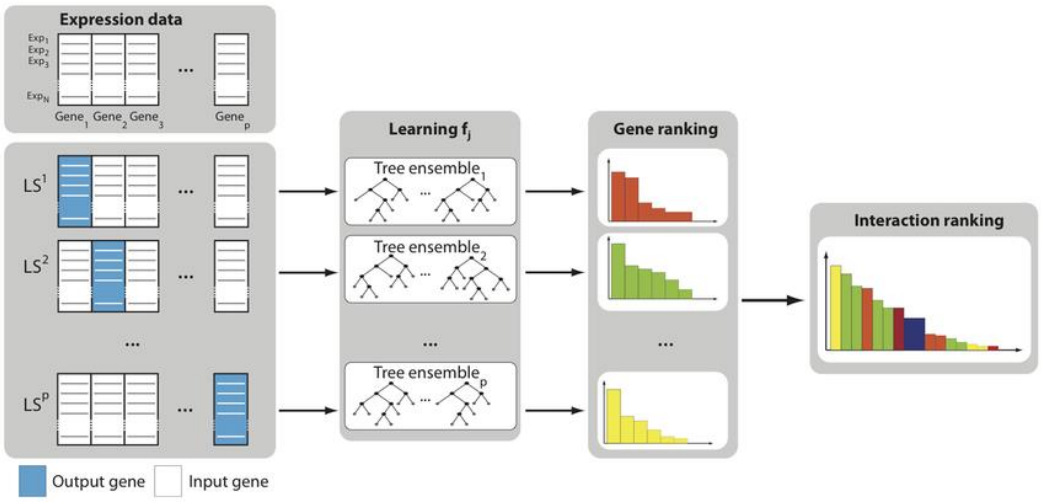
➢ Therefore, here the ***network inference*** problem is equivalent to a ***feature selection*** problem.

- Importantly, **variable (i.e., gene) importance can be computed from a tree** that allows to rank the input features according to their relevance for predicting the output. GENIE3 uses a measure which at each test node computes the total reduction of the variance of the output variable due to the split.

- The **overall importance of one variable** is computed by **summing the importance values of all tree nodes where this variable is used to split**. Those attributes that are not selected at all obtain a zero value of their importance, and those that are selected close to the root node of the tree typically obtain high scores.
- Attribute importance measures can be easily extended to ensembles, simply by **averaging importance scores over all trees** in the ensemble.

▪ **GENIE3 procedure**

1. For each gene j = 1, … ,p, a learning sample LS^j is generated with expression levels of j as output values and expression levels of all other genes as input values.
2. A function f_j is learned (with RF) from LS^j and a local ranking of all genes except j is computed.
3. The p local rankings are then aggregated to get a global ranking of all regulatory links.



❖ Ensemble Learning

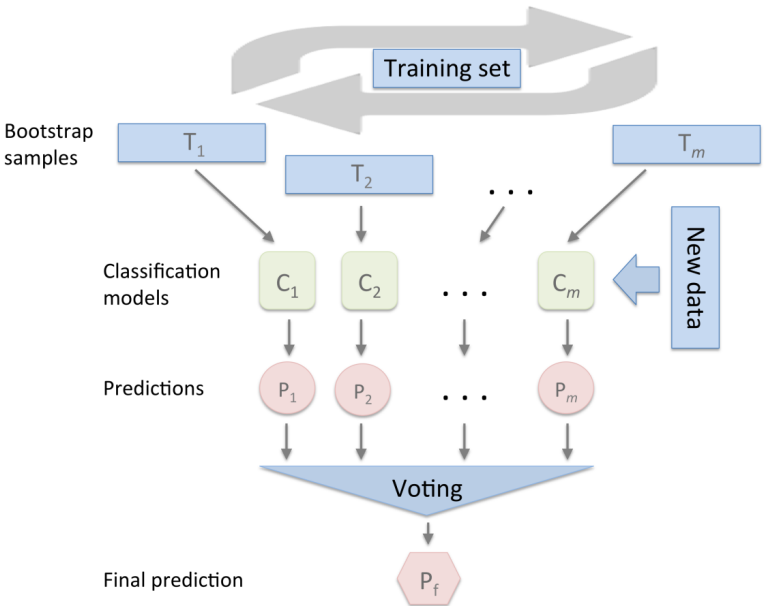
- Goal: to combine *weak* models (classifiers or regressions) into a final model that has a better generalization performance than the individual models.

▪ Ensemble method

- Ensemble methods use multiple learning algorithms (e.g., decision tree, logistic regression, etc) to obtain better predictive performance.
- **Two major types of ensemble learning** approaches: **Bagging** and **Boosting**

❖ Bagging (L. Breiman, 1994)

- Building multiple models (e.g., classifiers C_1, C_2, …, C_m) **on the same learner** using bootstrap samples of the original training sets (T_1, T_2, …, T_m) → Aggregating prediction results (e.g., majority voting in classification) for the final model

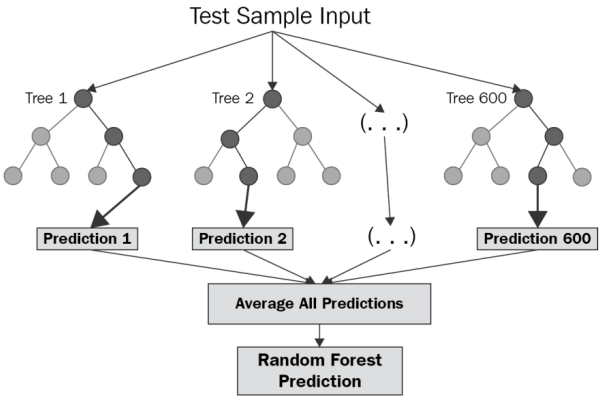


❖ Random Forests

- A popular ensemble learner with **bagging** approach
- Combining individual trees (weak learners) to build random forests (strong learner)

▪ Steps

- Draw a random **bootstrap** sample of size *n* (choose n random samples out of total *n* samples with replacement)
- Make **weak** decision trees from the bootstrap samples with two **hyperparameters**:
  - Maximum depth of the tree: *d*
  - The number of trees in the forest: *k*
- Split input data using the best feature to maximize the information gain.
- Repeat above steps for *d* features for *k* trees
- Aggregate the prediction of each tree by **majority voting (in classification)** or **averaging (variable weight scores in regression)**

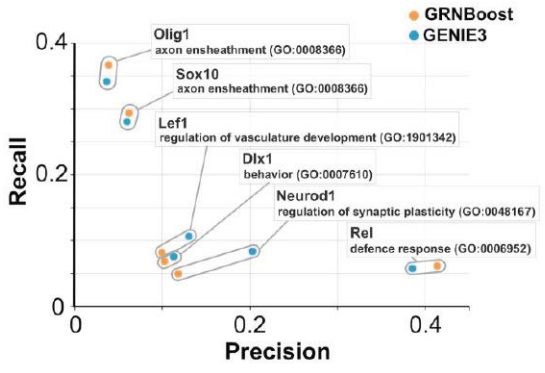


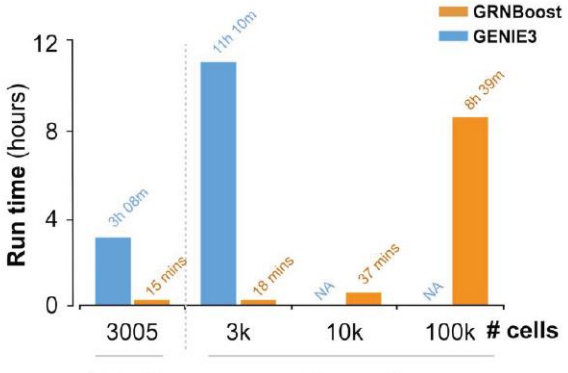
▪ Pros and Cons

- Don't need to prune the random forest in general, since the ensemble model is quite robust to the noise from individual decision trees
- The larger the number of trees *k*, the better the performance of the random forest
- Large computational cost for large *k*

❖ GRNBoost *Nature Methods 14:1083 (2017)*

- GRNBoost is based on the same concept as GENIE3 but using the **gradient-boosting machines** (GBM). Boosting is an ensemble learning strategy.
- GRNBoost uses **stumps** (regression **trees of depth 1**) **as the base learner**.

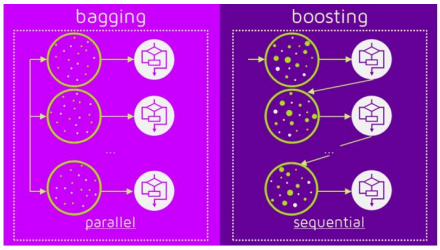




❖ GRNBoost2 *Bioinformatics 35:2159 (2019)*

- GRNBoost2 employs a regularized stochastic variation on GBMs. It equips GBM regressions with a **heuristic *early-stopping* regularization strategy** using out-of-bag improvement estimates.
- Each new decision tree is trained in function of a random subset of observations (90%, hence stochastic), whereas the remaining (10%, out-of-bag) observations are used to calculate an estimate of the loss function improvement entailed by adding that tree to the ensemble.
- When the average of the last *n* improvement values drops below 0, the early-stopping criterion is met and no more trees are added to the ensemble.
- Regressions that do not display net improvement early on are aborted and thus **prevented from causing useless computational workload**.

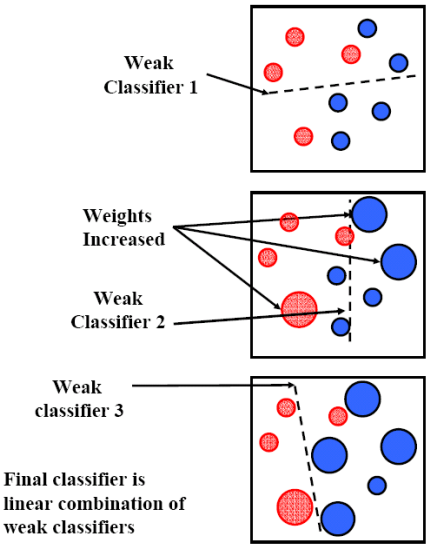
❖ Boosting



- Boosting is different from bagging. In boosting, we consider the mistakes of previous predictors and **train the new predictors on those mistakes** and then repeat the process till we get a better fit.

❖ AdaBoost (Adaptive Boosting) [Y. Freund & R. Shapire 1995]

- ***Iteratively reweight your dataset***, placing ***higher weights on the examples you are getting wrong***. At each iteration, refit and add the result to ensemble.

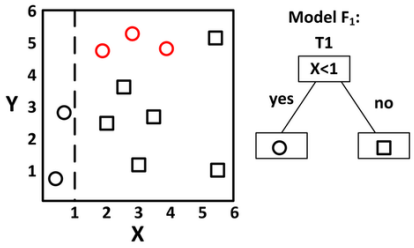


Algorithm

1. Start by applying some method to the learning data, where each observation is assigned an equal weight.
2. Compute the predicted classifications, and assign greater weight to those observations that were difficult to classify (where the misclassification rate was high), and lower weights to those that were easy to classify (where the misclassification rate was low).
3. Then apply the classifier again to the weighted data (or with different misclassification costs), and continue with the next iteration (application of the analysis method for classification to the re-weighted data).

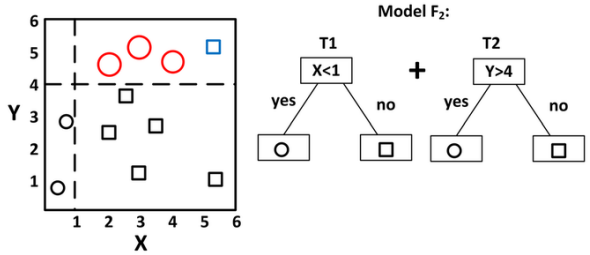
▪ A simple example of visualizing boosting with trees.

Iteration 1



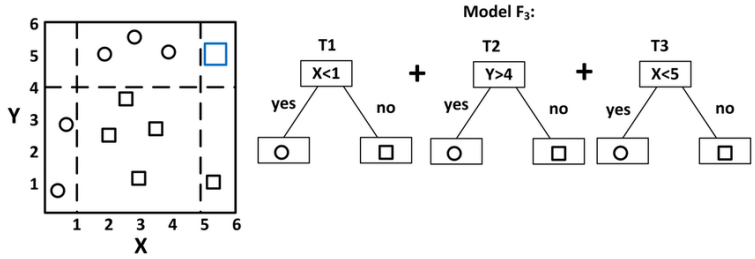
- Boosting is a framework that iteratively improves *any* weak learning model. **In practice** however, **boosted algorithms almost always use decision trees as the base-learner**.
- Whereas random forests build an ensemble of deep independent trees, **Boosting machines** build an **ensemble of shallow and weak successive trees** with **each tree learning and improving on the previous**.

Iteration 2



- When combined, these many weak successive trees produce a powerful "committee" that are often hard to beat with other algorithms.
- Fits consecutive trees where each solves for the net loss of the prior trees. Results of new trees are applied partially to the entire solution.
- Final model is the **linear combination of weak models** with **weighted votes for each of the base models.**

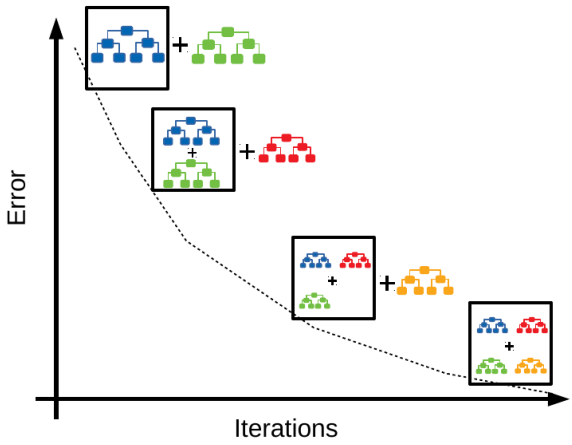
Iteration 3



Scientific Reports 8:1 (2018)

❖ Gradient Boosting Machines (GBM) [J. Friedman 1999]

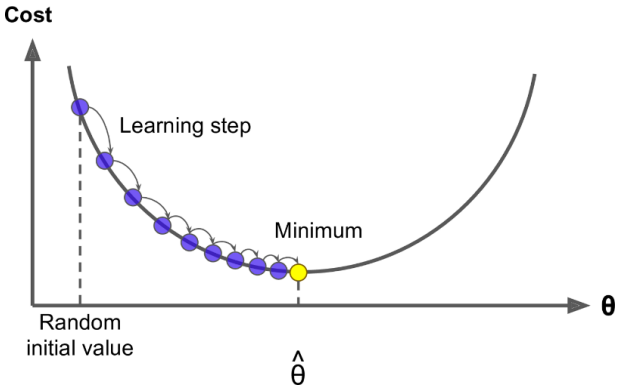
- The basic principle is same for both AdaBoost and Gradient Boost. **The differences is how the new predictor learns from the old one**.
- **Adaboost learns the weights of weak predictors** during the learning process. It **keeps adding +ve and -ve weights to predictors** about certain data points till we have predictors that can combine to give a better result.
- **GBM generates predictors** during the learning process. Instead of adding any weights to predictors, **wrong predicted data points are considered as a new training set** and the new predictor tries to fit these data points making a new model. It **keeps fitting wrongly predicted data points with the new predictor till lesser predictions are wrong** and then use all predictors together to predict output by voting or averaging.
- GBM uses **gradient descent algorithm** which can **optimize any differential loss function**. Each tree in GBM is a successive gradient descent step.
- **GBM = Gradient Descent + Boosting**
- In GBM instead of reweighting used in AdaBoost, each tree is fit to the negative gradients of the previous tree.
- Basic elements of GBM: loss function, weak learner, additive model
- Improvement of basic GBM: tree constraints, shrinkage, random sampling, penalized learning (=regularization)



http://tvas.me/articles/2019/08/26/Block-Distributed-Gradient-Boosted-Trees.html

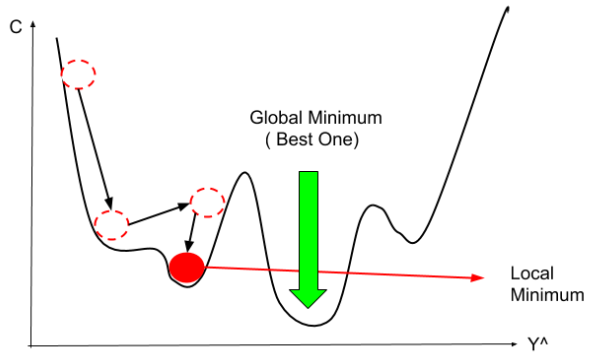
❖ Gradient Descent (경사하강법)

- Many algorithms, including decision trees, focus on minimizing the residuals and, therefore, emphasize the mean squared error (MSE) loss function. ***Gradient boosting machines*** can be generalized to loss functions other than MSE.
- ***Gradient descent*** is a very generic optimization algorithm capable of finding optimal solutions to a wide range of problems. The general idea of gradient descent is to tweak parameters iteratively in order to minimize a loss function.
- Suppose you are a downhill skier racing your friend. A good strategy to beat your friend to the bottom is to **take the path with the steepest slope**. This is exactly what gradient descent does - it measures the local gradient (기울기) of the loss function for a given set of parameters and **takes steps in the direction of the descending gradient**.
- Once the gradient is zero, we have reached the minimum.
- Gradient descent can be performed on **any loss function that is differentiable (미분이 가능한)**. Consequently, this allows GBMs to optimize different loss functions as desired.

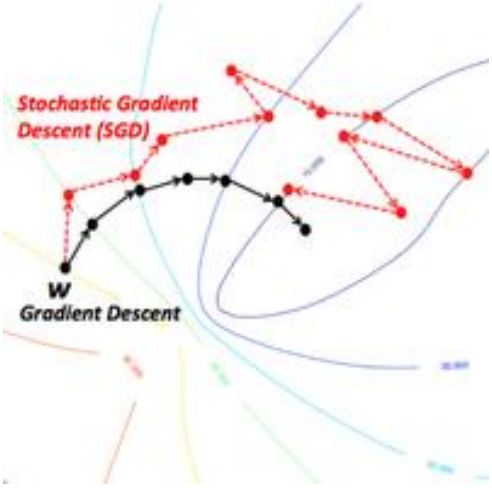


http://uc-r.github.io/gbm_regression

- An important parameter in gradient descent is the size of the steps which is determined by the ***learning rate*** **(**기울기 보폭**).**
- **Not all cost (loss) functions are convex (bowl shaped). There may be local minimas**, plateaus, and other irregular terrain of the loss function that makes **finding the global minimum difficult**.
- **If the learning rate is too small**, then the algorithm will **take many iterations** to find the minimum. On the other hand, **if the learning rate is too high**, you **might jump cross the minimum** and end up further away than when you started.
- ***Stochastic gradient descent*** enables to find near the global minimum with much less iterations **by sampling a fraction of the training observations** (typically without replacement) and growing the next tree using that subsample.



https://www.mltut.com/stochastic-gradient-descent-a-super-easy-complete-guide/

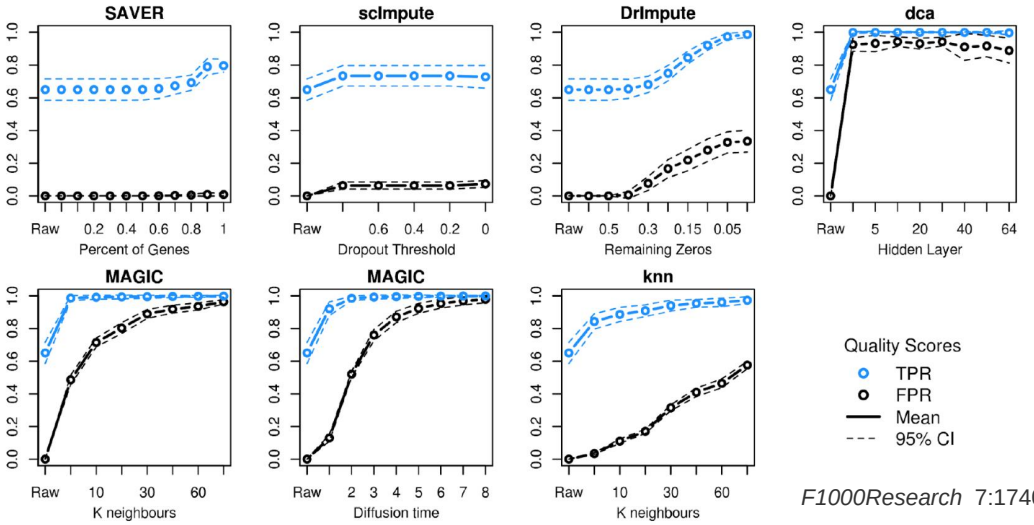


https://golden.com/wiki/Stochastic_gradient_descent_(SGD)

❖ Is imputation of zero-inflated scRNA-seq data helpful for network inference?

- Data smoothing based methods, MAGIC, knn-smooth and dca, generated many false-positives.
- Imputation of single-cell RNA-seq data introduces circularity that can generate false-positive results. Thus, statistical tests applied to imputed data should be treated with care.
- Model-based imputation methods typically generated fewer false-positives but this varied greatly depending on the diversity of cell-types in the sample.
- SAVER was the least likely to generate false or irreproducible results, thus should be favored over alternatives if imputation is necessary.

➢ **False positive and true positive gene-gene correlations (p < 0.05 Bonferroni multiple testing correction) as imputation parameters are changed.**



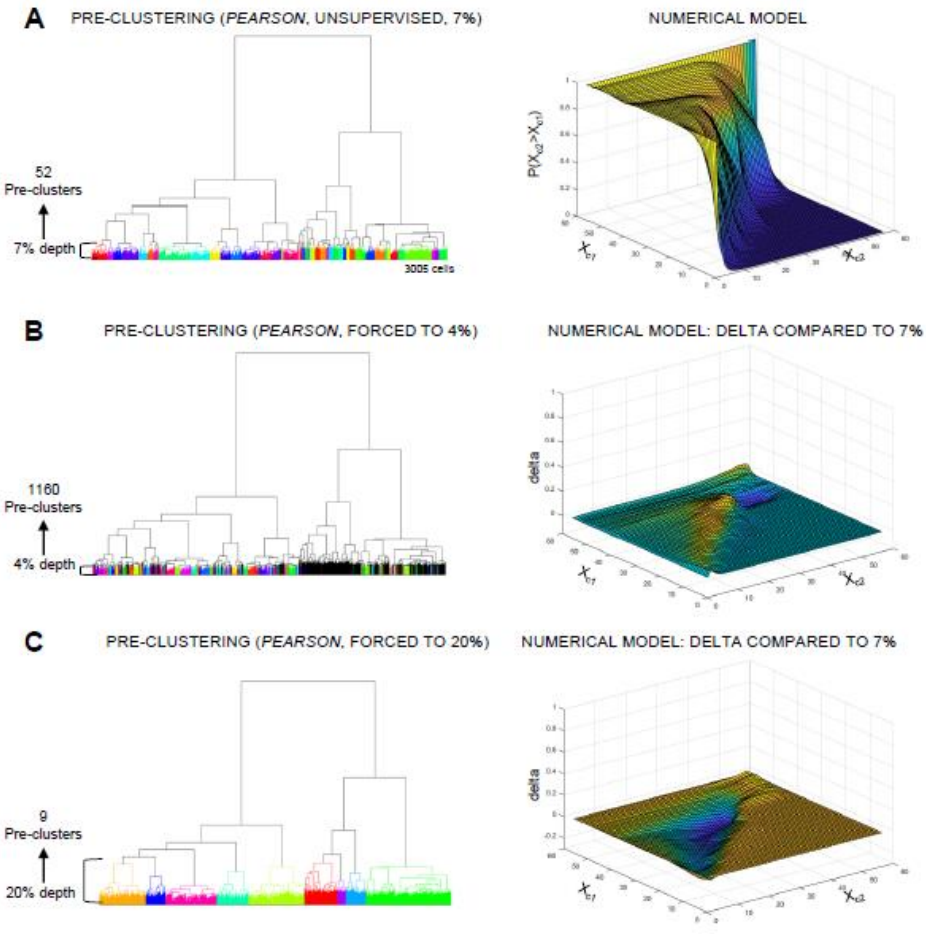
F1000Research 7:1740 (2019)

❖ bigSCale method for scRNA-seq data transformation

Genome Research 28:878 (2018)
Genome Biology 20:110 (2019)

▪ Step 1: Preclustering and numerical modeling

- To handle the noise and sparsity of scRNA-seq data, bigSCale method **uses large sample sizes to estimate an accurate numerical probabilistic model of noise**.
- **Preclustering cells into groups sharing highly similar expression profiles**, which are next **treated as biological replicates to allow evaluation of the noise**.
- Preclustering procedure: (1) read/count data normalization (2) $\log_{10}(X+1)$ transformation (3) gene normalization to avoid severe effect of highly expressed genes on clustering (4) clustering cells with Pearson correlation and hierarchical clustering.
- Different cutting depths give different numerical models. It finds the **deepest possible cut** (10-20% of total tree height in general) in the tree to ensure that **only highly similar cells are grouped together**. → final clusters
- At this stage, **the cells within each group are treated as replicates**, assuming their **changes of expression to be solely due to noise and not to biological differences**.
- **All within-group pairwise comparisons between cells** are enumerated in order **to determine** how rare/common (i.e., assigning a ***P*-value**) **each combination of expression values** is. Specifically, if a cluster contains *n* cells, it produces $C(n,2) = n*(n-1)/2$ combinations of cells. Each of these combinations contain *k* couples of expression values $(Xcell_1, Xcell_2)$, where *k* is equal to the total number of genes and $Xcell_1$, $Xcell_2$ is the expression of a gene in the two compared cells.
- **The numerical model is robust to the different tree cut**. Difference in numerical probabilistic models between default 7% cut and forced 4% cut or 20% cut is marginal.



(A) (Left) Default, unsupervised heuristic sets a cut of 7% of the total dendrogram depth, which results in 52 pre-clusters. (Right) The numerical model calculated using the 52 pre-clusters. ***Xc1* and *Xc2* represent the expression (in a binned UMIs grid) of a given gene *X* in two cells *c1* and *c2* belonging to the same pre-cluster**.

The cumulative distribution plot estimates the frequency, hence **likelihood, of an expression change**.

(B-C) The difference between the numerical model of 4% cut and 7% cut (B. Right) or 20% (C. Right) is **marginal**.

Genome Research 28:878 (2018)

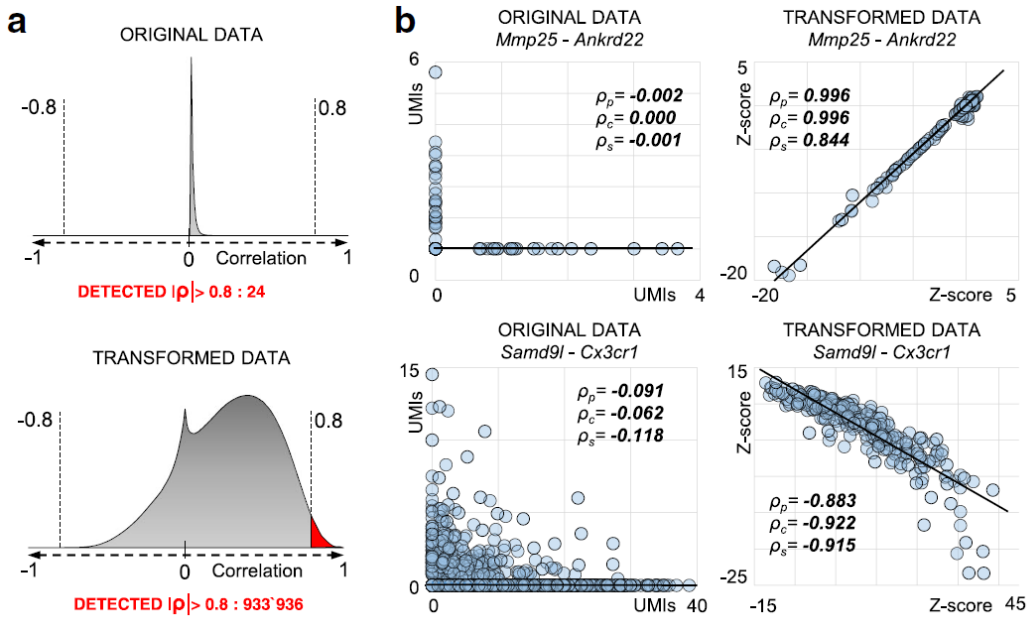
- **Step 2: Differential expression analysis**

  - The **numerical model of noise** is used to identify differential expression (DE) between groups.
  - After clustering the cells to the highest feasible granularity, we used *bigSCale* to **run an iterative DE analysis between all pairs of clusters**. For x clusters, this results in a total of $x \times (x-1)/2$ unique comparisons, each yielding a ***Z*-score for each gene** that indicates the **likelihood of an expression change between two clusters**.
  - Thus, if we started with **(10 clusters)** × **(*k* genes)**, we end up with **[45 × *k*] matrix of Z-scores**.

- **Step 3: Network inference using Z-score**

  - We then compute **correlations between genes using *Z*-scores** instead of expression values.
  - Therefore, **linear correlations in the *Z*-score space can reflect non-linear correlations in the original expression space**. Hence, **Pearson (or Spearman) correlation coefficient is recommended** to measure association between genes.

Genome Biology 20:110 (2019)



Transformed single-cell data allow detection of hidden correlations.

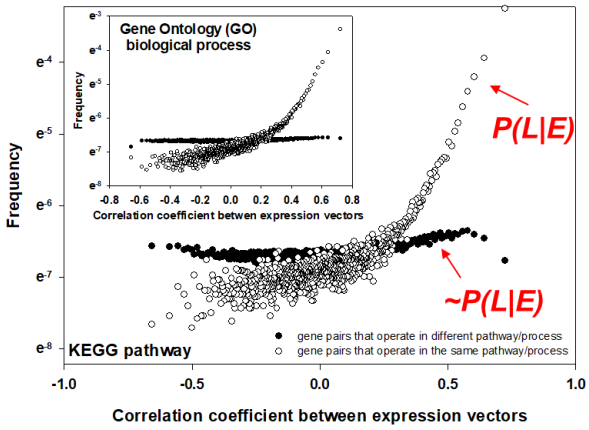
a. Distribution of *Pearson* correlations ρ_p in normalized expression data (7697 microglia cells) or in the *Z*-score space. We detect only 24 correlations $|\rho_p| > 0.8$ in the first scenario, but almost one million $|\rho_p| > 0.8$ in the *Z*-score space.

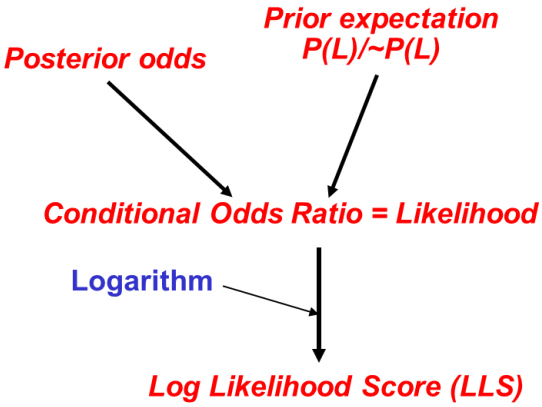
b. Examples of correlations using either expression values or *Z*-score-transformed data (ρ_p *Pearson*, ρ_c *Cosine*, ρ_s *Spearman*). Due to drop-out events and other artifacts, the positive correlation between *Mmp25* and *Ankrd22* is only exposed using *Z*-scores. Similarly for the negative correlation between *Samd9l* and *Cx3cr1*.

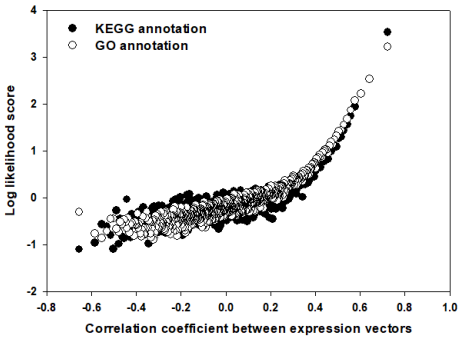
Genome Biology 20:110 (2019)

❖ Benchmarking co-functional association using Bayesian statistics

- We use Bayesian statistics to measure Likelihood of being associated.







Posterior Odds (belief *after* the observed data)

$$\text{Log Likelihood Scores (LLS)} = \ln\left(\frac{P(L \mid E) / \sim P(L \mid E)}{P(L) / \sim P(L)}\right)$$

Prior Expectation (Belief *before* the observed data)

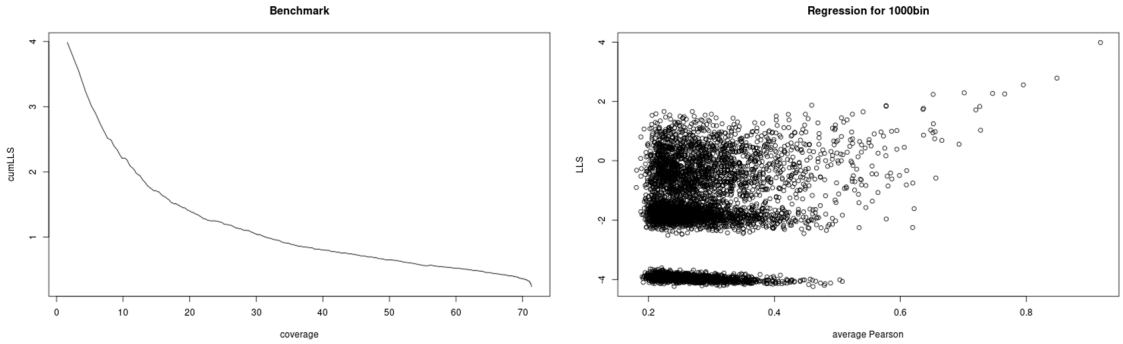
L: linkage between two genes
E: an evidence by given data
If LLS = 0, the likelihood of two gene's association is no better than random chance

❖ How to make a benchmarking data set from pathway database

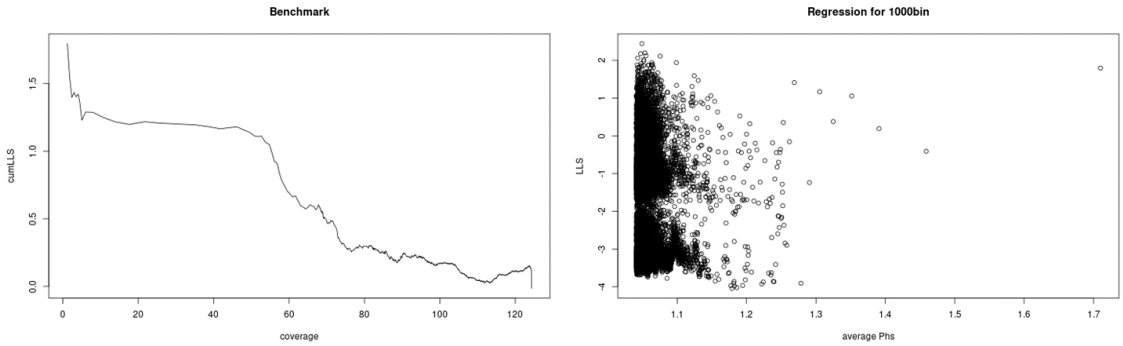
- Collect pairs of genes that belong to the same pathway.
- Use pathway annotation DBs (Gene Ontology biological process, KEGG pathway, MetaCyc, …).
- What makes a good pathway annotation DB for network modeling?
  - Frequent update
  - comprehensive
  - Evidence codes
- From pathway annotation to pathway links for network training: for example, a pathway has 4 member genes (gene A, B, C, D). Then we can make the following training samples by the pathway
  - A – B
  - A – C
  - A – D
  - B – C
  - B – D
  - C – D

➢ **Coexpression network by bigSCale was compare with proportionality score GSE99254 (~12k T cells FACS sorted from NSCLC patients)**

- **bigSCale transformation → PCC**

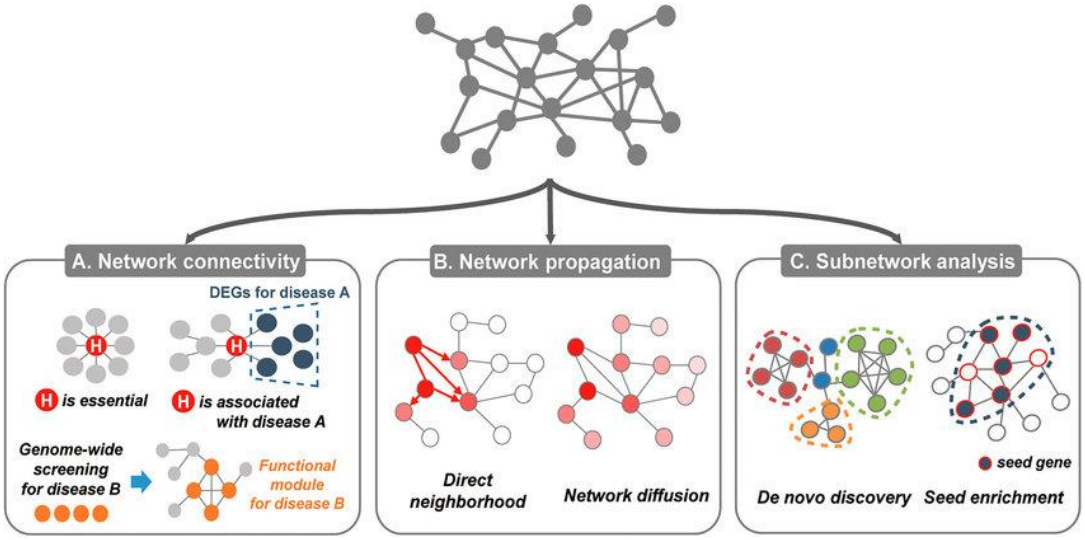


- **Standard preprocessed data → Proportionality**



❖ **Hypothesis generation using "static/integrated" gene/protein networks**

1. **Network connectivity**: Hub genes tend to be functionally more important (e.g., essential genes)
2. **Network propagation**: Genes for the same phenotype (e.g., disease) tend to be connected in the interactome. Thus, novel disease genes can be inferred by propagated information from neighboring disease genes.
3. **Subnetwork analysis**: Functional or disease modules can be represented as subnetworks of tightly connected genes



(*Animal Cells and Systems. 21:1-7, 2017*)

❖ Hypothesis generation in single-cell network biology

1. Hypothesis from subnetwork (co-expression module or regulon) activity analysis

1) Co-expression modules associated with specific context or cell-type (use WGCNA)

2) Regulon activity profiles of cell states (use SCENIC)

2. Hypothesis from network topology analysis

1) Changes in Centrality: Marker genes shows higher centrality in associated cell type.

2) Changes in Neighbors (Targets):

- The lose of co-expression between cell-type specific regulators and their normal targets causes Impairment of the cell function.
- Lineage regulators change targets in different cell types at the stage of differentiation.

3) Changes in Modularity

3. Hypothesis from genotype-network association

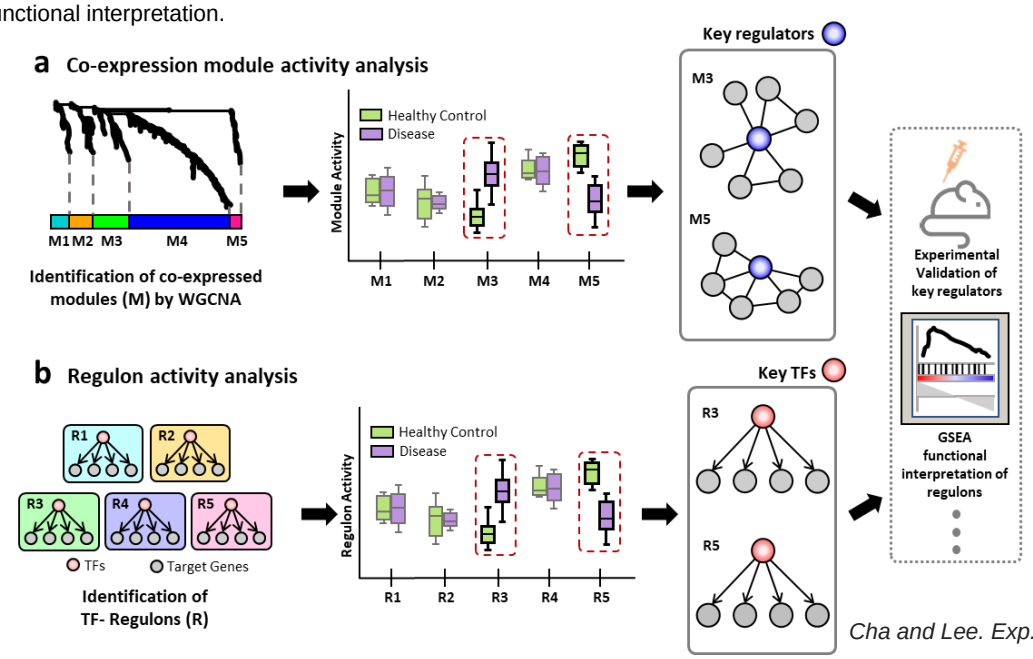
1) Co-expression QTL

❖ Hypothesis from subnetwork analysis

(a) Weighted correlation network analysis (WGCNA) on scRNA-seq data generally reveals multiple **modules** (M1-5) of co-expressed genes with various size. **Activity of modules can be measured by average gene expression level**. Module activity may significantly differ between cells from different states (e.g., cells of disease samples versus those of healthy control), which suggest that this **co-expressed module** is **associated with the disease state** and may **contain key regulators for the disease**, often those with **high network centrality**.

(b) Transcription factor (TF)-target interaction inference generates a set of regulons (R1-5) that are regulated genes by each TF. Comparison of **regulon activity** between healthy and disease states, similarly to module activity, can suggest its **association with disease state**. Then, the **TF** for the associated regulon is predicted to be a **key regulator**.

(c) These candidate regulators often go into experimental validation and gene set enrichment analysis (GSEA) for functional interpretation.



Cha and Lee. Exp. Mol. Med. (2020)

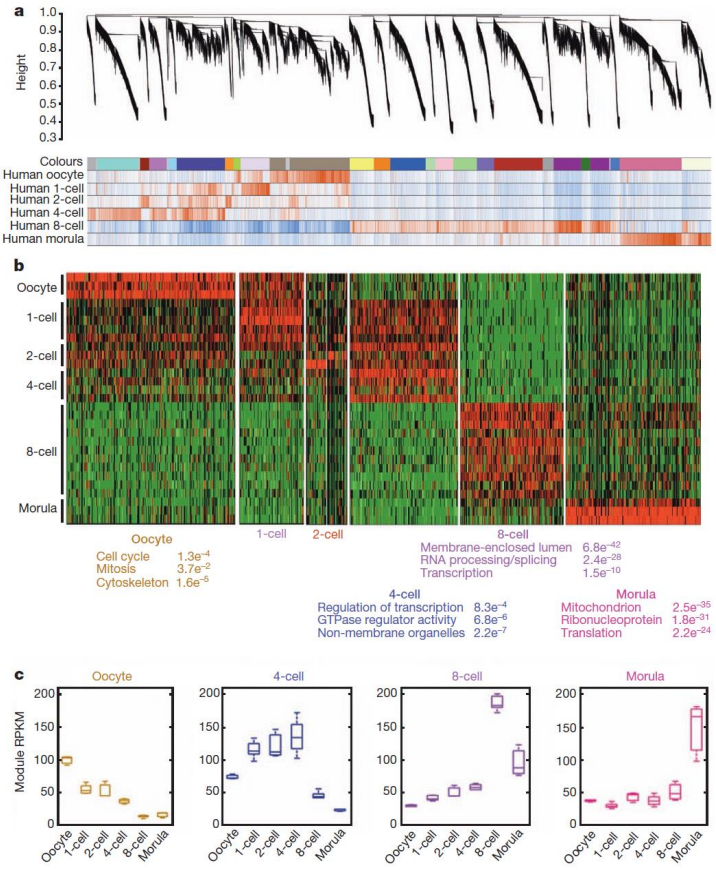
❖ Co-expression module associated with specific context or cell type (*Nature 500:593, 2013*)

- Find WGCNA coexpressed modules from scRNA-seq from various development stages of human embryo.
- Measure mean expression of all genes of a module for each context (developmental stage).
- **Each stage can be delineated concisely by a few modules**.
- We can **identify key regulators in the module** (e.g. hub gene; TF)

a. WGCNA modules correspond to branches and are labelled by colors as indicated by the first color band underneath the tree. Remaining color bands reveal highly correlated (red) or anti-correlated (blue) transcripts for the particular stage.

b. Heatmap showing relative expression of **7,313 genes in 7 representative stage-specific modules** across all samples. As each developmental window only has one or two highly correlated modules, the modules were assigned biological names. Top three representative gene ontology terms and their associated functional enrichment *P*-values are shown below.

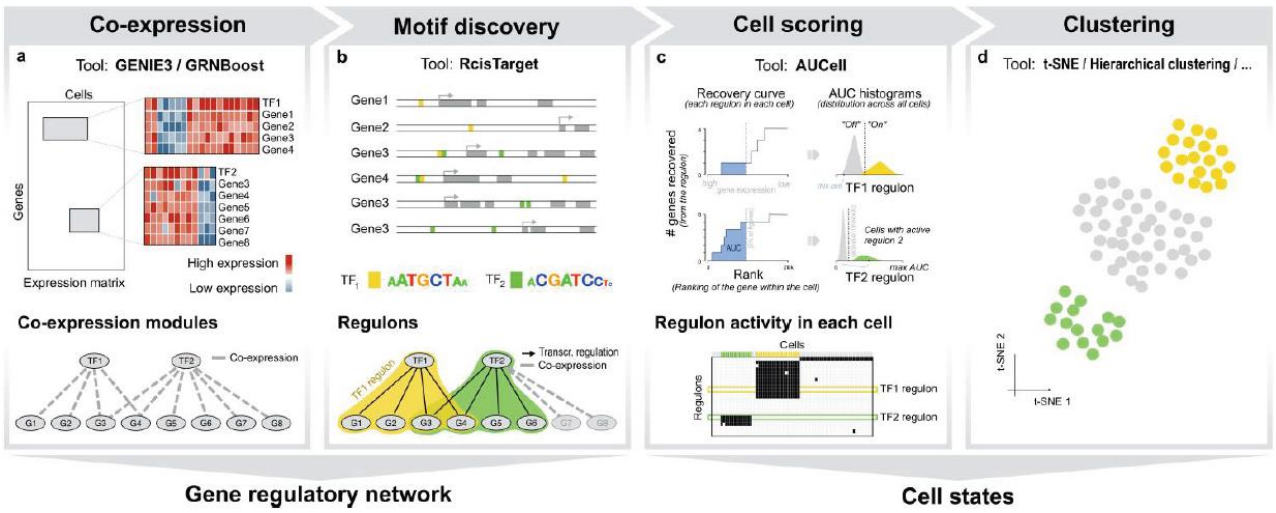
c. Boxplots showing the distribution of module expression (**meanRPKM of all genes within a given module**) for different cell types.



❖ SCENIC (single-cell regulatory network inference and clustering) *Nature Methods 14:1083 (2017)*

- **To overcome the high noise and sparsity of scRNA-seq data**, SCENIC **uses single-cell gene expression as well as cis-regulatory sequences**. SCENIC workflow consists of 3 steps:

1. Sets of genes that are coexpressed with TFs are identified using **GENIE3 or GRNBoost**.
2. To identify putative direct-binding targets, each coexpression module is subjected to cis-regulatory motif analysis using **RcisTarget**. Only modules with significant motif enrichment of the correct upstream regulator are retained. → ***Regulon***
3. **AUCell** scores the **activity of each regulon in each cell**, thereby yielding a **binarized activity matrix with reduced dimensionality**, which **can be useful for downstream analyses**. For example, clustering based on this matrix identifies cell types and states based on the shared activity of a regulatory subnetwork. Since the regulon is scored as a whole, instead of using the expression of individual genes, this approach is robust against dropouts.

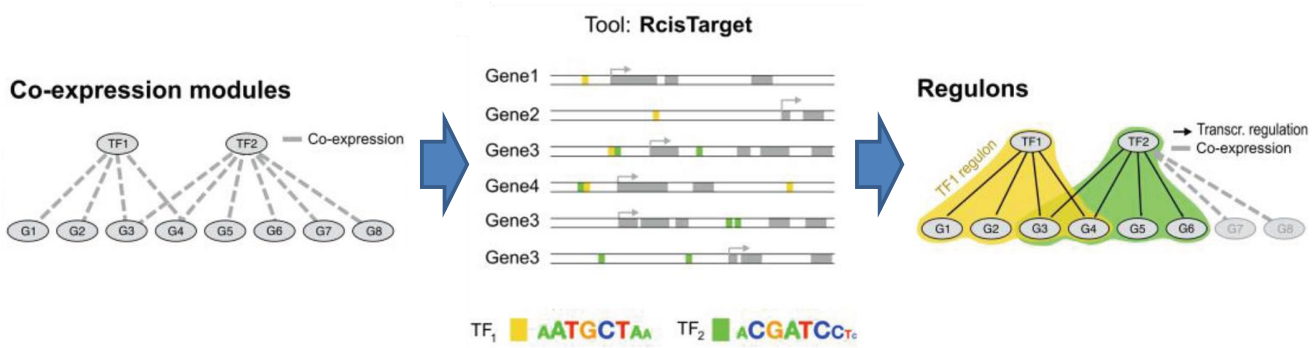


❖ RcisTarget

- RcisTarget is based on two steps.

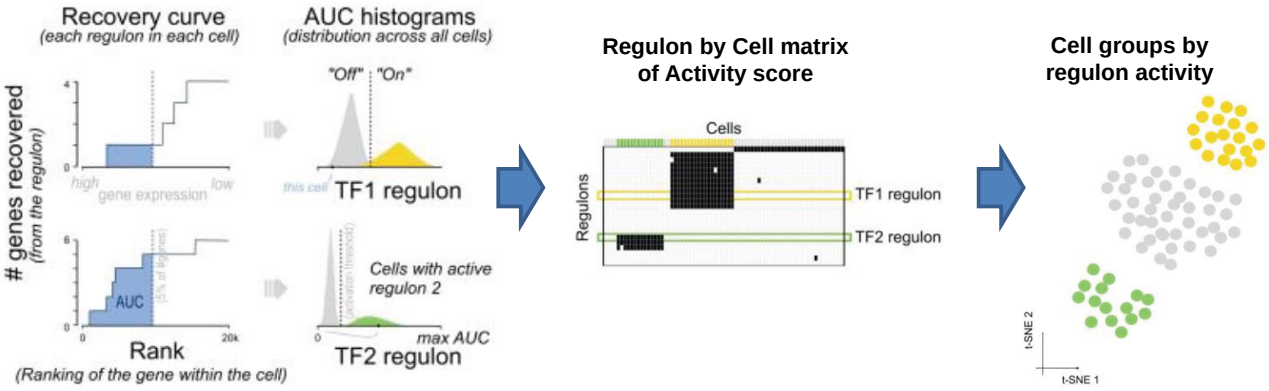
1. **Identification of enriched TF-binding motif across the genes of Regulon**. For each TF, RcisTarget selects DNA motifs that are significantly over-represented in the surroundings of the transcription start site (TSS) of the target genes. This is achieved by applying a recovery-based method on a database that contains genome-wide cross-species rankings for each motif. The motifs that are annotated to the corresponding TF and obtain a normalized enrichment score (NES) > 3.0 are retained.
2. **Prediction of target genes by enriched motif**. (i.e., genes in the target gene set that have the enriched motif).

- **The final GRN = TF-target by expression patterns ∩ TF-target by enriched motif**
- There could be negative-correlated TF modules. However, these modules are generally less numerous and showed very low motif enrichment. For this reason, we take only positive-correlated targets.



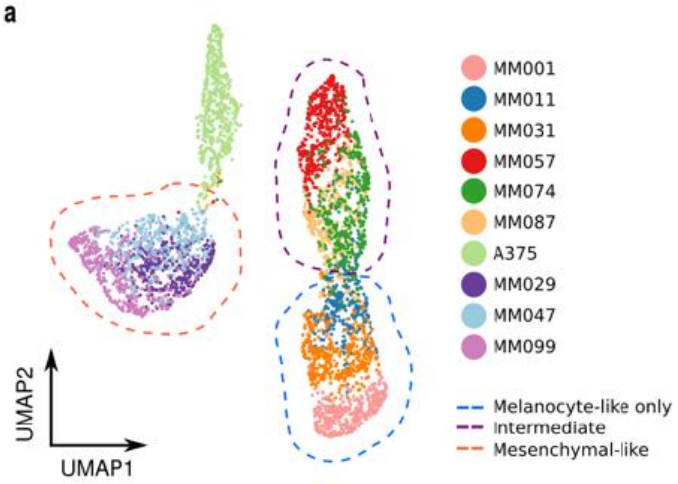
❖ AUCell

- AUCell can identify cells with **active regulons** in single-cell RNA-seq data.
- AUCell scoring method is based on a recovery analysis where the ***x*-axis** is the ranking of **all genes** based on expression level (genes with the same expression value, e.g., '0', are randomly sorted); and the ***y*-axis** is the number of genes recovered from the input set (**regulon genes**).
- AUCell then uses an area under the recovery curve (**AUC**) to calculate **whether a critical subset of the input gene set is enriched at the top of the ranking for each cell**.
- The output of this step is a **matrix with the AUC score for each regulon** (of each TF) **in each cell**. We use either the AUC scores (across regulons) directly as continuous values to cluster single cells, or we generate a binary matrix using a cutoff of the AUC score for each regulon.
- Clustering cells for regulon activity profiles can group cell types, suggesting that network activity score can **complement to expression data** in single-cell analysis.

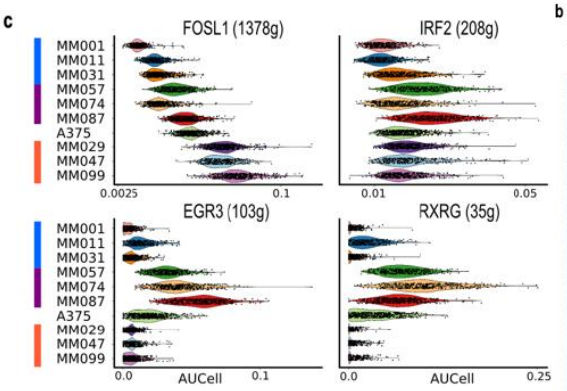


❖ **Study of cancer cell states based on regulon activity** (*bioRxiv* https://doi.org/10.1101/715995)

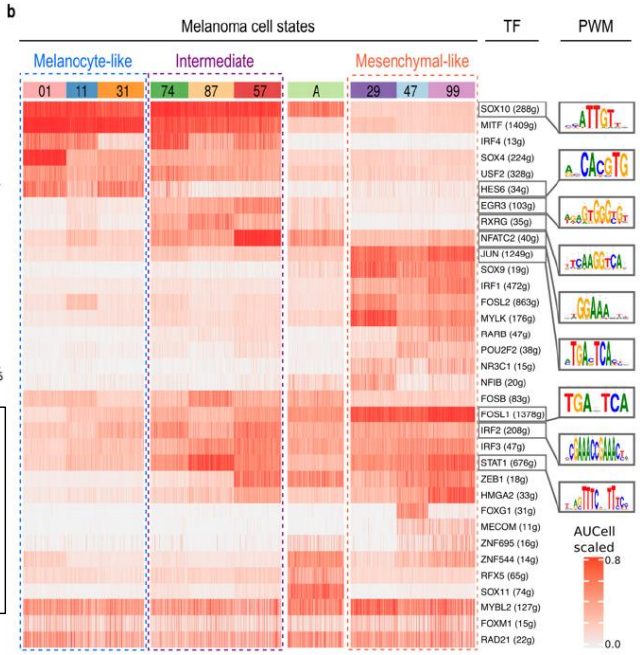
- Melanoma cells with a melanocytic phenotype can switch to a mesenchymal-like phenotype.
- Searching for intermediate state between them and decipher their underlying GRN, regulon-based analysis was applied on scRNA-seq of patient-derived melanoma cultures.
- GRN we identified may serve as a new putative target to prevent the switch to mesenchymal cell state and thereby, acquisition of metastatic and drug resistant potential.
- SCENIC predicts transcription factors (TFs) governing each melanoma cell state, alongside candidate transcription factor target genes (regulon).
- SCENIC yields a **regulon-cell matrix with regulon activities across all single cells**, and provides therefore an alternative dimensionality reduction. A UMAP visualization based on the regulon-cell matrix **reveals three candidate cell states** in an unsupervised manner, recapitulating findings based on count-cell matrix.



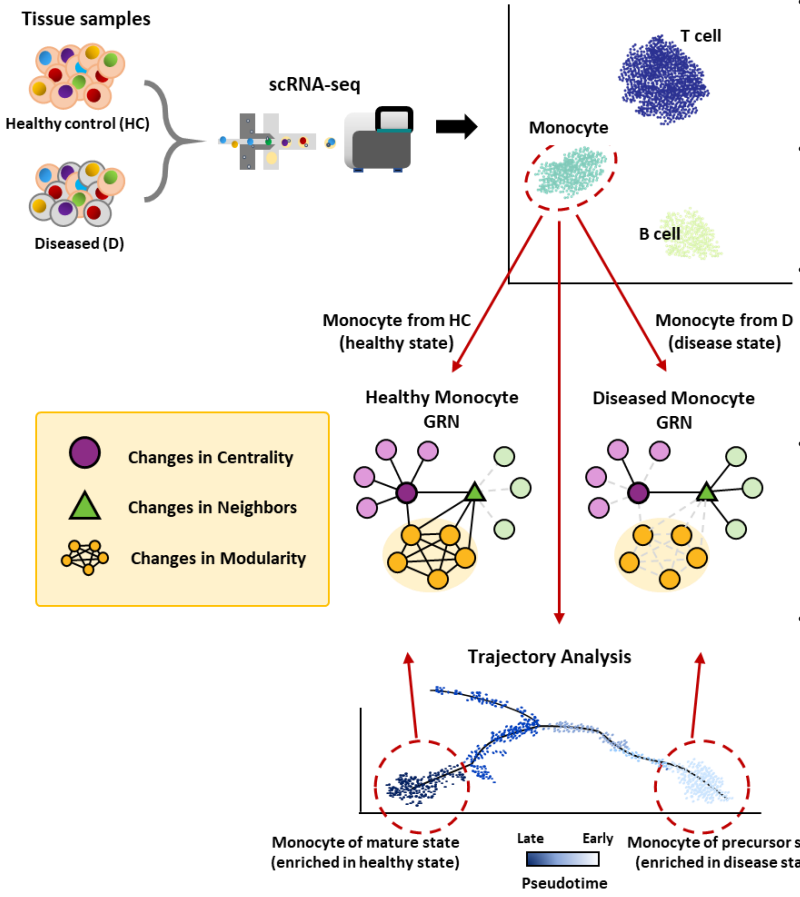
- Regulon activity analysis **revealed some regulons specific to each state**.
- The intermediate state shares several regulons with the melanocyte-like cell state or mesenchymal-like cell state.
- Some regulons are specific for intermediate state, including EGR3, RXRG and NFATC2. These TFs have previously been linked to a more aggressive/dedifferentiated phenotype in cancer and/or in melanoma specifically.
- GSEA of EGR3 targets: vasculature development and stem cells.
- GSEA of NFAT2 targets: wounding response, EMT and stemness



Violin plots showing the activity of SCENIC regulons specific for the intermediate cultures (bottom) and shared between the intermediate and the mesenchymal-like cultures, as measured by AUCell for all melanoma cultures.



❖ Hypothesis from network topology analysis



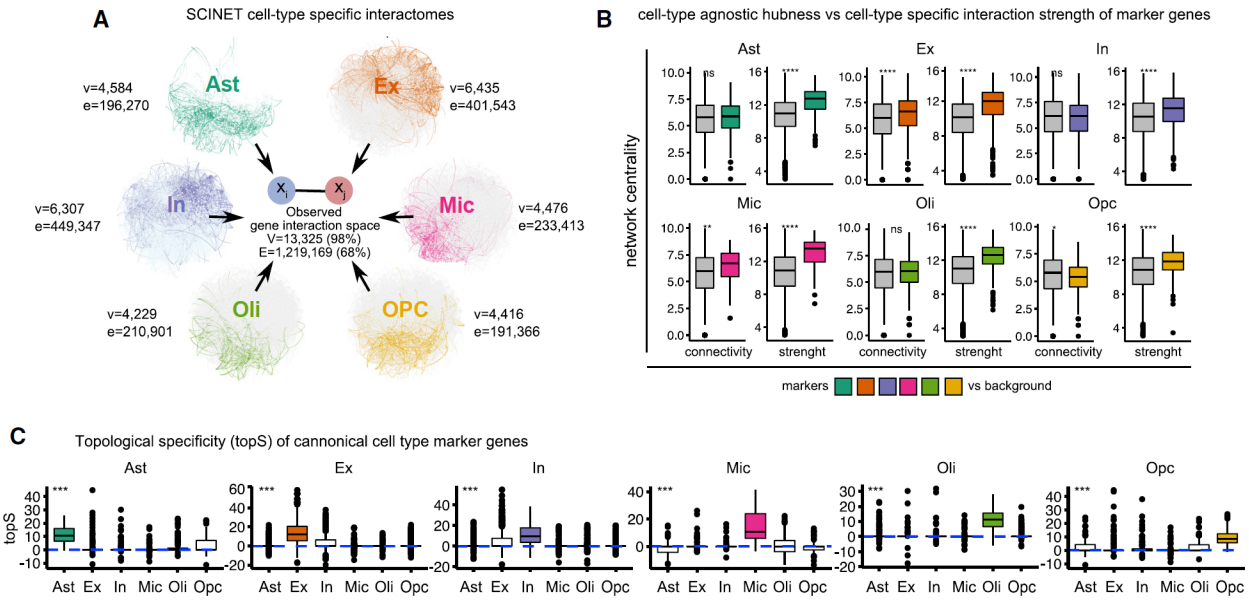
- Inference of co-regulatory from transcriptome profiles of cells from two distinct states (healthy control versus disease state) will construct different GRNs.
- Genes that show changes in three types of network topology are likely to be associated with the state: centrality, neighbors, and modularity.
- For example, correlation analyses for monocytes from healthy and disease samples may generate different network, and changes in three types of topology between healthy and disease states will be examined for every genes.
- Similarly, networks for different developmental time along with topological analysis would suggest disease-associated genes, because many disease states are associated with defect in development.
- For example, defect in maturation of monocyte into functional dendritic cells would result in immune disorders.

Cha and Lee. Exp. Mol. Med. (2020)

❖ Cell-type specific changes in centrality

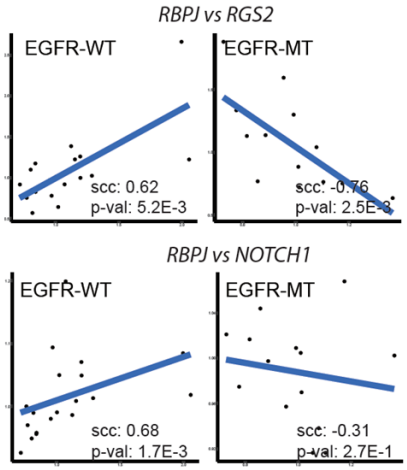
Cell Systems 9:559 (2019)

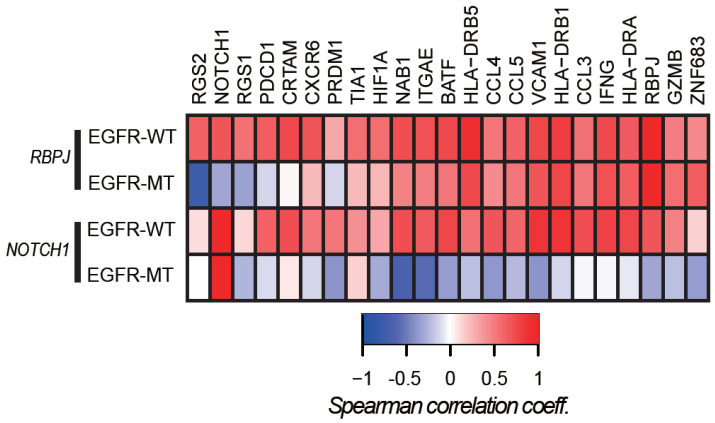
- Gene network for 6 brain cell types (astrocyte, excitatory neuron, inhibitory neuron, microglia, oligodendrocyte, oligodendrocyte progenitor) and a global network by their integration.
- In general, **marker genes do not show higher centrality** than other genes (cell-type agnostic centrality by global network). However, they **exhibit a significantly higher cell-type-specific centrality**.
- **Topological specificity** (topS): **measure of cell-type-specific centrality**



❖ Cell-type specific changes in neighbors (targets)

- Tissue-resident memory CD8+ T cells (Trm) are important for anti-tumor activity.
- Lung cancer patients with EGFR mutation (EGFR-MT) do not response to cancer immunotherapy compared with those with wild-type EGFR (EGFR-WT).
- EGFR-MT contains fewer tumor-infiltrating Trm cell than EGFR-WT, indicating that Trm cell state is impaired in EGFR-MT.
- NOTCH-RBPJ complex is a key regulator for Trm.
- In EGFR-MT, correlations between *NOTCH* (or *RBPJ*) and genes for differentiation and homeostasis of Trm are dysregulated.
- **The lose of co-expression between cell-type specific regulators and their normal targets causes Impairment of the cell function**.

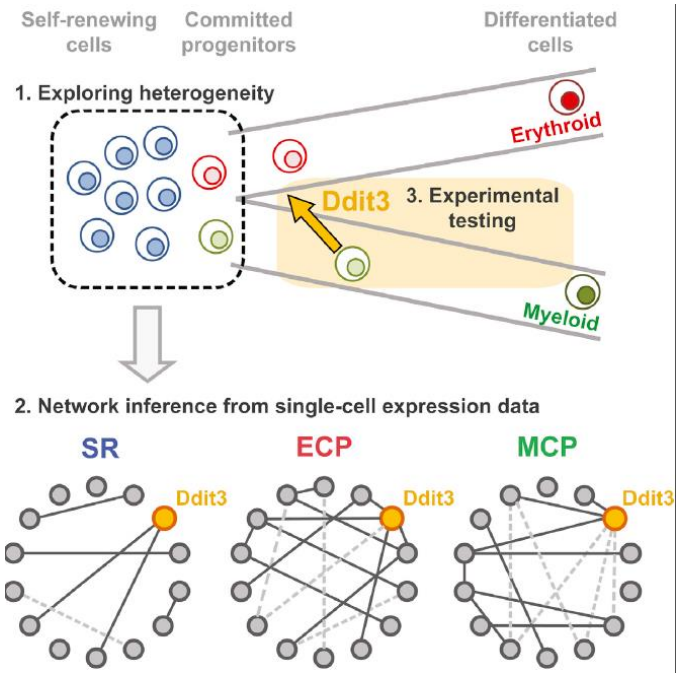


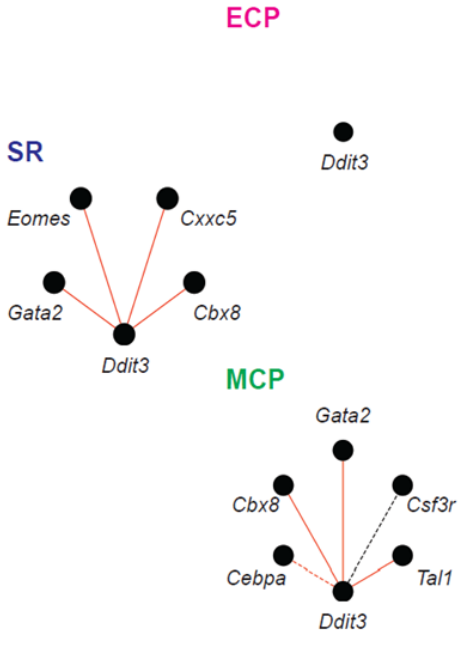


Cho et al. manuscript was submitted

❖ Cell-type specific changes in neighbors (targets)

- GRNs for self-renewing cells, erythroid committed progenitors and myeloid-committed progenitors, and demonstrated that **DDIT3 changes its targets in three different GRNs**.
- These results suggest that **DDIT3 is a lineage regulator**.
- Later, DDIT3 was experimentally validated as a lineage regulator.



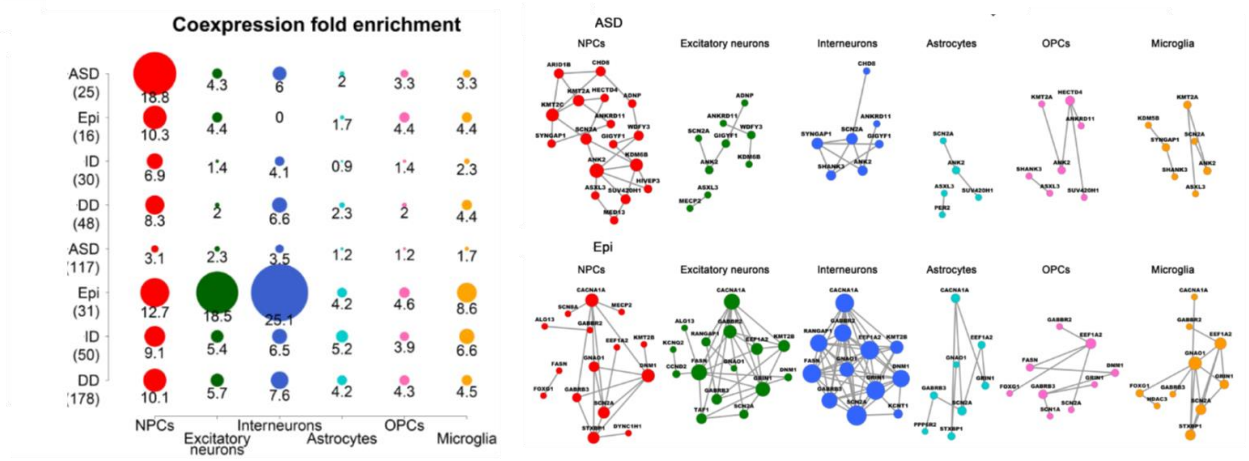


Cell Reports 10:1503 (2015)

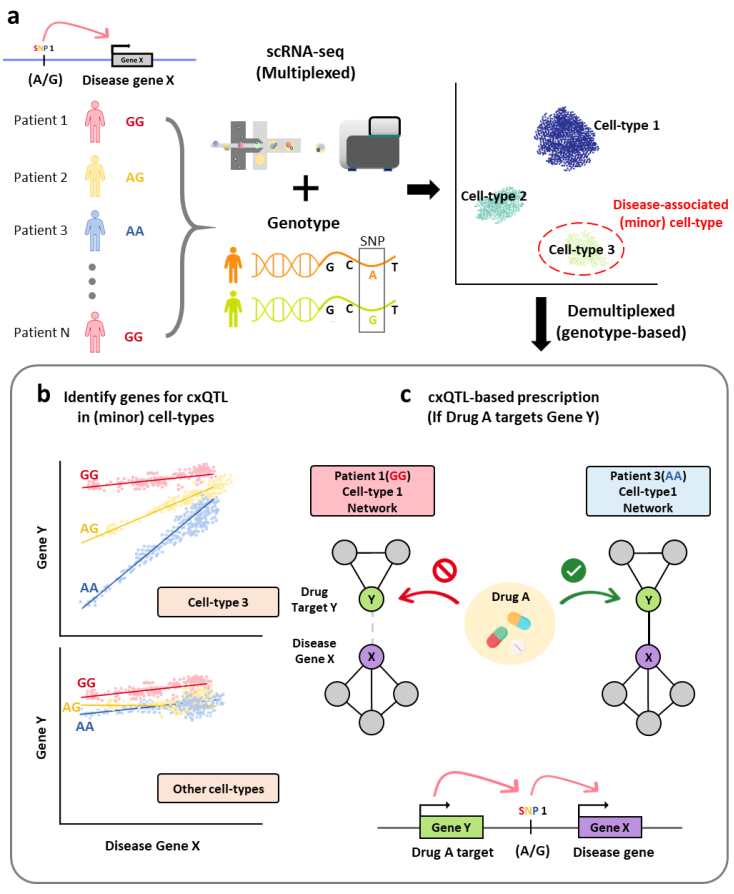
❖ Cell-type specific changes in modularity

Genome Res. 30:835 (2020)

- **Disease-associated genes tend to be connected by cell-type specific interactions**.
- Many neurodevelopmental disorder (NDD) genes have been identified through de novo mutation studies. However, effects of NDD genes on specific brain cell type are still elusive.
- **Co-expression enrichment** (a measure of **modularity**) among genes for autism spectrum disorder (ASD), epilepsy, intellectual disability (ID) and developmental delay (DD) in six **different brain cell types** (neural progenitor cells, excitatory neurons, interneurons, astrocytes, oligodendrocyte progenitor cells, microglia) were tested using scRNA-seq data and found cell-type specific effect of NDD genes.
- The results suggest that **disease genes tend to interact with cell-type-specific preference, with preferential cell types being targeted by the different disease classes**. For example, ASD and epilepsy genes specifically effect on NPCs and interneurons, respectively.
- **Cell-type-specific modularity of disease genes** reveals **cell-type-specific pathophysiology**.



❖ Hypothesis from genotype-network association



- Majority of disease-associated SNPs exert phenotypic effect via action of expression quantitative trait loci (eQTL) because most of them are located within noncoding regions.
- The eQTLs have long been suggested to exert its influence in a cell-specific manner.
- As scRNA-seq can provide transcriptome data for multiple cell types of the given tissue simultaneously, it can greatly facilitate cell-type-specific eQTL analysis (**Fig. a**).
- Interestingly, **some eQTL effects of a gene can be modified by expression of another gene** (**Fig. b**), called ***co-expression QTL***, because they turned out to affect co-regulatory relationship between two genes.
- For example, **effect of gene X eQTL depends on the expression of gene Y (e.g. Y is a TF for X)**.
- Single-cell transcriptome data from each person can be sufficient to infer gene-gene correlation, building personalized GRN. Thus, we may test whether personal genetic variations affect disease risk or drug response by altering co-regulatory interactions.
- If a co-regulatory interaction between **a disease gene (X)** and a **drug target (Y)** that affects the disease gene activity is modulated by a co-expression QTL, this genotype information would be utilized in tailored prescription for individual patients in the future (**Fig. c**).

Cha and Lee. Exp. Mol. Med. (2020)