PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Representing high throughput expression profiles via perturbation barcodes reveals compound targets

Tracey M. Filzen[1], Peter S. Kutchukian[2], Jeffrey D. Hermes[3], Jing Li[4], Matthew Tudor[5]*

1 Medical Writing, Merck Research Laboratories, Upper Gwynedd, Pennsylvania, United States of America, **2** Informatics, Merck Research Laboratories, Boston, Massachusetts, United States of America, **3** Screening & Protein Sciences, Merck Research Laboratories, North Wales, Pennsylvania, United States of America, **4** Screening and Compound Profiling, Merck Research Laboratories, Kenilworth, New Jersey, United States of America, **5** Informatics, Merck Research Laboratories, West Point, Pennsylvania, United States of America

* matthew_tudor@merck.com

Tracey M. Filzen, Peter S. Kutchukian, Jeffrey D. Hermes, Jing Li, and Matthew Tudor. Representing high throughput expression profiles via perturbation barcodes reveals compound targets.
PLOS Computational Biology, 13(2):e1005335, February 2017

Perturbation Barcodes

Overview

Goal Representation learning for transcriptomic analysis.

Perturbation Barcodes

Overview

Goal Representation learning for transcriptomic analysis.

Methods Metric Learning Network

Perturbation Barcodes

Overview

Goal	Representation learning for transcriptomic analysis.
Methods	Metric Learning Network
Data	L1000 Data, 2 Cell Lines, 3700 Compounds.

Perturbation Barcodes

Overview

Goal	Representation learning for transcriptomic analysis.
Methods	Metric Learning Network
Data	L1000 Data, 2 Cell Lines, 3700 Compounds.
Evaluation	Efficacy in various predictive or clustering tasks.

Perturbation Barcodes

Overview

Goal	Representation learning for transcriptomic analysis.
Methods	Metric Learning Network
Data	L1000 Data, 2 Cell Lines, 3700 Compounds.
Evaluation	Efficacy in various predictive or clustering tasks.
Code	`https://github.com/matudor/siamese`

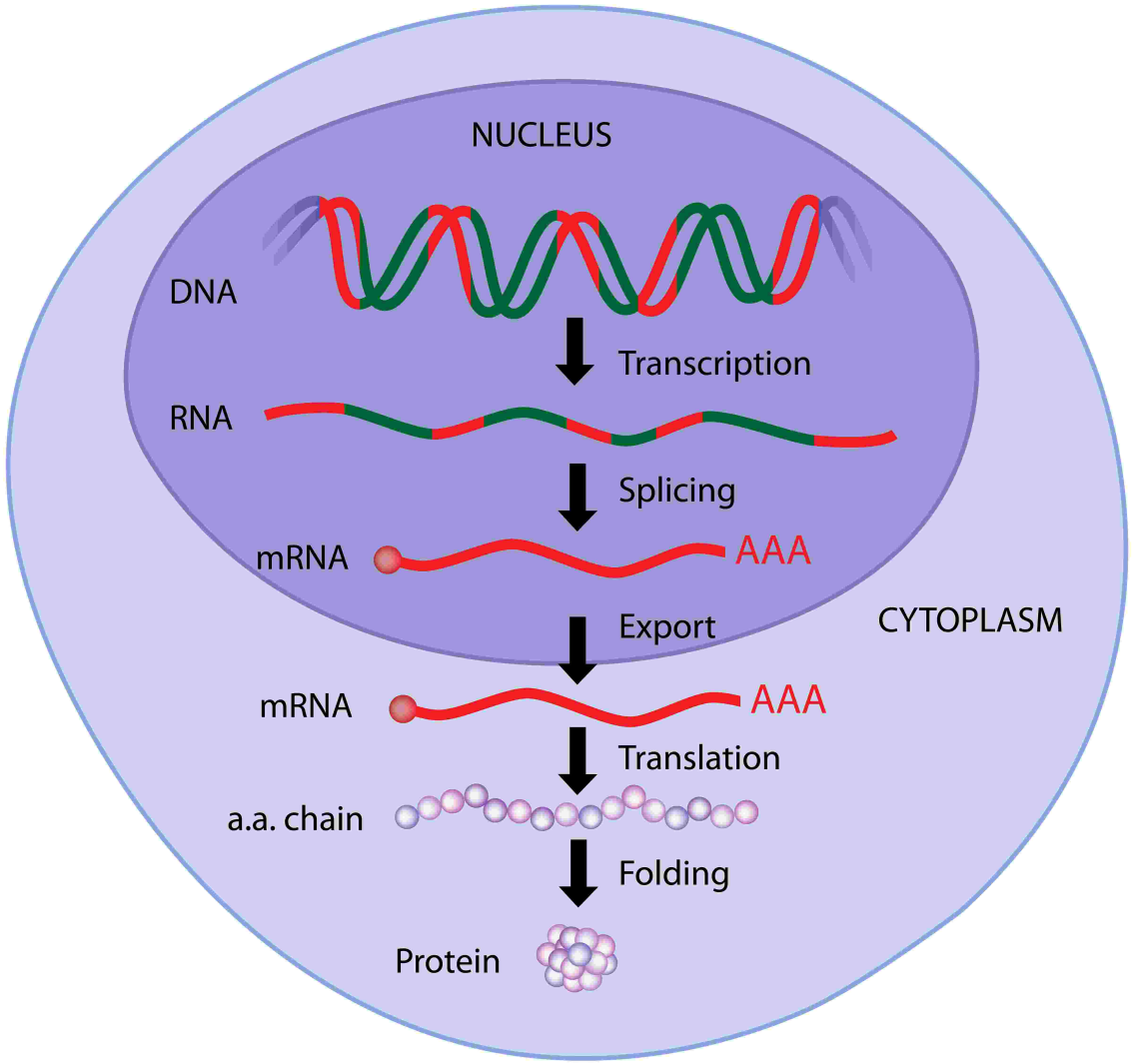
Gene Expression Data

Overview

Transcriptomic/Gene expression data measures counts of transcription factors inside perturbed cells.

Gene Expression Data

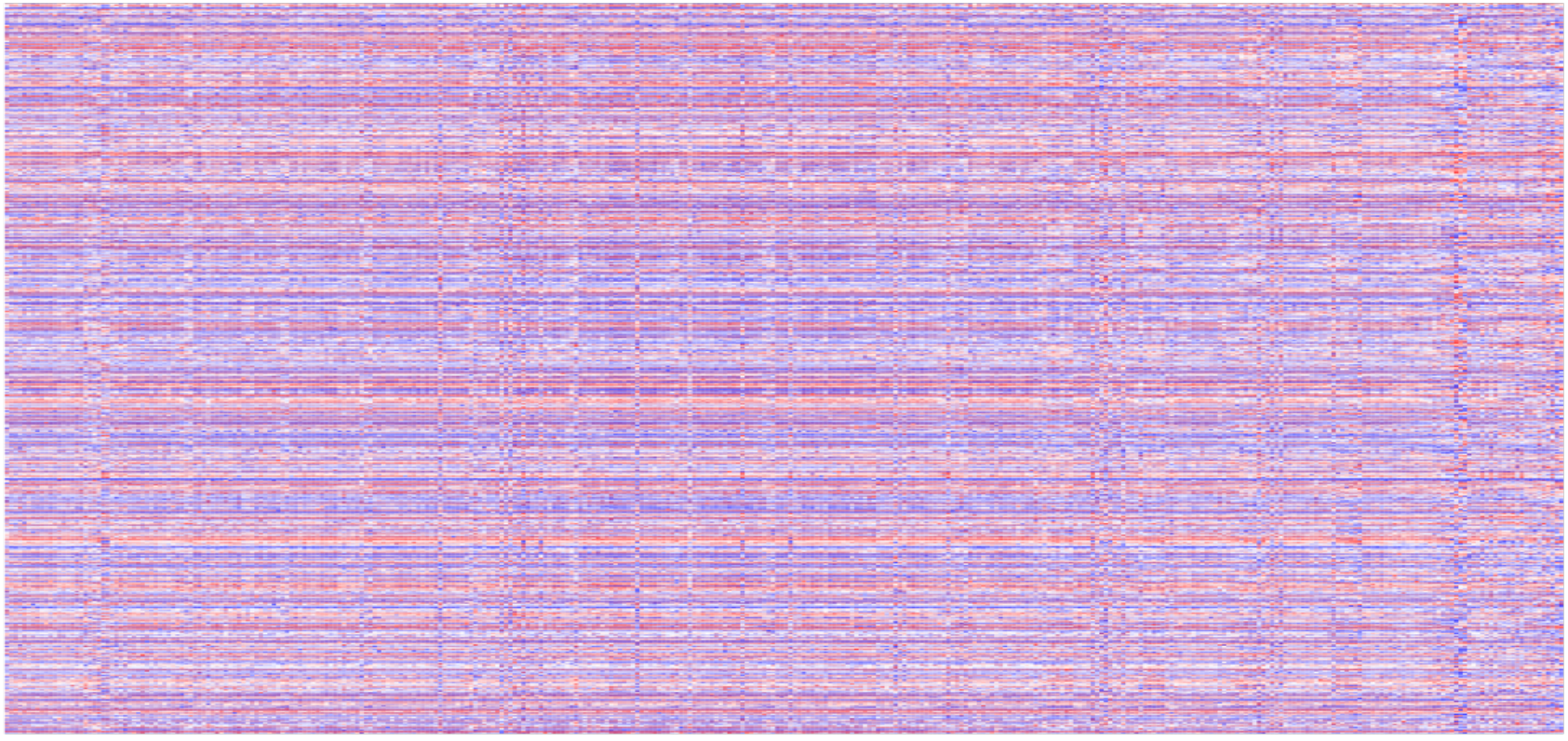
Overview



http://www.biology-questions-and-answers.com/images/Protein-Synthesis.jpg

Gene Expression Data

Overview



Gene Expression Data

Why we care

Gene expression vectors give a picture of the internal state of the cell at the time of measurement.

Datasets

Genometry Data (private)

- Two cell lines (PC3, ME180).
- 3699 known or potential bioactive compounds used (6h).
- Some compounds had multiple replicates, others did not.
- 14 months total of data collection.
- After quality control, 7573 total L1000 vectors.

Datasets

Genometry Data (private)

- Two cell lines (PC3, ME180).
- 3699 known or potential bioactive compounds used (6h).
- Some compounds had multiple replicates, others did not.
- 14 months total of data collection.
- After quality control, **7573** total L1000 vectors.

Datasets

LINCS Data (public)

- 15 cell lines
- 273 compounds, 2 doses, 24h.
- Data more heavily pre-processed.
- Data available: `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE70138`

Model

Side Task Representation Learning

Biological replicates vary widely:

Model

Side Task Representation Learning

Biological replicates vary widely:

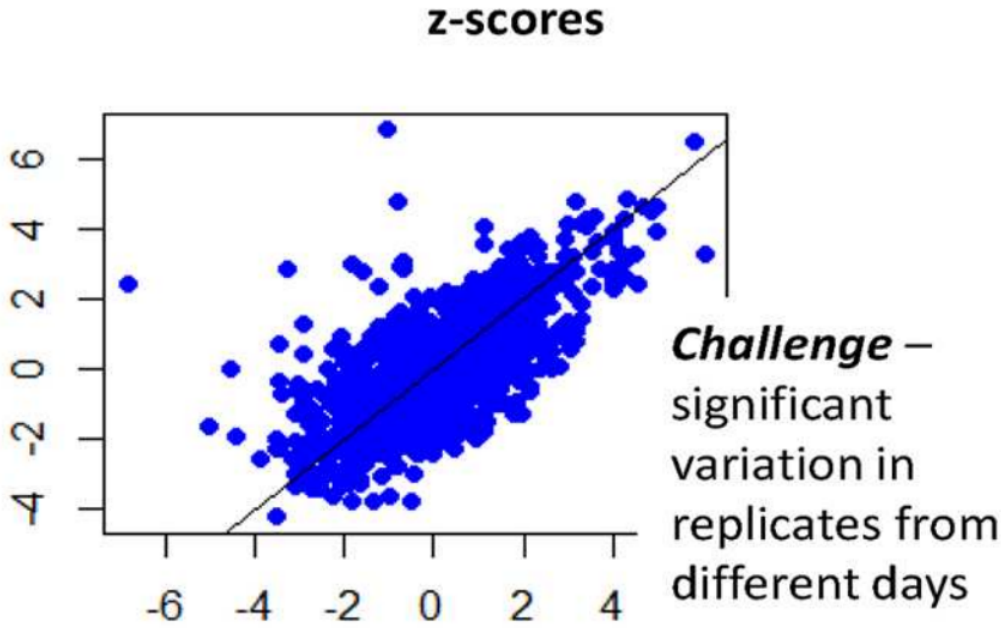


Figure 1 (partial), Page 4

Model

Side Task Representation Learning

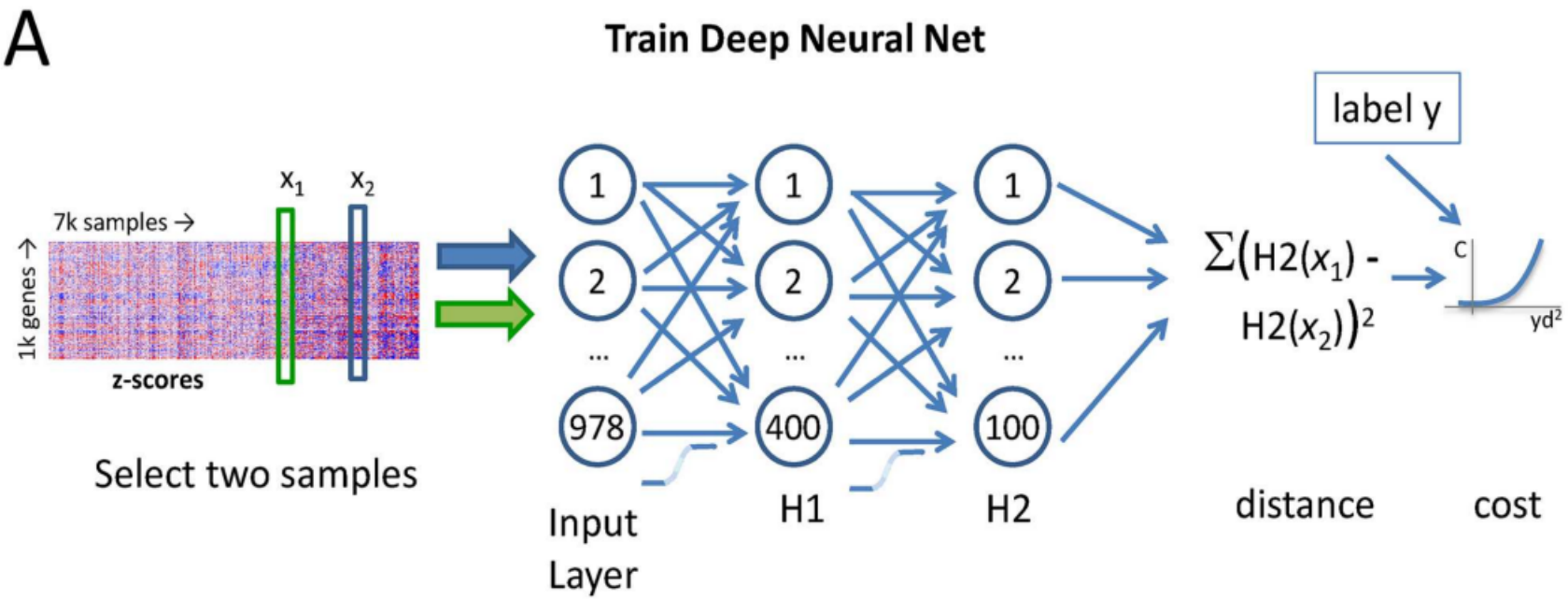
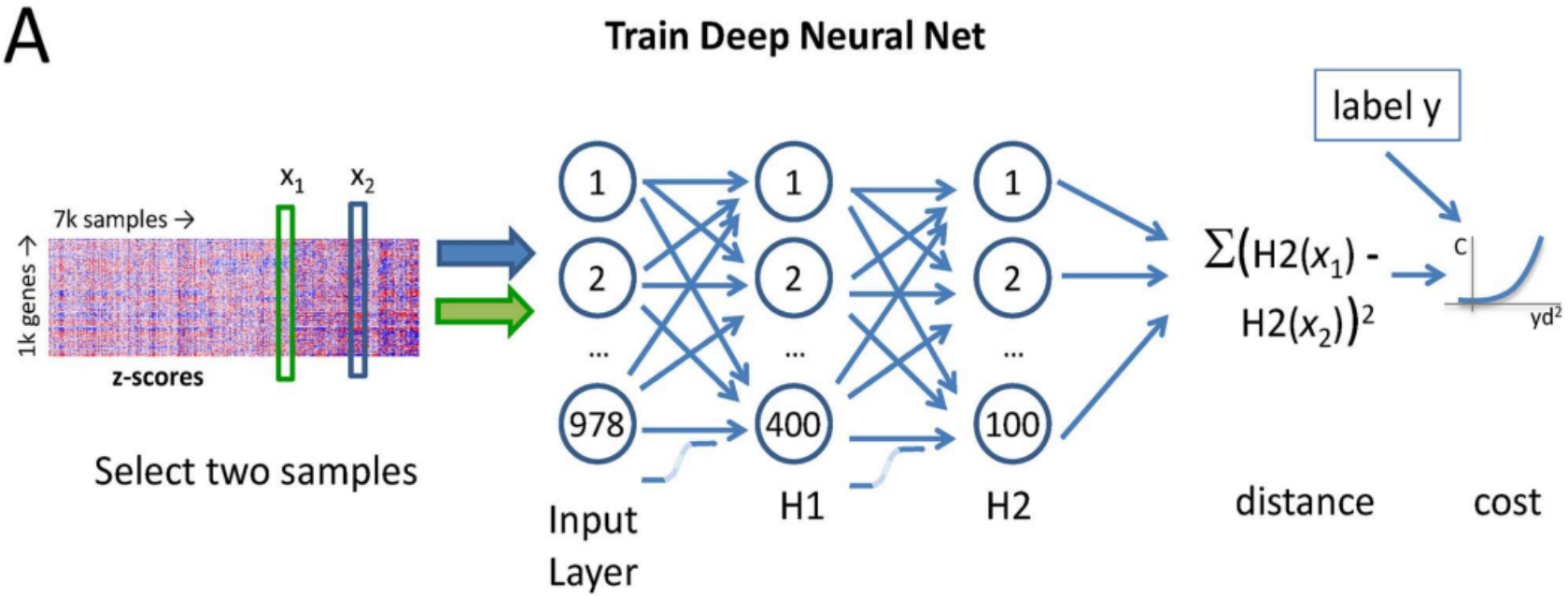


Figure 2 (partial), Page 6

Model

Side Task Representation Learning



$$c = \text{softplus}\left(1 - y\left(5 - \|\text{H2}(\boldsymbol{x}_1) - \text{H2}(\boldsymbol{x}_2)\|^2\right)\right).$$

Figure 2 (partial), Page 6

Model

Side Task Representation Learning

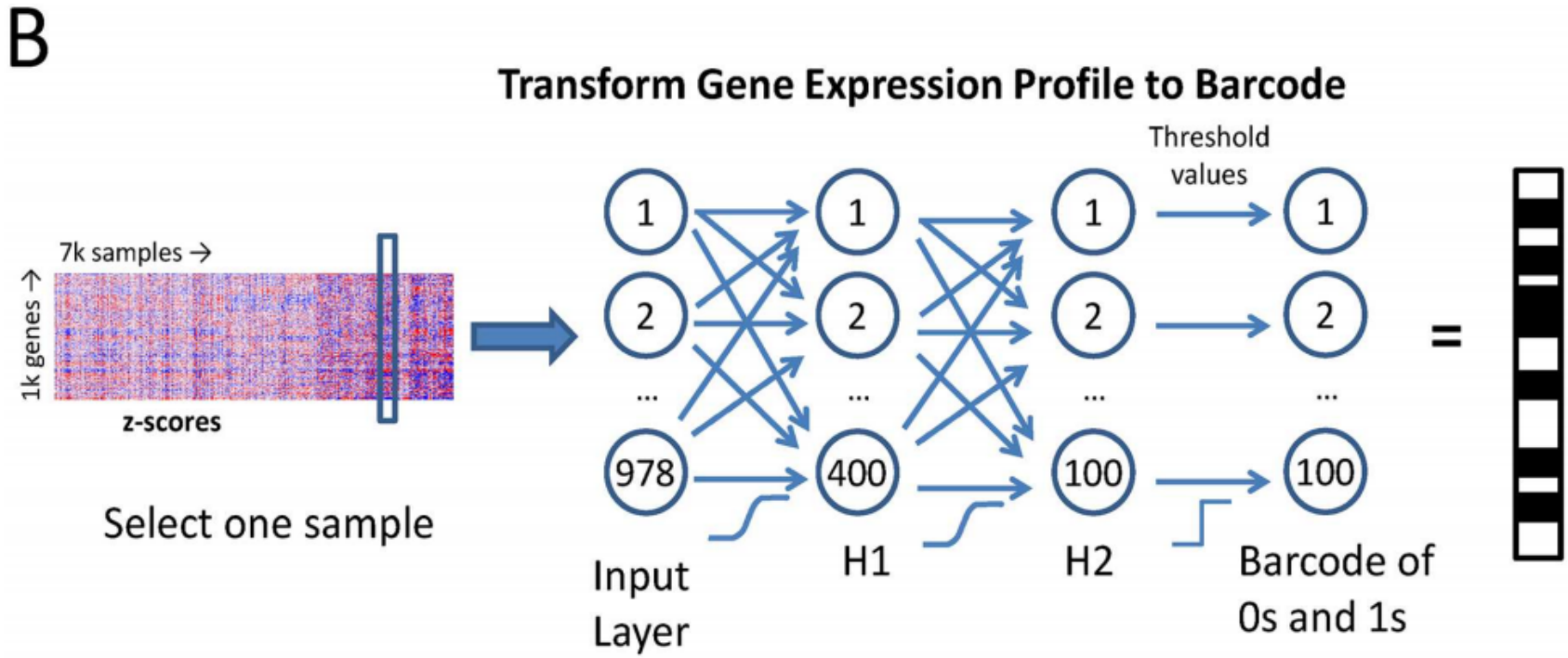


Figure 2 (partial), Page 6

Evaluation

Comparitive Tasks

Median rank of replicates How many samples on average are closer to a query sample than its replicates?

Evaluation

Comparitive Tasks

Median rank of replicates How many samples on average are closer to a query sample than its replicates?

Distance by shared target How different are the distances between pairs sharing a target vs. those that do not (t-statistic).

Evaluation

Comparitive Tasks

Median rank of replicates How many samples on average are closer to a query sample than its replicates?

Distance by shared target How different are the distances between pairs sharing a target vs. those that do not (t-statistic).

Structural clustering overlap How similar are compound structure based clustering and expression based clustering?

Evaluation

Comparitive Tasks

Median rank of replicates How many samples on average are closer to a query sample than its replicates?

Distance by shared target How different are the distances between pairs sharing a target vs. those that do not (t-statistic).

Structural clustering overlap How similar are compound structure based clustering and expression based clustering?

Correlation of HTS profiles How similar are phenotypic screens to gene expression induced representations?

Evaluation

Comparitive Tasks

Median rank of replicates How many samples on average are closer to a query sample than its replicates?

Distance by shared target How different are the distances between pairs sharing a target vs. those that do not (t-statistic).

Structural clustering overlap How similar are compound structure based clustering and expression based clustering?

Correlation of HTS profiles How similar are phenotypic screens to gene expression induced representations?

Promiscuity prediction How well do the various representations predict compound promiscuity?

Evaluation

Comparitive Tasks

Table 1. Performance of learned perturbation barcodes compared to z-scores and GSEA scores.

metric	z-score	GSEA	perturbation barcode
Median rank of replicates (of 7573)	225	72	**24**
Distance by shared target, t statistic	-1	-38	**-43**
Structural clustering overlap with expression clustering	0.01	0.03	**0.17**
Correlation of HTS profiles with expression	0.04	0.02	**0.12**
Promiscuity prediction by SVR, R^2	0.21	0.16	**0.34**

Evaluation

Exploratory Evaluation

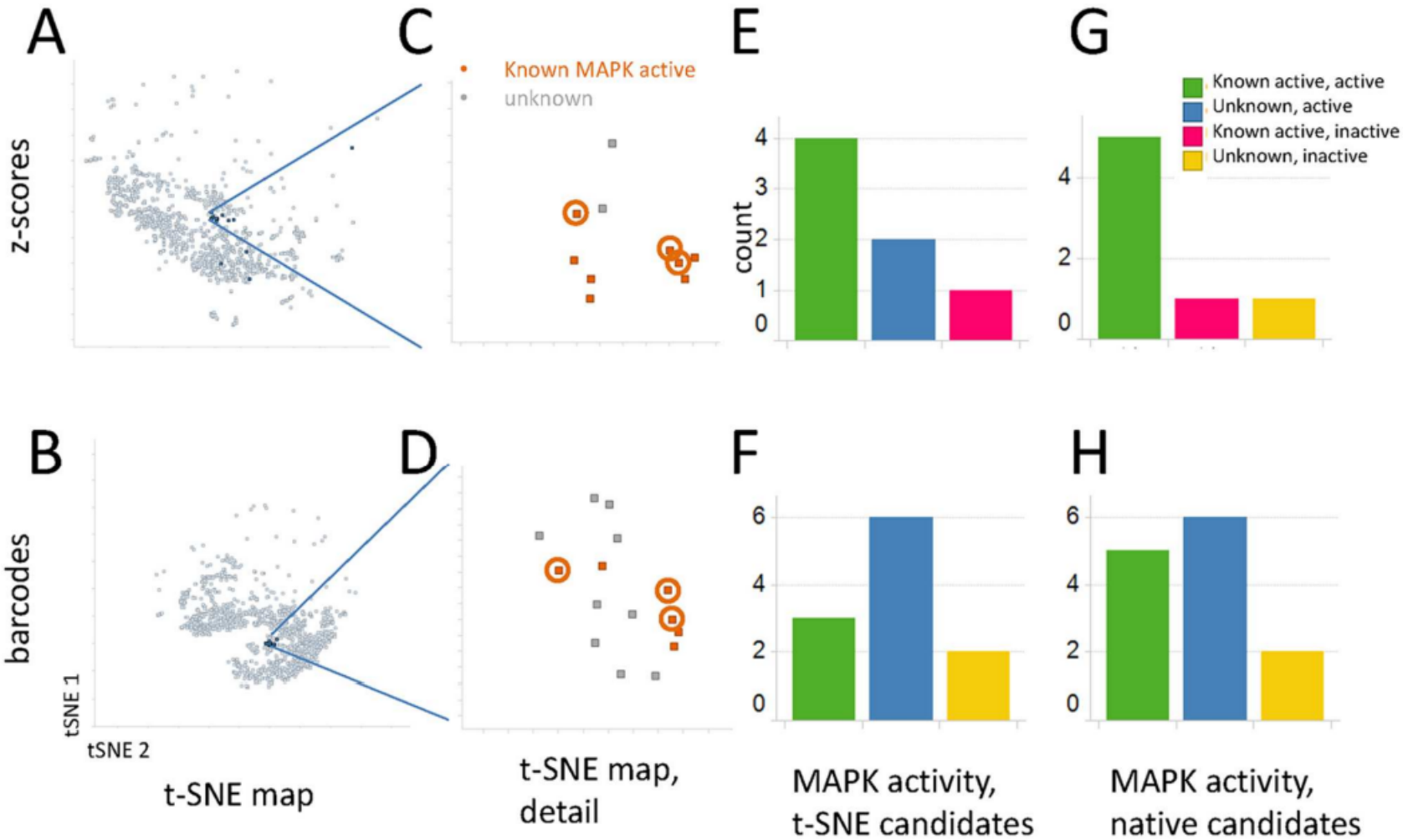


Figure 3, Page 11

Concluding Thoughts

Pros:

- Broad evaluation

Cons:

Concluding Thoughts

Pros:

- Broad evaluation
- Good choice of side task

Cons:

Concluding Thoughts

Pros:

- Broad evaluation
- Good choice of side task

Cons:

- Questions remain about network design.

Concluding Thoughts

Pros:

- Broad evaluation
- Good choice of side task

Cons:

- Questions remain about network design.
- Main results on private data.

Concluding Thoughts

Pros:

- Broad evaluation
- Good choice of side task

Cons:

- Questions remain about network design.
- Main results on private data.
- Small population size raises concerns about generalizability.

Concluding Thoughts

Pros:

- Broad evaluation
- Good choice of side task

Cons:

- Questions remain about network design.
- Main results on private data.
- Small population size raises concerns about generalizability.
- Generalizability not really tested.

Conclusion

Tracey M. Filzen, Peter S. Kutchukian, Jeffrey D. Hermes, Jing Li, and Matthew Tudor. Representing high throughput expression profiles via perturbation barcodes reveals compound targets.

PLOS Computational Biology, 13(2):e1005335, February 2017.