Supplement to Using Neural Networks for Reducing the Dimensions of Single-Cell RNA-Seq Data

Chieh Lin [1,*], Siddhartha Jain [2], Hannah Kim [3] and Ziv Bar-Joseph [1, 3,*]

1. Machine Learning Department, School of Computer Science, Carnegie Mellon University
2. Computer Science Department, School of Computer Science, Carnegie Mellon University
3. Computational Biology Department, School of Computer Science, Carnegie Mellon University

*To whom correspondence should be addressed.

1 Supplementary Methods

1.1 Details of clustering methods we compared to

1.1.1 SINCERA

We modify the demo.R file of SINCERA to perform clustering on whole dataset (without dimensionality reduction). The data processing steps before clustering are removed since they are not applicable to our dataset. We also set the parameters for the clustering as the default parameter in demo.R.

1.1.2 SNN-Cliq

We use default parameter for SNN-Cliq clustering on whole dataset (without dimensionality reduction).

1.1.3 pcaReduce

We apply pcaReduce to the whole dataset. We set the starting reduced dimension to 3 times the number of cell types. Maximum probability is selected as the merge method.

1.1.4 SIMLR

We use the python implementation of SIMLR with default parameters. The default parameter for the number of neighbors (30 neighbors) is not applicable when the number of cell is too low. In this case we set the number to be half of the number of input cells. redWe also try 20 or 40 neighbors for performance comparison.

2 Supplementary Tables

Supplementary Table 1: Summary of the 33 datasets

No.	doi	data_accession#	#sample	#cell type	tissue/cell type
1	10.1038/nature12172	GSE41265	18	1	BMDC
2	10.1186/gb-2013-14-4-r31	GSE42268	77	9	ESC
3	10.1126/science.1245316	GSE45719	317	24	embryonic cells
4	10.1038/nbt.3102	E-MTAB-2805	288	14	ESC
5	10.1186/s13059-016-0950-z	GSE76483	159	14	DRG
6	10.1073/pnas.1402030111	GSE47835	71	12	ESC MEF
7	10.1038/nature13173	GSE52583	201	26	distal lung epithelium
8	10.1016/j.stem.2014.11.005	GSE55291	94	20	iPS TTF ESC
9	10.1101/gr.177725.114	GSE57249	56	20	embryonic cells
10	10.1101/gr.171645.113	GSE60297	174	16	thymus TEC
11	10.1126/science.aaa1934	GSE60361	3005	15	celebral-cortex
12	10.15252/msb.20156198	GSE60768	107	11	ESC NSC
13	10.1038/nbt.3154	GSE61470	15	14	ESC PS NP HF
14	10.1038/cr.2015.149	GSE63576	209	14	DRG
15	10.1016/j.cub.2015.01.034	GSE64960	69	18	granulosa
16	10.1016/j.cell.2015.04.044.	GSE65525	8669	12	ESC
17	10.1016/j.devcel.2015.09.009.	GSE66202	91	19	kidney
18	10.1038/nbt.3443	GSE70844	83	14	neuron
19	10.1016/j.cell.2015.11.009	GSE75107	166	10	CNS/Th17
20	10.1016/j.cell.2015.11.009	GSE75108	136	9	LN/Th17
21	10.1016/j.cell.2015.11.009	GSE75109	139	17	spleen LN/Th17
22	10.1016/j.cell.2015.11.009	GSE75110	130	17	spleen LN/Th17
23	10.1016/j.cell.2015.11.009	GSE75111	151	17	spleen LN/Th17
24	10.1038/ncomms10220	GSE74923	194	31	cancer
25	10.1038/nature17997	GSE67120	181	30	HSC
26	10.1186/s12974-016-0581-z	GSE79510	45	21	brain
27	10.1038/celldisc.2016.10	GSE70605	145	30	embryonic cells
28	10.1182/blood-2016-05-716480	GSE81682	1920	21	HSC
29	10.1172/JCI77378	GSE66578	6	17	lung
30	10.1038/ni.3437	GSE74596	203	19	thymus
31	10.1038/ni.3412	GSE77029	64	22	bone marrow
32	10.1016/j.devcel.2016.02.020	GSE65924	70	16	embryonic cells
33	10.1038/ncomms11075	GSE70657	135	16	HSC

Supplementary Table 2: The datasets for each query cell type in the retrieval analysis

cell type	dataset No.
HSC (Hematopoietic stem cells)	25 28 33
4cell	3 9 27
ICM (Inner Cell Mass)	9 27
spleen	21 22 23
8cell	3 27
neuron	5 11 14 18 19 23 26
zygote	3 9 27
2cell	3 9 27
ESC (Embryonic stem cell)	4 6 8 12 13 16

Supplementary Table 3: Average clustering performance of different scoring metrics with 2 cell types in testing set for 20 clustering experiments (using different random initialization. Homo: homogeneity, Comp: Completeness, Vmes: v-measure, ARI: adjusted random index, AMI: adjusted mutual information, FM: Fowlkes-Mallows

Feature	**Homo**	**Comp**	**Vmes**	**ARI**	**AMI**	**FM**	**Averagge**
original	0.889	0.891	0.889	0.891	0.881	0.973	0.902
pca 2	0.947	0.946	0.946	0.947	0.941	0.979	0.951
tsne 2	0.038	0.032	0.034	0.017	0.015	0.589	0.121
ica 2	0.785	0.795	0.787	0.787	0.78	0.959	0.816
nmf 2	0.649	0.656	0.648	0.632	0.629	0.892	0.684
pca 5	0.947	0.946	0.946	0.947	0.941	0.979	0.951
tsne 5	0.051	0.056	0.048	0.015	0.027	0.613	0.135
ica 5	0.569	0.603	0.577	0.57	0.559	0.89	0.628
nmf 5	0.735	0.745	0.736	0.73	0.717	0.925	0.765
pca 10	0.906	0.912	0.909	0.902	0.905	0.976	0.918
tsne 10	0.025	0.019	0.021	-0.009	0.002	0.56	0.103
ica 10	0.201	0.29	0.214	0.19	0.183	0.806	0.314
nmf 10	0.612	0.604	0.6	0.586	0.574	0.872	0.641
pca 50	0.889	0.891	0.889	0.891	0.881	0.973	0.902
nmf 50	0.349	0.394	0.363	0.336	0.332	0.834	0.435
pca 100	0.889	0.891	0.889	0.891	0.881	0.973	0.902
nmf 100	0.302	0.357	0.315	0.291	0.28	0.796	0.39
pcaReduce	0.813	0.82	0.815	0.803	0.81	0.96	0.837
SIMLR_20	0.918	0.914	0.915	0.909	0.909	0.963	0.921
SIMLR_30	0.854	0.861	0.853	0.852	0.84	0.951	0.868
SIMLR_40	0.815	0.827	0.816	0.807	0.8	0.933	0.833
SNN-Cliq	0.0	0.0	0.0	0.0	0.0	0.0	0.0
sincera_hc	0.97	0.961	0.965	0.98	0.959	0.994	0.972
Dense 1layer 100	0.966	0.955	0.96	0.978	0.953	0.993	0.967
Dense 1layer 796	0.936	0.931	0.933	0.939	0.925	0.977	0.94
Dense 2layer 796/100	0.961	0.949	0.955	0.975	0.947	0.992	0.963
PPITF 1layer 696+100	0.956	0.943	0.949	0.968	0.941	0.989	0.958
PPITF 2layer 696+100/100	**0.976**	**0.966**	**0.971**	**0.984**	**0.965**	**0.996**	**0.976**
Dense 1layer 100 pretrain	0.934	0.919	0.926	0.935	0.917	0.976	0.935
Dense 1layer 796 pretrain	0.936	0.931	0.933	0.939	0.925	0.977	0.94
Dense 2layer 796/100 pretrain	0.97	0.961	0.965	0.98	0.959	0.994	0.972
PPITF 1layer 696+100 pretrain	0.936	0.931	0.933	0.939	0.925	0.977	0.94
PPITF 2layer 696+100/100 pretrain	0.97	0.961	0.965	0.98	0.959	0.994	0.972

Supplementary Table 4: Average clustering performance of different scoring metrics with 6 cell types in testing set for 20 clustering experiments (using different random initialization. Homo: homogeneity, Comp: Completeness, Vmes: v-measure, ARI: adjusted random index, AMI: adjusted mutual information, FM: Fowlkes-Mallows

Feature	**Homo**	**Comp**	**Vmes**	**ARI**	**AMI**	**FM**	**Averagge**
original	0.724	0.879	0.789	0.62	0.709	0.747	0.745
pca 2	0.828	0.853	0.839	0.755	0.809	0.822	0.818
tsne 2	0.229	0.218	0.222	0.131	0.166	0.321	0.215
ica 2	0.819	0.846	0.831	0.747	0.799	0.815	0.809
nmf 2	0.597	0.698	0.641	0.473	0.566	0.628	0.601
pca 5	0.746	0.881	0.804	0.65	0.732	0.766	0.763
tsne 5	0.053	0.054	0.053	-0.002	-0.005	0.229	0.064
ica 5	0.73	0.867	0.789	0.625	0.715	0.747	0.745
nmf 5	0.674	0.782	0.717	0.6	0.646	0.722	0.69
pca 10	0.738	0.885	0.8	0.63	0.723	0.756	0.755
tsne 10	0.062	0.056	0.059	-0.001	-0.0	0.2	0.063
ica 10	0.659	0.844	0.732	0.582	0.64	0.729	0.698
nmf 10	0.703	0.807	0.743	0.631	0.684	0.743	0.719
pca 50	0.746	0.864	0.798	0.65	0.733	0.759	0.758
nmf 50	0.538	0.714	0.607	0.459	0.511	0.637	0.578
pca 100	0.708	0.89	0.781	0.612	0.693	0.75	0.739
nmf 100	0.488	0.638	0.542	0.389	0.458	0.588	0.517
pcaReduce	0.769	0.848	0.803	0.693	0.753	0.78	0.774
SIMLR_20	0.767	0.743	0.754	0.609	0.719	0.699	0.715
SIMLR_30	0.79	0.796	0.791	0.671	0.75	0.752	0.758
SIMLR_40	0.798	0.784	0.79	0.641	0.753	0.727	0.749
SNN-Cliq	0.716	0.895	0.787	0.661	0.702	0.772	0.755
sincera_hc	0.773	**0.931**	0.838	0.738	0.761	0.827	0.811
Dense 1layer 100	0.854	0.851	0.852	0.769	0.824	0.827	0.83
Dense 1layer 796	**0.858**	0.862	**0.86**	**0.783**	**0.835**	**0.839**	**0.839**
Dense 2layer 796/100	0.858	0.842	0.849	0.771	0.821	0.825	0.828
PPITF 1layer 696+100	0.839	0.855	0.846	0.745	0.811	0.812	0.818
PPITF 2layer 696+100/100	0.854	0.85	0.851	0.767	0.824	0.824	0.828
Dense 1layer 100 pretrain	0.858	0.845	0.851	0.775	0.827	0.829	0.831
Dense 1layer 796 pretrain	0.848	0.839	0.843	0.759	0.817	0.819	0.821
Dense 2layer 796/100 pretrain	0.846	0.839	0.842	0.756	0.815	0.815	0.819
PPITF 1layer 696+100 pretrain	0.835	0.845	0.839	0.763	0.809	0.823	0.819
PPITF 2layer 696+100/100 pretrain	0.849	0.831	0.839	0.762	0.812	0.817	0.818

Supplementary Table 5: Average clustering performance of different scoring metrics with 8 cell types in testing set for 20 clustering experiments (using different random initialization. Homo: homogeneity, Comp: Completeness, Vmes: v-measure, ARI: adjusted random index, AMI: adjusted mutual information, FM: Fowlkes-Mallows

Feature	**Homo**	**Comp**	**Vmes**	**ARI**	**AMI**	**FM**	**Averagge**
original	0.701	0.883	0.777	0.55	0.68	0.683	0.712
pca 2	0.792	0.84	0.815	0.686	0.772	0.754	0.776
tsne 2	0.257	0.246	0.251	0.131	0.19	0.275	0.225
ica 2	0.787	0.84	0.812	0.679	0.766	0.751	0.772
nmf 2	0.553	0.651	0.596	0.383	0.521	0.524	0.538
pca 5	0.814	0.893	**0.85**	0.701	0.793	**0.774**	0.804
tsne 5	0.053	0.06	0.056	-0.005	-0.008	0.193	0.058
ica 5	0.808	0.884	0.843	0.692	0.789	0.764	0.797
nmf 5	0.675	0.771	0.715	0.538	0.647	0.649	0.666
pca 10	0.758	0.889	0.816	0.627	0.741	0.725	0.759
tsne 10	0.073	0.066	0.069	-0.001	-0.003	0.153	0.06
ica 10	0.686	0.853	0.759	0.55	0.664	0.675	0.698
nmf 10	0.704	0.816	0.753	0.599	0.683	0.694	0.708
pca 50	0.757	0.882	0.812	0.622	0.736	0.722	0.755
nmf 50	0.588	0.736	0.651	0.484	0.563	0.612	0.606
pca 100	0.74	0.887	0.805	0.609	0.719	0.716	0.746
nmf 100	0.516	0.676	0.582	0.388	0.486	0.545	0.532
pcaReduce	0.779	0.861	0.817	0.681	0.763	0.751	0.776
SIMLR_20	0.763	0.759	0.759	0.58	0.717	0.662	0.707
SIMLR_30	0.79	0.786	0.788	0.634	0.753	0.703	0.742
SIMLR_40	0.766	0.77	0.767	0.594	0.73	0.671	0.717
SNN-Cliq	0.546	0.893	0.668	0.435	0.53	0.618	0.615
sincera_hc	0.739	**0.929**	0.821	0.674	0.723	0.766	0.775
Dense 1layer 100	0.816	0.799	0.806	0.658	0.77	0.722	0.762
Dense 1layer 796	**0.841**	0.852	0.846	**0.718**	**0.813**	0.774	**0.807**
Dense 2layer 796/100	0.819	0.801	0.809	0.659	0.771	0.722	0.763
PPITF 1layer 696+100	0.823	0.854	0.837	0.693	0.793	0.758	0.793
PPITF 2layer 696+100/100	0.821	0.807	0.813	0.668	0.778	0.73	0.77
Dense 1layer 100 pretrain	0.839	0.817	0.827	0.677	0.792	0.737	0.782
Dense 1layer 796 pretrain	0.826	0.83	0.827	0.694	0.794	0.754	0.788
Dense 2layer 796/100 pretrain	0.809	0.815	0.811	0.657	0.773	0.724	0.765
PPITF 1layer 696+100 pretrain	0.83	0.866	0.846	0.706	0.806	0.769	0.804
PPITF 2layer 696+100/100 pretrain	0.839	0.821	0.829	0.693	0.797	0.75	0.788

Supplementary Table 6: Mean performance of different scoring metrics with mean clustering performance for all number of cell types for 20 clustering experiments (using random subsets). Homo: homogeneity, Comp: Completeness, Vmes: v-measure, ARI: adjusted random index, AMI: adjusted mutual information, FM: Fowlkes-Mallows

Feature	**Homo**	**Comp**	**Vmes**	**ARI**	**AMI**	**FM**	**Averagge**
original	0.775	0.882	0.819	0.698	0.761	0.811	0.791
pca 2	0.85	0.881	0.863	0.793	0.836	0.857	0.847
tsne 2	0.176	0.166	0.17	0.098	0.124	0.394	0.188
ica 2	0.806	0.841	0.821	0.75	0.791	0.85	0.81
nmf 2	0.614	0.685	0.644	0.52	0.588	0.698	0.625
pca 5	0.815	0.896	0.85	0.746	0.802	0.833	0.824
tsne 5	0.048	0.051	0.048	-0.0	0.002	0.342	0.082
ica 5	0.692	0.798	0.736	0.626	0.678	0.797	0.721
nmf 5	0.697	0.786	0.732	0.64	0.675	0.78	0.718
pca 10	0.79	0.89	0.833	0.714	0.779	0.82	0.804
tsne 10	0.052	0.046	0.049	-0.003	0.001	0.303	0.075
ica 10	0.499	0.68	0.562	0.434	0.48	0.729	0.564
nmf 10	0.684	0.76	0.712	0.623	0.657	0.78	0.703
pca 50	0.771	0.876	0.815	0.7	0.757	0.813	0.789
nmf 50	0.489	0.617	0.538	0.427	0.466	0.691	0.538
pca 100	0.77	0.887	0.817	0.702	0.756	0.817	0.791
nmf 100	0.426	0.576	0.478	0.356	0.399	0.651	0.481
pcaReduce	0.782	0.855	0.814	0.731	0.77	0.835	0.798
SIMLR_20	0.81	0.806	0.807	0.704	0.779	0.786	0.782
SIMLR_30	0.81	0.816	0.811	0.726	0.78	0.811	0.792
SIMLR_40	0.803	0.803	0.801	0.697	0.77	0.791	0.778
SNN-Cliq	0.671	0.898	0.752	0.604	0.652	0.744	0.72
sincera_hc	0.822	**0.937**	0.87	0.797	0.809	0.868	0.851
Dense 1layer 100	0.885	0.876	0.88	0.819	0.858	0.864	0.864
Dense 1layer 796	0.883	0.883	0.882	0.824	0.862	**0.873**	**0.868**
Dense 2layer 796/100	0.882	0.868	0.875	0.815	0.852	0.861	0.859
PPITF 1layer 696+100	0.879	0.887	0.882	0.818	0.857	0.868	0.865
PPITF 2layer 696+100/100	0.889	0.881	**0.885**	0.823	**0.864**	0.867	0.868
Dense 1layer 100 pretrain	0.879	0.863	0.87	0.806	0.848	0.859	0.854
Dense 1layer 796 pretrain	0.874	0.868	0.87	0.809	0.849	0.861	0.855
Dense 2layer 796/100 pretrain	0.877	0.873	0.874	0.808	0.853	0.857	0.857
PPITF 1layer 696+100 pretrain	0.87	0.879	0.873	0.81	0.849	0.864	0.858
PPITF 2layer 696+100/100 pretrain	**0.89**	0.877	0.883	**0.826**	0.862	0.869	0.868

Supplementary Table 9: The average testing performance of NN training with different parameter settings of SGD

parameter setting	**accuracy**	**loss**
learning rate: 0.01 momentum: 0.5 decay: 0.001	0.925	0.355
learning rate: 0.01 momentum: 0.5 decay: 1e-06	0.925	0.3399
learning rate: 0.01 momentum: 0.9 decay: 0.001	0.925	0.355
learning rate: 0.01 momentum: 0.9 decay: 1e-06	0.925	0.3399
learning rate: 0.1 momentum: 0.5 decay: 0.001	0.95	0.3509
learning rate: 0.1 momentum: 0.5 decay: 1e-06	0.95	0.3396
learning rate: 0.1 momentum: 0.9 decay: 0.001	0.95	0.3509
learning rate: 0.1 momentum: 0.9 decay: 1e-06	0.95	0.3396

Supplementary Table 7: Average ratio of mean absolute error compared to 0 drop-out rate

missing rate	**drop-out rate**	**ratio of mean absolute error**
0.05	0	1
0.05	0.01	0.999
0.05	0.03	1.015
0.05	0.05	1.029
0.1	0	1
0.1	0.01	1.005
0.1	0.03	1.019
0.1	0.05	1.034

Note that the ratio is compared to the 0 drop-out rate with the same missing rate. This experiment shows that applying drop-out after imputation does not improve the performance (reduce the mean absolute error) and yields similar results.

Supplementary Table 8: The cell types of the datasets used in training and clustering experiment

BMDC (Bone Marrow-derived Dendritic Cells)
ES (embryonic stem cells)
PrE (primitive endoderm)
late2cell
earlyblast
midblast
8cell
4cell
16cell
mid2cell
lateblast
zygote
2cell
fibroblast
C57 2cell
BXC (liver cells)

Supplementary Table 10: The p-value comparison of some examples of highly ranked GO function for some of the cell types found both by deeplift and our method

Cell type	**GO function**	**p-value of DeepLift**	**p-value of our method**
ES	Factor: E2F-3;	1.31E-06	4.44E-15
ES	Factor: IRF6;	1.65E-02	5.78E-12
BDMCs	immune system process	5.10E-12	1.56E-05
BDMCs	positive regulation of immune system process	5.91E-09	1.59E-11
BDMCs	response to cytokine	1.63E-03	2.06E-06
fibroblast	Focal adhesion	-	3.16E-06
fibroblast	acetyltransferase complex	-	2.85E-20
zygote	regulation of cell proliferation	-	6.85E-07
zygote	cell junction	-	7.36E-09
zygote	TF: foxd3	-	significant

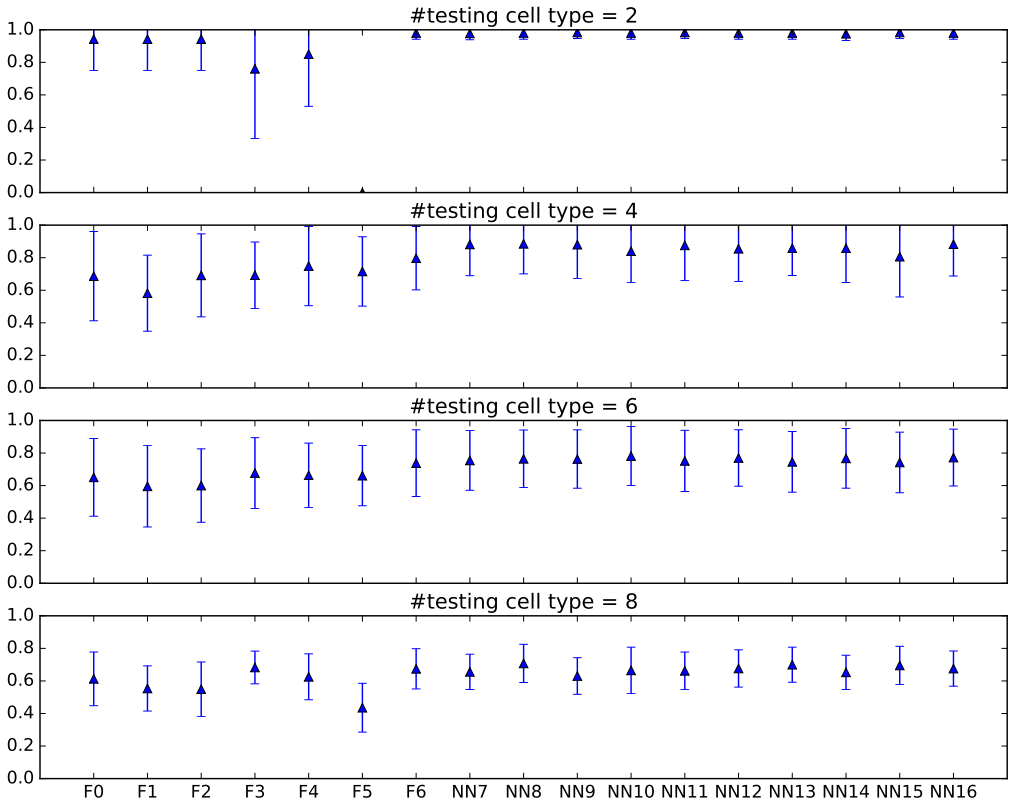
Supplementary Table 11: Significant gene list and repeat counts for pretrained model (100 dense nodes)

rpl27a	36	rpl23a	4	srsf7	3	rplp2	2	rps17	1	rplp1	1
rps12	24	rplp0	4	eef1b2	3	klhl7	2	s100a6	1	laptm5	1
rpl37a	21	rpsa	4	arpc1b	2	ccdc78	2	psmb8	1	lgals1	1
rpl15	12	cox7c	4	bard1	2	ncl	2	rpl26	1	mapt	1
rpl7l1	9	rps24	4	rps16	2	aak1	1	atp5b	1	mpv17l	1
loxl2	9	rps14	4	calm1	2	rps19	1	cd47	1	mrpl43	1
pwwp2a	6	rab3a	4	polr2l	2	rps29	1	coro1a	1	naca	1
ankfy1	6	stmn3	4	rps11	2	rrm2	1	cp	1	nsf	1
uba52	6	rps4x	4	rpl28	2	sec62	1	ddx5	1	pfn1	1
serinc1	6	ppia	3	itm2b	2	serinc3	1	eif3f	1	prex2	1
slc25a4	5	kif5c	3	rpl36	2	shc2	1	eno2	1	psat1	1
cd3eap	5	hsp90ab1	3	rpl18	2	sub1	1	gsn	1	rac2	1
app	5	rpl10	3	rpl34	2	syt1	1	hsp90aa1	1	rpl10a	1
fau	5	arhgdib	3	cpm	2	tecr	1	hspe1	1	ldha	1
exosc2	5	rpl4	3	gapdh	2	trim28	1	igf2bp1	1	wnk1	1
slc25a5	5										

Supplementary Table 12: GO analysis results for the significant genes in pretrain models (top 50 results by p-value)

p_value	term_id	description
6.77E-32	GO:0003735	structural constituent of ribosome
6.68E-30	GO:0005840	ribosome
7.36E-30	KEGG:03010	Ribosome
1.87E-29	GO:0022626	cytosolic ribosome
5.45E-26	GO:0044445	cytosolic part
1.87E-25	GO:0044391	ribosomal subunit
1.42E-24	GO:0006412	translation
4.01E-24	GO:0043043	peptide biosynthetic process
1.49E-23	GO:0030529	intracellular ribonucleoprotein complex
1.55E-23	GO:1990904	ribonucleoprotein complex
7.69E-23	GO:0006518	peptide metabolic process
1.93E-22	GO:0043604	amide biosynthetic process
3.34E-21	GO:0003723	RNA binding
1.87E-20	GO:0043603	cellular amide metabolic process
2.65E-20	GO:0005829	cytosol
3.33E-19	GO:0005198	structural molecule activity
1.35E-18	GO:0044444	cytoplasmic part
3.69E-18	GO:1903561	extracellular vesicle
4.07E-18	GO:0043230	extracellular organelle
5.80E-18	GO:0022625	cytosolic large ribosomal subunit
2.41E-17	GO:0070062	extracellular exosome
2.76E-17	GO:1901566	organonitrogen compound biosynthetic process
9.21E-17	GO:0044822	poly(A) RNA binding
5.12E-16	GO:0043232	intracellular non-membrane-bounded organelle
5.12E-16	GO:0043228	non-membrane-bounded organelle
1.44E-15	GO:1901564	organonitrogen compound metabolic process
5.05E-15	GO:0032991	macromolecular complex
6.74E-15	GO:0015934	large ribosomal subunit
8.28E-15	GO:0005737	cytoplasm
2.44E-13	GO:0031982	vesicle
4.82E-13	GO:0043226	organelle
2.04E-12	GO:0044421	extracellular region part
5.77E-12	GO:0005622	intracellular
2.29E-11	GO:0044424	intracellular part
2.49E-11	GO:0043229	intracellular organelle
7.85E-11	GO:0005925	focal adhesion
8.98E-11	GO:0005924	cell-substrate adherens junction
1.12E-10	GO:0030055	cell-substrate junction
1.34E-10	GO:0005912	adherens junction
2.06E-10	GO:0044422	organelle part
2.16E-10	GO:0070161	anchoring junction
2.52E-10	GO:0044446	intracellular organelle part
5.96E-10	GO:0005576	extracellular region
1.17E-09	GO:0022627	cytosolic small ribosomal subunit
9.13E-09	GO:0003676	nucleic acid binding
1.92E-08	HP:0012133	Erythroid hypoplasia
3.62E-08	GO:0043227	membrane-bounded organelle
9.01E-08	GO:0015935	small ribosomal subunit
1.32E-07	GO:1901363	heterocyclic compound binding
1.98E-07	GO:0042254	ribosome biogenesis

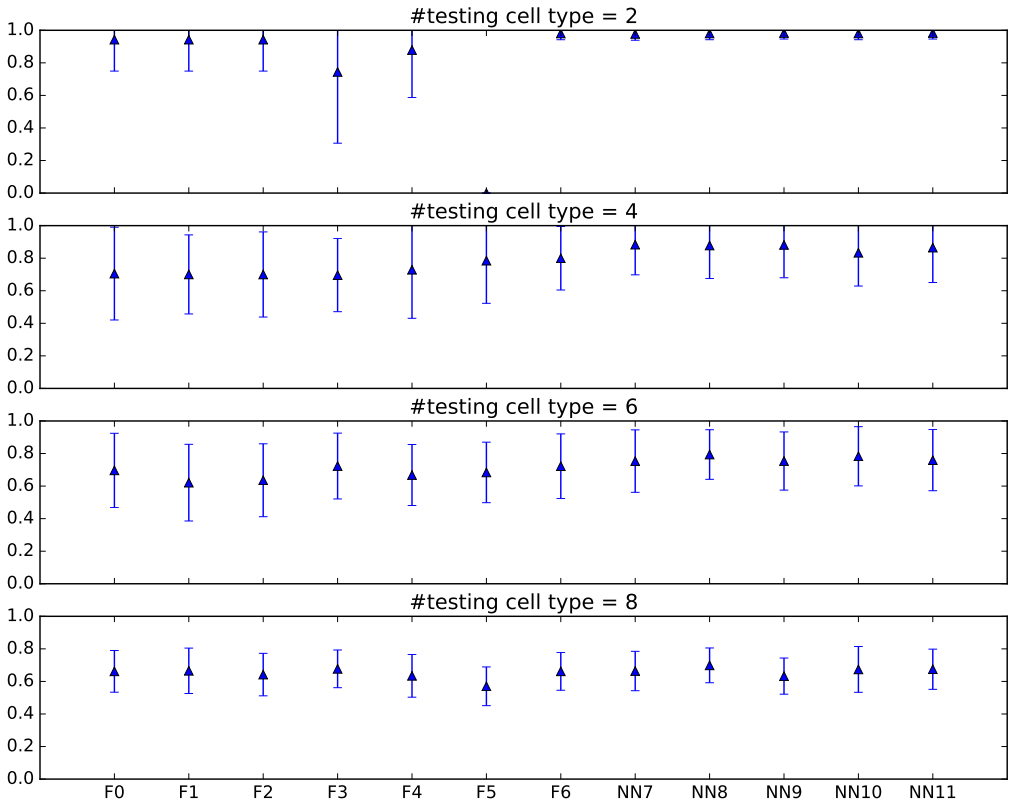
3 Supplementary Figures



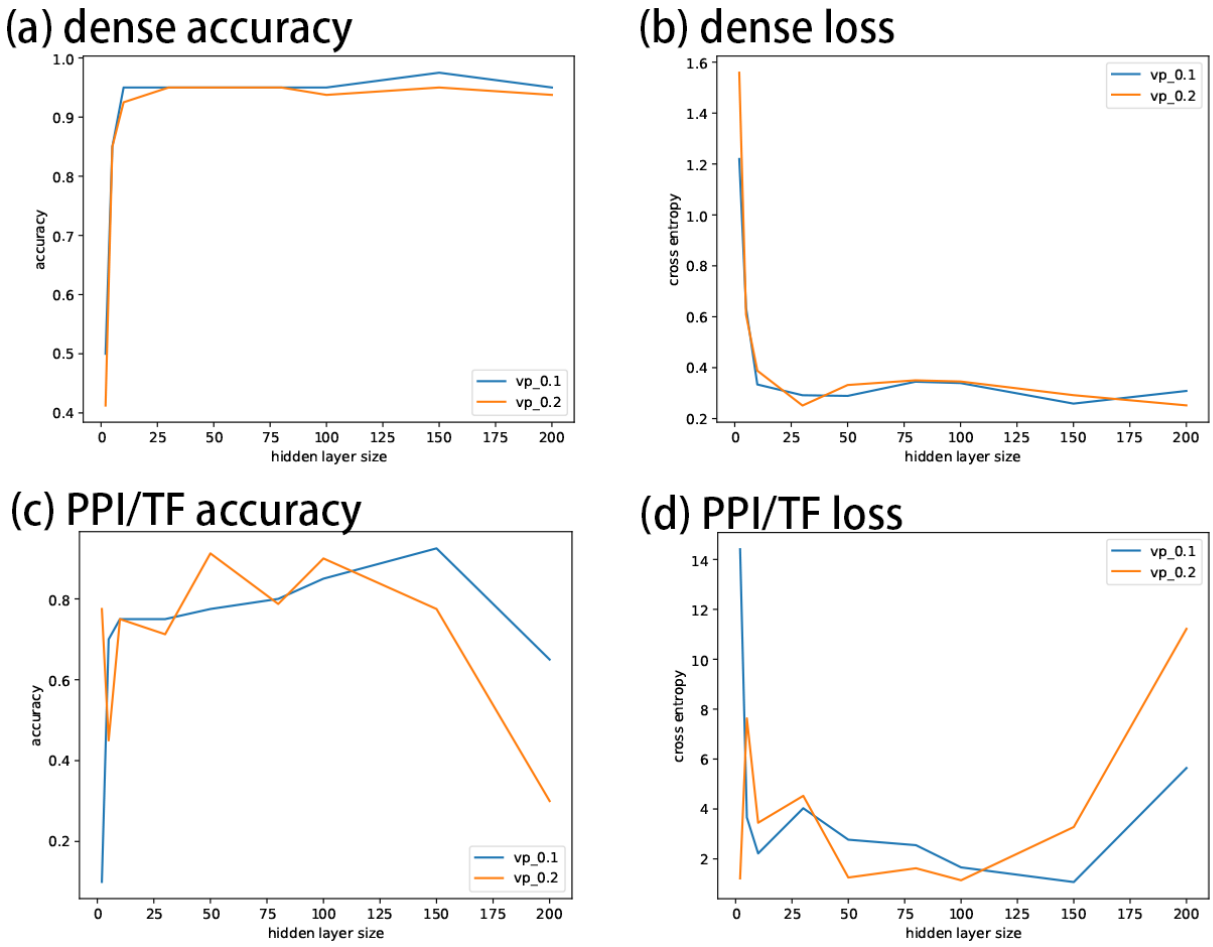
Supplementary Figure 1: The mean and standard error of ARI for some of the clustering results presented in Table 2 of the main paper and in Supplementary Table 3-6). See Supporting Table 13 below for methods represented by F0-F6 and NN7-NN16.

Supplementary Table 13: Abbreviations used for Supporting Figures 1 and 2

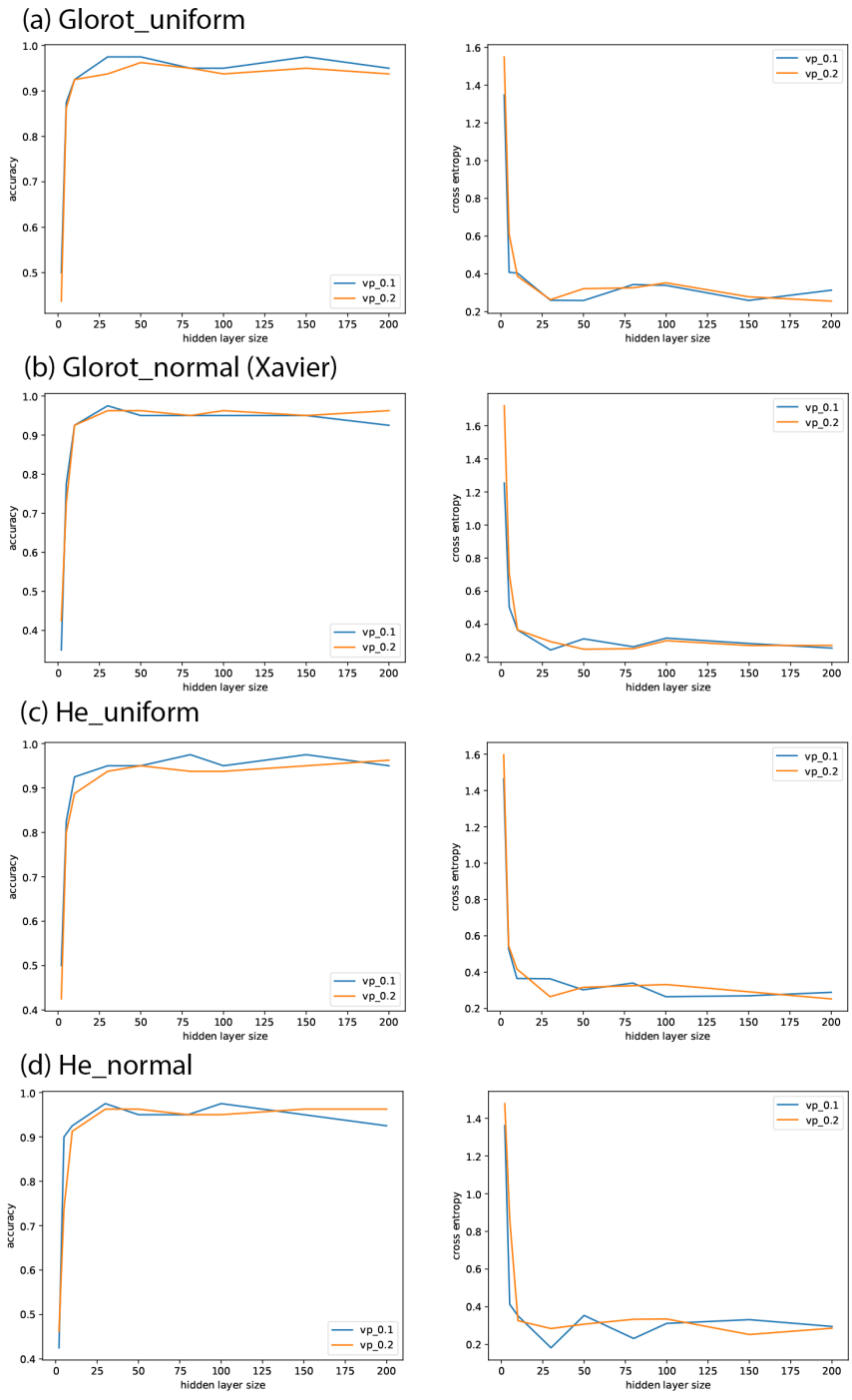
F0	original
F1	PCA 100
F2	PCA 796
F3	pcaReduce
F4	SIMLR
F5	SNN-Cliq
F6	SINCERA hierarchical clustering
NN7	Dense 1layer 100
NN8	Dense 1layer 796
NN9	Dense 2layer 796/100
NN10	PPITF 1layer 696+100
NN11	PPITF 2layer 696+100/100
NN12	Dense 1layer 100 pretrain
NN13	Dense 1layer 796 pretrain
NN14	Dense 2layer 796/100 pretrain
NN15	PPITF 1layer 696+100 pretrain
NN16	PPITF 2layer 696+100/100 pretrain



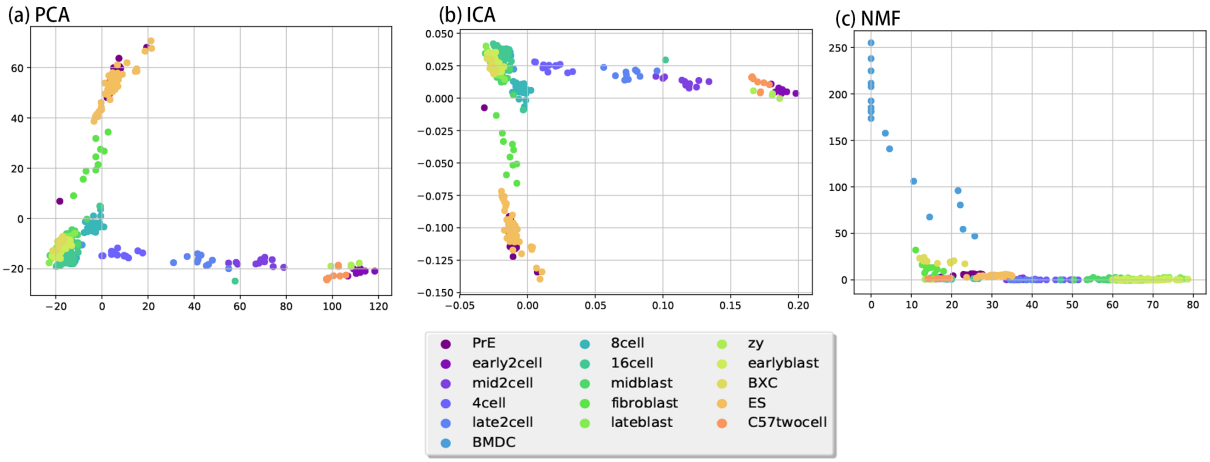
Supplementary Figure 2: The same experiment of Supplementary Figure 1 with data normalized. The results are similar.



Supplementary Figure 3: Average test performance for 10 random train / test splits of NN for various architectures and number of internal nodes. We tested the following number for the dense layer: 2, 5, 10, 30, 50, 80, 100, 150 and 200. VP is the percentage of cells left for testing. (a) Accuracy of a single dense layer (b) Cross entropy loss for that architecture (c) Accuracy for changing the number of nodes in the dense layer of the PPI/TF model (d) Cross entropy for that model



Supplementary Figure 4: Average testing performance for 10 random train / test splits of NN for various initialization methods. VP is the percentage of cells left for testing. (a) Accuracy and loss of glorot uniform initialization (which is the method used in the paper) (b) Accuracy and loss of glorot normal (Xavier) initialization (c) Accuracy and loss of uniform He initialization (d) Accuracy and loss of normal He initialization



Supplementary Figure 5: The 2D visualization of reduced dimensions. Other figures are in supplementary. (a) 2D TSNE visualization of original data (b) 2D PCA visualization of SIMLR-transformed data (c) 2D TSNE visualization of the transformed data by 2 layer PPI/TF NN (d) 2D PCA visualization of the transformed data by 2 layer PPI/TF NN