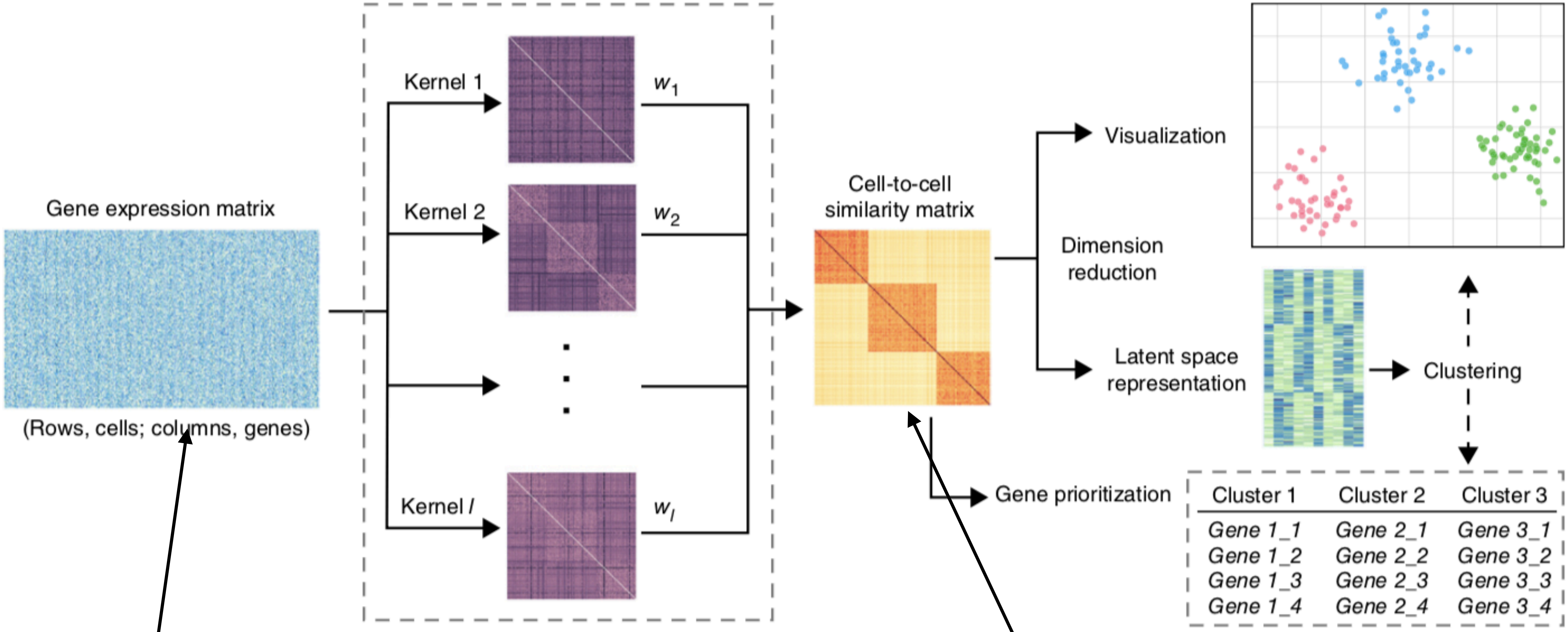
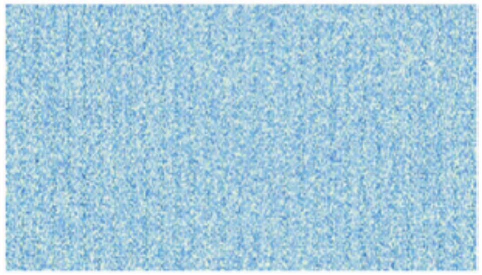
SIMLR

by 谢欣承



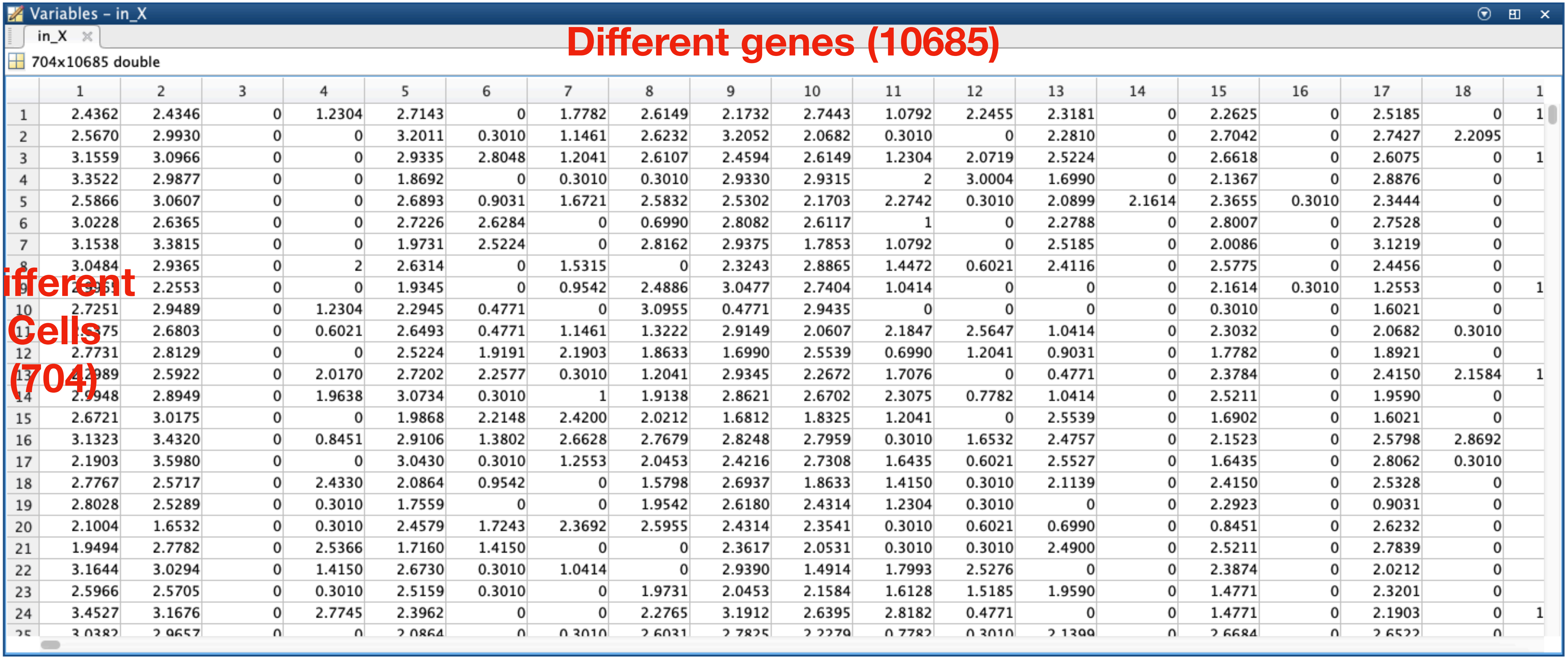
Given N×M matrix, SIMLR solves for N×N matrix

Gene expression matrix



(Rows, cells; columns, genes)

measured by gene expression levels and so on



Variables – in_X

in_X

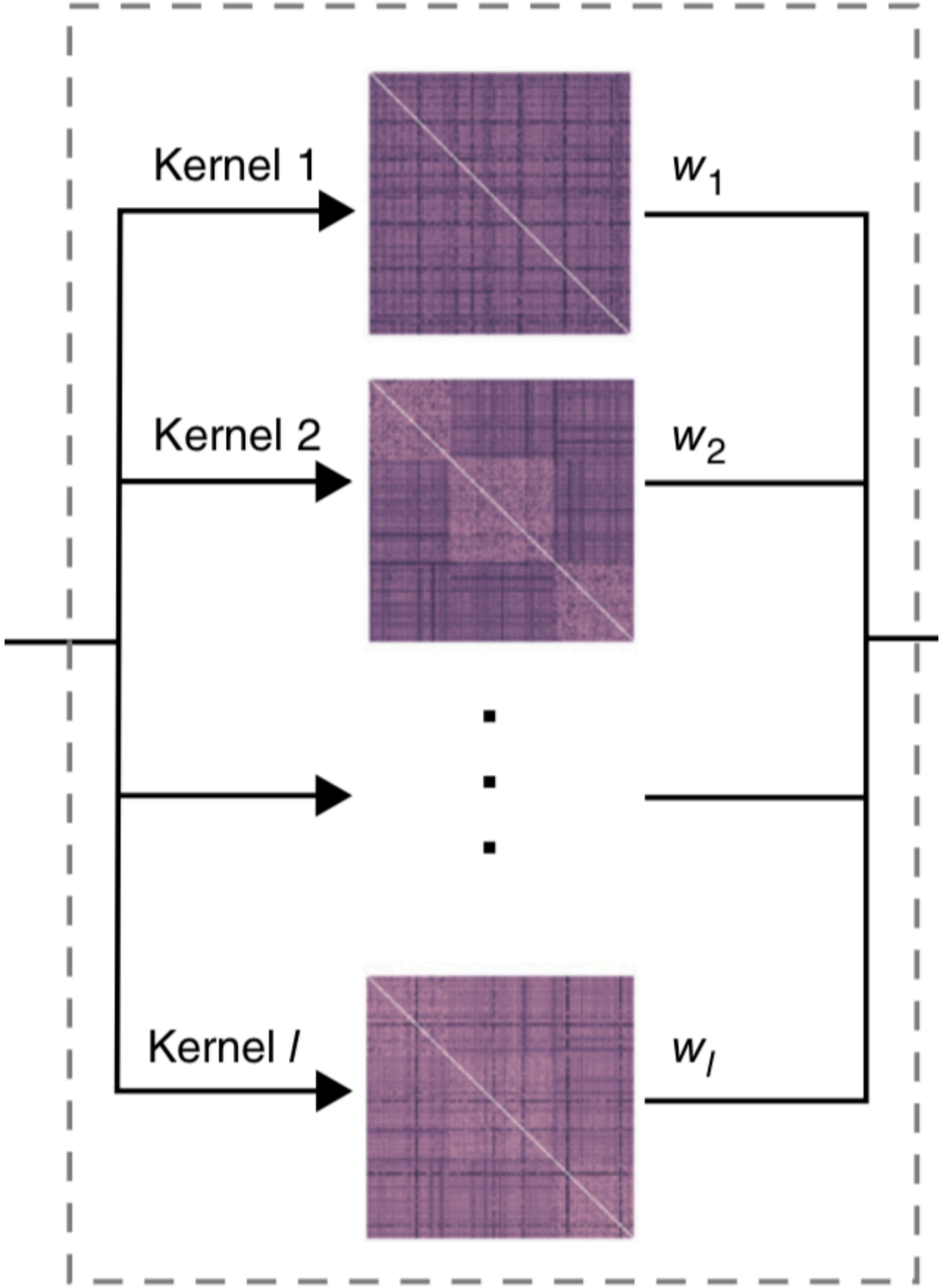
704x10685 double

Different genes (10685)

Different Cells (704)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	1
1	2.4362	2.4346	0	1.2304	2.7143	0	1.7782	2.6149	2.1732	2.7443	1.0792	2.2455	2.3181	0	2.2625	0	2.5185	0	1
2	2.5670	2.9930	0	0	3.2011	0.3010	1.1461	2.6232	3.2052	2.0682	0.3010	0	2.2810	0	2.7042	0	2.7427	2.2095	
3	3.1559	3.0966	0	0	2.9335	2.8048	1.2041	2.6107	2.4594	2.6149	1.2304	2.0719	2.5224	0	2.6618	0	2.6075	0	1
4	3.3522	2.9877	0	0	1.8692	0	0.3010	0.3010	2.9330	2.9315	2	3.0004	1.6990	0	2.1367	0	2.8876	0	
5	2.5866	3.0607	0	0	2.6893	0.9031	1.6721	2.5832	2.5302	2.1703	2.2742	0.3010	2.0899	2.1614	2.3655	0.3010	2.3444	0	
6	3.0228	2.6365	0	0	2.7226	2.6284	0	0.6990	2.8082	2.6117	1	0	2.2788	0	2.8007	0	2.7528	0	
7	3.1538	3.3815	0	0	1.9731	2.5224	0	2.8162	2.9375	1.7853	1.0792	0	2.5185	0	2.0086	0	3.1219	0	
8	3.0484	2.9365	0	2	2.6314	0	1.5315	0	2.3243	2.8865	1.4472	0.6021	2.4116	0	2.5775	0	2.4456	0	
9	[illegible]	2.2553	0	0	1.9345	0	0.9542	2.4886	3.0477	2.7404	1.0414	0	0	0	2.1614	0.3010	1.2553	0	1
10	2.7251	2.9489	0	1.2304	2.2945	0.4771	0	3.0955	0.4771	2.9435	0	0	0	0	0.3010	0	1.6021	0	
11	[illegible]	2.6803	0	0.6021	2.6493	0.4771	1.1461	1.3222	2.9149	2.0607	2.1847	2.5647	1.0414	0	2.3032	0	2.0682	0.3010	
12	2.7731	2.8129	0	0	2.5224	1.9191	2.1903	1.8633	1.6990	2.5539	0.6990	1.2041	0.9031	0	1.7782	0	1.8921	0	
13	2.2989	2.5922	0	2.0170	2.7202	2.2577	0.3010	1.2041	2.9345	2.2672	1.7076	0	0.4771	0	2.3784	0	2.4150	2.1584	1
14	2.9948	2.8949	0	1.9638	3.0734	0.3010	1	1.9138	2.8621	2.6702	2.3075	0.7782	1.0414	0	2.5211	0	1.9590	0	
15	2.6721	3.0175	0	0	1.9868	2.2148	2.4200	2.0212	1.6812	1.8325	1.2041	0	2.5539	0	1.6902	0	1.6021	0	
16	3.1323	3.4320	0	0.8451	2.9106	1.3802	2.6628	2.7679	2.8248	2.7959	0.3010	1.6532	2.4757	0	2.1523	0	2.5798	2.8692	
17	2.1903	3.5980	0	0	3.0430	0.3010	1.2553	2.0453	2.4216	2.7308	1.6435	0.6021	2.5527	0	1.6435	0	2.8062	0.3010	
18	2.7767	2.5717	0	2.4330	2.0864	0.9542	0	1.5798	2.6937	1.8633	1.4150	0.3010	2.1139	0	2.4150	0	2.5328	0	
19	2.8028	2.5289	0	0.3010	1.7559	0	0	1.9542	2.6180	2.4314	1.2304	0.3010	0	0	2.2923	0	0.9031	0	
20	2.1004	1.6532	0	0.3010	2.4579	1.7243	2.3692	2.5955	2.4314	2.3541	0.3010	0.6021	0.6990	0	0.8451	0	2.6232	0	
21	1.9494	2.7782	0	2.5366	1.7160	1.4150	0	0	2.3617	2.0531	0.3010	0.3010	2.4900	0	2.5211	0	2.7839	0	
22	3.1644	3.0294	0	1.4150	2.6730	0.3010	1.0414	0	2.9390	1.4914	1.7993	2.5276	0	0	2.3874	0	2.0212	0	
23	2.5966	2.5705	0	0.3010	2.5159	0.3010	0	1.9731	2.0453	2.1584	1.6128	1.5185	1.9590	0	1.4771	0	2.3201	0	
24	3.4527	3.1676	0	2.7745	2.3962	0	0	2.2765	3.1912	2.6395	2.8182	0.4771	0	0	1.4771	0	2.1903	0	1
25	3.0382	2.9657	0	0	2.0864	0	0.3010	2.6031	2.7825	2.2279	0.7782	0.3010	2.1399	0	2.6684	0	2.6522	0	

An example from Kolodziejczyk dataset



Kernels are calculated using Gaussian kernels

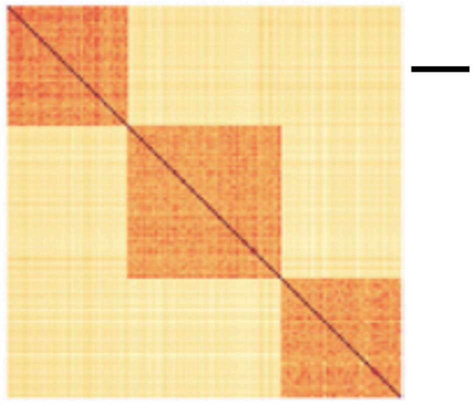
$$K\left(c_i, c_j\right) = \frac{1}{\epsilon_{ij}\sqrt{2\pi}} \exp\left(-\frac{\| c_i - c_j \|_2^2}{2\epsilon_{ij}^2}\right)$$

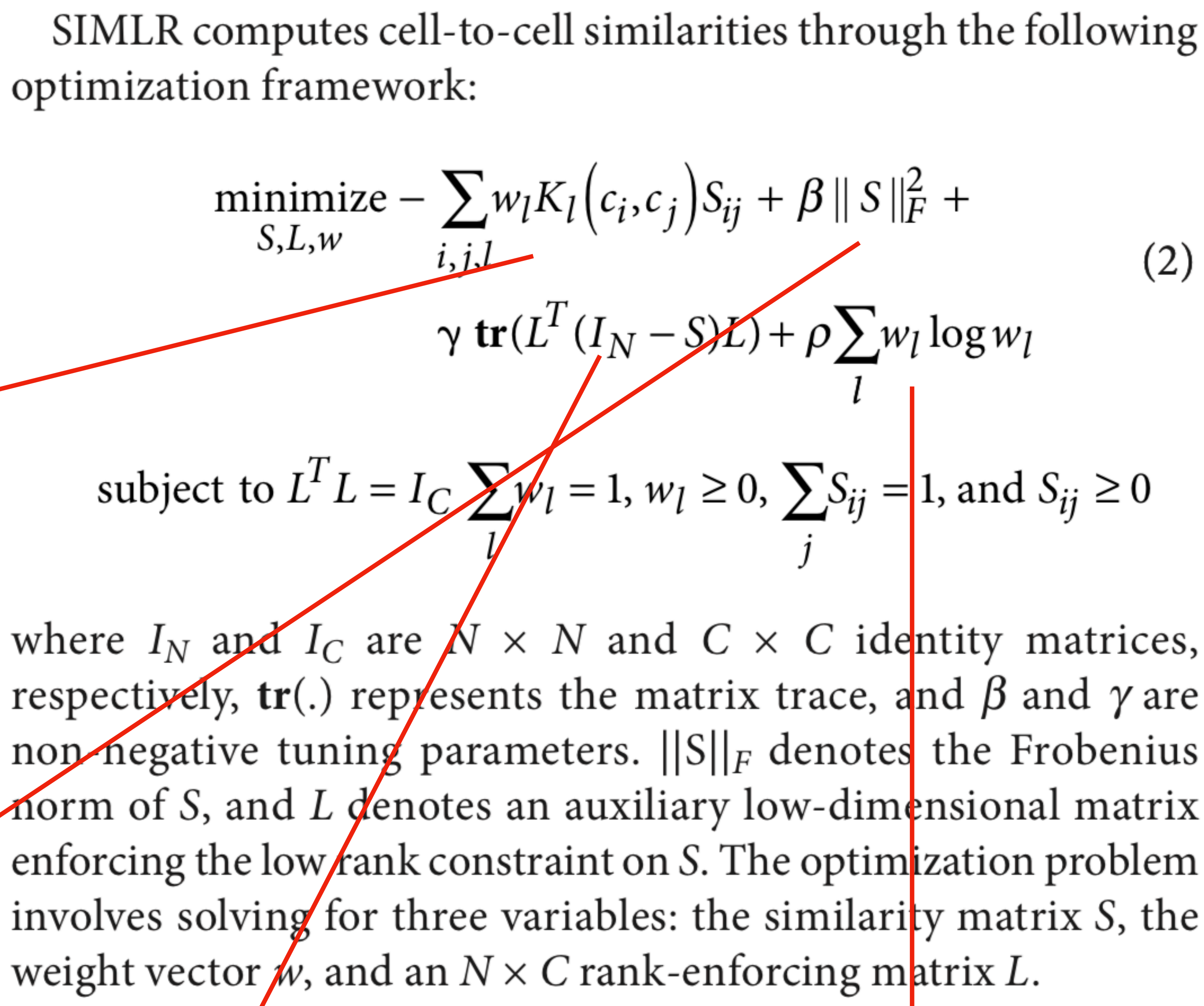
where $||c_i - c_j||_2$ is the Euclidean distance between cell i and cell j. The variance, $\in_{ij}$, can be calculated with different scales:

$$\mu_i = \frac{\sum_{l \in \mathrm{KNN}(c_i) \| c_i - c_j \|_2}}{k}, \; \varepsilon_{ij} = \frac{\sigma\left(\mu_i + \mu_j\right)}{2}$$

where $\mathrm{KNN}(c_i)$ represents cells that are top k neighbors of the cell i.

Cell-to-cell similarity matrix





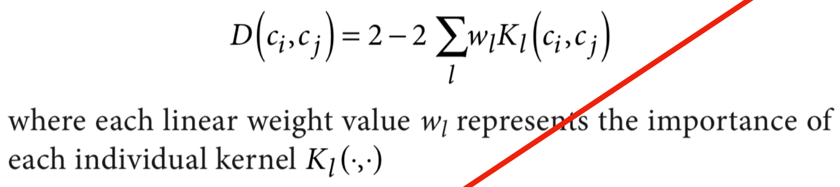
SIMLR computes cell-to-cell similarities through the following optimization framework:

$$\underset{S,L,w}{\text{minimize}} -\sum_{i,j,l} w_l K_l\left(c_i,c_j\right) S_{ij} + \beta \| S \|_F^2 + \gamma\, \mathbf{tr}\left(L^T\left(I_N - S\right)L\right) + \rho \sum_l w_l \log w_l \qquad (2)$$

$$\text{subject to } L^T L = I_C, \sum_l w_l = 1, w_l \geq 0, \sum_j S_{ij} = 1, \text{ and } S_{ij} \geq 0$$

where I_N and I_C are $N \times N$ and $C \times C$ identity matrices, respectively, $\mathbf{tr}(.)$ represents the matrix trace, and β and γ are non-negative tuning parameters. $||S||_F$ denotes the Frobenius norm of S, and L denotes an auxiliary low-dimensional matrix enforcing the low rank constraint on S. The optimization problem involves solving for three variables: the similarity matrix S, the weight vector w, and an $N \times C$ rank-enforcing matrix L.

Learned similarity S between two cells should be small if the distance between them is large



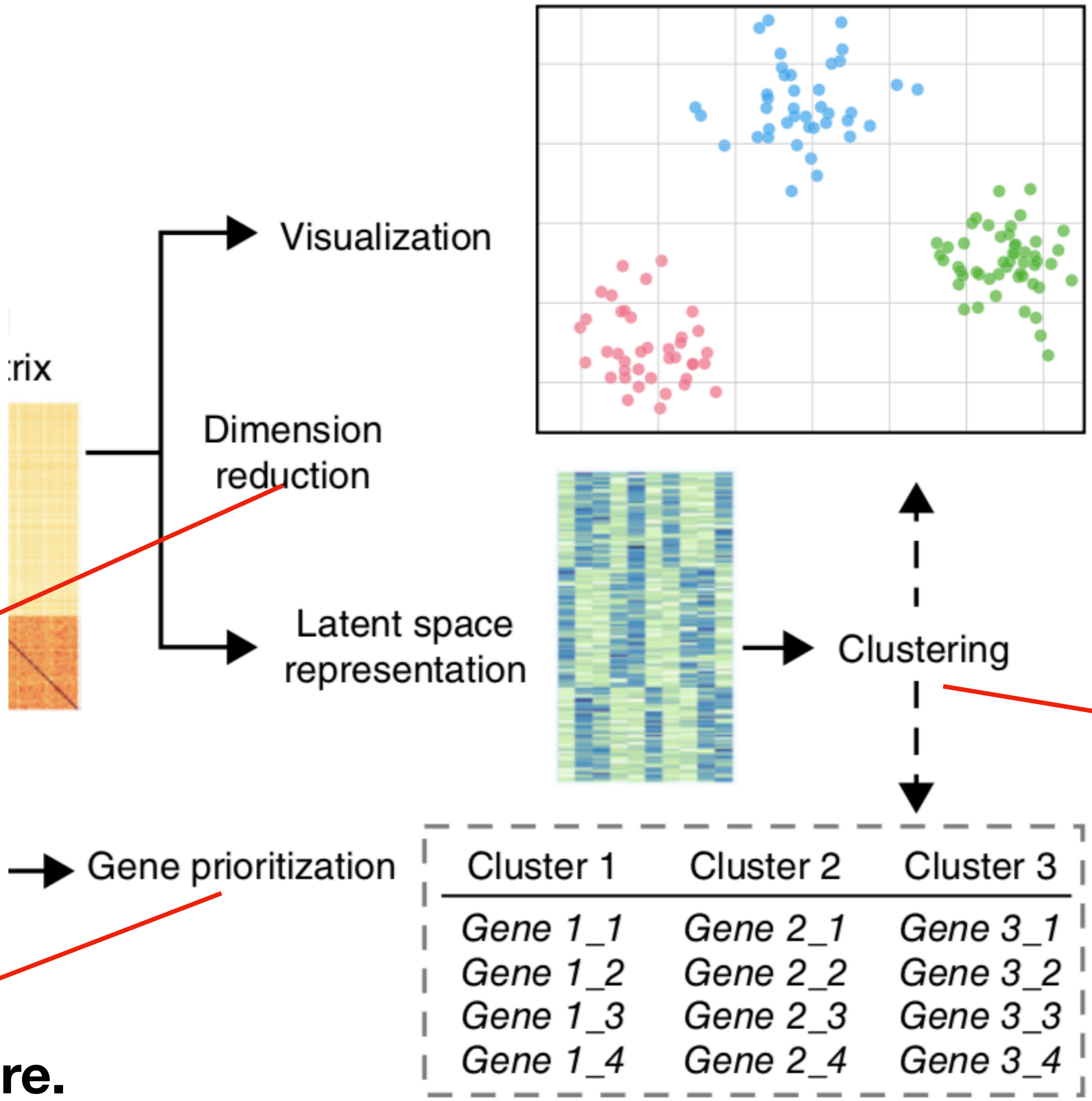
$$D\left(c_i,c_j\right) = 2 - 2\sum_l w_l K_l\left(c_i,c_j\right)$$

where each linear weight value w_l represents the importance of each individual kernel $K_l(\cdot,\cdot)$

Regularization term preventing S from becoming identity matrix

Enforces S to be rank C matrix

Constraints avoiding selection of single kernel



Cluster 1	Cluster 2	Cluster 3
Gene 1_1	*Gene 2_1*	*Gene 3_1*
Gene 1_2	*Gene 2_2*	*Gene 3_2*
Gene 1_3	*Gene 2_3*	*Gene 3_3*
Gene 1_4	*Gene 2_4*	*Gene 3_4*

Using t-SNE to project similarity matrix into lower dimension

k-means and so on

Using Laplacian score. The higher the score, the more important the gene is to globally differentiate the subpopulations of cells

The End