VNiVERSiDAD D SALAMANCA

Trabajo Fin de Grado

ANÁLISIS ESTADÍSTICO DE DATOS TRANSCRIPTÓMICOS COMPLEJOS DE MUESTRAS HUMANAS Y USO DE UN MÉTODO DE DECONVOLUCIÓN PARA IDENTIFICAR TIPOS CELULARES ESPECÍFICOS

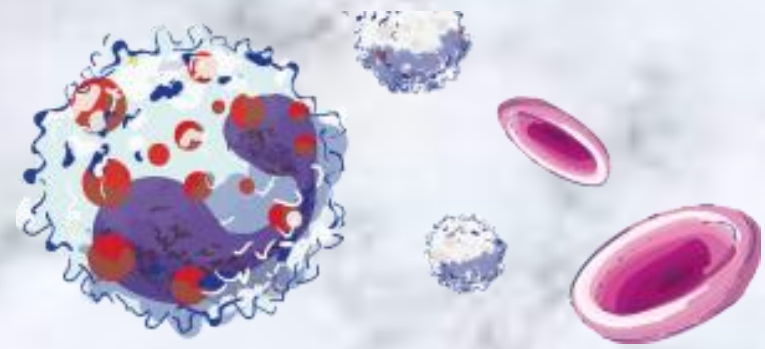
STATISTICAL ANALYSIS OF LARGE-SCALE COMPLEX TRANSCRIPTOMIC DATA OF HUMAN SAMPLES AND USE OF A DECONVOLUTION METHOD TO IDENTIFY SPECIFIC CELL TYPES

Laura Gutiérrez García

Tutores

José Manuel Sánchez Santos

Javier de las Rivas Sanz



GRADO DE ESTADÍSTICA

FACULTAD DE CIENCIAS

Trabajo Fin de Grado

ANÁLISIS ESTADÍSTICO DE DATOS TRANSCRIPTÓMICOS COMPLEJOS DE MUESTRAS HUMANAS Y USO DE UN MÉTODO DE DECONVOLUCIÓN PARA IDENTIFICAR TIPOS CELULARES ESPECÍFICOS

STATISTICAL ANALYSIS OF LARGE-SCALE COMPLEX TRANSCRIPTOMIC DATA OF HUMAN SAMPLES AND USE OF A DECONVOLUTION METHOD TO IDENTIFY SPECIFIC CELL TYPES

Laura Gutiérrez García

Tutores

Dr. José Manuel Sánchez Santos

Dr. Javier de las Rivas Sanz

Dr. Javier de las Rivas Sanz

Dr. José Manuel Sánchez Santos

Laura Gutiérrez García

Salamanca, 2020

LISTA DE TABLAS

LISTA DE GRÁFICOS

ANEXO: Gráficos y Tablas Suplementarios

En este anexo se presentan los gráficos y tablas suplementarios que complementan al Trabajo de Fin de Grado: *ANÁLISIS ESTADÍSTICO DE DATOS TRANSCRIPTÓMICOS COMPLEJOS DE MUESTRAS HUMANAS Y USO DE UN MÉTODO DE DECONVOLUCIÓN PARA IDENTIFICAR TIPOS CELULARES ESPECÍFICOS.*

Los primeros gráficos (*S1-S4*) hacen referencia a la representación de las proporciones observadas en el conjunto de datos simulado y en el GSE64385 mostrando específicamente en cada muestra los porcentajes relativos a los tipos celulares y visualizando la gran diversidad de unas muestras a otras.

A continuación, se encuentran los gráficos (*S6-S15*) que proporciona el método *linseed* al ejecutar su guion de R. En ellos, se observan algunos de los pasos de su procedimiento al realizar la deconvolución (Nivel de significación, SVD, simplex) junto con otros gráficos de proporciones (por muestras, t-SNE y comparación con las observadas) en ambos *datasets*, salvo el gráfico referente al t-SNE que únicamente se visualiza en los datos simulados a causa del reducido número de muestras en los reales.

Posteriormente, aparecen el gráfico de correlaciones con los resultados de las proporciones estimadas con la mediana en *linseed* (*S16*) y los *heatmaps* y el gráfico de correlación del método *deconICA* obtenidos mediante las puntuaciones estimadas para así poder compararlos con los porcentajes (*S17-S19*).

Por último, se detallan las tablas (*S1-S5*) que recogen las medidas del error y divergencia por muestra, así como los valores máximos de RMSE en cada *dataset* y su representación gráfica en un *boxplot* (*S20-S21*). Además, también se presentan los resultados del test no paramétrico de Wilcoxon para la divergencia según la salida del programa R (*S6*).

GRÁFICOS PROPORCIONES CELULARES POR MUESTRA

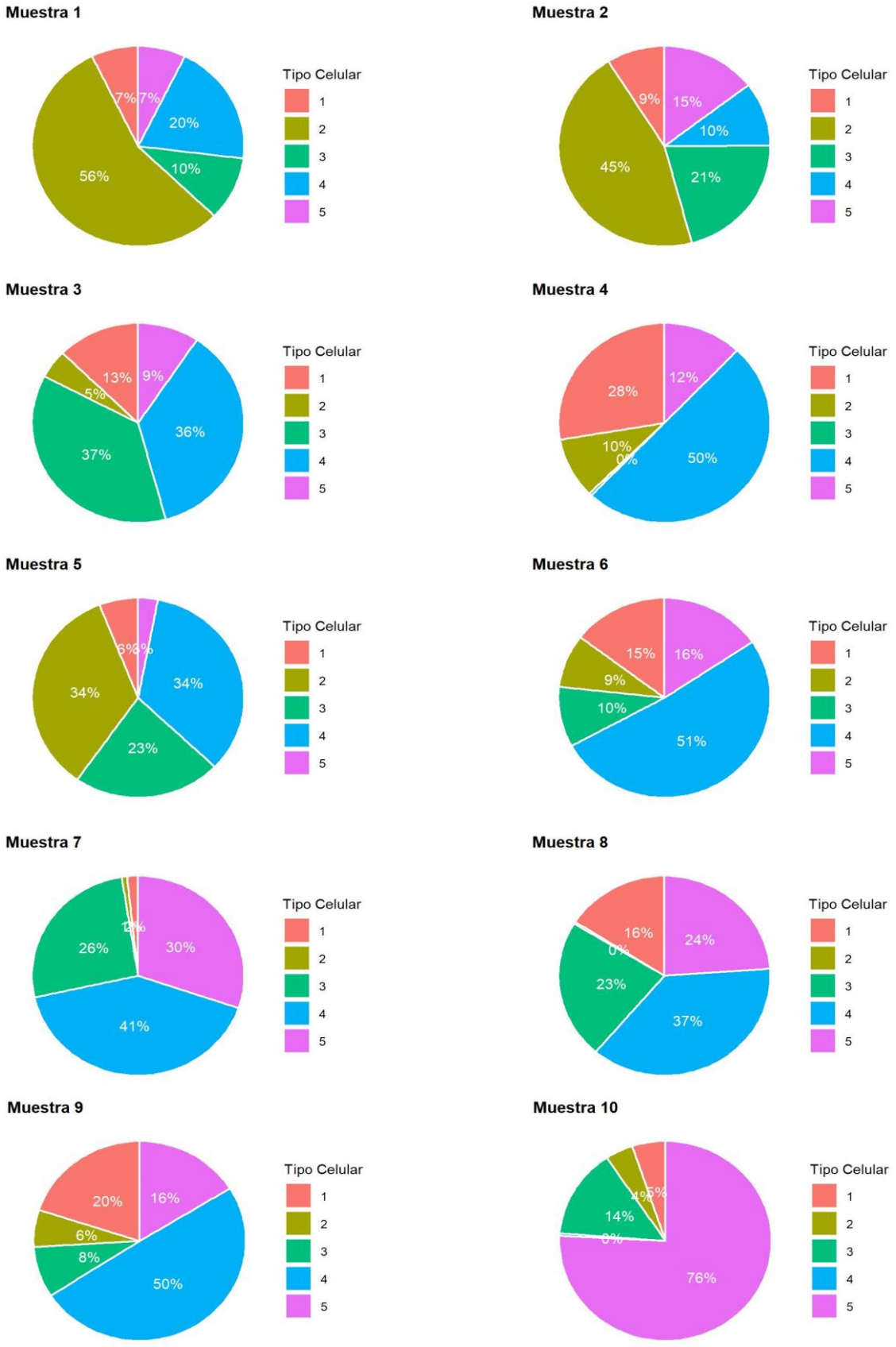


Gráfico S1. Proporciones muestras simuladas (I)

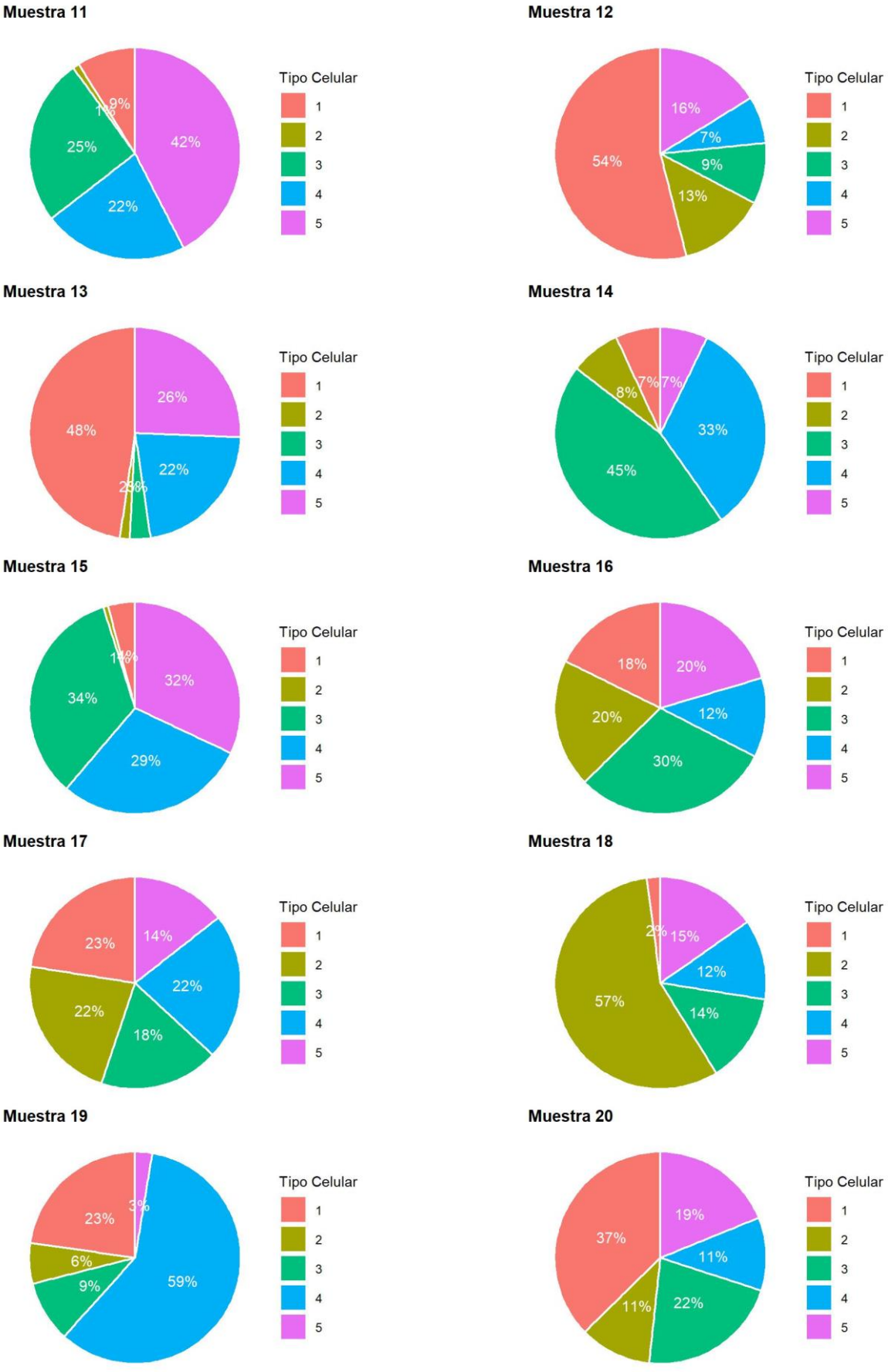


Gráfico S2. Proporciones muestras simuladas (II)

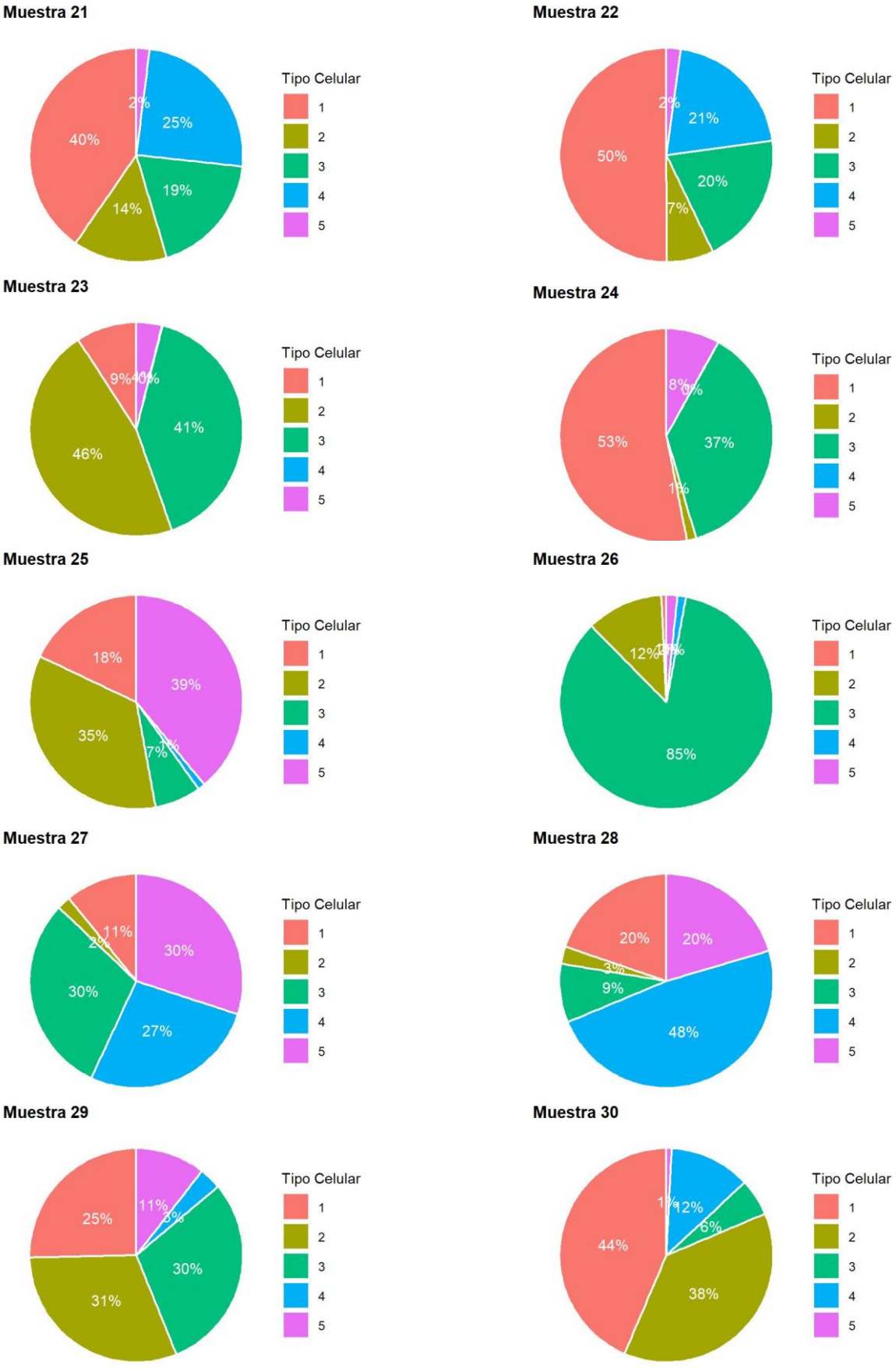
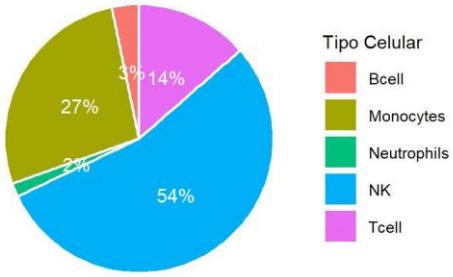
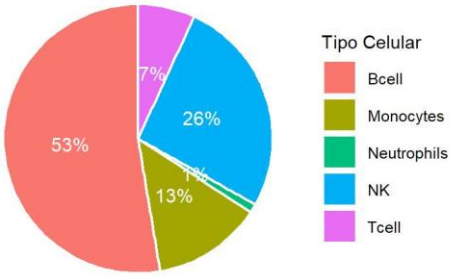


Gráfico S3. Proporciones muestras simuladas (III)

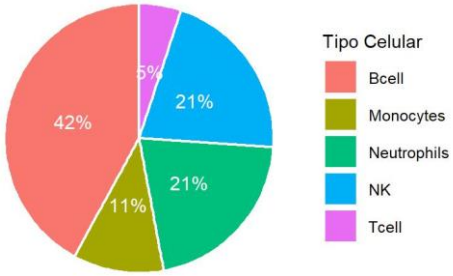
Muestra 3



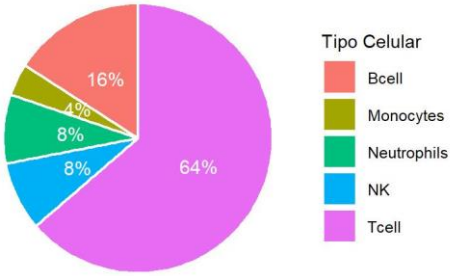
Muestra 4



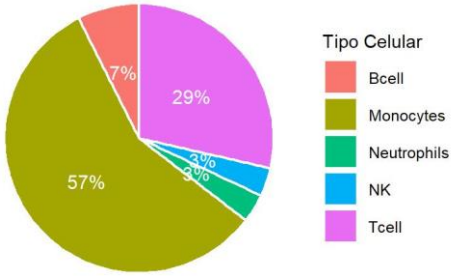
Muestra 5



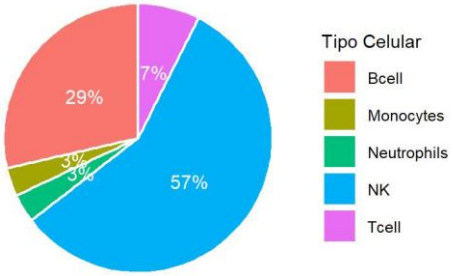
Muestra 6



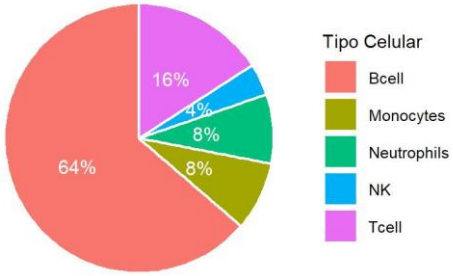
Muestra 7



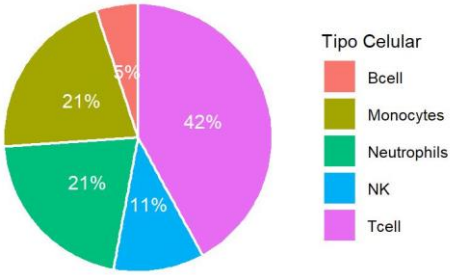
Muestra 8



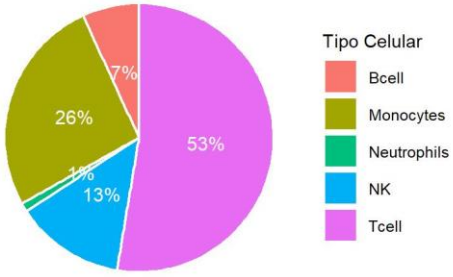
Muestra 9



Muestra 10



Muestra 11



Muestra 12

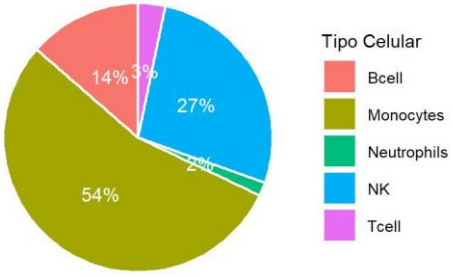


Gráfico S4. Proporciones muestras GSE64385
(excluyendo las dos primeras correspondientes al tipo celular HCT116)

GRÁFICOS PROCEDIMIENTO LINSEED

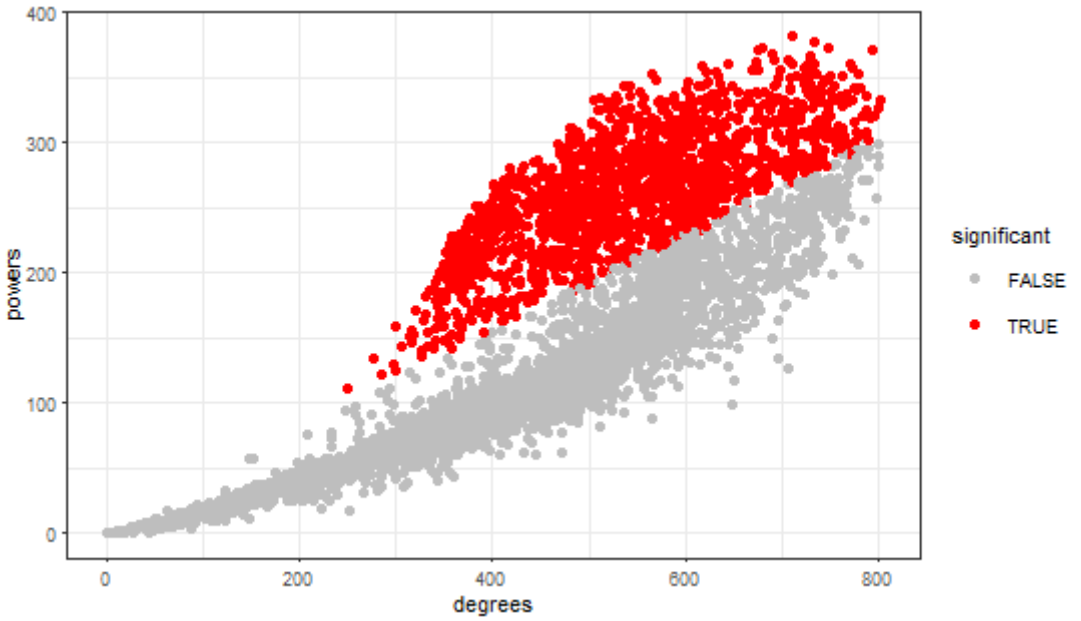


Gráfico S5. Nivel de significación en la correlación de Pearson en *linseed* (Simulación)

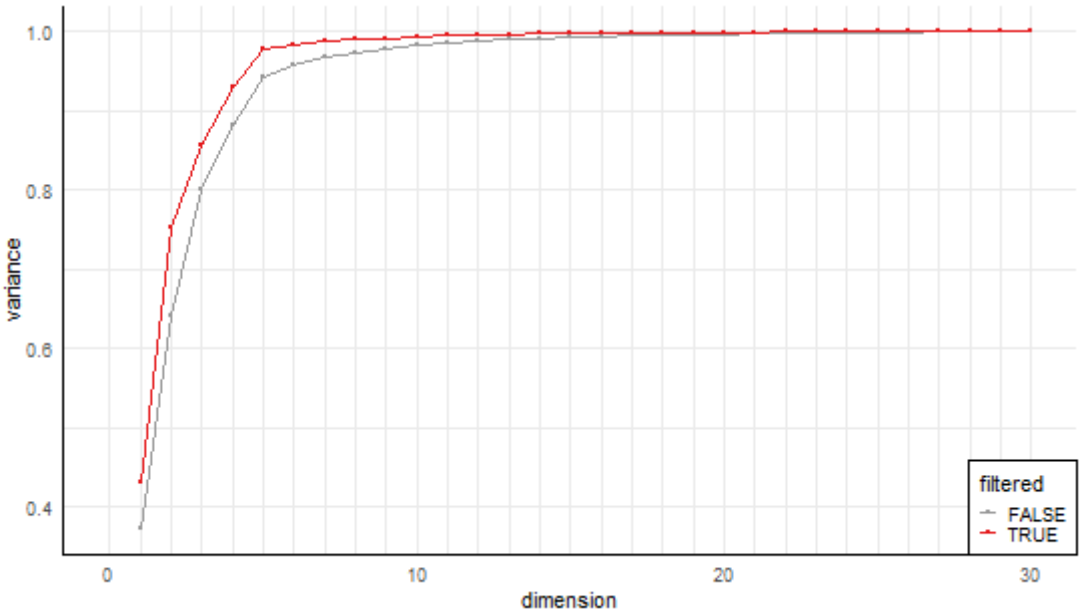


Gráfico S6. Determinación del número de tipos celulares con SVD en *linseed* (Simulación)

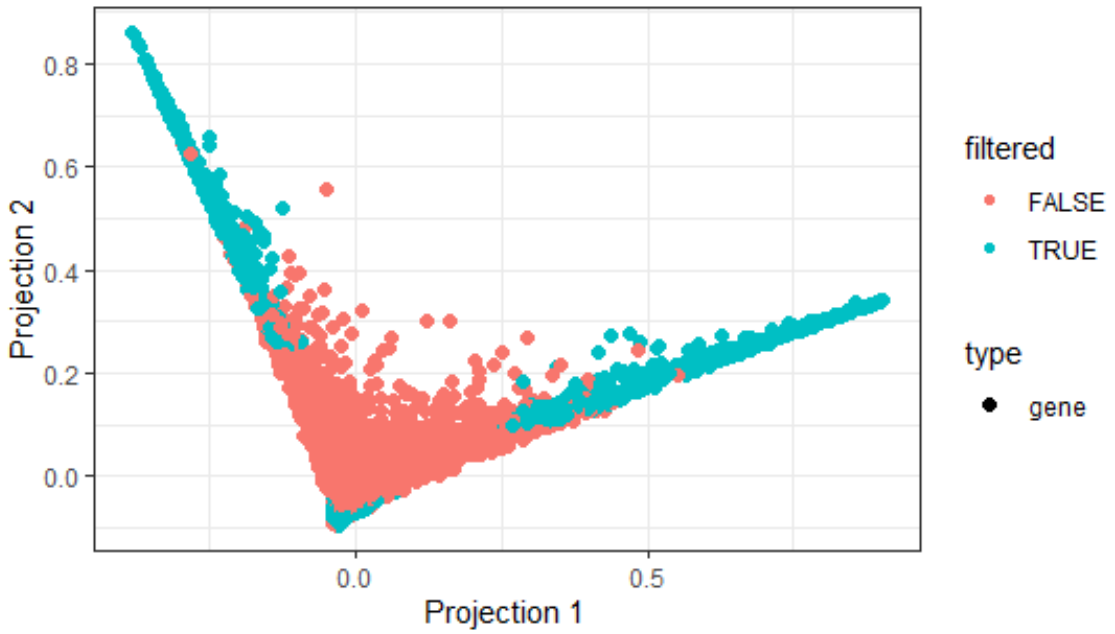


Gráfico S7. Proyección datos simulados en el espacio simplex en *linseed*

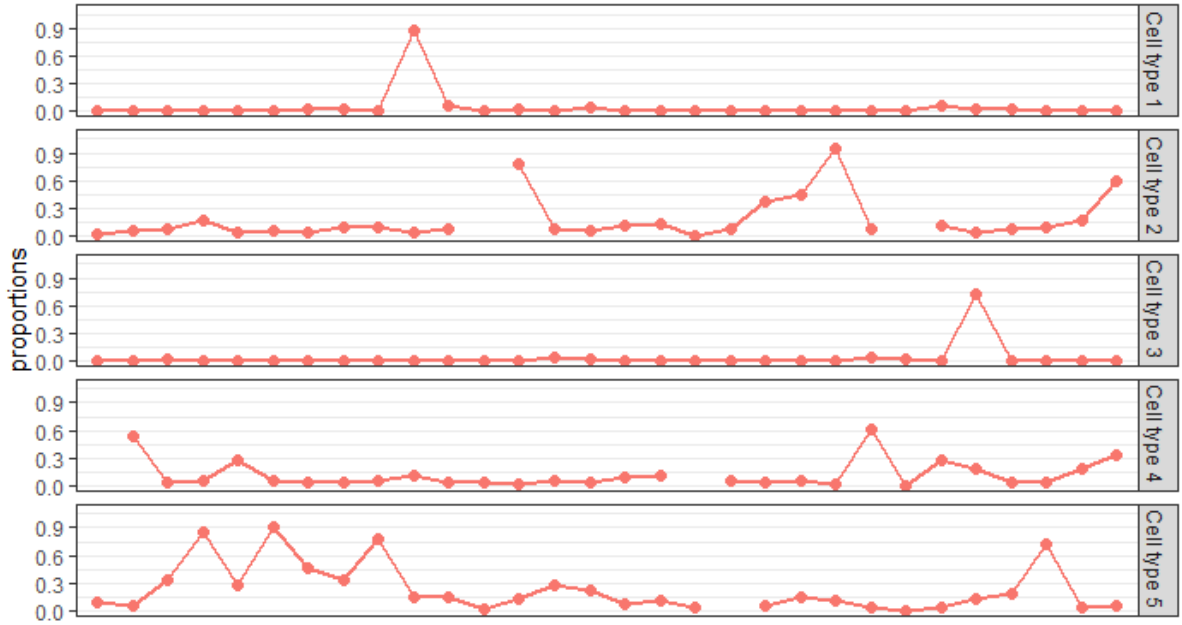


Gráfico S8. Proporciones estimadas para cada tipo celular en *linseed* (Simulación)

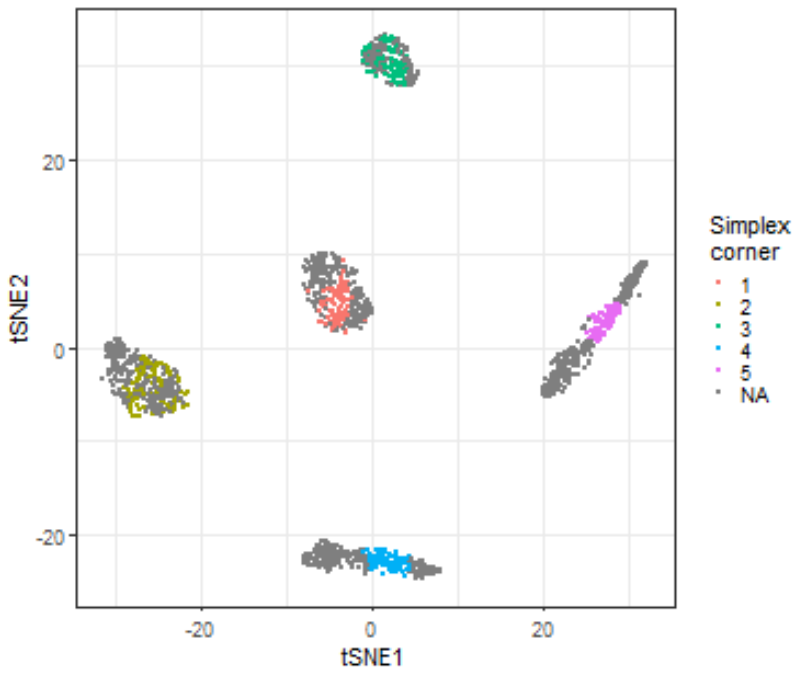


Gráfico S9. Representación t-SNE para los datos simulados en *linseed*

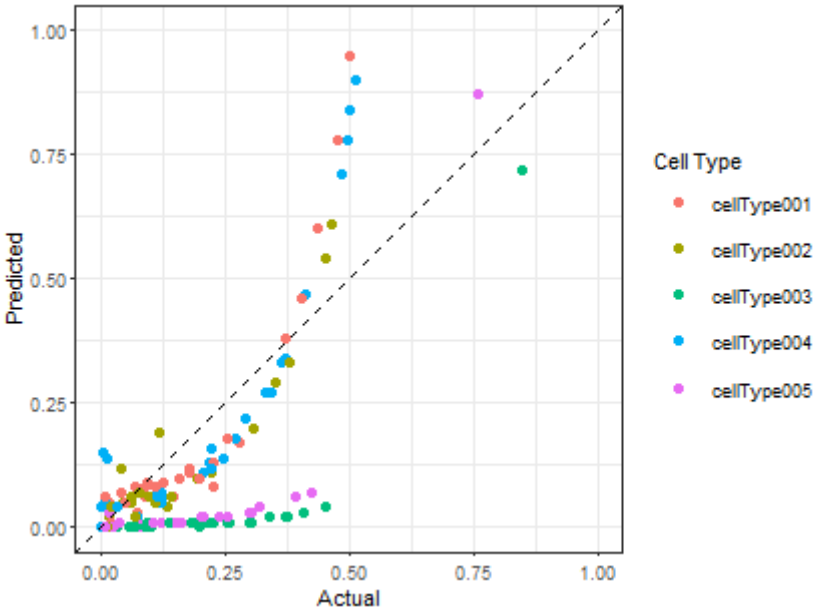


Gráfico S10. Comparación proporciones estimadas y observadas en *linseed* (Simulación)

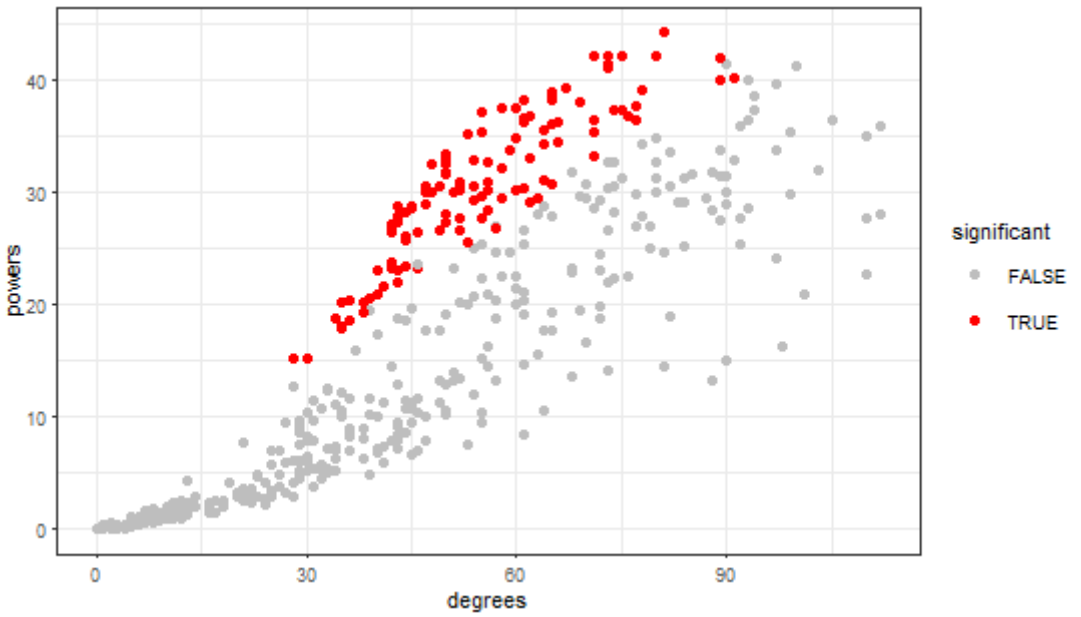


Gráfico S11. Nivel de significación en la correlación de Pearson en *linseed* (GSE64385)

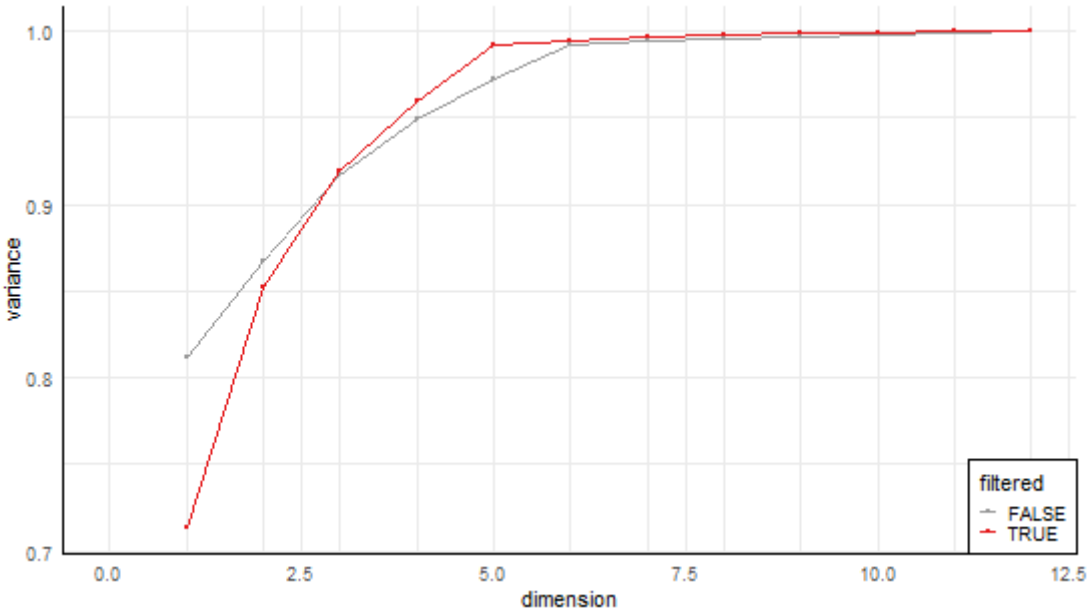


Gráfico S12. Determinación del número de tipos celulares con SVD en *linseed* (GSE64385)

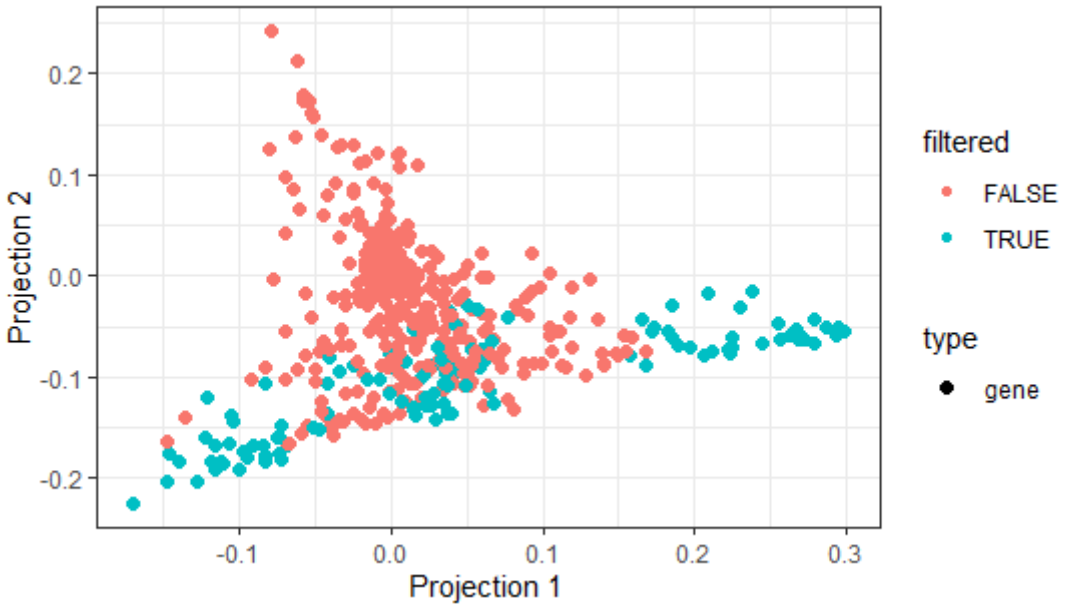


Gráfico S13. Proyección de los datos del GSE64385 en el espacio simplex en *linseed*

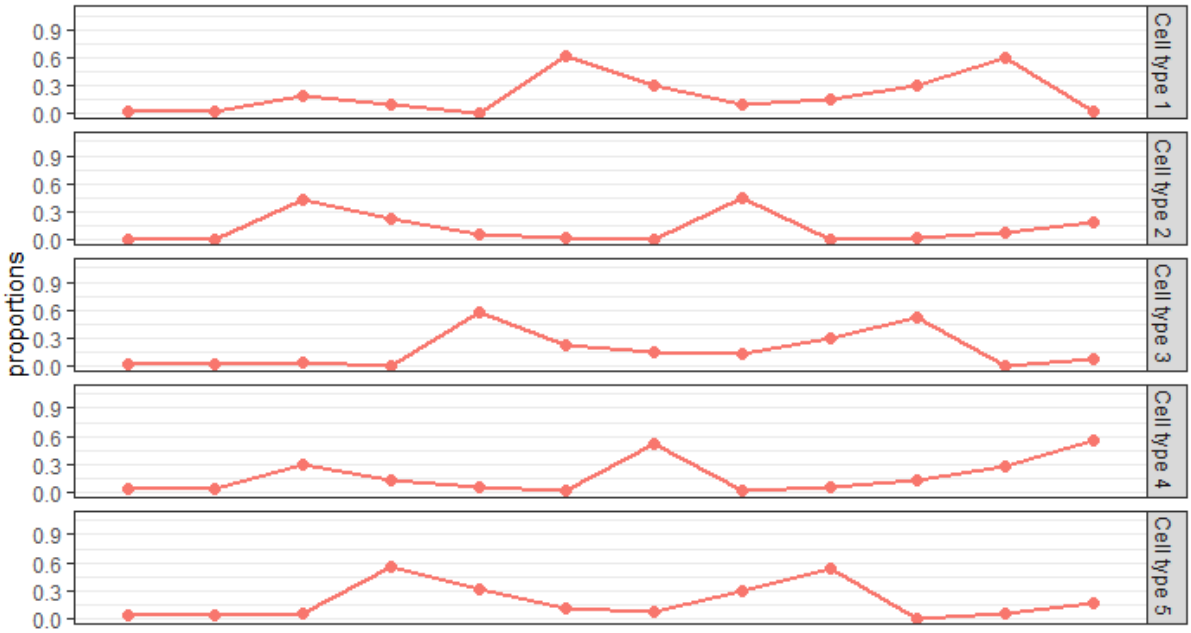


Gráfico S14. Proporciones estimadas para cada tipo celular en *linseed* (GSE64385)

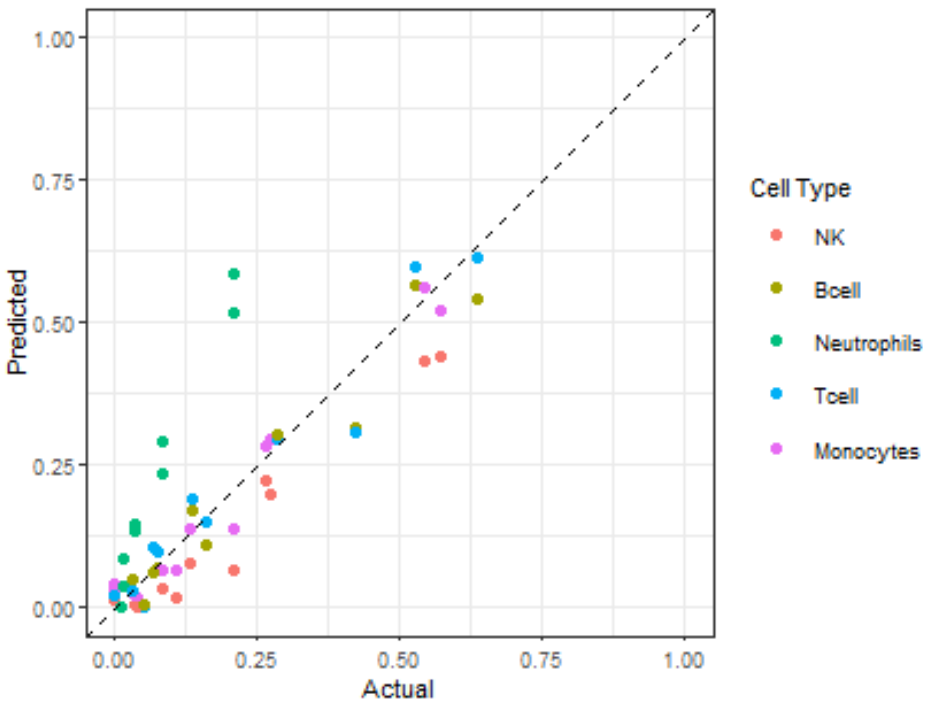


Gráfico S15. Comparación proporciones estimadas y observadas en *linseed* (GSE64385)

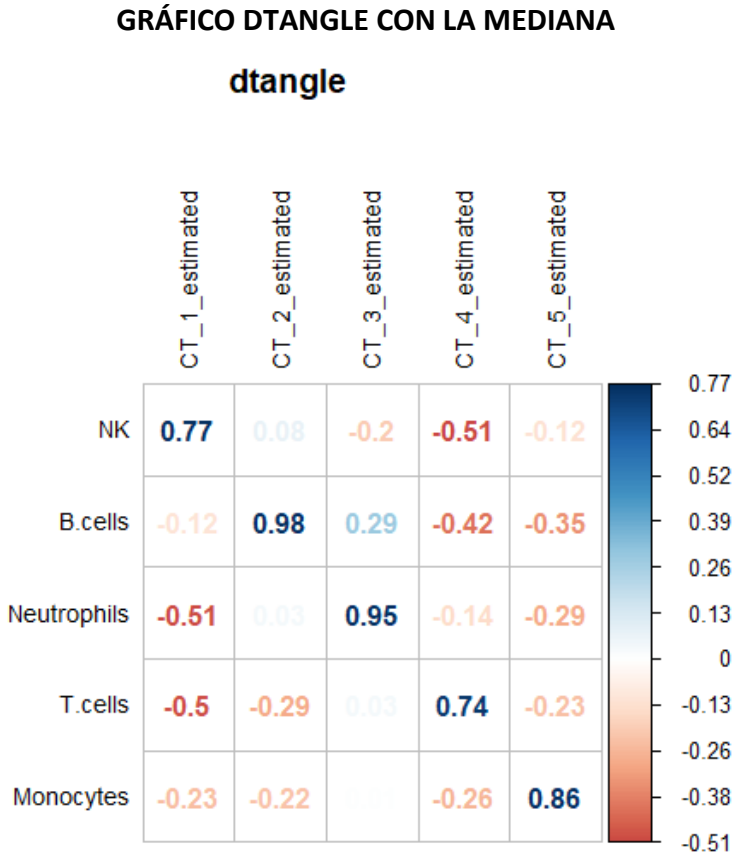


Gráfico S16. Gráfico correlaciones con la mediana *dtangle* (GSE64385)

GRÁFICOS DECONICA PUNTUACIONES

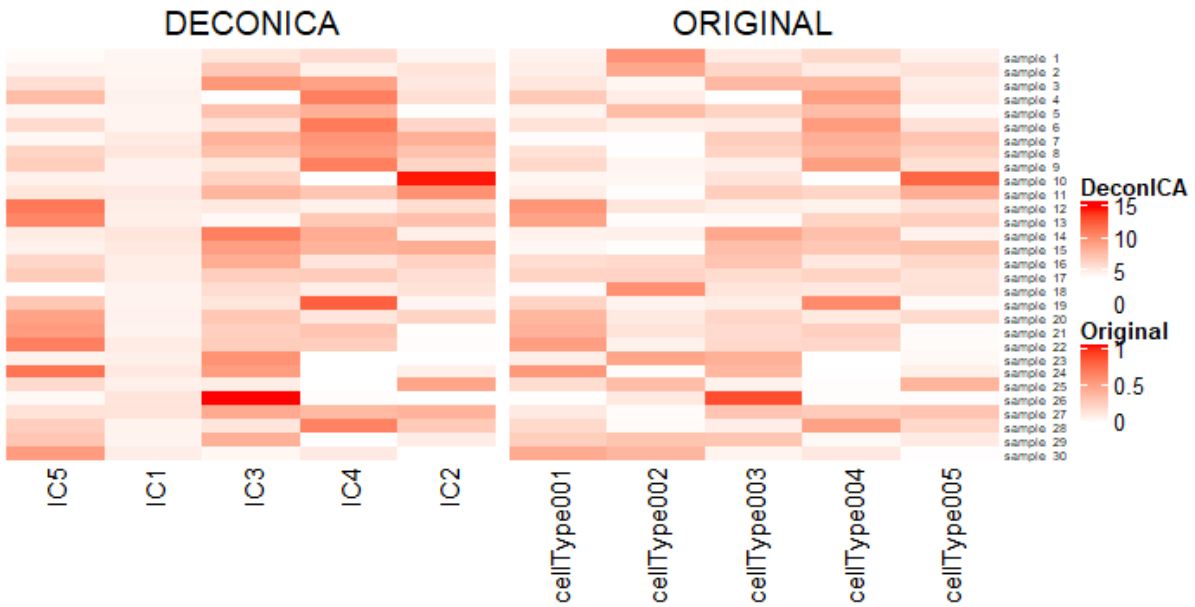


Gráfico S17. Heatmap deconICA puntuaciones (Simulación)

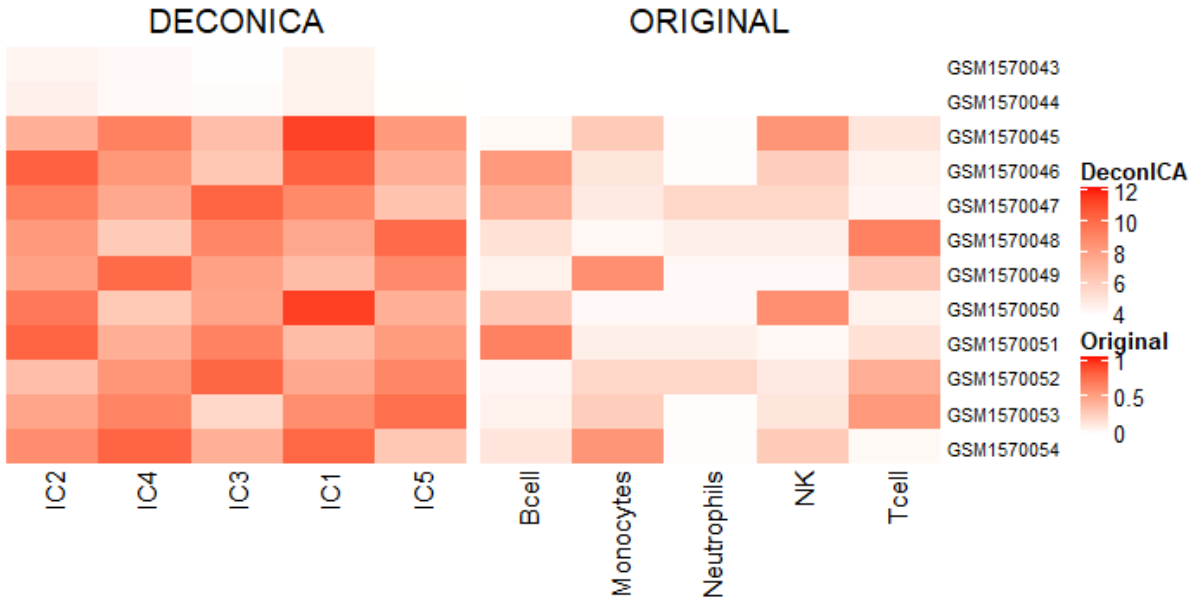


Gráfico S18. Heatmap deconICA puntuaciones (GSE64385)

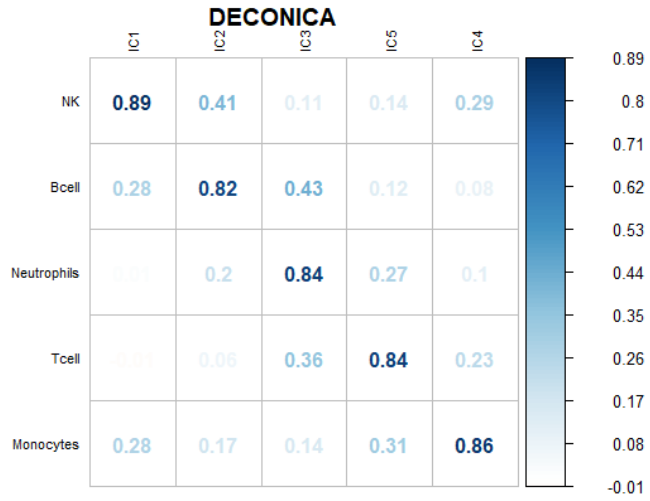


Gráfico S19. Correlación puntuaciones *deconICA* (GSE64385)

EVALUACIÓN ERROR Y DIVERGENCIA

Tabla S1. RMSE por muestras en cada método en los datos simulados

	CIBERSORT	***DTANGLE***	***MIND***	***LINSEED***	***DECONICA***
S1	0,0236	0,2213	0,0011	0,1749	0,1879
S2	0,0531	0,2568	0,0020	0,1861	0,1409
S3	0,0153	0,0971	0,0006	0,2020	0,0807
S4	0,0231	0,2382	0,0025	0,1438	0,1031
S5	0,0185	0,0993	0,0008	0,1164	0,1088
S6	0,0384	0,2361	0,0016	0,1802	0,0956
S7	0,0384	0,1976	0,0006	0,2258	0,1066
S8	0,0312	0,2146	0,0018	0,1908	0,0860
S9	0,0255	0,2442	0,0010	0,1717	0,0959
S10	0,0657	0,1292	0,0013	0,0849	0,1746
S11	0,0459	0,1933	0,0013	0,1861	0,0956
S12	0,0396	0,2268	0,0011	0,2057	0,1005
S13	0,0290	0,2398	0,0013	0,1891	0,1081
S14	0,0119	0,1940	0,0025	0,2083	0,0959
S15	0,0154	0,0774	0,0022	0,2180	0,0964
S16	0,0106	0,1206	0,0015	0,1715	0,0342
S17	0,0059	0,0701	0,0014	0,1184	0,0363
S18	0,0286	0,2176	0,0008	0,1945	0,1900
S19	0,0554	0,2118	0,0012	0,1738	0,1259
S20	0,0177	0,2451	0,0021	0,2034	0,0528
S21	0,0433	0,2172	0,0018	0,1493	0,0720
S22	0,0301	0,2425	0,0012	0,1969	0,1013
S23	0,0265	0,1359	0,0011	0,2193	0,1531
S24	0,0316	0,2340	0,0017	0,2598	0,1349
S25	0,0247	0,1135	0,0017	0,1596	0,1141
S26	0,0298	0,0866	0,0012	0,1017	0,1888
S27	0,0114	0,0592	0,0023	0,1994	0,0801
S28	0,0274	0,2457	0,0021	0,1775	0,0987
S29	0,0235	0,0792	0,0013	0,1584	0,0874
S30	0,0265	0,1161	0,0008	0,0871	0,1265

Tabla S2. RMSE por muestras en cada método en el GSE64385

	CIBERSORT	***DTANGLE***	***MIND***	***LINSEED***	***DECONICA***
S1	0,2865	0,2306	0,2209	0,0279	0,2002
S2	0,2763	0,2301	0,2224	0,0312	0,2003
S3	0,0776	0,1619	0,2167	0,0590	0,1584
S4	0,1830	0,1989	0,2057	0,0313	0,1521
S5	0,2450	0,1667	0,1586	0,1838	0,1080
S6	0,1788	0,1024	0,1262	0,0724	0,2000
S7	0,1385	0,1895	0,1946	0,0544	0,1841
S8	0,1265	0,1816	0,2316	0,0767	0,1690
S9	0,2808	0,2357	0,2223	0,1025	0,1993
S10	0,1822	0,0621	0,0619	0,1546	0,1107
S11	0,0755	0,0893	0,1022	0,0335	0,1565
S12	0,1182	0,1929	0,2198	0,0497	0,1654

Tabla S3. Muestras máximo RMSE
(eliminando las dos primeras muestras en el GSE64385)

RMSE	***SIMULACIÓN***		***GSE64385***	
Método	**Muestra**	**Valor**	**Muestra**	**Valor**
CIBERSORT	10	0.0657	7	0.2808
DTANGLE	2	0.2568	7	0.2357
MIND	4	0.0025	6	0.2316
LINSEED	24	0.2598	3	0.1838
DECONICA	18	0.1899	4	0.2

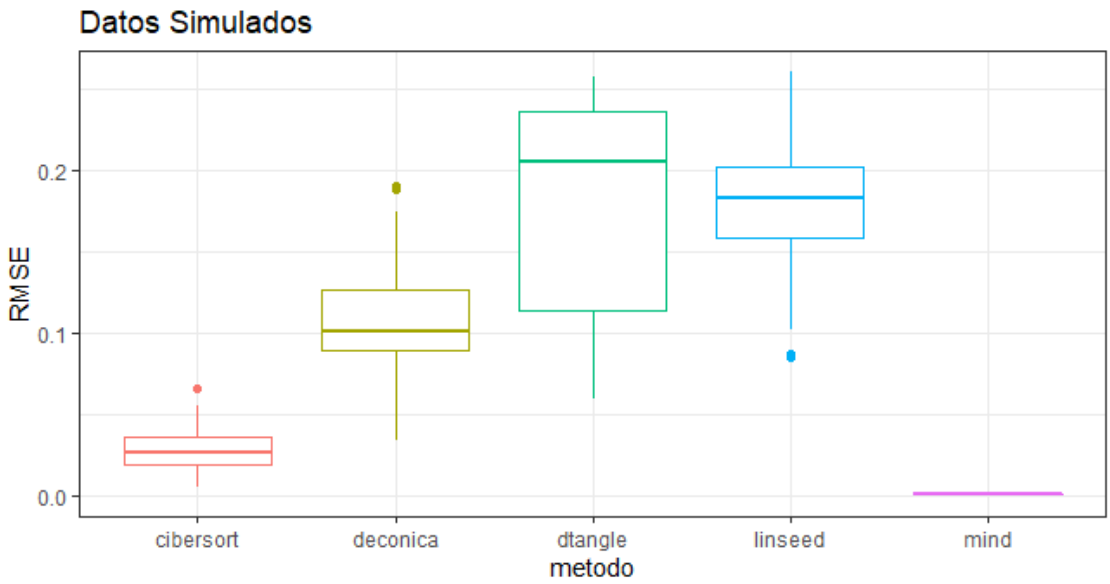


Gráfico S20. Boxplot RMSE Datos Simulados

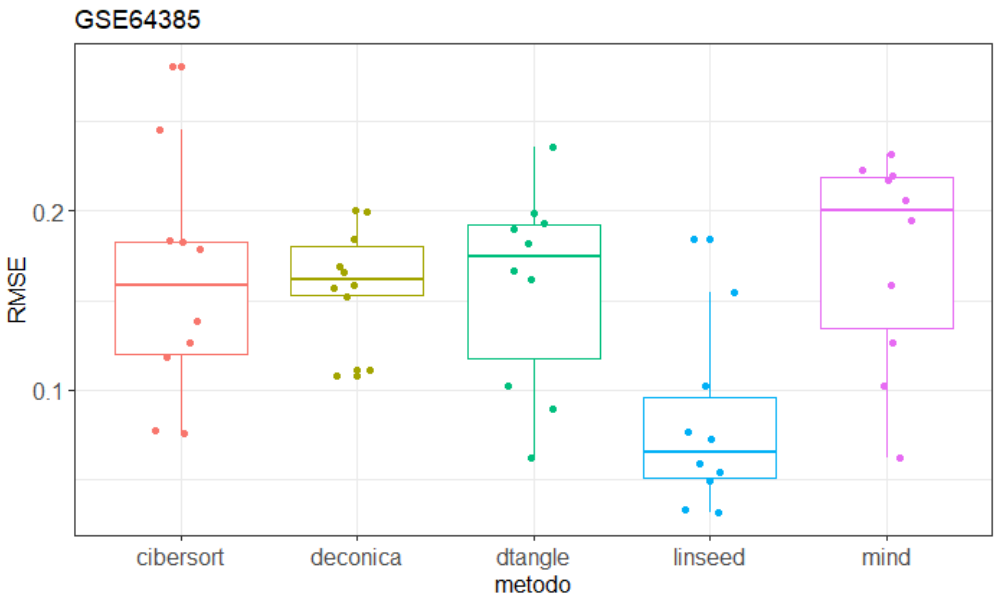


Gráfico S21. Boxplot con puntos RMSE GSE64385

Tabla S4. Divergencia Kullback-Leibler en los datos simulados

	CIBERSORT	***dtangle***	***MIND***	***LINSEED***	***DECONICA***
S1	0,0106	0,7703	0,0000	0,4474	0,4810
S2	0,0316	0,9008	0,0001	0,5156	0,2737
S3	0,0074	0,2835	0,0000	0,5891	0,1562
S4	0,0772	0,7343	0,0027	0,2881	0,5133
S5	0,0190	0,2086	0,0000	0,3055	0,3366
S6	0,0175	0,8403	0,0001	0,4824	0,1178
S7	0,0296	0,4057	0,0000	0,8999	0,6671
S8	0,0889	0,5221	0,0002	0,8032	0,6558
S9	0,0091	0,8462	0,0000	0,4393	0,1558
S10	0,1050	0,3945	0,0009	0,6680	0,7270
S11	0,0356	0,4332	0,0001	0,8446	0,4430
S12	0,0215	0,8060	0,0000	0,6357	0,1420
S13	0,0444	0,6858	0,0001	0,4677	0,4632
S14	0,0020	0,5141	0,0002	0,5874	0,1897
S15	0,0318	0,0884	0,0001	1,0452	0,5730
S16	0,0022	0,1880	0,0000	0,5720	0,0246
S17	0,0006	0,1160	0,0000	0,3719	0,0255
S18	0,0563	0,7606	0,0000	0,5548	0,5735
S19	0,0441	0,7036	0,0001	0,4312	0,3271
S20	0,0043	0,7311	0,0001	0,6106	0,0452
S21	0,0258	0,5763	0,0002	0,3650	0,2219
S22	0,0430	0,7920	0,0000	0,5405	0,2792
S23	0,2208	0,2770	0,0011	0,7691	1,1156
S24	0,2398	0,6104	0,0007	0,7902	1,2715
S25	0,0509	0,2744	0,0001	0,5106	0,4792
S26	0,0784	0,2391	0,0001	0,3793	0,9307
S27	0,0135	0,0880	0,0002	0,7832	0,3060
S28	0,0209	0,8019	0,0001	0,4824	0,2553
S29	0,0334	0,1506	0,0001	0,5110	0,2263
S30	0,0198	0,2708	0,0001	0,1560	0,5873

Tabla S5. Divergencia Kullback-Leibler GSE64385

	CIBERSORT	***dtangle***	***MIND***	***LINSEED***	***DECONICA***
S1	15,0475	14,5098	14,4284	14,3851	14,2893
S2	14,9573	14,5066	14,4391	14,4147	14,2899
S3	0,2957	0,5967	0,9243	0,0627	0,6815
S4	0,6279	0,8953	0,9854	0,0203	0,6102
S5	0,8240	0,5395	0,5764	0,4894	0,2171
S6	0,6614	0,1999	0,2511	0,1571	0,5571
S7	0,6382	0,6617	0,6118	0,1639	0,7137
S8	0,5415	0,7043	0,9816	0,1533	0,6104
S9	1,0805	0,7115	0,5929	0,2667	0,5665
S10	0,4418	0,0742	0,1310	0,3937	0,2330
S11	0,3152	0,3981	0,4621	0,0409	0,6163
S12	0,5143	1,0552	1,2102	0,1358	0,6994

Tabla S6. Test de Wilcoxon para la divergencia en los datos simulados y en el GSE64385

SIMULACIÓN				
	cibersort	**deconica**	**dtangle**	**linseed**
deconica	1.9e-08	-	-	-
dtangle	1.9e-08	0.579	-	-
linseed	1.9e-08	0.063	0.579	-
mind	1.9e-08	1.9e-08	1.9e-08	1.9e-08
GSE64385				
	cibersort	**deconica**	**dtangle**	**linseed**
deconica	1.000	-	-	-
dtangle	1.000	1.000	-	-
linseed	0.020	0.088	0.088	-
mind	1.000	1.000	0.387	0.088