Spectral clustering of Big Data in genetics: applications to RNA-seq data

Tsvetelin Zaevski, Ognyan Kunchev, Dean Palejev, Eugenia Stoimenova
Institute of Mathematics and Informatics, Bulgarian Academy of Sciences

14 септември 2017 г.

Обща схема на спектралната клъстеризация:

1. Нека $W \in R^{n\times n}$ със стойности w_{ij}, където n е броят на елементите, е съответната матрица на подобие. Нека също $d_i = \sum_{j=1,\dots,n} w_{ij}$ и D е диагоналната матрица, конструирана чрез d_i. Нека k е броят на клъстерите.
2. Изчисляваме Лапласиана $L = D - W$.
3. Намираме първите k собствени вектора $u_1, \dots, u_k$ на L.
4. Нека $U \in R^{n\times k}$ е матрицата конструирана от векторите $u_1, \dots, u_k$ като колони.
5. Нека $y_i \in R^k$, for $i = 1, \dots, n$, е съответният i-ти ред на U.
6. Клъстерираме точките $y_i, i = 1, \dots, n$, в k клъстера, $C_1, C_2, \dots, C_k$, използвайки k-means algorithm.
7. Клъстерите, $A_1, A_2, \dots, A_k$, на първоначалните дани се получават чрез $A_i = \{j, y_j \in C_i\}$.

Мишките на Bottomly:

1. 21 мишки – 10 от тях са здрави, а 11 болни.
2. 36536 гена, 8824 от тях имат изразяване повече от 8.
3. Има много методи, които разделят гените на важни и неважни.

1	803	433	469	585	321	301	461	309	374	781
56	55	27	44	32	47	40	40	30	27	46
58	32	19	18	44	22	17	24	29	15	34
78	860	528	401	584	401	331	431	341	480	930
88	653	365	414	454	331	473	548	413	395	1153
93	40	17	49	47	36	28	31	23	20	36
120	81	54	171	44	21	63	27	24	14	164
125	54	15	52	33	28	14	31	25	28	37
126	117	97	41	162	218	165	200	287	152	181
127	182	112	129	108	93	99	114	93	99	195
131	2173	1139	1683	1489	1463	1502	1794	1330	1286	2436

1. Поставяме очкваните важни и неважни гени като две отделни множества
2. Генерираме корелациони матрици – една чрез използване на $(2, 2)$ бета разпределение и друга с параметри $(2, 5)$.
3. Заменяме първоначалната корелационна матрица с генерираната. За всеки важен ген изчисляваме коефициент, който е сумата от квадратите на корелационните коефициенти между него и всички неважни гени. Използваме същото за неважните гени.
4. Има някои неважни гени, които имат много малка корелация с останалите (по малка от 0.03). Тези гени ги поставяме близо до неважните.
5. Избираме различни нива на значимост – $l = 0.6; 0.7; 0.8; 0.9; 0.98$ – и променяме корелационната матрица.
6. Матрицата на подобие – (1) евклидово разстояние за важните гени и (2) модифицираната корелационна матрица за останалите разстояния.

Резултати – използваме четири модела. Важните гени варират от 470 до 1145.
Броят на важните гени след клъстеризацията за бета разпределение $(2, 2)$ варира между

1. 471 и 3417 за модела на Anders and Huber (2010),
2. 1000 и 3753 за модела на Robinson et al. (2010),
3. 849 и 3655 за модела на Ritchie et al. (2015),
4. 1120 и 3844 за модела на M.I. Love and Anders (2014).

Броят на важните гени след клъстеризацията за бета разпределение $(2, 5)$ варира между

1. 486 и 817 за модела на Anders and Huber (2010),
2. 1010 и 1322 за модела на Robinson et al. (2010),
3. 845 и 1180 за модела на Ritchie et al. (2015),
4. 1155 и 1464 за модела на M.I. Love and Anders (2014).

1. Нека M е множеството от важните гени (след клъстерирането).
2. Нека M_1 е множеството от тези важни гени, които преди клъстерирането не са били важни.
3. Нека N е множеството от неважните гени.

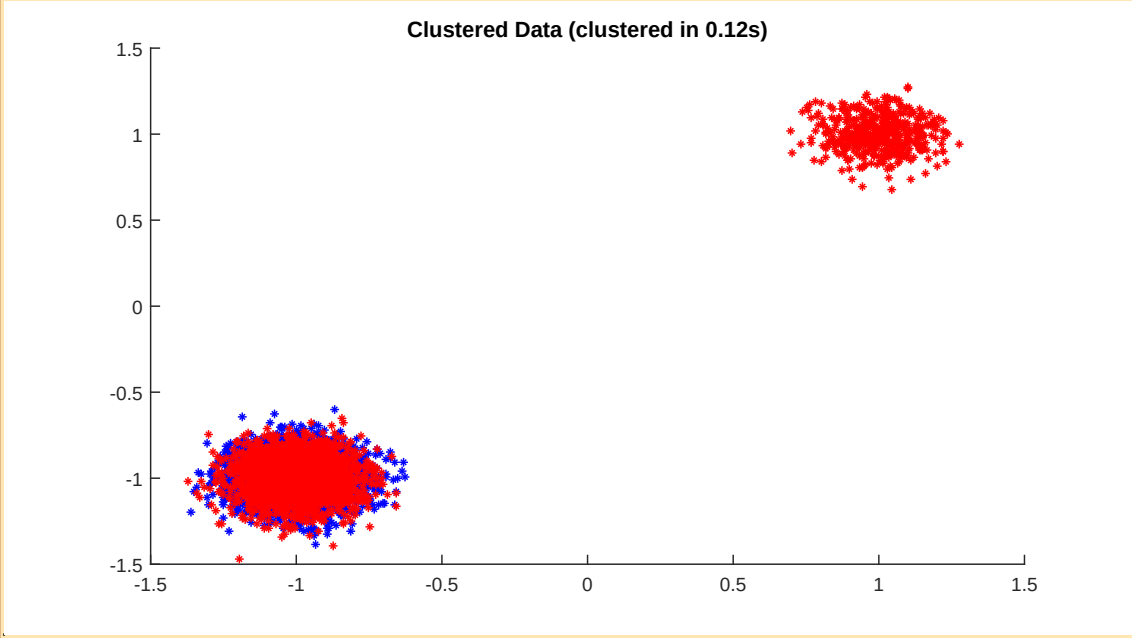
Следните две неравенства са верни – те обясняват математическата логика на алгоритъма.

$$\begin{aligned} \min\{\max\{|C(M, M_1)|\}\} > l \\ \max\{\max\{|C(M, N)|\}\} < l \end{aligned} \tag{1}$$

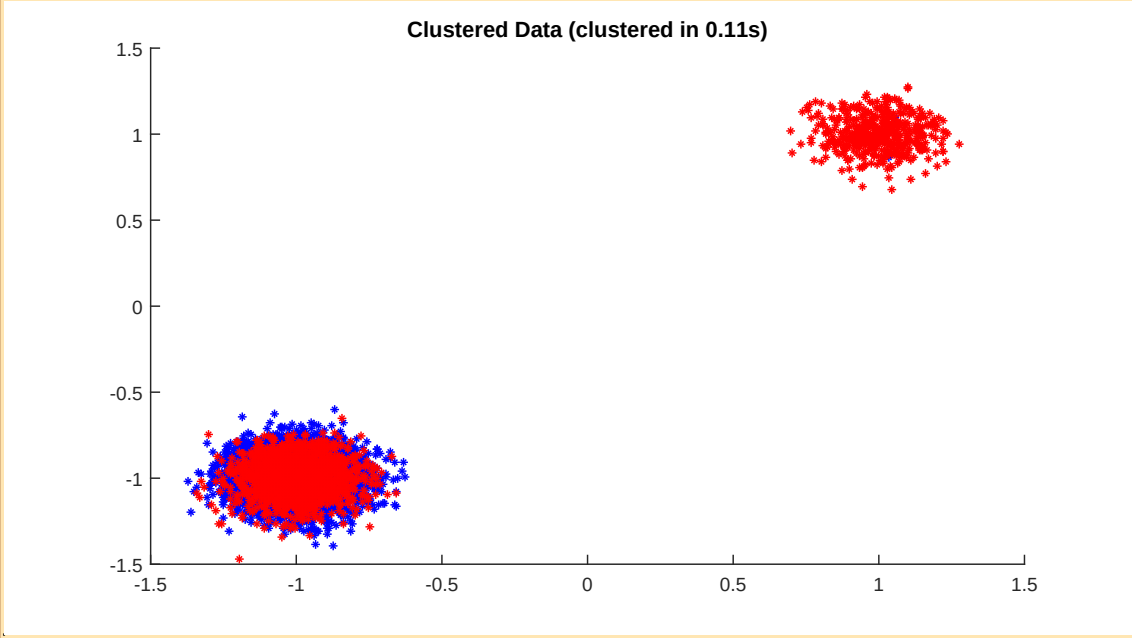
където l е нивото на значимост. Това означава, че

1. За всеки важен ген, за който първоначално сме мислели, че не е важен, съществува поне един друг важен ген, такъв че корелацията между двата е по-голяма от нивото на значимост l.
2. За всеки неважен ген няма нито един важен, такъв че корелацията между тях е по-голяма от нивото на значимост l.

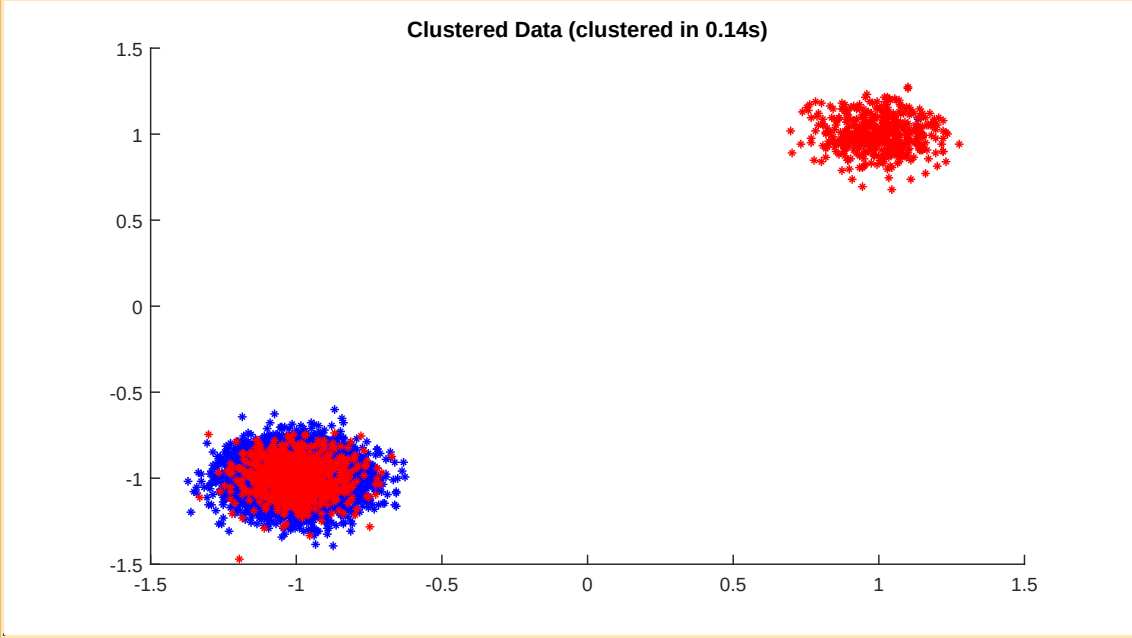
Beta parameters $(2, 2)$, significant level $l = 0.6$ important genes – 470



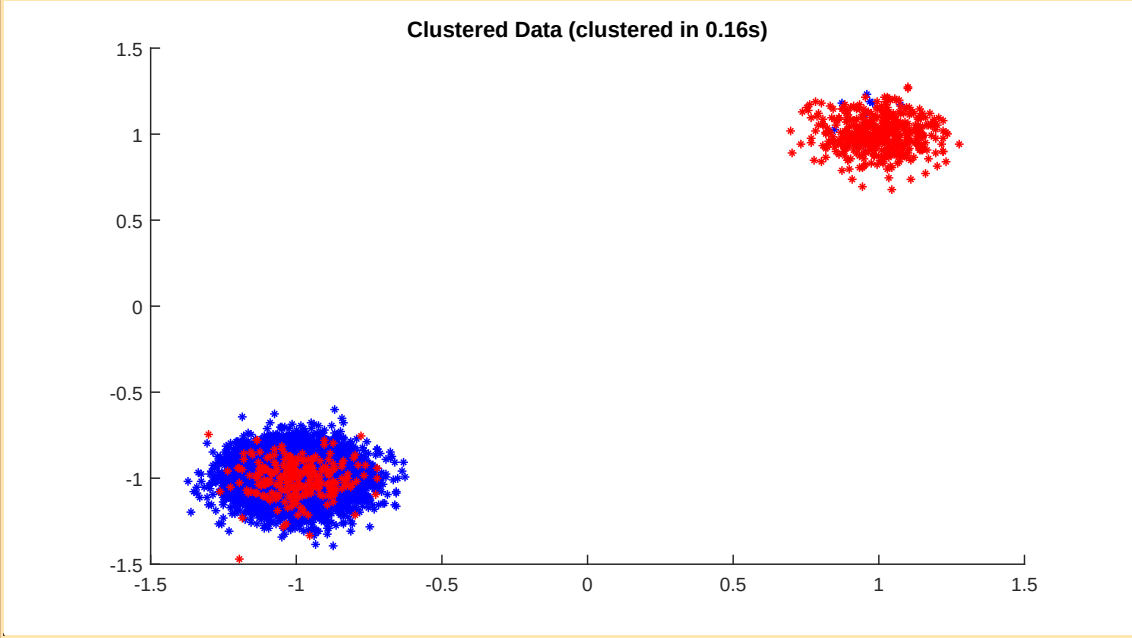
Beta parameters $(2, 2)$, significant level $l = 0.7$ important genes $-$ 470



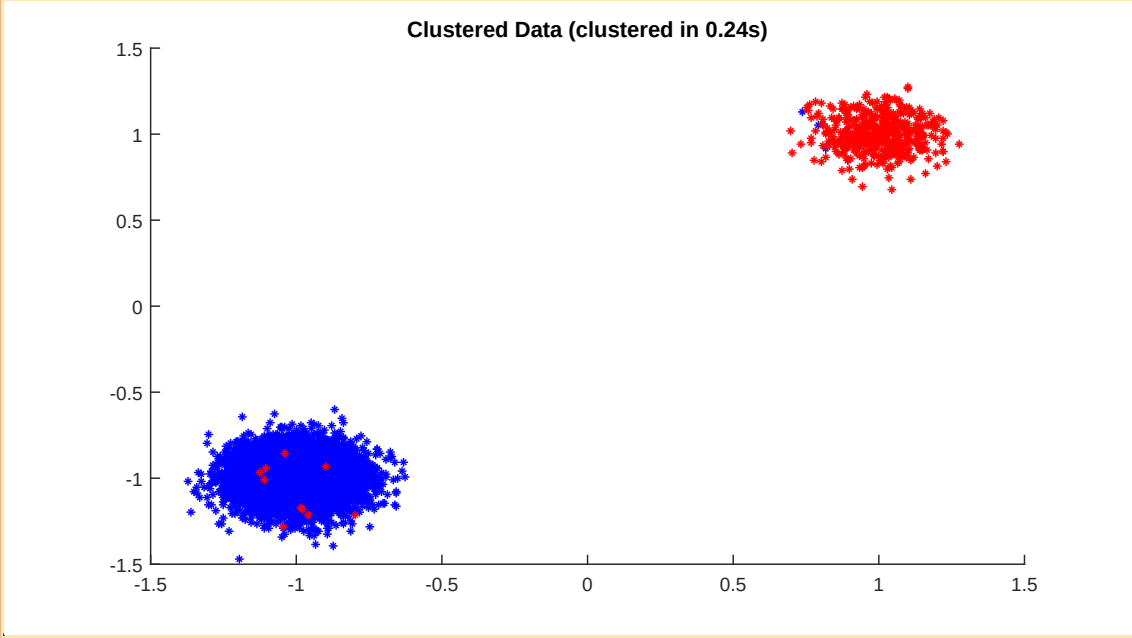
Beta parameters $(2, 2)$, significant level $l = 0.8$ important genes – 470



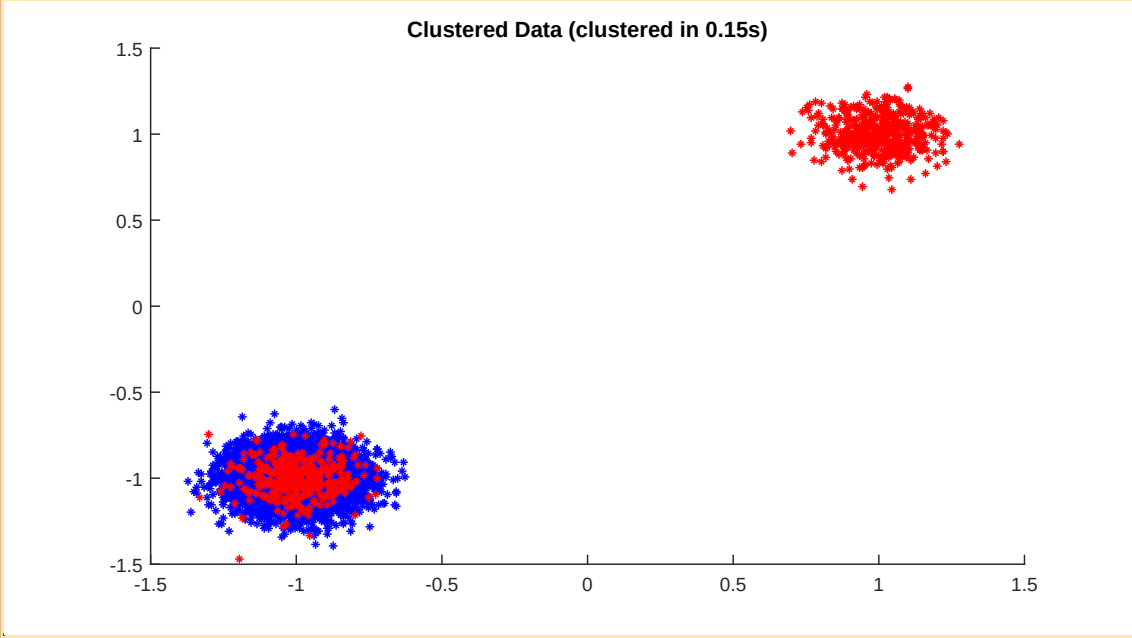
Beta parameters $(2, 2)$, significant level $l = 0.9$ important genes – 470



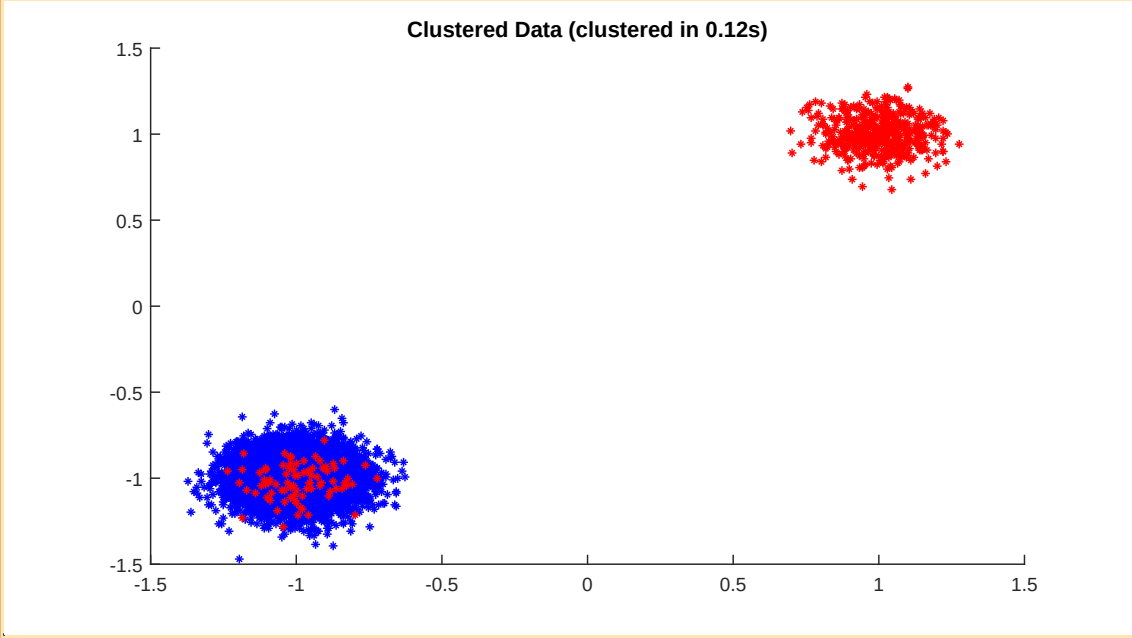
Beta parameters $(2, 2)$, significant level $l = 0.98$ important genes $-$ 470



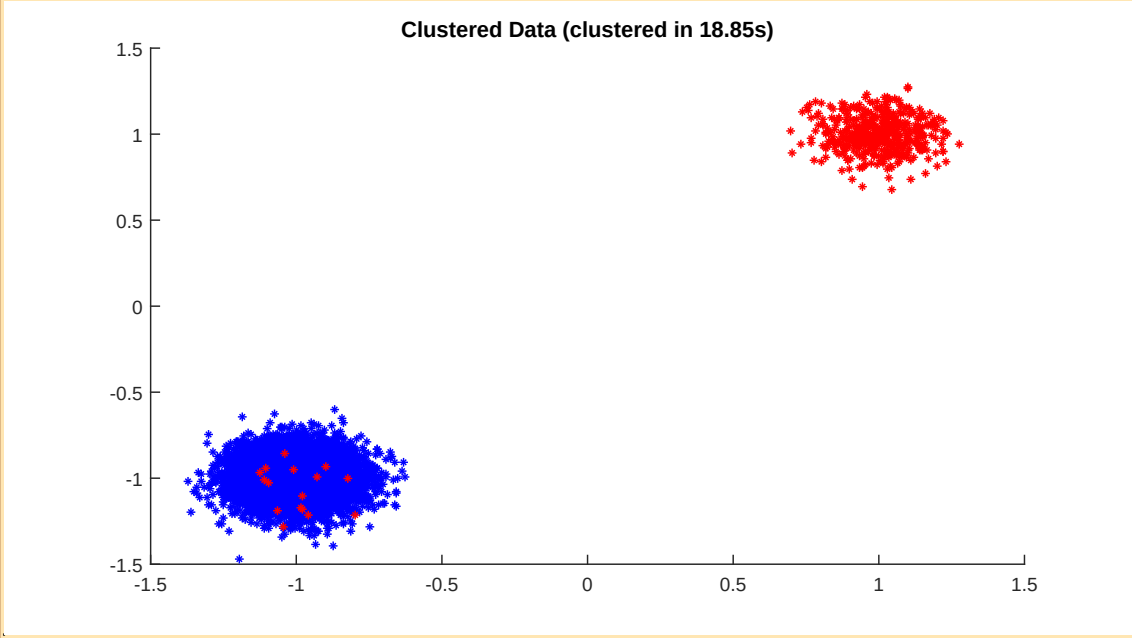
Beta parameters $(2, 5)$, significant level $l = 0.6$ important genes – 470



Beta parameters $(2, 5)$, significant level $l = 0.7$ important genes – 470



Beta parameters $(2, 5)$, significant level $l = 0.8$ important genes – 470



parameters	Expected important			Expected not important		
	Total	Imp.	Not imp.	Total	Imp.	Not imp.
(2,2) 0.6	470	470	0	8354	2947	5407
(2,2) 0.7	470	467	3	8354	1786	6568
(2,2) 0.8	470	469	1	8354	853	7501
(2,2) 0.9	470	460	10	8354	240	8114
(2,2) 0.98	470	461	9	8354	10	8344
(2,5) 0.6	470	470	0	8354	347	8007
(2,5) 0.7	470	469	1	8354	81	8273
(2,5) 0.8	470	470	0	8354	16	8338

Литература

Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106+, October 2010. ISSN 1465-6906.

W. Huber M.I. Love and S. Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, 15(4):550, 2014. doi: 10.1186/s13059-014-0550-8.

Matthew E. Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47, April 2015. ISSN 1362-4962. doi: 10.1093/nar/gkv007. URL http://dx.doi.org/10.1093/nar/gkv007.

Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, January 2010. ISSN 1460-2059. doi: 10.1093/bioinformatics/btp616. URL http://dx.doi.org/10.1093/bioinformatics/btp616.