Difficult genes and their impact on RNA-Seq data analysis

Alicja Szabelska-Beręsewicz Joanna Zyprych-Walczak
Idzi Siatkowski Michał Okoniewski

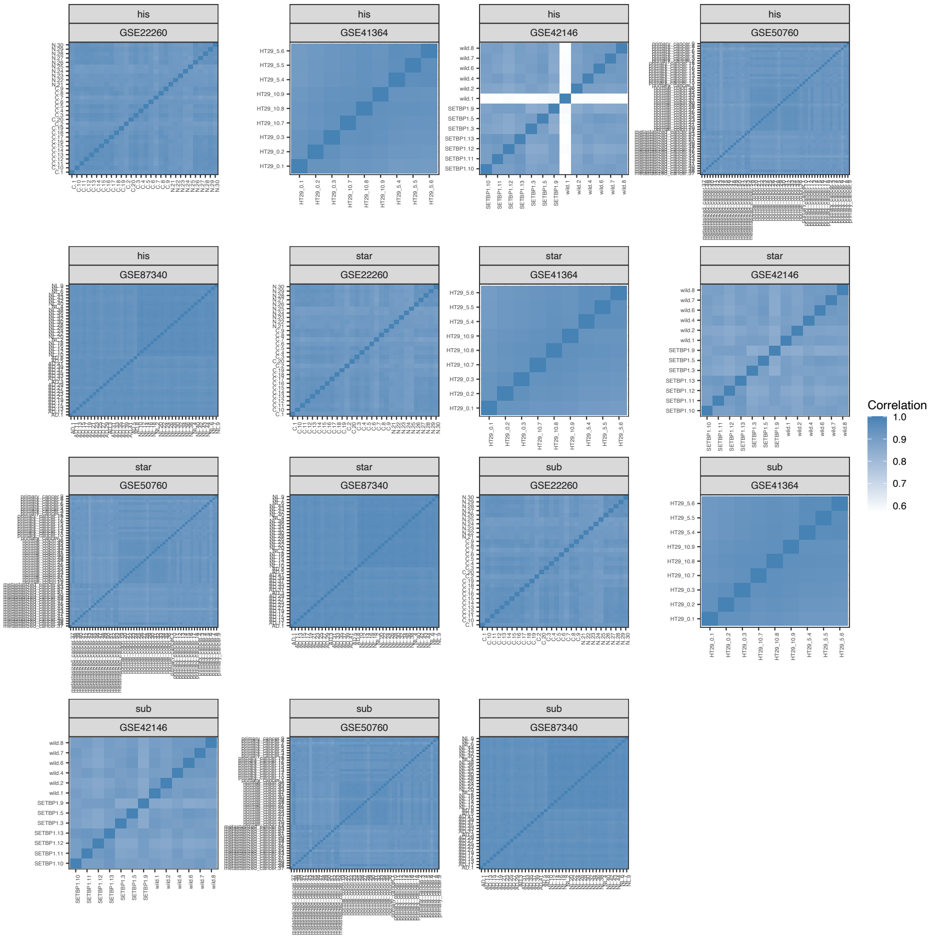
Contents

1 Introduction

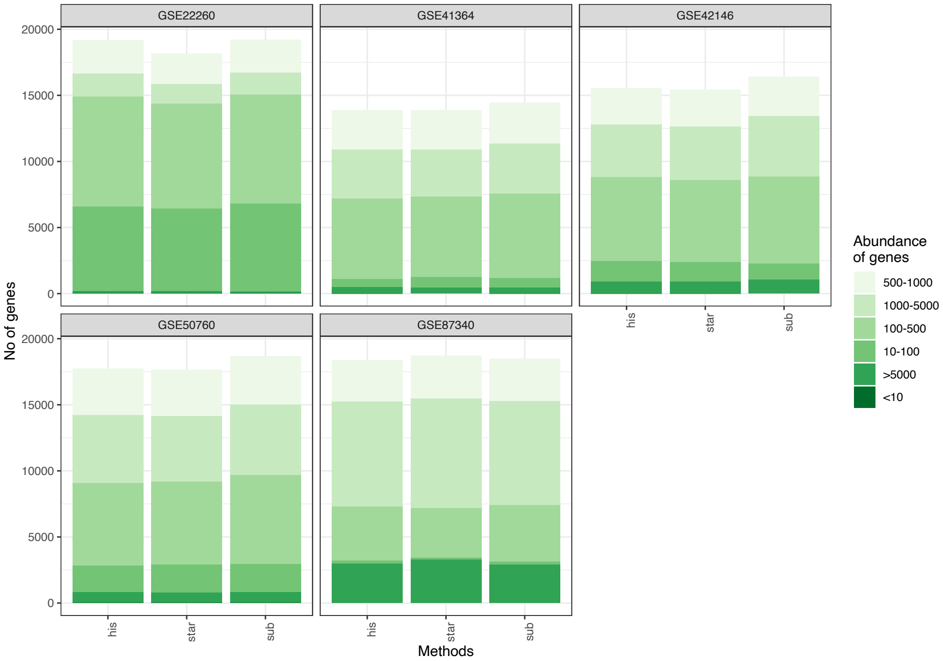
The issue being analysed is pinpointing the genes that cause systematically artifactual results in the analysis of RNA-seq. Such genes cannot be reliably measured and detected as differentially expressed. In particular the problem occurs, when popular genome aligners do not agree in the number and distribution of reads assigned to such genes. It causes confusion in reproducible data analysis. When such difficult genes are those of particular biological interest, it may distort the biological interpretation of the whole experiment. When difficult genes are the key ones in human metabolic pathways, the distorted results may be confusing for the further research in genomic personalised medicine.

2 Preliminary analysis

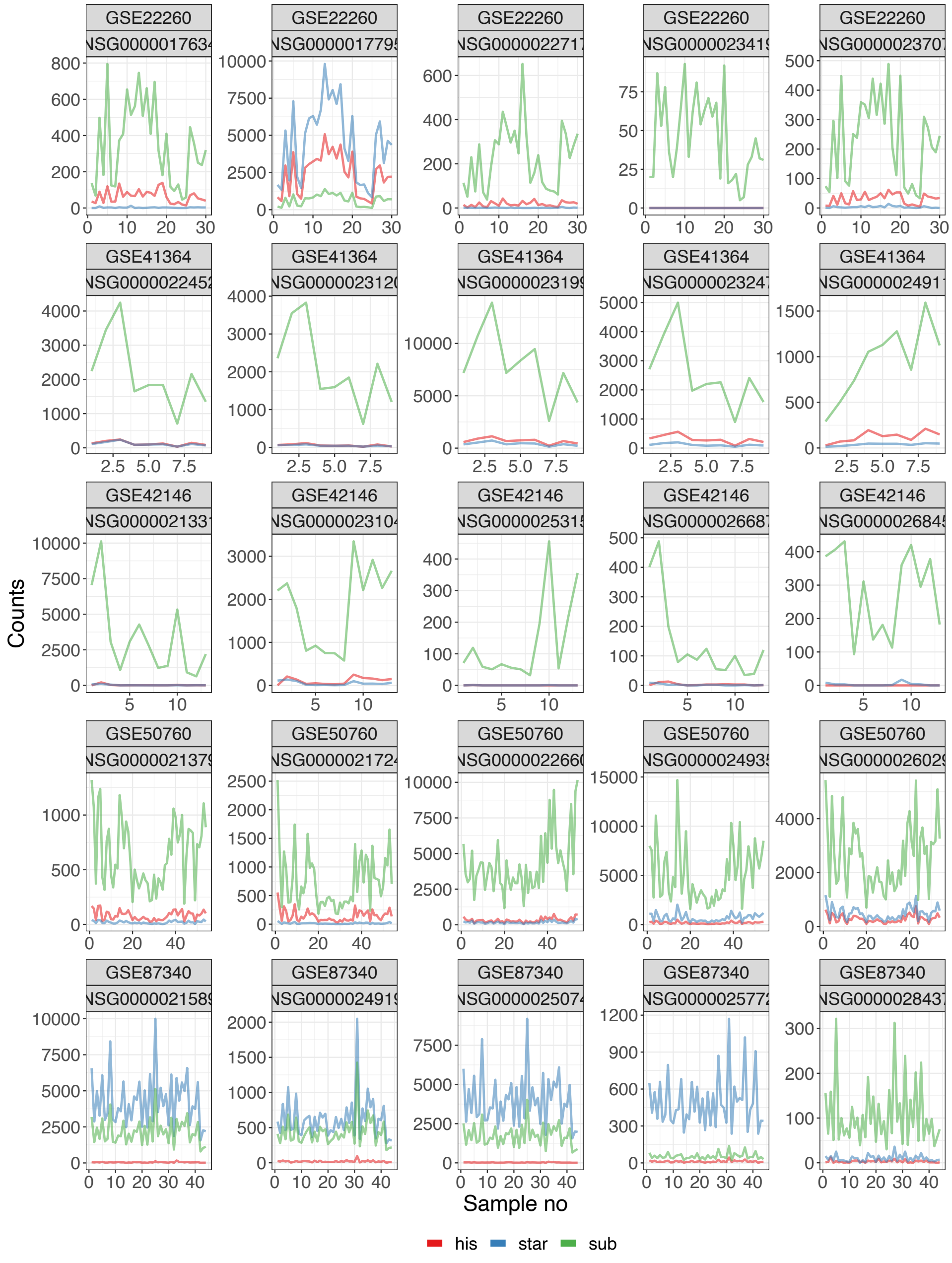
2.1 Heatmaps



2.2 Barplots


GSE22260
GSE41364
GSE42146
GSE50760
GSE87340
20000
15000
10000
5000
0
No of genes
his
star
sub
Methods
Abundance of genes
500-1000
1000-5000
100-500
10-100
>5000
<10

3 Difficult genes



4 Description of difficult genes

contingency table for methods edgeR, DESeq, limma + voom i limma + vst for dataset dataset4

	m.T	m.N	m.T	m.N	m.T	m.N	m.T	m.N
g.T	10.09	28.28	2.32	12.47	9.24	28.80	7.29	24.1
g.N	18.95	42.68	20.86	64.35	19.37	42.59	20.21	48.4

contingency table for methods edgeR, DESeq, limma + voom i limma + vst for dataset dataset5

	m.T	m.N	m.T	m.N	m.T	m.N	m.T	m.N
g.T	20.83	56.17	6.29	39.77	19.14	55.08	17.11	55.35
g.N	10.26	12.73	15.30	38.62	11.71	14.07	12.62	14.92

contingency table for methods edgeR, DESeq, limma + voom i limma + vst for dataset dataset6

	m.T	m.N	m.T	m.N	m.T	m.N	m.T	m.N
g.T	21.08	50.96	10.15	42.51	20.60	49.19	19.12	49.40
g.N	11.76	16.20	14.05	33.23	12.84	17.37	13.41	18.07

contingency table for methods edgeR, DESeq, limma + voom i limma + vst for dataset dataset7

	m.T	m.N	m.T	m.N	m.T	m.N	m.T	m.N
g.T	3.96	41.02	0.79	19.99	3.15	41.46	2.13	32.83
g.N	8.97	46.05	6.74	72.47	8.46	46.93	9.17	55.86

contingency table for methods edgeR, DESeq, limma + voom i limma + vst for dataset dataset8

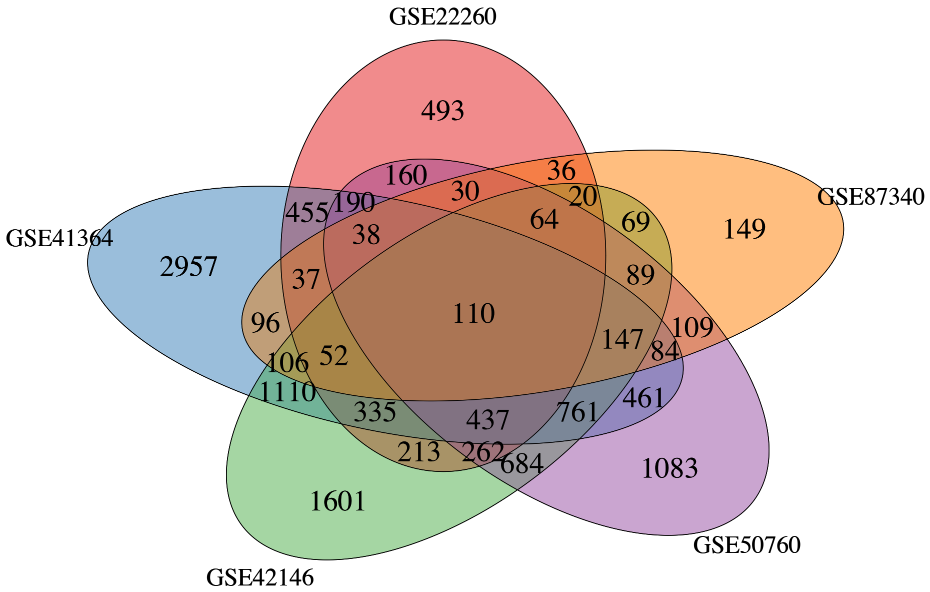
	m.T	m.N	m.T	m.N	m.T	m.N	m.T	m.N
g.T	16.01	62.67	9.32	56.37	15.45	62.35	14.41	63.72
g.N	8.06	13.26	9.77	24.54	8.00	14.20	8.27	13.61

Table 1: Percentage of significant genes due to mappers and groups across each dataset

	Dataset									
	GSE22260		GSE41364		GSE42146		GSE50760		GSE87340	
Mappers / Groups	yes	no	yes	no	yes	no	yes	no	yes	no
yes	10.09	28.28	16.01	62.67	3.96	41.02	20.83	56.17	21.08	50.96
no	18.95	42.68	8.06	13.26	8.97	46.05	10.26	12.73	11.76	16.20

Table 2: Percentage of significant genes due to mappers and groups across each dataset

Datasets	Mappers					
	Hisat		Star		Subread	
	% of all	% of DEG	% of all	% of DEG	% of all	% of DEG
GSE22260	0.10	17.93	0.25	27.11	0.20	22.75
GSE41364	11.60	23.38	13.34	26.19	12.62	25.39
GSE42146	11.73	20.70	11.61	20.43	14.52	25.92
GSE50760	10.08	18.04	9.90	17.67	11.40	20.63
GSE87340	0.44	4.81	0.44	4.22	0.77	7.27



5 Machine learning

5.1 Classification errors

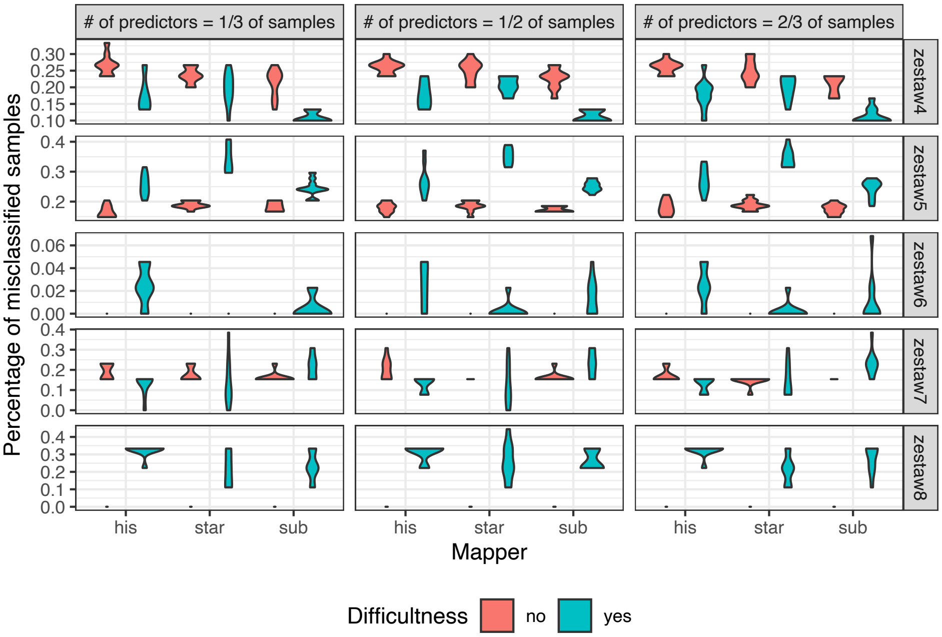


Table 3: Average AUC values for 10 simulations for considered datasets and mappers

Dataset	No of pred	Mapper/Difficultness					
		Hisat		Star		Subread	
		yes	no	yes	no	yes	no
GSE22260	10	0.575	0.893	0.569	0.904	0.717	0.945
	15	0.598	0.901	0.636	0.884	0.668	0.939
	20	0.610	0.903	0.905	0.879	0.908	0.913
GSE41364	18	0.890	0.881	0.893	0.814	0.899	0.877
	27	0.910	0.878	0.908	0.811	0.906	0.870
	36	0.897	0.858	0.906	0.795	0.901	0.865
GSE42146	14	1.000	0.991	1.000	1.000	1.000	0.998
	22	1.000	1.000	1.000	1.000	1.000	1.000
	29	1.000	1.000	1.000	1.000	1.000	1.000
GSE50760	4	0.773	0.977	0.913	0.838	0.899	0.788
	6	0.795	0.988	0.908	0.970	0.894	0.874
	8	0.865	0.995	0.948	0.969	0.918	0.811
GSE87340	3	1.000	0.886	1.000	0.930	1.000	0.958
	4	1.000	0.944	1.000	0.998	1.000	1.000
	6	1.000	0.945	1.000	0.990	1.000	0.996