Difficult genes and their impact on RNA-Seq data analysis

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

Contents

1 Introduction **1**

2 Preliminary analysis **1**

2.1 Heatmaps 1

2.2 Library sizes 2

2.3 Structure of counts 4

3 Difficult genes **6**

4 Description of difficult genes **7**

4.1 Characteristics of DGs 7

4.2 Number of DGs 7

4.3 Number of DGs with mappers separately 9

4.4 Venn diagrams 9

5 Machine learning **9**

5.1 Classification errors 9

5.2 AUC values 10

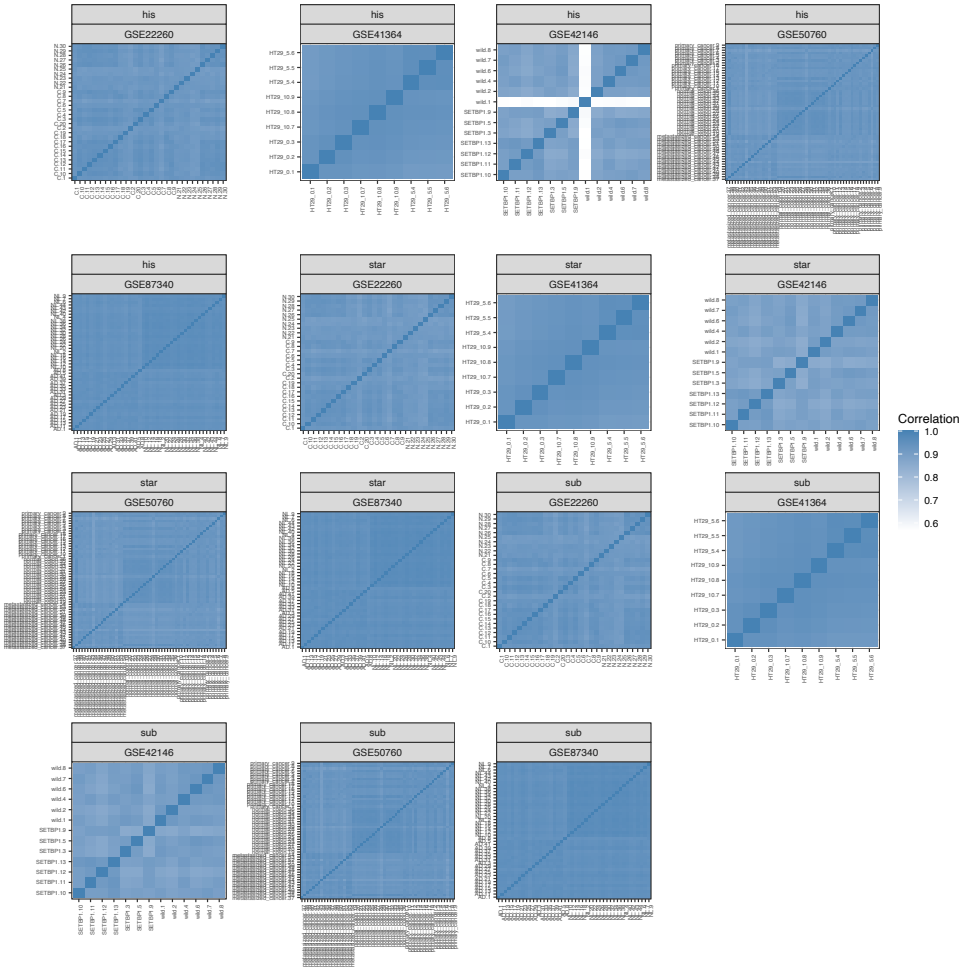
1 Introduction

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

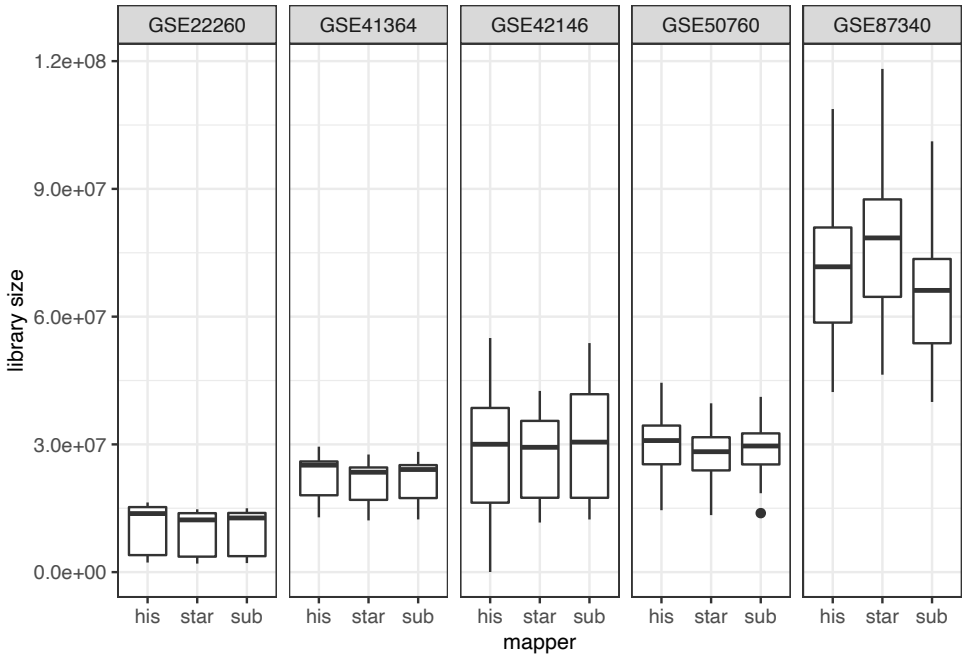
2.1 Heatmaps

These plots are checking whether experimental design is proper.



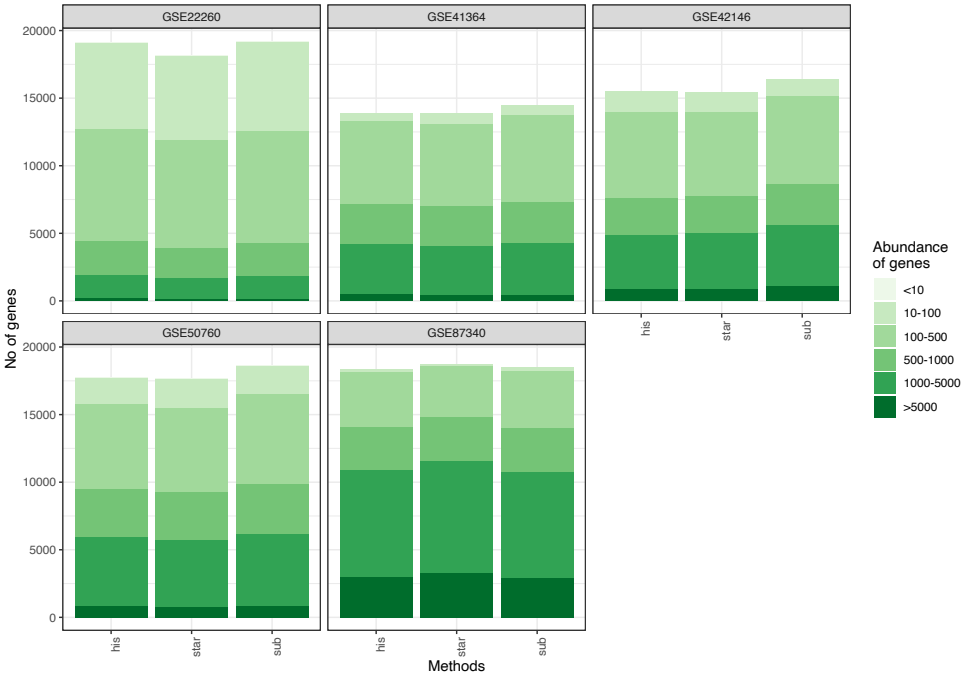
2.2 Library sizes

Distribution of library sizes in each dataset and for each maper.

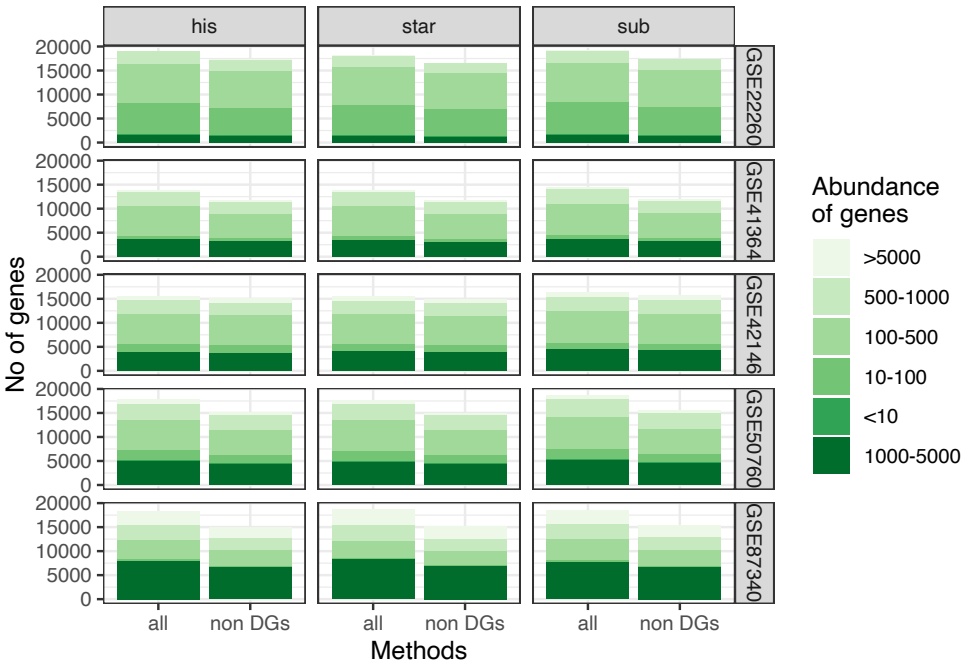

GSE22260
GSE41364
GSE42146
GSE50760
GSE87340
1.2e+08
9.0e+07
6.0e+07
3.0e+07
0.0e+00
library size
his
star
sub
mapper

2.3 Structure of counts

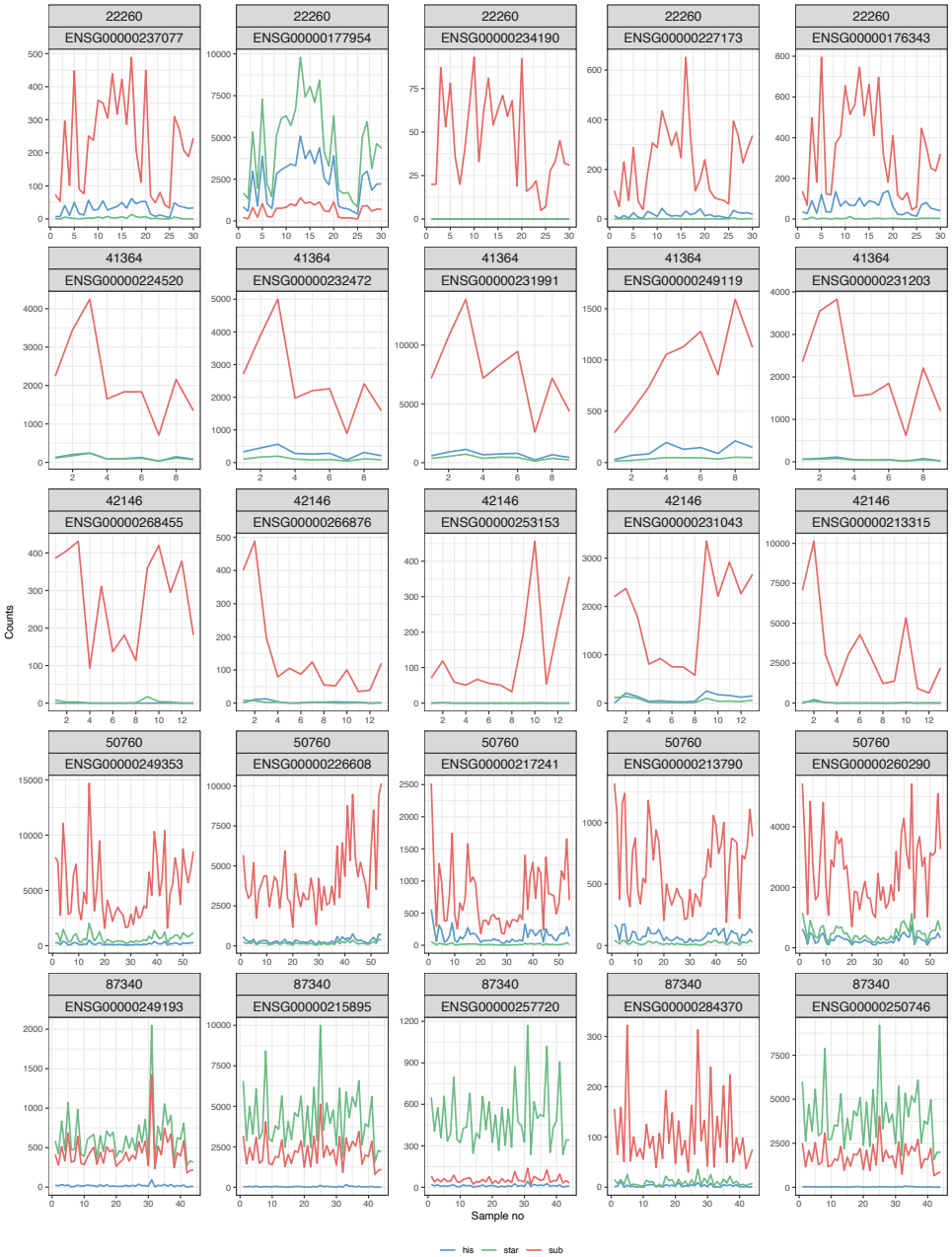
2.3.1 Barplots without division



2.3.2 Barplots with division



3 Difficult genes



4 Description of difficult genes

4.1 Characteristics of DGs

Loading annotation for exons.

Table 1: Characteristics of DG genes found in each dataset. Columns represent Ex.min - the length of the shortest exon, Ex.max - the length of the longest exon, Ex.mean - the mean exon length, Ex.no - the total number of exons as well as Tr.length - the transcript length. Each value was calculated as a mean value across all transcripts in datasets with division to DG and other genes.

Dataset	Type	Ex.min	Ex.max	Ex.median	Tr.length	Ex.no	Tr.no	Pseudogenes.per
GSE22260	DG	251	506	335	764	1	1	58
GSE22260	non DG	78	483	156	994	4	5	4
GSE41364	DG	132	492	242	780	2	1	39
GSE41364	non DG	80	473	156	930	4	4	8
GSE42146	DG	285	480	341	678	1	1	57
GSE42146	non DG	81	469	159	897	4	4	9
GSE50760	DG	140	475	252	744	2	1	42
GSE50760	non DG	76	478	154	1008	4	5	4
GSE87340	DG	169	484	274	757	2	1	55
GSE87340	non DG	82	487	161	960	4	4	5

4.2 Number of DGs

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

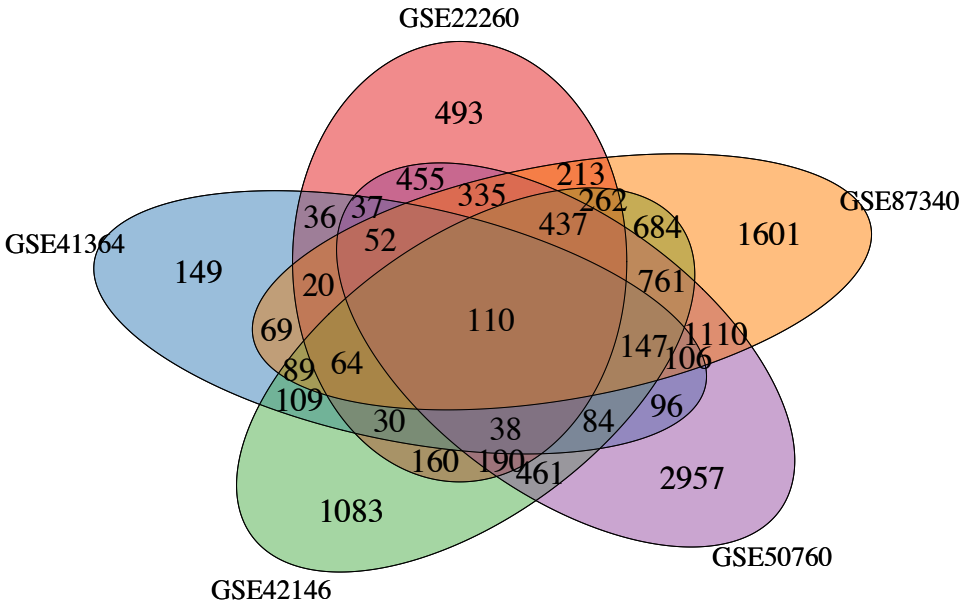
	Dataset									
	GSE22260		GSE41364		GSE42146		GSE50760		GSE87340	
Groups \ Mappers	yes	no	yes	no	yes	no	yes	no	yes	no
yes	10.09	28.28	3.96	41.02	16.01	62.67	21.08	50.96	20.83	56.17
no	18.95	42.68	8.97	46.05	8.06	13.26	11.76	16.20	10.26	12.73

Table 3: 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	10.08	18.04	9.90	17.67	11.40	20.63
GSE42146	0.44	4.81	0.44	4.22	0.77	7.27
GSE50760	11.73	20.70	11.61	20.43	14.52	25.92
GSE87340	11.60	23.38	13.34	26.19	12.62	25.39

4.3 Number of DGs with mappers separately

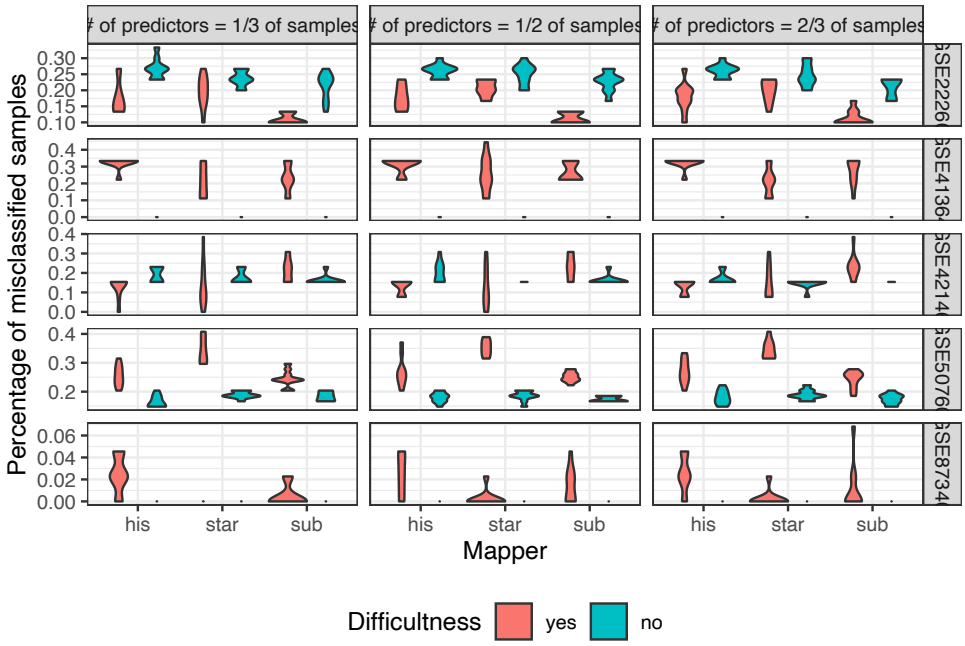
4.4 Venn diagrams



5 Machine learning

5.1 Classification errors

Before this code you have to run codes from file “errors_joint_classifier.R”



5.2 AUC values

Table 4: 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	0.893	0.575	0.904	0.569	0.945	0.717
	15	0.901	0.598	0.884	0.636	0.939	0.668
	20	0.903	0.610	0.879	0.905	0.913	0.908
GSE41364	3	0.886	1.000	0.930	1.000	0.958	1.000
	4	0.944	1.000	0.998	1.000	1.000	1.000
	6	0.945	1.000	0.990	1.000	0.996	1.000
GSE42146	4	0.977	0.773	0.838	0.913	0.788	0.899
	6	0.988	0.795	0.970	0.908	0.874	0.894
	8	0.995	0.865	0.969	0.948	0.811	0.918
GSE50760	18	0.881	0.890	0.814	0.893	0.877	0.899
	27	0.878	0.910	0.811	0.908	0.870	0.906
	36	0.858	0.897	0.795	0.906	0.865	0.901
GSE87340	14	0.991	1.000	1.000	1.000	0.998	1.000
	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