Experiment 3949: AEmond

Nucleomics Core

November 21, 2021

Contents

Nucleomics Core, VIB
Herestraat 49
3000 LEUVEN,
Post Box 816
Belgium

VIB NUCLEOMICS CORE

Contact info:
nucleomics@vib.be
+32/16/373128
+32/16/373129
www.nucleomics.be

1 Experimental design

Organism: *Linum_usitatissimumJGI_152022v1.0*
Platform: *NovaSeq*
Library Prep Kit: *TruSeq*
Sequencing Kit: *NovaSeq6000 S1 Flowcell 100 cycles*
Fragments: *single end / fr-firststrand*

File	Sample	Lane	Barcode
A2@01@S48@ACTAAGAT@CCGCGGTT@L001	A2@01	L001	CCGCGGTT—ACTAAGAT
A2@01@S48@ACTAAGAT@CCGCGGTT@L002	A2@01	L002	CCGCGGTT—ACTAAGAT
A3@02@S49@GTCGGAGC@TTATAACC@L001	A3@02	L001	TTATAACC—GTCGGAGC
A3@02@S49@GTCGGAGC@TTATAACC@L002	A3@02	L002	TTATAACC—GTCGGAGC
A7@S50@CTTGGTAT@GGACTTGG@L001	A7	L001	GGACTTGG—CTTGGTAT
A7@S50@CTTGGTAT@GGACTTGG@L002	A7	L002	GGACTTGG—CTTGGTAT
B3@04@S51@TCCAACGC@AAGTCCAA@L001	B3@04	L001	AAGTCCAA—TCCAACGC
B3@04@S51@TCCAACGC@AAGTCCAA@L002	B3@04	L002	AAGTCCAA—TCCAACGC
B6@S52@CCGTGAAG@ATCCACTG@L001	B6	L001	ATCCACTG—CCGTGAAG
B6@S52@CCGTGAAG@ATCCACTG@L002	B6	L002	ATCCACTG—CCGTGAAG
B8@S53@TTACAGGA@GCTTGTCA@L001	B8	L001	GCTTGTCA—TTACAGGA
B8@S53@TTACAGGA@GCTTGTCA@L002	B8	L002	GCTTGTCA—TTACAGGA
E3@07@S54@GGCATTCT@CAAGCTAG@L001	E3@07	L001	CAAGCTAG—GGCATTCT
E3@07@S54@GGCATTCT@CAAGCTAG@L002	E3@07	L002	CAAGCTAG—GGCATTCT
E4@08@S55@AATGCCTC@TGGATCGA@L001	E4@08	L001	TGGATCGA—AATGCCTC
E4@08@S55@AATGCCTC@TGGATCGA@L002	E4@08	L002	TGGATCGA—AATGCCTC
E6@S56@TACCGAGG@AGTTCAGG@L001	E6	L001	AGTTCAGG—TACCGAGG
E6@S56@TACCGAGG@AGTTCAGG@L002	E6	L002	AGTTCAGG—TACCGAGG
F1@10@S57@CGTTAGAA@GACCTGAA@L001	F1@10	L001	GACCTGAA—CGTTAGAA
F1@10@S57@CGTTAGAA@GACCTGAA@L002	F1@10	L002	GACCTGAA—CGTTAGAA
F4@11@S58@AGCCTCAT@TCTCTACT@L001	F4@11	L001	TCTCTACT—AGCCTCAT
F4@11@S58@AGCCTCAT@TCTCTACT@L002	F4@11	L002	TCTCTACT—AGCCTCAT
F7@S59@GATTCTGC@CTCTCGTC@L001	F7	L001	CTCTCGTC—GATTCTGC
F7@S59@GATTCTGC@CTCTCGTC@L002	F7	L002	CTCTCGTC—GATTCTGC
G1@S60@TCGTAGTG@CCAAGTCT@L001	G1	L001	CCAAGTCT—TCGTAGTG
G1@S60@TCGTAGTG@CCAAGTCT@L002	G1	L002	CCAAGTCT—TCGTAGTG
G2@S61@CTACGACA@TTGGACTC@L001	G2	L001	TTGGACTC—CTACGACA
G2@S61@CTACGACA@TTGGACTC@L002	G2	L002	TTGGACTC—CTACGACA
G5@S62@TAAGTGGT@GGCTTAAG@L001	G5	L001	GGCTTAAG—TAAGTGGT
G5@S62@TAAGTGGT@GGCTTAAG@L002	G5	L002	GGCTTAAG—TAAGTGGT
H4@S63@CGGACAAC@AATCCGGA@L001	H4	L001	AATCCGGA—CGGACAAC
H4@S63@CGGACAAC@AATCCGGA@L002	H4	L002	AATCCGGA—CGGACAAC
H6@S64@ATATGGAT@TAATACAG@L001	H6	L001	TAATACAG—ATATGGAT
H6@S64@ATATGGAT@TAATACAG@L002	H6	L002	TAATACAG—ATATGGAT
H7@S65@GCGCAAGC@CGGCGTGA@L001	H7	L001	CGGCGTGA—GCGCAAGC
H7@S65@GCGCAAGC@CGGCGTGA@L002	H7	L002	CGGCGTGA—GCGCAAGC
M4@S66@AAGATACT@ATGTAAGT@L001	M4	L001	ATGTAAGT—AAGATACT
M4@S66@AAGATACT@ATGTAAGT@L002	M4	L002	ATGTAAGT—AAGATACT
M5@S67@GGAGCGTC@GCACGGAC@L001	M5	L001	GCACGGAC—GGAGCGTC
M5@S67@GGAGCGTC@GCACGGAC@L002	M5	L002	GCACGGAC—GGAGCGTC
M6@S68@ATGGCATG@GGTACCTT@L001	M6	L001	GGTACCTT—ATGGCATG
M6@S68@ATGGCATG@GGTACCTT@L002	M6	L002	GGTACCTT—ATGGCATG
N1@S69@GCAATGCA@AACGTTCC@L001	N1	L001	AACGTTCC—GCAATGCA

Continued on next page

File	Sample	Lane	Barcode
N1@S69@GCAATGCA@AACGTTCC@L002	N1	L002	AACGTTCC—GCAATGCA
N5@S70@GTTCCAAT@GCAGAATT@L001	N5	L001	GCAGAATT—GTTCCAAT
N5@S70@GTTCCAAT@GCAGAATT@L002	N5	L002	GCAGAATT—GTTCCAAT
N6@S71@ACCTTGGC@ATGAGGCC@L001	N6	L001	ATGAGGCC—ACCTTGGC
N6@S71@ACCTTGGC@ATGAGGCC@L002	N6	L002	ATGAGGCC—ACCTTGGC

2 Data quality assessment

3 Data quality assessment

3.1 Yield Statistics

The table provides details per sample and per read direction about the number of bases, number of reads and read length. In case that the fraction of reads of a sample deviates with a factor 3 from the expected fraction (i.e., 2.04 %), the data of that sample are indicated in red.

	# Bases	*# Fragments*	*Read Length*	*% Fragments*
A2@L001	1,830,770,000	18,307,700	100	2.76
A2@L002	1,827,165,500	18,271,655	100	2.76
A3@L001	1,230,615,600	12,306,156	100	1.86
A3@L002	1,247,120,400	12,471,204	100	1.88
A7@L001	1,586,428,100	15,864,281	100	2.39
A7@L002	1,583,301,700	15,833,017	100	2.39
B3@L001	1,522,133,800	15,221,338	100	2.3
B3@L002	1,531,846,500	15,318,465	100	2.31
B6@L001	1,300,139,300	13,001,393	100	1.96
B6@L002	1,301,684,700	13,016,847	100	1.96
B8@L001	1,319,747,400	13,197,474	100	1.99
B8@L002	1,335,397,300	13,353,973	100	2.02
E3@L001	1,266,479,700	12,664,797	100	1.91
E3@L002	1,276,541,200	12,765,412	100	1.93
E4@L001	1,382,506,900	13,825,069	100	2.09
E4@L002	1,390,043,800	13,900,438	100	2.1
E6@L001	1,583,111,600	15,831,116	100	2.39
E6@L002	1,580,387,800	15,803,878	100	2.38
F1@L001	1,545,993,700	15,459,937	100	2.33
F1@L002	1,557,749,000	15,577,490	100	2.35
F4@L001	1,756,281,200	17,562,812	100	2.65
F4@L002	1,739,354,600	17,393,546	100	2.62
F7@L001	1,178,147,500	11,781,475	100	1.78
F7@L002	1,193,420,700	11,934,207	100	1.8
G1@L001	1,347,447,800	13,474,478	100	2.03
G1@L002	1,356,638,500	13,566,385	100	2.05
G2@L001	1,167,001,600	11,670,016	100	1.76
G2@L002	1,176,803,900	11,768,039	100	1.78
G5@L001	1,588,967,800	15,889,678	100	2.4
G5@L002	1,594,146,200	15,941,462	100	2.41
H4@L001	1,289,951,700	12,899,517	100	1.95
H4@L002	1,296,603,000	12,966,030	100	1.96
H6@L001	1,483,555,300	14,835,553	100	2.24
H6@L002	1,487,791,700	14,877,917	100	2.25

Continued on next page

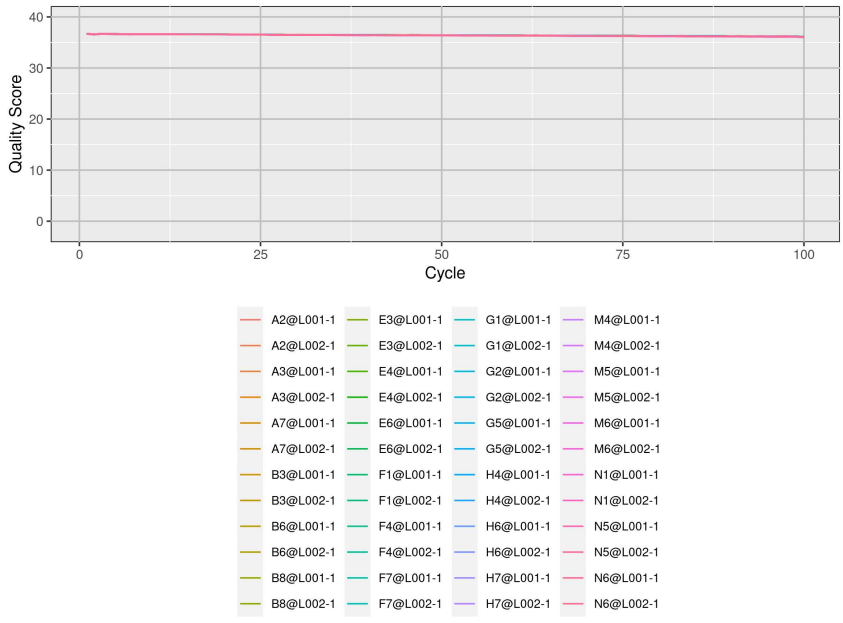
	# Bases	# Fragments	Read Length	% Fragments
H7@L001	1,343,634,800	13,436,348	100	2.03
H7@L002	1,356,412,100	13,564,121	100	2.05
M4@L001	1,296,187,500	12,961,875	100	1.96
M4@L002	1,296,651,500	12,966,515	100	1.96
M5@L001	1,061,303,900	10,613,039	100	1.6
M5@L002	1,076,831,200	10,768,312	100	1.62
M6@L001	1,343,375,500	13,433,755	100	2.03
M6@L002	1,346,437,600	13,464,376	100	2.03
N1@L001	1,108,866,300	11,088,663	100	1.67
N1@L002	1,125,992,600	11,259,926	100	1.7
N5@L001	1,446,824,300	14,468,243	100	2.18
N5@L002	1,443,612,000	14,436,120	100	2.18
N6@L001	1,072,975,200	10,729,752	100	1.62
N6@L002	1,093,029,600	10,930,296	100	1.65
Total	66,267,409,600	662,674,096	NA	NA
In total	66,267,409,600	662,674,096		

3.2 Accuracy Statistics

Base calling accuracy, measured by the Phred quality score (Q score), is the most common metric used to assess the accuracy of a sequencing platform. It indicates the probability that a given base is called incorrectly by the sequencer:

$$P(\text{error}) = 10^{\frac{-\text{Qscore}}{10}}$$

When sequencing quality reaches Q30, virtually all of the reads will be perfect (having zero errors and ambiguities). This is why Q30 is considered a benchmark for quality in next-generation sequencing. The following figure shows for each sample the average base quality per cycle, as calculated by the ShortRead 1.40.0 package from Bioconductor (http://www.bioconductor.org) [9]. Reported quality scores typically decline with cycle, in an accelerating manner.



3.3 Library Complexity

Sequencing outcomes of good quality should closely reflect the original library complexity.

3.3.1 Relative A,C,G,T Content

The table below provides details of the percentage of base calls per sample and per read direction. Base frequencies should accurately reflect the expected frequencies of the sequenced regions. For many applications, frequencies for A/T and G/C should be comparable.

	Forward read				
	A	*C*	*G*	*T*	*N*
A2@L001	25.81	25.35	23.17	25.68	0
A2@L002	25.8	25.35	23.18	25.67	0
A3@L001	26.15	25.29	22.79	25.77	0
A3@L002	26.13	25.3	22.8	25.77	0
A7@L001	26.02	25.3	22.94	25.74	0
A7@L002	26.01	25.3	22.94	25.74	0
B3@L001	26	25.32	22.89	25.79	0
B3@L002	25.99	25.33	22.89	25.79	0
B6@L001	26.29	25.23	22.67	25.81	0
B6@L002	26.28	25.24	22.67	25.81	0
B8@L001	26.24	25.23	22.65	25.88	0
B8@L002	26.22	25.24	22.66	25.88	0
E3@L001	26.23	25.25	22.74	25.78	0
E3@L002	26.21	25.26	22.74	25.78	0
E4@L001	25.96	25.43	22.91	25.69	0
E4@L002	25.95	25.44	22.92	25.69	0
E6@L001	26.09	25.37	22.8	25.74	0
E6@L002	26.08	25.37	22.82	25.73	0
F1@L001	26.13	25.24	22.81	25.82	0
F1@L002	26.11	25.25	22.82	25.81	0
F4@L001	26.27	25.08	22.74	25.91	0
F4@L002	26.26	25.09	22.75	25.9	0
F7@L001	26.3	25.17	22.63	25.9	0
F7@L002	26.29	25.18	22.64	25.89	0
G1@L001	25.97	25.13	23.08	25.82	0
G1@L002	25.96	25.14	23.08	25.82	0
G2@L001	26.69	25.06	22.33	25.93	0
G2@L002	26.67	25.07	22.33	25.93	0
G5@L001	25.94	25.34	22.97	25.75	0
G5@L002	25.93	25.35	22.97	25.75	0
H4@L001	26.28	25.28	22.65	25.78	0
H4@L002	26.27	25.29	22.66	25.78	0
H6@L001	26.35	25.29	22.54	25.82	0
H6@L002	26.33	25.31	22.54	25.82	0
H7@L001	26.54	25.17	22.42	25.88	0
H7@L002	26.52	25.17	22.43	25.88	0
M4@L001	27.43	24.77	21.32	26.48	0
M4@L002	27.42	24.78	21.32	26.47	0
M5@L001	27.2	24.8	21.52	26.48	0
M5@L002	27.18	24.81	21.53	26.47	0
M6@L001	27.45	24.62	21.3	26.62	0
M6@L002	27.43	24.64	21.3	26.62	0
N1@L001	27.17	24.93	21.53	26.36	0
N1@L002	27.15	24.94	21.55	26.37	0
N5@L001	26.77	25.03	21.85	26.35	0

Continued on next page

	Forward read				
	A	*C*	*G*	*T*	*N*
N5@L002	26.76	25.05	21.85	26.34	0
N6@L001	27.39	25	21.33	26.28	0
N6@L002	27.38	25	21.34	26.28	0

3.3.2 Uniform Frequency of Reads

The table below provides information whether some reads are overrepresented in the output. The columns give data about the percentage of unique reads and the frequency of the most observed read.

	Forward read	
	Unique Reads [%]	*Max Number*
A2@L001	27.86	7,826
A2@L002	27.42	7,688
A3@L001	33.17	5,226
A3@L002	32.65	5,256
A7@L001	29.15	6,777
A7@L002	28.77	6,710
B3@L001	31.48	5,760
B3@L002	30.98	5,628
B6@L001	32.86	4,937
B6@L002	32.51	5,105
B8@L001	33.37	5,077
B8@L002	32.81	5,159
E3@L001	31.13	5,678
E3@L002	30.67	5,710
E4@L001	30.47	5,407
E4@L002	29.97	5,542
E6@L001	29.33	6,798
E6@L002	28.94	6,638
F1@L001	31.65	5,667
F1@L002	31.19	5,715
F4@L001	28.71	6,331
F4@L002	28.38	6,409
F7@L001	34.85	4,446
F7@L002	34.38	4,690
G1@L001	32.91	4,769
G1@L002	32.46	4,467
G2@L001	32.74	4,471
G2@L002	32.26	4,483
G5@L001	30.02	5,877
G5@L002	29.58	5,778
H4@L001	31.8	5,077
H4@L002	31.42	5,011
H6@L001	29.95	6,296
H6@L002	29.49	6,182
H7@L001	32.33	5,305
H7@L002	31.78	5,550
M4@L001	34.16	2,399
M4@L002	33.68	2,369
M5@L001	38.65	1,702

Continued on next page

	Forward read	
	Unique Reads [%]	*Max Number*
M5@L002	38.1	1,718
M6@L001	35.44	1,893
M6@L002	34.87	1,937
N1@L001	37.77	2,027
N1@L002	37.14	2,003
N5@L001	32.26	3,025
N5@L002	31.85	3,113
N6@L001	36.04	3,252
N6@L002	35.43	3,318
Average	32.18	4,837.54

3.3.3 Diversity in Read Quality

Following figures show the distributions of the average read quality, as calculated by the ShortRead 1.40.0 package from Bioconductor (http://www.bioconductor.org) [9]. The figures provide insight whether good quality bases are grouped together in a few reads or are spread over the entire library. Samples with consistently good quality reads have unimodal, strong peaks near the right of the panel.

3.4 Data transfer

The data are stored in 48 compressed fastq-file(s) with extension '.fastq.gz'.

4 Read preprocessing

4.1 Materials and methods

The fastq-files are processed to remove as much technical artifacts as possible. The following preprocessing steps were performed:

1. *Quality trimming:* We trim low quality ends ($< Q20$) with FastX 0.0.14 [5]. Reads that are shorter than 35bp after trimming are removed.
   The following table shows the number of remaining reads after the quality trimming step:

SAMPLE	INITIAL # READS	# REMOVED READS (%)	# REMAINING READS (%)
A2@L001-1	18,307,700	10,796 (0.06)	18,296,904 (99.94)
A2@L002-1	18,271,655	10,797 (0.06)	18,260,858 (99.94)
A3@L001-1	12,306,156	7,654 (0.06)	12,298,502 (99.94)
A3@L002-1	12,471,204	7,647 (0.06)	12,463,557 (99.94)
A7@L001-1	15,864,281	10,148 (0.06)	15,854,133 (99.94)
A7@L002-1	15,833,017	10,251 (0.06)	15,822,766 (99.94)
B3@L001-1	15,221,338	9,648 (0.06)	15,211,690 (99.94)
B3@L002-1	15,318,465	9,831 (0.06)	15,308,634 (99.94)
B6@L001-1	13,001,393	8,595 (0.07)	12,992,798 (99.93)
B6@L002-1	13,016,847	8,784 (0.07)	13,008,063 (99.93)
B8@L001-1	13,197,474	7,480 (0.06)	13,189,994 (99.94)
B8@L002-1	13,353,973	7,412 (0.06)	13,346,561 (99.94)
E3@L001-1	12,664,797	8,163 (0.06)	12,656,634 (99.94)
E3@L002-1	12,765,412	8,002 (0.06)	12,757,410 (99.94)
E4@L001-1	13,825,069	8,273 (0.06)	13,816,796 (99.94)
E4@L002-1	13,900,438	8,056 (0.06)	13,892,382 (99.94)
E6@L001-1	15,831,116	10,349 (0.07)	15,820,767 (99.93)
E6@L002-1	15,803,878	10,466 (0.07)	15,793,412 (99.93)
F1@L001-1	15,459,937	9,557 (0.06)	15,450,380 (99.94)
F1@L002-1	15,577,490	9,702 (0.06)	15,567,788 (99.94)
F4@L001-1	17,562,812	11,693 (0.07)	17,551,119 (99.93)
F4@L002-1	17,393,546	11,702 (0.07)	17,381,844 (99.93)
F7@L001-1	11,781,475	7,302 (0.06)	11,774,173 (99.94)
F7@L002-1	11,934,207	7,350 (0.06)	11,926,857 (99.94)
G1@L001-1	13,474,478	8,113 (0.06)	13,466,365 (99.94)
G1@L002-1	13,566,385	8,128 (0.06)	13,558,257 (99.94)
G2@L001-1	11,670,016	11,206 (0.1)	11,658,810 (99.9)
G2@L002-1	11,768,039	11,277 (0.1)	11,756,762 (99.9)
G5@L001-1	15,889,678	9,637 (0.06)	15,880,041 (99.94)
G5@L002-1	15,941,462	9,865 (0.06)	15,931,597 (99.94)
H4@L001-1	12,899,517	9,987 (0.08)	12,889,530 (99.92)
H4@L002-1	12,966,030	9,988 (0.08)	12,956,042 (99.92)
H6@L001-1	14,835,553	10,750 (0.07)	14,824,803 (99.93)
H6@L002-1	14,877,917	10,714 (0.07)	14,867,203 (99.93)
H7@L001-1	13,436,348	10,257 (0.08)	13,426,091 (99.92)
H7@L002-1	13,564,121	10,179 (0.08)	13,553,942 (99.92)
M4@L001-1	12,961,875	15,061 (0.12)	12,946,814 (99.88)
M4@L002-1	12,966,515	15,159 (0.12)	12,951,356 (99.88)
M5@L001-1	10,613,039	8,961 (0.08)	10,604,078 (99.92)
M5@L002-1	10,768,312	9,208 (0.09)	10,759,104 (99.91)
M6@L001-1	13,433,755	13,011 (0.1)	13,420,744 (99.9)
M6@L002-1	13,464,376	13,247 (0.1)	13,451,129 (99.9)
N1@L001-1	11,088,663	10,035 (0.09)	11,078,628 (99.91)
N1@L002-1	11,259,926	10,235 (0.09)	11,249,691 (99.91)
N5@L001-1	14,468,243	10,744 (0.07)	14,457,499 (99.93)
N5@L002-1	14,436,120	10,849 (0.08)	14,425,271 (99.92)
N6@L001-1	10,729,752	11,916 (0.11)	10,717,836 (99.89)
N6@L002-1	10,930,296	12,547 (0.11)	10,917,749 (99.89)

2. *Adapter trimming:* The adapters are trimmed only at the end (at least 10bp overlap and 90% match) with

cutadapt 1.15 [8]. Reads that are shorter than 35bp after adapter trimming are removed.
The following table shows the number of remaining reads after the adapter trimming step:

SAMPLE	# QUALITY TRIMMED READS	# REMOVED READS (%)	# REMAINING READS (%)
A2@L001-1	18,296,904	1,441 (0.01)	18,295,463 (99.99)
A2@L002-1	18,260,858	1,428 (0.01)	18,259,430 (99.99)
A3@L001-1	12,298,502	408 (0)	12,298,094 (100)
A3@L002-1	12,463,557	451 (0)	12,463,106 (100)
A7@L001-1	15,854,133	644 (0)	15,853,489 (100)
A7@L002-1	15,822,766	642 (0)	15,822,124 (100)
B3@L001-1	15,211,690	639 (0)	15,211,051 (100)
B3@L002-1	15,308,634	574 (0)	15,308,060 (100)
B6@L001-1	12,992,798	528 (0)	12,992,270 (100)
B6@L002-1	13,008,063	473 (0)	13,007,590 (100)
B8@L001-1	13,189,994	265 (0)	13,189,729 (100)
B8@L002-1	13,346,561	230 (0)	13,346,331 (100)
E3@L001-1	12,656,634	318 (0)	12,656,316 (100)
E3@L002-1	12,757,410	276 (0)	12,757,134 (100)
E4@L001-1	13,816,796	552 (0)	13,816,244 (100)
E4@L002-1	13,892,382	525 (0)	13,891,857 (100)
E6@L001-1	15,820,767	856 (0.01)	15,819,911 (99.99)
E6@L002-1	15,793,412	855 (0.01)	15,792,557 (99.99)
F1@L001-1	15,450,380	876 (0.01)	15,449,504 (99.99)
F1@L002-1	15,567,788	964 (0.01)	15,566,824 (99.99)
F4@L001-1	17,551,119	950 (0.01)	17,550,169 (99.99)
F4@L002-1	17,381,844	856 (0)	17,380,988 (100)
F7@L001-1	11,774,173	914 (0.01)	11,773,259 (99.99)
F7@L002-1	11,926,857	1,001 (0.01)	11,925,856 (99.99)
G1@L001-1	13,466,365	1,162 (0.01)	13,465,203 (99.99)
G1@L002-1	13,558,257	1,204 (0.01)	13,557,053 (99.99)
G2@L001-1	11,658,810	816 (0.01)	11,657,994 (99.99)
G2@L002-1	11,756,762	824 (0.01)	11,755,938 (99.99)
G5@L001-1	15,880,041	823 (0.01)	15,879,218 (99.99)
G5@L002-1	15,931,597	797 (0.01)	15,930,800 (99.99)
H4@L001-1	12,889,530	944 (0.01)	12,888,586 (99.99)
H4@L002-1	12,956,042	981 (0.01)	12,955,061 (99.99)
H6@L001-1	14,824,803	692 (0)	14,824,111 (100)
H6@L002-1	14,867,203	736 (0)	14,866,467 (100)
H7@L001-1	13,426,091	804 (0.01)	13,425,287 (99.99)
H7@L002-1	13,553,942	797 (0.01)	13,553,145 (99.99)
M4@L001-1	12,946,814	975 (0.01)	12,945,839 (99.99)
M4@L002-1	12,951,356	989 (0.01)	12,950,367 (99.99)
M5@L001-1	10,604,078	1,043 (0.01)	10,603,035 (99.99)
M5@L002-1	10,759,104	1,036 (0.01)	10,758,068 (99.99)
M6@L001-1	13,420,744	1,530 (0.01)	13,419,214 (99.99)
M6@L002-1	13,451,129	1,564 (0.01)	13,449,565 (99.99)
N1@L001-1	11,078,628	208 (0)	11,078,420 (100)
N1@L002-1	11,249,691	204 (0)	11,249,487 (100)
N5@L001-1	14,457,499	596 (0)	14,456,903 (100)
N5@L002-1	14,425,271	556 (0)	14,424,715 (100)
N6@L001-1	10,717,836	629 (0.01)	10,717,207 (99.99)
N6@L002-1	10,917,749	614 (0.01)	10,917,135 (99.99)

3. *Quality filtering:* Using FastX 0.0.14 and ShortRead 1.40.0, we remove polyA-reads (more than 90% of

the bases equal A), ambiguous reads (containing N), low quality reads (more than 50% of the bases $< Q25$) and artifact reads (all but 3 bases in the read equal one base type).
The following table shows the number of reads that are filtered and the number of remaining reads after the filtering step:

SAMPLE	# TRIMMED READS	# FILTERED READS (%)	# REMAINING READS (%)
A2@L001-1	18,295,463	7,675 (0.04)	18,287,788 (99.96)
A2@L002-1	18,259,430	7,821 (0.04)	18,251,609 (99.96)
A3@L001-1	12,298,094	5,336 (0.04)	12,292,758 (99.96)
A3@L002-1	12,463,106	5,401 (0.04)	12,457,705 (99.96)
A7@L001-1	15,853,489	6,679 (0.04)	15,846,810 (99.96)
A7@L002-1	15,822,124	6,894 (0.04)	15,815,230 (99.96)
B3@L001-1	15,211,051	6,544 (0.04)	15,204,507 (99.96)
B3@L002-1	15,308,060	6,801 (0.04)	15,301,259 (99.96)
B6@L001-1	12,992,270	5,790 (0.04)	12,986,480 (99.96)
B6@L002-1	13,007,590	5,663 (0.04)	13,001,927 (99.96)
B8@L001-1	13,189,729	5,524 (0.04)	13,184,205 (99.96)
B8@L002-1	13,346,331	5,522 (0.04)	13,340,809 (99.96)
E3@L001-1	12,656,316	5,452 (0.04)	12,650,864 (99.96)
E3@L002-1	12,757,134	5,557 (0.04)	12,751,577 (99.96)
E4@L001-1	13,816,244	5,930 (0.04)	13,810,314 (99.96)
E4@L002-1	13,891,857	6,017 (0.04)	13,885,840 (99.96)
E6@L001-1	15,819,911	6,967 (0.04)	15,812,944 (99.96)
E6@L002-1	15,792,557	7,043 (0.04)	15,785,514 (99.96)
F1@L001-1	15,449,504	6,773 (0.04)	15,442,731 (99.96)
F1@L002-1	15,566,824	7,026 (0.05)	15,559,798 (99.95)
F4@L001-1	17,550,169	7,846 (0.04)	17,542,323 (99.96)
F4@L002-1	17,380,988	7,688 (0.04)	17,373,300 (99.96)
F7@L001-1	11,773,259	4,988 (0.04)	11,768,271 (99.96)
F7@L002-1	11,925,856	5,240 (0.04)	11,920,616 (99.96)
G1@L001-1	13,465,203	5,815 (0.04)	13,459,388 (99.96)
G1@L002-1	13,557,053	5,906 (0.04)	13,551,147 (99.96)
G2@L001-1	11,657,994	5,501 (0.05)	11,652,493 (99.95)
G2@L002-1	11,755,938	5,678 (0.05)	11,750,260 (99.95)
G5@L001-1	15,879,218	6,796 (0.04)	15,872,422 (99.96)
G5@L002-1	15,930,800	6,820 (0.04)	15,923,980 (99.96)
H4@L001-1	12,888,586	5,883 (0.05)	12,882,703 (99.95)
H4@L002-1	12,955,061	6,031 (0.05)	12,949,030 (99.95)
H6@L001-1	14,824,111	6,628 (0.04)	14,817,483 (99.96)
H6@L002-1	14,866,467	6,666 (0.04)	14,859,801 (99.96)
H7@L001-1	13,425,287	6,101 (0.05)	13,419,186 (99.95)
H7@L002-1	13,553,145	6,229 (0.05)	13,546,916 (99.95)
M4@L001-1	12,945,839	6,743 (0.05)	12,939,096 (99.95)
M4@L002-1	12,950,367	6,700 (0.05)	12,943,667 (99.95)
M5@L001-1	10,603,035	4,836 (0.05)	10,598,199 (99.95)
M5@L002-1	10,758,068	4,894 (0.05)	10,753,174 (99.95)
M6@L001-1	13,419,214	6,422 (0.05)	13,412,792 (99.95)
M6@L002-1	13,449,565	6,458 (0.05)	13,443,107 (99.95)
N1@L001-1	11,078,420	5,085 (0.05)	11,073,335 (99.95)
N1@L002-1	11,249,487	5,297 (0.05)	11,244,190 (99.95)
N5@L001-1	14,456,903	6,396 (0.04)	14,450,507 (99.96)
N5@L002-1	14,424,715	6,589 (0.05)	14,418,126 (99.95)
N6@L001-1	10,717,207	5,383 (0.05)	10,711,824 (99.95)
N6@L002-1	10,917,135	5,477 (0.05)	10,911,658 (99.95)

4. *Removal of contaminants:* Using bowtie 2.3.3.1, we identify reads that align to phix_illumina and remove them.

SAMPLE	# PAIRED READS	# FILTERED READ (%)	# REMAINING READS (%)
A2@L001-1	18,287,788	5 (0)	18,287,783 (100)
A2@L002-1	18,251,609	1 (0)	18,251,608 (100)
A3@L001-1	12,292,758	99 (0)	12,292,659 (100)
A3@L002-1	12,457,705	103 (0)	12,457,602 (100)
A7@L001-1	15,846,810	0 (0)	15,846,810 (100)
A7@L002-1	15,815,230	0 (0)	15,815,230 (100)
B3@L001-1	15,204,507	10 (0)	15,204,497 (100)
B3@L002-1	15,301,259	8 (0)	15,301,251 (100)
B6@L001-1	12,986,480	3 (0)	12,986,477 (100)
B6@L002-1	13,001,927	4 (0)	13,001,923 (100)
B8@L001-1	13,184,205	4 (0)	13,184,201 (100)
B8@L002-1	13,340,809	0 (0)	13,340,809 (100)
E3@L001-1	12,650,864	3 (0)	12,650,861 (100)
E3@L002-1	12,751,577	7 (0)	12,751,570 (100)
E4@L001-1	13,810,314	2 (0)	13,810,312 (100)
E4@L002-1	13,885,840	1 (0)	13,885,839 (100)
E6@L001-1	15,812,944	55 (0)	15,812,889 (100)
E6@L002-1	15,785,514	40 (0)	15,785,474 (100)
F1@L001-1	15,442,731	2 (0)	15,442,729 (100)
F1@L002-1	15,559,798	2 (0)	15,559,796 (100)
F4@L001-1	17,542,323	7 (0)	17,542,316 (100)
F4@L002-1	17,373,300	5 (0)	17,373,295 (100)
F7@L001-1	11,768,271	5 (0)	11,768,266 (100)
F7@L002-1	11,920,616	4 (0)	11,920,612 (100)
G1@L001-1	13,459,388	10 (0)	13,459,378 (100)
G1@L002-1	13,551,147	6 (0)	13,551,141 (100)
G2@L001-1	11,652,493	2 (0)	11,652,491 (100)
G2@L002-1	11,750,260	1 (0)	11,750,259 (100)
G5@L001-1	15,872,422	3 (0)	15,872,419 (100)
G5@L002-1	15,923,980	2 (0)	15,923,978 (100)
H4@L001-1	12,882,703	4 (0)	12,882,699 (100)
H4@L002-1	12,949,030	5 (0)	12,949,025 (100)
H6@L001-1	14,817,483	7 (0)	14,817,476 (100)
H6@L002-1	14,859,801	12 (0)	14,859,789 (100)
H7@L001-1	13,419,186	2 (0)	13,419,184 (100)
H7@L002-1	13,546,916	7 (0)	13,546,909 (100)
M4@L001-1	12,939,096	14 (0)	12,939,082 (100)
M4@L002-1	12,943,667	16 (0)	12,943,651 (100)
M5@L001-1	10,598,199	0 (0)	10,598,199 (100)
M5@L002-1	10,753,174	0 (0)	10,753,174 (100)
M6@L001-1	13,412,792	3 (0)	13,412,789 (100)
M6@L002-1	13,443,107	2 (0)	13,443,105 (100)
N1@L001-1	11,073,335	14 (0)	11,073,321 (100)
N1@L002-1	11,244,190	6 (0)	11,244,184 (100)
N5@L001-1	14,450,507	0 (0)	14,450,507 (100)
N5@L002-1	14,418,126	1 (0)	14,418,125 (100)
N6@L001-1	10,711,824	7 (0)	10,711,817 (100)
N6@L002-1	10,911,658	4 (0)	10,911,654 (100)

4.2 Results

Following statistics and figures can be used to verify whether data preprocessing could successfully remove technical artifacts and enhance data quality.

Yield Statistics: The table provides details about the number of bases, number of reads and read length.

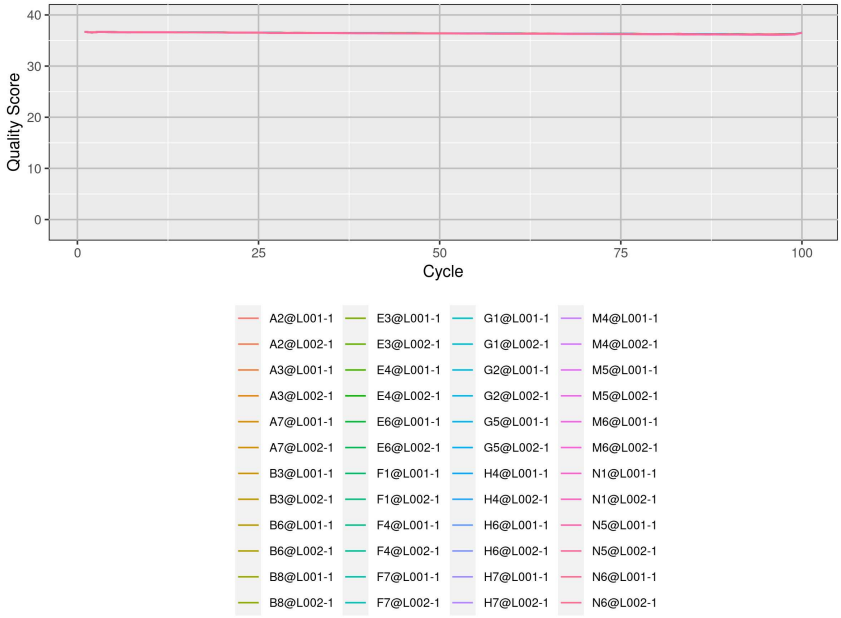
In case that the fraction of reads of a sample deviates with a factor 3 from the expected fraction (i.e., 2.08 %), the data of that sample are indicated in red.

	# Bases	*# Fragments*	*Read Length*	*% Fragments*
A2@L001	1,822,789,190	18,287,783	99.67	2.76
A2@L002	1,819,211,551	18,251,608	99.67	2.76
A3@L001	1,224,748,548	12,292,659	99.63	1.86
A3@L002	1,241,185,620	12,457,602	99.63	1.88
A7@L001	1,578,525,978	15,846,810	99.61	2.39
A7@L002	1,575,319,640	15,815,230	99.61	2.39
B3@L001	1,514,380,422	15,204,497	99.6	2.3
B3@L002	1,524,000,114	15,301,251	99.6	2.31
B6@L001	1,293,420,492	12,986,477	99.6	1.96
B6@L002	1,294,895,656	13,001,923	99.59	1.96
B8@L001	1,313,177,801	13,184,201	99.6	1.99
B8@L002	1,328,761,224	13,340,809	99.6	2.02
E3@L001	1,260,835,654	12,650,861	99.66	1.91
E3@L002	1,270,851,468	12,751,570	99.66	1.93
E4@L001	1,377,113,122	13,810,312	99.72	2.09
E4@L002	1,384,655,206	13,885,839	99.72	2.1
E6@L001	1,575,594,246	15,812,889	99.64	2.39
E6@L002	1,572,865,244	15,785,474	99.64	2.39
F1@L001	1,537,768,415	15,442,729	99.58	2.33
F1@L002	1,549,402,348	15,559,796	99.58	2.35
F4@L001	1,746,575,627	17,542,316	99.56	2.65
F4@L002	1,729,724,529	17,373,295	99.56	2.62
F7@L001	1,171,188,771	11,768,266	99.52	1.78
F7@L002	1,186,353,726	11,920,612	99.52	1.8
G1@L001	1,340,697,472	13,459,378	99.61	2.03
G1@L002	1,349,823,785	13,551,141	99.61	2.05
G2@L001	1,158,219,156	11,652,491	99.4	1.76
G2@L002	1,167,903,593	11,750,259	99.39	1.78
G5@L001	1,582,003,879	15,872,419	99.67	2.4
G5@L002	1,587,114,137	15,923,978	99.67	2.41
H4@L001	1,282,993,456	12,882,699	99.59	1.95
H4@L002	1,289,577,441	12,949,025	99.59	1.96
H6@L001	1,475,650,837	14,817,476	99.59	2.24
H6@L002	1,479,842,220	14,859,789	99.59	2.25
H7@L001	1,334,709,738	13,419,184	99.46	2.03
H7@L002	1,347,381,416	13,546,909	99.46	2.05
M4@L001	1,283,909,177	12,939,082	99.23	1.95
M4@L002	1,284,332,261	12,943,651	99.22	1.96
M5@L001	1,048,889,115	10,598,199	98.97	1.6
M5@L002	1,064,202,984	10,753,174	98.97	1.62
M6@L001	1,327,959,968	13,412,789	99.01	2.03
M6@L002	1,330,952,666	13,443,105	99.01	2.03
N1@L001	1,099,291,956	11,073,321	99.27	1.67

Continued on next page

	# Bases	# Fragments	Read Length	% Fragments
N1@L002	1,116,255,481	11,244,184	99.27	1.7
N5@L001	1,437,903,866	14,450,507	99.51	2.18
N5@L002	1,434,676,451	14,418,125	99.51	2.18
N6@L001	1,064,788,965	10,711,817	99.4	1.62
N6@L002	1,084,622,087	10,911,654	99.4	1.65
In total	65,867,046,699	661,859,165		

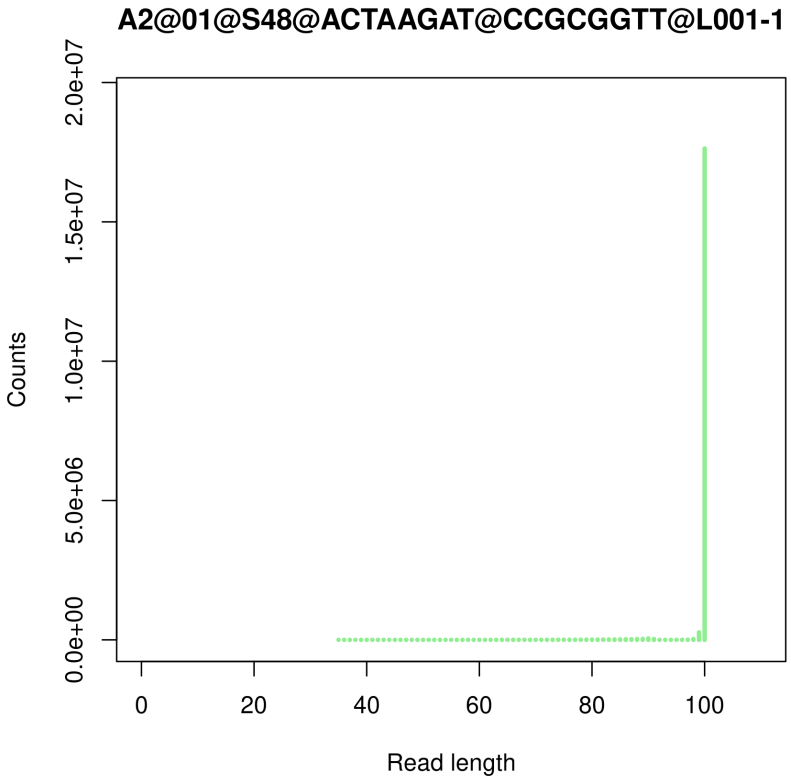
Accuracy Statistics: The following figures show for each sample the average base quality per cycle, as calculated by the ShortRead 1.40.0 package from Bioconductor (http://www.bioconductor.org) [9]. Reported quality scores typically decline with cycle, in an accelerating manner.



Following figures show the distributions of the average read quality, as calculated by the ShortRead 1.40.0 package from Bioconductor (http://www.bioconductor.org) [9]. The figures provide insight whether good quality bases are grouped together in a few reads or are spread over the entire library. Samples with consistently good quality reads have unimodal, strong peaks near the right of the panel.

Forward reads

Read Length Distribution: The figures show for each sample the distribution of read length. One figure is given here, you can find the others in the annex.



4.3 Data transfer

The preprocessed reads are stored in 48 compressed fastq-file(s) with extension '.pp.fastq.gz'.

5 Mapping of RNA-seq data

In this section we align the preprocessed reads to the reference genome of Linum_usitatissimumJGI_152022v1.0 (LinumusitatissimumJGI152022v10).

5.1 Materials and methods

The alignment files are generated in 3 steps.

1. *Read mapping:* The reads are aligned to the reference genome with STAR 2.5.2b ([4]). As parameter options we use: `--outSAMprimaryFlag OneBestScore --twopassMode Basic --alignIntronMin 50 --alignIntronMax 50000 --outSAMtype BAM SortedByCoordinate`. Unmapped reads are kept in a separate file.
2. *Quality filtering:* With samtools 1.5 we remove reads from the alignment that are non-primary mappings or have a mapping quality ≤ 20 [6].
3. *Sorting and indexing:* With samtools 1.5 we sort the reads from the alignment according to the chromosomes and index the resulting bam-files.

Unmapped reads and reads from the final alignments are stored in separate fastq-files (see section about data transfer).

5.2 Results

5.2.1 Mapping statistics

Following table provides an overview of the mapping efficiency. It shows the number of mapped reads prior to and after the filtering of the mapped reads. Take note that the sequencing coverage of the reference genome is **71.59%**.

Sample	*# Preprocessed reads*	*# unmapped reads (%)*	*# Mapped reads (%)*	*# Mapped filtered reads (%)*
A2@L001	18,287,783	189,071 (1.03)	18,098,712 (98.97)	15,982,168 (87.39)
A2@L002	18,251,608	185,782 (1.02)	18,065,826 (98.98)	15,952,948 (87.41)
A3@L001	12,292,659	117,840 (0.96)	12,174,819 (99.04)	10,750,694 (87.46)
A3@L002	12,457,602	117,449 (0.94)	12,340,153 (99.06)	10,891,170 (87.43)
A7@L001	15,846,810	139,042 (0.88)	15,707,768 (99.12)	13,879,640 (87.59)
A7@L002	15,815,230	138,001 (0.87)	15,677,229 (99.13)	13,849,229 (87.57)
B3@L001	15,204,497	139,578 (0.92)	15,064,919 (99.08)	13,175,948 (86.66)
B3@L002	15,301,251	139,132 (0.91)	15,162,119 (99.09)	13,259,701 (86.66)
B6@L001	12,986,477	112,867 (0.87)	12,873,610 (99.13)	11,344,549 (87.36)
B6@L002	13,001,923	111,206 (0.86)	12,890,717 (99.14)	11,359,865 (87.37)
B8@L001	13,184,201	118,064 (0.90)	13,066,137 (99.10)	11,535,961 (87.50)
B8@L002	13,340,809	118,085 (0.89)	13,222,724 (99.11)	11,673,857 (87.50)
E3@L001	12,650,861	103,734 (0.82)	12,547,127 (99.18)	11,088,158 (87.65)
E3@L002	12,751,570	102,750 (0.81)	12,648,820 (99.19)	11,180,327 (87.68)
E4@L001	13,810,312	118,233 (0.86)	13,692,079 (99.14)	12,109,693 (87.69)
E4@L002	13,885,839	116,782 (0.84)	13,769,057 (99.16)	12,178,992 (87.71)
E6@L001	15,812,889	131,307 (0.83)	15,681,582 (99.17)	13,815,718 (87.37)
E6@L002	15,785,474	130,434 (0.83)	15,655,040 (99.17)	13,793,190 (87.38)
F1@L001	15,442,729	138,871 (0.90)	15,303,858 (99.10)	13,525,239 (87.58)

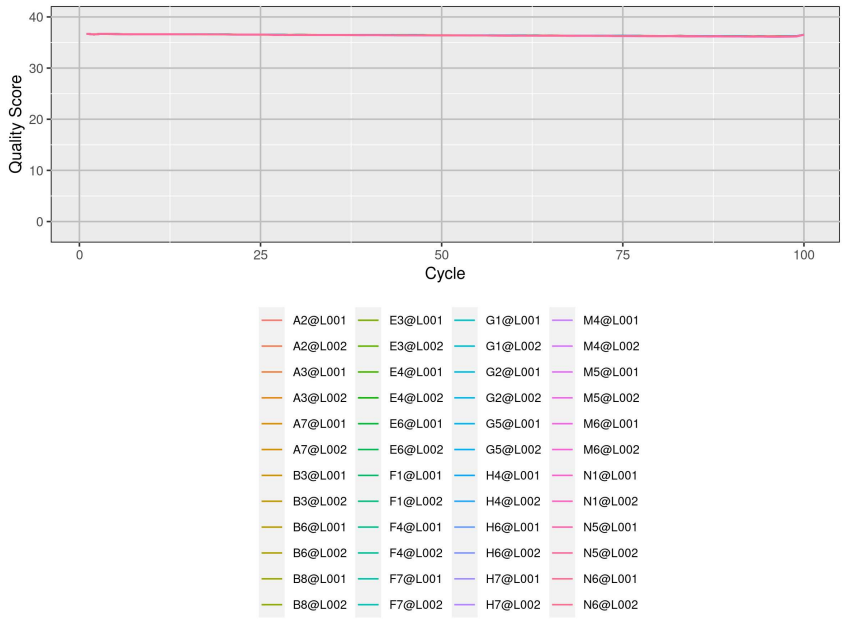
Continued on next page

Sample	# Preprocessed reads	# unmapped reads (%)	# Mapped reads (%)	# Mapped filtered reads (%)
F1@L002	15,559,796	139,424 (0.90)	15,420,372 (99.10)	13,634,596 (87.63)
F4@L001	17,542,316	155,575 (0.89)	17,386,741 (99.11)	15,335,936 (87.42)
F4@L002	17,373,295	152,152 (0.88)	17,221,143 (99.12)	15,196,317 (87.47)
F7@L001	11,768,266	99,498 (0.85)	11,668,768 (99.15)	10,318,241 (87.68)
F7@L002	11,920,612	99,304 (0.83)	11,821,308 (99.17)	10,450,467 (87.67)
G1@L001	13,459,378	136,948 (1.02)	13,322,430 (98.98)	11,799,067 (87.66)
G1@L002	13,551,141	137,323 (1.01)	13,413,818 (98.99)	11,880,533 (87.67)
G2@L001	11,652,491	97,153 (0.83)	11,555,338 (99.17)	10,142,027 (87.04)
G2@L002	11,750,259	96,272 (0.82)	11,653,987 (99.18)	10,223,779 (87.01)
G5@L001	15,872,419	169,155 (1.07)	15,703,264 (98.93)	13,907,950 (87.62)
G5@L002	15,923,978	166,191 (1.04)	15,757,787 (98.96)	13,953,906 (87.63)
H4@L001	12,882,699	98,793 (0.77)	12,783,906 (99.23)	11,300,097 (87.72)
H4@L002	12,949,025	97,750 (0.75)	12,851,275 (99.25)	11,361,911 (87.74)
H6@L001	14,817,476	119,635 (0.81)	14,697,841 (99.19)	12,966,373 (87.51)
H6@L002	14,859,789	118,217 (0.80)	14,741,572 (99.20)	13,002,294 (87.50)
H7@L001	13,419,184	112,541 (0.84)	13,306,643 (99.16)	11,711,623 (87.28)
H7@L002	13,546,909	111,994 (0.83)	13,434,915 (99.17)	11,825,453 (87.29)
M4@L001	12,939,082	93,518 (0.72)	12,845,564 (99.28)	11,408,213 (88.17)
M4@L002	12,943,651	92,412 (0.71)	12,851,239 (99.29)	11,416,730 (88.20)
M5@L001	10,598,199	68,632 (0.65)	10,529,567 (99.35)	9,353,756 (88.26)
M5@L002	10,753,174	69,127 (0.64)	10,684,047 (99.36)	9,490,716 (88.26)
M6@L001	13,412,789	86,121 (0.64)	13,326,668 (99.36)	11,891,585 (88.66)
M6@L002	13,443,105	85,137 (0.63)	13,357,968 (99.37)	11,920,907 (88.68)
N1@L001	11,073,321	79,405 (0.72)	10,993,916 (99.28)	9,764,520 (88.18)
N1@L002	11,244,184	78,826 (0.70)	11,165,358 (99.30)	9,914,603 (88.18)
N5@L001	14,450,507	112,860 (0.78)	14,337,647 (99.22)	12,766,515 (88.35)
N5@L002	14,418,125	111,674 (0.77)	14,306,451 (99.23)	12,741,544 (88.37)
N6@L001	10,711,817	84,630 (0.79)	10,627,187 (99.21)	9,346,474 (87.25)
N6@L002	10,911,654	85,207 (0.78)	10,826,447 (99.22)	9,520,099 (87.25)

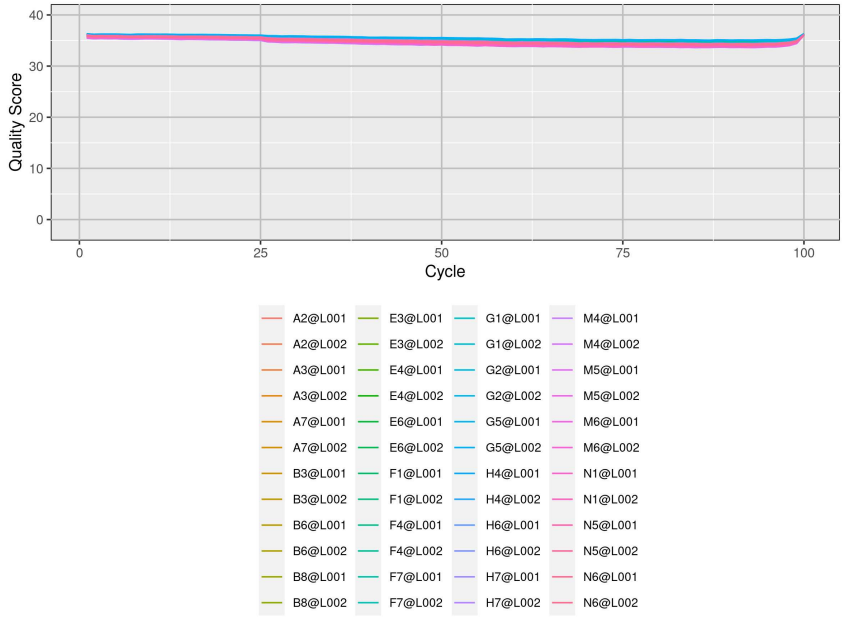
5.2.2 Quality difference between mapped and unmapped reads

Accuracy Statistics: The following figures show for each sample the average base quality per cycle, as calculated by the ShortRead 1.40.0 package from Bioconductor (http://www.bioconductor.org, [9]).

Forward reads of mapped fragments



Forward reads of unmapped fragments



Library Complexity: The table below provides details of the percentage of base calls per sample and per read direction. Base frequencies should accurately reflect the expected frequencies of the sequenced regions.

Mapped fragments

	Forward read				
	A	*C*	*G*	*T*	*N*
A2@L001	25.83	25.31	23.19	25.67	0
A2@L002	25.82	25.32	23.2	25.66	0

Continued on next page

	Forward read				
	A	*C*	*G*	*T*	*N*
A3@L001	26.2	25.24	22.79	25.77	0
A3@L002	26.18	25.26	22.79	25.77	0
A7@L001	26.07	25.25	22.93	25.74	0
A7@L002	26.06	25.26	22.95	25.74	0
B3@L001	26.06	25.25	22.9	25.79	0
B3@L002	26.05	25.26	22.9	25.79	0
B6@L001	26.37	25.18	22.65	25.81	0
B6@L002	26.36	25.18	22.66	25.81	0
B8@L001	26.3	25.18	22.64	25.88	0
B8@L002	26.28	25.19	22.65	25.88	0
E3@L001	26.29	25.2	22.74	25.76	0
E3@L002	26.27	25.22	22.74	25.77	0
E4@L001	26	25.39	22.93	25.68	0
E4@L002	26	25.4	22.93	25.67	0
E6@L001	26.16	25.31	22.8	25.73	0
E6@L002	26.14	25.32	22.82	25.73	0
F1@L001	26.17	25.2	22.82	25.82	0
F1@L002	26.15	25.21	22.83	25.81	0
F4@L001	26.33	25.02	22.74	25.91	0
F4@L002	26.31	25.03	22.75	25.91	0
F7@L001	26.36	25.12	22.61	25.91	0
F7@L002	26.35	25.13	22.63	25.9	0
G1@L001	25.99	25.09	23.11	25.82	0
G1@L002	25.97	25.1	23.1	25.82	0
G2@L001	26.79	24.97	22.27	25.97	0
G2@L002	26.78	24.99	22.26	25.97	0
G5@L001	25.96	25.31	22.98	25.75	0
G5@L002	25.95	25.32	22.98	25.75	0
H4@L001	26.35	25.23	22.64	25.78	0
H4@L002	26.34	25.24	22.64	25.77	0
H6@L001	26.42	25.24	22.52	25.82	0
H6@L002	26.4	25.26	22.52	25.82	0
H7@L001	26.63	25.1	22.37	25.9	0
H7@L002	26.61	25.11	22.38	25.9	0
M4@L001	27.53	24.72	21.22	26.53	0
M4@L002	27.52	24.74	21.23	26.52	0
M5@L001	27.28	24.76	21.39	26.58	0
M5@L002	27.26	24.77	21.4	26.57	0
M6@L001	27.52	24.6	21.18	26.7	0
M6@L002	27.5	24.62	21.18	26.7	0
N1@L001	27.26	24.88	21.44	26.42	0
N1@L002	27.24	24.89	21.45	26.42	0
N5@L001	26.85	24.99	21.8	26.36	0
N5@L002	26.83	25	21.81	26.36	0
N6@L001	27.56	24.9	21.22	26.31	0
N6@L002	27.55	24.91	21.23	26.31	0

Unmapped fragments

	Forward read				
	A	*C*	*G*	*T*	*N*
A2@L001	28.71	24.29	23.64	23.36	0

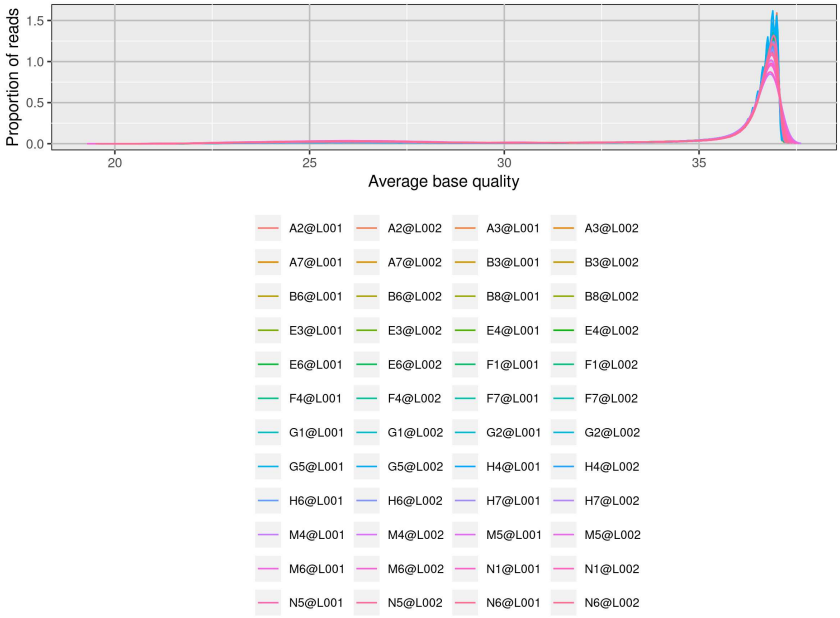
Continued on next page

	Forward read				
	A	*C*	*G*	*T*	*N*
A2@L002	28.61	24.35	23.7	23.34	0
A3@L001	29.2	24.08	22.79	23.92	0
A3@L002	29.07	24.12	22.92	23.88	0
A7@L001	28.67	24.14	22.97	24.23	0
A7@L002	28.61	24.15	22.99	24.25	0
B3@L001	28.95	24.29	22.28	24.47	0
B3@L002	28.81	24.37	22.34	24.47	0
B6@L001	29.13	23.98	22.18	24.71	0
B6@L002	28.96	24.06	22.25	24.73	0
B8@L001	28.9	23.98	22.43	24.7	0
B8@L002	28.77	24.04	22.54	24.65	0
E3@L001	29.29	23.69	22.33	24.69	0
E3@L002	29.22	23.68	22.41	24.68	0
E4@L001	29.03	24.1	22.55	24.32	0
E4@L002	28.9	24.09	22.71	24.3	0
E6@L001	29.22	24.09	22.41	24.28	0
E6@L002	29.05	24.17	22.44	24.33	0
F1@L001	29.15	24.03	22.52	24.3	0
F1@L002	28.98	24.09	22.63	24.3	0
F4@L001	29.24	23.97	22.19	24.61	0
F4@L002	29.09	24	22.31	24.6	0
F7@L001	29.02	24.07	22.5	24.42	0
F7@L002	28.87	24.15	22.59	24.39	0
G1@L001	28.6	24.32	23.51	23.58	0
G1@L002	28.49	24.32	23.61	23.58	0
G2@L001	29.73	24.05	21.53	24.69	0
G2@L002	29.6	24.09	21.64	24.68	0
G5@L001	28.51	24.49	23.75	23.26	0
G5@L002	28.44	24.54	23.8	23.22	0
H4@L001	29.64	23.83	21.84	24.69	0
H4@L002	29.42	23.93	21.93	24.72	0
H6@L001	29.6	23.82	21.91	24.67	0
H6@L002	29.43	23.9	22	24.67	0
H7@L001	29.23	24.13	22.07	24.56	0
H7@L002	29.08	24.18	22.17	24.57	0
M4@L001	30.56	24.14	20.5	24.8	0
M4@L002	30.3	24.19	20.64	24.87	0
M5@L001	29.13	24.35	21.35	25.16	0
M5@L002	28.96	24.47	21.44	25.13	0
M6@L001	30.09	23.93	20.8	25.18	0
M6@L002	29.83	23.99	20.94	25.24	0
N1@L001	29.94	24.34	20.93	24.8	0
N1@L002	29.77	24.3	21.11	24.83	0
N5@L001	29.22	24.24	21.43	25.12	0
N5@L002	29.02	24.28	21.53	25.17	0
N6@L001	29.51	24.64	21.04	24.81	0
N6@L002	29.46	24.65	21.12	24.77	0

Following figures show the distributions of the average read quality, as calculated by the ShortRead 1.40.0 package from Bioconductor (http://www.bioconductor.org, [9]). The figures provide insight whether good quality bases are grouped together in a few reads or are spread over the entire library. Samples with consistently good quality reads have unimodal, strong peaks near the right of the panel.

Forward reads of mapped fragments

Forward reads of unmapped fragments



5.3 Data transfer

- The reference data (fasta and gff) are stored in the compressed archive exp3949-Reference-LinumusitatissimumJGI152022v10.zip. The indexed fasta-file (.fasta.fai) is only used during visualization by genome browsers such as IGV.
- The alignments are merged per sample and are stored in 24 bam-files with extension '.mgs.bam' and 24 indexed bam-files with extensions '.mgs.bam.bai'. The indexed bam-files are only used during visualization by genome browsers such as IGV.
- The unmapped reads are stored in fastq-files with extensions '.unmp.fastq'.

6 Summarization of expression levels

In this section we count reads that overlap genes and as such derive expression levels.

6.1 Materials and methods

The expression levels are computed as follows:

1. *Counting per gene:* We count the number of reads in the alignments that overlap with the gene features, using featureCounts 1.5.3 [7]. As parameters we take: `-Q 0 -s 2 -t exon -g gene_id`. Some reads are not counted, since they can be attributed to more than one gene (ambiguous) or cannot be attributed to any gene (no feature). Compared to stranded libraries, the number of ambiguous counts are generally higher when unstranded libraries are used. The table below provides an overview of the number of non-counted and counted reads.

Sample	# mapped reads	# NoFeatures (%)	# Ambiguity (%)	# Counted (%)
A2	31935116	6443036 (20.18)	83724 (0.26)	25408356 (79.56)
A3	21641864	4807531 (22.21)	55548 (0.26)	16778785 (77.53)
A7	27728869	6047030 (21.81)	73810 (0.27)	21608029 (77.93)
B3	26435649	6031116 (22.81)	67881 (0.26)	20336652 (76.93)
B6	22704414	5391640 (23.75)	58161 (0.26)	17254613 (76)
B8	23209818	5370417 (23.14)	60703 (0.26)	17778698 (76.6)
E3	22268485	4904311 (22.02)	57744 (0.26)	17306430 (77.72)
E4	24288685	5146847 (21.19)	64041 (0.26)	19077797 (78.55)
E6	27608908	6232864 (22.58)	74324 (0.27)	21301720 (77.16)
F1	27159835	6093501 (22.44)	72197 (0.27)	20994137 (77.3)
F4	30532253	7159410 (23.45)	79249 (0.26)	23293594 (76.29)
F7	20768708	4908738 (23.64)	55431 (0.27)	15804539 (76.1)
G1	23679600	4977419 (21.02)	62544 (0.26)	18639637 (78.72)
G2	20365806	5522661 (27.12)	54171 (0.27)	14788974 (72.62)
G5	27861856	5859030 (21.03)	75250 (0.27)	21927576 (78.7)
H4	22662008	5355700 (23.63)	58010 (0.26)	17248298 (76.11)
H6	25968667	6283165 (24.2)	68957 (0.27)	19616545 (75.54)
H7	23537076	6092511 (25.88)	60520 (0.26)	17384045 (73.86)
M4	22824943	7251965 (31.77)	51505 (0.23)	15521473 (68)
M5	18844472	5705693 (30.28)	44386 (0.24)	13094393 (69.49)
M6	23812492	7417211 (31.15)	54400 (0.23)	16340881 (68.62)
N1	19679123	5880749 (29.88)	45005 (0.23)	13753369 (69.89)
N5	25508059	6930908 (27.17)	60071 (0.24)	18517080 (72.59)
N6	18866573	6201367 (32.87)	41112 (0.22)	12624094 (66.91)

2. *Merging with gene annotation:* The data of raw counts are merged with the reference gene annotation that was constructed in the previous section. The counts are available as 'raw counts' in an Excel sheet.

3. *Filtering genes:* We further remove 12348 genes for which all samples have less than 1 counts-per-million (absent genes) [12]. As such, we continue with 31123 genes.

4. *Within-sample normalization*: A gene's GC-content may have a strong sample-specific effect when counting reads from RNAseq-experiments [10]. We correct per sample for GC-content using full quantile normalization on bins of GC-content with the *EDASeq* package from Bioconductor [10].

5. *Between-sample normalization*: The main sources of sample-specific variation are the library size and RNA composition [11]. We correct for both factors using full quantile normalization with the *EDASeq* package from Bioconductor.

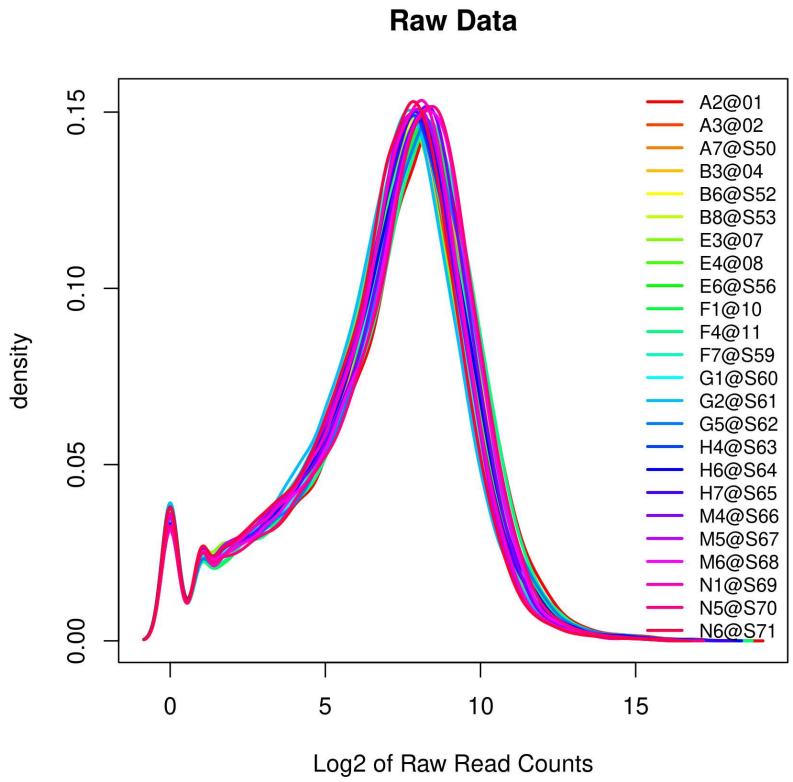
6. *FPKM values*: We divide for each sample the normalized counts by the total number of counts (in millions). Then we divide for each gene the scaled counts by the gene length (in kbp). As such we get the number of Fragments Per Kilobase of gene sequence and per Million fragments of library size.

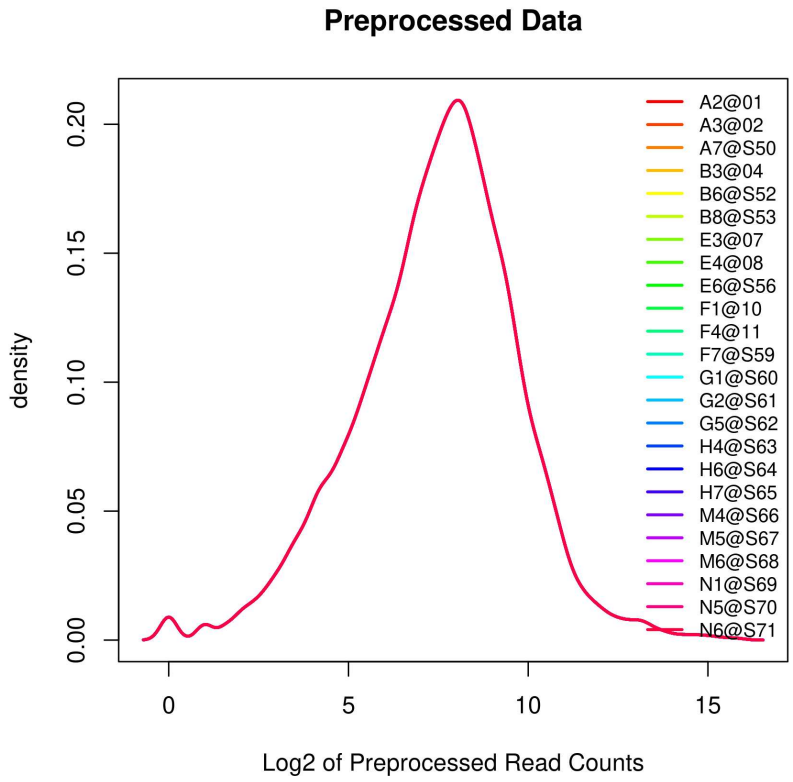
The raw counts, as well as the FPKM values, are available in Excel sheets (see section about data transfer). We should note that raw counts cannot be interpreted as absolute measures of abundance. The counts are for example not normalized for gene length, giving longer genes more chance to have higher counts. FPKM values are more easily interpreted as levels of expression, as they allow comparison of expression values both within as between samples. However, most statistical packages for RNA-seq will start from raw counts as input.

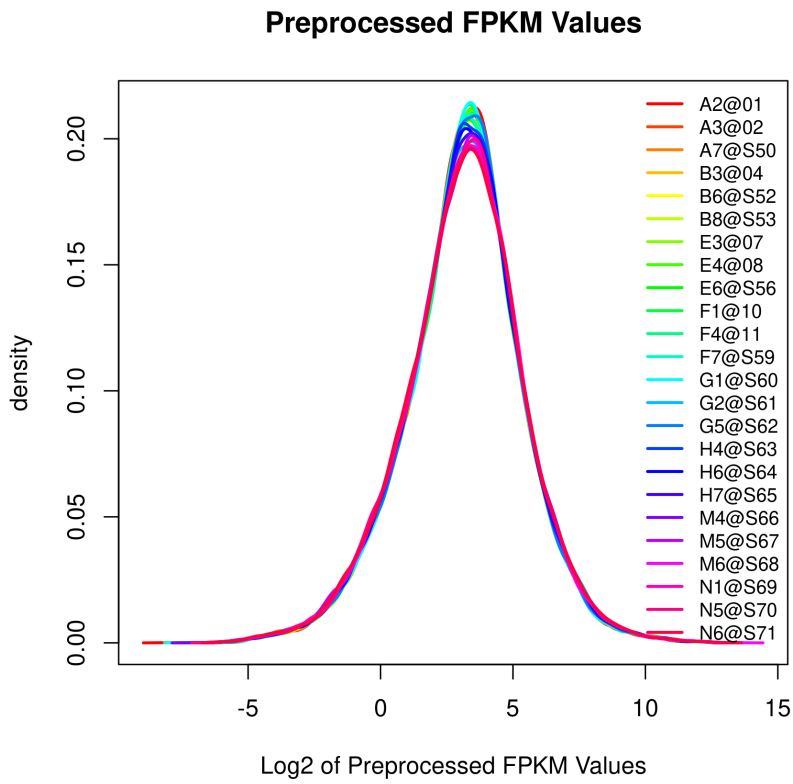
6.2 Results

6.2.1 Signal densities

We plot the densities of the raw counts, of counts after filtering and normalization, and of the final FPKM values.

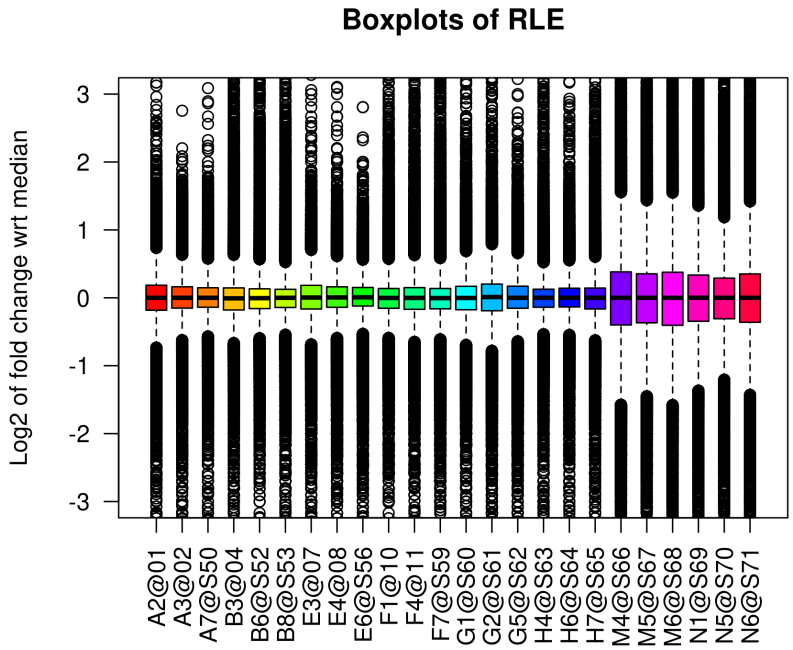



Preprocessed Data
density
0.00
0.05
0.10
0.15
0.20
0
5
10
15
Log2 of Preprocessed Read Counts
A2@01
A3@02
A7@S50
B3@04
B6@S52
B8@S53
E3@07
E4@08
E6@S56
F1@10
F4@11
F7@S59
G1@S60
G2@S61
G5@S62
H4@S63
H6@S64
H7@S65
M4@S66
M5@S67
M6@S68
N1@S69
N5@S70
N6@S71


Preprocessed FPKM Values
density
0.00
0.05
0.10
0.15
0.20
-5
0
5
10
15
Log2 of Preprocessed FPKM Values
A2@01
A3@02
A7@S50
B3@04
B6@S52
B8@S53
E3@07
E4@08
E6@S56
F1@10
F4@11
F7@S59
G1@S60
G2@S61
G5@S62
H4@S63
H6@S64
H7@S65
M4@S66
M5@S67
M6@S68
N1@S69
N5@S70
N6@S71

6.2.2 RLE-plot

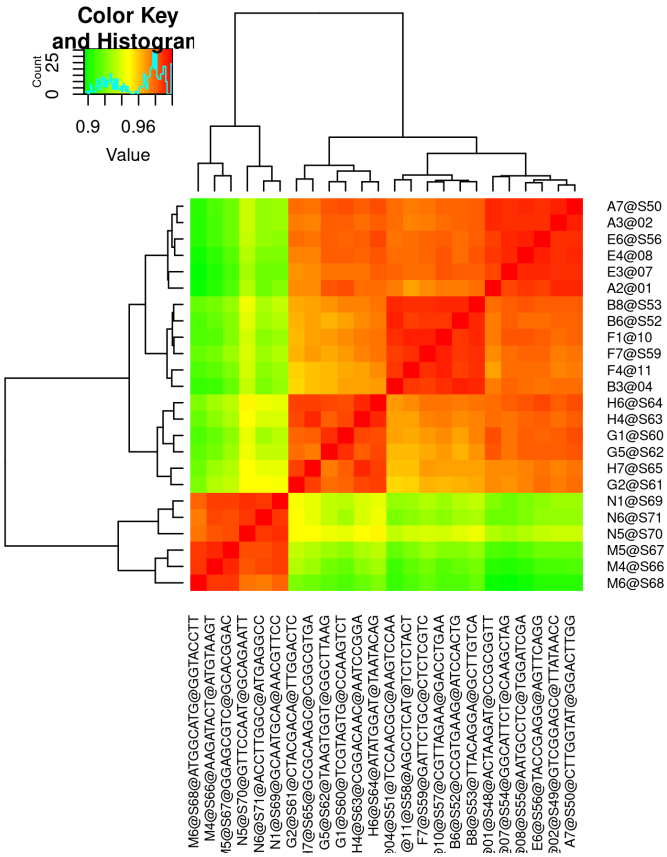
We construct a RLE (Relative Log Expression) plot using log2-scale normalized counts for each gene as expression level. For each gene and each sample, ratios are calculated between the expression level of a gene and the median expression of this gene across all samples of the experiment. For each sample, these relative expression values are displayed as a box plot.



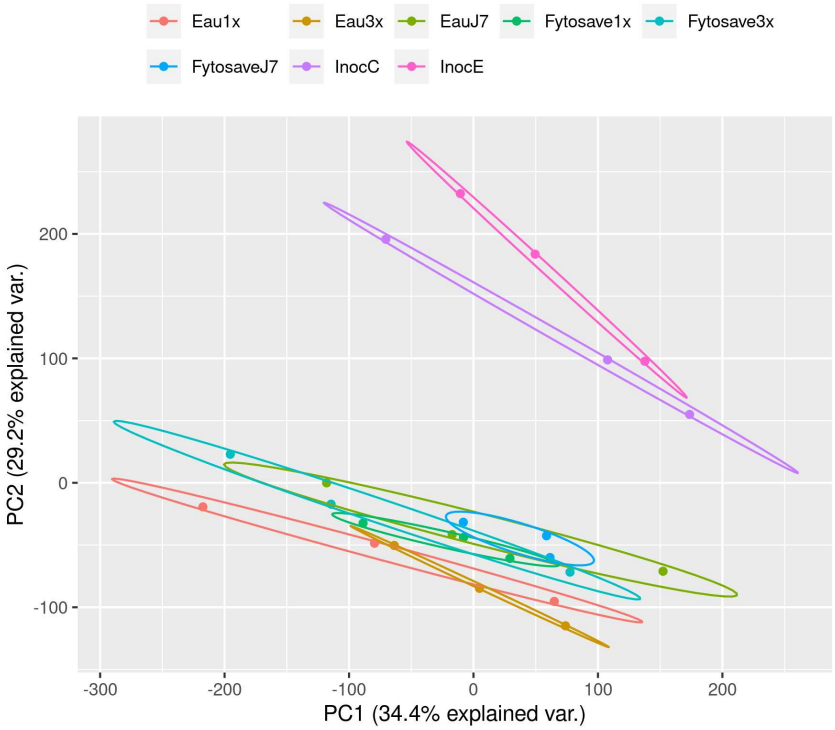
Since it is assumed that in most experiments relatively few genes are differentially expressed, the boxes should be similar in range and be centered close to 0 when the normalization was successful.

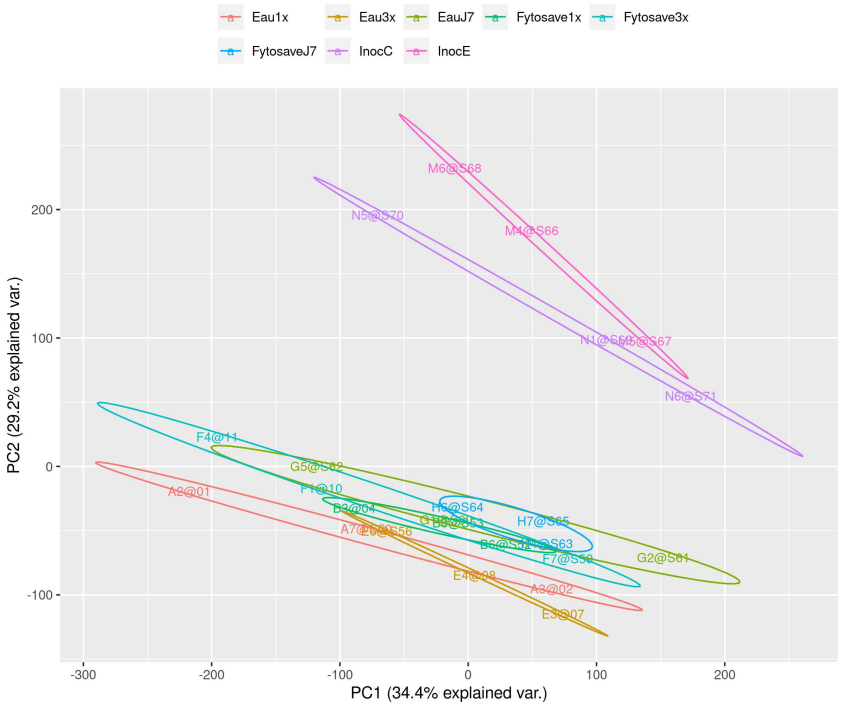
6.2.3 Latent sources of variation

We compute the Spearman-correlation between all samples using the normalized counts as expression values. A heatmap of the correlation matrix is shown below. The columns and rows of the correlation matrix are reorded such that similar samples form clusters. The presence of clusters with a biological interpretation is an indication that normalization could successfully remove technical artefacts.



Using the Principal Component Analysis (PCA), we can project the samples on a two-dimensional graph using the two first principal components that explain the best the biological variation between those samples.
Each point corresponds to a sample plotted by PC1 and PC2. The ellipses are 68% data ellipses for each of the groups of samples in the data. The plot can be used to examine the samples for outliers and other relationships. When normalization successfully removed technical artefacts, the relative distances should be biologically interpretable.





Nucleomics Core, VIB
Herestraat 49
3000 LEUVEN,
Post Box 816
Belgium

VIB NUCLEOMICS CORE

Contact info:
nucleomics@vib.be
+32/16/373128
+32/16/373129
www.nucleomics.be

6.3 Data transfer

The count data are stored in an Excel file *exp3949-RNAseqCounts.xlsx* with two sheets. The first sheet *Raw Counts* contains the counts per gene before filtering and normalization:

- **Gene ID** *(Column A)*: Unique gene identifier
- **Gene Name** *(Column B)*
- **Raw Counts** *(Columns C-Z)*: For each sample, the raw counts are given. Most software packages for statistical RNAseq analysis require raw counts as input data.
- **Additional gene annotation** *(Columns AA-AU)*

The second sheet *FPKM Values after Preprocessing* contains the FPKM values per gene after filtering and normalization:

- **Gene ID** *(Column A)*: Unique gene identifier
- **Gene Name** *(Column B)*
- **FPKM values** *(Columns C-Z)*: For each sample, the FPKM values are given, computed from filtered and normalized counts. These can be used for visual inspection and interpretation of your data. For statistical analysis, please start from raw counts.
- **Additional gene annotation** *(Columns AA-AU)*

7 Statistical comparative analysis

In this section we compare the expression levels of different sample groups and select genes that are differentially expressed.

7.1 Design

	Eau1x	Eau3x	EauJ7	Fytosave1x	Fytosave3x	FytosaveJ7	InocC	InocE
A2	1	0	0	0	0	0	0	0
A3	1	0	0	0	0	0	0	0
A7	1	0	0	0	0	0	0	0
B3	0	0	0	1	0	0	0	0
B6	0	0	0	1	0	0	0	0
B8	0	0	0	1	0	0	0	0
E3	0	1	0	0	0	0	0	0
E4	0	1	0	0	0	0	0	0
E6	0	1	0	0	0	0	0	0
F1	0	0	0	0	1	0	0	0
F4	0	0	0	0	1	0	0	0
F7	0	0	0	0	1	0	0	0
G1	0	0	1	0	0	0	0	0
G2	0	0	1	0	0	0	0	0
G5	0	0	1	0	0	0	0	0
H4	0	0	0	0	0	1	0	0
H6	0	0	0	0	0	1	0	0
H7	0	0	0	0	0	1	0	0
M4	0	0	0	0	0	0	0	1
M5	0	0	0	0	0	0	0	1
M6	0	0	0	0	0	0	0	1
N1	0	0	0	0	0	0	1	0
N5	0	0	0	0	0	0	1	0
N6	0	0	0	0	0	0	1	0

7.2 Materials and methods

Differentially expressed genes are selected in 4 steps:

1. *Statistical modeling:* We specify the design of the experiment as follows:

   $$\log(Count) = Eau1x \times \beta_1 + Eau3x \times \beta_2 + EauJ7 \times \beta_3 + Fytosave1x \times \beta_4 + Fytosave3x \times \beta_5 + FytosaveJ7 \times \beta_6 + InocC$$

   For each gene, the coefficients β are estimated with the edgeR 3.26.8 package of Bioconductor, by fitting a negative binomial generalized linear model (GLM) [12]. To estimate the models, we do not use the normalized counts directly, but work with offsets (see edgeR manual for more details).

2. *Hypothesis testing:* Using the model estimates, we can compute now the contrasts of primary interest:

   (a) Fytosave1x vs Eau1x (Fytosave1XvsEau1X)

   (b) Fytosave3x vs Eau3x (Fytosave3XvsEau3X)

   (c) FytosaveJ7 vs EauJ7 (FytosaveJ7vsEauJ7)

   (d) Eau1x vs Eau3x (Eau1XvsEau3X)

(e) Eau3x vs EauJ7 (Eau3XvsEauJ7)

(f) Fytosave1x vs Fytosave3x (Fytosave1XvsFytosave3X)

(g) Fytosave3x vs FytosaveJ7 (Fytosave3XvsFytosaveJ7)

(h) InocE vs InocC (InocEvsInocC)

(i) FytosaveJ7 vs InocE (FytosaveJ7vsInocE)

(j) EauJ7 vs InocC (EauJ7vsInocC)

They are defined as:

$$\begin{pmatrix} Fytosave1XvsEau1X \\ Fytosave3XvsEau3X \\ FytosaveJ7vsEauJ7 \\ Eau1XvsEau3X \\ Eau3XvsEauJ7 \\ Fytosave1XvsFytosave3X \\ Fytosave3XvsFytosaveJ7 \\ InocEvsInocC \\ FytosaveJ7vsInocE \\ EauJ7vsInocC \end{pmatrix} = \begin{pmatrix} -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \\ \beta_7 \\ \beta_8 \end{pmatrix}$$

With these contrasts, we test the null hypotheses:

$$H_0 : L^T \beta = 0$$

where L is the matrix of the contrasts and β is the vector of the parameter estimates from the model. We test whether these contrasts are significantly deviating from 0 with a GLM likelihood ratio test, as implemented in edgeR 3.26.8.

3. *Correcting for multiple testing:* The resulting p-values are corrected for multiple testing with Benjamini-Hochberg to control the false discovery rate (FDR) [2].

4. *Selecting differentially expressed genes:* Once p-values are computed, we define a p-value-based criterion for selecting genes. We need to make a trade-off between precision and recall, or otherwise false discovery rate and statistical power. One criterion can be to select all genes with a FDR-value less than 0.05. Or we can adopt the criterion that was used during the elaborate MAQC-I study and select genes based on $p < 0.001$ [3]. Both selection procedures can be combined with a cut-off on fold-change, by further constraining gene selection to genes with an absolute $\log_2$-ratio larger than 1.

7.3 Results

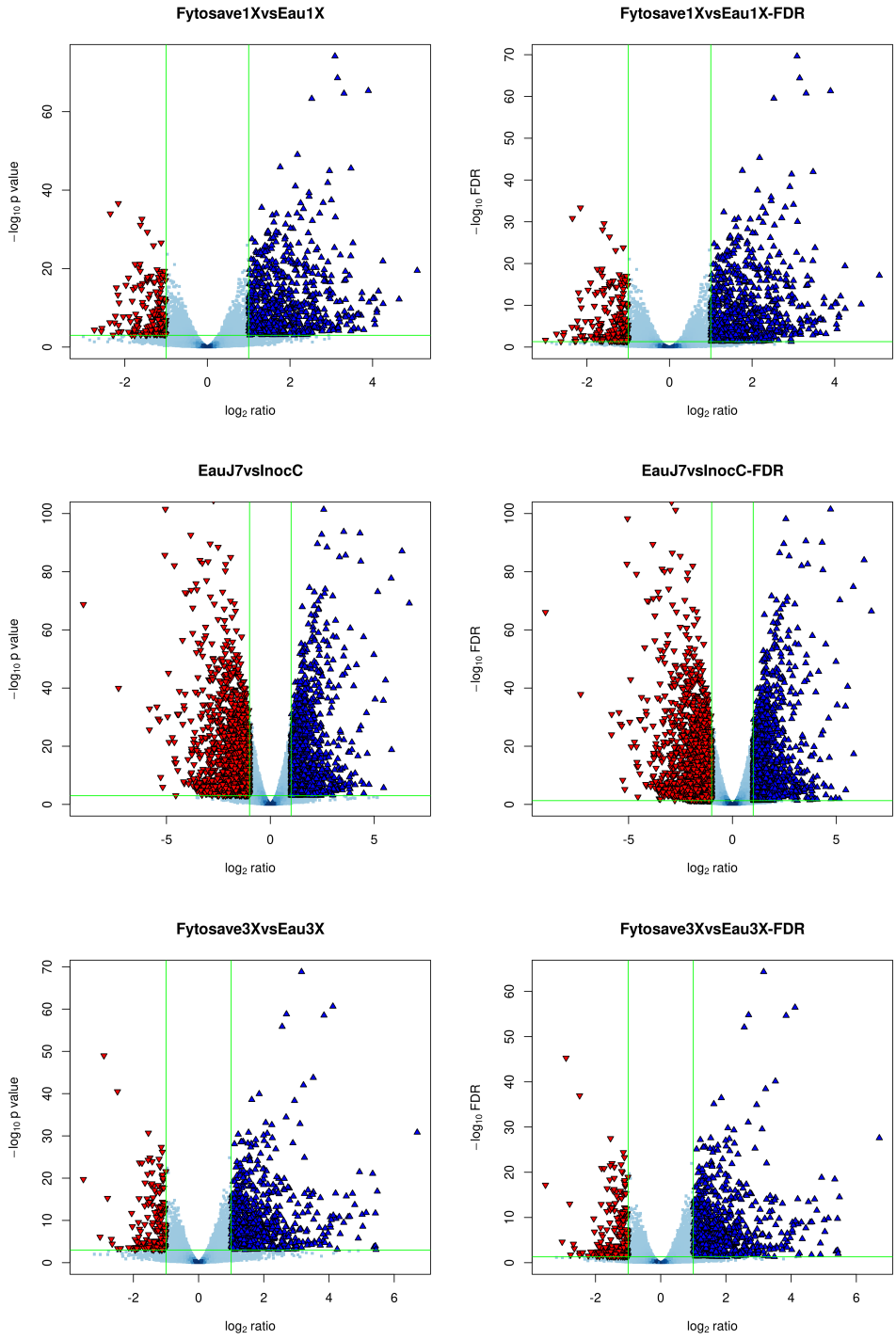
7.3.1 Number of differentially expressed genes

We find the following numbers of differentially expressed genes:

	Uncorr p-val < 0.001		Corr p-val < 0.05	
	log2-ratio < -1	log2-ratio > 1	log2-ratio < -1	log2-ratio > 1
Fytosave1XvsEau1X	176	653	235	723
Fytosave3XvsEau3X	167	646	226	720
FytosaveJ7vsEauJ7	49	186	51	190
Eau1XvsEau3X	39	24	40	24
Eau3XvsEauJ7	404	205	474	256
Fytosave1XvsFytosave3X	58	52	57	52
Fytosave3XvsFytosaveJ7	512	727	638	823
InocEvsInocC	180	276	238	350
FytosaveJ7vsInocE	2140	2434	2419	2669
EauJ7vsInocC	1574	1632	1827	1876

7.3.2 Visualization

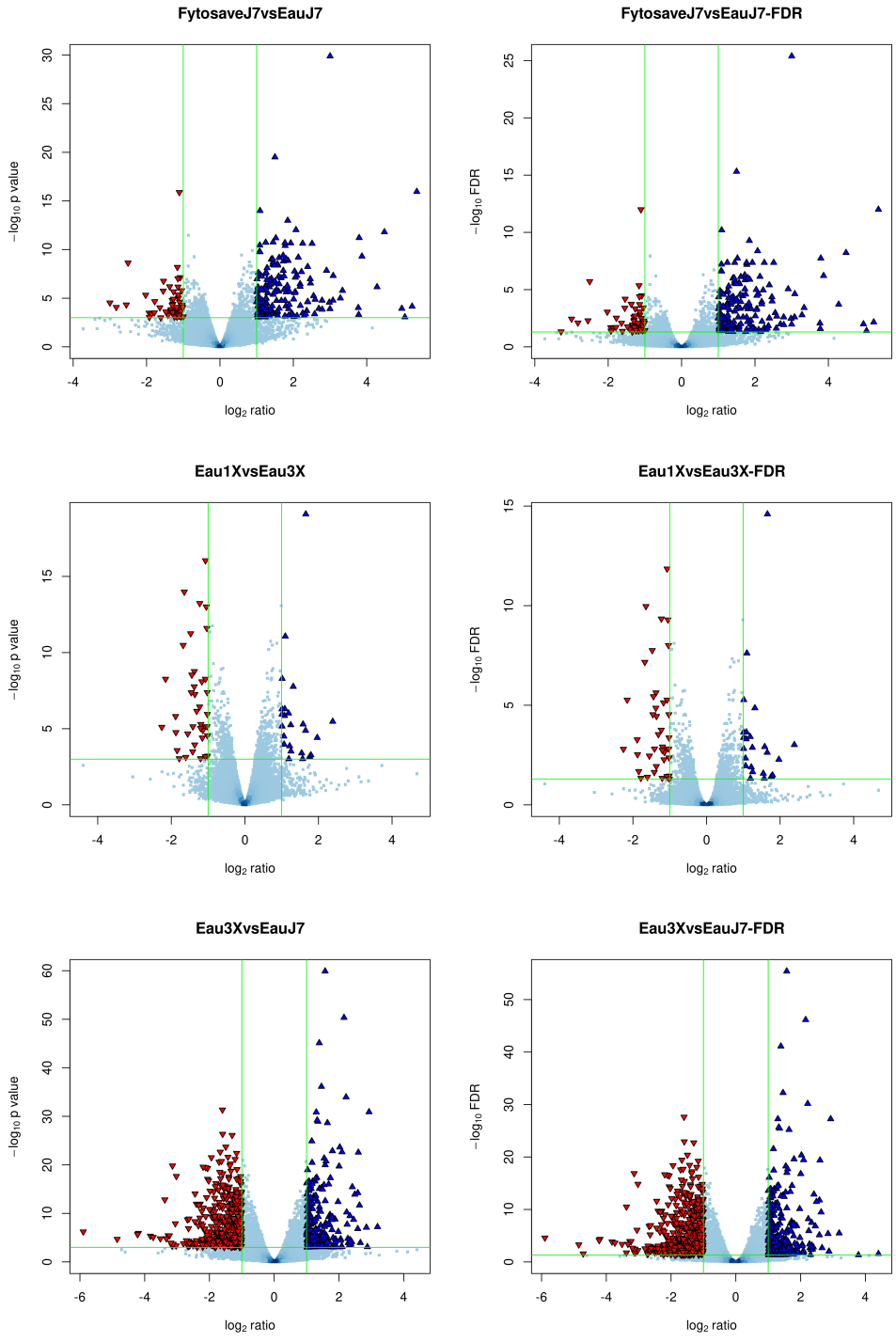
A **Volcano plot** visualizes the data along dimensions of biological and statistical significance, i.e. edgeR-estimated $\log_2$-ratios versus minus the $\log_{10}$ p-value. The dots are colored red and blue if they are classified as down- and up-regulated, respectively. The left figures use the uncorrected p-values, the right figures use the FDR-values.



Nucleomics Core, VIB
Herestraat 49
3000 LEUVEN,
Post Box 816
Belgium

VIB NUCLEOMICS CORE

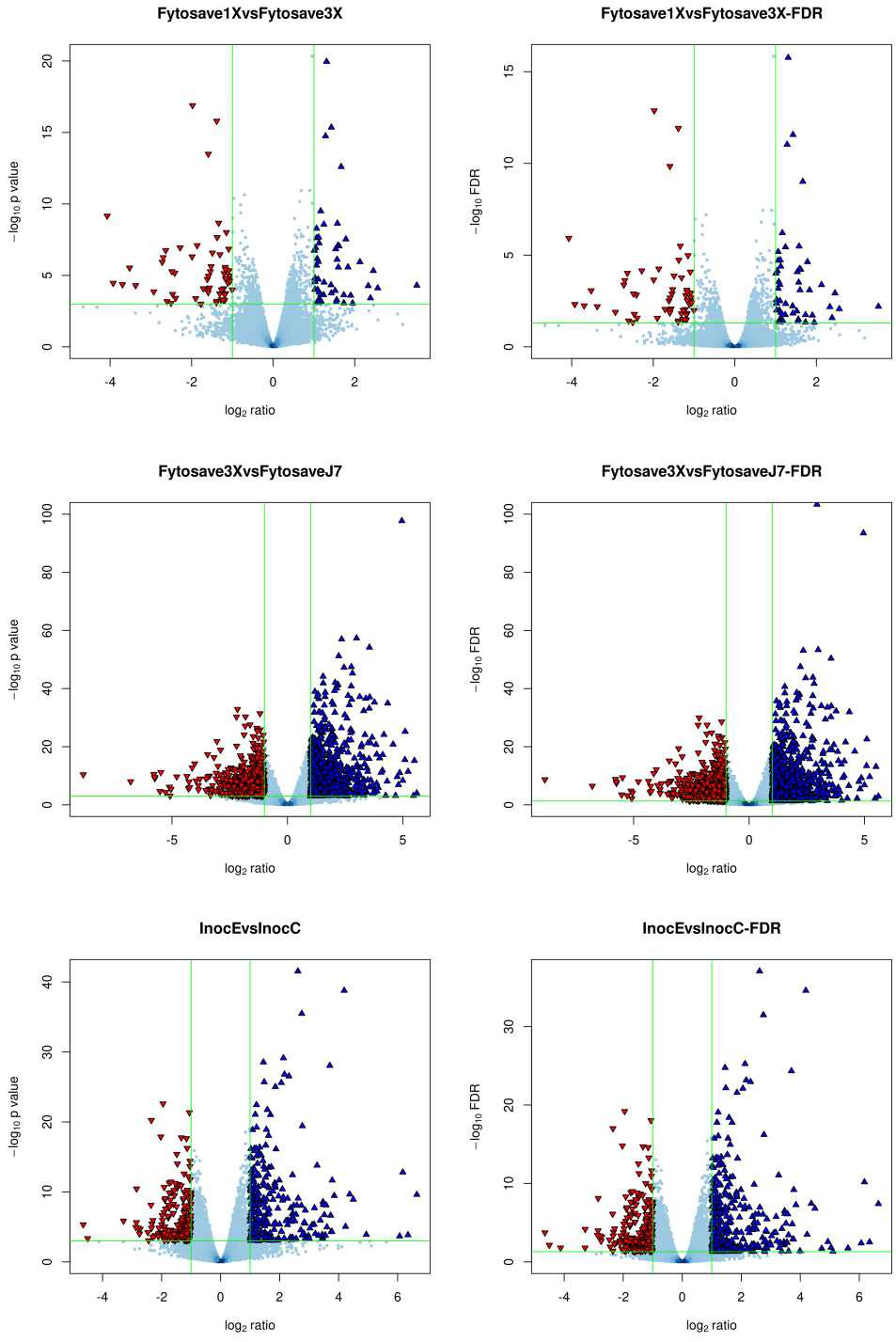
Contact info:
nucleomics@vib.be
+32/16/373128
+32/16/373129
www.nucleomics.be



Nucleomics Core, VIB
Herestraat 49
3000 LEUVEN,
Post Box 816
Belgium

VIB NUCLEOMICS CORE

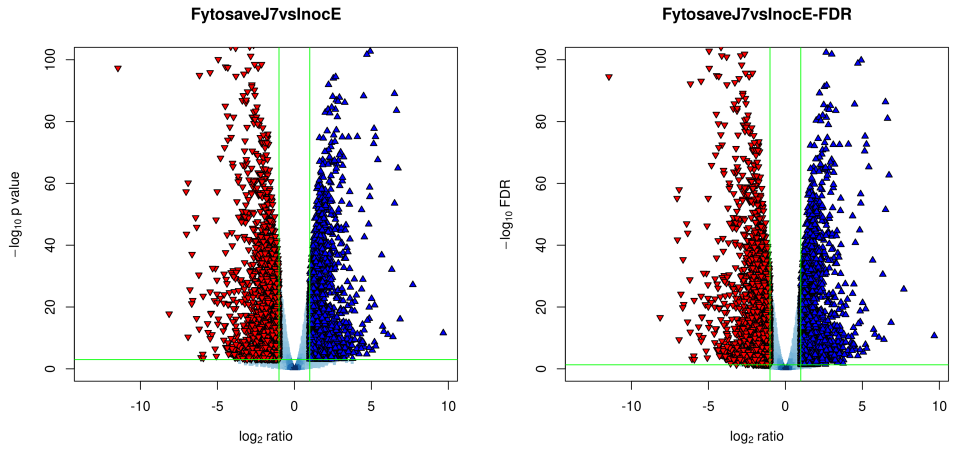
Contact info:
nucleomics@vib.be
+32/16/373128
+32/16/373129
www.nucleomics.be



Nucleomics Core, VIB
Herestraat 49
3000 LEUVEN,
Post Box 816
Belgium

VIB NUCLEOMICS CORE

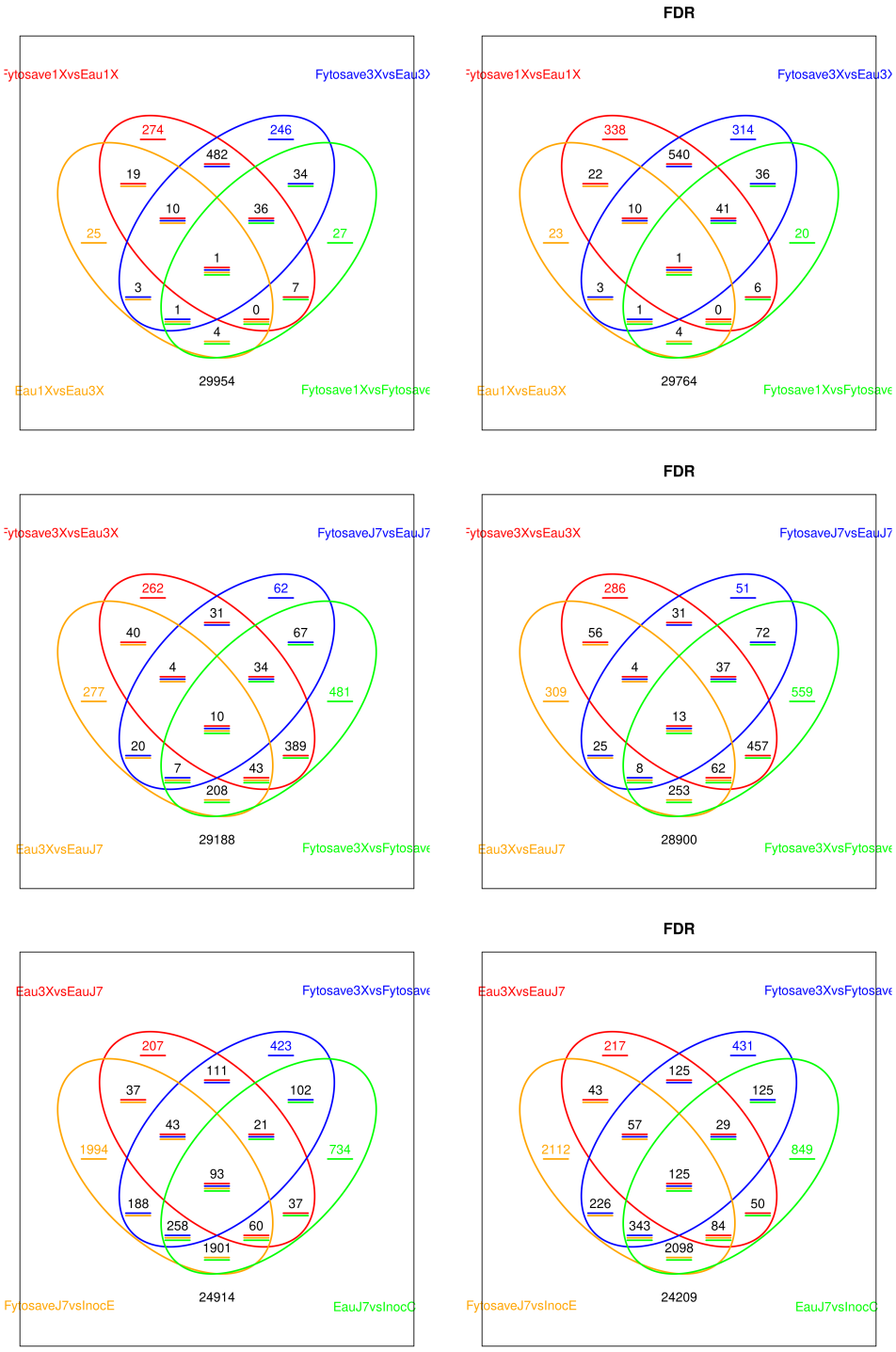
Contact info:
nucleomics@vib.be
+32/16/373128
+32/16/373129
www.nucleomics.be



7.3.3 Meta-analysis

The outcomes for the tested contrasts are compared by computing the overlap between sets of differentially expressed genes (using corrected p-values). The actual gene names can be found by putting adequate filters on columns in the provided Excel sheet. A hierarchical clustering provides also insight in relative expression profiles. Heatmaps are made according to the genes in the intersection or union of the different comparisons.

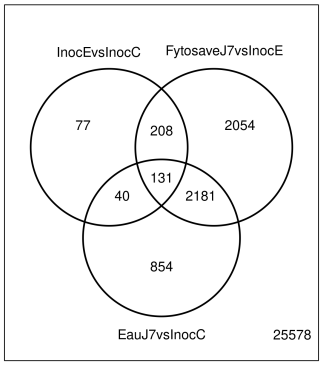
Differentially expressed genes from comparisons



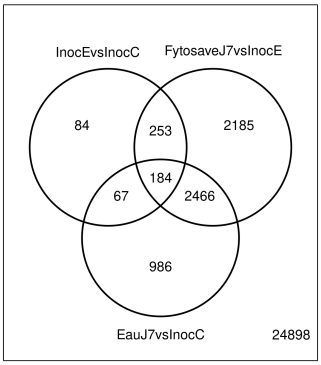
Nucleomics Core, VIB
Herestraat 49
3000 LEUVEN,
Post Box 816
Belgium

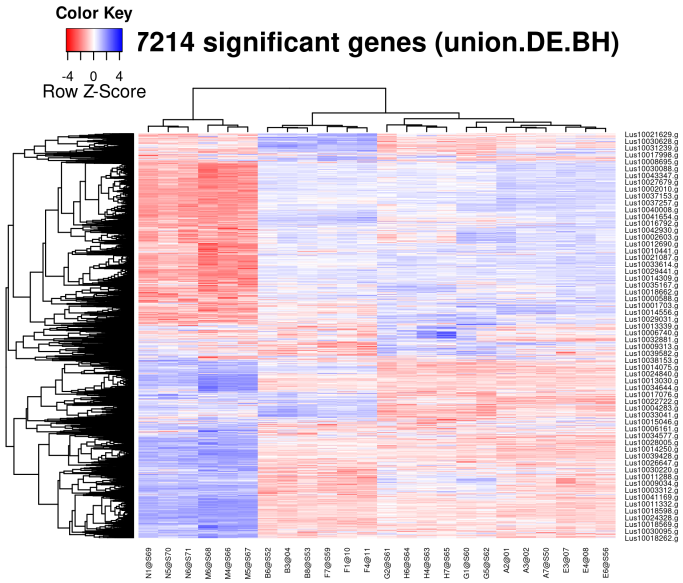
VIB NUCLEOMICS CORE

Contact info:
nucleomics@vib.be
+32/16/373128
+32/16/373129
www.nucleomics.be


InocEvsInocC
FytosaveJ7vsInocE
EauJ7vsInocC
77
208
2054
131
40
2181
854
25578

FDR


InocEvsInocC
FytosaveJ7vsInocE
EauJ7vsInocC
84
253
2185
184
67
2466
986
24898


Color Key
7214 significant genes (union.DE.BH)
-4 0 4
Row Z-Score

7.4 Data transfer

The results are stored in an Excel file *exp3949-StatisticalResults.xlsx*:

- **Gene ID** *(Column A)*: Unique gene identifier
- **Gene Name** *(Column B)*
- **Statistical Results** *(Columns C-AZ)*: Five columns per contrast.
  ■ *logFC*: The $\log_2$-ratio as calculated by edgeR.
  ■ *PValue*: The p-values of the likelihood ratio test, as calculated by edgeR.
  ■ *FDR*: The 'corrected p-values' as calculated by edgeR with Benjamini-Hochberg correction.
  ■ *DE*: We compare the p-value with a stringent cut-off of 0.001 and an absolute log2 fold-change > 1 and assign
    - **1:** the gene is up-regulated
    - **-1:** the gene is down-regulated
    - **0:** the gene is not differentially expressed

  ■ *DE.BH*: We compare the corrected p-values with a cut-off of 0.05 and an absolute log2 fold change > 1 and assign
    - **1:** the gene is up-regulated
    - **-1:** the gene is down-regulated
    - **0:** the gene is not differentially expressed
- **Additional gene annotation** *(Columns BA-BU)*
- **Flag** *(Column BV)*: Genes are flagged (i.e. having a 1) when the expression-value in at least one sample belongs to the 5% lowest of 5% highest values for that sample. In such case, our preprocessing may lead to large distortions of the FPKM values (i.e. exactly same value in several samples) and the estimated fold change. We always recommend qPCR-validation for all genes that take up an important role in your story, but emphasize the necessity even more for flagged genes. We expect that non-flagged genes will be confirmed by qPCR more often than flagged genes.

NOTE: Numerical values smaller than 1e-200 were replaced by '0' to comply with Excel limitations.

8 What's next?

The project may continue as follows:

1. *Signature validation*
   - Meta-analysis: play around with filters in the Excel and merge information from different comparisons into biologically sound signatures.
   - Exclude mapping artefacts: verify (and eliminate) some differentially expressed genes by checking the alignments in the Integrative Genomics Viewer (`www.broadinstitute.org/igv/`)
   - Technological validation: test (and eliminate) some differentially expressed genes in the same samples with another technology like qPCR/nCounter
   - Biological validation: test (and eliminate) some differentially expressed genes in new samples
2. *Explanatory modeling:* To get more insight, model the connection between differentially expressed genes and phenotype by qualitatively explaining which molecular mechanisms and functional pathways are affected.
   - Signature optimization: other gene selection methods (e.g. DESeq [1] or from your expert knowledge) may provide other lists of genes that nevertheless often probe the same pathways; combining them in one large signature may increase sensitivity when conducting pathway analysis [13][14].
   - Analysis of GO terms, pathways and gene regulatory with the *now relatively outdated* DAVID (`http://david.abcc.ncifcrf.gov`) or GSEA (`http://www.broadinstitute.org/gsea/`)
   - Other open access tools (`http://www.pathguide.org`), or Ingenuity Pathway Analysis (IPA).
   - Or using more recents web-tools incuding Enrichr (`http://amp.pharm.mssm.edu/Enrichr/`) or webgestalt (`http://bioinfo.vanderbilt.edu/webgestalt/`).
   - Regulatory Analysis, using iRegulon (`http://iregulon.aertslab.org`) to detect enriched Transcription Factors and their targets.

9 Acknowledgements

We ask you to acknowledge the Nucleomics Core in all papers to which our facility has contributed. Please refer to the Nucleomics Core as follows: *[Sequencing / Data /???] production and/or analysis was/were performed by VIB Nucleomics Core (`www.nucleomics.be`).*

References

[1] S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106+, Oct. 2010.

[2] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B*, 57:289–300, 1995.

[3] M. Consortium. The MicroArray quality control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology*, 24(9):1151–1161, Sept. 2006.

[4] A. D. et al. Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[5] HannonLab. Fastx-toolkit. *`http://hannonlab.cshl.edu/fastx_toolkit/index.html`*, 2010.

[6] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and samtools. *Bioinformatics (Oxford, England)*, 25(16):2078–2079, August 2009.

[7] Y. Liao, G. K. Smyth, and W. Shi. FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930, 2014.

[8] M. Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1), 2011.

[9] M. Morgan, M. Lawrence, and S. Anders. *ShortRead: Base classes and methods for high-throughput short-read sequencing data.*, 2009.

[10] D. Risso, K. Schwartz, G. Sherlock, and S. Dudoit. GC-content normalization for RNA-seq data. *BMC Bioinformatics*, 12(1):480+, Dec. 2011.

[11] M. Robinson and A. Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3):R25+, Mar. 2010.

[12] M. D. Robinson and G. K. Smyth. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23:2881–2887, Nov 2007.

[13] W. Shi, M. Bessarabova, D. Dosymbekov, Z. Dezso, T. Nikolskaya, M. Dudoladova, T. Serebryiskaya, A. Bugrim, A. Guryanov, R. J. Brennan, R. Shah, J. Dopazo, M. Chen, Y. Deng, T. Shi, G. Jurman, C. Furlanello, R. S. Thomas, J. C. Corton, W. Tong, L. Shi, and Y. Nikolsky. Functional analysis of multiple genomic signatures demonstrates that classification algorithms choose phenotype-related genes. *The pharmacogenomics journal*, 10(4):310–323, Aug. 2010.

[14] A. Statnikov and C. F. Aliferis. Analysis and computational dissection of molecular signature multiplicity. *PLoS Comput Biol*, 6(5):e1000790+, May 2010.