SUPPLEMENTARY INFORMATION for manuscript:

Strand-specific deep sequencing of the transcriptome

Ana P. Vivancos[1*]‡, Marc Güell[1]‡, Juliane C. Dohm[1,2]‡, Luis Serrano[1,3§] & Heinz Himmelbauer[1§]

[1]Centre for Genomic Regulation (CRG), UPF, C. Dr. Aiguader 88, 08003 Barcelona, Spain.
[2]Max Planck Institute for Molecular Genetics, Ihnestr. 63-73, 14195 Berlin, Germany.
[3]Institució Catalana de Recerca i Estudis Avancats (ICREA).

‡Equally contributing authors.

[§]Corresponding authors

* Present address: Vall d'Hebron Institute of Oncology, Vall d'Hebron University Hospital, Psg. Vall d'Hebron 119-129, 08035 Barcelona, Spain

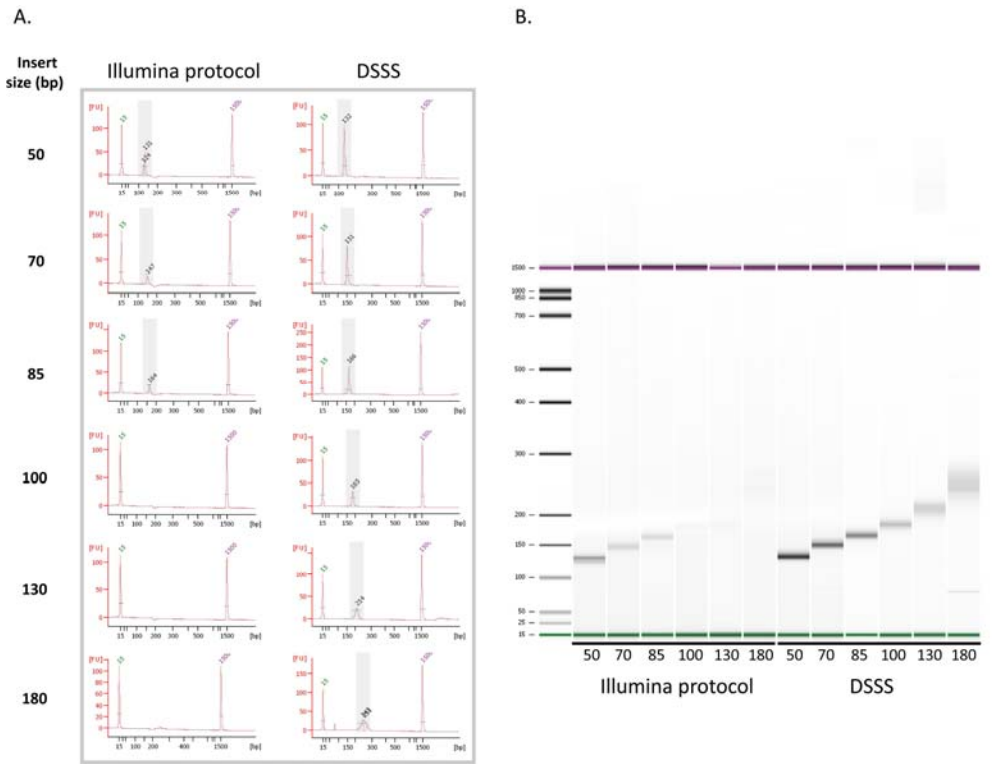
Email addresses:
APV: avivancos@vhio.net
MG: marc.guell@crg.es
JCD: juliane.dohm@crg.es
LS: luis.serrano@crg.es
HH: heinz.himmelbauer@crg.es

This file includes:

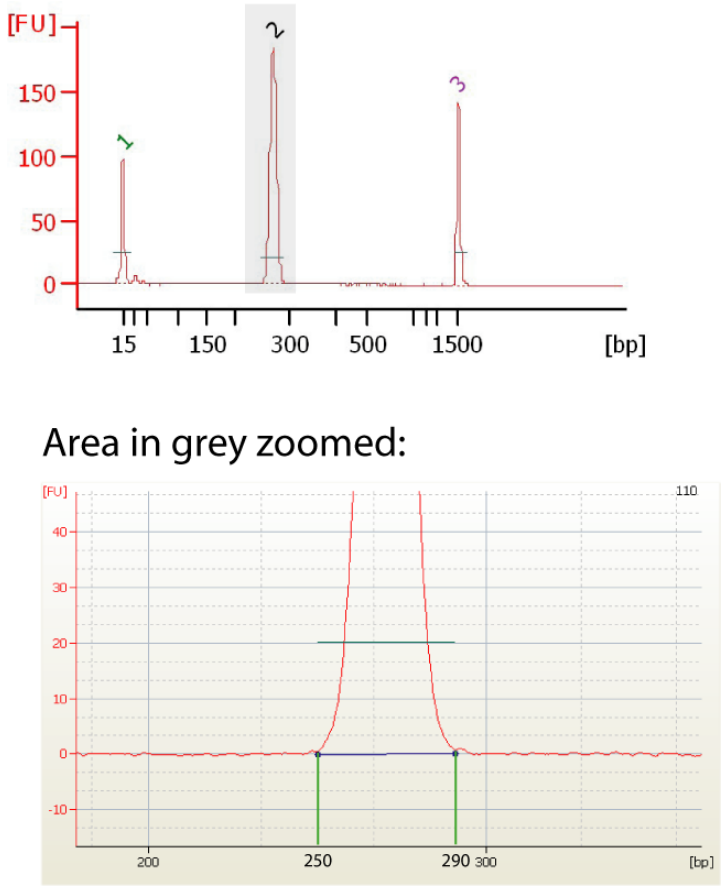
1- Supplementary Figures

2- Supplementary Tables

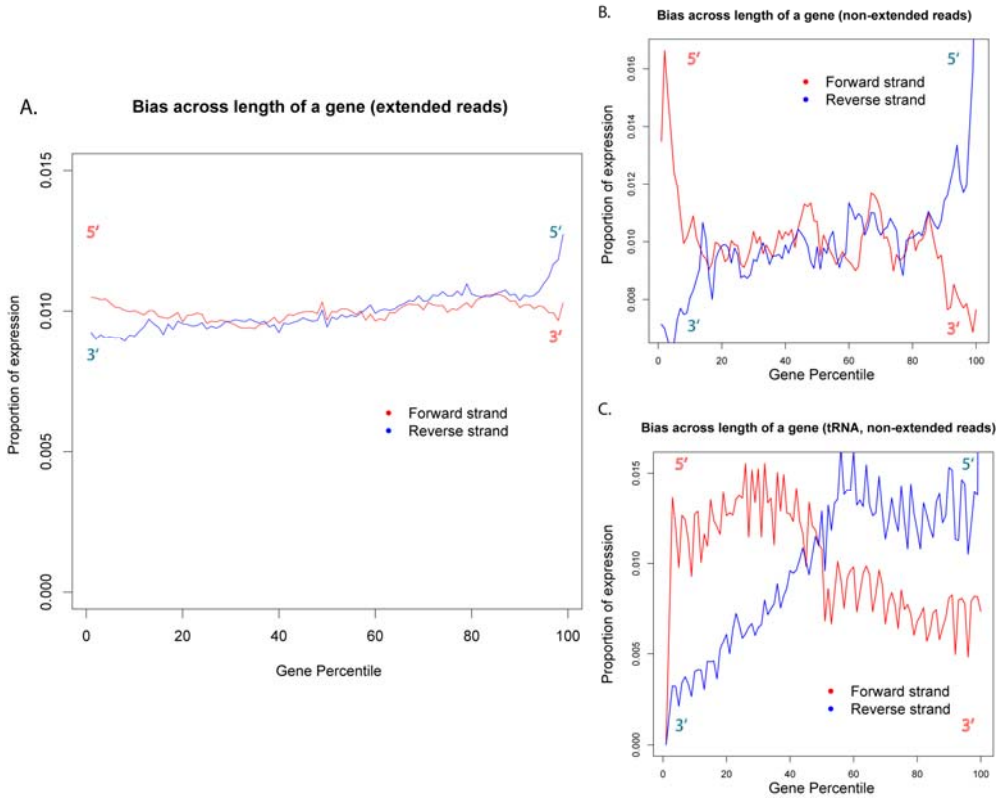
1-SUPPLEMENTARY FIGURES



Suppl. Fig. 1. Side-by side comparison of the Illumina small RNA sample preparation protocol and DSSS sample preparation for fragment size range 50-180 nt. Eight micrograms of total RNA from mouse brain was depleted of ribosomal RNA and fragmented to a size range of 60-200 nt. The fragmented and depleted sample was loaded onto a 10% polyacrylamide gel and we excised bands corresponding to fragments of 50, 70, 85, 100, 130 and 180 nt. After elution of the RNA contained within each of the gel slices, we split the samples and used them either for sample preparation either using the Illumina small RNA sample preparation kit, or the DSSS protocol. We show the final output of products for each of the samples**.** A) Electropherogram of each of the samples using the Agilent 2100 Bioanalyzer (Lab-on-a-chip DNA 1000)**.** B) Gel view of the electropherograms in A. Note the broader amplification bands in the DSSS protocol for inserts of 130 and 180 bp, the reason to this is the low resolution when excising these sizes in a 10% polyacrylamide gel. For the sequenced samples we performed purification on 6% PAGE (see Methods).

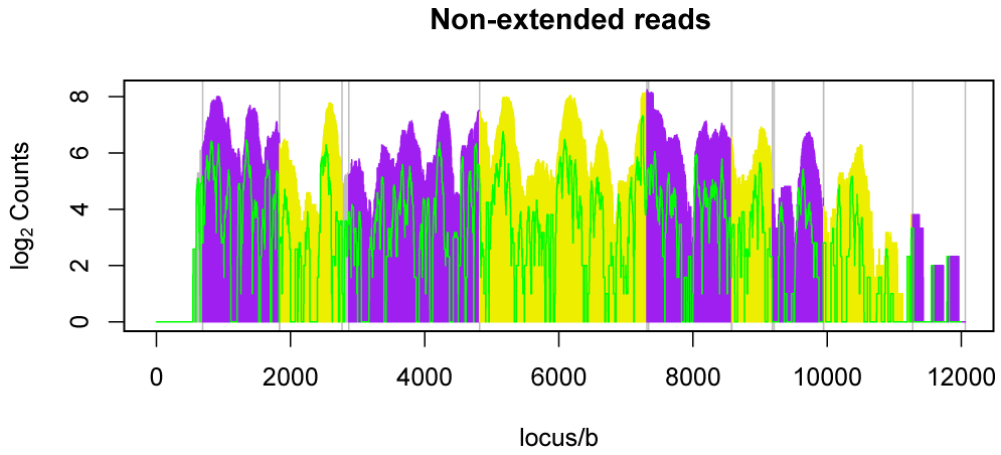


Suppl. Fig. 2. Insert size selection electropherogram, using the Agilent 2100 Bioanalyzer (Lab-on-a-chip DNA 1000). Peaks 1 and 3 correspond to size marker, peak 2 (270 nt) corresponds to library insert plus ligated 3' and 5' adapters. Adapters contribute 70 nt to the fragment size.

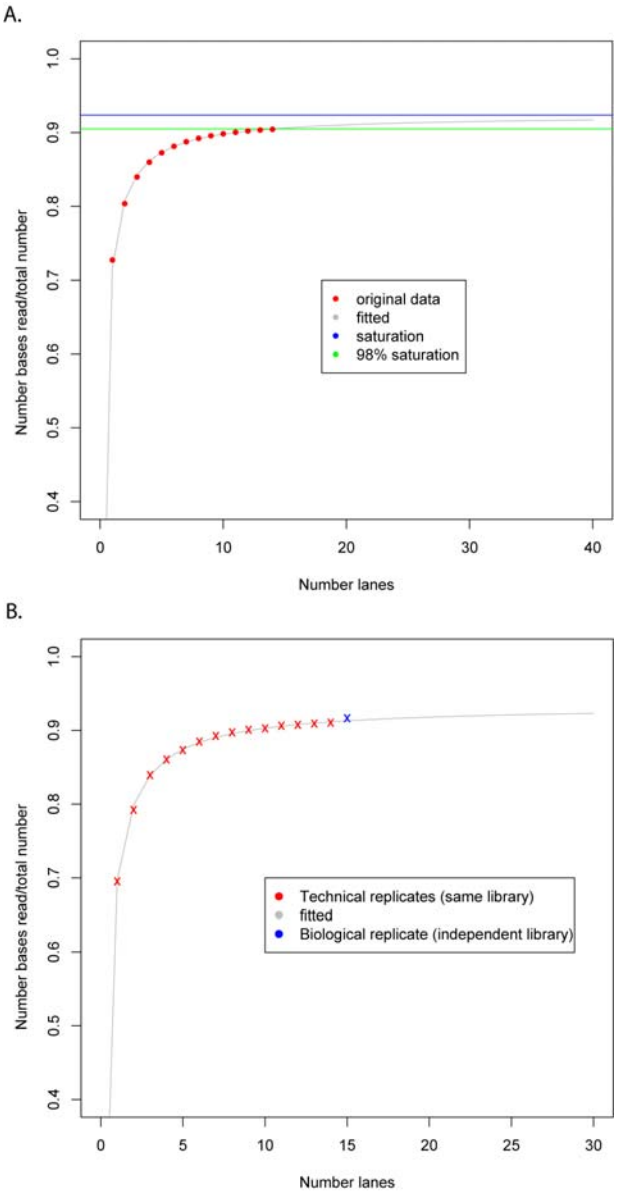


Suppl. Fig. 3. A). Coverage bias across the length of transcripts as observed with reads extended to 180 nt. Mean expression of every gene is interrogated in each percentile of the gene length using the log2 transformation of the mean number of reads mapping in such a percentile. The fraction of detecting each of the percentiles for any gene is computed. Even transcript coverage with slight overrepresentation of 5’ ends is detected. B) Coverage bias across the length of transcripts for all genes, as observed with reads without extension. C) tRNA genes only, to help understand the starting position of the bias. We chose all tRNAs (average length 79 bases). Only those present in polycistronic constructs longer than 180 nt were detected.

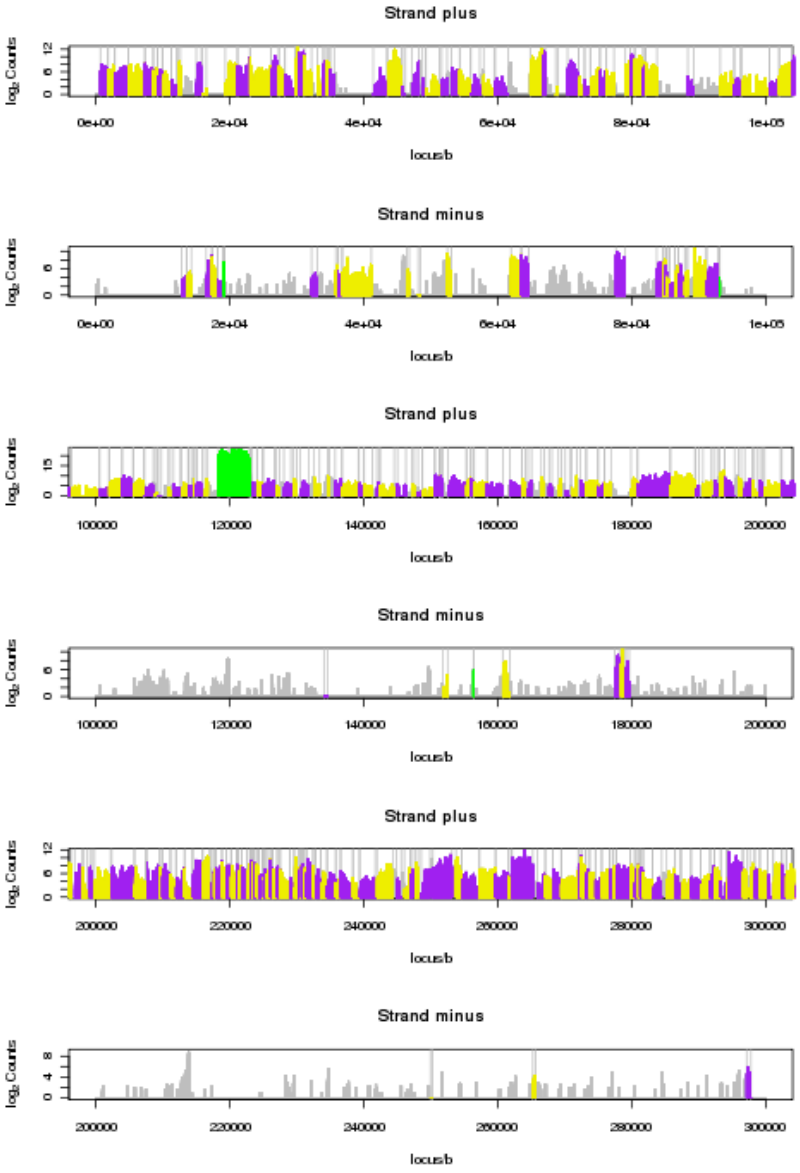
Deleted: ,

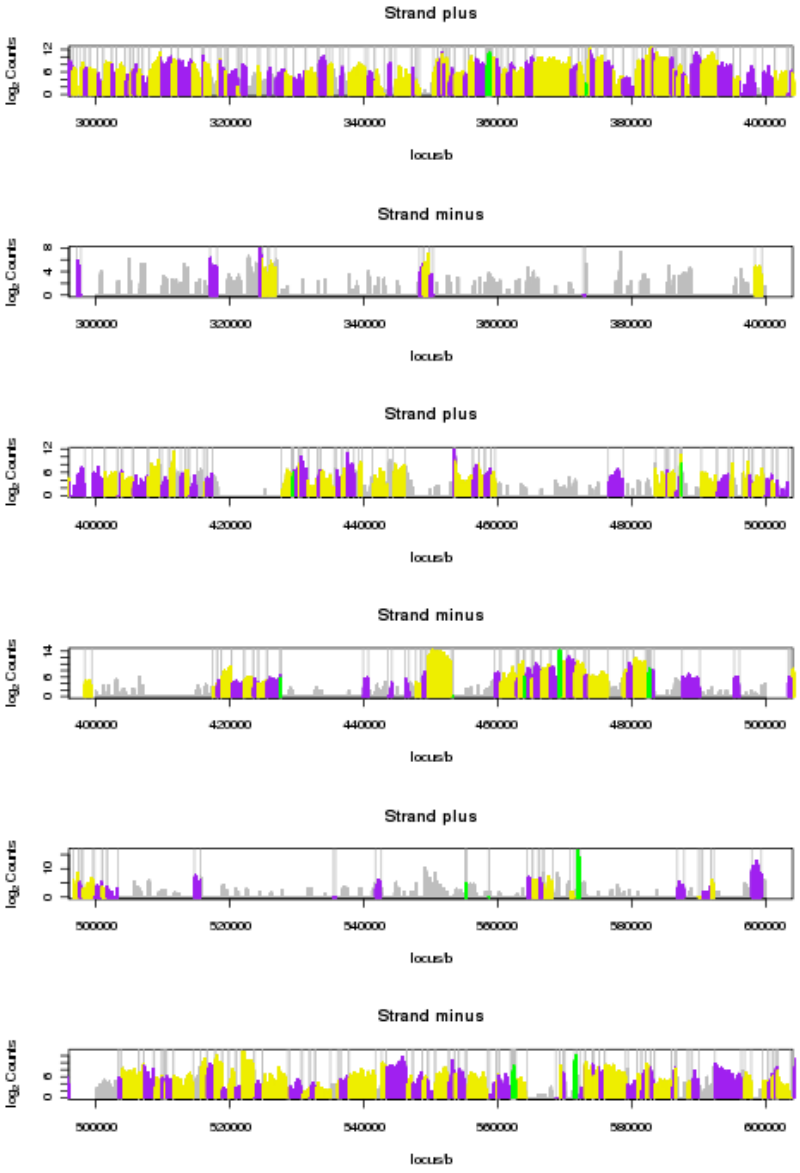


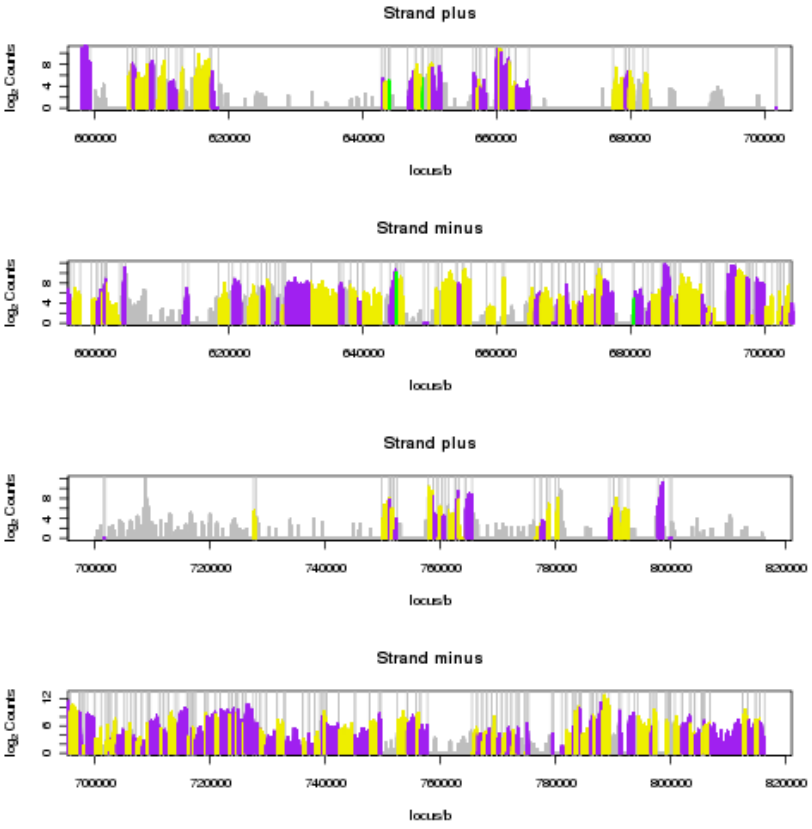
Suppl. Fig. 4. Comparison of DSSS with and without read extension on a section of the mycoplasma genome. *M. pneumoniae* strain M129 genome (genome coordinates 0-12000 bp; Forward strand). DSSS signals (extended reads) on alternately colored protein coding gene are shown in purple and yellow. Grey vertical lines mark gene boundaries. In green, DSSS signal without read extension. DSSS with and without read extension show 0.85 Pearson correlation. Without read extension, coverage is 20% decreased.



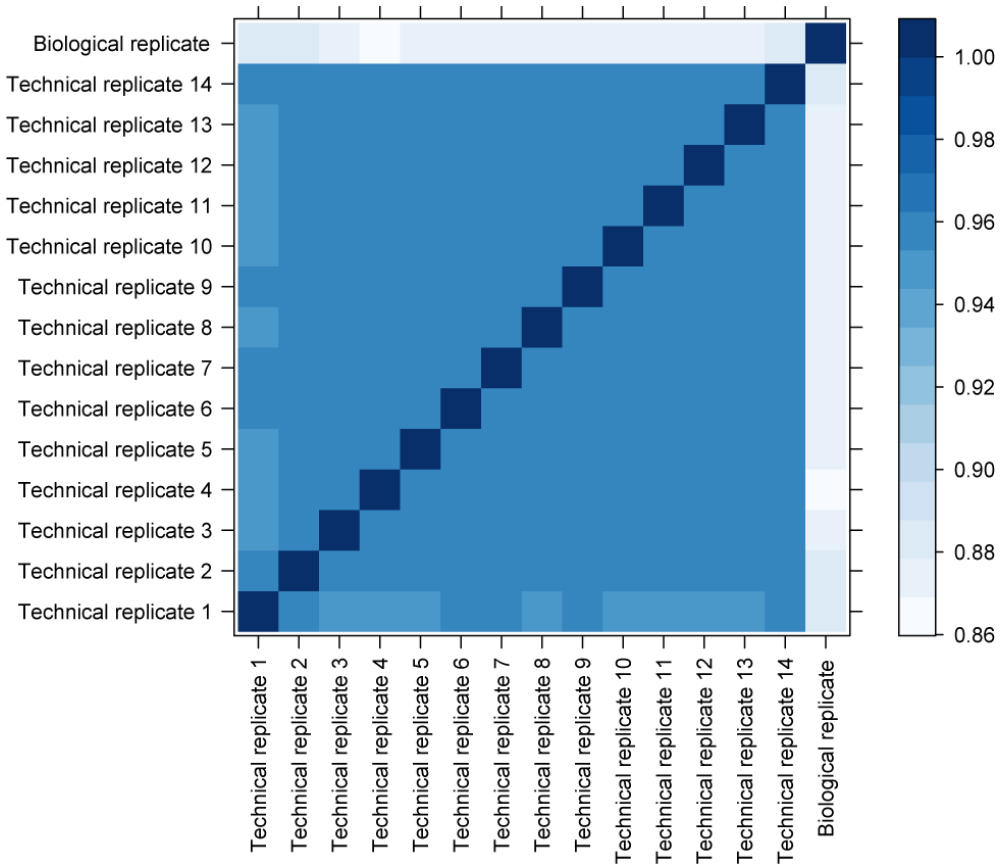
Suppl. Fig. 5. A**)** DSSS cumulative coverage plot. Number of bases mapped is computed after adding one more independent experiment (read dots are the mean value for all possible combinations of a certain number of experiments) and normalized to the length of all ORFs (ratio: Number of bases read/total number). Data is fitted to a two parameter hyperbole to detect the level of saturation. We assume complete saturation when the ratio reaches 98% of the hyperbole horizontal asymptote. Saturation is reached after the 14th lane. B) DSSS cumulative coverage plot combining different libraries. After adding data from one flowcell lane from an independent library, the coverage is not significantly increased.


Strand plus
log2 Counts
0
6
12
0e+00
2e+04
4e+04
6e+04
8e+04
1e+05
locus b
Strand minus
log2 Counts
0
6
0e+00
2e+04
4e+04
6e+04
8e+04
1e+05
locus b
Strand plus
log2 Counts
0
15
100000
120000
140000
160000
180000
200000
locus b
Strand minus
log2 Counts
0
6
100000
120000
140000
160000
180000
200000
locus b
Strand plus
log2 Counts
0
6
12
200000
220000
240000
260000
280000
300000
locus b
Strand minus
log2 Counts
0
4
8
200000
220000
240000
260000
280000
300000
locus b


Strand plus
log2 Counts
0 6 12
300000 320000 340000 360000 380000 400000
locusb
Strand minus
log2 Counts
0 4 8
300000 320000 340000 360000 380000 400000
locusb
Strand plus
log2 Counts
0 6 12
400000 420000 440000 460000 480000 500000
locusb
Strand minus
log2 Counts
0 6 14
400000 420000 440000 460000 480000 500000
locusb
Strand plus
log2 Counts
0 10
500000 520000 540000 560000 580000 600000
locusb
Strand minus
log2 Counts
0 6
500000 520000 540000 560000 580000 600000
locusb

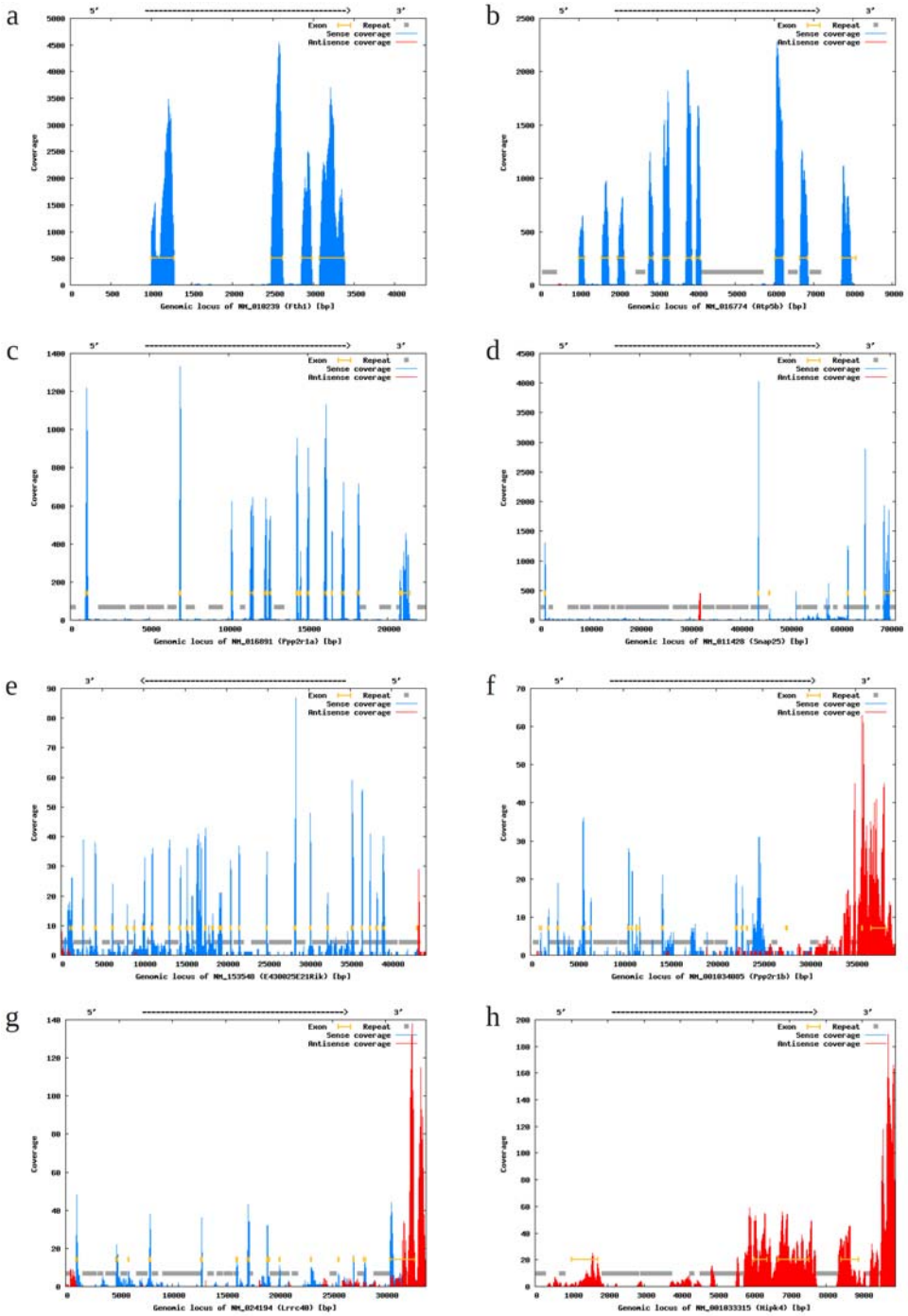


Suppl. Fig. 6. DSSS reference data for the *M. pneumoniae strain* M129 genome. Purple/yellow: Protein coding genes; Green: Genes encoding untranslated RNAs; Grey: Regions without annotated genes. Grey vertical lines mark gene boundaries.



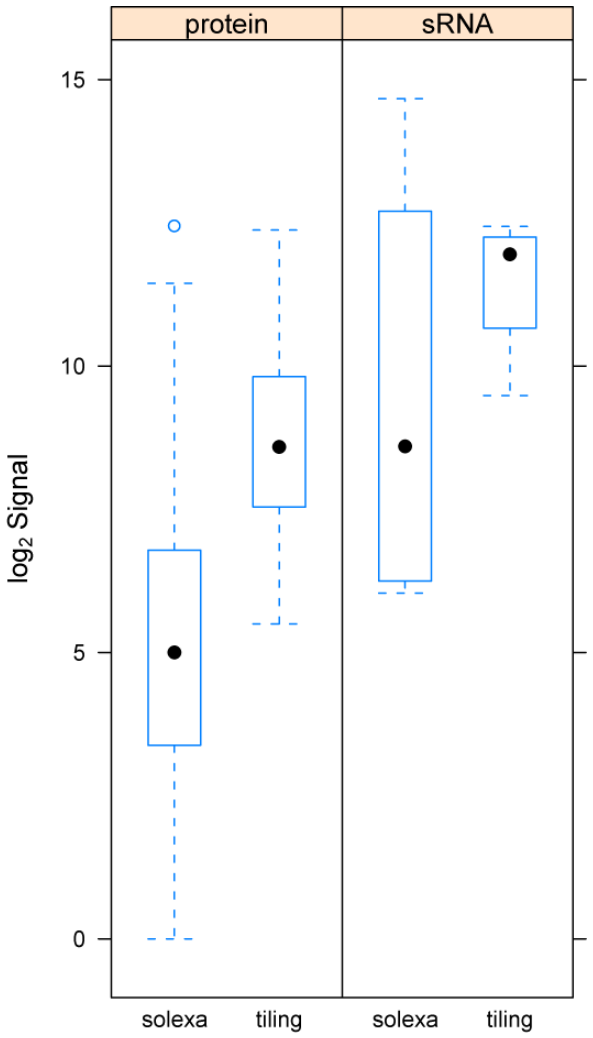
Suppl. Fig. 7. Correlation between technical and biological replicates. Each experiment refers to one lane, run on the Illumina GA II sequencing platform. Pair-wise Pearson correlation is calculated for individual pileups* generated from each independent dataset (single lane).

* pileup: vector with the length of the reference genome and each position indicates the number of reads mapped to the corresponding base.

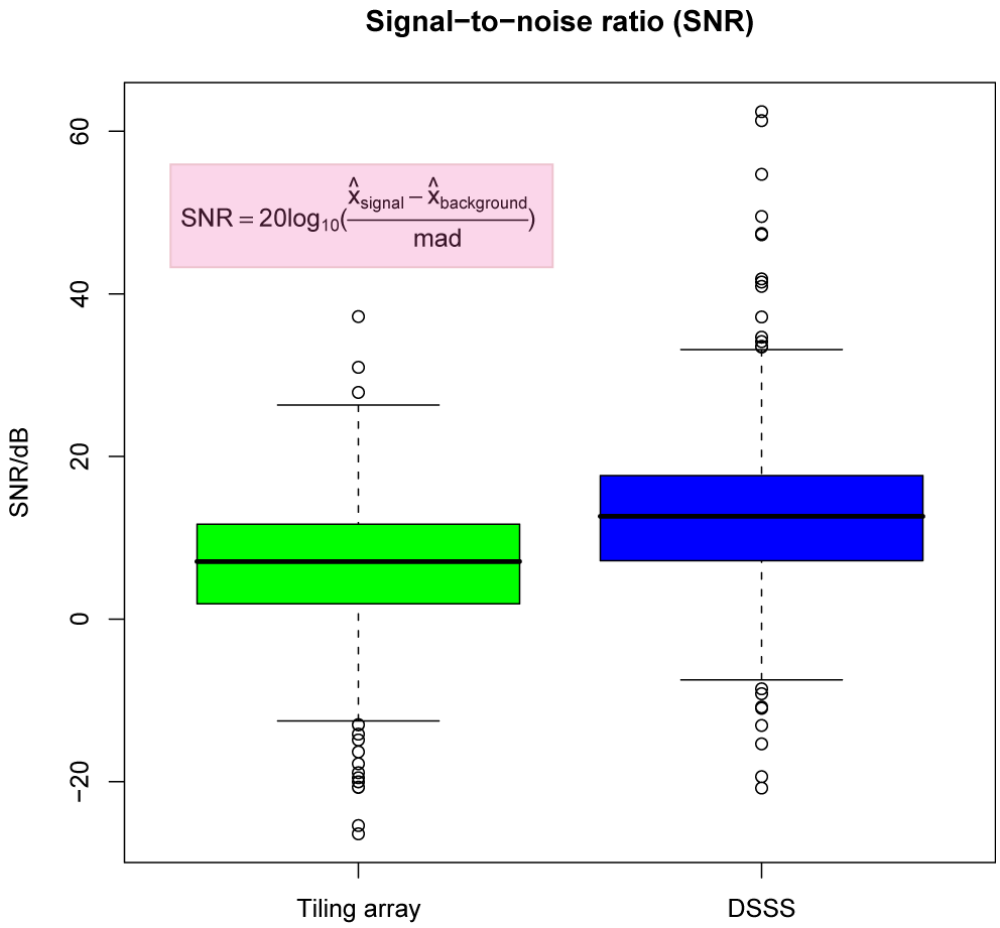


Suppl. Fig. 8. Analysis of mouse gene loci by DSSS. Reads were mapped against the repeat-masked genomic mouse sequences plus 1kbp upstream and 1kbp downstream of the gene locus.

Exon annotation (yellow) is derived from the mouse genome annotation. The coverage by DSSS reads is shown in blue (sense orientation) or red (antisense orientation), respectively. A). *Fth1*; B) *Atp5b*; C) *Ppp2ra1*; D) *Snap25*. The antisense signal detected at the *Snap25* locus co-localises with an *Rpl21* pseudogene. E) *E430025E21Rik*. In the mouse genome, *E430025E21Rik* is closely flanked by *Nsmce2* and *Sqle*, respectively (antisense signal at 5’ and at 3’ of *E430025E21Rik*). F) *Ppp2r1b*. Two isoforms for *Ppp2r1b* are available from RefSeq. The exon information for the long isoform was used for exon annotation. According to the mouse genome annotation, the long isoform overlaps with the *Sik2* gene. G) *Lrcc40*. The antisense signal at the 5’ end of *Lrcc40* can be attributed to transcription from the *Sfrs11* gene which is in head-to-head orientation with *Lrcc40*. There is no annotated feature in the mouse genome explaining the observed antisense transcription at the *Lrcc40* 3’ end. H) Antisense transcription at the *Hipk4* locus, in absence of any sense transcription of *Hipk4*.



Suppl. Fig. 9. DSSS dynamic range. Mean DSSS values and mean tiling array intensities were computed for annotated protein coding genes and annotated structural RNA genes (sRNAs, excluding tRNA and rRNA). Note that DSSS dynamic range spans several orders of magnitude more than tiling intensities.



Suppl. Fig. 10. Signal-to-noise ratio (SNR) for DSSS and tiling array. For every gene, the antisense median signal, assumed to be an estimation of the background, was subtracted from the sense median signal and divided by the median absolute deviation of the sense gene. The ratio has been expressed in decibels (dB). The formula used is depicted in the pink box. SNR is 5.6 dB higher in DSSS than in the tiling array.

SUPPLEMENTARY TABLES

Suppl. Table 1. Details of sequencing runs. Clusters (raw): Average number of clusters per tile; Clusters (PF): Average number of clusters per tile passing quality filter; % PF Clusters: Percentage of clusters passing quality filter; Lane yield: Bases sequenced passing the quality filter; % Mapped sequences: Percentage of mapped sequences.

(a) *M. pneumoniae* cDNA (5 flowcells, 14 lanes, single reads)

Run:	**30W5WAAXX**				
Lane	Clusters (raw)	Clusters (PF)	% PF Clusters	Lane yield (kb)	% Mapped sequences
6	181541 +/- 4950	75538 +/- 7430	41.62 +/- 4.03	377691	71.5
7	167815 +/- 3662	75444 +/- 6856	45.00 +/- 4.44	377224	71.5
8	173477 +/- 6394	70621 +/- 8081	40.74 +/- 4.68	353106	71.5
Run:	**30KTCAAXX**				
Lane	Clusters (raw)	Clusters (PF)	% PF Clusters	Lane Yield (kb)	% Mapped sequences
1	155131 +/- 7316	62196 +/- 5435	40.15 +/- 3.70	210473	97.6
2	156186 +/- 10160	68255 +/- 4032	43.84 +/- 3.30	245719	97.6
3	160550 +/- 6908	69478 +/- 4251	43.32 +/- 2.77	250123	97.6
5	158401 +/- 6572	72704 +/- 3182	45.95 +/- 2.28	261736	97.6
6	158278 +/- 7001	69567 +/- 3536	44.02 +/- 2.74	250442	97.5
7	159490 +/- 7026	67590 +/- 5009	42.46 +/- 3.64	243324	97.7
Run:	**30LCYAAXX**				
Lane	Clusters (raw)	Clusters (PF)	% PF Clusters	Lane Yield (kb)	% Mapped sequences
1	151813 +/- 7618	64962 +/- 5561	42.90 +/- 4.26	263681	86.0
2	149640 +/- 4758	70874 +/- 3655	47.35 +/- 1.72	290585	85.9
Run:	**30HGHAAXX**				
Lane	Clusters (raw)	Clusters (PF)	% PF Clusters	Lane Yield (kb)	% Mapped sequences
3	102181 +/- 8081	65034 +/- 2477	64.01 +/- 5.10	231783	86.1
4	101613 +/- 7818	65306 +/- 2927	64.58 +/- 4.68	235101	96.7
Run:	**311JJAAXX**				
Lane	Clusters (raw)	Clusters (PF)	% PF Clusters	Lane Yield (kb)	% Mapped sequences

8	111233 +/- 25667	59696 +/- 9726	55.51 +/- 10.52	238785	89.0

(b) *M. pneumoniae* cDNA, biological replicate, single reads

Run: **313P8AAXX**

Lane	Clusters (raw)	Clusters (PF)	% PF Clusters	Lane Yield (kb)	% Mapped sequences
1	112125 +/- 14959	77286 +/- 11473	69.09 +/- 6.52	309147	87.8

(c) Mouse cDNA, Solexa library with insert size of 200 nt, paired end reads

Run: **42YV6AAXX**

Lane/Read	Clusters (raw)	Clusters (PF)	% PF Clusters	Lane Yield (kb)	% Mapped sequences
1/1	154301 +/- 4738	122893 +/- 5476	79.77 +/- 4.94	796348	69.1
1/2	154301 +/- 4738	122893 +/- 5476	79.77 +/- 4.94	796348	69.9
2/1	156474 +/- 5542	126158 +/- 4028	81 +/- 2	817505	69.6
2/2	156474 +/- 5542	126158 +/- 4028	80.73 +/- 3.89	817505	70.2
3/1	157224 +/- 5929	127957 +/- 3161	81.49 +/- 3.53	829159	69.8
3/2	157224 +/- 5929	127957 +/- 3161	81.49 +/- 3.53	829159	70.1
4/1	156180 +/- 5678	127397 +/- 3234	81.68 +/- 3.69	825531	69.8
4/2	156180 +/- 5678	127397 +/- 3234	81.68 +/- 3.69	825531	69.9
6/1	155128 +/- 5771	126870 +/- 3251	81.92 +/- 4.04	822119	69.9
6/2	155128 +/- 5771	126870 +/- 3251	81.92 +/- 4.04	822119	70.1
7/1	157830 +/- 5142	128742 +/- 2312	81.67 +/- 3.26	834248	69.8
7/2	157830 +/- 5142	128742 +/- 2312	81.67 +/- 3.26	834248	70.1
8/1	155644 +/- 4396	126336 +/- 2487	81.26 +/- 3.34	818654	69.7
8/2	155644 +/- 4396	126336 +/- 2487	81.26 +/- 3.34	818654	70

Suppl. Table 2. qPCR confirmation of DSSS expression of selected *M. pneumoniae* genes, using SYBR Green real-time RT-PCR. The details on the genes, their position and strand on the chromosome as well as the sequences of the primers (designed using Primer3, annealing temperature of 60ºC, amplification product of 60-150bp) used for amplification are indicated:

start	end	strand	gi	name	Gene	qPCR primer 1 (5' to 3')	qPCR primer 2 (5' to 3')
41409	43409	+	13507774	-	mpn035	CCAGGACTTTGATGGCATTT	GGTGGAGTTGGCTACACGTT
571431	571801	-	13507739	-	mpns03	AACGCGGTAAACTCCACAAG	TGCCAATAAGCCATGTTCTG
4821	7340	+	13507743	gyrA	mpn004	TGAACATGTTGGCACTGGTT	TGAAAGCGCTCCTGGTACTT
692	1834	+	13507740	dnaN	mpn001	GAATTCCCCAGAATCCGTTT	GAAACAGCATGGGCAATCTT
651802	653949	-	13508270	clpB	mpn531	CACGCGGTGAAATTAAGGTT	CCCGCATAATGGTAAGTGCT
449573	452995	-	13508115	-	mpn376	CTCTAACCGCGAAAGGACAG	TTGGTTGGCATAACCCAAAT
118312	119824	+	13507739	-	mpnr01	CAGCTCGTGTCGTGAGATGT	TTGACGTCATCCCTTCCTTC
120057	122961	+	13507739	-	mpnr02	GTAGGCGATGGACAACAGGT	GCACAACGGATTTGCCTATT

Suppl. Table 3. Adapter and primer sequences used for DSSS library preparation. The 3' adapters are modified at their 3' end with an idT (inverted deoxythimidine) moiety, in order to prevent ligation with any phosphorylated 5' end.

Single-read DSSS	
5' adapter	5' GUUCAGAGUUCUACAGUCCGACGAUC 3'
3' adapter	5' P-UCGUAUGCCGUCUUCUGCUUGUidT 3'
RT primer	5' CAAGCAGAAGACGGCATACGA 3'
PCR primer GX1	5' CAAGCAGAAGACGGCATACGA 3'
PCR primer GX2	5' AATGATACGGCGACCACCGACAGGTTCAGAGTTCTACAGTCCGA 3'
Sequencing primer	5' CGACAGGTTCAGAGTTCTACAGTCCGACGATC 3'
Paired-end DSSS	
5' adapter	5' ACACUCUUUCCCUACACGACGCUCUUCCGAUCU 3'
3' adapter	5' P-AGAUCGGAAGAGCGGUUCAGCAGGAAUGCCGAGidT 3'
RT primer	5' CTCGGCATTCCTGCTGAACCGCTCTTCCGATC 3'
PE primer 1.1	5' AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT 3'
PE primer 2.0	5' CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT 3'
Sequencing primer Read 1	5' ACACTCTTTCCCTACACGACGCTCTTCCGATCT 3'
Sequencing primer Read 2	5' CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT 3'