RNA-seq analysis of *Pseudomonas* S5 genes associated with motility - Supplementary materials

As a supplement to the main text, we present in this document further investigations of the *Pseudomonas* S5 RNA-seq data. All analyses were conducted in R (R version 3.3.0 (2016-05-03)). We used the following packages: edgeR (Robinson et al., 2009), phia (De Rosario-Martinez, 2015) and vegan (Oksanen et al., 2016) for the statistical analyses, genoPlotR (Lionel et al., Kultima, and Andersson, 2010), ggplot2 (Wickhamm, 2009), pheatmap (Kolde, 2015) and VennDiagram (Chen, 2016) for the graphics, dplyr (Wickham and Francois, 2015), knitr (Xie, 2013), magrittr (Bache and Wickham, 2014) and reshape2 (Wickham, 2007)for data manipulation. The script from which this document is generated as well as additional Perl scripts used during the analysis are given at the end of this document.

Quality assessment and data exploration

Histograms count data (after log-normalization)

We checked first the distribution of the counts. We presented here the histograms after log-normalization of count data (histograms of RPKM values not shown, but available in the script). After log-normalization, the counts data seem approximately normally distributed.

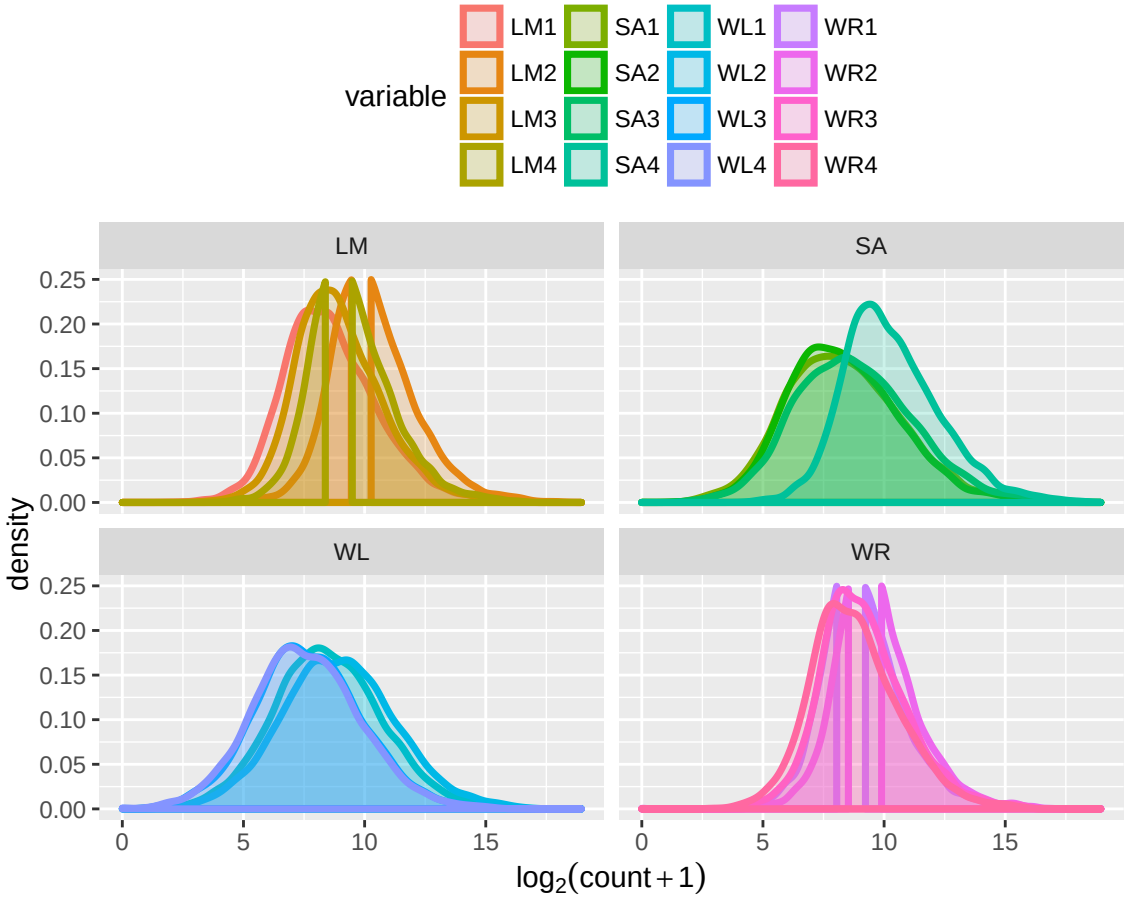


Figure S1. Density plot of log-normalized count data for the four experimental conditions.

Biological coefficient of variation

We use the plotBCV function "which shows the root-estimate, i.e., the biological coefficient of variation for each gene" (Chen et al. 2015) to plot the genewise biological coefficient of variation (BCV) against gene abundance (in log2 counts per million).

The y-axis represents the BCV. This latter is "the coefficient of variation with which the (unknown) true abun-dance of the gene varies between replicate RNA samples. It represents the CV that would remain between biological replicates if sequencing depth could be increased indefinitely. [...] [It] is reasonable to suppose that BCV is approximately constant across genes." (Chen et al., 2015). The black dots allow to appreciate the dispersion across reads (tags). Witch BCV plots, "estimation of genewise BCV allows observation of changes for genes that are consistent between biological replicates and giving less priority to those with inconsistent results" (Diray-Arce et al., 2015).

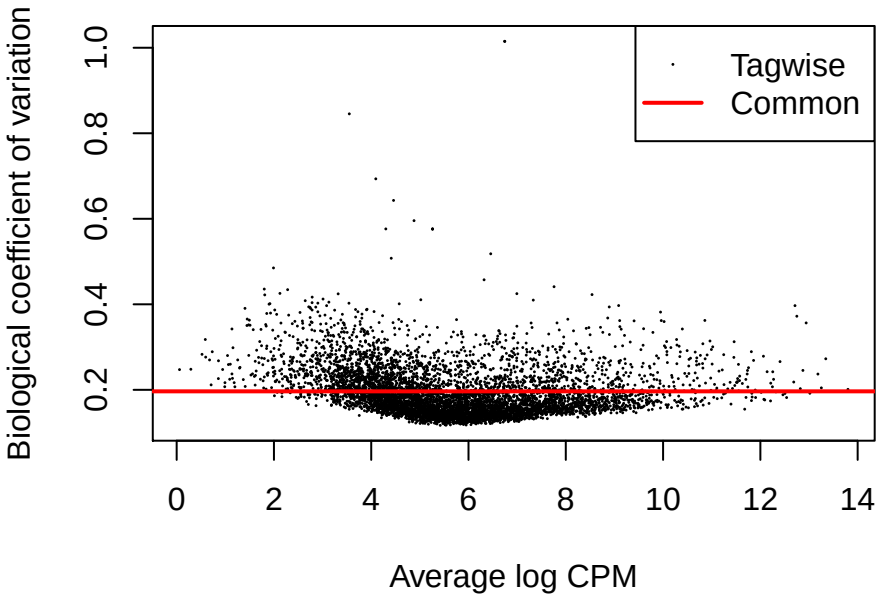


Figure S2. Plot of biological coefficient of variation.

Multidimensional scaling plot of distance between expression profiles

We used here the plotMDS function. This latter plots samples on a two-dimensional scatterplot so that distances on the plot approximate the expression differences between the samples. It "produces a plot in which distances between samples correspond to leading biological coefficient of variation (BCV) between those samples" (Chen et al. 2015).

Here, we could also check that the replicates for a given condition cluster well together. This is mostly the case, except for the replicate "SA4" that seems more distinct that the three other SA replicates.

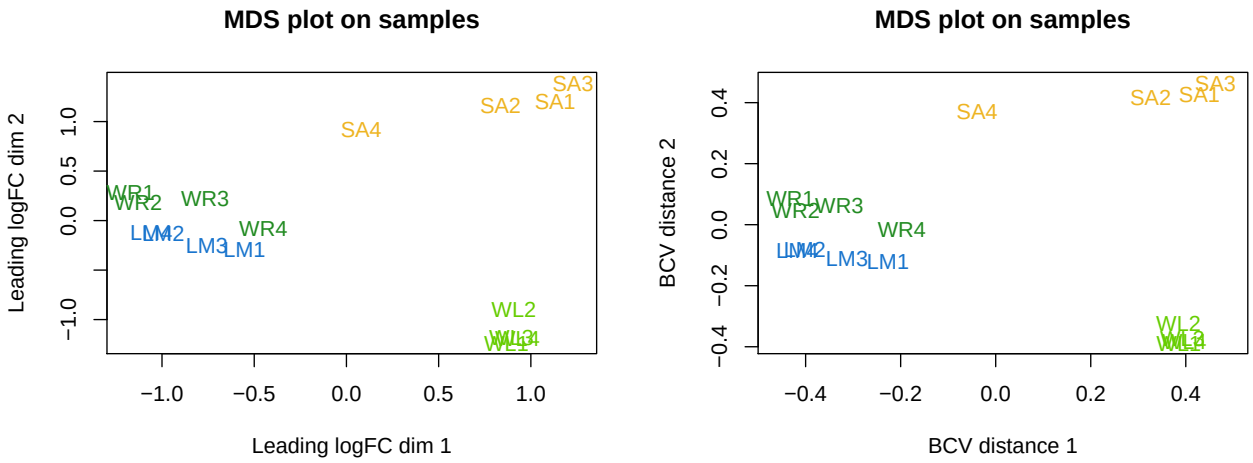


Figure S3. MDS plots for logFC (left) and BCV (right).

Multivariate analyses

PCA

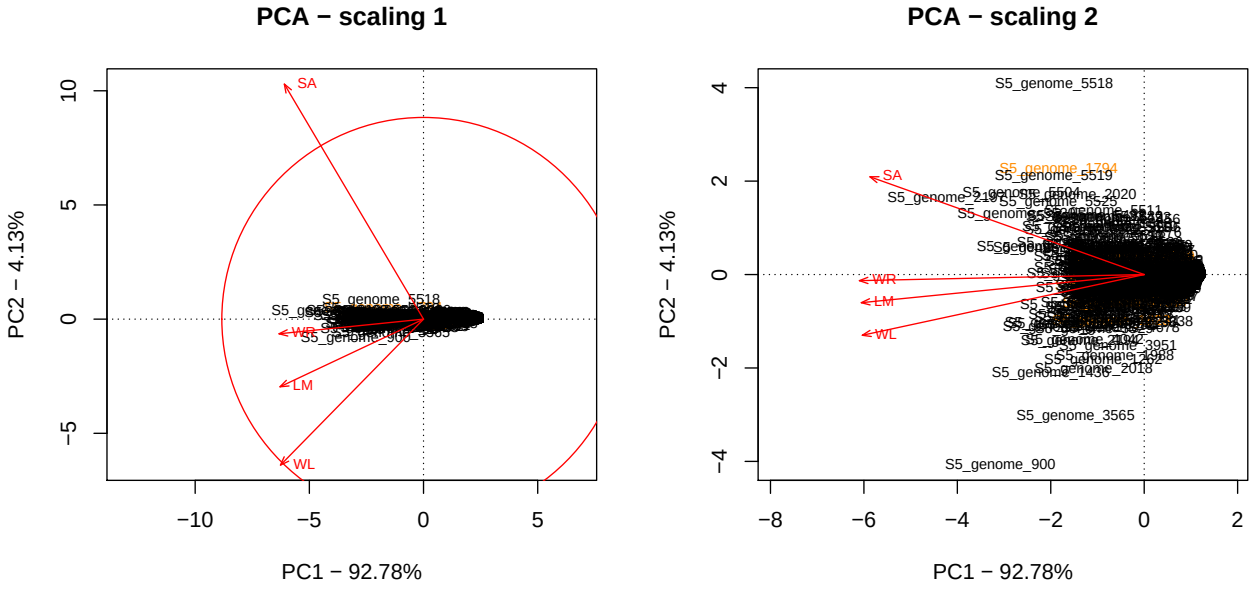


Figure S4. PCA plots for all genes and all conditions (mean data). Left: scaling 1 (angles are meaningless), right: scaling 2 (distances are meaningless).

PCA by condition (with coloured motility-associated genes)

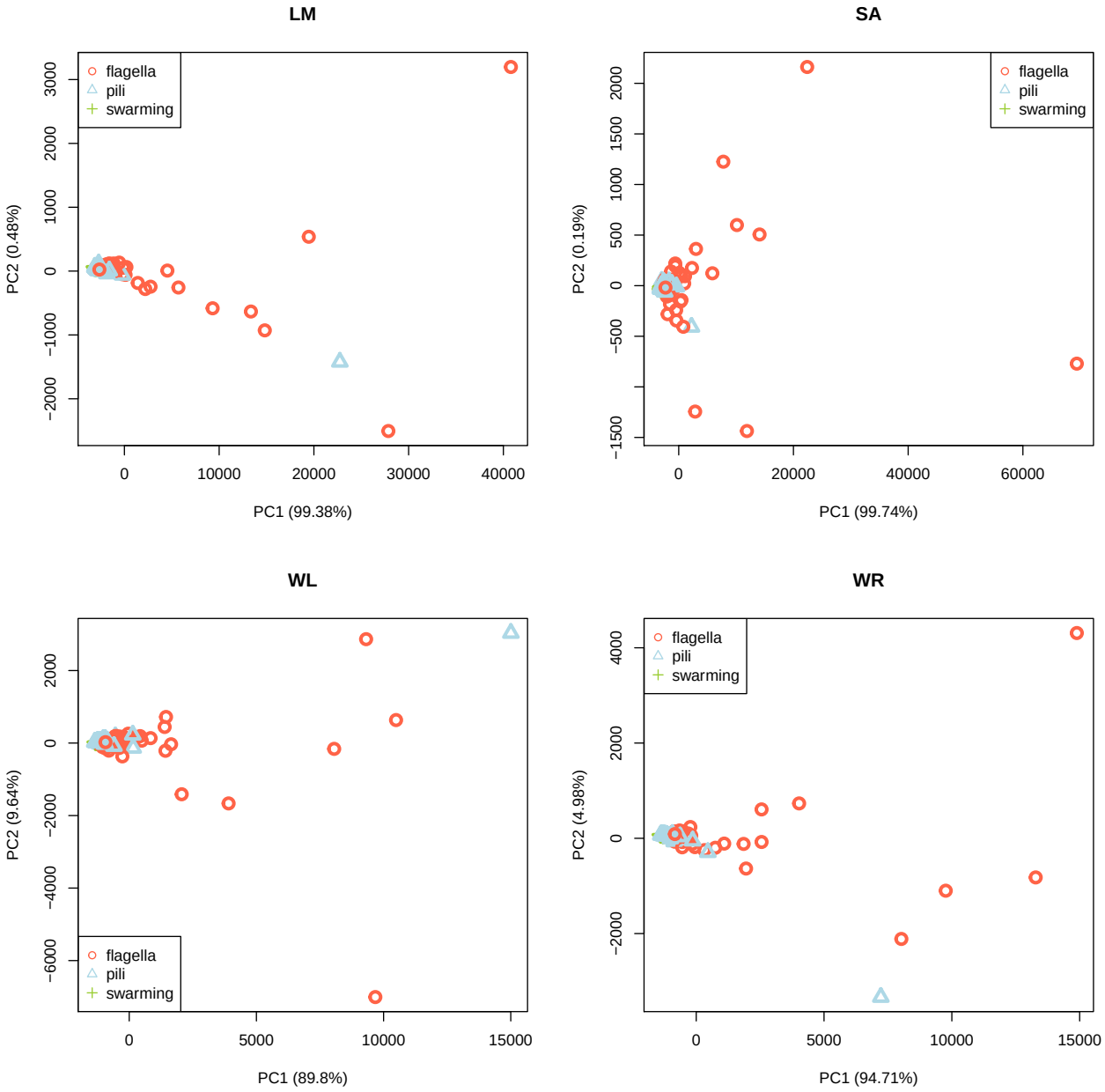


Figure S5. PCA plots for motility-associated genes only for all conditions separately.

Heatmap

Here we looked at the gobal level of expression across all replicates conditions of motility-associated genes. We observed that the replicates of a given experimental condition do not necessarily cluster together.

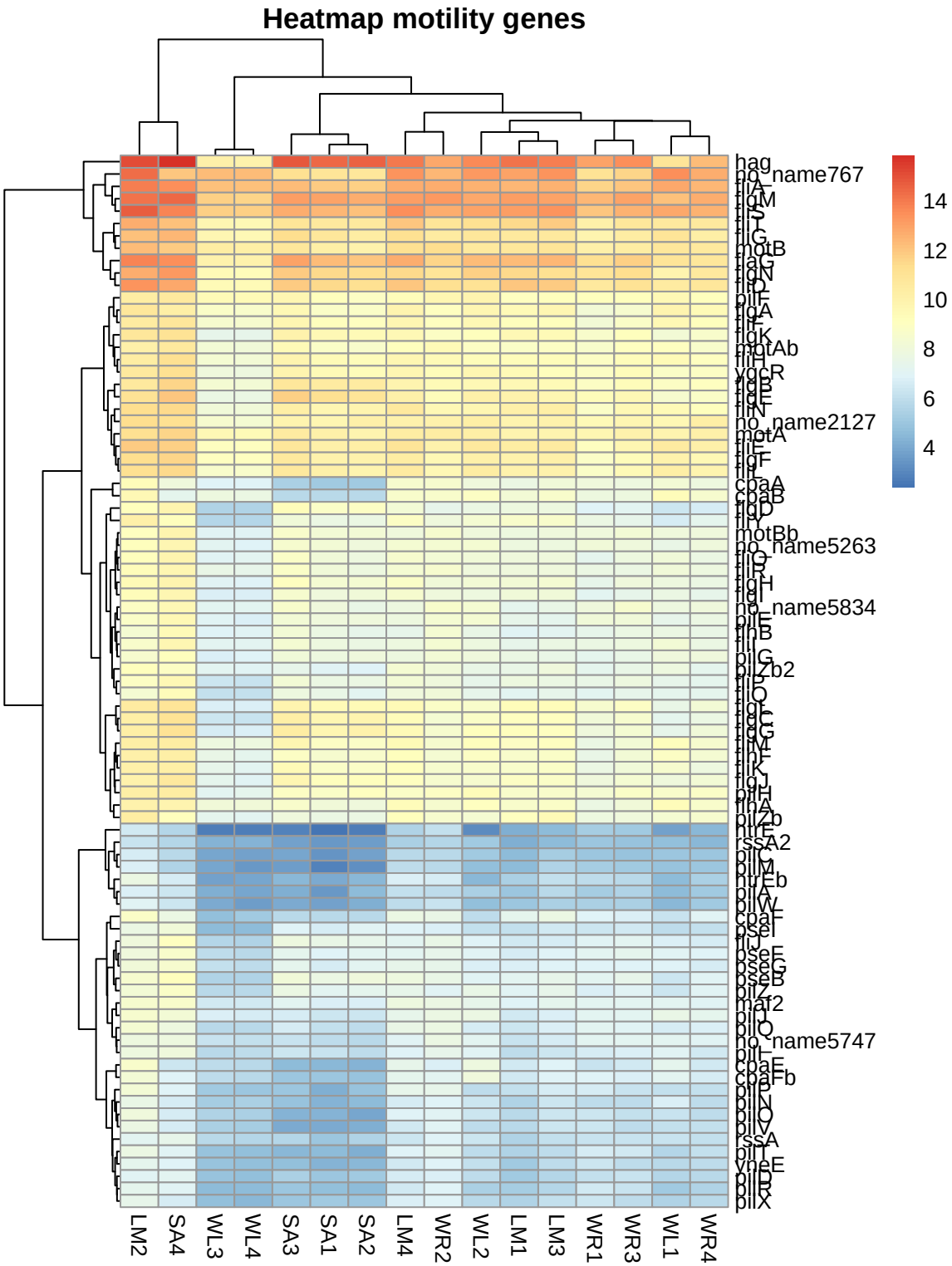


Figure S6. Heatmap for counts data for all replicates (motility-associated genes only).

Boxplot RPKM (without and with chemotaxis-associated genes)

Here, we compared the RPKM between all conditions (and also between all replicates). We also included chemotaxis genes (5 *che* family genes found by searching in the data frame exported from GenDB), to see if they exhibit the same pattern of expression level as one kind of motility (plots on the right here below).

We noticed that the level of expression (log of RPKM values) is quite homogeneous for the replicates of a given condition. Interestingly, the genes involved in chemotaxis seem more expressed in conditions where plant material is present, consistent with plant-oriented motility (as discussed in the main text). Further investigations would be needed (e.g. statistical tests, include more chemotaxis genes).

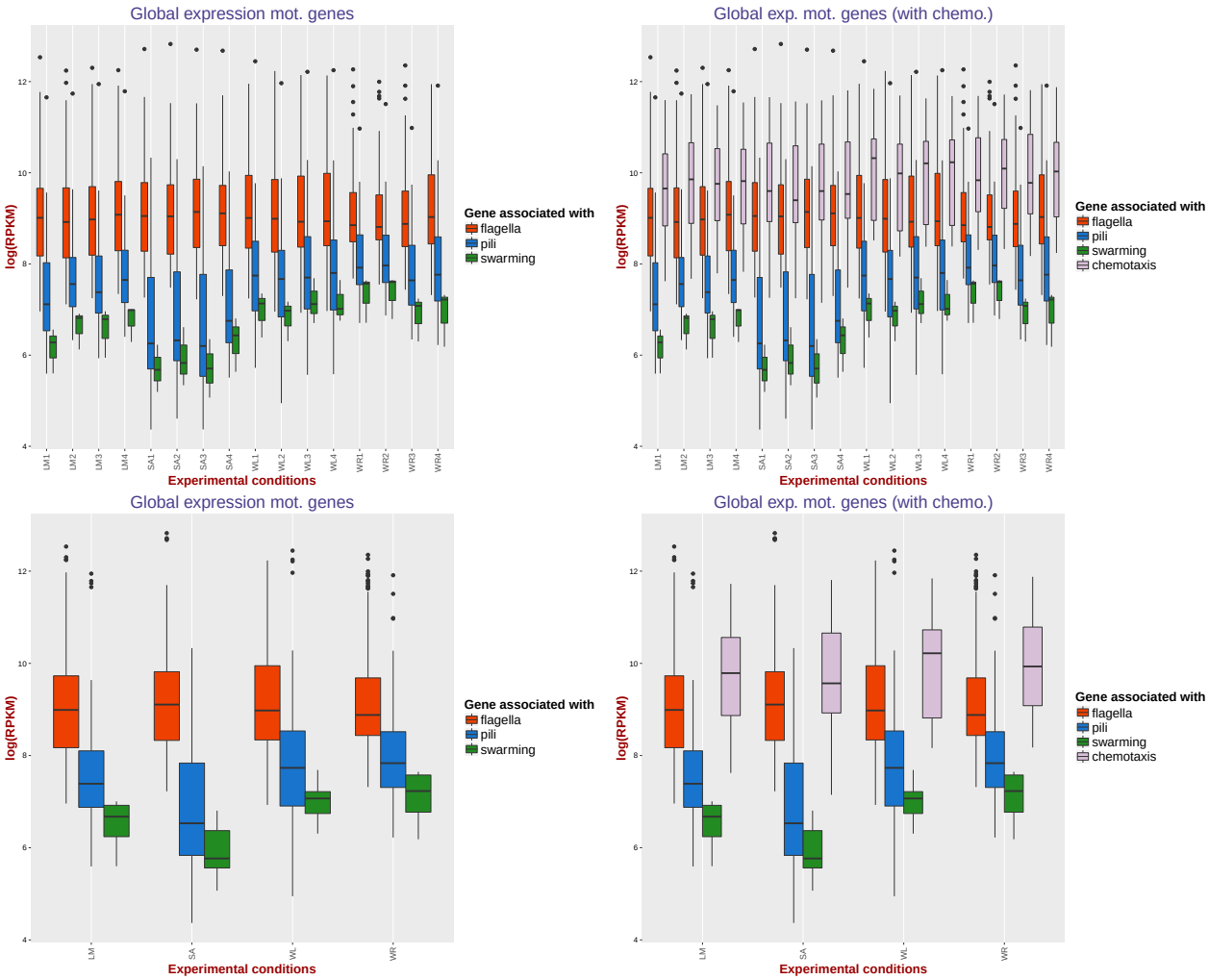


Figure S7. Boxplots of the log of RPKM values by motility type, for all replicates (top) or for the four conditions (bottom), without (left) and with (right) chemotaxis-associated genes.

Differential expression

In the same way as for the main text, we considered for these analyses only the genes exhibiting a statistically significant change in differential expression (adjusted p-values < 0.05).

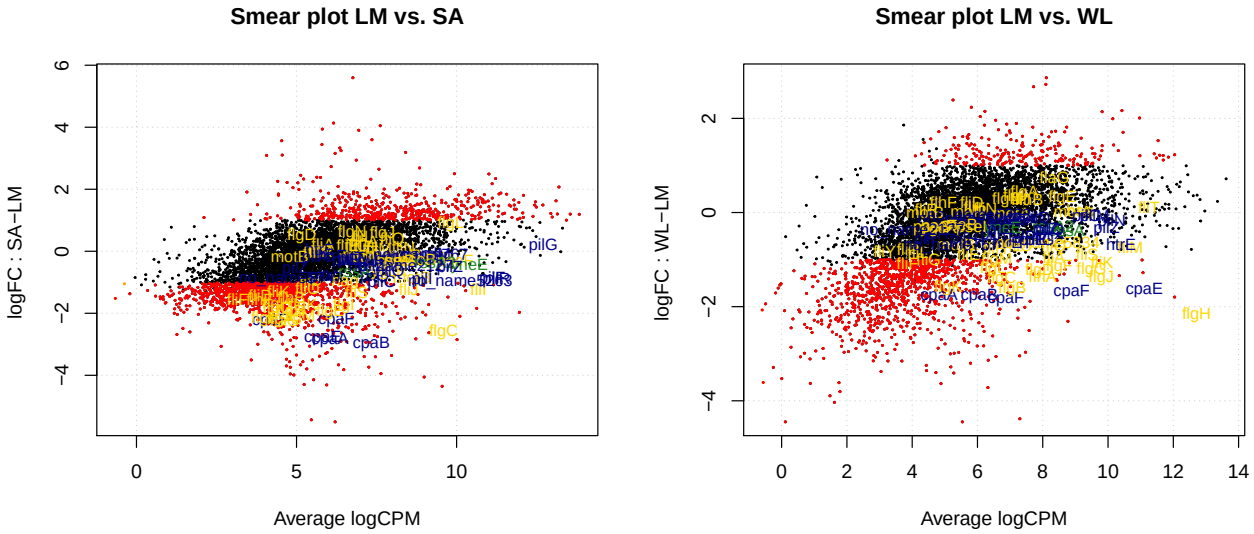
Histogram of p-values and plots logFC vs. logCPM (M vs. A)

MA plot: plot the log-fold change (i.e. the log of the ratio of expression levels for each gene between two experimental groups) against the log-concentration (i.e. the overall average expression level for each gene across the two groups).

Here, we drew "smear plots" (average logCPM in x-axis, logFC in y-axis) for all pairs of comparisons. We added to the plots the label of the motility-associated genes that we annotated (see figure legend). Please notice that the y-axis is not always on the same scale.

As they are neither particularly informative nor conclusive, histograms of adjusted p-values are not shown here (but the code is available in the R script).

On the smear plots here below, we noticed global variations of the change in gene expression. For example, it seems that gene expression varies slightly in LM vs. WR (less red dots). When root material is present (SA vs. WR, WL vs. WR), a global trend of upregulation is visible (more red dots in the upper part of the plot). It seems to be the opposite ("global" downregulation) in LM vs. WL. Broadly, we observed that the genes we annotated are not the ones that exhibit the most important changes in gene expression (not the highest on the y-axis) and have a broad-ranged level of expression (from middle to right part of the cloud of points). For the motility-associated genes, no clear trends emerge from these smear plots.



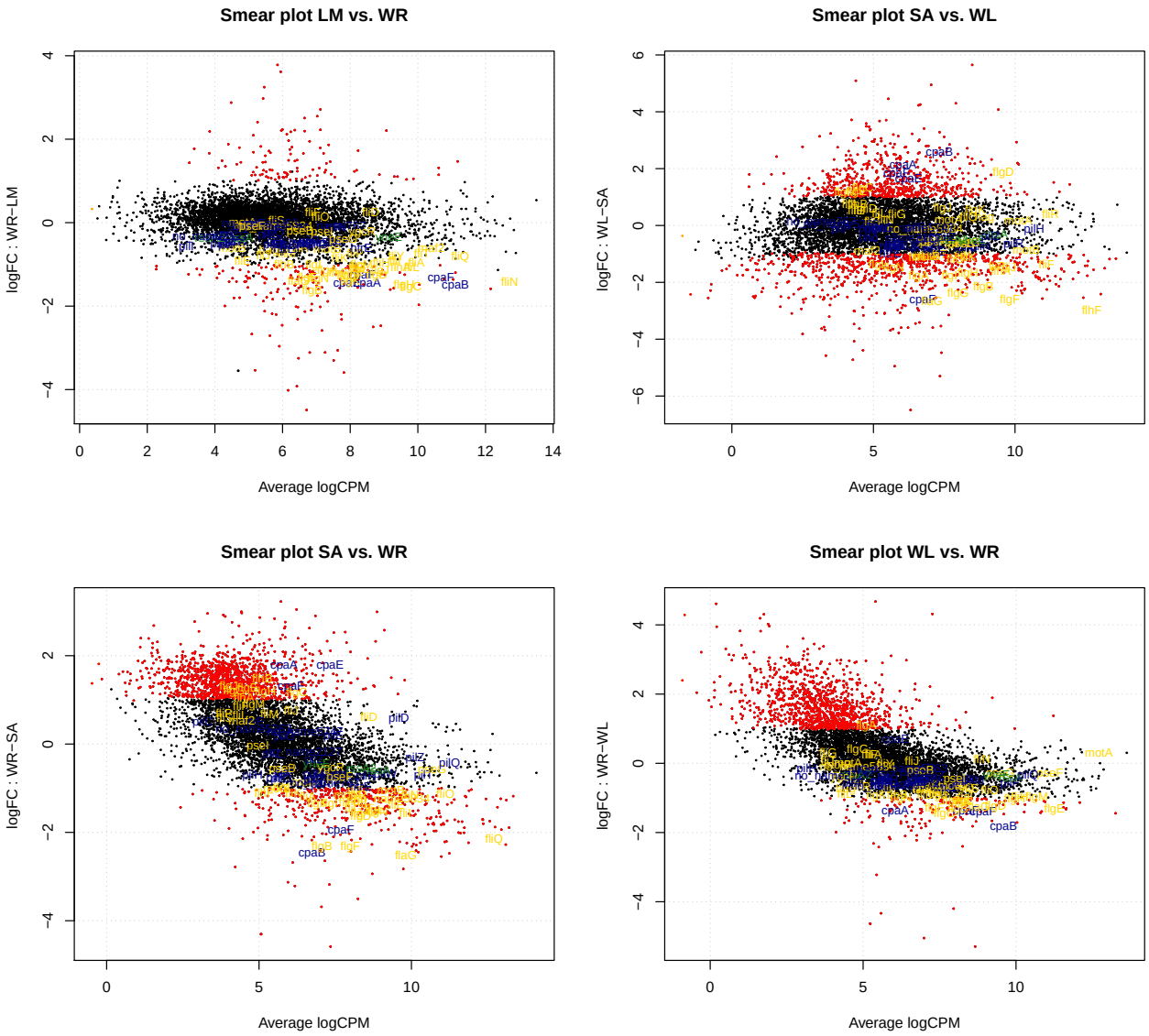


Figure S8. Smear plots for differential expression between all pairs. Genes with more than twofold change of expression shown in red. Motility-associated genes (labelled) shown in orange (flagellum-related), blue (pilus-related) and green (swarming-related). Please notice the different scales of the y-axis.

Scatterplot matrix: correlation between differential expression pairs

We drew scatterplot matrix to compare the differential expression between pairs of pairwise comparisons (motility-associated genes only). We noticed that change in differential expression is sometimes highly correlated (e.g. WR vs. SA and WL vs. SA or SA vs. LM and WL vs. SA), and sometimes not (e.g. SA vs. LM and WL vs. LM or WR vs. LM and WR vs. WL). As explained in the main text, we used this plot to decide which comparison to examine more in detail.

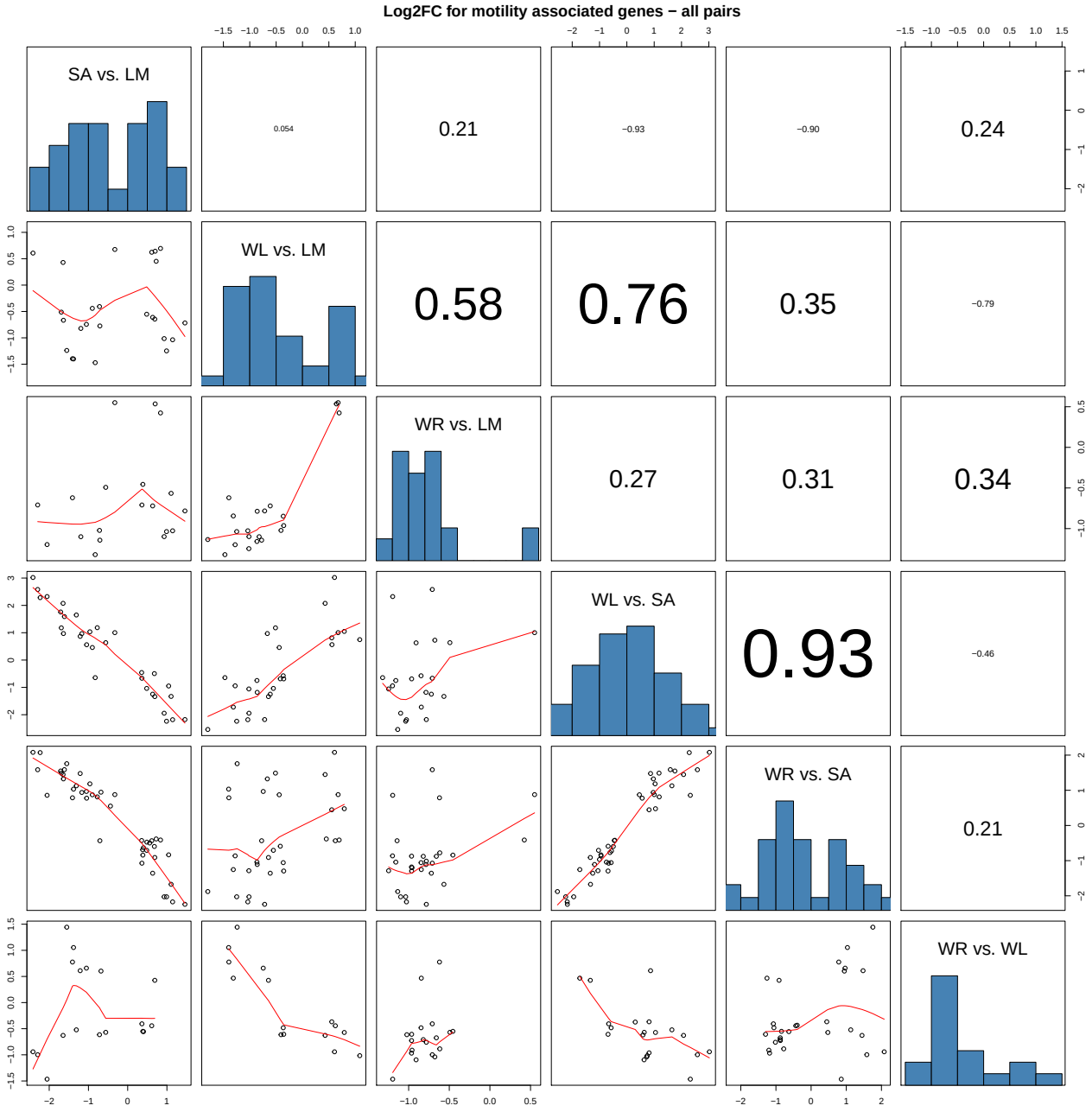


Figure S9. Scatterplot matrix: correlation between differential expression between pairs of conditions (motility-associated genes only.

Heatmap for all pairs of comparisons

Here, we used a heatmap for visualization of differential expression in all pairs of comparisons. We noticed that the profile of differential expression is sometimes very similar (e.g. SA vs. wR or SA vs. WL). For all tests of differential expression, we only retained the genes for which the adjusted p-value was below 0.05 (hence the grey cases).

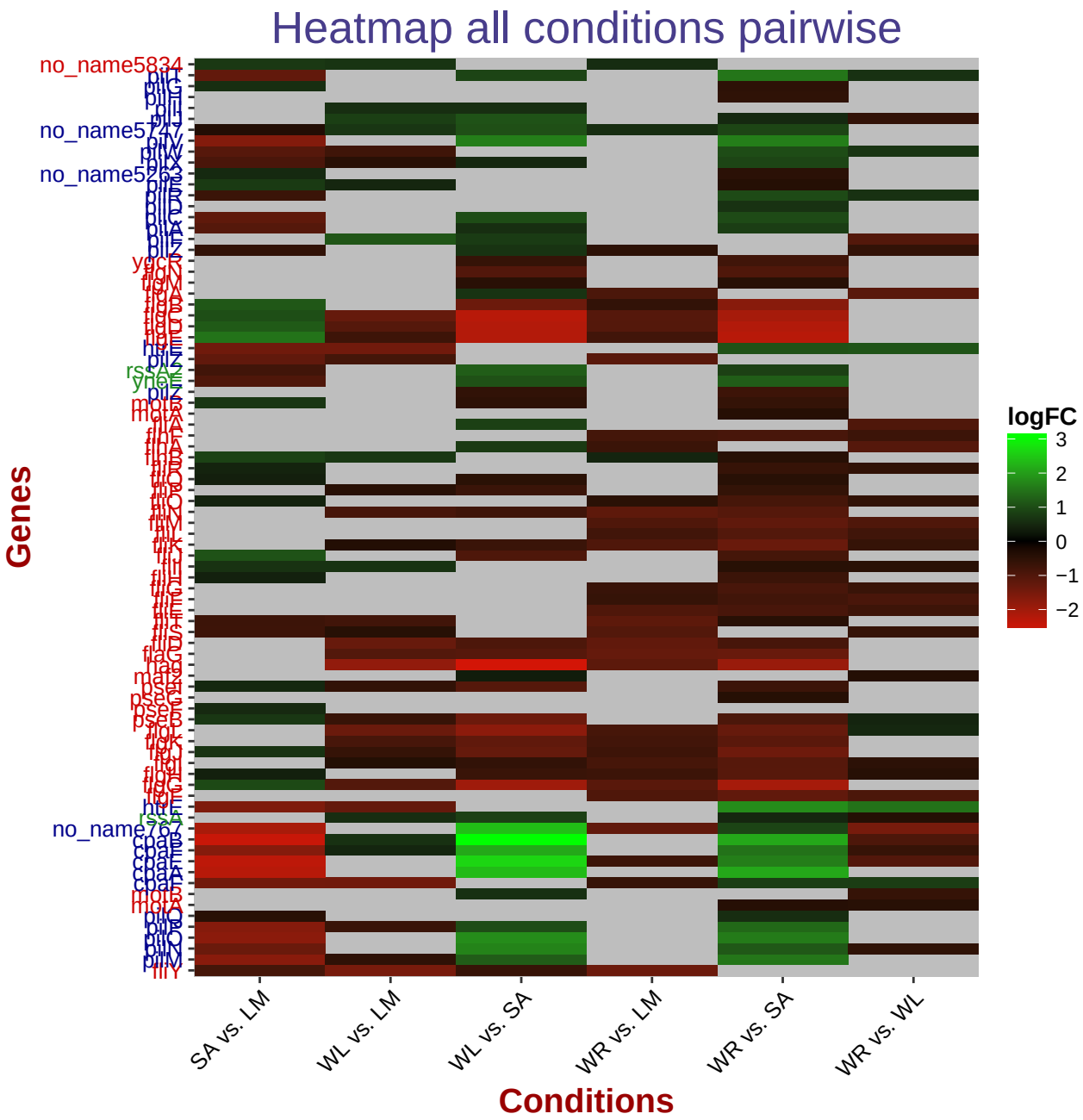
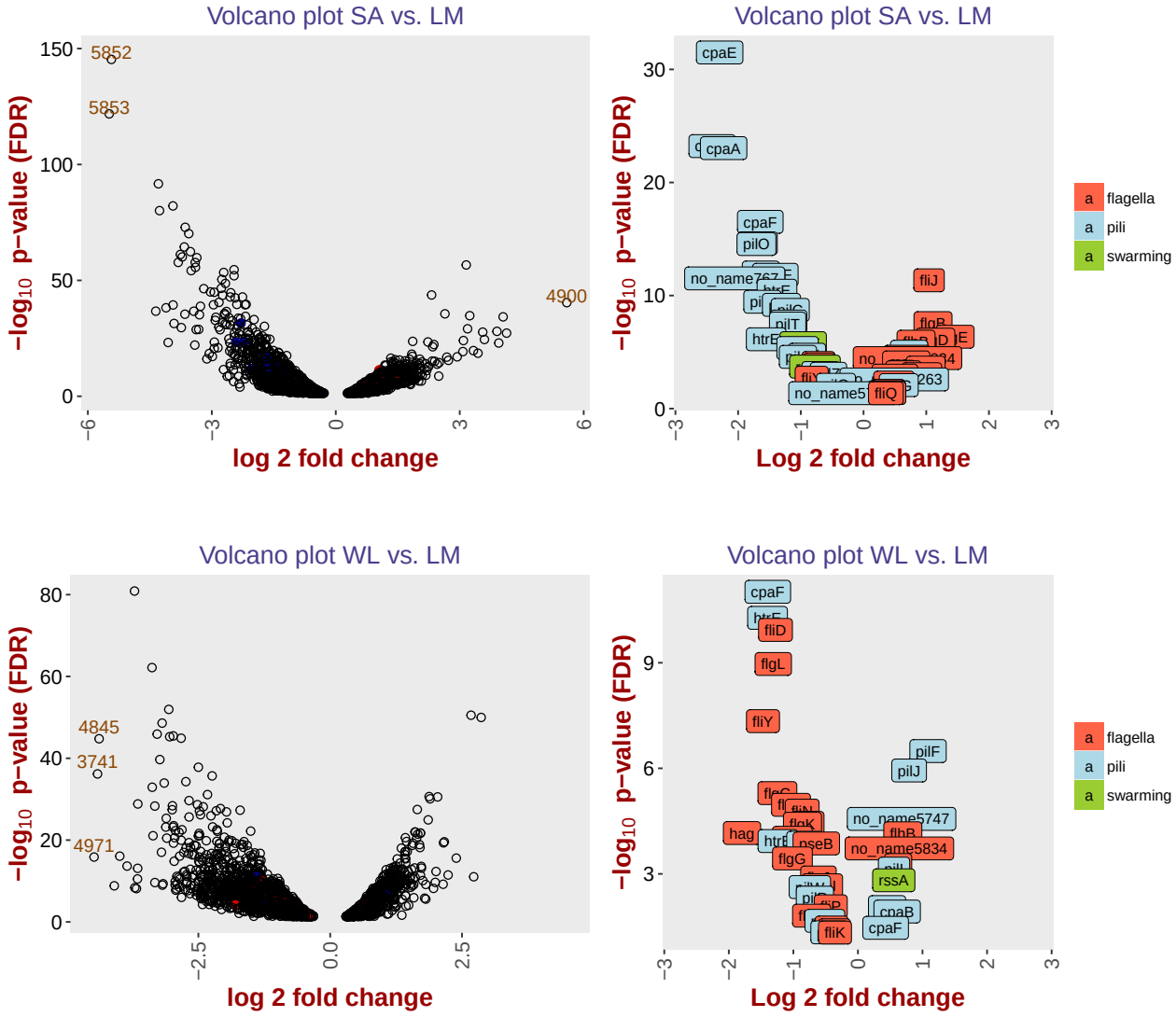


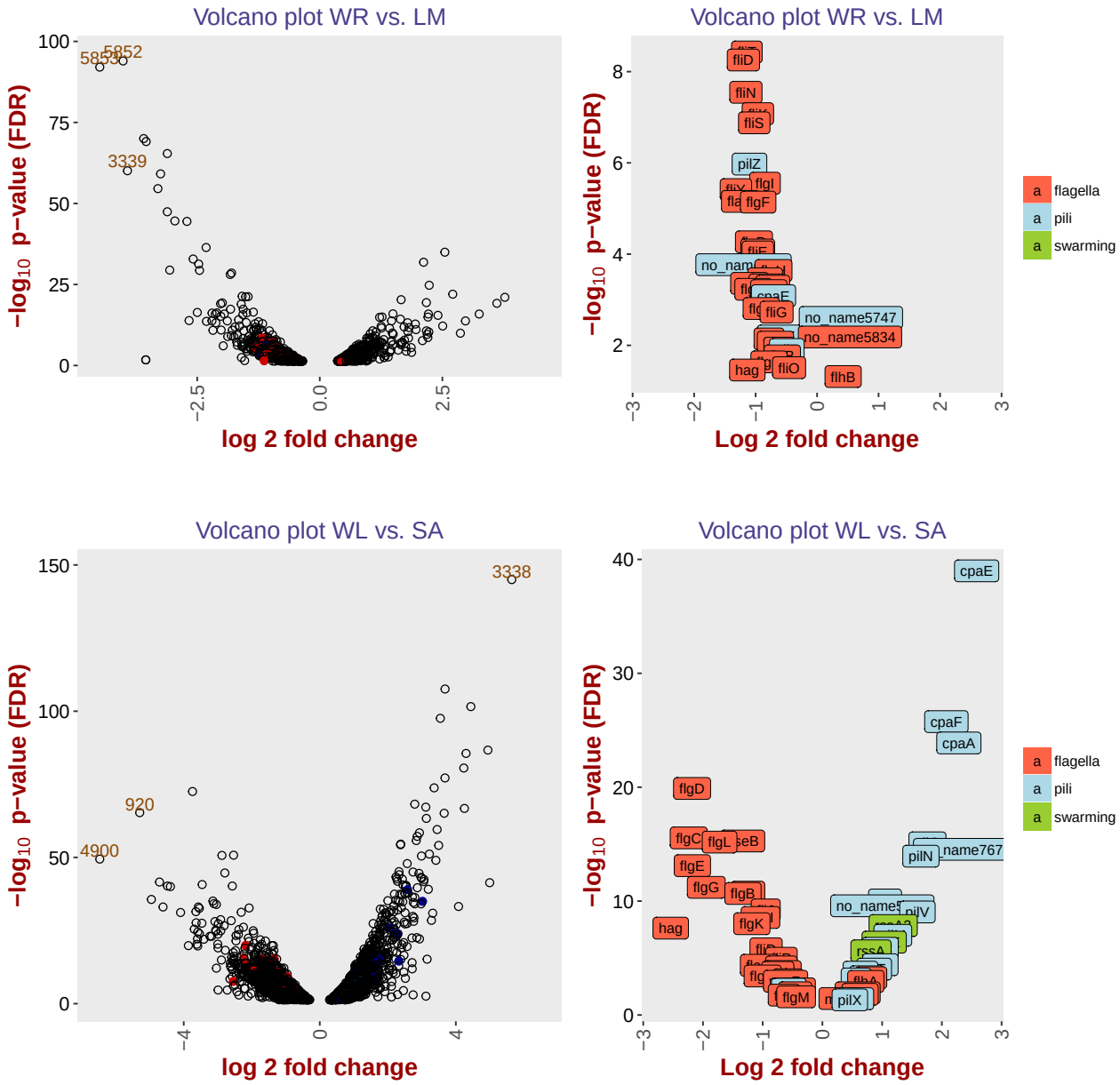
Figure S10. Heatmap of differential expression for all pairs of conditions (motility-associated genes only; only statistically significant genes with adjusted p-values < 0.05 are shown). Y-axis: genes are ordred according to their position on the chromosome.

Volcano plots: all genes and motility-associated genes

Next, we drew the volcano plots for all pairs of conditions. They provide more precise information than the heatmap. Because of time limitation, we could not discuss all pairs of conditions. We observed nonetheless that motility-associated genes are not the genes that exhibit the most important changes in expression (left column). The three genes with the most changing expression are labelled (e.g. 3353 corresponds to the CDS S5_genome_3353). We used an easy-to-use custom Perl script (provided at the end of this document; blast ouptuts available in the data folder) to investigate quickly the function of these genes (920: siderophore receptor; 3338: cytochrome oxidase; 3353, 5852: transport proteins; 4845: bacterioferritin-associated ferredoxin, 5853: import protein; all others: uncharacterized proteins).

We noticed also that the differential expression of flagella and pili in some cases shows a clear opposite pattern (e.g. SA vs. LM, WR vs. SA; discussed in the main text), although this tendency is not obvious in all pairwise comparisons (e.g. WL vs. LM). In particular, we noticed that the profile of WL vs. SA is very similar to the one of WR vs. SA discussed in detail in the main text. As discussed in the main text, genes specifically associated with swarming more often exhibit the same pattern of differential expression than the one of pilus-associated genes.




Volcano plot WR vs. LM
−log10 p−value (FDR)
log 2 fold change
5852
3339
Volcano plot WR vs. LM
−log10 p−value (FDR)
Log 2 fold change
fliD
fliN
fliS
pilZ
flgI
flgF
fliG
no_name5747
no_name5834
fliO
flhB
hag
a flagella
a pili
a swarming
Volcano plot WL vs. SA
−log10 p−value (FDR)
log 2 fold change
3338
920
4900
Volcano plot WL vs. SA
−log10 p−value (FDR)
Log 2 fold change
cpaE
cpaF
cpaA
flgD
flgE
flgG
flgK
hag
pilN
pilV
flgM
pilX
a flagella
a pili
a swarming

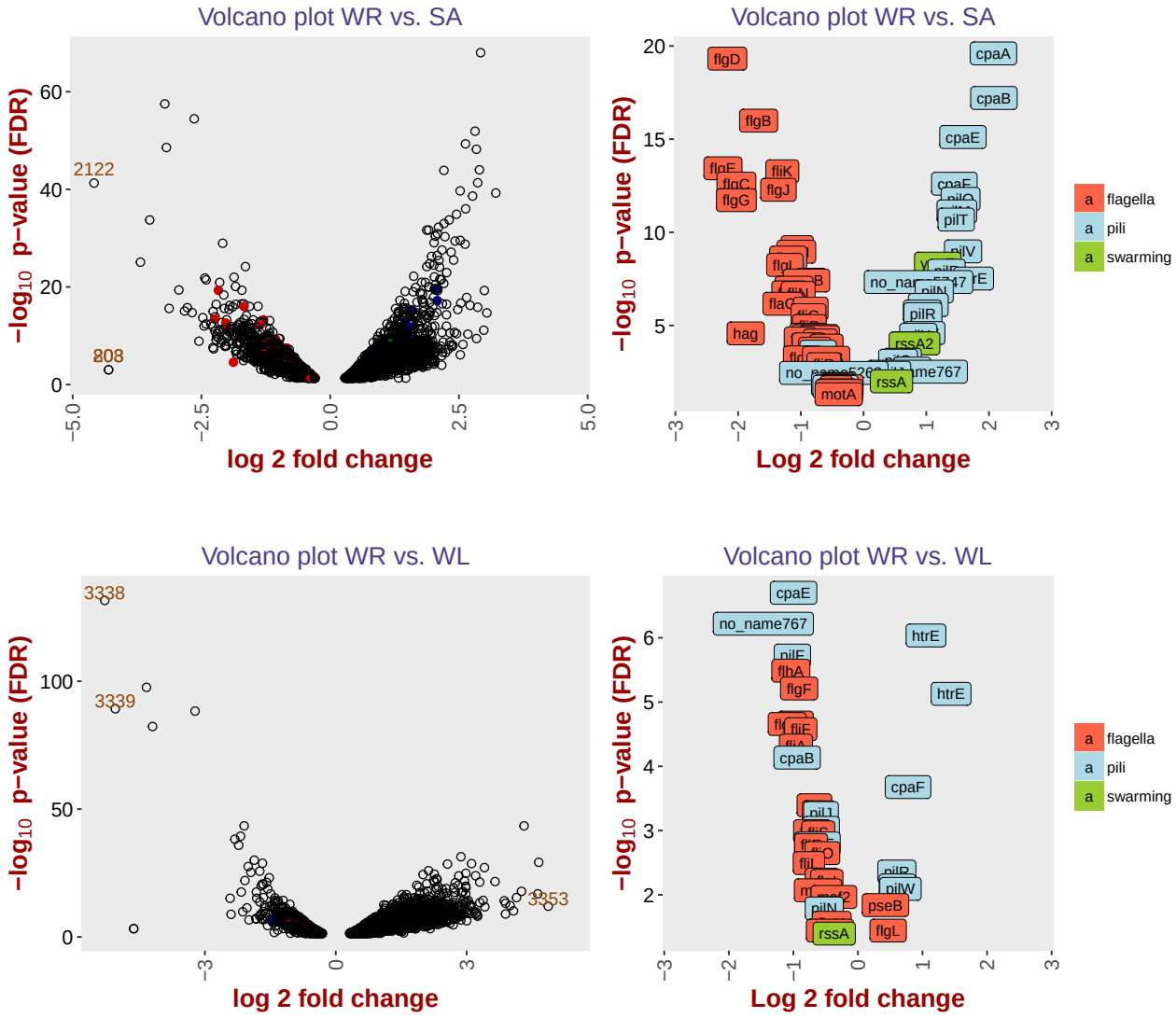


Figure S11. Volcano plots for all pairs of comparisons (left column: all genes differentially expressed in a statistically significant manner (FDR < 0.05); right column: only motility-associated genes).

Up- and downregulation for all pairs

Again, we focused at the up- and downexpression for all pairs of conditions. In fact, these plots show the same information as the volcano plots. Here, it is particularly apparent that the fold change of expression of the motility-associated genes is rarely more than twofold (dashed line).

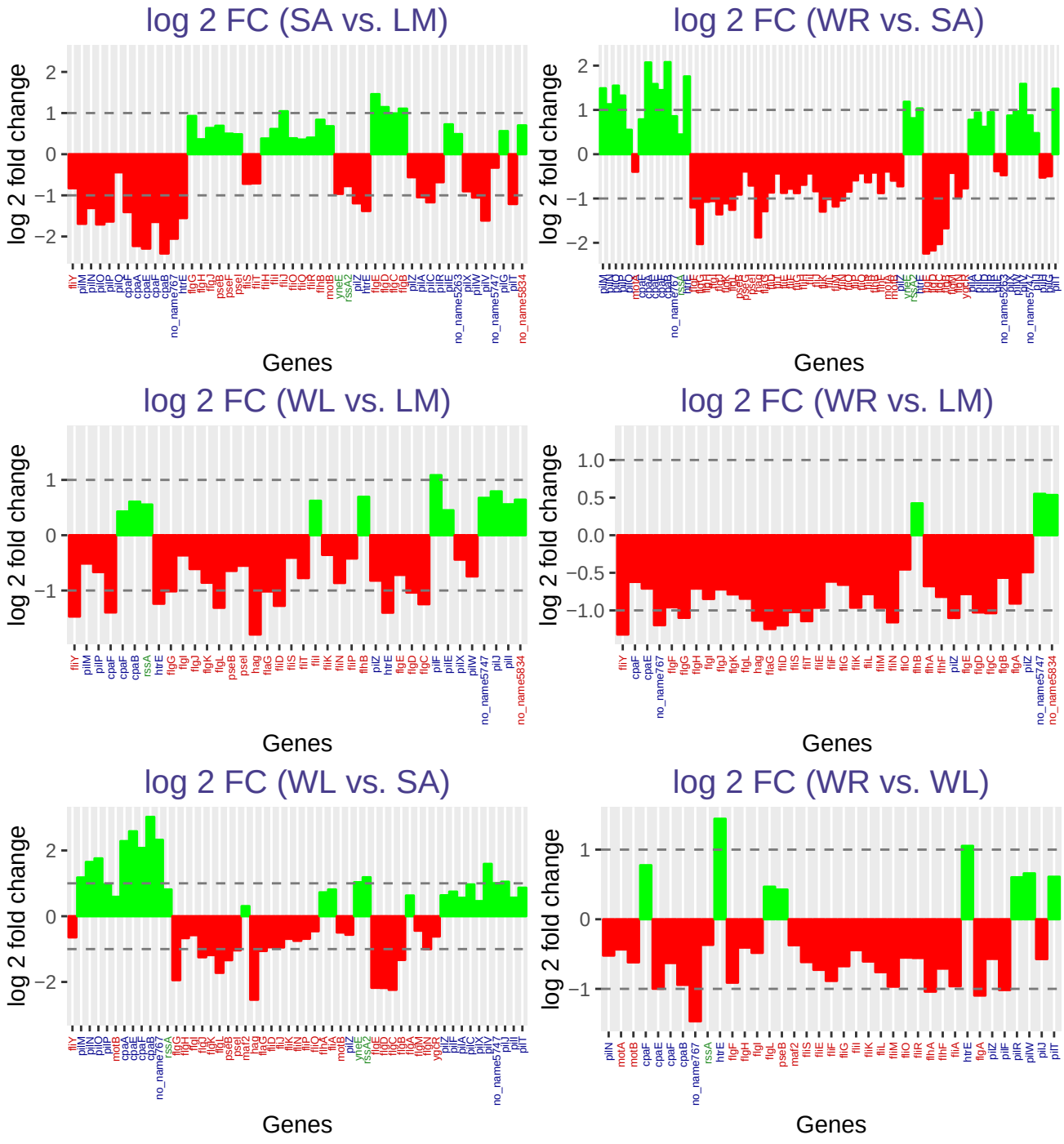


Figure S12. Barplot for all pairs of comparisons. Only motility-associated genes are shown. Dashed line indicating |logFC| = 1 (twofold change of expression). X-axis: genes ordered according to their chromosomal position.

Association between gene expression and other gene characteristics

GC content, purine content and gene length

Here we tried to see if some characteristics (GC content, purine content and length of the genes; computed with a short Perl script provided at the end of this document) of the genes could explain their level of expression (expressed in log of RPKM). We noted a clear inverse correlation between the GC content and the expression level as well as between the length of the gene and the expression level (assessed using Spearman's correlation coefficient). This correlation is stronger for the third codon position than for the first two codon positions (see plots and table below). GC content has already been reported to be associated with gene expression in other species and phyla, e.g. neem (Krishnan et al., 2011), chicken (Rao et al., 2013) or human (Vinogradov, 2005). But technological biases should not be overlooked. In our case, we do not exactly know which biases could skew our data, but for example it has been reported that "GC-rich and GC-poor fragments tend to be under-represented in RNA-Seq" (Risso et al., 2011).

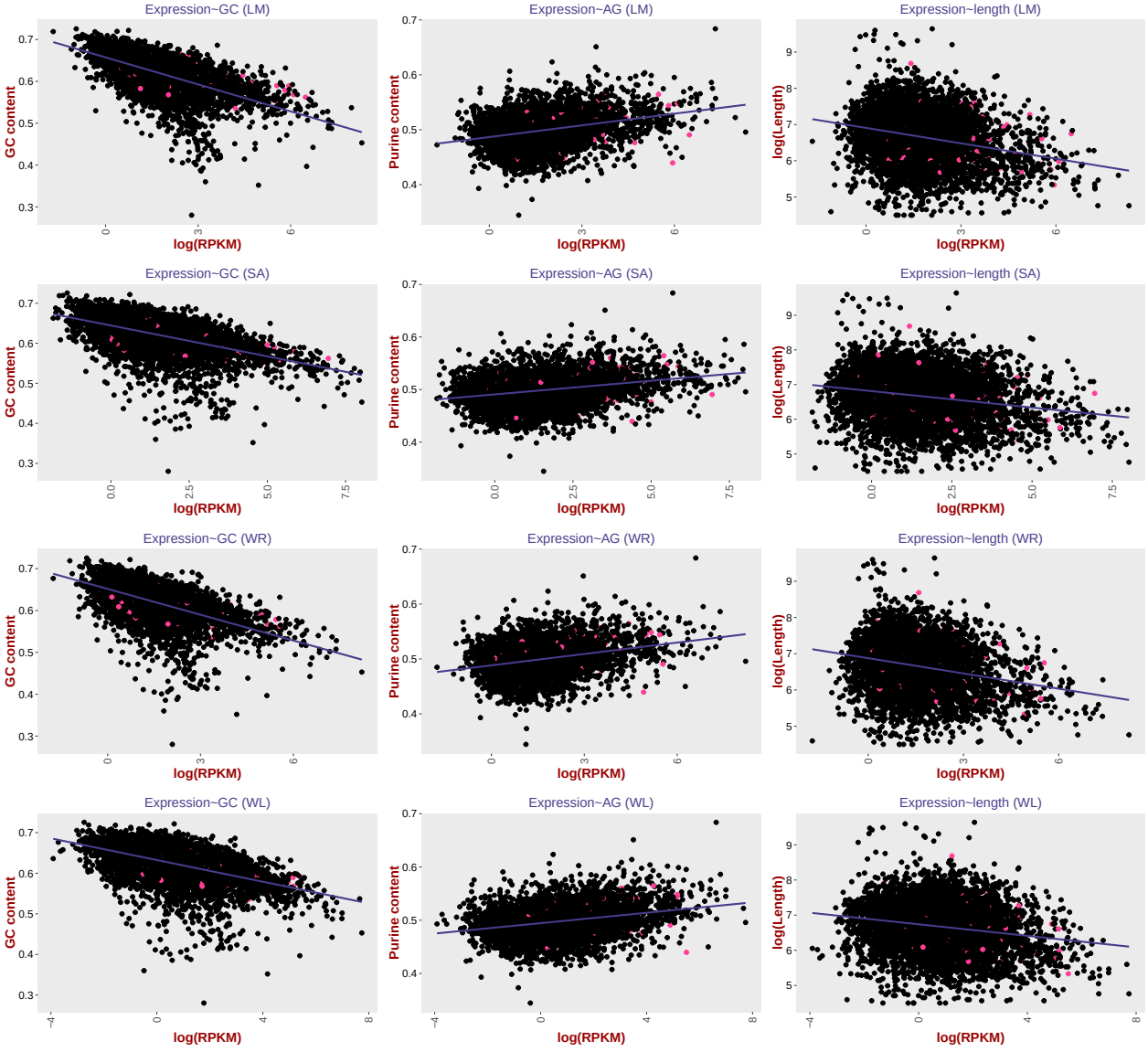


Figure S13. For each condition, plot showing log of RPKM values for all genes against i) GC content of the gene (left column), ii) purine content of the gene (mid column), iii) length of the gene (right column). Motility-associated genes are shown with pink dots.

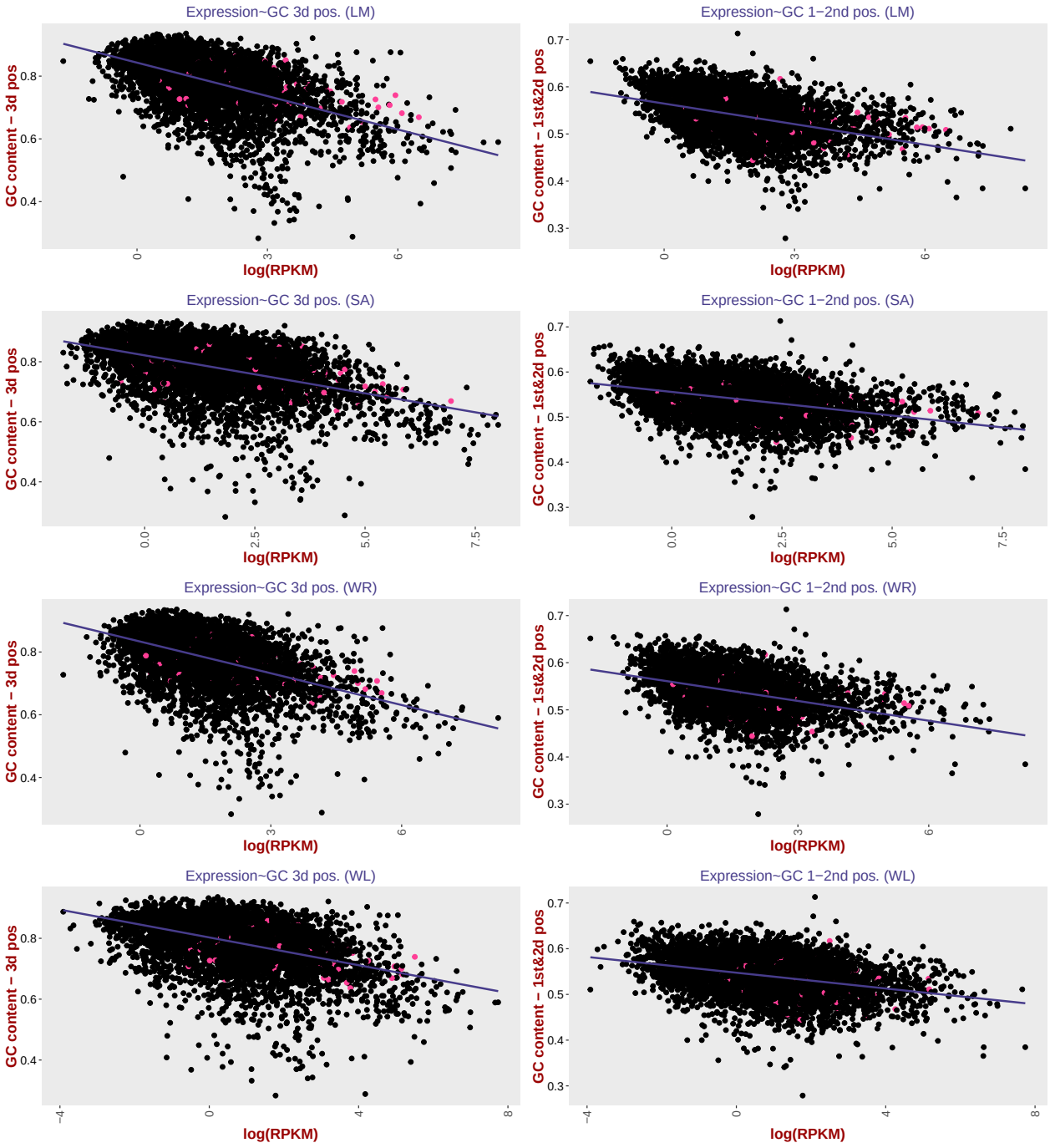


Figure S14. For each condition, plot showing log of RPKM values for all genes against i) GC content at the third codon position (left column), ii) GC content at the first two positions (right column). Motility-associated genes are shown with pink dots.

Correlations between GC content (global, first two codon positions, third codon position) across all conditions:

Correlation	*Spearman's corr. coeff.*	*p-value*
LM ~ GC-content	-0.64	< 2.2e-16
LM ~ GC-content (1st&2d pos.)	-0.44	1.1e-288
LM ~ GC-content (3d pos.)	-0.49	< 2.2e-16
SA ~ GC-content	-0.56	< 2.2e-16
SA ~ GC-content (1st&2d pos.)	-0.39	1.4e-220
SA ~ GC-content (3d pos.)	-0.43	4.2e-274
WL ~ GC-content	-0.51	< 2.2e-16
WL ~ GC-content (1st&2d pos.)	-0.32	1e-148
WL ~ GC-content (3d pos.)	-0.42	6.3e-262
WR ~ GC-content	-0.59	< 2.2e-16
WR ~ GC-content (1st&2d pos.)	-0.42	1.9e-252
WR ~ GC-content (3d pos.)	-0.45	4e-294

Table S15. Results of correlation tests (Spearman's coefficient) between GC content (global, first two positions and third position) and log of RPKM values for all genes for all experimental conditions separately.

After that, we also tried to see if a difference between leading and lagging strand was noticeable. This does not seem to be the case (maybe a slightly higher level of expression for genes on leading ("+") strand).

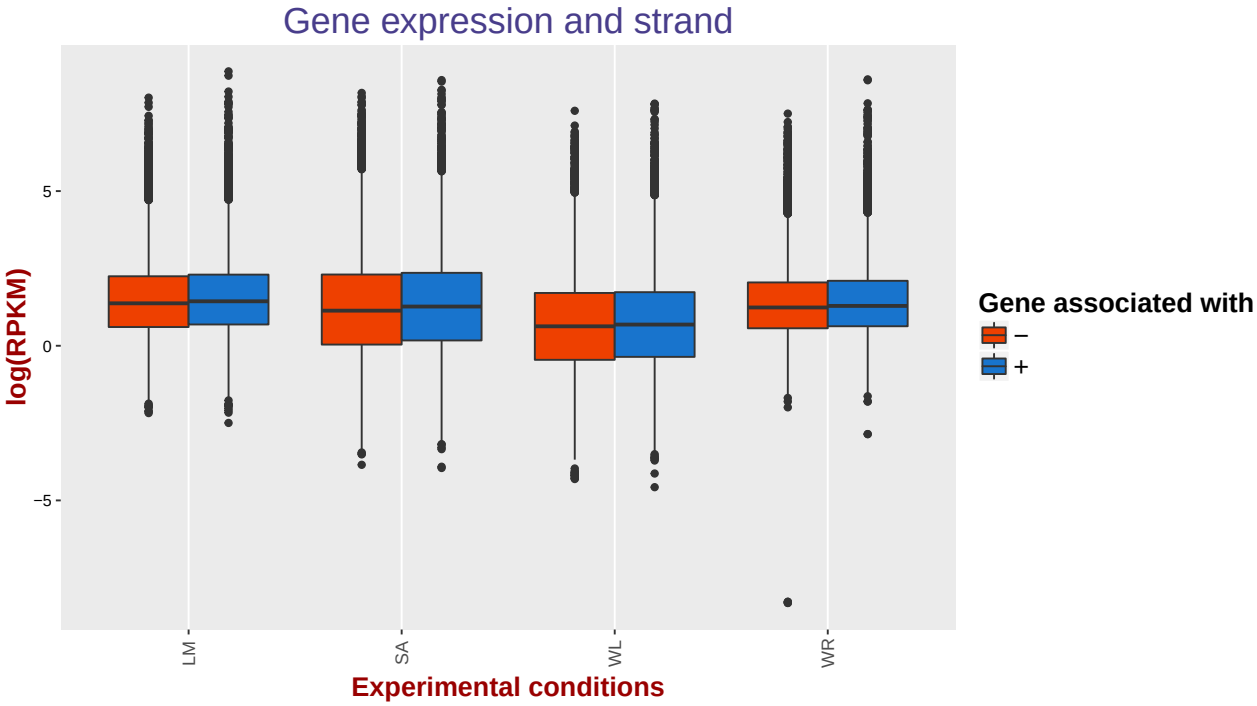


Figure S16. For each condition, plot showing log of RPKM values conditioned by the strand on which the gene is located (this information was not available for all, but for most of the genes).

Multivariate analyses

We also tried to use multivariate tools to visualize the contribution of "structural" parameters to variation of gene expression. We first used a symmetrical method (PCA). Then, we tried an asymmetrical method, redundancy analysis (RDA), that performs a multivariate multiple linear regression followed by PCA (Borcard et al., 2011). We still doubt that this method is appropriate for RNA-seq data. Only a small fraction of expression variation seems to be explained by "structural" parameters (see percents along the axis).

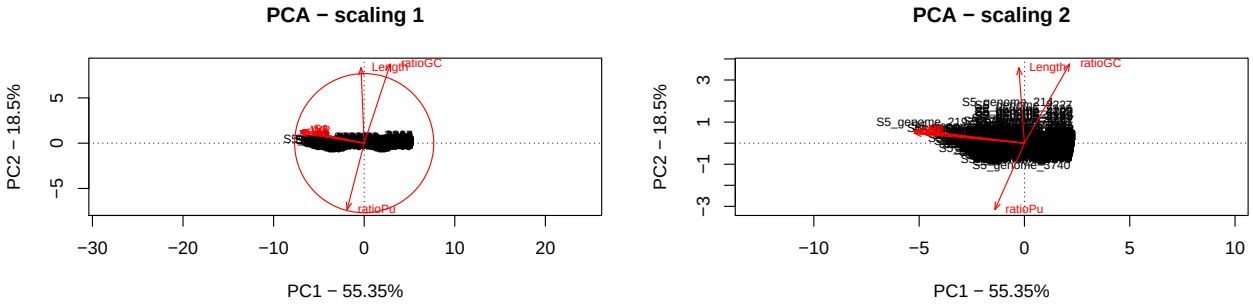


Figure S17. PCA plots for all genes and "structural parameters". Left: scaling 1 (angles are meaningless), right: scaling 2 (distances are meaningless).

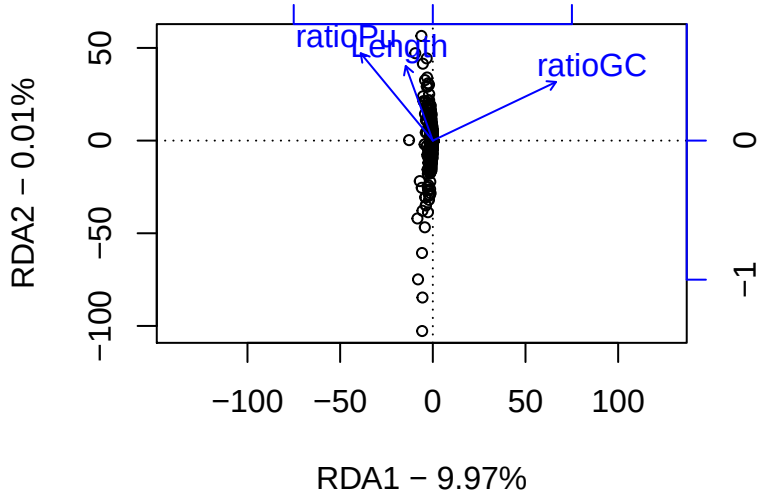


Figure S18. RDA plot of expression values regressed against "structural parameters".

KEGG pathways and GO categories

Here, we retrieved the KEGG pathways of *Pseudomonas fluorescens* Pf5 available on the KEGG database. In the first step, we "matched" the *Pseudomonas fluorescens* Pf5 genes with the ones of our *Pseudomonas* S5 (with BLAT, see Perl script at this end the document; although this is probably not the most optimal solution, it is fast and presumably convenient for explanatory purposes). This allowed us to associate most genes of *Pseudomonas* S5 with a pathway.

For the gene ontology (GO) categories, we did something "on the fly" as another group was already working with the time-consuming BLAST2GO. We retrieved the GO categories for *Pseudomonas aeruginosa* PAO1 genes, as we did not find GO data for the *Pseudomonas fluorescens* Pf5 on the Pseudomonas database (www.pseudomonas.com). We found the orthologous pairs of genes between *Pseudomonas aeruginosa* PAO1 and *Pseudomonas fluorescens* Pf5 genes. Thus we could retrieve GO of a large number of *Pseudomonas fluorescens* Pf5 genes. Then, we could associate GO to our *Pseudomonas* S5 genes as we had already linked *Pseudomonas protegens* Pf5 and *Pseudomonas* S5 genes (as described just here above).

GO categories

We brought together the categories associated with flagella or type IV pili under a "motility" category. We observed that motility is clearly not the most represented category.

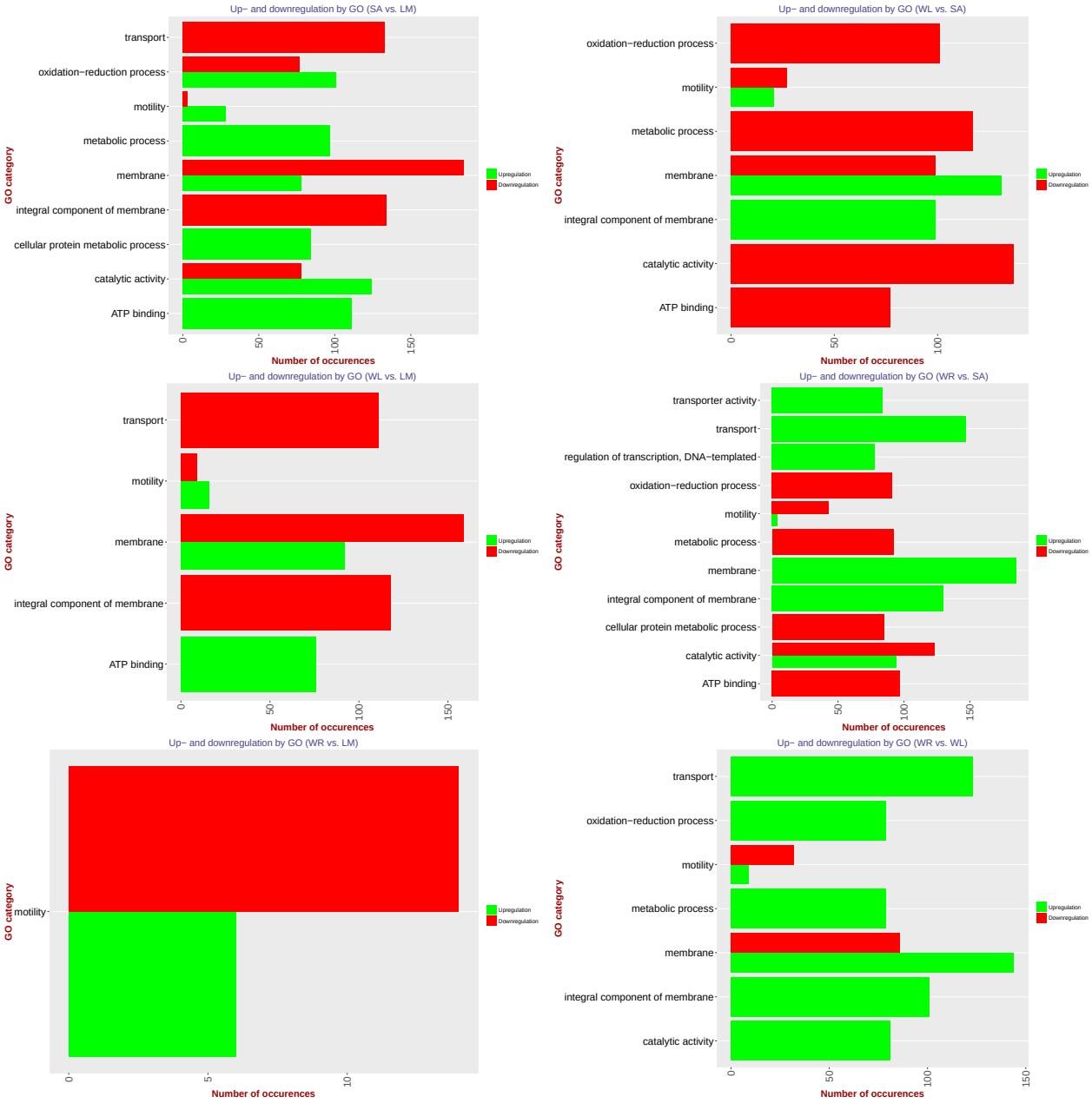


Figure S19. Barplots showing for each pairs of condition to which GO category the up- and downregulated genes belong. Threshold: 75 occurences of the GO category (motility added independently of the number of occurences, as explained in the text).

KEGG pathways

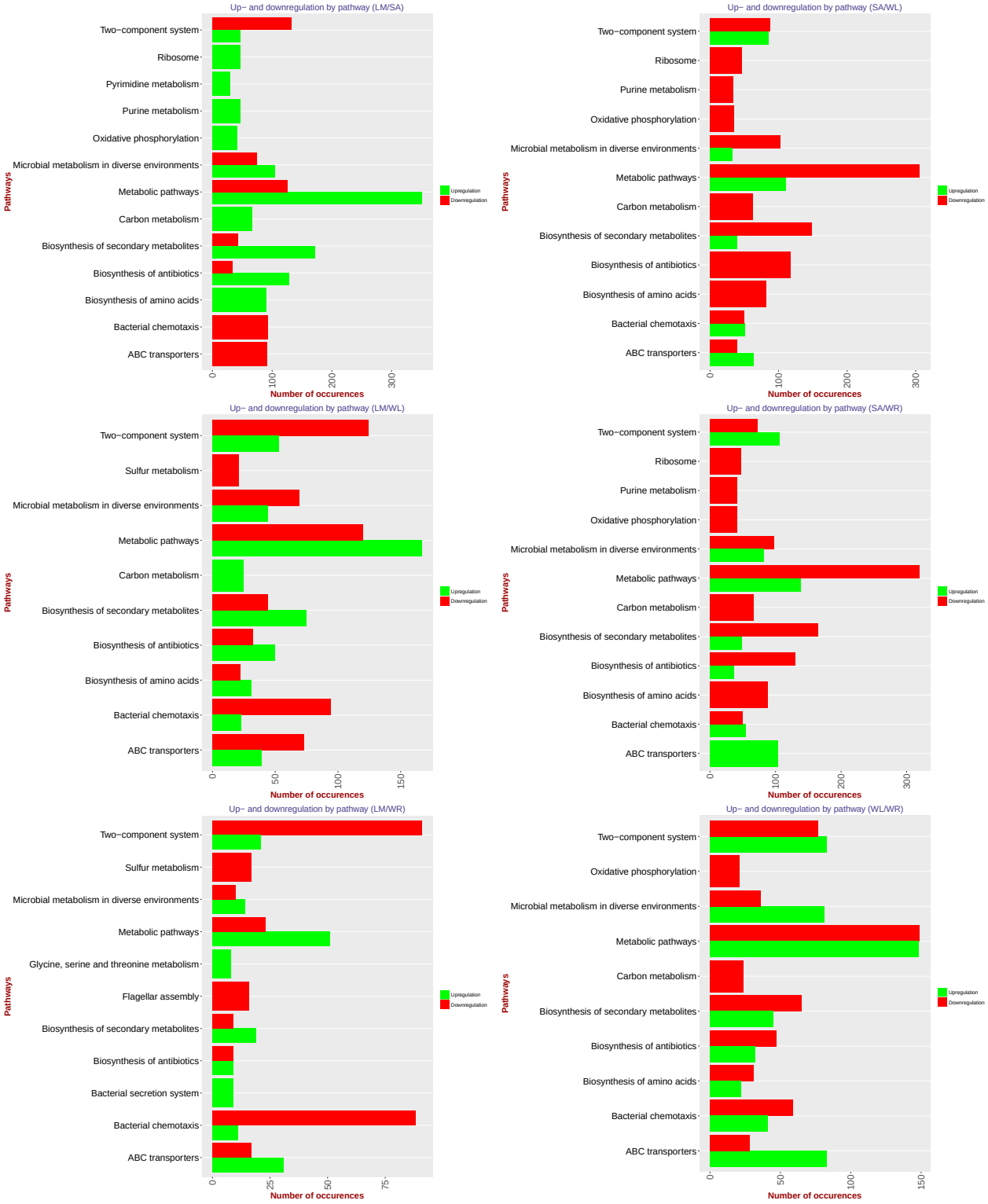


Figure S20. For all pairs of comparisons, barplots showing to which KEGG pathway the down- and upregulated genes belong.

Further statistical tests

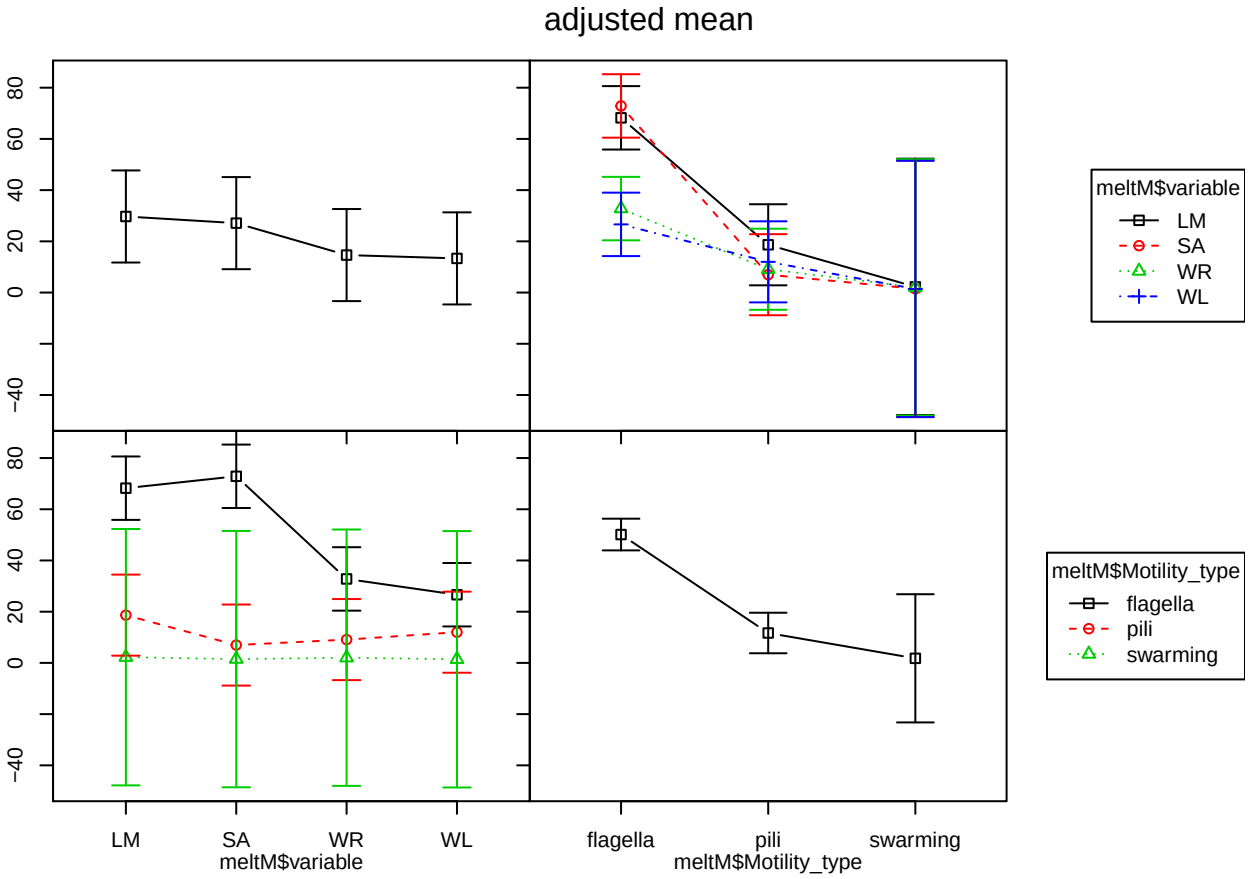


Figure S21. Interaction plots.

	Value	Df	Sum of Sq	F	Pr(>F)
LM-SA : flagella-pili	-16.32	1	2477.76	0.33	0.98
LM-WR : flagella-pili	25.89	1	6237.75	0.83	0.98
LM-WL : flagella-pili	34.94	1	11360.83	1.51	0.98
SA-WR : flagella-pili	42.21	1	16578.24	2.21	0.98
SA-WL : flagella-pili	51.26	1	24449.79	3.25	0.98
WR-WL : flagella-pili	9.05	1	762.19	0.10	0.98
LM-SA : flagella-swarming	-5.39	1	41.06	0.01	0.98
LM-WR : flagella-swarming	35.24	1	1755.41	0.23	0.98
LM-WL : flagella-swarming	40.78	1	2351.08	0.31	0.98
SA-WR : flagella-swarming	40.63	1	2333.41	0.31	0.98
SA-WL : flagella-swarming	46.17	1	3013.53	0.40	0.98
WR-WL : flagella-swarming	5.54	1	43.43	0.01	0.98
LM-SA : pili-swarming	10.93	1	162.89	0.02	0.98
LM-WR : pili-swarming	9.35	1	119.16	0.02	0.98
LM-WL : pili-swarming	5.84	1	46.51	0.01	0.98
SA-WR : pili-swarming	-1.58	1	3.41	0.00	0.98
SA-WL : pili-swarming	-5.09	1	35.32	0.00	0.98
WR-WL : pili-swarming	-3.51	1	16.78	0.00	0.98

Table S22. Test contrasts of factor interactions (experimental conditions and motility type).

Venn diagram

Here, we also tried to draw Venn diagram to help us visualize differential expression. Our trial was with liquid medium as reference. This was nonetheless not very conclusive.

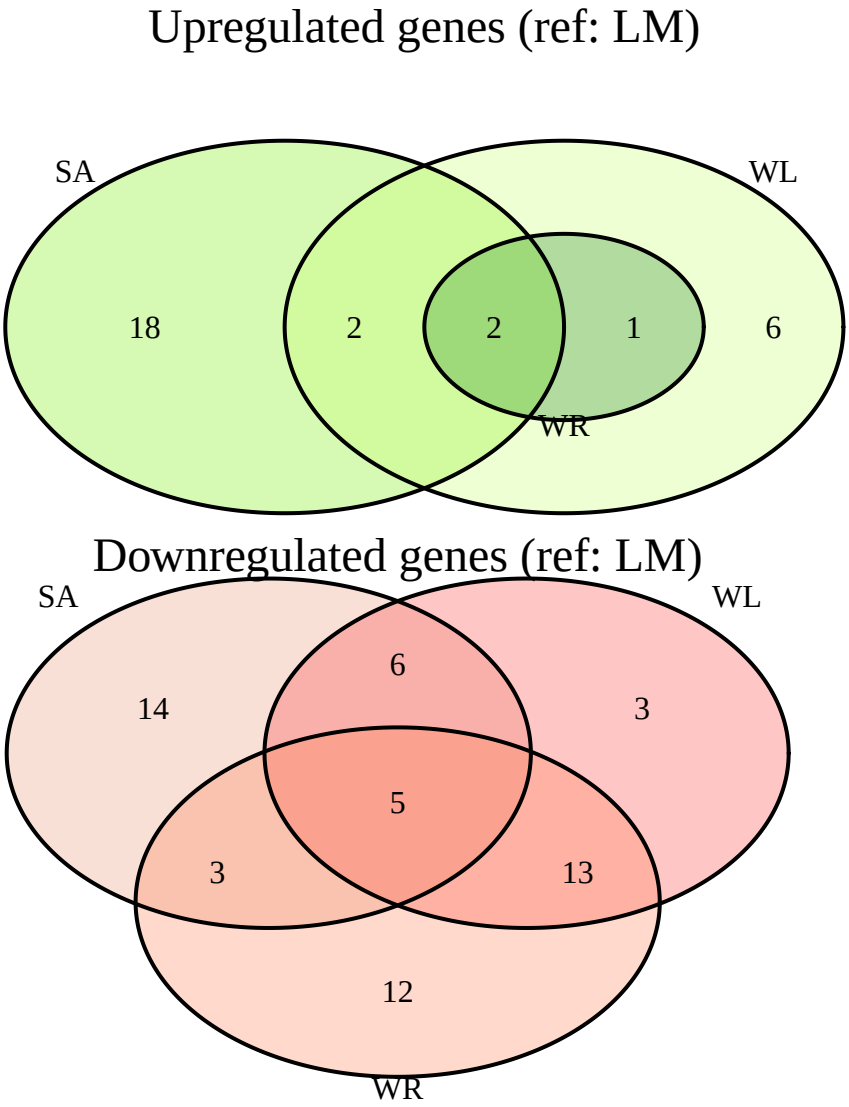


Figure S23. Example of Venn diagram for up- and downregulated genes. The number indicates the number of motility-associated genes up- (top) and downregulated (bottom) in the indicated condition when compared to LM.

Genome plots

Finally, we tried to visualize the clusters of motility-associated genes along the *Pseudomonas S5* genome with tools of the genoPlotR package (Lionel et al., Kultima, and Andersson, 2010). We compared their position in *Pseudomonas S5* genome and *Pseudomonas fluorescens* Pf5 genome (retrieved from http://www.pseudomonas.com and then processed in the terminal to obtain the optimal data shape; 58 motility-associated genes in common based on the gene name). Globally, the order of these genes is conserved for a large part of the motility-associated genes.

All anotated motility-associated genes

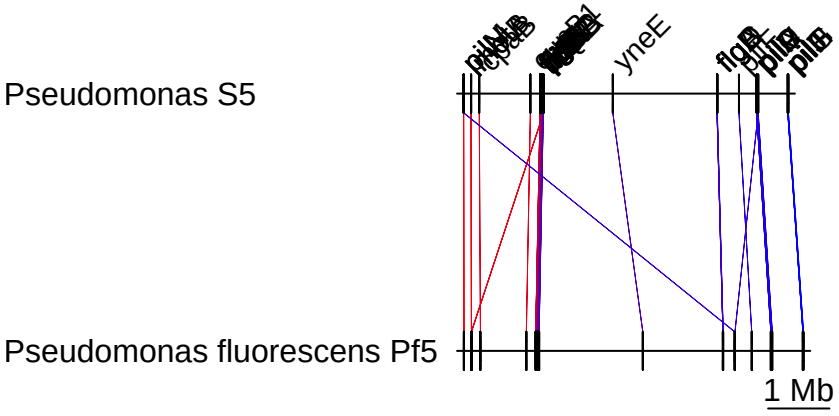


Figure S24. Genome plot for all motility-associated genes: comparison Pseudomonas protegens*S5 and* Pseudomonas fluorescens *Pf5.*

Flagellum-associated genes

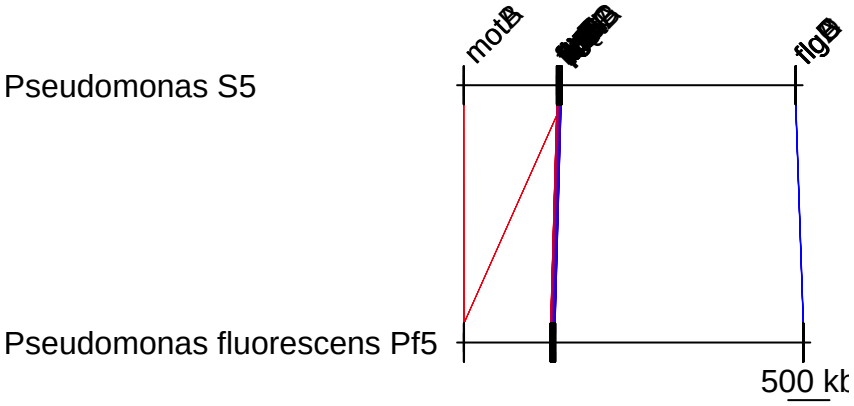


Figure S25. Genome plot for flagellum-associated genes: comparison Pseudomonas protegens *S5 and* Pseudomonas fluorescens *Pf5.*

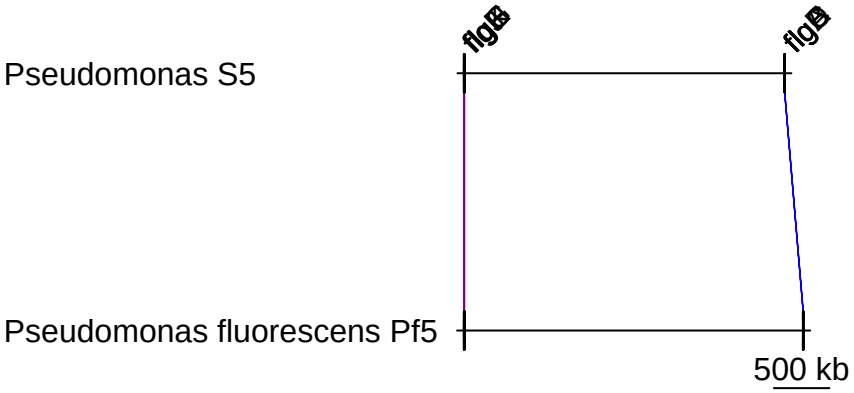


Figure S26. Genome plot for flg *family genes: comparison* Pseudomonas protegens *S5 and* Pseudomonas fluorescens *Pf5.*

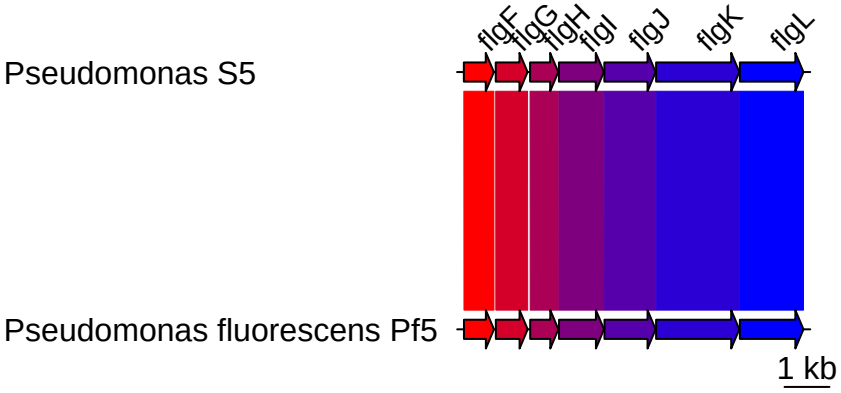


Figure S27. Genome plot for flg *family genes (close-up 1): comparison* Pseudomonas protegens *S5 and* Pseudomonas fluorescens *Pf5.*

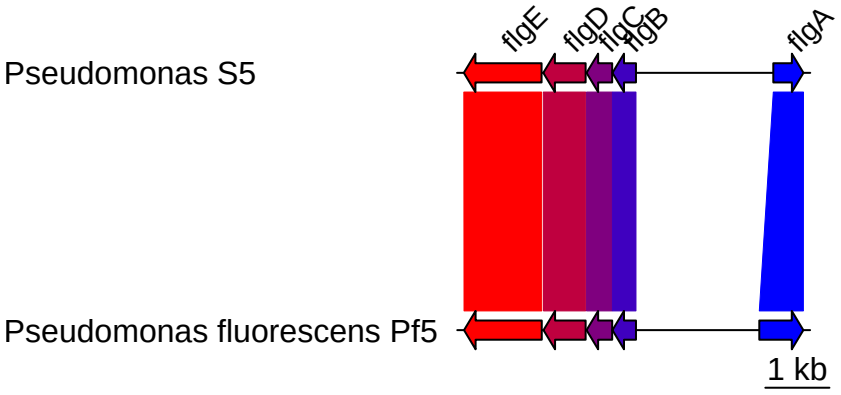


Figure S28. Genome plot for flg *family genes (close-up 2): comparison* Pseudomonas protegens *S5 and* Pseudomonas fluorescens *Pf5.*

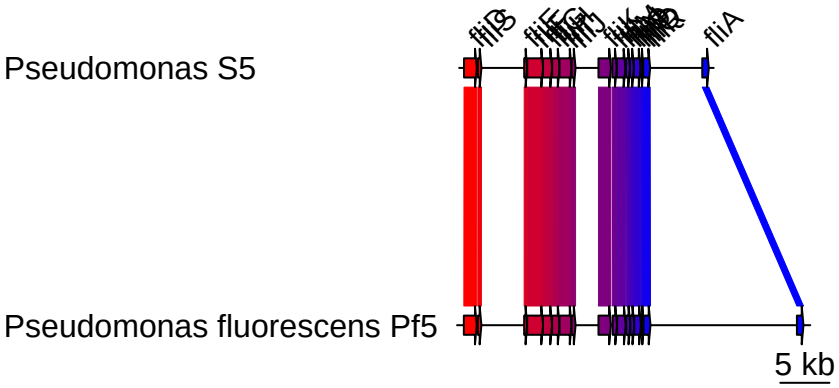


Figure S29. Genome plot for fli *family genes: comparison* Pseudomonas protegens *S5 and* Pseudomonas fluorescens *Pf5.*

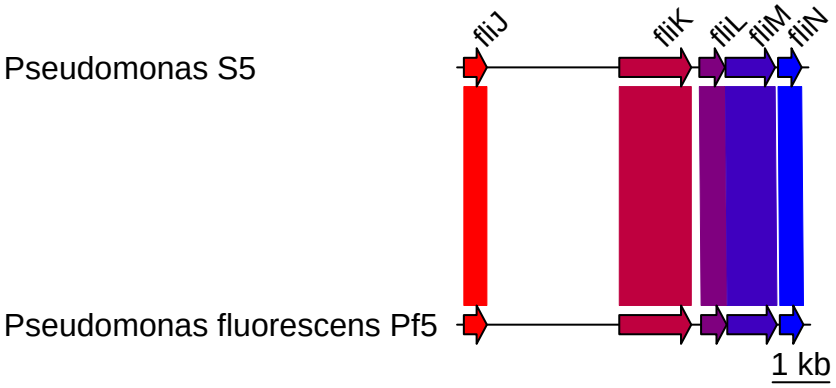


Figure S30. Genome plot for fli *family genes (close-up): comparison* Pseudomonas protegens *S5 and* Pseudomonas fluorescens *Pf5.*

Pilus-associated genes

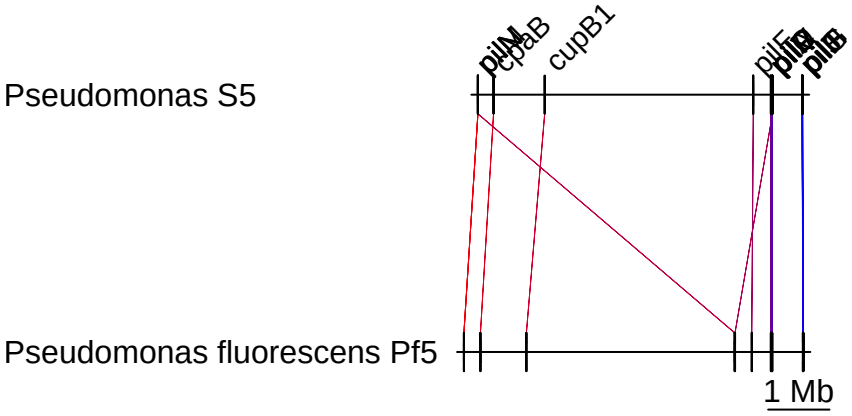


Figure S31. Genome plot for pilus-associated genes: comparison Pseudomonas protegens *S5 and* Pseudomonas fluorescens *Pf5.*

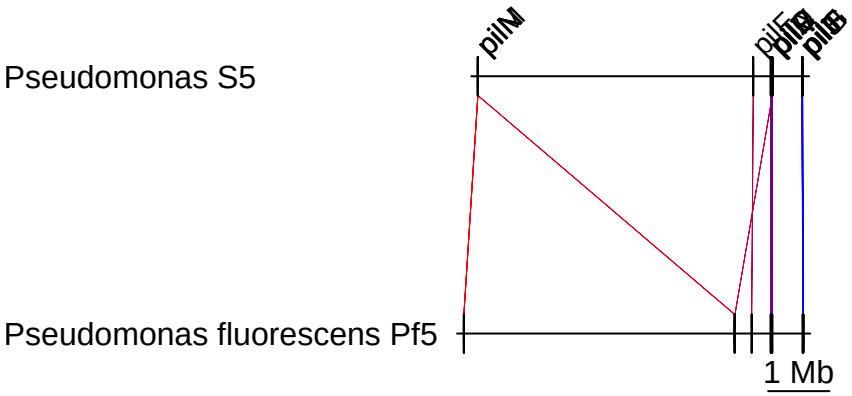


Figure S32. Genome plot for pil *family genes: comparison* Pseudomonas protegens *S5 and* Pseudomonas fluorescens *Pf5.*

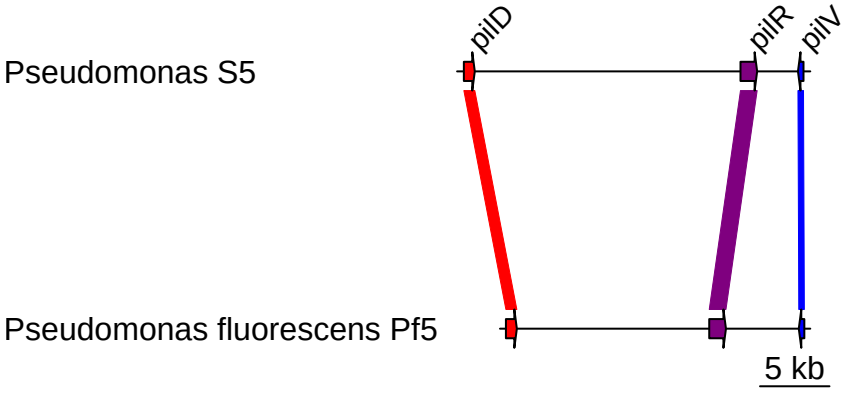


Figure S33. Genome plot for pil *family genes (close-up 1): comparison* Pseudomonas protegens *S5 and* Pseudomonas fluorescens *Pf5.*

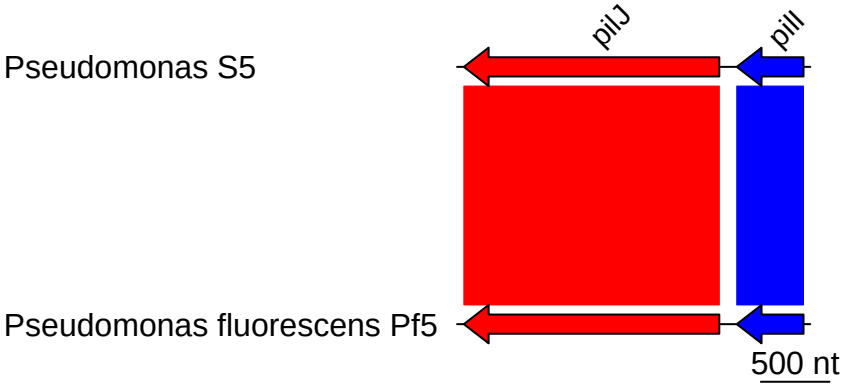


Figure S34. Genome plot for pil *family genes (close-up 2): comparison* Pseudomonas protegens *S5 and* Pseudomonas fluorescens *Pf5.*

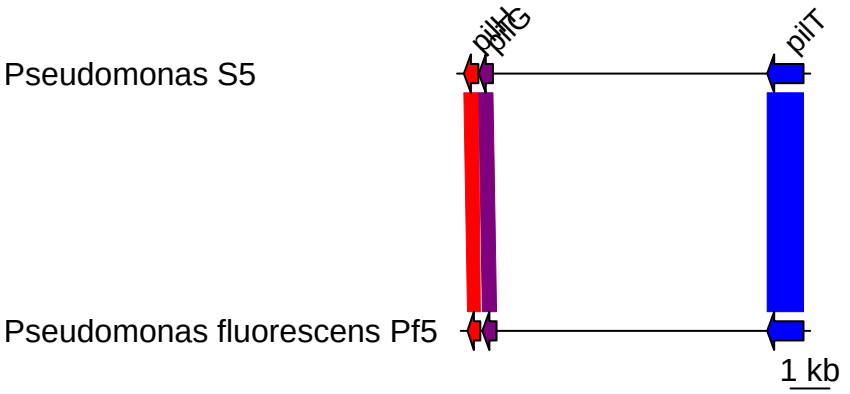


Figure S35. Genome plot for pil *family genes (close-up 3): comparison* Pseudomonas protegens *S5 and* Pseudomonas fluorescens *Pf5.*

MotA/MotB duplication ?

We also observed that the motor proteins of the flagellum (*motA* and *motB*) are duplicated in the *Pseudomonas* S5 genome that we sequenced. Interestingly, these genes have been reported to be present in two sets in other bacterial genome (*Pseudomonas aeruginosa*; Doyle et al. 2004)

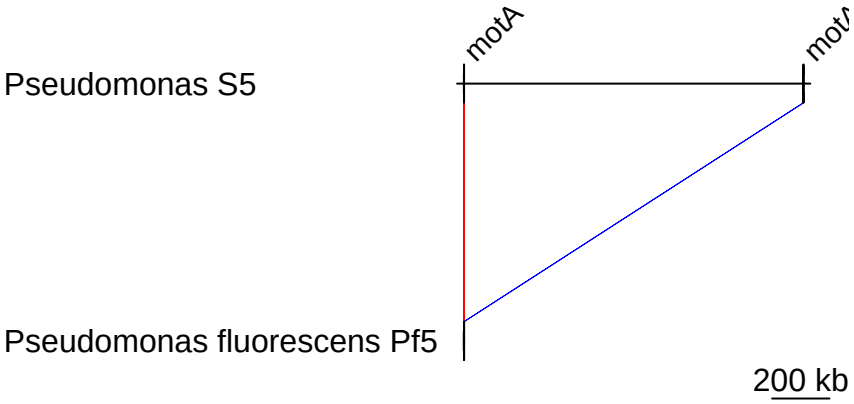


Figure S36. Genome plot for motA *genes: comparison* Pseudomonas protegens *S5 and* Pseudomonas fluorescens *Pf5.*

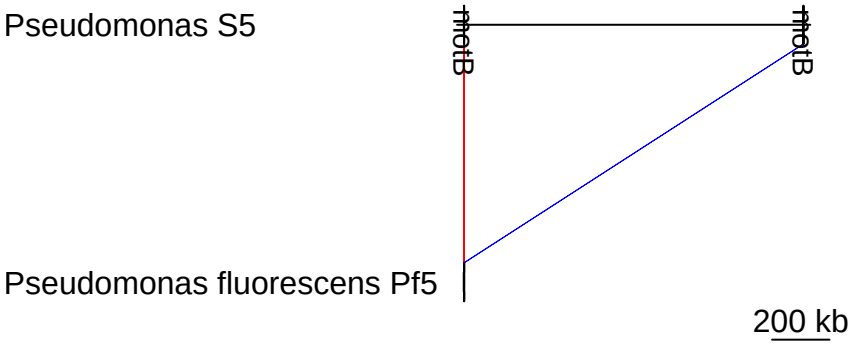


Figure S37. Genome plot for motB *genes: comparison* Pseudomonas protegens *S5 and* Pseudomonas fluorescens *Pf5.*

References

[1] S. M. Bache and H. Wickham. *magrittr: A Forward-Pipe Operator for R*. R package version 1.5. 2014.

[2] D. Borcard et al. *Numerical ecology with R*. Springer-Verlag New York, 2011.

[3] Y. Chen et al. "edgeR : differential expression analysis of digital gene expression data". In: *User ' s Guide* (2015), p. 104p.

[4] H. Chen. *VennDiagram: Generate High-Resolution Venn and Euler Plots*. R package version 1.6.17. 2016.

[5] H. De Rosario-Martinez. *phia: Post-Hoc Interaction Analysis*. R package version 0.2-1. 2015.

[6] L. Diray-Arce et al. "Transcriptome assembly, profiling and differential gene expression analysis of the halophyte Suaeda fruticosa provides insights into salt tolerance.". In: *BMC genomics* 16.1 (2015), p. 353.

[7] R. Kolde. *pheatmap: Pretty Heatmaps*. R package version 1.0.8. 2015.

[8] N. M. Krishnan et al. "De novo sequencing and assembly of Azadirachta indica fruit transcriptome". In: *Current Science* 101.12 (2011), pp. 1553-1561.

[9] G. Lionel et al., J. R. Kultima, et al. "genoPlotR: comparative gene and genome visualization in R". In: *Bioinformatics* 26.18 (2010), pp. 2334-2335.

[10] J. Oksanen et al. *vegan: Community Ecology Package*. R package version 2.3-5. 2016.

[11] Y. S. Rao et al. "Impact of GC content on gene expression pattern in chicken.". In: *Genetics, selection, evolution* 45.1 (2013), p. 9.

[12] D. Risso et al. "GC-Content Normalization for RNA-Seq Data". In: *BMC Bioinformatics* (2011), p. 17.

[13] M. D. Robinson et al. "edgeR: A Bioconductor package for differential expression analysis of digital gene expression data". In: *Bioinformatics* 26.1 (2009), pp. 139-140.

[14] A. E. Vinogradov. "Dualism of gene GC content and CpG pattern in regard to expression in the human genome: magnitude versus breadth". In: *Trends in Genetics* 21.12 (2005), pp. 633-639.

[15] H. Wickham. "Reshaping Data with the reshape Package". In: *Journal of Statistical Software* 21.12 (2007), pp. 1-20.

[16] H. Wickham and R. Francois. *dplyr: A Grammar of Data Manipulation*. R package version 0.4.3. 2015.

[17] H. Wickhamm. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009.

[18] Y. Xie. *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, 2013.