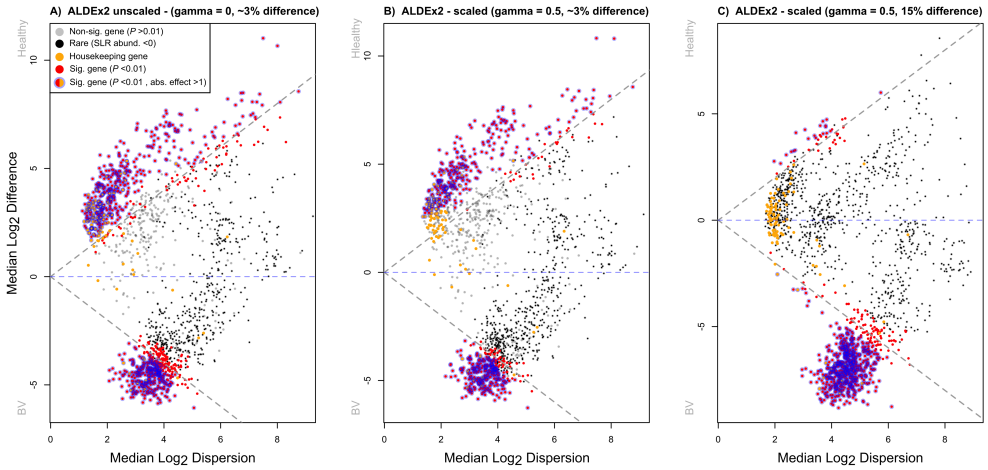
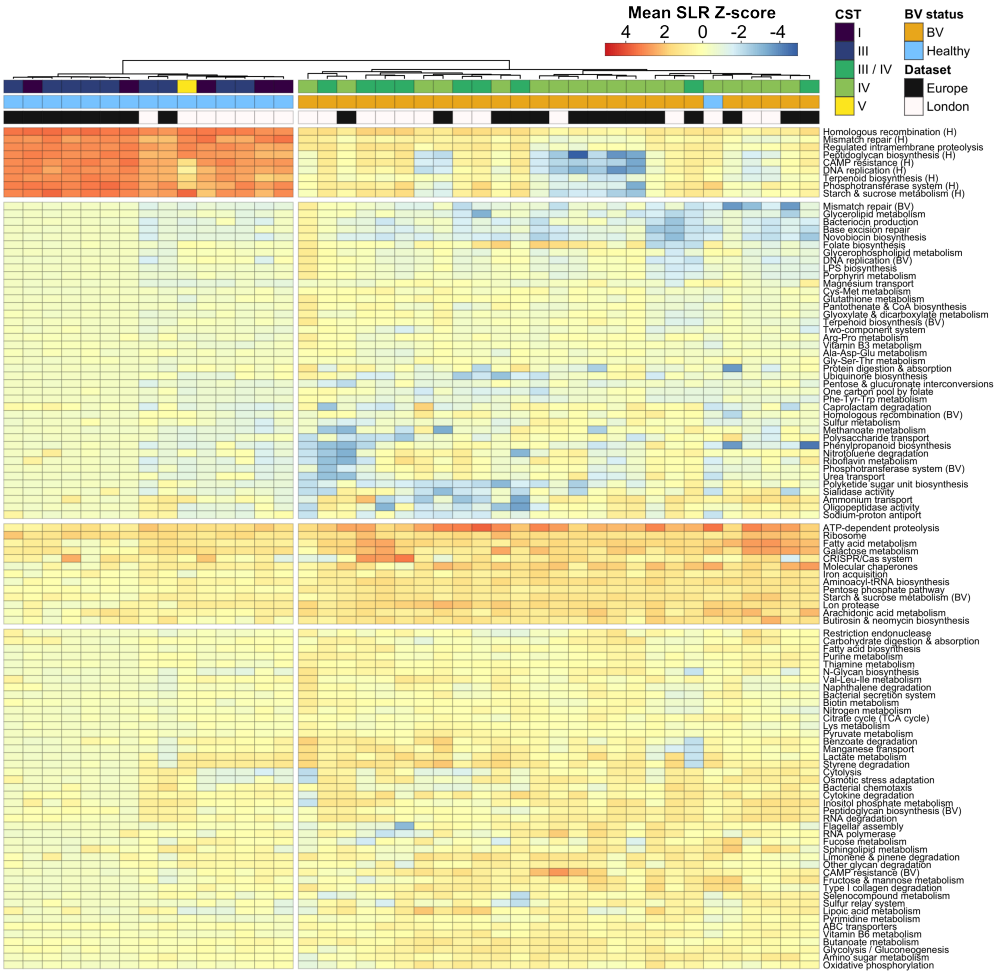
1 Supplementary material



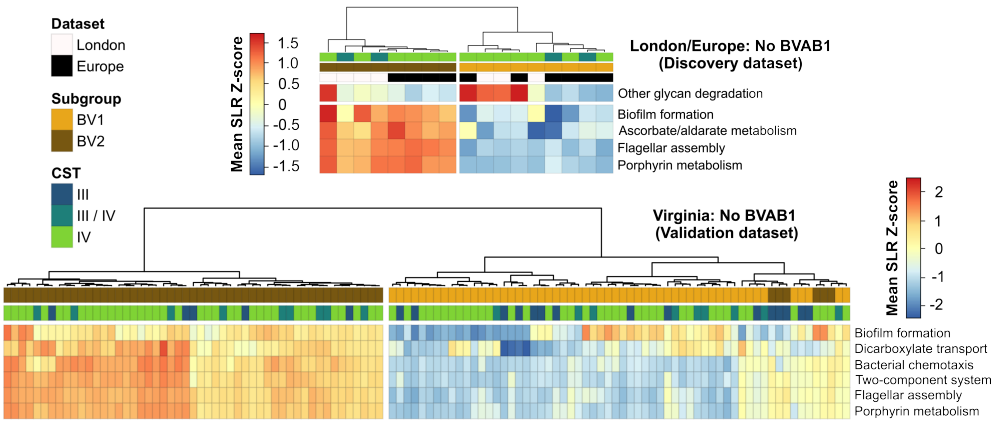
Supplementary Fig. S1 Using a between-group scale offset of 15 % is sufficient to centre a vaginal metatranscriptomic dataset: MW effect plots were produced for the London/Europe discovery dataset normalised using (**A**) ALDEx2 without scale simulation, (**B**) ALDEx2 operating with the default scale model (gamma = 0.5) or (**C**) ALDEx2 running with a full scale model (gamma = 0.5, 15 % scale offset between healthy and BV groups). Not accounting for scale results in a large number of false-positive results (**A**) and a clear error in the centre-point of the data which is, by definition, determined by the housekeeping features. Adding a small amount of scale uncertainty (**B**) increases within-group dispersion of all features such that many housekeeping functions are no longer differential; however, only the full scale model (**C**) correctly centres the housekeeping functions around the point of no difference between groups and corrects the large number of false-positive and -negative results. Colouring of data points as follows: non-significant features = grey ($P \geq 0.01$), non-significant, rare features = black (median SLR abundance ¡0), housekeeping features = orange (KO terms corresponding to 'Glycolysis', 'Ribosome' or 'Aminoacyl-tRNA biosynthesis'), significant features based on P-value alone = red ($P \leq 0.01$) and significant features based on P-value and effect size = red with blue outline ($P \leq 0.01$, absolute effect size ≥ 1). Grey dashed lines indicate equivalence of between- and within- group difference; blue dashed line indicates no difference between groups.



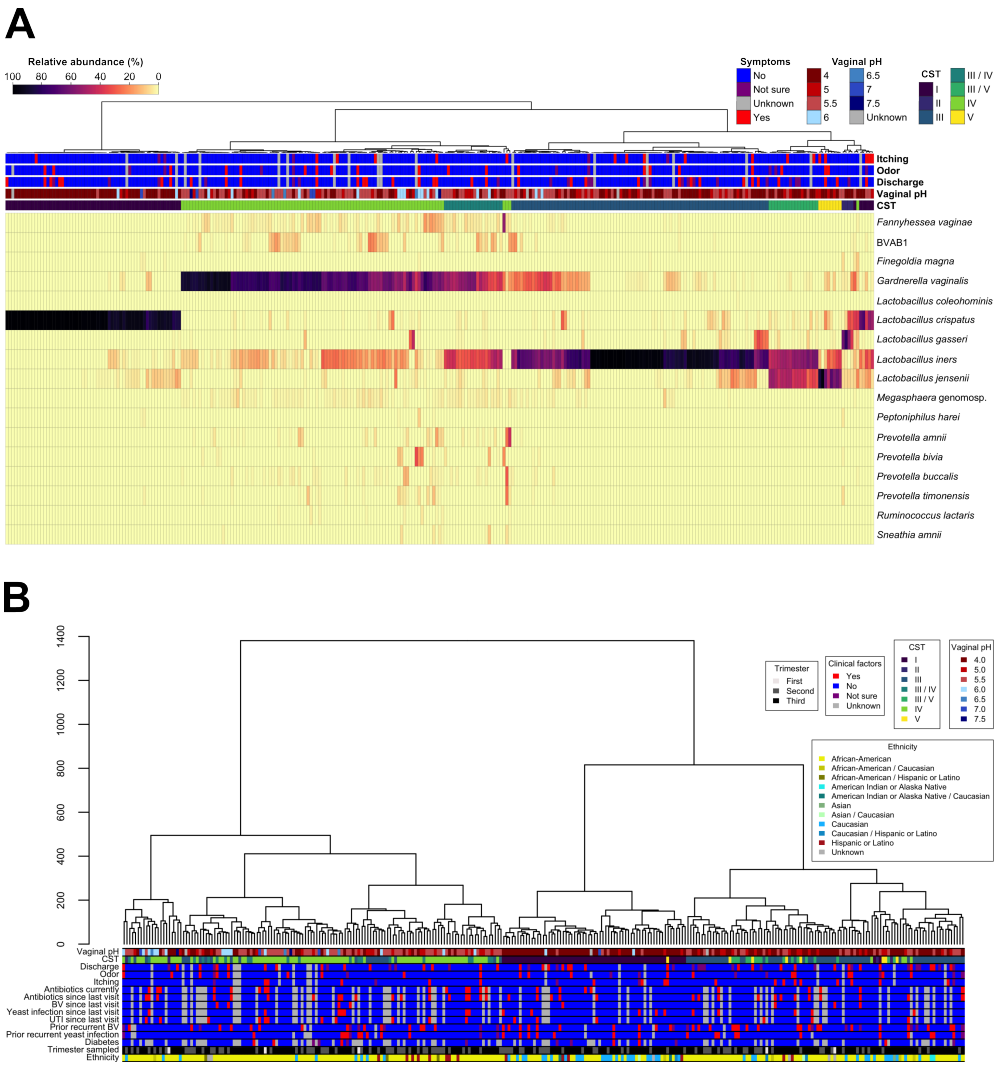
Supplementary Fig. S2 Differentially expressed pathways between health and BV – London/Europe dataset: Gene assignments were grouped by KEGG identifiers regardless of species for each sample and all differentially abundant KEGG pathways between healthy and BV metatranscriptomes in the London/Europe dataset are shown. All post-SRI absolute effect sizes and expected false discovery rates calculated by ALDEx2 are >1 and <0.01, respectively. CST, BV status and dataset indicated by colour bars.

Supplementary Table S1 Total gene and read counts for all genes assigned to the four KEGG orthology terms that represent the DltABCD operon. Data are separated to show gene and read counts for the London/Europe (discovery) and Virginia (validation) datasets.

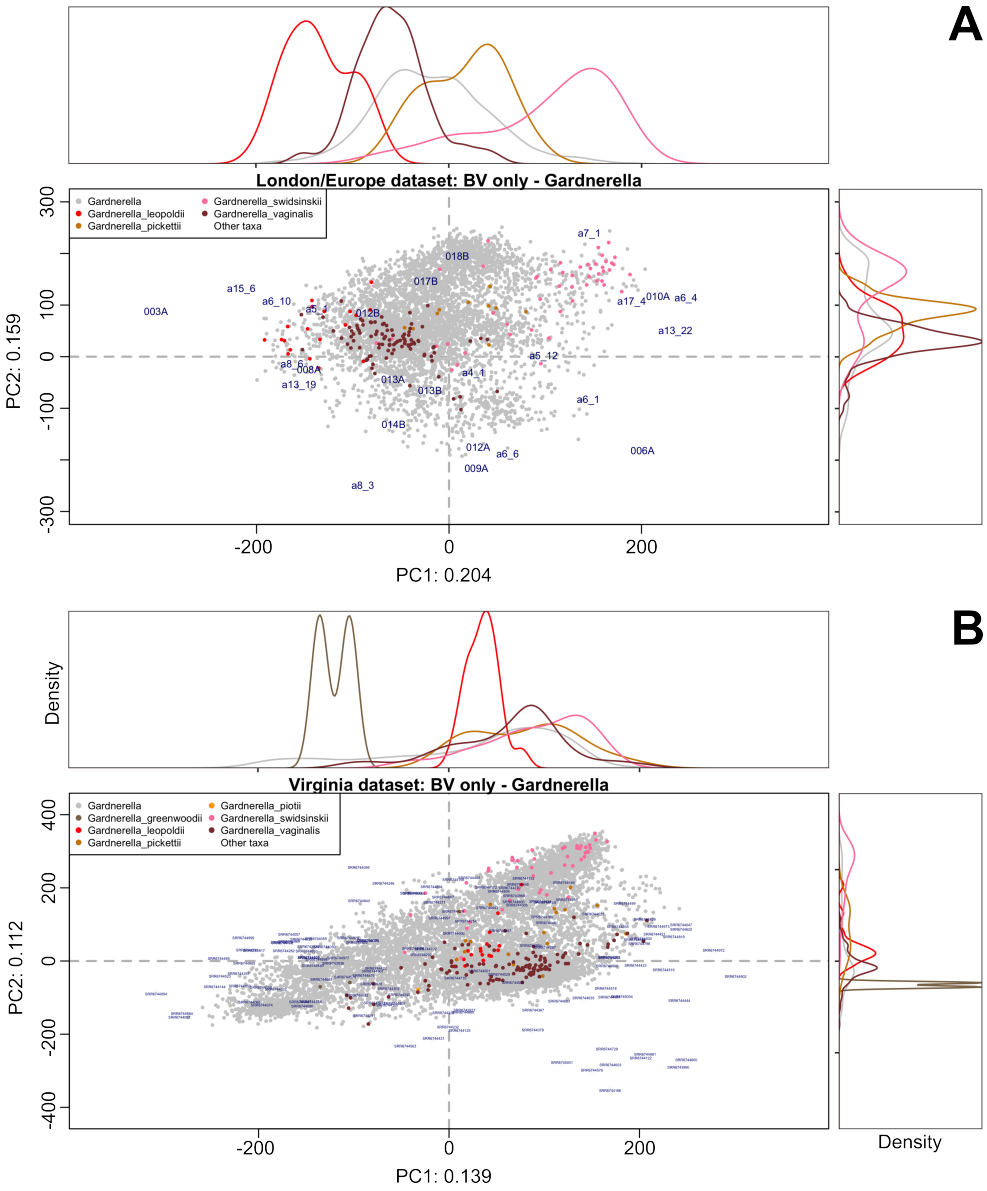
	London/Europe		Virginia	
Dataset	No. genes	No. reads	No. genes	No. reads
L. crispatus	5	115,858	8	7,191,584
L. iners	8	177,871	8	5,429,846
L. jensenii	2	79,900	7	1,613,482
L. gasseri	2	14,043	2	144,203
Unknown taxonomy	9	41,381	11	1,458,737



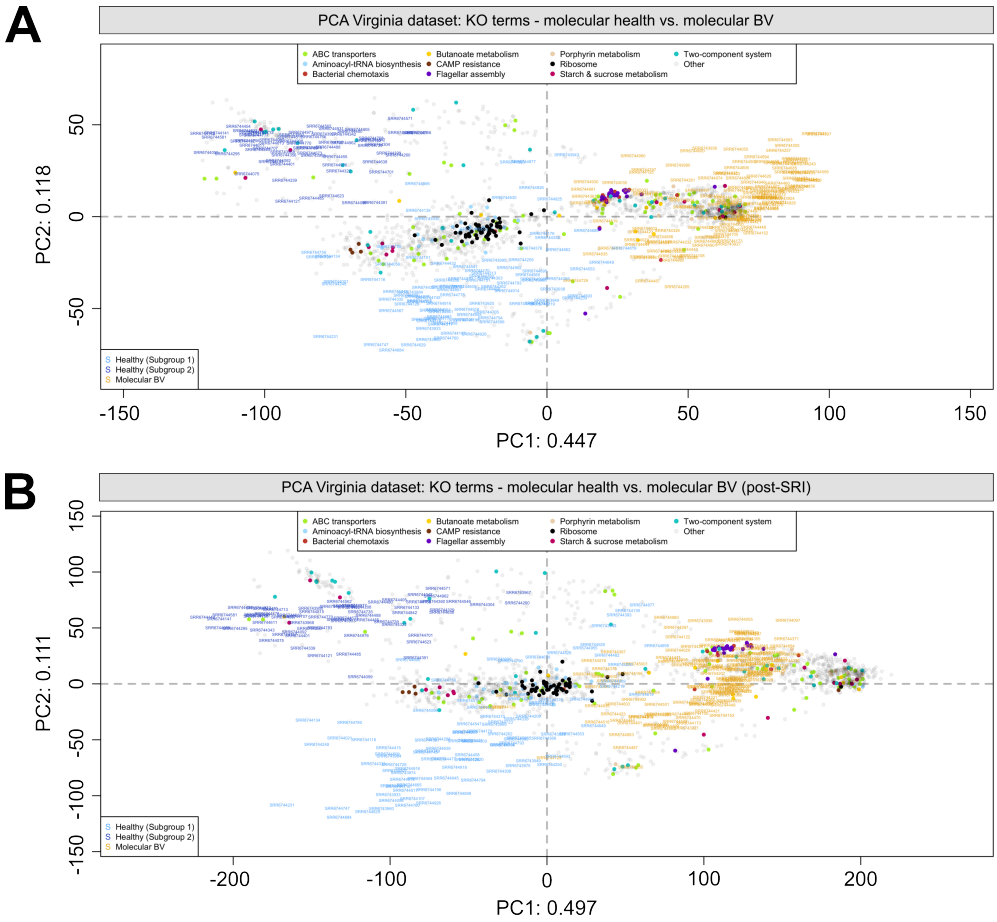
Supplementary Fig. S3 BV subgroup differences are not solely due to the presence of BVAB1: Differential expression analyses for both London/Europe (top) and Virginia (bottom) datasets were repeated after removing all genes with a taxonomic assignment to BVAB1. Differentially expressed pathways among samples are expressed as a Z-score calculated from the mean scaled log-ratio (SLR) values of all KO terms assigned to a given pathway (individual SLR values represent a median across 128 Monte-Carlo instances). All post-SRI absolute effect sizes and false discovery rates calculated by ALDEx2 were >1 and <0.01, respectively).



Supplementary Fig. S4 Definition of molecular states of health and BV by clustering of metatranscriptomes at the functional level: All reads from the Virginia dataset were aggregated by species (proportional abundance of expressed transcripts; **A**) or function (KO term; **B**), prior to hierarchical clustering of Euclidean distances using Ward's method. Colour bars indicate vaginal pH, CST, clinical metadata, and ethnicity. (**A**) Proportional abundances of expressed metatranscriptome reads with known taxonomy; species represented by $\geq$75 genes are shown. (**B**) Dendrogram of SLR-transformed, post-SRI metatranscriptome profiles aggregated by KO term, regardless of taxonomy.

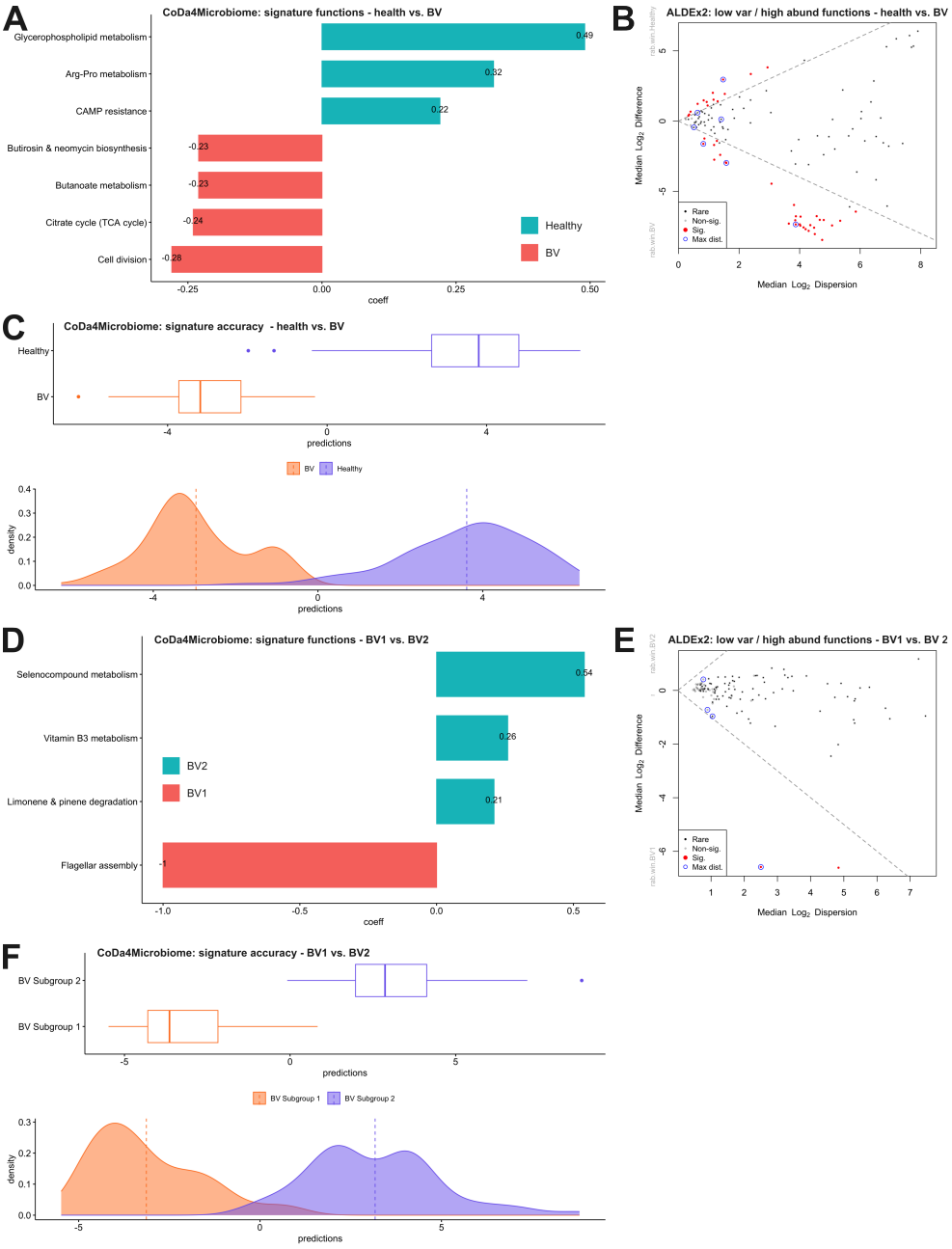


Supplementary Fig. S5 Individual *Gardnerella* species can be distinguished within metatranscriptome profiles: Metatranscriptome profiles from London/Europe (**A**) and Virginia (**B**) datasets were subject to SLR-transformation prior to PCA. All gene features were included in the ordination; however only reads assigned to *Gardnerella* are shown. Genes found to be unique to each of the named *Gardnerella* species by pangenome analyses are coloured.

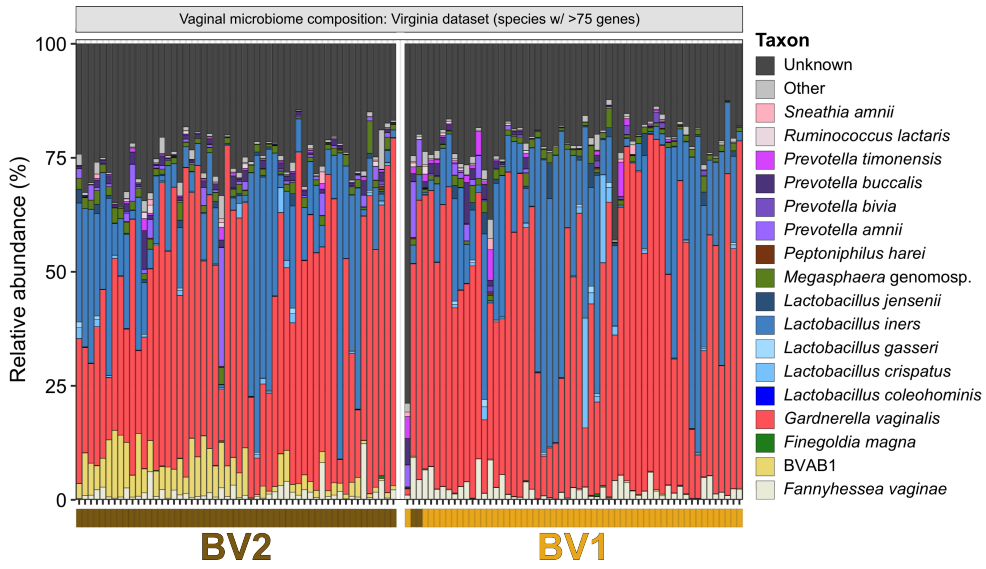


Supplementary Fig. S6 Correcting data asymmetry with SRI effectively centres genes of no-difference: KO-aggregated metatranscriptome profiles from the Virginia dataset underwent SLR-transformation either alone (**A**) or with SRI (**B**; 15 % scale difference, gamma = 0.5) prior to PCA. KO terms coloured by KEGG pathway assignments for a selection of relevant KEGG pathways. Samples coloured by group according to hierarchical clustering in Supplementary Figure S4.

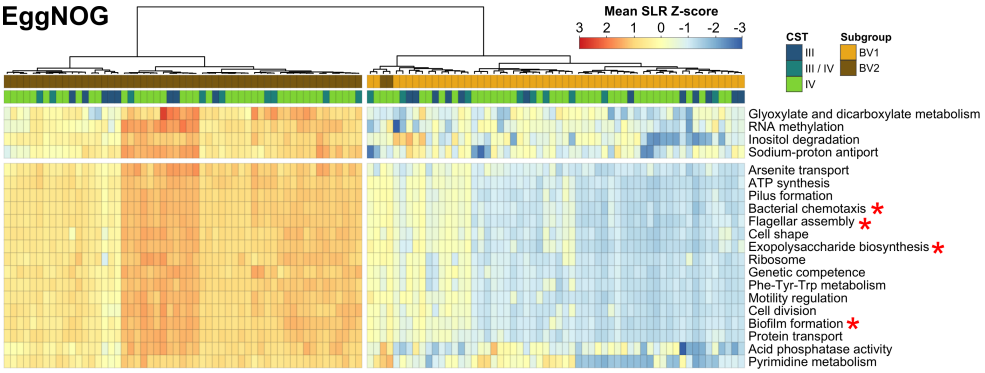
Supplementary Fig. S7 Major differentially expressed pathways of the vaginal metatranscriptome can be replicated across datasets: Metatranscriptome profiles from the Virginia dataset were aggregated by KO term prior to SLR-transformation and SRI using ALDEx2. Analysis was stratified by states of molecular health or BV and Z-scores of mean SLR values per pathway are plotted. Colour bars indicate group and CST. All post-SRI effect sizes and BH-corrected P-values calculated by ALDEx2 are $\leq$0.01 and $<$1 %, respectively.



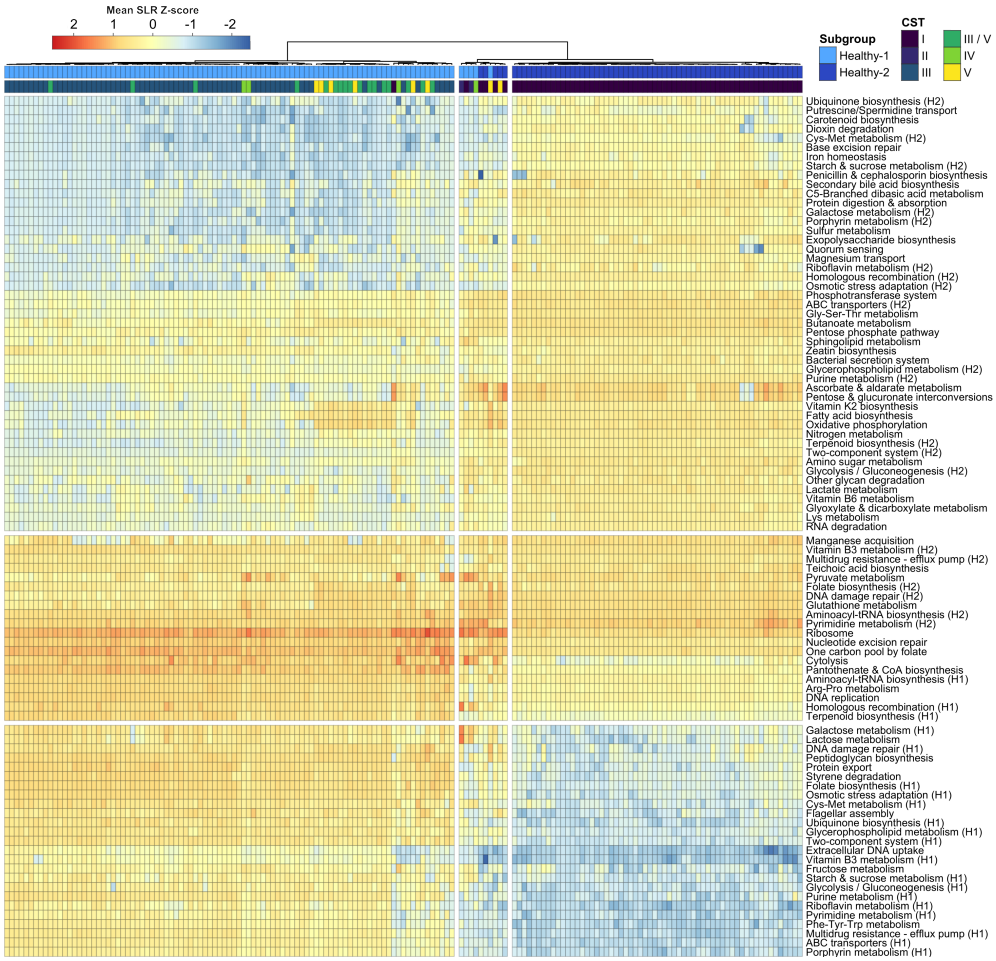
Supplementary Fig. S8 Prediction of signature functions complements differential expression analyses: Prediction of signature functions for health vs. BV (A-C) and BV subgroups (D-F) by the 'coda4microbiome' R package. Sets of functions identified in signature plots (A,D) exhibit high AUC values of 0.9975 and 0.9947, respectively, and are indicated on ALDEx2 effect plots in blue. Prediction plots (C,F) demonstrate a high degree of separation by group for these signatures, indicating that they define the respective groups with a high degree of accuracy. These results should be interpreted with caution, given that calculating the geometric mean from low-variance, high-abundance features does not eliminate the problematic data asymmetry shown in Suppl. Fig S1.



Supplementary Fig. S9 Proportional abundance of BVAB1 transcripts is not the sole driver of functional differences between molecular BV subgroups: Metatranscriptome profiles of samples belonging to the two main BV subgroups within the Virginia dataset are expressed as proportional abundance of expressed transcripts. Reads of known and unknown taxonomy are included, with reads corresponding to species represented by $\leq$75 genes grouped into 'Other'. Subgroup membership indicated below plot.



Supplementary Fig. S10 Functional differences between BV subgroups are not dependent on enzyme classification systems: Metatranscriptome profiles from the Virginia dataset were aggregated by KO term prior to SLR-transformation and SRI using ALDEx2. Analysis was stratified by membership of molecular BV subgroups and Z-scores of mean SLR values per pathway are plotted. Red asterisks indicate functions identified in the corresponding KEGG KO term analysis. Colour bars indicate group and CST. All post-SRI effect sizes and BH-corrected P-values calculated by ALDEx2 are $\leq$0.01 and $<$1 %, respectively.



Supplementary Fig. S11 Differentially expressed pathways between molecular health subgroups –Virginia dataset: Metatranscriptome profiles from the Virginia dataset were aggregated by KO term prior to SLR-transformation and SRI using ALDEx2. Analysis was stratified by membership of molecular health subgroups and Z-scores of mean SLR values per pathway are plotted. Colour bars indicate group and CST. All post-SRI effect sizes and BH-corrected P-values calculated by ALDEx2 are $\leq$0.01 and $<$1 %, respectively.