Compound Set Enrichment: A novel approach to analysis of primary HTS data.

Thibault Varin, Hanspeter Gubler, Christian N. Parker, Ji-Hu Zhang, Pichai Raman, Peter Ertl, Ansgar Schuffenhauer*

Novartis Institutes for BioMedical Research.

CH-4056 Basel, Switzerland & 250 Massachusetts Avenue, Cambridge MA 02139, USA

*thibault.varin@novartis.com

Most of data included in the article corresponds to the bioassay 893. We introduce in Supplementary material data for the 6 other bioassays. These data are commented in the article manuscript. This material is available free of charge via the Internet at http://pubs.acs.org.

Part IV : Bonferroni correction effects

An important parameter for statistical hypothesis tests is the sample size (number of compounds having a scaffold) that is considered for the prediction. The larger the sample size is, the lower the number of false positive and negative predictions[24]. As a consequence greater confidence can be assigned to the results from highly populated scaffolds. However, to determine the minimum sample size is not something trivial a priori. The Bonferoni correction was applied separately for each level (see Table 2 in the article for the corresponding critical level of significance per test) and then analyzed, *a posteriori*, the number of compounds for scaffolds predicted as actives in order to determine the impact of the Bonferroni correction on predicted significantly active scaffolds and to see if the number of compounds having these scaffolds is too small.

Figure S14 (KSP) and Figure S15 (BTP) represent for the bioassay 893 the number of scaffolds significantly active (α[PF]<=0.01) according to the number of compounds having these scaffolds (for the 6 other bioassays, report to Figures from S16 to S27).

This analysis has been done for active scaffolds without (column 1) and with the Bonferroni correction (column 2) and for levels 1, 2, 3, 4 and 5 of ST compound classification. As would be expected the Bonferroni correction dramatically reduces the number of scaffolds predicted as active. For both KSP and BTP, at each level of the ST classification, the proportion and / or the number of scaffolds having less than 10 compounds is much smaller with the Bonferroni correction than without it. For this reason, it was decided to analyze results without fixing a limit on the number of compounds per scaffold. With the Bonferroni correction, the proportion and / or the number of significant active scaffolds that are populated by less than 10 compounds are much higher for KSP than for BTP (except for the bioassay 883 for which results are similars). These

results suggest that the KS test is able to identify significant active classes even if few compounds have been tested.

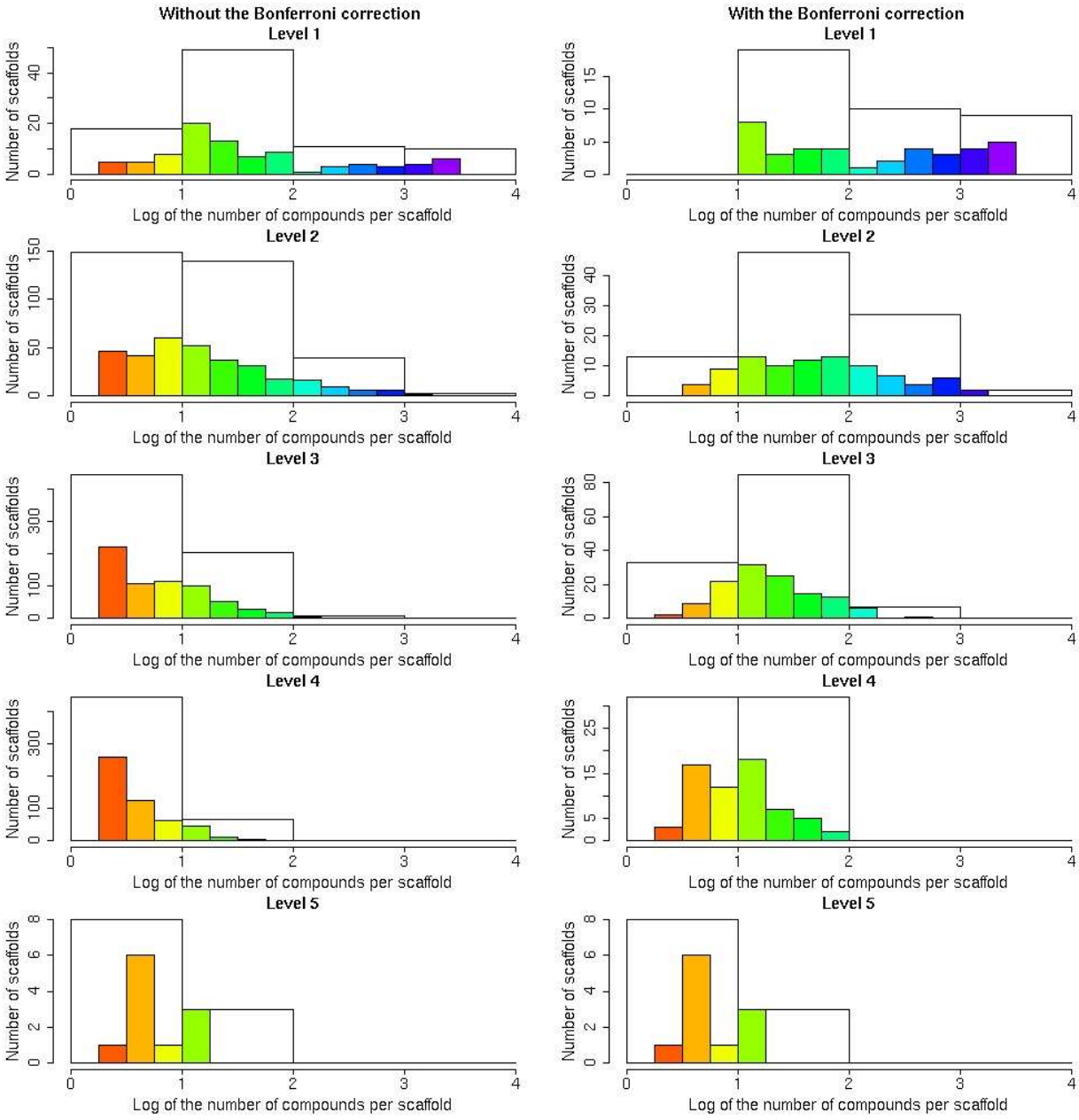

Figure S14. Size of KSP (KS Prediction) significantly active scaffolds for the bioassay 893. The x axis corresponds to the log of the number of compounds per scaffold. The coloured bars are sub-levels of the white bars. Colors are shown to facilitate comparison of results without (left) and with (right) the Bonferroni correction.

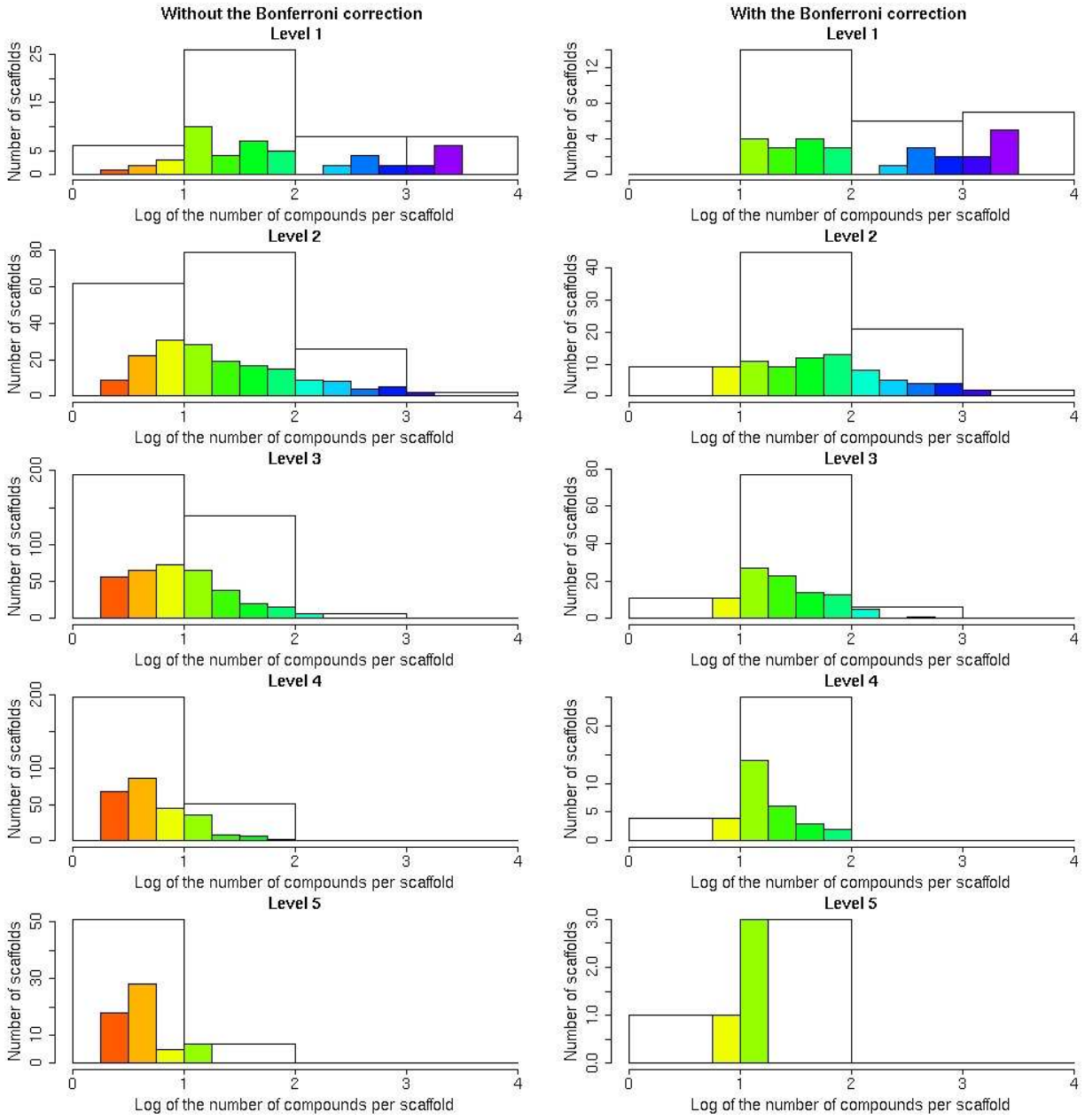

Figure S15. Size of BTP (Binomial Threshold Prediction) significantly active classes for the bioassay 893. The x axis corresponds to the log of the number of compounds per class. The coloured bars are sub-levels of the white bars. Colors are shown to facilitate comparison of results without (left) and with (right) the Bonferroni correction.

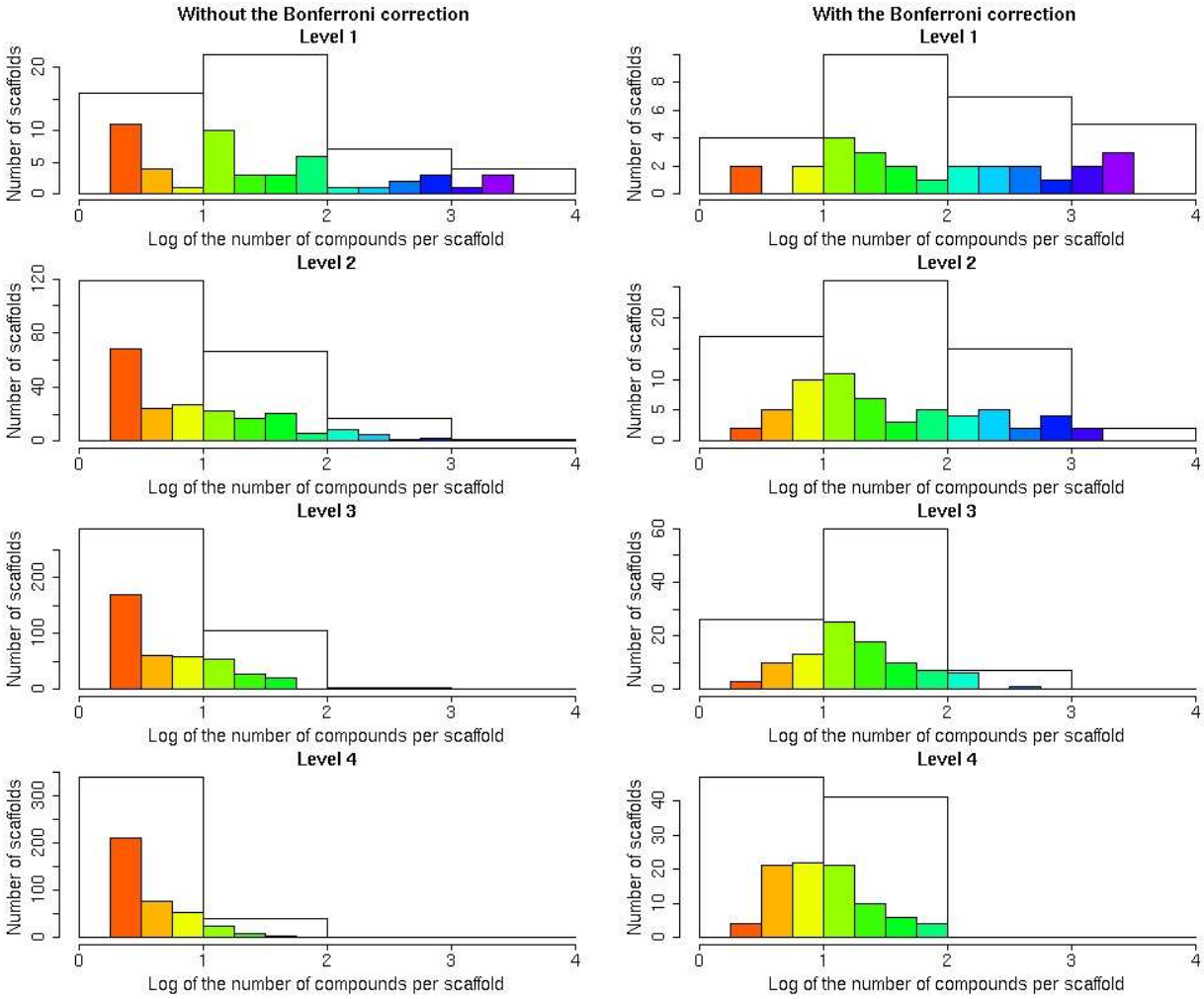

Figure S16 Size of KSP (KS Prediction) significantly active scaffolds for the bioassay 900. The x axis corresponds to the log of the number of compounds per scaffold. The coloured bars are sub-levels of the white bars. Colors are shown to facilitate comparison of results without (left) and with (right) the Bonferroni correction.

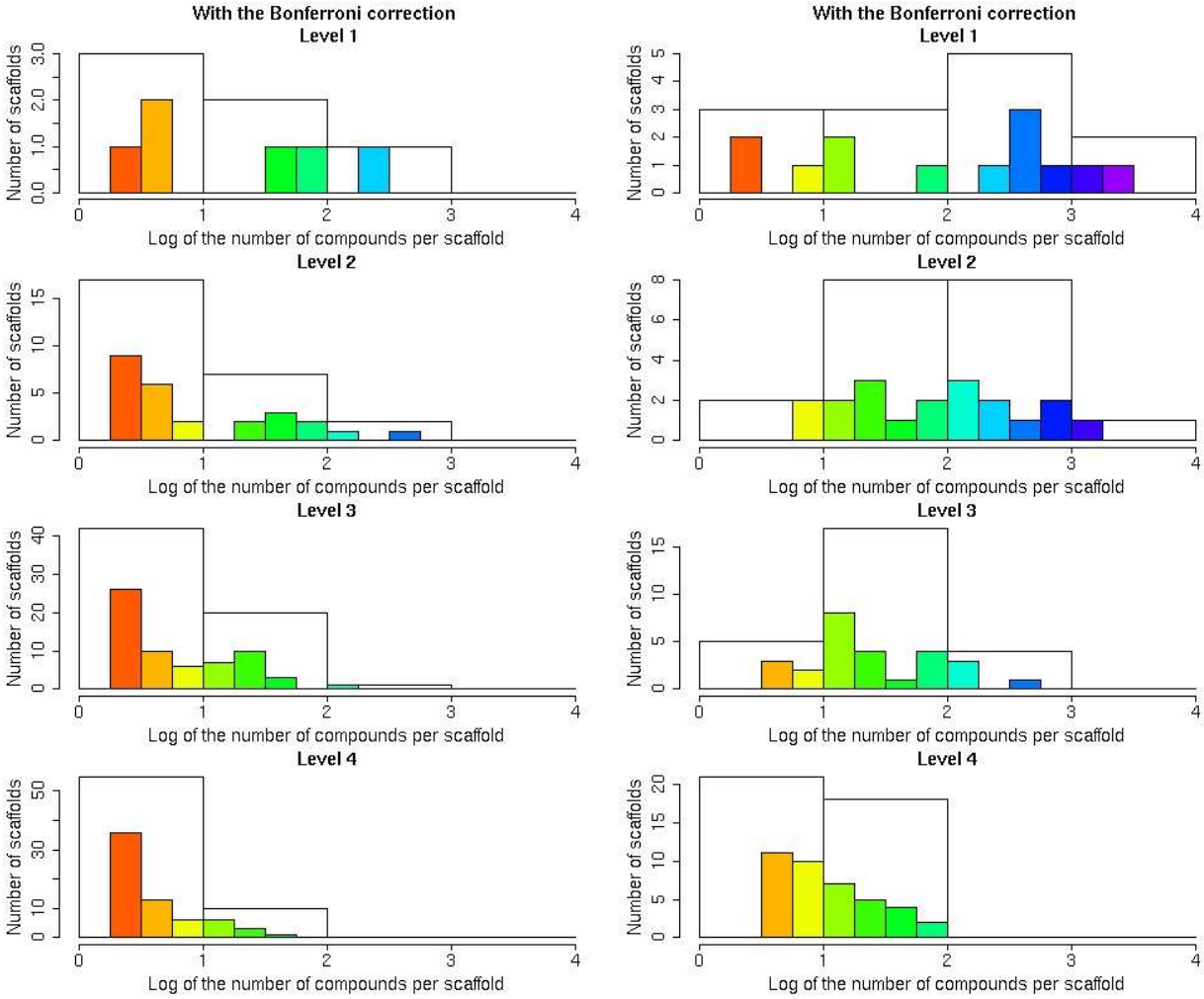

Figure S17 Size of BTP (Binomial Threshold Prediction) significantly active classes for the bioassay 900. The x axis corresponds to the log of the number of compounds per class. The coloured bars are sub-levels of the white bars. Colors are shown to facilitate comparison of results without (left) and with (right) the Bonferroni correction.

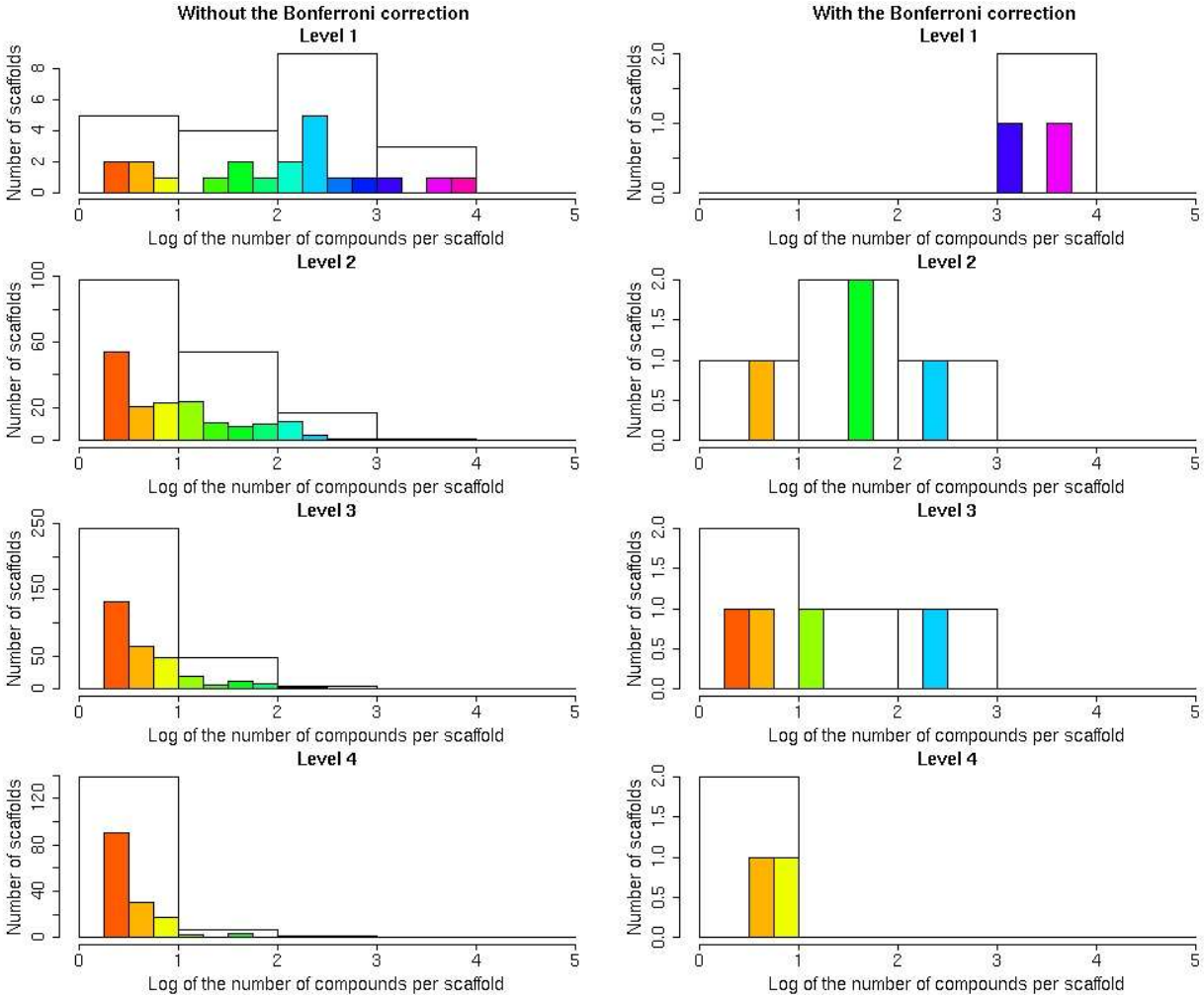

Figure S18 Size of KSP (KS Prediction) significantly active scaffolds for the bioassay 1634. The x axis corresponds to the log of the number of compounds per scaffold. The coloured bars are sub-levels of the white bars. Colors are shown to facilitate comparison of results without (left) and with (right) the Bonferroni correction.

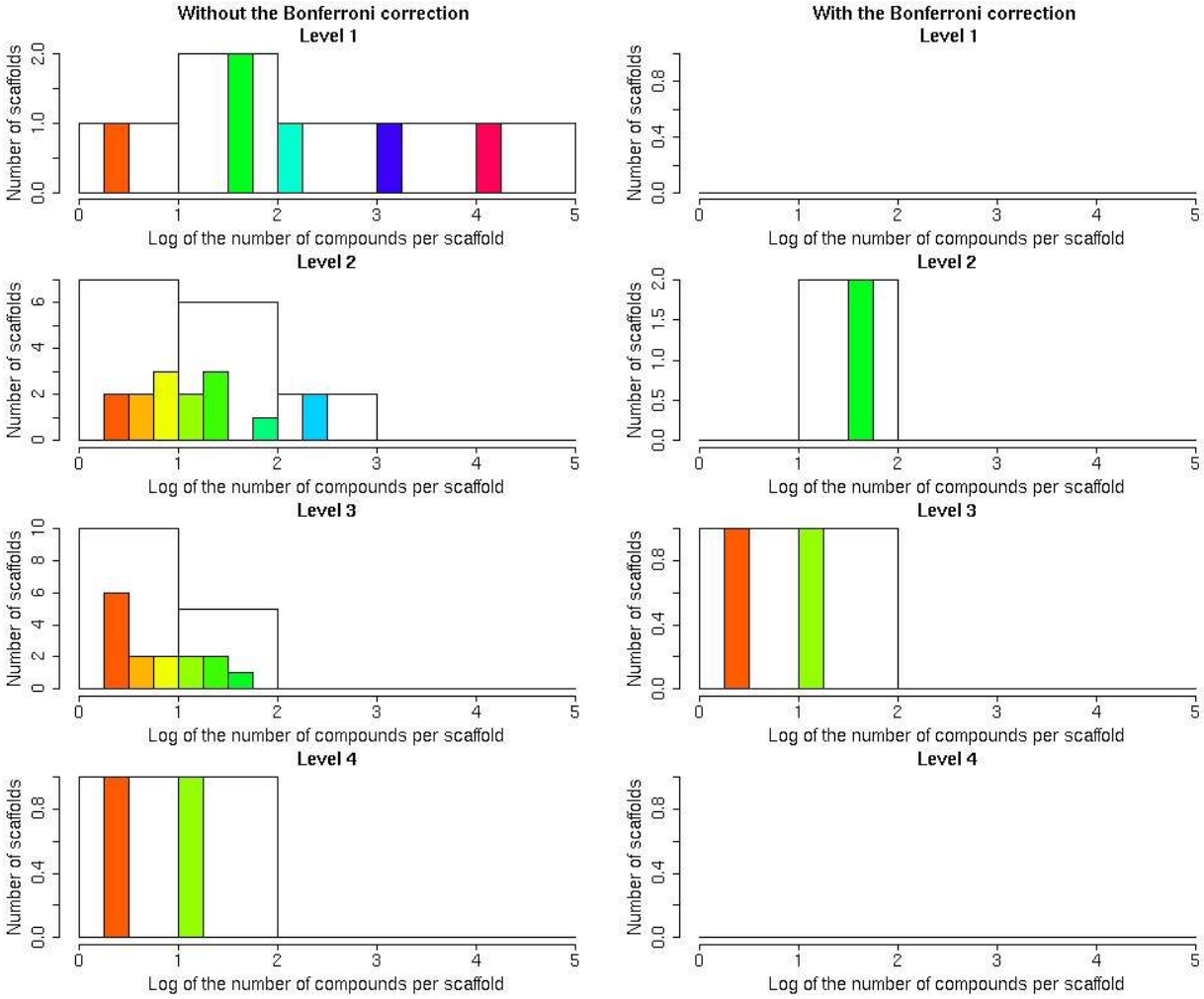

Figure S19 Size of BTP (Binomial Threshold Prediction) significantly active classes for the bioassay 1634. The x axis corresponds to the log of the number of compounds per class. The coloured bars are sub-levels of the white bars. Colors are shown to facilitate comparison of results without (left) and with (right) the Bonferroni correction.

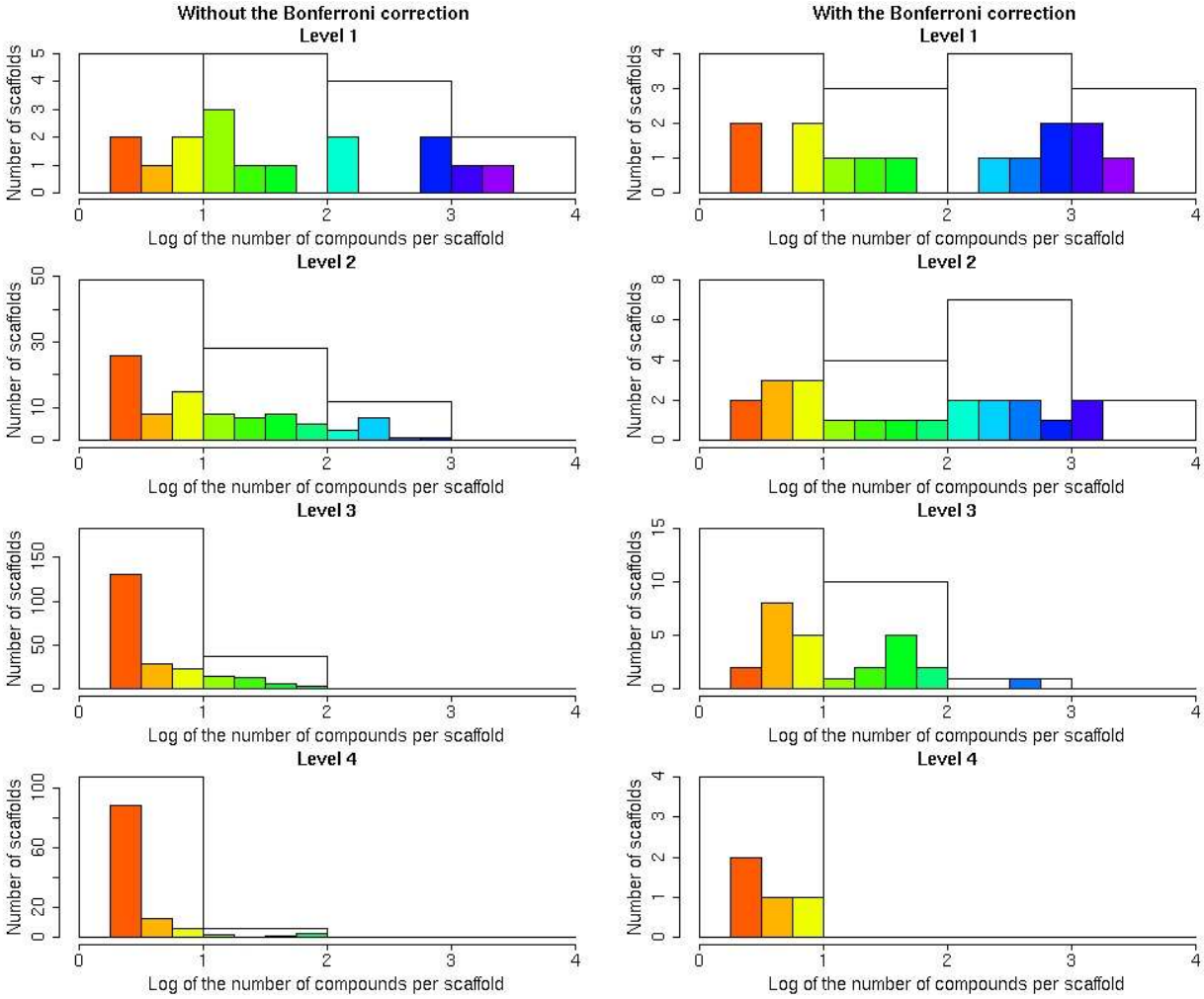

Figure S20 Size of KSP (KS Prediction) significantly active scaffolds for the bioassay 411. The x axis corresponds to the log of the number of compounds per scaffold. The coloured bars are sub-levels of the white bars. Colors are shown to facilitate comparison of results without (left) and with (right) the Bonferroni correction.

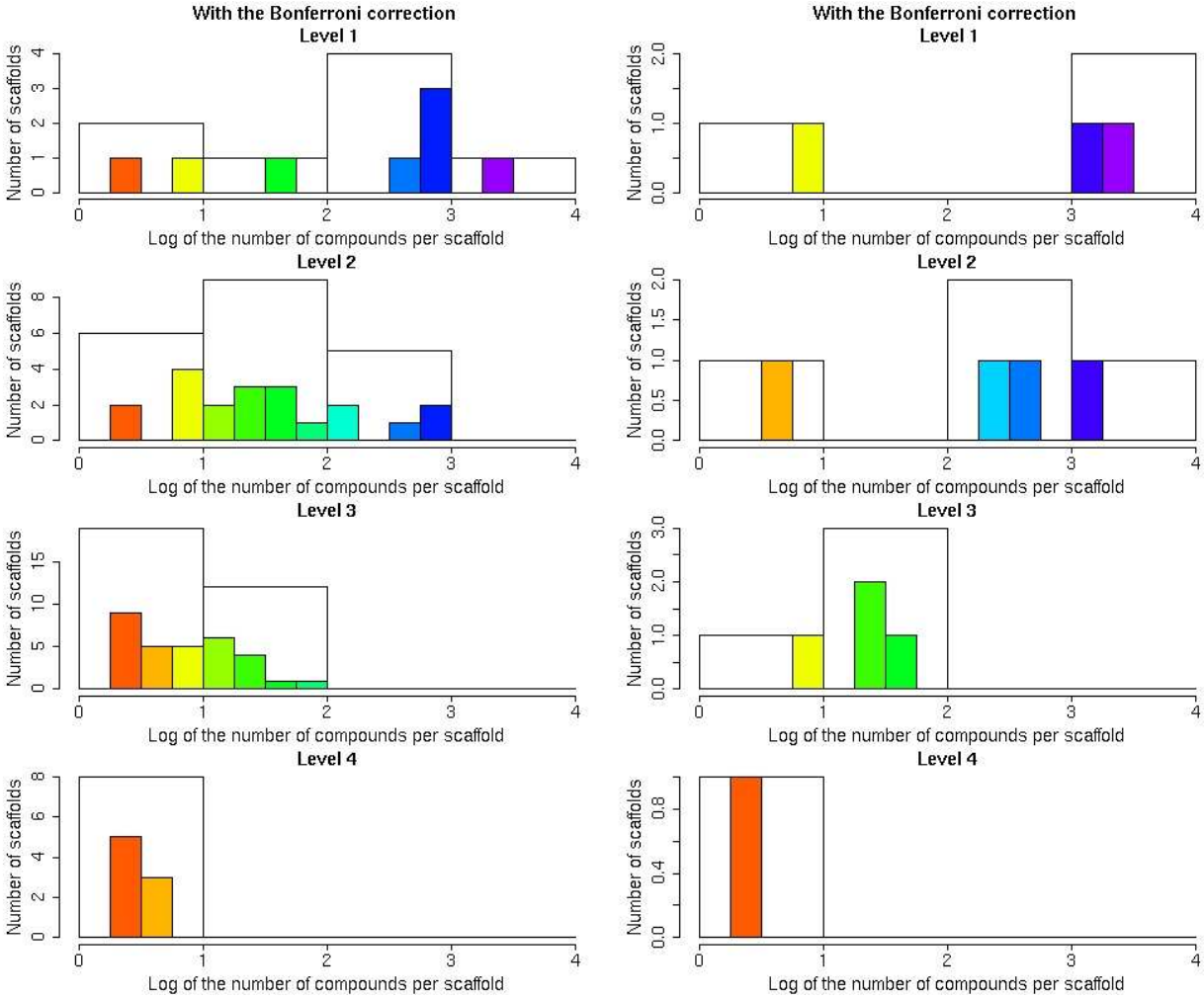

Figure S21 Size of BTP (Binomial Threshold Prediction) significantly active classes for the bioassay 411. The x axis corresponds to the log of the number of compounds per class. The coloured bars are sub-levels of the white bars. Colors are shown to facilitate comparison of results without (left) and with (right) the Bonferroni correction.

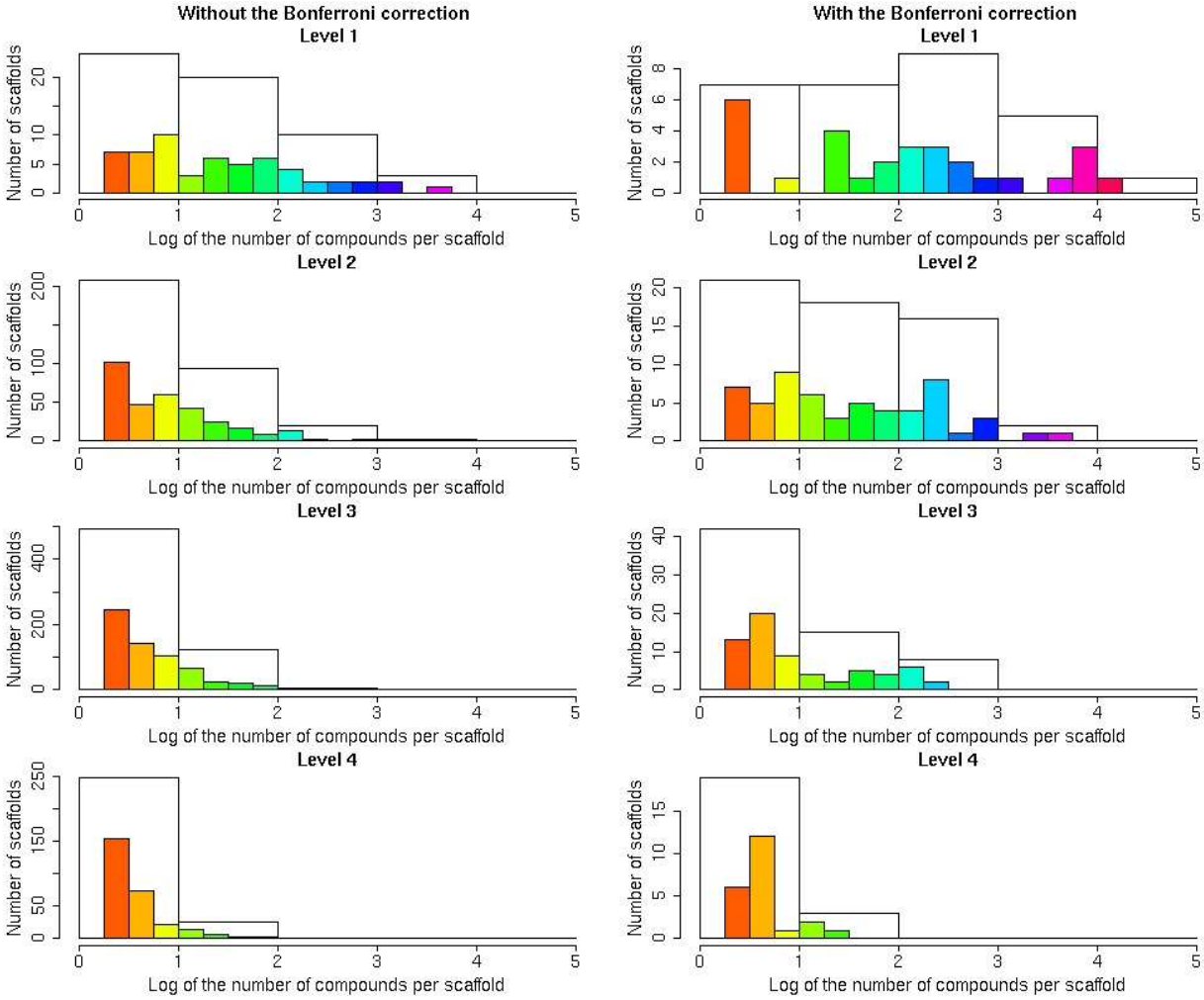

Figure S22 Size of KSP (KS Prediction) significantly active scaffolds for the bioassay 1379. The x axis corresponds to the log of the number of compounds per scaffold. The coloured bars are sub-levels of the white bars. Colors are shown to facilitate comparison of results without (left) and with (right) the Bonferroni correction.

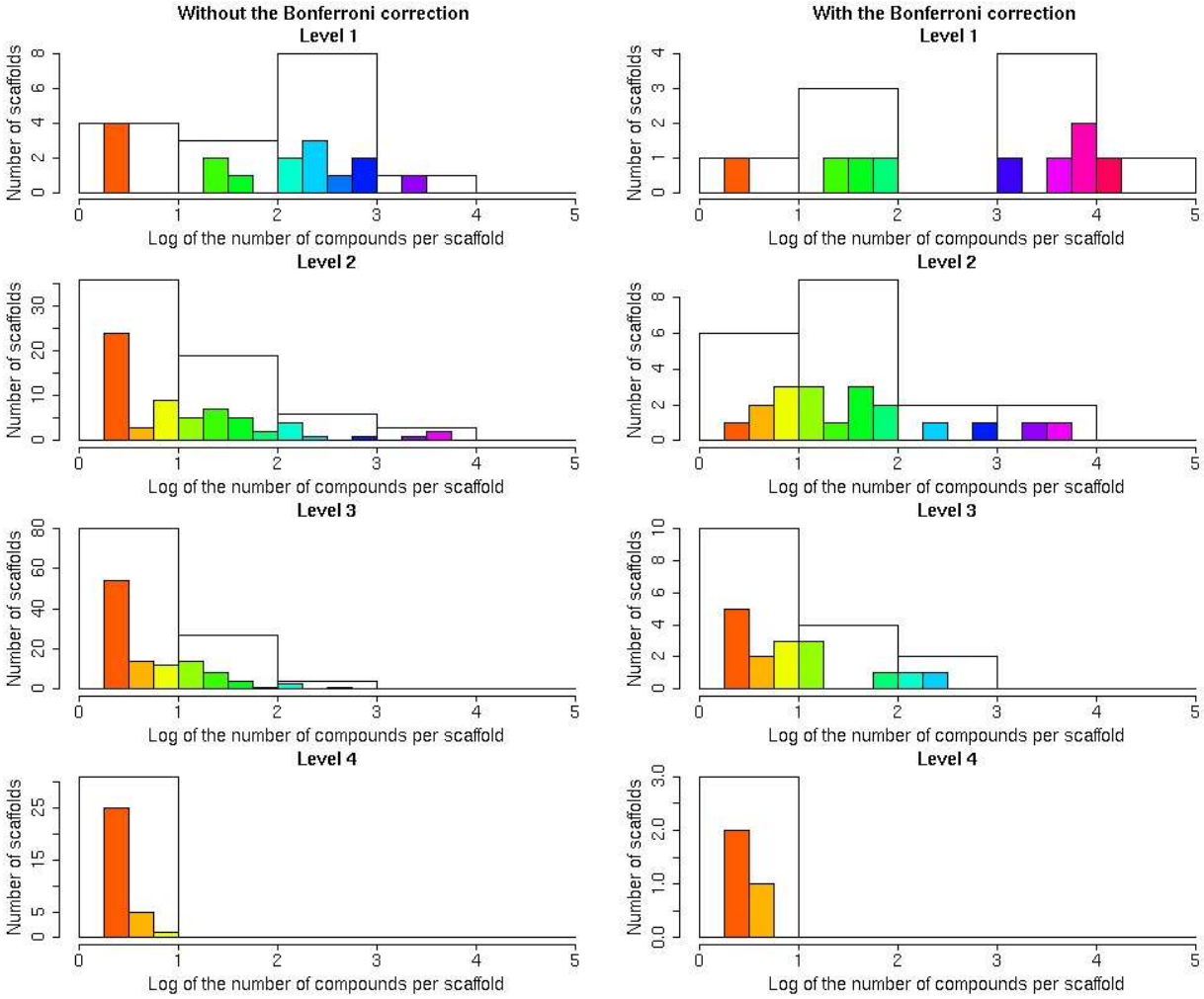

Figure S23 Size of BTP (Binomial Threshold Prediction) significantly active classes for the bioassay 1379. The x axis corresponds to the log of the number of compounds per class. The coloured bars are sub-levels of the white bars. Colors are shown to facilitate comparison of results without (left) and with (right) the Bonferroni correction.

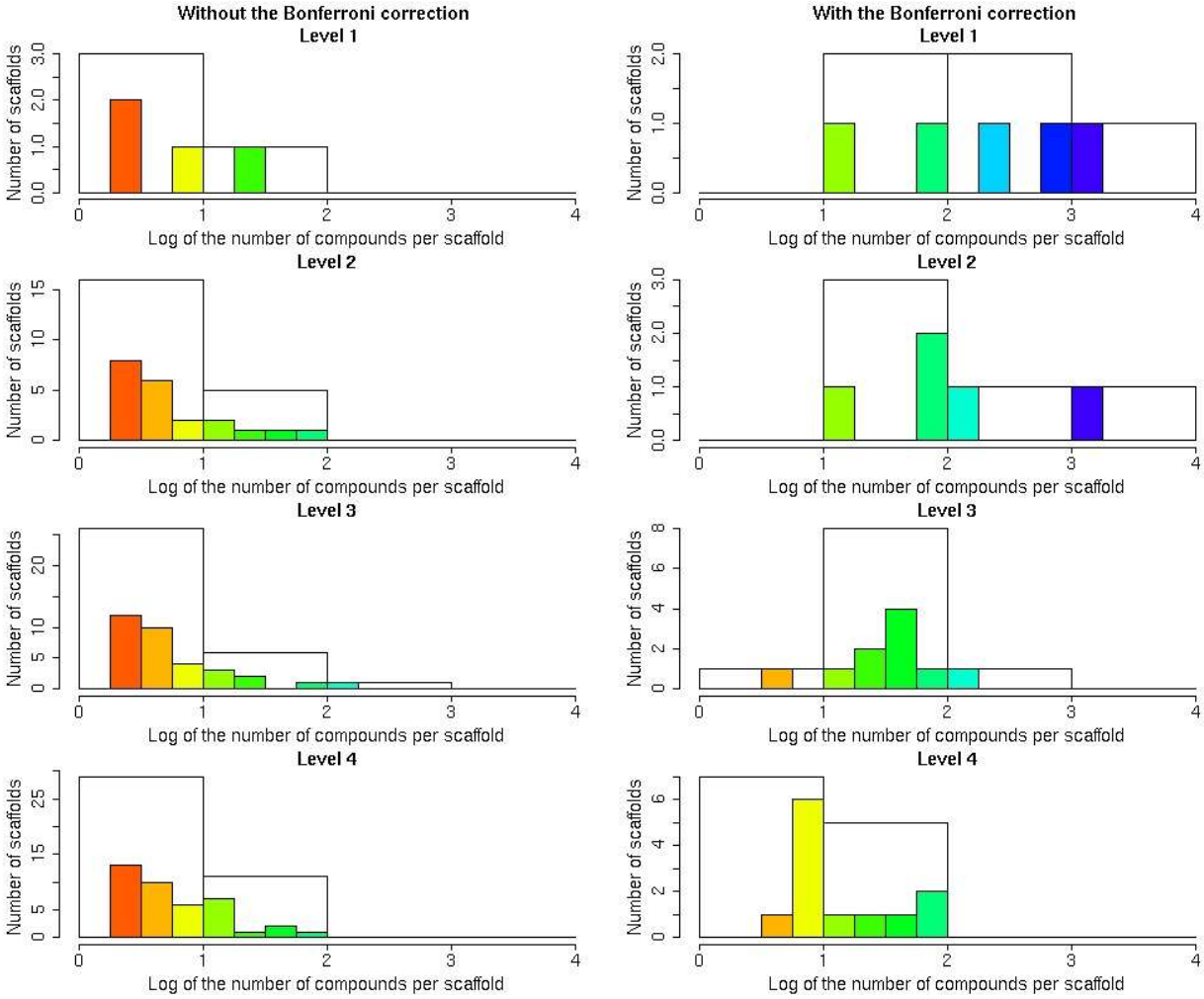

Figure S24 Size of KSP (KS Prediction) significantly active scaffolds for the bioassay 883. The x axis corresponds to the log of the number of compounds per scaffold. The coloured bars are sub-levels of the white bars. Colors are shown to facilitate comparison of results without (left) and with (right) the Bonferroni correction.

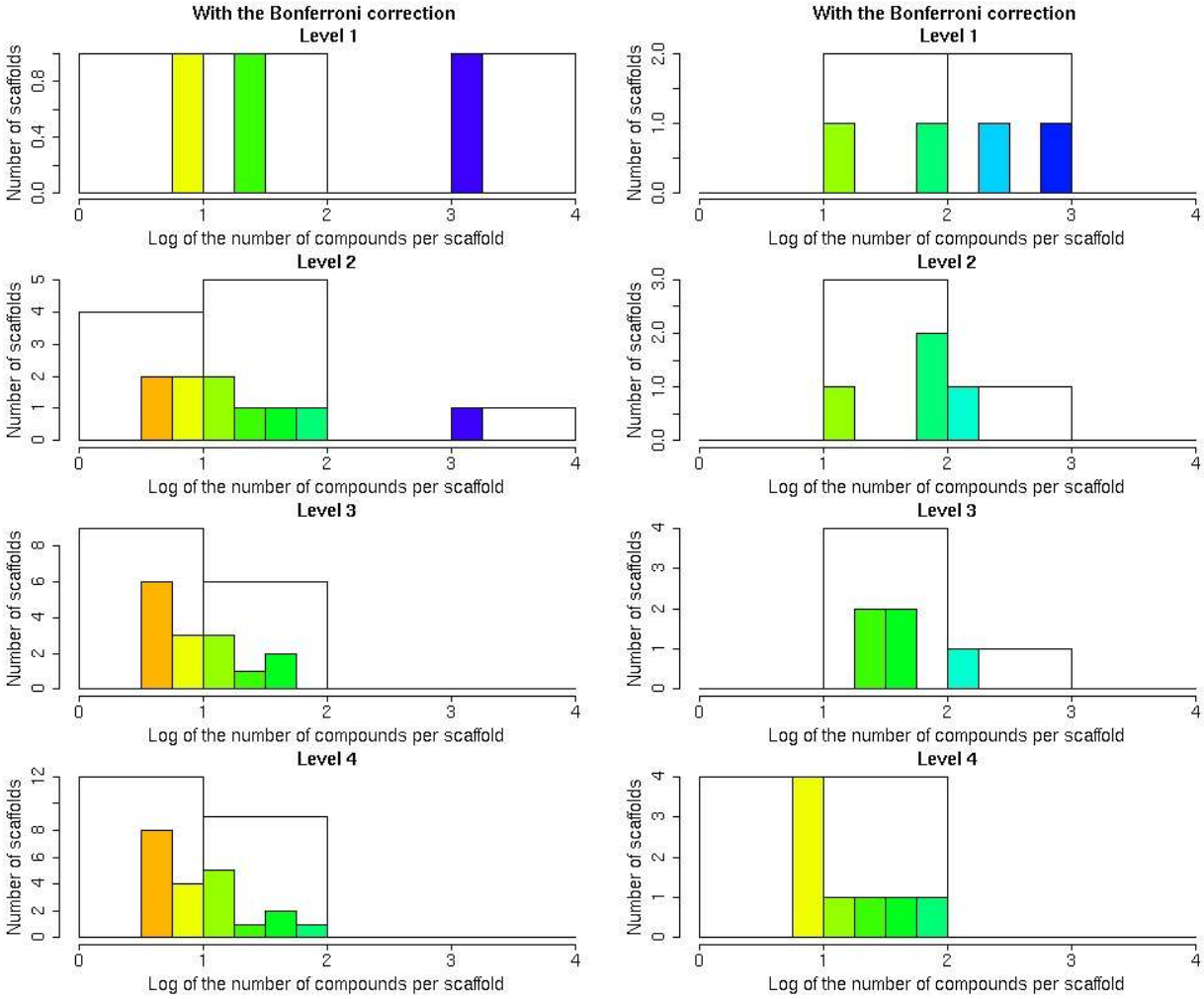

Figure S25 Size of BTP (Binomial Threshold Prediction) significantly active classes for the bioassay 883. The x axis corresponds to the log of the number of compounds per class. The coloured bars are sub-levels of the white bars. Colors are shown to facilitate comparison of results without (left) and with (right) the Bonferroni correction.

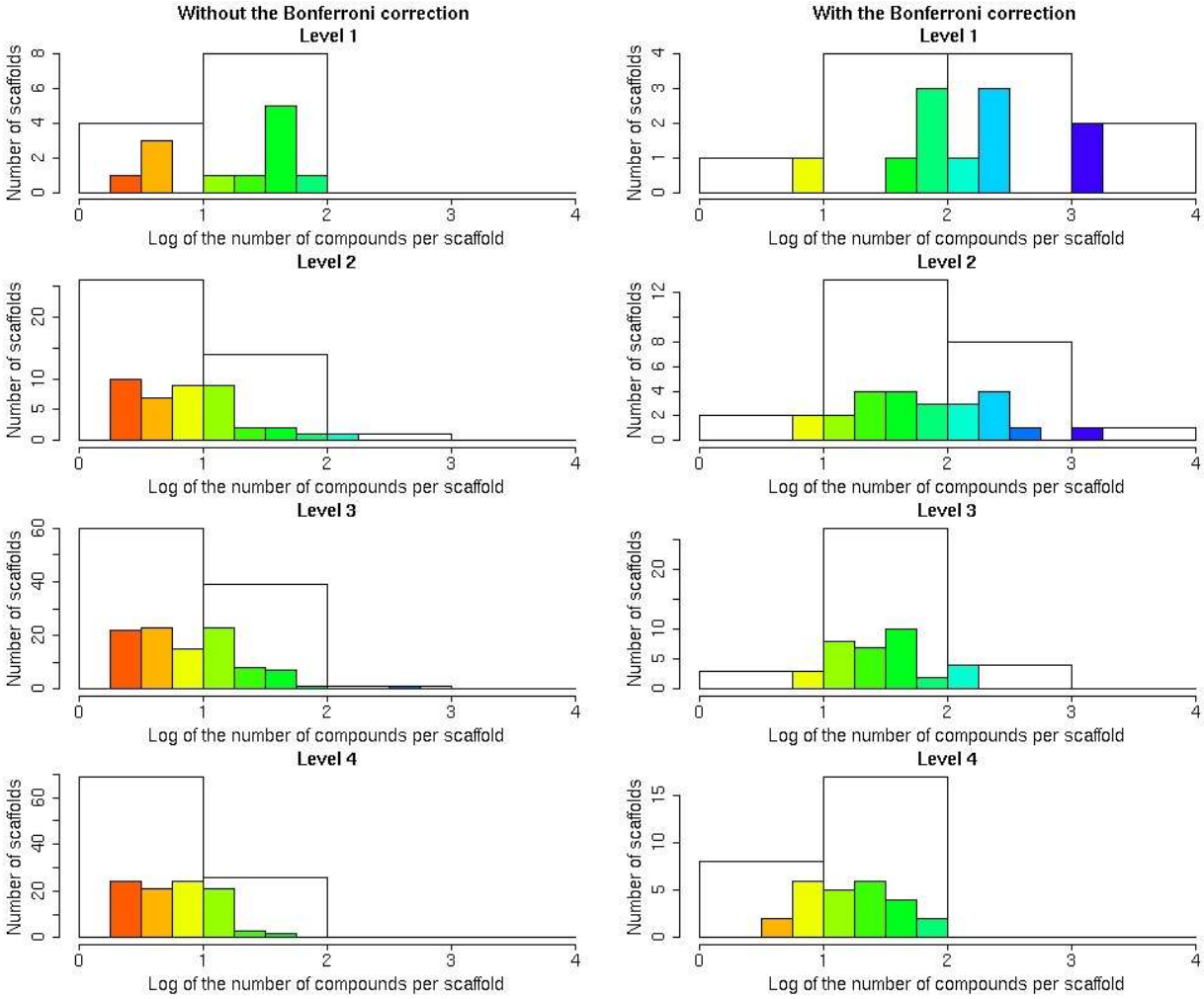

Figure S26 Size of KSP (KS Prediction) significantly active scaffolds for the bioassay 884. The x axis corresponds to the log of the number of compounds per scaffold. The coloured bars are sub-levels of the white bars. Colors are shown to facilitate comparison of results without (left) and with (right) the Bonferroni correction.

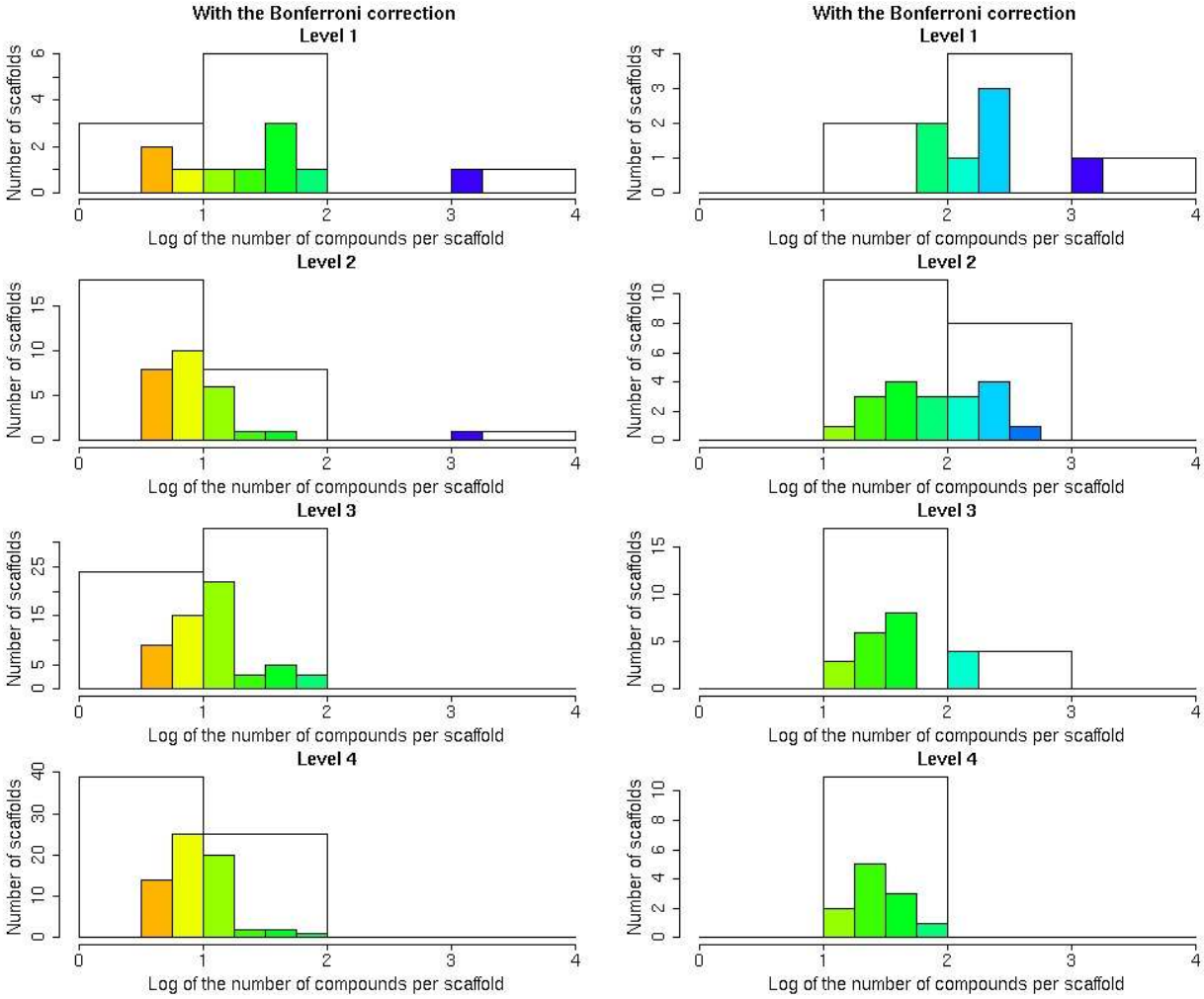

Figure S27 Size of BTP (Binomial Threshold Prediction) significantly active classes for the bioassay 884. The x axis corresponds to the log of the number of compounds per class. The coloured bars are sub-levels of the white bars. Colors are shown to facilitate comparison of results without (left) and with (right) the Bonferroni correction.