Supplementary
KOMB: K-core based characterization of Copy Number Variation in Microbiomes

Advait Balaji[1], Nicolae Sapoval[1], Charlie Seto[2], R.A. Leo Elworth[1], Michael G. Nute[1], Tor Savidge[2], Santiago Segarra[3,†] & Todd J. Treangen[1,†,#]

[1]*Department of Computer Science, Rice University, Houston, Texas, USA.*

[2]*Department of Pathology and Immunology, Baylor College of Medicine, Houston, TX, USA*

[3]*Department of Electrical and Computer Engineering, Rice University, Houston, Texas, USA.*

[†]These authors share senior authorship

[#]Corresponding author contact: treangen@rice.edu

Supplementary Data

SD1. L1 norm analysis of KOMB Profiles from HMP sites

We calculate the L1 norm of KOMB profiles both within and between each body site. For samples within in each body site, anterior nares had the highest average distance (1.15) followed by buccal mucosa (1.14), supragingival plaque (1.00) and stool (0.89). The aggregate distances (see implementation) between body sites were also calculated. We observed that anterior nares and buccal mucosa had the highest aggregate distance (1.09). Overall, anterior nares also had the greatest separation from supragingival plaque (1.08) and stool (0.66). Supragingival plaque and buccal mucosa were closest in terms of aggregate distance (0.32).

SD2. GO terms obtained from the HMP analyses can be found at: https://tinyurl.com/3s5fe99k

SD3. L1 norm analysis of KOMB Profiles fro Subjects in Gut-microbiome analysis

We calculated the average pairwise L1 norm for samples from each subject. Bugkiller (0.30), Scavenger (0.46), Tigress (0.34), and Daisy (0.38) showed higher variability in the early samples as compared to other subjects Alien (0.12) and Peacemaker (0.16) who exhibited fairly consistent profiles. We generally observe intra-sample similarity over the three time points and also observe some similarities between profiles based on gender also reported by previous studies [1, 2]. Aggregate profiles from Daisy and Tigress were closest to each other (0.26) than to any of the male subjects. The average distances of the male subjects to Daisy and Tigress were; Alien (0.58, 0.33), Bugkiller (0.62,0.37), Peacemaker (0.61,0.36) and Scavenger (0.52,0.27) respectively while the average pairwise distance between profiles from the male subjects was 0.12.

SD4. Kraken2 analysis on FMT samples can be found at: https://tinyurl.com/yhr9w8hv

SD5: KOMB analysis on Gut Microbiome data from cohort healthy, IBD and obese patients.

We thank the reviewer for the suggestion to run KOMB on the dataset given by Greenblum et. al (2015) [3]. To better understand if the KOMB topologies capture relevant CNVs, we decided to download the data from Danish and Spanish individuals analyzed in the study . We considered 258 experiments related to the study with 137 associated with healthy individuals, 44 associated with IBD and 77 associated with obese. The study found 24 different Kegg Ortholog (KO)-cluster pairs (KCs) across 6 different genome clusters to be associated with the IBD condition and 3 KO-cluster pairs across 2 different genome clusters. It is important to note here that these clusters were found by the authors to be specific to this dataset as the KCs found in the chinese cohort only yielded 3 of the 24 KCs pairs that were common to IBD samples and none in the obesity associated samples.

We first analyzed the KOMB profiles of healthy, IBD and obese samples. We calculated and plotted the median shell number for all experiments per sample type (Figure below). To compare the distribution of the medians, we calculated the Mann-Whitney U test and found that the p–value between healthy and IBD and healthy and obese were statsitically significant (0.002 and 0.001 respectively) wheareas between IBD and obese was not (0.18, not significant, n.s) as seen in Figure S2.

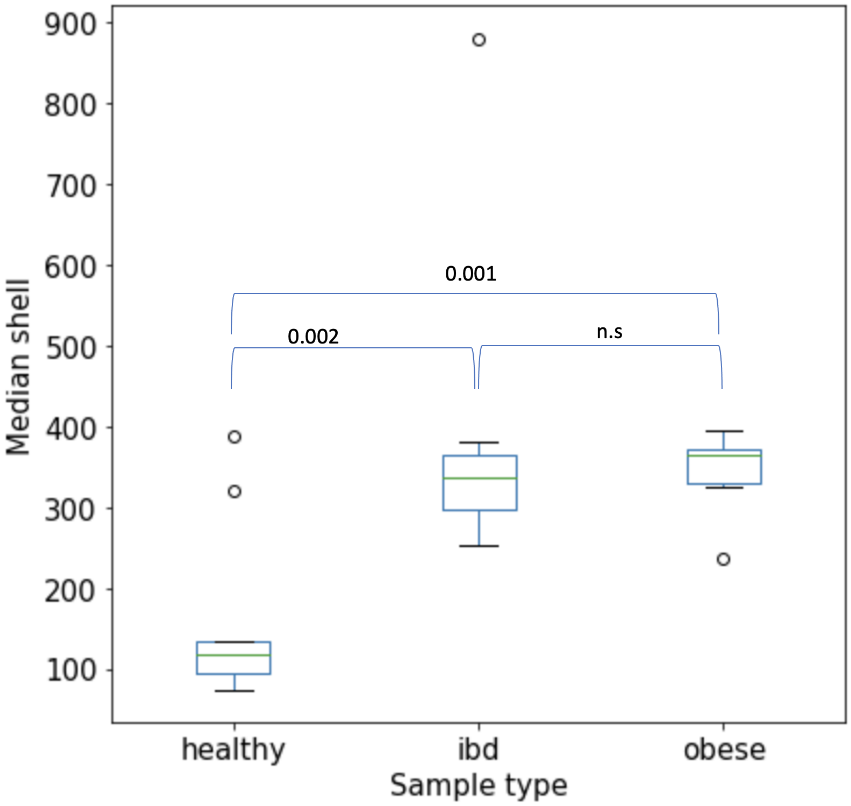


Figure S1. Box plots showing the median shell numbers of all samples belonging to healthy, IBD and obese individuals from the study. Values between box plots indicate the Mann-Whitney U p-values.

We then reasoned that given KOMB's ability to capture anomalous unitigs, we should observe a change in associated KCs in these unitigs. This is because the anomalous unitigs capture high core or high degree unitigs. Due to limitations of accessing KEGG FTP site (subscription only) and non-availability of online tools in the KEGG website to annotate a large number of DNA sequences with KOs, we developed an alternate strategy to annotate anomalous unitigs. First, all anomalous unitigs having length greater than 150bp from a given host-associated sample were pooled. The length cut-off was chosen as the reads were too small (75bp or 44bp) for confident protein assignment downstream. Second, taxonomic classification was performed using Kraken2 on the standard DB (2021) and the unitigs were separated based on the classification of gene clusters done in the study. Third, the classified unitigs were mapped to the UniRef100 DB using DIAMOND blastx in fast mode and top 1% of the hits were retained. Finally, we scraped all UniProt ids for a given KO from the website and mined for them in the DB hits. We then calculated the number of unitigs containing the KO by analyzing the diamond output. We computed a score to compare "enrichment" of KCs to sample-types by normalizing the number of unitigs containing KOs by the total unitigs belonging to the sample as described below in Eqn 1:

$$\text{Score} = \frac{\text{Number of unitigs having the associated KO assigned to the cluster}}{\text{Total number of anomalous unitigs assigned to the cluster}} \times 100 \tag{1}$$

From the data, we could correctly identify 16 out of the 24 KCs belonging to 3 out of the six clusters in the IBD samples. As seen in the Figure , we observed that we predominantly identified clusters having multiple genomes such as c2 and c5 for IBD and c49 for obese (though the presence of the KO in obese sample was very weak). We further identified that the score reflected the direction of enrichment for c2, c5 and c49 (increase in IBD and obese indicated by asterix) whereas in c55 (*Bifidobacterium adolescentis*) which was the only cluster with a single genome identified we could not capture the decrease in KOs as reported by the authors. Both c2 and c5 contain multiple IBD-associated *Bacteroides* species especially *Bacteroides vulgatus* and *Bacteroides uniformis* and their highly variable KOs were corroborated by running KOMB.

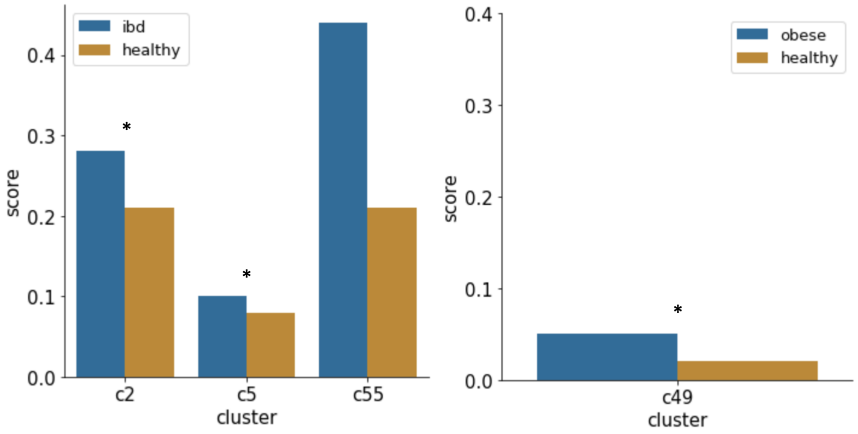


Figure S2. Score obtained per cluster of genomes on highly variable KOs Higher score indicates higher number of unitigs (normalized) had KOs beonging to the given cluster that were found enriched for disease phenotypes in hte study. Clusters with asterix indicates that the direction of enrichment reflects the conclusions drawn by Greenblum et al. (2015)

References

[1] JA Santos-Marcos, C Haro, A Vega-Rojas, JF Alcala-Diaz, H Molina-Abril, A Leon-Acuña, J Lopez-Moreno, BB Landa, M Tena-Sempere, P Perez-Martinez, et al. Sex differences in the gut microbiota as potential determinants of gender predisposition to disease. *Molecular nutrition & food research*. **63**: 1800870.

[2] F Fransen, AA van Beek, T Borghuis, B Meijer, F Hugenholtz, C van der Gaast-de Jongh, HF Savelkoul, MI de Jonge, MM Faas, MV Boekschoten, et al. The impact of gut microbiota on gender-specific differences in immunity. *Frontiers in immunology*. **8**: 754.

[3] S Greenblum, R Carr, and E Borenstein. Extensive strain-level copy-number variation across human gut microbiome species. *Cell*. **160**: 583–594.

Supplementary Tables

Table S1. Average and Standard deviation of the number of reads per sample type in the Human Microbiome Project (HMP) dataset.

Sample type	Number of Reads	
	Average	**Standard deviation**
Anterior nares	598662.96	626816.93
Bucccal Mucosa	6108735.22	8415476.78
Supragingival plaque	24883927.41	11842439.49
Stool	49012875.08	9566369.2

Table S2. Time and memory usage for KOMB and MetaCarvel . Shakya: Shakya et al (2013); HMP (Av); average across HMP samples, TGM(Av); average across Temporal Gut Microbiome samples and FMT (Av); average across FMT samples. For the average, samples having approximately the average number of reads were chosen as representatives for benchmarking . The timings for MetaCarvel include assembly and mapping for an accurate comparison to the KOMB pipeline. Both MetaCarel including the data preparation tools and KOMB were run with 20 threads and k-mer size 51.

	MetaCarvel			**KOMB**		
	Wall Clock	**CPU time**	**Memory**	**Wall Clock**	**CPU time**	**Memory**
Shakya	79m47s	1023m15s	22.21 GB	77m50s	1296m21s	25.29 GB
HMP (Av)	21m10s	310m25s	11.27 GB	17m44s	293m8s	9.59 GB
TGM (Av)	23m53s	420m53s	15.82 GB	15m17s	264m4s	13.22 GB
FMT (Av)	46m34s	627m53s	17.39 GB	57m3s	971m34s	13.65 GB

Supplementary Figures

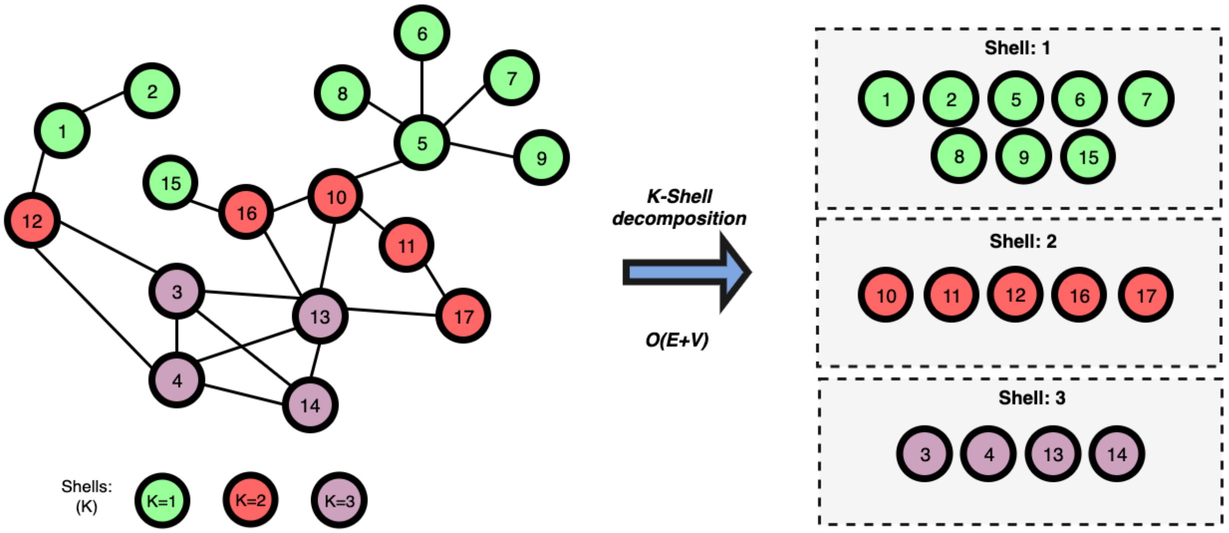


Figure S1. K-core decomposition of a graph into K-shells. The algorithm starts by considering all the vertices of degree 1. It iteratively removes those vertices and continues the execution on the resulting induced subgraph removing vertices having degree 1 after every iteration. Once the induced subgraph has no vertices of degree 1, this process stops and all discarded vertices are marked as belonging to the 1-shell (green). Then the process continues, now considering vertices of degree 2 to obtain the 2-shell (red) and, subsequently, the 3-shell (purple). The last shell is a dense subgraph of the original graph.

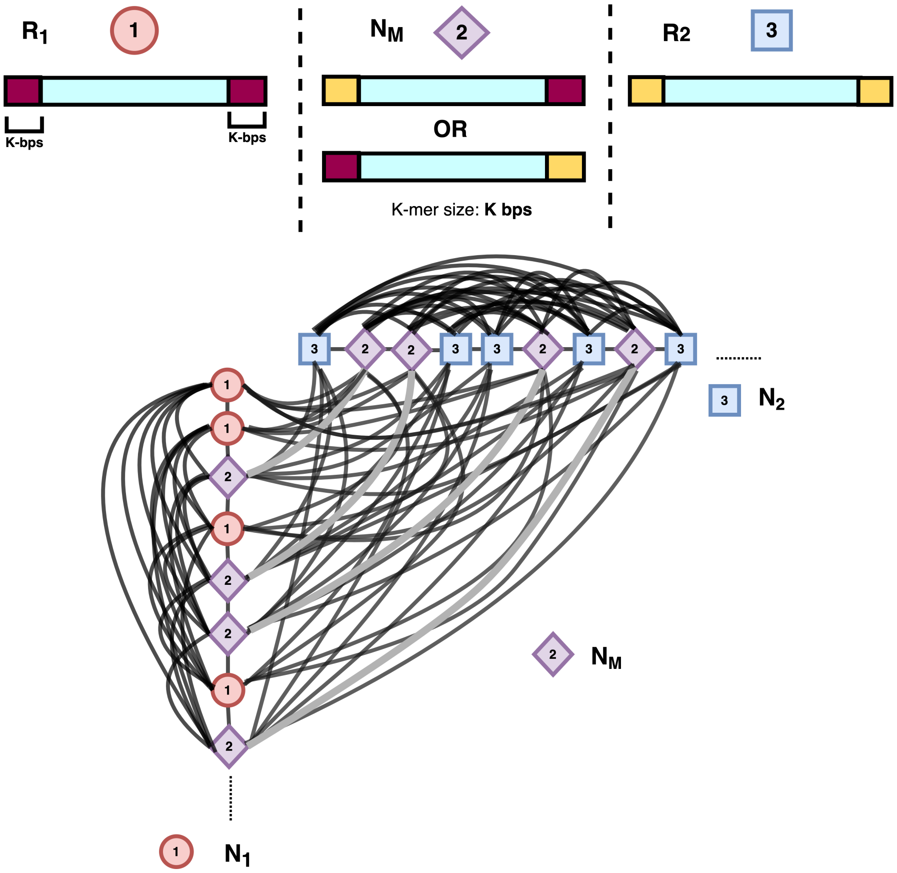


Figure S2. Types of unitigs in a genome with two repeat families and expected shell profiles in corresponding unitig graphs. The type of profile we observe depends on the relative lengths of the repeats and insert size. If the insert size is greater than the length of the repeat, the mixed repeats (N_m) will be connected to each other whereas if the insert size is smaller than the length of the repeat then it is not possible to map across the two mixed repeat unitigs and, hence, they will not be connected by an edge in the unitig graph. The black edges are present for both cases whereas the gray edges are only present when the repeat length is less than the insert length.

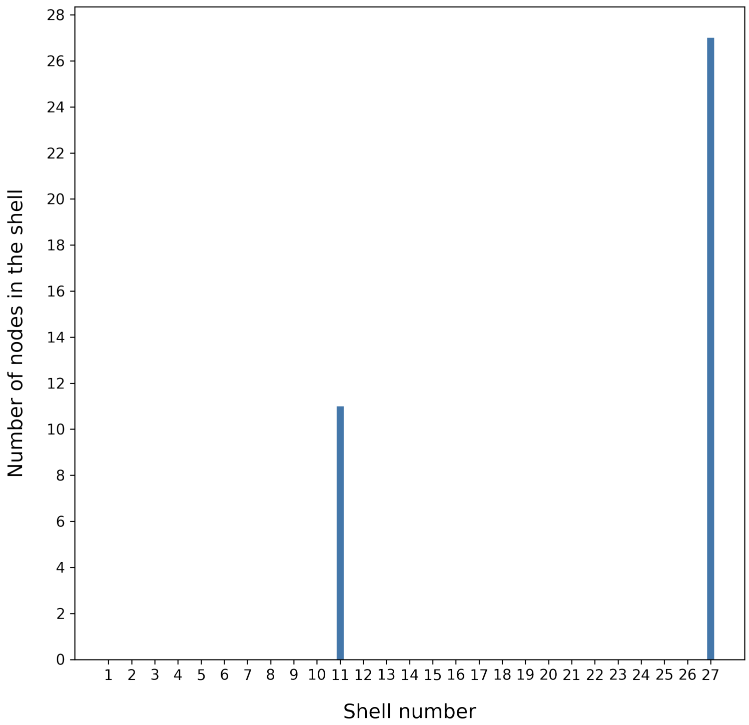
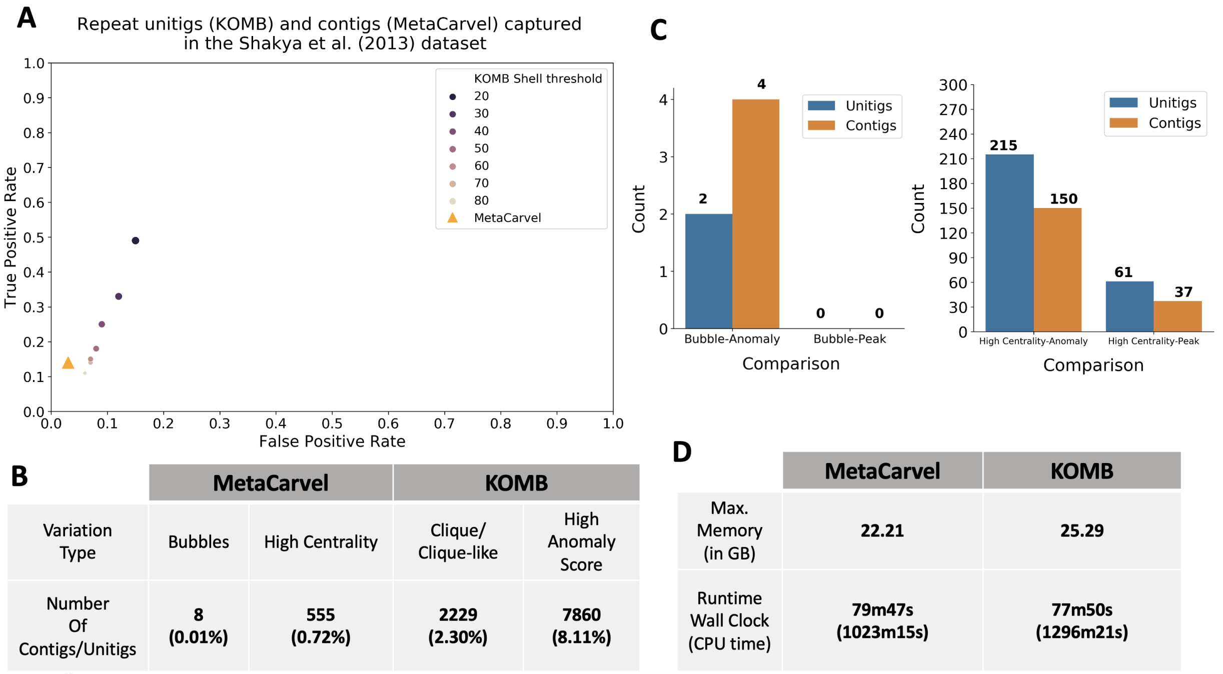


Figure S3. Validation of KOMB on simulated data. KOMB profiles on a random backbone with 10×400bp and 25×400bp identical repeats. We observe peaks at approximately the copy numbers of the repeats at shell numbers 11 and 27.



B

	MetaCarvel		KOMB	
Variation Type	Bubbles	High Centrality	Clique/ Clique-like	High Anomaly Score
Number Of Contigs/Unitigs	**8 (0.01%)**	**555 (0.72%)**	**2229 (2.30%)**	**7860 (8.11%)**

D

	MetaCarvel	KOMB
Max. Memory (in GB)	**22.21**	**25.29**
Runtime Wall Clock (CPU time)	**79m47s (1023m15s)**	**77m50s (1296m21s)**

Figure S4. Comparison of detection of repeats and variants in MetaCarvel and KOMB (A) TPR and FPR of repeat identification MetaCarvel contigs (triangle) and KOMB unitigs (dots). The shade of dots represent different shell thresholds for KOMB. (B) Table comparing number of bubbles and high centrality contig variants obtained through MetaCarvel and peaks (cliques/clique-like) as well as anomalous unitigs reported by KOMB. Only peaks afte shell 70 were considered. (C) Comparing the overlap of these four sets of variants to check the number of contigs or unitigs that were similar in content (*nucmer* $\geq$ *95%*). (D) Runtime and memory of MetaCarvel and KOMB

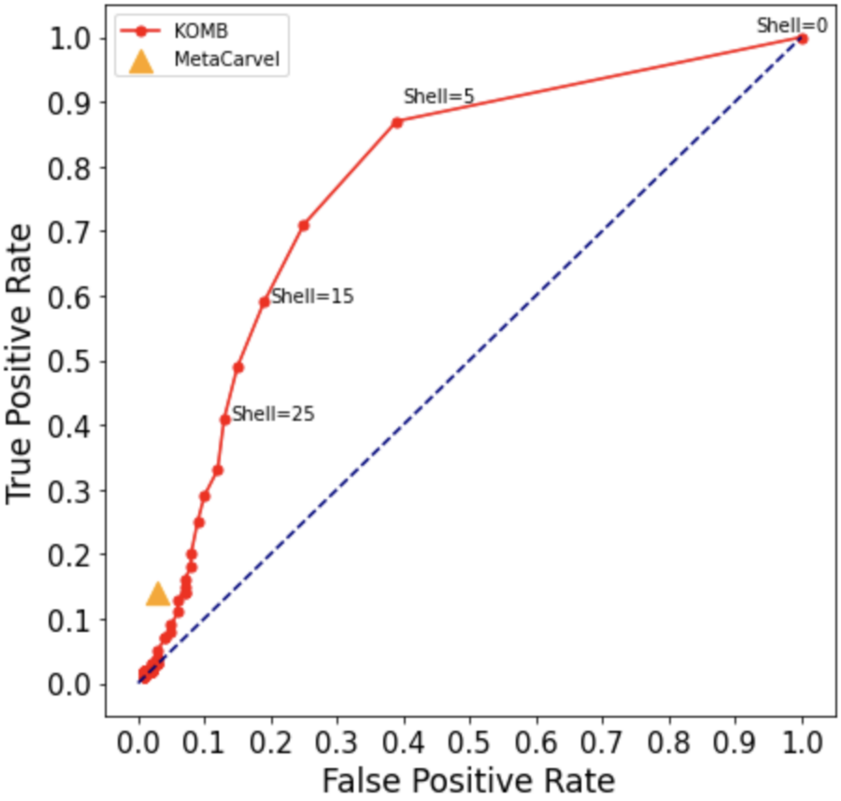


Figure S5. ROC curve of KOMB vs MetaCarvel Initial shells in KOMB contain low copy number repeats but also a lot of non-repeat unitigs that increases FPR.

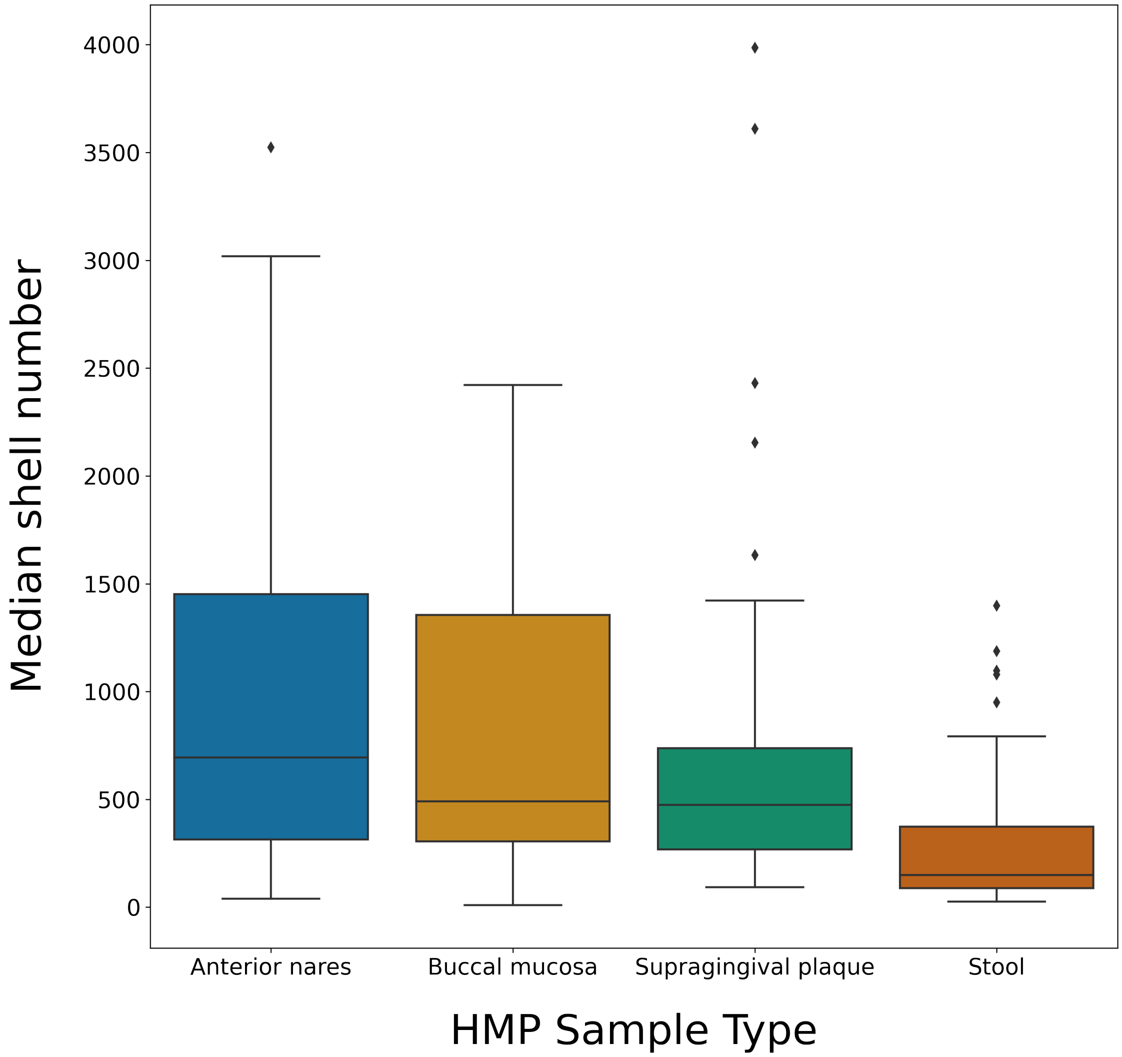
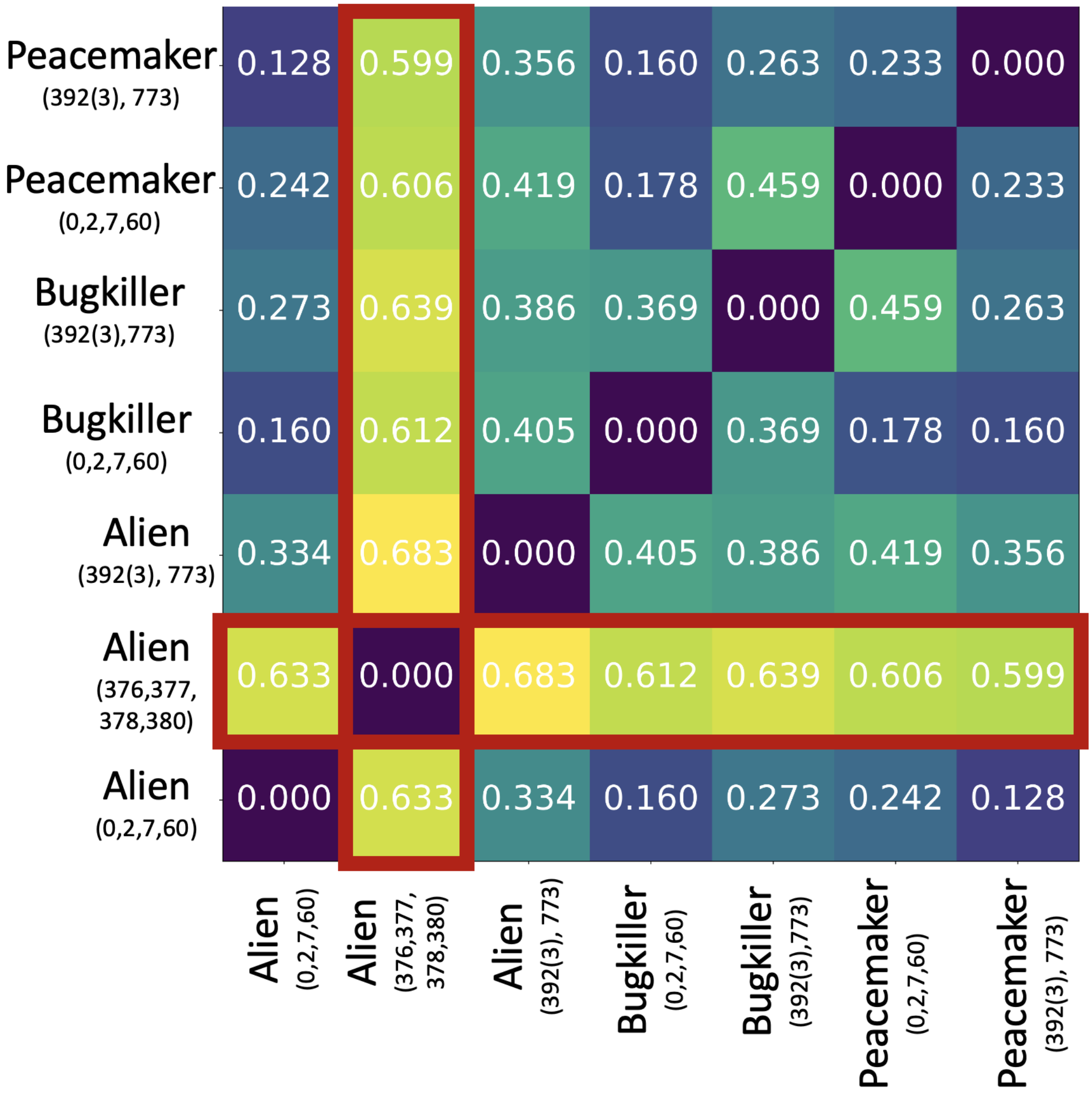


Figure S6. Median number of shells for samples in each body site Box plots shwing the median number of shells for 50 samples for each body site.



	Alien (0,2,7,60)	Alien (376,377, 378,380)	Alien (392(3), 773)	Bugkiller (0,2,7,60)	Bugkiller (392(3),773)	Peacemaker (0,2,7,60)	Peacemaker (392(3), 773)
Peacemaker (392(3), 773)	0.128	0.599	0.356	0.160	0.263	0.233	0.000
Peacemaker (0,2,7,60)	0.242	0.606	0.419	0.178	0.459	0.000	0.233
Bugkiller (392(3),773)	0.273	0.639	0.386	0.369	0.000	0.459	0.263
Bugkiller (0,2,7,60)	0.160	0.612	0.405	0.000	0.369	0.178	0.160
Alien (392(3), 773)	0.334	0.683	0.000	0.405	0.386	0.419	0.356
Alien (376,377, 378,380)	0.633	0.000	0.683	0.612	0.639	0.606	0.599
Alien (0,2,7,60)	0.000	0.633	0.334	0.160	0.273	0.242	0.128

Figure S7. Heatmap showing L1 norm of KOMB profiles in the Voigt et al. dataset. Each row and column represents a subject and days (four each) for which the samples are considered (in parenthesis). Day 392 had 3 samples in the dataset which are all considered here. The samples represented by Alien (Days 376, 377, 378, and, 380), also marked in red, are the ones collected during antibiotic perturbation. Higher total variation of probability denotes greater distance between two distributions. Days 0,2,7,60 correspond to the initial time points and Days 392(3) and Day 773 correspond to later time points.

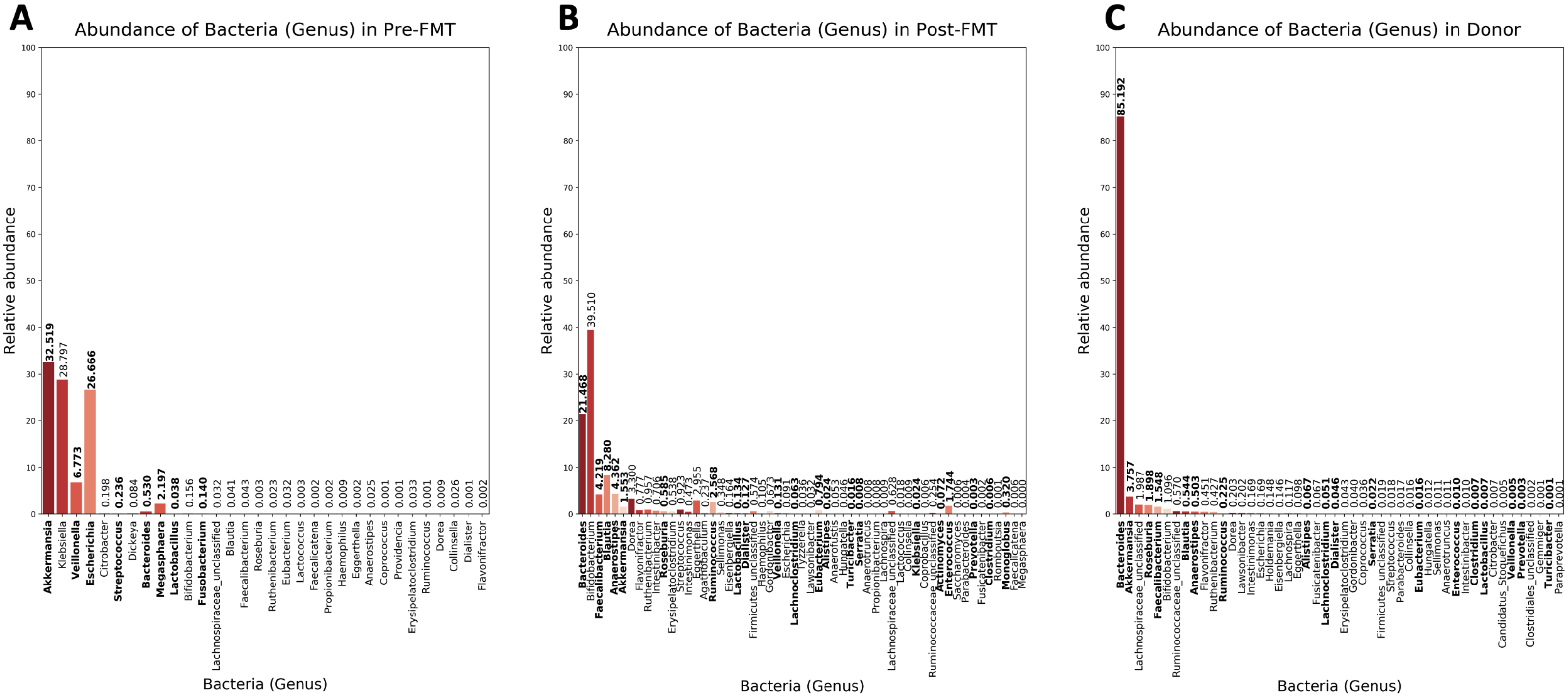


Figure S8. Average Relative abundance of taxa in FMT Samples Average Relative abundance of taxa (at genus level) as reported by MetaPhlAn3 for (A) Two Pre-FMT samples (B) Two Post-FMT samples (C) One Donor sample. In bold in (A) and (B) are the taxa marked by KOMB as anomalous in the respective samples. In (C) the taxa marked are the taxa found anomalous in Post-FMT by KOMB to indicate their abundances in the Donor sample.

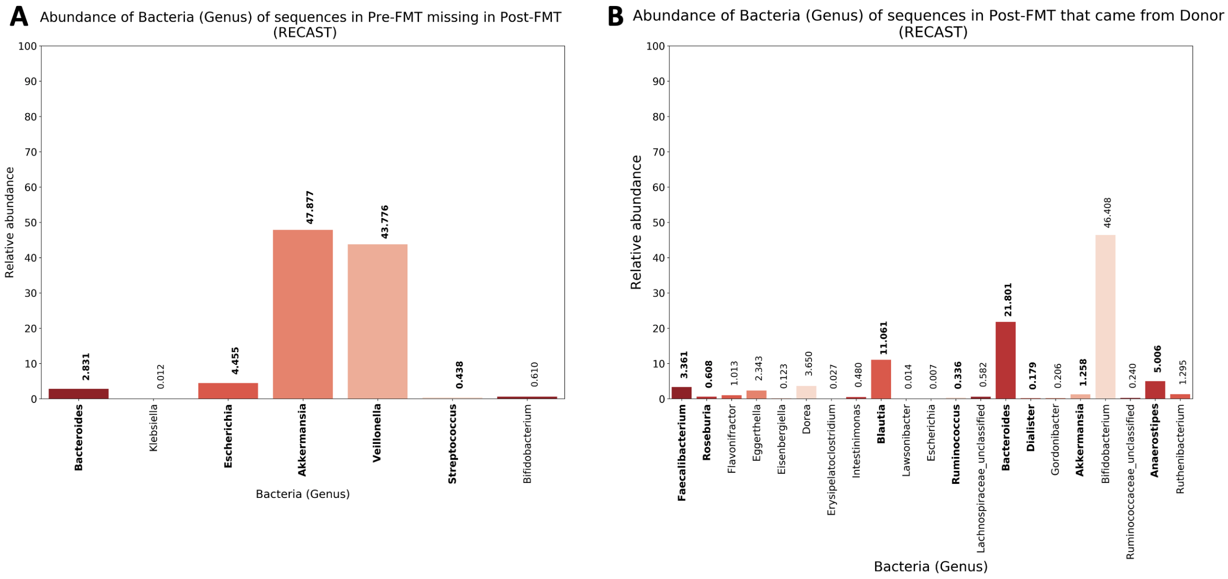


Figure S9. Comparison of taxa from anomalous unitigs in KOMB and taxa found to important by RECAST (A) Average relative abundances of taxa found in Pre-FMT that were be missing from Post-FMT by RECAST. (B) Average relative abundances of taxa in Post-FMT found to be from Donor by RECAST. In bold in (A) and (B) are the taxa marked by KOMB as anomalous in the respective samples. Some keystone taxa in (B) that we marked anomalous by KOMB but not present in RECAST results are discussed in text.