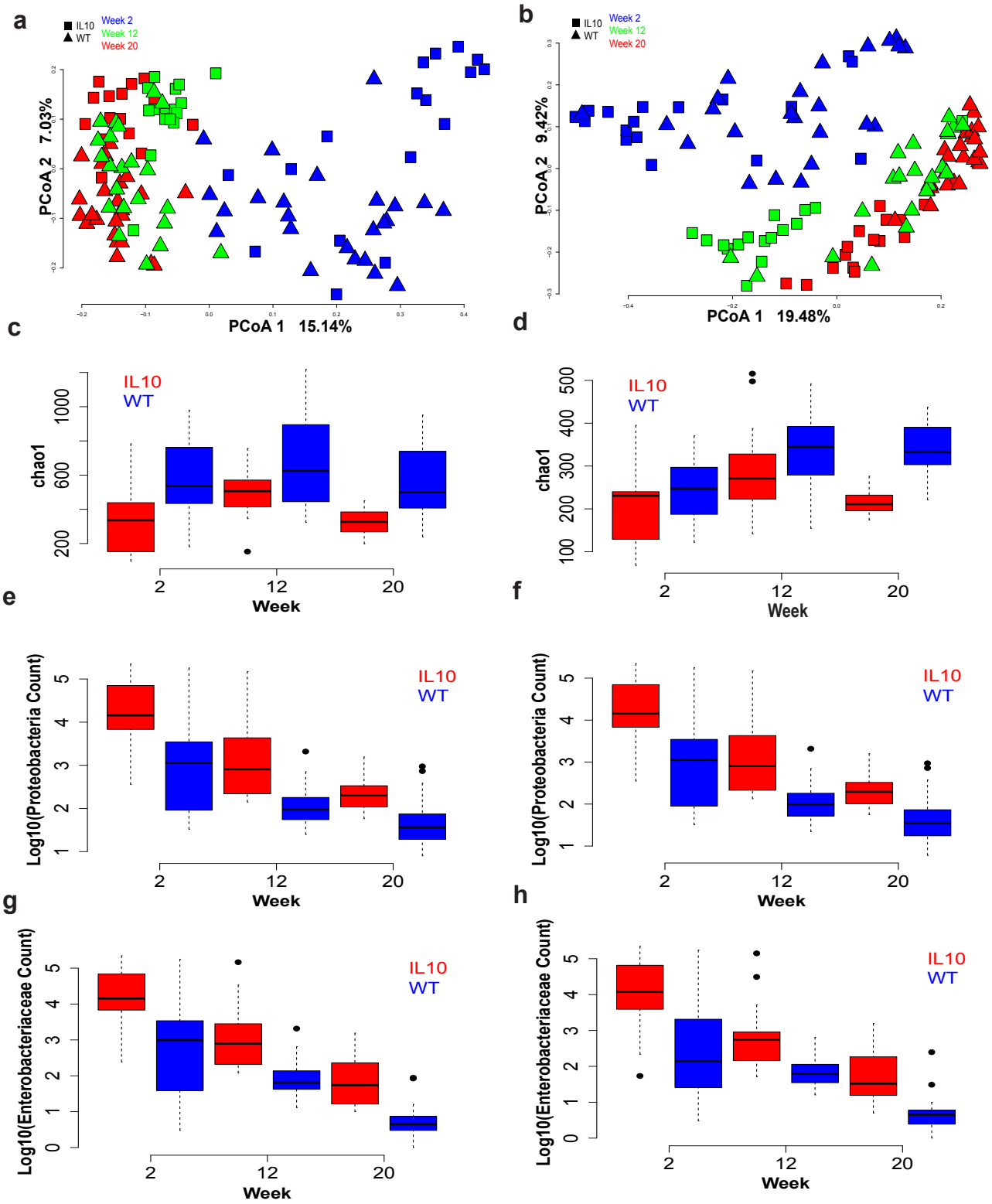
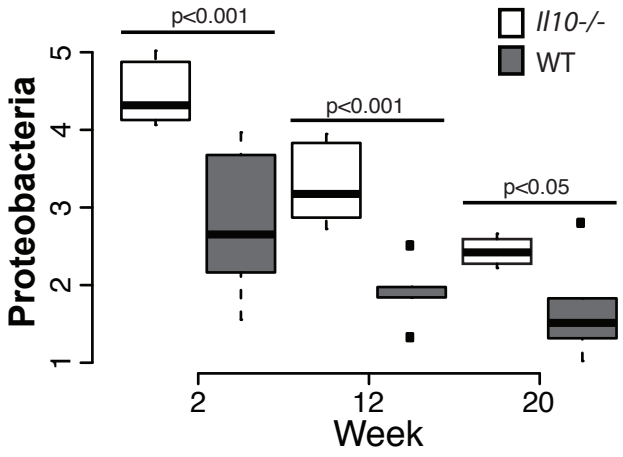


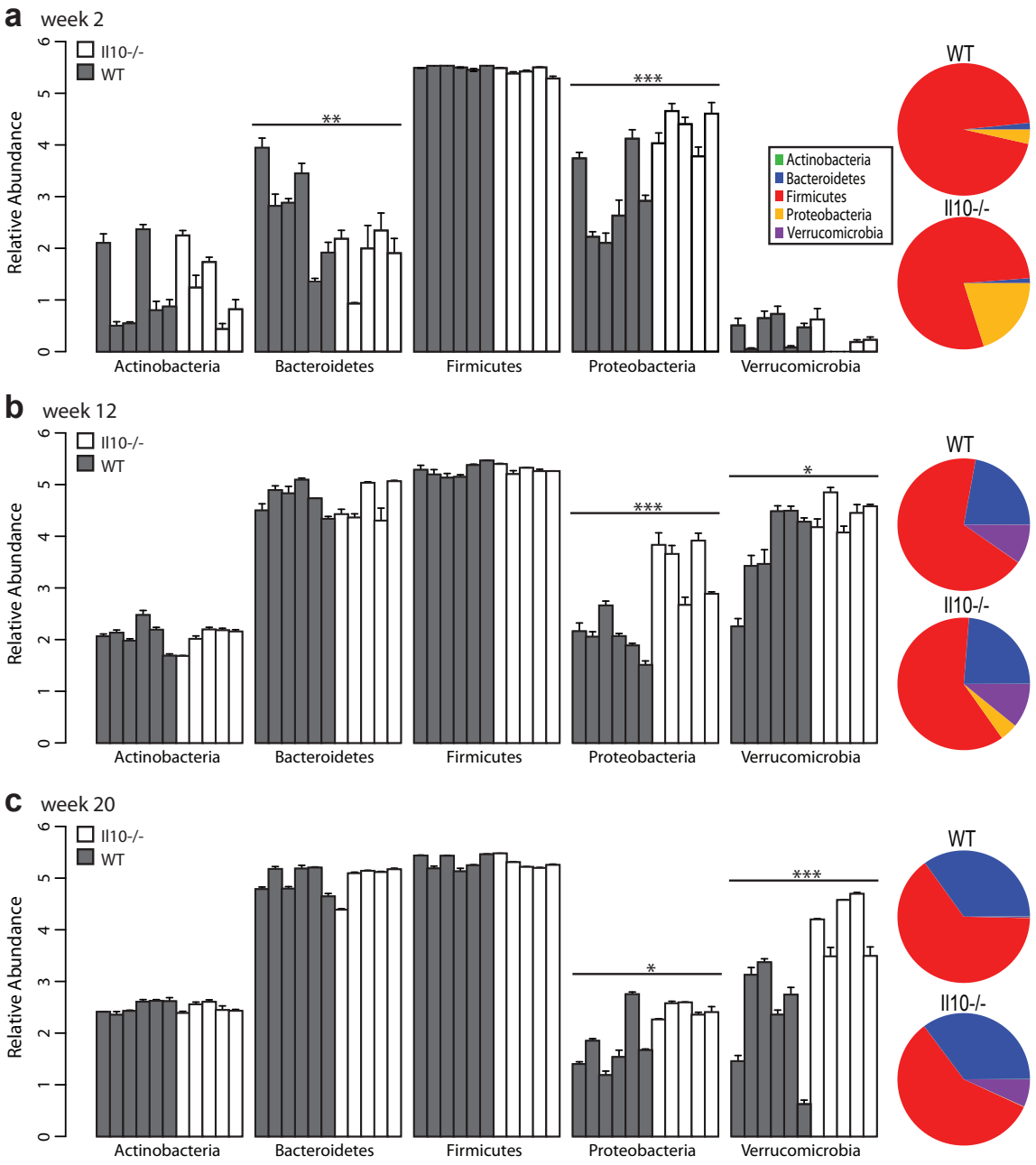

**Supplementary Figure 1 – Change in microbial community composition over time analyzed using QIIME pipeline**

**a**) Unweighted-Unifrac PCoA generated from QIIME de novo OTU picking approach (BIOM file rarefied to 15,941 sequences per sample) **b**) Unweighted-Unifrac PCoA generated from QIIME close-reference OTU picking approach using greengenes 13_5 release (BIOM file rarefied to 15,660 sequences per sample). **c**) Richness (choa1) calculated from QIIME de novo OTU picking approach, BIOM file rarefied to 15,941 sequences per sample. **d**) Richness (choa1) calculated from QIIME close-reference OTU picking approach, BIOM file rarefied to 15,660 sequences per sample. **e**) Relative abundance of Proteobacteria from QIIME de novo OTU picking approach, taxonomy assignment was done using RDP with minimum confidence set to 80%. **f**) Relative abundance of Proteobacteria from QIIME close-reference OTU picking approach, taxonomy assignment using greengenes 13_5 release. **g**) Relative abundance of Enterobacteriaceae from QIIME de novo OTU picking approach, taxonomy assignment was done using RDP with minimum confidence set to 80%. **h**) Relative abundance of Enterobacteriaceae from QIIME close-reference OTU picking approach, taxonomy assignment using greengenes 13_5 release.

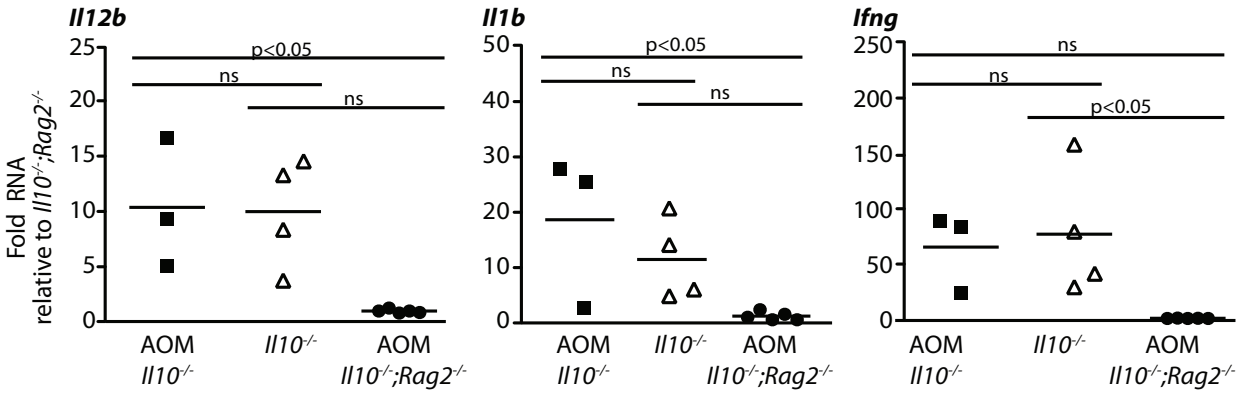



**Supplementary Figure 2 – Relative abundance of Proteobacteria over time in WT vs. *Il10*$^{-/-}$ mice**

Increased abundance of Proteobacteria in *Il10*$^{-/-}$ mice. Box and whisker plots show the minimum, first quartile, median, third quartile and maximum relative abundance (showing the median of each cage). FDR corrected *P* values from the mixed linear model. *Il10*$^{-/-}$ week 2 *n*=17, week 12 *n*=16, week 20 *n*=15; WT week 2 *n*=24, week 12 *n*=22, week 20 *n*=24.

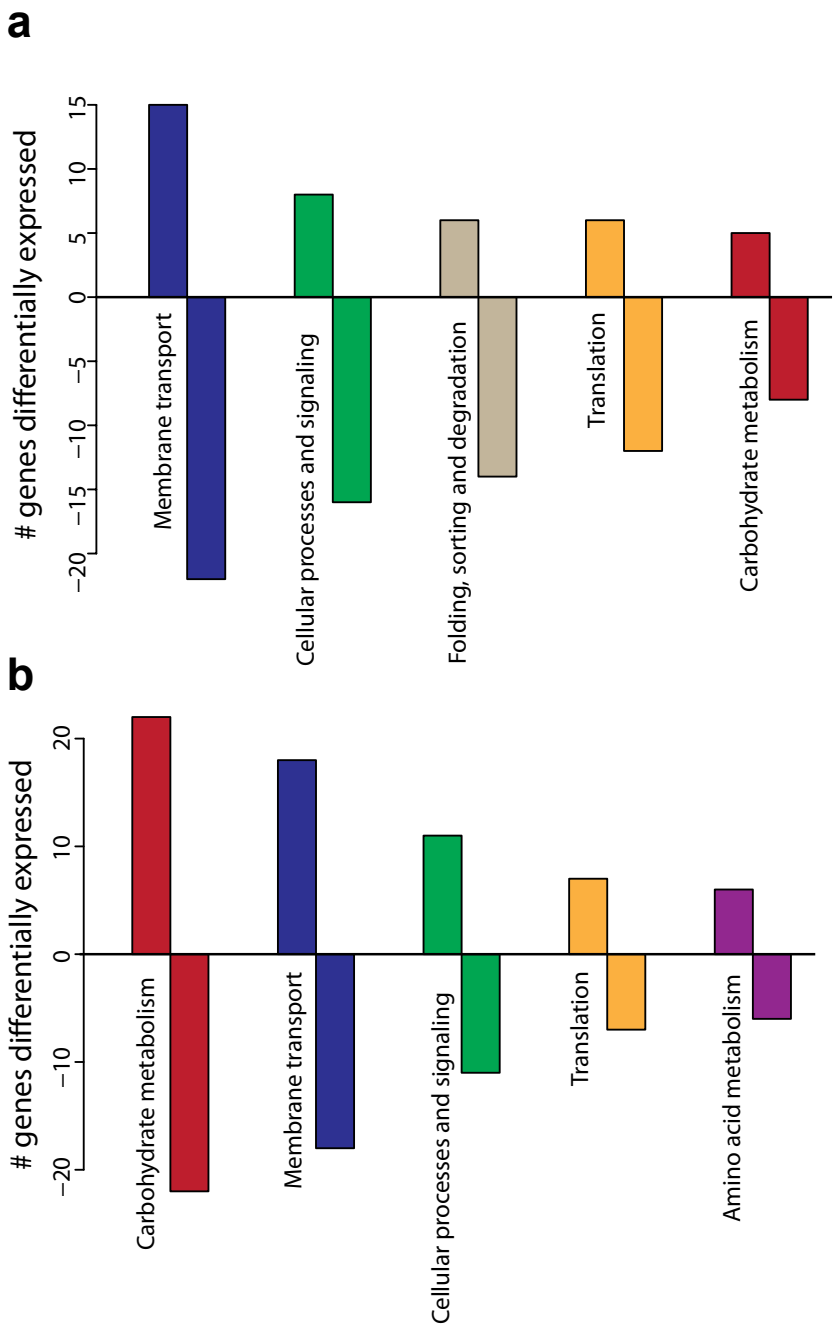



**Supplementary Figure 3 – Phylum-level distribution of WT vs. *Il10*$^{-/-}$ at each time-point**
**a**) 2 weeks, **b**) 12 weeks and **c**) 20 weeks after conventionalization. Each bar represents the mean + SEM of mice within each cage (i.e. cage corrected relative abundance). Each pie piece represents the mean percentage value from each group, WT or *Il10*$^{-/-}$. See Supplementary Data 1 for statistics on all phylum-level comparisons.



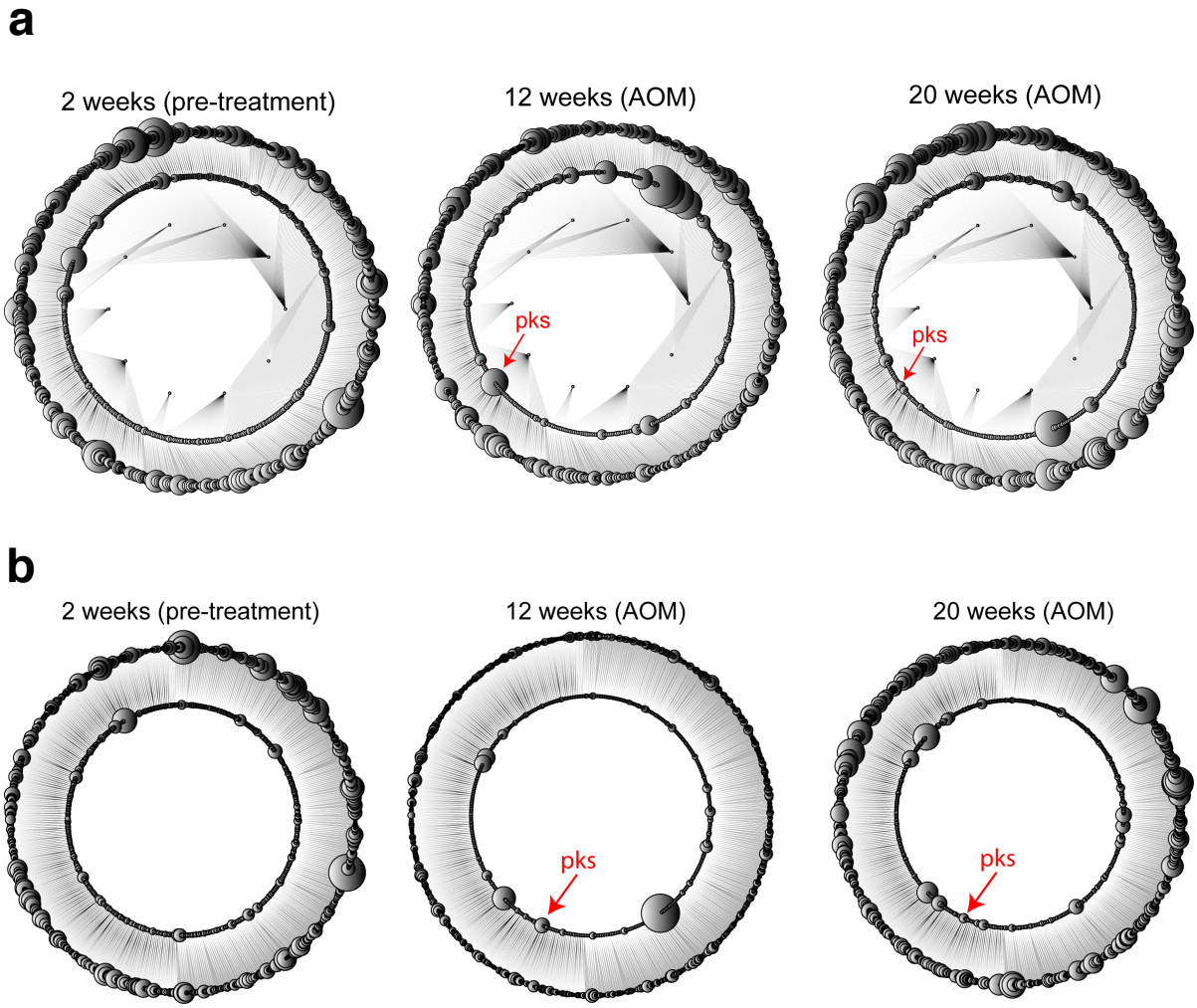

**Supplementary Figure 4 – Colonic tissue cytokine expression in *E. coli* NC101 mono-associated mice.**

These data were generated from the same experiment as Figure 4 in the main manuscript. RNA was extracted from distally colon biopsies using MMLV reverse transcriptase (Invitrogen) and qPCR amplification was performed in triplicate using SYBR green (Applied Biosystems) on an ABI 7900HT Real-Time PCR system. $C_T$ values were normalized to Gapdh to generate $\Delta C_T$ values, and fold changes were calculated by $\Delta\Delta C_T$ to the mean $\Delta C_T$ of the AOM/*Il10*$^{-/-}$;*Rag2*$^{-/-}$ group.

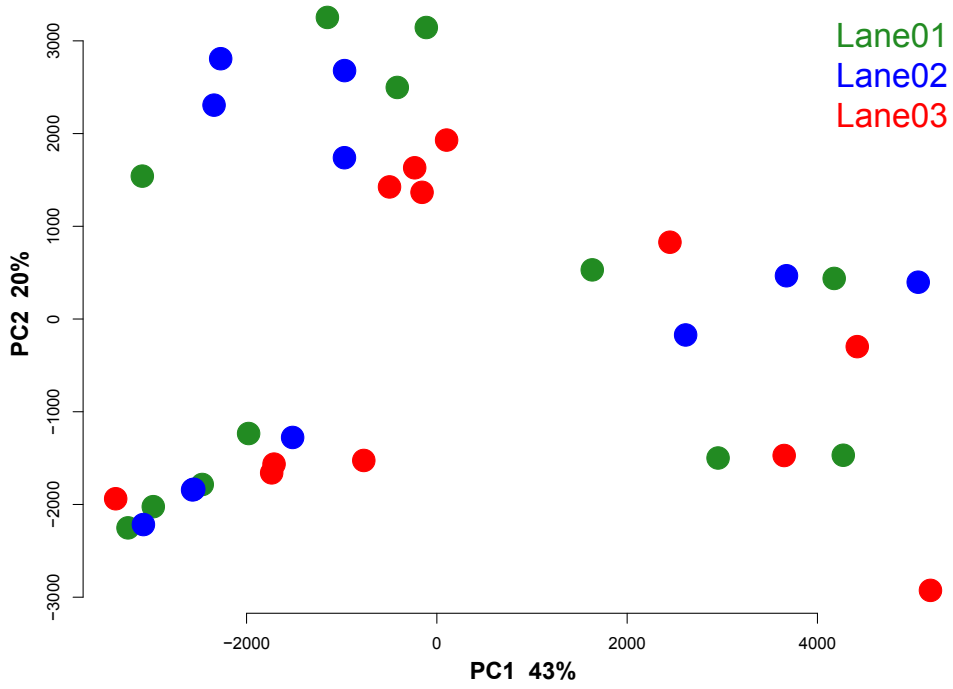



**Supplementary Figure 5 – Most represented KEGG pathways among all genotype/ disease groups**

Number of differentially expressed genes (FDR corrected *P*-value < 0.1) in the top 5 most represented KEGG pathways among all genotype/disease groups. Positive values on y-axis represent genes upregulated and negative values represent genes down-regulated relative to 2 week time-point at: **a**) 12 weeks and **b**) 20 weeks post-colonization. See Supplementary Data 3 for a list of genes changed among all groups at each time-point.

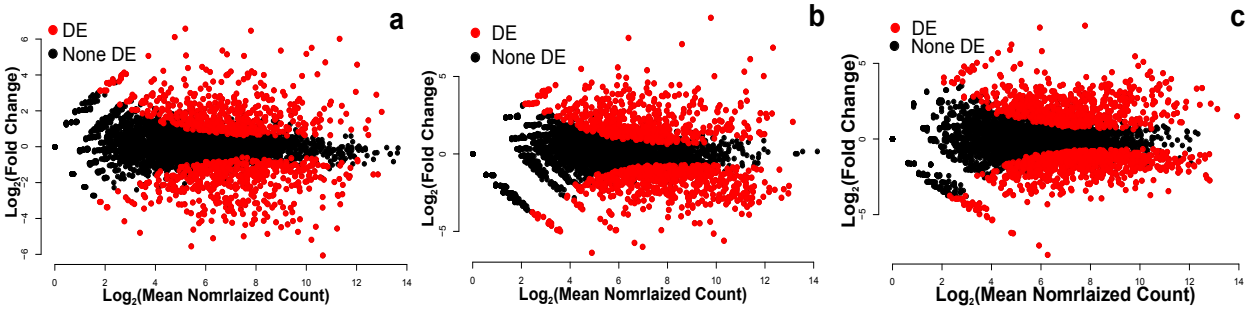



**Supplementary Figure 6 – Operon-based differential expression analysis reveals that the *pks* island is upregulated in the cancer microenvironment at 12 weeks**
For each predicted operon for each of the 10 contigs (shown in inner ring) in the NC101 genome (**a**) and each predicted operon of *E. coli* 536 (**b**), we evaluated a null hypothesis of no differential expression between *Il10*$^{-/-}$ and AOM/*Il10*$^{-/-}$ at each time point. *P*-values for this null hypothesis for each operon (second ring) were generated via the GAGE algorithm (see methods and supplemental data 6). At 2 weeks (left panel), animals had not yet been exposed to AOM and the calculated *P*-value for *pks* was 0.9999 in both **a** and **b**. At the 12 week time-point, the *pks* island had a *P*-value of $3.4 \times 10^{-5}$ (**a**) and $1.82 \times 10^{-5}$ (**b**), the 5$^{th}$ most differentially expressed operon in the *E. coli* NC101 and *E. coli* 536 genomes. At the 20 week time-point, the *pks* island had a *P*-value of 0.09 (**a**) and 0.035 (**b**) and was the 19$^{th}$ (**a**) and the 12$^{th}$ (**b**) most differentially expressed operon. Sizes of all circles reflect uncorrected *P*-values on a log scale. In determining *P*-values, only the 3 animals for which we had RNA-seq data for all 3 time-points were used.




**Supplementary Figure 7 – Principal Component Analysis plot showing no batch effect in the second RNA-seq experiment**

Principal Component Analysis plot constructed from the normalized *E. coli* gene counts from all samples and time points reveals no batch effect (PC1 one-way ANOVA *P*-value = 0.63, PC2 one-way ANOVA *P*-value = 0.72). Each symbol indicates an individual mouse at each time-point (green=Lane01 samples, blue=Lane02 samples, red=Lane03 samples)




**Supplementary Figure 8 – MA plots ($log_2$ fold change versus $log_2$ mean normalized counts for each transcripts)**

**a)** AOM/*Il10*$^{-/-}$ at 2 weeks vs. 12 weeks, **b)** *Il10*$^{-/-}$ at 2 weeks vs. 20 weeks and c**)** AOM/ *Il10*$^{-/-}$;Rag2$^{-/-}$ at 12 weeks vs. 20 weeks show that there is no bias in our differentially expression calls toward the high abundance transcripts. Differentially expressed transcripts (at FDR corrected $P < 0.1$) are shown in red.