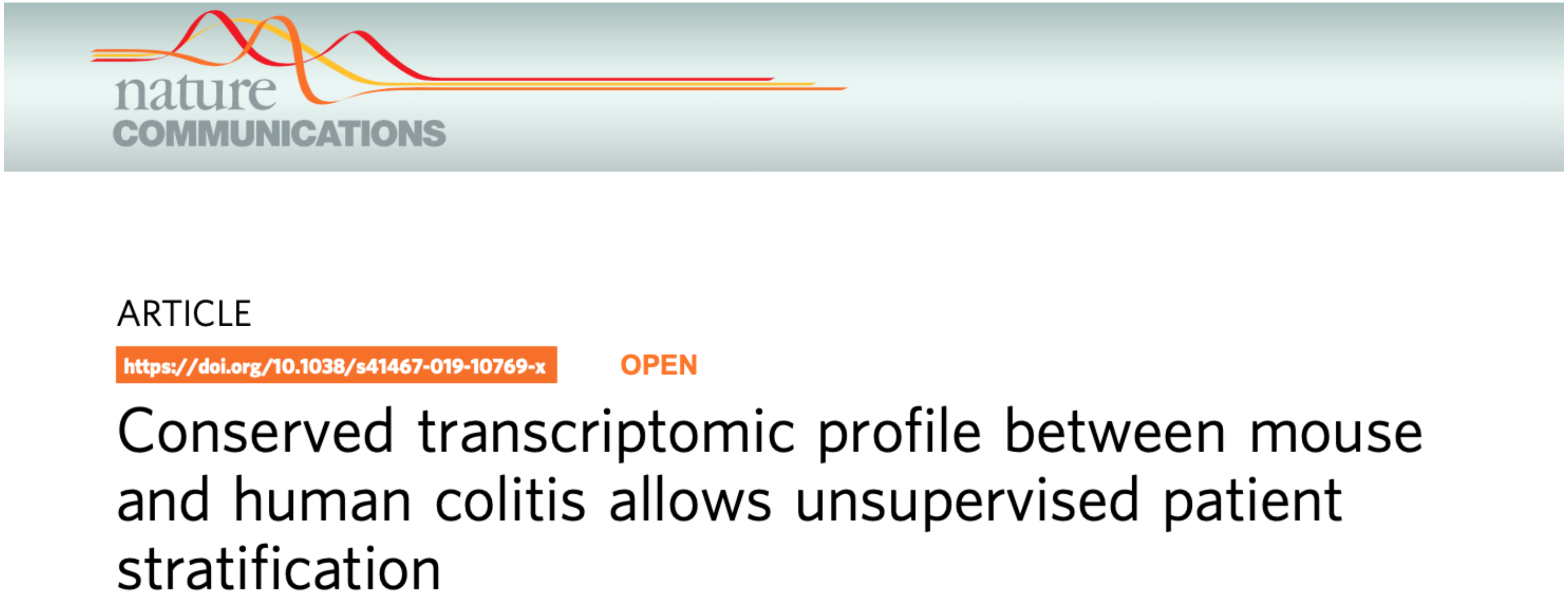
Reading an article on RNA-seq data analysis

RNA-seq data analysis

Paulo Czarnewski
https://czarnewski.github.io/czarnewski/index.html

How does it look in real life ?



nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-019-10769-x OPEN

Conserved transcriptomic profile between mouse and human colitis allows unsupervised patient stratification

Czarnewski et al (2019) *Nat Communications*

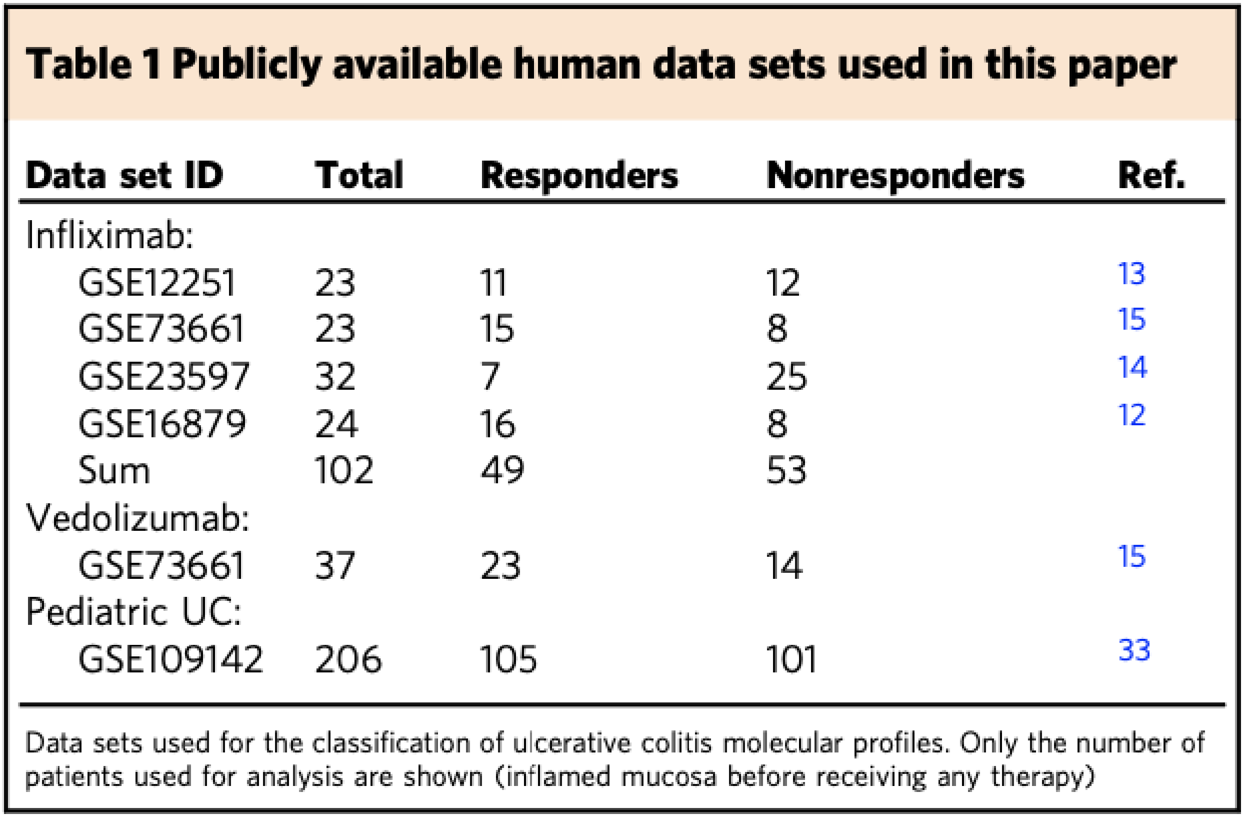
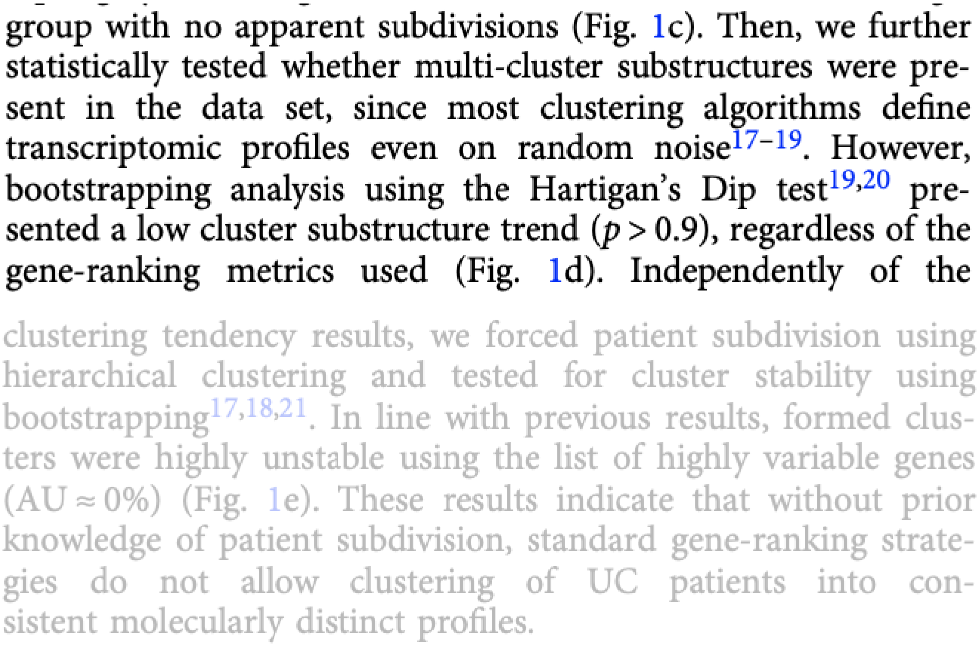


Table 1 Publicly available human data sets used in this paper

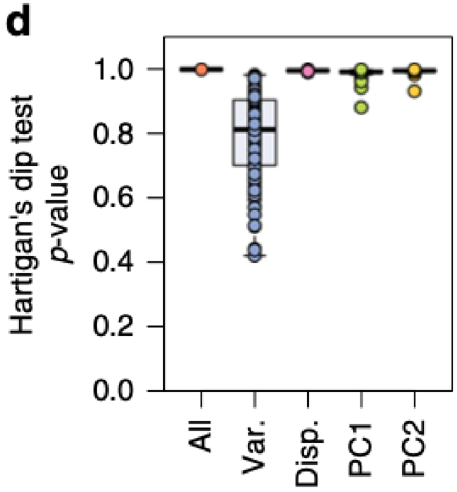
Data set ID	Total	Responders	Nonresponders	Ref.
Infliximab:				
GSE12251	23	11	12	13
GSE73661	23	15	8	15
GSE23597	32	7	25	14
GSE16879	24	16	8	12
Sum	102	49	53	
Vedolizumab:				
GSE73661	37	23	14	15
Pediatric UC:				
GSE109142	206	105	101	33

Data sets used for the classification of ulcerative colitis molecular profiles. Only the number of patients used for analysis are shown (inflamed mucosa before receiving any therapy)

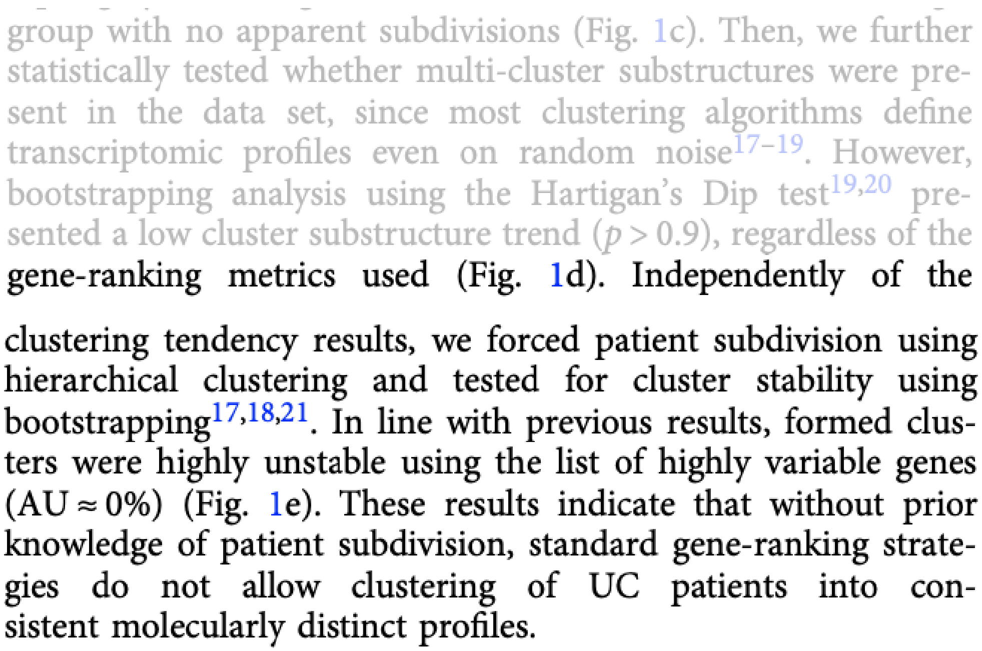
Czarnewski et al (2019) *Nat Communications*



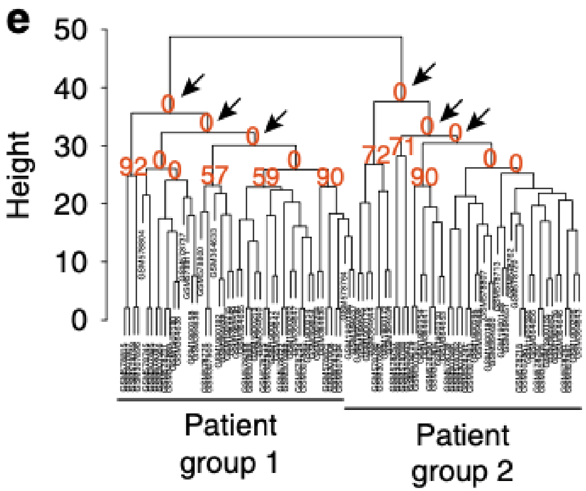
group with no apparent subdivisions (Fig. 1c). Then, we further statistically tested whether multi-cluster substructures were present in the data set, since most clustering algorithms define transcriptomic profiles even on random noise[17–19]. However, bootstrapping analysis using the Hartigan's Dip test[19,20] presented a low cluster substructure trend ($p > 0.9$), regardless of the gene-ranking metrics used (Fig. 1d). Independently of the clustering tendency results, we forced patient subdivision using hierarchical clustering and tested for cluster stability using bootstrapping[17,18,21]. In line with previous results, formed clusters were highly unstable using the list of highly variable genes (AU ≈ 0%) (Fig. 1e). These results indicate that without prior knowledge of patient subdivision, standard gene-ranking strategies do not allow clustering of UC patients into consistent molecularly distinct profiles.



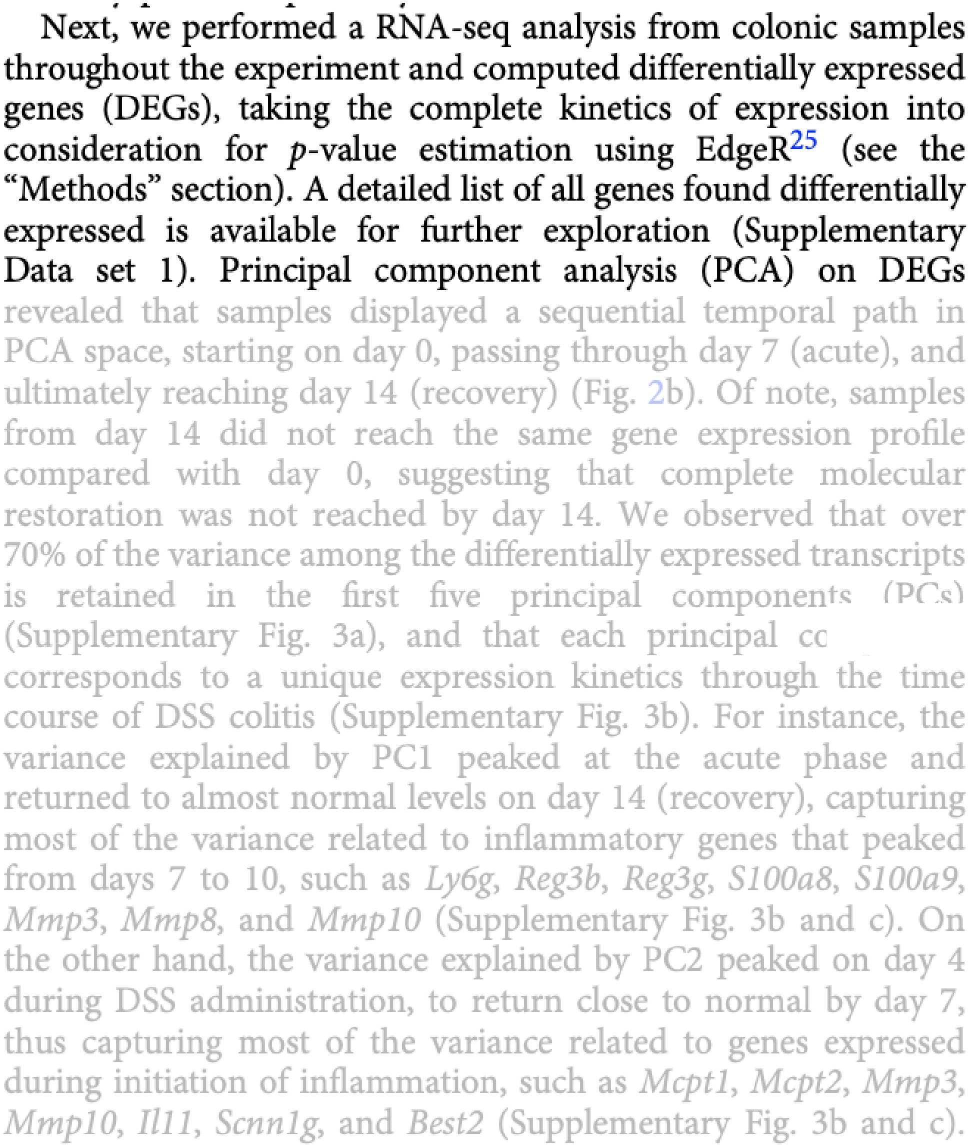
Czarnewski et al (2019) *Nat Communications*



group with no apparent subdivisions (Fig. 1c). Then, we further statistically tested whether multi-cluster substructures were present in the data set, since most clustering algorithms define transcriptomic profiles even on random noise[17–19]. However, bootstrapping analysis using the Hartigan's Dip test[19,20] presented a low cluster substructure trend ($p > 0.9$), regardless of the gene-ranking metrics used (Fig. 1d). Independently of the clustering tendency results, we forced patient subdivision using hierarchical clustering and tested for cluster stability using bootstrapping[17,18,21]. In line with previous results, formed clusters were highly unstable using the list of highly variable genes (AU $\approx$ 0%) (Fig. 1e). These results indicate that without prior knowledge of patient subdivision, standard gene-ranking strategies do not allow clustering of UC patients into consistent molecularly distinct profiles.

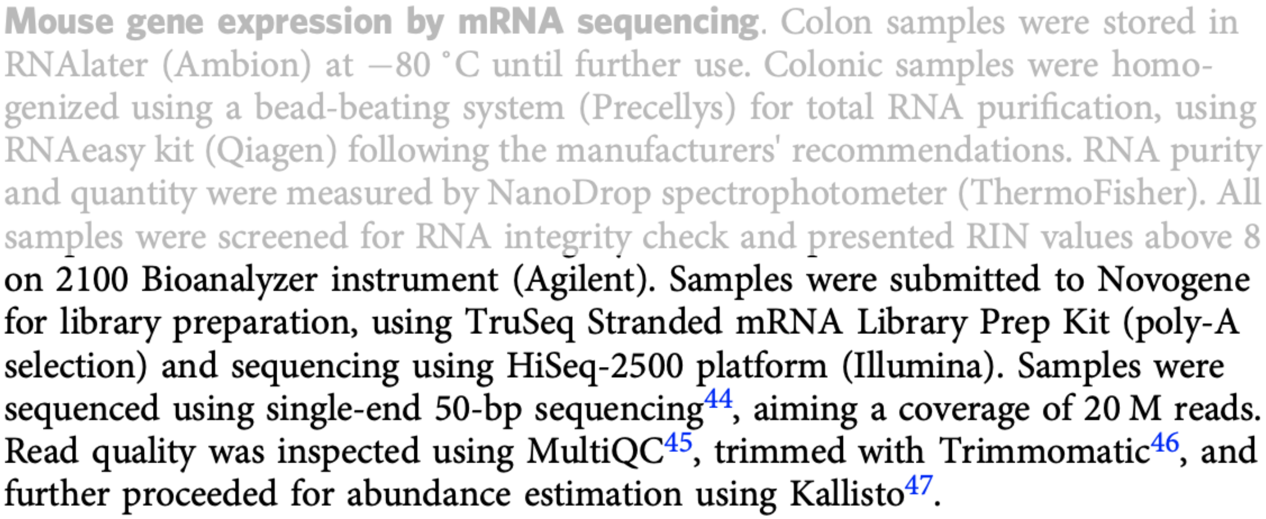


Czarnewski et al (2019) *Nat Communications*



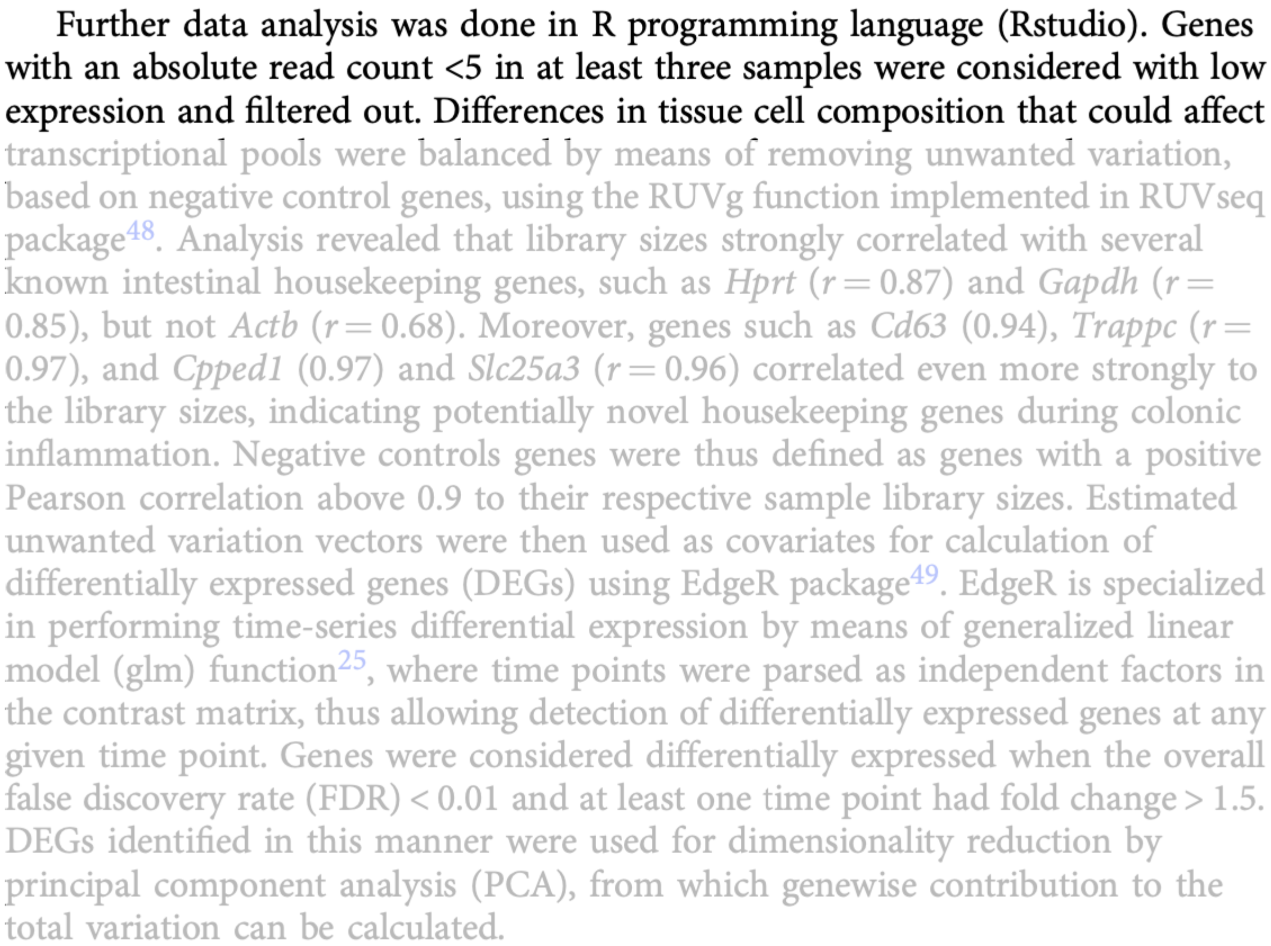
Next, we performed a RNA-seq analysis from colonic samples throughout the experiment and computed differentially expressed genes (DEGs), taking the complete kinetics of expression into consideration for *p*-value estimation using EdgeR[25] (see the "Methods" section). A detailed list of all genes found differentially expressed is available for further exploration (Supplementary Data set 1). Principal component analysis (PCA) on DEGs revealed that samples displayed a sequential temporal path in PCA space, starting on day 0, passing through day 7 (acute), and ultimately reaching day 14 (recovery) (Fig. 2b). Of note, samples from day 14 did not reach the same gene expression profile compared with day 0, suggesting that complete molecular restoration was not reached by day 14. We observed that over 70% of the variance among the differentially expressed transcripts is retained in the first five principal components (PCs) (Supplementary Fig. 3a), and that each principal cc corresponds to a unique expression kinetics through the time course of DSS colitis (Supplementary Fig. 3b). For instance, the variance explained by PC1 peaked at the acute phase and returned to almost normal levels on day 14 (recovery), capturing most of the variance related to inflammatory genes that peaked from days 7 to 10, such as *Ly6g*, *Reg3b*, *Reg3g*, *S100a8*, *S100a9*, *Mmp3*, *Mmp8*, and *Mmp10* (Supplementary Fig. 3b and c). On the other hand, the variance explained by PC2 peaked on day 4 during DSS administration, to return close to normal by day 7, thus capturing most of the variance related to genes expressed during initiation of inflammation, such as *Mcpt1*, *Mcpt2*, *Mmp3*, *Mmp10*, *Il11*, *Scnn1g*, and *Best2* (Supplementary Fig. 3b and c).

Czarnewski et al (2019) *Nat Communications*



Mouse gene expression by mRNA sequencing. Colon samples were stored in RNAlater (Ambion) at −80 °C until further use. Colonic samples were homogenized using a bead-beating system (Precellys) for total RNA purification, using RNAeasy kit (Qiagen) following the manufacturers' recommendations. RNA purity and quantity were measured by NanoDrop spectrophotometer (ThermoFisher). All samples were screened for RNA integrity check and presented RIN values above 8 on 2100 Bioanalyzer instrument (Agilent). Samples were submitted to Novogene for library preparation, using TruSeq Stranded mRNA Library Prep Kit (poly-A selection) and sequencing using HiSeq-2500 platform (Illumina). Samples were sequenced using single-end 50-bp sequencing[44], aiming a coverage of 20 M reads. Read quality was inspected using MultiQC[45], trimmed with Trimmomatic[46], and further proceeded for abundance estimation using Kallisto[47].

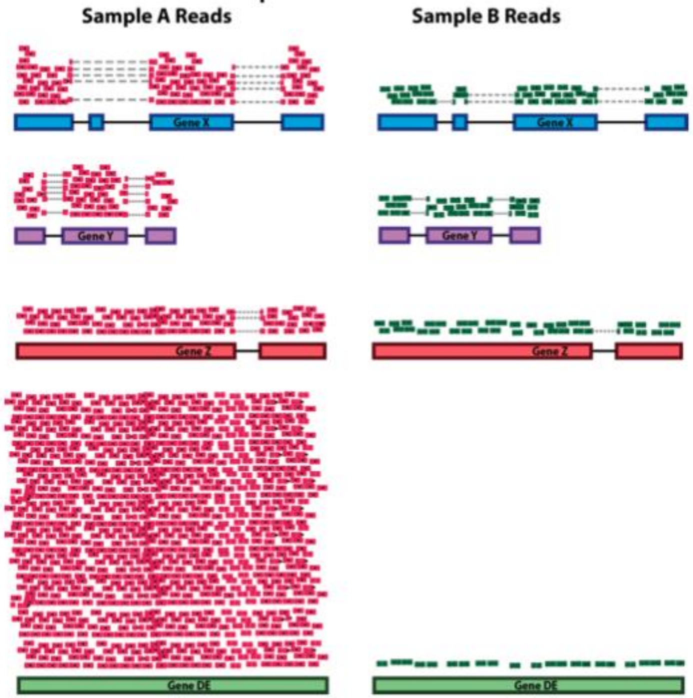
Czarnewski et al (2019) *Nat Communications*



Further data analysis was done in R programming language (Rstudio). Genes with an absolute read count <5 in at least three samples were considered with low expression and filtered out. Differences in tissue cell composition that could affect transcriptional pools were balanced by means of removing unwanted variation, based on negative control genes, using the RUVg function implemented in RUVseq package[48]. Analysis revealed that library sizes strongly correlated with several known intestinal housekeeping genes, such as *Hprt* ($r = 0.87$) and *Gapdh* ($r = 0.85$), but not *Actb* ($r = 0.68$). Moreover, genes such as *Cd63* (0.94), *Trappc* ($r = 0.97$), and *Cpped1* (0.97) and *Slc25a3* ($r = 0.96$) correlated even more strongly to the library sizes, indicating potentially novel housekeeping genes during colonic inflammation. Negative controls genes were thus defined as genes with a positive Pearson correlation above 0.9 to their respective sample library sizes. Estimated unwanted variation vectors were then used as covariates for calculation of differentially expressed genes (DEGs) using EdgeR package[49]. EdgeR is specialized in performing time-series differential expression by means of generalized linear model (glm) function[25], where time points were parsed as independent factors in the contrast matrix, thus allowing detection of differentially expressed genes at any given time point. Genes were considered differentially expressed when the overall false discovery rate (FDR) < 0.01 and at least one time point had fold change > 1.5. DEGs identified in this manner were used for dimensionality reduction by principal component analysis (PCA), from which genewise contribution to the total variation can be calculated.

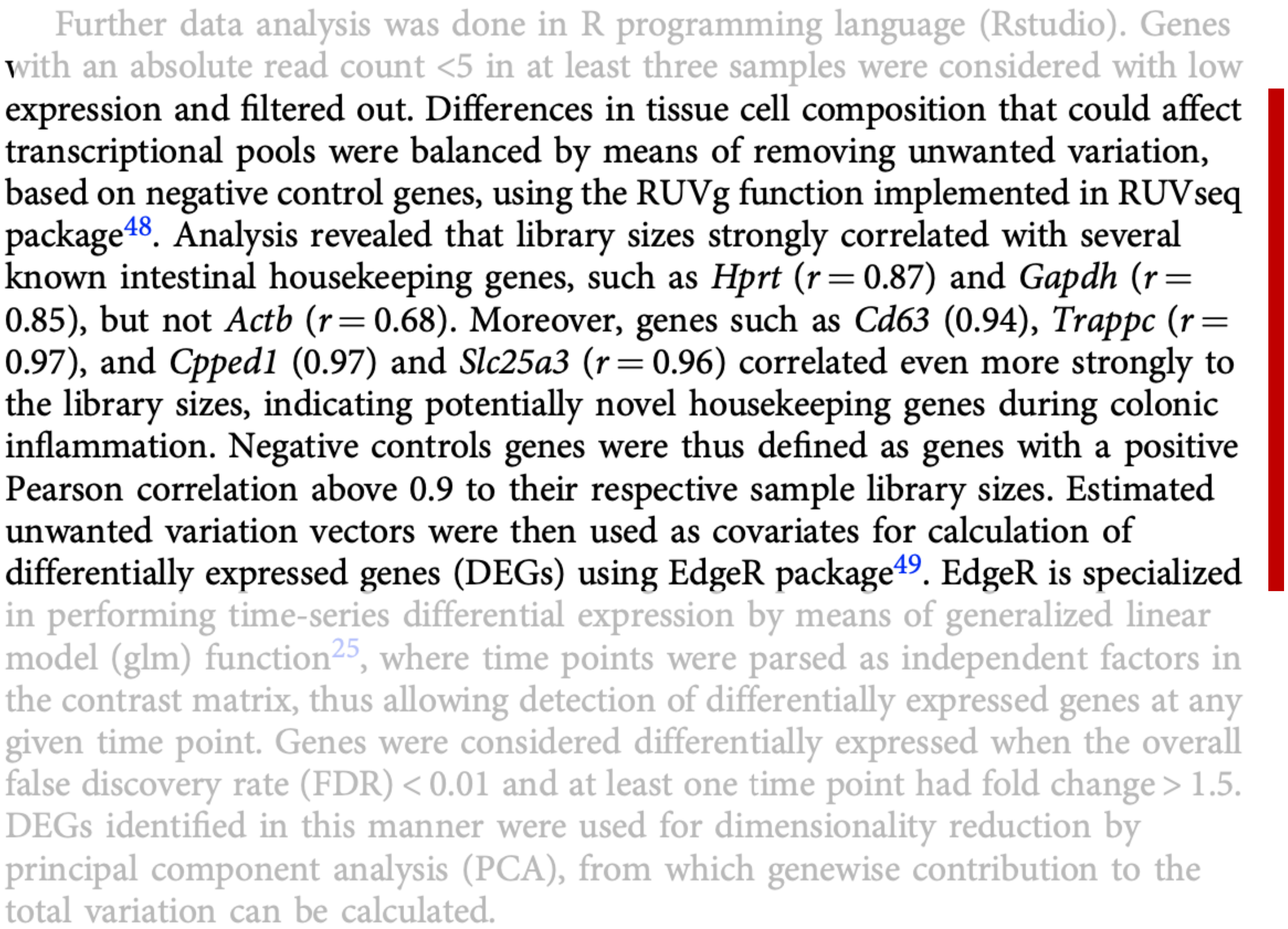
Czarnewski et al (2019) *Nat Communications*

- Control for compositional bias



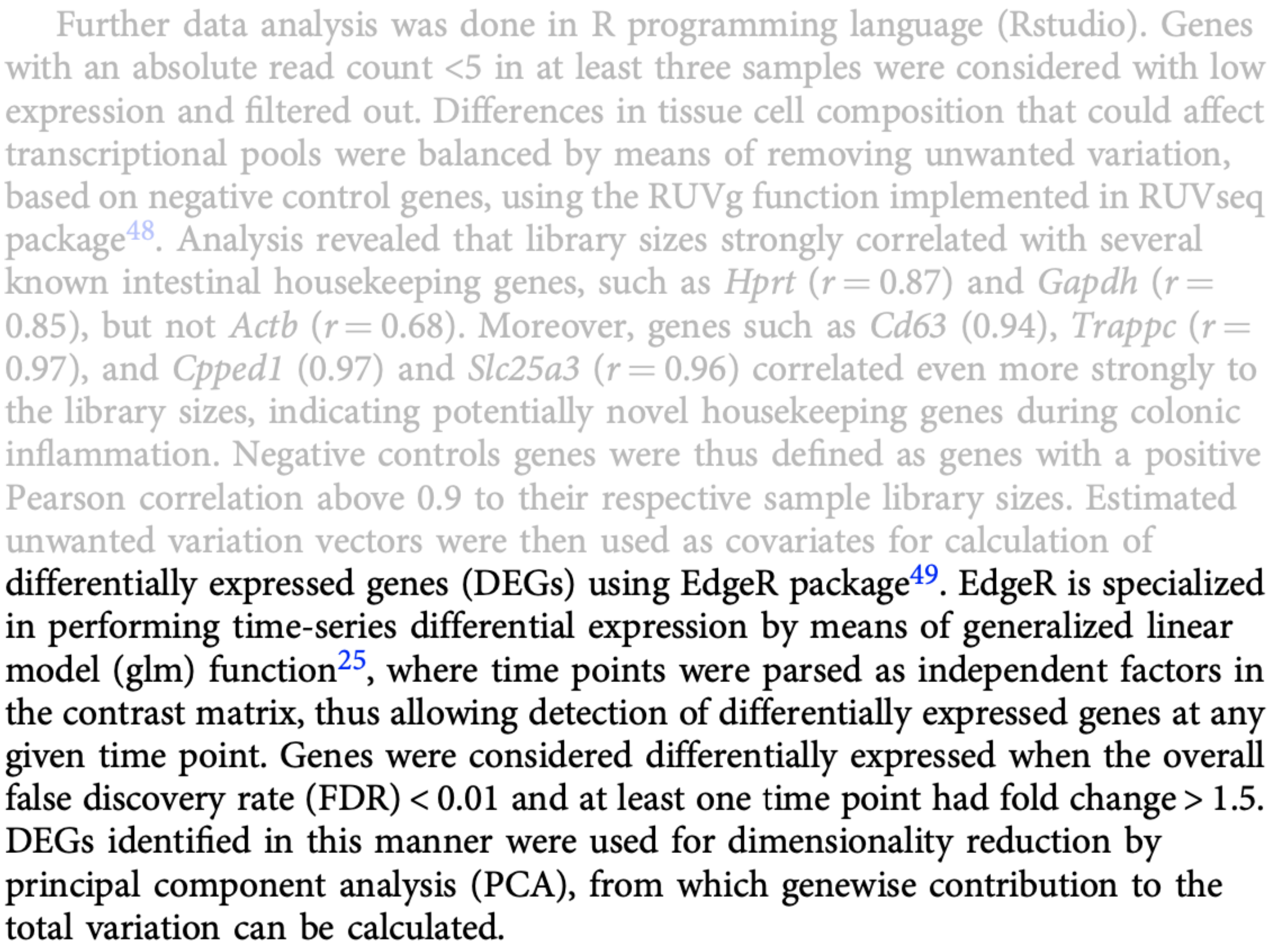
Evans, Ciaran, Johanna Hardin, and Daniel M. Stoebel. "Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions." Briefings in bioinformatics (2017)

Czarnewski et al (2019) *Nat Communications*



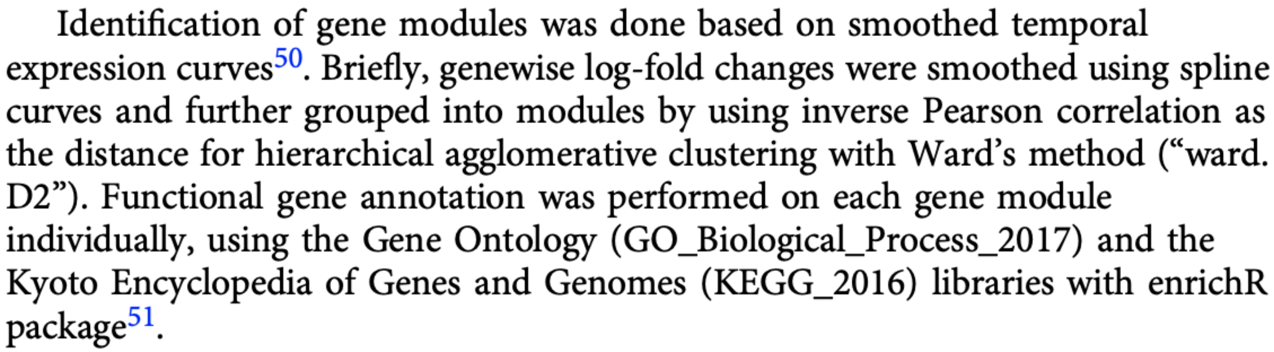
Further data analysis was done in R programming language (Rstudio). Genes with an absolute read count <5 in at least three samples were considered with low expression and filtered out. Differences in tissue cell composition that could affect transcriptional pools were balanced by means of removing unwanted variation, based on negative control genes, using the RUVg function implemented in RUVseq package[48]. Analysis revealed that library sizes strongly correlated with several known intestinal housekeeping genes, such as *Hprt* ($r = 0.87$) and *Gapdh* ($r = 0.85$), but not *Actb* ($r = 0.68$). Moreover, genes such as *Cd63* (0.94), *Trappc* ($r = 0.97$), and *Cpped1* (0.97) and *Slc25a3* ($r = 0.96$) correlated even more strongly to the library sizes, indicating potentially novel housekeeping genes during colonic inflammation. Negative controls genes were thus defined as genes with a positive Pearson correlation above 0.9 to their respective sample library sizes. Estimated unwanted variation vectors were then used as covariates for calculation of differentially expressed genes (DEGs) using EdgeR package[49]. EdgeR is specialized in performing time-series differential expression by means of generalized linear model (glm) function[25], where time points were parsed as independent factors in the contrast matrix, thus allowing detection of differentially expressed genes at any given time point. Genes were considered differentially expressed when the overall false discovery rate (FDR) < 0.01 and at least one time point had fold change > 1.5. DEGs identified in this manner were used for dimensionality reduction by principal component analysis (PCA), from which genewise contribution to the total variation can be calculated.

Czarnewski et al (2019) *Nat Communications*



Further data analysis was done in R programming language (Rstudio). Genes with an absolute read count <5 in at least three samples were considered with low expression and filtered out. Differences in tissue cell composition that could affect transcriptional pools were balanced by means of removing unwanted variation, based on negative control genes, using the RUVg function implemented in RUVseq package[48]. Analysis revealed that library sizes strongly correlated with several known intestinal housekeeping genes, such as *Hprt* ($r = 0.87$) and *Gapdh* ($r = 0.85$), but not *Actb* ($r = 0.68$). Moreover, genes such as *Cd63* (0.94), *Trappc* ($r = 0.97$), and *Cpped1* (0.97) and *Slc25a3* ($r = 0.96$) correlated even more strongly to the library sizes, indicating potentially novel housekeeping genes during colonic inflammation. Negative controls genes were thus defined as genes with a positive Pearson correlation above 0.9 to their respective sample library sizes. Estimated unwanted variation vectors were then used as covariates for calculation of differentially expressed genes (DEGs) using EdgeR package[49]. EdgeR is specialized in performing time-series differential expression by means of generalized linear model (glm) function[25], where time points were parsed as independent factors in the contrast matrix, thus allowing detection of differentially expressed genes at any given time point. Genes were considered differentially expressed when the overall false discovery rate (FDR) < 0.01 and at least one time point had fold change > 1.5. DEGs identified in this manner were used for dimensionality reduction by principal component analysis (PCA), from which genewise contribution to the total variation can be calculated.

Czarnewski et al (2019) *Nat Communications*

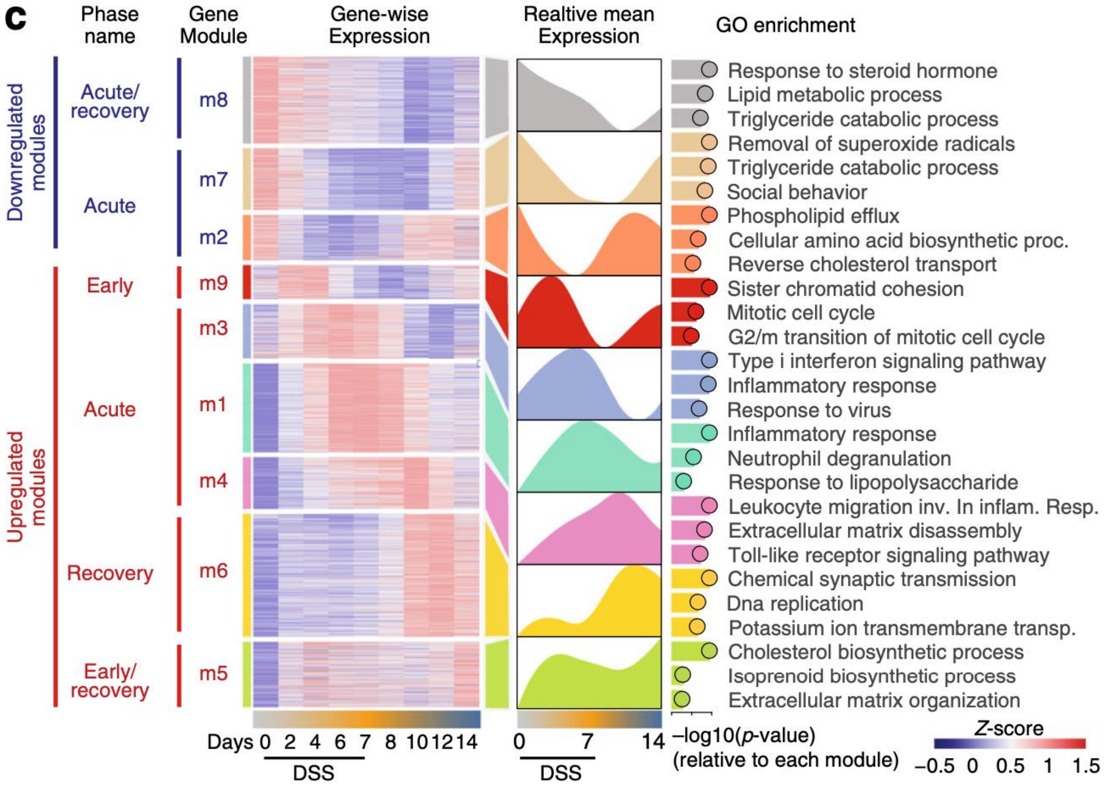


Identification of gene modules was done based on smoothed temporal expression curves[50]. Briefly, genewise log-fold changes were smoothed using spline curves and further grouped into modules by using inverse Pearson correlation as the distance for hierarchical agglomerative clustering with Ward's method ("ward. D2"). Functional gene annotation was performed on each gene module individually, using the Gene Ontology (GO_Biological_Process_2017) and the Kyoto Encyclopedia of Genes and Genomes (KEGG_2016) libraries with enrichR package[51].

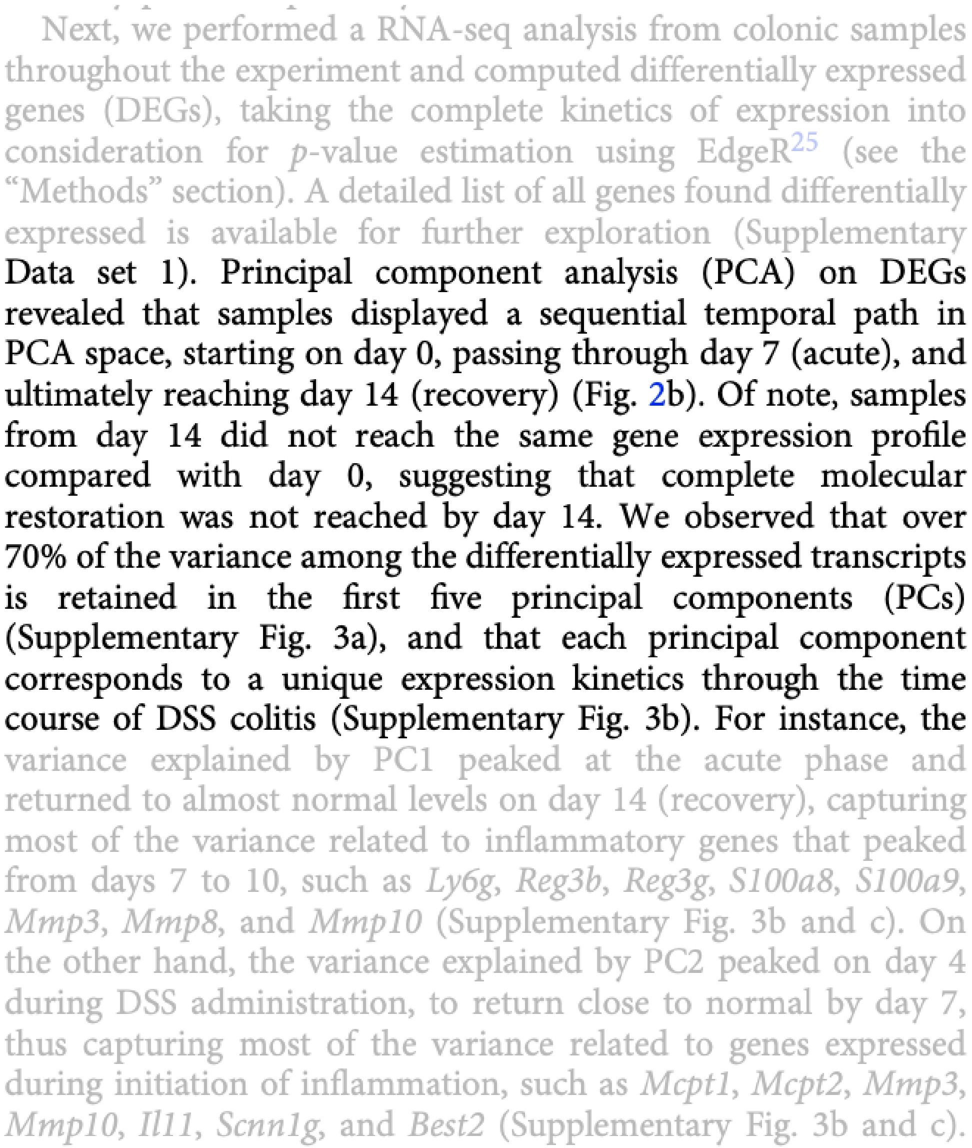
Czarnewski et al (2019) *Nat Communications*

Time-series analysis during colon inflammation and repair
(Fig2)

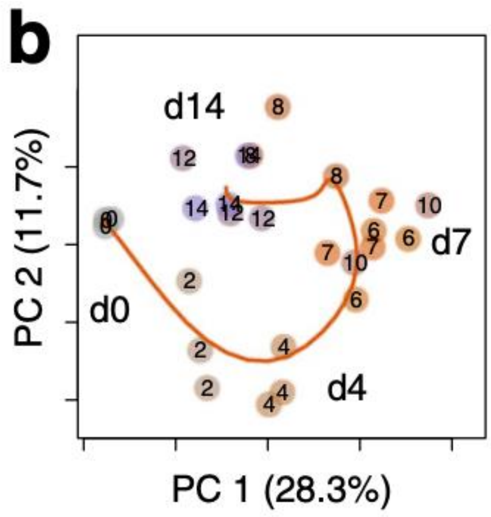
```
metadata$Group <- factor( metadata$Group , levels=c(“day00” , ...) )


y ~ Group
```

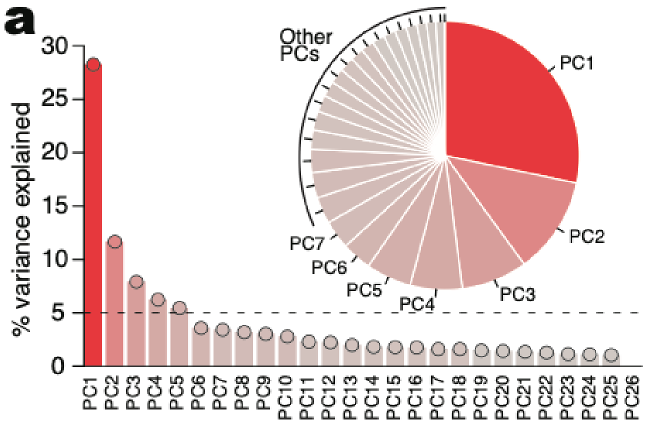


Czarnewski et al (2019) *Nat Communications*

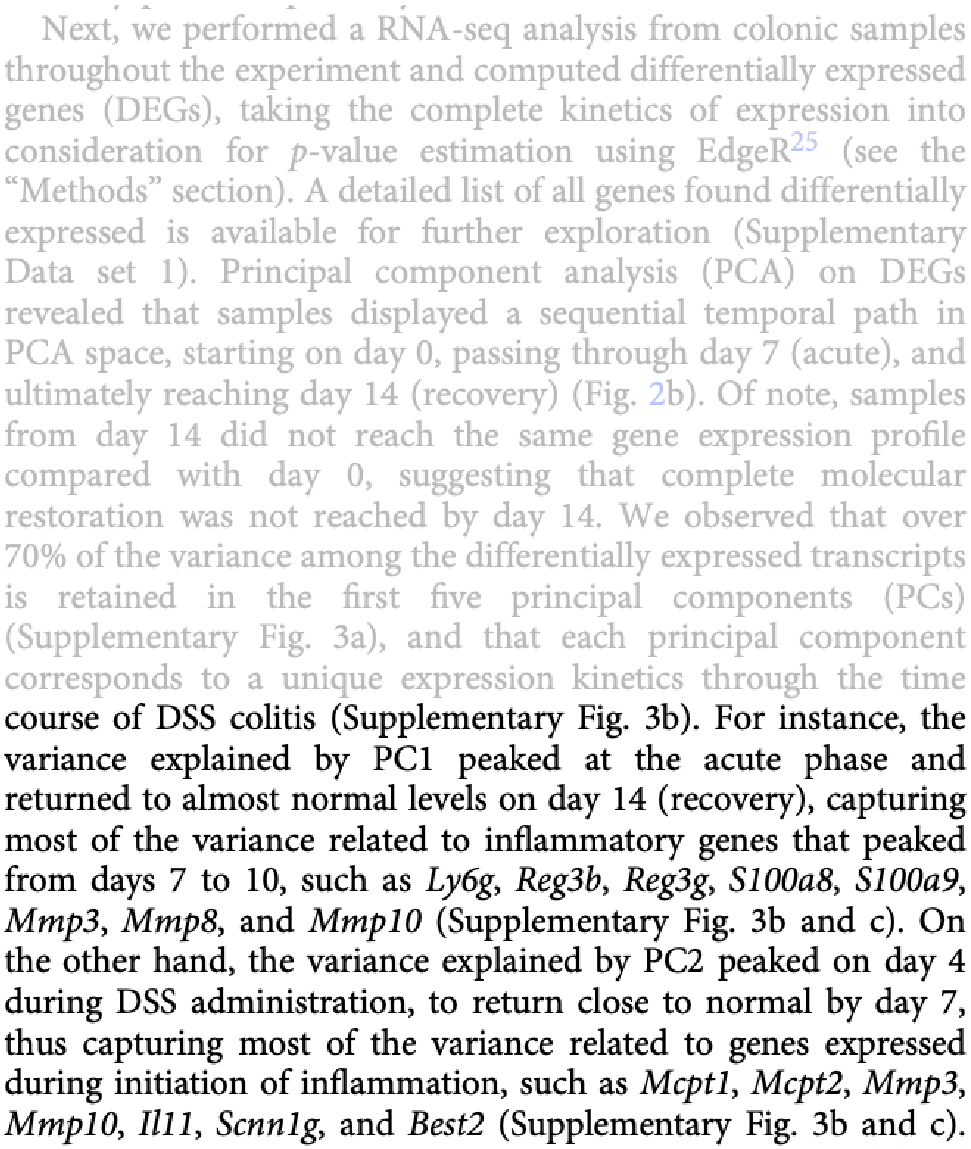


Next, we performed a RNA-seq analysis from colonic samples throughout the experiment and computed differentially expressed genes (DEGs), taking the complete kinetics of expression into consideration for *p*-value estimation using EdgeR[25] (see the "Methods" section). A detailed list of all genes found differentially expressed is available for further exploration (Supplementary Data set 1). Principal component analysis (PCA) on DEGs revealed that samples displayed a sequential temporal path in PCA space, starting on day 0, passing through day 7 (acute), and ultimately reaching day 14 (recovery) (Fig. 2b). Of note, samples from day 14 did not reach the same gene expression profile compared with day 0, suggesting that complete molecular restoration was not reached by day 14. We observed that over 70% of the variance among the differentially expressed transcripts is retained in the first five principal components (PCs) (Supplementary Fig. 3a), and that each principal component corresponds to a unique expression kinetics through the time course of DSS colitis (Supplementary Fig. 3b). For instance, the variance explained by PC1 peaked at the acute phase and returned to almost normal levels on day 14 (recovery), capturing most of the variance related to inflammatory genes that peaked from days 7 to 10, such as *Ly6g*, *Reg3b*, *Reg3g*, *S100a8*, *S100a9*, *Mmp3*, *Mmp8*, and *Mmp10* (Supplementary Fig. 3b and c). On the other hand, the variance explained by PC2 peaked on day 4 during DSS administration, to return close to normal by day 7, thus capturing most of the variance related to genes expressed during initiation of inflammation, such as *Mcpt1*, *Mcpt2*, *Mmp3*, *Mmp10*, *Il11*, *Scnn1g*, and *Best2* (Supplementary Fig. 3b and c).

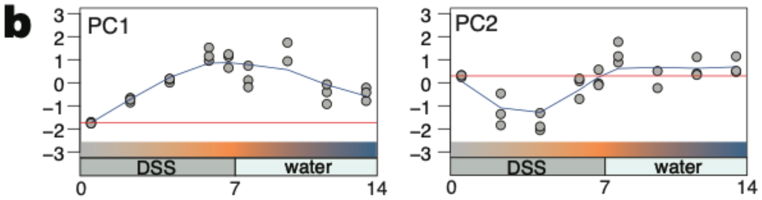


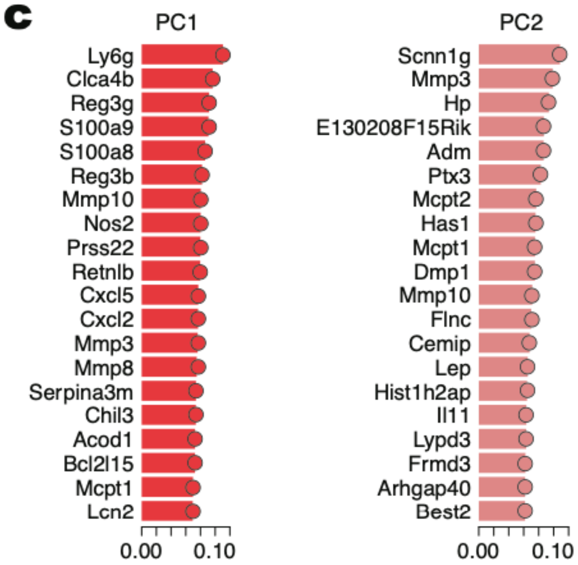


Czarnewski et al (2019) *Nat Communications*

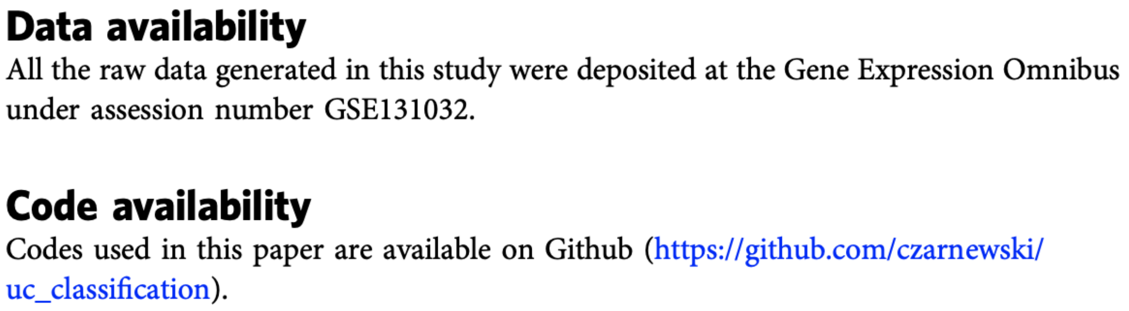


Next, we performed a RNA-seq analysis from colonic samples throughout the experiment and computed differentially expressed genes (DEGs), taking the complete kinetics of expression into consideration for *p*-value estimation using EdgeR[25] (see the "Methods" section). A detailed list of all genes found differentially expressed is available for further exploration (Supplementary Data set 1). Principal component analysis (PCA) on DEGs revealed that samples displayed a sequential temporal path in PCA space, starting on day 0, passing through day 7 (acute), and ultimately reaching day 14 (recovery) (Fig. 2b). Of note, samples from day 14 did not reach the same gene expression profile compared with day 0, suggesting that complete molecular restoration was not reached by day 14. We observed that over 70% of the variance among the differentially expressed transcripts is retained in the first five principal components (PCs) (Supplementary Fig. 3a), and that each principal component corresponds to a unique expression kinetics through the time course of DSS colitis (Supplementary Fig. 3b). For instance, the variance explained by PC1 peaked at the acute phase and returned to almost normal levels on day 14 (recovery), capturing most of the variance related to inflammatory genes that peaked from days 7 to 10, such as *Ly6g*, *Reg3b*, *Reg3g*, *S100a8*, *S100a9*, *Mmp3*, *Mmp8*, and *Mmp10* (Supplementary Fig. 3b and c). On the other hand, the variance explained by PC2 peaked on day 4 during DSS administration, to return close to normal by day 7, thus capturing most of the variance related to genes expressed during initiation of inflammation, such as *Mcpt1*, *Mcpt2*, *Mmp3*, *Mmp10*, *Il11*, *Scnn1g*, and *Best2* (Supplementary Fig. 3b and c).





Czarnewski et al (2019) *Nat Communications*



Data availability

All the raw data generated in this study were deposited at the Gene Expression Omnibus under assession number GSE131032.

Code availability

Codes used in this paper are available on Github (https://github.com/czarnewski/uc_classification).

Thank you. Questions?

Paulo Czarnewski