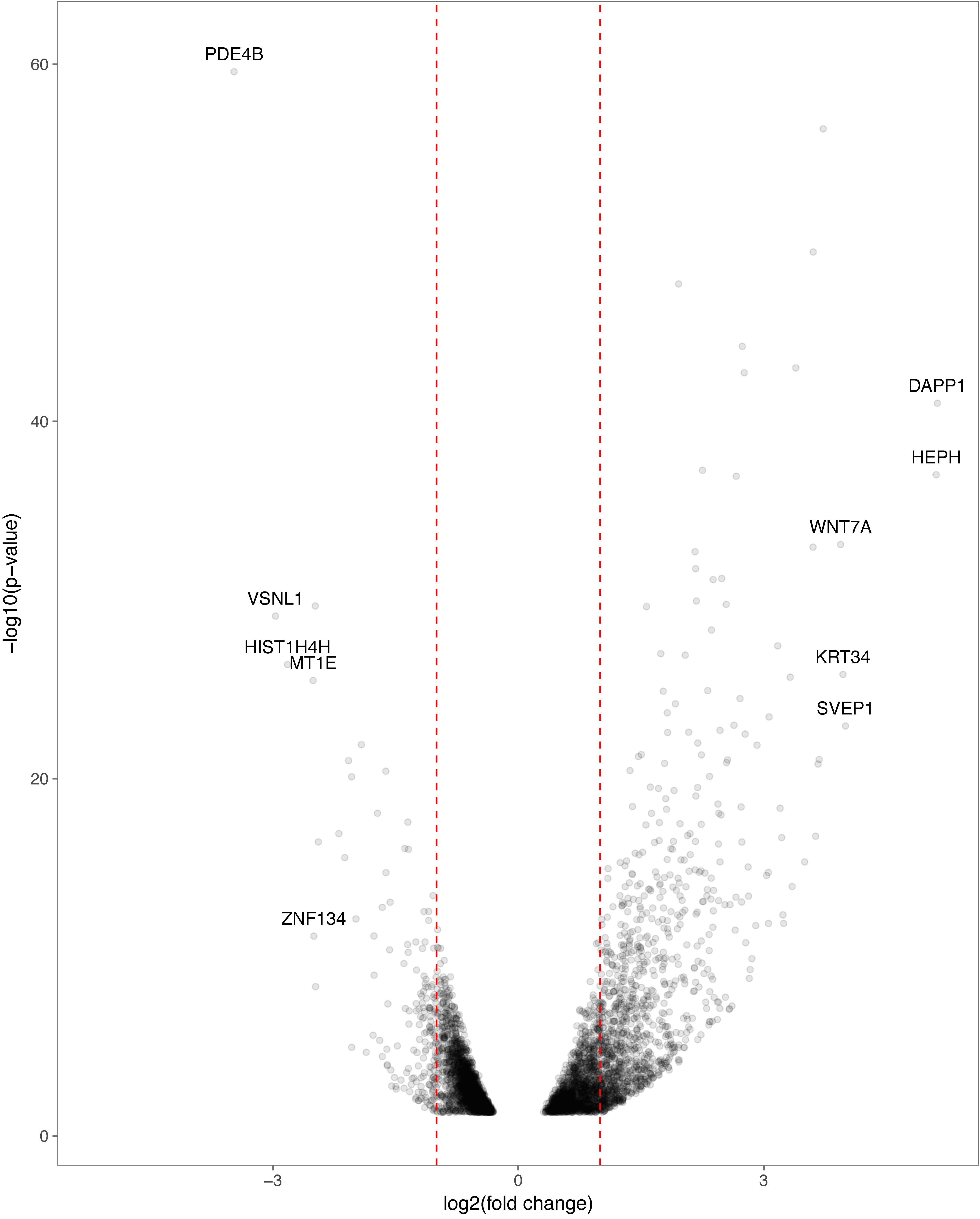


**Figure 1.** **A)** A network graph visualizing the results from the community analysis. Edge (line) direction is represented by color with edges originating from a node inheriting that nodes color. Edges between community nodes (large points) indicate that the GO terms representing the nodes are each other’s ancestors or offspring dependent on the direction of the edge. Edges between term nodes (small points) and community nodes indicate the terms inclusion in that community. **B)** A heatmap representing gene expression profiles in the detected communities. Communities are indicated by the color bar on the right side of the heatmap. Expression scaled to [0, 1] indicates the regularized log transformed expression values scaled between 0 and 1. **C-J)** Gene expression profiles for selected terms in a subset of the detected communities for parental and acid adapted cells. Legend abbreviations include: antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP−independent (APPEP); generation of precursor metabolites and energy (GPME); somatic recombination of immunoglobulin genes involved in immune response (SRIGIIR).



**Figure 2.** A volcano plot summarizing the results from the differential expression analysis. The five genes with the highest fold change in either direction are highlighted in the plot. Dotted red vertical lines represent a fold change of 2.

Results (in brief)

*Differential expression analysis*

Differential expression analysis discovered 4796 genes to be significantly (alpha < 0.05) regulated in acid adapted cells vs. their parental cell line with 1283 of these genes exhibiting a fold change > 2.

*Gene ontology term enrichment and community analysis*

Gene ontology term enrichment analysis resulted in 579 significantly (alpha < 0.05) enriched terms. Due to the fact that many of these terms are related to similar biological processes, we desired to summarize these results by identifying groups of similar terms. To achieve this we utilized a community detection algorithm3 to deduce similar terms within the GO graph of the significant terms and their ancestors. This resulted in the detection of 39 communities (figure 1A). In the communities detected we identified many terms associated with the known biology of acid adapted cells such as cell differentiation, cell death, and cell adhesion as well as identifying novel terms such as autophagy, cellular metabolism, and Wnt signaling (these can be changed accordingly) (figure 1A and 1C-J). Analysis of the gene expression patterns in each community revealed strong contrasts in gene expression between parental and acid adapted cells (figure Xb). Finally, analysis of individual expression patterns within each communities terms revealed both known and novel players in the biology of acid adapted cells (figure 1C-J).

Methods

*RNA sequencing and Bioinformatics*

RNA sequencing was performed in biological triplicate using Strand-specific TruSeq library preparation and Ribo-Zero ribosomal depletion. Tophat2, HTSeq, and DESeq2 were utilized for alignment, quantification, and differential expression analysis, respectively, with the hg19 genome and Ensembl v73. All bioinformatics were preformed in the R programming language7 and scripts related to the differential expression analysis, gene ontology analysis, and community detection analysis are publically available at <https://github.com/GranderLab/acidAdaptedRNAseq> as an R package facilitating reproduction of each analysis and the associated figure.

*Gene Ontology Term Enrichment*

Significant (alpha < 0.05) genes in the differential expression analysis were used for biological processes GO term enrichment analysis with the topGO software1. All quantified genes, defined as counts per million > 1 in at least 3 samples, were utilized as the gene universe. Terms with less than 5 annotated genes were not included in the significance testing procedure. Significance testing was performed using the classic Fisher method and the top 1000 terms with the lowest p-value were included in downstream analysis unless otherwise specified.

*Community analysis*

To understand the relationship between the significant GO terms, the GO graph was retrieved for all significant (alpha < 0.05) GO terms using the GOSim Bioconductor package2,8. The GO graph was then utilized for community detection via the spin-glass algorithm3 from the igraph package4 using a maximum of 200 possible spin states. The authority score was subsequently calculated for all nodes in the GO graph to determine the node with the highest in-degree for each community4,5. The GO graph was then collapsed on the nodes with the highest authority scores (community node) by merging all nodes into said node and simplifying the graphs edges. In cases when several nodes had identical authority scores equaling the max authority score for that community’s nodes, one was chosen at random to represent the community. All nodes not presently contained in the collapsed graph, were merged into the graph and edges were re-drawn between the merged nodes and the community node. Results were visualized with the ggraph package6.

1. Alexa A and Rahnenfuhrer J (2016). topGO: Enrichment Analysis for Gene Ontology.
2. Holger Froehlich, N. Speer, A. Poustka, Tim Beissbarth (2007). GOSim - An R-Package for Computation of Information Theoretic GO Similarities Between Terms and Gene Products. BMC Bioinformatics, 8:166
3. Statistical mechanics of community detection Jörg Reichardt and Stefan Bornholdt Phys. Rev. E 74, 016110 – Published 18 July 2006
4. Csardi G, Nepusz T: The igraph software package for complex network research, InterJournal, Complex Systems 1695. 2006.
5. J. Kleinberg. Authoritative sources in a hyperlinked environment. Proc. 9th ACM-SIAM Symposium on Discrete Algorithms, 1998. Extended version in Journal of the ACM 46(1999). Also appears as IBM Research Report RJ 10076, May 1997.
6. Thomas Lin Pedersen (2017). ggraph: An Implementation of Grammar of Graphics for Graphs and Networks. R package version 1.0.0. <https://CRAN.R-project.org/package=ggraph>
7. R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
8. Orchestrating high-throughput genomic analysis with Bioconductor. W. Huber, V.J. Carey, R. Gentleman, ... M. Morgan Nature Methods, 2015:12, 115.

*Community detection*:

1. It may be possible to merge some of the detected communities although, so far, I have not found a valid metric that would allow us to do this in a way that will still highlight many of the communities I think you want to highlight. I think it would be beneficial for the overall understanding of the network graph if there were less communities to visualize but it would probably be best for us to discuss this a bit first.
2. Figures C-J also give the opportunity to highlight specific genes if we want to do so. As it is now, I chose the 10 most significant genes per term. This makes the differences very clear in the visualization (which is beneficial because the plots are so small) but also results in the same gene being plotted several times (since it is included in several terms), like PLK2 for example.

*Differential expression analysis*:

1. I would consider replacing the PCA (figure 1C) and heatmap (Figure 5b) in the manuscript now with something more informative.
   1. A volcano plot (-log pvalue vs. log fold change) could for example be used instead of the PCA and would also give the opportunity to highlight some genes if so desired. This would make it easier to actually say a bit more in the text about the differentially expressed genes we found instead of jumping right into the gene ontology analysis.
   2. I think the difference in ubiquitin related genes (or autophagy, cell cycle, metabolism, etc.) could be better shown either in the figure pertaining to the gene ontology analysis in one of the C-J plots, or in a similar plot located elsewhere in the manuscript.
2. I noticed an error in the previous materials and methods concerning the genome version used in the RNAseq analysis. It has now been changed in the materials and methods above.