

Supplementary Information

The landscape of viral expression and host perturbations in Epstein-Barr virus-associated neoplasms

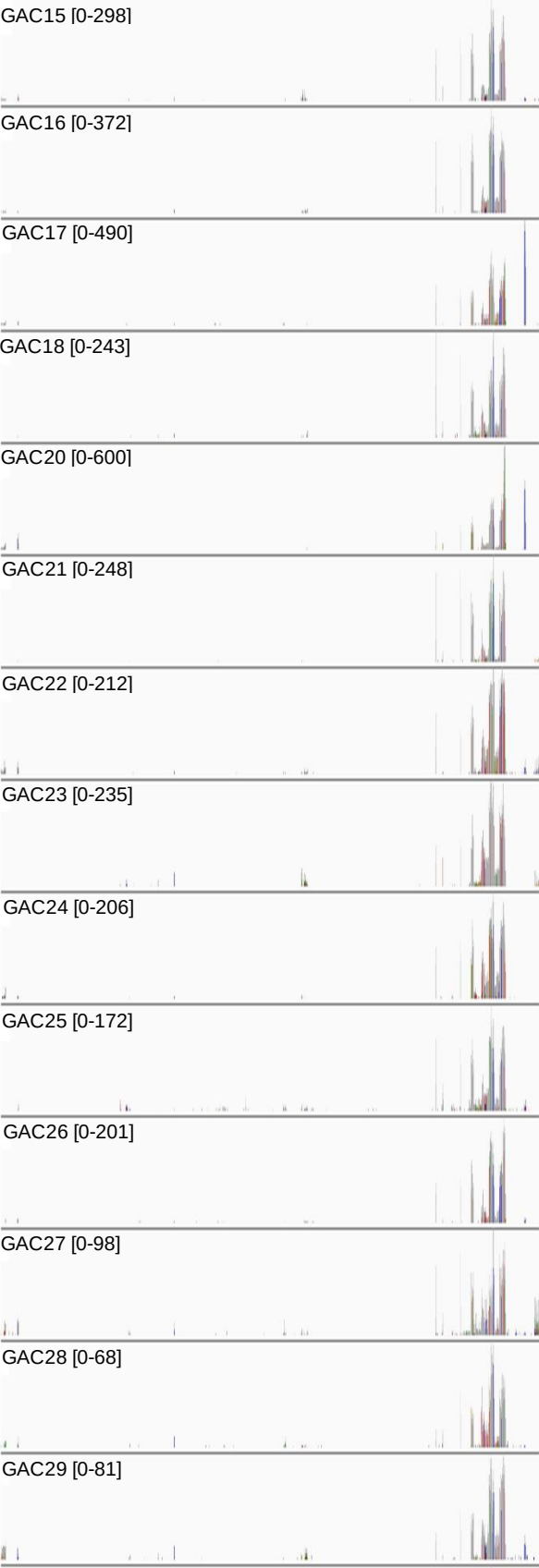
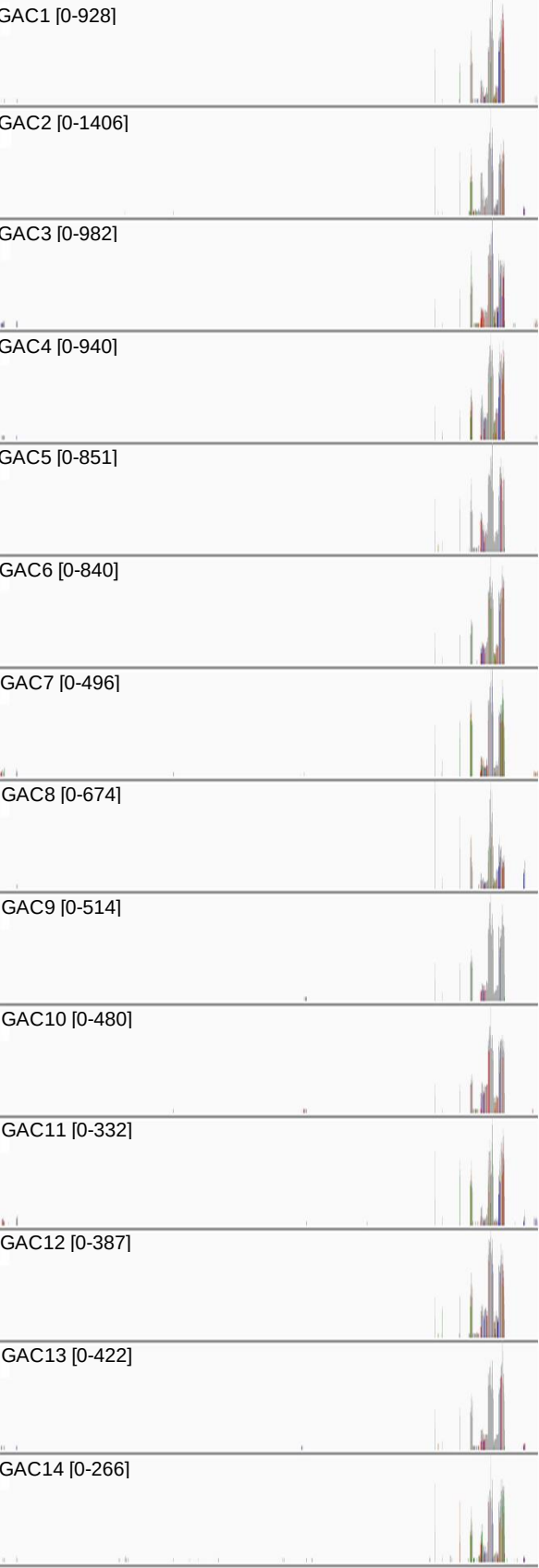
Contents

1.	EBV RNA.....	2
2.	<i>RPMS1</i> transcription start site	14
3.	BamHI-A/I splice-junctions	26
4.	Namalwa modified cell lines	44
5.	BamHI-A rightward elements	45
6.	Tpm-threshold	46
7.	EBV gene expression	47
8.	EBV-positive cells.....	55
9.	EBV reactivation	65
10.	Cancer cell identification.....	67
11.	Epithelial cell classification.....	74
12.	References	79

1. EBV RNA

EBV reads detected in each dataset was aligned to the EBV genome. Datasets were sorted in a decreasing order of the EBV ppm levels for each tumor type. Coverage plots were generated for all positive datasets (>10 ppm EBV RNA), numbers within brackets indicate the y-axis scale, total number of reads. Alignment to the BamHI digestion map is shown at the bottom of each page. The eBL datasets contained strand information which allowed for separation of the rightward and leftward reads. The y-axis scale for each rightward plot was maintained as reference for the leftward plot for each tumor. The eBL datasets were unlike the other studies, generated using a strand-specific sequencing library. In 15 out of 16 eBL-datasets, more than 99% of the RNA identified within the BamHI-A/I region were in a rightward direction (Figure 1d, third panel above/below baseline; Extended Data Table 1). The reads mapped to the unique mid-region of *BALF5* are therefore not likely to originate from the leftward transcript *BALF5*. Modified Namalwa cell lines with control plasmid (pTRE), plasmid containing inducible *RPMS1* long non-coding RNA (pTRE-RPMS1) and modified *RPMS1*-promoter with replaced inducible promoter (Pro-re) were treated with indicated amount of doxycycline. GAC, gastric adenocarcinoma; NPC, nasopharyngeal carcinoma; sBL, sporadic Burkitt's lymphoma; eBL, endemic Burkitt's lymphoma; ECL, EBV-associated tumor derived cell lines; LCL, lymphoblastoid cell lines.

GAC



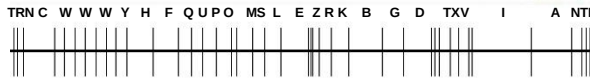
GAC19 [0-415]



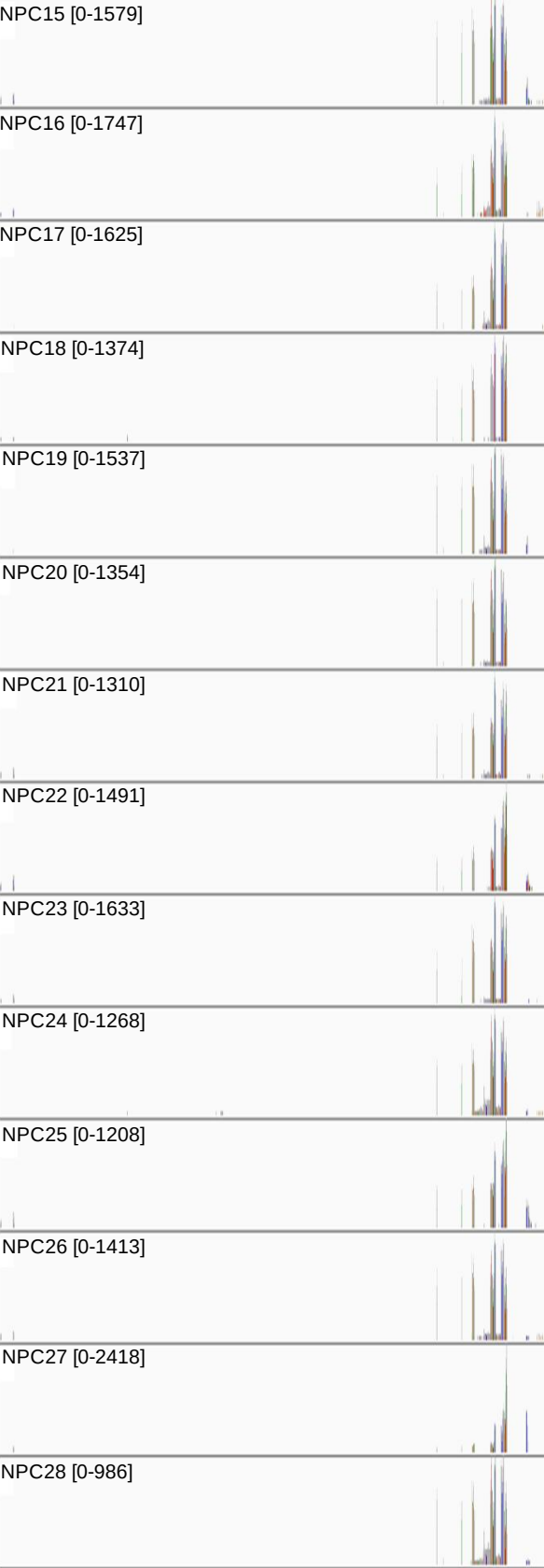
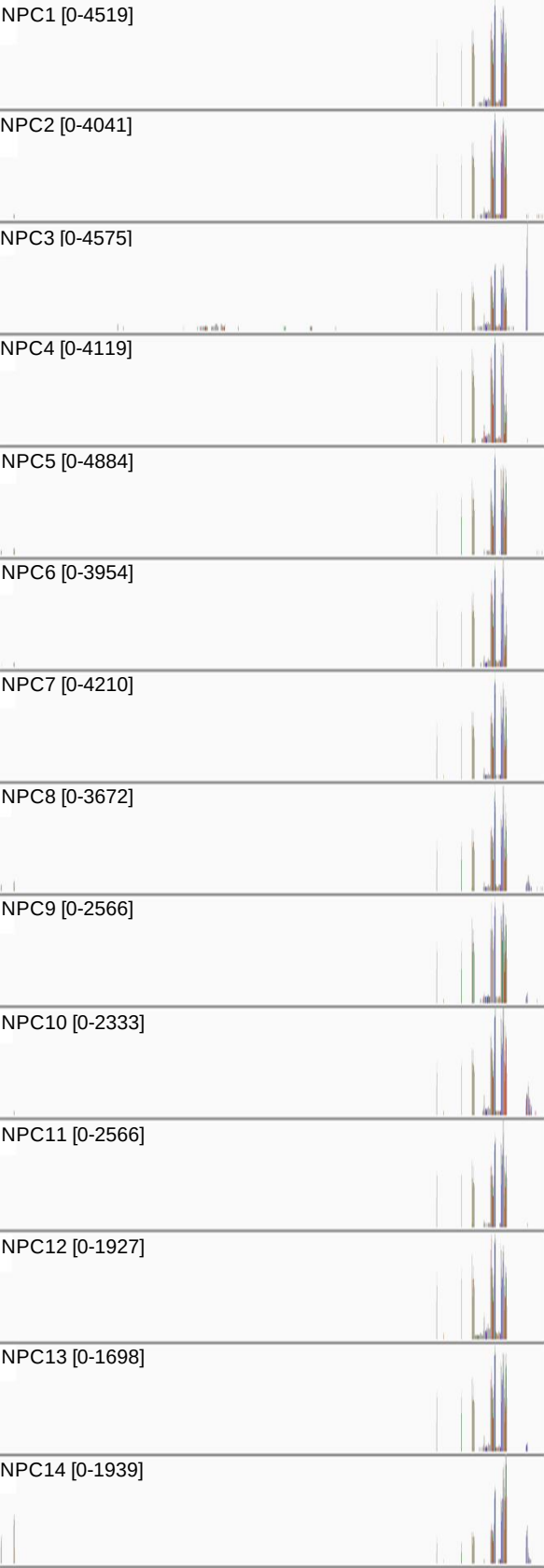
GAC30 [0-114]

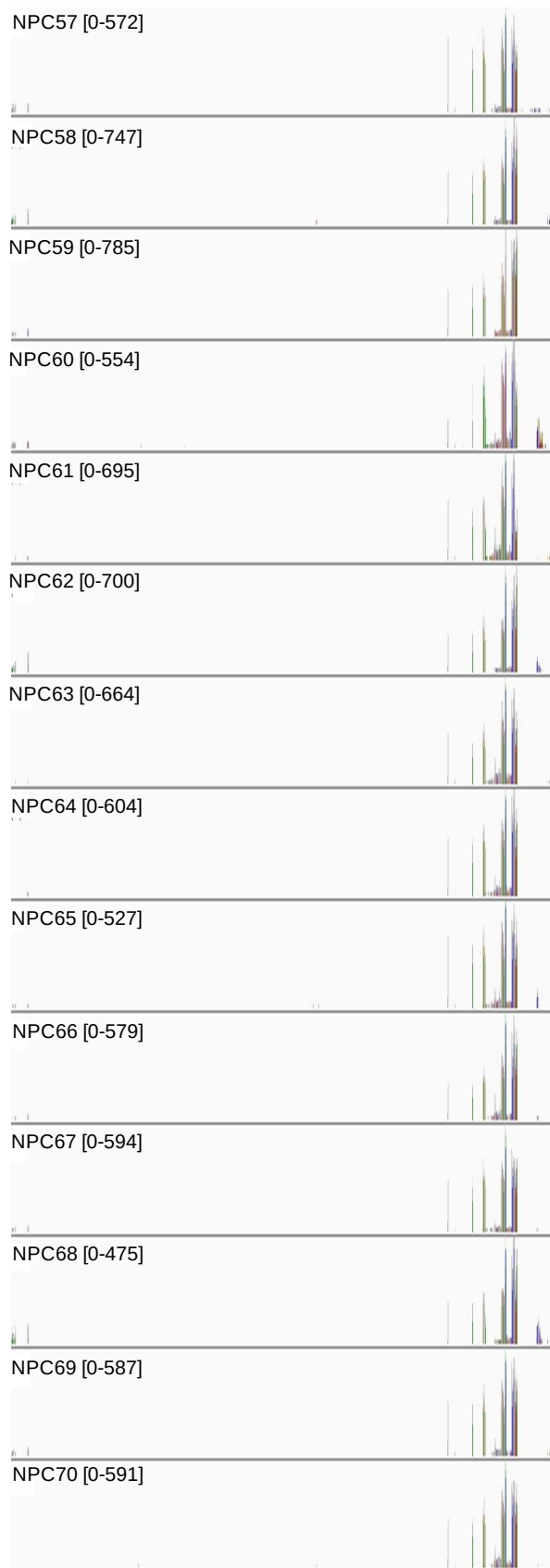


TRNC W W W Y H F Q U P O MS L E Z R K B G D TXV I A NTR

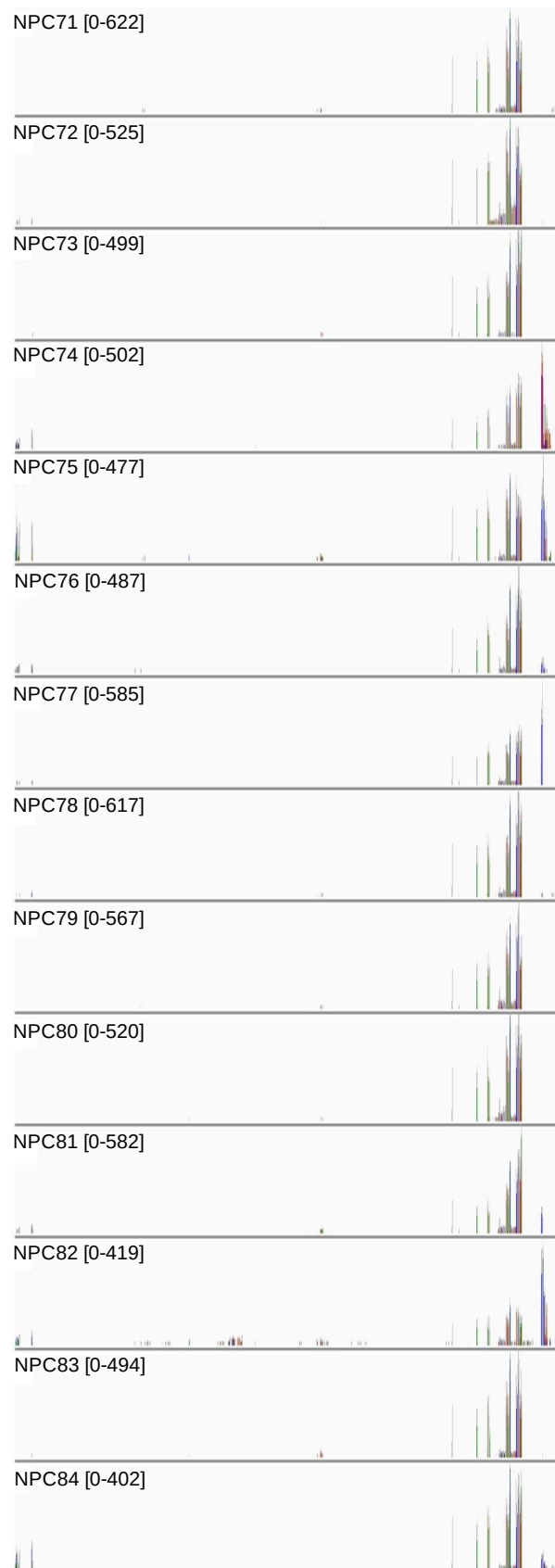


NPC





TRNC W W W Y H F Q U P O M S L E Z R K B G D T X V I A NTR



TRNC W W W Y H F Q U P Q M S L E R K B G D T X V I A NTR

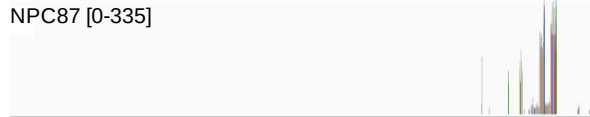
NPC85 [0-515]



NPC86 [0-338]



NPC87 [0-335]



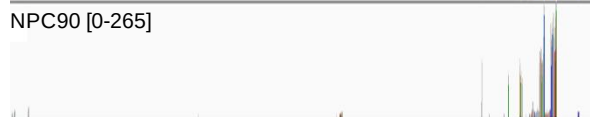
NPC88 [0-310]



NPC89 [0-377]



NPC90 [0-265]



NPC91 [0-301]



NPC92 [0-257]



NPC93 [0-480]



NPC94 [0-283]



NPC95 [0-299]



TRNC W W W Y H F Q U P O M S L E Z R K B G D T X V I A N T R

NPC96 [0-239]



NPC97 [0-340]



NPC98 [0-243]



NPC99 [0-212]



NPC100 [0-186]



NPC101 [0-181]



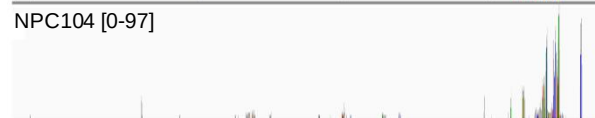
NPC102 [0-87]



NPC103 [0-144]



NPC104 [0-97]



NPC105 [0-71]



NPC106 [0-65]



TRNC W W W Y H F Q U P O M S L E Z R K B G D T X V I A N T R

sBL

sBL1 [0-3184]



sBL2 [0-1891]



sBL3 [0-2713]

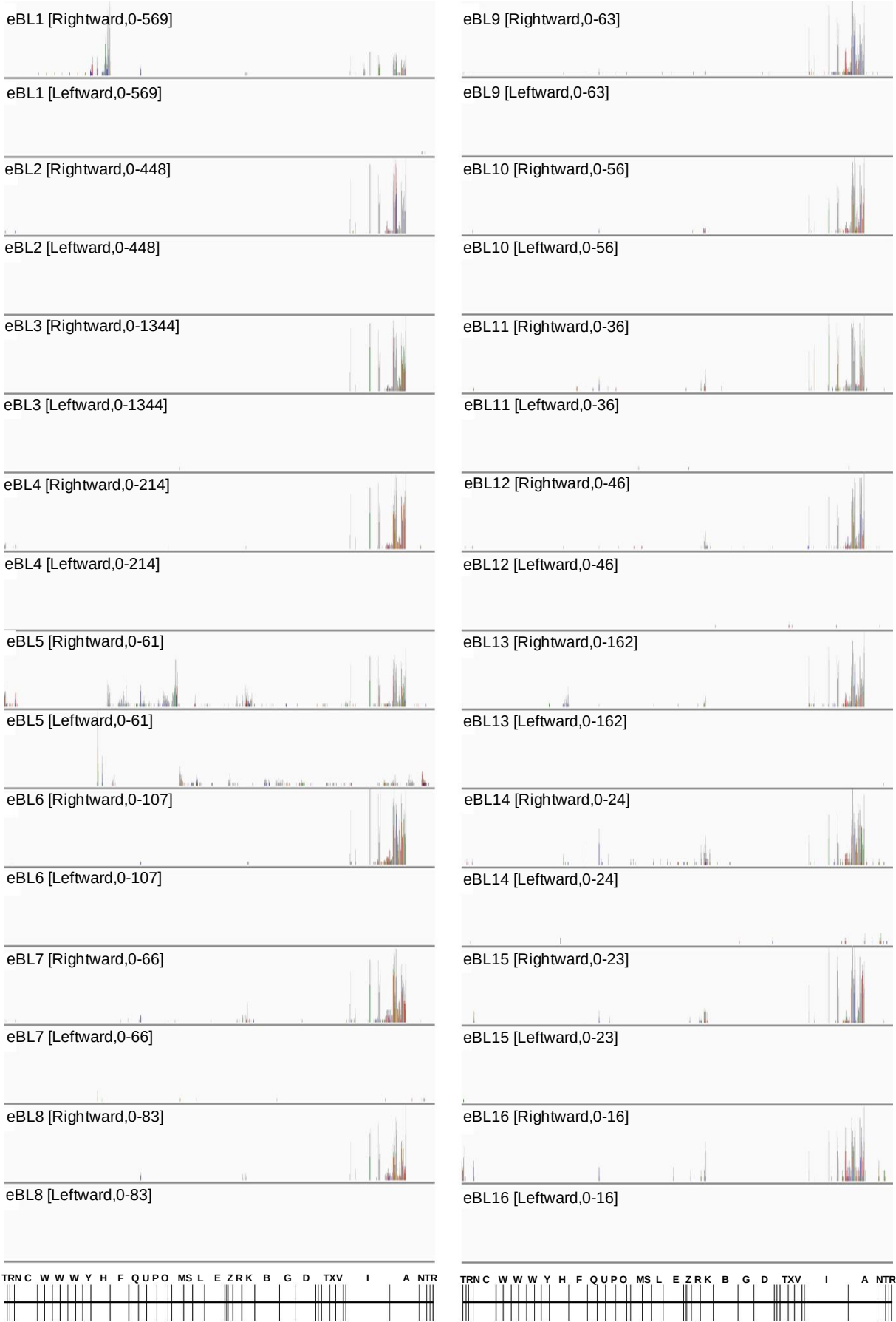


sBL4 [0-953]



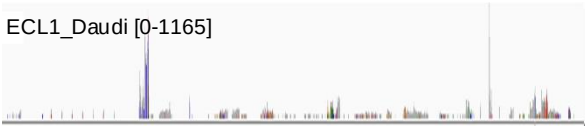
TRNC W W W Y H F Q U P O M S L E Z R K B G D T X V I A N T R

eBL



ECL

ECL1_Daudi [0-1165]



ECL2_C666-1 [0-939]



ECL3_Raji [0-338]



ECL4_Namalwa [0-588]



ECL5_BL58 [0-1698]



ECL6_BL41 [0-520]



LCL

LCL1 [0-37232]



LCL2 [0-62298]



LCL3 [0-17613]



LCL4 [0-17651]



LCL5 [0-12426]



LCL6 [0-15327]



LCL7 [0-20255]



LCL8 [0-19364]



LCL9 [0-7088]



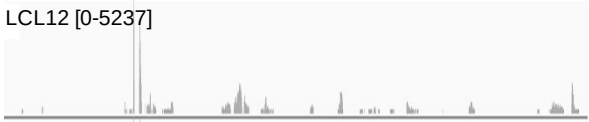
LCL10 [0-11465]



LCL11 [0-10245]



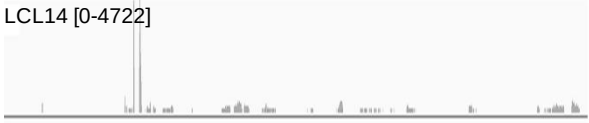
LCL12 [0-5237]



LCL13 [0-7705]



LCL14 [0-4722]



LCL15 [0-8557]



LCL16 [0-13065]



LCL17 [0-10058]



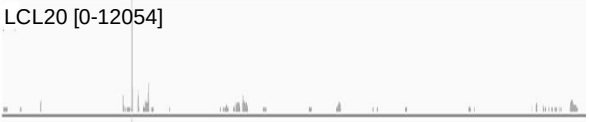
LCL18 [0-5325]



LCL19 [0-5136]



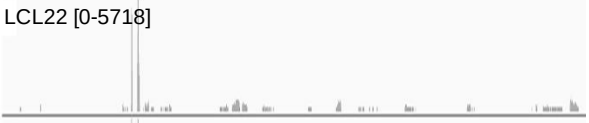
LCL20 [0-12054]



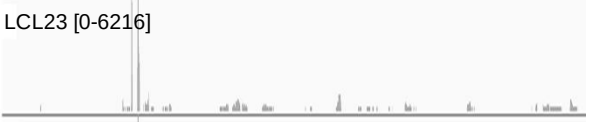
LCL21 [0-3434]



LCL22 [0-5718]



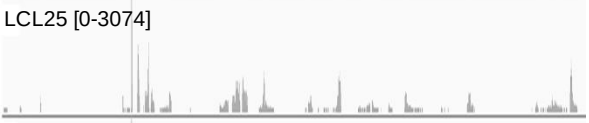
LCL23 [0-6216]



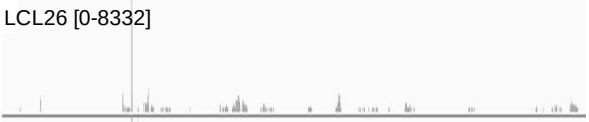
LCL24 [0-6617]



LCL25 [0-3074]



LCL26 [0-8332]

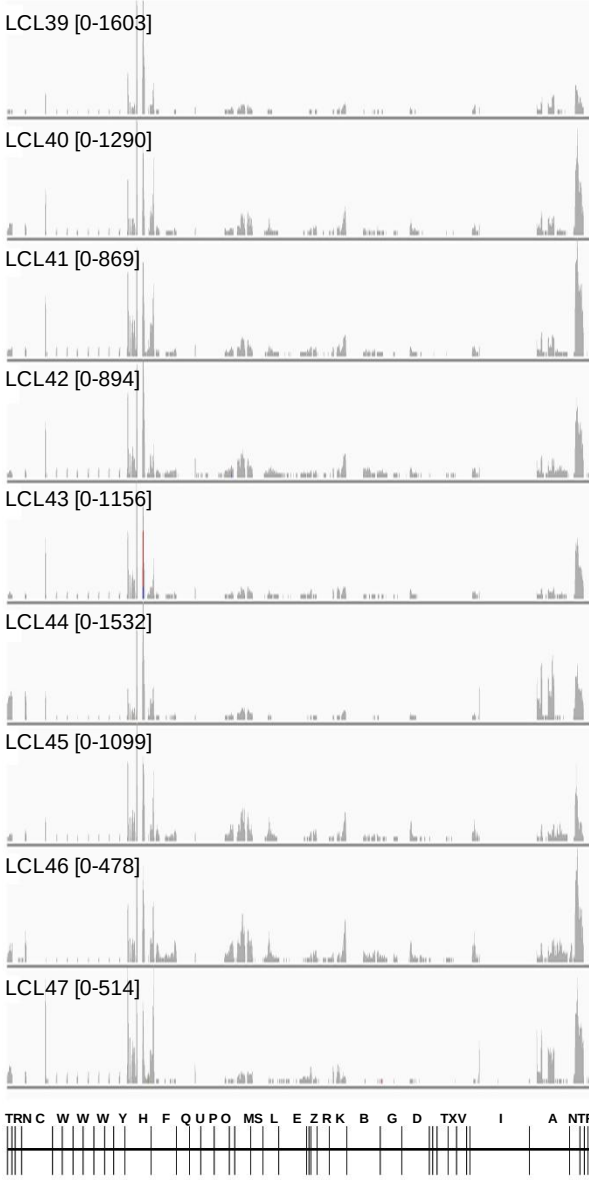
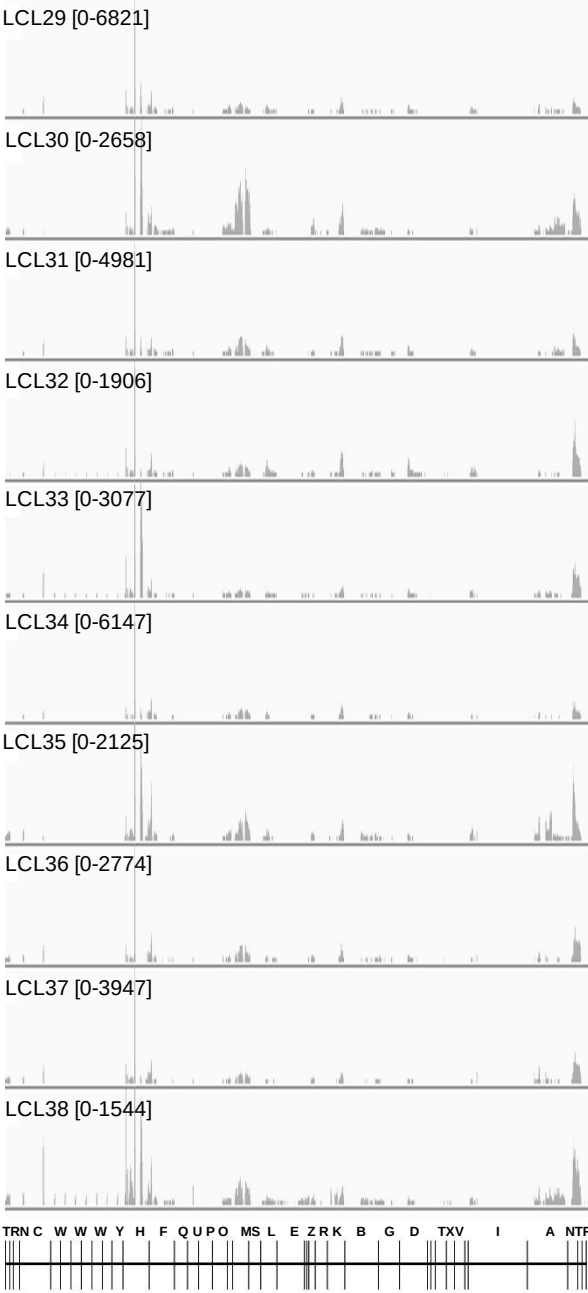


LCL27 [0-2983]

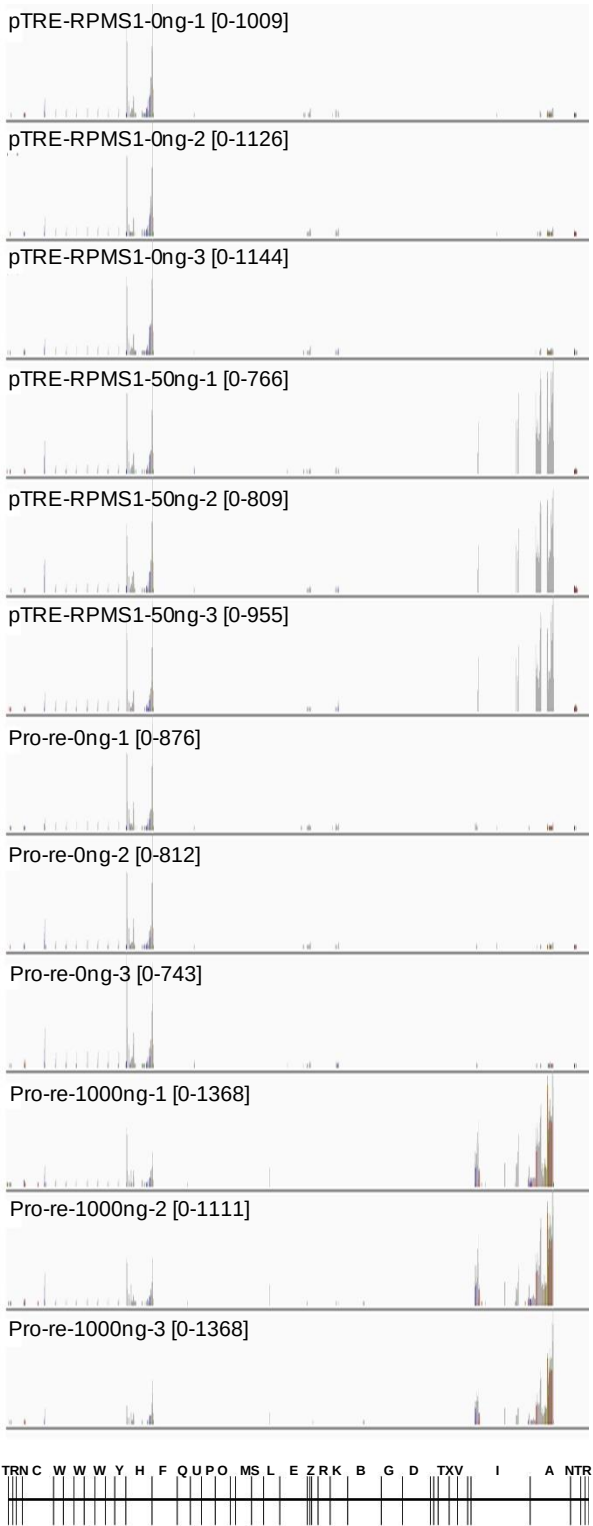
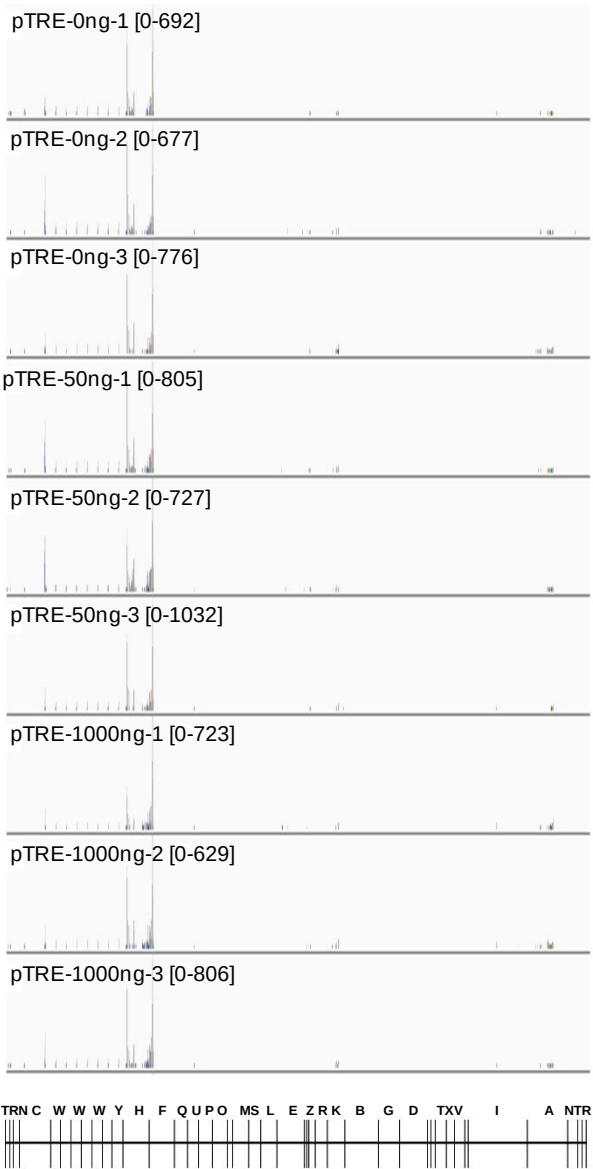


LCL28 [0-4870]





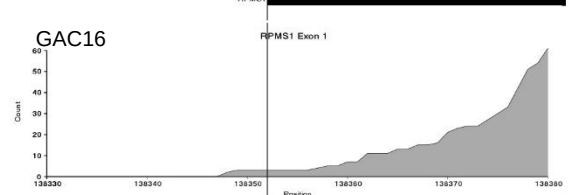
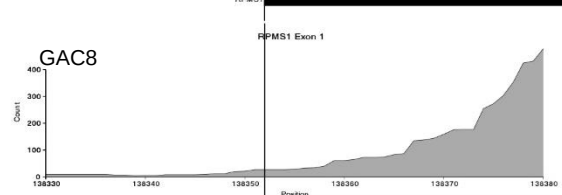
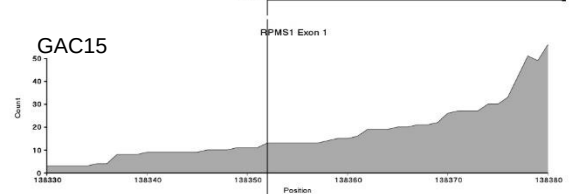
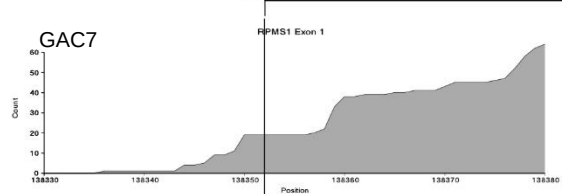
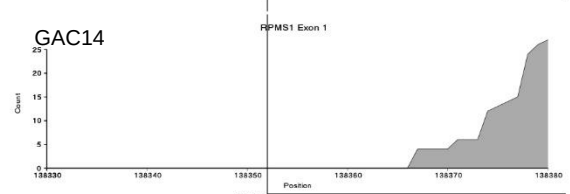
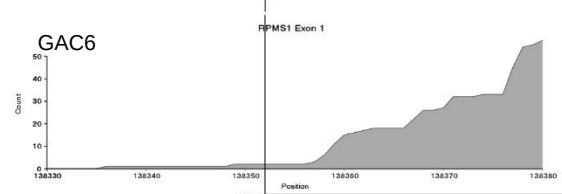
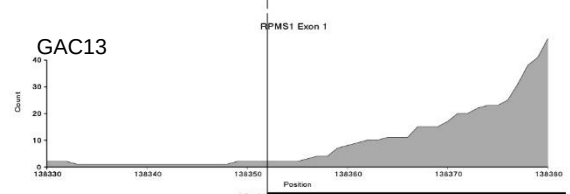
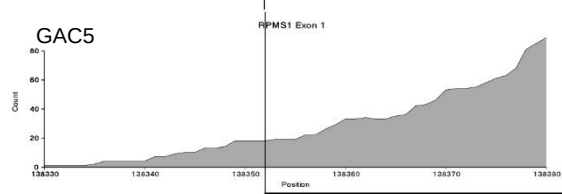
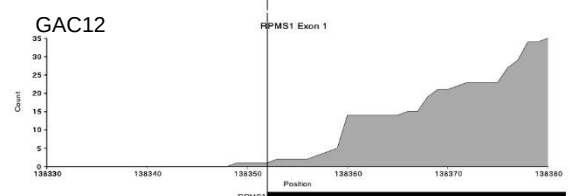
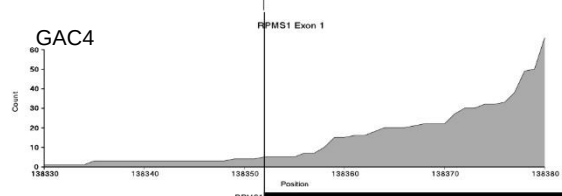
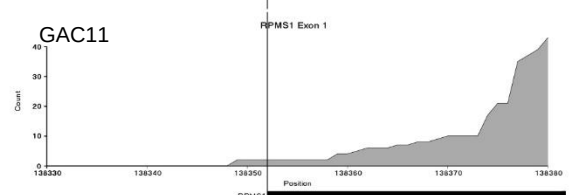
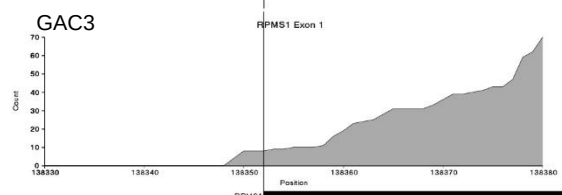
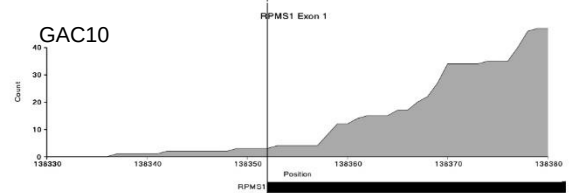
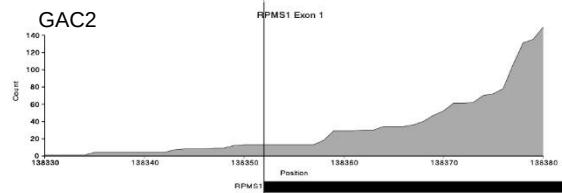
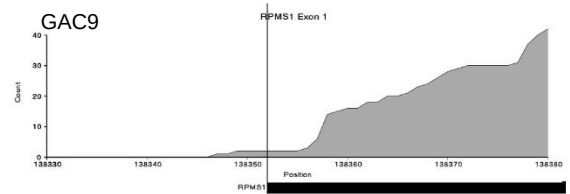
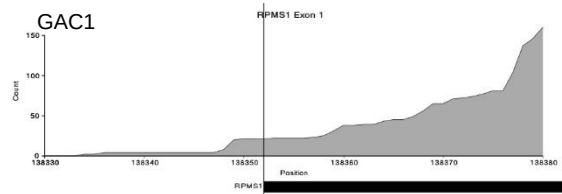
Modified Namalwa cell lines

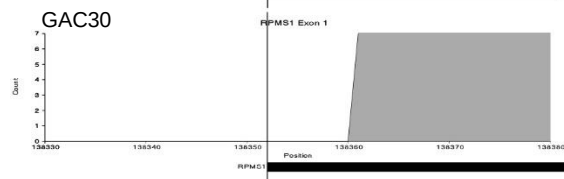
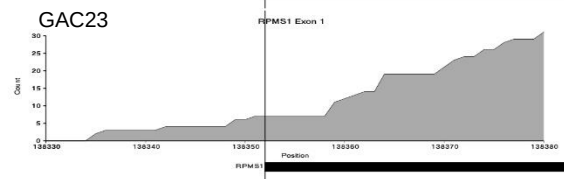
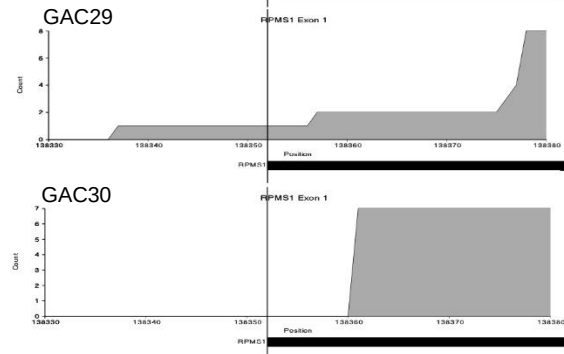
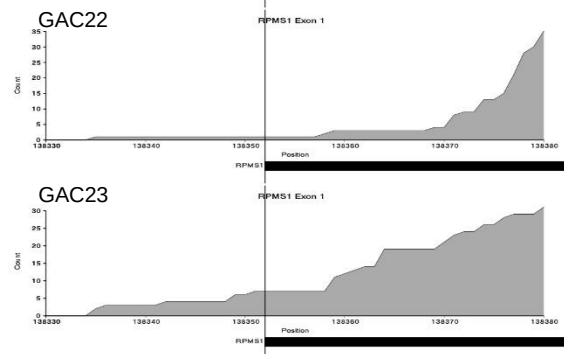
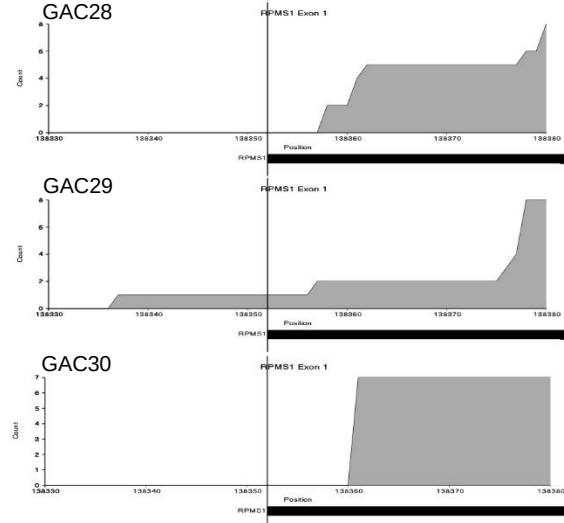
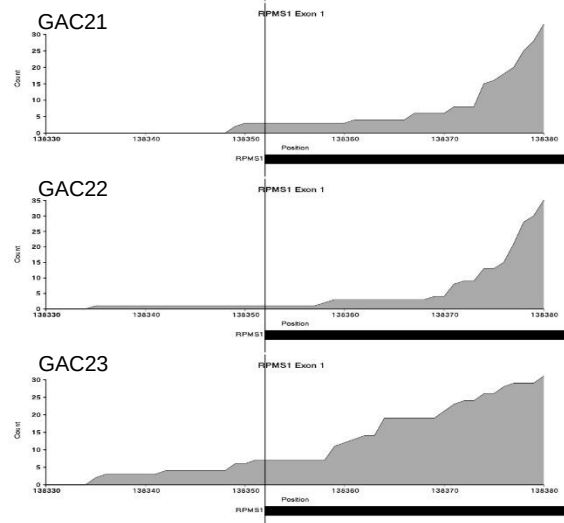
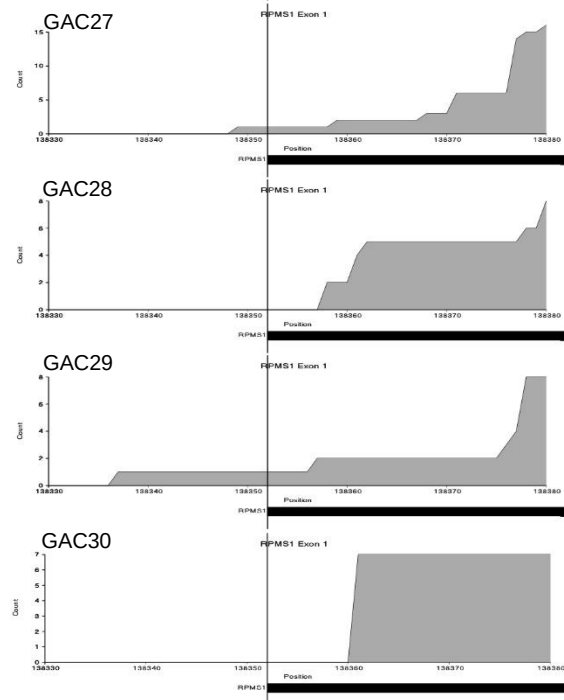
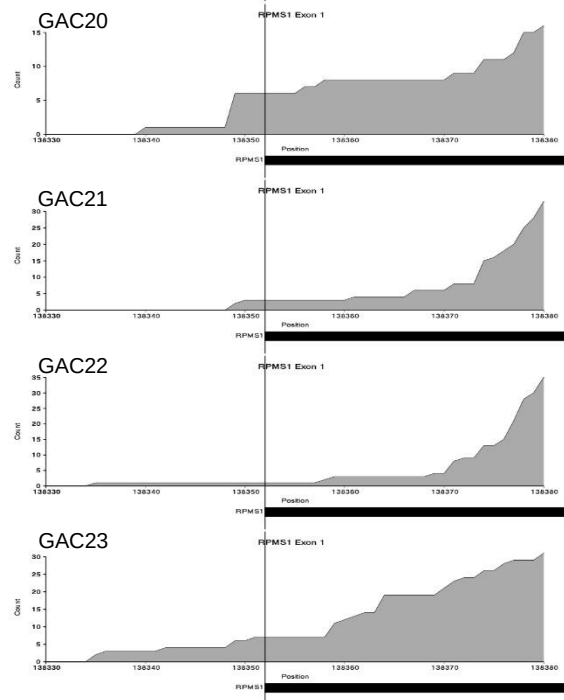
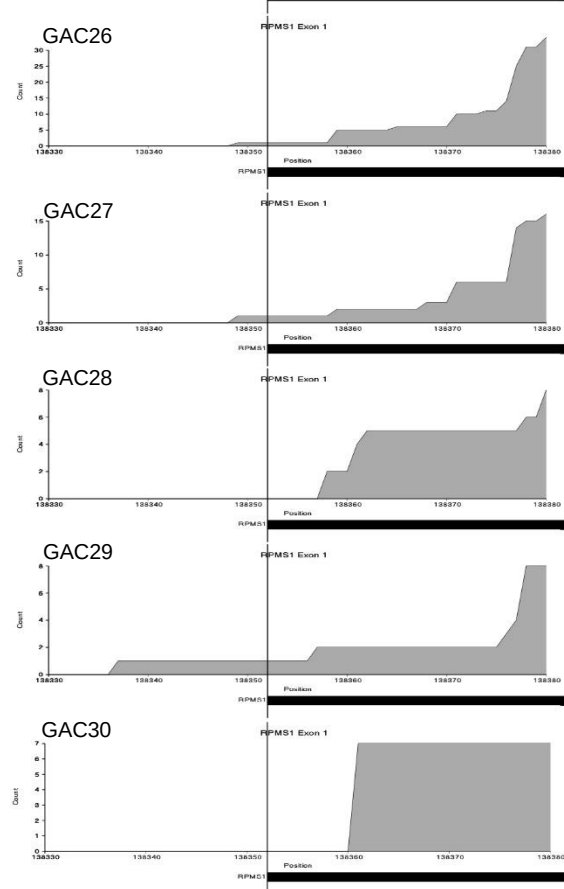
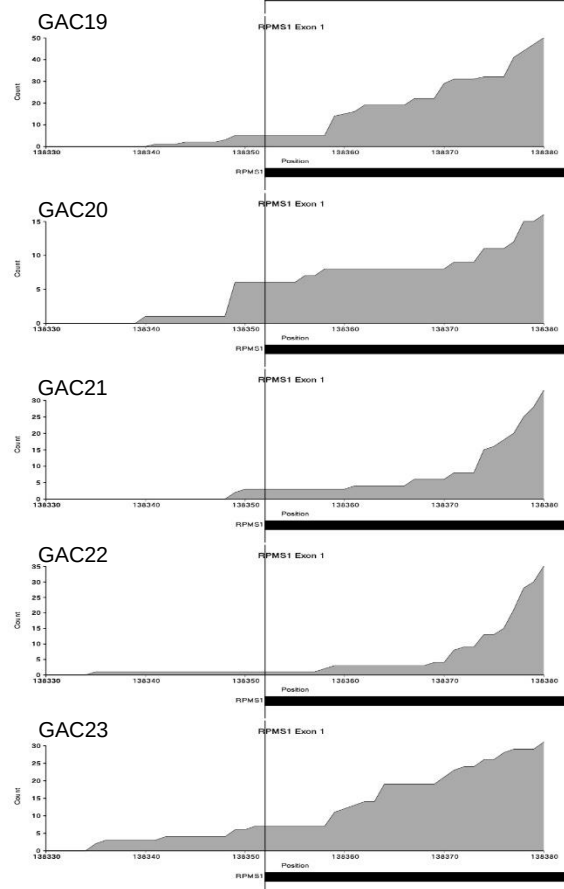
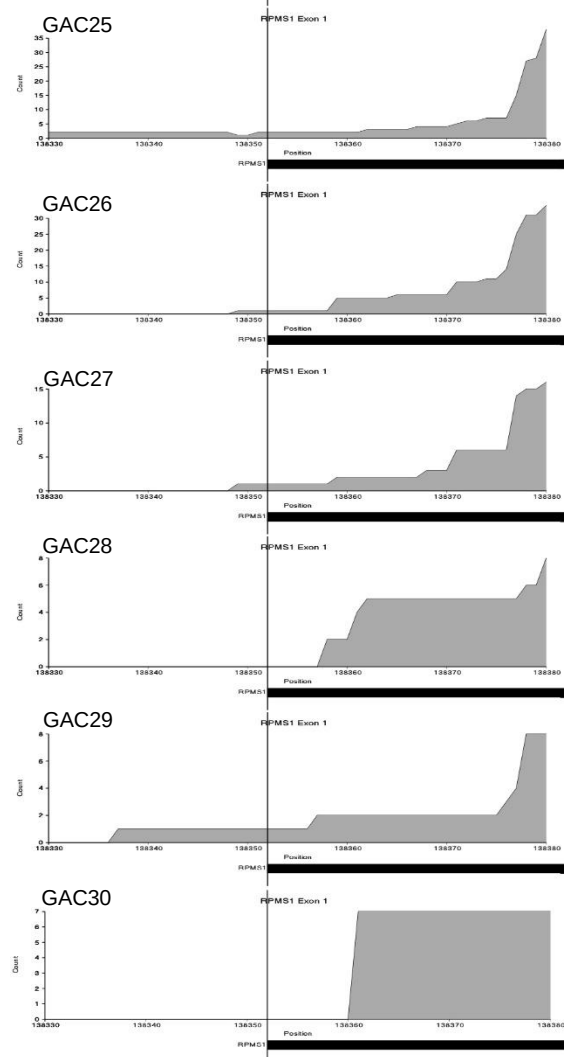
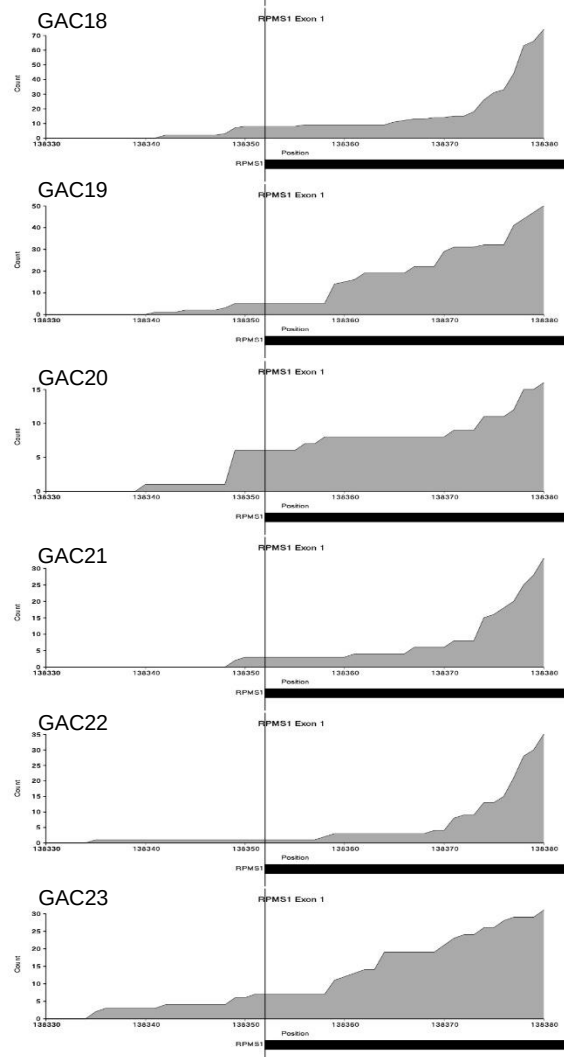
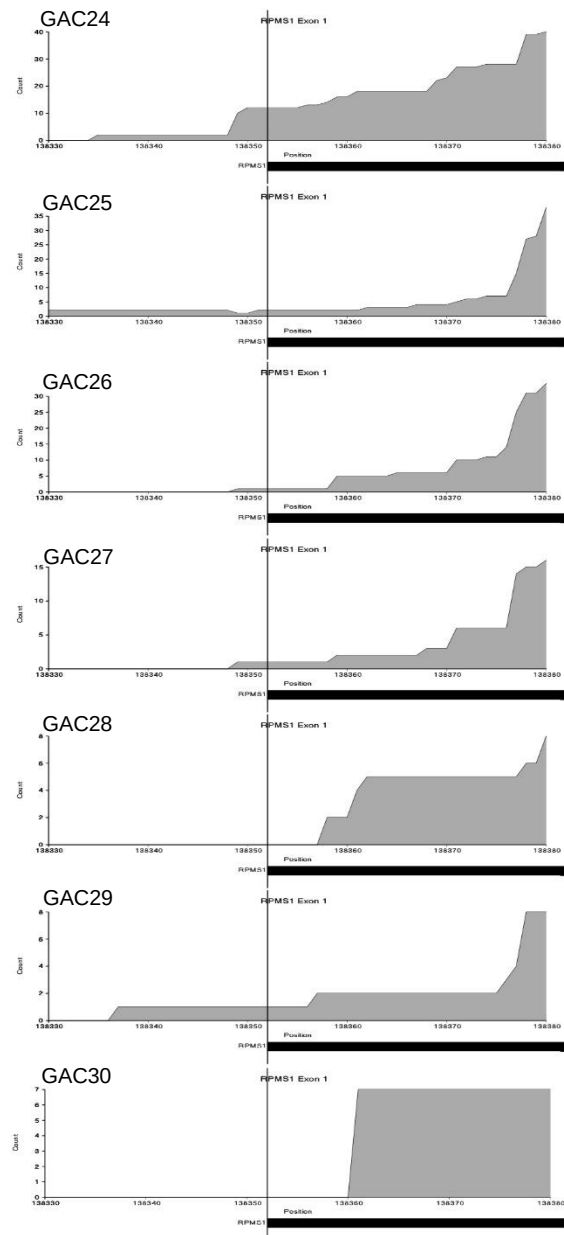
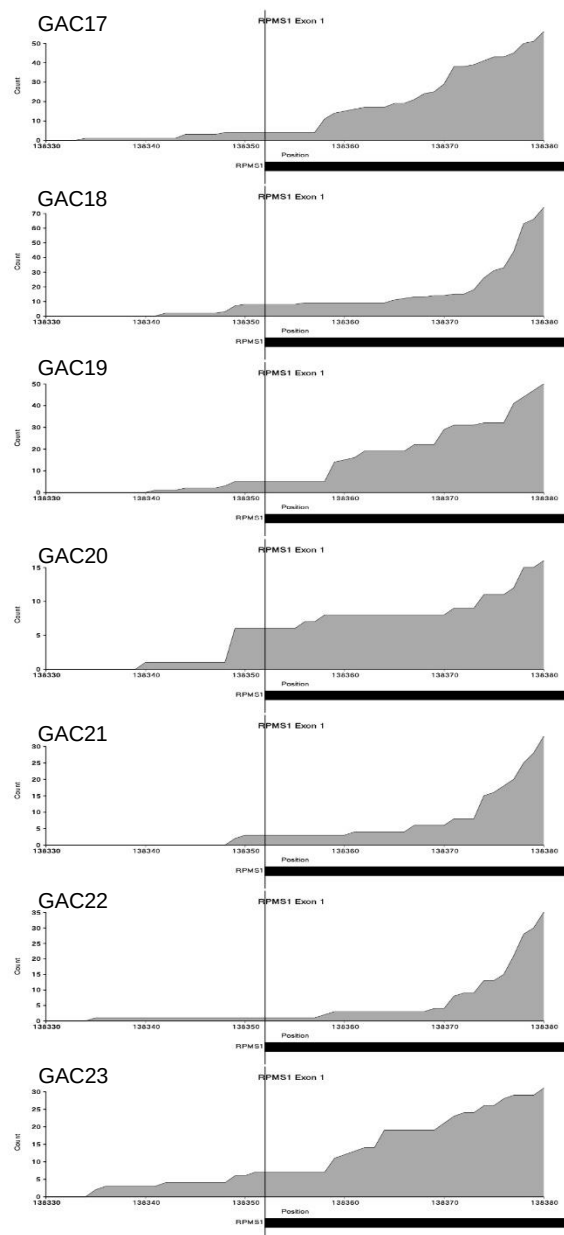


2. *RPMS1* transcription start site

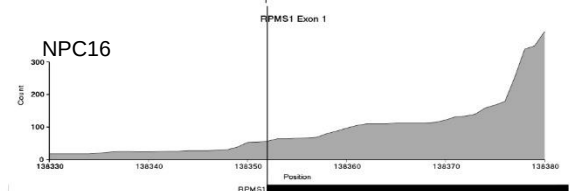
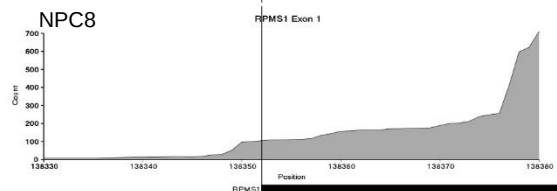
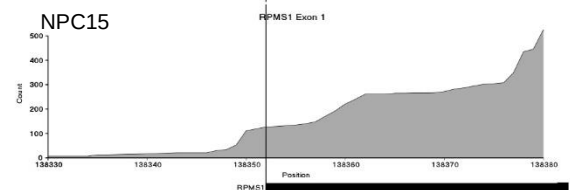
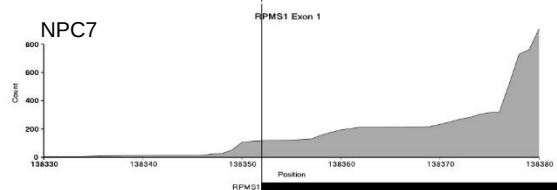
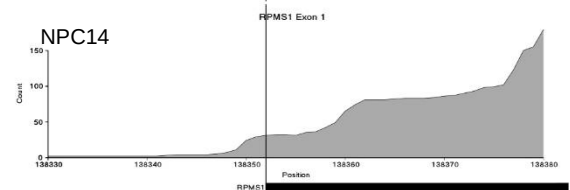
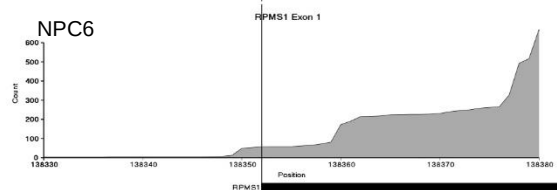
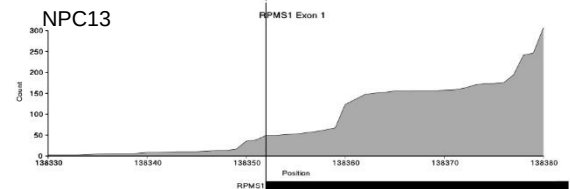
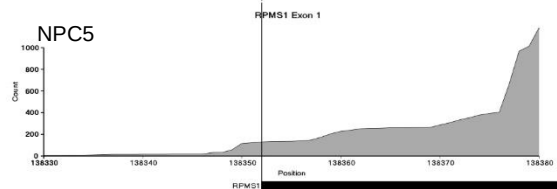
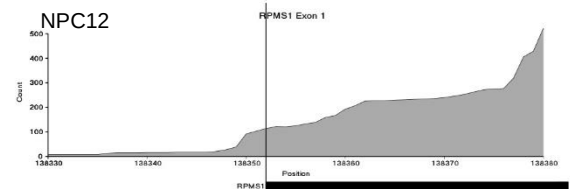
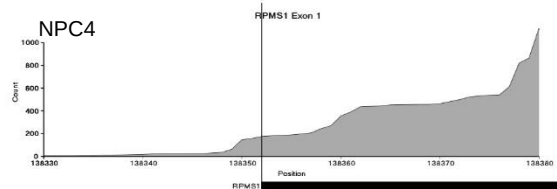
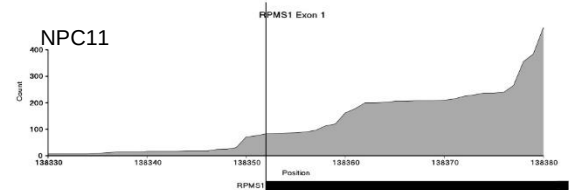
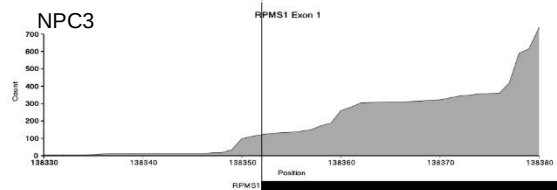
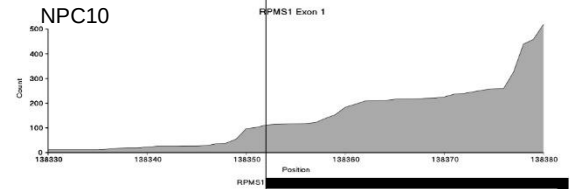
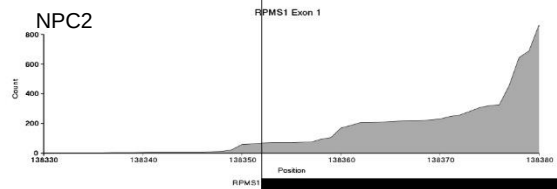
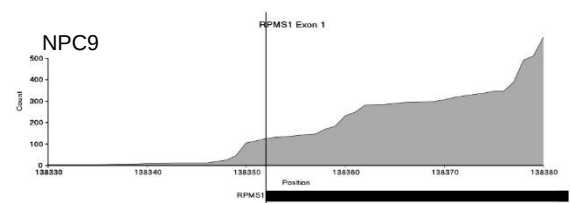
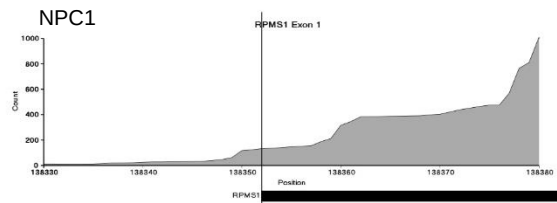
Sequencing data from all tumors originated from poly-adenylated enriched RNA. In such datasets a decline in coverage is expected at the 5'-end of the transcript due to sequencing library preparation bias. Coverage plot of the RNA detected in the region of the *RPMS1* transcription start site in EBV-positive datasets. The vertical line indicates the transcription start site of *RPMS1*. The black bar shows the upstream portion of the *RPMS1* exon 1. Numbers in the x-axis indicate position in the EBV genome. Number in the y-axis indicate the number of reads. A reduced coverage at the transcription start site of *RPMS1* was observed in almost all samples which indicates that the major RNA element is initiated at exon I. GAC, gastric adenocarcinoma; NPC, nasopharyngeal carcinoma; eBL, endemic Burkitt's lymphoma; sBL, sporadic Burkitt's lymphoma.

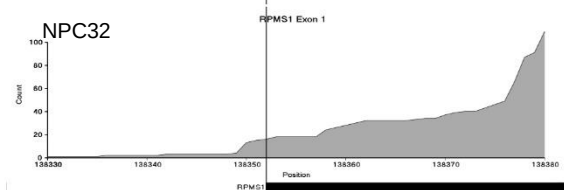
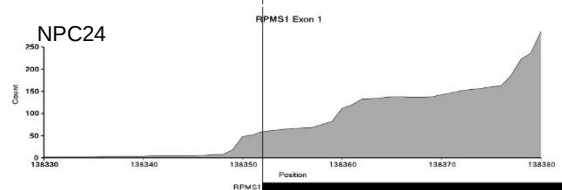
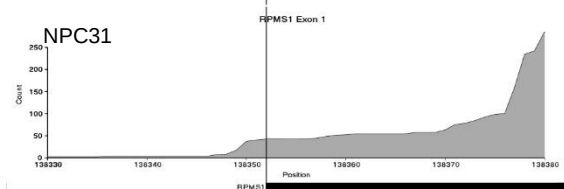
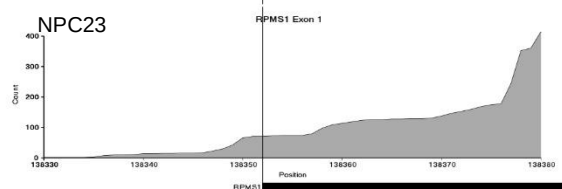
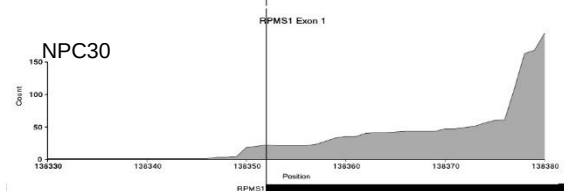
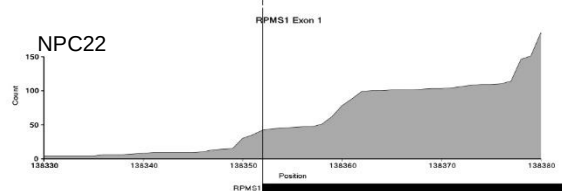
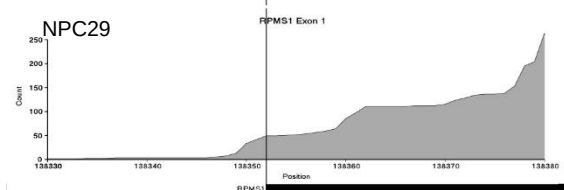
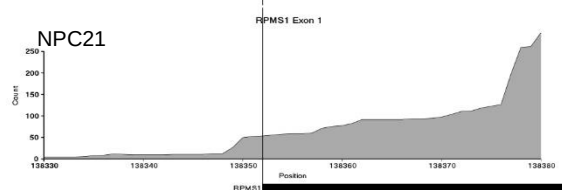
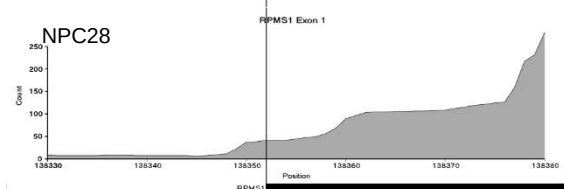
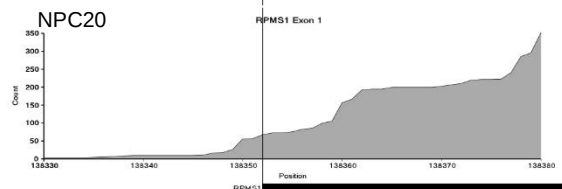
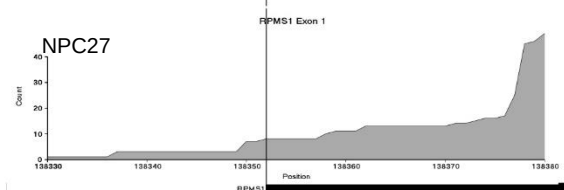
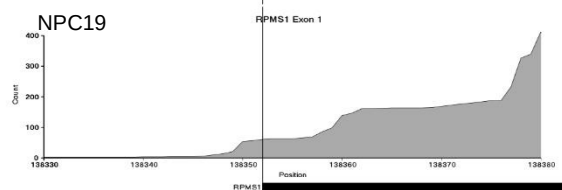
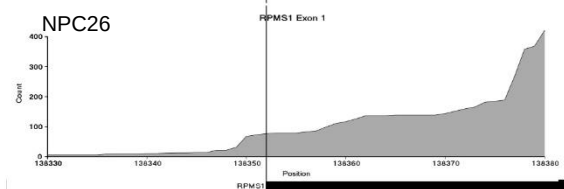
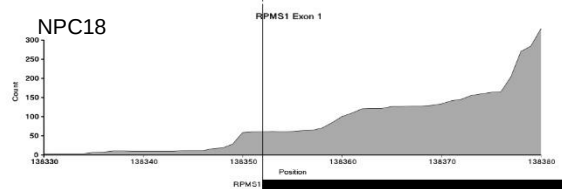
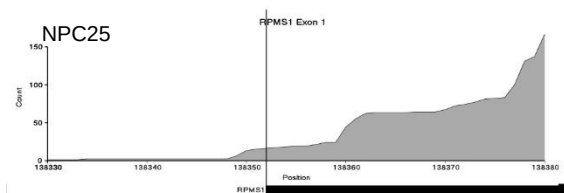
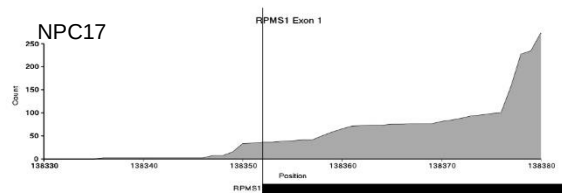
GAC

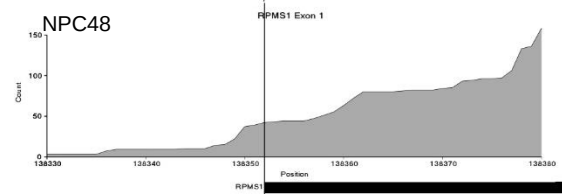
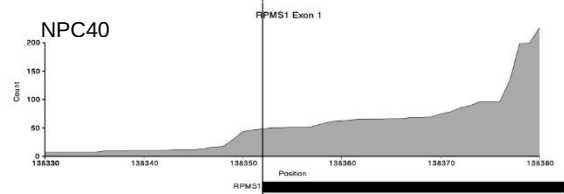
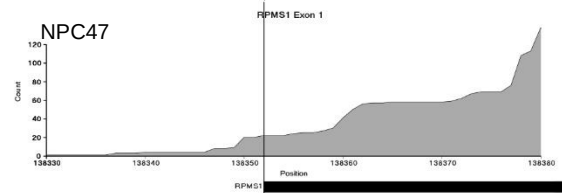
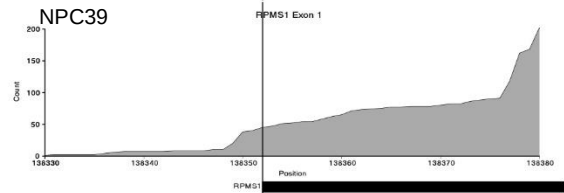
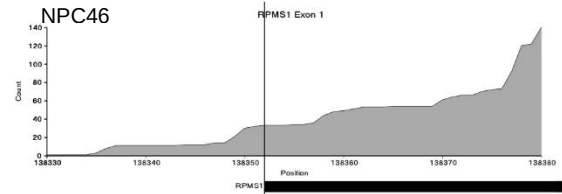
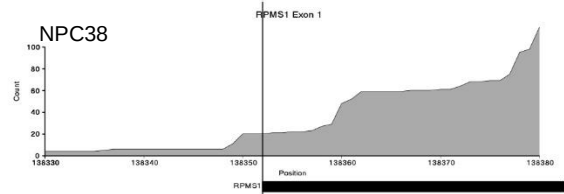
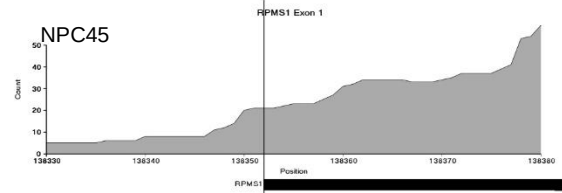
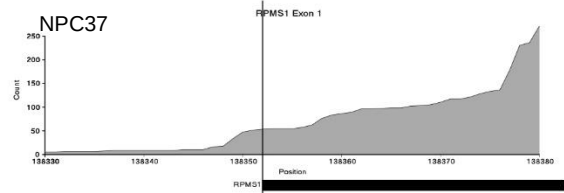
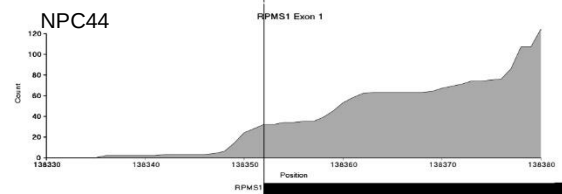
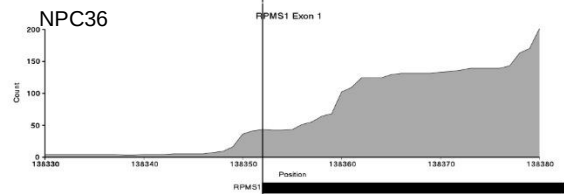
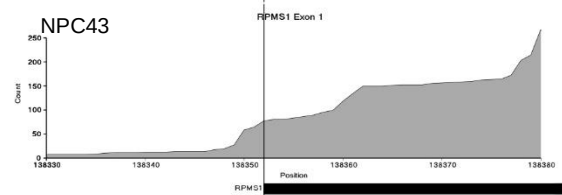
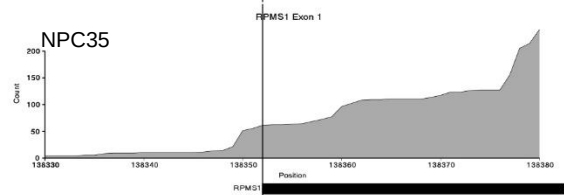
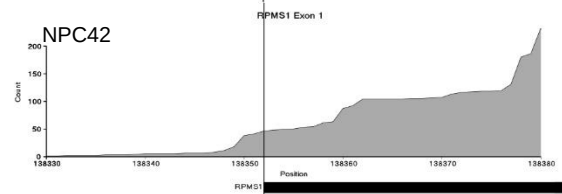
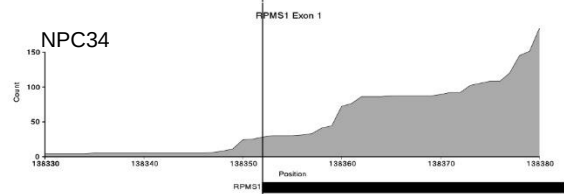
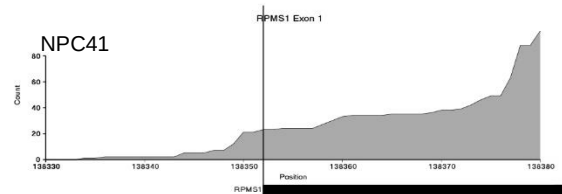
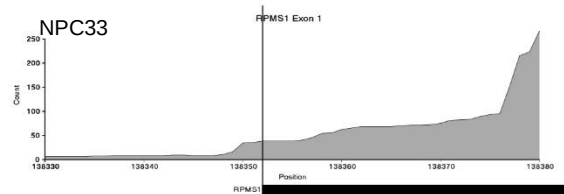


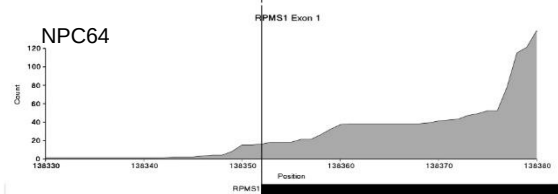
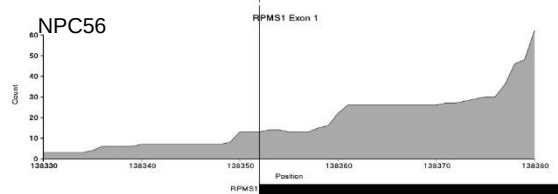
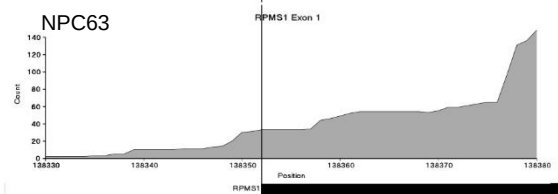
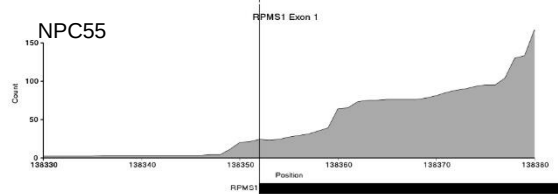
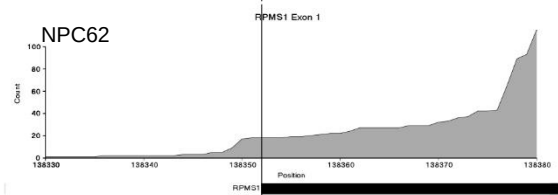
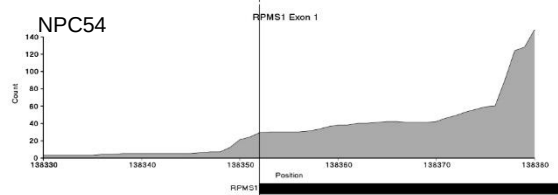
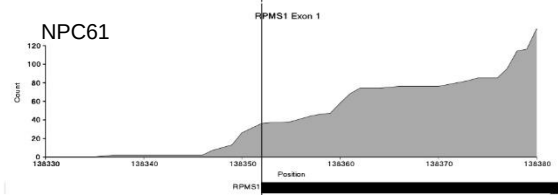
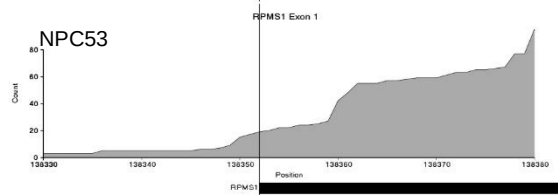
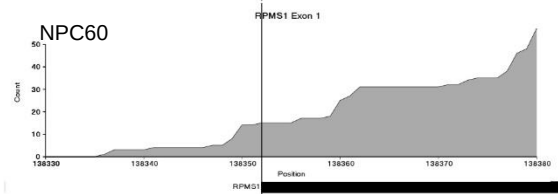
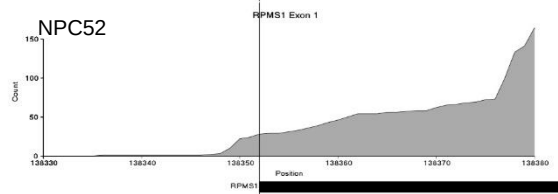
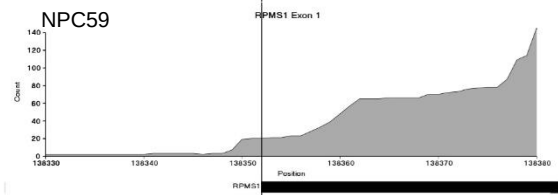
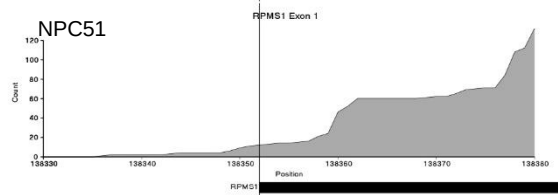
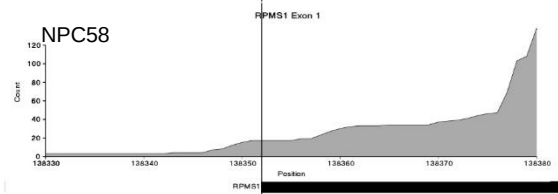
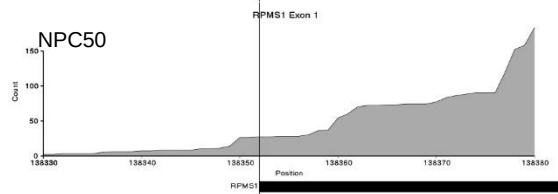
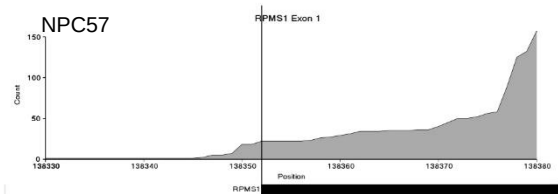
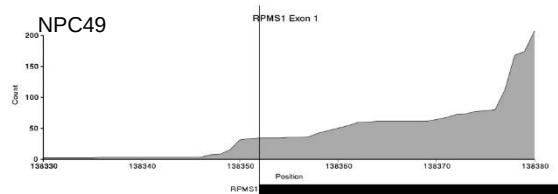


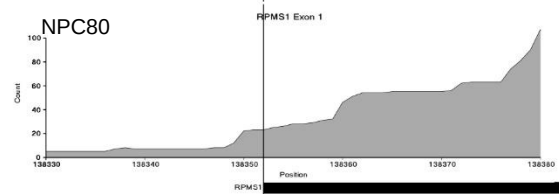
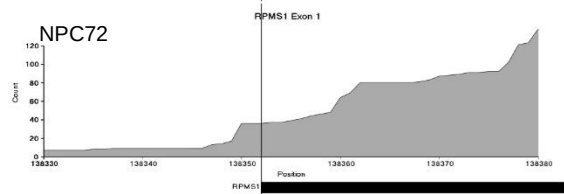
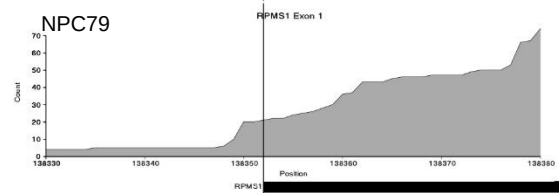
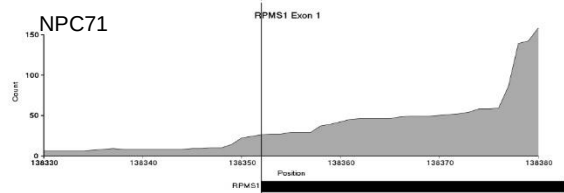
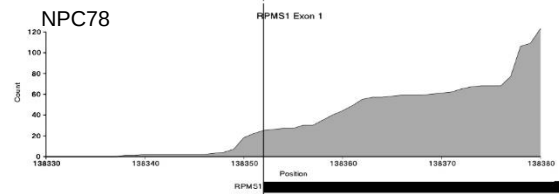
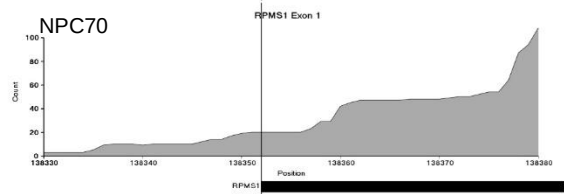
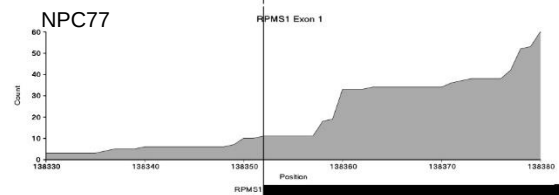
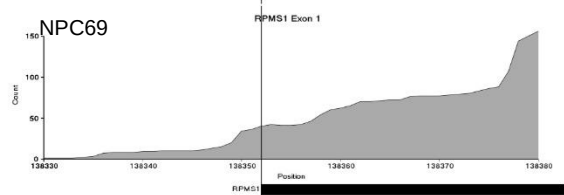
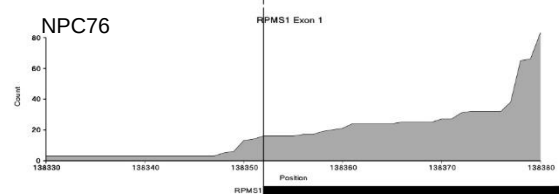
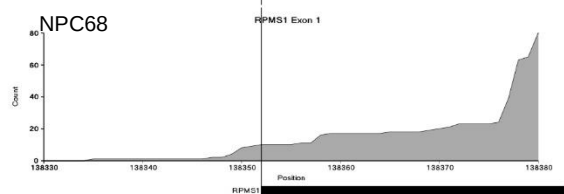
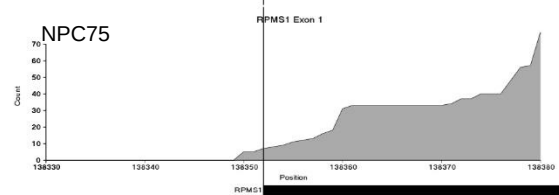
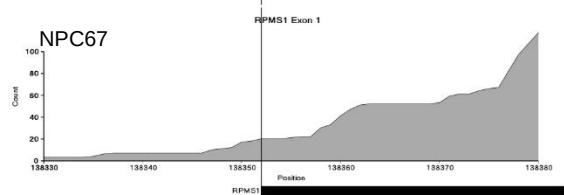
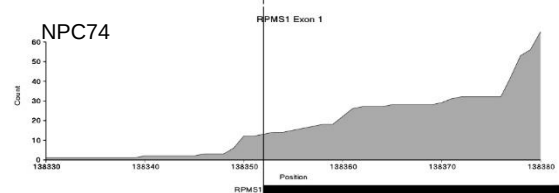
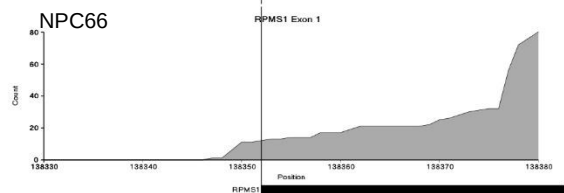
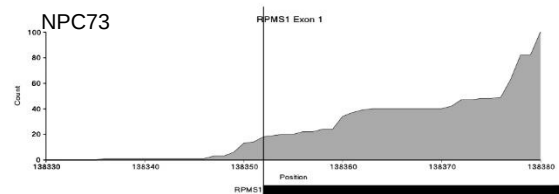
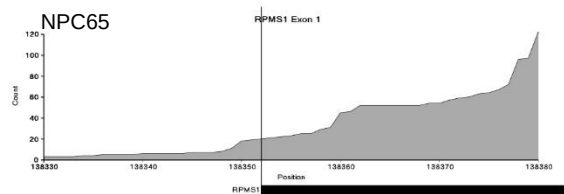
NPC

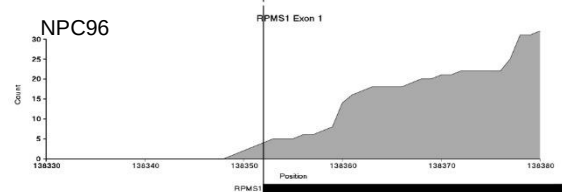
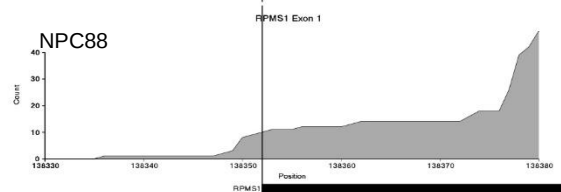
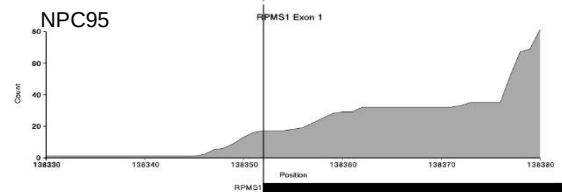
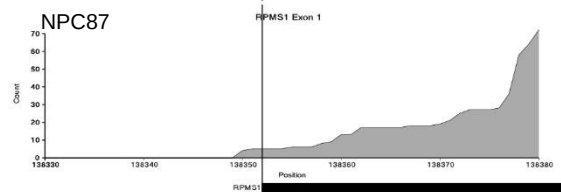
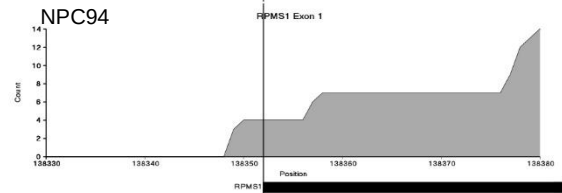
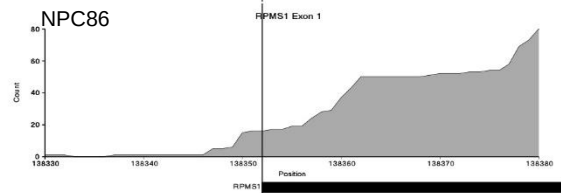
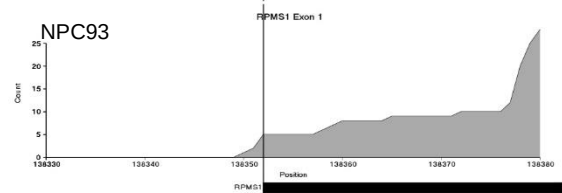
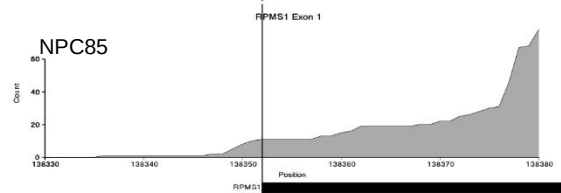
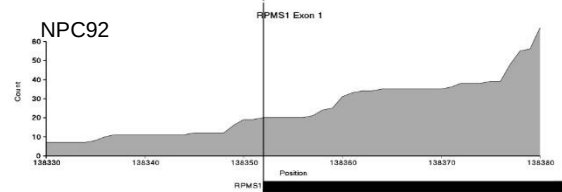
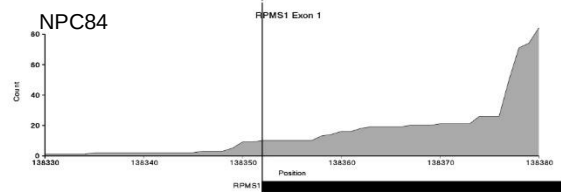
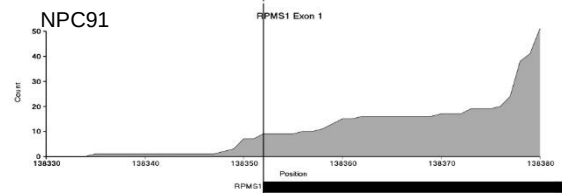
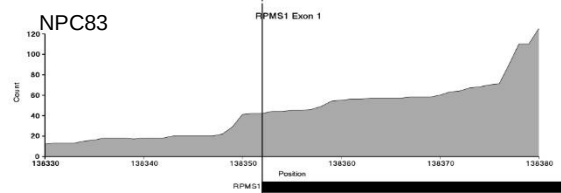
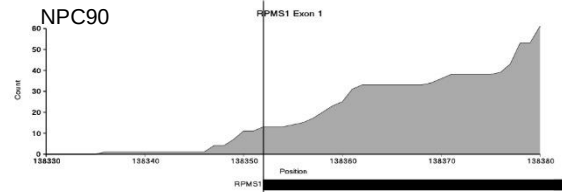
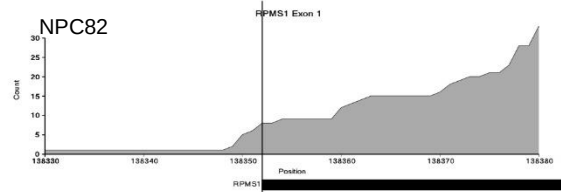
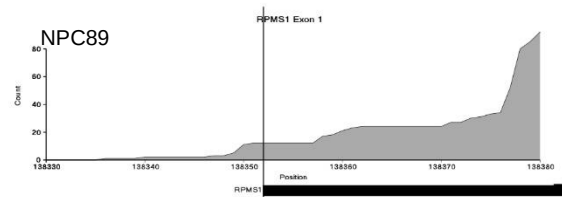
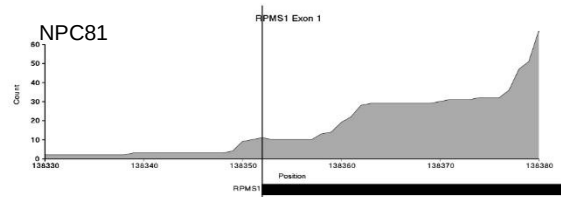


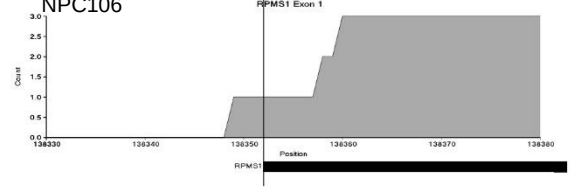
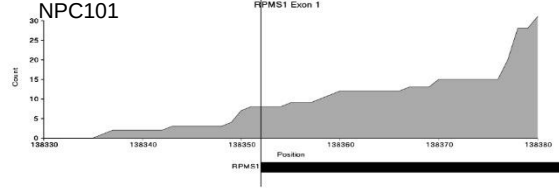
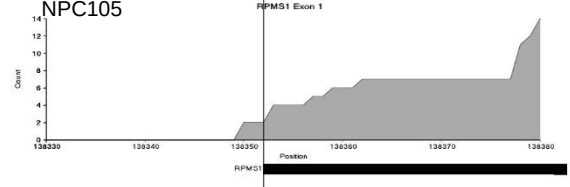
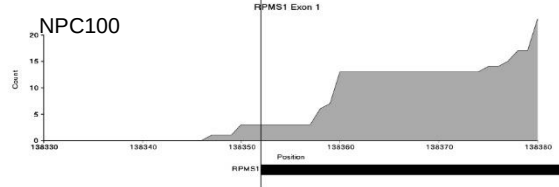
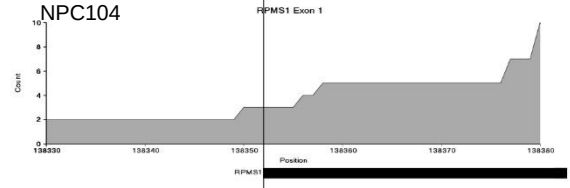
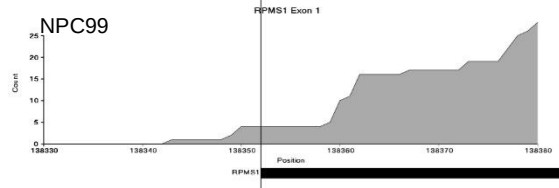
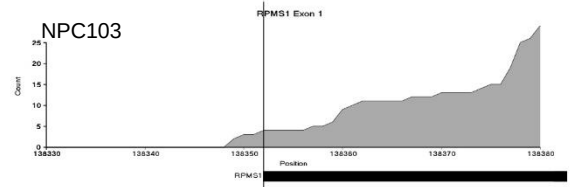
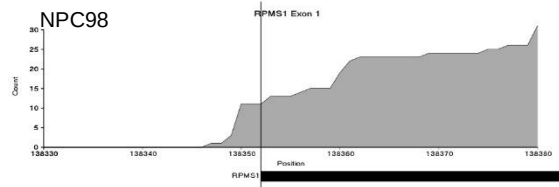
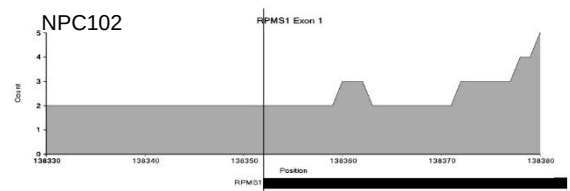
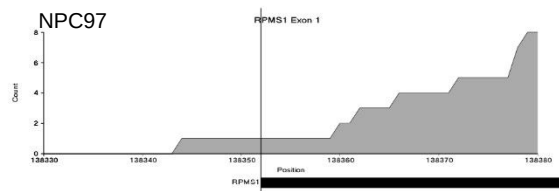




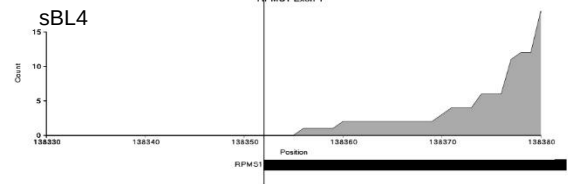
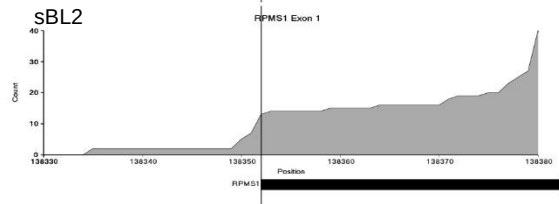
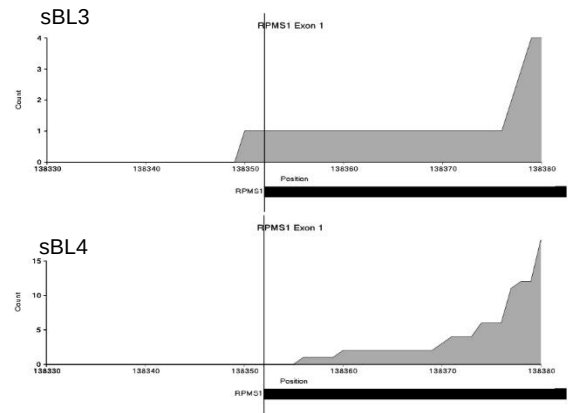
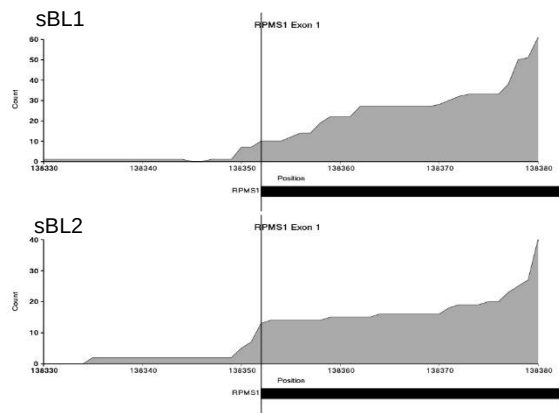




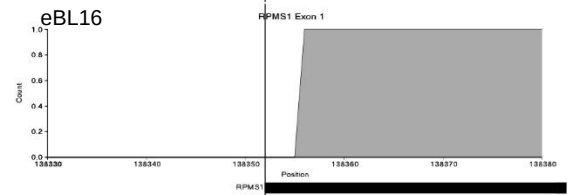
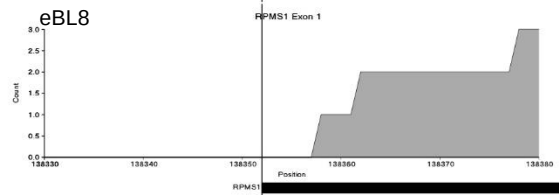
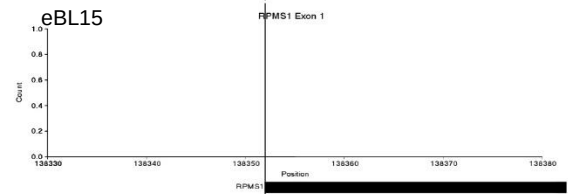
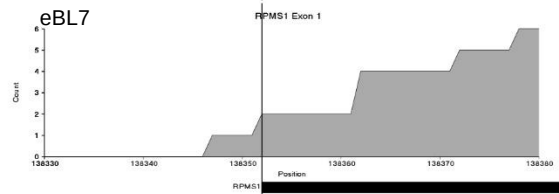
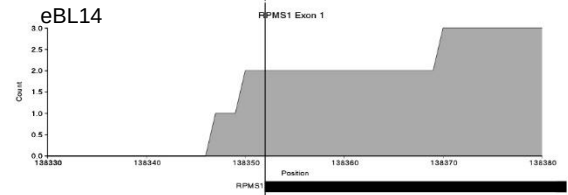
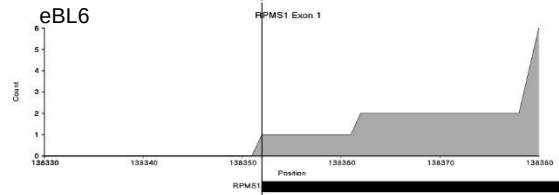
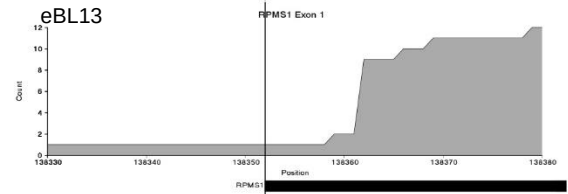
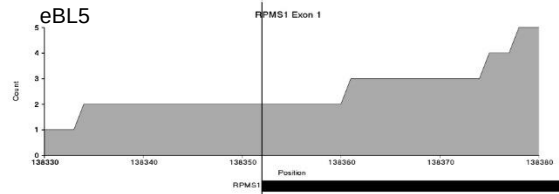
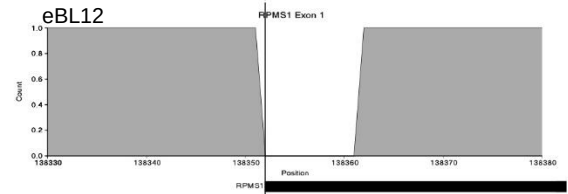
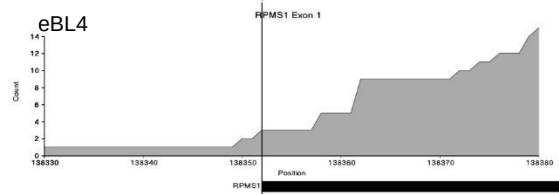
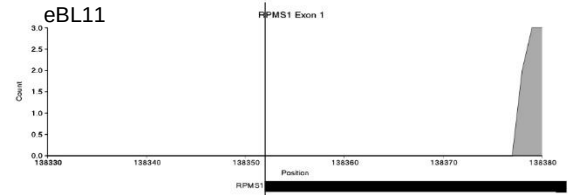
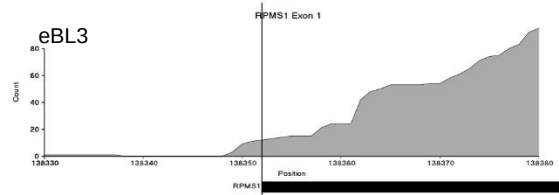
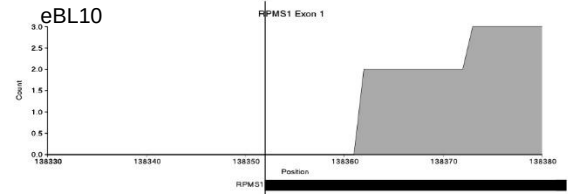
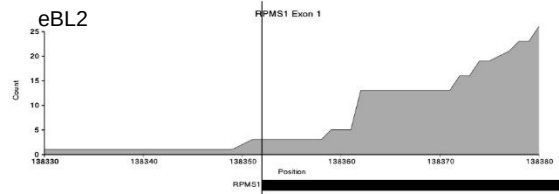
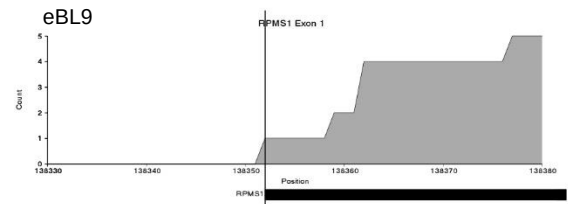
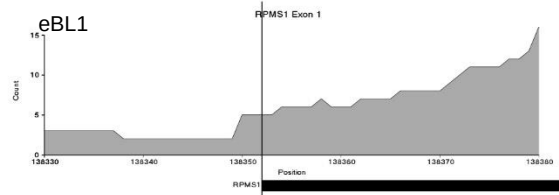




sBL



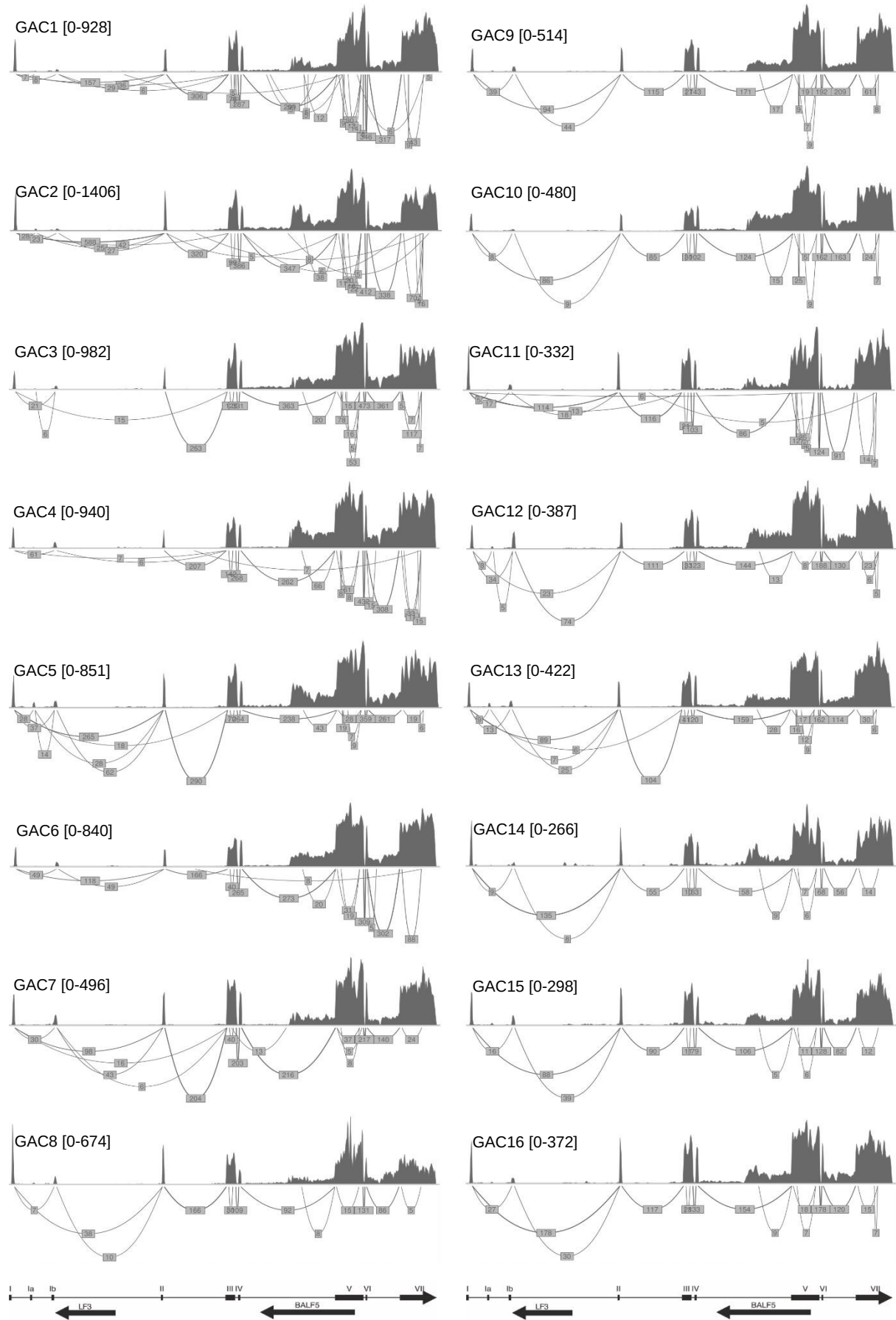
eBL

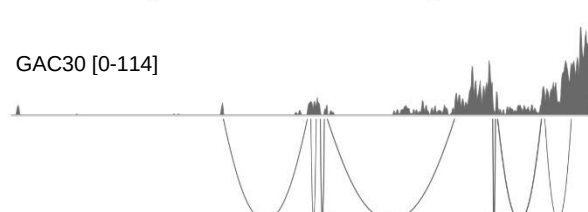
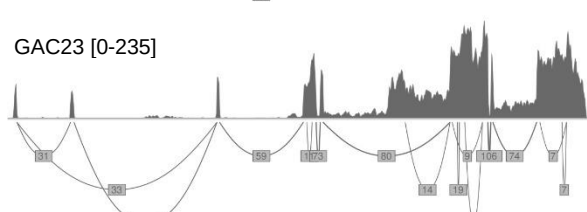
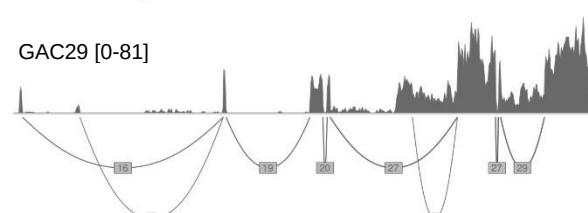
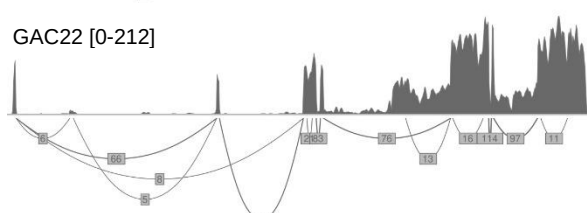
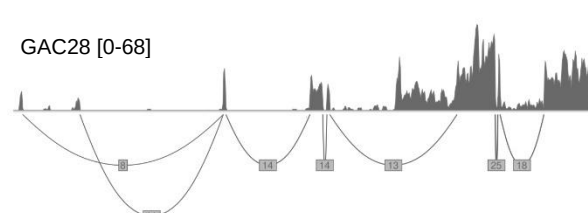
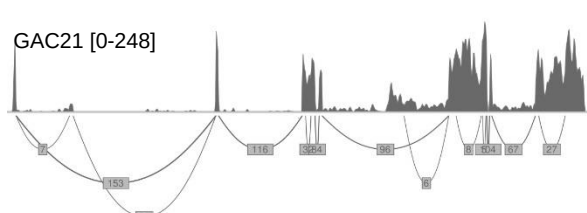
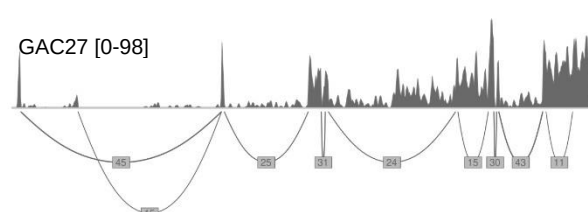
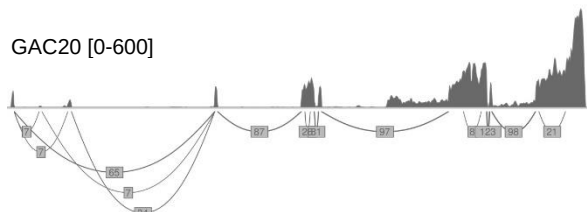
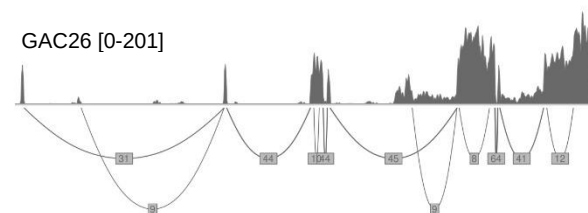
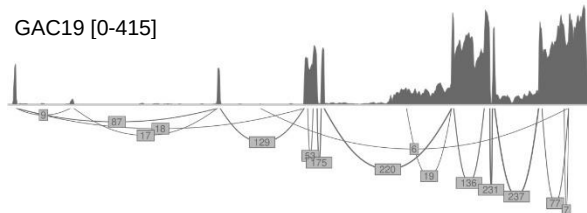
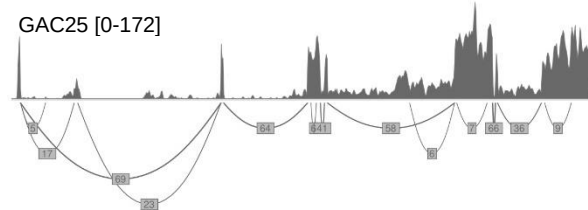
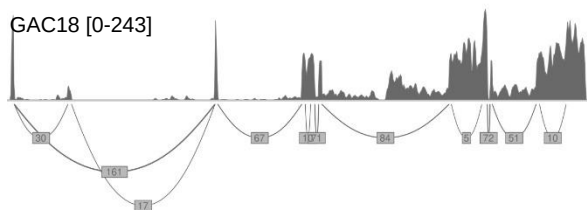
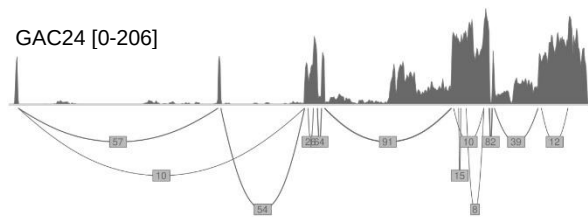
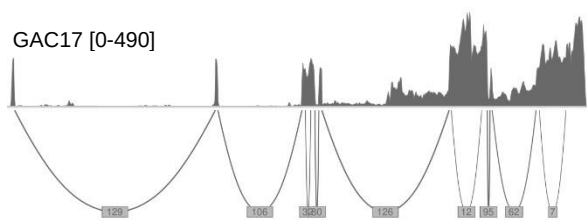


3. BamHI-A/I splice-junctions

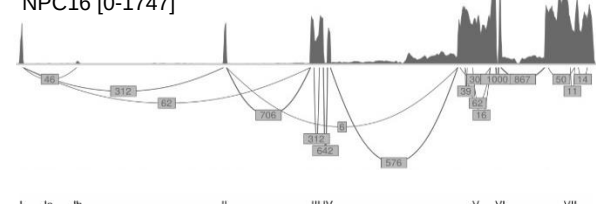
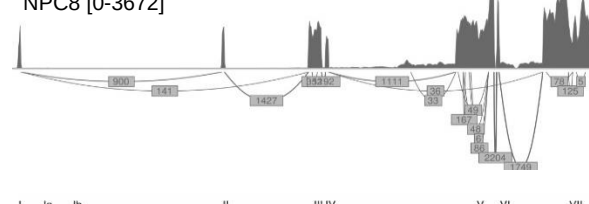
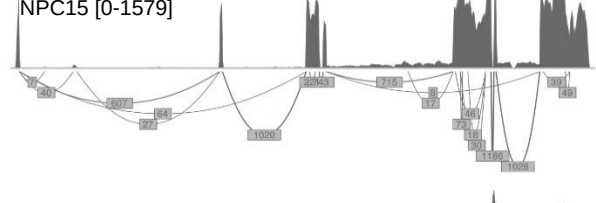
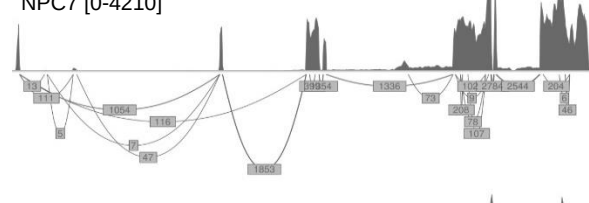
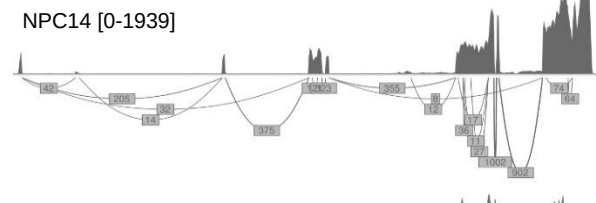
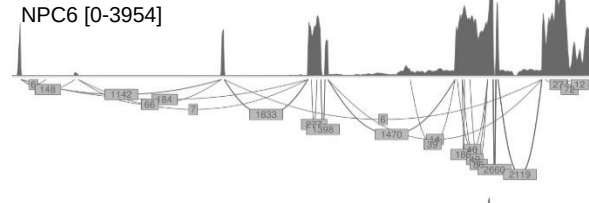
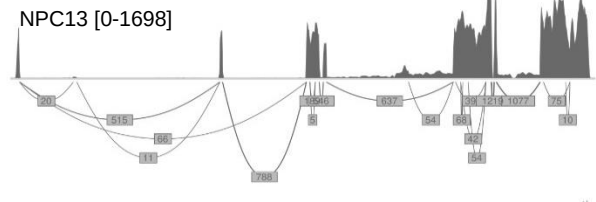
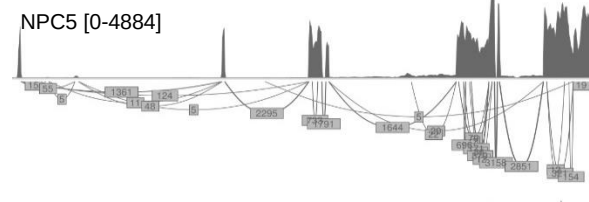
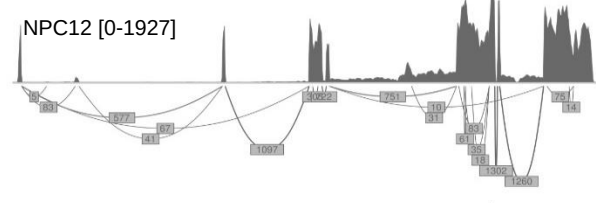
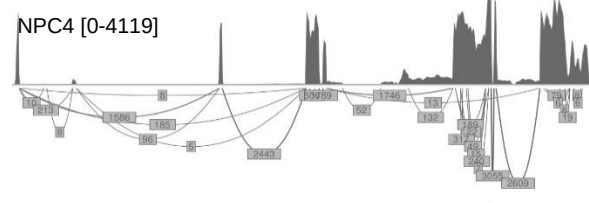
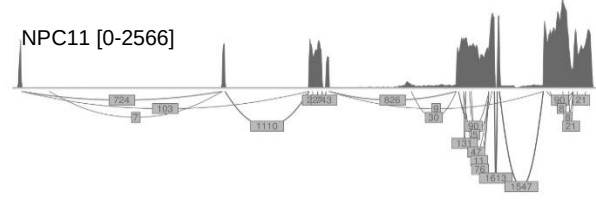
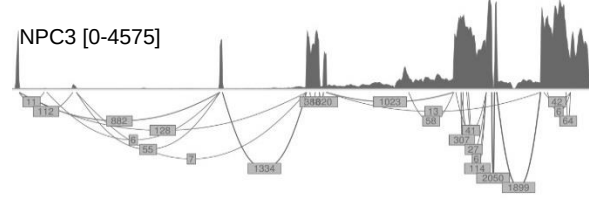
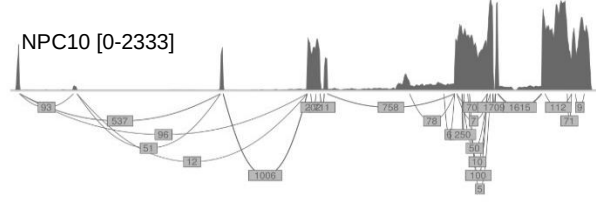
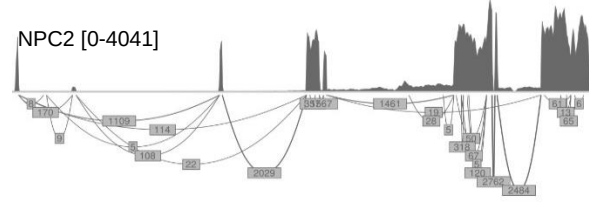
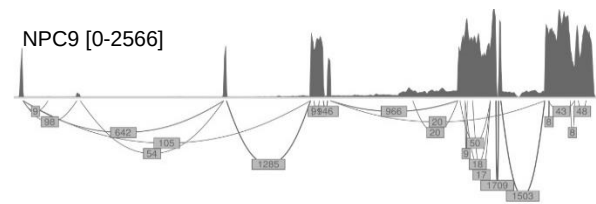
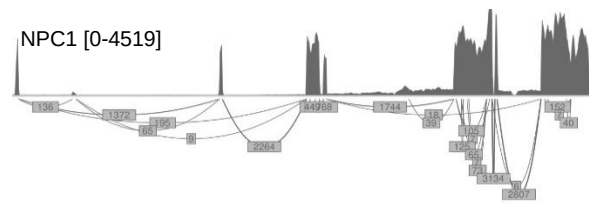
Coverage plot of RNA detected in the BamHI-A/I region of tumors and cell lines. Numbers within brackets indicate the y-axis scale, total number of reads. Connecting lines and numbers below baseline display the location and amount of splice-junction reads detected between exons. Splice-junction reads were abundantly detected within the BamHI-A/I region and showed that the most frequent splicing events occur between the major RPMS1 exons. Note that no splice-junction reads were found between exon I and exon II in LCLs. The LCLs were transformed using the B95-8 virus strain with a deletion in this region. It should be noted that exon II was not included in the reference genome, but splice-junction reads using the exon II splice donor or acceptor were readily detected. Unannotated exon II, accessory splice-sites Ia and Ib and intra-exonic splice-sites within exon III, V and VII were frequently detected and imply a high degree of splicing variants in RPMS1 as previously described¹. Rightward, but not leftward splicing was detected within the unique mid-region of BALF5 was detected. Modified Namalwa cell lines with control plasmid (pTRE), plasmid containing inducible RPMS1 long non-coding RNA (pTRE-RPMS1) and modified RPMS1-promoter with replaced inducible promoter (Pro-re) were treated with indicated amount of doxycycline. Alignment to the three genes (RPMS1, LF3 and BALF5) within regions with non-overlapping sequences is shown at the bottom of each page. GAC, gastric adenocarcinoma; eBL, endemic Burkitt's lymphoma; sBL, sporadic Burkitt's lymphoma; ECL, EBV-associated tumor derived cell lines; LCL, lymphoblastoid cell lines.

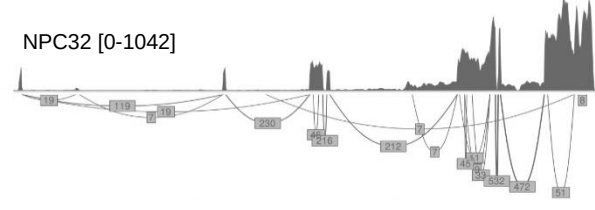
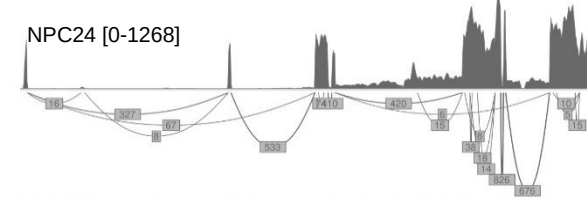
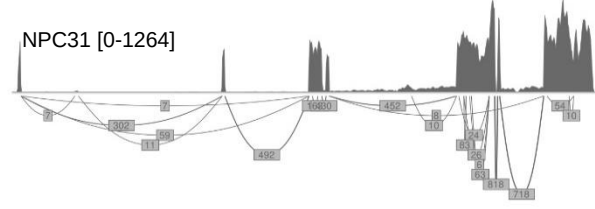
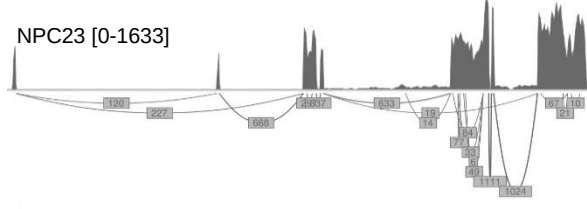
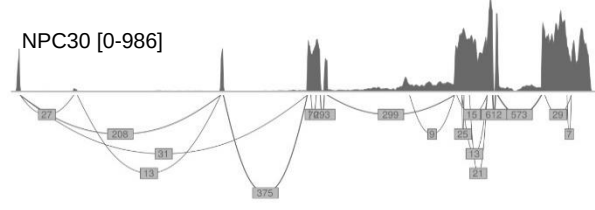
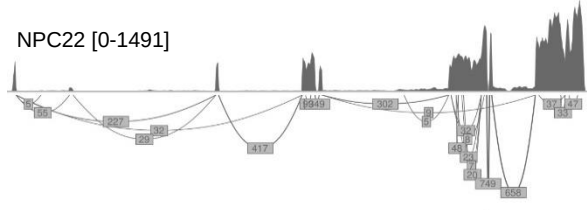
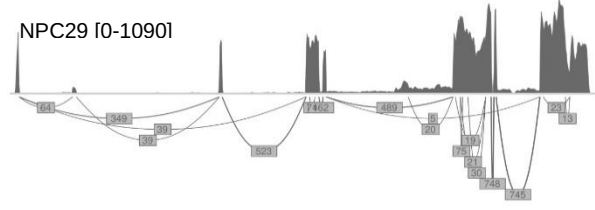
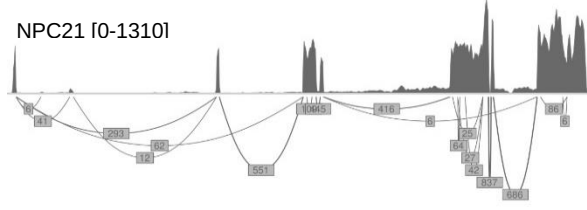
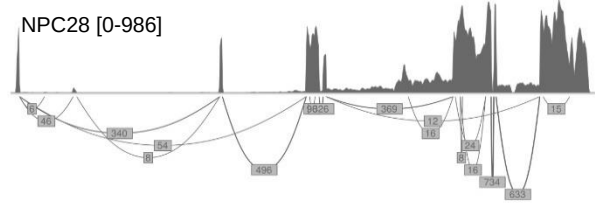
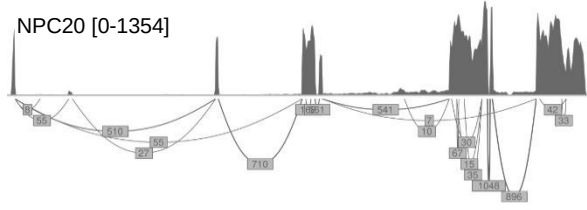
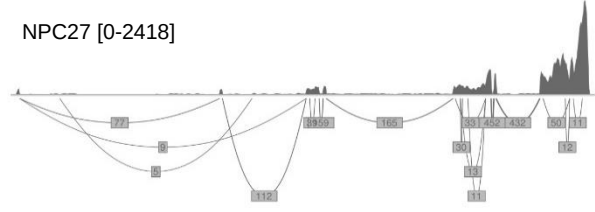
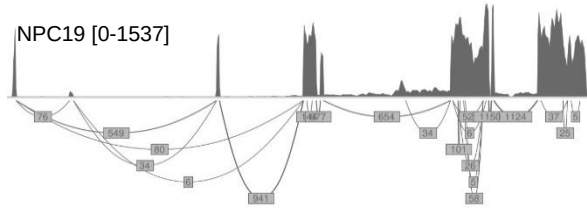
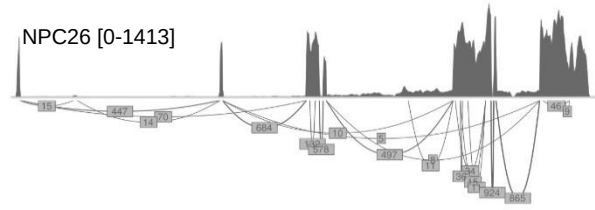
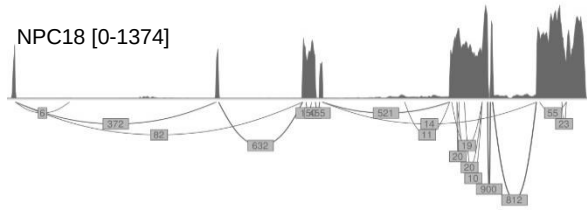
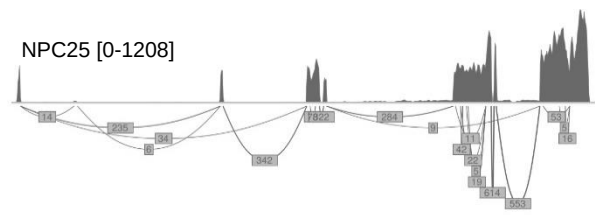
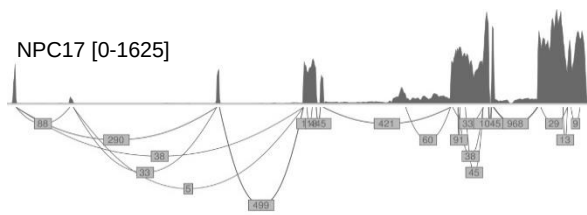
GAC

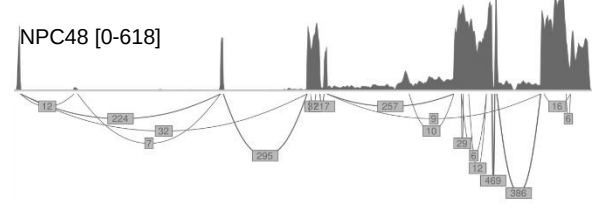
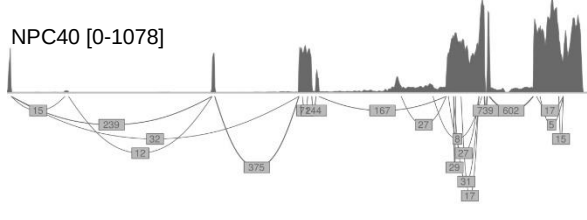
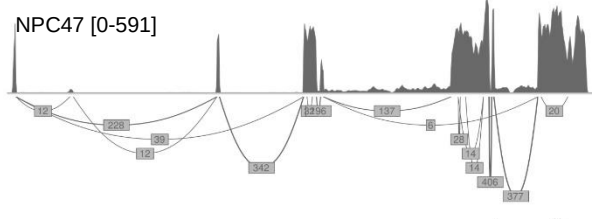
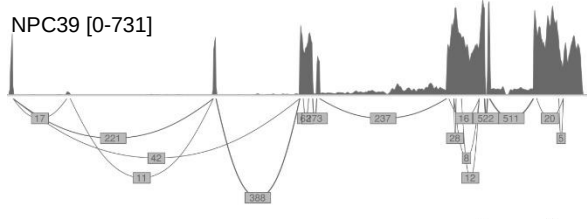
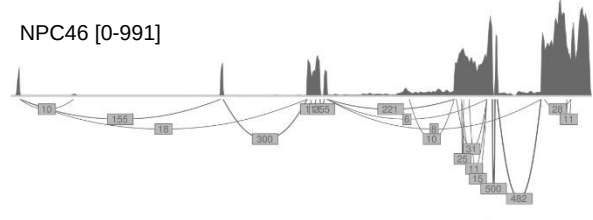
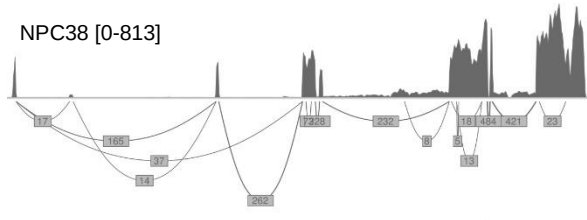
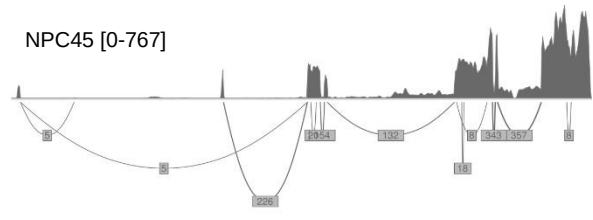
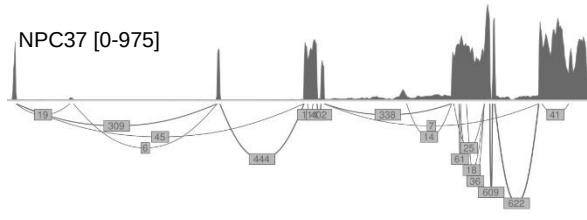
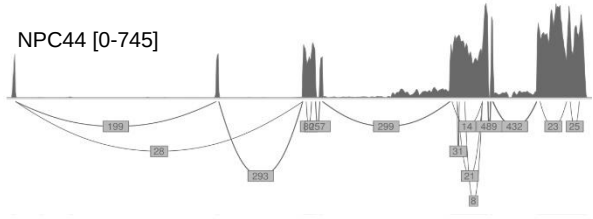
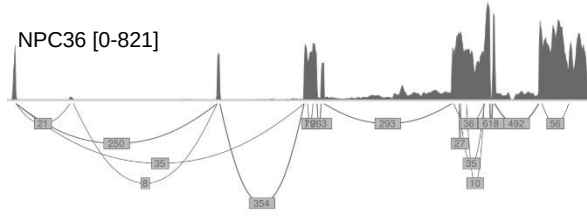
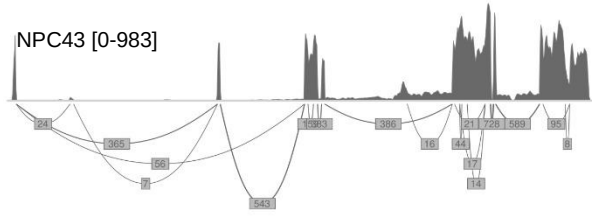
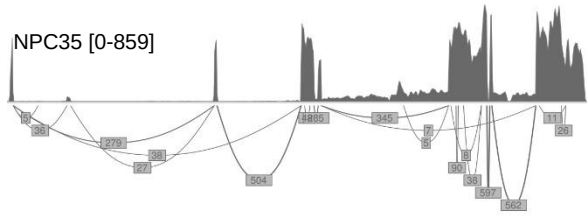
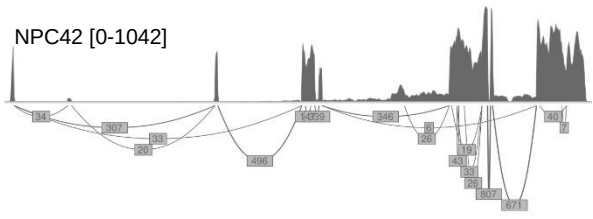
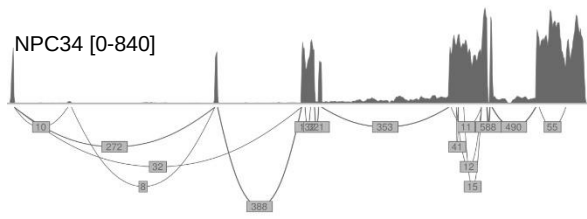
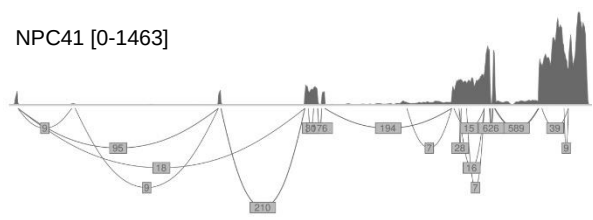
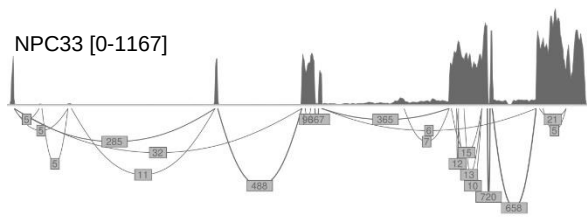


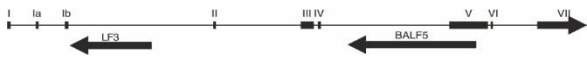
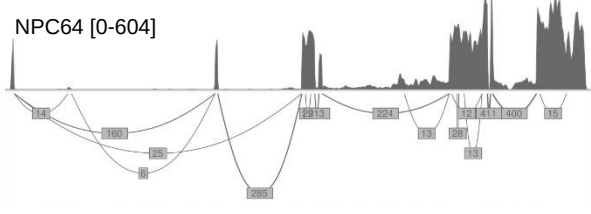
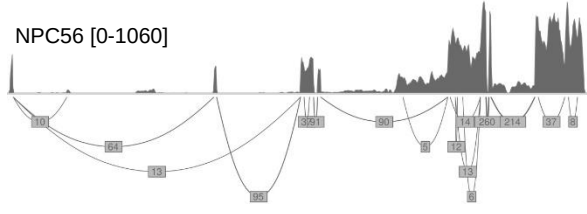
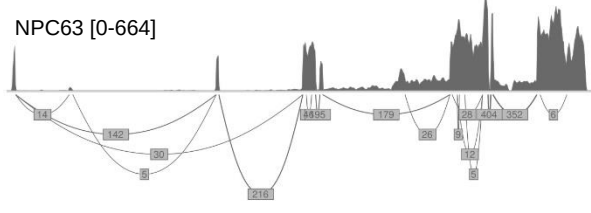
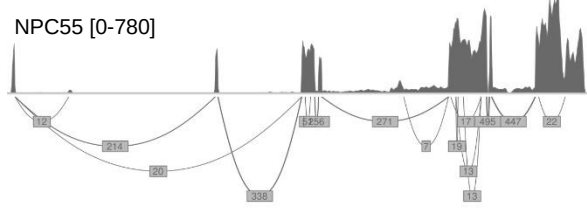
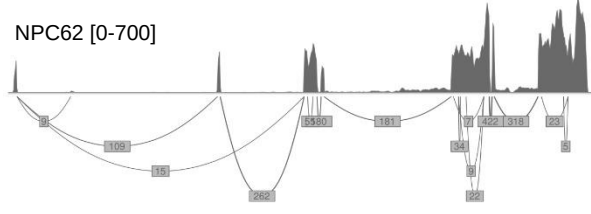
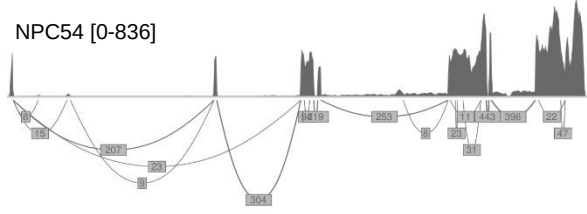
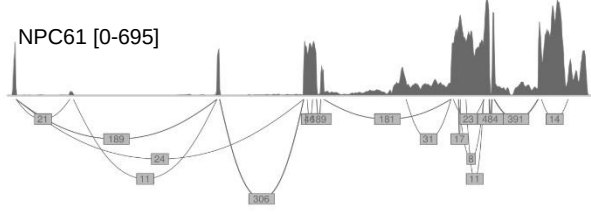
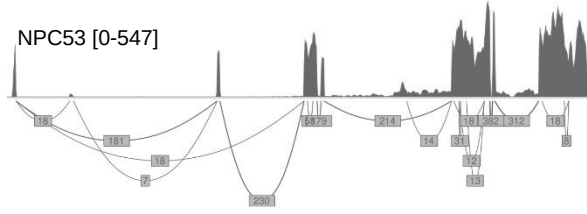
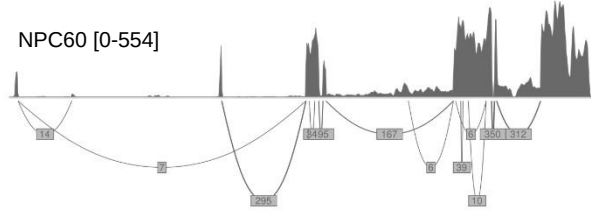
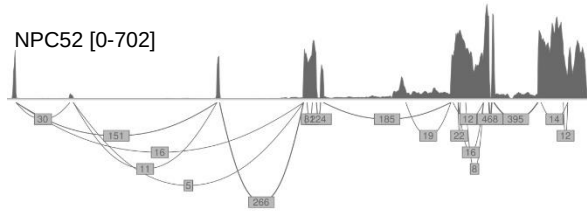
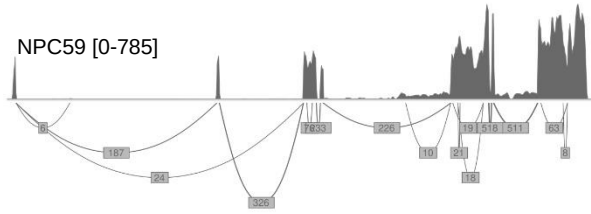
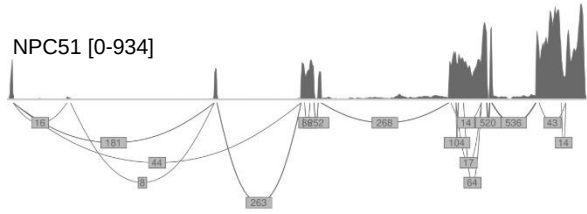
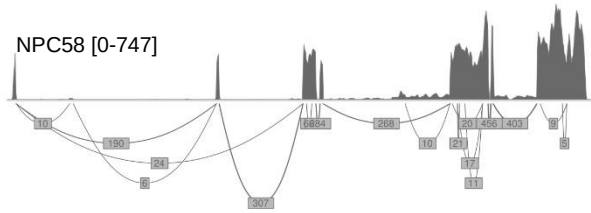
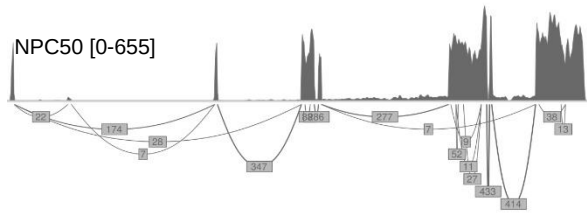
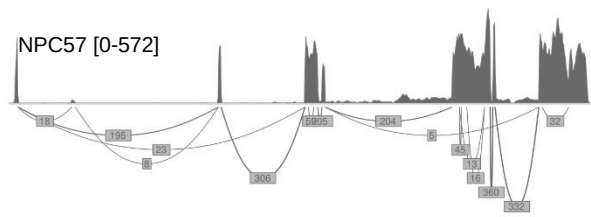
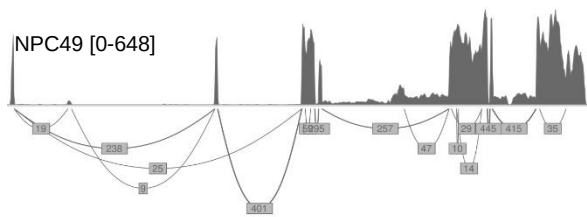


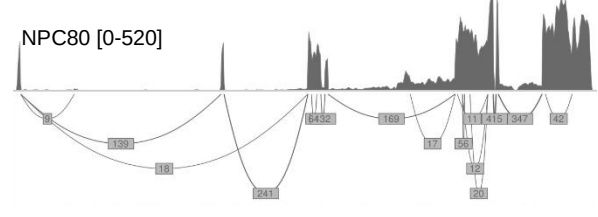
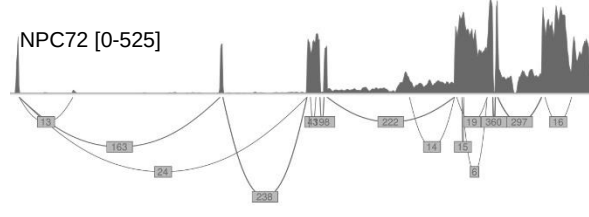
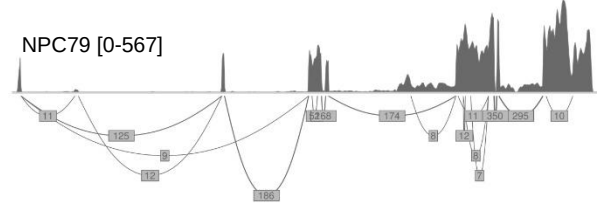
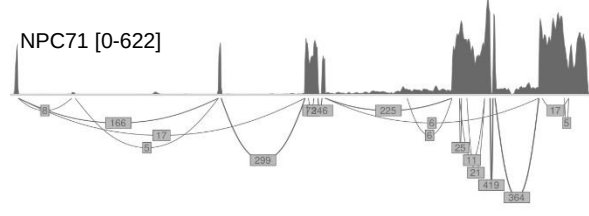
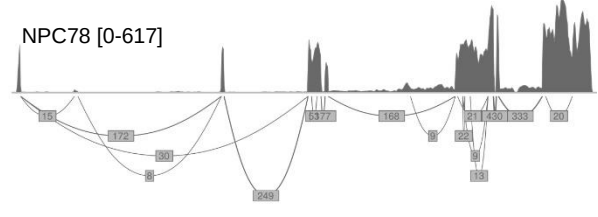
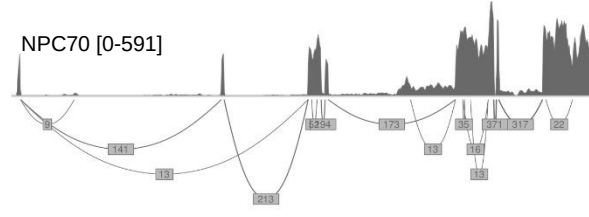
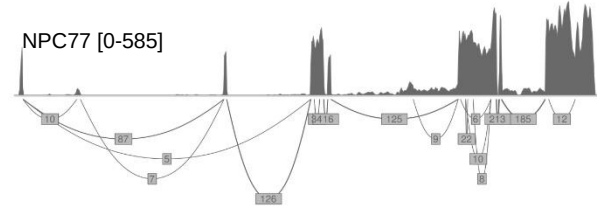
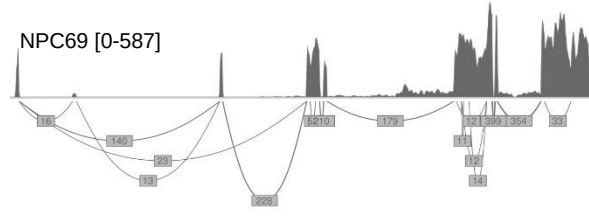
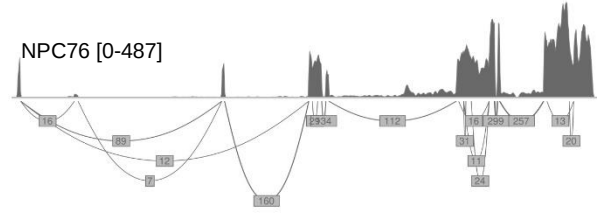
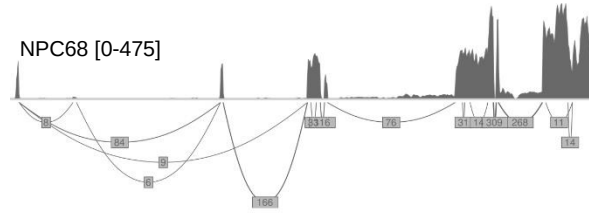
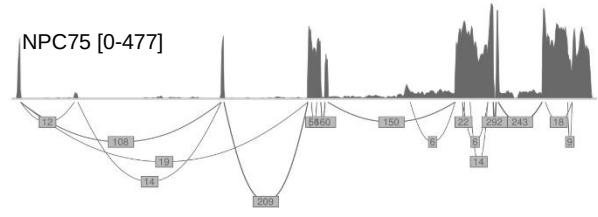
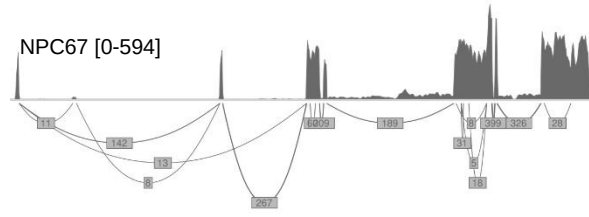
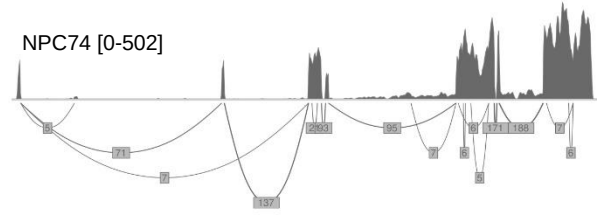
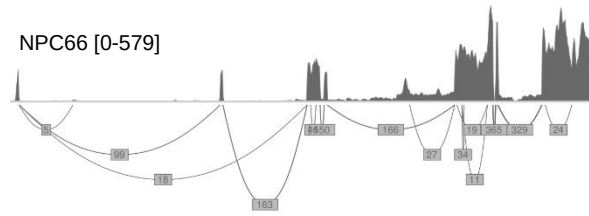
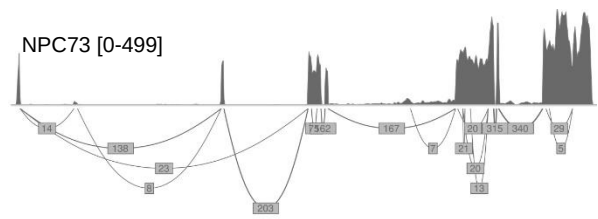
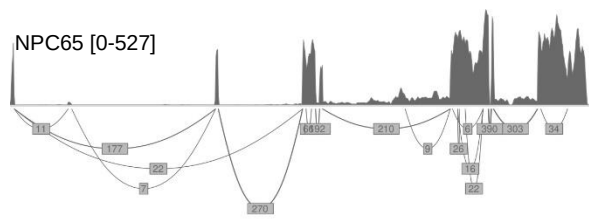
NPC

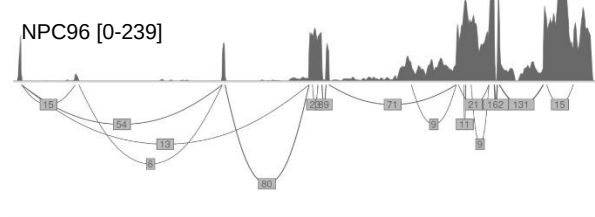
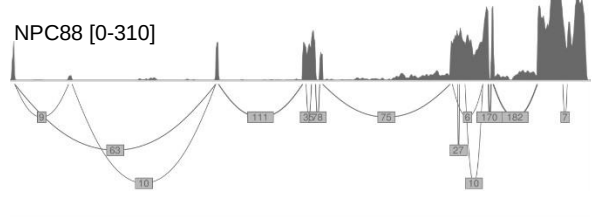
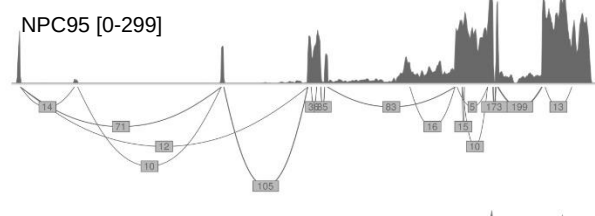
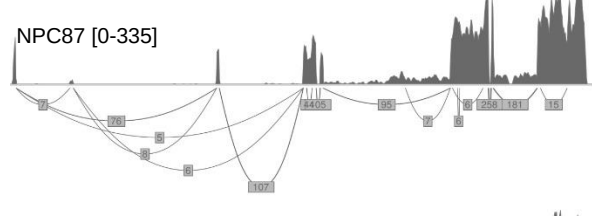
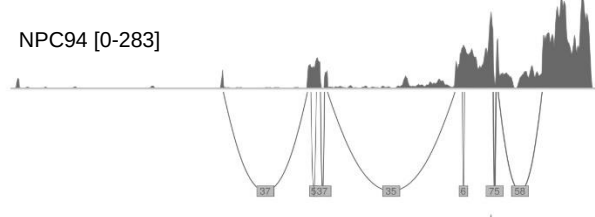
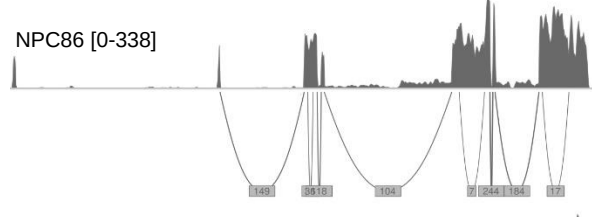
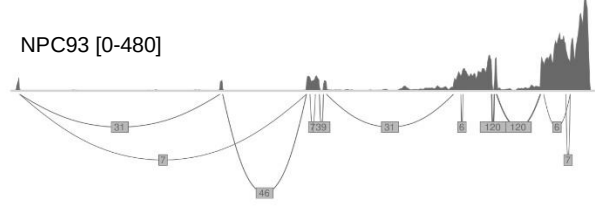
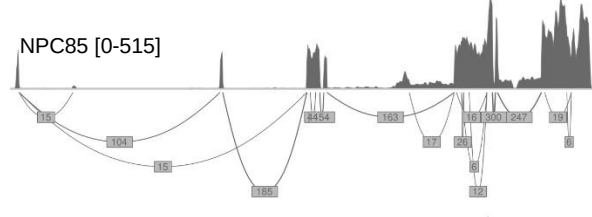
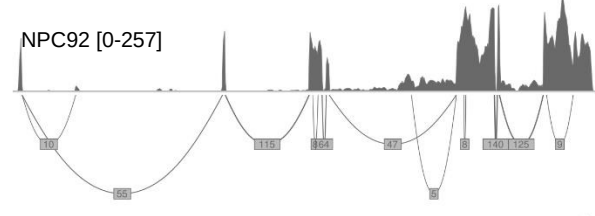
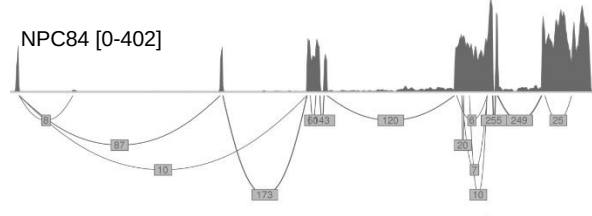
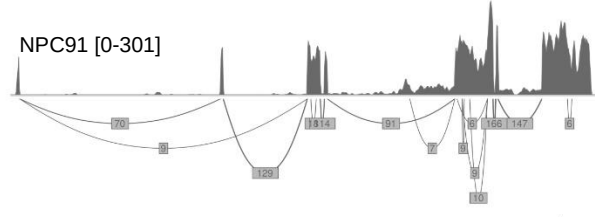
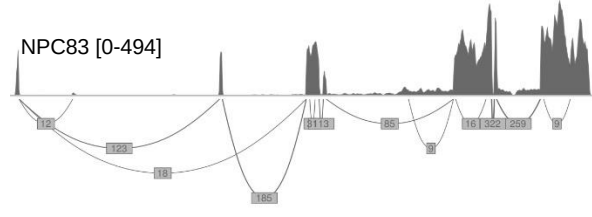
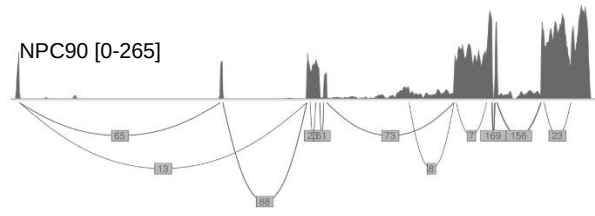
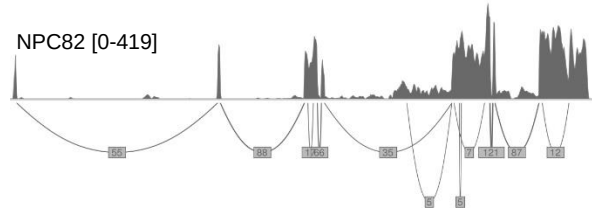
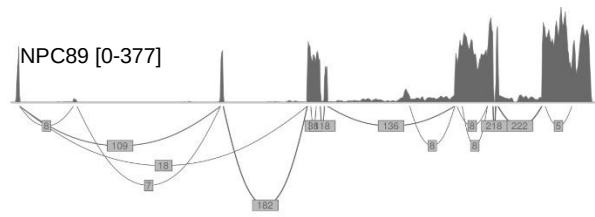
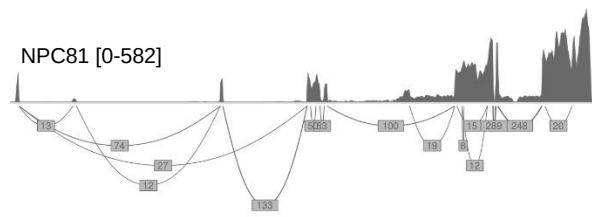


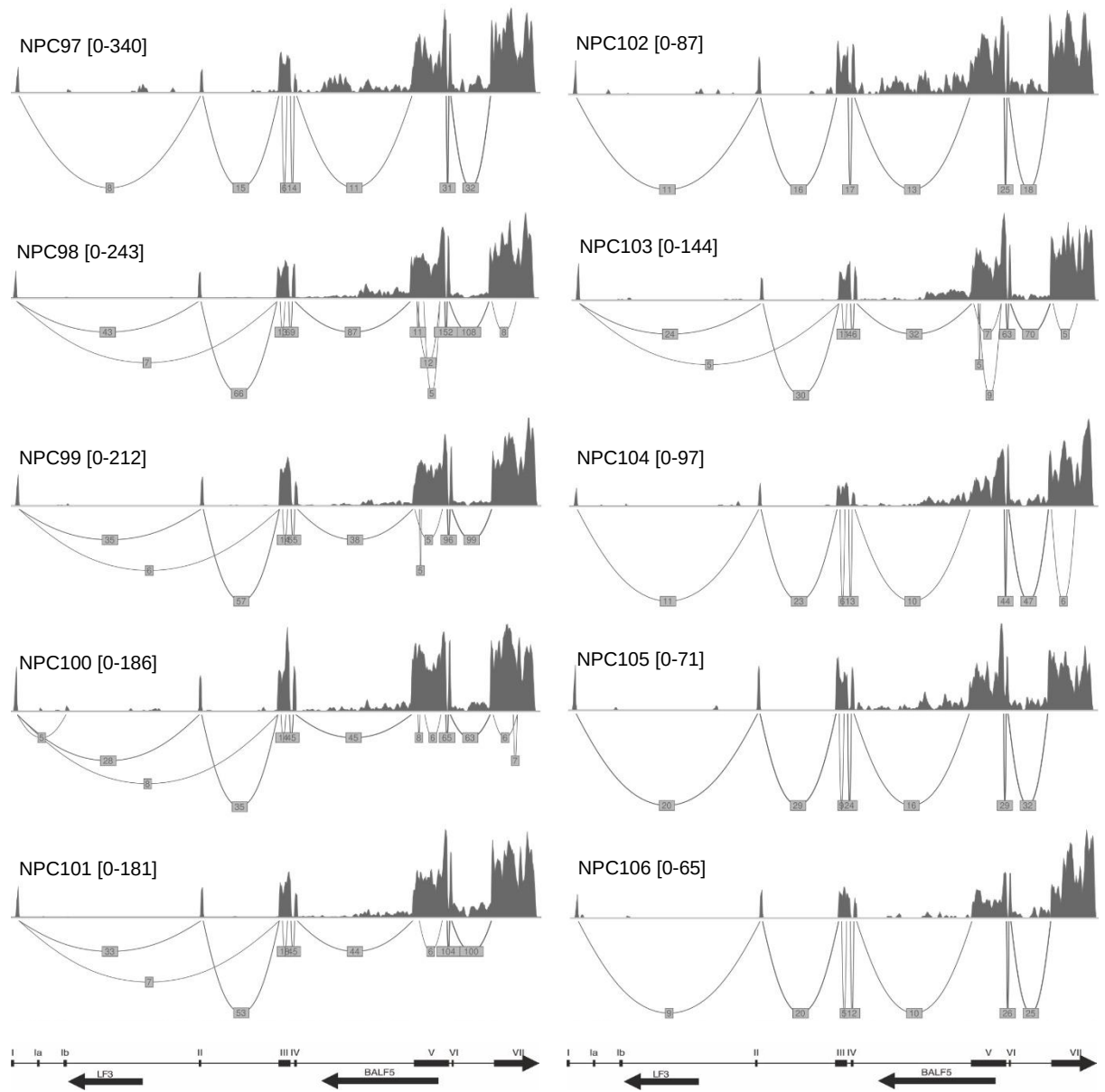




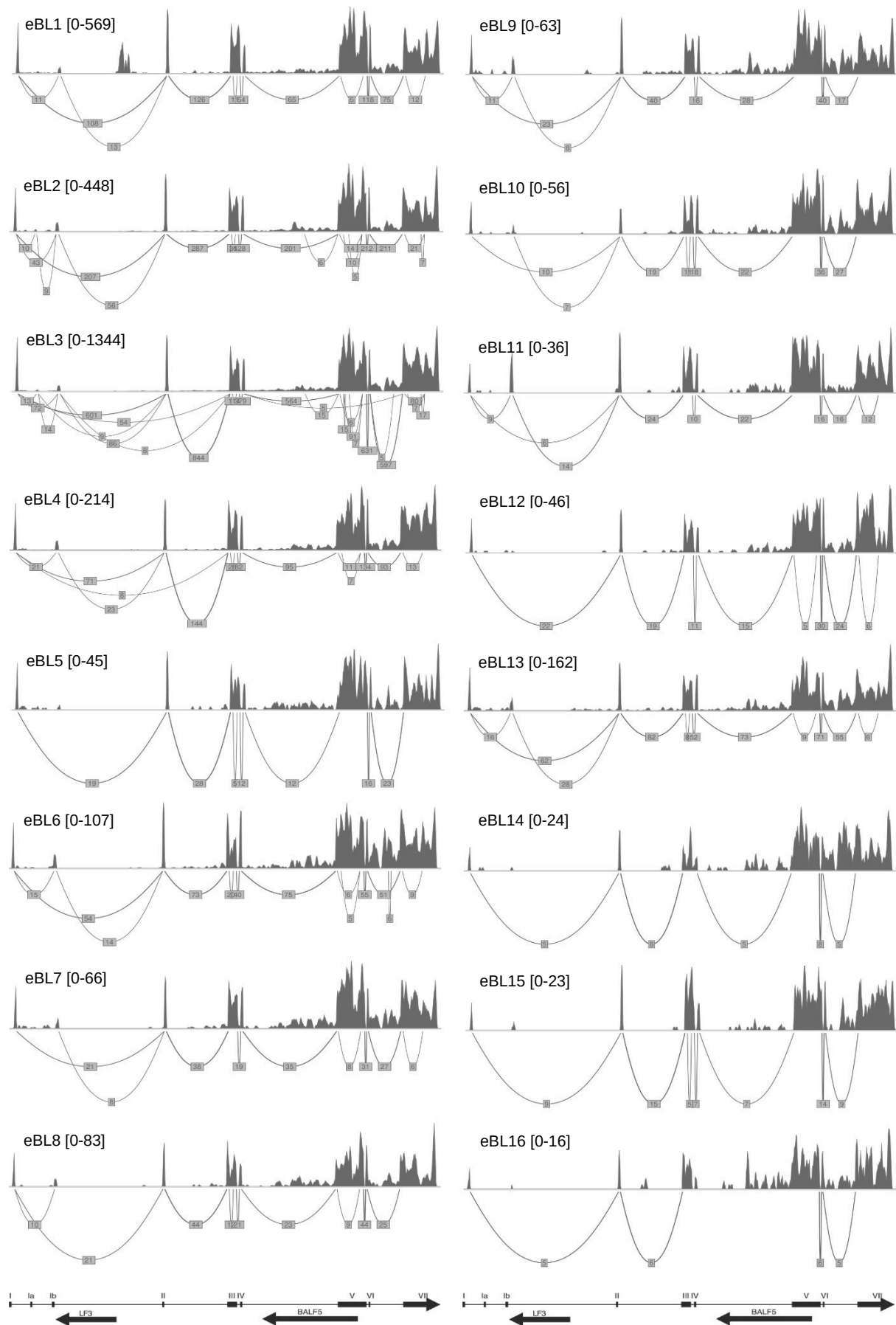






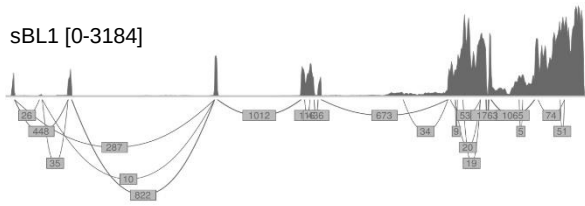


eBL

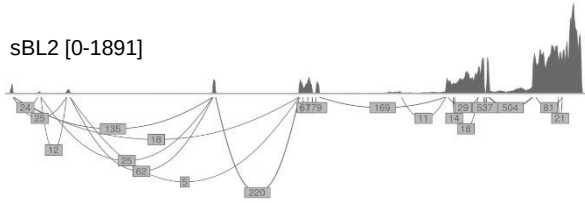


sBL

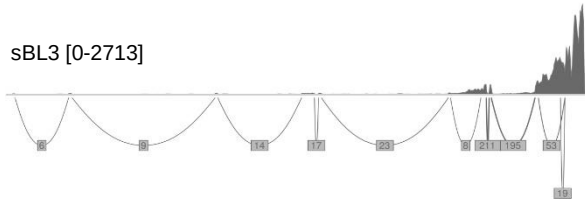
sBL1 [0-3184]



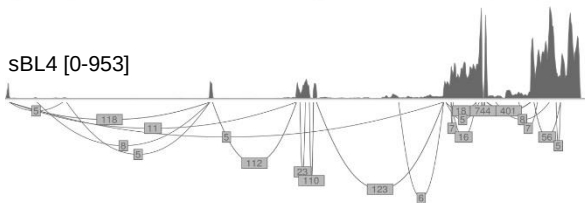
sBL2 [0-1891]



sBL3 [0-2713]

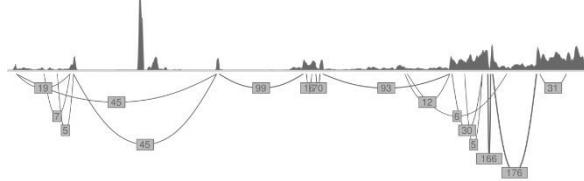


sBL4 [0-953]

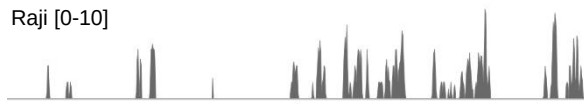


ECL

Daudi [0-1165]



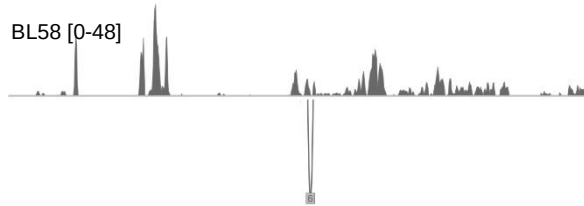
Raji [0-10]



Namalwa [0-53]



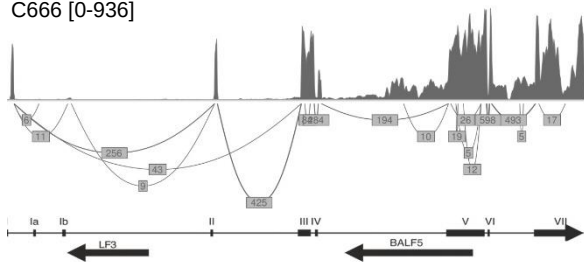
BL58 [0-48]



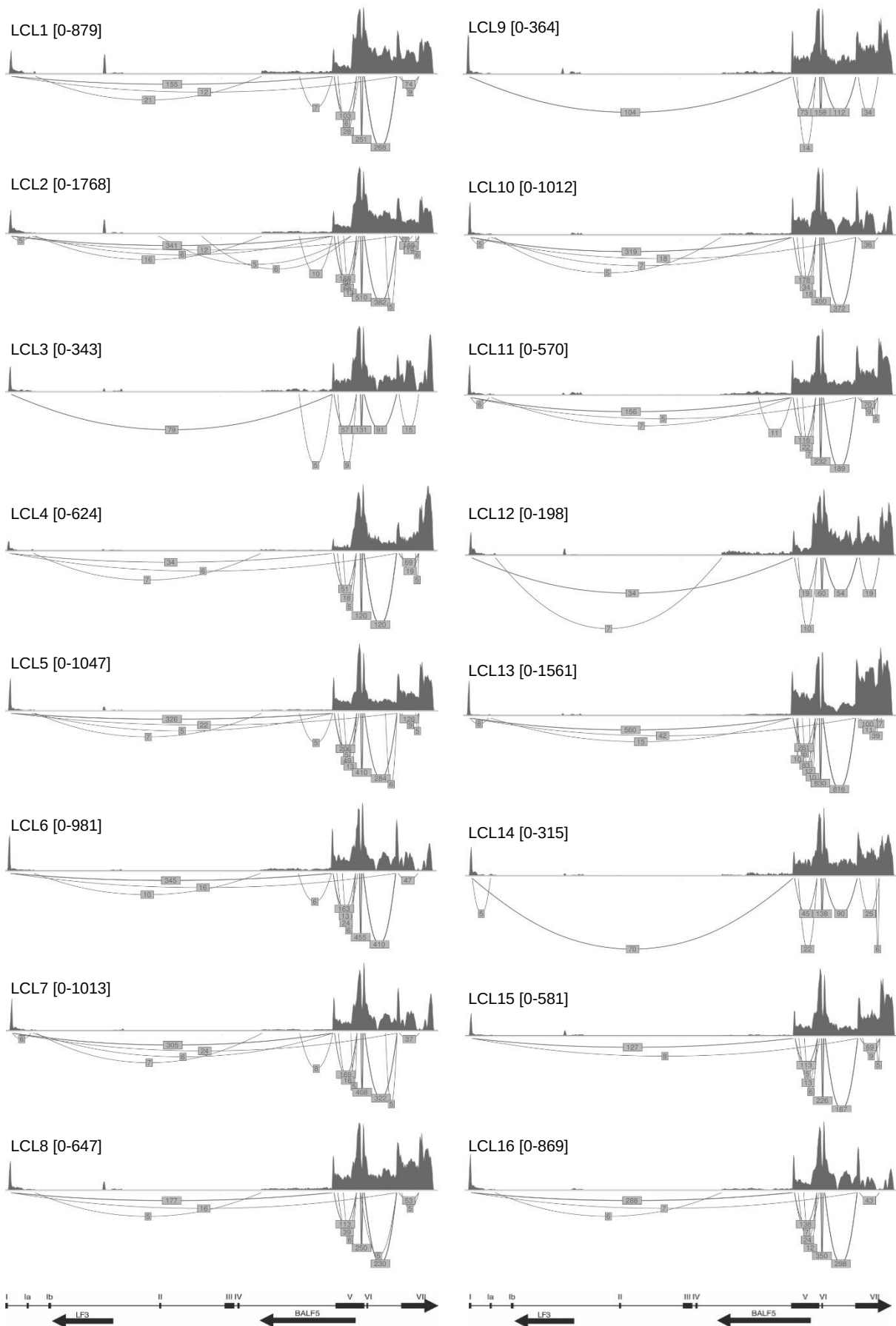
BL41 [0-125]

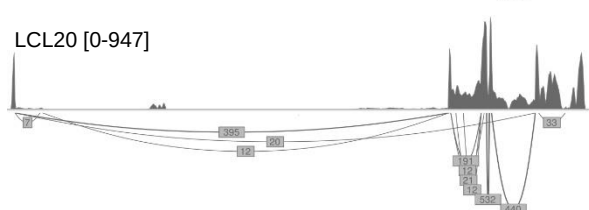
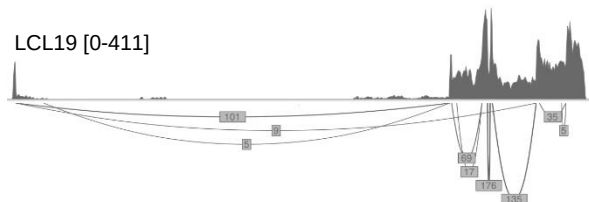
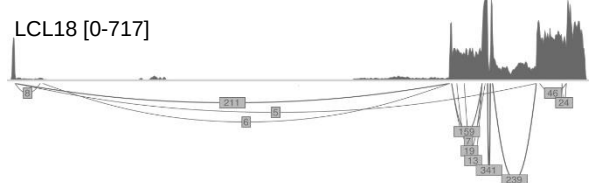
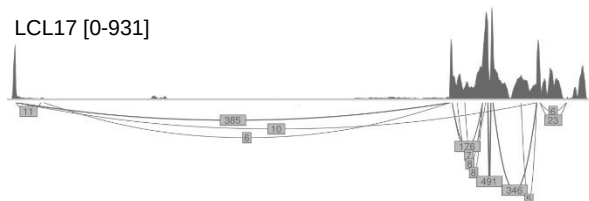


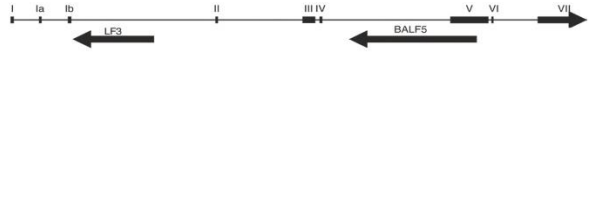
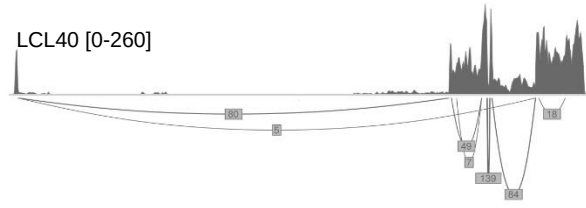
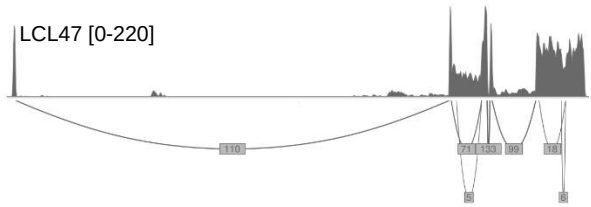
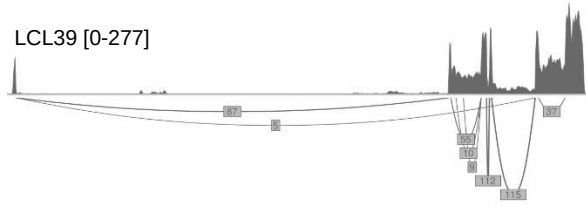
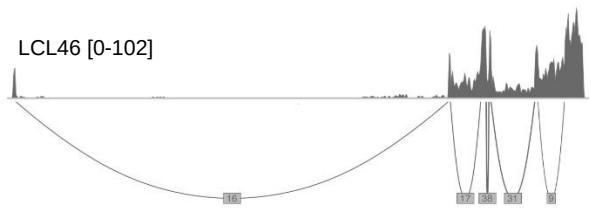
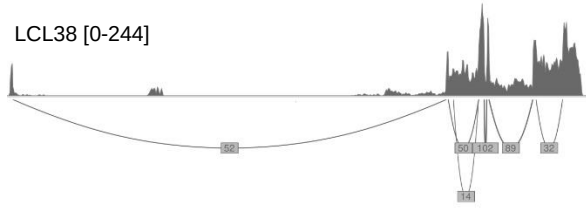
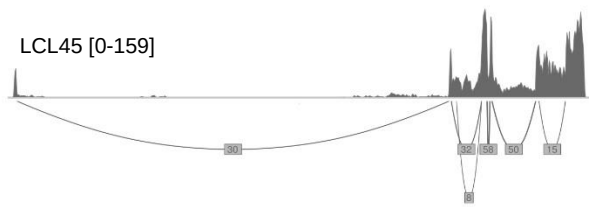
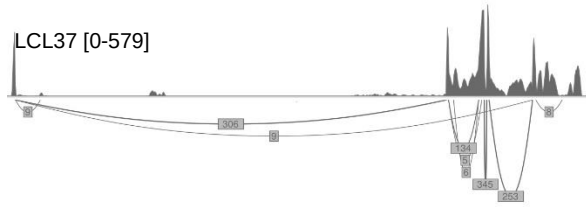
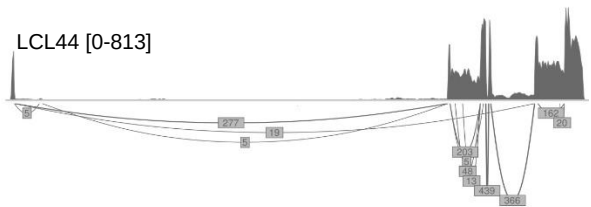
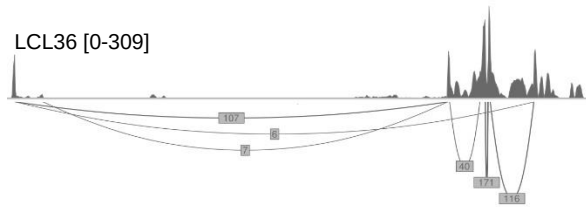
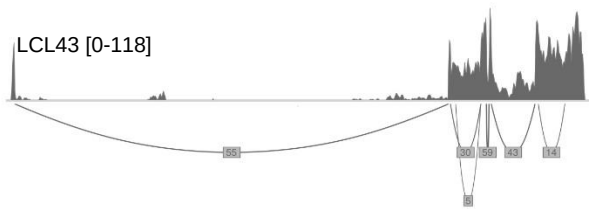
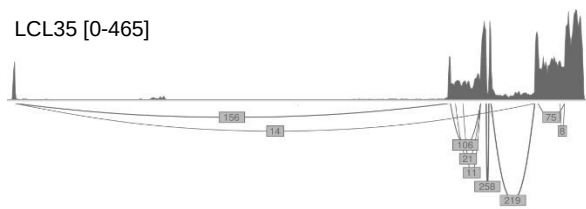
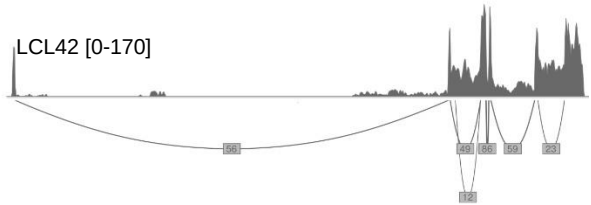
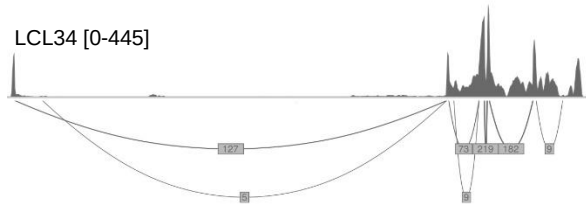
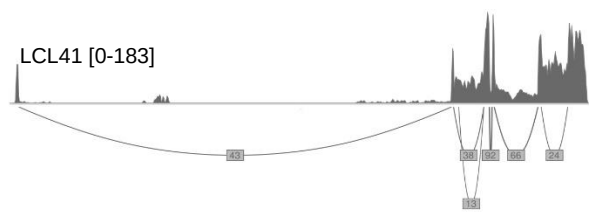
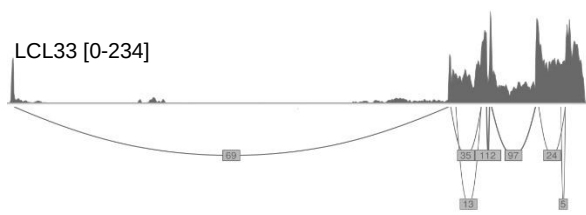
C666 [0-936]



LCL

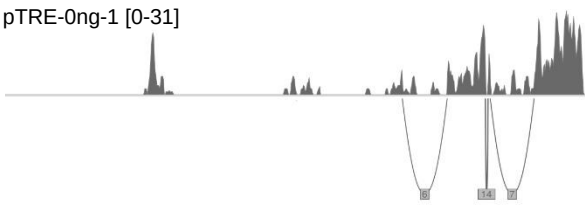




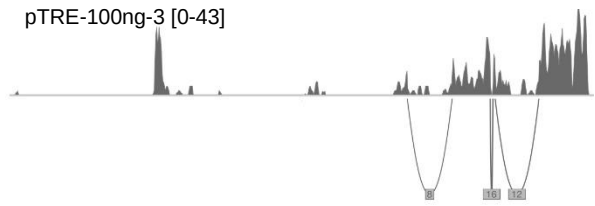


Modified Namalwa cell lines

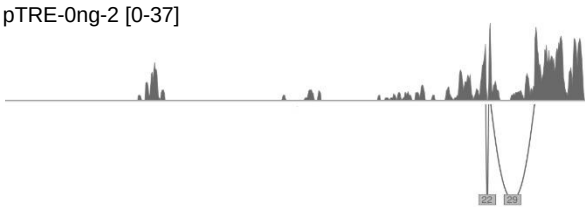
pTRE-0ng-1 [0-31]



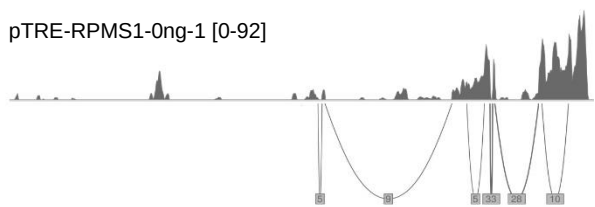
pTRE-100ng-3 [0-43]



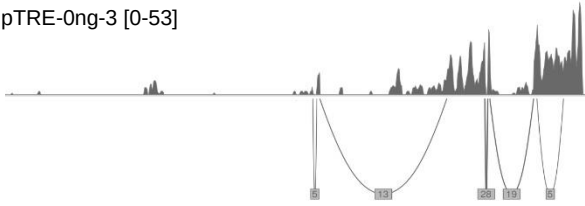
pTRE-0ng-2 [0-37]



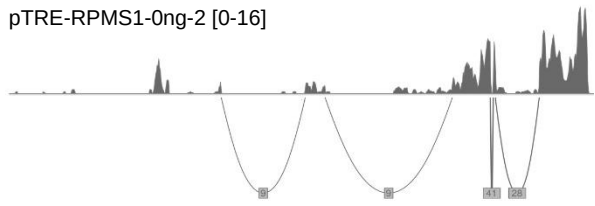
pTRE-RPMS1-0ng-1 [0-92]



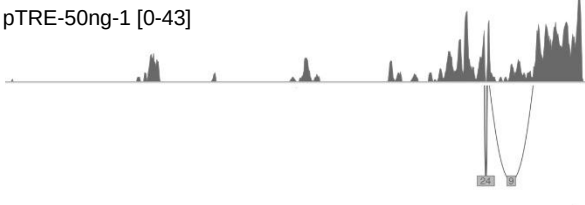
pTRE-0ng-3 [0-53]



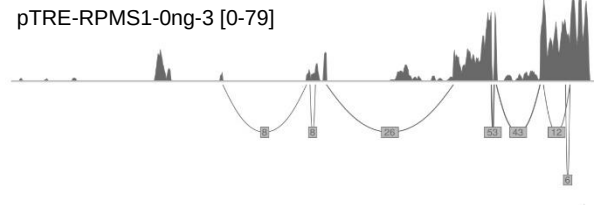
pTRE-RPMS1-0ng-2 [0-16]



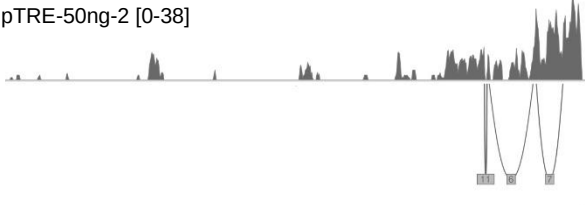
pTRE-50ng-1 [0-43]



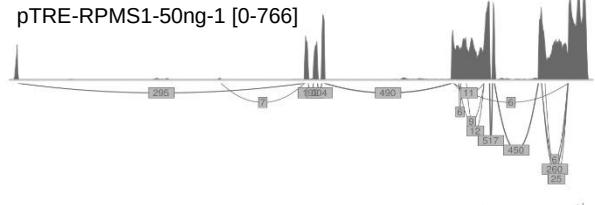
pTRE-RPMS1-0ng-3 [0-79]



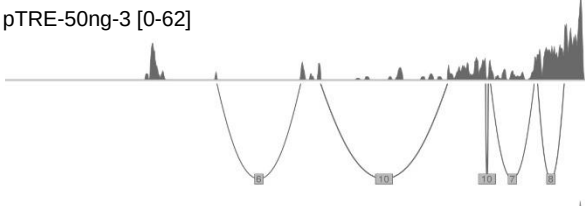
pTRE-50ng-2 [0-38]



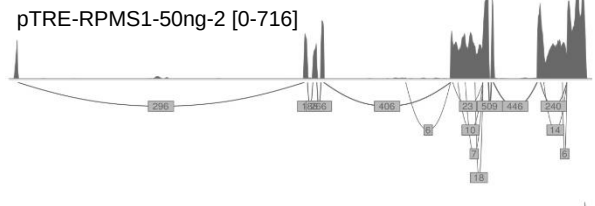
pTRE-RPMS1-50ng-1 [0-766]



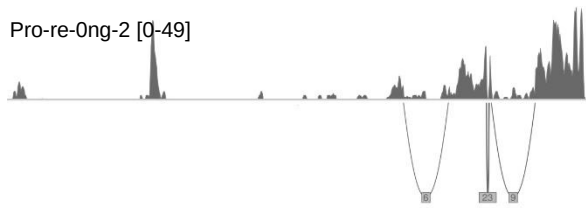
pTRE-50ng-3 [0-62]



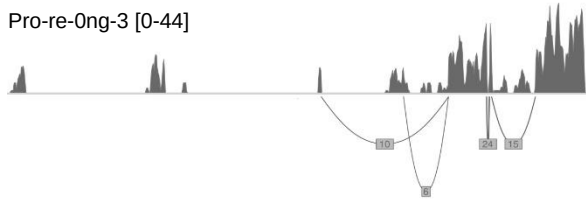
pTRE-RPMS1-50ng-2 [0-716]



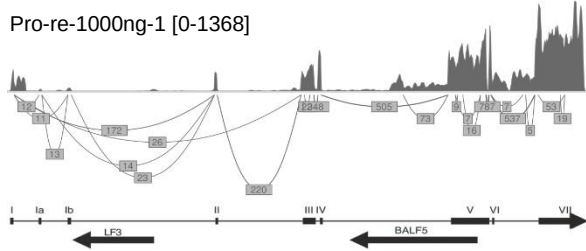
Pro-re-0ng-2 [0-49]



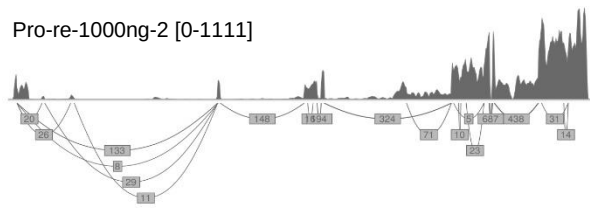
Pro-re-0ng-3 [0-44]



Pro-re-1000ng-1 [0-1368]



Pro-re-1000ng-2 [0-1111]



4. Namalwa modified cell lines

To generate a Tet-On 3G-Expressing stable Namalwa cell line, 1 µg of pCMV-Tet3G plasmid was transfected in 5.6×10^5 Namalwa cells by electroporation (Thermo Neon Transfection System). Positive cells were selected by G418 (800 µg/ml) over two weeks. To create the *RPMS1* promoter-replaced (pro-re) cell line, two sgRNA were designed against the promoter region of the *RPMS1* gene and cloned into the px458 vector. The modified px458 and linear template containing homology arms at both ends to the *RPMS1*-promoter, mCherry and the bidirectional promoter (pTRE3G-BI-mCherry) were co-transfected into the Tet-On 3G-Expressing stable Namalwa cell line by electroporation. Controls containing empty vector pTRE3G-BI-mCherry (pTRE) and pTRE3G-BI-mCherry-RPMS1 (pTRE-RPMS1), containing the exons of the long non-coding RNA were into the Tet-On 3G-Expressing stable Namalwa cell line by electroporation. pTRE, pTRE-RPMS1 and pro-re cells were sorted using red fluorescence five days post-transfection into 96-well plates and cultured for six weeks. The *RPMS1* promoter replaced by an inducible bidirectional promoter encoding mCherry, between the Cas9 target sites was confirmed by whole genome sequencing. Doxycycline-induction of *RPMS1* long non-coding RNA (pTRE-RPMS1) and *RPMS1* gene (pro-re) including miR-BARTs was tested at different concentrations to reach biological equivalent levels. Transcriptomes of induced pTRE-RPMS1 and pro-re cells were compared with pTRE-cells subjected to equivalent concentration of doxycycline.

5. BamHI-A rightward elements

We set out to characterize the unannotated genes corresponding to the peak observed in the unique mid-region of *BALF5*. Retention of the intron between exon IV and V in *RPMS1* has not been detected and would therefore not explain the RNA elements in this region¹. Considering that reads with polyA-stretches were almost exclusively detected in conjunction with the polyA-signal of *RPMS1*, we examined whether this novel transcript, henceforth called *BARE1* (BamHI-A rightward element 1), was co-expressed and co-terminated with *RPMS1*. In the Namalwa cell line, only low levels of RNA from the BamHI-A/I region were expressed (Supplementary Information; EBV RNA). The cell line contains two integrated copies of the EBV genome in human chromosome 1². Using CRISPR/Cas9 we genetically modified a clone by replacing the *RPMS1* promoter with a bidirectional doxycycline inducible promoter and mCherry. Transcriptome data from doxycycline induced cells showed that along with *RPMS1* activation the *BARE1* transcript was induced and the transcriptome pattern displayed resemblance with the EBV expression in the neoplasms, which argues that the transcripts are co-expressed (Supplementary Information; BamHI-A/I splice-junctions; promoter replaced, pro-re). To further confirm the *BARE1* transcripts we adapted a single-molecule sequencing method to characterize the splicing variants of each transcript. The cell line C666-1 which contained EBV expression with highest similarity to primary tumors was sequenced using nanopore sequencing (Extended Data Fig. 2c-e)¹. Almost all *BARE1* transcripts contained a splice junction at position 153,528 (in NC_007605.1) connected to the splice acceptor of exon V as previously observed in the splice-junction analysis. We can therefore conclude that *BARE1* is a rightward transcript with its own promoter and is initiated within intron IV of *RPMS1* and co-terminates with *RPMS1*. We continued to characterize the elements downstream of the first exon of *BARE1*. The following element, henceforth called *BARE2*, utilized the same splice donors and acceptors as *RPMS1* and *BARE1*. Finally, the RNA element occupying the intron between exon VI and VII, henceforth called *BARE3*, followed the same splicing pattern in terms of usage of alternative exons within exon VII. As observed in the peak analysis, *BAREs* were co-expressed with *RPMS1*, especially in GAC compared with the other neoplasms (Figure 2g, Extended Data Fig. 2a,b). We thus conclude that the major EBV-transcripts in primary neoplasms are the rightward transcripts *RPMS1* and *BAREs*.

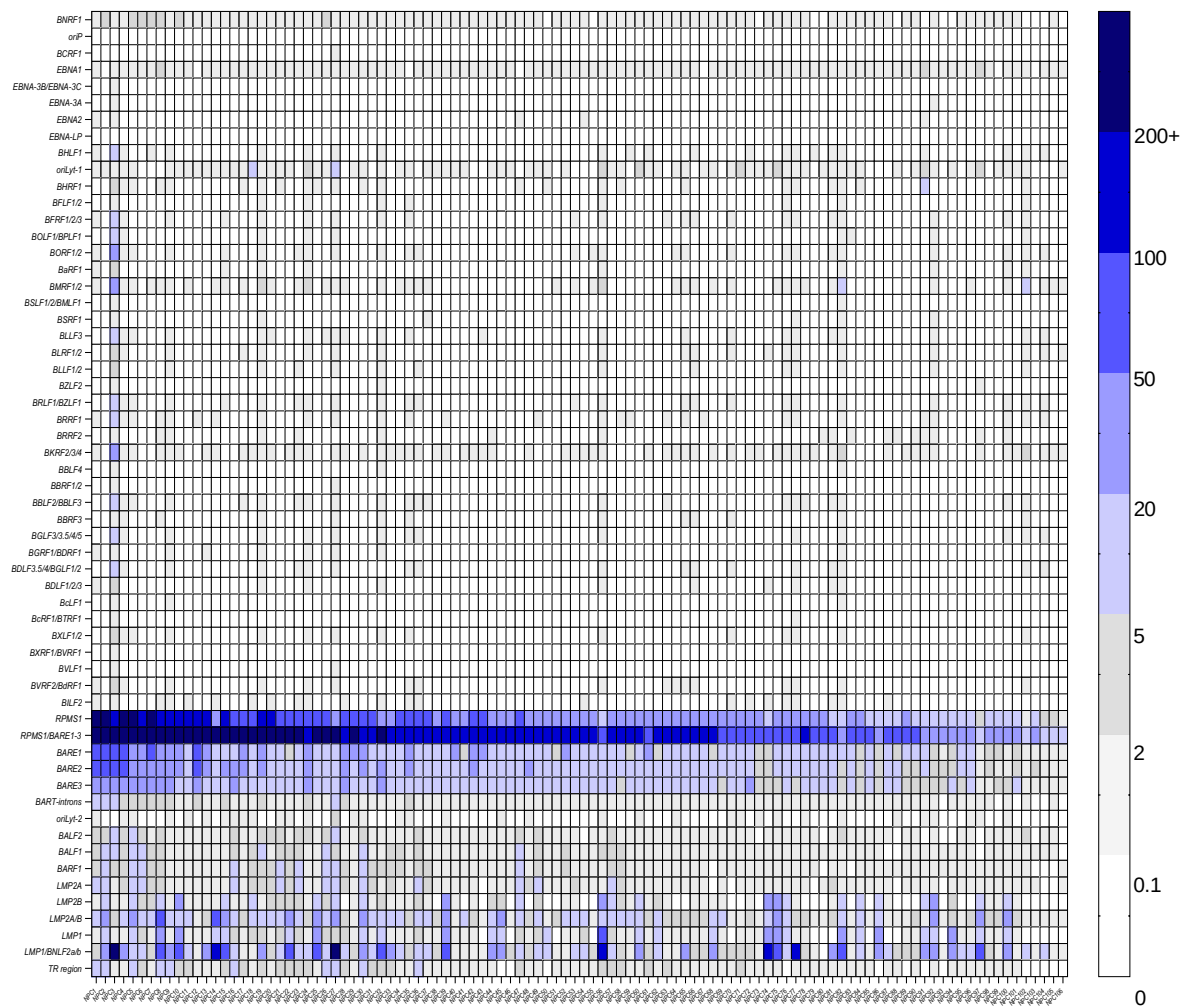
6. Tpm-threshold

Setting a threshold for which EBV genes to consider is dependent on a multitude of factors including tissue heterogeneity, viral reactivation, RNA quality, and bioinformatic processing. We applied experimental and bioinformatic analyses to guide us in setting the threshold. To calculate the number of *RPMS1/BAREs* transcripts the tpm-value corresponds to, we subjected Namalwa cells to *RPMS1* RNA fluorescence *in situ* hybridization (RNA-FISH) and RT-qPCR, and compared it with the transcriptome values (Extended Data Fig. 3a,b). The Namalwa cells had a *RPMS1/BAREs* tpm-value of 12 (Figure 2e, Extended Data Table 4; ECL). The RNA-FISH showed that the range of RNA transcripts varied between 0 and 26 copies, on average 7 transcripts per cell. Similarly, the RT-qPCR quantified the average number of *RPMS1/BAREs* molecules to 2 transcripts per cell. However, this estimation is only valid for these genes in this B-lymphocyte cell line with its low total mRNA content and homogeneity, compared with the heterogeneous primary neoplasms. Analysis of commonly used experimental and bioinformatic house-keeping genes in EBV-positive samples showed RNA encoding ribosomal proteins with average values ranging from 129-8,141 tpm and *HPRT1* with an average of 79 tpm (Extended Data Fig. 3c, Extended Data Table 3)³. Bioinformatic analysis of the distribution of tpm-values for multiple EBV genes in the different neoplasms showed a distribution into three groups, low (5-20 tpm), intermediate (20-50 tpm), and high (>50 tpm) expression signals (Extended Data Fig. 4). The majority of neoplasms have been subjected to deep sequencing. This makes it unlikely that even transcripts with low expression have been omitted in the analysis. Contrarily, rare viral transcripts in the cancer cells are detected, but cannot be discerned as they are diluted, for example by cells in the tissue undergoing viral reactivation. Therefore, in order to identify EBV-transcripts expressed at low levels in cancer cells we had to use scRNA-Seq datasets in which the signal from surrounding infected cells and stromal cells can be removed.

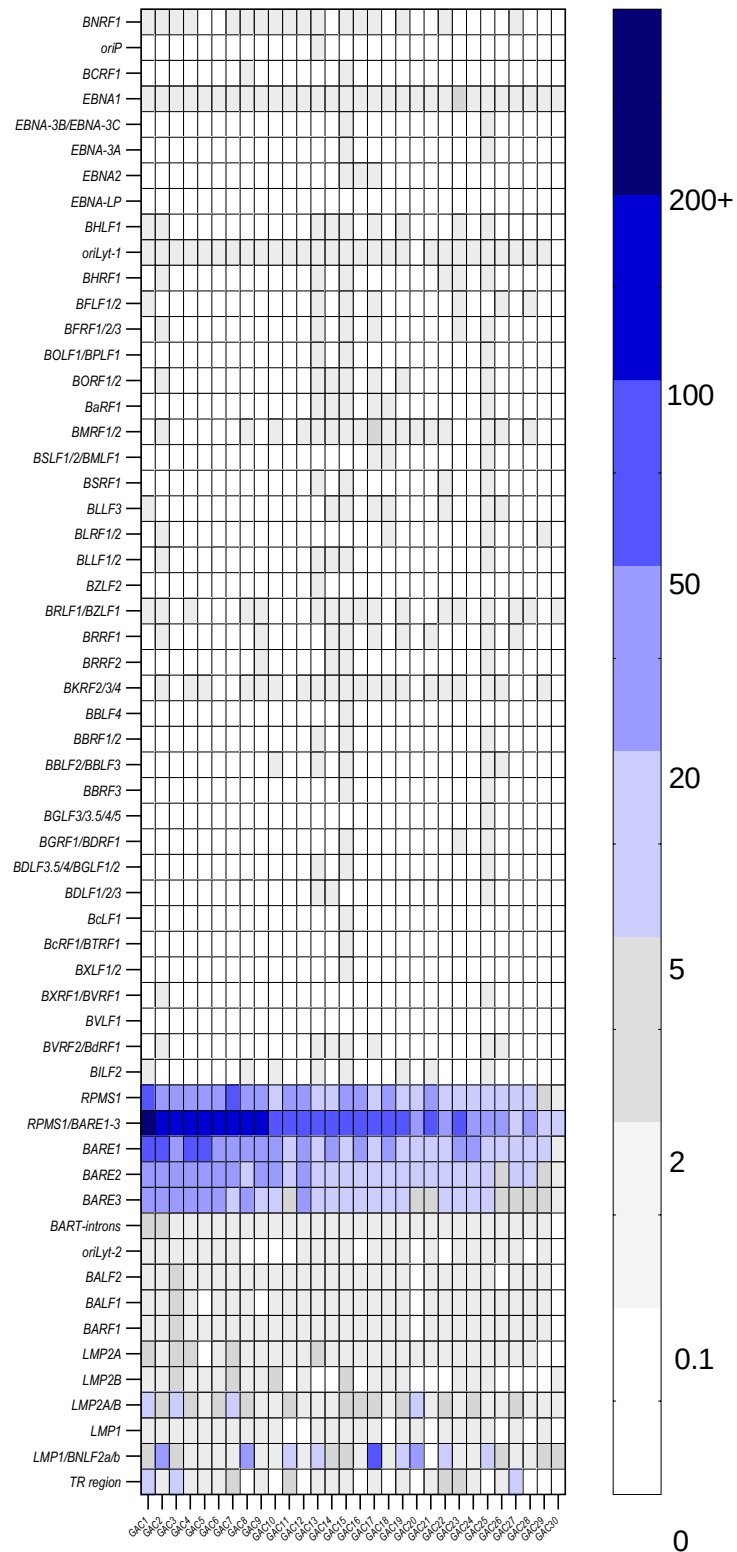
7. EBV gene expression

Tpm-values of EBV genes in individual datasets from NPC, GAC, eBL, sBL, ECL, LCL and modified Namalwa cell lines. Genes expressed at the range of 0.1-4.9 tpm are shown in grey. The EBV genes are ordered according to their position in the genome. tpm, transcripts per million; NPC, nasopharyngeal carcinoma; GAC, gastric adenocarcinoma; eBL, endemic Burkitt's lymphoma; sBL, sporadic Burkitt's lymphoma; ECL, EBV-associated tumor derived cell lines; LCL, lymphoblastoid cell lines

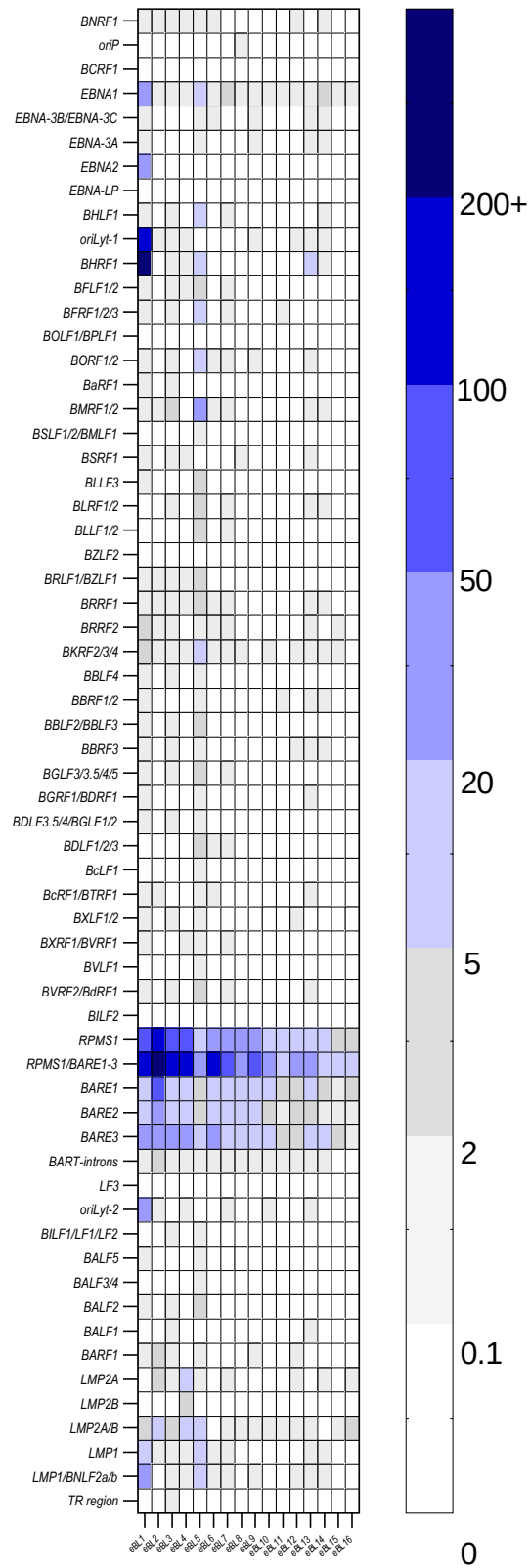
NPC



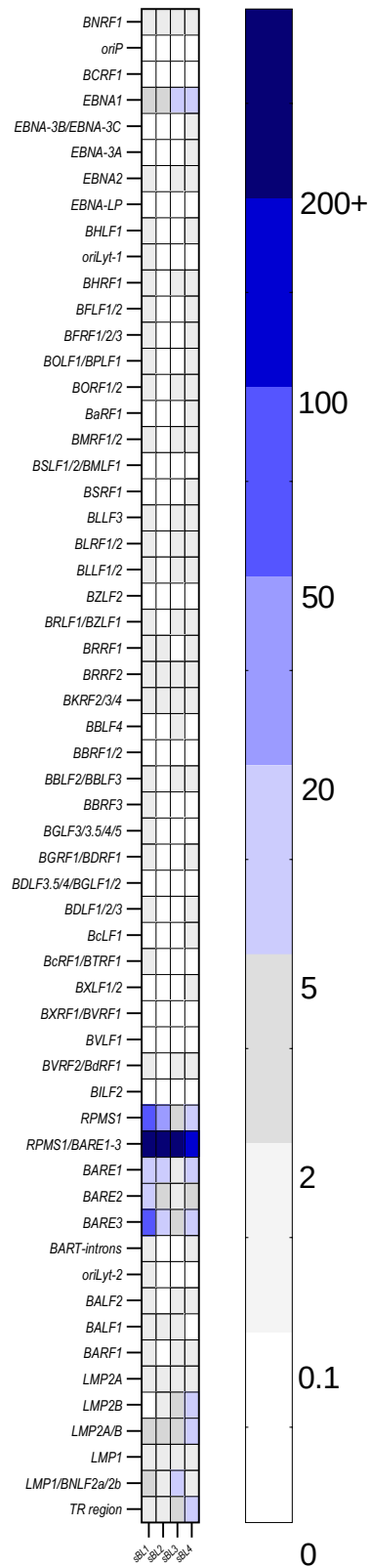
GAC



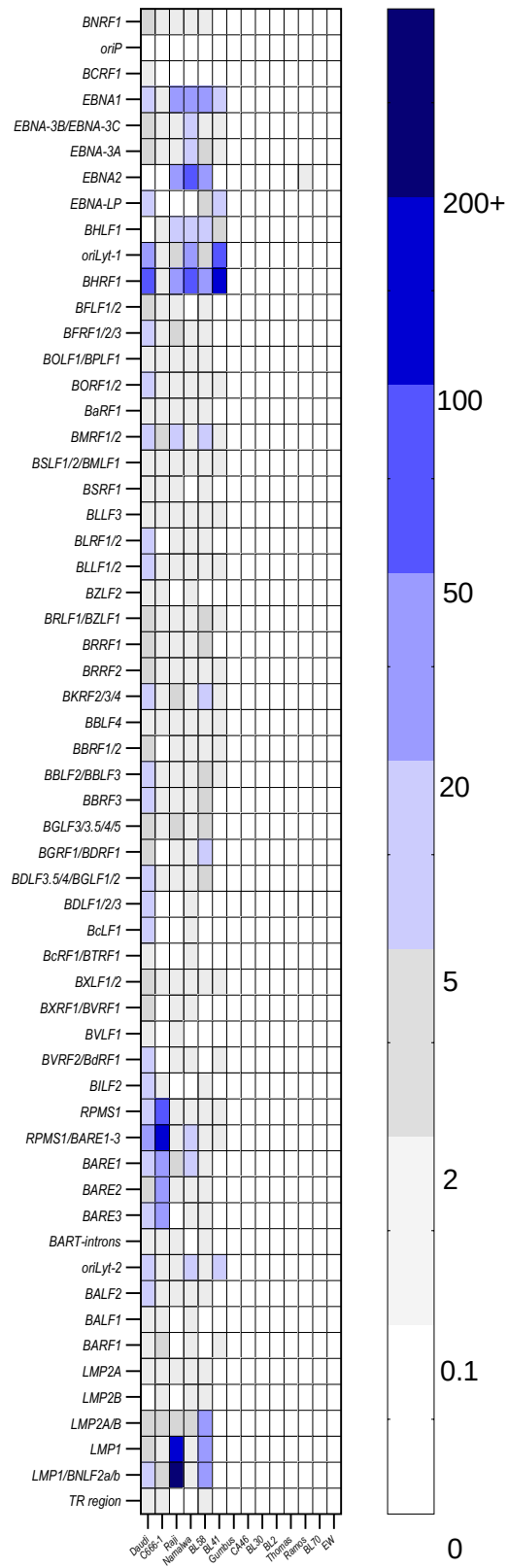
eBL



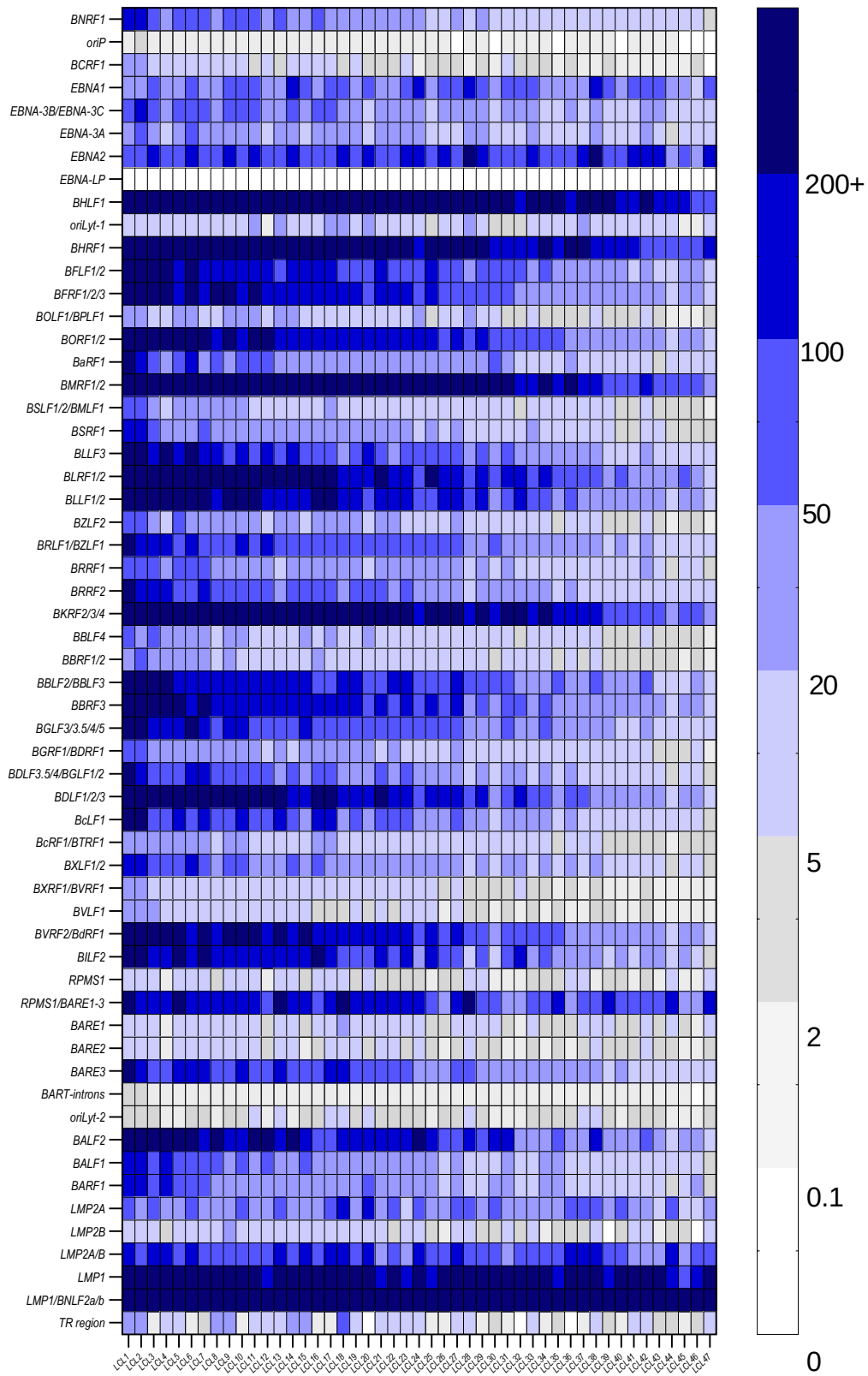
sBL



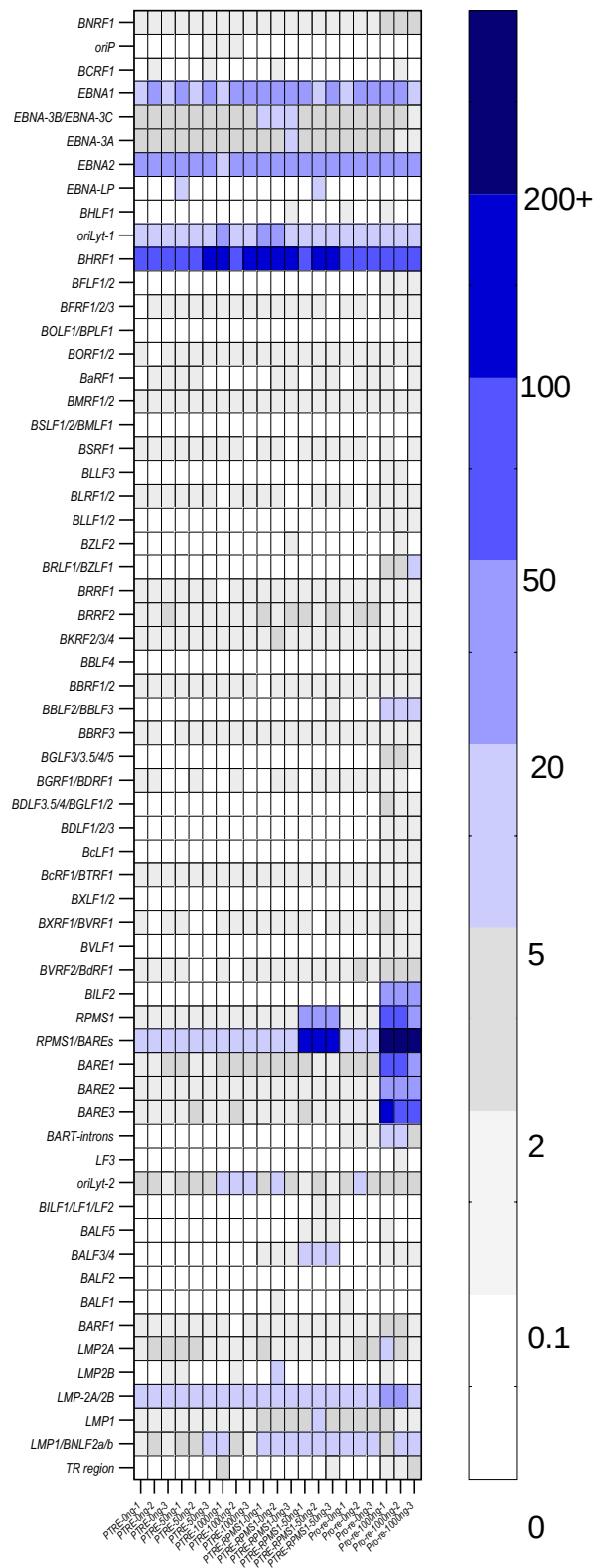
ECL



LCL



Modified Namalwa cell line



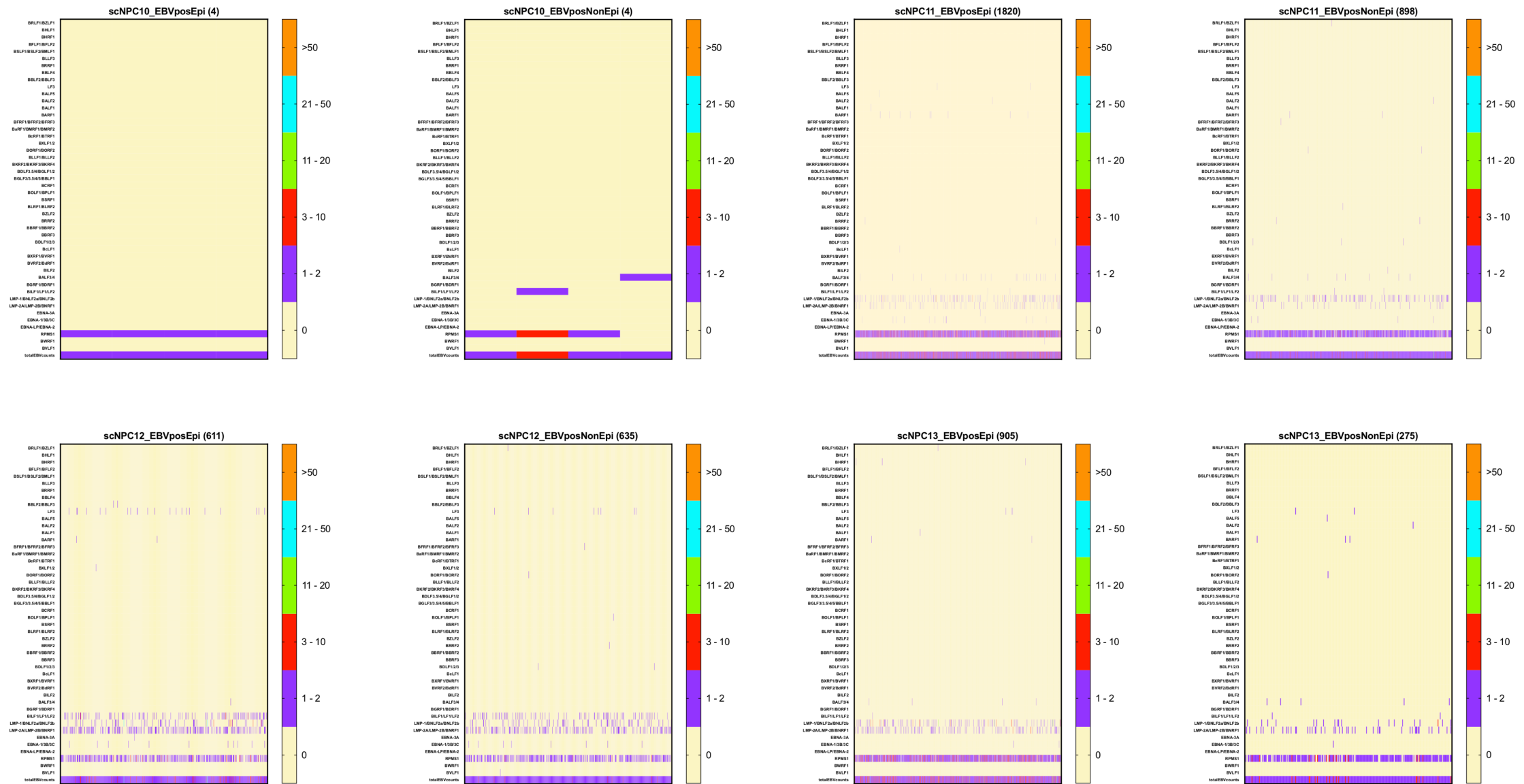
8. EBV-positive cells

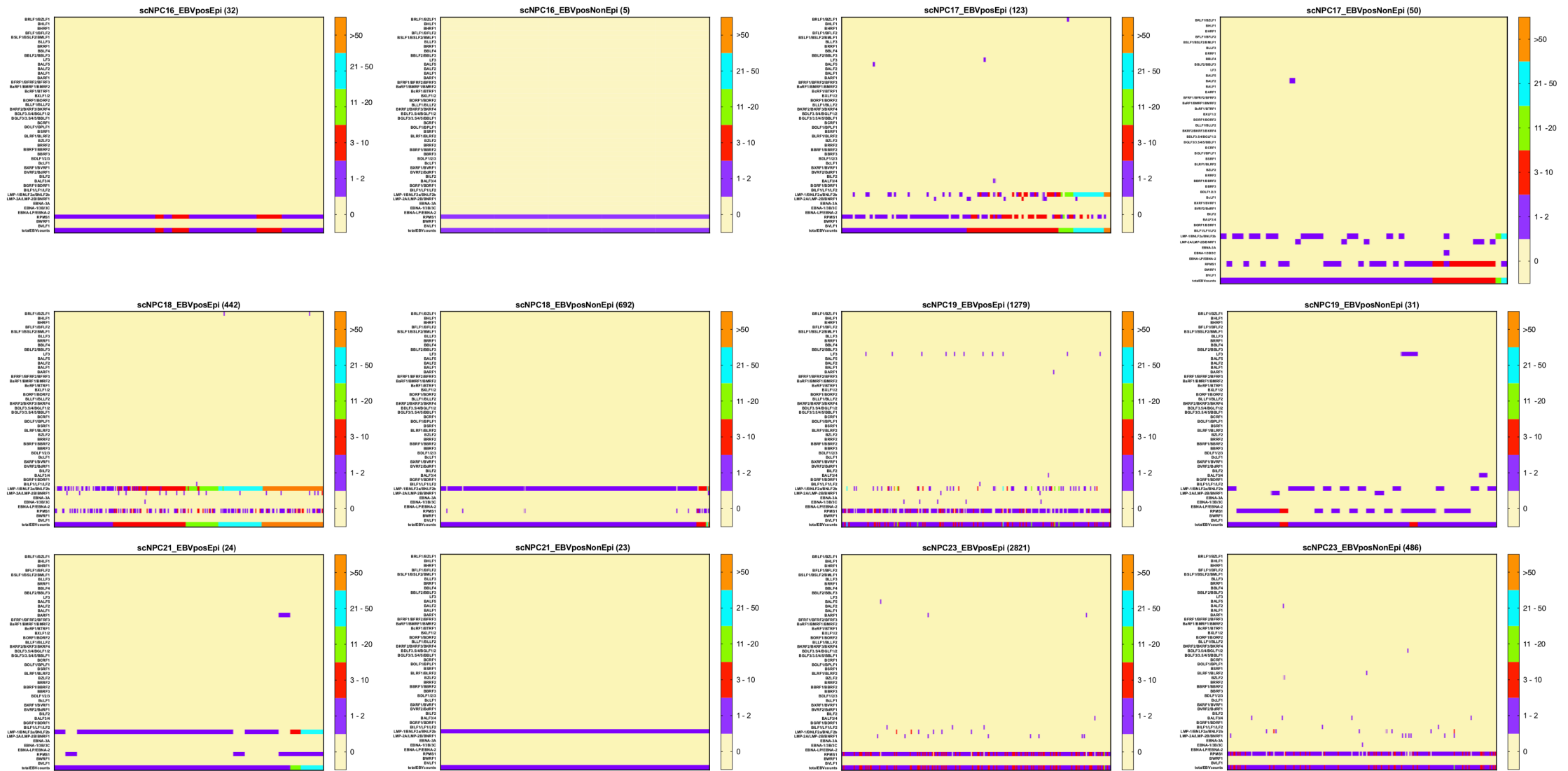
All EBV positive cells were included, the epithelial cells (Epi) and non-epithelial stromal cells (Non-Epi) were separated in every sample and the cell count is denoted within the brackets. The EBV genes are arranged by the gene groups; immediately early, early and late lytic, latent and unknown. The colour indicates the total number of RNAs detected in each cell.

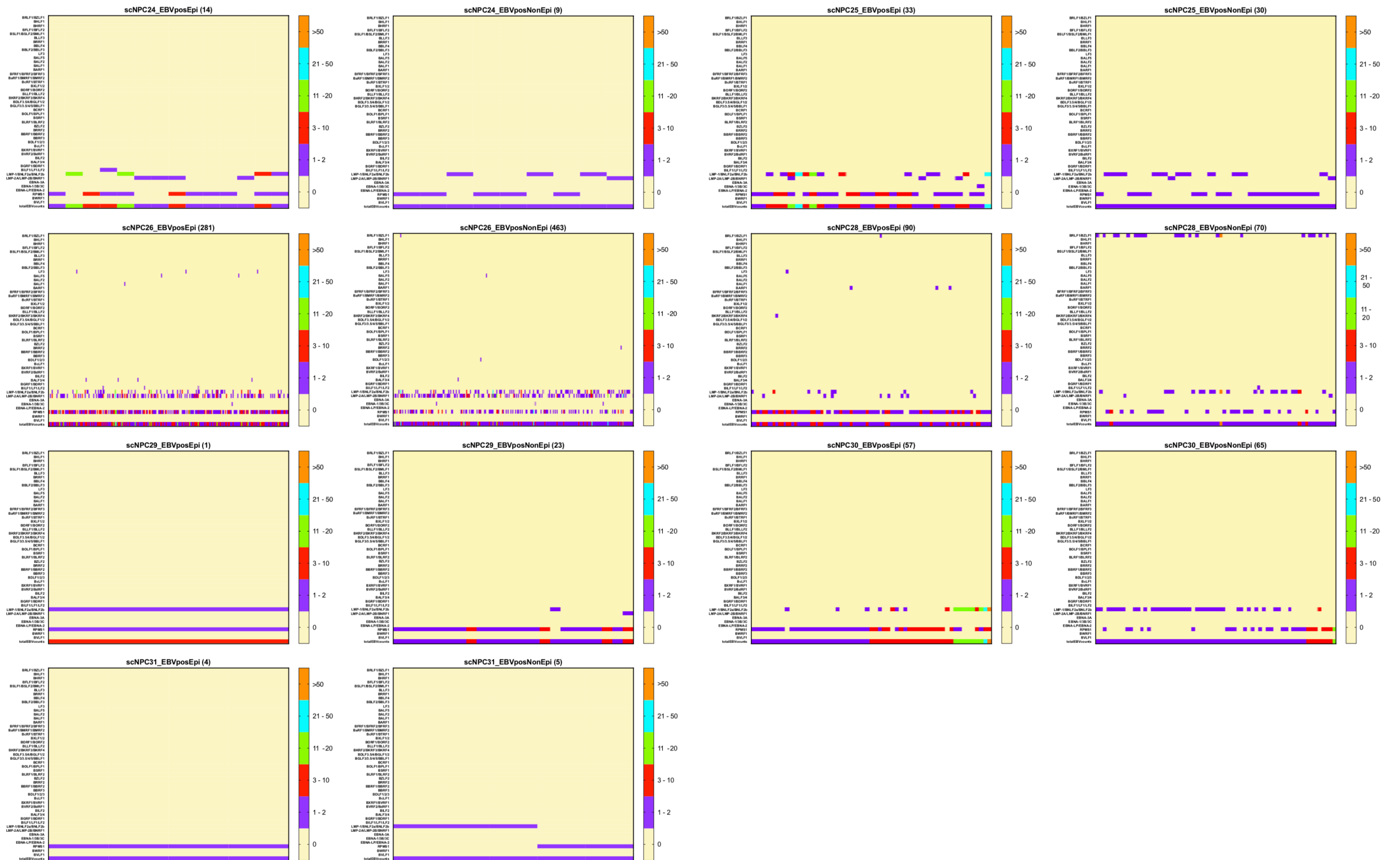
The EBV gene expression analyses of the majority of datasets have been limited due to the availability of bioinformatic adaptations for analysis of viral genomes. In the processing of scRNA-Seq data, reads mapping to overlapping genes would by default be discarded as they cannot be assigned to a single gene. Two different library preparation methods were used for generating the data, 10x Genomics 3' v2 (Study 1 and 2) or 5' v1 (Study 3 and 4), which influenced previous attempts to map EBV RNA, as many of the genes overlap. We have circumvented this problem by using a genome in which overlapping gene sections would be annotated to a single fused gene so that the EBV-reads could be detected and counted⁴.

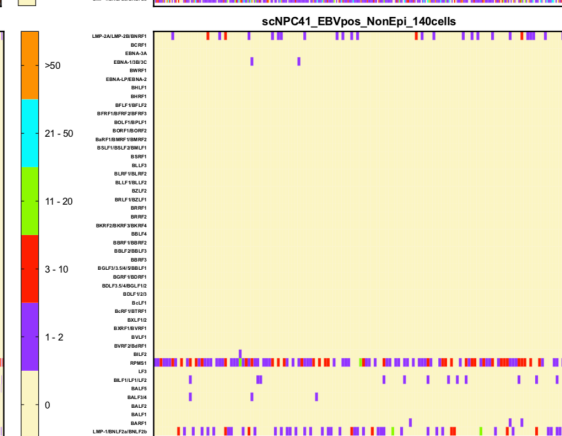
In accordance with bulk sequencing data a subset of tumors displayed a high expression of *LMP1/BNLF2*. The NPC samples, scNPC2, scNPC5 and scNPC18 had the highest *LMP1/BNLF2* levels (Extended Data Table 6). When ranking the tumors according to their *BRLF1/BZLF1* content, these three tumors were found amongst the top five samples, suggestive of frequent reactivation. These tumors also had the three highest fraction of EBV-expressing stromal cells (Figure 3a).

EBV-positive stromal cells were found in almost all EBV-positive scNPCs at various levels (0.1 to 49.5%). All types of stromal cells from the individual donors expressed EBV RNA at similar fractions (Extended Data Fig. 6b). Considering that EBV mainly has tropism for B-lymphocytes and to lesser extent T-lymphocytes and NK-cells these findings argue that the EBV RNA is derived from indiscriminate passive uptake from exosomes or apoptotic bodies of latent cells. The presence of *RPMS1/BAREs* with a predominantly nuclear distribution argues against exosomal excretion. Also, considering the extensive washing during the preparation of the cells and in some datasets an addition of FACS, the probability of the EBV RNA being derived from extracellular elements is highly unlikely.









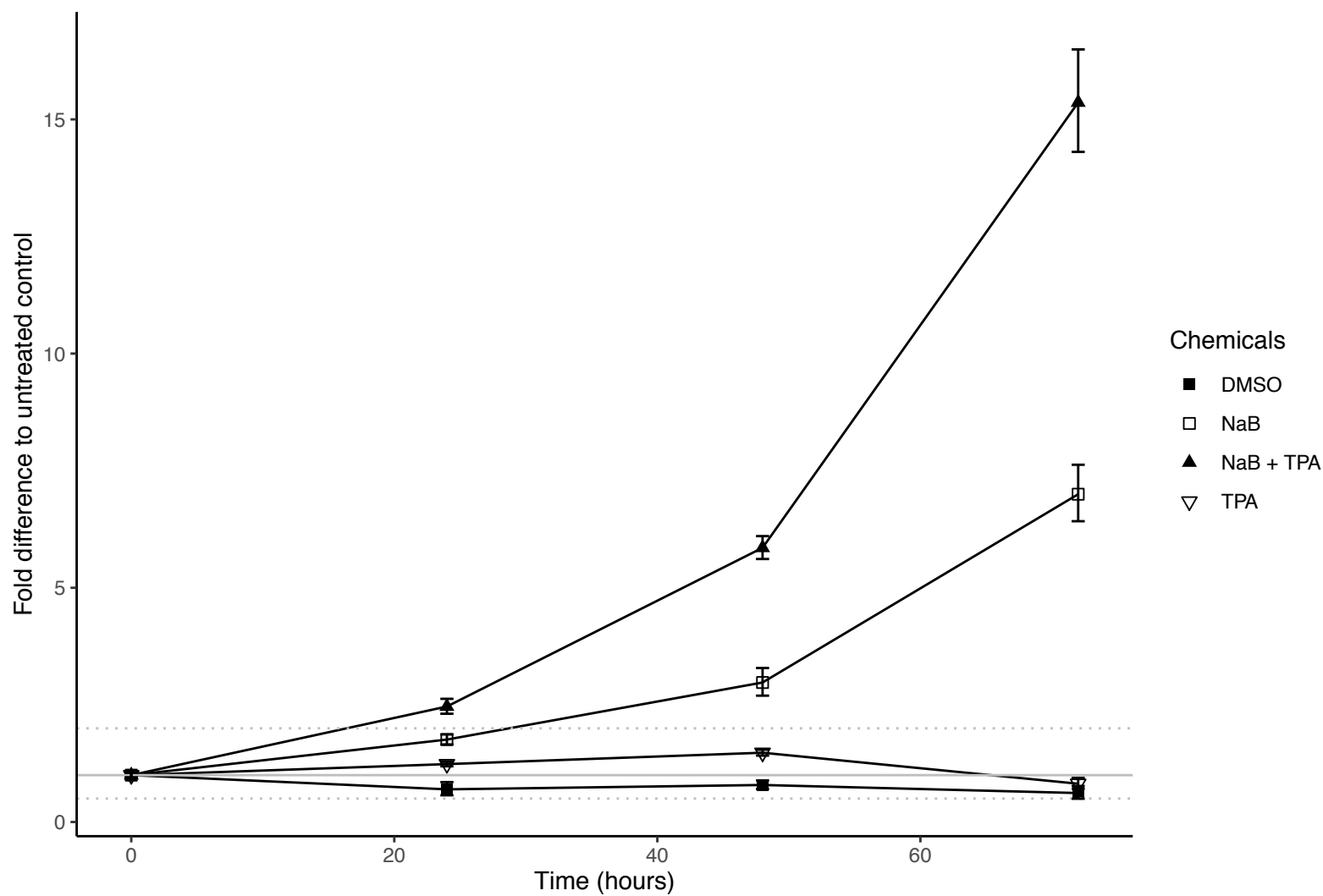


9. EBV reactivation

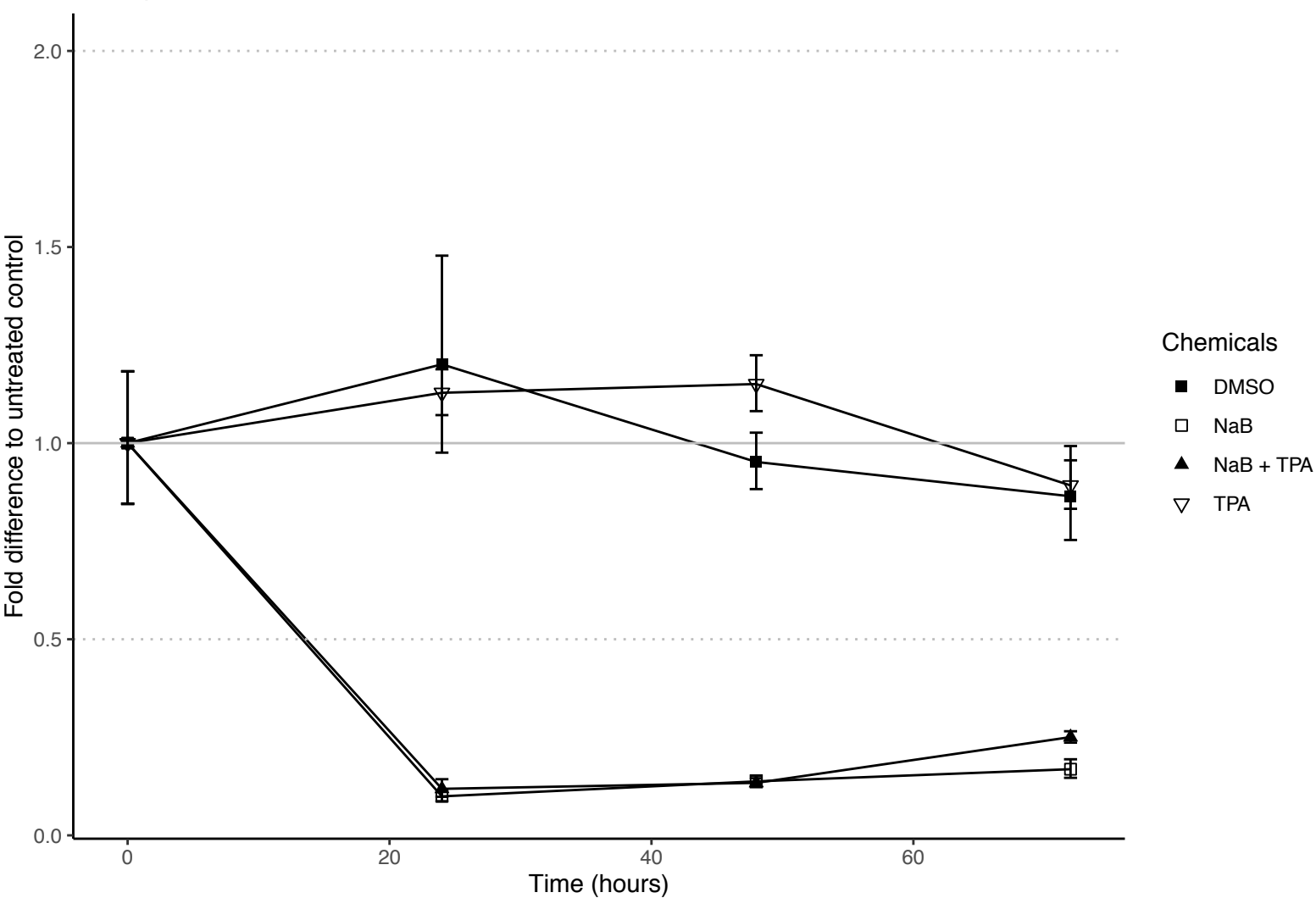
In stark contrast to the stromal cells with mirrored epithelial EBV expression, lytic EBV genes were detected in the stromal cells of scNPC28 and scNPC46, which were not found in their respective epithelial cells. These cells more likely represent infected stromal cells undergoing reactivation and not a passive uptake of EBV RNA. This is most apparent in scNPC28 where the EBV-positive cells have 1-3 detectable viral RNA molecules, but each cell is almost exclusively either expressing *RPMS1/BAREs*, *LMP1/BNLF2* or *BRLF1/BZLF1*. This mutual exclusivity indicates a separate state between the latent *RPMS1/BAREs* and/or *LMP1/BNLF2a/b* state and the lytic/reactivated *BRLF1/BZLF1* state. To confirm this phenomenon, we subjected *RPMS1/BAREs* expressing C666-1 cells to reactivation by addition of sodium butyrate and TPA. The RT-qPCR results showed an upregulation of *BZLF1* and a simultaneous downregulation of *RPMS1/BAREs*. The results imply that stromal cells in nasopharyngeal carcinoma can be infected and reactivated. However, considering the distribution of EBV RNA in the different cell types, the majority of the detected EBV RNA in the stromal cells are likely derived from an indiscriminate uptake of apoptotic bodies from cancer cells.

C666-1 cells were induced with chemicals for EBV reactivation, 3 mM sodium butyrate (NaB), 20 ng/ml 12-O-tetradecanoylphorbol-13-acetate (TPA) or a combination of both inducers. *BZLF1* (a) and *RPMS1* (b) expression levels were measured by RT-qPCR after 24 h, 48 h and 72 h of treatment. Each point represents the mean and the error bars the standard deviation obtained from three technical replicates. Horizontal unbroken grey line (y-intercept = 1) indicates equal gene expression level as in untreated control at respective time point. Dotted lines mark doubling (fold difference = 2) and halving (fold difference = 0.5) of the number of transcripts.

A Expression of BZLF1



B Expression of RPMS1

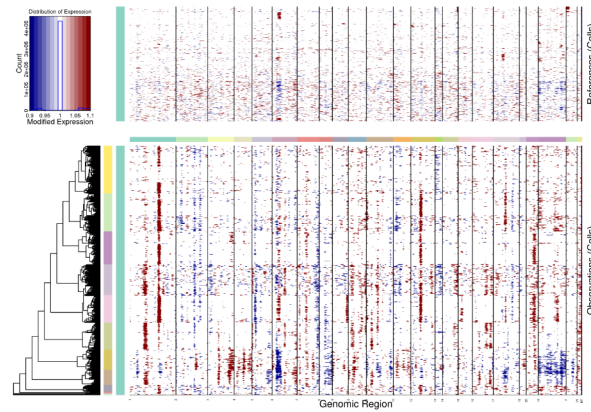


10. Cancer cell identification

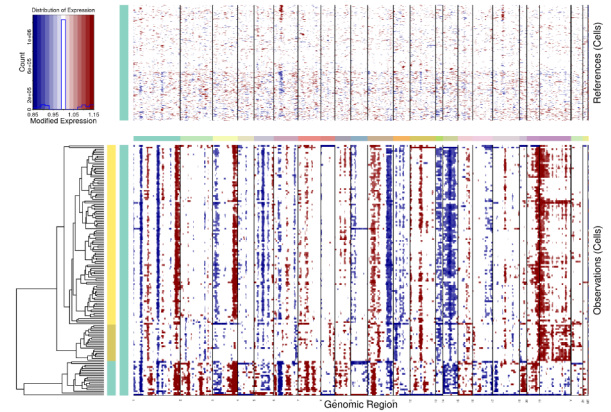
The major perturbations observed in comparison of bulk sequencing datasets originates from inter-patient differential gene expression. Subtle dysregulations involved in NPC development caused by EBV-expression therefore requires that the comparison is between healthy epithelial cells and cancer cells from the same patient. The classification of healthy epithelial cells and cancer cells requires a non-epithelial baseline as a reference. The two non-epithelial cell types found in all tumors are B-cells (except for NPC4) and T-cells. Considering that B-cells act as the main reservoir of latent EBV, a subset of 513 T-cells (except NPC41 in which only 274 T-cells were detected) were randomly selected to represent the reference baseline. Cells from individual tumors annotated by SingleR as either epithelial cells or T-cells were included for analysis using InferCNV⁵. The tumors with discordant EBV-status based on EBER-ISH, antibody staining and EBV RNA detection were not included in the analysis. The epithelial cells from each tumor were subsequently compared to the T-cells from the same sample. A minority of samples contained only tumor or healthy cells. InferCNV is the community standard for implying tumor status of all epithelial cells, similar to how a whole genome sequencing analysis would perform. The tool lacks definitive parameters for the designation of tumor cell status for individual cells. Therefore, we established a quantitative method to separate the healthy epithelial cells from the cancer cells. InferCNV converts each cell's whole transcriptomic expression profile into Modified Expression. This modified expression is then centered around a value of 1 with an assumed Gaussian background distribution. Using this matrix and calculating the standard deviation, i.e. variance (σ), of each cell results in a quantifiable parameter. This arbitrary determination could be assisted by hierarchical clustering methods supplied by InferCNV. However, as hierarchical clustering is heuristic in nature, it would still result in an arbitrary determination of the exact node at which cancer cells deviate from healthy cells. Instead, the variance of each cell's modified expression can act as a proxy for cancer probability with the caveat that the output is a continuous non-negative value range, rather than a percent probability. This is why the reference cells, in our case T-cells, have their variance of modified expression measured and used as a reference point for a normal distribution of variance.

In order to determine the cut-off for healthy or cancer cells we proportionally removed cells based on their proximity to the median until the lower variance subpopulation reflects the reference population, i.e. only the epithelial cells with the highest and lowest variance were classified into high-confidence cancer cells and high-confidence non-cancer cells, respectively. While the methodology performs well when a good ratio of underlying healthy and cancer cells exists, the top fraction of cells' absolute variance distribution is shifted closer toward the reference population's variance distribution if the number of cancer cells is lacking. Conversely, if the number of healthy cells is lacking, this analysis is non-functional, i.e. the bottom fraction of cells' absolute variance distribution is shifted away from the reference population's variance distribution. To avoid possible wrongful influence on downstream analysis, samples with too few healthy cells and cancer cells were omitted from further analysis.

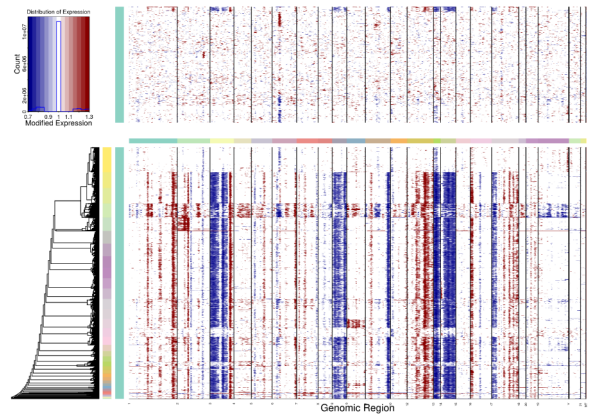
scNT1



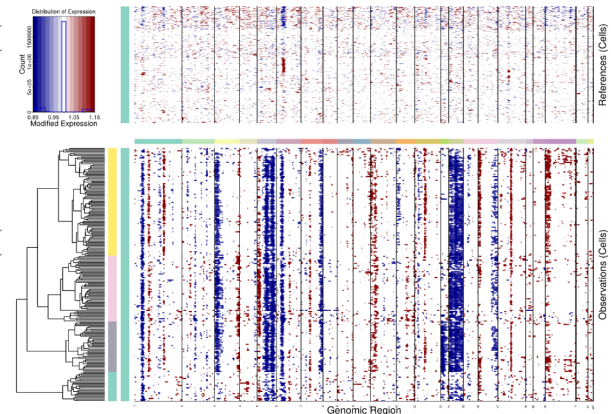
scNPC4



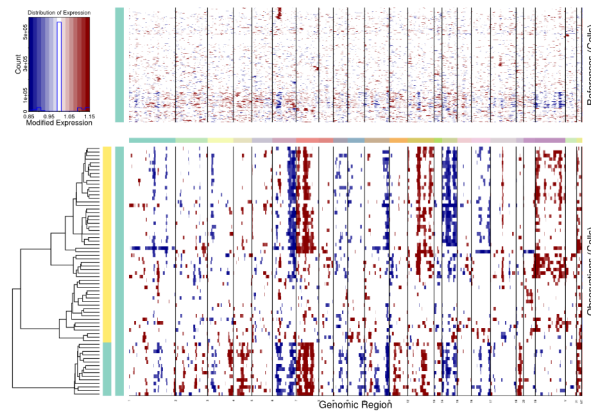
scNPC1



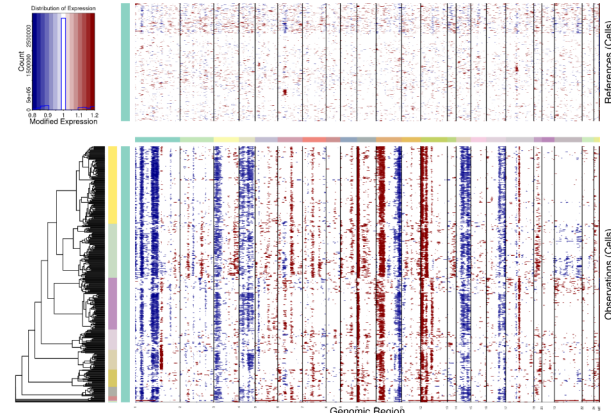
scNPC5



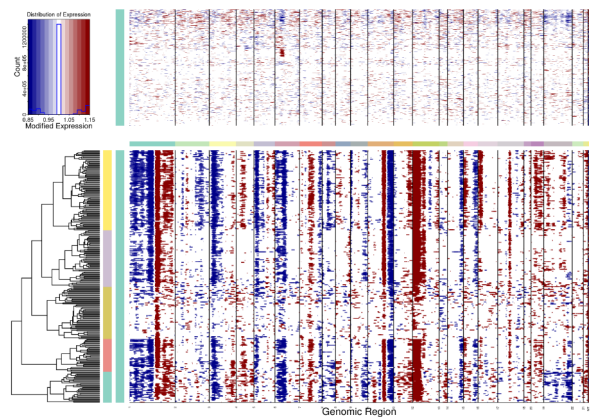
scNPC2



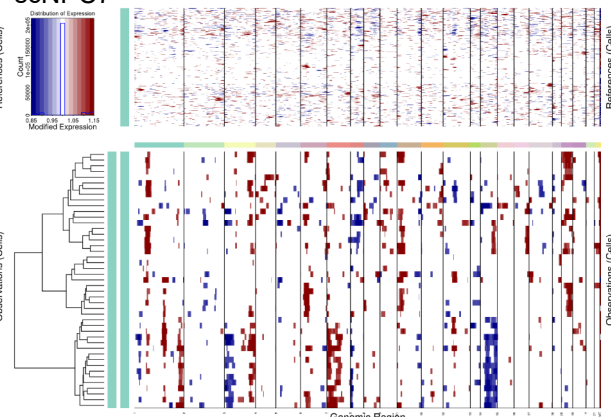
scNPC6



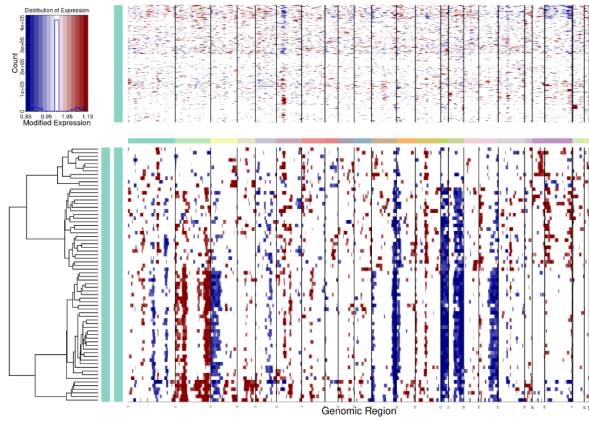
scNPC3



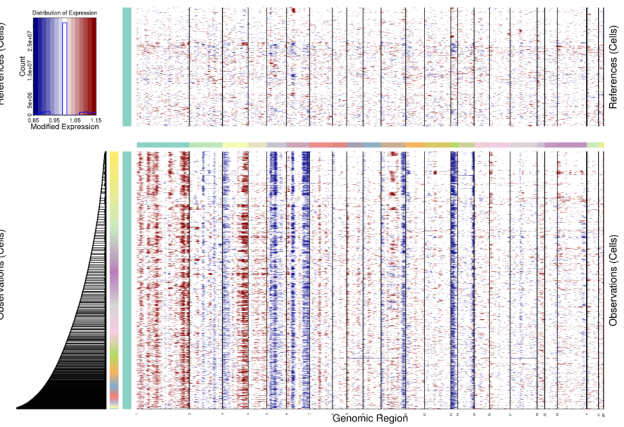
scNPC7



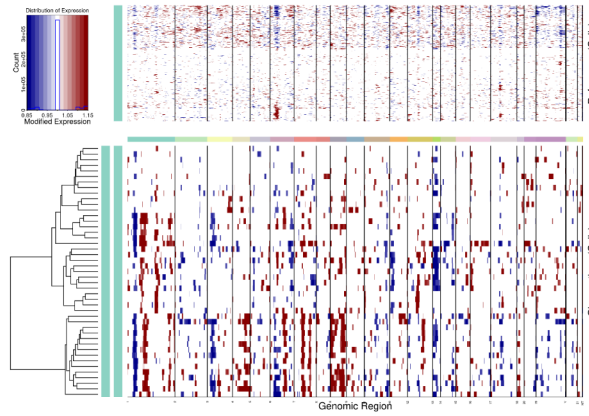
scNPC8



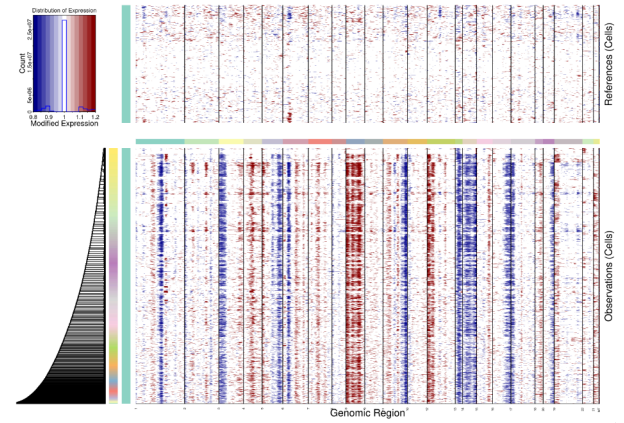
scNPC12



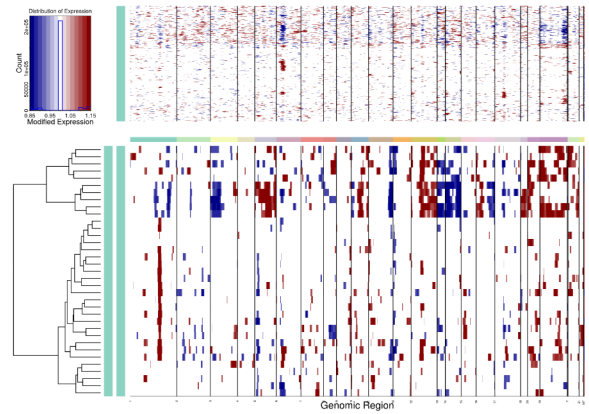
scNPC9



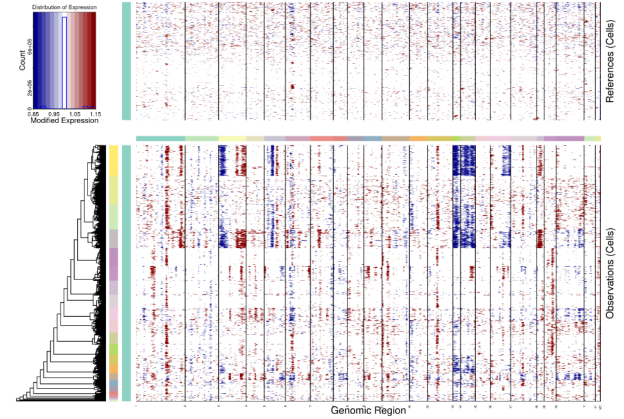
scNPC13



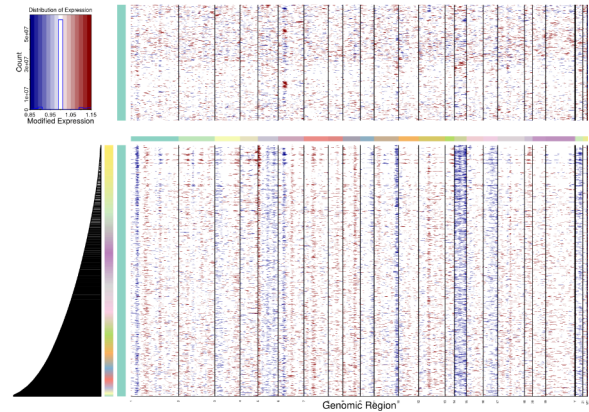
scNPC10



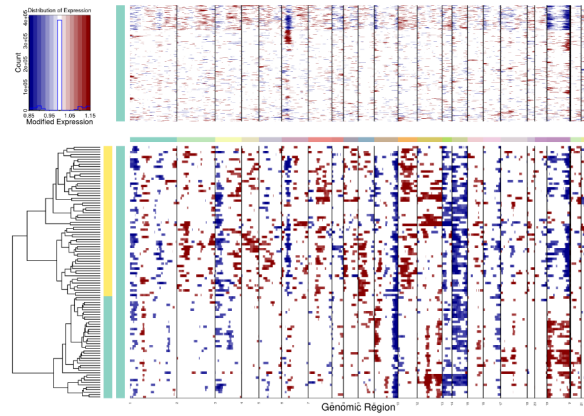
scNPC14



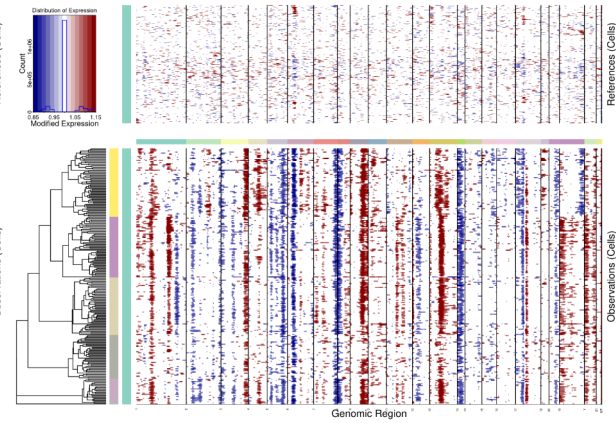
scNPC11



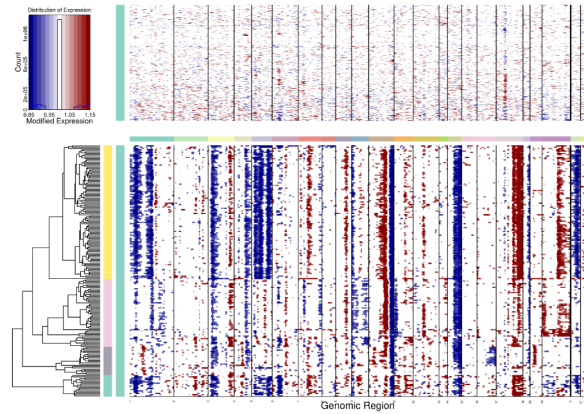
scNPC16



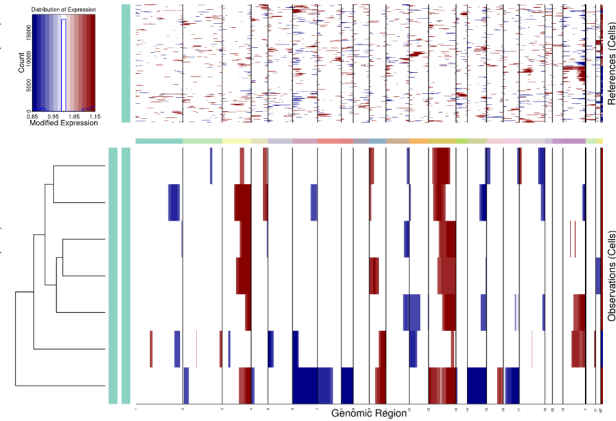
scNPC20



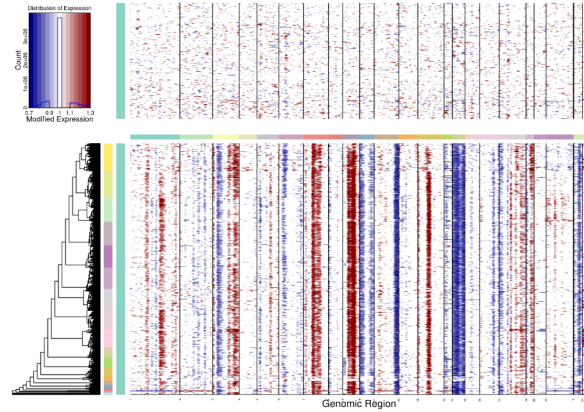
scNPC17



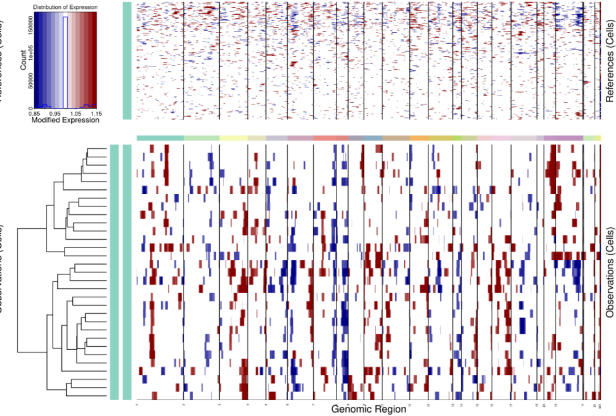
scNPC21



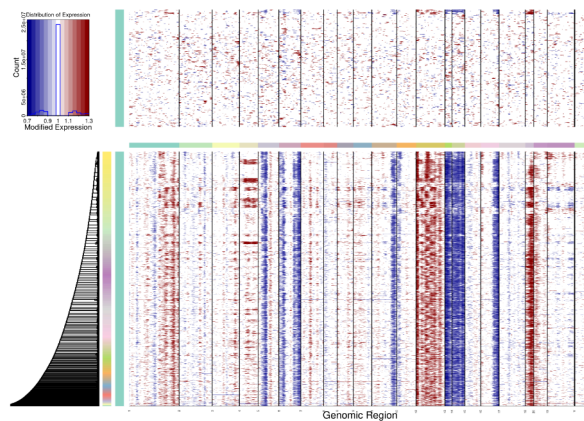
scNPC18



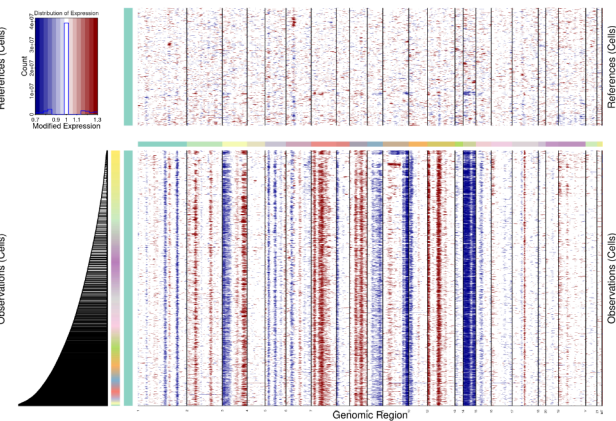
scNPC22



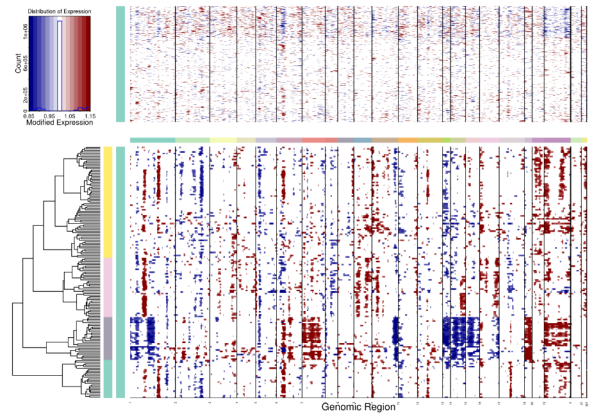
scNPC19



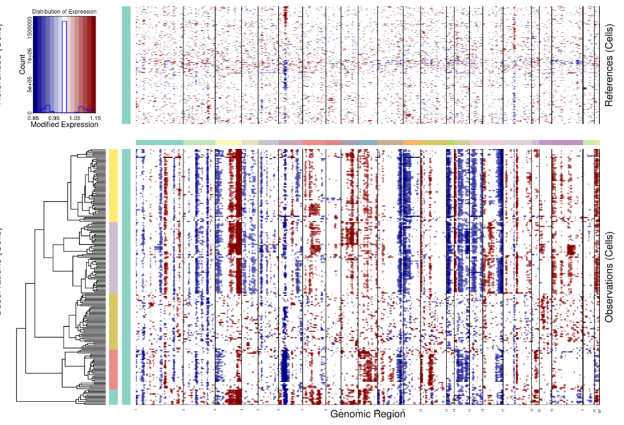
scNPC23



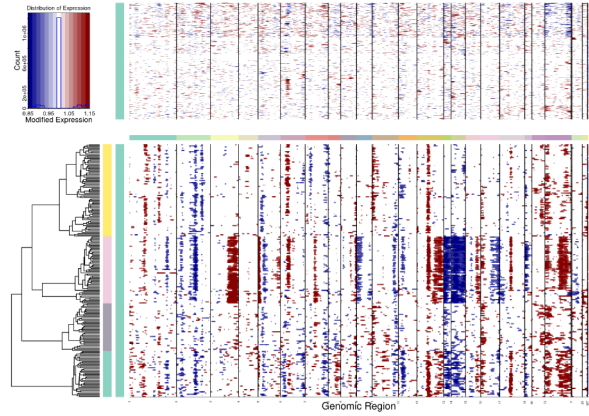
scNPC24



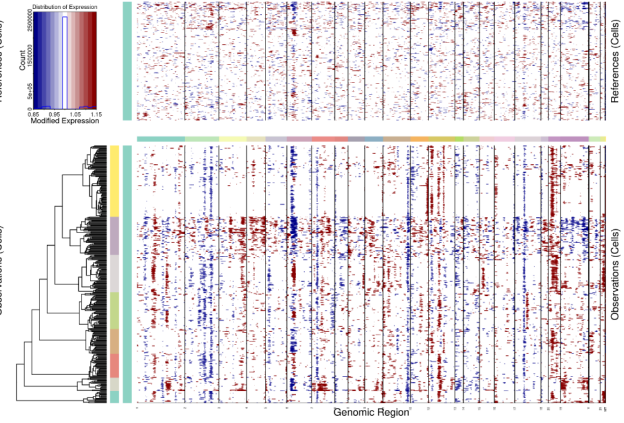
scNPC28



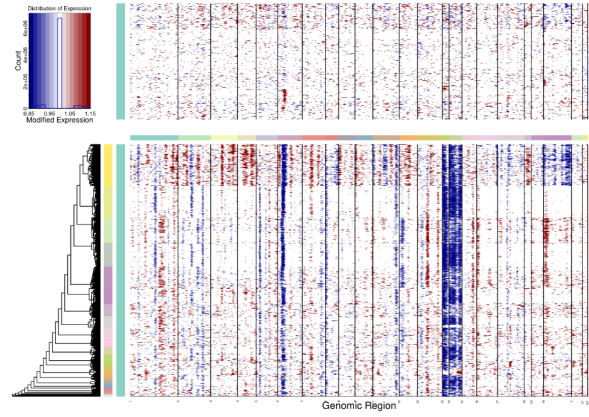
scNPC25



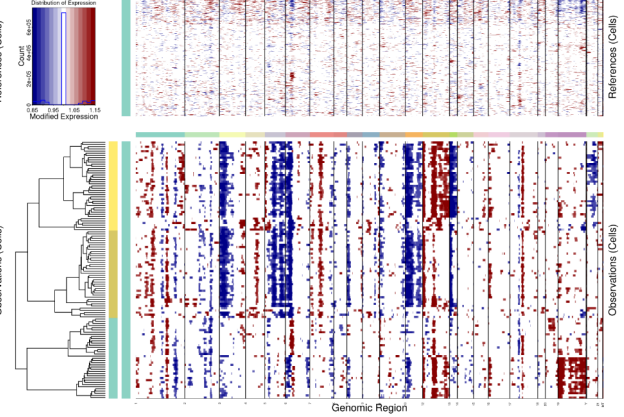
scNPC29



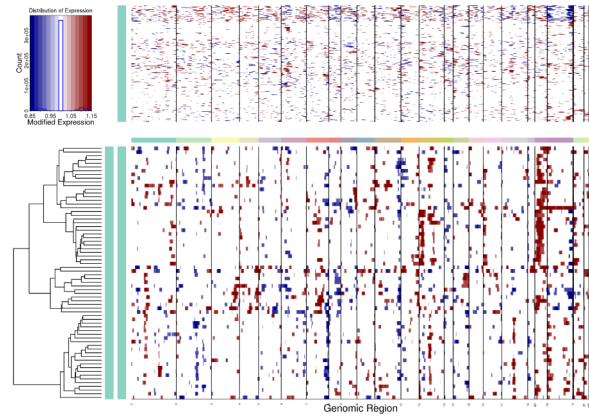
scNPC26



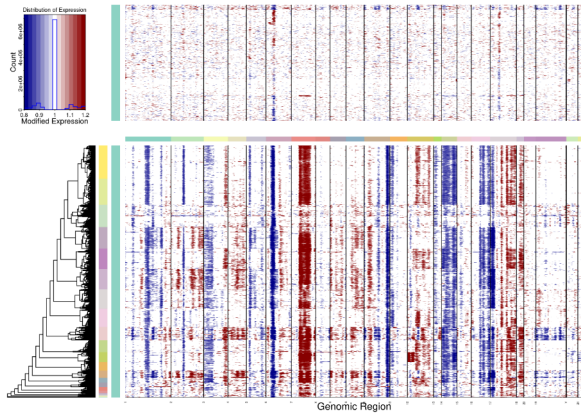
scNPC30



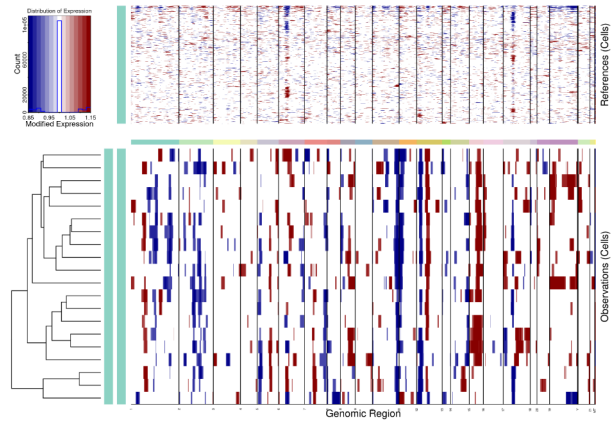
scNPC27



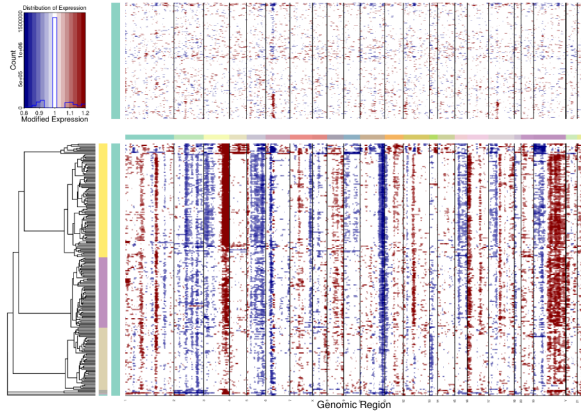
scNPC32



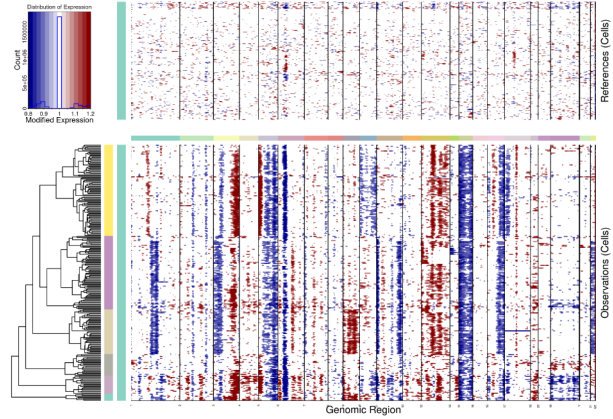
scNPC36



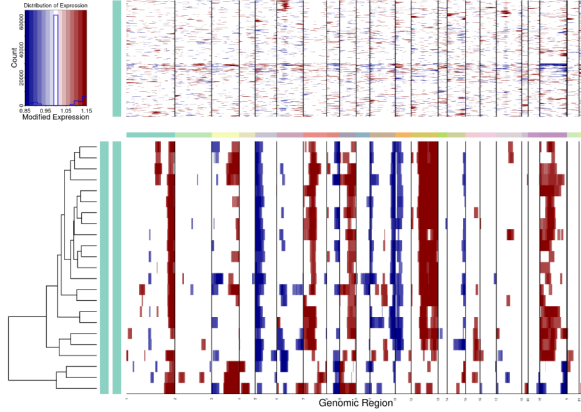
scNPC33



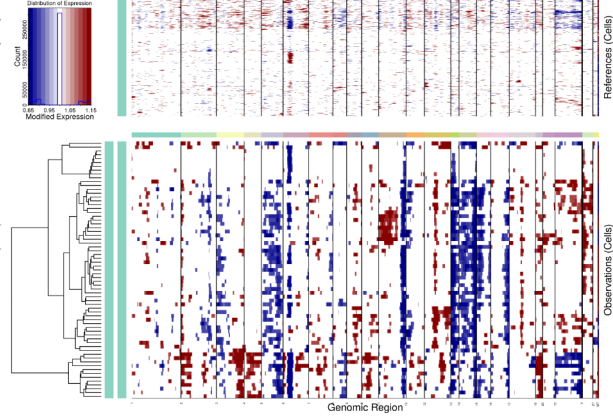
scNPC37



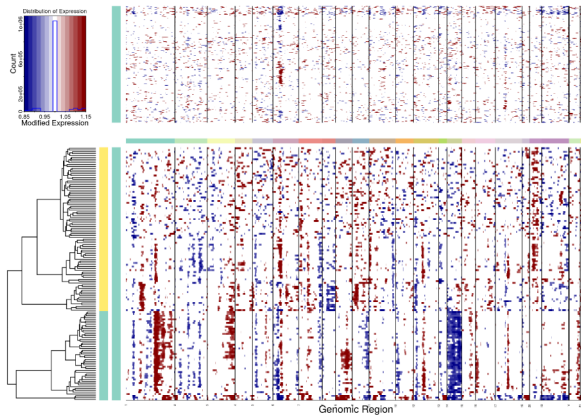
scNPC34



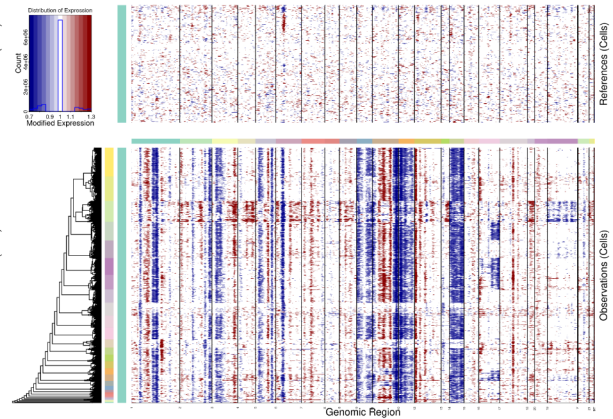
scNPC38



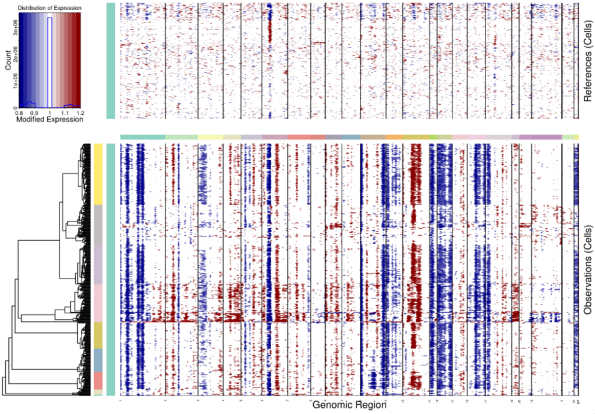
scNPC35



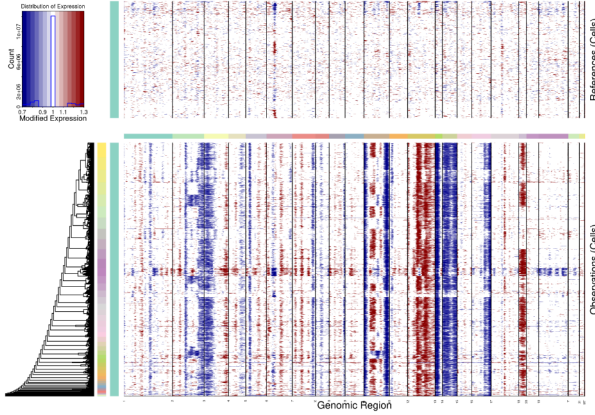
scNPC39



scNPC40

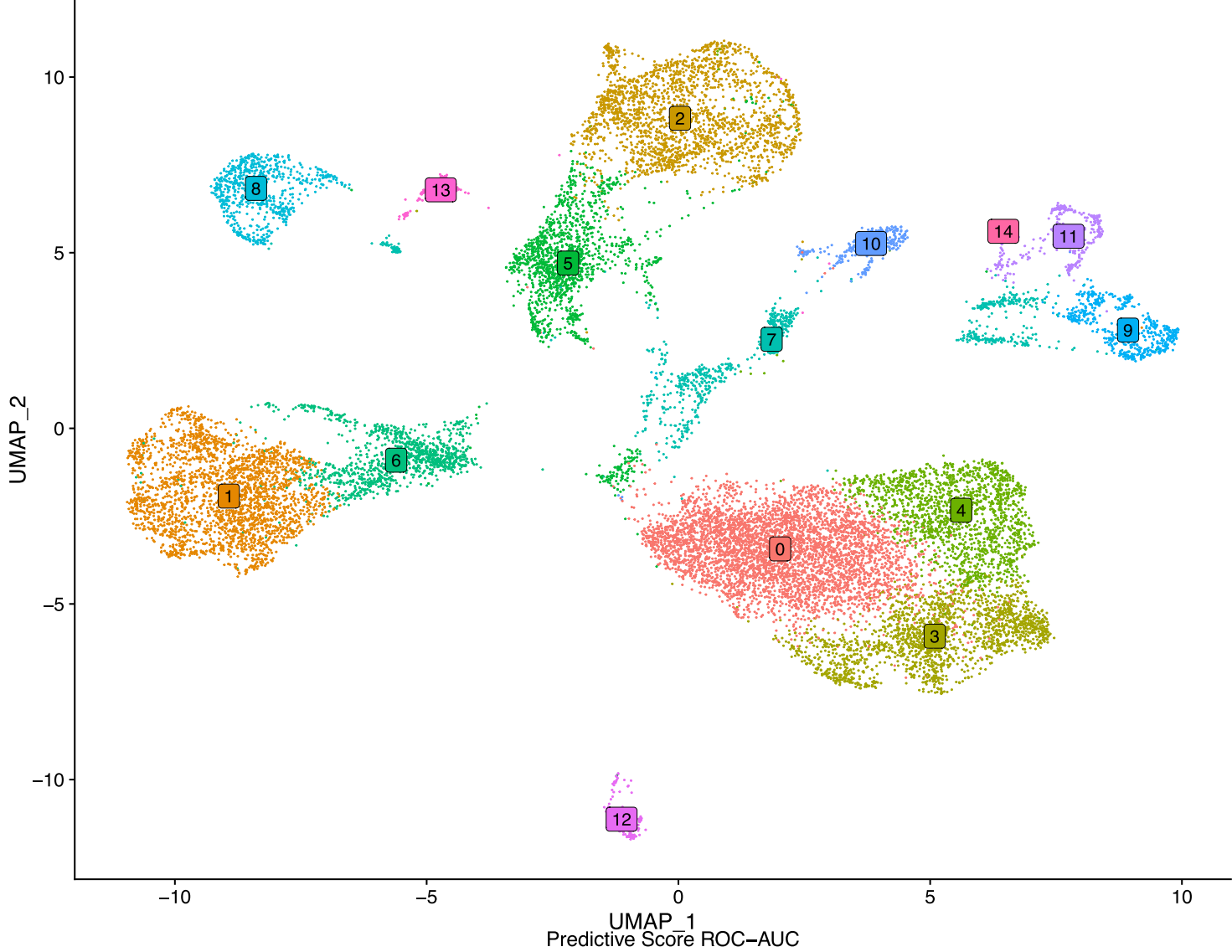


scNPC41



11. Epithelial cell classification

The epithelium around the nasopharynx acts as a transition between the stratified epithelium of the mouth and nose, and the airway epithelium. The pseudostratified epithelium is characterized by the presence of various epithelial cell types. Three scRNA-Seq studies had together 7 non-cancerous patient biopsies of the nasopharynx from healthy subjects. An automated annotation of cell types is commonly achieved by using SingleR, which is based on micro-array data of various cell types⁶. This tool is sufficient for most applications in which major cell types needs to be separated, for example T-cells and epithelial cells. However, due to the complexity of the pseudostratified epithelial tissue, further classification is required. After preprocessing the merged epithelial cells based on respective study, each of the clusters corresponding to different epithelial cell types were analyzed for specific marker genes⁷. Applying the ROC-AUC (Receiver Operating Characteristic Area Under Curve) using the marker genes, we noticed that no cluster had a strong signal for cells typical of the lower airway, i.e. ionocytes and alveolar cells. Instead, the cell types most notable were the basal cells, the secretory cells (merged from club cells and goblet cells), and the ciliated cells. Interestingly, the suprabasal cells are not very pronounced which can be explained by the fact that suprabasal are usually transition states between basal and differentiated epithelium. Using this method we have established the baseline for each of the epithelial cell types within healthy individuals as well characterizing the transcriptomic differences between basal and differentiated cells.



	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10	Cluster 11	Cluster 12	Cluster 13	Cluster 14	
C1orf194 C20orf85 C2orf40 C6orf135 C6orf142 DYNLL2 DYNLL2B2 ERICH3 FAM183A FOXJ1 AQP5 BPIFA1 BPIFB1 MSMB MUC5AC KRT5 KRT6A KRT13 KRT17 KRT19 KRT4 KRT7	0.338	0.759	0.729	0.477	0.406	0.421	0.536	0.328	0.571	0.266	0.684	0.253	0.627	0.610	0.206	Basal
KRT5 KRT6A KRT13 KRT17 KRT19	0.600	0.141	0.612	0.573	0.523	0.387	0.177	0.632	0.497	0.799	0.847	0.776	0.835	0.646	0.993	Suprabasal
KRT4 KRT7	0.385	0.448	0.571	0.527	0.541	0.308	0.351	0.627	0.892	0.878	0.639	0.530	0.601	0.546	0.911	Columnar
SCGB1A1 SCGB3A1 TP63	0.432	0.420	0.741	0.501	0.464	0.436	0.419	0.548	0.425	0.860	0.496	0.541	0.385	0.551	0.608	Secretory
C1orf194 C20orf85 C2orf40 C6orf135 C6orf142 DYNLL2 DYNLL2B2 ERICH3 FAM183A FOXJ1 AQP5 BPIFA1 BPIFB1 MSMB MUC5AC KRT5 KRT6A KRT13 KRT17 KRT19 KRT4 KRT7	0.354	0.713	0.633	0.519	0.458	0.354	0.442	0.489	0.604	0.617	0.404	0.430	0.485	0.566	0.451	Ciliated
ASCL3 ATP8V1G3 CLCNKB FOXJ1 FOXP1 FOXP2 PABRES2 SCG2 SCG3 STAP1 TSPAN19	0.476	0.505	0.488	0.510	0.500	0.474	0.485	0.503	0.565	0.633	0.494	0.520	0.506	0.542	0.491	Ionocyte
AGER CTSE NAPSA PGC RNASE1 SFTAA1 SFTAA2 SFTPB SFTPC SFTPD TCRH	0.458	0.547	0.480	0.514	0.499	0.456	0.484	0.481	0.543	0.580	0.581	0.580	0.591	0.556	0.627	Alveolar
DSG1 FABP5 FLG GNA2 KRT14 KRT6A KRT6B KRT6C SPRR1A SPRR1B SPRR2B SPRR2C SPRR2E SPRR4 TCRH	0.393	0.551	0.645	0.446	0.429	0.432	0.441	0.471	0.454	0.641	0.843	0.947	0.672	0.572	0.999	Stratified
DSG1 DSG2 DSG3 TGM1 TGM5	0.400	0.530	0.687	0.654	0.550	0.334	0.354	0.354	0.515	0.494	0.642	0.512	0.541	0.646	0.288	Spinous
DSG1 DSG2 DSG3 KRT10 KRT19 TGM1 TGM3	0.428	0.463	0.572	0.607	0.603	0.335	0.386	0.482	0.699	0.610	0.419	0.586	0.553	0.528	0.514	Granular
IVL LCE-Genes SPRR-Genes	0.439	0.426	0.688	0.506	0.470	0.487	0.416	0.487	0.494	0.520	0.545	0.941	0.507	0.498	1.000	Cornified
Unknown		Basal	Basal	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	Secretory	Suprabasal	Stratified	Suprabasal	Unknown	Stratified	

12. References

- 1 Holmqvist, I., Backerholm, A., Tian, Y., Xie, G., Thorell, K. & Tang, K. W. FLAME: long-read bioinformatics tool for comprehensive spliceome characterization. *RNA* **27**, 1127-1139 (2021). <https://doi.org:10.1261/rna.078800.121>
- 2 Lawrence, J. B., Villave, C. A. & Singer, R. H. Sensitive, high-resolution chromatin and chromosome mapping in situ: presence and orientation of two closely integrated copies of EBV in a lymphoma line. *Cell* **52**, 51-61 (1988). [https://doi.org:10.1016/0092-8674\(88\)90530-2](https://doi.org:10.1016/0092-8674(88)90530-2)
- 3 Hoang, V. L. T. *et al.* RNA-seq reveals more consistent reference genes for gene expression studies in human non-melanoma skin cancers. *PeerJ* **5**, e3631 (2017). <https://doi.org:10.7717/peerj.3631>
- 4 Ziegler, P. *et al.* A primary nasopharyngeal three-dimensional air-liquid interface cell culture model of the pseudostratified epithelium reveals differential donor- and cell type-specific susceptibility to Epstein-Barr virus infection. *PLoS Pathog* **17**, e1009041 (2021). <https://doi.org:10.1371/journal.ppat.1009041>
- 5 Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396-1401 (2014). <https://doi.org:10.1126/science.1254257>
- 6 Aran, D. *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat Immunol* **20**, 163-172 (2019). <https://doi.org:10.1038/s41590-018-0276-y>
- 7 Ruiz Garcia, S. *et al.* Novel dynamics of human mucociliary differentiation revealed by single-cell RNA sequencing of nasal epithelial cultures. *Development* **146** (2019). <https://doi.org:10.1242/dev.177428>