# All cell lines in Trips-viz with good triplet periodicity score

Here are the studies that I used that are already available in Trips-viz.

I selected ones that have good triplet periodicity score (>=0.5).

| **Study short name and doi** | **Cell line** | **Number of samples** | **GSE/PRJNA** | **Triplet periodicity** |
| --- | --- | --- | --- | --- |
| Fijalkowska, 2017  <https://doi.org/10.1093/nar/gkx469> | HCT116 (colon cancer cells) | 4 (init + elong) | GSE87328 | 0.76 |
| Crappe, 2015 <https://doi.org/10.1093/nar/gku1283> | HCT116 (colon cancer cells) | 2 | GSE58207 | 0.68 |
| Gameiro18  <https://doi.org/10.1016/j.celrep.2018.07.021> | MCF10A-ER-Src (breast cancer cells) | 25 | GSE114794 | 0.57 |
| Ji16  <https://doi.org/10.7554/eLife.08890> | MCF10A-ER-Src (breast cancer cells) | 26 (init + elong) | GSE65885 | 0.58 |
| Lauria18  <https://doi.org/10.1371/journal.pcbi.1006169> | MCF7 (breast cancer cells) | 1 | GSE111866 | 0.52 |
| Gawron16  <https://doi.org/10.15252/msb.20156662> | Jurkat (T-cell leukaemia, peripheral blood) | 1 (init) | GSE74279 | 0.56 |
| Wolfe15  <https://doi.org/10.1038/nature13485> | KOPT-K1 (Childhood T acute lymphoblastic leukaemia cells) | 4 | GSE56887 | 0.68 |
| Hsieh12  <https://doi.org/10.1038/nature10912> | PC3 (prostate cancer cells) | 6 | GSE35469 | 0.52 |
| Guo14  <https://doi.org/10.1186/s13059-014-0409-z> | U2OS (Bone Osteosarcoma Epithelial Cells) | 3 | GSE51584 | 0.74 |
| Goodarzi16  <https://doi.org/10.1016/j.cell.2016.05.046> | CN34-parental and CN34-LM1a (breast cancer cell line) | 24 | GSE77347 | 0.58 |

72+24=96 samples

* I used the Trips-viz algorithm to predict uORF (upstream), uoORF (upstream overlapping), nORF (nested), dORF (downstream), doORF (downstream overlapping), ORFs in lncRNA.
* For each type of non canonical ORF, I manually choose the rank for which the translation signal of ORF looks reliable (I provide plots of the top and bottom examples of non-canonical ORFs, **Fig.1** and **Fig.2**).

| **Type** | **Rank threshold** |
| --- | --- |
| uORF | 6200 |
| uoORF | 800 |
| nORF | 950 |
| lncORFs | 2800 |
| dORF | 350 |
| doORF | 110 |

* I added genomic coordinates and aggregated predicted ORFs based on that (e.g. the same ORF can be uORF in one transcript and nORF in another).
* It resulted in 10997 unique peptides (11204 non-canonical ORFs, file = ncORFs\_with\_global\_coo.txt).
* 110 of peptides from HLA peptidomics can be found in 94 ncORFs (peptides\_found\_in\_ncORFs.txt, I plotted some of their profiles with good ranks, **Figure 3**).

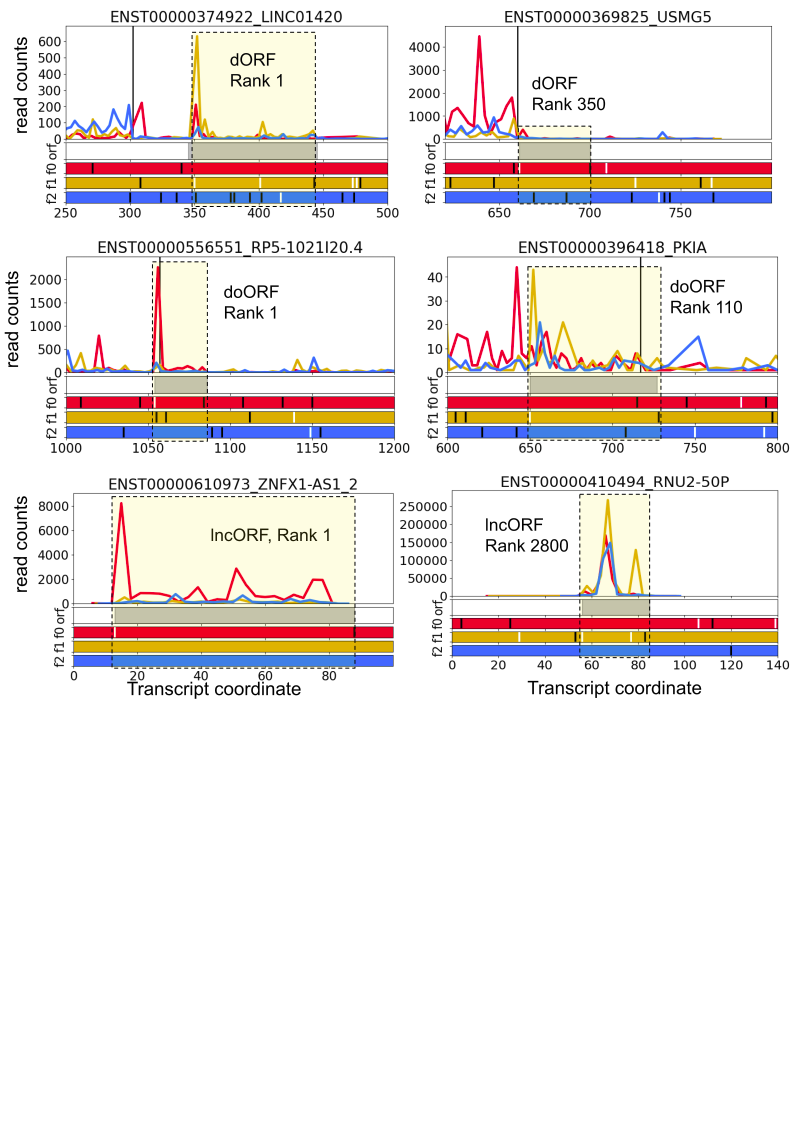


Fig.1. Subcodon ribo-seq profiles of top- and bottom-ranked non-canonical ORFs (dORF, doORF, ORFs in lncRNAs). Profiles are differentially coloured by a supported reading frame. Bottom colour bars show ORF plot where white lines are AUGs, black - stop codons. Predicted ORFs are depicted as grey bars and also highlighted in yellow.

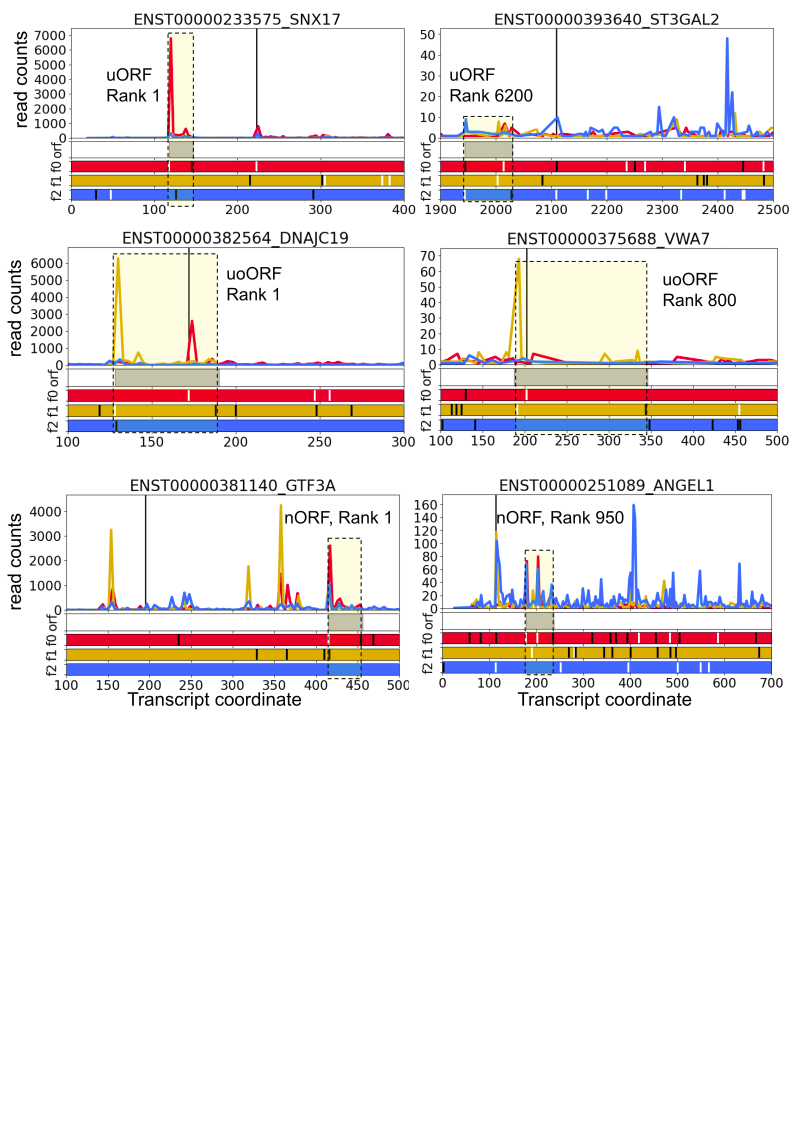


Fig.2. Subcodon ribo-seq profiles of top- and bottom-ranked non-canonical ORFs (uORF, uoORF, nORF). Profiles are differentially coloured by a supported reading frame. Bottom colour bars show ORF plot where white lines are AUGs, black - stop codons. Predicted ORFs are depicted as grey bars and also highlighted in yellow.

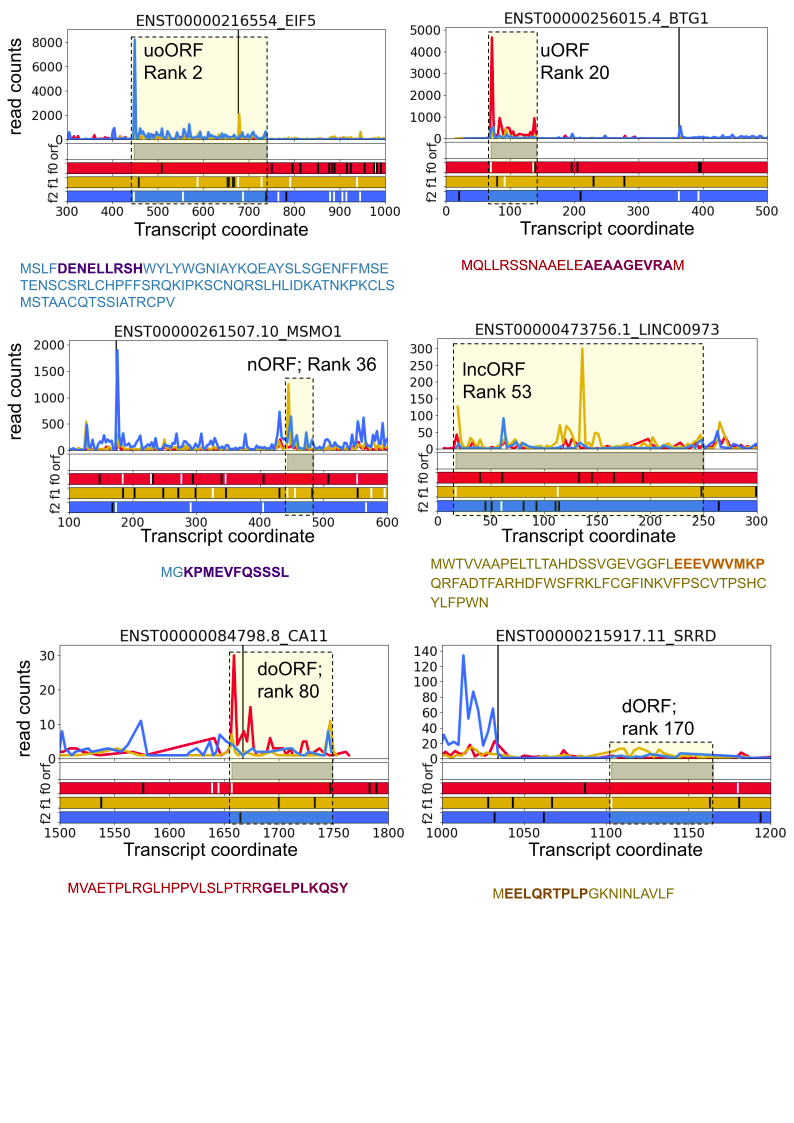


Fig.3 Subcodon ribo-seq profiles of ncORFs for which there are peptides found in HLA peptidomics. Profiles are differentially coloured by a supported reading frame. Bottom colour bars show ORF plot where white lines are AUGs, black - stop codons. Predicted ORFs are depicted as grey bars and also highlighted in yellow. Amino acid sequence of ncORF-encoded peptide are show at the bottom of each profile and HLA peptide is show in bold and different color.

# Only HCT116 cell line non-canonical ORFs

In Trips-viz there are only 2 studies with HCT116 cell line (6 samples, mapped on GENCODE v25).

Therefore I added more studies and reprocessed raw reads and mapped them on GENCODE v41.

Here are the studies that I used. Notice that only 3 of them with 20 samples have decent triplet periodicity score, > 0.5 which is an important measure of quality of the data because Trips-viz ORF predictor relies on triplet periodicity for detection of translated ORFs.

| **Study short name and doi** | **Number of samples** | **GSE/PRJNA** | **Triplet periodicity** | **Adapter sequence** |
| --- | --- | --- | --- | --- |
| Fijalkowska, 2017  <https://doi.org/10.1093/nar/gkx469> | 4 | GSE87328 | 0.76 | AGATCGGAAGAGCACAC |
| Crappe, 2015 <https://doi.org/10.1093/nar/gku1283> | 2 | GSM1403308 | 0.77 | AGATCGGAAGAGCACAC |
| Babaian, 2020 <https://doi.org/10.1016/j.celrep.2020.107611> | 14 | PRJNA602544 | 0.82 | CTGTAGGCACCATCAAT |
| Calviello, 2021  <https://doi.org/10.1093/nar/gkab287> | 8 | GSE157063 | 0.23 | TAGACAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC |
| Shuaikun Su, 2022 <https://doi.org/10.1093/nar/gkac538> | 23 | GSE188574 | 0.35 | AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC |

Therefore I decided to go with another option - taking multiple already available cancer cell lines in Trips-viz with good triplet periodicity.