**内源性病毒潜在新抗原多肽筛选结题报告**

目录

[项目目标 2](#_Toc22230057)

[项目成员 2](#_Toc22230058)

[项目时间 2](#_Toc22230059)

[项目背景 2](#_Toc22230060)

[项目总结 3](#_Toc22230061)

[附录 5](#_Toc22230062)

[方法详细流程及说明 5](#_Toc22230063)

[附件列表 15](#_Toc22230064)

# 项目目标

筛选来源于内源性病毒（HERV）的泛癌肿肿瘤新抗原，并针对结直肠癌（COAD）进行优化

# 项目成员

大数据中心：丁远彤，李华平，杨悦羽霄

吉诺因：张乐

# 项目时间

2019.3-2019.10

# 项目背景

肿瘤特异性抗原（TSA），即在正常组织中不表达，而仅在肿瘤组织表达的抗原，是备受关注的肿瘤免疫疗法之一。肿瘤特异性抗原的来源可分为三种，一是外源性来源，如个别由外源病毒引起的肿瘤，可以选择病毒产生的抗原作为新抗原，但病毒介导的肿瘤仅占肿瘤中较小部分，因此对大部分肿瘤不适用；二是突变来源，由于肿瘤组织突变率一般高于正常组织，因此可能产生肿瘤特异性突变蛋白成为新抗原，这是目前新抗原的主流鉴定方式；三是特异性表达来源，在近期的此篇文献中：

Laumont, C.M., Vincent, K., Hesnard, L., Audemard, É., Bonneil, É., Laverdure, J.P., Gendron, P., Courcelles, M., Hardy, M.P., Côté, C. and Durette, C., 2018. Noncoding regions are the main source of targetable tumor-specific antigens. Science translational medicine, 10(470), p.eaau5516.

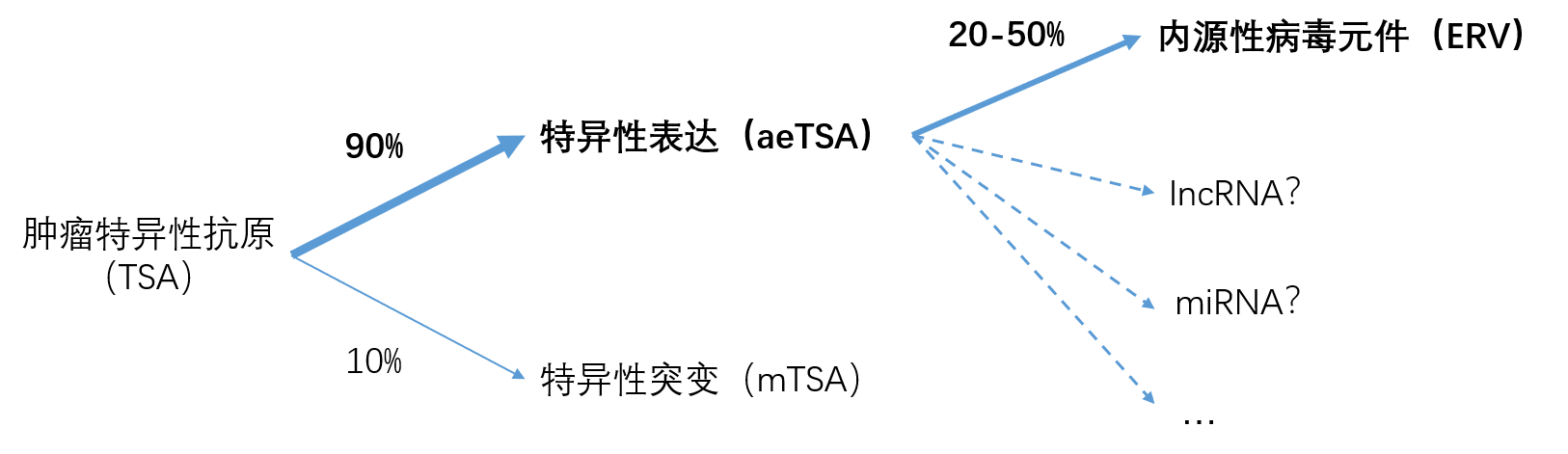
作者提出实际肿瘤特异性抗原仅有少部分来自于特异性突变，而高达90%TSA来源于不存在突变，但在肿瘤中产生特异性非正常表达的蛋白。而这些特异性表达抗原的主要来源之一为内源性病毒元件（ERV），见图 1。

图 1 肿瘤特异性抗原分类

基于此思路，本项目中我们尝试基于人类内源性病毒元件（HERV）来筛选肿瘤新抗原。

# 项目总结

本项目为从人体3173个内源性病毒（HERV）中筛选出潜在的肿瘤新抗原，筛选原则包括评估HERV及所筛选多肽精细区域在正常/肿瘤组织中的差异表达，评估相关多肽与HLA0201的结合力，以及确认相关多肽在正常组织中无其它来源表达或仅有低表达。基于以上原则综合考量，共筛选出30条多肽推荐可进行下一步免疫原性验证实验。

简要筛选流程见图 2，筛选条件及证据总结见下方表格 1，多肽详细信息见附表3。 图片包含 文字, 地图

描述已自动生成

图 2简要筛选流程图

表格 1：筛选证据总结

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 在肿瘤/正常组织中有差异表达 | | 在正常组织中无其它来源表达或低表达 | | 结合力 | |
|  | RNA证据1 | 多肽证据IEDB2 | RNA证据3 | 多肽证据UniPort4 | 预测证据5 | 质谱证据6 |
| YMRTLLDSI | √ | - | √ | X | √ | √ |
| GQVPLNPFSFTL | √ | - | √ | X | √ | √ |
| KELQLFSV | √ | - | √ | - | √ | √ |
| QLFSVIVHL | √ | - | √ | - | √ | - |
| VMVKHLILA | √ | - | √ | - | √ | - |
| SLWNSPVFV | √ | - | √ | - | √ | - |
| LLFPHPNLLSL | √ | - | √ | - | √ | - |
| YLFSESVYL | √ | - | √ | - | √ | - |
| FLHPDLLSL | √ | - | √ | - | √ | - |
| YLSSESVYL | √ | - | √ | - | √ | - |
| SLLPGEPLQKV | √ | - | √ | - | √ | - |
| KLFGQKGYRV | √ | - | √ | - | √ | - |
| RLLQLYPLFV | √ | - | √ | - | √ | - |
| FLYPKSDSV | √ | - | √ | - | √ | - |
| YLTSESVYL | √ | - | √ | - | √ | - |
| LLFTLPVYTV | √ | - | √ | - | √ | - |
| FLYPKSEAFRL | √ | - | √ | - | √ | - |
| TLLPNPNPPL | √ | - | √ | - | √ | - |
| LVFPHLNPQV | √ | - | √ | - | √ | - |
| YLVILLFTV | √ | - | √ | - | √ | - |
| SLLPFSLTQSL | √ | - | √ | - | √ | - |
| SLHTDVHEI | √ | - | √ | - | √ | - |
| SLYWGQVAL | √ | - | √ | - | √ | - |
| YIQEFRHLTL | √ | - | √ | - | √ | - |
| HISPFLVSV | √ | - | √ | - | √ | - |
| HLSPFPHTA | √ | - | √ | - | √ | - |
| TLTDDIPPL | √ | - | √ | - | √ | - |
| FLLLYTLKV | √ | - | √ | - | √ | - |
| YLMCLLLKL | √ | - | √ | - | √ | - |

\*KELQLFSVIVHL, GKELQLFSVIVHL以短肽KELQLFSV形式进行在正常组织中无其它来源表达或低表达的RNA证据分析。

证据定义：

1. √：该多肽来源HERV在肿瘤/正常组织中存在显著差异表达（pan-cancerlog2FC>4 且 -log10pvalue>50，表达量正则化公式为，p-value以双边的t-test检验计算得到；COADlog2FC>4 且 -log10pvalue>20，计算方式同上），且该多肽片段精细区域在肿瘤/正常组织中存在显著差异表达（仅考虑COAD，log2FC>2，表达量由匹配上的read条数表示）；对于存在证据2的多肽，该多肽片段区域在肿瘤/正常组织中存在显著差异表达（仅考虑COAD）
2. √：该多肽在IEDB数据库中仅存在于肿瘤样本中

-：该多肽在IEDB数据库中不存在（由于IEDB数据库鉴定多肽时通常使用UniPort建库，因此非UniPort中的HERV多肽一般无法被鉴定到）

1. √：所有可能表达出该多肽的RNA序列组合在121个pan-cancer正常组织样本中非HERV来源的平均总表达量<1 reads
2. √：该多肽存在UniPort中，但唯一来源为HERV；或该多肽存在UniPort中，但其来源基因TPM<1（TCGA和GTEx样本）;在H161细胞系中？？

-：该多肽不存在于UniPort数据中

1. √：NetMHCpan4.0 预测为“结合”（strong bind或weak bind）且PSSM aff<500的多肽，或EPIC分值>0.5的多肽
2. √：在肿瘤细胞系质谱中可被鉴定到，鉴定通过标准：

X:在质谱中可被鉴定到，但未通过

-：在质谱中未被鉴定到

**综上**，推荐以上**29**条多肽进行免疫原性实验。筛选步骤细节见后附说明。

其它结论：

1. 从HERV的泛癌肿表达分析中发现结直肠癌中HERV表达较特殊，后续如需验证HERV来源新抗原，可考虑选择其它癌肿如肾透明细胞癌。
2. IEDB数据库对HERV来源新抗原筛选帮助有限，推测原因为IEDB多肽鉴定所使用数据库以UniPort为主，HERV来源多肽通常不在鉴定范围内，因此如需使用IEDB，可能需获取原始质谱数据重新建库鉴定。

# 附录

## 方法详细流程及说明

完整筛选过程见图 3，或见附件4：图片包含 文字, 地图

描述已自动生成

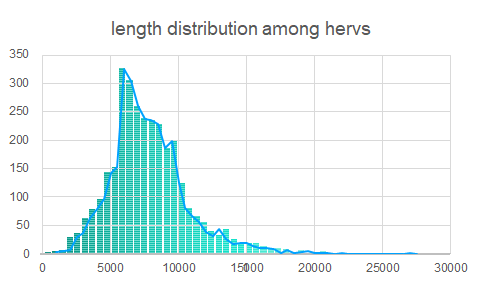
图 3完整筛选步骤流程图

1. HERV来源

我们选取了3173个HERV，其位置、分类及结构信息来自于此篇文献（2016），是我们目前找到的最新关于HERV的分类文章，HERV长度分布见图 4a：

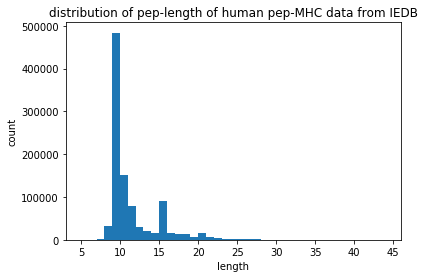
Vargiu L, Rodriguez-Tomé P, Sperber GO, Cadeddu M, Grandi N, Blikstad V, Tramontano E, Blomberg J. Classification and characterization of human endogenous retroviruses; mosaic forms are common. Retrovirology. 2016 Dec;13(1):7.

图 4 HERV长度分布及IEDB多肽长度分布



**数量**

**长度（bp）**



**a.**

**b.**

Figure . 3173条HERV长度频数分布（a）IEDB数据库中多肽片段长度频数分布（b）

1. 表达量分析
   1. HERV表达量

HERV表达量由HervQuant软件分析RNAseq数据获得，软件来源文献：

Smith CC, Beckermann KE, Bortone DS, De Cubas AA, Bixby LM, Lee SJ, Panda A, Ganesan S, Bhanot G, Wallen EM, Milowsky MI. Endogenous retroviral signatures predict immunotherapy response in clear cell renal cell carcinoma. The Journal of clinical investigation. 2018 Oct 2;128(11).

HervQuant对版本和参数较敏感，所有参数已与作者确认，并在2个样本中重复出作者结果。另除HervQuant之外，我们也选择了另一内源性病毒表达量分析软件ViGen进行对比，两者结果显著相关，在后续分析中仍使用HervQuant。

STAR参数设置：  
--outFilterMultimapNmax 10   
--outFilterMismatchNmax 7

上述HervQuant文献中有8678个TCGA肿瘤样本内源性病毒表达分析结果，我们额外加入了127个正常组织样本进行HERV差异表达筛选，得到66条存在显著差异表达的HERV（log2FC > 4, -log10pvalue > 50 ），见图 5，其中t-test使用了 Welch’s t-test。此外，使用非参数检验Mann-Whitney rank test所获结果一致。

TCGA normal samples : 127（25 types of cancers）

TCGA cancer samples: 8678（23 types of tissue cell）

由于后续实验将在结直肠癌中展开，因此我们也分析了HERV在结直肠癌样本中的差异表达，得到335条HERV（log2FC > 4, -log10pvalue > 20），见图 6。

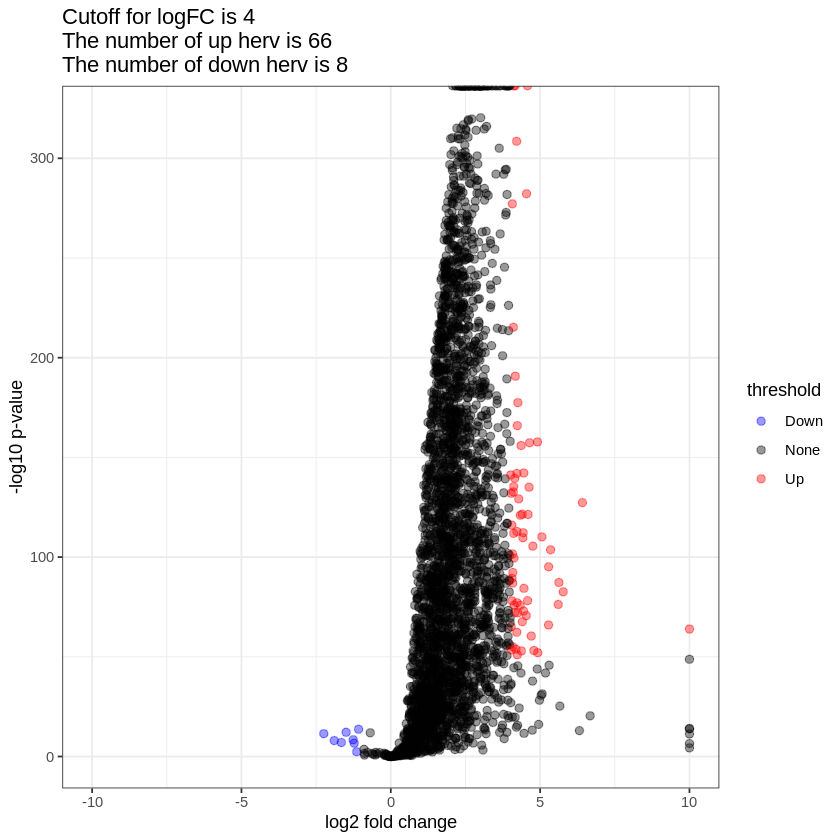
合并两者结果，在泛肿瘤样本和COAD样本中均有显著差异表达的HERV数量为13条，所有HERV表达量结果详见附表

图 5：TCGA-pncancer肿瘤/正常样本HERV表达火山图。红色区域的点为筛选出的66个肿瘤中显著高表达的HERV。

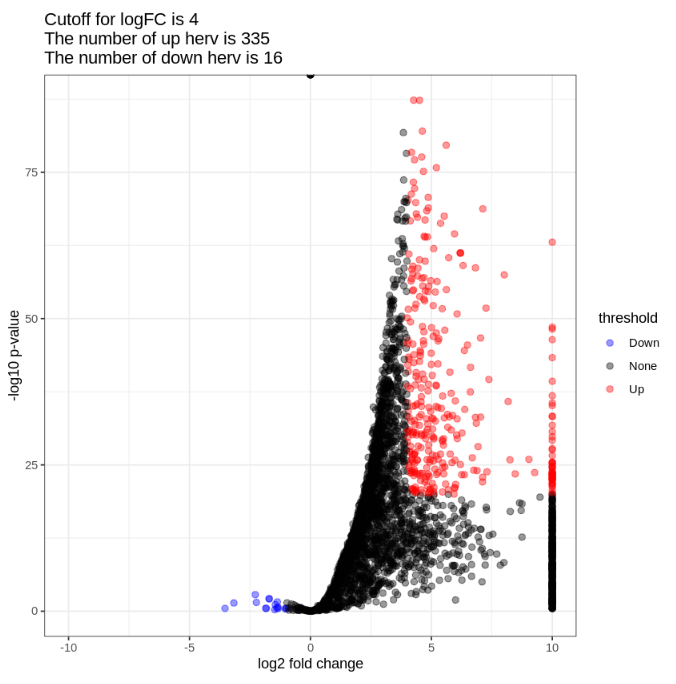


图 6：TCGA-COAD肿瘤/正常样本HERV表达火山图。红色区域的点为筛选出的335个肿瘤中显著高表达的HERV

* 1. Peptide site-wise精细区域表达量

前面2a）步骤中，我们考虑每中herv区域整体表达量在肿瘤和正常样本中是否存在显著差异，然而这些区域长度为最大约28000bp，但其中每个位点不一定均匀表达。而我们的多肽长度仅为21-45bp，因此可针对这些多肽片段精细区域再次验证肿瘤与非肿瘤样本中的表达量。

对上一步所选存在泛肿瘤差异表达的66个HERV，用paired的结直肠癌样本call出每个位点的tumor和normal的表达量。得到显著高表达且在肿瘤中平均表达量为0的位点之后（log2FC>1且 avg\_tumor>0，图 7），把连续的位点连成mRNA序列，得到268条mRNA区域。用滑动窗口的方式切出每个mRNA中正反链上所有长度为21-45bp的片段（即翻译后7-15AA），翻译成为多肽片段（若遇到终止子则停止翻译，仅保留长度满足7-15AA的片段。从这些区域中共切出525,130条多肽，其中与COAD差异表达的13条HERV区域重合的共83,451条。

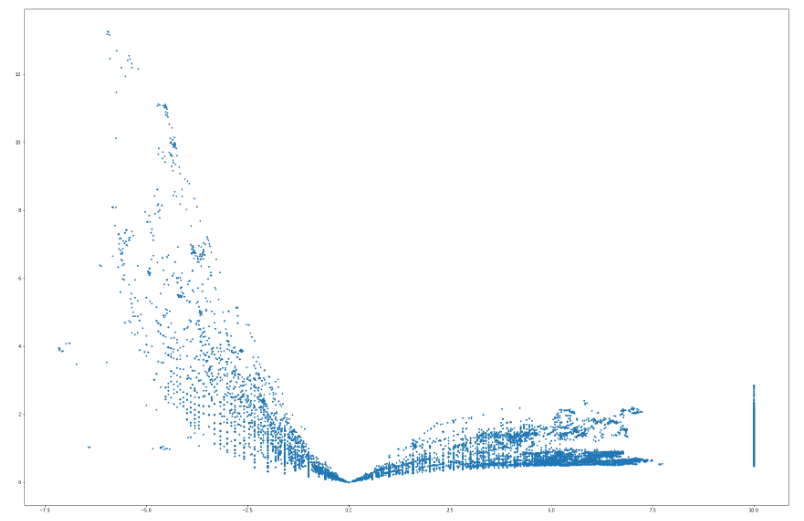


图 7：66HERV所有位点差异表达。其中每个点是HERV上的每个位点，保留肿瘤表达量>0且log2FC>1的点（计算方法同上）。

1. 结合力预测

我们使用NetMHCpan，PSSM以及EPIC预测它们与HLA0201的结合力，筛选出NetMHCpan4.0 预测为“结合”（strong bind或weak bind）且PSSM aff<500的多肽，或EPIC分值>0.5的多肽，共943条。

1. 排除正常组织表达

随后为了确认这些多肽是否在正常组织中是否无表达或仅有低表达，以及是否有HERV之外的正常基因表达来源，我们从RNA和多肽层面分别进行确认。在多肽层面，我们比较这些多肽是否在UniPort中存在，若不存在，则说明该多肽仅来源于HERV，若存在，则需验证该HERV的其它来源基因是否在正常组织中低表达。在RNA层面，我们考虑可能翻译出该多肽的所有RNA序列组合，在正常样本RNAseq数据中寻找匹配的reads，确认其在RNA层面是否不表达或低表达。

* 1. UniPort

UniPort是一个集合了人体正常表达的蛋白質序列和功能信息資源。从数据库获取人体蛋白质序列及序列注释信息，将分析得到的多肽序列与Uniport数据集比较，得到条数。如果在Uniport数据集并未发现该多肽序列，则认为该多肽在正常人体内不表达。Uniport数据集可能包含少量HERV和抗生素等序列杂质，可通过查看序列ID的注释信息判断来源基因。

* 1. RNAseq

最终分析得到的候选多肽，依照氨基酸简并性转变为可能的核苷酸序列，并在121个TCGA正常样本的bam文件直接进行搜索，查找统计到个体可能表达候选多肽的数目（排除herv区域的表达），最终如果平均在每个正常样本的表达不到1，则认为该多肽在正常人体内不表达。正常样本指肿瘤病人的正常组织取样测序样本。

1. 质谱
2. IEDB筛选
   1. 多肽库构建

用滑动窗口的方式切出每个HERV中正反链上所有长度为21-45bp的片段（即翻译后7-15AA），翻译成为多肽片段（若遇到终止子则停止翻译，仅保留长度满足7-15AA的片段），共得到153,579,605条独特多肽。

* 1. 多肽库与IEDB中多肽比对

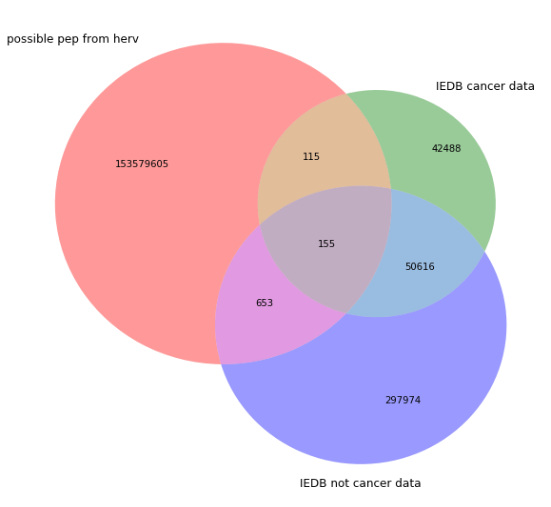
所有多肽片段与IEDB中肿瘤/非肿瘤多肽数据进行比对，筛选出仅在肿瘤数据中表达的多肽共115条，多肽与IEDB重合情况见图 8。

所获得115条多肽片段来自于215条HERV，对应关系如图 9a与图 9b所示，每条多肽对应HERV数量在1-93之间，每条HERV可包含多肽数量在1-5条。对应93条HERV的多肽为HHVSQDGLDLL。

所获115条多肽并非每条都可由HERV编码区表达得到，其中31条位于HERV编码区且同向（含两端LTR、gag、pro、pol和env），即可能来自于HERV编码片段，其余多肽30条位于编码区但不同向，54条位于非编码区（图 9c，图 10）。

图 8：HERV多肽与IEDB数据集重合情况

**Candidates**



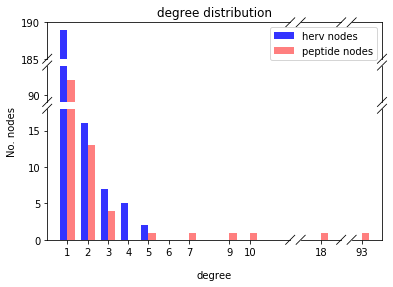
**IEDB cancer data**

**IEDB non-cancer data**

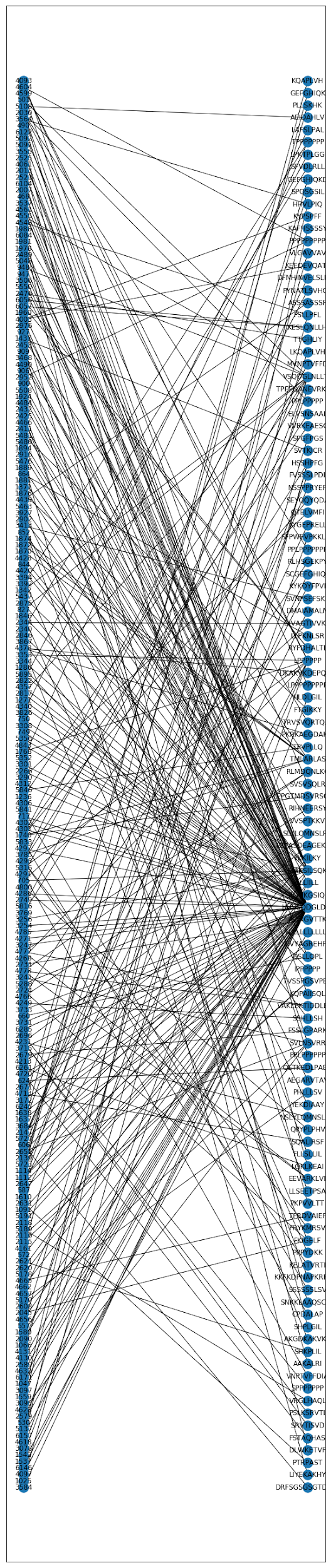
**All possible HERV peptides**

Figure ．所有内源性病毒可能表达的7-15AA多肽（红色圈）与IEDB中肿瘤样本多肽（绿色圈），以及IEDB中所有非肿瘤样本多肽（蓝色圈，含不确定类型样本）的交集。红色圈出部分115条HERV多肽仅在肿瘤中有表达，其它样本中无表达，即成为候选多肽进行下一步筛选。

图 9：IEDB中115条仅在肿瘤中表达的HERV基本情况



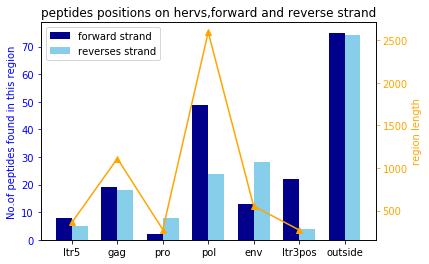
**a.**



**b.**

**215 HERV**

**115 peptides**



**c.**

**HHVSQDGLDLL**

Figure ．a.HERV与多肽对应频数分布图，蓝色柱为每条HERV对应从IEDB筛选出的115条多肽数量（1至5条），红色柱为每条筛选出的115条多肽可来源的HERV数量（1至93条）。b. HERV与多肽对应关系图，左侧列为215条HERV，右侧列为115条多肽。c. IEDB中筛选出的115条多肽在HERV结构区域的分布（由于每条多肽可来自于不同HERV，会被多次计算，因此此图中总频数大于115）。

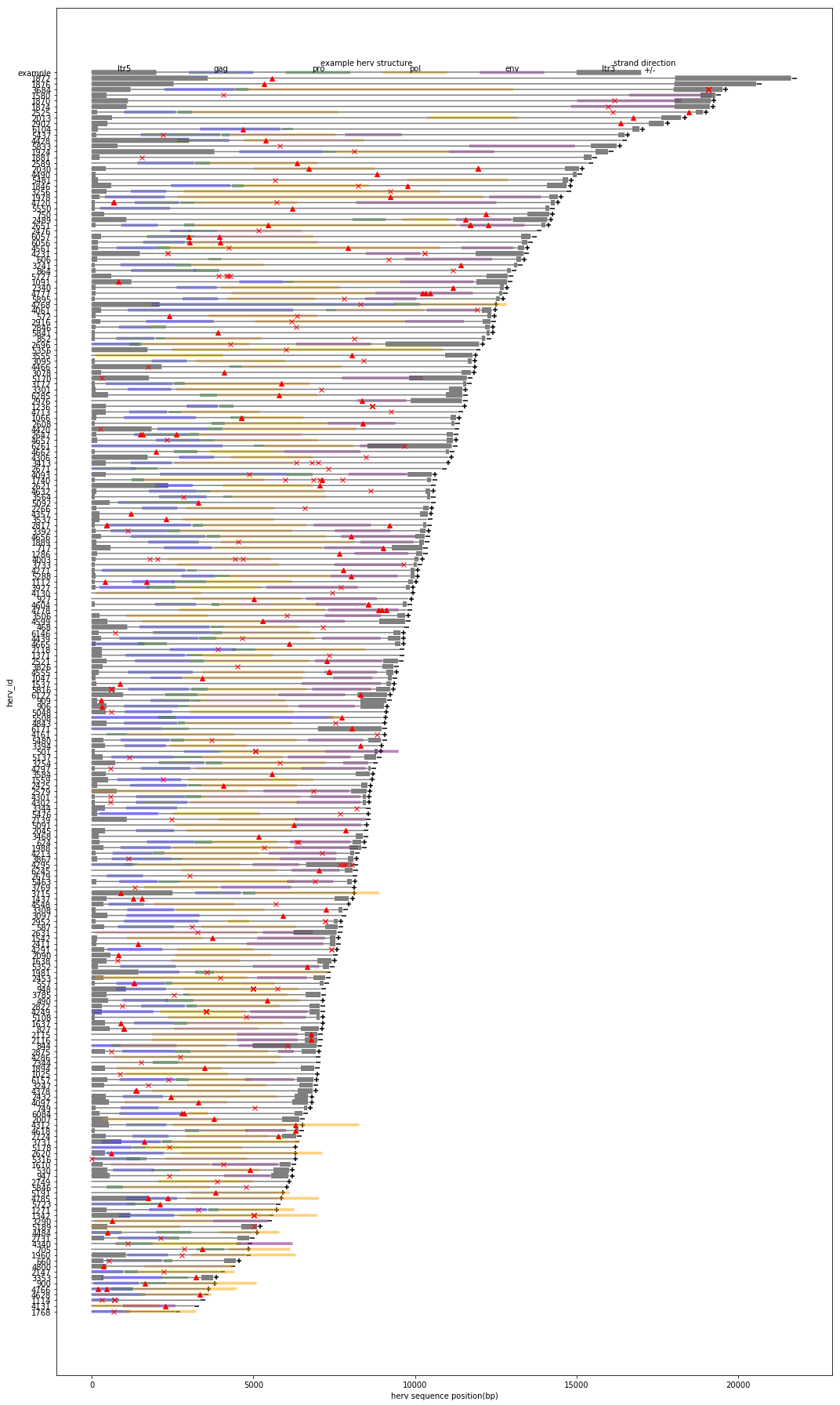


图 10：115条多肽在215条HERV结构中分布情况。灰色区域为HERV两端LTR，蓝色、绿色、黄色、红色区域分别为gag、pro、pol和env蛋白编码区。右端箭头方向代表HERV方向。红色三角和叉分别代表正向和反向链上的多肽位置。

* 1. 多肽表达量筛选

此步骤考虑在IEDB中筛选到的115条多肽序列在COAD中的表达量，在36组结直肠癌病人的成对数据中求出每一条多肽的平均表达量，考虑肿瘤和正常细胞中的平均表达量差异作火山图。筛选显著高表达的多肽序列（log2FC>2，图 11）。

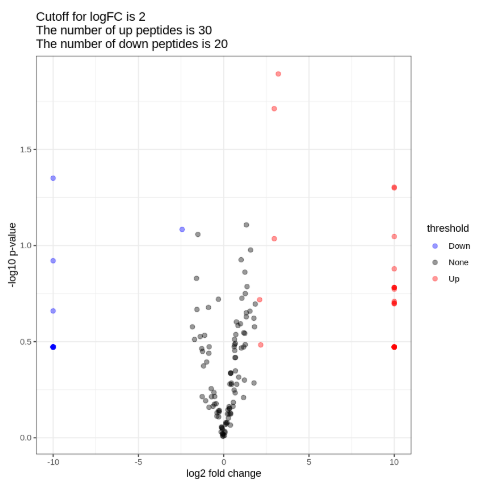


图 11：115条多肽COAD中表达差异。TCGA中36对结直肠癌肿瘤/正常样本多肽平均表达量的差异表达火山图，图中每个点代表一组多肽及可能来源的HERV区域组合（即若一条多肽可能被多个HERV区域表达，会以多个点表示）。

1. 文献中来自HERV的潜在新抗原

Laumont CM, Vincent K, Hesnard L, Audemard É, Bonneil É, Laverdure JP, Gendron P, Courcelles M, Hardy MP, Côté C, Durette C. **Noncoding regions are the main source of targetable tumor-specific antigens.** Science translational medicine. 2018 Dec 5;10(470):eaau5516.

该文献从4个白血病（B-ALL）和3个肺癌患者中鉴定出4条可能来自于内源性病毒区域的多肽，其中3条在我们使用的HERV reference所产生的多肽库中未找到，在原文中来源推测除ERV外也可能为intron或UTR，因此被排除。另外1条在IEDB中未找到，且在质谱中未找到。此条多肽所处HERV区域在泛肿瘤表达分析（logFC = 1.358855439，pvalue=1.29E-144）与COAD表达分析中均有显著差异表达（logFC = 2.077793308，pvalue=8.62E-18），但未达我们流程的筛选标准（logFC>4）因此被排除。

表格 2：文献来源多肽

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Sample | TSA Seq | expr (rphm) | TEC/mTEC expr(rphm) | IEDB | Position | TSA Origin | Length (aa) | MHC I | Our peptide library | HERV\_id |
| B-ALL | KILILLQSL | 22.67 | 0.34 | No | chr5:132450600-132450626 | ERE or Intron | 9 | A\*02:01 | No | -- |
| **B-ALL** | **KISLYLPAL** | **9.45** | **0.00** | **No** | **chr8:144861684-144861710** | **ERE** | **9** | **A\*02:01** | **present** | **2982（outside）** |
| B-ALL | TSFAETWMK | 7.73 | 0.00 | No | chr7:43947484-43947510 | Intron or ERE | 9 | A\*11:01 | No | -- |
| Lung tumor | SSASQLPSK | 696.73 | 0.00 | No | chr16:19430493-19430519 | ERE or 5'UTR | 9 | A\*11:01 | No | -- |

## 附件列表

附表1：HERV筛选结果

附表2：多肽详细信息

附表3：候选多肽final list，含多肽序列，多肽来源HERV，在COAD tumor/normal中精细区域表达量，来源HERV在肿瘤和正常组织中表达量（pan-cancer和COAD），三种软件预测结果，质谱结果，是否在uniport中存在，来源基因的表达量

附表4：候选多肽所有可能来源RNA序列组合编码及在121个TCGA正常样本中RNAseq过滤结果

附件4：完整筛选步骤流程图