内源性病毒分析

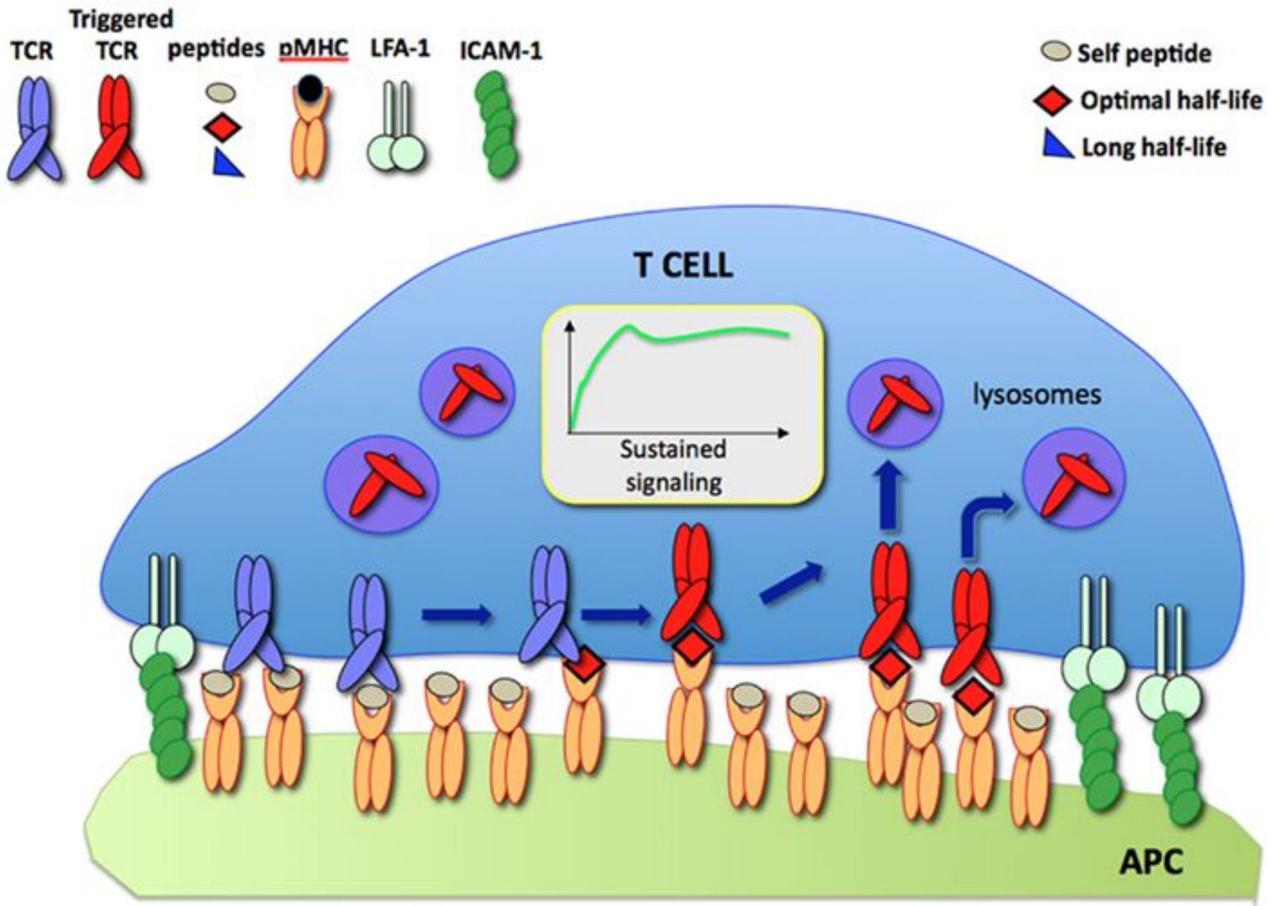
丁远彤 李华平 杨悦羽霄 2019/10/30

前情提要

herv 人类内源性病毒

antigen 抗原

peptides 多肽序列
+ HLA + TCR
- > 激活免疫反应



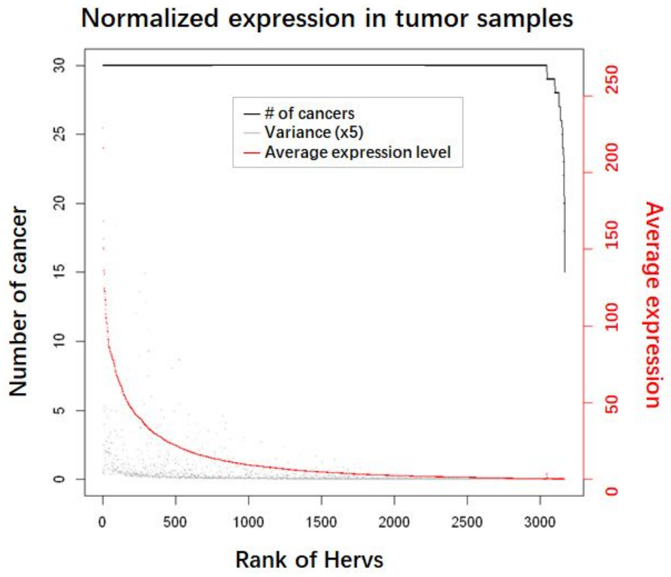
数据说明

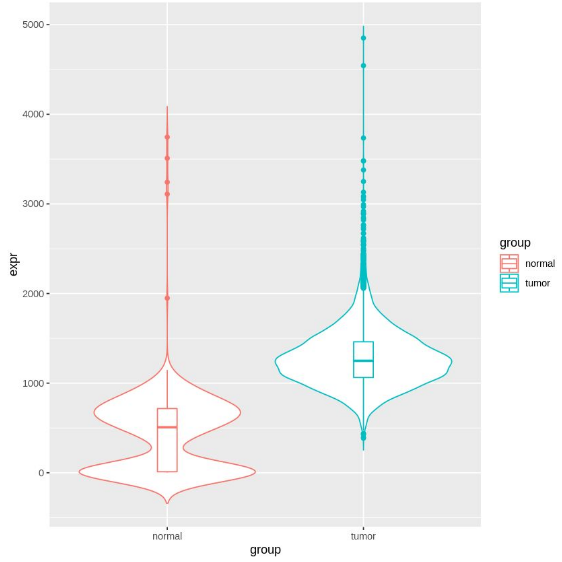
TCGA tumor 8470 样本，25 种肿瘤 JCI121476.sdt12.csv

（除去 LAML，STAD herv表达量异常低，且样本均来自其中1个或2个项目，可能由操作差异导致表达量偏低，故排除）

TCGA normal 534 样本，normalized_normal_534.csv

TCGA COAD paired-sample 32对样本，paired-samples.txt





SDT12:

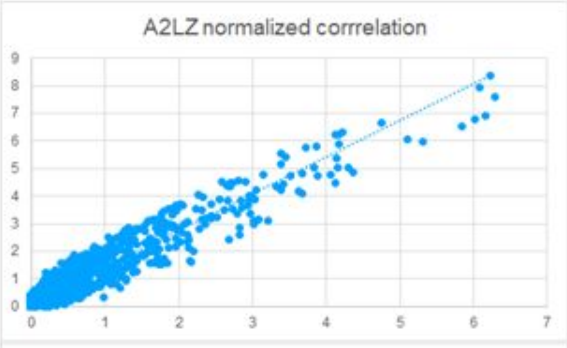
Raw expression matrices were normalized to hERV counts per million total FASTQ reads and log2 transformed，

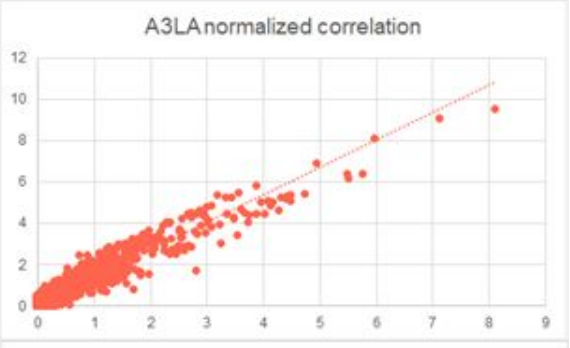
total reads 为 L， herv mapped reads 为N， 并以2为底求log，类似于TPM

$$X = log2^{(\frac{N*10^{\wedge}6}{L}+1)}$$

hervquant 效果验证及复现

- 所有参数/版本与作者确认
- 选取2个TCGA肿瘤样本复现结果
  - correlation ~ 96%
- 可复现


A2LZ normalized corrrelation

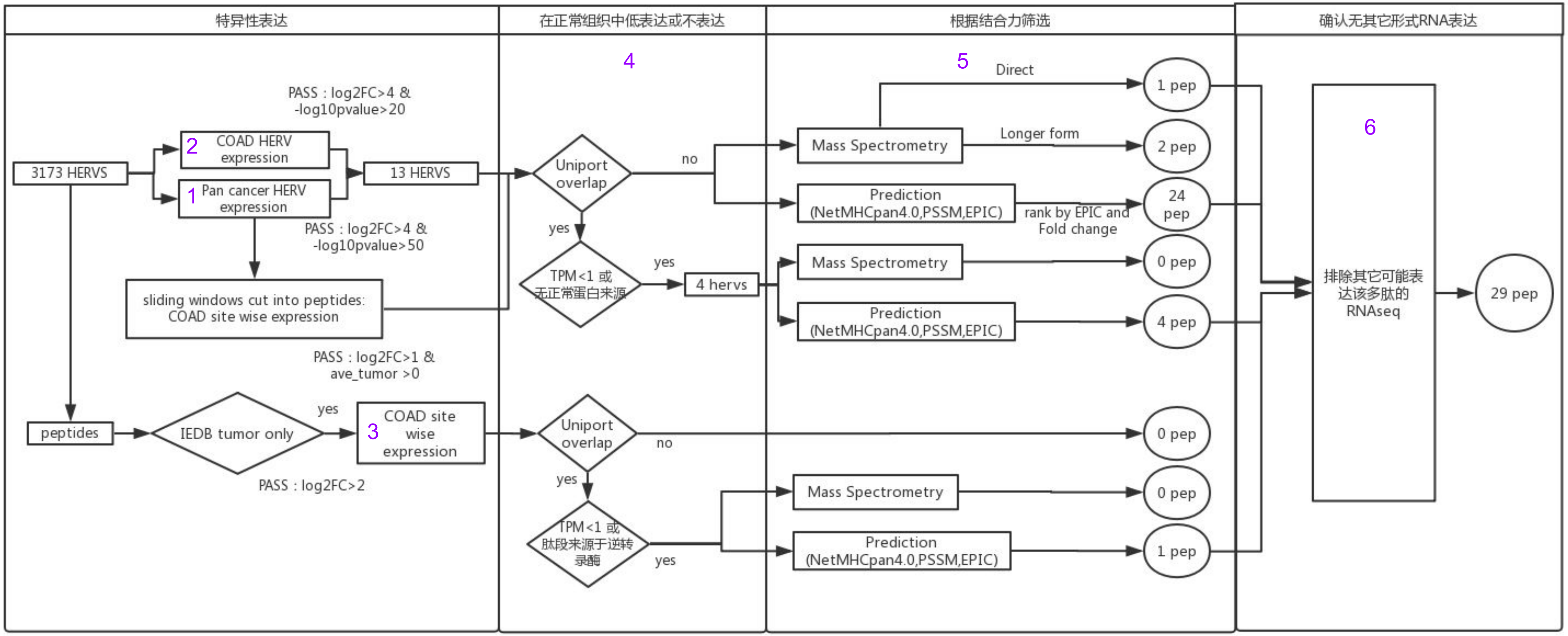

A3LA normalized correlation

Sample ID	Pearson Correlation Coefficient for normalized result	P-value	Spearman‘s rank Correlation Coefficient for normalized result	P-value
TCGA-C5-A2LZ-01A	0.995052	3.017e-07	0.7142857	0.05759
TCGA-IR-A3LA-01A	0.9631543	0.0001216	0.9940298	5.296e-07

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).

差异表达 + 特异结合 ——> 激活免疫

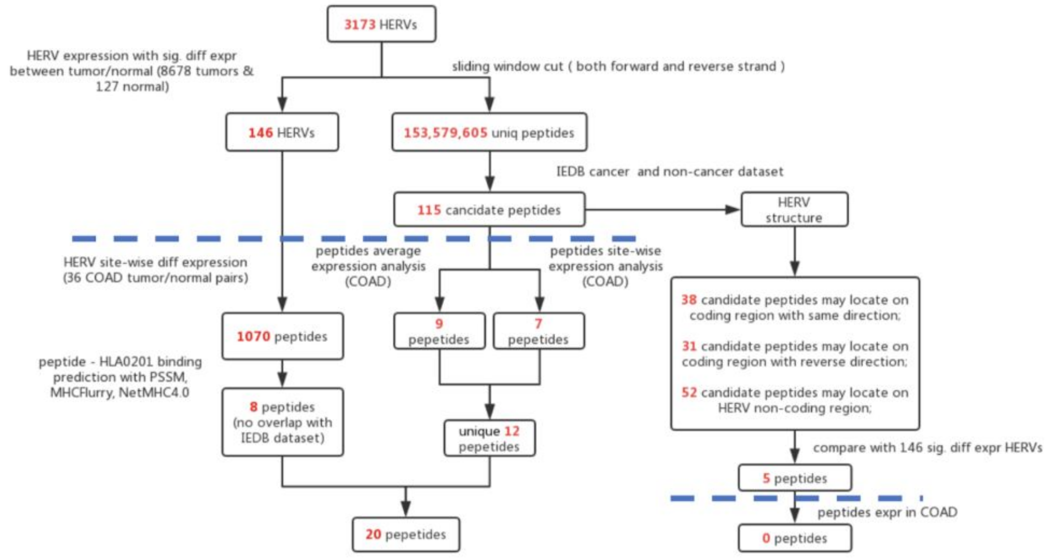
总流程思路图

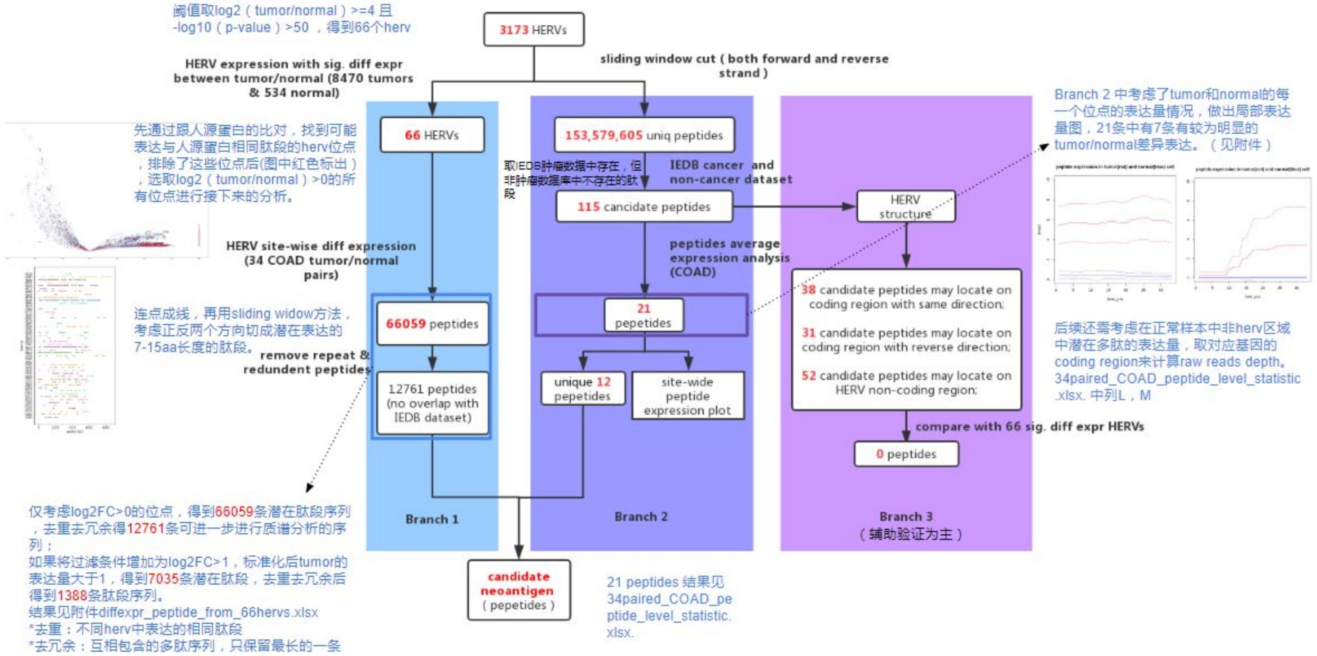


1-1.pancancer_8470_534_herv_level_statistic.csv
1-2.pancancer_highlydiff_66herv.csv
2-1.COAD_281_30_herv_level_statistic.csv
2-2.COAD_highlydiff_335herv.csv
3.115_peptide_IEDB.xlsx

4.diffexpr_peptide_from_66hervs_nofilter.xlsx
5.SelectedpepList.xlsx

6.diffexpr_peptide_from_66hervs_nofilter.xlsx


3173 HERVs
HERV expression with sig. diff expr between tumor/normal (8678 tumors & 127 normal)
sliding window cut (both forward and reverse strand)
146 HERVs
153,579,605 uniq peptides
IEDB cancer and non-cancer dataset
115 cancidate peptides
HERV structure
HERV site-wise diff expression (36 COAD tumor/normal pairs)
peptides average expression analysis (COAD)
peptides site-wise expression analysis (COAD)
1070 peptides
9 pepetides
7 pepetides
38 candidate peptides may locate on coding region with same direction;
31 candidate peptides may locate on coding region with reverse direction;
52 candidate peptides may locate on HERV non-coding region;
peptide - HLA0201 binding prediction with PSSM, MHCFlurry, NetMHC4.0
8 peptides (no overlap with IEDB dataset)
unique 12 pepetides
compare with 146 sig. diff expr HERVs
5 peptides
peptides expr in COAD
20 pepetides
0 peptides


3173 HERVs
HERV expression with sig. diff expr between tumor/normal (8470 tumors & 534 normal)
sliding window cut (both forward and reverse strand)
66 HERVs
153,579,605 uniq peptides
IEDB cancer and non-cancer dataset
115 cancidate peptides
HERV structure
HERV site-wise diff expression (34 COAD tumor/normal pairs)
peptides average expression analysis (COAD)
66059 peptides
21 pepetides
remove repeat & redundent peptides
12761 peptides (no overlap with IEDB dataset)
unique 12 pepetides
site-wide peptide expression plot
38 candidate peptides may locate on coding region with same direction;
31 candidate peptides may locate on coding region with reverse direction;
52 candidate peptides may locate on HERV non-coding region;
compare with 66 sig. diff expr HERVs
0 peptides
Branch 1
Branch 2
Branch 3
（辅助验证为主）
candidate neoantigen
(pepetides)

阈值取log2（tumor/normal）>=4 且 -log10（p-value）>50 ，得到66个herv

先通过跟人源蛋白的比对，找到可能表达与人源蛋白相同肽段的herv位点，排除了这些位点后(图中红色标出），选取log2（tumor/normal）>0的所有位点进行接下来的分析。

连点成线，再用sliding widow方法，考虑正反两个方向切成潜在表达的7-15aa长度的肽段。

取IEDB肿瘤数据中存在，但非肿瘤数据库中不存在的肽段

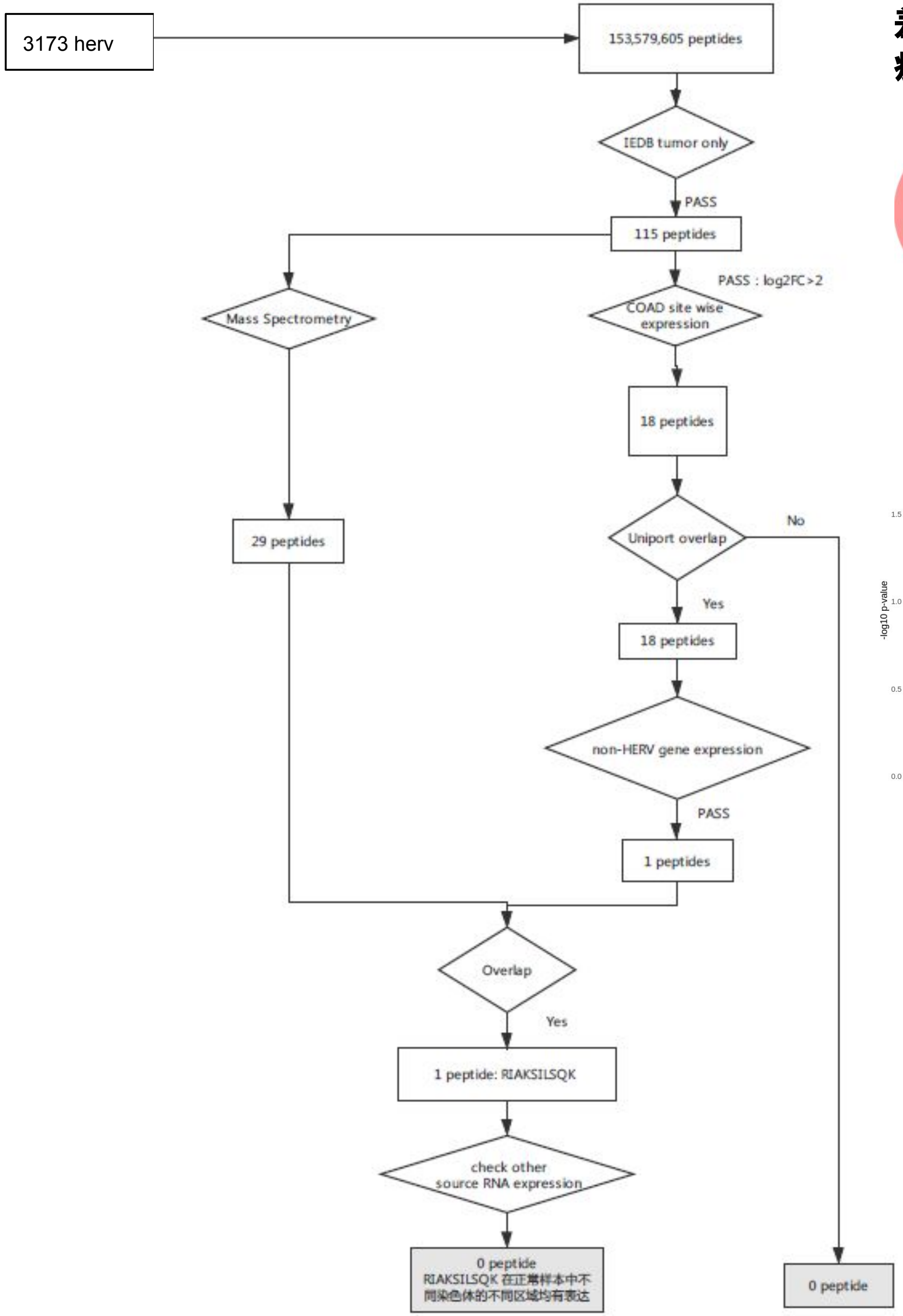
仅考虑log2FC>0的位点，得到66059条潜在肽段序列，去重去冗余得12761条可进一步进行质谱分析的序列；
如果将过滤条件增加为log2FC>1，标准化后tumor的表达量大于1，得到7035条潜在肽段，去重去冗余后得到1388条肽段序列。
结果见附件diffexpr_peptide_from_66hervs.xlsx
*去重：不同herv中表达的相同肽段
*去冗余：互相包含的多肽序列，只保留最长的一条

21 peptides 结果见34paired_COAD_peptide_level_statistic.xlsx.

Branch 2 中考虑了tumor和normal的每一个位点的表达量情况，做出局部表达量图，21条中有7条有较为明显的tumor/normal差异表达。（见附件）

后续还需考虑在正常样本中非herv区域中潜在多肽的表达量，取对应基因的coding region来计算raw reads depth。34paired_COAD_peptide_level_statistic.xlsx. 中列L，M

分析流程图



3173 herv

153,579,605 peptides

IEDB tumor only

PASS

115 peptides

Mass Spectrometry

29 peptides

PASS : log2FC>2

COAD site wise expression

18 peptides

Uniport overlap

No

Yes

18 peptides

non-HERV gene expression

PASS

1 peptides

Overlap

Yes

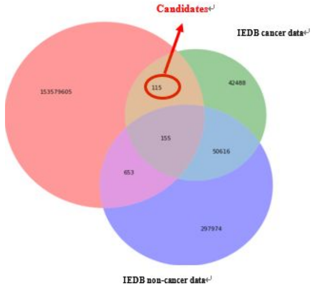
1 peptide: RIAKSILSQK

check other source RNA expression

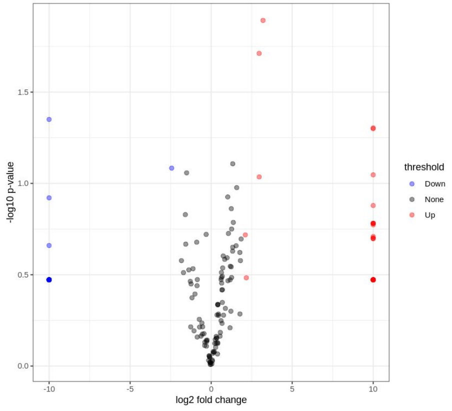
0 peptide
RIAKSILSQK 在正常样本中不同染色体的不同区域均有表达

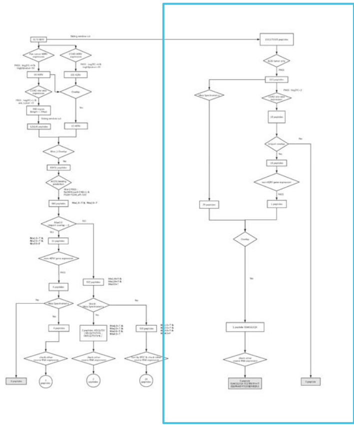
0 peptide

差异表达 + 特异结合 ——> 激活免疫



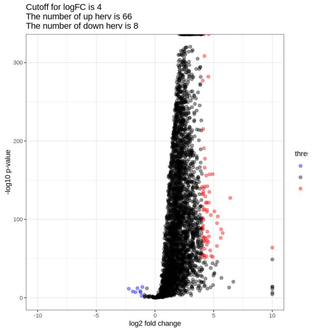
Cutoff for logFC is 2
The number of up peptides is 30
The number of down peptides is 20

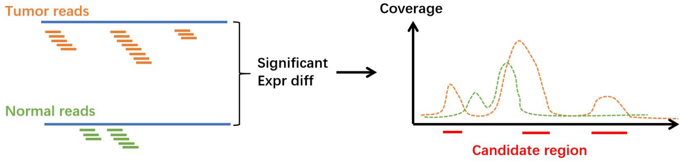




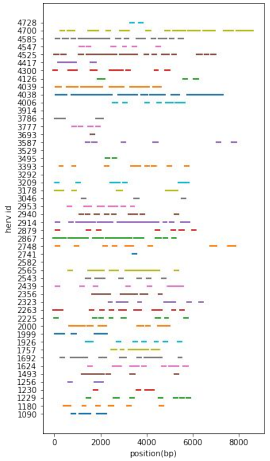
BGI华大

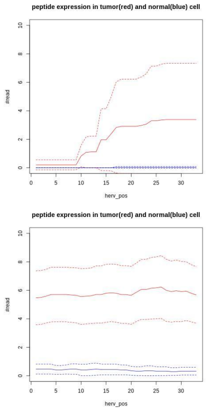
分析流程图

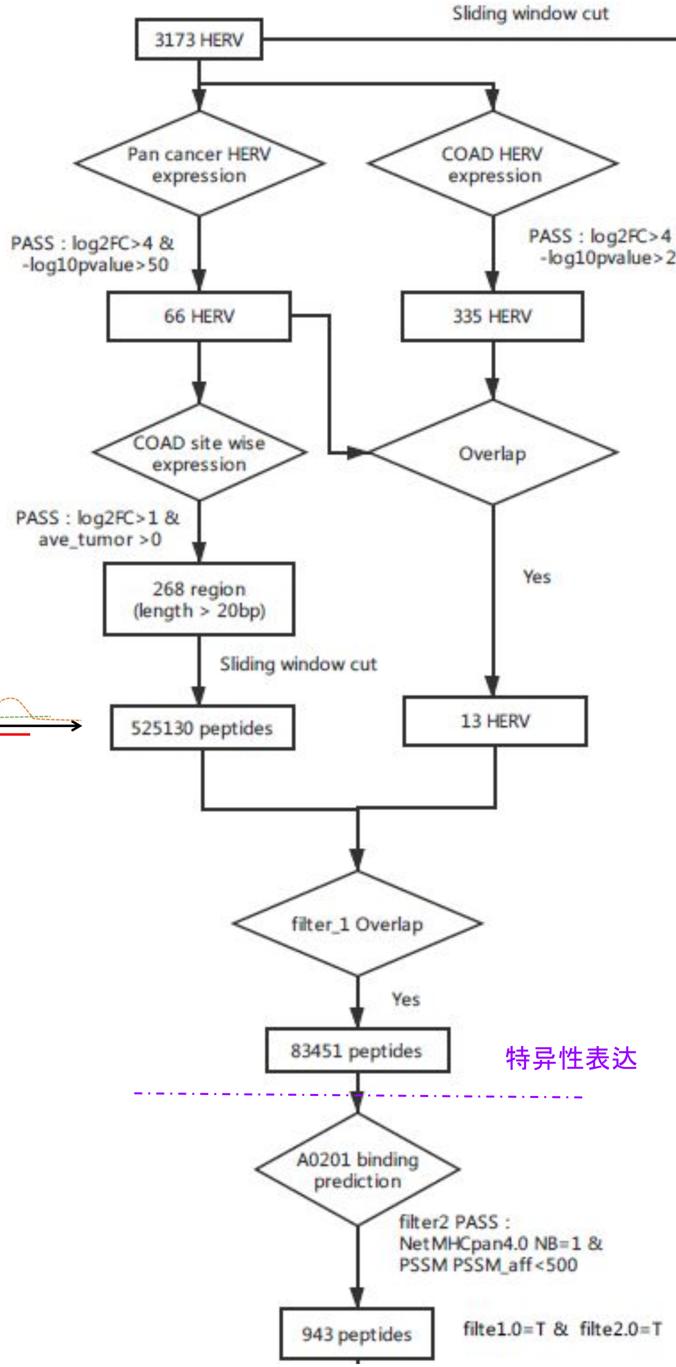




Need tumor RNAseq raw data
Testing with colon cancer samples first







3173 HERV

Sliding window cut

Pan cancer HERV expression

COAD HERV expression

PASS：log2FC>4 & -log10pvalue>50

PASS：log2FC>4 & -log10pvalue>20

66 HERV

335 HERV

COAD site wise expression

Overlap

PASS：log2FC>1 & ave_tumor >0

268 region (length > 20bp)

Yes

Sliding window cut

525130 peptides

13 HERV

filter_1 Overlap

Yes

83451 peptides

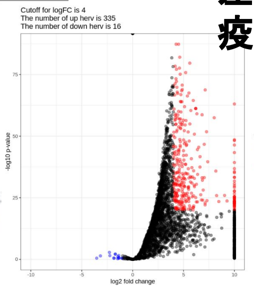
特异性表达

A0201 binding prediction

filter2 PASS：NetMHCpan4.0 NB=1 & PSSM PSSM_aff<500

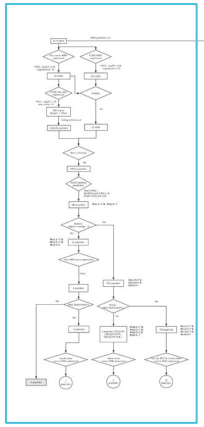
943 peptides

filte1.0=T & filte2.0=T

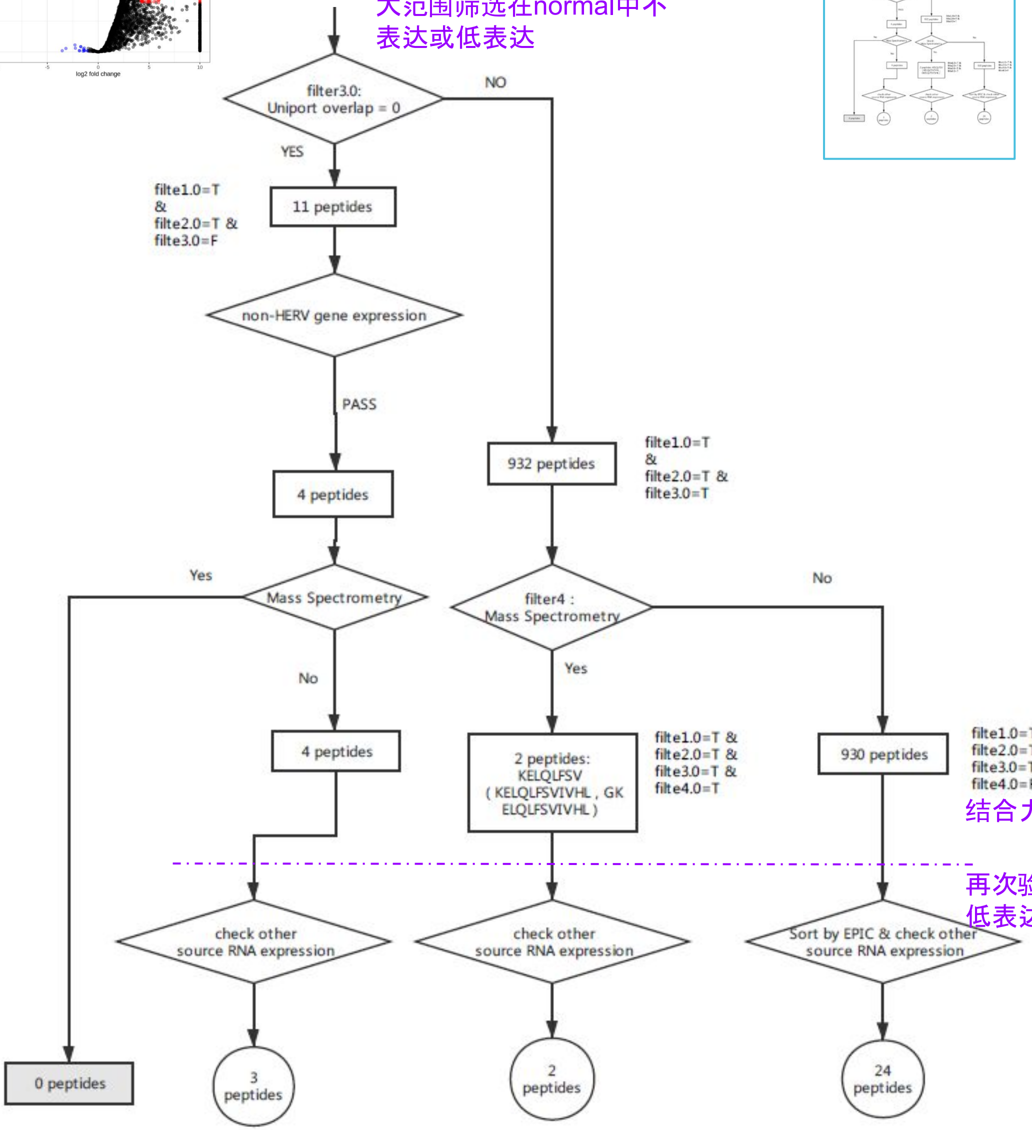


差异表达 + 特异结合 ——> 激活免疫

*normal指TCGA的normal样本集



大范围筛选在normal中不表达或低表达



filter3.0: Uniport overlap = 0

NO

YES

filte1.0=T & filte2.0=T & filte3.0=F

11 peptides

non-HERV gene expression

PASS

4 peptides

932 peptides

filte1.0=T & filte2.0=T & filte3.0=T

Yes

Mass Spectrometry

filter4：Mass Spectrometry

No

No

Yes

4 peptides

2 peptides: KELQLFSV (KELQLFSVIVHL , GK ELQLFSVIVHL)

filte1.0=T & filte2.0=T & filte3.0=T & filte4.0=T

930 peptides

filte1.0=T & filte2.0=T & filte3.0=T & filte4.0=F

结合力预测筛选

check other source RNA expression

check other source RNA expression

Sort by EPIC & check other source RNA expression

再次验证在normal低表达

0 peptides

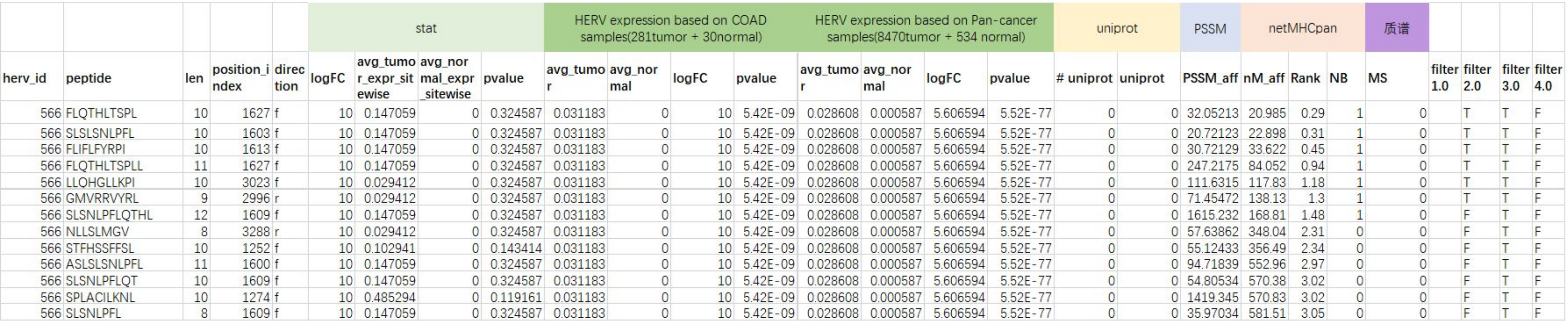
3 peptides

2 peptides

24 peptides

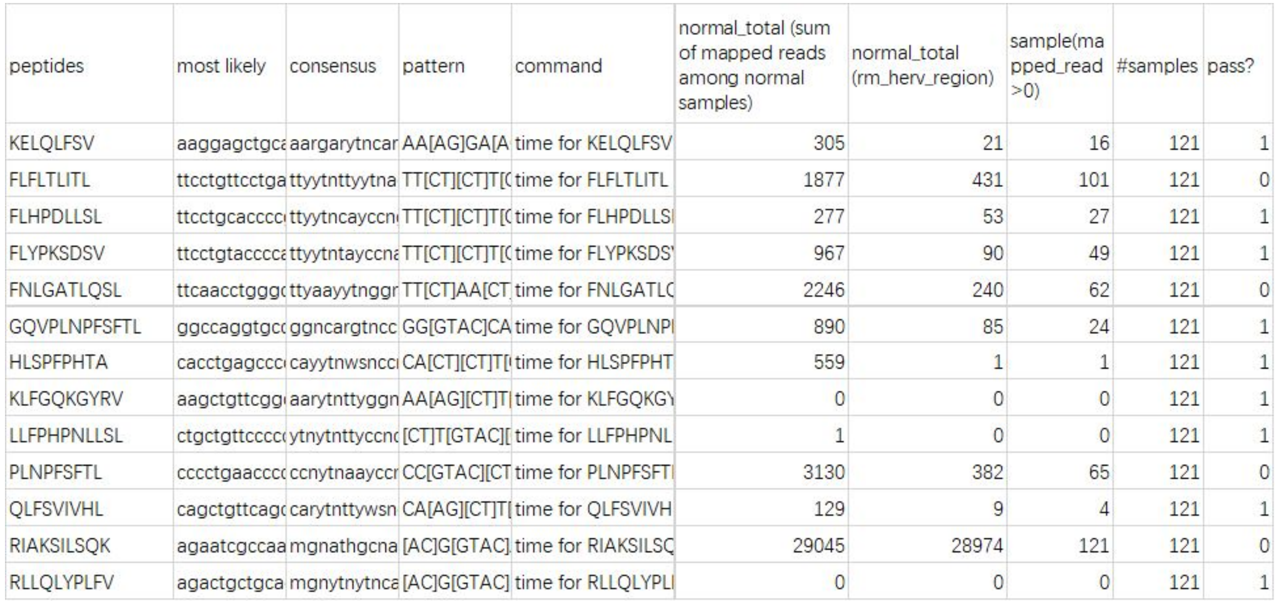
BGI华大

4.diffexpr_peptide_from_66hervs_nofilter.xlsx

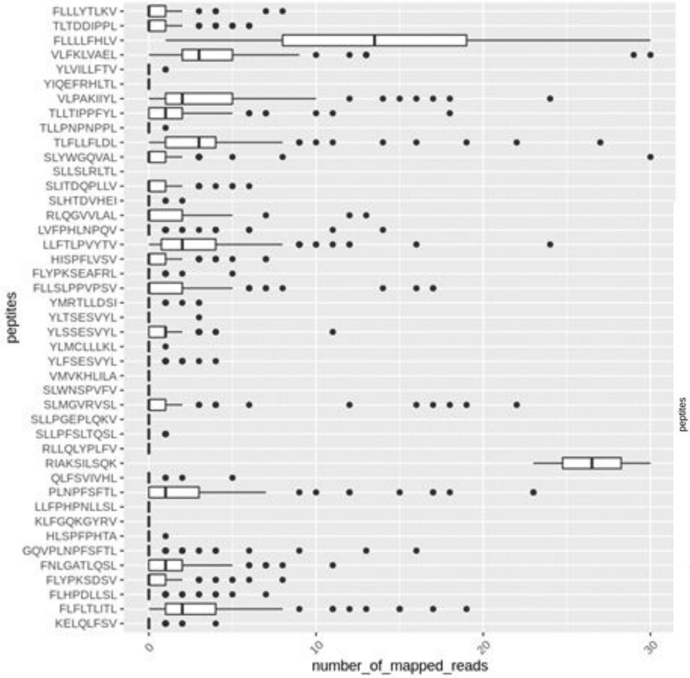


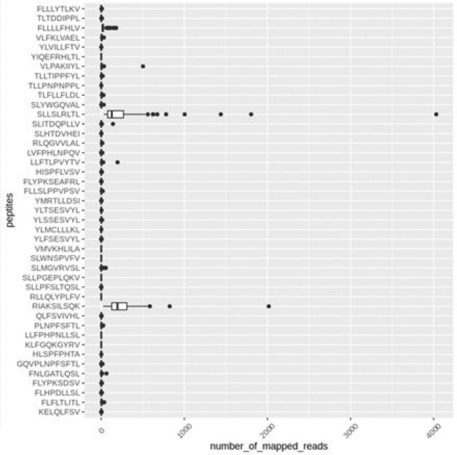
					stat				HERV expression based on COAD samples(281tumor + 30normal)				HERV expression based on Pan-cancer samples(8470tumor + 534 normal)				uniprot		PSSM	netMHCpan			质谱				
herv_id	peptide	len	position_index	direction	logFC	avg_tumor_expr_sitewise	avg_normal_expr_sitewise	pvalue	avg_tumor	avg_normal	logFC	pvalue	avg_tumor	avg_normal	logFC	pvalue	# uniprot	uniprot	PSSM_aff	nM_aff	Rank	NB	MS	filter 1.0	filter 2.0	filter 3.0	filter 4.0
566	FLQTHLTSPL	10	1627	f	10	0.147059	0	0.324587	0.031183	0	10	5.42E-09	0.028608	0.000587	5.606594	5.52E-77	0	0	32.05213	20.985	0.29	1	0		T	T	F
566	SLSLSNLPFL	10	1603	f	10	0.147059	0	0.324587	0.031183	0	10	5.42E-09	0.028608	0.000587	5.606594	5.52E-77	0	0	20.72123	22.898	0.31	1	0		T	T	F
566	FLIFLFYRPI	10	1613	f	10	0.147059	0	0.324587	0.031183	0	10	5.42E-09	0.028608	0.000587	5.606594	5.52E-77	0	0	30.72129	33.622	0.45	1	0		T	T	F
566	FLQTHLTSPLL	11	1627	f	10	0.147059	0	0.324587	0.031183	0	10	5.42E-09	0.028608	0.000587	5.606594	5.52E-77	0	0	247.2175	84.052	0.94	1	0		T	T	F
566	LLQHGLLKPI	10	3023	f	10	0.029412	0	0.324587	0.031183	0	10	5.42E-09	0.028608	0.000587	5.606594	5.52E-77	0	0	111.6315	117.83	1.18	1	0		T	T	F
566	GMVRRVYRL	9	2996	r	10	0.029412	0	0.324587	0.031183	0	10	5.42E-09	0.028608	0.000587	5.606594	5.52E-77	0	0	71.45472	138.13	1.3	1	0		T	T	F
566	SLSNLPFLQTHL	12	1609	f	10	0.147059	0	0.324587	0.031183	0	10	5.42E-09	0.028608	0.000587	5.606594	5.52E-77	0	0	1615.232	168.81	1.48	1	0		F	T	F
566	NLLSLMGV	8	3288	r	10	0.029412	0	0.324587	0.031183	0	10	5.42E-09	0.028608	0.000587	5.606594	5.52E-77	0	0	57.63862	348.04	2.31	0	0		F	T	F
566	STFHSSFFSL	10	1252	f	10	0.102941	0	0.143414	0.031183	0	10	5.42E-09	0.028608	0.000587	5.606594	5.52E-77	0	0	55.12433	356.49	2.34	0	0		F	T	F
566	ASLSLSNLPFL	11	1600	f	10	0.147059	0	0.324587	0.031183	0	10	5.42E-09	0.028608	0.000587	5.606594	5.52E-77	0	0	94.71839	552.96	2.97	0	0		F	T	F
566	SLSNLPFLQT	10	1609	f	10	0.147059	0	0.324587	0.031183	0	10	5.42E-09	0.028608	0.000587	5.606594	5.52E-77	0	0	54.80534	570.38	3.02	0	0		F	T	F
566	SPLACILKNL	10	1274	f	10	0.485294	0	0.119161	0.031183	0	10	5.42E-09	0.028608	0.000587	5.606594	5.52E-77	0	0	1419.345	570.83	3.02	0	0		F	T	F
566	SLSNLPFL	8	1609	f	10	0.147059	0	0.324587	0.031183	0	10	5.42E-09	0.028608	0.000587	5.606594	5.52E-77	0	0	35.97034	581.51	3.05	0	0		F	T	F

6.peptide_reads_inNormal.xlsx

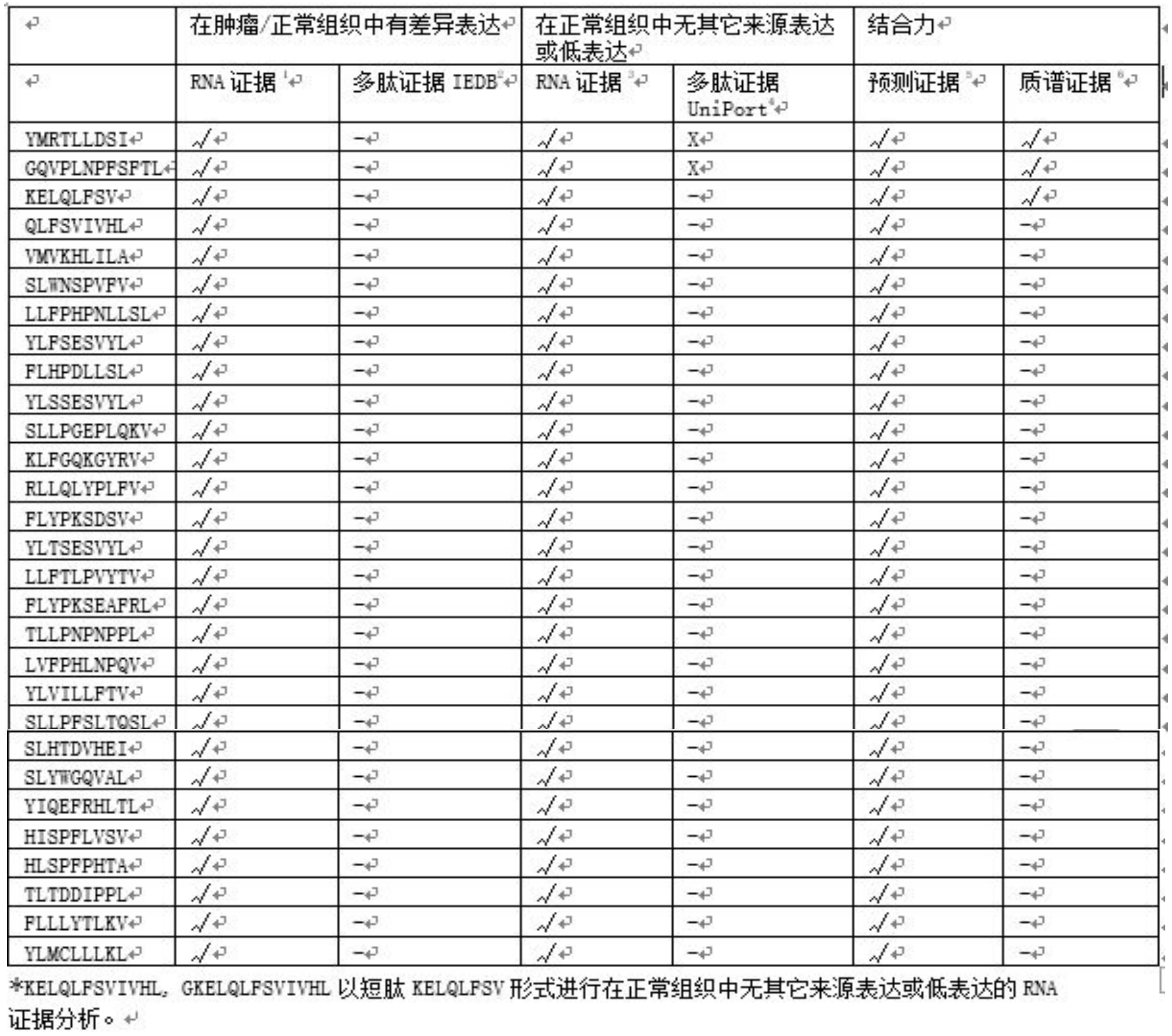


peptides	most likely	consensus	pattern	command	normal_total (sum of mapped reads among normal samples)	normal_total (rm_herv_region)	sample(mapped_read >0)	#samples	pass?
KELQLFSV	aaggagctgca	aargarytncar	AA[AG]GA[A	time for KELQLFSV	305	21	16	121	1
FLFLTLITL	ttcctgttcctga	ttyytnttyytna	TT[CT][CT]T[C	time for FLFLTLITL	1877	431	101	121	0
FLHPDLLSL	ttcctgcacccc	ttyytncayccn	TT[CT][CT]T[C	time for FLHPDLLSL	277	53	27	121	1
FLYPKSDSV	ttcctgtacccca	ttyytntayccna	TT[CT][CT]T[C	time for FLYPKSDSV	967	90	49	121	1
FNLGATLQSL	ttcaacctgggc	ttyaayytnggn	TT[CT]AA[CT	time for FNLGATLQ	2246	240	62	121	0
GQVPLNPFSFTL	ggccaggtgcc	ggncargtncc	GG[GTAC]CA	time for GQVPLNPI	890	85	24	121	1
HLSPFPHTA	cacctgagccc	cayytnwsncc	CA[CT][CT]T[	time for HLSPFPHT	559	1	1	121	1
KLFGQKGYRV	aagctgttcgg	aarytnttyggn	AA[AG][CT]T[	time for KLFGQKGY	0	0	0	121	1
LLFPHPNLLSL	ctgctgttcccc	ytnytnttyccn	[CT]T[GTAC][	time for LLFPHPNL	1	0	0	121	1
PLNPFSFTL	cccctgaacccc	ccnytnaayccn	CC[GTAC][CT	time for PLNPFSFTI	3130	382	65	121	0
QLFSVIVHL	cagctgttcagc	carytnttywsn	CA[AG][CT]T[	time for QLFSVIVH	129	9	4	121	1
RIAKSILSQK	agaatcgccaa	mgnathgcna	[AC]G[GTAC]	time for RIAKSILSQ	29045	28974	121	121	0
RLLQLYPLFV	agactgctgca	mgnytnytnca	[AC]G[GTAC]	time for RLLQLYPLI	0	0	0	121	1





最终结果



	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 IEDB[2]	RNA 证据[3]	多肽证据 UniPort[4]	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。

Thanks