踩坑记录-GB_ACC转换基因-不允许有重复的‘row.names‘-R语言零基础基因/数据差异分析（三）

已于 2022-06-22 15:53:40 修改

编辑

R语言零基础基因/数… 专栏收录该内容

5 篇文章

摘要 本文指导如何将GB_ACC转换为基因名，提供两种方法，推荐使用简单的方法避免繁琐操作。涉及排序、去重、正则表达式处理及CSV格式校验，确保数据无重复'row.names'问题。

文章目录

GB_ACC转换成基因

直接利用GEO分析

在上个系列中，我们记住了 GB_ACC ，但制作热图，需要我们将其转换成基因名，如果你已经转换好了或者不需要制作热图，前参考后续系列。

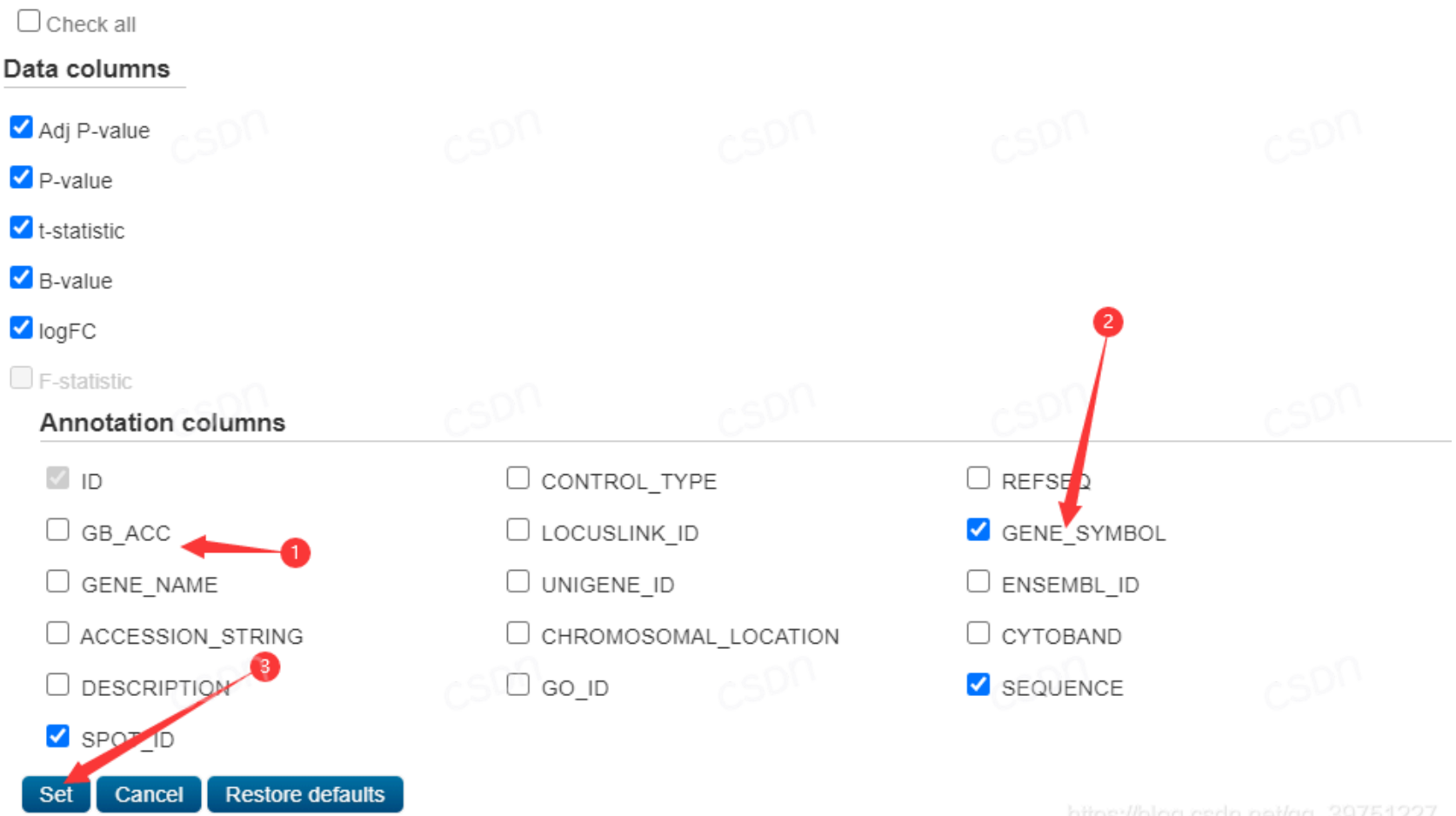
方法1（推荐）

如下图示意，将GB_ACC转换成基因名。

Top differentially expressed genes

Download full table Select columns

ID	adj.P.Val	P.Value	t	B	logFC	GENE_SYMBOL	SEQUENCE	SPOT_ID
A_23_P10385	0.0000107	3.20e-10	145.4	12.24	6.22	DTL	GGAAGATGTTA...	A_23_P10385
A_23_P256956	0.0000107	3.67e-10	141.4	12.21	7.63	KIF20A	TCAAGCCTTG...	A_23_P256956
A_24_P322354	0.000018	1.70e-09	103.9	11.69	5.51	SKA1	ACCTGAAATG...	A_24_P322354
A_23_P302672	0.000018	1.70e-09	-103.9	11.68	-4.34	DDIT4L	ATCAATGTTGT...	A_23_P302672
A_33_P3374205	0.000018	2.37e-09	97.2	11.54	5.87	MKI67	ATGTGCTCTG...	A_33_P3374205
A_24_P346855	0.000018	2.39e-09	97	11.53	4.73	MKI67	TCTCACCCTG...	A_24_P346855
A_33_P3376116	0.000018	2.57e-09	95.6	11.5	3.78	SPC24	TCCAGGAAATT...	A_33_P3376116
A_23_P251421	0.000018	2.79e-09	94	11.46	4.42	CDCA7	ATTTACTTGCA...	A_23_P251421
A_22_P00006...	0.000018	3.01e-09	-92.6	11.43	-4.73	LOC644135	GATTACAAGTT...	A_22_P00006231
A_33_P3392187	0.000018	3.25e-09	91.2	11.39	4.47	CCDC85A	TCCAAATGCAA...	A_33_P3392187
A_23_P63789	0.000018	3.40e-09	90.4	11.37	5.2	ZWINT	TCAAAGATTCA...	A_23_P63789



上图选项框勾选后，就会出现对应的列在列表中，如下图勾选 GENE_SYMBOL后的效果。

Top differentially expressed genes ?

Download full table Select columns

ID	adj.P.Val	P.Value	t	B	logFC	GENE_SYMBOL	SEQUENCE	SPOT_ID
A_23_P10385	0.0000107	3.20e-10	145.4	12.24	6.22	DTL	GGAAGATGTTA...	A_23_P10385
A_23_P256956	0.0000107	3.67e-10	141.4	12.21	7.63	KIF20A	TCAAGCCTTG...	A_23_P256956
A_24_P322354	0.000018	1.70e-09	103.9	11.69	5.51	SKA1	ACCTGAAATG...	A_24_P322354
A_23_P302672	0.000018	1.70e-09	-103.9	11.68	-4.34	DDIT4L	ATCAATGTTGT...	A_23_P302672
A_33_P3374205	0.000018	2.37e-09	97.2	11.54	5.87	MKI67	ATGTGCTCTG...	A_33_P3374205
A_24_P346855	0.000018	2.39e-09	97	11.53	4.73	MKI67	TCTCACCCTG...	A_24_P346855
A_33_P3376116	0.000018	2.57e-09	95.6	11.5	3.78	SPC24	TCCAGGAAATT...	A_33_P3376116
A_23_P251421	0.000018	2.79e-09	94	11.46	4.42	CDCA7	ATTTACTTGCA...	A_23_P251421
A_22_P00006...	0.000018	3.01e-09	-92.6	11.43	-4.73	LOC644135	GATTACAAGTT...	A_22_P00006231
A_33_P3392187	0.000018	3.25e-09	91.2	11.39	4.47	CCDC85A	TCCAAATGCAA...	A_33_P3392187
A_23_P63789	0.000018	3.40e-09	90.4	11.37	5.2	ZWINT	TCAAAGATTCA...	A_23_P63789
A_24_P413884	0.0000186	3.83e-09	88.2	11.31	5.78	CENPA	TAGTTTGTGAG...	A_24_P413884
A_21_P0003965	0.0000186	4.17e-09	-86.7	11.26	-4.84		CAATTTTAATG...	A_21_P0003965
A_23_P35219	0.0000186	4.48e-09	85.5	11.23	5.18	NEK2	AGGAGTGAGA...	A_23_P35219
A_33_P3316273	0.0000186	5.94e-09	-80.8	11.07	-3.56	CCL3	TGCTTTTGTTC...	A_33_P3316273
A_22_P00018...	0.0000186	6.06e-09	-80.5	11.06	-4.19		ATTCAGCTCTG...	A_22_P00018082
A_22_P00001...	0.0000186	6.07e-09	80.4	11.06	5.26	LINC00842	CCCATATTCTT...	A_22_P00001017

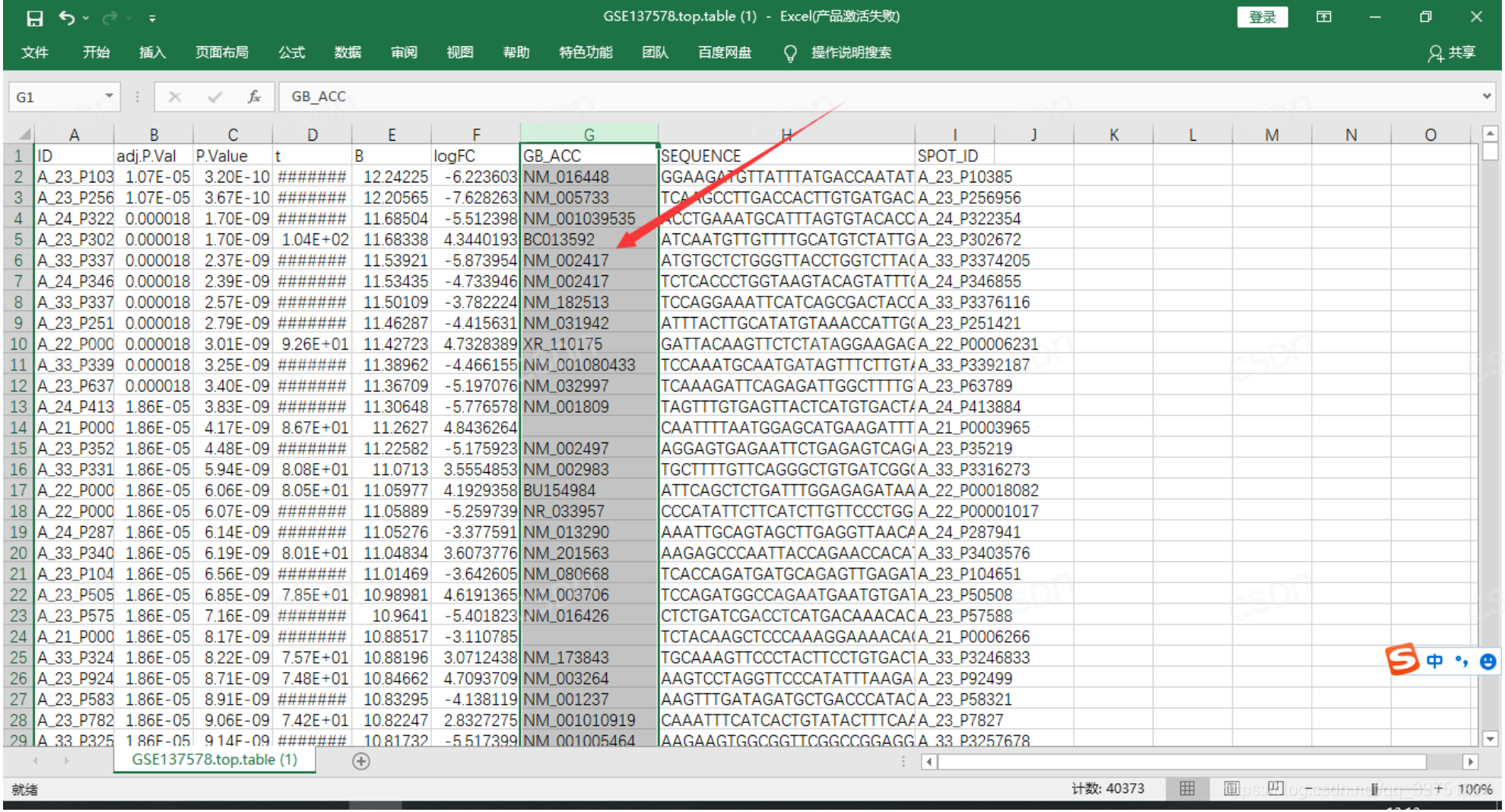
如此就可以实现GB_ACC与基因名的对应。

GB_ACC的第二种方法（繁琐）

我并**不推荐这种方法**，它十分的繁琐且对新手不友好，如果你是新手，**请忽视掉它。**

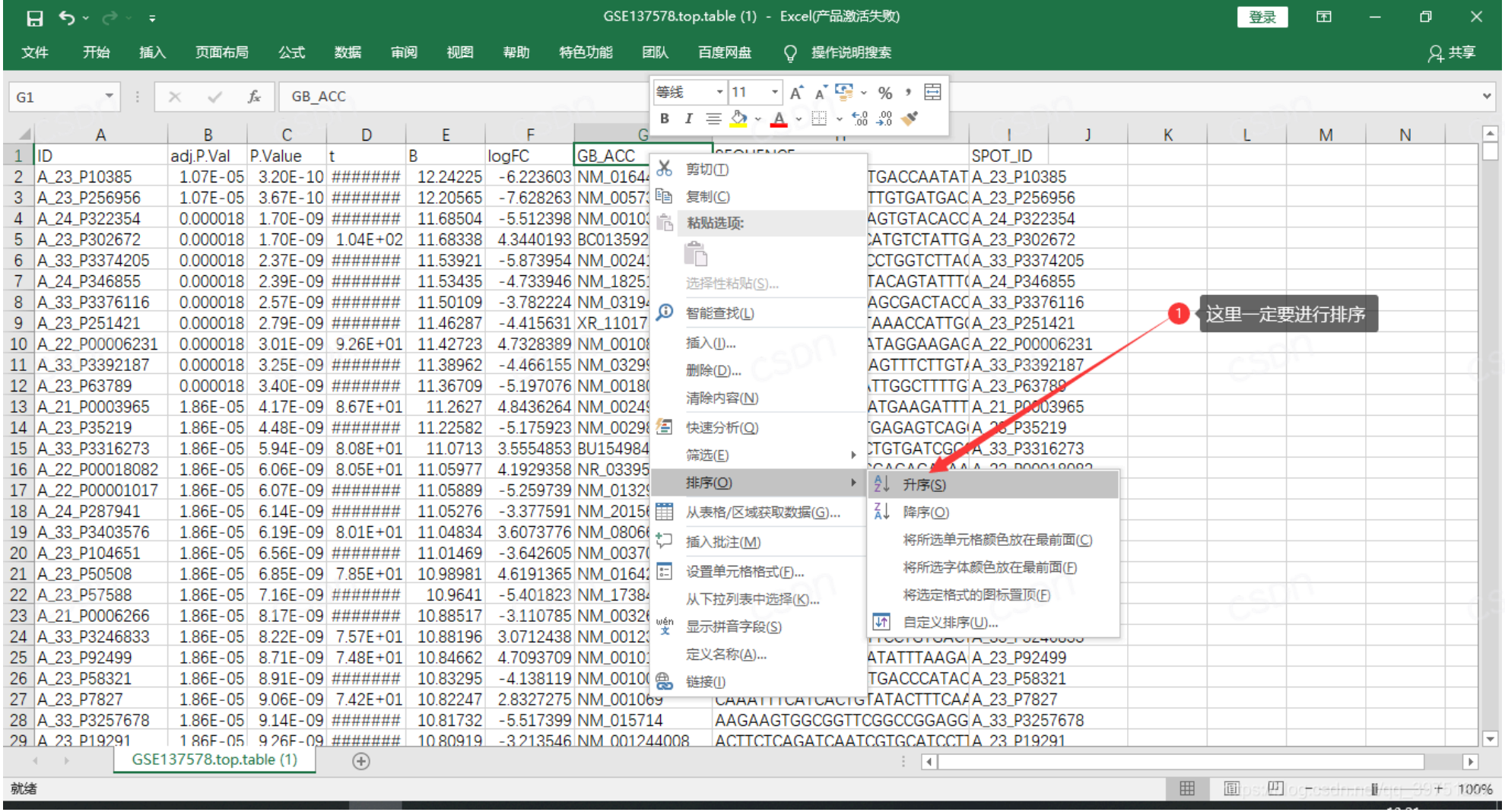
在后面的文章中，如果**特别指明（请忽视）说明是不重要**且**无伤大雅**的。

要复制的列：

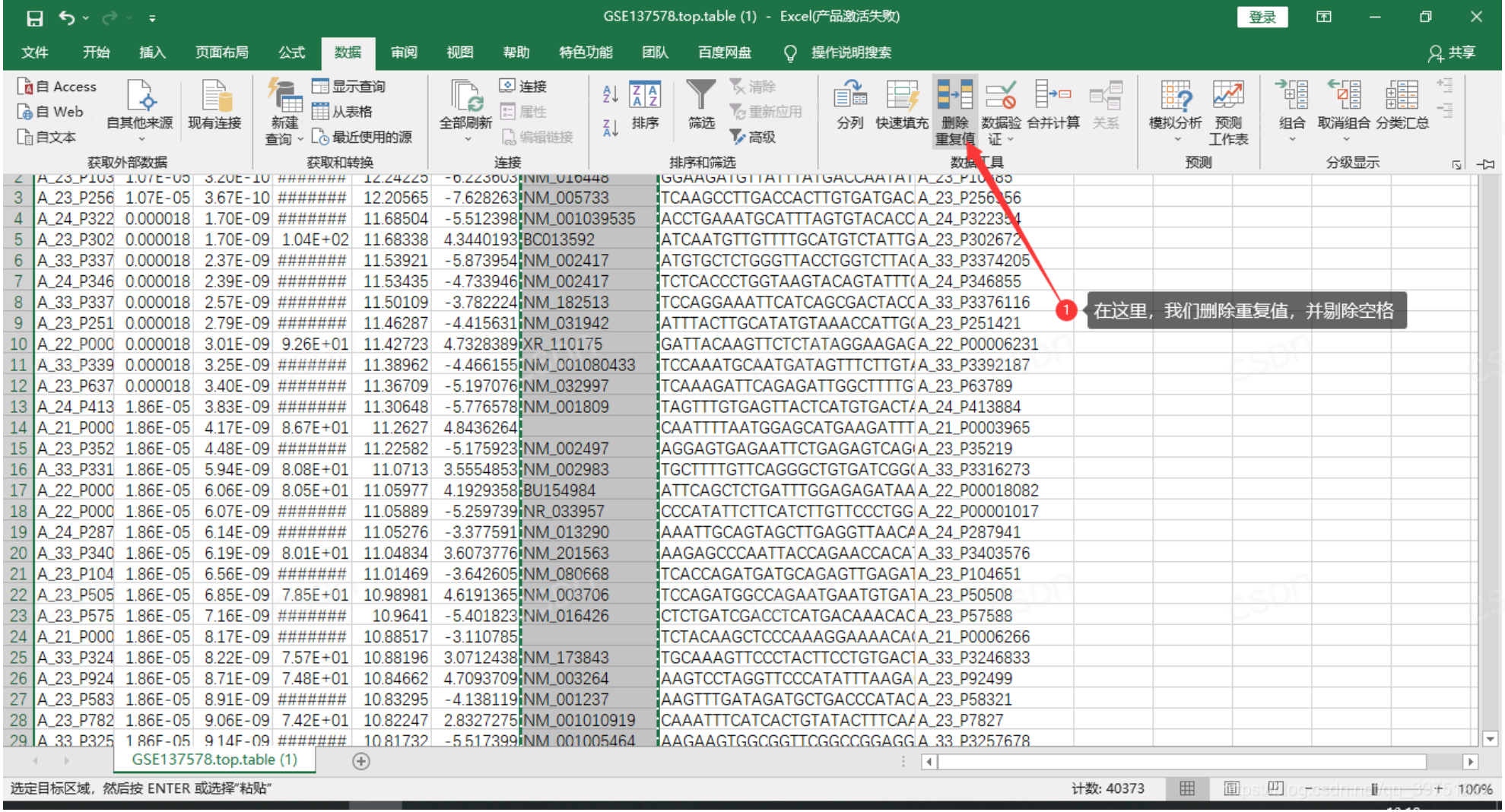


	A	B	C	D	E	F	G	H	I
1	ID	adj.P.Val	P.Value	t	B	logFC	GB_ACC	SEQUENCE	SPOT_ID
2	A_23_P103	1.07E-05	3.20E-10	#######	12.24225	-6.223603	NM_016448	GGAAGATGTTATTTATGACCAATAT	A_23_P10385
3	A_23_P256	1.07E-05	3.67E-10	#######	12.20565	-7.628263	NM_005733	TCAAGCCTTGACCACTTGTGATGAC	A_23_P256956
4	A_24_P322	0.000018	1.70E-09	#######	11.68504	-5.512398	NM_001039535	ACCTGAAATGCATTTAGTGTACACC	A_24_P322354
5	A_23_P302	0.000018	1.70E-09	1.04E+02	11.68338	4.3440193	BC013592	ATCAATGTTGTTTTGCATGTCTATTG	A_23_P302672
6	A_33_P337	0.000018	2.37E-09	#######	11.53921	-5.873954	NM_002417	ATGTGCTCTGGGTTACCTGGTCTTAC	A_33_P3374205
7	A_24_P346	0.000018	2.39E-09	#######	11.53435	-4.733946	NM_002417	TCTCACCCTGGTAAGTACAGTATTTC	A_24_P346855
8	A_33_P337	0.000018	2.57E-09	#######	11.50109	-3.782224	NM_182513	TCCAGGAAATTCATCAGCGACTACC	A_33_P3376116
9	A_23_P251	0.000018	2.79E-09	#######	11.46287	-4.415631	NM_031942	ATTTACTTGCATATGTAAACCATTGC	A_23_P251421
10	A_22_P000	0.000018	3.01E-09	9.26E+01	11.42723	4.7328389	XR_110175	GATTACAAGTTCTCTATAGGAAGAG	A_22_P00006231
11	A_33_P339	0.000018	3.25E-09	#######	11.38962	-4.466155	NM_001080433	TCCAAATGCAATGATAGTTTCTTGTA	A_33_P3392187
12	A_23_P637	0.000018	3.40E-09	#######	11.36709	-5.197076	NM_032997	TCAAAGATTCAGAGATTGGCTTTTG	A_23_P63789
13	A_24_P413	1.86E-05	3.83E-09	#######	11.30648	-5.776578	NM_001809	TAGTTTGTGAGTTACTCATGTGACTA	A_24_P413884
14	A_21_P000	1.86E-05	4.17E-09	8.67E+01	11.2627	4.8436264		CAATTTTAATGGAGCATGAAGATTT	A_21_P0003965
15	A_23_P352	1.86E-05	4.48E-09	#######	11.22582	-5.175923	NM_002497	AGGAGTGAGAATTCTGAGAGTCAG	A_23_P35219
16	A_33_P331	1.86E-05	5.94E-09	8.08E+01	11.0713	3.5554853	NM_002983	TGCTTTTGTTCAGGGCTGTGATCGG	A_33_P3316273
17	A_22_P000	1.86E-05	6.06E-09	8.05E+01	11.05977	4.1929358	BU154984	ATTCAGCTCTGATTTGGAGAGATAA	A_22_P00018082
18	A_22_P000	1.86E-05	6.07E-09	#######	11.05889	-5.259739	NR_033957	CCCATATTCTTCATCTTGTTCCCTGG	A_22_P00001017
19	A_24_P287	1.86E-05	6.14E-09	#######	11.05276	-3.377591	NM_013290	AAATTGCAGTAGCTTGAGGTTAACA	A_24_P287941
20	A_33_P340	1.86E-05	6.19E-09	8.01E+01	11.04834	3.6073776	NM_201563	AAGAGCCCAATTACCAGAACCACAT	A_33_P3403576
21	A_23_P104	1.86E-05	6.56E-09	#######	11.01469	-3.642605	NM_080668	TCACCAGATGATGCAGAGTTGAGAT	A_23_P104651
22	A_23_P505	1.86E-05	6.85E-09	7.85E+01	10.98981	4.6191365	NM_003706	TCCAGATGGCCAGAATGAATGTGAT	A_23_P50508
23	A_23_P575	1.86E-05	7.16E-09	#######	10.9641	-5.401823	NM_016426	CTCTGATCGACCTCATGACAAACAC	A_23_P57588
24	A_21_P000	1.86E-05	8.17E-09	#######	10.88517	-3.110785		TCTACAAGCTCCCAAAGGAAAACAC	A_21_P0006266
25	A_33_P324	1.86E-05	8.22E-09	7.57E+01	10.88196	3.0712438	NM_173843	TGCAAAGTTCCCTACTTCCTGTGACT	A_33_P3246833
26	A_23_P924	1.86E-05	8.71E-09	7.48E+01	10.84662	4.7093709	NM_003264	AAGTCCTAGGTTCCCATATTTAAGA	A_23_P92499
27	A_23_P583	1.86E-05	8.91E-09	#######	10.83295	-4.138119	NM_001237	AAGTTTGATAGATGCTGACCCATAC	A_23_P58321
28	A_23_P782	1.86E-05	9.06E-09	7.42E+01	10.82247	2.8327275	NM_001010919	CAAATTTCATCACTGTATACTTTCAA	A_23_P7827
29	A_33_P325	1.86E-05	9.14E-09	#######	10.81732	-5.517399	NM_001005464	AAGAAGTGGCGGTTCGGCCGGAGG	A_33_P3257678

但在复制选择的列之前，我们需要对**它进行排序**，这个非常重要

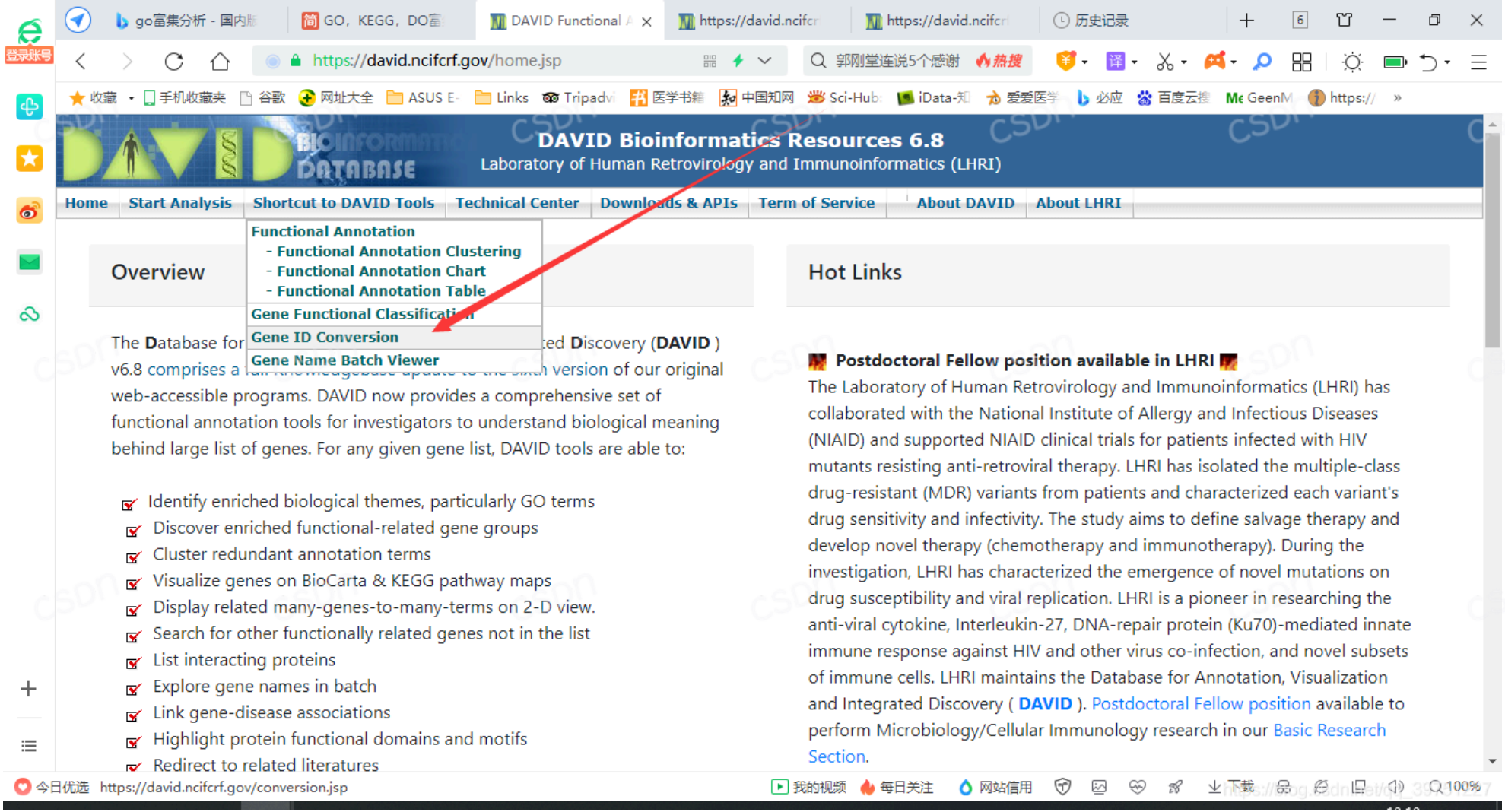


再删除重复，**如图**

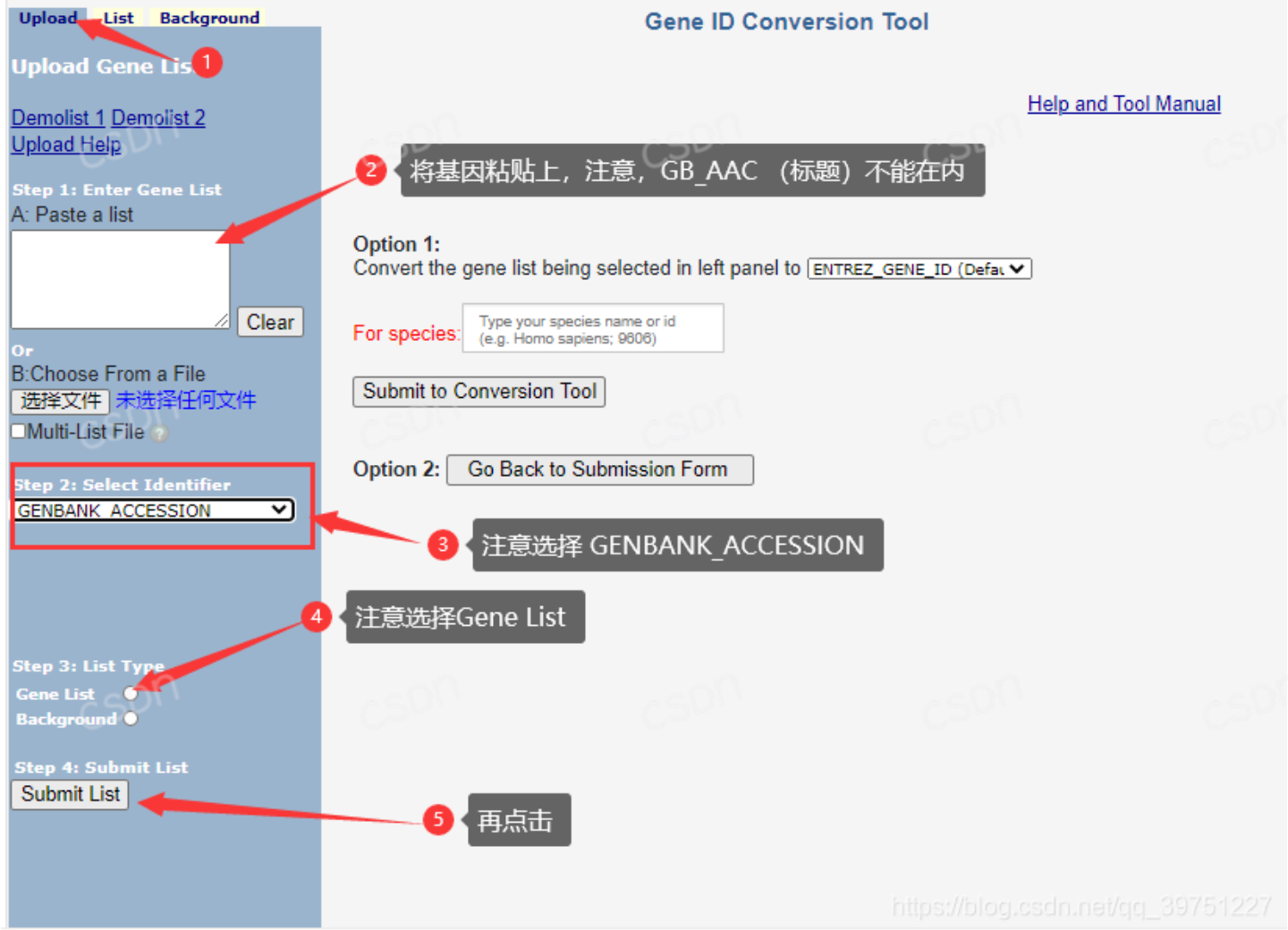


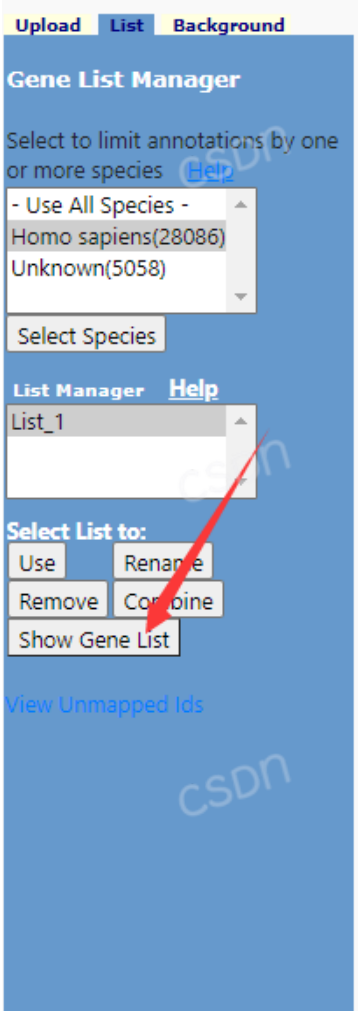
之后打开David,

地址：https://david.ncifcrf.gov/

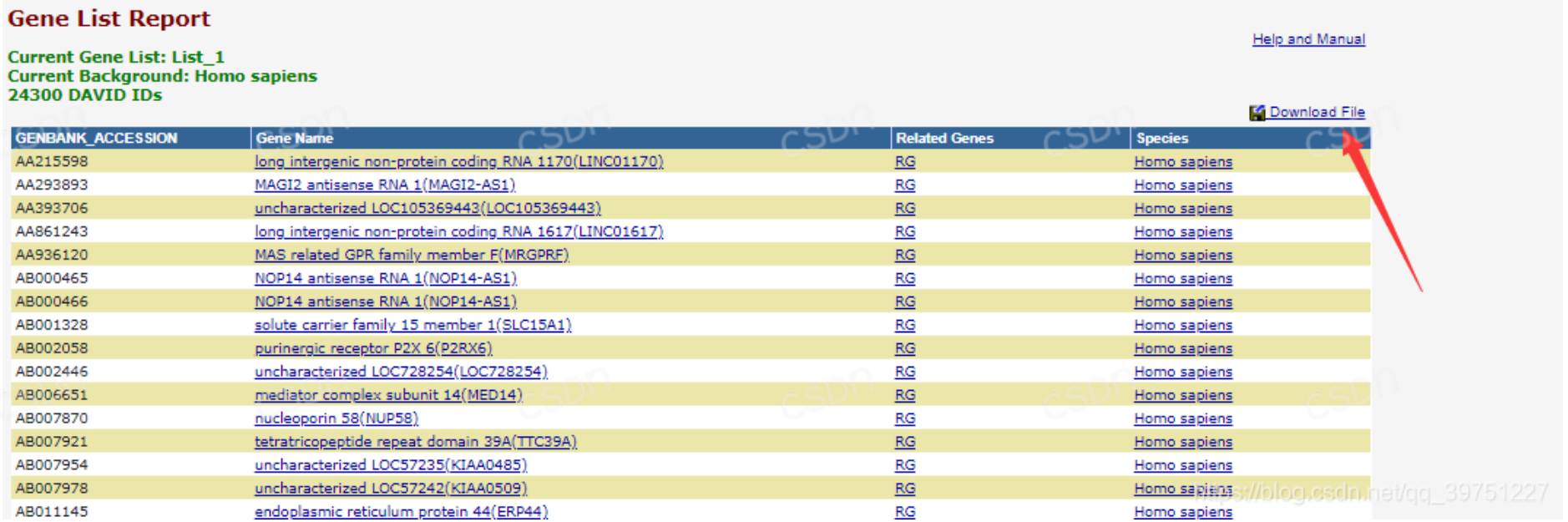


由于我做的是人类基因，所以下图步骤3选择**可能不同**，请自己选择





之后我们可点击下载 文件



Gene List Report

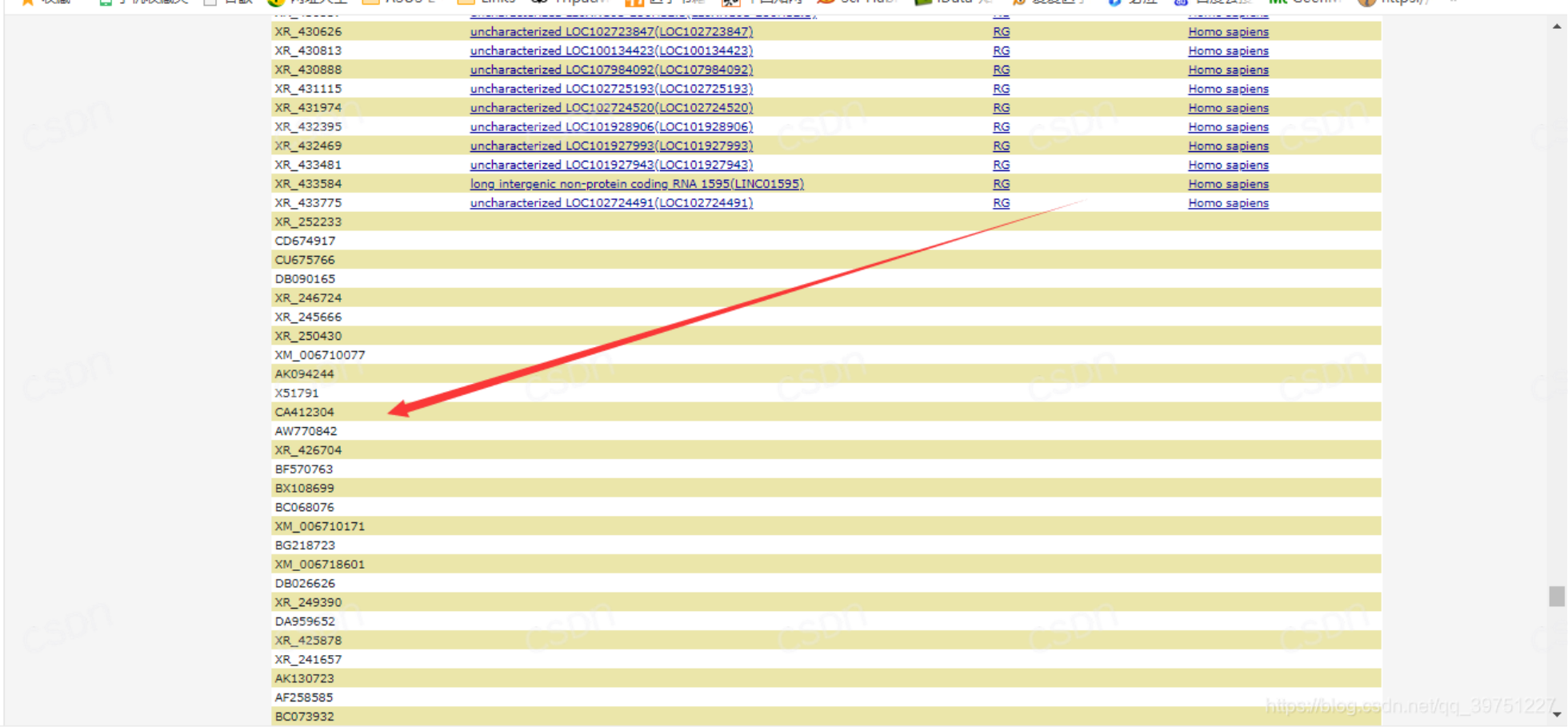
Help and Manual

Current Gene List: List_1
Current Background: Homo sapiens
24300 DAVID IDs

Download File

GENBANK_ACCESSION	Gene Name	Related Genes	Species
AA215598	long intergenic non-protein coding RNA 1170(LINC01170)	RG	Homo sapiens
AA293893	MAGI2 antisense RNA 1(MAGI2-AS1)	RG	Homo sapiens
AA393706	uncharacterized LOC105369443(LOC105369443)	RG	Homo sapiens
AA861243	long intergenic non-protein coding RNA 1617(LINC01617)	RG	Homo sapiens
AA936120	MAS related GPR family member F(MRGPRF)	RG	Homo sapiens
AB000465	NOP14 antisense RNA 1(NOP14-AS1)	RG	Homo sapiens
AB000466	NOP14 antisense RNA 1(NOP14-AS1)	RG	Homo sapiens
AB001328	solute carrier family 15 member 1(SLC15A1)	RG	Homo sapiens
AB002058	purinergic receptor P2X 6(P2RX6)	RG	Homo sapiens
AB002446	uncharacterized LOC728254(LOC728254)	RG	Homo sapiens
AB006651	mediator complex subunit 14(MED14)	RG	Homo sapiens
AB007870	nucleoporin 58(NUP58)	RG	Homo sapiens
AB007921	tetratricopeptide repeat domain 39A(TTC39A)	RG	Homo sapiens
AB007954	uncharacterized LOC57235(KIAA0485)	RG	Homo sapiens
AB007978	uncharacterized LOC57242(KIAA0509)	RG	Homo sapiens
AB011145	endoplasmic reticulum protein 44(ERP44)	RG	Homo sapiens

注意，有些GB_ACC是没有基因的，所以我们需要对下载的文件进行进一步排序。



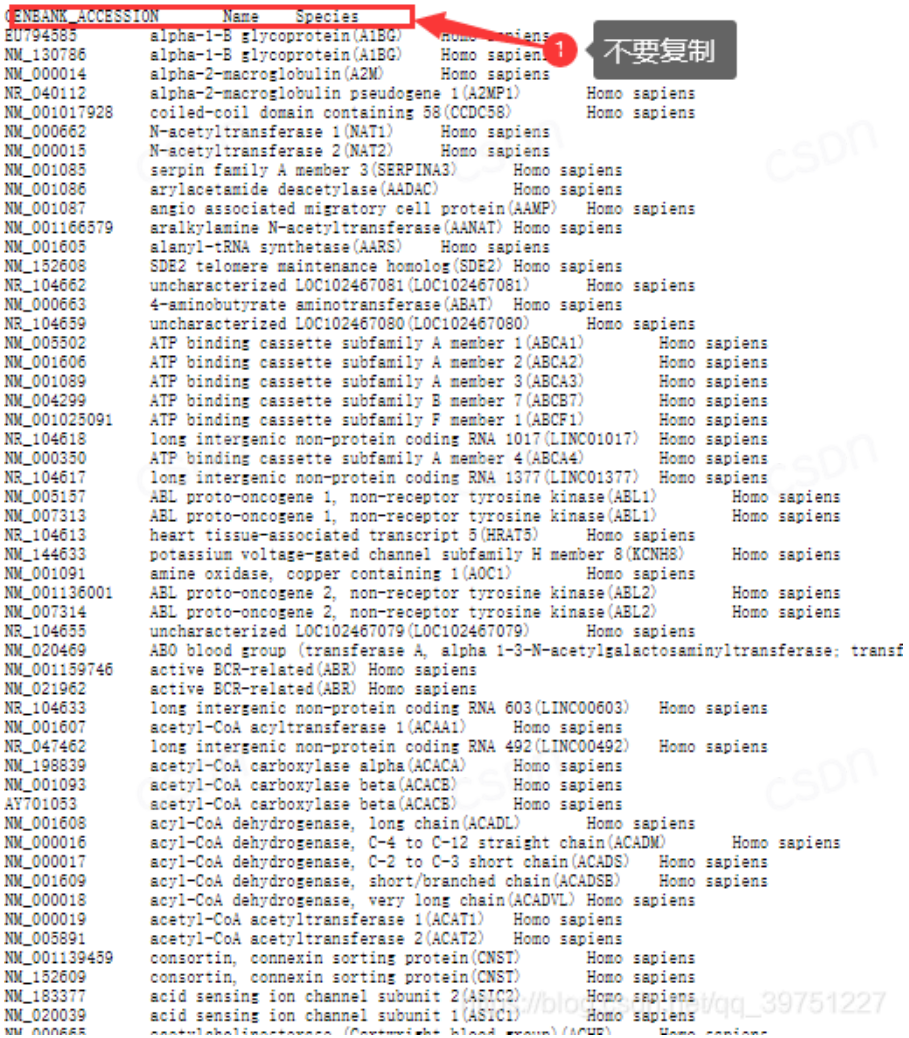
XR_430626	uncharacterized LOC102723847(LOC102723847)	RG	Homo sapiens
XR_430813	uncharacterized LOC100134423(LOC100134423)	RG	Homo sapiens
XR_430888	uncharacterized LOC107984092(LOC107984092)	RG	Homo sapiens
XR_431115	uncharacterized LOC102725193(LOC102725193)	RG	Homo sapiens
XR_431974	uncharacterized LOC102724520(LOC102724520)	RG	Homo sapiens
XR_432395	uncharacterized LOC101928906(LOC101928906)	RG	Homo sapiens
XR_432469	uncharacterized LOC101927993(LOC101927993)	RG	Homo sapiens
XR_433481	uncharacterized LOC101927943(LOC101927943)	RG	Homo sapiens
XR_433584	long intergenic non-protein coding RNA 1595(LINC01595)	RG	Homo sapiens
XR_433775	uncharacterized LOC102724491(LOC102724491)	RG	Homo sapiens
XR_252233			
CD674917			
CU675766			
DB090165			
XR_246724			
XR_245666			
XR_250430			
XM_006710077			
AK094244			
X51791			
CA412304			
AW770842			
XR_426704			
BF570763			
BX108699			
BC068076			
XM_006710171			
BG218723			
XM_006718601			
DB026626			
XR_249390			
DA959652			
XR_425878			
XR_241657			
AK130723			
AF258585			
BC073932			

基因的排序（请忽视）

准备工具

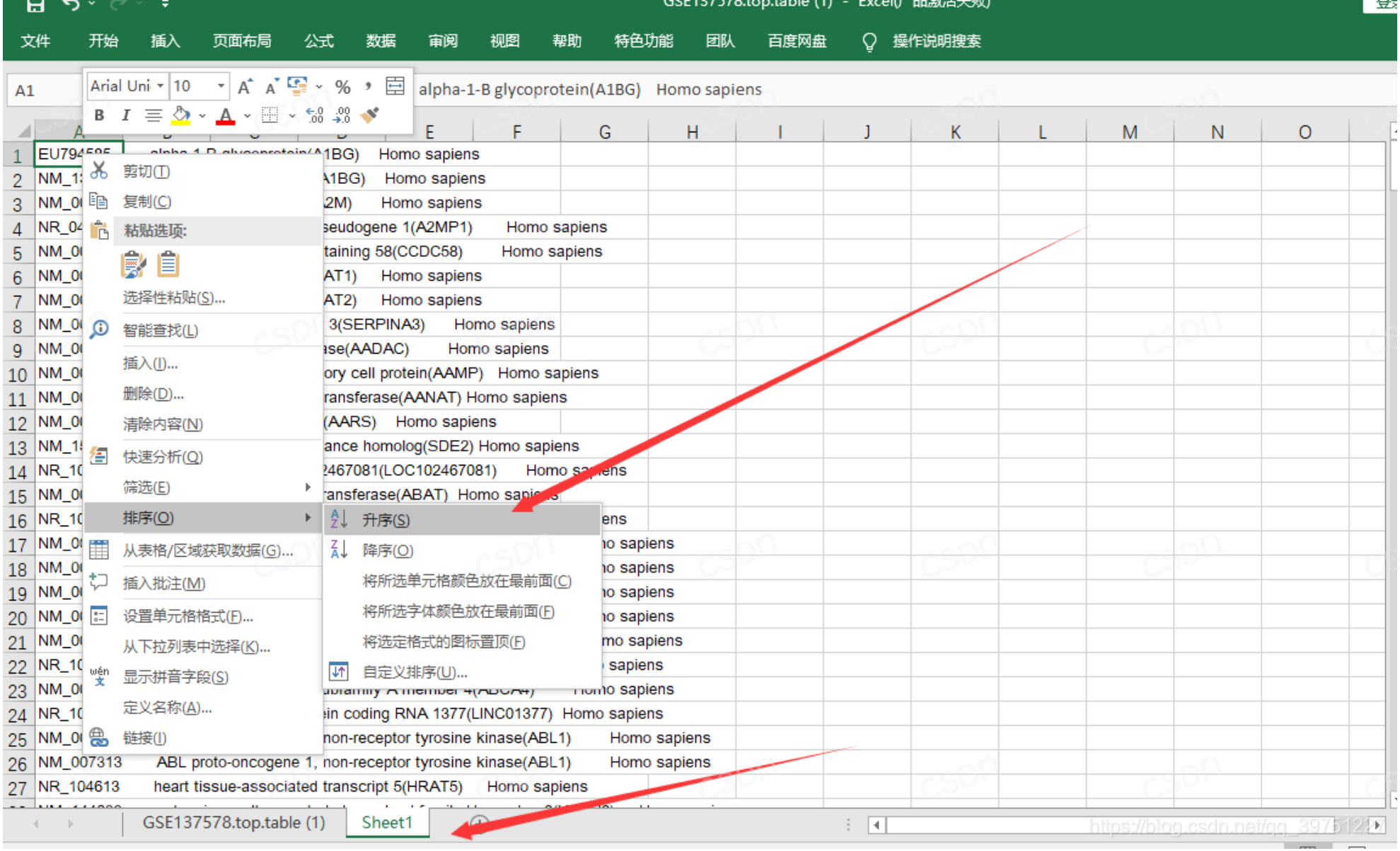
Sublime Text,请自行下载

除如图所示外，其他全复制。



GENBANK_ACCESSION	Name	Species
EU794585	alpha-1-B glycoprotein(A1BG)	Homo sapiens
NM_130786	alpha-1-B glycoprotein(A1BG)	Homo sapiens
NM_000014	alpha-2-macroglobulin(A2M)	Homo sapiens
NR_040112	alpha-2-macroglobulin pseudogene 1(A2MP1)	Homo sapiens
NM_001017928	coiled-coil domain containing 58(CCDC58)	Homo sapiens
NM_000662	N-acetyltransferase 1(NAT1)	Homo sapiens
NM_000015	N-acetyltransferase 2(NAT2)	Homo sapiens
NM_001085	serpin family A member 3(SERPINA3)	Homo sapiens
NM_001086	arylacetamide deacetylase(AADAC)	Homo sapiens
NM_001087	angio associated migratory cell protein(AAMP)	Homo sapiens
NM_001166579	aralkylamine N-acetyltransferase(AANAT)	Homo sapiens
NM_001605	alanyl-tRNA synthetase(AARS)	Homo sapiens
NM_152608	SDE2 telomere maintenance homolog(SDE2)	Homo sapiens
NR_104662	uncharacterized LOC102467081(LOC102467081)	Homo sapiens
NM_000663	4-aminobutyrate aminotransferase(ABAT)	Homo sapiens
NR_104659	uncharacterized LOC102467080(LOC102467080)	Homo sapiens
NM_005502	ATP binding cassette subfamily A member 1(ABCA1)	Homo sapiens
NM_001606	ATP binding cassette subfamily A member 2(ABCA2)	Homo sapiens
NM_001089	ATP binding cassette subfamily A member 3(ABCA3)	Homo sapiens
NM_004299	ATP binding cassette subfamily B member 7(ABCB7)	Homo sapiens
NM_001025091	ATP binding cassette subfamily F member 1(ABCF1)	Homo sapiens
NR_104618	long intergenic non-protein coding RNA 1017(LINC01017)	Homo sapiens
NM_000350	ATP binding cassette subfamily A member 4(ABCA4)	Homo sapiens
NR_104617	long intergenic non-protein coding RNA 1377(LINC01377)	Homo sapiens
NM_005157	ABL proto-oncogene 1, non-receptor tyrosine kinase(ABL1)	Homo sapiens
NM_007313	ABL proto-oncogene 1, non-receptor tyrosine kinase(ABL1)	Homo sapiens
NR_104613	heart tissue-associated transcript 5(HRAT5)	Homo sapiens
NM_144633	potassium voltage-gated channel subfamily H member 8(KCNH8)	Homo sapiens
NM_001091	amine oxidase, copper containing 1(AOC1)	Homo sapiens
NM_001136001	ABL proto-oncogene 2, non-receptor tyrosine kinase(ABL2)	Homo sapiens
NM_007314	ABL proto-oncogene 2, non-receptor tyrosine kinase(ABL2)	Homo sapiens
NR_104655	uncharacterized LOC102467079(LOC102467079)	Homo sapiens
NM_020469	ABO blood group (transferase A, alpha 1-3-N-acetylgalactosaminyltransferase; transf	
NM_001159746	active BCR-related(ABR)	Homo sapiens
NM_021962	active BCR-related(ABR)	Homo sapiens
NR_104633	long intergenic non-protein coding RNA 603(LINC00603)	Homo sapiens
NM_001607	acetyl-CoA acyltransferase 1(ACAA1)	Homo sapiens
NR_047462	long intergenic non-protein coding RNA 492(LINC00492)	Homo sapiens
NM_198839	acetyl-CoA carboxylase alpha(ACACA)	Homo sapiens
NM_001093	acetyl-CoA carboxylase beta(ACACB)	Homo sapiens
AY701053	acetyl-CoA carboxylase beta(ACACB)	Homo sapiens
NM_001608	acyl-CoA dehydrogenase, long chain(ACADL)	Homo sapiens
NM_000016	acyl-CoA dehydrogenase, C-4 to C-12 straight chain(ACADM)	Homo sapiens
NM_000017	acyl-CoA dehydrogenase, C-2 to C-3 short chain(ACADS)	Homo sapiens
NM_001609	acyl-CoA dehydrogenase, short/branched chain(ACADSB)	Homo sapiens
NM_000018	acyl-CoA dehydrogenase, very long chain(ACADVL)	Homo sapiens
NM_000019	acetyl-CoA acetyltransferase 1(ACAT1)	Homo sapiens
NM_005891	acetyl-CoA acetyltransferase 2(ACAT2)	Homo sapiens
NM_001139459	consortin, connexin sorting protein(CNST)	Homo sapiens
NM_152609	consortin, connexin sorting protein(CNST)	Homo sapiens
NM_183377	acid sensing ion channel subunit 2(ASIC2)	Homo sapiens
NM_020039	acid sensing ion channel subunit 1(ASIC1)	Homo sapiens
NM_000665	acetylcholinesterase (Cartwright blood group)(ACHE)	Homo sapiens

先将文件贴到新的excel中，**进行排序**



排序完成之后，再复制全部（ctrl + A全选，ctrl + C 复制）新建一个文本文档并用 sublime text打开，粘贴基因。

正则表达式

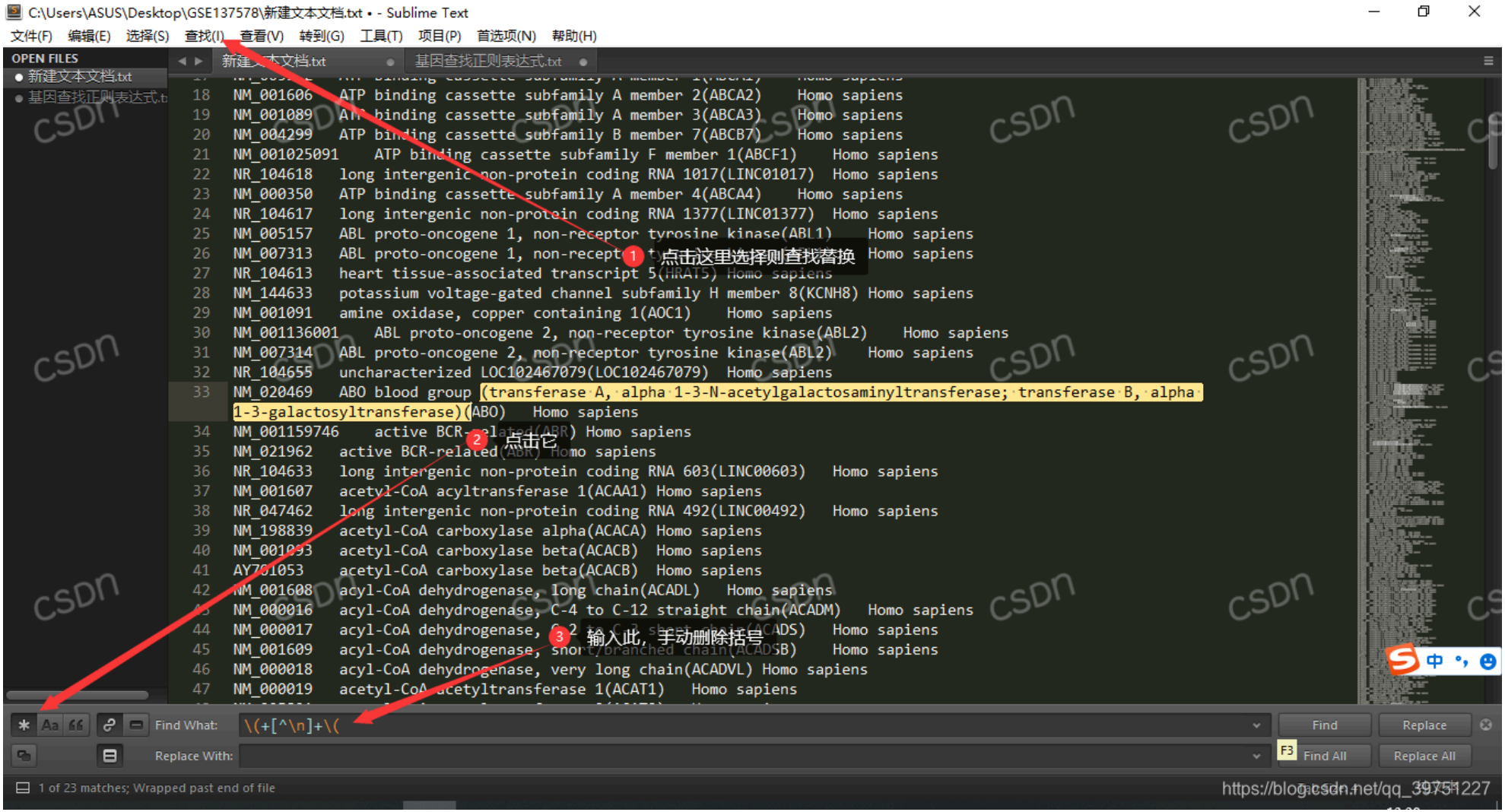
```java
查询多个括号:
\(+[^\n]+\(
```

```
删除括号左右侧:
\)+[^\n]+\n
\n+[^\n]+\(
```

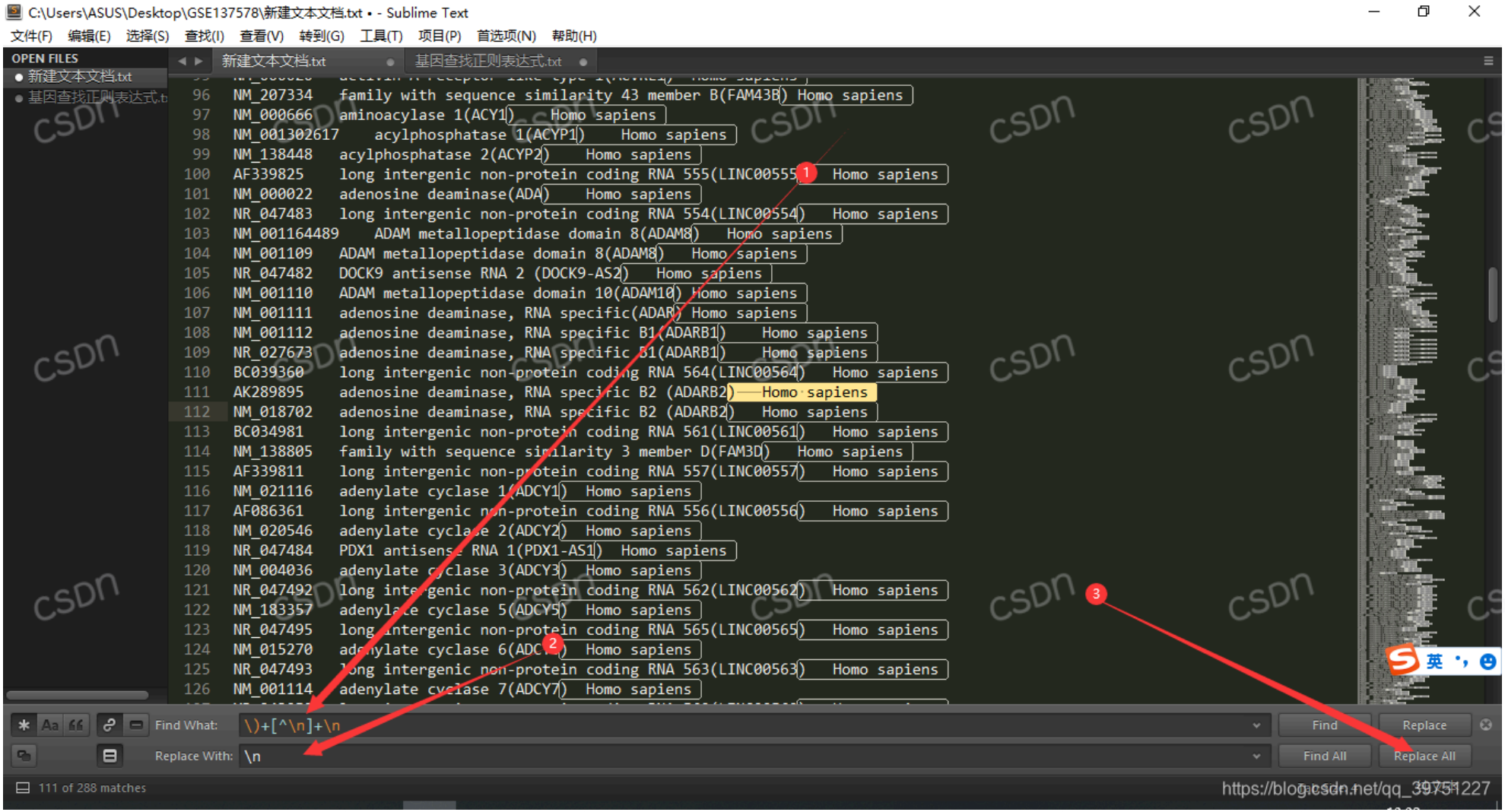
点击查询，找到查找和替换，然后逐个输入上方正则表达式（就是上面每行表达式，注意**中文是注解行，就别输入了**）替换成换行符（也就是写"\n",注意不带引号）。

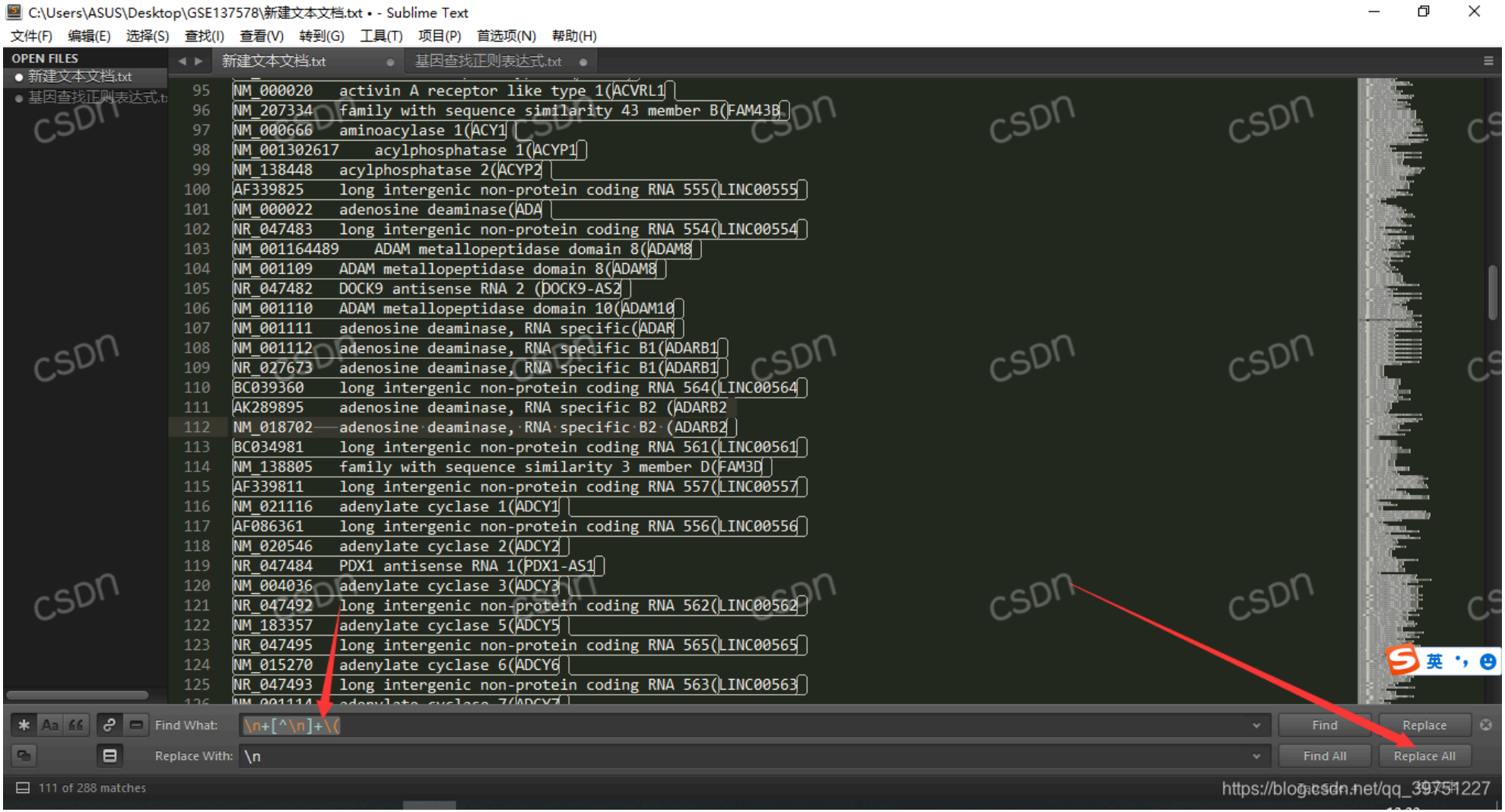
一般来说括号内是基因，有的 **基因说明** 中，括号内不一定是基因，请在进行上面正则表达式操作之前，**手动删除用括号括起来但里面又不是基因的文本（连带它的括号一起删了）**。

注意输入正则表达式顺序已排好，依次操作就好，别整错顺序了。



上图注意按顺序点击。

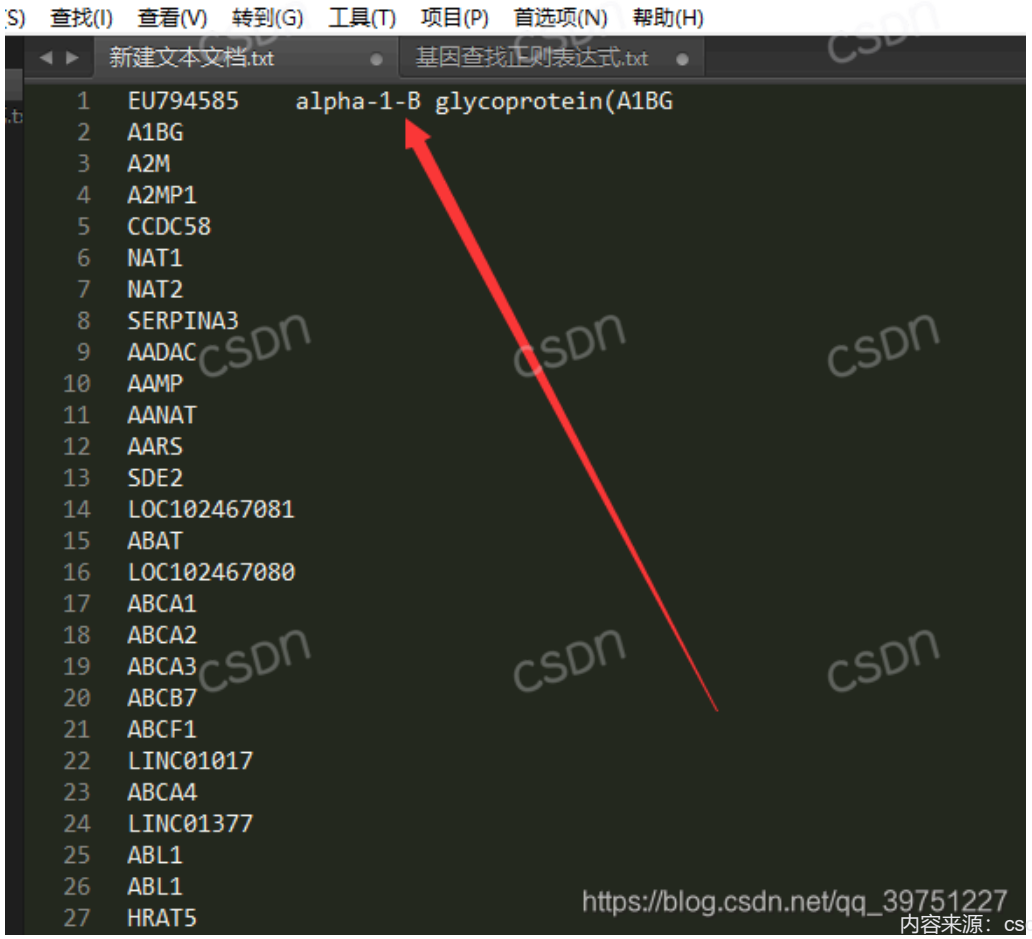




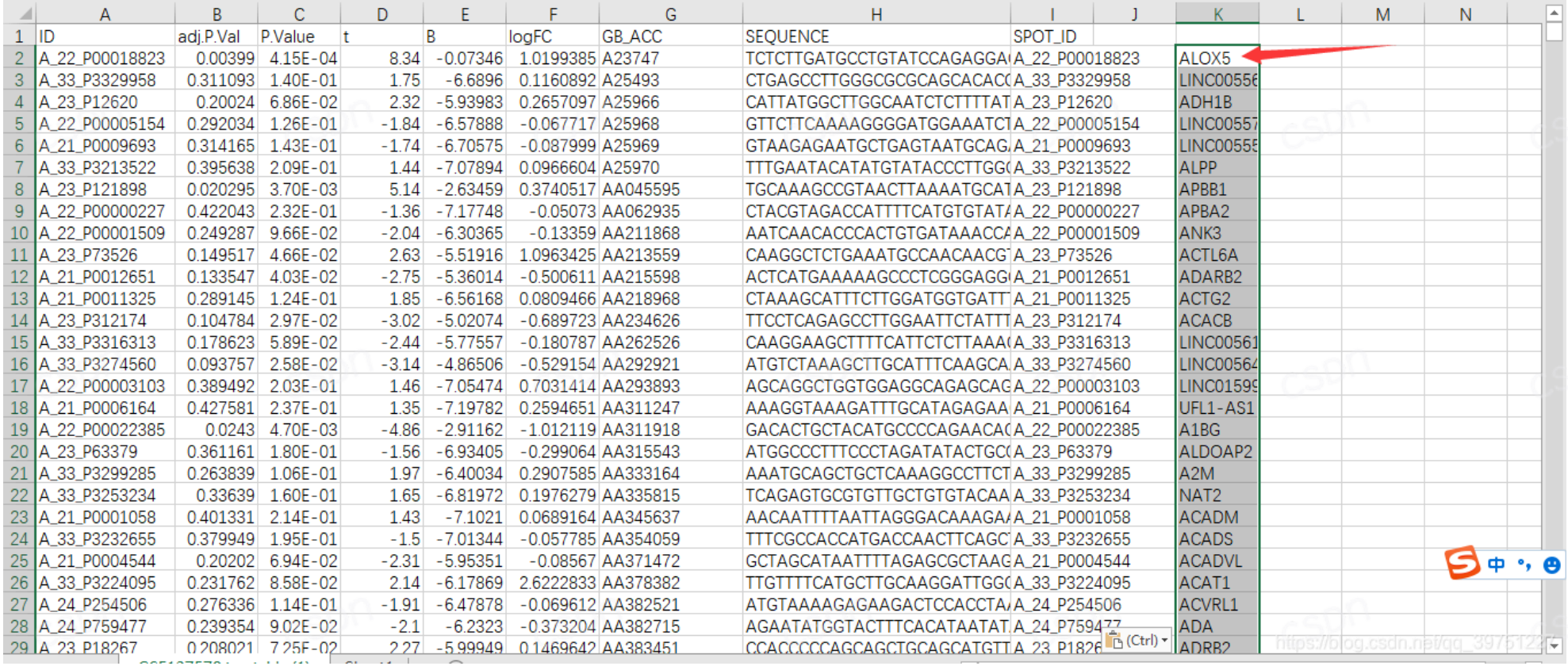
上**两个图**示范了一个正则表达式的用法，请**自行完成**后续正则表达式的操作。

删除末尾多余文字

删除开头多余文字



之后全部复制贴回对应行即可，另存为**csv文件**，我们这里命名为**GSE_HM**



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	ID	adj.P.Val	P.Value	t	B	logFC	GB_ACC	SEQUENCE	SPOT_ID					
2	A_22_P00018823	0.00399	4.15E-04	8.34	-0.07346	1.0199385	A23747	TCTCTTGATGCCTGTATCCAGAGGA	A_22_P00018823		ALOX5			
3	A_33_P3329958	0.311093	1.40E-01	1.75	-6.6896	0.1160892	A25493	CTGAGCCTTGGGCGCGCAGCACAC	A_33_P3329958		LINC00556			
4	A_23_P12620	0.20024	6.86E-02	2.32	-5.93983	0.2657097	A25966	CATTATGGCTTGGCAATCTCTTTTAT	A_23_P12620		ADH1B			
5	A_22_P00005154	0.292034	1.26E-01	-1.84	-6.57888	-0.067717	A25968	GTTCTTCAAAAGGGGATGGAAATCT	A_22_P00005154		LINC00557			
6	A_21_P0009693	0.314165	1.43E-01	-1.74	-6.70575	-0.087999	A25969	GTAAGAGAATGCTGAGTAATGCAG	A_21_P0009693		LINC00555			
7	A_33_P3213522	0.395638	2.09E-01	1.44	-7.07894	0.0966604	A25970	TTTGAATACATATGTATACCCTTGG	A_33_P3213522		ALPP			
8	A_23_P121898	0.020295	3.70E-03	5.14	-2.63459	0.3740517	AA045595	TGCAAAGCCGTAACTTAAAATGCAT	A_23_P121898		APBB1			
9	A_22_P00000227	0.422043	2.32E-01	-1.36	-7.17748	-0.05073	AA062935	CTACGTAGACCATTTTCATGTGTATA	A_22_P00000227		APBA2			
10	A_22_P00001509	0.249287	9.66E-02	-2.04	-6.30365	-0.13359	AA211868	AATCAACACCCACTGTGATAAACCA	A_22_P00001509		ANK3			
11	A_23_P73526	0.149517	4.66E-02	2.63	-5.51916	1.0963425	AA213559	CAAGGCTCTGAAATGCCAACAACG	A_23_P73526		ACTL6A			
12	A_21_P0012651	0.133547	4.03E-02	-2.75	-5.36014	-0.500611	AA215598	ACTCATGAAAAAGCCCTCGGGAGG	A_21_P0012651		ADARB2			
13	A_21_P0011325	0.289145	1.24E-01	1.85	-6.56168	0.0809466	AA218968	CTAAAGCATTTCTTGGATGGTGATT	A_21_P0011325		ACTG2			
14	A_23_P312174	0.104784	2.97E-02	-3.02	-5.02074	-0.689723	AA234626	TTCCTCAGAGCCTTGGAATTCTATTT	A_23_P312174		ACACB			
15	A_33_P3316313	0.178623	5.89E-02	-2.44	-5.77557	-0.180787	AA262526	CAAGGAAGCTTTTCATTCTCTTAAA	A_33_P3316313		LINC00561			
16	A_33_P3274560	0.093757	2.58E-02	-3.14	-4.86506	-0.529154	AA292921	ATGTCTAAAGCTTGCATTTCAAGCA	A_33_P3274560		LINC00564			
17	A_22_P00003103	0.389492	2.03E-01	1.46	-7.05474	0.7031414	AA293893	AGCAGGCTGGTGGAGGCAGAGCAG	A_22_P00003103		LINC01599			
18	A_21_P0006164	0.427581	2.37E-01	1.35	-7.19782	0.2594651	AA311247	AAAGGTAAAGATTTGCATAGAGAA	A_21_P0006164		UFL1-AS1			
19	A_22_P00022385	0.0243	4.70E-03	-4.86	-2.91162	-1.012119	AA311918	GACACTGCTACATGCCCCAGAACA	A_22_P00022385		A1BG			
20	A_23_P63379	0.361161	1.80E-01	-1.56	-6.93405	-0.299064	AA315543	ATGGCCCTTTCCCTAGATATACTGC	A_23_P63379		ALDOAP2			
21	A_33_P3299285	0.263839	1.06E-01	1.97	-6.40034	0.2907585	AA333164	AAATGCAGCTGCTCAAAGGCCTTCT	A_33_P3299285		A2M			
22	A_33_P3253234	0.33639	1.60E-01	1.65	-6.81972	0.1976279	AA335815	TCAGAGTGCGTGTTGCTGTGTACAA	A_33_P3253234		NAT2			
23	A_21_P0001058	0.401331	2.14E-01	1.43	-7.1021	0.0689164	AA345637	AACAATTTTAATTAGGGACAAAGAA	A_21_P0001058		ACADM			
24	A_33_P3232655	0.379949	1.95E-01	-1.5	-7.01344	-0.057785	AA354059	TTTCGCCACCATGACCAACTTCAGC	A_33_P3232655		ACADS			
25	A_21_P0004544	0.20202	6.94E-02	-2.31	-5.95351	-0.08567	AA371472	GCTAGCATAATTTTAGAGCGCTAAG	A_21_P0004544		ACADVL			
26	A_33_P3224095	0.231762	8.58E-02	2.14	-6.17869	2.6222833	AA378382	TTGTTTTCATGCTTGCAAGGATTGGC	A_33_P3224095		ACAT1			
27	A_24_P254506	0.276336	1.14E-01	-1.91	-6.47878	-0.069612	AA382521	ATGTAAAAGAGAAGACTCCACCTA	A_24_P254506		ACVRL1			
28	A_24_P759477	0.239354	9.02E-02	-2.1	-6.2323	-0.373204	AA382715	AGAATATGGTACTTTCACATAATAT	A_24_P759477		ADA			
29	A_23_P18267	0.208021	7.25E-02	2.27	-5.99949	0.1469642	AA383451	CCACCCCCAGCAGCTGCAGCATGTT	A_23_P1826		ADRB2			

总之，**一定要注意排序**，排序是**都有的数据**（这里是GB_ACC）

不允许有重复的’ row.names’ 解决问题

在R studio中，我们发现输入以下代码可能报错：

r 一键获取完整项目代码 | 复制

```r
df = read.csv("文件目录", header = T, row.names = 1)
```

一键获取完整项目代码 | 复制

```
不允许有重复的'row.names'解决问题
```

出错原因是**数据格式**不对，但这在网上**很少解释清楚**，常错的原因有以下几点：

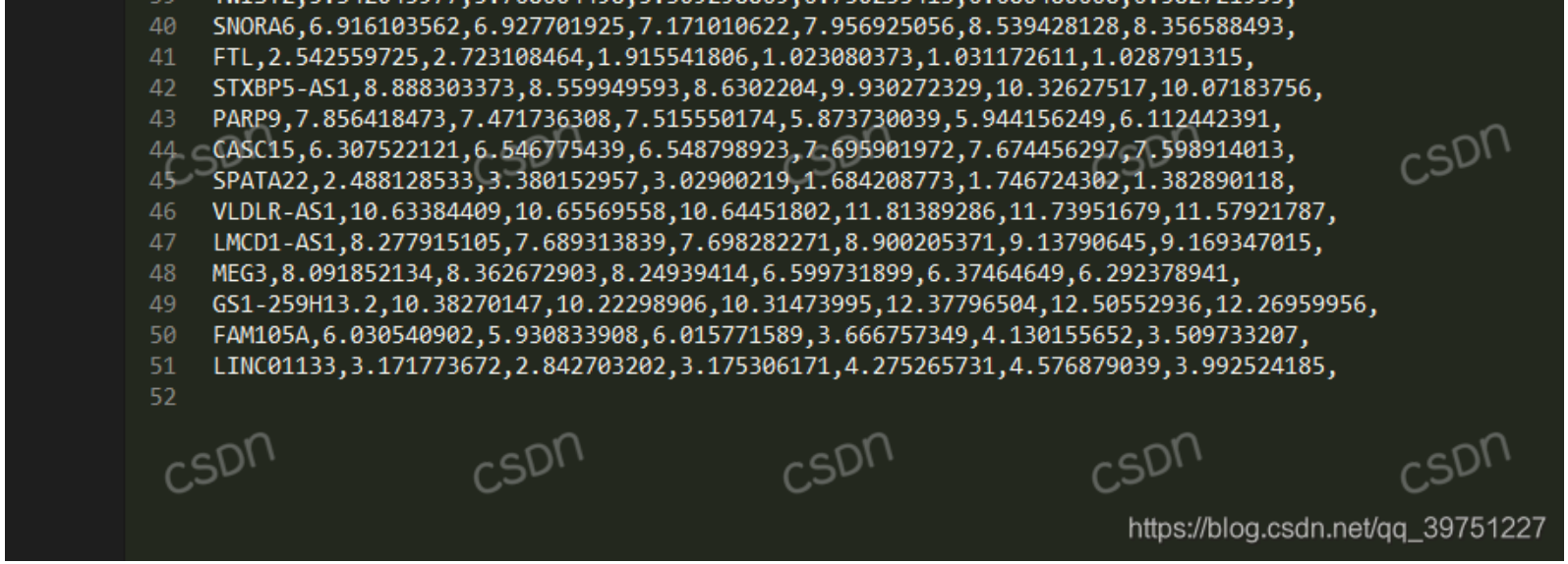
```
1.第一行有重复名
2.CSV文件格式错误
```

原因1用Excel**修改第一行重复名称**即可。

我们介绍2如何解决，这里我们先用 **sublime text** （软件，自行下载）打开文件修改即可，**（事实上，任何文本编辑器都可以）**。

我们需要分析的csv的数据格式如下：

```
1. 每行末尾必须有英文符号的逗号
2. 最后一行需要换行（即保留一行空白）
3. 如果存在末行连续逗号，要删除，如图。
4. 注意，开头没有空格
```



这样问题就基本解决了。