


A.

The file "9606_abund.txt" gives, for each human protein (Gn column), copy numbers. These roughly measure the average amount of each protein in a typical human cell.

Inspect data

Taxid	Ensembl_protein	Gn	Mean-copy-number
i64	str	str	f64
min: 9,606 max: 9,606 unique: 1	unique: 18,832	unique: 18,991	min: 0 max: 22,306.39 unique:
9606	ENSP00000263100	A1BG	885.188
9606	ENSP00000282641	A1CF	19.016
9606	ENSP00000282641	A1CF	19.016
9606	ENSP00000282641	A1CF	19.016
9606	ENSP00000323929	A2M	1114.564
9606	ENSP00000323929	A2M	1114.564
9606	ENSP00000323929	A2M	1114.564
9606	ENSP00000323929	A2M	1114.564
9606	ENSP00000299698	A2ML1	90.762
9606	ENSP00000299698	A2ML1	90.762

   53,641 rows, 4 columns 10 / page  Page 1 of 5,365  [Download](#)

****A1.1.**** How many protein/copy-number pairs are in the file? (Single numerical value)

53641

****A1.2.**** How many unique copy number values are there in the file? (Single numerical value)

16240

****A1.3.**** How many pairs of protein and copy number values are in the file? (Single numerical value)

19566

****A1.4.**** Please also answer the first three questions using a single command line operation in linux.

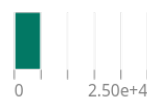
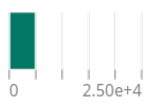

Answer: You can run the following cli commands in nushell (<https://www.nushell.sh/>)

```
#A1.1
open data/9606_abund.txt | from tsv | length
53641
#A1.2
open data/9606_abund.txt | from tsv | get `Mean-copy-number` | uniq | length
#A1.3
open data/9606_abund.txt | from tsv | each {|x| $"($x.Gn)_($x.Mean-copy-number)"} | uniq | length
```

****A2.**** Compute the mean and standard deviation of copy numbers for all proteins (considering unique pairs only) first as a single number for all proteins (two numerical values) and then for each protein separately (Table in tsv/csv).

For all proteins, mean is 89.26967075557874 and std is 415.88515929358914.

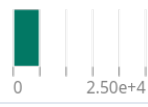
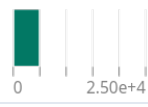
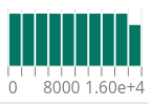
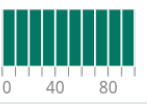
for each protein:

Gn	Mean-copy-number	mean	std
str	f64	f64	f64 (nulls: 7770)
unique: 18,991			
C11orf91	11.014	11.014	0
AMBRA1	17.092	17.092	0
CIPC	10.217	10.217	0
RSPH3	2.694	2.6940000000000004	0
VAV1	29.248	29.248	0
LPAR4	4.079	4.079	0
GTPBP8	29.005	29.005	0
PTTG1IP	84.806	84.806	0
MND1	22.982	22.982	0
PPM1D	10.6	10.6	0

19,566 rows, 4 columns 10 / page

Page 1 of 1,957 Download

****A3.**** Calculate the percentile rank (in terms of average copy number ranks) for each protein. (i.e. for protein X, where is it in the ranks from top (0%) to bottom (100%) in terms of abundance) (Table in csv/tsv).

Gn	Mean-copy-number	mean	rank	percentile_rank
str	f64	f64	f64	f64
unique: 18,991				
ERVMER34-1	3.853	3.853	3323.5	16.986098333844424
SLC8A1	32.358	32.358	12973	66.30379229275275
POLR2J3	11.528	11.528	7663.5	39.167433302667895
GPATCH11	33.683	33.683	13205.5	67.49207809465399
GTF2H2C_2	18.013	18.013	9721.5	49.68567923949709
SPRTN	20.914	20.914	10472	53.5214146989676
TRPV4	4.691	4.691	3998	20.433404886026782
PRAMEF6	1.288	1.288	1040.5	5.3178983951753045
GFOD2	15.533	15.533	9006	46.02882551364612
POU2F2	7.211	7.211	5659	28.92262087294286

19,566 rows, 5 columns 10 / page

Page 1 of 1,957 Download

Please also give the top ten proteins (highest abundance) as a list with the associated numerical values.

[('ALB', 100.0), ('HBA2', 99.99488909332516), ('HBB', 99.98977818665031), ('LALBA', 99.98466727997547), ('TMSB4X', 99.97955637330062), ('IGLJ1', 99.96933455995094), ('IGLL5', 99.96933455995094), ('IGLC1', 99.96933455995094), ('CSN1S1', 99.95911274660125), ('GAPDH', 99.9540018399264)]

B.

Proteins can contain one or more “domains” that are regions in the sequence that correspond to a particular function. The same domain can be seen in many different proteins and proteins can have many domains (i.e. it is a many-to-many relationship).

Here you’ll need to use the file “9606_gn_domains.txt” in combination with the file above (that is we want you to combine data from both files). This file gives, for each protein (Gn column) each domain (Domain column) that is present inside it.

Domains:

Gn str unique: 19,151	Domain str unique: 6,493	Start i64 min: 1 max: 34,256 unique: 3,730	End i64 min: 13 max: 34,347 unique: 3,791	Eval f64 min: 0 max: 200 unique:
A1BG	Ig	127	201	0.38
A1BG	Ig	217	300	3e-15
A1BG	Ig	31	110	0.0000082
A1BG	Ig	403	490	0.0019
A1BG	SpaA	327	352	44
A1CF	DND1_DSRM	447	523	2.3e-24
A1CF	RRM	138	199	4.4e-7
A1CF	RRM	233	296	6.7e-11
A1CF	RRM	58	124	2.4e-16
A2M	A2M	738	828	4.5e-31

65,884 rows, 5 columns 10 / page

Page 1 of 6,589 Download

Combined:

Taxid i64 min: 9,606 max: 9,606 unique: 1	Ensembl_protein str unique: 18,080	Gn str unique: 18,114	Mean-copy-number f64 min: 0 max: 22,306.39 unique:	Domain str unique: 6,419	Start i64 min: 1 max: 34,256 unique: 3,668	End i64 min: 13 max: 34,347 unique: 3,732	Eval f64 min: 0 max: 200 unique:
9606	ENSP00000263100	A1BG	885.188	Ig	127	201	0.38
9606	ENSP00000263100	A1BG	885.188	Ig	217	300	3e-15
9606	ENSP00000263100	A1BG	885.188	Ig	31	110	0.0000082
9606	ENSP00000263100	A1BG	885.188	Ig	403	490	0.0019
9606	ENSP00000263100	A1BG	885.188	SpaA	327	352	44
9606	ENSP00000282641	A1CF	19.016	DND1_DSRM	447	523	2.3e-24
9606	ENSP00000282641	A1CF	19.016	DND1_DSRM	447	523	2.3e-24
9606	ENSP00000282641	A1CF	19.016	DND1_DSRM	447	523	2.3e-24
9606	ENSP00000282641	A1CF	19.016	RRM	138	199	4.4e-7
9606	ENSP00000282641	A1CF	19.016	RRM	138	199	4.4e-7

195,856 rows, 8 columns 10 / page

Page 1 of 19,586 Download

****B1.**** What is the domain with the highest average abundance (i.e. across all copies of the domain in all proteins)? (single string value and two numerical values)

Domain with most abundance is Serum_albumin domain with an average abundance of 13314.00017073171

****B2.**** Compute the mean and standard deviation of domain average abundance for each protein domain (i.e. by summing abundance values of all versions of these domains) by combining these two files also, compute the percentile rank values as above (One table).

Domain str unique: 6,419	mean f64	std f64 (nulls: 839)	rank f64	percentile_rank f64
Serum_albumin	13314.00017073171	10821.545569013993	6419	100
Casein_kappa	4953.42		6418	99.98442124941579
ApoA-II	4771.998375	1632.1788792282662	6417	99.96884249883159
ApoC-I	4531.031	0	6416	99.9532637482474
Keratin_2_tail	3745.121		6415	99.93768499766318
Transthyretin	3512.367		6414	99.92210624707897
Casein	2989.446		6413	99.90652749649477
Gp_dh_C	2827.9868	3807.6067370665655	6411.5	99.88315937061847
Gp_dh_N	2827.9868	3807.6067370665655	6411.5	99.88315937061847
Apo-CII	2716.309	0	6410	99.85979124474217

6,419 rows, 5 columns 10 / page

☐ Page 1 of 642 ☐ [Download](#)