Molecular grammars of intrinsically disordered regions that span the human proteome

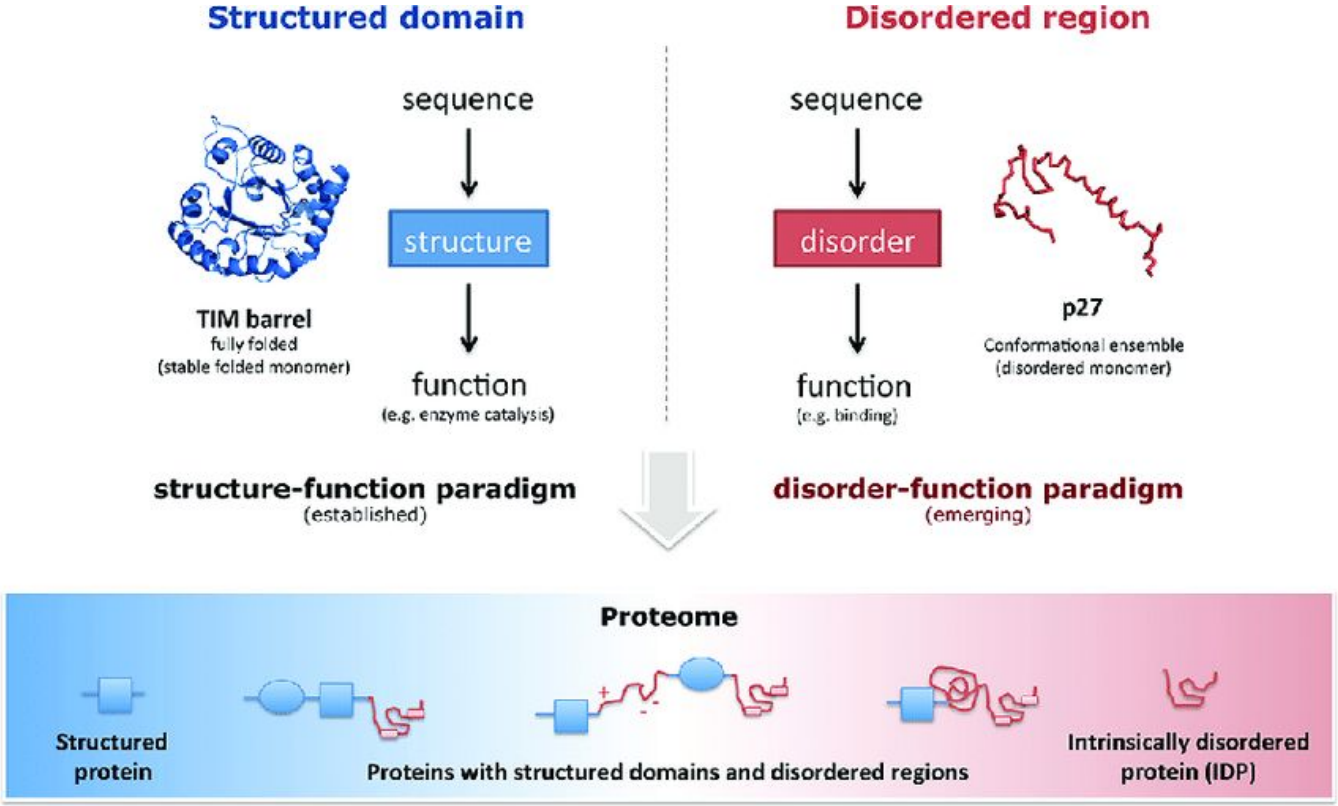
Kiersten M. Ruff, Matthew R. King, Alexander W. Ying Vicky Liu, Avnika Pant, Whitney E. Lieberman, Min Kyung Shinn, Xiaolei Su, Cigall Kadoch, Rohit V. Pappu

LifeLU reading group

presented by Özdeniz Dolu

10.04.2025

Intrinsically Disordered Region



Intrinsically Disordered Region

- In human proteome, more than 50% of proteins contain at least one IDR
- IDRs show conformational heterogeneity and poor sequence conservation

Authors attempt to associate molecular grammars with IDRs:

- "IDR-specific molecular grammars are defined jointly by the non-random amino acid composition and the non-random patterning of distinct pairs of amino acid types with respect to one another."

Proposed Approach

NARDINI+ Algorithm: Combination of NARDINI algorithm and compositional analyses inspired by other work in the literature.

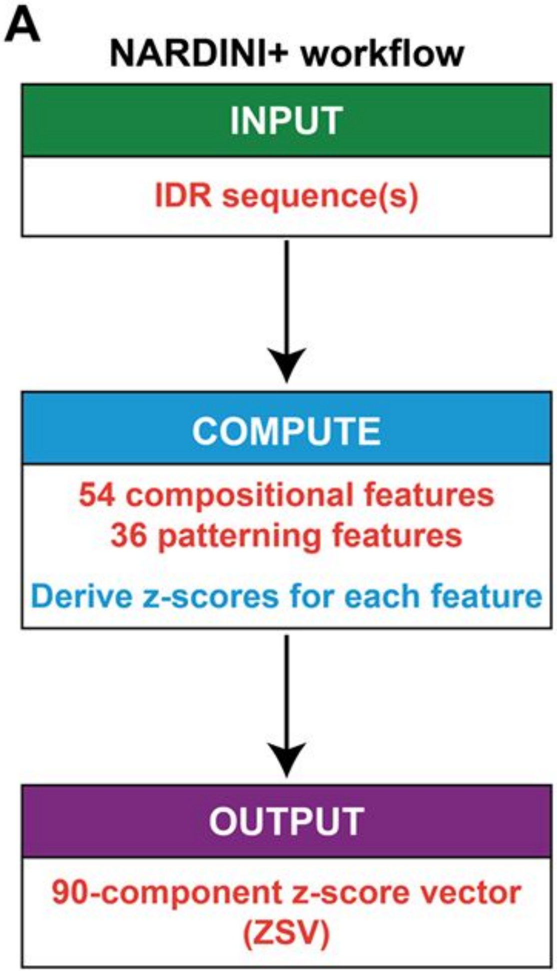
Through unsupervised learning methods and clustering, characterize IDRs in human proteome based on molecular grammars.

GIN: Grammars Inferred using NARDINI+

IDRome-spanning basis set, 30 clusters, unique fingerprint for each cluster

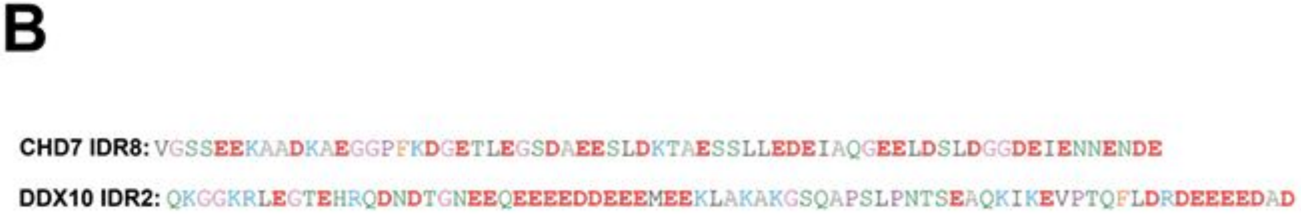
NARDINI+

- 36 non-random binary patterns computed by NARDINI algorithm.
- Statistics on alphabet symbols used in the IDR yield 54 compositional features.
- Combined into a z-score vector.


A
NARDINI+ workflow
INPUT
IDR sequence(s)
COMPUTE
54 compositional features
36 patterning features
Derive z-scores for each feature
OUTPUT
90-component z-score vector
(ZSV)

NARDINI+ (Compositional Features)

- The fraction of each of the 20 naturally occurring residues (20 features)
- Fractions of polar, aliphatic, aromatic, (Lys + Arg), (Asp + Glu), charged residue, residues that promote chain expansion, disorder promoting residues (8 features)
- “20 values that quantify the presence of specific residue or RG patches”
- Arg/Lys ratio, Glu/Asp ratio
- Net charge per residue (NCPR) (4 features)
- Apparent isoelectric point (pI), Kyte-Doolittle hydrophobicity

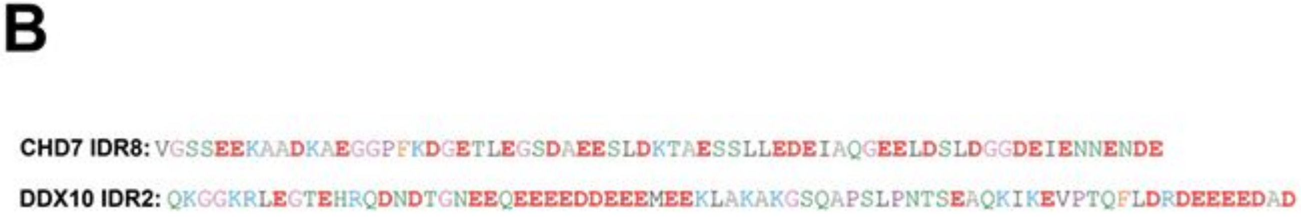


CHD7 IDR8: VGSSEEKAADKAEGGPFKDGETLEGSDAEESLDKTAESSLLEDEIAQGEELDSLDGGDEIENNENDE

DDX10 IDR2: QKGGKRLEGTEHRQDNDTGNEEQEEEEDDEEEMEEKLAKAKGSQAPSLPNTSEAQKIKEVPTQFLDRDEEEEDAD

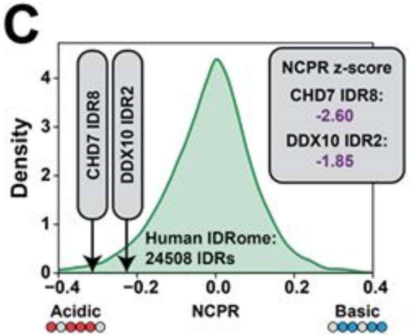
NARDINI+ (Compositional Features)

- The fraction of each of the 20 naturally occurring residues (20 features)
- Fractions of polar, aliphatic, aromatic, (Lys + Arg), (Asp + Glu), charged residue, residues that promote chain expansion, disorder promoting residues (8 features)
- “20 values that quantify the presence of specific residue or RG patches”
- Arg/Lys ratio, Glu/Asp ratio
- Net charge per residue (NCPR) (4 features)
- Apparent isoelectric point (pI), Kyte-Doolittle hydrophobicity



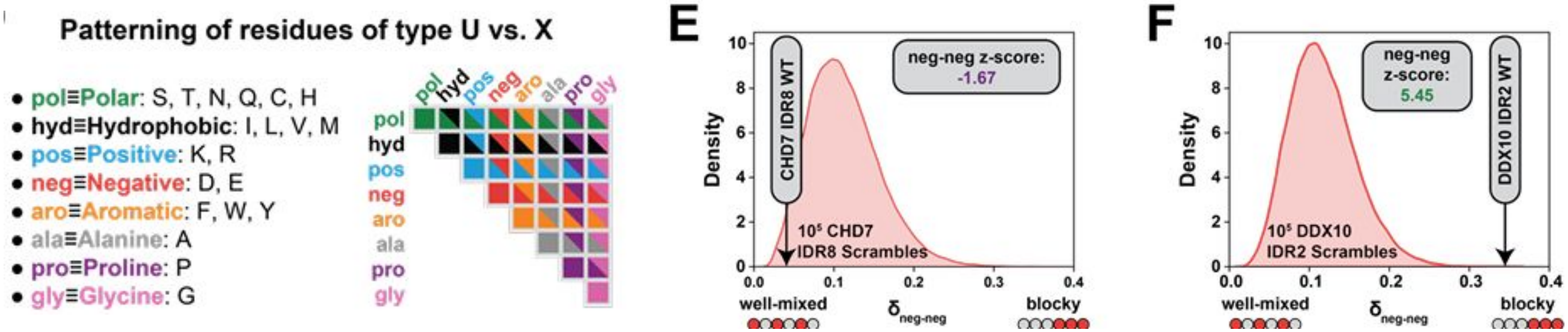
CHD7 IDR8: VGSSEEKAADKAEGGPFKDGETLEGSDAEESLDKTAESSLLEDEIAQGEELDSLDGGDEIENNENDE

DDX10 IDR2: QKGGKRLEGTEHRQDNDTGNEEQEEEEDDEEEMEEKLAKAKGSQAPSLPNTSEAQKIKEVPTQFLDRDEEEEDAD



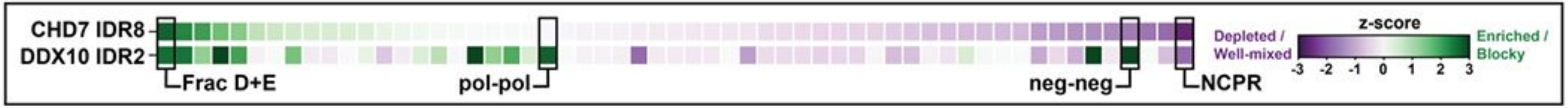
NARDINI+ (Binary Patterning Features)

- 36 binary features computed by NARDINI algorithm.
- Grouping of residues into 8 groups: polar, hydrophobic, positively charged, negatively charged, aromatic, alanine, glycine, and proline
- “For each unique pair of residue types U and X, the NARDINI algorithm computes a parameter δUX that quantifies the extent to which U and X are mixed or segregated along the linear sequence.”



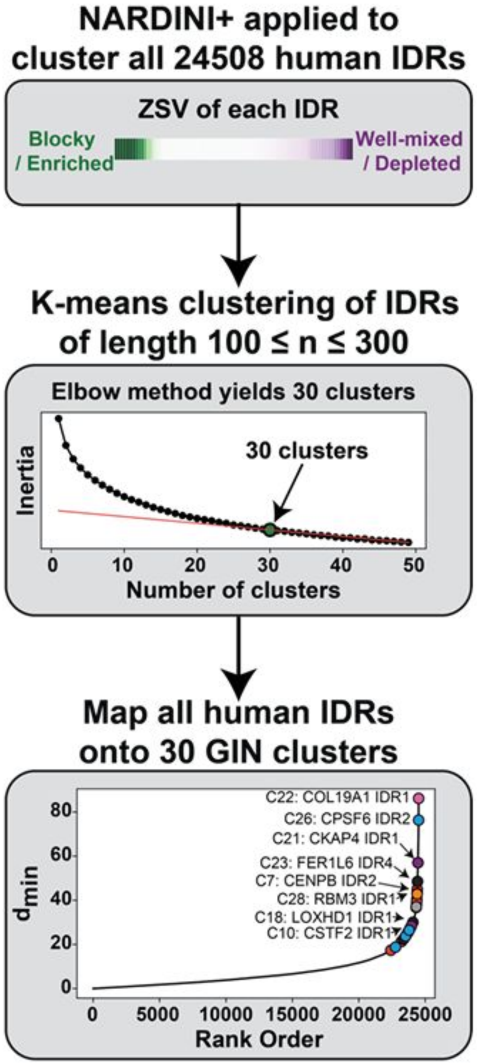
NARDINI+ (An Example)

Calculated features for 2 IDRs. Zero features are omitted.



Grammars Inferred using NARDINI+ (GIN)

- 4,529 IDRs in preferred length range (out of 24508) used to cluster.
- K-means clustering with K = 30
- **All human IDRs** then mapped to clusters.
- 30 clusters are referred as **GIN clusters**.



NARDINI+ applied to cluster all 24508 human IDRs

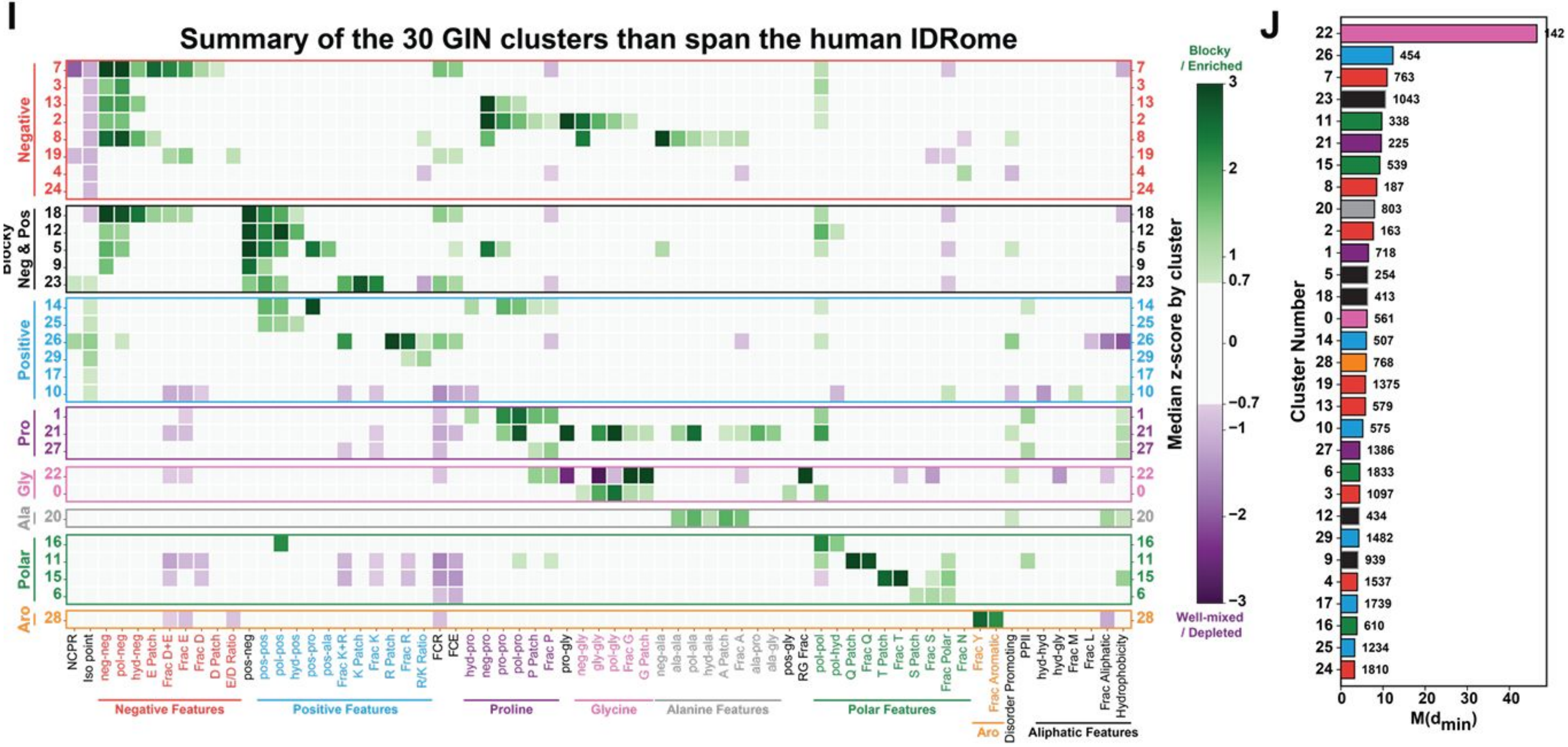
ZSV of each IDR

K-means clustering of IDRs of length 100 ≤ n ≤ 300

Elbow method yields 30 clusters

Map all human IDRs onto 30 GIN clusters

Grammars Inferred using NARDINI+ (GIN)

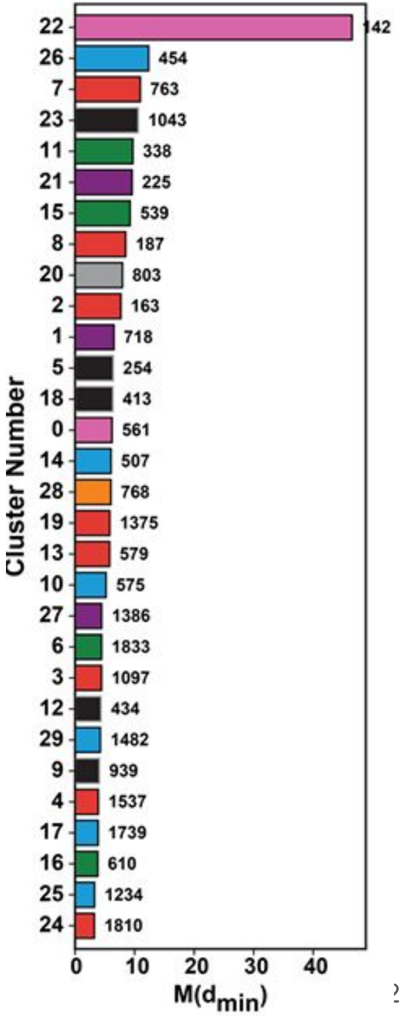


Grammars Inferred using NARDINI+ (GIN)

Higher M(dmin) (Median of minimum inter-cluster distance) values imply stronger mapping to clusters.

Higher number of IDRs in a cluster should result in a weaker mapping. However there are some exceptions.

For example, cluster 22 is very strongly mapped. 22 is uniquely defined by uniform distribution of Pro and Gly. Defining grammar for elastomers and collagens.



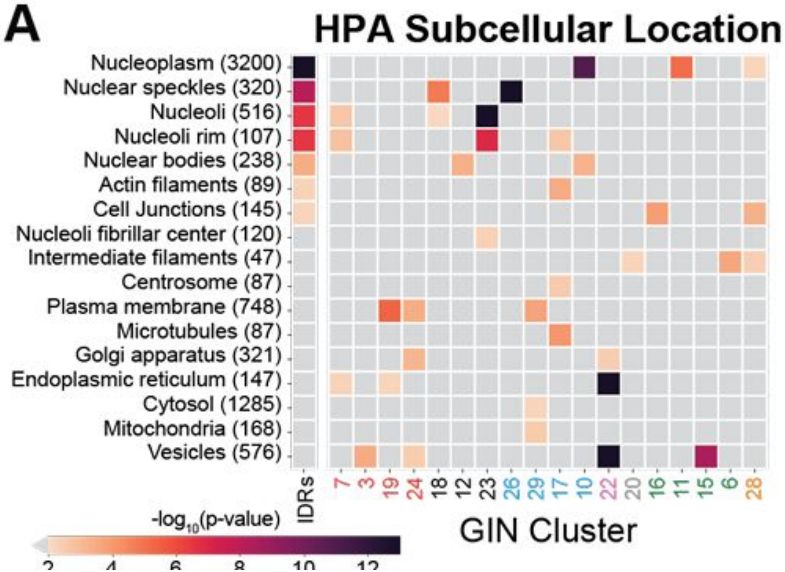
Results: GIN as a resource

- Authors provide 2 Google Colab notebooks.
- First notebook takes gene id or uniprot accessions and calculates GIN annotations and ZSV vectors.
- Second notebook takes IDRs as fasta or a list of sequences and make the calculations from scratch.

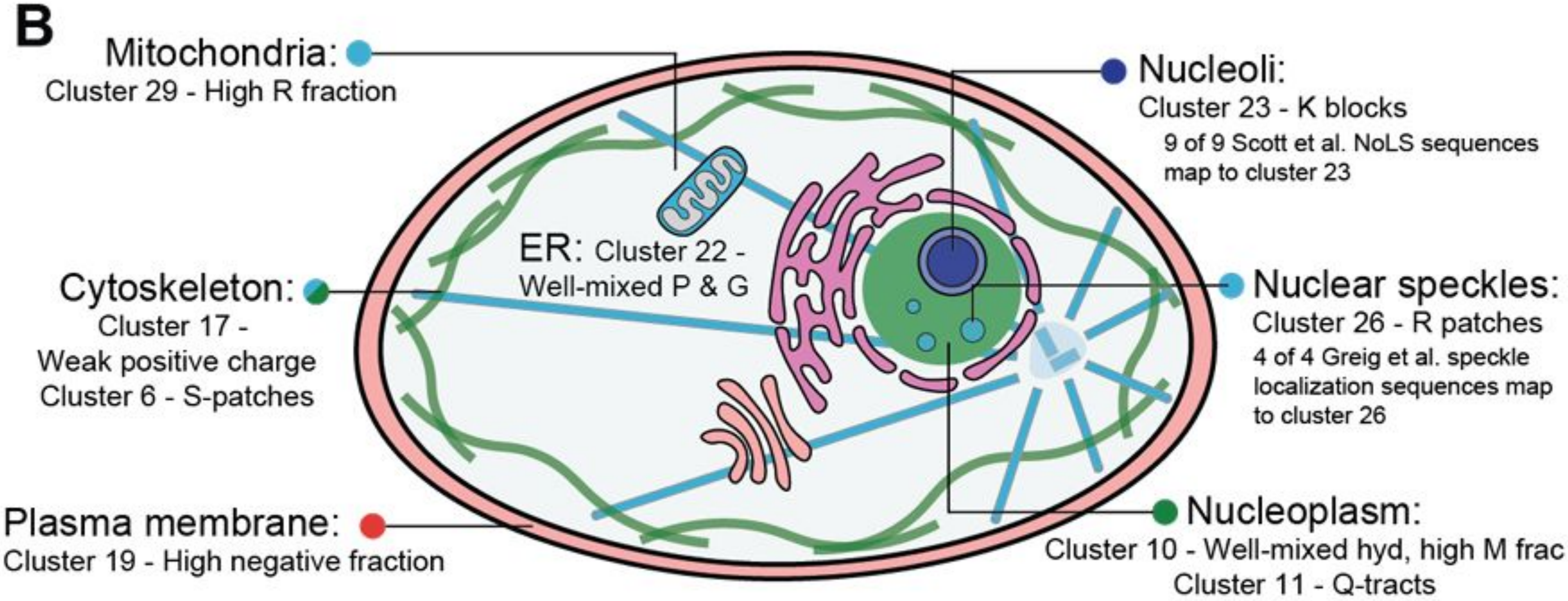
https://github.com/kierstenruff/RUFF_KING_Grammars_of_IDRs_using_NARDINI-

Results: GIN clusters -> Sub-nuclear Localization

- Annotations of subcellular locations of proteins with IDRs from the Human Protein Atlas (HPA)
- Focus on IDRs of length ≥ 70 and non-linker IDRs of length ≥ 50 and high dmin values.
- Cluster 26 strongly associated with Nuclear speckles
- Cluster 23 strongly associated with Nucleoli rim



Results: GIN clusters -> Sub-nuclear Localization



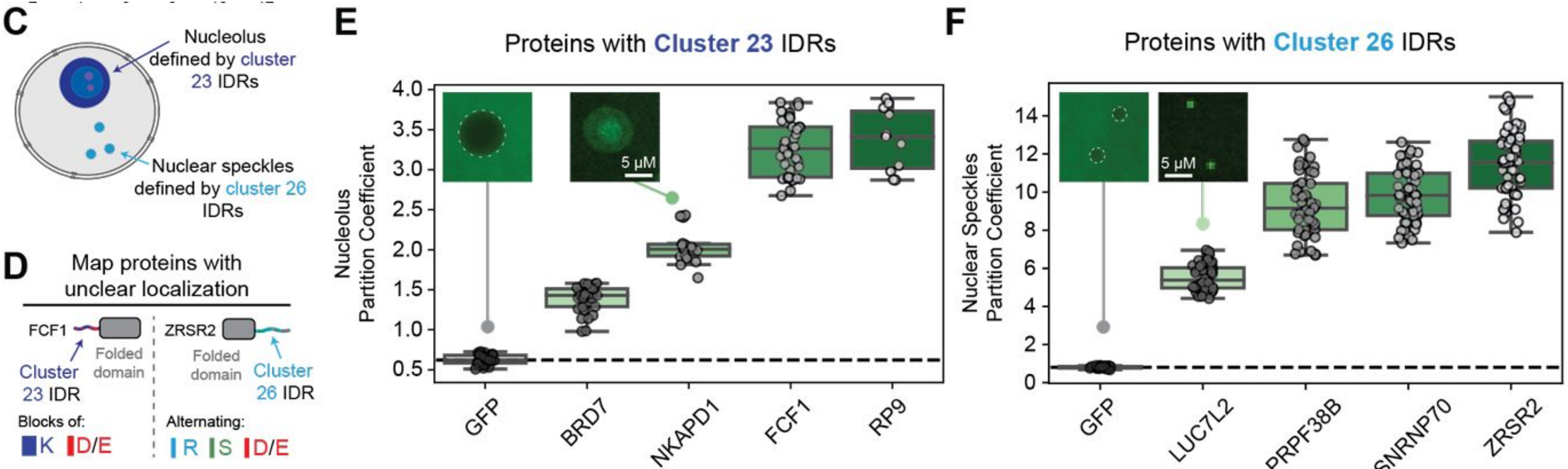
Results: GIN clusters -> Sub-nuclear Localization

For further testing the case for cluster 23 and 26:

- 5 monomeric proteins from HPA dataset whose IDRs are are clustered into either 23 or 26 are selected. These proteins have ambiguous annotations as to sub-nuclear localizations.
- Experiments are made on germinal vesicles (GVs) from live Xenopus laevis oocytes.
- Germinal Vesicle: Nucleus of an oocyte
- Oocyte: A developing egg

Results: GIN clusters -> Sub-nuclear Localization

- Except for GFP, proteins with cluster 23 are partitioned into Nucleolus and proteins with cluster 26 are partitioned into Nuclear Speckles.
- The **partition coefficient**, abbreviated ***P***, is defined as a particular ratio of the concentrations of a solute between the two solvents (a biphase of liquid phases) (Wikipedia)

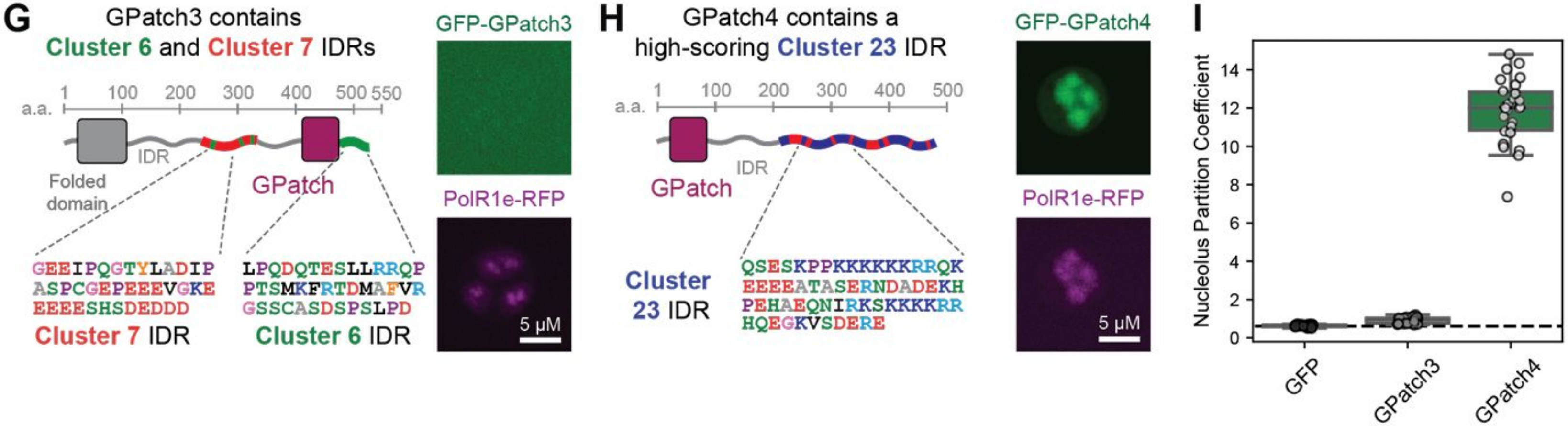


Results: GIN clusters -> Sub-nuclear Localization

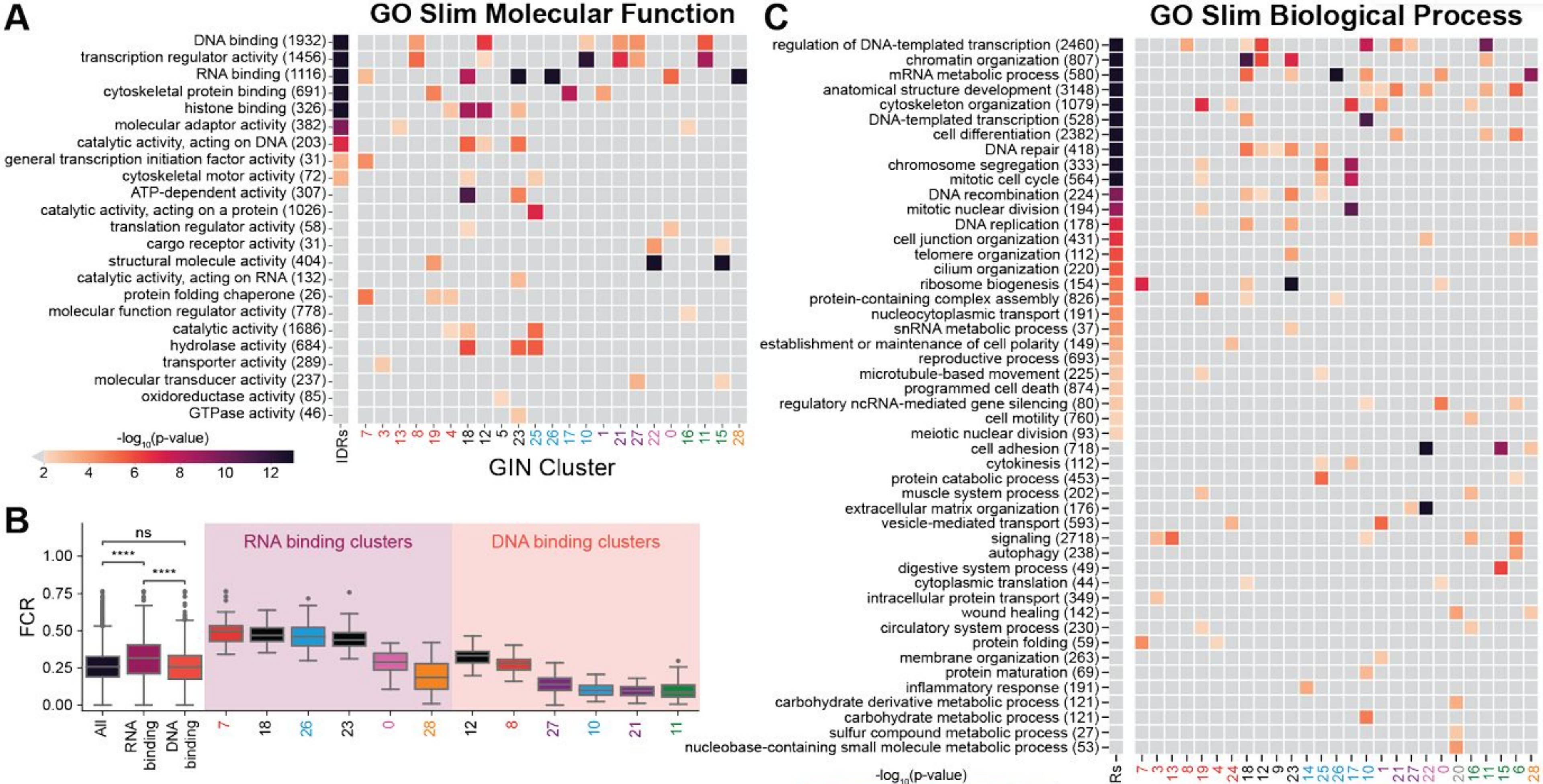
Two naturally occurring proteins GPatch3 and GPatch4 are investigated. Both contain GPatch domain. However, their localization preference are different.

Localization preferences of proteins same domains can be determined by IDRs.

Authors note that, in some cases the vice versa is true.



Results: GIN clusters -> MF and BP



Results: GIN clusters -> MF and BP

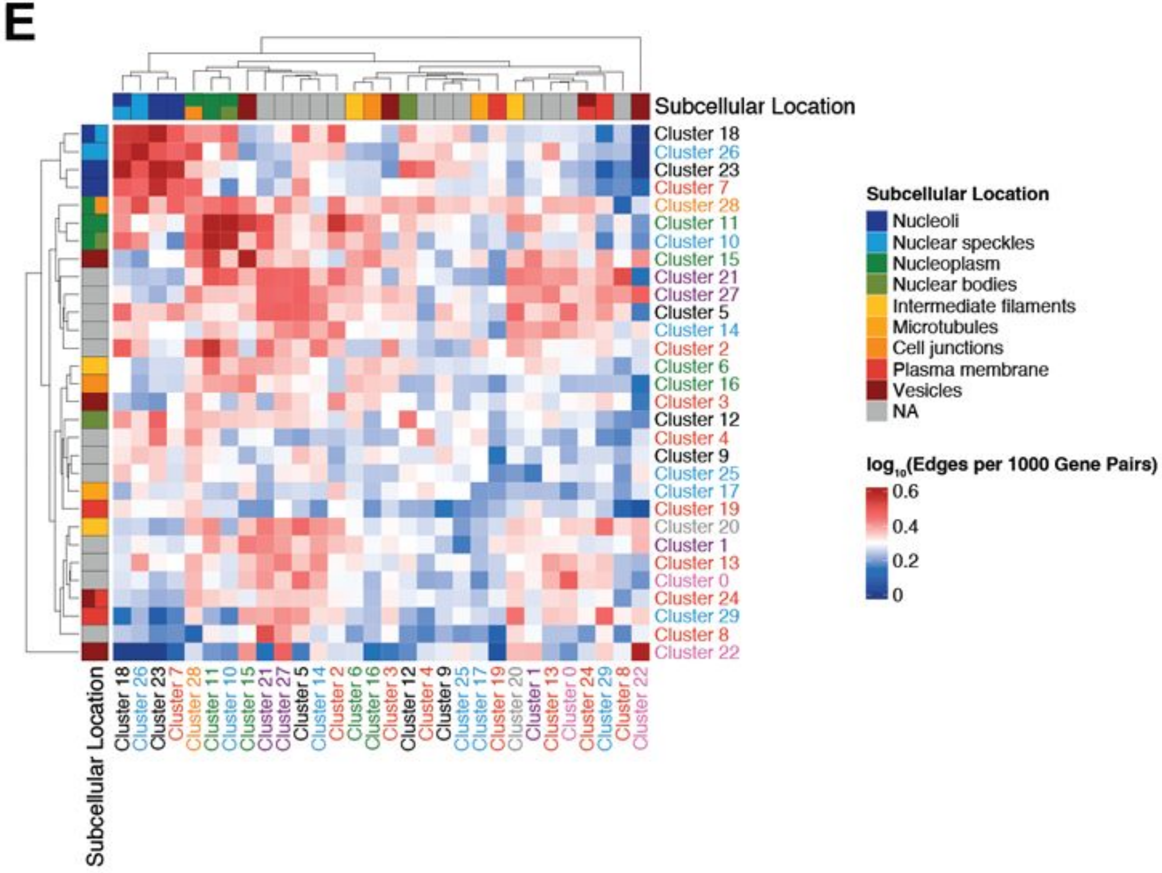
DepMap24Q4 Dataset

- “Genome-wide CRISPR knockout screening in over 1000 cancer cell lines”
- “Two genes (proteins) might be functionally linked if their fitness effects upon knockout across the cell lines are correlated”
- How proteins with IDR GIN cluster annotations are associated?

Results: GIN clusters -> MF and BP

Heatmap quantifying functional relationships based on **DepMap24Q4** dataset.

Higher value corresponds to more relationship.



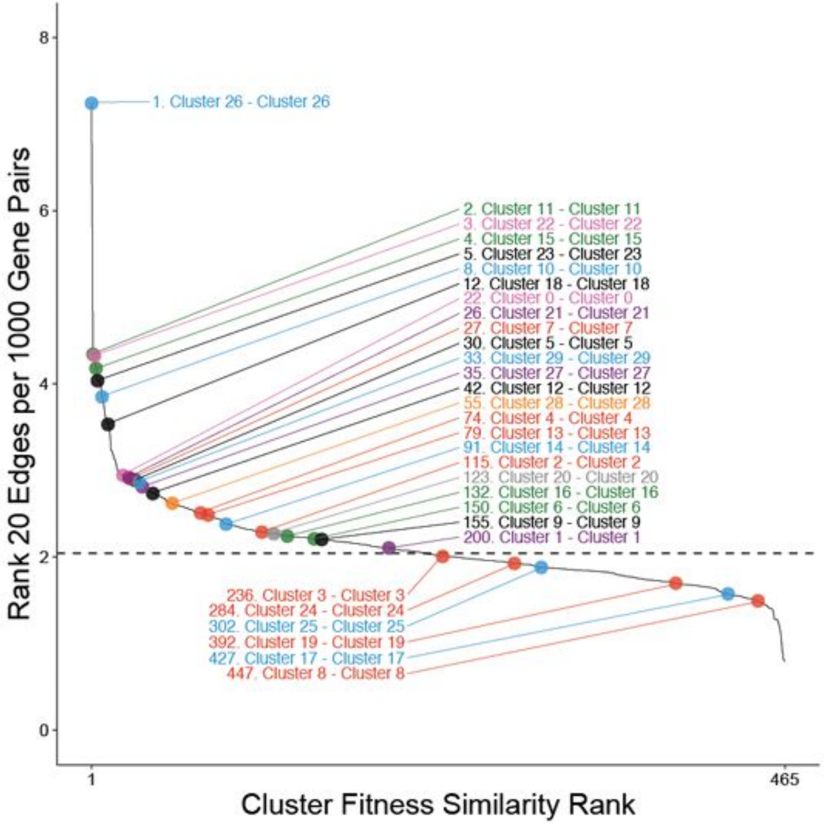
Results: GIN clusters -> MF and BP

Inter and intra cluster fitness correlations based on **DepMap24Q4**.

Only the intra-cluster data shown on right.

An interesting note:

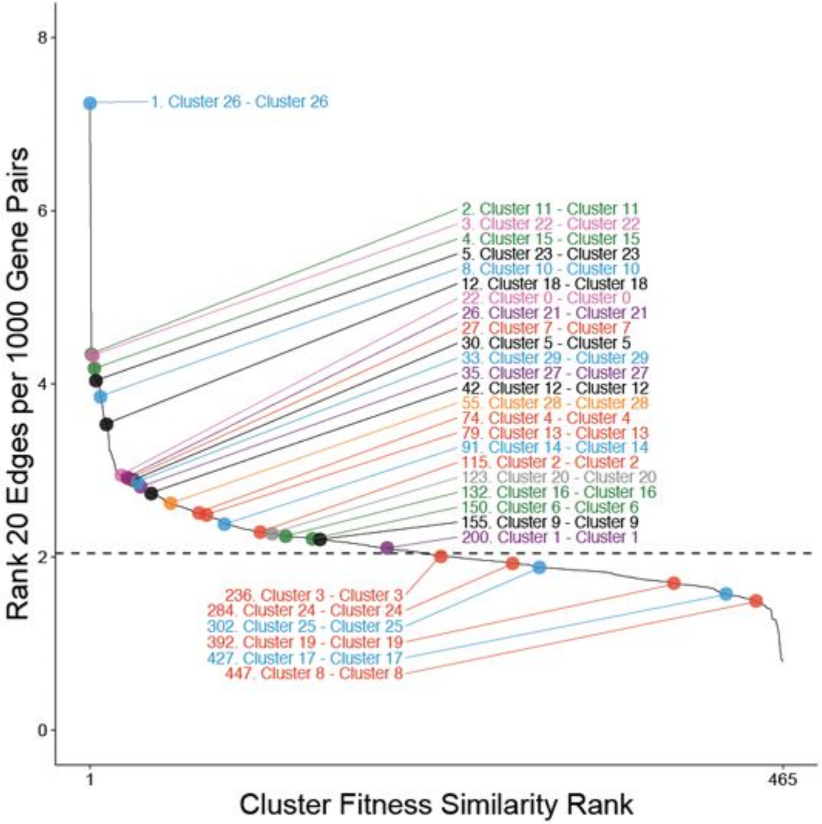
Top ranking clusters 26, 11, 22, 15, 23, and 10 are also highly enriched for specific subcellular locations. Suggesting that subcellular location might be an important signal for fitness data.



Results: GIN clusters -> MF and BP

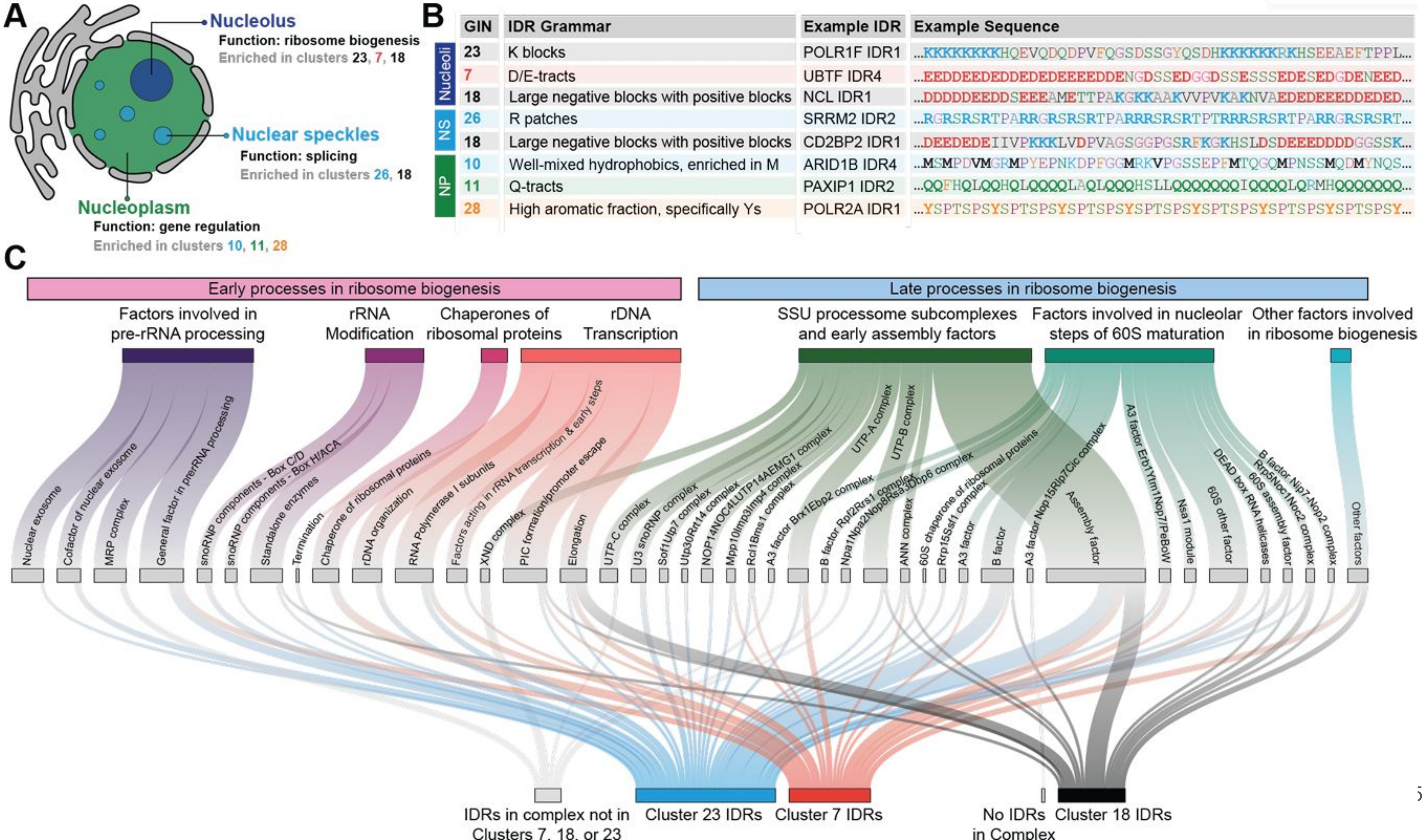
6 of 10 top-ranked gene (protein) fitness correlations are (within GIN clusters)

Genes/proteins that are associated some GIN clusters seem to have functional relationships.



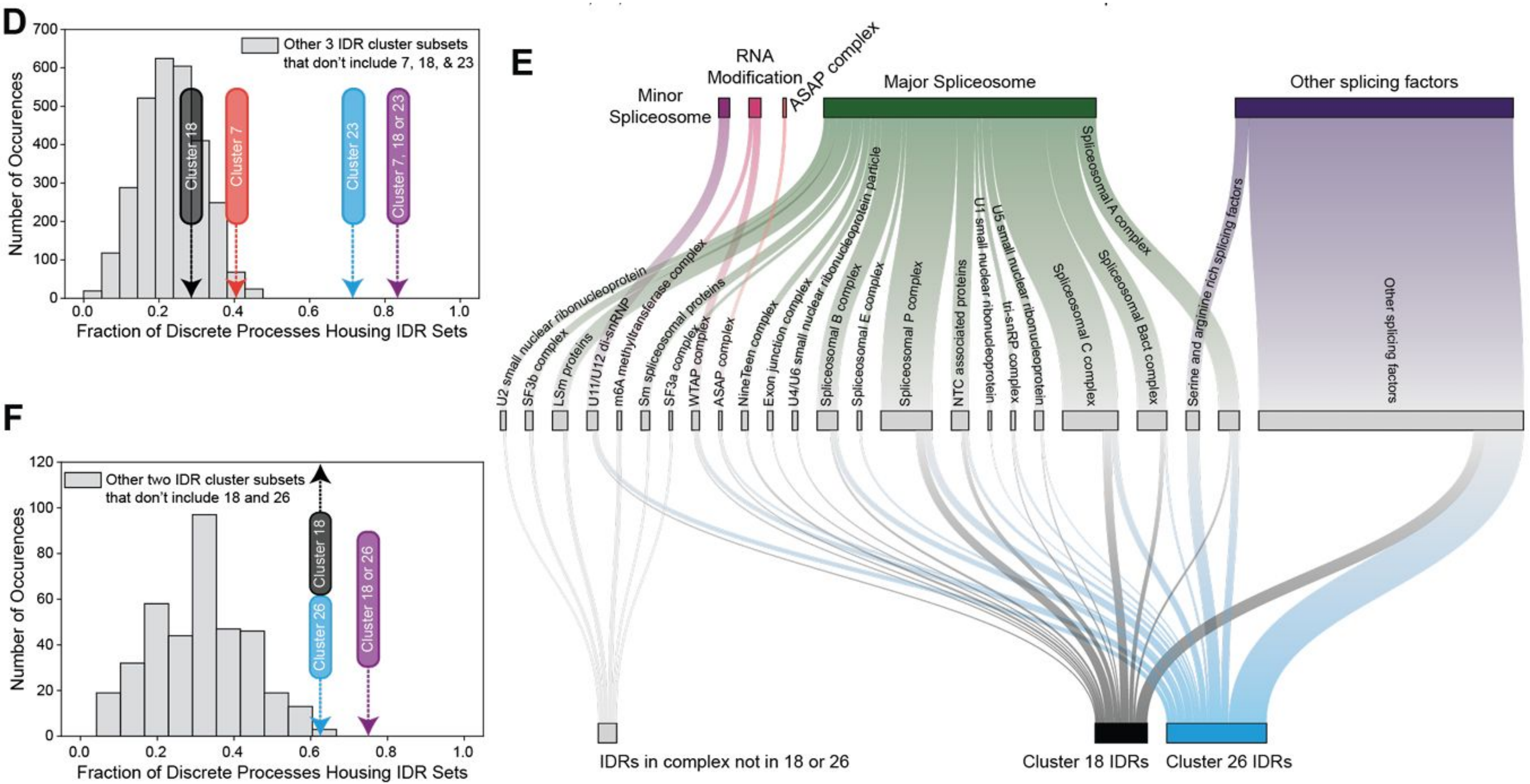
Results: GIN clusters -> Discrete Processes

Closer look into by combining known functions of subcellular locations and GIN cluster enrichment.


A
Nucleolus
Function: ribosome biogenesis
Enriched in clusters 23, 7, 18
Nuclear speckles
Function: splicing
Enriched in clusters 26, 18
Nucleoplasm
Function: gene regulation
Enriched in clusters 10, 11, 28
B

	GIN	IDR Grammar	Example IDR	Example Sequence
Nucleoli	23	K blocks	POLR1F IDR1	...KKKKKKKKHQEVQDQDPVFQGSDSSGYQSDHKKKKKKRKHSEEAEFTPPL...
	7	D/E-tracts	UBTF IDR4	...EEDDEEDEDDEDEDEEEEDDENGDSSEDGGDSSESSSEDESEDGDENEED...
	18	Large negative blocks with positive blocks	NCL IDR1	...DDDDDEEDDSEEEAMETTPAKGKKAAKVVPVKAKNVAEDEDEEEDDEDED...
NS	26	R patches	SRRM2 IDR2	...RGRSRSRTPARRGRSRSRTPARRRSRSRTPTRRRSRSRTPARRGRSRSRT...
	18	Large negative blocks with positive blocks	CD2BP2 IDR1	...DEEDEDEIIVPKKKLVDPVAGSGGPGSRFKGKHSLDSDEEEDDDDGGSSK...
NP	10	Well-mixed hydrophobics, enriched in M	ARID1B IDR4	...MSMPDVMGRMPYEPNKDPFGGMRKVPGSSEPFMTQGQMPNSSMQDMYNQS...
	11	Q-tracts	PAXIP1 IDR2	...QQFHQLQQHQLQQQQLAQLQQQHSLLQQQQQQQIQQQQLQRMHQQQQQQQ...
	28	High aromatic fraction, specifically Ys	POLR2A IDR1	...YSPTSPSYSPTSPSYSPTSPSYSPTSPSYSPTSPSYSPTSPSYSPTSPSY...

C
Early processes in ribosome biogenesis
Late processes in ribosome biogenesis
Factors involved in pre-rRNA processing
rRNA Modification
Chaperones of ribosomal proteins
rDNA Transcription
SSU processome subcomplexes and early assembly factors
Factors involved in nucleolar steps of 60S maturation
Other factors involved in ribosome biogenesis
Nuclear exosome
Cofactor of nuclear exosome
MRP complex
General factor in prerRNA processing
snoRNP components - Box C/D
snoRNP components - Box H/ACA
Standalone enzymes
Termination
Chaperone of ribosomal proteins
rDNA organization
RNA Polymerase I subunits
Factors acting in rRNA transcription & early steps
XND complex
PIC formation/promoter escape
Elongation
UTP-C complex
U3 snoRNP complex
Sof1Utp7 complex
Utp30Rrt14 complex
NOP14NOC4LUTP14AEMG1 complex
Mpp10Imp3Imp4 complex
Rcl1Bms1 complex
A3 factor Brx1Ebp2 complex
UTP-A complex
UTP-B complex
B factor Rpf2Rrs1 complex
Npa1Npa2Nop8Rsa3Dbp6 complex
ANN complex
60S chaperone of ribosomal proteins
Rrp15Ssf1 complex
A3 factor
B factor
A3 factor Nop15Rlp7Cic complex
Assembly factor
A3 factor Erb1Ytm1Nop7/PeBoW
Nsa1 module
60S other factor
DEAD box RNA helicases
60S assembly factor
Rrp5Noc1Noc2 complex
B factor Nip7-Nop2 complex
Other factors
IDRs in complex not in Clusters 7, 18, or 23
Cluster 23 IDRs
Cluster 7 IDRs
No IDRs in Complex
Cluster 18 IDRs


D
700
600
500
400
300
200
100
0
Number of Occurences
Other 3 IDR cluster subsets that don't include 7, 18, & 23
Cluster 18
Cluster 7
Cluster 23
Cluster 7, 18 or 23
0.0
0.2
0.4
0.6
0.8
1.0
Fraction of Discrete Processes Housing IDR Sets
E
Minor Spliceosome
RNA Modification
ASAP complex
Major Spliceosome
Other splicing factors
U2 small nuclear ribonucleoprotein
SF3b complex
LSm proteins
U11/U12 di-snRNP
m6A methyltransferase complex
Sm spliceosomal proteins
SF3a complex
WTAP complex
ASAP complex
NineTeen complex
Exon junction complex
U4/U6 small nuclear ribonucleoprotein particle
Spliceosomal B complex
Spliceosomal E complex
Spliceosomal P complex
NTC associated proteins
U1 small nuclear ribonucleoprotein
tri-snRP complex
U5 small nuclear ribonucleoprotein
Spliceosomal C complex
Spliceosomal Bact complex
Spliceosomal A complex
Serine and arginine rich splicing factors
Other splicing factors
IDRs in complex not in 18 or 26
Cluster 18 IDRs
Cluster 26 IDRs
F
120
100
80
60
40
20
0
Number of Occurences
Other two IDR cluster subsets that don't include 18 and 26
Cluster 18
Cluster 26
Cluster 18 or 26
0.0
0.2
0.4
0.6
0.8
1.0
Fraction of Discrete Processes Housing IDR Sets

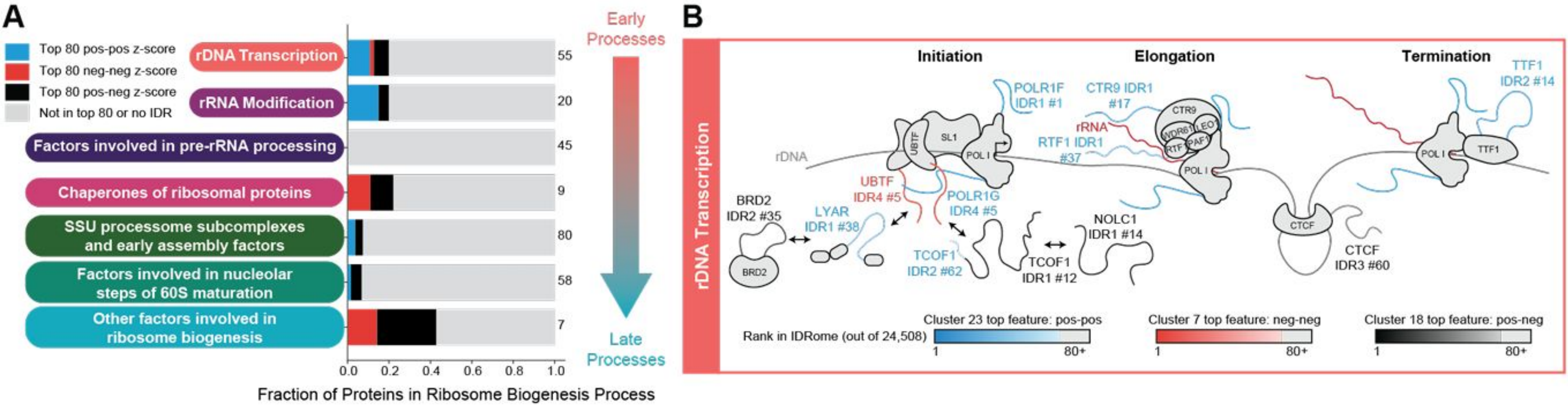
Results: Exceptional Grammars

All 24508 IDRs in human proteome is sorted by a “given” GIN cluster feature.

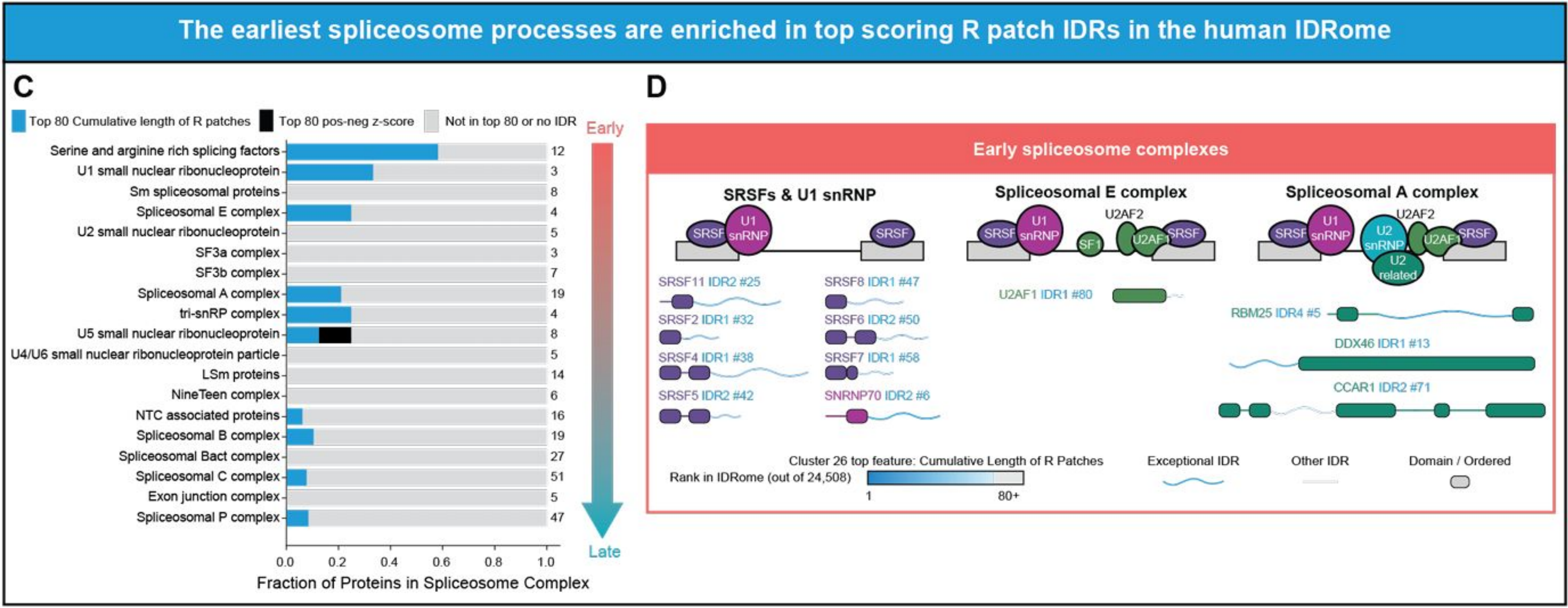
IDRs in 99th percentile are examined.

Results: Exceptional Grammars

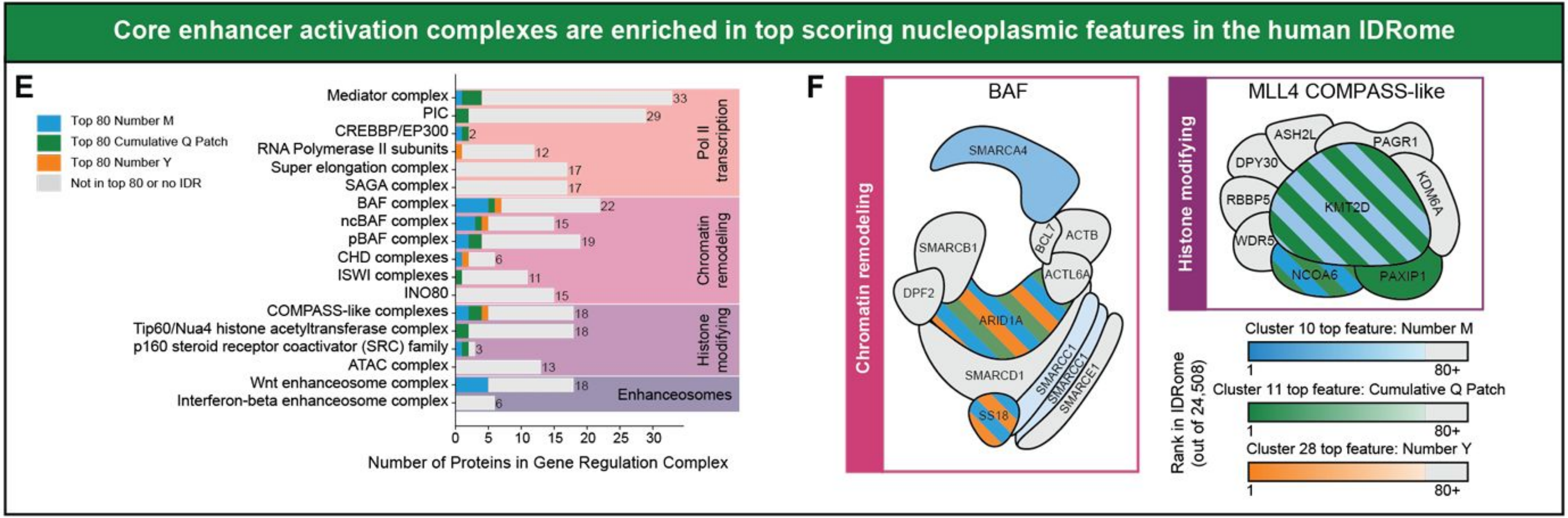
The earliest ribosome biogenesis processes are enriched in top scoring K blocks IDRs in the human IDRome (24,508 IDRs)



Results: Exceptional Grammars


The earliest spliceosome processes are enriched in top scoring R patch IDRs in the human IDRome

Results: Exceptional Grammars



Core enhancer activation complexes are enriched in top scoring nucleoplasmic features in the human IDRome

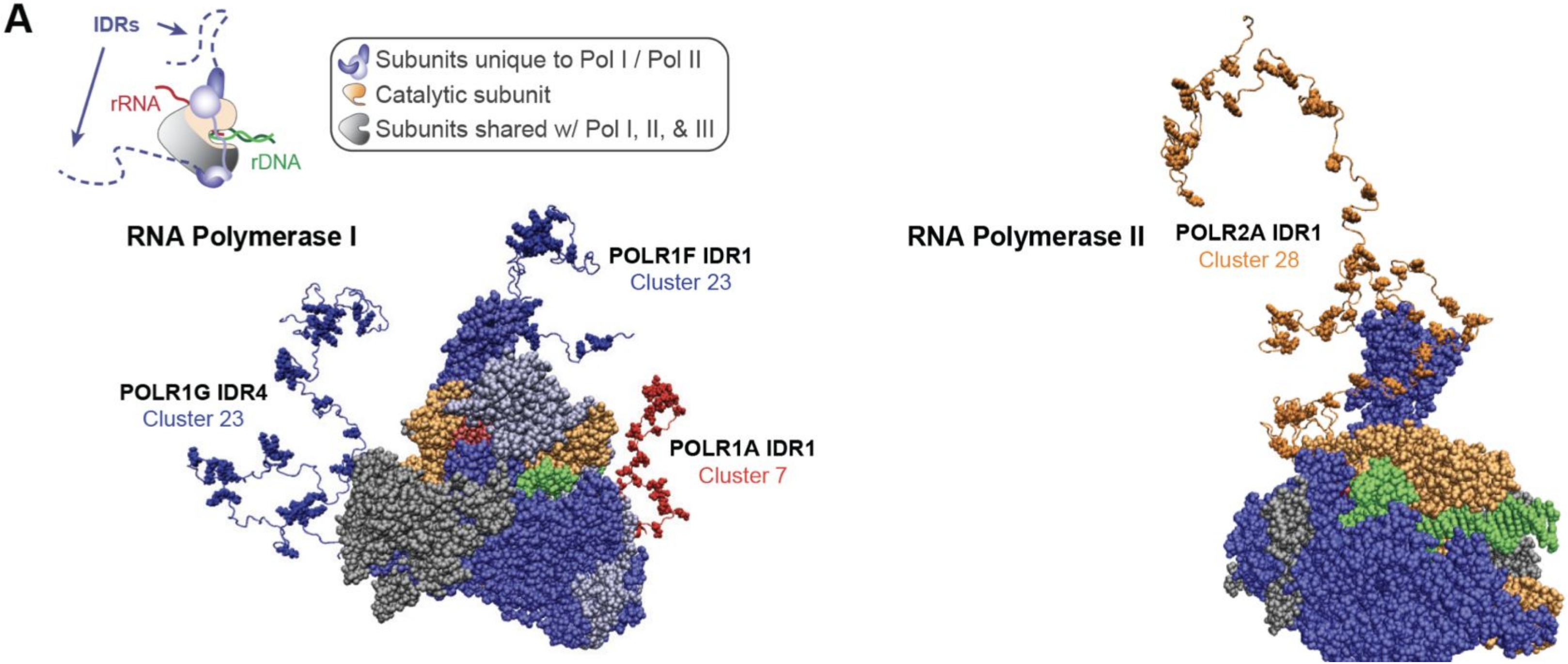
Results: RNA Pol I & II and evolutionarily conserved exceptional grammars

RNA polymerases play an essential role in transcription.

The subunit-specific IDRs of Pol I and Pol II have grammars with exceptional features that define clusters 23 and 28, respectively.

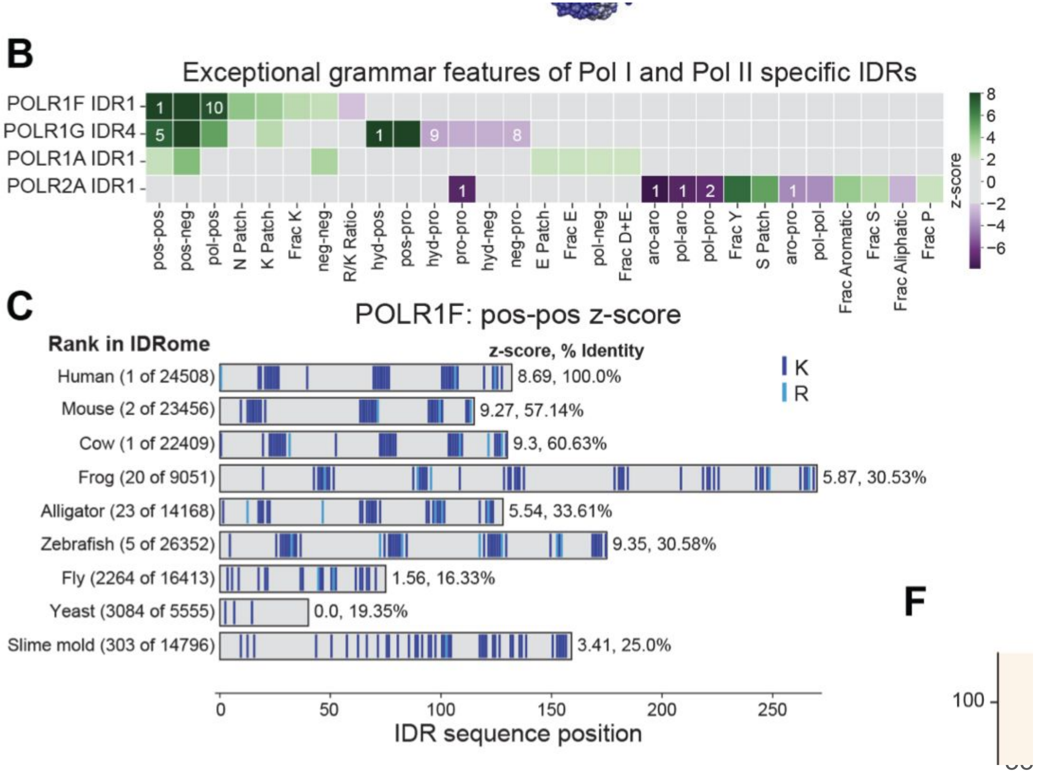
Orthologs of these IDRs throughout IDRomes of 8 species is investigated.

Results: RNA Pol I & II and evolutionarily conserved exceptional grammars



Results: RNA Pol I & II and evolutionarily conserved exceptional grammars

Although sequences differ throughout,, “exceptionality” of POLR1F is conserved in many species.

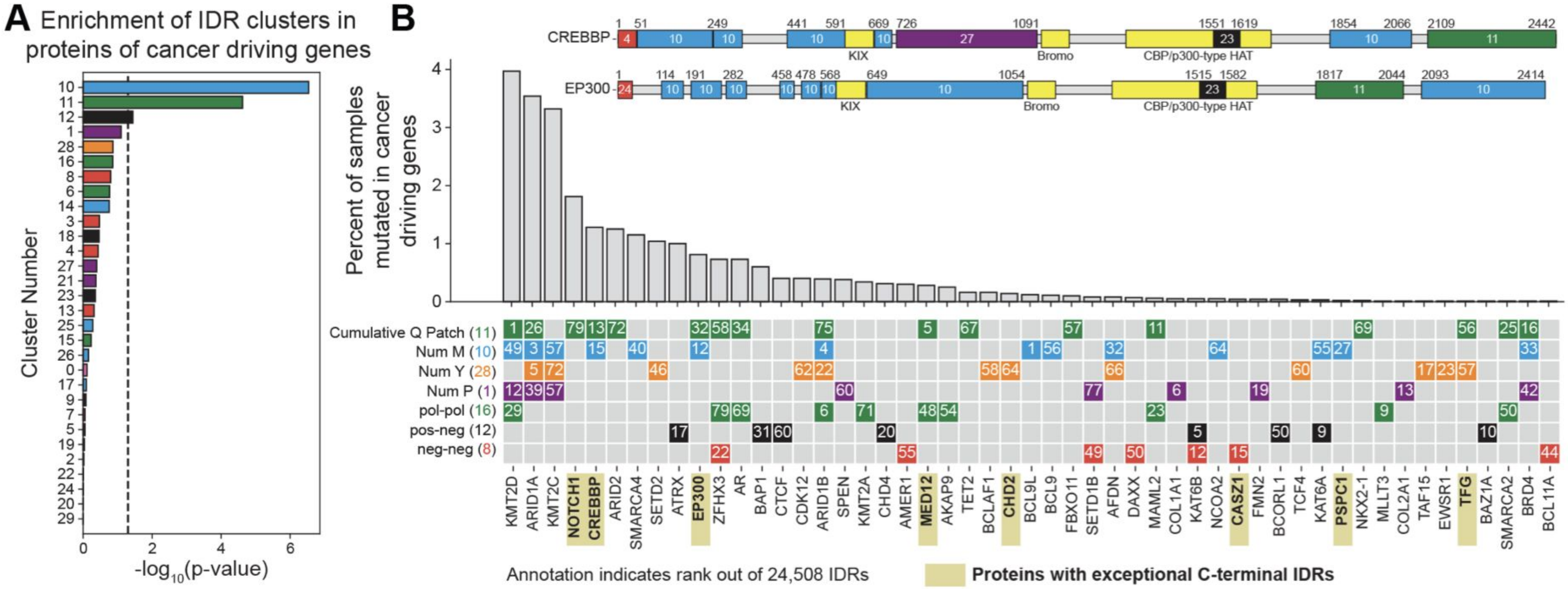


Results: RNA Pol I & II and evolutionarily conserved exceptional grammars

Figures and analyses containing further investigation and atomic simulations are also included in Figure 6. However, omitted from this presentation.

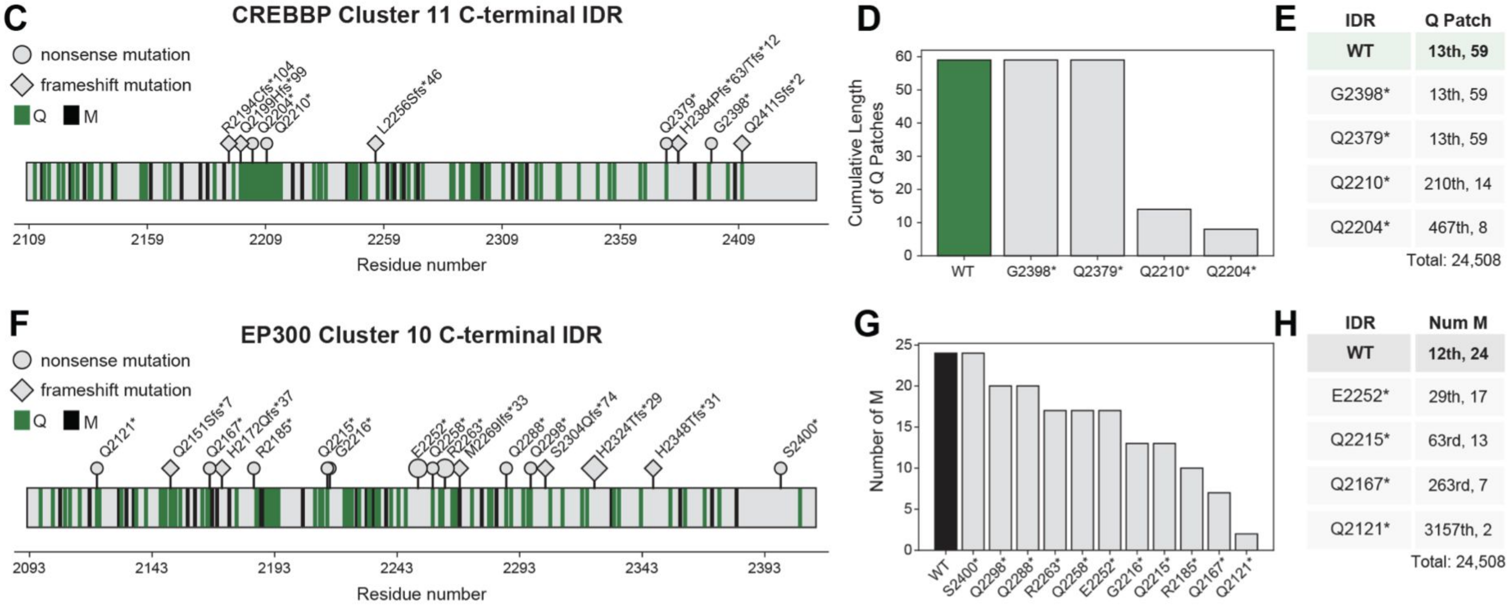
Results: Cancer mutations disrupt exceptional grammars

Cancer driver genes are investigated. Proteins associated with them are seen to be enriched in clusters 10, 11, 12.



Results: Cancer mutations disrupt exceptional grammars

Cancer driving mutations of C-terminus IDRs of CREBBP and EP300 can lead to “exceptionality loss” for grammars.



E

IDR	Q Patch
WT	**13th, 59**
G2398*	13th, 59
Q2379*	13th, 59
Q2210*	210th, 14
Q2204*	467th, 8

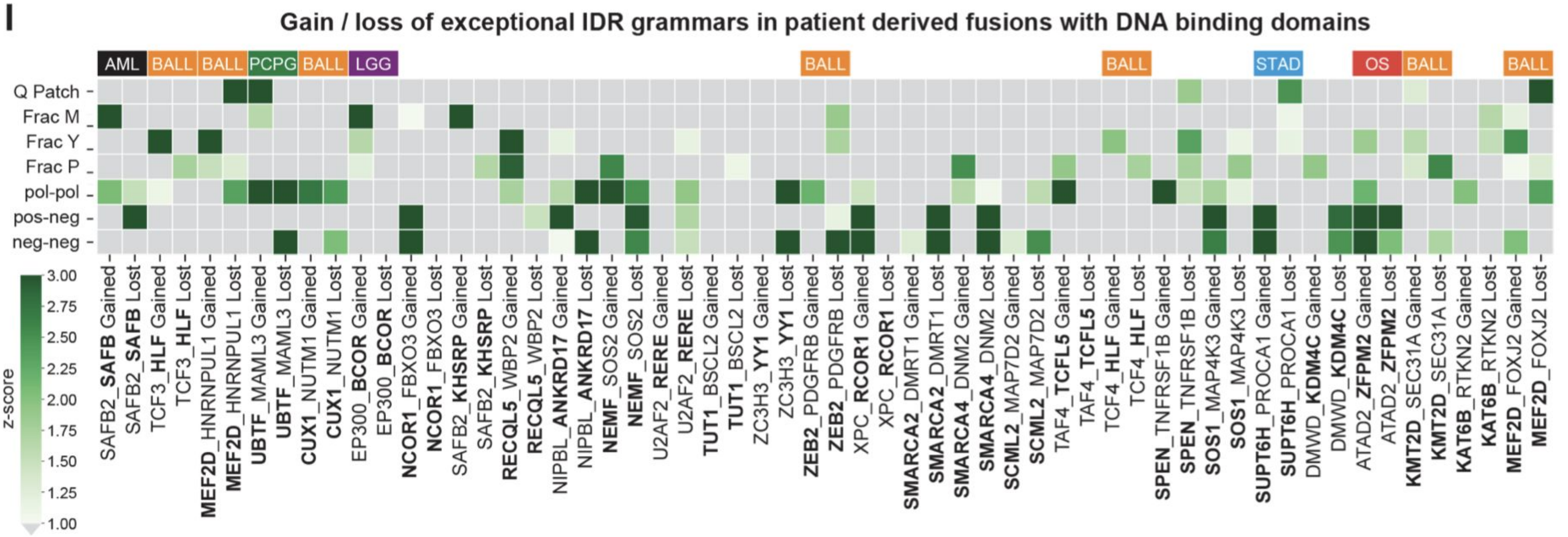
Total: 24,508

H

IDR	Num M
WT	**12th, 24**
E2252*	29th, 17
Q2215*	63rd, 13
Q2167*	263rd, 7
Q2121*	3157th, 2

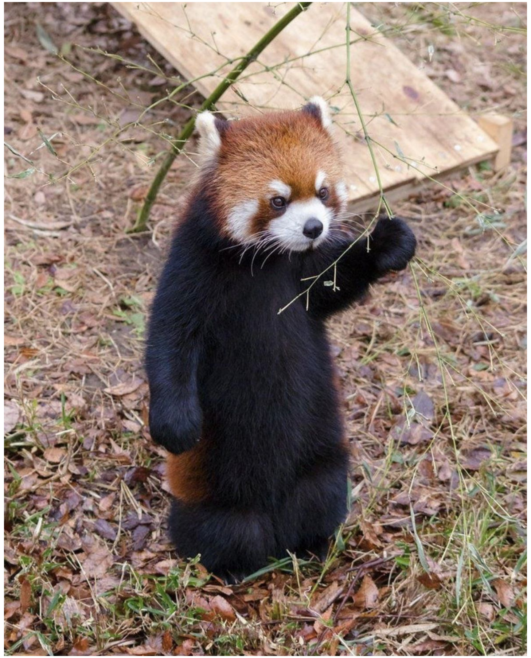
Total: 24,508

Results: Cancer mutations disrupt exceptional grammars



I

Gain / loss of exceptional IDR grammars in patient derived fusions with DNA binding domains



Thanks for listening