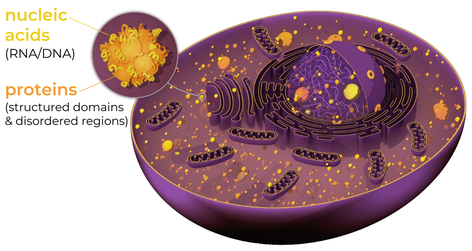
Rare missense variants in protein intrinsically disordered regions: impact on condensates and common diseases

Ruoyu Tian[1], Chi Zhang[1], Antonio Domingues[2], Isaac Klein[1]

1. Dewpoint Therapeutics, Boston, MA, USA; 2. Dewpoint Therapeutics GMBH, Dresden, Germany

Introduction

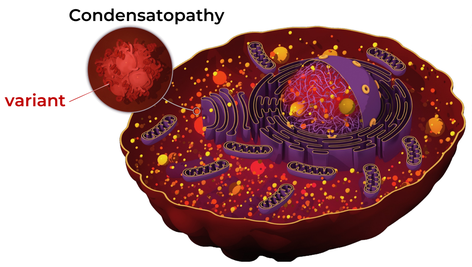
Biomolecular condensates are membrane-less compartments that concentrate a subset of biomolecules through phase separation from the surrounding cellular milieu and control numerous biological processes.



Functions

- Transcription
- Signaling
- RNA metabolism
- Stress response

Growing evidence supports that condensatopathies, aberrant condensates that serve as central nodes of dysregulation, drive the pathogenesis of various diseases, such as cancer, cardiomyopathy, neurodegeneration, and viral infections.



Aberrant condensates drive disease by:

- toxic gain of function
- loss of function of sequestered components
- inappropriate formation
- inappropriate dissolution
- altered material properties

Intrinsically disordered regions (IDRs) within proteins, classically considered undruggable, can contribute to condensate formation. Disease mutations located in IDRs perturb normal phase separation, leading to dysregulation of biological processes and pathogenesis. Here, we explored the role of genetic variation of IDRs in common diseases and their related traits to identify new condensate drug targets and biomarkers.

Methods

Whole-exome sequencing

- 454,787 whole-exome sequencing data from the UK Biobank participants was generated and quality controlled (QC-ed) by the Regeneron Genetics Center (RGC).[1]

Sample QC and population assignment

- We performed population assignment as described in ref. 2. We obtained 438,817 European samples for the downstream analyses.

Variant annotation

- Missense variants were annotated by the degree of deleterious effects (likely, possibly and unlikely), as described in ref. 1.
- We then used Metapredict[3] to predict 18,985 protein IDRs based on their amino acid sequence on the canonical transcript.

Gene-level exome-wide association analyses (ExWAS)

- To identify IDR-enriched risk genes, we performed gene-level burden tests (REGENIE[4], v3.1.1) by fitting rare variant burden to disease case-control status or quantitative phenotypes and controlled for population structures with the top 20 principal components, age and sex.
- We examined 16 ICD10-defined common diseases and 17 related quantitative traits in individuals of European genetic ancestries.
- In each gene, we aggregated rare missense variants into the following 4 groups: likely and possibly deleterious missense variants in IDRs and out of IDRs. Rare variants were tested at 3 minor allele frequency cutoffs: < 1e-3, < 1e-5 and singleton.
- Exome-wide significance threshold: $P < 2.5e\text{-}6 = 0.05/20{,}000$

Results – Common diseases

We tested the effect of IDR mutations in common diseases to identify condensate-related risk genes.

Category	Disease	Number of cases	Number of controls	Prevalence (European)	ICD10
Autoimmune	Rheumatoid arthritis	7,993	374,970	2.1%	M06
	Inflammatory bowel disease	6,480	376,483	1.7%	K50, K51
	Psoriasis	4,635	378,328	1.2%	L40
	Type 1 diabetes	3,995	378,968	1.0%	E10
Cardiovascular	Atrial Fibrillation	31,424	351,539	8.2%	I48
	Coronary artery disease	14,616	368,347	3.8%	I21
	Heart Failure	14,398	368,565	3.8%	I50
	Stroke	11,360	371,603	3.0%	I60, I61, I63, I64
Eye	Other retinal disorders	9,394	373,569	2.5%	H35
Metabolic	Type 2 diabetes	31,941	351,026	9.1%	E11
	Non-alcoholic steatohepatitis	5,354	377,609	1.4%	K758, K760
Neurological	Major depressive disorder	25,298	357,665	6.6%	F32, F33
	Anxiety disorder	18,018	364,945	4.7%	F41
	Insomnia	9,719	373,244	2.5%	G47
Respiratory	Asthma	39,857	343,106	10.4%	J45
	Chronic obstructive pulmonary disease	18,564	364,399	4.8%	J44

Showing diseases with ≥1% case prevalence

Exome-wide significant IDR enriched gene-disease associations ($P < 2.5e\text{-}6$)

Disease	Gene	Variant type	Region	OR (95%CI)	P-value
Inflammatory bowel disease	*ZNF366*	Possibly deleterious missense (singleton)	IDR	12.93 (4.56-36.64)	1.47e-6
			nonIDR	12.25 (0.31-485.41)	1.82e-1
Heart failure	*CEP290*	Likely missense (MAF<1e-5)	IDR	22.27 (6.27-79.13)	1.61e-6
			nonIDR	1.40 (0.85-2.33)	1.90e-1
Non-alcoholic steatohepatitis	*TGIF2*	Possibly deleterious missense (MAF<1e-5)	IDR	7.63 (3.36-17.33)	1.19e-6
			nonIDR	0.36 (0.00-37.87)	6.69e-1

*OR: odds ratio; CI: confidence interval.

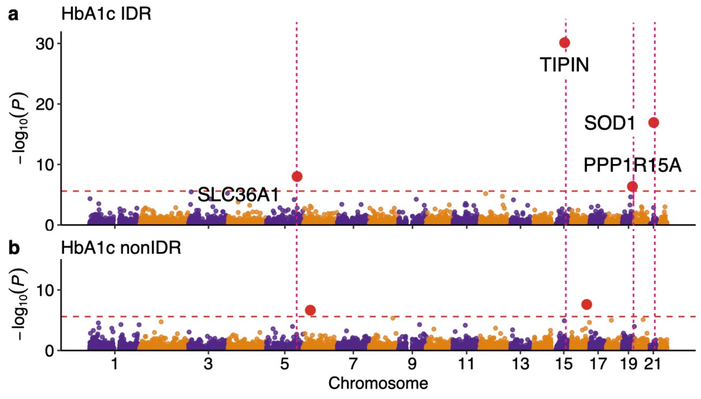
- 1/3 of the missense variants are located in IDRs across 80.9% protein-coding genes.
- A single missense variant (S332R) of *ZNF366* has been previously identified in a family study of inflammatory bowel disease.[5]
- Both ZNF366 and TGIF2 are transcription factors (TFs).
- CEP290 contains 13 coiled-coil domains.
- Given current knowledge on the role of TFs and coiled-coil domains in condensates,[6] our findings suggest that ZNF366, TGIF2 and CEP290 may be involved in condensates that serve as central nodes of dysregulation in common diseases, and that mutations within these genes affect disease risk through disrupting the normal function of these condensates.

Results - Quantitative traits

Gene-trait associations provide a proxy for disease risk and an opportunity to identify new disease markers.

Category	Trait
Body measure	BMI, waist-to-hip-ratio
Blood pressure	systolic blood pressure, diastolic blood pressure
Metabolic	glucose, hbA1c, IGF1
Lipid	HDL, LDL, total cholesterol, triglyceride, lipoprotein(a)
Liver function	alanine transaminase, aspartate transaminase
Kidney function	creatine, cystatin-C, urea

Rare likely deleterious missense variant burden for hbA1c



Manhattan plots of the -$\log_{10}P$ of the association of gene-level likely deleterious missense variant burden (a) in IDRs and (b) out of IDRs with hbA1c. Each dot represents a gene, and its genomic location is plotted on the x-axis. The horizontal red dashed line is the Bonferroni significant threshold, $P < 2.5e\text{-}6$. Red dots represent significant genes.

Amyotrophic lateral sclerosis mutant SOD1 aggregates into stress granules and induces alternative slicing. SOD activity is significantly reduced in type 2 diabetes and has negative correlation with hbA1c, which is regulated by *SOD1* and *SOD2* polymorphism.[7]

Conclusions

1. By whole-exome sequencing analyses in the UK Biobank, we identified *ZNF366*, *CEP290* and *TGIF2* as risk genes for inflammatory bowel disease, heart failure and non-alcoholic steatohepatitis, respectively. The burden of risk variants in IDRs may affect disease risk through disrupting the normal function of condensates.

2. We identified 90 IDR enriched gene-trait associations across 17 quantitative traits.

3. We propose that condensates serve as novel therapeutic targets for drug discovery and condensate modifying drugs (c-mods) can reverse the course of these complex diseases.

References

1. Backman JD, et al, Nature (2021)
2. Tian R, et al, medRxiv (2021)
3. Emenecker RJ, et al, Biophysical Journal (2021)
4. Mbatchou J, et al, Nature Genetics (2021)
5. Schiff ER, et al, Human Genetics, (2018)
6. Mitrea DM, et al, Cell Communication and Signalling (2016)
7. Flekac M, et al, BMC Medical Genetics (2008)