

## Rare Disease Data Center (RDDC)



Al Meets Bioinformatics: Revolutionizing Rare Disease Research



Your One-Stop Solution for Gene, Disease, and Model Data



Guangzhou Rare Disease Gene Therapy Alliance





## What is the Rare Disease Data Center?



The Rare Disease Data Center (RDDC) is a cutting-edge database co-developed by the AI Innovation Center of Tsinghua Pearl River Delta Research Institute and Cyagen Biosciences. It integrates open-source data from domestic and international sources, including epidemiology, drug development, disease-related gene maps, mutation sites, and rodent models. By combining genetic data resources with AI and bioinformatics technologies, RDDC deploys tools such as the Pathogenicity Predictor and RNA Splicer, significantly advancing rare disease research.

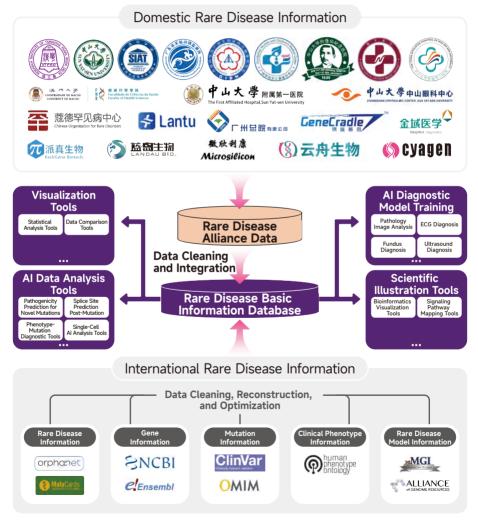


Figure 1. Data Sources and Applications

### **Core Applications**



### Comprehensive Information Retrieval

#### Genes

RDDC contains information on 64,099 genes across humans, mice, and rats.



#### Diseases

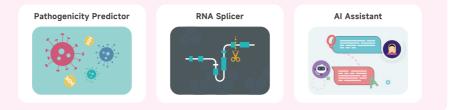
RDDC integrates data from Malacards, OMIM, Orphanet, ClinVar, and the Rare Disease Alliance, covering 21,933 diseases.



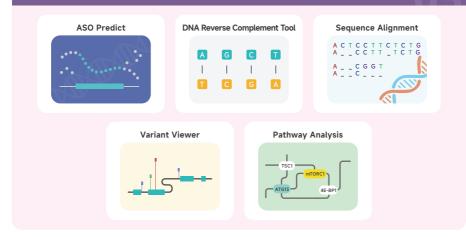
#### Mouse Models

RDDC includes 107,541 gene-edited mouse models from various literature sources.

### Al Prediction Tools: From genomics to proteomics, Al empowers life science research, helping you explore uncharted territories.



### Bioinformatics Analysis Tools: Fast, accurate, and intelligent tools to decode data and accelerate life science research.



### **Core Function Use Cases**



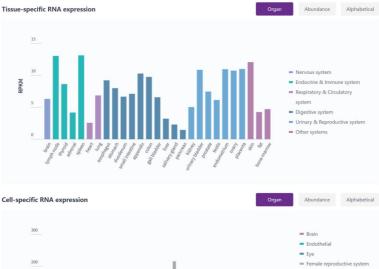
#### Diseases & Mutations

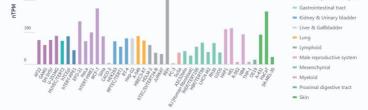
Genes are the building blocks of heredity, carrying instructions that define traits and functions. Mutations in these genes can disrupt protein function, leading to disease. Understanding how genetic mutations relate to diseases is essential for developing targeted therapies that can improve patient outcomes or even cure diseases.

#	Disease	Anatomical Category	Score \$	Mutations ‡
1	Li-Fraumeni Syndrome (LFS)	25	1764.42	2052(28996)
2	Papilloma of Choroid Plexus (CPP)		1464.26	30(428)
3	Osteogenic Sarcoma (OSRC)		1376.1	30(434)
4	Glioma Susceptibility 1 (GLM1)	<b>*</b> 3	1109.43	33(467)
5	Basal Cell Carcinoma 7 (BCC7)	65 © A	1107.87	30(428)
6	Adrenocortical Carcinoma, Hereditary (ADCC)		1100.02	30(428)
7	Bone Marrow Failure Syndrome 5 (BMFS5)		1057.85	26(354)

### Gene Expression

Gene expression measures how active genes are in specific tissues or cells. These patterns are crucial for understanding biological functions and the mechanisms behind diseases. Studying gene expression helps uncover links between genes and diseases, offering new avenues for prevention and treatment.





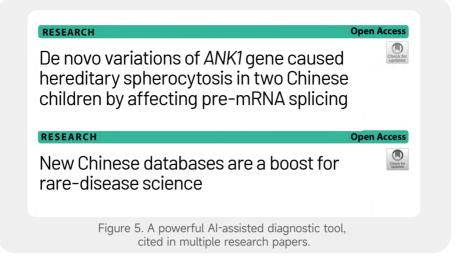
#### Figure 2. Rich structured data and visualizations.

Genomi	c mutations Protein mutations			
NM_0	00546.6 (NP_000537.3) (MANE) ~		Q Search ? TP53 S	Sequence
Y All	🖻 P. 704 🔄 LP. 189 💟 LB. 631 🔄 B. 8 💽 VUS: 673			Q
1	М Е <sup>3</sup> Е <sup>4</sup> Р <sup>7</sup> Q <sup>2</sup> S <sup>2</sup> D <sup>4</sup> Р <sup>3</sup> S <sup>6</sup> V <sup>6</sup> E <sup>3</sup> Р <sup>4</sup> Р <sup>5</sup> L <sup>6</sup> S <sup>2</sup> Q <sup>4</sup> ато сало сасастстваятся так стак стока сссстстваятся с	E T <sup>3</sup> F S <sup>5</sup> D L W K <sup>2</sup> GAAACATTTTCAGACCTATGGAAA	L <sup>3</sup> L <sup>3</sup> P <sup>8</sup> E N <sup>5</sup> N <sup>5</sup> CTACTTCCTGAAAACAAC	Exon 3
31	V <sup>5</sup> L <sup>5</sup> S <sup>3</sup> P <sup>8</sup> L <sup>6</sup> P <sup>6</sup> S <sup>5</sup> Q <sup>3</sup> A <sup>5</sup> M <sup>3</sup> D <sup>4</sup> D <sup>6</sup> L <sup>6</sup> M <sup>5</sup> L <sup>3</sup> S <sup>4</sup> GTTCTGTCCCCCTTGCCGTCCCAAGCAATGATGATGATGATGATGCTGTCC	P <sup>7</sup> D <sup>3</sup> D <sup>6</sup> I <sup>3</sup> E <sup>3</sup> Q <sup>6</sup> W' F <sup>2</sup> CCGGGACGATATTGAACAATGGTTC		Exon 4
61	D <sup>5</sup> E <sup>4</sup> A <sup>4</sup> P R <sup>5</sup> M <sup>5</sup> P <sup>4</sup> E <sup>3</sup> A <sup>6</sup> A <sup>3</sup> P <sup>4</sup> P <sup>+</sup> V <sup>6</sup> A <sup>5</sup> P <sup>3</sup> A <sup>5</sup> 6 A T 6 A A 6 C T C C C A 6 A A T 6 C A 6 A 6 C T 6 C T C C C C C C 6 T 6 C C C C T 6 C A		P <sup>4</sup> A <sup>6</sup> P <sup>5</sup> A <sup>7</sup> P <sup>4</sup> S <sup>7</sup> CCTGCACCAGCCCCCTCC	Exon 4
91	W <sup>5</sup> P <sup>3</sup> L <sup>6</sup> S <sup>3</sup> S <sup>4</sup> V <sup>7</sup> P <sup>7</sup> S <sup>6</sup> Q <sup>3</sup> K <sup>4</sup> T <sup>6</sup> Y <sup>4</sup> Q <sup>3</sup> G <sup>4</sup> S <sup>7</sup> S <sup>6</sup> S <sup>2</sup> S <sup>6</sup> S <sup>7</sup> S <sup>7</sup> S <sup>6</sup> S <sup>7</sup> S <sup>6</sup> S <sup>7</sup> S <sup>6</sup> S <sup>7</sup> <td< td=""><td>Y" G<sup>7</sup> F" R" L" G<sup>4</sup> F<sup>7</sup> L<sup>3</sup></td><td>H<sup>6</sup> S<sup>3</sup> G<sup>6</sup> T<sup>3</sup> A<sup>4</sup> K<sup>6</sup> CATTCT6666ACA66CCAA6</td><td>Exon 4</td></td<>	Y" G <sup>7</sup> F" R" L" G <sup>4</sup> F <sup>7</sup> L <sup>3</sup>	H <sup>6</sup> S <sup>3</sup> G <sup>6</sup> T <sup>3</sup> A <sup>4</sup> K <sup>6</sup> CATTCT6666ACA66CCAA6	Exon 4
121	S <sup>8</sup> V <sup>0</sup> T <sup>5</sup> C <sup>6</sup> T <sup>9</sup> Y <sup>5</sup> S <sup>7</sup> P <sup>6</sup> A <sup>6</sup> L <sup>6</sup> N <sup>9</sup> K <sup>*</sup> M <sup>6</sup> F <sup>*</sup> C <sup>9</sup> Q <sup>6</sup> тетотоваеттосае отастессе тоссе тоссе то асаа о атоттто се са а			Exon 5
151	$P^{0}  P^{+}  P^{7}  G^{9}  T^{9}  R^{7}  V^{9}  R^{+}  A^{7}  M^{1}  A^{6}  I^{-7}  Y^{7}  K^{6}  Q^{4}  S^{5}  CCCCCGCCCCGCCCCCGCCCCCCCCCCCCCCCCCCC$	Q <sup>7</sup> H <sup>4</sup> M <sup>3</sup> T <sup>0</sup> E <sup>6</sup> V <sup>7</sup> V <sup>+</sup> R <sup>4</sup> CAGCACATGACGGAGGTTGTGAGG		Exon 5
181	R <sup>6</sup> C <sup>4</sup> S <sup>3</sup> D <sup>3</sup> S <sup>3</sup> D <sup>3</sup> G <sup>3</sup> L <sup>3</sup> A <sup>7</sup> P <sup>*</sup> P <sup>5</sup> Q <sup>*</sup> H <sup>8</sup> L <sup>7</sup> I <sup>7</sup> R <sup>5</sup> coctoctcadata ocaatostcto accectecte accatetta tecaa		Y <sup>9</sup> L <sup>5</sup> D D <sup>6</sup> R <sup>5</sup> N <sup>5</sup> TATTTGGATGACAGAAAC	Exon 6
211	Т <sup>6</sup>		N <sup>0</sup> Y <sup>0</sup> M <sup>+</sup> C <sup>0</sup> N <sup>+</sup> S <sup>+</sup> A A C T A C A T G T G T A A C A G T	Exon 7
241	$ \begin{array}{cccccccccccccccccccccccccccccccccccc$		L <sup>4</sup> G <sup>5</sup> R <sup>8</sup> N <sup>5</sup> S <sup>5</sup> F <sup>*</sup>	Exon 8
271	E <sup>8</sup> V <sup>9</sup> R <sup>*</sup> V <sup>9</sup> C <sup>9</sup> A <sup>2</sup> C <sup>8</sup> P <sup>8</sup> G <sup>6</sup> R <sup>*</sup> D <sup>*</sup> R <sup>*</sup> R <sup>8</sup> T <sup>9</sup> E <sup>6</sup> E <sup>2</sup> 6 A 6 6 T 6 C 6 T 6 T T 7 6 T 6 C C T 6 T C C T 6 6 A 6 A A 6 A C C 6 6 C 6 C A C A 6 A 6		P <sup>8</sup> H <sup>4</sup> H <sup>7</sup> E <sup>8</sup> L <sup>5</sup> P <sup>3</sup> CCTCACCACGAGCTGCCC	Exon 8
301	P <sup>3</sup> G <sup>6</sup> S <sup>3</sup> T <sup>4</sup> K <sup>4</sup> R <sup>3</sup> A <sup>3</sup> L <sup>4</sup> P <sup>4</sup> N <sup>7</sup> N <sup>6</sup> T <sup>6</sup> S <sup>2</sup> S <sup>4</sup> S <sup>3</sup> P <sup>6</sup> CCA066846CACTA86C68C6CT0CCC48C8AC8C6C6CCCTCCCCC	Q* P6 K K K <sup>2</sup> P <sup>7</sup> L <sup>5</sup> D <sup>5</sup>	G <sup>5</sup> E <sup>4</sup> Y <sup>4</sup> F <sup>2</sup> T <sup>5</sup> L <sup>5</sup> 0 0 A 0 A A T A T T T C A C C C T T	Exon 9
331	Q <sup>6</sup> 1 <sup>6</sup> R <sup>4</sup> G <sup>6</sup> R <sup>4</sup> E <sup>3</sup> R <sup>8</sup> F <sup>4</sup> E <sup>7</sup> M <sup>5</sup> F <sup>5</sup> R <sup>9</sup> E <sup>3</sup> L <sup>6</sup> N E <sup>5</sup> CAGATCCGTGGGCGTGAGCGCTTCGAGATGTTCCGAGACCTGAATGAG		A <sup>5</sup> G <sup>3</sup> K E <sup>2</sup> P <sup>2</sup> G <sup>9</sup> 6 C T 6 6 6 A A 6 6 A 6 C C A 6 6 6	Exon 10
	G <sup>5</sup> S <sup>7</sup> R <sup>4</sup> A <sup>3</sup> H <sup>3</sup> S <sup>5</sup> S <sup>3</sup> H <sup>6</sup> L K <sup>3</sup> S <sup>2</sup> K <sup>2</sup> K <sup>6</sup> G <sup>5</sup> O <sup>5</sup> S	T <sup>5</sup> S <sup>2</sup> R <sup>6</sup> H <sup>3</sup> K <sup>2</sup> K <sup>4</sup> L <sup>4</sup> M <sup>6</sup>	a u st a a al	Exon 11

Figure 3. Intuitive sequence and mutat	tion search functionality.
--	----------------------------

lighlight the dentical items	<b>↑</b> TP53 ⊗	⇔Trp53 ⊗	†DMD ⊗	†SMN1 ⊗	÷SMN2 ⊗
Genetic					
Chromosome 🕑	GRCh38-Chr17	GRCm39-Chr11	GRCh38-ChrX	GRCh38-Chr5	GRCh38-Chr5
Location 🕖	17 p13.1	11 q83	X p21.2	5 q13.2	5 q13.2
Position	7,668,421-7,687,490	69,471,174-69,482,699	31,119,222-33,339,388	70,924,941-70,966,375	70,049,523-70,090,528
Direction		+		+	+
Gene Length	19070 bp	11526 bp	2220167 bp	41435 bp	41006 bp
MW (kDa) 🔞	43.65 kDa	43.46 kDa	426.75 kDa	31.85 kDa	31.85 kDa
Orthologs	Trp53   Tp53	🗄 ТР53   🔝 Тр53	Dmd   Dmd	Smn1 Smn1	Smn1 Smn1
Related Mouse Models	Trpszimited Trpszimited Trpszimited Trpszimite Trpszimite Trpszimite Trpszimite Trpszimite Trpszimite Trpszimite Trpszimite Trpszimite Trpszimite	TypSzimily TypSzimilia TypSzimilia TypSzimilia TypSzimilia TypSzimilia TypSzimilia TypSzimilia TypSzimilia TypSzimilia TypSzimilia TypSzimilia	Dragettik Dragettik Kv Dragettik Kv View all	Smithatid Smithatia Smithatia Smithatia Smithatia Smithatia Smithatia Smithatia Smithatia Smithatia Smithatia Smithatia Smithatia	Smn1 <sup>th</sup> Likke Smn1 <sup>th</sup> Likke
Transcript					
Transcript ID	NM_000546.6 MANE	NM_011640.3 MANE	NM_004006.3 MANE	NM_000344.4 MANE	NM_017411.4 MANE
Length (nt)	2512 (nt)	1772 (nt)	13992 (nt)	1482 (nt)	1482 (nt)
Exon Count	11	11	79	9	9
CDS (bp)	1182 bp	1173 bp	11058 bp	885 bp	885 bp
Length (aa)	393 (aa)	390 (aa)	3685 (aa)	294 (aa)	294 (aa)
Gene Expression (	0				
Tissue	spleen (13.139) lymph node (13.04)	thymus adult (69.138) ovary adult (56.493)	heart (7.034) fat (4.274)	bone marrow (23.326) testis (21.559)	bone marrow (22.74) testis (21.494)

Figure 4. Detailed data comparison tools.



### AI-Assisted Prediction Case Study: Duchenne Muscular Dystrophy (DMD)

### Clinical Features

Duchenne Muscular Dystrophy (DMD) is an X-linked recessive disorder caused by mutations in the DMD gene, the largest gene in the human genome (2.5 Mb). Approximately 60-65% of cases result from large deletions of one or more exons, while 20% are caused by single nucleotide variants, including frameshift, nonsense, missense, and insertion-deletion mutations. In-frame mutations typically lead to Becker Muscular Dystrophy (BMD), a less severe form, whereas frameshift mutations result in severe DMD phenotypes. Whole-genome sequencing of patient samples revealed a hemizygous mutation in the DMD gene: c.4675-2A>G, causing aberrant splicing.

Category	Details	Category	Details	
Gene	DMD	Normal Population Frequency		
Chromosomal Location	chrX: 32398799	, ,		
		ACMG Pathogenicity Analysis	Pathogenic	
Transcript/Exon	NM_004006; Exon34		Duchenne Muscular Dystrophy (XLR)	
Nucleotide/ Amino Acid	c.4675-2A>G	Disease/Phenotype	Becker Muscular Dystrophy (XLR) Dilated Cardiomyopathy 3B (XL)	
Homozygous/ Heterozygous	Heterozygous	Variant Origin	Spontaneous	

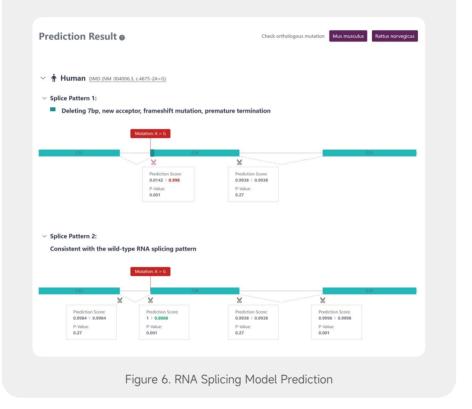
Table 1. Whole Exome Sequencing Results for DMD Patients

### Al Prediction Results

Using RDDC's AI tools—Pathogenicity Predictor and RNA Splicer—the pathogenicity and RNA splicing effects of the mutation were predicted.

### The mutation may cause two splicing abnormalities:

- Exon 34 skipping.
- A 7-base deletion (non-triplet), leading to a frameshift and premature termination.



### Validation Results

Literature reports confirm the same mutation in DMD carriers, generating a novel splice acceptor site (Hofstra et al., 2004), consistent with RDDC's predictions. To date, over 4,000 pathogenic DMD mutations have been reported. RNA Splicer's predictions for dozens of DMD splicing mutations show high concordance with published results.

# Empowering Rare Disease Research with Al-Driven Insights

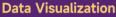
### **Al-Driven Insights**

Leveraging vast datasets on diseases, genes, and mutations, RDDC employs machine learning and deep learning algorithms to develop tools for rare disease research and genetic disorder diagnosis.



### **Core Data Focus**

Centered on disease, gene, and animal model data, RDDC provides essential support for preclinical research and drug development in rare diseases.



۲

### With advanced data visualization and comprehensive comparison tools, RDDC enhances researchers' efficiency in information retrieval, addressing the shortcomings of traditional biological databases that prioritize data over user experience.

### **Targeting Mutations**

Focusing on the relationship between mutations and phenotypes, RDDC utilizes AI tools to explore the pathogenic mechanisms of genetic rare diseases, helping users pinpoint target genes and mutations with precision.

### **Efficient AI Assistance**

Integrating proprietary database APIs with large language models, RDDC allows users to query information conversationally, significantly reducing the learning curve and improving accessibility.

### Guangzhou Rare Disease Gene Therapy Alliance

- C Phone: 18028568424
- Email: rddc-support@tsinghua-gd.org
- Swebsite: https://rddc.tsinghua-gd.org/

