腺相关病毒衣壳蛋白结合亲和力筛选和结构预测

AAV Capsid Binding Affinity Screening and Structure Prediction

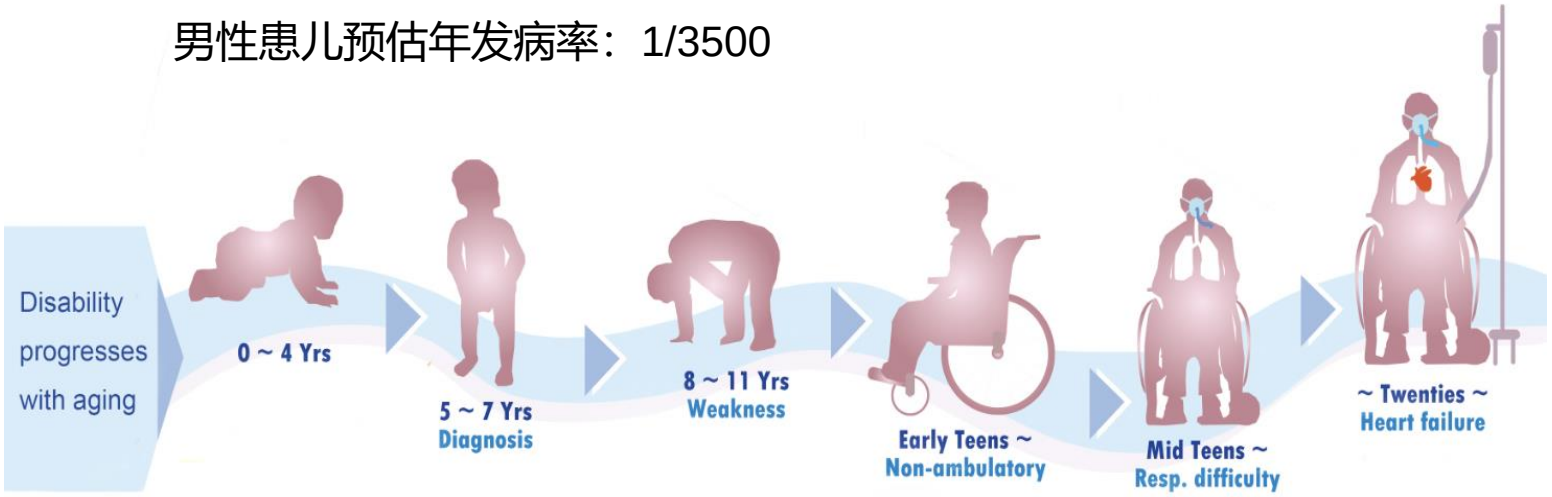
汇报人：刘路旋

组员：闫泽淇、张博洋、罗小山

杜氏肌营养不良症（Duchenne Muscular Dystrophy）

X连锁隐性遗传疾病，以进行性肌肉退化和无力为主要特征

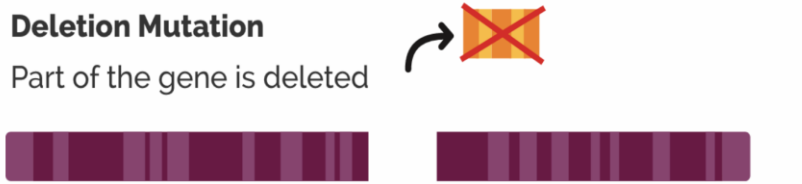
男性患儿预估年发病率：1/3500



60–70%

Deletion Mutation

Part of the gene is deleted



5–15%

Duplication Mutation

Part of the gene is repeated

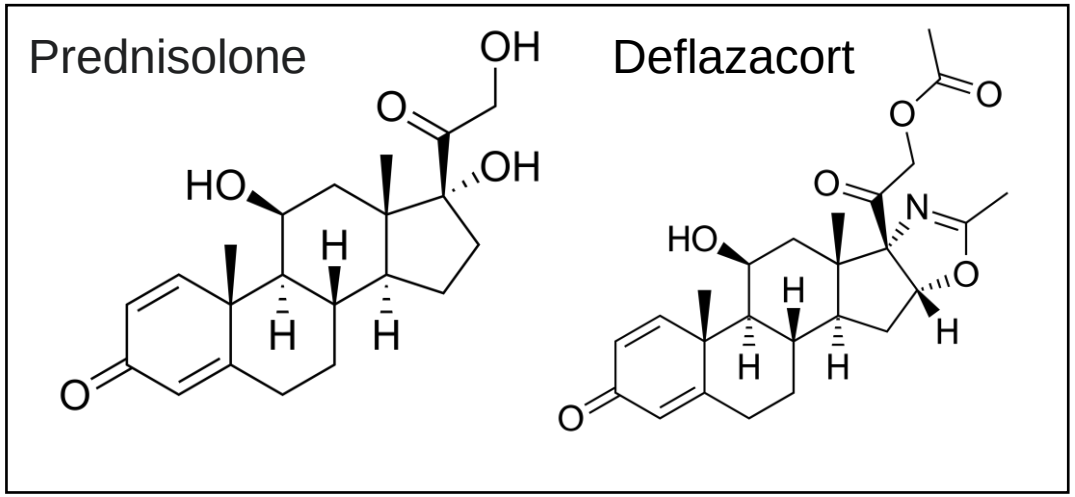
Other Changes

20% are point mutations,
small deletions or insertions

DMD的临床治疗策略

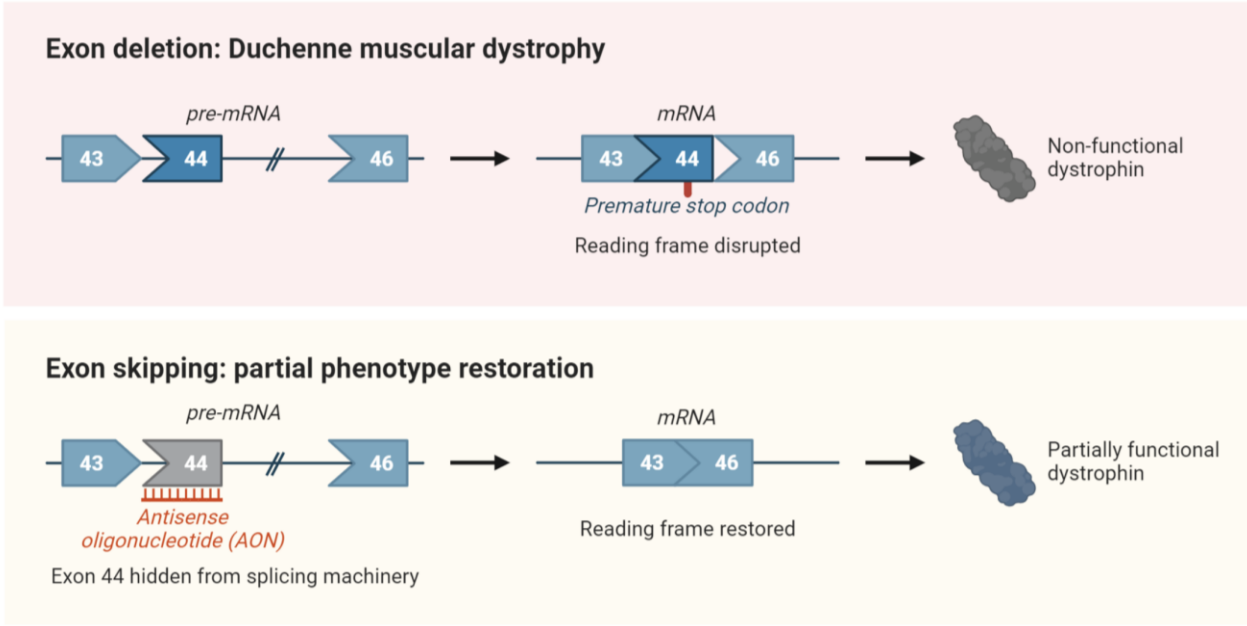
北京大学 PEKING UNIVERSITY

糖皮质激素



X Can preserve muscle strength
but their benefit is symptomatic and temporary

外显子跳跃疗法

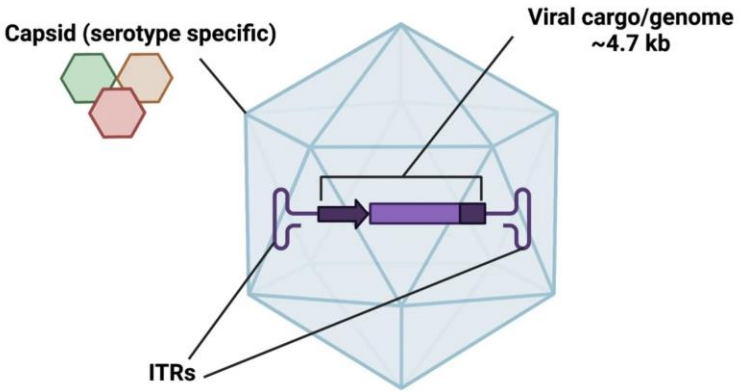


X Restore the reading frame to produce a shortened but partially functional dystrophin protein
But benefit only 10%-15% of DMD patients
efficiency of dystrophin restoration is often low

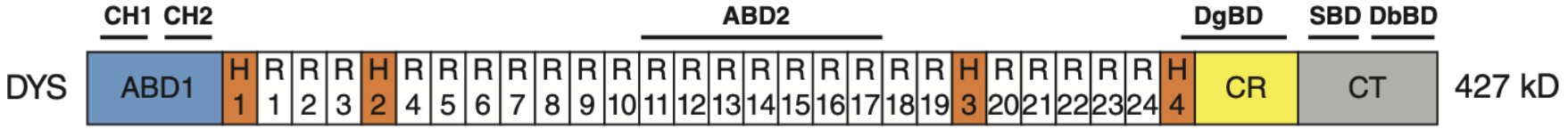
杜氏肌营养不良症的微型抗肌萎缩蛋白基因疗法

北京大学 PEKING UNIVERSITY

安全且天然的肌肉趋向性



可覆盖广泛的患者人群，
且不受突变类型限制



Name	Microdystrophin transgene structure	AAV serotype	Promoter	Notes
Elevidys Delandistrogene moxeparvovec SRP-9001 (Sarepta Tx)	N ABD H1 1 2 3 H2 24 H4 CR C	AAVrh74	MHCK7	Approved by US FDA, 2024
GNT0004 (Genethon/Sarepta Tx)	N ABD H1 1 2 3 H2 24 H4 CR C	AAV8	Spc5-12	
PF-06939926 Fordadistrogene movaparvovec (Pfizer)	N ABD H1 1 2 H3 22 23 24 H4 CR C	AAV9	MSP	Discontinued
SGT-001 (Solid Biosciences)	N ABD H1 1 16 17 23 24 H4 CR C	AAV9	CK8	Deprioritzed: focusing on SGT-003
RGX-202 (REGENXBIO)	N ABD H1 1 2 3 H2 24 H4 CR CT C	AAV8	Spc5-12	

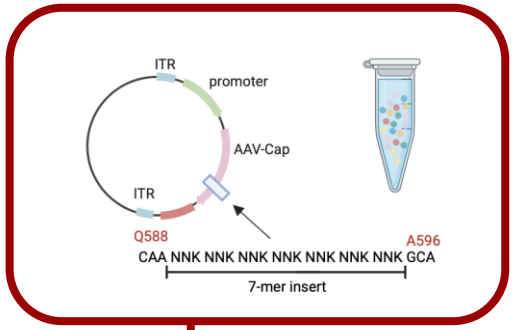
高病毒剂量 ⟶ 副作用

Liver injury, hypertransaminasaemia, rhabdomyolysis, immune-mediated myositis, myocarditis, increased transaminases, left ventricular dysfunction, gamma-glutamyl transferase and transaminase increased…

Quan Q, Gao et al. 2015
Katarzyna Chwalenia et al. 2025

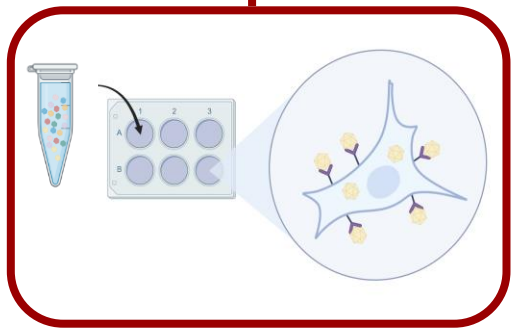
工作流程

构建AAV9 7-mer library



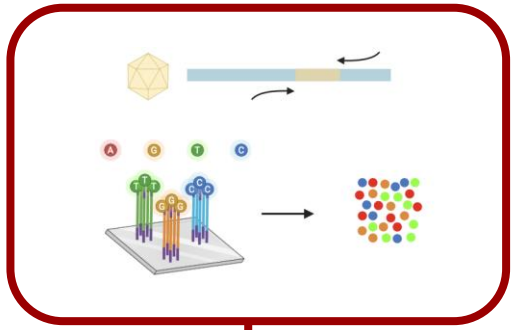
Step1

Step2



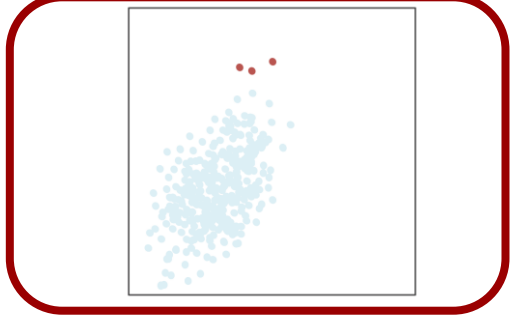
受体结合筛选

高通量测序



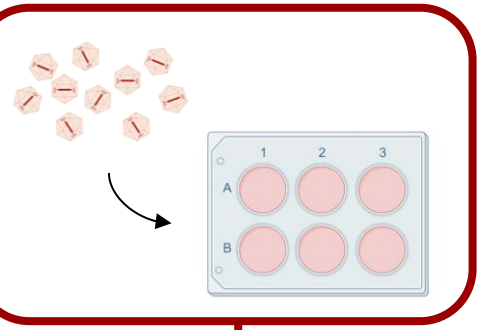
Step3

Step4



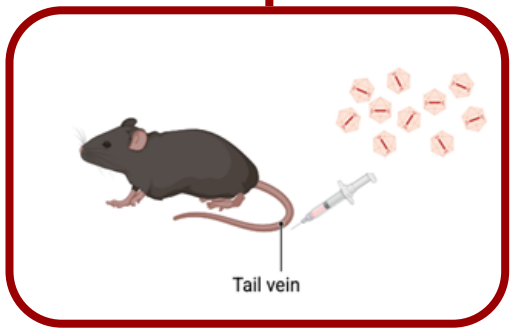
阳性克隆鉴定

体外验证



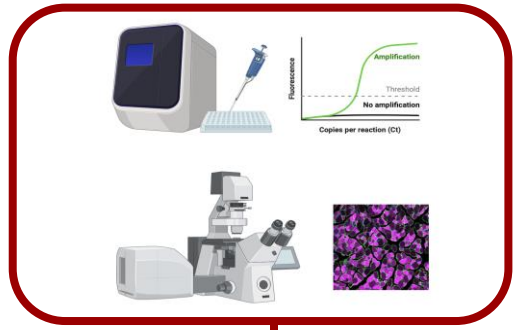
Step5

Step6



病毒注射

体内验证

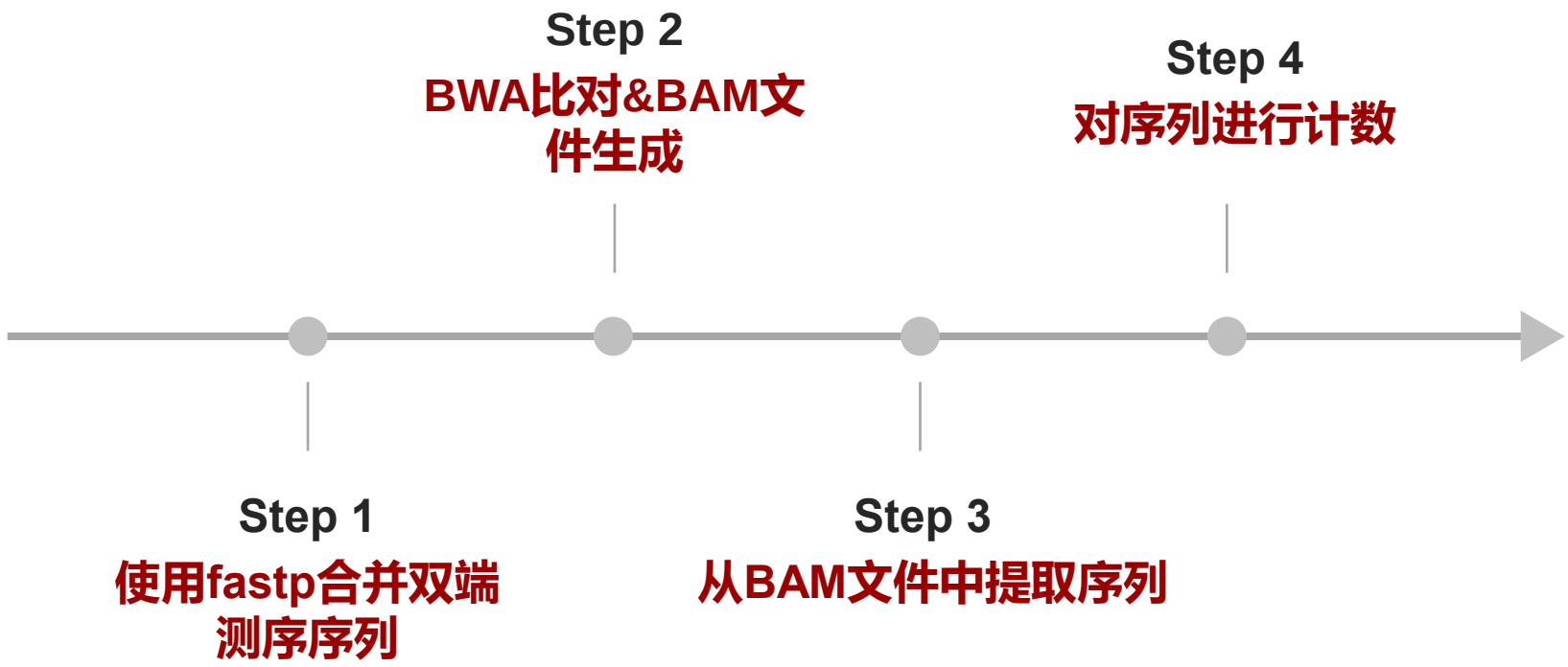


Step7

1 测序数据预处理流程

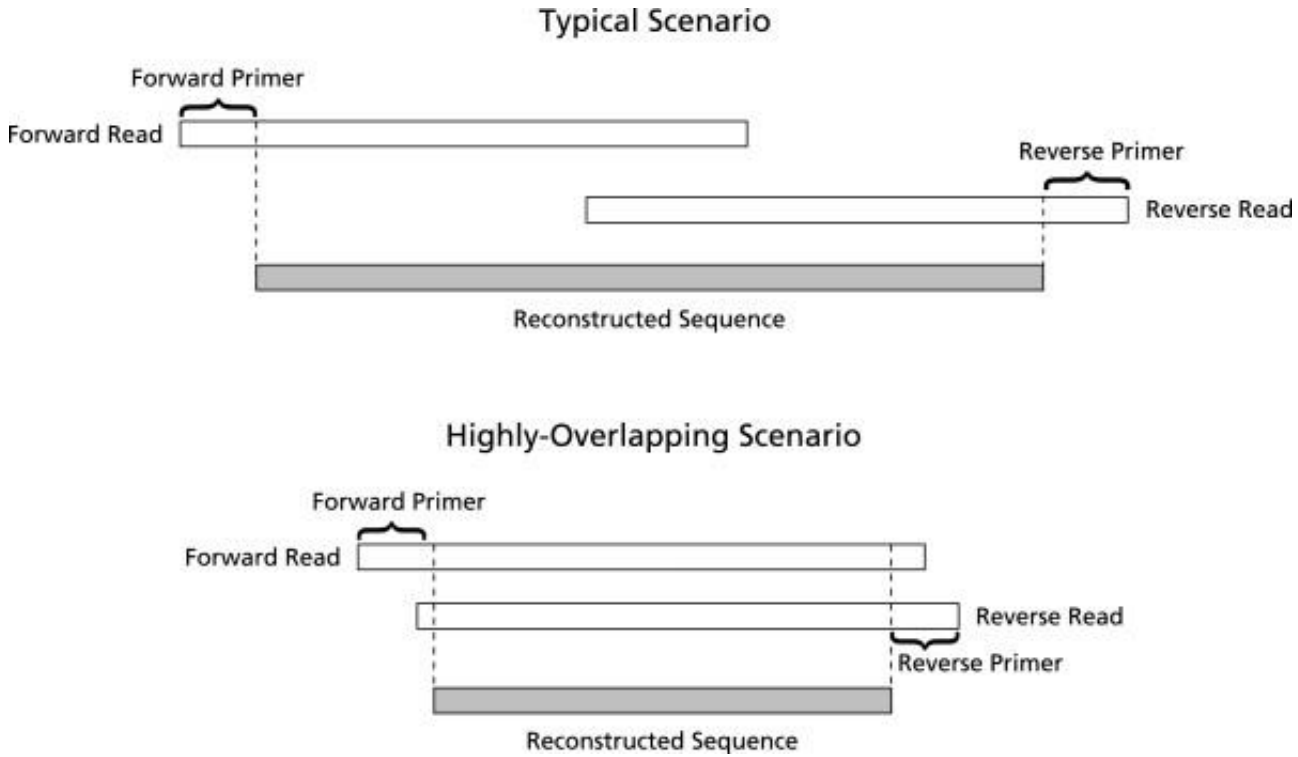
北京大学 PEKING UNIVERSITY

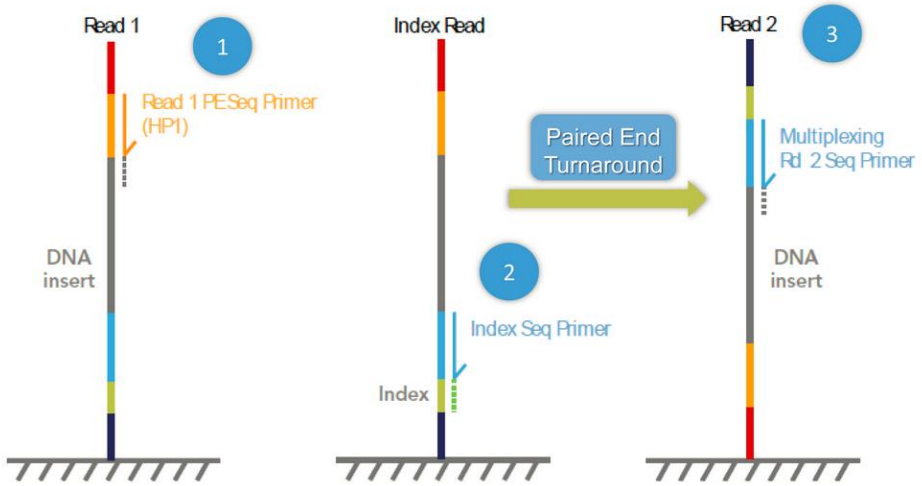
Workflow Overview



1 测序数据预处理流程

Step 1: 使用fastp合并双端测序序列






```
fastp -i "$r1" -I "$r2" -m \
    --merged_out "$sid.merge.fastq.gz" \
    --unpaired1 "$sid.unpaired1.fastq.gz" \
    --unpaired2 "$sid.unpaired2.fastq.gz" \
    --failed_out "$sid.failed.fastq.gz" \
    -w 4 \
    -h "$sid.merge.html" \
    -j "$sid.merge.json"
```

Step 2: BWA比对 & BAM文件生成

BWA (Burrows-Wheeler Aligner) 是一种高效的生物信息学比对工具，用于将DNA序列（短读或长读）比对到大型参考基因组（如人类基因组）。它基于 Burrows-Wheeler变换（BWT） 和 FM 索引技术，实现了高压缩率和快速检索。

BWA 包含三种主要算法：

- **BWA-backtrack**：适合 ≤100bp 的短序列（早期Illumina数据）。
- **BWA-SW**：支持70bp到几百万bp的长序列。
- **BWA-MEM**：推荐的通用算法，适合 ≥70bp 的高质量序列，速度快且准确度高。

这里使用的 **BWA-MEM** 是一种新的比对算法，用于将序列读段或长查询序列与大型参考基因组（如人类）比对。它自动选择 local 和 end-to-end 比对，支持成对端读段 (paired-end reads) 并执行嵌合比对。该算法对测序误差具有鲁棒性，适用于从 70bp 到几百万碱基的广泛序列长度。对于 100 bp 序列的定位，BWA-MEM 表现优于迄今为止多种最先进的读段比对器。

该算法先使用最大精确匹配 (maximal exact matches, MEM) 进行 seeding alignments，再使用 SW (affine-gap Smith-Waterman) 算法进行 seeds 的延伸。

北京大学 PEKING UNIVERSITY

Step 2: BWA比对 & BAM文件生成

SAM (Sequence Alignment/Map) 格式是用于存储和描述高通量测序比对结果的标准格式。SAM 文件是纯文本格式，而 **BAM (Binary Alignment/Map)** 文件是其二进制形式，二进制文件占用的磁盘空间比文本文件小，同时运算速度快。非常多的比对软件可以生成SAM格式，如 HISAT2、bwa、minimap2 等。samtools 专门用于处理 SAM 和 BAM 文件的工具集，包括排序、索引、格式转换等。

BAM文件由两个主要部分组成:

- 头部（Header）：可选部分，包含关于数据源、参考序列和比对方法的信息。
- 比对部分（Alignment Section）：每一行对应一个测序读段的比对信息，包含多个字段，如读段ID、比对位置、比对质量值等。

Align to reference (BWA-MEM)

```
bwa mem -t 4 -R "@RG\tID:$sid\tSM:$sid" "$REF_FASTA" "$merged_file" | \
    samtools sort -@ 2 -o "$BAM_DIR/$sid.merge.sort.bam"
```

Sort & index BAM files (samtools)

```
samtools index "$bam_file"
```

北京大学 PEKING UNIVERSITY

Step 3 & 4: 从BAM文件中提取序列& 频率计数

Target Region:

3811 - 3855

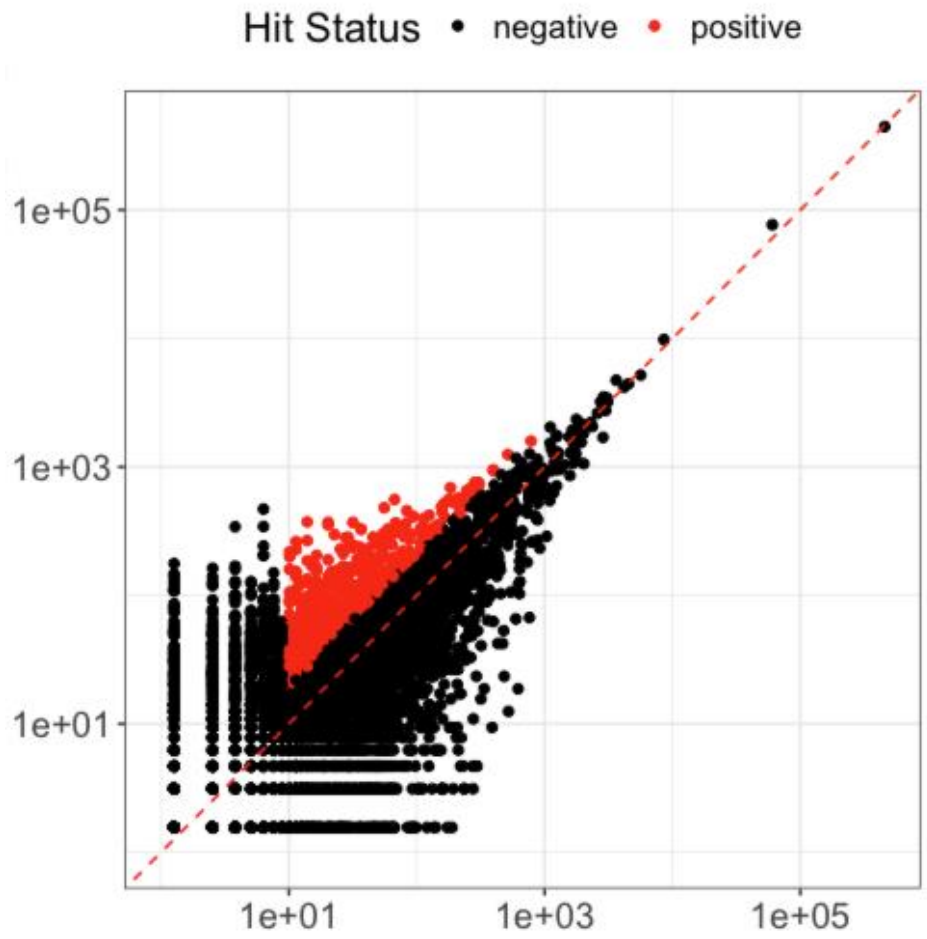
- Extract region from BAM
- Filter: MAPQ ≥ 30, full coverage
- Translate DNA → AA

```python
def extract_sequences_vectorized(read_align_position_list, read_sequence, start_pos, end_pos):
    positions = np.array(read_align_position_list)
    read_positions = positions[:, 0]
    read_positions[read_positions == None] = 0
    ref_positions = positions[:, 1]
    ref_positions[ref_positions == None] = 0
    extract_start_pos = np.where(ref_positions >= start_pos-1)[0][0]
    extract_end_pos = extract_start_pos + end_pos - start_pos + 1
    extracted_sequence = read_sequence[extract_start_pos:extract_end_pos]
    return extracted_sequence
```

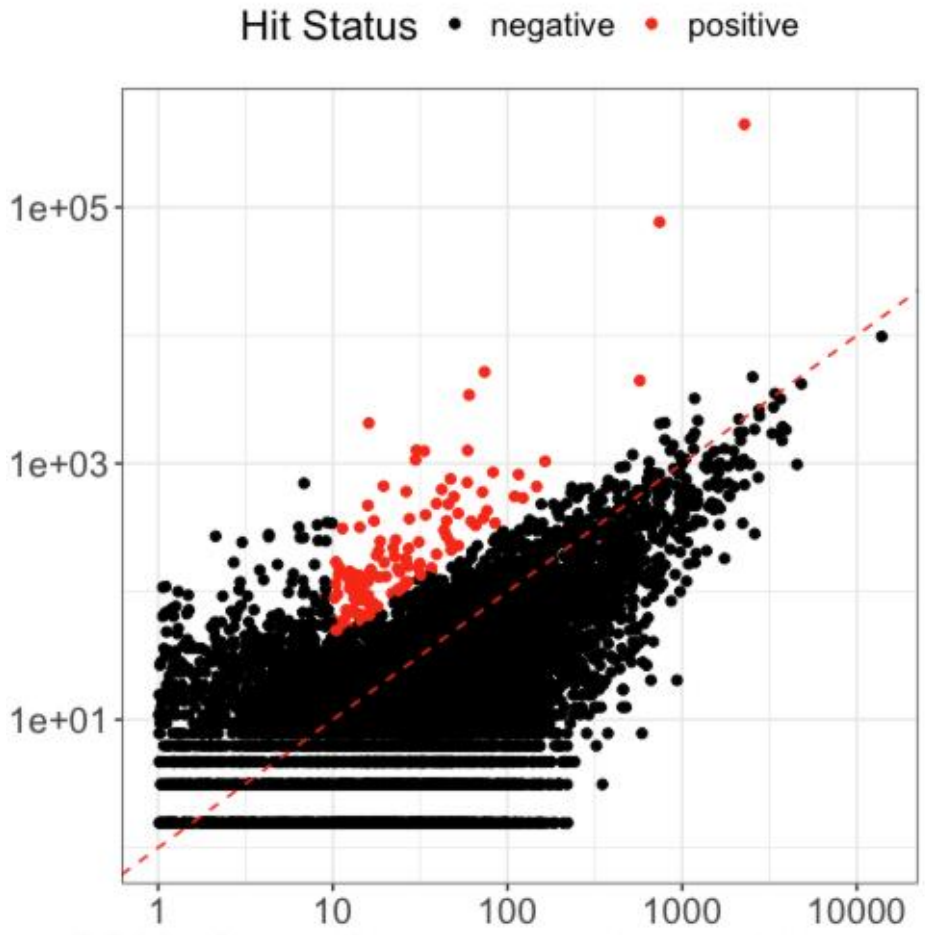
2 富集AAV Variants的系统发育分析

$$\log\mathrm{FC}_i = \log_2\left(\frac{\mathrm{RPM}_i^{\mathrm{output}}}{\mathrm{RPM}_i^{\mathrm{input}}}\right)$$

定义阳性序列



- RPM>10, logFC>2



- RPM>10, logFC>4

2 Clustal Omega多序列比对&系统发育树构建

北京大学 PEKING UNIVERSITY

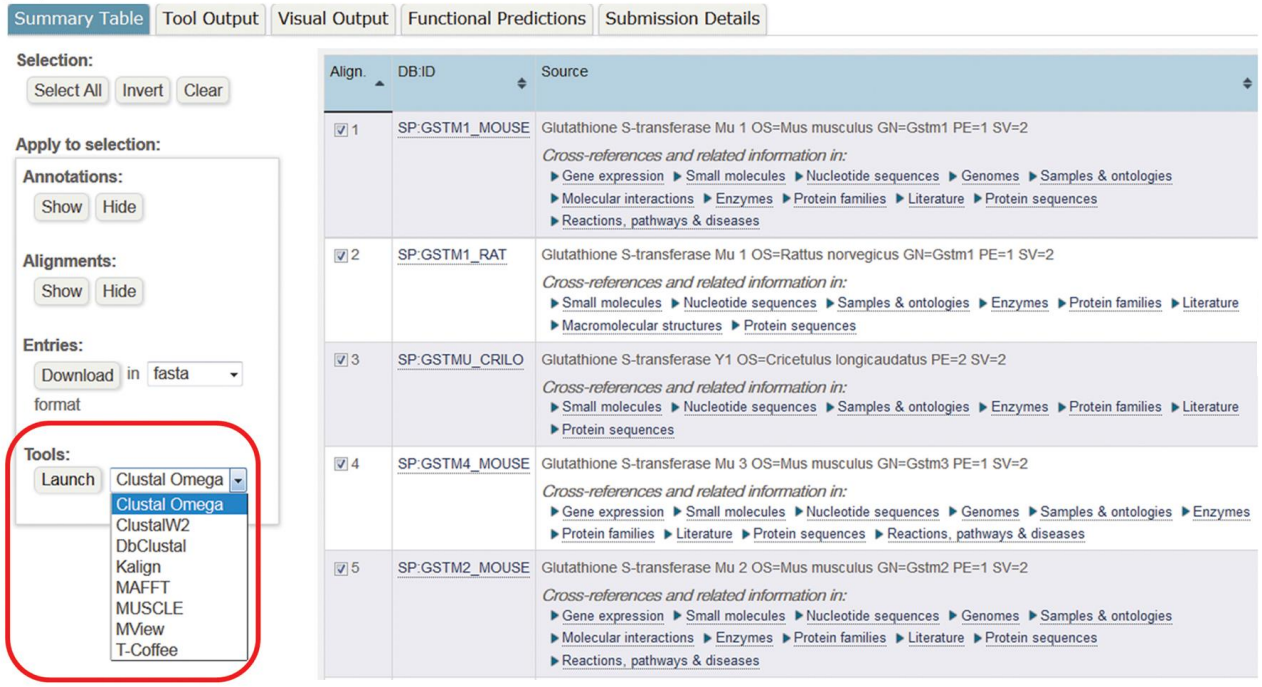
EMBL-EBI home | Services | Research | Training | About us | EMBL-EBI

Clustal Omega

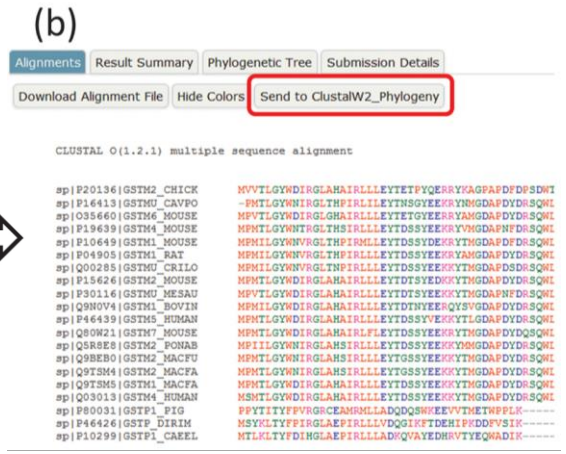
Multiple Sequence Alignment (MSA)

Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between **three or more** sequences

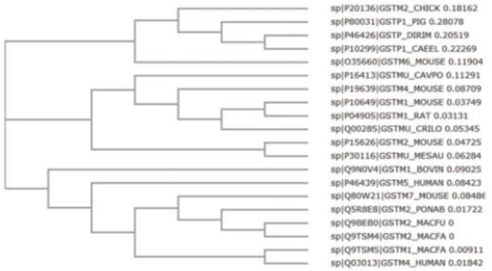
(a)



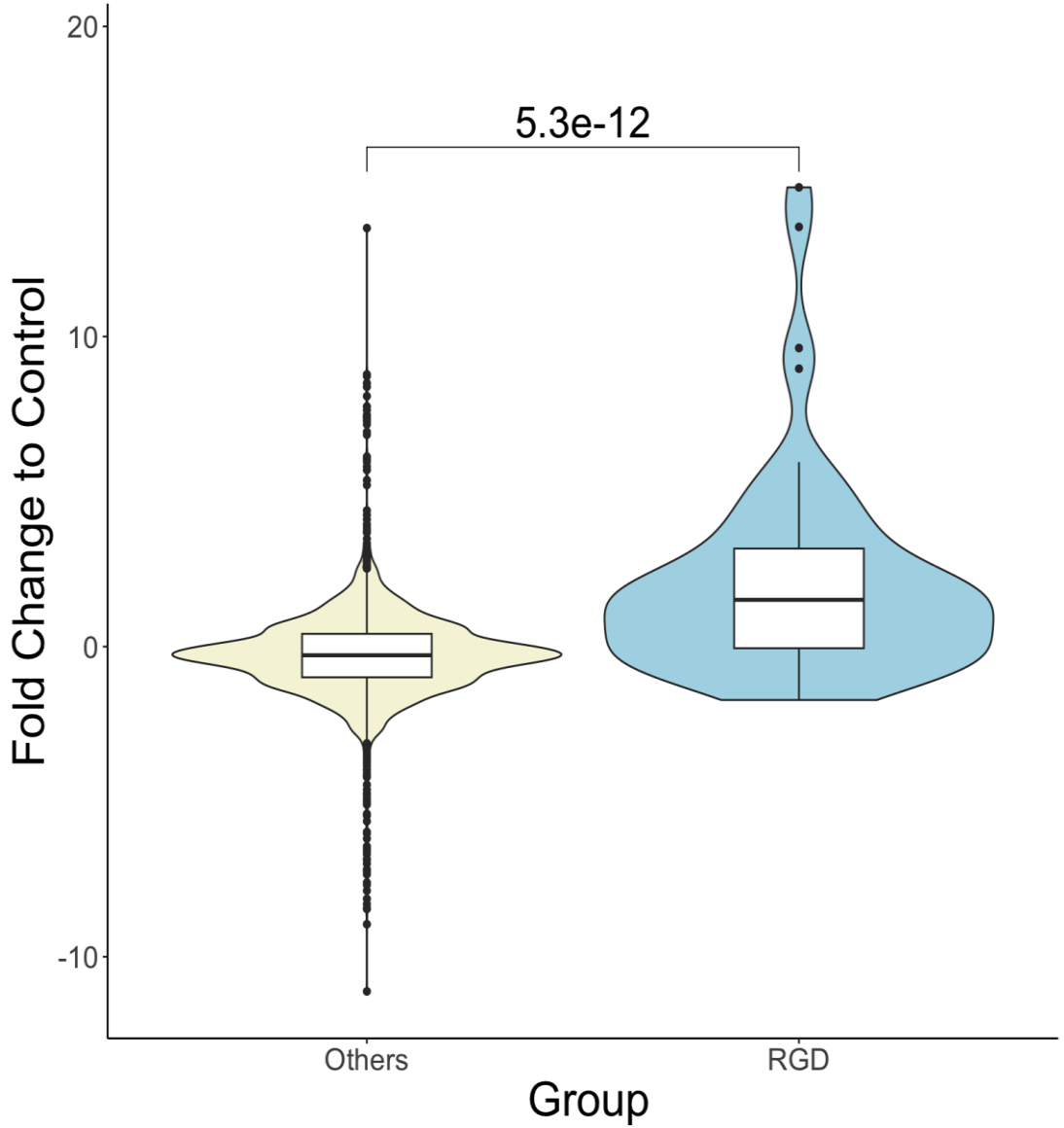
(b)



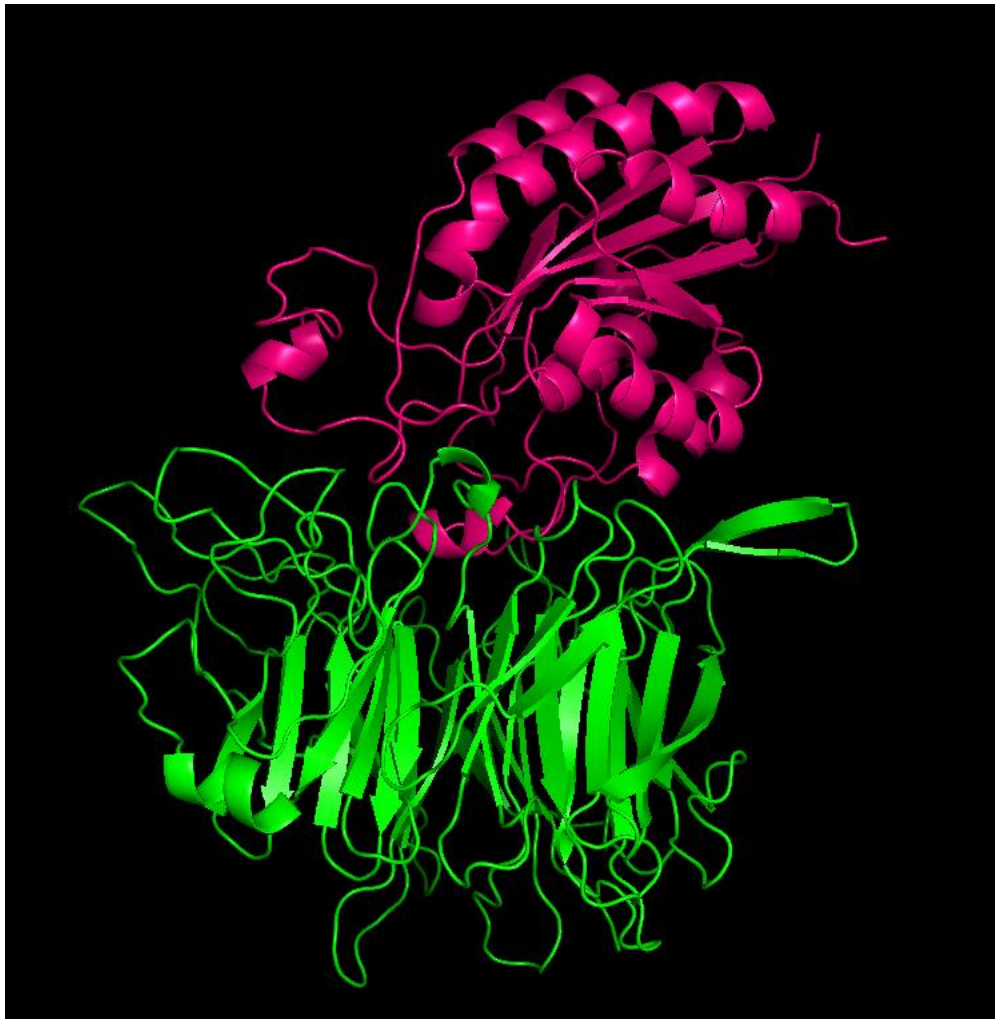
(c)

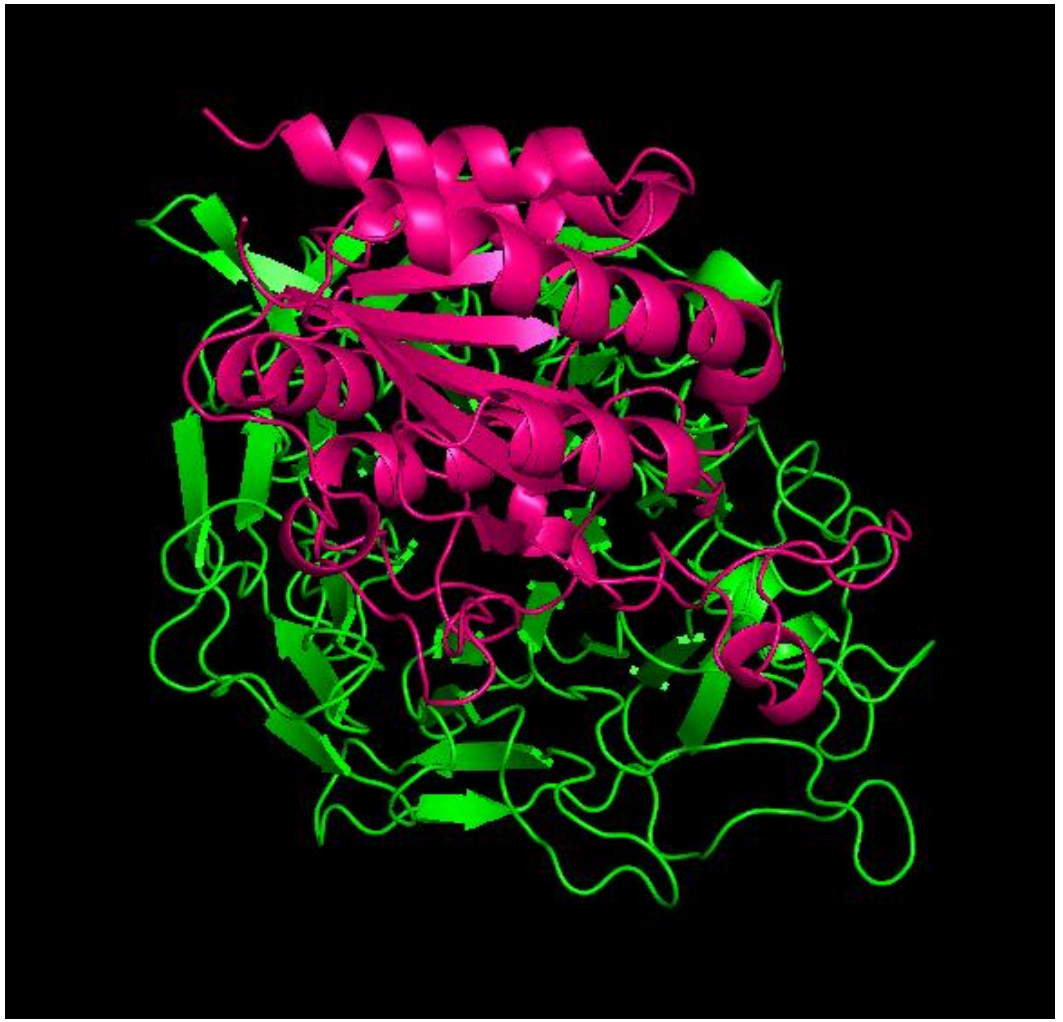


含RGD基序的序列在阳性序列中显著富集

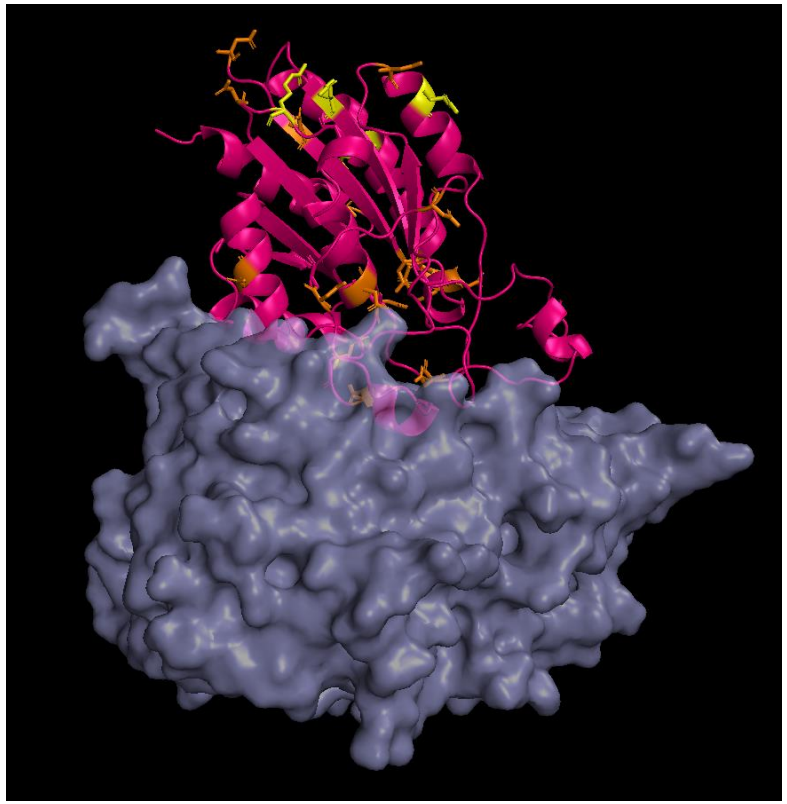


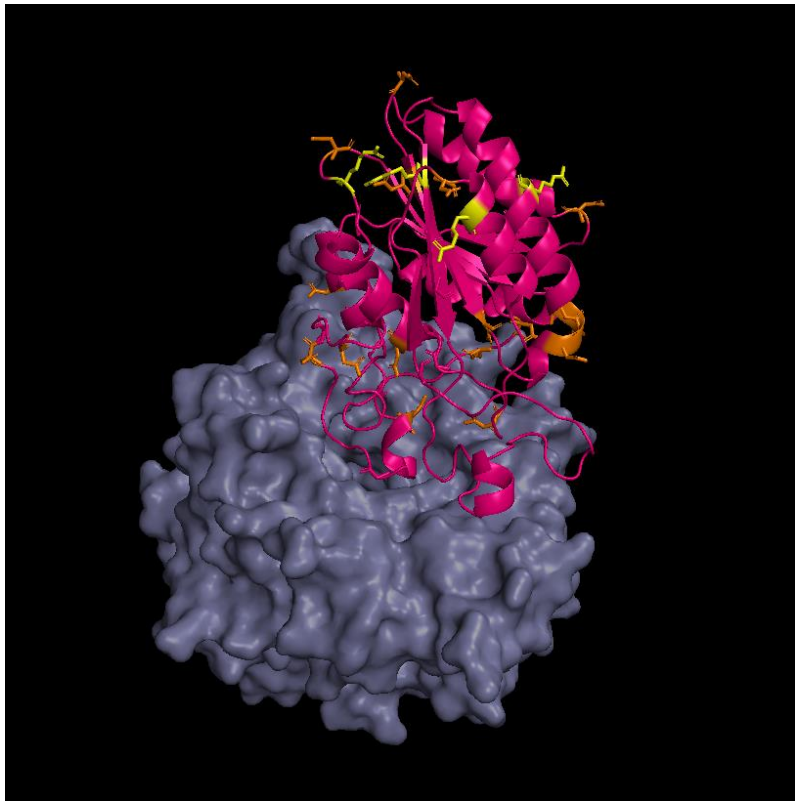
3 基于ZDock的结构预测

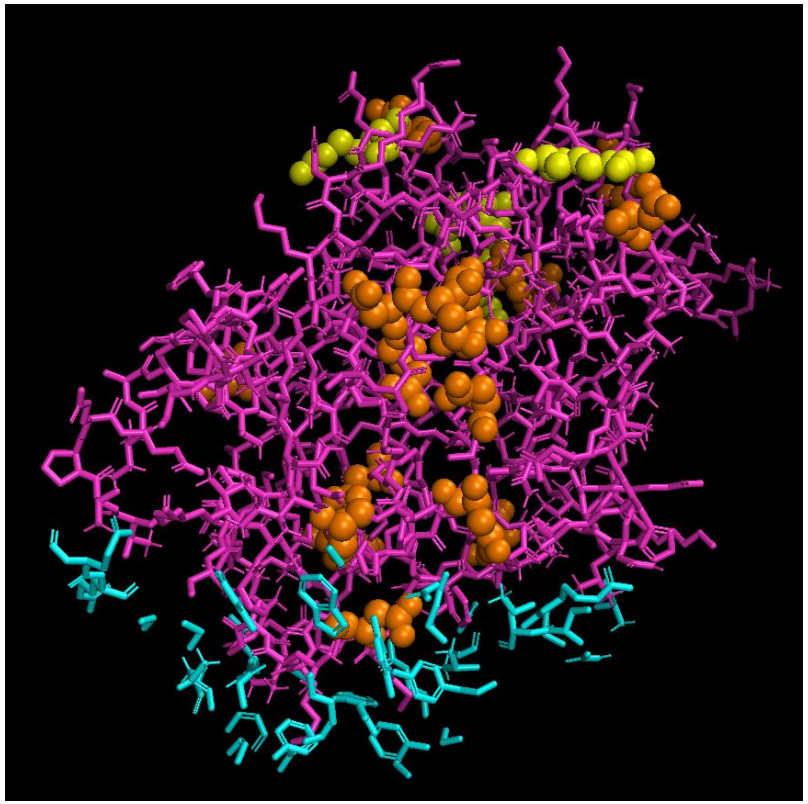




3 使用PyMOL进行可视化







3 基于AlphaFold3的结构预测

ipTM = 0.52
pTM = 0.59

pLDDT 评估的是**每条蛋白链自身结构的可信度。**图中深蓝色表示90+，浅蓝色表示70-90，黄色50-70，红色50-
pTM 评估的是**整个复合物整体组装的可信度，可以简单理解为pLDDT的全局版**（不过不是单纯平均）
ipTM评估**蛋白与蛋白之间相互作用界面的质量**，一般来说如果ipTM高于0.6，意味着预测的结合界面非常可靠

3 基于AlphaFold3的结构预测

ipTM = 0.92
pTM = 0.93

4 后续实验

1. In vitro functional validation

   候选AAV是否可以高效感染受体高表达细胞

2. In vivo muscle transduction in mice

   给小鼠注射候选AAV，观察其是否可以在体高效转导小鼠骨骼肌

Our Group

刘路旋：**课题**提供，背景&已有**结**果，**汇报**

闫泽淇：**测**序数据**处**理 pipeline

张博洋：AAV 系**统发**育分析

罗小山：AAV 与受体蛋白**对**接**结**构**预测**

共同完成 PPT 制作

Thanks for your attention!