



# Report

 汇报人: Lilian



## 文献来源


### nature communications

[Explore content](#) ▾ [About the journal](#) ▾ [Publish with us](#) ▾

[nature](#) > [nature communications](#) > [articles](#) > article

Article | [Open Access](#) | [Published: 01 December 2020](#)

## Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure

[Jan Zrimec](#), [Christoph S. Börlin](#), [Filip Buric](#), [Azam Sheikh Muhammad](#), [Rhongzen Chen](#), [Verena Siewers](#),  
[Vilhelm Verendel](#), [Jens Nielsen](#), [Mats Töpel](#) & [Aleksej Zelezniak](#) 

深度学习表明，基因表达编码在共同进化的相互作用基因调控结构的所有部分



## Abstract

- ✓ 了解控制基因表达的遗传调控密码是分子生物学的一个重要挑战。
- ✓ 文章对超过 20,000 个 mRNA 数据集应用深度学习，以检查从细菌到人类的 7 种模式生物中控制 mRNA 丰度的遗传调控密码。
- ✓ 在研究所涉及的所有生物体中，可以直接从 DNA 序列预测 mRNA 丰度，其中高达 82% 的转录水平变异编码在基因调控结构中。通过在基因调控结构中寻找 DNA 调控基序，作者发现基序相互作用可以解释 mRNA 水平的整个动态范围。

## Introduction

- 基因表达控制着所有生物的发育、适应、生长和繁殖；转录调控在过去几十年中一直是研究的核心领域。
- 但是人们无法量化 DNA 密码在多大程度上决定了 mRNA 的丰度，也无法理解这些信息是如何在 DNA 中编码的；缺乏这种定量理解阻碍了通过简单地**操纵四个 DNA 核苷酸的序列来准确控制 mRNA 和蛋白质水平的潜力。**

### **序列控制表达水平**

- 多种生物体中蛋白质和 mRNA 水平之间的强烈一致性表明**转录是蛋白质丰度的主要决定因素；**
- **mRNA 转录是通过基因调控结构控制的**，该结构由编码区和顺式调控区组成，包括启动子、非翻译区 (UTR) 和终止子，**每个区域都编码大量与 mRNA 水平相关的信息**，目前仍不清楚非编码区是如何协同工作来调节基因的表达水平的。

### **序列上部分区域的协同工作机制未知**



## Introduction

- 目前的天然和合成方法在研究基因调控结构不同部分及其表达联合调控之间的整体关系的能力方面存在根本性的局限性



本篇文章，搜集来自包括酵母、智人等7种模式生物的20000多个RNA-seq实验分析中得到的100,000多个天然序列，利用**深度神经网络**学习天然序列并预测基因的水平。

验证部分：使用 **GFP 荧光测量实验** 证明了我们的模型在酿酒酵母中用任何所需基因指导基因表达工程的潜力。

*本文工作*

对于每个生物体，编码区和调控区是根据转录本和 ORF 边界提取的。DNA序列被独热编码 (one-hot encoding)，UTR序列被零填充到指定长度，密码子频率被归一化为概率，并且计算了**8个mRNA稳定性变量**，包括：长度5'-UTR、ORF 和 3'-UTR 区域的含量、5'-和 3'-UTR 区域的 GC 含量以及 ORF 中每个密码子位置的 GC 含量。



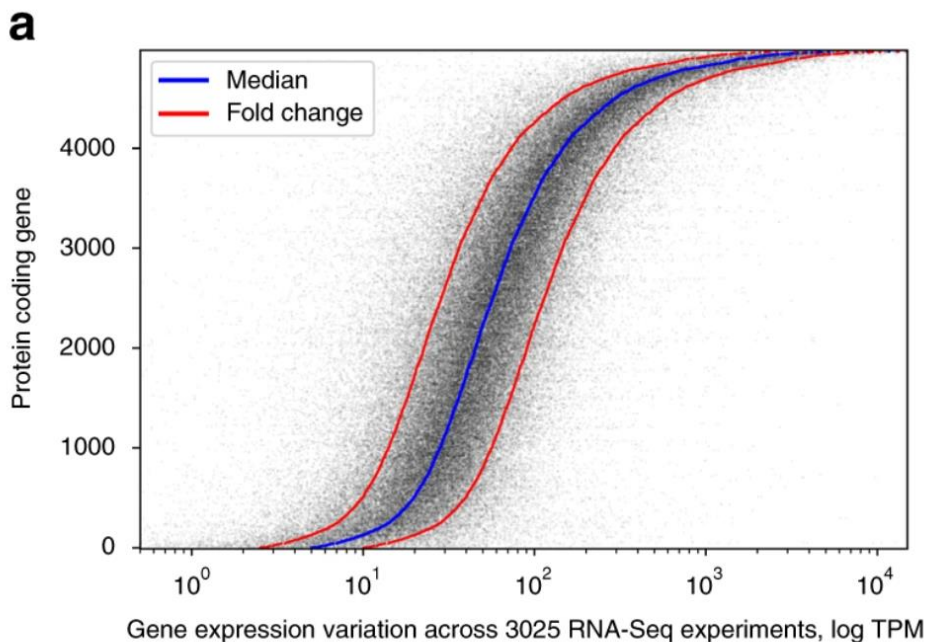
- 1、通过质量控制的实验进行过滤
- 2、将原始 mRNA 数据转化为每百万转录本 (TPM) 计数
- 3、去除 mRNA 输出为零 (TPM < 5) 的基因
- 4、mRNA 计数经过 Box-Cox 转化

通过上述数据处理获得建模数据集，包括成对的基因调控结构解释变量和 mRNA 丰度响应变量。

## ■ The dynamic range of gene expression levels is encoded in the DNA sequence

基因表达水平的动态范围编码在 DNA 序列中

为了探索 DNA 序列和基因表达水平之间的关系，选用包含 3025 个高质量酿酒酵母 RNA-Seq 实验的数据集，涵盖了来自 2365 个独特研究的大部分可用实验条件。

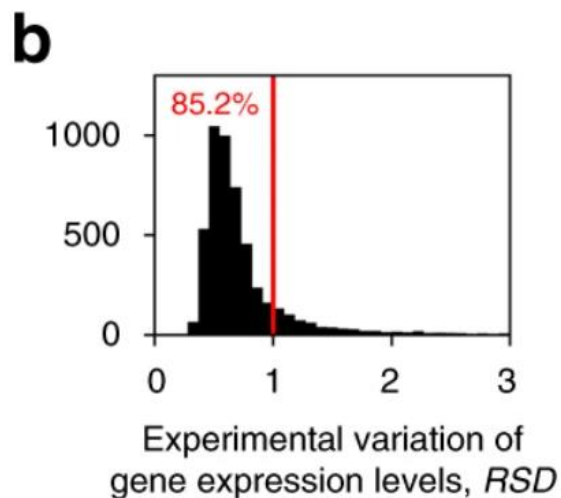


79% 的酵母蛋白编码基因表达水平变化范围在1 倍以内；

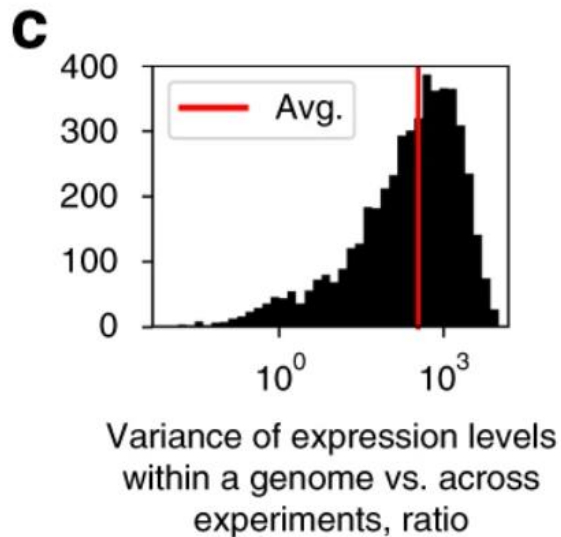
3025 次 RNA-seq 实验中蛋白质编码基因的表达水平（每百万转录本，TPM）



在整个生物条件范围内变化最大的基因（图 1b:  $RSD > 1$ ）在代谢过程、运输和应激反应中显著富集，而最稳定的基因（ $RSD < 1$ ）显著富集于 TFIID 型组成型启动子。



以RSD表示的基因表达水平的实验变化



85% 的基因在 1 个相对标准偏差内  
( $RSD = \sigma/\mu$ )

所有基因的平均 TPM 值的动态范围跨越 4 个数量级，整个基因组内表达水平的方差平均比实验中每个基因的方差高 340 倍

基因组内表达水平的差异与实验中每个基因的表达水平差异之间的比率分布





使用 DNA 序列信息作为输入，建立了一个基于深度卷积神经网络 (CNN) 的回归模型，能够识别序列中的功能性 DNA 基序，并训练模型以预测中值基因表达水平。

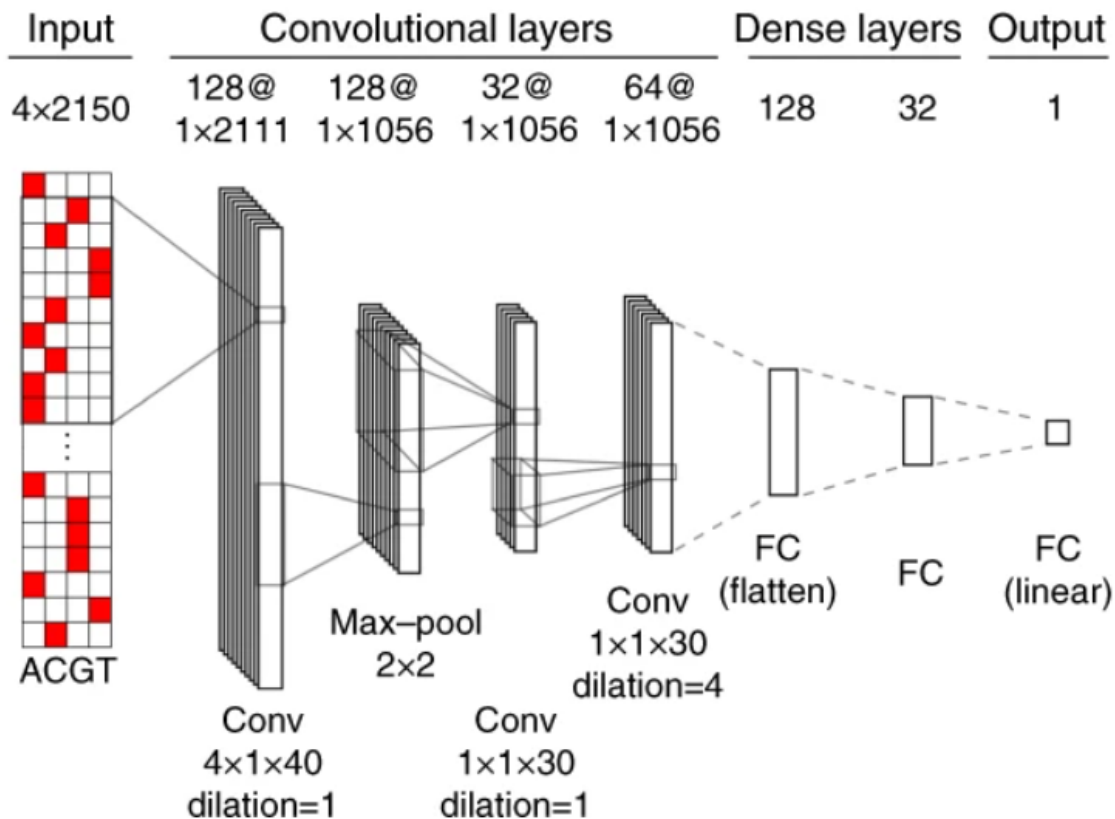
### 测试了不同的神经网络架构



表明产生最佳结果的架构是同时训练的 CNN (3 层) -FC (2层)



e



指定了表示参数layers、stride、kernels、filters、max-pooling和dilation大小的值

批量标准化和权重丢弃在所有层之后应用，在 CNN 层之后应用最大池化。

使用了具有均方误差 (MSE) 损失函数的 Adam 优化器和具有 uniform127 权重初始化的 ReLU 激活函数。

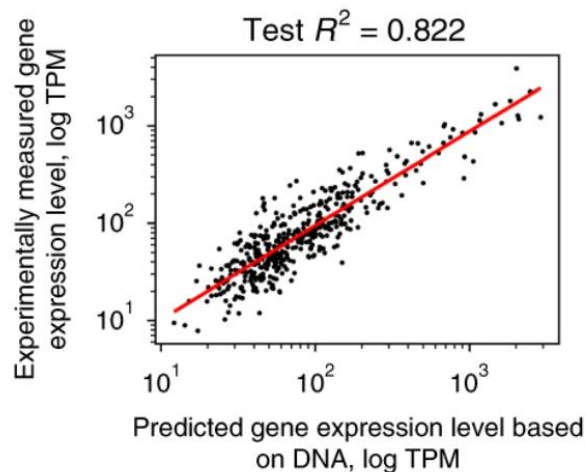
总共有 26 个超参数通过 Hyperopt v0.1.1128 使用树形结构的 Parzen 估计器方法在默认设置下进行了 1500 次迭代的优化。

由于基于片段的转录本丰度估计的技术标准化偏差，除了**mRNA计数**和**ORF长度**之间，变量之间没有发现显著的成对相关性。

为了避免与基于读数的测序相关的潜在技术偏差，mRNA 水平针对基因长度偏差进行了校正。共有 3433 个基因序列用于训练模型，381 个用于调整模型超参数，424 个用于测试。在优化模型后，对测试集的预测性能 ( $R^2$ 测试 = 0.822,  $p$ 值 <  $1e-16$ ) 表明 DNA 编码了大多数有关 mRNA 表达水平的信息。

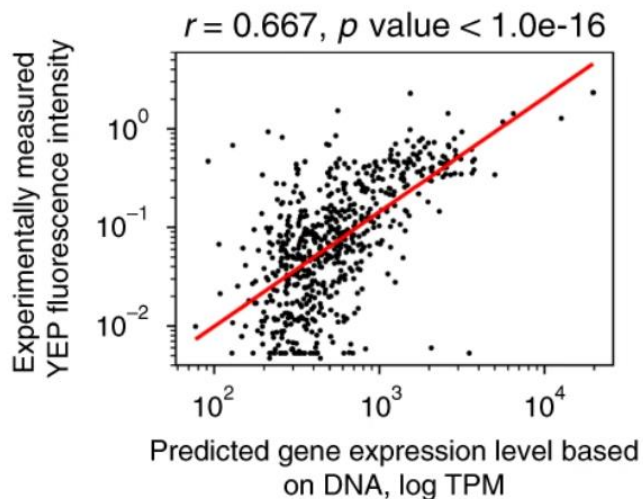
用酿酒酵母实验确定（真实）与预测的表达水平在保留的测试数据集 ( $n = 425$ ) 上建立模型。  
红线表示最小二乘拟合。

**f**





g



比较实验荧光测量值与预测的表达水平  
( $n = 625$ )。  
红线表示最小二乘拟合。

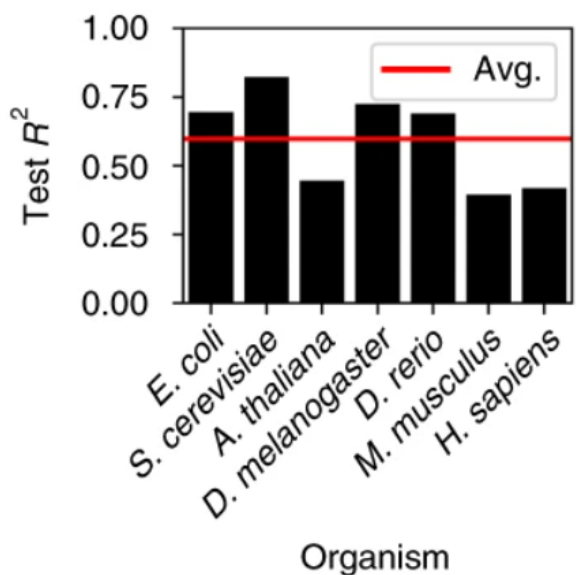
验证

数据集包括在 10 种不同条件下记录在合成结构中的约 900 个天然酵母启动子的测量活动，具有单个强终止子 (ADH1) 和 YFP 荧光报告器。

在所有 10 种条件下，基于合成构建体的 DNA 序列和 YFP 密码子频率推断的 mRNA 水平预测与实验 YFP 读数高度一致 (Pearson 的  $r$  从 0.570 到 0.718,  $p$  值  $< 1e-16$ )

为了研究**是否也可以从其他模式生物中的 DNA 序列预测 mRNA 丰度**，处理了来自 1 个原核和 5 个真核模式生物的额外 18,098 个 RNA-Seq 实验，步骤与酵母相同。

总体而言，对高等真核生物的预测不太准确，这可能归因于转录复杂性的增加，例如由于可变剪接、组织间表达差异以及远距离增强子相互作用在目前的模型中。因此，预测性能与模型生物的基因组复杂性相关。



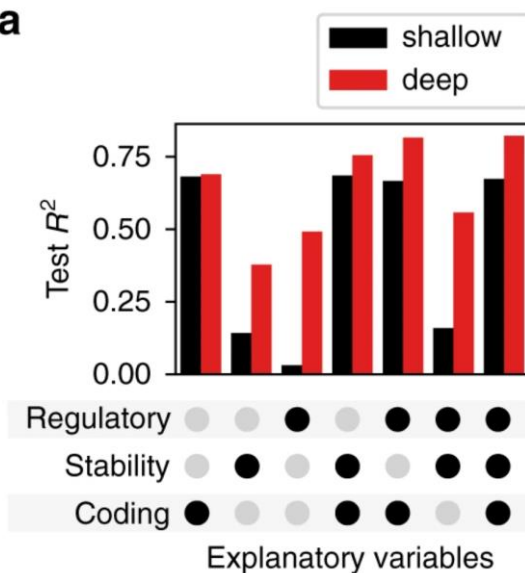
模式生物体的平均性能为 0.6 证实了所有生物体中的大多数 mRNA 表达差异可以直接从 DNA 预测。

## Coding and cis-regulatory regions jointly contribute to gene expression prediction

### 编码区和顺式调控区共同促进基因表达预测

与完整模型类似，在启动子、5' -UTR、3' -UTR 和终止子区域及其组合上独立训练了多个 CNN 模型。

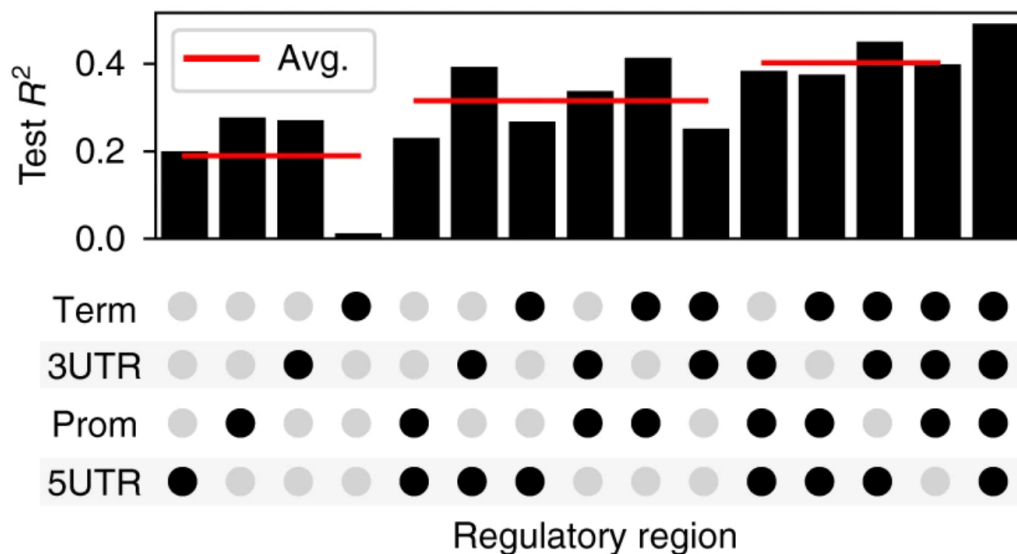
使用各种回归算法进行了浅层建模，包括多元线性回归、弹性网络、随机森林和具有嵌套交叉验证的支持向量机等，来证明使用深度卷积网络的合理性；



测试  $R^2$  具有编码（密码子概率）、mRNA 稳定性和顺式调节区域的不同组合，分别使用浅层（黑色）和深层建模（红色）



b



- ✓ 单独一个调控区域可以解释不到 28% 的 mRNA 丰度水平变化, 但当使用调控区域的组合时, 每个区域都有助于预测 mRNA 水平并提高模型性能
- ✓ 在所有四个调节区域上训练的模型约占 mRNA 丰度变化的 50%

**整个基因调控结构对于控制基因表达水平很重要。**

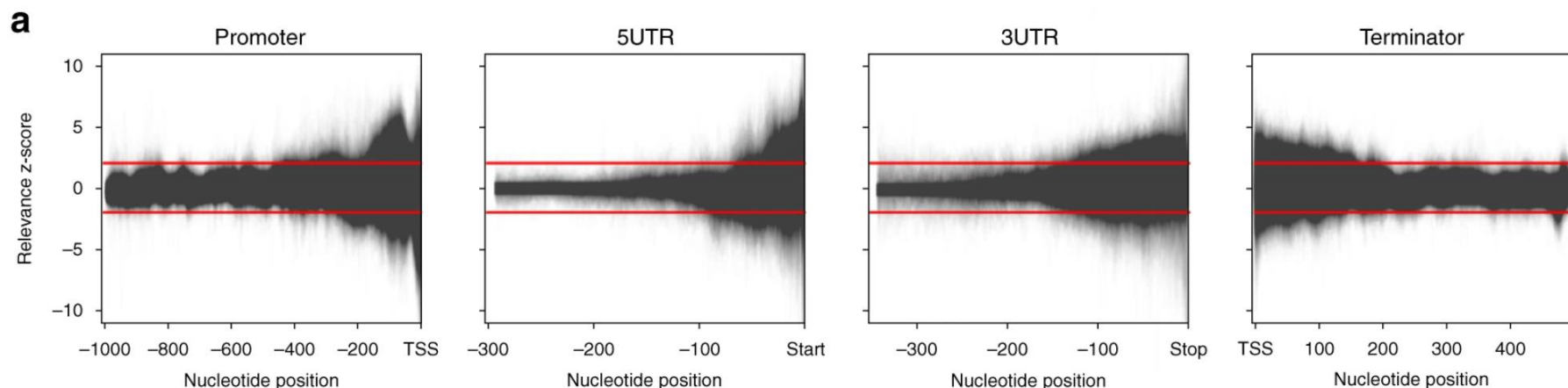


## Deep learning identifies specific DNA positions controlling gene expression levels

### 深度学习识别控制基因表达水平的特定DNA位置

开发了一个pipeline来评估每个特定核苷酸位置的相关性预测的基因表达水平，为了确定最能预测基因表达水平的DNA序列的特定部分。

对于每个基因，沿着其调节DNA序列移除了10个碱基对的滑动窗口，并将封闭序列的预测与原始未封闭序列的预测进行比较。与原始数据显着偏离（超过 $\pm 2$ 个标准偏差）的输入DNA序列的封闭部分被认为与基因表达变化最相关。



通过查询深度模型获得的跨顺式调控区域序列的相关性概况

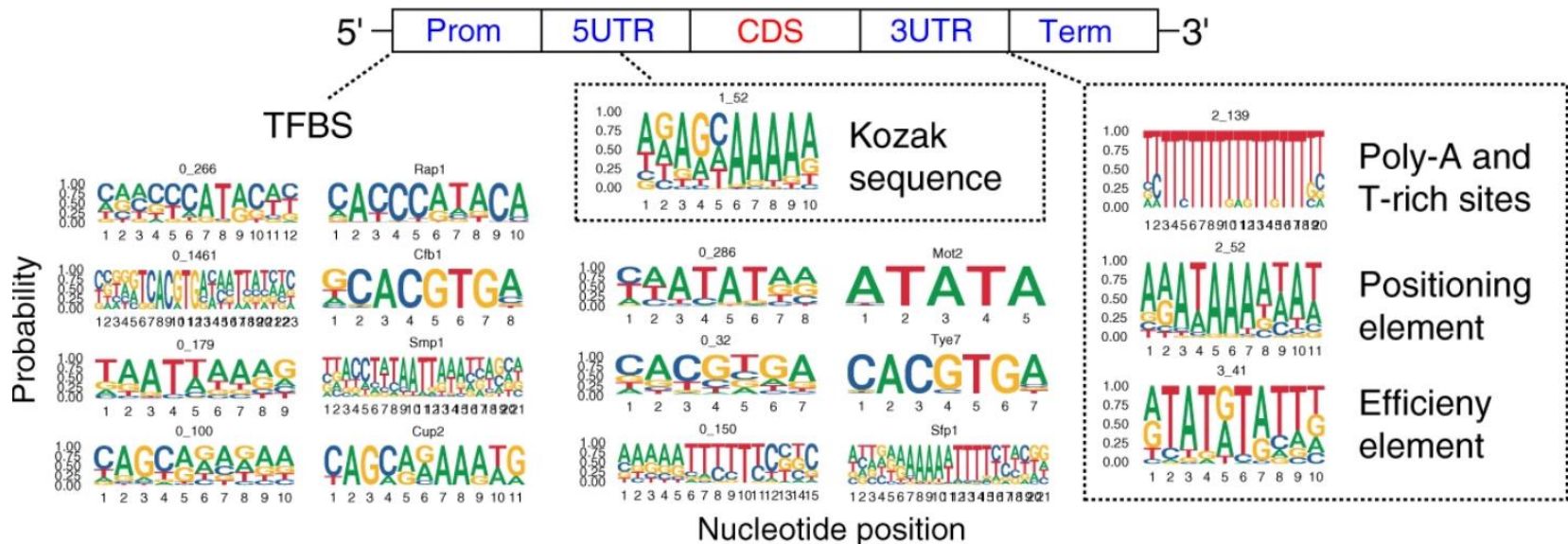


每个基因的启动子中的 214 个碱基对 (bp)、5'-UTR 中的 74 bp、3'-UTR 中的 94 bp 和终止子区域中的 127 bp 显著影响其表达水平的预测

使用聚类 and 比对从一组所有显著相关的 DNA 序列中确定了对于预测表达水平很重要的特定调控 DNA 基序。在所有 4 个调节区域中发现了超过 2200 个表达相关的调节 DNA 基序，且大多数基序对于每个特定区域都是唯一的。

➡ 每个调控区域都包含与基因表达水平相关的独特信息

f



在与已发表的基序和序列元件相对应的所有顺式调节区域中发现的调节 DNA 基序的示例



文章的最后对所确定的序列motif进行了分析，通过统计分析基序之间的相互作用，发现 2 到 6 个基序的 9,962 种组合比在基因区域中单独出现的频率更高。作者发现在统计上更有可能在基因中一起出现而不是单独出现的基序组合，形成了一种特有的模式，并将这些模式称为“调节规则”（regulatory rules）。

基序共现规则还定义了比单个基序更具体的密码子使用范围，进一步支持文章的结果，即整个基因调控结构，包括编码区和非编码区，是一个共同进化的相互作用单元，协同影响并控制着基因表达。