



透過基於後綴數組的特徵工程優化 CRISPR-Cas9 裂解效率預測

Optimizing CRISPR-Cas9 Cleavage Efficiency
Prediction via Suffix Array-based Feature Engineering

指導教授：賀保羅 專題成員：林映汝

Table of contents

01

Introduction

02

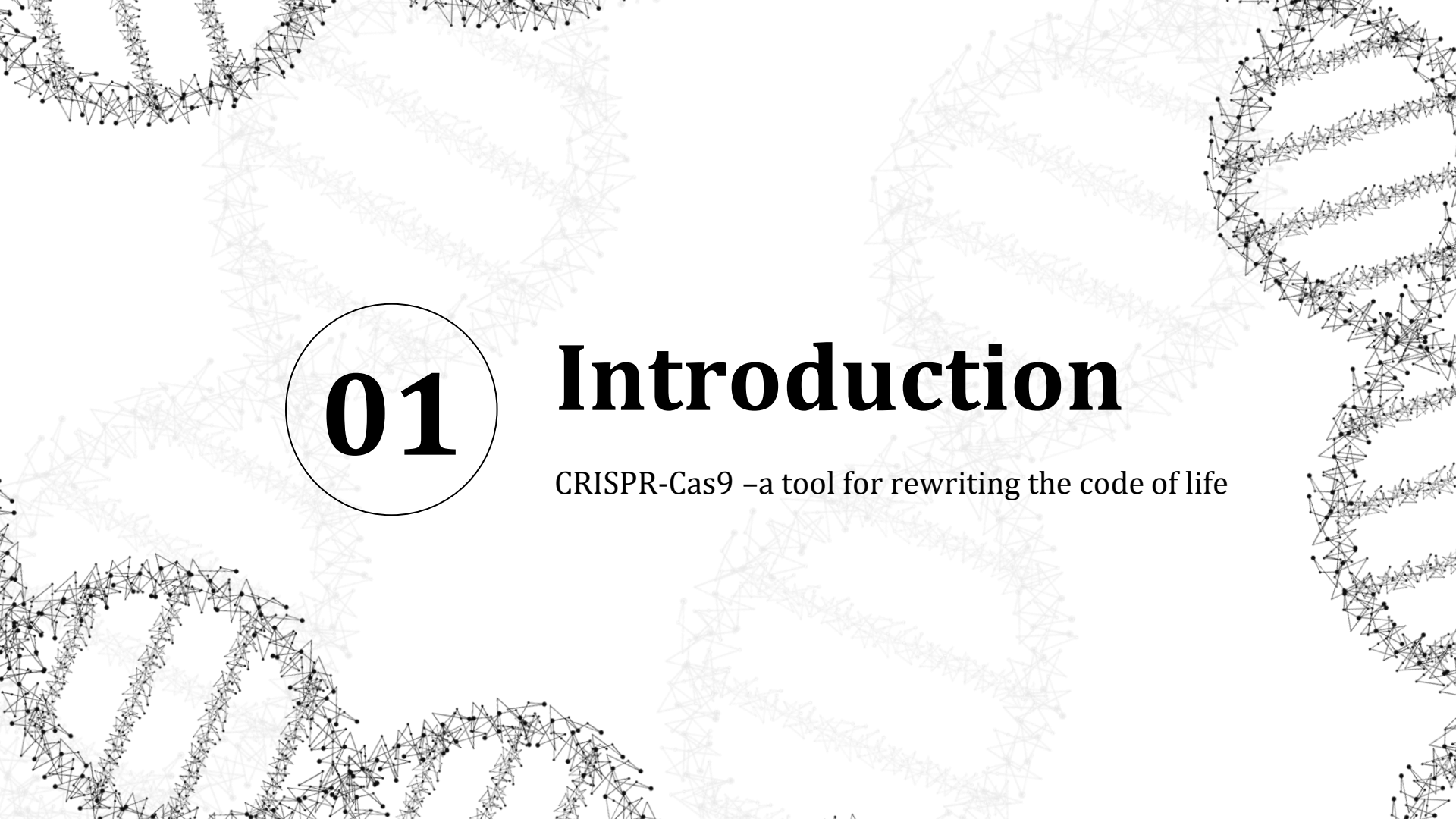
Methodology

03

Results

04

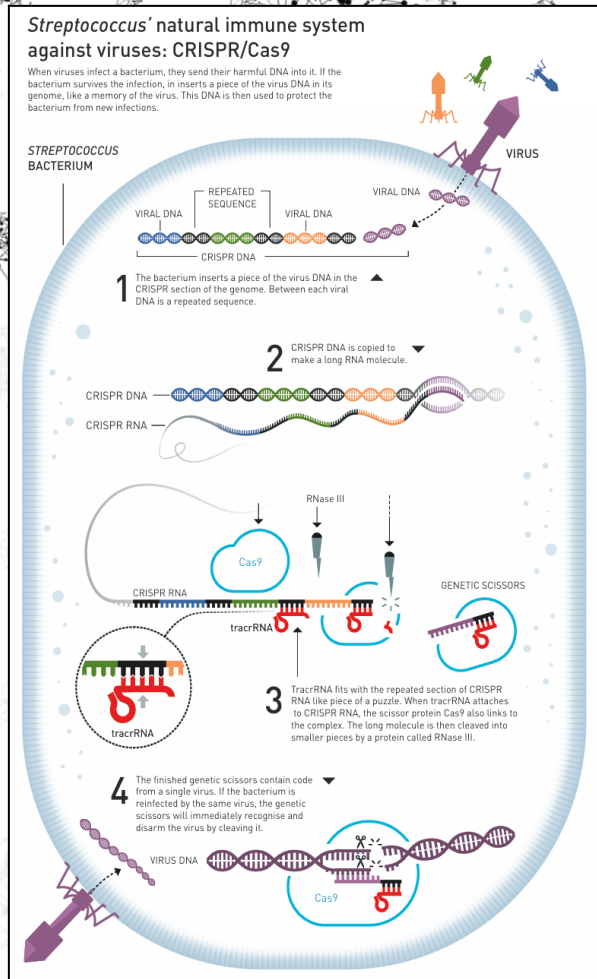
Conclusion

The background features a complex, abstract design. It includes several overlapping, semi-transparent helical structures that resemble DNA double helices. These are composed of interconnected black dots (nodes) and thin black lines (edges), creating a network-like appearance. The helices are arranged in a way that they seem to spiral around the central text area. The overall color scheme is black and white, with the background elements being lighter than the main text.

01

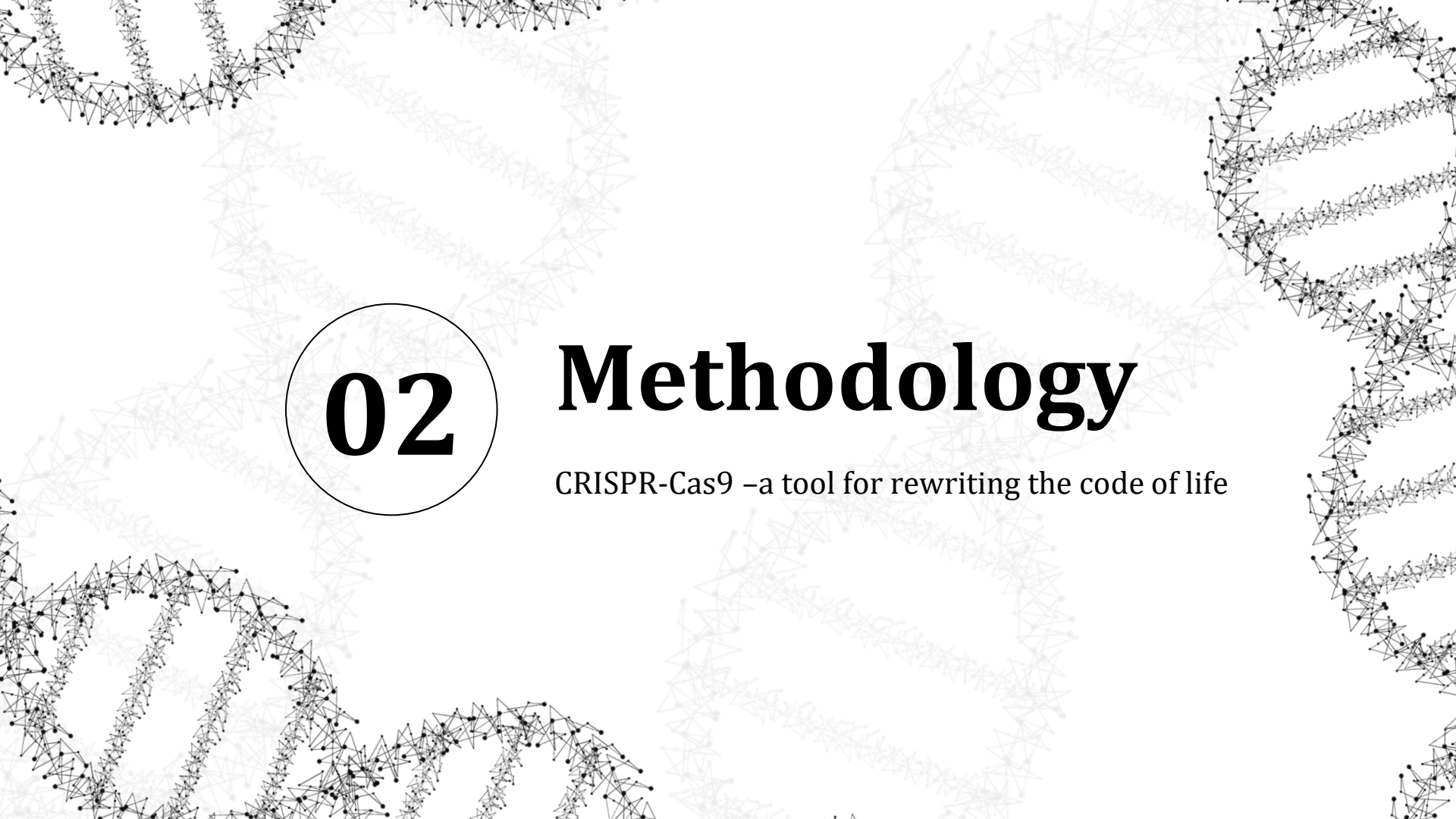
Introduction

CRISPR-Cas9 –a tool for rewriting the code of life



THE NOBEL PRIZE IN CHEMISTRY 2020

裂解（cleavage）過程中潛在的
脫靶效應（off-target effect）危機
多源於sgRNA（single guide RNA）
與目標DNA的匹配問題



02

Methodology

CRISPR-Cas9 –a tool for rewriting the code of life

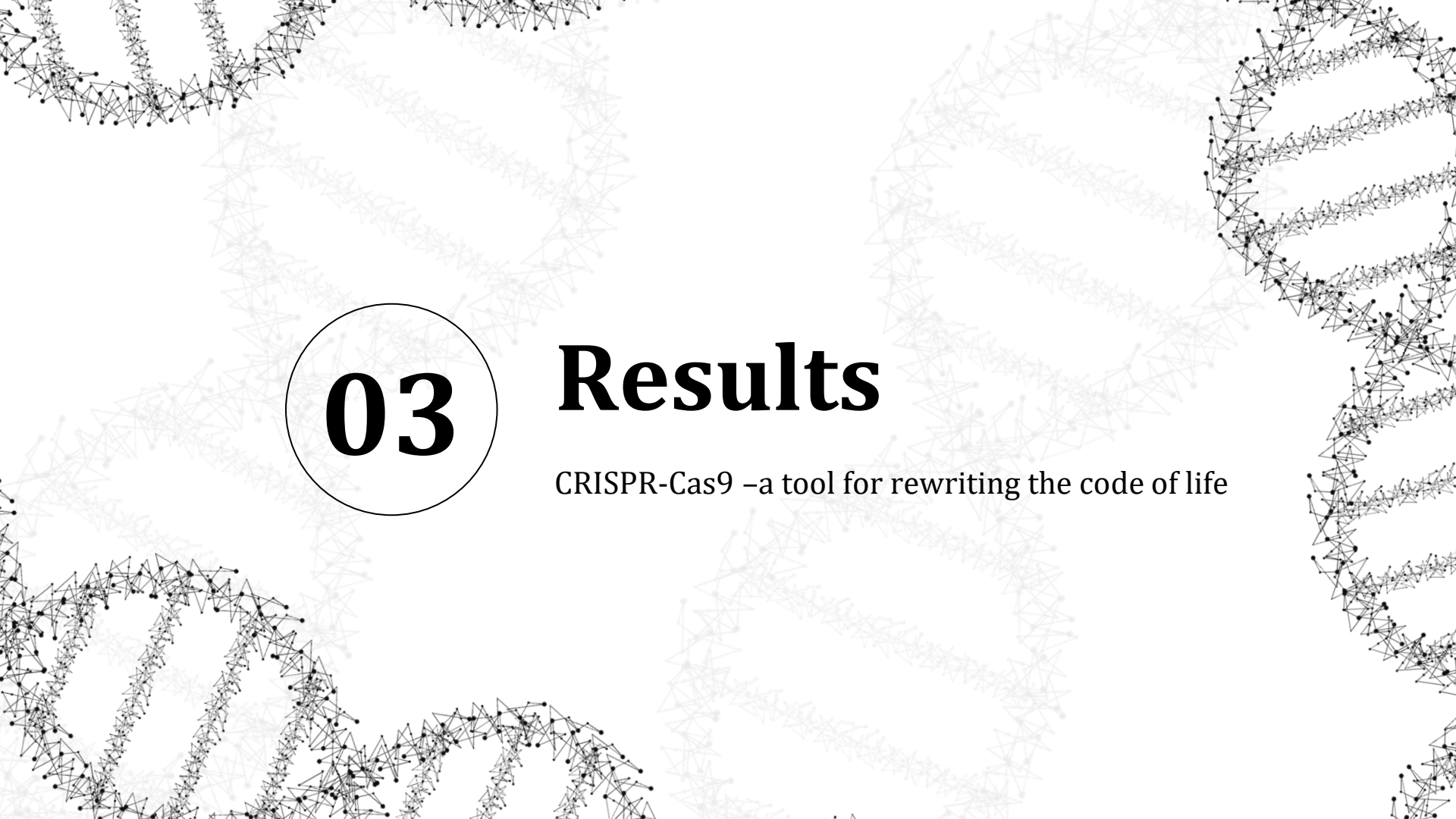
Methodology

Preprocessing:

- human reference genome (hg38) FASTA fetch
- suffix array construction

Genome-wide sgRNA Occurrence:

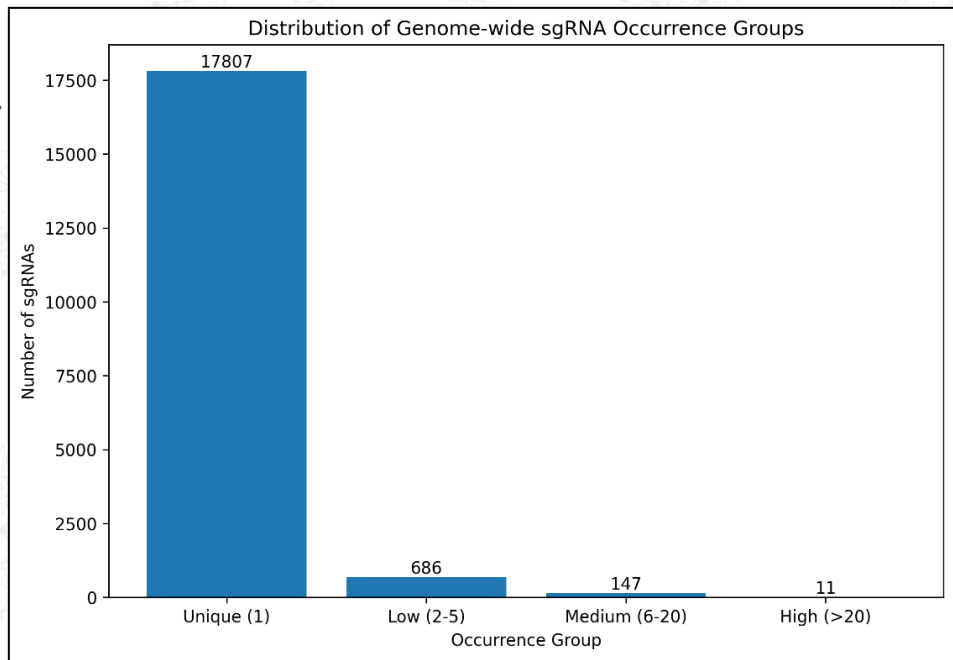
- binary search-based sgRNA(a total of 18651 records) matching
- feature engineering for machine learning

The background features a complex, abstract design. It includes several overlapping, light gray helical structures that resemble DNA double helices. These are interspersed with a network of black dots connected by thin lines, creating a mesh-like pattern. The overall aesthetic is scientific and modern.

03

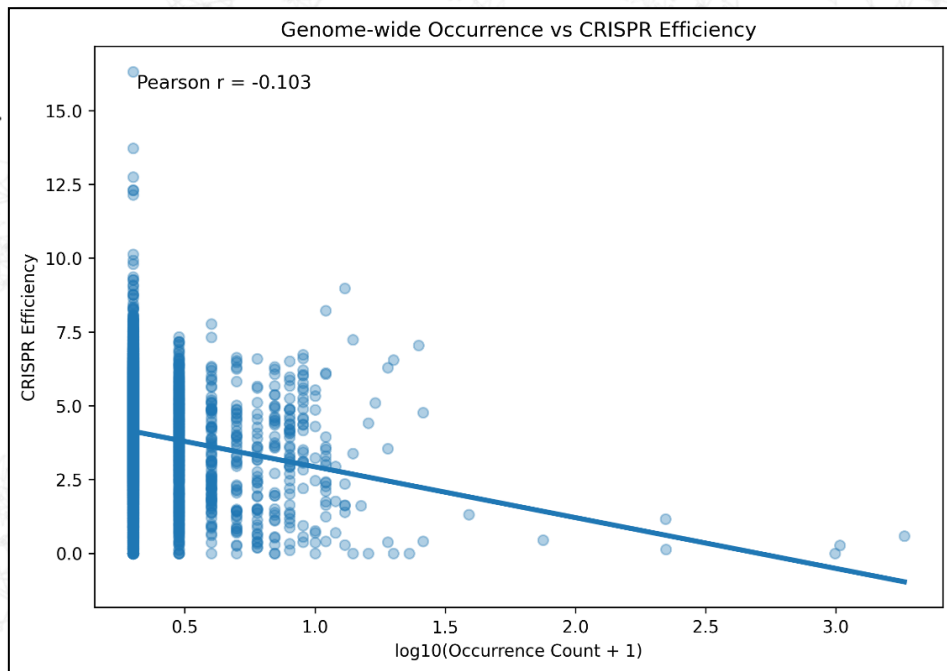
Results

CRISPR-Cas9 –a tool for rewriting the code of life



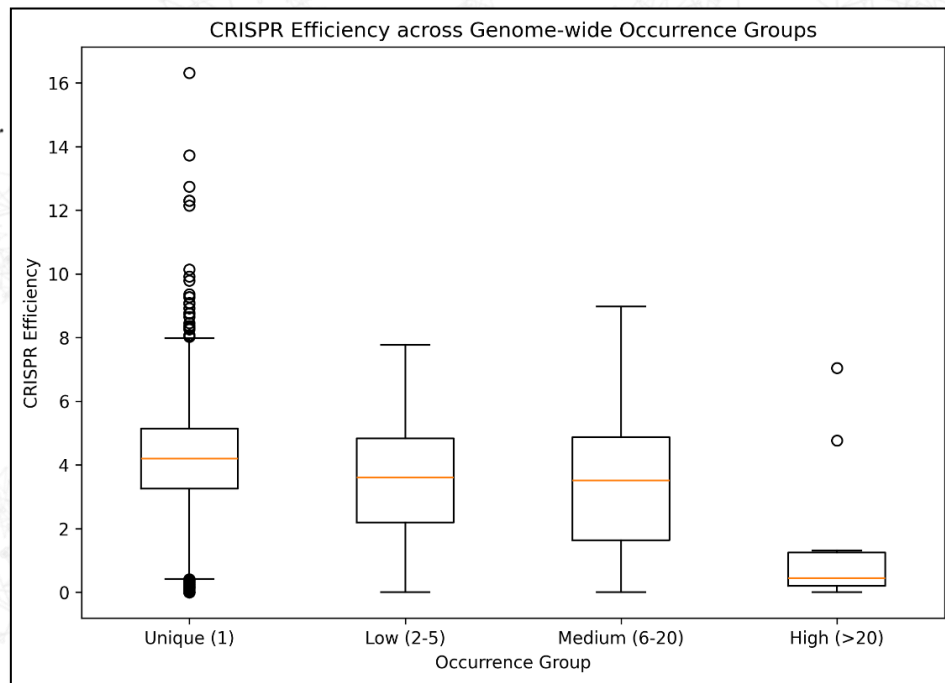
圖一、sgRNA在人類基因組中的分布情形

- 大部分sgRNA唯一（unique）→多數與目標DNA具高度專一性
- 少部分sgRNA重複程度高→ off-target風險程度高



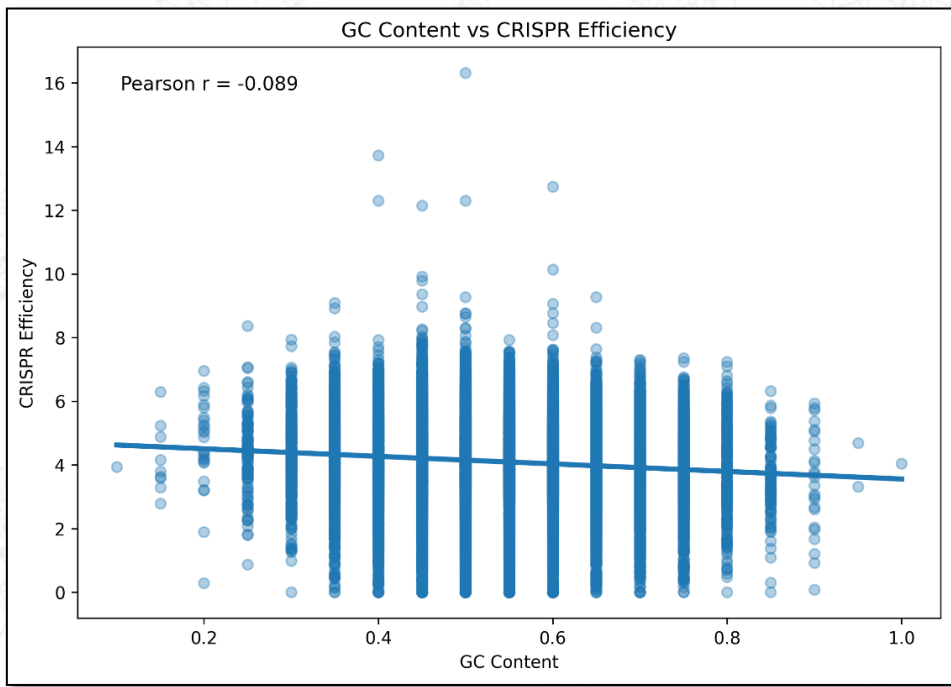
圖二、sgRNA的重複程度是否影響CRISPR裂解效率預測

兩者呈弱負相關 → 重複程度高的sgRNA可能降低基因編輯效率



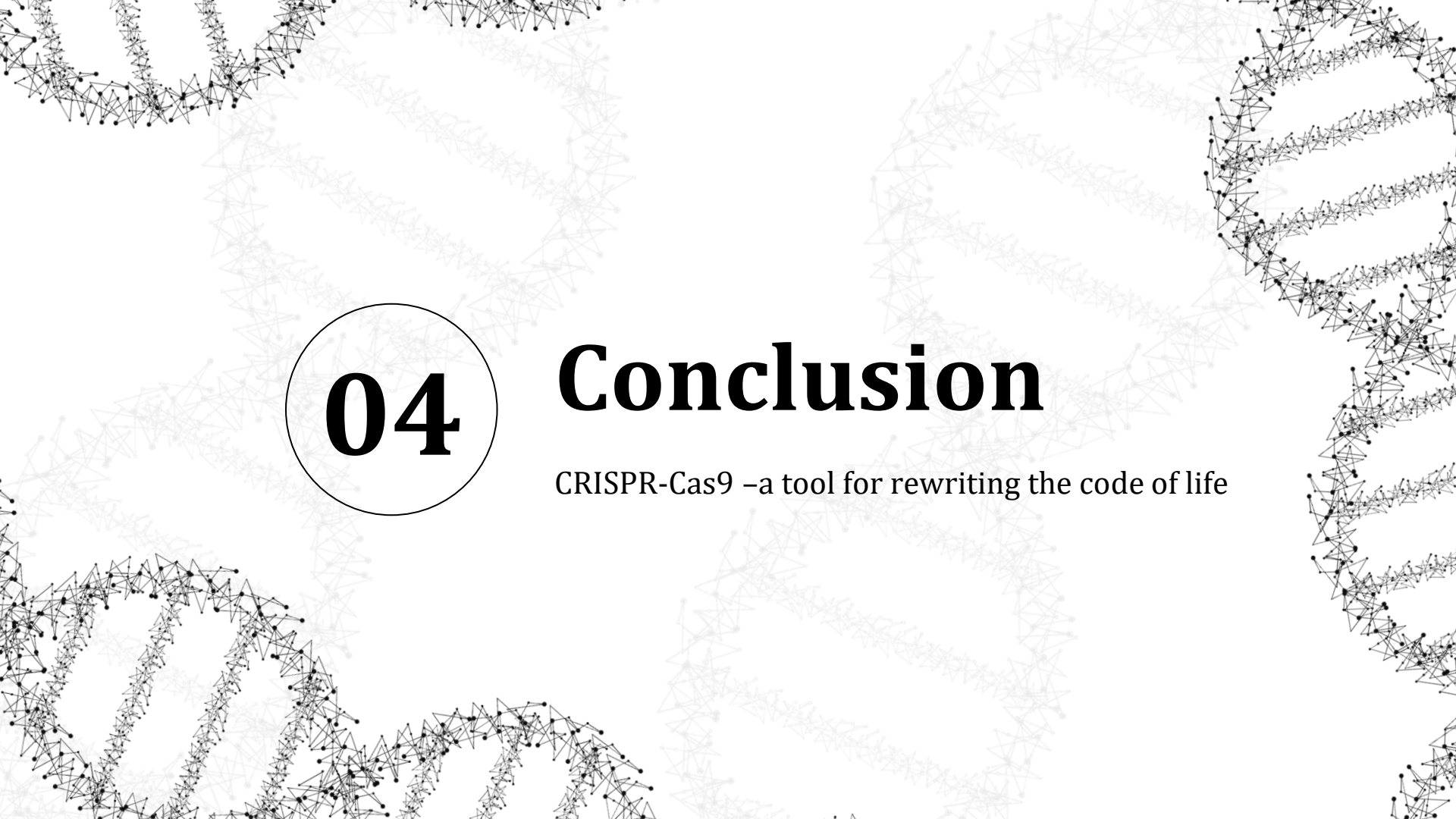
圖三、比較不同occurrence group間之CRISPR裂解效率

high occurrence group整體efficiency較低。此外，其分布相對集中，顯示重複程度高的sgRNA普遍呈現較低的編輯效率，而非僅由少數極端值所造成。



圖四、分析GC content與CRISPR裂解效率的關係

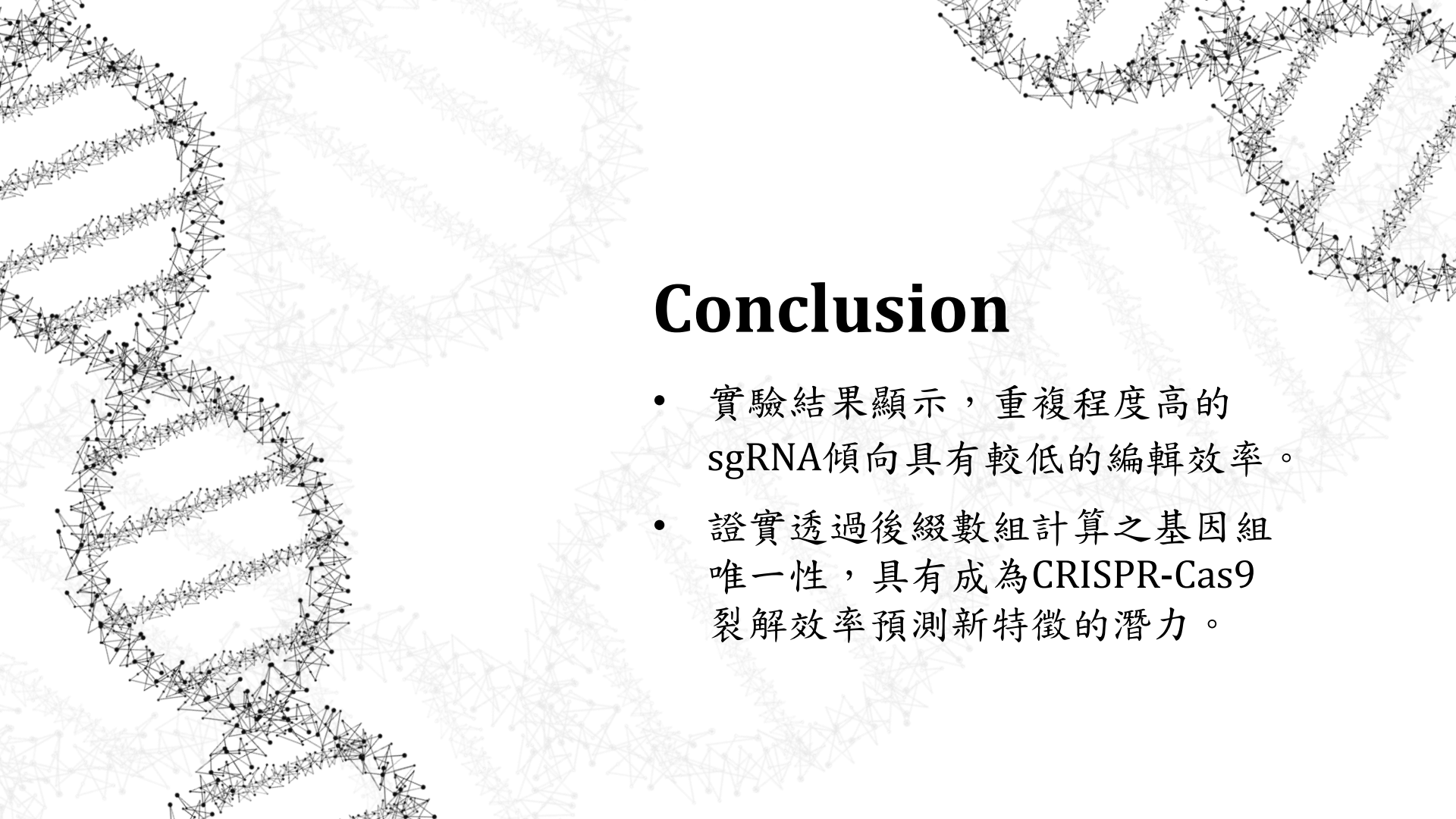
兩者呈弱負相關 → 傳統sequence-level feature無法充分解釋基因編輯效率（對照圖二的occurrence feature）



04

Conclusion

CRISPR-Cas9 –a tool for rewriting the code of life



Conclusion

- 實驗結果顯示，重複程度高的 sgRNA 傾向具有較低的編輯效率。
- 證實透過後綴數組計算之基因組唯一性，具有成為 CRISPR-Cas9 裂解效率預測新特徵的潛力。