

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

Background & Motivation

- CRISPR-Cas9—a tool for rewriting the code of life這把編輯過程簡單明瞭且高效準確的基因剪刀曾於2020年榮獲諾貝爾化學獎。
- 其裂解(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。
- 現有多數CRISPR裂解效率預測的模型主要聚焦於sgRNA本身的序列特徵，例如GC content、鹼基組成與PAM sequence等。
- 其中「目標序列於整個基因組中的上下文關係」較少被考慮。
- 本研究希望透過建立uniqueness score驗證「sgRNA的genome-wide occurrence與CRISPR裂解效率預測具潛在關聯性」。

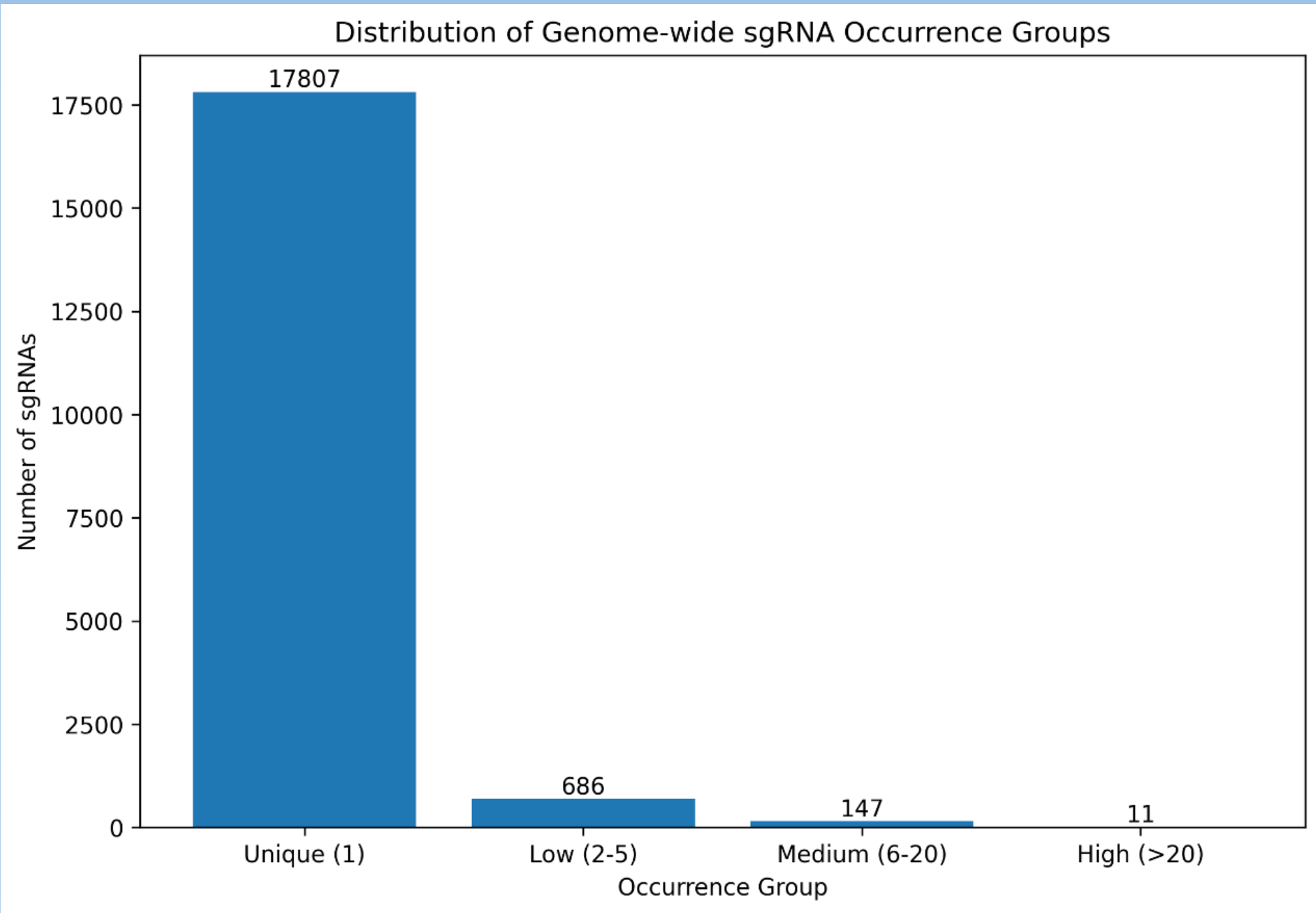
Methodology & Results

Preprocessing:

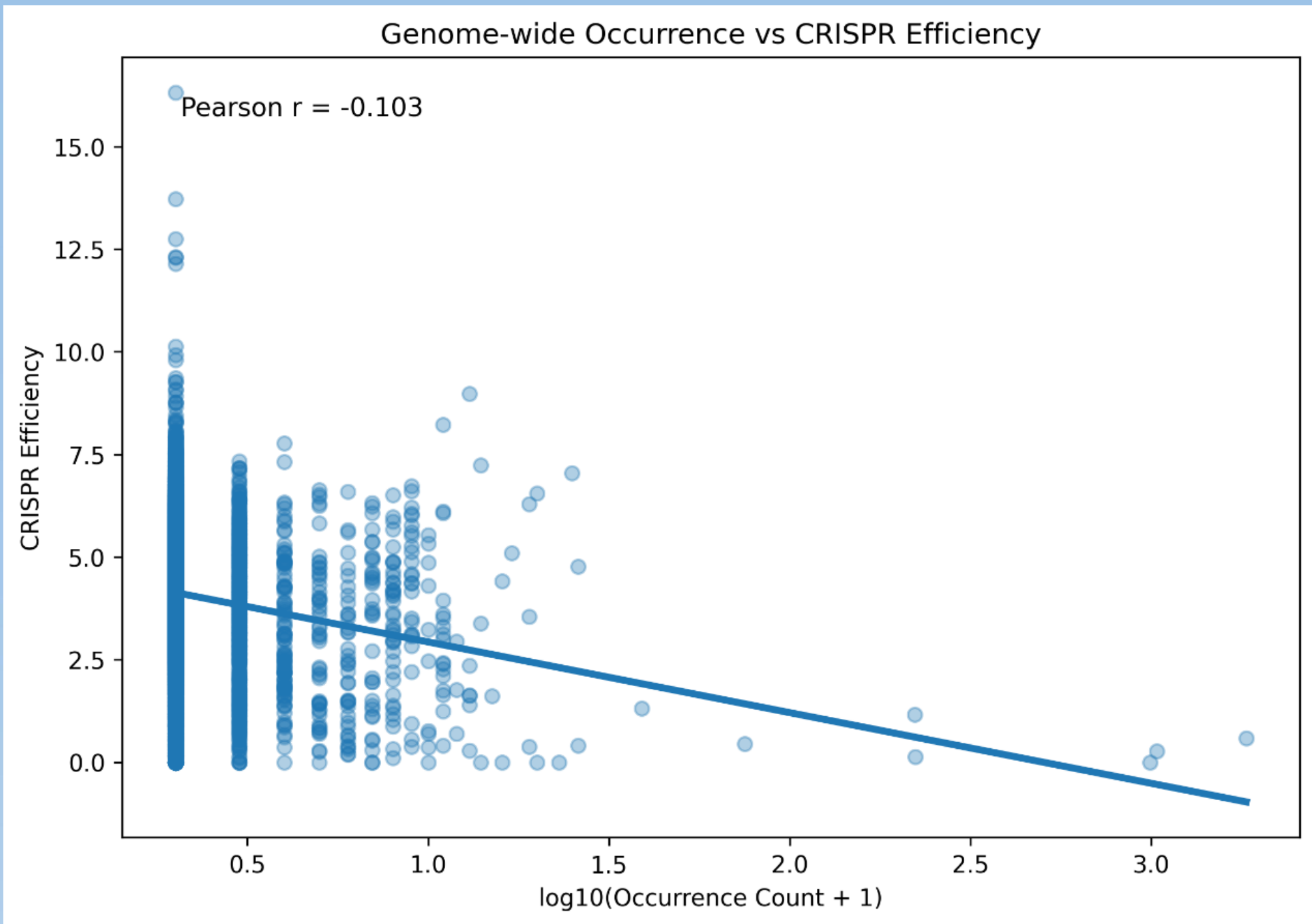
- human reference genome (hg38) FASTA fetch
- suffix array construction

Genome-wide sgRNA Occurrence:

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

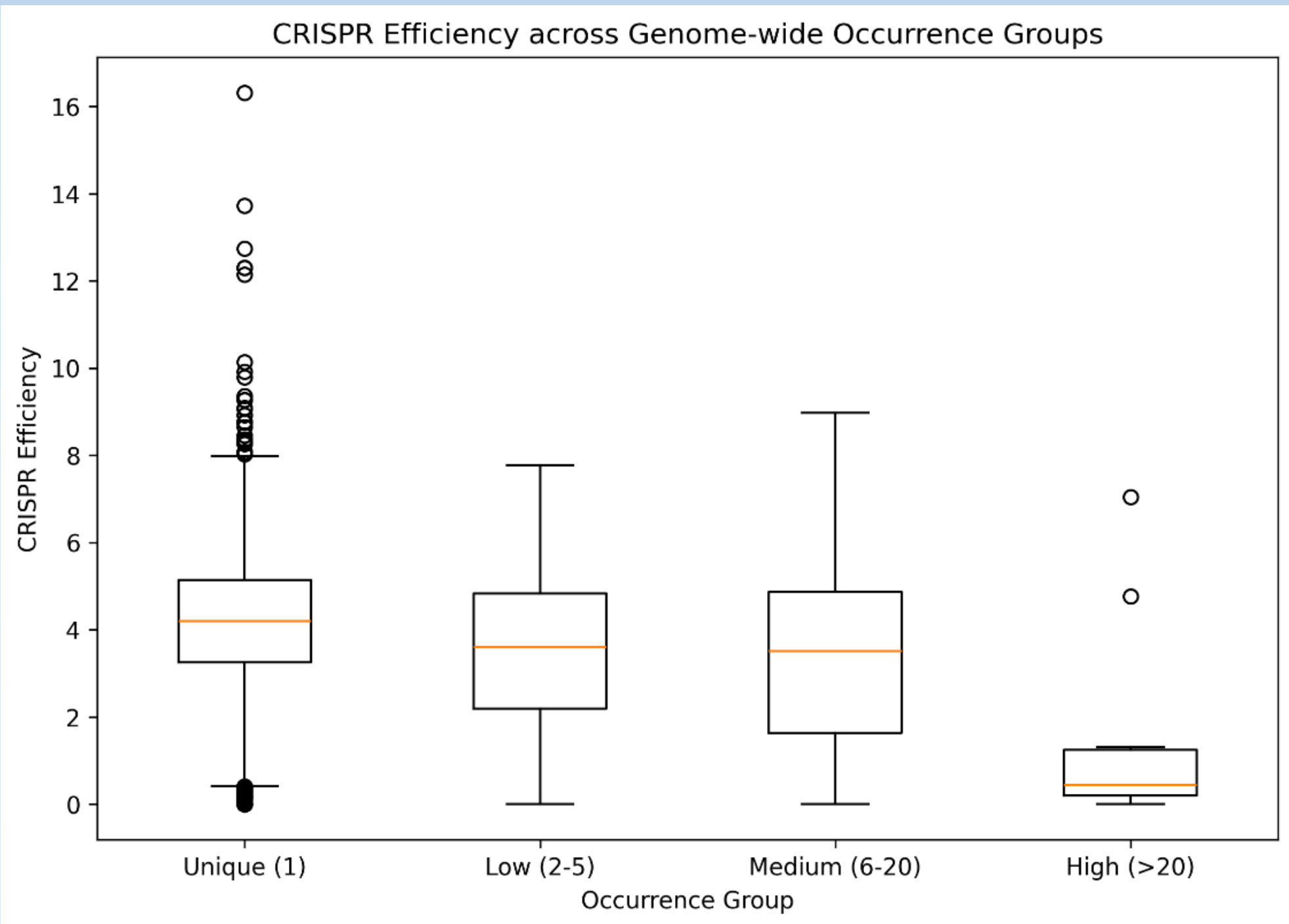


圖一、sgRNA在人類基因組中的分布情形



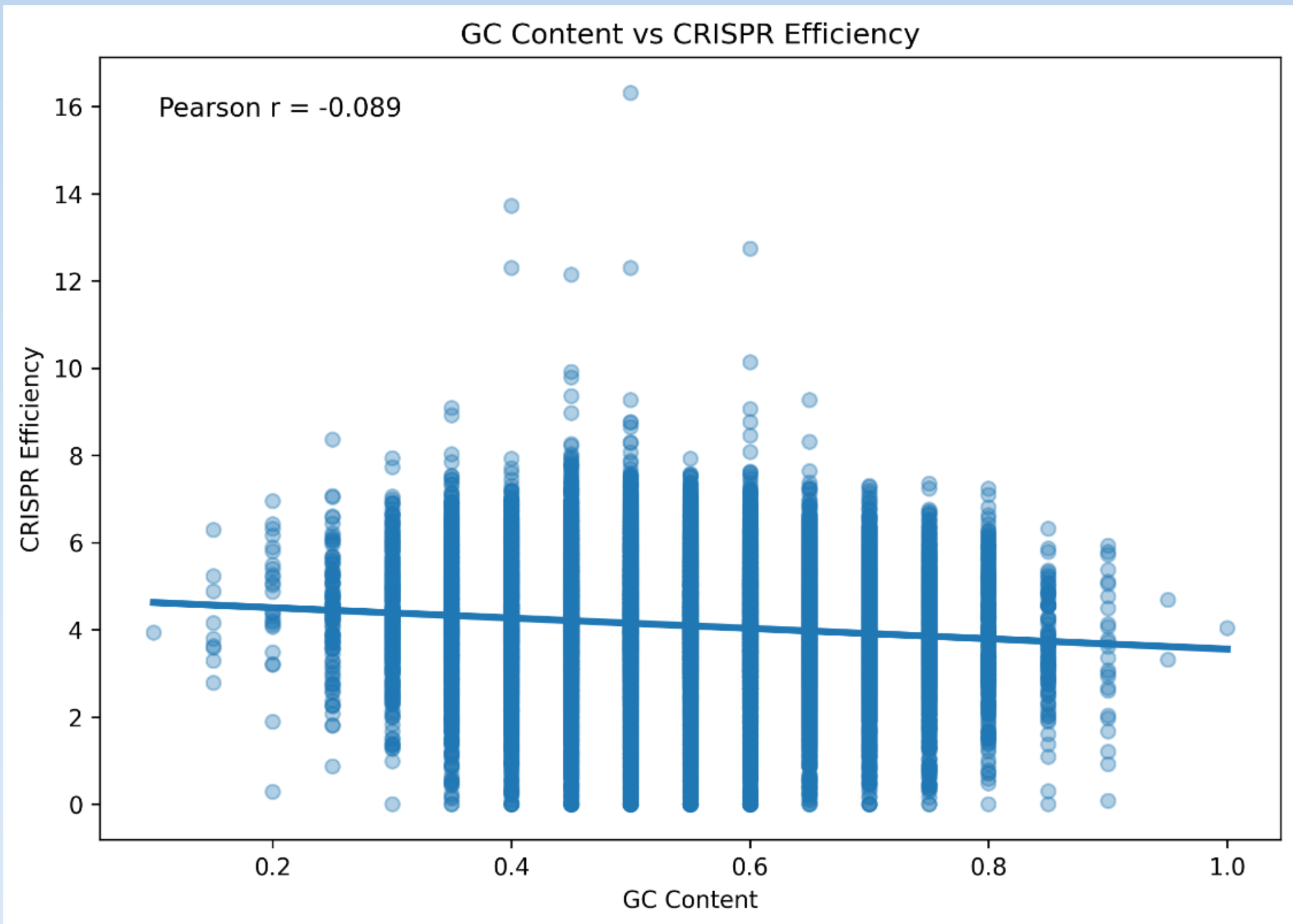
圖二、sgRNA的重複程度是否影響CRISPR裂解效率預測

- 大部分sgRNA唯一(unique) → 多數與目標DNA具高度專一性 兩者呈弱負相關 → 重複程度高的sgRNA可能降低基因編輯效率
- 少部分sgRNA重複程度高 → off-target風險程度高



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)

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。