

De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds

《基於Hi-C輔助從頭組裝-埃及斑蚊基因體組裝生成染色體尺度之長片段》

/ Paper Reproduction

NCCUCS 楊明翰 108753203

/ 論文覆現實驗

NCCUCS 曾偉綱 108753122

NCCUCS 王神鐸 107753048

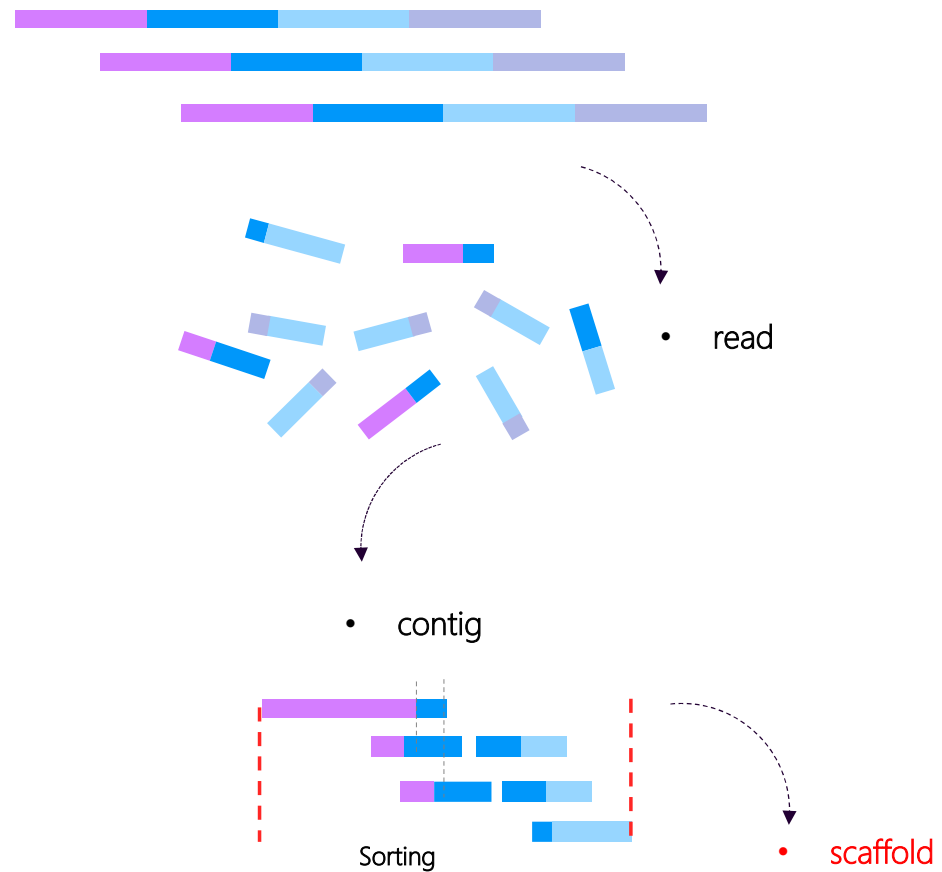
■ https://github.com/1081-Bioinformatics/finalproject-3ddna_108

Part 1 / Introduction

NGS 次世代 定序

把原本的DNA打成小段的碎片之後利用重複的頭尾關係，從較短的片段組合回長序列。而所謂的從頭組裝，是指不依靠既有的參考基因組 (Reference Genome) 直接從短片段組合出定序結果。

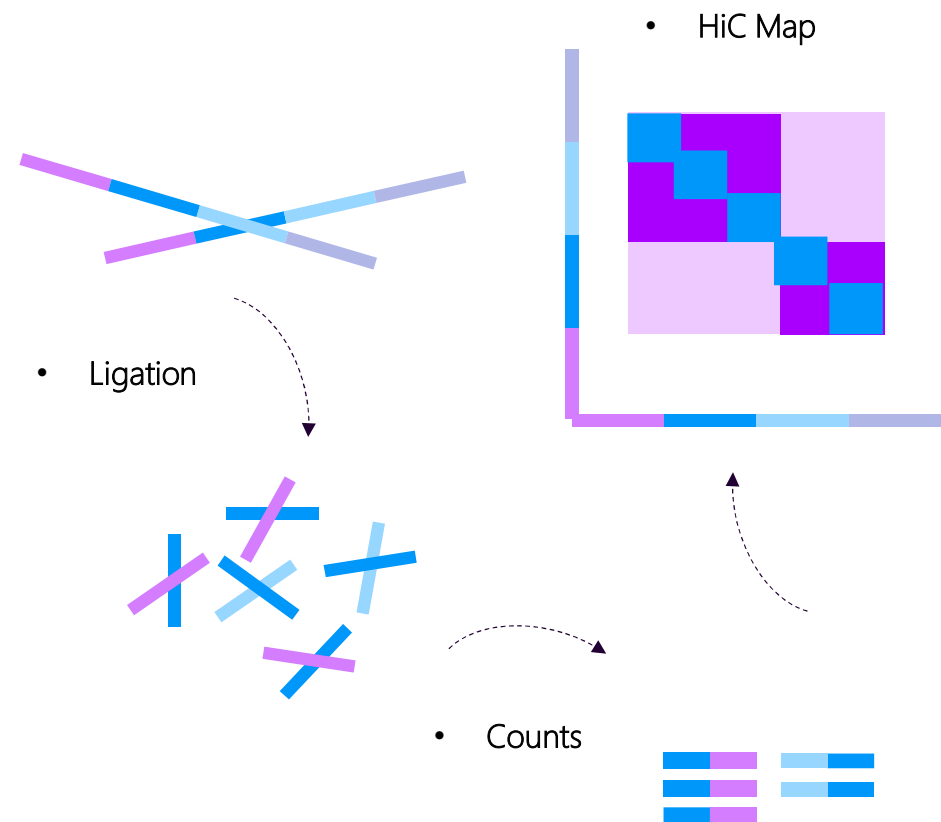
因為快速和便宜，雖然單一樣本的準確率不如第一代測序Sanger Method，但依靠大量的數據可以更方便的投入疫區有更多醫療和商業價值。



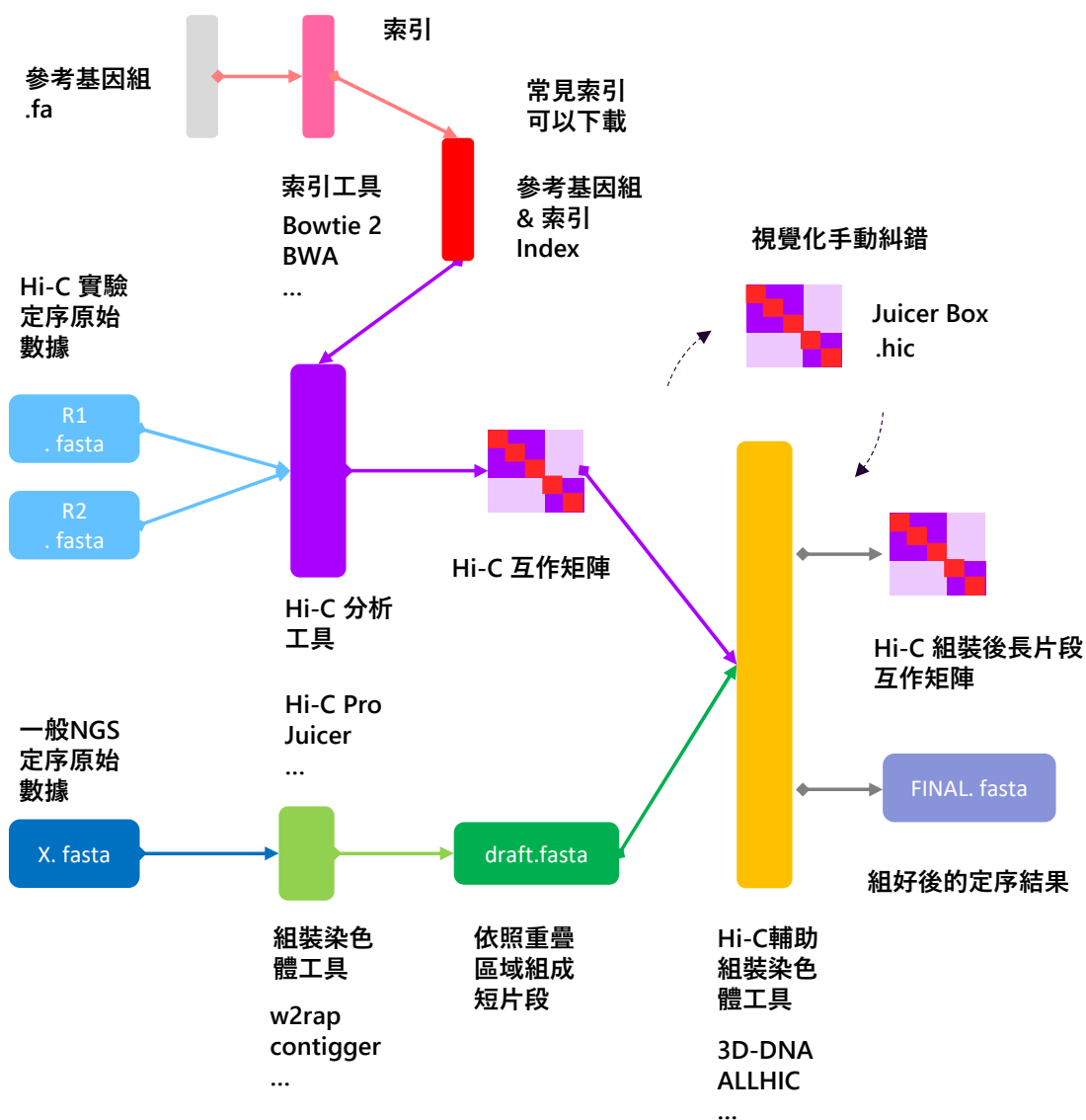
Hi-C 染色體結構 捕獲

主要是利用染色體之間存在實際上的折疊三維空間關係，利用交聯反應捕捉後，計算次數，最後得到能夠反應三維遠近關係的二維矩陣

額外多了3維的空間關係資訊除了去了解染色體的各種性質之外也可以利用較近的片段容易被交聯而數量較多，越遠則越少的關係，去輔助染色體定序的組裝排序。

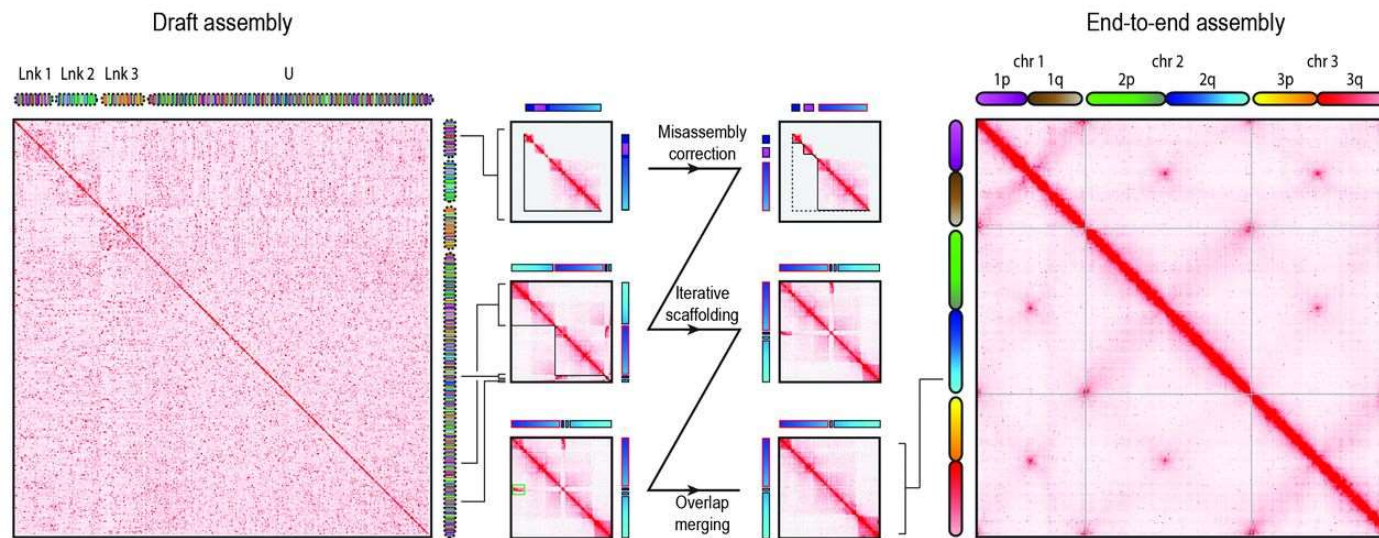


整體分析流程如右圖，本次專案覆
現實驗會從進入3D-DNA之後開始，
直接使用原論文提供的輸入前處理
數據



實驗目標設計

Fig. 1 Starting with a draft assembly, we used Hi-C data to correct misjoins, scaffold, and merge overlaps, thereby generating an assembly of the *Ae. aegypti* mosquito genome with chromosome-length scaffolds.



本論文利用組裝前，中，後的 Hi-C圖表，去展現出輔助組裝後長片段在互作圖的表現，會明顯清晰很多。利用這部分的現象，想要簡易的覆現這篇論文，可以把組裝前中後的圖表進行比較。

Part 2 / Environment Setup

3D-DNA 環境需求與 安裝

3D-DNA 需要的使用的程式包含了, Java Python , Shell 這三種語言。本專案安裝的部分是事先寫好每一步驟的安裝腳本，再逐一執行。

如果想要使用原始的3D-DNA進行實驗覆現，可以利用版控跳到第745779b 版本

[< /Source Code >](#)

- https://github.com/1081-Bioinformatics/finalproject-3ddna_108/tree/master/code/install_3ddna_dependence

Install Oracle Java 8

Using JDK version jdk-8u231-linux-x64 (change if you need for other hardware.)

```
sudo sh install_jdk8.sh
```

Install GNU Tools (gawk , coreutils , parallel)

```
sudo sh install_gnuutil.sh
```

Install Lastz

```
sudo source install_lastz.sh
```

Install Python & lib

Install 3D-DNA Package

```
git clone https://github.com/theaidenlab/3d-dna
cd 3d-dna
```

For the old stable version (This version using run-pipeline.sh to run.):

```
git checkout 745779b
```


AaegL2 & Human 實驗數據準備

	組裝草稿 (待組序列)	Hi-C 合併的片段數量統計 (輔助組裝 上游分析為 Juicer)
AaegL2 埃及斑蚊	GSE95797_AaegL2.fast a.gz	GSE95797_AaegL2.mnd.txt. gz
Human 人類	GSE95797_Hs1.fasta	GSE95797_Hs1.mnd.txt

[< /Source Code >](#)

- https://github.com/1081-Bioinformatics/finalproject-3ddna_108/tree/master/code/download_data

[< /Tips to Speed up>](#)

- `axel -n 30 ftp://...`

AaegL

Download AaegL2.mnd.txt (After unzip will be 76GB)

```
sudo sh download_GSE95797_AaegL2.mnd.txt.gz.sh
```

Download AaegL2.fasta

This code was provide by 3ddna project the same file in project [./supp/get-AaegL2.sh](#)

```
sudo sh download_GSE95797_AaegL2.fasta.gz.sh
```

Human

Download GSE95797_Hs1.mnd.txt

```
sudo sh download_GSE95797_Hs1.mnd.txt.gz.sh
```

Download GSE95797_Hs1.fasta

```
sudo sh download_GSE95797_Hs1.fasta.gz.sh
```

3D-DNA 程式除錯

$$\frac{y}{x} \longrightarrow \frac{(y+1)}{(x+1)}$$

執行舊版的時候有時候會遇到

fatal: division by zero attempted

為了順利進行分析，以拉普拉斯平滑修正了出錯的地方。

< /Source Code >

- https://github.com/1081-Bioinformatics/finalproject-3ddna_108/tree/master/code/fixbug_code

```
if (($1!=prev1 || $2!=prev2 || $3!=prev3) && FNR!=1)
{
    print prev1, prev2, prev3, (count+1)/(slength[prev1]*slength[prev2]+1), (count+1)
    count = 0
}
```

Part 3 / Run Analysis

3D-DNA 執行

```
$ chmod +x run-asm-pipeline.sh
```

```
$ nohup ./run-asm-pipeline.sh GSE95797_Hs1.fasta  
GSE95797_Hs1.mnd.txt > run_GSE95797_Hs1.log 2>&1 &
```

(新版)

[< /Source Code >](#)

- https://github.com/1081-Bioinformatics/finalproject-3ddna_108/tree/master/code/run_analysis

3D-DNA 運行情況

- Google Cloud Platform
n1-standard-16 (16 vCPUs, 60 GB memory, 1TB HDD)
- OS Version
Ubuntu 16.04.6 LTS (Xenial Xerus)
- Run both AaegL2&Human
- CPU Usage Plot ([The analysis process run 5 Days](#))



Part 4 / Output Result

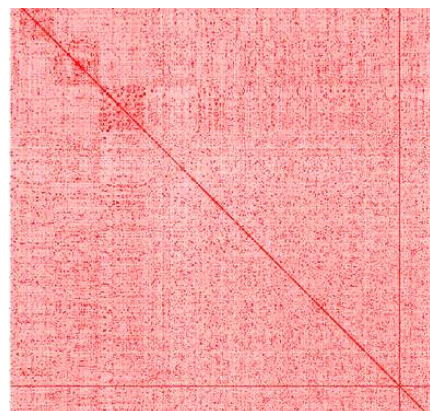
- Hi-C Figure
https://github.com/1081-Bioinformatics/finalproject-3ddna_108/tree/master/results
- .hic File
https://github.com/1081-Bioinformatics/finalproject-3ddna_108/tree/master/data

埃及斑蚊 Hi-C圖 變化

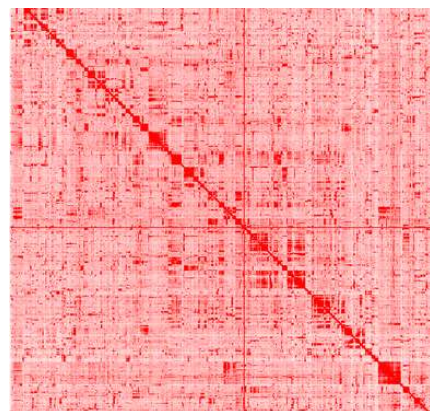
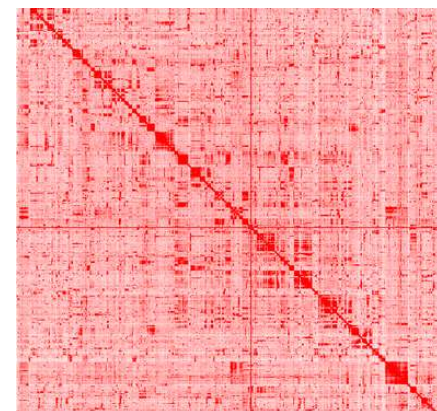
從左上，右上，左下，右下依序是從草稿Hi-C是視覺化圖，一直到最後一輪迭代的Hi-C圖。視覺圖有由本次專案輸出的.hic檔經Juicer Box 視覺化。

可以明顯的比較出來，隨著3D-DNA正確的參照Hi-C的資訊進行長片段組裝迭代，Hi-C圖有越來越清晰的傾向，表示隨著組裝序列越長，越有回復到原本的序列。

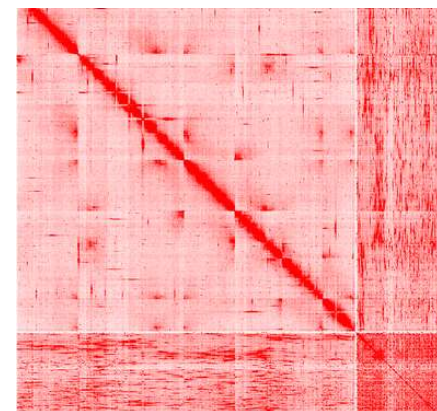
AaegL2.2.draft.hic



AaegL2.2.0.hic



AaegL2.2.1.hic

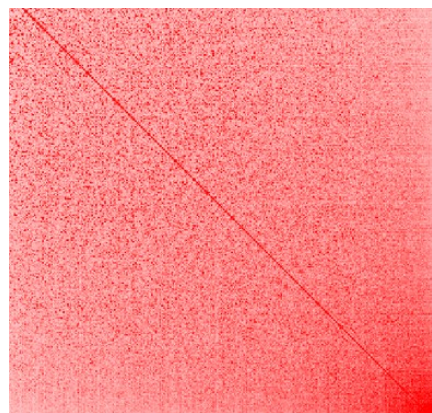


AaegL2.2.2.hic

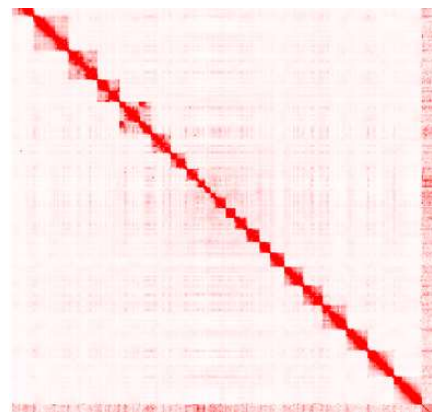
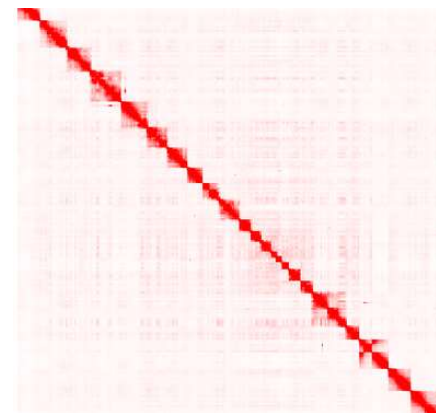
人類 Hi-C圖 變化

與前面一樣，迭代次數增加，Hi-C圖有越來越清晰的傾向，恢復的資訊越多。從以上兩組Hi-C圖的視覺化，就能夠清楚知道其作用。

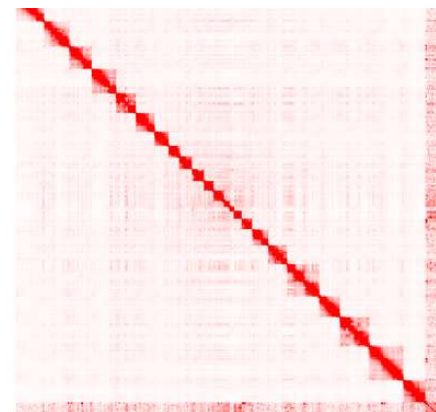
GSE95797_Hs1.draft.hic



GSE95797_Hs1.0.hic



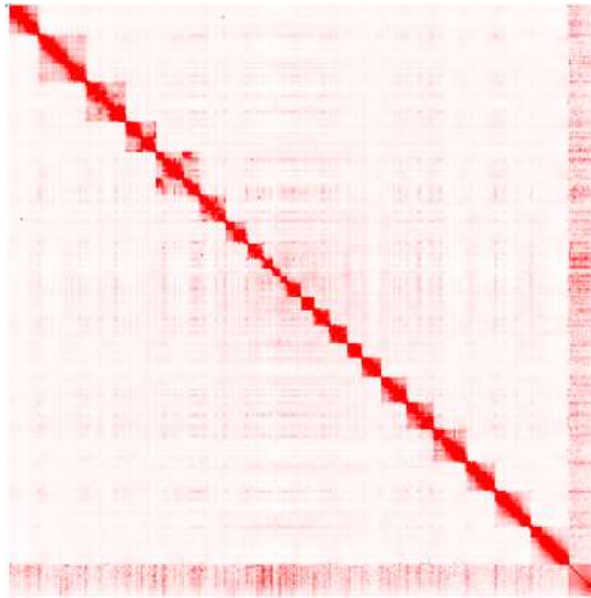
GSE95797_Hs1.1.hic



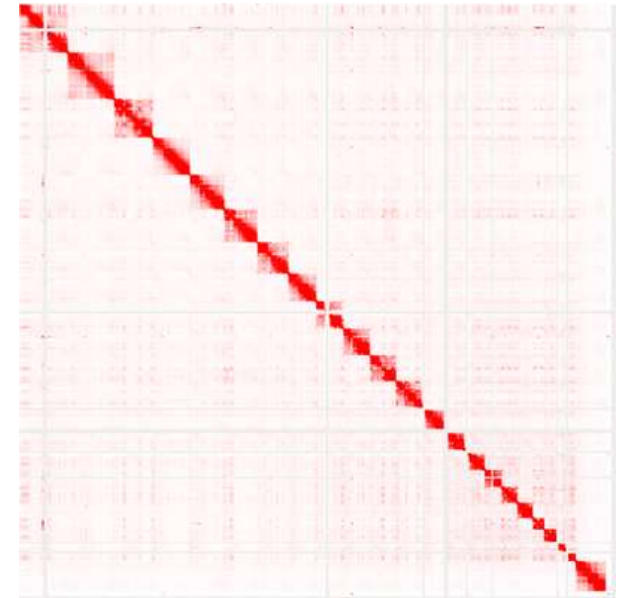
GSE95797_Hs1.2.hic

人類 Hi-C圖 比較

左邊是本論文的資料經過輔助組裝後的結果，右邊是2014年其他人實驗所繪製的Hi-C圖，兩者具有類似結構



- GSE95797_Hs1.2.hic



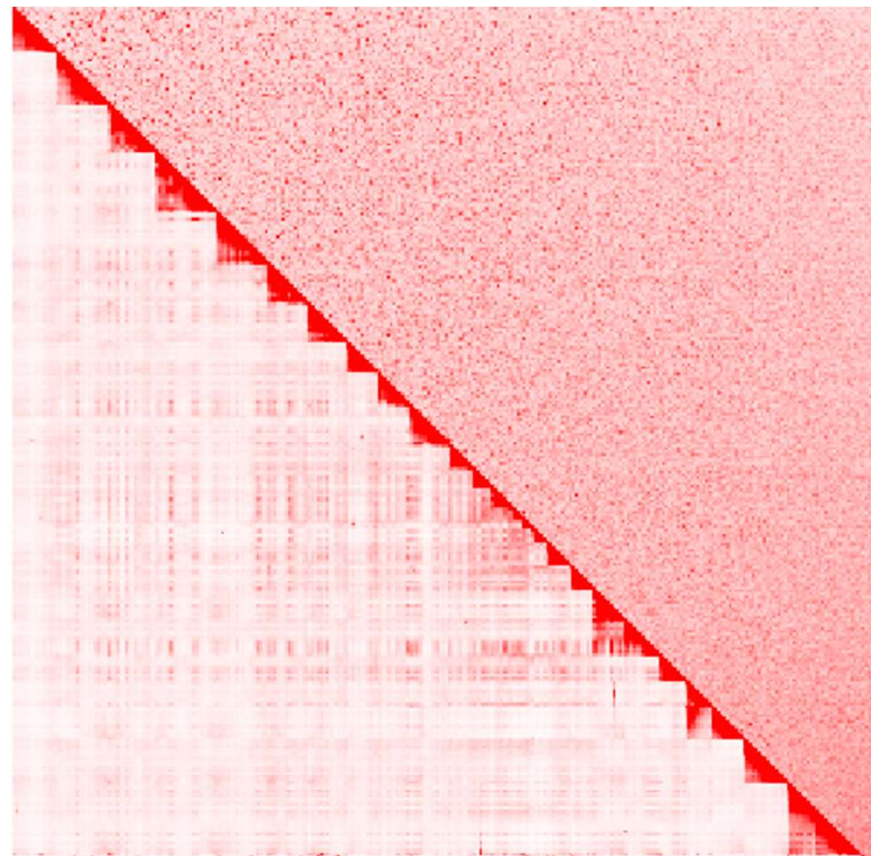
- GM12878_Cell_2014

(HIC001(148M) Cell. 2014 Dec
18;159(7):1665-80. doi:
10.1016/j.cell.2014.11.021. Epub 2014 Dec
11.)

人類 Hi-C圖 疊圖比較

左下半較清晰的是組裝後的圖，右
上半是尚未組裝的圖。

GSE95797_Hs1.draft.hic



GSE95797_Hs1.2.hic

結論

本專案是部分
<De novo assembly of the Aedes
aegypti genome using Hi-C yields
chromosome-length scaffolds> 論文再
現實驗。

並且得到了由Hi-C輔助組裝的序列。其
中，3D-DNA透過修正與排序疊代調整
組裝得長片段，右邊列出最後三點總結。

01

Hi-C能夠有效協助序列組裝，
3D-DNA組裝效果佳。

02

3D-DNA 計算效能消耗大。
如果需要在有限時間內完成，
需要預留更多時間。

03

3D-DNA 部分程式碼應該
可以更加完善

參考資料

1. THE CENTER FOR GENOME ARCHITECTURE Baylor College of Medicine & Rice University http://aidenlab.org/assembly/manual_180322.pdf
2. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. <https://www.ncbi.nlm.nih.gov/pubmed/28336562>
3. Hi-C辅助基因组安装 <https://www.jianshu.com/p/95792fbde9c3>



THANKS