BỘ GIÁO DỤC VÀ ĐÀO TẠO

VIỆN HÀN LÂM
KHOA HỌC VÀ CÔNG NGHỆ VN

HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

LÊ TÙNG LÂM

NGHIÊN CỨU ĐÁNH GIÁ MỘT SỐ PHƯƠNG PHÁP CHÚ GIẢI HỆ GEN LỤC LẠP

LUẬN VĂN THẠC SĨ HỆ THỐNG THÔNG TIN

Hà Nội, ngày 01/10/2023

BỘ GIÁO DỤC VÀ ĐÀO TẠO

VIỆN HÀN LÂM KHOA HỌC VÀ CÔNG NGHỆ VN

HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

LÊ TÙNG LÂM

NGHIÊN CỨU ĐÁNH GIÁ MỘT SỐ PHƯƠNG PHÁP CHÚ GIẢI HỆ GEN LỤC LẠP

LUẬN VĂN THẠC SĨ HỆ THỐNG THÔNG TIN

Mã số: 8 48 01 04

NGƯỜI HƯỚNG DẪN KHOA HỌC

1. TS. Nguyễn Thị Phương Thảo

Hà Nội, ngày 01/10/2023

LỜI CAM ĐOAN

Tôi xin cam đoan đề tài nghiên cứu trong luận văn này là công trình nghiên cứu của tôi dựa trên những tài liệu, số liệu do chính tôi tự tìm hiểu và nghiên cứu. Chính vì vậy, các kết quả nghiên cứu đảm bảo trung thực và khách quan nhất. Đồng thời, kết quả này chưa từng xuất hiện trong bất cứ một nghiên cứu nào. Các số liệu, kết quả nêu trong luận văn là trung thực nếu sai tôi hoàn chịu trách nhiệm trước pháp luật.

Hà Nội, ngày 22 tháng 12 năm 2023

Học viên thực hiện

Lê Tùng Lâm

LỜI CẢM ƠN

Đầu tiên em xin gửi lời cảm ơn đến TS. Nguyễn Thị Phương Thảo – giảng viên hướng dẫn đã tận tình giúp đỡ, hướng dẫn em hoàn thành tốt luận văn này.

Em cũng cảm ơn lãnh đạo/các đồng nghiệp Viện Công nghệ Sinh học, Trung tâm Giám định ADN và Phòng Tin sinh học đã giúp đỡ em về thiết bị phân tích và tạo điều kiện để em có thể hoàn thành khoá học và luận văn này.

Em cũng xin chân thành cảm ơn các thầy cô giáo, phòng Đào tạo tại Học Viện Khoa học Công nghệ đã tận tình chỉ bảo, tạo điều kiện cho em hoàn thành bài luận văn của mình. Qua đây, em cũng gửi lời cảm ơn tới gia đình, bạn bè đã động viên, khuyến khích và tạo điều kiện cho em trong suốt quá trình học tập cũng như trong quá trình làm luận văn.

Do còn hạn chế nhiều về kiến thức, kinh nghiệm và thời gian tìm hiểu nên luận văn chắc chắn còn nhiều thiếu sót. Em rất mong sẽ nhận được nhiều đóng góp của thầy, cô để có thể hoàn thiện hơn bài luận văn này. Và em cũng hy vọng rằng đây sẽ là tài liệu bổ ích cho những người quan tâm về lĩnh vực này, mọi chi tiết cần điều chỉnh, bổ sung xin liên hệ tới letunglam1991@gmail.com.

Em xin chân thành cảm ơn!

Hà Nội, ngày 22 tháng 12 năm 2023

Học viên thực hiên

Lê Tùng Lâm

MỤC LỤC

DANH MỤC BẢNG BIỂU

DANH MỤC HÌNH VẼ

MỞ ĐẦU

Ngày nay, nhờ sự phát triển của công nghệ giải trình tự gen, việc giải trình tự toàn bộ hệ gen không còn khó khăn nữa. Đặc biệt với những hệ gen nhỏ như lục lạp thì việc giải trình tự, lắp ráp, chú giải hệ gen lục lạp trở nên tương đối dễ dàng. Tuy nhiên, như đã biết trên hệ thống ngân hàng gen NCBI vẫn còn rất nhiều hệ gen lục lạp được lắp ráp, chú giải sai sót mặc dù đó là những hệ gen đã được nghiên cứu kỹ lưỡng. Một số lỗi phổ biến như: gen bị cắt ngắn, thêm vào những phần mở rộng không mong muốn của các exon, bỏ sót các gen đã biết, lựa chọn sai các chuỗi mã hoá, các khung đọc mở được giả định là gen chức năng… Việc chú giải gen chức năng của lục lạp rất quan trọng, việc này giúp ích cho các nhà nghiên cứu về phân loài có thể áp dụng để phân loại chính xác các cây thực vật gần gũi trong cùng chi, họ; việc chú giải sai có thể dẫn đến một hệ quả domino khi những người nghiên cứu sau sử dụng những kết quả chưa chính xác này cho những nghiên cứu của mình. Tính đến thời điểm hiện tại chưa có công cụ chú giải hệ gen lục lạp nào có ưu thế và chưa có bước tiến lớn nào trong việc nâng cao thuật toán chú giải hệ gen lục lạp vì số lượng hạn chế các nhà khoa học về khoa học máy tính, thuật toán tin sinh học phát triển những thuật toán mới cho việc này.

Tính đến nay chỉ có một số công cụ hỗ trợ chú giải lục lạp như : Dual Organellar GenoMe Annotator (DOGMA); Chloroplast Genome Annotation, Visualization, Analysis, and GenBank Submission (CPGAVAS & CPGAVAS2) ; GeSeq ;Verdant. Tuy nhiên, chúng đều có những ưu điểm và khuyết điểm riêng. Việc khảo sát, đánh giá những phần mềm này có ý nghĩa quan trọng nhằm nâng cao chất lượng chú giải gen chức năng trong hệ gen lục lạp. Tiến tới việc đề xuất những thuật toán mới hiệu quả hơn thuật toán cũ.

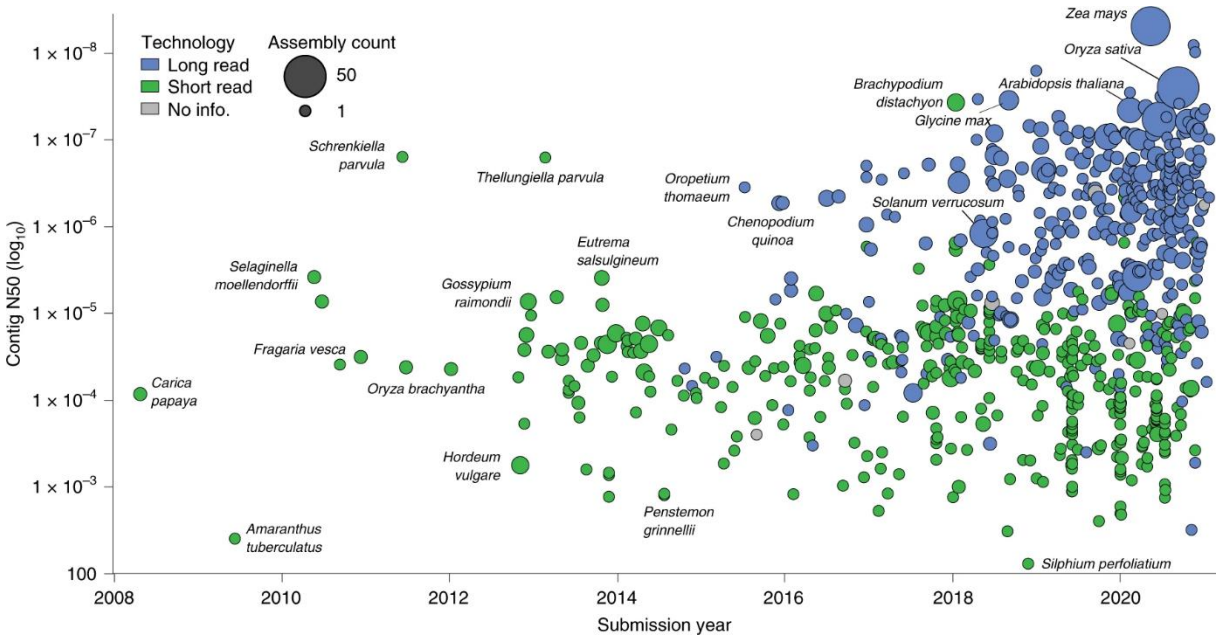
Khóa luận được bố cục như sau:

- Chương 1: Tổng quan về các nghiên cứu liên quan hệ gen thực vật nói chung, lục lạp nói riêng, tổng quan về quy trình phân tích hệ gen lục lạp đặc biệt là lắp ráp và chú giải hệ gen lục lạp
- Chương 2: Nghiên cứu về các phương pháp phân tích chú giải hệ gen lục lạp, tiêu biểu là 3 phương pháp CPGAVAS2, Geseq và PGA
- Chương 3: Lựa chọn các dữ liệu đầu vào, cài đặt các công cụ cần thiết và thực hiện so sánh các phương pháp.
- Chương 4: KẾT LUẬN

1 CHƯƠNG 1: TỔNG QUAN LÝ THUYẾT

1.1 *Tổng quan tình hình nghiên cứu hệ gen thực vật*

Trên thế giới các nghiên cứu về hệ gen học đã bắt đầu phát triển từ những năm cuối thế kỷ 20 khi có sự ra đời của các thiết bị giải trình tự thế hệ thứ nhất, điển hình là việc nghiên cứu và thành lập hệ gen người tham chiếu bắt đầu từ năm 1990, hoàn thành vào năm 2003[1]. Tiếp sau đó là sự ra đời của công nghệ giải trình tự thế hệ mới những năm đầu thế kỷ 21 đã thúc đẩy sự phát triển của nhánh nghiên cứu hệ gen học. Đối với thực vật nói riêng những nghiên cứu đầu tiên về hệ gen của loài cây mô hình *Arabidopsis thaliana* khi sử dụng dữ liệu giải trình tự thế hệ mới đầu thế kỷ 21 là nghiên cứu tiền đề cho việc phát triển hệ gen học và tiến hoá thực vật.[2], [3] Đến năm 2008 rất nhiều hệ gen thực vật khác nhau được công bố và đề cập đến trong nghiên cứu của tác giả Gupta. [4]. Trong những năm gần đây khi có sự phát triển vũ bão của công nghệ giải trình tự thế hệ mới đặc biệt là công nghệ giải trình tự thế hệ thứ 3 và thứ 4, các công bố liên quan tới hệ gen thực vật ngày càng gia tăng. Trong khoảng 20 năm phát



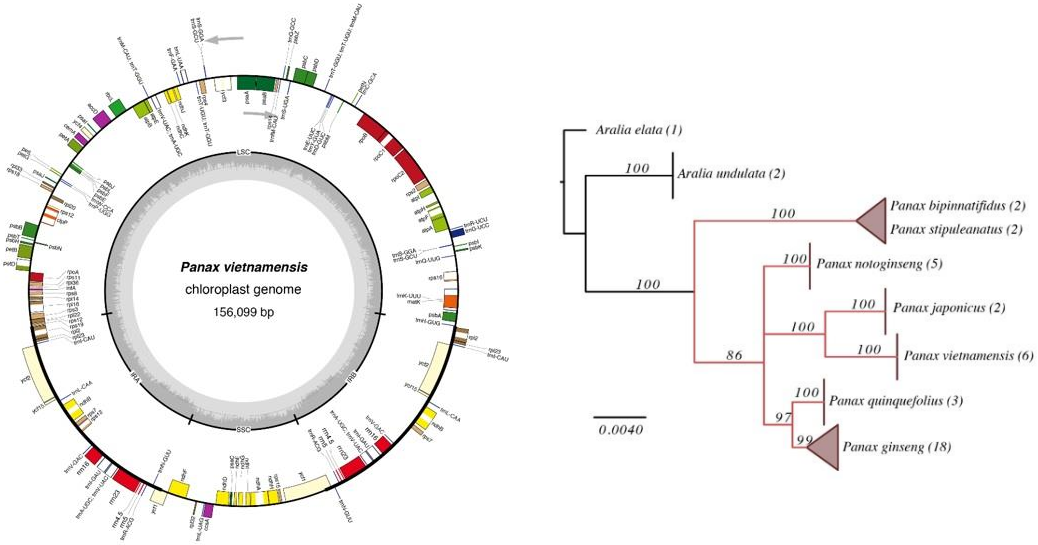
Ảnh 1-1: Thống kê về số lượng hệ gen thực vật được công bố trong 20 năm qua [5]

triển của công nghệ giải trình tự đã có trên 1000 loài thực vật được giải trình tự gen. Cung cấp một lượng thông tin khổng lồ và hữu ích cho những nhà nghiên cứu về thực vật học cũng như tiến hoá.[5], [6]

Nhìn vào hình 1-1 chúng ta có thể thấy rằng sự số lượng hệ gen thực vật được công bố tăng mạnh từ những năm 2014 khi Illumina ra mắt những hệ máy thông lượng cao của họ HiSeq, NovaSeq. Đặc biệt từ 2016 có sự tham gia của những hãng giải trình tự đoạn dài giúp gia tăng số lượng và chất lượng của hệ gen thực vật.

Những đóng góp về hệ gen thực vật đã giúp các nhà phân loại thực vật phân loài chính xác các loài về đúng nhánh của chúng. Năm 2011, chi *Psilanthus* có quan hệ gần gũi đã được gộp vào *Coffea*. Tuy nhiên, kết quả thu được vào năm 2017 - dựa trên 28.800 SNP - chỉ ra rằng không có hỗ trợ phát sinh gen đáng kể cho sự hợp nhất này.[7] Thêm vào đó những nghiên cứu về gen trong công bố của Yves Bawin năm 2021 chỉ ra rằng *Coffea canephora* và *C. eugenioides* đã được xác nhận là loài tổ tiên giả định của *C. arabica*. Những loài này rất có thể đã được lai tạo từ khoảng 1,08 triệu đến 543 000 năm trước, trùng với các thời kỳ biến động môi trường, có thể gây ra sự thay đổi phạm vi của các loài tổ tiên tạo điều kiện cho sự xuất hiện của *C. arabica*.[8]

Ở Việt Nam cũng có những nghiên cứu về hệ gen thực vật nói chung là lục lạp nói riêng giúp ích cho việc phân loài, chọn giống, bảo tồn những loài



Ảnh 1-2: Kết quả giải trình tự lục lạp sâm ngọc linh và phân loài sâm ngọc linh trong nghiên cứu của GS. Nông Văn Hải và các cộng sự

thực vật quý hiếm. Điển hình như nghiên cứu về lục lạp của loài sâm ngọc linh đặc hữu của Việt Nam của Gs. Nông Văn Hải và các công sự. Trong nghiên cứu này nhóm nghiên cứu đã tìm kiếm được 4 chỉ thị có tiềm năng làm mã vạch phân tử cho phân loại sâm Ngọc Linh và các loài khác thuộc chi Nhân sâm. [9]–[11]

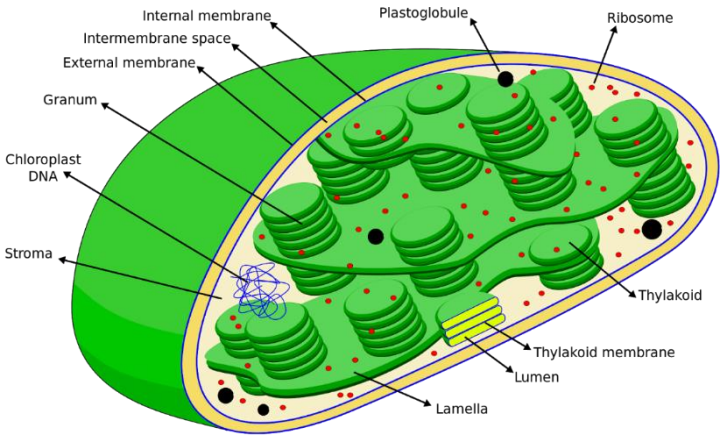
Như vậy, vai trò của việc nghiên cứu hệ gen thực vật nói chung và hệ gen lục lạp nói riêng là rất quan trọng. Tuy nhiên, hiện nay phương pháp phân tích hệ gen lục lạp có rất nhiều. Và chưa có nhiều nghiên cứu về việc so sánh, đánh giá những phương pháp này. Vì vậy, việc tiến hành so sánh đánh giá các phương pháp này là rất cần thiết.

1.2 *Cấu trúc hệ gen lục lạp và ý nghĩa khoa học*

Lục lạp là một đơn vị chức năng trong tế bào và đóng vai trò là bào quan quang hợp chỉ có ở thực vật và tảo. Nhờ có lục lạp mà thực vật, tảo có thể chuyển hóa năng lượng ánh sáng thành lượng tích trong chất hữu cơ. Ở thực vật, lục lạp có trong các bộ phận xanh của cây, trong đó có nhiều nhất là ở lá. Người đầu tiên phát hiện ra lục lạp là Julius von Sachs (1832–1897) - một nhà thực vật học và tác giả của nhiều cuốn sách giáo khoa cơ bản.

Lục lạp cũng có cấu trúc màng hai lớp với màng ngoài rất dễ thấm còn màng trong thấm rất ít và ở giữa 2 lớp màng này có một khoang giữa màng. Màng trong bao bọc một vùng không có màu xanh lục, được gọi là Stroma. Stroma là nơi diễn ra các phản ứng của pha tối và nó giống như chất nền matrix của ty thể, có chứa các enzyme, ARN, AND và các ribosome. Các ribosome là các hạt hình cầu có kích thước 15 - 20 cm. Nó ở trong chất nền cùng với các hạt tinh bột với kích thước khác nhau.

Trong lục lạp có chứa đến 80% loại protein không hòa tan có liên kết với lipit ở dạng lipoprotein. Clorophyl là một trong những thành phần thuộc hệ sắc tố quang hợp của lục lạp, bao gồm diệp lục a và diệp lục b . Các phân tử clorophyl có cấu trúc không đối xứng gồm một đầu ưa nước được do 4 vòng pirol xếp xung quanh nguyên tử magie tạo thành và một đuôi dài là mạch kị nước.



Ảnh 1-3: Cấu tạo của lục lạp

Bên cạnh Corophyl, Caroic cũng là những sắc tố khác màu có trong lục lạp, tuy nhiên, nó thường bị màu lục của clorophyl che lấp. Chúng chỉ có cơ hội xuất hiện vào mùa thu, thời điểm mà lượng Clorophyl bị sụt giảm đi khá nhiều. Ở tảo và thực vật thủy sinh thì sắc tố quang hợp là Phicobilin. Đây là nhóm sắc tố đóng vai trò quan trọng trong việc hấp thụ ánh sáng lục (550 nm) và vàng (612 nm) trong ánh sáng mặt trời.

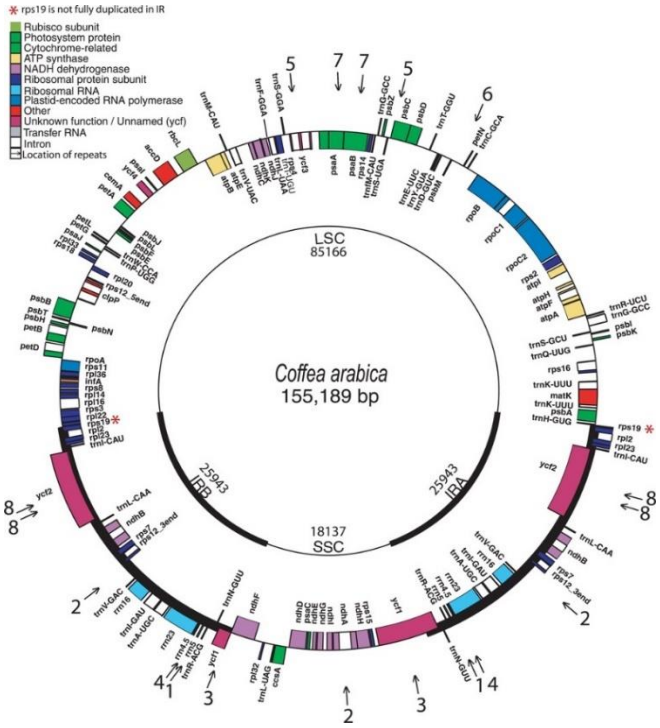
Ngoài ra, trong lục lạp cũng có chứa axit nucleic, ARN (hàm lượng từ 2 - 4 % khối lượng khô), ADN (0,2 - 0,5% khối lượng khô), các chất truyền năng lượng, enzim, NADP, cytocrom, plastokinon, reductasa, atp-sintetase, plastoxiamin, ferredonxin và các enzim của chu trình calvin.

Bảng các thành phần hóa học của lục lạp:

Chất	**Hàm lượng %**	**Các cấu thành**
Protein	35 - 55	80% không hòa tan
Lipit	20 - 30	Mỡ 50%, colin 46%, sterin 20%, sáp 16%, photphatit 2-7%, etanolamin 8%
Gluxit	Thay đổi	Tinh bột, đường có photphat
Clorophyl	9	Clorophyl α 75%, Clorophyl β75%
Carotinoit	4.5	Xantophyl 75%, carotin 25%
ARN	2 - 4	
ADN	**0.2 - 0.5**	

Mặc dù chỉ chiếm 0.2 - 0.5% thành phần của lục lạp nhưng bộ gen lục lạp lại có ý nghĩa rất lớn trong việc nghiên cứu tiến hoá và di truyền. Hệ gene lục lạp nói chung là hệ gen lục lạp của cà phê nói riêng có cấu trúc là hệ DNA dạng vòng gồm 4 phần: vùng sao chép đơn dài (LSC - long single copy section), vùng sao chép đơn ngắn (SSC - short single copy section) và 2 vùng lặp lại đảo ngược IRA và IRB. Các đoạn lặp đảo ngược có độ dài rất khác nhau, mỗi đoạn dài từ 4.000 đến 25.000 cặp bazơ.[11] Sự lặp lại nghịch đảo ở thực vật có xu hướng ở giới hạn trên của phạm vi này, mỗi lần lặp lại có chiều dài 20.000–25.000 cặp bazơ.[9] [13] Các vùng lặp đảo ngược thường chứa ba RNA ribosome và hai gen tRNA, nhưng chúng có thể được mở rộng hoặc thu nhỏ để chứa ít nhất bốn hoặc nhiều nhất là trên 150 gen.

Bộ gen lục lạp của cà phê là một phân tử DNA hình tròn có kích thước 155 189 bp với cấu trúc bốn phần đặc trưng của phần lớn các nhiễm sắc thể lục lạp thực vật trên cạn. Nó bao gồm hai vùng lặp lại đảo ngược (IRa và IRb) 25 943 bp được phân tách bằng các vùng sao chép đơn lớn (LSC) và nhỏ (SSC) lần lượt là 85 166 và 18 137 bp. Tỷ lệ các chuỗi protein, RNA vận chuyển (tRNA), RNA ribosome (rRNA), trình tự intron và liên gen lần lượt là 51%, 2%, 6%, 9% và 32%. Trong số 130 gen có trong bộ gen, 112 gen hiện diện dưới dạng một bản sao duy nhất và 18 gen được sao chép trong IR. Vùng mã hóa bao gồm 79 gen protein, 29 tRNA và 4 rRNA. Bộ gen lục lạp cà phê có 59,35% trình tự mã hóa, trong đó 51,76% mã hóa cho protein. Mười tám gen chứa intron, 15 gen có hai exon và ba gen có ba exon. Mười hai gen mã hóa protein và sáu tRNA có intron. Một phần gen rps19 được nhân đôi tại ranh giới IRA–LSC do sự mở rộng của IR. Sự sao chép tương tự các phần của rps19 xảy ra ở tất cả các thành viên của họ Solanaceae ngoại trừ thuốc lá. Ngoài ra, trong trường hợp cà phê, các nhà khoa học quan sát thấy gen infA còn nguyên vẹn, trong khi đó nó là gen giả ở



Ảnh 1-4: Cấu trúc hệ gen lục lạp loài cà phê arabica

thuốc lá và ở hầu hết các thành viên khác của họ Solanaceae. Hàm lượng AT và GC của bộ gen lục lạp cà phê lần lượt là 63% và 37%, rất giống với hàm lượng của lúa, ngô, cam quýt, bông và thuốc lá.[12]

1.3 *Công nghệ giải trình tự NGS và dữ liệu giải trình tự NGS*

Thuật ngữ "Next generation sequencing - giải trình tự thế hệ tiếp theo" thể hiện rằng công nghệ giải trình tự đã bước sang một giai đoạn mới, công nghệ mới, đột phá về công suất, giá thành cũng như chất lượng giải trình tự. Hiện nay, NGS đã có đến thế hệ thứ 4. Thế hệ thứ 2 là thế hệ giải trình tự đoạn ngắn của các hãng như: Illumina, MGI, Genemind, Ion Torrent… Thế hệ thứ 3 là thế hệ giải trình tự đoạn dài bằng công nghệ SMRT sequencing – giải trình tự thời gian thực của hãng Pacbio và hãng Oxford Nanopore công bố rằng họ là thế hệ giải trình tự thế hệ thứ 4 – công nghệ giúp giải trình tự được những đoạn trình tự Ultra-longread trong những thiết bị nhỏ gọn, linh hoạt và thời gian nhanh.[13]

Các phương pháp giải trình tự thế hệ thứ hai có thể được chia thành hai loại chính, giải trình tự bằng phương pháp lai và giải trình tự bằng phương pháp tổng hợp (SBS). Phương pháp SBS còn xa hơn nữa là công nghệ giải trình tự Sanger, không có đầu cuối dideoxy, kết hợp với các chu kỳ tổng hợp, hình ảnh và phương pháp lặp đi lặp lại để kết hợp các nucleotide bổ sung trong chuỗi ngày càng tăng. Nếu chỉ đánh giá sơ qua thì có thể nghĩ rằng những phương pháp mới này có chi phí đắt đỏ, nhưng thực ra những phản ứng giải trình tự được chạy song song hàng trăm nghìn phản ứng cùng một lúc, ở các thể tích nanoliter, picoliter hoặc zeptoliter trong các con con chip/flow-cell nhỏ; do đó chi phí cho mỗi nucleotide là rất thấp. Các công nghệ được cải tiến liên tục, cho độ chính xác lớn hơn, đoạn đọc dài hơn, thu nhỏ kích thước chip giải trình tự, tăng mật độ trên mỗi diện tích chip vì vậy chi phí giải trình tự đang hơn nữa.

Bảng 1-1: Bảng so sánh các công nghệ giải trình tự phổ biến hiện nay

STT	Nền tảng	Thế hệ	Nguyên lý	Kich thước đoạn đọc (bp)	Công suất tối đa	Ref.
1	Ion Torrent	Thế hệ thứ hai	Nguyên lý giải trình tự bán dẫn ion phát hiện ion H+ được tạo ra trong quá trình kết hợp nucleotide.	200–400	50 Gb	[14], [15]
2	Illumina	Thế hệ thứ hai	Giải trình tự pha rắn trên bề mặt cố định tận dụng sự hình thành mảng vô tính bằng cách sử dụng công nghệ kết thúc có thể đảo ngược độc quyền để giải trình tự quy mô lớn nhanh chóng và chính xác bằng cách sử dụng các dNTP có nhãn đơn, được thêm vào chuỗi axit nucleic.	36–300	6000 Gb	[14], [15]
3	DNA nanoball sequencing	Thế hệ thứ hai	Phép lai oligo nẹp với khuếch đại sau PCR từ các thư viện giúp hình thành các vòng tròn. ssDNA hình tròn này hoạt động như mẫu DNA để tạo ra một chuỗi DNA dài tự lắp ráp thành một quả cầu nano DNA chặt chẽ. Chúng được thêm vào tế bào dòng được phủ aminosilane (tích điện dương) để cho phép liên kết theo khuôn mẫu của các hạt nano DNA. Các bazơ được gắn thẻ huỳnh quang được tích hợp vào chuỗi DNA và việc giải phóng thẻ huỳnh quang được ghi lại bằng kỹ thuật hình ảnh.	50–150	6000 Gb	[24,25]
4	PacBio Onso system	Thế hệ thứ hai	Hóa học giải trình tự bằng liên kết (SBB) sử dụng các nucleotide tự nhiên, sự kết hợp không có sẹo trong các điều kiện tối ưu hóa	100–200		

			để liên kết và mở rộng. (https://www.pacb.com/technology/sequencing-by-bind/, truy cập vào ngày 1 tháng 9 năm 2023).			
5	Single-molecule real-time sequencing (SMRT)	Thế hệ thứ ba	Các đoạn DNA dài được định vị trong các giếng nơi DNA polymerase có quá trình xử lý cao được gắn trước. Các giếng được tiếp xúc với các nucleotide có nhãn huỳnh quang, khi kết hợp sẽ phát ra tín hiệu huỳnh quang. Hệ thống phát hiện quang học được lập trình để thu tín hiệu và phân tử nhanh chóng khuếch tán.	average 10,000–16,000	66.5Gb	[15], [16]
6	Nanopore DNA sequencing	Thế hệ thứ "tư"	Phương pháp này dựa vào sự tuyến tính hóa của các phân tử DNA hoặc RNA và khả năng di chuyển của chúng qua một lỗ sinh học gọi là "lỗ nano", có chiều rộng 8 nanomet. Tính di động điện di cho phép chuỗi axit nucleic tuyến tính đi qua, từ đó có khả năng tạo ra tín hiệu dòng điện.	average 10,000–30,000	14Tb	[14], [15], [17]

1.4 *Các định dạng file thường gặp trong khi xử lý dữ liệu hệ gen lục lạp*

Công nghệ giải trình tự ngày càng phát triển, dữ liệu giải trình tự ngày càng được tạo ra với số lượng lớn, trong thời gian ngắn, độ chính xác cao. Do đó, vai trò của ngành Công nghệ thông tin nói chung, tin sinh học nói riêng ngày càng quan trọng. Để có thể lưu trữ, xử lý được lượng dữ liệu khổng lồ từ các hệ thống giải trình tự là không đơn giản. Với dạng dữ liệu từ máy giải trình tự xuất ra thông thường sẽ là dạng dữ liệu văn bản có cấu trúc: bam/fastq/fasta và một số dạng file log. Trong file dữ liệu có chứa các thông tin cơ bản như: thiết bị giải trình tự, thời gian giải trình tự, trình tự đoạn đọc, chất lượng của từng đoạn đọc, toạ độ vị trí của đoạn đọc được tổng hợp trên chip giải trình tự….

1.4.1 *Fastq – file trình tự chứa thông tin chất lượng trình tự*

Theo định nghĩa: Định dạng FASTQ là định dạng dựa trên văn bản để lưu trữ các trình tự sinh học (thường là trình tự nucleotide) và điểm chất lượng tương ứng của nó. Cả ký tự thứ tự và điểm chất lượng đều được mã hóa bằng một ký tự ASCII duy nhất để ngắn gọn.[18]

Ban đầu nó được phát triển tại Viện Wellcome Trust Sanger để kết hợp trình tự được định dạng FASTA và dữ liệu chất lượng của nó, nhưng gần đây đã trở thành tiêu chuẩn trên thực tế để lưu trữ đầu ra của các công cụ giải trình tự thông lượng cao như Máy phân tích bộ gen Illumina.

Một tệp FASTQ cơ bản có bốn trường được phân tách bằng dòng trên mỗi chuỗi:

- Trường 1: bắt đầu bằng ký tự '@' và theo sau là mã định danh trình tự và mô tả tùy chọn (như dòng tiêu đề FASTA).
- Trường 2: là trình tự của đoạn đọc thô.
- Trường 3: bắt đầu bằng ký tự '+' và được theo sau tùy ý bởi cùng một mã định danh trình tự (và bất kỳ mô tả nào).

- Trường 4: mã hóa các giá trị chất lượng cho chuỗi trong Trường 2 và phải

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

chứa cùng số ký hiệu như các chữ cái trong chuỗi.

Ảnh 1-5: Mô tả định dạng file fastq điển hình

1.4.2 *Fasta – file chứa dữ liệu trình tự*

Định dạng FASTA là định dạng dựa trên văn bản để biểu thị trình tự nucleotide hoặc trình tự peptide, trong đó các cặp bazơ hoặc axit amin được biểu thị bằng mã một chữ cái. Trình tự ở định dạng FASTA bắt đầu bằng mô tả một dòng, theo sau là dòng dữ liệu trình tự. Dòng mô tả được phân biệt với dữ liệu trình tự bằng ký hiệu lớn hơn (">") ở cột đầu tiên. Khuyến nghị rằng tất cả các dòng văn bản có độ dài ngắn hơn 80 ký tự.

Ví dụ của 1 file fasta

```
>NC_008535.1 Coffea arabica chloroplast, complete genome
TGGGCGAACGACGGGAATTGAACCCGCGCATGGTGGATTCACAATCCACTGCCTTGATCCACTTGGCTAC
ATCCGCCCCTCTACTCTATTTTTATATTTTTTTATTTCATATTCGAACAATTTCTTTACTTTTCTTTAAA
TCTTTAAAATTAAAAAAAAAACATCTATCTATATTTAAGTACAATTACTACTAAAATAACCAAATAAAAA
AATAAATAAAGGAGCAATAAGACCCTCTTATCTTAAGAGAATAAGAAGGAAATTATTGCTCCTTTATTTT
TCAATAACTCTTATACAATAAGACTAACGTCTTATCCATTTACAGATGGAGCATCTATAGCAGCTAGGTC
TAGAGGGAAGTTATGAGCATTACGTTCATGCATAACTTCCATACCAAGGTTAGCGCGGTTAATGATATCC
GCCCAAGTATTAATTACACGACCTTGACTATCAACTACAGATTGGTTGAAATTAAACCCGTTTAGGTTGA
```

Các trình tự dự kiến sẽ được thể hiện trong mã axit amin và axit nucleic IUB/IUPAC tiêu chuẩn, với những ngoại lệ sau:

- Chữ cái viết thường được chấp nhận và được ánh xạ thành chữ hoa;
- một dấu gạch ngang có thể được sử dụng để biểu thị vị trí bị gaps – vị trí không có trính tự xác định - ;
- trong trình tự axit amin, U và * là các chữ cái được chấp nhận (xem bên dưới).

- bất kỳ chữ số nào trong chuỗi truy vấn phải được loại bỏ hoặc thay thế bằng mã chữ cái thích hợp (ví dụ: N cho dư lượng axit nucleic chưa biết hoặc X cho dư lượng axit amin chưa xác định).

1.4.3 *Genbank file (.gb, .gbk)*

Genbank file là một định dạng file text được giới thiệu bởi NCBI. Nhằm mục đích để người dùng có thể up load thông tin trình tự gen lên trên ngân hàng Genbank. Cấu trúc file genbank gồm rất nhiều trường thông tin, Bảng 1-2 tóm tắt một số trường thông tin hay dùng như sau:

Bảng 1-2: Danh sách các trường thông tin trong cấu trúc file genbank (.gb, .gbk)[19]

Locus Name	Tên locus ban đầu được thiết kế để giúp nhóm các mục có trình tự tương tự: ba ký tự đầu tiên thường được chỉ định sinh vật; ký tự thứ tư và thứ năm được sử dụng để hiển thị các ký hiệu nhóm khác, chẳng hạn như sản phẩm gen; đối với các mục được phân đoạn, ký tự cuối cùng là một trong chuỗi các số nguyên tuần tự.
Sequence Length	Số cặp bazơ nucleotide (hoặc dư lượng axit amin) trong bản ghi trình tự. Trong ví dụ này, độ dài chuỗi là 5028 bp. Không có giới hạn tối đa về kích thước của trình tự có thể được gửi tới GenBank. Bạn có thể gửi toàn bộ bộ gen nếu bạn có một đoạn trình tự liền kề từ một loại phân tử.
Molecule Type	Loại phân tử được giải trình tự. Trong ví dụ này, loại phân tử là DNA. Mỗi bản ghi GenBank phải chứa dữ liệu trình tự liền kề từ một loại phân tử đơn lẻ. Các loại phân tử khác nhau được mô tả trong tài liệu về Sequin và có thể bao gồm DNA bộ gen, RNA bộ gen, RNA tiền thân, mRNA (cDNA), RNA ribosome, RNA chuyển, RNA hạt nhân nhỏ và RNA tế bào chất nhỏ.
Modification Date	Ngày trong trường LOCUS là ngày sửa đổi lần cuối. Bản ghi mẫu hiển thị ở đây được sửa đổi lần cuối vào ngày 21 tháng 6 năm 1999.
DEFINITION	Mô tả ngắn gọn về trình tự; bao gồm thông tin như sinh vật nguồn, tên gen/tên protein hoặc một số mô tả về chức năng của

	trình tự (nếu trình tự không mã hóa). Nếu trình tự có vùng mã hóa (CDS), phần mô tả có thể được theo sau bởi từ hạn định tính đầy đủ, chẳng hạn như "các đĩa CD hoàn chỉnh".
ACCESSION	Mã định danh duy nhất cho bản ghi trình tự. Số gia nhập áp dụng cho bản ghi hoàn chỉnh và thường là sự kết hợp của (các) chữ cái và số, chẳng hạn như một chữ cái theo sau là năm chữ số (ví dụ: U12345) hoặc hai chữ cái theo sau là sáu chữ số (ví dụ: AF123456). Một số phần bổ sung có thể dài hơn, tùy thuộc vào loại bản ghi trình tự. Số gia nhập không thay đổi ngay cả khi thông tin trong hồ sơ được thay đổi theo yêu cầu của tác giả.
GI	Trong trường hợp này, số nhận dạng trình tự "GenInfo Identifier" dành cho trình tự nucleotide. Nếu một chuỗi thay đổi theo bất kỳ cách nào thì số GI mới sẽ được gán.
Organism	Tên khoa học chính thức của sinh vật nguồn (chi và loài, nếu phù hợp) và dòng dõi của nó, dựa trên sơ đồ phân loại phát sinh gen được sử dụng trong Cơ sở dữ liệu phân loại NCBI. Nếu dòng dõi hoàn chỉnh của một sinh vật rất dài thì dòng viết tắt sẽ được hiển thị trong bản ghi GenBank và dòng dõi hoàn chỉnh sẽ có trong Cơ sở dữ liệu phân loại.
REFERENCE	Các ấn phẩm của các tác giả của trình tự thảo luận về dữ liệu được báo cáo trong hồ sơ. Các tài liệu tham khảo được tự động sắp xếp trong bản ghi dựa trên ngày xuất bản, hiển thị các tài liệu tham khảo cũ nhất trước tiên.
FEATURES	Thông tin về gen và sản phẩm gen cũng như các vùng có ý nghĩa sinh học được báo cáo trong trình tự. Chúng có thể bao gồm các vùng của chuỗi mã hóa protein và phân tử RNA, cũng như một số tính năng khác.
source	Tính năng bắt buộc trong mỗi bản ghi tóm tắt độ dài của trình tự, tên khoa học của sinh vật nguồn và số ID Taxon. Cũng có thể bao gồm các thông tin khác như vị trí bản đồ, chủng, bản

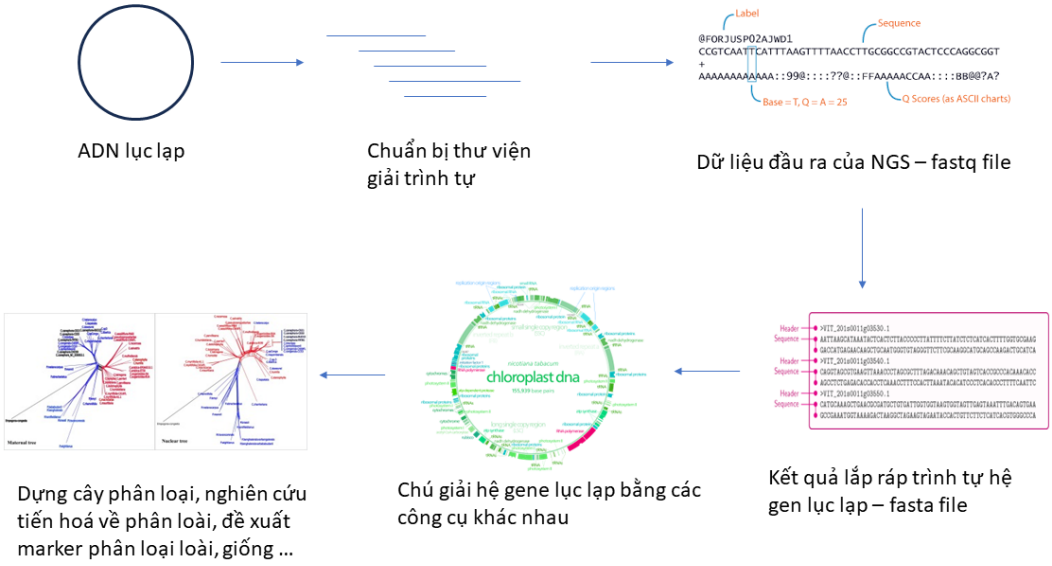
	sao, loại mô, v.v. nếu được người gửi cung cấp.
Taxon	Một số nhận dạng duy nhất ổn định cho đơn vị phân loại của sinh vật nguồn. Số ID phân loại được gán cho từng đơn vị phân loại (loài, chi, họ, v.v.) trong Cơ sở dữ liệu phân loại NCBI.
CDS	Trình tự mã hóa; vùng nucleotide tương ứng với trình tự axit amin trong protein (vị trí bao gồm codon bắt đầu và kết thúc). Tính năng CDS bao gồm dịch mã axit amin. Người gửi cũng được khuyến khích chú thích đặc điểm mRNA, bao gồm vùng chưa được dịch 5' (5'UTR), trình tự mã hóa (CDS, exon) và vùng chưa được dịch 3' (3'UTR).
<1..206	Khoảng cơ sở của đặc điểm sinh học được chỉ ra ở bên trái, trong trường hợp này là đặc điểm CDS. (Tính năng CDS được mô tả ở trên và khoảng cơ sở của nó bao gồm các codon khởi đầu và kết thúc.) Các tính năng có thể hoàn chỉnh, một phần ở đầu 5', một phần ở đầu 3' và/hoặc trên chuỗi bổ sung. Ví dụ: tính năng hoàn chỉnh được viết đơn giản là n..m
protein_id	Số nhận dạng trình tự protein, tương tự như số Phiên bản của trình tự nucleotide. ID protein bao gồm ba chữ cái, theo sau là năm chữ số, dấu chấm và số phiên bản. Nếu có bất kỳ thay đổi nào đối với dữ liệu trình tự (thậm chí chỉ một axit amin), số phiên bản sẽ tăng lên nhưng phần gia nhập sẽ vẫn ổn định (ví dụ: AAA98665.1 sẽ thay đổi thành AAA98665.2).
translation	Sự dịch mã axit amin tương ứng với trình tự mã hóa nucleotide (CDS). Trong nhiều trường hợp, các bản dịch mang tính khái niệm. Lưu ý rằng tác giả có thể chỉ ra liệu CDS dựa trên bằng chứng thực nghiệm hay phi thực nghiệm.
gene	Một vùng sinh học được quan tâm được xác định là một gen và được đặt tên. Khoảng cơ sở cho đặc điểm gen phụ thuộc vào đặc điểm 5' và 3' xa nhất. Các ví dụ bổ sung về bản ghi thể hiện mối quan hệ giữa các đặc điểm gen và các đặc điểm khác như mRNA và CDS là AF165912 và AF090832.

complement	Chỉ ra rằng gene này nằm trên sợi bổ sung hay sợi gốc

Trong rất nhiều trường thông tin như vậy: một số trường thông tin cần lưu ý khi trích xuất thông tin từ file này là trường thông tin về vị trí, tên gene và trình tự gene.

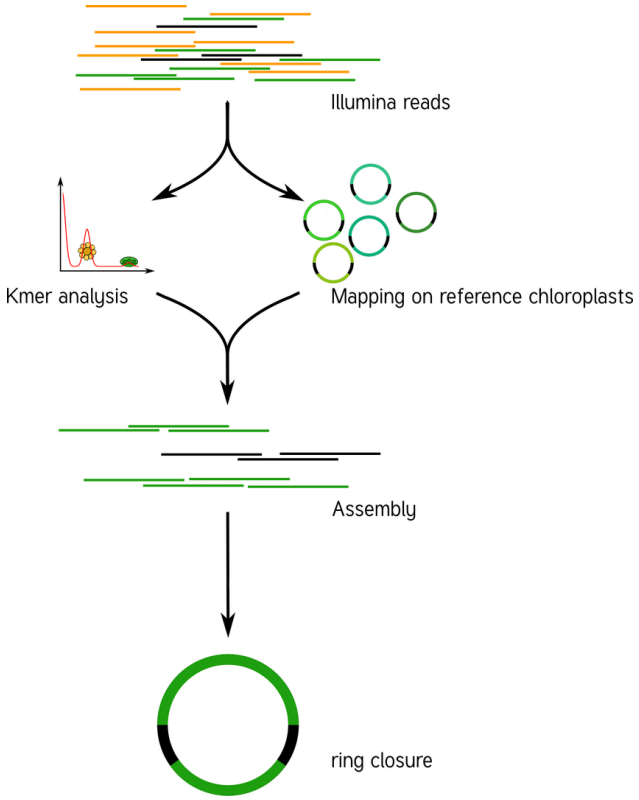
1.5 *Quy trình phân tích hệ gen lục lạp*

Bộ gen lục lạp thường được sử dụng trong nghiên cứu sinh học quần thể và thực vật học vì sự đơn giản của cấu trúc bộ gen hình tròn, sự di truyền chủ yếu là dòng vô tính của nó dọc theo dòng mẹ cũng như số lượng bản sao cao trong tế bào[20]. Bộ gen lục lạp thường được cho là có độ bảo thủ cao, số lượng variants thấp sự biến đổi trình tự và do đó việc sử dụng bộ gen được chủ yếu giới hạn trong các nghiên cứu ở phạm vi liên loài và liên họ[21]. Những phân tích so sánh gần đây của trình tự lục lạp hoàn chỉnh cho thấy nhận thức về sự biến đổi thấp của lục lạp trong loài là sai khi nhìn vào quy mô bộ gen. Kane và cộng sự đề xuất rằng toàn bộ bộ gen lục lạp có thể được sử dụng làm mã vạch để xác định các giống cây trồng[22]. Hơn nữa, sử dụng một hoặc một số vùng của bộ gen lục lạp không phù hợp để mô tả mức độ biến đổi của bộ gen lục lạp. Vì vậy, việc sử dụng bộ gen lục lạp hoàn chỉnh chắc chắn là cách tốt nhất để khai thác thông tin về tiến hoá.



Ảnh 1-6: Quy trình phân tích hệ gen lục lạp.

Quy trình phân tích hệ gen lục lạp được mô tả như trong Ảnh 1-6. Thông qua dữ liệu giải trình tự toàn bộ hệ gen của các loài thực vật, các nhà nghiên cứu đã thống kê được rằng có khoảng 5% số đoạn đọc trở lên có nguồn gốc từ lục lạp.[23] Điều này cung cấp một cách khác để có được bộ gen lục lạp. Thay vì tách riêng lục lạp bằng phương pháp sucrose gradient hoặc high salt, sau đó tách chiết ADN lục lạp và giải trình tự; trình tự ADN lục lạp trong các dữ liệu giải trình tự toàn bộ hệ gen WGS sẽ được dóng hàng vào bộ gen lục lạp tham chiếu (có thể là bộ gen sẵn có hoặc bộ gen của loài gần nó). Như vậy phương pháp dựa trên dóng hàng là một trong những phương pháp phổ biến được lựa chọn để thực hiện so sánh trình tự trong những năm gần đây. Tuy nhiên, vì cấu trúc và chức năng trong bộ gen có thể khác nhau nên các phương pháp dựa trên sự liên kết như vậy có thể trở thành không đáng tin cậy đối với các đơn vị phân loại mà không có họ hàng gần gũi tồn tại với bộ gen lục lạp chất lượng cao.



Ảnh 1-7: Mô tả cơ bản về workflow xử lý dữ liệu và lắp ráp trình tự hệ gen lục lạp[24]

Vì vậy, phương pháp lắp ráp *denovo* hệ gen lục lạp là phương án tối ưu hơn khi muốn tạo ra hệ gen lục lạp. Hiện nay phổ biến nhất là sử dụng bảng tần số *k-mers*. *K-mers* là một chuỗi con chính xác của chuỗi DNA có độ dài xác định *(k)*, tần số của nó trong một tập hợp các chuỗi DNA có thể được tính một cách đơn giản.[24] Ứng dụng thống kê trên việc chia sẻ k-mer giữa các mẫu cho phép ước tính về khoảng cách di truyền.[25]

Từ bảng tần số *k-mer* có thể vẽ được biểu đồ phân phối tần số *k-mer* và cho thấy số lượng *k-mer* xuất hiện ở mỗi tần số trong tập dữ liệu. Đây là cơ sở cho việc lắp ráp của công cụ plasmidSPAdes[26] và Recycler[27] hai công cụ được đánh giá là tốt khi thực hiện lắp ráp các trình tự hệ gen ti thể hoặc hệ gen lục lạp.

Để chú giải hệ gen lục lạp các nhà nghiên cứu thường sử dụng các công cụ phổ biến như: DOGMA, Verdant, CPGAVAS2, GeSeq, PGA… với các cách

tiếp cận chú giải khác nhau. Tuy nhiên, có thể chia thành 2 phương pháp cơ bản: sử dụng các công cụ BLAST để so sánh trình tự giữa hệ gen tham chiếu (hệ gen đã có sẵn, được nghiên cứu đầy đủ) và hệ gen đích; sử dụng HMM profile để so sánh.

Sau khi có kết quả chú giải hệ gen, các nhà nghiên cứu có thể sử dụng trình tự các gene đã được chú giải hoặc toàn bộ trình tự hệ gen để thực hiện dựng cây phân loài và nghiên cứu tiến hoá. Thông thường cây phân loài được suy luận khi sử dụng phương pháp Maximum Likelihood và mô hình TamuraNei [72, 73]. Các kết quả này thực sự có ý nghĩa trong việc áp dụng định danh hoặc phân loại thực vật. Ngoài ra có thể sử dụng những kết quả này để áp dụng vào việc chọn, tạo giống các loài thực vật có giá trị kinh tế cao như sâm ngọc linh[9]–[11], đông trùng hạ thảo, cà phê…

Hệ thống ngân hàng gen NCBI vẫn còn rất nhiều hệ gen lục lạp được lắp ráp, chú giải sai sót mặc dù đó là những hệ gen đã được nghiên cứu kỹ lưỡng. Một số lỗi phổ biến như: gen bị cắt ngắn, thêm vào những phần mở rộng không mong muốn của các exon, bỏ sót các gen đã biết, lựa chọn sai các chuỗi mã hoá, các khung đọc mở được giả định là gen chức năng… Việc chú giải gen chức năng của lục lạp rất quan trọng, việc này giúp ích cho các nhà nghiên cứu về phân loài có thể áp dụng để phân loại chính xác các cây thực vật gần gũi trong cùng chi, họ; việc chú giải sai có thể dẫn đến một hệ quả domino khi những người nghiên cứu sau sử dụng những kết quả chưa chính xác này cho những nghiên cứu của mình. Tính đến thời điểm hiện tại chưa có phần mềm chú giải hệ gen lục lạp nào có ưu thế và chưa có bước tiến lớn nào trong việc nâng cao thuật toán chú giải hệ gen lục lạp vì số lượng hạn chế các nhà khoa học về khoa học máy tính, thuật toán tin sinh học phát triển những thuật toán mới cho việc này. Đến nay chỉ có một số công cụ hỗ trợ chú giải lục lạp như : Dual Organellar GenoMe Annotator (DOGMA)[3]; Chloroplast Genome Annotation, Visualization, Analysis, and GenBank Submission (CPGAVAS & CPGAVAS2) [4]; GeSeq [5]) ;Verdant [6], PGA. Tuy nhiên, chúng đều có những ưu điểm và

khuyết điểm riêng. Việc khảo sát, đánh giá những phần mềm này có ý nghĩa quan trọng nhằm nâng cao chất lượng chú giải gen chức năng trong hệ gen lục lạp.

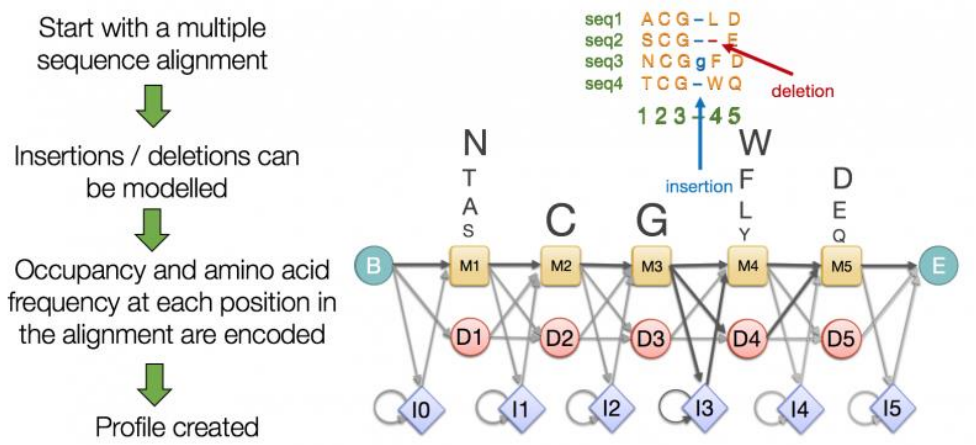
Hiện nay, CPGAVAS2, GeSeq, PGA là các công cụ nổi bật đã được một số nhà nghiên cứu khác đánh giá cao trong những nghiên cứu riêng lẻ [28]–[30]. Nhưng chưa có nghiên cứu cụ thể sao so sánh trực tiếp 03 công cụ này. Trong khuôn khổ của luận văn này tôi thực hiện so sánh, đánh giá công cụ CPGAVAS2, GeSeq, PGA trên một tập dữ liệu cụ thể: Hệ gen lục lạp Cà phê Arabica.

Cà phê Arabica là một loại cây công nghiệp mang lại giá trị cao của Thế giới cũng như Việt Nam – năm 2022, Việt Nam đạt giá trị xuất khẩu hơn 250 triệu USD – nhưng cho đến nay cũng chưa có nghiên cứu cụ thể nào về chọn giống cà phê arabica thông qua hệ gen lục lạp ở Việt Nam. Như vậy, việc đưa ra một phương pháp tối ưu cho việc chú giải hệ gen lục lạp cũng sẽ đóng góp một phần nhỏ cho việc gia tăng sản lượng và giá trị cà phê Việt Nam thông qua các nghiên cứu về chọn, tạo giống cà phê.

2 CHƯƠNG 2: CÁC PHƯƠNG PHÁP CHÚ GIẢI HỆ GEN LỤC LẠP

Có 2 phương pháp cơ bản để chú giải hệ gen lục lạp: sử dụng các công cụ BLAST để so sánh trình tự giữa hệ gen tham chiếu (hệ gen đã có sẵn, được nghiên cứu đầy đủ) và hệ gen đích; sử dụng HMM profile để so sánh.

Trong tin sinh học, BLAST (basic local alignment search tool) là một thuật toán và chương trình để so sánh thông tin trình tự sinh học cơ bản, chẳng hạn như trình tự axit amin của protein hoặc nucleotide của trình tự DNA và/hoặc RNA[31]. Tìm kiếm bằng BLAST cho phép các nhà nghiên cứu so sánh trình tự protein hoặc nucleotide (được gọi là truy vấn) với thư viện hoặc cơ sở dữ liệu về trình tự và xác định trình tự cơ sở dữ liệu giống với trình tự truy vấn trên một



Ảnh 2-1: Mô tả quá trình hình thành HMM profile

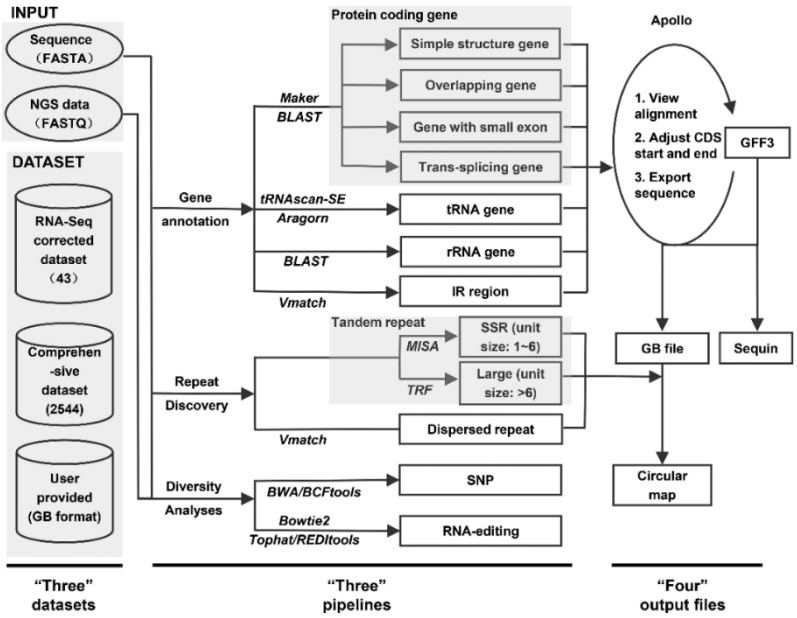
ngưỡng nhất định. Một số công cụ điển hình như: DOGMA[32], Verdant[33], CPGAVAS[34]…

Các mô hình Hidden markov profile (HMM) là một trong những quy trình thành công nhất để phát hiện sự tương đồng giữa các protein. HMM profile là một biến thể của HMM liên quan cụ thể đến trình tự sinh học. HMM profile biến việc liên kết nhiều chuỗi thành một hệ thống tính điểm dành riêng cho vị trí, hệ thống này có thể được sử dụng để dóng hàng các chuỗi và tìm kiếm cơ sở dữ liệu cho các chuỗi tương đồng từ xa[35]. Tận dụng thực tế là các vị trí nhất

định trong sự sắp xếp trình tự có xu hướng có các sai lệch trong đó các phần dư có nhiều khả năng xảy ra nhất và có khả năng khác nhau về xác suất chứa phần chèn hoặc phần xóa. Việc thu thập thông tin này mang lại khả năng phát hiện các điểm tương đồng thực sự tốt hơn so với các phương pháp dựa trên BLAST truyền thống, phương pháp này xử phạt các hành vi thay thế, chèn và xóa như nhau, bất kể chúng xuất hiện ở đâu khi thực hiện dóng hàng.

Dưới đây tôi xin trình bày thuật toán của những đại diện tiêu biểu cho hai phương pháp chú giải lục lạp này gồm: CPGAVAS2 và PCA (sử dụng BLAST để tìm kiếm các gen tương đồng) và GeSeq sử dụng kết hợp profile HMM và BLAST để tìm kiếm gene chức năng, chú giải hệ gen.

2.1 *Thuật toán CPGAVAS/CPGAVS2*



Ảnh 2-2:Quy trình phân tích của CPGAVAS2. 3 Step 3-3-4

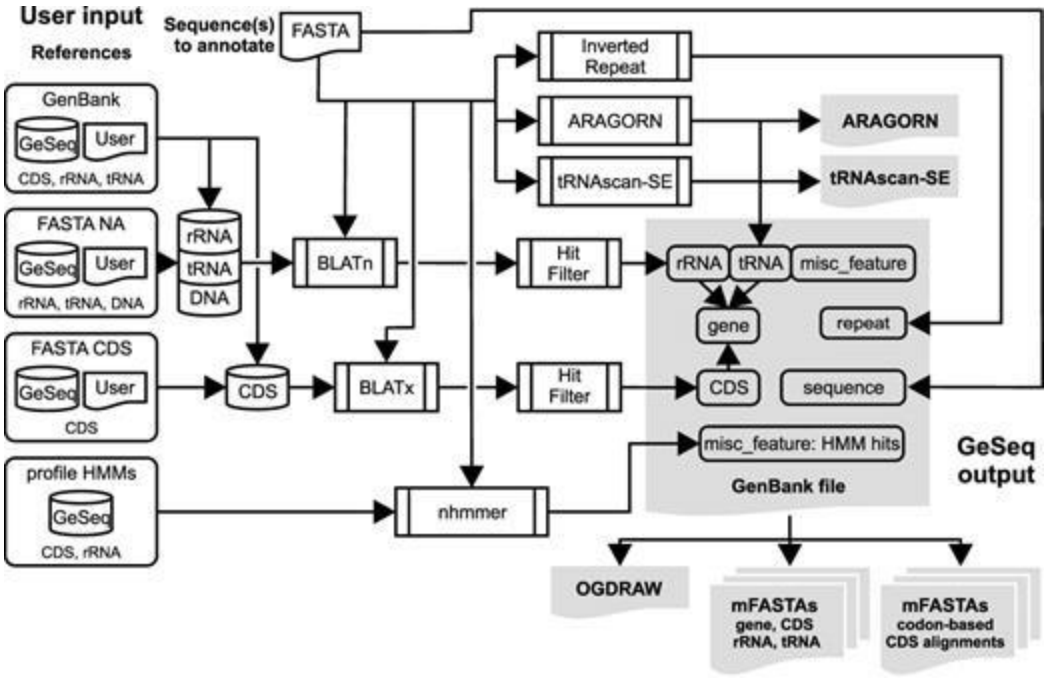
CPGAVAS2 lấy chuỗi plastome ở định dạng FASTA và dữ liệu NGS tùy chọn ở định dạng FASTQ làm đầu vào. Các bộ dữ liệu, quy trình phân tích và tệp đầu ra được được hiển thị trong Ảnh 2-2 và được tóm tắt dưới dạng step-workflow: 'ba', 'ba' và 'bốn'. Step 'Ba' đầu tiên cho thấy rằng CPGAVAS2 hỗ trợ ba bộ dữ liệu khác nhau để chú giải bộ gen, chẳng hạn như bộ dữ liệu đã có dữ liệu RNA-seq (43-tập dữ liệu plastome), tập dữ liệu công cộng đầy đủ (tập dữ

liệu 2544-plastome lấy ở trên ngân hàng NCBI) và trình tự do người dùng cung cấp[29]. Bước ‘ba’ thứ hai chỉ ra rằng CPGAVAS2 hỗ trợ ba loại quy trình, cụ thể là chú thích bộ gen, lặp lại nhận dạng và phân tích đa dạng (thăm dò). Step 'bốn' thông tin rằng CPGAVAS2 tạo ra bốn loại đầu ra: tệp GFF3 để chỉnh sửa thủ công bằng các trình chỉnh sửa, chẳng hạn như Apollo; một tệp đồ họa hiển thị các gen và các đoạn lặp lại có chú thích; một tệp ở định dạng GenBank; và một bộ tệp dữ liệu có cấu trúc người dùng có thể tải lên ngân hàng gen NCBI.

Thuật toán tìm kiếm các gene chức năng được tối ưu và trình bày trong nghiên cứu của Liu khi công bố về công cụ CPGAVAS bao gồm 4 bước cơ bản như sau:

- Bước 1 nhóm các chuỗi protein, cDNA và “gen rRNA” thành các nhóm tương đồng dựa trên cơ sở dữ liệu GenBank và sau đó tạo cơ sở dữ liệu có khả năng tìm kiếm cho mỗi nhóm.
- Bước 2 tạo các protein tham chiếu và một tập dữ liệu cDNA + “gen rRNA” tham chiếu cho mỗi chuỗi bộ gen truy vấn đầu vào.
- Bước 3, các chuỗi protein tham chiếu, cDNA và “gen rRNA” được ánh xạ tới trình tự bộ gen bằng các chương trình Blastx, Blastn, protein2genome và est2genome.
- Bước 4, các vùng lặp đảo ngược được xác định bằng công cụ phần mềm vmatch với các tham số mặc định. Và tRNA được xác định bằng tRNAscan với các tham số do người dùng chỉ định.[34]

2.2 *Thuật toán GeSeq*



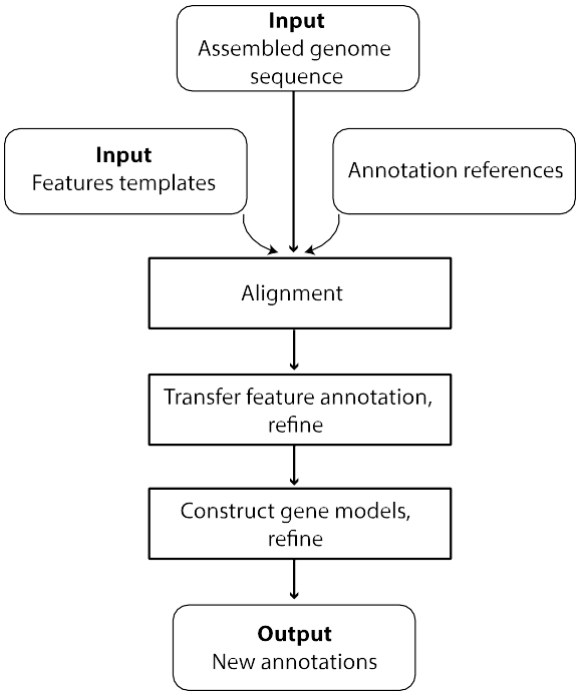
Ảnh 2-3: Thuật toán GeSeq

Người dùng cung cấp (các) trình tự FASTA axit nucleic để chú thích và chọn hoặc cung cấp các trình tự axit nucleic tham chiếu ở định dạng GenBank hoặc FASTA. Dựa trên các chuỗi tham chiếu đã chọn hoặc được tải lên, GeSeq xây dựng cơ sở dữ liệu BLAT không mã hóa protein (rRNA, tRNA và DNA) và cơ sở dữ liệu BLAT mã hóa protein (CDS), thực hiện tìm kiếm so sánh trình tự ADN ('BLATn') và BLAT ('BLATx') tìm kiếm tương ứng trình tự ADN sang protein. GeSeq chú thích từ các lượt truy cập được lọc các lớp rRNA, tRNA, CDS và gen tiềm năng. “Các gen tiềm năng” là kết quả của các lần so sánh tRNA, rRNA và CDS và bao gồm các intron (nếu có). Các lần so sánh DNA được chú thích là 'misc_features' hoặc, cách khác, là 'primer_bind'. Ngoài ra, người dùng có thể kích hoạt tìm kiếm nhmmer bằng cách chọn các HMM hồ sơ của các chuỗi CDS và rRNA (hiện chỉ có lục lạp) làm tài liệu tham khảo. Tất cả các lần so sánh hồ sơ HMM đều được chú thích dưới dạng misc_features để hỗ trợ quản lý thủ công. Theo tùy chọn, người dùng có thể gọi ARAGORN hoặc tRNAscan-SE để chú thích *de novo* các tRNA và tự tìm kiếm BLASTN để phát hiện cặp lặp đảo ngược (IR) thường thấy trong bộ gen lục lạp. Đầu ra GeSeq tối

thiểu (tất cả các tệp đầu ra được dán nhãn màu xám) là tệp GenBank chứa tất cả các chú thích và phần giải thích của nó bởi OGDRAW để đánh giá nhanh. Ngoài ra, người dùng có thể chọn các đầu ra tùy chọn bổ sung, bao gồm các tệp đa FASTA riêng biệt ('mFASTAs') chứa các chuỗi chú thích thuộc về các lớp gen, CDS, rRNA và tRNA. Nếu một số trình tự được tải lên để chú thích trong cùng một công việc thì các mFASTA kết hợp cho tất cả các trình tự có chú thích của bốn lớp cũng được cung cấp để tải xuống và theo tùy chọn, việc sắp xếp dựa trên codon có thể được tạo ra cho tất cả các trình tự CDS được chú thích có hoặc không có GenBank được chọn hoặc tải lên.

Để cung cấp cho GeSeq một bộ tham chiếu chất lượng cao về các trình tự hệ gen lục lạp tham chiếu, tác giả đã chọn bộ gen lục lạp hoàn chỉnh của 34 loài thực vật trải rộng trên toàn bộ phạm vi phân loại từ rêu đến cây giống. Sau đó, họ tạo ra nhiều cách sắp xếp cho từng gen mã hóa protein và rRNA, đồng thời quản lý các cách sắp xếp này theo cách thủ công[36]. Do đối với nhiều loài trong số 34 loài, không có xác nhận thực nghiệm về gen hoặc chú thích intron exon, nên một số hệ gen được chú thích khi sử dụng dữ liệu từ các sinh vật được nghiên cứu chuyên sâu về biểu hiện gen lục lạp, như *Arabidopsis thaliana, Nicotiana tabacum, Oenothera elata và Zea mays*. Bộ dữ liệu này được gọi là MPI-MP chloroplast reference set.

2.3 *Thuật toán Chloe*



Ảnh 2-4: Mô hình mô tả quy trình phân tích của Chloe

Công cụ sẽ lấy đầu vào là trình tự bộ gen sẽ được chú thích, một tập hợp các bộ gen tham chiếu và chú thích của chúng cũng như danh sách các mẫu dành cho các đặc điểm được chú thích. Mẫu chứa những kỳ vọng trước về các tính năng, bao gồm cả thứ tự của chúng so với các tính năng khác và độ dài điển hình của chúng. Chloe thực hiện ba bước tuần tự: sắp xếp toàn bộ bộ gen, chiếu chú

Species	Common name	Genome sequence		RNA-seq accession
		Accession	Reference	
Arabidopsis thaliana	Mouse-ear cress	AP000423.1	Sato et al., 1999	SRR1292877
Spinacia oleracea	Spinach	NC_002202.1	Schmitz-Linneweber et al., 2001	SRR3680359
Nicotiana tabacum	Tobacco	Z00044.2	Shinozaki et al., 1986	SRR6149304
Medicago truncatula	Barrelclover	JX512022.1	Gurdon & Maliga, 2014	SRR504353
Zea mays	Maize	NC_001666.2	Maier et al., 1995	SRR3089594
Oryza sativa	Rice	NC_001320.1	Hiratsuka et al., 1989	DRR147591
Nymphaea mexicana	Mexican waterlily	NC_024542.1	Yang et al., 2014	SRR10158656
Liriodendron chinense	Chinese tulip poplar	NC_030504.1	B. Li et al., 2016	SRR9949011
Amborella trichopoda	NA	NC_005086.1	Goremykin, 2003	SRR9087794
Zamia furfuracea	Cardboard palm	NC_026040.1	NA	SRR5894356
Gnetum montanum	NA	NC_021438.1	Mao et al., 2017	SRR5908685
Ginkgo biloba	Ginkgo	NC_016986.1	C.-P. Lin et al., 2012	SRR10311648
Picea glauca	White spruce	KT634228.1	Jackman et al., 2015	SRR2134666
Marsilea crenata	NA	NC_022137.1	Gao et al., 2013	NA
Azolla caroliniana	Water fern	MF177093.1	F.-W. Li et al., 2018	SRR5499388
Selaginella moellendorffii	Lycophyte	FJ755183.1	Smith, 2009	SRR4762345
Physcomitrella patens	Moss	AP005672.1	Sugiura et al., 2003	SRR3350460
Marchantia paleacea	Liverwort	NC_001319.1	Logacheva et al., 2009	SRR3926770
Anthoceros formosae	Hornwort	NC_004543.1	Kugita et al., 2003	ERR2040977

Ảnh 2-5: Danh sách các dữ liệu được lựa chọn để xây dựng cơ sở hệ gen tham chiếu của Chloe

thích và dự đoán và sàng lọc mô hình gen.

Tương tự như CPGAVAS2 Chloe cũng xây dựng cơ sở dữ liệu hệ gen tham chiếu dựa trên những trình tự tốt, có kết quả RNA-seq. Công cụ sử dụng 19 hệ gen lục lạp có chất lượng tốt, có kết quả RNA-seq. Đặc biệt những loài này được lựa chọn đại diện cho sự đa dạng phát sinh loài của thực vật trên cạn, bao gồm các dòng chính trong nhánh: rêu, rêu, dương xỉ, thực vật lycophytes, thực vật hạt trần, thực vật hạt kín sớm, thực vật hai lá mầm và thực vật một lá mầm. Cơ sở dữ liệu này sẽ cung cấp bằng chứng cho việc chú giải tương đối đáng tin cậy của các loài có liên quan chặt chẽ (hệ gen cùng loài, loài gần hoặc cùng họ). Hơn nữa, các dữ liệu lục lạp này đều có đủ dữ liệu RNA-Seq để xác minh ranh giới exon-intron. Ngoài ra, hầu hết các loài đều là thực vật được nghiên cứu rộng rãi và do đó thường có sẵn các bằng chứng thực nghiệm khác.[37]

Dóng hàng

Chloe sử dụng mảng hậu tố để căn chỉnh nhanh các kết quả khớp chính xác giữa hai bộ gen. Tiền tố của một chuỗi (ví dụ: 'ATCG') là một chuỗi ký tự (chuỗi con) liền kề trong chuỗi 'ATCG' xuất hiện ở đầu, bao gồm 'A', 'AT', 'ATC', 'ATCG'. Hậu tố của một chuỗi là chuỗi con xuất hiện ở cuối, bao gồm 'ATCG', 'TCG', 'CG', 'G' và các vị trí bắt đầu của chúng là 1, 2, 3, 4. Mảng hậu tố là danh sách các chuỗi vị trí bắt đầu của các hậu tố, được sắp xếp theo thứ tự bảng chữ cái (Shrestha et al. 2014). Ví dụ: mảng hậu tố của 'ATCG' là 1, 3, 4, 2. Mảng hậu tố là một cấu trúc dữ liệu đơn giản, hiệu quả và được áp dụng phổ biến cho các ứng dụng tin sinh học, chẳng hạn như công cụ căn chỉnh trình tự theo cặp MUMmer4. Giai đoạn dóng hàng của Chloe được thực hiện như sau:

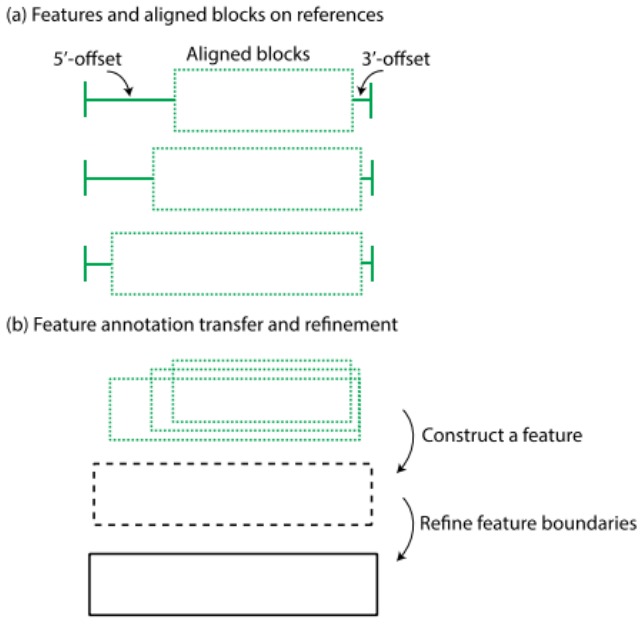
- Dóng hàng bằng cách hợp nhất các mảng hậu tố; xử lý từng sợi một cách độc lập
- Lấy các khối được dóng hàng dài nhất từ mảng tiền tố cho tới khối hậu tố đã được dóng hàng

- Lấp đầy khoảng trống giữa các khối được dóng hàng bằng cách tìm kiếm các kết quả khớp ngắn hơn
- Hợp nhất những khối đã dòng hàng liền kề cách đều nhau ở cả hai hệ gen.

Điều này sẽ cho kết quả những khối dóng hàng giữa hai hệ gen có thể có những điểm sai khác nhưng sẽ không có những khoảng trống.

Di chuyển chú giải

- Đối với mỗi chức năngđược chú giải, kết quả chú giải sẽ được "thuyên chuyển" sang những khối được dóng hàng
- Các chú giải được di chuyển mang bản ghi về độ lệch từ đầu 5' và 3' của các chức năng mà chúng bắt nguồn từ đó (nếu đó là CDS chuỗi mã hóa)
- Xây dựng những nhóm được chú thích theo quy trình sau: với mỗi nucleotide trong trình tự hệ gene, đếm số lượng chú thích được di chuyển của mỗi loại chức năng được dóng hàng tại vị trí đó
- Dóng hàng cho mỗi nhóm chú thích
- Xây dựng bảng chú thích chức năng trên hệ gen đích dựa trên những bản mẫu đã được dóng hàng
- Xác định ranh giới của các vùng chức năng bằng cách dự đoán cận biên dựa trên độ lệch 3' và 5' dựa trên các chú giải được di chuyển



Ảnh 2-6: Mô tả phương thức di chuyển chú giải

Xây dựng mô hình gene

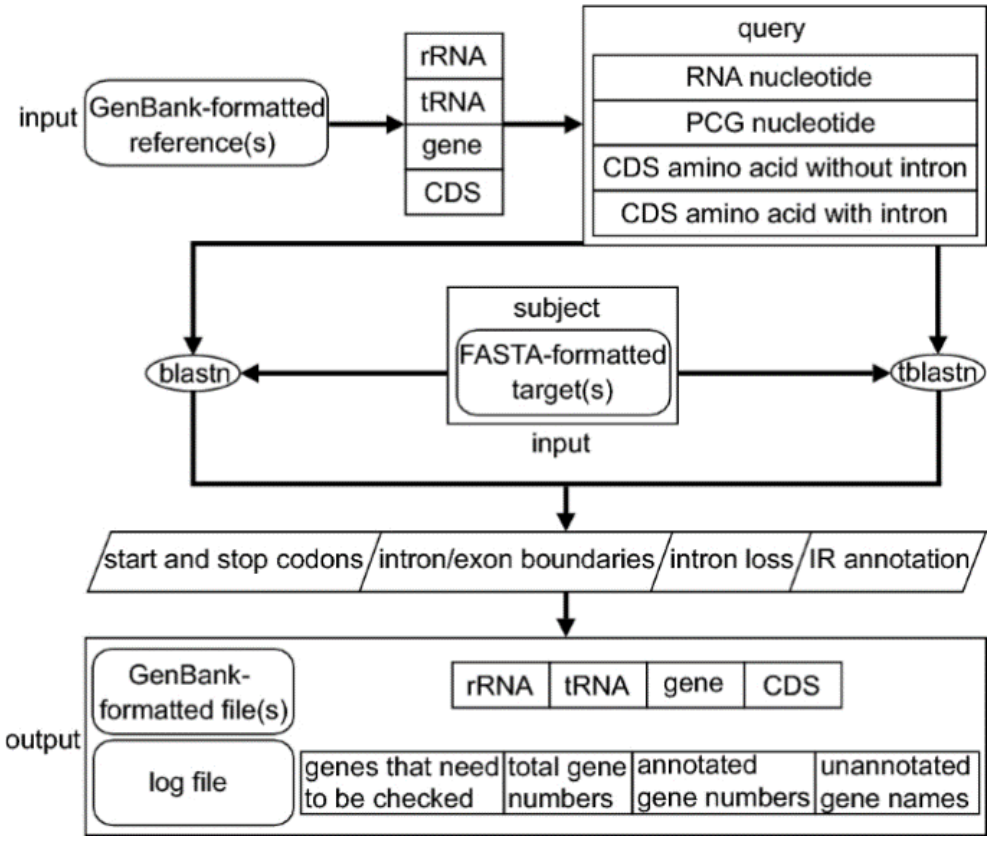
- Thu thập những chức năng liên quan vào mô hình gene
- Nếu là một gene mã hoá, tính toán vị trí của exon cuối cùng và tìm kiếm vị trí bộ ba kết thúc.
- Kiểm tra những chức năng có thể cạnh nhau (ranh giới intron-exon, các gen cạnh nhau …)
- Xác định các exon liên tiếp gần nhau cùng pha
- Tìm kiếm các vị trí có thể là bộ ba mở đầu, bao gồm cả bộ ba "ACG" và "GTG" – những bộ ba hiếm gặp hoặc được tìm kiếm từ cơ sở dữ liệu RNA

Tôi nhận thấy rằng Chloe là một thuật toán tối ưu, có một số cải tiến, đặc biệt cơ sở dữ liệu lục lạp tham chiếu của Chloe được cho là đầy đủ, trải dài trên nhiều loài thực vật. Đã được so sánh có hiệu quả khi so sánh với các công cụ chú giải lục lạp khác[24]. Ngoài ra, Chloe được tích hợp cùng với GeSeq, vì vậy tôi lựa chọn phương án sử dụng công cụ GeSeq kết hợp thuật toán Chloe để chú

giải hệ gen lục lạp. Với dự đoán phương án này sẽ kết hợp được điểm mạnh của 2 thuật toán và cho kết quả tốt.

2.4 *Thuật toán PGA*

PGA (Plastid Genome Annotator), một công cụ độc lập sử dùng dòng lệnh, có thể thực hiện chú thích hàng loạt hệ gen lục lạp nhanh chóng, chính xác và linh hoạt. Công cụ sử dụng cho các hệ gen mục tiêu mới được chú giải dựa trên các hệ gen tham chiếu đã được chú thích rõ ràng. Ngược lại với các công cụ hiện có, PGA sử dụng các hệ gen lục lạp tham chiếu làm truy vấn và các hệ gen lục lạp đích làm chủ đề để xác định vị trí gen - gọi là phương pháp tìm kiếm BLAST truy vấn ngược.



Ảnh 2-7: Mô tả thuật toán chú giải của PGA

PGA xác định chính xác ranh giới gen và intron cũng như sự mất mát intron. Chương trình xuất ra các tệp có định dạng GenBank cũng như tệp nhật ký để hỗ trợ người dùng xác minh chú thích.

Ảnh 3-8 thể hiện sáu bước được tiến hành để chú thích các plastome: (1) Chuẩn bị các plastome tham chiếu theo định dạng GenBank; (2) Chuẩn bị các plastomes đích có định dạng FASTA; (3) Tạo cơ sở dữ liệu tham khảo; (4) Tìm kiếm BLAST; (5) Xác định ranh giới đối tượng; (6) Tạo GenBank và tệp nhật ký[28].

Định vị gen

Các gen rRNA và tRNA trong hệ gene lục lạp đích được tìm kiếm bằng BLASTN trên cơ sở dữ liệu lục lạp tham chiếu. Đối với các gen mã hóa protein (PCG), BLASTN và TBLASTN được sử dụng để tìm kiếm. Bất kì PCG (protein coding gene) có tỉ lệ tương đồng TBLASTN lớn hơn mức giá trị ngưỡng (mặc định là 40%) sẽ được chú thích trong hệ gene lục lạp đích. Nếu sử dụng nhiều hơn một hệ gene tham chiếu, mỗi rRNA, tRNA hoặc PCG có giá trị tương đồng cao nhất bằng công cụ BLAST/TBLASTN sẽ được chọn để xác định cặp trình tự liên tục có giá trị tương đồng cao (High scoring segment pair - HSP) trong hệ gene đích.

Trong một số trường hợp đặc biệt như với các gen *rpl16, petB* và *petD*. Mỗi gen này có một exon đầu tiên ngắn (6–9 bp) và exon thứ hai dài hơn nhiều. Điều này dẫn đến việc khi sử dụng BLAST/TBLASTN thì sẽ dễ dàng tìm kiếm vùng exon số 2, với những vùng exon ngắn sẽ vô cùng khó khăn. Bởi vì, vùng exon số 1 sẽ trở thành một vùng có độ tương đồng rất cao với nhiều trình tự ở trong các cơ sở dữ liệu tham chiếu.

Thuật toán tìm kiếm ranh giới

Để tìm kiếm ranh giới giữa các vùng một cách chính xác, 3 thuật toán được áp dụng: xác định bộ ba mở đầu và bộ ba kết thúc; xác định vị trí các ranh giới intron-exon và xác định vùng intron bị mất; xác định ranh giới các vùng lặp

đảo. Những kết quả cặp trình tự liên tục có giá trị tương đồng cao (HSPs) được tìm kiếm bằng TBLASTN được sử dụng làm dữ liệu ban đầu. PGA sau đó sẽ xác định bộ ba mở đầu và bộ ba kết thúc cho các gene mã hoá protein. Để xác định bộ ba kết thúc, PGA sẽ bắt đầu từ điểm kết thúc 5’ của HSPs và bộ ba kết thúc đầu tiên được tìm thấy sẽ được xác định là bộ ba kết thúc. Bộ ba mở đầu sẽ được xác định nếu nằm trong một số trường hợp sau: nếu bộ ba đầu tiên của HSP là ATG thì sẽ xác định là bộ ba mở đầu; nếu bộ ba đầu tiên của HSP không phải là methionine (ATG) thì PGA sẽ tìm kiếm ATG trên trình tự HSP, nếu gặp trình tự đầu tiên thì sẽ xác định đó là bộ ba mở đầu; nếu không tìm thấy trình tự ATG thì PGA sẽ xác định 4 trình tự trên trình tự tham chiếu để làm probe tìm kiếm từ phải sang trái.

3 CHƯƠNG 3: CÁC THỰC NGHIỆM VÀ KẾT QUẢ

3.1 *Dữ liệu thử nghiệm*

Nhờ những tiến bộ của công nghệ giải trình tự và lưu trữ. Đã có hơn 4000 bộ gene lục lạp được công bố trên toàn thế giới[38]. Trong suốt 2 thập kỷ phát triển, đã có nhiều nỗ lực của các nhà khoa học để thành lập những cơ sở dữ liệu chuyên biệt cho hệ gene dạng vòng nói chung và hệ gen lục lạp nói riêng. Có thể kể đến một số cơ sở dữ liệu tiêu biểu dưới đây:

Organelle Genome Resources

Cơ sở dữ liệu Organelle Genome Resources là một phần của hệ cơ sở dữ liệu sinh học lớn nhất thế giới NCBI. Cơ sở dữ liệu OGR chứa toàn bộ những hệ gen dạng vòng như mtDNA và hệ gen lục lạp. Cho đến nay, đã có 3020 bộ gen lục lạp từ thực vật trên cạn đã được giải trình tự và lưu trữ trong cơ sở dữ liệu NCBI https://www.ncbi.nlm.nih.gov/genome/browse#!/organelles/, truy cập lần cuối ngày 01 tháng 09 năm 2023). Tuy nhiên, những dữ liệu trên NCBI thường không được kiểm duyệt cẩn thận. Các trình tự được đưa lên NCBI sẽ chỉ được kiểm duyệt bằng nhân viên của NCBI nếu trình tự đó được đăng kí vào hệ cơ sở dữ liệu tham chiếu Refseq. Đặc biệt là với những hệ gene nhỏ như hệ gene lục lạp. Có nhiều trình tự không đầy đủ hoặc lắp ráp chất lượng kém được tìm thấy trên ngân hàng gene NCBI.

ChloroplastDB

Cơ sở dữ liệu bộ gen lục lạp (ChloroplastDB) là một cơ sở dữ liệu dựa trên nền tảng web, dành cho các bộ gen plastid được giải trình tự đầy đủ, chứa các trình tự gen, protein, DNA và RNA, vị trí gen, vị trí chỉnh sửa RNA (http://chloroplast .cbio.psu.edu/)[39]. ChloroplastDB cung cấp các chú thích thống nhất, tìm kiếm tên gen, BLAST và chức năng tải xuống cho các gen và trình tự gen được mã hóa bằng lục lạp. Người dùng có thể truy xuất tất cả các chuỗi chỉnh hình bằng một lần tìm kiếm bất kể tên gen trong GenBank. ChloroplastDB hỗ trợ rất nhiều cho nghiên cứu so sánh về tiến hóa trình tự bao

gồm những thay đổi về hàm lượng gen, cách sử dụng codon, cấu trúc gen và các sửa đổi sau phiên mã như chỉnh sửa RNA. Các bộ protein chỉnh hình được TribeMCL phân loại và mỗi bộ được gán một tên gen tiêu chuẩn.

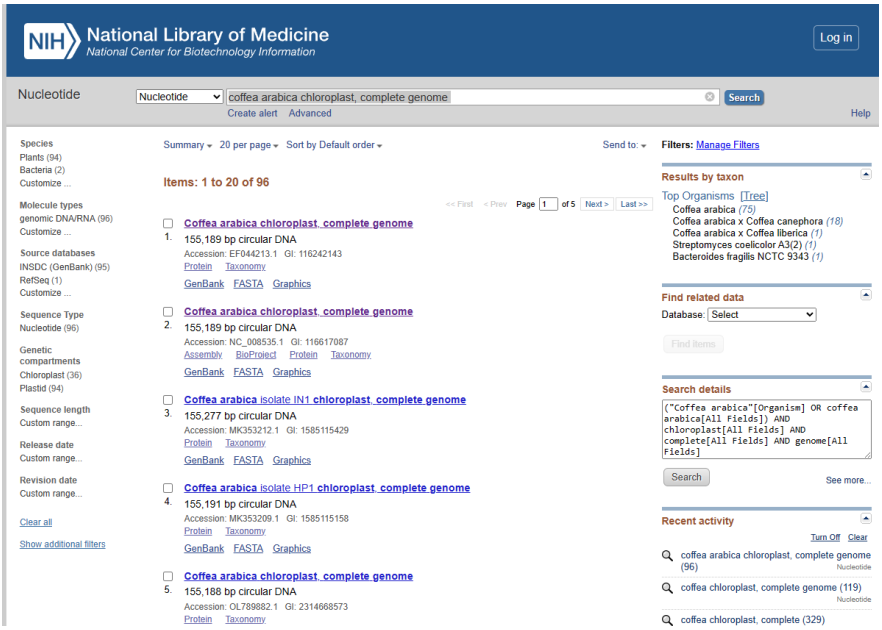
CpGDB (Chloroplast Genome Database)

CpGDB (Chloroplast Genome Database)[40] có địa chỉ tại gndu.ac.in/CpGDB/index.aspx là cơ sở dữ liệu miễn phí, cho phép tìm kiếm và tải xuống trình tự bộ gen lục lạp hoàn chỉnh, trình tự gen riêng lẻ và hồ sơ đặc trưng của các loài thực vật khác nhau. Cơ sở dữ liệu còn cho phép so sánh độ dài gen và vị trí của các gen khác nhau giữa các loài thực vật thuộc cùng họ hoặc khác họ. Phiên bản CpGDB 1.0 được các nhà khoa học Ấn Độ tại trường Đại học Guru Nannak Dev ra mắt vào ngày 29 tháng 02 năm 2020. Theo thống kê trên trang web của cơ sở dữ liệu hiện nay trên cơ sở dữ liệu có 3823 bộ gen lục lạp hoàn chỉnh, thuộc 256 họ. (Dữ liệu được lấy trên trang chủ của cơ sở dữ liệu, truy cập lần cuối ngày 01/09/2023).[40]

Ngoài ra, còn có một số cơ sở dữ liệu lục lạp đã cũ và không còn hoạt động nữa như GOBASE, PCIR (Database of Plant Chloroplast Inverted Repeats). Sau quá trình khảo sát tôi nhận thấy, hiện nay chỉ có cơ sở dữ liệu của NCBI là có trình tự cập nhật. Các cơ sở dữ liệu lục lạp khác như CpGDB hay ChloroplastDB vẫn sử dụng các trình tự ở trên cơ sở dữ liệu NCBI. Các cơ sở dữ liệu này cung cấp thêm cho người dùng một số công cụ phân tích. Tuy nhiên, trong phạm vi luận văn chỉ cần các trình tự hệ gen lục lạp để làm đầu vào cho các công cụ. Do đó, cơ sở dữ liệu NCBI Organelle Genome Resources là thích hợp nhất để khai thác. Mặc dù vậy, đã có những báo cáo cho thấy chất lượng trình tự trên ngân hàng NCBI không đồng đều, nhiều trình tự kém chất lượng vẫn được đưa lên vì vậy, cần thiết có công tác sàng lọc dữ liệu đầu vào.

3.2 *Sàng lọc dữ liệu đầu vào*

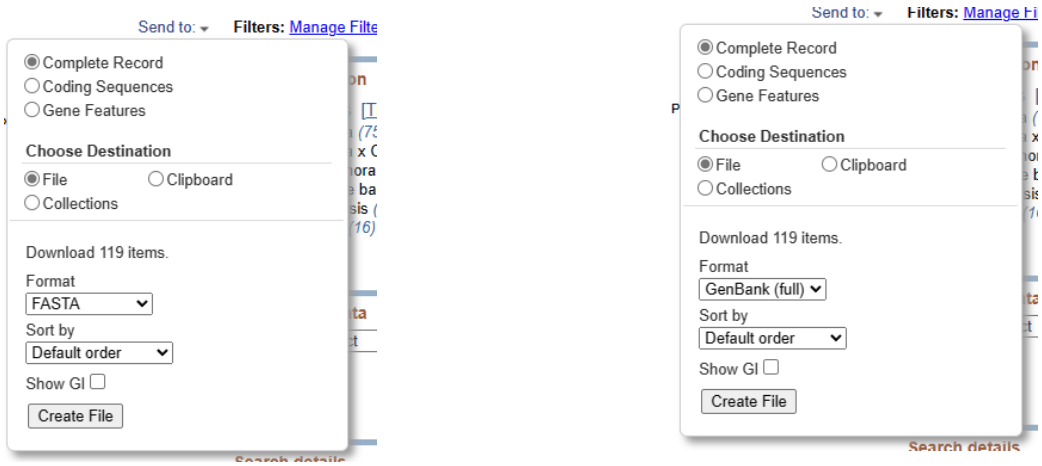
Như đã phân tích ở mục 3.1 dữ liệu đầu vào sẽ được khai thác trên cơ sở dữ liệu NCBI Organelle Genome Resources. Các trình tự hệ gen lục lạp sẽ được



Ảnh 3-1: Kết quả tìm kiếm trình tự lục lạp đầy đủ của loài cà phê arabica

tìm kiếm với từ khoá như sau: "*coffea arabica chloroplast, complete genome*". Tổng cộng có 96 trình tự được tìm thấy dựa vào từ khoá trên.

Để thực hiện lấy trình tự đã lắp ráp của các mẫu, người dùng cần bấm vào "Sent to" ở góc bên phải màn hình. Và lựa chọn thiết đặt như sau:



Ảnh 3-2: Thiết đặt tải về trình tự để phân tích

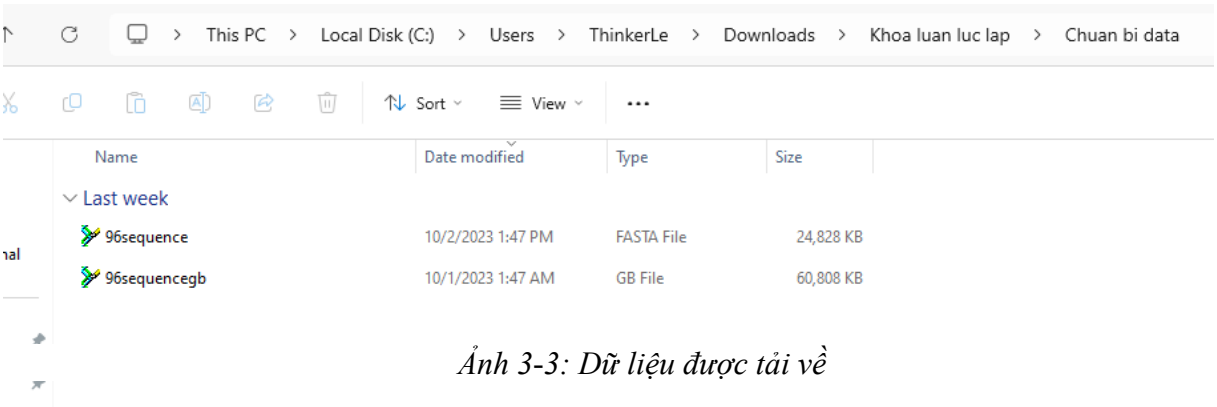
Lựa chọn thiết đặt như vậy có ý nghĩa như sau:

- **Complete Record:** tải về toàn bộ recorde
- **File:** Lựa chọn tải file
- **Fasta:** Lựa chọn tải về trình tự fasta hệ gen lục lạp
- **GenBank (full):** Lựa chọn tải về các bản ghi đã chú giải của các trình tự hệ gen lục lạp

Để phân tích hệ gene lục lạp cần sử dụng 2 loại dữ liệu: trình tự hệ gen lục lạp (định dạnh fasta) và bản ghi kết quả chú giải hệ gen lục lạp (định dạng .gb – genebank). Đồng thời, như đã phân tích ở trên cơ sở dữ liệu NCBI có số lượng trình tự lớn nên không thể tránh khỏi sai sót trong quá trình kiểm duyệt chất lượng trình tự. Vì vậy, nhằm lựa chọn ra những trình tự có chất lượng ổn định để tiến hành so sánh đánh giá ở bước sau, hai loại dữ liệu sẽ được tải về đó là: trình tự hệ gen và file genebank – chứa thông tin đã chú giải của các trình tự hệ gen.

Dữ liệu sau khi được tải về sẽ được xử lý bằng tay, kiểm tra từng trình tự để tìm ra những trình tự có chất lượng tốt dựa vào một số tiêu chí sau:

- Số lượng gene



Ảnh 3-3: Dữ liệu được tải về

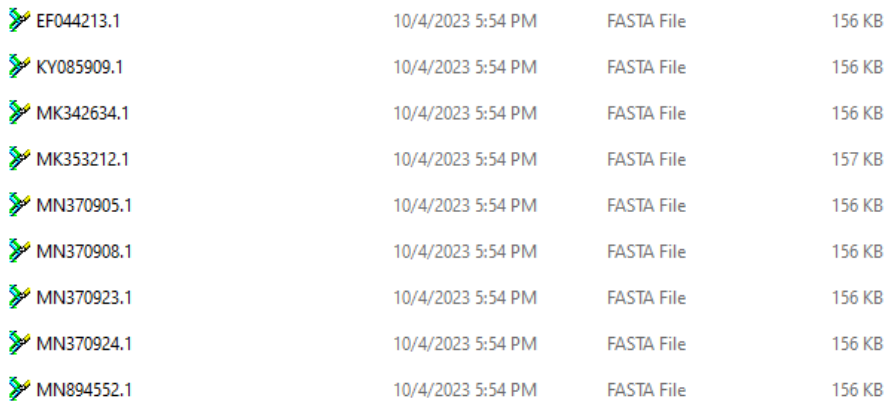
- Cấu trúc trình tự hệ gen: kích thước các vùng SSC, LSC và IR

Sau khi kiểm tra tôi đưa ra lựa chọn 10 trình tự như bảng dưới:

Bảng 3-1: Bảng tổng hợp trình tự sử dụng để so sánh, đánh giá trong luận văn

STT	Genbank ID	Tên loài	Kích thước hệ gen (bp)
1	*EF044213.1*	*Coffea Arabica isolate*	*155 189*
2	*KY085909.1*	*Coffea Arabica isolate*	*155 188*
3	*MK342634.1*	*Coffea Arabica isolate CH3*	*155 191*
4	*MK353209.1*	*Coffea Arabica isolate HP1*	*155 191*
5	*MK353212.1*	*Coffea Arabica isolate IN1*	*155 277*
6	*MN370905.1*	*Coffea Arabica isolate DASF1*	*155 134*
7	*MN370908.1*	*Coffea Arabica isolate GESF1*	*155 059*
8	*MN370924.1*	*Coffea Arabica isolate WEG1*	*155 134*
9	*MN894552.1*	*Coffea Arabica isolate TGG1*	*155 133*
10	*MN370923.1*	*Coffea Arabica isolate CM*	*155 189*

Để chuẩn bị file đầu vào cho các công cụ, tôi thực hiện trích xuất từng trình tự hệ gene thành các file đơn lẻ.



Bảng 3-2: Trình tự hệ gen lục lạp theo từng Genbank ID

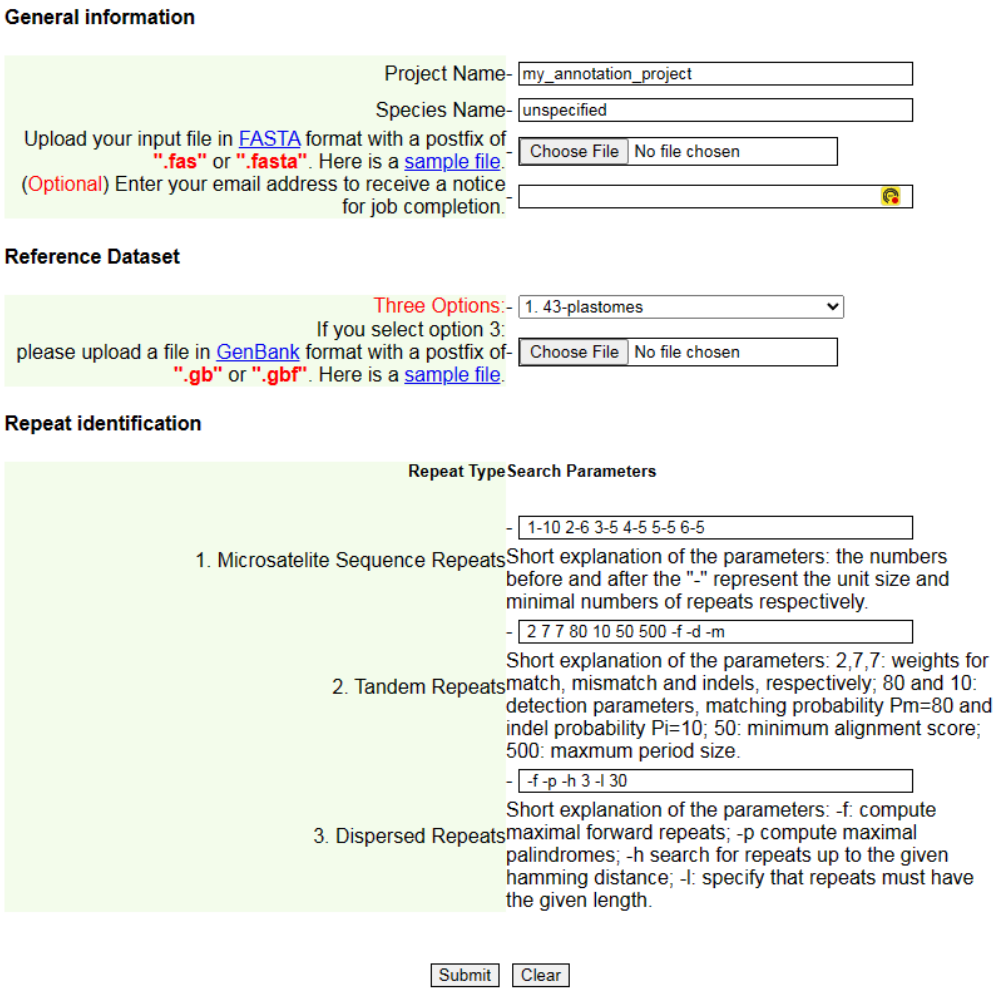
3.3 *Các thực nghiệm*

3.3.1 *Chú giải bằng công cụ CPGAVAS2*

Để thực hiện phân tích được trên công cụ CPGAVAS2, người dùng thực hiện truy cập vào trang web của công cụ:

http://47.96.249.172:16019/analyzer/home

Giao diện công cụ hiển thị như hình dưới.



Ở mục **General Information** – Thông tin cơ bản người dùng cần nhập vào một số thông tin sau:

- Project Name: Tên của dự án
- Species Name: Tên của loài lục lạp cần phân tích

- Upload your input file in FASTA format with a postfix of ".fas" or ".fasta". Here is a sample file.: Vị trí để tải lên trình tự file fasta của hệ gene lục lạp
- (Optional) Enter your email address to receive a notice for job completion.: Tuỳ chọn nhập vào email để công cụ thông báo tình trạng phân tích

Mục **Reference Dataset** – Dữ liệu tham chiếu. Tại mục này người dùng có 3 lựa chọn cơ sở dữ liệu: 43 – plastomes: bộ 43 hệ gene đã có kết quả RNA-Seq; 2544 plastomes: 2544 hệ gen trên cơ sở dữ liệu NCBI; Custom reference in Genbank format: dữ liệu tham chiếu người dùng tự đưa lên để so sánh.

Mục **Repeat identification**: tại đây người dùng có thể thay đổi các tham số để tìm kiếm các vùng lặp đảo trong hệ gen

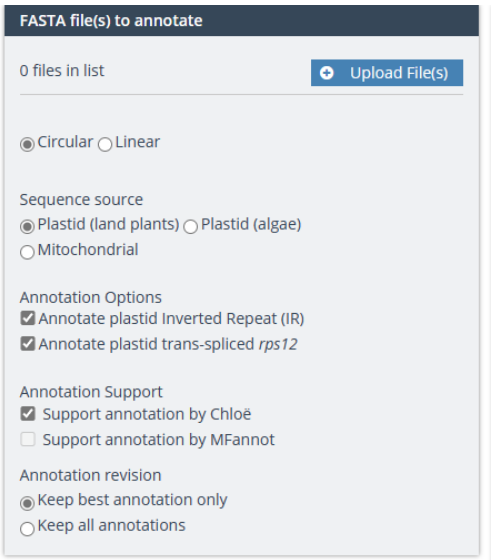
- **Microsatelite Sequence Repeats:** đoạn lặp lại microsatellite là các đoạn lặp lại có kích thước từ 1-6 nucleotide và tổng độ dài không quá 100 nucleotide. Giá trị tham số mặc định: ***1-10 2-6 3-5 4-5 5-5 6-5***. Giải thích tham số: các số trước và sau dấu "-" tương ứng là kích thước đơn vị và số lần lặp lại tối thiểu.
- **Tandem Repeats:** Các chuỗi lặp lại trong hệ gen. Giá trị tham số mặc định: ***2 7 7 80 10 50 500 -f -d -m*** .Giải thích ngắn gọn các tham số: 2,7,7: trọng số lần lượt là khớp, không khớp và indel; 80 và 10: tham số phát hiện, xác suất trùng khớp Pm=80 và xác suất indel Pi=10; 50: điểm căn chỉnh tối thiểu; 500: kích thước thời gian tối đa.
- **Dispersed Repeats:** Các vùng lặp lại phân tán. Giá trị tham số mặc định: **-f -p -h 3 -l 30** .Giải thích các tham số: -f: tính số lần lặp chuyển tiếp tối đa; -p tính toán các bảng màu tối đa; -h tìm kiếm lặp lại đến khoảng cách hamming nhất định; -l: chỉ định rằng các lần lặp lại phải có độ dài nhất định.

3.3.2 *Chú giải bằng công cụ GeSeq*

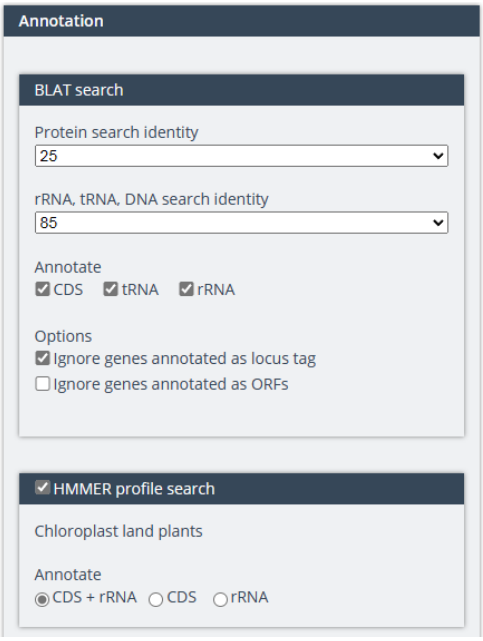
Geseq là một công cụ có nền tảng là web. Địa chỉ để truy cập vào công cụ như sau:

https://chlorobox.mpimp-golm.mpg.de/geseq.html

Để thực hiện chú giải người dùng cần thực hiện lựa chọn như sau:

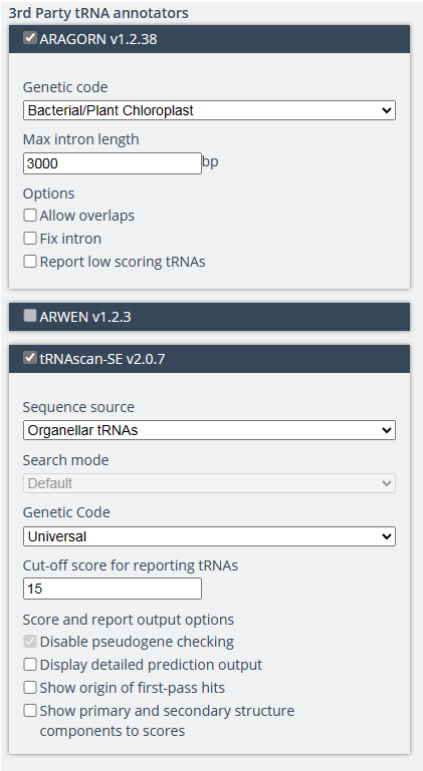


Mục **FASTA files to annotate** chứa những tuỳ chọn cho file fasta khi tải lên công cụ. Ở đây tôi lựa chọn **Circular;** Sequence source là **Plastid (land plants)**; Annotation Options: **Annotated plastid Inverted Repeat (IR)**; **Annote plastid tran-spliced *rps12*** ; Annotation support tôi lựa chọn **Support annotation by Chloe** là sử dụng thuật toán Chloe; Annotation revision lựa chọn

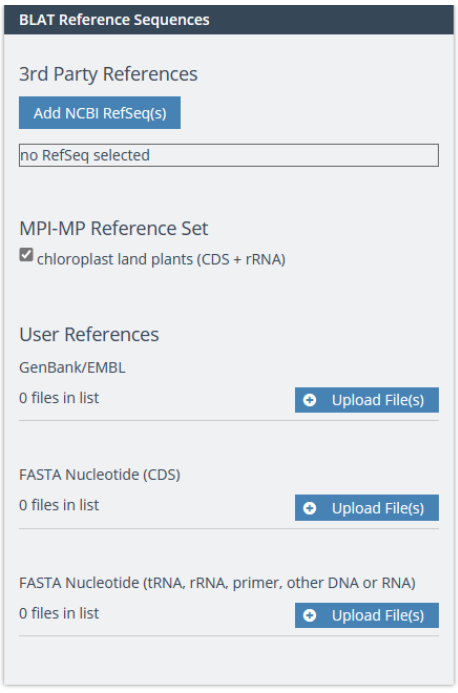


Keep best annotation only.

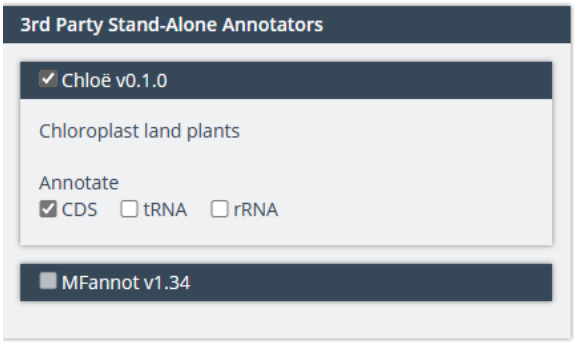
Mục **Annotation** tôi lựa chọn tham số cho 2 phương pháp tìm kiếm **BLAT search** và **HMMER profile search**. Tại mục BLAT các tham số đều được thiết đặt theo mặc định: Protein search identity là 25; rRNA, tRNA, DNA search identity là 85. Mục HMMER profile search chúng tôi lựa chọn **Annotate CDS + rRNA**.



Tại mục **3rd Party tRNA annotators** tôi lựa chọn cả 2 công cụ là **ARAGORN v1.2.38** và **tRNAscan-SE v2.07**



Mục **BLAT Reference Sequenchers** tôi lựa chọn **MPI-MP Reference Set** làm cơ sở dữ liệu tham chiếu



Chúng thôi sử dụng **3rd Party Stand-Alone Annotators** là **Chloe v0.1.0** khi **Annotate CDS**.

Kết quả khi trả về sẽ trả về định dạng file Genbank.

3.3.3 *Chú giải bằng công cụ PGA*

Các thực nghiệm chú giải bằng công cụ PGA được thực hiện trên máy tính hệ điều hành Ubuntu 20.04 phiên bản kernel 5.14, ram 96GB.

Một số công cụ yêu cầu được cài đặt như sau:

- Công cụ BLAST+

Công cụ BLAST đã có sẵn trên trong danh mục package của hệ điều hành Ubuntu vì vậy để cài đặt công cụ người dùng chỉ cần cài đặt thông qua câu lệnh:

sudo apt install ncbi-blast+

Để kiểm tra công cụ đã cài đặt thành công hay chưa người dùng sử dụng câu lệnh:

blastn -version

Kết quả trả về là:

blastn: 2.12.0+

Package: blast 2.12.0, build Mar 8 2022 16:19:08

Như vậy là công cụ BLAST+ đã cài đặt thành công.

Tiếp theo tôi cài đặt công cụ PGA thông qua tổ hợp câu lệnh CMD

git clone https://github.com/quxiaojian/PGA.git

vim ~/.bashrc

export PATH=/home/xxx/PGA:$PATH

source ~/.bashrc

chmod a+rwx PGA.pl

Để đánh giá xem công cụ đã cài đặt thành công hay chưa tôi sử dụng câu lệnh để thực hiện chạy với bộ dữ liệu test

perl PGA.pl -r test/angiosperms/reference -t test/angiosperms/target

Kết quả trả về như sau:

PGA.pl Plastid Genome Annotator
Copyright (C) 2020 Xiao-Jian Qu
Email: quxiaojian@sdnu.edu.cn

2023.09.01 22:23:18 || Begin extracting annotations from reference!
2023.09.01 22:23:19 || Finish extracting annotations from reference!

2023.09.01 22:23:19 || Begin annotating the 1 sequence: Rosa_roxburghii
2023.09.01 22:23:19 || Begin blasting reference to sequence!
Warning: [blastn] Examining 5 or more matches is recommended
Warning: [blastn] Examining 5 or more matches is recommended

Warning: [tblastn] Examining 5 or more matches is recommended
Warning: [tblastn] Examining 5 or more matches is recommended
2023.09.01 22:23:20 || Finish blasting reference to sequence!
2023.09.01 22:23:27 || Finish annotating the 1 sequence: Rosa_roxburghii
2023.09.01 22:23:27 || Finish annotating all sequences and total elapsed time is: 19 seconds!

Như vậy, công cụ PGA đã được cài đặt thành công. Đối với 2 công cụ CPGAVAS2 và GeSeq đều là công cụ trên nền tảng web vì vậy tôi không phải cài đặt và thiết lập môi trường chạy.

perl PGA.pl -r reference -t target -i 1000 -p 40 -q 0.5,2 -o gb -f circular -l warning

Giải thích các giá trị tham số

[-h -help] thông tin trợ giúp.

[-r -reference] tham số này để khai báo vị trí chứ file trình tự Genbank đã formatted của trình tự hệ gen lục lạp tham chiếu. tham số này là bắt buộc

[-t -target] tham số này để khai báo vị trí chứ file trình tự Fasta của trình tự hệ gen lục lạp đích. tham số này là bắt buộc

[-i -ir] tham số khai báo độ lớn tối thiểu của vùng lặp đảo. giá trị mặc định là 1000.

[-p -pidentity] tham số dùng để khai báo độ tương đồng của trình tự gene. Giá trị mặc định là 40. Bất kì PCGs (protein coding gene – gene mã hoá protein) với TBLASTN thấp hơn giá trị này sẽ được thống kê trong file log và sẽ không được chú giải.

[-q -qcoverage] Tham số dùng để khai báo độ bao phủ. Giá trị mặc định là (0.5,2) bất kì PCGs nào có độ bao thủ thấp hơn 0.5 và lớn hơn 2 đều không được chú giải và liệt kê vào file log

[-o -out] tham số dùng để định nghĩa thư mục đầu ra

[-f -form] tham số dùng để định nghĩa loại hệ gene đầu vào circular hoặc linear

[-l -log] tham số dùng để định nghĩa cấu trúc file log. Giá trị mặc định là warningCấu trúc dữ liệu đầu ra của các phần mềm chú giải lục lạp

3.4 *Kết quả thử nghiệm*

Kết quả so sánh đặc điểm 3 phương pháp chú giải CpGAVAS2, GeSeq, PGA và so sánh thời gian chạy trên 10 trình tự đầy đủ của lục lạp loài Cà phê Arabia được tóm lược trong hình 3-1.

Bảng 3-2: So sánh tính năng một số công cụ:

STT	Công cụ	Hệ điều hành	Giao diện người dùng	Thời gian / 1 bộ gen	Có thể chạy dc nhiều hệ gen/ lần	Cách tiếp cận
1	CpGAVAS2	Windows, Linux, Mac	Web	~ 1 hour	không	So sánh hệ gen và cơ sở dữ liệu tham chiếu, blast/blastx
2	GeSeq	Windows, Linux, Mac	Web	~ 60s	có	So sánh hệ gen và cơ sở dữ liệu tham chiếu, mô hình HMM
3	PGA	Windows, Linux, Mac	Command line	19.3s*	có	So sánh cơ sở tham chiếu và hệ gen, blast/blastx

*giá trị trung bình của 10 lần chạy

Nhận thấy rằng, 3 công cụ trên đều có thể sử dụng được đa nền tảng; trong đó 2 công cụ CpGAVAS2 và GeSeq có giao diện người dùng là web, còn PGA là một công cụ độc lập sử dụng cmd. Sau quá trình thiết lập và chạy công cụ thì nhận thấy công cụ PGA cho kết quả nhanh chóng. Chỉ trung bình là 19.2s so với xấp xỉ 60s của công cụ GeSeq. Riêng công cụ CpGAVAS2 cho kết quả trả về rất lâu khoảng trên 1 tiếng cho 1 hệ gen lục lạp. Thêm vào đó công cụ GeSeq và PGA có khả năng phân tích

nhiều hệ gen cùng 1 lúc. Chỉ cần thiết lập 1 file fasta chứa nhiều trình tự hệ gen hoặc thiết lập một thư mục chứa nhiều trình tự hệ gen các công cụ có thể xử lý lần lượt các hệ gen với thời gian nhanh chóng.

Bảng 3-3: Bảng tổng hợp số gene chú giải được bằng 3 công cụ chú giải.

	cpGAVAS2						**GeSeq**						**PGA**					
Mẫu	tRNAs	PCGs	rRNA	Gene có introns	Gene có exon	Tổng số	tRNAs	PCGs	rRNA	Gene có introns	Gene có exon	Tổng số	tRNAs	PCGs	rRNA	Gene có introns	Gene có exon	Tổng số
EF044213.1	29	79	4	15	3	130	29	79	4	15	3	130	29	79	4	15	3	130
KY085909.1	28	80	5	13	3	129	28	80	4	13	3	128	28	80	4	13	3	128
MK342634.1	28	81	5	13	5	132	28	79	4	15	3	129	28	79	4	15	3	129
MK353209.1	30	78	4	14	3	129	30	82	4	14	3	133	30	82	4	14	3	133
MK353212.1	28	80	4	13	5	130	29	82	4	13	4	132	28	82	4	13	4	131
MN370905.1	30	80	5	13	4	132	29	81	4	13	3	130	29	80	4	13	4	130
MN370908.1	28	78	4	15	4	129	29	80	4	14	3	130	30	80	4	14	3	131
MN370924.1	28	78	4	13	4	127	29	81	4	13	3	130	29	81	4	13	3	130
MN894552.1	28	79	4	13	3	127	28	79	4	14	4	129	28	78	4	14	4	128
MN370923.1	29	80	4	16	3	132	30	80	4	15	3	132	29	80	4	16	3	132

Bảng 3-2 tổng hợp số gen chú giải được bằng 3 công cụ chú giải. Bảng số liệu cho thấy số lượng gene chú giải được thông qua các công cụ không có nhiều sự sai khác. Đặc biệt là giữa hai công cụ GeSeq và PGA. Các mẫu EF044213.1, KY085909.1, MK342634.1, MK353209.1, MN370905.1, MN370924.1, MN370923.1 đều có số lượng gen chú giải được là giống nhau giữa hai

công cụ PGA và GeSeq. Các mẫu MK353212.1, MN370908.1, MN894552.1 chỉ có sự sai khác là 1 gene. Đối với công cụ CPGAVAS2 có sự sai khác với kết quả chú giải bằng 2 công cụ PGA và GeSeq 8 trên 10 mẫu, số lượng mẫu sai khác là từ 1-2 gen trên một mẫu. Tuy vậy, số lượng gen này đều nằm ở trong khoảng chấp nhận được. Khi số lượng gen trên hệ gen lục lạp vào khoảng 130 gen.

Để đánh giá thêm về chất lượng chú giải tôi thực hiện đánh giá trên một số tiêu chí như sau:

- Với nhóm gen tRNAs/PCGs không có instrons: số lượng gen chú thích bị thiếu; gen chú thích sai; ranh giới gen chú thích sai; gen chú thích chính xác
- Với nhóm gen tRNAs/PCGs có instrons/ rRNAs: số lượng gene chú thích bị thiếu exon; exon bị chú thích sai; ranh giới exon bị chú thích sai; số lượng gene chú thích exon đúng

Dựa vào bảng kết quả tổng hợp kết quả chú giải theo các tiêu chí (Bảng 3-2), tôi nhận thấy phần mềm PGA và GeSeq có kết quả tương tự nhau. Trong đó GeSeq hơi kém hơn một chút ở một số tiêu chí như số lượng gen tRNAs không có instrons bị thiếu với PGA trung bình là 0.2 còn với GeSeq là 0.4 hoặc ranh giới chú thích PCGs không có introns bị sai của GeSeq là 0.44 còn PGA là 0.4. Kết quả này cũng tương tự như nghiên cứu của Xiao-Jian Qu năm 2019.[28]

Bảng 3-4: Bảng tổng hợp kết quả chú giải theo các tiêu chí

		Coffea Arabica		
		CPGAVAS2	**GeSeq**	**PGA**
tRNAs không có	MG (số lượng gen chú thích bị thiếu)	1.1	0.4	0.2
	WG (gen chú thích sai)	0.3	0.2	0

instrons	WGB (ranh giới gen chú thích sai)	1.7	1.7	0
	CG (gen chú thích chính xác)	15.7	17.2	17.9
tRNAs có instrons	ME số lượng gene chú thích bị thiếu exon	1.2	0	0
	WE: exon bị chú thích sai	0.3	0	0
	WEB: ranh giới exon bị chú thích sai	0.8	0.2	0.15
	CE số lượng gene chú thích exon đúng	7.3	8.5	8.4
PCGs không có introns	MG: (số lượng gen chú thích bị thiếu)	1.15	0.25	0.15
	WG: (gen chú thích sai)	0.1	0	0
	WGB: (ranh giới gen chú thích sai)	0.4	0.44	0.4
	CG: (gen chú thích chính xác)	76.5	77.7	78.5
PCGs có introns	ME: số lượng gene chú thích bị thiếu exon	1.1	0	0
	WE: exon bị chú thích sai	0	0	0
	WEB: ranh giới exon bị chú thích sai	4.3	0.1	0.1
	CE số lượng gene chú thích exon đúng	15.4	18.9	19.2
rRNAs	MG: (số lượng gen chú thích bị thiếu)	0.4	0	0
	WG: (gen chú thích sai)	0	0	0
	WGB: (ranh giới gen chú thích sai)	1.25	0.55	0.5
	CG: (gen chú thích chính xác)	2.6	3.4	3.55

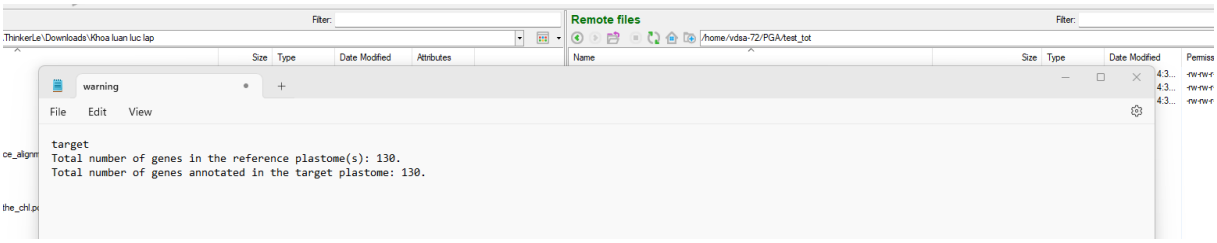
Trong đó, đa phần sai sót của phần mềm PGA đều đến từ việc xác định sai bộ ba mở đầu, nguyên nhân có thể vì thuật toán PGA có xu hương lựa chọn bộ ba ATG xuất hiện đầu tiên là khung đọc mở. GeSeq cho kết quả tốt hơn khi xác định bộ ba mở đầu khi cho phép cả ATG, GTG, ACG và TTG là bộ ba mở đầu.

Khi so sánh PGA và GeSeq nhận thấy rằng PGA yêu cầu người dùng phải tự xác định hệ gen làm tham chiếu. Điều đó dẫn đến việc nếu lựa chọn hệ gen tham chiếu không tốt hoặc hệ gen đó không hoàn chỉnh sẽ ảnh hưởng đến kết quả chú giải. Trong nghiên cứu của mình tác giả Xiao-Jian Qu cũng đã khuyến cáo người sử dụng như sau:

- Người dùng nên kiểm tra cẩn thận hệ gen lục lạp tham chiếu. PGA được đóng gói với một số plastome được chú thích thích hợp, và do đó nó người dùng có thể sử dụng PGA để chú thích lại một hệ gen đã được sử dụng làm tham chiếu.[28]
- Ngươi dùng không nên chú thích các plastome có độ hoàn thiện cao bằng cách sử dụng plastome tham chiếu hoàn chỉnh, vì BLAST có thể chú thích một số gen dư thừa.[28]

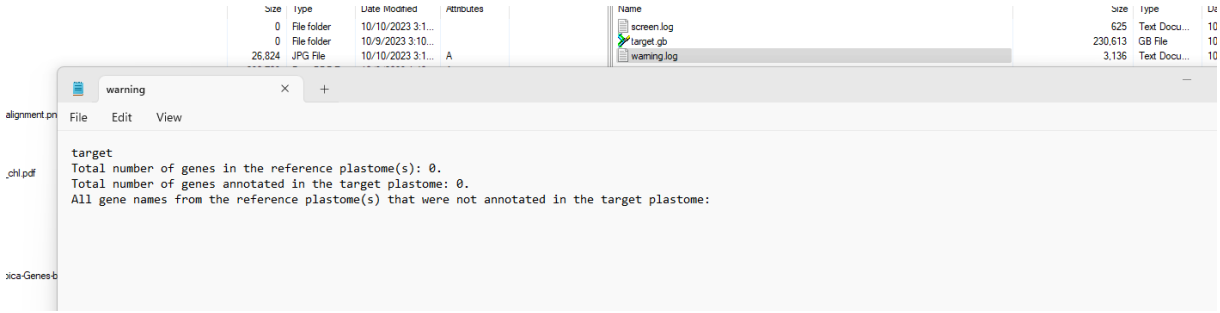
Để chứng minh cho việc này tôi thực hiện so sánh kết quả khi sử dụng 2 hệ gen tham chiếu khác nhau. Hệ gen tốt tôi vẫn lựa chọn hệ gen có mã số *NC_008535.1* đang là hệ gen được xác định là hệ gen tham chiếu trên ngân hàng gen NCBI. Hệ gen kém chất lượng tôi sử dụng hệ gen Coffea-arabica-cultivar-Red-Bourbon-isolate-ID1545 có mã số Genbank ID là *CM061598.1*. Hệ gen *CM061598.1* là một hệ gen chất lượng khá kém khi có kích thước vùng lặp đảo chỉ 3015 nucleotide . Đây là một giá trị kém khi hệ gen lục lạp nói chung vùng lặp đảo có kích thước khoảng 25000 nucleotide.

Kết quả khi sử dụng trình tự tham chiếu *NC_008535.1* tôi thu được kết quả



Ảnh 3-4:Kết quả sử dụng trình tự tham chiếu chất lượng tốt để chú giải hệ gen bằng PGA

Kết quả khi sử dụng trình tự tham chiếu CM061598.1 tôi thu được kết quả



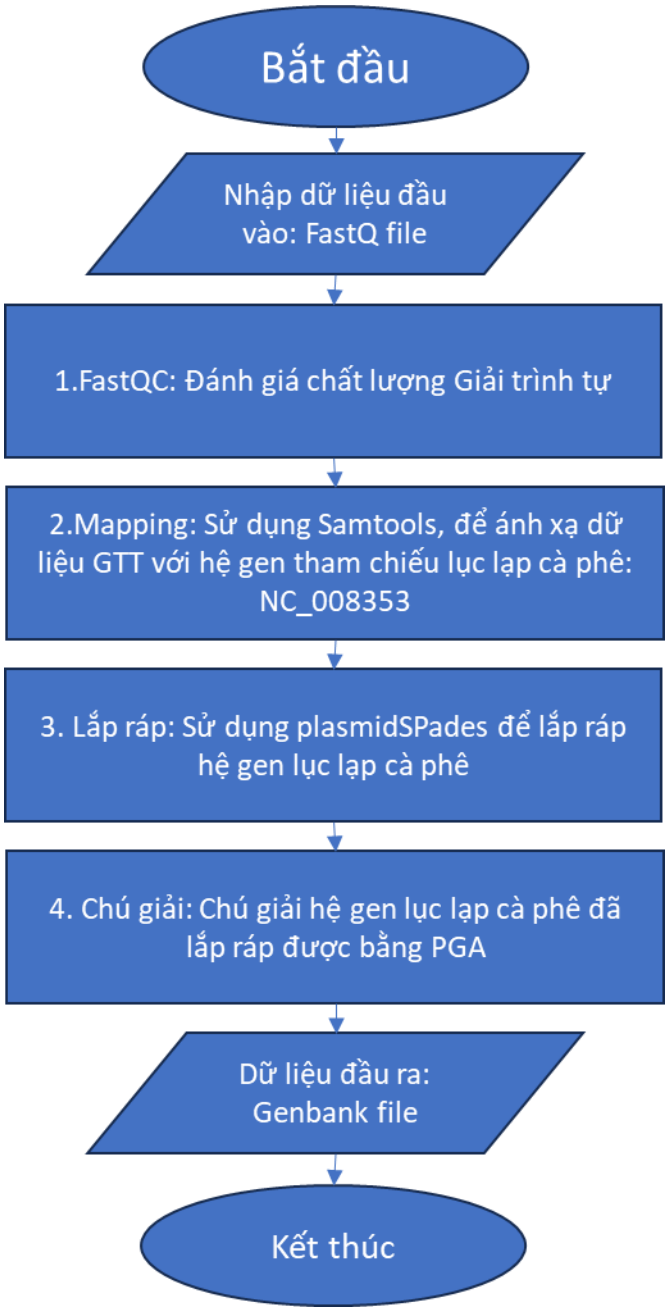
Ảnh 3-5: Kết quả sử dụng trình tự tham chiếu kém chất lượng để chú giải hệ gen bằng PGA

Dễ dàng có thể nhận thấy được rằng khi sử dụng trình tự tham chiếu chất lượng kém tôi không thể chú giải được hệ gen lục lạp, trong khi sử dụng trình tự tham chiếu chất lượng cao tôi thu được đầy đủ 130 gen ở trong hệ gen của mẫu *EF044213.1*.

Như vậy, qua quá trình khảo sát với 10 hệ gen lục lạp của loài cà phê arabica tôi nhận thấy: công cụ GeSeq sử dụng thuật toán Chloe và PGA khi lựa chọn hệ gen tham chiếu tốt sẽ cho kết quả chú giải tốt tương đương nhau. Tuy nhiên, hạn chế của PGA là người dùng phải tự xác định hệ gen tham chiếu đầu vào, do đó với những nhà nghiên cứu chưa có nhiều kinh nghiệm trong tin sinh có thể sẽ lựa chọn hệ gen tham chiếu đầu vào chưa tốt dẫn đến kết quả chú giải chưa tốt. Vì vậy, tôi khuyến cáo với những nhà nghiên cứu nên sử dụng công cụ GeSeq sử dụng thuật toán Chloe và phương pháp HMMER profile để chú giải những hệ gene của mình.

3.5 *Xây dựng quy trình tự động lắp ráp và phân tích hệ gen lục lạp*

Để thuận tiện cho người dùng có thể thực hiện chú giải hệ gen lục lạp từ dữ liệu NGS sử dụng công cụ PGA, tôi đã xây dựng một quy trình tích hợp các công cụ thực hiện các bước của quá trình phân tích hệ gene lục lạp như sau:



Ảnh 3-6: Quy trình tự động lắp ráp trình tự hệ gen lục lạp và chú giải bằng PGA.

Đầu vào là dữ liệu giải trình tự xuất ra từ hệ máy giải trình tự thế hệ 2 – Illumina /MGI/Thermofisher. FastQC[41] được sử dụng để đánh giá chất lượng giải trình tự. Sau quá trình đánh giá chất lượng giải trình tự, công cụ minimap2[42] được sử dụng để ánh xạ dữ liệu giải trình tự với trình tự tham chiếu NC_00835 và Samtools [43] được dùng để xử lý dữ liệu sau ánh xạ. Dữ liệu sau khi ánh xạ sẽ được chuyển thành dạng file fastq để làm đầu vào cho

công cụ lắp ráp plasmidSPAdes[26] - một công cụ phổ biến cho việc lắp ráp hệ gen dạng vòng. Đầu ra của công cụ plasmidSPAdes là trình tự lục lạp đã lắp ráp sẽ được chứa trong file *.scaffold.fasta. Đây sẽ là đầu vào cho công cụ PGA để

```
(base) vdsa-72@vdsa72-desktop:~/SRR$ cat luclap.ubuntu.sh
mkdir 1.fastqc
cd ./1.fastqc
for i in ../*.fastq; do n=$(basename $i .fastq); mkdir $n; fastqc -t 24 --nogroup -o $n $i; done
cd ..
mkdir 2.mapping_reads/
cd ./2.mapping_reads/
cp ../*.fastq .
for i in *.fastq; do n=$(basename $i .fastq) ; minimap2 -ax sr -t 24 ../ref/NC_008535.1.fasta $i > $n.sam; done
for i in *.sam; do n=$(basename $i .sam) ; samtools view -bSh $i > $n.bam; done
for i in *.bam; do n=$(basename $i .bam) ; samtools view -bh -F 4 $i > $n.mapped.bam; done
for i in *.mapped.bam; do n=$(basename $i .mapped.bam); samtools fastq $i > $n.mapped.fastq; done
cd ..
mkdir 3.assembly
cd ./3.assembly/
cp ../2.mapping_reads/*.mapped.fastq .;
mkdir scaffolds_fasta
for i in *.mapped.fastq; do n=$(basename $i .mapped.fastq); spades --plasmid --phred-offset 33 --12 $i -o $n ;
cp ./$n/scaffolds.fasta ./scaffolds_fasta/$n.scaffolds.fasta; done
cd ..
mkdir 4.annotation
cd ./4.annotation/
cp -r ../3.assembly/scaffolds_fasta .
for i in *.mapped.bam; do n=$(basename $i .mapped.bam); perl /home/vdsa-72/PGA/PGA.pl -r /home/vdsa-72/PGA/ref_coffea -t ./scaffolds_fasta ; done
cd ..
(base) vdsa-72@vdsa72-desktop:~/SRR$
```

Ảnh 3-7: Code trong flie linux.ubuntu.sh

chú giải hệ gene cà phê. Dưới đây là đoạn code được viết dưới dạng file sh:

Thực nghiệm với một tập dữ liệu mẫu:

Tôi thực hiện thực nghiệm một một tệp dữ liệu giải trình tự mẫu: SRR25655514 – đây là một tập dữ liệu gốc giải trình tự từ hệ máy HISEQ4000 trên ngân hàng SRA của NCBI. Bộ dữ liệu này bao gồm 21,397,912 đoạn đọc, tổng số base là 2.7 Gigabase; trình tự thuộc về loài cà phê chè (*Coffea Arabica*),

```
(base) vdsa-72@vdsa72-desktop:~/SRR$ tree
.
├── luclap.ubuntu.sh
├── ref
│   └── NC_008535.1.fasta
└── SRR25655514.fastq

1 directory, 3 files
```

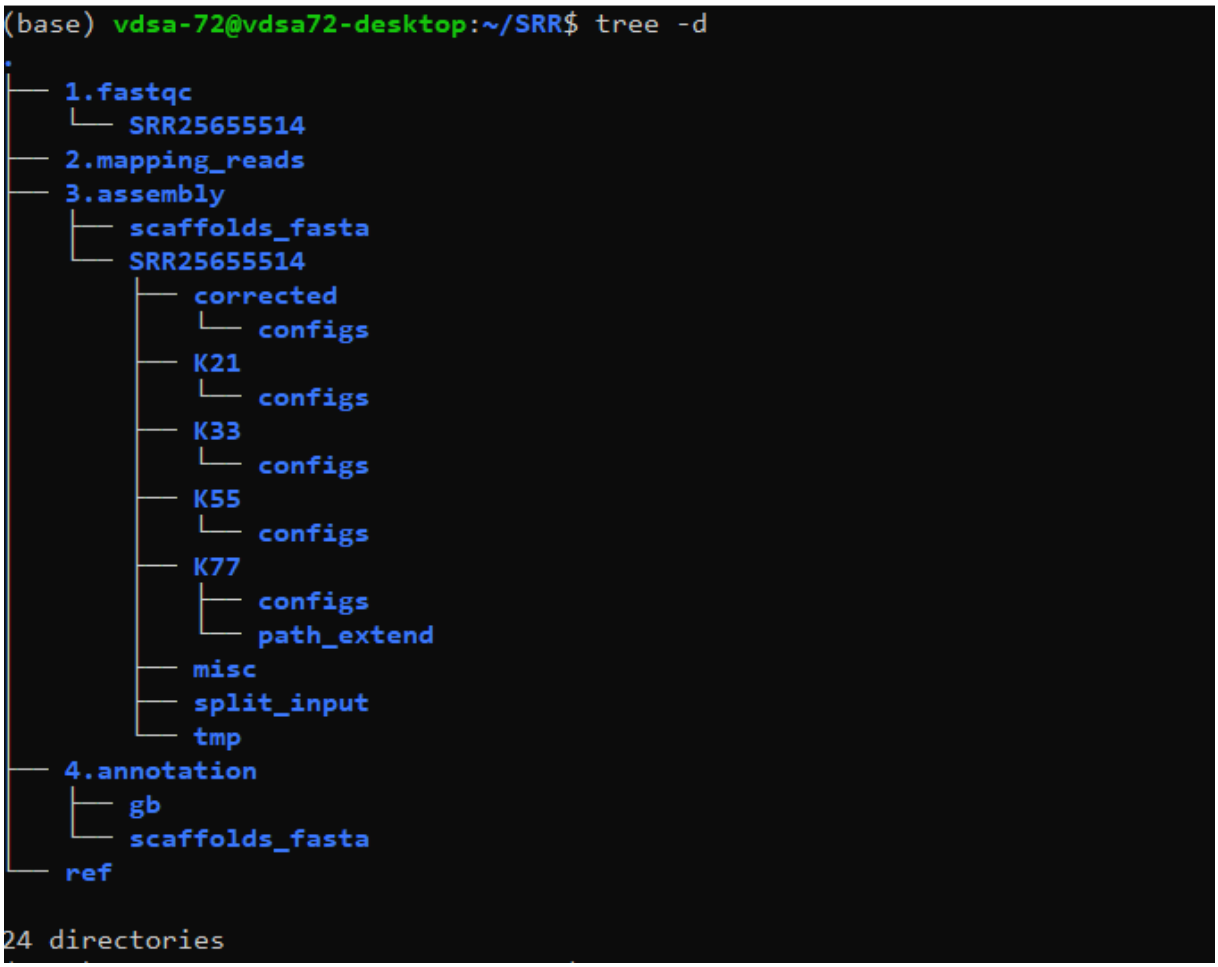
Ảnh 3-8: Chuẩn bị dữ liệu phân tích tự động

được thu tại Uracao, Colombia; dữ liệu được giải trình tự vào năm 2022.

Chúng tôi chuẩn bị file như sau:

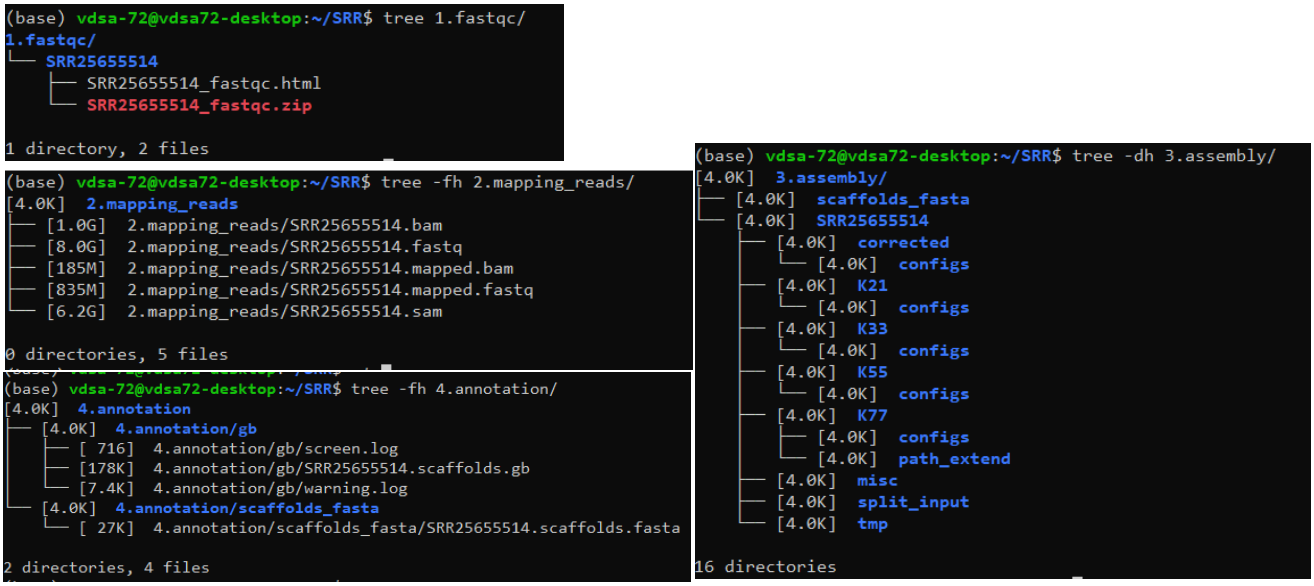
- SRR25655514.fastq là file dữ liệu giải trình tự
- Luclap.ubuntu.sh là file chứa quy trình phân tích
- Thư mục ref chứa trình tự tham chiếu của lục lạp loài cà phê: NC_008535

Sau khi chuẩn bị tôi chạy file ***luclap.ubuntu.sh*** và nhận về kết quả như sau:



Ảnh 3-9: Cây thư mục tạo ra sau quá trình phân tích tự động

Sau khoảng 20 phút cho toàn bộ quá trình tôi thu được kết quả từ đánh giá chất lượng giải trình tự, ánh xạ đoạn đọc vào trình tự tham chiếu, lấy ra những trình tự tương đồng với hệ gen tham chiếu, lắp ráp và chú giải hệ gen.



Ảnh 3-10: Danh sach các file tạo ra sau quá trình phân tích tự động

Như vậy, thông qua luận văn này tôi đã tìm hiểu được các phương pháp chú giải hệ gene lục lạp, có những so sánh ban đầu giữa 3 phương pháp CPGAVAS2/GESEQ và PGA. Chúng có những điểm mạnh điểm yếu riêng, đồng thời thông qua đó đề xuất một quy trình tự động phân tích và chsu giải hệ gene lục lạp cho cây cà phê.

4 CHƯƠNG 4: KẾT LUẬN

Trong luận văn này tôi đã tìm hiểu quy trình phân tích hệ gen lục lạp; khảo sát 03 cơ sở dữ liệu hệ gen lục lạp: NCBI, ChloroplastDB và CpGDBD; Đã thực nghiệm trên tập 10 trình tự đầy đủ của 10 hệ gen lục lạp Cà phê Arabica trên 03 công cụ CPGAVAS2, GeSeq, PGA và đưa ra một số kết luận sau:

1. Công cụ CPGAVAS2 có hạn chế là phân tích lâu, không phân tích được nhiều trình tự cùng một lúc. Có chất lượng chú giải thấp hơn so với GeSeq và PGA. Lý do có khả năng do CPGAVAS2 sử dụng cơ sở dữ liệu tham chiếu cũ.
2. Công cụ GeSeq và PGA cho kết quả chú giải hệ gen lục lạp tương đương nhau về hiệu suất cũng như là độ chính xác.
3. Công cụ PGA có hạn chế là phải sử dụng 1 trình tự có sẵn để làm database phân tích. Do đó: kết quả phân tích sẽ phụ thuộc nhiều vào việc người dùng lựa chọn trình tự để so sánh.
4. GeSeq khi sử dụng kết hợp cả phương pháp BLAST và mô hình HMM; đồng thời sử dụng cơ sở dữ liệu tham chiếu, thuật toán Chloe cho kết quả phân tích tốt. Không yêu cầu người dùng cung cấp trình tự tham chiếu. Do đó, phù hợp với những người dùng ít hiểu biết về tin sinh học, chú giải những loài mới hoặc loài có ít dữ liệu tham chiếu.
5. Xây dựng quy trình tự động từ khâu kiểm tra chất lượng giải trình tự, xử lý dữ liệu giải trình tự, lắp ráp và chú giải cho riêng dữ liệu giải trình tự lục lạp cây cà phê.

5 KIẾN NGHỊ VÀ GIẢI PHÁP

1. Trong luận văn sử dụng những trình tự để so sánh có chất lượng tương đối tốt, cùng một loài. Do đó, kết quả có thể chưa được phổ quát.
2. Cần mở rộng tập dữ liệu so sánh cho nhiều đối tượng hơn.

6 TÀI LIỆU THAM KHẢO

[1] International Human Genome Sequencing Consortium, "Finishing the euchromatic sequence of the human genome," *Nature*, vol. 431, no. 7011, Art. no. 7011, Oct. 2004, doi: 10.1038/nature03001.

[2] S. Sato, Y. Nakamura, T. Kaneko, E. Asamizu, and S. Tabata, "Complete structure of the chloroplast genome of Arabidopsis thaliana," *DNA Res. Int. J. Rapid Publ. Rep. Genes Genomes*, vol. 6, no. 5, pp. 283–290, Oct. 1999, doi: 10.1093/dnares/6.5.283.

[3] The Arabidopsis Genome Initiative, "Analysis of the genome sequence of the flowering plant Arabidopsis thaliana," *Nature*, vol. 408, no. 6814, Art. no. 6814, Dec. 2000, doi: 10.1038/35048692.

[4] P. K. Gupta and Y. Xu, "Genomics of Major Crops and Model Plant Species," *Int. J. Plant Genomics*, vol. 2008, p. 171928, 2008, doi: 10.1155/2008/171928.

[5] R. A. Marks, S. Hotaling, P. B. Frandsen, and R. VanBuren, "Representation and participation across 20 years of plant genome sequencing," *Nat. Plants*, vol. 7, no. 12, Art. no. 12, Dec. 2021, doi: 10.1038/s41477-021-01031-8.

[6] Y. Sun, L. Shang, Q.-H. Zhu, L. Fan, and L. Guo, "Twenty years of plant genome sequencing: achievements and challenges," *Trends Plant Sci.*, vol. 27, no. 4, pp. 391–401, Apr. 2022, doi: 10.1016/j.tplants.2021.10.006.

[7] J.-C. Charr *et al.*, "Complex evolutionary history of coffees revealed by full plastid genomes and 28,800 nuclear SNP analyses, with particular emphasis on Coffea canephora (Robusta coffee)," *Mol. Phylogenet. Evol.*, vol. 151, p. 106906, Oct. 2020, doi: 10.1016/j.ympev.2020.106906.

[8] Y. Bawin *et al.*, "Phylogenomic analysis clarifies the evolutionary origin of Coffea arabica," *J. Syst. Evol.*, vol. 59, no. 5, pp. 953–963, 2021, doi: 10.1111/jse.12694.

[9] L. T. T. Hiền, H. D. Boer, V. Manzanilla, H. V. Huân, and N. V. Hải, "Genome sequencing in plants and the genus Panax L.," *Vietnam J. Biotechnol.*, vol. 14, no. 1, Art. no. 1, Mar. 2016, doi: 10.15625/1811-4989/14/1/9286.

[10] L. T. Hương *et al.*, "Application of DNA barcodes in identification of ginseng samples in the genus Panax L.," *Vietnam J. Biotechnol.*, vol. 15, no. 1, Art. no. 1, 2017, doi: 10.15625/1811-4989/15/1/12321.

[11] V. Manzanilla, A. Kool, L. Nguyen Nhat, H. Nong Van, H. Le Thi Thu, and H. J. de Boer, "Phylogenomics and barcoding of Panax: toward the identification of ginseng species," *BMC Evol. Biol.*, vol. 18, no. 1, p. 44, Apr. 2018, doi: 10.1186/s12862-018-1160-y.

[12] "The complete nucleotide sequence of the coffee (Coffea arabica L.) chloroplast genome: organization and implications for biotechnology and phylogenetic relationships amongst angiosperms - PMC." Accessed: Oct. 10, 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3473179/

[13] E. L. van Dijk, H. Auger, Y. Jaszczyszyn, and C. Thermes, "Ten years of next-generation sequencing technology," *Trends Genet. TIG*, vol. 30, no. 9, pp. 418–426, Sep. 2014, doi: 10.1016/j.tig.2014.07.001.

[14] "Overview of Next Generation Sequencing Technologies - PMC." Accessed: Oct. 10, 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6020069/

[15] M. T. Pervez, M. J. ul Hasnain, S. H. Abbas, M. F. Moustafa, N. Aslam, and S. S. M. Shah, "A Comprehensive Review of Performance of Next-Generation Sequencing Platforms," *BioMed Res. Int.*, vol. 2022, p. 3457806, Sep. 2022, doi: 10.1155/2022/3457806.

[16] J. F. Thompson and P. M. Milos, "The properties and applications of single-molecule DNA sequencing," *Genome Biol.*, vol. 12, no. 2, p. 217, Feb. 2011, doi: 10.1186/gb-2011-12-2-217.

[17] E. Check Hayden, "Nanopore genome sequencer makes its debut," *Nature*, Feb. 2012, doi: 10.1038/nature.2012.10051.

[18] P. J. A. Cock, C. J. Fields, N. Goto, M. L. Heuer, and P. M. Rice, "The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants," *Nucleic Acids Res.*, vol. 38, no. 6, pp. 1767–1771, Apr. 2010, doi: 10.1093/nar/gkp1137.

[19] "GenBank Sample Record." Accessed: Oct. 10, 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html

[20] J. D. Palmer and D. B. Stein, "Conservation of chloroplast genome structure among vascular plants," *Curr. Genet.*, vol. 10, no. 11, pp. 823–833, Jul. 1986, doi: 10.1007/BF00418529.

[21] S. Izan, D. Esselink, R. G. F. Visser, M. J. M. Smulders, and T. Borm, "De Novo Assembly of Complete Chloroplast Genomes from Non-model Species Based on a K-mer Frequency-Based Selection of Chloroplast Reads from Total DNA Sequences," *Front. Plant Sci.*, vol. 8, 2017, Accessed: Oct. 10, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpls.2017.01271

[22] "Ultra-barcoding in cacao (Theobroma spp.; Malvaceae) using whole chloroplast genomes and nuclear ribosomal DNA - PubMed." Accessed: Oct. 10, 2023. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/22301895/

[23] F. T. Bakker *et al.*, “Herbarium genomics: plastome sequence assembly from a range of herbarium specimens using an Iterative Organelle Genome Assembly pipeline,” *Biol. J. Linn. Soc.*, vol. 117, no. 1, pp. 33–43, Jan. 2016, doi: 10.1111/bij.12642.

[24] “fast, lock-free approach for efficient parallel counting of occurrences of k-mers | Bioinformatics | Oxford Academic.” Accessed: Oct. 10, 2023. [Online]. Available: https://academic.oup.com/bioinformatics/article/27/6/764/234905

[25] “Alignment-free genetic sequence comparisons: a review of recent approaches by word analysis | Briefings in Bioinformatics | Oxford Academic.” Accessed: Oct. 10, 2023. [Online]. Available: https://academic.oup.com/bib/article/15/6/890/177681

[26] D. Antipov, N. Hartwick, M. Shen, M. Raiko, A. Lapidus, and P. A. Pevzner, “plasmidSPAdes: assembling plasmids from whole genome sequencing data,” *Bioinformatics*, vol. 32, no. 22, pp. 3380–3387, Nov. 2016, doi: 10.1093/bioinformatics/btw493.

[27] “Recycler: an algorithm for detecting plasmids from de novo assembly graphs | Bioinformatics | Oxford Academic.” Accessed: Oct. 10, 2023. [Online]. Available: https://academic.oup.com/bioinformatics/article/33/4/475/2623362

[28] X.-J. Qu, M. J. Moore, D.-Z. Li, and T.-S. Yi, “PGA: a software package for rapid, accurate, and flexible batch annotation of plastomes,” *Plant Methods*, vol. 15, no. 1, p. 50, May 2019, doi: 10.1186/s13007-019-0435-7.

[29] L. Shi *et al.*, “CPGAVAS2, an integrated plastome sequence annotator and analyzer,” *Nucleic Acids Res.*, vol. 47, no. W1, pp. W65–W73, Jul. 2019, doi: 10.1093/nar/gkz345.

[30] “Evaluation of chloroplast genome annotation tools and application to analysis of the evolution of coffee species - PMC.” Accessed: Oct. 10, 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6561552/

[31] S. McGinnis and T. L. Madden, “BLAST: at the core of a powerful and diverse set of sequence analysis tools,” *Nucleic Acids Res.*, vol. 32, no. Web Server issue, pp. W20–W25, Jul. 2004, doi: 10.1093/nar/gkh435.

[32] “Automatic annotation of organellar genomes with DOGMA | Bioinformatics | Oxford Academic.” Accessed: Oct. 10, 2023. [Online]. Available: https://academic.oup.com/bioinformatics/article/20/17/3252/186263

[33] “Verdant: automated annotation, alignment and phylogenetic analysis of whole chloroplast genomes | Bioinformatics | Oxford Academic.” Accessed:

Oct. 26, 2023. [Online]. Available: https://academic.oup.com/bioinformatics/article/33/1/130/2525692

[34] C. Liu *et al.*, "CpGAVAS, an integrated web server for the annotation, visualization, analysis, and GenBank submission of completely sequenced chloroplast genome sequences," *BMC Genomics*, vol. 13, no. 1, p. 715, Dec. 2012, doi: 10.1186/1471-2164-13-715.

[35] M. Madera and J. Gough, "A comparison of profile hidden Markov model procedures for remote homology detection," *Nucleic Acids Res.*, vol. 30, no. 19, pp. 4321–4328, Oct. 2002.

[36] M. Tillich *et al.*, "GeSeq – versatile and accurate annotation of organelle genomes," *Nucleic Acids Res.*, vol. 45, no. W1, pp. W6–W11, Jul. 2017, doi: 10.1093/nar/gkx391.

[37] "Assembly, annotation and analysis of chloroplast genomes — the UWA Profiles and Research Repository." Accessed: Oct. 10, 2023. [Online]. Available: https://research-repository.uwa.edu.au/en/publications/assembly-annotation-and-analysis-of-chloroplast-genomes

[38] J. Zhou *et al.*, "Chloroplast genomes in Populus (Salicaceae): comparisons from an intensively sampled genus reveal dynamic patterns of evolution," *Sci. Rep.*, vol. 11, no. 1, Art. no. 1, May 2021, doi: 10.1038/s41598-021-88160-4.

[39] "ChloroplastDB: the Chloroplast Genome Database - PubMed." Accessed: Oct. 10, 2023. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/16381961/

[40] "CpGDB : A Comprehensive Database of Chloroplast Genomes - PMC." Accessed: Oct. 10, 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7196173/

[41] "Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data." Accessed: Oct. 24, 2023. [Online]. Available: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[42] "Minimap2: pairwise alignment for nucleotide sequences | Bioinformatics | Oxford Academic." Accessed: Oct. 24, 2023. [Online]. Available: https://academic.oup.com/bioinformatics/article/34/18/3094/4994778

[43] H. Li *et al.*, "The Sequence Alignment/Map format and SAMtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, Aug. 2009, doi: 10.1093/bioinformatics/btp352.