Lắp ráp và chú giải hệ gen vi tảo biển dị dưỡng *Schizochytrium mangrovei* PQ6 của Việt Nam

Nguyễn Văn Lâm, Phạm Quang Huy, Nguyễn Quốc Đại, Hoàng Minh Hiền, Đặng Diễm Hồng, Lê Văn Sơn, Chu Hoàng Hà, Trương Nam Hải, Nguyễn Cường*

Viện Công nghệ sinh học, Viện Hàn lâm KH&CN Việt Nam

Ngày nhận bài 13.3.2015, ngày chuyển phản biện 18.3.2015, ngày nhận phản biện 20.4.2015, ngày chấp nhận đăng 24.4.2015

Vi tảo biển dị dưỡng *Schizochytrium mangrovei* PQ6 là loài/chủng đặc hữu của Việt Nam được phân lập tại huyện đảo Phú Quốc, tỉnh Kiên Giang. Loài vi tảo này có tiềm năng ứng dụng rất lớn để tổng hợp axit béo không bão hòa đa nối đôi (Polyunsaturated Fatty Acids-PUFAs), đặc biệt là axit docosahexaenoic (DHA; C226 n-3). Nhằm mục đích xây dựng và chú giải hệ gen, từ đó tìm ra những con đường trao đổi chất có liên quan đến quá trình tổng hợp axit béo và DHA, đề tài đã tiến hành giải trình tự, lắp ráp *de novo* và chú giải hệ gen loài vi tảo *Schizochytrium mangrovei* PQ6. Sử dụng thiết bị đọc trình tự thế hệ mới Illumina Miseq để giải trình tự ADN hệ gen của loài vi tảo biển này thu được 54200442 đoạn trình tự ngắn (reads). Trình tự thô được tiền xử lý rồi đem lắp ráp *de novo* và thu được hệ gen có kích thước khoảng 59,26 Mb với 2601 đoạn trình tự dài contigs với N50 là 59034 bp. Sau khi dự đoán gen, đã thu được 4128 mô hình gen, trong đó có 3970 gen có sự tương đồng trên cơ sở dữ liệu NR (non-reduntdant) và 2383 gen được chú giải chức năng trên cơ sở dữ liệu GO (Gene Othology). Trong số 16 gen liên quan đến quá trình tổng hợp axit béo tìm được, đã phát hiện có 4 gen liên quan đến quá trình sinh tổng hợp PUFAs. Các kết quả thu được nêu trên mới chỉ là những thông tin ban đầu về hệ gen *S. mangrovei* PQ6 nhưng có ý nghĩa quan trọng cho các nghiên cứu tiếp theo, đặc biệt liên quan đến quá trình tổng hợp PUFAs và DHA.

***Từ khóa:* DHA, giải trình tự thế hệ mới, lắp ráp de novo, PUFAs, Schizochytrium mangrovei PQ6, vi tảo biển dị dưỡng.**

Chỉ số phân loại 1.6

Đặt vấn đề

Schizochytrium mangrovei PQ6 được phân lập ở huyện đảo Phú Quốc, tỉnh Kiên Giang năm 2006-2008 [1]. Đây là chủng vi tảo dị dưỡng đặc hữu của Việt Nam thuộc chi *Schizochytrium,* là một trong những nguồn sản xuất DHA tiềm năng [2], với hàm lượng lipit tổng số có thể đạt đến 70% so với sinh khối khô tế bào [3]. Sinh khối *Schizochytrium* được sử dụng với nhiều mục đích khác nhau như: bổ sung vào thức ăn giúp tăng hàm lượng DHA ở tôm và luân trùng [4, 5]; thay thế dầu cá giúp tăng khả năng miễn dịch, chất lượng thương phẩm và trọng lượng của cá hồi [6-8]. Ngoài ra, sinh khối của loại tảo này cũng còn được sử dụng như là một chất kích thích sinh sản nhân tạo cho hải sâm nhiệt đới *Holothuria scabra* [9]. Hơn nữa, sinh khối *Schizochytrium* còn được ứng dụng trong sản xuất thực phẩm và thực phẩm chức năng cho con người và động vật nuôi. Nhiều nghiên cứu đã chỉ ra rằng, bổ sung hàm lượng DHA được chiết xuất từ tảo vào trong thực phẩm như trứng, thịt gà, sữa sẽ giúp làm giảm nguy cơ mắc một số bệnh liên quan đến tim mạch ở người khi sử dụng các thực phẩm nêu trên [10, 11]... Hiện nay, ở Việt Nam đã có một số nghiên cứu về đặc điểm sinh học, sinh hóa; tách chiết các axit béo PUFAs; biodiessel và một số chất có hoạt tính sinh học từ sinh khối chủng PQ6 được công bố [12-15].

Ngày nay, với công nghệ giải trình tự mới (Next generation sequencing - NGS) được phát triển mạnh mẽ, cho phép đọc trình tự toàn bộ hệ gen với độ bao phủ rất cao, giúp tiết kiệm đáng kể chi phí và thời gian

* *Tác giả chính: Tel: 0916110333, Email: cuongnguyen@ibt.ac.vn*

GENOME ASSEMBLY AND ANNOTATION OF THE HETEROTROPHIC MICROALGA SCHIZOCHYTRIUM MANGROVEI PQ6 IN VIET NAM

Summary

Heterotrophic marine microalga *Schizochytrium mangrovei* PQ6 is an endemic strain of Vietnam which is isolated in Phu Quoc island, Kien Giang Province, Viet Nam. This species has a great potential of applications to synthesise polyunsaturated fatty acids (PUFAs), especially docosahexaenoic acid (DHA; C226 n-3). This study aims at constructing and annotating the genome, and then finding unsaturated fatty acid biosynthesis pathways. The whole DNA of *S. mangrovei* PQ6 has been sequenced by Illumina Miseq® to obtain 54200442 reads. After preprocessing and *de novo* assembly, the authors have obtained a draft genome of *S. mangrovei* PQ6 containing 2601 contigs with the length arranging from 501 bp to 320542 bp (~59.26 Mb), and with N50 as 59034 bp and the GC content of 45.19%. Applying gene prediction methods, the authors have been able to identify 4128 gene models. The results have shown that 3970 genes and 2383 genes have been functionally annotated on NCBI-NR database and Gene Ontology database respectively. Moreover, the authors have also found 16 PUFA-related genes. These results is only the initial information about *S. mangrovei* PQ6 genome, but they are really improtant for future studies, especially for the studies relating to PUFA and DHA systhesis.

Keywords: ***de novo assembly, DHA, heterotrophic marine microalga, next generation sequencing, PUFAs, Schizochytrium mangrovei PQ6.***

Classification number 1.6

so với công nghệ giải trình tự Sanger trước đây. Chính đặc điểm quan trọng này đã giúp cho NGS trở thành một công cụ không thể thiếu trong các nghiên cứu về toàn bộ hệ gen, hay phân tích hệ gen phiên mã của một sinh vật nào đó. Vì vậy, chúng tôi đã tiến hành nghiên cứu ứng dụng công nghệ giải trình tự thế hệ mới để giải trình tự, lắp ráp và chú giải hệ gen của chủng vi tảo biển *S. mangrovei* PQ6 nhằm làm sáng tỏ con đường sinh tổng hợp axit béo, đặc biệt là những con đường sinh tổng hợp PUFAs của chủng vi tảo đặc hữu này.

Dữ liệu và phương pháp nghiên cứu

Dữ liệu nghiên cứu

Chủng vi tảo biển dị dưỡng *Schizochytrium mangrovei* PQ6 được phân lập tại đảo Phú Quốc, tỉnh Kiên Giang (2006-2008), được lưu giữ trong bộ sưu tập giống của Phòng Công nghệ tảo, Viện Công nghệ sinh học. Chúng tôi sử dụng dịch vụ đọc trình tự bằng máy Ilumina Miseq® của Công ty cổ phần Vật tư khoa học Biomedic (Mỹ).

Phương pháp nghiên cứu

Quy trình nghiên cứu được thực hiện theo các bước sau: đánh giá và tiền xử lý dữ liệu; lắp ráp *de novo*; dự đoán và chú giải chức năng gen.

Đánh giá và tiền xử lý dữ liệu: dữ liệu trình tự thu được từ thiết bị đọc trình tự thế hệ mới được đánh giá và tiền xử lý bằng cách sàng lọc các trình tự có độ dài ngắn (nhỏ hơn 101 bp), chất lượng thấp (QC < 30) với phần mềm FastQC[1] và Trimmomatic [16].

Lắp ráp de novo và đánh giá chất lượng lắp ráp: các trình tự đã được xử lý được đưa vào lắp ráp *de novo* để thu được các đoạn trình tự dài liên tục gọi là contigs bằng các phần mềm Velvet [17] và Edena [18], với các tham số đã được tối ưu. Chất lượng lắp ráp được đánh giá dựa trên các thông số như: kích thước hệ gen, N50 và hiệu suất lắp ráp bằng phần mềm Quast [19] và Bowtie2 [20]. Các kết quả lắp ráp được hợp nhất lại với nhau theo nguyên tắc giữ lại các contigs có độ tương đồng cao bằng phần mềm MIX [21].

Dự đoán và chú giải chức năng gen: tập các contigs thu được từ quá trình lắp ráp sẽ được đưa vào dự đoán gen dựa trên hai phương pháp: có căn cứ (evidence based) và *ab-initio*. Đối với phương pháp evidence based, dữ liệu trình tự protein của các loài tương tự (bảng 1) được sử dụng trong quá trình dự đoán mô

[1]http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

hình gen. Các phần mềm được sử dụng là Exonerate [22] và Blast [23]. Trong khi đó, phương pháp *ab-initio* sử dụng các mô hình gen đã biết của các loài để dự đoán mô hình gen của vi tảo biển dị dưỡng *Schizochytrium mangrovei* PQ6. Các phần mềm được sử dụng là Augustus [24] và SNAP [25]. Các tập mô hình gen được dự đoán sẽ được hợp nhất bằng phần mềm Maker [26].

Bảng 1: thông tin về các loài đã được sử dụng trong quá trình dự đoán gen

Loài	Số lượng trình tự proteins	Nguồn
Chlamydomonas reinhardtii	16709	JGI[a]
Aureococcus anophagefferens	11501	JGI[a]
Ectocarpus siliculosus	16753	NCBI[b]
Hyaloperonospora arabidopsidis	14321	VBI[c]
Phaeodactylum tricornutum	10025	JGI[a]
Phytophthora infestans	18140	BROAD[d]
Phytophthora ramorum	15743	JGI[a]
Phytophthora sojae	26584	JGI[a]
Pythium ultimum	15322	MSU[e]
Thalassiosira pseudonana	11390	JGI[a]

a) http://www.jgi.doe.gov; b)http://www.ncbi.nlm.nih.gov/;
c) http://vmd.vbi.vt.edu/; d) http://www.broadinstitute.org/;
e) http://pythium.plantbiology.msu.edu/

Tập gen sau khi hợp nhất được chú giải bằng cách so sánh lên cơ sở dữ liệu NR (non-redundant) trên NCBI (National center for biotechnology information) bằng phần mềm Blast [23] với evalue là 10^{-6}. Những gen có sự tương đồng trên NR tiếp tục được chú giải chức năng trên cơ sở dữ liệu Gene Ontology [27] và tìm ra các con đường sinh tổng hợp axit béo trên cơ sơ dữ liệu KEGG [28] (Kyoto encyclopedia of genes and genomes) bằng việc sử dụng phần mềm Blast2GO [29].

Kết quả và thảo luận

Đánh giá và tiền xử lý dữ liệu

Sử dụng thiết bị đọc trình tự thế hệ mới, chúng tôi đã thu được tổng số trình tự thô là 54200442. Sau quá trình tiền xử lý, thu được 51767972 (chiếm 95,5%) trình tự có chất lượng tốt (QC > 30) có độ dài là 101 bp (bảng 2).

Bảng 2: bảng tổng hợp chất lượng dữ liệu trước và sau khi tinh sạch

	Số đoạn trình tự	Độ dài (bp)	%GC
Dữ liệu thô	54200442	30-101	43
Sau tinh sạch	51767972 (95,5%)	101	43

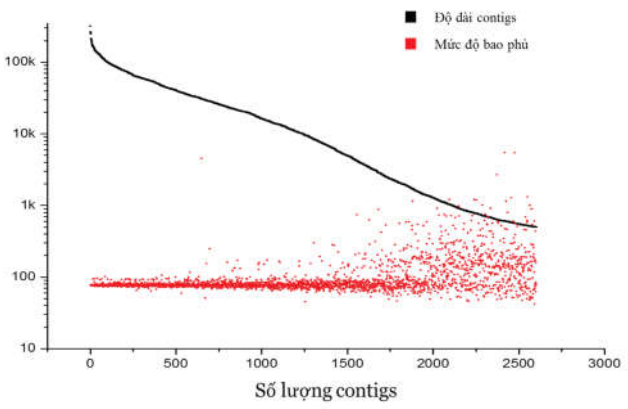
Lắp ráp de novo và đánh giá chất lượng lắp ráp

Khi sử dụng hai phần mềm Velvet và Edena, hai tập contigs được lắp ráp *de novo* có các chỉ số khá tương đồng nhau như sau: kích thước hệ gen lần lượt là 60,1 Mb và 60,9 Mb; N50 là 47528 bp và 35414 bp và hầu hết các đoạn trình tự đều được sử dụng trong quá trình lắp ráp với hiệu suất tương ứng là 97,48% và 99,25% (bảng 3).

Bảng 3: kết quả lắp ráp de novo hệ gen chủng PQ6

Các chỉ số	Velvet	Edena	MIX
Kích thước hệ gen (mb)	60,1	60,9	59,29
Số lượng contigs	7710	7181	2601
Contigs lớn nhất (bp)	193575	189386	320452
Contigs ngắn nhất (bp)	121	151	501
Số lượng contigs ≥ 100 kb	67	28	101
Số lượng contigs ≥ 20 kb	980	1050	912
Số lượng contigs ≥ 1 kb	2524	3399	2110
N50	47528	35,14	59034
Tỷ lệ G + C (%)	45,19	45,10	45,19
Tỷ lệ trình tự ánh xạ ngược lại (%)	97,48	99,25	99,00

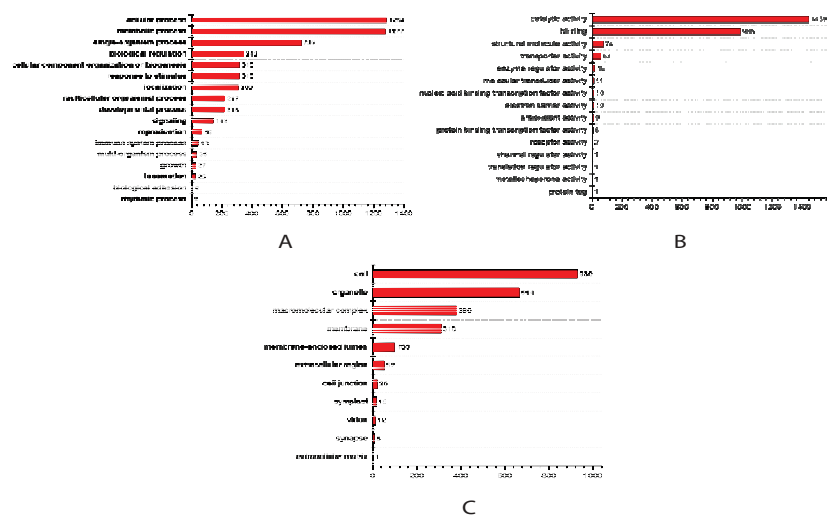
Sau khi hợp nhất các kết quả nghiên cứu thu được với hai phần mềm nêu trên, chúng tôi thu được tập contigs gồm 2601 contigs có độ dài từ 501 bp đến 320452 bp; kích thước hệ gen là 59,26 Mb; N50 là 59034 bp và hiệu suất lắp ráp là 99% (bảng 3). Biểu diễn độ dài contigs và mức độ bao phủ tương ứng với từng cotigs được thể hiện ở hình 1. Kết quả trên hình 1 cho thấy, hầu hết các contigs đều có độ bao phủ (coverage) từ 80x trở lên. Điều này chứng tỏ hệ gen có chất lượng lắp ráp tốt, độ tin cậy cao và đủ điều kiện để thực hiện bước tiếp theo - bước dự đoán và chú giải chức năng gen.



Hình 1: biểu đồ biểu diễn độ dài contigs và mức độ bao phủ tương ứng với từng contigs

Kết quả dự đoán và chú giải chức năng gen

Kết quả dự đoán gen đã thu được 4128 gen, trong đó có 3970 gen tương đồng với ít nhất một trình tự trên cơ sở dữ liệu NR, mức độ tương đồng từ 40-100%. Chúng tôi tiếp tục so sánh 3970 gen nêu trên với cơ sở dữ liệu Gene Ontology [27] bằng phần mềm Blast2GO [29]. Kết quả thu được cho thấy, có 2383 gen được chú giải trên Gene Ontology [27] (hình 2). Ngoài ra, khi so sánh với cơ sở dữ liệu KEGG [28], chúng tôi đã phát hiện có 16 gen mã hóa cho 12 enzyme tham gia vào 20 giai đoạn khác nhau của quá trình sinh tổng hợp axit béo, trong đó có 4 gen mã hóa cho 4 enzym (ec:4.2.1.17; ec:1.1.1.100; ec:1.3.1.38; ec:2.3.1.16) liên quan trực tiếp đến quá trình sinh tổng hợp PUFAs.



Hình 2: biểu đồ số lượng gen có mã GO tham gia

[A: quá trình sinh học (Biological process); B: chức năng phân tử (Molecular function); C: thành phần tế bào (Cellular component)]

Kết luận

Từ các kết quả nghiên cứu được trình bày ở trên, chúng tôi rút ra một số kết luận như sau:

Đề tài đã lắp ráp thành công hệ gen của loài vi tảo biển dị dưỡng *Schizochytrium mangrovei* PQ6 với kích thước hệ gen khoảng 59,26 Mb. Dự đoán được 4128 gen, trong đó có 3970 đã được chú giải trên cơ sở dữ liệu NR của NCBI và có 2383 gen trong 3970 gen có liên quan đến ít nhất một thuật ngữ GO trên trên cơ sở dữ liệu Gene Ontology [27]. Đặc biệt hơn, khi so sánh với cơ sở dữ liệu KEGG [28], chúng tôi đã phát hiện ra 16 gen có liên quan đến quá trình sinh tổng hợp axit béo, trong đó có 4 gen liên quan trực tiếp đến quá trình sinh tổng hợp axit béo không bão hòa (PUFAs). Đây mới chỉ là những kết quả ban đầu giúp chúng ta có cái nhìn tổng quan về hệ gen của loài vi tảo *Schizochytrium mangrovei* PQ6, cung cấp những cơ sở khoa học cho các nghiên cứu sâu hơn về hệ gen của loài tảo này cũng như những thông tin di truyền có liên quan đến các con đường sinh tổng hợp axit béo.

Tài liệu tham khảo

[1] Đặng Diêm Hồng, Hoàng Lan Anh, Ngô Thị Hoài Thu (2008), "Phân lập được vi tảo biển dị dưỡng *Schizochytrium* giàu DHA ở vùng biển huyện đảo Phú Quốc", *Tạp chí Sinh học*, **30(2)**, pp.50-55.

[2] Lewis T.E, Nichols P.D, McMeekin T.A (1999), "The biotechnological potential of thraustochytrids", *Mar Biotechnol (NY)*, **1(6)**, pp.580-587.

[3] Yaguchi T, Tanaka S, Yokochi T, Nakahara T, Higashihara T (1997), "Production of high yields of docosahexaenoic acid by *Schizochytrium* sp. strain SR21", *Journal of the American Oil Chemists' Society*, **74(11)**, pp.1431-1434.

[4] Barclay W, Abril R, Abril P, Weaver C, Ashford A (1998), "Production of docosahexaenoic acid from microalgae and its benefits for use in animal feeds", *World Rev Nutr Diet*, **83**, pp.61-76.

[5] Estudillo-del Castillo C, Gapasin R.S, Leaño E.M (2009), "Enrichment potential of HUFA-rich thraustochytrid *Schizochytrium* mangrovei for the rotifer Brachionus plicatilis", *Aquaculture*, **293(1-2)**, pp.57-61.

[6] Carter C.G, Bransden M.P, Lewis T.E, Nichols P.D (2003), "Potential of thraustochytrids to partially replace fish oil in Atlantic salmon feeds", *Mar Biotechnol (NY)*, **5(5)**, pp.480-92.

[7] Miller M.R, Nichols P.D, Carter C.G

(2007), "Replacement of fish oil with thraustochytrid *Schizochytrium* sp. L oil in Atlantic salmon parr (Salmo salar L) diets", *Comp Biochem Physiol A Mol Integr Physiol*, **148(2)**, pp.382-92.

[8] Rainuzzo J.R, Reitan K.I, Olsen Y (1997), "The significance of lipids at early stages of marine fish: a review", *Aquaculture*, **155(1-4)**, pp.103-115.

[9] Battaglene S.C, Seymour J.E, Ramofafia C, Lane I (2002), "Spawning induction of three tropical sea cucumbers, Holothuria scabra, H. fuscogilva and Actinopyga mauritiana", *Aquaculture*, **207(1-2)**, pp.29-47.

[10] Calder P.C (2004), "Long-chain n-3 fatty acids and cardiovascular disease: further evidence and insights", *Nutrition Research*, **24(10)**, pp.761-772.

[11] Din J.N, Newby D.E, Flapan A.D (2004), "Omega 3 fatty acids and cardiovascular disease-fishing for a natural treatment", *BMJ*, **328(7430)**, pp.30-5.

[12] Hoang M.H, Ha N.C, Thom Le T, Tam L.T, Anh H.T, Thu N.T, Hong D.D (2014), "Extraction of squalene as value-added product from the residual biomass of *Schizochytrium mangrovei* PQ6 during biodiesel producing process", *J Biosci Bioeng*, **118(6)**, pp.632-9.

[13] Hoàng Thị Minh Hiền, Lê Thị Thơm, Nguyễn Cẩm Hà, Lương Hồng Hạnh, Hoàng Thị Lan Anh, Ngô Thị Hoài Thu, Đặng Diễm Hồng (2013), "Nghiên cứu quá trình tách chiết lipit tổng số và axít béo tự do cho sản xuất dầu omega-3 và omega-6 từ sinh khối vi tảo biển dị dưỡng *Schizochytrium mangrovei* PQ6", *Tạp chí Sinh học*, pp.9.

[14] Hong D.D, Mai D.T, Thom Le T, Ha N.C, Lam B.D, Tam L.T, Anh H.T, Thu N.T (2013), "Biodiesel production from Vietnam heterotrophic marine microalga *Schizochytrium mangrovei* PQ6", *J Biosci Bioeng*, **116(2)**, pp.180-5.

[15] Hong D.D, Anh H.T.L, Thu N.T.H (2011), "Study on biological characteristics of heterotrophic marine microalga-Schizochytrium mangrovei PQ6 isolated from Phu Quoc Island, Kien Giang province, Viet Nam", *Journal of Phycology*, **47(4)**, pp.944-954.

[16] Bolger A.M, Lohse M, Usadel B (2014), "Trimmomatic: a flexible trimmer for Illumina sequence data", *Bioinformatics*, **30(15)**, pp.2114-2120.

[17] Zerbino D.R (2010), "Using the Velvet de novo assembler for short-read sequencing technologies", *Current protocols in bioinformatics/editoral board, Andreas D. Baxevanis ... [et al.]*, pp.Unit-11.5.

[18] Hernandez D, François P, Farinelli L, Østerås M, Schrenzel J (2008), "De novo bacterial genome sequencing: Millions of very short reads assembled on a desktop computer", *Genome Research*, **18(5)**, pp.802-809.

[19] Gurevich A, Saveliev V, Vyahhi N, Tesler G (2013), "QUAST: quality assessment tool for genome assemblies", *Bioinformatics*, **29(8)**, pp.1072-1075.

[20] Langmead B, Salzberg S.L (2012), "Fast gapped-read alignment with Bowtie 2", *Nature methods*, **9(4)**, pp.357-359.

[21] Soueidan H, Maurier F, Groppi A, Sirand-Pugnet P, Tardy F, Citti C, Dupuy V, Nikolski M (2013), "Finishing bacterial genome assemblies with Mix", *BMC Bioinformatics*, **14(Suppl 15)**, pp.S16-S16.

[22] Slater G.S, Birney E (2005), "Automated generation of heuristics for biological sequence comparison", *BMC bioinformatics*, **6(1)**, pp.31.

[23] Altschul S.F, Gish W, Miller W, Myers E.W, Lipman D.J (1990), "Basic local alignment search tool", *Journal of Molecular Biology*, **215(3)**, pp.403-410.

[24] Stanke M, Steinkamp R, Waack S, Morgenstern B (2004), "AUGUSTUS: a web server for gene finding in eukaryotes", *Nucleic acids research*, **32(suppl 2)**, pp.W309-W312.

[25] Korf I (2004), "Gene finding in novel genomes", *BMC bioinformatics*, **5(1)**, pp.59.

[26] Cantarel B.L, Korf I, Robb S.M.C, Parra G, Ross E, Moore B, Holt C, Sánchez Alvarado A, Yandell M (2008), "MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes", *Genome Research*, **18(1)**, pp.188-196.

[27] Ashburner M, Ball C.A, Blake J.A, Botstein D, Butler H, Cherry J.M, Davis A.P, Dolinski K, Dwight S.S, Eppig J.T, Harris, Hill D.P, Issel-Tarver L, Kasarskis A, Lewis S, Matese J.C, Richardson J.E, Ringwald M, Rubin G.M, Sherlock G (2000), "Gene Ontology: tool for the unification of biology", *Nature genetics*, **25(1)**, pp.25-29.

[28] Kanehisa M, Goto S (2000), "KEGG: kyoto encyclopedia of genes and genomes", *Nucleic acids research*, **28(1)**, pp.27-30.

[29] Conesa A, Götz S (2008), "Blast2GO: A comprehensive suite for functional analysis in plant genomics", *International journal of plant genomics*.