Tạp chí Công nghệ Sinh học **17**(3): 449-454, 2019

XÂY DỰNG CƠ SỞ DỮ LIỆU HỆ GEN CÁ TRA VIỆT NAM

Nguyễn Hoàng Vũ, Nguyễn Thành Phương, Lê Thị Nguyên Bình, Kim Thị Phương Oanh✉

Viện nghiên cứu hệ Gen, Viện Hàn lâm Khoa học và Công nghệ Việt Nam

✉ Người chịu trách nhiệm liên lạc. E-mail: ktpoanh@gmail.com

Ngày nhận bài: 12.02.2019
Ngày nhận đăng: 17.9.2019

TÓM TẮT

Các nghiên cứu sinh học phân tử có vai trò quan trọng trong ngành thủy sản, góp phần nâng cao chất lượng giống một cách hiệu quả. Gần đây, cùng với sự phát triển của công nghệ giải trình tự thế hệ mới, nghiên cứu hệ gen được phát triển mạnh mẽ, trong đó việc tổ chức và quản lý dữ liệu giữ một vị trí thiết yếu. Sau khi giải trình tự toàn bộ hệ gen loài cá tra Việt Nam (*Pangasianodon hypophthalmus*), chúng tôi đã tiến hành phân tích và chú giải bộ gen cá tra. Để có thể khai thác dữ liệu này một cách hiệu quả, chúng tôi đã xây dựng một cơ sở dữ liệu cho toàn bộ các dữ liệu thu được. Cơ sở dữ liệu được xây dựng trên nền tảng các phần mềm mã nguồn mở theo mô hình kiến trúc ba lớp (giao diện, dịch vụ và cơ sở dữ liệu) với giao diện sử dụng thuận tiện qua trình duyệt Web. Người sử dụng có thể tra cứu các dữ liệu trình tự và dữ liệu chú giải cũng như hiển thị trực quan các trình tự thông qua trình duyệt hệ gen JBrowse. Cơ sở dữ liệu này là nguồn thông tin quan trọng, tiền đề cho những nghiên cứu sâu hơn về chức năng và nâng cao chất lượng di truyền của cá tra.

Từ khóa: cơ sở dữ liệu, hệ gen cá tra, JBrowse, Pangasianodon hypophthalmus, tin sinh học

ĐẶT VẤN ĐỀ

Cá tra (*Pangasianodon hypophthalmus*) thuộc họ cá tra (Pangasiidae), bộ cá da trơn hay cá nheo (Siluriformes). Cá tra nuôi là một trong những loài cá đặc hữu của vùng lưu vực sông Mê Kông (Việt Nam, Thái Lan, Lào, Campuchia), có giá trị kinh tế lớn và được nuôi phổ biến ở vùng này và một số nước khác thuộc khu vực miền nam châu Á. Việt Nam là nước có sản lượng cá tra nuôi *P. hypophthalmus* lớn nhất thế giới và xuất khẩu sang 140 nước trên thế giới, trong đó có Mỹ, EU, Trung Quốc, các nước ASEAN, Mexico và Brazil. Theo thống kê từ Tổng cục Thủy sản, năm 2017 diện tích thả nuôi cá tra hơn 5.230 ha; sản lượng đạt hơn 1,2 triệu tấn. Kim ngạch xuất khẩu cá tra năm 2017 đạt 1,78 tỷ USD, đóng góp hơn 21% tổng giá trị xuất khẩu của ngành thủy sản.

Để sản xuất cá tra mang lại hiệu quả cao và xuất khẩu theo hướng bền vững, ngoài việc tổ chức lại sản xuất, ngành thủy sản cần phải kiểm soát dịch bệnh, nâng cao chất lượng sản phẩm cá tra để đáp ứng được yêu cầu của thị trường và bảo vệ thương hiệu cá tra Việt Nam trên thị trường quốc tế. Nền tảng cho chiến lược phát triển này là công tác giống nhằm nâng cao chất lượng di truyền của loài cá có giá trị kinh tế cao này. Một trong những vấn đề quan trọng đối với công tác giống là thông tin về đặc điểm cấu trúc phân tử của bộ gen (genome) của cá tra. Nghiên cứu genome sẽ cung cấp những thông tin chính xác nhất cho việc xác định các tính trạng quan trọng, như: tính kháng bệnh, tính chống chịu đối với điều kiện môi trường, các tính trạng liên quan đến năng suất, chất lượng sản phẩm của cá tra. Hơn nữa, nghiên cứu genome cá tra sẽ cung cấp thông tin nhằm nghiên cứu di truyền quần thể, quản lý quần đàn, phát triển DNA barcoding truy xuất nguồn gốc.

Để có chiến lược phát triển lâu dài nghề nuôi một số loài cá kinh tế, nhiều nước trên thế giới đã đầu tư mạnh cho nghiên cứu cơ bản, giải mã và phân tích hệ genome và transcriptome. Ví dụ như: phân tích transcriptome ở cá hồi (Tymchuk *et al*., 2009), cá bơn (Vera *et al*., 2013), cá song (Huang *et al*., 2011), cá nheo Mỹ (Liu *et al*., 2016; Wang *et al*., 2010), cá rô phi (Huang *et al*., 2012)... Từ những nghiên cứu cơ bản này mở ra khả năng cho hàng loạt nghiên cứu ứng dụng, trong đó quan trọng nhất là tìm kiếm các chỉ thị phân tử liên quan đến tính trạng quan tâm như tính trạng tăng trưởng, sức sinh sản và kháng bệnh. Trước nghiên cứu này, dữ liệu về gen cá tra (*P. hypophthalmus*) được lưu giữ trong

Genbank/NCBI vẫn còn rất ít ỏi, thống kê mới nhất trên NCBI website truy cập ngày 21/5/2018 với Taxonomy ID: 310915 chỉ bao gồm: 267 trình tự nucleotide, 239 trình tự protein suy diễn (trong đó rất nhiều trình tự trùng lặp, ví dụ như cytochrome oxidase subunit I) và bộ gen ty thể (NCBI Reference Sequence: NC_021752.1). Do vậy, để tạo tiền đề cho các nghiên cứu về hệ gen cá tra, góp phần cho công tác nghiên cứu và ứng dụng công nghệ sinh học trong thủy sản, chúng tôi thực hiện đề tài nghiên cứu giải mã genome và transcriptome của cá tra trong khuôn khổ đề tài cấp nhà nước.

Sử dụng công nghệ giải trình tự thế hệ mới (Next-Generation Sequencing) với hệ thống của Illumina, chúng tôi đã tiến hành giải mã genome từ mẫu tinh trùng của cá tra, giải mã transcriptome từ mô cơ của cá tra. Khối lượng dữ liệu thu từ máy giải trình tự thế hệ mới lên tới hàng trăm Gbp. Từ dữ liệu này chúng tôi đã tiến hành lắp ráp và chú giải bộ gen cá tra. Để giúp các nhà khoa học có thể khai thác dữ liệu genome cá tra một các dễ dàng và hiệu quả, chúng tôi tiến hành xây dựng cơ sở dữ liệu genome cá tra bao gồm toàn bộ các dữ liệu mà chúng tôi đã giải mã. Cơ sở dữ liệu mà chúng tôi xây dựng sẽ cho phép những nhà nghiên cứu quan tâm khai thác thông tin di truyền hữu ích để nghiên cứu các gen chức năng và các nghiên cứu khác. Song song với việc xây dựng cơ sở dữ liệu này, chúng tôi cũng đồng thời chia sẻ dữ liệu genome/ transcriptome của cá tra trên hệ thống ngân hàng gen chung của thế giới NCBI với mã số BioProject ID, PRJNA448819. Cơ sở dữ liệu riêng của cá tra giúp cho những nhà nghiên cứu chuyên môn sâu dễ dàng tìm kiếm thông tin riêng biệt của riêng loài cá này.

NGUYÊN LIỆU VÀ PHƯƠNG PHÁP

Chuẩn bị dữ liệu genome

Trong quá trình tiến hành giải trình tự và phân tích tin sinh học cho cá tra, chúng tôi đã cho ra các dạng dữ liệu khác nhau gồm dữ liệu trình tự và dữ liệu chú giải.

Dữ liệu trình tự genome được lắp ráp từ dữ liệu giải trình tự thô (dạng fastq) bằng phần mềm Platanus (Kajitani *et al*., 2014). Các dữ liệu này đều được lưu dưới định dạng file FASTA. Trong định dạng file FASTA, mỗi một đoạn trình tự được đánh dấu bằng một dòng bắt đầu bằng ký tự '>' và tên đoạn trình tự; các dòng sau đó là nội dung trình tự này (các ký tự đại diện cho các nucleotide).

Dữ liệu chú giải genome có định dạng GFF (Reese *et al*., 2010). GFF là một định dạng file chuẩn để chứa các đặc trưng genome dưới dạng file văn bản. GFF là viết tắt của Generic Feature Format. File GFF chỉ gồm các kí tự; có 9 cột cách nhau bằng dấu tab. GFF có nhiều bản; bản gần đây nhất là GFF3. GFF3 không tương thích với bản GFF2 trước nó. Định dạng chính thông của GFF3 được mô tả trên trang web Sequence Ontology (http://www.sequenceontology.org/). Cho trường hợp dữ liệu cá tra, file GFF của chúng tôi chứa các thông tin chú giải của các đoạn trình tự: tên trình tự, vị trí bắt đầu và kết thúc, cùng các thông tin chú giải cho đoạn trình tự.

Các nền tảng và công cụ được sử dụng

Trong quá trình xây dựng cơ sở dữ liệu, chúng tôi đều sử dụng các phần mềm mã nguồn mở phổ biến để thuận tiện cho việc cài đặt cũng như thay đổi sau này. Toàn bộ cơ sở dữ liệu được xây dựng trên môi trường hệ điều hành Linux. Hệ thống có mô hình kiến trúc ba lớp. Tầng thứ 1 gồm giao diện tương tác giữa người dùng và cơ sở dữ liệu. Trong trường hợp cơ sở dữ liệu cá tra, giao diện Web được thiết kế trên nền tảng Drupal. Tầng thứ 2 gồm các phần mềm dịch vụ web và quản lý cơ sở dữ liệu. Các phần mềm Apache cùng PHP được sử dụng làm nền tảng kết nối cho tầng này. Tầng thứ 3 gồm hệ thống cơ sở dữ liệu bên dưới. Cho dữ liệu cá tra, hệ thống cơ sở dữ liệu MySQL được sử dụng để lưu trữ dữ liệu.

Khi người dùng tương tác với giao diện Web, các thông tin truy vấn sẽ được đưa về cơ sở dữ liệu MySQL và kết quả truy vấn sẽ được hiển thị lại cho người sử dụng trên giao diện Web. Cấu trúc ba lớp của hệ thống được minh họa trên hình 1.

Với dữ liệu định dạng file fasta, phần mềm JBrowse (Buels *et al*., 2016) được sử dụng để cung cấp giao diện tương tác dạng Web cho những dữ liệu này. JBrowse là một trình duyệt dữ liệu trình tự được sử dụng phổ biến cho nhiều cơ sở dữ liệu sinh học phân tử. JBrowse được phát triển trên nền tảng HTML5 và Javascript. JBrowse có tốc độ hiển thị nhanh, cho phép nhúng vào trong trang web dễ dàng, hỗ trợ nhiều trình duyệt Web khác nhau; đồng thời cung cấp nhiều tính năng hỗ trợ hiển thị dữ liệu trình tự.

Quá trình xây dựng cơ sở dữ liệu

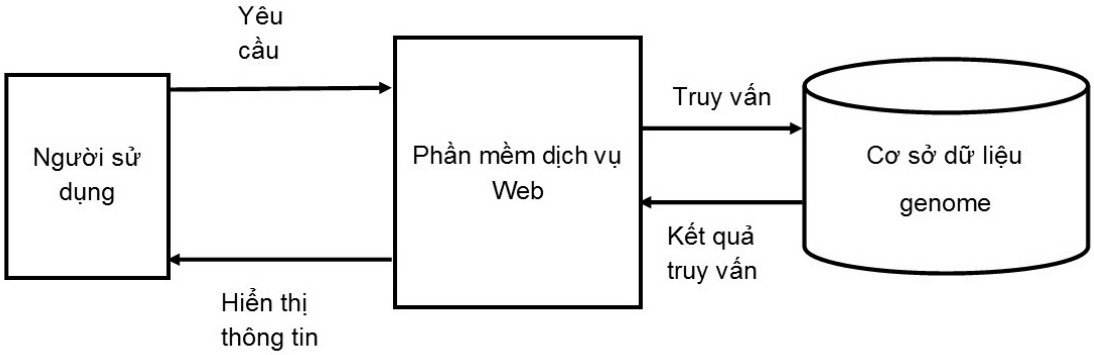
Với các dữ liệu dạng gff, xls và vcf; các dữ liệu này đều đã có sẵn dạng bảng. Do đó, với mỗi loại dữ liệu, chúng tôi xây dựng một bảng trong SQL với số

cột và định dạng dữ liệu cột tương ứng. Quá trình tạo bảng được tiến hành bằng các lệnh SQL trong giao diện dòng lệnh của MySQL trên Linux.

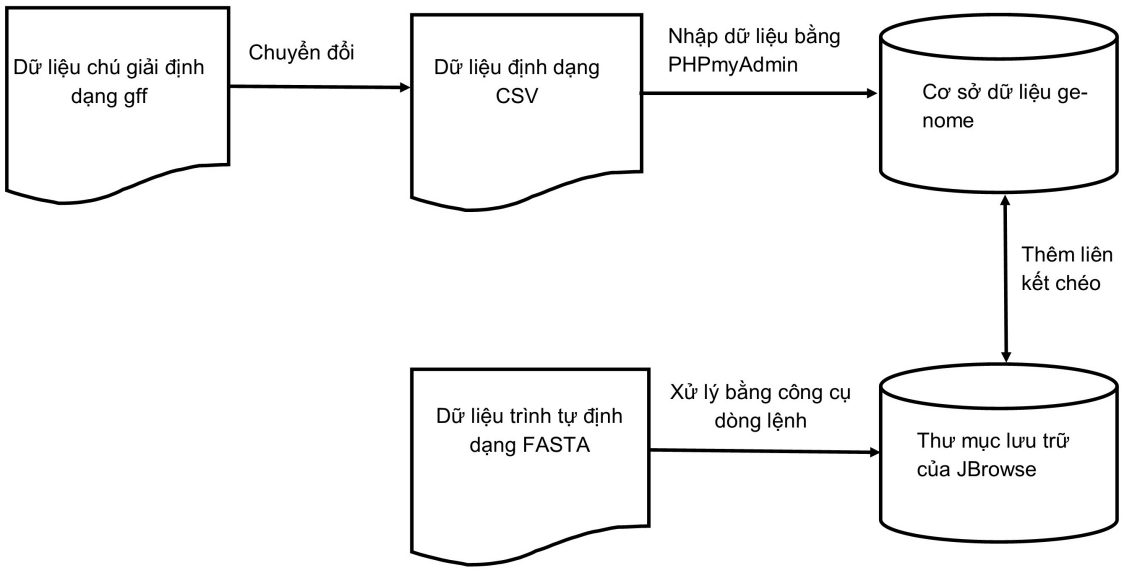
Để nhập dữ liệu vào cơ sở dữ liệu, trước hết các dữ liệu bảng (gff, xls và vcf) đều được chuyển hết về định dạng CSV (dữ liệu được mở bằng Microsoft Excel và xuất ra CSV). Sau đó, các file CSV này được nhập vào bảng tương ứng trong cơ sở dữ liệu bằng giao diện phpMyAdmin.

Với những dữ liệu trình tự có định dạng file fasta, dữ liệu được đưa trực tiếp vào thư mục lưu trữ của JBrowse trên máy chủ và được xử lý bằng các công cụ dòng lệnh của JBrowse. Trình duyệt JBrowse được cài lên cùng hệ thống máy chủ với cơ sở dữ liệu vào giao diện web nhưng cũng có thể chạy trên một máy chủ độc lập nếu cần.

Quy trình xử lý dữ liệu để đưa vào cơ sở dữ liệu được biểu diễn trong hình 2.



Hình 1. Sơ đồ cơ sở dữ liệu.



Hình 2. Quy trình xử lý dữ liệu.

KẾT QUẢ

Cơ sở dữ liệu được truy cập thông qua giao diện web ở trên chính máy chủ chứa cơ sở dữ liệu. Cơ sở dữ liệu của hệ gen cá tra đã được lắp ráp thành 563 scaffold (ký hiệu từ sc0000001 đến sc0000563), trong đó scaffold dài nhất là 37,5Mbp. Cơ sở dữ liệu hiển thị kết quả dự đoán và chú giải hệ gen, bao gồm 28.580 gen.

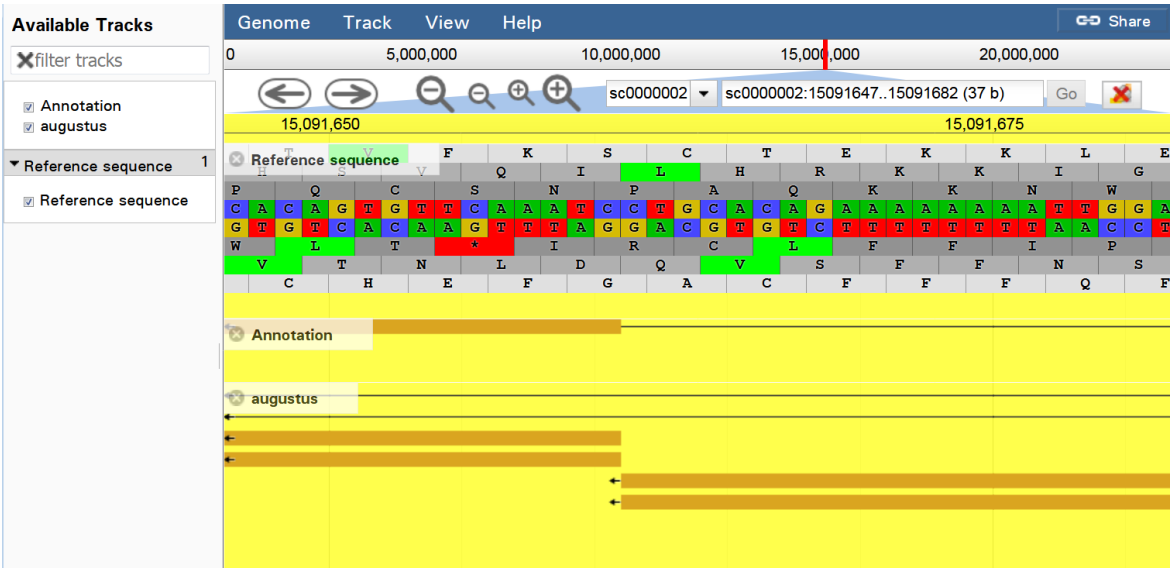
Sau khi truy cập vào trang web, người sử dụng cần phải nhập vào tên truy cập và mật khẩu đã được cấp. Sau khi đăng nhập, sẽ có ô để nhập từ khóa tìm kiếm. Khi người dùng nhập một từ khóa tìm kiếm vào ô tương ứng, một truy vấn sẽ được gửi đến cơ sở dữ liệu và trang web sẽ hiện ra kết quả truy vấn dưới dạng bảng. Mỗi trình tự có liên quan đến từ khóa tìm kiếm được hiển thị trên một dòng (Hình 3).

Đồng thời, cột cuối cùng của kết quả sau truy vấn là đường dẫn đến đoạn trình tự tương ứng trong trình duyệt JBrowse. Đoạn trình tự sẽ được đánh dấu khi hiển thị bằng JBrowse. Người sử dụng có thể dùng các công cụ có sẵn trong JBrowse để tiến hành xem chi tiết trình tự cùng các chú giải kèm theo (Hình 4).

Page **1** 2 3 >>>Last

Scaffold	Start	End	Type	Strand	Attribute	Jbrowse
sc0000002	15077226	15100210	CDS	-	Note=Insulin-degrading enzyme	View
sc0000003	9143133	9151274	CDS	+	Note=Insulin-like growth factor-binding protein 5	View
sc0000003	9159827	9172763	CDS	-	Note=Insulin-like growth factor-binding protein 2	View
sc0000003	11861995	11864346	CDS	+	Note=Insulin-induced gene 2 protein	View
sc0000003	9159827	9172763	CDS	-	Note=Insulin-like growth factor-binding protein 2	View
sc0000003	18991478	19002941	CDS	+	Note=Insulin receptor substrate 2-A	View
sc0000005	85032	88211	CDS	-	Note=Insulinoma-associated protein 1	View
sc0000006	12108751	12117085	CDS	-	Note=Insulin-like growth factor 2 mRNA-binding protein 1	View
sc0000007	13393127	13397113	CDS	+	Note=Insulin-like growth factor-binding protein-like 1	View
sc0000008	11606409	11610932	CDS	-	Note=Insulin gene enhancer protein isl-2b	View

Hình 3. Ví dụ kết quả truy vấn. Các kết quả chú giải liên quan sẽ được hiển thị dưới dạng bảng gồm: cột thứ nhất là tên của scaffold; cột thứ hai và thứ ba là vị trí bắt đầu (Start) và kết thúc (End) của đoạn mã hóa protein; cột thứ tư (Type) chỉ rõ loại trình tự là CDS; cột thứ năm (Strand) là chiều mã hóa của sợi DNA ; cột thứ sáu (Attibute) là kết quả chú giải gen; và cột thứ bảy là đường dẫn đến trình tự hiển thị (View) bằng Jbrowse.



Hình 4. Hiển thị trình tự trên JBrowse.

THẢO LUẬN

Trong quá trình xây dựng cơ sở dữ liệu, toàn bộ các phần mềm và công cụ chúng tôi sử dụng đều có bản quyền mã nguồn mở. Điều này tạo nhiều thuận lợi trong quá trình phát triển cơ sở dữ liệu cũng như cho phép chỉnh sửa, mở rộng cơ sở dữ liệu một cách dễ dàng trong tương lai. Đồng thời, các dự án cơ sở dữ liệu genome sau này cũng có thể áp dụng hệ thống phần mềm tương tự mà không phải lo chi phí cao về bản quyền phần mềm.

Cơ sở dữ liệu của chúng tôi cung cấp những tiện ích cơ bản cho người dùng khi tra cứu cơ sở dữ liệu, đồng thời sử dụng nền tảng JBrowse, một trình duyệt genome được sử dụng cho nhiều cơ sở dữ liệu trên thế giới và mang tính phổ cập cao. Người dùng đã quen với giao diện JBrowse từ trước có thể dễ dàng sử dụng Jbrowse được cài đặt trên máy chủ cơ sở dữ liệu để xem dữ liệu cá tra.

Giao diện sử dụng được thiết kế theo tiêu chí gọn nhẹ, trực quan và dễ sử dụng. Giao diện truy vấn các thành phần cơ sở dữ liệu cũng có thể dễ dàng được mở rộng trong tương lai nếu có nhu cầu tìm kiếm chuyên biệt hóa hơn.

Cơ sở dữ liệu cùng trang web cũng đã được xây dựng theo hướng sẵn sàng mở rộng cho trường hợp có thêm những dữ liệu sinh học phân tử khác của cá tra được đưa vào hoặc cho trường hợp cần bổ sung dữ liệu phân tử của một số loài khác.

KẾT LUẬN

Chúng tôi đã xây dựng thành công cơ sở dữ liệu genome cá tra (*Pangasius hypophthalmus*) nhằm mục đích phục vụ nghiên cứu và ứng dụng trong tương lai. Cơ sở dữ liệu có thể được truy cập và tìm kiếm thông qua giao diện Web đồng thời tích hợp trình duyệt JBrowse để hiển thị dữ liệu trình tự. Cơ sở dữ liệu được đưa lên trang web tại địa chỉ http://catfish.genome.ac.vn.

Lời cảm ơn: *Công trình này là một nhiệm vụ của đề tài cấp nhà nước "Phân tích hệ gen biểu hiện (exome + transcriptome) của cá tra nhằm phát triển chỉ thị phân tử phục vụ chọn giống cá tra theo hướng tăng trưởng" do Bộ Nông nghiệp và Phát triển nông thôn cấp kinh phí thực hiện*.

TÀI LIỆU THAM KHẢO

Buels R, Yao E, Diesh CM, Hayes RD, Munoz-Torres M, Helt G, Goodstein DM, Elsik CG, Lewis SE, Stein L, Holmes IH (2016) JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol.* 17:66

Huang CW, Li YH, Hu SY, Chi JR, Lin GH, Lin CC, Gong HY, Chen JY, Chen RH, Chang SJ, Liu FG, Wu JL (2012) Differential expression patterns of growth-related microRNAs in the skeletal muscle of Nile tilapia (Oreochromis niloticus). *J Anim Sci.* (12):4266-79

Huang Y, Huang X, Yan Y, Cai J, Ouyang Z, Cui H, Wang P, Qin Q (2011) Transcriptome analysis of orange-spotted grouper (Epinephelus coioides) spleen in response to Singapore grouper iridovirus. *BMC Genomics* 12: 556.

Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, Yabana M, Harada M, Nagayasu E, Maruyama H, Kohara Y, Fujiyama A, Hayashi T, Itoh T (2014) Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res* 8:1384-95.

Liu Z, Liu S, Yao J, Bao L, Zhang J, Li Y, Jiang C, Sun L, Wang R, Zhang Y, Zhou T, Zeng Q, Fu Q, Gao S, Li N, Koren S, Jiang Y, Zimin A, Xu P, Phillippy AM, Geng X, Song L, Sun F, Li C, Wang X, Chen A, Jin Y, Yuan Z, Yang Y, Tan S, Peatman E, Lu J, Qin Z, Dunham R, Li Z, Sonstegard T, Feng J, Danzmann RG, Schroeder S, Scheffler B, Duke MV, Ballard L, Kucuktas H, Kaltenboeck L, Liu H, Armbruster J, Xie Y, Kirby ML, Tian Y, Flanagan ME, Mu W, Waldbieser GC (2016) The channel catfish genome sequence provides insights into the evolution of scale formation in teleosts. *Nat Commun.* 7:11757

Reese MG, Moore B, Batchelor C, Salas F, Cunningham F, Marth GT, Stein L, Flicek P, Yandell M, Eilbeck K (2010) A standard variation file format for human genome sequences. *Genome Biol.* 11(8): R88

Tymchuk W, Sakhrani D, Devlin R (2009) Domestication causes large-scale effects on gene expression in rainbow trout: analysis of muscle, liver and brain transcriptomes. *Gen Comp Endocrinol* 164(2-3): 175-183.

Vera M, Alvarez-Dios JA, Fernandez C, Bouza C, Vilas R, Martinez P (2013) Development and Validation of Single Nucleotide Polymorphisms (SNPs) Markers from Two Transcriptome 454-Runs of Turbot (Scophthalmus maximus) Using High-Throughput Genotyping. *Int J Mol Sci* 14(3): 5694-5711.

Wang S, Abernathy J, Waldbieser G, Lindquist E, Richardson P, Lucas S, Wang M, Li P, Thimmapuram J, Liu L, Vullaganti D, Kucuktas H, Murdock C, Small B, Wilson M, Liu H, Jiang Y, Lee Y, Chen F, Lu J, Wang W, Peatman E, Xu P, Somridhivej B, Baoprasertkul P, Quilang J, Sha Z, Bao B, Wang Y, Wang Q, Takano T, Nandi S, Liu S, Wong L, Kaltenboeck L, Quiniou S, Bengten E, Miller N, Trant J, Rokhsar D, Liu ZJ, Catfish Genome Consortium. (2010). Assembly of 500,000 inter-specific catfish expressed sequence tags and large scale

gene-associated marker development for whole genome association studies. *Genome Biol* 11 (1): R8.

DATABASE CONSTRUCTION FOR VIETNAMESE CATFISH GENOME

Nguyen Hoang Vu, Nguyen Thanh Phuong, Le Thi Nguyen Binh, Kim Thi Phuong Oanh

Institute of Genome Research, Vietnam Academy of Science and Technology

SUMMARY

Molecular biological research plays an important role in aquaculture, contributes to the improvement of broodstocks efficiently. Recently, with the development of next-generation sequencing (NGS) technology, genomic studies have been rapidly increased, in which data organisation and management hold a crucial position. After obtaining NGS sequencing data of Vietnamese catfish (*Pangasianodon hypophthalmus*), we have analysed and annotated the catfish genome, from which we have constructed a database for efficient usage. The database is built upon open source software following a three-layer model (interface, Web service and database) with a convenient interface through Web browsers. Users can look up sequence and annotation data as well as visualize sequences through the Jbrowse genome browser. This database is important resource for functional genome and genetic improvement of the catfish.

Keywords: bioinformatics, database, genome, JBrowse, Pangasianodon hypophthalmus