Tạp chí Công nghệ Sinh học **15**(3): 433-439, 2017

ỨNG DỤNG CÔNG NGHỆ GIẢI TRÌNH TỰ GEN THẾ HỆ MỚI VÀ CÁC PHẦN MỀM TIN SINH HỌC TRONG VIỆC ĐÁNH GIÁ SƠ BỘ BIẾN THỂ DI TRUYỀN Ở NGƯỜI BỆNH TỰ KỶ VIỆT NAM

Nguyễn Thu Hiền[1,2], Nguyễn Thị Thanh Ngân[1], Nguyễn Thị Kim Liên[1], Nguyễn Ngọc Lan[1], Nguyễn Văn Tụng[1], Thành Ngọc Minh[3], Phan Văn Chi[4], Nguyễn Huy Hoàng[1,✉]

[1]*Viện Nghiên cứu hệ gen, Viện Hàn lâm Khoa học và Công nghệ Việt Nam*
[2]*Học viện Khoa học và Công nghệ, Viện Hàn lâm Khoa học và Công nghệ Việt Nam*
[3]*Bệnh viện Nhi trung ương, Bộ Y tế*
[4]*Viện Công nghệ sinh học, Viện Hàn lâm Khoa học và Công nghệ Việt Nam*

✉ Người chịu trách nhiệm liên lạc. E-mail: nhhoang@igr.ac.vn

Ngày nhận bài: 26.10.2016
Ngày nhận đăng: 07.01.2017

TÓM TẮT

Tự kỷ là một hội chứng rối loạn phát triển của hệ thần kinh. Bệnh được biểu hiện bằng những khiếm khuyết về tương tác xã hội, khó khăn về giao tiếp và các hành vi sở thích hạn chế, lặp đi lặp lại. Tỷ lệ mắc bệnh ở trẻ nam nhiều hơn trẻ nữ và có xu hướng ngày càng tăng nhanh trên thế giới. Hiện nay chưa có phương pháp chữa trị dứt điểm cho các triệu chứng của bệnh tự kỷ. Các nghiên cứu trên thế giới cho thấy rằng tự kỷ là một trong bệnh có yếu tố di truyền chiếm từ 40-80%, và do nhiều gen liên quan. Nguy cơ di truyền của bệnh có liên quan đến ảnh hưởng kết hợp của các biến thể khác nhau. Giải trình tự vùng mã hóa - Whole exome sequencing (WES) đã xác định hàng chục nghìn biến thể gen trong mỗi exome ở nhiều bệnh đa gen như: tim mạch, thần kinh Vì thế, WES đang được coi là hướng đi đúng đắn để nghiên cứu di truyền bệnh tự kỷ. Bằng cách ứng dụng các phần mềm tin sinh học chuyên sâu như BWA (Burrows-Wheeler Alignment Tool); Picard; GATK (Genome Analysis Tool Kit), SnpEff, SnpSift, PolyPhen-2, nghiên cứu này đưa ra một quy trình cơ bản nhất để xác định các biến thể di truyền ở người bệnh tự kỷ. Đây là nghiên cứu đầu tiên sử dụng phương pháp WES để phân tích mối liên quan di truyền với bệnh nhân tử kỷ ở Việt Nam. Kết quả của nghiên cứu này làm cơ sở để định hướng cách thức phân tích số liệu WES.

Từ khóa: *Bệnh di truyền; giải trình tự gen thế hệ mới; giải trình tự vùng mã hóa; tin sinh học; tự kỷ*

MỞ ĐẦU

Tự kỷ (Autism Spectrum Disorders -(ASD)) thuộc một nhóm các rối loạn thần kinh, không đồng nhất về mặt di truyền.Tự kỷ được biểu hiện ra ngoài bằng những khiếm khuyết về tương tác xã hội, khó khăn về giao tiếp ngôn ngữ và phi ngôn ngữ, hành vi, sở thích và hoạt động mang tính hạn hẹp, lặp đi lặp lại (Butler *et al*., 2015). Ngoài những triệu chứng lâm sàng cổ điển cụ thể, có khoảng 31% bệnh nhân bị khuyết tật trí tuệ, 20-25% có triệu chứng co giật (Canitano, 2007; Liu, Takumi, 2014; Srivastava, Schwartz, 2014). Một số bệnh thường thấy đi kèm với ASD bao gồm rối loạn lo âu (White *et al.,* 2009), rối loạn giấc ngủ, rối loạn tiêu hóa (Valicenti-McDermott *et al.,* 2006) và các phản ứng bất thường gây kích thích cảm giác (Rogers *et al*., 2003). Điều đáng nói là hiện nay chưa có phương pháp chữa trị dứt điểm cho các triệu chứng của bệnh tự kỷ. Các biện pháp được áp dụng hiện nay chỉ để giảm các triệu chứng về hành vi, các loại thuốc nhằm giảm sự hung hăng, lo âu, trầm cảm…(Smith *et al*., 2010). Ước tính mới nhất cho thấy rằng ASD ảnh hưởng đến khoảng 1 trong 68 trẻ em và tỷ lệ mắc bệnh ở nam giới chiếm ưu thế so với nữ (4:1) (Butler *et al*., 2015).

Nguy cơ di truyền của bệnh được đề xuất có liên quan đến ảnh hưởng kết hợp của các biến thể khác nhau (Inoue *et al*., 2015). Trong những nghiên cứu ở những cặp song sinh, sự đồng nhất kiểu hình của ASD ở những cặp song sinh cùng trứng chiếm 70-90%, trong khi tỉ lệ này ở những cặp song sinh khác trứng chỉ 0-30% (Rosenberg *et al*., 2009; Ronald,

Hoekstra, 2014). Các nghiên cứu cho thấy rằng, anh chị em trong cùng một gia đình có một bệnh nhân mắc bệnh có nguy cơ cao lên tới 25% so với dân số nói chung (Chahrour *et al*., 2012). Tự kỷ được coi là một trong những rối loạn thần kinh có tính di truyền cao (Chahrour *et al*., 2012). Yếu tố môi trường cũng có những tương tác với yếu tố sơ di truyền và gây ra những thay đổi bất thường trong sự phát triển tế bào thần kinh, phát triển trí não, và liên kết chức năng (Sener *et al*., 2016).

Giải trình tự vùng mã hóa - Whole exome sequencing (WES) là một ứng dụng của công nghệ giải trình tự thế hệ mới để xác định các biến thể trên tất cả các vùng mã hóa, hoặc exon của gen được biết đến. Vì thế WES đã được sử dụng rộng rãi trong các nghiên cứu lâm sàng vài năm gần đây, đặc biệt trong việc xác định các gen bệnh di truyền theo Mendel (Sener *et al*., 2016). Hàng chục nghìn biến thể gen có thể được xác định trong mỗi exome trong nhiều bệnh phức tạp như: tim mạch, thần kinh,... Trí tuệ là một tính trạng cực kỳ phức tạp do nhiều gen quy định, những nghiên cứu ảnh hưởng của thay đổi các gen liên quan đến trí tuệ dẫn đến thiểu năng trí tuệ cũng như tự kỷ cần được tiến hành ở mức độ hệ gen, nhất là hệ gen biểu hiện (exome). WES đang được coi là hướng đi đúng đắn để nghiên cứu di truyền bệnh tự kỷ. Phương pháp này giúp xác định điều kiện di truyền cụ thể với những trường hợp còn nghi ngờ về mặt lâm sàng, cho thấy tầm quan trọng của sự mất một phần chức năng của gen trong hội chứng tự kỷ (Yu *et al*., 2013). Thành công của phương pháp giải trình tự vùng mã hóa (WES) trong việc phát hiện những đột biến và xác định các gen gây bệnh tự kỷ đã được chứng minh bởi nhiều nghiên cứu (Sener *et al.,* 2016).

Tuy nhiên, việc áp dụng công nghệ giải trình tự gen thế hệ mới đi cùng với một vấn đề cần giải quyết đó chính là việc phân tích khối lượng dữ liệu khổng lồ. Một dữ liệu hệ gen cần được phân tích, so sánh, khai thác với các trình tự tham chiếu. Để giải quyết vấn đề này, các công cụ tin sinh đã được phát triển và ứng dụng rộng rãi. Một số công cụ tin sinh phổ biến hiện nay trong lĩnh vực này như BWA (Burrows-Wheeler Alignment Tool) (Li, Durbin, 2009), Picard,GATK (Genome Analysis Toolkit),… Nghiên cứu này báo cáo phương pháp phân tích các biến dị di truyền ở người bệnh tự kỷ Việt Nam bằng phương pháp WES và các công cụ tin sinh hiện đại. Đây có thể coi là nghiên cứu đầu tiên tại Việt Nam trong lĩnh vực nghiên cứu di truyền bệnh tự kỷ bằng phương pháp giải trình tự gen thế hệ mới.

NGUYÊN LIỆU VÀ PHƯƠNG PHÁP

Đối tượng tham gia

Các bệnh nhân được khám, xét nghiệm và chẩn đoán bởi các bác sĩ Khoa thần kinh của Bệnh viện Nhi Trung ương. Thủ tục lấy mẫu tuân thủ đúng theo Hội đồng Y đức của Bệnh viện Nhi Trung ương.

Phương pháp

Tách chiết DNA

DNA tổng số được tách chiết từ máu toàn phần của bệnh nhân ASD và gia đình được tách chiết bằng bộ kit QIAamp DNA Blood Mini Kit – QIAGEN (Đức).

Giải trình tự

Mẫu DNA được giải trình tự trên máy giải trình tự thế hệ mớiIllumina Hiseq/Nextseq của hãng Illumina (USA).

Phân tích dữ liệu

Thư viện DNA được chuẩn bị theo hướng dẫn của bộ kit Agilent SureSelect Target Enrichment của hãng Illumina (Mỹ) dựa trên việc sử dụng các mồi cARN có chiều dài khoảng 120 mer để lựa chọn các khu vực cần quan tâm và làm giàu khu vực đó để chuẩn bị thư viện đoạn gen dùng trong giải trình tự gen thế hệ mới (Next Generation Sequencing – NGS).

Thư viện DAN được chuẩn bị theo 4 bước chính

1- Từ gDNA được phân cắt thành những phân đoạn nhỏ.

2- Chuẩn bị thư viện cùng với adaptor và index có trình tự đặc thù. Các phân đoạn DNA được ligase với adaptor và mẫu dò trong buffer HY BUFFER.

3- Hỗn hợp mẫu và đầu dò được gắn vào các hạt bead và được giữ lại trên giá kim loại. Các phân đoạn còn lại sẽ bị loại bỏ.

4- Hỗn hợp DNA+mẫu dò+hạt bead được rửa sạch để loại bỏ mẫu dò và hạt bead. Các đoạn DNA tinh sạch, đạt yêu cầu chất lượng sẽ được đưa vào máy đọc trình tự.

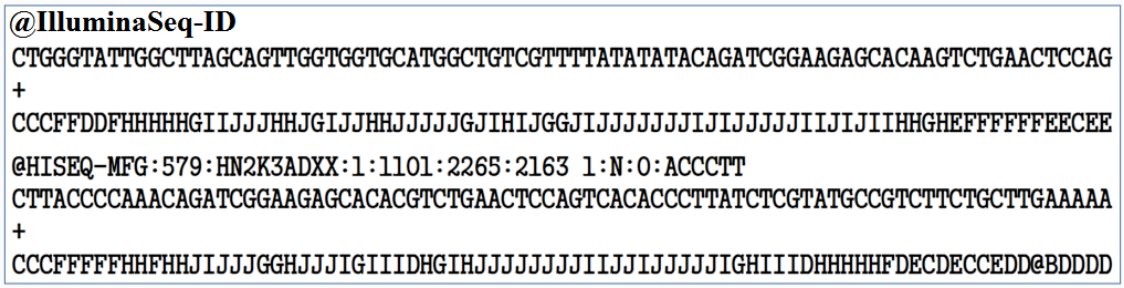
Thư viện DNA sau đó được giải trình tự trên máy giải trình tự mới. Dữ liệu trình tự được sắp xếp và so sánh với ngân hàng gen người (hg19) bằng phần mềm BWA phiên bản 0.7.10. (Li, Durbin, 2009). Bản sao phân tử được loại bỏ bằng cách sử dụng Picard v1.118. Dữ liệu sau đó được phân tích bằng Genome Analysis Toolkit v3.4 để tìm tất cả những vị trí có sự thay đổi alen với tần số thống kê

cao, bao gồm SNPs, đoạn thêm, mất ngắn và CNVs (McKenna, Hanna *et al.,* 2010). Biến thể được chú giải bằng phần mềm SnpEff v4.1 và các cơ sở dữ liệu dbSNP v142, 1000Genome, ClinVar, ESP nhằm xác định ảnh hưởng của biến thể (Cingolani *et al*., 2012) . Để chọn lọc được những biến thể tiềm năng, dữ liệu được lọc qua các bước lọc như sau. Đầu tiên, các biến thể có giá trị MQ < 40 bị loại bỏ. Thứ hai, các biến thể có giá trị Sift_Pred được đánh dấu là "Damaging (D)" hoặc "NA ('.')" được giữ lại. Thứ ba, chọn lọc các biến thể thay thế. Thứ tư, loại bỏ những biến thể đã được được biết đến trong ngân hàng dữ liệu SNPs 142.

KẾT QUẢ

Kiểm định chất lượng

Sau khi đưa mẫu vào máy giải trình tự gen, việc chạy máy kết thúc sẽ cho dữ liệu thô đầu tiên. Để đánh giá, kiểm soát chất lượng và nhận diện các lỗi trong dữ liệu thì việc đầu tiên chính là kiểm định chất lượng, bước này đặc biệt quan trọng vì nó đảm bảo cho các bước phân tích tiếp theo. Đối với máy giải trình tự gen thế hệ mới Illumina thì số liệu thô ban đầu được định dạng dưới file fastq, file này bao gồm 4 dòng ví dụ như hình 1.



Hình 1. Hình ảnh minh họa file kiểm định chất lượng. Dòng 1: ID-tên kí hiệu cho thông tin nhận dạng mẫu; Dòng 2: trình tự nucleotide; Dòng 3: dòng định danh điểm chất lượng - dấu cách (+); Dòng 4: dòng điểm chất lượng.

Điểm chất lượng (Phred quanlity score chart) thể hiện tính chính xác của mỗi nucleotide. Trong giải trình gen thế hệ mới (Next generation sequencing - NGS) mỗi nucleotide có một chất lượng xác suất riêng được tính bằng thuật toán phred và mã hóa bằng ký tự ASCII (ASCII character code = phred quanlity value +33) theo chuẩn phred (quanlity of phred score-Q), số Q càng cao thì độ chính xác cũng càng cao. Ví dụ, nếu có điểm Q chất lượng khoảng 30 thì các lỗi đọc base là 1 trong 1000. Điểm chất lượng được tính theo công thức $Q = -10\log_{10}P$, trong đó P là xác suất của các lần đọc sai sót.

Dữ liệu thu được từ máy giải trình tự gen được định dạng dưới dạng file fastq. Kết quả cho thấy các mẫu đều thu được số trình tự đọc (read) rất lớn, với độ dài tổng số các mẫu cao, lên tới 10,7 Gb của mẫu T09 (Bảng 1), hàm lượng GC từ 47% trở lên. Ở đây tỷ lệ %GC trên toàn bộ trình tự trong mẫu phân bố đạt chuẩn với tỷ lệ trung bình %GC của hệ gen phân tích (tỷ lệ % GC > 15% là đạt chuẩn – theo hãng Illumina). Tỷ lệ Q30 đều trên 95% (tỷ lệ đọc có điểm chất lượng Phred trên 30) và Q20 trên 97% (tỷ lệ đọc có điểm chất lượng Phred trên 20) (Bảng 1).

Bảng 1. Bảng thông tin chất lượng đọc.

Tên mẫu	Tổng base (bp)	Tổng số trình tự đọc	%GC	Q20 (%)	Q30(%)
T01	7,898,618,342	78,204,142	47.8	98.2	97.0
T02	9,005,484,816	89,163,216	47.4	98.3	97.2
T03	8,005,656,526	79,263,926	47.7	98.3	97.1
T06	8,615,935,896	85,306,296	47.6	97.9	96.
T07	9,140,252,146	90,497,546	47.	97.1	95.
T08	9,496,766,794	94,027,394	47.5	97.2	95.6
T09	10,724,544,206	106,183,606	47.4	97.9	96.6

Gióng hàng dữ liệu với hệ gen tham chiếu hg19 và loại bỏ vị trí phân tử trùng lặp

BWA (Burrows-Wheeler Alignment Tool) là một chương trình phần mềm liên kết trình tự các gen nhỏ khác nhau với một bộ gen tham khảo lớn, ví dụ như gen người. Chương trình này bao gồm 3 thuật toán BWA-backtrack, BWA-SW và BWA-MEM. Thuật toán đầu tiên BWA-backtrack được thiết kế cho việc đọc chuỗi trình tự Illumina có kích thước 100 bp trở xuống, trong khi 2 thuật toán kia dùng cho các trình tự có khả năng đọc cao hơn, dao động từ 70 bp đến 1 Mbp. BWA-MEM và BWA-SW chia sẻ các chức năng tương tự nhau, ví dụ như hỗ trợ khả năng đọc cao và sắp xếp các trình tự. Tuy nhiên, BWA-MEM là chương trình mới nhất và được khuyến cáo dùng cho các kết quả có yêu cầu chất lượng, độ chính xác cao, và nhanh hơn. Thêm vào đó, BWA-MEM còn có hiệu suất tốt hơn so với BWA-backtrack trong khoảng đọc 70-100 bp.

Đối với tất cả các thuật toán của BWA, việc cần thiết đầu tiên là phải cấu trúc được FM-index cho các gen tham khảo (sử dụng lệnh index). Các thuật toán sắp xếp được thực hiện theo lệnh "aln/samse/sample", "bwasw" đối với BWA-SW và "mem" đối với BWA-MEM.

Picard là bộ công cụ được xây dựng trên nền tảng Java nhằm thao tác trên tập tin định dạng SAM, BAM. Picard MarkDuplicates sẽ kiểm tra việc sắp xếp dữ liệu trong tập SAM và BAM qua đó cung cấp vị trí các phân tử trùng lặp.

Bảng 2 cho thấy sử dụng công cụ BWA cho khả năng gióng hàng tốt, trên 99,8% dữ liệu được gióng hàng thành công với trình tự tham chiếu hg19. Sau khi sử dụng Picard để loại bỏ phân tử trùng lặp, 97 - 98% số đoạn trình tự được giữ lại, trong đó có 72 – 77% dữ liệu được ánh xạ vào vùng gen quan tâm (Bảng 2).

Bảng 2. Kết quả gióng hàng.

Tên mẫu	Số đoạn trình tự gióng hàng thành công	Số đoạn trình tự gióng hàng thành công sau khi loại bỏ phân tử trùng lặp	Số đoạn trình tự được ánh xạ vào vùng gen quan tâm
T01	78,092,641	76,441,302	57,234,763
T02	89,037,208	86,413,065	66,873,193
T03	79,188,077	76,975,824	58,228,513
T06	85,237,890	83,203,213	61,971,614
T07	90,427,239	88,256,633	66,092,691
T08	93,956,665	91,994,667	68,498,820
T09	106,049,469	103,164,496	74,784,161

Xác định và chú giải biến thể

GATK là bộ công cụ phân tích hệ gen được phát triển tại Viện Broad để phân tích dữ liệu trình tự có thông lượng cao. Gói phần mềm này cung cấp một loạt các công cụ phân tích khác nhau, tập trung chính vào việc phát hiện các biến thể và kiểu gen cũng như nhấn mạnh vào việc cung cấp dữ liệu có độ chính xác cao.

Để tăng độ tin cậy của quá trình phân tích các biển thể được phát hiện, chúng tôi sử dụng phần mềm GATK để loại bỏ những biến thể giả. Chỉ tiêu cần áp dụng lọc các biến thể indel là: QD < 2.0, FS > 200.0, với các biến thể SNP là: |QD < 2.0 || FS > 60.0|.

Trong đó QD (QualByDepth) là độ tin cậy khi gọi tên biến thể, được tính bằng chiều sâu của mỗi trình tự đọc hỗ trợ cho một biến thể. Chỉ số này được tính theo công thức QUAL/AD. Chỉ số Qual là tổng điểm chất lượng của nucleotide tại vị trí xảy ra biến thể và AD là số lượng allen chứa vị trí xảy ra biến thể bao gồm cả allen chưa lọc và allen tham chiếu.

FS (Strand bias estimated using Fisher's Exact Test) là giá trị của phép thử Fisher's Exact nhằm xác định độ lệch chuỗi trong các đoạn trình tự (có những variant chỉ được phát hiện trên sợi xuôi hoặc trên sợi ngược). Giá trị FS càng cao thì đoạn trình tự càng có khả năng bị lệch. Các thông số được lựa chọn dựa theo khuyến cáo của phần mềm GATK.

Phần mềm SnpEff sử dụng để phân chia các biến thể thành các nhóm theo mức độ ảnh hưởng chức năng của biến thể (Bảng 3). Đây là công cụ chú thích và dự báo ảnh hưởng của các biến thể gen (như thay đổi amino acid). Dữ liệu đầu vào của công cụ này là

các biến thể được dự đoán (SNPs, chèn, xóa và MNPs), là kết quả của giải trình tự, và có định dạng VCF (Variant Call Format). Trong dữ liệu đầu ra, SnpEff sẽ phân tích các biến đầu vào để chú giải và tính toán các tác động mà các biến thể có thể tạo ra trên gen. SnpEff đưa ra các kết quả như sau: kiểu gen và các điểm bị ảnh hưởng bởi biến thể; vị trí của các biến thể; làm thế nào mà các biến thể ảnh hưởng đến quá trình tổng hợp protein; so sánh với các dữ liệu khác để tìm các biến thể đã biết (Bảng 3).

Bảng 3. Kết quả xác định và chú giải biến thể.

Tên biến thể	Mẫu T01	Mẫu T02	Mẫu T03	Mẫu T06	Mẫu T07	Mẫu T08	Mẫu T09
Tổng SNP	**103,84**	**105,091**	**103.809**	**104,497**	**104.022**	**103.954**	**107.192**
Biến thể đồng nghĩa	11,488	11,539	11.322	11,417	11.276	11.447	11.664
Biến thể sai nghĩa	10,546	10,734	10.540	10,456	10.423	102	10.644
Thêm bộ mã hóa kết thúc	78	80	95	95	84	34	97
Mất bộ ba mã kết thúc	38	31	36	38	39	37	42
Tổng số biến thể thơm bớt	**14,843**	**15.581**	**14.898**	**15,077**	**14.943**	**14.793**	**16.192**
Đột biến lệch khung đọc	284	279	273	283	276	275	306
Thêm bộ ba mã hóa	163	156	148	148	158	155	154
Mất bộ ba mã hóa	207	207	174	178	185	185	198
% tìm thấy trên dbSNP142	**97.3**	**97.2**	**97.4**	**97.3**	**97.3**	**97.3**	**97.1**

Kết quả, chúng tôi đã thu được 6 nhóm biến thể, trong đó có đến hơn 97% số biến thể đã có sẵn trong ngân hàng dbSNP142.

Sau quá trình lọc, những gen/đột biến được giữ lại thỏa mãn các điều kiện:

- Gen có khả năng gây ra bệnh liên quan thần kinh
- Có chỉ số MQ>40 (mapping quality)
- SIFT_Pred=D, PolyPhen 2 _ Pred =D (Damaging)
- Biến thể thay thế
- Đột biến không có trong cơ sở dữ liệu dbSNP 142

MQ là chỉ số đánh giá chất lượng gióng hàng được tính theo công thức MQ= $-10\log_{10}P$ với P là xác suất đoạn trình tự bị gióng hàng sai vị trí. Với MQ = 40, xác suất gióng hàng sai lệch là 1/10000, có nghĩa là cứ 10.000 đoạn trình tự được gióng hàng thì chỉ có 1 đoạn trình tự bị gióng hàng sai. Độ chính xác tương đương 99,99%.

Với công cụ SIFT, các nhà phân tích có thể dự đoán xem một sự thay thế amino acidcó khả năng ảnh hưởng đến chức năng của protein hay không, dựa trên sự tương đồng về trình tự và tương tự hóa lý (Physico-chemical) giữa các amino acid thay thế. Dữ liệu cung cấp cho mỗiamino acid thay thế là chỉ số và dự đoán định tính (hoặc dung nạp hoặc gây hại). Chỉ số này là tỉ lệ mà amino acid được thay thế có dung nạp hay không, vì vậy chỉ số gần với mức 0 tương tự với việc sẽ gây hại. Dự đoán định tính sẽ được đưa ra từ chỉ số, như vậy sự thay thế với chỉ số <0.05 được gọi là gây hại và ngược lại sẽ là dung hợp.

Công cụ PolyPhen-2 dự đoán sự ảnh hưởng của amino acid thay thế trên cấu trúc và chúc năng của protein sử dụng sự tương đồng về trình tự, chú thích Pfam, cấu trúc 3D, từ PDB, và một số cơ sở dữ liệu và công cụ khác (bao gồm cả DSSP, ncoils…). Chỉ số PolyPhen - 2 đưa ra xác suất mà việc thay thế là có hại, vì vậy giá trị gần với mức 1 sẽ được hiểu như là có hại (chú ý rằng điều này ngược hẳn với SIFT). Dự đoán định tính dựa trên tỉ lệ dương tính giả (False Positive Rate hay còn gọi là tỉ lệ báo động giả) của việc phân loại phương thức được sử dụng để dự đoán. Theo hướng dẫn của phần mềm đánh giá này, các biến thể có điểm đánh giá trong khoảng 0.957 đến 1 được cho là có hại (D - porobably damaging); thang điểm trong khoảng 0.453 - 0.956 là có thể gây hại (P – possibly damaging) và các biến thể có điểm đánh giá trong khoảng 0 - 0.452 là an toàn (B - 0,0.452).

Vì vậy, trong bảng 4, các biến thể bị đánh giá là có ảnh hưởng đến chức năng protein (SIFT_Pred=D và PolyPhen 2 _ Pred =D (Damaging)) được giữ lại.

Vì mục tiêu của nghiên cứu là tìm ra các biến thể mới nằm trong các gen tiềm năng liên quan đến bệnh tự kỷ nên số lượng biến thể được xác định trong cơ sở ngân hàng dữ liệu đa hình đơn

nucleotide (The Single Nucleotide Polymorphism Database - dbSNP) được bỏ qua. dbSNP là một kho lưu trữ mở, bao gồm thông tin các biến thể di truyền trong và giữa các loài khác nhau được phát triển bởi Trung tâm thông tin Công nghệ sinh học (National Center for Biotechnology Information - NCBI) phối hợp với Viện Nghiên cứu quốc gia về gen người (National Human Genome Research Institute - NHGRI). dbSNP được biết đến là các đa hình trung tính, đa hình liên quan đến một kiểu hình cụ thể (Sherry *et al*., 1999). Hiện nay, chưa có một ngân hàng SNP nào cho bệnh tự kỷ. Vì thế, sau các bước lọc, số lượng biến thể đã được loại bỏ đáng kể. Chỉ còn nhiều nhất là 19 biến thể đáng quan tâm ở mẫu T06 , 10, 15, 12, 14, 16, 8 biến thể ở các bệnh nhân T07, T08, T09, T01, T02, T03. Đây chính là những dữ liệu quan trọng cho các nghiên cứu tiếp theo.

Bảng 4. Số lượng đột biến trong các mẫu sau mỗi bước lọc.

Dữ liệu	T06	T07	T08	T09	TO1	TO2	TO3
Dữ liệu gốc	119574	118965	118774	123386	118687	120672	118707
Thuộc gen có tiềm năng gây bệnh và MQ>40	16325	16118	16389	16747	16498	16478	16495
SIFT_Pred=D Và PolyPhen 2 _ Pred =D	319	305	319	304	330	342	309
Effect=missense	319	305	319	304	330	342	309
Không có trong dbSNP 142	19	10	15	12	14	16	8

KẾT LUẬN

Bằng cách áp dụng các công cụ tin sinh chuyên dụng, khối lượng dữ liệu khổng lồ các biến thể được thu gọn đáng kể. Các biến thể di truyền trên các gen tiềm năng từ người bệnh tự kỷ Việt Nam được đưa ra một các chính xác nhất. Nghiên cứu này đưa ra một quy trình đơn cơ bản nhất để xác định các biến thể di truyền ở người bệnh tự kỷ. Kết quả này làm tiền đề cho những nghiên cứu tiếp theo sâu hơn đối với nghiên cứu di truyền bệnh này.

Lời cảm ơn: *Công trình nghiên cứu này được thực hiện bằng sự hỗ trợ kinh phí của đề tài "Giải trình tự toàn bộ vùng mã hóa (exome) ở bệnh nhân tự kỷ Việt Nam", mã số: VAST02, 2015-2016, TS. Nguyễn Huy Hoàng làm chủ nhiệm, thuộc các hướng KHCN ưu tiên cấp Viện Hàn lâm Khoa học và Công nghệ Việt Nam.*

TÀI LIỆU THAM KHẢO

Butler MG, Rafi SK, Hossain W, Stephan DA, Manzardo AM (2015) Whole exome sequencing in females with autism implicates novel and candidate genes. *Int J Mol Sci* 16(1): 1312-1335.

Canitano R (2007) Epilepsy in autism spectrum disorders. *Eur Child Adolesc Psychiatry* 16: 61–66.

Chahrour MH, Yu TW, Lim ET, Ataman B, Coulter ME, Hill RS, Stevens CR, Schubert CR; ARRA Autism Sequencing Collaboration, Greenberg ME, Gabriel SB, Walsh CA (2012) Whole-exome sequencing and homozygosity analysis implicate depolarization-regulated neuronal genes in autism. *PLoS Genet* 8(4): e1002635.

Sener EF, Canatan H, Ozkul Y (2016) Recent Advances in Autism Spectrum Disorders: Applications of Whole Exome Sequencing Technology. *Psychiatry Investig* 13(3): 255–264.

Inoue E, Watanabe Y, Xing J, Kushima I, Egawa J, Okuda S, Hoya S, Okada T, Uno Y, Ishizuka K, Sugimoto A, Igeta H, Nunokawa A, Sugiyama T, Ozaki N, Someya T (2015) Resequencing and Association Analysis of CLN8 with Autism Spectrum Disorder in a Japanese Population. *PLoS One* 10(12): e0144624.

Li H and Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14): 1754-1760.

Liu X and Takumi T (2014) Genomic and genetic aspects of autism spectrum disorder. *Biochem Biophys Res Commun* 452(2): 244-253.

Rogers SJ, Hepburn S, Wehner E (2003) Parent reports of sensory symptoms in toddlers with autism and those with other developmental disorders. *J Autism Dev Disord* 33(6): 631-642.

Ronald A and Hoekstra R (2014) Progress in Understanding the Causes of Autism Spectrum Disorders and Autistic Traits: Twin Studies from 1977 to the Present Day. *Springer, New York*: 33-65.

Rosenberg RE, Law JK, Yenokyan G, McGready J, Kaufmann WE, Law PA (2009) Characteristics and concordance of autism spectrum disorders among 277 twin pairs. *Arch Pediatr Adolesc Med* 163(10): 907-914.

Sherry ST, Ward M, Sirotkin K (1999) dbSNP - database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Research* 9(8): 677–679.

Smith CL, Bolton A, Nguyen G (2010) Genomic and epigenomic instability, fragile sites, schizophrenia and autism. Curr Genomics. *Curr Genomics* 11: 447–469.

Srivastava AK and Schwartz CE (2014) Intellectual disability and autism spectrum disorders: causal genes and molecular mechanisms. *Neurosci Biobehav Rev* 46: 161–174.

Valicenti-McDermott M, McVicar K, Rapin I, Wershil BK, Cohen H, Shinnar S (2006) Frequency of gastrointestinal symptoms in children with autistic spectrum disorders and association with family history of autoimmune disease. *J Dev Behav Pediatr* 27(2 Suppl): S128-136.

White SW, Oswald D, Ollendick T, Scahill L (2009) Anxiety in children and adolescents with autism spectrum disorders. *Clin. Psychol. Rev.* 29: 216-229.

Y Yu TW, Chahrour MH, Coulter ME, Jiralerspong S, Okamura-Ikeda K, Ataman B, Schmitz-Abe K, Harmin DA, Adli M, Malik AN, D'Gama AM, Lim ET, Sanders SJ, Mochida GH, Partlow JN, Sunu CM, Felie JM, Rodriguez J, Nasir RH, Ware J, Joseph RM, Hill RS, Kwan BY, Al-Saffar M, Mukaddes NM, Hashmi A, Balkhy S, Gascon GG, Hisama FM, LeClair E, Poduri A, Oner O, Al-Saad S, Al-Awadi SA, Bastaki L, Ben-Omran T, Teebi AS, Al-Gazali L, Eapen V, Stevens CR, Rappaport L, Gabriel SB, Markianos K, State MW, Greenberg ME, Taniguchi H, Braverman NE, Morrow EM, Walsh CA. (2013) Using whole-exome sequencing to identify inherited causes of autism. *Neuron* 77(2): 259-273.

PRELIMINARY ASSESSMENT OF VARIATIONS IN VIETNAMESE PATIENTS WITH AUTISM SPECTRUM DISORDERS BY WHOLE-EXOME SEQUENCING AND BIOINFORMATICS SOFTWARE

Nguyen Thu Hien[1,2], Nguyen Thi Thanh Ngan[1], Nguyen Thi Kim Lien[1], Nguyen Ngoc Lan[1], Nguyen Van Tung[1], Thanh Ngoc Minh [3], Phan Van Chi[4], Nguyen Huy Hoang[1]

[1]*Institute of Genome Research, Vietnam Academy of Science and Technology*
[2]*Graduate University of Science and Technology, Vietnam Academy of Science and Technology*
[3]*National Hospital of Pediatrics, Ministry of Health*
[4]*Institute of biotechnology, Vietnam Academy of Science and Technology*

SUMMARY

Autism is a developmental disorder of the central nervous system. The disease is manifested by impairments of social interaction, difficulty with communication and restricted and repetitive behaviors. Boys are more likely to be diagnosed with ASD than girls and the incidence rate is trending in the world. However, there is no definite cure for the symptoms of autism so far. Previous studies have showed that autism is a hereditary disease with the causes from genetic factors accounted for 40-80% and related to many genes. Genetic risk of the disease is related to the combined effects of different variants. Sequencing the coding region - Whole exome sequencing (WES) has identified tens of thousands of genes variants in each exome in many multi-gene disease such as cardiovascular, neurological. Therefore, WES is being considered as the right and effective method in the study of genetics of the autism. By applying intensive bioinformatics programs, including BWA (Burrows-Wheeler Alignment Tool); Picard; GATK (Genome Analysis Toolkit), SnpEff, SnpSIFT, PolyPhen-2, this study describes a basic procedure to determine the genetic variations in the people with autism. It is noted that this is the first report on the application of WES method in research of the autism in Vietnam. The results obtained in the present study could be used as a basic guide for the WES data analysis.

Keywords: *Autism; bioinformatics; genetic diseases; next generation sequencing, whole exome sequencing*