

Phân tích biến dị trong hệ gen của 99 cá thể người dân tộc Kinh

Nguyễn Cường*, Nguyễn Đức Long, Nguyễn Nhật Hưng
Trịnh Thị Xuân, Chu Hoàng Hà, Trương Nam Hải

Viện Công nghệ sinh học, Viện Hàn lâm Khoa học và Công nghệ Việt Nam

Ngày nhận bài 27.7.2015, ngày chuyển phản biện 29.7.2015, ngày nhận phản biện 31.8.2015, ngày chấp nhận đăng 3.9.2015

Nghiên cứu này khai thác thông tin về các biến dị di truyền trong bộ gen của 99 người Kinh, được chiết xuất từ dữ liệu của dự án 1.000 hệ gen. Hai tập biến dị đã được lọc ra: tập các biến dị xuất hiện duy nhất trong hệ gen người Kinh (1.895.398 biến dị) (KHVonly) và tập biến dị xuất hiện với tần suất cao hơn 90% trong hệ gen người Kinh (1.556.750 biến dị) (KHV90). Trong đó, biến dị đa hình đơn nucleotide (SNP - Single Nucleotide Polymorphism) chiếm gần 99% tổng số biến dị; có sự tương đồng về phân bố của biến dị trên các nhiễm sắc thể (NST) trong 2 tập KHVonly và KHV90 với NST số 2 mang nhiều biến dị nhất lần lượt là 161.224 và 132.508 biến dị. Mật độ của các biến dị trên các NST không đồng đều, thể hiện ở độ biến động cao về mật độ biến dị tại các vị trí khác nhau trên NST (0-70 ngàn SNP/1 Mb) và có xu hướng tập trung cao ở 2 đầu mút của NST thường hoặc ở giữa NST giới tính. Phân tích vùng chức năng của hệ gen đã dự đoán rằng, chức năng gen nằm trong vùng exon sẽ bị tác động bởi các biến dị trong tập KHVonly nhiều hơn bởi các biến dị trong tập KHV90, lần lượt là 157.311 và 66.123 chức năng, nhưng không tìm thấy SNP gây bệnh trong 2 tập biến dị liên quan đến 27 bệnh di truyền, có thể là do tiêu chí chọn cá nhân có kiểu hình khỏe mạnh tham gia dự án 1.000 hệ gen. Các kết quả trong nghiên cứu này đã được tải lên cơ sở dữ liệu trực tuyến về biến dị di truyền của người Việt tại địa chỉ <https://vsnp.ibt.ac.vn>. Kết quả của nghiên cứu là nền tảng cho việc mô tả chi tiết toàn bộ hệ gen của người Kinh, là nguồn tài nguyên phong phú và có giá trị cho những nghiên cứu sắp tới về các biến dị di truyền, bệnh di truyền cũng như xác định danh tính ở cộng đồng người Việt.

Từ khóa: bệnh di truyền, dự án 1.000 hệ gen, giải trình tự gen thế hệ mới, hệ gen người Kinh, khoa học hình sự, tin sinh học.

Chỉ số phân loại 1.6

Đặt vấn đề

Các biến dị di truyền trong hệ gen người [bao gồm SNP, các đoạn chèn/xóa (DIP/Indel - Deletion/Insertion Polymorphism) cũng như những biến dị cấu trúc] có mối liên hệ mật thiết với các bệnh di truyền và sự đa dạng của loài người [1-3]. Các nghiên cứu chỉ ra rằng, có những biến dị hiếm gặp, gây ra những biểu hiện rõ rệt và chỉ ở những tính trạng nhất định như bệnh Huntington hay bệnh xơ nang (Cystic fibrosis), tuy nhiên có những căn bệnh gây ra bởi các biến dị di truyền mà cơ chế ảnh hưởng của chúng tới các chức năng của cơ thể rất phức tạp như bệnh đái đường, bệnh tim và các bệnh khác liên quan đến nhận thức [4, 5]. Các biến dị di truyền ảnh hưởng đến sự phát triển bệnh cũng như phản ứng của bệnh nhân đối với các loại thuốc và chất hóa học khác nhau [6, 7].

Nhằm tìm hiểu đầy đủ sự liên kết cũng như tác động

tổng hợp của các biến dị di truyền lên chức năng của hệ gen người, dự án 1.000 hệ gen đã được triển khai thực hiện. Dự án là sự hợp tác của các nhóm nghiên cứu đa ngành từ các viện nghiên cứu hàng đầu trên thế giới nhằm xây dựng một bản đồ hệ gen hoàn chỉnh của con người, áp dụng công nghệ giải trình tự gen thế hệ mới (NGS - Next Generation Sequencing) [8]. Với công nghệ NGS, việc giải trình tự các hệ gen có kích thước lớn như hệ gen người đã tiết kiệm được chi phí, thời gian và có độ chính xác cao hơn phương pháp giải trình tự thế hệ cũ [9, 10]. Được bắt đầu từ năm 2008, dự án đến nay đã hoàn thành giai đoạn 3, giải mã hơn 2.500 hệ gen người thuộc các cộng đồng, dân tộc khác nhau trên thế giới [8]. Trong tương lai, việc kết hợp đồng thời dữ liệu về các biến dị di truyền, các chú giải chức năng đầy đủ của hệ gen người cùng với dữ liệu về hệ phiên mã sẽ là lời giải cho tác động của các biến dị

*Tác giả liên hệ: Email: cuongnguyen@ibt.ac.vn

GENOME ANALYSIS OF 99 KINH PEOPLE

Summary

The authors report the whole genome analysis of 99 Kinh people extracted from 1,000 Genome Project data. They classified variants into two sets: variants that were uniquely present in Kinh genomes (KHVonly -1,895,398 variants) and variants that were present in more than 90% of Kinh samples (KHV90 -1,556,750 variants), in which 99% were SNP - Single Nucleotide Polymorphism. Both of the variants sets shared a similar pattern in the distribution of variants across chromosomes with chromosome 2 carrying highest number of variants (KHVonly: 161,224; KHV90: 132,508). The distribution of variants in a chromosome was not even, presented by the wide range of density values 0-70,000 SNP/1 Mb with the trend of focusing on both ends of chromosomes (autosomal chromosomes) or focusing in the center (sex chromosomes). Functional region analysis predicted that gene functions located in the exon regions were affected more by variants in KHVonly (157,311 functions) than by variants in KHV90 (66,123 functions). The authors found that no pathogenic SNP related to 27 genetic diseases, due to the criteria of choosing individuals with healthy appearance for sequencing of 1,000 genomes project. All the results of this study were uploaded to the web browser developed by the research group at the address "<https://vsnp.ibt.ac.vn>". The results are the foundation for detailed characterization of whole genome of Kinh people, providing an informative and valuable resource for upcoming researches on genetic variations, identification, and genetic diseases of Vietnamese population.

Keywords: *bioinformatics, forensics, genetic diseases, Kinh people genome, Next Generation Sequencing (NGS), 1,000 genome project.*

Classification number 1.6

di truyền đối với sự biểu hiện gen [11].

Việt Nam là nước có dân số đứng thứ 13 trên thế giới, chủ yếu là người Kinh (chiếm trên 85%) nên bộ gen của người Kinh cũng đã được giải trình tự trong dự án 1.000 hệ gen. Việc giải trình tự và lắp ráp đầy đủ hệ gen là bước đầu tiên trong việc thông hiểu các thông tin chứa trong các trình tự của hệ gen [12]. Vì vậy, trong nghiên cứu này chúng tôi đã thực hiện khai thác dữ liệu của dự án 1.000 hệ gen nhằm trích xuất các thông tin về biến dị di truyền trong hệ gen của người Kinh được giải trình tự trong dự án và tiến hành các phân tích cơ bản về phân bố và chức năng của các loại biến dị trong hệ gen của người Kinh. Việc chú giải hệ gen một cách đầy đủ và chất lượng sẽ là một định hướng tốt cho những nghiên cứu sau này.

Dữ liệu và phương pháp nghiên cứu

Dữ liệu

Dữ liệu được công bố về hệ gen người Kinh trong dự án 1.000 hệ gen (<http://www.1000genomes.org/>) phiên bản 20130502 (<ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/>) đáp ứng được các tiêu chuẩn chất lượng của dự án (độ bao phủ trên 4x). Những phân tích sâu đối với 2 tập biến dị: các biến dị chỉ xuất hiện ở cộng đồng người Kinh (KHVonly) và các biến dị xuất hiện với tần suất trên 90% ở cộng đồng người Kinh (KHV90) được thực hiện.

Phương pháp nghiên cứu

Tổng quan dữ liệu: công cụ VCFtools [13] được dùng để tách riêng dữ liệu di truyền của 99 người Kinh khỏi tập dữ liệu biến dị của thế giới. Từ dữ liệu trích xuất được, chúng tôi sử dụng công cụ được phát triển bởi Phòng Tin sinh học, Viện Công nghệ sinh học để lọc và phân loại các biến dị di truyền thành hai tập: KHVonly và KHV90. Các phân tích và so sánh cơ bản về phân bố và tần suất xuất hiện của biến dị trên NST được thực hiện.

Chú giải ảnh hưởng chức năng: công cụ SNPeff [14] được sử dụng với tùy chỉnh mặc định nhằm phân tích ảnh hưởng chức năng của hai tập biến dị. Các ảnh hưởng này được chú giải theo mức độ ảnh hưởng, vùng chức năng và tác động cụ thể đến chức năng hệ gen dựa vào thuật ngữ mô tả trình tự tiêu chuẩn (Sequence Ontology) [15].

Biến dị liên quan đến bệnh di truyền: nhằm xác định biến dị liên quan tới bệnh di truyền, danh sách 27 bệnh di truyền đang được nghiên cứu tại Việt Nam tham khảo từ danh mục của Lê Đình Lương (bảng 1). Từ cơ sở dữ liệu ClinVar (NCBI) [16], thông tin về các biến dị được phân loại là biến dị gây bệnh liên quan đến danh sách 27 bệnh trên được trích xuất ra nhờ công cụ phát triển tại Phòng Tin sinh học trước khi đem so sánh với các tập biến dị KHVonly và KHV90. Chúng tôi cũng đã xây dựng và phát triển cơ sở dữ liệu trực tuyến nhằm tạo thuận tiện cho việc tra cứu, dựa trên hệ thống phần mềm Gbrowser (Generic Genome Browser) [17].

Bảng 1: danh sách các bệnh di truyền hiện đang được tập trung nghiên cứu tại Việt Nam tham khảo danh mục của Lê Đình Lương

STT	Bệnh di truyền	STT	Bệnh di truyền
1	Alpha-1-antitrypsin deficiency	15	Fragile X syndrome
2	Amyotrophic lateral sclerosis	16	Hemophilia
3	Alzheimer's disease	17	Hereditary Hemochromatosis
4	Ataxia telangiectasia	18	Huntington's disease
5	Gaucher disease	19	Myotonic dystrophy
6	Inherited breast and ovarian cancer	20	Neuro fibromatosis type 1
7	Hereditary nonpolyposis colon cancer	21	Phenylketonuria
8	Charcot - Marie - Tooth	22	Adult Polycystic Kidney Disease
9	Congenital adrenal hyperplasia	23	Sickle cell disease
10	Cystic fibrosis	24	Spinocerebellar ataxia, type 1
11	Duchenne muscular dystrophy	25	Spinal muscular atrophy
12	Dystonia	26	Thalassemias
13	Fanconi anemia	27	Tay-Sachs Disease
14	Factor V		

Kết quả nghiên cứu

Tổng quan dữ liệu

Trong lần công bố gần nhất của dự án 1.000 hệ gen, chỉ có dữ liệu về hệ gen của 99 người Kinh trong số 124 người được đưa vào danh sách giải trình tự, với độ bao phủ trung bình của dữ liệu lên hệ gen người tham chiếu (GChR37) theo báo cáo của dự án là 8,7x với độ bao phủ tối thiểu là 4x và tối đa là 19x. Dữ liệu từ file giải trình tự được trích xuất và đối chiếu với thông tin liên quan đến những người trong danh sách thì thấy 99 người Kinh được công bố chỉ bao gồm các cặp bố mẹ và các cá nhân độc lập (bảng 2). Trong số 25 hệ gen không được công bố, có 23 hệ gen không có bất kỳ thông tin nào về việc giải trình tự, 2 hệ gen còn lại có thông tin về việc giải trình tự nhưng không đạt chất lượng, 22/25 hệ gen không được công bố là của người con trong các gia đình tham gia dự án. Ngoài ra, hệ gen

của người anh/em trong cặp anh - em đã không được công bố. Chính vì thế, việc phân tích các tính trạng di truyền dựa vào huyết thống (bộ ba bố - mẹ - con) đã không thể thực hiện được.

Bảng 2: tổng quan dữ liệu về cộng đồng người Kinh trong dự án 1.000 hệ gen

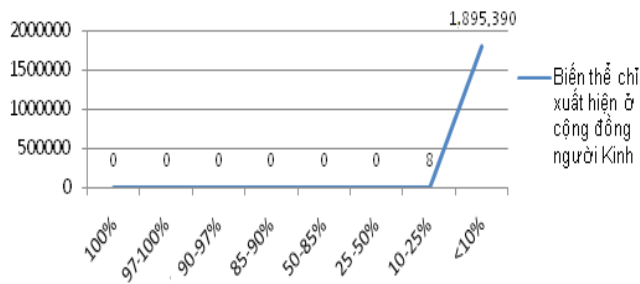
Dữ liệu	Tổng số	Gia đình	Nam			Nữ			Anh/Chị/Em	
Số mẫu giải trình tự	124	65	Bố	Con	Cá nhân*	Mẹ	Con	Cá nhân	Nam	Nữ
			22	14	25	21	8	34	1	1
			61			63			2	
Số mẫu công bố	99	21	Bố	Con	Cá nhân	Mẹ	Con	Cá nhân	Nam	Nữ
			21	0	25	20	0	33	1	0
			46			53			1	

*Cá nhân là những người độc lập, không có quan hệ huyết thống với những người Kinh còn lại được giải trình tự hệ gen trong dự án

Tổng quan về tập biến dị KHVonly được trích xuất từ file dữ liệu chung của dự án được thể hiện trong bảng 3. Số lượng biến dị là gần 1,9 triệu, trong đó số lượng SNP chiếm gần 99% (1,89 triệu), số lượng Indel và biến dị cấu trúc ít hơn rất nhiều, lần lượt là 5.942 và 1.045. Trong số những NST thường, NST số 2 chứa nhiều biến dị nhất, chiếm 8,5% tổng số biến dị, trong đó có 160.901 SNP; 239 Indel và 84 biến dị cấu trúc. NST 22 chứa ít biến dị nhất với 23.849 SNP, 33 Indel và 7 biến dị cấu trúc. NST giới tính X mang lượng Indel nhiều nhất trong tất cả các NST, gần 3.500 Indel, chiếm khoảng 60% tổng số lượng Indel, nhiều gấp 350 lần số lượng Indel trên NST giới tính Y và gấp hơn 14 lần lượng Indel trên NST số 2. Tổng biến dị của toàn bộ các cá thể được giải trình tự trong dự án 1.000 hệ gen cũng được chúng tôi tổng hợp từ file dữ liệu chung và đưa vào so sánh trong bảng 3.

Bảng 3: số lượng các biến dị trong tập KHVonly so sánh với số lượng biến dị trên thế giới

NST	SNP	Indel	Biến dị cấu trúc	Tổng biến dị	Tổng biến dị trên thế giới
1	144.467	186	83	144.736	6.468.094
2	160.901	239	84	161.224	7.081.600
3	130.323	151	67	130.541	5.832.276
4	126.336	164	71	126.571	5.732.585
5	118.541	166	66	118.773	5.265.763
6	110.514	147	61	110.722	5.024.119
7	104.016	152	63	104.231	4.716.715
8	103.169	140	73	103.382	4.597.105
9	79.524	95	43	79.662	3.560.687
10	87.191	118	48	87.357	3.992.219
11	90.289	145	60	90.494	4.045.628
12	84.742	124	50	84.916	3.868.428
13	62.919	97	27	63.043	2.857.916
14	57.776	91	37	57.904	2.655.067
15	54.355	86	20	54.461	2.424.689
16	61.103	83	58	61.244	2.697.949
17	51.080	62	25	51.167	2.329.288
18	50.075	58	32	50.165	2.267.185
19	39.245	54	22	39.321	1.832.506
20	41.265	52	23	41.340	1.812.841
21	24.351	41	13	24.405	1.105.538
22	23.849	33	7	23.889	1.103.547
X	82.392	3.447	11	85.850	3.468.093
Y	1.716	11	1	1.728	62.042
Tổng	1.888.423	5.942	1.045	1.895.398	84.739.838



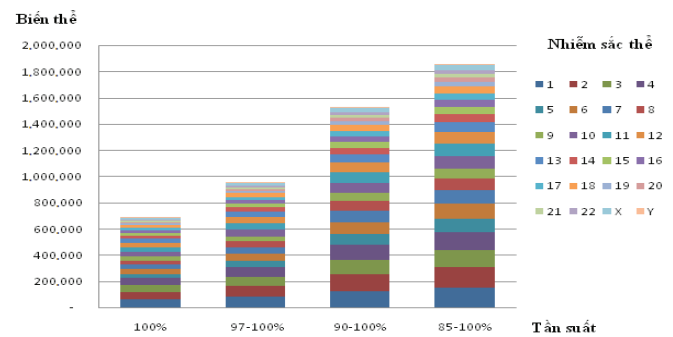
Hình 1: tần suất các biến dị chỉ xuất hiện trong cộng đồng người Kinh

Tần suất xuất hiện của các biến dị trong tập KHVonly đã được tính toán và kết quả cho thấy không có biến dị nào xuất hiện với tần suất 100% trong hệ gen của cộng đồng người Kinh được giải trình tự (hình 1). Biến dị chỉ bắt đầu xuất hiện ở khoảng tần suất dưới 25% và chủ yếu là ở dưới 10%. Trong khoảng từ 10-25%, chỉ có 8 biến dị xuất hiện (bảng 4), số lượng biến dị với tần suất dưới 10% là rất lớn, khoảng 1,8 triệu biến dị.

Bảng 4: danh sách 8 biến dị xuất hiện trong tập KHVonly với tần suất 10-25%

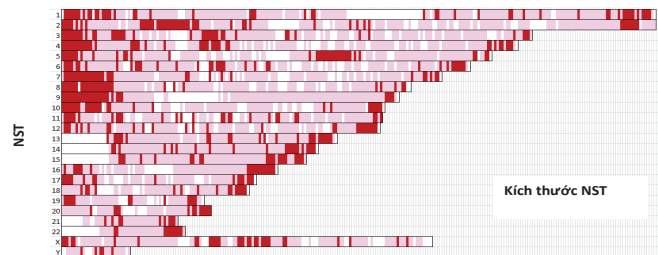
Biến dị	NST	Vị trí	Tần suất (%)	Dự đoán ảnh hưởng chức năng
rs373032880	7	18796400	10,1	Rất thấp
rs372829488	7	155810198	12,1	Rất thấp
rs377210607	9	21158496	10,1	Rất thấp
rs371006249	11	8333813	11,1	Rất thấp
rs367788580	11	112037409	10,1	Rất thấp
rs373524703	13	66185412	12,1	Rất thấp
rs371883058	14	51908186	11,1	Rất thấp
rs547676368	15	72111947	10,1	Rất thấp

Thống kê cho thấy, khoảng 685 ngàn biến dị xuất hiện ở 100% hệ gen người Kinh tham gia dự án (hình 2), bằng một nửa lượng biến dị có tần suất xuất hiện từ 90-100% (1,55 triệu) và bằng 1/3 lượng biến dị có tần suất xuất hiện từ 85% (1,85 triệu). Chúng tôi thực hiện phân tích sâu đối với tập 1,53 triệu biến dị có tần suất xuất hiện trên 90% (KHV90).



Hình 2: tần suất xuất hiện biến dị di truyền trong cộng đồng người Kinh được giải trình tự trong dự án 1.000 hệ gen

Kết quả phân tích mật độ biến dị trên các NST thể hiện sự không đồng đều, có những khu vực trên NST có mật độ biến dị rất cao (50-70 ngàn SNP/1 Mb), trong khi có những vùng không có sự xuất hiện của một biến dị nào như phần đầu của NST 13, 14 và 15 (hình 3). Các NST thường đều có xu hướng tập trung cao biến dị ở hai đầu mút của NST, trừ các NST 13-15, 21, 22 (biến dị xuất hiện ít ở đầu NST), còn các NST giới tính X và Y có biến dị tập trung ở vị trí giữa NST.



Hình 3: phân bố của biến dị di truyền trên NST

Giá trị một ô tương đương 1 Mb. Các ô màu đỏ thể hiện vùng NST tập trung nhiều biến dị (nằm trong 20% các ô có lượng biến dị nhiều nhất của NST đó). Các ô màu trắng thể hiện vùng NST tập trung ít biến dị (nằm trong 20% các ô có lượng biến dị ít nhất của NST đó). Các ô màu hồng thể hiện vùng NST có mật độ trung bình.

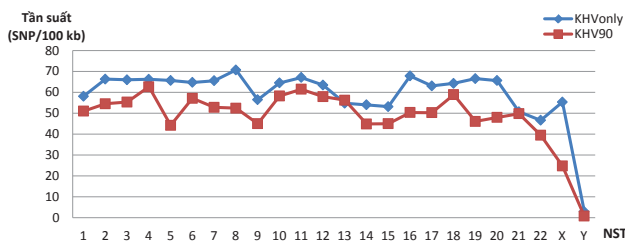
Bảng 5 thể hiện sự tương quan giữa tập biến dị KHV90 với tập KHVonly. Lượng biến dị trong tập KHV90 (1,53 triệu) ít hơn so với tập KHVonly (1,89 triệu), tuy nhiên có sự tương đồng về phân bố các biến dị trên NST. Riêng NST 13, lượng biến dị trong tập KHV90 (64.711) tập trung nhiều hơn so với lượng biến dị trong tập KHVonly (63.043). Số lượng biến dị tập trung trên NST giới tính của tập KHV90 (X: 38.461; Y: 446) ít hơn nhiều lần so với số lượng biến dị của tập KHVonly (X: 85.850; Y: 1.728).

Bảng 5: so sánh số lượng các biến dị trong tập KHV90 và KHVonly

NST	SNP	Indel	Biến dị cấu trúc	Tổng biến dị KHV90	Tổng biến dị KHVonly
1	106.944	19.973	95	127.012	144.736
2	111.843	20.601	64	132.508	161.224
3	92.622	16.891	53	109.566	130.541
4	101.990	17.714	41	119.745	126.571
5	67.197	12.715	48	79.960	118.773
6	83.050	14.579	76	97.705	110.722
7	70.747	13.225	56	84.028	104.231
8	65.836	10.798	40	76.674	103.382
9	54.209	9.228	41	63.478	79.662
10	66.487	12.339	61	78.887	87.357
11	70.147	12.782	42	82.971	90.494
12	65.150	12.335	37	77.522	84.916
13	54.546	10.139	26	64.711	63.043
14	40.344	7.762	28	48.134	57.904
15	38.472	7.568	54	46.094	54.461
16	38.766	6.639	52	45.457	61.244
17	33.584	7.189	23	40.796	51.167
18	33.999	6.923	31	40.953	50.165
19	22.615	4.590	12	27.217	39.321
20	25.669	4.556	17	30.242	41.340
21	19.501	4.426	17	23.944	24.405
22	16.902	3.328	9	20.239	23.889
X	31.376	7.058	27	38.461	85.850
Y	430	16	0	446	1.728
Tổng	1.312.426	243.374	950	1.556.750	1.895.398

Do các NST khác biệt về kích thước nên tần suất xuất hiện của biến dị trên các NST được sử dụng để xác định độ tập trung của các biến dị KHV90 và KHVonly (hình 4). Tần suất xuất hiện của biến dị trên một NST được tính theo công thức:

$$\text{Tần suất} = \frac{\text{Tổng số biến thể}}{\text{Kích thước NST}}$$



Hình 4: so sánh tần suất xuất hiện trên NST của các biến dị trong 2 tập KHV90 và KHVonly

Do có lượng biến dị ít hơn, nên tần suất xuất hiện trung bình của các biến dị KHV90 so với các biến

dị KHVonly cách biệt nhau rất nhiều, lần lượt là 59,1 SNP/100 kb và 48,6 SNP/100 kb. Phần lớn các biến dị KHVonly xuất hiện trên các NST thường với cùng tần suất (NST 2-7; 10-12; 18-20), trong khoảng 65 SNP/100 kb. Tần suất biến dị cao nhất trong tập KHVonly là trên NST số 8, khoảng 70 SNP/100 kb. Tần suất thấp nhất ghi nhận được là 46 SNP/100 kb ở NST thường số 22.

Tần suất trung bình của biến dị KHV90 trên NST thường có sự dao động lớn, từ 44,84 SNP/100 kb (NST 14) đến 58,86 SNP/100 kb (NST 18). Tần suất xuất hiện của biến dị KHV90 trên NST 4, 13 và 21 lần lượt là 62,66 SNP/100 kb; 56,2 SNP/100 kb và 49,75 SNP/100 kb, tương đương với tần suất xuất hiện của biến dị KHVonly (lần lượt là 66,3 SNP/100 kb; 54,76 SNP/100 kb và 50,71 SNP/100 kb). Đối với NST giới tính, tần suất xuất hiện biến dị của tập KHV90 và KHVonly rất khác biệt, lần lượt là 55,3 SNP/100 kb và 24,7 SNP/100 kb trên NST X.

Phân tích ảnh hưởng của SNP lên chức năng của hệ gen

SNP là dạng biến dị phổ biến nhất, chiếm hơn 90% tổng số biến dị của hệ gen, vì thế việc phân tích ảnh hưởng chức năng của các biến dị trong tập KHVonly và tập KHV90 đã được thực hiện. Ý nghĩa của các mức độ ảnh hưởng chức năng được tóm lược trong bảng 6.

Bảng 6: mức độ ảnh hưởng của biến dị lên hệ gen

Ảnh hưởng	Ý nghĩa	Ví dụ
Nghiêm trọng	- Ảnh hưởng nghiêm trọng tới protein, gây ra sự cắt cụt protein, mất chức năng hoặc khởi động quá trình phân hủy các protein chưa trưởng thành	- Mất đoạn lớn trên NST (hơn 1%) - Mất đoạn lớn exon - Biến dị chèn codon kết thúc
Trung bình	- Làm giảm hiệu năng của protein	- Biến dị gây thay đổi acid amin - Biến dị gây thay đổi acid amin, đồng thời gây mất một hoặc nhiều acid amin khác
Thấp	- Ảnh hưởng không đáng kể hoặc không làm thay đổi hoạt động của protein	- Biến dị gây thay đổi bộ ba mã hóa nhưng không làm thay đổi acid amin
Rất thấp	- Xảy ra ở các biến dị không mã hóa protein hoặc nằm trong vùng không mã hóa - Việc dự đoán là rất khó khăn hoặc không có bằng chứng về ảnh hưởng	- Biến dị ở trong vùng gen không có phiên mã - Biến dị ở vùng intron - Biến dị ở vùng Downstream của một gen

Kết quả phân tích các ảnh hưởng chức năng của tập biến dị KHVonly được tóm tắt trong bảng 7. Hầu hết các biến dị có mức độ ảnh hưởng thấp hoặc rất

thấp tới chức năng của hệ gen. Với 1,9 triệu biến dị có thể khiến cho khoảng 2,2 ngàn chức năng của hệ gen có nguy cơ bị ảnh hưởng nghiêm trọng và hơn 53 ngàn chức năng có mức độ ảnh hưởng trung bình. NST số 1 có số lượng các chức năng bị ảnh hưởng nghiêm trọng lớn nhất (211 chức năng) trong khi NST giới tính Y ít bị các ảnh hưởng nghiêm trọng (1 chức năng). Tổng cộng có hơn 49 ngàn gen có khả năng đột biến Missense, hơn 28 ngàn gen có khả năng bị đột biến Silent. Đột biến Nonsense khá ít (995 tác động), trong đó 2 NST giới tính X và Y là thấp nhất, lần lượt là 7 và 0. Đặc biệt, số lượng các chức năng có khả năng xảy ra đột biến Nonsense cao nhất ở NST 19 (130 chức năng), cao hơn 30% so với NST có kích thước lớn nhất là NST số 1 (105 chức năng) và cao hơn mức trung bình 6,19 lần (21 chức năng/NST).

Bảng 7: ảnh hưởng chức năng của các biến dị trong tập KHVonly

NST	Mức độ ảnh hưởng*				Tác động về chức năng		
	Nghiêm trọng	Trung bình	Thấp	Rất thấp	Missense	Nonsense	Silent
1	211	5.045	392.682	504.776	4.751	105	2.541
2	136	4.219	520.757	539.777	3.872	65	2.177
3	130	3.300	371.313	475.477	2.869	54	1.583
4	78	2.009	225.453	368.108	1.927	49	1.058
5	79	2.434	232.650	359.840	1.959	38	1.252
6	114	2.663	270.815	314.500	2.483	38	1.357
7	128	2.383	299.083	360.598	2.220	50	1.351
8	113	2.257	269.603	346.947	2.063	39	1.047
9	66	2.017	199.796	235.743	1.955	34	1.030
10	111	2.060	223.831	268.688	1.947	45	999
11	118	3.181	190.163	358.616	3.110	32	1.966
12	116	2.806	258.513	338.662	2.558	57	1614
13	26	815	77.587	133.066	787	16	394
14	50	1.979	116.775	233.196	1.772	19	930
15	76	1.760	162.695	225.180	1.571	18	855
16	141	2.594	121.838	278.575	2.280	50	1.492
17	92	3.094	137.558	269.896	2.823	62	1.743
18	42	983	112.578	168.774	853	28	391
19	251	3.669	87.359	221.158	3.456	130	1.986
20	69	1.223	1.081	131.103	1.226	34	750
21	24	601	58.193	80.889	568	12	449
22	56	1.186	59.683	107.710	1.161	13	601
X	29	1.404	119.659	196.610	1.143	7	796
Y	1	12	340	2.871	12	0	3
Tổng	2.257	53.694	4.510.005	6.520.760	49.366	995	28.365

*Các mức độ ảnh hưởng được mô tả dựa vào thuật ngữ miêu tả trình tự gen tiêu chuẩn (Sequence Ontology)

Ảnh hưởng chức năng của các biến dị trong tập KHV90 cũng đưa ra những kết quả tương tự nhưng với số lượng ít hơn rất nhiều (bảng 8). Cụ thể, số lượng các ảnh hưởng chức năng ở mức nghiêm trọng và trung bình có thể xảy ra bởi các biến dị này lần lượt là 711 và 13.853, bằng 30% số lượng chức năng bị ảnh hưởng

cùng mức gây ra bởi các biến dị trong tập KHVonly. Trong khi đó, các ảnh hưởng chức năng bị ảnh hưởng ở mức thấp và rất thấp gây ra do các biến dị trong tập KHV90 giảm không đáng kể (lần lượt là 3,7 triệu và 5,07 triệu chức năng) so với các ảnh hưởng cùng mức gây ra bởi các biến dị trong tập KHVonly (lần lượt là 4,5 triệu và 6,5 triệu chức năng).

Khi tiến hành so sánh về cơ chế tác động, chúng tôi nhận thấy sự khác biệt rất rõ rệt về tỷ lệ các đột biến ở 2 tập biến dị. Với tập biến dị KHV90, số lượng chức năng bị đột biến Silent và Missense là gần tương đương nhau (lần lượt là 11.973 chức năng và 10.108 chức năng). Các tác động gây đột biến Silent nhiều hơn tác động gây đột biến Missense, đối ngược với tác động của biến dị trong tập KHVonly. Các biến dị cũng ít có khả năng gây ra đột biến Nonsense trên NST thường và giới tính. Số lượng các chức năng có khả năng bị đột biến Nonsense cao đặc biệt ở NST 11 (12 chức năng), trong khi ở các NST khác, số lượng các chức năng bị cùng tác động về chức năng chỉ dao động từ 0-4.

Bảng 8: ảnh hưởng chức năng của các biến dị trong tập KHV90

NST	Mức độ ảnh hưởng*				Tác động về chức năng		
	Nghiêm trọng	Trung bình	Thấp	Rất thấp	Missense	Nonsense	Silent
1	60	1.346	403.884	419.322	1.061	2	1.189
2	36	1.264	515.794	431.967	794	3	812
3	24	860	261.376	381.048	606	0	680
4	20	569	202.175	319.224	410	0	515
5	31	565	146.005	225.776	337	0	435
6	7	1.486	258.686	268.319	691	1	778
7	39	790	240.370	271.057	457	1	458
8	38	547	211.854	243.188	326	0	381
9	22	441	150.502	181.048	389	2	453
10	45	500	194.751	230.691	376	0	521
11	82	791	142.353	300.910	753	12	907
12	33	607	189.333	274.479	508	1	645
13	10	179	86.279	131.915	123	2	181
14	21	365	82.309	182.468	310	0	421
15	24	389	139.571	181.935	339	0	486
16	29	500	89.889	205.899	395	1	594
17	55	786	130.466	210.414	576	4	692
18	11	235	89.893	141.678	184	0	201
19	44	755	49.195	141.229	680	4	679
20	18	239	456	87.271	232	0	308
21	16	162	41.718	75.806	134	0	104
22	32	279	43.544	86.659	274	1	344
X	14	198	40.952	82.115	153	0	189
Y	0	0	119	751	0	0	0
Tổng	711	13.853	3.711.474	5.075.169	10.108	34	11.973

*Các mức độ ảnh hưởng được mô tả dựa vào thuật ngữ miêu tả trình tự gen tiêu chuẩn (Sequence Ontology)

Ảnh hưởng của hai tập biến dị đến chức năng hệ gen được thể hiện ở bảng 9. Phân tích vùng phân bố của các biến dị trong tập KHVonly cho thấy, phần lớn các SNP nằm trong vùng intron, chiếm đến gần 40% số lượng SNP. Các biến dị có vùng chức năng chưa phân định (NONE) chiếm hơn 36% tổng số SNP. Chỉ khoảng 1,5% số lượng biến dị nằm trong vùng exon (157.311 chức năng). Dựa vào thuật ngữ mô tả trình tự gen, các biến dị ở các vùng khác nhau sẽ có những mức độ ảnh hưởng tới chức năng khác nhau. Các biến đổi trong vùng exon và vùng cắt intron (splice site) được đánh giá là có mức độ ảnh hưởng nghiêm trọng tới chức năng của hệ gen.

Bảng 9: so sánh về phân bố của các ảnh hưởng chức năng gây bởi các biến dị

Vùng chức năng	KHVonly		KHV90	
	Tổng số chức năng	Trung bình (%)	Tổng số chức năng	Trung bình (%)
Downstream	651.640	6,363	498.909	6,264
Exon	157.311	1546	66.123	0,862
Intergenic	865.951	9,702	771.995	10,818
Intron	4.230.085	38,927	3.308.411	38,282
None	4.473.256	36,520	3.696.253	36,939
Splice Site Acceptor	437	0,005	230	0,003
Splice Site Donor	673	0,008	159	0,002
Splice Site Region	11.568	0,125	6.685	0,081
Transcript	26.111	0,184	25	0,001
Upstream	607.973	6,034	503.599	6,283
UTR 3'	45.640	0,439	26.949	0,337
UTR 5'	16.804	0,167	9.853	0,133

Phân tích các SNP liên quan đến các bệnh di truyền

Hiện nay, có rất nhiều cơ sở dữ liệu sinh học lưu trữ thông tin về các biến dị liên quan đến các bệnh di truyền. Trong nghiên cứu này, chúng tôi sử dụng cơ sở dữ liệu ClinVar (NCBI) để lọc ra những biến dị gần nhãn “gây bệnh” trong danh sách 27 bệnh di truyền đang được nghiên cứu tại Việt Nam từ 2 tập biến dị KHV90 và KHVonly. Kết quả cho thấy, không có biến dị nào trong 2 tập biến dị này có nguy cơ gây bệnh trong danh sách 27 bệnh đã được liệt kê.

Cơ sở dữ liệu trực tuyến về gen người Việt

Dữ liệu về hệ gen người Kinh có thể truy cập tại địa chỉ <https://vsnp.ibt.ac.vn>. Người dùng có thể tìm kiếm dữ liệu bằng nhiều cách khác nhau như nhập mã biến dị, vị trí trên NST, tên bệnh di truyền, tên gen... Hệ thống sẽ tự động liên kết với các cơ sở dữ liệu quốc tế như OMIM, NCBI, dbSNP và trích xuất dữ liệu cho

người dùng dễ dàng tra cứu các thông tin liên quan. Hiện tại, các chức năng cơ bản của hệ thống cơ sở dữ liệu trực tuyến đã hoạt động, các chức năng nâng cao vẫn đang được phát triển.

Thảo luận

Như vậy, 99 hệ gen của người Kinh đã được giải trình tự và công bố trong dự án 1.000 hệ gen. Tuy nhiên, dữ liệu hệ gen của 23 người con trong các gia đình tham gia dự án chưa được công bố khiến cho việc phân tích các SNPs và Indels liên quan đến huyết thống vẫn chưa thực hiện được. Với 8 biến dị xuất hiện với tần suất 10-25% (khoảng tần suất cao nhất được ghi nhận) và 1,89 triệu biến dị xuất hiện với tần suất dưới 10%, thể hiện sự đặc trưng cho từng cá thể trong cộng đồng người Việt. Kết quả cho thấy sự đa dạng cao trong hệ gen của các cá thể trong cộng đồng người Kinh, nhưng cũng rất đặc trưng khi so với các cộng đồng khác. Các biến dị trong tập KHVonly có tần suất không cao, tuy nhiên việc kiểm chứng ở một quần thể lớn hơn cần được thực hiện để kết quả đạt được có độ tin cậy cao và có ý nghĩa thống kê hơn nữa. Chúng tôi kỳ vọng rằng khi triển khai ở quần thể lớn, những biến dị KHVonly có tần suất cao trên 75% đặc trưng cho người Kinh sẽ được phát hiện thêm.

Mỗi biến dị có thể bị biến đổi theo nhiều hướng, mà sự biến đổi đó có tác động tới hệ gen ở các mức độ khác nhau. Với công cụ SNPEff, thông tin thực nghiệm về các biến đổi được thu thập từ các cơ sở dữ liệu như dbSNP, OMIM... sau đó được tổng hợp lại, tính toán và báo cáo một cách chi tiết về vị trí, mức độ ảnh hưởng cũng như tác động cụ thể của từng biến đổi xảy ra ở những biến dị mong muốn [14]. Tuy nhiên, sự liên kết giữa vị trí của biến dị với mức độ ảnh hưởng của nó tới chức năng hệ gen đã được thực hiện nhưng chỉ mang tính tương đối, nhằm mục đích chuẩn hóa và đồng bộ hóa kết quả của quá trình phân loại các ảnh hưởng chức năng [18]. Trong thời gian tới, chúng tôi sẽ tiến hành so sánh các chương trình chú giải chức năng khác nhau để tìm ra câu trả lời đúng nhất.

Khi so sánh về ảnh hưởng chức năng của 2 nhóm biến dị, chúng tôi nhận thấy các biến dị trong tập KHVonly gây ra nhiều ảnh hưởng nghiêm trọng hơn so với các biến dị xuất hiện trong tập KHV90. Kết quả này có thể được giải thích dựa vào tiêu chí lựa chọn cá nhân tham gia giải trình tự của dự án 1.000 hệ gen, đó là lựa chọn những cá nhân có kiểu hình khỏe mạnh, không có biểu hiện mắc một loại bệnh nào rõ rệt [8].

Việc tìm kiếm các biến dị gây bệnh dựa vào cơ sở dữ liệu ClinVar đã không đưa lại kết quả nào. Điều này có thể được lý giải bằng phương pháp lựa chọn biến dị trên ClinVar: chỉ những biến dị được gắn nhãn “gây bệnh” mới được chọn, tức là những người mang biến dị này chắc chắn sẽ mắc các căn bệnh di truyền được liệt kê.

Những kết quả của nghiên cứu này là bước đi đầu tiên trong việc xây dựng một bản mô tả chi tiết về toàn bộ hệ gen của người Việt, phục vụ cho các nghiên cứu sau này về biến dị như bệnh di truyền và xác định danh tính. Trong các nghiên cứu sắp tới, thông tin từ các Indel liên quan đến bệnh di truyền về giới tính trên NST X sẽ được đi sâu phân tích. Việc so sánh các biến dị của người Kinh với những bộ kit sử dụng chỉ thị biến dị cũng sẽ được tiến hành để phát hiện đầy đủ các SNP và Indel có nguy cơ gây bệnh ở cộng đồng người Kinh.

Những hiểu biết về các biến dị di truyền, đặc biệt là về SNP, mang đến hy vọng về các phương pháp điều trị đích và phát triển thuốc cá thể hóa [19]. Chúng tôi mong rằng, những thông tin thu thập được sẽ giúp định hướng các nghiên cứu thực nghiệm nhằm phát triển và chuẩn hóa các bộ kit chẩn đoán bệnh di truyền, lai ghép di truyền và ung thư, từ đó nâng cao sức khỏe của người Kinh nói riêng và người Việt nói chung. Các kết quả và thông tin liên quan sẽ được cập nhật thường xuyên trên cơ sở dữ liệu trực tuyến của Viện Công nghệ sinh học, Viện Hàn lâm Khoa học và Công nghệ Việt Nam tại địa chỉ <https://vsnp.ibt.ac.vn>.

Lời cảm ơn

Chúng tôi xin chân thành cảm ơn sự hỗ trợ về kinh phí từ nhiệm vụ KH&CN theo chức năng của Phòng thí nghiệm trọng điểm công nghệ gen, Viện Công nghệ sinh học, Viện Hàn lâm Khoa học và Công nghệ Việt Nam (mã số: NV05-PTNTĐ 2015).

Tài liệu tham khảo

[1] The ENCODE Project Consortium (2012), “An integrated encyclopedia of DNA elements in the human genome”, *Nature*, **489**(7414), pp.57-74.

[2] Gonzaga-Jauregui C, Lupski J.R, Gibbs R.A (2012), “Human Genome Sequencing in Health and Disease”, *Annual Review of Medicine*, **63**(1), pp.35-61.

[3] Korbel J.O, et al (2007), “Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome”, *Science*, **318**(5849), pp.420-426.

[4] Singh M, et al (2011), “SNP-SNP interactions within APOE gene influence plasma lipids in postmenopausal osteoporosis”, *Rheumatol Int*, **31**(3), pp.421-423.

[5] Manolio T.A, Brooks L.D, Collins F.S (2008), “A HapMap harvest of insights into the genetics of common disease”, *The Journal of Clinical Investigation*, **118**(5), pp.1590-1605.

[6] Rauch A, et al (2010), “Genetic variation in IL28B is associated with chronic hepatitis C and treatment failure: a genome-wide association study”, *Gastroenterology*, **138**(4), pp.1338-1345, pp.1345.e1-7.

[7] Pang G.S, et al (2009), “Predicting potentially functional SNPs in drug-response genes”, *Pharmacogenomics*, **10**(4), pp.639-653.

[8] Khurana E, et al (2013), “Integrative annotation of variants from 1092 humans: application to cancer genomics”, *Science*, **342**(6154), p.1235587.

[9] McKernan K.J, et al (2009), “Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding”, *Genome Res*, **19**(9), pp.1527-1541.

[10] Soon W.W, Hariharan M, Snyder M.P (2013), “High-throughput sequencing for biology and medicine”, *Molecular Systems Biology*, **9**, pp.1-14.

[11] Haraksingh R.R, Snyder M.P (2013), “Impacts of Variation in the Human Genome on Gene Regulation”, *Journal of Molecular Biology*, **425**(21), pp.3970-3977.

[12] Moore B, Fan G, Eilbeck K (2010), “SOBA: sequence ontology bioinformatics analysis”, *Nucleic Acids Research*, **38**(suppl 2), pp.W161-W164.

[13] Danecek P, et al (2011), “The variant call format and VCFtools”, *Bioinformatics*, **27**(15), pp.2156-2158.

[14] Pablo Cingolani A.P, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J Land, Xiangyi Lu, Douglas M Ruden (2012), “A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3”, *Fly (Austin)*, **6**(2), pp.80-92.

[15] Gene Ontology Consortium (2015), “Gene Ontology Consortium: going forward”, *Nucleic Acids Res*, **43**(Database issue), pp.D1049-D1056.

[16] Melissa J, Landrum J.M.L, George R Riley, Wonhee Jang, Wendy S Rubinstein, Deanna M Church, Donna R Maglott (2013), “ClinVar: public archive of relationships among sequence variation and human phenotype”, *Nucleic Acids Research*, pp.1-6.

[17] Donlin M.J (2009), “Using the Generic Genome Browser (GBrowse)”, *Current Protocols in Bioinformatics*, **28**.

[18] Eilbeck K, et al (2009), “Quantitative measures for the management and comparison of annotated genomes”, *BMC Bioinformatics*, **10**(1), p.67.

[19] Laing R.E, et al (2011), “The role and impact of SNPs in pharmacogenomics and personalized medicine”, *Curr Drug Metab*, **12**(5), pp.460-486.