

Nghiên cứu ứng dụng tin sinh học xác định biến dị di truyền trên bệnh nhân bệnh bạch cầu dòng tủy mạn tính

Hoàng Quốc Huy¹, Trịnh Thị Xuân², Phạm Quang Huy¹, Nguyễn Văn Lâm¹, Trần Công Hoàng³, Vũ Thị Bích Hương³, Lê Quang Tường³, Hoàng Thị Hồng³, Bạch Quốc Khánh³, Nguyễn Anh Trí³, Dương Quốc Chính³, Nguyễn Cường^{1*}

¹Phòng Tin sinh học, Viện Công nghệ sinh học, Viện Hàn lâm KH&CN Việt Nam

²Khoa Công nghệ thông tin, Viện Đại học mở Hà Nội

³Khoa Di truyền và Sinh học phân tử, Viện Huyết học - Truyền máu Trung ương

Ngày nhận bài 16/1/2017, ngày chuyển phản biện 20/1/2017, ngày nhận phản biện 9/2/2017, ngày chấp nhận đăng 23/2/2017

Bệnh bạch cầu dòng tủy mạn tính (chronic myeloid leukemia - CML) là một bệnh ác tính hệ tạo máu thuộc nhóm hội chứng tăng sinh tủy mạn ác tính, đặc trưng bởi sự tăng sinh các tế bào dòng bạch cầu hạt biệt hóa. Hiện nay Việt Nam đã có thuốc để điều trị bệnh CML nhưng do xuất hiện các biến dị di truyền trong hệ gen làm tăng khả năng kháng thuốc của bệnh nhân. Vì vậy, việc xác định các biến dị có vai trò quan trọng trong điều trị bệnh. Nghiên cứu đã tiến hành tách chiết RNA trên tổ hợp gen BCR/ABL1 gây nên bệnh CML trên 10 mẫu tủy từ các bệnh nhân CML khác nhau tại Viện Huyết học - Truyền máu Trung ương, sau đó giải trình tự bằng máy giải trình tự thế hệ mới Illumina MiSeq, đánh giá và tiền xử lý dữ liệu, ánh xạ trình tự vào hệ gen tham chiếu và xác định các biến dị trên từng bệnh nhân. Kết quả cho thấy, trên 10 mẫu giải trình tự của các bệnh nhân khác nhau thu được tổng số 18 biến dị, trong đó có 2 biến dị xuất hiện trên 8 bệnh nhân ở vị trí 133747603, 133747604 và 3 biến dị xuất hiện trên 6 bệnh nhân ở vị trí 133755142, 133755143, 133755146; đặc biệt có 4 biến dị ở vị trí 133738357, 133748391, 133738356, 133759413 đã được chú giải trên dbSNP của NCBI (National Center for Biological Information), trong đó đáng chú ý có 2 biến dị ở vị trí 133738357 (rs121913461) và 133748391 (rs121913457) gây kháng thuốc Imatinib. Những kết quả này có ý nghĩa quan trọng đối với các nghiên cứu tiếp theo liên quan đến vai trò của các biến dị với đột biến kháng thuốc.

Từ khóa: biến dị di truyền, CML, giải trình tự thế hệ mới.

Chỉ số phân loại: 1.6

Đặt vấn đề

CML là một loại bệnh ung thư máu thuộc nhóm tăng sinh tủy mạn ác tính, đặc trưng bởi sự tăng nhanh của các tế bào dòng bạch cầu hạt biệt hóa, hệ quả là số lượng bạch cầu tăng cao ở máu ngoại vi với đủ các giai đoạn của dòng bạch cầu hạt. Chuyển đoạn nhiễm sắc thể (NST) số 9 và 22 dẫn đến việc proto-oncogene ABL1 trên (NST số 9) gắn với gen BCR trên (NST số 22) tạo thành tổ hợp gen BCR/ABL1 (gọi là NST Philadelphia, viết tắt là Ph) gặp ở trên 90% bệnh nhân CML. Đây cũng là khâu quan trọng nhất của cơ chế bệnh sinh của CML và cũng là cơ sở để tìm ra thuốc điều trị nhắm đích cho bệnh nhân CML. Điều trị nhắm đích cho bệnh nhân CML có chuyển đoạn gen BCR/ABL1 bằng các thuốc ức chế hoạt tính tyrosin kinase là lựa chọn điều trị hàng đầu hiện nay. Mục tiêu điều trị là đạt được và duy trì tình trạng lui bệnh mức độ phân tử. Thuốc ức chế hoạt tính tyrosin kinase thế hệ thứ 1

là Imatinib (biệt dược là Glivec) là lựa chọn điều trị hàng đầu hiện nay. Tuy nhiên thực tế điều trị cho thấy không phải mọi bệnh nhân CML đều đáp ứng tốt với Imatinib. Nhiều nghiên cứu lâm sàng cho thấy, ở giai đoạn mạn tính có khoảng 33% bệnh nhân kháng thuốc (ở mức độ tế bào và di truyền) và tăng lên ở nhóm bệnh nhân trong giai đoạn tăng tốc [1, 2]. Kháng Imatinib thứ phát là do sự xuất hiện các đột biến thứ phát ở tổ hợp gen BCR/ABL1 dẫn đến giảm khả năng gắn và ức chế tyrosine kinase của protein BCR/ABL1. Vì vậy việc xác định các biến dị di truyền có vai trò quan trọng trong điều trị bệnh.

Ngày nay, các công ty thương mại đã cho ra đời các thế hệ máy đọc trình tự dựa trên nhiều công nghệ mới (Next Generation Sequencing - NGS). Các kỹ thuật đọc trình tự gen đã trở nên đơn giản và nhanh hơn nhờ sự ứng dụng huỳnh quang phân tích tự động [3, 4]. Các công nghệ đọc trình tự thế hệ mới luôn hướng tới mục

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

Application of bioinformatics for identification of genetical variants in chronic myeloid leukemia patients

Summary

Chronic myeloid leukemia (CML) is a malignant disease of hematopoietic systems, belongs to the proliferative marrow syndrome group, and is characterized by the proliferation of differentiated granulocytes of cell lines. Nowadays, Vietnam has some drugs for treatment of CML, but the appearance of genetic variants is likely to increase the drug resistance of patients. Therefore, the identification of variants plays an important role in CML treatment. In this study, the authors extracted RNA on BCR/ABL1 gene, which causes CML disease, from marrow samples of ten CML patients at National Institute of Hematology and Blood Transfusion (NIHBT). After that, RNA samples of each patient was sequenced by using Illumina MiSeq sequencing technology, then they carried out preprocessing data, aligned high quality sequences to the reference genome, and identified variants of the patient. The results showed that there were 18 variants detected from ten CML patients, including 2 variants at loci 133747603, 133747604 found on 8 patients; 3 variants at loci 133755142, 133755143, 133755146 found on 6 patients. Among them, four variants 133738357, 133748391, 133738356, 133759413 were annotated on dbSNP database of NCBI (National Center for Biotechnology Information); remarkably, 2 variants at loci 133738357 (rs121913461) and 133748391 (rs121913457) were reported to be resistant to Imatinib drug in CML patients. The results in this study will provide valuable information for further studies into the roles of the variants with drug resistance.

Keywords: CML, genetical variants, next generation sequencing.

Classification number: 1.6

tiêu làm tăng dung lượng (throughput), giảm thời gian và giá thành [5]. Chính đặc điểm quan trọng này giúp cho NGS trở thành một công cụ không thể thiếu trong các nghiên cứu hệ gen để tìm ra các biến dị di truyền, từ đó tìm được ảnh hưởng của các biến dị tác động đến bệnh lý như thế nào. Tuy nhiên, do công nghệ NGS có thông lượng lớn có thể lên tới vài tỷ basepair, nên việc nghiên cứu ứng dụng các công cụ tin sinh học để phân tích dữ liệu NGS là cần thiết. Vì vậy, chúng tôi đã tiến hành nghiên cứu xây dựng phương pháp xác định biến dị di truyền bằng công cụ tin sinh học đối với dữ liệu thông lượng cao của thiết bị NGS, từ đó có thể làm sáng tỏ đột biến gây kháng thuốc điều trị ở bệnh nhân CML.

Dữ liệu và phương pháp nghiên cứu

Dữ liệu nghiên cứu

Đối tượng nghiên cứu là dữ liệu giải trình tự RNA tổng số được tách chiết từ tủy xương của 10 bệnh nhân CML khác nhau do Khoa Di truyền và Sinh học phân tử (Viện Huyết học - Truyền máu Trung ương) thực hiện. Dữ liệu trình tự RNA được gọi là các “trình tự đọc” (read) sau đó sẽ được giải trình tự theo phương pháp paired-end bằng máy giải trình tự thế hệ mới Illumina MiSeq® (Illumina, San Diego, CA, USA).

Phương pháp nghiên cứu

Quy trình nghiên cứu được thực hiện theo các bước sau: đánh giá và tiền xử lý dữ liệu; ánh xạ trình tự (reads) vào hệ gen tham chiếu; xác định các biến dị sau khi ánh xạ.

Đánh giá và tiền xử lý dữ liệu: dữ liệu trình tự thô thu được từ thiết bị đọc trình tự thế hệ mới sẽ được đánh giá chất lượng với phần mềm FastQC (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) và được tiến hành tiền xử lý bằng cách cắt 15 base đầu của tất cả các trình tự nhằm loại bỏ các adapter, sau đó tiến hành tinh sạch các trình tự có độ dài ngắn (nhỏ hơn 36 bp), chất lượng thấp (QC < 30) bằng phần mềm Trimmomatic [6].

Ánh xạ trình tự vào hệ gen tham chiếu và đánh giá chất lượng ánh xạ: các trình tự sau khi tiền xử lý chất lượng sẽ được ánh xạ với hệ gen tham chiếu bằng phần mềm STAR [7], ưu điểm của STAR là độ nhạy cao

cho các SNP và INDEL với việc phát hiện khoảng nối giữa các exon và intron gọi là splice junctions. Hệ gen tham chiếu được sử dụng là của NST số 9 (chr9. fa) trên người (*Homo sapiens*) có mã số hg19 (ftp://ussdftp.illumina.com/Homo_sapiens/UCSC/hg19/). Ảnh xạ được thêm hoặc thay thế (Add or Replace Read Groups) nhằm gộp tất cả các read trong một lần chạy của máy giải trình tự vào một nhóm (ReadGroups), đặc trưng cho dữ liệu RNAseq. Việc “đánh dấu trùng lặp” (Mark Duplicated) để loại bỏ các read được ánh xạ ở cùng một vị trí trên hệ gen tham khảo gây nhiễu kết quả trong việc xác định các biến dị. Cuối cùng, tiến hành “sắp xếp” (Sorted) các read theo vị trí trên hệ gen. Toàn bộ 3 bước phân tích ánh xạ nêu trên được thực hiện bằng phần mềm Picard (<https://broadinstitute.github.io/picard/command-line-overview.html>). Bước tiếp theo, các trình tự được ánh xạ sẽ được chia tách thành các đoạn N-cigar (các đoạn intron giữa các exon) bằng phần mềm SplitNCigarReads trong bộ công cụ GATK (Genome Analysis Toolkit) [8]. Sau đó, tiến hành đánh giá chất lượng ánh xạ trình tự bằng phần mềm Qualimap [9].

Xác định và chọn lọc các biến dị: các biến dị được xác định và chọn lọc bằng bộ công cụ GATK. Việc xác định các biến dị bằng gói script Haplotypecaller có sẵn trong GATK sẽ giúp tối ưu việc tìm ra các đoạn N-cigar và các base ánh xạ không chính xác (soft-clipped bases), nhằm nâng cao chất lượng biến dị được tìm thấy. Cuối cùng, nhóm nghiên cứu tiến hành sàng lọc các biến dị bằng gói phần mềm SelectVariant tích hợp sẵn trong bộ công cụ GATK, với tham số lần lượt là: độ sâu các trình tự ánh xạ vào hệ gen tham chiếu ($DP > 15$) và điểm chất lượng ánh xạ ($QUAL > 30$).

Kết quả và thảo luận

Đánh giá và tiền xử lý dữ liệu

Sử dụng thiết bị đọc trình tự thế hệ mới Illumina MiSeq, đồng thời tiến hành tiền xử lý dữ liệu thô bằng phần mềm Trimmomatic nhằm loại bỏ những đoạn trình tự ngắn, chất lượng thấp thu được từ việc giải trình tự 10 mẫu bệnh nhân CML khác nhau. Số lượng đoạn trình tự tinh sạch được giữ lại chiếm khoảng 80% tổng số dữ liệu thô cho thấy việc đọc trình tự cho chất lượng tốt, phù hợp cho việc ánh xạ trình tự vào hệ gen tham chiếu (bảng 1).

Bảng 1: thống kê số lượng trình tự đọc trước và sau tinh sạch của 10 mẫu

bệnh nhân

Mã bệnh nhân	Kiểu dữ liệu	Số lượng đoạn trình tự (% tinh sạch)	Độ dài (bp)	%GC
51452	Dữ liệu thô	2.332.542	32-151	55
	Dữ liệu tinh sạch	1.816.884 (77,9%)	36-136	55
51559	Dữ liệu thô	1.301.958	32-151	55
	Dữ liệu tinh sạch	1.093.532 (84%)	36-136	54
52492	Dữ liệu thô	1.044.932	32-151	55
	Dữ liệu tinh sạch	802.384 (76,8%)	36-136	55
52548	Dữ liệu thô	1.924.288	32-151	55
	Dữ liệu tinh sạch	1.565.196 (81,3%)	36-136	55
52550	Dữ liệu thô	62.944	32-151	55
	Dữ liệu tinh sạch	49.120 (78%)	36-136	54
52558	Dữ liệu thô	83.128	32-151	55
	Dữ liệu tinh sạch	74.028 (89%)	36-136	54
52592	Dữ liệu thô	1.083.088	32-151	55
	Dữ liệu tinh sạch	913.416 (84,3%)	36-136	54
52598	Dữ liệu thô	81.674	32-151	55
	Dữ liệu tinh sạch	72.872 (89,2%)	36-136	54
52660	Dữ liệu thô	1.950.530	32-151	55
	Dữ liệu tinh sạch	1.503.496 (77%)	36-136	55
52700	Dữ liệu thô	82.086	32-151	54
	Dữ liệu tinh sạch	67.666 (82,4%)	36-136	54

Ánh xạ các trình tự vào hệ gen tham chiếu

Sau khi ánh xạ các trình tự vào hệ gen tham chiếu, tiến hành kiểm tra chất lượng ánh xạ thông qua phần mềm Qualimap, tỷ lệ ánh xạ vào hệ gen tham chiếu chiếm trung bình 99% cho thấy chất lượng ánh xạ rất cao (bảng 2).

Bảng 2: đánh giá chất lượng ánh xạ vào hệ gen tham chiếu trên 10 mẫu bệnh nhân

Mã bệnh nhân (SID)	Số lượng trình tự	Số lượng trình tự được ánh xạ	Tỷ lệ ánh xạ (%)
51452	549.759	545.909	99,30
51559	399.923	397.939	99,50
52492	232.319	229.952	98,98
52548	522.917	519.994	99,44
52550	22.084	21.926	99,28
52558	33.788	33.772	99,95
52592	287.302	285.407	99,34
52598	33.708	33.692	99,95
52660	519.468	514.590	99,06
52700	30.278	30.167	99,63

Xác định và chọn lọc biến dị

Việc xác định và chọn lọc các biến dị được tiến hành bằng cách sử dụng 2 phần mềm Haplotypecaller và SelectVariant trong bộ công cụ GATK để chọn ra các biến dị với chất lượng ánh xạ lớn hơn 30 (QUAL)

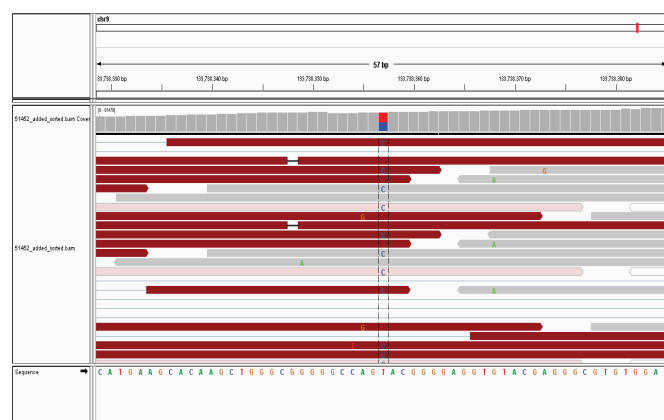
và độ sâu ánh xạ lớn hơn 15 (DP). Sau đó, đối chiếu trên cơ sở dữ liệu dbSNP của NCBI thu được kết quả như bảng 3.

Bảng 3: danh sách các biến dị trên 10 mẫu bệnh nhân

Mã bệnh nhân (SID)	Vị trí biến dị (POS)	Nucleotide trên hệ tham chiếu (REF)	Nucleotide ánh xạ (ALT)	Chất lượng ánh xạ (QUAL)	Độ sâu ánh xạ (DP)	dbSNP	
						Mã SNP	Mã bệnh nhân (SID)
51452	133738357	T	C	63738.77	4563	rs121913461	Kháng Imatinib
	133747603	G	GTCTGC,GC	11185.73	1133		
	133747604	AGTAAG	A	12292.73	1123		
	133755142	G	T	601.77	83		
	133755143	A	G	601.77	83		
	133755146	T	A	601.77	83		
51559	133895596	G	T	287.78	18		
	133747603	G	GC	7843.73	1027		
	133747604	AGTAAG	A	8000.73	1024		
	133748391	T	C	83930.77	5936	rs121913457	Kháng Imatinib
	133755142	G	T	241.84	53		
	133755143	A	G	241.84	53		
52492	133755146	T	A	241.84	53		
	133747513	C	A	599.77	117		
	133747603	G	GC	7948.73	976		
	133747604	AGTAAG	A	8358.73	971		
52548	133895596	G	T	243.78	20		
	133738357	T	C	107716.8	4380		
	133747603	G	GC	8246.73	1022		
	133747604	AGTAAG	A	8750.73	1020		
	133755142	G	T	466.77	35		
	133755143	A	G	466.77	35		
	133755146	T	A	469.77	35		
	133755354	T	C	46.77	43		
52550	133755360	T	C	46.77	43		
	133755987	T	C	99.77	19		
	133755988	G	C	99.77	19		
52558	133747603	G	GC	1658.73	223		
	133747604	AGTAAG	A	1674.06	222		
52592	133738356	G	C	48022.77	2094	rs121913458	
	133747603	G	GTCTGC,GC	10706.73	1069		
	133747604	AGTAAG	A	11477.73	1060		
	133755142	G	T	235.77	50		
	133755143	A	G	235.77	50		
	133755146	T	A	238.77	50		
	133828908	T	C	779.77	18		
	133828916	T	G	779.77	18		
52598	133828920	C	T	779.77	18		
	133747603	G	GC	1749.73	289		
52660	133747604	AGTAAG	A	1858.73	289		
	133747513	C	A	747.77	271		
	133747603	G	GC	5253.73	935		
	133747604	AGTAAG	A	5489.73	927		
	133755142	G	T	372.03	75		
	133755143	A	G	372.03	75		
	133755146	T	A	372.03	75		
	133759413	C	T	106.77	18	rs587778013	
52700	133828908	T	C	824.77	19		
	133828916	T	G	781.77	18		
	133828920	C	T	781.77	17		
52700	133755142	G	T	152.03	19		
	133755143	A	G	152.03	19		
	133755146	T	A	152.03	19		

Kết quả thu được, có 5 biến dị lặp lại với tần số cao ở các vị trí 133747603, 133747604, 133755142, 133755143, 133755146 có điểm chất lượng (QUAL) và độ sâu ánh xạ (DP) cao. Đặc biệt, khi đối chiếu trên cơ sở dữ liệu dbSNP, có 4 biến dị đã được chú giải ở các vị trí 133738357 (với mã rs là rs121913461), 133748391 (rs121913457), 133738356 (rs121913458) và 133759413 (rs587778013). Trong đó có 2 vị trí 133738357 (rs121913461) và 133748391 (rs121913457) là các đột biến gây hiệu ứng kháng thuốc Imatinib. Các biến dị còn lại có độ tin cậy chưa cao do QUAL và DP đều thấp.

Từ kết quả biến dị, phần mềm IGV được sử dụng để quan sát ánh xạ vào hệ gen tham chiếu với các biến dị tìm được (hình 1).



Hình 1: minh họa kết quả ánh xạ trình tự và xác định biến dị với dữ liệu NGS ở mẫu bệnh nhân mang mã số 51452 tại vị trí 133738357 có tham chiếu T, trình tự ánh xạ C

Kết luận

Nghiên cứu đã xây dựng thành công phương pháp xác định các biến dị và ứng dụng trên 10 mẫu bệnh nhân với kết quả thu được 18 biến dị, trong đó có 5 biến dị lặp lại với tần số cao ở các vị trí 133747603, 133747604, 133755142, 133755143, 133755146; 4 biến dị được ghi nhận trên cơ sở dữ liệu dbSNP với mã ID lần lượt là rs121913461, rs121913457, rs121913458, rs587778013. Đặc biệt, 2 biến dị có ý nghĩa lâm sàng gây kháng thuốc (vị trí 133738357 với mã rs121913461 và 133748391 với mã rs121913457). Đây là những kết quả ban đầu giúp chúng ta tìm ra phương pháp xác định các biến dị di truyền bằng công cụ tin sinh học, từ đó cung cấp những kiến thức cơ sở để tìm ra các ảnh hưởng của các biến dị đó trên bệnh nhân trong các nghiên cứu tiếp theo.

Lời cảm ơn

Nghiên cứu này được tài trợ bởi đề tài thuộc các hướng KH&CN ưu tiên cấp Viện Hàn lâm KH&CN Việt Nam, mã số VAST02.05/15-16. Các tác giả xin trân trọng cảm ơn.

Tài liệu tham khảo

[1] S. Branford, Z. Rudzki, S. Walsh, I. Parkinson, A. Grigg, J. Szer, K. Taylor, R. Herrmann, J.F. Seymour, C. Arthur, D. Joske, K. Lynch, T. Hughes (2003), "Detection of BCR-ABL mutations in patients with CML treated with imatinib is virtually always accompanied by clinical resistance, and mutations in the ATP phosphate-binding loop (P-loop) are associated with a poor prognosis", *Blood*, **102**(1), pp.276-283.

[2] S. Soverini, A. Hochhaus, F.E. Nicolini, F. Gruber, T. Lange, G. Saglio, F. Pane, M.C. Müller, T. Ernst, G. Rosti, K. Porkka, M. Baccarani, N.C.P Cross and G. Martinelli (2011), "BCR-ABL kinase domain mutation analysis in chronic myeloid leukemia patients treated with tyrosine kinase inhibitors: recommendations from an expert panel on behalf of European LeukemiaNet", *Blood*, **118**(5), pp.1208-1215.

[3] O. Olsvik, J. Wahlberg, B. Pettersson, M. Uhlén, T. Popovic, I.K. Wachsmuth and P.I. Fields (1993), "Use of automated sequencing of polymerase chain reaction-generated amplicons to identify three

types of cholera toxin subunit B in *Vibrio cholerae* O1 strains", *Journal of Clinical Microbiology*, **31**(1), pp.22-25.

[4] E. Pettersson, J. Lundeberg, and A. Ahmadian (2009), "Generations of sequencing technologies", *Genomics*, **93**(2), pp.105-111.

[5] E.E. Schadt, S. Turner and A. Kasarskis (2010), "A window into third-generation sequencing", *Human Molecular Genetics*, **19**(R2), pp.R227-R240.

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

[7] A. Dobin, C.A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson and T.R. Gingeras (2013), "STAR: ultrafast universal RNA-seq aligner", *Bioinformatics*, **29**(1), pp.15-21.

[8] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M.A. DePristo (2010), "The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data", *Genome Research*, **20**(9), pp.1297-1303.

[9] F. García-Alcalde, K. Okonechnikov, J. Carbonell, L.M. Cruz, S. Götz, S. Tarazona, J. Dopazo, T.F. Meyer, and A. Conesa (2012), "Qualimap: evaluating next-generation sequencing alignment data", *Bioinformatics*, **28**(20), pp.2678-2679.