

TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI

ĐỒ ÁN TỐT NGHIỆP

**Xây dựng hệ thống phát hiện và chú giải đột biến gen
của bệnh nhân ung thư**

ĐỒ XUÂN TÙNG

tung.dx183851@sis.hust.edu.vn

Ngành Kỹ thuật máy tính

Giảng viên hướng dẫn: TS. Nguyễn Hồng Quang

Chữ kí GVHD

Khoa:

Kỹ thuật máy tính

Trường:

Công nghệ thông tin và Truyền thông

HÀ NỘI, 02/2023

LỜI CẢM ƠN

Lời đầu tiên, em xin chân thành cảm ơn và biết ơn sâu sắc đến nhà trường, các thầy cô giảng viên của Đại học Bách Khoa Hà Nội đã tận tình giảng dạy, truyền đạt những kiến thức quý báu trong suốt 5 năm học tập và nghiên cứu tại trường, giúp em có nền tảng kiến thức vững chắc trước khi bước ra khỏi cánh cổng trường Đại học Bách Khoa Hà Nội và bước tiếp chặng đường tương lai phía trước.

Đặc biệt, em xin chân thành cảm ơn thầy giáo, TS. Nguyễn Hồng Quang, giảng viên Trường Công nghệ Thông tin và Truyền thông, Đại học Bách Khoa Hà Nội đã trực tiếp định hướng, hướng dẫn, hỗ trợ và động viên em trong suốt quá trình thực hiện đồ án tốt nghiệp, giúp em hoàn thành đồ án một cách tốt nhất.

Em cũng xin gửi lời cảm ơn tới anh Vũ Trung Dũng và bạn Phan Thị Lệ Hằng đã hỗ trợ em trong quá trình nghiên cứu và xử lý dữ liệu trong đồ án. Cũng nhân dịp này, em xin cảm ơn tới gia đình, người thân và bạn bè đã luôn giúp đỡ, động viên em trong suốt quá trình học tập và hoàn thành đồ án tốt nghiệp.

Em xin chân thành cảm ơn!

TÓM TẮT NỘI DUNG ĐỒ ÁN

Ung thư là nguyên nhân gây tử vong xếp thứ hai trên thế giới dù là bệnh không lây nhiễm. Ở Việt Nam, phần lớn các bệnh nhân ung thư được phát hiện ở giai đoạn muộn, không có nhiều cơ hội chữa trị dẫn tới tử vong. Do đó, việc tầm soát, sớm phát hiện ung thư là chìa khóa giúp cho việc điều trị ung thư có hiệu quả cao, giảm gánh nặng đối với xã hội. Với sự phát triển của khoa học kỹ thuật, các công cụ giải trình tự bộ và phát hiện biến thể gen ra đời. Cùng với đó, nguồn dữ liệu khổng lồ về ung thư do các tổ chức uy tín trên thế giới phát triển công khai đã hỗ trợ các bác sĩ chuẩn đoán, tìm kiếm thuốc điều trị chính xác bệnh ung thư, tăng cơ hội khỏi bệnh.

Ở Việt Nam, việc phát hiện và hỗ trợ bác sĩ tìm kiếm thông tin liên quan đến bệnh ung thư ngày càng trở nên tương đối phổ biến trong các cơ sở y tế trên cả nước. Tuy nhiên, các trang thiết bị phát hiện đột biến thường có chi phí cao, độ chính xác của các phương pháp phát hiện đột biến và công cụ hỗ trợ các bác sĩ tìm kiếm thông tin thuốc điều trị liên quan là những rào cản có thể nhắc tới.

Vì vậy, mục tiêu của đồ án là nghiên cứu, ứng dụng và xây dựng hệ thống phát hiện và chú thích biến thể gen liên quan tới 50 gen gây bệnh ung thư từ dữ liệu giải trình tự gen cung cấp bởi Đại học Y Hà Nội. Dựa trên nền tảng lý thuyết, các kỹ thuật, công cụ phát hiện đột biến với độ chính xác cao và tập dữ liệu ung thư uy tín, đã được kiểm chứng và công bố trên các trang web của COSMIC và NCBI. Hệ thống là công cụ hỗ trợ các bác sĩ chuẩn đoán và tìm thuốc chính xác trong điều trị bệnh ung thư.

MỤC LỤC

CHƯƠNG 1. GIỚI THIỆU ĐỀ TÀI.....	1
1.1 Đặt vấn đề.....	1
1.2 Mục tiêu và phạm vi đề tài.....	1
1.3 Định hướng đề tài	2
1.4 Bố cục đồ án	2
CHƯƠNG 2. NỀN TẢNG LÝ THUYẾT	4
2.1 Ngữ cảnh của bài toán.....	4
2.2 Các kiến thức sinh học cơ bản.....	4
2.2.1 Gen.....	4
2.2.2 Đột biến gen	5
2.2.3 Giải trình tự gen.....	6
2.2.4 Quy trình phân tích dữ liệu giải trình tự NGS.....	8
2.3 Các công cụ phát hiện đột biến gen	12
2.3.1 DeepVariant	12
2.3.2 MuSE	13
2.3.3 Samtools và bcftools	14
2.4 Cơ sở dữ liệu COSMIC	15
CHƯƠNG 3. PHƯƠNG PHÁP ĐỀ XUẤT.....	19
3.1 Tổng quan giải pháp.....	19
3.2 Dữ liệu đầu vào	21
3.3 Phát hiện đột biến	22
3.3.1 Kiểm soát chất lượng và tiền xử lý dữ liệu	22
3.3.2 Căn chỉnh trình tự với bộ gen tham chiếu.....	24
3.3.3 Gọi tên đột biến bằng Bcftools	25

3.4 Chú giải đột biến	27
3.4.1 Định dạng file VCF	27
3.4.2 Lọc và biểu diễn đột biến theo định dạng chuẩn	28
3.4.3 Tra cứu đột biến và tìm kiếm thuốc điều trị.....	32
CHƯƠNG 4. ĐÁNH GIÁ THỰC NGHIỆM.....	34
4.1 Kết quả với mẫu bệnh nhân BN-081	34
4.1.1 Kiểm soát chất lượng và tiền xử lý dữ liệu	34
4.1.2 Kết quả phát hiện và chú thích đột biến gen	36
4.1.3 So sánh với kết quả mẫu	37
4.2 Kết quả với mẫu bệnh nhân BN-082	40
4.2.1 Kiểm soát chất lượng và tiền xử lý dữ liệu	40
4.2.2 Kết quả phát hiện và chú thích đột biến gen	42
4.2.3 So sánh với kết quả mẫu	43
4.3 Kết quả với mẫu bệnh nhân BN-084	44
4.3.1 Kiểm soát chất lượng và tiền xử lý dữ liệu	44
4.3.2 Kết quả phát hiện và chú thích đột biến gen	46
4.3.3 So sánh với kết quả mẫu	47
4.4 Đánh giá	48
4.5 Đề xuất cải thiện tìm kiếm thuốc điều trị.....	48
CHƯƠNG 5. KẾT LUẬN	51
5.1 Kết luận	51
5.2 Hướng phát triển trong tương lai	51
TÀI LIỆU THAM KHẢO.....	53
PHỤ LỤC.....	55
.1 Các kiến thức sinh học liên quan.....	55
.1.1 Các loại đột biến gen.....	55

.1.2 Bộ gen tham chiếu	55
.1.3 Độ sâu của giải trình tự	55

DANH MỤC HÌNH VẼ

Hình 2.1	Cấu tạo DNA	5
Hình 2.2	Đột biến gen gây bệnh hồng cầu lưỡi liềm	6
Hình 2.3	Phương pháp giải trình tự Sanger ra đời năm 1977	7
Hình 2.4	Chi phí giải trình tự giảm sau khi NGS ra đời	8
Hình 2.5	Quy trình phân tích dữ liệu NGS.[3]	9
Hình 2.6	Biểu đồ "Chất lượng dữ liệu đầu vào" sau khi chạy FASTQC .	10
Hình 2.7	Công cụ BWA-MEM trong căn chỉnh trình tự DNA	11
Hình 2.8	Chi tiết các bước trong DeepVariant	13
Hình 2.9	Luồng gọi đột biến của MuSE	14
Hình 2.10	Luồng hoạt động của Samtools và Bcftools	15
Hình 2.11	Cơ sở xây dựng dữ liệu COSMIC	16
Hình 2.12	Tổng quan dữ liệu COSMIC	17
Hình 3.1	Tổng thể quá trình phát hiện đột biến gen	20
Hình 3.2	Tổng thể quá trình chú giải đột biến gen	21
Hình 3.3	Các trường thông tin trong dữ liệu giải trình tự gen	22
Hình 3.4	Báo cáo kết quả đọc chất lượng giải trình tự	23
Hình 3.5	Phân phối chất lượng trình tự ở file dữ liệu chất lượng kém . .	23
Hình 3.6	Ví dụ về căn chỉnh một đoạn trình tự với bộ tham chiếu	24
Hình 3.7	Luồng hoạt động gọi tên đột biến bằng Bcftools	26
Hình 3.8	Phần tiêu đề trong file VCF	27
Hình 3.9	Phần nội dung trong file VCF	28
Hình 3.10	Các công cụ hỗ trợ của Ensembl phát triển	30
Hình 3.11	Kết quả đầu ra chứa thông tin biến thể theo định dạng HGVS .	31
Hình 3.12	Ví dụ tra cứu thông tin gen bằng thông tin Transcript	32
Hình 3.13	Trích lọc dữ liệu đột biến với 50 gen mục tiêu	33
Hình 4.1	Báo cáo tổng quan về dữ liệu đầu vào của BN-081	34
Hình 4.2	Biểu đồ phân phối chất lượng đoạn đọc trên các đoạn trình tự của BN-081	35
Hình 4.3	Báo cáo tổng quan về dữ liệu sau khi xử lý BN-081	35
Hình 4.4	Biểu đồ phân phối chất lượng đoạn đọc sau khi xử lý BN-081 .	36
Hình 4.5	Kết quả file VCF chứa danh sách biến thể của BN-081	36
Hình 4.6	Kết quả phát hiện và chú thích đột biến BN-081	37
Hình 4.7	Báo cáo tổng quan về dữ liệu đầu vào của BN-082	40

Hình 4.8	Biểu đồ phân phối chất lượng đoạn đọc trên các đoạn trình tự của BN-082	41
Hình 4.9	Báo cáo tổng quan về dữ liệu sau khi xử lý BN-082	41
Hình 4.10	Biểu đồ phân phối chất lượng đoạn đọc sau khi xử lý BN-082 .	42
Hình 4.11	Kết quả file VCF chứa danh sách biến thể của BN-082	42
Hình 4.12	Kết quả phát hiện và chú thích đột biến BN-082	43
Hình 4.13	Báo cáo tổng quan về dữ liệu đầu vào của BN-084	44
Hình 4.14	Biểu đồ phân phối chất lượng đoạn đọc trên các đoạn trình tự của BN-084	45
Hình 4.15	Báo cáo tổng quan về dữ liệu sau khi xử lý BN-084	45
Hình 4.16	Biểu đồ phân phối chất lượng đoạn đọc sau khi xử lý BN-084 .	46
Hình 4.17	Kết quả file VCF chứa danh sách biến thể của BN-084	46
Hình 4.18	Kết quả phát hiện và chú thích đột biến BN-084	47
Hình 4.19	Trang web của cơ sở dữ liệu ung thư OncoKB	49
Hình 4.20	Tra cứu thông tin biến thể với OncoKB	50
Hình .1	Ví dụ độ sâu giải trình tự	55

DANH MỤC BẢNG BIỂU

Bảng 2.1	Các công cụ để xác định biến thể	12
Bảng 2.2	Một số trường quan trọng trong dữ liệu đột biến và thuốc . . .	18
Bảng 3.1	50 gen mục tiêu có liên quan đến bệnh ung thư	21
Bảng 4.1	Bảng danh sách kết quả phát hiện đột biến trên 50 gen gây bệnh ung thư của BN-081	37
Bảng 4.2	Bảng danh sách kết quả trên các gen phát hiện của BN-082 . .	43
Bảng 4.3	Bảng danh sách kết quả trên các gen phát hiện của BN-084 . .	48