illumina®

DRAGEN for Illumina DNA Prep with Enrichment Dx

Tài liệu hướng dẫn về sản phẩm cho NovaSeq 6000Dx

Tài liệu số 200014776 v02
Tháng 9 năm 2022
DÙNG CHO CHẨN ĐOÁN IN VITRO

Lịch sử sửa đổi

Tài liệu	Ngày	Mô tả thay đổi
200014776 v02	Tháng 9 năm 2022	Định dạng của tệp bản kê khai đã được chỉnh sửa từ dạng văn bản (*.txt) sang dạng BED (*.bed) trong hướng dẫn tạo lần chạy. Chỉnh sửa từ tệp VCF đồng thuận thành tệp VCF trong phần đầu ra phân tích.
200014776 v01	Tháng 8 năm 2022	Đã thêm: Phần Cài đặt. Phần Lọc nhiễu hệ thống. Đã cập nhật hướng dẫn tạo lần chạy để bao gồm thông tin chi tiết hơn. Đã sửa lỗi đánh máy và lỗi ngữ pháp. Đã chỉ rõ rằng các hướng dẫn được áp dụng cho ứng dụng khi sử dụng với Thiết bị NovaSeq 6000Dx. Đã cập nhật thông tin về nội dung tệp đầu ra VCF.
200014776 v00	Tháng 3 năm 2022	Phát hành lần đầu.

Mục lục

Tổng quan

Ứng dụng DRAGEN™ for Illumina® DNA Prep with Enrichment Dx thực hiện quá trình tách đoạn, tạo tệp FASTQ, ánh xạ đoạn đọc và căn chỉnh với hệ gen tham chiếu, cũng như phát hiện biến thể tùy thuộc vào quy trình công việc phân tích được chọn.

Phương pháp phân tích

DRAGEN for Illumina DNA Prep with Enrichment Dx thực hiện tách đoạn, tạo FASTQ, ánh xạ đoạn đọc và căn chỉnh với hệ gen tham chiếu tùy thuộc vào quy trình công việc được chọn:

- Tạo FASTQ
- Tạo FASTQ và VCF cho dòng mầm
- Tạo FASTQ và VCF cho soma

Tạo FASTQ

Các trình tự đã lắp ráp được ghi vào các tệp FASTQ theo từng mẫu. Tệp FASTQ là tệp văn bản chứa dữ liệu giải trình tự và điểm chất lượng cho một mẫu duy nhất. Với mỗi mẫu, các tệp FASTQ riêng biệt được tạo cho từng làn tế bào dòng chảy và cho từng đoạn đọc giải trình tự. Tên mẫu được chỉ định trong quá trình thiết lập lần chạy sẽ được đưa vào tên tệp FASTQ. Tệp FASTQ là đầu vào chính để căn chỉnh. Bước đầu tiên của quá trình tạo FASTQ là tách đoạn. Quá trình tách đoạn sẽ gán các cụm đi qua bộ lọc cho một mẫu bằng cách so sánh mỗi trình tự đọc chỉ thị với các trình tự chỉ thị đã được chỉ định cho lần chạy. Bước này không xem xét giá trị chất lượng nào. Đoạn đọc chỉ thị được xác định bằng các bước sau:

- Các mẫu được đánh số bắt đầu từ 1 dựa trên thứ tự mẫu được liệt kê cho lần chạy.
- Mẫu số 0 được dành riêng cho các cụm không được chỉ định cho một mẫu.
- Các cụm được chỉ định cho một mẫu khi trình tự chỉ thị khớp chính xác hoặc khi có tối đa một lần không khớp trên mỗi đoạn đọc chỉ thị.

Phần mềm bao gồm tính năng nén ORA để nén các tệp FASTQ. Khi sử dụng định dạng ORA (*.ora), tổng kiểm md5 của nội dung FASTQ được giữ nguyên sau một chu kỳ nén và giải nén để đảm bảo nén không mất dữ liệu.

Căn chỉnh & ánh xạ DNA

Giai đoạn đầu tiên của ánh xạ là tạo các hạt giống từ đoạn đọc, và sau đó tìm các vị trí khớp chính xác trong hệ gen tham chiếu. Các kết quả này được tinh chỉnh bằng cách chạy thuật toán Smith-Waterman đầy đủ tại những vị trí có mật độ hạt giống khớp cao nhất. Thuật toán đã được kiểm chứng này hoạt động bằng cách so sánh từng vị trí của đoạn đọc với tất cả các vị trí ứng viên trên hệ gen tham chiếu. Các phép so sánh này tương ứng với một ma trận các khả năng căn hàng giữa đoạn đọc và hệ gen tham chiếu. Đối với mỗi vị trí căn chỉnh ứng viên này, thuật toán Smith-Waterman sẽ tạo ra các mức điểm được

dùng để đánh giá xem đường căn chỉnh tốt nhất đi qua ô ma trận đó đạt được bằng cách: khớp hoặc không khớp nucleotide (di chuyển theo đường chéo), vùng xóa (di chuyển theo chiều ngang), hay vùng chèn (di chuyển theo chiều dọc). Vị trí khớp giữa đoạn đọc và tham chiếu sẽ được cộng điểm thưởng, và vị trí không khớp hoặc có biến đổi indel sẽ bị tính điểm phạt. Đường đi có tổng số điểm cao nhất xuyên qua ma trận chính là phương án căn chỉnh được lựa chọn.

Các giá trị cụ thể được chọn để tính điểm trong thuật toán này giúp cân bằng giữa các cách giải thích khác nhau cho một kết quả căn chỉnh—ví dụ: ưu tiên xem đó là một đột biến indel hay là một hoặc nhiều SNP, hoặc ưu tiên một kết quả căn chỉnh không bị cắt đoạn. Các giá trị tính điểm mặc định của DRAGEN phù hợp để căn chỉnh các đoạn đọc có độ dài trung bình với toàn bộ hệ gen người cho các ứng dụng phát hiện biến thể. Bất kỳ bộ tham số tính điểm Smith-Waterman nào cũng chỉ là một mô hình ước lệ về đột biến gen và lỗi giải trình tự. Các giá trị tính điểm căn chỉnh được điều chỉnh khác nhau có thể phù hợp hơn cho một số ứng dụng.

Phát hiện biến thể dòng mầm DRAGEN

DRAGEN Germline Small Variant Caller nhận đầu vào là các đoạn đọc DNA đã được ánh xạ và căn chỉnh và phát hiện các SNP cũng như indel thông qua sự kết hợp giữa phát hiện theo cột và lắp ráp kiểu đơn bội *de novo* tại chỗ.

Đầu tiên, các vùng tham chiếu có thể phát hiện sẽ được nhận diện dựa trên phạm vi căn chỉnh đầy đủ. Trong các vùng tham chiếu này, một lượt quét nhanh các đoạn đọc đã sắp xếp sẽ giúp xác định các vùng hoạt động, vốn là các vùng tập trung xung quanh các cột dữ liệu chồng lấp có bằng chứng về biến thể. Các vùng hoạt động này được đệm thêm một khoảng trình tự đủ rộng để bao phủ các nội dung không khớp với tham chiếu quan trọng ở lân cận. Nếu có bằng chứng về indel, các vùng hoạt động sẽ được đệm thêm diện tích.

Các đoạn đọc đã căn chỉnh sẽ được cắt gọn trong phạm vi mỗi vùng hoạt động và lắp ráp thành một đồ thị De Bruijn. Các cạnh của các đoạn đọc được cắt gọn này được gán trọng số dựa trên số lần quan sát được, với trình tự tham chiếu đóng vai trò là khung xương. Sau khi làm sạch và đơn giản hóa đồ thị, tất cả các đường đi từ điểm nguồn đến điểm đích sẽ được trích xuất dưới dạng các kiểu đơn bội ứng viên. Mỗi kiểu đơn bội sau đó được căn chỉnh theo thuật toán Smith-Waterman với hệ gen tham chiếu để xác định các biến thể mà nó đại diện. Tập hợp các sự kiện này có thể được bổ sung bằng phương pháp phát hiện dựa trên vị trí. Đối với mỗi cặp đoạn đọc-kiểu đơn bội, xác suất P(r|H) của quan sát đoạn đọc đó khi giả định kiểu đơn bội là mẫu gốc thực sự sẽ được ước tính bằng mô hình Markov ẩn cặp (HMM).

Bằng cách quét theo vị trí tham chiếu trên vùng hoạt động, các kiểu gen ứng viên được hình thành từ các tổ hợp lưỡng bội của các sự kiện biến thể (SNP hoặc indel). Đối với mỗi sự kiện (bao gồm cả tham chiếu), xác suất có điều kiện P(r|e) của quan sát mỗi đoạn đọc chồng lấn được ước tính là giá trị P(r|H) tối đa cho các kiểu đơn bội hỗ trợ sự kiện đó. Các giá trị này được kết hợp thành xác suất có điều kiện P(r|e1e2) cho một kiểu gen (cặp sự kiện) và nhân lên để tạo ra xác suất có điều kiện P(R|e1e2) của quan sát toàn bộ các đoạn đọc chồng lấp. Sử dụng Công thức Bayes, xác suất hậu nghiệm P(e1e2|R) của mỗi kiểu gen lưỡng bội sẽ được tính toán, và kiểu gen có xác suất cao nhất sẽ được phát hiện.

Ở chế độ gVCF (được sử dụng để phát hiện biến thể đa mẫu có khả năng mở rộng), DRAGEN Germline Small Variant Caller có thể chạy trên từng mẫu để tạo ra một tệp phát hiện biến thể hệ gen trung gian (gVCF). Sau đó, tệp gVCF này có thể được sử dụng để xác định kiểu gen chung cho nhiều mẫu một cách hiệu quả, cho phép xử lý tăng dần nhanh chóng các mẫu và mở rộng quy mô cho các cỡ đoàn hệ lớn.

Vì DRAGEN Germline Small Variant Caller sở hữu các thuật toán có khả năng phân biệt hiệu quả giữa các lỗi tương quan và các biến thể thực sự, nên các quy tắc lọc rất đơn giản.

Phát hiện biến thể soma DRAGEN

DRAGEN Somatic Small Variant Caller nhận đầu vào là các đoạn đọc DNA đã được ánh xạ và căn chỉnh và phát hiện các SNV và indel thông qua việc lắp ráp kiểu đơn bội *de novo* tại chỗ trong một vùng hoạt động.

Đầu tiên, các vùng tham chiếu có thể phát hiện sẽ được nhận diện dựa trên phạm vi căn chỉnh đầy đủ. Trong các vùng tham chiếu này, một lượt quét các đoạn đọc đã sắp xếp sẽ giúp xác định các vùng hoạt động, vốn là các vùng tập trung xung quanh các cột dữ liệu chồng lấp có bằng chứng về biến thể trong các đoạn đọc từ khối u. Các vùng hoạt động này được đệm thêm một khoảng trình tự đủ rộng để bao phủ các nội dung không khớp với tham chiếu quan trọng ở lân cận. Nếu có bằng chứng về indel, các vùng hoạt động sẽ được đệm thêm diện tích.

Các đoạn đọc đã căn chỉnh sẽ được cắt gọn trong phạm vi mỗi vùng hoạt động và lắp ráp thành một đồ thị De Bruijn. Các cạnh của các đoạn đọc được cắt gọn này được gán trọng số dựa trên số lần quan sát được, với trình tự tham chiếu đóng vai trò là khung xương. Sau khi làm sạch và đơn giản hóa đồ thị, tất cả các đường đi từ điểm nguồn đến điểm đích sẽ được trích xuất dưới dạng các kiểu đơn bội ứng viên. Mỗi kiểu đơn bội sau đó được căn chỉnh theo thuật toán Smith-Waterman với hệ gen tham chiếu để xác định các biến thể mà nó đại diện. Đối với mỗi cặp đoạn đọc-kiểu đơn bội, xác suất P(r|H) của quan sát đoạn đọc được ước tính bằng mô hình Markov ẩn cặp (HMM) khi giả định kiểu đơn bội là mẫu gốc thực sự.

Để xác định điểm số TLOD, DRAGEN Somatic Small Variant Caller trước tiên quét theo vị trí tham chiếu cho từng sự kiện soma ứng viên cũng như sự kiện tham chiếu trên khắp vùng hoạt động. Xác suất có điều kiện P(r|e) của quan sát mỗi đoạn đọc chồng lấn được ước tính là giá trị P(r|H) tối đa cho các kiểu đơn bội hỗ trợ sự kiện đó. Các giá trị này được kết hợp thành xác suất có điều kiện P(r|E) cho một giả thuyết sự kiện E, bao gồm một hỗn hợp giữa alen tham chiếu và alen soma ứng viên trên một dải tần suất alen có thể xảy ra và nhân lên để tạo ra xác suất có điều kiện P(R|E) của quan sát toàn bộ các đoạn đọc chồng lấp. Từ đó, điểm số TLOD được tính toán làm bằng chứng cho thấy sự hiện diện của một alen THAY THẾ tại một lô-cut nhất định trong mẫu khối u.

Tạo lần chạy

Thực hiện các bước sau để thiết lập một lần chạy trong Illumina Run Manager trên NovaSeq 6000Dx hoặc bằng cách sử dụng trình duyệt trên máy tính kết nối mạng. Dữ liệu mẫu có thể được nhập thủ công hoặc bằng cách nhập một bảng thông tin mẫu.

Cài đặt ứng dụng & lần chạy

1. Từ màn hình Runs (Lần chạy), chọn **Create Run** (Tạo lần chạy).
2. Chọn ứng dụng DRAGEN for Illumina DNA Prep with Enrichment Dx, sau đó chọn **Next** (Tiếp theo).
3. Trên màn hình Run Settings (Cài đặt lần chạy), nhập tên lần chạy. Tên lần chạy giúp xác định lần chạy từ bước giải trình tự đến bước phân tích.
4. **[Không bắt buộc]** Nhập mô tả cho lần chạy để giúp xác định lần chạy rõ hơn.
5. Hãy đảm bảo rằng Bộ kit chuẩn bị thư viện được chọn là một bộ kit chuẩn bị thư viện Illumina DNA Prep with Enrichment Dx.
6. Chọn bộ kit adapter chỉ thị mong muốn.
7. Nhập độ dài đoạn đọc.
   Đoạn đọc 1 và Đoạn đọc 2 có giá trị mặc định là 151 chu kỳ.
   Chỉ thị 1 và Chỉ thị 2 có giá trị cố định là 10 chu kỳ.
8. **[Không bắt buộc]** Nhập ID ống thư viện.
9. Chọn **Next** (Tiếp theo).

Dữ liệu mẫu

Sử dụng bảng trên màn hình Sample Data (Dữ liệu mẫu) để nhập thông tin mẫu thủ công. Hoặc, chọn **Import Samples** (Nhập mẫu) để tải lên thông tin mẫu. Để biết thêm thông tin về việc nhập thông tin mẫu, tham khảo phần *Nhập mẫu* trên trang 5.

Nhập mẫu theo cách thủ công

1. Nhập ID mẫu duy nhất vào trường Sample ID (ID mẫu).
2. Sử dụng **Plate – Well Position** (Khay – Vị trí giếng) để chọn vị trí giếng.
   Các trường i7 Index (Chỉ thị i7), Index 1 (Chỉ thị 1), i5 Index (Chỉ thị i5) và Index 2 (Chỉ thị 2) sẽ tự động được điền.
3. **[Không bắt buộc]** Nhập tên thư viện.
4. Thêm các hàng và lặp lại bước 1–3 khi cần cho đến khi tất cả mẫu đã được thêm vào bảng.
5. Chọn **Next** (Tiếp theo).

Nhập mẫu

Một tệp mẫu (*.csv) có sẵn để tải xuống trên màn hình Sample Data (Dữ liệu mẫu) khi lập kế hoạch một lần chạy trong Illumina Run Manager bằng trình duyệt trên máy tính kết nối mạng.

1. Chọn **Download Template** (Tải mẫu xuống) để tải xuống một tệp CSV trống.
2. Từ tệp CSV, nhập thông tin mẫu và lưu tệp.
   Tệp bảng thông tin mẫu định dạng CSV bao gồm các cột dữ liệu sau: Sample ID (ID mẫu), Plate - Well Position (Khay - Vị trí giếng), **Optional** Library Name (Tên thư viện không bắt buộc).
3. Chọn **Import Samples** (Nhập mẫu) rồi duyệt đến vị trí của Tệp CSV.
4. Chọn **Next** (Tiếp theo).

Cài đặt phân tích

1. Chọn quy trình công việc phân tích mong muốn:
   - Tạo FASTQ
   - Quy trình công việc tạo FASTQ và VCF cho dòng mầm
   - Quy trình công việc tạo FASTQ và VCF cho soma
2. **[Không bắt buộc]** Nếu muốn, hãy chọn ô **Generate ORA compressed FASTQs** (Tạo FASTQ nén bằng ORA) để bật chế độ nén FASTQ bằng ORA.
3. **[Quy trình công việc tạo VCF]** Sử dụng menu thả xuống **Manifest File Selection** (Chọn tệp bản kê khai) để chọn tệp bản kê khai.
   Tệp bản kê khai là dữ liệu đầu vào bắt buộc cho DRAGEN for Illumina DNA Prep with Enrichment Dx. Tệp bản kê khai là tệp BED (*.bed) được phân tách bằng tab, dùng để chỉ định tên và vị trí của các vùng tham chiếu được nhắm mục tiêu.
4. **[Quy trình công việc tạo FASTQ và VCF cho somat]** Sử dụng menu thả xuống **Noise File Selection** (Chọn tệp nhiễu) để chọn một tệp nhiễu.
   Có thể chỉ định một tệp BED với mức độ nhiễu theo vị trí để lọc bỏ nhiễu hệ thống. Để biết thêm thông tin, tham khảo phần *Lọc nhiễu* trên trang 6.
5. Chọn **Next** (Tiếp theo).

Xem lại lần chạy

1. Trên màn hình Review (Xem lại), hãy kiểm tra lại thông tin đã nhập ở các màn hình Run Settings (Cài đặt lần chạy), Sample Data (Dữ liệu mẫu) và Analysis Settings (Cài đặt phân tích).
2. Chọn **Save** (Lưu).
   Lần chạy được lưu trên tab Planned (Được lên kế hoạch) trên màn hình Runs (Lần chạy).

Cài đặt

Chọn ứng dụng trên màn hình Applications (Ứng dụng) để xem các thiết lập hiện tại và thay đổi cài đặt.

Cấu hình

Màn hình cấu hình hiển thị các cài đặt ứng dụng sau:

- **Library Prep Kits** (Bộ kit chuẩn bị thư viện)—Hiển thị bộ kit chuẩn bị thư viện mặc định cho ứng dụng. Cài đặt này không thể thay đổi.
- **Index Adapter Kits** (Bộ kit adapter chỉ thị)—Hiển thị bộ kit adapter chỉ thị mặc định cho ứng dụng. Cài đặt này không thể thay đổi.
- **Read lengths** (Độ dài đoạn đọc)—Độ dài đoạn đọc được đặt mặc định là 151 cho ứng dụng, nhưng có thể thay đổi trong quá trình tạo lần chạy.
- **Manifest and Noise Files** (Tệp bản kê khai và nhiễu)—Tải lên và thay đổi cài đặt cho các tệp bản kê khai và nhiễu.
  - Chọn **Upload File** (Tải lên tệp) để tải lên tệp sử dụng trong phân tích.
  - Chọn nút radio **Default** (Mặc định) để đặt tệp làm tệp bản kê khai hoặc nhiễu mặc định được chọn trong quá trình tạo lần chạy khi chọn ứng dụng.
  - Chọn hộp kiểm **Enabled** (Đã bật) để hiển thị tệp trong menu thả xuống trong quá trình tạo lần chạy.

Các quyền

Sử dụng các hộp kiểm trên màn hình Permissions (Quyền) để quản lý quyền truy cập của người dùng cho ứng dụng.

Lọc nhiễu

Có tính năng lọc nhiễu hệ thống khi sử dụng quy trình công việc soma. Bộ lọc có thể được sử dụng trong chế độ Khối u-Bình thường, nhưng đặc biệt hữu ích cho các lần chạy Chỉ-Khối u khi không có mẫu bình thường khớp.

Tệp BED nhiễu hệ thống cần được tạo ra từ các mẫu bình thường. Khuyến nghị xây dựng các tệp nhiễu hệ thống theo từng quy trình chuẩn bị thư viện, hệ thống giải trình tự, và bộ bảng cụ thể. Khuyến nghị sử dụng khoảng 50 mẫu bình thường để tạo tệp nhiễu.

Đầu ra phân tích

DRAGEN for Illumina DNA Prep with Enrichment Dx lưu các thông tin sau trong thư mục phân tích. Chỉ các quy trình công việc dòng mầm và soma mới tạo tệp PDF.

- Tệp bản kê khai được sử dụng
- Phiên bản phần mềm
- ID mẫu
- Tổng số đoạn đọc đã căn chỉnh
- Tỷ lệ phần trăm đoạn đọc đã căn chỉnh trên mỗi mẫu
- Số lượng SNV được phát hiện trên mỗi mẫu
- Số lượng indel được phát hiện trên mỗi mẫu
- Thống kê độ phủ

Tệp đầu ra phân tích

Ứng dụng tạo ra các tệp đầu ra sau đây. Các tệp chính xác được tạo ra phụ thuộc vào quy trình công việc phân tích được sử dụng. Các tệp đầu ra được đặt trong thư mục phân tích.

Tệp đầu ra	Mô tả
FASTQ (*.fastq.gz hoặc *.fastq.ora)	Tệp trung gian chứa các phát hiện base đã được chấm điểm chất lượng. Tệp FASTQ là đầu vào chính cho bước căn chỉnh. Nếu chọn nén ORA, tên tệp sẽ phản ánh điều này.
Tệp Alignment BAM (*.bam)	Chứa các đoạn đọc đã được căn chỉnh cho một mẫu cụ thể.
Tệp Genome VCF (*.gvcf.gz)	Chứa kiểu gen cho từng vị trí, dù được phát hiện là biến thể hay tham chiếu.
Tệp VCF (*.vcf.gz)	Chứa các biến thể được phát hiện tại mỗi vị trí.
Báo cáo số liệu lần chạy (*.csv)	Chứa các số liệu chất lượng của lần chạy, bao gồm tổng hiệu suất dữ liệu thu được và điểm Q30.

Tệp FASTQ

FASTQ (*.fastq.gz, *.fastq.ora) là một định dạng tệp văn bản chứa các phát hiện base và giá trị chất lượng cho mỗi đoạn đọc. Mỗi tệp chứa những thông tin sau:

- Mã nhận dạng mẫu
- Trình tự
- Dấu cộng (+)
- Điểm chất lượng Phred được mã hóa theo định dạng ASCII + 33

Mã nhận dạng mẫu được định dạng như sau.

```
@Instrument:RunID:FlowCellID:Lane:Tile:X:Y
ReadNum:FilterFlag:0:SampleNumber
Ví dụ:
@SIM:1:FCX:1:15:6329:1045 1:N:0:2
TCGCACTCAACGCCCTGCATATGACAAGACAGAATC
+
<>;##=><9=AAAAAAAAAA9#:<#<;<<<????#=
```

Tệp BAM

Tệp BAM (*.bam) là phiên bản nhị phân nén của tệp SAM (ánh xạ căn chỉnh giải trình tự), được dùng để biểu diễn các trình tự đã căn chỉnh có độ dài lên đến 128 Mb. Tệp BAM sử dụng định dạng tên tệp `SampleName_S#.bam`. # là số mẫu, được xác định theo thứ tự các mẫu được liệt kê trong lần chạy. Ở chế độ đa nút, ký hiệu S# luôn được đặt là S1, bất kể thứ tự của mẫu.

Tệp BAM gồm có phần đầu và phần căn chỉnh:

- Phần đầu—Chứa thông tin về toàn bộ tệp, chẳng hạn như tên mẫu, độ dài mẫu, và phương pháp căn chỉnh. Các căn chỉnh trong phần căn chỉnh sẽ được liên kết với thông tin cụ thể trong phần đầu.
- Căn chỉnh—Chứa tên đoạn đọc, trình tự đoạn đọc, chất lượng đọc, thông tin căn chỉnh và thẻ tùy chỉnh. Tên đoạn đọc bao gồm nhiễm sắc thể, tọa độ bắt đầu, chất lượng căn chỉnh, và chuỗi mô tả khớp.

Phần căn chỉnh bao gồm các thông tin sau cho mỗi đoạn đọc hoặc cặp đoạn đọc:

- AS: Chất lượng căn chỉnh của đoạn đọc kết đôi.
- RG: Nhóm đoạn đọc, cho biết số lượng đoạn đọc của một mẫu cụ thể.
- BC: Thẻ mã vạch, dùng để chỉ định ID mẫu đã được tách đoạn gắn với đoạn đọc.
- SM: Chất lượng căn chỉnh của đoạn đọc đơn.
- XC: Chuỗi mô tả khớp.

- XN: Thẻ tên amplicon, ghi lại ID amplicon liên kết với đoạn đọc. Tệp chỉ thị BAM (*.bam.bai) cung cấp chỉ thị cho tệp BAM tương ứng.

Tệp VCF

Tệp Variant call format (*.vcf) chứa thông tin về các biến thể được tìm thấy ở các vị trí cụ thể trong hệ gen tham chiếu.

Phần đầu của tệp VCF bao gồm phiên bản định dạng tệp VCF, phiên bản trình phát hiện biến thể và danh sách các chú giải được sử dụng trong phần còn lại của tệp. Phần đầu của VCF cũng bao gồm tệp hệ gen tham chiếu và tệp BAM. Dòng cuối cùng trong phần đầu chứa tiêu đề cột cho các dòng dữ liệu. Mỗi dòng dữ liệu trong tệp VCF chứa thông tin về một biến thể duy nhất.

Bảng 1 Các tiêu đề trong tệp VCF

Tiêu đề	Mô tả
CHROM	Nhiễm sắc thể của hệ gen tham chiếu. Các nhiễm sắc thể xuất hiện theo cùng thứ tự như trong tệp FASTA tham chiếu.
POS	Vị trí một base duy nhất của biến thể trên nhiễm sắc thể tham chiếu. Đối với biến thể đơn nucleotide (SNV), vị trí này là base tham chiếu có biến thể. Đối với indel, vị trí này là base tham chiếu ngay trước biến thể.
ID	Số rs (SNP tham chiếu) cho SNP lấy từ `dbSNP.txt`, nếu có. Nếu có nhiều số rs tại cùng vị trí, danh sách sẽ được phân tách bằng dấu chấm phẩy. Nếu không có mục nhập dbSNP tại vị trí này, sẽ sử dụng ký hiệu giá trị thiếu ('.').
REF	Hệ gen tham chiếu. Ví dụ: một đột biến mất một nucleotide T được biểu diễn là kiểu gen tham chiếu TT và kiểu gen thay thế T. Một biến thể đơn nucleotide từ A sang T được biểu diễn là kiểu gen tham chiếu A và kiểu gen thay thế T.
ALT	Các alen khác với đoạn đọc tham chiếu. Ví dụ: một vùng chèn thêm một T duy nhất được biểu diễn là kiểu gen tham chiếu A và kiểu gen thay thế AT. Một biến thể đơn nucleotide từ A sang T được biểu diễn là kiểu gen tham chiếu A và kiểu gen thay thế T.
QUAL	Điểm chất lượng theo thang Phred do trình phát hiện biến thể gán. Điểm càng cao thì độ tin cậy về biến thể càng lớn và xác suất lỗi càng thấp. Với điểm chất lượng Q, xác suất lỗi ước tính là 10-(Q/10). Ví dụ: bộ phát hiện Q30 có tỷ lệ lỗi 0,1%. Nhiều trình phát hiện biến thể gán điểm chất lượng dựa trên mô hình thống kê, thường cao hơn so với tỷ lệ lỗi quan sát được.

Bảng 2 Các chú thích trong tệp VCF

Tiêu đề	Mô tả
FILTER	Nếu tất cả các bộ lọc đều đạt, giá trị PASS sẽ được ghi trong cột bộ lọc. Các mục nhập FILTER trong quy trình công việc dòng mầm có thể bao gồm: • **DRAGENSnpHardQUAL**—Áp dụng nếu điểm QUAL của biến thể SNP không đạt ngưỡng • **DRAGENIndelHardQUAL**—Áp dụng nếu điểm QUAL của biến thể indel không đạt ngưỡng • **LowDepth**—Vị trí bị lọc vì độ sâu phạm vi không đạt ngưỡng • **LowGQ**—Vị trí bị lọc vì chất lượng kiểu gen không đạt ngưỡng • **PloidyConflict**—Kiểu gen từ trình phát hiện biến thể không phù hợp với mức bội thể của nhiễm sắc thể • **base_quality**—Vị trí bị lọc vì chất lượng base trung vị của các đoạn đọc thay thế tại lô-cut này không đạt ngưỡng • **filtered_reads**—Vị trí bị lọc vì tỷ lệ quá lớn các đoạn đọc đã bị lọc bỏ • **fragment_length**—Vị trí bị lọc vì chênh lệch tuyệt đối giữa độ dài phân đoạn trung vị của các đoạn đọc thay thế và tham chiếu tại lô-cut này vượt ngưỡng • **low_depth**—Vị trí bị lọc vì độ sâu đoạn đọc quá thấp • **low_frac_info_reads**—Vị trí bị lọc vì tỷ lệ đoạn đọc thông tin thấp hơn ngưỡng • **low_normal_depth**—Vị trí bị lọc vì độ sâu đoạn đọc của mẫu bình thường quá thấp • **long_indel**—Vị trí bị lọc vì chiều dài indel quá lớn • **mapping_quality**—Vị trí bị lọc vì chất lượng ánh xạ trung vị của các đoạn đọc thay thế tại lô-cut này không đạt ngưỡng • **multiallelic**—Vị trí bị lọc vì có hơn hai alen thay thế đi qua ngưỡng LOD của khối u • **non_homref_normal**—Vị trí bị lọc vì kiểu gen của mẫu bình thường không phải là tham chiếu đồng hợp tử • **no_reliable_supporting_read**—Vị trí bị lọc vì không có đoạn đọc soma hỗ trợ đáng tin cậy • **panel_of_normals**—Được ghi nhận trong ít nhất một mẫu trong tệp VCF của bảng nhập giá trị bình thường • **read_position**—Vị trí bị lọc vì trung vị khoảng cách giữa đầu/cuối đoạn đọc và lô-cut này thấp hơn ngưỡng • **RMxNRepeatRegion**—Vị trí bị lọc vì toàn bộ hoặc một phần alen biến thể là lặp lại của tham chiếu • **strand_artifact**—Vị trí bị lọc vì sai lệch theo chiều sợi nghiêm trọng • **str_contraction**—Vị trí bị lọc do nghi ngờ lỗi PCR, trong đó alen thay thế ít hơn một đơn vị lặp so với tham chiếu • **too_few_supporting_reads**—Vị trí bị lọc vì có quá ít đoạn đọc hỗ trợ trong mẫu khối u • **weak_evidence**—Điểm biến thể soma không đạt ngưỡng

Tiêu đề	Mô tả
FILTER (tiếp)	Các mục nhập FILTER trong quy trình công việc soma có thể bao gồm: • **base_quality**—Vị trí bị lọc vì chất lượng base trung vị của các đoạn đọc thay thế tại lô-cut này không đạt ngưỡng • **filtered_reads**—Vị trí bị lọc vì tỷ lệ quá lớn các đoạn đọc đã bị lọc bỏ • **fragment_length**—Vị trí bị lọc vì chênh lệch tuyệt đối giữa độ dài phân đoạn trung vị của các đoạn đọc thay thế và tham chiếu tại lô-cut này vượt ngưỡng • **low_depth**—Vị trí bị lọc vì độ sâu đoạn đọc quá thấp • **low_frac_info_reads**—Vị trí bị lọc vì tỷ lệ đoạn đọc thông tin thấp hơn ngưỡng • **low_normal_depth**—Vị trí bị lọc vì độ sâu đoạn đọc của mẫu bình thường quá thấp • **long_indel**—Vị trí bị lọc vì chiều dài indel quá lớn • **mapping_quality**—Vị trí bị lọc vì chất lượng ánh xạ trung vị của các đoạn đọc thay thế tại lô-cut này không đạt ngưỡng • **multiallelic**—Vị trí bị lọc vì có hơn hai alen thay thế đi qua ngưỡng LOD của khối u • **non_homref_normal**—Vị trí bị lọc vì kiểu gen của mẫu bình thường không phải là tham chiếu đồng hợp tử • **no_reliable_supporting_read**—Vị trí bị lọc vì không có đoạn đọc soma hỗ trợ đáng tin cậy • **panel_of_normals**—Được ghi nhận trong ít nhất một mẫu trong tệp VCF của bảng nhập giá trị bình thường • **read_position**—Vị trí bị lọc vì trung vị khoảng cách giữa đầu/cuối đoạn đọc và lô-cut này thấp hơn ngưỡng • **RMxNRepeatRegion**—Vị trí bị lọc vì toàn bộ hoặc một phần alen biến thể là lặp lại của tham chiếu • **strand_artifact**—Vị trí bị lọc vì sai lệch theo chiều sợi nghiêm trọng • **str_contraction**—Vị trí bị lọc do nghi ngờ lỗi PCR, trong đó alen thay thế ít hơn một đơn vị lặp so với tham chiếu • **too_few_supporting_reads**—Vị trí bị lọc vì có quá ít đoạn đọc hỗ trợ trong mẫu khối u • **weak_evidence**—Điểm biến thể soma không đạt ngưỡng • **systematic_noise**—Vị trí bị lọc dựa trên bằng chứng về nhiễu hệ thống trong các mẫu bình thường

Tiêu đề	Mô tả
INFO	Các mục nhập INFO trong quy trình công việc dòng mầm có thể bao gồm: • **AC**—Số lượng alen trong kiểu gen đối với mỗi alen THAY THẾ (ALT), theo đúng thứ tự được liệt kê. • **AF**—Tần suất alen đối với mỗi alen THAY THẾ, theo đúng thứ tự được liệt kê. • **AN**—Tổng số lượng alen trong các kiểu gen đã được phát hiện. • **DB**—Thành viên của dbSNP. • **FS**—Giá trị p theo thang Phred sử dụng kiểm định chính xác Fisher để phát hiện sai lệch theo chiều sợi. • **QD**—Độ tin cậy/Chất lượng biến thể theo độ sâu đọc. • **R2_5P_bias**—Điểm dựa trên sai lệch cặp đọc và khoảng cách từ đầu mồi 5. • **SOR**–Tỷ số chênh lệch đối xứng của bảng liên hợp 2x2 để phát hiện sai lệch theo chiều sợi. • **DP**—Độ sâu đoạn đọc xấp xỉ (thông tin và không thông tin); một số đoạn đọc có thể đã bị lọc dựa trên chất lượng ánh xạ, v.v. • **END**—Vị trí kết thúc của khoảng. • **FractionInformativeReads**—Tỷ lệ đọc thông tin trên tổng số đọc. • **MQ**—Chất lượng ánh xạ RMS. • **MQRankSum**— Giá trị Z-score từ phép kiểm định tổng hạng Wilcoxon, so sánh chất lượng ánh xạ của các đoạn đọc Thay thế so với Tham chiếu. • **ReadPosRankSum**—Giá trị Z-score từ phép kiểm định tổng hạng Wilcoxon, so sánh sai lệch vị trí của các đoạn đọc Thay thế so với Tham chiếu. • **SOMATIC**—Có ít nhất một biến thể tại vị trí này là soma. Các mục nhập INFO trong quy trình công việc soma có thể bao gồm: • **DP**—Độ sâu đoạn đọc xấp xỉ (thông tin và không thông tin); một số đoạn đọc có thể đã bị lọc dựa trên chất lượng ánh xạ, v.v. • **END**—Vị trí kết thúc của khoảng. • **FractionInformativeReads**—Tỷ lệ đọc thông tin trên tổng số đọc. • **MQ**—Chất lượng ánh xạ RMS. • **MQRankSum**— Giá trị Z-score từ phép kiểm định tổng hạng Wilcoxon, so sánh chất lượng ánh xạ của các đoạn đọc Thay thế so với Tham chiếu. • **ReadPosRankSum**—Giá trị Z-score từ phép kiểm định tổng hạng Wilcoxon, so sánh sai lệch vị trí của các đoạn đọc Thay thế so với Tham chiếu. • **AQ**—Điểm nhiễu hệ thống. • **hotspot**—Vị trí soma đã biết, được sử dụng để tăng độ tin cậy trong việc phát hiện. • **SOMATIC**—Có ít nhất một biến thể tại vị trí này là soma.

Tiêu đề	Mô tả
FORMAT	Cột định dạng liệt kê các trường được phân tách bằng dấu hai chấm. Ví dụ: GT:GQ. Các trường trong quy trình công việc dòng mầm có thể bao gồm: • **AD**—Độ sâu alen (chỉ tính các đoạn đọc thông tin trong tổng số đoạn đọc) cho alen tham chiếu và alen biến thể theo thứ tự liệt kê. • **AF**—Tỷ lệ alen cho các alen biến thể theo thứ tự liệt kê. • **DP**—Độ sâu đoạn đọc xấp xỉ (các đoạn đọc có MQ=255 hoặc cặp lỗi sẽ bị lọc bỏ). • **F1R2**—Số lượng đoạn đọc trong hướng cặp F1R2 hỗ trợ mỗi alen. • **F1R2**—Số lượng đoạn đọc trong hướng cặp F2R1 hỗ trợ mỗi alen. • **GP**—Xác suất hậu nghiệm theo thang Phred cho kiểu gen, như được định nghĩa trong thông số kỹ thuật VCF. • **GQ**—Chất lượng kiểu gen. • **GT**—Kiểu gen. 0 tương ứng với base tham chiếu. 1 tương ứng với mục nhập đầu tiên trong cột ALT. Và tiếp tục như vậy cho các mục nhập kế tiếp. Dấu gạch chéo (/) cho biết không có thông tin định pha. • **MB**—Thống kê thành phần theo mẫu để phát hiện sai lệch cặp đọc. • **PL**—Xác suất hợp lý theo thang Phred đã chuẩn hóa cho kiểu gen, như được định nghĩa trong thông số kỹ thuật VCF. • **PRI**—Xác suất tiên nghiệm theo thang Phred cho kiểu gen. • **PS**—Thông tin ID định pha vật lý, mỗi ID duy nhất trong một mẫu (chứ không áp dụng giữa các mẫu) kết nối các bản ghi trong một nhóm định pha. • **SB**—Thống kê thành phần theo mẫu, bao gồm Kiểm định chính xác Fisher để phát hiện sai lệch theo chiều sợi. • **SQ**—Chất lượng soma. Các trường trong quy trình công việc soma có thể bao gồm: • **AD**—Độ sâu alen (chỉ tính các đoạn đọc thông tin trong tổng số đoạn đọc) cho alen tham chiếu và alen biến thể theo thứ tự liệt kê. • **AF**—Tỷ lệ alen cho các alen biến thể theo thứ tự liệt kê. • **DP**—Độ sâu đoạn đọc xấp xỉ (các đoạn đọc có MQ=255 hoặc cặp lỗi sẽ bị lọc bỏ). • **F1R2**—Số lượng đoạn đọc trong hướng cặp F1R2 hỗ trợ mỗi alen. • **F1R2**—Số lượng đoạn đọc trong hướng cặp F2R1 hỗ trợ mỗi alen. • **GT**—Kiểu gen. 0 tương ứng với base tham chiếu. 1 tương ứng với mục nhập đầu tiên trong cột ALT. Và tiếp tục như vậy cho các mục nhập kế tiếp. Dấu gạch chéo (/) cho biết không có thông tin định pha. • **MB**—Thống kê thành phần theo mẫu để phát hiện sai lệch cặp đọc. • **PS**—Thông tin ID định pha vật lý, mỗi ID duy nhất trong một mẫu (chứ không áp dụng giữa các mẫu) kết nối các bản ghi trong một nhóm định pha. • **SB**—Thống kê thành phần theo mẫu, bao gồm Kiểm định chính xác Fisher để phát hiện sai lệch theo chiều sợi. • **SQ**—Chất lượng soma.
SAMPLE	Cột mẫu cung cấp các giá trị được chỉ định trong cột FORMAT.

Tệp Genome VCF

Tệp Genome VCF (*.gvcf.gz) tuân theo một tập hợp quy ước để biểu diễn tất cả các vị trí trong hệ gen theo định dạng gọn nhẹ hợp lý. Các tệp gVCF bao gồm toàn bộ các vị trí trong vùng quan tâm trong một tệp duy nhất cho mỗi mẫu. Tệp gVCF hiển thị không phát hiện tại các vị trí không đi qua tất cả bộ lọc. Một thẻ kiểu gen (GT) có giá trị ./. biểu thị trạng thái không phát hiện.

Xem kết quả phân tích

Các lần chạy đang diễn ra được hiển thị trong tab Active (Đang hoạt động). Các lần chạy đã hoàn thành được hiển thị trong tab Completed (Đã hoàn tất). Tham khảo Tài liệu hướng dẫn về sản phẩm NovaSeq 6000Dx (tài liệu số 200010105) để biết thêm thông tin về cách xem kết quả.

Hỗ trợ kỹ thuật

Để được hỗ trợ kỹ thuật, hãy liên hệ với bộ phận Hỗ trợ kỹ thuật của Illumina.

Trang web: www.illumina.com
Email: techsupport@illumina.com

Các số điện thoại hỗ trợ kỹ thuật của Illumina

Khu vực	Số miễn cước	Quốc tế
Úc	+61 1800 775 688	
Áo	+43 800 006249	+43 1 9286540
Bỉ	+32 800 77 160	+32 3 400 29 73
Canada	+1 800 809 4566	
Trung Quốc		+86 400 066 5835
Đan Mạch	+45 80 82 01 83	+45 89 87 11 56
Phần Lan	+358 800 918 363	+358 9 7479 0110
Pháp	+33 8 05 10 21 93	+33 1 70 77 04 46
Đức	+49 800 101 4940	+49 89 3803 5677
Hồng Kông, Trung Quốc	+852 800 960 230	
Ấn Độ	+91 8006500375	
Indonesia		0078036510048
Ireland	+353 1800 936608	+353 1 695 0506
Ý	+39 800 985513	+39 236003759
Nhật Bản	+81 0800 111 5011	
Malaysia	+60 1800 80 6789	
Hà Lan	+31 800 022 2493	+31 20 713 2960
New Zealand	+64 800 451 650	
Na Uy	+47 800 16 836	+47 21 93 96 93
Philippines	+63 180016510798	
Singapore	1 800 5792 745	
Hàn Quốc	+82 80 234 5300	
Tây Ban Nha	+34 800 300 143	+34 911 899 417

Khu vực	Số miễn cước	Quốc tế
Thụy Điển	+46 2 00883979	+46 8 50619671
Thụy Sĩ	+41 800 200 442	+41 56 580 00 00
Đài Loan, Trung Quốc	+886 8 06651752	
Thái Lan	+66 1800 011 304	
Vương quốc Anh	+44 800 012 6019	+44 20 7305 7197
Hoa Kỳ	+1 800 809 4566	+1 858 202 4566
Việt Nam	+84 1206 5263	

Các bảng dữ liệu an toàn (SDS)—Có trên trang web của Illumina tại địa chỉ support.illumina.com/sds.html.

Tài liệu hướng dẫn về sản phẩm—Có thể tải xuống từ support.illumina.com.

Illumina
5200 Illumina Way
San Diego, California 92122 U.S.A.
+1.800.809.ILMN (4566)
+1.858.202.4566 (ngoài khu vực Bắc Mỹ)
techsupport@illumina.com
www.illumina.com

Illumina Netherlands B.V.
Steenoven 19
5626 DK Eindhoven
Hà Lan

Nhà bảo trợ tại Úc
Illumina Australia Pty Ltd
Nursing Association Building
Level 3, 535 Elizabeth Street
Melbourne, VIC 3000
Úc

DÙNG CHO CHẨN ĐOÁN IN VITRO