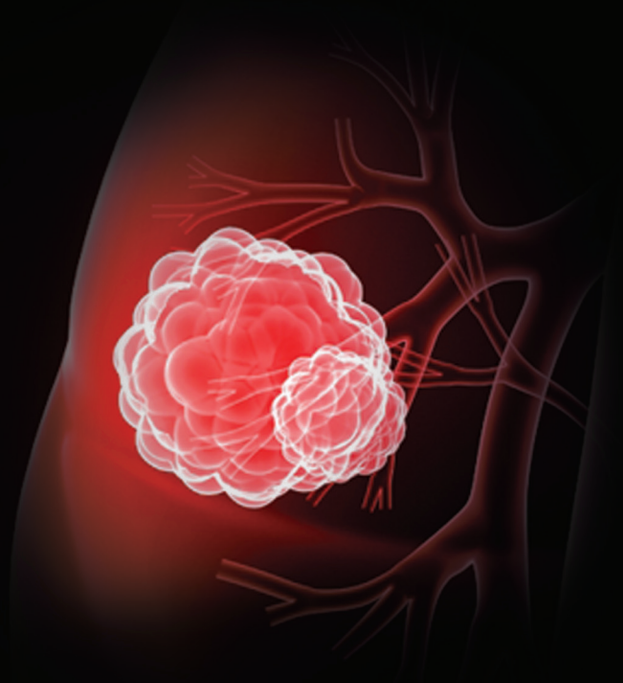
Application Note

BRCA1 と BRCA2 における変異の同定および比較



Application note on analyzing amplicon sequencing data and identifying cancer relevant variants using CLC Genomics Workbench

BRCA1 と BRCA2 における変異の同定および比較

Multiplicom の BRCA MASTR assay を Ion Torrent のシーケンシング技術および CLC Genomics Workbench のデータ解析と組み合わせることで、BRCA1 および BRCA2における癌関連変異を高感度かつ特異的に同定し、アノテーション付けを行うことができます。このアプリケーションノートでは、CLC Genomics Workbench を用いて、これら2つのよく研究された癌抑制遺伝子において、アンプリコンのシーケンシングデータをどのように解析して癌に関係した 変異 を同定するのかを紹介します。

データ

14例のDNAサンプルをマルチプレックスシーケンシングしたIon Torrent のデータ(sff フォーマット)
- Hg19におけるターゲット領域の位置情報を含む GFFファイル[1]
- バーコード情報とサンプルに対応する名前の含まれるExcelシート[2]
- ユニバーサルプライマー配列[3]
- 特異的なPCRプライマー配列[4]

11例のDNAサンプルと3例のコントロールサンプルについて、Multiplicom の BRCA MASTR v2.1 assay を用いて BRCA1 および BRCA2 の全てのエクソンを増幅しました。Ion Torrent のプロトコールに従って、断片化処理、バーコード配列付加、シーケンシングを行いました。

[1] customerservice@multiplicom.com にリクエストすることで入手可能

[2] Ion XpressTM Barcode Adapters kits 4471250, 4474009

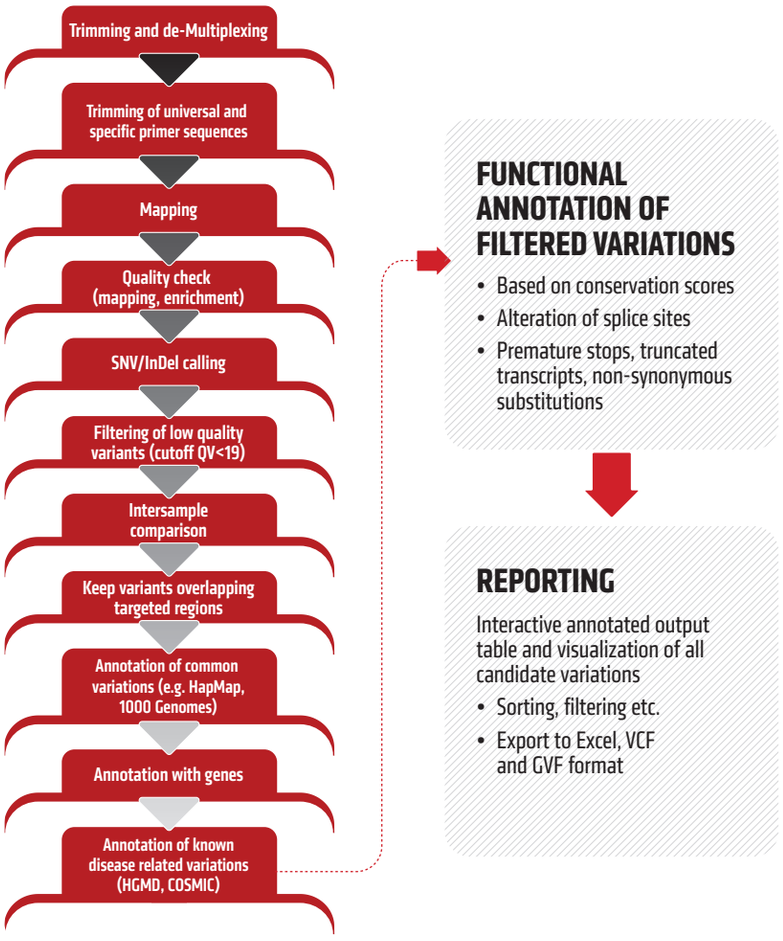
[3] Multiplicom

[4] Multiplicom

※オランダの LGTC (Leiden Genome Technology Center) において H.Buermans と彼のチームにより行われた研究です。

解析のワークフロー

血液サンプルからゲノムDNAを抽出しました[5]。Multiplicomのプロトコール[6]に従って、BRCA MASTR v2.1 Assay によりBRCA1 と BRCA2 の全てのコード領域を増幅し、ユニバーサルタグ配列を付加しました。ユニバーサルPCR (Short Read Amplification Kit, Multiplicom) を行った後、得られたアンプリコンライブラリーを、個人ごとにプールしました。各ライブラリーを、酵素的に断片化処理し[7]、精製しました[8]。引き続き、バーコードアダプター[9]をライゲーションし、異なるバーコードライゲーションライブラリーをプールしました。Ion OneTouch System Template Kit (100bp) を用いて、付属のプロトコールに従い、ゲル上で断片を100bp で size selection しました[10]。次いで、Ion316TM chipと Ion Sequencing Kit (100bp chemistry) を用いて、Ion TorrentTM Personal Genome MachineTM (PGMTM) によりシーケンシングしました。



[5] QiAamp DNA Blood kit, Qiagen

[6] Multiplicom

[7] Ion XpressTM Plus Fragment Library Kit, Life Technologies

[8] AMPure® XP, Agencourt

[9] Ion XpressTM Barcode Adaptors 1-16 Kit, Life Technologies

[10] MinElute Gel Extraction kit, Qiagen

ヒトリファレンスデータのダウンロード・準備

リファレンスデータは、"Download Genome" ツールを用いて準備することができます。この例では、Homo sapiens (hg19) を選択し、配列、遺伝子のアノテーション、dbSNP, COSMIC, HapMap, 1000 Genomes Project に含まれる変異 をダウンロードしました。データ量が大きいため、ダウンロードには時間がかかります。

ターゲット領域の外に誤ってマップされたリードは variant コーリングにおいて偽陽性につながる可能性があります。リードは、シーケンスのエラー、リファレンスゲノム上の類似した配列、非特異的な増幅などにより誤ってマッピングされる可能性があります。

解析の途中で、これらのアーティファクトをフィルタリングして取り除く選択肢をもつために、、ターゲット領域のアノテーション付けを行います。このために、ターゲット領域の情報を持つファイルをインポートします。

それぞれのトラックは、トラックリストを作成することにより、一緒に閲覧することができます。次のよ

うなトラック　からなるトラックリストを作成します。

- Homo_sapiens.GRCh37.66.dna.toplevel（配列）
- Homo_sapiens.GRCh37.66.gtf.gz_CDS
- Homo_sapiens.GRCh37.66.gtf.gz_Gene
- ターゲット領域

トラックは、ドラッグ アンド ドロップを用いて、除いたり、加えたり、配置を変えたりすることができます。

リードのトリミングと de-multiplexing

ヒトリファレンス配列へリードを精度よくマップするために、すべてのリードを、マッピングの前にトリムする必要があります。そのために、関連するアダプター配列とプライマー配列のすべてをworkbenchにインポートします。

二つのライブラリー特異的なアダプター配列にサンプル特異的なバーコードが隣接しています。これらのアダプターは、de-multiplex の際の "リンカー" としてデザインされています。それらは、解析の間に自動的にトリムされます。

de-multiplexを行うために、"Process Tagged Sequences" ツールを使用します。このとき、サンプルに対応するバーコードおよびサンプル名を含むファイルを用います。

結果として、それぞれのサンプルにもとづいて名前がつけられた14個のフォルダーが作成されます。それぞれのフォルダーには、指定されたバーコードにより同定されたすべてのリードが含まれます。次に、それぞれのフォルダーの中のリードに対して、ユニバーサルタグプライマーとターゲット特異的なプライマーがトリムされます。

リード配列のマッピング

リード配列は、"Read Mapper"を用いてリファレンスゲノム (hg19)にマップされます。それぞれのサンプルのリード配列は、指定したパラメータ設定を用いて、別々にヒトリファレンス配列にマップされます（insertion costs=2, deletion costs=2, length fraction=0.5, similarity=0.8)。

すべてのサンプルを同時に解析するために、マッピングをバッチモードで実行します。新たに得られたトラックを、作成した トラックリスト に加えます。

Enrichment

"Target Regions Statistics"ツールでは、enrichment領域の特異度とカバレッジについての統計値を計算します。

Variant calling

Probabilistic Variant Detection11 を用いて、すべてのサンプルにおいて、マッピングされたリード配列中の短い挿入・欠失およびSNVsを同定します。結果として得られるトラックは、average base quality <19 および、

図1: バーコード6のサンプルに見いだされたBRCA2 の 4塩基の欠失は乳がんとの関連が知られていた（HGMD）

Name	Sequence
Universal primer Tag 1	AAGACTCGGCAGCATCTCCA
Universal primer Tag2:	GCGATCGTCACTGTTCTCCA
Target-specific primers	Available on request by email at customerservice@multiplicom.com

Table 1: Subset of some of the identified unknown somatic variants

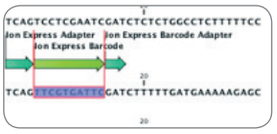


図2: Ion Express library preparation kit の部分としてのアダプター配列とバーコード配列の構成

比較のためaverage base quality <16でフィルターします。結果として得られたトラックは、作成した トラックリスト にドラッグ アンド ドロップで追加します。

[11]
- Minimum coverage: 10
- Variant present in forward and reverse reads: No
- Maximum expected variations: 2
- Variant Probability: 90.0
- 454/Ion Torrent correction: No
- Use only specific matches: Yes

変異のフィルタリング、アノテーション付け、および比較

異なるサンプルで得られた 変異 は、"Compare Variants within Group" ツールを用いて比較することができます。各々の 変異アレルに対して、このツールを用いることで、アレルが見いだされるサンプル数を数え、リストを作成することができます。すべてのサンプルに存在する変異アレルを得るために、frequency thresholdは 0%で解析を行います

フィルタリングとアノテーション付けは、自動的なワークフローによって実行することができます。本解析では、次のように精査しました。

- "Filter against Overlapping Annotations" により、ターゲット領域にオーバーラップするすべての変異を抽出します。
- "Annotate from Variant Database" により変異データベース（例えば、HapMap および 1000人ゲノムプロジェクト）に対してフィルタリングを行い、変異候補を同定します。
- "Annotate from Overlapping Annotations" により候補の 変異 のアノテーション付けを行い、起こりうるfunctional consequences と既知の変異に基づいて優先順位を決定します。
- "Annotate from Variant Database" により、変異データベース（例え

Sample	# Reads after demultiplexing	Avg. length	# Reads after trimming	Avg. length	% Reads mapped	% Specificity	Avg. coverage	# Variants > QV19	% Concordance	# Variants >QV16	% Concordance
Barcode 1	127,360	142	95,460	116.5	89.7	99.6	339	43	0.97	49	1.0
Control 1	280,229	138	213,803	115	83.22	99.67	620	30	1.0	36	1.0
Control 2	162,518	137	122,559	115	83.6	99.7	360	15	1.0	21	1.0
Control 4	212,396	137	161,301	114.6	83.55	99.67	471.8	42	1.0	45	1.0

Table 2: Overview of summary statistics from selected samples[14]

ば、COSMIC12 および HGMD13 - HGMDデータベースにアクセスするためにはライセンスが必要です)からの情報を用いて既知の変異をアノテーション付けします。
- "Amino Acid Changes" は、アミノ酸の置換について調べます。

結果を、作成された トラックリスト に追加します。

[12] Forbes et al. (2011) Nucl. Acids Res. 39 (suppl 1): D945-D950

[13] Stensen et al. (2009) Genome Med. 1:13

ここまでで、すべての変異は、オーバーラップする遺伝子の名前、アミノ酸置換、起こりうるスプライスの影響、conservation scores、変異データベース(この例では、COSMIC およびHGMD)へのリンク、データセットの 変異の由来とともにアノテーション付けされます。さらに、HapMap または 1000人ゲノムプロジェクトのデータ中にvariantが見いだされた場合、アノテーション付けされます。

評価された変異の比較

1つの癌サンプルおよび3つのコントロールサンプルからの変異のいくつかは以前にサンガーシーケンシングにより評価されました。今回得られた変異をこれらのデータセットと比較し、解析の感度を評価します。

Results

Enrichment

トリムされたリードのおよそ90%がヒトリファレンスゲノム (hg19) にマップされました。99.6%の整列されたリードは、部分的にまたは完全に増幅された領域にオーバーラップしました。平均のカバレッジは 330から620の間でした。ターゲット領域の全ての塩基の99.9-100%が少なくとも100xのカバレッジがありました。

変異の比較

ホモポリマーのエラー補正を行わず、 quality cut-off を19とし、すべてのサンプルを合わせて、ターゲット領域とオーバーラップする87個のvariantアレルを見つけました。34個の変異は、1000人ゲノムプロジェクトのデータで既知であり、31個の変異は、HapMapの中で既知でした。5個の変異は、COSMICの中に見いだされ、このうち3個が1000人ゲノムプロジェクトのデータでも見いだされました。これらは、癌関連の変異である可能性があります。16個の変異がHGMDに存在しており、そのうち14個が癌と関連していました。42個の変異(InDels, MNVsおよびSNVs)は、上記のデータベース中に見出すことができず、また、参照した方法によっても検出することができませんでした。これらの変異は偽陽性である可能性がありますが、それらは、さらなる評価の対象として検討されるべきでしょう。しかしながら、そのうち9個がデータベース中の変異と部分的にオーバーラップしており、これらが本当に真の陽性である可能性が高いと考えられます。

454/Ion Torrent homopolymer error correction のために、HGMDおよびCOSMICに見いだされる乳がんに関連する4塩基の欠失および挿入を検出することができませんでした。

3個のサンプル(Control 1, Control 2 および Control 4)をコントロールとし、残りのサンプルをケースとして使い、 フィッシャーの正確確率検定を用いて各々の変異アレルが確率的に癌に特異的と言えるかどうかを比較したところ、どの変異アレルも 0.05というカットオフ値を超えませんでした。しかしながら、36個の変異アレルはコントロールサンプルのいずれでも見いだされず、また、そのうち16個は、アノテーション付けられたいずれのデータベースでも既知ではありませんでした。

評価後の変異との比較

Table2では、ホモポリマーエラー補正を用いず、quality threshold の設定を16にすることにより、CLC Genomics Workbench が100%の感度を達成できることを示しています。代わりにquality threshold の設定を19にすることにより、一つの偽陰性が生じますが、ホモポリマーの領域でコールされる変異の数は増加しました。

[14]
- % Reads Mapped -(マップされたリードの数)/(トリムされた後の総リード数)
- Specificity - ターゲットにマップされたリード vs. マップされたアンプリコン
- Avg. Coverage - あるポジションまたは領域が平均何リードでカバーされているか
- % Concordance ^ (検出された true positive)/(サンガーシーケンスで validateされた true positive)

CLC bio · EMEA
Finlandsgade 10-12
Katrinebjerg · DK-8200 Aarhus N
Denmark
Phone: +45 7022 5509

CLC bio · Americas
10 Rogers St # 101
Cambridge · MA 02142
USA
Phone: +1 (617) 945 0178

CLC bio · AsiaPac
69 · Lane 77 · Xin Ai Road · 7th fl.
Neihu District · Taipei · Taiwan 114
Taiwan
Phone: +886 2 2790 0799