次世代シーケンサを用いた疾患関連変異の探索

多発家系および罹患同胞を対象とした、次世代シーケンサを用いた全エクソーム解析の実例

ポイント

- 全エクソーム解析の結果得られた数万個の変異からどのように疾患に関連する変異を絞り込むか
- どのようなフリーのソフトを組み合わせると解析ができるか
- トリオの全エクソームの変異情報と、罹患同胞の 16 遺伝子の変異情報から、疾患関連候補を 2 つに絞り込むことができた

1．目的

次世代シーケンサを用いた疾患関連遺伝子の変異（SNV/Indel）解析において、変異検出以降の絞り込みは重要である。目的変異の絞り込みには、遺伝子型や既知の変異の情報の活用が効果的であると考えられる。我々は遺伝子型と既知の変異の情報をアノテーション付けするソフト QmergeVCF を作製し、アノテーションに基づく絞込みの効果を検証するため以下の解析を行った。あるレーバー先天黒内障（LCA）が多発しているパキスタン家系において、Falk らによる研究で、この家系には NMNAT1 の 9 番目のバリンをメチオニンに置換させる変異が生じており、この変異をホモで持つとき LCA を発症することが報告されている [1]。公開されているこの家系の Exome sequence データから、弊社の resequencing パイプライン（Reseq パイプライン）（図 1）で変異を検出し、QmergeVCF（図 3）でアノテーションして絞り込みを行った結果と Falk らの論文の結果を比較し、変異の絞り込みの効果を確認した。

2．使用したデータ

公開されているパキスタン家系トリオ（III-3、III-5、IV-1）の exome sequence データ（.fastq）を使用した（図 2）[3]。

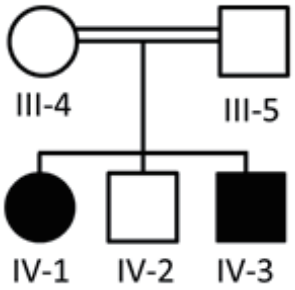


図 2　家系図

Illumina HiSeq 2000 で pair-end シーケンスをされた、リード長が 101 bp、リード数が約 76,000,000 ～ 88,000,000 のデータを用いた。

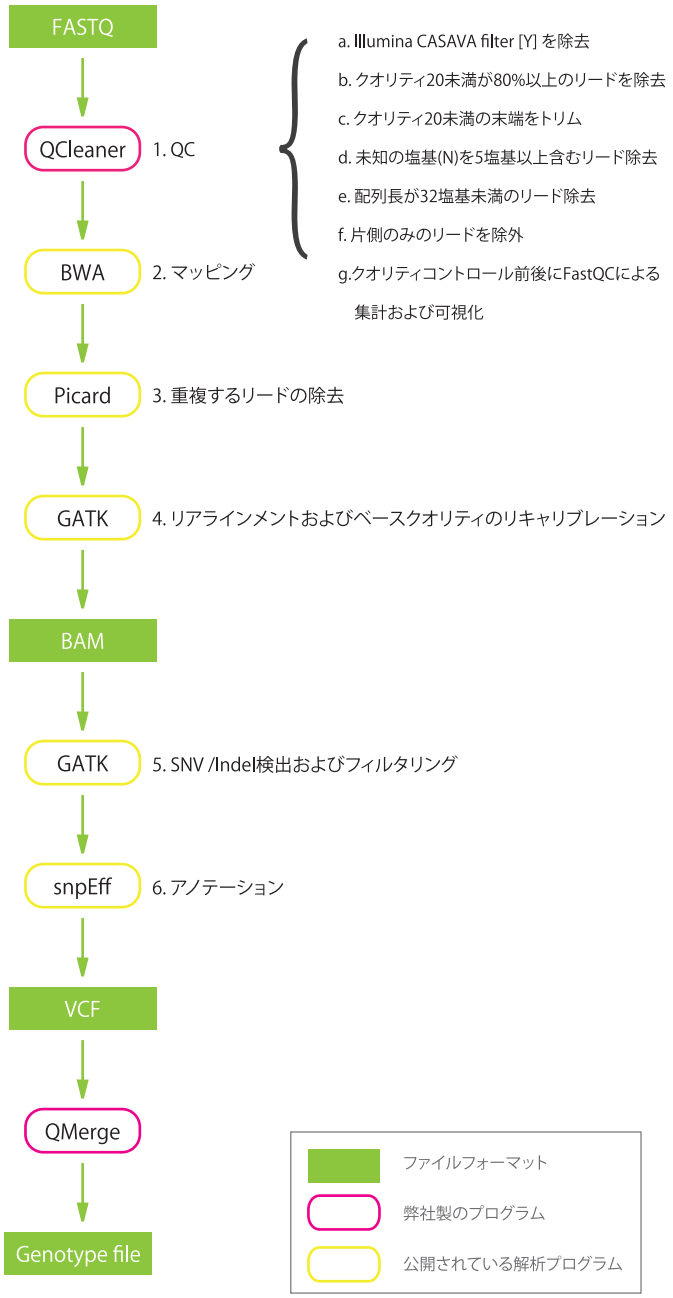


図 1　Reseq パイプライン

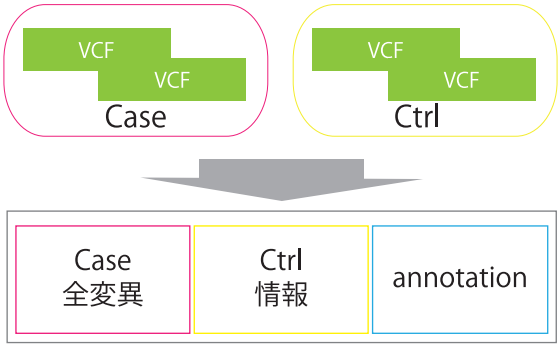


図 3　Qmerge の機能イメージ

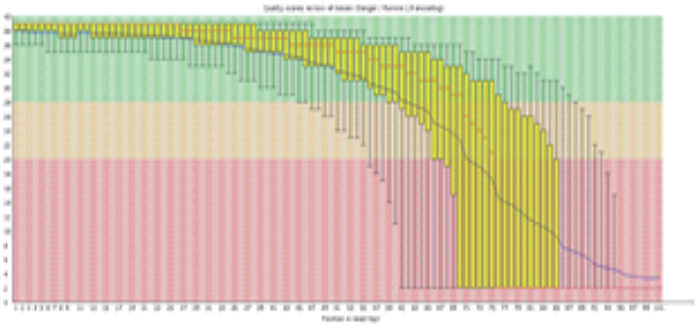
・複数の VCF ファイルを入力
・Case サンプルの変異を基準に、Ctrl サンプルの情報を付与
・Ctrl サンプルで変異が検出されていない場合、リファレンスと同一か、シーケンスされていないか判定

3．解析手法

3.1．クオリティ・コントロール（QC）

FastQC を用いてクオリティを確認し、弊社の QCleaner を用いて 6 項目のフィルタリングを行った（図 1）[3]。QC は低クオリティのリードに起因する偽陽性の発生を防ぐため重要だと考えられている。実際に、リードのクオリティを確認した結果、3’ 末端側に低クオリティの塩基を持つリードが含まれていることが分かった（図 4）。

a　QC 前



b　QC 後

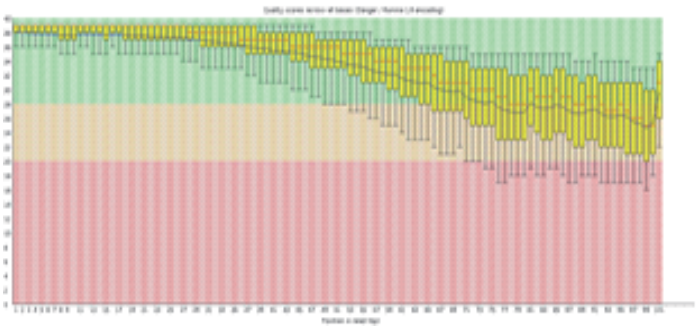


図 4　QC 前後のリードのクオリティ

3.2．マッピングと変異検出

bwa を用いて、クリーニングしたリードをリファレンス配列（hg19、1000 人ゲノムプロジェクトのスキャフォルド配列およびデコイ配列 version 5）にマッピングした。マッピング結果（.bam）から、Picard を用いて重複リードの除去、GATK を用いてリアライメント、及びリキャリブレーションを行った。その後 GATK を用いて変異を検出し、変異コールのクオリティ情報を付与した。最後に snpEff を用いて変異に遺伝子情報を付与した（図 1）。その結果、マッピング率は 98.6 ～ 99.8%、カバレージ x10 でターゲット領域の 26.2 ～ 80.6% をシーケンスできていた。また、各サンプルから約 49,000 ～ 81,400 個の変異が検出された（表 1）。

a　マッピング率

		average coverage	mapped region(bp)
IV-1	coverage(x10)	53.5	82,527,978
	(x30)	68.5	52,138,602
III-4	coverage(x10)	15.3	24,233,674
	(x30)	35.1	1,404,782
III-5	coverage(x10)	42.8	74,773,265
	(x30)	59.9	40,374,945

b　変異数

	SNV	Indel
IV-1	71,949	9,471
III-4	46,675	2,332
III-5	71,660	8,507

表 1　マッピング率と変異数

3.3．疾患関連変異候補の探索

3 サンプルの変異情報ファイル（.vcf）から、QmergeVCF を用いて Case を基準に Control の変異情報を検索し、ファイルの統合と変異情報の付与を行った [4]。得られた変異のうち、既知の変異（dbSNP131）、アレル頻度の高い変異（1000 人ゲノムプロジェクト、AF>0.05）を除外した。次に、GATK により付与したクオリティフィルタで変異検出の信頼度が低い変異をフィルタリングした。さらに snpEff により予測およびアノテーションされた変異がタンパク質の機能に与える影響（インパクト）が小さい変異を除外した。また、Control は LCA を発症しておらず、遺伝形式は劣性遺伝であると考えられることから、Control においてヘテロ接合、Case においてホモ接合の変異を選び出し、その結果 16 個の疾患関連候補遺伝子を得た（図 5）。

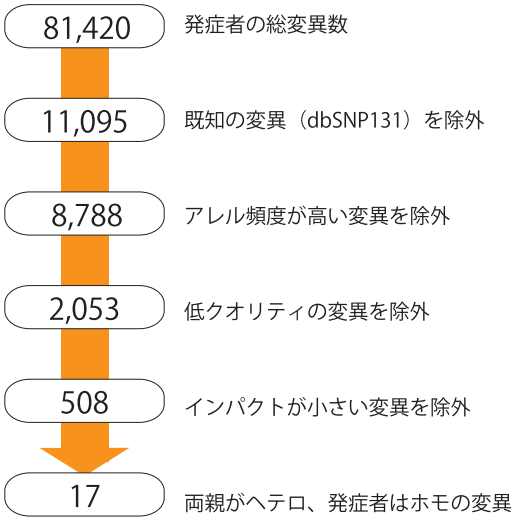


図 5　トリオを用いた絞り込み

3.4．同胞発症例によるバリデーション

解析を行っている家系内には、トリオ解析で採用した IV-1 の他に、LCA を発症している同胞（IV-3）と未発症の同胞（IV-2）がおり、IV-3 は IV-1 と同じ疾患関連変異をホモ接合で持つ一方で、IV-2 はこの変異を持たないか、ヘテロ接合で持つと考えられる（図 2）。そこで、この 3 人の変異を比較するため、3 人の exome sequence データから検出した変異のうち 3.3 節で得られた 16 個の遺伝子上の変異をすべて抜き出し、3.3 節と同様に 3 サンプルの変異情報を統合した。その後、既知の変異、アレル頻度、クオリティフィルタ、そしてインパクトによって絞り込み、最後に LCA を発症している IV-1 と IV-3 の両方でホモ接合だが、発症していない IV-2 ではヘテロ接合または野生型である変異を選び出した。最終的に 2 個の疾患関連候補遺伝子に絞り込むことができた（図 6）。

図 6　発症同胞を用いた絞り込み

この 2 個のうち 2 つは Falk らの論文でこの家系の LCA 原因遺伝子として報告された NMNAT1 で、9 番目のバリンをメチオニンに置換する変異であった。

4．まとめ

Reseq パイプラインと QmergeVCF を用いてトリオ解析を行った結果、疾患関連変異を絞り込むことで、16 個の候補遺伝子を得た。この候補遺伝子に対して、同胞発症例を対象に解析を行うことで、候補遺伝子を 2 個に絞ることができた。この候補変異はこの家系において LCA を引き起こしていると報告された NMNAT1 上の変異を含んでいた。

以上の結果から、疾患関連変異探索において、本解析で用いた dbSNP やアレル頻度など、既知の変異のデータベース、推定遺伝形式に基づく絞り込みも効果的であることが示された。

今後、より大規模なデータベースや、特定の人種のデータベースの利用により、さらに解析精度を高められると考えられる。また、本解析で用いた QmergeVCF は、各サンプルの変異情報の統合や、変異情報、遺伝子型などのアノテーションの付与を行うことで変異の絞り込みに寄与し、複数サンプルの比較による疾患遺伝子の探索に役立つツールであることが示された。

5．参考文献

1. Falk M., Zhang Q., et al., NMNAT1 mutations cause Leber congenital amaurosis. Nat Genet. 2012 September; 44(9): 1040–1045.
2. Available from DNA Data Bank of Japan, Sequence Read Archive, Accession number SRP013517 (http://trace.ddbj.nig.ac.jp/DRASearch/study?acc=SRP013517)

6．ソフトウェア

3. QCleaner（http://amelieff.jp/software）
4. QmergeVCF (http://amelieff.jp/software)