# 2　遺伝性疾患の網羅的エクソームシークエンスデータの解析法

重水　大智
理化学研究所統合生命医科学研究センター
医科学数理研究グループ

# Analysis for Whole Exome Sequencing Data in Hereditary Diseases

Daichi SHIGEMIZU

*RIKEN Center for Integrative Medical Sciences*
*Laboratory for Medical Science Mathematics*

**要　旨**

次世代シークエンス技術の目覚ましい発展により，全ゲノムシークエンスに比べ低コストでより深く遺伝子コード領域をシークエンスする全エクソームシークエンス解析が疾患関連変異を同定する解析の応用として盛んに行われ始めている．一方，次世代シークエンサーで読まれる配列は比較的短いため（50－150bp），ショートリードを標準ゲノム配列に対して正確にマッピングし，高精度にバリアントを同定することが難しく，洗練された手法の開発が課題のひとつとしてあげられている．

本稿では次世代シークエンサーで読まれたショートリードのデータから高精度に一塩基置換（SNV，一致率99.9％），挿入・欠失（Indel，一致率97.3％）の同定に成功した手法の紹介と，家系データを用いた遺伝性疾患原因変異探索の際の注意点とそのアプローチの仕方を紹介する．

**キーワード**：網羅的エクソームシークエンス，遺伝形式，家系解析

## はじめに

次世代シークエンサーはゲノムを断片化し，各断片を両端から約100 bp程度シークエンスする．得られた膨大な塩基配列は標準ゲノム配列に計算機でマッピングされ，ゲノム上での位置が決定される．シークエンスされた配列は読み取り長が短くマッピングが困難なため，誤ったバリアントの同定を引き起こす問題（偽陽性：false positive; FP，偽陰性：false negative; FN）が指摘されている．特にメンデル遺伝病（単一遺伝子疾患）の場合，疾患原因変異は1つであるためFPやFNを極力抑えることが原因解明への重要な鍵となる．

疾患原因変異の同定法として大きく二通りが考えられる．一つが家系情報を基に遺伝形式を想定し同定する方法，もう一つが家系情報なしの独立な発症者集団と非発症者集団を比較し同定する方法である．後者は多重検定補正をクリアする統計学的有意差が認められる変異を見出すために十分な検体数が必要とされ，前者の解析法が現在の主流となっている．

## バリアントコールと精度評価

### 1．シークエンスデータ

理化学研究所の倫理委員会で承認された検体

**Reprint requests to**: Daichi SHIGEMIZU
RIKEN Center for Integrative Medical Sciences
1－7－22 Suehiro－cho Tsurumi－ku,
Yokohama 230－0045 Japan

**別刷請求先**：〒230－0045　横浜市鶴見区末広町1－7－22
理化学研究所統合生命医科学研究センター
重水大智

RK001の末梢血を使用した．シークエンスは，ヒト全エクソンキャプチャキット Agilent 社の SureSelect V4 を用いエクソン領域の DNA 断片を抽出した後，Illumina 社の HiSeq2000 を用い各断片の両端から 101bp（ペアエンド法）シークエンスした．得られた約 7Gbp のショートリードの 98.5 %が標準ゲノム配列へマッピングされ，そのうち約 5 %が重複する PCR 断片として除去された．標準ゲノム配列へのマッピング，重複する PCR 断片の除去はそれぞれ BWA[1] と SAMtools[2] によって実装された．

### 2．SNV，Indel コール

ヒト全エクソンキャプチャキットで設計されたプローブ領域（on-target）における平均リードカバレッジは 70.7 reads/base で，約 4.1 %が SNV，Indel コールに必要な最低リード数を下回った．バリアントコールには独自に開発した VCMM（Variant Caller with Multinomial probabilistic Model）[3][4] を使用した．シークエンスデータのマッピングから SNV，Indel コールまで完全自動化された解析パイプラインにより，検体あたり 5 時間程度で解析が行える（200 CPU 使用時）．

### 3．精度評価

コールした SNV の精度評価にはエクソーム領域に着目して多型を多くカバーしている Illumina Human Exome BeadChip を使用した．190，197 ヵ所が評価対象となり，その一致率は 99.97 %であった（FP = 0.0036 %，FN = 0.0084 %）．Indel に対しては大規模な精度評価が困難で，コールした Indel を対象にサンガーシークエンス法で検証し，精度を見積もった（一致率：94.67 %，75 ヵ所中 71 ヵ所が一致）．

## 変異候補の絞り込み

### 1．public，in-house データ

遺伝子コード領域における SNV 数は検体あたり 20,000 前後と言われ[5]，メンデル遺伝病（単一

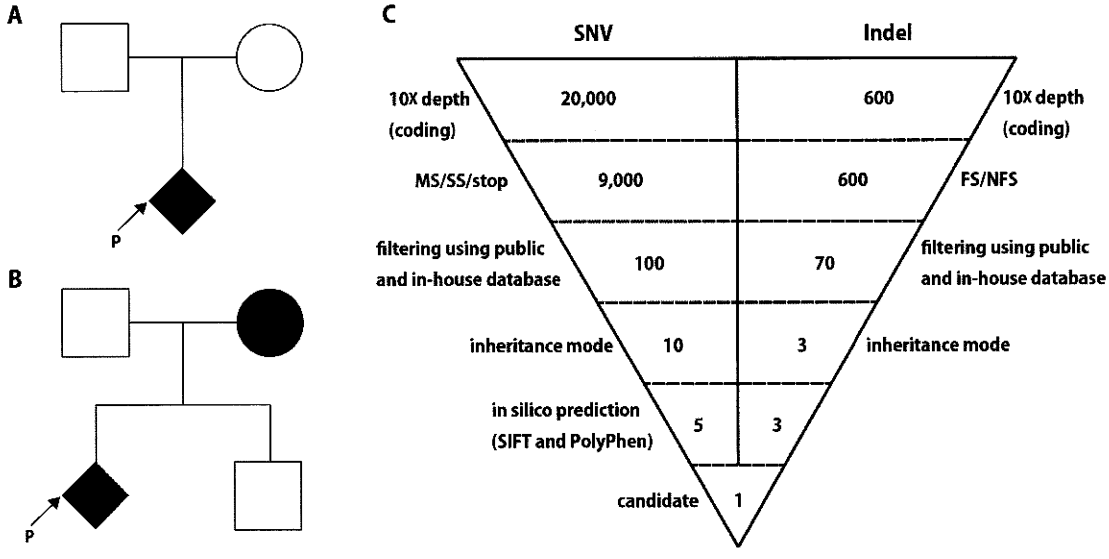


**図 1**　劣性遺伝または *de novo* 変異が想定される家系図（A）と優性遺伝が想定される家系図（B）と，家系解析の流れ（C）．MS: Missense, SS: Splice Site variant, stop: Nonsense, FS: Frameshift, NFS: Non Frameshift

遺伝子疾患）の場合，その中の一つが疾患原因変異である．当然，家系情報だけで絞り込むことは難しい．よって通常，疾患と関連がないバリアント（例えばSNP）を用いた絞り込みが行われ，dbSNP[6]，1000 genomes project[7]，NHLBI Exome Variant Server[8]等の公共データベースが利用される．しかしながら検体あたり300弱の変異候補が残り，近年，独自にシークエンスしたデータ（in-house data）をさらなる絞り込みに使用するケースが増えてきている．この絞り込みが試薬特異的に起こるエラーの除去や集団特異的なSNPの除去を可能にすることが期待され，実際約1,200検体分のin-house dataで絞り込みを行った結果，300弱の変異候補が70程度まで絞り込めた．

### 2．家系図と遺伝形式

家系図から推定される遺伝形式として劣性遺伝形式，*de novo*変異（**図1A**），優性遺伝形式（**図1B**）がある．**図1A**の*de novo*変異または劣性遺伝形式（ホモ接合と複合ヘテロ接合）が想定される家系は比較的変異候補の絞り込みが可能であるが，**図1B**の優性遺伝形式が想定される家系は候補の絞り込みが困難で，ある程度家系内の検体をシークエンスする必要がある．

## おわりに

全ての家系解析において変異候補が見つかるとは限らない．候補を絞り込めない場合もあれば，逆に見つからない場合もある．前者の場合，家系内の追加検体のシークエンスにより改善が見込めるが，後者の場合WESがターゲットとしている領域外に疾患原因変異がある可能性がある．またSNV，short Indel以外の，例えばコピー数バリアント（CNV）やlong Indelが原因変異の場合，現在のプロトコルでは検出が困難である．これらの問題は未だ解決されておらず，今後の検討課題である．

## 参考文献


1）Li H and Durbin R: Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754-1760, 2009.

2）Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup: The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078-2079, 2009.

3）Fujimoto A, Nakagawa H, Hosono N, Nakano K, Abe T, Boroevich KA, Nagasaki M, Yamaguchi R, Shibuya T, Kubo M, Miyano S, Nakamura Y and Tsunoda T: Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing. Nat Genet 42: 931-936, 2010.

4）Shigemizu D, Fujimoto A, Akiyama S, Abe T, Nakano K, Broevich KA, Yamamoto Y, Furuta M, Kubo M, Nakagawa H and Tsunoda T: A practical method to detect SNVs and indels from whole genome and exome sequencing data. Sci Rep 3: 2161, 2013.

5）Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA and Shendure J: Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet 12: 745-755, 2011.

6）Smigielski EM, Sirotkin K, Ward M and Sherry ST: dbSNP: a database of single nucleotide polymorphisms. Nucleic Acids Res 28: 352-355, 2000.

7）Via M, Gignoux C and Burchard EG: The 1000 Genomes Project: new opportunities for research and social challenges. Genome Med 2: 3, 2010.

8）Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, Gabriel S, Rieder MJ, Altshuler D, Shendure J, Nickerson DA, Bamshad MJ; NHLBI Exome Sequencing Project, Akey JM: Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature 493: 216-220, 2013.