

平成 27 年度 国内外における遺伝子診療の実態調査報告書

研究・医療面において有用なデータシェアリング基盤の検討

2016 年 3 月 31 日

 株式会社三菱総合研究所

目 次

1. 目的.....	1
2. 概要.....	2
3. 調査の方法.....	3
3.1 調査のフロー.....	3
3.2 調査対象.....	4
4. データシェアリング基盤に関する調査.....	6
4.1 既存データシェアリング基盤の整理.....	6
4.2 研究のための既存データシェアリング基盤.....	8
4.3 医療のための既存データシェアリング基盤（ClinVar）.....	23
4.4 ClinVar に関連する研究プロジェクト（ClinGen）.....	57
5. データシェアリング基盤に関する検討.....	74
5.1 国内有識者から得られた情報・意見.....	74
5.2 研究のためのデータシェアリング基盤.....	77
5.3 医療のためのデータシェアリング基盤.....	80
6. まとめ.....	83
参考文献.....	84

1. 目的

【データシェアリングに関する国際動向】

ゲノム研究の周辺領域では、研究によって得られたデータの共有（以降、「データシェアリング」という。）を行っている。データシェアリングは、研究コミュニティの目標を効率的に達成するとともに、研究によって得られたデータの検証や利用を促進することによって、共有されたデータを健康医療分野の研究に役立てることを目指している。この考え方は、ヒトゲノム計画実施中に採用された配列データの共有方針を始まりとして、米国の NIH や英国の Wellcome Trust が先導役となって推進している。

【わが国での取組】

わが国では、健康・医療戦略推進本部が、「平成 28 年度 医療分野の研究開発関連予算の概算要求のポイント」⁽¹⁾において、「3. 世界最先端の医療の実現に向けた取組」の「⑤ 疾病克服に向けたゲノム医療実現化プロジェクト」として、以下のようなデータシェアリングを含めた取組を進めることとしている。

（ア）ゲノム医療実現推進プラットフォーム事業

- ✓ 既存のバンク等をプラットフォームとして構築するとともに、その研究基盤を利活用し、以下の研究開発を推進
 - 糖尿病、循環器疾患等、多くの国民が罹患する一般的な疾患を対象とした研究開発
 - 疾患予防や治療の最適化に向けた発症予測法の確立 等

（イ）臨床ゲノム情報統合データベース整備事業

- ✓ 患者リクルートと全ゲノム解析等の実施
- ✓ 臨床ゲノム情報統合データベースの管理運営

（ウ）AMED によるマッチング・データシェアリングポリシー策定

- ✓ 患者リクルートと全ゲノム解析等の実施
- ✓ ワンストップサービス（基盤活用の仲介等）
- ✓ ポリシー策定（データ集約・IC フォーマット等） 等

【目的】

これらの実現のためには、（ア）と（イ）ではデータシェアリング基盤の構築、（ウ）では研究プロジェクト実施に対しての支援機能の確立と、データシェアリングポリシーの策定が求められている。本調査の目的は、それらの基盤及び支援機能の要件整理及び AMED によるデータシェアリングポリシーの策定を支援することである。

2. 概要

本調査では、有識者に対するインタビュー調査や先行事例に対する文献調査を行い、さらに、それらに基づいて、ゲノム医療実現推進プラットフォーム事業及び臨床ゲノム情報統合データベース整備事業で構築されることとなるデータシェアリング基盤の要件整理、AMED によるデータシェアリングポリシーの策定支援を実施した。

研究に用いられるデータシェアリング基盤の先行事例に関する調査としては、米国の NCBI 及び欧州の EMBL-EBI が構築・運営するものを中心に、取り扱われるデータや、データの登録方法、研究者に対する公開方法についての調査を行った。その整理結果を「4.1 章 既存データシェアリング基盤の整理」に整理する。

ゲノム医療に係るデータシェアリング基盤のうち、NCBI の dbGaP と EMBL-EBI の EGA は、誰でもデータを閲覧可能な「非制限公開」と、閲覧許可を与えられた者のみがデータを閲覧可能な「制限公開」というメカニズムをになっている（わが国では、科学技術振興機構のバイオサイエンスデータベースセンターによる JGA 等が、それらに対応するデータシェアリング基盤としてサービスを提供している）。これらに加えて、AMED のデータシェアリングポリシーには「制限共有」というメカニズムが検討されている。これは、「データマネジメントプランに基づいてデータベースに登録することにより、データマネジメントプランに記載された研究者及びデータアクセス申請を承認された研究者間で共有する」ということであり、研究実施者が、自らが作成したデータを用いて研究成果を出すためのメカニズムとして位置づけられる。本調査では、そのような共有方法の事例として、英国の DECIPHER 等を取り上げた。これらの調査結果については「4.1 章 研究のための既存データシェアリング基盤」に示す。

さらに、本調査では、遺伝子検査等の医療を支えるデータシェアリング基盤として注目を集めている ClinVar と、ClinVar に係る研究プロジェクトである ClinGen を対象に、目的や実施内容、予算、体制、取り扱われるデータについて調査を実施した。これらの調査結果については「4.3 章 医療のための既存データシェアリング基盤」に示す。

これらの調査結果を踏まえ、「⑤ 疾病克服に向けたゲノム医療実現化プロジェクト」に示されているデータシェアリング基盤及び支援機能の要件検討を行った結果を「5. 章 データシェアリング基盤に関する検討」に示す。この章には、国内有識者から得られた情報・意見も整理する。

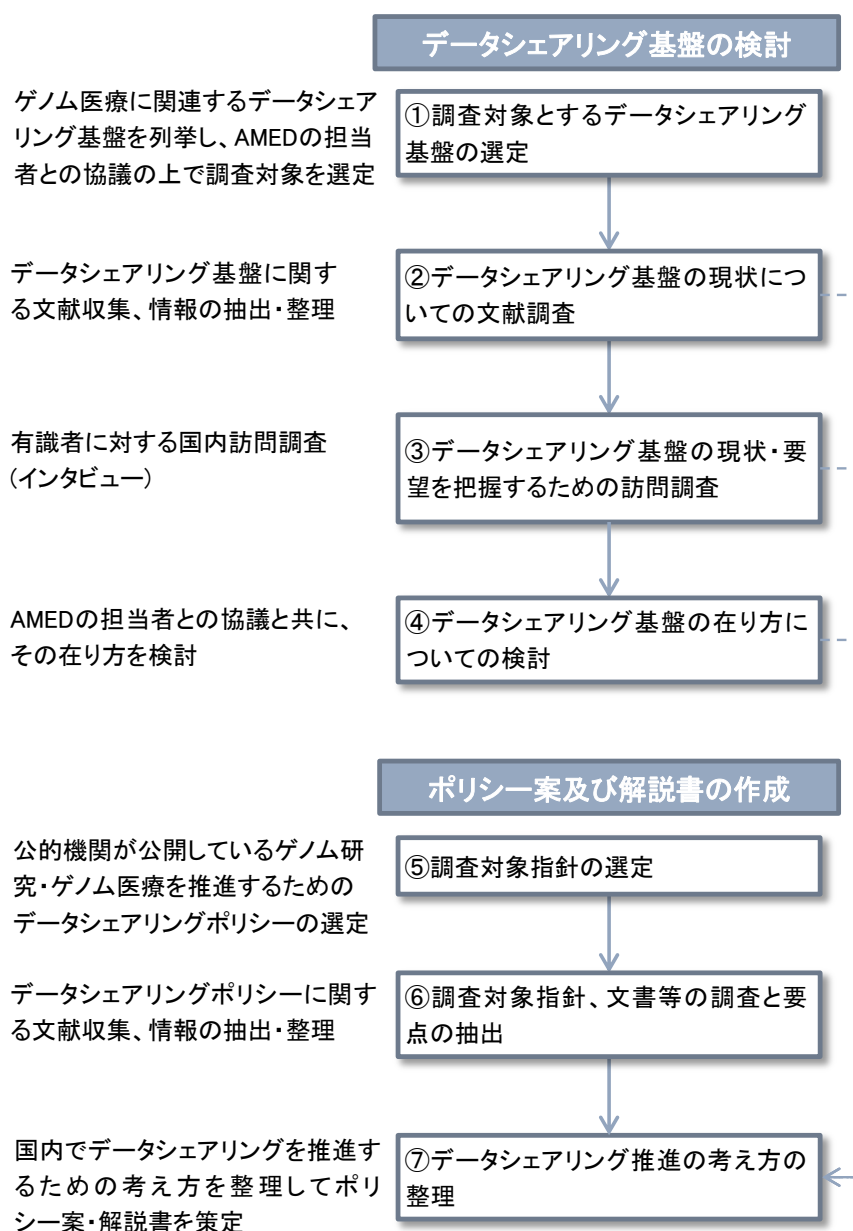
そして、本調査では、AMED によるデータシェアリングポリシーの作成に対しての支援を行った。上述の調査に加え、NIH や Wellcome Trust といった欧米の主要なファンディングエージェンシー等のポリシーやガイドライン等に関する調査を行い、AMED における検討の材料を用意した。

3. 調査の方法

3.1 調査のフロー

本調査の手順を以下に示す（図 3-1）。①～④と⑤～⑨は同時並行的に進行した。欧米のデータシェアリング基盤に関しての調査結果と国内有識者に対するインタビュー結果をもとに、AMED が今後整備していくデータシェアリング基盤について検討し、そのデータシェアリング基盤を念頭に置きつつ AMED によるデータシェアリングポリシー策定の支援を行った。

図 3-1 調査のフロー



3.2 調査対象

(1) データシェアリング基盤に関するインタビュー調査

表 3-1 に、本調査においてインタビュー対象者とした国内有識者を示す。本インタビューでは、わが国のデータシェアリング基盤・データシェアリングポリシーに対する意見、海外のデータシェアリング基盤・データシェアリングポリシーに関する情報を収集した。これらの調査結果を踏まえて、AMED が整備する予定であるデータシェアリング基盤に関する検討（「4. 章 データシェアリング基盤に関する調査」、「5. 章 データシェアリング基盤に関する検討」）及び AMED によるデータシェアリングポリシー策定に対しての支援を行った。

表 3-1 インタビュー対象者（五十音順）

研究機関	所属・役職	名前 (敬称略)
国立国際医療研究センター	研究所 遺伝子診断治療開発研究部 部長	加藤 規弘
慶應義塾大学	医学部 臨床遺伝学センター 教授	小崎 健次郎
東京大学医科学研究所	ヒトゲノム解析センター ゲノム医科学分野 教授	柴田 龍弘
国立循環器病研究センター	創薬オミックス解析センター 室長	高橋 篤
東京大学	大学院医学系研究科 小児科学分野 准教授	滝田 順子
国立国際医療研究センター	研究所 遺伝子診断治療開発研究部 室長	竹内 史比古
東京大学	大学院医学研究系研究科 人類遺伝学教室 教授	徳永 勝士
理化学研究所	統合生命医科学研究センター ゲノムシーケンス解析研究チーム チームリーダー	中川 英刀

(2) データシェアリング基盤に関する文献調査

以下に示すデータシェアリング基盤を調査対象として、それらに係る WEB ページや論文を参照した（括弧付き番号は参考文献に対しての参照番号）。

- 米国 NCBI が構築・運用するデータシェアリング基盤
 - ✓ dbGaP、ClinVar、SRA、Trace Archive、BioSample、BioProject、dbSNP、dbVar、GEO
（これらの位置づけに関しては 4.1 章で整理、dbGaP と ClinVar⁽⁴³⁾⁽⁴⁴⁾については 4.1 に概説）
- 欧州 EMBL-EBI が構築・運用するデータシェアリング基盤
 - ✓ EGA、ENA、EVA、DGVa、ArrayExpress
（これらの位置づけに関しては 4.1 章で整理、EGA については 4.1 章に概説）
- 欧州 Sanger Institute が構築・運用する DECIPHER（4.1 章に概説）⁽³⁹⁾⁽⁴⁰⁾
- 国際的なアライアンスである GA4GH による MatchMaker Exchange（4.1 章に概説）⁽⁴¹⁾
- 英国 Genomics England による GeCIP（4.1 章に概説）⁽⁴²⁾

(3) データシェアリングポリシーに関する文献調査

主として以下の機関が公開しているデータシェアリングポリシーを調査対象として、それらに係る WEB ページや公開されている文書を参照した。

- NIH⁽¹⁾⁽²⁾⁽³⁾⁽⁴⁾⁽⁵⁾⁽⁶⁾⁽⁷⁾
- National Cancer Institute⁽⁸⁾⁽⁹⁾⁽¹⁰⁾
- Wellcom Trust⁽¹¹⁾⁽¹²⁾⁽¹³⁾⁽¹⁴⁾⁽¹⁵⁾⁽¹⁶⁾⁽¹⁷⁾⁽¹⁸⁾⁽¹⁹⁾
- Wellcome Trust Sanger Institute⁽²⁰⁾
- International Cancer Genome Consortium (ICGC)⁽²¹⁾⁽²²⁾⁽²³⁾⁽²⁴⁾⁽²⁵⁾⁽²⁶⁾⁽²⁷⁾

以上に加え、以下に示す文書を調査対象に加えた。

- 第 5 期科学技術基本計画⁽²⁸⁾
- ゲノム医療実現推進協議会「中間とりまとめ」⁽²⁹⁾
- OSTP の指針⁽³⁰⁾⁽³¹⁾⁽³²⁾
- RCUK Common Principles in Data Policy⁽³³⁾
- P-DIRECT データ共有ガイドライン⁽³⁴⁾
- NBDC ヒトデータ共有ガイドライン ver. 2.0⁽³⁵⁾
- インフォームドコンセント参考事例⁽³⁶⁾⁽³⁷⁾

4. データシェアリング基盤に関する調査

本章では、米国及び欧州の主要なデータシェアリング基盤に関して、登録の対象となるデータの種類と各データシェアリング基盤の役割を整理する（4.1 章）。次に、研究に用いられることを主眼においたデータシェアリング基盤を対象としてデータの共有方法について説明する（4.2 章）。加えて、上述の通り、比較的新しい取組である制限共有に関しても、DECIPHER や MatchMaker Exchange、GeCIP の事例について取り上げる。さらに、GDC といった新規情報基盤の構築に関する新しい取組も取り上げる。最後に、遺伝子検査等の医療を支えるためのデータシェアリング基盤である ClinVar とそれに係る研究プロジェクトである ClinGen について説明する（4.3 章）。

4.1 既存データシェアリング基盤の整理

本章では、米国及び欧州におけるゲノム医療関連の主要なデータシェアリング基盤群の役割分担を整理する。ここでは、米国の National Center for Biotechnology Information (NCBI) と欧州の European Bioinformatics Institute (EBI) が構築・運営するデータシェアリング基盤を整理の対象とする。NCBI は、National Institutes of Health (NIH) の下にある United States National Library of Medicine (NLM) の一部門であり、分子生物学に係るデータシェアリング基盤群を構築・運用している。EBI は、European Molecular Biology Laboratory (EMBL) の一部門であり、欧州の研究コミュニティに対して NCBI と同様の役割を果たす存在である。

表 4-1 は、それらのデータシェアリング基盤群を整理したものである。この表の列は、まず、NCBI と EBI とで分かれており、さらに、非制限公開と制限公開との区別で分かれている。行はデータの種類によって分かれている。以下は、ゲノム研究に係るデータとして、ここで取り上げられたものである。

- 次世代シーケンサーにより生成されるデータ（図中の「NGS データ」）
- サンガー法に基づくシーケンサーにより生成されるデータ（図中の「キャピラリーリード」）
- 計測対象に関するデータ（図中の「サンプル」）、研究に関するデータ（図中の「スタディ」）
- 集団中のバリエーションに関するデータ（図中の「バリエーション」）
- 集団中の構造多型に関するデータ（図中の「構造多型」）
- 遺伝子発現やエピジェネティクス等に関するデータ（図中の「遺伝子発現・エピジェネティクス」）
- GWAS 等の個人を特定し得る情報を含むデータ（図中の「GWAS や WES/WGS 等の大規模研究データ、分子診断アッセイの結果、等」）
- 遺伝子診断等の医療に有益な情報を含むバリエーションのデータ（図中の「病的意義のあるバリエーション」）

NCBI のデータシェアリング基盤は、取り扱うデータの種類で分けることができる。GWAS 等のデータを取り扱うものとしては dbGaP が存在しており、非制限公開と制限公開のメカニズムを有している。SRA や GEO は個人を特定し得る情報を含むデータを取り扱うが、それらは dbGaP の機能を用いて制限公開を実現させている。NCBI には、遺伝子診断等の医療

に有益なバリエーション情報を取り扱うデータシェアリング基盤として ClinVar が存在している。EBI のデータシェアリング基盤群は、制限公開のメカニズムを有する EGA と、非制限公開のメカニズムを有するデータシェアリング基盤群（ENA、EVA、DGVa、ArrayExpress）に分けることができる。EGA は、GWAS や網羅的なゲノム解析等で得られる個人識別可能性が生じ得るデータを制限公開するためのデータシェアリング基盤として位置づけられる。EBI では、EVA がそのデータを受け付けているが、EVA から公開することはせず、登録者からの許可を得た上で、ClinVar の登録形式に変換し、ClinVar に登録している。

表 4-1 ゲノム医療に係る既存データシェアリング基盤

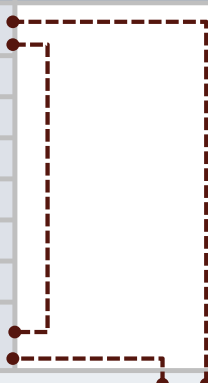
	米国, NCBI		欧州 (EU), EMBL-EBI	
	非制限公開	制限公開	非制限公開	制限公開
NGSリード	SRA		ENA	EGA
キャピラリーリード	Trace Archive			
アノテートされた配列	GenBank			
サンプル	BioSample			
スタディ	BioProject			
多型・バリエーション	dbSNP			
構造多型	dbVar			
遺伝子発現・エピジェネティクス	GEO		EVA	
			DGVa	
		ArrayExpress		
GWASやWES/WGS等の大規模研究データ、分子診断アッセイの結果、等	dbGaP			
病的意義のあるバリエーション	ClinVar			

表 4-2 は、上述のデータシェアリング基盤群のうち、バリエーションに係るものを整理したものである。この表の列は、まず、バリエーションのデータを取り扱うものか、次世代シーケンサー（NGS）により生成されるデータを扱うもので分かれており、前者は、さらに、集団中のバリエーションに関するデータを取り扱うものと個人別のバリエーションに関するデータ（遺伝子型や表現型）を扱うものに分かれている。この表の列は、病的意義があり遺伝子検査等に用いられるものと、そうでないものに分かれている。以下に、この表に含まれているデータシェアリング基盤が取り扱うデータのフォーマットや、特徴を整理する（以下の番号は、図中の番号と対応）。

① 個人を特定し得る情報を含む NGS データを取り扱うデータシェアリング基盤

FASTQ 形式のリードデータ（配列パターンと計測のクオリティ情報を含むデータ）や、リファレンスゲノム配列に対してアライメントされた結果をバイナリに圧縮した BAM 形式のデータを取り扱うデータシェアリング基盤である。dbGaP にも①の印が付けられているが、上述の通り、dbGaP がそれらの格納先というわけではなく、SRA¹ がデータの格納先となっており、dbGaP の制限公開システムが用いられる仕組みとなっている。

¹ 予算的制約のため、SRA は 2011 年に、一旦、運用を終了している。現在は運営されており dbGaP のメカニズムの下に制限公開を行っている。SRA からのデータダウンロードには専用ツール（SRA toolkit）が用いられる。

- ② 個人を特定し得る情報を含むバリエントを取り扱うデータシェアリング基盤
バリエントや表現型のデータを取り扱うデータシェアリング基盤である。データの登録や配布のために様々なフォーマットをサポートしており、そのなかでも VCF 形式（リファレンス配列に対するバリエントの位置や遺伝子型、その他のアノテーションを含むデータ）を主なサポート対象としている。
- ③ 疾患に関わる情報付きのバリエントデータを取り扱うデータシェアリング基盤
遺伝子検査等の医療に寄与する情報が付与されたバリエントのデータを取り扱うデータシェアリング基盤（具体的には“ClinVar”）である。登録される個々の多型には、個人を容易に識別し得るような情報は含まれない。データの登録にはエクセルファイル等の形式を、配布には VCF 等の形式をサポートしている。

表 4-2 バリエントに係る既存データシェアリング基盤

バリエント		NGSデータ
集団中に認められたバリエントを列挙 (個人別には集積せず)	個人の持つ バリエント(複数)を 表現型とともに 個人別に集積	
バリエントを集積 (病的意義は不問)		
病的意義のあるバリエントのみを集積		

制限公開の
区分あり

4.2 研究のための既存データシェアリング基盤

本章では、米国と欧州における「研究のために用いられるデータシェアリング基盤」のうち、制限公開のメカニズムを有するものと、制限共有に類似するメカニズムを有するものについて説明する。制限公開に関しては、NCBI による dbGaP と EBI による EGA に着目し、データ登録手順と制限公開データの提供に際して求められる手続きについて説明する。制限共有に関しては、Sanger Institute による DECIPHER、GA4GH による MatchMaker Exchange、Genomics England による GeCIP に着目し、その共有の手続きについて説明する。また、デー

タシェアリングに関連して、米国の National Cancer Institute (NCI) では、最近、大量のデータを蓄積・解析する環境としてクラウドを構築し、研究者に対して提供する試みを進めている。本章の最後に、その試みの一つである Genomic Data Commons (GDC) について説明する。

(1) dbGaP (NCBI) 及び EGA (EBI)

dbGaP と EGA は、閲覧許可が与えられた者のみがデータを閲覧することができる「制限公開」という仕組みを持っている。データシェアリング基盤に格納されているデータへの利用要請を、データシェアリング基盤に紐づく Data Access Committee (DAC) が審査し、利用要請に対する審査に通った者がデータを使用できる、という流れがこのメカニズムの骨格である。

本章では、まず、これらのデータシェアリング基盤に対するデータ登録について説明し、次に、制限公開データ提供の手続きについて説明する。

1) データ登録

図 3-1 は、dbGaP のデータ登録手順を説明した図である。以下に、その手順を説明する。

① NIH へのコンタクト

NIH のファンディングを受けた研究か否かで、登録に関わる条件が異なる。NIH のファンディングを受けていない研究からの登録に関しては、その科学的価値と、登録されるデータを利用可能とするために必要な計算資源をもとに、その登録を判断する。以下にその判断基準を示す。

- ✓ 科学的価値：NIH 傘下研究機関にとっての研究テーマの重要性、論文発表計画、品質、量
- ✓ 登録されるデータのサイズ：登録されるデータ (NGS リード) が 1 テラバイトより大きい場合、データ登録者と NLM との間で、そのコストをカバーするための調整が必要となる (データ登録者が、NLM への贈与として、そのコストを提供可能としている一つまり、データ登録者がコストを賄えば、そのデータ登録が許されることとなっている)

② 研究情報の登録

データ登録者が、研究に関しての情報を登録する。

③ データの登録

データ登録者は、メタデータ (Study、Samples 等) や表現型データ、表現型と同意に関するデータ²を作成する必要がある。データ登録者は、それらのデータに加え、アライメントデータ (BAM) やバリエーションデータ (VCF) 等を、dbGaP のデータ登録ポータル³を用いてそれらのデータを登録する⁴。

² 研究参加者ごとに、参加した研究とは別の研究でのデータ利用に関する同意が得られているか否かの情報を含むデータ。

³ データ登録ポータルの URL は、データ登録に先立って、dbGaP 側からデータ登録者に送付される。

⁴ データ登録手順の詳細や、登録を受け付けるデータフォーマット、それらに関するよく見受けられる間

④ 提供開始

dbGaP がアクセッション番号の付与とプレビューサイトの構築を行い、データ登録者がプレビューサイトをチェックして問題がなければ提供開始となる。データ提供にあたっては、データ登録から提供開始までに少なくとも 8 週間を要する（データに問題が見つかった場合には、さらに時間を要する）⁴。

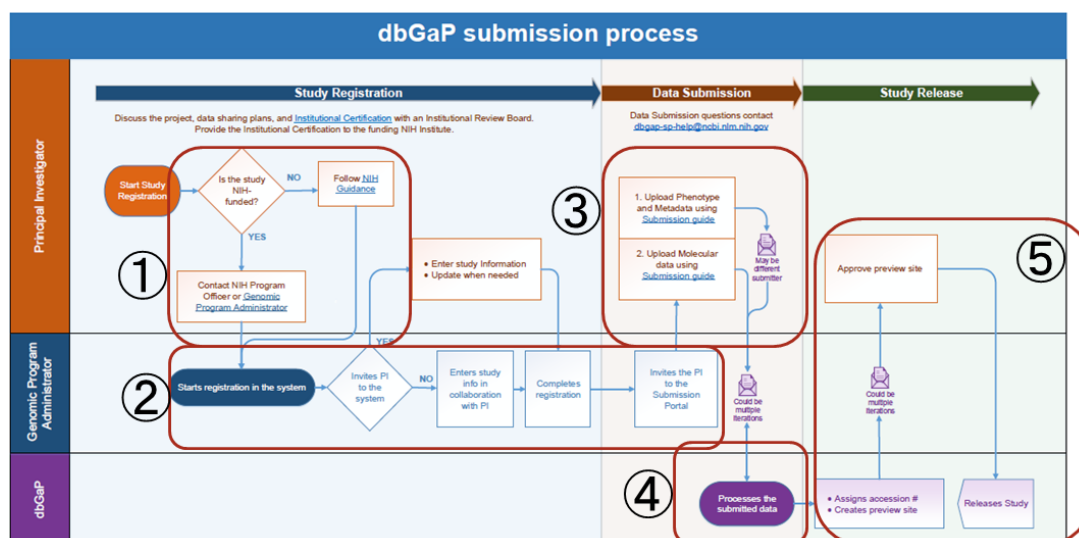


図 4-1 dbGaP のデータ登録手順

(出所 : http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/GetPdf.cgi?document_name=HowToSubmit.pdf)

EGA のデータ登録手順は以下の通りである。

① EGA の仕様に準拠したファイルの作成

配列データ、アレイデータ、表現型データを作成し、EGA が提供するツールである EgaCryptor によって暗号化する。配列データに関しては、CRAM や BAM、FASTQ といった標準フォーマットのほか、SFF 等のシーケンサーが出力するフォーマットを受け付ける仕様となっており、登録に推奨されるフォーマットも指定されている⁵。表現型データに関しては、特にフォーマットが指定されていないが、Experimental Factor Ontologies⁶を用いることが推奨されている⁷。さらに、データが破損していないことや改ざんされていないことを確認するために、上述の圧縮されたデータ及び非圧縮のデータに対しての MD5 チェックサム⁸を作成する。

② EGA のデータアップロード

FTP もしくは Aspera⁹を用いて、EGA にデータをアップロードする。

違い等に関しては、以下のファイル群に示されている。

(http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/GetZip.cgi?zip_name=dbGaP_SubmissionPackage.zip)

⁵ <https://www.ebi.ac.uk/ega/submission/sequence> を参照

⁶ <https://www.ebi.ac.uk/efo/>

⁷ <https://www.ebi.ac.uk/ega/submission/phenotypes>

⁸ 入力データに対して作成された 128 ビットのハッシュ値。

⁹ IBM 傘下の Aspera 社が提供する高速ファイル転送ツール。

③ メタデータの登録

Study、Samples、Data Access Committee (DAC)、Data Access Agreement (DAA) や登録されるデータセットに関するメタデータを登録する。登録されるデータの種類によっては、Analysis (BAM や VCF の作成といった解析に関するメタデータ) や Experiments (計測機器やライブラリに関するメタデータ) の記載内容が異なる。後述する通り、EGA では、データ登録者に紐づく形で DAC を形成することとなる。その DAC に関する情報をこの時点で登録する。これらの登録には、WEB ページをユーザインタフェースとする EGA Webin を用いた方法と、ユーザーが作成したプログラム等による方法¹⁰が用意されている。

2) 制限公開データ提供の手続き

図 4-2 は、制限公開データを利用するに際しての基本的な手順を示したものである。利用者は、利用したいデータがどのデータであるかを調べ、必要なデータが見つかった場合には、利用目的を明示した利用申請を DAC に申請し、申請が承認された後にデータを利用することが可能となる。

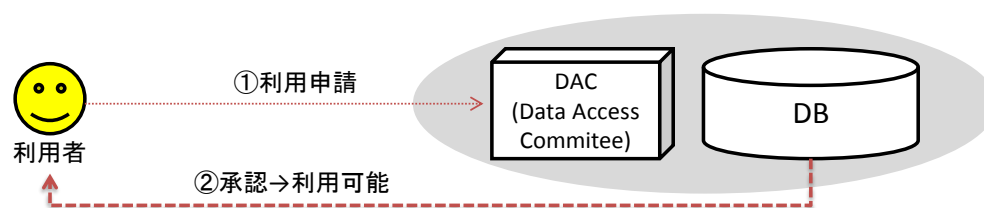


図 4-2 制限公開データを利用するための基本手順

dbGaP と EGA が用いている制限公開メカニズムは、上述の通り、基本骨格は同じであるが、DAC の形態や運用方法が異なっている (図 4-3)。

dbGaP の形態や構築方法、運用方法は以下の通りである。

【DAC の形態】

- 上級連邦職員 (Senior federal employees) で構成される
- 各メンバーは、科学的専門知識、倫理的専門知識、ヒトを研究対象とする研究における専門知識を有している者である

【DAC の構築方法】

- メンバーの任期は、NIH 内の招集した機関 (つまり、DAC が設置される機関¹¹) によって決定することができる
- 特定の専門性を持つコンサルタントを、DAC に関わるミーティングに参加させたり、DAC の顧問として招集したりすることが可能である (コンサルタントは連邦職員である必要はないが、DAC のメンバー選定に関与してはならない)

¹⁰ HTTP プロトコルを用いて XML データをサーバと受送信する方法。

¹¹ DAC の一覧が示されている。複数の DAC が存在する。 (https://gds.nih.gov/pdf/NIH_DACs_Chairs.pdf)

【DAC の運用】

- DAC は、GDS Policy に示されている仕様及び研究プログラムからの独自要請に基づいて、dbGaP に格納されているデータへのアクセスを審査し、承認・否認する
- 利用申請には、データ登録機関によって禁止されている利用方法を含めてはいけない
- 利用申請は、公開されているウェブページで閲覧可能でなければならない
- 利用者は、DAC に利用申請を提出し（図中の①）、承認された場合にデータが利用可能となる（図中の②）
- 審査を担当する DAC は、データの登録に関わった組織の DAC である
- dbGaP へのアクセス申請に対しての応答時間には幅があり得るが、全ての DAC はできるだけタイムリーに応答をするように務めるべきとされている

EGA の形態や構築方法、運用方法は以下の通りである。

【DAC の形態】

- 一般的に、DAC は、同じ組織に属する 1 人以上から構成される
- DAC の構成メンバーは、利用者がデータへアクセスする事に係る決定に責任を有する者である
- 複数のデータが、1 つの DAC の配下に存在し得る

【DAC の構築方法】

- DAC を構築するためには、EGA へのデータ登録過程において、詳細を登録する必要がある（DAC 名、構成員名、データセットについて情報）
- DAC 構築の際には、以下に示す書類の提出が必要である
 - ✓ データ利用申請の入力フォーム（Data Access Application Form）：利用者が、利用を希望するデータセット、利用目的、実行可能性等を示すためのフォーム
 - ✓ データ利用同意書（Data Access Agreement：DAA）：データセットをどのように保管するかや、そのデータセットと用いた研究成果をいつから論文文化していいか（Embargo）といったデータ利用に係る諸条件
 - ✓ アクセスポリシー関連書類（EGA DAC Access Policy Document）：管理者ツール（EGA DAC admin tools）を用いて、アカウント作成や権限設定をする方法を示したドキュメント

【DAC の運用方法】

- EGA DAC admin tools を用いて、EGA における新しいアカウントの作成やアクセス権限の設定が可能である
- 利用者は、まず、データ生産者が所属する組織（コンソーシアム等を含む）が保有する DAC に利用申請を提出し（図中の①）、承認された場合に利用可能となる（図中の②）

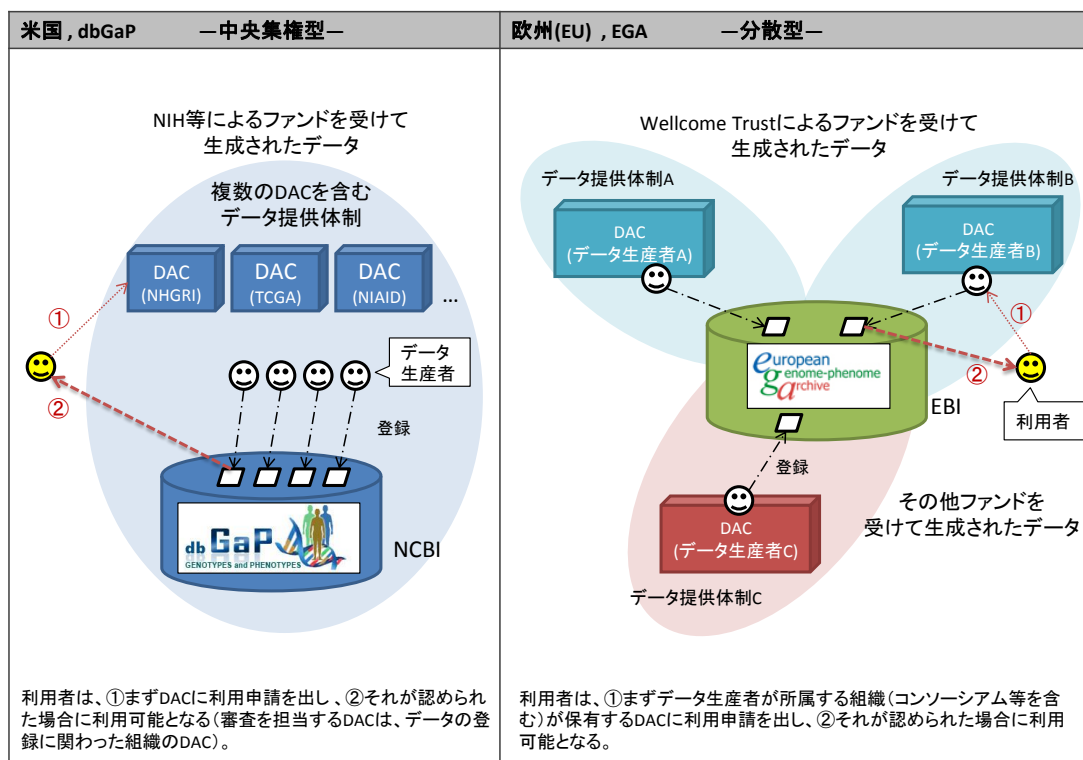


図 4-3 dbGaP と EGA のデータ公開手順の比較

(2) DECIPHER (Sanger Institute)

DECIPHER¹²は、研究参加者（患者等）が持っているバリエーションの遺伝子型と表現型を蓄積し、解析結果の可視化や、類似する遺伝子型・表現型を有する他者の同定、疾患に影響をおよぼす可能性がある遺伝子の探索といった機能を提供するデータシェアリング基盤である⁽⁴⁰⁾。DECIPHER に登録されている研究参加者のデータは、51,500 人分に及んでいる（40 か国の臨床遺伝学施設からの登録を受けている）¹³⁽³⁹⁾。そのうち、17,000 人以上のデータを同意に基づいて匿名で一般公開しており、その他の 18,000 人分のデータを DECIPHER の研究共同体において共有（非公開）している。

¹² <https://decipher.sanger.ac.uk/>

¹³ これらは、Chatzimichali EA et al. Hum Mutat. 2015 Oct;36(10):941-9.執筆時の数である。

DECIPHER におけるデータシェアリングのメカニズムは図 4-4 の通りである。そのルールを以下に示す。

- 研究協力者が公開を明記した同意書に同意した場合に、匿名化されたデータは一般公開される可能性がある
- データ提供者は、研究協力者の一つあるいは複数のバリエーションについて共有状態を設定可能である

Private : 公開・共有されない（データ提供者のみが閲覧可能）
 Managed : 一つあるいは複数の研究共同体での共有（データ登録者同士で共有を契約；メンバー以外は閲覧不可）
 Public : 非制限公開

上述の通り、DECIPHER は、遺伝子型・表現型が類似する研究参加者を同定する機能を用意していることや、データを登録した研究者同士で適宜その共有を行えることから、研究目的が合致する限られた範囲でのデータ共有を促進するインフラとして捉えられることができる。

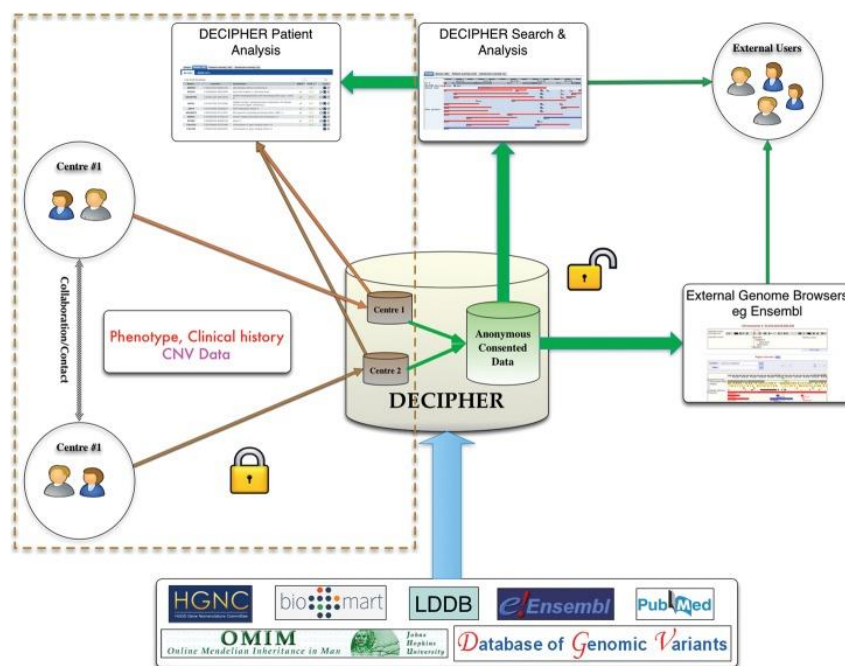


図 4-4 DECIPHER におけるデータシェアリングのメカニズム
 （出所：Swaminathan GJ et al. Hum Mol Genet. 2012 Oct 15;21(R1):R37-44. Epub 2012 Sep 8.）

(3) MatchMaker Exchange (GA4GH)

Matchmaker Exchange¹⁴は、それぞれのデータシェアリング基盤を超えて、研究者同士がデータを共有する仕組みであり、現状、12 のデータシェアリング基盤が参加している（図 4-5：上述の DECIPHER も含まれる）¹⁵。その共有のメカニズムは、最終的には研究者間で連絡を取り合った上で共有が成立するというものであり、研究目的が合致するコミュニティ内でのデータ共有という点で DECIPHER と類似する考え方となっている。このメカニズムにおける手続きの流れは以下の通りである（図 4-6）⁽⁴¹⁾。

- (1) ある研究者がデータシェアリング基盤に研究協力者のデータを登録する（図中の 1 及び 2）
- (2) 別のデータシェアリング基盤に登録されているデータに対して検索をかける（図中の 3a）
- (3) 類似するものがあった場合に双方のデータ登録者に通知する（図中の 3b、4）
- (4) 双方が連絡を取り合う

この 3a と 3b の通信を標準化しプログラムで自動的に処理を行うために API（Application Programming Interface）の開発や、研究参加者同士の類似性を計算するためにアルゴリズムの開発が進められている¹⁶。

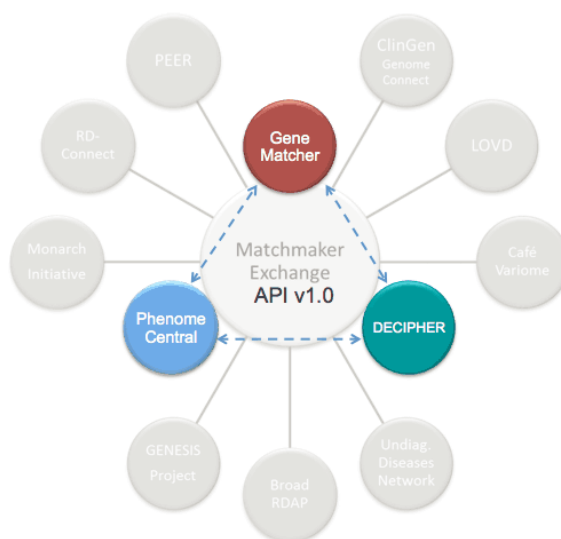


図 4-5 MatchMaker Exchange に参加するデータシェアリング基盤
(出所：http://www.matchmakerexchange.org/i_am_a_clinician_laboratory.html)

¹⁴ <http://www.matchmakerexchange.org/>

¹⁵ 現状、12 のデータシェアリング基盤が参加しているが、その中でこの共有の仕組みで結ばれているのは、DECIPHER (<https://decipher.sanger.ac.uk/>)、GeneMatcher (<https://genematcher.org/>)、PhenomeCentral (<https://phenomecentral.org/>) の 3 者間のみである。

¹⁶ 遺伝子型もしくは表現型、あるいはそれらの両方を入力データとして研究参加者の類似度を評価する。表現型や遺伝子型のマッチングには、OMIM、Orphanet、HPO、HGNC の識別子や用語、オントロジーが利用される。

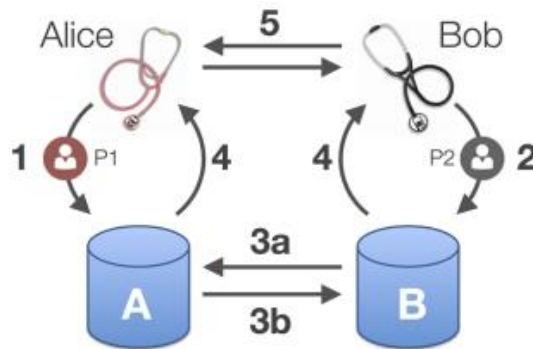


図 4-6 MatchMaker Exchange における共有までの手順

(出所：Buske OJ et al. Hum Mutat. 2015 Oct;36(10):922-7. doi:10.1002/humu.22850.)

(4) GeCIP (Genomics England)

Genomics England では、100,000 Genomes Project で得られたデータを解釈し、医療に活用するための活動として、Genomics England Clinical Interpretation Partnership (GeCIP) を進めている⁽⁴²⁾。GeCIP では、研究者、医療従事者、研修者を構成員として、研究テーマごとにドメインを形成し、そのメンバー内で優先的にデータを共有している¹⁷ (図 4-7)。現時点において、10 種類のがんに関するドメイン群、13 種類の希少疾患に関するドメイン群、教育・トレーニング、電子的な記録、希少疾患に関わる国際協力、倫理学・社会学、医療経済学、機械学習等、疾患の種類に関わらないドメイン群が存在している¹⁸。

【GeCIP の目的】

- 研究・医療におけるフローの最適化
 - ✓ 臨床データ及びサンプルの収集
 - ✓ 臨床医や患者へ返却するためのレポートやデータの解釈
- 臨床現場におけるゲノム医療の改善や理解に繋がる研究
- トレーニング環境の充実

【研究ドメインの構築方法】

- 研究者、医療従事者、研修者が研究ドメインを構成する（研究者、医療従事者はサンプル、データ、スキル等を提供する）
- その構成メンバーが、特定の疾患あるいは特定の課題を取り扱う「ドメイン」を Genomics England に提案する（希少疾患、がん、感染症、バリエーションの解釈等についてのドメインが存在する）
- 研究ドメインは、リーダー（UK を拠点とする者）を推薦し、そのリーダーが NHS と契約を交わす
- 英国の組織だけでなく、海外の組織からの参加を認めている（ただし「重要な研究を行う組織」というただし書きが付いている）
- 民間からの参加を認めていないが、後述の通り、GeCIP とのコンソーシアム形成は推

¹⁷ <http://www.genomicsengland.co.uk/about-gecip/gecip-domains/>

¹⁸ 2016 年 3 月の状況 (<http://www.genomicsengland.co.uk/about-gecip/gecip-domains/>)

奨めている

【ドメイン内における優先的なデータ共有】

- あるドメインは、そのドメイン内における 100,000 Genomes Project のデータに対して、6 ヶ月間は優先的にアクセスすることができる
- ドメイン内のデータ利用に係る研究実施・成果出版については費用を支払う必要はない
- UK 外の研究者及びドメインのメンバーではない研究者は、データのアクセス申請を出すことはできるが、6 ヶ月間はデータ利用を待つ必要があり、さらに使用料を支払う必要がある

【優先的データ共有以外の活動内容】¹⁹

- 研究テーマとして取り扱われるべき希少疾患を提案できる
- 希少疾患として認定される基準の変更を提案できる
- 研究テーマの登録を行うことができる(希少疾患に関するドメインと研究テーマ横断的ドメインのメンバーのみ)
- ドメインが申請しているグラントの共有する
- WEB アプリケーションである PanelApp²⁰を用いた、希少疾患を診断するための遺伝子パネルのキュレーションと共有を行うことができる

【民間とのパートナーシップ】

- GeCIP の目的に合致した場合、前競争的 (Pre-competitive) な状況において民間とのコンソーシアムを形成できる
- 民間も知財のライセンスを受けることができる

¹⁹ <http://www.genomicsengland.co.uk/about-gecip/for-gecip-members/>

²⁰ Randboud University Medical Center, Illumina TruGenome Predisposition Screen、Emory Genetics Laboratory、UKGTN のリソースからパネルに含まれる遺伝子の候補を列挙し、それらに対してキュレーションを行う。GeCIP は、ドメイン外部の専門家にもこの活動に参加することを希望している。

(<http://www.genomicsengland.co.uk/panelapps-new-look/>)

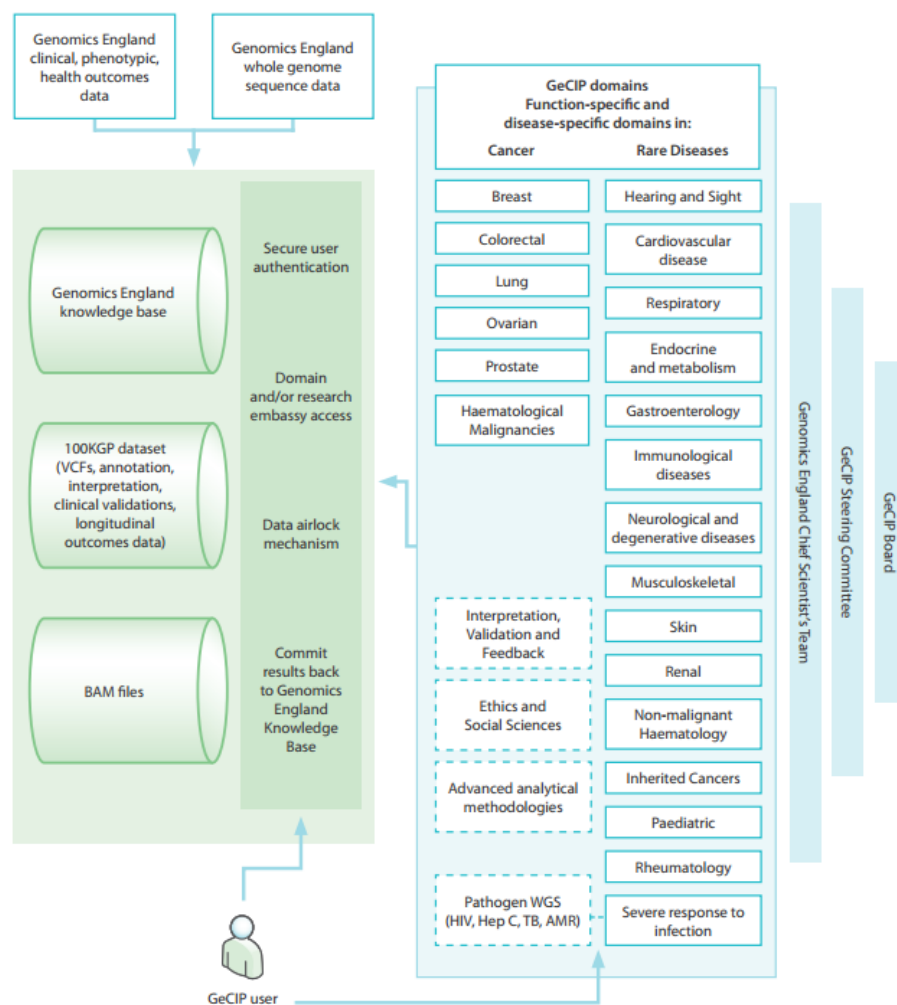


図 4-7 GeCIP のドメインと共有されるデータ

(出所： http://www.genomicsengland.co.uk/wp-content/uploads/2014/11/GeCIP-Guidance_FINAL.pdf)

(5) GDC (NCI)

米国の National Cancer Institute (NCI) は、米国におけるがん研究を主導する NIH の下部組織であり、同じく NIH の下部組織である National Human Genome Research Institute (NHGRI) と共同で、ゲノム解析技術によって、がんに対する包括的な理解を促進する大規模プロジェクトである The Cancer Genome Atlas (TCGA) 等、大規模なゲノム解析プロジェクトを進めている。プロジェクトで生成されるデータが大規模化するに伴い、個々の研究者がデータ解析のための計算資源を自前で確保することが困難になってきたことや、データ転送にも大きな時間がかかることから、それらを克服するための方策としてクラウドコンピューティング（以下、「クラウド」という。）の活用を検討している。そもそも、TCGA の開始後、すぐに、dbGaP では大規模なデータを取り扱うことが困難であることが判明し²¹、University of California, Santa Cruz が TCGA のためのデータシェアリング基盤である Cancer Genomics Hub (CGHuB) を構築し、そこにデータが移されてデータ共有が進められてきたという経緯が

²¹ プロジェクト完了の 2016 年までに 2.5PB のデータが生成される予定。

ある。このような経緯から、NCI が検討を進めているクラウドでは、TCGA のデータのような大規模なデータを取り扱えるものとなっている。NCI は、クラウドに関する取組として Cancer Genomics Cloud Pilot (CGC) と Genomics Data Commons (GDC) を打ち出している²²。前者にはパイロット的な側面もあり、計算能力やデータ蓄積量に対する高い要求に対する評価等が主目的となっている²³。後者は、がん研究コミュニティに対して、TCGA 等の大規模データを活用することや、自らが保有するデータを解析することができる環境を提供するものである²⁴。本章では GDC について概説する。

GDC は、University of Chicago を中心に開発中のクラウドであり、2016 年 7～9 月に本サービス開始する予定となっている。GDC は、TCGA をはじめとする NCI のプロジェクトによって生成されたデータやユーザーが保有するデータを利活用する環境を提供する（図 4-8）。以下は、GDC が提供する機能である。

- TCGA 等、大規模データの蓄積（標準化された形式で蓄積）
- ユーザーがデータにアクセスする機能の提供
- API の提供（データへのアクセスをプログラムで自動化するため）
- データ登録機能の提供
- 登録されているデータや解析結果を整理・可視化する機能の提供
- ID 管理の機能
- セキュリティの確保
- 外部のデータシェアリング基盤（dbGaP 等）からのデータ取り込み機能
- 基本的なデータ解析機能（バリエーションの同定や遺伝子発現の定量）

²² <https://cbiit.nci.nih.gov/ncip/nci-cancer-genomics-cloud-pilots>

²³ クラウドの活用を検討するためのパイロットプロジェクトであり、以下の 3 機関が別個にクラウドを構築し、それぞれが解析ツールの開発や API の開発、データ規模増大（10～100 倍）に伴う設備投資・運用コストの評価、セキュリティ検討等を実施する。CGC の構想は 2013 年発表され、2014 年に以下に示す機関において、その構築・運用が決定、2016 年から外部の研究者にサービスを開始する予定である。

- Broad Institute（Google Cloud Platform を活用）
- Institute for Systems Biology（Google Cloud Platform を活用）
- Seven Bridges Genomics（Amazon Web Services を活用）

CGC はパイロット的な位置づけに終わらず、GDC と連携して、がん研究コミュニティに更なる高い計算能力を提供するために、サービスを一般公開する予定となっている。

²⁴ 将来的には、がん以外の研究に対して利用範囲を広げる予定となっている。

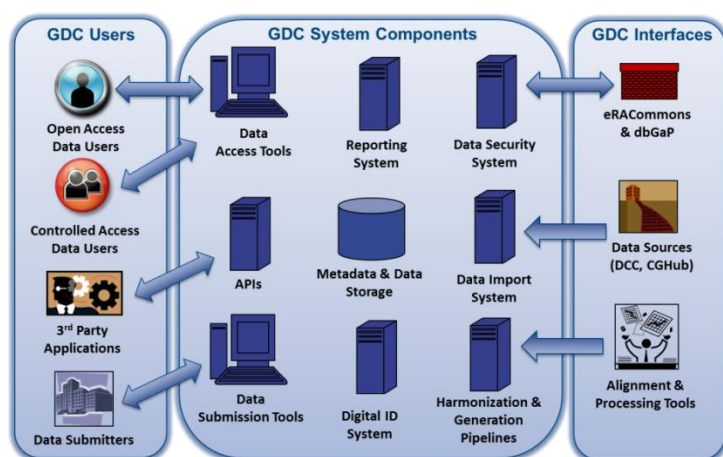


図 4-8 GDC が提供するサービス
(出所：GDC 開発グループからの提供資料)

ユーザーは、WEB ページ上のポータルサイト（図 4-9）からデータの検索やダウンロード、アップロードが可能である。ユーザーが GDC にアップロードしたデータの統計情報（臨床データ、サンプルに関するデータ、配列等の分子生物学関連データや、解析の進捗状況）は、そのポータルサイトから確認できるようになっている。

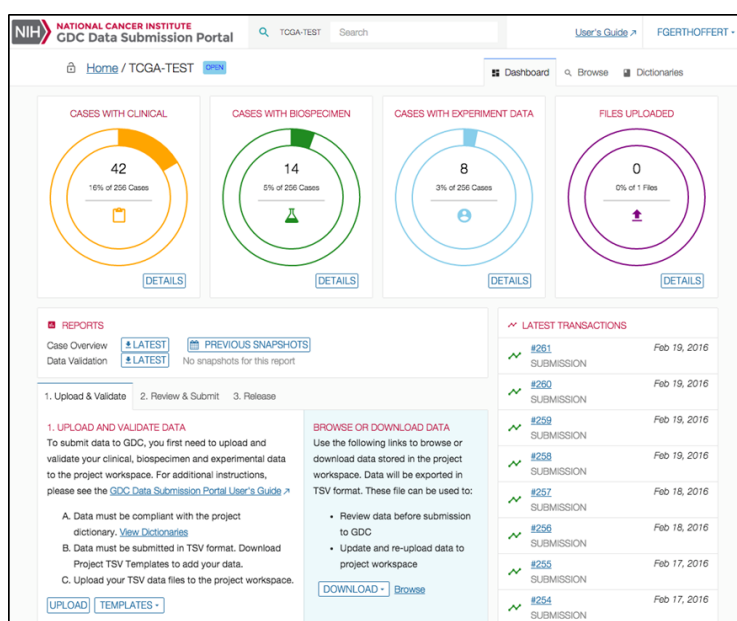


図 4-9 GDC のデータ登録ポータルサイト
(出所：GDC 開発グループからの提供資料)

また、UNIX 系の計算機リテラシーを有するユーザーのために、コマンドラインインターフェース²⁵も用意されている。上述のポータルサイトと同様に、データのアップロードやダウンロードを行うことができる。

²⁵ UNIX 系 OS で一般的に用いられているターミナルプログラムを用いたユーザインタフェースである。コマンドを入力することで OS や OS 上で稼働するプログラムを操作するものである。

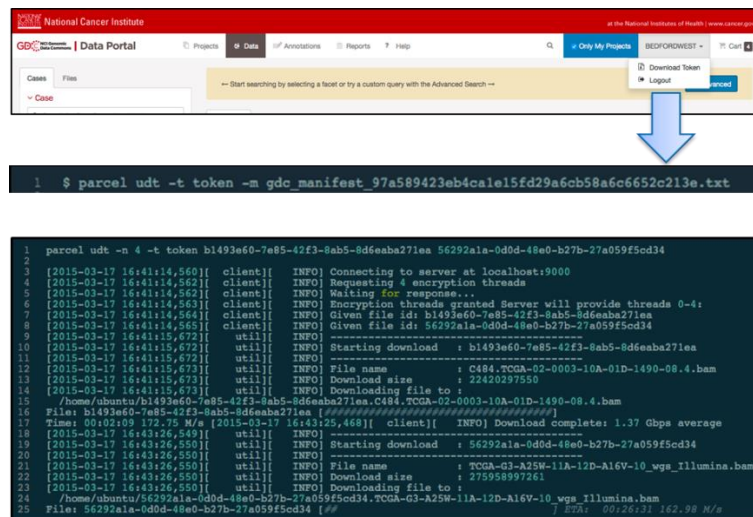


図 4-10 GDC におけるデータ転送のコマンドラインインターフェース
(出所：GDC 開発グループからの提供資料)

上述の通り、GDC は、バリエーションの同定や遺伝子発現の定量といった基本的な解析機能をユーザーに対して提供する。この機能は、一連の解析手順をパイプライン化したものとなっており、BAM からのリードの抽出、リファレンスゲノムへの再マッピングといった機能も含まれている（図 4-10）。バリエーションの同定に関しては、GATK による生殖細胞系列向けの解析パイプラインに加え、Baylor College of Medicine、University of Washington が開発する体細胞向けの解析パイプラインを提供する（図 4-11 に示したパイプラインは、Baylor College of Medicine による 1 本と University of Washington による 2 本の合計 3 本）。

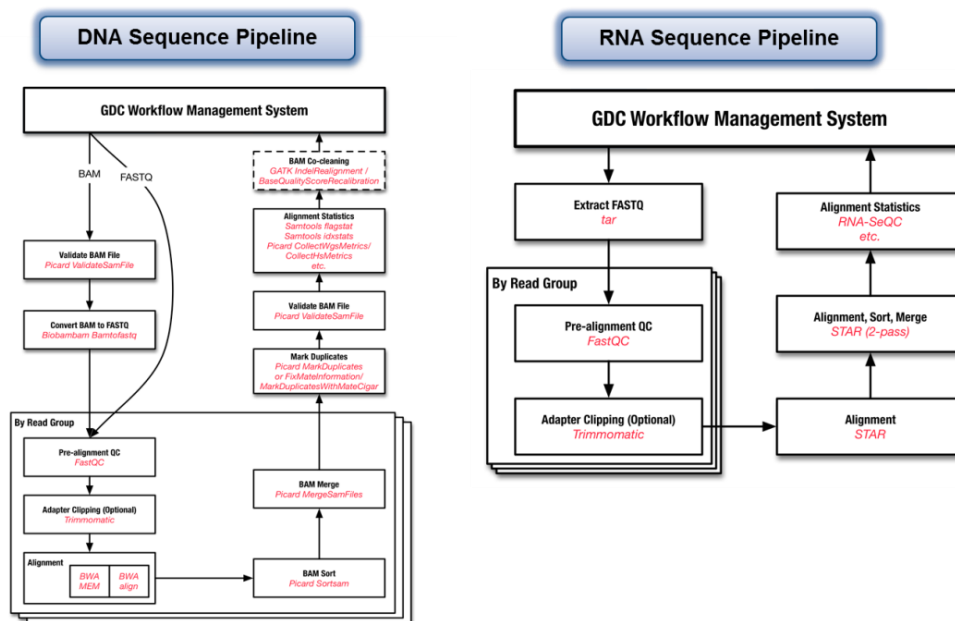


図 4-11 バリアント同定パイプライン及び遺伝子発現量同定パイプライン
(出所：GDC 開発グループからの提供資料)

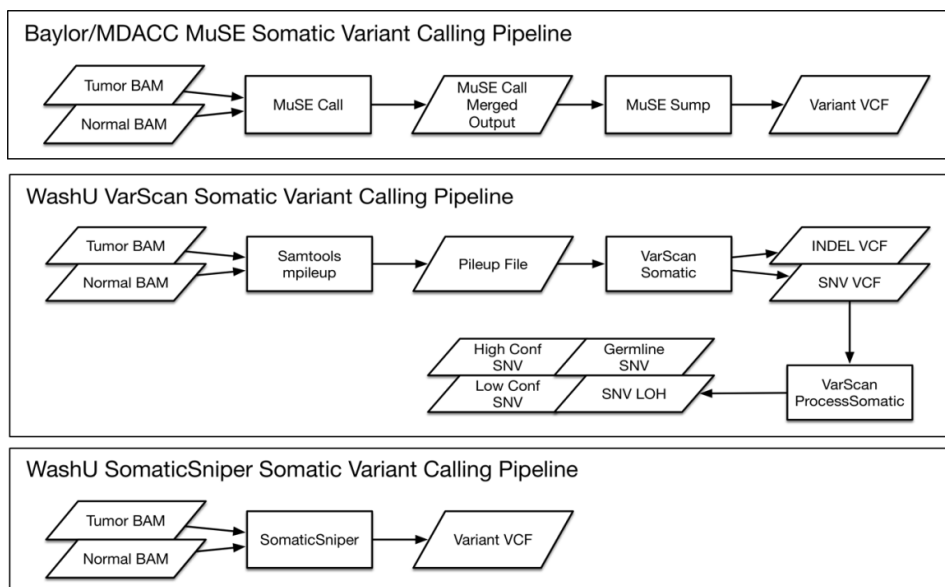


図 4-12 体細胞向けのバリアント同定パイプライン
(出所：GDC 開発グループからの提供資料)

4.3 医療のための既存データシェアリング基盤（ClinVar）

本章では、遺伝子検査等の医療応用を目的としたデータシェアリング基盤である ClinVar について説明する（ClinVar に係る研究プロジェクトである ClinGen については 4.4 章で説明する）。本章の内容は以下の通りである。

- ClinVar の目的と実施内容の概要（「(1) 目的・概要」）
- ClinVar の運用開始年次と現在の運用体制（「(2) ClinVar の運用」）
- WEB ページ及び FTP サイトからのデータ提供方法（「(3) データ提供方法」）
- 登録されたデータの臨床的な重要性（「(4) 臨床的な重要性」）
- 登録されたデータの信憑性（「(5) レビューステータス」）
- 登録されたデータ数と大きさ（「(6) データ登録数・サイズ」）
- データの登録を許されている機関及び実際に登録した具体的な機関名（「(7) データ登録者」）
- データの登録方法（「(8) データの登録方法」）

(1) 目的・概要

ClinVar は、医療において重要なバリエーションと表現型との関係性を取り扱うデータシェアリング基盤である。ClinVar が取り扱うデータ、データシェアリングの仕組み、その仕組みの狙いを以下に整理する。

- ClinVar が取り扱うバリエーションは、具体的には、遺伝子バリエーションがほぼ症状を決定する遺伝子に関するバリエーションである。
- ClinVar は、他のデータベースから抽出されたデータや、研究室や検査会社から登録されたデータを蓄積し、それらを非制限的に公開することで、医療等の場面で広く活用できるようにしている⁽⁴³⁾（図 4-13）。
- より多くの情報源から情報を集積・統合することにより、バリエーションと表現型に関する解釈の一致・不一致を把握できるようにし、さらに、その一致・不一致やエキスパートパネル等によるキュレーションによってバリエーションと表現型との関係性に信頼度を付与できるようにしている。
- バリエーションには、診断等に有用である必要最小限の情報が付与されているのみである。付与される情報は、表現型、表現型の機能的・医学的な重要性に対する解釈、解釈に対する評価方法・基準、バリエーションの同定手法、これらに関するエビデンス等となっている。
- 利用者は、WEB ページや FTP サイトから簡単にデータにアクセスすることができる。
- データシェアリングの観点から、データ登録者と利用者との間のコミュニケーションを促進し、データ登録者の貢献をしっかりと評価するために、データ登録者毎のデータ登録数・コンタクト先を WEB ページから公開している。

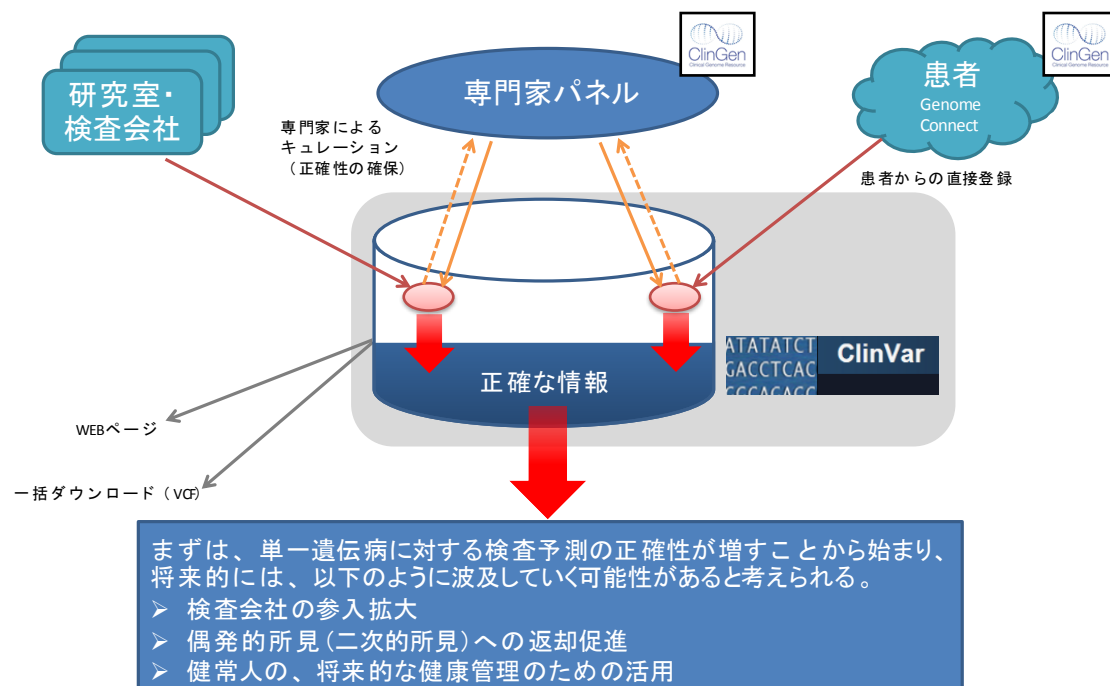


図 4-13 ClinVar の概要
(ロゴの出所：<http://www.ncbi.nlm.nih.gov/clinvar/>)

上述の通り、ClinVar は、医療・研究コミュニティにおけるデータシェアリング基盤であり、インターネットを介してデータ登録者とデータ利用者がデータを共有する仕組みとなっている。表 4-3 に、データの利用・登録に関して要点を整理する。

表 4-3 ClinVar におけるデータ利用・登録の要点
対象者・使い方・工夫・狙い

表 4-3 ClinVar におけるデータ利用・登録の要点 対象者・使い方・工夫・狙い	
データ利用	● 利用者は、遺伝子検査を行っている医療関係者や、ゲノミクス・遺伝学に関わっている研究者 ● 利用者は、ClinVar を遺伝子検査等におけるワークフローやアプリケーションに組み込むことが可能 ● ClinVar が大量のデータを蓄え、データの正確性が向上することによって、将来的には以下のような道が開ける可能性がある (図 4-13) ✓ 検査会社が遺伝子検査市場に参入する可能性の拡大 ✓ 偶発的所見 (二次的所見) への対応可能性の向上 ✓ 健常人の健康管理 (将来的に罹る可能性がある病気に備えた健康管理等) (これらの詳細については、(4) 臨床的な重要性」、「(5) レビューステータス」を参照)

	対象者・使い方・工夫・狙い
データ登録	● ClinVar 外部からのデータ登録者は研究室や臨床検査会社（表 4-14） ● 大規模にデータ登録を行うために、エクセルやフラットファイルにデータを記載してアップロードすることによってデータを登録する仕組みと、小規模にデータ登録を行うために、ポータルサイトにデータ登録用のページを提供 ✓ データ登録用のページでは、データ登録者がウィザードを用いながら簡単にデータを登録 ✓ エクセルやフラットファイルでのデータ登録にはデータを格納するためのテンプレートファイル提供（表 4-16） ● テンプレートファイルには、詳細な情報を格納するための完全版と、最小限の情報だけを登録するためのライト版が存在 ✓ 完全版には個人毎のデータを含めることが可能 ✓ 最小版には個人を区別できない集約されたデータ（頻度情報等）を含めることが可能 ✓ どちらのテンプレートファイルを用いるかはデータ登録者次第であり、個人毎のデータまでを登録したくない場合にはライト版を用いればよい ● ClinVar 内部（NCBI 内に設置された ClinVar チームのメンバー）からも、OMIM 等のデータベースから抽出したデータを蓄積（これらの詳細については「(6) データ登録数・サイズ」、「(7) データ登録者」を参照）

(2) ClinVar の運用

ClinVar のプロトタイプは 2012 年 11 月に運用を開始し、2013 年 4 月に正式運用が始まっている。その開発・運用の資金は、NLM の所内研究プログラム (Intramural Research Program) からのものであり、NCBI 内に設置された ClinVar チームにより担われている。研究室や検査会社から登録されるデータに対してのキュレーションに ClinVar チームは関わっているものの、データ登録者による解釈の部分に対しては関与しないという線引きがなされている（この点に関しては、「(8) データの登録方法」において再述する）。現時点、このキュレーションには 4 名が関与しているとのことである²⁶。

²⁶ Melissa J. Landrum 博士からの情報（Melissa J. Landrum 博士は、ClinVar に関する論文の執筆者⁽⁵⁰⁾）。この他に 13 名のソフトウェア開発者が関わっているとのこと。ただし、ClinVar に何名が関わっているかということに関しては正確な数を明言するのは難しいとのこと（関係者は、dbSNP、dbVar、dbGaP の開発・運営を掛け持ちしているため）。

(3) データ提供方法

ClinVar は、以下に示す 2 通りの方法で利用者にデータを提供する。

- 『WEB ページからの提供』

利用者は、利用者にとって興味があるレコードを検索し、レコードの内容を提供するページにアクセスして内容を閲覧する。

- 『FTP サイトからの提供』

利用者は ClinVar に蓄積されている大量のデータを一括でダウンロードする。

以下に、まず『WEB ページからの提供』について説明し、次に『FTP サイトからの提供』について説明する。『WEB ページからの提供』では、「WEB ページから利用者に提供されるデータ構成の骨格」、「WEB ページの構成」について説明する。『FTP サイトからの提供』では、FTP から提供されるデータの種類と形式、データサイズについて説明する。

『WEB ページからの提供』

利用者は、NCBI の WEB ページに備え付けられている検索機能を用いて、ClinVar に格納されているデータレコードのなかから利用者が必要な情報を検索することができる（図 4-14）。

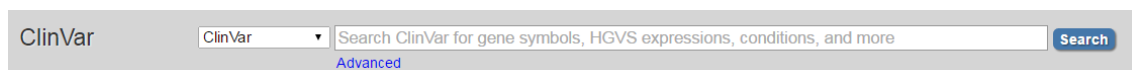


図 4-14 ClinVar の WEB ページに有る検索機能

（出所：<http://www.ncbi.nlm.nih.gov/clinvar/>）

最も単純な検索方法は、遺伝子シンボル、配列中のバリエーション²⁷、dbSNP におけるバリエーションの ID（rs number）、疾患名、登録者名をキーワードとして検索することである²⁸。この検索機能は、NCBI における他の検索機能と同様にブール検索（AND や OR を組み合わせた検索）や、フィールドを指定しての検索をサポートしている²⁹。図 4-15 は、その検索結果を表示した WEB ページである。このページには、ヒトゲノムのリファレンス配列である GRCh37³⁰において、17 番染色体の 43000000 番目の塩基から 44000000 番目の塩基に存在するバリエーションを検索した結果が一覧表示されている。図中の A がその一覧であり、各行が検索にヒットした個々のレコードである。レコードは、染色体における個々のバリエーション毎にデータをまとめたものであり、一覧の各列は、バリエーションの位置、ゲノム配列上の位置においてそのバリエーションとオーバーラップする遺伝子、症状、頻度、臨床的な重要性（後述）、

²⁷ Human Genome Variation Society (HGVS)に定められたバリエーションの表現方法である。ゲノムや RNA、遺伝子をコードした DNA、タンパク質の配列におけるバリエーションの位置と、アレルの配列パターンを表現する。その表現は、リファレンス配列のアクセッション番号と、側鎖の向き、その側鎖上での位置、リファレンス配列におけるアレルと、もう一方のアレルから構成される（例：CCDS4702.1:c.123C>T）。

（<http://www.ncbi.nlm.nih.gov/variation/hgvs/>）

²⁸ <http://www.ncbi.nlm.nih.gov/clinvar/docs/help/>

²⁹ データを構成する要素の最小単位。ClinVar の場合、例えば、研究名、データ登録者名、バリエーションの ID、アレルの ID、データ登録日等がフィールドとなっている。

³⁰ Genome Reference Consortium (GRC)によって 37 番目に構築されたヒトゲノムのリファレンス配列。

レビューステータス（後述）を示している。図中の **B** は、検索結果に対するフィルターであり、フィルターを選択することで（リンクをクリックすること）で、該当結果のみに絞り込むことができる。フィルターとしては、遺伝子の種類、臨床的な重要性、レビューステータス、バリエントの複雑度（単一、ハプロタイプ等）、アレルが同定された細胞の種類（生殖細胞系列、体細胞）等を用いることができる。

この列のリンクをクリックすることで、該当のバリエントに関するデータがまとめられたページへ移動

	Variation Location	Gene(s)	Condition(s)	Frequency	Clinical significance (Last reviewed)	Review status
1.	NM_014798.2(PLEKHM1):c.296+1G>A GRCh37: Chr17:43555265 GRCh38: Chr17:45477899	PLEKHM1	Osteopetrosis autosomal recessive 6		Pathogenic (Apr 1, 2007)	no assertion criteria provided
2.	GRCh38/hg38_17q21.31(chr17:456295-20-46111134)x1 GRCh37: Chr17:43706886-44176717 GRCh38: Chr17:45629520-46111134	CRHR1 , MAPT , CRHR1-IT1 , SPPL2C , STH , KANSL1 , MAPT-AS1 , MAPT-IT1 , MGC57346 , CRHR1	See cases		Pathogenic (Aug 12, 2011)	criteria provided, single submitter
3.	GRCh38/hg38_17q21.31(chr17:455783-81-46273786)x1 GRCh37: Chr17:43655747-44351152 GRCh38: Chr17:45578381-46273786	CRHR1 , MAPT , LRRC37A , CRHR1-IT1 , SPPL2C , STH , KANSL1 , KANSL1-AS1 , MAPT-AS1 , MAPT-IT1	See cases		Pathogenic (Aug 12, 2011)	criteria provided, multiple submitters, no conflicts
4.	GRCh38/hg38_17q21.31(chr17:456264-35-46110126)x1 GRCh37: Chr17:43703801-44187492 GRCh38: Chr17:45626435-46110126	CRHR1 , MAPT , CRHR1-IT1 , SPPL2C , STH , KANSL1 , MAPT-AS1 , MAPT-IT1 , MGC57346 , CRHR1	See cases		Pathogenic (Mar 24, 2014)	no assertion criteria provided

図 4-15 ClinVar における検索結果のページ

（出所：[http://www.ncbi.nlm.nih.gov/clinvar/?term=17\[chr\]+AND+43000000%3A44000000\[chrpos37\]](http://www.ncbi.nlm.nih.gov/clinvar/?term=17[chr]+AND+43000000%3A44000000[chrpos37])）

ヒットの一覧における「バリエントの位置」の列にあるリンクをクリックすることで、バリエントに関するデータがまとめられたページに移動する。上述の通り、レコードは、染色体における個々のバリエント毎にデータがまとめられたものであり、登録されたデータを統合したものである（図 4-16 の“Variant”）。ClinVar に登録されるデータは、バリエントと表現型に関する情報を含んだものであり、登録者によってその内容に差異が存在し得るものである（図中の“Variant Phenotype Submitter”）。レコード群のなかから、バリエント（染色体上の同じ位置に存在し、さらに同じアレルから構成されるバリエント）と表現型の組合せが同じであるものを抽出して作られたレコードが、図中の中段にある“Variant Phenotype”である。このなかから、さらに、同じバリエントだけを抽出して統合したものが、上述の“Variant”（図の最上段にあるレコード）である。“Variant Phenotype Submitter”のレコードは“SCV”の文字列から始まるアクセッション番号（例：SCV000000010）を、“Variant Phenotype”のレコ

ードは“RCV”の文字列から始まるアクセッション番号（例：RCV0000000050）を持っている。“Variant”のレコードは、それらのようなアクセッション番号を持っておらず、HGVS で定められた表記を持っており、その表記によって区別できるようになっている³¹。以下、“Variant Phenotype Submitter”のレコードを SCV レコード、“Variant Phenotype”のレコードを RCV レコードという。

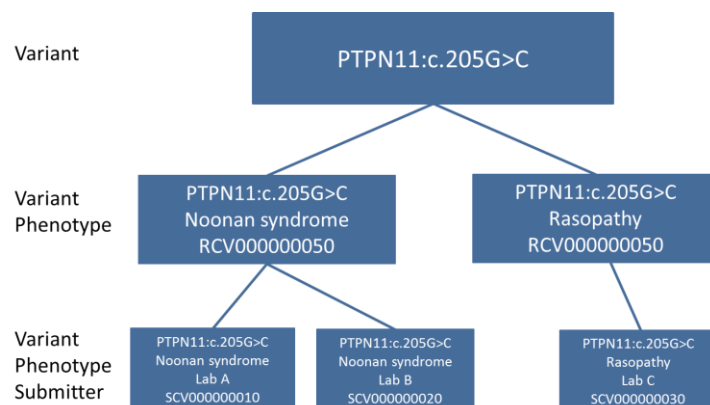


図 4-16 ClinVar において WEB ページから利用者に提供されるデータ構成の骨格
 （出所：ClinVar:A Central Repository for Interpretations of Clinically Relevant Variants、Human
 Variome Project、Melissa J Landrum）⁽⁴⁴⁾

ClinVar は、レコードの内容を利用者に提供するために、以下に示す 2 種類のページを用意している（図 4-17）。これらのページは、ほぼ同様の内容を利用者に提供するが、一部異なっている。その相違を表 3-1 にまとめる。

- Variation report
 上述の“Variant”に関するレコードを提供するためのページ（同じバリエントに関して“Variant Phenotype”を統合したレコード、つまり、“Variant Phenotype Submitter”のデータも含まれる）
- Record report
 RCV レコードを提供するためのページ（同じバリエントに関して SCV レコードを統合したページ）

³¹つまり、“Variant”のレコードの場合、HGVS で定められた表記がアクセッション番号の役割を果たす。

Variation report

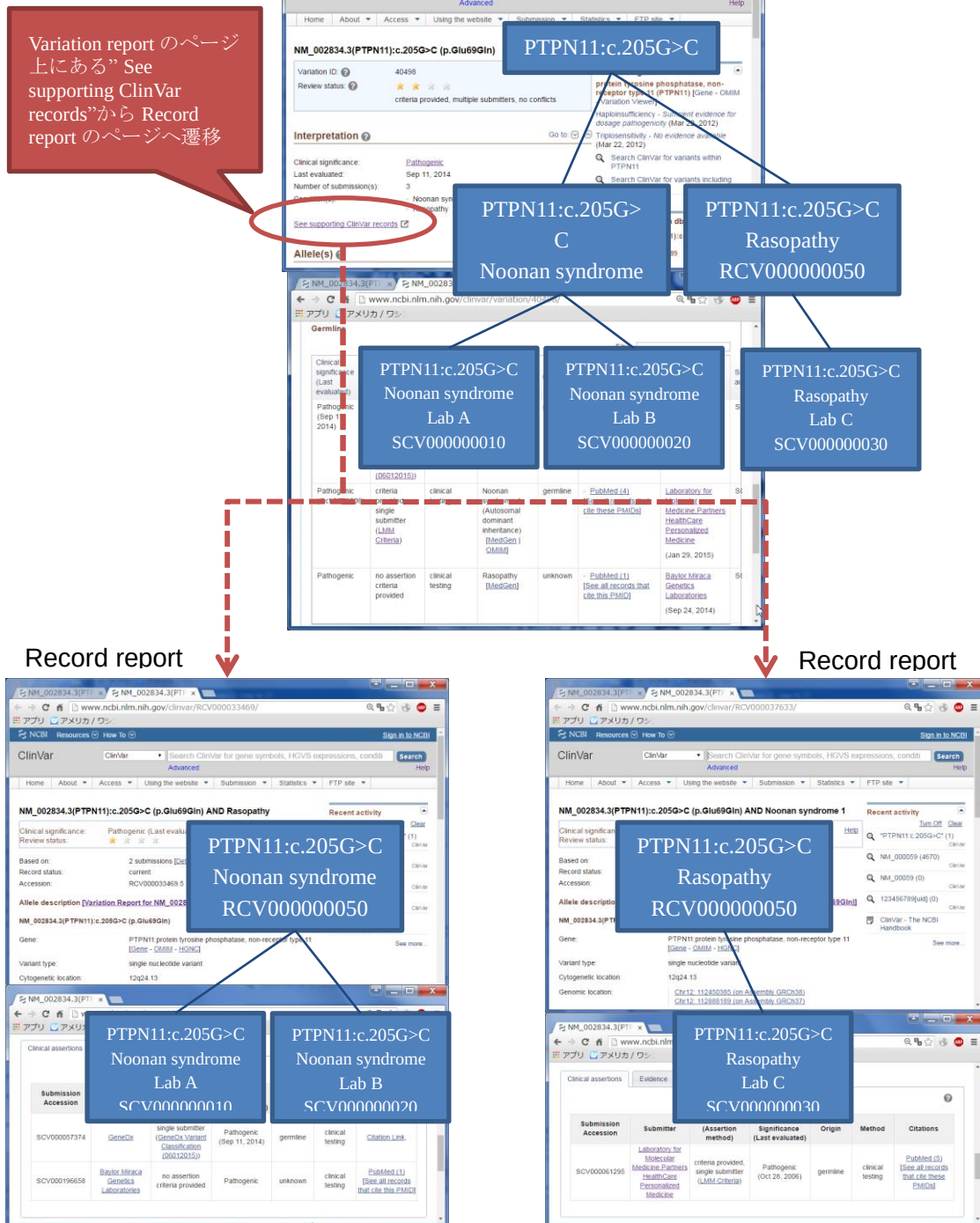


図 4-17 ClinVar の主要な WEB ページ (Variation/Record report)

(出所: <http://www.ncbi.nlm.nih.gov/clinvar/variation/40498/>、
<http://www.ncbi.nlm.nih.gov/clinvar/RCV000037633/>、
<http://www.ncbi.nlm.nih.gov/clinvar/RCV000033469/>)

RCV レコードはバージョンの概念を持っている。ClinVar には古いバージョンのレコードも保存されている（例えば、上述の RCV000000050 には、2 つのバージョン（RCV000000050.1 と RCV000000050.2 が存在している³²）。Record report のページは、それらのバージョンに対して（つまり、これらの 2 つレコードのそれぞれについて）存在する。

表 4-4 Variation report と Record report とにおける内容の相違
(出所：http://www.ncbi.nlm.nih.gov/clinvar/docs/compare_displays/)

両者で相違がある箇所	Record report のページ	Variation report のページ
ヘテロ接合体の組み合わせやハプロタイプが ClinVar に登録されており、それらを構成する単一のバリエーションも ClinVar に登録されている場合	無し	単一のバリエーションも ClinVar に登録されている場合、臨床的な重要性（Clinical significance）を、WEB ページ上の“ <i>Assertions for related alleles.</i> ”というタグが付与された場所に表示する
登録された RCV レコード群における、疾患を引き起こす性質（Pathogenicity）の比較方法	ACMG で定義されている 5 つのレベルとの厳格な比較に基づき、あらゆる差違を不一致（Conflict）として報告する	Record report の場合よりもゆるい比較を行う： “Pathogenic”と “Likely pathogenic”や、 “Benign”と“Likely benign”は不一致とはしない
症状（Condition）に関する記載の詳細さ	WEB ページ上の“Condition(s)”という領域に情報を表示するが、それ以上の明確な記載はない	臨床的な正しさについての提示（Clinical assertions）の場所に、各登録者により入力された症状を記載する
Record report へのリンクの有無	無し	“See supporting ClinVar Record”と表記された場所に、RCV レコード群へのリンクを提供する

上述の通り、Record report のページと Variation report のページとでは表示内容が似通っているため、これ以降には、Variation report のページ内容について説明する（Record report のページについては省略する）。図 4-18 及び表 4-5 はページの上部、図 4-20～図 4-22 及び表 4-6 はページの下部に関する画面キャプチャーとその説明である。

³² “.”を区切り文字としてアクセッション番号の末尾にバージョンを付与したものが、バージョン付きのアクセッション番号として取り扱われる。

NM_000059.3(BRCA2):c.8351G>A (p.Arg2784Gln)

Variation ID: 52559
Review status: ★★☆☆ criteria provided, multiple submitters, no conflicts

Interpretation

Clinical significance: [Uncertain significance](#)
Last evaluated: Oct 30, 2015
Number of submission(s): 5
Condition(s):

- Breast-ovarian cancer, familial 2 [\[MedGen - OMIM\]](#)
- Hereditary breast and ovarian cancer syndrome [\[MedGen - Orphanet\]](#)
- Hereditary cancer-predisposing syndrome [\[MedGen\]](#)
- Breast cancer [\[MedGen\]](#)

[See supporting ClinVar records](#)

Allele(s)

NM_000059.3(BRCA2):c.8351G>A (p.Arg2784Gln)

Allele ID: 67227
Variant type: single nucleotide variant
Cytogenetic location: 13q13
Genomic location:

- Chr13: 32370421 (on Assembly GRCh38)
- Chr13: 32944558 (on Assembly GRCh37)

Other names:

- 8579G>A

Protein change: R2784Q
HGVS:

- NG_012772.3:g.59942G>A
- NM_000059.3:c.8351G>A
- NC_000013.11:g.32370421G>A (GRCh38)

[...more](#)
Links:
dbSNP: [80359076](#)
NCBI 1000 Genomes Browser: [rs80359076](#)
Molecular consequence: NM_000059.3:c.8351G>A: missense variant [\[Sequence Ontology SO:0001583\]](#)
Allele frequency: GO-ESP 0.00008 (A)

1 Affected gene
breast cancer 2 (BRCA2) [\[Gene - OMIM - Variation View\]](#)
Haploinsufficiency - [Sufficient evidence for dosage pathogenicity](#) (Jul 6, 2012)
Triplosensitivity - [No evidence available](#) (Jul 6, 2012)
Search ClinVar for variants within BRCA2
Search ClinVar for variants including BRCA2

Variant frequency in dbGaP

NM_000059.3(BRCA2):c.8351G>A (p.Arg2784Gln)
GRCh37 Chr13:32944558

	Called variants	Potential variants
Sample count	1 of 2952	2 of 40906

Called variants are samples submitted to dbGaP that have the variant allele. Potential variants are SRA runs that display the allele in at least 30% of the reads covering the position, and have 10 or more passing reads covering the position.

Browser views
[RefSeqGene](#)
[Variation Viewer \[GRCh38 - GRCh37\]](#)
[UCSC \[GRCh38/hg38 - GRCh37/hg19\]](#)

Related information
[dbSNP](#)
[Gene](#)
[MedGen](#)
[OMIM](#)
[PMC](#)
[PubMed](#)

図 4-18 Variation report のページ (上部)
(出所: <http://www.ncbi.nlm.nih.gov/clinvar/variation/52559/>)

表 4-5 Variation report のページ（上部）の表示内容
 (出所：http://www.ncbi.nlm.nih.gov/clinvar/docs/variation_report/)

領域	表示内容
A	バリエントの HGVS 表記やレビューステータス等
B	解釈 (Interpretation) に関わるデータとして、臨床的な重要性、臨床的な重要性が更新された日付、登録数、登録された症状群、Record report へのリンク
C	ゲノム配列におけるバリエントの位置、dbSNP での ID、アレル頻度等
D	ゲノム配列上でバリエントとオーバーラップする遺伝子に対するリンク (NCBI における Gene のページや OMIM のページ、Haploinsufficiency や Triplosensitivity といった遺伝子の性質に対する現状の解釈に関する ClinGen のページ (Dosage Sensitivity Map) 等)
E	dbGaP におけるバリエントの頻度 (図 4-19)
F	NCBI の Variation Viewer や、UCSC Genome Browser といった可視化ツールへのリンク
G	dbGaP や MedGen といったデータシェアリング基盤に対するリンク

図 4-19 は、図 4-18 における E の部分である (図 4-18 は“NM_000059.3(BRCA2):c.8351G>A (p.Arg2784Gln)”のページから、図 4-19 は“NM_000041.2(APOE):c.[137T>C;388T>C] – Haplotype”から取られた画面キャプチャーである)。この部分は、dbGaP に格納されているデータに基づき計算されたバリエントの頻度を表示する³³。この計算に使用された研究 (Study) の一覧は、ClinVar の FTP サイトから入手可能である (後述)。

³³ この頻度は、dbGaP のデータから計算されたものであり、特定の集団 (人種等) で計算された頻度ではないため。計算方法は以下の通りである。

- VCF の場合は、少なくとも 1 つのバリエントがあれば集計対象とする
- リードデータの場合は、リードの 30% もしくは 10 本以上が、そのバリエントを含む場合に集計対象とする

Variant frequency in dbGaP ?

NM_000041.3(APOE):c.388T>C
(p.Cys130Arg)
GRCh37 Chr19:45411941

	Called variants	Potential variants
Sample count	3452 of 11445	6545 of 24665

NM_000041.3(APOE):c.137T>C (p.Leu46Pro)
GRCh37 Chr19:45411110

	Called variants	Potential variants
Sample count	81 of 10299	185 of 47255

Called variants are samples submitted to dbGaP that have the variant allele. Potential variants are SRA runs that display the allele in at least 30% of the reads covering the position, and have 10 or more passing reads covering the position.

図 4-19 dbGaP におけるバリエントの頻度
(出所 : <http://www.ncbi.nlm.nih.gov/clinvar/variation/17870/>)

Assertion and evidence details

Go to: [?] [?]

Clinical assertions Summary evidence Supporting observations

Germline

Filter:

Clinical significance (Last evaluated)	Review status (Assertion method)	Collection method	Condition(s) (Mode of inheritance)	Origin	Citations	Submitter - Study name (Last submitted)	Submission accession
Uncertain significance (Oct 30, 2015)	criteria provided, single submitter (Invitae Variant Classification Sherloc (09022015))	clinical testing	Hereditary breast and ovarian cancer syndrome (MedGen Orphanet)	germline	PubMed (1) [See all records that cite this PMID]	Invitae (Jan 6, 2016)	SCV000073501.4
Uncertain significance (Jan 20, 2015)	criteria provided, single submitter (Ambry Autosomal Dominant and X-Linked criteria (9/4/14))	clinical testing	Hereditary cancer-predisposing syndrome (MedGen)	germline		Ambry Genetics (Feb 13, 2015)	SCV000184238.2
Uncertain significance (Jun 1, 2014)	no assertion criteria provided	research	Breast cancer (Autosomal dominant inheritance) (MedGen)	germline		CSER, CC, NCGI, University of Washington Medical Center - ESP 6500 variant annotation Study description (Aug 28, 2014)	SCV000190140.1
Uncertain significance (Aug 21, 2012)	no assertion criteria provided	clinical testing	Breast-ovarian cancer, familial 2 (MedGen OMIM)	germline		Sharing Clinical Reports Project (SCRIP) (Aug 8, 2013)	SCV000109233.2
Uncertain significance (May 29, 2002)	no assertion criteria provided	clinical testing	Breast-ovarian cancer, familial 2 (MedGen OMIM)	germline		Breast Cancer Information Core (BIC) (BRCA2) (Mar 28, 2014)	SCV000147330.1

図 4-20 Variation report のページ (下部 : “Clinical assertions”のタブ)
(出所 : <http://www.ncbi.nlm.nih.gov/clinvar/variation/52559/>)

NM_000059.3(BRC x

www.ncbi.nlm.nih.gov/clinvar/variation/52559/#summary-evidence

Assertion and evidence details

Go to: [icon]

Clinical assertions Summary evidence Supporting observations

Filter:

Submitter	Families	Individuals	Allele origin	Ethnicity	Geographic origin	Citations and Databases	Description
Total for all submitters	not provided	5	germline	Central/Eastern European; Western European; Western European, Central/Eastern European, German, Hungarian	not provided		
Ambry Genetics	not provided	not provided	germline	not provided	not provided	Other citation	not provided
Breast Cancer Information Core (BIC) (BRCA2)	not provided	4	germline	Central/Eastern European; Western European; Western European, Central/Eastern European, German, Hungarian	not provided	not provided	not provided
CSER, CC, NCGI, University of Washington Medical Center	not provided	not provided	germline	not provided	not provided	not provided	not provided
Invitae	not provided	not provided	germline	not provided	not provided	PubMed Other citation	This sequence change replaces ... Full description
Sharing Clinical Reports Project (SCRCP)	not provided	1	germline	not provided	not provided	not provided	not provided

Last Updated: Apr 21, 2016

図 4-21 Variation report のページ（下部：“Summary evidence”のタブ）
（出所：<http://www.ncbi.nlm.nih.gov/clinvar/variation/52559/>）

NM_000059.3(BRC x

www.ncbi.nlm.nih.gov/clinvar/variation/52559/#supporting-observations

Assertion and evidence details

Go to: [icon]

Clinical assertions Summary evidence Supporting observations

Filter:

Submitter	Allele origin	Individuals	Phenotypes (Affected status)	Ethnicity	Geographic origin	Citations	Description
Ambry Genetics	germline	not provided	Hereditary cancer-predisposing syndrome (unknown)	not provided	not provided	Other citation	not provided
Breast Cancer Information Core (BIC) (BRCA2)	germline	1	Breast-ovarian cancer, familial 2 (yes)	not provided	not provided	not provided	not provided
Breast Cancer Information Core (BIC) (BRCA2)	germline	1	Breast-ovarian cancer, familial 2 (yes)	Central/Eastern European	not provided	not provided	not provided
Breast Cancer Information Core (BIC) (BRCA2)	germline	1	Breast-ovarian cancer, familial 2 (yes)	Western European	not provided	not provided	not provided
Breast Cancer Information Core (BIC) (BRCA2)	germline	1	Breast-ovarian cancer, familial 2 (yes)	Western European, Central/Eastern European, German, Hungarian	not provided	not provided	not provided
CSER, CC, NCGI, University of Washington Medical Center	germline	not provided	Breast cancer (unknown)	not provided	not provided	not provided	not provided
Invitae	germline	not provided	Hereditary breast and ovarian cancer syndrome (unknown)	not provided	not provided	PubMed Other citation	not provided
Sharing Clinical Reports Project (SCRCP)	germline	1	Breast-ovarian cancer, familial 2 (not provided)	not provided	not provided	not provided	not provided

図 4-22 Variation report のページ（下部：“Supporting observations”のタブ）
（出所：<http://www.ncbi.nlm.nih.gov/clinvar/variation/52559/>）

表 4-6 Variation report のページ（下部）の表示内容
 (出所：http://www.ncbi.nlm.nih.gov/clinvar/docs/variation_report/)

タブ名	表示内容
Clinical assertions	登録されたレコード毎に、臨床的な重要性、最終更新日、レビューステータス、検査の方法、症状（登録者によって提示される臨床的な重要性に遺伝様式が関与する場合には、その遺伝様式も記載）、MedGen や Orphanet、OMIM 等へのリンク、検査対象となった細胞の種類（生殖細胞系列あるいは体細胞といった細胞の種類）、引用、登録者、研究プロジェクト名、最終登録日を表示する。後述する通り、データの登録者は、臨床的な重要性を提示する方法についてのドキュメントを添付する必要がある。タブにおける「臨床的な重要性」には、そのドキュメントに対するリンクが添えられる（図 4-23）。登録者に関しては、登録者の氏名や住所、登録数、その遺伝子毎・症状毎の統計や、遺伝子検査毎の統計を表示するページへのリンクが添えられる（図 4-24）。
Summary evidence	登録者毎に、登録者名、登録した家族数、登録した個人数、検査対象となった細胞の種類（同上）、登録された人種とその地域（Geographic origin）、引用文献及びデータベースへのリンク、登録者についての説明を表示する。
Supporting observations	登録者、個体数、表現型、人種、その地域、検査対象となった細胞の種類を組み合わせたものを観察結果として、観察結果毎にその内容を表示する。図 4-22 にある通り、ある登録者から、複数の観察結果が登録されていることが分かる。

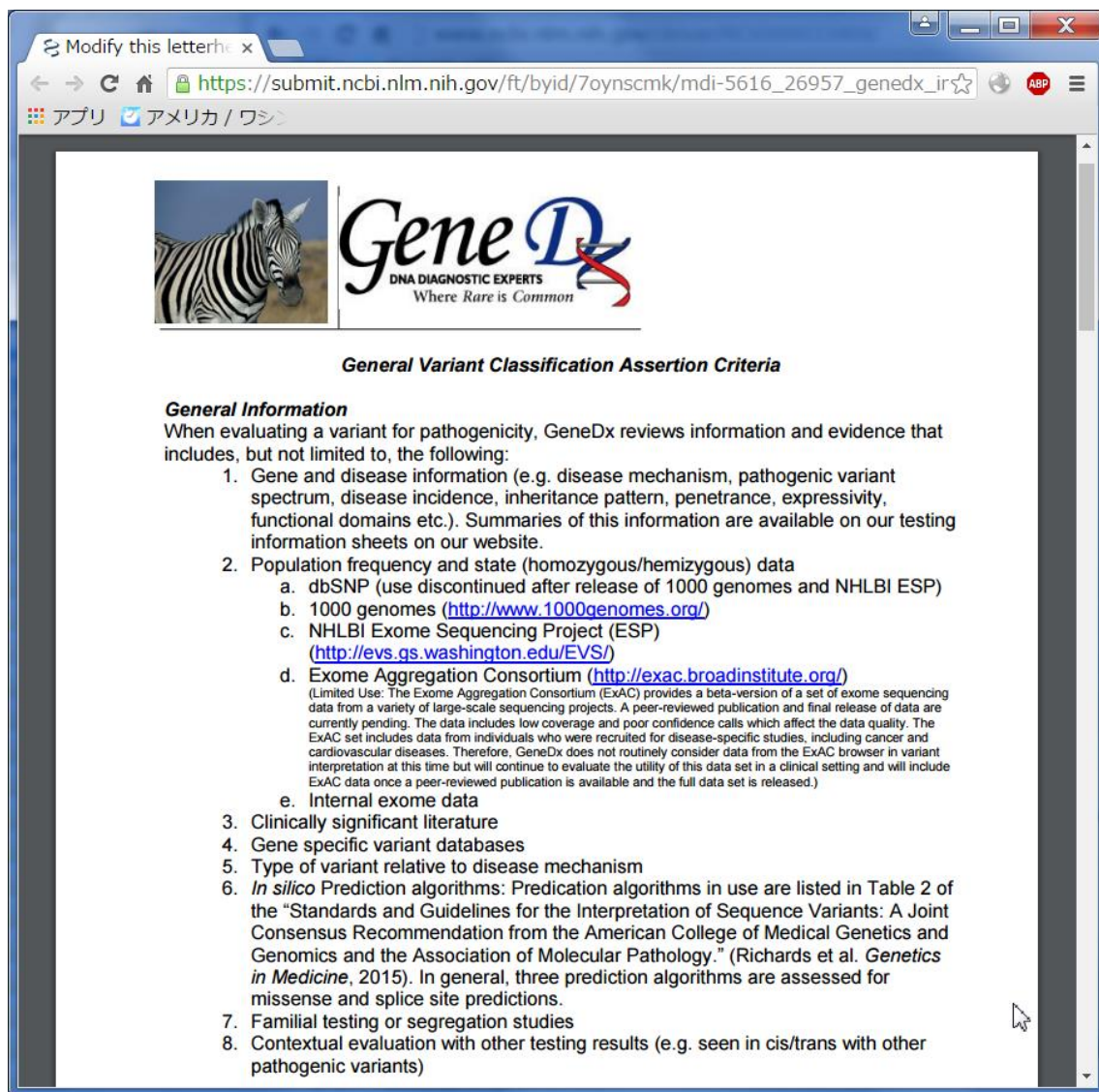


図 4-23 臨床的な重要性を提示する方法についてのドキュメントのページ
(Variation report のページからのリンク)

(出所 : https://submit.ncbi.nlm.nih.gov/ft/byid/7oynscmk/mdi-5616_26957_genedx_interprules_final_061215.pdf)

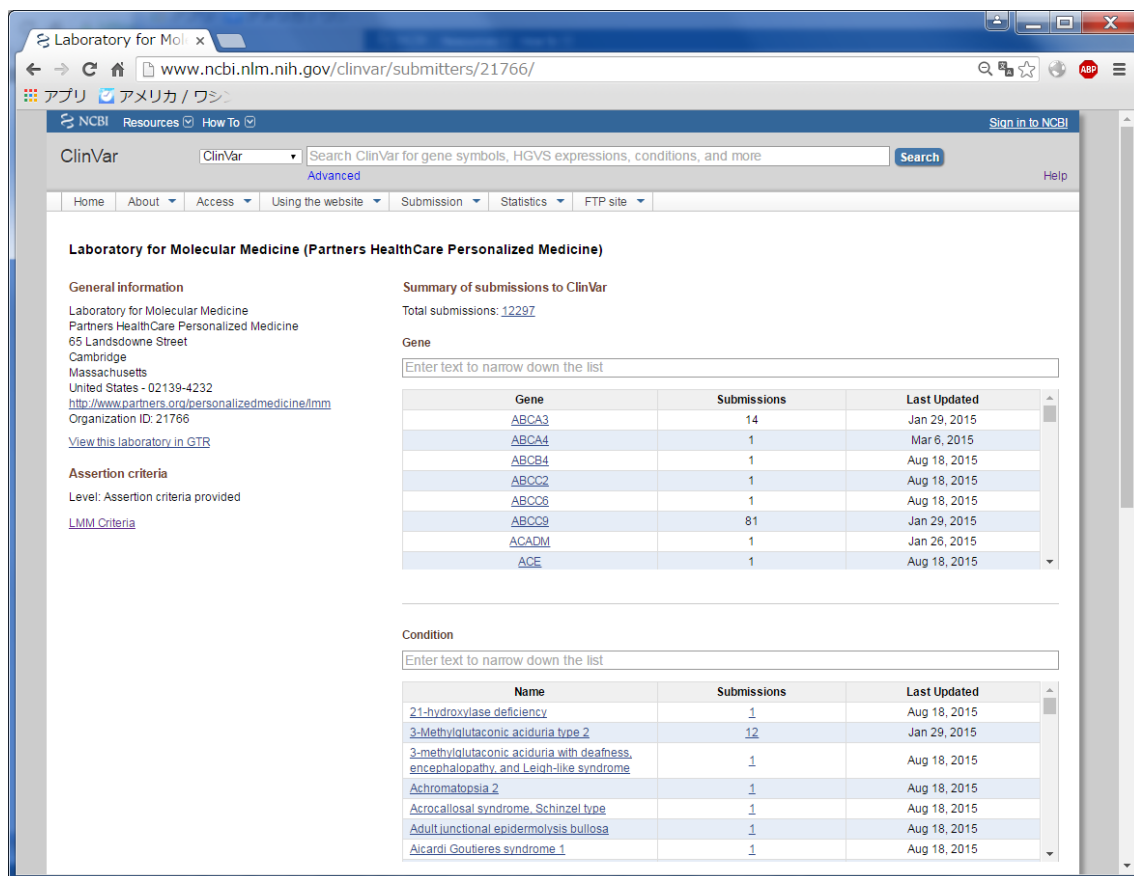


図 4-24 登録者に関する統計のページ (Variation report のページからのリンク)
(出所: <http://www.ncbi.nlm.nih.gov/clinvar/submitters/21766/>)

『FTP サイトからの提供』

ClinVar の FTP サイトからは、ClinVar に格納されているデータ、データ構造の定義が示されているファイル、データ登録のテンプレートとなるファイル、ClinVar に関する各種ドキュメント (開発の構想やプレゼンの資料)、エキスパートパネルに関する資料を提供する (図 4-25)。データやデータ構造の定義を格納したファイルには、ダウンロードした者が、データが破損していないかを確認するために、MD5 チェックサムを格納したファイルが付随している。

以下に、FTP サイトから提供されるファイルをまとめる。

- バリエントに関するデータ (VCF、タブ区切り、XML の 3 種類)

- ✓ VCF のファイル

ヒトゲノムのリファレンス配列毎に作成されている (現状、GRCh37 と GRCh38 についてフォーマットのファイルが作成されている)。GRCh37 についてのデータは、vcf_GRCh37 という名前のディレクトリ配下に置かれている (表 4-7 に示されるファイル群が置かれている)。このなかの clinvar.vcf には、臨床的な重要性や疾患名、提示に関する基準 (Assertion criteria) の有無等に関するデータが含まれる (common_and_clinical.vcf³⁴等のファイルには含まれていない)。

³⁴ 臨床的な意味に関してのエビデンスが付与された生殖細胞系列のバリエントを格納するファイル (表 4-7)。

- ✓ タブ区切りフォーマットのファイル
ClinVar に登録されたデータを統合・整理したものであり、WEB ページから閲覧できる大部分のデータを含んでいる (表 4-8)。これらのファイル群は、`tab_delimited` という名前のディレクトリ配下に置かれている。
- ✓ XML フォーマットのファイル
RCV レコードの全データを格納している。これらのファイル群は、`xml` という名前のディレクトリ配下に置かれており、毎月、更新されている。このデータ構造に関しての定義は `clinvar_public.xsd` に示されている^{35,36}。
- 疾患名や、遺伝子と疾患との関係性に関するファイル
 - ✓ 疾患名に関するファイル (`disease_names`)
Genetic Testing Registry (GTR) と ClinVar で使用される疾患名に関するデータが格納されている。そのデータは、疾患名毎に、その疾患名が用いられている情報ソース (Human Phenotype Ontology 等)、コンセプト ID³⁷、MIM³⁸ の ID 等が関連付けられている。ファイルは毎日更新される。
 - ✓ 遺伝子と疾患との関係性に関するファイル (`gene_condition_source_id`)
遺伝子と疾患との関係性を収めたファイルであり、その関係性が、NCBI が提供する遺伝子関連のデータベース (ClinVar、Gene、GTR、MedGen) に含まれることになる。この関係性は、OMIM や GeneReviews、NCBI のスタッフによるキュレーションによって得られたものである。データは、遺伝子の ID 毎に、遺伝子のシンボル、疾患名、コンセプト ID、MIM の ID 等が関連付けられている。ファイルは毎日更新される。
 - ✓ コンセプト ID の履歴が格納されたファイル (`ConceptID_history.txt`)
コンセプト ID の履歴が格納されたファイルである。
- データ登録の際に使用されるテンプレートファイル
`submission_templates` という名前のディレクトリ配下に置かれている (後述)。
- バリエーションの頻度を計算するために用いられた研究の一覧を格納したファイル
ClinVar における Variation report のページには、dbGaP に格納されているデータに基づくバリエーションの頻度が表示されている。その頻度の計算に用いられた研究 (Study) の一覧が、`dbGaP_frequency_study_list.html` と `dbGaP_frequency_study_list.txt` に格納されている。
- ClinVar に関するドキュメント類
ClinVar の構想に関するドキュメントや、プレゼンテーションが `document_archives` と

³⁵ XSD は、XML のスキーマ (データの構造: データに含まれる値とその型、それらの関係性) を定義するフォーマット。

³⁶ XSD の履歴は、`xsd_public` というディレクトリの配下に置かれている。`xsd_public` の配下にある `variation_report` というディレクトリには、“Variant”のレコードについての XSD も置かれている (これは利用者に対しての新しいデータ提供方法として 2015 年 7 月に追加された)。登録されたレコード (アクセッション番号が SCV で始まるレコード) の XSD は、`clinvar_submission.xsd` であり、その履歴は `xsd_submission` という名前のディレクトリに置かれている。

³⁷ 疾患名を管理するための ID。

³⁸ ヒト遺伝子とそれらが関わる表現型に関する情報をソースである Mendelian Inheritance in Man (MIM) における ID。OMIM は MIM のオンライン版。

いうディレクトリ及び **presentations** というディレクトリに格納されている。

- エキスパートパネルを形成するためのドキュメント類

ClinVar では、後述する通り、3 つ星のレビューレベルを得るためには、エキスパートパネルによって、データのキュレーションが行われる必要がある。エキスパートパネルは、データ登録者が提案する専門家によって形成されるものであり、ClinGen によって認められる必要がある。ClinGen というディレクトリには、その申請書³⁹（エキスパートパネルのメンバー等を記入するためのドキュメント）と、ClinGen によって認められたエキスパートパネル⁴⁰が提出した申請書及びキュレーション方法に関するドキュメントが置かれている。



図 4-25 ClinVar の FTP サイト
(出所 : <ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/>)

³⁹ Request_for_Expert_Panel_or_Practice_Guideline_Review_Level_Designation .docx

⁴⁰ 現在、CFTR2、ENIGMA、InSiGHT、PharmGKB から提出されているエキスパートパネルに関するドキュメントがこのディレクトリに置かれている。

表 4-7 ClinVar の FTP サイトから提供される VCF ファイル
(出所: http://www.ncbi.nlm.nih.gov/variation/docs/human_variation_vcf/)

ファイル名の接頭辞	更 新 頻度	内容
common_and_clinical.vcf	毎週	臨床的な意味に関してのエビデンスが付与された生殖細胞系列のバリエーションを格納するファイル
common_and_clinical_papu.vcf		X 染色体・Y 染色体上に存在する擬似的な常染色体領域等 (Pseudoautosomal region (PAR), Alternate loci, patch sequences and unlocalized or Unplaced contigs (papu)) に含まれるバリエーションを格納したファイル (“common_and_clinical” に含まれないものを格納)
common_no_known_medical_impact.vcf		遺伝性疾患への影響度 (Pathogenicity rating) が分かっていないものを格納
clinvar.vcf		Locus-Specific Databases (LSDB)、遺伝子検査室 (Genetic testing laboratories)、OMIM 等からのチャンネルから少なくとも 1 つのバリエーションが登録されている rs (dbSNP にいてリファレンスとして取り扱われている多型 ⁴¹) を格納したファイル

⁴¹ http://www.ncbi.nlm.nih.gov/SNP/get_html.cgi?whichHtml=how_to_submit#REFSNP

表 4-8 ClinVar の FTP サイトから提供されるタブ区切り形式のファイル
(出所: <ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/README.txt>)

ファイル名の接頭辞	更新頻度	内容
gene_specific_summary.txt	毎月	遺伝子に関するデータを整理した結果を格納するファイルであり、遺伝子毎に、該当する遺伝子とゲノム配列上で重なるバリエーション数や、バリエーションに関しての登録数等を格納する
variant_summary.txt	毎月	バリエーションに関するデータを整理した結果を格納するファイルであり、該当するバリエーションとゲノム配列上で重なる遺伝子や、臨床的な重要性、該当するバリエーションに対する遺伝子検査が存在するか (Genetic Testing Registry (GTR) に登録されているもの) 等を格納する (ClinVar におけるアレルの ID 毎にデータを格納する)
cross_references.txt	毎月	バリエーションからの ClinVar 以外のデータベースへのリンクを格納する (リンクは、ClinVar におけるアレルの ID 毎に設定されたものである)
var_citationsvar.txt	毎週	バリエーションからの文献への引用を格納する (ClinVar におけるアレルの ID 毎にリンクを設定する)
summary_of_conflicting_interpretations.txt	2016 年 1 月から提供	登録者間 (2 者間) における登録内容の相違を格納する (「臨床的な重要性 (Clinical significance)」や「レビューステータス」等に差異があったバリエーションについて差異の内容を列挙する)
special_requests (ディレクトリ)	現在は更新停止	登録者に関するデータを格納したファイル等が置かれているディレクトリ

(4) 臨床的な重要性

ClinVar は、登録されるバリエントの臨床的な重要性（Clinical significance）として表 4-9 に示す値を選択可能としている⁴²。

表 4-9 ClinVar に登録される臨床的な重要性
(出所：<http://www.ncbi.nlm.nih.gov/clinvar/docs/clinsig/>)

値	説明
Benign	ACMG/AMP ⁴³ 推奨による、メンデル遺伝病（Mendelian disorders）に関わるバリエントを分類するための用語 ⁽⁴⁵⁾ 疾患を引き起こすわけではないバリエントであることを意味する
Likely benign	ACMG/AMP ⁴³ 推奨による、メンデル遺伝病（Mendelian disorders）に関わるバリエントを分類するための用語 ⁽⁴⁵⁾ Benign である可能性が高いバリエントであることを意味する（ ⁽⁴⁵⁾ では、90%以上の確かさにあるバリエントをこれに分類するとしている）
Uncertain significance	ACMG/AMP ⁴³ 推奨による、メンデル遺伝病（Mendelian disorders）に関わるバリエントを分類するための用語 ⁽⁴⁵⁾ Benign、Likely benign、Likely pathogenic、Pathogenic に分類されないものに、Uncertain significance を与える。
Likely pathogenic	ACMG/AMP ⁴³ 推奨による、メンデル遺伝病（Mendelian disorders）に関わるバリエントを分類するための用語 ⁽⁴⁵⁾ Pathogenic である可能性が高いバリエントであることを意味する（ ⁽⁴⁵⁾ では、90%以上の確かさにあるバリエントをこれに分類するとしている）
Pathogenic	ACMG/AMP ⁴³ 推奨による、メンデル遺伝病（Mendelian disorders）に関わるバリエントを分類するための用語 ⁽⁴⁵⁾ 疾患を引き起こすバリエントに対しての表現
Drug response	疾患には影響を及ぼさないが、薬に対する反応に影響を及ぼすバリエント（今後、薬の反応に対してより詳細な値（CPIC ⁴⁴ 推奨の用語）を追加する予定）
Association	GWAS によって同定されたバリエントに対して、さらに臨床的な重要性が加えられたもの
Risk factor	疾患を引き起こすわけではないがそのリスクを高めるバリエント
Protective	疾患（感染を含む）のリスクを下げるバリエント
Affects	乳糖不耐性（Lactose intolerance）のような、疾患以外の表現形にかかわるバリエント

⁴² 登録者によって「臨床的な重要度」に対して特定の遺伝様式が指定された場合、“Clinical assertions”のタブには、その遺伝様式が表示される（図 4-20）。

⁴³ American College of Medical Genetics and Genomics(ACMG) (<https://www.acmg.net/>)、
Association for Molecular Pathology (AMP) (<http://www.amp.org/>)

⁴⁴ <https://cpicpgx.org/>

値	説明
Conflicting data from submitters	コンソーシアムから登録されるバリエーションに対して付与される表現：コンソーシアム内で解釈に不一致がある場合、この表現を用いる
Other	この表現は、ClinVar が提示するオプションに、適切な表現がない場合に用いられる。この場合、ClinVar のスタッフが、データ登録者に対して、ClinVar のオプションに加えるべき表現があるかを問い合わせることになる。
not provided	臨床的な重要性に対しての解釈を含まない場合に用いられる表現 ⁴⁵

SCV レコード群を統合して RCV レコードを作成する際に、SCV レコード群の臨床的な重要性に相違が見受けられる場合がある。それらの間で、ACMG 及び AMP が推奨する表現（Benign、Likely benign、Uncertain significance、Likely pathogenic、Pathogenic）だけが用いられており、それらに差異があった場合、RCV レコードでは、“Conflicting interpretations of pathogenicity”（臨床的な重要性に関しての解釈に不一致がある）という表現を用いる。“Pathogenic”と“Likely pathogenic”、“Benign”と“Likely benign”、といった差異には、“Conflicting”という言葉を用いない（表 4-10）。

⁴⁵ この表現の使用は、以下のような場合に用いられる（いずれの場合も、疾患に関する重要度が不明であることが前提）。

- バリエーションに関する論文発表のために、そのバリエーションを ClinVar に登録する際
- 機能的な重要度を登録する際（例えば、タンパク質レベルではその差異が見いだせないような場合）
- クリニックや臨床医によって表現型等に関する付加的な情報（バリエーションを持っている個人に関する情報）が登録される場合

表 4-10 登録者間において相違があった場合の臨床的な重要性の表現例

(出所: <http://www.ncbi.nlm.nih.gov/clinvar/docs/clinsig/>)

SCV に含まれる表現 (例)	RCV における表現
Pathogenic と Likely pathogenic を含む表現	Pathogenic/LikelyPathogenic
(Pathogenic もしくは Likely pathogenic もしくは Benign もしくは Likely benign) と Uncertain significance を含む表現	Conflicting interpretations of pathogenicity
(Pathogenic もしくは Likely pathogenic) と (Benign もしくは Likely benign) を含む表現	Conflicting interpretations of pathogenicity
Benign と Likely benign を含む表現	Benign/Likely benign
ACMG/AMP 推奨表現のどれかと、それ以外の表現を含む表現 (例えば、Uncertain significance と Risk factor を含む表現)	Uncertain significance, Risk factor
ACMG/AMP 推奨表現が一致しておらず、さらに ACMG/AMP 推奨以外の用語を含む表現 (例えば、Pathogenic と Uncertain significance と Risk factor を含む表現)	Conflicting interpretations of pathogenicity, Risk factor
ACMG/AMP 推奨表現が含まれておらず、ACMG 及び AMP が推奨していない複数個の用語を含む表現 (例えば、Drug response と Risk factor)	Drug response, Risk factor
データ登録者であるコンソーシアム内において、どの ACMG/AMP 推奨表現が適切であるかの共通認識が固まっていない	Conflicting data from submitters

(5) レビューステータス

Variation report 及び Record report のページには、登録者によって提示された臨床的な重要性の信頼度を示すために、レビューステータスが表示されている。このレビューステータスは、個々の登録 (SCV レコード) に与えられるものであり、デフォルトでは“single submitter - criteria not provided”となっている。レビューステータスは、星なしから、星 4 個 (4 stars) の 5 段階に分かれる。星なし、星 1 つ、星 3 つ、星 4 つは、個々の SCV レコードに対して定めることができる。星 2 つは、複数の SCV レコードを比較することによって定めることができる。表 4-11 に、これらの定め方をまとめる。

表 4-11 レビューステータスの決定方法

(出所: http://www.ncbi.nlm.nih.gov/clinvar/docs/review_guidelines/)

レビューレベル	概要及び要件	ClinGen の関わり
—	なし	なし
★ (1 star)	臨床的な重要性を判定するために用いられた基準（その基準を示した文章）が、ある 1 人の登録者から登録された場合にこのレビューステータスを与える（登録者は以下の要件を満たす必要がある） ● アレルもしくは遺伝子型の分類方法に関するドキュメントを作成すること（この分類方法は、エビデンスに対する広範なレビュー（現在のガイドライン（Practice guidelines）もしくはさらに厳しい基準と一致するレビュー）に基づくこと） ● 単一遺伝子疾患（メンデル遺伝病）に対して、少なくとも 3 段階の意義付け（Pathogenic、Uncertain significance、Benign）がなされていること、または他の適切な臨床的な重要性が提示されていること ● バリエントに関する提示（Pathogenic 等の臨床的な重要性）とその基準が示された文献に対する引用や、その電子媒体を登録すること（ClinVar のホームページから、それらに対するリンクが貼られる） ● バリエントの分類に関するエビデンスや論理的な根拠（引用、ケース数、エビデンスを要約した文章等）、またはコンタクト先を登録すること ClinVar では、登録される基準に対して詳細な審査は行わない	なし
★★ (2 stars)	複数の登録者によって臨床的な重要性が提示されており、さらにそれらに一致しない場合にこのレビューステータスを与える（不一致がある場合には、1 star となる）	なし

レビューレベル	概要及び要件	ClinGen の関わり
★★★ (3 stars)	エキスパートパネルによってレビュー・承認された場合にこのステータスを与える ● 登録者は、エキスパートパネルのメンバーに関する情報を申請書 (ClinVar Expert Panel request form) に記入し、clinvar@ncbi.nlm.nih.gov 宛に送付し、それが認められなければならない ● エキスパートパネルには医療の専門家（疾患に関わる遺伝子を保有する患者に思いやりを持つ者）、医学に携わる遺伝学者（Medical geneticists）、臨床検査施設の診断医・分子病理医（Molecular pathologists）、対象となる疾患や遺伝子、機能分析（Functional assay）、統計解析の専門家が含まれることが推奨される ● エキスパートパネルのメンバーは、複数の学術機関・民間研究機関から構成されること ● エキスパートパネルのメンバーは国際的に活躍する人物であり、発表論文や、研究コミュニティに対する長年の貢献によって認められている人物であること ● ClinGen は、一つの遺伝子に対して一つのエキスパートパネルが形成されることを望む（その領域のエキスパートがそこに含まれることも望まれる） ● ある遺伝子に対して既存のエキスパートパネルが存在すれば、それらと一緒にエキスパートパネルを構築することを検討する必要がある、そうするのでなければ、もう一つのエキスパートパネルを構築する正当性を示す必要がある ● 1star の要件を満たしていること ● エキスパートパネル間に存在するあらゆる金銭的な利益相反の存在と、その管理方法を通知すること ClinGen はエキスパートパネルの構築を推奨しており、その構築に関して、早い段階から相談に応じること	登録者から申請されたエキスパートパネルは、Clinical Genome Resource (ClinGen) program によってレビュー・承認される必要がある⁽⁴⁶⁾

レビューレベル	概要及び要件	ClinGen の関わり
★★★★ (4 stars)	ガイドラインに合致する場合にこのレビューステータスを与える ● 登録者は、申請書 (ClinVar Practice Guideline request form) に必要事項を記入し、 clinvar@ncbi.nlm.nih.gov に送付し、 申請し、ClinGen のレビューにより判断される ● 1 star 及び 3 star の要件を満たしていること ● エビデンスの確からしさを評価するシステムについて説明すること ● 公表に先立って、バリエントが疾患に関わることを決定することに対してどのような外部のレビュープロセスを用いたかを説明すること ● バリエントに対する解釈を、どのようにして常に更新し続けるかを説明すること	登録されたバリエントは、ClinGen Steering Committee によってレビュー・承認される必要がある

(6) データ登録数・サイズ

表 4-12 に、2016 年 1 月 25 日時点での、ClinVar に登録されたデータに関する統計を示す。登録数（つまり SCV レコードの数）は約 17 万 3 千件であり、その内、臨床的な重要性を判定するための基準（Assertion criteria）が付与されている数は約 8 万 5 千件（全登録件数の約半数）となっている。

表 4-12 登録されているデータ数に関する統計（2016/1/25 の統計）
（出所：<https://www.ncbi.nlm.nih.gov/clinvar/submitters/>）

分類		件数
登録数（SCV レコード数）		173,252
臨床的な重要性を判定するための基準（Assertion criteria）が付与されている登録数		85,663
ユニークなバリエーション数	ユニークなバリエーション数	135,430
	解釈が付与されているバリエーション数	121,834
	提示のための基準が付与されているバリエーション数	70,816
	★★★★ ガイドライン（Practice guideline）に合致するバリエーション数（ClinGen Steering Committee によってレビューされたバリエーション）	23
	★★★ エキスパートパネルによってレビューされたバリエーション数	3,620
	★★ 複数の登録者によって、提示のための基準が付与されており、それらの内容が一致しないバリエーション数	7,359
	★ 提示のための基準が付与されているバリエーション数	56,879
ゲノム配列上でオーバーラップする遺伝子群と関連付けられるバリエーション数※ （※ゲノム配列上で 2 つ以上の遺伝子領域がオーバーラップしている場合に、この部分に存在するバリエーション数）		8,749
遺伝子数	バリエーションと関連付けられる遺伝子数	26,613
	1 つの遺伝子に特異的であるバリエーション※を持つ遺伝子数 （※ゲノム配列上で、ある一つの遺伝子の領域中のみに存在するバリエーションの数）	4,492
	1 つのタンパク質をコードする遺伝子に特異的であるバリエーション※を持つ遺伝子数 （※ゲノム配列上で、タンパク質をコードするある一つの遺伝子の領域中のみに存在するバリエーションの数）	4,406
	1 つ以上のバリエーションと関連付けられる遺伝子数※ （※複数のバリエーションと染色体上でオーバーラップする遺伝子の数）	26,498
登録者数		440

ClinVar の FTP サイトに置かれているファイルの内、登録されたデータを蓄積したファイル、つまり ClinVar の主要かつサイズが大きいデータは以下の通りである。

表 4-13 ClinVar の FTP サイトに置かれている主要なファイルのデータサイズ

(出所 : <ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/>)

ファイルの種類	サイズ
XML フォーマットのファイル※	143Mbyte
タブ区切りフォーマットのファイル※※	15Mbyte
VCF のファイル※※※	1,914Mbyte

※ RCV レコードの全データを格納しているファイル。ClinVarFullRelease_2016-03.xml.gz のファイルサイズをサイズとした。

※※ ClinVar に登録されたデータを統合・整理したものであり、WEB ページから閲覧できる大部分のデータを含んでいるファイル (RCV レコードの一部分)。以下のファイルの合計値をこのサイズとした。

gene_specific_summary_2016-03.txt.gz

summary_of_conflicting_data_2016-03.txt.gz

summary_of_conflicting_interpretations_2016-03.txt.gz、variant_summary_2016-03.txt.gz

※※※ ヒトゲノムのリファレンス配列毎に作成されているファイル (RCV レコードの一部分)。GRCh37 と GRCh38 について作成されているファイルの合計値をこのサイズとした。対象となったファイルは以下の通り (GRCh37 について作成された物と GRCh38 について作成されたものの両方)。

clinvar_20160302.vcf.gz

clinvar_20160302_papu.vcf.gz

common_and_clinical_20160302.vcf.gz

common_and_clinical_20160302_papu.vcf.gz

common_no_known_medical_impact_20160302.vcf.gz

common_no_known_medical_impact_20160302_papu.vcf.gz

(7) データ登録者

ClinVar にデータを登録することができるものは、以下に示す通りとなっている⁴⁶。

- 臨床検査施設 (Clinical testing labs)
- 研究者 (Researchers)
- 遺伝子検査を提供するクリニック (Genetics clinics)
- ゲノム上のバリエーションの位置を取り扱うデータシェアリング基盤 (Locus-specific databases)
- エキスパートパネル (Expert panels)
- ガイドラインを策定するグループ (Groups establishing practice guidelines)

現在、ClinVar にデータを登録するために、バリエーションに対する臨床的な重要性を判定するための手段として認められている手法は表 4-14 に示される 3 通りとなっている。

⁴⁶ <http://www.ncbi.nlm.nih.gov/clinvar/docs/submit/#who>

表 4-14 ClinVar にデータを登録するために
バリエントに対する臨床的な重要性を判定するための手段
(出所： <http://www.ncbi.nlm.nih.gov/clinvar/intro/>)

手段	説明	ClinVar からの利用者に情報を提供する際に用いられる用語
臨床検査 (Clinical testing)	CLIA もしくは ISO 15189 を取得した検査室における臨床検査の結果の一部として報告される臨床的な重要性 (Clinical significance)	clinical testing
研究 (Research)	ヒトを対象とする研究プロジェクトの一部として、バリエントに対して付与される情報	research
文献からの抽出 (Extraction from the literature)	表現型に影響を及ぼすバリエントに関して、論文中に含まれている報告 (著者による記述には修飾を加えずに抽出される)	literature only

表 4-15 は、ClinVar に対するデータ登録者の一部を示したものである。2016/3/31 における全登録者数は 482 である。この表には、登録数が多い上位の 20 者のみを示す。以下は、この上位の 20 者を分類する。

- 他のデータシェアリング基盤
OMIM⁴⁷、GeneReviews⁴⁸、Sharing Clinical Reports Project⁴⁹、Breast Cancer Information Core (BIC)⁵⁰、Retina International⁵¹
- 検査会社
GeneDx⁵²、Invitae⁵³、Laboratory for Molecular Medicine; Partners HealthCare Personalized Medicine⁵⁴、Ambry Genetics⁵⁵
- 研究室
Emory Genetics Laboratory; Emory University School of Medicine⁵⁶、Genetic Services Laboratory; University of Chicago、Cardiovascular Biomedical Research Unit; Royal Brompton and Harefield NHS Foundation Trust⁵⁷

⁴⁷ <http://www.ncbi.nlm.nih.gov/omim>

⁴⁸ <http://www.ncbi.nlm.nih.gov/books/NBK1116/>

⁴⁹ <https://www.clinicalgenome.org/data-sharing/sharing-clinical-reports-project-scrp/>

⁵⁰ <https://research.nhgri.nih.gov/bic/>

⁵¹ <http://www.retina-international.org/>

⁵² <http://www.genedx.com/>

⁵³ <https://www.invitae.com/en/>

⁵⁴ <http://personalizedmedicine.partners.org/laboratory-for-molecular-medicine/>

⁵⁵ <http://www.ambrygen.com/>

⁵⁶ <http://geneticslab.emory.edu/>

⁵⁷ <http://www.rbht.nhs.uk/about/our-work/publications/08-09/research/cardiovascular-biomedical-research/>

- コンソーシアム、研究プロジェクト、研究共同体
 ISCA site 1/6/4⁵⁸、International Society for Gastrointestinal Hereditary Tumours⁵⁹、Biesecker Laboratory - ClinSeq Project⁶⁰ (ClinSeq は米国の複数大学による研究プロジェクトである CSER の一部であるため、このカテゴリに分類)
 ※ 34 位にも CSER からの登録がランキングされている (CSER_CC_NCGL; University of Washington)
 ※ 141 位には、American College of Medical Genetics and Genomics (ACMG) からの登録がランキングされている。ClinVar に登録されている Practice guideline レベルの登録 (23 件) は、今のところ、全て ACMG からの登録となっている。
- 非営利の研究機関
 Inova Translational Medicine Institute; Inova Health System⁶¹
 ※ 76 位にも非営利の研究機関である HudsonAlpha Institute からの登録がランキングされている。

⁵⁸ http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000205.v2.p1

⁵⁹ <http://insight-group.org/>

⁶⁰ <https://www.genome.gov/10000356/biesecker--group/>

⁶¹ <http://www.inova.org/itmi/home>

表 4-15 ClinVar に対するデータ登録者 (2016/3/31 付の統計における登録数上位の 20 者)

(出所 : http://www.ncbi.nlm.nih.gov/clinvar/docs/submitter_list/)

登録者	最も高いレビュー ステータス	登録数	解釈が付与さ れた登録数	遺伝 子数	最終更新 日
<u>OMIM; Johns Hopkins University</u>	-	26518	26517	3951	Mar 18, 2016
<u>GeneDx</u>	Assertion criteria	23735	22377	650	Oct 23, 2015
<u>Emory Genetics Laboratory; Emory University School of Medicine</u>	Assertion criteria	16047	15982	1297	Jun 09, 2015
<u>Invitae</u>	Assertion criteria	13115	8425	334	Jan 27, 2016
<u>Laboratory for Molecular Medicine; Partners HealthCare Personalized Medicine</u>	Assertion criteria	12297	12207	405	Aug 18, 2015
<u>Genetic Services Laboratory; University of Chicago</u>	Assertion criteria	10413	10413	818	Sep 15, 2015
<u>Ambry Genetics</u>	Assertion criteria	9995	9995	204	Nov 25, 2015
<u>ISCA site 1</u>	Assertion criteria	5563	5563	24927	Jun 21, 2014
<u>GeneReviews</u>	-	5165	5149	614	Mar 03, 2016
<u>ISCA site 6</u>	Assertion criteria	3690	3690	21110	Jun 21, 2014
<u>ISCA site 4</u>	Assertion criteria	2983	2983	21035	Jun 21, 2014
<u>Inova Translational Medicine Institute; Inova Health System</u>	Assertion criteria	2692	4	171	Dec 01, 2015
<u>International Society for Gastrointestinal Hereditary Tumours</u>	Expert panel	2368	2368	9	Jan 28, 2016
<u>Biesecker Laboratory - ClinSeq Project</u>	Assertion criteria	2215	2215	107	Mar 01, 2016
<u>Sharing Clinical Reports Project</u>	-	2145	2144	2	Jul 28, 2014
<u>Breast Cancer Information Core (BIC)</u>	-	2001	1976	1	Mar 28, 2014
<u>Breast Cancer Information Core (BIC)</u>	-	1792	1758	1	Mar 28, 2014
<u>Tuberous sclerosis database - Tuberous sclerosis 2 (TSC2)</u>	-	1671	0	2	Aug 09, 2013
<u>Retina International</u>	-	1549	0	21	Dec 20, 2012
<u>Cardiovascular Biomedical Research Unit</u>	Assertion criteria	1509	1508	14	Oct 08, 2014

(8) データの登録方法

以下に、ClinVar に対してデータを登録するために必要な要件と手順を示す。

- 登録者

上述の通り、臨床検査施設、研究者、遺伝子検査を提供するクリニック、ゲノム上のバリエーションの位置を取り扱うデータシェアリング基盤、エキスパートパネル、ガイドラインを策定するグループが、ClinVar に対するデータ登録を行えることになっている。

- 登録できるデータ（解釈に関して）

ClinVar が登録を受け付けるデータ（バリエーションについての解釈）は、個人を容易に識別できるレベルのものではない（ある症状を持つ個人毎のものではない）。登録者は、症状を持つ集団と、症状を持たない集団との違いのような集約された形（Aggregate data）か、NIH のガイドラインにそって個人を特定できないようにした上で症状・バリエーションを持つ個人のデータをエビデンスとして登録することが可能とされている。

- 登録できるデータ（バリエーションに関して）

単一のバリエーション、複数のバリエーションで構成されるハプロタイプやヘテロ接合体の組み合わせ（Compound heterozygote）を 1 レコードとして登録可能である。後述する通り、登録はエクセルファイル等を用いて行われるが、1 ファイルに複数のバリエーションを記載することが可能である。ClinVar は、個体レベルのバリエーション群、例えば、ある個体が持つ全バリエーションの登録は受け付けていない⁶³。

- 登録されるデータに含まれる必要がある要素

ClinVar は、ある症状に関連するバリエーションについて、その臨床的あるいは機能的な重要性に関する解釈を登録する必要があるとしている。

- 登録できないデータ

その一方で、GWAS で得られた結果⁶²や、臨床的な重要性が不明であるもの⁶³、プログラムによる臨床的な重要性の判定結果が付与されたバリエーションのデータは受け付けないとしている（ただし、アレル頻度に基づいて Benign と分類されたバリエーションの登録は受け付けている）。

- 登録方法

データ登録のためのポータルサイト⁶⁴を用いる。まず、データの登録者が所属する組織をそのポータルサイトに登録し、その登録が認められた後に、ポータルサイトが利用可能となる。一つのバリエーションに関するデータを登録するだけであれば、ポータルサイトにあるウィザード⁶⁵を用いることができる。複数のバリエーションに関するデータを登録する場合、ClinVar が用意するテンプレートファイル（エクセルファイル）に必要事項を記入してアップロードする。

⁶² dbGaP に登録することになっている。

⁶³ dbSNP、dbVar に登録することになっている。

⁶⁴ <https://submit.ncbi.nlm.nih.gov/clinvar/>

⁶⁵ WEB ページ上で、インタラクティブに値の入力を行える仕組み。

- 登録に用いるファイル

エクセルファイル、タブ区切りもしくはコンマ区切りフォーマットで作成されたファイル、XML フォーマットのファイルでデータ登録が可能である。エクセルファイルに関しては、表 4-16 に示す 2 種類が用意されている。このエクセルファイルは、複数のタブから構成されている。図 4-26 は、ライト版の画面キャプチャーである。ライト版に含まれているタブは、そのファイルに関する説明 (READ_ME)、登録者に関するデータ入力用 (SubmissionInfo)、バリエーションに関するデータ入力用 (Variant)、エビデンスに関するデータ入力用 (ExpEvidence)、以前の登録されたデータの削除用 (Deletes) となっている。図 4-27 は、Variant タブの画面キャプチャーである。各列は、記載項目 (フィールド) となっており、最初の 2 行には、入力値の名称と、それに関する詳細な説明、記載例が示されている。登録者は、3 行目から自身のデータを記載していく (バリエーションについて解釈を、1 行につき 1 つずつ記載していく)。詳細版には、ExpEvidence のタブは含まれていない。その代わりに含まれているタブは、個人毎のデータ入力用 (CaseData)、表現型に関する詳細の入力用 (PhenotypicDetails)、動物実験や In vitro の実験結果の入力用 (FunctionalEvidence) となっている。タブ区切りもしくはコンマ区切りフォーマット及び XML フォーマットでデータを作成する方法は、データ登録に関する説明のページ (<http://www.ncbi.nlm.nih.gov/clinvar/docs/submit/>) に示されている通りであり、XML フォーマットに関しては、そのデータ構造の定義ファイルである XSD も提供されている (ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/clinvar_submission.xsd)。これらのファイルが含む必要がある値、含むことが推奨される値を表 4-17 に示す。

- レビューステータスの獲得方法

登録者は、1 star、3 star、4 star を獲得することができる。その具体的な基準は上述の通りである (表 4-11)。

- 論文発表に先立つアクセッション番号の取得

アクセッション番号は、バリエーションについての個々の解釈に対して付与される。一度に登録された全てのデータを一まとめにして、それらに一つのアクセッション番号を与える仕組みは持っていない (つまり、「登録」という行為に対しては、アクセッション番号が与えられるわけではない)。アクセッション番号を事前に入手するためには、バリエーションに関するデータを記載するためのタブ (Variant) にある“Private comment”という名前の列に、「公表のためにアクセッション番号が必要である」ことを記載する必要がある。“Submission name”という名前の列に、その登録名を記載しておけば、ClinVar のホームページにおいて、その登録名をキーワードに用いて検索することによって、その登録名が与えられたデータ群をヒットさせることが可能となる。また、“Release status”という名前の列に、“Hold until published”と記載しておけば、論文発表までその解釈は公開されない。登録者もしくは論文誌が ClinVar にその論文発表がなされたことを通知するか、ClinVar がその論文発表に気が付く或いはデータ登録の 6 ヶ月後に、そのデータは ClinVar から公開されることになる (ClinVar は、その公開に先立って、データ登録者に事前にその旨を通知する)。

- 登録されたデータの取扱

登録されたデータは、他のデータと統合された形で FTP サイトから公開される (図 4-25 : 上述の通り、RCV レコードとして XML ファイルに格納されて提供される)。登録された個人の年齢や人種の地域 (Geographic origin) は、詳細な情報が特定でき

ないように、曖昧な表現に変換されて公開されることになる。このデータの一部が、VCF やタブ区切りフォーマットのファイルに格納されて FTP サイトからも公開される。新規のバリエーションは、dbSNP もしくは dbVar に登録される。ClinVar に登録されるデータは、NIH's Genomic Data Sharing policy を遵守して取り扱われる。

表 4-16 データ登録に使用されるエクセルファイルのテンプレートの種類

(出所 : <http://www.ncbi.nlm.nih.gov/clinvar/docs/submit/>)

テンプレートファイル名	使用方法
ライト版 (Lite) SubmissionTemplateLite.xlsx	● 最小限の登録 (Minimal submissions) ● 集約されたデータのみ (Aggregate data only) ● HGVS 表記によってバリエーションを記載 (Variants described only with HGVS expressions)
完全版 (Full) SubmissionTemplate.xlsx	● 最小限もしくは詳細な登録 (Minimal or detailed submissions) ● 集約されたデータもしくは個人毎のデータの登録 (Aggregate or case data) ● HGVS 表記以外に、染色体上の位置やバンドによってバリエーションを記載 (Variants described with HGVS expressions, chromosome coordinates, or cytogenetic descriptions)

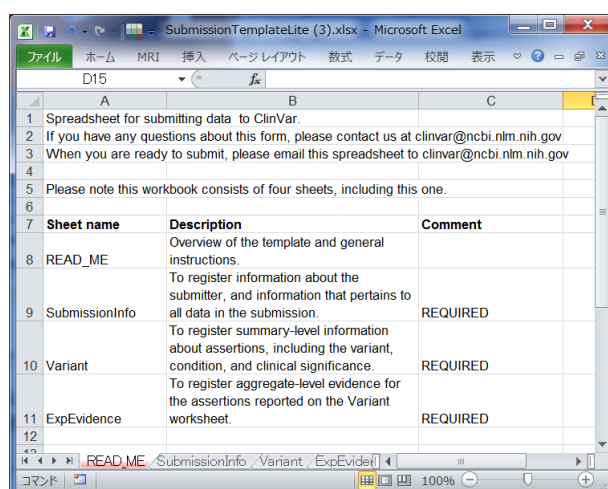


図 4-26 データ登録に使用されるエクセルファイルのテンプレートの構造

(ライト版の場合)

(出所 : <http://www.ncbi.nlm.nih.gov/clinvar/docs/submit/>)

Variant	Condition			Clinical significance	
HGVS name	Condition ID type	Condition ID value	Preferred condition name	Clinical significance	Date last evaluated
Required. The full HGVS nucleotide expression for the variant. (http://www.hgvs.org/mutnomen/examplesDNA.html, http://www.hgvs.org/mutnomen/standards.html). Can also be used for a compound heterozygote or a haplotype, e.g. c.[76A>C;105G>A] or c.[76A>C][105G>A]. For variation represented cytogenetically, enter "cytogenetic" in this column and the description (e.g. 11q14.3 deletion) in the Location column.	Required if "Preferred condition name" is not supplied. Please specify the database for the ID in the Condition ID value column. Allowed values are: OMIM, MeSH, MedGen, UMLS, Orphanet, HPO. If you use a different source, please check with us first.	Required if "Preferred condition name" is not supplied. Please specify the identifier from the database referenced in the Condition ID type column. If there are multiple values, they all need to be from the same source, and separated by a semicolon. See http://www.ncbi.nlm.nih.gov/clinvar/docs/faq_submitters/#examples for examples of identifiers. Values may include both diseases and phenotypes (but not genes).	Required if "Condition ID Type", "Condition ID Value" are not supplied. We strongly encourage describing condition by specific database identifiers. If there is no database identifier for the reported condition, please provide your own term in this column. If you do not wish to make an assertion about condition, please enter "not provided." In cases where the Clinical significance is "benign" and no condition is associated with the variant, please submit "not specified".	Required. Click cell for drop-down list; allowed values include Pathogenic, Likely pathogenic, Uncertain significance, Likely benign, Benign, association, drug response, confers sensitivity, protective, risk factor, other, not provided. If "other" or "drug response", please provide more information in the next column.	Required, if available: date the clinical significance of the variant was last evaluated by the submitter (not the date the patient was evaluated in the clinic). If not available, leave blank. If only month/year is provided we will convert it to the first day of the month. If only year is provided it will be converted to the first day of the first month of the year. Please use the format yyyy-mm-dd.
#Examples are in lines 6-9. Please start your submission in line 10.					
EXAMPLE: NC_000007.13:g.117119262_117119264delCTC			CFTR-related disorders	Uncertain significance	2013/2/1
EXAMPLE: NM_000218.2:c.775C>T			Long QT syndrome, LQT1 subtype	Pathogenic	2013/2/1
EXAMPLE: NM_007294.3:c.2586C>T MedGen		C0346153		Benign	2013/2/1
EXAMPLE: NM_000059.3:c.6238delT MedGen		C0346153		Pathogenic	2013/2/1

図 4-27 データ登録に使用されるエクセルファイルのテンプレートの内容
(ライト版の Variant タブ)

(出所: <http://www.ncbi.nlm.nih.gov/clinvar/docs/submit/>)

表 4-17 登録されるデータを構成する値の概要

(出所: <http://www.ncbi.nlm.nih.gov/clinvar/docs/submit/>)

	入力値の概要
登録すべき	- 登録者 (組織名、人名、この登録に関して公的に貢献を受けるべき者の人名) - バリエントの位置 (HGVS 表記、染色体上の位置による表記、バンドによる表記) - 症状 - 臨床的あるいは機能的な重要度 - データを収集した手法 (臨床検査、研究、等) - バリエントが同定された細胞の種類 (生殖細胞系列か体細胞か) - 疾患発症の有無
登録が推奨される	- バリエントが観察された数 - バリエント・症状の遺伝様式 - 個体レベルの接合性 (遺伝子型がヘテロかホモか) - 家族歴の有無 - 民族、地域

4.4 ClinVar に関連する研究プロジェクト (ClinGen)

本章では、ClinVar に関連するプロジェクトである ClinGen について説明する。本章の内容は以下の通りである。

- ClinGen の目的 (「(1) 目的」)
- ファンドと実施体制 (「(2) ファンド・体制の概要」)
- 各ワーキンググループの目的・実施内容の概要 (「(3) ワーキンググループ」)
- ClinVar と ClinGen におけるデータのやり取り・ClinGen におけるデータ生成 (「(4) ClinVar との連携」)
- 患者 (研究参加者) による直接的なデータ登録 (「(5) GenomeConnect」)

(1) 目的

ClinGen は、臨床医、研究者及び患者が情報を共有し、その知識を体系化することによって患者に提供される医療を向上させることを大きな目的としている⁶⁶。

ClinGen の重要なテーマとして、ClinVar を中心として (図 4-28)、それを中心に流通するデータフォーマットの標準化や、バリエーションと疾患との関係性に関する理解の深化、それらを取り扱うためのソフトウェア (ツールやデータベース) の開発と ClinGen によって整備される情報の普及が挙げられている⁶⁸。具体的なゴールを、以下に示す⁶⁷。

- 臨床及び研究利用のために、臨床医、研究者及び患者が提供するゲノムデータと表現型データを、集中化されたデータベース (Centralized database) によって共有する
- バリエーションに対しての臨床的なアノテーション・解釈を標準化する
- 遺伝子及びバリエーションのキュレーションのために、エビデンスに基づいた専門家による合意形成手段を実装する
- 様々な人々のバリエーションについての理解を深めることによって、世界規模で遺伝子検査に対する理解を深める
- バリエーションを解釈するスループットを上げるために機械学習アルゴリズムを開発する
- 遺伝子及びバリエーションに関する情報が医学的な対処法の判断にどれだけ使えるか (この英語表記は“Medical actionability”である。以下、「アクションナビリティ」という。) を評価する
- 電子医療記録を中心とするエコシステム (EHR ecosystem) においてゲノムに関する知識を活用できるようにするために、その体系化とアクセス手段を提供する
- 上述のコミュニティが、これらの集合知とリソースを制約なく利用できるようにするために普及活動を行う

⁶⁶ <https://clinicalgenome.org/genomeconnect/about-clingen/>

⁶⁷ <https://clinicalgenome.org/about/>

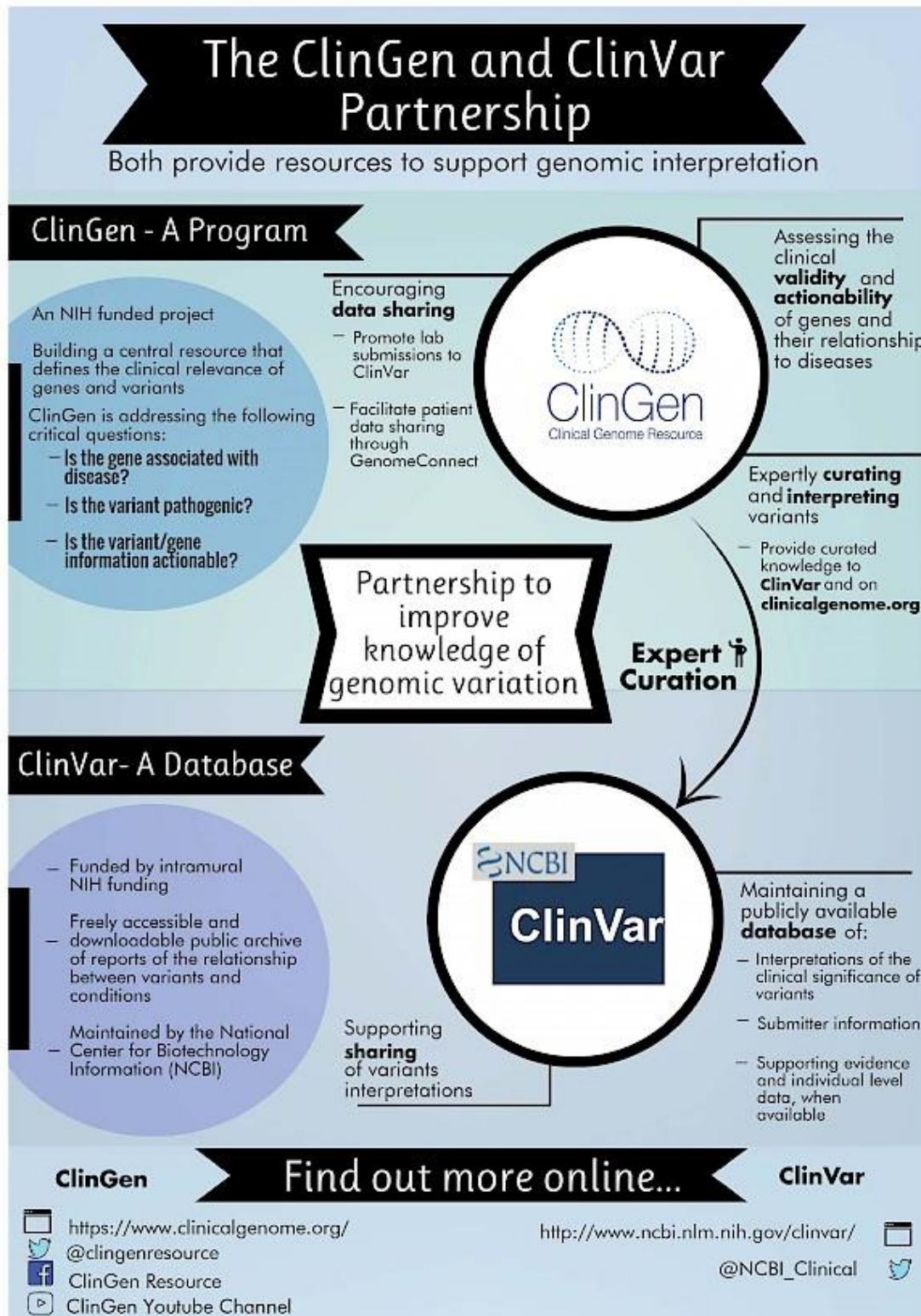


図 4-28 ClinVar と ClinGen との関係性
(出所 : <https://www.clinicalgenome.org/about/about-the-clingen-and-clinvar-partnership/>)

(2) ファンド・体制の概要

ClinGen は、NHGRI と National Institute of Child Health and Human Development (NICHD) の下にある Eunice Kennedy Shriver National Institute からのファンドにより、2013 年 9 月に活動を開始した⁶⁸。現在は、NCI からのファンドも得ている。活動期間は 4 年間となっており、上述の 3 機関からの資金は合計して年間あたり 2 千 5 百万ドルとなっている。ClinGen の下には、複数のワーキンググループ（適宜、再構成されている）、運営委員会、外部パネルが形成されている（総勢で約 200 名の体制となっている）。ClinVar の活動は、上述の ClinVar と密接な関係にあるため、NCBI のメンバーも ClinGen の活動に参加している。

(3) ワーキンググループ

現状では、ワーキンググループは、キュレーション、教育、情報技術、アドバイザリー、ポリシーの 5 つに分類されている⁶⁹。この分類とワーキンググループの関係を以下に示す。

- キュレーション
 - ✓ Actionability
 - ✓ Clinical Domain
 - ✓ Gene Curation
 - ✓ Genomic Variant
 - ✓ Phenotyping
- 教育
 - ✓ Education, Engagement & Counseling
 - ✓ EHR
 - ✓ CADRe (Consent & Disclosure Recommendations)
- 情報技術
 - ✓ IT Standards & Data Submission
 - ✓ Data Model
 - ✓ EHR
 - ✓ Informatics & Analysis
 - ✓ Software Alignment
- アドバイザリー
 - ✓ External Scientific Panel
 - ✓ Steering Committee
- ポリシー
 - ✓ EHR
 - ✓ CADRe (Consent & Disclosure Recommendations)

⁶⁸

[https://www.nih.gov/news-events/news-releases/new-nih-funded-resource-focuses-use-genomic-variants-medical-car](https://www.nih.gov/news-events/news-releases/new-nih-funded-resource-focuses-use-genomic-variants-medical-care)

⁶⁹

<https://clinicalgenome.org/working-groups/>

ここに示されている通り、複数の分類に跨るワーキンググループが存在する。以下に、これらのワーキンググループについて説明する（図 4-29）。External Scientific Panel と Steering Committee は、特別なテーマを持ったグループではないため、ここでは説明を省く。

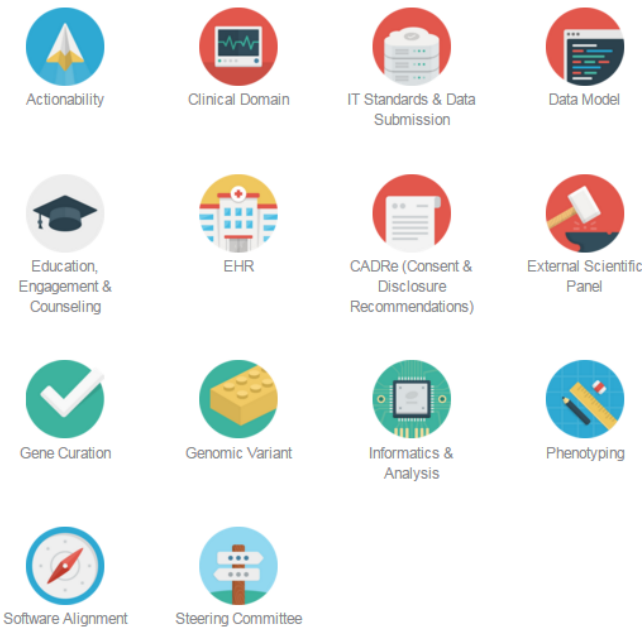


図 4-29 ClinGen のワーキンググループ
(出所：<https://clinicalgenome.org/working-groups/>)

ワーキンググループ名	Actionability
分類	キュレーション
委員長	James P. Evans, MD, PhD Katrina Goddard, PhD
概要	● 重篤な疾患を引き起こす遺伝子であり、さらにその危険性が分かっている場合にその回避あるいは緩和ができる遺伝子を特定することを目的とする。 ● “Clinical actionability”というカテゴリを定義するための厳格かつ標準化された手順を開発し、それに該当する遺伝子と疾患を指定する。さらに、それらのために準定量的な指標を開発する。この活動の結果は ClinGen のホームページから公開されている⁷⁰。

⁷⁰ 遺伝子毎に Actionability や遺伝子と疾患との関係性、Dosage Sensitivity を列挙した一覧が提供されている。(<https://www.clinicalgenome.org/clingen-tools-and-work-products/clingen-curated/>)

ワーキンググループ名	Clinical Domain
分類	キュレーション
委員長	Jonathan Berg, MD, PhD Sharon E. Plon, MD, PhD
概要	● ClinGen の具体的な目標の一つとして、様々な臨床分野で専門家チームを形成・発展させ、遺伝子と疾患の関係や、バリエーションが疾患に対して及ぼす影響に対しての臨床的な妥当性を評価することである。 ● 特定の疾患に対するエキスパートである臨床医、検査室の診断医、研究者から構成されるサブワーキンググループを構成し、その疾患に関わるバリエーションを対象としてアノテーションプロトコルの標準化を行う。サブワーキンググループのメンバーは、コミュニティに含まれている組織の代表者から構成される。以下のサブワーキンググループが存在する。 ✓ 遺伝性がん (Hereditary Cancer) ✓ 循環器系 (Cardiovascular Domain) ✓ ファーマコゲノミクス (Pharmacogenomics) ✓ 先天性の代謝異常 (Inborn Errors of Metabolism) ✓ 体細胞がん (Somatic Cancer)

ワーキンググループ名	Gene Curation
分類	キュレーション
委員長	Jonathan Berg, MD, PhD Christa Lese Martin, PhD, FACMG
概要	● エビデンスに基づいて、遺伝子と疾患との関係性を評価する手法を開発することを目的としている。 ● ある遺伝子によって特定の疾患を引き起こされることを決定づけるために、そのエビデンスを評価するフレームワークを打ち立て、キュレーションの標準化と必要なエビデンスのタイプと量を定める。 ● ClinGen のアプリケーション開発チームに、キュレーションのために用いられるユーザインタフェースの仕様とガイダンスを提供する。 ● さらに、ClinGen におけるキュレーションを監督し、Clinical Domain ワーキンググループの活動を支援する。

ワーキンググループ名	Genomic Variant
分類	キュレーション
委員長	Christa Lese Martin, PhD, FACMG Sharon E. Plon, MD, PhD Heidi Rehm, PhD
概要	● 委員長と、構造多型を含むバリエントに関する研究コミュニティの代表者とが、バリエントに対する解釈とキュレーション方法についてコンセンサスを打ち立てるための活動を行う。特に重要性の高い疾患については、タスクチームを形成して望む（以下のような疾患が対象となっている：Congenital Muscular Dystrophy、Developmental Delay、Rett syndrome、Angelman syndrome、Early infantile epileptic encephalopathy、Mowat-Wilson Syndrome、The RASopathies）。 ● また、ACMG Interpreting Sequence Variant Guidelines⁽⁴⁸⁾を定量的な方法によって補完する事も目標としている。 ● 遺伝性疾患に対するバリエントの影響（Pathogenicity）の度合いを、ガイドラインに沿って人手で導き出すことは労力がかかる作業であるため、その作業を支援するために PATHOGENICITY CALCULATOR 等⁷¹の WEB アプリケーションを開発している。 ● また、Dosage Sensitivity⁷²に関わるバリエントや染色体領域をシステムティックに評価し、評価結果をマイクロアレイの設計や臨床の意思決定に用いることができるように整備する活動を行っている。この活動による情報を、ClinGen Dosage Sensitivity Map から公開している⁷³。

⁷¹ PATHOGENICITY CALCULATOR (<http://calculator.clinicalgenome.org/site/cg-calculator>、<http://calculator.clinicalgenome.org/redmine/attachments/download/252/ASHGUseCaseGLA-2.pdf>)、ALLELE REGISTRY (<http://reg.clinicalgenome.org/site/cg-registry>)

⁷² 2 倍体生物において遺伝子の一方のコピーが機能を損失した場合に表現型に与える影響 (Haploinsufficiency) と、遺伝子のコピーが追加されて 2 つ以上になることによって表現型に与える影響 (Triplosensitivity) (11)

⁷³ <http://www.ncbi.nlm.nih.gov/projects/dbvar/clingen/>

ワーキンググループ名	Phenotyping
分類	キュレーション
委員長	David Miller, MD, PhD
概要	● ClinGen 関連のリソースに対して、表現型データの収集及び登録を支援する。 ● ClinVar に対する疾患及び表現型の登録（標準化された登録）や、Clinical Domain ワーキンググループの各サブドメインに対してオントロジーの選択を支援する。

ワーキンググループ名	Education, Engagement & Counseling
分類	教育
委員長	W. Andrew Faucett, MS, LGC Erin Rooney Riggs, MS, CGC
概要	● ClinGen プロジェクトのあらゆる側面において、教育、アウトリーチ、リソース開発を通じ、コミュニティとの繋がりを育むことを目的として、以下のような活動を行う。 ✓ NCBIやClinGen リソースへのデータ登録に関心を持っている検査室、研究者、臨床医のために、倫理審査委員会に対して申請するためのツールキットの作成・管理 ✓ ゲノム研究コミュニティや世間に対して、ClinGen のイニシアチブや、データシェアリング、その活用、キュレーション方法等をまとめた教育資料の作成 ✓ 患者主導型のデータ登録レジストリである GenomeConnect（後述）の継続的なサポート ✓ ClinGen のホームページや、Twitter や Facebook といったソーシャルメディアを活用したオンライン上のプレゼンス向上

ワーキンググループ名	EHR
分類	教育 情報技術 ポリシー
委員長	Marc S Williams, MD
概要	● Electronic Health Record (EHR) 及びその関連システムの提供者や患者が、ClinGen のリソースに確実にアクセスできるようにするために、ClinGen のリソースを設計する。以下のような活動を行う。 ✓ EHR の有意な利用 (Meaningful use) のステージ 3⁷⁴で提案されている E-resource/Infobutton standards を用いて、EHR から ClinGen のリソースにアクセスする戦略の設計 ✓ 臨床における主要なステークホルダー（その組織化・設計に携わる関係者）を結びつけることによる、ClinGen のリソースが持つ利便性の最適化 ✓ 少なくとも 1 つの EHR システムについては、ClinGen リソース (ClinGen が関わっているデータベース等⁷⁵) へのアクセス戦略を実装 ✓ その実装に関して、他のシステムに対する汎用性の調査 ✓ その戦略で示されている内容を維持するために必要なリソースと、それを実行するために必要な手順についての計画の調整

⁷⁴ The American Reinvestment & Recovery Act (ARRA)による“Health Information Technology for Economic and Clinical Health (HITECH) Act”において、アメリカ国内において相互運用可能 HER の有効利用方法が提案されている。この提案は Centers for Medicare & Medicaid Services (CMS)と、the Office of the National Coordinator for Health IT (ONC)が先導役となっており、EHR の適用のためや、イノベーションのためや、特定のヘルスケアプロバイダーに負荷がかかり過ぎないようにするために、CMS はその専門家や病院に対してインセンティブとして資金提供を行っている。この取組内容は、3つのステージに分けられている (stage1~3)。

(<http://www.cdc.gov/ehrmeaningfuluse/introduction.html>)

⁷⁵ <https://clinicalgenome.org/tools/web-resources/>

ワーキンググループ名	CADRe (Consent & Disclosure Recommendations)
分類	教育 ポリシー
委員長	W. Andrew Faucett, MS, LGC Kelly Ormond, MS, CGC, LGC
概要	● ある遺伝子・バリエントによって引き起こされる疾患の回避法や緩和法等 (Actionability) を患者に報告する場合に関わる倫理的・法的・社会的な問題 (Ethical, Legal, and Social Issues : ELSI) を取り扱う。 ✓ 臨床実装に際しての遺伝子やバリエントを対象とした評価 (評価対象はキュレーション済みの遺伝子やバリエント、臨床実装に際しての ELSI を踏まえた評価を行う) ✓ 評価規定の策定 ✓ 遺伝子やバリエントの種類に応じた最善な方法 (患者に対して検査の前後に行われる教育・遺伝カウンセリングの方法) の模索、 ✓ それらの提案に対する関係者の反応を調べるための調査 (遺伝カウンセリング等の従事者を対象とした調査) を行う。

ワーキンググループ名	IT Standards & Data Submission
分類	情報技術
委員長	Karen Eilbeck, MSc, PhD Melissa Landrum, PhD
概要	● ClinVar のような公共レポジトリへのデータ登録を容易にすることを目的とする。 ● データ登録メカニズムやデータモデルの要件や標準化に対する要求及び利用に関して、Clinical Domain ワーキンググループとのオープンかつ継続的なコミュニケーションを行い、そこからデータのインプットやサンプルに基づいて分析を行い、問題を特定して Data Model ワーキンググループ (後述) 及び ClinVar にフィードバックする (Data Model ワーキンググループとのやり取りは毎月)。 ● さらに、研究機関や Clinical Domain ワーキンググループと、登録プロセスや、データの表現方法の標準化や新規データ項目について議論する。 ● また、研究機関やデータ利用者による多大な要望に答えられるように、NCBI との効率的にコミュニケーションを図れる仕組みを提供する。

ワーキンググループ名	Data Model
分類	情報技術
委員長	Larry Babb Chris Bizon, PhD
概要	● データの構造・内容に関わる ClinGen のワーキンググループ（データを取り扱う全てのワーキンググループ）に対して関連情報を提供するため、概念モデルのコア部分とデータディクショナリを開発する。 ● データ構造の中心的な概念や属性、用語が ClinGen の活動において一貫性が取れるように定義し、さらに、GlinGen から公開される情報の一貫性を担保することを目的する。

ワーキンググループ名	Informatics & Analysis
分類	情報技術
委員長	Carlos Bustamante, PhD
概要	● ClinGen におけるデータの取得・解析・普及のための活動を行う。 ● ClinGen におけるインフォマティクス関連活動のマップ（どのような活動が行われているかを整理したもの⁷⁶）の確定、ClinGenDB（遺伝子・バリエーションに対するキュレーションを行うためのデータベース）及びその関連データソースにアクセスするためのツール（User-facing tool）の設計・開発、ClinVar からのデータ抽出手順やツールの確立、ClinGenDB に対しての同意に基づくデータアクセスレベルの実装、データ解析のための新手法の開発、ケースレベルのデータや個人データを取り扱うためのデータベースの開発、臨床的エビデンスや実験的エビデンスが限られている場合にバリエーションの臨床的な重要性を予測する機械学習手法の開発を行っている。 ● 機械学習に関しては、Gene-Level Modeling として、特定の遺伝子を対象として非同義置換の効果を予測する手法の開発、Genome-Wide Modeling として、ノンコーディング領域やイントロン領域にあるバリエーションや、稀なバリエーション、同義置換を含む効果を予測する手法の開発を行っている。 ● また、Experimental Validation として、CRISPR/Cas9 ベースのゲノム編集技術等の実験的手法によってバリエーションの機能的な影響を予測する手法の開発も行っている。

⁷⁶ <http://interactive.clinicalgenome.org/clingen-resource-system-map/>

ワーキンググループ名	Software Alignment
分類	情報技術
委員長	Selina Dwight, PhD
概要	ClinGen のミッションに関わる活動の効果・効率を最大化することを目的として、各ワーキンググループの目的・ニーズ・ソフトウェア面での挑戦の足並みをそろえるための ClinGen 内におけるコミュニケーションを担う。

(4) ClinVar との連携

ClinVar は、ClinGen の内部的なデータベースである ClinGenDB（遺伝子・バリエントをキュレーションするためのデータベース）とは独立した存在ではあるが、それらの間では頻繁にデータが交換されている。ClinVar は、ClinGen にとって、バリエントに関する 1 次的なデータソースの役割を果たしている（図 4-30）。ClinVar からは、ClinVar に対するデータの登録者や、様々なリンクが付与されたデータを ClinGen に渡している。ClinGenDB からは、キュレーションを受けたバリエントのデータを ClinVar に渡している。

ClinVar/ClinGen System Interactions

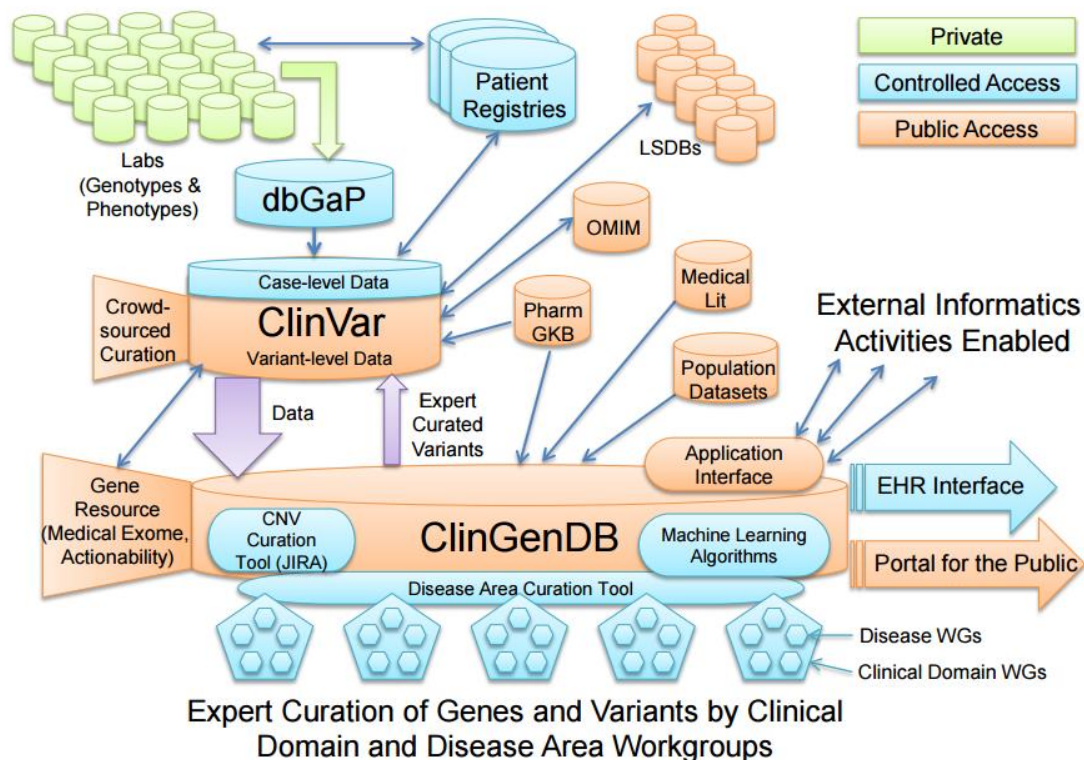


図 4-30 ClinVar と ClinGenDB との間におけるデータの流れ

(出所 : https://www.clinicalgenome.org/site/assets/files/2750/heidi_rehm_clingen_overview.pdf)

ClinVar から ClinGenDB にデータが渡されると、そのデータに係る複数の作業が発生する。遺伝子と疾患との関係性に関しては、キュレーション用のプリケーション⁷⁷を介して、Clinical Domain ワーキンググループと専門家がデータにアクセスして臨床的な妥当性を評価する⁷⁸。その一連の流れを以下に示す。

1. キュレーターがエビデンス（文章化・公開されているもの）を検索して評価する
2. キュレーターが仮の分類（Preliminary classification）を割り当てる
3. ClinGen によって選ばれた専門家によって形成されるグループがこの分類を審査する
4. 審査に基づいて ClinGen としての分類を決定する

この一連の流れにおいて、キュレーターと専門家のグループは、バリエーションに関する情報を得るために様々な情報源を活用する。情報源は、論文検索サイトである PubMed や Google Scholar、バリエーションのデータベースである ClinVar、LOVD⁷⁹、ExAC⁸⁰等、エビデンスとして認められるものである。仮の分類については、キュレーターが以下の 5 つを主要項目として簡潔にまとめる。

Probands	: 遺伝性疾患を有する個体が含まれる家系を見出すきっかけとなった個人（遺伝性疾患を有する個人）の数
Independent Publication Times	: 着目している遺伝子と疾患との関係性についての論文の数 : 着目している遺伝子と疾患との関係性が最初に見出されてからの時間（何年が経っているか）
Experimental evidence points	: 実験的なエビデンス
Contradictory Evidence	: 確認しようとしている遺伝子と疾患との関係性に対して一致しない有力なエビデンスが存在していないか；存在しているのであればその概要

これらの各項目に基づいて、遺伝子と疾患との関係性を採点する。採点には、図 4-31 の表が用いられる。上述した主要項目に基づいてスコアを求め、スコアに基づいて仮の分類（Limited、Moderate、Strong、Definitive）を決定する。

⁷⁷ Pathogenicity Calculator (<http://calculator.clinicalgenome.org/site/cg-calculator>) や Allele Registry (<http://reg.clinicalgenome.org/site/cg-registry>) が含まれる。それらは、Baylor College of Medicine のチームによって開発が進められている Genboree をベースとしたツールである (<http://www.genboree.org/site/>)。現状、開発版が公開されており (http://genboree.org/genboreeKB/genboree_kbs?project_id=acmg-apitest)、ログインせずとも閲覧は可能となっている（編集のためにはアカウントを取得する必要がある）。

⁷⁸ <https://www.youtube.com/watch?v=Q8CB1kKlwF8>

⁷⁹ <http://www.lovd.nl/3.0/home>

⁸⁰ <http://exac.broadinstitute.org/>

Summary Matrix									
Assertion criteria	Description	Number of Points							
		0	1	2	3	4	5	6	7
# Probands	Total # of unrelated probands with variants that provide convincing evidence for disease causality across all curated literature	N/A	1-3	4-6	7-9	10-12	13-15	16-18	19+
Experimental evidence points	Points given based on the gene-level functional evidence supporting a role for this gene in disease	0	1	2	3	4	5	6+	
# Independent Publications	# of Independent publications identifying human variants in the gene in association with disease	N/A	1	2	3	4	5+		
Time (yrs)	# of years since first publication reporting a disease association (If ≥ 2 publications $\rightarrow$ then 1 is max score for time)	this yr	1-3 yr	≥ 3 yr					
Is there valid contradictory evidence?		Y/N?	Classification		Total Score		Assertion:		
Description of Contradictory Evidence:			Limited: Moderate: Strong: Definitive:		0-8 9-12 13-16 17-20				

図 4-31 遺伝子と疾患との関係性を採点するための採点表
(出所: <https://www.youtube.com/watch?v=Q8CB1kKlwF8>)

上述した一連の流れを経て得られた仮の分類に対しては、さらに専門家が審査し、必要であれば修正を加えて、最終的な分類を決定する。その結果は、ClinGen のホームページから閲覧可能となっている (図 4-32、図 4-33)。

Clinical Validity Classification Summaries				
Classifications current as of January 2015				
HGNC Gene Symbol	Disease curated	Orphanet ID	OMIM ID	Clinical Validity Classification
AGTR2	X-linked non-syndromic intellectual disability	ORPHA777	N/A	Contradictory (refuted)
AKAP9	Romano-Ward syndrome	ORPHA101016	611820	Limited
ARSD	Chondrodysplasia punctata	N/A	N/A	No Reported Evidence
C1QB	Immunodeficiency due to an early component of complement deficiency	ORPHA169147	613652	Moderate Strong
CD3E	Severe combined immunodeficiency	ORPHA183660	615615	Strong*
CHD1L	Renal or urinary tract malformation (CAKUT)	ORPHA93545	N/A	Limited
DICER1	Pleuropulmonary blastoma	ORPHA64742	601200	Definitive
FGFR3	Achondroplasia	ORPHA15	100800	Definitive
NGLY1	Alacrima-choreoathetosis-liver dysfunction syndrome	ORPHA04454	615273	Strong Moderate
NHP2	Dyskeratosis Congenita	ORPHA1775	613987	Limited
PALB2	Hereditary breast cancer	ORPHA227535	114480	Definitive
PMS2	Pancreatic cancer	N/A	N/A	No Reported Evidence
RAD51C	Fanconi anemia	ORPHA84	613390	Moderate Limited
RAD51D	Hereditary breast cancer	ORPHA227535	614291	Contradictory (disputed)
RPS10	Diamond-Blackfan anemia	ORPHA124	613308	Limited
RPS24	Diamond-Blackfan anemia	ORPHA124	610629	Moderate
SCN4B	Long QT Syndrome	ORPHA768	611819	Limited
SMAD3	Aneurysm-osteoarthritis syndrome	ORPHA284984	613795	Definitive
TMPO	Familial or idiopathic dilated cardiomyopathy	ORPHA154	613740	Contradictory (disputed)
WRAP53	Dyskeratosis Congenita	ORPHA1775	613988	Limited

図 4-32 ClinGen から提供されている遺伝子と疾患との関係性
(出所: <https://www.youtube.com/watch?v=Q8CB1kKlwF8>)

Gene/Disease Pair: SMAD3/Aneurysm-Osteoarthritis Syndrome

Curator 6

Assertion Criteria	Criteria Description	Number of Points						
		0	1	2	3	4	5	6
# Probands	Total # of curated <i>unrelated</i> probands with variants that provide convincing evidence for disease causality	N/A						6
Experimental Evidence Points	# of points assigned for gene-level experimental evidence supporting a role for this gene in disease							6
# Publications	# of curated independent publications reporting human variants in the gene under consideration	N/A					5	
Time (yrs)	# of years since initial report defining a gene-disease association (If ≤ 2 pubs, then max score for time = 1)			2				
Is there valid contradictory evidence?		No	Classification		Total Score	Calculated Classification		
Description of Contradictory Evidence			Limited	2-8	19	Curator Classification		
			Moderate	9-12				
			Strong	13-16		Definitive		
			Definitive	17-20				

Clinical PMIDs:

21217753
22167769
23272854
22803640
23554019
21778426

Experimental PMIDs:

20506210
21217753
11285272
19478656
23585475
23782924

Experimental PMIDs:

24257415
15623506

図 4-33 遺伝子と疾患との関係性に対する採点結果
(SMAD3 (遺伝子) と Aneurysm-Osteoarthritis Syndrome (疾患))
(出所: <https://www.youtube.com/watch?v=Q8CB1kKlwF8>)

(5) GenomeConnect

ClinGen は、患者が直接的に ClinGen の活動に参画する手段として GenomeConnect と呼ばれる患者向けのポータルサイトを支援している⁽⁴⁷⁾(図 4-34)。このポータルサイトを通じて、患者（以降、研究参加者）は自らの病歴・遺伝子検査結果を ClinGenDB 登録することが可能であり、研究者と自らのデータを安全に共有することが可能となる。ClinGenDB は文献や ClinVar 等からのデータを受け入れているが、GenomeConnect はそれらと並ぶデータソースとして位置づけられる。GenomeConnect の狙いは、他のプロジェクトやレジストリに取って代わるということではなく、既存の枠組みを支援し、発展に導くことにある。

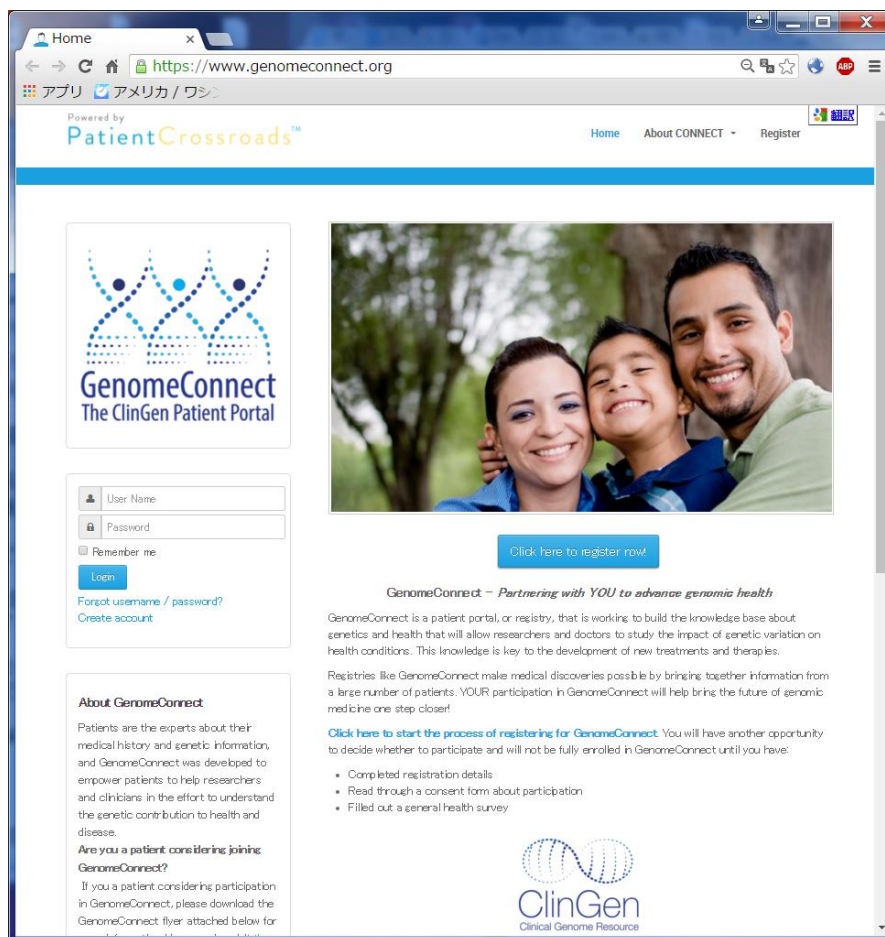


図 4-34 GenomeConnect のトップページ
(出所 : <https://www.genomeconnect.org/>)

GenomeConnect は、2014 年 10 月に発足し、Geisinger Health System⁸¹が保有する倫理審査委員会からの監督を経て、存続・発展している。GenomeConnect は、遺伝子型や表現型のデータを研究参加者から収集し、それらを ClinVar 等のオープンなデータベースの構築に結び付けている。研究参加者は、パソコンやタブレット、スマートフォンを通じて、オンラインでアカウントを作成し、オンライン上でインフォームドコンセントに同意する（10～17 歳の未成年者には親の同意が必要となる）（図 4-35、図 4-36）。同意が成立した後、研究参加者は自らの健康に関しての一般的な質問を受け、自身の遺伝子検査結果が示された PDF を GenomeConnect のポータルにアップロードする。2016 年 1 月の時点で、360 名がその研究参加に同意している。

⁸¹ 米国の非営利医療研究機関 (<https://www.geisinger.org/>)

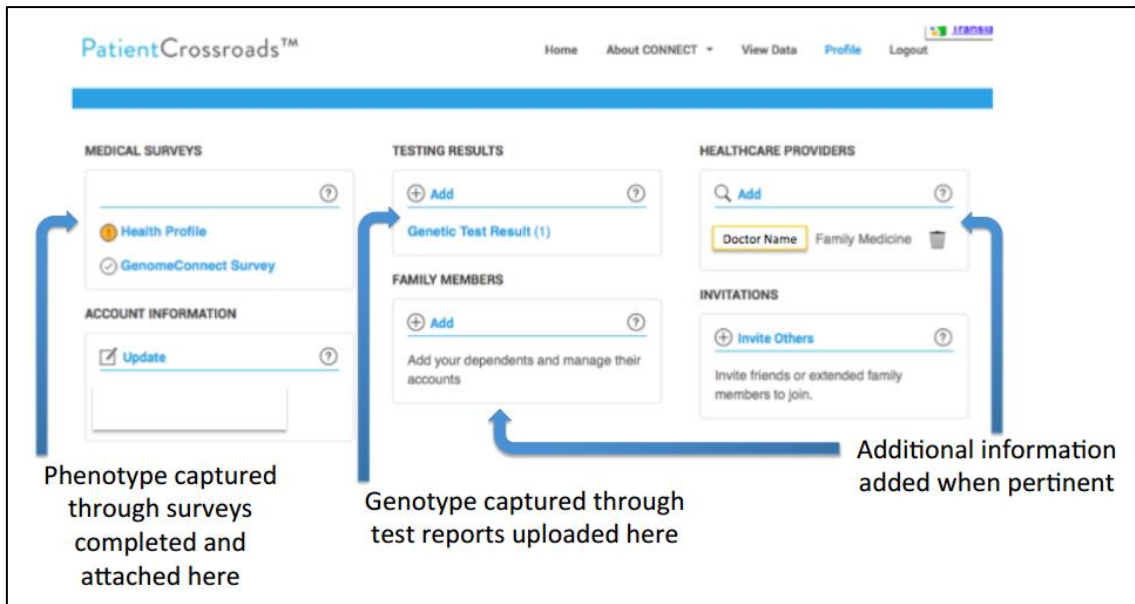


図 4-35 GenomeConnect におけるデータ登録ページ

(出所 : https://www.clinicalgenome.org/site/assets/files/2296/acmg_2015_poster_3_16_15.pdf)

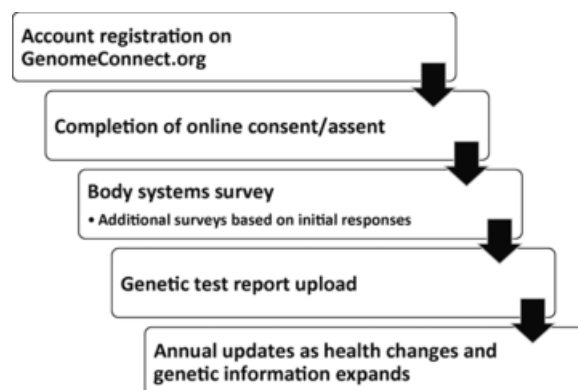


図 4-36 GenomeConnect における研究への参加のフロー

(出所 : <http://onlinelibrary.wiley.com/doi/10.1002/humu.22838/full>)

GenomeConnect に登録される表現型は、Human Phenotype Ontology (HPO)⁸²に基づいて記述されたものである。これにより、研究参加者がその内容を理解しやすいようにし、さらに医師や研究者がそのデータを活用しやすいように均一な表現でのデータ蓄積が可能となっている。また、ClinGen は、GenomeConnect のために、研究参加者が自らの健康記録を登録するためのシステムを開発している。入力を完了させるまでの時間を短縮するために、回答により質問を変更する仕組み (Branching logic) を組み込むことで回答率の向上を図っている。

研究参加者は、GenomeConnect に対して、遺伝型を PDF の形でアップロードし、GenomeConnect の担当者が審査して 1 つのレポートから最大で 10 個のバリエーションを抽出する。認証を受けた遺伝子カウンセラー等の専門家がこうしたキュレーションを行うことで、適切・均一的にデータが共有されることになる。GenomeConnect は、研究参加者のプライバシー保護を重要な優先事項としており、保護すべき情報は研究者と共有せず、

⁸² 症状等を標準化された用語を用いて分類するためのシステム (<http://human-phenotype-ontology.github.io/>)

PatientCrossroads 社⁸³がセキュリティに配慮して保管しておくことになっている。

GenomeConnect によって収集された遺伝子型及び表現型は、匿名化された形で、公的なデータシェアリング基盤で共有される。ClinVar にデータを登録している機関は、ClinVar の WEB ページに列挙されている⁸⁴。研究参加者に対して遺伝子検査を行った機関がそのリストに含まれていれば、GenomeConnect から ClinVar へのデータの登録は行われない。より詳細な遺伝子型・表現型のデータは dbGaP に転送されることになる⁸⁵。将来的には、研究参加者によるゲノムデータの直接的なアップロードや、セキュアなクラウド型サーバが利用できるようになる予定である。さらに、研究参加者に対する検査の解釈や、後向きの分析 (Retrospective analysis) のために、必要であれば、GenomeConnect に登録されていることや、追加的な情報の提供に応じる意志があることが、研究参加者に対する検査を行った機関に通知される。

⁸³ 研究参加者を対象としたレジストリを専門とする企業。同企業は、米国における医療情報の相互運用性と説明責任に関する法律 (Health Information Portability and Accountability Act : HIPAA) が定める最新のルール (Security Rule) を遵守している。また、NIH に対して、連邦政府による情報セキュリティの管理に関する法律 (Federal Information Security Management Act : FISMA) に準拠したプログラムを提供している。
(<https://www.patientcrossroads.com/>)

⁸⁴ <http://www.ncbi.nlm.nih.gov/clinvar/submitters/>

⁸⁵ 研究参加者のデータが GenomeConnect の外部に転送される場合には、その個人の特定に繋がる情報が除去され、再特定可能な ID が割り振られることになる。また、研究参加者が全ゲノム解析やエクソーム解析の結果を保有している場合には、研究者や研究機関が研究参加者に対して利用を申請できるか (つまり、データ利用について同意を提示しているか) が、データの特記事項として加えられる。

5. データシェアリング基盤に関する検討

上述の通り、今後、「ゲノム医療実現推進プラットフォーム事業」と「臨床ゲノム情報統合データベース整備事業」が推し進められることとなっている。本調査では、「4. 章 データシェアリング基盤に関する調査」の調査結果と国内有識者に対するインタビュー結果をもとに、これらの事業に求められるデータシェアリング基盤の機能要件(概念レベルでの要件)と、想定されるデータの流れ、課題採択者・外部委託先のインセンティブ、研究コミュニティ・医療従事者が受ける恩恵について検討・整理した。

本章では、まず、「5.1 章 国内有識者から得られた情報・意見」において、国内有識者から得られた情報と意見をまとめる。次に、「5.2 章 研究のためのデータシェアリング基盤」と「5.3 章 医療のためのデータシェアリング基盤」において、それらの基盤に対しての検討結果をまとめる。

5.1 国内有識者から得られた情報・意見

以下に、国内有識者に対するインタビュー結果をテーマ毎にまとめる。

【データシェアリングの重要性に関して】

- データシェアリングを単なる理念や目標として捉えず、データシェアリングの目的をデータの提供者や利用者の互いが医療上・研究上の実益を受けるような「互惠的利他」という観点から捉えることが重要である。
- 研究資源としての価値や、改正個人情報保護法を含め個人の識別可能性に鑑みて、データの登録・利用に関してのルール作りが大切である。公的資金で収集されたデータの公共性と、データを集積した研究者のエフォートのバランスを勘案することや、アクセス権限のコントロールや、データの不正使用防止策の組み込むことが重要となる。全てのデータを国際的な範囲でデータを無制限に共有する必要はない。発表前のデータであれば、データの共有に先立って、生成者・利用者間での合意が大切である。
- データシェアリングの場が、研究者同士で意見交換する場となる。アイデアがあれば仲間を見つけることができ、さらに、指導も受けられるコミュニティであれば、そこに参加することが、データシェアリングに参加するインセンティブとなる。また、それが次の世代を教育する場にもなる。
- サンプル数が小規模なデータでゲノム研究を行っても得られる成果は限られるため、データシェアリングをしてより大きなデータで研究をできる方が、研究コミュニティとして得られる成果は大きくなる。
- 難病や小児がん、家族性の疾患に関しては、データをできるだけシェアリングして、ある程度の大きいデータを形成しなければ、なかなか成果が得られない。
- 特に大型のファンドによって実施される研究プロジェクトの場合には、データシェアリングが義務化されるべきである。その研究期間が数年間あれば、その間はファンドを提供した側もその研究に寄り添って（モニタリング・チェック・支援）、データシェアリングを促すべきである。

【海外の状況に関して】

- 諸外国においては、遺伝子型と表現型の関連性を整理・検証したデータベースを構築し、臨床応用や、研究の発展のために、医療者・研究者が相互に活用するための仕組み作りが進められている。一方、日本ではそのような仕組み作りがほとんど行われておらず、グランドデザインが求められている。
- がんに係る体細胞変異を収集したデータベース（COSMIC）は存在するが、そこには ClinVar レベルの記述は格納されていない。
- その DB はクリニカルシーケンスの参考になるが、バリエーションの意味付け等、よりクリティカルな情報については Foundation Medicine のような企業が商用データベースとして整備している。
- EGA に制限公開データのアクセス要求を出してもレスポンスが悪いことがある（そのアクセス要求は、DAC（「4.2 章 研究のための既存データシェアリング基盤」を参照）の担当者に対してのメールに依存するため、その担当者がメールを見逃す等によって対応が遅れる可能性がある）。
- 現状、（インタビュー対象者は）EGA にデータを登録しているが、それがスタンダードだからということだけが、その理由である。他のデータシェアリング基盤との比較の上で EGA を選択したわけではない。
- Pan-Cancer Analysis of Whole Genomes（PCAWG）では、3つの Mutation caller を 6,000（～3000 症例）近い全ゲノムデータに対して適用し、それらの評価を行っている。それぞれの Mutation caller について WG が存在し、その検討を行っている。6つ以上のデータセンターが仮想マシン（Virtual Machine：VM）を作り、6,000 近い全ゲノムを共有し、データ解析を行っている。各データセンターが 1PB のディスクスペースを用意したが、それでは十分でなく、Amazon のクラウドも利用している。現状、それでも容量が不足している状況である。

【日本の研究力を確保・向上させるための取組に関して】

- 仮に、日本国内のファンドでデータ生産を実施し、それを ICGC に登録することになったとして、その日本国内のファンドに係るポリシーと、ICGC のポリシーに相違があった場合、日本国内のポリシーを優先させることは可能である。その旨を ICGC に伝え、理解してもらえばよい。実際、ドイツは、自国で得られたデータを海外に出せない事になっており、ICGC においてもそれが尊重されているはずである。
- ゲノム研究の成果を上げるためには、どれだけ大きなグループを作れるか、あるいはグループに加わるかが重要である。そのための国内外のネットワーク作りが必要である。その意味では日本は全般的に遅れを取っている。
- 日本だけでデータシェアリングをしても限界がある。アジアの国々を巻き込むことも一つの方法として考えられる。
- 全てのデータを国際的な範囲でデータを無制限に共有する必要はない。発表前のデータであれば、データの共有に先立って、生成者・利用者間での合意が必要である。
- ゲノムサイエンスが発展途上の分野であることから、全ゲノム情報等 DB の維持・管理には第一線の研究者が関わるのが望ましい。同時に、これらの研究者が一定のエフォートを使って、研究活動が続けることを認め、さらに推奨すべきである。DB 事業を担当する研究者の研究活動は、DB システムの中長期的な継続的な発展と改良の

ために有益であり、また担当者のインセンティブを保つことに繋がる。

- **Mutation calling** の研究開発もアカデミックな研究成果の一部となり得るが、その部分を AMED が行うのであれば、ある程度覚悟を決め、常にアルゴリズムを更新しながら、スピード感を持って研究開発を進めるべきである。それ相応の資金・人員が必要となる。複数の **Mutation caller** 研究開発グループを設け、それらを複数同時に走らせてもよいのではないだろうか。**Mutation caller** の開発は、大きなコンピュータを動かすノウハウや、シーケンス解析の経験、技術的問題点等の知識を有した企業が行ってもよいのではないだろうか。

【データシェアリングのインセンティブに関して】

- ICGC においては、データの登録やそのモラトリアム（猶予期間）に関する取り決めを守っているかをチェックする機構が、論文誌側に実装されている。**Nature**、**Cell**、**Science** といった主要誌においては、その取り決めを守っていない研究成果がリジェクトされる事になっている。
- ICGC は、8～10 ヶ月程度に 1 回ワークショップを開催しているが、ICGC データベースにデータを登録していない研究者は、そのワークショップで発表することができない。このワークショップの聴衆には、エディターやレビュアーがいるため、研究コミュニティに自身の研究成果をいち早く公表できプレゼンスを示すことができる。
- データ登録には手間がかかるため、その手間にもしっかりと対応したことを研究の評価に加える事はよいと考えられる。
- 研究公募の提案時に、データシェアリングを行う意志が盛り込まれていない場合や、それに掛かる費用を積んでいない場合、それを減点してもよいのではないか。

【データシェアリング基盤に備わっているとよい具体的な機能に関して】

- 新たな研究テーマを想起するために、どのようなゲノムデータが存在するかの探索的検討を行うべきである。その探索的検討のための機能は、国全体の研究活動を加速すると期待される。誰が、どのような研究テーマで、どのようなサンプルから、どのようなゲノムデータを生成して、それを登録したかが分かるようにしておくといよい。また潜在的に有用と考えられるゲノムデータについてどのようなインフォームドコンセントが得られているかが確認できるとよい。
- 情報の専門家以外の研究者にとっても使いやすいデータベースである方がありがたい（1つの入り口から様々なデータにアクセスでき、ダウンロードできた方がよい。つまり、情報の専門家以外でも使えるポータルサイトになっている方がよい）。
- 情報の専門家以外には、データ登録は大変な作業である。手続きも大変であり、登録作業が遅れる。
- 大量のデータを登録するにはプログラムを書く必要があり敷居が高い。
- JGA と EGA に互換性があり、JGA にデータを登録すれば論文文化に必要なアクセッション番号を取得できるのであれば便利である。
- EGA のアクセッション番号を JGA で検索できると便利である。
- データ登録を代行してくれるサービスがあればありがたい。
- データ登録を容易にすることを目的とした研究開発のために、データを共有してもよい。

【その他】

- がん研究のコミュニティにおいては、データ登録は当たり前になっている。
- 成人のがんと小児がんとでは事情が異なるため、異なる研究領域として分けて考える必要がある（成人のがんは比較的に病態がはっきりしているが、小児がんはそうではないことが多い）。
- 特に、症例が少ない疾患においては、データシェアリングの必要性が高い。
- 体細胞変異は、Mutation caller によって得られる結果が異なる傾向が高い。つまり、体細胞変異に着目するがん研究の場合、生殖細胞系列バリエーションに着目する研究に比べて、解析に用いる Mutation caller によって得られる結果に差異が出てくる。がん以外の多くの疾患に対しては、生殖細胞系列バリエーション用のパイプラインを用いることができると考えられる。

5.2 研究のためのデータシェアリング基盤

本章には、「ゲノム医療実現推進プラットフォーム事業」に求められるデータシェアリング基盤に関しての検討結果を示す。

図 5-1 は、その検討結果をまとめた図である。この図にある通り、この基盤の目的は、公募の課題採択者が「迅速なデータシェアリングを実施することにより研究開発を促進すること」であり、課題採択者に対しては「解析後すぐに全ゲノム情報等を AMED 支援機能に登録」することを期待するものである。そのために、課題採択者を支援する様々な機能や、課題採択者がデータを登録するための様々なインセンティブが必要であると考えられる。以下に、その機能やインセンティブについての検討結果を整理する。ここでは、このためのデータシェアリング基盤を「全ゲノム情報等 DB」ということにする。

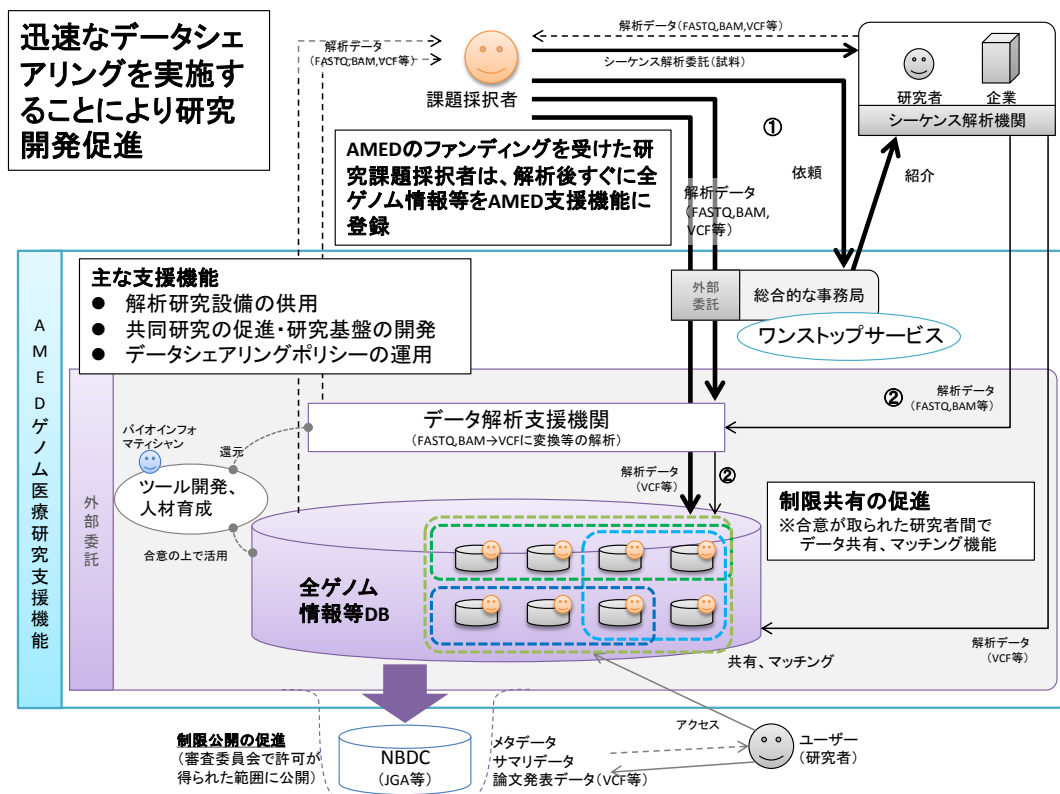


図 5-1 「ゲノム医療実現推進プラットフォーム事業」に関する検討

(1) 全ゲノム情報等 DB に求められる機能

以下に、全ゲノム情報等 DB に求められる機能を整理する。

● 総合的な事務局（ワンストップサービス）

バリエーションと疾患との関係性を明らかにするといったゲノム研究の主目的に取り組むためには、ゲノム配列決定や、バリエーションの同定といった基本的な解析を行う必要があり、そのためにはシーケンサーや大規模な計算機が必要である。あらゆる研究者にとってそれらを十分に確保することが困難であるという現状がある。そこで、全ゲノム情報等 DB がその不足分を補うことができれば、多くの課題採択者にとって、全ゲノム情報等 DB に対してデータを登録するインセンティブになると考えられる。図中の①は、課題採択者からのデータ、試料あるいは依頼を表している。この「依頼」は、課題採択者が「総合的な事務局」に対して、シーケンス解析機関の紹介あるいはデータ解析支援機関への解析を依頼する流れである。このような課題採択者からの依頼を、総合的な事務局が一元的に捌くことによって、つまりワンストップサービスを実現させることによって、全体的なコストの削減とスムーズなサービスが提供できると考えられる。この総合的な事務局の運営は、AMED からの外部委託を受けた機関が行う。以下に、課代採択者のタイプと、タイプ別に求められるサービスを整理する。図 5-1 の②はシーケンス解析機関及びデータ解析支援機関からのデータの流れであり、データの登録先が全ゲノム情報等 DB となっている。

表 5-1 課題採択者のタイプとタイプ別に求められるサービス

課題採択者のタイプ	求められるサービス
十分な配列決定能力を有していない課題採択者	十分な配列決定能力を有している機関（図中の「シーケンス解析機関」）の紹介
十分なデータ解析決定能力を有していない課題採択者	データ解析支援機能によるデータ解析（FASTQ や BAM から VCF への変換、等）
十分な配列決定能力・データ解析決定能力を有している課題採択者	全ゲノム情報等 DB への直接的なデータ登録

- データ解析支援機能

課題採択者が保有しているデータを、共有に適した形に変換する作業を代行する機能である。全ゲノム情報等 DB では、バリエントのデータが中心的な共有対象となるため、リードやアライメント結果からバリエントを同定する解析（つまり、FASTQ や BAM から VCF への変換）は、多くの課題採択者にとって有用であると考えられる。課題採択者からのニーズに応じて解析パイプラインを用意し、それを用いた作業支援を行う。

- 制限共有機能

全ゲノム情報等 DB にデータを登録した課題採択者を中心に、全ゲノム情報等 DB に登録したデータを共有する機能である。課題採択者以外であっても、AMED 及びデータ登録者の承認が得られれば、データの共有が認められる（そのためには、データの提供など、ゲノム医療実現推進プラットフォーム事業に対する価値提供が求められる）。この機能を有効化するためには、全ゲノム情報等 DB 内において、ある課題採択者によって登録されたデータが、別の課題登録者によって登録されたデータと関係していることが分かるようにする必要がある。また、AMED が、データを登録した課題採択者同士、新たな共同研究を育めるように支援するマッチング機能（検索機能や、登録されているデータの概要一覧等）を持つことも、その有効化に繋がると考えられる。

- ツール開発・人材育成

全ゲノム情報等 DB の構築・運営を行う機関（AMED からの外部委託を受けた機関）や課題採択者、日本の研究コミュニティにおいてそれらのテーマに関する第一人者が、ゲノム医療実現推進プラットフォーム事業の下にワーキンググループを形成し、研究開発に関する方向性や実施内容を議論し、議論に従ってその整備を進める。この研究開発には、全ゲノム情報等 DB に登録されたデータを活用できることとする（データの活用に関しては、データの登録者にその了解を得ることとする）。例えば、生殖細胞系列から得られた配列データに対するバリエントの同手法は世界的にも標準化されつつあるが、体細胞から得られた配列データに対するバリエントの同手法はまだ研究の途上であるため、この後者に対するツール開発が、最初のテーマ候補になると考えられる。ただし、HudsonAlpha のように解析速度面において GATK を凌駕するシステムを開発する動きもある。このツール開発や、データ解析支援機能の体制に、若手のバイオインフォマティシャンを加える事でそれらの人材育成にも繋がると考えられる。

(2) 全ゲノム情報等 DB に求められるインセンティブ

以下に、課題採択者、外部委託先（全ゲノム情報等 DB の構築・運営を行う機関、総合的な事務局）のインセンティブと、研究コミュニティが受けるベネフィットをまとめる。

- 課題採択者のインセンティブ
 - ✓ 電算資源（スパコン等）が利用可能
 - ✓ 情報解析サービスが利用可能（標準化パイプライン（FASTQ あるいは BAM→VCF）の提供を検討）
 - ✓ 共同研究の促進・拡大（データ提供者と利用者との合意形成支援、マッチング機能）
※データ利用における裁量の保証（課題採択者にある）
 - ✓ 日本人標準ゲノム参照配列・頻度情報が利用可能（公募採択時に協力推奨）
 - ✓ アクセションナンバーの取得に当たっての負担軽減（論文文化支援）
- 外部委託先のインセンティブ
 - ✓ 研究開発の実施（登録データを用いたツール開発）
 - ✓ 電算資源（スパコン等）の拡充
 - ✓ バイオインフォマティクス等の人材の定着
- 研究コミュニティとしてのベネフィット
 - ✓ オープン・制限公開データ等を介した研究の促進（NBDC（JGA 等）を介して）
 - ✓ インフォマティクス等によって得られる成果のナレッジシェアリング

5.3 医療のためのデータシェアリング基盤

本章には、「臨床ゲノム情報統合データベース整備事業」に求められるデータシェアリング基盤に関しての検討結果を示す。図 5-2 は、その検討結果をまとめた図である。この図にある通り、この基盤の目的は、クリニカルシーケンスによって得られたバリエーション情報及び臨床情報を医療で活用するために、医療・研究コミュニティでそれらの情報を含むデータを共有することである。これに登録されるデータは、最終的には、公開して問題のないと考えられるデータが非制限公開されることとなる。この基盤に登録されるデータは、研究的な価値よりも医学的な価値（遺伝子検査で参照できるという価値）が高いものであり、研究者にとっては、登録のためのデータセットを作成する事に、必ずしも強い動機があるとは考えにくい。そのために、「全ゲノム情報等 DB」の場合と同様に、課題採択者を支援する様々な機能や、課題採択者がそのような登録をするための様々なインセンティブが必要であると考えられる。以下に、その機能やインセンティブについての検討結果を整理する。ここでは、このためのデータシェアリング基盤を「臨床ゲノム情報統合 DB」ということにする。

アント毎やバリエーション-表現型の関係性毎に整理する機能（図 4-16）や、キュレーションレベルを計算する機能（表 4-11）、登録者の一覧を外部公開する機能（表 4-15）が求められる。

（2）臨床ゲノム情報統合 DB に求められるインセンティブ

以下に、課題採択者、外部委託先（臨床ゲノム情報統合 DB の構築・運営を行う機関）のインセンティブと、研究コミュニティが受ける恩恵をまとめる。

- 課題採択者のインセンティブ
 - ✓ データ登録をした研究者は次の公募採択時に加点される等の優遇処置
 - ✓ 登録者・登録関係者ごとの登録数の公表（貢献度の公表）
 - ✓ 比較的簡便な登録方法の提供
- 外部委託先のインセンティブ
 - ✓ エキスパートパネルや WG のメンバーを公表
 - メンバーは世の中に専門家として認知される
 - 外部委託先は、エキスパートパネルや WG のハブとなり、情報流通の中心となる
- 医療従事者のインセンティブ
 - ✓ わが国におけるゲノム医療実装のための基盤
 - 自らが専門とする疾患以外を対象にコミュニティーサービスとして遺伝子検査を行っている医師に対して、その専門外の検査に関する情報を提供することで、その労力を削減する利点がある
 - 病的バリエーションのデータベースの充実は、アカデミアに加えて、臨床検査会社による遺伝子診断業務を支援する基盤になると考えられる

6. まとめ

今後、AMED は「ゲノム医療実現推進プラットフォーム事業」及び「臨床ゲノム情報統合データベース整備事業」といった事業をはじめとして、データシェアリングのための基盤及びポリシーの整備・施行を行うこととなっている。本調査（「研究・医療面において有用なデータシェアリング基盤の検討」）では、その整備・施行を支援することを目的として、欧米におけるデータシェアリング基盤及びポリシーについての調査を行った。

本調査では、欧米におけるゲノム医療に関わるデータシェアリング基盤及びポリシーを対象として、それらで取り扱われるデータの種類やフォーマット、公開メカニズム（制限公開／非制限公開）、データの登録手順や、登録までの期間、制限公開データに対する利用申請手順について整理した。NIH や Wellcome Trust による制限公開（dbGaP や EGA をその基盤とする制限公開）では、データが生成されてから短い期間の内にそれらが公開されることになっており、さらに、世界中のあらゆる研究者・機関が利用を申請可能であるため、研究のためのインフラと人材を揃えている研究グループ（欧米に存在する一部のグループ）にとっては有利に働く場合があることが国内外の有識者から指摘されている。その一方で、DECIPHER では、研究者同士の同意に基づいて、それらの間でのみデータを共有できる基盤が構築・運用されている。MatchMaker Exchange は、その取組を複数の基盤を横断する形に拡張したものであり、合意形成の方法だけでなく、類似する研究参加者を検索する手法（研究参加者が保有しているバリエーション群や表現型で類似度を計算する手法）や、インターネット越しにその検索・合意形成を行うための API の開発も行っている。また、GeCIP では、研究対象とする疾患の種類（がんや希少疾患の種類毎）や、研究テーマ（バリエーションの解釈と研究参加者へのフィードバックや ELSI 等）毎にドメインが形成され、そこに所属する目的を同じくする研究者（主に英国の研究者）の間で、データ生成後の一定期間は 100K Genome Project のデータを共有できることとしている。DECIPHER や MatchMaker Exchange、GeCIP に実装されているデータを共有するためのメカニズムは、研究目的を同じくする研究者・研究グループの間で、自らの意思に基づいてデータを共有する相手を決め、その相手との間でデータの価値を確保するためのものである。GeCIP では、さらに、国内の研究者を保護する仕組みがそこに組み込まれている。これらのような制限的にデータを共有する方法（制限共有）は、日本の研究コミュニティにとっての重要なオプションであると考えられる。

この調査結果を受け、「ゲノム医療実現推進プラットフォーム事業」及び「臨床ゲノム情報統合データベース整備事業」で整備されることになる基盤の要件、課題採択者及び外部委託先のインセンティブ、研究コミュニティや医療従事者が受ける恩恵について、洗い出して整理し、AMED によるデータシェアリングポリシーの作成に対して支援を行った。NIH や Wellcome Trust といった欧米の主要なファンディングエージェンシーや NBDC のガイドライン等、AMED によるポリシーに関連する情報を収集し、それらから AMED によるポリシーの関連箇所を抽出した。それらの材料等を踏まえた検討の経緯や、それらの材料に対するインデックス（どの資料の何ページ目に、AMED によるポリシーの関連箇所があるか）や運用のための注意点を整理することで、AMED によるデータシェアリングポリシーの作成に寄与できた。

参考文献

- (1) National Institutes of Health Genomic Data Sharing Policy、NIH、(2014)
- (2) Supplemental Information to the National Institutes of Health Genomic Data Sharing Policy、NIH、(2014)
- (3) データシェアリング関係ポリシー一覧、NIH、
https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_policies.html
- (4) Final NIH Genomic Data Sharing Policy、NIH、(2014)
- (5) Example Plan addressing Key Elements for a Data Sharing Plan under NIH Extramural Support、NIH、http://grants.nih.gov/grants/sharing_example_data_sharing_plan.doc
- (6) NIH's genomic data sharing policy: timing and tradeoffs、Trends in Genetics、Jorge L. Contreras、(2014)
- (7) Prepublication data release, latency, and genome commons、Science、Jorge L. Contreras、(2010)
- (8) NCI Grant Policies、NCI、
<http://www.cancer.gov/grants-training/grants-management/nci-policies>
- (9) Genomic Data Sharing (GDS) Policy、NCI、
<http://www.cancer.gov/grants-training/grants-management/nci-policies/genomic-data>
- (10) The National Cancer Institute Policy Ensuring Public Availability of Results from NCI-supported Clinical Trials、NCI、
<http://grants.nih.gov/grants/guide/notice-files/NOT-CA-15-011.html>
- (11) Data management and sharing、Wellcome Trust、
<http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/Data-management-and-sharing/index.htm>、(2010)
- (12) Policy on data management and sharing、Wellcome Trust、
<http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTX035043.htm>、(2010)
- (13) Q&A: Wellcome Trust policy on data management and sharing、Wellcome Trust、
<http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/Data-management-and-sharing/WTX035045.htm>
- (14) Guidance for researchers: Developing a data management and sharing plan、Wellcome Trust、
<http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTD002753.htm>、(2012)
- (15) Policy on intellectual property and patenting、Wellcome Trust、
<http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTD002762.htm>
- (16) Sharing Data from Large-scale Biological Research Projects: A System of Tripartite Responsibility (the Fort Lauderdale Principles)、Wellcome Trust、(2003)
- (17) Prepublication data sharing (the Toronto statement)、Nature、Toronto International Data Release Workshop Authors、(2009)
- (18) Large-scale genetics research、Wellcome Trust、
<http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/Large-scale-genetics-research/index.htm>

- (19) Statement on genome data release (the Bermuda principles)、Wellcome Trust、
<http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WT002751.htm>
- (20) DATA SHARING POLICY、Wellcome Trust Sanger Institute、(2015)
- (21) Goals, Structure, Policies and Guidelines、ICGC、
<https://icgc.org/icgc/goals-structure-policies-guidelines>
- (22) E. Consortium Policies and Guidelines、ICGC、
<https://icgc.org/icgc/goals-structure-policies-guidelines/e-consortium-policies-and-guidelines>
- (23) E.1 Informed Consent, Access and Ethical Oversight、ICGC、
<https://icgc.org/icgc/goals-structure-policies-guidelines/e1-informed-consent-access-and-ethical-oversight>、(2012)
- (24) E.2 Data Release Policies、ICGC、
<https://icgc.org/icgc/goals-structure-policies-guidelines/e2-data-release-policies>
- (25) E.3 Publication Policy、ICGC、
<https://icgc.org/icgc/goals-structure-policies-guidelines/e3-publication-policy>、(2011)
- (26) E.4 Intellectual Property Policy、ICGC、
<https://icgc.org/icgc/goals-structure-policies-guidelines/e4-intellectual-property-policy>
- (27) E.9 Data Management、ICGC、
<https://icgc.org/icgc/goals-structure-policies-guidelines/e9-data-management>
- (28) 第5期科学技術基本計画、首相官邸 閣議決定、(2016)
- (29) 中間とりまとめ、首相官邸健康・医療戦略推進本部 ゲノム医療実現推進協議会、(2015)
- (30) 研究データ共有ポリシーに関する調査、科学技術振興機構 情報企画部、(2015)
- (31) Office of Science and Technology Policy ホームページ、White House、
<https://www.whitehouse.gov/administration/eop/ostp>
- (32) OSTP Increasing Access to the Results of Federally Funded Scientific Research、White House、(2013)
- (33) RCUK Common Principles on Data Policy、Research Councils UK、(2015)
- (34) P-DIRECT データ共有ガイドライン、文部科学省、(2015)
- (35) NBDC ヒトデータ共有ガイドライン ver. 2.0、科学技術振興機構 バイオサイエンスデータベースセンター、(2015)
- (36) バイオバンク・ジャパン説明同意文書、東京大学医科学研究所 バイオバンク・ジャパン、
https://biobankjp.org/public/pdf/panf_v4.pdf
- (37) 東北メディカル・メガバンク機構地域住民コホートの同意文書、東北大学東北メディカル・メガバンク機構、
http://www.megabank.tohoku.ac.jp/chco/rinrimen/pdf/doi_setsumei_20150929.pdf
- (38) 平成 28 年度 医療分野の研究開発関連予算の概算要求のポイント、首相官邸 健康・医療戦略推進本部、(2015)
- (39) Facilitating collaboration in rare genetic disorders through effective matchmaking in DECIPHER、Human Mutation、Chatzimichali EA et al.、(2015)
- (40) DECIPHER: database for the interpretation of phenotype-linked plausibly pathogenic sequence and copy-number variation、Nucleic Acids Res.、Bragin E et al.、(2014)
- (41) The Matchmaker Exchange API: automating patient matching through the exchange of structured phenotypic and genotypic profiles.、Hum Mutat.、Buske OJ et al.、(2015)

- (42) Genomics England Clinical Interpretation Partnership (GeCIP), Genomics England, (2014)
- (43) ClinVar: public archive of relationships among sequence variation and human phenotype, Nucleic Acids Res., Landrum MJ et al., (2014)
- (44) ClinVar: A Central Repository for Interpretations of Clinically Relevant Variants, Human Variome Project, Melissa J Landrum, (2014)
- (45) Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology., Genetics in Medicine, Richards S et al., (2010)
- (46) ClinGen: The Clinical Genome Resource - Heidi Rehm, Human Variome Project, Heidi Rehm, (2014)
- (47) GenomeConnect: Matchmaking Between Patients, Clinical Laboratories, and Researchers to Improve Genomic Knowledge, Human Mutation, Kirkpatrick BE et al., (2015)
- (48) Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology., Genetics in Medicine, Richards S et al., (2015)
- (49) Cumulative Haploinsufficiency and Triplosensitivity Drive Aneuploidy Patterns to Shape the Cancer Genome, Cell, Teresa Davoli, (2014)
- (50) ClinVar: public archive of interpretations of clinically relevant variants., Nucleic Acids Research, Landrum MJ et al., (2016)

平成 27 年度 国内外における遺伝子診療の実態調査報告書

2016 年 3 月

株式会社 三菱総合研究所
人間・生活研究本部
科学・安全政策研究本部
経営コンサルティング本部
金融イノベーション事業本部

TEL (03)6705-6025