

ゲノム医療推進に向けた試験的運用・調査

- 欧州・米国におけるデータシェアリング
- データシェアリングの動向 — ClinVar/ClinGenについて —
- 参考資料

株式会社 三菱総合研究所

欧州・米国におけるデータシェアリング

【コンテンツ】

- ① 欧州・米国におけるデータシェアリング基盤の位置づけ
 - データシェアリング基盤が取り扱う内容、非制限公開/制限公開
 - ゲノム医療研究の過程(ステージ)における各データシェアリング基盤の役割と関係性
- ② 欧州・米国における制限公開データ提供の手続き
 - 利用申請の審査体制、審査方法、提供方法、ファンドとの関係性
- ③ まとめ・論点整理

これ以降のページに有るNCBI及びEBIのDBのロゴは、NCBI及びEBIのHPからの出典
出典)NCBIのロゴ: <http://www.ncbi.nlm.nih.gov/>の配下にあるHP

EBIのロゴ: <http://www.ebi.ac.uk/>の配下にあるHP

① 欧州・米国における主要データシェアリング基盤の位置付け(1/2)

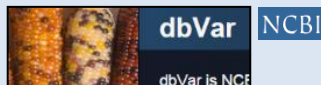
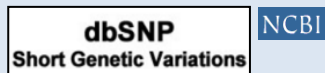
変異

集団中に認められた変異を列挙
(個人別には集積せず)

個人の持つ
変異(複数)を
表現型とともに
個人別に集積

NGSデータ

変異を集積
(病的意義は
不問)



小さな多型・変異のアレル・
ゲノム配列上位置



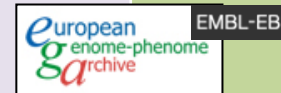
構造多型・変異のアレル・
ゲノム配列上位置



サンプル毎の遺伝子
型とそのサンプルの
表現型、統計解析結
果等(GWAS等の入
力・解析結果を含む)



NGSによって生成された
DNA配列データ(リード、
マッピング結果)

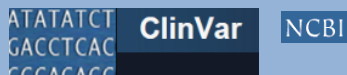


(EGAは、サンプル毎の
遺伝子型とそのサンプ
ルの表現型の他に、
DNA配列データや発現
データ等も取り扱う)



各種DNA配列データ(キャ
ピラリーシーケンサー・
NGSによって生成された
DNA配列データ、アセンブ
ル結果)

病的意義のある
変異のみを
集積



病気に係る変異(主に単一遺伝子病)、その表現型・エビデンスの集約結果
(現状、GWASによって得られたレビューを受けていない変異は含まれない)



Clinical Studies ClinVarに対応するものとして計画されている

制限公開の
区分あり

NCBI NCBIは、Public Law 100-607で、NIHの付属機関として設置される

EMBL-EBI EBIは、イギリス政府とEMBLとの協定*に基いて設立される

* Agreement between the Government of the United Kingdom of Great Britain and Northern Ireland and the European Molecular Biology Laboratory Concerning the European Bioinformatics Institute)

① 欧州・米国におけるデータシェアリング基盤の位置づけ(2/2)

医療価値が高いデータ(遺伝子変異がほぼ症状を決定する遺伝子に関する変異を集約・整理して提供; 非制限公開)

	欧州 (EBI)	米国 (NCBI)
解析結果・エビデンス ・変異と表現型との関係性 ・医学的な分類や、その確からしさ、そのサポートデータ、データ登録者	Clinical Studies	

■ 個別研究成果のデータシェアリング基盤からのデータ登録

■ CLIAあるいはISO 15189といった基準を満たしている臨床検査室、
■ 研究、■ 文献からのデータ登録

個別研究成果(生成されたデータ・解析結果の蓄積・提供; 非制限公開/制限公開)

	欧州 (EBI)	米国 (NCBI)
遺伝子型・表現型 ・サンプル毎の遺伝子型とその表現型 ・それらに対する統計解析結果(多型・変異と表現型との関係性等)		

ここに格納されているデータを共有すれば、更に大規模な研究や、着目点を変えた研究が可能

■ 研究者・研究チームからのデータ登録(共通基盤を活用した研究成果)

- ・ 計測プラットフォームの開発(SNPデータを用いた遺伝子型判定のためのマイクロアレイの設計等)
- ・ GWAS、リファレンスゲノムを用いたWES/WGS
- ・ コホート研究、ケースコントロール研究

共通基盤(リファレンスとして用いられるデータの蓄積・提供; 非制限公開)

	欧州 (EBI)	米国 (NCBI)
多型・変異 ・多型・変異のデータ(ゲノム中の位置やアレル、頻度情報、遺伝子型判定のリファレンスとして使われる)	EVA  	dbSNP Short Genetic Variations 

これらの他に、リファレンスゲノムや遺伝子構造アノテーションのデータベースも存在

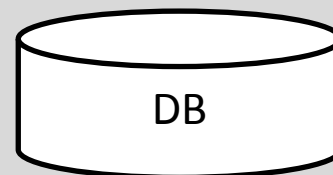
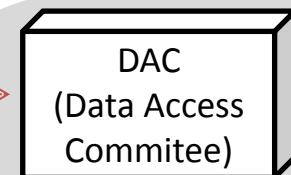
② 欧州・米国における制限公開データ提供の手続き(1/2)

制限公開データ利用の基本的な流れ



利用者

①利用申請



②承認→利用可能

欧州(EU), EGA

【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
- それは、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

【構築】

- DACを構築するためには、EGAへのデータ登録過程において、その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
 - ・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化していいか(embagoe)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する方法を示したドキュメント

【運用】

- 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ウ権限の設定が可能である

米国, dbGaP

【形態】

- 上級連邦職員(senior Federal employees)で構成される
- それらは、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達である

【構築】

- そのメンバーの任期は、それらを招集した機関(つまり、DACが設置される機関)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、ミーティングへの参加や、コンサルテーションのために招集することが可能(連邦職員であってもなくてもよいが、DACのメンバー選定に関わってはならない)

【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法を含めてはいけけない
- 利用申請は、公開されているウェブページで見つけられることが可能でなければならない(must be identified)
- DACは、NIH職員による計画的な監査(programmatic oversight)の申請を審査し、承認・否認する
- dbGaPへのアクセス申請に対しての応答時間には幅が有り得るが、全てのNIH DACはできるだけタイムリーにその応答をするように務めるべき

② 欧州・米国における制限公開データ提供の手続き(2/2)

欧州(EU), EGA

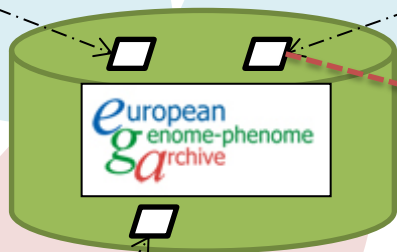
—分散型—

Wellcome Trustによるファンドを受けて生成されたデータ

データ提供体制A



データ提供体制B

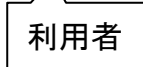


登録



データ提供体制C

その他ファンドを受けて生成されたデータ



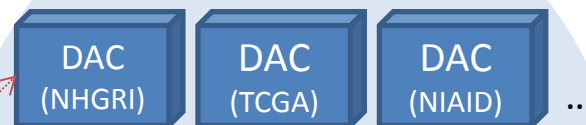
利用者は、①まずデータ生産者が所属する組織(コンソーシアム等を含む)が保有するDACに利用申請を出し、②それが認められた場合に利用可能となる。

米国, dbGaP

—中央集権型—

NIH等によるファンドを受けて生成されたデータ

複数のDACを含むデータ提供体制

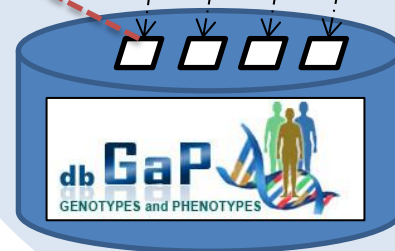


①

②



登録



利用者は、①まずDACに利用申請を出し、②それが認められた場合に利用可能となる(審査を担当するDACは、データの登録に関わった組織のDAC)。

まとめ

1. ゲノム医療に係るデータシェアリング基盤を大枠で整理
 - 集団中に認められた変異を列挙したもの
 - ・ 変異を集積したもの (dbSNP, EVA等)
 - ・ 病的意義のある変異のみを集積したもの (ClinVar等)
 - 個人の持つ変異を表現型とともに個人別に集積したもの (dbGaP, EGA等)
2. dbGaP, EGA等には制限公開の区分がある
 - 米国は中央集権型 (登録者は利用申請・承認に関知せず)
 - 欧州は分散型 (登録者は利用申請・承認に関知)

論点

1. 日本で整備すべきデータシェアリング基盤
2. 制限公開の方法 (中央集権型、分散型、その他?)

データシェアリングの動向

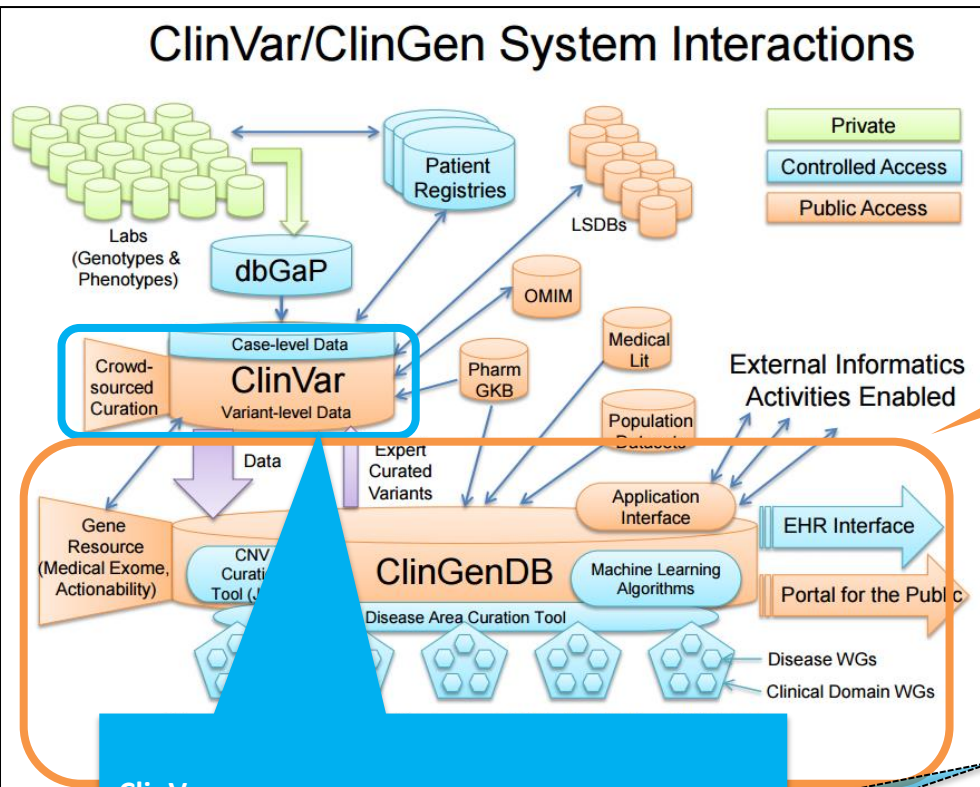
— ClinVar/ClinGenについて —

【コンテンツ】

- ① ClinVar/ClinGenの概要
それぞれの実施内容の概要、実施主体、予算規模、人員規模
- ② ClinVarとClinGenの関係性
それぞれの役割、関係性
- ③ まとめ・論点整理

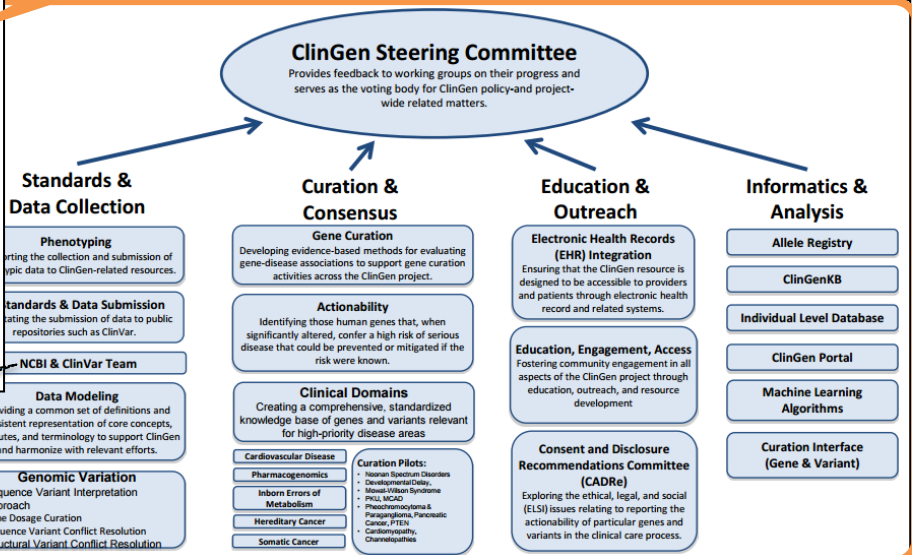
① ClinVar/ClinGenの概要

- ゲノムの変異をより深く理解するためにデータシェアリングと協働的キュレーションが必要
- ClinVarは、遺伝子変異がほぼ症状を決定する遺伝子に関する変異を集積するレポジトリ（登録者の自主的は判断によるキュレーションを許容した上で、変異の病原性に関するデータを受け付けている）
- ClinGenにより、科学雑誌におけるピア・レビューのごとき評価が加えられる（科学雑誌にエディターがいて、論文の内容をキュレーションするように、ClinGenでは、研究者がゲノム研究の成果をキュレーションする。そのキュレーション方法を研究・体系化している）
- ClinGenにある15のワーキンググループで、各テーマにそって、そのような活動をしている。活動内容は、研究的なものであり、システム構築が主目的ではない



ClinGen
 NHGRIとNICHDが3つの研究グループに助成、それがコンソーシアムを形成
 2013年からの4年間
 \$25 million (3グループ分の合計額)
 15のワーキンググループ、運営委員会、外部パネルで構成 (総勢約200名)

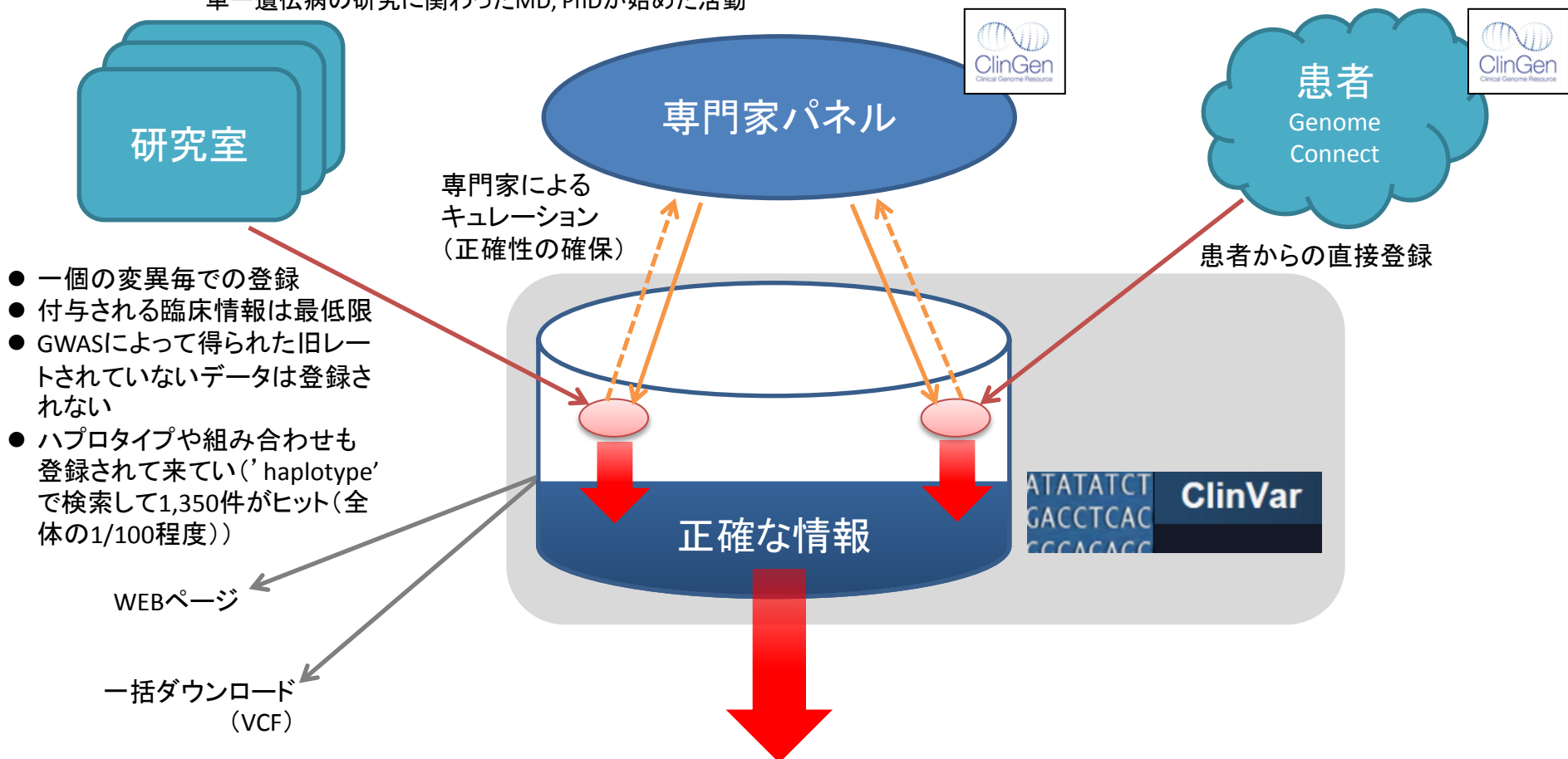
ClinVar
 NCBIにより構築される
 (ClinGenの委員がClinVarの開発に加わったり、NCBIのスタッフがClinGenに参加したり、密なやり取りの上で開発が進められている)



出典) https://www.clinicalgenome.org/site/assets/files/1457/plon_clingendb.pdf

② ClinVarとClinGenの関係性

単一遺伝病の研究に関わったMD, PhDが始めた活動



まずは、単一遺伝病の検査
予測の正確性が増すことにより、将来的には、

- 検査会社の参入可能性拡大
- 偶発的所見への対応可能性の向上
- 健常人の、将来的な健康管理のための活用

まとめ

1. ClinVarは、遺伝子変異がほぼ症状を決定する遺伝子に関する変異を集積するためのレポジトリ
 - NCBIにより構築・運営されている
 - 日常の遺伝子診療にあまねく利用されている
2. ClinGenは、病気に関わる変異のキュレーションを向上させるための様々な活動
 - NHGRIやNICHDのファンドを受けている4年間のプロジェクト
 - キュレーションの体系化や実施が活動の中心であるが、それに関わる様々な活動も含まれる

論点

1. ClinVarを構築するためには、そのコンテンツが必要
コンテンツ作成のためにはその体制(特に、キュレーション体制)が必要
(人材確保の必要性)
2. 今後の研究によって得られるデータはもちろんのこと、臨床検査のラボ等に蓄積されている情報を、いかに迅速に共有し医療に活用していくか
3. 人的ネットワークをどの様に深めていくか(データの登録・評価・活用等)

参考資料

【コンテンツ】

- ① 欧州・米国における主要データシェアリング基盤
 - 主要データシェアリング基盤と流通しているデータ
(データの内容・フォーマット)
- ② ClinGenの注目活動
 - 1) データシェアリング基盤の整備(ClinVar)
 - 2) キュレーション方法の検討・実施
 - 3) 機械学習手法の開発
 - 4) 一般患者の直接的な研究参加(GenomeConnect)

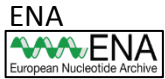




① 欧州・米国における主要データシェアリング基盤

欧州 (EBIが整備しているものを整理)

EBI: イギリス政府とEMBLとの協定*に基いて設立される


米国 (NCBIが整備しているものを整理)


NCBI: Public Law 100-607で、NIHの付属機関として設置される


	格納データ	主要登録フォーマット	主要提供フォーマット
 ENA European Nucleotide Archive	2008年～(母体のEMBL-Bankは1980年代にサービス開始) 各種DNA配列データ(キャピラリーシーケンサー・NGSによって生成されたDNA配列データ、アセンブル結果)	XML, EMBL-Bank形式	CRAM, XML, FASTA, FASTQ
 EVA	小さな多型・変異のアレル・ゲノム配列上位置	VCF	VCF, CSV
 EGA European genome-phenome archive	サンプル毎の遺伝子型とそのサンプルの表現型、統計解析結果を制限公開 (その他に、DNA配列データや発現データ等も取り扱う)	VCF, BAM, NGSデータフォーマット	VCF, BAM, NGSデータフォーマット
 EGVa DGVA archive	構造多型・変異のアレル・ゲノム配列上位置	EXCEL, TSV	GVF
 Clinical Studies	ClinVarに対応するものとして画されている	?	?

* Agreement between the Government of the United Kingdom of Great Britain and Northern Ireland and the European Molecular Biology Laboratory Concerning the European Bioinformatics Institute)

	格納データ	主要登録フォーマット	主要提供フォーマット
 dbSNP Short Genetic Variations	1999年～ 小さな多型・変異(50塩基未満)のアレル・ゲノム配列上位置	EXCEL, VCF	FASTA, ASN1, XML, BED
 dbGaP Genotypes and Phenotypes	2006年～ サンプル毎の遺伝子型とそのサンプルの表現型、統計解析結果(GWAS等の入力・解析結果を含む)	EXCEL, FASTQ, VCF, CEL	VCF, 行列, (indfmt, matrixfmt) PLINK
 SRA Sequence Read Archive	2007年～ NGSによって生成されたDNA配列データ(リード、マッピング結果) (制限公開データへのアクセスには、dbGaPへのアクセス申請が必要)	FASTQ, BAM	SRA, cSRA
 dbVar dbVar is NCBI	2010年～ 大きな多型・変異(50塩基以上)のアレル・ゲノム配列上位置	EXCEL, VCF	VCF, GVF, XML, CSV/TSV
 ClinVar	2013年～ 変異と疾患(主に単一遺伝子病)、その表現型・エビデンスの集約結果 (現状、GWASによって得られたレビューを受けていない変異は含まれない)	EXCEL, XML, CSV/TSV	VCF, XML, CSV/TSV

 集団中に認められた多型・変異 (個人別ではない)

 個人の持つ多型・変異(複数)と表現型 (個人別)

 医学的に意義のある多型・変異・表現型

主な流通フォーマット

- FASTA, FASTQ : 配列データを格納するためのフォーマット(FASTQはクオリティデータを含むことができ、NGSのリードデータを格納する事に用いられる)
- BAM : リードのアライメント結果をバイナリで格納するためのフォーマット
- VCF : ゲノム配列における多型・変異の位置情報・配列情報、サンプルの遺伝子型を格納するためのフォーマット(病的意義・頻度情報等を付加可能)
- GVF : ゲノム配列における多型・変異の位置情報・配列情報を格納するためのフォーマット
- SRA, cSRA, CRAM : リードデータ(FASTA/FASTQ、シーケンサーの出力フォーマット)、リードのアライメント結果をまとめてバイナリで格納するためのフォーマット
- BED : ゲノム配列に対するアノテーション情報を格納するためのフォーマット
- CEL : Affymetrix社の遺伝子型決定用チップを用いて得られる生データ(装置に付属するソフトウェアが出力するデータ)

②ClinGenの注目活動

ClinGenの注目活動(次項以降に紹介)

1)



Data Sharing

ClinGenは、匿名化されたゲノムデータは自由に活用・アクセスできるべきであるとしており、データシェアリング基盤へのデータ登録を促進する活動を行っている

2)



Knowledge Curation

ClinGenは、遺伝子と疾患との関係性の深さや、変異が表現型に係る程度をアノテーションし解釈するためのキュレーション基盤を整備している

3)



Machine Learning

ClinGenは、臨床的・実験的証拠が制限されている場合に対して、多型・変異の臨床的意味の予測を支援するための機械学習手法を開発している

4)



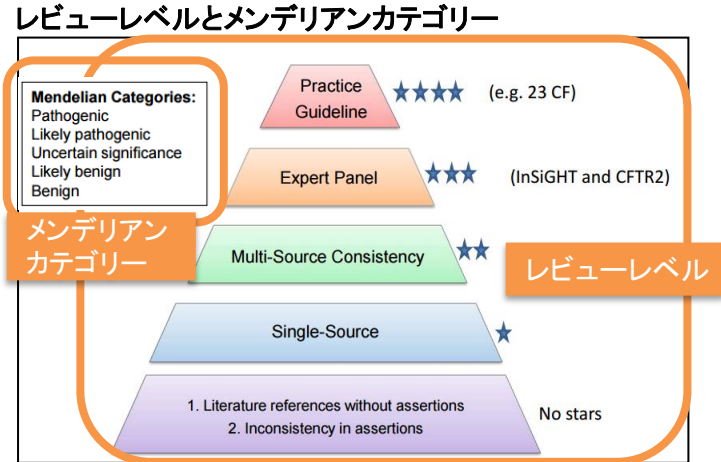
ClinGenは、患者が多型・変異と表現型とを、医療供給者や研究者と共有するための基盤を整備している。

② ClinGenの注目活動 1) データシェアリング基盤の整備 (ClinVar)

多型のレコード

レビューレベル

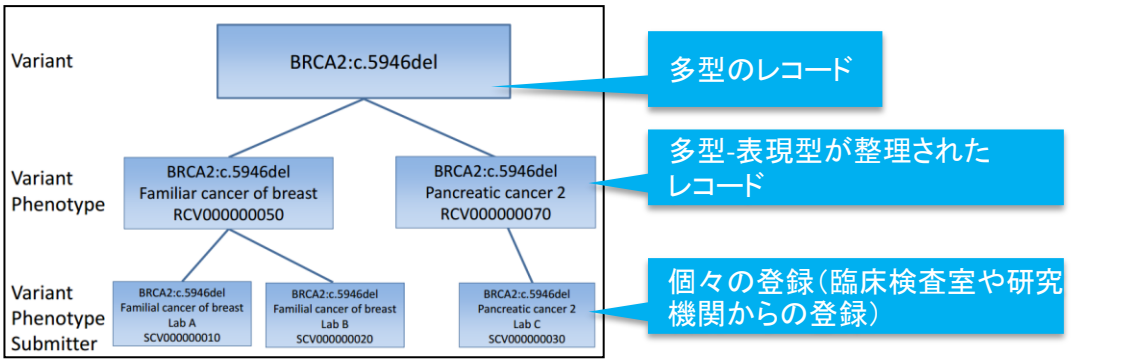
メンデルアン カテゴリー



- ClinVarはNCBIによって整備されているデータシェアリング基盤であるがClinGenもそれに関与
- 医学的に意義がある多型・変異とそれに関連付けられる表現型の情報を集約
- 更に、個々のレコードには、以下の情報が付与される
 - レビューレベル(どの程度のレビューを受けているか; 医学的信頼性を判断するための指標)
 - メンデルアンカテゴリー(多型・変異の医学的な意味(疾患に対する影響度))

多型-表現型が整理されたレコード(RCV)

データ構成



② ClinGenの注目活動 2) キュレーション方法の検討・実施

- NGSの登場により、遺伝子検査の対象となる遺伝子数が増大し、その検査方法も多様化(単一遺伝子、遺伝子パネル、エクソーム、ゲノム、トランスクリプトーム、エピゲノム)
- ACMGは、その状況の変化を受けて、変異を解釈する方法の標準化・ガイドライン作成を行う(この活動に、ClinGenのメンバーも参加)
- 得られている証拠を基にして多型を分類するためのスコアリングルールを作成
- 多型の分類(Benign/Pathogenic等)に対して、得られている証拠の種類と、その分類の確からしさの整理

多型のスコアリングルール

Pathogenic	(i) 1 Very strong (PVS1) AND (a) ≥1 Strong (PS1–PS4) OR (b) ≥2 Moderate (PM1–PM6) OR (c) 1 Moderate (PM1–PM6) and 1 supporting (PP1–PP5) OR (d) ≥2 Supporting (PP1–PP5) (ii) ≥2 Strong (PS1–PS4) OR (iii) 1 Strong (PS1–PS4) AND (a) ≥3 Moderate (PM1–PM6) OR (b) 2 Moderate (PM1–PM6) AND ≥2 Supporting (PP1–PP5) OR (c) 1 Moderate (PM1–PM6) AND ≥4 supporting (PP1–PP5)
Likely pathogenic	(i) 1 Very strong (PVS1) AND 1 moderate (PM1–PM6) OR (ii) 1 Strong (PS1–PS4) AND 1–2 moderate (PM1–PM6) OR (iii) 1 Strong (PS1–PS4) AND ≥2 supporting (PP1–PP5) OR (iv) ≥3 Moderate (PM1–PM6) OR (v) 2 Moderate (PM1–PM6) AND ≥2 supporting (PP1–PP5) OR (vi) 1 Moderate (PM1–PM6) AND ≥4 supporting (PP1–PP5)
Benign	(i) 1 Stand-alone (BA1) OR (ii) ≥2 Strong (BS1–BS4)
Likely benign	(i) 1 Strong (BS1–BS4) and 1 supporting (BP1–BP7) OR (ii) ≥2 Supporting (BP1–BP7)
Uncertain significance	(i) Other criteria shown above are not met OR (ii) the criteria for benign and pathogenic are contradictory

分類の確からしさを評価するフレーム

	Benign			Pathogenic		
	Strong	Supporting	Supporting	Moderate	Strong	Very strong
Population data	MAF is too high for disorder BA1/BS1 OR observation in controls inconsistent with disease penetrance BS2			Absent in population databases PM2	Prevalence in affecteds statistically increased over controls PS4	
Computational and predictive data		Multiple lines of computational evidence suggest no impact on gene /gene product BP4 Missense in gene where only truncating cause disease BP1 Silent variant with non predicted splice impact BP7 In-frame indels in repeat w/out known function BP3	Multiple lines of computational evidence support a deleterious effect on the gene /gene product PP3	Novel missense change at an amino acid residue where a different pathogenic missense change has been seen before PM5 Protein length changing variant PM4	Same amino acid change as an established pathogenic variant PS1	Predicted null variant in a gene where LOF is a known mechanism of disease PVS1
Functional data	Well-established functional studies show no deleterious effect BS3		Missense in gene with low rate of benign missense variants and path. missenses common PP2	Mutational hot spot or well-studied functional domain without benign variation PM1	Well-established functional studies show a deleterious effect PS3	
Segregation data	Nonsegregation with disease BS4		Cosegregation with disease in multiple affected family members PP1	Increased segregation data →		
De novo data				De novo (without paternity & maternity confirmed) PM6	De novo (paternity and maternity confirmed) PS2	
Allelic data		Observed in trans with a dominant variant BP2 Observed in cis with a pathogenic variant BP2		For recessive disorders, detected in trans with a pathogenic variant PM3		
Other database		Reputable source w/out shared data = benign BP6	Reputable source = pathogenic PP5			
Other data		Found in case with an alternate cause BP5	Patient's phenotype or FH highly specific for gene PP4			

出典) Richards et al. 2015

② ClinGenの注目活動

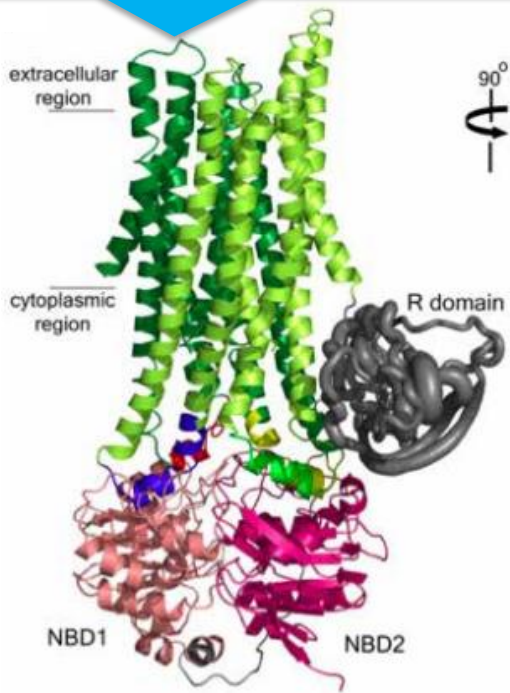
3) 機械学習手法の開発

- 医学的・実験的証拠が常に十分であるとは限らないため、マニュアルキュレーションを支援するための機械学習手法の開発を行う
 - 以下に示すような開発が存在
 - Gene-Level: 特定の遺伝子を対象と定め、それらにおける非同義置換の効果を予測する
 - Genome-Wide Modeling: 全ゲノムに亘る変異の機能・医学的重要度を予測する(ノンコーディング領域、イントロン、同義置換を含む)
- ※機械学習手法の他に、実験的手法の開発も行っている(例: CRISPER/Cas9の活用)

変異がもたらす影響を予測する手法の開発

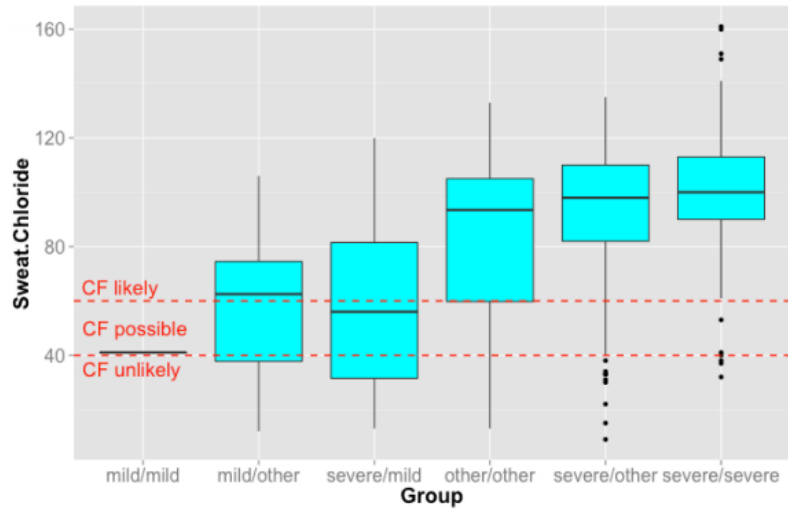
【例1】

- 変異が構造の安定性にもたらす影響の予測
- 配列、構造、進化のデータ等を組み合わせて予測精度を高める(SIFT, PolyPhen-2, OANTHER, POSE, 等)
- 学習データは、CFTR2、1000G、ClinVar等から取得



【例2】

- 変異が表現型に及ぼす影響を予測
- 新生児からフォローアップしている600人近くを対象として表現型のデータベースを構築、それを手法開発に活用
- タンパク質立体構造中のどこに変異があるか(内部・表面)及び同義・非同義置換でその影響度を分類(Severe/Mild/Other)



Mutation classes:
Severe: located in the inner parts of the protein;
Mild: located in the outer part of the protein;
Other: all non-missense mutations.

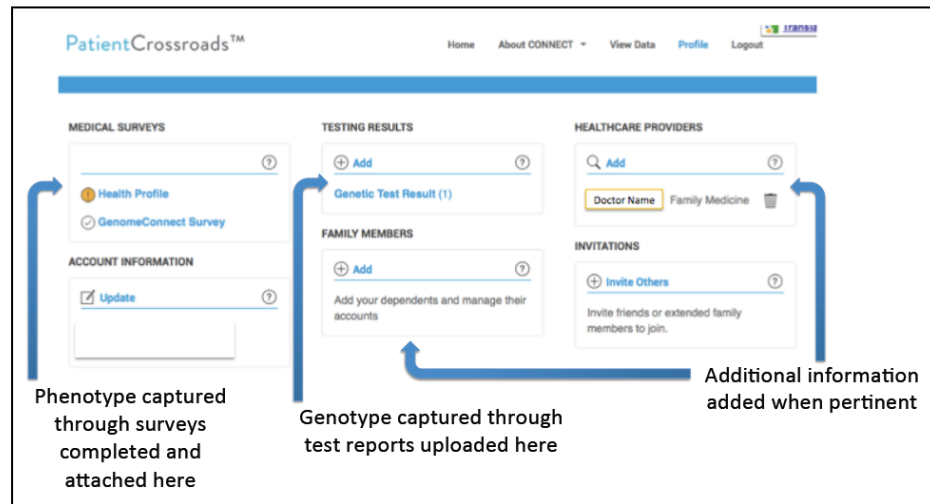
出典) https://www.clinicalgenome.org/site/assets/files/1991/nsgc_207_plenary_faucett_ormond_clingen_thurs_plenary_for_posting.pdf

② ClinGenの注目活動 4) 一般患者の直接的な研究参加

- GenomeConnectは、研究者、臨床検査室、医療供給者に次ぐ第4のリソースとして、一般患者からの直接的なデータ登録を受け付ける取り組み（データは、ClinVar/ClinGenに取り込まれる）
- Patient Crossroads社が提供するプラットフォームを用いて構築されている



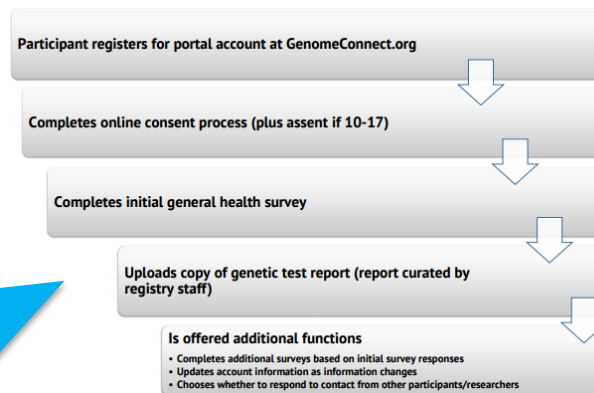
データ取得のためのWEBページ



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

- 1) ポータルサイト (GenomeConnect.org) でのアカウント作成
- 2) オンラインでの同意
- 3) オンラインでの一般的な健康情報の入力
- 4) 遺伝子検査結果のアップロード（アップロードされた物は、GenomeConnectのスタッフによる確認を受ける）
- 5) 適宜的な実施内容
 - 3)に加えての追加入力
 - アカウント情報の更新
 - 他の参加者や研究者からのコンタクトの受付

参加者の実施内容



出典) https://www.clinicalgenome.org/site/assets/files/2296/genomeconnect_ashg_poster_2014.pdf