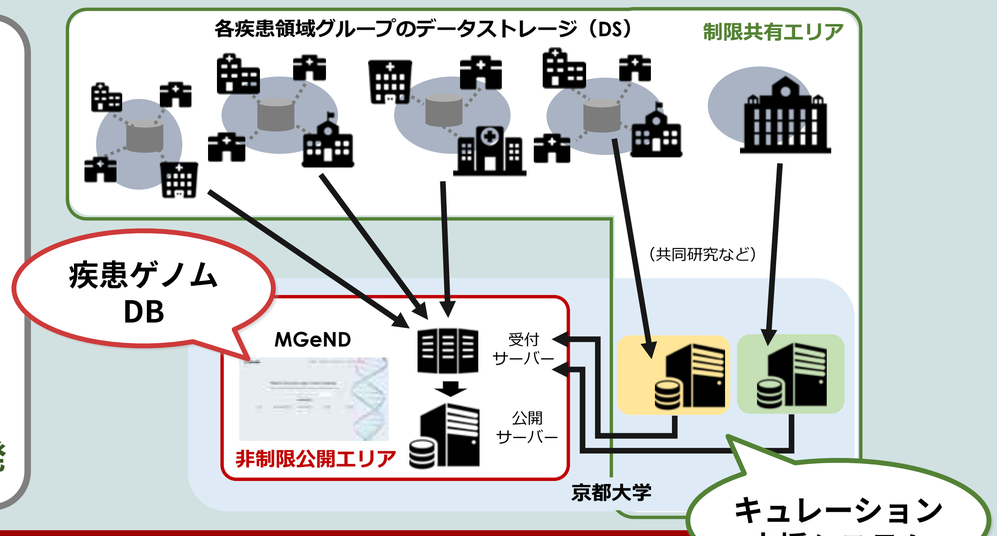
日本人疾患ゲノム情報統合データベースMGeNDとキュレーション支援システムの開発

鎌田 真由美, 中津井 雅彦, 小島 諒介, 奥野 恭史（京都大学大学院 医学研究科）

研究の概要

- ゲノム医療では、解析で得られるバリアントに対する臨床的な解釈付け（キュレーション）に基づき治療方針決定を行う
- 適切なキュレーションには、民族集団における遺伝的背景の違いを考慮する必要がある
  - ➔ 日本人集団から得られるゲノム情報と疾患特異性・臨床特性情報を集約した統合データベース **MGeND （Medical Genomics Japan Variant Database）** の開発
- これまでに、「がん」「希少・難治性疾患」「認知症」「感染症」「難聴」を主な対象疾患とし、各疾患症例より得られた一塩基・構造バリアントとHLA領域多型の頻度情報を公開
- キュレーションには膨大な情報の包括的解釈が必要 ➔ **AI・NLP技術によるサポートシステムの開発**

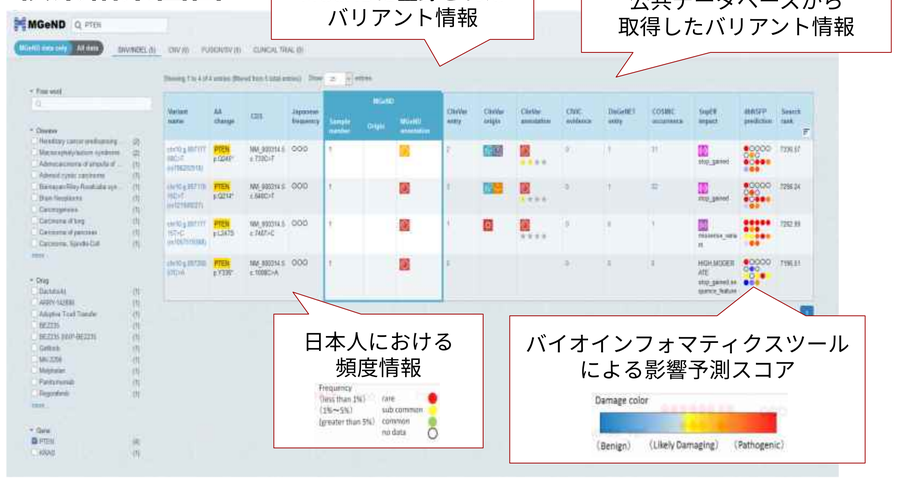


疾患ゲノム情報統合データベースMGeNDの開発

MGeND https://mgend.med.kyoto-u.ac.jp

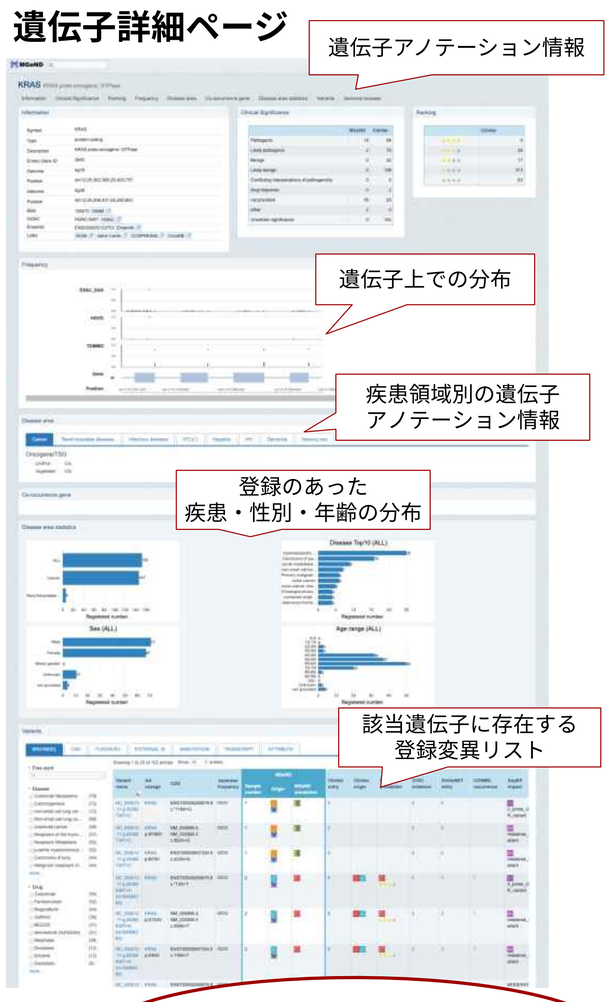
トップ画面

検索結果画面



変異詳細ページ

遺伝子詳細ページ

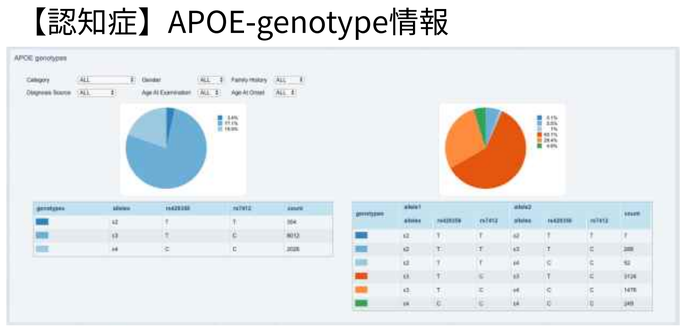


現在、HIV薬剤耐性変異画面やGWAS画面などを開発中！

疾患固有画面

【感染症】HLAアリル頻度情報

【認知症】APOE-genotype情報



データ登録フォーマット

SNV/INDEL	• ClinVar登録フォーマットに準拠した項目の設定 • 個人情報に考慮した統計値（頻度）での登録にも対応 • TSV/XML/VCF形式に対応
GWAS	• GWAS解析の共有に必要な項目を策定 • TSV形式に対応
HLA	• HLAタイピング解析で標準的に用いられるテーブル形式を踏襲 • Xlsx(TSV)形式に対応

登録可能なデータ

- ◆ 疾患名（標準的な記載方法のもの）
- ◆ 遺伝子名
- ◆ 遺伝型情報（Genotype)
- □ １～数箇所程度のSNV・SNPまたはp値<10^{-4} のSNP すべて
- □ 年齢（層）
- □ 性別（「不明」「混合」等を含む）
- □ 解析手法など

登録データ数（2019年3月現在）

Disease area	Variants	GWAS	HLA allele data
Cancer	25,075(9041)	---	---
Rare/Intractable disease	13,513 (2459)	---	---
Infectious disease	2 (2)	155,098 (22,358)	1,306 (999)
Hearing loss	122 (122)	---	---
Dementia	7631(7760) APOE: 12,298 (5196)	410 (414)	---
Others	2721 (1998)	14,324,204 (0)	3261 (0)

疾患関連遺伝子キュレーション支援システムの開発

バリアントに対するキュレーションは、**ゲノム医療のボトルネック**

膨大な量の文献からエビデンスを調査する必要

幅広いデータベース情報を包括的に参照し、理解する必要

専門家でも1変異のキュレーションに約１時間，エキソームデータの臨床的評価には 20～40時間が必要であるとも言われている…

論文からの知識抽出

変異と疾患の解釈に重要そうな論文をNLP技術を用いて自動的にピックアップ！

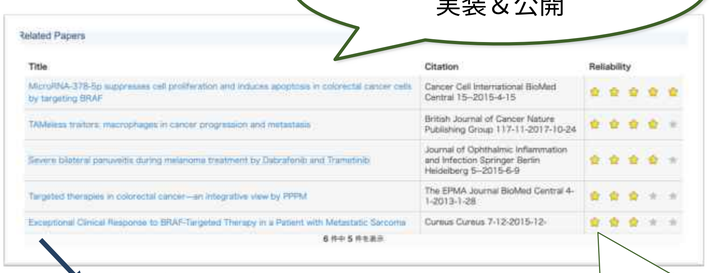
PubMed 論文

固有表現抽出
テキストから「変異」「遺伝子」「疾患」を抽出

関係抽出
抜き出したエンティティ間の関係を推定

関連する論文のリスト

制限共有システムとして実装＆公開

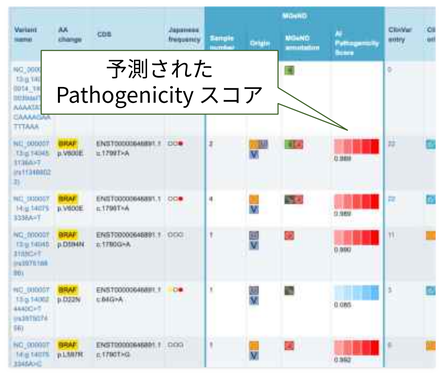


各論文のアブストラクト

関連度・重要度をスコア化し、スコアノア高いものから順にリストアップ

臨床的意義の推定

バリアントの疾患関連性をAIにより予測



謝辞：本研究は、AMED「臨床ゲノム情報統合データベース整備事業」（課題番号19kk0205013h0004）の支援を受け、実施しています。