

生成AI時代における ゲノム解析基盤開発

国立がん研究センター

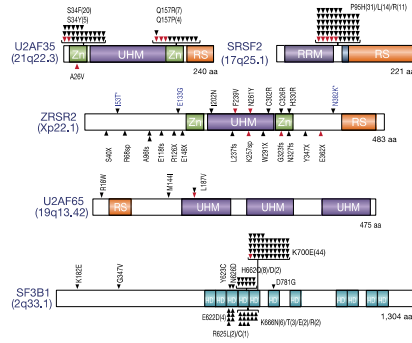
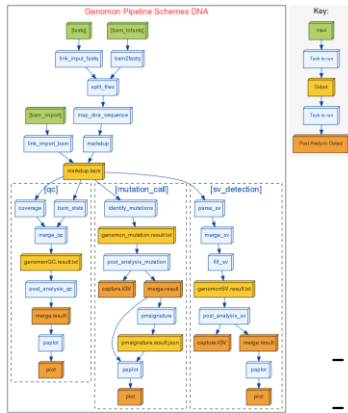
白石 友一

自己紹介

がんゲノム解析パイプラインの開発

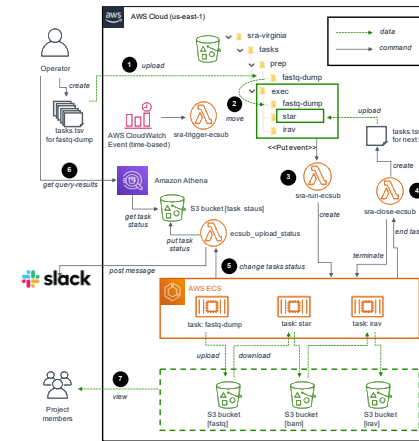


Genomon



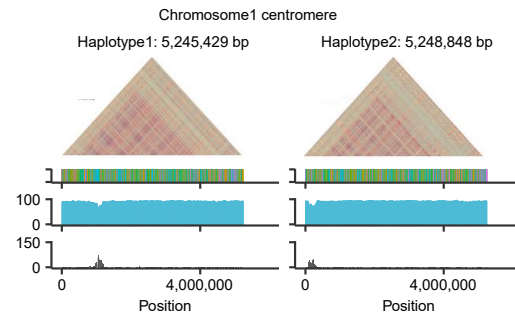
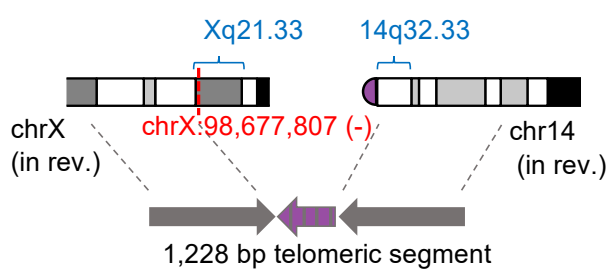
- Yoshida et al., Nature, 2011
- Sato et al., Nature Genet., 2013
- Kataoka et al. Nature, 2016

公共オミクスデータからの知識発見基盤



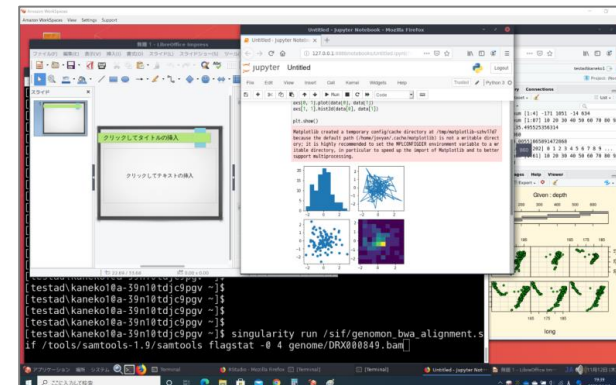
- Shiraishi et al. Genome Res., 2018
- PCAWG et al., Nature, 2020
- Shiraishi et al., Nature Comm., 2022
- Mateos et al., npj SBA, 2024.
- Iida et al., Nature Comm., 2025

難読領域（セントロメア）におけるゲノム構造異常解析



- Shiraishi et al., NAR, 2023
- Nakamura et al., npj GM, 2024
- Sugawa et al., in preparation

保険診療のゲノムデータの利活用基盤 (がんゲノム情報管理センター C-CAT)



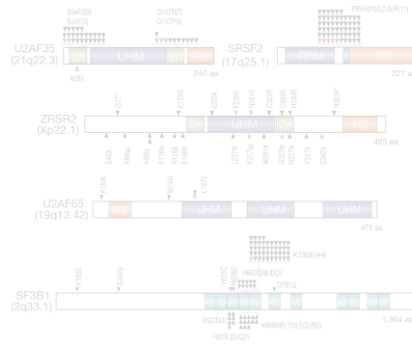
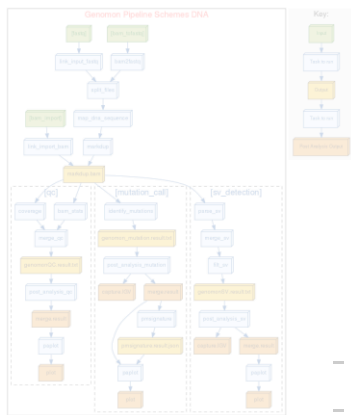
利活用解析基盤
C-CAT CALICO

自己紹介

ゲノム解析パイプラインの開発

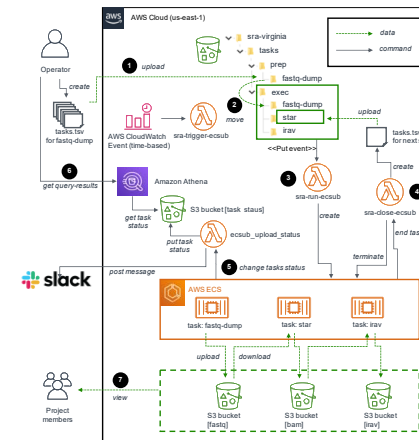


Genomon



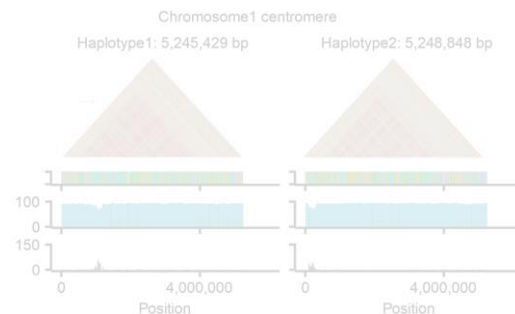
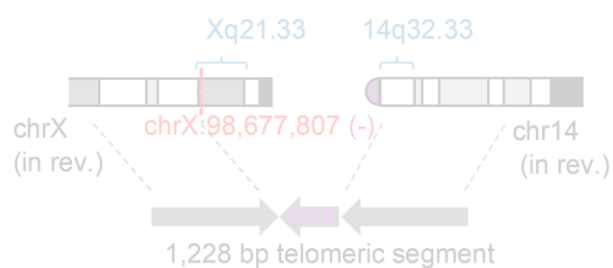
- Yoshida et al., Nature, 2011
- Sato et al., Nature Genet., 2013
- Kataoka et al. Nature, 2016

公共オミクスデータからの知識発見基盤



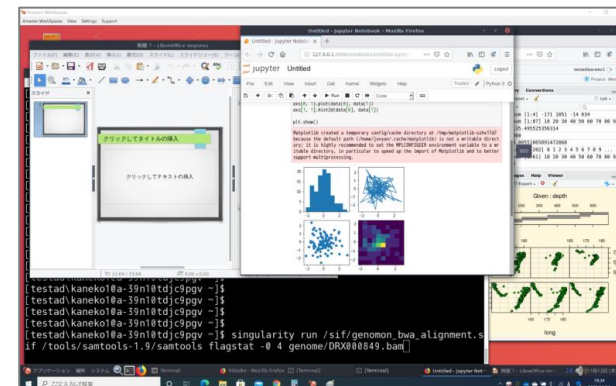
- Shiraishi et al. Genome Res., 2018
- PCAWG et al., Nature, 2020
- Shiraishi et al., Nature Comm., 2022
- Mateos et al., npj SBA, 2024.
- Iida et al., Nature Comm., 2025

難読領域（セントロメア）におけるゲノム構造異常解析



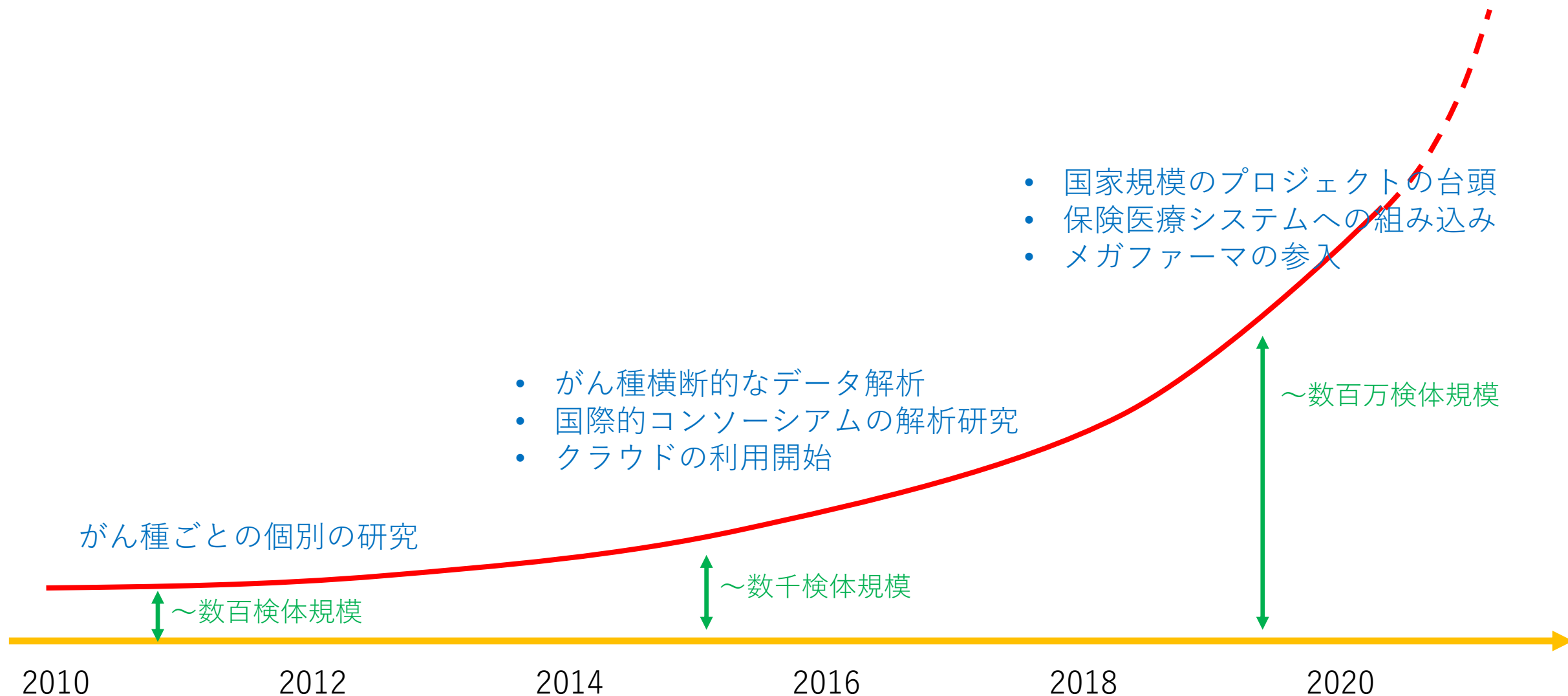
- Shiraishi et al., NAR, 2023
- Nakamura et al., npj GM, 2024
- Sugawa et al., in preparation

保険診療のゲノムデータの利活用基盤 (がんゲノム情報管理センター C-CAT)



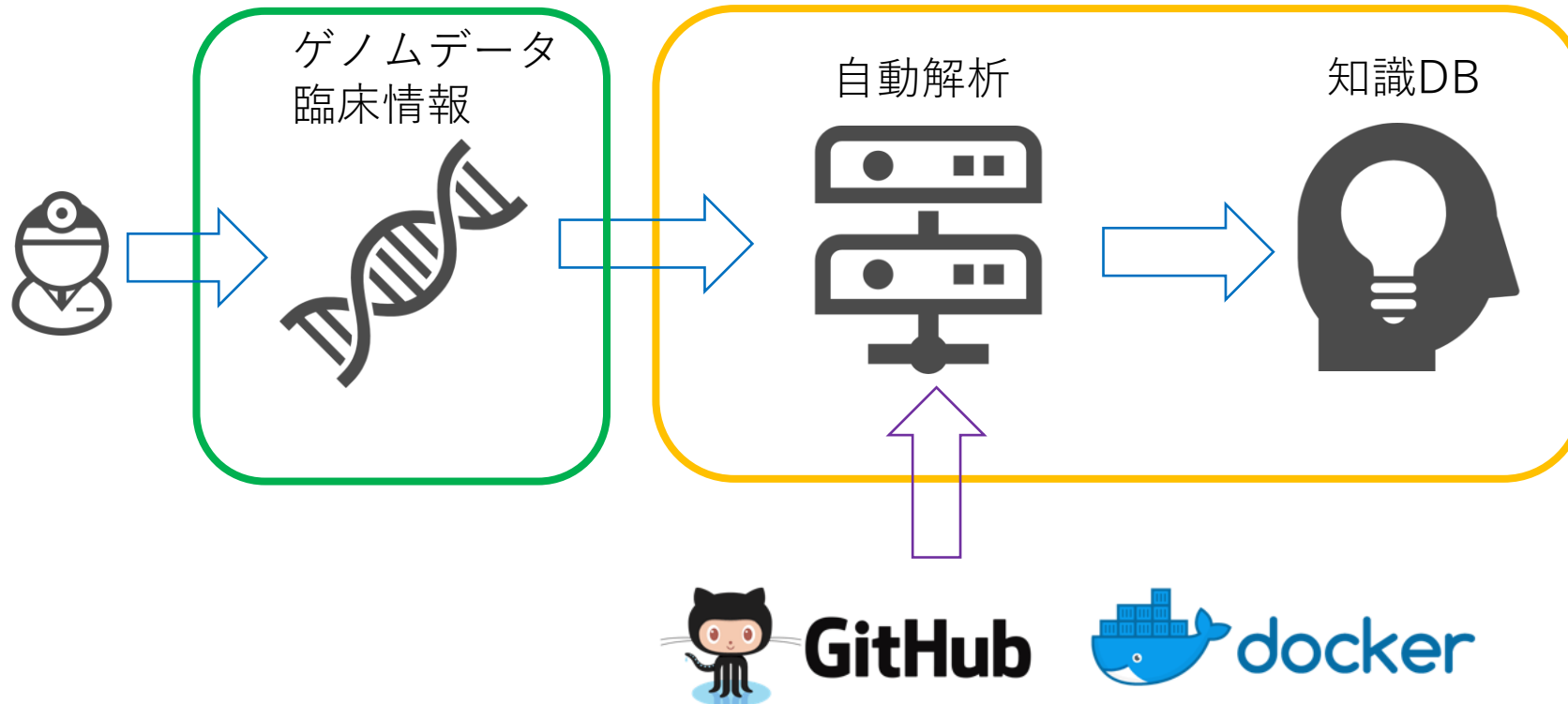
利活用解析基盤
C-CAT CALICO

ゲノムデータの数 は 年々 大きくなっている



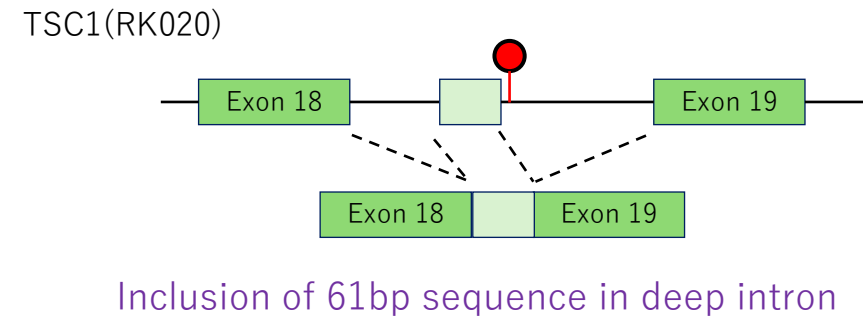
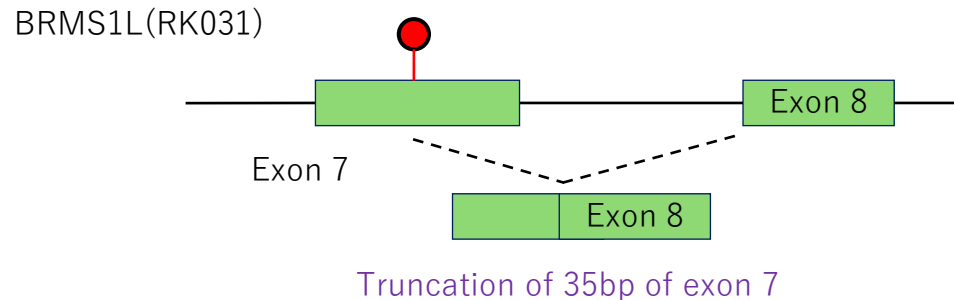
ゲノムデータ利活用の方法

- 「知識」が自律的に蓄積していくシステム。

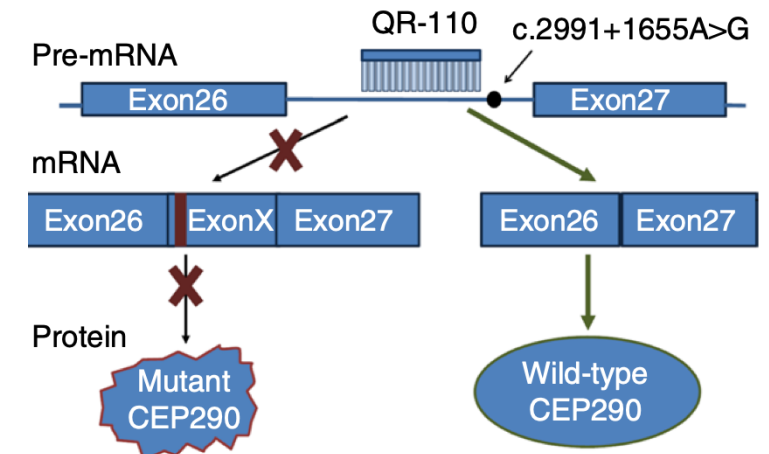


Splice-Site Creating Variant, SSCV

Splice-site creating variant (変異によって新しいsplice siteが形成)

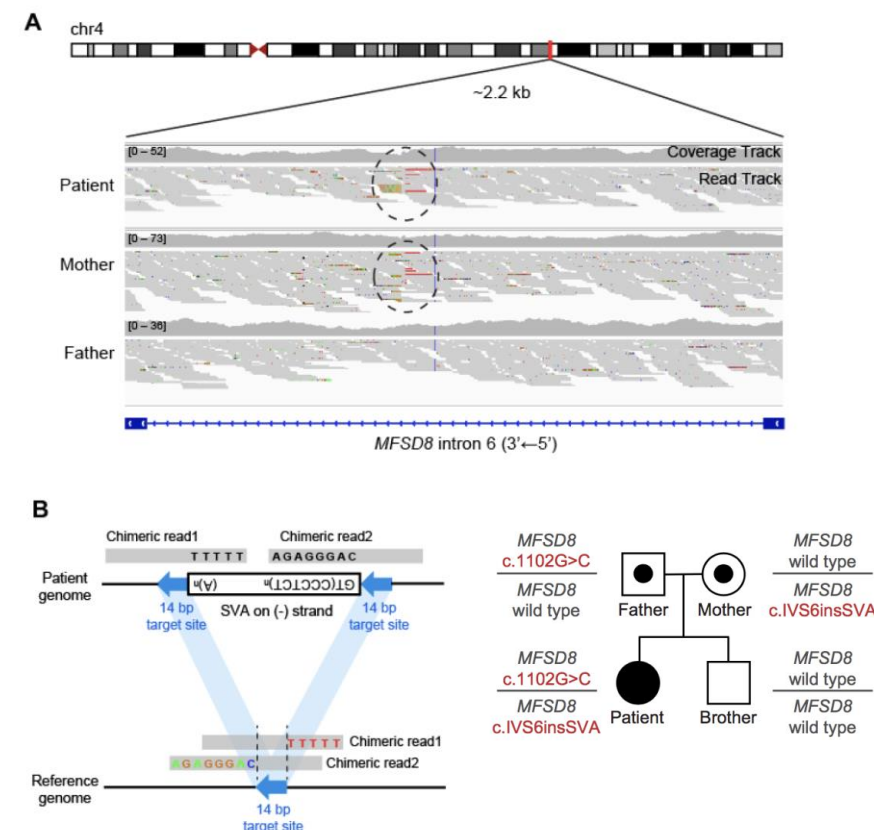


- 全ゲノム解析で検出される変異の解釈
 - 深部イントロンにおいて偽エキソンを生成する変異の同定
- 核酸医薬（アンチセンスオリゴヌクレオチド）による治療の可能性
 - レーバー先天性黒内障の *CEP290* の c.2991+1655A>G を対象とする核酸医薬 (Cideciyan et al., Nature Medicine, 2019)。
 - N-of-1 medicine! (Kim et al., Nature, 2023).



Cideciyan et al., Nature Medicine, 2019

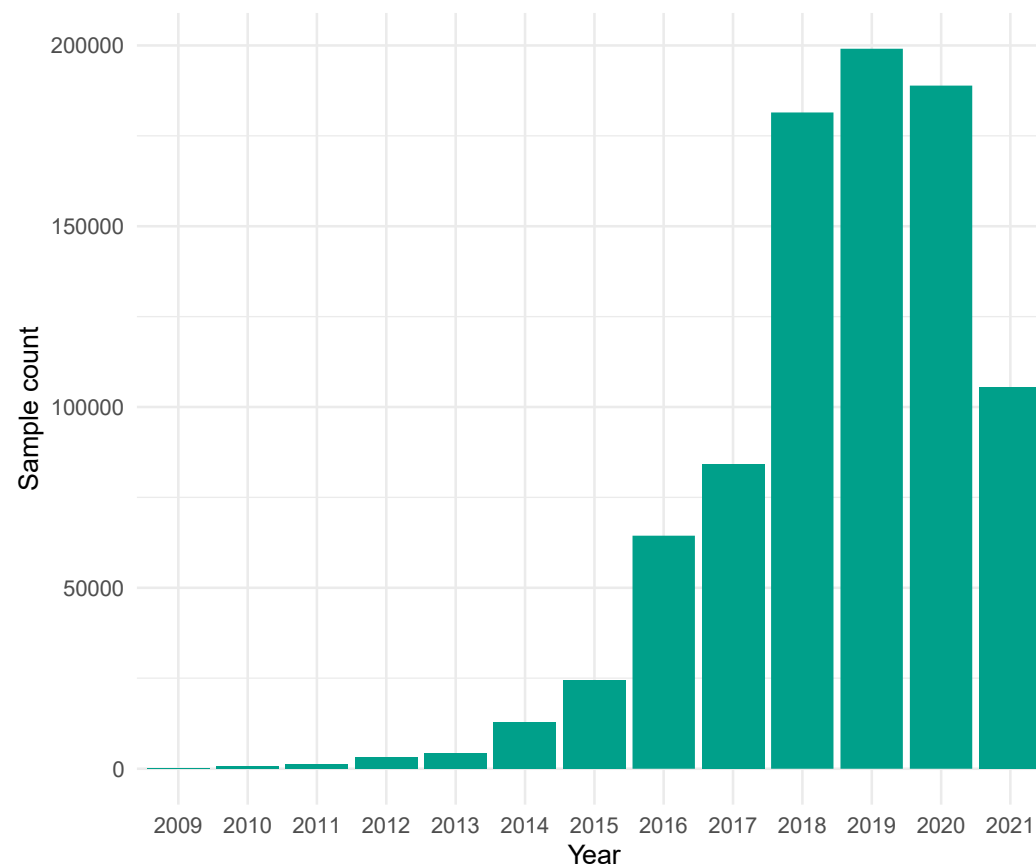
N-of-1 medicine (超希少疾患の個別化医療)



- MFSD8遺伝子における2つのゲノム変異
 - 一つは深部イントロンでスプライシング異常を引き起こす変異
- アメリカのボストン小児病院で「Milasen」という核酸医薬が開発された (Kim et al., NEJM, 2019)

トランスクリプトームのみを用いたSSCVの検出？

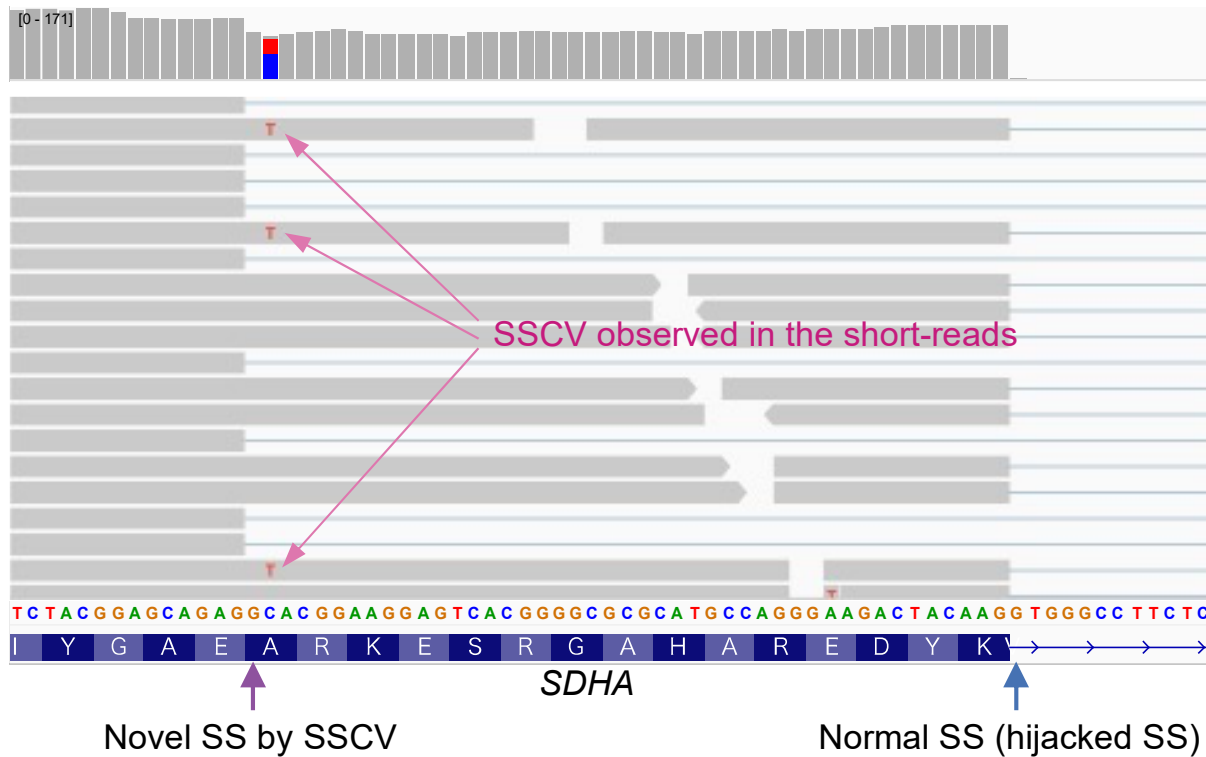
- ゲノム・トランスクリプトームのペアがあれば比較的容易
 - ゲノム・トランスクリプトームがペアで与えられているデータセットは希少である。
 - TCGA: 約10,000ペア。
- トランスクリプトームデータだけでスプライシング変異を検出することはできる？
 - INSDC (SRA, EGA, DRA)の100万件規模のトランスクリプトームデータが使える！



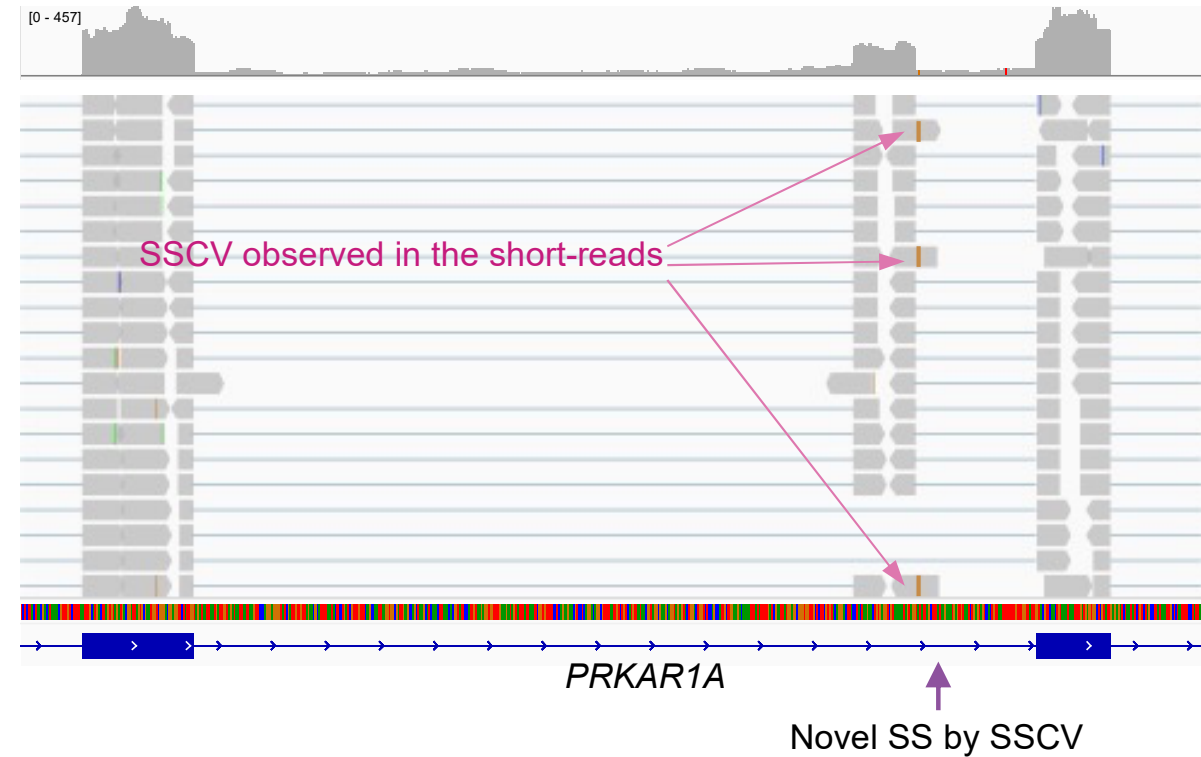
100万件以上のRNA-seqが
「申請の必要がなく」利用可能！

RNA-seqのリードからSSCVが観測できる

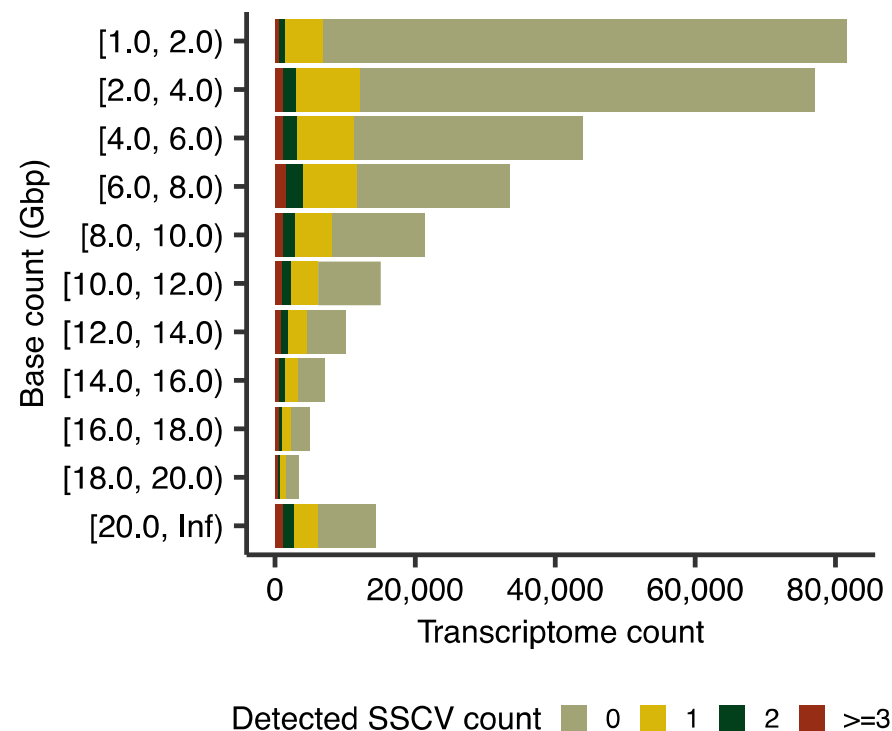
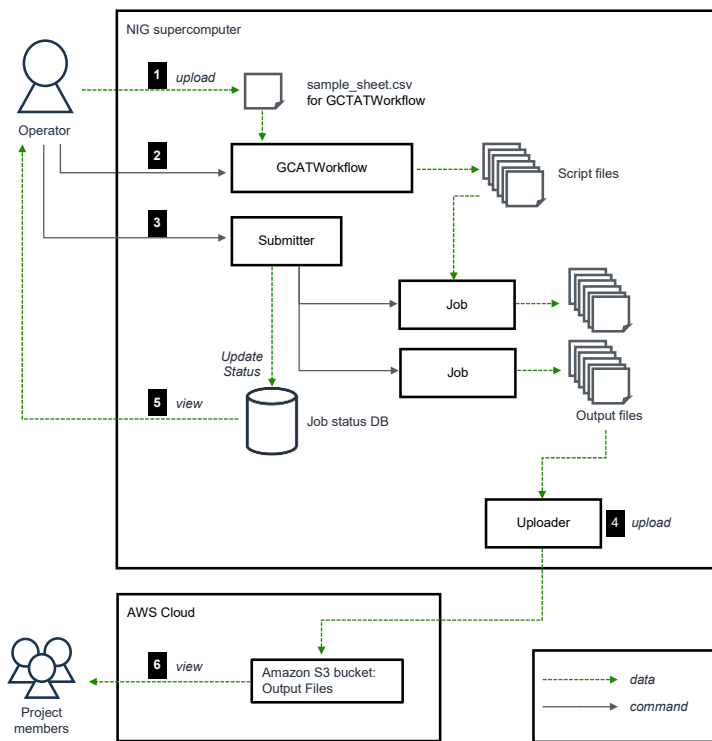
TCGA-DX-A1L4 (partial exon truncation by SSCV)



TCGA-OR-A5JY (cryptic exon inclusion by SSCV)



INSDCの30万件のRNA-seqのスクリーニング



- DDBJ supercomputerを用いて312,315のRNA-seqからの大規模SSCVスクリーニングを実行。
- Coding遺伝子において30,130のSSCVを検出。

SSCV DB

SSCV DB About

SSCV DB



Towards autonomous acquisition of splicing variants

14,960

SRA project number

310,699

SRA runID number

28,945

SSCV (distinct) number

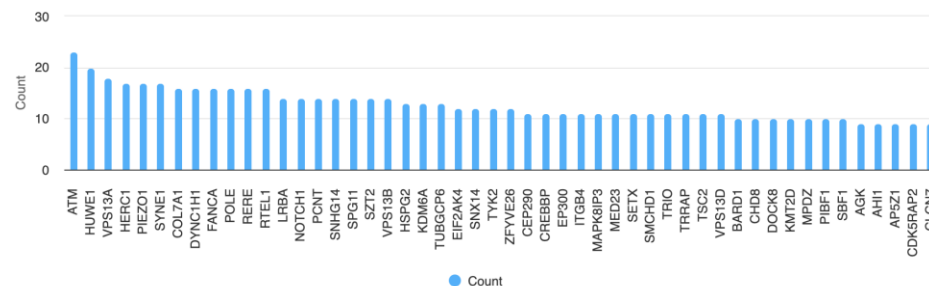
Search for gene

Search

Examples: TP53, SMARCA4, NRF2

Disease associated gene ▾

Number of distinct SSCVs (disease associated genes)



CREBBP

Cancer Gene Census

ClinGen Dosage Sensitivity

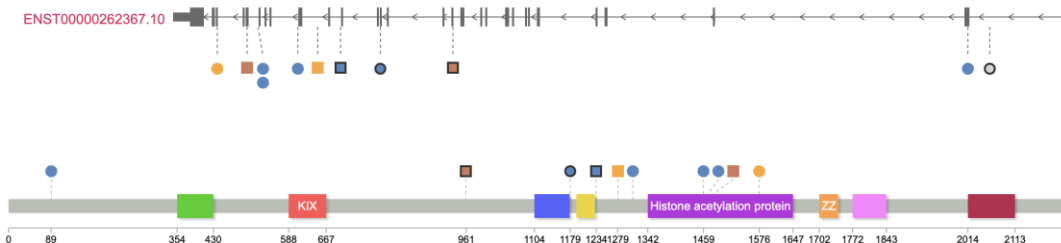
Clinical Genomics DataBase

OMIM

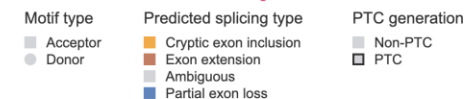
Gene Alias RSTS1, MKHK1, CBP, KAT3A, RSTS

Region (GRCh38) chr16: 3725053-3880713

CREBBP, chr16



* The Gencode ID marked in red signifies a MANE Select or MANE Plus Clinical transcript.

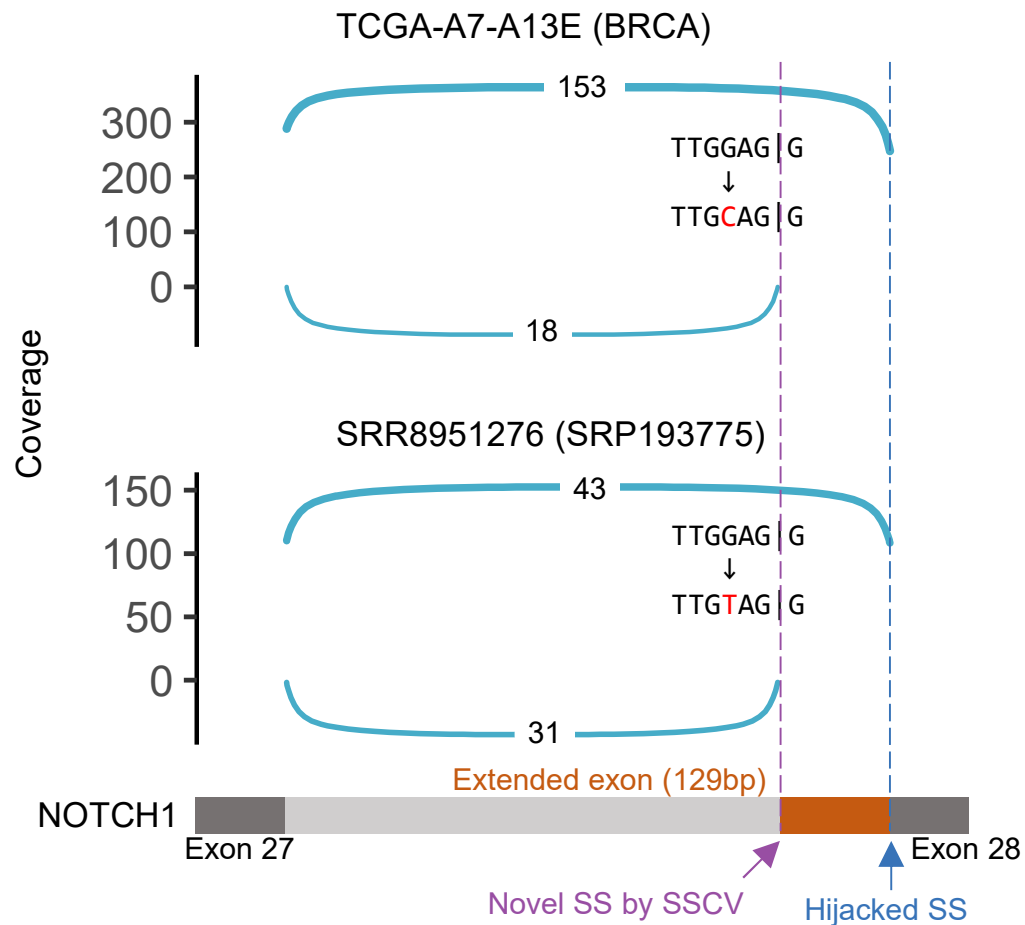


Or google "SSCVDB"

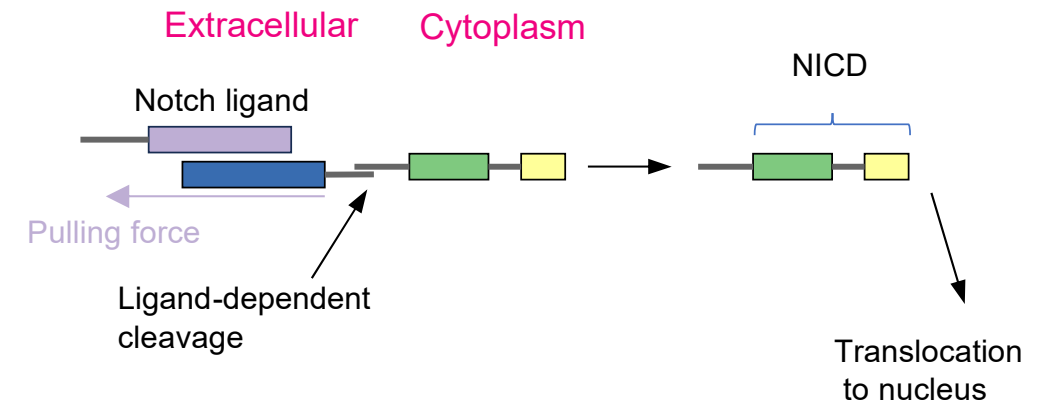
Iida et al., Nature Communications, 2025

NOTCH1における新規gain-of-function 変異

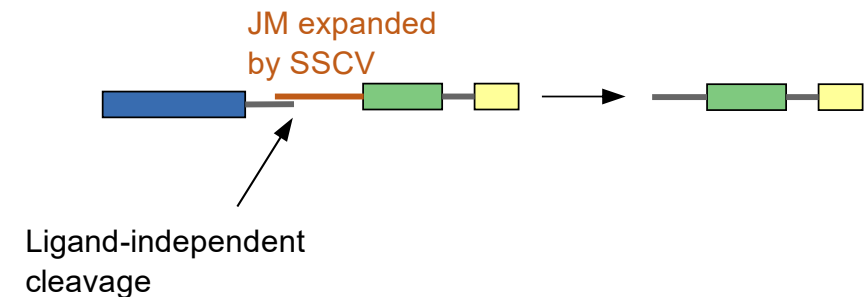
SSCVによるIn-frame insertionの生成



NOTCH1

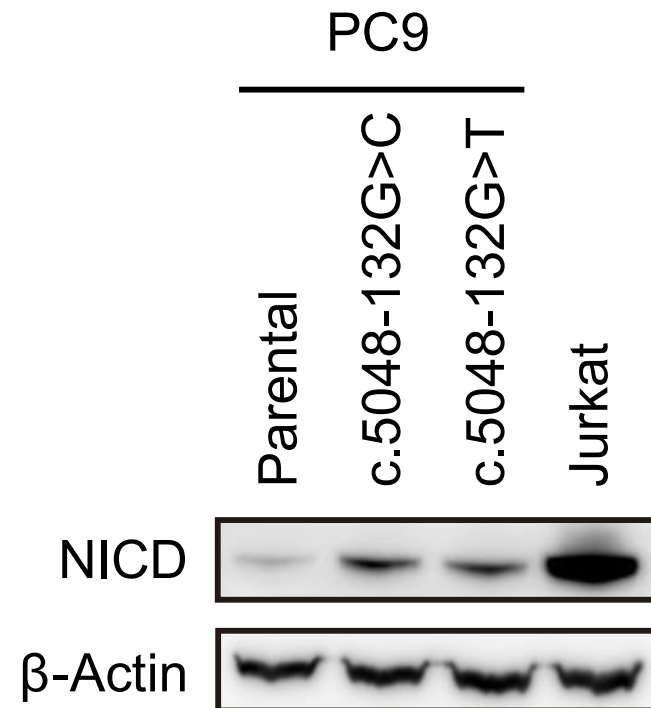


NOTCH1 JME by SSCV



NOTCH1のSSCVのゲノム編集モデルの構築、 核酸医薬による制御

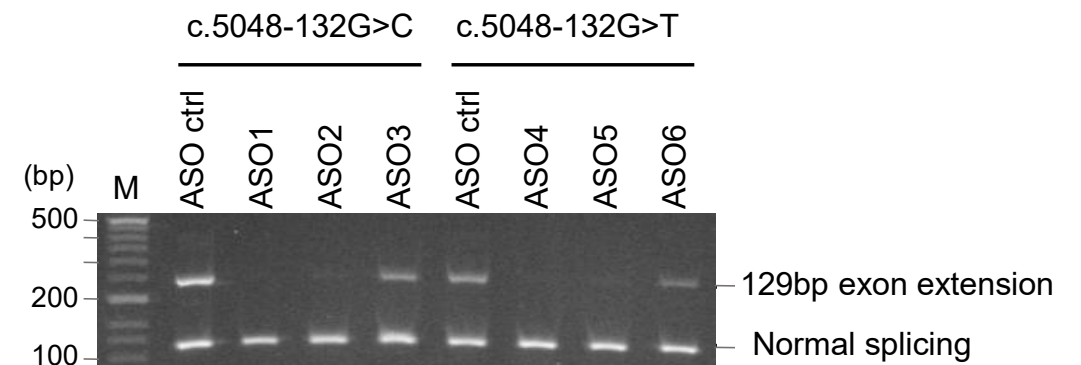
SSCVのゲノム編集モデル



核酸医薬によるスプライシングの制御

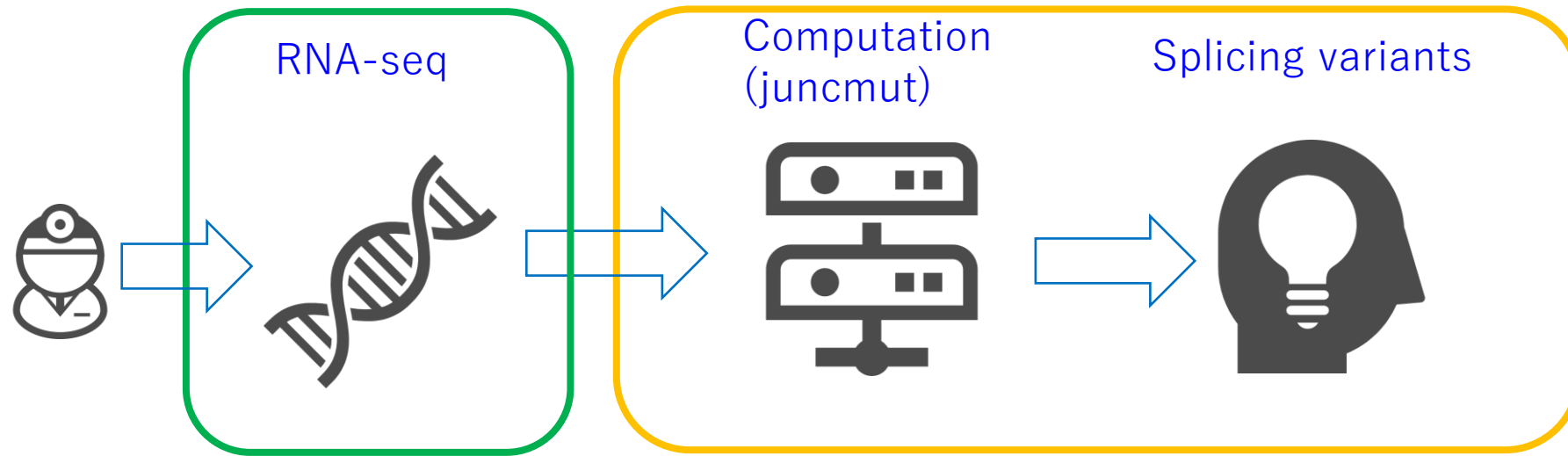
Reference: AAATGACCTTATTTTGGAGGGCGGGATGGAGGAGAGTGG
c.5048-132G>C: AAATGACCTTATTTTGCAGGGCGGGATGGAGGAGAGTGG
c.5048-132G>T: AAATGACCTTATTTTGTAGGGCGGGATGGAGGAGAGTGG

ASO1 3' AAAACGTCCTCCGCTACCTCTCTC 5'
ASO2 3' ATAAAACGTCCTCCGCTACCTCC 5'
ASO3 3' ACTGGAATAAAACGTCCTCCGCT 5'
ASO4 3' AAAACATCCCGCTACCTCTCTC 5'
ASO5 3' ATAAAACATCCCGCTACCTCC 5'
ASO6 3' ACTGGAATAAAACATCCCGCT 5'



自律的な「知識」の獲得

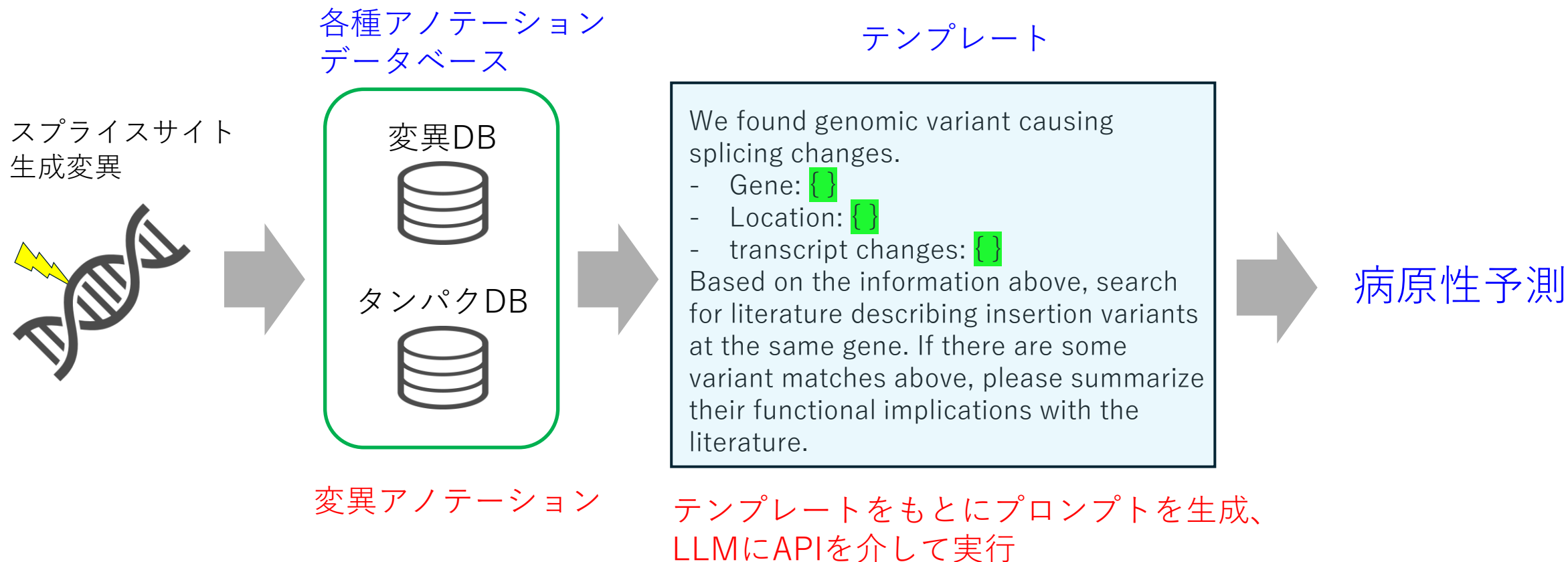
- スプライシング変異（創薬ターゲット）がRNA-seqデータの蓄積に合わせて「自律的」に収集されるインフラストラクチャー



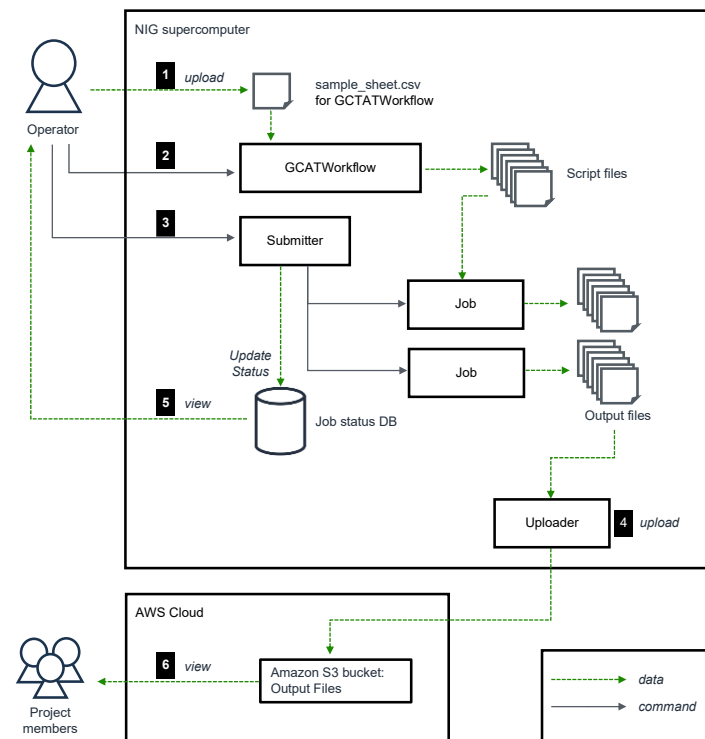
Juncmutの特性

- RNA-seqのみでゲノム変異が検出できる。
- データのグループ分けが不要なく、1検体だけで実行可能である。
- 正答率: ~95%, 感度: 25~35%.

LLMを用いた病原性予測の概要



30万件以上のRNA-seqを使ったSSCVのスクリーニング

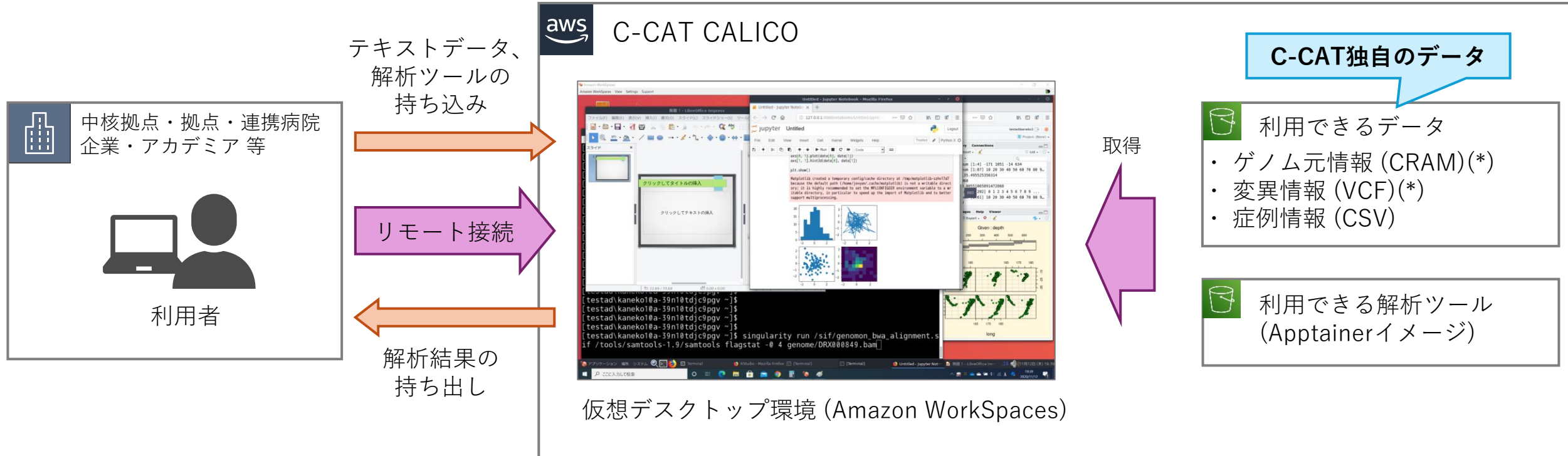


- 半年ほど、DDBJ supercomputerで計算を続ける。
- DDBJ supercomputerには、SRA, EGA, DRAのデータがミラーリングされていた！
- 計算はほとんど STAR によるアラインメントの部分（95%以上）。
- もし、STARによるアラインメントがなされていたら、計算コストは大幅に短縮されていた。

C-CAT CALICO (CALculation & Investigation ClOud) とは

遺伝子パネル検査のゲノム元情報 (FASTQ, CRAM)に基づいて多種多様な解析するための仮想デスクトップ環境を、企業、アカデミア向けに提供するサービスです。

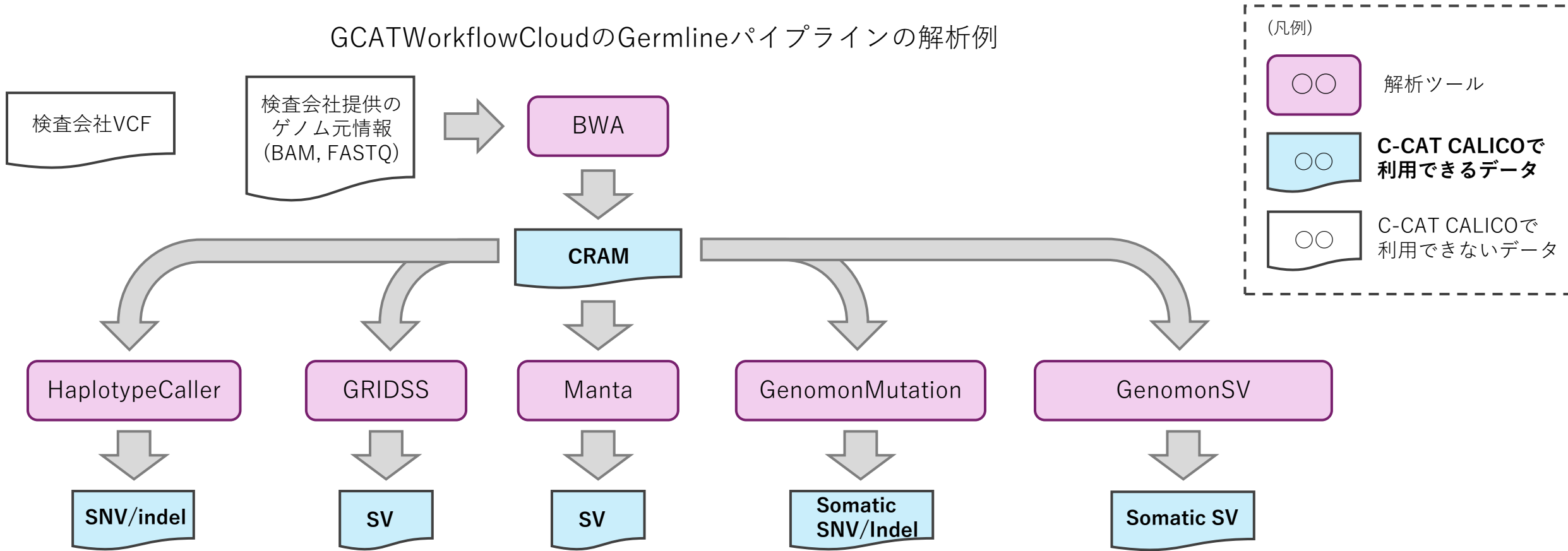
症例数: 57,184 (F1: 49,969、NOP: 7,215) (2025/3時点)



- 仮想デスクトップ環境内にゲノム元情報や変異情報をダウンロードし、C-CATが準備した解析ツールを使って、解析を行うことができます。

C-CAT CALICOで利用できるゲノム・変異データ

- 統一化された解析パイプラインによる変異検出・偽陽性除去・アノテーション付与。
- 利用者はそれらの処理結果を使い、即座に様々な集計を開始することができます！



GCATWorkflowCloud: <https://github.com/ncc-ccat-gap/GCATWorkflowCloud>

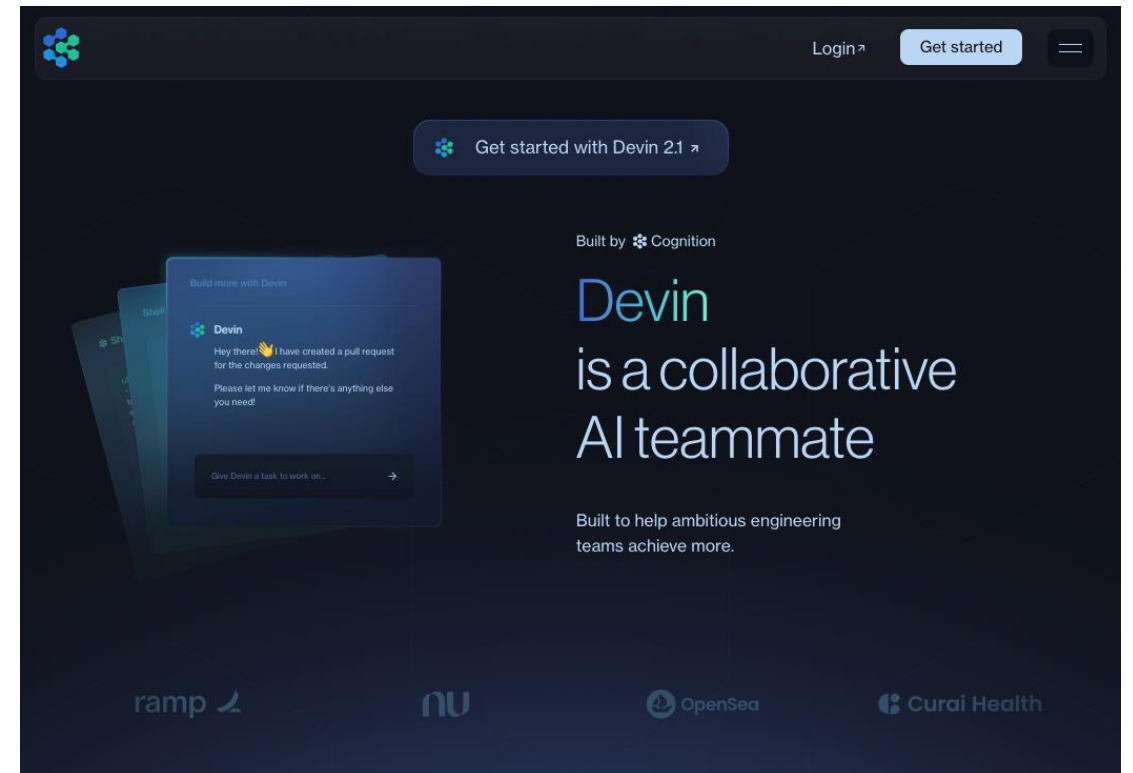
ソフトウェア開発の方法の変化

対話形式でコードを生成（Cursor）



<https://www.cursor.com/en>

自律的にソフトウェアを生成・改善(Devin)



<https://devin.ai/>

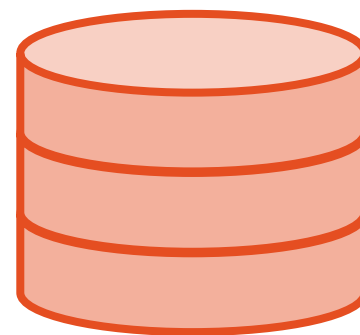
今後のゲノムの変遷（予想）

現在 研究者自らがレポジトリ上のデータを解析し、知識の発見を行っている（AIに相談はある）。

1～3年後 研究者はAIに解析の指示や結果の承認を行い、実際の解析の多くのをAIが担うようになる。

5年後以降 AIが自律的にゲノムデータを解析し、新たな知識を自ら発見・報告する段階に移行する（研究者はAIのメンテナンス）。

ゲノムデータレポジトリ
(一次加工済み)



解析



LLM

助言



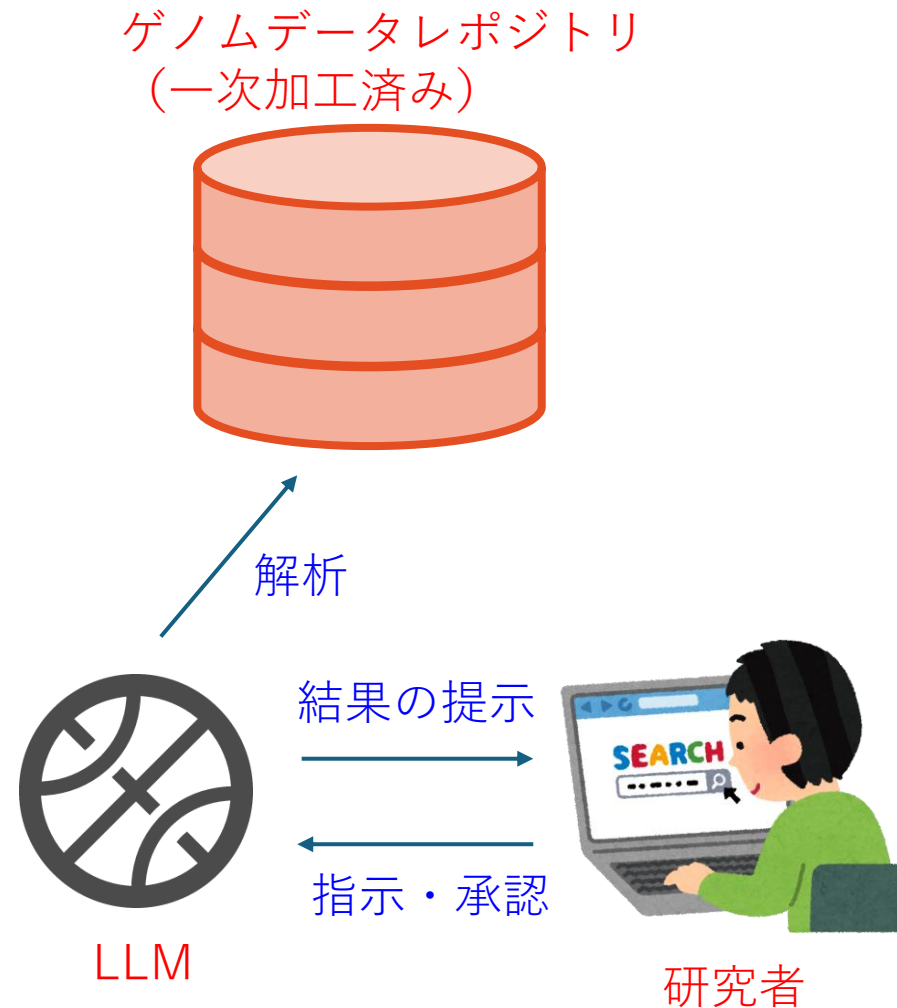
研究者

今後のゲノムの変遷（予想）

現在 研究者自らがレポジトリ上のデータを解析し、知識の発見を行っている（AIに相談はある）。

1～3年後 研究者はAIに解析の指示や結果の承認を行い、実際の解析の多くのをAIが担うようになる。

5年後以降 AIが自律的にゲノムデータを解析し、新たな知識を自ら発見・報告する段階に移行する（研究者はAIのメンテナンス）。

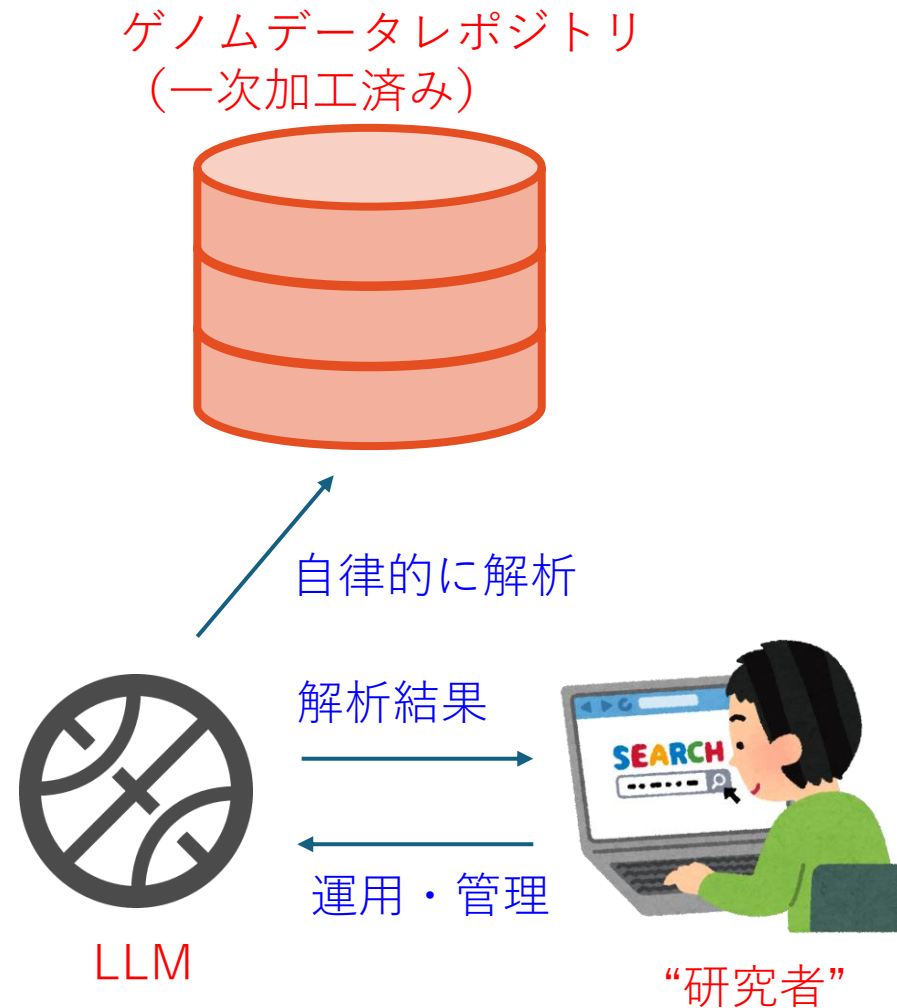


今後のゲノムの変遷（予想）

現在 研究者自らがレポジトリ上のデータを解析し、知識の発見を行っている（AIに相談はある）。

1～3年後 研究者はAIに解析の指示や結果の承認を行い、実際の解析の多くのをAIが担うようになる。

5年後以降 AIが自律的にゲノムデータを解析し、新たな知識を自ら発見・報告する段階に移行する（“研究者”はAIの運用・管理）。



まとめ

- 生成AIの到来により、解析・知識発見のプロセスは今後さらに自律化・自動化が進むだろう。
- データレポジトリにおけるデータの標準化や規格化が極めて重要である。
- AIが解析ジョブを効率的に実行できるスーパーコンピュータ環境の整備が求められる。
- その結果、研究者や大学の役割や存在意義も大きく変化することが予想される。

Acknowledgement

- 国立がん研究センター
 - ゲノム解析基盤開発分野
 - 岡田 愛
 - 千葉 健一
 - 飯田 直子
 - 分子病理学分野
 - 小林 祥久
- Funding
 - AMED革新的がん医療実用化研究事業
 - 統合化推進プログラム（DICP）



Portal site

