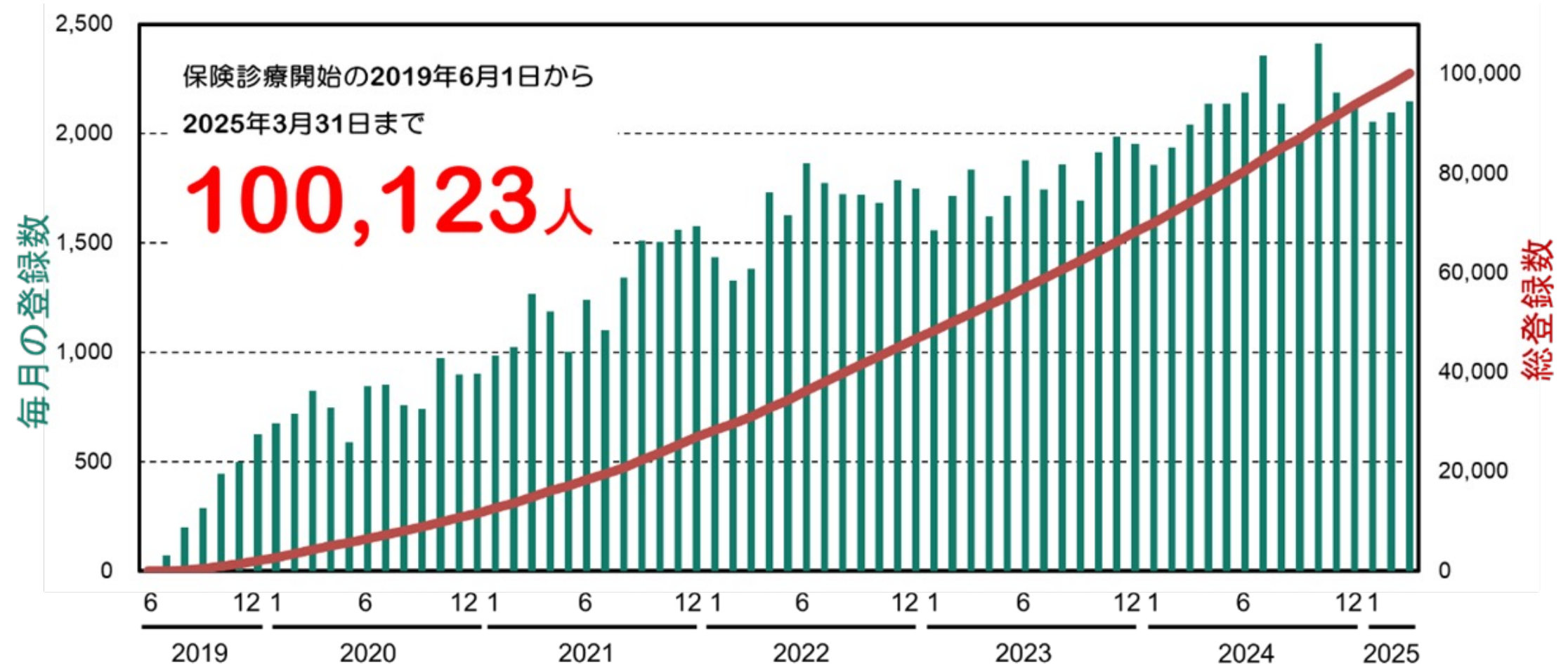


大規模言語モデルを活用した 病的スプライシング変異データベースの 自律的構築

国立がん研究センター・研究所
ゲノム解析基盤開発分野 白石 友一

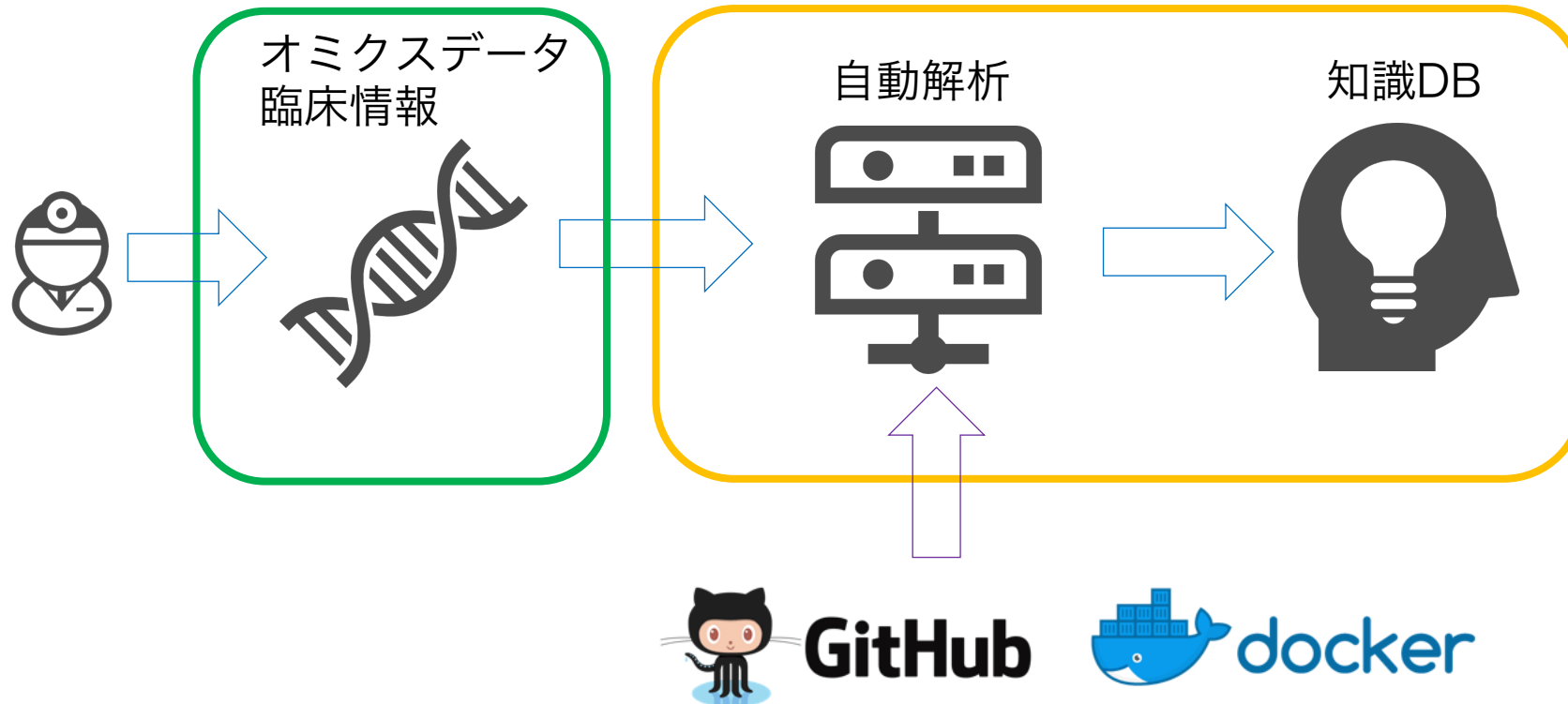
ゲノムデータの蓄積（C-CATを例に）



保険診療でのがん遺伝子パネル検査のC-CATへの登録患者数が10万例に到達

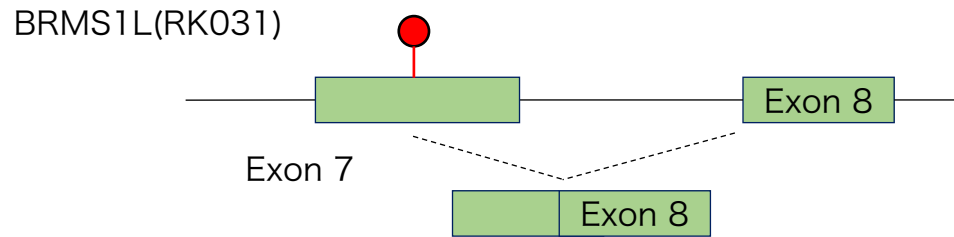
ゲノムデータ利活用の方法

- 「知識」が自律的に蓄積していくシステム。

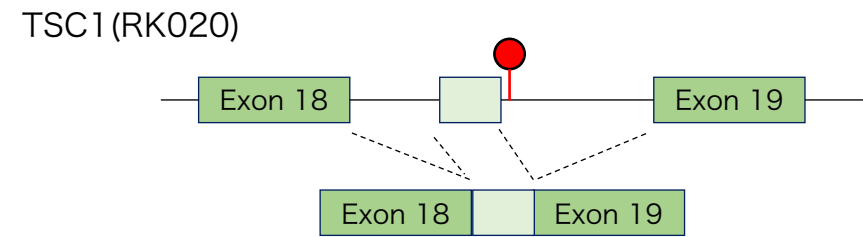


Splice-Site Creating Variant, SSCV

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

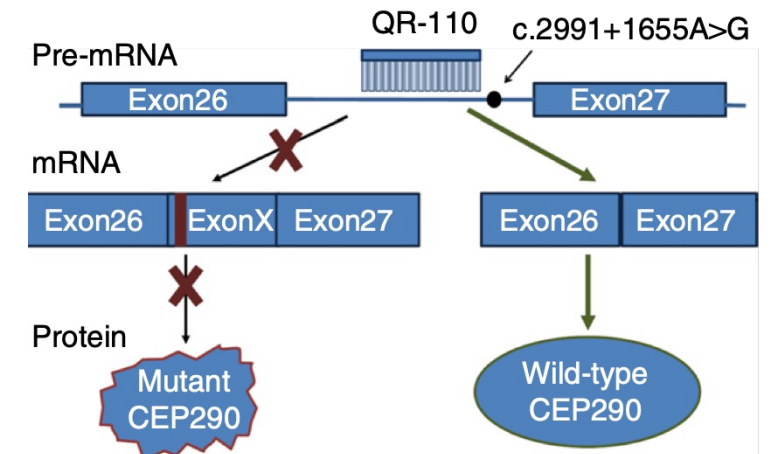


Truncation of 35bp of exon 7



Inclusion of 61bp sequence in deep intron

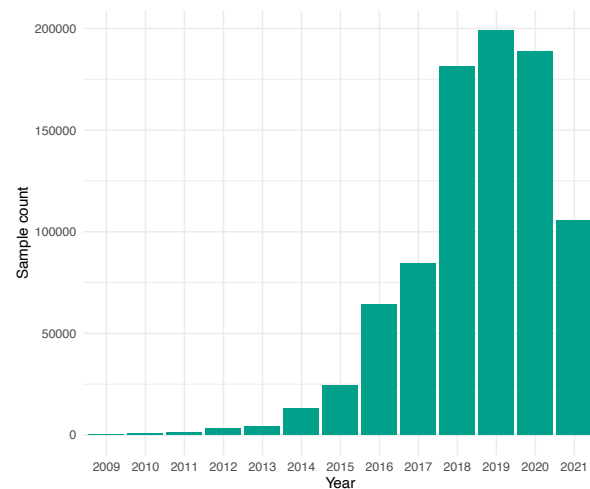
- 全ゲノム解析で検出される変異の解釈
 - 深部イントロンにおいて偽エキソンを生成する変異の同定
- 核酸医薬（アンチセンスオリゴヌクレオチド）による治療の可能性
 - レーバー先天性黒内障の *CEP290* の c.2991+1655A>G を対象とする核酸医薬(Cideciyan et al., Nature Medicine, 2019)。
 - N-of-1 medicine! (Kim et al., Nature, 2023).



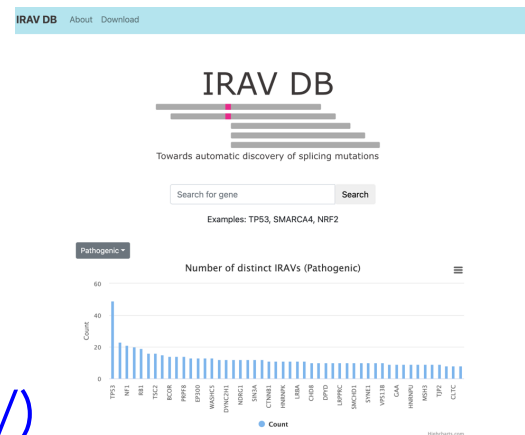
Cideciyan et al., Nature Medicine, 2019

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を両方利用したスプライシング変異の検出？
 - SAVNet (Shiraishi et al., Genome Research, 2018, PCAWG et al., Nature, 2020)
 - 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.
- 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？
 - Sequencing Read Archiveの数十万件のトランスクリプトームデータが使える！
 - Intron retentionを引き起こす変異 (Shiraishi et al., Nature Communications, 2022)



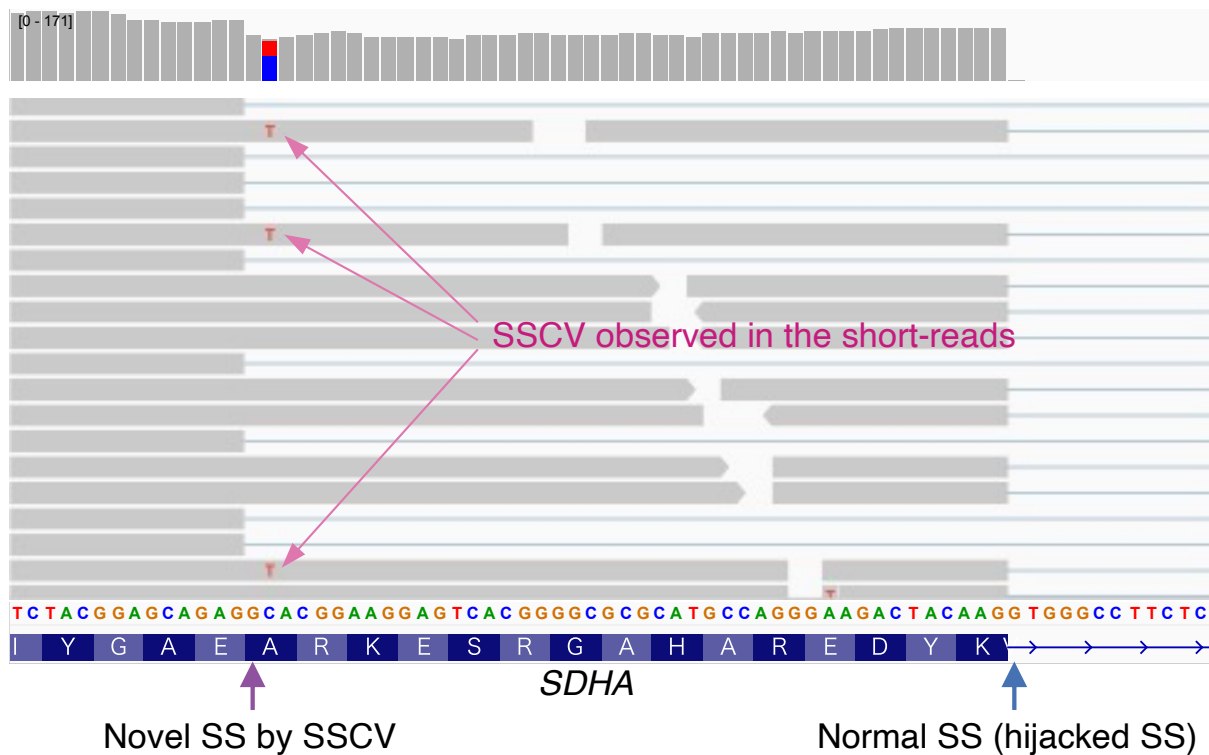
100万件以上のRNA-seqが「申請の必要がなく」利用可能！



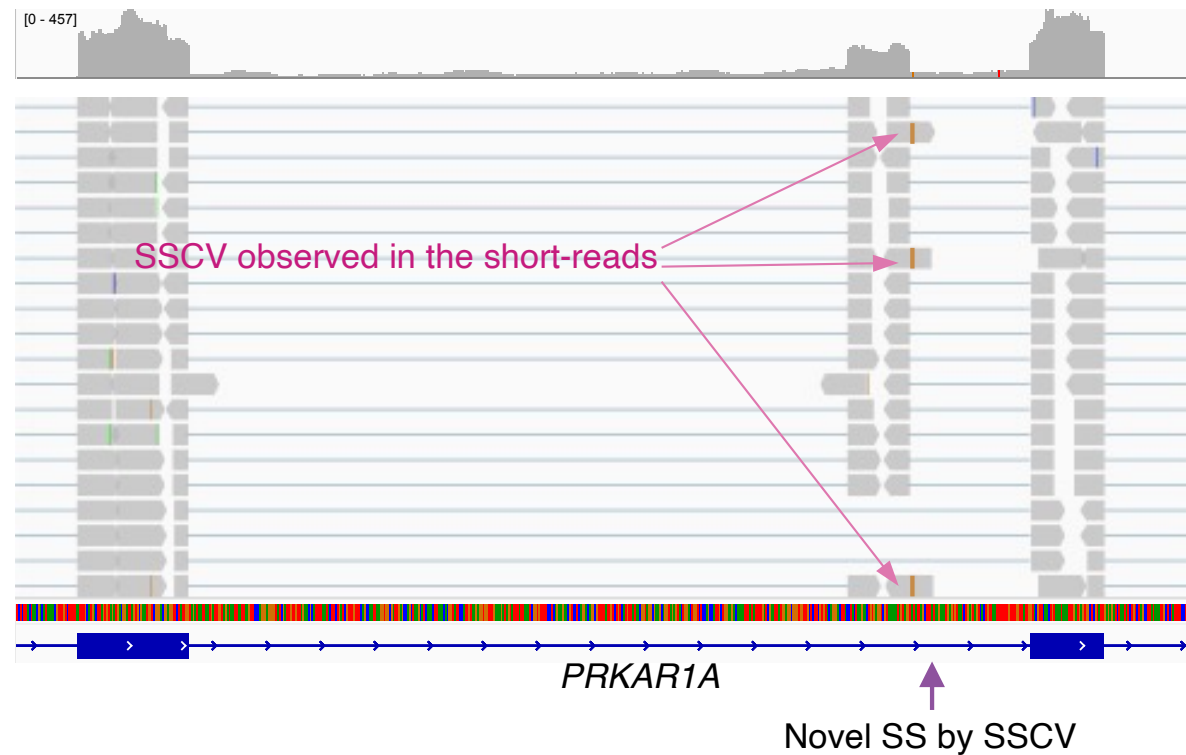
IRAV DB (<https://iravdb.io/>)

RNA-seqだけからSSCVを検出する？

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



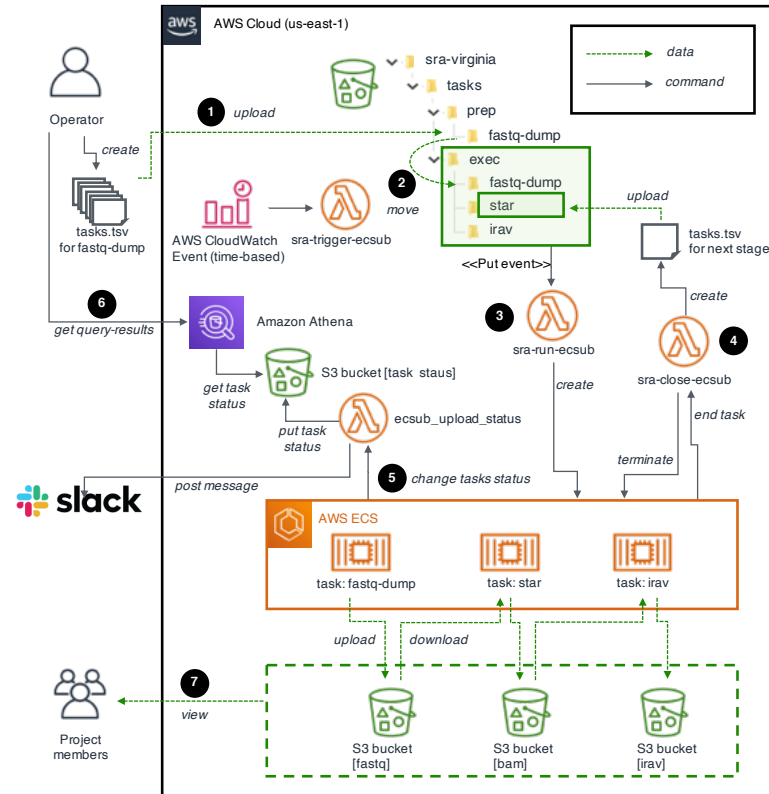
TCGA-OR-A5JY (cryptic exon inclusion by SSCV)



RNA-seqのリードからSSCVが観測できる！

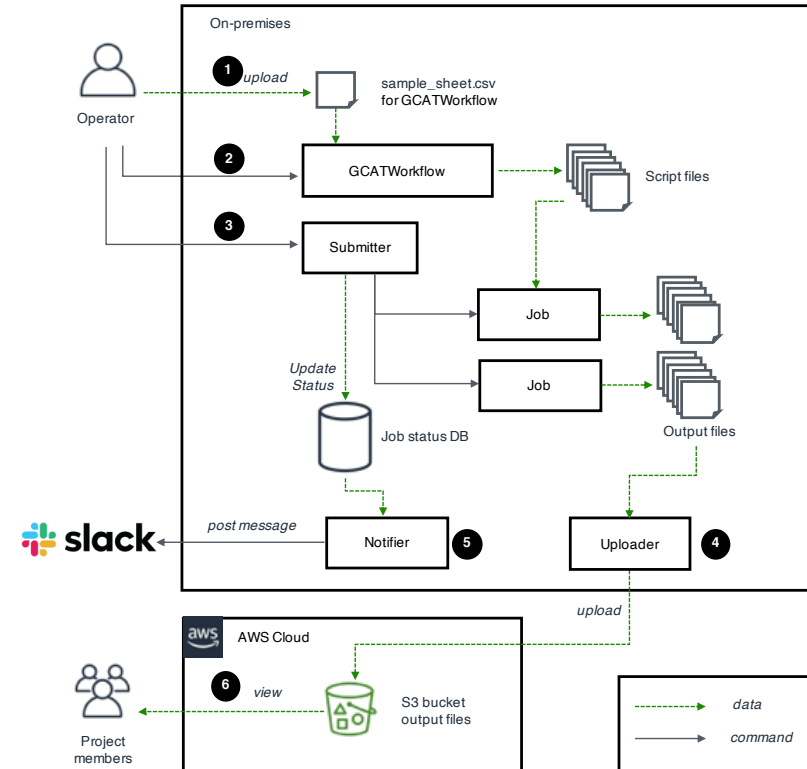
Sequence Read Archiveの30万件のRNA-seqのスクリーニング

Amazon Web Services



- 色々な最適化の上 \$0.1~0.2 / sampleに。
- サーバーの調達・管理の必要がない。
- きめ細やかなプラットフォーム構築が可能。

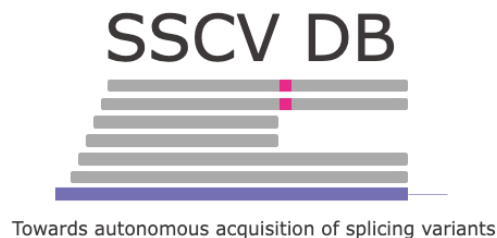
遺伝学研究所 スパコン



- SRAのデータの大部分がミラーリングされている。
- 気楽に「やり直し」ができる。
- 安い。

SSCV DB

SSCV DB About Download



24,458

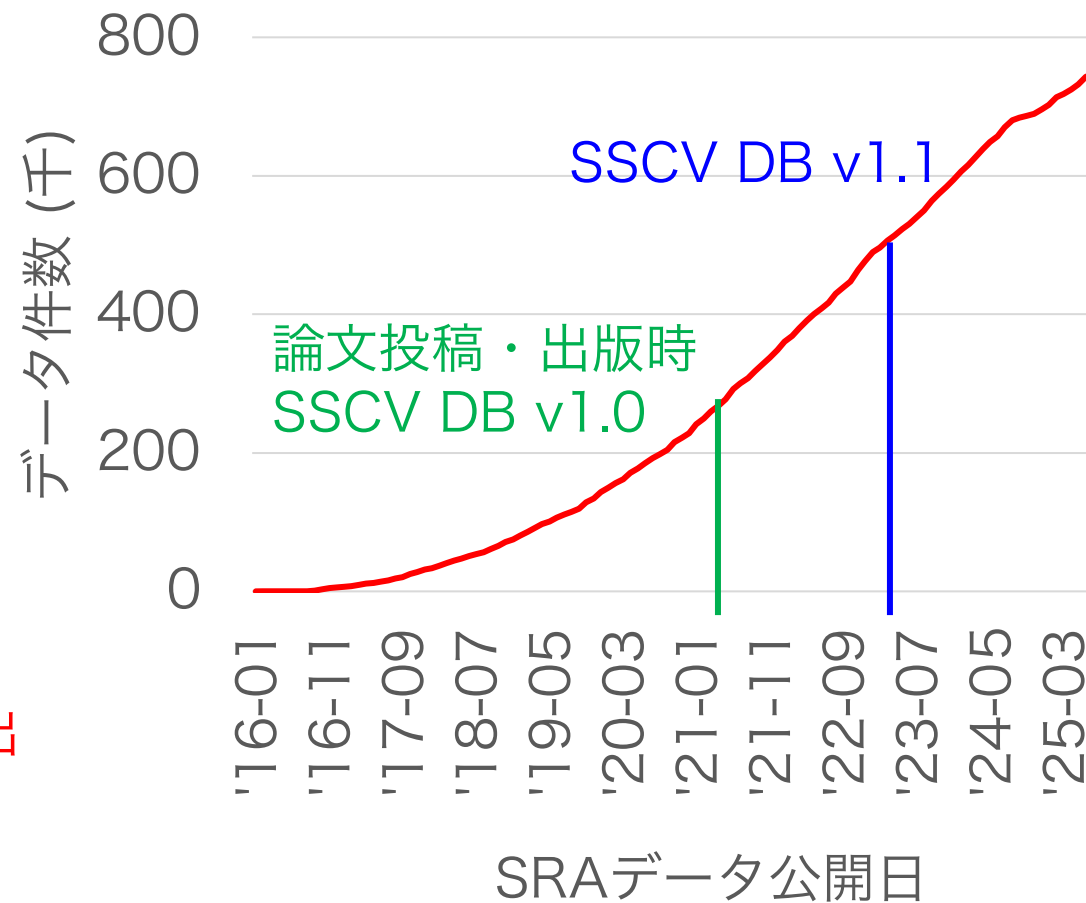
SRA project number

525,323

SRA runID number

48,569

SSCV (distinct) number



現在52万件のRNA-seqを処理し、48,569のSSCVを検出



Iida et al., Nature Communications, 2025

CREBBP

CREBBP

Cancer Gene Census

ClinGen Dosage Sensitivity

Clinical Genomics DataBase

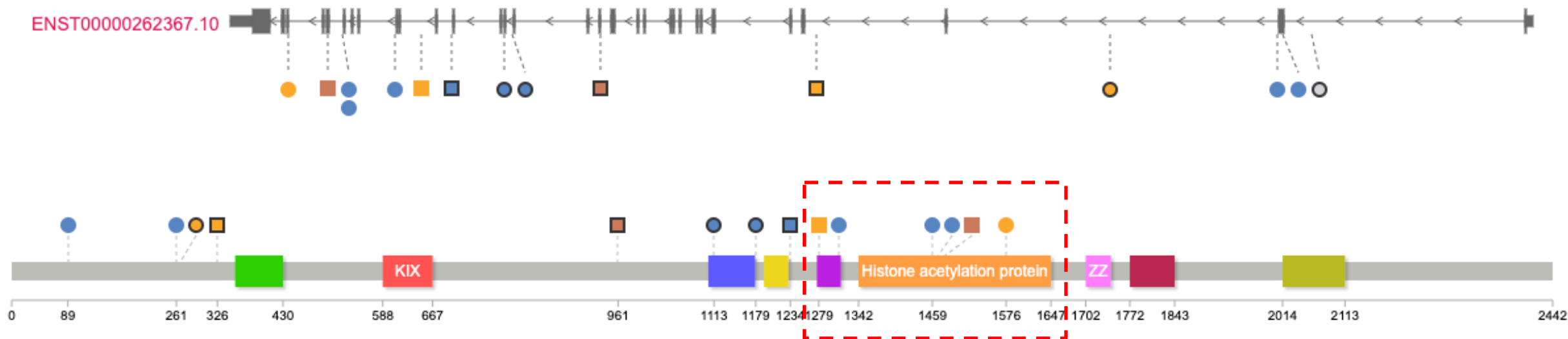
OMIM

Gene Alias CBP, KAT3A, MKHK1, RSTS, RSTS1

Region (GRCh38) chr16: 3725053-3880713

CREBBP, chr16

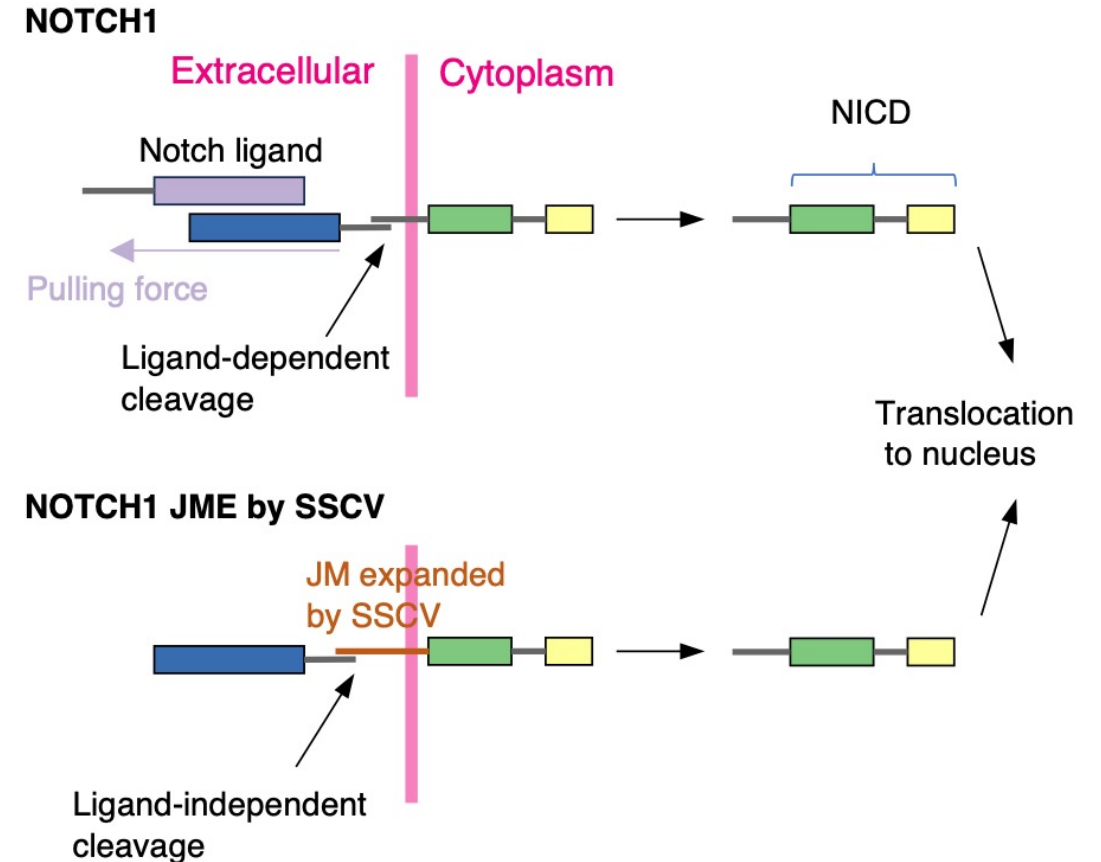
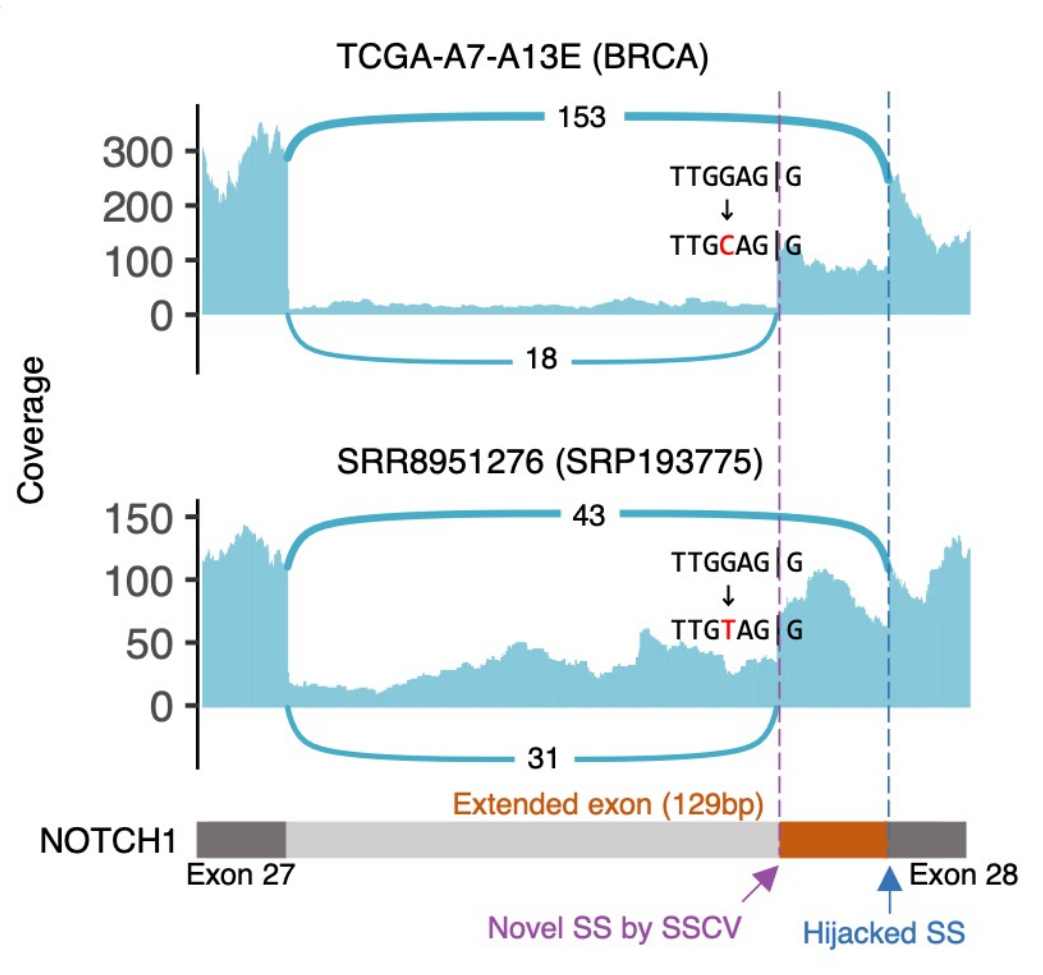
ENST00000262367.10



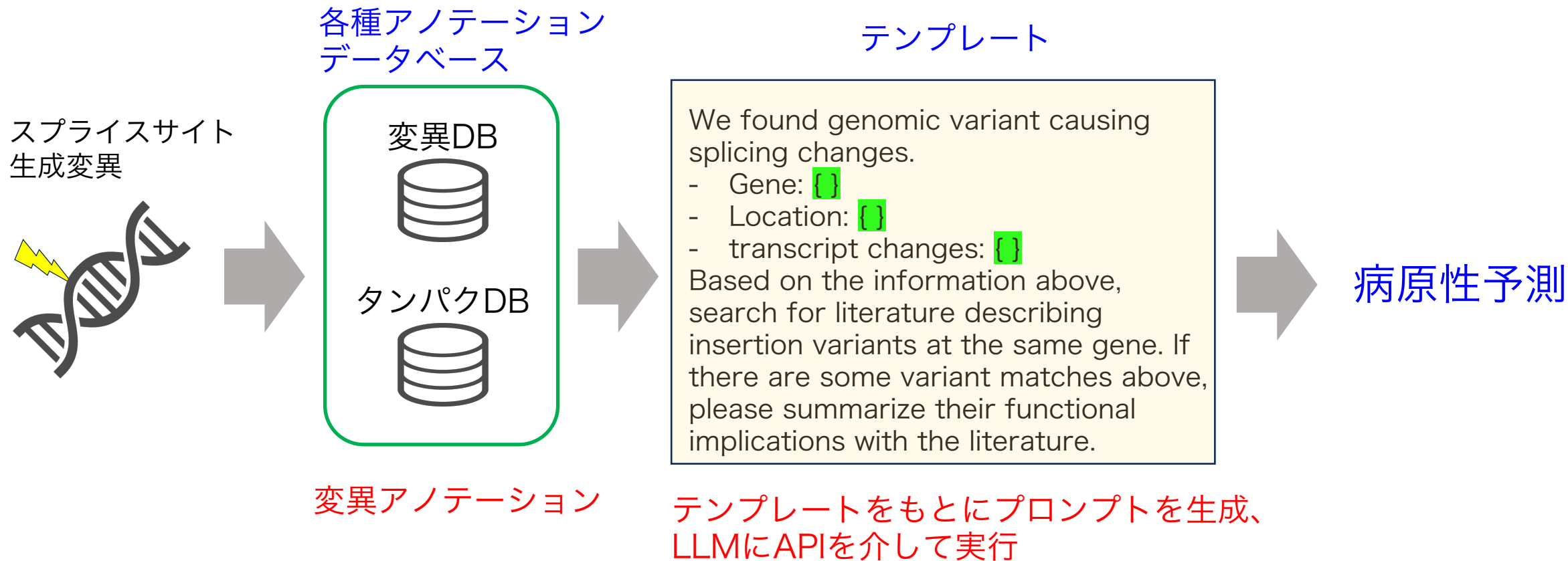
KAT domainにおけるin-frameを引き起こすSSCV

NOTCH1における新規gain-of-function変異

SSCVによるIn-frame insertionの生成

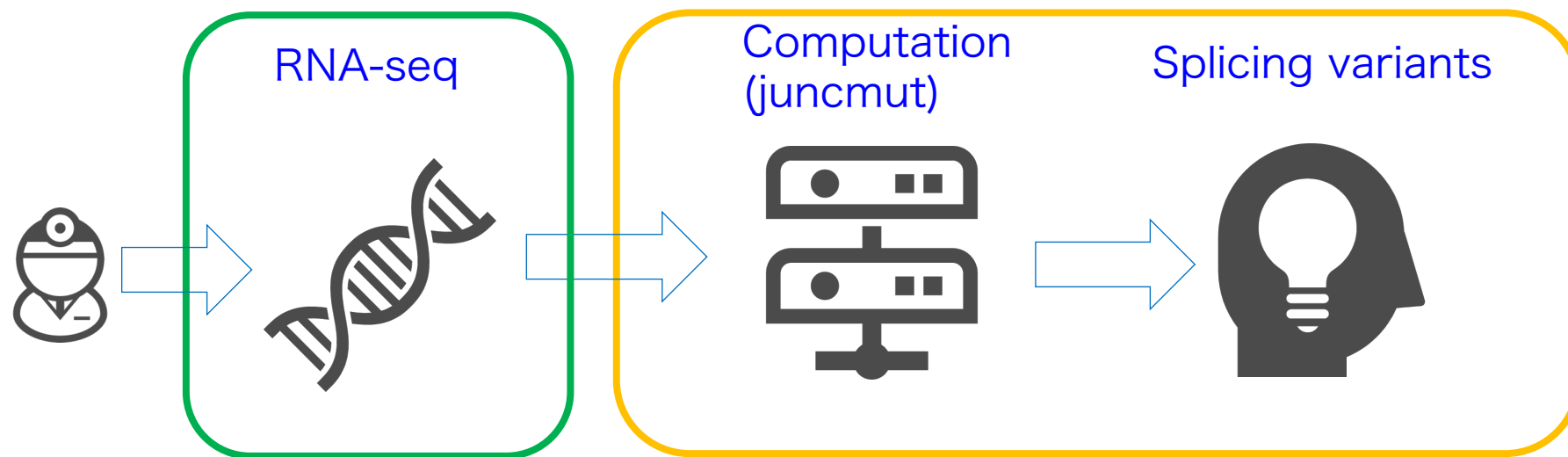


LLMを用いた病原性予測の概要



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%.

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
 - ラウル・マテオス
 - 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

