

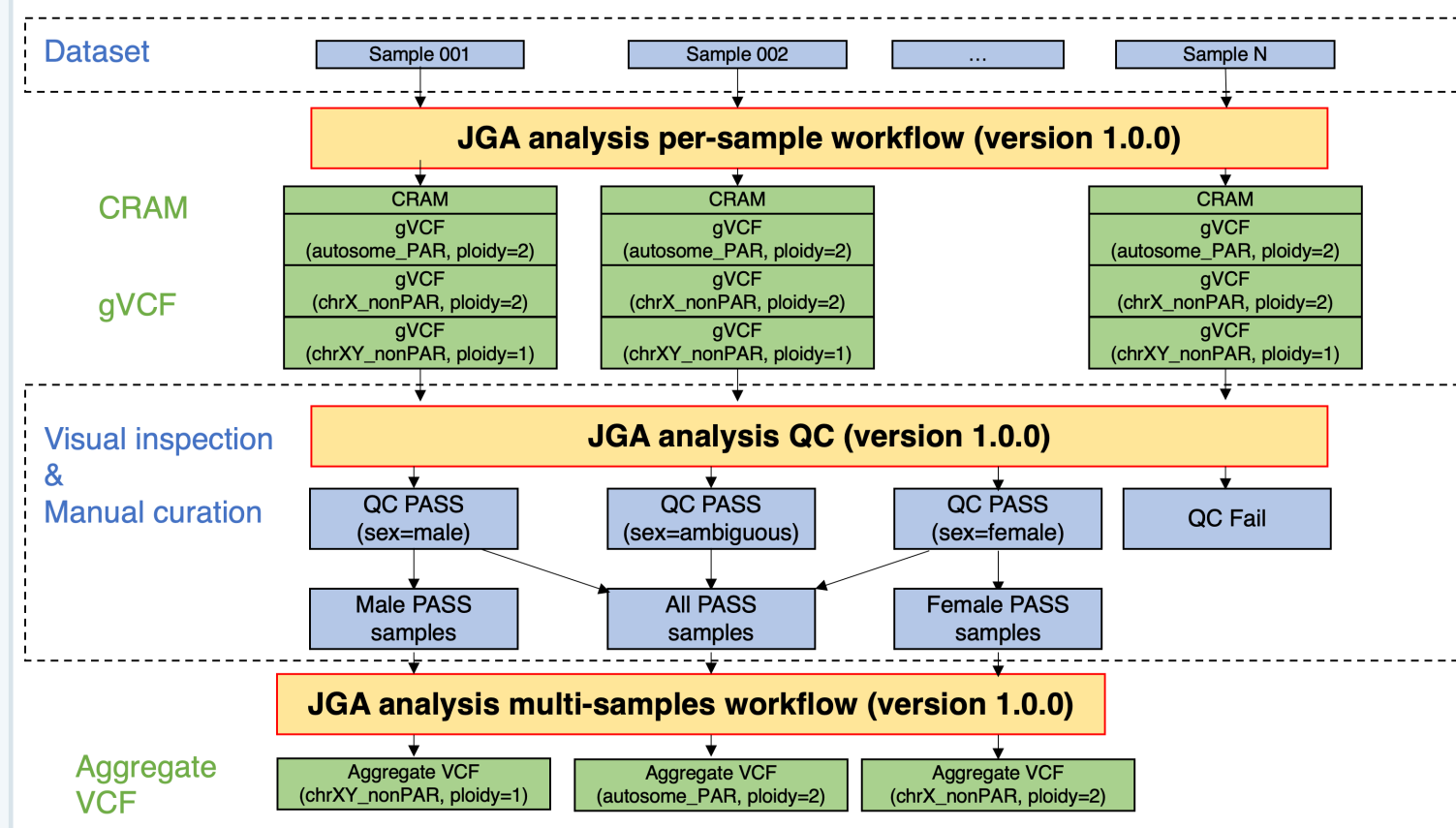
JGAデータ加工とTogoImputation

山本謙太郎¹、石井学¹、八谷剛史¹、藤澤貴智²、三橋信孝²、川嶋実苗²

¹株式会社ゲノムアナリティクスジャパン（GAJ）、²情報・システム研究機構 国立遺伝学研究所（NIG）

JGAデータ加工

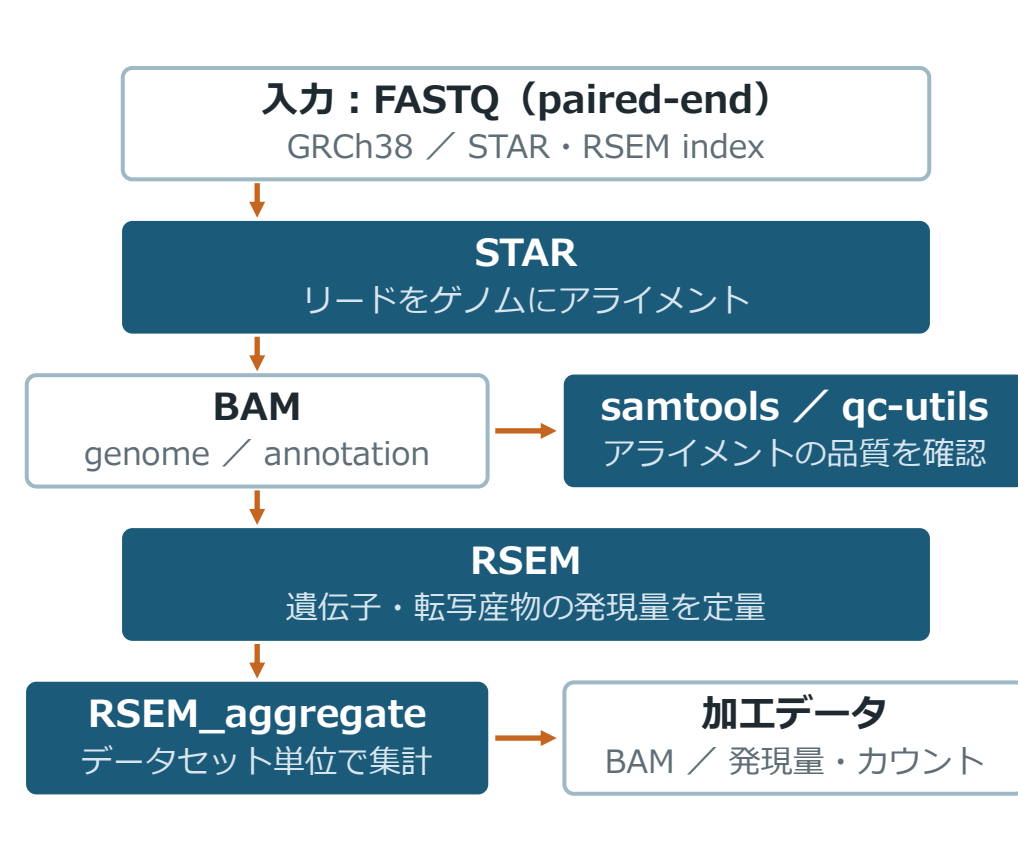
生殖系列の全ゲノムシーケンスタデータの加工



※引用： <https://github.com/ddbj/jga-analysis/blob/main/JGA-germline-WGS-workflow.png>

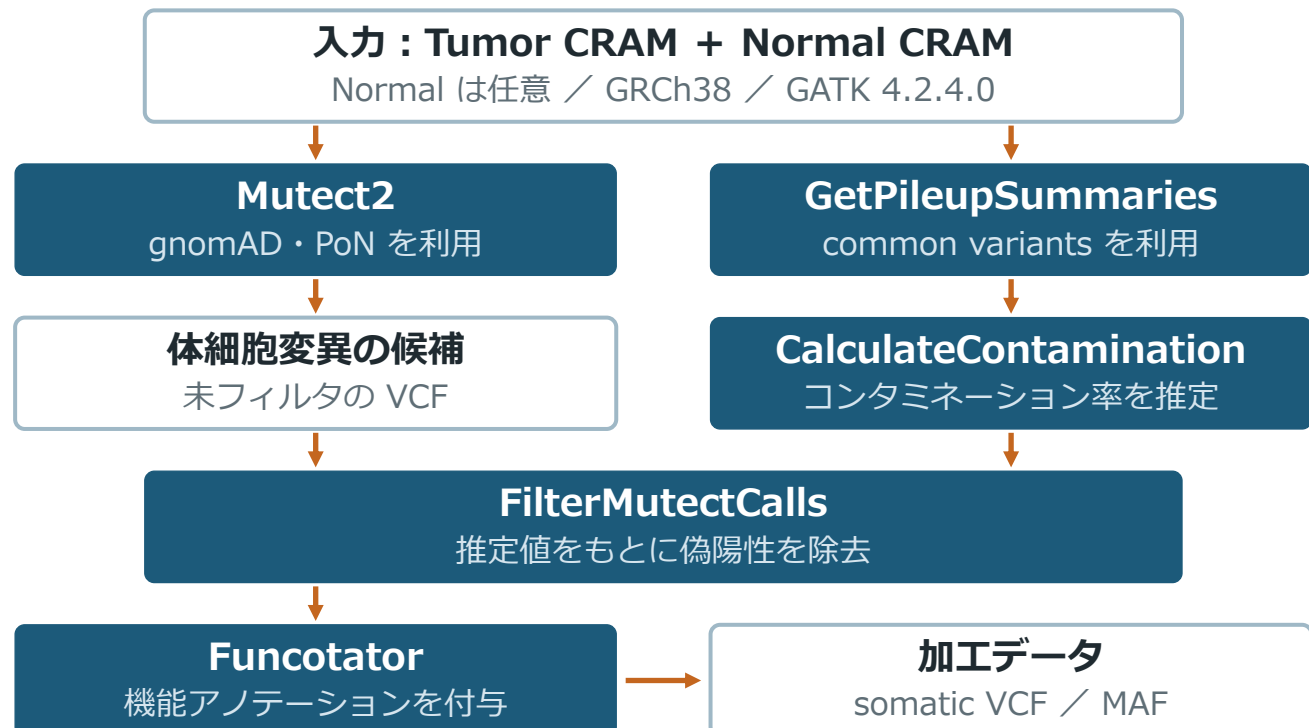
- 解析環境
NIGスパコンのヒトゲノムデータを扱える閉じた計算環境
- CWLワークフロー (jga-analysis : per-sample workflow)
GATK best practice に準拠し、CWL で実装・公開
- ① per-sample : FASTQ を GRCh38 にアライメント、サンプルごとにバリエーション検出
- ② QC : coverage 分布で外れ値を除外し、ゲノムデータから性別を推定
- ③ multi-samples : joint call と VQSR を実施
- 得られる加工データ
CRAM : アライメント結果 (サンプルごと)
gVCF : バリエーション検出結果 (サンプルごと)
aggregate VCF : バリエーション検出結果 (データセットごと)
sites-only aggregate VCF : 個人ごとのデータを除いた統計情報

RNA-Seqワークフローを用いたデータ加工



- 解析環境
NIGスパコンのヒトゲノムデータを扱える閉じた計算環境
- CWLワークフロー (jga-analysis : rna-seq workflow)
rna-seq-pipeline-per-sample.wdl と rsem_aggregate_results.wdl に従い、CWL で実装・公開
- ① アライメント : STAR で FASTQ をゲノムにマッピングし、BAM を出力
- ② 品質確認 : samtools と qc-utils でアライメント結果を確認
- ③ 定量・集計 : RSEM で発現量を定量し、RSEM_aggregate でデータセット単位に集計
- 得られる加工データ
BAM : アライメント結果 (サンプルごと)
rsem_genes_tpm / expected_count : 遺伝子ごとの発現量・カウント
rsem_transcripts_tpm / expected_count : 転写産物ごとの発現量・カウント
rsem_transcripts_isopct : アイソフォームの比率

がん変異検出ワークフローを用いたデータ加工



- 解析環境
NIGスパコンのヒトゲノムデータを扱える閉じた計算環境
- CWLワークフロー (jga-analysis : somatic-short-variant)
GATK best practice に準拠し、CWL で実装・公開
- ① 変異検出 : Mutect2 で腫瘍検体から体細胞変異の候補を検出
- ② 品質評価 : GetPileupSummaries CalculateContamination でコンタミネーション率を推定
- ③ 絞り込み・注釈 : FilterMutectCalls で偽陽性を除去し、Funcotator で機能アノテーションを付与
- 得られる加工データ
somatic VCF : 偽陽性を除去した体細胞変異 (サンプルごと)
annotated MAF : 機能アノテーションを付与した変異一覧 (サンプルごと)

加工済み（データセット数）実績

公開済み（加工データのDataset ID）

- 生殖系列の全ゲノムシーケンス
・ JGAD000670
・ JGAD000689
・ JGAD000688
・ JGAD000687
・ JGAD000690
・ JGAD000758
■ Imputation Panel
・ JGAD000679
・ JGAD000867
・ JGAD000868

公開準備中（解析対象のDataset ID）

- 生殖系列の全ゲノムシーケンス ※multi-samples : Biggest Practices
・ JGAD000220 (BBJ1K), JGAD000495 (BBJ2K)
・ JGAD000831, JGAD000832, JGAD000833, JGAD000834
- RNA-Seqワークフロー
・ JGAD000798, JGAD000463, JGAD000696, JGAD000095, JGAD000111, JGAD000122, JGAD000139, JGAD000301, JGAD000376, JGAD000407, JGAD000446, JGAD000653, JGAD000457
- がん変異検出ワークフロー
・ JGAD000252, JGAD000235, JGAD000234, JGAD000116, JGAD000696, JGAD000095

成果物の提供

- ① 利用申請
- ② 審査・承認
- ③ 取得・利用

- 加工データの利用方法
① 利用申請 : 研究データ一覧から対象を選び、NBDCヒトデータ利用申請書を提出
② 審査・承認 : ヒトデータ審査委員会が審査し、承認後に利用方法が通知される
③ 取得・利用 : JGAからダウンロード。
※元データ（加工前）が承認済みであっても、加工データの利用申請は必要

Imputation Panel

パネル名	リファレンス	収載サンプル
GRCh37.BBJ1K_1KGP	GRCh37	BBJ 1,037名 + 1KGP 2,504名 (計 3,541名)
GRCh37.BBJ1K_1KGP-EAS	GRCh37	BBJ 1,037名 + 1KGP 東アジア 504名 (計 1,541名)
GRCh38.BBJ1K	GRCh38	BBJ 1,026名 (WGS)
GRCh38.BBJ2K	GRCh38	BBJ 1,964名 (WGS)

※ 出典 : NBDCヒトデータベース「加工済みデータ一覧 (Imputation reference)」
<https://humandb.bclis.jp/processed-data-imputation>

まとめ

- 生殖系列WGS・がん・RNA-SeqのワークフローをCWLで実装し、jga-analysisとして公開
- データセットによらず同一手順で加工し、結果の比較可能性と再現性を確保
- 加工データは元データに紐づけてJGAに登録、元データと併せて利用申請することで利用可能
- 生殖系列WGSのaggregate VCFはTogoImputationの参照パネルの基盤となる

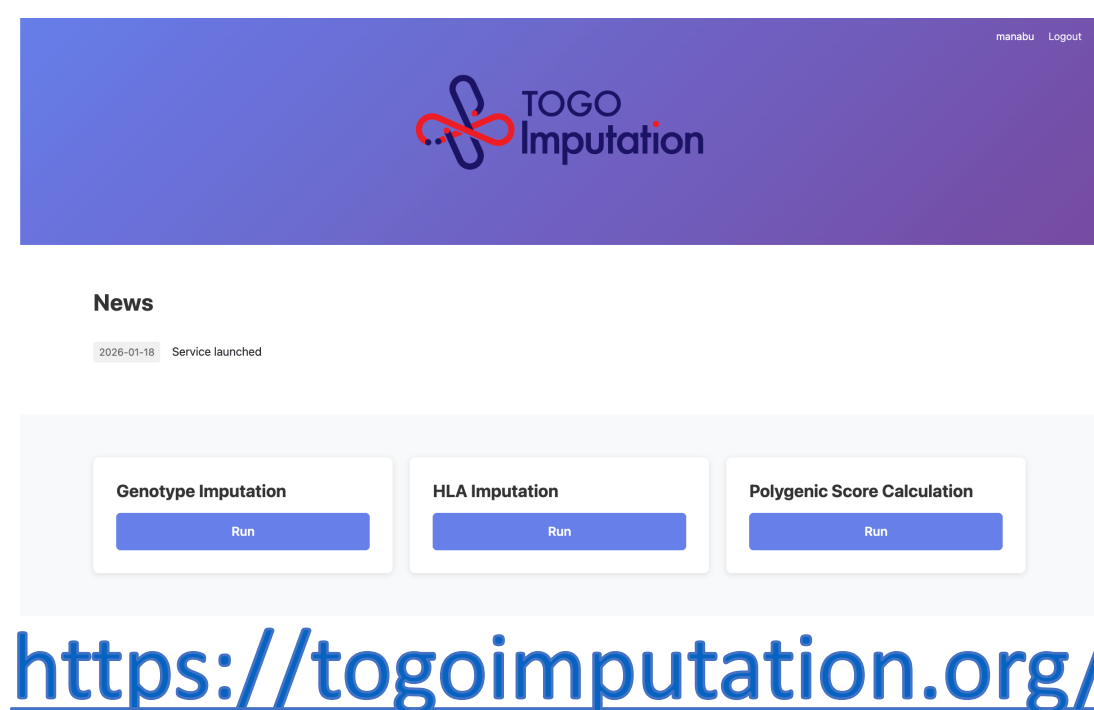
TogoImputation

概要

TogoImputation は、togoimputation.org から利用できるWeb解析サービスです。

- ・ Genotype Imputation
- ・ HLA Imputation
- ・ PGS Calculation

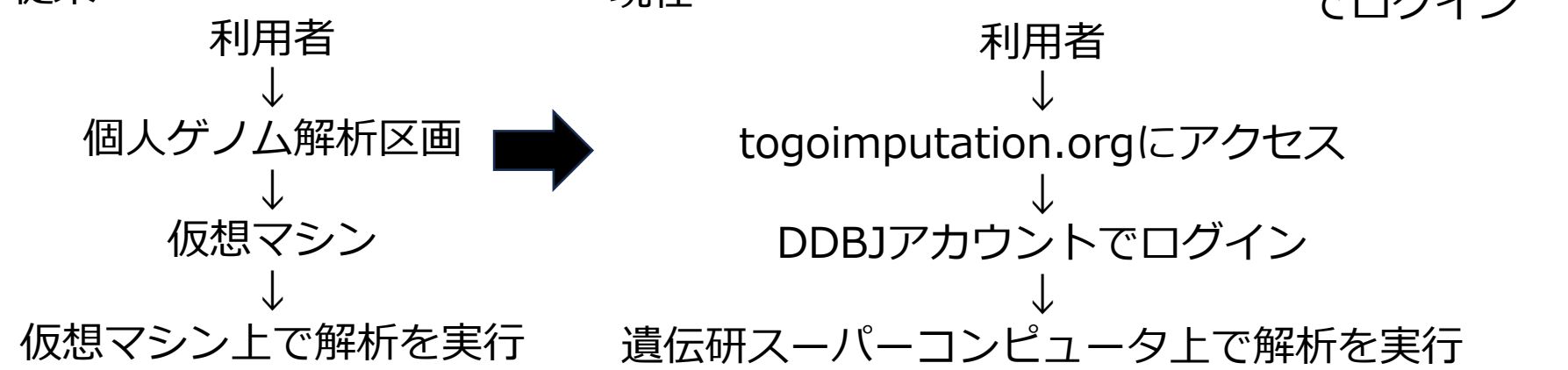
Webブラウザから解析条件を設定し、解析を実行・結果を取得できます。



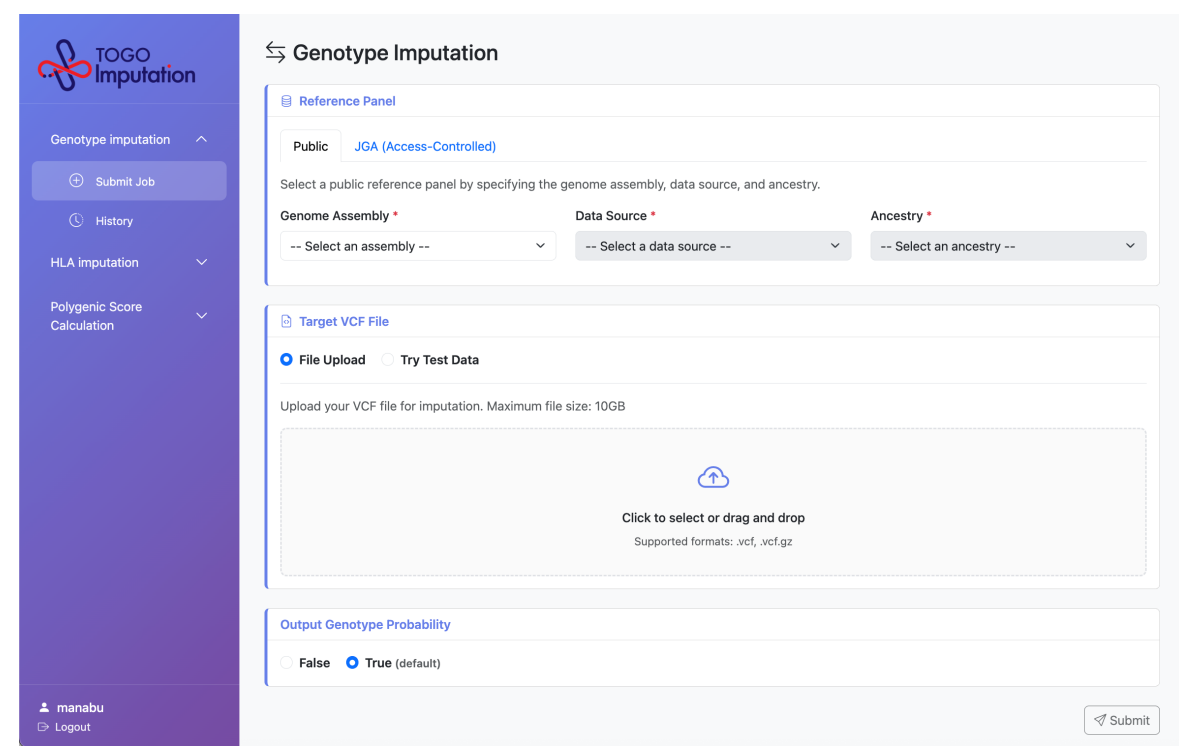
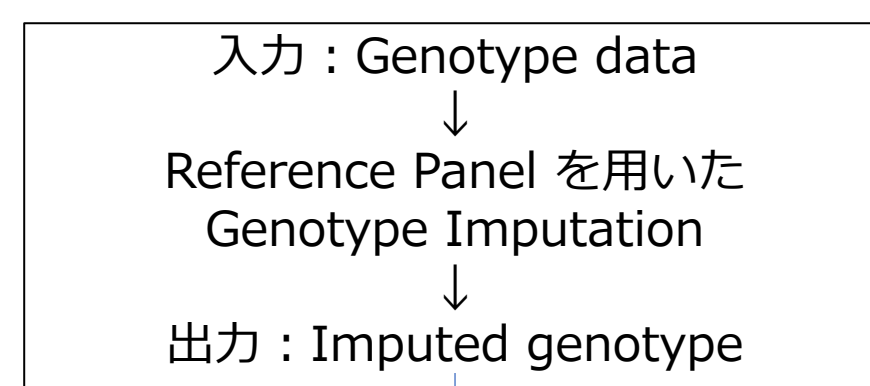
<https://togoimputation.org/>

構成の刷新

解析環境の貸し出し型から、Webサービス型へ移行

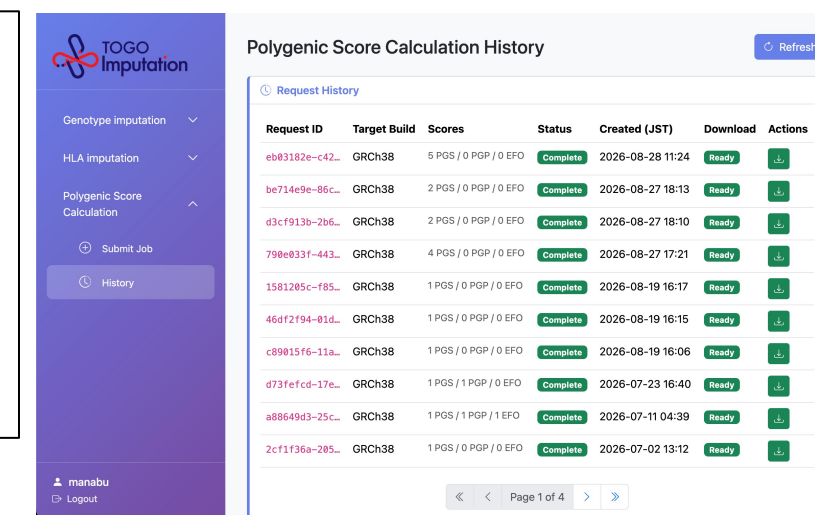
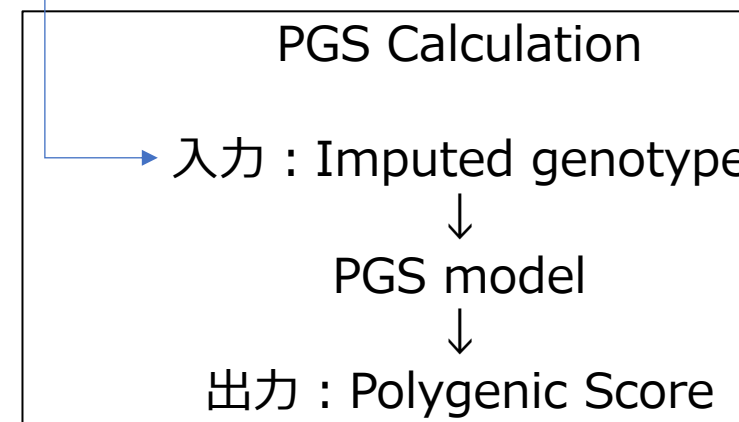
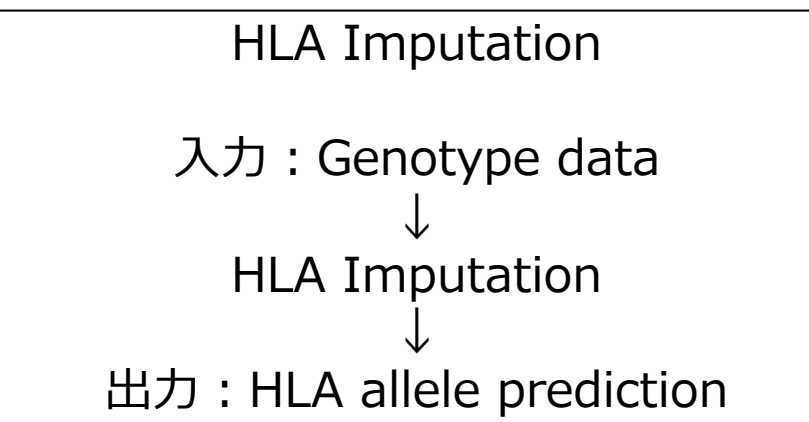


Genotype Imputation



解析設定画面：参照パネルを選択してジョブを実行

HLA Imputation ・ PGS Calculation



PGS Calculation : 計算結果の確認

利用方法

- ① togoimputation.org にアクセス
- ② 解析データをアップロード
- ③ 解析機能・条件を選択
- ④ ジョブを実行し、結果を取得

まとめ・今後の展望

まとめ

- ・ TogoImputationをWebサービスとして刷新
- ・ Genotype Imputation、HLA Imputation、PGS Calculation を提供

今後

- ・ 対応する参照パネル・解析モデル・利便性を拡充
- ・ 新たな解析機能の追加

