

データ提供機関	バイオバンク・ジャパン (BBJ)
対象	患者（難病やがん以外の疾患、出生地が47都道府県に均等に散らばるように選択）：6,000名
Platform	Illumina [NovaSeq 6000]
ソース	バイオバンク・ジャパン第2コホートから選定した対象者の末梢血（5検体は唾液）から抽出したDNA
ライブラリ作製方法（キット名）	TruSeq DNA PCR-Free Sample Prep Kit
断片化の方法	超音波断片化
ライブラリ構築方法	Paired-end
リード長（除：バーコード、アダプタ、プライマー、リンカー）	150bp
インサートサイズ	550bp
クオリティコントロール方法	以下の条件で全ゲノムシーケンス解析を実施 - インサートサイズピークが400-600bpにあることを確認 - QV30以上の塩基の割合が75%以上 - FASTQCによる重複リード除去後の総塩基数が900億塩基以上 アライメントとhaplotypcaller後に以下のサンプルを解析から除外した - 唾液サンプル - 性染色体のDepthが臨床情報の性別と矛盾するサンプル - bamから収集したメトリック（Depth、Coverage、Mismatch Rate）が異常値を示すサンプル - 性染色体については、Depthから性染色体異常（XO、XXX、XXY、XYY）の可能性があるサンプル バリエーションコールの結果は以下のフィルタリングを実行した - VQSRの結果をVCFのFILTERフィールドにセット（常染色体のみ）
重複するリードの除去方法	MarkDuplicates (GATK4.1.0.0) 互換アルゴリズム (Parabricks 3.6.1 fq2bam)
マッピング方法	bwa mem (v0.7.15) 互換アルゴリズム (Parabricks 3.6.1 fq2bam)
マッピングの際のリファレンス配列	GRCh38 (+HLA+decoy)
平均カバー率（Depth）	33.8（常染色体）
変異検出方法	HaplotypeCaller (GATK4.1.0.0) 互換アルゴリズム (Parabricks 3.6.1 haplotypcaller)
総データ量	552TiB (FASTQ、CRAM、gVCF)
利用にあたっての制限事項	