A Practical Workflow for Correcting Kit-Specific Effects in Whole-Exome Sequencing Data

Laura Jarosz , Marcel Ochocki , Julia Merta , Lajos Pusztai , Michal Marczyk *

Posted Date: 21 April 2026

doi: 10.20944/preprints202604.1444.v1

Keywords: batch effect; whole-exome sequencing; exome capture kit; imputation; genetic variants

 doi:10.20944/preprints202604.1444.v1

Article

A Practical Workflow for Correcting Kit-Specific Effects in Whole-Exome Sequencing Data

Laura Jarosz [1], Marcel Ochocki [1], Julia Merta [1], Lajos Pusztai [2] and Michal Marczyk [1,2,*]

[1] Department of Data Science and Engineering, Silesian University of Technology, Gliwice, Poland

[2] Yale Cancer Center, New Haven, CT, USA

* Correspondence: michal.marczyk@polsl.pl

Abstract

Large-scale, multi-center projects have become common in the era of rapid technological development, but protocol standardization remains challenging. In whole-exome sequencing (WES), various exome enrichment kits exhibit variable efficiency across genomic regions, leading to systematic, non-biological batch effects, much stronger than other technical factors. We propose a workflow to minimize the effect of WES capture inconsistencies in single-nucleotide variation (SNV) data. The pipeline consists of quality control, mapping to the genome, variant calling, joint genotyping, and imputing genotypes using reference haplotypes. Variants are then aggregated into gene-level features measuring the burden of deleterious mutations. Finally, a gene-level imputation is performed using a customized algorithm. Namely, if the detection rate of a gene is low in samples enriched with a given capture kit, but high in samples enriched with other kits, missing values in the former group are imputed, as such differences are unlikely to reflect true biology. As a benchmark, we conducted a study on over a thousand breast cancer cases across 11 cohorts, using 8 exome capture kits. We demonstrated that the proposed pipeline leads to a considerable decrease in the batch effect signal, potentially increasing the likelihood of finding true biological signals. The workflow is publicly available here: https://github.com/ZAEDPolSl/WESworkflow.

Keywords: batch effect; whole-exome sequencing; exome capture kit; imputation; genetic variants

1. Introduction

Technological advances in recent decades have enabled large-scale sequencing projects, allowing researchers to analyze numerous, diverse cohorts [1]. The 1000 Genomes Project (1kGP) is a pioneering population-level initiative that provided one of the first global references for human genetic variation [2]. A more cancer-focused example is The Cancer Genome Atlas (TCGA), which characterized the genomic landscape of 33 cancer types by collecting over 20,000 samples [3]. The most recent project is the UK Biobank, which includes over 500,000 participants, making it one of the largest genetic data resources [4]. Despite the invaluable benefits of expanding analytical capabilities, these projects also have some drawbacks. When designing a study at such a large scale, incorporating multiple sites is often necessary to provide feasibility. To illustrate, there are around 17 million containers with blood, urine, and saliva specimens held in the UK Biobank laboratories, highlighting the logistical complexity and infrastructure required to succeed with such a venture. As the number of collaborating sites increases, protocol standardization becomes more challenging [5]. Carrying out the experiments and processing hundreds of samples in different environments introduces technical variability, known as batch effects. This non-biological variation may considerably confound the results and lead to incorrect conclusions [6].

Whole-exome sequencing (WES) enables access to all protein coding sequences of the human genome, significantly reducing the costs and computational effort compared to whole-genome sequencing (WGS) [7]. In WES, technical variability can arise at multiple stages of the study. For example, a specimen preservation method involving fixation in formalin and embedding in paraffin

 doi:10.20944/preprints202604.1444.v1

(FFPE) is known to induce cytosine deamination, leading to C>T transition artifacts [8]. During the library preparation, differences in PCR can cause amplification biases, such as sequencing duplicates or inaccurate representation of the original sequence - especially in GC- and AT-rich regions [5,9]. Different tools used in the bioinformatic pipeline may lead to divergent results in mutation detection [10]. Finally, the use of different exome capture kits results in systematic coverage variability, with inefficiencies in certain target regions and an increased false negative rate [11]. Recently, we compared multiple technical and clinical characteristics of the samples and found that the highest variance was due to the capture efficiency differences between the various versions of the kits, which led to systematic gene-level detection biases in WES data [12]. These biases manifest as near-binary gene detection patterns, where variants in a given gene are discovered in almost all the samples enriched with one capture kit but none of those processed in another.

There are numerous methods for batch effect correction in transcriptomics. ComBat reduces the effects of known batches using the empirical Bayesian framework [13]. Limma facilitates batch effect removal by applying linear modeling techniques [14]. Harmony method, designed for single-cell RNA sequencing, adjusts for the batches while preserving biological signals [15]. However, to date, we are not aware of any comprehensive, dedicated framework to overcome these issues in variant data, since their nature is somehow different from transcriptomic data. Nevertheless, there are some methods that can be used to reduce some specific technical artifacts. Differential depth of coverage can be normalized through downsampling the overrepresented reads. GC-bias can be corrected with a dedicated model [16]. Multiple tools have been developed for dealing with artifacts in FFPE samples [17–20]. However, none of the described methods can be applied to WES variant data in terms of addressing the variability arising from differences in exome capture. To fulfill that gap, we propose a workflow aimed at minimizing batch effects primarily driven by differential enrichment efficiency in WES SNV data.

Our pipeline starts conventionally, with raw sequencing reads quality control, trimming, and mapping to the reference genome. The next step consists of calling the variants and joint genotyping the samples with a restriction to the custom regions of the genome. Variants are then post-processed and prepared for the genotype imputation with the reference haplotypes method. Subsequent aggregation into gene-level average allele fraction weighted by CADD Phred-like score allows to measure the burden of deleterious mutations within the genes. Finally, a gene-level feature imputation is performed, applying our custom algorithm. The core rule states that if the detection rate of the gene is extremely low in samples processed with the same exome capture kit but substantially higher in samples enriched with other kits, the missing values in the former group are imputed, as such differences are unlikely to reflect true biology. For benchmarking purposes, we conducted a study on over a thousand breast cancer cases across 11 cohorts enriched using 8 exome capture kits.

2. Materials and Methods

2.1. Pipeline Overview

The proposed method (Figure 1) is an end-to-end framework that integrates all processing steps from raw sequencing reads to a gene-level variant feature matrix, explicitly reducing batch effects driven by differences in exome capture efficiency. Pipeline is initiated by quality check (QC) of FASTQ files with a focus on sequence/base quality scores and adapter content. Depending on the QC results, reads are either trimmed first or directly aligned to the reference genome, followed by duplicate marking. Subsequently, variants are detected with DeepVariant, which generates both VCF and GVCF files for each sample independently [21]. GLnexus is utilized to merge and genotype GVCFs into a single cohort-level VCF [22]. Both of these steps are restricted to regions specified in a BED file built on NCBI RefSeq exonic coordinates [23]. Genotype imputation is carried out using Beagle with a reference panel built from phased VCF files from the 1kGP [24–26]. Feature calculation step utilizes both imputed and non-imputed files, as well as CADD prescored database, aggregating

 doi:10.20944/preprints202604.1444.v1

the variants into CADD-weighted average allele fraction (CWAF) [27]. Subsequent feature imputation is performed on the gene-level matrix (samples × genes), with a mask defining missing-not-at-random (MNAR) entries, which are selectively imputed. The implementation supports parallel execution across samples or chromosomes, depending on the current step. The code of the workflow is publicly available here: https://github.com/ZAEDPolSl/WESworkflow.

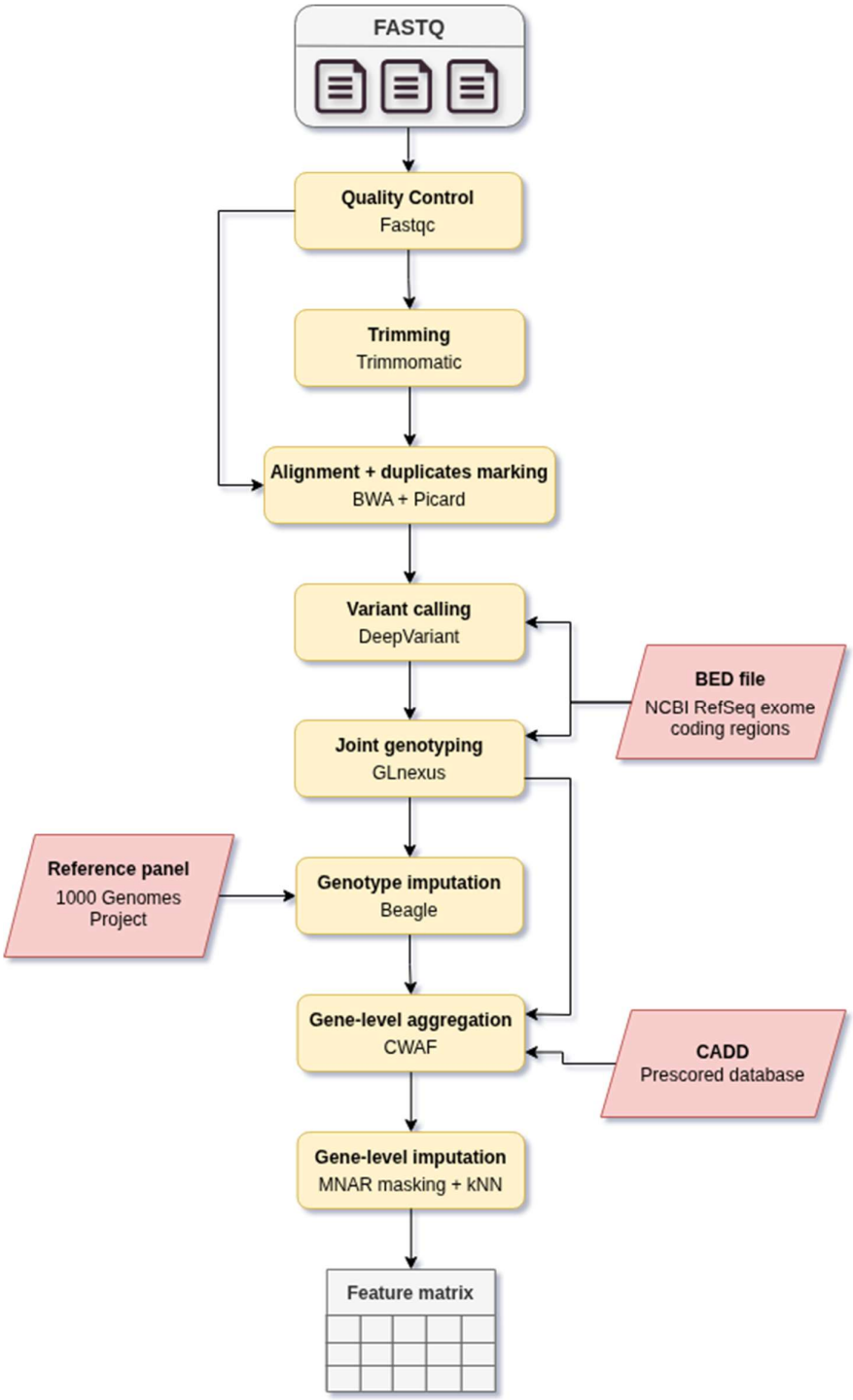

Figure 1. Overview of the proposed pipeline. QC, quality control; CWAF, CADD-weighted average allele fraction; MNAR, missing not at random.

 doi:10.20944/preprints202604.1444.v1

2.2. Reads Quality Control and Mapping

Quality check of raw sequencing reads is carried out with fastQC (v0.12.1) [28] and MultiQC is used for reporting [29]. Low-quality reads are trimmed with Trimmomatic (v0.39; SLIDINGWINDOW:4:20), removing the 3′ end when the average Phred quality within a 4-base window drops below 20. Reads shorter than 45 bp after trimming are discarded, and if detected, adapters are removed (ILLUMINACLIP) [30]. Burrow-Wheeler Aligner (BWA; v0.7.17-r1188) is used for mapping to the reference GRCh38.p14 genome (Ensembl primary assembly release 109) with *-T 0* flag to increase sensitivity [30,31]. Resulting SAM files are converted to BAM format and coordinate-sorted with Samtools (v1.19.2) [32]. Sequencing duplicates marking is performed with Picard MarkDuplicates (v3.3.0) to ensure redundant reads are ignored in downstream steps [33]. Additionally, for data available only as GRCh37-aligned BAM files, coordinates are lifted over to GRCh38 using CrossMap (v0.7.0) with an Ensembl chain file (2014-07-25) [34].

2.3. Variant Calling and Joint Genotyping

Exonic regions were derived from the UCSC hg38 refGene database (release 2020-08-17). Coordinates with ±5 bp flanking were extracted from transcript records and merged to obtain a non-redundant set of genomic intervals, excluding non-canonical contigs and patch sequences. The resulting BED file is used to restrict variant calling and joint genotyping to target regions, thereby reducing the computational effort and improving scalability.

For the mutation detection step, DeepVariant (v1.8.0, model_type=WES) is applied [35]. GVCF output is enabled by default, and the resulting per-sample files are subsequently merged and jointly genotyped with GLnexus (v1.4.1-0-g68e25e5; DeepVariantWES config), yielding a cohort-level, multi-sample VCF [36].

2.4. Genotype Imputation

During the pre-imputation processing, multiallelic sites in the multi-sample VCF are split, SNVs extracted, left-aligned, and records with at least one sample having alternative allele depth (AD) > 5 retained. The multi-sample VCF is partitioned by chromosome, with all subsequent steps up to feature imputation executed in parallel, only for computational purposes. Genotype imputation is carried out with Beagle (v5.5), which has been shown to considerably reduce the computational cost without sacrificing the precision [24]. To ensure maximal concordance with the imputation reference panel, query genotypes are conformed by allele matching and strand alignment, with unresolved records being excluded. Subsequently, haplotypes are phased across the cohort and used for genotype estimation and imputation of missing variants [37].

The reference panel was constructed from 1000 Genomes Project VCF files (30x on GRCh38), for each chromosome separately [38]. Samples (n=3,202) were restricted to individuals of European ancestry (n=632) to match the benchmark cohort and maximize imputation accuracy, as haplotype structure varies considerably across populations [39]. However, reference panel can also be constructed for other ancestral groups.

2.5. Variants Post-Processing and Filtering

As a result of genotype imputation, all variants present in the reference panel are represented in the output VCF, including those previously absent in the query cohort. Consequently, the imputed VCF contains additional records, while those unresolved during genotype conforming are absent from the final dataset. Therefore, directly observed variants present in the non-imputed VCF are annotated using ANNOVAR with multiple annotation sources (Tab. A1) and the CADD prescored database (v1.7) [40,41]. The resulting VCF files are filtered to retain variants annotated as *exonic*, *splicing*, *exonic;splicing*, or *ncRNA_exonic;splicing*. This procedure yielded a final set of variants selected for downstream analyses. The imputed VCF is subsequently restricted to this variant set, resulting in a substantial reduction of file size and computational burden in downstream processing.

 doi:10.20944/preprints202604.1444.v1

2.7. Gene-Level Aggregation

Missing CADD Phred-like scores are imputed using k-nearest neighbors (kNN; k = 5) implemented in *scikit-learn* (v1.7.2) Python (v3.12.12) library [42]. The algorithm is run using additional variant-level features: global allele frequency, variant type, MetaSVM prediction, and ClinVar significance annotations to provide potentially better imputation.

For each sample, variants are aggregated per gene with the information derived from both the original and imputed VCF files. Allele fractions (AF) are calculated primarily as the ratio of alternative allele read counts (ALT) to the total allelic depth (sum of reference (REF) and alternative allele counts) at a given position ($AF = ALT/(REF + ALT)$). In cases where AF could not be determined, estimated allele dosage (DS) from the imputed VCF is used ($AF = DS/2$). If this measure is not available as well, AF is set to zero. When germline variants are analyzed, variants having AF > 0.2 are retained in further analysis.

For each gene and sample, a CADD-weighted average allele frequency (CWAF) was computed as the weighted average of variant-level AF values with CADD Phred-like scores serving as weights (1):

$$CWAF_{g,p} = \frac{\Sigma_{i \in g} AF_{i,p} \cdot CADD_i}{\Sigma_{i \in g} CADD_i} \quad (1)$$

where *g,p* denote gene and patient, and *i* indexes variants within gene *g*. The resulting gene-level feature matrix (samples × genes) was used for downstream analyses.

2.8. Gene-Level Feature Imputation

Gene-level imputation is performed in R (v4.5.1) environment. Upon that, samples are clustered with *PARC* (v0.40; *knn=30, jac_std_global=0.15, resolution_parameter=1, random_seed=43*) implemented in Python [43].

A missing-not-at-random (MNAR) mask is defined using gene-level detection rates (DR) across clusters. DR is computed as the proportion of samples with non-zero CWAF values per gene and cluster. Gene-cluster pairs, with a low DR ($DR < a$) in a given cluster and a high DR (DR $> b$) in at least one other cluster with at least 30 members, are flagged as under-detected. These flags are propagated to the sample level, and only missing CWAF entries corresponding to flagged gene-cluster groups are marked as MNAR (Figure 2). Thresholds a and *b* are estimated by modeling DR across all the clusters as a Gaussian mixture using *dpGMM* (v0.1.9) [44].

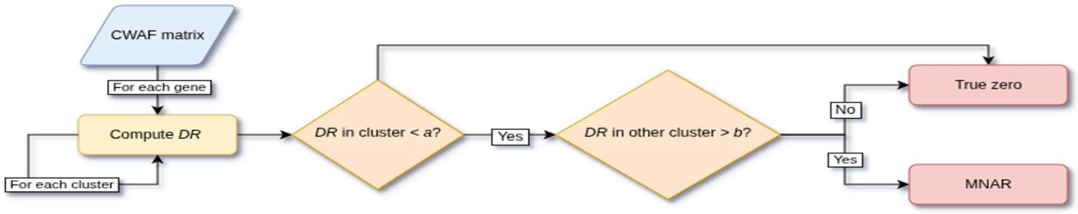

Figure 2. MNAR masking scheme. DR - detection rate.

Gene-level imputation is carried out using a masked cosine similarity-based kNN approach. For each sample and MNAR entry, samples from the same cluster are excluded from candidates from which to impute to avoid within-cluster bias. Cosine similarity is computed using non-MNAR genes shared between candidates and sample to impute, thereby preventing similarity driven by shared missingness. The minimum overlap (*min_overlap=50*) of common features is required. The nearest neighbors (*k=5*) are used to impute a similarity-weighted average. The procedure is parallelized across samples using the *future* package [45].

 doi:10.20944/preprints202604.1444.v1

2.9. Benchmark Study

2.9.1. Samples and Data

A total of 1,077 samples from 11 datasets enriched with 8 exome capture kits were analyzed (Table 1). WES data were obtained from germline breast cancer samples of Caucasian ancestry.

Table 1. Analyzed datasets with the corresponding exome capture kits and aliases. SureSelect kits are manufactured by Agilent, whereas SeqCap kits are produced by Nimblegen (now Roche).

Dataset	Samples	Exome capture kit	Kit alias
PRJNA412025	10	SureSelect Human All Exon V4	SureSelectV4
BEAUTY	106		
PRJNA516884	9	SureSelect Human All Exon V5	SureSelectV5
EGAD00001002747	205[1]		
PRJNA824495	38	SureSelect XT Clinical Research	SureSelectXTClin
PRJNA1085200	19	SureSelect Human All Exon V6	SureSelectV6
PRJNA851929	8		
EGAD50000000770	49	SureSelect XT Human All Exon V7	SureSelectXTV7
TCGA	566[2]	SeqCap EZ Exome V2	SeqCapV2
EGAD00001003137	4	SeqCap EZ Exome V3	SeqCapV3
Yale	61	IDT xGen Exome Research Panel	IDT

[1]SureSelectV5=93, SureSelectXTClin=123; [2]SeqCapV2=457, SeqCapV3=109.

2.9.2. Benchmark Analysis

The workflow described in Sections 2.2-2.8 was applied with default settings. Gene-level aggregation (2.7) was additionally performed, omitting the information from the imputed VCF, thereby enabling to benchmark genotype imputation performance.

To conduct the visual assessment of the underlying CWAF structure, uniform manifold approximation projection (UMAP) was used with the R package *uwot* (v0.2.3; *n_neighbors = 15, min_dist = 0.5, metric = cosine, n_epochs = 1500, nn_method = nndescent*). Additionally, two batch effect metrics were computed to evaluate the performance of the correction methods, with the same set of parameters in all cases and dataset serving as batch label. Local Inverse Simpson's Index (LISI) obtained with lisi package (v1.0; *perplexity=15*), quantifies batch mixing, with higher values indicating improved diversity of batches among the neighbours for a given data point [21]. K-nearest neighbors batch effect test (kBET) performs Pearson's χ^2 test to compare the distribution of batches in the local neighborhood to their global distribution, rejecting the null hypothesis when significant dissimilarities are observed. The *kBET* (v0.99.6; *k0=15, heuristic=FALSE, do.pca=FALSE*) library was used for statistical testing, with *RANN* package (v2.6.2; *k=k0+1*) utilized to compute the neighborhood matrix [46,47]. Reported in the study values reflect an average of LISI scores, or kBET null hypothesis rejection rate, derived from all the samples in the analyzed dataset.

Genes with differentiated CWAF across the clusters were evaluated with the Kruskal-Wallis test, and a false discovery rate (FDR) adjustment was applied to the obtained p-values. Top 10% significant genes identified for the data before the genotype and CWAF imputation were selected for subsequent analysis of DR.

 doi:10.20944/preprints202604.1444.v1

3. Results

3.1. Variant-Level Processing

For 8 datasets, we were able to acquire raw FASTQ files, while for three cohorts (TCGA, EGAD00001002747, and EGAD00001003137), only BAM files were available. A total of 300 FASTQ file pairs were processed during QC of raw sequencing reads, of which 164 did not pass the control due to poor read-end quality. Low-quality reads were removed (BEAUTY, PRJNA516884) and adapter trimming was performed (EGAD50000000770) for these samples. Reads stored in BAM files were mapped to the GRCh37 reference genome; thus, for 775 files, we lifted over the genomic coordinates to the GRCh38 version.

1,077 GVCF files were jointly genotyped, resulting in a single multi-sample VCF containing 483,851 records and 501,467 variants. During genotype imputation preprocessing, 10,816 multiallelic sites were split, and 13,139 variants were realigned to the reference, thereby obtaining a standardized representation of detected germline mutations. After SNV extraction and allelic depth filtering, 440,344 variants were retained, corresponding to a 87.81% reduction in variants compared to the initial dataset. During genotype conformation to the reference panel, variants were removed from the query VCF, 69,966 due to reference mismatch and 3,095 due to undeterminable strand orientation. Genotype imputation recovered approximately 2% (n=228,957) of missing genotypes across the analyzed cohort.

3.2. Gene-Level Processing

3.2.1. Variant Aggregation and Genotype Imputation Performance

Following the gene-level aggregation, a CWAF matrix of 1,077 rows representing samples and 18,346 columns corresponding to genes with at least 1 variant was obtained for both non- and genotype-imputed data. UMAP visualization revealed clearly separated groups, with TCGA samples enriched using SeqCap EZ capture kits forming the most distinct community and SureSelect/IDT samples being more similar (Figure 3a,b). Following genotype imputation, population structure became more diffuse, as reflected by an increase in LISI from 1.85 to 2.08, indicating reduced inter-group differences (Figure 3c,d).

 doi:10.20944/preprints202604.1444.v1

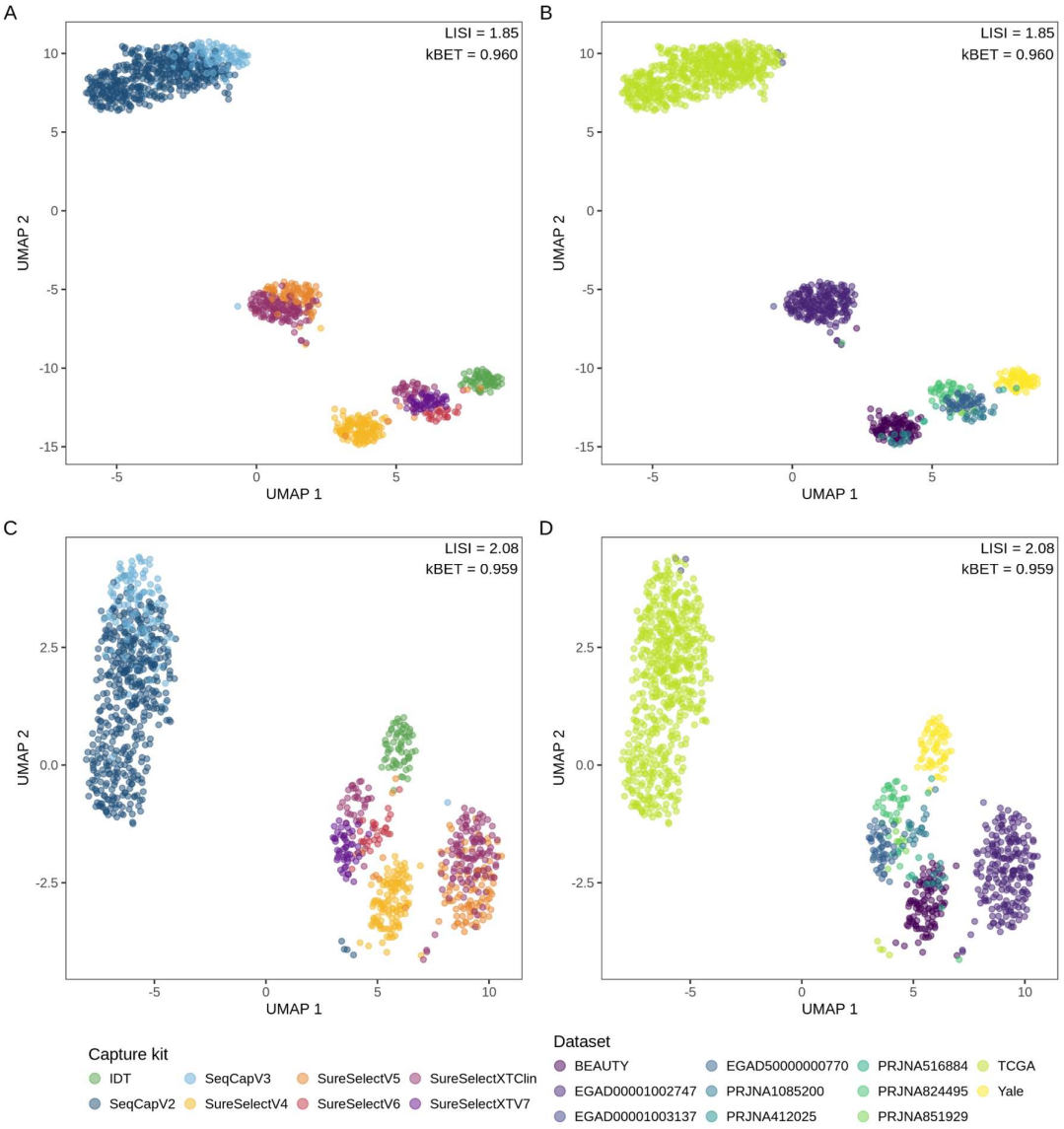

Figure 3. UMAP embeddings before (**a**,**b**) and after (**c**,**d**) genotype imputation, colored by: (**a**,**c**) Exome enrichment kit; (**b**,**d**) Dataset. Batch effects decreased as indicated by LISI improvement and kBet decrease..

3.2.2. Sample Clustering

Several capture kits showed noticeably similar CWAF profiles, forming overlapping communities in the UMAP space, such as SureSelectV6, SureSelectXTV7, and partially SureSelectXTClin or SeqCapEZV2 and SeqCapEZV3. An interesting observation was made for SureSelectXTClin, where samples from PRJNA824495 dataset exhibited evidently different locations than EGAD00001002747 samples generated with the same capture kit, reportedly (Figure 3c,d). Furthermore, most datasets were reported to use exclusively one exome enrichment kit, but individual samples were observed to group with different kits than specified. We therefore hypothesized that these samples may correspond to externally processed cases included in the cohorts, or were possibly misannotated. Following these findings, samples were clustered to ensure that gene-level imputation was performed within the technically homogenous groups sharing similar CWAF profiles rather than based on provided kit annotations. Clustering parameters for genotype-imputed data were empirically selected based on visual inspection of the assignments projected onto the UMAP, to ensure clear separation of major sample groups, while avoiding overclustering. Six

 doi:10.20944/preprints202604.1444.v1

clusters of samples sharing similar CWAF profiles were identified, comprising 374, 207, 197, 118, 115, and 66 members, respectively (Figure 5a and Table 3).

Table 3. Cluster assignments for genotype-imputed data, with regard to exome enrichment kits used. Numbers in the header reflect cluster numbers.

Capture kit	0	1	2	3	4	5
IDT						61
SeqCapV2	261	1	195			
SeqCapV3	111		2			
SureSelectV4	1			114		1
SureSelectV5		90		4	3	2
SureSelectV6					26	1
SureSelectXTV7					49	
SureSelectXTClin	1	116			37	1
Total	**374**	**207**	**197**	**118**	**115**	**66**

3.2.3. MNAR Masking and Gene-Level Imputation

For each gene, DR values across all groups were jointly modeled using a Gaussian mixture model, yielding eight thresholds at the intersections of adjacent components (Figure 4). To find optimal cutoffs for gene-level imputation, the grid search method was applied with $a \in (0.077, 0.131, 0.218)$, $b \in (0.797, 0.877, 0.933)$ and the objective to maximize LISI and minimize kBET, while keeping the fraction of imputed features reasonably low.

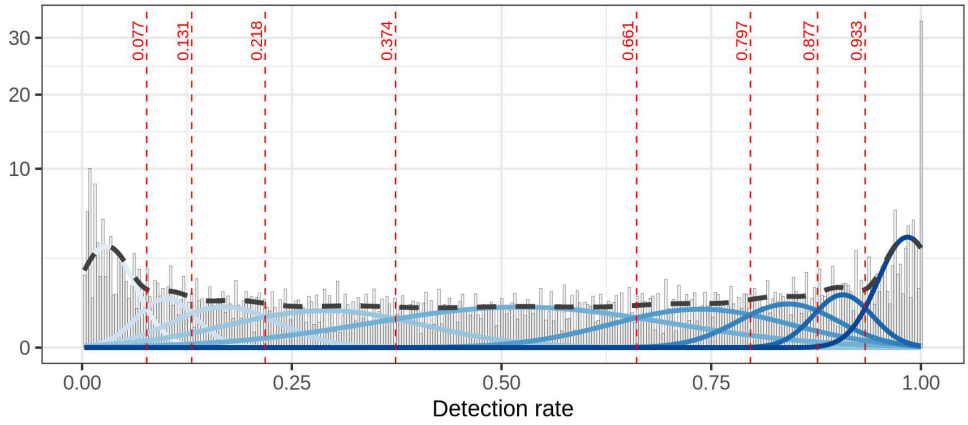

Figure 4. Gaussian mixture modeling of global gene DR. Blue curves represent individual model components, with the black dashed curve reflecting the full model. Estimated candidate thresholds for CWAF imputation were marked with a red dashed line.

The highest of candidate values for a (a=0.218), and the lowest one for b (b=0.797), was found to provide the best performance metrics, yielding a 0.35 increase in LISI ($2.08 \rightarrow 2.43$) and 0.127 decrease in kBET ($0.959 \rightarrow 0.832$) (Figure 5). Imputed CWAF values accounted for 0.47% (n=93,781) of all gene-level feature matrix entries. In the UMAP space, a considerably less pronounced separation between the clusters was observed, with samples forming a more integrated embedding (Figure 4b). Compared to raw data without the proposed imputations, the workflow resulted in an increase of 0.58 (31.35%) in LISI and a 0.128 (13.33%) decrease in the kBET rejection rate.

 doi:10.20944/preprints202604.1444.v1

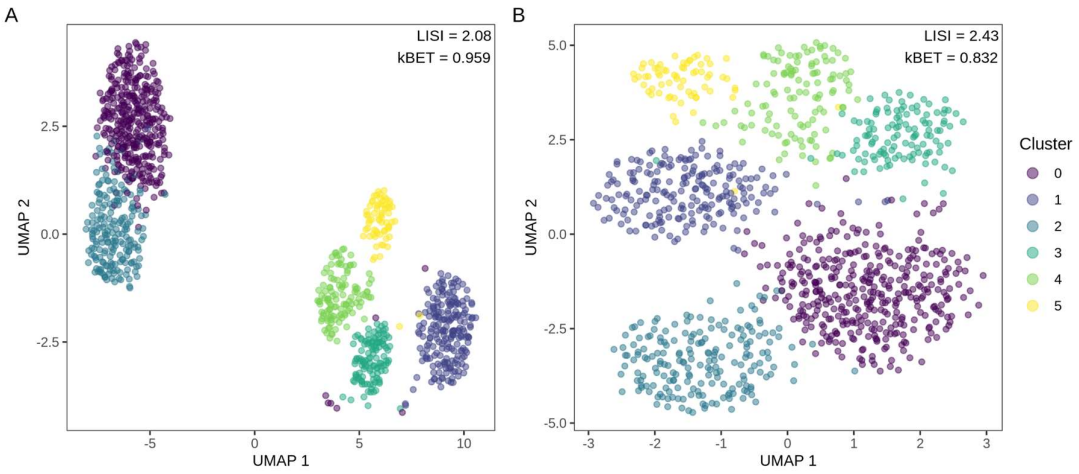

Figure 5. UMAP space colored by cluster assignments derived from CWAF values; before **(a)** and after **(b)** gene-level imputation. LISI increase (2.08→2.43) and kBET decrease (0.959→0.832) indicate successful batch effect reduction.

3.2.4. Gene Detection

Given the previously reported nearly binary detection patterns across genes distinguishing exome capture kits [12], a comparison analysis was performed on the benchmark cohort. The evaluation was conducted at each processing stage: prior to imputation, following genotype imputation, and after gene-level imputation. CWAF differences between the clusters were assessed using statistical testing. Initially, 2,875 genes showed significant differences, decreasing to 2,705 after genotype imputation, and 2,680 after CWAF imputation, yielding a total of 195 decrease. Top 10% (n=286) of initially identified, significant genes were selected for the downstream DR analysis. Four groups of detection patterns were found using k-means clustering.

We observed cluster-specific detection patterns, with the genes being detected in all the samples within the given cluster, but absent in others (Figure 6). Genotype imputation resulted in a modest DR improvement on the level of individual genes (Figure 6, row 3-4). Gene-level imputation resulted in a noticeable global increase in gene detectability under the binary detection scenario (Figure 6, row 2-3); however, as expected, it had no effect when DR differences between clusters were not sufficiently large (Figure 6, row 1).

 doi:10.20944/preprints202604.1444.v1

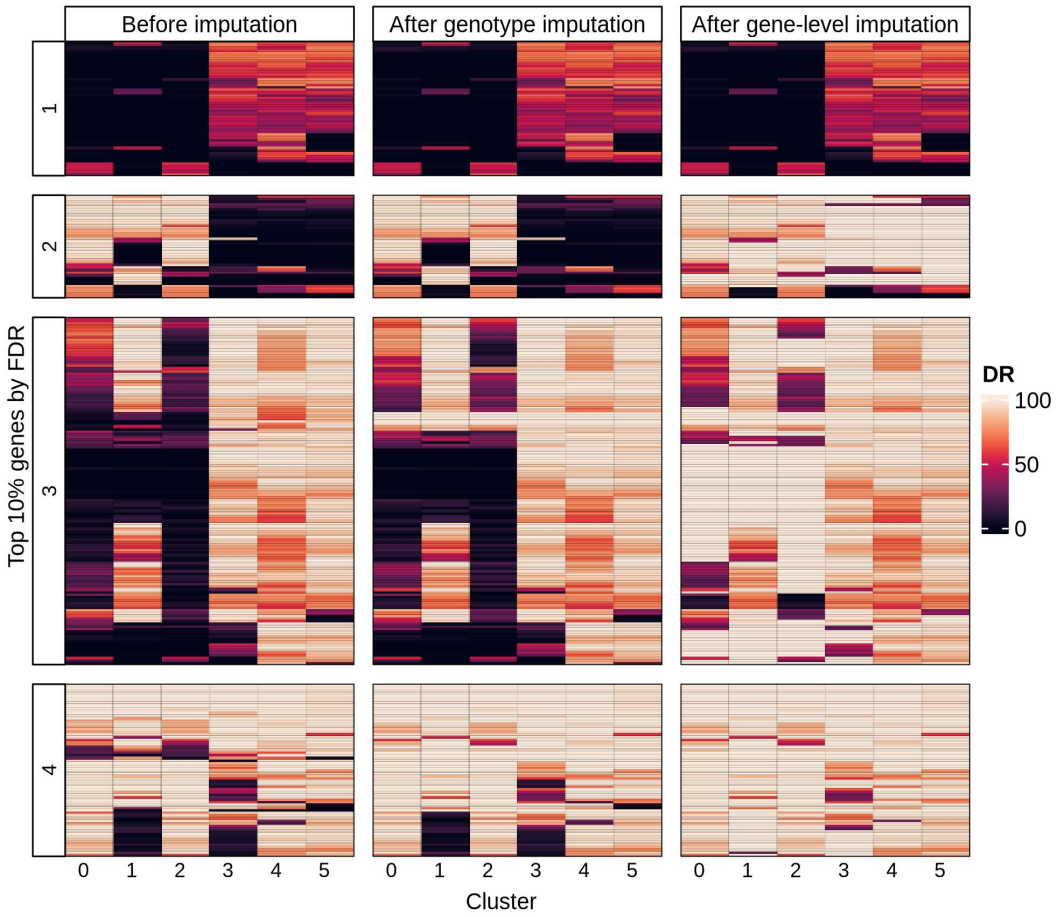

Figure 6. Comparison heatmap of gene detection rate (DR) across the sample clusters at each processing stage. 10% of the most significant CWAF differentiating genes before the genotype and gene-level imputation were selected and clustered into 4 groups corresponding to heatmap rows.

4. Discussion

Our work emphasizes the need for addressing the technical variability arising from the use of different exome enrichment kits. We reported that at the gene level, these biases manifest as near-binary gene detection patterns, where variants in a given gene are discovered in almost all samples enriched with one capture kit but none of those processed in another, highlighting the gene-level imputation importance [12]. There are dozens of exome capture kits available on the market, with each manufacturer offering multiple versions. The largest producers include Roche (formerly NimbleGen), Agilent, and Illumina, with more recent platforms provided by Twist Bioscience and Integrated DNA Technologies (IDT). In another study investigating four solution-based exome capture kits, Sulonen et al. report that regions captured with NimbleGen kits aligned more accurately with their targets, compared to Agilent methods. However, none of the considered kits captured all the Consensus Coding Sequence (CCDS) annotated exons [48]. Shigemizu et al. demonstrated that the Agilent XT method outperforms in target enrichment efficiency and sensitivity, while Illumina and NimbleGen excel in clinical application purposes [49]. In the most recent study, López et al. analyzed 10 commercial exome enrichment kits from seven manufacturers. Designs of the targets covered 89% to 100% of CCDS regions. However, empirical evaluation of efficiency revealed that at 10X coverage depth, this fraction decreased to 76-95% and further to 47-92% at 20X [50]. To improve consistency across capture kits, we restrict variant calling and joint genotyping to a common set of canonical protein-coding regions. This eliminates alternative contigs or kit-specific extensions of target exome regions, enforcing a shared observation space.

 doi:10.20944/preprints202604.1444.v1

Although the exome capture kit is a major driver of forming the UMAP communities, we observed that samples assigned to SureSelectXTClin were located as two distinct groups corresponding to their dataset of origin (PRJNA824495 and EGAD00001002747). Our further investigation revealed potential annotation inconsistencies, suggesting a different kit was used for both of them. The first version of SureSelectXTClin (2014) was based on the SureSelectV5 panel, whereas SureSelectXTClinV2 (2017) was based on the SureSelectV6, according to the manufacturer's specification [51]. In the UMAP space, samples EGAD00001002747 (deposited in 2016) overlapped with V5 samples, while PRJNA824495 (deposited in 2022) aligned with V6 samples. This strongly suggests that the two datasets correspond to different versions of the SureSelectXTClin, fully explaining the spurious capture kit subdivision. These findings highlight the risk of relying solely on derived metadata labels and emphasize the importance of data-supported clustering to characterize the technical structure more precisely.

Despite the wide range of batch correction methods developed for transcriptomics, their application to variant-derived features, such as CWAF, is limited. After normalization, gene expression levels can be treated as continuous variables, with batch effects often manifesting as mean and variance shifts, which enables their correction with statistical methods [52]. In contrast, variant call data exhibit a more complex structure, containing various data types such as coordinates, genotypes, alleles, frequencies, or likelihoods. There are also numerous classes of genomic variation, including single-nucleotide variants (SNVs), indels, multi-nucleotide variants (MNVs), structural variants, and complex rearrangements [53]. Moreover, a single genomic variation can be represented in multiple ways [54]. After the variant aggregation, CWAF values follow a sparse, zero-inflated, multimodal distribution with peaks around expected allele ratios (0, 0.5, and 1). These characteristics violate the assumptions of commonly used methods. For example, ComBat requires approximately continuous, normally distributed data, with a ComBat-seq modification assuming a negative binomial distribution. Limma method assumes batch effects to be linearly additive. Furthermore, none of these methods explicitly account for the observed non-random missingness (MNAR), where the absence of a variant often reflects a lack of detection.

Given these limitations, alternative approaches that account for the structure of variant-derived data are required, particularly in terms of batch effects driven by the use of different exome enrichment kits. In this context, the proposed workflow addresses these biases by involving the enforcement of shared observation space, joint genotyping, genotype imputation, and data-driven clustering combined with MNAR-aware gene-level imputation. The results indicate that this strategy reduces gene detection discrepancies, thereby improving the sample comparability in multi-source gene-level variant data. Although UMAP embeddings are not intended as definitive evidence, they provide a useful representation of global and local data structure. No substantial sample mixing was observed; however, clusters exhibited reduced compactness and increased dispersion, suggesting a decreased differentiating signal. These findings are consistent with batch effect metrics, which indicate improved batch integration within the local neighbourhood structure.

The DeepVariant–GLnexus workflow was selected as an alternative to traditional pipelines such as GATK, as previous benchmarking studies have reported improved robustness of DeepVariant-based variant calling and efficient scalability of GLnexus for large cohorts [55–57]. In the context of multi-source WES data, joint genotyping provides a consistent representation of variant sites across the cohort, reducing sample-specific calling variability, which is particularly important in terms of mitigating technical biases.

Several limitations of this study should be acknowledged. First, the proposed workflow is strongly dependent on the variability of the analyzed cohorts. As the number and diversity of exome capture kits increase, the observable feature space expands, and more MNAR events can be properly identified. Furthermore, as the results of the kNN application, the imputed values are inherently dependent on the composition of the dataset and can vary when the cohort structure changes, unlike reference-based methods. However, this also enables the adaptation to the underlying data structure, which is beneficial for heterogeneous datasets. Another limitation of the proposed approach is the

 doi:10.20944/preprints202604.1444.v1

clustering sensitivity to parameter selection. Overly coarse clusters may mask meaningful signals, while too granular clusters may lead to overcorrection. In addition, our workflow is primarily applicable to single-ancestry cohorts. A uniform population structure was necessary to enable a proper assessment of underlying technical variation, as CWAF is a direct transformation of AF, known to differ significantly across ancestral groups [58]. Inclusion of multiple ancestries would introduce a dominant genetic signal as previously demonstrated [12]. So, in multi-ancestry settings, the MNAR masking algorithm, which is essentially based on comparing AF-derived features, may be confounded by differences between the genetic populations rather than true detection biases. Moreover, the genotype imputation accuracy is strongly determined by the quality and size of the reference panel. Although the 1kGP panel included in the pipeline is a widely used resource, derived from high-quality sequencing data, the sample size (n=632) is limited. Larger reference panels, such as TOPMed, which include over 130,000 individuals, could provide substantially improved performance [59]. However, their integration into the proposed workflow is currently limited by the lack of readily accessible GRCh38-compatible releases to run locally.

To address these limitations, future work should focus on incorporating the ancestry covariate into the framework. During the gene-level imputation, this would require restricting the MNAR masking and kNN imputation to samples with a shared ancestral background. This, in turn, would also facilitate the inclusion of larger and more diverse datasets, thereby improving the robustness and generalizability of the proposed workflow.

Author Contributions: Conceptualization, M.M. and L.J.; methodology, M.M. and L.J.; software, L.J.; validation, M.M. and L.J.; formal analysis, L.J.; investigation, L.J, M.O, J.M., L.P. and M.M.; resources, M.O. and L.P.; data curation, L.J. and M.O; writing—original draft preparation, L.J.; writing—review and editing, M.M., M.O., and J.M.; visualization, L.J.; supervision, L.P. and M.M.; project administration, M.M.; funding acquisition, M.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research and APC were funded by the National Science Centre, Poland, grant number 2023/50/E/NZ2/00583 to LJ, MO, JM, and MM.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets analyzed in this study were obtained from multiple sources, including publicly available repositories and controlled-access databases. Publicly available data were retrieved from the Sequence Read Archive (SRA) under accession numbers PRJNA412025, PRJNA516884, PRJNA824495, PRJNA1085200, and PRJNA851929. Controlled-access data were obtained from the European Genome-phenome Archive (accessions: EGAD00001002747, EGAD50000000770, EGAD00001003137), The Cancer Genome Atlas, and the Database of Genotypes and Phenotypes (BEAUTY, accession: phs001050.v1.p1). Access to controlled datasets requires authorization from the respective data access committees. Additional data from the Yale cohort are available on request from the corresponding author due to specific data use agreements. Several samples from the Yale cohort are deposited on SRA database under accession number PRJNA1087680.

Acknowledgments: This study includes data from the database of Genotypes and Phenotypes (dbGaP) under accession number phs001050.v1.p1. We acknowledge the contributions of the study submitters: Mayo Clinic Center for Individualized Medicine, Nadia's Gift Foundation, John P. Guider, The Pharmacogenomics Research Network (U10GM 61388-15), Mayo Clinic Cancer Center (CA15083-40A2), and Mayo Clinic Breast SPORE (P50CA 116201-9; Goetz, Ingle, Kalari, Suman). This study makes use of whole exome data (EGAD00001002747) generated by Gustave Roussy, Unicancer and Institut Curie and registered jointly in the metaBreast database for scientific community. The authors acknowledge the data producers and the data access committee (EGAC00001000574 DAC) for providing access to the dataset EGAD00001003137, generated by the Translational Breast Cancer Genomic and Therapeutics Lab, Peter MacCallum Cancer Centre (Melbourne, Australia) [60]. The authors acknowledge the data producers and the data access committee (EGAC00001002391 DAC) for providing access to the dataset EGAD50000000770, generated by the Medical University of Gdansk (Gdansk, Poland).

 doi:10.20944/preprints202604.1444.v1

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

1kGP	The 1000 Genomes Project
AF	Allele fraction
CCDS	Consensus coding sequence
CWAF	CADD-weighted allele fraction
DR	Gene detection rate
FDR	False discovery rate
FFPE	Formalin-fixed paraffin-embedded
kBET	k-nearest neighbor Batch Effect Test
kNN	K-nearest neighbors
LD	Linear dichroism
LISI	Local inverse Simpson's index
MNAR	Missing-not-at-random
PARC	Phenotyping by accelerated refined community-partitioning
SNV	Single nucleotide variation
TCGA	The Cancer Genome Atlas
UMAP	Uniform manifold approximation and projection
WES	Whole-exome sequencing
WGS	Whole-genome sequencing

Appendix A

Appendix A.1

Table A1. Annotation databases used in ANNOVAR.

Database	Version
RefGene	hg38 2020-08-17
ExAC	v0.3
gnomAD exomes	v4.1
1000 Genomes	2015-08
dbSNP	build 150
ClinVar	2024-12-1
dbNSFP	v4.7a

 doi:10.20944/preprints202604.1444.v1

References

1. E. R. Mardis, "DNA sequencing technologies: 2006-2016," *Nat. Protoc.*, vol. 12, no. 2, pp. 213–218, Feb. 2017, doi: 10.1038/nprot.2016.182.
2. A. Auton *et al.*, "A global reference for human genetic variation," *Nature*, vol. 526, no. 7571, pp. 68–74, Oct. 2015, doi: 10.1038/nature15393.
3. J. N. Weinstein *et al.*, "The Cancer Genome Atlas Pan-Cancer Analysis Project," *Nat. Genet.*, vol. 45, no. 10, pp. 1113–1120, Oct. 2013, doi: 10.1038/ng.2764.
4. C. Sudlow *et al.*, "UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age," *PLoS Med.*, vol. 12, no. 3, p. e1001779, Mar. 2015, doi: 10.1371/journal.pmed.1001779.
5. A. R. Buckley *et al.*, "Pan-cancer analysis reveals technical artifacts in TCGA germline variant calls," *BMC Genomics*, vol. 18, no. 1, p. 458, Jun. 2017, doi: 10.1186/s12864-017-3770-y.
6. "Next-generation data filtering in the genomics era | Nature Reviews Genetics." Accessed: Mar. 19, 2026. [Online]. Available: https://www.nature.com/articles/s41576-024-00738-6
7. J. Majewski, J. Schwartzentruber, E. Lalonde, A. Montpetit, and N. Jabado, "What can exome sequencing do for you?," *J. Med. Genet.*, vol. 48, no. 9, pp. 580–589, Sep. 2011, doi: 10.1136/jmedgenet-2011-100223.
8. Q. Guo, E. Lakatos, I. A. Bakir, K. Curtius, T. A. Graham, and V. Mustonen, "The mutational signatures of formalin fixation on the human genome," *Nat. Commun.*, vol. 13, no. 1, p. 4487, Sep. 2022, doi: 10.1038/s41467-022-32041-5.
9. H. Krehenwinkel, M. Wolf, J. Y. Lim, A. J. Rominger, W. B. Simison, and R. G. Gillespie, "Estimating and mitigating amplification bias in qualitative and quantitative arthropod metabarcoding," *Sci. Rep.*, vol. 7, no. 1, p. 17668, Dec. 2017, doi: 10.1038/s41598-017-17333-x.
10. B. Pan *et al.*, "Assessing reproducibility of inherited variants detected with short-read whole genome sequencing," *Genome Biol.*, vol. 23, no. 1, p. 2, Jan. 2022, doi: 10.1186/s13059-021-02569-8.
11. D. P. Wickland *et al.*, "Impact of variant-level batch effects on identification of genetic risk factors in large sequencing studies," *PLoS ONE*, vol. 16, no. 4, p. e0249305, Apr. 2021, doi: 10.1371/journal.pone.0249305.
12. L. Jarosz, J. Dai, M. Ochocki, J. Merta, L. Pusztai, and M. Marczyk, "Kit-Specific and Other Non-Biological-Based Biases in Germline Whole-Exome Analysis at the Gene Level," presented at the 17th International Conference on Bioinformatics Models, Methods and Algorithms, Mar. 2026, pp. 487–496. Accessed: Mar. 27, 2026. [Online]. Available: https://www.scitepress.org/PublicationsDetail.aspx?ID=A1n7u6b9T84=&t=1
13. W. E. Johnson, C. Li, and A. Rabinovic, "Adjusting batch effects in microarray expression data using empirical Bayes methods," *Biostatistics*, vol. 8, no. 1, pp. 118–127, Jan. 2007, doi: 10.1093/biostatistics/kxj037.
14. M. E. Ritchie *et al.*, "limma powers differential expression analyses for RNA-sequencing and microarray studies," *Nucleic Acids Res.*, vol. 43, no. 7, pp. e47–e47, Jan. 2015, doi: 10.1093/nar/gkv007.
15. I. Korsunsky *et al.*, "Fast, sensitive and accurate integration of single-cell data with Harmony," *Nat. Methods*, vol. 16, no. 12, pp. 1289–1296, Dec. 2019, doi: 10.1038/s41592-019-0619-0.
16. Y. Benjamini and T. P. Speed, "Summarizing and correcting the GC content bias in high-throughput sequencing," *Nucleic Acids Res.*, vol. 40, no. 10, p. e72, May 2012, doi: 10.1093/nar/gks001.
17. M. Diossy *et al.*, "Strand Orientation Bias Detector to determine the probability of FFPE sequencing artifacts," *Brief. Bioinform.*, vol. 22, no. 6, p. bbab186, Nov. 2021, doi: 10.1093/bib/bbab186.
18. D. Heo *et al.*, "DEEPOMICS FFPE, a deep neural network model, identifies DNA sequencing artifacts from formalin fixed paraffin embedded tissue with high accuracy," *Sci. Rep.*, vol. 14, no. 1, p. 2559, Jan. 2024, doi: 10.1038/s41598-024-53167-0.
19. M. Tellaetxe-Abete, B. Calvo, and C. Lawrie, "Ideafix: a decision tree-based method for the refinement of variants in FFPE DNA sequencing data," *NAR Genomics Bioinforma.*, vol. 3, no. 4, p. lqab092, Oct. 2021, doi: 10.1093/nargab/lqab092.
20. M. Ikegami *et al.*, "MicroSEC filters sequence errors for formalin-fixed and paraffin-embedded samples," *Commun. Biol.*, vol. 4, no. 1, p. 1396, Dec. 2021, doi: 10.1038/s42003-021-02930-4.
21. R. Poplin *et al.*, "A universal SNP and small-indel variant caller using deep neural networks," *Nat. Biotechnol.*, vol. 36, no. 10, pp. 983–987, Oct. 2018, doi: 10.1038/nbt.4235.

 doi:10.20944/preprints202604.1444.v1

22. M. F. Lin *et al.*, "GLnexus: joint variant calling for large cohort sequencing," Jun. 11, 2018, *bioRxiv*. doi: 10.1101/343970.
23. N. A. O'Leary *et al.*, "Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D733–D745, Jan. 2016, doi: 10.1093/nar/gkv1189.
24. B. L. Browning, Y. Zhou, and S. R. Browning, "A One-Penny Imputed Genome from Next-Generation Reference Panels," *Am. J. Hum. Genet.*, vol. 103, no. 3, pp. 338–348, Sep. 2018, doi: 10.1016/j.ajhg.2018.07.015.
25. B. L. Browning, X. Tian, Y. Zhou, and S. R. Browning, "Fast two-stage phasing of large-scale sequence data," *Am. J. Hum. Genet.*, vol. 108, no. 10, pp. 1880–1890, Oct. 2021, doi: 10.1016/j.ajhg.2021.08.005.
26. "A global reference for human genetic variation | Nature." Accessed: Mar. 25, 2026. [Online]. Available: https://www.nature.com/articles/nature15393
27. P. Rentzsch, D. Witten, G. M. Cooper, J. Shendure, and M. Kircher, "CADD: predicting the deleteriousness of variants throughout the human genome," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D886–D894, Jan. 2019, doi: 10.1093/nar/gky1016.
28. "Andrews, S. (2010) FastQC A Quality Control Tool for High Throughput Sequence Data. - References - Scientific Research Publishing." Accessed: Mar. 25, 2026. [Online]. Available: https://www.scirp.org/reference/referencespapers?referenceid=4024153
29. P. Ewels, M. Magnusson, S. Lundin, and M. Käller, "MultiQC: summarize analysis results for multiple tools and samples in a single report," *Bioinformatics*, vol. 32, no. 19, pp. 3047–3048, Oct. 2016, doi: 10.1093/bioinformatics/btw354.
30. A. M. Bolger, M. Lohse, and B. Usadel, "Trimmomatic: a flexible trimmer for Illumina sequence data," *Bioinformatics*, vol. 30, no. 15, pp. 2114–2120, Aug. 2014, doi: 10.1093/bioinformatics/btu170.
31. "Ensembl 2022 | Nucleic Acids Research | Oxford Academic." Accessed: Mar. 25, 2026. [Online]. Available: https://academic.oup.com/nar/article/50/D1/D988/6430486?login=true
32. H. Li *et al.*, "The Sequence Alignment/Map format and SAMtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, Aug. 2009, doi: 10.1093/bioinformatics/btp352.
33. "MarkDuplicates (Picard)," GATK. Accessed: Mar. 25, 2026. [Online]. Available: https://gatk.broadinstitute.org/hc/en-us/articles/360037052812-MarkDuplicates-Picard
34. H. Zhao, Z. Sun, J. Wang, H. Huang, J.-P. Kocher, and L. Wang, "CrossMap: a versatile tool for coordinate conversion between genome assemblies," *Bioinformatics*, vol. 30, no. 7, pp. 1006–1007, Apr. 2014, doi: 10.1093/bioinformatics/btt730.
35. R. Poplin *et al.*, "A universal SNP and small-indel variant caller using deep neural networks," *Nat. Biotechnol.*, vol. 36, no. 10, pp. 983–987, Oct. 2018, doi: 10.1038/nbt.4235.
36. M. F. Lin *et al.*, "GLnexus: joint variant calling for large cohort sequencing," Jun. 11, 2018, *bioRxiv*. doi: 10.1101/343970.
37. B. L. Browning, X. Tian, Y. Zhou, and S. R. Browning, "Fast two-stage phasing of large-scale sequence data," *Am. J. Hum. Genet.*, vol. 108, no. 10, pp. 1880–1890, Oct. 2021, doi: 10.1016/j.ajhg.2021.08.005.
38. M. Byrska-Bishop *et al.*, "High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios," Feb. 07, 2021, *bioRxiv*. doi: 10.1101/2021.02.06.430068.
39. M. Leitwein, M. Duranton, Q. Rougemont, P.-A. Gagnaire, and L. Bernatchez, "Using Haplotype Information for Conservation Genomics," *Trends Ecol. Evol.*, vol. 35, no. 3, pp. 245–258, Mar. 2020, doi: 10.1016/j.tree.2019.10.012.
40. K. Wang, M. Li, and H. Hakonarson, "ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data," *Nucleic Acids Res.*, vol. 38, no. 16, p. e164, Sep. 2010, doi: 10.1093/nar/gkq603.
41. M. Schubach, T. Maass, L. Nazaretyan, S. Röner, and M. Kircher, "CADD v1.7: using protein language models, regulatory CNNs and other nucleotide-level scores to improve genome-wide variant predictions," *Nucleic Acids Res.*, vol. 52, no. D1, pp. D1143–D1154, Jan. 2024, doi: 10.1093/nar/gkad989.
42. F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, no. 85, pp. 2825–2830, 2011.

 doi:10.20944/preprints202604.1444.v1

43. S. V. Stassen, D. M. D. Siu, K. C. M. Lee, J. W. K. Ho, H. K. H. So, and K. K. Tsia, "PARC: ultrafast and accurate clustering of phenotypic data of millions of single cells," *Bioinformatics*, vol. 36, no. 9, pp. 2778–2786, May 2020, doi: 10.1093/bioinformatics/btaa042.
44. J. Zyla, K. Szumala, A. Polanski, J. Polanska, and M. Marczyk, "*dpGMM*: A new R package for efficient and robust Gaussian mixture modeling of 1D and 2D data," *J. Comput. Sci.*, vol. 95, p. 102811, Apr. 2026, doi: 10.1016/j.jocs.2026.102811.
45. H. Bengtsson, "A Unifying Framework for Parallel and Distributed Processing in R using Futures," *R J.*, vol. 13, no. 2, pp. 273–291, Nov. 2021, doi: 10.32614/RJ-2021-048.
46. M. Büttner, Z. Miao, F. A. Wolf, S. A. Teichmann, and F. J. Theis, "A test metric for assessing single-cell RNA-seq batch correction," *Nat. Methods*, vol. 16, no. 1, pp. 43–49, Jan. 2019, doi: 10.1038/s41592-018-0254-1.
47. G. Jefferis, S. Kemp, S. Arya, and D. Mount, *RANN: Fast Nearest Neighbour Search (Wraps ANN Library) Using L2 Metric*. R. [Online]. Available: https://github.com/jefferislab/rann
48. A.-M. Sulonen *et al.*, "Comparison of solution-based exome capture methods for next generation sequencing," *Genome Biol.*, vol. 12, no. 9, p. R94, Sep. 2011, doi: 10.1186/gb-2011-12-9-r94.
49. "Performance comparison of four commercial human whole-exome capture platforms | Scientific Reports." Accessed: Mar. 20, 2026. [Online]. Available: https://www.nature.com/articles/srep12742
50. F. V. López, J. J. Ashton, G. Cheng, and S. Ennis, "A systematic analysis of contemporary whole exome sequencing capture kits to optimise high-coverage capture of CCDS regions," *NAR Genomics Bioinforma.*, vol. 7, no. 3, p. lqaf115, Sep. 2025, doi: 10.1093/nargab/lqaf115.
51. "Agilent Technologies | Agilent." Accessed: Apr. 10, 2026. [Online]. Available: https://www.agilent.com/en/
52. Y. Yu, Y. Mai, Y. Zheng, and L. Shi, "Assessing and mitigating batch effects in large-scale omics studies," *Genome Biol.*, vol. 25, no. 1, p. 254, Oct. 2024, doi: 10.1186/s13059-024-03401-9.
53. E. Garrison, Z. N. Kronenberg, E. T. Dawson, B. S. Pedersen, and P. Prins, "A spectrum of free software tools for processing the VCF variant call format: vcflib, bio-vcf, cyvcf2, hts-nim and slivar," *PLOS Comput. Biol.*, vol. 18, no. 5, p. e1009123, May 2022, doi: 10.1371/journal.pcbi.1009123.
54. A. Tan, G. R. Abecasis, and H. M. Kang, "Unified representation of genetic variants," *Bioinformatics*, vol. 31, no. 13, pp. 2202–2204, Jul. 2015, doi: 10.1093/bioinformatics/btv112.
55. T. Yun, H. Li, P.-C. Chang, M. F. Lin, A. Carroll, and C. Y. McLean, "Accurate, scalable cohort variant calls using DeepVariant and GLnexus," *Bioinformatics*, vol. 36, no. 24, pp. 5582–5589, Apr. 2021, doi: 10.1093/bioinformatics/btaa1081.
56. Y. A. Barbitoff, R. Abasov, V. E. Tvorogova, A. S. Glotov, and A. V. Predeus, "Systematic benchmark of state-of-the-art variant calling pipelines identifies major factors affecting accuracy of coding sequence variant discovery," *BMC Genomics*, vol. 23, no. 1, p. 155, Feb. 2022, doi: 10.1186/s12864-022-08365-3.
57. V. Pinto, L. Sousa, and C. Silva, "Variant calling in genomics: A comparative performance analysis and decision guide," *PLOS One*, vol. 21, no. 2, p. e0339891, Feb. 2026, doi: 10.1371/journal.pone.0339891.
58. G. T. Marth, E. Czabarka, J. Murvai, and S. T. Sherry, "The Allele Frequency Spectrum in Genome-Wide Human Variation Data Reveals Signals of Differential Demographic History in Three Large World Populations," *Genetics*, vol. 166, no. 1, pp. 351–372, Jan. 2004, doi: 10.1534/genetics.166.1.351.
59. D. Taliun *et al.*, "Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program," *Nature*, vol. 590, no. 7845, pp. 290–299, Feb. 2021, doi: 10.1038/s41586-021-03205-y.
60. P. Savas *et al.*, "The Subclonal Architecture of Metastatic Breast Cancer: Results from a Prospective Community-Based Rapid Autopsy Program 'CASCADE,'" *PLoS Med.*, vol. 13, no. 12, p. e1002204, Dec. 2016, doi: 10.1371/journal.pmed.1002204.