# Review paper notes

**Review Topic:** Genomic variant benchmark: if you cannot measure it, you cannot improve it

\*\*Must know before reading the documentation

What is genome sequencing and what is it’s relationship with gene association?

In the future of [personalized medicine](https://en.wikipedia.org/wiki/Personalized_medicine), whole genome sequence data may be an important tool to guide therapeutic intervention. The tool of [gene sequencing](https://en.wikipedia.org/wiki/DNA_sequencing) at [SNP](https://en.wikipedia.org/wiki/Single-nucleotide_polymorphism) level is also used to pinpoint functional variants from [association studies](https://en.wikipedia.org/wiki/Association_studies) and improve the knowledge available to researchers interested in [evolutionary biology](https://en.wikipedia.org/wiki/Evolutionary_biology), and hence may lay the foundation for predicting disease susceptibility and drug response.

**Focus of this review:**

In this review, we discuss the available benchmark datasets and their utility. Additionally, we focus on the most recent benchmark of genes with medical relevance and challenging genomic complexity.

Benchmark datasets containing well-established variant calls are needed to develop, optimize, and analytically validate variant detection methods, which can ultimately be translated into systematic research studies and clinical practices. Such benchmark datasets are useful for a diverse range of researchers involved in sequencer manufacturing, library preparation, bioinformatics method development, and clinical studies.

**Key points of the review:**

A genome variant benchmark dataset serves as a vital resource for evaluating the effectiveness and accuracy of newly developed variant callers. To ensure a comprehensive assessment, it is crucial for the benchmark dataset to include relevant information about the specific genomic regions associated with these variants. This distinction is essential as it helps differentiate these regions from those that were not considered in the benchmark dataset, such as regions that are not assembled or lack support from multiple callers.

To ensure a comprehensive assessment, it is crucial for the benchmark dataset to include relevant information about the specific genomic regions associated with these variants. This distinction is essential as it helps differentiate these regions from those that were not considered in the benchmark dataset, such as regions that are not assembled or lack support from multiple callers.

Although attaining 100% accuracy and sensitivity is challenging. To achieve this, creating a benchmark dataset involves utilizing diverse sequencing technologies such as long-reads, short-reads, and linked-reads, each with varying insert sizes and high coverage.

**Few important terms to understand this review:**

* Genetic Variants: Bivinno Gene er moddhe sequencing er difference. Nucleotide level eo ei difference thakte pare abar aro boro level e giyeo ei difference dekha jete pare.

Some common types of genetic variants:

**Single Nucleotide Polymorphisms (SNPs):** SNPs are the most common type of genetic variant and involve the substitution of one nucleotide (A, T, C, or G) for another at a specific position in the DNA sequence.

**Insertions and Deletions (Indels):** Indels are genetic variants where one or more nucleotides are inserted into or deleted from the DNA sequence.

**Copy Number Variations (CNVs):** CNVs are variations in the number of copies of a particular DNA segment, which can range from kilobases to megabases in size.

* Variant callers: Variant callers are software tools or algorithms used in genomics to identify and characterize genetic variants from DNA sequencing data.

Sequencing techniques: Linked read, long read, short read

**Methods introduced in this review:**

**Benchmark database sites:**

Genome in a Bottle Project (GIAB): publicly available dataset site

* GIAB is a well-known project that aims to create high-confidence benchmark datasets for genomics research. It involves extensively characterizing specific human genomes to provide a comprehensive reference for sequencing and variant calling.
* GIAB datasets include data generated from various technologies, such as Illumina short-read sequencing, Pacific Biosciences (PacBio) long-read sequencing, and other technologies.
* GIAB provides not only aligned sequencing data (BAM files) but also information about validated genetic variants in these genomes. These validated variants serve as a "gold standard" for evaluating the accuracy of variant calling algorithms.

**Platinum Genome:**

* Platinum Genome is another initiative focused on creating benchmark datasets for genomics research. It involves deep sequencing and extensive characterization of specific genomes.
* Platinum Genome datasets are derived from the genomes of individuals whose DNA has been sequenced to high coverage with Illumina sequencing technology.

Some information of benchmark datasets:

**Benchmarking process:**

In this review the process described for genomic variant benchmark is ,

First, gene sequencing is done in the lab using real DNA sample.

After sequencing, essential analysis steps including read alignment or genome assembly should be performed together with variant calling [22–24] to obtain a result that can then be compared against the benchmark dataset.

This is done over specific SNVs or SVs benchmark methods that compare the results with that of the benchmark and provide a clear metric on the calls (precision and recall/sensitivity) of the analysis. This provides a comprehensive insight if the sequencing technology and the computational pipeline being tested are reliable.

Depending on the sequencing method of choice, some parameters need to be adjusted during the evaluation of the data, for example, the minimum size of the variants for optical mapping or the constraints of sequenced resolved representation. Over the past years, this process not only helped establish pipelines and technologies, but also drove the development of new sequencing approaches [15].

Data curation:

**Short-read sequencing** (e.g., Illumina) is still one of the most accurate technologies and is well established to identify SNVs and especially indels [25, 26]. However, short-read methods often struggle to characterize repetitive regions of the genome and have become known to also have limitations for SV detection.

multiple different benchmark datasets creation

1. Using simulated data pipelines

Problem: suffer from a simplistic representation of real genomics data.

1. Synthetic benchmark datasets: overcome this simplicity issue partially but often still suffer from a less reliable representation of the underlying challenges found in real datasets.
2. recently, benchmark datasets created with fully characterized “real” data.

The benefits from using real datasets are highly dependent on how accurately it was characterized. This characterization is the most challenging part of creating such a benchmark dataset, but it best represents the challenges that scientists face when analysing their genomic samples.

**Considerate things when benchmarking: (for this docs)**

**Overview of genomic benchmarks:**

In the construction of a benchmark dataset, several other criteria are carefully considered. For instance, regions where callers exhibit systematic errors or misrepresent genotypes are typically excluded [11]. Similarly, areas demonstrating high complexity may also be excluded to ensure a more accurate representation of genomic variations [14]. **These selection criteria help** to enhance the quality and reliability of the benchmark dataset, providing researchers with a robust and comprehensive resource for **evaluating variant calling algorithms**.

Choosing sequencing technology

When it comes to choosing the sequencing approach, any technology can be used, from short reads (e.g., Illumina) to long read (e.g., ONT: Oxford Nanopore Technologies or PacBio: Pacifc Biosciences), or even optical mapping (Bionano) or other technologies [20, 21].

After sequencing, essential analysis steps including read alignment or genome assembly should be performed together with variant calling [22–24] to obtain a result that can then be compared against the benchmark dataset.

This provides a comprehensive insight if the sequencing technology and the computational pipeline being tested are reliable.

Over the past years, this process not only helped establish pipelines and technologies, but also drove the development of new sequencing approaches [15].

The curation of such benchmark datasets is often a tedious process and typically involves multiple sequencing technologies to leverage their strengths, overcome each other’s limitations, and avoid any systematic errors from a single platform.

There have been different attempts to enhance or produce benchmark datasets including the Genome in a Bottle Consortium (GIAB) using alignment, and de novo assembly [13, 19] and Platinum Genome [30], which are limited to specific regions.

They used two technologies in the latest version of GIAB (v.4.2.1) for sequencing seven samples (HG001, HG002, HG003, HG004, HG005, HG006, and HG007) with linked [32] and long-read [33] to characterize segmental duplication and hard-to-align regions which were traditionally often avoided. Using the previously mentioned technologies, the authors were able to add 16% more exonic regions, many of which are medically important [14].

Additionally, they established a benchmark across the Major Histocompatibility Complex (MHC) region [34].

**Benchmarks for challenging medically relevant genes:**

Mandelker et al. introduced one of the first lists of medically relevant genes that were difficult or impossible to analyze via the sequencing technology of their time [29]. The challenge was that short reads could not resolve these genes due to their complexity (e.g., repetitive regions). Around 88% of these genes were then included in the GIAB benchmark database version 4.2.1 thanks to the use of linked- and long-read technologies [14]. Using HiFi long-read [32], Wagner et al. assembled a list of 5175 genes from COSMIC, OMIM, HGMD, and ClinVar databases, adding to those genes commonly tested in inherited diseases [16] (Fig.  3B).

The Challenging Medically Relevant Genes (CMRGs) is a benchmark dataset created specifically for the purpose of testing new tools [16]. It is a list of medically relevant genes of high complexity. For a gene to be included, the presence of its whole sequence plus 20 kb fanking regions on both sides is required in a single assembled contig. Only 273 out of 395 genes described in the CMRG study by Wagner et al. were fully resolved and included in the CMRG benchmark dataset [16]. It is worth mentioning that at least 15% of the gene body of 99% of the 273 genes are either challenging to sequence or contain challenging variants to detect (due to the low mappability of these regions and the presence of repeats). The other 122 genes of the list of 395 genes are absent from the benchmark set for various reasons. One reason is the shortcomings of the reference genomes, which include gaps in the reference sequence, being resolved only on one of the references but not the other (as happened for the KCNJ18 gene), or duplications in HG002 compared to GRCh38 [16].

**Towards clinical usability and implementation of variation benchmarks:**

In contrast to previous genome-wide efforts, the CMRGs benchmark dataset (Fig. 4) is not focused on resolving the largest portion of the genome, but challenging genes with medical importance that were not fully resolved (Fig.  3A) in previous benchmark efforts from GIAB [16].

Each gene in the list of CMRGs has been studied in-depth in the literature and is related to one or **multiple diseases.** This is also exemplified by the number of ClinVar (i.e., a database to include variations and their impact on diseases) variants that are overlapping these genes (Fig. 4C). Out of 386 total resolved genes found in GRCh38 alone, 208 of them are correlated with neuronal diseases, such as KBG syndrome (associated with the genes ANKRD11 and CDH15 from CMRGs), neurofibromatosis (APOBEC1), filamentary keratitis (FLG), and spinal muscular atrophy (SMN1), based on the GeneCards database [44] (Fig. 4A). Additionally, 117 of the genes in the CMRGs list are correlated with blood, eye, and immune diseases. Genes related to respiratory, skeletal, nephrological, and skin diseases are also included in this list.

Cardiovascular **disease** is another category also covered in the CMRG gene list. That includes atypical coarctation of aorta (RNF213), right bundle branch block (TRPM4), and pseudo-von Willebrand disease (GP1BA), as described in the GeneCards database. The third category that we mention here is immune diseases. Autoimmune lymphoproliferative syndrome (CASP10, CD4), neutropenia (G6PC3, ANKRD11, TYMS), anemia, autoimmune hemolytic (CD4, RHCE), and chronic granulomatous disease (NCF1) were found related to this category in the literature. Importantly, NCF1 is known to be associated with 20% of cases of chronic granulomatous disease. Finally, several genes in the CMRG list had been surveyed in cancer-related studies, including colon adenocarcinoma (BAX), colorectal cancer (BAX, BRAF), hepatocellular carcinoma (AXIN1, TERT), adenocarcinoma (BRAF, H19), and prostate cancer (BAX, PTEN). Phosphatase and tensin homolog (PTEN), a tumor suppressor gene, is mutated in many cancers [49]. Specifically, it is commonly inactivated or lost in breast and ovarian cancers [50]. This gene may be a target for tandem duplications generating out-of-frame exon duplications [51].

Importantly, several hundreds of ClinVar variants intersect with the CMRGs list, as depicted in Fig. 4C. All in all, the CMRGs list with their genetic variations is a precious resource paving the way for answering a diverse range of clinical research questions.

**Emerging sequencing technologies:**

There is a cyclic reliance between the development of benchmark datasets and the emergence of new sequencing technologies and bioinformatic methods to detect variants. In turn, once these new technologies are established, they are often integrated to form novel benchmarks. In this section, we describe novel approaches and technologies in DNA sequencing.

While short-read technologies such as Illumina’s exome sequencing are widely used for studying many genetic diseases because of their low cost and high accuracy, they still have their limitations [1, 52, 53]. Other sequencing technologies such as long-read sequencing technology have the potential to delineate a large number of SVs that could be contributing to some diseases and were undetected with short-read sequencing [54]. Long-read sequencing, regardless of whether it is PacBio or ONT, has played a pivotal role in addressing key challenges in genomics. It has been instrumental in filling gaps within the human genome, allowing us to obtain a comprehensive understanding of its structure for the first time in history. The application of long-read sequencing technology has had far-reaching implications, particularly in the field of medicine. It has played a crucial role in resolving numerous medical cases by facilitating the identification of **disease-causing genetic variations**, allowing for accurate diagnosis, such as solving CGG-repeat expansion in the fragile X gene [17, 48, 57–62].

Here, we describe five novel sequencing technologies: AVITI, Illumina Complete Long-Read technology, TeLL-seq, SBB, and Ultima (Table 2).

**Impact of reference genome:**

One major challenge for creating genomic benchmark datasets is that they are often dependent on the quality of the underlying reference genome. The reference genome is a fundamental resource for biomedical research, human genetics, and clinical studies.

Emerging improved reference genomes offer a great opportunity to have more accurate benchmark datasets of genetic variations.

**Challenges of benchmark curation: overview of tools:**

Creating a benchmark dataset is a challenging task. It requires high-quality DNA samples, on which multi-platform sequencing machines should be run. In a typical pipeline for developing benchmark datasets, the first step is to map the DNA sequencing reads to the reference genome.

Tere are two different techniques for benchmarking: the alignment and the genome assembly approaches.

**Discussion:**

In this review, we highlighted the emergence and development of genomics benchmark datasets, while describing the ones currently available in both healthy and patient samples. Furthermore, we briefly reviewed new technologies that might lead to further improvements of the currently available benchmark datasets. This review brings a special focus on the recently released challenging medically relevant genes (CMRGs) benchmark study, which encompasses 386 genes that are challenging due to the complexity of their genomic location. Benchmark tools give clearer evidence whether variants identified by the tool being tested are present or not in the benchmark dataset, and on the correctness of the variant representation.

Despite these direct challenges, there is a huge gain from keeping benchmark datasets up to date and continuing to challenge the genomics and bioinformatics field. The most obvious one is creating a dataset that combines SNVs, indels, and SVs. While this was first done at the CMRGs benchmark, we are still currently lacking methods to simultaneously identify SNVs and SVs, and thus also benchmark tools that can do the same. We reported only one benchmark dataset describing variants identified in a somatic cell lineage as a benchmark set.

Over this review, we focused only on genome-based variant benchmark datasets, but this is of course only the tip of the iceberg. Furthermore, a key point is that this would also allow the development of samples carrying pathogenic variants, which are potentially key to pushing the medical genomics field forward. The selection of the individual or disease is of course challenging, and as discussed before, a tumor/normal sample would be probably a good start in this direction. Another aspect is of course also the development of genomic. benchmarks across key non-human model species such as plants, animals, and fungi. These could be key for genomics and clinical research, but also for economic reasons.

FAQs:

1. **Why is data curation needed?**

Ans: The curation of such benchmark datasets is often a tedious process and typically involves multiple sequencing technologies to leverage their strengths, overcome each other’s limitations, and avoid any systematic errors from a single platform.

Depending on the sequencing method of choice, some parameters need to be adjusted during the evaluation of the data, for example, the minimum size of the variants for optical mapping or the constraints of sequenced resolved representation. Over the past years, this process not only helped establish pipelines and technologies, but also drove the development of new sequencing approaches [15].

1. **Why genomic benchmark datasets are essential?**

Genomic benchmark datasets are essential to driving the field of genomics and bio‑ informatics. They provide a snapshot of the performances of sequencing technologies and analytical methods and highlight future challenges. However, they depend on sequencing technology, reference genome, and available benchmarking methods. benchmark datasets that play a pivotal role in evaluating RNA-sequencing tools.

Summary: This review paper highlights about why benchmark datasets are needed, also about gene sequencing and different genomic benchmarked datasets.

The paper highlight about some gene sequencing technologies and tools, and which one is more promising when benchmarking a dataset e.g. GIAB datasets was established using alignment, and de novo assembly sequencing process and then enhanced by using linked-read sequencing in addition to short-read and enlarging the benchmarked regions.