The Journal of Molecular Diagnostics, Vol. 25, No. 1, January 2023

ELSEVIER

the Journal of Molecular Diagnostics
jmdjournal.org

SPECIAL ARTICLE

Recommendations for the Use of *in Silico* Approaches for Next-Generation Sequencing Bioinformatic Pipeline Validation

A Joint Report of the Association for Molecular Pathology, Association for Pathology Informatics, and College of American Pathologists

Eric J. Duncavage,[*†] Joshua F. Coleman,[*‡] Monica E. de Baca,[*§] Sabah Kadri,[*¶] Annette Leon,[*‖] Mark Routbort,[*,**] Somak Roy,[*††] Carlos J. Suarez,[*‡‡] Chad Vanderbilt,[*§§] and Justin M. Zook[*¶¶]

From the In Silico *Pipeline Validation Working Group of the Clinical Practice Committee,* Association for Molecular Pathology, Rockville, Maryland; the Department of Pathology and Immunology,[†] Washington University School of Medicine, St. Louis, Missouri; the Department of Pathology,[‡] University of Utah, Salt Lake City, Utah; Pacific Pathology Partners,[§] Seattle, Washington; the Department of Pathology,[¶] Anne and Robert H Lurie Children's Hospital of Chicago, Chicago, Illinois; Color Health,[‖] Burlingame, California; the Department of Hematopathology,** MD Anderson Cancer Center, Houston, Texas; the Department of Pathology and Laboratory Medicine,[††] Cincinnati Children's Hospital, Cincinnati, Ohio; the Department of Pathology,[‡‡] Stanford University, Palo Alto, California; the Department of Pathology,[§§] Memorial Sloan Kettering Cancer Center, New York, New York; and the Biomarker and Genomic Sciences Group,[¶¶] National Institute of Standards and Technology, Gaithersburg, Maryland*

Accepted for publication September 28, 2022.

Address correspondence to Eric J. Duncavage, M.D., Department of Pathology and Immunology, Washington University in St. Louis, 660 S. Euclid Ave., St. Louis, MO 63110. E-mail: eduncavage@wustl.edu.

In silico approaches for next-generation sequencing (NGS) data modeling have utility in the clinical laboratory as a tool for clinical assay validation. *In silico* NGS data can take a variety of forms, including pure simulated data or manipulated data files in which variants are inserted into existing data files. *In silico* data enable simulation of a range of variants that may be difficult to obtain from a single physical sample. Such data allow laboratories to more accurately test the performance of clinical bioinformatics pipelines without sequencing additional cases. For example, clinical laboratories may use *in silico* data to simulate low variant allele fraction variants to test the analytical sensitivity of variant calling software or simulate a range of insertion/deletion sizes to determine the performance of insertion/deletion calling software. In this article, the Working Group reviews the different types of *in silico* data with their strengths and limitations, methods to generate *in silico* data, and how data can be used in the clinical molecular diagnostic laboratory. Survey data indicate how *in silico* NGS data are currently being used. Finally, potential applications for which *in silico* data may become useful in the future are presented. *(J Mol Diagn 2023, 25: 3—16; https://doi.org/10.1016/j.jmoldx.2022.09.007)*

Support for this project was exclusively provided by the Association for Molecular Pathology.

Disclosures: To provide active management of potential perceived and/or actual conflicts of interest (COIs), a Working Group cochair without relevant conflicts was appointed, and COI disclosures were requested from and/or provided by all authors throughout all phases of the consensus manuscript development process. E.J.D. is employed by the Washington University School of Medicine and is a cofounder of P&V Licensing LLC, which manufactures *in silico* proficiency testing material. J.F.C. is employed by the University of Utah. M.E.d.B. is employed by Pacific Pathology Partners and is a member of the Board of Governors, Informatics and Diversity, Equity and Inclusion Committees and the Council on Education and is Co-Chair of the Artificial Intelligence Committee at College of American Pathologists (CAP). S.K. is employed by and has disclosed stock options at AbbVie and was previously employed by the Anne and Robert H. Lurie Children's Hospital of Chicago. A.L. was previously employed by

https://doi.org/10.1016/j.jmoldx.2022.09.007

Next-generation sequencing (NGS)–based molecular diagnostics have rapidly proliferated for both molecular somatic and germline applications, necessitating standards and guidelines for their commonplace use in the clinical laboratory.[1–3] A major advantage of broad NGS-based panels is that NGS can identify variants in a large number of genes; however, the analytical validation of NGS panels can be challenging, as it is difficult to obtain physical samples with the vast number of variants capable of being detected by the assay. Instead, most laboratories will sequence a representative number of samples or cell lines with known variants and rely on sequencing metrics (ie, the fraction of targets with sufficient coverage) to infer performance across most sequenced positions as part of analytical assay validation.[4]

Several approaches have been developed to supplement real physical samples for analytical validation, including spiking in synthetic DNA and *in silico* data. Synthetic DNA with common or challenging variants of clinical interest can be added to reference samples (eg, from Genome in a Bottle[5]) to help validate the method and bioinformatics parts of the assay, but these are limited to short-read sequencing and certain classes of variants.[5–9] *In silico* NGS validation testing can take many forms and has been adopted by many clinical laboratories, and commercial proficiency testing programs are now available (eg, College of American Pathologists).[10] Although many previous guidelines do not discuss *in silico* data, two previous guidelines for validation of oncology NGS panels and bioinformatics pipelines recommended using *in silico* data during the optimization and familiarization process and envisioned increasing use of *in silico* data to augment real samples for some pathogenic mutations.[3,4] In this article, the Working Group focuses on what constitutes *in silico* NGS testing, how *in silico* data files can be generated, how *in silico* testing can be used in clinical NGS assay validation, and the limitations of *in silico* testing compared with physical samples.

In silico NGS data may be broadly defined as any data that have been artificially manipulated or generated. For example, *in silico* data may be generated *de novo* by simulating reads from reference sequence data (purely simulated data) (Figure 1A and Supplemental Table S1).[11–40] More commonly, in clinical laboratories, *in silico* data have been generated by manipulating existing NGS data files (Table 1).[9,41–43] For example, two data files from physical samples may be mixed at various ratios to simulate variants with different variant allele frequencies to evaluate the bioinformatic pipeline performance across a greater number of variant allele frequencies than may be obtained with physical samples (mixing sample data) (Figure 1B).[44] Data from a single physical sample may also be downsampled to simulate the effects of lower coverage depths on variant calling (downsampling FASTQ files) (Figure 1C).[44,45] *In silico* data may also be generated by manipulating read-level data within data files generated from physical samples (manipulated assay data) (Figure 1D and Table 1). For example, variants such as single-nucleotide variants (SNVs), insertions/deletions (indels), or even structural variants may be inserted into laboratory data files (BAM or FASTQ) to assess the ability of the bioinformatics pipeline to correctly identify and annotate the variant.[9,43,46] This approach may be especially useful to determine the performance characteristics of a bioinformatics pipeline for the detection of variants for which there are few physical samples available, such as in the case of rare disorders or rare, clinically relevant somatic alterations. *In silico* approaches may also assess the bioinformatic pipeline's ability to detect challenging variants for which it is difficult to find physical samples, such as medium-sized indels between 15 and 1000 bp[47] or dinucleotide substitutions. *In silico* NGS data have primarily been used for SNVs and indels, but there are opportunities for expanding to microsatellite instability (MSI), tumor mutation burden (TMB),[48] structural variants,[9,30–32,37,43] copy number variations,[41–43,49] RNA sequencing[50] [see *RNA Sequencing* below for references for FUSIM, Polyester (*https://bioconductor.org/packages/release/bioc/html/polyester.html*, last accessed September 1, 2022), rlsim (*https://github.com/sbotond/rlsim*, last accessed September 1, 2022), RNASeqReadSimulator (*https://github.com/davidliwei/RNASeqReadSimulator*, last accessed September 1, 2022), SimCT (*https://github.com/jaudoux/simct*, last accessed September 1, 2022), and bisulfite sequencing[13] (*https://github.com/BeyondTheSky/BSSim*, last accessed September 1, 2022)], and metagenomics.[18] Finally, *in silico* approaches may be used when changes are made only to the bioinformatics pipeline and not to the wet laboratory components of the assay; in this scenario, data previously generated by the laboratory are reanalyzed with the new version of the bioinformatics pipeline (data reanalysis) (Figure 1E).

Color Health and is currently employed by Sema4. M.R. is employed by MD Anderson Cancer Center. S.R. is employed by the Cincinnati Children's Hospital, is the founder and owner of StarBioInfo LLC, which provides paid consultation services in the domain of genomics, bioinformatics, and clinical informatics, and is a member of the CAP Genomic Medicine Resource Committee (GMRC) and the Clinical Laboratory Standards Institute (CLSI) MM25 Bioinformatics Working Group. C.J.S. is employed by the Stanford University School of Medicine and is a member of the CAP GMRC. C.V. is employed by the Memorial Sloan Kettering Cancer Center and is a consultant for and owns stock options at Paige AI. J.M.Z. is employed by the National Institute of Standards and Technology.

Standard of practice is not defined by this article, and there may be alternatives. See *Disclaimers* for further details.

The *In Silico* Pipeline Validation Working Group of the Clinical Practice Committee, Association for Molecular Pathology (AMP), was cochaired by E.J.D. and J.M.Z. with organizational representation from the College of American Pathologists (S.R.) and the Association for Pathology Informatics (M.E.d.B.). The AMP 2021 Clinical Practice Committee consisted of Jane Gibson (Chair), Fatimah Nahhas, Steven Sperber, Rashmi Goswami, Michael Kluk, Susan Hsiao, David Eberhard, Joseph Yao, Blake Buchan, Joshua Coleman, Elaine Gee, Andrés Madrigal, and Jack Tung.

Current address of A.L., Sema4, Stamford, CT.

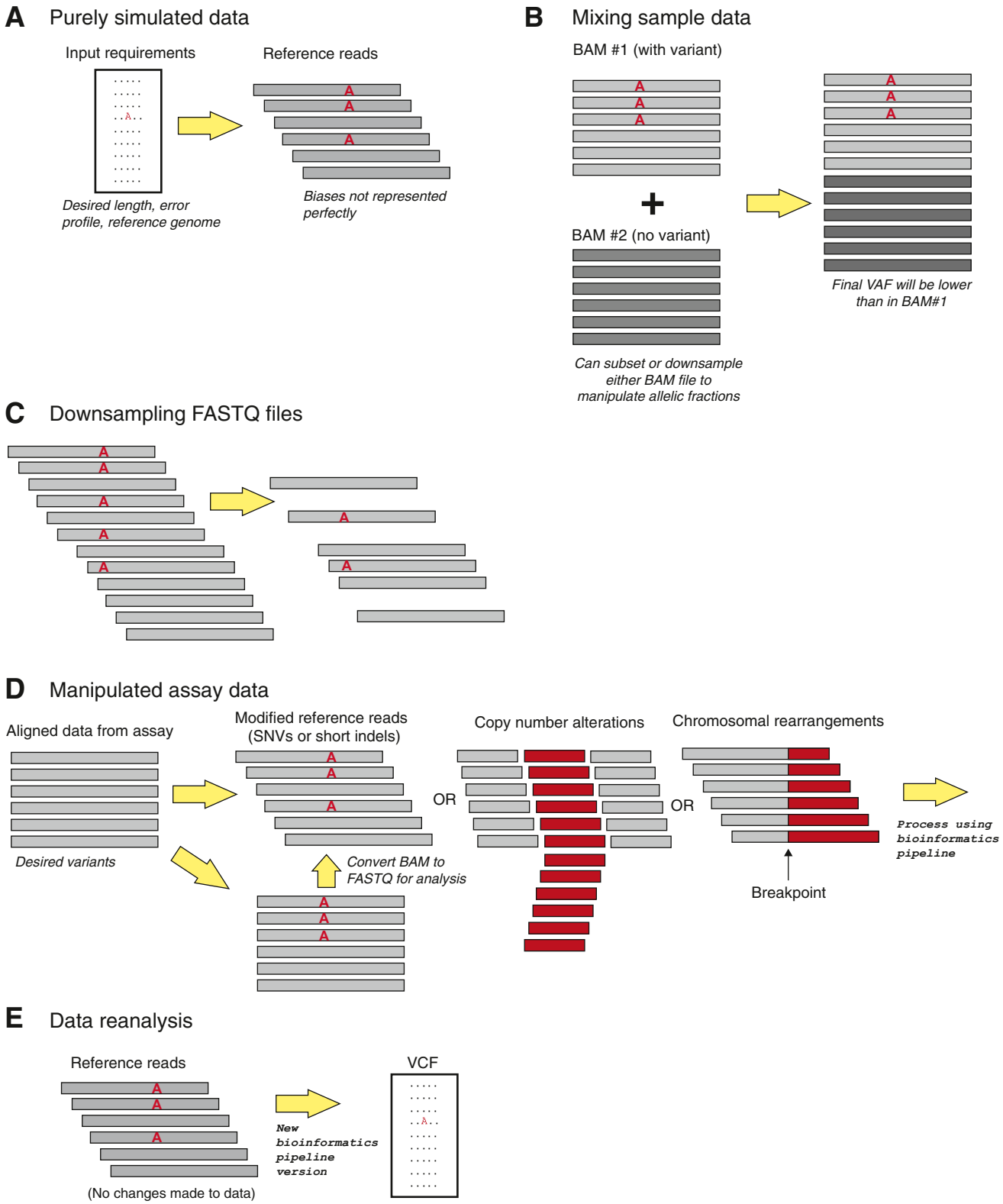

Figure 1 Types of *in silico* data. **A:** Purely simulated data can simulate almost any type of variant but cannot mimic all types of sequencing biases. **B:** Mixing sample data combines real data from more than one sample to represent particular variant allele frequencies (VAFs). **C:** Downsampling FASTQ files can be used to test the effect of lower coverage on variant calling performance. **D:** Manipulated assay data are one of the most common types of *in silico* data generated by clinical laboratories, because they use real data from the laboratory, which are modified to introduce variants, such as single-nucleotide variants (SNVs), copy number alterations, and chromosomal rearrangements. **E:** Data reanalysis can be used to help validate bioinformatics pipeline changes with existing data. Indel, insertion/deletion.

Table 1 Bioinformatics Tools that Produce Manipulated Assay Data

Name	GitHub URL*	Summary
Bamgineer (no version given, last update July 30, 2020)	*https://github.com/pughlab/bamgineer*	Simulates haplotype-phased, allele-specific copy number variants in an existing BAM file. Requires legacy dependencies samtools version 1.2 and pysam version 0.8.4.[41]
BAMSurgeon version 1.3	*https://github.com/adamewing/bamsurgeon/*	Simulates SNVs, small indels, and structural variants in an existing BAM file.[9]
insiM version 1.0	*https://github.com/thesushantpatil/insiM*	Simulates SNVs, small indels, and duplication events in an existing BAM file. Outputs a paired-end FASTQ file.[42]
VarBen (no version given, last update May 15, 2021)	*https://github.com/nccl-jmli/VarBen*	Simulates SNVs, indels, and structural variants in an existing BAM file.[43]

*These are examples of software packages available at the time of this writing and not comprehensive lists. Inclusion does not represent an organizational endorsement by the Association for Molecular Pathology of any individual product or service. All websites last accessed April 12, 2022.

Indel, insertion/deletion; SNV, single-nucleotide variant.

To better assess the technical aspects, limitations, and advantages of *in silico* variant simulation in the clinical laboratory, the Association for Molecular Pathology (AMP) convened a panel of subject matter experts to examine the topic. In this article, the Working Group explores the different types of *in silico* data files, how they can be used by clinical laboratories, and their advantages and disadvantages. Data from an Association for Molecular Pathology survey reflects how *in silico* files are being used by laboratories today. Finally, the Working Group provides recommendations and future directions for the use of *in silico* NGS data.

Types of *in Silico* Data

In silico data can be a powerful tool, but it is important to understand the strengths and limitations of the variety of types of *in silico* data and which aspects of the pipeline they can help validate. Different types of data will be useful for identifying different sources of errors (eg, systematic sequencing errors versus alignment errors), mimicking different types of variants (eg, small variants versus structural variants), and testing different variant origins (eg, germline versus acquired/somatic).

Data Reanalysis

When a change is only made to the bioinformatics pipeline, a laboratory can help validate the change by using existing, unmodified data from its assay on a variety of samples, including reference materials and clinical samples. This approach has the strength of using existing, real data that have all the biases and errors of the method used. The laboratory can test the ability to reproduce any variants detected by the previous version of the pipeline but does not test performance for any variants not previously detected in its samples.

Purely Simulated Data

Purely simulated data are generated from scratch by generating reads of the desired length and error profile from a reference genome modified to contain the variants of interest. These data are often used in the initial testing of bioinformatics tools because it is relatively easy to generate a large number of variants of practically any type. These data are particularly useful for complex variant types, where samples are scarce, and methods are immature. These data generally overestimate performance because read simulators normally do not model noise or biases of sequencing technologies perfectly, especially in repetitive regions where variants are challenging to detect (eg, at homopolymers). It is also difficult to model the biases and coverage distributions of the targeted sequencing methods most commonly used in clinical testing, and it is hard to model pre-analytical artifacts like those introduced by fixing formalin-fixed, paraffin-embedded tissues. Software packages used to generate purely simulated data are listed in Supplemental Table S1.

Manipulated Assay Data

One of the most common *in silico* strategies used for clinical assays is to modify the laboratory's real data by introducing variants into a fraction of the reads at one or more positions of interest. This approach has the advantage of working well with targeted or whole genome sequencing data and maintains many of the error profiles and biases of real data. It can also enable validation of a much larger number of variants of different types in different genome contexts at different allele fractions than real samples allow. It is generally important to convert these BAM files back to FASTQ so that the pipeline is tested from the initial alignment step. Databases, such as ClinVar, can be used to obtain important pathogenic variants in genes of interest and test the ability of

the pipeline to pick up these variants. There are important limitations of this approach: i) because it relies on reads being correctly mapped before modifying the reads, it will not model all mapping errors that can occur in difficult-to-map regions, such as genes with pseudogenes; ii) it does not model errors caused when a variant introduces poison motifs (eg, an expansion of a homopolymer or tandem repeat that causes higher sequencing error rates, or the introduction of a GGT sequence motif that can cause systematic sequencing errors)[51]; iii) it can be challenging to model larger variants that mimic real data perfectly (eg, modeling the coverage drop and breakpoints caused by large deletions or the sequencing errors inside large insertions); and iv) some sequencing technologies have raw forms of data, such as Ion Torrent (Thermo Fisher Scientific, Waltham, MA) flow data, raw reads from PacBio HiFi (Pacific Biosciences, Menlo Park, CA), and fast5 files from nanopore sequencing (Oxford Nanopore Technologies, Oxford, UK), which cannot be easily manipulated by current tools to introduce variants. Table 1 describes several tools that have been developed to generate manipulated assay data.

Mixing Sample Data

Another common *in silico* strategy for somatic clinical sequencing assays is to mix real data (eg, BAM or FASTQ files) from two samples together at different fractions, which can help assess the allele fraction detection limit for variants. This strategy can mix BAM or FASTQ files from two well-characterized normal samples (eg, two individuals from the Genome in a Bottle Consortium) such that the large number of germline variants in one sample, but not the other, will mimic somatic variants at a low allele fraction. This enables testing detection limits for a large number of variants of different types and in different genome contexts. However, mixing two normal samples like this does not mimic the complexity of many somatic variants that occur in tumors (eg, copy number variations and large structural variations). Also, sample mixing has the potential to generate unrealistic numbers of somatic variants, most of which are known germline variants, which can cause issues with some variant callers. There will also be some vertically complex (or in *trans* variants, where there are different variants on the two alleles, which can result in more than two alleles when mixing samples) and horizontally complex variants (or in *cis* variants, where there are multiple nearby variants on the same allele, which can cause challenges with representing variants when mixing samples), and the results need to be interpreted cautiously in these regions. Also, normal cell lines will often have variants present in a fraction of the cells that can appear to be false-positive somatic variants if these are not well characterized in the reference samples. Mixing two normal genomes will typically not simulate variants of clinical interest, so laboratories can also mix tumor genomes that contain variants of interest with the corresponding normal sample to test their ability to detect these variants at different allele fractions.

Downsampling FASTQ Files

To test a pipeline's ability to detect variants at different coverage levels, high-coverage FASTQ files can be downsampled (ie, a fraction of the reads can be randomly selected from a higher coverage data set). In general, it is important to randomly sample the desired fraction of reads from the entire FASTQ files and not just the first X percentage of reads. For paired-end sequencing, it is important to select both pairs of each read that is selected [eg, using a tool like seqtk (*https://github.com/lh3/seqtk*, last accessed May 19, 2022)]. When performing downsampling, the approach should reflect the desired conditions to be tested (eg, check that the desired coverage metrics are obtained after downsampling, and if variant calling is typically done from sequencing of a single library, then downsampling should be performed from a single library). This strategy can identify limitations in a pipeline's ability to detect variants at lower coverage levels, although it requires data containing the variants of interest at the allele fractions of interest using one of the above strategies.

Modifying Reference Genome

An *in silico* strategy that is more used for haploid genomes (eg, bacteria) than human genomes is to modify the reference genome to which reads are being aligned. When the reference genome is changed, the individual being sequenced should have a variant called at that location (assuming the individual matches the original reference). These variants will generally appear to be homozygous variants in autosomes of diploid genomes, which are easier to detect than heterozygous variants. Therefore, this approach works best for haploid samples or for haploid chromosomes, like chromosome X and chromosome Y in males outside the pseudoautosomal regions. One exception to this is that if a diploid individual has a heterozygous variant at a position, and the reference is changed to match the variant, then the variant would be reversed (eg, a C>T SNV would change to a T>C SNV, or a 2-bp deletion would change to a 2-bp insertion).

Applications for *in Silico* Data

The type of *in silico* approach that is used should depend on the exact clinical application of the test. Some common scenarios, with recommended *in silico* test designs, are described below and summarized in Table 2. It is important to perform these *in silico* tests in replicates or by using multiple samples from different sequencing runs (two at a minimum) so that sample- or run-specific biases do not influence the results.

Table 2 Recommended Applications for Different Types of *in Silico* Data

Purpose	Type of *in silico* data used*
Benchmark bioinformatics tools	Purely simulated data; manipulated assay data (where applicable)
Validation of bioinformatics pipeline:	
(a) New variants	(a) Manipulated assay data
(b) Limit of detection	(b) Mixing multiple samples; downsampling FASTQ files
(c) Lowest number of sequencing reads	(c) Downsampling FASTQ files
Assessment of the pipeline after updates or version changes:	
(a) Laboratory protocol changes	(a) Biological samples. *In silico* data can be supplementary depending on changes.
(b) Changes do not affect limits of tools in the pipeline	(b) Existing assay data. *In silico* data can be supplementary depending on changes.
(c) Changes affect the limits of tools in the pipeline	(c) Manipulated assay data in addition to existing assay data where applicable.
Proficiency testing	Manipulated assay data
Variant annotations	Manipulated assay data; VCF file manipulation

*Please see *Types of in Silico Data* and Recommendations in the article for discussions related to limitations of each type.

Assay Development Phase and Benchmarking Bioinformatics Tools

One of the foundational aspects of bioinformatics pipeline design is selection of bioinformatics tools for the detection of a given variant type. Because of the large variety of test designs and the large selection of tools available for every variant type, laboratories often test multiple tools to evaluate each tool's performance for their particular assay, sometimes called benchmarking. Herein, the question is whether the software accurately detects the variant in the data even when the assay captures and sufficiently covers the variant of interest. Developers of bioinformatic software also require sequencing data to test their code and ensure comparable performance with similar packages.[46,47] Whether designing assays or software, one might use either purely simulated data (eg, generated by example software packages) (Supplemental Table S1) or manipulated data files from prior sequencing runs of the same assay (example software packages) (Table 1).

For example, benchmarking tools for detection of copy number or structural variants might use purely simulated data for initial assessments. This permits bulk interrogation of a broad range of alterations, including those that might be encountered otherwise only rarely in biological samples. Kadri et al[47] utilized *in silico* data to benchmark the performance of a custom indel variant calling software module, generating multiple data sets to separately control for insert size and indel location.

Notably, however, the performance with respect to purely simulated data sets should not be conflated with actual analytical sensitivity or specificity. One might reasonably expect that if the bioinformatics tool being evaluated does not perform well with simulated data, it will perform worse when run on data from real biological material. Simulated data generally overestimate tool performance because the data cannot capture all biases and errors present in real data. Therefore, the purely simulated *in silico* data can be used only to understand some of the performance limitations of the software and are not ultimately suited for assay validations on their own (see below). Alternatively, one could deliberately simulate errors that emulate systematic artifacts, thereby probing the potential limits of assay specificity.

Manipulated data files from the assay can also be used for benchmarking (Table 2).[9,41–44] For example, insertion and deletion variants of increasing sizes can be introduced into existing BAM files, after which the BAM files should be converted back to FASTQ for concerted assessment of both alignment and variant calling stages of the pipeline. This might help benchmark the variant size limitations at which the performance of the tools decreases. Mixing data files from multiple samples can help introduce a large number of apparent somatic or mosaic variants at lower allelic fractions, as described in the section above. Both in the development phase and in subsequent validation, this technique is helpful to test the limit of detection of the bioinformatics pipeline, but notably, it will not test performance in regions where no variants already exist in the samples. Spencer et al[44] outline use of this approach in comparing sensitivity among a set of different variant calling software packages. Cheng et al[52] report mixing varying amounts of tumor sequencing reads into data from patient-matched normal tissue, ultimately determining the limit of sensitivity for a quality assurance check on tumor-normal contamination within their workflow.

Assessment of the Bioinformatics Pipeline during Assay Validation

As mentioned above, biological samples are not always available to cover all genes, exons, and/or other regions of interest tested by an NGS assay. Thus, after the experimental part of the assay is thoroughly validated with a

sufficient number of samples, *in silico* variants can be introduced across a wider range of genes to test the performance of the bioinformatics pipeline. The types of *in silico* data that can be used for this purpose are manipulated assay data or else sequencing files generated by mixing multiple samples, depending on the purpose as described below. Use of data generated purely *in silico* is not recommended for this critical phase of assay validation.

To test variants not already represented by biological validation samples, one may introduce variants in existing assay BAM files. For example, laboratories performing whole exome sequencing might supplement overall validation of the assay with manipulated assay data that contain rare variants, in particular subsets of genes that are reported clinically. Another exemplary use is the introduction of variants at low allele fraction, near to the edges of the amplicons (for amplicon-based assays), or in areas of low sequencing coverage.[44,45] As mentioned above, another method of evaluating variant calling performance in regions of low coverage entails downsampling existing sequencing data from the assay to fewer reads, until the point where the pipeline starts missing known variants in the sample. This provides some empirical data on the lowest number of acceptable reads for each sample type, variant type, allelic fraction, and sequence context (ie, GC enriched and low complexity). In one of the early published reports of a clinical NGS assay validation, Cottrell et al[45] describe downsampling variant positions with low allelic frequency to assess limit of detection as a function of coverage. The authors also showcase use of *in silico* sample mixtures for interrogation of the limit of detection. Other investigators have reported use of *in silico* sequencing data modeling to explore the relationship between coverage and sensitivity, which may be helpful when the latter is especially critical, such as for the design of a circulating tumor DNA assay.[53]

In addition to assessing technical performance of variant calling as described above, the tertiary analysis software can be tested for its performance in annotating challenging variants (eg, complex variants in *cis* and larger insertions or deletions). This can be done with manipulated assay data that are run through the entire pipeline, or, in a more targeted approach, by manipulating VCF files for different types of variants and testing the performance of the annotation software alone. This could include testing accuracy of converting VCF to Human Genome Variation Society (HGVS), which can be particularly challenging for complex variants. Similarly, variants may be introduced primarily for the purpose of testing therapy-assignment rules, clinical trial recommendations, and other clinical annotations. However, performing comparison of a laboratory's annotation to the correct annotation is often challenging and may need to be manually checked. In addition, laboratories should not rely solely on known pathogenic variants in databases like ClinVar to do *in silico* validation of their clinical interpretation pipeline. Detailed recommendations for use of *in silico* data for tertiary analysis are outside the scope of this article.

Please note that a similar caveat applies to these applications as that mentioned above for the assay development phase; namely, samples modified *in silico* may ultimately overestimate the performance of bioinformatic tools. Although data fabricated purely *de novo* may represent the greater concern, biological data edited *in silico* should still be regarded cautiously. All pitfalls and failure modes associated with the use of such data are difficult to predict, but users should broadly consider the consequences of false-positive or false-negative results that might result and whether these pose a substantial threat to either analytical sensitivity or specificity. To mitigate this possibility, *in silico* samples cannot be used solely during bioinformatics validation, as emphasized by Jennings et al.[4] Unmodified biological samples in sufficient numbers are critical to determining real-world performance of the analysis pipeline.

Assessment of the Pipeline after Updates or Version Changes to Software/Tools/Databases

The type of data best used for validation of updates to the bioinformatics pipeline depend on the type and scope of the updates. If the bioinformatics updates follow an experimental update, such as probes or any changes to the assay protocols, then additional biological samples are needed, and *in silico* data alone will not suffice for this. The data can supplement the biological samples, however, as described above.

If no such changes are made to the laboratory protocols, but minor changes are made to the bioinformatics pipeline, then running the newly configured pipeline on an appropriate number of existing sample data files from the laboratory's assay may be sufficient. However, if the pipeline changes affect the intrinsic limitations of the software (eg, detecting variants at lower frequencies or increasing the size of insertion and deletion variants that can be detected), then additional biological samples, as well as *in silico* data (eg, manipulated assay data files), could be used to assist in validation of the new capabilities.

Proficiency Testing and Quality Monitoring

Regular proficiency testing is important to monitor the performance of an NGS assay. For example, a proficiency test available commercially from College of American Pathologists provides *in silico* manipulated assay data to test the bioinformatics pipelines for certain assays. In addition to using external proficiency testing, internal proficiency testing standards can also be set up. However, internal proficiency tests need to be evaluated for accuracy before using them to evaluate the bioinformatics pipeline performance.

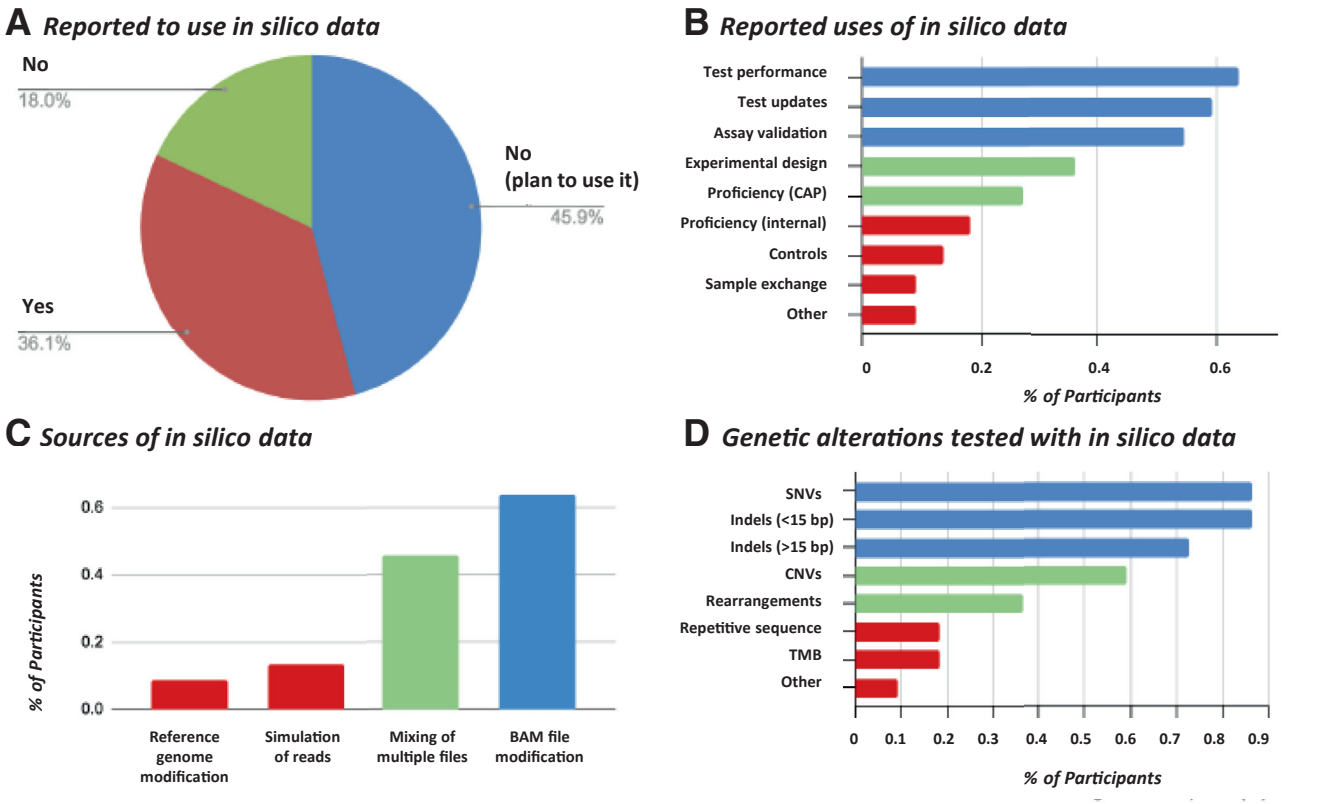

Figure 2 Aspects of *in silico* generated data files as part of clinical pipelines reported in the survey. **A:** Selected Association for Molecular Pathology survey data on current *in silico* data use. Most responding laboratories indicated that they were either currently using (36.1%) or planned to use (45.9%) *in silico* data. **B:** For laboratories currently using *in silico* data, the three most common reported uses included test performance, test updates, and assay validation. **C:** Survey data indicated that the most common sources of *in silico* data were from mixing of multiple files and modification of BAM files. **D:** Survey data indicating the types of variants laboratories currently model with *in silico* data. CAP, College of American Pathologists; CNV, copy number variation; Indel, insertion/deletion; SNV, single-nucleotide variant; TMB, tumor mutation burden.

AMP Laboratory Practices Survey

The Working Group designed a survey composed of 23 questions with two objectives: to explore the existing practice regarding the use of *in silico* data for the bioinformatics pipeline development, optimization, and validation, and to understand the limitations and challenges associated with its use. This survey was offered to both members and nonmembers of AMP, distributed via multiple communications channels (eg, AMP member listserv, direct email, and social media), and was open from October 16, 2019, to November 15, 2019. Responses obtained from 61 participants were de-identified before the survey data were analyzed. Analysis of the survey data was performed using the R version 3.6.1 environment (The R Project for Statistical Computing; *https://cran.r-project.org*).

In this survey, 43% of participants self-identified as molecular genetics professionals, 27% self-identified as clinical laboratory directors, 21% self-identified as bioinformaticians, and 18% self-identified as pathologists. The rest of the participants' roles in clinical laboratories included supervisor, medical director, technologist or technician, and scientist or physician. Most participants were employed by academic laboratories (54%), followed by private (11%) and non-profit (10%) organizations. A total of 31% of laboratories have performed NGS-based testing for >5 years, 21% for 4 to 5 years, 18% for 2 to 3 years, 15% for <1 year, 8% for 1 to 2 years, and 7% for 3 to 4 years. About 74% of these laboratories employed a bioinformatician or software engineer. Of those laboratories, 31% have more than four full-time bioinformaticians, 27% have one, 20% have one to two, and 18% have two to four. These professionals were mostly PhDs (69%), whereas 27% had a master's degree and 4% had a bachelor's degree.

Several sequencing platforms were reported to be used for clinical purposes, according to the responses in this survey. Illumina (San Diego, CA) was used by most participants (82%), followed by Thermo Fisher Scientific (35%), and others, such as Oxford Nanopore (7%), 10x Genomics (Pleasanton, CA) (3%), and PacBio (2%). Approximately 84% of participants were performing bioinformatics analysis of clinical results at the time of the

survey, 11% were in the process of validation, and 5% were not doing bioinformatics. Of these participants, 62% had developed their internal bioinformatics pipeline (including open-source software), and 56% were using bioinformatics software provided by the sequencing vendor. From commercially available bioinformatics solutions for secondary analysis (alignment and variant calling), the most commonly reported to be in use were Torrent Suite (39%), MiSeq Reporter (Illumina, San Diego, CA) (31%), and Illumina BaseSpace (16%) (Supplemental Figure S1).

Of participants, 36% already use *in silico* generated NGS data files (Figure 2A), most commonly to test pipeline performance, technical limitations, updates, and assay validation (Figure 2B). Of participants, 46% were planning to but not yet using *in silico* data, and 18% were neither using nor planning to use it. Some of the issues reported in using *in silico* data files were difficult to find or download files (39% of responses), *in silico* data are not considered adequate for a robust validation (21%), difficult to load files in the workflow (20%), and *in silico* data do not represent assay correctly (18%).

Use of *in silico* generated data was reported mostly for DNA sequencing-based assays (82%), with somatic (64%) or germline (59%) applications. RNA and circulating free DNA were also an application but with less frequency. Modifications of real BAM files by introducing variants in the reads (manipulated assay data) were the most common methods to generate *in silico* data (64%), although other methods were also used (Figure 2C). Variant sources included internal databases (59%), Catalogue of Somatic Mutations in Cancer (41%), gnomAD (8%), ClinVar (8%), Human Gene Mutation Database (4%), and others, such as scientific literature. The most common simulated variant types were SNVs and small indels (<15 bp) (Figure 2D).

Approximately 28% of participants reported making changes to the pipeline based on results from *in silico* files only versus 72% who did not. Cost of pipeline revalidation was reported to be <$5000 by 25% of participants, $10,000 to $20,000 by 25%, $5000 to $10,000 by 18%, $20,000 to $50,000 by 16%, and >$50,000 by 15%. Assay or pipeline revalidation was performed once a year by 20% of laboratories, twice a year (18%), fewer than once a year (16%), more than twice a year (16%), once every 2 years (15%), and never (15%).

Opportunities for Expanding *in Silico* Modeling and Future Directions

Copy Number Variants

Copy number variants are a clinically important class of genetic alterations that have diagnostic, therapeutic, and prognostic significance in the management of cancer and constitutional disorders. Although different bioinformatics algorithms have been developed to identify copy number variations, including copy neutral events, they are generally more challenging to detect than small variants, particularly when copy number alterations (CNAs) are present at a subclonal (tumor) or mosaic (germline) level. It is therefore important to perform a comprehensive validation of a clinical NGS assay designed to identify CNAs. In contrast to single-nucleotide variants and small insertions and deletions, CNAs are relatively less common and therefore are harder to procure in sufficient numbers to perform a comprehensive validation. Therefore, methods to generate *in silico* data sets for CNAs are highly desirable for supplementing a clinical NGS validation.

At this time of literature review, there are few *in silico* tools and algorithms for CNA simulation that insert numerical losses and gains as well as allele-specific loss of heterozygosity in aligned sequences (BAM and CRAM formats) from real samples. The general principles of these algorithms focus on the sampling of random reads, with or without allele specificity, from specified regions of the aligned sequences and modify the BAM file to add excess or remove reads. Such changes result in the introduction of copy number gains (or amplifications) or losses, respectively, in the BAM file, which can be used for evaluating the performance of CNA calling algorithms. Bamgineer[41] is a recently published algorithm that can introduce user-defined, allele-specific CNA at any desired level into a BAM file. The algorithm accounts for the read pairs in paired-end sequencing data when sampling the reads from the BAM file. This approach attempts to preserve the original biases in the BAM file and better mimic CNAs in real samples. The algorithm can be applied in many use cases, such as simulating CNAs at low allelic burden in cell-free DNA samples and subclonal CNA detection.

VarBen is a new comprehensive *in silico* variant simulation algorithm that introduces a wide variety of genetic alterations in a BAM file, including SNVs, insertions, deletions, large structural variants, including copy number alterations, duplications, and balanced and unbalanced translocations.[43] Other software packages capable of simulating or editing copy number alterations are listed in Table 1 and Supplemental Table S1. Although not published as a formal tool, a study by Ellingford et al[49] also demonstrated the use of random read sampling strategy from specific regions of their targeted panel's BAM file to simulate copy number gains and losses. It was used to supplement the validation of a CNA detection using a targeted panel.

Translocation (Gene Fusion) Assessment

Translocations are commonly detected in clinical NGS assays by targeted DNA/RNA sequencing, RNA sequencing, or whole genome sequencing.[54] At the DNA level, most translocations occur in introns, which may contain repeats or low complexity regions that are difficult to analyze. Similarly, translocations must result in sufficient numbers of gene fusion transcripts to be detected by RNA sequencing. These issues make extensive validation of translocation detection performance critical for NGS assays. However, for

many translocations, such as those in *ROS1*, *RET*, or *NTRK*, it can be difficult to find sufficient numbers of cases with translocations to fully test translocation detection in the assay validation process. For this reason, the use of *in silico* modified data would be useful. Several software packages purport to simulate translocations, both *de novo*[30–32,37] (Supplemental Table S1) and via modification of existing BAM files[9,43] (Table 1). Given the complexity of the sequence alterations involved, however, a degree of caution is warranted when using such tools.

RNA Sequencing

RNA sequencing is becoming more common in the clinical laboratory to detect translocation/fusion events, to measure gene-level expression, to resolve variants of uncertain significance, and to measure allele-specific expression.[55] RNA itself is highly labile, making replicate testing of physical samples for quality control purposes difficult. Multiple RNA sequencing simulation tools exist[50] [Polyester (*https://bioconductor.org/packages/release/bioc/html/polyester.html*, last accessed September 1, 2022), rlsim (*https://github.com/sbotond/rlsim*, last accessed September 1, 2022), RNASeqReadSimulator (*https://github.com/davidliwei/RNASeqReadSimulator*, last accessed September 1, 2022), and simCT (*https://github.com/jaudoux/simct*, last accessed September 1, 2022)] but were not specifically considered for inclusion in Supplemental Table S1. The lack of peer reviewed publication for some of these cited tools speaks to further opportunities for development in this domain.

Clonality Assessment

Some laboratories have developed assays to detect dominant clones of blood cells with identical or highly conserved sequence that can indicate a neoplasm with clonally rearranged T-cell receptor or IgH.[56] Currently, clinical NGS assays detect clonality by comparing individual reads with a standard database of nucleic acid sequences. In addition to clonality, these assays can use alignment search functions to detect rearrangements of the variable and joining regions. Once a clonal sequence has been characterized, subsequent tests can be performed in which low levels of the clonal sequence can be searched as a method for minimal residual disease testing.[57] Development of software that can simulate these events *in silico* would be valuable to laboratories implementing these assays.

TMB and MSI Testing

Both TMB and MSI have been shown to be important markers of therapeutic response in cancer, and reporting of these two metrics is now included in somatic cancer panels. TMB and MSI represent measurements calculated from the observed number of somatic variants per megabase of DNA and expansion of dinucleotide repeats in specific regions of the genome, respectively. Both types of underlying events (somatic variants or microsatellite expansions) can be simulated by current *in silico* genome modeling tools, but the Working Group was unaware of software designed specifically to model TMB or MSI.

Microbiome Assessment

NGS data can be efficiently used to assess a multitude of microbial species in a single assay and have been established as a viable technique for multiple clinical scenarios in which traditional microbiological assessment does not identify a causal species.[58] The traditional techniques have utilized amplification of the 16S ribosomal RNA sequence to search for the presence of bacteria species. However, shotgun metagenomic sequencing is now being explored for detection of occult infection in clinical practice.[59] A variety of novel and existing *in silico* tools could help assess clinical performance characteristics of microbiome/metagenomic sequencing. Although such tools were not specifically sought for inclusion in Supplemental Table S1, Shcherbina[18] describes use of FASTQsim, a general read simulator, for generation of metagenomic data sets.

Minimal Residual Disease Testing

Techniques to overcome the error rate of modern next-generation sequencing have been developed to enable detection of variants at variant fractions well below 1% in clinical samples.[60,61] The strategy to achieve this is to use unique molecular indexes to label each DNA molecule before amplification. These unique molecular indexes are then amplified, and variants are detected on the basis of a composite of the amplified segments associated with the unique molecular index. To date, no software has been developed to simulate minimal residual disease testing *in silico* by incorporating unique molecular indexes with spiked-in variants. As this form of testing proliferates, such techniques would be of tremendous value to the clinical community.

Long-Read Sequencing Methods

Until recently, long-read sequencing methods from Pacific Biosciences and Oxford Nanopore Technologies were expensive and/or had an unacceptably high error rate for most clinical applications. However, long reads can enable accurate calling of variants challenging for short reads, such as genes with homologous genes or pseudogenes and structural variants, as well as phasing of variants over long stretches of the genome.[62,63] As long reads become increasingly cost-effective and accurate for clinical applications, techniques for simulating and editing variants in long-read files will need to be developed in addition to the methods that simulate long reads *de novo* in Supplemental Table S1.

Table 3 Recommendations for Use of *in Silico* Data Files within the Clinical Laboratory

Recommendation*	Description
1	The laboratory may use *in silico* data files to supplement NGS analytical validation, particularly to assess analytical sensitivity or false-negative rates for specific variants; however, *in silico* data files cannot supplant the use of physical samples (eg, patient samples).
2	The laboratory should understand the functional limitations of the type(s) of *in silico* data being utilized.
3	The laboratory should understand the limitations of most *in silico* data for assessing performance in particular genome contexts and variant types susceptible to systematic sequencing errors (eg, homopolymers and tandem repeats) and mapping errors (eg, genes with pseudogenes).
4	The laboratory may use *in silico* samples for testing required for minor updates to clinical bioinformatics software pipelines.
5	Commercial vendors and internal pipeline developers should include options in their analysis pipelines to facilitate easier *in silico* data file import and analysis by clinical laboratories.

*Please see article for discussions related to each best practice recommendation listed. The limitations of each type of *in silico* data are discussed in *Types of in Silico Data* in this article.

NGS, next-generation sequencing.

General Recommendations

On the basis of the survey results and a directed review of the available published literature, the Working Group developed several expert consensus opinion general recommendations for *in silico* data, summarized in Table 3.

Recommendation 1: The Laboratory May Use *in Silico* Data Files to Supplement NGS Analytical Validation, Particularly to Assess Analytical Sensitivity or False-Negative Rates for Specific Variants; However, *in Silico* Data Files Cannot Supplant the Use of Physical Samples (eg, Patient Samples)

In general, *in silico* data are best used to characterize false-negative rates (sensitivity) rather than false-positive rates (specificity). For example, the College of American Pathologists Molecular Pathology Checklist[64] requires sequencing of physical samples as part of the test validation process.

Recommendation 2: The Laboratory Should Understand the Functional Limitations of the Type(s) of *in Silico* Data Being Utilized

The limitations of each type of *in silico* data are discussed in *Types of* in Silico *Data* in this article. It is critically important to understand the functional limitations of the type(s) of *in silico* data being utilized in NGS bioinformatics pipeline validation and the potential downstream impacts on establishing and/or monitoring assay performance characteristics to avoid serious pitfalls. For example, *in silico* data generated by modifying existing data files may better reflect the systematic sequencing errors, off-target reads, paired-end distance, and coverage variability across the genes targeted by clinical sequencing panels when compared with *de novo* simulated *in silico* data. In general, multiple data sets generated by one's own laboratory with the standard workflow should be used.

Recommendation 3: The Laboratory Should Understand the Limitations of Most *in Silico* Data for Assessing Performance in Particular Genome Contexts and Variant Types Susceptible to Systematic Sequencing Errors (eg, Homopolymers and Tandem Repeats) and Mapping Errors (eg, Genes with Pseudogenes)

It is important to understand limitations of most *in silico* data for assessing performance in particular genome contexts and variant types. In particular, even modifying real data files will not mimic some systematic errors, such as in homopolymers and tandem repeats. Because modifying real data depends on proper mapping of reads, it also usually cannot assess errors in difficult to map regions or segmental duplications, such as genes with pseudogenes or highly homologous genes, like *PRSS1*, *PMS2*, and *SMN1*/*SMN2*, or genes with errors in GRCh38, like *CBS*, *U2AF1*, and *KCNE1*.[65,66]

Recommendation 4: The Laboratory May Use *in Silico* Samples for Testing Required for Minor Updates to Clinical Bioinformatics Software Pipelines

In silico data that are used for testing updates or version changes to software/tools/databases should use existing data from the laboratory.

Recommendation 5: Commercial Vendors and Internal Pipeline Developers Should Include Options in Their Analysis Pipelines to Facilitate Easier *in Silico* Data File Import and Analysis by Clinical Laboratories

This will enable broader use of *in silico* data. Like any bioinformatics software, data simulation software packages vary in terms of the required inputs and expected outputs, license terms, operating system compatibility and software dependencies, regularity of bug fixes and maintenance, and ease of installation and use. In addition to features and usability, the quality and community acceptance of a particular

software should be considered before adoption in a clinical sequencing workflow, although admittedly, these may be difficult to assess rigorously.

Discussion

In this article, the Working Group describes the utility of *in silico* data in the clinical NGS laboratory, how *in silico* data files are made, and potential future applications for *in silico* data. AMP's current practices survey data indicate that 35% of clinical laboratories are already using *in silico* data, with another 45% planning to start using it. For the purpose of validation testing, *in silico* data have multiple advantages over physical samples in that they can be used to simulate difficult to find gene variants or class specific classes of variants, such as medium-sized insertions that are often difficult to detect by NGS pipelines. A recent study finding one in seven pathogenic variants is challenging to detect with NGS highlights the importance of evaluating performance for challenging variants, and *in silico* data can help with some but not all types of challenging variants.[7] *In silico* data can also be used to simulate multiple low-frequency (low variant allele frequency) variants that can be used to derive the limit of detection performance of NGS pipelines. *In silico* data files may also be engineered to contain multiple variants, allowing for a single sample to interrogate multiple different regions of a targeted panel or different classes of variants. As *in silico* data applications increase, it is important to understand the strengths and limitations of these data as described in this article. As a resource, the Working Group provides a list of example tools available at the time of this publication that can be used to generate *in silico* data (Supplemental Table S1).

The Working Group acknowledges that published evidence about utilization of *in silico* samples in NGS validation continues to evolve and emerge. This article and included recommendations will be reviewed approximately 2 years after publication by the AMP Clinical Practice Committee to determine the need for an update based on the evidence available at that time.

Disclaimers

The Association for Molecular Pathology (AMP) Clinical Practice Guidelines and Reports are developed to be of assistance to laboratory and other health care professionals by providing guidance and recommendations for particular areas of practice. The Guidelines or Reports should not be considered inclusive of all proper approaches or methods, or exclusive of others. The Guidelines or Reports cannot guarantee any specific outcome, nor do they establish a standard of care. The Guidelines or Reports are not intended to dictate the treatment of a particular patient. Treatment decisions must be made on the basis of the independent judgment of health care providers and each patient's individual circumstances. The AMP makes no warranty, express or implied, regarding the Guidelines or Reports and specifically excludes any warranties of merchantability and fitness for a particular use or purpose. The AMP shall not be liable for direct, indirect, special, incidental, or consequential damages related to the use of the information contained herein.

Acknowledgments

The Association for Molecular Pathology (AMP) *In Silico* Pipeline Validation Working Group thanks Mrudula Pullambhatla (AMP) and Robyn Temple-Smolkin (AMP) for support and contributions to the development of this article.

Supplemental Data

Supplemental material for this article can be found at *http://doi.org/10.1016/j.jmoldx.2022.09.007*.

References

1. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, Voelkerding K, Rehm HL: Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med 2015, 17:405–424
2. Li MM, Datto M, Duncavage EJ, Kulkarni S, Lindeman NI, Roy S, Tsimberidou AM, Vnencak-Jones CL, Wolff DJ, Younes A, Nikiforova MN: Standards and guidelines for the interpretation and reporting of sequence variants in cancer: a joint consensus recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. J Mol Diagn 2017, 19:4–23
3. Roy S, Coldren C, Karunamurthy A, Kip NS, Klee EW, Lincoln SE, Leon A, Pullambhatla M, Temple-Smolkin RL, Voelkerding KV, Wang C, Carter AB: Standards and guidelines for validating next-generation sequencing bioinformatics pipelines. J Mol Diagn 2018, 20:4–27
4. Jennings LJ, Arcila ME, Corless C, Kamel-Reid S, Lubin IM, Pfeifer J, Temple-Smolkin RL, Voelkerding KV, Nikiforova MN: Guidelines for validation of next-generation sequencing–based oncology panels: a joint consensus recommendation of the Association for Molecular Pathology and College of American Pathologists. J Mol Diagn 2017, 19:341–365
5. Zook JM, McDaniel J, Olson ND, Wagner J, Parikh H, Heaton H, Irvine SA, Trigg L, Truty R, McLean CY, De La Vega FM, Xiao C, Sherry S, Salit M: An open resource for accurately benchmarking small variant and reference calls. Nat Biotechnol 2019, 37:561–566
6. He HJ, Stein EV, Konigshofer Y, Forbes T, Tomson FL, Garlick R, Yamada E, Godfrey T, Abe T, Tamura K, Borges M, Goggins M, Elmore S, Gulley ML, Larson JL, Ringel L, Haynes BC, Karlovich C, Williams PM, Garnett A, Ståhlberg A, Filges S, Sorbara L, Young MR, Srivastava S, Cole KD: Multilaboratory assessment of a new reference material for quality assurance of cell-free tumor DNA measurements. J Mol Diagn 2019, 21:658–676
7. Lincoln SE, Hambuch T, Zook JM, Bristow SL, Hatchell K, Truty R, Kennemer M, Shirts BH, Fellowes A, Chowdhury S, Klee EW, Mahamdallie S, Cleveland MH, Vallone PM, Ding Y, Seal S, DeSilva W, Tomson FL, Huang C, Garlick RK, Rahman N, Salit M,

Kingsmore SF, Ferber MJ, Aradhya S, Nussbaum RL: One in seven pathogenic variants can be challenging to detect by NGS: an analysis of 450,000 patients with implications for clinical sensitivity and genetic test implementation. Genet Med 2021, 23:1673−1680
8. Sims DJ, Harrington RD, Polley EC, Forbes TD, Mehaffey MG, McGregor PM, Camalier CE, Harper KN, Bouk CH, Das B, Conley BA, Doroshow JH, Williams PM, Lih CJ: Plasmid-based materials as multiplex quality controls and calibrators for clinical next-generation sequencing assays. J Mol Diagn 2016, 18:336−349
9. Ewing AD, Houlahan KE, Hu Y, Ellrott K, Caloian C, Yamaguchi TN, Bare JC, P'Ng C, Waggott D, Sabelnykova VY, Kellen MR, Norman TC, Haussler D, Friend SH, Stolovitzky G, Margolin AA, Stuart JM, Boutros PC: Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. Nat Methods 2015, 12:623−630
10. Duncavage EJ, Abel HJ, Merker JD, Bodner JB, Zhao Q, Voelkerding KV, Pfeifer JD: A model study of in silico proficiency testing for clinical next-generation sequencing. Arch Pathol Lab Med 2016, 140:1085−1091
11. Huang W, Li L, Myers JR, Marth GT: ART: a next-generation sequencing read simulator. Bioinformatics 2012, 28:593−594
12. Frampton M, Houlston R: Generation of artificial FASTQ files to evaluate the performance of next-generation sequencing pipelines. PLoS One 2012, 7:e49110
13. Xie Q, Liu Q, Mao F, Cai W, Wu H, You M, Wang Z, Chen B, Sun ZS, Wu J: A Bayesian framework to identify methylcytosines from high-throughput bisulfite sequencing data. PLoS Comput Biol 2014, 10:e1003853
14. Cao MD, Ganesamoorthy D, Zhou C, Coin LJM: Simulating the dynamics of targeted capture sequencing with CapSim. Bioinformatics 2018, 34:873−874
15. Caboche S, Audebert C, Lemoine Y, Hot D: Comparison of mapping algorithms used in high-throughput sequencing: application to Ion Torrent data. BMC Genomics 2014, 15:264
16. Li Y, Han R, Bi C, Li M, Wang S, Gao X: DeepSimulator: a deep simulator for nanopore sequencing. Bioinformatics 2018, 34:2899−2908
17. Li Y, Wang S, Wang S, Bi C, Qiu Z, Li M, Gao X: DeepSimulator1.5: a more powerful, quicker and lighter simulator for nanopore sequencing. Bioinformatics 2020, 36:2578−2580
18. Shcherbina A: FASTQSim: platform-independent data characterization and in silico read generation for NGS datasets. BMC Res Notes 2014, 7:533
19. Balzer S, Malde K, Lanzén A, Sharma A, Jonassen I: Characteristics of 454 pyrosequencing data-enabling realistic simulation with flowsim. Bioinformatics 2010, 26:i420−i425
20. McElroy KE, Luciani F, Thomas T: GemSIM: general, error-model based simulator of next-generation sequencing data. BMC Genomics 2012, 13:74
21. Angly FE, Willner D, Rohwer F, Hugenholtz P, Tyson GW: Grinder: a versatile amplicon and shotgun sequence simulator. Nucleic Acids Res 2012, 40:e94
22. Yuan X, Zhang J, Yang L: IntSIM: an integrated simulator of next-generation sequencing data. IEEE Trans Biomed Eng 2017, 64:441−451
23. Lau B, Mohiyuddin M, Mu JC, Fang LT, Asadi NB, Dallett C, Lam HYK: LongiSLND: in silico sequencing of lengthy and noisy datatypes. Bioinformatics 2016, 32:3829−3832
24. Luo R, Sedlazeck FJ, Darby CA, Kelly SM, Schatz MC: LRSim: a linked-reads simulator generating insights for better genome partitioning. Comput Struct Biotechnol J 2017, 15:478−484
25. Holtgrewe M: Mason—A Read Simulator for Second Generation Sequencing Data. Berlin, Germany: FU Berlin, 2010
26. Yang C, Chu J, Warren RL, Birol I: NanoSim: nanopore sequence read simulator based on statistical characterization. Gigascience 2017, 6:1−6
27. Stephens ZD, Hudson ME, Mainzer LS, Taschuk M, Weber MR, Iyer RK: Simulating next-generation sequencing datasets from empirical mutation and sequencing models. PLoS One 2016, 11:e0167047
28. Wei ZG, Zhang SW: NPBSS: a new PacBio sequencing simulator for generating the continuous long reads with an empirical model. BMC Bioinformatics 2018, 19:177
29. Ono Y, Asai K, Hamada M: PBSIM: PacBio reads simulator - toward accurate genome assembly. Bioinformatics 2013, 29:119−121
30. Hu X, Yuan J, Shi Y, Lu J, Liu B, Li Z, Chen Y, Mu D, Zhang H, Li N, Yue Z, Bai F, Li H, Fan W: pIRS: profile-based Illumina pair-end reads simulator. Bioinformatics 2012, 28:1533−1535
31. Xia Y, Liu Y, Deng M, Xi R: Pysim-sv: a package for simulating structural variation data with GC-biases. BMC Bioinformatics 2017, 18:53
32. Bartenhagen C, Dugas M: RSVSim: an R/Bioconductor package for the simulation of structural variations. Bioinformatics 2013, 29:1679−1681
33. Xing Y, Dabney AR, Li X, Wang G, Gill CA, Casola C: SECNVs: a simulator of copy number variants and whole-exome sequences from reference genomes. Front Genet 2020, 11:82
34. Chen S, Han Y, Guo L, Hu J, Gu J: SeqMaker: a next generation sequencing simulator with variations, sequencing errors and amplification bias integrated, The Institute of Electrical and Electronics Engineers (IEEE) International Conference on Bioinformatics and Biomedicine (BIBM); 2016. pp. 835−840
35. Baker EAG, Goodwin S, McCombie WR, Mendivil Ramos O: SiLiCO: a simulator of long read sequencing in PacBio and Oxford Nanopore. bioRxiv 2016, [Preprint] doi:10.1101/076901
36. Stöcker BK, Köster J, Rahmann S: SimLoRD: simulation of long read data. Bioinformatics 2016, 32:2704−2706
37. Yue JX, Liti G: SimuG: a general-purpose genome simulator. Bioinformatics 2019, 35:4442−4444
38. Pattnaik S, Gupta S, Rao AA, Panda B: SInC: an accurate and fast error-model based simulator for SNPs, indels and CNVs coupled with a read generator for short-read sequence data. BMC Bioinformatics 2014, 15:40
39. Bolognini D, Sanders A, Korbel JO, Magi A, Benes V, Rausch T: VISOR: a versatile haplotype-aware structural variant simulator for short- and long-read sequencing. Bioinformatics 2020, 36:1267−1269
40. Kim S, Jeong K, Bafna V: Wessim: a whole-exome sequencing simulator based on in silico exome capture. Bioinformatics 2013, 29:1076−1077
41. Samadian S, Bruce JP, Pugh TJ: Bamgineer: introduction of simulated allele-specific copy number variants into exome and targeted sequence data sets. PLoS Comput Biol 2018, 14:e1006080
42. Patil SA, Mujacic I, Ritterhouse LL, Segal JP, Kadri S: insiM: in silico mutator software for bioinformatics pipeline validation of clinical next-generation sequencing assays. J Mol Diagn 2019, 21:19−26
43. Li Z, Fang S, Zhang R, Yu L, Zhang J, Bu D, Sun L, Zhao Y, Li J: VarBen: generating in silico reference data sets for clinical next-generation sequencing bioinformatics pipeline evaluation. J Mol Diagn 2021, 23:285−299
44. Spencer DH, Tyagi M, Vallania F, Bredemeyer AJ, Pfeifer JD, Mitra RD, Duncavage EJ: Performance of common analysis methods for detecting low-frequency single nucleotide variants in targeted next-generation sequence data. J Mol Diagn 2014, 16:75−88
45. Cottrell CE, Al-Kateb H, Bredemeyer AJ, Duncavage EJ, Spencer DH, Abel HJ, Lockwood CM, Hagemann IS, O'Guin SM, Burcea LC, Sawyer CS, Oschwald DM, Stratman JL, Sher DA, Johnson MR, Brown JT, Cliften PF, George B, McIntosh LD, Shrivastava S, Nguyen TT, Payton JE, Watson MA, Crosby SD, Head RD, Mitra RD, Nagarajan R, Kulkarni S, Seibert K, Virgin HW IV, Milbrandt J, Pfeifer JD: Validation of a next-generation sequencing assay for clinical molecular oncology. J Mol Diagn 2014, 16:89−105

46. Balan J, Jenkinson G, Nair A, Saha N, Koganti T, Voss J, Zysk C, Barr Fritcher EG, Ross CA, Giannini C, Raghunathan A, Kipp BR, Jenkins R, Ida C, Halling KC, Blackburn PR, Dasari S, Oliver GR, Klee EW: SeekFusion - a clinically validated fusion transcript detection pipeline for PCR-based next-generation sequencing of RNA. Front Genet 2021, 12:739054
47. Kadri S, Zhen CJ, Wurst MN, Long BC, Jiang ZF, Wang YL, Furtado LV, Segal JP: Amplicon Indel Hunter is a novel bioinformatics tool to detect large somatic insertion/deletion mutations in amplicon-based next-generation sequencing data. J Mol Diagn 2015, 17:635−643
48. Makrooni MA, O'Sullivan B, Seoighe C: Bias and inconsistency in the estimation of tumour mutation burden. BMC Cancer 2022, 22:840
49. Ellingford JM, Campbell C, Barton S, Bhaskar S, Gupta S, Taylor RL, Sergouniotis PI, Horn B, Lamb JA, Michaelides M, Webster AR, Newman WG, Panda B, Ramsden SC, Black GCM: Validation of copy number variation analysis for next-generation sequencing diagnostics. Eur J Hum Genet 2017, 25:719−724
50. Bruno AE, Miecznikowski JC, Qin M, Wang J, Liu S: FUSIM: a software tool for simulating fusion transcripts. BMC Bioinformatics 2013, 14:13
51. Meacham F, Boffelli D, Dhahbi J, Martin DIK, Singer M, Pachter L: Identification and correction of systematic error in high-throughput sequence data. BMC Bioinformatics 2011, 12:451
52. Cheng DT, Mitchell TN, Zehir A, Shah RH, Benayed R, Syed A, Chandramohan R, Liu ZY, Won HH, Scott SN, Rose Brannon A, O'Reilly C, Sadowska J, Casanova J, Yannes A, Hechtman JF, Yao J, Song W, Ross DS, Oultache A, Dogan S, Borsu L, Hameed M, Nafa K, Arcila ME, Ladanyi M, Berger MF: Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): a hybridization capture-based next-generation sequencing clinical assay for solid tumor molecular oncology. J Mol Diagn 2015, 17:251−264
53. Deveson IW, Gong B, Lai K, LoCoco JS, Richmond TA, Schageman J, et al: Evaluating the analytical validity of circulating tumor DNA sequencing assays for precision oncology. Nat Biotechnol 2021, 39:1115−1128
54. Duncavage E, Schroeder M, O'Laughlin M, Wilson R, MacMillan S, Bohannon A, et al: Genome sequencing as an alternative to cytogenetic analysis in myeloid cancer. N Engl J Med 2021, 384:924−935
55. Marco-Puche G, Lois S, Benítez J, Trivino JC: RNA-Seq perspectives to improve clinical diagnosis. Front Genet 2019, 10:1152
56. Boyd SD, Marshall EL, Merker JD, Maniar JM, Zhang LN, Sahaf B, Jones CD, Simen BB, Hanczaruk B, Nguyen KD, Nadeau KC, Egholm M, Miklos DB, Zehnder JL, Fire AZ: Measurement and clinical monitoring of human lymphocyte clonality by massively parallel V-D-J pyrosequencing. Sci Transl Med 2009, 1:12ra23
57. Logan AC, Vashi N, Faham M, Carlton V, Kong K, Buño I, Zheng J, Moorhead M, Klinger M, Zhang B, Waqar A, Zehnder JL, Miklos DB: Immunoglobulin and T cell receptor gene high-throughput sequencing quantifies minimal residual disease in acute lymphoblastic leukemia and predicts post-transplantation relapse and survival. Biol Blood Marrow Transplant 2014, 20:1307−1313
58. Salipante SJ, Sengupta DJ, Rosenthal C, Costa G, Spangler J, Sims EH, Jacobs MA, Miller SI, Hoogestraat DR, Cookson BT, McCoy C, Matsen FA, Shendure J, Lee CC, Harkins TT, Hoffman NG: Rapid 16S rRNA next-generation sequencing of polymicrobial clinical samples for diagnosis of complex bacterial infections. PLoS One 2013, 8:e65226
59. Ivy MI, Thoendel MJ, Jeraldo PR, Greenwood-Quaintance KE, Hanssen AD, Abdel MP, Chia N, Yao JZ, Tande AJ, Mandrekar JN, Patel R: Direct detection and identification of prosthetic joint infection pathogens in synovial fluid by metagenomic shotgun sequencing. J Clin Microbiol 2018, 56:e00402−e00418
60. Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, Loeb LA: Detection of ultra-rare mutations by next-generation sequencing. Proc Natl Acad Sci U S A 2012, 109:14508−14513
61. Duncavage EJ, Jacoby MA, Chang GS, Miller CA, Edwin N, Shao J, Elliott K, Robinson J, Abel H, Fulton RS, Fronick CC, O'Laughlin M, Heath SE, Brendel K, Saba R, Wartman LD, Christopher MJ, Pusic I, Welch JS, Uy GL, Link DC, DiPersio JF, Westervelt P, Ley TJ, Trinkaus K, Graubert TA, Walter MJ: Mutation clearance after transplantation for myelodysplastic syndrome. N Engl J Med 2018, 379:1028−1041
62. Gorzynski JE, Goenka SD, Shafin K, Jensen TD, Fisk DG, Grove ME, et al: Ultrarapid nanopore genome sequencing in a critical care setting. N Engl J Med 2022, 386:700−702
63. Wenger AM, Peluso P, Rowell WJ, Chang PC, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND, Töpfer A, Alonge M, Mahmoud M, Qian Y, Chin CS, Phillippy AM, Schatz MC, Myers G, DePristo MA, Ruan J, Marschall T, Sedlazeck FJ, Zook JM, Li H, Koren S, Carroll A, Rank DR, Hunkapiller MW: Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol 2019, 37:1155−1162
64. College of American Pathologists (CAP): Molecular Pathology Accreditation Checklist. Northfield, IL: College of American Pathologists Press, 2021
65. Wagner J, Olson ND, Harris L, McDaniel J, Cheng H, Fungtammasan A, et al: Curated variation benchmarks for challenging medically relevant autosomal genes. Nat Biotechnol 2022, 40:672−680
66. Aganezov S, Yan SM, Soto DC, Kirsche M, Zarate S, Avdeyev P, Taylor DJ, Shafin K, Shumate A, Xiao C, Wagner J, McDaniel J, Olson ND, Sauria MEG, Vollger MR, Rhie A, Meredith M, Martin S, Lee J, Koren S, Rosenfeld JA, Paten B, Layer R, Chin C-S, Sedlazeck FJ, Hansen NF, Miller DE, Phillippy AM, Miga KH, McCoy RC, Dennis MY, Zook JM, Schatz MC: A complete reference genome improves analysis of human genetic variation. Science 2022, 376:eabl3533