White Paper

The importance of expert curation in clinical NGS testing

High-touch, expert curation methods are essential to provide consistent and accurate biological and clinical variant interpretation

Introduction

Over the last decade, the field of molecular diagnostics has undergone tremendous transformation, catalyzed by the clinical implementation of next-generation sequencing (NGS). Today, commercial molecular diagnostic labs are competing to offer an ever-expanding menu of NGS tests. But there's an inherent problem: as NGS panels grow, test interpretation becomes more complex and time-consuming.

Throughout the industry, single gene tests and small gene panels are being replaced with large, comprehensive gene panels and even whole exome and genome sequencing. This shift is placing immense pressure on workflow efficiency.

When a molecular diagnostic lab moves from a 50-gene panel to a 300-gene panel, the number of variants generated in a single test increases dramatically. And, as the range of testing options soars, most molecular diagnostic labs no longer have genetic scientists with expertise in every test indication. Under these conditions, variant interpretation is the bottleneck delaying the uptake of NGS in routine clinical practice and patient management.

The accelerated development of complex or larger NGS panels have made this a challenging time for oncologists to effectively incorporate test results into routine patient care. Keeping up with new data on the clinical utility of each test, as well as the increasing availability of evidence-based clinical guidelines and primary literature beyond some of the most common cancer types, like lung cancer, is difficult to accomplish. Adding to the pressure, oncologists have less time than ever to glean what is most relevant and useful from molecular diagnostic reports and often find it challenging to interpret NGS test results for their clinical use in the context of the patient's history.

Therefore, it falls on the molecular diagnostic labs to do the heavy lifting to filter, prioritize, and gather critical information on detected variants in a matter of days or even hours and summarize the data in a concise report, highlighting clinically relevant information backed up by scientific evidence.

What information do oncologists need in a molecular diagnostic report?

While molecular diagnostic reports differ by cancer indication, test ordered, and individual lab offerings, oncologists generally look for answers to the following questions:

- Does the patient have a variant or biomarker present that affects diagnosis or prognosis?
- What is the level of evidence supporting the detected variants' diagnostic and prognostic significance as it applies to professional guidelines (NCCN®, WHO, ELN etc.)?
- Does the patient have a variant or biomarker that can be targeted with approved therapies or clinical trials?
- What is the level of evidence supporting the detected variants' therapeutic implications? (For example, is the "matched" therapy an approved drug in the patient's disease, an off-label drug, or investigational drug in a clinical trial?)
- Does the patient have co-occurring variants with contradictory outcomes or drug interactions?
- If a novel gene or variant of unknown significance is detected, what information exists to help better understand its role in disease?
- What is the scientific and medical literature supporting the biological and clinical evidence and can this be transparently reviewed and verified?

For a molecular diagnostic lab to be able to answer each of these questions, there is significant overhead in acquiring, analyzing, managing, and updating an enormous amount of clinical and genomic knowledge. Newly described or uncommonly mutated genes will inevitably reveal novel variants, and each variant that is reported must be assessed to determine whether it has a known pathogenic role, or if it is a variant of unknown significance.

▷

In 2019, more than 1 million medical papers were published [1]. This equates to approximately a new publication every 30 seconds. The rapidly changing nature of science poses a formidable challenge. There are scores of new research articles appearing daily, specific to particular variants, and it is often not feasible for a molecular diagnostic lab to keep abreast of every new publication as they become available. Additionally, necessary information about genes and variants is dispersed across many different databases, requiring substantial time to acquire, aggregate, and analyze the information with semantic consistency. Interpretations of variants based on outdated or missing aggregated knowledge increases the risk of variant misclassification.

The rise of clinical decision support systems

Clinical decision support systems, or CDSS, represent a paradigm shift in molecular diagnostic testing. Used to augment lab directors in their complex decision-making processes, CDSS translate NGS variants into biological and actionable clinical insights using curated evidence and computerized algorithms from a centralized knowledgebase.

However, a CDSS, which can be offered as a software or service, is only as good as its underlying curated evidence. Neither molecular diagnostic labs nor oncologists or geneticists who care for patients can afford to rely on curated evidence that is outdated, incomplete, unproven, sluggish, inflexible, or prone to gaps. Too much is at stake.

The quality of a knowledge base comes down to how it is curated, structured, and maintained.

Automated vs. manual curation methods

Data curation is the process of turning independently created data sources into structured and semantic-consistent unified data sets ready for analytics, using domain experts to guide the process. The curation process encompasses identifying, acquiring, storing, prioritizing, analyzing, and transforming data. Moreover, the process of clinical and phenotypic data curation is a laborious, intellectually intensive activity that demands deep subject matter expertise and significant personnel resources. Using the most appropriate method at each step in the curation process is critical for operating at scale and coping with the ever-growing body of literature and databases. Combining human aspects and machinable methods should be used to their advantages when appropriate. Whereas a human can decipher deep unstructured content with high granularity and precision, machinable methods can take over data conversions and repetitive tasks and can structure content with semantic consistency and build relational models.

For a molecular diagnostic lab to be able to effectively adapt different methods during a curation process, a team of ontologists, data scientists, and scientific domain experts is needed. However, assembling such a team incurs high labor costs and pricey subscription fees to journals and databases.

The process of identifying and quickly prioritizing relevant publications for curation is commonly carried out by querying articles in PubMed and manually screening the content for applicability. Machine learning is proving to be a valuable approach to enhance and accelerate this process. Machine learning-based approaches provide an improvement over rule-based approaches as it is more amenable to new projects. Training sets allow the system to be taught to find articles containing relevant information.

Using natural language processing and machine learning approaches further facilitates “seeding” relationships from articles in order to describe phenotypic and genotypic relationships. These methods allot the human curators and domain experts more time to focus on challenging tasks, such as the deeper curation required to build highly precise and organized information relationships that can be provided to users.

Manual curation remains the cornerstone to review deep, unstructured biological, phenotypic, and complex clinical outcome data including graphics, full text, and supplementary material. Human judgment remains critical for the analysis and capture of complex relationships, interactions, and contradictory evidence. For clinical decision support tools, the high-touch human review process ensures high accuracy, high specificity, relevance, context, and consistency.

Different approaches to manual curation

Various approaches exist for manually analyzing and capturing content to keep a knowledge base dynamically active and useful. We will discuss two approaches performed by QIAGEN Digital Insights:

- **Transformation of unstructured data into computable units** – Manually curates and models scientific literature and professional guidelines with semantic consistency and captures biological, phenotypic, therapeutic, and outcomes data into machine readable formats to allow computing of content which can be contextualized based on the provided metadata.
- **Topic-based curation resulting in expert-written summaries** – searches for information aligning to a topic that is selected, synthesized and summarized – all conducted by a domain expert exercising human judgment.

A molecular diagnostic lab can use either approach or a combination of the two based on their own needs and the needs of their customers. However, in both approaches, continuous manual curation is required to ensure that molecular diagnostic labs are provided with the latest information to interpret NGS variants.

Transformation of unstructured data into computable units

This approach places human curation and judgment upstream of well-tuned algorithms for computing relevant content in context of metadata, such as a provided diagnosis. The curation team is trained on a highly detailed and specific process to read articles and extract the data, converting unstructured information into structured data units that are stored in the knowledge base. Once the data has been structured, it can be subjected to computation to derive information about submitted variants. The advantages are:

- Every variant in every gene across the genome can be classified in the context of the disease or phenotype since they are assessed against the computable data in the knowledge base.
    - Contextualized content can be displayed in a CDSS software for every variant across the genome to gain insights into the biological effect of a variant, clinical cases, relevant clinical insights, such as prognostic and diagnostic relevant information, biomarker-driven treatments, and clinical trials.
- The rules that determine the variant classification are transparent in the application.
- The user can transparently review the referenced information supporting the classification and revise the classification, if appropriate.

Overall, this approach eliminates the need for researching, retrieving, and analyzing information, saving a molecular diagnostic lab a significant amount of time and money.

Topic-based curation resulting in expertly written summaries

The topic-based approach allows domain expert scientists to apply expert judgment at the time of classification. Literature surrounding each topic is collected, and the curators analyze the literature and create summaries that can be used by the molecular diagnostic labs.

- Rather than presenting the full collection of literature for a given alteration, the most relevant literature is collected, analyzed, and summarized.
- The user receives fully referenced, easy-to-read summaries that describe the impact of the alteration in the context of the specific disease, essentially eliminating the need to analyze the literature and write the summaries themselves.
- The summaries are written in a regular, structured format, so that the user always knows where to look for each type of information (e.g. molecular impact, incidence, prognostic relevance, etc).
- A software can transparently display the expertly written variant- and disease-specific summaries at time of variant classification.

In the topic-based approach, the selection of genes and cancer types to curate is organic, but recent developments in this approach have allowed curation teams to identify genes and cancer types submitted most frequently so that these may be prioritized for regular updates.

Combining the two approaches

The two approaches — computing of aggregated content for each patient case and expertly written summaries — are highly complementary to each other. The combination of these two methods is particularly valuable when analyzing the sequencing results from large NGS panels. A molecular diagnostic lab can rely on the prioritization and selection of clinically or biologically relevant variants based on computable content and then use a topic-based approach to gain a deeper understanding on the biological or clinical relevance with expertly written summaries. Moving forward, the combination of both approaches will be extremely beneficial to molecular diagnostic labs as they expand test menus and grow caseload volume.

QIAGEN's advanced curation processes ensure complete and up-to-date biological and clinical relevance

QIAGEN Digital Insights is the leading provider of genomic content knowledge, based on advanced curation methods to ensure relevance and accuracy of insights. The structured content is used in various aspects in the QIAGEN Clinical Insights (QCI®) portfolio that offers three different clinical decision solutions: QCI Interpret, QCI Precision Insights, and QCI Interpret One.

QIAGEN's software solutions, QCI Interpret and QCI Interpret One, are powered by the QIAGEN Knowledge Base, which stores computable structured units of information composed of proprietary, open source, and licensed biological and clinical content that has been aggregated, integrated, and curated for biomedical relevance. It has been used by researchers, clinicians, and pharmaceutical companies for more than 20 years and has been cited in more than 30,000 scientific publications. QCI Precision Insights is a professional clinical interpretation service powered by a world-class team of molecular biologists and oncologists who translate molecular data specific to each patient into state-of-the-art clinical insights. With over a decade of content aggregation, curation, and interpretation of somatic variants for oncology, QCI Precision Insights has interpreted more than 200,000 cases with coverage of over 1,000 cancer-related genes and 200,000 unique variants across more than 1,000 different cancer subtypes. The same professional interpretation service can be accessed through the QCI Interpret One software.

QIAGEN's Knowledge Base is built with a manually curated ontology

The core data models in the QIAGEN's Knowledge Base are designed for flexibility, scale, accuracy of representation, and computability (algorithm-friendly), thus serving as models for a broad range of artificial intelligence and statistical computational approaches. Domain experts and ontologists create structured vocabulary systems called ontologies to define the data sources and types that are relevant for a biological impact, disease, or clinical use case. The QIAGEN Knowledge Base is built on QIAGEN's own comprehensive ontology, which is manually curated. Using an ontology helps to uniformly model relationships between different entities, such as the relationship between a variant, the gene that it resides in, and the observed phenotype.

QIAGEN's QCI products make use of the underlying ontology system and curated findings to identify new and existing relationships between identified variants, disease phenotypes, and biological processes. A dedicated team keeps the ontology up-to-date as new concepts are identified and described.

QCI Interpret	QCI Precision Insights	QCI Interpret One
Clinical decision support software for variant interpretation	Professional interpretation service for somatic variants in oncology	Clinical decision support software combined with professional interpretation service for somatic variants in oncology
Computing based on aggregated structured units for each patient case	Expert-written summaries based on selected topics	Computing based on aggregated structured units for each patient case together with expert-written summaries

QIAGEN's transformation of unstructured data at scale

For the transformation of unstructured data into computable units, QIAGEN Digital Insights employs more than 200 experts worldwide to develop curation protocols, curation methods, data models, quality control systems, and custom-developed curation tools that work with the QIAGEN Knowledge Base ontology, ensuring consistent representation of biomedical data and relationships. The curation tools allow manual extraction of biological and clinical insights from over 4,000 scientific journals, drug labels, clinical trials, and professional guidelines monthly. The extracted biological and clinical information is integrated into the QIAGEN Knowledge Base according to the ontologies.

To operate at such a large scale without compromising quality, QIAGEN Digital Insights has developed sophisticated and streamlined curation processes and protocols. Throughout the curation process QIAGEN Digital Insights employs both manual and computational methods to iteratively aggregate, prioritize, and review data sources and compile curated findings under a rigorous quality management system.

All domain expert curators have a MD or PhD background with experience in both genetics and clinical research. Curators go through several months of training, with proficiency testing prior to entry of new variant findings and gene annotations into the QIAGEN Knowledge Base. During this training period, all their work is reviewed until it is comparable to the work of an experienced curator, at which time the new curator becomes certified.

The quality control review of the curated content in the QIAGEN Knowledge Base consists of both extensive automated testing and manual spot-checking of scientific content. First, manual scientific spot-checking is used to verify fidelity of curation, allowing curators to improve with feedback. Second, automated quality control (QC) testing is performed to identify any content logic discrepancies and systematic errors during and after the curation process. Finally, end-to-end testing in the QCI applications are performed to ensure the curated content is represented as intended in the application.

Using a database founded on manual curation of complex biological and clinical evidence provides high accuracy of variant classifications. By investing the time and resources into the QIAGEN Knowledge Base, QIAGEN Digital Insights has made it possible for molecular diagnostic labs to find disease-causing variants faster, gain deeper insights into diagnostic, prognostic, or therapeutic actionability, and significantly reduce the time needed to interpret genomic variants and other data without compromising quality. It is precisely this differentiator—QIAGEN Digital Insights' team of expert curation—that makes the QIAGEN Knowledge Base a trusted resource in modern genetic testing.

The QIAGEN Knowledge Base in QCI Interpret

QCI Interpret uses sophisticated algorithms to compute the structured clinical evidence and disease knowledge in the QIAGEN Knowledge Base into scientific and clinically relevant content within based on the uploaded genomic profile, phenotypes, or diagnosis, age, and gender. At scale, the software displays the downstream biological effect of each variant, its possible corresponding impact on disease physiology, a differential clinical diagnosis, and potential sensitivity or resistance to an array of available therapeutic options and matched clinical trials. All of these steps are automated, running rapidly in the background, and the eventual output provides a detailed explanation for the algorithmic reasoning that led to the given conclusion.

In addition, QCI Interpret provides a computed classification based on ACMG and AMP/ASCO/CAP criteria. QCI Interpret respects the role of human judgment and experience. Any tool that incorporates genomic guidelines should also emphasize transparency. QCI Interpret allows users to click through to relevant source material and report information based on their own expertise in cases where they may disagree with a variant assessment or the supporting clinical evidence. The transparent scientific evidence review makes QCI Interpret a true CDSS; it does not replace traditional variant interpretation and reporting with "black box" automation, but rather gives clinicians the resources needed to make consistent and informed decisions.

QIAGEN's curation methods resulting in expert-written summaries

QIAGEN Digital Insights builds upon a decade of domain expertise in professional interpretation services for users, allowing them to outsource the literature research, interpretation, and variant classifications, as well as the write-ups of expert oncologist-reviewed summaries to accelerate test turnaround time. The on-demand curation services, such as the pre-curation of content for variants for a targeted gene, allow labs to further accelerate their interpretation of variants.

The PhD scientists are trained to a common system and set of criteria and standards. The domain experts conduct the curation utilizing the topic-based approach, and all scientists undergo a lengthy training program to ensure consistency and quality. Any content generated by junior scientist is subject to review by a senior scientist, and all content is subject to a regular quality control process, whereby senior scientists peer-review a subset of reports that have been handled by other curators on a weekly basis.

The curators analyze the literature and create summaries along with their references, storing the information in a hierarchical manner. The hierarchical method of storing information minimizes redundancy in the database and enhances the efficiency of the system, allowing maximum reuse of data and scalability. Information is added to the database daily, based on both on-demand reporting and managed pre-curation. The workflow system assesses the currency of the curation for all variants and gene-disease contexts, automatically queuing tasks for scientists. Most variants and gene-disease relationships are reviewed and updated on at least an annual basis. New drug approvals are incorporated within a business day, and NCCN treatment guideline updates are incorporated within a week.

The report content is highly standardized, using distinct sections that guide the oncologist to the relevant section of interest; customers know where to look for particular types of information and have confidence that the literature was analyzed in a rigorous and standardized manner. This type of rigor is difficult to achieve with crowdsourced data. Even with renowned experts submitting information and analysis, the format and level of information is nearly impossible to standardize without a cohesive team, working from a common protocol. A detailed set of reporting logic governs the report content assembly and includes handling of interactions among variants within a case that may affect drug sensitivity, drug resistance, prognosis, or diagnosis.

The hierarchical storage of the summaries allows very specific data to be assembled for a variant or variant type (e.g. FLT3 ITDs) in the context of a disease. General information about the role of the gene or incidence of mutation types in a gene can be applied more broadly for all variants or variant types across the same cancer type.

The PhD scientists use the summarized evidence to classify actionability based on the 2017 AMP/ASCO/CAP guidelines for standardizing the interpretation and reporting of sequence variants in cancer. Since this is manually performed, classification is determined via expert judgment, which is particularly valuable when classifying emerging biomarkers with limited or conflicting evidence.

Then, practicing oncologists review all clinical content before the results are returned to the customer. This allows for the delivery of concise oncologist-reviewed interpretation summaries for each biomarker in the context of the cancer sub-type, providing information on the mutation's molecular characteristics, role in disease, and therapeutic, prognostic, and diagnostic implications with applicable references.

QIAGEN's expert-written summaries fuels QCI's professional interpretation services

QIAGEN's professional interpretation service, QCI Precision Insights, delivers report content based on expert-written summaries of the latest biological, diagnostic, prognostic, and therapeutic evidence in context of the tumor profile, treatment, and country-specific clinical trials. For rare or novel variants, QCI Precision Insights' variant scientist team perform the in-depth research, curation and interpretation, thus eliminating the time-consuming step for molecular diagnostic labs of writing up variant and disease specific comments and assembling final report content.

The power of aggregated knowledge combining computation and human expert judgement

QCI Interpret One combines the advantages of computing aggregated content for each patient case and expert-written summaries. Using sophisticated algorithms to compute over the structured clinical evidence and disease knowledge in the QIAGEN Knowledge Base allows for dynamic prioritization and automatic classification of all variants in the patient's molecular profile at scale. QCI Interpret One has an in-software option to submit clinically and biologically relevant variants to QIAGEN's professional interpretation services, who apply the topic-based approach to research and summarize the biological and clinical significance, which allows users to gain a deeper understanding into the characteristics and clinical relevance of a mutation.

An advantage of applying both curation methods is the confidence in classifying variants as variant of unknown significance (VUS). After initially computing a variant as a VUS based on the content in the QIAGEN Knowledge Base, a domain expert subsequently searches for evidence using the topic-based approach. The manual verification of a VUS provides a user higher confidence in distinguishing a true VUS from a potentially biologically or clinically relevant biomarker

The hybrid curation approach further allows the scalable interpretation of rare or novel variants. QCI Interpret One's on-demand clinical curation and interpretation services does the research, curation, and interpretation for users, replacing labor-intensive processes with automated simplicity. QIAGEN's professional interpretation services provide a "second set of human eyes" on the otherwise computed variant classification, allowing labs to be confident and more accurate in their variant classifications.

Conclusion

Interpreting genetic variants at scale continues to challenge evidence-based medicine. To overcome the major bottleneck of accurately interpreting an individual's genetic variants from larger panels and even whole exome and genomes requires sophisticated curation methods and processes to find, prioritize, transform, and constantly update biologically and clinically relevant publications at scale.

QIAGEN Digital Insights has unparalleled experience in content curation. As the leading provider of genomic content knowledge, QIAGEN's variant interpretation software and service take advantage of different curation methods to accurately transform the literature into biological and clinical insights. Ultimately, the aggregated knowledge ensures users receive timely, accurate, reproducible, and consistent content to confidently interpret variants at scale and support evidence-based medicine.

References:

1. Landhuis, E. Scientific Literature: Information Overload. Nature. 2016 July **21**;7612:457-58.

Learn more about QCI and the QIAGEN Knowledge Base at
www.digitalinsights.qiagen.com/products-overview/clinical-insights-portfolio/

QCI Interpret and QCI Interpret One are evidence-based decision support software intended as an aid in the interpretation of variants observed in genomic next-generation sequencing data. The software evaluates genomic variants in the context of published biomedical literature, professional association guidelines, publicly available databases, annotations, drug labels, and clinical trials. Based on this evaluation, the software proposes a classification and bibliographic references to aid in the interpretation of observed variants. The software is NOT intended as a primary diagnostic tool by physicians or to be used as a substitute for professional healthcare advice. Each laboratory is responsible for ensuring compliance with applicable international, national, and local clinical laboratory regulations and other specific accreditations requirements.

QCI Precision Insights does not provide medical services, nor is any QIAGEN employee engaged in the practice of medicine for or on behalf of QIAGEN. QCI Precision Insights report content is for professional medical and scientific use only.

Ordering and Technical Support **bioinformaticssales@qiagen.com**
Website **digitalinsights.qiagen.com/**