tetrascience

Solution Brief

NGS Data Management

Next-generation sequencing (NGS) has revolutionized genomics, unlocking unprecedented insights into biological systems. Yet, extracting the full value from NGS data remains a challenge. Scientists must navigate fragmented data storage, inconsistent metadata, and inefficient workflows—barriers that slow discovery, hinder collaboration, and limit analytical depth.

To maximize the potential of NGS, organizations need a scalable solution that transforms raw sequencing data into actionable insights. The Tetra Scientific Data and AI Cloud™ streamlines NGS workflows by automating metadata management, centralizing data access, and enhancing quality control. By integrating seamlessly with existing informatics platforms and advanced analytics tools, TetraScience empowers scientists to move beyond data wrangling and focus on scientific breakthroughs.

This solution brief examines the key challenges of NGS data management and demonstrates how TetraScience provides a more efficient, scalable, and insight-driven approach to genomics.

The Challenge

From collecting data to generating insights, NGS users need to navigate a series of obstacles to analyze their data. These challenges fall into three main categories:

1. Metadata Management and Context

The importance of metadata in genomics cannot be overstated. Metadata provides crucial context for the raw genomic data, enabling its interpretation and reuse. However, scientists face persistent hurdles:

- **Manual metadata association:** Linking metadata with instrument and analysis data is time-consuming and prone to errors.
- **Difficult data retrieval:** Without proper contextual information, locating specific datasets becomes inefficient and cumbersome.
- **Inconsistent metadata:** Inconsistencies in metadata formatting and structure make it difficult to aggregate and compare results across multiple experiments.

2. Data Access and Quality

Organizations often struggle to access and fully leverage their NGS data due to systemic inefficiencies:

- **Siloed data storage:** NGS data is fragmented across multiple isolated systems, limiting accessibility and collaboration. This also prevents scientists from associating NGS data with outputs from other instruments.
- **Cumbersome data transfers:** Manual data movement between platforms is inefficient, increasing the risk of errors and inconsistencies.
- **Limited integration with modern tools:** Many platforms lack seamless connectivity with advanced analytics and AI applications, restricting the potential for deeper insights and automation.

3. Quality Control and Performance Monitoring

Ensuring high-quality sequencing data requires comprehensive quality control (QC) measures, but several obstacles stand in the way:

- **Disconnected QC processes:** Variant calling results are often difficult to link with QC and supply chain data, leading to incomplete assessments.
- **Lack of instrument performance tracking:** Organizations struggle to monitor instrument parameters and their impact on sequencing accuracy over time.
- **Absence of historical benchmarking:** Without access to aggregated historical data, establishing consistent quality standards and detecting long-term trends remains a challenge.

The Solution

TetraScience streamlines the entire NGS data workflow, enabling scientists and data scientists to spend less time managing data and more time making discoveries (Figure 1). The Tetra Scientific Data and AI Cloud automates tedious tasks, provides seamless access to high-quality data, and enables deeper analytical insights.

1. Context for Tracking and Retrieval

- **Automated labeling and linking:** All files associated with an experiment are automatically labeled and linked, eliminating the need for manual annotation and reducing errors.
- **Intuitive identifiers:** Input and output files are named using clear, human-readable identifiers, streamlining data tracking and retrieval.
- **Metadata-driven searchability:** Metadata can be leveraged to locate relevant files quickly and aggregate results across multiple experiments, significantly improving research efficiency.

2. Data at Your Fingertips

- **Unified data access:** Centralized access to raw data, metadata, and analysis results ensures that all scientists can retrieve and use the information they need without delays.
- **Seamless third-party compatibility:** Data is made available to third-party analysis and AI applications via API.
- **Cross-instrument analysis:** Data from different sequencing instruments can be analyzed together, enabling researchers to identify trends, optimize workflows, and derive deeper insights.

3. Actionable Insights for Quality

- **Automated quality control:** New sequencing data is continuously compared against historical datasets, ensuring consistency and accuracy.
- **Real-time benchmarking:** Instant quality assessments against predefined standards allow for early issue detection and timely corrective action.
- **Long-term performance monitoring:** Continuous tracking of instrument trends helps identify gradual performance shifts, anticipate maintenance needs, and prevent costly failures.

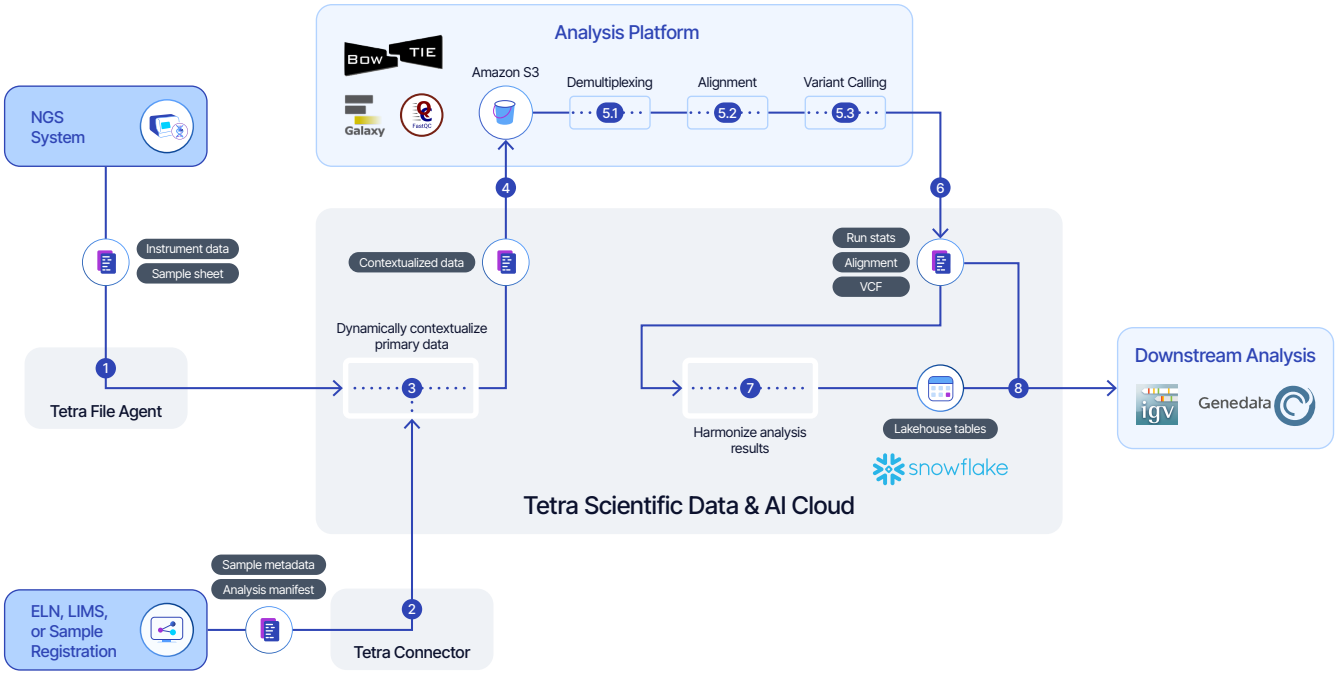

Figure 1. Tetra workflow for NGS data management.

1. **Data ingestion:** Instrument data (FASTQ files), run metadata, sequencing statistics, and sample sheets are ingested into the Tetra Scientific Data and AI Cloud via the Tetra File Agent.
2. **Metadata collection:** Sample and analysis information is retrieved from the ELN or LIMS through a Tetra Connector.
3. **Contextualization:** A pipeline enriches the primary NGS data with experiment-specific metadata, such as sample name and run ID.
4. **Data transfer:** The contextualized data is uploaded to the customer's S3 bucket, ensuring accessibility from EC2 or other Linux-based systems.
5. **Analysis:** The data is processed using the customer's preferred software (e.g., Bowtie, Galaxy, FastQC, DIY pipelines).
6. **Results transfer:** Alignment files (BAM/SAM), run stats, and variant data (VCF) are transferred to the Tetra Scientific Data and AI Cloud.
7. **Data harmonization:** Analysis results are harmonized using an NGS data schema (see Figure 2) and aggregated into lakehouse tables, accessible via Snowflake data sharing.
8. **Data utilization:** Third-party applications (e.g., IGV, Genedata) consume the lakehouse tables (or individual analysis files), which can be combined with data from other instruments for further insights.

Common Sequencing Ontology	system_name	system
	instrument_model	
	flow_cell_type	flow_cell
	flow_cell_configuration	
	system_id	run
	flow_cell_id	
	run_datetime	
	read_length_configuration	
	run_mode	
	percent_phix	
	percent_q30	
	cluster_density_target	
	cluster_density	
	run_id	channel
	signal_intensity	
	percent_accuracy_per_cycle	
	percent_phasing	
	percent_pre_phasing	
	sample_name	sample
	sample_barcode	
	sample_organism	
	sample_tissue	
	library_type	
	library_preparation_kit	
	input_amount	
	sample_extraction_date	
Use Case Ontology	sample_id_list[]	analysis
	alignment_tool	
	reference_genome	
	annotation_database	
	sample_id	variant
	run_id	
	analysis_id	
	chromosome_id	
	position	
	strand	
	reference_allele	
	alternative_allele	
	read_depth	

Figure 2. Example ontology for NGS data. This particular ontology provides a structured framework for organizing and analyzing NGS data to identify variants relative to a reference genome. It enables efficient searches across sample metadata, sequencing run parameters, variant characteristics, and mapping metrics.

Application Areas of the Solution

TetraScience accelerates and improves NGS workflows across the biopharmaceutical value chain, providing scientists with engineered sequencing data to unlock new insights and optimize processes.

Discovery and Research

With larger, richer, and more accessible sequencing datasets, researchers can uncover new patterns that drive advancements in biomedical science. Meta-analyses of rare genetic variants can lead to the identification of **novel drug targets and biomarkers**, accelerating the development of precision medicine approaches. Furthermore, **longitudinal studies of microbiome dynamics** offer a deeper understanding of disease progression, treatment efficacy, and the emergence of drug resistance. Comparative analyses of **drug response mechanisms** across diverse preclinical models refine translational medicine strategies, ultimately improving drug development and therapeutic outcomes.

Development

Ensuring the safety and efficacy of advanced therapeutics, such as cell and gene therapies, requires a comprehensive analysis of sequencing data throughout the product lifecycle. By analyzing sequence-based lot release data, organizations can identify **trends in product variability**, enabling more precise quality assessments and process optimization. Tracking mutations, recombination events, and molecular degradation in advanced therapies ensures their **long-term safety and efficacy**. For CRISPR-based products, NGS plays a critical role in meeting **regulatory safety requirements** by enabling the detection and quantification of potential off-target mutations, thus supporting the development of safer gene-editing therapeutics.

Manufacturing and Quality

Optimizing quality control in biopharmaceutical manufacturing is crucial for ensuring consistency and regulatory compliance in biologics production. By leveraging large-scale NGS data, manufacturers can minimize batch-to-batch variability and enhance **sequencing reliability** in Good Manufacturing Practice (GMP) environments. Regular monitoring of genomic stability in production cell lines helps maintain **consistent drug product quality**, minimizing deviations that could impact therapeutic efficacy. Moreover, advanced NGS analysis enables the **detection of contaminants** and adventitious agents in sequencing runs, mitigating the risk of batch failures and ensuring adherence to regulatory standards.

Conclusion

The Tetra Scientific Data and AI Cloud optimizes NGS data management to accelerate discoveries and ensure consistent, high-quality results. It minimizes manual data processing and errors, allowing scientists and data scientists to focus on higher-value work. Automated workflows and purpose-engineered datasets boost efficiency and drive deeper insights, while proactive quality control and instrument monitoring enhance data reliability. Centralized data management fosters seamless collaboration within and across teams, transforming NGS data into a powerful, accessible asset. With TetraScience, biopharma organizations can innovate with greater speed, confidence, and precision.

Ready to transform your NGS data management? Get Started