Mini Review

Sebastian H. Eck*

Challenges in data storage and data management in a clinical diagnostic setting

https://doi.org/10.1515/labmed-2018-0310
Received May 14, 2018; accepted July 5, 2018; previously published online August 10, 2018

Abstract: The implementation of next-generation sequencing (NGS) in a clinical diagnostic setting opens vast opportunities through the ability to sequence all genes contributing to a certain morbidity simultaneously at a cost and speed that is superior to traditional sequencing approaches. On the other hand, the practical implementation of NGS in a clinical diagnostic setting involves a variety of new challenges, which need to be overcome. Among these are the generation, analysis and storage of unprecedented amounts of data, strict control of sequencing performance, validation of results, interpretation of detected variants and reporting. In the following sections, key aspects of data management and integration will be discussed. In particular, issues of data storage, data analysis using in-house IT infrastructure vs. data analysis employing cloud computing and the need for data integration from different sources will be covered.

Keywords: cloud computing; data integration; data management; next-generation sequencing (NGS).

Genetic testing volume

The rise of next-generation sequencing (NGS) methods can be considered as a disruptive technology for genetic testing [1, 2]. The volume of genetic data generated per test rapidly increased. Prior to NGS, the diagnostics of the majority of rare disorders followed a step-wise paradigm where different genes involved in the disorder were sequenced sequentially. NGS enables larger approaches like the sequencing of gene panels (i.e. all genes known to contribute to certain disorders are sequenced simultaneously, typically 5–100 genes) and the sequencing of the clinical exome (i.e. all genes known to be involved in any Mendelian disorder, 2500–6000 genes). Even the sequencing of all protein-coding regions (i.e. whole exome, ~20,000 genes) or the complete genome is possible (Figure 1). Consequently, the amount of data that needs to be pre-processed, analyzed, interpreted and stored per case also dramatically increased (Table 1). There are now several guidelines for the use and quality control of NGS in clinical diagnostics [3–6], yet the issue of efficient data handling, integration and long-term storage are more rarely covered.

Long-term handling and storage of NGS data

The huge amount of raw and processed data from NGS warrants new strategies for the long-term storage and accessibility of these data [4]. There are three possible approaches:

1. Complete storage of raw data

In this approach, all raw and processed files from the sequencing instrument (Figure 2) and the analysis pipeline are stored in long-term storage. This includes raw base call files (.bcf) or their equivalent for non-Illumina instruments used to generate .fastq files as well as the aligned sequences, preferably in the binary sequence alignment map format (.bam, [7]) and the complete output of the analysis pipeline. This may consist of the detected variants (variant call format, .vcf), quality metrics, coverage analysis, annotations and all parameter settings used to obtain these results. The advantage of this approach is the complete transparency and reproducibility of the results, should circumstances warrant a re-inspection of a particular case. On the other hand, this approach poses a huge burden on the backup and data storage pipeline due to the amount of data and that some data are saved partially redundantly. For example, .bam files may be generated using .fastq files, which in turn can be generated using .bcf files. To reduce the redundancy while maintaining the absolute level of reproducibility, the software versions and settings used when the case was first analyzed need to be saved as well.

*Correspondence: Sebastian H. Eck, Center for Human Genetics and Laboratory Medicine Dr. Klein, Dr. Rost and Collegues, Lochamer Strasse 29, 82152 Martinsried, Germany, E-Mail: sebastian.eck@medizinische-genetik.de
2. Storage of data needed to repeat analysis

The second approach tries to omit saving data directly from the instruments, and instead focuses on all data needed to repeat the analysis. This typically means that the .fastq files and/or .bam files along all parameter settings and pipeline components are archived, while primary data from the instruments like .bcf files are not. This approach relaxes the conditions of storage requirements somewhat, while still maintaining the ability to repeat all primary analysis steps if necessary. The only part that is not reproducible is the actual base calling, i.e. the translation of instrument output into continuous reads.

3. Storage of results and parameter settings

In the last approach, only processed results like detected variants and the parameter setting with which they were obtained are stored, all intermediate files used to generate the results are deleted after analysis. This includes .fastq and .bam files. As a consequence, results cannot be easily reproduced. Instead, the sample has to be re-sequenced in an additional instrument run, potentially requiring new primary material ethylenediaminetetraacetic acid (EDTA) blood. While this sounds like a disadvantageous strategy, it could be more attractive for large-scale data sets like whole genome and exome data. Due to further declining sequencing cost, the financial burden of storage space for all samples can outweigh the cost for re-sequencing of few selected cases where re-analysis is necessary.

All approaches share the requirement of recording pipeline versions and parameter settings of how the results were obtained.

### Cloud vs. in-house data analysis

Cloud computing, by definition, is a way to organize computing and storage resources so that multiple users may rent certain amounts of CPU hours and storage according to their specific need [8]. Since the inception, cloud computing concepts have extended to additionally provide platforms, i.e. pre-configured computing environments, as well as already installed and ready-to-use software. Major cloud computing providers include Amazon, Google and Microsoft, but also data resources and large scale data sets such as GenBank, the 1000 Genomes project [9] and the Exome Aggregation Consortium data (ExAC [10]) are

---

**Table 1:** Output and approximate required disk space to store an instrument run (run folder size, source Illumina bulletin 01/2018).

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Run type</th>
<th>Output Gigabases</th>
<th>Run folder size Gigabytes</th>
</tr>
</thead>
<tbody>
<tr>
<td>MiSeq</td>
<td>2 × 150 bp</td>
<td>5</td>
<td>16–18</td>
</tr>
<tr>
<td>MiSeq</td>
<td>2 × 250 bp</td>
<td>8</td>
<td>22–24</td>
</tr>
<tr>
<td>MiSeq</td>
<td>2 × 300 bp</td>
<td>15</td>
<td>22–26</td>
</tr>
<tr>
<td>NextSeq500</td>
<td>2 × 150 bp Mid output</td>
<td>35</td>
<td>22–26</td>
</tr>
<tr>
<td>NextSeq500</td>
<td>2 × 150 bp High output</td>
<td>120</td>
<td>60–70</td>
</tr>
<tr>
<td>HiSeq2500</td>
<td>2 × 250 bp Rapid run</td>
<td>150</td>
<td>80–100</td>
</tr>
<tr>
<td>HiSeq2500</td>
<td>2 × 125 bp High output</td>
<td>500</td>
<td>295–310</td>
</tr>
<tr>
<td>NovaSeq</td>
<td>2 × 150 bp S2 flowcell</td>
<td>1000</td>
<td>730</td>
</tr>
<tr>
<td>NovaSeq</td>
<td>2 × 150 bp S4 flowcell</td>
<td>2500</td>
<td>2190</td>
</tr>
</tbody>
</table>

---

**Figure 1:** Volume of different genetic tests, size in base pairs (bp).

**Figure 2:** Illumina raw data folder structure (Illumina bcl2fastq guide v2, page 6).
stored in clouds. Different cloud concepts may be classified by services that are offered; broad categories are Data-as-a-Service, Infrastructure-as-a-Service, Software-as-a-Service and Platform-as-a-Service. An overview with example providers for each category and possible services can be seen in Table 2. The defining feature of cloud computing models is “elasticity”, which in this context describes the ability to access exactly the amounts of resources needed for a specific task. Two typical problems when considering investments in IT infrastructure are under-provisioning, i.e. the lack of computational power and/or storage to efficiently handle a task, and problems of over-provisioning, i.e. mostly idle resources not needed routinely. These problems may be solved by employing cloud solutions as only the resources specifically needed are rented and paid for.

In genomics research, cloud computing solutions are not only used for primary analysis in large-scale projects like the 1000 Genomes Project [9], the Exome Aggregation Consortium [10] or the Cancer Genome Atlas [11], they also enable global access to these data sets. This allows re-analysis of the data which can assure reproducibility and exploitation for different projects by, for example, providing reference data sets.

The main reasoning for choosing a cloud computing approach is to reduce the burden of storage and computing requirements that need to be maintained in-house, and to have access to a sufficient resource when needed.

On the other hand, maintaining an in-house server structure for primary data analysis has the advantage of having full control over the computing systems. New software may easily be deployed and utilized, which may be difficult depending on the chosen cloud computing model. Another advantage is that all data remain in an in-house network and no large data sets need to be transferred, reducing the cost and requirements for the cable connection. This comes at the price of having to install, update and maintain the in-house servers. This approach is also not readily scalable so that an increase in sequencing capacity often warrants new investments in the IT infrastructure for storage and analysis.

To summarize, both approaches have their respective advantages and caveats. Smaller scale data sets like primarily multi-gene panel sequencing can be efficiently handled on in-house servers. The capital investment for the necessary IT infrastructure is manageable and all data are generated, processed and stored in the same place. On the other hand, cloud computing solutions are more suited for the analysis of larger scale data sets, ranging from primarily exome sequencing to whole genome sequencing data analysis. As cloud solutions are readily scalable, the risk of wrong IT infrastructure investments is eliminated.

### Data analysis and integration

Efficient data handling goes far beyond the problems of which data need to be stored. In state-of-the-art diagnostics, the integration of different data sources, both external and internal, from different tests, workflows and clinical specialties are vital for optimal patient care. In the following paragraphs two examples of integrated diagnostics will be discussed.

The first example is the diagnostic routing for non-small cell lung carcinoma (NSCLC, Figure 3) where data from different tests and the cooperation between different clinical specialties are combined for an optimal performance [12]. The workflow is as follows:

<table>
<thead>
<tr>
<th>Cloud service</th>
<th>Acronym</th>
<th>Example services</th>
<th>Example providers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data-as-a-Service</td>
<td>DaaS</td>
<td>Access to large data sets</td>
<td>GenBank, ExaC, 1000 Genomes</td>
</tr>
<tr>
<td>Infrastructure-as-a-Service</td>
<td>IaaS</td>
<td>Virtualized, computational infrastructure (compute power and/or storage)</td>
<td>Amazon Web Service AWS, Amazon EC2, Amazon S3, Google Cloud Platform, IBM Cloud</td>
</tr>
<tr>
<td>Software-as-a-Service</td>
<td>SaaS</td>
<td>Software applications, open source analysis pipelines</td>
<td>Illumina Base Space, DNAnexus</td>
</tr>
<tr>
<td>Platform-as-a-Service</td>
<td>PaaS</td>
<td>Programming language environments, web servers, databases</td>
<td>Amazon Web Services Software Development Kit (AWS SDK), Google App Engine</td>
</tr>
</tbody>
</table>
In pathology, tumor diagnosis and classification of the tumor tissue is performed on the basis of the primary or re-biopsy material. After morphological examination, multi-gene panel sequencing is performed to identify driver and passenger mutations. Based on the results, a molecular profile is created with the aim of identifying therapeutic options. The molecular profile can be used in a follow-up by liquid biopsy or liquid profiling from peripheral blood. Targeted ctDNA analysis is performed for the monitoring of minimal residual disease or therapy resistance. Tumor markers complement and increase the sensitivity of the follow-up. In case of clinical deterioration, therapy/drug resistance or conspicuous imaging results, a re-biopsy with a morphological examination and molecular tumor profile is recommended.

The second example constitutes the integration of systematic phenotype information in the diagnostics of rare Mendelian disorders, in particular unexplained intellectual disability and developmental delay (ID/DD) [13, 14]. Rare disorders often manifest during infancy or early childhood, yet are difficult to diagnose due to high phenotypic variability and masking by auxiliary symptoms. The majority of rare disease patients remain without diagnosis or have to endure a diagnostic odyssey before the diagnosis could be obtained. Current routine diagnostic testing for patients with ID/DD of unclear etiology consists of karyotyping and array comparative genomic hybridization (array CGH) analysis. However, the major cause of severe ID/DD are de novo point mutations [15], which cannot be detected by these genome-wide low-resolution tests. A further approach in routine diagnostics is targeted sequencing of single genes with disease association, which is best suited for highly distinct clinical conditions that are caused by just one or a few genes. Many rare pediatric conditions including neurodevelopmental disorders are characterized by a large genetic heterogeneity, so testing each candidate gene individually is not feasible, whereas sequencing all genes simultaneously, such as in a whole exome sequencing (WES) approach, enables a fast and comprehensive genetic analysis [16]. Moreover, trio analysis of the WES data of the affected child and both unaffected parents enables benign familial genetic variants to be filtered out and de novo variants that are only present in the child to be identified.

By broad NGS-based diagnostic approaches like WES, the amount of clinical data increases tremendously, which necessitates the integration of phenotype and genotype data for fast and precise data interpretation and generation of conclusive genetic reports. To record phenotypic information in a standardized way, the Human Phenotype Ontology (HPO, [17–19]) was developed and is now widely employed. The HPO provides a hierarchic system of standardized vocabulary of phenotypic abnormalities encountered in human diseases and their semantic relationships, which can be stored in a relational database for easy query. Patient’s clinical features can thereby be used for filtering and evaluating genetic variants as described in the Multiple Integration and Data Annotation Study (MIDAS, [20]), which aims to accelerate WES data analysis by automated variant prioritization based on the patients’ phenotype. Besides, it would be beneficial for diagnostic purposes to integrate and correlate NGS data with additional genetic data or further resources like results from biochemical

Figure 3: Diagnostic routing and interplay of different tests in the diagnosis of NSCLC.
testing or medical imaging to support or exclude certain diagnoses (conf. MIDAS concept, Figure 4).

These examples show that strategies for long-term data storage, data handling and integration need to be developed, as these points are crucial for the success of future genetic testing in a clinical laboratory. The ever-expanding amounts of data generated per test warrant a more and more sophisticated approach to the data to extract the most useful information and exploit synergies of different data sets and clinical specialties by means of integrative diagnostic concepts.

**Acknowledgments:** The author would like to thank Oliver Wachter for the input to the NSCLC workflow and Yasemin Dinçer for the input to the integration of HPO terms and critical revision of the manuscript.

**Author contributions:** The author has accepted responsibility for the entire content of this submitted manuscript and approved submission.

**Research funding:** Bayerisches Staatsministerium für Wirtschaft, Infrastruktur, Verkehr und Technologie: Funder Id: 10.13039/501100005017, Grant Number: MED-1603-0001.

**Employment or leadership:** None declared.

**Honorarium:** None declared.

**Competing interests:** The funding organization(s) played no role in the study design; in the collection, analysis, and interpretation of data; in the writing of the report; or in the decision to submit the report for publication.

**References**