Micronbrane Medical

The Current Landscape of Clinical Bioinformatics

Bottlenecks and Alternatives

The current landscape of Clinical Bioinformatics – bottlenecks and alternatives

Next-generation sequencing (NGS) approaches are expanding from their niche applications in clinical settings, as they prove to be robust alternatives to biochemical, mass spectrometry-based or Sanger sequencing-based techniques of pathogen discovery. By enabling the parallel and high-throughput sequencing of large numbers of individual nucleic acids (typically 10^5 to 10^9 DNA/RNA molecules per sequencing run), NGS applications provide high-resolution genomic information that accounts for heterogeneous genetic traits and fast-tracks the in-house assembly of microbial genomes (Dullanto and Dekker, 2017). Sequencing costs in NGS-based pipelines represent about 80% of the total costs in the value-chain (Fig. 1), but these can be significantly optimized with an effective host DNA depletion method and with a processing pipeline that fast-tracks sample output.

In our previous White Paper, we showed how the products offered by Micronbrane Medical can streamline an affordability of up to 75% in sequencing costs and over 50% in the overall costs of NGS-based pathogen detection, while also significantly reducing the turn-around times by providing accurate results in up to 24 hours.

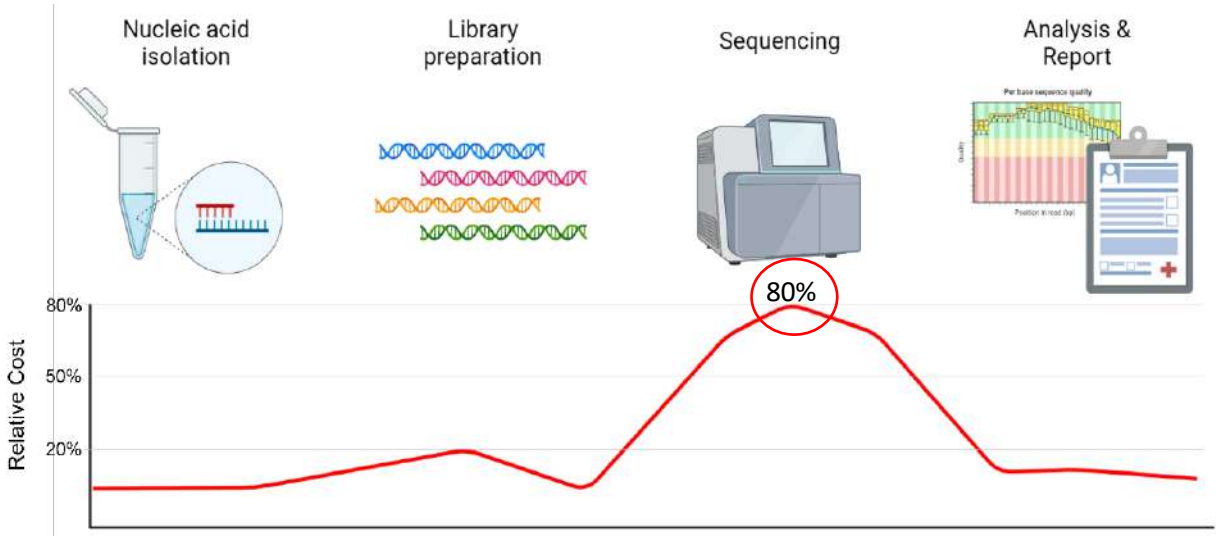

Figure 1. Typical NGS workflow and an estimation of the relative cost of each step

Despite not representing the highest fraction of the costs (Fig. 1), the single-molecule resolution provided by NGS generates a significant amount of raw data that can be quite complex to handle, as it requires sophisticated bioinformatics software, cutting edge computational hardware and specialized human resources. All these factors entail several layers of logistical and financial constraints that need to be accounted for to accelerate the integration of NGS-based pipelines in healthcare systems.

Bottlenecks of Clinical Bioinformatics: a multi-layered issue

Clinical bioinformatics is still faced with many challenges and constraints that prevent it to achieve enough maturity for widespread utilization. **From poorly curated and highly scattered genomic resources to a poor offer on ready-to-use software and integrative frameworks, much is yet to be optimized in this regard.** For instance, a recent survey made by Yale University revealed that inadequate training and the lack of proper software were the two major challenges biomedical researchers faced in analysing NGS data (Garcia-Milian et al., 2018). For healthcare practitioners, these bottlenecks are expected to increase several-fold. The most relevant limitations involved with the handling of large DNA/RNA datasets outputted from contemporary NGS pipelines are listed below, including their prospective influence on costs, for all stakeholders across the Clinical Microbiology value-chain, and the estimated impacts on clinical output.

Bottleneck #1	Cost	Impact
Extensive database curation	$	+++

There is a large variety of genomic resources with incomplete reference databases, dubious taxonomic assignments (often considering only partial sequence homologues) and sequence data that is often agnostic to serovar-level genetic polymorphisms (Schlaberg et al., 2017). This results in NGS-based pathogen detection pipelines that are of dubious relevancy for clinical practice, due to their propensity for false-positive and false-negative results, a direct consequence of inconsistent taxonomical terminologies and data formats, as well as of the overrepresentation of sequence data

from clinically common pathogens in detriment of emerging pathogens (e.g., in the scope of an emerging epidemic/pandemic outbreak) (Schlaberg et al., 2017).

Given its confounding impact in clinical diagnosis, high-fidelity databases become a critical logistical constraint and proper database engineering and optimization must be a high-priority task for stakeholders operating in the Clinical Bioinformatics ecosystem. Indeed, highly customized databases are increasingly sought after in this field, and they must strive for:

(i) an ongoing curation of the taxonomical assignments of individual pathogens, considering up-to-date phylogenetic relationships and the typically high inherent genetic diversity of microbial life, by resorting to independent and reproducible classification tools.

(ii) proper benchmarking of confounding factors in pathogen detection, for instance by highlighting microbial pathogens with unusual sequence homologies and/or close taxonomic relationships (that could lead to a higher output of false-positive results).

(iii) a quick and effective tracking of potential misannotated and/or misrepresented pathogenic species.

(iv) a prioritization of database accuracy instead of size, as large databases often come at the cost of reduced annotation quality.

Bottleneck #2	Cost	Impact
Specialized software and poor integrability	$$	+++

Clinical bioinformatics is still highly reliant on “deep computing” approaches, mostly consisting of gated, poorly standardized and code-based software lacking optimized human-computer interfaces and integrability with centralized network systems. Both these aspects have been highly disregarded in the sparse commercial solutions that are currently available in the market, which has prevented clinical NGS to achieve the required maturity for widespread utilization in clinical settings (Ahmed et al., 2014).

Optimized human-computer interfaces coupled with online and centralized data streaming capabilities that are embedded in simple and ready-to-use push-to-operate applications should be regarded as standards in software packages designed to streamline the handling of large and complex NGS datasets for pathogen detection. Particular attention should be paid to the intended end-user, which are clinicians, health

technicians and other medical staff that often lack strong informatics backgrounds and, thus, require efficient communication protocols with the bioinformatics software. In that vein, according to Ahmed et al. (2014) there are four critical aspects that must be accommodated steps while using any type of software in clinical practice:

(i) simple installation of the software's compiler and editor, which are necessary to run and execute the scripts and often represent the largest limitation associated with the use of specialized software by medical staff

(ii) a friendly graphical user interface, avoiding command line-based interfaces, that strives to be easily deployable and usable

(iii) an integrated bioinformatics software framework that can natively fetch and combine the data from the sequencer with the patients' metadata from centralized systems such as Laboratory Information Management Systems (LIMS)

(iv) proper documentation associated with the installation and use of the sotware application, as well as to assist with common troubleshooting issues.

Bottleneck #3	Cost	Impact
Data handling and management	**$$$**	**++**

Data storage of large NGS datasets entails both logistical and financial impacts in clinical contexts. Regulators enforce healthcare institutions to store patient data for decades-long tenures, which for large NGS datasets can become highly cumbersome and expensive due to the ever-growing number of patients, even if sophisticated methods of data compression are implemented (Calabrese and Cannataro, 2016). Traditionally, storage and archiving of clinical data is done on-site, in hard-drives or local servers, which in addition of representing a cumbersome logistical affair, it also forefronts significant cybersecurity and data integrity implications.

In this context, **cloud computing may facilitate the handling and management of large datasets produced by clinical bioinformatics pipelines, as it facilitates the efficient storage, retrieval and integration of clinical data in a high-throughput manner**. While there are a handful of cloud-based bioinformatics frameworks that are currently in implementation for both academic and commercial purposes, those focusing on streamlining the dynamic storage of large volumes of clinical data are the ones that better fit this purpose. Specifically, "*Data as a Service"* or "*Infrastructure as a Service"* are relevant service models that are offered in cloud-hosted computing infrastructures

that ensure the dynamic, interactable and user-controlled handling and storage of clinical data in virtual servers with specific computational capabilities and/or storage capacities (Calabrese and Cannataro, 2016).

Bottleneck #4	Cost	Impact
Specialized human resources	**$$$**	**+++**

The complex hardware and software requirements of the current clinical NGS landscape requires highly specialized staff for data handling and management, as well as to produce the desired clinical outputs. For this reason, and many others, ever since Clinical Bioinformatics has been regarded as a separate field of knowledge within the Medical Informatics domain (Belmont and Shaw, 2016), highly specialized training and certification programmes have emerged to meet the demands of the ever-increasing technical and scientific requirements of this field. From data wrangling and outputting to software and hardware troubleshooting, it is not both realistic nor feasible to expect that physicians, pathologists and other medical staff can easily accumulate these highly specialized functions. The alternative, then, is to either attract and hire highly specialized staff to accommodate these roles, which may entail significant costs to healthcare institutions, **or to prioritize optimized software packages that require minimal handling and troubleshooting and that outsource most of their hardware requirements off-site (i.e., via cloud-computing).**

The current Clinical Bioinformatics landscape

The current offer on Clinical Bioinformatics software is characterized by a sparse array of commercial and open-source software packages with significant limitations, such as:

(i) only working within specific operating systems (i.e., 16sPIP only works in a Linux environment)

(ii) being designed to target only a small subset of human pathogens (with bacterial pathogens being the most common)

(iii) being compatible with old-generation, low throughput sequencing solutions (e.g., Sanger sequencing)

(iv) lacking LIMS integrability

The table below shows some examples of popular Clinical Bioinformatics software solutions designed for pathogen detection.

	SmartGene IDNS	MicroSEQ™	16sPIP	MicroBridge
IP holder (Country)	SmartGene Services (Switzerland)	Applied Biosystems™ (United States)	ICDC (China)	Applied Biosystems™ (United States)
Type of sequencing	Sanger sequencing	Amplicon sequencing	Amplicon sequencing	Sanger sequencing
Commercial/public	Commercial	Commercial	Public	Public
Quality control	Partial	Based on type strains only	Partial	Phenotypic data integration
Curation	Weekly	Periodically	Periodically	Periodically
Paywall?	Yes	Yes	No	No
Database size (no. of reference sequences)	112,000	2,000	252,567	6,500
LIMS Integrated?	No	No	No	No

As an aim to address shortcomings and various limitations of current solutions Micronbrane Medical has started to develop PaRTI-Cular™, proprietary bioinformatics analysis software, which is specifically designed to accelerate data analysis and enhance reporting for more informed clinical decision in clinical diagnostics of infectious diseases.

More information about PaRTI-Cular™ can be found here:
https://micronbrane.com/#products

If you have interest to test PaRTI-Cular™ or join our development team, please contact us at info@micronbrane.com

References

Ahmed, Z., Zeeshan, S., & Dandekar, T. (2014). Developing sustainable software solutions for bioinformatics by the "Butterfly" paradigm. F1000Research, 3. https://doi.org/10.12688%2Ff1000research.3681.2

Belmont, J. W., & Shaw, C. A. (2016). Clinical bioinformatics: emergence of a new laboratory discipline. Expert Review of Molecular Diagnostics, 16(11), 1139-1141. https://doi.org/10.1080/14737159.2016.1246184

Calabrese, B., & Cannataro, M. (2016). Cloud Computing in Bioinformatics: current solutions and challenges. https://doi.org/10.7287/peerj.preprints.2261v1

Dulanto Chiang, A., & Dekker, J. P. (2020). From the pipeline to the bedside: advances and challenges in clinical metagenomics. The Journal of infectious diseases, 221(Supplement_3), S331-S340. https://doi.org/10.1093/infdis/jiz151

Garcia-Milian, R., Hersey, D., Vukmirovic, M., & Duprilot, F. (2018). Data challenges of biomedical researchers in the age of omics. PeerJ, 6, e5553. https://doi.org/10.7717/peerj.5553

Schlaberg, R., Chiu, C. Y., Miller, S., Procop, G. W., Weinstock, G., Professional Practice Committee and Committee on Laboratory Practices of the American Society for Microbiology, & Microbiology Resource Committee of the College of American Pathologists. (2017). Validation of metagenomic next-generation sequencing tests for universal pathogen detection. Archives of Pathology and Laboratory Medicine, 141(6), 776-786. https://doi.org/10.5858/arpa.2016-0539-RA