myNEO Therapeutics

The generation and application of simulated sequencing data for tool benchmarking and performance assessment

myNEO BV
Ottergemsesteenweg-Zuid 808
9000 Ghent, Belgium

www.myneotx.com
contact@myneotx.com

Introduction

High-throughput sequencing has increased our knowledge on the human genome tremendously. The technique allows us to rapidly elucidate DNA and RNA sequences resulting in accumulating amounts of available genomic and transcriptomic data becoming available. The scale and complexity of the data poses increasing challenges for the accuracy and efficiency of pre- and post-data-processing methods as well as for the interpretation of the ensuing results. The evolution in sequencing technology has thus gone hand in hand with continuous improvement of existing and development of novel computational tools covering different aspects of sequencing data analysis such as read alignment, variant calling, and expression analysis, among others. Often these tools are tweaked for specific sequencing platforms (e.g. short- vs long-read sequencing) and many are accompanied by unique features to try to distinguish them from competing methods. The different approaches applied in distinct software packages makes them prone to tool-specific biases and errors, potentially leading to low concordance of the obtained results[1,2]. Tool comparison and benchmarking are thus of utmost importance to chart the accuracy and performance of each tool to identify the computational method best suited for your study. The remainder of the paper will focus on the application of benchmarking and performance assessment in the context of variant calling algorithms.

Reliable characterization of a tool requires so called 'gold standard' datasets with known status of the variants present (i.e. the ground-truth). Unfortunately obtaining physical samples harboring every potential variant is rarely possible. Most research groups circumvent this by processing multiple representative samples (e.g. the Genome in a Bottle samples[3]) and rely on general sequencing statistics and the (limited) available variants to evaluate the performance of their tool. Alternatively, synthetic DNA representing the variants of interest can be spiked into one or more reference samples. However, this limits the number variants for which performance can be assessed, and becomes more complex and difficult to achieve for specific variant types (e.g. translocations, transposable element insertion, ...).

Simulated sequencing data

Another interesting approach is the use of simulated or in-silico sequencing data. Using such artificial sequencing samples is more cost efficient than using physical samples since it omits the need to sequence novel samples. Other benefits over physical samples include the ability to assess any number and type of variants, and the possibility to easily evaluate the impact of different sample and sequencing-inherent features such as sample purity and sequencing coverage. However, since any in-silico sample represents a simplification of a real sample, and inclusion of every potential bias inherent to sequencing and physical samples is unachievable, the former will always lack components present in the latter.

Depending on the application in-silico sequencing data can be generated in different ways. Sequencing data can be simulated by mixing reads obtained from multiple physical samples, or by manipulating real sequencing data by introducing variants[4]. The major advantage of the latter two

approaches lies in the fact that any biases and noise inherent to physical samples are retained in the simulated sample. While sample mixing only increases the number of variants and the range of variant allele frequencies for which the tool can be benchmarked, inserting custom variants in real sequencing data has the added value of expanding the pool of variant types and allowing performance testing on (a larger set of) clinically relevant variants.

In-silico sequencing data can also be generated de novo by extracting reads from a reference genome. In the context of variant benchmarking the genome can be modified to present a set of variants of interest. This approach is especially useful when one wants to benchmark a tool on complex and rare variants for which few samples are available. De novo simulated data, however, generally does not model biases and noise perfectly, but allows customization of most sample and sequencing-related parameters such as coverage, read length, variant allele frequency, among others.

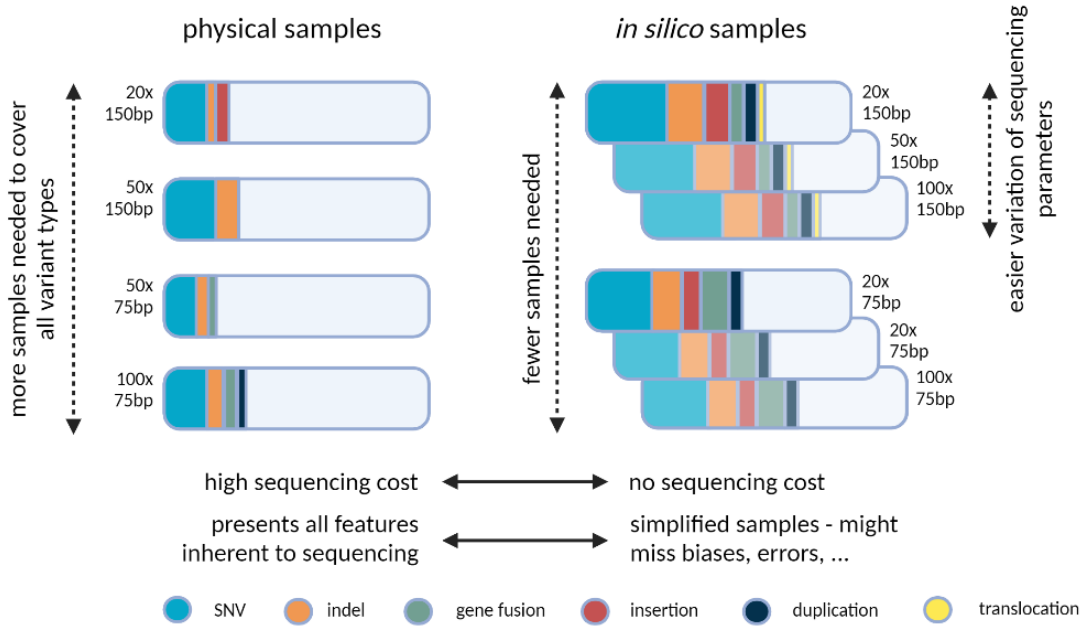

Figure 1: Comparison of physical samples and in-silico samples.

Genomic sequencing simulators with variant-introduction functionality

Various tools for generating in-silico data have been developed over the years. The description below will be restricted to de novo simulators capable of generating samples containing a set of customizable variants. An exhaustive list of other DNA sequencing simulators can be found in literature[4–8].

The composition of a sequencing dataset is largely directed by the type of sample and sequencing run. For example, Illumina instruments are more prone to substitution errors, while indels are the

main source of error on the IonTorrent platform [9,10]. Other parameters that can significantly impact downstream sequencing results include read length, sequencing depth, base quality scores and GC-bias. Most simulators estimate these parameter settings from an empirical dataset, but provide the user to customize individual settings. This approach has the benefit that the parameters set will more match closely profiles observed from physical samples, which is especially advantageous in the context of modeling of sequencing errors. Examples of such DNA simulators are ART, (d)wgsim, ReadSim, FASTQsim, EAGLE, BEAR and SInC, among others[11–18]. Other tools require all parameter values to be set manually by the user, for which CuReSim and ArtificialFastqGenerator are the most common ones [19,20]. Sequencing platform compatibility and the ability to simulate paired- and/or single-end reads are other differentiating aspects of sequencing simulators. Although most can simulate in-silico data for multiple platforms (e.g. EAGLE, ART, wgsim, ...), others such as pIRS, SinC and SimSeq are limited to the Illumina sequencing platform [11,13,16,18,21,22].

If one wishes to apply an in-silico dataset for variant calling benchmarking, an important simulator selection criteria is the type of variants it can introduce. Although most of the simulators described above are capable of mutating reads to harbor SNVs and indels, workflows compatible with structural variants are less common. pIRS, EAGLE, and dwgsim are examples of simulators having functionality to introduce insertions, with the latter two also being compatible with translocations [12,16,21].

In comparison to DNA simulators fewer computational tools to generate in-silico RNA sequencing datasets are available. The crucial aspect here is the read distribution along a transcript to approximate the potential 3'-5' bias inherent to some of the sequencing platforms, and the ability to reflect the transcript expression profiles of physical RNA samples. Most of the available RNA simulators such as Polyester, BEERS(2) and CAMPAREE take these into account, but none currently possess the functionality to introduce variants[23–26].

Use cases for in-silico generated data

Validating and benchmarking a bioinformatics tool is an important process in its development. The ease of creation of in-silico data in a cost-efficient manner allows the generation of large artificial sample sets. This enables developers to assess in-depth the accuracy and performance of their algorithms in a reproducible manner. By creating dataset pool using a constant set of design parameters and varying a single setting at a time makes it easier to pinpoint a tools weaknesses. In the context of a variant calling simulated data can for example be helpful to validate a tools performance on a more extensive set of (rare) variants, and to assess the impact of for example low(er) covered samples. Other use cases include regular quality monitoring and evaluating the impact of package updates in case the computational method depends on third-party software tools.

Conclusion

Altogether, sequencing simulators can be a cost-efficient alternative to physical data during the development stages of sequencing related computational tools. The versatility of the in-silico

datasets that can be generated provide a powerful way to track the effect of changes to an analysis workflow and assess its sensitivity and accuracy.

About myNEO Therapeutics

myNEO Therapeutics is a distinguished biopharmaceutical powerhouse, dedicated to pioneer breakthrough immunotherapies to fight cancer. myNEO Therapeutics leverages its ImmunoEngine platform to tap into novel promising immunotherapy tumor targets as well as enable patient-centric tumor profiling and therapy developments. In the past 5 years, myNEO Therapeutics has closed several partnerships with major biopharma companies to identify and prioritize tumor neoantigens, bring those into development, and profile patient response datasets within clinical studies. In parallel, myNEO has several internal programs developing targets in the dark genome – named camyotopes™ – which have the potential to unlock immunotherapy for large patient populations who currently do not respond. The company was founded in 2018, and raised its capital from different VC funds in Europe.

Interested in more information about myNEO Therapeutics? Contact us!

Lien Lybaert, PhD.
Chief Development Officer
lien.lybaert@myneotx.com

References

1. Barbitoff, Y. A., Abasov, R., Tvorogova, V. E., Glotov, A. S. & Predeus, A. V. Systematic benchmark of state-of-the-art variant calling pipelines identifies major factors affecting accuracy of coding sequence variant discovery. *BMC Genomics* **23**, 1–17 (2022).
2. Krishnan, V. *et al.* Benchmarking workflows to assess performance and suitability of germline variant calling pipelines in clinical diagnostic assays. *BMC Bioinformatics* **22**, 1–17 (2021).
3. Zook, J. M. *et al.* Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nature Biotechnology 2014 32:3* **32**, 246–251 (2014).
4. Duncavage, E. J. *et al.* 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. *Journal of Molecular Diagnostics* **25**, 3–16 (2023).
5. Escalona, M., Rocha, S. & Posada, D. A comparison of tools for the simulation of genomic next-generation sequencing data. *Nat Rev Genet* **17**, 459 (2016).
6. Zhao, M., Liu, D. & Qu, H. Systematic review of next-generation sequencing simulators: computational tools, features and perspectives. *Brief Funct Genomics* **16**, 121–128 (2017).
7. Alosaimi, S. *et al.* A broad survey of DNA sequence data simulation tools. *Brief Funct Genomics* **19**, 49 (2020).
8. Milhaven, M. & Pfeifer, S. P. Performance evaluation of six popular short-read simulators. *Heredity (Edinb)* **130**, 55 (2023).

9. Dohm, J. C., Lottaz, C., Borodina, T. & Himmelbauer, H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res* **36**, (2008).
10. Nakamura, K. *et al.* Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res* **39**, (2011).
11. Huang, W., Li, L., Myers, J. R. & Marth, G. T. ART: a next-generation sequencing read simulator. *Bioinformatics* **28**, 593 (2012).
12. GitHub - nh13/DWGSIM: Whole Genome Simulator for Next-Generation Sequencing. https://github.com/nh13/DWGSIM.
13. Li et al. GitHub - lh3/wgsim: Reads simulator. https://github.com/lh3/wgsim.
14. Lee, H. *et al.* Error correction and assembly complexity of single molecule sequencing reads. *bioRxiv* 006395 (2014) doi:10.1101/006395.
15. Shcherbina, A. FASTQSim: Platform-independent data characterization and in-silico read generation for NGS datasets. *BMC Res Notes* **7**, 1–12 (2014).
16. GitHub - sequencing/EAGLE: Enhanced Artificial Genome Engine: next generation sequencing reads simulator. https://github.com/sequencing/EAGLE.
17. Johnson, S., Trost, B., Long, J. R., Pittet, V. & Kusalik, A. A better sequence-read simulator program for metagenomics. *BMC Bioinformatics* **15**, 1–10 (2014).
18. Pattnaik, S., Gupta, S., Rao, A. A. & 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* **15**, 1–9 (2014).
19. Frampton, M. & Houlston, R. Generation of Artificial FASTQ Files to Evaluate the Performance of Next-Generation Sequencing Pipelines. *PLoS One* **7**, e49110 (2012).
20. Caboche, S., Audebert, C., Lemoine, Y. & Hot, D. Comparison of mapping algorithms used in high-throughput sequencing: Application to Ion Torrent data. *BMC Genomics* **15**, 1–16 (2014).
21. Hu, X. *et al.* pIRS: Profile-based Illumina pair-end reads simulator. *Bioinformatics* **28**, 1533–1535 (2012).
22. Earl, D. *et al.* Assemblathon 1: A competitive assessment of de novo short read assembly methods. *Genome Res* **21**, 2224 (2011).
23. Frazee, A. C., Jaffe, A. E., Langmead, B. & Leek, J. T. Polyester: simulating RNA-seq datasets with differential transcript expression. *Bioinformatics* **31**, 2778–2784 (2015).
24. Brooks, T. G. *et al.* BEERS2: RNA-Seq simulation through high fidelity in-silico modeling. *Brief Bioinform* **25**, (2024).
25. Grant, G. R. *et al.* Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). *Bioinformatics* **27**, 2518–2528 (2011).
26. Lahens, N. F. *et al.* CAMPAREE: a robust and configurable RNA expression simulator. *BMC Genomics* **22**, 1–12 (2021).