Factors for consideration in *de novo* sequencing

Summary

- Next-generation sequencing (NGS)-based *de novo* sequencing has revolutionized the ability to obtain primary genetic information from a tremendous number of previously uncharacterized organisms, due to its affordability and speed.
- A variety of technical, biological, and computational issues—such as increasing read length, overall genome complexity, and improvements to data analysis algorithms—need to be considered to optimize the use of *de novo* sequencing.
- Advances in NGS technology for *de novo* sequencing are enabling the discovery of a vast amount of novel genomic information.

Introduction

NGS continues to yield insights into the genomics of an expanding number of organisms. For many of these organisms, little to no prior genetic information has been available, particularly for viruses, bacteria, and fungi. In some cases, the existing genetic information is of such low quality that it is not particularly useful. When the genomic data for a particular organism are either unavailable or of insufficient quality, *de novo* sequencing can be used to generate or update it. *De novo* sequencing constitutes sequencing a novel genome, in the absence of a reference sequence, and subsequent assembly of the reads to form a map of the complete genome. This powerful technique enables the determination of many novel genome sequences at an unprecedented rate.

A reference genome, or reference assembly, is a representative example of the genome of a particular species, and the goal of *de novo* sequencing is to generate a reference sequence to be used for more detailed genetic studies. It is typically a mosaic, because it is usually not constructed from a single individual of a species. For example, the human genome assembly is constructed from 13 individuals, and it contains more than a million single-nucleotide polymorphisms (SNPs) [1]. A reference genome is considered to be a guide that is a good approximation of the genetic information for a particular species. However, the quality and availability of reference sequences can vary greatly. Many species have no reference sequences at all, while others may have only rudimentary ones that are insufficient for more detailed genomic studies due to a variety of technical or biological issues.

The construction of a reference genome is based on assembling shorter sequencing reads into longer contiguous sequences (contigs) using computational tools like genome assemblers. The main goal is to generate an overall physical map that does not contain gaps and is an accurate representation of the genome. Traditionally, reference genomes have been constructed with Sanger sequencing, which has longer reads than NGS, but this approach can be cost prohibitive. NGS has revolutionized the ability to perform *de novo* sequencing because of its affordability. However, some technical issues (e.g., short reads) still need to be improved in order for the method to be broadly applicable. Some of the challenges for genome assembly posed by short reads include the increased impact of base-calling errors and difficulty with aligning reads in repetitive regions of the genome. To some extent, these deficiencies can be addressed with the use of paired-end reads, mate-pair sequencing approaches, and optimal genome assembly algorithms [2].

Considerations

Several technical, biological, and computational factors need to be considered in order for *de novo* sequencing with NGS to be successful. Of primary concern is the error rate and read length of the sequencing method. Longer sequencing reads, such as those obtained from Sanger sequencing or single-molecule sequencing technologies [3], significantly improve *de novo* sequencing, yet they have a higher error rate compared to shorter reads and are generally more expensive. The typically shorter reads produced by NGS data can be more accurate, but they can also result in more difficult genome assembly. Eukaryotic genomes are the most challenging to assemble with shorter read lengths, due to their overall length and complexity. Accurate genome assembly depends on the ability to order the reads correctly, have no gaps, and resolve repetitive regions. Longer sequencing reads can make all of these tasks easier. Developments in NGS technology are currently focused on methods to overcome some of these issues. Many different types of advances are enabling longer read lengths, such as improvements in sequencers, protocols, computational methodologies, and algorithms.

Computational and biological factors are also very important to evaluate when considering *de novo* sequencing. Algorithms and genome assemblers are being developed at a rapid pace, and there are a variety of options to choose from, all with distinct advantages and disadvantages [2]. Additional computational issues, including storage, hardware, overall efficiency, and specific software, should be taken into account. Smaller genomes, like those of microbes, are much easier to sequence than larger, more complex eukaryotic genomes. Repetitive genomic segments are particularly challenging for *de novo* sequencing because it can be difficult to detect the repeat sequence during genome assembly. Some of these repeats can be nearly 100% identical and thousands of base pairs long, such that they cannot be bridged by short reads. In addition, the GC content of a genome can affect data quality. For example, GC bias has been reported to lead to poor coverage at certain thresholds [4]. Strong GC bias fragments genome assembly due to low coverage of reads in GC-poor or GC-rich regions of the genome, regardless of the assembly algorithms used. Therefore, improving GC coverage during library preparation can facilitate *de novo* assembly. Invitrogen™ Collibri™ Physically Sheared (PS) DNA Library Prep Kits, when used in combination with Invitrogen™ Collibri™ Library Amplification Master Mix (2X), significantly reduces GC bias in libraries sequenced on Illumina™ systems compared to equivalent methods, making it an ideal choice for *de novo* sequencing.

The combined use of several sequencing technologies is referred to as a hybrid approach, and it is becoming more prevalent because it can solve many of the technical issues previously associated with *de novo* sequencing. Recent studies have shown this approach to be feasible even for the highly complex human genome [5]. A hybrid approach generally includes the use of long reads (with their higher error rate) to build a scaffold, and short reads (with their lower error rate) to fill in and improve the quality of a new genome. Longer reads are easier for overall genome assembly but are more expensive than shorter reads. In general, a hybrid approach may be a reasonable compromise with regard to costs, error rates, read lengths, and quality.

Conclusions

- NGS-based *de novo* sequencing can uncover novel genetic information from a variety of organisms, and this application has grown at an impressive rate.
- Several technical, biological, and computational factors need to be considered in order to optimize this approach, including how to increase overall read length, manage the degree of genome complexity, and select the best genome assembler for the data.
- Recent advances will enable NGS-based *de novo* sequencing to generate detailed genetic maps of more complex organisms in shorter times and at lower cost than previous methods.
- Collibri PS DNA library preparation kits provide even coverage regardless of GC content, making them an ideal choice for library preparation for *de novo* sequencing.

References

1. The Genome Reference Consortium. ncbi.nlm.nih.gov/grc/human (accessed July 2019).
2. Khan AR, Pervez MT, Babar ME et al. (2018) A comprehensive study of *de novo* genome assemblers: current challenges and future prospective. *Evol Bioinform Online* 14:1176934318758650.
3. Sakai K, Naito K, Ogiso-Tanaka E et al. (2015) The power of single molecule real-time sequencing technology in the *de novo* assembly of a eukaryotic genome. *Sci Rep* 5:16780.
4. Chen YC, Liu T, Yu CH et al. (2013) Effects of GC bias in next-generation-sequencing data on *de novo* genome assembly. *PLoS One* 8(4):e62856.
5. Mostovoy Y, Levy-Sakin M, Lam J et al. (2016) A hybrid approach for *de novo* human genome sequence assembly and phasing. *Nat Methods* 13(7):587–590.

Ordering information

Product		Quantity	Cat. No.
DNA-Seq kits for Illumina systems			
Collibri ES DNA Library Prep Kits	with CD Indexes	24 preps	A38605024
	with CD Indexes	96 preps	A38607096
	with UD Indexes, Set A (1-24)	24 preps	A38606024
	with UD Indexes, Set B (25-48)	24 preps	A43605024
	with UD Indexes, Set C (49-72)	24 preps	A43606024
	with UD Indexes, Set D (73-96)	24 preps	A43607024
Collibri PCR-Free ES DNA Library Prep Kits	with CD Indexes	24 preps	A38545024
	with CD Indexes	96 preps	A38603096
	with UD Indexes, Set A (1-24)	24 preps	A38602024
	with UD Indexes, Set B (25-48)	24 preps	A43602024
	with UD Indexes, Set C (49-72)	24 preps	A43603024
	with UD Indexes, Set D (73-96)	24 preps	A43604024
Collibri PS DNA Library Prep Kits	with CD Indexes	24 preps	A38612024
	with CD Indexes	96 preps	A38614096
	with UD Indexes, Set A (1-24)	24 preps	A38613024
	with UD Indexes, Set B (25-48)	24 preps	A43611024
	with UD Indexes, Set C (49-72)	24 preps	A43612024
	with UD Indexes, Set D (73-96)	24 preps	A43613024
	with UD Indexes, Set A-D (1-96)	96 preps	A38614196
Collibri PCR-Free PS DNA Library Prep Kits	with CD Indexes	24 preps	A38608024
	with CD Indexes	96 preps	A38610096
	with UD Indexes, Set A (1-24)	24 preps	A38609024
	with UD Indexes, Set B (25-48)	24 preps	A43608024
	with UD Indexes, Set C (49-72)	24 preps	A43609024
	with UD Indexes, Set D (73-96)	24 preps	A43610024
	with UD Indexes, Set A-D (1-96)	96 preps	A38615196

CD = combinatorial dual, UD = unique dual

invitrogen

Ordering information (continued)

Product	Quantity	Cat. No.
RNA-Seq kits for Illumina systems		
Collibri Stranded RNA Library Prep Kit for Illumina Systems	24 preps	A38994024
	96 preps	A38994096
Collibri Stranded RNA Library Prep Kit for Illumina Systems with H/M/R rRNA Depletion Kit	24 preps	A39003024
	96 preps	A39003096
ERCC RNA Spike-In Mix	1 kit	4456740
ERCC ExFold RNA Spike-In Mixes	1 kit	4456739
Library quantification		
Collibri Library Quantification Kit	100 rxns	A38524100
	500 rxns	A38524500
Qubit 4 Fluorometer, with WiFi	1 fluorometer	Q33238
Qubit 4 NGS Starter Kit, with WiFi	1 kit	Q33240
Library amplification		
Collibri Library Amplification Master Mix	50 rxns	A38539050
	250 rxns	A38539250
Collibri Library Amplification Master Mix with Primer Mix	50 rxns	A38540050
	250 rxns	A38540250

H/M/R = human/mouse/rat

Find out more at **thermofisher.com/collibri**

ThermoFisher SCIENTIFIC

COL23574 0919