Marine Genomics 30 (2016) 3–13

Contents lists available at ScienceDirect

Marine Genomics

journal homepage: www.elsevier.com/locate/margen

Next-generation biology: Sequencing and data analysis approaches for non-model organisms

Rute R. da Fonseca [a,*], Anders Albrechtsen [a], Gonçalo Espregueira Themudo [c], Jazmín Ramos-Madrigal [b], Jonas Andreas Sibbesen [a], Lasse Maretty [a], M. Lisandra Zepeda-Mendoza [b], Paula F. Campos [b,d], Rasmus Heller [a], Ricardo J. Pereira [b]

[a] The Bioinformatics Centre, Department of Biology, University of Copenhagen, Copenhagen, Denmark
[b] Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark
[c] Section of Forensic Genetics, Department of Forensic Medicine, University of Copenhagen, Copenhagen, Denmark
[d] CIMAR/CIIMAR, Centro Interdisciplinar de Investigação Marinha e Ambiental, Universidade do Porto, Rua dos Bragas 177, 4050-123 Porto, Portugal

ARTICLE INFO

Article history:
Received 30 November 2015
Received in revised form 23 March 2016
Accepted 26 April 2016
Available online 13 May 2016

Keywords:
RADseq
RNAseq
Targeted sequencing
Genotype likelihoods
Comparative genomics
Population genomics

ABSTRACT

As sequencing technologies become more affordable, it is now realistic to propose studying the evolutionary history of virtually any organism on a genomic scale. However, when dealing with non-model organisms it is not always easy to choose the best approach given a specific biological question, a limited budget, and challenging sample material. Furthermore, although recent advances in technology offer unprecedented opportunities for research in non-model organisms, they also demand unprecedented awareness from the researcher regarding the assumptions and limitations of each method.

In this review we present an overview of the current sequencing technologies and the methods used in typical high-throughput data analysis pipelines. Subsequently, we contextualize high-throughput DNA sequencing technologies within their applications in non-model organism biology. We include tips regarding managing unconventional sample material, comparative and population genetic approaches that do not require fully assembled genomes, and advice on how to deal with low depth sequencing data.

1. Introduction

High-throughput sequencing, more broadly referred to as next-generation sequencing (NGS), has become essential for modern day research within biological sciences, particularly in evolutionary biology. Since the assembly of the first complete genome using Sanger capillary sequencing in 1977 (Sanger et al., 1977), large technological improvements have been made and different methods have been developed. These methods aim to increase the sequencing throughput as well as the quality and length of the reads, while decreasing the time and cost of the process in such a way that it seems that virtually any biological question can be asked given enough data. However, the increase in data production comes with a cost: the corresponding data analysis approaches require a more detailed knowledge on the caveats and drawbacks of each method.

The literature of evolutionary biology has traditionally been dominated by model species such as mammals and drosophilids, for which fully sequenced and well-annotated genomes have been available for years. The recent advent of high-throughput sequencing opened the use of genomic approaches to the study of non-model organisms, allowing the test of generalizations based on a limited number of model species, and unlocking new research programs in fields related to evolutionary biology, such as phylogenomics and population genomics. The choice of the sequencing approach needs to take into account the evolutionary time scale of the biological question (Fig. 1). For example, transcriptome data (RNA-seq) has been used to produce hundreds of protein alignments that resolved deep phylogenetic relationships in Metazoans (Smith et al., 2011) and in plants (Wickett et al., 2014), providing key insights into how characters such as development, morphology or genome structures evolve throughout the tree of life (Dunn et al., 2014). For taxa that have diverged at relatively deep time scales, up to hundreds of millions of years of evolution, targeted sequencing of highly conserved genomic regions (named ultra-conserved elements or UCEs) have been used to establish well-resolved phylogenies of large species radiations such as Amniotes (Faircloth et al., 2012), vertebrates (Lemmon et al., 2012), birds (Prum et al., 2015) or mammals (McCormack et al., 2012). For bacterial species, which have a simple genomic structure, whole genome sequencing has been used to ask similar questions (Ziemert et al., 2014). Targeted sequencing can be extended to micro-evolutionary time scales by designing targets specific for coding and non-coding genomic regions, based on partial genomes of

* Corresponding author.
E-mail address: fonseca@binf.ku.dk (R.R. da Fonseca).

http://dx.doi.org/10.1016/j.margen.2016.04.012

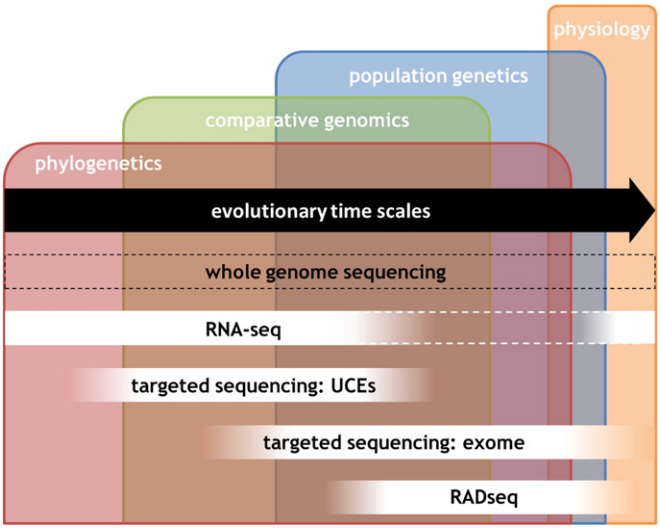

Fig. 1. Application of different high-throughput sequencing methods to different evolutionary time scales. Research applications related to evolutionary biology (colored polygons) address biological questions at different, but overlapping, evolutionary time scales (black arrow), spanning from hundreds of millions of years of evolution between Phyla (left), to generations between populations, individuals or cells (right). In the absence of whole genome sequencing, different high-throughput sequencing methods (white bars) provide a cost- and time-efficient alternative for non-model species. Benefits and limitations of each method depend on the time scale relevant to each biological question.

closely related species (e.g. the domestication of maize (da Fonseca et al., 2015), and diversification of palm trees (Heyduk et al., 2016)). At shorter evolutionary time scales, methods such as Restriction-site Associated DNA sequencing (RADseq), in which thousands of genomic regions spread throughout the genome are sequenced, have been used to establish phylogenetic relationships between closely related species and populations, despite the large confounding effect of interspecific gene flow and incomplete lineage sorting (e.g. diversification and hybridization in oaks (Eaton et al., 2015)). In order to generate summary statistics for population genetics in the absence of a reference genome, Gayral et al. (Gayral et al., 2013) established a pipeline for transcriptome data that controls for paralogue genes and for variation in gene expression among individuals and loci. Using this method, population genomics studies across animals have shown that levels of genetic diversity within a species seem to be largely determined by its ecological strategy, such as propagule size and fecundity (Romiguier et al., 2014), rather than geographic range or invasive status. The same methods have been applied to data from multiple populations within the same species, to understand how genetic drift and positive selection contribute to divergence patterns across the genome (Tsagkogeorga et al., 2012), and to identify genes repeatedly evolving under selection across multiple independent populations (Pereira et al., n.d.). Analysis of polymorphisms (e.g. F_{ST}) estimated from transcriptome data from different species (Renaut et al., 2013) or from populations within a species (Carneiro et al., 2014) have also been used to determine genomic areas of high differentiation that could harbor genes involved in local adaptation and genetic barriers to gene flow, offering insights into the genetic basis of speciation. At a more recent evolutionary time scale, RNA-seq can be used to describe genes involved in physiological adaptation of populations to different environments (e.g. in corals (Barshis et al., 2013)) or to physiological variation between individuals or cells.

Despite these broad applications of high-throughput sequencing, choosing the most appropriate method to address a specific biological question requires considering the benefits and limitations of each method.

2. Sample quality and preparation

The type, quality and quantity of the tissue samples used in genomic analyses has a great impact in the final results, both in terms of quantity and quality of reads obtained after sequencing. The success of high-throughput sequencing approaches is highly dependent on the use of high molecular weight DNA/RNA samples, and this is only possible if fresh or carefully stored tissue samples are used for nucleic acid isolation. Frozen collection and transportation can be challenging, and in fields like marine genomics, this is not always feasible. For some species (mammals, birds, fish) tissues like blood or muscle might be a good source of DNA. Plant tissues rich in resins, gums, and polyphenolics should be avoided (Abu Almakarem et al., 2012). When the specimen size is very small (e.g. small marine invertebrates) the whole specimen or even a pool of several specimens might be required to obtain enough genetic material for subsequent analyses. For microbes (bacteria, fungi, diatoms, microalgae) single species isolates are needed to ensure the required amounts of pure DNA for NGS sequencing (but laboratorial culturing is only possible for a restricted number of species).

Ideally, when in the field, samples should be collected and immediately stored either in liquid nitrogen (preferred), at −20 °C (using a freezer or dry ice) or in a chemical preservative (RNAlater type solution), a solution that rapidly permeates the tissue and protects cellular nucleic acids in unfrozen tissue samples, as these materials are susceptible to fast post-mortem degradation or degradation following collection from living specimens. If possible, several subsamples should be obtained per specimen as a back-up procedure in case something goes wrong between collection and sample processing, or to allow for high coverage sequencing. In cases where the species of interest is extinct, very rare or difficult to collect in the field, museum specimens might also be used to obtain genetic material. Such specimens are also valuable for time series analysis (e.g. (Bi et al., 2013)). Tissue samples stored in ethanol (internal organs, muscle), dried (beaks, bones), taxidermized (skin, nails, hair), or frozen in tissue banks are alternatives to freshly collected samples. The DNA obtained from these samples is usually of inferior quality (lower molecular weight and concentration) and the amount of external contaminants will be higher.

Crucial for obtaining high molecular weight DNA is also the choice of extraction method used for nucleic acid isolation, appropriate for tissue type and preservation method (Campos et al., 2009). Another extremely important preliminary step is to obtain all legal permits and documentation associated with collection of samples. Large sequencing consortiums like Genome 10 K have established very strict protocols for tissue collection and storage (Wong et al., 2012) to maximize the information obtained for each specimen.

3. Restriction-site Associated DNA sequencing (RADseq)

RADseq (Restriction-site Associated DNA sequencing) was developed as a method for the simultaneous discovery and genotyping of tens of thousands of genome-wide markers through a reduced representation protocol (Baird et al., 2008). It can be used for population genetic analyses or for building genetic maps. Two features of RADseq have made it popular for population genetic studies on non-model species: i) it does not require a reference genome and ii) it provides a cost-efficient way of genotyping genome-wide markers in many individuals. RADseq is flexible in that the restriction enzyme can be chosen so as to fit the desired reduction factor, i.e. the proportion of the genome that is sequenced. Hence the whole continuum between sequencing a few loci (e.g. a few tens of thousands) at high coverage or many loci (e.g. many hundreds of thousands) at lower coverage is accessible in a RADseq study. For example, a single Illumina lane can be used to sequence 100 individuals at 150,000 RAD loci, providing up to 10× mean coverage per locus per sample. Such data has been used to identify fresh-water adaptation in sticklebacks (Hohenlohe et al., 2010) and to infer the phylogeny of African cichlids (Wagner et al., 2013). While