Trends in Parasitology

CellPress

Forum

Unleashing the Potential of Public Genomic Resources to Find Parasite Genetic Data

Ricardo Jorge Lopes,[1,*] Antonio Muñoz Mérida,[1] and Miguel Carneiro[1]

Genetic data generated by high-throughput sequencing and deposited in public databases are increasing exponentially. A substantial amount of these data is generated from wild animals, and can contain information from nontarget organisms, such as parasites. Methodologies that leverage this available information are warranted and can help to answer questions of general interest in parasitology.

High-throughput sequencing by several next-generation sequencing (**NGS**, see Glossary) methodologies has increased exponentially the amount of genetic data available from a multitude of species, ranging from whole-genome or transcriptome sequencing to reduced representations of the genome (e.g., targeted capture or restriction-site-associated DNA sequencing, such as RADSeq). A large percentage of these data is deposited in public databases, and most importantly, there is now a wealth of data from individuals from model organisms. These data are maintained by several organizations, with a strong prominence of NCBI (**National Center for Biotechnology Information**). The primary archive for original high-throughput sequencing data is the **Sequence Read Archive** (SRA) which stores raw sequence data from multiple technologies, including Illumina, 454, IonTorrent, Complete Genomics, PacBio, and Oxford Nanopore[i]. There are other databases with sequencing data, but they are mostly composed of curated and/or assembled data (e.g., WGS – Whole-Genome Shotgun database). The amount of data that has been deposited in the SRA database is increasing exponentially as NGS is becoming widely used and the cost per nucleotide base is fast decreasing. This database now holds more than 15^{14} bases (Figure 1), an increase of more than 500% from 2014 to the present time[ii], with a large proportion of genetic data from whole-genome sequencing approaches (Figure 1). Given the tendency of DNA sequencing to become even more affordable, whole-genome experimental designs will likely continue to increase at this exponential rate in the future. It is also important to highlight that the source of more than half of the genetic data (number of bases) available in the SRA database is from vertebrates and invertebrates, thus representing a large variety of potential hosts (Figure 1).

In the majority of sequencing methodologies, genomic reads from nontarget organisms are common, and researchers are aware of their presence. Considering that research questions and hypotheses require mostly data from target species, it is natural that their bioinformatic pipelines are made to filter and treat nontarget data as contamination [1,2]. One example is the approach provided by Orosz [3] that acknowledges the presence of parasite genetic material as contamination and, therefore, identifies this as a problem to be solved before further analysis. However, from a parasitologist's point of view, the SRA database can be an important repository of information concerning parasites, and recent articles use this approach successfully to retrieve important genetic data. For example, a query on the unmapped reads from the DNA and RNA sequencing associated with the first assembly and annotation of the *Bos taurus* genome found many contigs from several species of Spirurid nematodes and a blood-borne parasite, *Babesia bigemina* [4]. In the case of the nematodes, most of these data represented either new or previously identified, but unsequenced, species. In another recent study, a query on Whole-Genome Shotgun and Transcriptome Shotgun Assembly databases, also curated by the NCBI, found 20 907 contigs of apicomplexan origin in 51 animal genome datasets, from Gregarinasina, Coccidia, Piroplasmida, and Haemosporida [5]. Importantly, for most of these taxa no molecular data had been available previously. In this case, the results provided valuable insights concerning the links between hosts and parasite community, as supplementary approaches to environmental DNA or metagenomics analysis [6].

Simple and user-friendly *in silico* strategies to identify these reads have the potential to provide valuable genetic and genomic information from parasites, with the added value of not requiring *de novo* sequencing specifically targeting the parasites' genomes. On the other hand, it is important to understand that retrieving

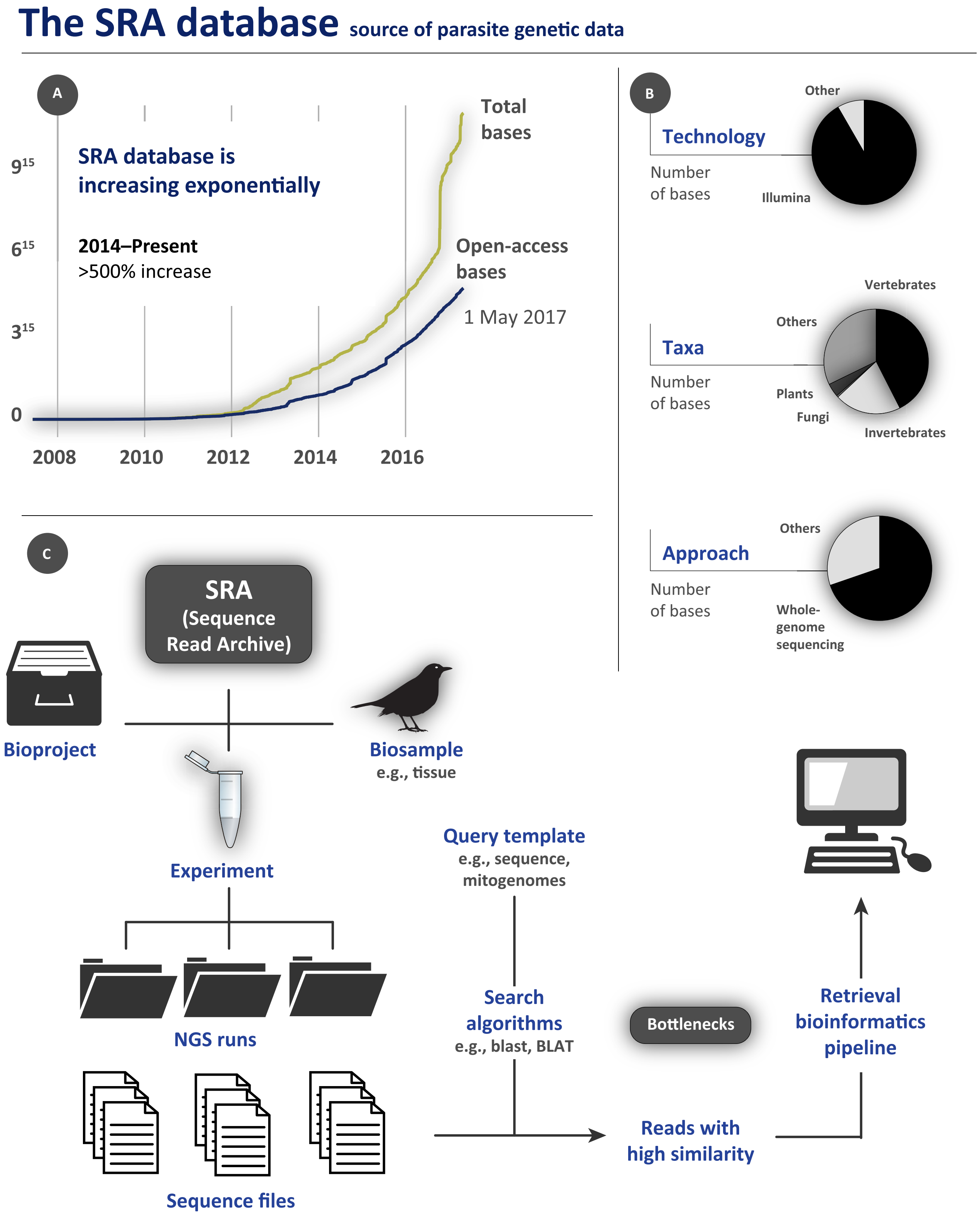
Glossary

Basic Local Alignment Search Tool (BLAST): an example of alignment search algorithms, that finds regions of similarity between biological sequences.

National Center for Biotechnology Information (NCBI): since 1988, the mission of the NCBI is to develop new information technologies to help in researching molecular and genetic processes. This includes the curation and automation of several databases and the implementation of several bioinformatic tools.

Next generation sequencing (NGS): a general term used to encompass a number of methodologies, focused on high-throughput sequencing, that parallelize the sequencing process, allowing the production of thousands or millions of genetic sequences.

Sequence Read Archive (SRA): NCBI database which stores sequence data obtained from NGS technologies, especially data concerning the project (Bioproject), sample of tissue (Biosample), and the files of results obtained in each run related to a given experiment.



Trends in Parasitology

(See figure legend on the bottom of the next page.)