Quantity but also quality: choosing a next-generation sequencing approach to address specific questions in systematics

Sonet Gontran¹, Smits Nathalie², Virgilio Massimiliano³, Nagy Zoltán T.¹, Winant Virginie², Huyse Tine², Backeljau Thierry¹,³ and De Meyer Marc²

¹ OD Taxonomy and Phylogeny (JEMU), Royal Belgian Institute of Natural Sciences, 29 Vautierstraat, B-1000 Brussels, Belgium.
² Department of Biology (JEMU), Royal Museum for Central Africa, 13 Leuvensesteenweg, B-3080 Tervuren, Belgium.
³ Evolutionary Ecology Group, University of Antwerp, 171 Groenenborgerlaan, B-2020 Antwerp, Belgium.

CONTEXT

Next-generation sequencing (NGS) technologies offer many new opportunities to study non-model organisms by analysing many more DNA markers than possible with traditional Sanger sequencing. Choosing the NGS approach that will best tackle a scientific question is not straightforward because many parameters have to be evaluated (the table below only presents some of them).

OBJECTIVE

As an addition to the reviews comparing NGS methods in general (e.g. Lemmon & Lemmon 2013), we want to evaluate how three specific scientific questions can be investigated using three different NGS methods.

RESULTS

Comparative table of three pilot NGS projects supported by JEMU in 2014 (see below).

<table>
<thead>
<tr>
<th>Project</th>
<th>Taxonomic group</th>
<th>Objective</th>
<th>NGS approach → general application</th>
<th>Requirements</th>
<th>Starting DNA material</th>
<th>Price per sample x number of samples processed</th>
<th>Timing (excl. data analysis)</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>genotyping by sequencing (GBS) →</td>
<td>population genetics</td>
<td>0.3-3 µg (quantified by an intercalating dye) free of RNA and contaminants</td>
<td>43 € x 192 samples = 8256 € (incl. library preparation, NGS run, SNP calling)</td>
<td>2 weeks of library preparation queue for outsourced NGS run: 2-4 months</td>
<td>raw data: 40 Gb (reads of 1 x 100 bp) processed dataset: 200K SNPs (expected)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>restriction-site associated DNA tags (RAD) → pop. genetics &amp; shallow-level phylogeny</td>
<td>availability of a reference genome is a plus</td>
<td>0.3-3 µg (quantified by an intercalating dye) free of RNA and contaminants</td>
<td>173 € x 16 samples = 2768 € (incl. library preparation, NGS run)</td>
<td>1-2 weeks of library preparation queue for outsourced NGS run: 1-2 months</td>
<td>raw data: 3.1 Gb (reads of 2 x 250 bp) processed dataset: 10 Mb = 650 kb x 16 samples (2714 loci &amp; 21K SNPs) missing data: 64% of nucleotides</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>anchored phylogenomics based on hybrid enrichment → deeper phylogeny</td>
<td>availability of several reference genomes for the taxonomic group of interest. set of probes hybridizing with selected nuclear regions. Kit for vertebrates by Lemmon et al. (2012).</td>
<td>0.1-2 µg (quantified by Qubit) free of RNA (alien DNA is less critical here)</td>
<td>220 € x 15 samples = 3300 € (incl. enrichment, library preparation, NGS run, data filtering and assembly)</td>
<td>2 days of DNA extraction and quantification queue for outsourced NGS run: 1-4 months</td>
<td>raw data: 100 Gb (reads of 2 x 150 bp) processed dataset: ~9 Mb = 390 loci x 1.6 kb (~assembly size) x 15 samples (17K SNPs) missing data: 1.5% of nucleotides</td>
</tr>
</tbody>
</table>

DISCUSSION

+ NGS is useful to distinguish sequences of the host from those of the parasite (full genome available for Schistosoma).
+ Compared to other NGS approaches, GBS allows a better sequencing depth and the analysis of more specimens but - provides data for less loci and PCR duplicates can not be removed.
+ RAD is a genome-wide exploratory tool, providing a higher proportion of homologous sequences for specimens that are more closely related.
- Here, the high proportion of missing data in the final dataset is due to both: • a considerable divergence between some specimens and • a less successful sequencing of some samples.
+ This approach is able to capture a set of nuclear loci throughout the genome showing various substitution rates.
+ It produces a large dataset with limited missing data.
- The set of markers to capture has to be optimized and necessitates the availability of several full genomes for the taxonomic group that is investigated.

CONCLUSION

Even if it is tempting to explore all possibilities offered by NGS, technology-driven research projects applied to non-model organisms risk to deliver a large amount of data that cannot be interpreted reliably. Here we chose to minimize the cost and optimize the expected dataset, not only in number of markers and samples but also according to the data already available for the organisms under study.

REFERENCES