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

Sonet Gontran[1], Smitz Nathalie[2], Virgilio Massimiliano[2], Nagy Zoltán T.[1], Winant Virginie[2], Huyse Tine[2], Backeljau Thierry[1,3] and De Meyer Marc[2]

[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.

Joint Experimental Molecular Unit
Africa Tervuren JEMU museum
GORA-MOLCOL
belspo

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).

Project	*Bulinus* sp. ©Kane et al. 2008	*Ceratidis* sp.©NHM 2000 (USNM specimen)	*Helophis schoutedeni* ©Vaclav Gvozdik
Taxonomic group	snail (Planorbidae: *Bulinus truncatus* and *B. globosus*)	fruitfly (Tephritidae: *Ceratitis fasciventris, C. anonae, C. rosa,*)	snake (Colubridae: subfamily Natricinae)
Objective	population genetics of 2 intermediate hosts of the human blood fluke (*Schistosoma*)	resolving the species complex of fruit pest species (the "FAR" species complex)	phylogeny of water snakes
NGS approach → general application	genotyping by sequencing (GBS) → population genetics	restriction-site associated DNA tags (RAD) → pop. genetics & shallow-level phylogeny	anchored phylogenomics based on hybrid enrichment → deeper phylogeny
Requirements	preliminary optimisation (choice of restriction enzyme, optimal ratio adapter/DNA) availability of a reference genome is a plus	availability of a reference genome is a plus	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).
Starting DNA material	0.3-3 µg (quantified by an intercalating dye) free of RNA and contaminants	0.3-3 µg (quantified by an intercalating dye) free of RNA and contaminants	0.1-2 µg (quantified by Qubit) free of RNA (alien DNA is less critical here)
Price per sample x number of samples processed	43 € x 192 samples = 8256 € (incl. library preparation, NGS run, SNP calling)	173 € x 16 samples = 2768 € (incl. library preparation, NGS run)	220 € x 15 samples = 3300 € (incl. enrichment, library preparation, NGS run, data filtering and assembly)
Timing (excl. data analysis)	2 weeks of library preparation queue for outsourced NGS run: 2-4 months	1-2 weeks of library preparation queue for outsourced NGS run: 1-2 months	2 days of DNA extraction and quantification queue for outsourced NGS run: 1-4 months
Output	raw data: 40 Gb (reads of 1 x 100 bp) processed dataset: 200K SNPs (expected)	raw data: 3.1 Gb (reads of 2 x 250 bp) processed dataset: 10 Mb = 650 kb x 16 samples (2714 loci & 21K SNPs) missing data: 64% of nucleotides	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

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

Kane RA, Stothard JR, Emery AM & Rollinson D (2008) **Molecular characterization of freshwater snails in the genus *Bulinus*: a role for barcodes?** Parasit Vectors 1: 1-15.
Lemmon EM & Lemmon AR (2013) **High-throughput genomic data in systematics and phylogenetics.** Annu Rev Ecol Evol Syst 44: 99-121.
Lemmon AR, Emme SA & Lemmon EM (2012) **Anchored hybrid enrichment for massively high-throughput phylogenomics.** Syst Biol 61: 727-744.