Empirical Assessment of RAD Sequencing for Interspecific Phylogeny

Astrid Cruaud,*[†,1] Mathieu Gautier,[†,1,2] Maxime Galan,[1] Julien Foucaud,[1] Laure Sauné,[1] Gwenaëlle Genson,[1] Emeric Dubois,[3] Sabine Nidelet,[3] Thierry Deuve,[4] and Jean-Yves Rasplus[1]

[1]INRA, UMR1062 CBGP, Montferrier-sur-Lez, France

[2]Institut de Biologie Computationnelle, 95 rue de la Galéra, 34095 Montpellier, France

[3]Montpellier GenomiX, c/o Institut de Génomique Fonctionnelle, Montpellier, France

[4]MNHN, UMR7205 OSEB, Muséum National d'Histoire Naturelle, Paris, France

[†]These authors contributed equally to this work.

***Corresponding author:** E-mail: cruaud@supagro.inra.fr.

Associate editor: Emma Teeling

Abstract

Next-generation sequencing opened up new possibilities in phylogenetics; however, choosing an appropriate method of sample preparation remains challenging. Here, we demonstrate that restriction-site-associated DNA sequencing (RAD-seq) generates useful data for phylogenomics. Analysis of our RAD library using current bioinformatic and phylogenetic tools produced 400× more sites than our Sanger approach (2,262,825 nt/species), fully resolving relationships between 18 species of ground beetles (divergences up to 17 My). This suggests that RAD-seq is promising to infer phylogeny of eukaryotic species, though potential biases need to be evaluated and new methodologies developed to take full advantage of such data.

Key words: **phylogenomics, next-generation sequencing, restriction-site associated DNA sequencing, empirical data, Insecta, Carabidae.**

Brief Communication

Next-generation sequencing opened up new possibilities for phylogenetics, allowing rapid and cost-effective generation of millions of reads from different loci on nonmodel species; however, choosing an appropriate method of library construction remains challenging. Consequently, Sanger sequencing of a few genes is still widely used to infer species phylogenies (McCormack et al. 2013). Sequencing complete eukaryotic genomes to infer species phylogenies remains expensive, time consuming, and unrealistic. Sequencing mitochondrial genomes is more affordable, but trees can be misleading due to introgression and heteroplasmy. Thus, methods based on reduced representations of the nuclear genome appear most adequate to generate large amounts of data across many individuals at reasonable costs.

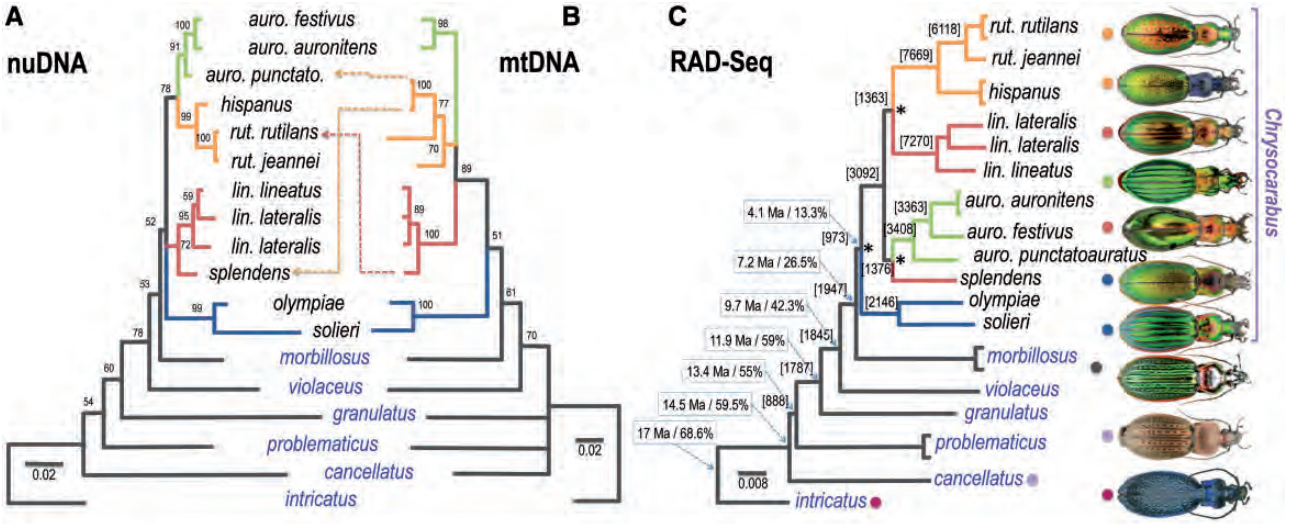
Restriction-site-associated DNA sequencing (RAD-seq, Baird et al. 2008; supplementary fig. S1, Supplementary Material online) has been used to infer the recent evolutionary history (<3 My) of few organisms (e.g., Jones et al. 2013; Nadeau et al. 2013). However, with increasing genetic distances, mutations in the restriction sites may reduce the number of orthologous loci, making RAD-seq inappropriate to infer deeper relationships (McCormack et al. 2012). Conversely, recent in silico studies suggested that a sufficient number of markers could be obtained from distant species (up to 60 My old, Rubin et al. 2012). Here, we empirically tested this prediction by comparing the power of Sanger and RAD sequencing approaches to resolve relationships between 18 nonmodel species of ground beetles (*Carabus*), whose divergences ranged from 1.2 to 17 My (supplementary table S1, Supplementary Material online).

Details regarding data generation and analysis are described in the supplementary materials, Supplementary Material online.

Sanger—After a lengthy process of screening loci for variability, primer design and PCR optimization, maximum likelihood (ML) analyses of sequences from three mitochondrial and six nuclear markers led to poorly resolved and conflicting topologies, highlighting possible mitochondrial introgression between species (fig. 1*A* and *B*).

RAD-seq—After 4 days of library preparation following the protocol by Etter et al. (2011) (*Pst*I enzyme), 2 weeks of sequencing on one lane of a HiSeq 2000 flowcell and a week of data processing on a standard computer (using Stacks; Catchen,et al. 2011), we obtained 400× the volume of Sanger data and 270× more informative sites. Our data set resulted from a stringent loci selection to ensure for homology and minimize the amount of missing data. More than half of the individuals should have sequences for a loci to be included in our analysis, and the number of mismatches allowed when merging loci from all individuals varied from 4 to 10 (parameter *n*, cstacks)

Whatever the value of *n*, ML analysis of the RAD-seq data sets (up to 25,425 loci, i.e., one every 12,000 nt, supplementary table S2, Supplementary Material online) produced the same fully resolved topology (fig. 1C and supplementary fig. S2, Supplementary Material online),



D

	Sanger		RAD ($n_{cstacks}$ = 10)	
Matrix :	*Carabus* (18 ind./13 spp.)	*Chrysocarabus* (12 ind./7 spp.)	*Carabus* (18 ind./13 spp.)	*Chrysocarabus* (12 ind./7 spp.)
Number of loci :	9		25 425	
Alignment length :	5 584 bp		2 262 825 bp	
Data set size :	100 512 nt	67 008 nt	40 730 850 nt	27 153 900 nt
Variable sites :	1030 (18.4%)	535 (9.6%)	367 813 (16.3%)	216 300 (9.6%)
Informative sites :	495 (8.9%)	264 (4.7%)	132 186 (5.8%)	91 420 (4.0%)
Missing data :				
without old DNA sample	21.8%	17.9%	26.9%	13.3%
of old DNA sample (*C. olympiae*)	65.5%		75.4%	
Sample preparation and sequencing :	1 week / ca. 1 200 USD (without marker screening)		3 weeks / ca. 3 600 USD	
Data Analysis (CBGP cluster) :	2-3 days		1-2 weeks	

Fig. 1. ML phylogenies (RAxML 7.2.8-ALPHA, Stamatakis 2006; 10,000 bootstrap replicates) (*A*) from the six nuclear (3,659 bp) and (*B*) three mitochondrial (1,925 bp) markers sequenced using Sanger technology. Nodes with bootstrap supports (BP) < 50% are collapsed (see also supplementary fig. S4, Supplementary Material online). Dashed arrows show possible cases of mitochondrial introgression. (*C*) ML phylogeny (RAxML, 10,000 bootstrap replicates, $n_{cstacks} = 10$) inferred from the RAD data sets (the same topology was recovered whatever the value of *n*, see supplementary fig. S2, Supplementary Material online). BP = 100 for all nodes but three (stars) for which BP = 99. Synapomorphies supporting nodes as calculated with MacClade 4.08 (Maddison DR and Maddison WP 2005) are between brackets. Node mean ages, taken from Deuve et al. (2012), and missing data are given as follows: ages/percentage of missing data. (*D*) Comparison between Sanger and RAD-seq data sets. Percentages of missing data were obtained using Geneious 6.1.6. (Drummond et al. 2010). Percentages of variable and informative sites were calculated using MEGA 5.2.2 (Tamura et al. 2011). See details for all data sets in supplementary table S2, Supplementary Material online.

different from the nuclear and mitochondrial Sanger trees. These two last topologies were rejected by statistical tests ($P < 0.001$). Although significant loss of RAD markers occurred for the oldest DNA sample (78.3%–75.4%, *Carabus olympiae*, extraction performed in 1998), enough signal remained for its placement. Loss of RAD markers also occurred with increasing genetic distances, though enough information was retained to accurately resolve the relationships within *Carabus* (Deuve et al. 2012). When *n* was set to 10, more polymorphic loci were included and missing data reached 68.6% for a divergence of 17 My (fig. 1C and supplementary fig. S2, Supplementary Material online). Preliminary tests are thus recommended to optimize the amount of allowed mismatches within a RAD locus to avoid loci overkill while ensuring loci homology across individuals. Nevertheless, percentages of missing data were comparable between Sanger and RAD-seq data sets (fig. 1D, supplementary table S2, Supplementary Material online). Furthermore, missing data did not bias reconstruction, as topologies obtained from the complete matrices (loci shared by all 18 species) were similar to those inferred from the incomplete matrices (supplementary fig. S3, Supplementary Material online).

Overall, this study illustrates the power of RAD-seq to infer shallow relationships, and we believe that this approach may be generalized to many groups (~90% of the insect genomes range between 100 and 800 Mb, supplementary fig. S5, Supplementary Material online). Here, we used only one partition and relied on current evolutionary models to analyze RAD-seq data. Testing alternative partitioning schemes or developing more appropriate models to deal with the specificities of such data (e.g., modeling gain/loss of restriction sites) might be promising to improve inferences.

Finally, RAD-seq definitely opens new avenues for phylogeneticists but possibly also a Pandora's box of analytical issues. These issues will need to be explored to avoid being

misled by systematic (Lemmon EM and Lemmon AR 2013) or other biases (Arnold et al. 2013), which may differ depending on the level of genetic divergence among samples.

Supplementary Material

Supplementary figures S1–S5 and tables S1 and S2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org.)

Acknowledgments

The authors thank the SeqGen plateform (CeMEB Labex, France) for facilitating access to the Covaris S220 instrument and the CBGP HPC computational platform on which analyses were performed. The authors also thank E. Artige (CBGP) for her help on the field, I. Meusnier (CBGP) for her assistance for lab work, P.A. Gagnaire for P1 adapters, K. Gharbi and C. Eland (GenePool, UK) for their assistance with transferring RAD-seq protocol to CBGP and two anonymous reviewers for their valuable comments on a previous version of this manuscript. This work was based upon financial support received from the division "Plant Health and the Environment" of the French National Institute for Agricultural Research (INRA) and from the network "Library of Life" funded by the National Center for Scientific Research (CNRS), the National Museum of Natural History (MNHN), the French National Agronomic Institute (INRA), and the French Alternative Energies and Atomic Energy Commission (CEA) (Genoscope). Sanger sequences were deposited in GenBank under accession numbers KJ158754-KJ158836. RAD-seq data sets can be downloaded from http://www1.montpellier.inra.fr/CBGP/NGS/Files/Datasets_Cruaud_et_al_2014_RAD_MBE.zip.

References

Arnold B, Corbett-Detig RB, Hartl D, Bomblies K. 2013. RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Mol Ecol.* 22:3179–3190.

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* 3:e3376.

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH. 2011. Stacks: Building and genotyping loci de novo from short-read sequences. *G3* 1:171–182.

Deuve T, Cruaud A, Genson G, Rasplus J-Y. 2012. Molecular systematics and evolutionary history of the genus *Carabus* (Col. Carabidae). *Mol Phylogenet Evol.* 65:259–275.

Drummond AJ, Ashton B, Buxton S, Cheung M, Cooper A, Heled J, Kearse M, Moir R, Stones-Havas S, Sturrock S, et al. 2010. Geneious v6.1.6. San Francisco (CA): Biomatters Inc. [cited 2014 Feb 17]. Available from: http://www.geneious.com.

Etter PD, Bassham S, Hohenlohe PA, Johnson EA, Cresko WA. 2011. SNP discovery and genotyping for evolutionary genetics using RAD sequencing. In: Orgogozo V, Rockman MV, editors. Molecular methods for evolutionary genetics. New York: Humana Press. p. 157–178.

Jones JC, Fan SH, Franchini P, Schartl M, Meyer A. 2013. The evolutionary history of *Xiphophorus* fish and their sexually selected sword: a genome-wide approach using restriction site-associated DNA sequencing. *Mol Ecol.* 22:2986–3001.

Lemmon EM, Lemmon AR. 2013. High-throughput genomic data in systematics and phylogenetics. *Annu Rev Ecol. Evol Syst.* 44:99–121.

Maddison DR, Maddison WP. 2005. MacClade 4: Analysis of phylogeny and character evolution. Version 4.08a. Sunderland (MA): Sinauer Associates [cited 2014 Feb 17]. Available from: http://macclade.org.

McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT. 2013. Applications of next-generation sequencing to phylogeography and phylogenetics. *Mol Phylogenet Evol.* 66:526–538.

McCormack JE, Maley JM, Hird SM, Derryberry EP, Graves GR, Brumfield RT. 2012. Next-generation sequencing reveals population genetic structure and a species tree for recent bird divergences. *Mol Phylogenet Evol.* 62:397–406.

Nadeau NJ, Martin SH, Kozak KM, Salazar C, Dasmahapatra KK, Davey JW, Baxter SW, Blaxter ML, Mallet J, Jiggins CD. 2013. Genome-wide patterns of divergence and gene flow across a butterfly radiation. *Mol Ecol.* 22:814–826.

Rubin BER, Ree RH, Moreau CS. 2012. Inferring phylogenies from RAD sequence data. *PLoS One* 7:e33394.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688–2690.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol.* 28:2731–2739.