Next generation sequencing for characterizing biodiversity: promises and challenges

François Pompanon, Sarah Samadi

Genetica DOI 10.1007/s10709-015-9816-7

Next Generation Sequencing for characterizing biodiversity: Promises and challenges

François Pompanon[1, 2] and Sarah Samadi[3]

[1] Univ. Grenoble Alpes, Laboratoire d'Ecologie Alpine, F-38000 Grenoble, France

[2] CNRS, Laboratoire d'Ecologie Alpine, F-38000 Grenoble, France

[3] Muséum national d'Histoire naturelle, UMR7205 CNRS-EPHE-MNHN-UPMC, ISyEB, CP26, 57 rue Cuvier, F-75231 Paris Cedex 05

Summary

DNA barcoding approaches are used to describe biodiversity by analysing specimens or environmental samples in taxonomic, phylogenetic and ecological studies. While sharing data among these disciplines would be highly valuable, this remains difficult because of contradictory requirements. The properties making a DNA barcode efficient for specimen identification or species delimitation are hardly reconcilable with those required for a powerful analysis of degraded DNA from environmental samples. The use of Next Generation Sequencing methods open up the way towards the development of new markers (e.g., multilocus barcodes) that would overcome such limitations. However, several challenges should be taken up for coordinating actions at the interface between taxonomy, ecology, molecular biology and bioinformatics in order to develop methods and protocols compatible with both taxonomic and ecological studies.

Key-words: DNA barcoding, Next Generation sequencing, taxonomy, metabarcoding, metagenomics

The large exploratory projects based on DNA sequencing that began in the early 1990s mainly focused on genomics and its medical applications (*e.g.* complete human genome and the 1000 genomes project, The 1000 Genomes Project Consortium 2010). Besides this scope, the Barcode of Life initiative (BoL, http://www.barcodeoflife.org) aimed at describing the worldwide diversity of species. The idea was to develop a universal diagnostic tool that could be used for identifying species with a variety of applications (*e.g.* ecology, agronomy, forensic) and could also speed up the description of unknown biodiversity (Hebert et al. 2003). Thus, DNA barcoding relies on the characterization of a standard genomic region (i.e. the DNA barcode) that can be compared to that of reference specimens identified by taxonomists. This involves the constitution of a DNA-library linked to collections of specimen-vouchers available for identification by taxonomists (Puillandre et al. 2012). Ideally, the range of specimens documented in the database should also covers type-specimens in order to ensure the use of adequate nomenclatures including when needed the revision of the classification (Puillandre et al. 2011). The standard barcode defined for animals is the *COI* gene (Hebert et al. 2003), while *rbcL* and *matK* are used for plants (CBOL Plant Working Group 2009). Ecologists quickly used the DNA barcoding approach but frequently with other (i.e., non standard) barcodes that better suit the characterisation of degraded DNA (Valentini et al. 2009). This allowed the recent development of DNA metabarcoding (Taberlet et al. 2012), which is the automated identification of species from a single environmental sample (e.g., water, soil, faeces). Researchers from different disciplines are taking advantage of the barcoding approach using different DNA barcodes to address questions related to ecology, phylogeny or taxonomy. As an outcome, the results obtained by one discipline are hardly exploitable by the others, while we expect that sharing data would be highly valuable. For example, databases relating the genetic information to taxa might be used as references when characterising ecosystem biodiversity; new taxa discovered in ecological surveys should be taken into account quickly in phylogenetic and taxonomic studies. Thus, the development of methods compatible for systematics and ecology appears to be crucial.

The recent developments of DNA barcoding *sensu stricto* mainly dealt with the completion of taxonomic databases and the definition of markers taxonomically pertinent (e.g. Kwong et al. 2012; Kvist et al. 2013), while the approaches relying on the characterisation of environmental samples (e.g. metagenomics, metabarcoding) were taking advantage of Next Generation Sequencing (NGS, Shendure and Ji 2008) to describe the diversity of complex (i.e. multispecies) samples. The efficiency of all these approaches is currently limited by the completion of taxonomic reference databases and by the choice of standard barcodes with sufficient taxonomic coverage and resolution. Selecting DNA-barcodes requires making a compromise between technical constraints and the quality of phylogenetic signal over the targeted range of diversity. However the efficiency of the selected DNA-barcode for species identification also depends on the completeness of the databases for the targeted taxa at the targeted taxonomic rank. In many applications of DNA-barcoding availability of data over an adequate taxonomic coverage is a major limitation. For example Puillandre et al. (2009) showed that the failure of identification of gastropods egg-masses using DNA-barcodes mainly results from the poor taxonomic coverage of DNA-sequences databases for the most diverse families of gastropods. Indeed most of the genetic data are gathered from databases such as GenBank in which, in addition to the not obligatory link to voucher specimens, the sampling scheme does not reflect the biological classifications (figure 1A) but the research efforts that are strongly biased toward model organisms (Figure 1B). Similarly the genetic data available across taxa in such databases are also not standardized. Figure 1B illustrates the biases resulting from the absence of a taxonomic sampling scheme in the data used for identification through DNA-barcodes. Some taxa are overrepresented because they are model organisms, while large compartments are not at all represented. Whatever the identification method used, the absence of a hierarchical sampling scheme, corresponding to the accepted classification, only allow the identification of specimens belonging to those well-studied taxa and the data do not allow recovering the classifications. The Barcoding of life project aims at standardizing the markers across taxa but also at covering the taxon diversity at the species rank. Ideally, such databases would cover the hierarchical classification with

replications of data at each classification rank (i.e. several specimens sequenced within each species, several species within each genus, several (if not all) genera per family etc …). One limitation of the Barcoding of life project is the small number of genetic markers documented in the database. The resolution of such a few number of markers would hardly both allow an identification at the species level and at deeper classification ranks. As a consequence, relationship or identification inferred from a unique marker may be distorted (illustrated in Figure 1C). The new sequencing methods potentially allow recovering several barcodes for each sample. Combining such sequencing methods with a taxonomic sampling that covers the hierarchical classification potentially allow identification at every taxonomic rank (Figure 1D). However, the constitution of such databases will also probably lead to question accepted classification (as figured in the Figure 1D). There are yet only few examples of such approaches. Among the rare available studies, (Cruaud et al. 2014) showed the potential of RAD-sequencing to resolve phylogeny for divergence up to 17My that were poorly resolved with a standard approach with 9 genetic markers. This study illustrates the loss of orthologous loci among species in correlation with the depth of the divergence (as drawn in Figure 2). This study shows the power of the genome-wide approaches for systematics with potential application for identification purposes in other fields. Indeed, if such studies are designated to adequately cover the classification, they will not only allow the revision of accepted classifications but also will offer a bulk of potential small DNA-barcodes for identification purposes in other field than systematics.

More generally, the use of NGS would allow developing markers compatible for systematics and ecology. This has been a difficulty until now because the properties making a DNA barcode efficient for taxonomic studies are hardly reconcilable with those making it powerful for the analysis of ecological samples (Valentini et al. 2009). For example, short barcodes are required for a robust analysis of degraded DNA from environmental samples, while longer barcodes from independently evolving parts of the genome are required to get more taxonomic information and phylogenetic signal. The huge number of sequences generated by NGS make now possible to characterize multilocus barcodes (i.e., combinations of several short barcodes) allowing more comprehensive genome surveys (i.e., covering a larger proportion of genomes). Using a combination of short fragments would lead to a higher resolution even from environmental samples or dried specimen from museum collection with degraded DNA. For example, multiplexed barcodes are used for assessing omnivorous diets from faecal samples (De Barba et al. 2013). Such approach would reveal enough genetic variability to get enough taxonomic resolution and phylogenetic signal, and the results produced would be exploitable by all disciplines.

Another difficulty of the current barcoding approaches is the characterization by sequencing following a PCR amplification of the DNA fragment. The PCR generates biases and errors (Coissac et al. 2012), and impose the definition of DNA barcodes in regions surrounded by highly conserved priming sites. Several NGS-based approaches would allow overcoming these problems. DNA capture methods (e.g., Hodges et al. 2007, 2009; Avila-Arcos et al. 2011) select target fragments without PCR using probes complementary to a conserved region within the barcode region of interest. This approach was successfully conducted to retrieve partial to full-length mitochondrial genomes from degraded DNA of museum specimens (Mason et al. 2011). Hancock-Hanser et al. (2013) showed that nuclear loci can also be captured, and that an efficient cross-species capture is possible provided sequences are less than 12% divergent. A second approach relies on the use of Restriction site Associated DNA Sequences (RAD-Seq, Baird et al. 2008). While these methods cannot be used for studying degraded DNA from environmental samples, they could be powerful for delimiting species and building phylogenies (Cariou *et al.* 2013; Cruaud et al 2014; Viricel et al. 2014) (Figure 2). Another possibility is the shotgun sequencing of environmental samples, which would allow assembling the complete genomes of organelles (i.e., mitochondrion and chloroplast) together with the sequence of multiple copy genes such as ribosomal RNA. These data are already used for building large scale phylogenies (e.g., Roquet et al. 2013) and could be produced for characterising environmental samples. Moreover, besides the information obtained from target regions, this method produces millions of short nuclear sequence reads that contain information usable for taxonomic or functional

assignment. This metagenomic approach (e.g., Tringe et al. 2005) can then be applied for characterizing simultaneously the taxonomic and functional diversity from environmental samples (Edwards et al. 2013, see also Taberlet et al. 2012).

The use of NGS is not free of technical constraints that are related, for example, to the difficulty of producing quantitative data from environmental samples (Pompanon et al. 2012) or to PCR and sequencing errors. Potential biases due to the production of errors should be considered when analysing data and setting up the experiment (Coissac et al. 2012; Pompanon et al. 2012), for example by choosing the number of loci and specimens allowing an adequate coverage for each sequenced fragment and by running programs dedicated to the removal of erroneous sequence reads.

Moreover, some conceptual and methodological limitations to the use of NGS exist. The potential use of multilocus barcodes would allow a less simplistic definition of taxa based on a larger number of independent characters, but we expect the persistence of a discrepancy between taxa, because complete reference genomes will be lacking for large parts of the tree of life for still a long time. A second limitation is related to the homology of markers between taxa and then to the depth of the phylogenetic signal. A part of the data produced by NGS approaches will not be standard. For example, the millions of short reads produced in a shotgun NGS will represent a variable part of a genome that is not always comparable among taxa. The treatment of such data in comparative studies is not trivial and testing hypotheses on the homology of markers would require a pertinent taxonomic sampling according to the phylogenetic depth considered. Another limitation is due to the impact of missing data in reference databases on the quality of taxonomic or phylogenetic inferences. This impact, which is already important when using standard barcodes, is increased with NGS approaches. All methods used for specimen identification or species discovery/delimitation require a good estimate of within- and between-groups variability. Thus it is important to have a good representativeness of between groups variability (e.g., all species of a genus) and of within-group variability (e.g., covering the whole distribution area of a species with, ideally, several individuals per population). This involves setting up sequencing strategies based on the analysis of large sample size. A problem is that the methods currently used for taxon identification or delimitation are either monolocus (e.g. Pons et al 2006; Puillandre et al 2011; Birky 2013) and adapted to the analysis of large datasets, or multilocus (e.g. O'Meara 2010; Carstens and Dewey 2010) but limited to the analysis of a reduced number of loci and specimens. It is now necessary to develop methods for analysing efficiently many markers for many individuals.

In this context, two major challenges should be taken up. On one hand, we should complete reference databases by increasing their taxonomic coverage. This should be the case for standard barcodes, standard regions such as complete mitochondrial genomes (Dettai et al. 2012) but also non-standard genomic data (partial genomes, RAD-Seq data, etc.). On the other hand, the potential of NGS approaches relies on the development of methods of taxonomic assignment and phylogenetic inference taking into account sequencing errors, DNA degradation (e.g. from environmental samples or museum specimen), and exploiting the information from standard markers but also from partial genomic data (Coissac et al. 2012). Taking up these challenges requires coordinating actions at the interface between taxonomy, ecology, molecular biology and bioinformatics for developing methods and protocols compatible with both taxonomic and ecological studies. This will also involve the development of reference databases coherent with NGS approaches and making the link between genetic data, reference specimens and long-term DNA collections that could be used in the future for the development of new ecological studies.

Acknowledgements

This article results from a think tank “Prospective Genomique environmentale” inititated by the French Centre National de la Recherche scientifique. The authors thank Eric Coissac, Régis Debruyne, Frédéric Delsuc, Catherine Hänni, Sébastien Lavergne, Morgane Ollivier, Eric Pante, Nicolas Puillandre, Jean-Yves Rasplus and Pierre Taberlet for providing food for thought, and Régis Debruyne for help in making Figure 2.

Conflict of interest

The authors declare that they have no conflict of interest

References

Avila-Arcos MC, Cappellini E, Romero-Navarro JA, Wales N, Moreno-Mayar JV, Rasmussen M, Fordyce SL, Montiel R, Vielle-Calzada J-P, Willerslev E, Gilbert MTP (2011) Application and comparison of large-scale solution-based DNA capture-enrichment methods on ancient DNA. Scientific Reports 1: 74

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

Birky Jr CW (2013) Species detection and identification in sexual organisms using population genetic theory and DNA sequences. PloS one 8: e52544

Cariou M, Duret L, Charlat S (2013) Is RAD-seq suitable for phylogenetic inference? An *in silico* assessment and optimization. Ecol Evol 3: 846–85. doi:10.1002/ece3.512

Carstens BC, Dewey TA (2010) Species delimitation using a combined coalescent and information-theoretic approach: an example from North American *Myotis* bats. Systematic Biology, 59: 400-414

CBOL Plant Working Group (2009) A DNA barcode for Land Plant. PNAS 106:12794-12797

Coissac E, Riaz T, Puillandre N. (2012) Bioinformatic challenges for DNA metabarcoding of plants and animals. Mol Ecol 21:1834–1847

Cruaud A, Gautier M, Galan M, Foucaud J, Sauné L, Genson G, Dubois E, Nidelet S, Deuve T, Rasplus JY. (2014) Empirical Assessment of RAD Sequencing for Interspecific Phylogeny. Mol Biol Evol, 31: 1272-1274

De Barba M, Miquel C, Boyer F, Mercier C, Rioux D, Coissac E, Taberlet P (2013) DNA metabarcoding multiplexing and validation of data accuracy for diet assessment: application to omnivorous diet. Mol Ecol Res 14: 306-323

Dettai A, Gallut C, Brouillet S, Pothier J, Lecointre G, Debruyne R (2012) Conveniently pre-tagged and pre-packaged: extended molecular identification and metagenomics using complete metazoan mitochondrial genomes. PloS one 7:e51263

Edwards A, Pachebat JA, Swain M, Hegarty M, Hodson AJ, Irvine-Fynn TDL, Rassner SME, Sattler B (2013) A metagenomic snapshot of taxonomic and functional diversity in an alpine glacier cryoconite ecosystem. Environ Res Lett 8: 035003

Hancock-Hanser BL, Frey A, Leslie MS, Dutton PH, Archer FI, Morin PA (2013) Targeted

multiplex next-generation sequencing: advances in techniques of mitochondrial and nuclear DNA sequencing for population genomics. Mol Ecol Res 13: 254-268

Hebert PD, Cywinska A, Ball SL (2003) Biological identifications through DNA barcodes. Proc Royal Soc London Series B: Biological Sciences 270:313-321

Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW, Middle CM, Rodesch MJ, Albert TJ, Hannon GJ, McCombie WR (2007) Genome-wide in situ exon capture for selective resequencing. Nature Genetics 39:1522–1527

Hodges E, Rooks M, Xuan ZY, Bhattacharjee A, Benjamin Gordon D, Brizuela L, Richard McCombie W, Hannon GJ (2009) Hybrid selection of dis- crete genomic intervals on custom-designed

microarrays for mas- sively parallel sequencing. Nature Protocols 4: 960–974
Kvist S (2013) Barcoding in the dark?: a critical view of the sufficiency of zoological DNA barcoding databases and a plea for broader integration of taxonomic knowledge. Molecular Phylogenetics and Evolution 69: 39–45.
Kwong S, Srivathsan A, Meier R (2012) An update on DNA barcoding: low species coverage and numerous unidentified sequences. Cladistics 28: 639-644.
Mason VC, Li G, Helgen KM, Murphy WJ (2011) Efficient cross-species capture
hybridization and next-generation sequencing of mitochondrial genomes from noninvasively sampled museum specimens. Genome Research 21 : 1695-704
O'Meara BC (2010) New heuristic methods for joint species delimitation and species tree inference. Syst Biol 59: 59-73
Pompanon F, Deagle BE, Symondson WOC, Brown DS, Jarman SN, Taberlet P (2012). Who is eating what: diet assessment using next generation sequencing. Mol Ecol 21:1931–1950.
Pons J, Barraclough TG, Gomez-Zurita J, Cardoso A, Duran DP, Hazell S, Kamoun S, Sumlin WD, Vogler AP (2006) Sequence-based species delimitation for the DNA taxonomy of undescribed insects. Systematic Biology 55: 595-609
Puillandre N, Bouchet P, Boisselier-Dubayle MC, Brisset J, Buge B, Castelin M, Chagnoux S, Christophe T, Corbari L, Lambourdière J, Lozouet P, Marani G, Rivasseau A, Silva N, Terryn Y, Tillier S, Utge J, Samadi S (2012) New taxonomy and old collections: integrating DNA barcoding into the collection curation process. Mol Ecol Res 12:396-402
Puillandre N, Macpherson E, Lambourdière J, Cruaud C, Boisselier-Dubayle MC, Samadi S (2011) Barcoding type specimens helps to identify synonyms and an unnamed new species in *Eumunida* Smith, 1883 (Decapoda: Eumunididae). Invertebrate Systematics 25:322-333
Puillandre N, Strong E, Bouchet P, Boisselier MC, Couloux A, Samadi S. (2009) Identifying gastropod spawn from DNA barcodes: possible but not yet practicable. Mol Ecol Res 9 :1311-1321
Roquet C, Thuiller W, Lavergne S (2013) Building megaphylogenies for macroecology: taking up the challenge. Ecography 36:13-26
Shendure J, Ji H (2008) Next-generation DNA sequencing. Nature Biotechnology 26: 1035–1045
Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerlslev E (2012) Towards next-generation biodiversity assessment using DNA metabarcoding. Mol Ecol 21: 2045– 2050
The 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061-1073
Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC, Bork P, Hugenholtz P, Rubin EM (2005) Comparative metagenomics of microbial communities. Science 308: 554–557
Valentini A, Pompanon F, Taberlet P (2009) DNA barcoding for ecologists. Trends Ecol Evol 24: 110–117
Viricel A, Pante E, Dabin W, Simon-Bouhet (2014) Applicability of RAD-tag genotyping for interfamilial comparisons: empirical data from two cetaceans. Mol Ecol Res 14 : 597–605

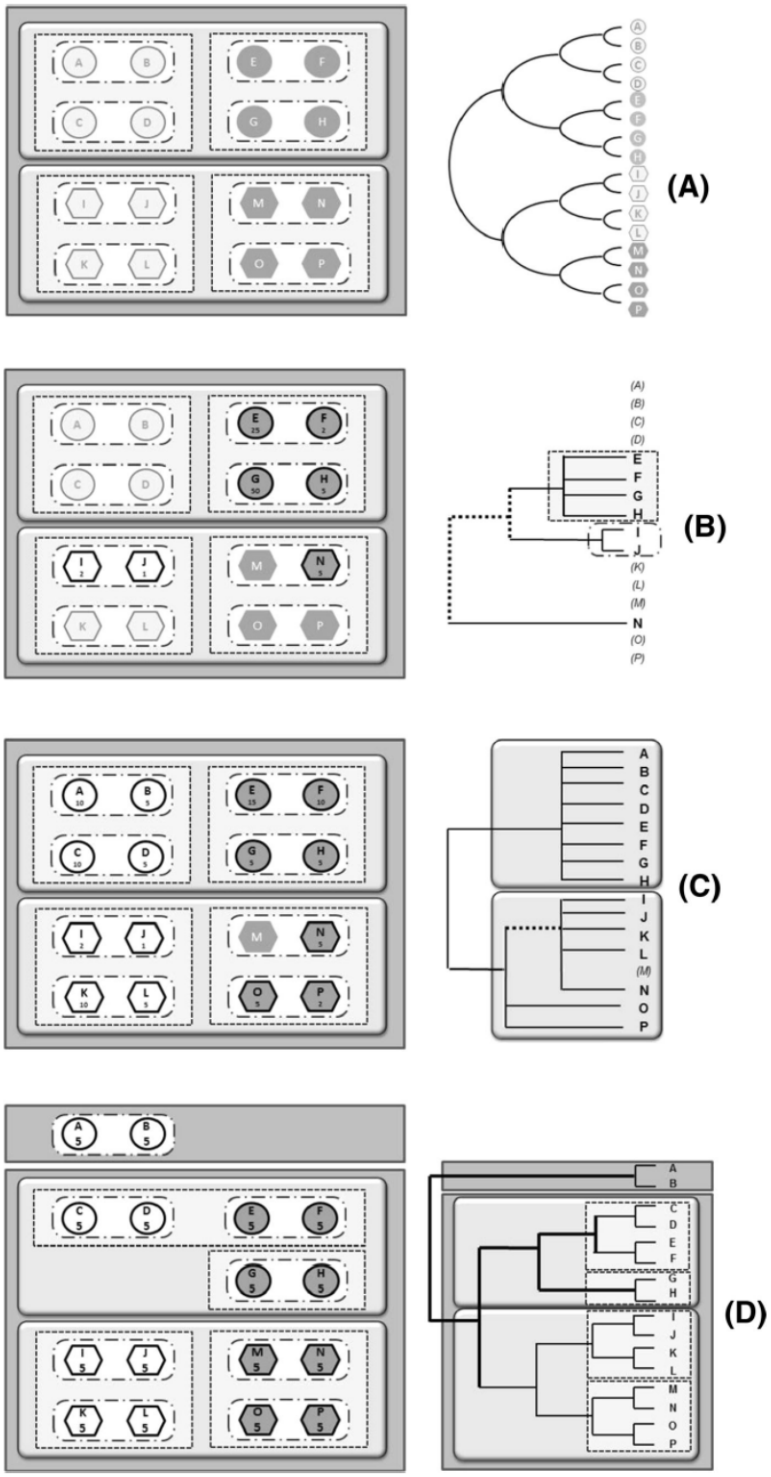

Figure 1. Sampling, DNA-barcoding and taxonomy with Next Generation sequencing.

On the left is the accepted classification with letters referring to taxa; sampled taxa are surrounded by a bold border and associated numbers refer to the number of individuals sampled. On the right are the phylogenies inferred from the DNA markers studied. (A) current classification and phylogenetic relationships expected following this classification. (B) Phylogenetic relationships inferred using data available through public data bases. Taxon-sampling is heavily distorted in favour of model organisms. The resulting tree does not allow testing the adequacy of the accepted classification. (C) Phylogenetic relationships inferred using data available through DNA-barcoding database in a case of a completed DNA-barcoding campaign. Most terminal taxa are present but for a single genetic maker. The resulting tree is not resolved for most taxonomic ranks. (D) Phylogenomic approach coupled with a DNA-barcoding taxon-sampling scheme. All taxa are covered, multiplication of markers allow resolving the tree at all phylogenetic depths. The obtained tree allows a revision of the classification. Discoveries of new taxa might still question this classification

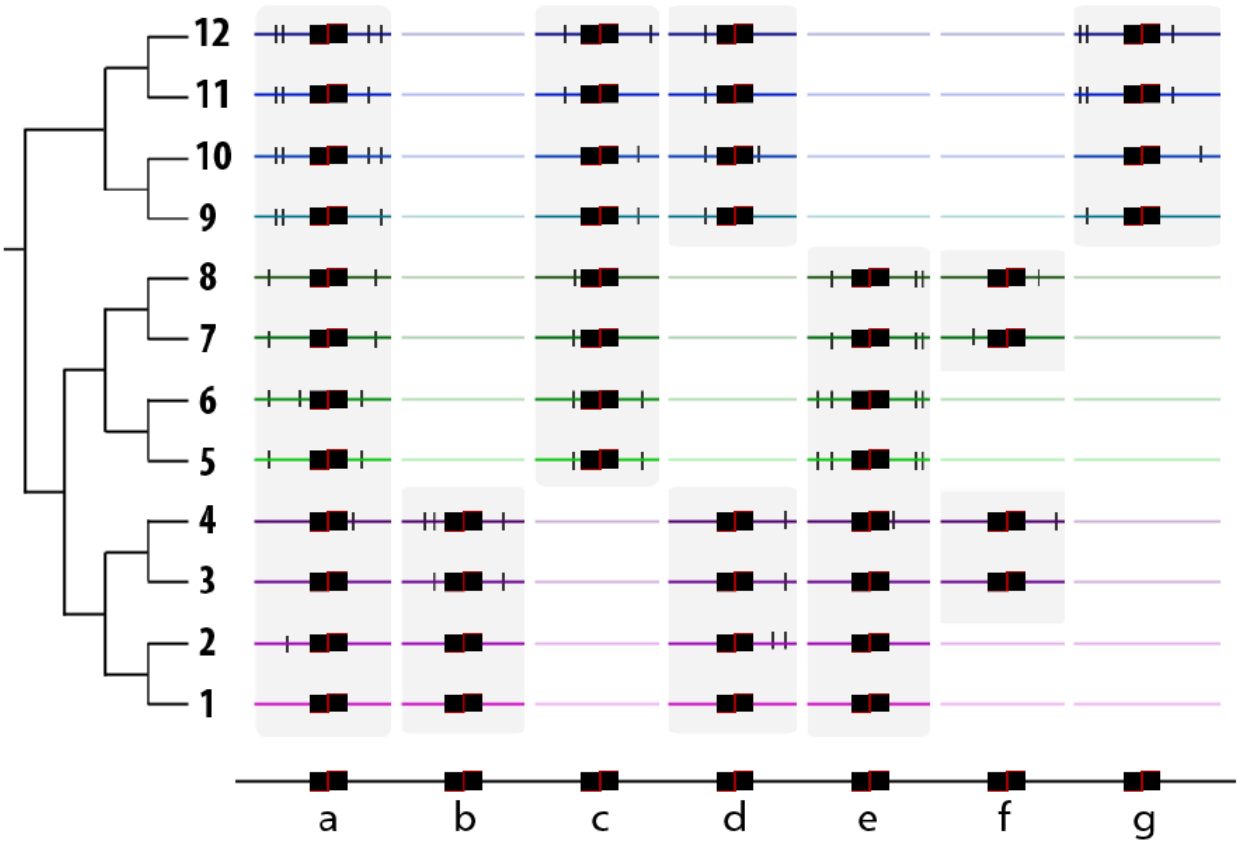

Figure 2. Use of RAD-Sequencing to resolve phylogenies.
The method generates a large number of polymorphic markers. The taxonomic sampling scheme allows inferring groups of homologous markers, and then allows defining how these markers can be used according to depth of the phylogenetic signal. For example 'a' is present and homologous for all specimens studied, while 'd' and 'g' are only informative for the first clade.