EasyAmp: a simple pipeline to genotype multilocus amplicon data from High Throughput Sequencing

Miguel Camacho-Sanchez (0000-0002-6385-7963)
Urbanización Los Maldonados 5, 41807, Espartinas, Sevilla
miguelcamachosanchez@gmail.com

ABSTRACT

I present a simple pipeline to genotype multilocus amplicon data for diploid genetic markers from one or multiple species simultaneously, from massive parallel sequencing, such as Illumina MiSeq. Genotypes are written to ready-to-use formats for downstream phylogenetic or population genetic analysis. Code is available at github.com/csmiguel/easyamp.

Keywords: amplicon genotyping, parallel tagged sequencing, targeted amplicon sequencing.

Massive parallel sequencing of amplicons has successfully been used for addressing biological questions in phylogenetics (Barrow, Ralicki, Emme, & Lemmon, 2014; Bybee et al., 2011; O'Neill et al., 2013), population genetics (Camacho-Sanchez et al., 2018) or for studying the diversity within gene families (Marmesat, Schmidt, Saveljev, Seryodkin, & Godoy, 2017). The technique relies on the amplification of target regions with specific primers and a second PCR in which individual barcodes are added to each sample (Bybee et al., 2011). This allows obtaining large amounts of data at a reduced cost using a fraction of a single high-throughput sequencing (HTS) run compared to Sanger sequencing, plus the possibility of identifying alleles at the read level (Camacho-Sanchez et al., 2018; Lemmon & Lemmon, 2013; Marmesat et al., 2017; O'Neill et al., 2013).

Genotyping amplicon data from HTS for phylogenetics and population genetics is challenging and there is no consensus effective protocol. All approaches include a demultiplexing step using the individual barcodes and a second step of read classification into loci, and sometimes the challenging task of differentiating alleles. Bybee *et al.* 2011, for a phylogeny of Pancrustacea, developed a first approach in which reads were classified into loci according to blast (Altschul, Gish, Miller, Myers, & Lipman, 1990) hits to a local database, followed by the novo assembly to determine consensus sequences. Barrow *et al.* 2013, used a similar approach in a frog phylogeny to build per-locus assemblies in which single nucleotide polymorphisms were identified and alleles were phased visually. Camacho-Sanchez *et al.* 2018, for a population genetics study of rats, mapped reads to a close reference and determined alleles visually. O'Neill *et al*. 2013, for studying the phylogenetic structure in salamanders, scripted a complex pipeline to classify reads into loci using blast, followed by multiple sequence alignments per-locus, error filtering and haplotype phasing. A more challenging application is the determination of alleles in complex gene families, such as in the MHC complex (Marmesat et al., 2017; Sommer, Courtiol, & Mazzoni, 2013). A tool called AmpliSAS, efficiently retrieves alleles in complex gene families (Sebastian, Herdegen, Migalska, & Radwan, 2016). AmpliSAS can also be used for genotyping single-copy genes (Forcina *et al.* under review). However, the complexity of gene families for which AmpliSAS was initially designed also implies more complex parameters, and output files which are not straight-ready for downstream analysis of population structure nor phylogenetics.

Here, I have developed a series of scripts which allow easy genotyping of multilocus amplicon data for diploid genetic markers (Figure 1). Users need to provide

minimal input data: primer sequences, names of loci and Illumina R1 and R2 reads already demultiplexed per sample. The outputs generated are easily interpretable (tables with genotypes and sequences in FASTA) and ready to be fed to downstream analysis of phylogenetics or population structure. They are written in the popular programming language R (R Core Team, 2020) and a some basic shell scripting, which allows users to easily customize scripts, if necessary. First, cutadapt 2.1 (Martin, 2011) is used to demultiplex and trim reads per locus. Only reads having the expected forward and reverse primers in R1 and R2 are retained. This allows removing non-targeted amplicons. Second, DADA2 (Callahan et al., 2016), an R package popularized for determining variants in metabarcoding experiments, is used to determine all possible alleles per locus. DADA2 has a high power of resolution, being able to determine reliably variants in a pool of amplicons which only differ in as little as 1 nucleotide. Samples from different species can be genotyped altogether, although if coverage is very low, allele determination in DADA2 could benefit from pooling samples from the same species in different batches. Third, a custom script removes spurious variants. That is, those not passing given thresholds of allele balance and coverage (O'Leary, Puritz, Willis, Hollenbeck, & Portnoy, 2018). Third, the resulting matrix is used to call genotypes. Current version only supports diploid data. The genotype call of an individual for a given locus is heterozygous (*ab*) when 2 alleles are present with sufficient coverage, and homozygous (*aa*) when only 1 allele with sufficient reads is present. If there is only 1 allele and the read count is low it can cause calling false homozygotes (O'Leary et al., 2018). For that reason, hemizygotes (*a-)* are called when the effective read count (details in scripts) falls below 5. In this case, according to a binomial distribution, observing at least 5 reads of one allele and 0 of any other returns at least a probability of 0.97 of that genotype call being homozygous, assuming all alleles are amplified with the same affinity. Fourth, several outputs are generated: a table with genotypes, read count per allele, all alleles in FASTA and all alleles for all samples in FASTA. Optional, further filtering can be done to prepare output data for population genetics analysis. These involve removal of loci with no data, filter individuals according to their missing data, filter loci according to their missing data, removal of monomorphic loci and removal of loci deviating from Hardy-Weinberg equilibrium. Final outputs will include filtered genotypes as a table in plain text, a *genind* object (Jombart, 2008) in R, and formatted for STRUCTURE (Pritchard, Stephens, & Donnelly, 2000) (Figure 1). Full instructions and a version of the pipeline can be found

in https://github.com/csmiguel/easyamp and in ZENODO (DOI:10.5281/zenodo.3945855).

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. doi: 10.1016/S0022-2836(05)80360-2

Barrow, L. N., Ralicki, H. F., Emme, S. A., & Lemmon, E. M. (2014). Species tree estimation of North American chorus frogs (Hylidae: Pseudacris) with parallel tagged amplicon sequencing. *Molecular Phylogenetics and Evolution*, *75*(1), 78–90. doi: 10.1016/j.ympev.2014.02.007

Bybee, S. M., Bracken-Grissom, H., Haynes, B. D., Hermansen, R. A., Byers, R. L., Clement, M. J., … Crandall, K. A. (2011). Targeted amplicon sequencing (TAS): A scalable next-gen approach to multilocus, multitaxa phylogenetics. *Genome Biology and Evolution*, *3*(1), 1312–1323. doi: 10.1093/gbe/evr106

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, *13*(7), 581–583. doi: 10.1038/nmeth.3869

Camacho-Sanchez, M., Quintanilla, I., Hawkins, M. T. R., Tuh, F. Y. Y., Wells, K., Maldonado, J. E., & Leonard, J. A. (2018). Interglacial refugia on tropical mountains: Novel insights from the summit rat (*Rattus baluensis*), a Borneo mountain endemic. *Diversity and Distributions*, *24*(9), 1252–1266. doi: 10.1111/ddi.12761

Forcina, G., Camacho-Sanchez, M., Tuh, F., Moreno, S., Lonard, J. Markers for Genetic Change. under review..

Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, *24*(11), 1403–1405. doi: 10.1093/bioinformatics/btn129

Lemmon, E. M., & Lemmon, A. R. (2013). High-Throughput Genomic Data in Systematics and Phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, *44*(1), 99–121. doi: 10.1146/annurev-ecolsys-110512-135822

Marmesat, E., Schmidt, K., Saveljev, A. P., Seryodkin, I. V., & Godoy, J. A. (2017). Retention of functional variation despite extreme genomic erosion: MHC allelic

repertoires in the Lynx genus. *BMC Evolutionary Biology*. doi: 10.1186/s12862-017-1006-z

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, *17*, 10–12. doi: http://dx.doi.org/10.14806/ej.17.1.200

O'Leary, S. J., Puritz, J. B., Willis, S. C., Hollenbeck, C. M., & Portnoy, D. S. (2018). These aren't the loci you'e looking for: Principles of effective SNP filtering for molecular ecologists. *Molecular Ecology*, (June), 3193–3206. doi: 10.1111/mec.14792

O'Neill, E. M., Schwartz, R., Bullock, C. T., Williams, J. S., Shaffer, H. B., Aguilar-Miguel, X., … Weisrock, D. W. (2013). Parallel tagged amplicon sequencing reveals major lineages and phylogenetic structure in the North American tiger salamander (*Ambystoma tigrinum*) species complex. *Molecular Ecology*, *22*(1), 111–129. doi: 10.1111/mec.12049

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945–959. doi: 10.1111/j.1471-8286.2007.01758.x

R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. Retrieved from https://www.r-project.org/

Sebastian, A., Herdegen, M., Migalska, M., & Radwan, J. (2016). Amplisas: A web server for multilocus genotyping using next-generation amplicon sequencing data. *Molecular Ecology Resources*, *16*(2), 498–510. doi: 10.1111/1755-0998.12453

Sommer, S., Courtiol, A., & Mazzoni, C. J. (2013). MHC genotyping of non-model organisms using next-generation sequencing: a new methodology to deal with artefacts and allelic dropout. *BMC Genomics*, *14*(1), 542. doi: 10.1186/1471-2164-14-542