Ferrandiz-Rovira M, Bigot T, Allainé D, Callait-Cardinal M-P, Cohas A.

Large-scale genotyping of highly polymorphic loci by next generation sequencing:

how to overcome the challenges to reliably genotype individuals?

Amplicon-based NGS data processing

The steps to obtain reliable individual genotypes are described below and are summarized in Table 3. All steps are carried out using two custom Python scripts which are freely available at https://github.com/tbigot/alFinder, accompanied by example files and dataset. A user friendly web interface (http://pbil.univ-lyon1.fr/software/alFinder/config_generator/alFinder_config_step1.php) is available to apply the procedure.

In the case where alleles have been previously described, the user can process directly through all the steps described below. In the case where alleles have not been previously described, the user should conduct a first post-processing procedure to identify alleles' sequences. To this purpose user should follow the steps described below but retaining only reads equal to the expected allele length in step 1.3 and spiking step 3. A second post-processing procedure can then be conducted (following all steps described below) to maximize the number of retained reads and amplicons thanks to indels' assignment (see step 3).

Step 1. Assignment of reads to loci and individuals, elimination of singletons, elimination of reads with inappropriate sizes

Only reads containing perfect analogous individual tags (forward and reverse) and the minimal number of bp allowing to distinguish with no ambiguity between the forward and reverse primers (i.e. the first x-bp of the different primers used) instead of complete primers are retained from the FASTA files produced as a result of the sequencing run (Step1.1). The use of x-bp of the primers instead of the complete primers was designed to maximize the number of retained reads per individual. The number of bp used is flexible and depends on the design of the primers and on the degree of stringency needed. After cutting off tags and primers, the library file is compressed by removing singletons (i.e. variants represented by a single read) (Step1.2) as well as reads with less than 95% or more than 105% of the expected allele length (Step1.3) to decrease the opportunity for inclusion of PCR chimeras and to maximize the number of reads with indels retained.

Step 2. Elimination of amplicons with insufficient coverage

Given that a minimal number of reads are required to obtain a reliable genotype and that several identical reads must be present within a given amplicon to obtain a reliable allele (Galan *et al.*, 2010), the minimal number of reads per amplicon needs to be calculated to ensure a negligible probability of missing alleles. The model proposed by Galan *et al.*, (2010) was used to assess this number per amplicon. In this model, the confidence level (*f*) to determine a correct genotype depends on three components: (1) *r*, the minimum required number of copies of the given allelic variant within an amplicon; (2) *n*, the total number of reads for a given amplicon; and (3) *m*, the maximum number of alleles within an amplicon (i.e. *m* is fixed to 2 for one locus in a diploid species). The program "Negative Multinomial" (implemented by Galan *et al.*, (2010) and freely available online at http://www.lirmm.fr/~caraux/Bioinformatics/NegativeMultinomial/) is used to determine the minimum value of *n* for a confidence level (*f*) of at least 95%. Thus, amplicons with less reads than the total number of reads for a given amplicon (*n*) are discarded at this stage.

Step 3. Determination of alleles

Variants corresponding to previously described alleles are identified and named following the original names of the alleles. Remaining variants are classified in two groups: (1) variants with a length corresponding to the one of the previously described alleles (i.e. correct length variants) and (2) variants with a length different than the one corresponding to the previously described alleles (i.e. incorrect length variants).

To increase the number of assigned reads thanks to the use of sequencing errors, all variants are aligned using the progressive alignment (Feng and Doolittle 1987) with the default aligning parameters of the CLC Sequence Viewer software free trial version 6.7.1. Since the primary artefacts generated by amplicon-based NGS techniques differ among platforms used (e.g. 454 and

Ion Torrent primary generate indels, SOLiD generates A-T bias and Illumina generates substitutions (Glenn 2011)), this step is more efficent in retaining additional reads for 454 and Ion Torrent instruments than for other technologies. Studies using SOLiD or Illumina instruments could thus skip this step until further implementations may allow the exploitation of artefacts generated by these techniques.

First, among correct length variants, variants presenting indels that lead to a change in all amino-acids following the indels (i.e. variants with one insertion and one deletion resulting from sequencing errors) are assigned manually to a previously identified alleles or to another correct length variant. Other correct length variants are not assigned since they could either be artefacts (i.e. indels) or they can be true alleles produced by mutation-selection effects. Assigning such variants could lead to the underestimation of allelic diversity. Second, incorrect length variants presenting indels are assigned manually to previously identified alleles or to correct length variants. Assignation of variants with indels allow to increase the evidence that the more frequent variant present in a given amplicon is a true allele (Stutz and Bolnick 2014). This procedure could be especially useful when true alleles are represented by few reads due to a lower sequencing efficiency relative to other alleles (Sommer *et al.*, 2013).

Step 4. Determination of homozygous and heterozygous amplicons

To determine homozygous and heterozygous amplicons, the model proposed by Hohenlohe *et al.,* (2010), and later used by Etter *et al.,* (2011), is used to calculate, given the sequencing error rate, the likelihood of each possible genotype given all the retained variants of an amplicon. An homozygous or an heterozygous genotype is assigned to each amplicon based on a likelihood ratio test between the most likely homozygous and the most likely heterozygous genotypes with one degree of freedom. If the likelihood ratio test is not significant, no genotype is assigned.

REFERENCES

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 (eds) Molecular methods for evolutionary genetics, Humana Press: Totowa. Vol 772, pp 157-178.

Feng DF, Doolittle RF (1987). Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J Mol Evol* 25: 351-360.

Galan M, Guivier E, Caraux G, Charbonnel N, Cosson J-F (2010). A 454 multiplex sequencing method for rapid and reliable genotyping of highly polymorphic genes in large-scale studies. *BMC Genomics* 11: 296.

Glenn TC (2011). Field guide to next-generation DNA sequencers. *Mol Ecol Resour* 11: 759-769.

Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA (2010). Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genet* 6: e1000862.

Sommer S, Courtiol A, Mazzoni CJ (2013). MHC genotyping of non-model organisms using next-generation sequencing: a new methodology to deal with artefacts and allelic dropout. *BMC Genomics* 14: 542.

Stutz WE, Bolnick DI (2014). Stepwise threshold clustering: a new method for genotyping MHC loci using next-generation sequencing technology. *PLoS One* 9: e100587.