Sequencing type and molecular markers

Dr Stéphane Joost – Oliver Selmoni (Msc)
Laboratory of Geographic Information Systems (LASIG)
Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

Introduction

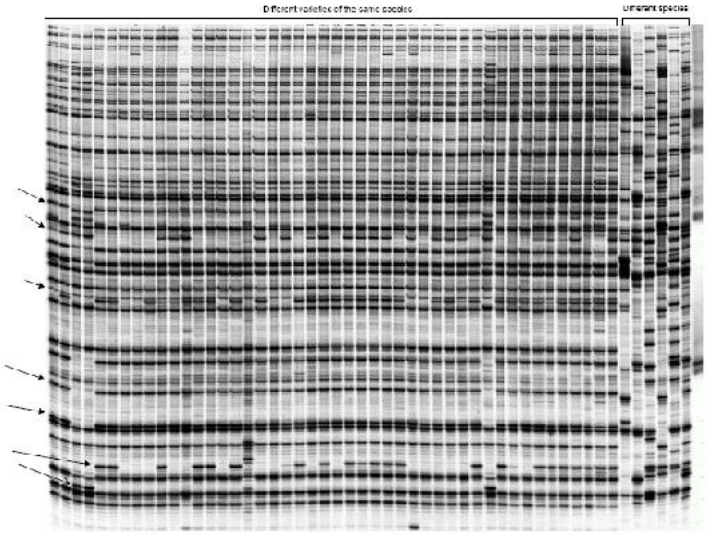
- Recent and upcoming advances in high throughput DNA sequencing leads to ever increasing availability of genomic sequences
- The limiting factor in future studies will no longer be the molecular laboratory work, but the development of statistical, bioinformatics and modelling tools for identifying both genes under selection, and the environmental factors acting as selective pressures
- The genomic tools available have characteristics we must take into account to design landscape genomic studies

- Rellstab, C., Gugerli, F., Eckert, A. J., Hancock, A. M., & Holderegger, R. (2015). A practical guide to environmental association analysis in landscape genomics. *Molecular Ecology*, *24*(17), 4348–4370. doi:10.1111/mec.13322
- Manel, S., Poncet, B. N., Legendre, P., Gugerli, F., & Holderegger, R. (2010). Common factors drive adaptive genetic variation at different spatial scales in Arabis alpina. Molecular Ecology, 19, 3824–3835. doi:DOI 10.1111/j.1365-294X.2010.04716.x

Overview of advantages and drawbacks of the main genomic resources available for landscape genomics studies

AFLPs

- Amplified fragment length polymorphisms
- Until recently AFLP was the method of choice to obtain large numbers of molecular markers for non-model organism genomic studies (up to ~2’000)
- It does not require prior sequenced-based information
- AFLP markers are bi-allelic, dominant and they usually cover the entire genome although they sometimes tend to cluster around centromeres
- A recurring issue associated with AFLP is fragment size homoplasy → nonhomologous AFLP fragments co-migrate



- A variant of the AFLP protocol is the Diversity Array Technology (DArT)
- Up to several thousands of DNA polymorphisms can be detected in a single hybridization assay on a microarray slide
- The major advantage of DArTs over AFLPs: their sequences are easily accessible

Microsatellites

- Microsatellites are codominant and generally multiallelic
- This makes them useful to monitor decreases in intrapopulation genetic variability observed in the vicinity of adaptive genes or to identify particular alleles specifically associated with environmental variables
- However, microsatellites have a high mutation rate and a complex mutation pattern, characteristics which can be difficult to accommodate when searching for selection of signatures using traditional population genomics models
- Moreover, microsatellites can be sparse in the genome of some species and thus difficult to find
- Up to now, the development of hundreds of microsatellites was time-consuming and expensive, and these markers were also not particularly amenable to massively parallel genotyping
- As a result, microsatellite resources were sufficient to be exploited in a population genomics context with model species only
- The increased availability of high-throughput sequencing data will facilitate microsatellite discovery and typing in non-model species
- Based on microsatellite data, Sork et al. (2010) detected climatically associated genetic variation in populations of **valley oak** in California, suggesting that the potential for future adaptation in the face of climate change is limited in this long-lived species

SNPs

- SNPs are the most abundant type of polymorphism in genomes
- For example, on average there is one SNP every Kb in the 3-billion-base human genome
- They are usually biallelic and evolve according to a simple infinite sites mutation model
- One of the major drawbacks of SNPs is their susceptibility to ascertainment bias, i.e. the bias introduced by using a subset of the studied individuals or populations for marker discovery purposes and which can lead to a skew in the distribution of allelic frequencies
- Detecting SNPs also requires a priori information on the studied genome sequence, but once this task is completed, SNPs present a high potential for an automated high-throughput analysis at a moderate cost
- Fortunately, next-generation sequencing technologies boosted the use of SNPs for both model and non-model organims
- Turner et al. (2010) investigated the genetic basis of adaptation to serpentine soils in Arabidopsis lyrata using about 8 millions SNPs

High density SNP chips

- E.g. Illumina 50k, 600k
- Specific to species
- Often used in domestic animals
- Made of selected SNPs to monitor loci related to traits involved in productive aspects (milk production, hair growth, temperature regulation)
- Not a ~random set of loci

Sequencing techniques

RAD-Seq

- In restriction-site associated DNA sequencing (RAD-Seq), the complexity of the genome is reduced using restriction enzymes
- The flanking regions of restriction sites are sequenced by NGS
- This approach has successfully been applied to pooled population samples
- RAD-Seq identifies fewer polymorphisms, from a few thousand to tens of thousands, as compared to millions of SNPs when using whole-genome Pool-Seq
- Controversy about the use of RAD-Seq without proper understanding of the biology and genome structure of the species to be studied

See …

Lowry DB, S Hoban, JL Kelley, KE Lotterhos, LK Reed, MF Antolin, and A Storfer. 2016. Breaking RAD: An evaluation of the utility of restriction site associated DNA sequencing for genome scans of adaptation. *Mol. Ecol. Resources.* 17:142–52.

… **and answers published in MER**

Pool-Seq

- Pool-Seq is a cost-effective method of NGS because the DNAs of several individuals are pooled before sequencing
- E.g. in plants, 10 individuals per plot can be sampled and their DNA then pooled before sequencing
- This approach can lead to accurate SNP allele frequency estimates
- As a drawback, individual multi-locus genotypes and information on heterozygosity are inaccessible
- As many landscape genomics tools can handle population allele frequencies, the use of whole-genome Pool-Seq is an attractive option, but only BAYENV yet accounts for the variance introduced by variation in sequencing coverage in Pool-Seq
- Whole-genome Pool-Seq data have only rarely been used in landscape genomics so far
- Does not permit to fully exploit local environmental variation

Conclusion: main distinction

- Study adaptation for conservation purposes
  - AFLP or Microsatellites are key (cheaper)
  - Reinforce the feasability of using genetic data in conservation (absolutely not the case yet, academia excepted)
- Fundamental research to understand the mechanisms of adaptation
  - SNPs
    - WGS, HD (modulation according to funding)
    - RADSeq

Exercise

- Data filtering: MAF, missingness
- Spatial genetic variation (F-stats, Ht, Fis, etc.)
- Population structure in a spatial context