Preface

Population genomics is a recently emerged discipline, which aims at understanding how evolutionary processes influence genetic variation across genomes. Its underlying principle involves genome-wide screening of several hundred or thousand loci across many individuals from several populations, in order to disentangle locus-specific (e.g., mutation, recombination, selection) from genome-wide (e.g., genetic drift, gene flow, inbreeding) effects. With such wealth of data, classical population genetics statistics, such as population differentiation or linkage disequilibrium, can then be studied as variables varying along the genome to understand which evolutionary forces drive its evolutionary dynamics.

In the past, characterizing polymorphism at the genome scale has been greatly facilitated by technological advances like capillary electrophoresis, which allows the simultaneous genotyping of numerous loci, such as Amplified Fragment Length Polymorphism (AFLP) markers, or DNA chips for Single Nucleotide Polymorphism (SNP) typing. Today, in the era of cheaper next-generation sequencing (NGS), population genomics takes on a new dimension and is meant to soar. Because it is not as daunting anymore to obtain whole genome data for any species of interest, population genomics is now conceivable in a wide range of fields, from medicine and pharmacology to ecology and evolutionary biology. However, because of the lack of reference genome and of enough a priori data on the polymorphism, population genomics analyses of populations will still involve higher constraints for researchers working on nonmodel organisms, as regards the choice of the genotyping/sequencing technique or that of the analysis methods. This observation guided our selection of chapters, which purposely put emphasis on various protocols and methods that are applicable to species where genomics resources are still scarce.

Setting up a population genomics study implies facing three main challenges. First, one has to devise a sampling and/or experimental design suitable to address the biological question of interest. Then, one has to implement the best genotyping or sequencing method to obtain the required data, given the time and cost constraints as well as the other genetic resources already available. Finally, one has to make the most of the (generally huge) dataset produced and use appropriate analysis methods to reach a biologically relevant conclusion.

The first challenge is addressed in Part I of this book, which is entitled "Sampling and Experimental Design." Chapter 1 is primarily concerned with the optimization of sampling and explains how basing a sampling design on model-based stratification using the climatic and/or biological spaces may be more efficient than basing it on the geographic space. Chapter 2 deals with the largely overlooked issue of sample tagging in NGS and presents bioinformatic tools for creating sets of tag and managing multiplexes of samples. Chapter 3 covers all aspects of SNP discovery in nonmodel organisms using Roche transcriptome sequencing, from tissue choice to publication of the work. Chapter 4 introduces ISIF, a program that helps design a successful AFLP experiment and choose the best restriction enzymes and combinations of selective bases based on a priori information from a close reference genome.

Part II, "Producing Data" brings together lab protocols aiming at generating high-quality genotyping or sequencing data. Chapter 5 focuses on Diversity Arrays Technology (DArT), a high-throughput genotyping technique that has already proven its worth for a wide range of nonmodel species, and Chapter 6 reports two protocols allowing the isolation and sequencing of AFLP fragments of interest for subsequent analysis.

When NGS is a reasonable option, many labs decide to outsource the sequencing task to a platform/company that can afford to acquire and maintain next-generation sequencers. Moreover, the fast evolution of NGS techniques is likely to make the current protocols quickly obsolete. Therefore, we decided not to emphasize such protocols unless they address a very specific problem. For example, Chapter 7 deals with whole-genome sequencing of ancient (or degraded) DNA, Chapter 8 describes transcriptome sequencing using the 454 platform of Roche, and Chapter 9 is dedicated to the practical implementation of paired-end Illumina sequencing of RAD (Restriction-site Associated DNA) fragments to reveal SNP markers in nonmodel organisms.

Part III entitled "Analyzing Data" compiles new statistical methods and bioinformatic tools to meet the last challenge pertaining to data management and analysis in population genomics. Chapter 10 presents RawGeno, a free program designed to automatically score AFLP profiles and identify reliable markers. Chapter 11 gives an overview of the most efficient methods allowing haplotype reconstruction. Chapter 12 tackles the issue of allele versus paralog determination in 454 transcriptomic data, and proposes a series of bioinformatic scripts to identify true allelic clusters in new datasets. Chapter 13 addresses the common problem of multiple-testing related to the typically huge numbers of loci analyzed in population genomics datasets, which can lead to the detection of many false positives. Chapter 14 outlines the analytical specificities of RAD-seq and similar data and introduces a new computational pipeline called Stacks for the analysis of RAD-seq data in organisms with and without a reference genome. Chapter 15 describes METAPOP, a computer application that provides an analysis of gene and allelic diversity in subdivided populations from molecular genotype or coancestry data.

One of the central goals of population genomics is to identify loci underlying phenotypic variation or adaptation in natural populations and this is reflected in the last three chapters of this book. Chapter 16 presents the R-DetSel software package, which implements a coalescent-based method to detect markers that deviate from neutral expectation in pairwise comparisons of diverging populations. Chapter 17 shows how allele distribution models can be exploited to detect loci displaying signatures of selection and illustrates this approach in cichlid fish using the software MATSAM. Finally, Chapter 18 presents a new method allowing the quantification of the genetic component of phenotypic variance in wild populations using phenotypic trait values and multilocus genetic data available simultaneously for a sample of individuals from the same population.

Whether presenting a specific protocol or an overview of several methods, each chapter aims at providing guidelines to help choose and implement the best experimental or analytical strategy for a given purpose. In this respect, the Notes section is particularly valuable since it gathers practical information and tips rarely highlighted in scientific articles. The methods and protocols described in these chapters were selected because they are likely to be of interest to a wide readership and we hope that that they will contribute to the success of many population genomics studies in the future.

Grenoble, France

François Pompanon
Aurélie Bonin

http://www.springer.com/978-1-61779-869-6

Data Production and Analysis in Population Genomics
Methods and Protocols
Pompanon, F.; Bonin, A. (Eds.)
2012, XI, 337 p., Hardcover
ISBN: 978-1-61779-869-6
A product of Humana Press