Measuring aphid clonal line abundance after experimental evolution

Lucas Nell

Spring 2017

Introduction

Experimental evolution studies have been instrumental in informing our understanding of the processes shaping evolution (Elena and Lenski, 2003). Most of such studies have been carried out in microbes (reviewed in Dettman et al., 2012; Jerison and Desai, 2015), and have provided insights on such diverse and fundamental themes as historical contingency, evolutionary innovation, parallel evolution, and adaptation (Blount et al., 2008; Barrick et al., 2009; Toll-Riera et al., 2016; Voordeckers and Verstrepen, 2015; Gerstein et al., 2012).

Similar experiments have been carried out in invertebrates (Gompert and Messina, 2016; Chandler, 2014; Burke et al., 2010; Kang et al., 2016; Rouchet and Vorburger, 2014), although such studies are comparatively rare. Experimental evolution studies in insects typically utilize limited numbers of clonal or inbred lines (e.g., Rouchet and Vorburger, 2014; Kang et al., 2016) and characterize experimental populations by either (a) measuring the distribution of specific phenotypes (e.g., Rouchet and Vorburger, 2014) or (b) sequencing pooled DNA ("Pool-seq"; e.g., Burke et al., 2010). The former requires that the researcher manipulate the environment such that a specific phenotype is predicted to change. Moreover, if initial experimental populations have a continuous distribution of phenotypes or if any significant degree of phenotypic plasticity exists, this method is not likely to provide accurate estimates of how distributions of individuals change through time.

Pool-seq, however, is an accurate, cost-effective method to measure allele frequencies in populations (Gautier et al., 2013; Futschik and Schlötterer, 2010) and to identify loci associated with traits (Rubin et al., 2010; Bastide et al., 2013). However, Pool-seq's advantageous accuracy-to-cost ratio is only present when there are many pooled individuals (> 40) and when depth of coverage is high ($> \times 50$). Because sequencing error is difficult to distinguish from rare alleles, Pool-seq is not ideal when trying to detect these low-frequency alleles (Schlötterer et al., 2014). Additionally, Pool-seq of whole-genome sequencing provides much unnecessary information if an association study is not the ultimate goal.

One way to reduce genome complexity is to use restriction site-associated DNA sequencing ("RADseq"). RADseq approaches use restriction enzymes to break apart the genome at specific locations determined by the enzyme's binding site sequence. Although some use RADseq to refer

to one specific methodology, here I use the more inclusive definition of RADseq by Andrews et al. (2016), who define RADseq as all methods "... that rely on restriction enzymes to determine the set of loci to be sequenced" (Andrews et al., 2016, p 81). Many iterations of RADseq exist and each has particular situations where they are most appropriate, such as 2bRAD or ezRAD when many cut sites are required or double-digest RAD (ddRAD) when sampling complex genomes or needing extreme flexibility (Andrews et al., 2016). Genotyping-by-sequencing ("GBS") does not usually provide as many cut sites as some other methods, but it is a particularly low-effort, low-cost RADseq methodology that requires no specialized equipment for sample preparation (Elshire et al., 2011). I have decided on GBS going forward, and below I will outline why its aforementioned attributes are particularly suitable for my experiment.

The purpose of this paper is to assess pooled GBS as a method to estimate the abundance of clonal lines of aphids after experimental evolution. In the proposed experiment, I seek to measure the repeatability of evolution by starting replicate cages (2m × 1m × 1m) with 50 each of the same 10 aphid clones, allowing aphids to compete in cages for ~6 months, then determining the relative abundances of the starting clonal lines at the end of the experiment. Clonal lines will also be genotyped individually to allow us to distinguish between them in pooled samples. Aspects of this specific experiment that make it suitable for pooled GBS are as follows:

Pooling: Each experimental population will contain ≫1,000 individuals when allele frequencies are sought at the end of the experiment. Individual-based estimates (e.g., using microsatellites) would require ~50–100 individuals to be sampled from the population (Figure 1), which would take huge amounts of preparation time. Pool-seq would allow me to only prepare a single or a few samples per experimental replicate, and, given adequate coverage, should provide accurate allele frequencies (Schlötterer et al., 2014).

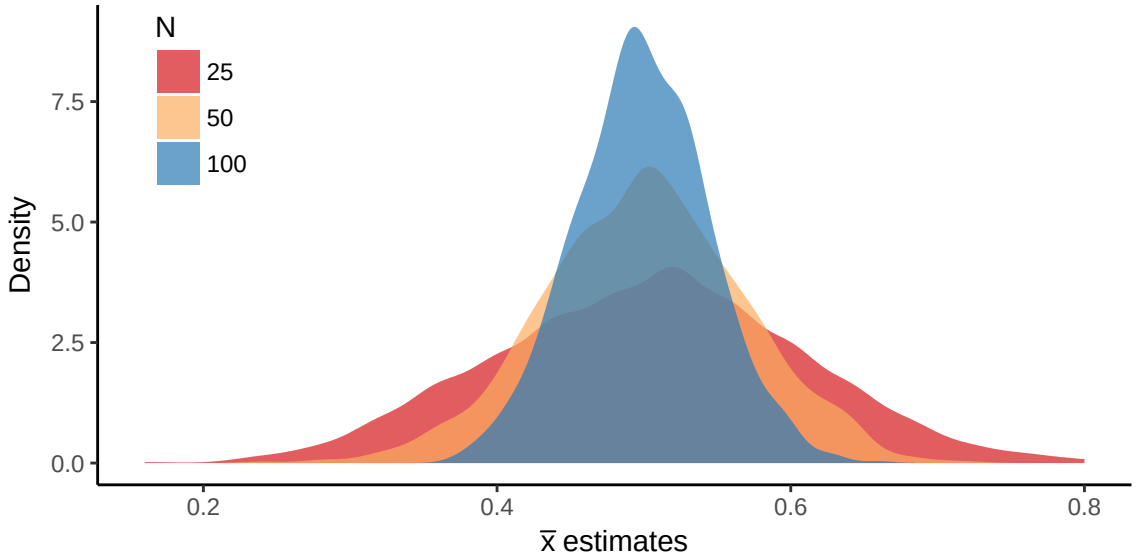

Figure 1: Simulated estimates of mean genotype abundance in a population for a given sample size. Samples were randomly drawn from a population of 1,000 with $\mu = 0.5$. N is the number of samples drawn from the population, and distributions are for 1,000 simulations. See here for this figure's code.

GBS: My lab group has limited access to wet lab supplies and cannot conduct lengthy sample preparation procedures. Because I also should not need many cut sites to distinguish between clonal lines, GBS should serve my purposes adequately.

Individually, Pool-seq and GBS would likely allow us to estimate aphid abundances, but the accuracy of combining Pool-seq with GBS to estimate allele frequencies from populations of known genotypes has never been examined (to this author's knowledge). I will measure accuracy using simulations of pooled GBS from variants derived from the aphid reference genome.

Methods

All of my code was created in R version 3.3.3 (R Core Team, 2017), and C++ code was implemented using the *'Rcpp'* package (Eddelbuettel, 2013). Code for the simulations can be found here.

Initial steps

I first input the aphid reference genome fasta file into R, where I first digested it with the restriction enzyme of my choice (*ApeKI* for now). Longer digested fragments are not as likely to be sequenced (Andrews et al., 2016; Elshire et al., 2011), so I filtered for size based on the distribution of fragment sizes seen by Elshire et al. (2011).

Creating variants

I next created 10 versions of the aphid genome, each containing SNPs, to simulate 10 different clonal lines. SNP abundance and characteristics where present were based on two estimates of population-genomic diversity from Bickel et al. (2013): $\theta_w = 0.0050$ and $\theta_\pi = 0.0045$. From these estimates, I estimated the proportion of segregating sites and the nucleotide diversity at segregating sites, respectively. I will calculate the proportion of segregating sites first.

θ_w is calculated as follows (Watterson, 1975):

$$\theta_w = \frac{K}{a_n} \tag{1}$$

where K is the proportion of segregating sites. Variable a_n is below:

$$a_n = \sum_{i=1}^{n-1} \frac{1}{i} \tag{2}$$

where n is the number of individuals sampled. Thus the the proportion of segregating sites is simply $K = \theta_w a_n$. For 10 samples and $\theta_w = 0.0050$, $K = 0.01414$.

Next, I will estimate the nucleotide diversity at segregating sites. θ_π is calculated using the following equation (Nei and W. H. Li, 1979):

$$\theta_\pi = \sum_{ij} x_i x_j \pi_{ij} \tag{3}$$

Here, x_i and x_j represent the frequencies of the i^{th} and j^{th} unique sequences respectively and π_{ij} represents the proportion of divergent sequence between the i^{th} and j^{th} unique sequences. If I assume that all n lines will be unique sequences—a safe assumption if whole genomes are considered—then the above equation can be expressed as follows:

$$\theta_\pi = \frac{1}{n^2} \sum_{ij} \pi_{ij} \tag{4}$$

Then, since the number of total pairwise combinations between n sequences is $\binom{n}{2}$, we can calculate $\bar{\pi}$, the mean proportional sequence divergence between any two sequences, as such:

$$\bar{\pi} = \frac{\sum_{ij} \pi_{ij}}{\binom{n}{2}} \tag{5}$$

Some simple arithmetic gives us...

$$\sum_{ij} \pi_{ij} = \binom{n}{2} \bar{\pi} \tag{6}$$

Now I insert this into equation 4:

$$\theta_\pi = \frac{1}{n^2} \binom{n}{2} \bar{\pi} \tag{7}$$

Solving for $\bar{\pi}$ yields the following:

$$\bar{\pi} = \frac{\theta_\pi n^2}{\binom{n}{2}} \tag{8}$$

Since I have already calculated the proportion of segregated sites, only considering segregating sites will make my simulations more simple. Thus I will calculate the mean divergence at segregated sites only. To only consider segregated sites, I divide the whole expression by the proportion of segregated sites. This leaves me with the proportional nucleotide divergence between two sequences at segregating sites, $\bar{\pi}_s$:

$$\bar{\pi}_s = \frac{\theta_\pi n^2}{K\binom{n}{2}} \tag{9}$$

Using the estimates above, $\bar{\pi}_s = 0.7072$. At each segregating (i.e., SNP) site, I inserted random nucleotide frequencies that matched closest to this average pairwise difference.

Preparing sequences

The last step in R was to prepare sequences for simulation. This involved making sure all sequences were at least as long as the anticipated read length (100bp), adding barcodes, and removing sections of the fragments that won't be sequenced (inner portions far from cut sites) (Elshire et al., 2011; Davey et al., 2011).

Simulating Illumina reads

I used the ART sequencing simulator (Huang et al., 2012) to simulate single-end, 100bp Illumina reads. Because I only wanted fragments to be sequenced from the ends (i.e., no further digestion), I simulated paired-end reads with the mean size of DNA fragments set to 200bp (100bp from each end, the maximum size output from the "Preparing sequences" step) with a standard deviation of zero. I also forced R2 reads to have the same error and mapping quality profile as R1.

I simulated 100× coverage for all samples across the entire genome.

Downstream analyses

I next aligned the pooled sample and all separated individual samples to the aphid genome using BWA-MEM (H. Li, 2013). Individual samples were created by filtering the `fastq` file output from ART by sample name. Thus all the reads found in a given sample `fastq` file were also present in the pooled `fastq` file. Those resulting `SAM` files were summarized using the mpileup function in samtools, resulting in a single `mpileup` file. That file was input to Popoolation2 (Kofler et al., 2011), which creates a much more concise summary of the output, a `sync` file.

Figure 2 contains an overview of this simulation process.

Abundance calculation

From the sync file, I estimated how many alleles a sample could have contributed to the pooled sample at a given location (from 0 to 2) based on what nucleotides were present in the pooled and sample's alignment at that location. I then summed these counts for each sample to create their total allele counts. Abundances were each sample's allele counts divided by the total allele counts for all samples.

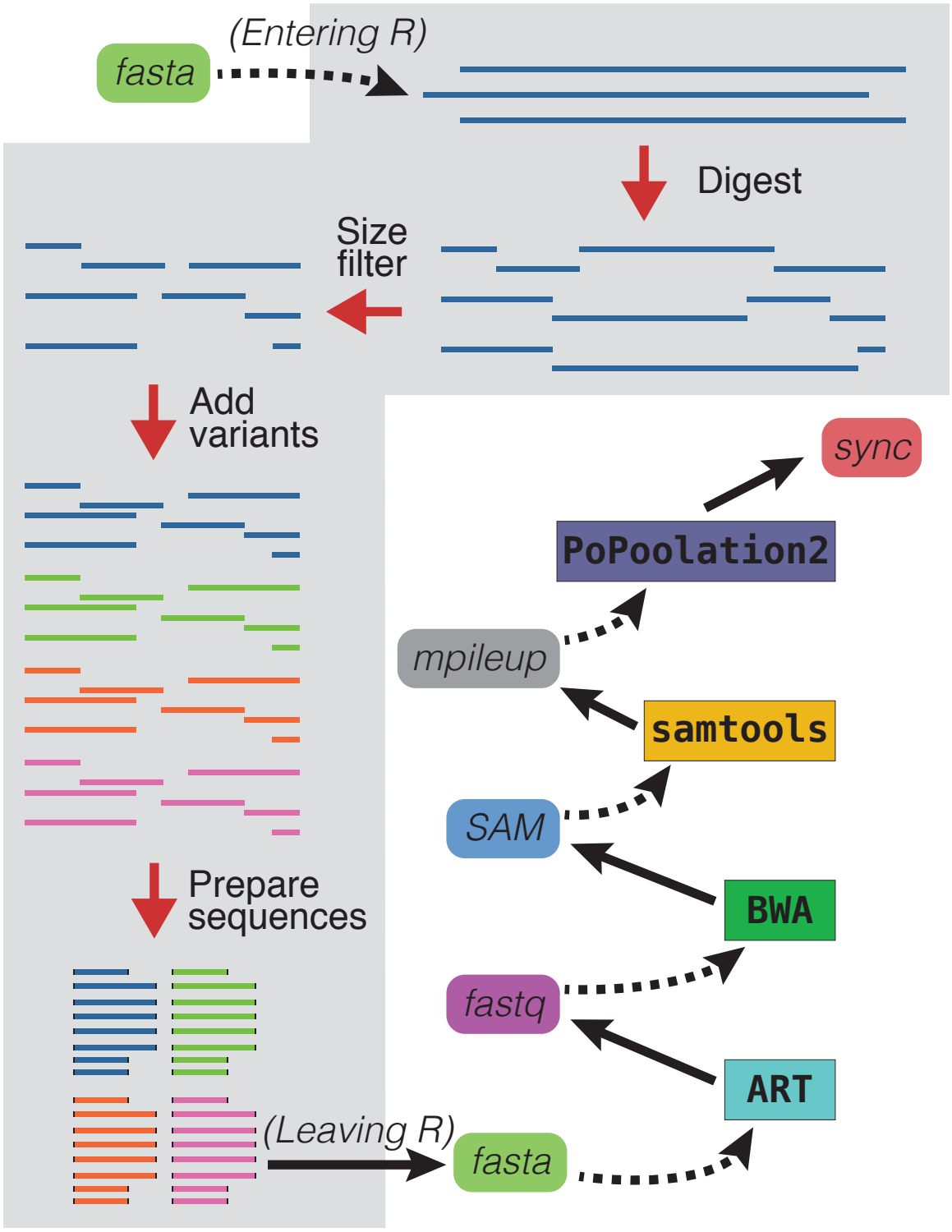

Figure 2: Overview of methods used to simulate GBS data and summary. The gray box indicates steps run entirely in R, each represented by a red arrow. Black dashed arrows represent file inputs to a program, while solid black arrows show output from a program. Rounded rectangles are file formats, while rectangles are programs. ART = ART sequencing simulator, BWA = Burrows-Wheeler Aligner

Results

There were a total of 212,350,000 reads simulated by the ART simulator. These reads, when aligned to the aphid genome, covered 9,795,971 unique positions. The distribution of read depths had a small peak at very low depths, with a very large peak at slightly less than 1,000×, and another, smaller peak at ∼1,750× (Figure 3).

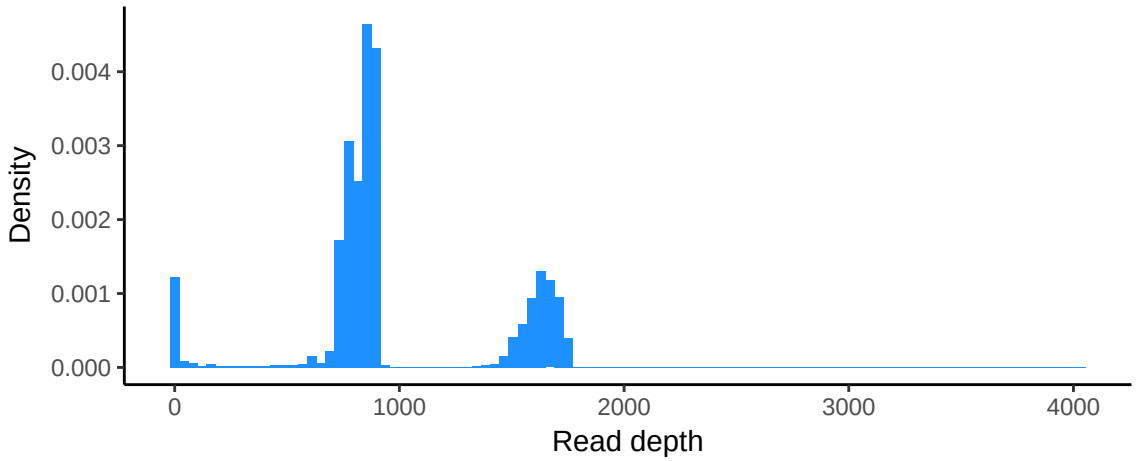

Figure 3: Histogram of read depths for simulated reads aligned to the aphid reference genome.

There were 151,572 locations containing SNPs. Using these locations to estimate abundances resulted in estimates very near actual values (Figure 4).

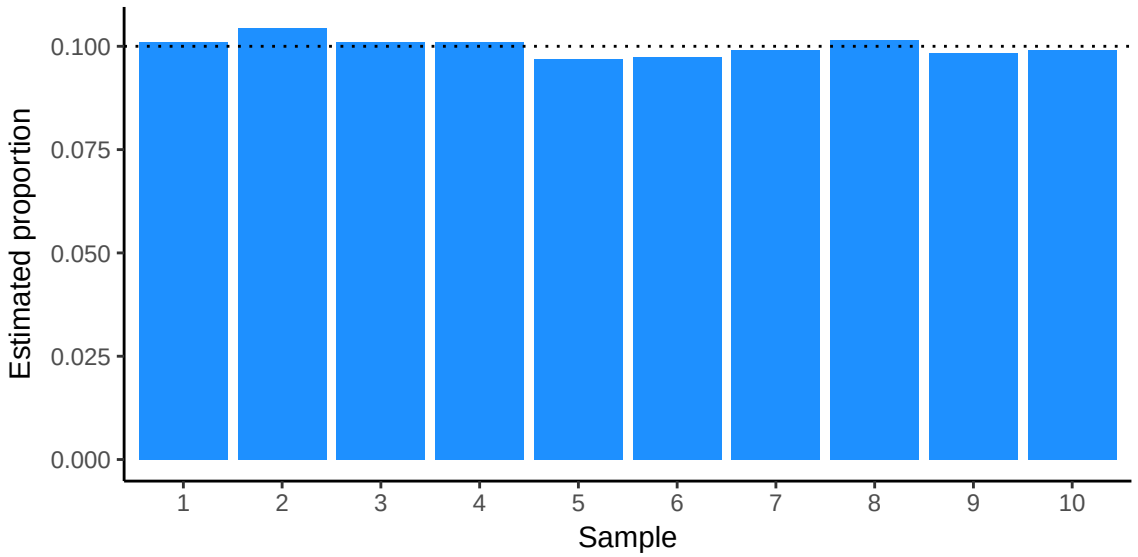

Figure 4: Estimated proportions from simulated data for 10 samples. The dotted line represents the actual abundance, which is the same for all samples.

Discussion

Simulated pooled GBS data resulted in abundance estimates nearly identical to correct values. However, multiple sources of error were not simulated here, and would need to be if a proper test of this method is to be made. First, I simply filtered the pooled DNA `fastq` file to get reads for the individual sample reads. This means that the same sequencing error that occurred in the pooled reads occurred in the samples' reads, which is unrealistic. Second, I created variants after digesting the genome, so mutations occurring in restriction enzyme binding sites, a common problem for RADseq data (Andrews et al., 2016), was not introduced into my simulations. Third, PCR bias was partially replicated when I filtered by fragment size, but I did not simulate an actual PCR process. Other biases may be introduced if a full PCR process is implemented. Fourth, I only introduced SNPs into genomic variants, but more complex genomic changes (e.g., indels, inversions) may necessitate more complex analyses. Fifth, no individuals were heterozygous in my simulations.

If pooled GBS remains reasonably accurate with these additional sources of error, there are other tests that I need to run to optimize how I use this method. All samples were evenly distributed in the pooled DNA, but this will not likely be the case in reality. Assessing how rare clonal lines are detected using this approach will be a key goal going forward, particularly because it is a predicted weakness of pooled RADseq methods (Schlötterer et al., 2014). Read depth will be particularly important for determining rare allele frequencies, but it might also have overall effects on allele frequency estimates. This should also be assessed.

References

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., and Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics* 17 (2), 81–92.

Barrick, J. E., Yu, D. S., Yoon, S. H., Jeong, H., Oh, T. K., Schneider, D., Lenski, R. E., and Kim, J. F. (2009). Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. *Nature* 461 (7268), 1243–1247.

Bastide, H., Betancourt, A., Nolte, V., Tobler, R., Stöbe, P., Futschik, A., and Schlötterer, C. (2013). A genome-wide, fine-scale map of natural pigmentation variation in *Drosophila melanogaster*. *PLOS Genetics* 9 (6), e1003534.

Bickel, R. D., Dunham, J. P., and Brisson, J. A. (2013). Widespread selection across coding and noncoding DNA in the pea aphid genome. *G3: Genes|Genomes|Genetics* 3 (6), 993–1001.

Blount, Z. D., Borland, C. Z., and Lenski, R. E. (2008). Historical contingency and the evolution of a key innovation in an experimental population of *Escherichia coli*. *Proceedings of the National Academy of Sciences of the USA* 105 (23), 7899–7906.

Burke, M. K., Dunham, J. P., Shahrestani, P., Thornton, K. R., Rose, M. R., and Long, A. D. (2010). Genome-wide analysis of a long-term evolution experiment with Drosophila. *Nature* 467 (7315), 587–590.

Chandler, C. H. (2014). Parallel genome-wide fixation of ancestral alleles in partially outcrossing experimental populations of *Caenorhabditis elegans*. *G3: Genes|Genomes|Genetics* 4 (9), 1657–1665.

Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., and Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics* 12 (7), 499–510.

Dettman, J. R., Rodrigue, N., Melnyk, A. H., Wong, A., Bailey, S. F., and Kassen, R. (2012). Evolutionary insight from whole-genome sequencing of experimentally evolved microbes. *Molecular Ecology* 21 (9), 2058–2077.

Eddelbuettel, D. (2013). *Seamless R and C++ integration with Rcpp*. New York, NY, USA: Springer.

Elena, S. F. and Lenski, R. E. (2003). Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. *Nature Reviews Genetics* 4 (6), 457–469.

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., and Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLOS ONE* 6 (5), e19379.

Futschik, A. and Schlötterer, C. (2010). The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. *Genetics* 186 (1), 207–218.

Gautier, M., Foucaud, J., Gharbi, K., Cezard, T., Galan, M., Loiseau, A., Thomson, M., Pudlo, P., Kerdelhué, C., and Estoup, A. (2013). Estimation of population allele frequencies from next-generation sequencing data: pool-versus individual-based genotyping. *Molecular Ecology* 22 (14), 3766–3779.

Gerstein, A. C., Lo, D. S., and Otto, S. P. (2012). Parallel genetic changes and nonparallel gene–environment interactions characterize the evolution of drug resistance in yeast. *Genetics* 192 (1), 241–252.

Gompert, Z. and Messina, F. J. (2016). Genomic evidence that resource-based trade-offs limit host-range expansion in a seed beetle. *Evolution* 70 (6), 1249–1264.

Huang, W., Li, L., Myers, J. R., and Marth, G. T. (2012). ART: a next-generation sequencing read simulator. *Bioinformatics* 28 (4), 593–594.

Jerison, E. R. and Desai, M. M. (2015). Genomic investigations of evolutionary dynamics and epistasis in microbial evolution experiments. *Current Opinion in Genetics & Development* 35, 33–39.

Kang, L., Aggarwal, D. D., Rashkovetsky, E., Korol, A. B., and Michalak, P. (2016). Rapid genomic changes in *Drosophila melanogaster* adapting to desiccation stress in an experimental evolution system. *BMC Genomics*, 1–11.

Kofler, R., Pandey, R. V., and Schlötterer, C. (2011). PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics* 27 (24), 3435–3436.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*. arXiv: 1303.3997.

Nei, M. and Li, W. H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the USA* 76 (10), 5269–5273.

R Core Team (2017). *R: a language and environment for statistical computing*. 3.3.3. R Foundation for Statistical Computing. Vienna, Austria.

Rouchet, R. and Vorburger, C. (2014). Experimental evolution of parasitoid infectivity on symbiont-protected hosts leads to the emergence of genotype specificity. *Evolution* 68 (6), 1607–1616.

Rubin, C.-J., Zody, M. C., Eriksson, J., Meadows, J. R. S., Sherwood, E., Webster, M. T., Jiang, L., Ingman, M., Sharpe, T., Ka, S., Hallböök, F., Besnier, F., Carlborg, Ö., Bed'hom, B., Tixier-Boichard, M., Jensen, P., Siegel, P., Lindblad-Toh, K., and Andersson, L. (2010). Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature* 464 (7288), 587–591.

Schlötterer, C., Tobler, R., Kofler, R., and Nolte, V. (2014). Sequencing pools of individuals — mining genome-wide polymorphism data without big funding. *Nature Reviews Genetics* 15 (11), 749–763.

Toll-Riera, M., San Millan, A., Wagner, A., and MacLean, R. C. (2016). The genomic basis of evolutionary innovation in *Pseudomonas aeruginosa*. *PLOS Genetics* 12 (5), e1006005–22.

Voordeckers, K. and Verstrepen, K. J. (2015). Experimental evolution of the model eukaryote *Saccharomyces cerevisiae* yields insight into the molecular mechanisms underlying adaptation. *Current Opinion in Microbiology* 28 (C), 1–9.

Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology* 7 (2), 256–276.