Using "identity by heterozygosity (IBH)" to detect clonemates under prevalent clonal reproduction in multicellular diploids

Lei Yu[1*], John J. Stachowicz[2, 3], Katie DuBois[2, 3], Thorsten B. H. Reusch[1]

[1]Marine Evolutionary Ecology, GEOMAR Helmholtz Centre for Ocean Research Kiel, Kiel, Germany
[2]Center for Population Biology, University of California, Davis, CA, USA
[3]Department of Evolution and Ecology, University of California, Davis, CA, USA
[*]Corresponding author, lyu@geomar.de

Abstract

Clonal reproduction, the production of asexual modular units with the ability to live independently under multicellularity, is a very common life history. Genetic markers have been widely used to detect clonemates tracing back to a single sexual event. However, the currently available methods cannot work in cases where all collected samples belong to a few or even one single clone. Here we propose the concept of identity by heterozygosity (IBH), meaning that two diploid genomes are identically heterozygous at some genetic markers such as SNPs (single-nucleotide polymorphisms). Based on IBH, we develop IBH similarity (IBH_s), a novel measure of genetic similarity between two diploid genomes. The core of our IBH method for detecting clonemates is to use IBH_s in combination with the cutoff of 0.95. Two samples with IBH_s >0.95 are considered members of the same clone, while two samples with IBH_s <=0.95 are considered to represent different clones. The method requires a relatively large number of marker loci where the two target samples are identically heterozygous (i.e., N_{IBH}, on the order of >=3,000). Another potential application of IBH is to detect possible parent-descendant pairs under selfing because this reproductive mode leads to a predictable loss of heterozygosity. Our IBH method extends the detection of clonemates to any pair of samples, and will allow the detection of clonemates in extreme cases where all collected samples belong to one single clone.

Key words: identity by heterozygosity, IBH similarity, clonal reproduction, clonemate

Introduction

Clonal reproduction, the production of asexual modular units with the ability to live independently, can happen by means of asexual budding, fragmentation or branching. This life history is widespread in multicellular animals, plants and fungi, shared by 66.5% of flora (Honnay & Bossuyt, 2005), and 40% of animal phyla (Buss, 1983). A ramet (Harper & White, 1974) is the basic modular unit with an ability to live independently. A genet (synonymous with clone) is the product of a single zygote from gamete fusion (Noble et al., 1979), and comprises all ramets that emerge during its lifetime (i.e., clonemates).

Ramets belonging to the same clone vs. different clones are often physically indistinguishable. Once they obtain autonomy and break off the parental module, genetic markers such as microsatellites or AFLP have been widely used for detecting clonemates (Arnaud-Haond et al., 2007). A ramet can be genotyped with multiple independent markers, and the genotypes at all these loci collectively form a multilocus genotype (MLG). In case of co-dominant markers such as microsatellites, identical or nearly identical MLGs are then interpreted as being members of the same clone or genet. However, repeated MLGs are not definitive evidence for clonal reproduction (Halkett et al., 2005), if for example there is low (or no) allelic diversity at particular loci. In order to calculate the probability of finding identical MLGs resulting from distinct zygotes, the population allelic frequencies are required, and the population is assumed to be under Hardy-Weinberg equilibrium (Parks & Werth, 1993). Evidently, estimates of such population-level frequencies are inaccurate in populations dominated by a few clones (Arnaud-Haond et al., 2007), and meaningless if a single site is dominated by one clone (Smith et al., 1992; Reusch et al., 1999; Arnaud-Haond et al., 2012; Japaud et al., 2015), thereby rendering distinction of clonemates from non-clonemates difficult.

Moreover, members of the same clone could also have slightly different MLGs, due to a combination of somatic mutations and genotyping errors (Douhovnikoff & Dodd, 2003; Yu et al., 2020). To overcome this issue and also to work with dominant markers that do not allow for the application of MLGs, an alternative suite of approaches is the analysis of pairwise genetic distances. Ramet pairs from the same clone will show much lower levels of genetic distances than those from different clones, and thus the frequency distribution of distances will be bimodal rather unimodal. Then, a threshold can be defined to separate the

intra-clone vs. inter-clone distances (Douhovnikoff & Dodd, 2003; Bailleul et al., 2016; Meirmans, 2020). However, it is impossible to define the threshold without inter-clone distances, and again, such methods cannot work on cases where all sampled ramets belong to one single clone.

In the genomic era when whole-genome data can be easily obtained, we propose to detect clonemates in diploid species by directly assessing the genetic similarity between two target samples based on the thousands of segregating genetic markers throughout their genomes. Heterozygous sites are the only informative ones because they show the potential for segregation in sexual reproduction, while remaining identically heterozygous in clonal reproduction. Here we propose the concept of identity by heterozygosity (IBH), meaning that two diploid genomes are identically heterozygous at some genetic markers such as SNPs (single-nucleotide polymorphisms). Based on IBH, we developed IBH similarity (IBH_s), a novel measure of genetic similarity between two diploid genomes.

IBH method is most suitable for large number of SNPs obtained from whole-genome resequencing, restriction-site associated DNA (RAD) sequencing, or SNP arrays. The goal of this study is to test the performance of our IBH method using seagrass *Zostera marina* clones of known clonal history and/ or known age.

Materials and methods

Detecting clonemates based on identity by heterozygosity (IBH)

For a pair of samples in multicellular diploids, N_{IBH} denotes the number of the genetic markers where the two samples are identically heterozygous. We define $IBH_X = N_{IBH}$ / (number of the heterozygous loci in the sample named "X"). IBH similarity (IBH_s) is defined as the smaller value of IBH_X for the two target samples. If assuming no somatic mutations, a pair of clonemates would share all the heterozygous genotypes (Fig. 1), which leads to $IBH_s = 1$. Somatic mutations and genotyping errors will cause the IBH_s to be slightly smaller.

Threshold (N_{IBH} and IBH_s) decided based on a *Z. marina* clone as old as 750 to 1,500 years old

To set the threshold for N_{IBH} and IBH_s, respectively, we applied the method to the SNP data set in Yu et al. (2020). Yu et al. (2020) conducted whole-genome resequencing of 24 ramets of a seagrass *Z. marina* clone located in Finland. The clone was estimated to be between 750

and 1,500 years old. Here the clone was named "FIN", and the 24 ramets were named from FIN01 to FIN24, respectively. The SNP data set in Yu et al. (2020) contained 38,831 genomic sites before excluding the ones where all 24 ramets are identically heterozygous (see Extended Data Fig. 2 in Yu et al., 2020).

To obtain a proper threshold for N_{IBH} and IBH_s, respectively, we randomly selected 1/10, 1/100 and 1/1000 from the 38,831 SNPs, based on which the pairwise IBH_s was calculated for each pair of ramets. Each selection was repeated 100 times, and the IBH_s values from the 100 repeats were pooled together to plot a histogram (Fig. 2).

Seagrass *Zostera marina* clones with known ancestry

Two *Z. marina* clones were selected from the eight clones cultured in adjacent 300-l outdoor flow-through mesocosms at Bodega Marine Laboratory (BML) under ambient light and temperature conditions (Hughes et al., 2009). Each mesocosm was initiated with a single terminal shoot collected from one of eight different locations in Bodega Harbor, California, in July 2004. The genetic distinctiveness of each of the original shoots was verified before planting. Six shoots (= ramets) were collected from each clone for genomic analysis in 2021, after the clones had grown in mesocosm for 17 years. One clone was named "GREEN", and the six collected shoots were named G01, G02, G06, G07, G10 and G16, respectively. The other clone was named "RED", and the six collected shoots were named R02, R03, R05, R08, R09 and R10, respectively.

DNA extraction, whole-genome resequencing, quality check and filtering of raw Illumina reads

Genomic DNA was extracted using NucleoSpin plant II kit from Macherey-Nagel following the instructions. The plant tissue (160-200 mg fresh weight) was ground with lN_2. Library construction and Illumina sequencing (PE150, paired end 150 bp) were conducted by BGI (Beijing Genomics Institute, Hong Kong). The quality of the raw Illumina reads was assessed by FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). BBDuk (https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbduk-guide/) was used to remove adapter contamination and conduct a basic quality filtering: (1) reads with more than one Ns were discarded (maxns=1); (2) reads shorter than 50 bp after trimming were discarded (minlength=50); (3) reads with average quality below 10 after trimming were discarded (maq=10). FastQC was used for second round of quality check for the filtered reads.

SNP calling and filtering

The quality-filtered Illumina reads were mapped against the chromosome-level *Z. marina* reference genome V3.1 (Ma et al., 2021) using BWA MEM (Li & Durbin, 2009). The alignments were converted to BAM format and then sorted using Samtools (Li et al., 2009). The MarkDuplicates module in GATK4 (Van der Auwera & O'Connor, 2020) was used to identify and tag duplicate reads in the BAM files. Sequencing depth and percentage of covered genome region for each ramet were calculated using Samtools and Bedtools (Quinlan & Hall, 2010), respectively. HaplotypeCaller (GATK4) was used to generate GVCF format file for each ramet, and all the GVCF files were combined by CombineGVCFs (GATK4). GenotypeGVCFs (GATK4) was used to call genetic variants.

BCFtools (Li, 2011) was used to remove SNPs within 20 base pairs of an indel or other variant type. Then we extracted only bi-allelic SNPs located in the six main chromosomes (2,969,820 SNPs). INFO annotations were extracted using VariantsToTable (GATK4). SNPs meeting one or more than one of the following criteria were marked by VariantFiltration (GATK4): MQ < 40.0; FS > 60.0; QD < 13.0; MQRandSum > 2.5 or MQRandSum < -2.5; ReadPosRandSum < -2.5; ReadPosRandSum > 2.5; SOR > 3.0; DP > 1192.45 (2 * average DP). Those SNPs were excluded by SelectVariants (GATK4). A total of 1,105,628 SNPs were retained. VCFtools (Danecek et al., 2011) was used to convert individual genotypes to missing data when GQ < 30 or DP < 20. Individual homozygous reference calls with one or more than one reads supporting the variant allele, and individual homozygous variant calls with one or more than one reads supporting the reference allele, were also converted to missing data using a custom Python3 script. Finally, we kept only genomic sites showing polymorphism (604,066 SNPs).

Calculation of IBH measures

IBH measures were calculated by a custom Phython3 script for each pair of *Z. marina* ramets. We only focused on the nucleotide positions in the reference genome where both samples had available genotype calls (not missing data).

Results

For the *Z. marina* clone in Yu et al. (2020), the pairwise N_{IBH} ranged from 29,338 to 32,756. All ramet pairs displayed IBH_s greater than 0.95 (Fig. 2a). Based on the subsets of the SNP data set (Fig. 2b-d), we found that the minimum IBH_s increased when N_{IBH} increased. When N_{IBH} ranged from 2,808 to 3,407, the minimum IBH_s was 0.96. Hence, a threshold of 0.95 would be proper. We suggest to use our method only when N_{IBH} is on the order of >=3000, and the threshold for IBH_s can be set to 0.95.

The whole-genome resequencing data of the 12 cultured *Z. marina* ramets covered >93% of the reference genome, and reached >66x coverage (Table 1). The pairwise N_{IBH} ranged from 108,585 to 288,606. All ramet pairs within the clone "GREEN" displayed IBH_s greater than 0.95 (Fig. 3). As for the clone "RED", we detected one "contaminant" ramet (i.e., R10), obviously not belonging to that clone. The five ramet pairs composed of R10 and one of the other five "RED" ramets showed IBH_s ranging from 0.4103 to 0.4114, indicating that R10 does not belong to this clone. Apart from R10, the ramet pairs among the other five "RED" ramets displayed IBH_s greater than 0.95 (Fig. 2). The ramet pairs composed of ramets from different clones showed IBH_s much smaller than 0.95, ranging from 0.4790 to 0.4803.

Discussion

Many species can propagate clonally to such an extent that the whole population is dominated by one or a few clones, such as fungi (Smith et al., 1992), seagrass species *Zostera marina* (Reusch et al., 1999) and *Posidonia oceanica* (Arnaud-Haond et al., 2012) and corals (Japaud et al., 2015). The situation of having a collection of samples belonging to one single clone (Yu et al., 2020) is thus not as rare as previously thought. The currently available methods cannot work in this extreme case (Arnaud-Haond et al., 2007). We have proposed the concept of identity by heterozygosity (IBH), meaning that two diploid genomes are identically heterozygous at some genetic markers. Based on IBH, we have developed IBH similarity (IBH_s), a novel measure of pairwise genetic similarity. The core of our IBH method for detecting clonemates is to use IBH_s in combination with the cutoff of 0.95 when N_{IBH} is on the order of >=3000. Two samples with IBH_s >0.95 are considered members of the same clone, while two samples with IBH_s <=0.95 are considered to represent different clones. Importantly, this method can be applied to any pair of ramets even if there is only one genet occurring within a "population"-level sampling. To the best of our knowledge, this is the only method that can work in cases where all collected ramets belong to one single clone.

One possible reason for the "contaminant" ramet (i.e., R10) in the clone “RED” is that R10 was produced by sexual event between members of the clone “RED” (i.e., selfing). In the first 15 years of growth, this would be the only way for contamination to occur, as the clones were cultured in isolated containers. Under selfing, each of the heterozygous loci in the clone “RED” can either remain identically heterozygous or change to a homozygous state in R10 with equal probabilities. As a consequence, the IBH_{R10} for the five ramet pairs composed of R10 and one of the other five “RED” ramets should be close to 1. However, that is not the case, as IBH_{R10} ranges from 0.4320 to 0.4331. Hence, R10 is not likely to be derived from selfing. In late 2019 or early 2020, these clones were transferred and planted in separate containers within in a single larger tank (~4 m diameter). All clones were in the same tank for at most one year. It is possible that either (a) outcrossing occurred and seeds set and germinated or (b) shoots dislodged and re-attached or were able to grow between pots.

The application of IBH_s requires a sufficient number of genetic markers where the two target samples are identically heterozygous (i.e., N_{IBH}). With the development of sequencing technology, it is now easy to obtain thousands of genetic markers using whole-genome resequencing, restriction-site associated DNA (RAD) sequencing, or SNP arrays. One should be careful when N_{IBH} is small. For example, with $N_{IBH} = 30$, even 2 heterozygous genotypes caused by somatic mutations or genotyping errors will lead to IBH_s of 0.94, and thus the two samples would be wrongly categorized as different genets. We suggest this method will be reliable only when N_{IBH} is on the order of >=3,000, based on comparing our results from different subsets of SNPs (Fig. 2).

IBH can potentially be used to detect possible parent-descendant pairs under selfing, because this reproductive mode leads to a predictable loss of heterozygosity. Under selfing, on average half of the heterozygous loci in the parent will change to a homozygous state in the descendants. If not considering the ramet-specific heterozygous loci derived from mutations and genotyping errors, all the heterozygous loci in the descendant will be identically heterozygous in the parent, while there will be a number of loci that are only heterozygous in the parent. The predicted loss in heterozygosity can potentially be used to detect possible parent-descendant pairs under selfing.

Our IBH method takes advantage of the availability of the larger number of genetic markers in the genomic era, and extends the detection of clonemates to any pair of samples. It has filled the gap in detecting clonemates in extreme cases where all collected samples belong to one single clone.

Samping permit
The original sampling of *Zostera marina* clones was conducted under the permit no. SC-4874.

Data availability
DNA sequence data are available in the NCBI short read archive, BioProject no. PRJNA806459, SRA accession no. SRR18000159- SRR18000170.

Code availability
Custom-made computer code is available at Github https://github.com/leiyu37/Detecting-clonemates.

Acknowledgements
This study was supported by a 4-year PhD scholarship from the China Scholarship Council (CSC) to L.Y. (No. 201704910807). We thank D. Gill for lab assistance. We thank Y. Li for the comments on the early version of the manuscript.

Authors' contributions
L.Y. and T.B.H.R. conceived and designed the study, J.J.S and K.D. cultured and provided the samples of the two 17-year-old seagrass *Zostera marina* clones, L.Y. extracted DNA and analyzed the sequencing data, L.Y. performed the bioinformatic analyses, L.Y. and T.B.H.R. wrote the paper, with input from all authors.

Competing interests
The authors declare that they have no competing interests.

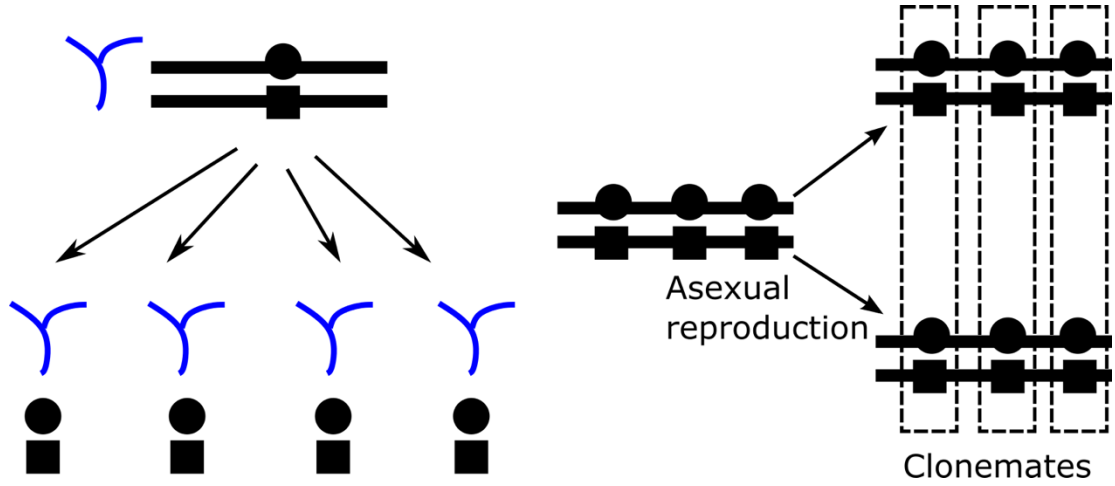

Figure 1: The maintenance of heterozygous genotypes under asexual reproduction. Under asexual reproduction, the heterozygous loci in the parent will remain identically heterozygous in all asexual descendants, except for rare somatic mutations.

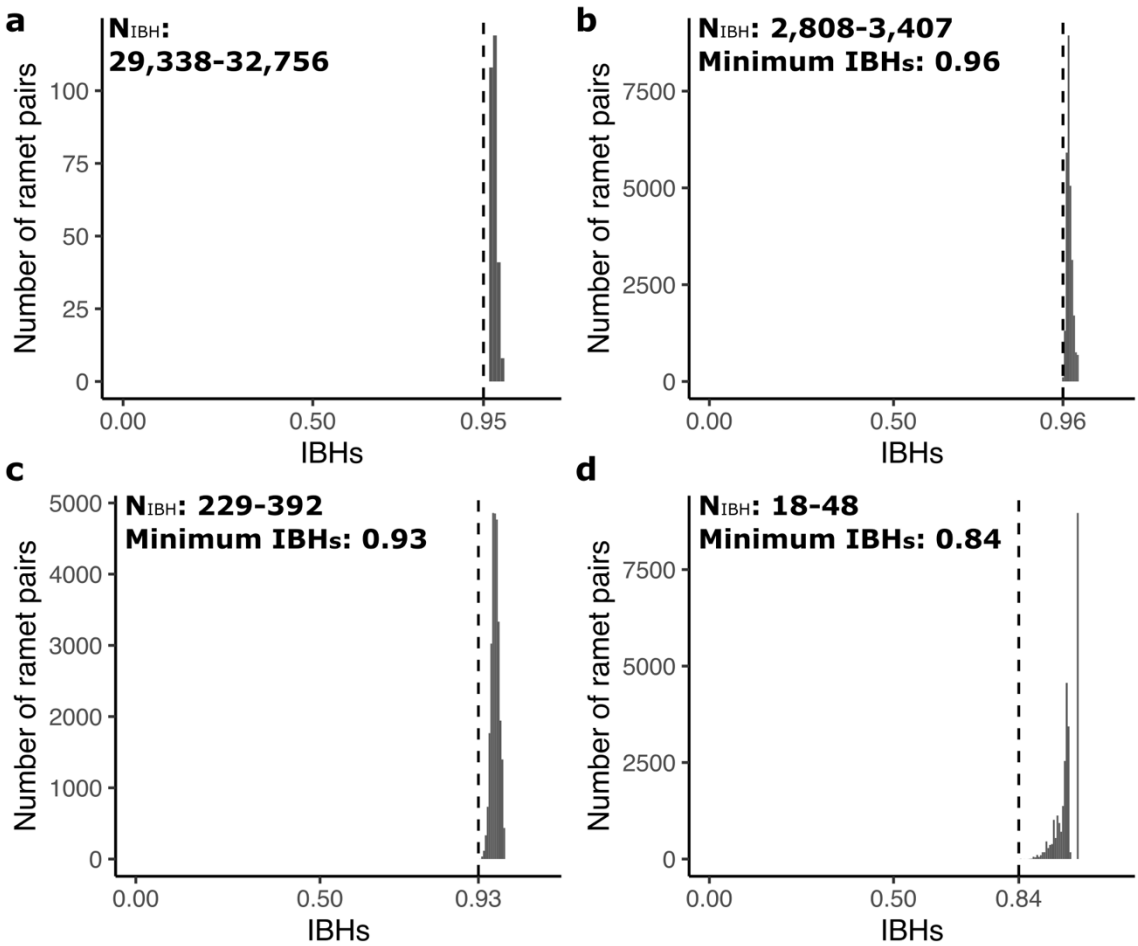

Figure 2: Frequency distribution of IBH similarity (IBH_s) for *Zostera marina* ramet pairs among the 24 ramets beloging to a clone of 750 to 1,500 years old. a, Frequency distribution based on all 38,831 SNPs. b, Frequency distribution based on 1/10 of the SNPs. c, Frequency distribution based on 1/100 of the SNPs. d, Frequency distribution based on 1/1000 of the SNPs.

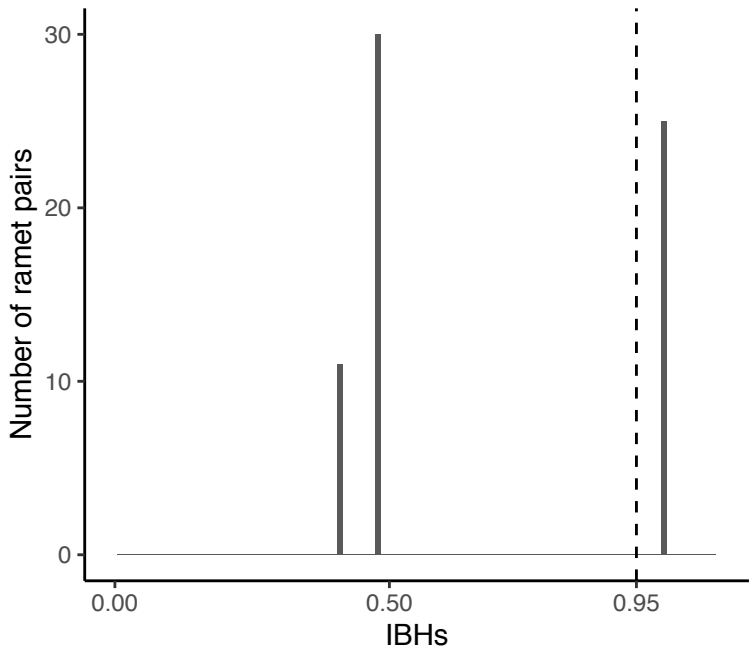

Figure 3: Frequency distribution of IBH similarity (IBH_s) for ramet pairs among the 12 cultured *Zostera marina* ramets.

Table 1: Coverage of the whole-genome resequencing data for the two cultured *Zostera marina* clones.

Sample	Average depth for the covered region	Genome region covered (%)
G01	68.64	93.33
G02	80.42	93.66
G06	79.05	93.72
G07	69.94	93.36
G10	77.49	93.52
G16	73.31	93.26
R02	81.55	93.86
R03	85.13	94.06
R05	68.14	93.57
R08	66.44	93.51
R09	71.73	93.59
R10	67.97	93.69

References

Arnaud-Haond, S., Duarte, C. M., Diaz-Almela, E., Marbà, N., Sintes, T. & Serrão, E. A. (2012). Implications of extreme life span in clonal organisms: millenary clones in meadows of the threatened seagrass *Posidonia oceanica*. *PLoS ONE, 7*(2), e30454. doi:10.1371/journal.pone.0030454

Arnaud-Haond, S., Duarte, C. M., Alberto, F. & Serrao, E. A. (2007). Standardizing methods to address clonality in population studies. *Mol. Ecol., 16*(24), 5115-5139. doi:10.1111/j.1365-294X.2007.03535.x

Bailleul, D., Stoeckel, S. & Arnaud-Haond, S. (2016). RClone: a package to identify MultiLocus Clonal Lineages and handle clonal data sets in r. *Methods Ecol. Evol., 7*(8), 966-970. doi:10.1111/2041-210X.12550

Buss, L. W. (1983). Evolution, development, and the units of selection. *Proc. Natl. Acad. Sci. U.S.A, 80*(5), 1387-1391. doi:10.1073/pnas.80.5.1387

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., . . . Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics, 27*(15), 2156-2158. doi:10.1093/bioinformatics/btr330

Douhovnikoff, V. & Dodd, R. S. (2003). Intra-clonal variation and a similarity threshold for identification of clones: application to Salix exigua using AFLP molecular markers. *Theor. Appl. Genet., 106*(7), 1307-1315. doi:10.1007/s00122-003-1200-9

Halkett, F., Simon, J. & Balloux, F. (2005). Tackling the population genetics of clonal and partially clonal organisms. *Trends Ecol. Evol. , 20*(4), 194-201. doi:10.1016/j.tree.2005.01.001

Harper, J. L. & White, J. (1974). The demography of plants. *Annu. Rev. Ecol. Syst., 5*(1), 419-463. doi:10.1146/annurev.es.05.110174.002223

Honnay, O. & Bossuyt, B. (2005). Prolonged clonal growth: escape route or route to extinction? *Oikos, 108*(2), 427-432. doi:10.1111/j.0030-1299.2005.13569.x

Hughes, R. A., Stachowicz, J. J. & Williams, S. L. (2009). Morphological and physiological variation among seagrass (*Zostera marina*) genotypes. *Oecologia, 159*(4), 725-733. doi:10.1007/s00442-008-1251-3

Japaud, A., Bouchon, C., Manceau, J. & Fauvelot, C. (2015). High clonality in *Acropora palmata* and *Acropora cervicornis* populations of Guadeloupe, French Lesser Antilles. *Mar. Freshw. Res., 66*(9), 847-851. doi:10.1071/MF14181

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics, 27*(21), 2987-2993. doi:10.1093/bioinformatics/btr509

Li, H. & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics, 25*(14), 1754-1760. doi:10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics, 25*(16), 2078-2079. doi:10.1093/bioinformatics/btp352

Ma, X., Olsen, J. L., Reusch, T. B. H., Procaccini, G., Kudrna, D., Williams, M., . . . Van de Peer, Y. (2021). Improved chromosome-level genome assembly and annotation of the seagrass, *Zostera marina* (eelgrass). *F1000research, 10*. doi:10.12688/f1000research.38156.1

Meirmans, P. G. (2020). genodive version 3.0: Easy-to-use software for the analysis of genetic data of diploids and polyploids. *Mol. Ecol. Resour., 20*(4), 1126-1131. doi:10.1111/1755-0998.13145

Noble, J. C., Bell, A. D. & Harper, J. L. (1979). The population biology of plants with clonal growth: I. The morphology and structural demography of *Carex arenaria*. *J. Ecol., 67*, 983-1008. doi:10.2307/2259224

Parks, J. C. & Werth, C. R. (1993). A study of spatial features of clones in a population of bracken fern, *Pteridium aquilinum* (Dennstaedtiaceae). *Am. J. Bot., 80*(5), 537-544. doi:10.1002/j.1537-2197.1993.tb13837.x

Quinlan, A. R. & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics, 26*(6), 841-842. doi:10.1093/bioinformatics/btq033

Reusch, T. B. H., Boström, C., Stam, W. T. & Olsen, J. L. (1999). An ancient eelgrass clone in the Baltic. *Mar. Ecol. Prog. Ser., 183*, 301-304. doi:10.3354/meps183301

Smith, M. L., Bruhn, J. N. & Anderson, J. B. (1992). The fungus *Armillaria bulbosa* is among the largest and oldest living organisms. *Nature, 356*(6368), 428-431. doi:10.1038/356428a0

Van der Auwera, G. A. & O'Connor, B. D. (2020). *Genomics in the cloud: using Docker, GATK, and WDL in Terra*: O'Reilly Media.

Yu, L., Boström, C., Franzenburg, S., Bayer, T., Dagan, T. & Reusch, T. B. H. (2020). Somatic genetic drift and multilevel selection in a clonal seagrass. *Nat. Ecol. Evol., 4*(7), 952-962. doi:10.1038/s41559-020-1196-4