

Títol del treball: Parameter validation for a genomics population analysis

Nom estudiant: Paloma Marie Buj Douirin

Correu electrònic: palomabujdouirin@gmail.com

Grau en: Doble titulació en biologia i biotecnologia

Nom del tutor: Jordi Viñas de Puig

Correu electrònic: jordi.vinas@udg.edu

INDEX

RESUM	i
RESUMEN	iii
ABSTRACT	v
REFLEXIONS SOBRE ÈTICA, SOSTENIBILITAT I PERSPECTIVA DE GÈNERE	vi
INTRODUCTION	1
1.1. The project	1
1.1.1. Sarda sarda (Bonito)	1
1.1.2. Importance of the genetic variability in non-model species of commercial interest, especially fisheries.	1
1.2. Restriction associated DNA sequencing	2
1.2.1. The development of RADseq	2
1.2.2. Double digest RADseq (ddRADseq)	6
1.2.3. The sequencing revolution	8
1.3. The Stacks software	11
1.3.1. The parameters	19
1.4. The Galaxy platform	21
GOALS: parameter validation	22
MATERIALS AND METHODS	23
2.1. Previously: sample collection and processing	23
2.2. Tutorials	23
2.3. Using stacks on Galaxy	24
2.4. The parameter setting	25
RESULTS	28
3.1. Polymorphic and variable sites	28
3.2. Coverage	29
DISCUSSION	33
CONCLUSIONS	36
BIBLIOGRAPHY	37

RESUM

Un estudi de Ollé, J. i Viñas, J. te com a finalitat ajudar crear una normativa per la regulació del Bonítol Atlàntic ja que es tracta d'una espècie molt sobre pescada que es troba al Mediterrani, les costes de la península Ibèrica i del nord-oest del continent africà. Al tractar-se d'una espècie d'interès comercial, es necessari determinar quines poblacions en formen part per tal de poder crear un pla de conservació de la biodiversitat i evitar la pèrdua de variabilitat genètica d'aquesta espècie.

Per tal de poder realitzar això, s'han col·lectat mostres de 92 individus de : Tunísia, Espanya, nord de Portugal, sud de Portugal, Marroc, Mauritània, Senegal i Costa D'Ivori. I s'ha realitzat la tècnica de seqüenciació del DNA associada a restricció (RADseq) que permet genotipar tots els individus i veure diferències genètiques entre poblacions. Aquesta informació genètica seqüenciada ha de passar per un programa d'assemblatge anomenat Stacks del qual en formen part els paràmetres m , M i n , que determinen com es produiran aquest assemblatges i influencien sobre la cobertura i el nombre de polimorfismes retrobats.

Per tant, en aquest estudi, s'investiga com influencien aquest paràmetres m , M i n , sobre una representació petita de les poblacions, es a dir tres individus per població, per tal de determinar quins son els valors més eficients a l'hora de recuperar el màxim de llocs polimòrfics veritables amb una alta cobertura de les seqüències. De tal manera que es generen onze testos amb diferents combinacions dels paràmetres corresponents als valors mitjans del rang de valors que poden prendre.

Després de la recollida de dades de cobertura i llocs polimòrfics obtingudes per Stacks, s'empra la prova estadística de Kruskal-Wallis, per a dades no paramètriques, que desvela diferències significatives entre els testos ja sigui per les dades polimòrfiques com per les de cobertura. En el cas de la cobertura s'observa que depèn principalment del paràmetre m , que tot hi haver diferències significatives entre el major i el menor valor assignat a m , les dades ja presenten una alta cobertura. Finalment les dades dels llocs polimòrfics revela moltes més diferències significatives entre grups que depenen sobretot de M i n .

La conclusió més important d'aquest projecte és que per cada grup de dades sorgides de RADseq, s'ha de realitzar un pas previ de validació de paràmetres ja que els valors òptims poden canviar en funció de l'espècie. En el cas de les dades utilitzades en aquest projecte, no hi ha una única combinació de paràmetres correcta i per tant les resultats obtinguts serveixen de guia per a futurs projectes en els que s'utilitzaran dades de Bonítol Atlàntic o

espècies properes, a l'hora de decidir quins valors són més òptims per aquell conjunt de dades.

RESUMEN

Un estudio de Ollé, J. y Viñas, J. tiene como finalidad ayudar a crear una normativa para la regulación del Bonito Atlántico, ya que se trata de una especie muy sobreexplotada que se encuentra en el Mediterráneo, las costas de la península Ibérica y el noroeste del continente africano. Al tratarse de una especie de interés comercial, es necesario determinar qué poblaciones forman parte de ella para poder crear un plan de conservación de la biodiversidad y evitar la pérdida de variabilidad genética de esta especie.

Con el fin de llevar a cabo esto, se han recolectado muestras de 92 individuos de Túnez, España, norte de Portugal, sur de Portugal, Marruecos, Mauritania, Senegal y Costa de Marfil. Se ha realizado la técnica de secuenciación de ADN asociada a restricción (RADseq), que permite genotipar a todos los individuos y observar las diferencias genéticas entre poblaciones. Esta información genética secuenciada debe pasar por un programa de ensamblaje llamado Stacks, en el cual se incluyen los parámetros m , M y n , que determinan cómo se producirán estos ensamblajes y afectan a la cobertura y el número de lugares polimórficos encontrados.

Por lo tanto, en este estudio se investiga cómo influyen estos parámetros m , M y n en una representación pequeña de las poblaciones, es decir, tres individuos por población, para determinar cuáles son los valores más eficientes a la hora de recuperar el máximo de loci polimórficos verdaderos con una alta cobertura de las secuencias. De esta manera, se generan once ensayos con diferentes combinaciones de los parámetros correspondientes a los valores promedio del rango de valores que pueden tomar.

Después de recopilar los datos de cobertura y loci polimórficos obtenidos por Stacks, se utiliza la prueba estadística de Kruskal-Wallis, para datos no paramétricos, que revela diferencias significativas entre los ensayos tanto en los datos polimórficos como en la cobertura. En el caso de la cobertura, se observa que depende principalmente del parámetro m , y aunque existen diferencias significativas entre el valor mayor y menor asignado a m , los datos ya presentan una alta cobertura. Finalmente, los datos de los lugares polimórficos revelan muchas más diferencias significativas entre grupos, que dependen principalmente de M y n .

La conclusión más importante de este proyecto es que para cada grupo de datos surgidos de RADseq, se debe realizar un paso previo de validación de parámetros, ya que los valores óptimos pueden cambiar en función de la especie. En el caso de los datos utilizados en este proyecto, no hay una única combinación de parámetros correcta, por lo que los resultados

obtenidos sirven como guía para futuros proyectos en los que se utilizarán datos de Bonito Atlántico o especies similares, a la hora de decidir qué valores son más óptimos para ese conjunto de datos.

ABSTRACT

A study by Ollé, J. and Viñas, J. aims to help create guidelines for the regulation of Atlantic Bonito, as it is an overfished species found in the Mediterranean, the coasts of the Iberian Peninsula, and the northwestern African continent. Being a commercially important species, it is necessary to determine which populations are part of it in order to create a conservation plan for biodiversity and prevent the loss of genetic variability in this species.

To achieve this, samples have been collected from 92 individuals from Tunisia, Spain, northern Portugal, southern Portugal, Morocco, Mauritania, Senegal, and Ivory Coast. The restriction associated DNA sequencing (RADseq) technique has been employed to genotype all individuals and observe genetic differences among populations. This sequenced genetic information needs to go through an assembly program called Stacks, which includes the parameters m , M , and n that determine how these assemblies will be produced and influence the coverage and number of detected polymorphic sites.

Therefore, this study investigates how these parameters m , M , and n influence a small representation of the populations, namely three individuals per population, to determine the most efficient values for recovering the maximum number of true polymorphic loci with high sequence coverage. Consequently, eleven tests are generated with different combinations of parameters corresponding to the mean values within the range of possible values.

After collecting coverage and polymorphic loci data obtained from Stacks, the non-parametric statistical Kruskal-Wallis test is used, which reveals significant differences between the tests, both in terms of polymorphic data and coverage. In the case of coverage, it is observed that it mainly depends on the parameter m . Although there are significant differences between the highest and lowest assigned values of m , the data already shows high coverage for all the m values. Finally, the polymorphic site data reveal many more significant differences between groups, primarily dependent on M and n .

The most important conclusion of this project is that for each set of RADseq data, a prior parameter validation step must be carried out since optimal values can vary depending on the species. In the case of the data used in this project, there is no single correct combination of parameters. Therefore, the obtained results serve as a guide for future projects using Atlantic Bonito data or related species when deciding the most optimal values for that specific data set.

REFLEXIÓ SOBRE ÈTICA

Quan es realitzen estudis poblacionals en espècies salvatges s'han de realitzar mostrejos al camp, en aquest punt es troba una discussió ètica sobre com es realitza el mostreig, com es tracten els animals involucrats i el sofriment o lesions que poden patir.

Els peixos, tractats en aquest TFG, es veuen afectats per l'antropocentrisme, ja que en les últimes dècades, el benestar dels animals en recerca a millorat però s'ha centrat cap als mamífers per que son animals més semblants a l'humà de manera que tenim és empatia cap a ells (Mather, 2019). En peixos ossis s'ha demostrat que tenen un sistema de recepció del dolor similar al dels mamífers i canvien el comportament i fisiologia enfront a un estímul dolorós (Sneddon, 2015)

Per tant, queda demostrat que els peixos son capaços de sentir dolor tant com els mamífers així que l'investigador ha de ser responsable del seu comportament cap a aquests animals i vetllar pel seu benestar.

REFLEXIÓ DE SOSTENIBILITAT

La sostenibilitat es veu involucrada en la recerca ja sigui en la conservació d'ecosistemes a l'hora de realitzar mostrejos al camp o en el moment de fer experiments al laboratori.

Una manera de evitar el malbaratament d'energies i recursos és evitant crear residus que s'han de tractar ja sigui reutilitzant materials o reciclant-los. Per això en els darrers anys s'han desenvolupat eines d'experimentació *in silico* que presenten avantatges en front a l'experimentació *in vivo* habitual dels laboratoris, com serien; menor risc per als essers vius, dades il·limitades, més econòmiques, en menys tems i recursos. Tot hi que la principal limitació es que sol s'aproxima a la realitat (Badano, 2021).

Per tant, els assajos computacionals son molt útils per tal de reduir el temps que es passa al laboratori ja que, fent un estudi previ *in silico* permet acotar el nombre d'experiments necessaris amb animals vius.

REFLEXIÓ SOBRE PERSPECTIVA DE GÈNERE

Avui en dia encara existeixen diferències entre homes i dones en l'espai laboral que poden arribar a desencadenar en assetjament a la feina, actituds masclistes, menor representació de dones en llocs de lideratge, bretxes salarials... Aquestes disparitats poden ser degudes a factors com ara estereotips de gènere, biaixos inconscients, manca de models a seguir i barreres institucionals.

Aquest fets es corroboren amb estudis que indiquen que les dones es troben poc representades com a post-doctorades en els camps de l'enginyeria i ciències naturals a més que son més propenses a treballar a temps parcial que els homes (Waijjer et al., 2016). Un altre estudi demostra que dones tenen menys probabilitats de continuar una carrera com a investigadores editorials enfront als homes. A més, a les disciplines biomèdiques, els homes tenen un 25% més de probabilitats que les dones de ser l'últim autor d'una publicació, cosa que suggereix que els homes tendeixen a tenir més càrrecs que les dones (Boekhout et al., 2021).

En resum, és crucial abordar les diferències de gènere en la recerca científica i treballar cap a la igualtat d'oportunitats. Això implica fomentar entorns inclusius i lliures de discriminació, promoure polítiques d'igualtat de gènere a les institucions acadèmiques i científiques.

INTRODUCTION

1.1. The project

The ICCAT (International Commission for the Conservation of the Atlantic Tuna) is an intergovernmental fishing organization responsible for the conservation of the tunas and related species from the Atlantic ocean and adjacent seas (ICCAT·CICTA·CICAA, n.d.). Unfortunately, the ICCAT commission does not have a plan for the conservation of the Atlantic Bonito therefore, Ollé, J., & Viñas J., launched a project aimed at the genotyping of this species in order to create a plan for the conservation of the populations, biodiversity and genetic variability.

1.1.1. Atlantic Bonito (*Sarda sarda*, Bloch 1793)

Atlantic bonito (*Sarda sarda*, Bloch 1793) is an epi-pelagic species of small tuna. It has a common size of 50 cm fork length, weighs about 2 kg and lives approximately 5 years in the wild. It lives between 80 to 200 m of depths and preys sardine, anchovy, mackerel and small pelagic species (Nøttestad et al., 2013). It is found in the tropical and subtropical Atlantic Ocean, Gulf of Mexico, Mediterranean Sea and Dead Sea.

Although, Atlantic bonito is a fish of economic interest since it is one of the most fished averaging 30.000 tons a year from 1950 to 2019, there are no regulations in place to guarantee an international cooperation for its conservation. In 2019, 67% of the Bonitos fished came from the Atlantic ocean and the Mediterranean sea, being the latest where most bonitos were fished (Mourato et al., 2021).

For wild species of fish, genetic studies and assessments have increased in recent years, since that information is relevant for the management of stocks. Many species of marine wildlife have tight regulations due to those studies however, the Atlantic Bonito is not one of them. The main reason for the conservation of wild population of fisheries species is the loss of genetic variability due to overexploitation that, in the end, can lead to the decline of their adaptation capacities (Viñas et al., 2011).

1.1.2. Importance of the genetic variability in non-model species of commercial interest, especially fisheries.

Thus, the new genomic approaches enable non-model species to be genotyped and discover their genetic variance in front of the different areas where they are fished, allowing to discover markers that characterize the populations and identify local adaptation or speciation events (Rodríguez-Ezpeleta et al., 2016). Therefore, genomic population studies are essential for the correct fishery management for long term fishery sustainability (Mejía-

Ruíz et al., 2020), thanks to the identification of conservation units (Rodríguez-Ezpeleta et al., 2016).

A widely used genomic approach for genotyping large non-model marine populations is, Restriction DNA Associated Sequencing (RADseq) that allows the obtention of whole genome SNP without the whole genome sequencing, making it a more cost-effective approach.

1.2. Restriction associated DNA sequencing

At the beginning of the sequencing revolution, single nucleotide polymorphism (SNP) chip microarray-based platforms were the low-cost genotyping methods most used but they required prior knowledge of sequence and variability that ultimately produced some bias and hindered the detection of rare or population specific variants (Peterson et al., 2012). Other markers of variation used are microsatellites and indel polymorphisms however they have a long processing time, are expensive and only generate a few working markers. Thus, an alternative could be whole genome sequencing (WGseq) but in eukaryotes this method is not suitable since they have complex and large genomes (Peterson et al., 2012) (Davey & Blaxter, 2010).

Nevertheless, restriction associated DNA sequencing (RADseq, also known as genotyping-by-sequencing) is a new technique for SNP detection that solves those problems since it only sequences certain areas of the genome that are adjacent to the restriction sites. This allows to reduce the complexity of the genome and still deliver high throughput data of genetic population markers and reduce the cost of library generation by 5 to 10 fold compared to commercial libraries (Inbar et al., 2020). This method can be used in many areas such as; population genetics inferences, association mapping, genetic mapping, and in estimation of allele frequencies (Shen, 2019b). However, in the case of large evolutionary distances, there is a higher variability of restriction sites between taxa, that makes harder securing enough homologous loci at distant phylogenetic scales (Dodsworth et al., 2019).

As mentioned above, the RADseq method consist on the digestion of the genome with specific restriction enzymes and sequencing the DNA next to those sites, making the loci basically random and short. There is the possibility of partial selection of those loci using methylation-sensitive enzymes (Dodsworth et al., 2019).

1.2.1. The development of RADseq

Eric Jonson and his team first developed the restriction site associated DNA (RAD) marker genotyping technique to find markers across the genome in an easy way for both model and non-model organisms (Miller et al., 2007) (*RAD-Seq Genotypes Less, But Offers More*, 2011).

It consists on the digestion of the genomic DNA with a specific restriction enzyme (figure 1.1) and ligation to biotinylated linkers (figure 1.2). After that, the DNA is cut into smaller fragments than the distance between restriction sites (figure 1.3) so, only the fragments that are closest to de restriction site are attached to the linkers (Miller et al., 2007).

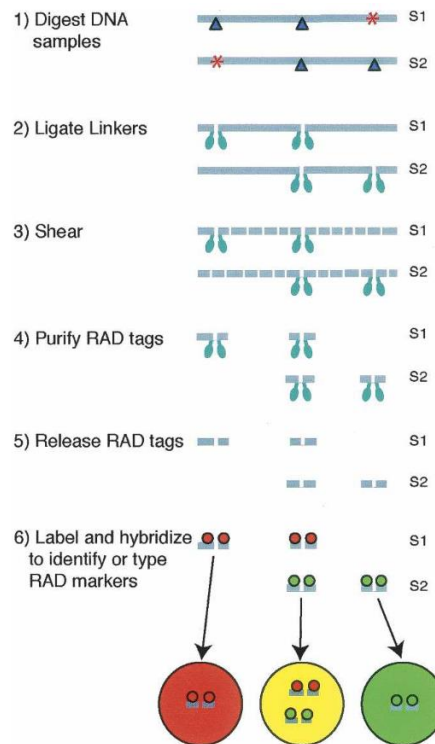


Figure 1. Schematic process for the obtention of RADtags (Miller et al., 2007).

Then, the rest of the DNA is removed at the same time that the fragments are immobilized in streptavidin beads (figure 1.4). Another digestion is used with the same restriction enzyme to release the fragments from the beads (figure 1.5). During this process, the DNA tags flanking this specific restriction sites are isolated (Miller et al., 2007).

Afterwards, these RAD markers can be paralleled screened by detecting differential hybridization patterns of RADtag samples on a microarray (figure 1.6) (Miller et al., 2007) allowing for the mapping of natural variation and mutations in many distinct organisms (RAD-Seq Genotypes Less, But Offers More, 2011).

The quantity of distinct informative markers associated with an enzyme relies on the frequency of SNPs and the genome's size. In the case of a 6-base pair sequence-recognizing restriction enzyme, it is projected that an informative RAD marker will occur approximately once every 100kb on average among individuals or strains with an average SNP frequency of 0.5%.(Miller et al., 2007).

Interestingly, a year after the development of this method, the Illumina Genome Analyzer™ System was launched to the public and they took the opportunity to adapt their RADtag libraries to be compatible with the sequencing platform, resulting in RADseq allowing for massive parallel sequencing and the discovery of thousands of SNPs (*RAD-Seq Genotypes Less, But Offers More*, 2011).

Restriction site associated sequencing (RADseq) is similar to the; restriction fragment length polymorphisms (RFLPs), amplified fragment length polymorphisms (AFLPs) and random amplified polymorphic DNA (RAPD) analyses since they all employ restriction enzymes in order to reduce the complexity of the genome. Nonetheless, RADseq presents some advantages in front of the other methods as it surpasses them in its ability to identify, verify and score markers concurrently and to robustly identify which markers derive from each site (Davey & Blaxter, 2010) (Shen, 2019a).

RADSeq is applicable on any design type of crosses and in wild populations. It facilitates not only genotyping and SNP discovery but also permits more intricate analyses, such as quantitative genetic and phylogeographic studies (Davey & Blaxter, 2010).

RADseq utilizes a similar methodology to RADtags. Initially, the genome undergoes shearing into fragments with sticky ends using a specific restriction enzyme (figure 2.A). These fragments are then connected to P1 adapters, which not only match the ends of the target fragments but also incorporate a Molecular Identifier (MID) sequence for individual identification, these adapters bind to an Illumina flow cell (figure 2.B). Next, the labelled restriction fragments from multiple individuals are combined and randomly trimmed to produce fragments with an average length of a few hundred base pairs (figure 2.C). These trimmed fragments are subsequently connected to a second P2 adapter and amplified through PCR using P1 and P2 primers (figure 2.D) (Davey & Blaxter, 2010).

It is important to note that the P2 adapter possesses a distinct 'Y' configuration, which remains unresponsive to the P2 primer until it undergoes amplification by the P1 adapter (figure 2.E). This guarantees that all amplified fragments consist of the P1 adapter, the MID, a partial restriction site, a few hundred bases of flanking sequence, and the P2 adapter. (Davey & Blaxter, 2010).

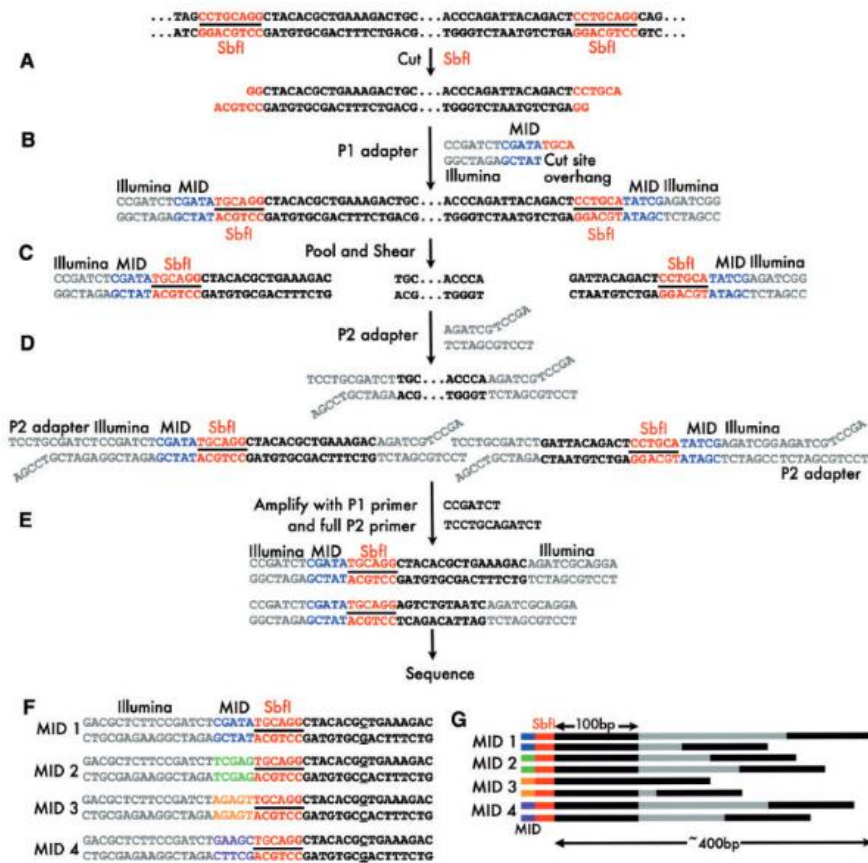


Figure 2. Schematic view of the Restriction Associated DNA sequencing process (Davey & Blaxter, 2010).

Ultimately, the fragments are prepared for sequencing. However, before that, they undergo size selection to retain fragments ranging from 200 to 500 bases. Subsequently, the RADseq library is sequenced using the Illumina platform (figure 2.F). The resulting sequence originates from the MID within the P1 adapter and extends across the restriction enzyme site (figure 2.G). This process generates a dataset of sequences downstream of the restriction sites, known as RAD tags, which represent a significantly reduced portion of the original genome. In the case of a symmetric restriction site, two RAD tags are generated from each site. After sequencing, the sequences from each individual are segregated using the MID for analysis and interpretation (Davey & Blaxter, 2010).

Originally, the Illumina platform enabled sequencing up to 150 bases, allowing screening of approximately 300 bases surrounding each restriction site for polymorphisms (Davey & Blaxter, 2010). Nowadays, adapter-ligated restriction fragments are fragmented within the range of 300 to 700 base pairs for compatibility with the Illumina platform (Shen, 2019a).

For wild populations, it is anticipated that the level of diversity is around 0.1%. Therefore, approximately 20-30% of all restriction sites are expected to be accompanied by a polymorphism within the adjacent 200-300bp of sequence (Davey & Blaxter, 2010).

The primary purpose of RADseq is to identify polymorphisms in the form of restriction site presence-absence, SNPs, and indels located in the sequence flanking the restriction site. If a reference genome is available, the sequence reads can be aligned to it, allowing for the automatic correction of sequencing errors present in the reads. On the other hand, in the absence of a reference genome the RAD tags are analysed de novo. This involves grouping identical reads into unique sequences, treating them as potential alleles. These clusters typically exhibit a few mismatches among them. By comparing the counts of each base at each position, SNPs and indels can be identified between alleles at the same genomic location, and errors can be corrected. Moreover, genuinely homozygous or heterozygous alleles tend to have relatively high read counts, while errors display lower counts (Davey & Blaxter, 2010).

1.2.2. Double Digest RADseq (ddRADseq)

The double digest restriction associated DNA sequencing (ddRADseq) technique is based on the RADseq method explained above therefore, the protocol is similar in terms of addition of adapters, amplification, and sequencing, but differs in the use of two restriction enzymes instead of one. One of the restriction enzymes has a frequent restriction site in the genome and the other is rarer. Additionally, there is a precise size selection step prior to library amplification to further reduce representation of the genome during the library construction step (Magbanua et al., 2023).

Therefore, ddRADseq offers some advantages over RADseq such as elimination of random shearing and end repair of genomic DNA by utilizing two restriction enzymes simultaneously. This approach results in a minimum of 5 fold reduction in library production cost. Moreover, the accurate selection of genomic fragments based on size enables enhanced control over the specific regions represented in the final library at a finer scale. By combining precise and consistent size selection with sequence-specific fragmentation, ddRADseq creates sequencing libraries that solely comprise a subset of genomic restriction digest fragments generated by cuts made with both restriction enzymes. (Peterson et al., 2012).

Since this technique uses two restriction enzymes, not many fragments are of the desired length during the size selection step thus, there is allow likelihood of sampling both directions of the same restriction site. This in return means that the duplicate sampling

region is reduced on top of reducing the number of reads required to achieve a high-confidence SNP sampling related to a certain restriction enzyme cut size (Peterson et al., 2012).

Additionally, there is a shared bias in the representation of regions, which tends to favor fragments that are closer to the average size selection. As a result, independent samples have the tendency towards recovering the same genomic regions. Consequently, the recovery of these regions occurs in a similar order across all individual samples. Even samples with lower read recovery counts, before reaching saturation, will still exhibit a significant number of well-covered regions in common (Peterson et al., 2012).

When comparing simple RADseq to ddRADseq, the latter offers enhanced flexibility and robustness in region recovery while requiring fewer resources in terms of economics, genomic material, and time. (Peterson et al., 2012).

In figure 3 can be observed the representation of the distinctions between traditional RADseq and ddRADseq. The most obvious difference is the use of two restriction enzymes in the case of ddRADseq that allows to have a lower complexity of the genome since the fragment size is smaller and avoids sequencing of both sides of the restriction sites. Thus, in figure 3.B the sequence reads exclude very far (b) or very close (a) restriction sites (Peterson et al., 2012).

The number of reads in this library is expected to be higher for regions that closely match the desired size-selection target. This means that the read counts will likely be similar between individuals, as shown by the yellow and green bars in figure 3.B. In other words, if the genomic interval is close to the size selected, then there will be a high representation of sequence reads for all individuals, on the contrary, if the size of the genomic interval is higher or lower, then the representation of those reads in the library will be reduced (Peterson et al., 2012).

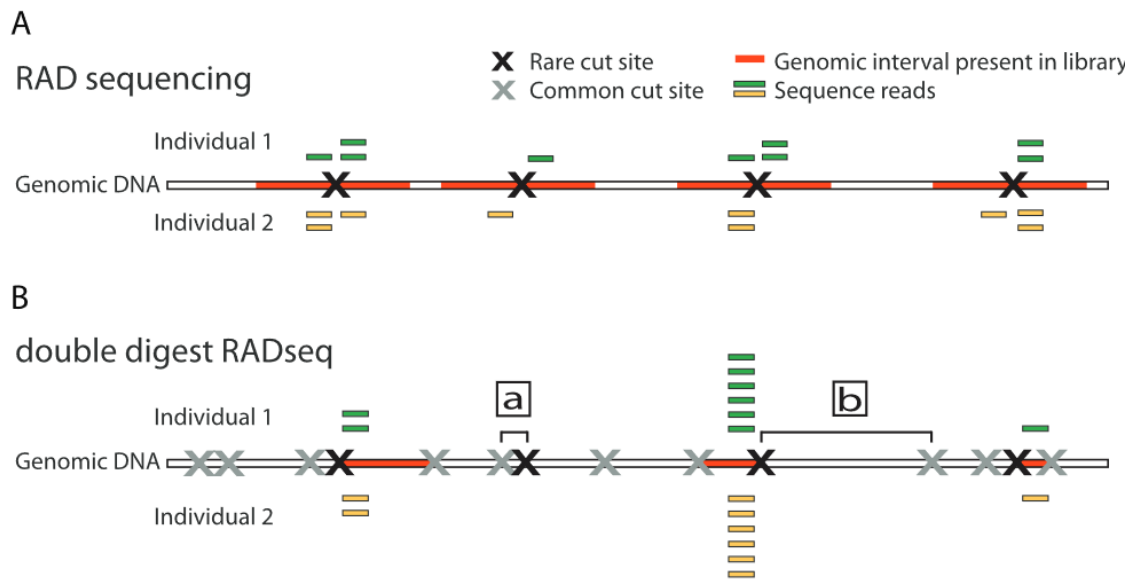


Figure 3. Schematic representation for the comparison of RADseq and ddRADseq (Peterson et al., 2012).

The RADseq method can be used in many areas such as in association mapping, population genetics inferences, genetic mapping, and in estimation of allele frequencies (Davey & Blaxter, 2010).

1.2.3. The sequencing revolution

In order to understand how organisms' function, it is important to understand their physiology to a molecular level. Ever since the discovery of the double-helical structure of DNA in 1953 (Watson & Crick, 1953) there has been an increased interest in understanding this molecule in a deeper level, hence the emergence of sequencing technologies.

The first generation of sequencing started with the Sanger method in 1975 (Sanger & Coulson, 1975) known as a sequencing by synthesis method (Kumar et al., 2019), after that the method was perfected in 1997 and used to sequence the first genome of a bacteriophage ϕ X174 (Sanger et al., 1977). The same year, Maxam and Gilbert's chemical chain termination method for DNA sequencing (Maxam & Gilbert, 1977) was also introduced (van Dijk et al., 2018). These techniques consist in mimicking the replication of DNA in the cell (McCombie et al., 2019).

These techniques operate by employing specific combinations of nucleotides in four distinct enzymatic extension reactions. In each reaction, a different dideoxynucleotide is incorporated, which acts as a stop signal when it is added to the growing chain. Initially, radioactive labeling was used for the dideoxynucleotides, but later, fluorescent labeling was adopted (Hu et al., 2021). After the completion of enzymatic reactions, a mixture of

molecules of different lengths is obtained due to the stopping action of dideoxynucleotides at various positions in the DNA chain. These molecules are then subjected to denaturation on a polyacrylamide gel, resulting in the formation of fragment ladders in separate lanes. Exposure of the gel to an X-ray film allows the visualization of these fragments. Each lane represents a different reaction mix with a specific dideoxynucleotide, and the fragments differ in length by one nucleotide. The sequencing pattern is read in a bottom-to-top fashion, enabling the deduction of the DNA sequence starting from the shortest fragment (McCombie et al., 2019).

However, this approach has its shortcomings since only one sequencing reaction can be analysed at a time, limiting the throughput (Hu et al., 2021). Therefore, in the initial projects where these techniques were used, they focused on the determination of single genes of small genomes (McCombie et al., 2019).

Nonetheless, the Human Genome Project started in October 1990 and was completed in April of 2003, being the largest collaborative biological project to date (van Dijk et al., 2018). From this project other collaborative projects emerged such as 1000 Genomes Project, the Cancer Genome Atlas, and the Human Microbiome Project (Green et al., 2015)

The second wave of sequencing technologies, also known as next-generation sequencing (NGS) emerged in the early two-thousands and are closely related to Sanger sequencing from the fundamental usage of enzymological underpinnings (McCombie et al., 2019). The advancements in these technologies have made it possible to sequence numerous individual DNA molecules in parallel, resulting in high throughput and short-read results. This is achieved by first fragmenting the DNA molecules and then amplifying and sequencing them. As a result, millions of these fragments are sequenced simultaneously when the complementary chain is synthesized. However, after sequencing, these short reads need to be reassembled for further analysis. (Hu et al., 2021).

The two prominent NGS platforms are Illumina and Ion Torrent. Illumina's NovaSeq 6000 offers paired-end reads with a read length of 150 bp while Ion Torrent's Ion 530 Chip provides a longer read length of 600 bp. Among these platforms, Illumina is considered the most competitive option because it offers the best value in terms of both cost and time (Kumar et al., 2019). Thus, commercial platforms that employ massively parallel sequencing are built on the principle of sequencing by synthesis. These platforms follow a common set of essential steps, which include DNA fragmentation, DNA end-repair, adapter ligation, surface attachment, and in-situ amplification. (Hu et al., 2021) (McCombie et al., 2019).

In order to reconstruct the fragmented sequence and differentiate between single nucleotide variants, indels, and sequencing errors, the utilization of read alignment algorithms is necessary. Currently, there are numerous platforms available that offer these services, each employing distinct principles to align individual short reads with an existing genome assembly for example STACKS (McCombie et al., 2019).

However, the sequencing by synthesis platforms have limitations in the length of sequence they can read. This is due to the accumulation of background noise at each step of nucleotide incorporation and detection, which increases with each cycle. (McCombie et al., 2019). Furthermore, the utilization of short reads in sequencing can result in mis-assemblies and gaps, particularly when the repeated sequences are longer than the read length provided by the platforms. Moreover, while single-nucleotide variations (SNVs) and short indels can be accurately identified, detecting larger structural variations can pose challenges. Additionally, regions with high GC% content are often inefficiently amplified by PCR, which can potentially compromise the sequencing process. (van Dijk et al., 2018).

In the recent years, a new wave of sequencing methods has emerged, creating the third generation of NGS capable to deal with some of the issues mentioned above. These approaches have the capacity of long-read (between >1kb to 2Mb) and real-time sequencing (Kumar et al., 2019)

These long-read technologies are Pacific Biosciences released in 2011 (single-molecule real-time sequencing) and Oxford Nanopore released in 2014 (introduced nanopore sequencing). The improvements that they offer are; absence of PCR amplification, real-time sequencing process, minimal library preparation steps and production of long reads (Hu et al., 2021) (van Dijk et al., 2018).

In figure 1 are represented the advantages that the third wave of NGS has over the second wave. The figure 4.A image represents two identical regions that have in between a sequence, when using short-reads, those identical sequences can be assembled in the same contig (blue) and the contig (green) that is in the middle cannot be placed either upstream or downstream of the contig 2 (blue). Similar problems happen with structural variants that involve repetitive regions. But in the case of using long-reads the hole sequence is sequenced in one long read and no ambiguities can cooccur (van Dijk et al., 2018).

In the figure 4.B image, two different haplotype single nucleotide polymorphisms (A or C) or larger variations (red or blue) that are too far to be read in the same short-read, can lead to ambiguous trajectories that result in fragmented assemblies (van Dijk et al., 2018). The figure 4.C image, has the representation of mRNA transcripts where short-reads will create

a pool of sequences in between exons therefore, some alternative splicing variants will be detected but the combination of exon junctions in the individual transcripts is lacking. However, long-read sequencing covers the whole transcript (van Dijk et al., 2018)..

Finally, the figure 4.D image represents how PCR amplification in regions with extreme GC content is inefficient, thus these regions will be poorly covered. Alternatively, Single-molecule real-time sequencing and nanopore long-read sequencing technologies do not require PCR amplification, solving this problem (van Dijk et al., 2018).

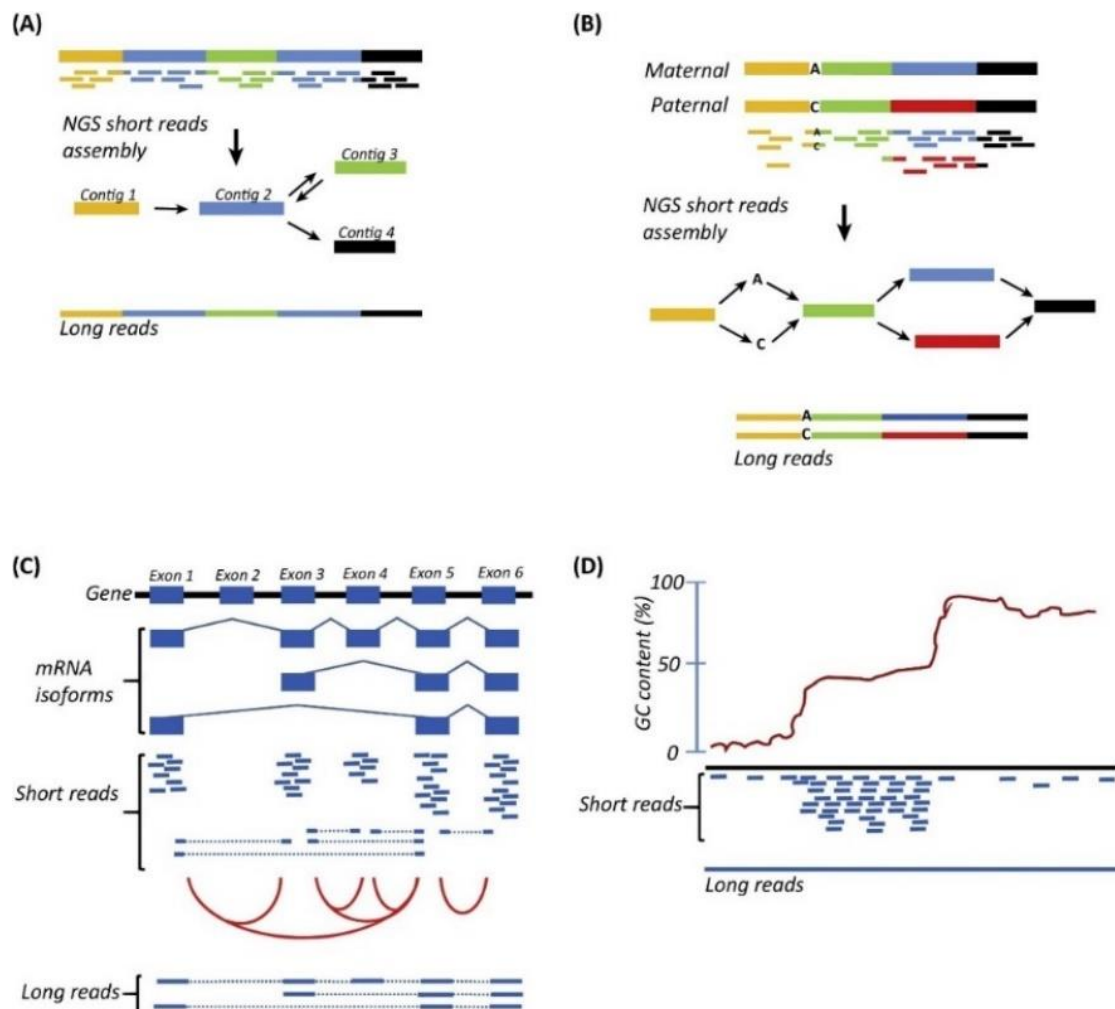


Figure 4. Representation of four instances where long-read sequencing is more advantageous over short-read sequencing (van Dijk et al., 2018).

1.3. The Stacks software

To be able to regenerate the genome sequence and find differences between individuals and populations, software systems are needed. In this case, for RADseq data, the STACKS software is a good option. A study from LaCava et al. compared six different software used in literature to process ddRadSeq data. Those software being: ABySS, CD-HIT, Stacks, Stacks2, Velvet and VSEARCH. In their research they simulated data from the Arabidopsis

thaliana and Homo sapiens genomes and compared *de novo* assemblies for each one of those software, similarly as the data from this study where the sequenced genome of *Sarda sarda* is not available. In their simulations, LaCava et al. simulated different mutation rates and types of mutations, and then used the six assemblers to the simulated datasets, varying the assembly parameters (LaCava et al., 2020).

In order to create data for their simulation they used the *in silico* ddRAD digests of *A. thaliana* and *H. sapiens* genomes., where the size selection was between 350 and 400 bp (LaCava et al., 2020).

The study revealed that ABySS did not successfully recover any authentic genome fragments. Velvet and VSEARCH exhibited poor performance across most simulations. Stacks and Stacks2 generated accurate assemblies for simulations containing SNPs, but their performance declined when insertion and deletion mutations were introduced. In contrast, CD-HIT consistently recovered a significant proportion of genuine genome fragments and emerged as the most reliable assembler in the study (LaCava et al., 2020).

During LaCava et al. literature revision, they found that Stacks is the most commonly used program for assembly, however it is not the most efficient according to their findings. In comparison to CD-HIT, Stacks and Stacks2 were able to recover true genomes well in the absence of allelic variation, however they performed less well than CD-HIT for both the *A. thaliana* and *H. sapiens* genomes when mutations were present (LaCava et al., 2020). Their main issues are; in the case of Stacks indel polymorphisms caused under-assembly of reads and, in the case of Stacks2, failure to recover a substantial fraction of true genome fragments, in other words, incomplete assemblies (Marrano et al., 2020) (LaCava et al., 2020). On the contrary, CD-HIT was the best assembler since it recovered a high proportion of real genome fragments and its assemblies typically were similar to the original genome fragments (LaCava et al., 2020).

Additionally, Velvet and VSEARCH encountered difficulties and failed either to recover all genome fragments or assembling reads into the correct number of genome fragments (LaCava et al., 2020).

Since CD-HIT and Stacks2 are the highest performing assemblers, it is interesting to note that they use a different algorithm, meaning that assembly algorithm is not a reliable criterion for select software for *de novo* assembly of genotyping by sequencing data (LaCava et al., 2020).

In this case the Stacks software was chosen since it is the most commonly used and has proven to be reliable for the type of data available. Also, in this study the platform Galaxy is

employed, since it is an interface that allows to use the Stacks program without having to download it and python programming skills are not needed. Therefore, in Galaxy there is both Stacks2 and Stacks de novo assembly options however, Stacks is the only one that has the parameter selection option, that is needed for the validation that is intended to be done in this project.

Stacks is a software employed to identify and genotype loci in a set of individuals either *de novo* or by comparison to a reference genome using short-read sequence data. During this process Stacks can recover thousands of single nucleotide polymorphism (SNP) markers adjacent to the restriction enzyme site (Catchen et al., 2011).

Stacks, a popular software tool, offers a range of modules for diverse applications. The software comprises several components tailored to different stages of analysis. These modules encompass read preprocessing (process_radtags), merging reads into loci within individuals (ustacks for de novo merging and pstacks for reference-based merging), merging loci across individuals (cstacks), and selecting loci and variants for subsequent analysis (genotypes and populations) among other programs described in figure 5 (Díaz-Arce & Rodríguez-Ezpeleta, 2019). The most interesting programs for this parameter validation analysis are cstacks and ustacks since they are the ones in charge of setting the parameters of interest; m , M and n .

Program	Description	Inputs
process_radtags.pl	Cleans raw Illumina reads, outputs FASTA/FASTQ files.	Raw Illumina reads
ustacks (unique stacks)	Builds loci <i>de novo</i> and detects haplotypes in one individual.	Cleaned FASTA/FASTQ files
cstacks (catalog stacks)	Merges loci from multiple individuals to form a catalog.	ustacks, tab-separated files
sstacks (search stacks)	Matches loci from an individual against a catalog.	ustacks and cstacks, tab-separated files
markers.pl	Calls mappable markers from parental loci.	None
index_radtags.pl	Indexes the database for use by the web interface.	None
denovo_map.pl	Executes ustacks on each individual, builds a catalog with cstacks, and matches individuals against the catalog with sstacks. Calls markers with markers.pl and indexes the database with index_radtags.pl.	Cleaned FASTA/FASTQ files
genotypes.pl	Calls genotypes in a map cross population and outputs markers for use by JoinMap or r/QTL.	None
pstacks (population stacks)	Takes cleaned reads aligned to a reference genome, builds stacks based on the genomic locations of the reads, and detects haplotypes in one individual.	Bowtie or SAM sequence alignments
ref_map.pl	Executes pstacks on each individual, builds a catalog with cstacks, and matches individuals against the catalog with sstacks. Calls markers with markers.pl and indexes the database with index_radtags.pl.	Cleaned FASTA/FASTQ files
sort_read_pairs.pl	Given a set of Stacks data and a set of cleaned, paired-end Illumina reads, outputs one FASTA file for each stack consisting of the paired-end reads associated with reads in that stack.	ustacks output files, cleaned FASTA/FASTQ files
load_sequences.pl	Loads a set of loci-associated sequences (e.g., RNA-seq ESTs) into the database.	FASTA file containing sequences
export_catalog.pl	Exports sequences from the database, including loci and loci-related sequences.	None

Figure 5. Stacks component programs with a brief description and inputs needed (Catchen et al., 2011).

In order to obtain clean sequence data in FASTA or FASTQ output files, needed for Stacks, the `process_radtags.pl` program uses a sliding window to analyse each read quality, if the average quality score within a window drops below 90% confidence, then Stacks discards the read. Stacks can correct some errors such as isolated errors in the restriction cut site sequence or in the barcode (Catchen et al., 2011). Moreover, Stacks uses the maximum likelihood statistical model in order to identify sequence polymorphisms and tell them apart from sequencing errors (Catchen et al., 2011).

In figure 5, `ustacks` is described as the program that builds loci *de novo* and detects haplotypes in each individual. As depicted in figures 6A to 6E, `ustacks` initially utilizes the short-reads to construct stacks that have exact matches. Subsequently, it disassembles the sequence of each stack into k-mers and stores them in a dictionary. This is followed by another disassembly of each stack into k-mers, and the k-mer dictionary is used to generate a list of potentially matching stacks (figure 6B). In the next step, `ustacks` combines the matched stacks to create putative loci and compares secondary reads that were not initially assigned to a stack against these putative loci to increase the stack depth. An SNP model implemented in `ustacks` examines each locus and nucleotide position for polymorphisms (figure 6C and 6D). Finally, `ustacks` generates a consensus sequence and records SNP and haplotype data.(Catchen et al., 2011).

On the other hand, `cstacks` is described in figure 5 as the catalog stacks since it is the one in charge to merge the loci from multiple individuals to form a catalog, in other words, all the reads of each locus are merged into a new consensus sequence catalog containing all the possible loci and alleles of the population, also seen in figure 6F (Catchen et al., 2011).

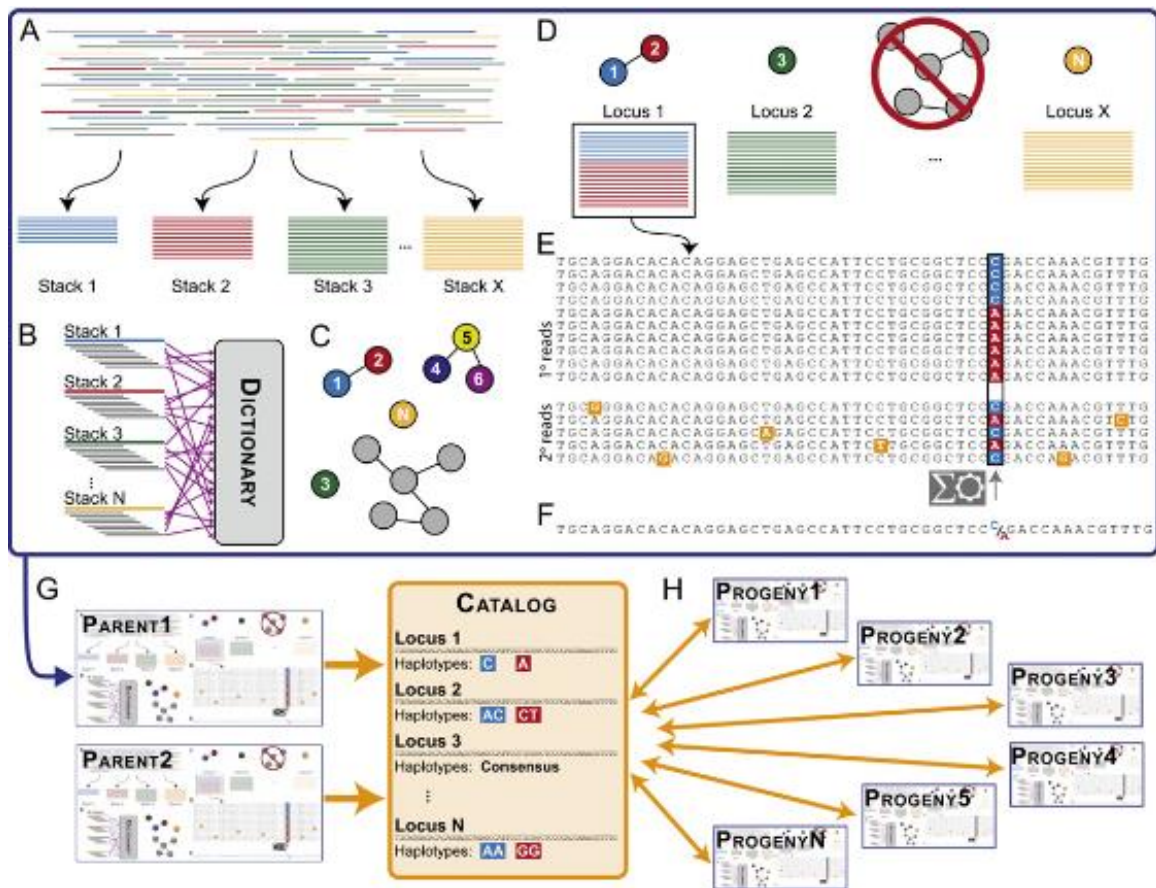


Figure 6. Stacks schematic of every step to obtain the loci and haplotypes of the population (Catchen et al., 2011).

To be more precise, the whole Stacks inner workings are illustrated in figure 6, below is a detailed explanation of this whole process in order to facilitate the understanding of its operation and subsequently of the parameters used and their importance for the assembly and SNP discovery.

For the purpose of identifying loci, the *ustacks* program reads the clean sequences and creates unique, exactly matching stacks. The stacks that do not encompass a threshold (that is configurable through the stack depth parameter) of reads are disassembled as they cannot be distinguished from stacks that contain sequencing errors (figure 6.A). Therefore, the reads in a stack are named primary reads and the ones discarded are called secondary reads (Catchen et al., 2011).

The *ustacks* program initially calculates the average coverage depth and subsequently identifies stacks that exceed two standard deviations from the mean. These high-depth stacks, as well as those located one nucleotide away from them, are excluded from subsequent analysis due to their frequent representation of repetitive elements. (Catchen et al., 2011).

In order to identify polymorphic loci, *ustacks* uses a k-mer search algorithm that measures the distance between stacks. This distance parameter can be adjusted based on the genetic characteristics of the dataset, such as the level of polymorphism and the length of the reads. Typically, only a small number of nucleotide differences are allowed within a defined distance (Catchen et al., 2011).

To perform the analysis, *ustacks* breaks down the sequence of each stack into overlapping fragments of equal length, called k-mers. Each k-mer represents a specific portion of the sequence. For example, the first k-mer spans nucleotides 1 to k, the second k-mer spans nucleotides 2 to k + 1, and so on. The program automatically determines the optimal length for the k-mers based on the allowed nucleotide difference. Longer k-mer lengths result in more specific k-mers that require fewer comparisons with other reads. These k-mers are then stored in a dictionary for further processing. All this process is represented by figure 6.B (Catchen et al., 2011).

In *Stacks*, the k-mer search algorithm compares pairs of stacks based on matching criteria. The merged stacks represent potential loci and are visualized as a graph (figure 6.C). In this graph, each unique stack is a node, and the edges connecting them are weighted based on the nucleotide distances. Each potential locus forms a separate group within the graph that contains all the stacks. (Catchen et al., 2011).

In figure 6.D the *ustacks* program examines the secondary reads that were not initially assigned to a stack and compares them with the putative loci to enhance the stack depth. Within *ustacks*, an SNP model evaluates each locus at every nucleotide position to identify potential polymorphisms. (Catchen et al., 2011).

The merging of stacks in *ustacks* is done in multiple iterations, gradually allowing for a larger nucleotide difference between stacks. Initially, stacks that differ by only one nucleotide are merged, followed by stacks with a two-nucleotide difference, and finally, stacks differing by three nucleotides. This iterative process groups together stacks within the specified distance based on their nucleotide variations. Subsequently, the secondary reads that were previously excluded are compared to the putative loci using the k-mer search algorithm (Catchen et al., 2011)

A slightly larger nucleotide distance is allowed in this comparison, typically two nucleotides more than the within-individual distance parameter. The goal is to find matches between the secondary reads and the defined loci, if the secondary reads that do not have a clear match to a specific locus, then they are definitely discarded. By the end of this process, *Stacks* has successfully created a set of putative loci using the merged high-confidence

stacks. The depth of each locus is further strengthened by incorporating the secondary reads that align with these loci, providing additional support and confidence to the identified loci. This whole process is represented in figure 6.E (Catchen et al., 2011).

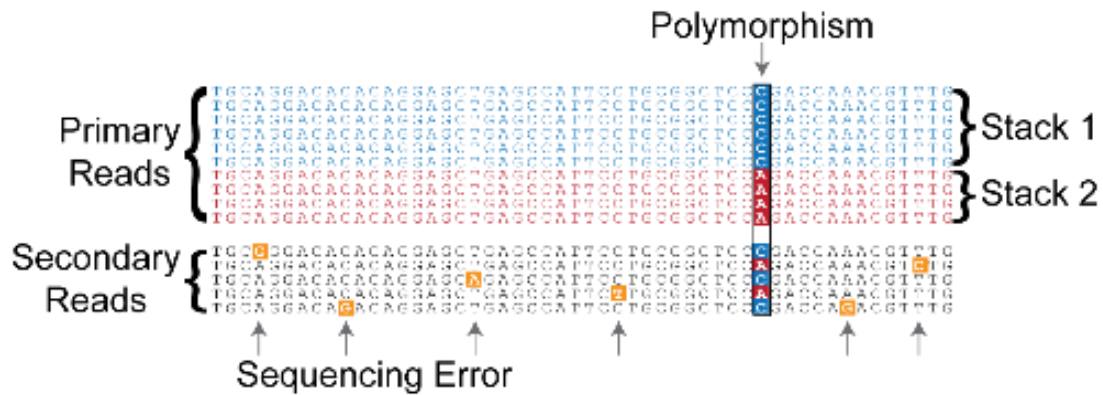


Figure 7. A more detailed representation of figure 6.E (*Stacks: Stacks: Parameter Tutorial*, n.d.).

In the subsequent stage (figure 6. E and F), *ustacks* utilizes a maximum likelihood framework to identify polymorphisms within loci and deduce alleles. It examines each putative locus, systematically evaluating each nucleotide position. By analyzing columns in a two-dimensional matrix of stacked sequencing reads, SNPs are detected. Haplotypes, on the other hand, are determined by examining rows (figure 7). These inferred haplotypes establish alleles for each locus, acting as genetic markers. Lastly, *Stacks* generates a consensus sequence for each locus (Catchen et al., 2011).

Once loci have been constructed for an individual using *Stacks*, the following step involves aggregating these loci into a Catalog that represents the entire population. This aggregation is performed by the *cstacks* program, which reads the output from *ustacks* and merges loci into the Catalog. The process begins with the initialization of the Catalog using the loci from the first individual, and then each additional individual is merged into the Catalog one by one. To match individual loci with those in the Catalog, *cstacks* utilizes the same k-mer search algorithm employed by *ustacks*. However, in this case, each locus in the k-mer dictionary is represented by the set of k-mers derived from each haplotype at that locus (Catchen et al., 2011).

When two loci match, *cstacks* combines their SNPs in the Catalog. In cases where the SNPs have conflicting alleles, indicating variations between the Catalog and the merging locus, the merge fails, and *cstacks* issues a warning. To accommodate the newly merged SNPs, *cstacks* adjusts its haplotype calls. The between-individual distance parameter in *cstacks* allows for mismatches during the merging of loci into the Catalog, thus accommodating variations between individuals in the population (Catchen et al., 2011).

In order to determine the presence of specific locus/haplotype combinations in each individual of the population, the sstacks program, also known as search stacks, compares all individuals, including parents and progeny, against the Catalog. Using each haplotype present in the Catalog, sstacks creates a hash table. It then compares all haplotypes from an individual and identifies any matches (Catchen et al., 2011).

Loci that match multiple loci in the Catalog are excluded, as this would introduce ambiguity in determining the true matching locus. However, it is possible for multiple loci to uniquely match the same Catalog tag. These cases may indicate the presence of repetitive sequences in the progeny that are not found in the parents. This is illustrated in the figure 6. G and H (Catchen et al., 2011).

Although, in this figure the sstacks is used to compare parents and progeny to the Catalog, this program is also used for population genetics, a better illustration for it is figure 8, where every sample of the population is compared to the Catalog.

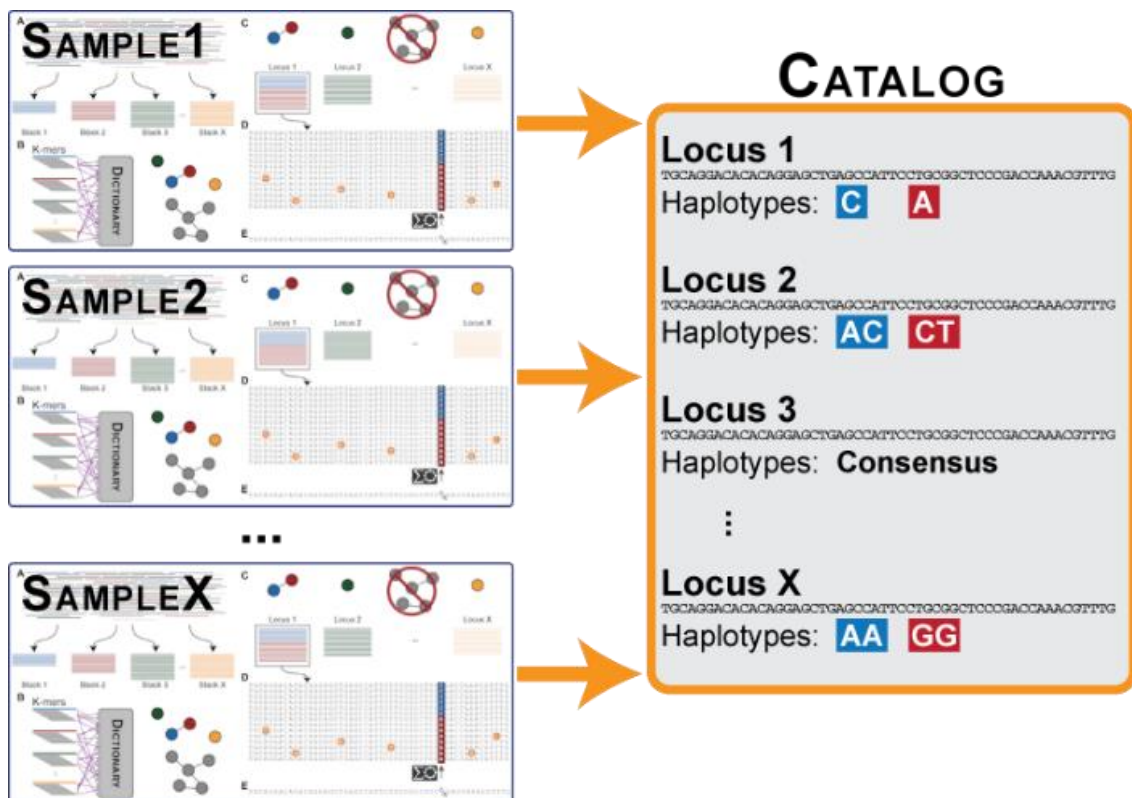


Figure 8. Image representing the comparison of each sample to the Catalog in order to determine specific locus and haplotypes (Stacks: Stacks: Parameter Tutorial, n.d.).

In addition to the data processing steps in Stacks, there are supplementary pipelines that can be utilized. Among them is the clone_filter pipeline, which facilitates the removal of PCR clones. PCR clones are identical sequence fragments that can arise during the amplification

process involved in generating RADseq libraries, particularly when paired-ends are utilized (Díaz-Arce & Rodríguez-Ezpeleta, 2019).

In summary, Stacks is a powerful tool for analysing RAD-seq data, allowing for the identification of numerous informative markers that provide valuable insights into population genetics (Catchen et al., 2011).

1.3.1. The parameters

As explained above, the program Stacks uses several different parameter to achieve the final goal of finding new SNP allowing for the categorising of populations due to differences in the genome of the samples taken. These parameters are in the `denovo_map.pl` pipeline that wraps the programs `ustacks` and `cstacks` (*Stacks: Stacks: Parameter Tutorial*, n.d.).

Minimum stack depth of coverage: m (Figure 6.A)

This parameter in the `ustacks` program determines the threshold for constructing loci in each individual or sample of the population. It regulates the minimum number of matching reads required to form an initial stack. In this process, reads that match exactly and represent alleles, albeit with some errors, are considered primary reads and are stacked together. However, if a stack falls below the m value, the allele is not formed, and the corresponding reads are temporarily set aside as secondary reads (*Stacks: Stacks: Parameter Tutorial*, n.d.).

It is important to note that this number should be adequately selected, if the value is too low then PCR errors would be considered in the stacks and if it is too high then some alleles would be left out. Therefore, the resolution of the sequencing should be taken into account (*Stacks: Stacks: Parameter Tutorial*, n.d.). Also, setting the m parameter too high or too low can lead to either an under-merging or an over-merging of reads, respectively. It is important to find the optimal balance to achieve accurate and informative results (Díaz-Arce & Rodríguez-Ezpeleta, 2019).

Distance allowed between stacks: M (Figure 6.B-C-D)

The second stage of the analysis is done parting from the stacks of alleles (similar with each other) generated previously and are put together into a locus. In this case, the distance allowed between stacks parameter is the number of nucleotides that can differ from two stacks in order to join them. The differences in these nucleotides can be due to polymorphisms or to sequencing errors (*Stacks: Stacks: Parameter Tutorial*, n.d.).

If the parameter is too low then the SNP of a certain locus cannot be identified and it will appear as two different loci. If it is too high then repetitive sequences to chain together into a nonsensical loci (*Stacks: Stacks: Parameter Tutorial*, n.d.).

Then the secondary reads are incorporated once the loci are reconstructed aligning them with a more permissive nucleotide mismatch value. This helps differentiate between polymorphisms and sequencing errors (*Stacks: Stacks: Parameter Tutorial*, n.d.).

Distance allowed between Catalog loci: n (Figure 6.G)

Once each sample of the data set has built the loci, all of this data is merged into a Catalog using the *cstacks* program and it contains all the loci and alleles of the population. (*Stacks: Stacks: Parameter Tutorial*, n.d.)

In the case of setting this parameter too low, some loci from different samples will appear as different loci when in reality they are the same locus and some fixed differences in a population are missed. On the contrary, if the value of the parameter is too high, then similar loci can form a non-sensical locus in the Catalog. It is usually set as the same value as the M parameter. (*Stacks: Stacks: Parameter Tutorial*, n.d.)

To sum up During the process of merging reads into loci within individuals, two important parameters are taken into account. The first parameter, m , known as the minimum required read coverage depth, determines the threshold for creating a stack or a group of identical reads. The second parameter, M , specifies the maximum number of allowable mismatches between stacks or groups of identical reads in order to classify them as different alleles of the same locus. When merging loci between individuals, the primary parameter of consideration is, n , the maximum number of allowed mismatches between loci from different individuals, which determines if they are considered homologs (Díaz-Arce & Rodríguez-Ezpeleta, 2019).

In a recent study utilizing the *Stacks* software, the impact of read filtering, loci assembly, polymorphic site selection, PCR clones, RAD-loci assembly parameters, and SNP selection on marker quantity and genetic differentiation was investigated. The study revealed several notable findings. Firstly, it was discovered that a larger number of polymorphic loci does not necessarily correlate with higher genetic differentiation. Factors such as the presence of PCR duplicates, chosen assembly parameters, and selected SNP filtering criteria were observed to influence both the quantity of recovered polymorphic loci and the level of genetic differentiation. Moreover, the effects of these factors were found to vary among different datasets, suggesting that the adoption of a universal and systematic protocol for RAD-seq data analysis may overlook significant insights into population differentiation. It

is important to note that PCR clones, RAD-loci assembly parameters, and SNP selection based on minor allele frequency (MAF) threshold are influential factors that affect population differentiation inferences using RAD-seq derived SNPs. However, their impact can vary across species, geographic scales, and different group comparisons. Therefore, relying solely on a systematic method for parameter selection may restrict the comprehensive understanding of genetic differentiation across diverse contexts (Díaz-Arce & Rodríguez-Ezpeleta, 2019).

1.4. The Galaxy platform

The galaxy project started in 2005, and its purpose is to offer to scientist of around the world a platform that connects them and allows them to process their data in a simple fashion. It envelops about 8000 analysis software packages that can be used without the need of any programming or Python language skills. From all the programs offered in Galaxy, Stacks is one of them. The Galaxy project is supported by several grants from different countries, highlighting the collective effort to create and maintain this platform (The Galaxy Community, 2022).

Galaxy has several servers across the globe, those are in Europe, United States and Australia and it can be used without the need for a download, set up or fee. Additionally, the Galaxy Training Network has more than 230 tutorials for virtual learning for free (The Galaxy Community, 2022).

In the 2022 update, notable advancements in the Galaxy platform encompass an enhanced user interface tailored for launching extensive analyses involving numerous files. Additionally, interactive tools facilitating exploratory data analysis have been introduced, alongside a comprehensive array of machine learning tools (The Galaxy Community, 2022).

GOALS: parameter validation

Therefore, the goal of this TFG is to validate and optimise the parameters m , M and n of the Stacks software for a small sample size of the Atlantic Bonito in order to recover a maximum of SNPs, thus a maximum of genetic variability. This is important, since in the literature, and as mentioned before, those parameters need to be tailored to each data set because we are genotyping a non-model species that does not have its genome sequenced.

In the long term, the optimized parameters will be used on all the data to produce a better genotyping of the individuals thus, differentiating better the populations. This will allow in the future to assess the management and conservation of the Atlantic Bonito populations so genetic variability is not lost. Additionally, for similar datasets of related species this information could be useful at the moment of parameter setting.

MATERIALS AND METHODS

2.1. Previously: sample collection and processing

In order to do the Stacks assembly, there is previous work done to obtain and process the samples. Therefore, 92 Atlantic Bonitos were gathered in several locations: Tunis, Spain, north of Portugal, south of Portugal, Morocco, Mauritania, Senegal and Côte D'Ivoire. A small muscle tissue or portion of the fin was excised, preserved in 96% ethanol, and shipped to the Laboratori d'Ictiologia Genètica of the Universitat de Girona (Ollé & Viñas, 2022).

For the sample preparation, first, total DNA extraction was performed using DNeasy Blood & Tissue Kit (Qiagen), including an RNase A digestion step. Then, DNA quantity and integrity were assessed using a Qubit dsDNA HS Assay Kit (Life Technologies) and a 1% agarose gel. Subsequently, a total amount of 100ng of DNA per sample was digested with two endonucleases (PstI and MspI) for the RAD-seq library preparation (Ollé & Viñas, 2022).

Secondly, adaptor and padding sequences were added to each fragment and paired-end sequencing of 151bp was realized using Illumina NovaSeq 6000 S4. Once the reads were sequenced, padding sequences and restriction sites were removed following the sequencing service directions with `gbstrim.pl`. After removing the padding sequences and restriction sites, reads with different lengths were trimmed to 80bp using `Process_radtags` in `STACKS2` (Ollé & Viñas, 2022).

The data already demultiplexed was used to validate the parameters on `STACKS`. Three individuals of every location were randomly selected, except for Tunis and Mauritania that were not included for the parameter validation. This reduced sample size permits doing a higher number of in silico experiments, since the more data is used, the longer it takes for `STACKS` to run. The locations selected correspond to the following acronyms: PRT_N (North of Portugal), PRT_S (South of Portugal), ESP (Spain), MOR (Morocco), SEN (Senegal) and CIV (Côte d'Ivoire).

2.2. Tutorials

As mentioned before, Galaxy has a Training Program to help scientist get familiar with the platforms and the thousands of tools there are available. To this end, four tutorials were followed in order to get familiar with Galaxy. Those tutorials being:

- A short introduction to Galaxy (Galaxy Training Materials) (Syme & Soranzo, 2023) accessed on November 17th, 2022.
- Galaxy 101 for everyone (Galaxy Training Materials), (Fouilloux et al., p. 101, 2023) accessed November 16th, 2022.

- Introduction to Genomics and Galaxy (Galaxy Training Materials) (Clements & Gallardo, 2023) accessed November 22nd, 2022.
- RAD-Seq Reference-based data analysis (Galaxy Training Materials) (Bras, 2023) accessed November 28th, 2022.
- RAD-Seq de-novo data analysis (Galaxy Training Materials) (Bras, 2023) accessed on December 12th, 2022.

The data used came from Community-Driven Data Analysis Training for Biology (Batut et al., 2018) for all of the tutorials.

2.3. Using stacks on Galaxy

As mentioned before, Galaxy is a very user-friendly platform that allows scientist to process their data easily. The interface, as seen in figure 9, has three principal segments. To the left is the search tool engine where all of the available programs can be found. In the middle the tool used can be visualized and the settings are available to be changed. To the right is the history component where the uploaded data, all the tools used are and results are stored in order of creation. Finally, at the top there are the work flow manager, the visualizer, the shared data, the help and the user parameters. And at the top right corner is the available storage on the server, each user gets 250 Gb for free 250 Gb.

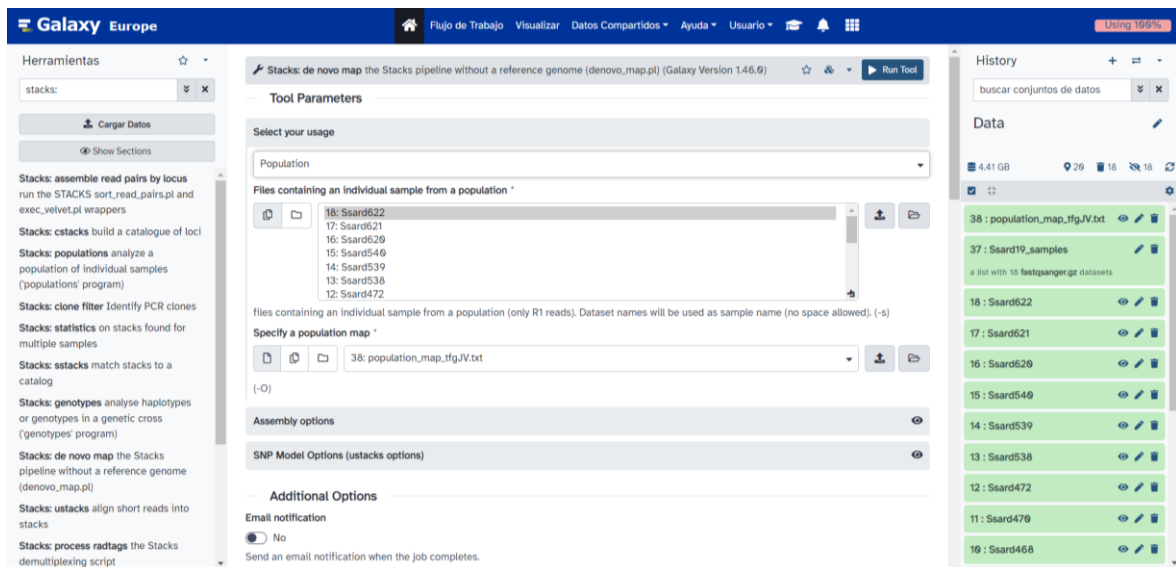
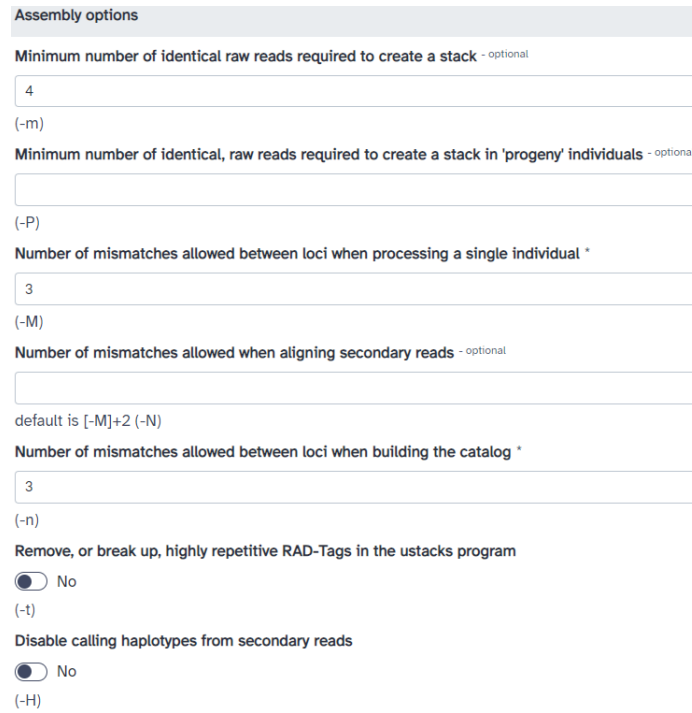


Figure 9. Galaxy's platform main menu where the Stacks: de novo map tool is selected.

Additionally, in figure 9 appears all the Stacks components needed to visualize and process the data for assembly. In this case, Stacks: de novo map, is selected since is the one used.

The usage selected is population, and 18 files corresponding to the sample ddRadseq demultiplexed data is selected. Then the population map is also selected, this file contains

the names of every sample and to which population it corresponds to. Subsequently, the assembly options are selected, in this case only the parameters m , M and n are specified. Finally, the SNP model with a χ^2 of 0.05 significance level to tell apart homozygote from heterozygote is set. This step is repeated for all the combinations of parameters needed.



Assembly options

Minimum number of identical raw reads required to create a stack - optional

4

(-m)

Minimum number of identical, raw reads required to create a stack in 'progeny' individuals - optional

(-P)

Number of mismatches allowed between loci when processing a single individual *

3

(-M)

Number of mismatches allowed when aligning secondary reads - optional

default is [-M]+2 (-N)

Number of mismatches allowed between loci when building the catalog *

3

(-n)

Remove, or break up, highly repetitive RAD-Tags in the ustacks program

No

(-t)

Disable calling haplotypes from secondary reads

No

(-H)

Figure 10. Image of the Galaxy interface for setting the parameters for the Stacks assembly where only the m , M and n are specified. The other two parameters are optional and left in blank so Stacks uses the default setting.

Once STACKS has run, several files are produced however only the `denovo_map_log` (that gives the coverage of each sample) and the summary of summary statistics for each population (where the variant and polymorphic sites are retrieved) are the documents of interest. Once the information from those files, and for each combination of parameters is compiled in to an excel sheet, for the generation of the box plots and grouped columns graphs.

2.4. The parameter setting

Subsequently, the statistical analysis is performed in the coverage and polymorphic sites data. These statistics are performed using the Jamovi software (The jamovi project, 2022) that utilizes the R software (R Core Team, 2021) and packages in order to create the plots and statistical analysis.

The packages used from R are *car* (Companion to Applied Regression) (Fox et al., 2023), *ggplot2* (Create Elegant Data Visualisations Using the Grammar of Graphics) (Wickham et

al., 2023), *ggstatsplot* ('ggplot2' Based Plots with Statistical Details) (Patil & Powell, 2023) and ClinicoPath jamovi Module (Balci, 2022.)

The statistical analysis performed on Jamovi corresponds to a non-parametric one-way ANOVA, also known as Kruskal-Wallis. This test is used to determine if there are any significant differences between mean values of each test. The hypotheses are:

- H_0 : states that there are not differences between groups.
- H_1 : states that there are statistical differences between groups.

When the corresponds to $p_{value} > 0.05$ the H_0 (null hypothesis) is rejected meaning that there are differences between groups.

The parameters that are set for this analysis are in table 1 and they correspond to the values in the middle, that can be set for each parameter as explained by Paris et al., 2017. They explain the main value for each parameter and the range of numbers they can be set at.

Parameter	Main value	Range	Description
m	3	3 to 7	Minimum number of raw reads required to form a stack (Paris et al., 2017).
M	2	1 to 8	Number of mismatches allowed between stacks to merge them into putative locus (Paris et al., 2017).
n	1	=M	Number of mismatches allowed between stacks during construction of the Catalog (Paris et al., 2017).

Table 1. Brief description of each parameter, range and main value for each one of them as described by Paris et al., 2017.

Parameter	Parameter setting for each test										
	Test_1	Test_2	Test_3	Test_4	Test_5	Test_6	Test_7	Test_8	Test_9	Test_10	Test_11
m	3	4	4	4	5	5	5	6	6	6	7
M	5	3	4	5	3	4	5	3	4	5	3
n	5	3	4	5	3	4	5	3	4	5	3

Table 2. All of the tests conducted on Galaxy with the Stacks software have a specific parameter combination that is specified in the table.

Since the main values set in table 1 are the most used in literature, here it was decided to use the values that remain in the middle of their range, as seen in table 2, and see how they perform. Originally there were 9 experiments (from Test_2 to Test_10) and the other two experiments were added later on in order to get a more robust analysis.

Therefore, the Stacks software was run 11 times corresponding to each test, for the same 18 samples of those six locations, and each time the coverage of each sample, the variable and polymorphic sites for the populations were retrieved.

RESULTS

3.1. Polymorphic and variable sites

The variant sites correspond to all of those single nucleotide variations (SNV) that occur in any type of cell and are rare in the population. Single nucleotide polymorphism (SNP) are found in germline DNA and are at least present in more than 1% of the population (*SNP & SNV Genotyping / NGS & Array Techniques*, n.d.) (Solem et al., 2015).

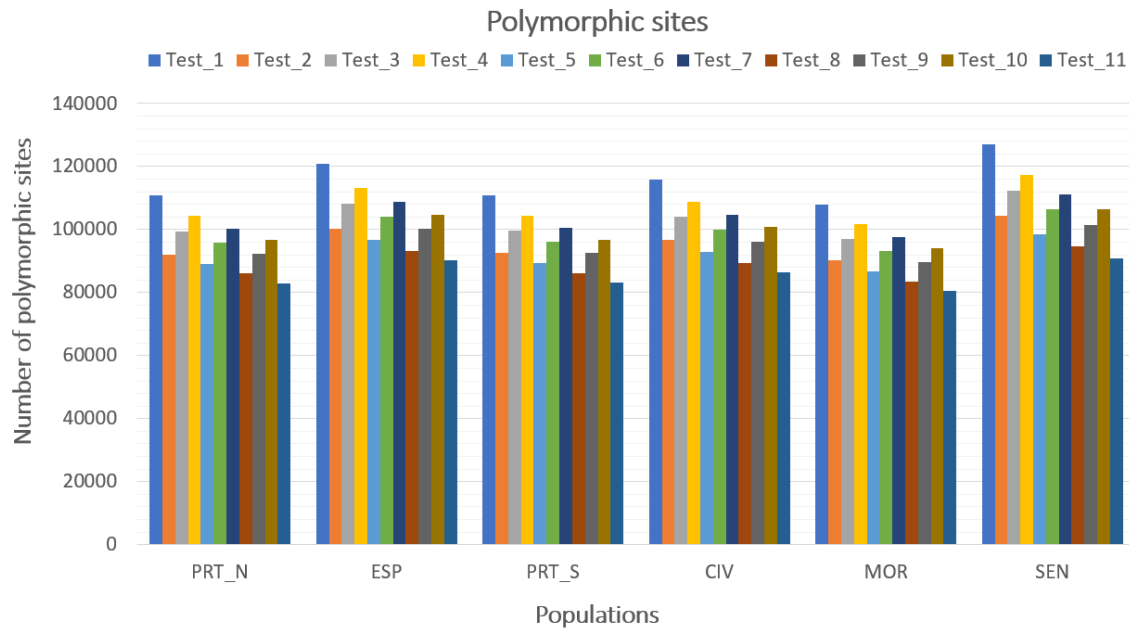


Figure 11. Grouped columns graph where the abscissa axis corresponds to the populations, named by their acronyms (PRT_N=North of Portugal, PRT_S=South of Portugal, ESP=Spain, MOR=Morrocco, SEN=Senegal and CIV=Côte d'Ivoire). For every population there is a column representing each tests that are identified by colours indicated in the legend. The ordinate axis corresponds to the number of polymorphism sites recovered for each test.

Following this definition in the figures 11 and 12 there are the visual representation of the amount of SNP and SNV recovered respectively for each population. It is observed that both graphics follow the same tendency, meaning that both parameters represent the same genetic variability.

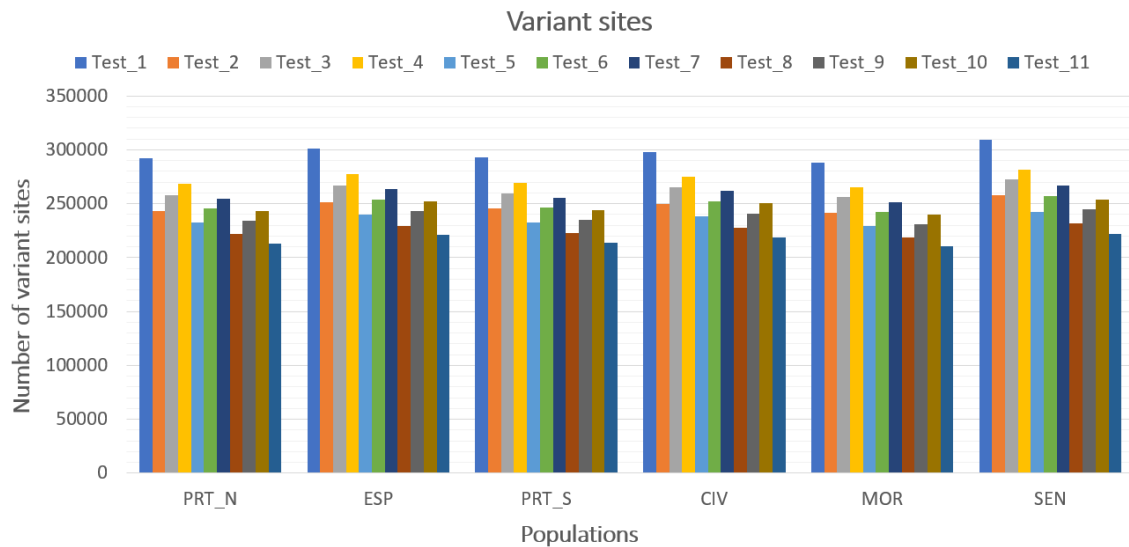


Figure 12. Grouped columns graph where the abscissa axis corresponds to the populations, named by their acronyms (PRT_N=North of Portugal, PRT_S=South of Portugal, ESP=Spain, MOR=Morocco, SEN=Senegal and CIV=Côte d’Ivoire). For every population there is a column representing each tests that are identified by colours indicated in the legend. The ordinate axis corresponds to the number of variant sites recovered for each test.

Taking into account that variant sites are less reliable for this type of analysis since they represent rare variants due to mutations therefore not representing correctly the population (Karki et al., 2015). Moreover, the most used in population genomics are SNPs since they allow us to differentiate between populations, those are the ones selected for the subsequent analysis.

3.2. Coverage

The coverage corresponds to the amount of times a sequence is read, in STACKS the number of raw reads required to form an allele depends on the value of m (Paris et al., 2017). The tendency is that increasing the m value, the coverage depth also increases (Figure E) however the number of polymorphisms recovered decreased (Figure D).

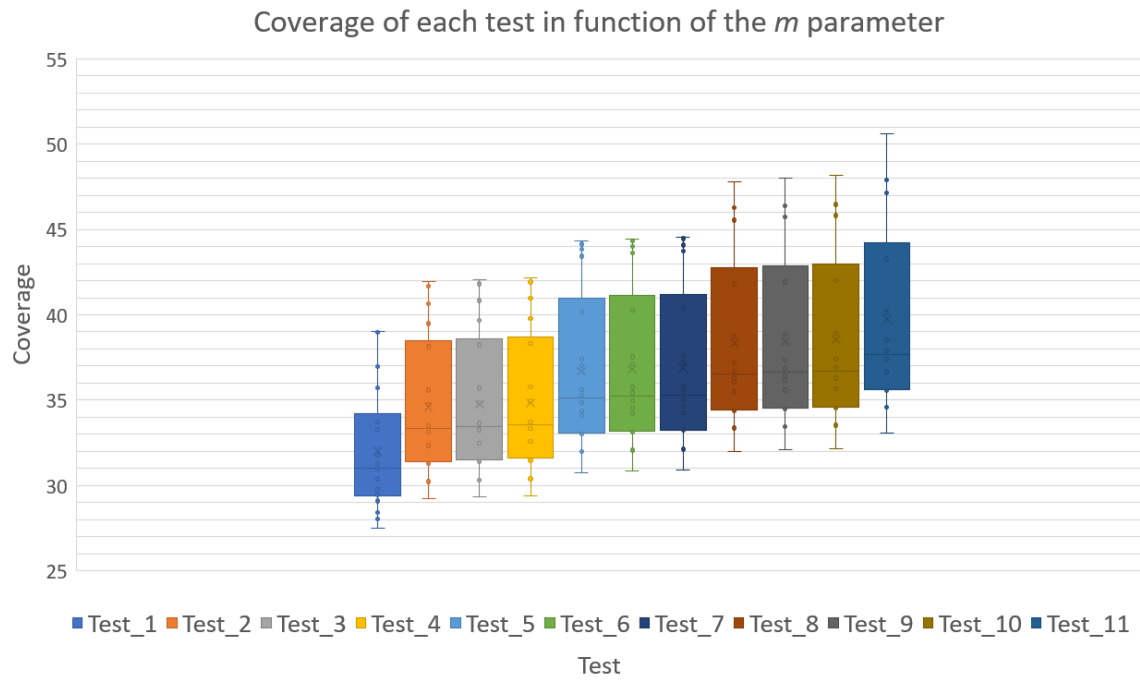


Figure 13. Boxplot graph where the abscissa axis corresponds to the tests differentiated by their colour and are in order of the boxes. The ordinate axis corresponds to the coverage of every test that includes the same 18 samples.

STACKS gives the coverage for each individual in every test however, the polymorphic sites are recovered for each population. Therefore, each test for the coverage has 18 values while (figure 13), each test from the polymorphism sites has 6 values corresponding to the populations analysed (figure 14).

The representation of both box plots of figures 13 and 14 are interpreted the same way. The X represents the mean of every box, the line inside the box represents the median, the minimum and maximum values in the data are represented by the bottom and top fences. The dots along the box represent the data values. The first quartile (Q1) is in between the bottom of the box and the median line, the third quartile (Q3) is in between the median line and the top of the box.

The outliers would be represented by dots outside of the top and bottom fences but in neither of the box plots (figures 13 and 14) are such values.

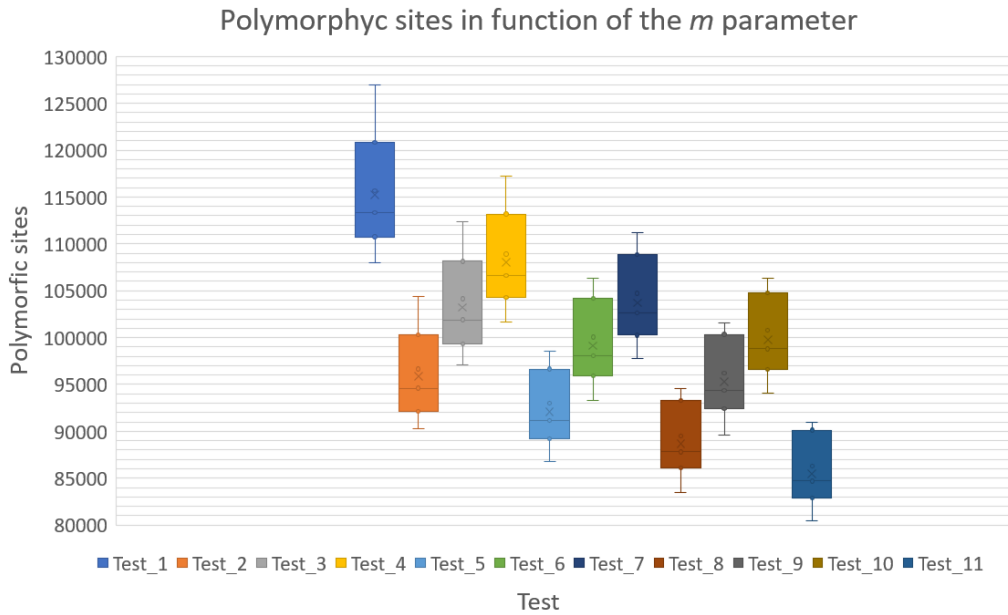


Figure 14. Boxplot graph where the abscissa axis corresponds to the tests differentiated by their colour and are in order of the boxes. The ordinate axis corresponds to the polymorphic sites of every test that includes the same 18 samples.

To corroborate the differences between tests, the Kruskal-Wallis (nonparametric) test was performed since the normality assumption of the one-way ANOVA are not met (Shapiro-wilk test not shown). The results are shown in the figure 15 and 16 for coverage and polymorphisms respectively.

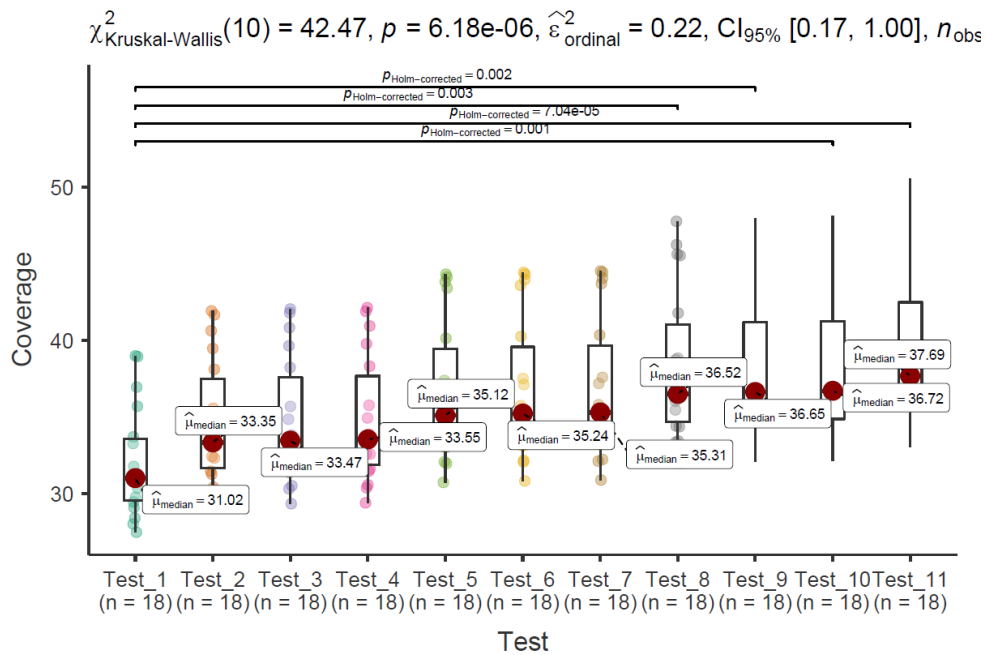


Figure 15. Boxplot graph for the visual representation of the Kruskal-Wallis analysis of the coverage data. All the tests are represented by different boxes and the significant differences are highlighted by the top lines with the corresponding p_{value} .

In the case of the coverage data, the Kruskal-Wallis test has a $p_{\text{value}} > 0.001$ ($\chi^2 = 42.5$, $df = 10$). The H_0 is rejected, meaning that there are differences between tests, those can be observed in the figure 14, where the statistically different test correspond to the Test_1 in front of the Test_8 ($p_{\text{Holm-corrected}} = 0.003$), Test_9 ($p_{\text{Holm-corrected}} = 0.002$), Test_10 ($p_{\text{Holm-corrected}} = 0.001$) and Test_11 ($p_{\text{Holm-corrected}} = 0.0000704$).

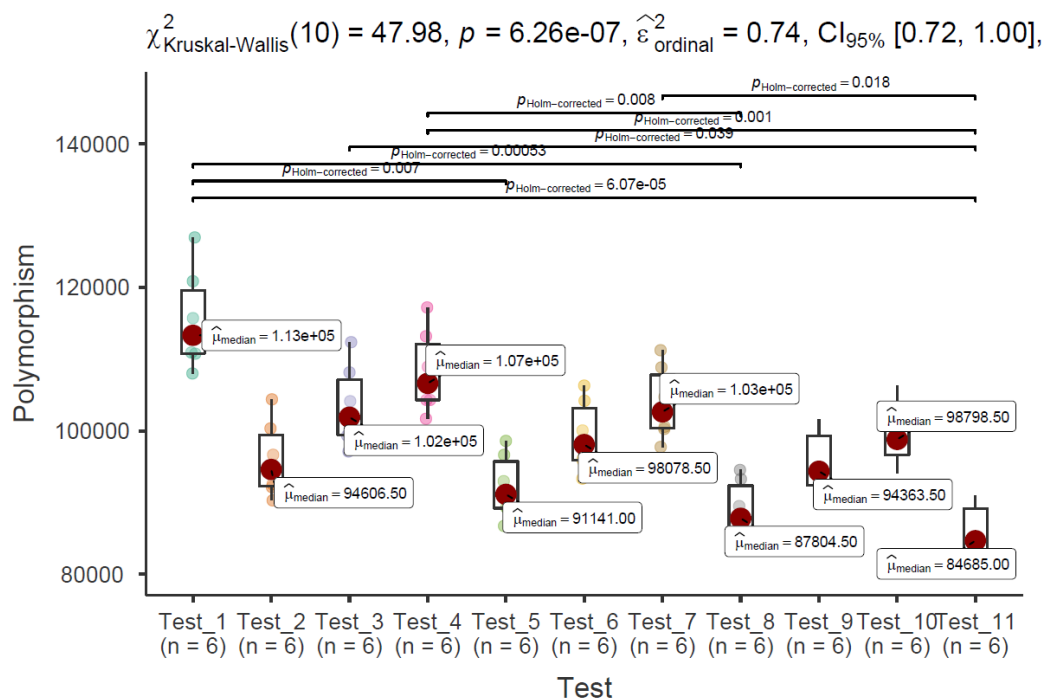


Figure 16. Boxplot graph for the visual representation of the Kruskal-Wallis analysis of the polymorphism site data. All the tests are represented by different boxes and the significant differences are highlighted by the top lines with the corresponding p_{value} .

In the case of the polymorphic site data, the Kruskal-Wallis test has a $p_{\text{value}} > 0.001$ ($\chi^2 = 48$, $df = 10$). The H_0 is rejected, meaning that there are differences between tests, those can be observed in the figure 15. Where the significant results are: Test_1 vs Test_5 ($p_{\text{Holm-corrected}} = 0.007$), Test_1 vs Test_8 ($p_{\text{Holm-corrected}} = 0.00053$), Test_1 vs Test_11 ($p_{\text{Holm-corrected}} = 0.0000607$), Test_3 vs Test_11 ($p_{\text{Holm-corrected}} = 0.039$), Test_4 vs Test_8 ($p_{\text{Holm-corrected}} = 0.008$), Test_4 vs Test_11 ($p_{\text{Holm-corrected}} = 0.001$) and Test_7 vs Test_11 ($p_{\text{Holm-corrected}} = 0.018$).

DISCUSSION

The polymorphic sites are the most commonly used to differentiate between populations in genotyping analysis. Thus, even if in the figures 11 and 12 demonstrate the same tendency in site recovery, which is to be expected, the subsequent analyses are done with the polymorphic site values.

Additionally, it is likely to call true SNPs with a high coverage (Paris et al., 2017). In figure 13 is represented the coverage for each combination of parameters and it is very evident that the parameter m is the one that influences the coverage, that is because it directly instructs the `ustacks` program the minimum number of reads to create a stack, thus directly inferring on the depth of the stacks.

In the literature it is not uncommon to use coverages below 10x (Paris et al., 2017) however, another study used $m=5$ and the mean depth coverage was 50x for RADseq data in Atlantic Mackerel (Rodríguez-Ezpeleta et al., 2016). In the case of this project analysis, the mean of all the tests done retrieved a coverage higher than 30x meaning that the coverage is already high to begin with as demonstrated in figure 13.

Since the lower setting of $m=3$ has a mean depth coverage of 32.0x, and the higher setting of $m=7$ has a mean depth coverage of 39.7x, it is interesting to see if there are any significant differences between any of tests ($m=3$, $m=4$, $m=5$, $m=6$ and $m=7$). In order to do that, a Kruskal-Wallis (nonparametric) test was performed (figure 15) and statistical differences are found between Test_1 ($m=3$) and Test_8, Test_9, Test_10, Test_11 ($m=6$, $m=6$, $m=6$, $m=7$ respectively) thus, meaning that the coverage is statically higher when $m=6$ or 7 that when it is set at $m=3$. Although, as mentioned before, all tests have a high coverage.

If $m=1$, then all raw reads are considerate putative alleles and the secondary alleles disappear (Paris et al., 2017). Setting the m parameter too low augments the likelihood of erroneously labelling as stacks reads with convergent sequencing errors. However, if it is too high then the true alleles that are lacking coverage can be dropped, thus recognizing them as homozygous when in reality there are other alleles (Mastretta-Yanes et al., 2015). Therefore, m is useful at the moment of distinguishing real loci from sequencing errors but can also miscall as homozygous an heterozygous allele that is underrepresented (Paris et al., 2017).

In the figure 14 it is visible how the number of SNP drops as the m parameter increases thus, corresponding to the expected behaviour. Although, the polymorphic sites recovered also depend on the n and M parameters, since M is the number of mismatches allowed between stacks to create a putative locus and n is the number of mismatches allowed between locus

during the construction of the Catalog. If n is set to 0 then the secondary reads are discarded, this can be useful in the case of extremely low levels of polymorphisms where a high to moderate coverage is needed (Paris et al., 2017).

Paris et al. recommend setting the m parameter smaller than 5 due to the fact that for their datasets, $m=3$ was favourable for all of them (Paris et al., 2017). Furthermore, our data also coincides that $m=3$ already has a high mean depth coverage and it only statistically significantly improves if it is changed to 6 or 7 and, as it has already been discussed, these larger values can cause the drop of heterozygous alleles.

On the other hand, the polymorphisms enable us to tell apart populations that are less genetically differentiated. Also, it allows to identify the homozygotic and heterozygotic individuals and which variations are present in each population. The higher the polymorphism sites the more resolution for the population discrimination.

In this case, figure 14 demonstrates a tendency of the number of polymorphic sites to descent as the m increases however, for each set of tests with the same m setting ($m=4$ in Test_2, Test_3 and Test_4, $m=5$ in Test_5, Test_6 and Test_7, $m=6$ in Test_8, Test_9 and Test_10) the polymorphism sites increase at the same time as the M and n parameters.

To assess the differences between groups, a Kruskal-Wallis analysis is performed and reveals several significant differences (figure 16). The most different tests are Test_1 and Test_11 ($p_{\text{Holm-corrected}}=0.0000607$) as expected since they have the most extreme parameter settings. The other differences are; Test_1 vs Test_8 ($p_{\text{Holm-corrected}}=0.00053$), Test_4 vs Test_8 ($p_{\text{Holm-corrected}}=0.008$), Test_4 vs Test_11 ($p_{\text{Holm-corrected}}=0.001$), Test_7 vs Test_11 ($p_{\text{Holm-corrected}}=0.018$), Test_1 vs Test_5 ($p_{\text{Holm-corrected}}=0.007$) and Test_3 vs Test_11 ($p_{\text{Holm-corrected}}=0.039$) in the order of most to least statistically distinct.

The M parameters sets the number of mismatches allowed between stacks to merge them into a locus when processing an individual therefore, species that present high levels of polymorphisms need higher levels of M . In the case of setting the M too low, the alleles of a same loci will be interpreted as different loci and will not be merged, on the contrary if the M parameter is set too high, then repetitive sequences or paralogous loci will be merged together creating a large erroneous loci (Paris et al., 2017) (Mastretta-Yanes et al., 2015). Additionally, it is discussed in other studies that for larger sequence length (250bp), the higher the M parameter should be, in this case the length after the processing of the RAD sequences left them at 80bp (Paris et al., 2017).

The n parameter controls the mismatches of the same stacks allowed between individuals when constructing the Catalog. Thus, it is logical to set $n=M$ giving that M sets the number

of mismatches allowed between stacks in a same individual. Therefore, if it is expected to have only homozygous individuals in the population $n=M$ makes sense considering that the alleles in those cases would be fixed however, if heterozygotes are expected and fixed alleles will be rare, then $n=M-1$ would be better (Paris et al., 2017). Additionally, setting $n=0$ implies that the same loci for different individuals would be represented as different loci, finally, using a too high n value would create erroneous loci (Mastretta-Yanes et al., 2015).

In the data used here, seeing that the populations are constituted of only three individuals, it very provable to have homozygous individuals so the $n=M$ is logical, also most n values used are high ($n=4,5,6$) compared to the usual setting of $n=1$ or 2 (Paris et al., 2017).

CONCLUSIONS

With all of this information, and taking in to account that the coverage in all of the test is high enough to for the data used, the parameter setting that is most interesting for this study would be the one that recovers the highest number of polymorphisms without over estimating or creating erroneous loci. Because of this, the tests with the m parameter set too high ($m=6$ and $m=7$) are discarded since heterozygous loci could be dropped and misrepresented. From this, 7 tests remain and, from a polymorphic site pint of view, the only significantly different are Test_1 ($m=3, M=5, n=5$) and Test_5 ($m=5, M=3, n=3$).

The best one between the two tests selected would depend on the expectations of the analysis, in the case of having a less differentiated set of populations Test_5 should work fine since the M and n parameters allow for an efficient merging of the loci for the same individual and in the Catalog (it is expected a high number of fixed loci taking into account that there are only three individuals per population).

However, in the case of a larger population, as would be in the case of the actual study of population differentiation that is intended to be executed with the 92 samples collected, Test_1 that provides a higher M and n could be more useful because it is expected to have a higher amount of heterozygotic individuals.

Therefore, the best parameter setting would be between Test_1 and Test_5. Setting $m>5$ is not recommended for this type of data since it does not seem to provide any significant advantage in the coverage department thus, m between 3 and 5 would be ideal. On the other hand, setting M and n between 3 and 5 could help recover more polymorphic sites.

The overall conclusion is that the parameter setting influences greatly on the coverage and polymorphic site recovery and it is important to do a series of test before exploring a big data set. For species close to the Atlantic Bonito studied here, this analysis could be useful at the moment to decide which values to set these parameters. In this case, not only one parameter combination is valid and the decision making should take in to count the type of data provided, meaning that the number of individuals per population, the length of the sequences analysed and the intentions of the analysis are influential considerations to take in to account during the decision making.

BIBLIOGRAFIA:

- 2_1_10_1_BON_ENG.pdf. (n.d.). Retrieved 24 June 2023, from https://www.iccat.int/Documents/SCRS/Manual/CH2/2_1_10_1_BON_ENG.pdf
- Analysis for Clinicopathological Research*. (n.d.). Retrieved 3 July 2023, from <https://www.serdarbalci.com/ClinicoPathJamoviModule/>
- Badano, A. (2021). In silico imaging clinical trials: Cheaper, faster, better, safer, and more scalable. *Trials*, 22(1), 64. <https://doi.org/10.1186/s13063-020-05002-w>
- Batut, B., Hiltmann, S., Bagnacani, A., Baker, D., Bhardwaj, V., Blank, C., Bretaudeau, A., Brillet-Guéguen, L., Čech, M., Chilton, J., Clements, D., Doppelt-Azeroual, O., Erxleben, A., Freeberg, M. A., Gladman, S., Hoogstrate, Y., Hotz, H.-R., Houwaart, T., Jagtap, P., ... Grüning, B. (2018). Community-Driven Data Analysis Training for Biology. *Cell Systems*, 6(6), 752-758.e1. <https://doi.org/10.1016/j.cels.2018.05.012>
- Boekhout, H., van der Weijden, I., & Waltman, L. (2021). *Gender differences in scientific careers: A large-scale bibliometric analysis* (arXiv:2106.12624). arXiv. <https://doi.org/10.48550/arXiv.2106.12624>
- Bras, Y. L. (a, 28:11). *RAD-Seq de-novo data analysis* [Text]. Galaxy Training Network; Galaxy Training Network. <https://training.galaxyproject.org/training-material/topics/ecology/tutorials/de-novo-rad-seq/tutorial.html>
- Bras, Y. L. (b, 28:11). *RAD-Seq Reference-based data analysis* [Text]. Galaxy Training Network; Galaxy Training Network. <https://training.galaxyproject.org/training-material/topics/ecology/tutorials/ref-based-rad-seq/tutorial.html>
- Catchen, J. M., Amores, A., Hohenlohe, P., Cresko, W., & Postlethwait, J. H. (2011). Stacks: Building and genotyping Loci de novo from short-read sequences. *G3 (Bethesda, Md.)*, 1(3), 171–182. <https://doi.org/10.1534/g3.111.000240>
- Clements, D., & Gallardo, C. (, 01:04). *Introduction to Genomics and Galaxy* [Text]. Galaxy Training Network; Galaxy Training Network. <https://training.galaxyproject.org/training-material/topics/introduction/tutorials/galaxy-intro-strands/tutorial.html>
- Davey, J. W., & Blaxter, M. L. (2010). RADSeq: Next-generation population genetics. *Briefings in Functional Genomics*, 9(5–6), 416–423. <https://doi.org/10.1093/bfgp/elq031>
- Díaz-Arce, N., & Rodríguez-Ezpeleta, N. (2019). Selecting RAD-Seq Data Analysis Parameters for Population Genetics: The More the Better? *Frontiers in Genetics*, 10, 533. <https://doi.org/10.3389/fgene.2019.00533>
- Dodsworth, S., Pokorny, L., Johnson, M. G., Kim, J. T., Maurin, O., Wickett, N. J., Forest, F., & Baker, W. J. (2019). Hyb-Seq for Flowering Plant Systematics. *Trends in Plant Science*, 24(10), 887–891. <https://doi.org/10.1016/j.tplants.2019.07.011>
- Fouilloux, A., Goué, N., Barnett, C., Maroni, M., Nahorna, O., Clements, D., & Hiltmann, S. (, 34:29). *Galaxy 101 for everyone* [Text]. Galaxy Training Network; Galaxy Training Network. <https://training.galaxyproject.org/training-material/topics/introduction/tutorials/galaxy-intro-101-everyone/tutorial.html>
- Fox, J., Weisberg, S., Price, B., Adler, D., Bates, D., Baud-Bovy, G., Bolker, B., Ellison, S., Firth, D., Friendly, M., Gorjanc, G., Graves, S., Heiberger, R., Krivitsky, P., Laboissiere, R., Maechler, M., Monette, G., Murdoch, D., Nilsson, H., ... R-Core. (2023). *car: Companion to Applied Regression* (3.1-2). <https://cran.r-project.org/web/packages/car/index.html>
- Green, E. D., Watson, J. D., & Collins, F. S. (2015). Twenty-five years of big biology. *Nature*, 526(7571), 29–31. <https://doi.org/10.1038/526029a>
- Hu, T., Chitnis, N., Monos, D., & Dinh, A. (2021). Next-generation sequencing technologies: An overview. *Human Immunology*, 82(11), 801–811. <https://doi.org/10.1016/j.humimm.2021.02.012>
- ICCAT·CICTA·CICAA. (n.d.). Retrieved 24 June 2023, from <https://www.iccat.int/es/>

- Inbar, S., Cohen, P., Yahav, T., & Privman, E. (2020). Comparative study of population genomic approaches for mapping colony-level traits. *PLoS Computational Biology*, 16(3), e1007653. <https://doi.org/10.1371/journal.pcbi.1007653>
- jamovi—Open statistical software for the desktop and cloud. (n.d.). Retrieved 23 June 2023, from <https://www.jamovi.org/>
- Karki, R., Pandya, D., Elston, R. C., & Ferlini, C. (2015). Defining “mutation” and “polymorphism” in the era of personal genomics. *BMC Medical Genomics*, 8, 37. <https://doi.org/10.1186/s12920-015-0115-z>
- Kumar, K. R., Cowley, M. J., & Davis, R. L. (2019). Next-Generation Sequencing and Emerging Technologies. *Seminars in Thrombosis and Hemostasis*, 45(7), 661–673. <https://doi.org/10.1055/s-0039-1688446>
- LaCava, M. E. F., Aikens, E. O., Megna, L. C., Randolph, G., Hubbard, C., & Buerkle, C. A. (2020). Accuracy of *de novo* assembly of DNA sequences from double-digest libraries varies substantially among software. *Molecular Ecology Resources*, 20(2), 360–370. <https://doi.org/10.1111/1755-0998.13108>
- Magbanua, Z. V., Hsu, C.-Y., Pechanova, O., Arick, M., Grover, C. E., & Peterson, D. G. (2023). Innovations in double digest restriction-site associated DNA sequencing (ddRAD-Seq) method for more efficient SNP identification. *Analytical Biochemistry*, 662, 115001. <https://doi.org/10.1016/j.ab.2022.115001>
- Marrano, A., Palmer, A. E., & Moyers, B. T. (2020). Stacking up RADSeq assembly programs: From complete hit to completely abysmal. *Molecular Ecology Resources*, 20(2), 357–359. <https://doi.org/10.1111/1755-0998.13140>
- Mather, J. A. (2019). Ethics and Care: For Animals, Not Just Mammals. *Animals: An Open Access Journal from MDPI*, 9(12), 1018. <https://doi.org/10.3390/ani9121018>
- Maxam, A. M., & Gilbert, W. (1977). A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 74(2), 560–564. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC392330/>
- McCombie, W. R., McPherson, J. D., & Mardis, E. R. (2019). Next-Generation Sequencing Technologies. *Cold Spring Harbor Perspectives in Medicine*, 9(11), a036798. <https://doi.org/10.1101/cshperspect.a036798>
- Miller, M. R., Dunham, J. P., Amores, A., Cresko, W. A., & Johnson, E. A. (2007). Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research*, 17(2), 240–248. <https://doi.org/10.1101/gr.5681207>
- Ollé, J., & Viñas, J. (2022). Atlantic Bonito (*Sarda sarda*) Genomic ddRadSeq Analysis Confirms Population Differentiation across Northeast Atlantic and Mediterranean Locations—Implications for Fishery Management. *Biology and Life Sciences Forum*, 13(1), Article 1. <https://doi.org/10.3390/blsf2022013023>
- Paris, J. R., Stevens, J. R., & Catchen, J. M. (2017). Lost in parameter space: A road map for stacks. *Methods in Ecology and Evolution*, 8(10), 1360–1373. <https://doi.org/10.1111/2041-210X.12775>
- Patil (@patilindrajeets), I., & Powell, C. (2023). *ggstatsplot: 'ggplot2' Based Plots with Statistical Details* (0.11.1). <https://cran.r-project.org/web/packages/ggstatsplot/index.html>
- Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: An inexpensive method for *de novo* SNP discovery and genotyping in model and non-model species. *PLoS One*, 7(5), e37135. <https://doi.org/10.1371/journal.pone.0037135>
- RAD-Seq Genotypes Less, But Offers More*. (2011).
- Rodríguez-Ezpeleta, N., Bradbury, I. R., Mendibil, I., Álvarez, P., Cotano, U., & Irigoien, X. (2016). Population structure of Atlantic mackerel inferred from RAD-seq-derived SNP markers: Effects of sequence clustering parameters and hierarchical SNP selection. *Molecular Ecology Resources*, 16(4), 991–1001. <https://doi.org/10.1111/1755-0998.12518>

- Sanger, F., & Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, 94(3), 441–448. [https://doi.org/10.1016/0022-2836\(75\)90213-2](https://doi.org/10.1016/0022-2836(75)90213-2)
- Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12), 5463–5467. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC431765/>
- Shen, C.-H. (2019a). *Diagnostic Molecular Biology*. Academic Press.
- Shen, C.-H. (2019b). Genome and Transcriptome Analysis. In *Diagnostic Molecular Biology* (pp. 303–329). Elsevier. <https://doi.org/10.1016/B978-0-12-802823-0.00012-2>
- Sneddon, L. U. (2015). Pain in aquatic animals. *The Journal of Experimental Biology*, 218(Pt 7), 967–976. <https://doi.org/10.1242/jeb.088823>
- SNP & SNV Genotyping | NGS & array techniques*. (n.d.). Retrieved 27 June 2023, from <https://www.illumina.com/techniques/popular-applications/genotyping/snp-snv-genotyping.html>
- Solem, A. C., Halvorsen, M., Ramos, S. B. V., & Laederach, A. (2015). The potential of the riboSNitch in personalized medicine. *Wiley Interdisciplinary Reviews. RNA*, 6(5), 517–532. <https://doi.org/10.1002/wrna.1291>
- Stacks: Stacks: Parameter Tutorial*. (n.d.). Retrieved 18 January 2023, from https://catchenlab.life.illinois.edu/stacks/param_tut.php
- Syme, A., & Soranzo, N. (2013). *A short introduction to Galaxy* [Text]. Galaxy Training Network; Galaxy Training Network. <https://training.galaxyproject.org/training-material/topics/introduction/tutorials/galaxy-intro-short/tutorial.html>
- The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. (2022). *Nucleic Acids Research*, 50(W1), W345–W351. <https://doi.org/10.1093/nar/gkac247>
- van Dijk, E. L., Jaszczyszyn, Y., Naquin, D., & Thermes, C. (2018). The Third Revolution in Sequencing Technology. *Trends in Genetics: TIG*, 34(9), 666–681. <https://doi.org/10.1016/j.tig.2018.05.008>
- Viñas, J., Gordo, A., Fernández-Cebrián, R., Pla, C., Vahdet, Ü., & Araguas, R. M. (2011). Facts and uncertainties about the genetic population structure of Atlantic bluefin tuna (*Thunnus thynnus*) in the Mediterranean. Implications for fishery management. *Reviews in Fish Biology and Fisheries*, 21(3), 527–541. <https://doi.org/10.1007/s11160-010-9174-6>
- Waaijer, C. J. F., Sonneveld, H., Buitendijk, S. E., van Bochove, C. A., & van der Weijden, I. C. M. (2016). The Role of Gender in the Employment, Career Perception and Research Performance of Recent PhD Graduates from Dutch Universities. *PloS One*, 11(10), e0164784. <https://doi.org/10.1371/journal.pone.0164784>
- Watson, J. D., & Crick, F. H. C. (1953). The Structure of Dna. *Cold Spring Harbor Symposia on Quantitative Biology*, 18, 123–131. <https://doi.org/10.1101/SQB.1953.018.01.020>
- Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., Dunnington, D., Posit, & PBC. (2023). *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics* (3.4.2). <https://cran.r-project.org/web/packages/ggplot2/index.html>