Can we harness disease resistance by association directly in the wild sea beet?

Hélène Yvanne

A thesis submitted for the degree of Doctor of Philosophy (PhD) to the University of East Anglia

Earlham Institute

Norwich Research Park, Norwich, NR4 7UZ

March 2024

Abstract

Sugar beet (*Beta vulgaris vulgaris*), which contributes for 20% of the worldwide sugar production, is one of the most recently domesticated crops. It was domesticated from the wild sea beet, *Beta vulgaris maritima*, still growing on European coasts. The beet system is a model for maintaining a sustainable crop production in a scenario of human population growth and climate change. Crop wild relatives are now considered a source of traits to improve their domesticated counterparts, especially regarding disease resistance.

Construction of a sea beet pan-genome consisting of eleven sea beets from England, Denmark, France and Spain marks an improvement in both contiguity and completeness compared to published data.

These pan-genomic data, along with the whole-genome re-sequencing of hundreds of wild sea beets sampled mostly in England, facilitated a *k*-mer-based association study for beet rust resistance. Five candidate NLR loci are identified, among which one locus appears in distinct controlled inoculation trials involving different English rusts, as well as in naturally inoculated wild sampled material. This opens the door to the potential success of association genetic studies conducted on wild individuals in controlled conditions or even directly in their natural habitat.

Sea beet population structure is investigated on east (rich in sugar beet cultivation) and west coasts of England and highlights a higher genetic diversity in resistance genes than in non-resistance genes, particularly on the east coast. This suggests a potential impact of the presence of crops or their pathogens in this area. Moreover, the sea currents from Humberside towards East-Anglia could explain reduced northerly gene flow. Finally, measures of nucleotide diversity and differentiation at the five candidate NLR loci indicate that population genetic measures could be used to inform on the efficacy of candidate resistance genes.

Acknowledgments

This work was supported by the UKRI Biotechnology and Biological Sciences Research Council Norwich Research Park Biosciences Doctoral Training Partnership BB/M011216/1 as a CASE Award in collaboration with KWS SAAT SE & Co. KGaA.

I want to thank Prof. Juliane Dohm and Dr. Wilfried Haerty for reviewing this thesis.

I would like to thank the KWS company to have collaborated with this PhD project, particularly Dr. Knuth Weißleder, Jens Christoph Lein and Dr. Dieter Hackenberg. Thank you very much Jens and Dieter for your warm welcome in Germany.

I would like to acknowledge my supervisory team: Prof. Neil Hall, Dr. Mark M^{c}Mullan, Prof. Brande Wulff and Dr. Jose de Vega, for the guidance I have received throughout the PhD and for their wise advice. I would like to thank Brande Wulff for the resource for pan-genomics. I want to particularly thank Mark for his trust in me, and his enormous support, both scientifically and humanly.

I would like to thank the team I had the pleasure to work with at the Earlham Institute: Prof. Neil Hall, Dr. Mark M^{c}Mullan, Michelle Grey, Dr. Ross Low, Mumdooh Sabir, Dr. Matt Bawn, Dr. Sally Warring, Dr. Rowena Hill, Laura Bird, Santa Walker and Angel Chandwani. I would like to especially thank you for your patience and your great help in improving my English!

I want to thank Michelle Grey for her great help with DNA extractions and for the fun we had with the large-scale glasshouse trials! Additionally, I would like to thank Dominic Rogers, Mumdooh Sabir and Alessandro Colosio for their help with plant phenotyping.

I would like to thank the British Beet Research Organisation (BBRO), and particularly Kevin Sawford for helping with rust collection. Also, I would like to thank Dr. Thies Marten Heick and Prof. Isabelle De Cauwer for the sharing of mainland European sea beet seeds.

I want to thank the JIC Horticultural Services for their help with conducting the large-scale rust inoculation trials, especially Lewis Hollingsworth, Lionel Perkins and Timothy Glister.

I would like to thank my friends for the great time we spent in Norwich, around beers and board games, and for the support they provided me during this period of my life, especially Anne Jordan, Dr. Wouter van Bakel and Dr. Sathish Thiyagarajan. I want to thank all the persons I have not cited but who helped me to complete this project.

Finally, and very importantly, I want to thank my family for always being there for me and supporting every stage of my life, Cécile Manouvrier and Marion Yvanne.

Table of contents

Table of figures

Table of tables

Chapter I - General introduction

I - A. Introduction to the beet system and its cultivation

I - A. 1. Beet domestication

Sugar beet belongs to the *Beta* genus, the Amaranthaceae family and the Caryophyllales order. It is one of the three *Beta vulgaris* subspecies: *Beta vulgaris ssp. vulgaris* (sugar beet and all cultivated beets), *Beta vulgaris ssp. maritima* (sea beet), and *Beta vulgaris ssp. adanensis*. Among these subspecies, sea beet is the progenitor of all domesticated beets, including sugar beet, leaf beet, table beet and fodder beet, and sugar beet is the most recently derived[1]. While beets were already cultivated as vegetables around 500 BC in the eastern Mediterranean area[2], the latest domestication events were very recent. Leaf beet was the first crop to derive from the sea beet, and was already cultivated in the Roman times in the Mediterranean area, for food and medicinal purposes[3]. Table beet (or garden beet), selected for its swollen root, derived from the leaf beet[4]. Fodder beet was domesticated around a thousand years ago[3], and gave birth to sugar beets, which were selected and grown for sugar production for the first time by Franz Carl Achard in the beginning of the 19th century. He selected the white Silesian fodder beet for its high content in sugar, Silesia being an ancient European region, most of which is Poland today. This beet contained about 6% sucrose[5], which corresponds to a third of the sucrose content of current sugar beets[6]. Due to this late domestication and the very recent bottleneck event it went through, the genetic diversity of sugar beet is poor.

Using the Nei's genetic distance coefficient[7], assessing genetic distance depending on the occurrence of mutations and the application of genetic drift, Letschert *et al.*[8] calculated the genetic distance between several taxa inside the genus *Beta* and found that the subspecies *vulgaris* and *maritima* had a small genetic distance and were closely related[9] (**Figure 1**). This proximity must have been recognised by Linnaeus as he first placed both of them as different varieties in the same species in *Species plantarum*, in 1753, and only later, in 1762, did he take out the *maritima* variety to form the distinct species *Beta maritima*[9]. The *Beta vulgaris* species being a recently diverged taxon, all the taxa are cross-compatible, and the hybrids generated produce an abundance of seeds[2].

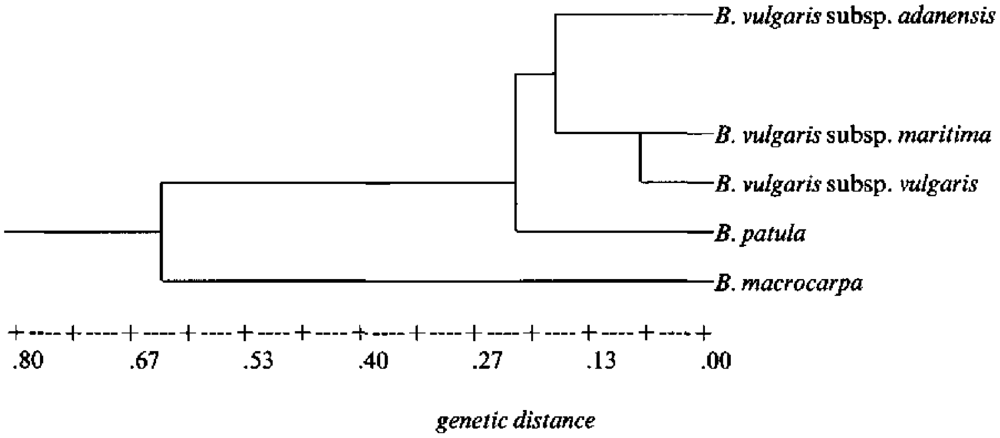

B. vulgaris subsp. adanensis
B. vulgaris subsp. maritima
B. vulgaris subsp. vulgaris
B. patula
B. macrocarpa
.80
.67
.53
.40
.27
.13
.00
genetic distance

Figure 1 - Dendogram representing the genetic distance (Nei's genetic distance coefficient) between five taxa of the genus Beta.

(Letschert J. et al., 1993)

In a recent study using a reference-free approach, Sandell *et al.*[10] explored the beet domestication history by measuring genomic distances between more than 600 wild and cultivated beets. Regarding sea beet accessions, two distinct groups were identified, Atlantic and Mediterranean. Furthermore, sugar beets were found to belong within the Mediterranean sea beet group, a sister to the Atlantic, revealing the genetic background of the wild beets at the origin of the domestication process, hundred years ago. It was suggested that the beet domestication probably occurred in the Greece area. In regard to the relationships between the members of the *Beta* genus, this study leads to questioning the potential belonging of *Beta vulgaris adanensis* to a *Beta* species on its own.

I - A. 2. An economically important crop

Sugar beet is one of the most cultivated crops in the world in terms of weight (FAOSTAT). It is responsible for 20% of global sugar production[11] and is the major source of sugar in Europe. It is principally grown in Russia, the United States, France and Germany (FAO).

Sugar beets are sown in spring and harvested in autumn, before the cold period, as cultivated beets require vernalisation to bolt. Once harvested and brought to nearby sugar factories, the sugar is extracted from the beets by diffusion, associating heat with disruption of the beet bits by slicers and a flow of counter-current water. The obtained juice containing 15 to 18% of sucrose is then purified using lime, carbon dioxide and sulphur dioxide. This is followed by an evaporation step and the generation of a syrup from which sugar is crystallised. Sucrose from the remaining molasses is recovered by chromatographic processes, or molasses are also used to feed animals and in production of yeast, citric acid, ethanol and fuel ethanol by fermentation[12].

Fuel ethanol from sugar beet molasses mostly produced in Europe is currently mixed with petrol and is expected to be a good alternative of traditional fuels to reduce the use and the dependence of depleting fossil fuels. One ton of sugar beet is estimated to produce approximately 90 litres of ethanol and one ton of beet molasses is estimated to produce approximately 260 litres of ethanol[13].

Due to its important agronomic value, the genome of the sugar beet has been investigated, and the first genome assembly was released in 2014 by Dohm *et al.*[14]. The current sugar beet reference genome assembly was published in 2022 by McGrath *et al*.[15]. On the wild side, the sea beet genome was first sequenced and assembled few years later than its cultivated counterpart, in 2019[16]. However, this work generated a draft assembly, and there is currently no high-quality reference genome for this subspecies.

I - A. 3. Challenges in maintaining a sustainable agricultural system

I - A. 3. a) Modern agriculture and its limitations

Presently, 40 to 50% of the Earth's land surface is devoted to agriculture, which represents more than 5,000 Mha including 100 Mha devoted to permanent crops, providing crops for several consecutive years[17]. Despite our focus on food production at an industrial scale, more than eight hundred million people still suffer from hunger all over the world today (FAOSTAT), and human population is growing. To successfully feed the planet, food production must be increased by approximately 70% by 2050 and by 200 – 300% by 2100. In 2009, one study led by the French Ministry of Food, Agriculture and Fisheries suggested that by 2050, cultivated lands could be multiplied by 1.7 to 2.5 in comparison with the cultivated lands in 2005. However, this would imply an encroachment on lands reserved for other uses, and, in the most aggressive scenario, a reduction of a third of the world's forests[18]. Therefore, the focus must be on sustainably increasing yield by intensification of agricultural practices, rather than increasing land areas. Moreover, this must be combined with the introduction of measures to reduce food waste and a change in human diet[19]. One way to increase yield is via the reduction of waste through loss of crops to pest and disease: this can be achieved by improving crop resistance against these threats.

I - A. 3. b) Climate change

The United Nations predict reaching 9.7 billion people by 2050 and 11.2 billion people by 2100[19]. Growth of cities, the increase of human activity, pollution and the loss of green spaces and forests are driving climate change. Among other phenomenon, climate change lead to an increase of the global average temperature of 1-1.2°C since 1850[20], the melting of glaciers, a higher frequency of severe weather, and an increase of greenhouse gas emissions. The agronomic industry must be prepared for these changes because they will modify the environment in which crops must grow. The Broom's Barn sugar beet crop-growth model aims to predict the impact of climate change on sugar beet yield from 2021 to 2050 in Europe[21]. The model shows that, on the one hand, higher temperatures will increase yield because foliage canopy will develop earlier but on the other hand, severe water stress will decrease yield. Moreover, France and Germany are the largest producers of sugar in Europe and these countries will be particularly impacted by high seasonal variation in water availability. Unpredictable drought conditions can impact yield to a greater extent than change in average yield. A more recent crop emergence simulation study for the period between 2020 and 2100 in the north of France predicted a year-to-year variability in emergence rate of sugar beets ranging from 0% to 85%, after testing five different sowing dates. Major causes of non-emergence would be non-germination, death of the seeds either because of clod blocking the seeds or of surface crusting induced by drought[22]. These worrying predictions in terms of climate change place the improvement of crop yield at the heart of our current concerns to sustain the growing human population.

I - A. 3. c) Cultures are threatened by plant pathogens

Over the centuries, crop pests and diseases have been, and still represent a major problem as they are ravaging crop fields. They are of different types: rodents, weeds, insects, nematodes, viruses, bacteria and fungi. One of the most ravaging crop pests is the aphid, colonizing a quarter of the plant species in temperate regions[23]. This insect causes hundreds of thousands of tons of lost wheat and potatoes and millions of tons of sugar beets per year in Europe[24]. It feeds on plant sap and can reduce the plant's growth by inducing a bad fructification, leaf decolouration, necrosis, leaf and fruit deformation and the formation of galls. Moreover, aphids excrete honeydew on the leaf surface, and this can be colonised by sooty moulds, thus hindering photosynthesis. The biggest economic repercussion caused by aphids is that they carry a lot of phytoviruses, highly reducing crop yields after infection[24]. This example shows that a single pest can cause substantial damage to a crop and can also vector other pathogens.

Regarding the sugar beet culture, crops are threatened by the same pathogens encountered in the wild. The most concerning ones, from an agricultural point of view, are viruses, bacteria, and fungi. The fungal pathogens include necrotrophic fungi such as the root-rot fungi[25], and biotrophic fungi which, although not killing the plant, significantly reduce the sugar yield: *Cercospora beticola*, *Ramularia beticola*, *Erysiphe betae* (powdery mildew) and *Uromyces beticola* (beet rust). A link has been shown between an early rust infection in sugar beets and the presence of local wild sea beets[26]. This sharing of pathogens makes sugar beet/sea beet an interesting system allowing host/pathogen coevolution studies between wild and agricultural plants. Rust infection (**Figure 2**) can decrease sugar beet yield by up to 11%[26] Rust infection leads to a decreased root yield and therefore sugar content, and can also lead to an increased transpiration, a decrease in protein synthesis and an attenuated photosynthesis[26], which are considerations that must be accounted for, given the changing climate.

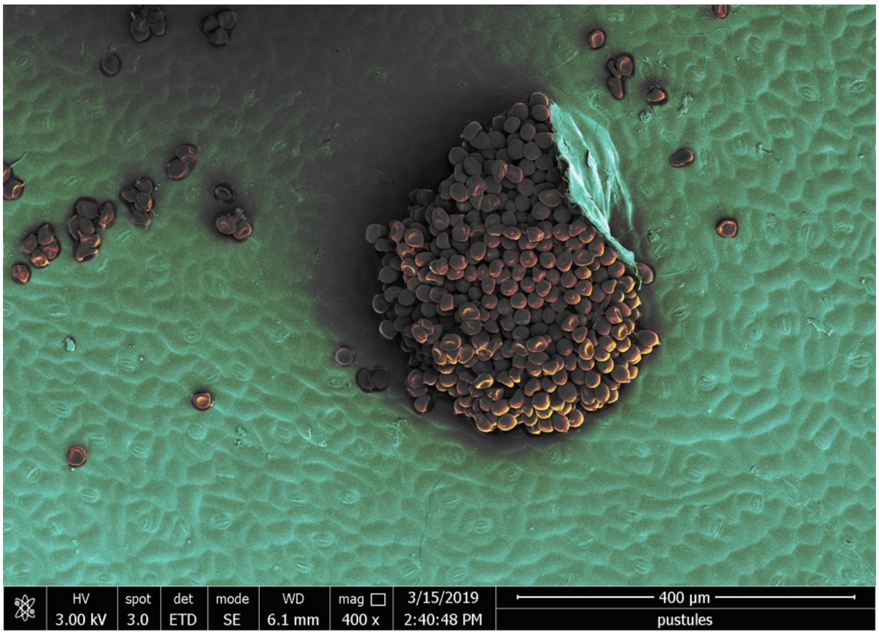


Figure 2 – Scanning Electron Microscope image of a rust pustule on a sugar beet leaf.

Credits for capture and colouration: Michelle Grey, Earlham Institute.

Uromyces beticola is an obligate biotrophic basidiomycete. It is an autoecious pathogen, which means that it can complete its entire life cycle on a same host[27]. Beet rust's life cycle is called macrocyclic as it is composed of five spore stages: urediospores, teliospores, basidiospores, pycniospores or spermatia and aeciospores[27]. This life cycle can be broken down into two infection stages: the first one setting up the early infection and the second one setting up the sexual stage[26] (**Figure 3**). Urediospores, spores produced by rust fungi, constitute the first infectious element of the fungus life cycle, and are produced in summer. These haploid binucleated spores penetrate the plant leaf by germinating through stomata. Once a susceptible plant is infected, urediospores can multiply by spreading from leaf to leaf and from plant to plant transported by wind and rain. When winter is coming, the sexual stage begins: urediospores transform into teliospores, which are diploid. These spores are unable to infect new tissues, but they can germinate into basidiospores, haploid spores generated by meiosis. Basidiospores can infect new leaves by germinating through the cuticula and the epidermis. A mycelium grows from these basidiospores, and pycniospores are produced, contained into an organ called pycnium. Pycniospores fuse with receptive hyphae from different pycnia. This fusion produces an aecium generating haploid binucleated aeciospores which can contaminate new leaves through the stomata. From there, the cycle life starts again with production of urediospores.

Usually, the first field crop infection comes from surrounding wild sea beets. As sugar beets are harvested in autumn, the rust life cycle only continues on sea beets during the winter and can infect sugar beet again when spring returns[26].

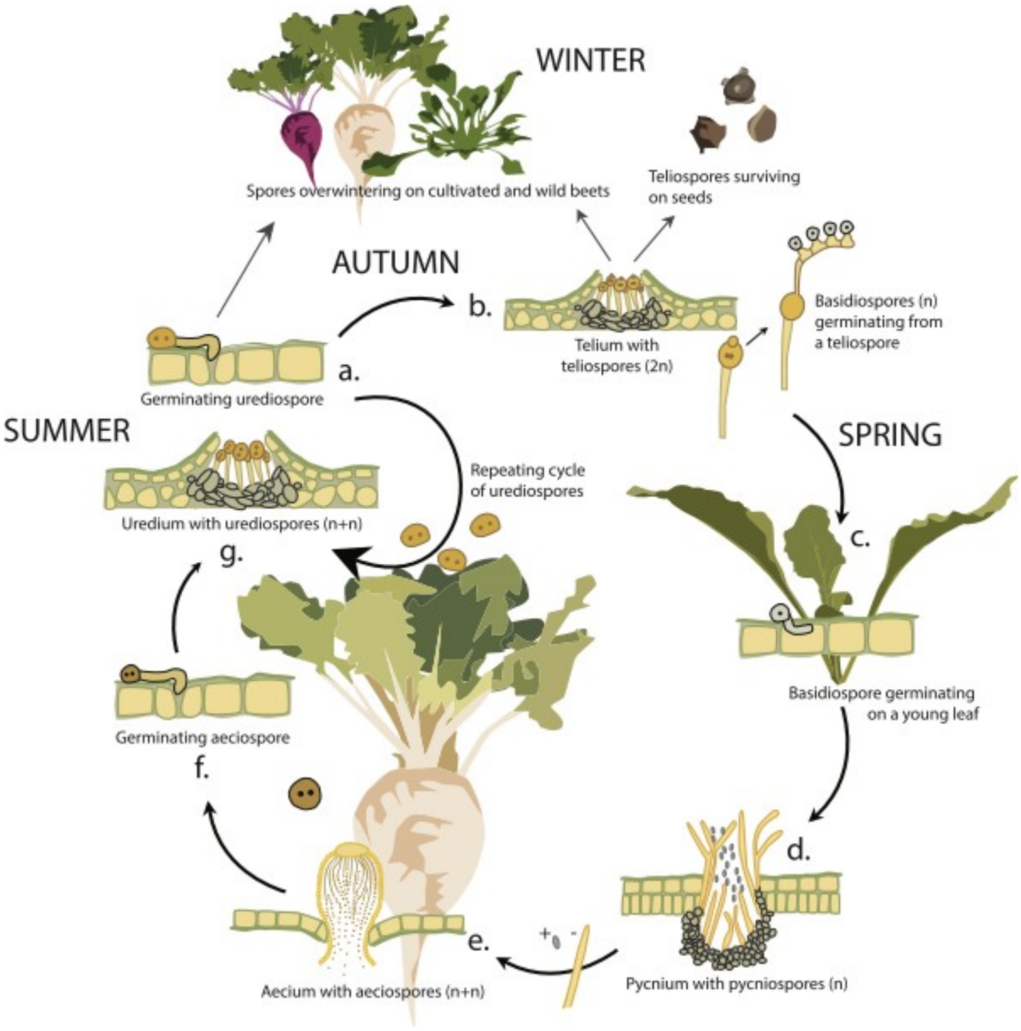

WINTER
Spores overwintering on cultivated and wild beets
Teliospores surviving on seeds
AUTUMN
b.
Telium with teliospores (2n)
Basidiospores (n) germinating from a teliospore
a.
Germinating urediospore
SUMMER
SPRING
Repeating cycle of urediospores
Uredium with urediospores (n+n)
g.
c.
Basidiospore germinating on a young leaf
Germinating aeciospore
f.
d.
e.
Aecium with aeciospores (n+n)
Pycnium with pycniospores (n)

Figure 3 - The life cycle of *Uromyces beticola*[26].

To summarise, the sustainability of crop cultures is threatened by many biotic and abiotic parameters including plant pathogens as well as environmental challenges such as climate change. This observation is even more true considering the sugar beet crop, whose very recent domestication implies a low genetic diversity, limiting its resources to face these challenges.

I - B. The plant/pathogen interaction

I - B. 1. Molecular bases of plant/pathogen interactions

To resist pathogen attacks, plants rely on a complex innate immune system which is organised into two defensive layers. The first is located on the cell surface: membrane receptors called PRRs (Pattern Recognition Receptors) are able to recognise a wide variety of pathogen-associated conserved molecules called PAMPs (Pathogen-Associated Molecular Patterns), such as bacterial flagellin. This first deployment of defensive weapons is called PTI, for pattern-triggered immunity.

To evade this recognition and improve their infection success, pathogens release small molecules called effectors. These effectors are released in plant cells or in the apoplast, and help and protect pathogen invasion through the manipulation of host components: they can for example inhibit the activity of plant hydrolytic enzymes, suppress plant immunity or trigger stomatal opening[28]. Plants have evolved a second layer of intracellular receptors which are able to specifically recognise these effectors; they are called NLR (Nucleotide-binding Leucine-rich Repeat) proteins[29]. This second layer of intracellular defence is called ETI for effector-triggered immunity. Plants encode hundreds of resistance genes and pathogens encode dozens to hundreds of pathogen effectors[30]. Pathogen effectors are also called avirulence proteins because of their molecular interactions in hosts: they induce an avirulent reaction when they are recognised by resistance host proteins[31].

NLR proteins have three main domains. In their N-terminus, most of the NLR proteins have either a coiled-coil (CC), a Toll/interleukin 1 receptor (TIR), or a RESISTANCE TO POWDERY MILDEW 8 (RPW8) domain. In their central part, they contain a nucleotide binding domain which controls the protein activation by switching from a bond to an ADP molecule to an ATP molecule[32]. Finally, NLR proteins are characterised by a C-terminal leucine-rich repeat (LRR) domain (**Figure 4**). LRRs consist of repetitions of a variable-sized motif including leucines and/or other hydrophobic residues, and are involved in protein/protein interactions. NLRs are classified according to their N-terminal domain, into TNLs (or TIR-NLRs), CNLs (or CC-NLRs), and RNLs (or RPW8-NLRs).

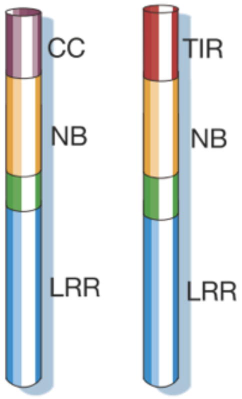


Figure 4 - Representation of the structure of NLR proteins (J. Dangle and J. Jones, 2001).

NLRs showing a coiled-coil domain in their N-terminal part are called CNLs or CC-NLRs whereas NLRs with a TIR domain are called TNLs or TIR-NLRs.

Recent observations highlight that the first (pattern-triggered immunity, extracellular) and second (effector-triggered immunity, intracellular) layers of plant immunity are actually not separate, but are interconnected and work together towards the ultimate immune response; a local cell death which prevents the pathogen from colonising new tissues. It has been shown in *Arabidopsis thaliana* that the immune extracellular receptors were required to generate an efficient ETI, and that they had a role in accelerating the hypersensitive response. However, the same was also true of NLR activation, which drives an increase in the presence of PTI proteins to sense extracellular pathogen components[33].

The foremost model of molecular interaction between host and pathogen suggests that recognition and binding between host resistance proteins and pathogen effectors stop the infection: the interaction is incompatible. This is called the gene-for-gene model. The pathogen is able to infect its host only if its avirulence factor is not recognised. This model was proposed by Harold Flor in 1951, and he suggested that "*for each gene determining resistance in the host there is a corresponding gene in the parasite with which it specifically interacts*"[34].

Since this model was proposed, it has been shown that the immune system was more complex than a two components interaction. NLR proteins can recognise pathogen effectors either directly or indirectly. One theory suggests that *R* proteins (immunity guards) are associated with guardee host proteins which are the targets of pathogen effectors. When an avirulence protein binds to its target host protein, a conformational change occurs, and this change is recognised by an *R* protein inducing the immune response[31].

Recent work on the interaction of effectors within networks posits that there are effector sensing/immune response initiating steps. NLR proteins can either function as singletons, i.e. they perform both functions, or they can form pairs[35]. More recent studies highlight the fact that NLR proteins interact within complex signalling networks, allowing a specific and effective immune response. As part of this network, some NLR proteins, called NLR sensors, directly or indirectly interact with pathogen effectors from a wide range of pathogens and need a co-receptor, called NLR helper, to translate the pathogen recognition signal into an immune response. The NLR helpers taking part in this network are genetically linked and belong to the NRC family (NLR required for cell-death), part of the CNLs. This family is present in the caryophyllales and asterids plant clades, and is thought to have diversified into a superclade, from a single pair of interacting NLRs (as present in the sugar beet genome), after the divergence between the kiwifruit and the rest of the asterids. The redundancy between the NLR helpers (**Figure 5**) is thought to be a strength, increasing the robustness of the immune system and enabling a rapid diversification of the NLR sensors[36].

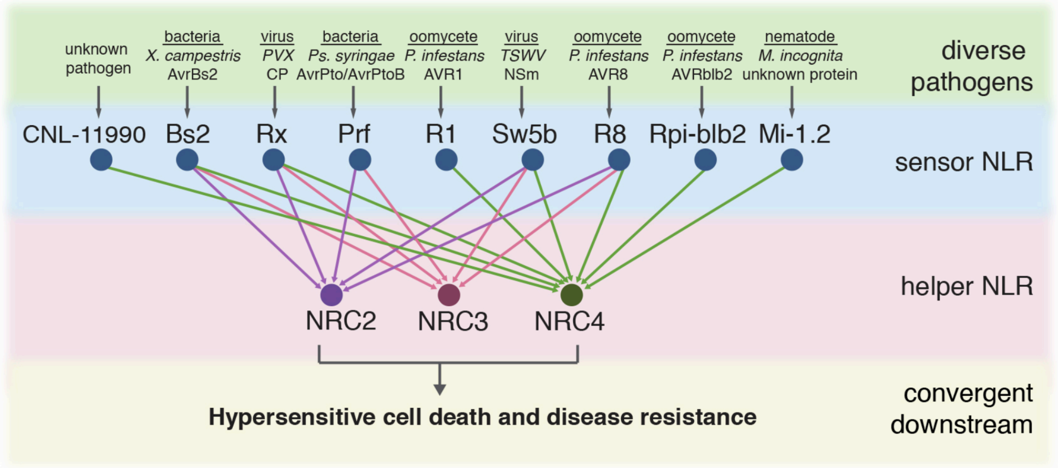


Figure 5 – A sensor-helper NLR network with redundancy of the NRC helper proteins in the *Nicotiana benthamiana* model organism (Wu *et al*., 2017).

The helper proteins NRC2, NRC3 and NRC4 are functionally redundant although displaying interaction specialities towards NLR sensors providing resistance against a wide range of pathogens.

The mechanism through which the activation of NLR genes leads to an immune response (i.e. programmed cell death), largely remains to be discovered, but recent studies bring major advancements. In *Arabidopsis thaliana*, the ZIR1 CNL gives resistance to *Xanthomonas* *campestris* and *Pseudomonas syringae*. Once activated after effectors recognition, this NLR undergoes an oligomerisation[37] and forms a pentameric complex called resistosome, by analogy with the mammalian inflammasome. This resistosome formation has been highlighted as well within the NRC network described above. It has been shown, in *Nicotiana benthamiana*, that the activation of both the Rx and the Bs2 NLR sensors from the CNL family, providing respective resistance against the *Potato virus X* and the *Xanthomonas campestris* pathogens (**Figure 5**), leads to the oligomerisation of their interacting NLR helper NRC2. This oligomerisation gives birth to a resistosome which associates with the plasma membrane[38]. Resistosome formation has been observed in TNLs as well. TIR-NLRs seem to work through homodimerization of the TIR domain. One example is the flax L6 NLR protein, which gives resistance to the flax rust *Melampsora lini*. Its activation leads to a TIR self-association[39].

While both CNLs and TNLs have been shown to form resistosomes after pathogen recognition, the way they induce the hypersensitive response seems to be class-dependant. On the CNL side, it has been shown that the ZIR1 resistosome was associating to the plasma membrane. It has been hypothesised that local cell death was triggered due to membrane disruption[40]. On the other hand, for TNLs, the resistosome formation is needed to activate the TIR NAD$^+$-cleaving enzymatic activity, which is required to transduce the pathogen recognition signal[41]. Further in the signalling pathway, TNLs require the EDS1 (enhanced disease susceptibility 1) protein, a basal component of plant resistance to biotrophic and hemi-biotrophic pathogens[42]. This protein interacts with the CNL helper NRG1 (N requirement gene 1) to induce the hypersensitive response[43].

The genome's composition in NLR genes varies between species[44], between individuals within the same species, and even within individuals from the same population[45]. This diversity is explained by the evolutionary history of this gene family and its rapid evolution. An analysis comparing 38 model organisms representing the six major taxonomic kingdoms revealed that while the TIR, nucleotide binding and LRR domains were present before the prokaryote/eukaryote split, the fusion of these key domains into NLR genes appeared later, in land plants (embryophytes). Complete TNL genes were first identified in bryophytes, whereas complete non-TIR NLRs were first found in lycophytes. This suggests that TNLs are the oldest class of NLRs[46]. The current NLR gene repertoire appears to result from the expansion of 23 ancestral genes (TNLs, CNLs and RNLs) in the last common ancestor of angiosperms[47]. CNLs underwent extensive expansion, while TNLs were lost in some groups such as the monocots, and RNLs did not experience significant expansion[47]. The expansion and diversification of NLR genes result from the process of gene birth and death[48], involving gene duplication events followed by functional diversification, or gene loss. This process is specifically active and rapid in clusters of NLRs, which are subject to tandem duplications. Due to their high sequence and number diversity, the fact that they can be organised in clusters[48] and their highly repetitive sequences[49], the annotation of NLR genes in genome assemblies is a challenging task requiring assemblies of high quality.

In sugar beet, a study based on the conserved nucleotide-binding domain of NLR genes (NB-ARC domain) identified 231 NB-ARC-like domains in a highly-continuous genome assembly, EL10. These domains were scattered over the nine chromosomes, with the majority of them present on the chromosomes 2, 3 and 7, and included 97 CNLs and 1 TNL[50].

I - B. 2. Plant/pathogen coevolution

Pathogens have evolved ways to avoid host recognition. Indeed, some pathogens have been shown to be able to produce molecules mimicking host molecules in order to evade host immune system[51]. They can also counteract the PRR-induced immune response through the action of virulence effectors[52]. In the same way that pathogens evolve to change their effectors, hosts also adapt to counteract pathogen strategies. These co-evolutionary processes generate new effector genes as well as resistance genes and this is thought to explain the diversity of NLR-binding accessory proteins[53].

These reciprocal changes in hosts and pathogens are evidence of coevolution. Coevolution is "*the process of reciprocal, adaptive genetic change in two or more species*"[51]. However, cultivated plants are known to lack genetic diversity by comparison with their wild ancestors. This is due to the bottleneck induced by, first, selecting relatively few wild individuals, second, the artificial selection process and the numerous rounds of crossing and inbreeding. In crops, the plant side of the coevolution is the responsibility of the breeder while the pathogens are free to evolve. In one of the oldest domesticated crops, maize (*Zea mays L.*), the domestication started in southwestern Mexico nine thousand years ago[54]. Although this crop is known to be genetically diverse, computer simulations have predicted that cultivated maize descended from a very small founder population of diverse progenitors: ten generations of approximately only twenty individuals[55].

This poor diversity in cultivars is also found in sugar beet. The calculation of genomic distances between wild and cultivated beets revealed a high level of relatedness in breeding material, among and between breeding companies.[10]

In 2008, Fénart *et al*. highlighted a lack of cytoplasmic diversity among cultivated beets. This was based on the analysis of four mitochondrial minisatellites between different taxa of the *Beta* section (**Figure 6**). Among the haplotypes studied, cultivated beets only presented one haplotype which is associated with the OwenCMS (a particular cytoplasmic male sterility only found in the cultivars) while the wild beets presented nine different mitochondrial haplotypes. They also performed an analysis of five nuclear microsatellite loci among these taxa and, measuring the allelic richness (number of alleles), showed that nuclear diversity was significantly lower in cultivated beets (Ar = 1.960) compared to sea beets (Ar = 2.87)[56].

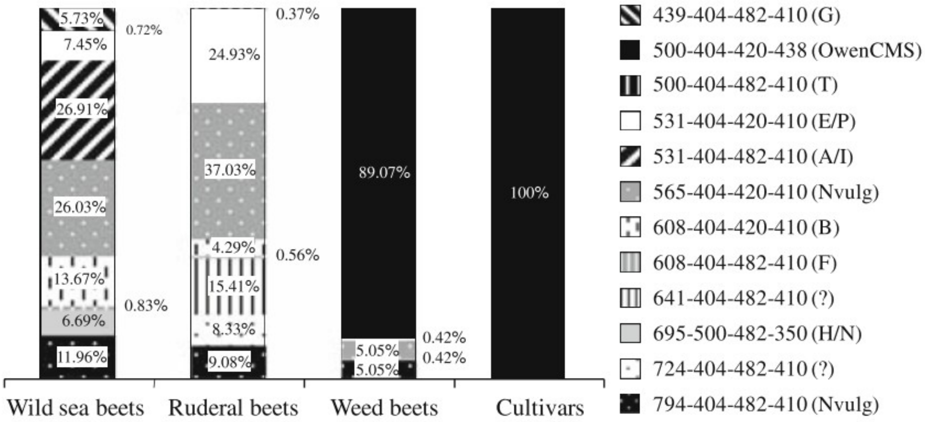


Figure 6 - Comparison of mitochondrial haplotypes between different taxa of the Beta section.

Each haplotype (on the right) is based on the combination of alleles from four mitochondrial minisatellites loci: 500, 404, 420 and 438 bp, for Tr1, Tr2, Tr3 and Tr4, respectively. The analysis has been done using 416 wild sea beets, 596 ruderal beets 481 weed beets and 147 cultivar individuals. (Fénart S. *et al*., 2008)

In a recent study[57] mapping hundreds of re-sequenced wild and agricultural beet genomes against a sugar beet genome assembly and calling for SNPs, genomic regions of low variation were identified in sugar beet. The enrichment of these regions in genes involved in processes agronomically interesting such as sugar transport and response to abiotic stresses, suggests the involvement of artificial selection in the observed loss of heterozygosity.

This lack of diversity goes hand-in-hand with a lack of resistance traits. Crops have become more susceptible to diseases because of artificial selection induced by the inbreeding processes and, in order to face this problem, the agricultural industry uses two lines of defence: a chemical solution through the use of pesticides and fungicides, and a genetic solution through the incorporation and deployment of resistance traits.

I - C. Methods to face crop pathogens

I - C. 1. Chemical methods

Chemical treatments are widely used in crop production to improve yield by reducing the losses due to competition with weeds and insects and infections caused by micro-organisms. A coating containing insecticides and fungicides is applied around the seeds sold to the farmers. Since 2005, neonicotinoids have been widely used in the EU in the composition of seed coating. This nicotine-like insecticide class is a systemic chemical spreading in the plant tissues, and it attacks the insect central nervous system. Since 2013, these insecticides have been strongly restricted by the European commission due to their deleterious effects on bee populations which represent a fundamental pillar in the ecosystem. Indeed, in 2015, it was shown that a neonicotinoid coating on oilseed rape seeds led to a reduction in the number of wild bee and solitary bees nesting sites. Moreover, the use of this neonicotinoid was negatively correlated with the growth and the reproduction of a bumblebee population[58].

Although the use of fungicides is efficient to control rust infection[26], it has also been associated with some disadvantages. Among fungicides used to counter beet rust infection, there are quinone outside inhibitors (QoI), as in the case of the pyraclostrobin fungicide. It belongs to the strobilurin fungicide group the member of which act as inhibitors of fungal mitochondrial respiration by binding the Qo site of the cytochrome b, which is the outer quinone oxidizing pocket. QoI prevents the electronic transfer to the cytochrome c, induces a lack of ATP (adenosine triphosphate) and subsequently a deficiency of fungal energy[59]. As this group of fungicides targets a single fungal site, it is subject to the risk of resistance development[60].

Another fungicide used in sugar beet fields to control rust is epoxiconazole, belonging to the triazole family, which are demethylation inhibitors (DMI) fungicides, sterol biosynthesis inhibitors. This fungicide family is believed to inhibit the cytochrome P450 which is involved, among others, in the steroid biosynthesis pathway. In fungi, this leads to a change of the membrane permeability and to the malfunction of membrane imbedded proteins[61]. In 2007, Taxvig *et al.* studied the toxic effects of epoxiconazole on mammals and showed that it was an endocrine disruptor[62]. The aim of the study was to analyse offspring of rats after exposure to this fungicide during gestation and lactation. Fetotoxic effects were observed and numerous dams were not able to deliver their pups. The observed effects on the offspring were a virilisation (anogenital distance increased in female foetuses and pups and in male foetuses, which is correlated with the degree of virilisation of genital development) and an affected reproductive development (reduced testes in males). This example shows how harmful fungicides can be for health and how it is important to focus on other strategies to combat plant pathogens infections.

Agrochemicals can be bad for human and animal health. Moreover, the efficacy of fungicides can be countered by the evolution of fungicide resistance. Indeed, even if the FRAC (Fungicide Resistance Action Committee) classes rust among plant pathogens with low risk of development of resistance to fungicides, it assigns to the triazole family a medium risk of resistance development and a high risk for the QoI fungicide family.

Due to these disadvantages and the recent loss of several chemicals, it is now vital to develop other strategies to face crop pathogens, such as improved biocontrol measures and of the breeding and judicious deployment of resistant varieties.

I - C. 2. Breeding for resistance and other traits

Introducing agronomically interesting traits by breeding has improved crop production by generating higher-yielding plants that grow faster and are more resistant to pests and herbicides. Moreover, the introduction of pathogen resistance by breeding has also allowed a reduced use of pesticides. However, plant breeding methods have limitations. First of all, it is longer to generate a resistant individual through selection and crossing stages than through transgenic methods[63]. Indeed, classical plant breeding consists of the selection of an interesting phenotypic trait in a population by crossing an individual of this population presenting the trait with a recipient line. Then, unwanted traits are removed by several back-crossings with the parental recipient line and selection of the offspring. Moreover, selecting one resistance gene to one particular pathogen can affect selection pressure on resistance genes to other pathogens[51]. Finally, intraspecific breeding leads to a lack of genetic diversity as inbred lines show a loss of heterozygosity linked with a loss of allelic diversity.

These observations lead to the conclusion that it is interesting to turn more on the search for traits in different species to gain genetic diversity. Moreover, turning to genetic modification saves time within the framework of the improvement of cultures.

I - C. 3. Bringing resistance from wild relatives

Crop wild relatives are an interesting source of genetic diversity, considering the large genetic diversity they harbour compared to their domesticated counterpart, and the propensity of their genes to adapt to a constantly changing environment. Breeders have been considering them from the beginning of the 20th century, mostly (at 80%) as a source of resistance to pests and diseases[64]. *Beta vulgaris ssp. maritima* was early considered as a source of interesting traits for *Beta vulgaris ssp. vulgaris* improvement, and an interesting trait was successfully brought to the crop for the first time in the 1910s by Munerati *et al.*, improving sugar beet resistance to *Cercospora beticola*[65]. Moreover, sea beet has also been used as a source of disease resistance with the deployment of the Rz2 gene, providing resistance against the rhizomania disease induced by a beet necrotic yellow vein virus. The CNL corresponding to the Rz2 resistance was identified in a wild sea beet population via a modified version of the mapping-by-sequencing method[66].

It is nonetheless difficult to consider wild relatives as species that can easily be bred with crops, even in the context of interbreedable *B. vulgaris* species. Indeed, they are so genetically rich in comparison to crops that the benefits associated with crossing them and selecting a particular trait is outweighed by the time involved in backcrossing to remove undesirable traits. Bad shape, woodiness of the roots are examples of these undesirable characters[11]. Therefore, breeders must find other ways to use this wild reservoir in an optimised way.

One of the ways to respond to this problem is the identification of resistance genes within a wild species and the cloning of these genes in the cultivated species. Resistance genes identification can be done through different manners, including association methods such as candidate-gene association studies or Genome Wide Association Studies (GWAS)[67]. The principle of these association studies is the highlighting of a correlation between genetic variants, most often SNPs (Single Nucleotide Polymorphisms), and a phenotype variation[68]. Applying these studies to wild accessions and landraces can help improving crops by allowing the identification of genomic regions associated with agronomically interesting traits.

Candidate-gene association studies involve speculating that a gene may be associated with a phenotype, such as a disease; identifying variants of this gene or of regions that could have an impact on the expression of this gene; genotyping the variants between two populations presenting two distinct phenotypes and statistically determining whether there is an association or not between the gene and the phenotype[69].

Next-generation sequencing methods started to be developed in the early 2000's and are much cheaper and faster than Sanger method. This step forward contributed to a reduction by 105 times of the cost per human genome sequenced and paved the way for more studies underpinned by sequencing[70]. Interestingly, the drop in the price of sequencing methods allowed more genomic analyses and better genome annotation of wild species[68]. A sea beet draft genome sequence was published in 2017[71]. This improvement enables the emergence of genome wide association studies, unbiased and more effective than candidate-gene methods because comparison of variants in most of the genome avoids having to pose a hypothesis about causal genes[67].

Although GWAS methods are innovative and promising, they are associated with different issues and constraints when performed in the wild[72]. One of the disadvantages of wild populations is that there are fewer genetic markers available in non-model than in model organisms, and wild population structures are less known[73]. Furthermore, it can be difficult to obtain a large number of independent individuals from the wild samples[68]. This can lead to the presence of false-positive or false-negative results that correspond to the highlighting of stronger or weaker associations, respectively, than they actually are[74]. Finally, these problems can be the source of difficulties to obtain reproducible results between independent studies[74].

The aim of the present work is to deepen the understanding of the genetics of the wild sea beet to be able to use it as a source for sugar beet resistance against the beet rust fungus, and to measure the gene flow of these potential candidate genes in natural wild populations across England. To do so, eleven sea beet genomes tested for rust resistance are selected in different populations in England, France, Spain and Denmark, for high-quality genome sequencing and assembling (see **Chapter II**). This collection of genomes is screened to generate a compilation of candidate NLR loci used in a second step as references to map associations retrieved from three large-scale rust inoculation experiments (see **Chapter III**). Indeed, a *k*-mer-based association genetics method is applied on sequencing data from approximately 500 sea beets selected for their extreme phenotypes in rust inoculation trials involving approximately 1,800 sea beets and three rust samples. The use of *k*-mers, sub-sequences of DNA with a given length noted *k*, in association genetic studies, helps to avoid the limitation of relying on a reference genome, which doesn't comprehensively represent the whole set of genetic variation which can be encountered in wild populations. Instead of mapping sequencing reads to a single genome and identifying association signals corresponding to SNPs, the partitioning of each genome into a collection of overlapping sub-sequences allows to compare to each other the sets of variations encountered in the hundreds of genomes tested. Finally, the population structure of the sea beets sampled in England is defined (see **Chapter IV**), and major population genetics statistics are measured to study the overall gene flow between populations, in a first step, and, in a second step, to zoom into five candidate resistance genes and study the population genetics patterns they exhibit.

Chapter II - Generating a sea beet pan-genome

II - A. Introduction

II - A. 1. The pan-genomic era

Whole genome sequencing started almost thirty years ago, with the 1.8 mega-base genome of the bacterium *Haemophilus influenzae*[75]. A few years later, the genome of the model *Arabidopsis thaliana* was the first plant genome to have been fully sequenced and assembled[76]. More than twenty times larger, with 3.2 giga bases, the human genome was fully sequenced shortly after, in 2001[77]. Those first whole-genome sequencing projects preceded a revolution in DNA sequencing. The early 2000s witnessed the development of what has been called “next generation sequencing”. This term groups sequencing techniques post Sanger sequencing technology. These more modern technologies are both more rapid and affordable, and quickly led to larger sequencing projects involving whole genomes.

Analysing the genome of single individuals soon becomes a limiting factor and hampered understanding species-wide genetic diversity. Within single species, two genomes can differ, both at the gene level, with variations such as gene presence/absence[78], SNPs, insertions, deletions, duplications, etc. as well as at the genome level with genomic rearrangements such as translocations and inversions[79]. Initially, this within species genetic diversity was determined using a single “reference” genome and then mapping multiple other individuals using much lower quality sequence data. These methods accurately capture SNP in single copy orthologous regions. However, using a single genome limits the observation of structural variants. Moreover, it limits Genome-Wide Association Studies (GWAS), where genes[80] or structural variants[81] involved in the phenotype of interest are not identified because they are not present in the reference genome. Consequently, a single genome, as accurate as its assembly can be, only captures a fraction of the intra-species genetic variation.

This observation led to the emergence of the concept of pan-genome in the early 2000s, in the vaccine development context. Tettelin *et al.*[82] were using reverse vaccinology, where the genomic sequence of the pathogen is bioinformatically parsed to predict candidate antigens, to develop a vaccine against the main cause of neonatal infections, the bacterium *Streptococcus agalactiae*. Several definitions have then been introduced, such as the core genome, referring to the collection of genes found in every individual of a given species, mainly assuring essential functions, and the dispensable genome (also referred as variable or accessory genome), containing the genes shared between a subset of genomes. The dispensable genome can be split into different categories being the soft-core, the shell and the cloud genomes. In the present work, the soft-core genome is referred as the content shared by more than 90% of the individuals. The extended soft-core genome is referred as the combination of the core and the soft-core genomes. The shell genome includes the content shared between 10 and 90% of the individuals. Finally, the cloud genome is the most variable part, corresponding to the content present in less than 10% of the individuals.

Tettelin and colleagues clearly showed that using the sequence of a single genome was impeding the development of universal vaccines, candidate vaccine antigens mainly being found in the dispensable genome. Moreover, they noticed that eight sequenced genomes were not sufficient to observe the full gene repertoire of *S. agalactiae*, and mathematically predicted that this would be reachable only after sequencing more than a hundred of genomes. From those observations, came up the concepts of opened and closed pan-genomes. A pan-genome is open when the sequencing of new individuals increases the size of the total gene collection known so far. Conversely, when the pan-genome has captured every single gene species-wide, this pan-genome is said to be closed.

Scientists were initially less interested in generating plant pan-genomes because, in addition to costs related to sequencing larger genomes, it was initially thought that the dispensable genome wasn't as significant as in bacteria. Moreover, the high content of repetitive sequences in plants make their genome more difficult to assemble[83]. Nevertheless, over the last decade, more and more plant pan-genomes were generated, both in the context of fundamental research with plant models such as *Arabidopsis thaliana*[76], as well as in the agricultural context with crop genomes (e.g. soybean[84], tomato[85], rice[86], wheat[87]). The first published plant pan-genome was of *Glycine soja*, the wild relative of the cultivated soybean[84]. The genomes of a geographically diverse panel of seven individuals have been short read sequenced, *de novo* assembled and compared, and generated a pan-genome with 80% of core genes and 20% of dispensable genes. Agronomic traits have been associated with this dispensable genome, such as seed composition, flowering and maturity time organ size and final biomass, and particularly NLR genes associated with biotic resistance. This highlighted the usefulness of pan-genomes to identify agronomically relevant genes. Nowadays, pan-genome sequencing is becoming a standard method and the list of achieved projects keeps growing (e.g. pearl millet (2023)[88], human (2023)[89]). In plants, the recent *Arabidopsis thaliana* pan-genomic study involving a diverse panel of 32 ecotypes from different continents[81] highlighted the importance of the dispensable genes for local adaptation, with genes mostly involved in stress response such as drought tolerance. The largest number of genomes sequenced for the generation of a pan-genome is so far attributed to the chickpea, with a total of 3,366 genomes, both from cultivated and wild individuals[90]. This allowed understanding the history of the domestication of this crop and will give considerable keys for chickpea breeding improvement.

Pan-genomes are then a considerable tool for breeding and crop improvement. Even though plant pan-genomes haven't been shown to have a dispensable genome larger than the core genome (i.e. 31.2% in *A. thaliana*[81]) as it is the case for some bacteria[91], seeing that genes involved in local adaptation and host/pathogen interactions are mainly present in this dispensable genome makes it extremely interesting to explore, especially for resistance/tolerance development in the crop breeding area. Indeed, if this dispensable genome tends to contain genes important for survival in different habitats, or against different pathogens, its size is less important than its content in terms of genes. It has indeed been shown that a large proportion of the genes providing resistance to plant pathogens are found in the dispensable genome (43% in *Brassica napus*[80], 47% in *A. thaliana*[92]).

Crop wild relatives constitute a fantastic source of novel agronomically important traits to improve crops, such as climate adaptation and pathogen resistance genes. Unfortunately, they are rarely sampled and sequenced with enough effort to see the extent of their dispensable genome. Instead, the crop plants that are sequenced to this level are already bottlenecked and unlikely to show the true level of difference between individuals. Taking a step further, genomic diversity across species has inspired the concept of the super-pan-genome[93]. The super-pan-genome is defined as the gathering of the pan-genomes of every species belonging to a given genus into a single super gene repertoire. This truly opens doors regarding crop improvement, generating a huge gene repertoire to parse for the identification of new agronomic traits, the genomes of crop wild relatives being far more genetically rich than their domesticated counterparts.

II - A. 2. The importance of generating a sea beet pan-genome

Sea beet is the wild ancestor of the sugar beet crop, the first source of sugar production in Europe[94] and the second source of sugar production in the world[95]. Due to its economic importance, it is a major crop and sugar beet breeding is important to sustainably maintain sugar production despite the emergence of new threats such as beet pathogens and climate change. Crop wild relatives, including sea beet[96], have been shown to constitute a fantastic source of genetic traits to improve their cultivated counterpart[97]. Moreover, a broader interest in this species resides in the fact that, because of the very recent domestication of the sugar beet, it is relatively easy to compare with its progenitor.

Unfortunately, the reference genome for the *Beta vulgaris maritima* subspecies is still only at the draft stage. The generation of a high-quality genome would constitute a great support to the beet breeding programs. Furthermore, the development of a pan-genome would allow greater comprehension of the genetic diversity available among wild reservoirs.

The present work has as principal aim identifying the utility of wild English sea beet as a source of resistance genes to improve sugar beet resistance against its rust pathogen (*Uromyces beticola*). In this context, the availability of multiple sea beet high-quality genome assemblies would significantly increase the chances of identifying resistance, when performing a Genome-Wide Association Study. Indeed, this would alleviate issues caused by gene absence or structural variations. Moreover, as resistance genes are expected to be a considerable part of the dispensable genome, the pan-genome notion is extremely well suited to study a panel of beets from diverse locations. Indeed, resistance genes are part of the genes permitting local adaptation[98], as they rapidly evolve[99] and can provide specific resistance towards local pathogens which are evolving different pathogenic components in distinct locations. The genome sequence of sea beets from different sites across Europe enables then the exploration of a large range of resistance components.

II - A. 3. Pan-genomes rely on a collection of high-quality genome assemblies

The utility of pan-genomes is twofold, it lies in their ability to describe the extent of a species' genetic reservoir by presence absence as well as allelic diversity. A pan-genome analysis relies on high-quality genome assemblies which can be compared with each other, requiring high contiguity, completeness, the absence of contamination, and great annotations.

High contiguity can be a specific challenge in plant genomes, as they are known to contain a large proportion of repetitive sequences[100]. Moreover, plant genomes are rich in duplications. Indeed, almost all plant lineages show whole genome duplication events[101]. Gene duplication produces paralogous genes, which can have a high sequence identity and can lead to assembly errors. Other plant features hampering error-free assemblies include their tendency to show high heterozygosity and ploidy, making it difficult to resolve haplotypes[102]. These problems can be addressed by utilising methodologies such as paired-end read sequencing, which consists of sequencing approximately 250 bp at both ends of a DNA library. Also, long-read methodologies, as provided by Pacific Biosciences or Oxford Nanopore, which produce reads that are regularly expected to exceed 10 kilobases in length[103]. The appearance of High-Fidelity (HiFi) long-read sequencing with Pacific Biosciences enabled the combination of the high accuracy of short-read sequencing with the benefits of long-read sequencing, using a technology involving a circularised DNA molecule around which the DNA polymerase passes multiple times. This allows the generation of consensus reads of an average length of 13.5 kilobases and an accuracy of 99.8%[104].

In the last years, assembly methods have evolved. Two main assembly graphs are encountered: the Overlap-Layout-Consensus (OLC) approach and the de Bruijn graph (DBG) approach[105]. With the OLC method, the sequencing reads are connected to each other when they share similar prefixes and suffixes. As a representation, nodes represent the sequencing reads, and edges are formed by the overlaps between them. The DBG approach deals better with short reads, being easier to resolve and less demanding in computing resources, but more sensitive to sequencing errors, it uses the concept of *k*-mers to link the sequencing reads[106]. Each sequencing read is split into every possible *k*-mer. Assuming that each *k*-mer is unique in the genome, in the graph representation, nodes stand for the overlaps between *k*-mers, corresponding to sequences of a length of *k*-1, and edges for the *k*-mers. Once the graph is generated, the best path between the reads is determined, and the remaining sequences are discarded. More recently, other genome assembly approaches deal with HiFi long-reads and take into account haplotypic information[107]. Phased assemblers have been developed, the two leading ones being HiCanu[108] and Hifi-asm[109], which are able to separate sequencing information into different alleles. Hifi-asm was shown to be the most efficient one, in terms of contiguity, rapidity and completeness[107]. It generates a string assembly graph, similar to the OLC approach but using exact overlaps between reads, and this graph only involves reads from a same haplotype.

Assembly "contiguity" is assessed by numerous statistics that describe the number and the length of the contigs generated. Statistics are used, such as the N50 which represents the size of the shortest contig at half of the total assembly length, or the L50 which is the count of the smallest number of contigs whose length sum makes up half of the assembly size.

While contiguity describes the proportion of the genome on large contigs, the quality of genome assemblies is also assessed by their completeness. One way to judge on the completeness of an assembly is by determining its content in a set of phylogenetically conserved orthologous single copy genes called BUSCO for Benchmarking Universal Single-Copy Orthologs[110].

Another aspect of the quality of an assembly is its purity. When assembling reads into a genome, contaminants can be incorporated and generate erroneous contigs. This contamination can occur at different stages of an experiment: during the material sampling step if the individual of interest is hosting another species, for example, or in the DNA extraction step, if cross-contamination occurs between different experiments. Moreover, human contamination has been shown to be present among published genome assemblies[111]. To avoid the impact of contamination of genomes used in different projects, it is important to ensure that they are pure, and tools have been developed to identify contaminant contigs. BlobTools[112] is a tool which partitions contigs into taxa, using the Blast tool[113] to search against the NCBI database, and, this way, allows the attribution of contigs to species, other than the one of interest. Moreover, BlobTools displays the GC percentage of each contig of the assembly, which is another way to identify erroneous contigs, as the GC percentage tends to differ between genomes.

The accelerating pace of genome sequencing and assembly doesn't unfortunately go hand in hand with the pace of genome annotation[114]. Indeed, *de novo* annotation of recently assembled genomes is a complex, costly and time-consuming process. To enable speeding up the process, a more affordable method has been developed, consisting in carrying over the annotation of a reference genome from the same or a related species onto the newly assembled genome, such as the Liftoff tool[114].

In the same way, the annotation of resistance genes is arduous as well, due to their highly repetitive sequence[115] and their low basal expression[116]. Recently, the NLR-Annotator tool has been developed to *de novo* annotate loci associated with NLRs, without the need of gene expression data[117]. NLR-Annotator parses the genome and searches for combinations of amino acid motifs extracted from a set of potato NLR proteins, after translating genomic fragments into the six possible reading frames. Then, the nucleotide positions of those motifs are identified and the distance between them is evaluated in order to generate a list of predicted NLR loci.

Given the need for genomic representatives of the sea beet wild plant to inform on resistance gene potential, the present work aims to provide the material for the generation of a sea beet pan-genome, i.e. multiple high-quality genome assemblies. Indeed, a strictly speaking pan-genome often visualised as a graph isn't generated here, as orthogroups presence/absence is analysed, but the eleven high-quality genome assemblies pave the way towards the generation of the first sea beet pan-genome. The results presented below describe these assemblies created from the sequencing of a set of eleven wild plants mostly sampled in England, but also from France, Denmark and Spain. RNA-sequencing was not used to individually annotate these genomes, but NLR genes were specifically annotated to facilitate the association analyses performed in the **Chapter III**.

II - B. Results

II - B. 1. Eleven sea beet genomes each represent an improvement in contiguity to the published genome

The genomes of eleven sea beets originating from different European locations have been sequenced and assembled in the present work. The pan-genome generated contains members of the two genetically distinct groups of sea beets described by Sandell and colleagues[10], namely the Atlantic group and the Mediterranean group. *k*-mer-based analyses on whole genomic data show a split of the samples into four clusters: eastern English sea beets (Gb_Humber_260, Gb_Norfolk_095 and Gb_Norfolk_426), north-western English sea beets (Gb_Merseyside_109 and Gb_Merseyside_206), south-eastern English sea beets (Gb_Essex_167, Gb_Essex_038 and Gb_Suffolk_251) grouped with Zealandian (Dk_Sjælland_406) and Breton (Fr_Bretagne_309) sea beets, and Mediterranean sea beets (Es_Catalonia_378) (**Figure 7**).

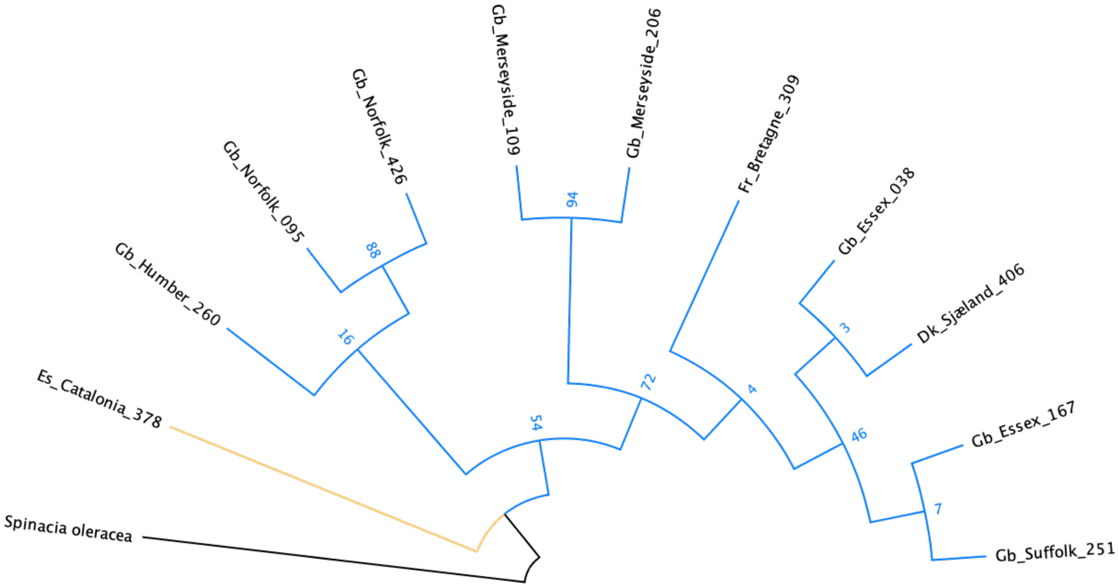


Figure 7 – *k*-mer-based comparison between the eleven sea beet genomes.

Mash distances have been calculated on whole genomic data, with k = 21. Genomes belonging to the Atlantic group described by Sandell *et al.* are indicated with blue branches, and the Spanish genome (Es_Catalonia_378), belonging to the Mediterranean group, is indicated with a yellow branch. The spinach (*Spinacia oleracea*) reference genome was used to root the tree. Bootstrap support values are based on 1,000 replicates.

To identify signs of contamination in each assembly, contigs were assigned to a taxon based on sequence similarity through a blast search against the NCBI database. Moreover, Blobtools[112] plots contig GC content against read coverage (**Figure 8**).

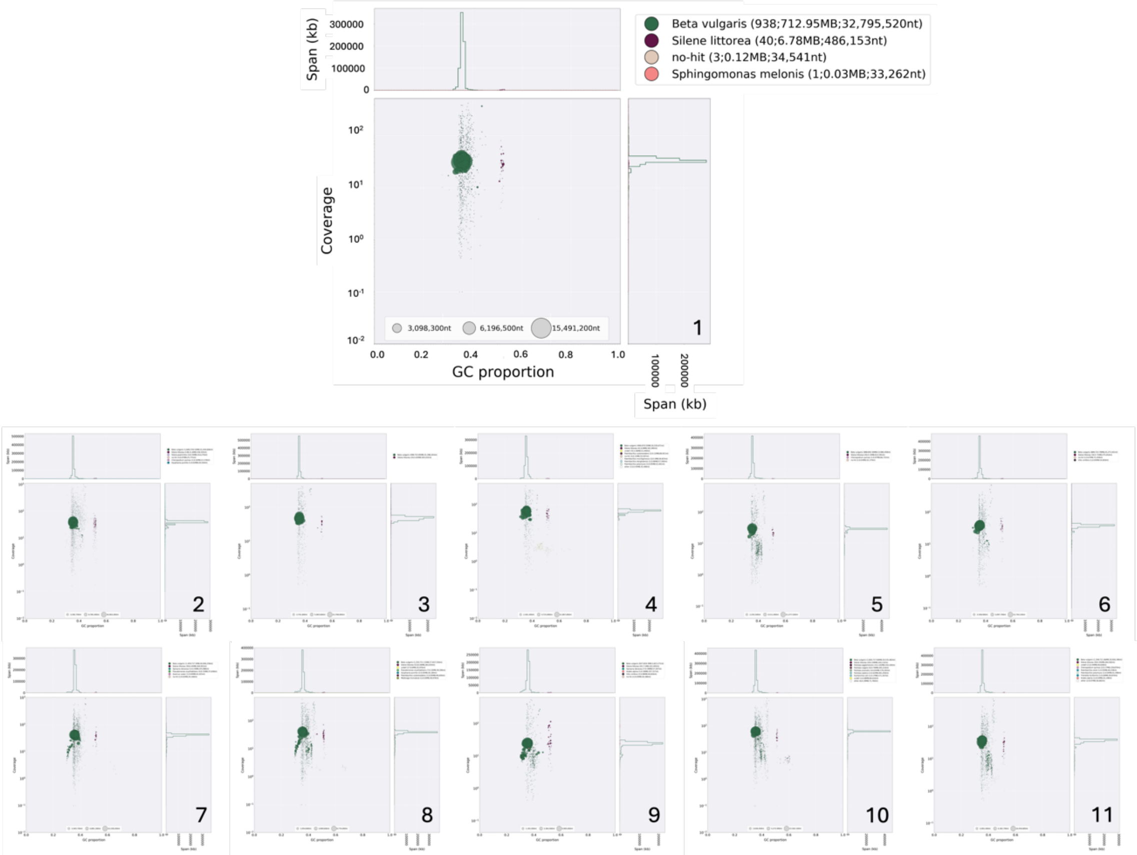


Figure 8 - BlobTools plots of the sea beet genome assemblies.

Contigs in the assemblies are depicted as circles, scaled proportional to sequence length and coloured by taxonomic annotation (at the species rank). Circles are positioned on the X-axis based on their GC proportion and on the Y-axis based on the average read coverage. 1: Gb_Norfolk_426; 2: Gb_Essex_038; 3: Gb_Norfolk_095; 4 Es_Catalonia_378; 5: Gb_Merseyside_109; 6: Gb_Humber_260; 7: Dk_Sjælland_406; 8: Fr_Bretagne_309; 9: Gb_Merseyside_206; 10: Gb_Suffolk_251; 11: Gb_Essex_167.

On a BlobTools plot, contamination is visible as contigs that Blast to unrelated species. Non target species identification can be caused by missing data in the database but is worth investigating where contaminant species live in the same environment, or are parasitic or endosymbiotic and could have been present in the DNA extraction. In these cases, however, because a parasite might be expected to have a different density of cells (a different number of genomes represented) the coverage associated with that contig may deviate from that of the focal species. Moreover, contamination may also differ in its GC proportion with the main proportion, as GC content tends to differ between genomes.

BlobTools results are comparable among eleven sea beet genomes. At the species level, the two predominant blast hists are for *Beta vulgaris* and *Silene littorea*, two species which both belong to the Caryophyllales order. There is no apparent sign of contamination among the studied genome assemblies as, in each plot, the majority of the assembly content is found at the same GC proportion as well as at the same coverage. There is, however, an exception for the contigs Blasting to the *Silene littorea* species, in every genome. Although those contigs show a higher GC, distinct from the main content, the observations that: it is present in every sequenced genome, that it belongs to the Caryophyllales order and that it is not from a parasitic or endosymbiotic organism suggest that these contigs don't correspond to contamination.

Prior to filtering, the sea beet genome Gb_Merseyside_206 did present signs of contamination at the classification level (**Figure 9 A,C**) as well as at the level of genome content (see **Figure 10**). This contamination is observed both at the species and at the order levels. These large blobs were present at a coverage 16 times greater than the expected genomic content (which is approximately 20x, **Table S1**). Contaminant contigs are attributed to Ascomycota fungi. To address this, the large contigs not associated with the Caryophyllales order were removed from this assembly (see **Figure 9 B,D)**. After removing contaminants, the BlobTools plot for Gb_Merseyside_206 (**Figure 9 B**) is similar than for the other genomes (**Figure 8**).

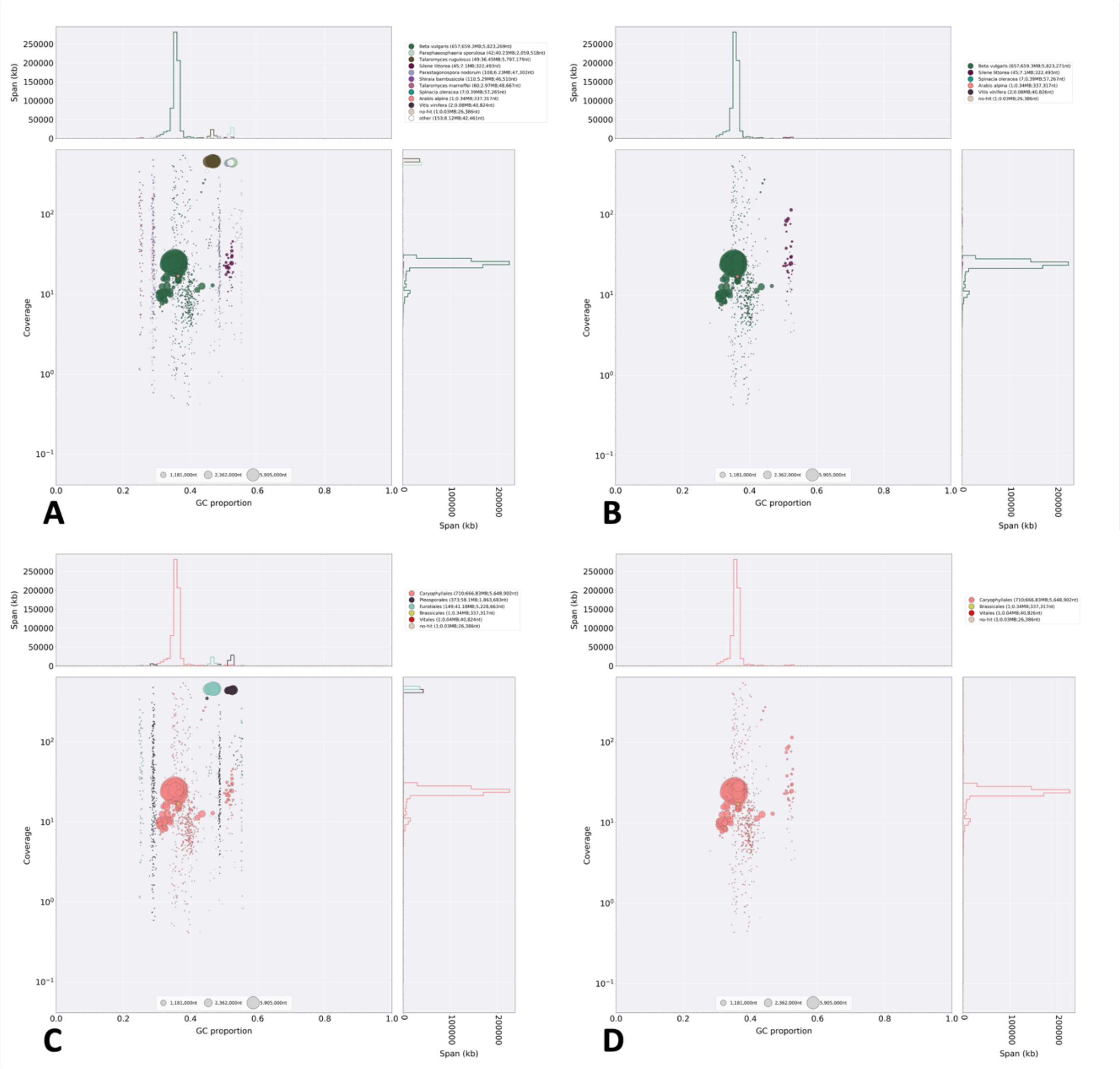


Figure 9 – Contaminated content was removed from the genome assembly Gb_Merseyside_206.

Contigs in the assemblies are depicted as circles, scaled proportional to sequence length and coloured by taxonomic annotation (at the species (A,B) or the order (C,D) rank). Circles are positioned on the X-axis based on their GC proportion and on the Y-axis based on the average read coverage. Plots on the left (A,C) show the content before contamination removal, and plots on the right (B,D) show the result post-decontamination.

To further assess the quality of the genome assemblies, a *k*-mer-based method to compare the assembly content and the sequencing reads content has been used; the comparison tool from the *K*-mer Analysis Toolkit (KAT[118]), was run on each genome and is interpreted here in terms of content in the reads that is present and absent in the assembly, as well as content depth, or multiplicity (**Figure 10**).

Interpreting these plots, firstly, a black peak (**Figure 10 panel 1, section A**) is present at a very low *k*-mer multiplicity. The black distribution represents those *k*-mers that are present in the reads but absent from the assembly, and this peak represents *k*-mers present at a low number of copies in the read set, and therefore, most likely correspond to sequencing errors. These erroneous *k*-mers have not been included in the genome assembly.

The red distribution represents the *k*-mer content, in the reads, that appears just once in the assembly. This should include most of the read content. If the genome Gb_Norfolk_426 is taken as an example, the red peak at a *k*-mer multiplicity approximately equivalent to the estimated sequencing depth (**Table S1**), i.e. at 39x, corresponds to the homozygous content (**Figure 10 panel 1, section C**). The other red peak at a *k*-mer multiplicity approximately half (~20x) corresponds to the heterozygous content (**Figure 10 panel 1, B**). The distinction between red peaks is not clear for few genomes: Gb_Merseyside_109, Fr_Bretagne_309 and Gb_Suffolk_251 (**Figure 10 panels 5, 8 and 9**), suggesting the ability of an assembler to partition haplotypic content may be reduced for these assemblies.

Interestingly, the observed depth is very close to the expected depth using the estimated genome size of Bmar (567 Mbp)[119], which is approximately 80% of the average size of the genome assemblies generated in the present work (average = 714 Mbp) (**Table S1**).

Within the red heterozygous peak, a black peak is present (**Figure 10 panel 1, D**). This corresponds to half of the heterozygous content which is not present in the genome assembly. In effect, this shows that both heterozygous and homozygous contents are only present once in the final assembly, and the black *k*-mers represent that heterozygous content in the reads, at half the expected depth, that are not present in the final assembly.

For each genome, the *k-mer* distribution is compatible with the distribution of a complete diploid individual. The *k*-mer comparison method used here provides a test for the presence of polyploid content among the sea beet genomes sequenced, as wild tetraploid beets have been observed, such as *Beta macrocarpa*. In the current case, there is no evidence for tetraploid content, as no extra peak with a higher *k*-mer multiplicity can be observed (**Figure 10 panel 1, section F**). A purple layer indicates the presence of some duplicated content which has only a marginal impact on these genomes.

The *k*-mer spectra can also inform on the presence of erroneously-joined contigs. Indeed, when joining contigs, assemblers can create new *k*-mers when these DNA fragments are not overlapping, hence present in none of the reads. This scenario would be observable through the *k*-mer spectra by the presence of *k*-mers present in the assembly but absent in the reads. They would be observable as a red bar on the rightmost part of the graph, where the *k*-mer multiplicity is equal to 0 (**Figure 10 panel 1, section E**). The analysis of the present *k*-mer spectra doesn't indicate the presence of erroneous content among any of the generated assemblies.

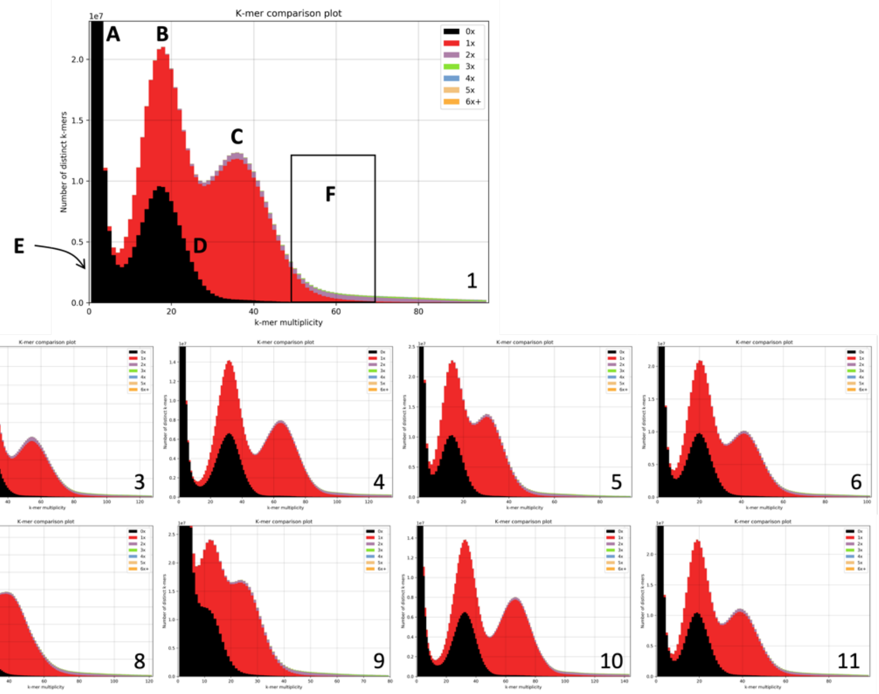


Figure 10 - k-mer comparison plots show expected k-mer spectra for good quality diploid genome assemblies.

For each genome, the plots describe the k-mer distribution among the PacBio sequencing reads. The number of distinct k-mers is plotted against the k-mer count. The colours of the peaks observed refer to the number of times the corresponding k-mer content is present in the genome assembly. 1: Gb_Norfolk_426; 2: Gb_Essex_038; 3: Gb_Norfolk_095; 4: Es_Catalonia_378; 5: Gb_Merseyside_109; 6: Gb_Humber_260; 7: Dk_Sjælland_406; 8: Fr_Bretagne_309; 9: Gb_Merseyside_206; 10: Gb_Suffolk_251; 11: Gb_Essex_167. In this analysis, k=27.

As seen previously, the *k*-mer spectra can be informative as to whether potential contaminant or erroneous content present in the reads has been removed from the assembly. This is exemplified in the genome Gb_Merseyside_206, which was previously shown by BLAST, coverage and GC content plotted by BlobTools, to have contamination (**Figure 9 A, C**). Here, we observe an extra red bump at a *k*-mer multiplicity approximately 20 times greater than for the first two peaks (**Figure 11 A**), and the expected coverage (see **Table S1**). This contaminant content has been removed, and this is observable here with a change in the colour of the bump from red (*k*-mers present in the assembly) to black (*k*-mers only present in the reads) (**Figure 11 B**). Due to their high *k*-mer multiplicity, the discarded sequences seem to correspond to highly repetitive content.

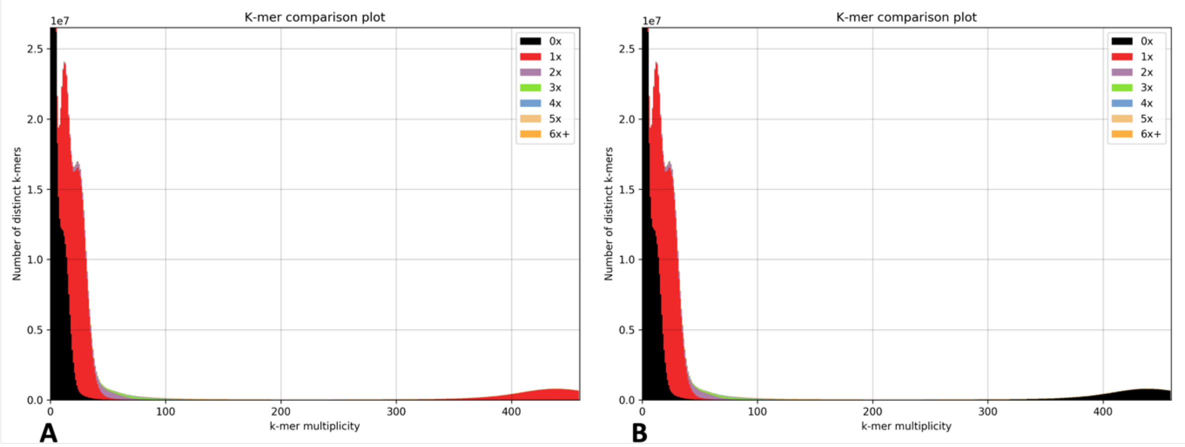
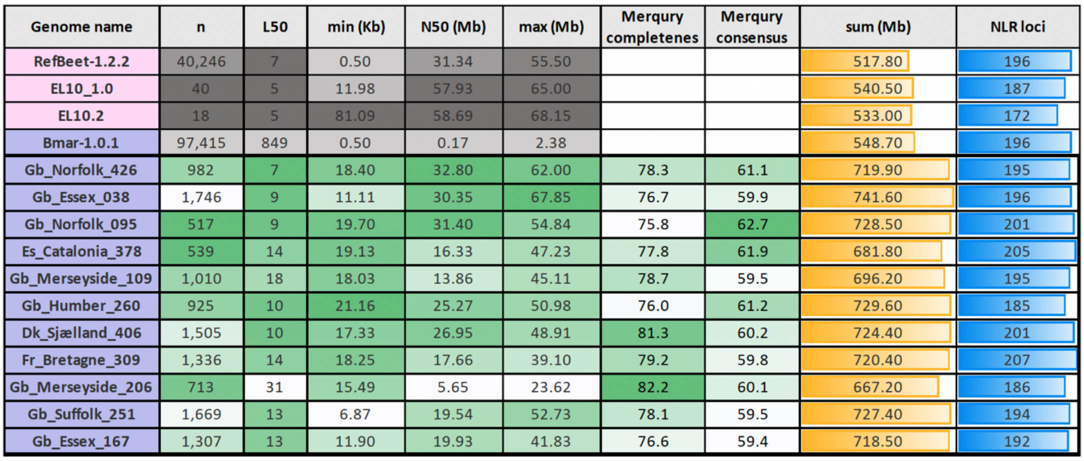


Figure 11 – Contaminated content has been removed from the genome D_206A.

The plots describe the *k*-mer distribution among the PacBio sequencing reads before (A) and after (B) that contaminated content has been removed. The number of distinct *k*-mers is plotted against the *k*-mer count. The colours of the peaks observed refer to the number of times the corresponding *k*-mer content is present in the genome assembly before (left) and after (right) removing contamination.

II - B. 2. Sea beet genome assemblies show comparable sizes and good levels of completeness

Eleven sea beet genomes are assessed against another sea beet genome (Bmar-1.0.1) and three sugar beet genomes (RefBeet-1.2.2, EL10_1.0 and EL10.2). Despite being more fragmented than the reference sugar beet genome (EL10.2), the eleven sea beet genomes show good levels of contiguity; N50 ranges from approximatively 5.6 to 32.8 Mb, considerably larger than for Bmar-1.0.1 (N50 = 172,314 bases), and the total contig number ranges from 517 to 1,746 (**Table 1**). The proportion of content (L50) ranges from 7 to 31 contigs.



Genome name	n	L50	min (Kb)	N50 (Mb)	max (Mb)	Merqury completenes	Merqury consensus	sum (Mb)	NLR loci
RefBeet-1.2.2	40,246	7	0.50	31.34	55.50			517.80	196
EL10_1.0	40	5	11.98	57.93	65.00			540.50	187
EL10.2	18	5	81.09	58.69	68.15			533.00	172
Bmar-1.0.1	97,415	849	0.50	0.17	2.38			548.70	196
Gb_Norfolk_426	982	7	18.40	32.80	62.00	78.3	61.1	719.90	195
Gb_Essex_038	1,746	9	11.11	30.35	67.85	76.7	59.9	741.60	196
Gb_Norfolk_095	517	9	19.70	31.40	54.84	75.8	62.7	728.50	201
Es_Catalonia_378	539	14	19.13	16.33	47.23	77.8	61.9	681.80	205
Gb_Merseyside_109	1,010	18	18.03	13.86	45.11	78.7	59.5	696.20	195
Gb_Humber_260	925	10	21.16	25.27	50.98	76.0	61.2	729.60	185
Dk_Sjælland_406	1,505	10	17.33	26.95	48.91	81.3	60.2	724.40	201
Fr_Bretagne_309	1,336	14	18.25	17.66	39.10	79.2	59.8	720.40	207
Gb_Merseyside_206	713	31	15.49	5.65	23.62	82.2	60.1	667.20	186
Gb_Suffolk_251	1,669	13	6.87	19.54	52.73	78.1	59.5	727.40	194
Gb_Essex_167	1,307	13	11.90	19.93	41.83	76.6	59.4	718.50	192

Table 1 - Sea beet pan-genome assembly metrics.

Principal assessment metrics of sea (purple) and sugar (pink) beet genome assemblies. n: number of contigs (present work) or scaffolds (published assemblies); L50: count of smallest number of contigs whose length sum makes up half of genome size; min: size of the shortest contig; N50: sequence length of the shortest contig at 50% of the total assembly length; max: size of the longest contig; sum: size of the assembly; NLR loci: number of NLR loci. To compare the different statistics across the beet genome assemblies, a grey (published assemblies) or green (newly assembled genomes) gradient is used (per metric), from low to high contiguity (light to dark, respectively). Genome sizes and content in NLR loci are compared across assemblies using, respectively, yellow and blue scales.

Merqury[120] is a *k*-mer-based program that uses sequencing reads set to assess the completeness and the quality of the assembled genomes. Completeness is measured as the ratio between the solid *k*-mers in an assembly against the solid *k*-mers in the read set, solid *k*-mers being distinct *k*-mers filtered for erroneous low copy *k*-mers. This value appears to be similar across the sea beet genomes. Interestingly there is a reduction in the completeness of the genome Gb_Merseyside_206, which was filtered for potential bacterial contamination (**Figure 9**, **Figure 11**). The Merqury consensus quality value is the probability of correctness in the assembly at a base level. This quality value is very similar between the sea beet genomes.

With the exception of Gb_Merseyside_206, the genomes assembled here are of similar quality in terms of contiguity and completeness. Gb_Merseyside_206's longest contig is approximately half the length of the median of the largest contigs from the other assemblies (49.9 Mb), and its L50 is more than twice the average L50 (13.45). Nevertheless, contiguity and completeness of each of these genomes is greater compared to the published sea beet genome (Bmar-1.0.1).

The sea beet genomes vary in size by approximately 10%, ranging from 667 to 742 Mb (mean=714 Mb). This is larger than the current sugar beet reference genome, EL10.2, with a size of 533 Mb. Furthermore, each one of those genomes is larger than the Bmar-1.0.1 sea beet genome, an assembly of 549 Mb.

To assess the completeness of the sea beet pan-genome, the Benchmarking Universal Single-Copy Orthologs (BUSCO[110]) software was used on sea beet genome assemblies generated in the present work in addition to previously released beet genomes (Bmar-1.0.1[119], RefBeet-1.2.2[14], EL10-1.0 and EL10.2_2[121]).

Complete and single-copy BUSCOs range from 90.9% to 93.0% across the eleven sea beet genome assemblies (**Figure 12**). These numbers are comparable to the levels of completeness found in published sugar beet genomes (RefBeet-1.2.2, EL10-1.0 and EL10.2_2), including the sugar beet reference genome (EL10.2_2). The proportion of complete and single copy orthologous genes is higher than the previously released sea beet draft genome assembly (Bmar-1.0.1). Moreover, the level of duplicated, fragmented, and missing BUSCOs are also higher in that published Bmar-1.0.1 assembly. This is consistent with the observation that the Bmar-1.0.1 assembly shows a higher number of contigs and a very lower N50 than the present sea beet assemblies (**Table 1**).

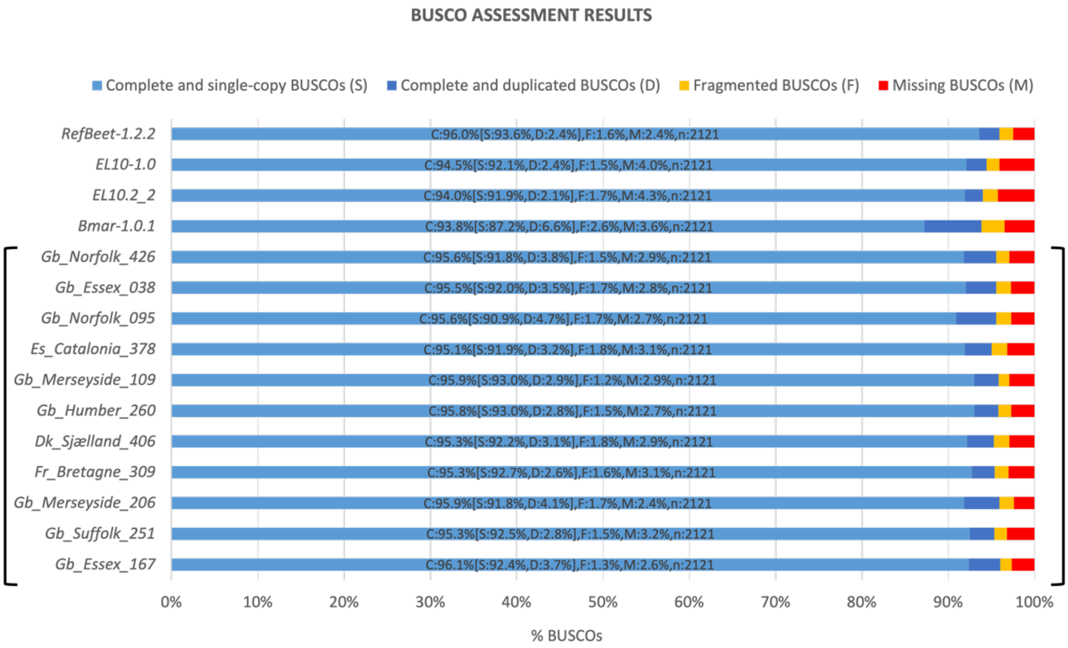


Figure 12 – BUSCO assessment results show the completeness of the sea beet pan-genome.

Percentage of universal single copy orthologs that are C: complete, S: single-copy, D: duplicated, F: fragmented, M: missing, n: gene number. Brackets are used to point out the eleven sea beet genome assemblies generated in the present work. BUSCO genome completeness is comparable between the present work and RefBeet-1.2.2, EL1-01.0, EL10.2_2 and Bmar-1.0.1 (three sugar beet genome assemblies and a sea beet draft genome assembly, respectively).

To summarise, the PacBio HiFi pan-genome generated in the present work is of higher contiguity and completeness than the currently available sea beet genome assembly Bmar-1.0.1 but not as contiguous as sugar beet genomes.

II - B. 3. The sea beet pan-genome content in terms of genes is comparable across samples

In order to explore the relationship of genome size variation and any impact on gene presence and absence across the pan-genome, genes were annotated lifting over genes from the reference sugar beet genome[15], and an orthogroup analysis was done. Orthogroups include both orthologs and paralogues. 99.7% of the annotated genes from every genome were classified into 19,454 orthogroups (**Table S2**). The distribution of the genes into a core/soft-core/shell and cloud genome whether these genes belong to orthogroups present in every genome, multiple genomes or a single one, is presented in the **Figure 13**. Overall, the number of genes per individual ranges from 21,157 to 21,237 and this number doesn't appear to be related to genome length. The majority of genes (92.3 to 93.6%) belongs to the core genome, as they are shared between all the eleven genomes. A small proportion of all genes (1.6%) are classified in the cloud genome.

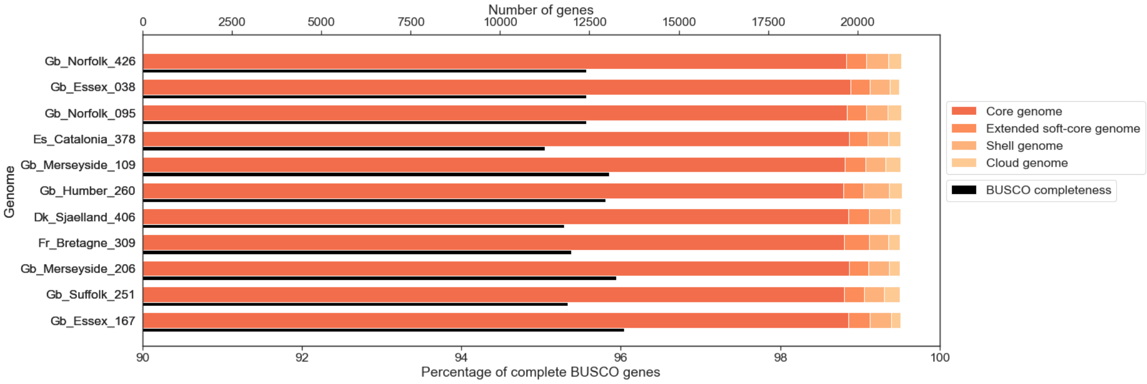


Figure 13 – Orthogroup sharing, across the eleven sea beet genomes.

Annotated genes in each genome are classified in core/soft-core/shell/cloud genome as whether they belong to orthogroups shared by all/all but one/two to nine/or present in a single genome. Black lines represent the percentage of completeness of each genome assembly, or complete (C) BUSCO genes.

In addition to observing gene presence/absence across the assemblies, the maintenance of the order of these genes in the genomes was investigated. The plot demonstrates synteny is largely maintained among genomes. However, it is not possible to say at this stage whether re-arrangements are biological or miss-assemblies (**Figure 14**).

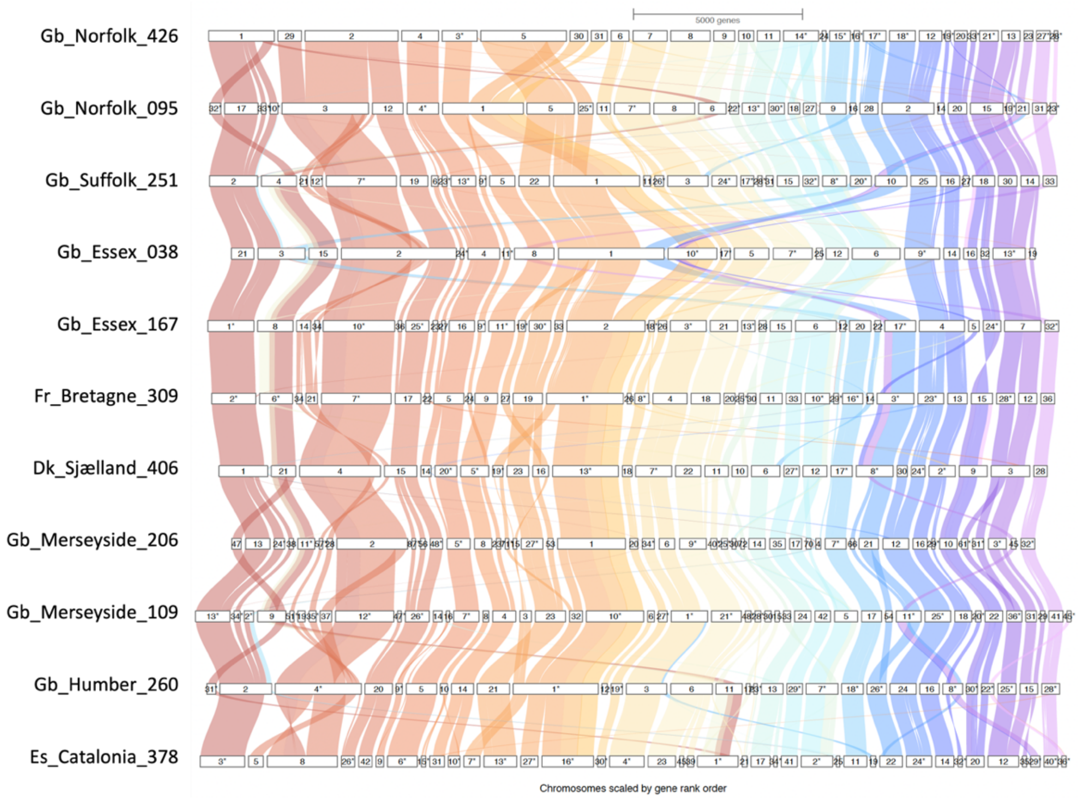


Figure 14 - Sea beet pan-genome synteny plot generated with GeneSpace.

One way to assess if a pan-genome is closed or open, i.e. if the complete diversity in terms of genes across a species is represented in the pan-genome or not, is to generate a saturation curve. Such a plot represents the relationship between the number of genes present in the pan-genome and the number of genomes investigated. The saturation (closure of the pan-genome) is reached when the curve reaches a plateau: the addition of new genomes into the study doesn't bring novelty in the gene collection. In the present study, the analysis of the presence/absence of orthogroups in the different genomes (**Figure 15**) shows a tendency to reach a plateau from a number of 10 genomes (pan genes). When considering orthogroups shared with every genome (core genes), the addition of extra genomes in the study reduces their number, as they may be absent from the novel orthogroup set. Here, the orthogroup analysis shows that 11 genomes allow the observation of a large number of pan genes (23,631) but isn't enough to constitute a closed sea beet pan-genome.

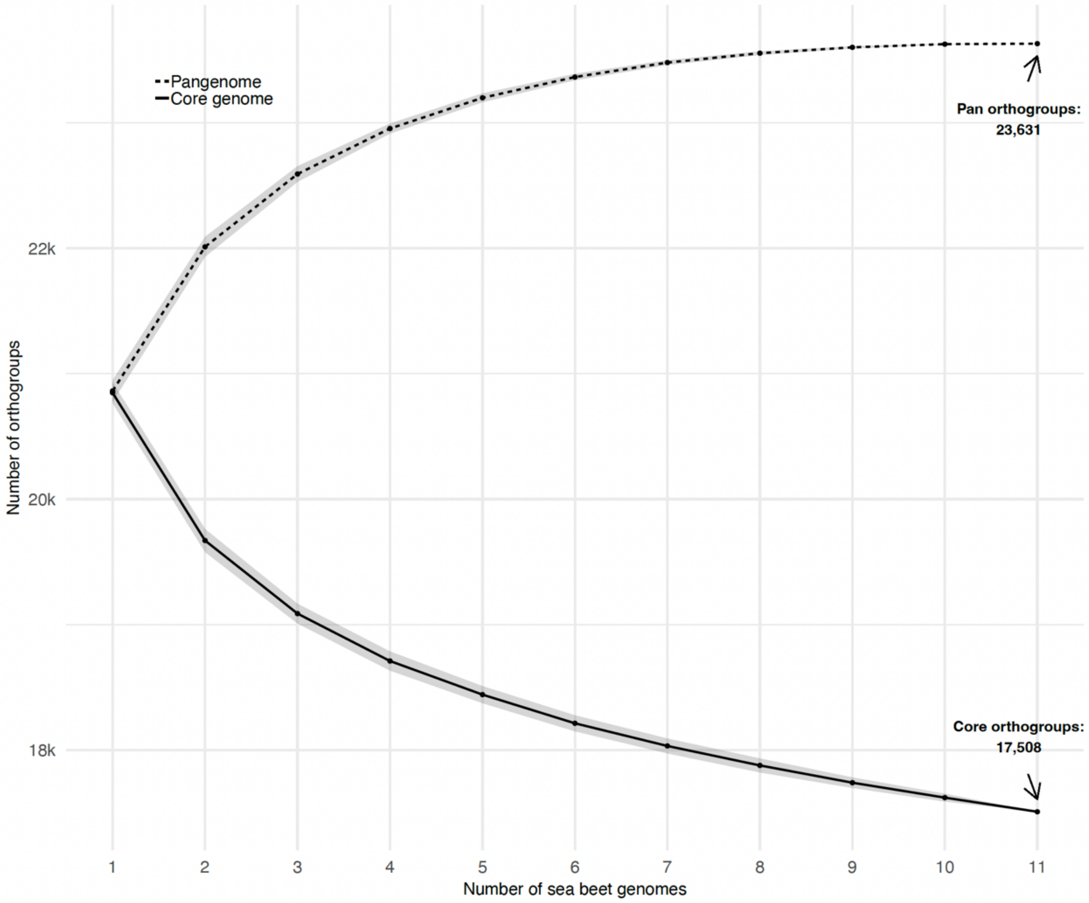


Figure 15 - Sea beet pan-genome gene saturation curve.

The number of orthogroups part of either the sea beet core genome or the pangenome is displayed in relation with the number of genomes studied.

II - B. 4. The sea beet pan-genome content is variable in terms of resistance genes

Gene presence/absence between genomes may be more restricted to specific gene types. In order to compare the NLR content between genomes, the NLR loci have been extracted with NLR-Annotator[117]. NLRs are the most abundant class of resistance genes in plants. They are important as they provide specific resistance through the recognition of pathogen effectors and trigger an immune response that could lead to a hypersensitive cell death, thus preventing the spread of the pathogen[29]. Also, because of their importance, they may be particularly prone to duplication and loss[122]. This often makes their number highly variable among genomes[123]. Interestingly, the number of NLR loci doesn't differ to a great extent between sea and sugar beet genome assemblies (**Table 1**).

The NLR loci were clustered in orthogroups with OrthoFinder[124] (**Table S3**), and the distribution of these orthogroups between genomes was studied (**Figure 16**). The core NLRome represents 60.9 to 67.5% of the total NLRome; the soft-core ranges from 71.5 to 79.5%; the shell NLRome ranges from 20.5 to 28.5% and, finally, only 4 genomes have a cloud NLRome (Gb_Essex_167=0.5%, Gb_Humber_260=1%, Gb_Merseyside_109=0.5% and Gb_Essex_038=0.5%).

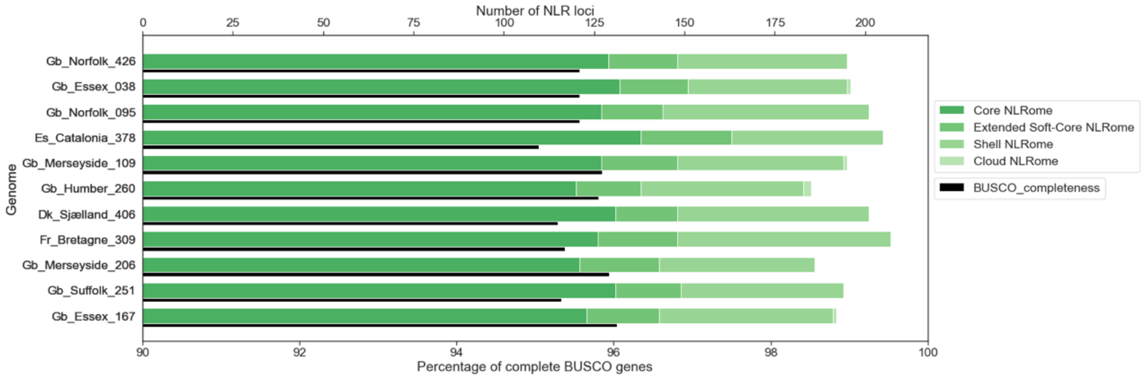


Figure 16 - Sea beet pan-genome NLR loci content.

Core NLRome: NLRs belonging to an orthogroup present in every genome, extended soft-core NLRome: NLRs belonging to an orthogroup present in 10 genomes, shell NLRome: NLRs belonging to an orthogroup present in 2 to 9 genomes, cloud NLRome: NLRs belonging to an orthogroup present in a single genome. Black lines represent the percentage of completeness of each genome assembly, or complete (C) BUSCO genes.

Differences in the number of NLR loci per individual could be both biological and/or methodological. To investigate the impact of the assembly quality on the number of NLR loci observed, the relationship between the number of NLRs and the completeness of the assemblies was assessed (**Figure 17, panel A**). The number of NLR loci identified in a genome does not increase with the completeness of its assembly. On the contrary, more complete genomes tend to show fewer NLR loci. However, this is observed in a very short range of completeness values (95.1% - 96.1%), and it would be thus overinterpreting that qualifying this result as a true negative correlation. In addition, the size of the assembly does not reflect its quality, in terms of completeness (**Figure 17, panel B**). There is no correlation between the genome assembly size and the number of NLR loci (**Figure 17, panel C**). The number of NLR genes is expected to vary between genomes in a same species, due to their rapid evolution leading to gene duplication and loss. These observations lead to the conclusion that the assembly quality impacts neither the number of NLRs retrieved nor the size of the genome assembly and therefore, suggests that differences in the numbers of NLRs between individuals is biological.

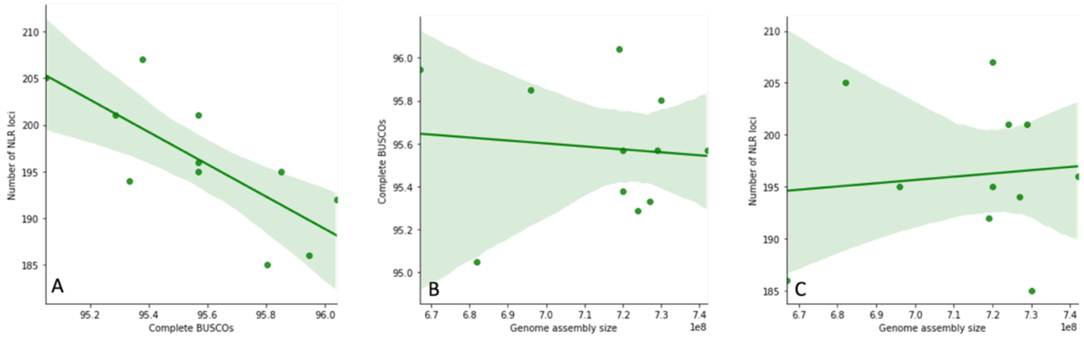


Figure 17 - The number of NLR loci is not correlated to the quality of the genome assembly.

A – Number of NLR loci in a genome plotted against the number of complete BUSCO genes. B – Number of complete BUSCO genes in a genome plotted against the size of the assembly. C – Number of NLR loci in a genome plotted against the size of the genome assembly.

Overall, the total number of NLR loci per genome is close to 200, ranging from 186 to 207 (**Table S4**). Differences in the number of loci between the core NLRomes correspond to NLRs belonging to a same orthogroup in a same genome, or duplicated NLRs. The Spanish genome (Es_Catalonia_378) has the largest core NLRome (largest number of duplicated NLR genes). However, it also has the smallest shell NLRome. This is perhaps notable given that this is the only Mediterranean genome.

There are very few NLRs found uniquely to a single individual (cloud NLR). Interestingly, they are found in sea beets which are not from geographically isolated locations to the rest of the individuals sequenced, as, for example, Gb_Essex_038 and Gb_Essex_167 both are from the Essex county.

When looking at the composition of the NLRome of each genome, in terms of gene classes, approximately half of the NLR loci extracted by NLR-Annotator corresponds to full-length genes coding proteins carrying the three characteristic NLR domains: either a coiled-coil (CC) or Toll/Interleukin 1 Receptor (TIR) N-terminal domain, a central NB-ARC domain, and a C-terminal leucine-rich repeat (LRR) domain (**Figure 18**). Among these full-length NLR loci, the large majority corresponds to CC-NBARC-LRR genes. However, TIR-NBARC-LRR are found in each of these genomes, from a single one, to five genes in the Gb_Essex_038 genome. Overall, the proportion of the different classes is the same between genomes, with approximately a third of the annotated NLR loci only consisting of the central NB-ARC and the terminal LRR domains, a small number of CC-NBARC loci, and approximately an eighth of the loci corresponding to NB-ARC domains alone. Intriguingly, the genomes Dk_Sjælland_406 and Es_Catalonia_378 genomes, show the presence of NLR loci harbouring an association of two N-termini domains. Indeed, these genomes show, respectively, the presence of a CC-TIR-NBARC-LRR locus, and a CC-TIR-NBARC-LRR and a CC-TIR-NBARC loci.

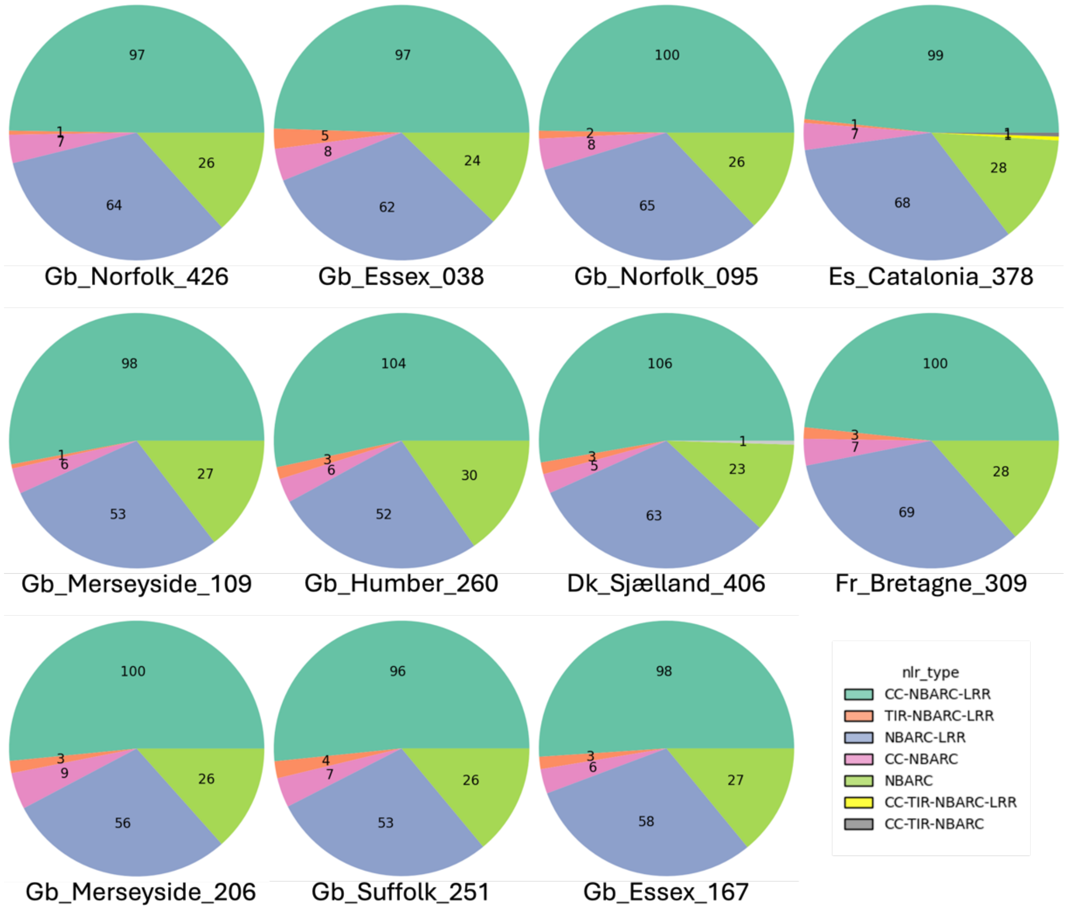


Figure 18 - Distribution of the NLR loci extracted in the eleven genomes into NLR classes.

The NLR loci extracted with NLR-Annotator are categorised according to which protein domain(s) they are encoding.

II - C. Discussion

II - C. 1. Assembly metrics allow quality assessment, but also highlight biological phenomena

The eleven sea beet genomes assembled in the present work are all of better quality compared to the published sea beet assembly, Bmar[119]. These assemblies are of higher contiguity and completeness, as evidenced by a lower number of contigs, a lower L50, a larger N50, and a higher percentage of BUSCOs. The percentage of BUSCO is a suitable measure of the completeness of the assemblies generated, as BUSCO can search for single copy orthologous genes without relying on an annotation. This way, the quality of the gene annotation doesn't impact this statistic.

These improvements are perhaps not surprising given the difference in the sequencing technology used. PacBio HiFi sequencing provided significantly longer reads than the Illumina technology used to generate the Bmar assembly. However, if compared to the reference sugar beet genome, the sea beet assemblies are of lower contiguity.

Genome purity assessment highlighted several contigs with a higher GC content. These high GC contigs were found in all 11 sea beet assemblies at the expected target depth and taxonomically identified as *Silene littorea*. *S. littorea* is a plant found around the Mediterranean sea and north Africa and also belongs within the Caryophyllales order. Salt tolerant and small, this species grows in sandy regions but is easily distinguishable from sea beet. This plant is not parasitic to beet and was not grown in the lab. The fact that these contigs are found in every sample suggests that blasting to this species could be due to a lack of representation of *Beta vulgaris maritima* in the NCBI databases and to the proximity between *S. littorea* and *B. maritima* genomes. To further investigate the nature of these patterns, it would be interesting to explore the published beet genomes and their sequencing reads to look for evidence of the presence of these *S. littorea* contigs.

While most of the orthogroups studied in this work appear to show a conserved synteny across the eleven genomes, some structural rearrangements are observed. Technical rearrangements can arise due to the heterozygous nature of the genome assembled. Indeed, structural differences can exist between homologous chromosomes. During the assembling process, the assembler can mix reads from these different haplotypes, resulting in a configuration which doesn't exist in either of the two haplotypes. Here, the nature of the rearrangements hasn't been investigated. However, discriminating between biological or technical phenomena could be done through a PCR (polymerase chain reaction) strategy: amplifying and analysing the segment where the rearrangement seems to happen would allow to know if two genes are indeed in proximity or if this configuration results from an error generated during the genome assembly process.

Given the differences in size and gene content between the sea beet genomes, it is perhaps important to begin to consider them within a pan-genomic context, using graphs[125]. However, in the present work, I will be using Gb_Norfolk_426 as the focal assembly, because it has the highest L50 and N50, and shows generally good results for the other quality assessment measures.

II - C. 2. Sea beet has a large and conserved core genome

Analysing published beet genomes, the number of genes annotated is on the order of 20-30 thousand (Bmar[119]: 27,662; EL10.1[121]: 24,255; EL10.2_2[121]: 21,587; RefBeet[14]: 27,421). The eleven sea beet genomes assembled here don't deviate from these observations, with a median number of annotated genes of 21,150 (**Table S5**). All genomes combined, this represents a total of 233,208 genes, distributed in 19,454 orthogroups, 91.2% of which are found in every genome. This highlights how the wild sea beet genome is conserved across diverse European locations, and this proportion of core genome is larger than for other pan-genomes, which show for example a core genome of 74.2% for the tomato[126], 70% for *Arabidopsis thaliana*[127], 64.3% for the hexaploid wheat[128], and 53.5% for the rice[129]. These numbers make the sea beet core genome appear surprisingly large, and this result could be explained by the pan-genome generated here only involving individuals belonging to the same subspecies, unlike the tomato study, which involved both wild and domesticated plants. Moreover, these results must be taken with caution, as firstly, only one representative of the sea beet Mediterranean group is analysed here, and the present pan-genome doesn't therefore truly cover the wild sea beet diversity, and secondly, these estimates are based on a lifted annotation.

It is a caveat of the present work that genomes were not de novo annotated and this will have reduced the potential for novel genes to be identified in each genome. However, these results are unlikely to be due to a too small number of genomes, as both the *A. thaliana* and the wheat pan-genomes only involve 19 genomes. Moreover, the exceptionally large core genome is probably not due to the lack of diversity among the sequenced individuals which are from different European locations. Using the sugar beet reference genome to lift over gene annotations to sea beet genome assemblies is an advantageous methodology for a time-efficient annotation, which was required in this work. However, the accuracy of these sea beet annotations is then not optimal as genes can be missing due to structural divergence or also because of the reliance on a single reference genome. Moreover, the estimation of the number of sea beet genomes needed to close the pan-genome is underestimated due to these missing genes. A better way would have been to use multiple available beet RNA-seq data to map to the new sea beet assemblies and perform a *de novo* transcriptome assembly. Then, these data could have been combined with the lift-over annotations, and this would have improved the annotations by by-passing the single-reference as well as the sub-species differences biases. Therefore, the present work can be taken as a first step towards the generation of eleven high-quality sea beet genome annotations.

II - C. 3. Variation in the resistance genes content is a mark of genomic adaptation to face constant changes in pathogen threats

The majority of the genes annotated in the eleven genomes, i.e. 92.9% (**Table S2**), are classified in orthogroups present in every genome. Also, 95.6% of these genes belong to orthogroups in the soft-core genome. This shows the level of conservation and relative contrast with the NLR loci, for which only 65.1% are part of the core NLRome, and 75.2% are part of the soft-core NLRome (**Table S4**). This clearly shows how NLR genes are much more variable between genomes compared to the total set of genes. The way these NLR genes are more different can also be observed through the proportion of these genes being part of the shell NLRome, i.e. 24.5% (**Table S4**), against only 3% when looking at the rest of the genes.

The totality of the NLR loci extracted from the eleven genomes, i.e. 2,157 loci, are classified into 215 orthogroups (**Table S3**). This is half the number of orthogroups identified for the *Arabidopsis thaliana* NLRome[92]. The *A. thaliana* NLRome is largely saturated, with 65 genomes involved, and a saturation estimated at approximately 40 genomes. The sea beet NLRome being still open, the size of its NRLome can't be compared with other species.

It is known that the proportion of NLR genes among the total number of annotated genes in a genome can vary considerably[123], ranging from 0.003% in bladderwort[130] to 2% in apple[131]. In the present study, the proportion of NLR loci represents on average 0.93% of the total number of annotated genes. This number is higher than that of the kiwifruit (0.256%[132]) and of *A. thaliana* (~0.5%[133]). This variation of the number of NLR loci across genomes has been explained by the fact that they are rapidly evolving through a high mutation rate coupled with positive selection[134], gene duplication[135], gene conversion[136] and unequal crossing-over[137].

Approximately half of the NLR loci extracted from the sea beet genomes are classified as CNLs, i.e. full NLR sequences, encoding the three characteristic domains, including a coiled-coil N-terminal domain. This represents 95 to 99% of the full-length NLR loci, and is consistent with the previous beet genomes observations having identified CNLs being the most prevalent class of NLRs. It was initially thought that the sugar beet genome wasn't containing TNLs[138]. Through functional gene annotation, Dohm *et al.* found for the first time a single TNL both in the sugar beet and in the spinach genomes[14]. They suggested that the presence of a single TNL was a feature of species belonging to the Amaranthaceae family, in contrast with the rosids and asterids clades which both show an expansion of this class of gene. The study of the NLR signatures in the reference sugar beet genome EL10 also identified a single TNL[139]. Despite not using RNA-seq in the annotation process, the present study identified multiple TNL genes. This suggests that Amaranthaceae could in fact contain multiple TNLs. The comparison between the sea beet published genome, the sea beet genomes generated here, and the published sugar beet genomes, in terms of NLR loci extracted with NLR-Annotator, leads as well to the suggestion that sea beet could harbour more TNLs than its cultivated counterparts. The Gb_Essex_038 genome harbours the largest number of TNLs (5), with twice the average of TNLs encountered among the eleven sea beet genomes.

Among the so far experimentally validated plant NLRs[140], the presence of both N-terminal CC and TIR domains is not observed. Thus, the CC-TIR NLR loci identified in the Es_Catalonia_378 and Dk_Sjælland_406 genomes are most likely the result of an erroneous annotation by NLR-Annotator.

In *Arabidopsis thaliana*, NLR genes have been observed to vary from a number of 167 to 251 across genomes[141]. According to this observation, the variation here is lower than expected, but this could be due to the fact that the *A. thaliana* accessions had a provenance covering a much wider distance.

French and Danish genomes have larger shell NLRomes than the English genomes, which is consistent with the expectation that these mainland European plants are outliers with respect to the 9 English plants sequenced. This suggests that NLR gene presence may have a regional component and it would be interesting to investigate the roles of these genes with respect to local pathogens (see **Chapter III**).

It is very interesting to observe a similar number of NLR loci predicted either on the sea and on the sugar beet genome assemblies. The bottleneck sugar beet has been through has decreased its genetic diversity including its diversity in resistance genes. Moreover, adaptation of these resistance genes to effectors released from pathogens encountered in fields is impeded by artificial selection. It could be expected that due to these restrictions, the NLR set among sugar beet genomes is fixed in its number of genes, and it is not the case in wild beet genomes which can undergo the process of NLRs duplications. However, the results observed here refute this hypothesis and highlight the cost of resistance genes and the underlying process of gene death[48]. Nevertheless, crop and wild NLR sets can differ in the way that the diversity could be lower in domesticated plants.

To conclude, the eleven sea beet genomes assembled and the NLR genes extracted from them constitute a great addition to the material available for the Beta research community and especially for the beet breeding community. In the present work, these data, combined with the re-sequencing of hundreds of wild plants, allow subsequent association genetics analyses aiming for the identification of agronomically important resistance genes. Future association genetics works would greatly benefit from a clear representation of genes presence/absence across the *Beta vulgaris maritima* species through the combination of the generation of additional sea beet genome assemblies as well as the creation of a sea beet graphical pan-genome.

II - D. Material and methods

Plant selection for the pan-genome generation

Plants sequenced with the HiFi technology were selected among the 589 sea beets grown in a temperature, light, and humidity-controlled environment, and to which a Danish rust sample was inoculated (see **Chapter III**). The criteria for selection were a combination of the best rust resistance with distinguishable agromorphological traits, and the geographic origin of the plants was also considered, assuring to have a panel of plants sampled from different locations, at country and European levels (**Table S6**). The agromorphological phenotypes recorded included the presence/absence of trichomes, the redness of different organs of the plants (leaves, petioles, leaf edges, veins), the leaf waxiness, the presence/absence of a stalk and the fact that the plant bolted or not during the course of the experiment. On the resistance side, apart from the beets from the Humberside population which all had signs of rust infection, all the plants selected for genome sequencing were resistant to the Danish rust, i.e. didn't develop symptoms of infection. Moreover, the rust resistance phenotype of the sisters (plants from the same mother, part of a different rust inoculation trial (see **Chapter III**)) was also considered, to include in the pan-genome plants with a good probability of carrying resistance genes, as this would be useful for a rust association analysis (see **Chapter III**).

Genomic DNA extraction and sequencing

Genomic DNA was extracted from the sea beet leaves using the Macherey-Nagel NucleoBond DNA/RNA/Protein kit according to the manufacturer's instructions. One modification was made to the protocol which included an overnight lysis on a mixer HC; StarLabs Smart Instrument set at 50°C and 450 rpm.

Genome sequencing was performed with the Single-Molecule Real-Time (SMRT) method on the Sequel II HiFi PacBio platform, by the Technical Genomics from the Earlham Institute for the genome Gb_Essex_038, and by Novogene for the ten other genomes. The sequencing depth is estimated between 19.9x and 47.7x (**Table S1**), with a mean of 35.6x. The samples Gb_Norfolk_095, Es_Catalonia_378, Gb_Suffolk_251 and Gb_Merseyside_206 were sequenced in two runs, and raw sequencing data were combined before data processing. Sequencing data has been submitted to the Sequence Read Archive (SRA) database from the National Center for Biotechnology Information (NCBI) under the BioProject PRJNA1209597.

Genome assembly and quality assessment

Genomes were de novo assembled using the hifiasm software[142], version 0.16.1, with the default parameters. The subsequent analyses were done on the primary contig file. The contigs have been ordered per size, from the largest one to the shortest one.

Contamination among the eleven assemblies was investigated with the default parameters of BlobTools[112] version 1.0.1, involving a Blast[113] (version 2.10) step against the NCBI database and a mapping and alignment step involving the programs minimap2[143] version 2.21 with the parameters -a and -x map-hifi, and samtools[144] version 1.13. Subsequently to running the BlobTools program, contigs corresponding to potential contaminants were removed. This translates in the removal of any contig which was not associated to the Caryophyllales order, when the proportion of the contamination in the combined mapped and unmapped reads was >= 1%.

k-mer content in the sequencing reads and in the respective assemblies were compared with the *K*-mer Analysis Toolkit[118] version 2.3.4, with default parameters.

Assembly metrics were obtained with the default parameters of the Abyss-fac tool from the Abyss[145] program, version 2.0.2. Assembly quality and completeness were measured with the Merqury program[120], version 1.3, with a *k* length of 21. When contigs were removed from the assembly due to contamination, the reads mapping to these contigs were removed from the Merqury analysis. Completeness of the assemblies generated, using single copy eudicotyledon annotated orthologs, was assessed with the Benchmarking Universal Single Copy Orthologs program[146], version 3.0, with the parameters --lineage eudicotyledons_odb10 -m geno -sp Arabidopsis.

Genomic data from published sugar beet and sea beet genomes were downloaded from the NCBI database; as the RefSeq GCF_000511025.2 assembly for RefBeet-1.2.2, the GenBank

GCA_002917755.1 assembly for EL10_1.0 and the EL10.2 assembly was downloaded from the Phytozome portal.

Genome annotation, orthology and synteny analyses

Genomic distances between the 11 genomes were measured and a tree was constructed using the mashtree[147] program (version 1.4.6). The spinach reference genome assembly BTI_SOV_V1 was downloaded from the NCBI database, as the GCF_020520425.1 RefSeq assembly. The mashtree_bootstrap.pl script was run with the flags --reps 1000 --min-depth 2 --kmerlength 21 --sketch-size 10000, to generated bootstrap supporting values from 1,000 relicates. The tree was rooted on the spinach genome using ape[148] (version 5.8) with the parameters resolve.root = TRUE and edgelabel = TRUE. The tree was visualised in FigTree (version 1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/)).

NLR-Annotator version 2.1[117] was used to annotate the NLR loci in the genome assemblies. Protein motifs from a set of potato NLRs were used to parse the translated assemblies and predict the presence of NLR loci.

Liftoff[149] (version 1.6.3) was used to annotate the genomes, transferring the annotations from the EL10.2_2 sugar beet assembly, with the options -a 1, -s 0.75, -polish.

OrthoFinder[150] (version 2.5.4) was used to classify NLR genes as well as the whole gene content into orthogroups, with the default parameters. The presence/absence matrix of these orthogroups among the eleven sea beet genomes was used to compute a gene saturation curve, using the panplots function in the plot_gene_content.R script from Dr. Rowena Hill, available on GitHub (https://github.com/Rowena-h/GaeumannomycesGenomics/blob/dc1c8f45fc835a43d31c4c7670794aea8fb532aa/07_comparative_genomics/plot_gene_content).

The synteny plot was generated with GeneSpace[151], using, as input files, genome annotations and protein sequences generated by Liftoff from the sugar beet EL10.2_2 genome assembly.

All computational experiments were carried out on a high-performance computing (HPC) cluster running AlmaLinux 9.5, utilising the x86_64 architecture. The cluster utilises SLURM

(Simple Linux Utility for Resource Management) to manage job scheduling and resource allocation, is equipped with a processor operating at a speed of 1.5 GHz and is provisioned with 503 GiB of RAM. The system supports 64 CPU cores.

Chapter III - Mining wild sea beet genetic diversity for rust resistance

III - A. Introduction

III - A. 1. Association genetics to identify resistance against plant pathogens

Association genetic studies attempt to statistically associate genetic polymorphism, usually SNPs, to a phenotype of interest. They involve the study of a panel of genetically diverse individuals harbouring differences in this phenotype. The association is said to be direct when the SNP identified is the true causative element for the phenotype, and indirect when the SNP identified segregates with the causative sequence due to the phenomenon of linkage disequilibrium (LD)[152].

Association genetics can either be conducted on candidate genes, which can correspond to a well-characterised gene family, for example, believed to be involved in the observed phenotype, or they can scan the entire genome, and are then called GWAS for Genome-Wide Association Studies[153]. Most of the association genetic studies conducted on crops focus on candidate genes[153], such as resistance genes.

Plant association genetic studies are mostly conducted on lines or crops. This comes with advantages such as the fact that they typically have low levels of heterozygosity. The power of the association study is stronger when considering homozygous individuals because it is not reduced by heterozygous individuals that carry both dominant and recessive alleles. However, working with agricultural material also comes with multiple disadvantages. Indeed, due to the domestication and the selection for agronomically important traits, crops are genetically bottlenecked and their level of inbreeding is high. On the one hand, this can benefit the study, as their genetic identity (fewer SNPs) increases the power to identify polymorphism associated with a phenotype. On the other hand, this also means that the genetic diversity present in crops is low, and therefore, there is less opportunity to identify interesting polymorphisms/traits as these have been lost from the breeding programs. Furthermore, precisely identifying causative variants may be difficult because of the high level of linkage disequilibrium found in crops. Indeed, large LD blocks may mask the impact of important polymorphisms and will reduce the potential to narrow down causative polymorphisms.

GWAS has been conducted on wild populations (e.g. black cottonwood[154], trout[155]), but this comes with multiple challenges borne out of much higher levels of diversity/heterozygosity, such as a lack of reproducibility, population structure, and the potential absence of a reference genome for a non-cultivated species[68]. Physiological differences are also important. For example, crops have been bred to be physiologically similar, to aid in harvesting and production purposes (e.g. short height of winter wheat[156]). These similarities don't exist to the same extent in wild plants (or wild crop relatives) and these differences can confound phenotyping efforts where traits are difficult to differentiate.

Wild-performed GWAS can lack replicability firstly because of sampling variations between studies, but also because allele frequencies are unmanipulated in natural populations, i.e. they differ from equilibrated frequencies formed by crosses between lines which make association studies more powerful[157]. It has been observed, in QTL mapping analyses, that replication was more successful within populations than across populations[74]. Non-replicability can be triggered by biological differences between populations such as differences in linkage patterns, differences in allele frequencies or context-dependency of the studied trait.

Population structure can have a significant impact on association studies, and can lead to the appearance of false-positive results[153], and to erroneous conclusions[68]. Individuals that belong to a same kinship-related subgroup within the sampled population can cause false positive associations, especially where the phenotype is predominantly present in a related subgroup in the analysis; then, any SNP enriched in this subgroup will show an association with the trait of interest. Controlling for population structure in an association study is thus of major importance.

Most of plant genome assemblies belong to crops, due to their importance in breeding programs[158]. Therefore, it is rarer for wild plants or crop relatives to have a reference genome, and this can hamper reference-based GWAS on these species. On the other hand, even when a reference genome is available, the trait of interest, such as resistance, can significantly differ in the panel of individuals studied, compared to this reference. The level of structural, and more local gene presence absence polymorphism in plants are only just beginning to be understood[159,160]. This is due in part to the prevalence of short read sequencing technology, which is unable to accurately assemble repetitive regions of the genome[160]. This may be particularly important for identifying loci associated with increased resistance as these genes have a high degree of presence absence and duplicated polymorphism and are difficult to assemble. Furthermore, even if a high contiguity reference is called, these resistance genes are likely to have a very different presence absence pattern in another individual[161].

Researchers have attempted to circumvent the costs associated with sequencing multiple references (and associated resequencing) by identifying SNPs that describe diversity in a panel of plants. However, these methods require a significant amount of understanding of the diversity of the focal species that is often not available for a non-crop. These methods are also open to ascertainment bias[162], which is more likely to be a problem in wild species.

Overreliance on the reference can be bypassed by the use of *k*-mers as a manner to measure difference between individuals. This way, instead of comparing every subject to a single reference genome to identify SNPs, every individual is compared to each other through *k*-mer presence/absence matrices, and this avoids any reference-lead polymorphism omissions, increases the chance of identifying significant associations and the power of the association analysis[163]. An additional advantage of *k*-mers is their ability to account for structural polymorphisms and insertions/deletions additionally to SNPs, as well as haplotypic information[163]. Indeed, if two sequence polymorphisms are close enough to appear on a same *k*-mer, it is then possible to get information about phenotypic variation lead by an haplotypic configuration of the variants.

The first *k*-mer-based association analysis in plants was conducted by Arora *et al.*[164] in 2019, and was called AgRenSeq, for Association genetics and Resistance gene enrichment sequencing. The *Aegilops tauschii* species, a wild relative having contributed to the D genome of bread wheat, was screened for resistance against the wheat stem rust *Puccinia graminis*.

A panel a 151 genetically diverse accessions was phenotyped for their level of resistance against six races of the rust pathogen. This work allowed the validation of the method through the detection of two previously cloned genes, acting as positive controls, and two additional candidate resistance genes.

The AgRenSeq method focuses on the most prevalent class of resistance genes in plants, the NLR genes. To do this, the method uses the RenSeq (Resistance gene enrichment Sequencing) methodology, which consists of the targeted sequencing of NLR sequences using probes retrieving sequences from as low as 80% of homology[165]. This enables deep sequencing of these resistance loci with a reduced cost. 51-mers are extracted from the sequencing reads and kept in the analysis when their correlation with rust resistance is positive. A linear regression model allows the generation of the *P* value reflecting the strength of the association, and the identification of candidate resistance genes is done through mapping the associated *k*-mers to the NLR sequences extracted from resistant accessions.

In 2022, the AgRenSeq original study was extended with whole-genome shotgun sequencing instead of resistance genes targeted sequencing, and the significantly associated *k*-mers were mapped against the reference genome of *Aegilops tauschii* and a *de novo* assembly of an accession anchored to the reference genome[166]. This whole-genome *k*-mer-based association approach was proven efficient to identify a candidate rust resistance gene discovered through AgRenSeq, as well as additional candidate genes responsible for various quantitative agronomically-interesting phenotypes such as the number of trichomes, the number of spikelets per spike or the resistance to the powdery mildew pathogen. This methodology allows firstly the inclusion of multiple phenotypes with an agronomic value within the same experiment, and, secondly, an opening to the possibility that the resistance isn't provided by an NLR gene, as it has been shown to be the case in multiple examples involving tandem protein kinases[167].

Although corresponding to a wild wheat-related species, the panel constituted by Arora *et al.* was formed by seeds from accessions obtained from multiple germplasm libraries and self-fertilised in glasshouses. This selfing step induced a decreased level of heterozygosity compared to individuals which would have otherwise been sampled in natural populations. As mentioned above, increasing homozygosity allows an improvement of the power of association but also increases linkage.

The present study aims to identify rust resistance genes in wild sea beets using a *k*-mer association method. It differs from Arora *et al.*'s study in the sense that the sea beet panel was formed by seeds directly collected from natural populations, and were not subject to any human intervention ahead of the inoculation trials. This choice of methodology is first explained by time and cost constraints, as well as a biological phenomenon being that the *Beta vulgaris* species is largely self-incompatible[168], making reducing heterozygosity more difficult to achieve. Finally, the aim of the present study, broader than the identification of resistance genes, is to test the validity of a *k*-mer-based association genetics method directly in natural populations. There is a level of experimental replication in the current work, by the inclusion of "sisters", when seeds sampled from a same mother plant are part of the same inoculation trial. Therefore, is then essential to control for population structure; unlike the KASP (Kompetitive Allele Specific PCR) markers used in Arora's method, this was achieved through a PCA (Principal Component Analysis) using a presence/absence matrix of 315 randomly selected *k*-mers from hundreds of re-sequenced individuals.

Focussing on the sequencing of NLR genes has the advantage of identifying an association with a single gene instead of a genomic region containing multiple paralogs. However, this comes as well with downsides, as mentioned above, being the limitation of the study to a single biological phenomenon and to a single gene family. In the present work, benefits from both methodologies are combined as whole genome shotgun sequencing is performed on a panel of rust inoculated and scored plants, and the utilisation of several good quality *de novo* genome assemblies of *Beta vulgaris maritima* (see **Chapter II**) allows mapping of significantly associated *k*-mers both on whole-genome sequences and specifically onto the NLR loci they contain.

Sea beet is a coastal species spread across European shorelines which shares genetic material between populations through airborne pollen and short-distance airborne or waterborne seeds dispersal. A pollen dispersal experiment involving weed beets measured a maximum distance of pollination of 9.6 km[169]. On the other hand, seeds can survive in water for several days and thus travel long distances[170]. Moreover, in a study involving 23 sites sampled along the northwest French coast, it has been shown that marine currents had a considerable impact on shaping the genetic diversity of wild sea beet populations[170].

Sampling a panel of genetically diverse plants is essential to give an association genetic study the best chances to succeed. In this optic, the aim of the present study was to screen for rust resistance diverse populations of English sea beets, showing differences in their resistance genes composition, both considering presence/absence as well as allelic polymorphisms. The material was collected from ten locations along the eastern coast, separated with a distance from 9.75 to 109 km, exceeding the 9.6 km precedingly reported as the distance threshold for pollination. Moreover, the addition to the seed panel of individuals collected in four sites on the west coast of England, geographically isolated from the east coast sites both considering pollen and seed dispersion, was thought to maximise the potential to sample across genetic groups.

Recently, Sandell *et al.*[10] used a whole-genome *k*-mer-based approach to observe genomic relationships between a diverse panel of 606 beets, including accessions of wild and cultivated beets as well as commercialised sugar beet lines, in order to get insight about the history of beet domestication. They highlighted a split between sea beets from Mediterranean and Atlantic coasts and showed that cultivated beets were genetically closer to the Mediterranean cluster and had most likely been domesticated from Greek populations. These results are particularly important considering beet cultivation, as they point out a paradox: cultivated beets derive from wild beets having evolved in the Mediterranean climate but are nowadays mostly cultivated in northern countries with cooler and wetter climatic conditions, the major producers in Europe being Germany and France (FAOSTAT). Moreover, beet domestication is a very recent event when comparing with other crop domestication history, and these differences between western European and Mediterranean climates may come with a lack of adaptation of cultivated beets to diseases local to cultivation sites, and an absence of the corresponding resistance genes in their gene pool.

These observations highlight how well-suited the sampled seed material is for the present work. Firstly, it emphasises the interest of mining wild Atlantic sea beets for agronomically important resistance genes that would provide resistance to fungi encountered in north western European countries. Indeed, Atlantic sea beets constitute the major part of the present panel. Secondly, it reveals two sea beet genetic groups, which are both represented in the panel, as a small number of Mediterranean Spanish individuals are included in the association genetic study.

Another factor needed for the success of an association study is the accuracy of the phenotype recorded. As an example, a specific scoring system to measure susceptibility to the wheat stem rust disease mentioned earlier exists, developed by Stakman *et al.*[171], which accounts for the level of sporulation, the size of the lesion and the presence of chlorosis and/or necrosis. Such a phenotyping scale would be more difficult to apply to rust infected beets, as lesion size and chlorosis or necrosis are less clearly observable than in the case of the wheat stem rust. Instead, recording the number of rust pustules, as a strategy previously adopted by Kristoffersen *et al.*[26], is a more appropriate way to score beet susceptibility.

Working in an agricultural context, an optimal phenotype would both measure the susceptibility to the pathogen as well as take into account the impact the fungus has on the crop yield. However, measuring sugar yield in beets after a rust inoculation trial would involve multiple time-consuming steps and wouldn't be feasible when considering experiments including several hundreds of plants. Moreover, even if sea and sugar beets have a relatively recent common ancestor and can interbreed, they can't be put on the same level when considering sugar yield, as sugar beets have been bred to develop optimal sugar production pathways that wouldn't be present in their wild relative.

III - A. 2. A large-scale rust inoculation experiment to mine English sea beet genetic diversity

With an aim to take advantage of the genetic diversity harboured by sugar beet's wild ancestor, the sea beet, an association analysis adapted from the AgRenSeq method[164] was carried out for resistance against local and non-local rusts. In more details, three large-scale rust inoculation trials were conducted on the Norwich Research Park from spring 2021 to spring 2022, each involving approximately six hundred wild sea beets. These wild individuals were sampled as seeds across east and west English coasts and a small number of mainland European regions (**Figure 19**), and grown on site under controlled conditions. In total, 20 sites were sampled: 10 sites from the east coast of England in the counties of Humberside, Norfolk, Suffolk and Essex; 4 sites from the west coast of England in the counties of Merseyside and Cheshire; 2 French sites in Brittany; 2 Spanish sites on the Mediterranean coast and 2 Danish sites in the Zealand island.

On the pathogen side, three beet rust samples were collected in three locations and utilised in three distinct inoculation trials. Two of them were provided by the BBRO, sampled from infected sugar beet fields either in Norfolk (hereafter referred as the N rust) or in the Lincolnshire (L rust), both from the eastern side of England. Additionally, a rust sample from another country, Denmark (D rust), was provided through collaboration (Thies Marten Heick, Aarhus University).

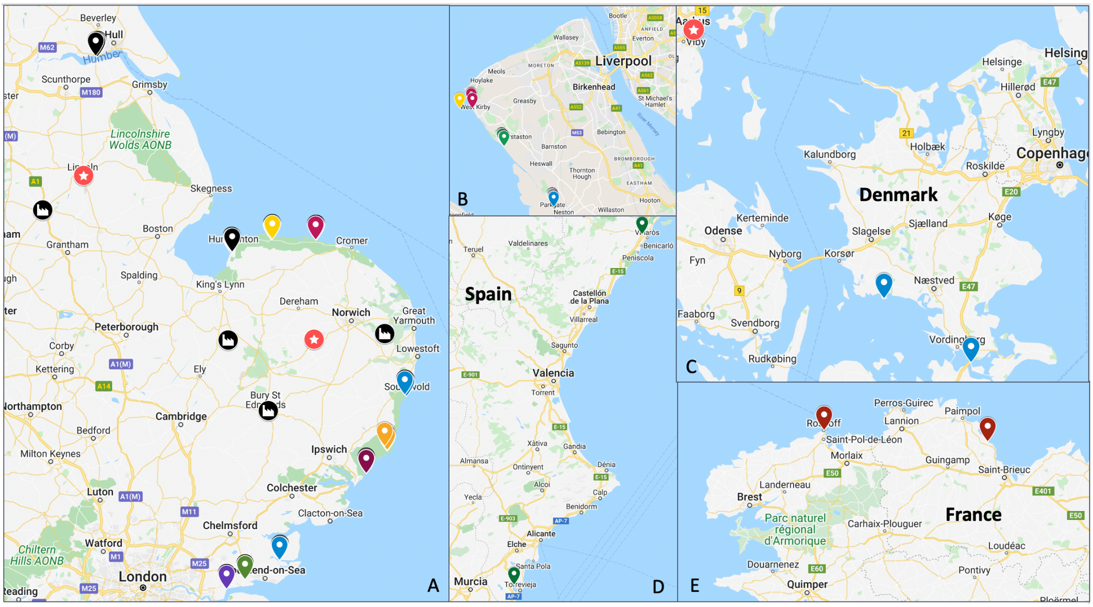


Figure 19 - Sea beet seed material and rust isolates were sampled across locations in Europe. Google My Maps.

Sea beet seeds were collected from English sites (A and B) over the summer of 2019, or provided by collaborating teams from Denmark (C) (Thies Marten Heick, Aarhus University) or from Spain and France (D and E) (Isabelle de Cauwer, Université de Lille). Coloured pins represent the sea beet sampling sites; factory symbols represent the sugar refinery locations in England (A) and pink stars represent the English and Danish beet rust sampling sites (A and C).

In each experiment, four individuals from the same mother plant were grown, with a mean of seven mothers from each site. Additionally, agricultural sugar beets were provided either by the BBRO or by the industrial partner, KWS (**Figure 20**). These were used as positive controls. Indeed, each KWS agricultural line is associated with a known rust susceptibility level that allows to assess the success of the rust inoculation method. Moreover, the sugar beets provided by the BBRO were rust susceptible. The number of plants from each site is comparable between the three experiments, with a difference in the last experiment, where the proportion of agricultural plants was increased to distribute more positive controls across trays. Overall, seeds from the same mother plants were used in the three experiments, with potential misses due to seeds that did not germinate.

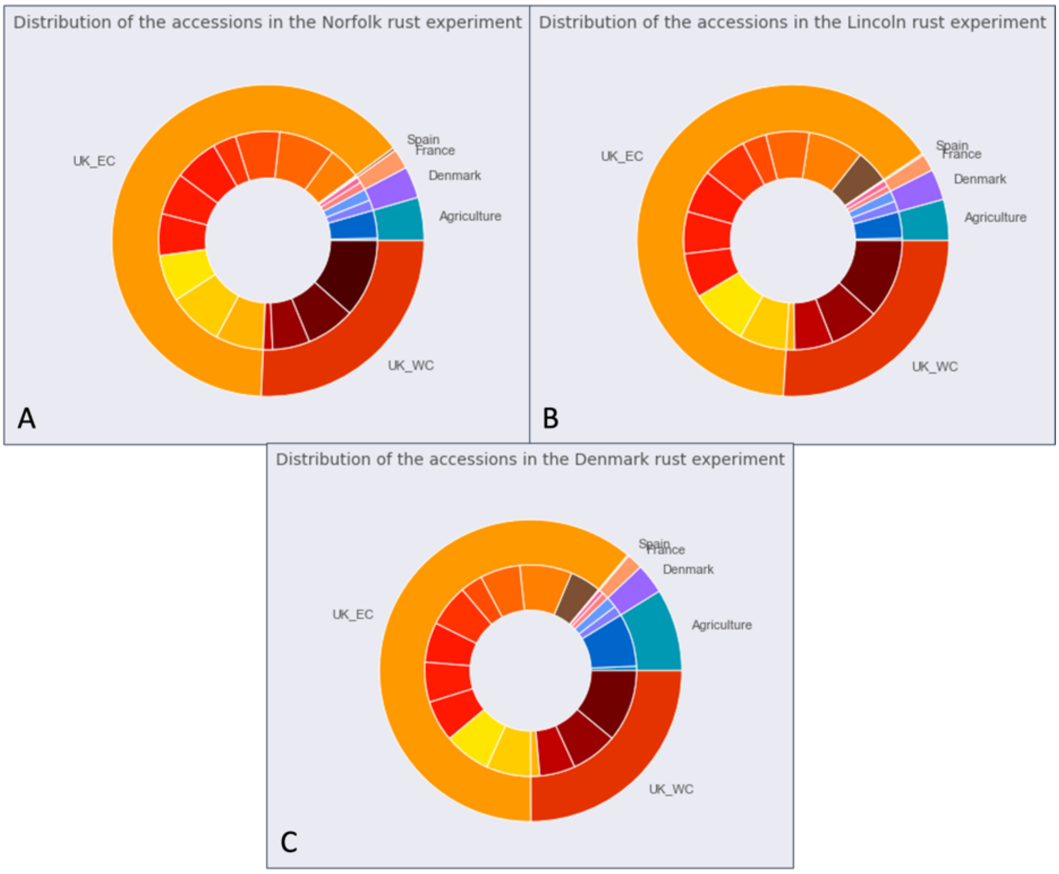


Figure 20 - Three inoculation experiments involve a similar composition of plant sampling sites.

Three experiments involving the rusts sampled in Norfolk (A, included 580 beets), Lincolnshire (B, included 562 beets) and Denmark (C, included 589 beets). UK_EC = east coast of England, UK_WC = west coast of England. Inner partitions reflect site use within broader regions listed as outer partitions.

A subset of sea beets from the three trials was selected for DNA extraction and short read whole genome sequencing (**Figure 21**): approximately 10% of the plants for the Norfolk and Lincolnshire experiments, and around 40% of the plants for the Denmark experiment. The most and least rust infected individuals were chosen, in order to have the best chance to generate two groups showing differences for rust resistance. A larger number of plants was sequenced in the experiment involving a Danish rust, in order to include additional phenotypes which were recorded among plants grown in more controlled conditions than the other two trials.

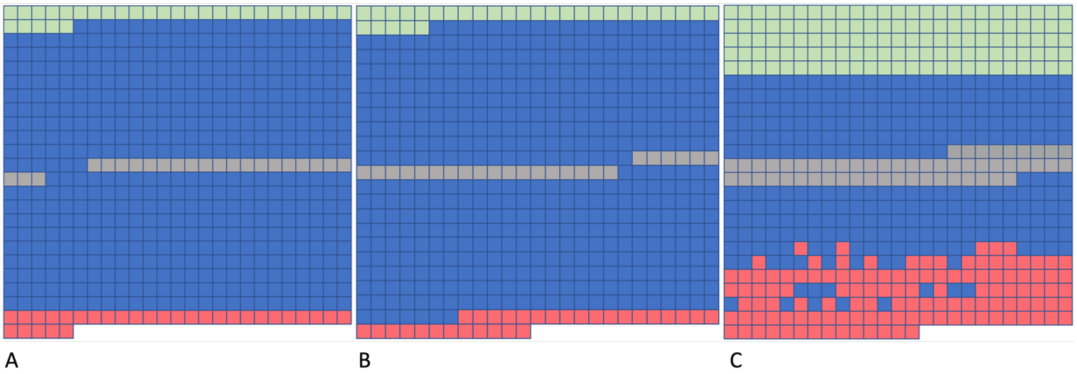


Figure 21 – Association genetics was based on genome sequencing of 370 sea beets from three large-scale rust inoculation experiments.

Plants are represented as squares in the three experiments inoculated with a rust from Norfolk (A), Lincolnshire (B), and Denmark (C). Plants were sequenced from those most resistant (green squares) and most susceptible (red squares) rust infected sea beets. Each square represents a plant part of the experiment, the squares being ordered per level of resistance, except for agricultural beets being represented in grey (which were not sequenced). The plants selected for genome sequencing are represented in green (resistant plants) and red (susceptible plants). The plants represented in blue (sea beets) and grey (sugar beets) were not selected for genome sequencing. 60 plants out of 580 (A) or 562 (B) were selected for sequencing in the Norfolk and Lincolnshire rust experiments, respectively, and 250 plants out of 589 were selected for sequencing in the Danish rust experiment. This third selection (C) included a broader sample of the distribution of resistant and susceptible plants and was also designed to include different agromorphological phenotypes.

The whole-genome short-read sequencing data were used to conduct an association genetics analysis for each rust strain. The methodology was adapted from the AgRenSeq method[164]: 51-mers are extracted from the sequencing data of each individual, pre-filtered based on their correlation with the rust phenotype, and a score reflecting the strength of the association between the presence of the 51-mer and the phenotype is calculated using a generalized linear model. This score is then plotted against a set of NLR orthogroups extracted from the eleven sea beet genome assemblies generated in the **Chapter II**, to identify which NLR orthogroups may provide resistance against the rust pathogen.

To summarise, the present study conducted a large-scale association genetic study adapted from the AgRenSeq method[164], involving inoculation of approximately 1,800 wild sea beets, with two local (UK) rust isolates and an additional isolate from Denmark. Considering the strategies used, such as involving a panel of Atlantic and Mediterranean individuals most likely sampled in populations harbouring a different genetic background; taking advantage of *k*-mers to identify polymorphisms associated with rust resistance; being able to map associated *k*-mers either on NLR loci or on whole genome data; retrieving the associated *k*-mers from a collection of NLR loci extracted from eleven good quality genome assemblies (see **Chapter II**), good conditions are met to find interesting loci for beet culture improvement.

In addition to the potential discoveries enabled by sequencing data from hundreds of sea beets, the phenotyping of almost two thousand wild plants in controlled inoculation trials can give significant insights about susceptibility differences between sites, populations, countries, and about host adaptation to local and/or non-local pathogens. Moreover, this analysis allows the comparison of the degree of susceptibility between wild and agricultural beets, and enables testing the assumption that wild plants would be overall more resistant than their domesticated counterparts.

This analysis could test hypotheses, such as "*Are sugar beets more rust susceptible than their wild counterparts*?", and reveal differences in the NLRs under selection when observing different locations. Indeed, the rust and the effectors it releases potentially differ between locations, and different NLR genes are expected to be subject to selection across populations, as it has previously been shown by Stam *et al.*[172], studying wild tomato populations. Evidence of selection of different resistance polymorphism is best analysed between eastern and western English populations of sea beets because of the predominance of sugar beets around the four beet factories in the east of England (**see Chapter IV**).

III - B. Results

III - B. 1. Large-scale sea beet rust inoculation experiments are replicable with different rust isolates

Rust pustules were observed on sea beet leaves in all three inoculation experiments, within the three weeks post inoculation (**Figure 22**).

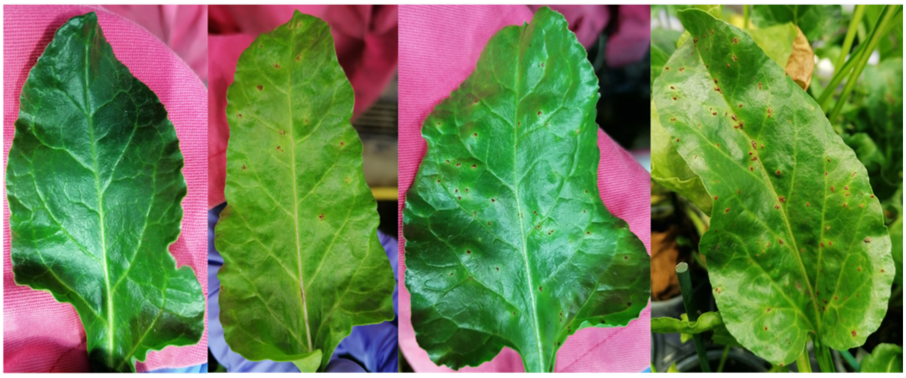


Figure 22 - Controlled rust inoculations reveal a range of susceptibilities across sea beet individuals.

Rust susceptibility observed through the number of rust pustules within the three weeks post inoculation.

Retrieving applicable susceptibility scores is paramount, so rust susceptibility was recorded both as the proportion of infected leaves (data not shown) as well as the total number of rust pustules per plant (**Figure 23**). More than 50% of the plants were infected in each experiment (Norfolk, 55%, Lincolnshire, 62%; Denmark, 77%). Interestingly, the experiment involving the Danish rust showed both a higher proportion of infected plants, as well as increased pustule load per plant (mean pustule number: Norfolk=12.4, Lincolnshire=16.3, Denmark=36.0). The most immediate observation, that English beets infected by Danish rust are impacted to a greater degree, was confirmed with a Mann-Whitney U test to compare the distribution of the two datasets (U=445482.0, p-value=4.47e^{-30}). This result is particularly interesting given predictions of local adaptation by the Red Queen hypothesis. Indeed, this hypothesis comes from an analogy with the queen from the Lewis Caroll*'s Through the Looking Glass* novel, who tells Alice *"Now, here, you see, it takes all the running you can do, to keep in the same place."* The hypothesis was proposed by Leigh Van Valen in the 70's, stating that in order to survive, species need to continuously evolve and adapt to their evolving environment, such as the evolution of their pathogens. Unfortunately, here, there are several differences between the experiments which prevent reliably identifying the cause for this difference of degree of infection (*i.e.* a higher number of agricultural plants in the D experiment).

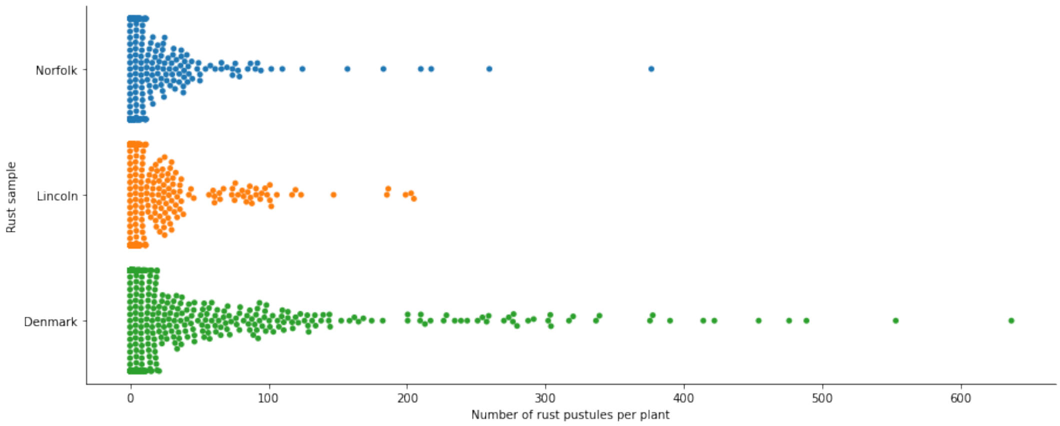


Figure 23 - Rust susceptibility per plant.

The susceptibility phenotype, or total number of pustules per plant, is presented per rust sample utilised in the inoculation trial. Norfolk trial = 554 sea beets, Lincolnshire trial = 538 sea beets, Denmark trial = 537 sea beets.

An interesting observation resides in the fact that crop plants were more infected than the wild beets, both in the proportion of infected plants and in the total number of rust pustules. This second statement was tested using a Mann-Whitney U test to compare the distributions of the two datasets (U=142478.0, p-value=1.18e^{-35}). This reinforces the idea that rust resistance is more prevalent in the wild than in cultivated sugar beet lines. Every sugar beet provided by KWS was infected in the N and D trials, and 23% of those were non-infected in the L trial. Regarding the average number of pustules, infected plants provided by KWS harboured, on average, 5.1 (63.29 vs 12.37), 2.9 (47.14 vs 16.33) and 4.6 times (166.5 vs 36.09) more pustules than wild plants (N, L and D, respectively). Moreover, the most infected plant from KWS was 1.7 (376.5 vs 217.5) and 1.3 (636.5 vs 488.5) times more infected than the most infected wild plant in the N and D trials, respectively. In the L trial, the most infected plant from the agricultural and wild groups harboured a similar number of pustules.

Looking more in detail in the rust susceptibility data, and comparing the beet phenotypes per site of provenance (**Figure 24**), the first observation is that sea beets from almost all the sites have been more sensitive to the Danish rust and more resistant to the rust from Norfolk. Interestingly, and inconsistent with local adaptation (Red Queen), the Danish sea beets show the same pattern.

Susceptibility also differs among sites. One site that exemplifies this is the Humber, the northernmost population on the east coast of England. For all rust strains tested, this population shows a greater average number of pustules than almost all the other sea beet populations (Spanish population set aside because of low numbers of individuals). This emphasises the importance of controlling population stratification in an association genetics analysis. Moreover, when considering the impact of the two British rusts, sea beets from the west English coast (from Merseyside and Cheshire) show more susceptibility than the sea beets from the east coast (Norfolk, Suffolk and Essex). Indeed, the mean number of rust pustules on plants from the Merseyside/Cheshire regions (west coast) was higher in the two trials (N=10.8, L=12.12) than when plants were originating from the east coast (Norfolk, Suffolk and Essex; N=6.40, L=7.17).

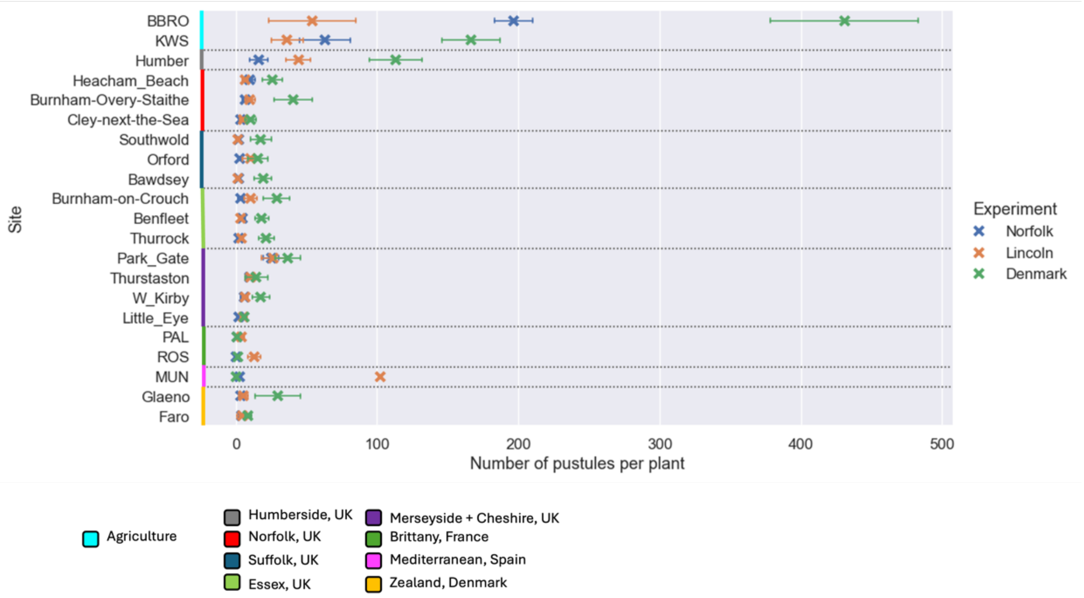


Figure 24 – Rust susceptibility has some trends at the site level.

Crosses represent the average of the number of pustules per plant per site. Sites are coloured by geographical regions (vertical bars). Sites may be grouped into those on the west of England (Merseyside, Cheshire), the east of England (Norfolk, Suffolk, Essex), the northeast of England (Humberside) and countries from mainland Europe. Rust susceptibility corresponds to the total number of rust pustules on each plant, recorded four weeks post inoculation. The absence of error bars for some sites reflects the low number of individuals from these sites involved in the study (e.g. Spain).

Susceptibility scores of KWS plants correlate somewhat with known KWS susceptibility metrics. In each case, a positive correlation is identified between the sugar beet lines' recorded susceptibility and their known susceptibility score (**Figure 25**). However, for the last experiment (**Figure 25, panel C)**, the correlation is less marked than for the other two experiments. Again, this pattern is interesting and could reflect the use of an overseas rust strain. Overall, these results allowed to validate the success of the rust inoculation step, especially considering the difficulties associated with recapitulating a correlation between a score and a direct count of pustules, in presumably different conditions.

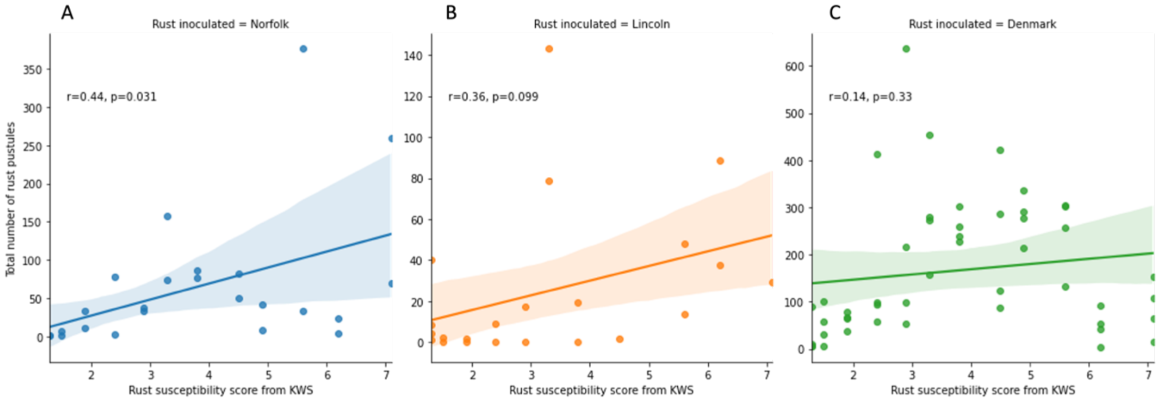


Figure 25 – Sugar beet rust susceptibility recapitulates expected KWS's trend.

The total number of pustules was recorded, at 25-26-27 DPI (days post inoculation), on KWS lines with a known rust susceptibility score. Twelve lines from KWS were integrated in the experiments, as duplicates in the two experiments involving an English rust, and as quadruplicates in the experiment involving a Danish rust sample. The correlation coefficients are r=0.44 (A), r=0.36 (B) and r=0.14 (C).

III - B. 2. Identification of NLR loci providing resistance against English beet rust samples

Beta vulgaris maritima constitutes a wild reservoir of yet unexplored resistance genes. By utilising the resistance phenotypes (above), the next aim is to use genetic diversity to identify loci providing resistance against the *Uromyces beticola* pathogen.

Genotyping was performed by re-sequencing the whole genomes of selected individuals from within each of the three replicates (N, L & D). However, the sequencing strategy was different between the Norfolk & Lincolnshire, and the Denmark replicates. In N & L replicates, from the approximately 600 plants in each trial, 60 plants were re-sequenced (from most resistant and most susceptible classes). In the D replicate, 250 plants were re-sequenced. Associations were analysed using *k*-mers (51-mers), and an association genetics analysis adapted from Arora *et al.*[173] was performed. *k*-mers positively correlated with resistance were analysed and tested through a regression analysis, and a PCA was used to control for population structure.

To identify resistance loci, *k*-mers were mapped to NLR loci extracted from the pan-genome generated in this work (see **Chapter II**), comprising eleven high-quality sea beet genomes and classified into orthogroups comprising both orthologs and paralogs. An average of 196 NLR loci was extracted per genome, with a total of 2,157 potential resistance-bearing loci explored, including 1,127 full-length NLRs and 1,030 truncated or single-domain NLR loci. In the interest of reducing complexity when plotting, the NLR loci from the multiple genomes were categorised into orthogroups (see **Chapter II**). The negative decimal logarithm of the *P* values of each *k*-mer was mapped against these orthogroups, with an aim of seeking loci showing numerous highly associated *k*-mers.

Standout associations of orthogroups with *k*-mers to a resistance phenotype are not obvious in Manhattan style plots (**Figure 26**, **Figure 27**). Orthogroups with peaks of interest could be categorised in two ways, i.e. they have one (or few) high association scores atop a sparsely populated peak of values (henceforth “max”). For example, the Norfolk rust experiment contains nine orthogroups with the highest association score. Alternatively, there are another three orthogroups that don’t receive the highest score, but that do have very well-defined peaks caused by high association scores throughout their range, producing a high mean score (henceforth “mean”). It is somewhat difficult to set a threshold *P* value given this pattern. However, in the present study, multiple replicated experiments have been run and orthogroup performance in multiple trials might be used to distinguish chance from a real underlying resistance association. Given the geographic proximity between the rusts used in the N & L replicates, orthogroup performance in both experiments are compared here using these broad definitions (e.g. max association score, and mean association score).

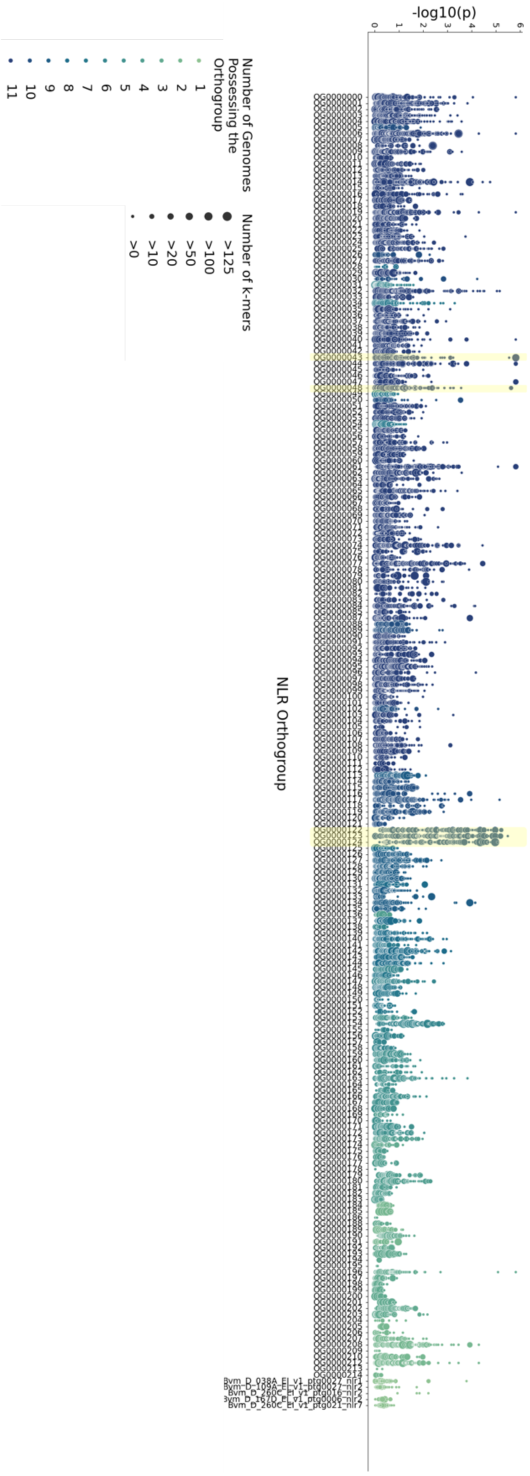


Figure 26 – NLR orthogroup Norfolk rust association scores.

k-mer-based association genetics on NLR loci extracted from 59 sea beets, looking for resistance against a rust sample collected in Norfolk. Each vertical line corresponds to a sea beet NLR orthogroup. Each dot represents *k*-mers positively associated with the resistance phenotype, sharing the exact same associations score. The larger the dot is, the greater the number of *k*-mers is. The strength of the association is expressed using the negative logarithm of the calculated *P* value (y-axis). The darker an orthogroup is coloured, the more sea beet genome assemblies possess it. OG0000043, OG0000048, OG0000122, OG0000123 and OG0000124 are highlighted in yellow, from left to right. *k* = 51.

Perhaps the clearest set of peaks in the Norfolk experiment is the three orthogroups with clear peaks defined by generally high association scores across their range: OG0000122, OG0000123 and OG0000124 (**Figure 26**). However, these three orthogroups are not immediately apparent in the Lincolnshire replicate (**Figure 27**). To account for orthogroup performance over a range of *k*-mer scores, the mean score per orthogroup is calculated in order to distinguish those orthogroups that receive one high (max) score from those that receive generally high scores across the whole range of the peak (mean). Interestingly, while those Norfolk *k*-mers from the mean orthogroups OG0000122, OG0000123 and OG0000124 don't appear to be the most associated *k*-mers with Lincolnshire replicate, their mean association scores are nevertheless among the highest in that trial (**Figure 28**).

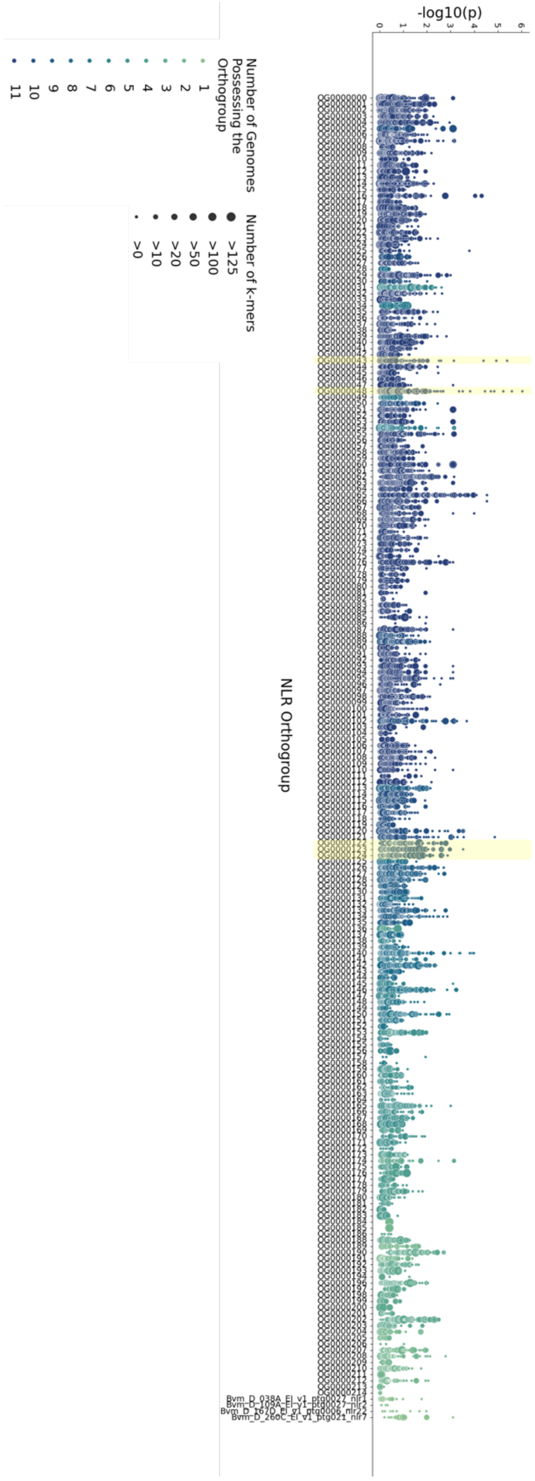


Figure 27 - NLR orthogroup Lincolnshire rust association scores.

k-mer-based association genetics NLR loci extracted from 60 sea beets, looking for resistance against a rust sample collected in the Lincolnshire. Each vertical line corresponds to a sea beet NLR orthogroup. Each dot represents *k*-mers positively associated with the resistance phenotype, sharing the exact same associations score. The larger the dot is, the greater the number of *k*-mers is. The strength of the association is expressed using the negative logarithm of the calculated *P* value (y-axis). The darker an orthogroup is coloured, the more sea beet genome assemblies possess it. OG0000043, OG0000048, OG0000122, OG0000123 and OG0000124 are highlighted in yellow, from left to right. *k* = 51.

However, they are not the highest peaks in the Norfolk trial (5.20, 5.47, 5.04, respectively for OG0000122, OG0000123 and OG0000124). Two orthogroups are among the highest peak scores (max) in both Norfolk and Lincolnshire replicates (**Figure 28**): OG0000048 displays the second highest score (5.61) in the Norfolk trial and the highest score in the Lincolnshire trial (6.014). OG0000043 shows the highest score (5.80) in the Norfolk trial and the second highest score with the Lincolnshire rust (5.373) (**Figure 26**, **Figure 27**).

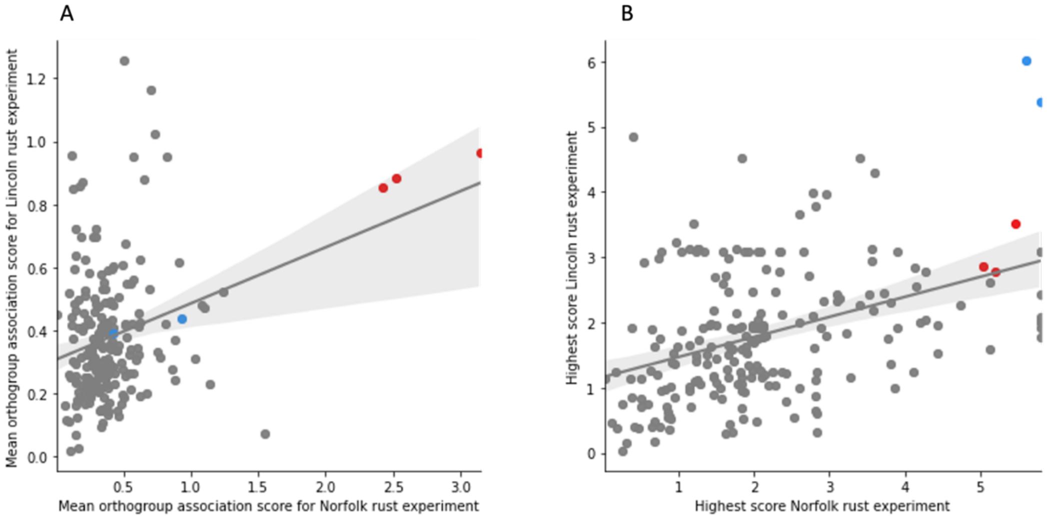


Figure 28 – Five NLR loci are associated with rust resistance in both English rust experiments.

Mean association score (A) or highest score (B) per NLR orthogroup with rust resistance compared between the experiments involving rust collected in Norfolk and rust collected in the Lincolnshire. The NLR orthogroups OG0000122, OG0000123 and OG0000124 are coloured in red and the orthogroups OG0000043 and OG0000048 are coloured in blue.

The *k* length of 51 was chosen primarily because it was demonstrated to be effective in the original AgRenSeq study. Additionally, a large size allows specificity of *k*-mer mapping, enabling the differentiation of NLR sequences with high sequence identity. A shorter *k* length, however, was also tested using the Norfolk experiment data. With this modified parameter, the three orthogroups OG0000122, OG0000123 and OG0000124 were retrieved, displaying a similar profile to that obtained with *k*=51 (data not shown). However, no additional peaks were identified. Consequently, 51 was selected as the *k* value for the remainder of the analyses.

III - B. 3. Association genetics for resistance against a rust isolate from mainland Europe

The sea beets sampled in this study mostly originate from the east and west coasts of England. It is reasonable to assume that these sea beets have co-evolved with local rust strains. Therefore, it is interesting to observe the performance of UK beets in response to a non-local rust strain and more specifically, if the resistant orthogroups are the same. Due to the nature of the use of a non-native (non-UK) rust isolate, the Danish rust trial included 589 mainly English sea beets grown and inoculated in contained (category 2) controlled conditions, of which 245 had their whole genome re-sequenced.

There was no apparent relationship between the well performing orthogroups in the Denmark trial with either the Norfolk or Lincolnshire trials (**Figure S1**). The NLR orthogroup showing the max association score in this Denmark trial is OG0000057, with a score of 11.39 (**Figure 29**), and this orthogroup is also ranked 21st when using the mean score per orthogroup. On the other hand, an orthogroup showing a high mean score throughout the range of its *k*-mers and appears as a peak is the orthogroup OG0000165 (mean score = 1.2, rank=2nd).

The three orthogroups OG0000122, OG0000123 and OG0000124, that were identified has having high mean scores in both N & L trials, are not among the highest associated NLRs in this Danish trial, with respective highest scores of 2.3, 1.94 and 1.58 and respective ranks of 104th, 132nd and 140th when considering the mean association score per orthogroup. Regarding the two orthogroups from N and L trials showing max association scores (OG0000043 and OG0000048), they too don’t stand out neither using the highest association scores (3.45 and 3.84) nor their rank (44th and the 49th).

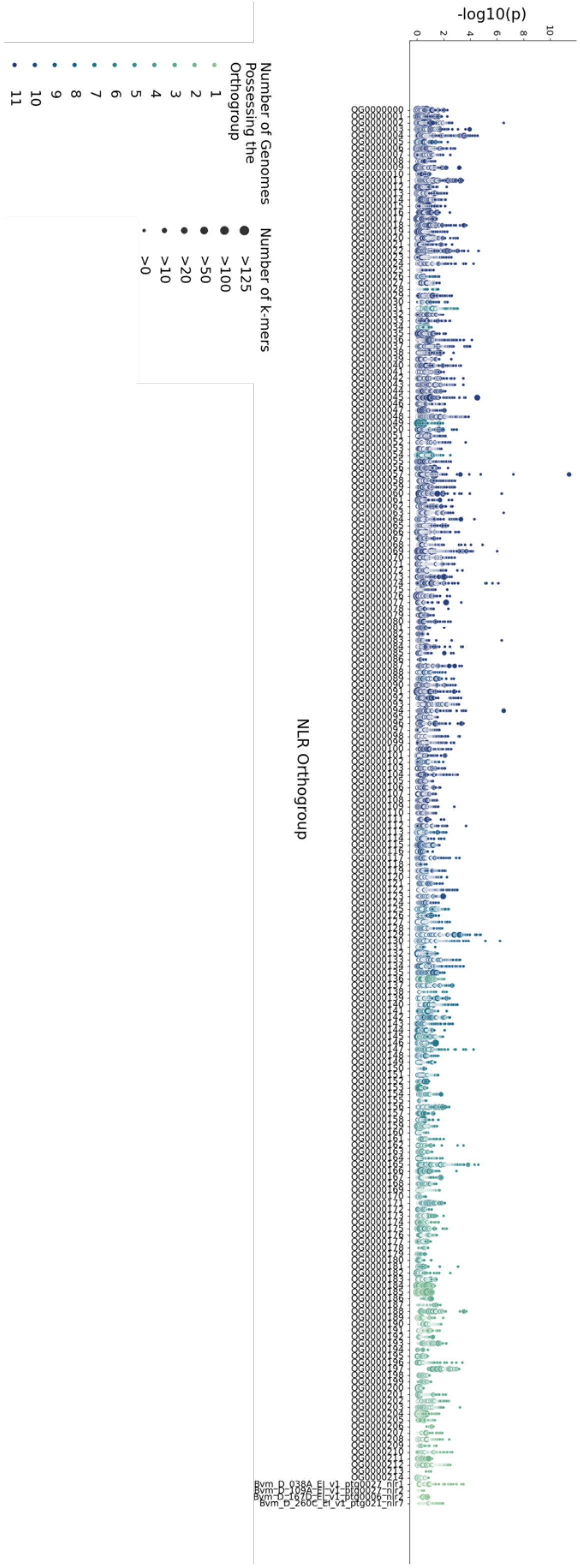


Figure 29 - *k*-mer-based association genetics NLR loci extracted from 245 sea beets, looking for resistance against a rust sample collected in Denmark.

Each vertical line corresponds to a sea beet NLR orthogroup. Each dot represents *k*-mers positively associated with the resistance phenotype, sharing the exact same associations score. The larger the dot is, the greater the number of *k*-mers is. The strength of the association is expressed using the negative logarithm of the calculated *P* value (y-axis). The darker an orthogroup is coloured, the more sea beet genome assemblies possess it. *k* = 51.

III - B. 4. Association genetics in nature

Can an association study directly carried on wild individuals identify resistance genes against the rust pathogen? A key aspect of the present work is not just to ascertain whether a *k*-mer-based approach can be used directly on wild hosts grown and inoculated in controlled conditions, but also if it is possible to skip controlled pathogen inoculation steps and consider natural pathogen infection. To answer this question and see if the mean and max NLR orthogroups OG0000122, OG0000123, OG0000124, OG0000043 and OG0000048 would be highlighted by an association study run directly in the wild, rust phenotypes were recorded on sea beets in their natural environment over the 2019 season. A total of 133 wild individuals from 14 sites in England along the western (4 sites) and eastern (10 sites) coasts were recorded as either infected or uninfected, depending on the presence of at least one rust pustule on their leaves (**Figure 30**).

Most of the sea beets randomly selected and phenotyped in their natural habitat didn't show any rust infection, but a considerable proportion (25.6%) were infected with the rust fungus (**Figure 31**). Interestingly, the proportion of beets harbouring signs of rust infection was greater (48.5%) on the west coast than on the east coast (17.4%).

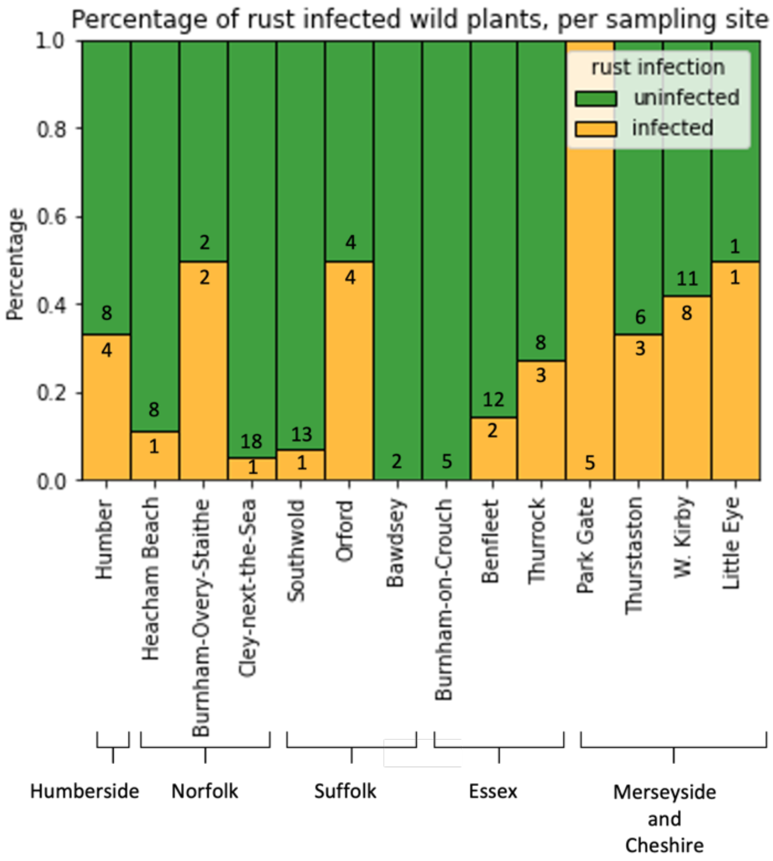


Figure 30 - Almost 30% of sea beets phenotyped in their natural habitat showed signs of rust infection.

Proportion of rust infected and rust non-infected plants from 133 wild sea beets sampled in 14 sites, on east and west English coasts. Any site having less than 5 beets phenotyped is not included in the statistics.

Whole-genome re-sequencing data from 118 wild individuals with a known rust phenotype were generated. On the *k*-mer-based association analysis, visually, three orthogroups stand out from the association plot as max peaks (**Figure 31**). The first two peaks are both new orthogroups, with respect to a resistance association; the third peak, however, is from the OG0000048 orthogroup and shows a high score (4.9). Interestingly, as well as having the highest max score, it is also ranked as the 16th orthogroup for the mean score alongside its associated *k*-mers.

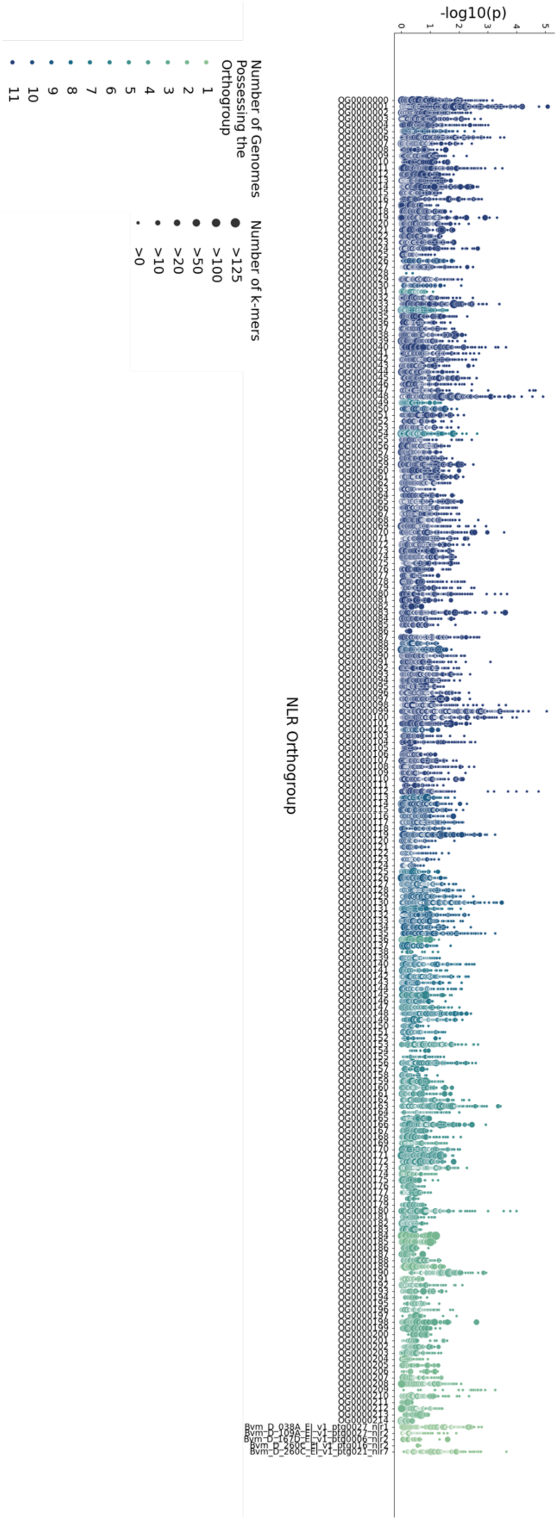


Figure 31 - *k*-mer-based association genetics NLR loci extracted from 118 wild sea beets, looking for resistance against rust.

Each vertical line corresponds to a sea beet NLR orthogroup. Each dot represents *k*-mers positively associated with the resistance phenotype, sharing the exact same associations score. The larger the dot is, the greater the number of *k*-mers is. The strength of the association is expressed using the negative logarithm of the calculated *P* value (y-axis). The darker an orthogroup is coloured, the more sea beet genome assemblies possess it. *k* = 51.

The NLR loci here associated with rust resistance in the different trials were then explored in more details. The orthogroups OG0000122, OG0000123 and OG0000124 are present once in each one of the eleven sea beet genome assemblies, except from the Spanish Es_Catalonia_378 (**Table S7**). Two of these loci correspond to full-length CC-NLRs (OG0000122 and OG0000123), and the third one corresponds to an NLR locus missing the N-terminal domain (OG0000124). Intriguingly, these three loci are physically linked. Indeed, in each genome, they are found on the same contig. The two full-length NLRs are separated by a distance ranging from 12,742 bp to 34,202 bp, depending on the genome observed. The truncated NLR is present on the opposite strand, upstream of the other loci by a distance ranging from 21,031 bp to 23,986 bp.

The two orthogroups showing high (max) association scores in Norfolk and Lincolnshire trials, i.e. OG0000048 and OG0000043, are not present on the same contig in any of the eleven genome assemblies. OG0000048 is a full-length CC-NLR present in all the eleven sea beet assemblies, and as paralogous genes in Dk_Sjælland_406 (**Table S7**). OG0000043 is also a full-length CC-NLR present in all the eleven assemblies, and as paralogous genes in Es_Catalonia_378 (**Table S7**).

OG0000057, identified as associated against the Danish rust sample, corresponds to an NLR orthogroup lacking the N-terminus part, only possessing the NBARC and LRR domains. This locus is prevalent in the eleven assemblies (**Table S7**), but the highest score isn't observed when the *k*-mers are mapped against the French genome, Fr_Bretagne_309, perhaps suggesting allelic diversity not present in that genome. OG0000165 corresponds to a CC-NLR, present in 5 genomes out of 11 (Gb_Essex_038, Gb_Norfolk_095, Gb_Merseyside_206, Dk_Sjælland_406 and Gb_Norfolk_426), as a single occurrence (**Table S7**). Interestingly, this orthogroup is present in the Danish genome Dk_Sjælland_406.

From the natural infection trial, the NLR locus displaying the highest score belongs to the orthogroup OG0000001 (5.07). It is a CC-NLR and also ranks as the 5th orthogroup for the mean score of its *k*-mers. It is the second largest orthogroup (41 genes). Loci from this orthogroup are present between 2 and 6 times in a same genome (**Table S7**) and are physically linked as they are systematically found on the same contig. The second highest score belongs to the OG0000099 orthogroup (5.03), a CC-NLR present in all assemblies (**Table**

S7) which, very interestingly, is also ranked as the 3^{rd} one for its mean *k*-mer score. These two orthogroups don't seem to be strongly associated with resistance against either the Norfolk or the Lincolnshire rust studies.

To summarise the wild association genetics analysis (phenotyped in their natural habitat), the three physically linked orthogroups which had shown high mean association scores both when inoculating plants with rusts from Norfolk and Lincolnshire do not stand out. However, one of the two NLRs displaying one of the highest (max) scores in these two previous experiments also carries a strong association in the wild analysis (**Figure 32**). This is a promising result as, not having validated any of the five associations made from controlled trials, this wild finding potentially adds weight to that result. Moreover, after testing and validating this wild candidate resistance gene, it would suggest the method utilised here could be an effective and low-cost way to quickly identify resistance genes directly in crop wild relatives.

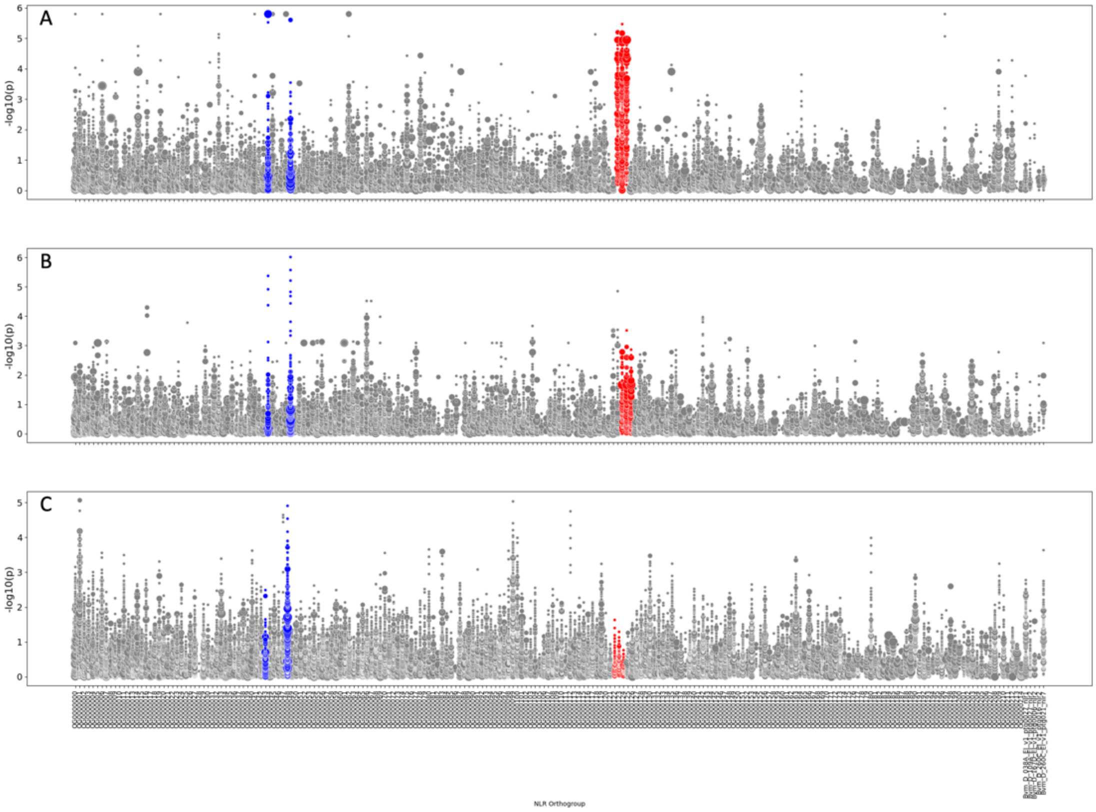


Figure 32 - Five NLR loci show association with resistance against different rust isolates.

k-mer-based association analyses between orthogroups of NLR loci and resistance against rust samples collected in Norfolk (A), Lincolnshire (B), or resistance against natural sea beet rust infections (C). *k* = 51. Each vertical line corresponds to a sea beet NLR orthogroup. Each dot represents *k*-mers positively associated with the resistance phenotype, sharing the exact same associations score. The larger the dot is, the greater the number of *k*-mers. The strength of the association is expressed using the negative logarithm of the calculated *P* value (y-axis). From left to right, the orthogroups OG0000043 and OG0000048 are coloured in blue and the orthogroups OG0000122, OG0000123 and OG0000124 are coloured in red.

III - C. Discussion

To summarise the main results observed through four association genetics analyses involving different beet rusts, three physically linked NLR loci have been identified as showing high mean association scores throughout their *k*-mers in both the trials involving the Norfolk and the Lincolnshire rusts. These loci belong to the orthogroups OG0000122, OG0000123 and OG0000124. On another side, two NLR loci belonging to the orthogroups OG0000043 and OG0000048 showed (max) association scores amongst the highest in the Norfolk and Lincolnshire trials. Importantly, one of them (OG0000048) was also identified in the trial conducted directly on naturally infected wild hosts. Although similarities were thus identified between independent UK analyses, this was not the case when considering the inoculation trial involving a non-UK rust. These main results are discussed here, as well as additional results, such as differences of infectivity between rust samples, differences of susceptibility between plants, the potential improvements of the method utilised here and further work which could be considered to continue this project.

One of the first results observed is the difference in infectivity and virulence between rust samples inoculated in the three trials. Indeed, the Danish rust performed better than the two UK rusts, both in its capacity to infect and in the degree of the infection observed. This is consistent with the hypothesis that coevolution with a native rust has preserved local resistance that is important to detect and react before the establishment of the invasion by Norfolk and Lincolnshire rusts. However, comparing UK rusts methods and the Danish rust method, we cannot account for the impact of the extra level of controlled environmental conditions (Category 2) required to infect with a non-native pathogen in the UK. If the crops incorporated into the trials are considered as controls and compared between experiments, increased infection in these controls suggest that the plant growth and incubation conditions played a role in the results observed due to these controls being more infected in the Denmark trial.

Following on the differences in beet susceptibilities, UK rust inoculations revealed a higher susceptibility for British wild beets originating from the west coast than the ones from the east coast. These results corroborate again the fact that hosts and pathogens coevolve at local scales and that it is probable that sea beets from the east coast are adapted to face a crop-pathogen attack when the rust is sampled in the same area. However, this pattern of increased susceptibility of western wild beets is also observed when considering natural infections of sea beets, as plants sampled in their natural environment were on average more often infected on the west than on the east coast. This observation suggests perhaps the differences in weather conditions, as this western region (Liverpool) is wetter than the sites bordering the east coast. This could then be a more favourable environment for the rust disease to settle the infection once the spores have landed on sea beet leaves. This sets the importance of identifying resistance genes *in situ* in a maritime weather environment that may be more conducive to the growth of fungi.

The trials highlighted a clear difference in the sea beet versus the crops rust susceptibility. This could correspond to a clear observation of the loss of genetic diversity during the bottleneck induced by domestication. However, conclusions can't be drawn due to differences between the two groups compared: the number of wild and crop plants is, first, very different, moreover, the sugar beets utilised were chosen for their rust susceptibility. Indeed, despite varying degrees of rust resistance, all KWS lines did get infected in at least one of the experiments. Furthermore, the plants provided by the BBRO also have a susceptible genotype. It could then be argued that no resistant sugar beets were incorporated in the trials, and that these results are biased towards susceptibility in control beets. Moreover, sugar beets' genomic background is from the Mediterranean sea beet group[10], thus, their set of resistance genes may be different in general from the Atlantic beets, not only because of a bottleneck, but the genetic background could be an important determiner for the way these plants respond to the pathogen.

One additional observation regarding the crops utilised resides in the way KWS's lines showed less correlation between KWS's susceptibility scores and the scores recorded in this work when the Danish rust was inoculated. This could be due to the nature of the rust utilised in

KWS's trials to evaluate the susceptibility of their lines: if the rust isolate tested is genetically closer to English rust than to Danish rust, it could account for the lack of reproducibility of the susceptibility results, and again highlights the importance of local co-evolution.

To close the phenotypic aspect of the results and touch now on the genotypic observations, the absence of clear association peak(s) on the Manhattan-like plots can't be ignored. Instead, the major outcome of the analysis is the differences in the most associated (max score or mean score) loci between experiments. The fact that English-Danish rust experiment outcomes were more different than the English-English ones ties in with the higher susceptibility of western beets towards the east Anglian rusts compared to eastern beets. This highlights indeed the phenomenon of coevolution and the spatial selection of the NLR genes to local pathogens. This aspect is noted as well when looking at the orthogroup OG0000165, performing well against the Danish pathogen, and, despite lacking in more than half the genome assemblies studied, is present in the Danish sea beet genome.

Interestingly, comparing outcomes between trials facilitated highlighting of five NLR loci potentially providing resistance against east England rust pathogens. Moreover, identifying one of these loci performing the association study directly in a wild setting reinforces on one hand the idea that this gene is of major importance to contribute to English rust resistance. On the other hand, this gives information about the efficiency of conducting association genetic studies directly on wild hosts, without setting any plant growth nor pathogen inoculation trials. Furthermore, these results are promising indications that results are not due to false-positive signals, as the identification of the same genes in independent studies involving different pathogens strengthens the probability that these loci actually correspond to true positive signals.

Looking in detail at the four full-length potential resistance genes identified, an intriguing detail is the prevalence of CC-NLRs and the absence of TIR-NLRs. This result must be put in the context of the low proportion of TIR-NLRs among the collection of NLRs harboured by the eleven sea beet genomes assembled in **Chapter II** (1 to 5 TIR-NLR per genome). Interestingly, truncated NLRs have been reported as being of great importance for pathogen resistance, playing the role of antagonists to their neighbouring full-length NLRs[174]. However, the location of this locus classified among the OG0000124 orthogroup in a haplotype with two other associated loci (OG0000122 and OG0000123) could in reality correspond to a single element part of this haplotype actually being involved in the resistance. The association analysis would then have highlighted a broad signal for this region, and the real signal could reside in only one of these loci. Another hypothesis would be that these physically linked genes have great chance to be co-expressed and to function together in the resistance process.

While the NLR repertoire extracted from only eleven sea beet genomes is analysed here, with a single representative of the Mediterranean group (Es_Catalonia_378), it is an exciting result to note that this particular genome is the only one missing the orthogroups OG0000122, OG0000123 and OG0000124. This concords with the differences of susceptibility phenotypes between the wild and cultivated beet groups: it makes sense that descendants from the Mediterranean split differ in their repertoire of NLRs compared to descendants from the Atlantic group, and that major resistance genes in one group misses in the other one. This has been demonstrated by Sandell *et al.*, with the example of the mildew pathogen: Mediterranean sea beets were shown to possess resistance against this local pathogen, lacking in the genome of Atlantic sea beets[10]. To come back to rust resistance, regarding the fact that each one of the three orthogroups lacks in the Spanish genome, it could be explained by their close proximity. It is indeed understandable that either genomes possess the three genes or none of them. It is also potentially promising from the perspective of introducing variation by breeding.

The *k*-mer-based method chosen in this study to search for resistance genes is different from a conventional GWAS analysis in the way that the association is performed directly comparing the set of *k*-mers between samples, instead of relying on genetic markers from a reference genome. This *k*-mer method is particularly suitable in the present work as sea beet is a wild species without a reference genome available, and the method focuses on NLR genes, allowing a rapid resistance gene identification. This is however also a downside of the present method, as resistance-associated elements situated at another location in the genome would be missed. Hypothesising that beet rust resistance is carried by NLR gene(s), a *k*-mer-based method is expected to perform better in a non-model or non-crop organism, and to allow a rapid identification of resistance.

Several hypotheses can explain the absence of a clear association peak in this study. First, "*the gene*" responsible for resistance against one or multiple inoculated rusts in the panel of sea beets could simply be absent in all the eleven genome assemblies. Alternatively, beet rust resistance could be a polygenic trait that the method used here couldn't identify. Importantly, the method applied in this work can be improved in several ways, one of them being the enlargement of the panel of NLR loci described, through the HiFi sequencing of additional genomes, extending the so-far generated pan-NLRome (**Chapter II**). Alternatively, this can be done by re-sequencing the panel of beets part of the experiments with a deeper coverage, to have the opportunity to map the associated *k*-mers against NLR genes extracted from many genomes having been phenotyped for their resistance. This last point raises another potential improvement in the design of the association analysis: in the current case, for logistical concerns, the pan-genome established in the **Chapter II** comes from the trial involving the Danish rust. A better design would include, in the collection of assemblies, genomes from plants subject to the inoculation of the two English rusts. On the rust side, constraints of time, space and resources limited the experiment to the use of samples from a single field. Reducing the pathogen input to a single isolate per experiment would reduce the number of effectors released inside the host, and, therefore, likely make an association signal appear clearer. Furthermore, numerous results in the present study are indicative of local adaptation to pathogen species and this increases the requirement to sequence the rust strains used, as well as perhaps survey for rust populations in the wild.

A key next step for the present work would consist in the test and validation of the five candidate loci potentially providing resistance against the English rust. To do so, one way would be to analyse the expression profile of these different NLR genes, and look at a potential change in the gene expression throughout pathogen infection, corroborant with the association with resistance that has been reported. If these analyses are conclusive, the final step would reside in the cloning of single and/or multiple loci into rust susceptible beets, to test on one hand the functionality of these genes as singletons and, on the other hand, to test the three loci OG0000122, OG0000123 and OG0000124 for their efficiency as a cluster of NLRs functioning together.

III - D. Material and methods

Plant sampling and phenotypes recording in natural environment

Fourteen coastal sites were visited over the summer 2019 to select approximately 15 plants and attach bags around flowers. A second visit few months later aimed to collect the bags containing seeds. During these two visits per site, leaf material from the plants sampled for their seeds was collected, and phenotypes were recorded for the presence/absence of leaf diseases. Collaborators from Aarhus University (Thies Marten Heick) and the University of Lille (Isabelle de Cauwer) kindly shared additional seeds from mainland European locations: Denmark, and France & Spain, respectively.

Rust sampling and bulking

The BBRO kindly provided two rust samples from Norfolk and Lincolnshire sugar beet fields. The sample from Norfolk (referred as N rust) was collected in October 2019 and the sample from the Lincolnshire (L rust) was collected over the 2020 and 2021 seasons. These samples have been bulked in the laboratory to increase the amount of inoculum available for the large-scale trials, through successive cycles of rust inoculation/collection using rust susceptible sugar beets provided by the BBRO. After collection, rust samples were desiccated for 3 days in a desiccator with desiccation granules, and stored at -80°C.

Two rust samples from Denmark (D rust) were kindly provided by Thies Marten Heick, respectively collected in 2018 and 2021.

The rust material utilised in this study is non-monogenic as it hasn’t been bottlenecked from a single spore, but corresponds to samples from the same location, i.e. same sugar beet field.

Large-scale beet growth in controlled environments

In average, seven mother plants per site were selected to provide four individuals included in the analysis (**Table S8**). The seeds were sowed in 9cm diameter pots using the “cereal mix” soil provided by the John Innes Centre Horticultural Services in April, May and November 2021, respectively, for the Norfolk, Lincolnshire and Denmark trials.

The beets taking part in the rust inoculation experiment with rust samples from Norfolk and the Lincolnshire were grown on the Norwich Research Park in glasshouses providing automatic watering for 3 months and 5 months, respectively. The plants were randomly dispatched on trays of approximately 15 plants. They were then moved in a level 2 containment glasshouse. Both glasshouses and containment rooms had light levels between 200 and 300 µmol.

The beets taking part in the experiment involving the Danish rust sample were grown in a level 2 containment building on the NBI, dispatched in 4 growth cabinets, in 44 trays of approximately 15 plants. The beets were randomly dispatched following an alpha lattice design (alpha Gendex module, http://www.designcomputing.net/), assuring that each replicate contained sea beets from each region as well as agricultural beets. The design was created with the following arguments: *v*=60 (number of treatments *v=ks)*, *r*=11 (number of replicates), *k*=30 (block size) and *s*=2 (number of blocks per replicate). The cabinets were set with cycles of 16 hours with 250 µmol of light at 16°C and 8 hours of dark at 14°C and 70% of relative humidity level.

Rust inoculation

After one week (N rust) or 4 weeks (L rust) in the containment glasshouse, respectively either 411.4 mg or 464.8 mg of agricultural rust stored at -80°C were thawed up at 42°C for 2 minutes and mixed in 195 ml of methoxy-nonafluorobutane (3MTM NovecTM 7100 Engineered Fluid). 5 ml of this solution was then sprayed onto each tray using an air gun. Each tray was then sprayed with water and enclosed in a plastic bag. The plants were incubated in another room at 14°C in the dark for 24 hours before the bags were removed, and the plants transported back to the glasshouse.

After 4 months of growth or 114 days of growth (D rust), 612.4 mg of rust stored at -80°C were thawed up at 42°C for 2 minutes and mixed with 220 ml of methoxy-nonafluorobutane (3MTM NovecTM 7100 Engineered Fluid). 5 ml of this solution was then sprayed onto each tray of approximately 15 plants using an air gun, while working in a microbiological safety cabinet. The plants were then incubated in the cabinets for 24 hours in the dark with 100%

of relative humidity level at 14°C. After this period, the cabinets were re-programmed with their initial settings.

Recording rust susceptibility

Rust pustules were counted by eye through two independent observations, at 27-28-29 (N rust), 27-28-29 (L rust) and 25-26-27 (D rust) DPI. The total number of pustules was recorded for each plant, from every leaf. A mean was then calculated from the two independent observations to obtain the final rust susceptibility score.

Leaf sampling

In order to carry out DNA extractions, leaf samples of each plant involved in the 3 experiments were collected using a leaf puncher at 35-36-37 (N rust), 32-33 (L rust) and 39-40-41 (D rust) DPI. 10 to 20 punches per plant were collected in Eppendorf tubes and stored at either -80°C or -20°C. The punches were done avoiding the main veins of the plants as well as rust pustules.

Plant selection and DNA extraction

A subset of 370 sea beets from the three trials (N=60, L=60, D=250) was chosen for DNA extraction, based on rust infection scores, as well as 150 wild sea beets from which biological material was sampled on the coast. A SNP analysis using markers developed on sugar beets was performed by KWS on the mother plants sampled on the coast. Utilising the results from this genotyping analysis as well as knowledge about sister relationships, pairs of sea beets were constituted in each trial, grouping the most related plants harbouring contrasting rust susceptibility phenotypes.

Genomic DNA from 520 sea beets was extracted using the Qiagen DNeasy Plant Mini Kit, according to the manufacturer’s instructions. The elution was done in 50 µl of Tris low EDTA buffer.

Sequencing

The sequencing of the whole genome of 520 sea beets was conducted by the Technical Genomics group at the Earlham Institute, on the Illumina NovaSeq 6000 platform (150 PE). The samples went through the LITE library preparation protocol[175] and were sequenced with a mean depth of 20x (**Figure S2**, **Table S9**), with two sequencing runs per sample, each on two lanes. The raw read files were combined before data processing. Sequencing data has been submitted to the Sequence Read Archive (SRA) database from the National Center for Biotechnology Information (NCBI) under the BioProject PRJNA1209597.

Association analysis

Raw reads were pre-processed with Trim Galore[176] version 0.4.0, with the parameter –length 70. For each association analysis, a *k*-mer presence/absence matrix was generated, jointly counting *k*-mers from the multiple read files using kmtricks[177]. The counting parameters were k=51, a=2, s=3, n=1, and s=5, with *k* being the *k*-mer length, a the hard-min or the minimal abundance of a *k*-mer in an individual's genome to be kept, s the soft-min (or solidity threshold) or the minimal abundance of a *k*-mer in an individual's genome to be kept during merge between individuals' partitions, n the share-min or the number of genomes the *k*-mer has to be solid in to be kept, and s the number of genomes the *k*-mer has to be present to be kept. In other words, in the present analysis, *k*-mers present only once in a reads file were removed, and the *k*-mers present twice were kept only if present at least three times in another reads file. Additionally, *k*-mers were discarded when present in less than 5 genomes.

Individuals with a sequencing depth lower than 5 were not kept in the analysis, leaving a total of 59, 60, 245 and 118 individuals, respectively, in the Norfolk, Lincolnshire, Denmark and natural rust infection analyses.

Association scores were attributed to each *k*-mer and mapped to NLR orthogroups extracted with NLR-Annotator[178] from the eleven genome assemblies generated in the **Chapter II** using the AgRenSeq_GLM pipeline (https://github.com/kgaurav1208/AgRenSeq_GLM). The rust susceptibility phenotypes were converted into negative numbers, in order to give the highest score to the most resistant plants (i.e. 0) and the lowest score to the most susceptible plants. The NLR orthogroups were flanked on either side of their sequence with 51 bp. Population structure was controlled providing the pipeline with a random subset of 315 *k*-mers from the presence/absence matrix.

Plotting *k*-mer-based associations

Low sequencing depth individuals were removed from the association analysis (only 245 kept for the Danish analysis).

Associated *k*-mers were first independently mapped to one set of NLR loci extracted from eleven sea beet genome assemblies. Then, NLR loci were classified into orthogroups, and the *k*-mers were plotted on a horizontal axis corresponding to the orthogroup they were attributed to. The size of the dot represents the number of *k*-mers with a same association score, in a single genome.

The different plots presented in this chapter were generated using Python (version 3.8.5).

All computational experiments were carried out on a high-performance computing (HPC) cluster running AlmaLinux 9.5, utilising the x86_64 architecture. The cluster utilises SLURM (Simple Linux Utility for Resource Management) to manage job scheduling and resource allocation, is equipped with a processor operating at a speed of 1.5 GHz and is provisioned with 503 GiB of RAM. The system supports 64 CPU cores.

Chapter IV - Exploring English sea beet population structure

IV - A. Introduction

IV - A. 1. Wild sea beet populations and their genetic diversity

Sea beets are wild maritime plants growing along Atlantic coasts from North Africa to the North Sea, as well as along Mediterranean coasts. Seeds are mainly dispersed by gravity close to the mother plant[179]. However, the colonisation of new habitats via ocean currents is possible since sea beet seeds are still viable after several weeks in the sea[180]. It has been argued, for example, that sea beet populations on Germanic Baltic coasts were probably derived from Danish populations[180]. This constitutes the principal mode of transfer of genetic material through long distances, as a study showed that pollen spread by wind rarely travels beyond 200 meters[181]. This way, the habitat and water currents have a great impact on colonisation and gene flow between populations.

The importance of sea currents in shaping sea beet population structure has been shown by Leys *et al.*[182], along with its mating system, and past climatic changes and anthropogenic pressures. Studying wild beets sampled along the Atlantic coast from Morocco to France, as well as in locations bordering the Strait of Gibraltar, they showed a decrease in the allelic richness and genetic diversity from the southern to northern coastal locations. This is consistent with a potential scenario of recolonisation from the Mediterranean region after the last glacial era. Moreover, the population structure of sea beet was shown to be consistent with an outcrossing mating system.

Due to the ability of beets to spread long distances, the fact that wild beets have, in some locations, been coexisting with bottlenecked domesticated beets for the past two hundred years, and that these subspecies can cross[183], it is reasonable to assume that crop genes could escape into nature through seeds[184] or pollen dispersal. This comes with two main concerns:

the impact on wild sea beet genetic diversity by the introgression of diversity from relatively bottlenecked cultivated beets, and the potential for the introduction of genetically modified genes into the wild. These considerations have been investigated through different studies in the last 30 years. Bartsch *et al.*[185] confirmed the presence of gene flow between cultivated and wild beets. They compared sea beets growing in the most important sugar beet seed production area in northeast Italy to sea beets sampled in different locations, out of Italy. Contrary to expectations, they showed that, in terms of Nei's estimated heterozygosity and proportion of polymorphic loci, sea beets growing close to the seed production area harboured a higher diversity than the control group. They suggested that gene flow over more than a hundred of years, from different sugar beet cultivars and different beet subspecies, has shaped the genetic diversity encountered in Italian wild beets without having necessarily decreased it. In a later study, by observing the gene flow between cliff top and drift line sea beet populations in South England, Cureton *et al.* [186] concluded that a transgene could more rapidly spread in the latter group due to proximity with water, and that, over and above preventing bolting in the crop, it is then of great importance to consider watercourses when designing potential future transgenic trials.

More recent estimates of diversity in cultivated and wild beets have utilised *k*-mer-based comparison methods[187] and were investigated by Sandell *et al.*[10]. A split between two groups of sea beets was highlighted: a Mediterranean group and an Atlantic group, with sugar beets constituting a sister group to the Mediterranean group. It was proposed that the strait of Gibraltar could be the cause of the separation between two distinct groups. The origin of sugar beet domestication was suggested to be located in Greece.

Studying genomic variants at a population level is important to understand mode of reproduction and spread, but more especially in this system because measures of diversity and linkage are important considerations for its potential for association genetics, for resistance and other traits. Indeed, identifying differences in the presence/absence and/or the allelic diversity of resistance genes between natural sea beet populations can help to understand how rare or frequent the contemporary resistant alleles are and potentially to predict their durability while developed in crops. It is yet not known but tantalising to determine whether specific population genetic patterns could even be used to survey and potentially identify genes without resorting to costly and time-consuming association methods.

IV - A. 2. Studying genetic diversity and gene flow between populations

Populations largely predominantly share their polymorphic sites and because of the random sampling of alleles from one generation to the next, these allele frequencies "drift", randomly over time[188]. The consequence of this is that the longer two populations are separated, the greater the allele frequency differences (on average) between them. This increasing genetic differentiation is countered by migration, which shares diversity among populations. These ideas were conceptualised using the differentiation and spread of diversity among island populations[189]. However, they are particularly useful concepts because the forces that underly diversity and differentiation are, for the most part random, and so understanding the population genetics of a species facilitates the distinction of chance from natural selection. For example, resistance polymorphism maintained in one population, for its role in defence against a pathogen, would be expected to have a lower differentiation (compared to the genome wide average) with another population impacted by the same pathogen genotype.

A general concept in population genetics used to describe populations is called the effective population size. This is a theoretical value which corresponds to the number of randomly breeding individuals that would be required to maintain the observed level of genetic diversity[188]. Consequently, this number is generally much smaller than the actual (census) population size[190] but can be a useful statistic to represent the level of diversity within different populations. The population effective size is important because, all else being equal, larger populations carry more genetic diversity and therefore, have more substrate for adaptation, and also the efficacy of natural selection is greater in larger populations because of the relative reduction of the impact of random genetic drift[188]. An approximation of the effective population size can be calculated by rearranging for nucleotide diversity (π), calculated as π = 4.Ne.u, with Ne being the effective population size and u the mutation rate per nucleotide[188].

In the 1960s, Wright devised different statistics to describe the structure of genetic variation between populations, called the F statistics[191]. They are today the mostly used statistics in population and evolutionary genetics. Among these statistics is the differentiation factor, or

F_{ST}. For a given mutation rate, the amount of differentiation between populations is influenced by genetic drift and the migration rate. F_{ST} reflects the differences in allele frequencies between two populations. The larger this index is (0-1), the more different allele frequencies are and, the more differentiated two populations are[192].

Considering the efficacy of natural selection by accounting for genetic diversity, as measured by SNPs, ignores recombination. Recombination occurs through the exchange of genetic material between homologous chromosomes during meiosis, via cross-over events creating novel combinations, or haplotypes [193]. Loci that are close to one another are generally less likely to be broken up by recombination and are said to be in Linkage Disequilibrium (LD). LD is calculated by measuring the probability of co-occurrence of two SNPs, comparing multiple individuals from a population. However, linkage disequilibrium can operate over larger distances in the presence of selection, preserving the haplotype, or after a recent selective sweep[194].

Rates of recombination can vary, as can the rate of sexual reproduction in many organisms. Scientists working on fungi have used correlation coefficient between two SNPs (r^2), sampled at the population level, to develop a way of measuring how frequently a species was having sex[195]. Plotting the r^2 value against the distance between two SNPs highlights the decrease in the strength of LD with increasing distances. In the fungal example, the rate of decay in linkage correlates with known rates of sex in that species. However, in plant populations, linkage decay is a composite of the rate of recombination (or outcrossing) and the level of diversity (effective population size) and has been used for assessing the utility of landraces as breeding stock[196]. Linkage disequilibrium has multiple impacts, notably in plant breeding, it is of benefit in marker-based association genetic studies, where genetic markers are used to identify genetically linked traits of interest. However, the ability of a breeder, or natural selection, to select a phenotype depends on the impact of genes on that haplotype. Where one gene may contribute to a phenotype, others may detract and so increased linkage, as is associated with small natural populations, or inbred crop varieties, will reduce the potential to observe the impact of a gene without the epistatic interactions of linked regions (Hill–Robertson interference[197]). In other words, the high linkage observed in small populations and inbred lines reduces the efficacy of selection, or the potential for the breeder to discern traits.

In addition to the utility of whole-genome sequencing of over five hundred sea beets from England, for association genetics, sampling naturally evolving wild populations at a single point in time allows an exploration of the population genetics of wild sea beets in England. This analysis aims to place among and compare the set of sea beets studied here with a collection of beet breeding material; to unravel the structure of the *Beta vulgaris maritima* species on east and west English coasts in terms of defining an estimation of the number of populations. Moreover, the level of recombination is studied (LD) among the defined populations. Finally, nucleotide diversity and genetic differentiation between these populations are measured at the whole genome level as well as focussing on NLR genes.

IV - B. Results

IV - B. 1. The 520 re-sequenced sea beets partition into the Atlantic and the Mediterranean groups

To position the re-sequenced sea beets from the present study (**Chapter III**) among wild and cultivated beet accessions whose genetic diversity has recently been explored[10], pairwise distances between sketches of randomly selected *k*-mers from each individual were calculated to generate a tree reflecting these genomic differences (**Figure 33**). The most remarkable observation is the split between Mediterranean (orange branches) and Atlantic (blue branches) sea beets described by Sandell *et al.*, with the majority of the present samples (pink leaves) belonging to the Atlantic group. The assignment of individuals to one or the other group is purely influenced by the geographic aspect: only the individuals sampled on the east Spanish coast are classified among the Mediterranean group.

Interestingly, one out of the 512 samples didn’t cluster in any of these two major groups, but, instead, grouped with the outgroups: the spinach crop and the wild species *Patellifolia procubens*, another species from the Amaranthaceae family and of the Betoideae sub-family. This individual was collected on a small island called Little Eye, close to the shore of West Kirby’s beach, in the Merseyside (west coast of England). The phenotypic similarities between *B. maritima* and *P. procubens* could explain the fact that seeds from the latter were included in the collection.

As previously noted in the original study, the breeding material from the different seed companies included showed a close relationship among and between breeding programs. Sugar beets are closer to Mediterranean sea beets than to the Atlantic group.

The sea beet genomes (blue leaves) selected for traits of interest, including rust resistance, and sequenced for the generation of a sea beet pan-genome (**Chapter II**) as well as for the rust association study (**Chapter III**), show a scattered distribution across the Atlantic side of the tree, in addition to one individual present from the Mediterranean lineage.

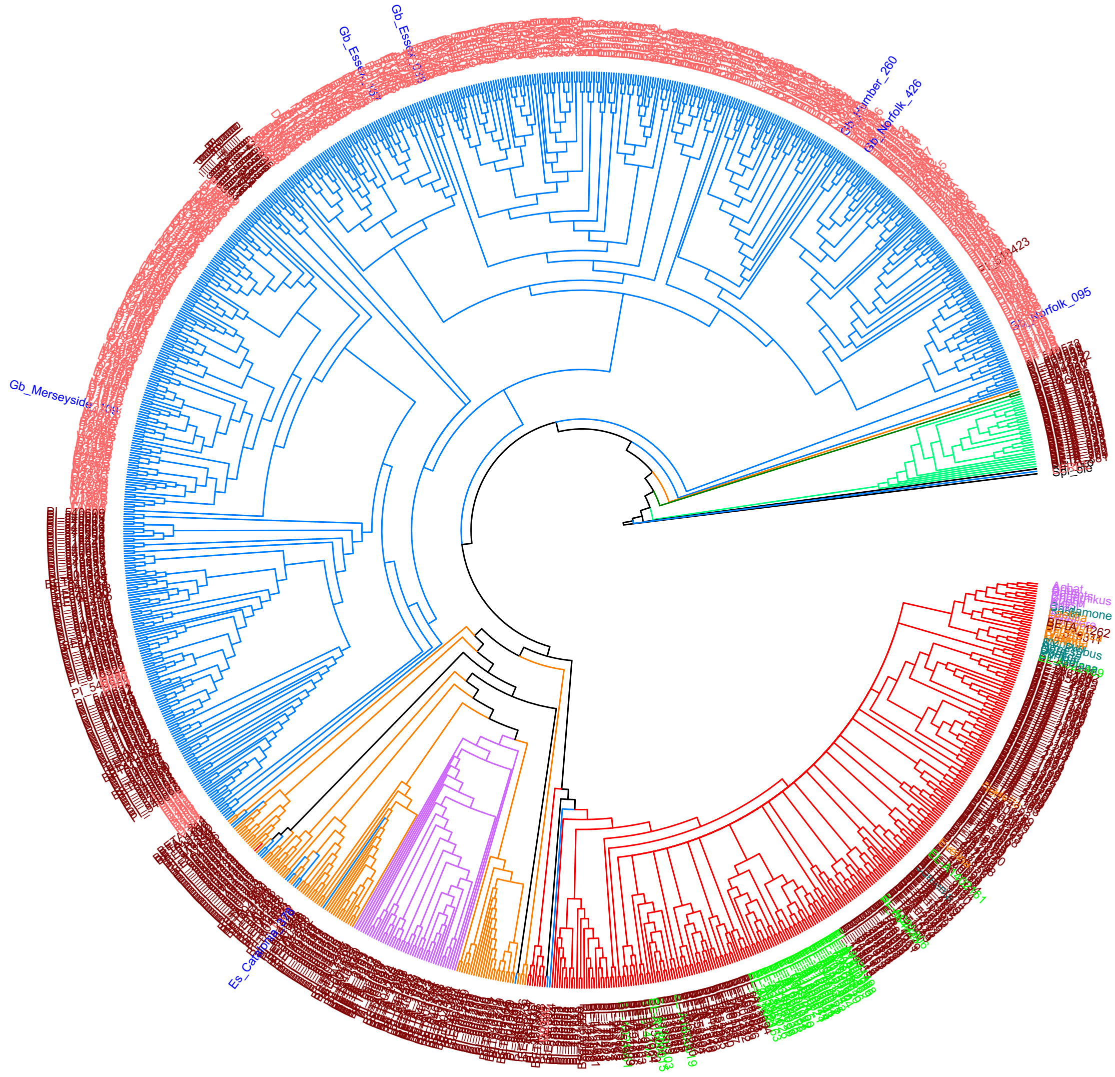


Figure 33 - Re-sequenced sea beets mostly fall in the Atlantic lineage described by Sandell *et al*..

Distances between MinHash sketches from whole genome sequencing data are represented in the format of a tree. Leaf colours distinguish 512 beets re-sequenced in the present work (pink) and 605 beets studied by Sandell *et al.* (brown), among which known sugar beet breeding programs are specified: green = East Lansing, orange = KWS SAAT SE = cyan: Syngenta and magenta = Strube Research[10]. Blue leaves indicate plants which genome has also been HiFi sequenced and assembled in the present work (**Chapter II**). Branch colours distinguish different beet species/subspecies: light green = *Beta macrocarpa*, dark green = *Beta patula*, blue = *Beta vulgaris maritima* from the Atlantic coast, orange = *Beta vulgaris maritima* from the Mediterranean coast, purple = *Beta vulgaris adanensis* and red = *Beta vulgaris vulgaris*. [10]The wild species *Pattelifolia procumbens* and the *Spinacia oleracea* species (spinach) were utilised as outgroups.

Branches of the tree to a large part recapitulate the expected pattern among sites. However, bifurcating trees are just one method of observing diversity and differentiation but unfortunately fall down where individuals are diploid, and could reasonably contain (hybrid) haplotypes from multiple sources and also where recombination operates to fuse branches among sexually reproducing individuals (introgression). Therefore, population genetic methods that describe individuals' assignment and population structure are used.

IV - B. 2. English and Danish sea beets constitute five populations

In order to explore genomic structure among wild sea beets and define populations, short sequencing reads from 161 individuals from coastal locations in east and west England as well as Denmark were mapped against the Gb_Norfolk_426 genome assembly, and variants were called. The SNPs identified were used to run the program PopCluster[198], which infers population structure by estimating the most likely number of populations (*K*) from a set of individuals, using the D_{LK2} and F_{STIS} estimators, and attributing each individual to one of these *K* clusters. For each of these *K* clusters, the clustering analysis is first done under a mixture model, and, in a second step, an admixture analysis is carried out to estimate individual admixture proportions. Both D_{LK2} and F_{STIS} values were used to estimate the most likely number of populations *K* to be 5 (**Figure 34**).

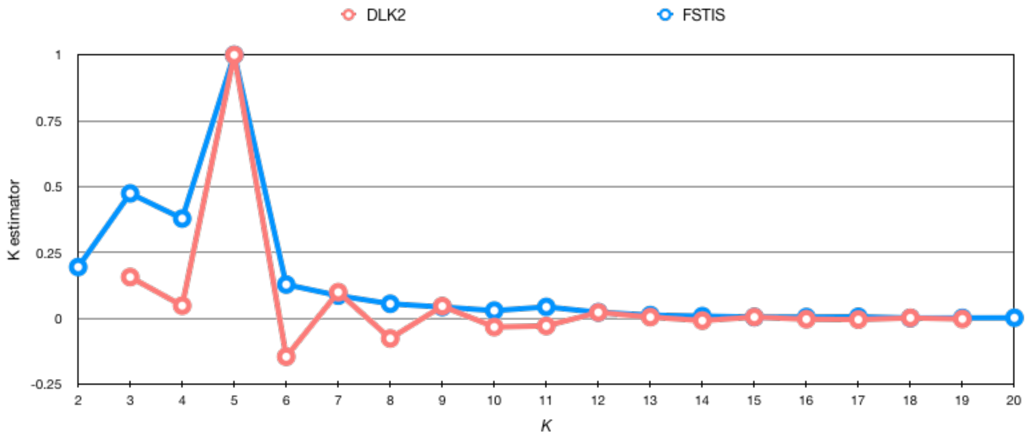


Figure 34 - DLK2 and FSTIS estimate the number of sea beet populations to be 5.

The PopCluster program was run with a *K* range from 2 to 20 (x axis). The *K* estimator represents the statistics D_{LK2} and F_{STIS} reported as a percentage.

The five populations defined by the PopCluster analysis distinguish the sea beets from Denmark on mainland Europe from, the Merseyside/Cheshire area (English west coast), and on the English east coast, the sea beets from Humberside, Norfolk, and from the Suffolk/Essex areas (**Figure 35**). These 5 populations are separated in well-defined groups, at least in the cases of Humberside, Zealand (Denmark) and Merseyside/Cheshire. On another side, while being grouped with Suffolk individuals in a single population, the Essex "sub-population" shows a lot of admixture from the Norfolk population. Actually, more individuals (93%) sampled in the county of Essex show admixture with the Norfolk population than individuals sampled in the Suffolk county (39%). This is despite the fact that Suffolk is geographically closer to Norfolk than Essex.

Each individual clusters in its sampling location except one individual sampled in the Essex county, which is classified by PopCluster in the Norfolk population. This is unlikely to be due to mislabelling of this sample, and is consistent with this large admixture from Norfolk into the Suffolk/Essex population. Also, when *K*=9, this sample is part of a small sub-population from the Norfolk population (**Figure 37 panel D**), and even in a smaller sub-population (2 individuals) when *K*=15 (**Figure S3**).

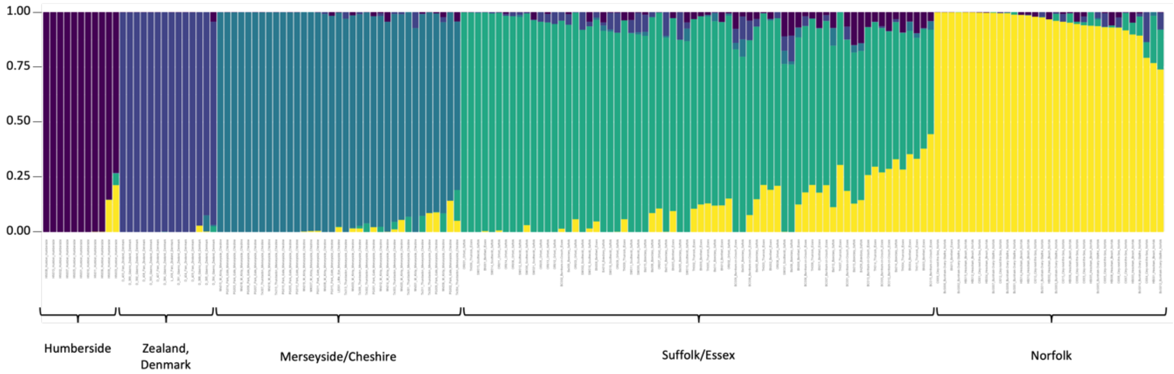


Figure 35 – Wild sea beet population structure identifies four English clusters.

The most likely number of populations was estimated as *K* = 5 (plum colour = Humberside, purple = Zealand, blue = Merseyside/Cheshire, green = Suffolk/Essex and yellow = Norfolk).

Five clusters were identified as most likely, but analysis of assignment and other clusters can provide information about where greatest differentiation exists. The first two genetic clusters that are resolved at *K*=2 are the east and west of England (**Figure 36 panel A**), and individuals part of these locations are never brought together in a same population, from *K*=2 to *K*=20 (**Figure 35**, **Figure 36**, **Figure 37**, **Figure S3**). The proper attribution to the west-coast samples to a single population starts at *K*=3 **(Figure 36 panel B**) until *K*=17 (**Figure 35**, **Figure 36**, **Figure 37**, **Figure S3**). The populations of Humber and Norfolk are the last to be resolved at *K*=4 (**Figure 36**).

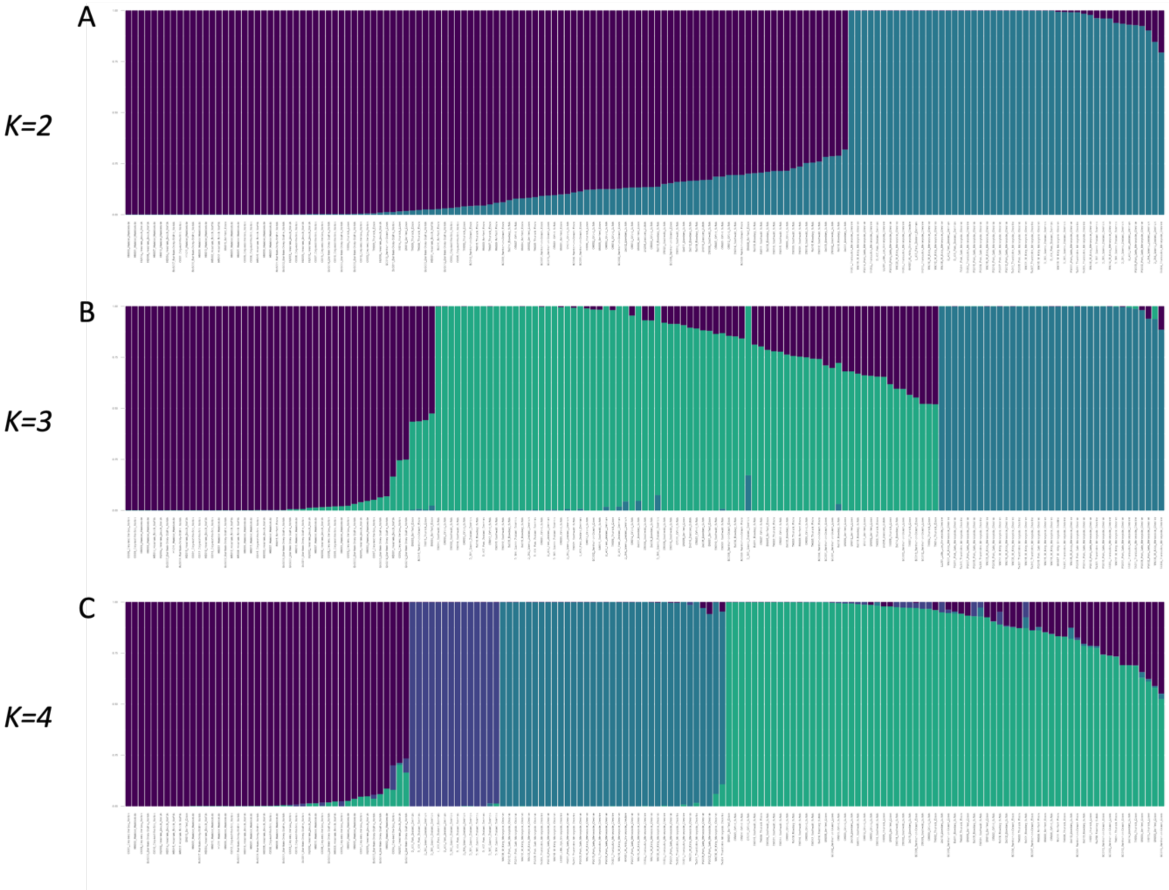


Figure 36 - Populations generated by PopCluster for *K*=2 to *K*=4.

When *K*=2 (A), populations are separated in an eastern UK group (purple) and a western UK group including the Denmark individuals (blue). When *K*=3 (B), Norfolk and Humberside are grouped in a same population (purple), Denmark, Suffolk and Essex in a second population (green), and the third population is made out of the western UK individuals (blue). When *K*=4 (C), Denmark constitutes a new separate population (dark blue).

The populations Humberside and Zealand (Denmark) remain intact until *K*=18 (**Figure 35**, **Figure 37**, **Figure S3**). Thereby, they seem to be genetically separated from the rest of the samples studied.

The population with the greatest assignment of individuals (at *K*=5), Suffolk/Essex, is the first one to be split into separate populations, when *K*=6 (**Figure 37 panel A**). In this configuration, a non-negligeable amount of admixture is observable, with more admixture from the Suffolk population in the Essex population than the other way round, perhaps indicating directionality in gene flow.

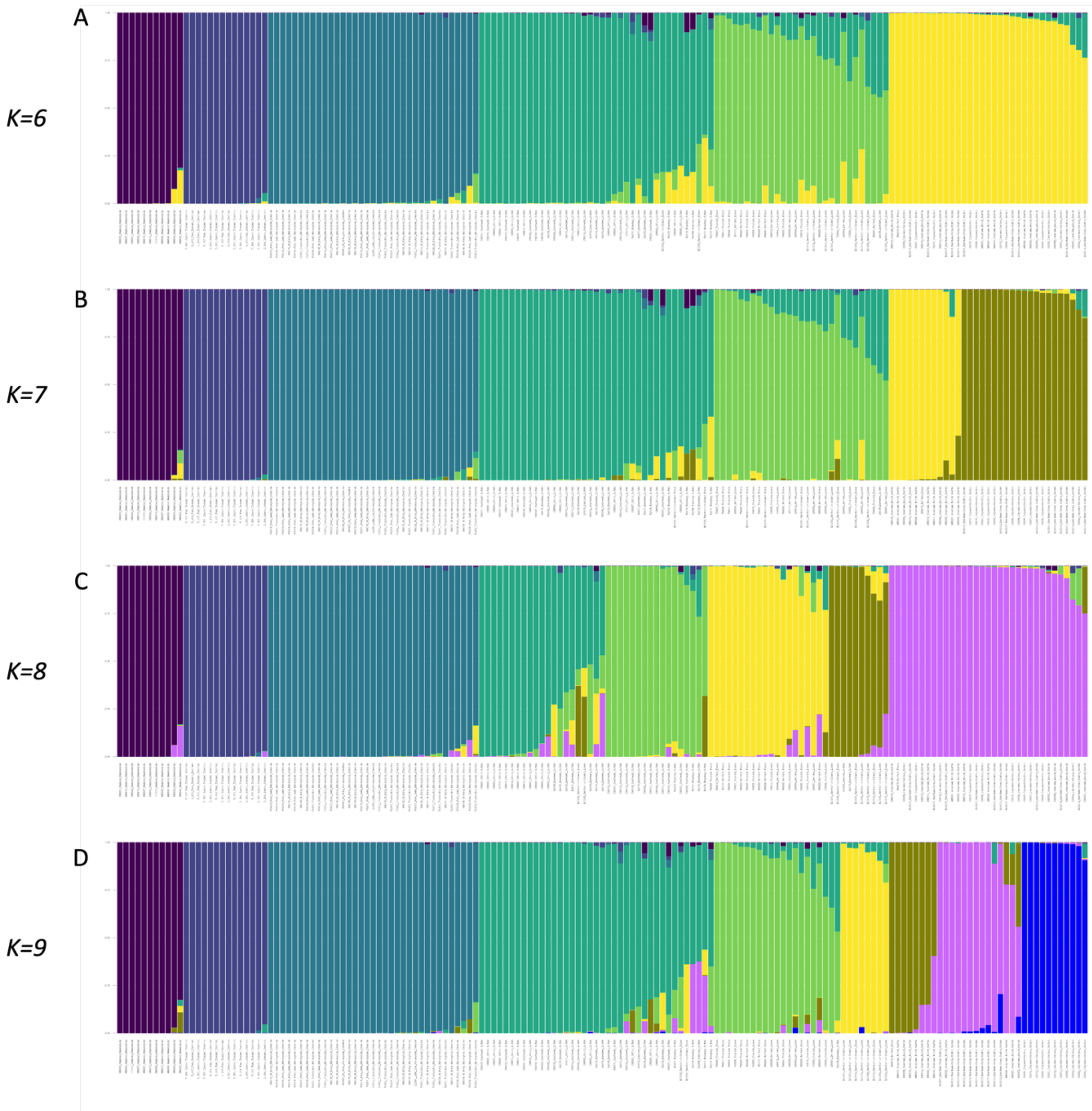


Figure 37 - Populations assignment generated by PopCluster for *K*=6 to *K*=9.

At *K*=6 (A), the groups of Suffolk (dark green) and Essex (light green) split in two distinct populations. At *K*=7 (B), the initial Norfolk population splits in two new populations (yellow and mustard). At *K*=8 (C), two populations are made out of the previous Suffolk population (dark green and light green), the Essex population is split as well in two distinct populations (yellow and mustard), and the Norfolk population from *K*=5 and *K*=6 is back to a single population (pink). When *K*=9 (D), the Suffolk group stands as a single population (dark green), the Essex group, similarly with *K*=8, is divided into two populations (light green and yellow), and the Norfolk individuals split in three distinct populations (mustard, pink and blue).

IV - B. 3. The linkage in wild sea beet is low

Using genotype data from re-sequencing reads of 147 sea beets mapped to the Gb_Norfolk_426 genome, linkage disequilibrium (LD) was analysed for the genotypes of four English populations (Norfolk, Humberside, Suffolk/Essex and Merseyside/Cheshire).

The analysis was carried out on the largest contig of the Gb_Norfolk_426 genome, using one SNP every 50 bp (maximum distance of 100 kb, **Figure 38**). The distance between two SNPs when the linkage disequilibrium drops to half of its maximum value corresponded to 1,700, 1,200, 1,500 and 1,300 bp, respectively, for the populations of Norfolk (**panel A**), Humberside (**panel B**), Suffolk/Essex (**panel C**) and Merseyside (**panel D**). Consistent with expectations of genetic diversity and effective population size, the Humberside population showed the highest LD measure (r^2 = min=0.23 and max=0.40, **Figure 38 panel B, Table S10**), and the Suffolk/Essex population showed the lowest LD (r^2 = min=0.05 and max=0.17, **Figure 38 panel C**).

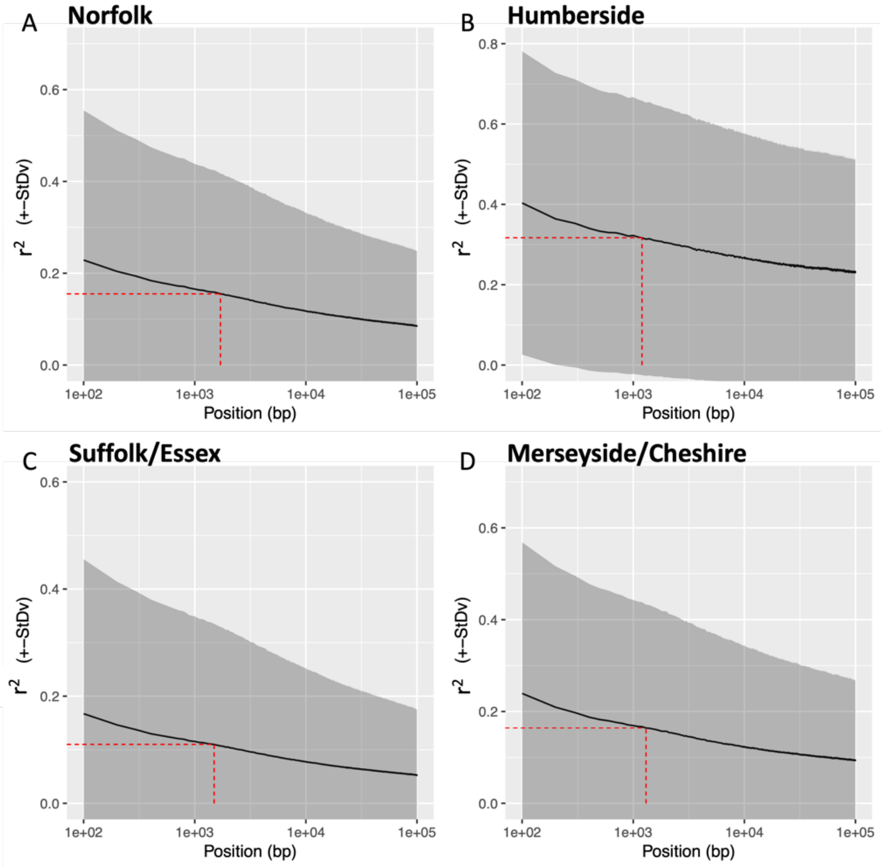


Figure 38 - The linkage disequilibrium measured in the largest contig drops quickly in sea beets.

The y axis represents the probability for two SNPs to be linked. The x axis represents, in a logarithmic scale, the distance between two SNPs, in bp. Linkage plots were done using windows increasing by 100 bp up until 100 kbp. The genotypes were thinned to 1 SNP per 50 bp. The 50% linkage for each population is close to 1,000 bp.

Secondly, to increase the granularity, the linkage analysis was run on non-thinned data over all contigs greater that 1Mb (maximum distance of 1 kb). The negative correlation between linkage disequilibrium and the distance separating two SNPs is clearly visible for each population (**Figure 39**), with a decrease in LD as the distance increases, marked with a notable drop, or LD decay, before reaching 100 bp.

The highest LD measured, for the closest SNPs, was, here again, found for the Humberside population, with an r^2 measure of 0.46 (**Figure 39 panel B**), as well as for the maximum distance separating two SNPs, i.e. 1 kb (r^2 = 0.28). On the other end of the spectrum, the Suffolk/Essex population showed the lowest LD both between SNPs within 10 bp (r^2 = 0.21) as well as at 1 kb (r^2 = 0.09) (**Figure 39 panel C**). Overall, the linkage disequilibrium is low in the four populations analysed.

The LD decay rate was measured as the chromosomal distance at which the average pairwise correlation coefficient (r^2) dropped to half its maximum value. LD decay rates of each of the four populations were estimated at ~40 bp (**Figure 39**).

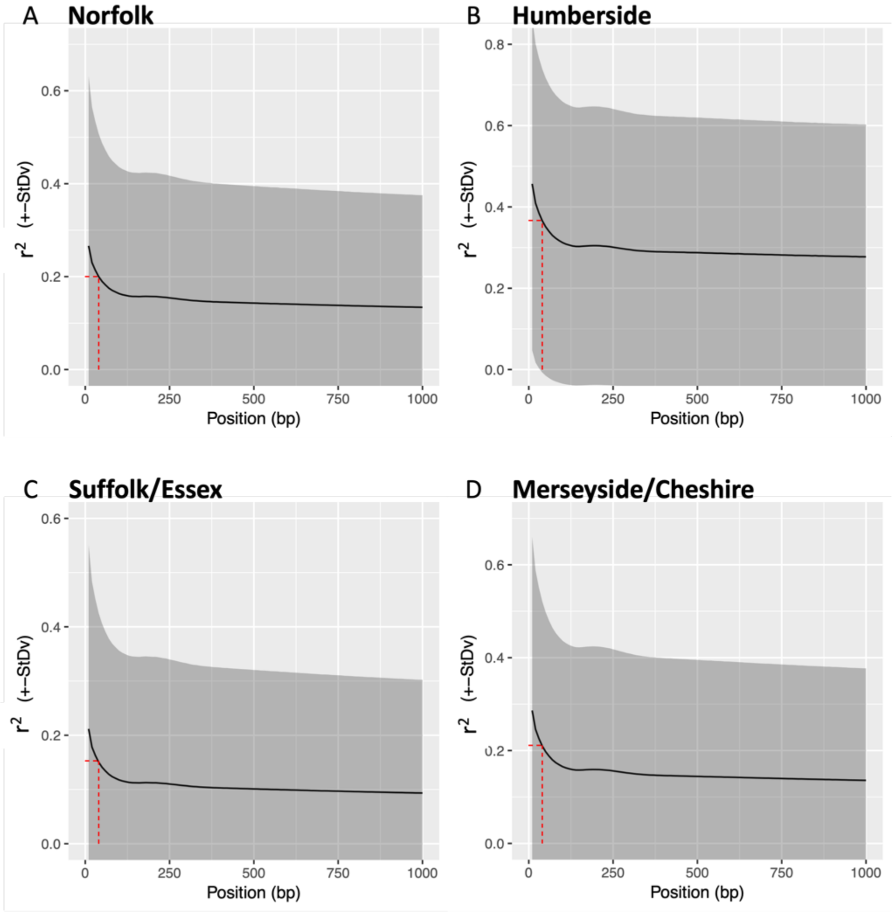


Figure 39 - The linkage disequilibrium drops rapidly in the four English sea beet populations.

The y axis represents the probability for two SNPs to be linked. The x axis represents the distance between two SNPs, in bp. Linkage plots were done using windows increasing by 10 bp up until 1,000 bp. The 50% linkage for each population is close to 40 bp.

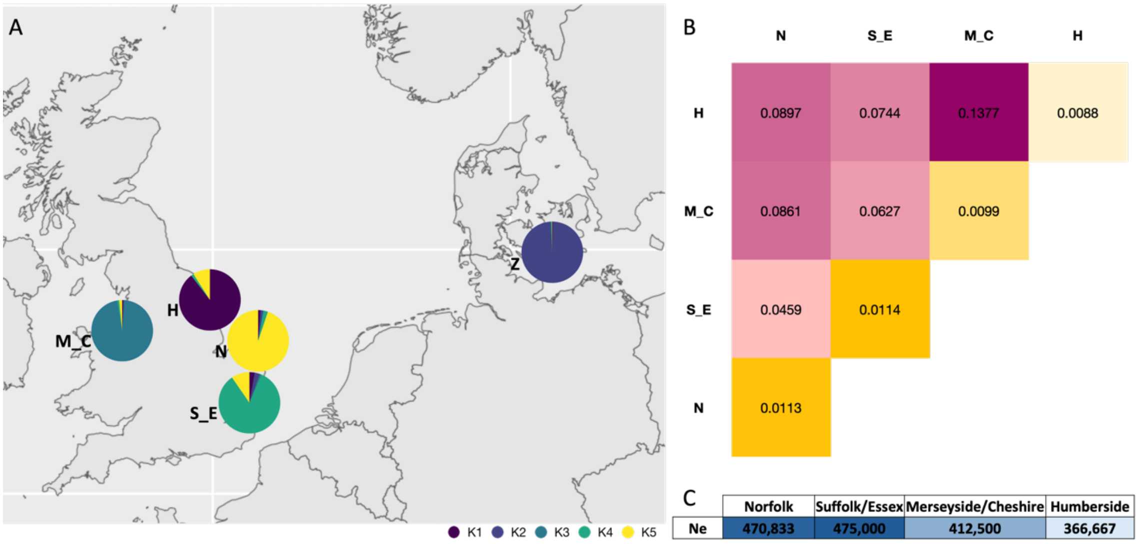
Overall, the rapid decay highlights a low linkage disequilibrium, at approximately 1 kb. This is the case both when the data was thinned to 1 SNP every 50 bp, over 100 kb, and the un-thinned data measured over 1 kb. The differences likely reflect distinctions in the window sizes, the data thinning as well as the maximum distance.

IV - B. 4. East and West England populations are differentiated and Humberside shows low genetic diversity

The admixture analysis highlights the genetic material exchange between the English eastern populations, and almost no admixture either in the Merseyside/Cheshire and the Zealand populations (**Figure 40 panel A**). The population of Suffolk/Essex is the most admixed (15.9%), followed by Humberside (10.4%), Norfolk (5.2%), Merseyside/Cheshire (3.6%) and Zealand (1.4%). The largest part of admixture comes from Norfolk, both for Humberside and the Suffolk/Essex populations.

Genotype data from 147 individuals from the four English populations were taken to measure genetic diversity and population differentiation (**Figure 40**). The nucleotide diversity doesn't differ to a large extent between the four English populations, but nevertheless highlights a difference between the east-coast populations: the Suffolk/Essex or the Norfolk populations, harbouring the highest diversity (nucleotide diversity = 0.0114 and 0.0113, respectively) and the population of Humberside harbouring the lowest diversity (nucleotide diversity = 0.0088; **Figure 40 panel B**). Measuring the nucleotide diversity of a population allows, together with the estimated mutation rate, to estimate its effective population size, or the number of individuals forming the breeding population which would be needed to explain the allelic diversity encountered in the population. The Suffolk/Essex population is then the largest one, with an effective population size of 475,000 individuals, followed closely by the Norfolk population counting 470,833 individuals. Humberside is the smallest population, with 366,667 individuals and, on the other side of the country, the Merseyside-Cheshire has an effective population size comparable to the east side, counting 412,500 individuals (**Figure 40 panel C**).

Pairwise genetic differentiation measures range between 0.0459 (Norfolk vs Suffolk/Essex) and 0.1377 (Humberside vs Merseyside/Cheshire) (**Figure 40 panel B**). The two most differentiated populations are then Humberside on the east coast and the Merseyside/Cheshire population on the west coast. Interestingly, there is more differentiation between Humberside and Norfolk populations, both on the east coast, than between the Norfolk and the western population (Merseyside/Cheshire). It is, moreover, noticeable that the Suffolk/Essex population shows the lowest indexes of differentiation with the other populations.



	N	S_E	M_C	H
H	0.0897	0.0744	0.1377	0.0088
M_C	0.0861	0.0627	0.0099	
S_E	0.0459	0.0114		
N	0.0113			

	Norfolk	Suffolk/Essex	Merseyside/Cheshire	Humberside
Ne	470,833	475,000	412,500	366,667

Figure 40 – Population admixture among the five sea beet populations, nucleotide diversity and gene flow from 147 individuals.

(A) Admixture for the population of wild sea beets estimated by PopCluster at *K*=5. (B) Matrix of nucleotide diversity (shaded yellow) and F_{ST} (shaded pink) values within and between the five populations estimated by PopCluster. N = Norfolk, S_E = Suffolk/Essex, M_C = Merseyside/Cheshire, H = Humberside. (C) Effective population size per English population. N_e is measured as the ratio between the mean nucleotide diversity and four times the estimated mutation rate in plants, 6.10^{-9} (from Schultz *et al.*, 1999).

IV - B. 5. Resistance genes are more polymorphic than other genes

Whole genome statistics provide estimates of levels of population genetic diversity and differentiation that obscure more local signals operating around genes that might indicate the presence of natural selection. The LD measured in sea beet is low, and this means that gene-based analyses can appropriately be conducted. Genotypic data of 21,273 genes from the 147 English individuals sampled in their natural environment and studied for population structure were analysed to calculate the nucleotide diversity, or π, and the fixation index, or F_{ST}.

Nucleotide diversity is used at the genome scale to approximate the effective population size. However, at a local level the presence of balancing selection operating to preserve diversity at resistance genes would be predicted to preserve variation (on average) relative to other genes[199]. Nucleotide diversity is significantly greater in NLR genes than when considering other genes on average (**Figure 41 panel A**). While nucleotide diversity in NLRs is also greater in three out of four individual populations (one-sided Wilcoxon-Mann-Whitney: Humberside, U=2232344.5, *P*=0.0189; Merseyside, U=2177741.0, *P*=0.0754; Norfolk, U=2243963.5, *P*=0.0135 and Suffolk/Essex, U=2288100.0, *P*=0.0032), only Suffolk/Essex remains significant after Bonferroni correction (α=0.0125). It is interesting that it is the western population (Merseyside/Cheshire) that shows no difference in nucleotide diversity between NLR genes and non-NLR genes, even prior to correction (**Figure 41 panel B**).

Overarching signals of, for instance, balancing selection, may be visible when observed over genes treated as groups because signals can outweigh the variance present among individual genes. Caution should be taken when considering signal to noise at the individual gene level as, for a given gene, noise may well overcome the signal. However, with this in mind, in **Chapter III**, five candidate NLRs for rust resistance were identified and here, these orthogroups are assessed for their adherence to the expected signals. There are two categories of NLR orthogroups: two NLR orthogroups with a high max *k*-mer association score in two controlled inoculation trials (English rusts) in addition to a natural rust inoculation trial (**Figure 41 panel B**, genes in blue), and three NLR orthogroups displaying a high mean association score through their associated *k*-mers (**Figure 41 panel B**, genes in red). For clarity, these orthogroups are hereafter referred as high-score orthogroups and high-mean orthogroups. One of the high-score orthogroups (OG0000048) displays a higher nucleotide diversity than the majority of the NLR genes, in all populations. The other high-score orthogroup falls within boxes that define the main distribution (50% of the data) in all populations except Humberside. High-mean orthogroups display a lower nucleotide diversity value than the median nucleotide diversity value of the other NLR genes. Of these three orthogroups, OG0000124 displays the lowest score in each population, OG0000122 the second lowest score and OG0000123 shows the least low value (**Table S11**).

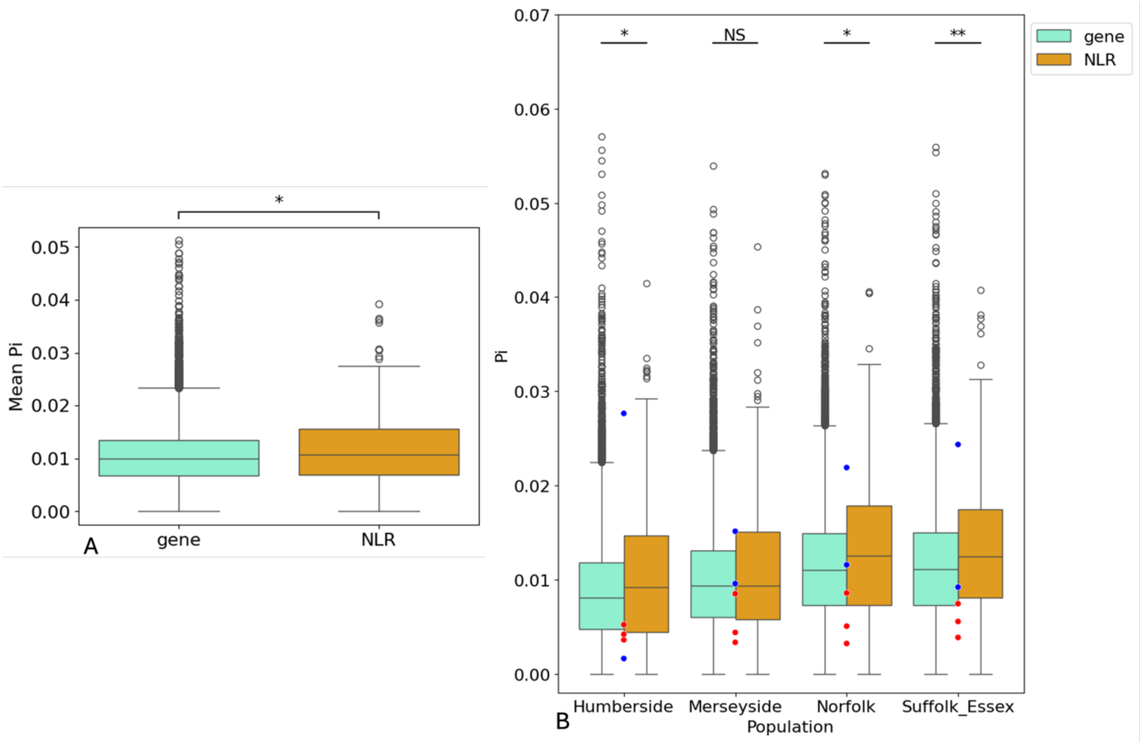


Figure 41 - Genetic diversity is overall higher in NLRs than in other genes.

(A) Mean nucleotide diversity (Pi) measured across all the four English populations, is significantly higher for NLRs (orange) than for other genes (green) (one-sided Wilcoxon-Mann-Whitney, U = 2253563.0, *P* = 0.0100). (B) Pi measured separately for each of the four English populations was higher for NLRs than for other genes for three out of four populations (asterisks: NS: 0.05<*P*≤1; *: 0.01<*P*≤0.05; **: 0.001<*P*≤0.01) but only Suffolk/Essex remains significant after Bonferroni correction (α=0.0083). The rust resistance association genetics NLR orthogroups (see **Chapter III**) are coloured in red (OG0000122, OG0000123 and OG0000124) and blue (OG0000043 and OG0000048).

The pairwise fixation indexes can also be calculated for each gene, and in that case the signal consistent with balancing selection (as expected for many resistance genes) corresponds to reduced differentiation on average (**Figure 42**). However, while the median of the F_{ST} is lower for NLR genes (0.074) than for non-NLR genes (0.078) these values are not significantly different between resistance and other genes (**Figure 42 panel A;** one-sided Wilcoxon-Mann-Whitney test: U = 1989245.0, *P* = 0.2202). Again, when looking in more detail at the results in pairwise population comparison, there is a tendency for NLR genes to be less differentiated between populations than other genes but these results are not significant after Bonferroni correction (see **Table 2**; **Figure 42 panel B**).

Interestingly, when considering the five candidate rust resistance NLRs highlighted in **Chapter III,** two different patterns emerge. First, within the high-value orthogroups, OG0000048 shows high F_{ST} values compared to the rest of the distribution, specifically and consistently when compared between eastern and western populations (**Figure 42 panel B, blue**). Second, in the high-mean orthogroups, with the exception of the Humberside-Suffolk/Essex comparison, OG0000122 and OG0000124 appear in the lower end of the main distribution (**Table S12**; **Figure 42 panel B, red**).

Populations	**Gene F_{ST}**	**NLR F_{ST}**	**Test statistic (U)**	***P* value**
Humberside – Merseyside	0.122	0.107	2006049.5	0.2828
Humberside – Norfolk	0.090	0.094	2136479.5	0.8298
Humberside – Suffolk/Essex	0.063	0.054	1871462.5	0.0157
Merseyside – Norfolk	0.075	0.069	1912568.0	0.0475
Merseyside – Suffolk/Essex	0.053	0.046	1928656.5	0.0693
Norfolk – Suffolk/Essex	0.038	0.041	2152146.0	0.8722

Table 2 - One-sided Wilcoxon-Mann-Whitney test results for pairwise genetic differentiation (F_{ST}) between resistance genes (NLR) and all other genes among the four English beet populations.

NLRs are not significantly less differentiated after Bonferroni correction with an alpha value at 0.083. Comparisons between east and west are shaded.

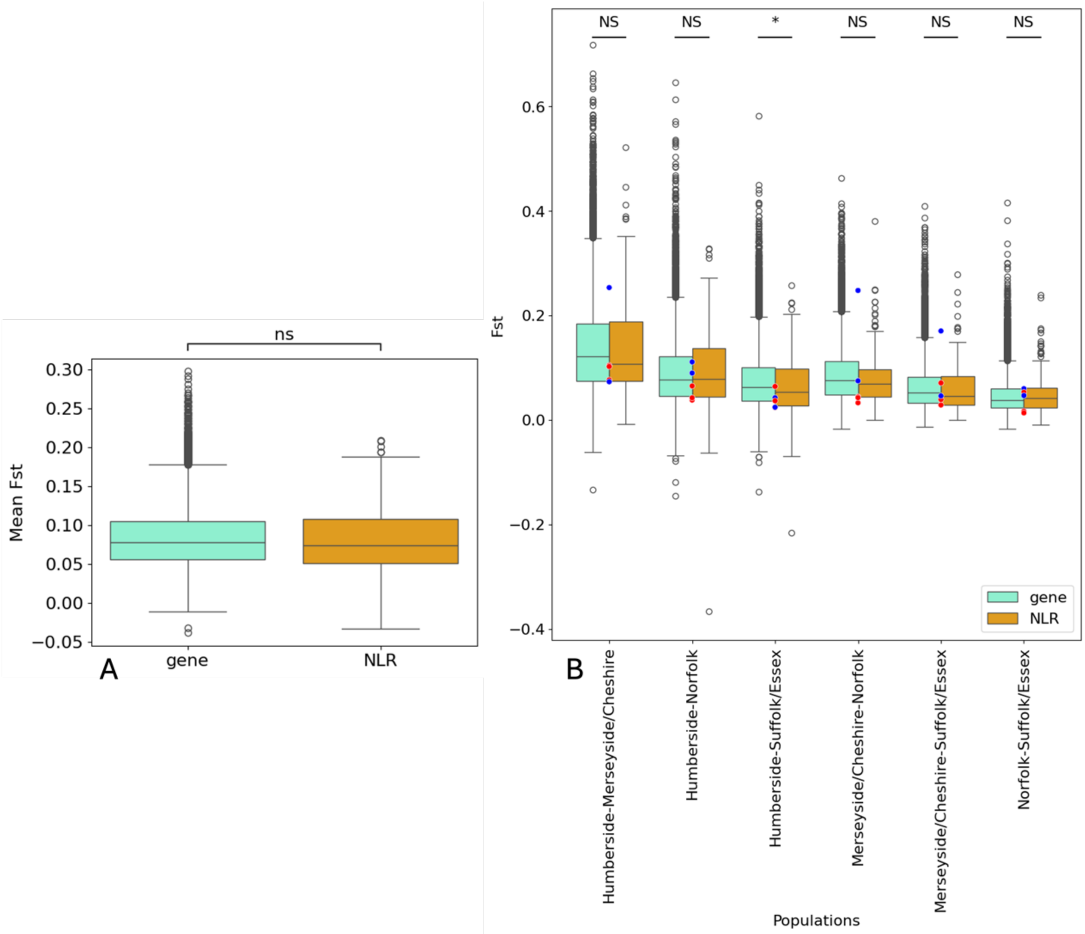


Figure 42 - The fixation index is overall lower in NLRs than in other genes.

(A) Mean F_{ST} measured from pairwise comparisons between the four English populations, either for NLRs (orange) or for other genes (green). (B) Pairwise F_{ST} measured between each of the four English populations, either for NLRs (orange) or for other genes (green). The well-performing orthogroups in association studies involving English rust samples (see **Chapter III**) are coloured in red (OG0000122, OG0000123 and OG0000124) or blue (OG0000043 and OG0000048) (asterisks: NS: 0.05<*P*≤1; *: 0.01<*P*≤0.05).

IV - C. Discussion

At a European scale, genomic distance between sea beets sampled in the present work confirmed the assignment to Mediterranean and Atlantic genetic groups, as first described by Sandell *et al.*[10]. The present study likely represents the largest genome-wide description of germplasm from the Atlantic lineage, given that both sugar beet and previous sea beet sequencing predominantly focus on diversity from the Mediterranean population[10]. On the one hand, from a breeding context, it might seem counterproductive to describe the genetic diversity of the Atlantic population, given that the genetic background of sugar beet is from the Mediterranean region. However, pathogen prevalence is linked to climate, and sugar beets are predominantly grown in Russia, France, and Germany, and contribute for 50% of the sugar consumption in the UK. Rust specifically favours humid conditions around 15–22 °C and is even supressed at temperatures above 26 °C[26]. Therefore, it is reasonable to assume that adaptation of the genes for resistance to rust have evolved, or been maintained, in those places where beets are now cultivated.

To summarise the results of this work analysing the genetic diversity of this large panel of Atlantic sea beets, five populations were defined: one containing individuals sampled in Denmark; a second population sampled on the west English coast, in the Cheshire county; and three populations on the east English coast, grouping individuals sampled in Humberside, Norfolk and Suffolk/Essex, from North to South. English populations showed an overall low linkage disequilibrium, and a consequent genetic diversity, with effective population sizes ranging from 367,000 (Humberside) to 475,000 (Suffolk/Essex) individuals. When observing differentiation between populations, the largest population (i.e. Suffolk/Essex) showed the lowest differentiation in pairwise comparisons with every other population, and the largest index was retrieved between Humberside and the Merseyside populations. Further analyses were carried out comparing resistance genes (NLRs) and the other genes, and showed a larger nucleotide diversity in NLR genes, except in the western population of Merseyside. Finally, the five resistance genes highlighted in the previous chapter (see **Chapter III**) showed contrasting, but potentially interesting patterns in their F_{ST} and nucleotide diversity measures. The present discussion aims to explore these results and propose an interpretation of the nucleotide diversity and fixation indexes measured for the different populations, involving the potential role of marine currents in shaping the structure of populations. Finally, the particular population genetics signals observed for one of the resistance genes lead to an opening regarding the promising potential use of these measures to identify interesting candidates.

IV - C. 1. English sea beet population structure, gene flow and ancestry analyses

Given the predominance of Mediterranean genotypes in germplasm collections, population genomic analyses of the wild Atlantic population perhaps represent the most cost-effective way to determine its potential benefit to sugar beet breeding. These methods both describe gene diversity as well as potentially reveal resistance associations (**Chapter III**) adapted to the environment beets are cultivated in. Furthermore, they reveal how diverse the eleven sea beet assemblies (see **Chapter II**) are in the diversity tree. This is a promising result when put in the context of the sea beet pan-genome generation, reflecting the solid basis this genetic collection is built on.

Studying the diversity among the sea beets sampled for the association project, more specifically analysing their population structure, defined five populations. These populations can be described geographically by distinguishing sea beets coming from: the Wirral peninsula on the west of England, away from commercial sugar beet cultivation, and moving from north to south on the east of England, Humberside, Norfolk, and Suffolk and Essex, and then the Danish Zealand region. Interestingly, the Humberside population stands out from the other populations in multiple aspects. It was noted that this population tended to be more susceptible to rust (see **Chapter III**) and clusters here into a single population (over a large range of *K* values). Moreover, compared to the four other populations established, Humberside is the smallest in terms of genetic diversity and, consequently, also has the largest linkage disequilibrium.

Marine currents along the east coast of England (**Figure 43**) provide insight into the diversity and differentiation encountered in this Humberside population. Indeed, the surface currents, involved in beet seed dispersal, move south down the coast from Humberside to East-Anglia (Norfolk, Suffolk, Essex), and there is no counter-current in direction toward Humberside. It is understandable, then, that genetic material can be exchanged through seed travel from Humberside to Norfolk and Suffolk/Essex, and less so in the other direction. This isolates the Humberside population from the other populations analysed.

The population encompassing counties of Suffolk and Essex, on the east coast, shows the largest effective population size, the lowest linkage disequilibrium, the lowest genetic differentiation with the other populations, as well as the highest level of admixture. This is again compatible with the North Sea currents (**Figure 43**), as the Suffolk/Essex population is likely to receive seeds from the two northern populations (as well as with mainland Europe), explaining the great nucleotide diversity and low differentiation with those populations. Intriguingly, it shows a large level of admixture with Norfolk, and this is especially the case for the samples collected in Essex rather than in Suffolk. This result is surprising as the Essex county is further in the south compared to Suffolk, when using Norfolk as a referential.

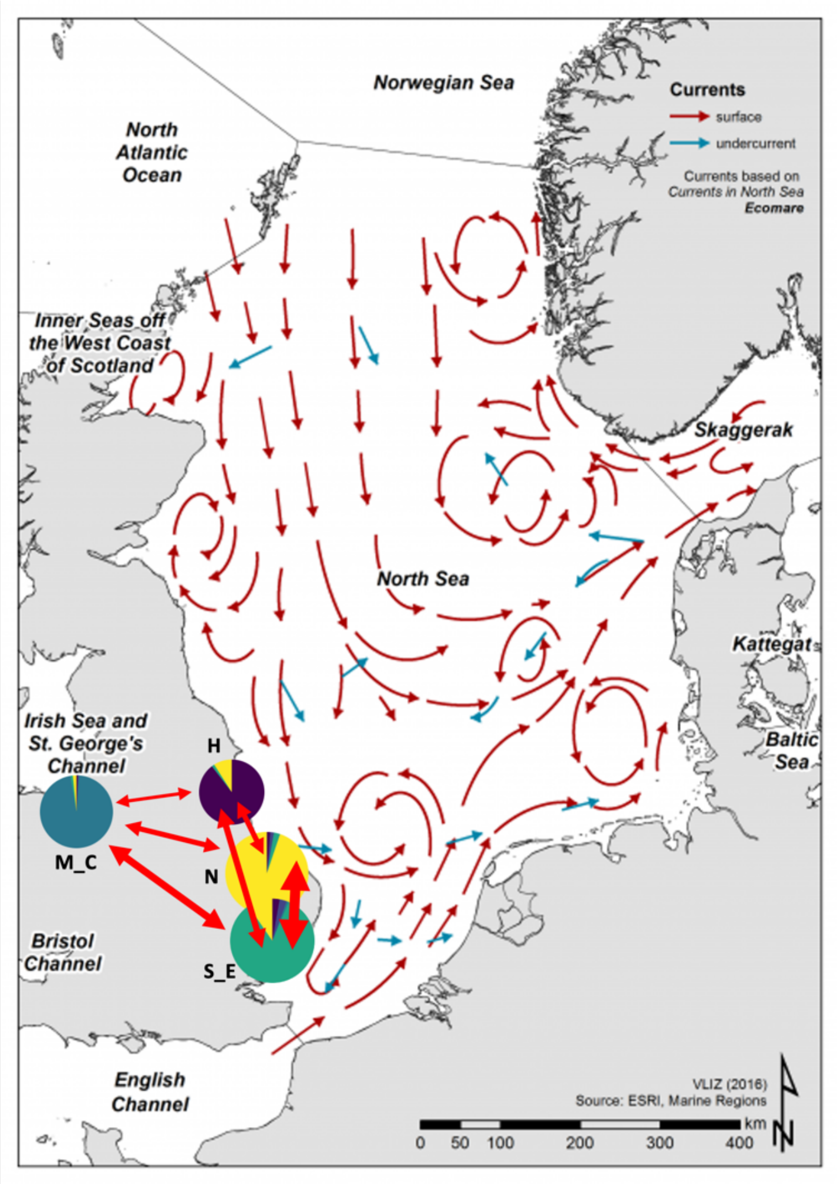


Figure 43 - Currents of the North Sea.

Population structure of East Anglian sea beets could be influenced by the marine currents, with northern currents moving southwards along the coast, hampering seed movement northwards towards Humberside population. Pie charts' size is scaled to represent the effective population size, and arrow width represents gene flow (inversely proportional to F_{ST}) between populations. H = Humberside, N = Norfolk, S_E = Suffolk/Essex, and M_C = Merseyside/Cheshire. Adapted from the original image created by Nathalie De Hauwere[200].

Marine currents would have shaped the population structure of east England sea beets, with the Norfolk population being the founder population. This would explain the large level of admixture encountered in Humberside and the Suffolk/Essex populations, and the unidirectional sea current along East-Anglia would explain the reduction in Humberside's genetic material over time due to isolation. Moreover, the samples from Essex showing more admixture from Norfolk, it could be possible that the colonisation of the Suffolk/Essex territory was done from south to north. Further investigation, utilising software that estimates isolation and bi-directional migration rates (IMa3[201]) and ancestral effective population (MSMC[202]) sizes would provide further information on the population genetics of this system as well as the split of Atlantic and Mediterranean lineages.

In regard to the patterns of genetic differentiation between the four English sea beet populations defined here, the largest fixation index was measured between the populations of Merseyside/Cheshire on the west coast and of Humberside on the east coast. Moreover, there is more differentiation between the populations of Humberside and Norfolk, which are geographically close, than between populations from east and west coasts. This first reflects the geographical constraints to genetic exchange between east and west coasts and, secondly, this could corroborate the potential isolation of the Humberside population, harbouring among the highest measures of F_{ST} with every other population. The high genetic differentiation between eastern and western coastal populations is likely influenced by gene flow between cultivated sugar beets and eastern sea beets, which does not occur on the west coast. While gene flow from crops to wild populations can lead to a reduction in genetic diversity, the high diversity observed in the Norfolk and Suffolk/Essex populations suggests that this gene flow has instead contributed to genetic enrichment. This enrichment may be due to the introduction of novel alleles into the wild beet genomes. On the other end, the Suffolk/Essex population, on the east coast, shows the lowest levels of differentiation with the other populations, supporting a high gene flow.

IV - C. 2. Genetic diversity in English sea beets

Linkage disequilibrium has been found to be very low in the four English populations of sea beets, with a 50% decay around 1 kb. This low LD in sea beet is very interesting when put in regards with the association analyses for beet rust resistance conducted in **Chapter III**. Indeed, unlike LD-based association analyses conducted on breeding material and using genetic markers, this low LD in the wild is interesting for association analyses as the associated loci can be considered as single genes. In the present case, this gives clues about the signal observed in three physically linked loci (OG0000122, OG0000123 and OG0000124), which are likely to be all three important for rust resistance, instead of only one being associated and the two others identified due to LD. LD estimates differed between the two methods used, likely because of the window sizes, SNP resolution and the data used. The long range LD analysis was done only on contig 1, whereas the 1kb analysis was done on the whole genome, and will have accounted better for lower linkage regions [203].

The low linkage encountered in wild beets contrasts with linkage observed in crops. For example, a study conducted on re-sequencing data from hundreds of rice landraces measured a LD decay rate above 100 kbp (123 kbp for *O. sativa ssp. indica* and 167 kbp for *O. sativa ssp. japonica*)[196], and another study on lettuce measured the LD decay rate at approximately 200 kbp[194]. This high linkage in crops is interesting for the conduction of marker-based association analyses.

The highest linkage disequilibrium was found in the Humberside population, and the lowest in the Suffolk/Essex population. These observations are consistent with expectations based on their effective population size and, thereby, with the diversity that they harbour; Humberside showing the lowest nucleotide diversity and Suffolk/Essex the highest. The effective population size measured for the wild beet (ranging from approximately 367,000 to 475,000) is larger than measured in any cultivated beets: approximately 16 times larger than for chard (Ne ≃ 25,000), 40 times larger than for the sugar and fodder beets (Ne ≃ 10,000), and approximately 66 times larger than for the table beet (Ne ≃ 6,000)[204]. This is consistent with the bottleneck induced by domestication, reducing the genetic diversity in crop by selecting for breeding traits. However, more detailed analyses of the effective population size (and directional migration rates) should be done using software such as IMa2[201] (or even of IMa3[201]).

Measures of variation in population genetics have been investigated in other wild species. For example, the wild tomato species *Solanum chilense* and *Solanum peruvianum* have been shown to have a high nucleotide diversity (ranging from π=0.55-1.10 or π=0.78-1.29, respectively) and a rapid LD decay[205], implying high effective population sizes (Ne = 504,000 and 687,000, respectively) and recombination rates. Moreover, this analysis of four populations per species, sampled in Peru and Chile, showed moderate levels of population differentiation, with an average F_{ST} of 0.20. While being large in wild beet and tomato, the nucleotide diversity estimates are slightly lower in the present study. The low LD measure in sea beets is consistent with its outcrossing mating system, unlike its relatives *B. macrocarpa*, *B. patula* and *B. v. adanensis*, being self-compatible[206]. Compared to wild tomato populations, sea beet populations show lower F_{ST} measures (between 0.05 and 0.14).

IV - C. 3. Can population genetics inform association genetics and breeding decisions?

The present study revealed a nucleotide diversity larger in NLR genes than when considering non-NLR genes. Immune genes are known to maintain high levels of genetic diversity, and this had been observed in a population study of Peruvian and Chilian wild tomato species *Solanum chilense*[98]. New mutations in NLR genes can facilitate the recognition of novel effectors and the deployment of the resistance machinery. These novel mutations are preserved by natural selection when they indeed facilitate the recognition of novel pathogen genotypes[134]. The preservation of genetic variation at resistance genes may be via balancing selection: that maintains alleles on account of how rare they are (negative frequency dependant selection) or because heterozygous individuals are resistant to a broader array of pathogens; or by directional selection that varies over space and time. Whereas genetic variation at most other genes is selected against by purifying selection[207,208].

In English sea beet, nucleotide diversity that is larger in NLRs is less marked in the western population. East coast wild populations grow in the presence of cultivated sugar beets, which could have a non-negligeable impact on the genetic diversity encountered in NLRs, either by crossing[184], or by natural selection from crop adapted pathogens[209]. Indeed, rust can infect all beet species, and was shown to evolve crop-specific effectors[209]. Thereby, sea beets must adapt to this potential larger reservoir of pathogen effectors, explaining the largest diversity in their NLRs on the east coast.

Another consequence of balancing selection on resistance genes is known as a higher effective migration rate[210,211]. The impact of selection for rare alleles, or high heterozygosity, is such that if novel alleles are introduced to a population, they will be disproportionately preserved by natural selection, on account of them being rare. This has the effect that it reduces the differentiation at resistance genes, in view of the benefits they bring to this population. However, this wasn't observed in the present study, either overall, or for most of the pairwise analyses carried out. The combination of increased nucleotide diversity, without reduced genetic differentiation, is more consistent with pathogen selection that varies spatially in time and space (different pathogens in different places)[208]. However, the trends in the signals (that are not significant) suggest this may be more a lack of statistical power.

Given the broader signals observed across NLRs as a group, attention turns to the specific signals associated with the five genes highlighted in the **Chapter III**. The variance associated with taking single gene statistics means that observations at this level should be taken with caution. However, the exploration of the data at this level is important to understand whether these signals can be used to inform on candidates' characteristics in a breeding panel. Population genetic measures were shown to be different depending on the genes studied. Two "high-mean" orthogroups (OG0000122 and OG0000124) tended to show nucleotide diversity and differentiation on the lower side between populations. If population genetic signals were to be used to predict gene performance, it would be consistent with what is observed for these genes that are on the lower end of the NLR polymorphism distribution, more conserved and shared between populations, to provide broad rust resistance across different locations. An important next step would be to look at the signal of non-synonymous polymorphism in these genes.

Among the "high-score" rust resistance orthogroups, the orthogroup OG0000048 shows high F_{ST} values (outliers) compared to the rest of the distribution, but only when comparing eastern and western populations. This is a particularly interesting and clearer result. This high gene differentiation between an area relatively free from beet cultivation and an area rich in beet cultivation is consistent with the respective absence/presence of crop-adapted pathogen effectors, to which specific resistance is evolving. There are two non-mutually exclusive hypothesises for this observation. First, that gene flow from the crop has introduced an NLR into the wild population at appreciable levels. If this is the case, the NLR will already be present in the crop. Second, if not already part of the crop, it is possible that this resistant gene is being selected for by pressure from the increased prevalence of crop pathogens in the East. This is a particularly interesting population genetic signal, exactly the sort of signal that provides useful information for breeders.

Due to the differences mentioned above, the “high-score” and “high-mean” genes could provide different defence strategies: the first ones may recognise population-specific effectors while the second ones may provide resistance against broader-range pathogens. This is intriguing and counter-intuitive, as the OG0000048 orthogroup (potentially providing population-specific resistance) was identified in three different association trials involving English rust.

The population genetics patterns presented in this work are especially visible for this “high-score” orthogroup OG0000048, which stands out when displaying outlier π and F_{ST} values in some analyses. This orthogroup is highly polymorphic, allowing a rapid adjustment to the emergence of new pathogen effectors, and is also significantly differentiated between locations. Thereby, this work provides optimistic insights about the use of population genetic measures to screen resistance genes and highlight potential candidates.

IV - D. Material and methods

To generate the phylogenetic tree, paired-end sequencing reads from 605 wild and cultivated beet genomes were downloaded from the BioProject PRJNA815240 of the NCBI SRA database. All samples (from Sandell *et al.*’s study [10] and the present work) with a sequencing depth above 5.5x were downsampled with seqtk (version 1.0, (https://github.com/lh3/seqtk)) to reach a sequencing depth of 5.5x. 512 sea beets were analysed out of the 520 re-sequenced ones from the present work, as 8 low-coverage individuals were removed from the analysis. For tree rooting, Illumina paired-end sequencing reads from the *Spinacia oleracea* and *Patellifolia procumbens* species were downloaded from the ENA database, respectively with the run accession numbers SRR869666 and SRR10224874, and downsampled to reach a depth of 5.5x with the seqtk tool. The mashtree program[147] (version 1.4.6) was used to measure genomic distances between the samples. 1,000 trees were generated, running 10 analyses with the mashtree_bootstrap.sh script and the options --reps 100 --numcpus 12 --tempdir temporary_directory -- --min-depth 2 --kmerlength 21 --sketch-size 10000. The 1,000 trees were then combined into a consensus tree with the IQ-TREE software[212] (version 2.2.2.2), using the options -nt AUTO and -con. The consensus tree was rooted on the spinach sample with the root() command from the ape package[148] (version 5.8) in R (version 4.4.1)[213]. Finally, colours were added to the tree with FigTree (version 1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/)).

SNPs were called from Illumina sequencing reads (**Chapter III**) mapped to the genome assembly Gb_Norfolk_426. BWA mem[214] (version 0.7.7) was used to map reads and SAMTOOLS[215] (version 1.5) and BCFTOOLS (version 1.3.1) were used to sort and remove duplicate reads and mpileup (-a DP,AD) to call variants (bcftools call -m).

For the population analysis, VCFtools[216] (version 0.1.13) was used to remove indels (due to frequent misidentification of errors as indels), filter SNPs to an individual minimum depth of 2, maximum depth of 40 and a minimum genotype quality of 30. SNP sites with more than two alleles were excluded as probable errors. Finally, sites that were missing in 30% or more individuals were also removed. The VCF file was thinned to keep 1 SNP every 10 kb.

The PopCluster program[198] (version 1.2.0.0) was used to estimate the number of populations, analysing 161 wild sea beets. The initial set of 520 short read sequenced sea beet genomes has been refined in order to: remove the very low depth samples (< 3x), remove any individual sampled in a site which would contribute to less than 5 individuals in the analysis, and remove any pseudo-replication problem. For this last condition, regarding the English samples, only parental individuals have been kept in the analysis (147 individuals). In the case of the mainland European samples, a single seedling sampled from each mother has been kept in the analysis (14 individuals). After this refining step, only English and Danish sea beets were studied. PopCluster was run with the following parameters: 161 individuals, a number of 62,442 loci, a weak scaling, 2 and 20 as the minimum and maximum *K*, respectively, 10 replicates per run, and the admixture model. The *K* estimation chart was generated with Numbers, and the bar chart representing the samples split into different populations was generated using *R*.

In order to estimate the level of linkage between genotypes (unphased) in each English population, VCFtools (version 0.1.13, parameter --geno-r2) was run at two scales: first, on data thinned to 1 SNP per 50 bp with a maximum distance of 100kb between genotypes. Second, on unthinned data, with a maximum distance of 1kb between genotypes. In the case of the thinned data, the analysis was conducted on the largest contig of the Gb_Norfolk_426 genome. The mean r2 was returned from each of these datasets in windows of 100bp (100kb maximum) and 10bp (1kb maximum).

Allelic data from 147 re-sequenced English individuals (see **Chapter III**), whose leaf material was collected directly in their natural environment, were utilised to get population genetics measures. VCFtools[216] (version 0.1.13) used SNPs data (see PopCluster analysis) to measure pairwise F_{ST} statistics between the four English populations defined by PopCluster, accordingly to Weir and Cockerham's calculation (1984). The calculations were performed on a windowed basis with a window of 5 or 50 kb and a step of 1 or 50 b. The nucleotide diversity was also calculated with VCFtools[216] (version 0.1.13) with a window of 5 kb and a step of 1 b.

All computational experiments were carried out on a high-performance computing (HPC) cluster running AlmaLinux 9.5, utilising the x86_64 architecture. The cluster utilises SLURM (Simple Linux Utility for Resource Management) to manage job scheduling and resource allocation, is equipped with a processor operating at a speed of 1.5 GHz and is provisioned with 503 GiB of RAM. The system supports 64 CPU cores.

General Discussion

The current PhD work provides an in-depth knowledge of the population and resistance genomics of sea beet populations. These data include the HiFi sequencing of 11 sea beet genomes with the long-read PacBio technology, with a mean depth of 36x, and the re-sequencing of the whole genome of 520 sea beets with the short-read Illumina technology, with a mean depth of 20x. These genomes originate from different locations across Europe, including Denmark, France and Spain but mostly England. These data have been screened for an association with rust resistance, a pathogen having a non-negligeable impact on sugar beet cultivation in England.

The eleven genome assemblies generated in this work represent an improvement compared to the published sea beet genome Bmar-1.0.1[16](L50 = 849, N50 = 0.17 Mb), both in terms of contiguity and completeness (L50 ranges from 7 to 31, and N50 from 5.7 to 32.8 Mb). They constitute a collection of high-quality material spanning the genetic diversity encountered in Atlantic sea beets, and this diversity they harbour is larger than present among breeding material. Moreover, most of this material (10 genomes out of 11), contrasts with the genetic background of domesticated sugar beets, derived from the Mediterranean sea beet group[10]. On the contrary, these wild beets, from the Atlantic group, are potentially better adapted to colder and wetter weather conditions. These reflections highlight a potential paradox of growing Mediterranean beets in an Atlantic climate, where most of the sugar beet cultivation is located. In addition to looking for traits in wild plants which haven't undergone a genetic bottleneck event (due to breeding), studying sea beets from England also has as an advantage that they most likely harbour resistance traits that function within this "Atlantic" climate, which includes resistance to pathogens encountered in this area. Indeed, fungal pathogens are more suited to these northern climates than to Mediterranean climates.

The eleven genome assemblies of a mean length of 714 Mb are explored and compared on the gene level in the present work within a pan-genomic perspective. A caveat should be added, as these gene analyses are likely downwards biased as the genomes were not re-annotated, except for NLR loci. However, the classification of annotated genes into orthogroups revealed a fraction of 91% (17,736 orthogroups) belonging to the sea beet core genome, and 9% (1,718 orthogroups) belonging to the accessory genome. On the resistance side, when observing potential NLR loci taking up, on average, 0.93% of the total number of genes, with 101 orthogroups, the core sea beet NLRome is smaller (47%) than the accessory NLRome (114 orthogroups). This highlights the importance of diversification in NLR genes to allow beets adapting to local pathogens, as well as pan-genome and *k*-mer based methodologies to analyse them. The present work suggests that, contrarily to what was previously hypothesised[14], sea beet genomes may harbour more than a single TIR-NLR. Indeed, potential TNL loci were predicted in more than one copy in multiple genomes encompassing the range of sea beet sampling locations.

The genomes generated here are of major importance and can open the door to the construction of a graph-based pan-genome, allowing clear representation of the diversity encountered in the wild. Moreover, these data facilitate the association genetic studies aiming to retrieve multiple agronomic traits from the wild, such as resistance to pathogens or climatic conditions specifically encountered in sugar beet cultivation areas.

The whole genome re-sequencing data from hundreds of sea beets, again, mostly from the Atlantic genomic background, facilitated the study of their population structure and gene flow, informing the understanding of resistance of a wild plant in relation to pathogens of its crop relative. Focusing on English and Danish samples, five populations were defined, corresponding to: the Danish, Merseyside/Cheshire (west of England), Humberside, Norfolk and Suffolk/Essex (east of England) populations. The fact that the Danish population splits apart from the English populations and shows low levels of admixture is perhaps expected given the location of the North Sea. However, this is also consistent with a previous study retrieving a separate ancestry for Danish sea beets, compared to wild beets sampled along Atlantic and north Mediterranean coasts[57].

The Humberside population appears as a standing-out population, due to it harbouring the highest linkage disequilibrium, the lowest effective population size, and some of the lowest gene flow measures with the other populations. These results indicate isolation of this population from the others, which could be explained by marine currents, flowing from North to South between the Humberside population and East Anglia (Norfolk and Suffolk). The low effective population size measured in Humberside is nevertheless far larger than measured in the sugar beet crop ($Ne \simeq 10{,}000$). The size of the present germplasm collection most certainly exceeds the size of a breeding panel.

Further gene-based analyses bring out a higher nucleotide diversity in resistance genes than in other genes. This result is promising for the use of sea beets as a source of resistance traits for the beet crop improvement. Interestingly, this result was less marked in the population from the west coast of England, which could indicate the impact of the presence of crop-adapted pathogens on the east coast to maintain wild genetic diversity, and the requirement for wild beets to develop specific resistance genes. However, it may also be impacted by gene flow from the crop.

The re-sequencing data from hundreds of sea beets, as well as the pan-genome, were used in the present study to search for candidate rust resistance genes via a *k*-mer-based association study. The pan-genome was utilised as a collection of potential NLR loci, to map *k*-mers for which an association to rust resistance was identified. This is the first *k*-mer-based association study conducted directly on wild-sampled individuals.

The observations from this analysis suggest potential sea beet-rust coevolution. Indeed, out of three rust inoculation trials, the sea beets, mostly sampled across England, were found to be more susceptible to the Danish rust than to the two English rusts. Moreover, the associated resistance loci were comparable across English rusts inoculations, but not between English and Danish rust experiments. However, this result is not clear cut because increased susceptibility to Danish rust is also observed in controls, and this suggests that this pattern may equally be the result of methodological differences.

The rust inoculation trials revealed agricultural beets as more susceptible than wild beets. Crops are bottlenecked and carry a different genetic background, due to the Mediterranean/Atlantic split, to the English sea beets, implying a potential difference in their resistance genes. Moreover, sea beets from the west coast were more susceptible than their counterparts from the east coast, in trials as well as in nature. In trials, this could be due to the location the rust was sampled from (east coast) and in nature, to the weather, which is wetter and, thus, more prone to fungal infections. Interestingly, this raises the point of designing future trials in conditions reflecting more the natural weather wild beets are subject to. Studies of wild germplasm are disadvantaged for a number of reasons but measuring the impact of a gene in the environment in which is it to be deployed may be one advantage.

On the association genetics side, the analysis didn't produce the clear peaks usually observed in crop-based studies, involving inbred lines. Instead, the method deployed here utilised signals across trials to identify genes of interest.

Two loci were identified based on their maximum association, one locus appeared in two controlled inoculations and one natural wild trial (each an English rust). Very interestingly, this work provides optimistic signs about a possibility to directly identify candidate resistance genes in the wild, only necessitating an overall scoring of pathogens presence/absence on the leaves of the plant and a collection of leaf material for sequencing. Randomly scoring 133 wild individuals, of which ¼ was infected, was sufficient to observe a signal directly in nature (replicated in the controlled trials). Population genomic outlier measures of genetic diversity and differentiation for the locus (identified in three analyses, two controlled and one wild) are encouraging for the potential to use population genetics to highlight candidate resistance genes. Outlier values were retrieved when comparing east and west English populations. This signal suggests either that this gene was introduced by crops to wild east coast populations, or that it is under selection due to the presence of crop-adapted pathogens on the east coast.

Three loci were identified reasoning that the mean association score carried by an NLR locus would be a strategic way to measure how well the locus in question is associated. These three loci are found within 30-50 kbp in the genome, two of them corresponding to full-length CC-NLRs and the third one, upstream and on the antisense strand, to a truncated NLR missing an N-terminal domain. The fact that they are all three identified despite the low linkage disequilibrium decay in sea beet (LD decay ≃ 1 kb) may indicate that they all are important and could function together. These three loci are present in ten HiFi genomes out of eleven: they are not present in the single representant of the Mediterranean clade, indicating that it is probable that these genes are missing from breeding materials. Further investigation into haplotype diversity, and local linkage could shed light on the mechanisms maintaining these three genes and the importance of maintaining them in a block for breeding.

In comparison with a reference-based GWAS, the present *k*-mer-based study enabled the identification of NLR sequences in wild sea beet that are absent from the sugar beet genome. A *k*-mer-based approach facilitates the identification of rare or novel alleles, as well as structural variants. It is particularly well-suited for screening NLR genes, as it does not rely on a reference genome, in which NLR genes could be misassembled due to their high identity with each other and their repetitive nature. The use of a long *k* length (51) minimises the likelihood of *k*-mers mapping to multiple NLR genes with high sequence identity. The pursuit of the present work, generating a sea beet pangenome representing the wild genetic diversity, will allow pan-genomic-based association analyses and a more comprehensive screening for novel resistance genes. Achieving this, however, will require the completion of the pangenome assembly.

To further the understanding of the present results, validation of the 5 NLR loci identified could involve using RNA-sequencing data to monitor their expression patterns consecutive to rust inoculation. The final validation will require the transformation of a susceptible beet background and the resistance evaluation through inoculation trials. Moreover, additional analyses could involve the measure of the strength and nature of selection on these potential candidate genes. Finally, investigating the resistance gene diversity in sea beet through the understanding of network dynamics would help comprehending the functioning of the three physically linked orthogroups associated with English rust resistance, crucial for a potential transfer to the sugar beet crop.

Bibliography

1. Lange W, Brandenburg WA, De Bock TSM. Taxonomy and cultonomy of beet (Beta vulgaris L.). *Botanical Journal of the Linnean Society*. 1999;130(1):81-96. doi:10.1006/bojl.1998.0250
2. Bosemark NO. Genetic poverty of the sugarbeet in Europe. In: *Proc Conf Broadening Genet Base Crops (1978) Pudoc, Wageningen*. ; 1979.
3. Biancardi E. Cultivated Offspring. In: *Beta Maritima*. Springer International Publishing; 2020:219-236. doi:10.1007/978-3-030-28748-1_9
4. Goldman IL, Navazio JP. History and Breeding of Table Beet in the United States. In: *Plant Breeding Reviews*. Wiley; 2002:357-388. doi:10.1002/9780470650202.ch7
5. de Bock ThSM. the Genus Beta: Domestication, Taxonomy and Interspecific Hybridization for Plant Breeding. *Acta Hortic*. 1986;(182):335-344. doi:10.17660/actahortic.1986.182.41
6. Lohaus G, Burba M, Heldt HW. Comparison of the contents of sucrose and amino acids in the leaves, phloem sap and taproots of high and low sugar-producing hybrids of sugar beet (Beta vulgaris L.). *J Exp Bot*. 1994;45(8):1097-1101. doi:10.1093/jxb/45.8.1097
7. Nei M. Genetic Distance between Populations. *Am Nat*. 1972;106(949):283-292.
8. Letschert JPW. *Beta Section Beta: Biogeographical Patterns of Variation, and Taxonomy*. Landbouwuniversiteit te Wageningen; 1993.
9. Letschert JPW. Beta section Beta: biogeographical patterns of variation, and taxonomy. *Wageningen Agricultural University Papers*. 1993;93-1.
10. Sandell FL, Stralis-Pavese N, McGrath JM, Schulz B, Himmelbauer H, Dohm JC. Genomic distances reveal relationships of wild and cultivated beets. *Nat Commun*. 2022;13(1):1-13. doi:10.1038/s41467-022-29676-9
11. Biancardi E, Panella LW, Lewellen RT. *Beta Maritima*.; 2012. doi:10.1007/978-1-4614-0842-0
12. Godshall MA. Sugar and Other Sweeteners. In: *Kent and Riegel's Handbook of Industrial Chemistry and Biotechnology*. Vol 34. Springer US; 2007:1657-1693. doi:10.1007/978-0-387-27843-8_35

13. Shapouri H, Salassi M, Fairbanks JN. The economic feasibility of ethanol production from sugar in the United States. *USDA*. Published online June 2006.

14. Dohm JC, Minoche AE, Holtgräwe D, et al. The genome of the recently domesticated crop plant sugar beet (Beta vulgaris). *Nature*. 2014;505(7484):546-549. doi:10.1038/nature12817

15. McGrath JM, Funk A, Galewski P, et al. A contiguous de novo genome assembly of sugar beet EL10 (Beta vulgaris L.). *DNA Res*. 2023;30(1):1-14. doi:10.1093/dnares/dsac033

16. Rodríguez del Río Á, Minoche AE, Zwickl NF, et al. Genomes of the wild beets Beta patula and Beta vulgaris ssp. maritima. *Plant Journal*. 2019;99(6):1242-1253. doi:10.1111/tpj.14413

17. Smith P, Martino D, Cai Z, et al. Agriculture. *Climate Change 2007: Mitigation*. 2007;45(09):45-5006-45-5006. doi:10.5860/choice.45-5006

18. Roudart L. Terres cultivables non cultivées : des disponibilités suffisantes pour la sécurité alimentaire durable de l'humanité. *Les publications du service de la statistique et de la prospective - Centre d'études et de prospective*. Published online 2009.

19. Crist E, Mora C, Engelman R. The interaction of human population, food production, and biodiversity protection. *Science (1979)*. 2017;264(April):260-264. doi:10.1126/science.aal2011

20. Malhi GS, Kaur M, Kaushik P. Impact of climate change on agriculture and its mitigation strategies: A review. *Sustainability (Switzerland)*. 2021;13(3):1-21. doi:10.3390/su13031318

21. Jones PD, Lister DH, Jaggard KW, Pidgeon JD. Future climate impact on the productivity of sugar beet (Beta vulgaris L.) in Europe. *Clim Change*. 2003;58(1-2):93-108. doi:10.1023/A:1023420102432

22. Lamichhane JR, Constantin J, Aubertot JN, Dürr C. Will climate change affect sugar beet establishment of the 21st century? Insights from a simulation study using a crop emergence model. *Field Crops Res*. 2019;238(March):64-73. doi:10.1016/j.fcr.2019.04.022

23. Dixon AFG, Kindlmann P, Leps J, Holman J. Why there are so few species of aphids, especially in the tropics. *American Naturalist*. 1987;129(4):580-592. doi:10.1086/284659

24. Dedryver CA, Le Ralec A, Fabre F. The conflicting relationships between aphids and men: A review of aphid damage and control strategies. *C R Biol*. 2010;333(6-7):539-553. doi:10.1016/j.crvi.2010.03.009

25. Hasselbring H. Fungus Diseases of Sugar Beet. *Botanical Gazette*. 1908;45(3):209-209. doi:10.1086/329507

26. Kristoffersen R, Hansen AL, Munk L, Cedergreen N, Jørgensen LN. Management of beet rust in accordance with IPM principles. *Crop Protection*. 2018;111(May 2017):6-16. doi:10.1016/j.cropro.2018.04.013

27. Aime MC, McTaggart AR, Mondo SJ, Duplessis S. Phylogenetics and Phylogenomics of Rust Fungi. *Adv Genet*. 2017;100:267-307. doi:10.1016/bs.adgen.2017.09.011

28. Hogenhout SA, Van Der Hoorn RAL, Terauchi R, Kamoun S. Emerging concepts in effector biology of plant-associated organisms. *Molecular Plant-Microbe Interactions*. 2009;22(2):115-122. doi:10.1094/MPMI-22-2-0115

29. Jones JDG, Dangl JL. The plant immune system. *Nature*. 2006;444(7117):323-329. doi:10.1038/nature05286

30. Karasov TL, Horton MW, Bergelson J. Genomic variability as a driver of plant-pathogen coevolution? *Curr Opin Plant Biol*. 2014;18(1):24-30. doi:10.1016/j.pbi.2013.12.003

31. Dangl JL, Jones JDG. Plant pathogens and integrated defence responses to infection. *Nature*. 2001;411(June).

32. Maruta N, Burdett H, Lim BYJ, et al. Structural basis of NLR activation and innate immune signalling in plants. *Immunogenetics*. 2022;74(1):5-26. doi:10.1007/s00251-021-01242-5

33. Yuan M, Jiang Z, Bi G, et al. Pattern-recognition receptors are required for NLR-mediated plant immunity. *Nature*. 2021;592(7852):105-109. doi:10.1038/s41586-021-03316-6

34. Thrall PH, Barrett LG, Dodds PN, Burdon JJ. Epidemiological and evolutionary outcomes in gene-for-gene and matching allele models. *Front Plant Sci*. 2016;6(JAN2016):1-12. doi:10.3389/fpls.2015.01084

35. Adachi H, Contreras M, Harant A, et al. An N-terminal motif in NLR immune receptors is functionally conserved across distantly related plant species. *Elife*. 2019;8:1-31. doi:10.7554/eLife.49956

36. Wu CH, Abd-El-Haliem A, Bozkurt TO, et al. NLR network mediates immunity to diverse plant pathogens. *Proc Natl Acad Sci U S A*. 2017;114(30):8113-8118. doi:10.1073/pnas.1702041114

37. Hu M, Qi J, Bi G, Zhou JM. Bacterial Effectors Induce Oligomerization of Immune Receptor ZAR1 In Vivo. *Mol Plant*. 2020;13(5):793-801. doi:10.1016/j.molp.2020.03.004

38. Contreras MP, Pai H, Tumtas Y, et al. Sensor NLR immune proteins activate oligomerization of their NRC helpers in response to plant pathogens. *EMBO J*. 2023;42(5):1-31. doi:10.15252/embj.2022111519

39. Bernoux M, Ve T, Williams S, et al. Structural and functional analysis of a plant resistance protein TIR domain reveals interfaces for self-association, signaling, and autoregulation. *Cell Host Microbe*. 2011;9(3):200-211. doi:10.1016/j.chom.2011.02.009

40. Wang J, Hu M, Wang J, et al. Reconstitution and structure of a plant NLR resistosome conferring immunity. *Science (1979)*. 2019;364(6435). doi:10.1126/science.aav5870

41. Wan L, Essuman K, Anderson RG, et al. TIR domains of plant immune receptors are NAD+-cleaving enzymes that promote cell death. *Science (1979)*. 2019;365(6455):799-803. doi:10.1126/science.aax1771

42. Wiermer M, Feys BJ, Parker JE. Plant immunity: The EDS1 regulatory node. *Curr Opin Plant Biol*. 2005;8(4):383-389. doi:10.1016/j.pbi.2005.05.010

43. Qi T, Seong K, Thomazella DPT, et al. NRG1 functions downstream of EDS1 to regulate TIR-NLR-mediated plant immunity in Nicotiana benthamiana. *Proc Natl Acad Sci U S A*. 2018;115(46):E10979-E10987. doi:10.1073/pnas.1814856115

44. Liang X, Dong J. Comparative-genomic analysis reveals dynamic NLR gene loss and gain across Apiaceae species. *Front Genet*. 2023;14. doi:10.3389/fgene.2023.1141194

45. Zhang Y, Edwards D, Batley J. Comparison and evolutionary analysis of Brassica nucleotide binding site leucine rich repeat (NLR) genes and importance for disease resistance breeding. *Plant Genome*. 2021;14(1). doi:10.1002/tpg2.20060

46. Yue J, Meyers BC, Chen J, Tian D, Yang S. Tracing the origin and evolutionary history of plant nucleotide-binding site–leucine-rich repeat (NBS-LRR) genes. *New Phytologist*. 2012;193(4):1049-1063. doi:10.1111/j.1469-8137.2011.04006.x

47. Shao ZQ, Xue JY, Wu P, et al. Large-Scale Analyses of Angiosperm Nucleotide-Binding Site-Leucine-Rich Repeat Genes Reveal Three Anciently Diverged Classes with Distinct Evolutionary Patterns. *Plant Physiol*. 2016;170(4):2095-2109. doi:10.1104/pp.15.01487

48. Michelmore RW, Meyers BC. Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. *Genome Res*. 1998;8(11):1113-1130. doi:10.1101/gr.8.11.1113

49. Bayer PE, Golicz AA, Tirnaz S, Chan CK, Edwards D, Batley J. Variation in abundance of predicted resistance genes in the Brassica oleracea pangenome. *Plant Biotechnol J*. 2019;17(4):789-800. doi:10.1111/pbi.13015

50. Funk A, Galewski P, McGrath JM. Nucleotide-binding resistance gene signatures in sugar beet, insights from a new reference genome. *Plant Journal*. 2018;95(4):659-671. doi:10.1111/tpj.13977

51. Woolhouse MEJ, Webster JP, Domingo E, Charlesworth B, Levin BR. Biological and biomedical implications of the co-evolution of pathogens and their hosts. *Nat Genet*. 2002;32(4):569-577. doi:10.1038/ng1202-569

52. Takken FLW, Goverse A. How to build a pathogen detector: Structural basis of NB-LRR function. *Curr Opin Plant Biol*. 2012;15(4):375-384. doi:10.1016/j.pbi.2012.05.001

53. Eitas TK, Dangl JL. NB-LRR proteins: Pairs, pieces, perception, partners, and pathways. *Curr Opin Plant Biol*. 2010;13(4):472-477. doi:10.1016/j.pbi.2010.04.007

54. Piperno DR. Assessing elements of an extended evolutionary synthesis for plant domestication and agricultural origin research. *Proc Natl Acad Sci U S A*. 2017;114(25):6429-6437. doi:10.1073/pnas.1703658114

55. Eyre-Walker A, Gaut RL, Hilton H, Feldman DL, Gaut BS. Investigation of the bottleneck leading to the domestication of maize. *Proc Natl Acad Sci U S A*. 1998;95(8):4441-4446. doi:10.1073/pnas.95.8.4441

56. Fénart S, Arnaud JF, De Cauwer I, Cuguen J. Nuclear and cytoplasmic genetic diversity in weed beet and sugar beet accessions compared to wild relatives: New insights into the genetic relationships within the Beta vulgaris complex species. *Theoretical and Applied Genetics*. 2008;116(8):1063-1077. doi:10.1007/s00122-008-0735-1

57. Felkel S, Dohm JC, Himmelbauer H. Genomic variation in the genus Beta based on 656 sequenced beet genomes. *Sci Rep*. 2023;13(1). doi:10.1038/s41598-023-35691-7

58. Rundlöf M, Andersson GKS, Bommarco R, et al. Seed coating with a neonicotinoid insecticide negatively affects wild bees. *Nature*. 2015;521(7550):77-80. doi:10.1038/nature14420

59. Karadimos DA, Karaoglanidis GS, Tzavella-Klonari K. Biological activity and physical modes of action of the Q o inhibitor fungicides trifloxystrobin and pyraclostrobin

against Cercospora beticola. *Crop Protection*. 2005;24(1):23-29. doi:10.1016/j.cropro.2004.06.004

60. Fernández-Ortuño D, Chen F, Schnabel G. Resistance to pyraclostrobin and boscalid in botrytis cinerea isolates from strawberry fields in the Carolinas. *Plant Dis*. 2012;96(8):1198-1203. doi:10.1094/PDIS-12-11-1049-RE

61. Menegola E, Broccia ML, Di Renzo F, Giavini E. Postulated pathogenic pathway in triazole fungicide induced dysmorphogenic effects. *Reproductive Toxicology*. 2006;22(2):186-195. doi:10.1016/j.reprotox.2006.04.008

62. Taxvig C, Hass U, Axelstad M, et al. Endocrine-disrupting activities In Vivo of the fungicides tebuconazole and epoxiconazole. *Toxicological Sciences*. 2007;100(2):464-473. doi:10.1093/toxsci/kfm227

63. *Genetically Modified Pest-Protected Plants: Science and Regulation*.; 2000. doi:10.17226/9795

64. Hajjar R, Hodgkin T. The use of wild relatives in crop improvement: A survey of developments over the last 20 years. *Euphytica*. 2007;156(1-2):1-13. doi:10.1007/s10681-007-9363-0

65. Stevanato P, Biaggi M De, Skaracis GN, Colombo M, Mandolino G, Biancardi E. The sea beet (Beta vulgaris L. ssp.maritima) of the adriatic coast as source of resistance for sugar beet. *Sugar Tech*. 2001;3(3):77-82. doi:10.1007/bf03014567

66. Capistrano-Gossmann GG, Ries D, Holtgräwe D, et al. Crop wild relative populations of Beta vulgaris allow direct mapping of agronomically important genes. *Nat Commun*. 2017;8. doi:10.1038/ncomms15708

67. Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet*. 2005;6(2):95-108. doi:10.1038/nrg1521

68. Santure AW, Garant D. Wild GWAS—association mapping in natural populations. *Mol Ecol Resour*. 2018;18(4):729-738. doi:10.1111/1755-0998.12901

69. Tabor HK, Risch NJ, Myers RM. Candidate-gene approaches for studying complex genetic traits: practical considerations. *Nat Rev Genet*. 2002;3(5):391-397. doi:10.1038/nrg796

70. Park ST, Kim J. Trends in next-generation sequencing and a new era for whole genome sequencing. *Int Neurourol J*. 2016;20:76-83. doi:10.5213/inj.1632742.371

71. Capistrano-Gossmann GG, Ries D, Holtgräwe D, et al. Crop wild relative populations of Beta vulgaris allow direct mapping of agronomically important genes. *Nat Commun*. 2017;8. doi:10.1038/ncomms15708

72. Santure AW, Garant D. Wild GWAS—association mapping in natural populations. *Mol Ecol Resour*. 2018;18(4):729-738. doi:10.1111/1755-0998.12901

73. François O, Martins H, Caye K, Schoville SD. Controlling false discoveries in genome scans for selection. *Mol Ecol*. 2016;25(2):454-469. doi:10.1111/mec.13513

74. Schielzeth H, Rios Villamil A, Burri R. Success and failure in replication of genotype–phenotype associations: How does replication help in understanding the genetic basis of phenotypic variation in outbred populations? *Mol Ecol Resour*. 2018;18(4):739-754. doi:10.1111/1755-0998.12780

75. Fleischmann RD, Adams MD, White O, et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science (1979)*. 1995;269(5223):496-512. doi:10.1126/science.7542800

76. Alonso-Blanco C, Andrade J, Becker C, et al. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell*. 2016;166(2):481-491. doi:10.1016/j.cell.2016.05.063

77. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature*. 2001;412(6846):565-566. doi:10.1038/35087627

78. Gerdol M, Moreira R, Cruz F, et al. Massive gene presence-absence variation shapes an open pan-genome in the Mediterranean mussel. *Genome Biol*. 2020;21(1):275. doi:10.1186/s13059-020-02180-3

79. Zhang X, Chen X, Liang P, Tang H. Cataloging Plant Genome Structural Variations. *Curr Issues Mol Biol*. Published online 2018:181-194. doi:10.21775/cimb.027.181

80. Dolatabadian A, Bayer PE, Tirnaz S, Hurgobin B, Edwards D, Batley J. Characterization of disease resistance genes in the Brassica napus pangenome reveals significant structural variation. *Plant Biotechnol J*. 2020;18(4):969-982. doi:10.1111/pbi.13262

81. Kang M, Wu H, Liu H, et al. The pan-genome and local adaptation of Arabidopsis thaliana. *Nat Commun*. 2023;14(1):1-14. doi:10.1038/s41467-023-42029-4

82. Tettelin H, Masignani V, Cieslewicz MJ, et al. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome." *Proc Natl Acad Sci U S A*. 2005;102(39):13950-13955. doi:10.1073/pnas.0506758102

83. Hamilton JP, Robin Buell C. Advances in plant genome sequencing. *Plant Journal*. 2012;70(1):177-190. doi:10.1111/j.1365-313X.2012.04894.x

84. Li YH, Zhou G, Ma J, et al. De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat Biotechnol*. 2014;32(10):1045-1052. doi:10.1038/nbt.2979

85. Gao L, Gonda I, Sun H, et al. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nat Genet*. 2019;51(6):1044-1051. doi:10.1038/s41588-019-0410-2

86. Shang L, Li X, He H, et al. A super pan-genomic landscape of rice. *Cell Res*. 2022;32(10):878-896. doi:10.1038/s41422-022-00685-z

87. Montenegro JD, Golicz AA, Bayer PE, et al. The pangenome of hexaploid bread wheat. *Plant Journal*. 2017;90(5):1007-1013. doi:10.1111/tpj.13515

88. Yan H, Sun M, Zhang Z, et al. Pangenomic analysis identifies structural variation associated with heat tolerance in pearl millet. *Nat Genet*. 2023;55(3):507-518. doi:10.1038/s41588-023-01302-4

89. Liao WW, Asri M, Ebler J, et al. A draft human pangenome reference. *Nature*. 2023;617(7960):312-324. doi:10.1038/s41586-023-05896-x

90. Varshney RK, Roorkiwal M, Sun S, et al. A chickpea genetic variation map based on the sequencing of 3,366 genomes. *Nature*. 2021;599(7886):622-627. doi:10.1038/s41586-021-04066-1

91. Tettelin H, Masignani V, Cieslewicz MJ, et al. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome." *Proc Natl Acad Sci U S A*. 2005;102(39):13950-13955. doi:10.1073/pnas.0506758102

92. Van de Weyer AL, Monteiro F, Furzer OJ, et al. A Species-Wide Inventory of NLR Genes and Alleles in Arabidopsis thaliana. *Cell*. 2019;178(5):1260-1272.e14. doi:10.1016/j.cell.2019.07.038

93. Khan AW, Garg V, Roorkiwal M, Golicz AA, Edwards D, Varshney RK. Super-Pangenome by Integrating the Wild Side of a Species for Accelerated Crop Improvement. *Trends Plant Sci*. 2020;25(2):148-158. doi:10.1016/j.tplants.2019.10.012

94. Iqbal MA, Saleem AM. Sugar beet potential to beat sugarcane as a sugar crop in pakistan. *American-Eurasian J Agric & Environ Sci*. 2015;15(1):36-44. doi:10.5829/idosi.aejaes.2015.15.1.12480

95. Řezbová H, Belová A, Škubna O. Sugar beet production in the European Union and their future trends. *AGRIS on-line Papers in Economics and Informatics*. 2013;(4). doi:10.22004/ag.econ.162299

96. Stevanato P, Biaggi M De, Skaracis GN, Colombo M, Mandolino G, Biancardi E. The sea beet (Beta vulgaris L. ssp.maritima) of the adriatic coast as source of resistance for sugar beet. *Sugar Tech*. 2001;3(3):77-82. doi:10.1007/bf03014567

97. Bohra A, Kilian B, Sivasankar S, et al. Reap the crop wild relatives for breeding future crops. *Trends Biotechnol*. 2022;40(4):412-431. doi:10.1016/j.tibtech.2021.08.009

98. Stam R, Silva-Arias GA, Tellier A. Subsets of NLR genes show differential signatures of adaptation during colonization of new habitats. *New Phytologist*. 2019;224(1):367-379. doi:10.1111/nph.16017

99. Borrelli GM, Mazzucotelli E, Marone D, et al. Regulation and evolution of NLR genes: A close interconnection for plant immunity. *Int J Mol Sci*. 2018;19(6). doi:10.3390/ijms19061662

100. Claros MG, Bautista R, Guerrero-Fernández D, Benzerki H, Seoane P, Fernández-Pozo N. Why assembling plant genome sequences is so challenging. *Biology (Basel)*. 2012;1(2):439-459. doi:10.3390/biology1020439

101. Clark JW, Donoghue PCJ. Whole-Genome Duplication and Plant Macroevolution. *Trends Plant Sci*. 2018;23(10):933-945. doi:10.1016/j.tplants.2018.07.006

102. Hamilton JP, Robin Buell C. Advances in plant genome sequencing. *Plant Journal*. 2012;70(1):177-190. doi:10.1111/j.1365-313X.2012.04894.x

103. Hu T, Chitnis N, Monos D, Dinh A. Next-generation sequencing technologies: An overview. *Hum Immunol*. 2021;82(11):801-811. doi:10.1016/j.humimm.2021.02.012

104. Wenger AM, Peluso P, Rowell WJ, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol*. 2019;37(10):1155-1162. doi:10.1038/s41587-019-0217-9

105. Rizzi R, Beretta S, Patterson M, et al. Overlap graphs and de Bruijn graphs: data structures for de novo genome assembly in the big data era. *Quantitative Biology*. 2019;7(4):278-292. doi:10.1007/s40484-019-0181-x

106. Li Z, Chen Y, Mu D, et al. Comparison of the two major classes of assembly algorithms: Overlap-layout-consensus and de-bruijn-graph. *Brief Funct Genomics*. 2012;11(1):25-37. doi:10.1093/bfgp/elr035

107. Wang Y. A Comparative Study of HiCanu and Hifiasm. *ACM International Conference Proceeding Series*. Published online 2022:100-104. doi:10.1145/3545839.3545855

108. Nurk S, Walenz BP, Rhie A, et al. HiCanu: Accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res*. 2020;30(9):1291-1305. doi:10.1101/GR.263566.120

109. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods*. 2021;18(2):170-175. doi:10.1038/s41592-020-01056-5

110. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. 2015;31(19):3210-3212. doi:10.1093/bioinformatics/btv351

111. Kryukov K, Imanishi T. Human contamination in public genome assemblies. *PLoS One*. 2016;11(9):1-11. doi:10.1371/journal.pone.0162424

112. Laetsch DR, Blaxter ML. BlobTools: Interrogation of genome assemblies. *F1000Res*. 2017;6:1287. doi:10.12688/f1000research.12232.1

113. Madden T. The BLAST Sequence Analysis Tool. *The NCBI Handbook*. Published online 2002:1-15.

114. Shumate A, Salzberg SL. Liftoff: Accurate mapping of gene annotations. *Bioinformatics*. 2021;37(12):1639-1643. doi:10.1093/bioinformatics/btaa1016

115. Bayer PE, Golicz AA, Tirnaz S, Chan CKK, Edwards D, Batley J. Variation in abundance of predicted resistance genes in the Brassica oleracea pangenome. *Plant Biotechnol J*. 2019;17(4):789-800. doi:10.1111/pbi.13015

116. Zhang W. NLR-annotator: A tool for de novo annotation of intracellular immune receptor repertoire. *Plant Physiol*. 2020;183(2):418-420. doi:10.1104/pp.20.00525

117. Steuernagel B, Witek K, Krattinger SG, et al. The NLR-annotator tool enables annotation of the intracellular immune receptor repertoire. *Plant Physiol*. 2020;183(2):468-482. doi:10.1104/pp.19.01273

118. Mapleson D, Accinelli GG, Kettleborough G, Wright J, Clavijo BJ. KAT: A K-mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics*. 2017;33(4):574-576. doi:10.1093/bioinformatics/btw663

119. Rodríguez del Río Á, Minoche AE, Zwickl NF, et al. Genomes of the wild beets Beta patula and Beta vulgaris ssp. maritima. *Plant Journal*. 2019;99(6):1242-1253. doi:10.1111/tpj.14413

120. Rhie A, Walenz BP, Koren S, Phillippy AM. Merqury: Reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol*. 2020;21(1):1-27. doi:10.1186/s13059-020-02134-9

121. McGrath JM, Funk A, Galewski P, et al. A contiguous de novo genome assembly of sugar beet EL10 (Beta vulgaris L.). *DNA Res*. 2023;30(1):1-14. doi:10.1093/dnares/dsac033

122. Jacob F, Vernaldi S, Maekawa T. Evolution and conservation of plant NLR functions. *Front Immunol*. 2013;4(SEP):1-16. doi:10.3389/fimmu.2013.00297

123. Barragan AC, Weigel D. Plant NLR diversity: the known unknowns of pan-NLRomes. *Plant Cell*. 2021;33(4):814-831. doi:10.1093/plcell/koaa002

124. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol*. 2015;16(1):1-14. doi:10.1186/s13059-015-0721-2

125. Eizenga JM, Novak AM, Sibbesen JA, et al. Pangenome Graphs. Published online 2020. doi:10.1146/annurev-genom-120219

126. Gao L, Gonda I, Sun H, et al. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nat Genet*. 2019;51(6):1044-1051. doi:10.1038/s41588-019-0410-2

127. Contreras-Moreira B, Cantalapiedra CP, García-Pereira MJ, et al. Analysis of plant pangenomes and transcriptomes with GET_HOMOLOGUES-EST, a clustering solution for sequences of the same species. *Front Plant Sci*. 2017;8(February):1-16. doi:10.3389/fpls.2017.00184

128. Montenegro JD, Golicz AA, Bayer PE, et al. The pangenome of hexaploid bread wheat. *Plant Journal*. 2017;90(5):1007-1013. doi:10.1111/tpj.13515

129. Wang W, Mauleon R, Hu Z, et al. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature*. 2018;557(7703):43-49. doi:10.1038/s41586-018-0063-9

130. Baggs EL, Monroe JG, Thanki AS, et al. Convergent loss of an EDS1/PAD4 signaling pathway in several plant lineages reveals coevolved components of plant immunity and drought response. *Plant Cell*. 2020;32(7):2158-2177. doi:10.1105/tpc.19.00903

131. Jia YX, Yuan Y, Zhang Y, Yang S, Zhang X. Extreme expansion of NBS-encoding genes in rosaceae. *BMC Genet*. 2015;16(1). doi:10.1186/s12863-015-0208-x

132. Wang T, Jia ZH, Zhang JY, Liu M, Guo ZR, Wang G. Identification and analysis of nbs-lrr genes in actinidia chinensis genome. *Plants*. 2020;9(10):1-12. doi:10.3390/plants9101350

133. Meyers BC, Kozik A, Griego A, Kuang H, Michelmore RW. Genome-wide analysis of NBS-LRR-encoding genes in Arabidopsis. *Plant Cell*. 2003;15(4):809-834. doi:10.1105/tpc.009308

134. Chen Q, Han Z, Jiang H, Tian D, Yang S. Strong positive selection drives rapid diversification of R-Genes in arabidopsis relatives. *J Mol Evol*. 2010;70(2):137-148. doi:10.1007/s00239-009-9316-4

135. Leister D. Tandem and segmental gene duplication and recombination in the evolution of plant disease resistance genes. *Trends in Genetics*. 2004;20(3):116-122. doi:10.1016/j.tig.2004.01.001

136. Bendahmane A, Querci M, Kanyuka K, Baulcombe DC. Agrobacterium transient expression system as a tool for the isolation of disease resistance genes: Application to the Rx2 locus in potato. *Plant Journal*. 2000;21(1):73-81. doi:10.1046/j.1365-313X.2000.00654.x

137. Meyers BC, Chin DB, Shen KA, et al. *The Major Resistance Gene Cluster in Lettuce Is Highly Duplicated and Spans Several Megabases*. Vol 10.; 1998. https://academic.oup.com/plcell/article/10/11/1817/5999460

138. Tian Y, Fan L, Thurau T, Jung C, Cai D. The Absence of TIR-Type Resistance Gene Analogues in the Sugar Beet (Beta vulgaris L.) Genome. *J Mol Evol*. 2004;58(1):40-53. doi:10.1007/s00239-003-2524-4

139. Funk A, Galewski P, McGrath JM. Nucleotide-binding resistance gene signatures in sugar beet, insights from a new reference genome. *Plant Journal*. 2018;95(4):659-671. doi:10.1111/tpj.13977

140. Kourelis J, Sakai T, Adachi H, Kamoun S. RefPlantNLR is a comprehensive collection of experimentally validated plant disease resistance proteins from the NLR family. *PLoS Biol*. 2021;19(10):1-26. doi:10.1371/journal.pbio.3001124

141. Van de Weyer AL, Monteiro F, Furzer OJ, et al. A Species-Wide Inventory of NLR Genes and Alleles in Arabidopsis thaliana. *Cell*. 2019;178(5):1260-1272.e14. doi:10.1016/j.cell.2019.07.038

142. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods*. 2021;18(2):170-175. doi:10.1038/s41592-020-01056-5

143. Li H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*. 2018;34(18):3094-3100. doi:10.1093/bioinformatics/bty191

144. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-2079. doi:10.1093/bioinformatics/btp352

145. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I. ABySS: A parallel assembler for short read sequence data. *Genome Res*. 2009;19(6):1117-1123. doi:10.1101/gr.089532.108

146. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. 2015;31(19):3210-3212. doi:10.1093/bioinformatics/btv351

147. Katz L, Griswold T, Morrison S, et al. Mashtree: a rapid comparison of whole genome sequence files. *J Open Source Softw*. 2019;4(44):1762. doi:10.21105/joss.01762

148. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*. 2019;35(3):526-528. doi:10.1093/bioinformatics/bty633

149. Shumate A, Salzberg SL. Liftoff: Accurate mapping of gene annotations. *Bioinformatics*. 2021;37(12):1639-1643. doi:10.1093/bioinformatics/btaa1016

150. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol*. 2015;16(1):1-14. doi:10.1186/s13059-015-0721-2

151. Lovell JT, Sreedasyam A, Schranz ME, et al. GENESPACE: syntenic pan-genome annotations for eukaryotes. *bioRxiv*. Published online 2022:2022.03.09.483468.

152. Bush WS, Moore JH. Chapter 11: Genome-Wide Association Studies. *PLoS Comput Biol*. 2012;8(12). doi:10.1371/journal.pcbi.1002822

153. Rafalski JA. Association genetics in crop improvement. *Curr Opin Plant Biol*. 2010;13(2):174-180. doi:10.1016/j.pbi.2009.12.004

154. Mckown AD, Klápště J, Guy RD, et al. Genome-wide association implicates numerous genes underlying ecological trait variation in natural populations of Populus trichocarpa. *New Phytologist*. 2014;203(2):535-553. doi:10.1111/nph.12815

155. Hecht BC, Campbell NR, Holecek DE, Narum SR. Genome-wide association reveals genetic basis for the propensity to migrate in wild populations of rainbow and steelhead trout. *Mol Ecol*. 2013;22(11):3061-3076. doi:10.1111/mec.12082

156. Würschum T, Langer SM, Longin CFH. Genetic control of plant height in European winter wheat cultivars. *Theoretical and Applied Genetics*. 2015;128(5):865-874. doi:10.1007/s00122-015-2476-2

157. Schielzeth H, Husby A. Challenges and prospects in genome-wide quantitative trait loci mapping of standing genetic variation in natural populations. *Ann N Y Acad Sci*. 2014;1320(1):35-57. doi:10.1111/nyas.12397

158. Dempewolf H, Baute G, Anderson J, Kilian B, Smith C, Guarino L. Past and future use of wild relatives in crop breeding. *Crop Sci*. 2017;57(3):1070-1082. doi:10.2135/cropsci2016.10.0885

159. Saxena RK, Edwards D, Varshney RK. Structural variations in plant genomes. *Brief Funct Genomic Proteomic*. 2014;13(4). doi:10.1093/bfgp/elu016

160. Yuan Y, Bayer PE, Batley J, Edwards D. Current status of structural variation studies in plants. *Plant Biotechnol J*. 2021;19(11):2153-2163. doi:10.1111/pbi.13646

161. Bush SJ, Castillo-Morales A, Tovar-Corona JM, Chen L, Kover PX, Urrutia AO. Presence-absence variation in A. thaliana is primarily associated with genomic signatures consistent with relaxed selective constraints. *Mol Biol Evol*. 2014;31(1):59-69. doi:10.1093/molbev/mst166

162. Geibel J, Reimer C, Weigend S, Weigend A, Pook T, Simianer H. How array design creates SNP ascertainment bias. *PLoS One*. 2021;16(3 March). doi:10.1371/journal.pone.0245178

163. Voichek Y, Weigel D. Identifying genetic variants underlying phenotypic variation in plants without complete genomes. *Nat Genet*. 2020;52(5):534-540. doi:10.1038/s41588-020-0612-7

164. Arora S, Steuernagel B, Gaurav K, et al. Resistance gene cloning from a wild crop relative by sequence capture and association genetics. *Nat Biotechnol*. 2019;37(2):139-143. doi:10.1038/s41587-018-0007-9

165. Jupe F, Witek K, Verweij W, et al. Resistance gene enrichment sequencing (RenSeq) enables reannotation of the NB-LRR gene family from sequenced plant genomes and rapid mapping of resistance loci in segregating populations. *Plant Journal*. 2013;76(3):530-544. doi:10.1111/tpj.12307

166. Gaurav K, Arora S, Silva P, et al. Population genomic analysis of Aegilops tauschii identifies targets for bread wheat improvement. *Nat Biotechnol*. 2022;40(3):422-431. doi:10.1038/s41587-021-01058-4

167. Klymiuk V, Coaker G, Fahima T, Pozniak CJ. Tandem protein kinases emerge as new regulators of plant immunity. *Molecular Plant-Microbe Interactions*. 2021;34(10):1094-1102. doi:10.1094/MPMI-03-21-0073-CR

168. De Cauwer I, Dufay M, Cuguen J, Arnaud JFÇ. Effects of fine-scale genetic structure on male mating success in gynodioecious Beta vulgaris ssp. maritima. *Mol Ecol*. 2010;19(8):1540-1558. doi:10.1111/j.1365-294X.2010.04586.x

169. Fénart S, Austerlitz F, Cuguen J, Arnaud JF. Long distance pollen-mediated gene flow at a landscape level: The weed beet as a case study. *Mol Ecol*. 2007;16(18):3801-3813. doi:10.1111/j.1365-294X.2007.03448.x

170. Fievet V, Touzet P, Arnaud JF, Cuguen J. Spatial analysis of nuclear and cytoplasmic DNA diversity in wild sea beet (Beta vulgaris ssp. maritima) populations: Do marine currents shape the genetic structure? *Mol Ecol*. 2007;16(9):1847-1864. doi:10.1111/j.1365-294X.2006.03208.x

171. Stakman EC, Stewart DM, Loegering WQ. Identification of physiologic races of Puccinia graminis var. tritici. *US Department of Agriculture, Agriculture Research Services*. Published online 1962:54.

172. Stam R, Silva-Arias GA, Tellier A. Subsets of NLR genes show differential signatures of adaptation during colonization of new habitats. *New Phytologist*. 2019;224(1):367-379. doi:10.1111/nph.16017

173. Arora S, Steuernagel B, Gaurav K, et al. Resistance gene cloning from a wild crop relative by sequence capture and association genetics. *Nat Biotechnol*. 2019;37(2):139-143. doi:10.1038/s41587-018-0007-9

174. Wu Z, Tian L, Liu X, Huang W, Zhang Y, Li X. The N-terminally truncated helper NLR NRG1C antagonizes immunity mediated by its full-length neighbors NRG1A and NRG1B. *Plant Cell*. 2022;34(5):1621-1640. doi:10.1093/plcell/koab285

175. Perez-Sepulveda BM, Heavens D, Pulford C V., et al. An accessible, efficient and global approach for the large-scale sequencing of bacterial genomes. *Genome Biol*. 2021;22(1):1-18. doi:10.1186/s13059-021-02536-3

176. Krueger F. Trim Galore!: A wrapper around Cutadapt and FastQC to consistently apply adapter and quality trimming to FastQ files, with extra functionality for RRBS data. *Babraham Institute*. Published online 2015. Accessed March 31, 2024. https://cir.nii.ac.jp/crid/1370294643762929691.bib?lang=en

177. Lemane T, Medvedev P, Chikhi R, Peterlongo P. kmtricks: efficient and flexible construction of Bloom filters for large sequencing data collections. *Bioinformatics Advances*. 2022;2(1):1-8. doi:10.1093/bioadv/vbac029

178. Steuernagel B, Witek K, Krattinger SG, et al. The NLR-Annotator tool enables annotation of the intracellular immune receptor repertoire. *Plant Physiol*. 2020;183(June):pp.01273.2019. doi:10.1104/pp.19.01273

179. De Cauwer I, Dufay M, Hornoy B, Courseaux A, Arnaud JF. Gynodioecy in structured populations: Understanding fine-scale sex ratio variation in Beta vulgaris ssp. maritima. *Mol Ecol*. 2012;21(4):834-850. doi:10.1111/j.1365-294X.2011.05414.x

180. Driessen S, Pohl M, Bartsch D. RAPD-PCR analysis of the genetic origin of sea beet (Beta vulgaris ssp. maritima) at Germany's Baltic Sea coast. *Basic Appl Ecol*. 2001;2(4):341-349. doi:10.1078/1439-1791-00061

181. Archimowitsch A. Control of Pollination in Sugar-Beet. *Botanical Review*. 1949;15(9):613-628.

182. Leys M, Petit EJ, El-Bahloul Y, Liso C, Fournet S, Arnaud JF. Spatial genetic structure in Beta vulgaris subsp. maritima and Beta macrocarpa reveals the effect of contrasting mating system, influence of marine currents, and footprints of postglacial recolonization routes. *Ecol Evol*. 2014;4(10):1828-1852. doi:10.1002/ece3.1061

183. Bartsch D, Pohl-Orf M. Ecological aspects of transgenic sugar beet: transfer and expression of herbicide resistance in hybrids with wild beets. *Euphytica*. 1996;91:55-58.

184. Arnaud JF, Viard F, Delescluse M, Cuguen J. Evidence for gene flow via seed dispersal from crop to wild relatives in Beta vulgaris (Chenopodiaceae): Consequences for the release of genetically modified crop species with weedy lineages. *Proceedings of the Royal Society B: Biological Sciences*. 2003;270(1524):1565-1571. doi:10.1098/rspb.2003.2407

185. Bartsch D, Lehnen M, Clegg J, Pohl-Orf M, Schuphan I, Ellstrand NC. Impact of gene flow from cultivated beet on genetic diversity of wild sea beet populations. *Mol Ecol*. 1999;8(10):1733-1741. doi:10.1046/j.1365-294X.1999.00769.x

186. Cureton AN, Newbury HJ, Raybould AF, Ford-Lloyd B V. Genetic structure and gene flow in wild beet populations: The potential influence of habitat on transgene spread and risk assessment. *Journal of Applied Ecology*. 2006;43(6):1203-1212. doi:10.1111/j.1365-2664.2006.01236.x

187. Ondov BD, Treangen TJ, Melsted P, et al. Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biol*. 2016;17(1):1-14. doi:10.1186/s13059-016-0997-x

188. Charlesworth B. Fundamental concepts in genetics: Effective population size and patterns of molecular evolution and variation. *Nat Rev Genet*. 2009;10(3):195-205. doi:10.1038/nrg2526

189. Hartl DL, Clark AG. *Principles of Population Genetics*. Vol Third Edition.; 1997.

190. Wright S. Evolution in Mendelian Populations. *Genetics*. 1931;16:97-159. doi:10.1093/genetics/16.2.97

191. Wright S. *The Interpretation of Population Structure by F-Statistics with Special Regard to Systems of Mating*. Vol 19.; 1965.

192. Holsinger KE, Weir BS. Genetics in geographically structured populations: Defining, estimating and interpreting FST. *Nat Rev Genet*. 2009;10(9):639-650. doi:10.1038/nrg2611

193. Barrett JC. Population Genetics and Linkage Disequilibrium. *Analysis of Complex Disease Association Studies: A Practical Guide*. Published online 2010:15-23. doi:10.1016/B978-0-12-375142-3.10002-1

194. Gupta PK, Rustgi S, Kulwal PL. Linkage disequilibrium and association studies in higher plants: Present status and future prospects. *Plant Mol Biol*. 2005;57(4):461-485. doi:10.1007/s11103-005-0257-z

195. Nieuwenhuis BPS, James TY. The frequency of sex in fungi. *Philosophical Transactions of the Royal Society B: Biological Sciences*. 2016;371(1706). doi:10.1098/rstb.2015.0540

196. Huang X, Wei X, Sang T, et al. Genome-wide asociation studies of 14 agronomic traits in rice landraces. *Nat Genet*. 2010;42(11):961-967. doi:10.1038/ng.695

197. Hill WG, Robertson A. The effect of linkage on limits to artificial selection. *Genet Res*. 1966;8(3):269-294. doi:10.1017/S0016672300010156

198. Wang J. Fast and accurate population admixture inference from genotype data from a few microsatellites to millions of SNPs. *Heredity (Edinb)*. 2022;129(2):79-92. doi:10.1038/s41437-022-00535-z

199. Ebert D, Fields PD. Host–parasite co-evolution and its genomic signature. *Nat Rev Genet*. 2020;21(12):754-768. doi:10.1038/s41576-020-0269-1

200. De Hauwere N. Marine Regions. (n.d.). World map with EEZs. Retrieved January 14, 2025, from https://www.marineregions.org/gazetteer.php?p=image&pic=115812.

201. Hey J, Chung Y, Sethuraman A, et al. Phylogeny estimation by integration over isolation with migration models. *Mol Biol Evol*. 2018;35(11):2805-2818. doi:10.1093/molbev/msy162

202. Schiffels S, Durbin R. Inferring human population size and separation history from multiple genome sequences. *Nat Genet*. 2014;46(8):919-925. doi:10.1038/ng.3015

203. Kaback DB, Guacci V, Barber D, Mahon JW. Chromosome Size-Dependant Control of Meiotic Recombination. *Science (1979)*. 1992;256.

204. Galewski P, McGrath JM. Genetic diversity among cultivated beets (Beta vulgaris) assessed via population-based whole genome sequences. *BMC Genomics*. 2020;21(1). doi:10.1186/s12864-020-6451-1

205. Arunyawat U, Stephan W, Städler T. Using multilocus sequence data to assess population structure, natural selection, and linkage disequilibrium in wild tomatoes. *Mol Biol Evol*. 2007;24(10):2310-2322. doi:10.1093/molbev/msm162

206. Letschert JPW, Lange W, Frese L, Van Den Berg RG. Taxonomy of Beta Section Beta. *Journal of Sugarbeet Research*. 1994;31(1 & 2):69-85. doi:10.5274/jsbr.31.1.69

207. Nielsen R. Molecular signatures of natural selection. *Annu Rev Genet*. 2005;39:197-218. doi:10.1146/annurev.genet.39.073003.112420

208. Spurgin LG, Richardson DS. How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. *Proceedings of the Royal Society B: Biological Sciences*. 2010;277(1684):979-988. doi:10.1098/rspb.2009.2084

209. McMullan M, Percival-Alwyn L, Sawford K, et al. Analysis of wild plant pathogen populations reveals a signal of adaptation in genes evolving for survival in agriculture in the beet rust pathogen (Uromyces beticola). *bioRxiv*. Published online 2021:1-18. doi:10.1101/2021.08.12.456076

210. Schierup MH. The Number of Self-Incompatibility Alleles in a Finite, Subdivided Population. *Genetics*. 1998;149(2):1153-1162. doi:10.1093/genetics/149.2.1153

211. Schierup MH, Vekemans X, Charlesworth D. The effect of subdivision on variation at multi-allelic loci under balancing selection. *Genet Res*. 2000;76(1):51-62. doi:10.1017/S0016672300004535

212. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol Biol Evol*. 2015;32(1):268-274. doi:10.1093/molbev/msu300

213. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Published online 2018.

214. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013;00(00):1-3.

215. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-2079. doi:10.1093/bioinformatics/btp352

216. Danecek P, Auton A, Abecasis G, et al. The variant call format and VCFtools. *Bioinformatics*. 2011;27(15):2156-2158. doi:10.1093/bioinformatics/btr330

Appendix

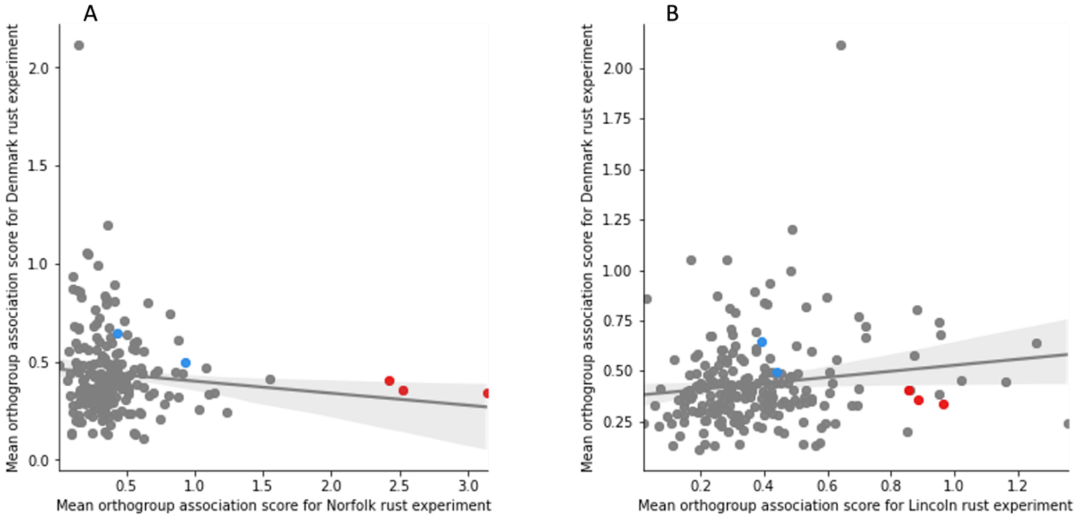


Figure S1 - There is no apparent parallel between the Danish and UK rust trials, in terms of mean score per NLR orthogroup.

Mean association score per NLR orthogroup with rust resistance compared between the trials involving the Danish and the Norfolk rust (A) or the Danish and the Lincolnshire rust (B). The NLR orthogroups OG0000122, OG0000123 and OG0000124 are coloured in red and the orthogroups OG0000043 and OG0000048 are coloured in blue.

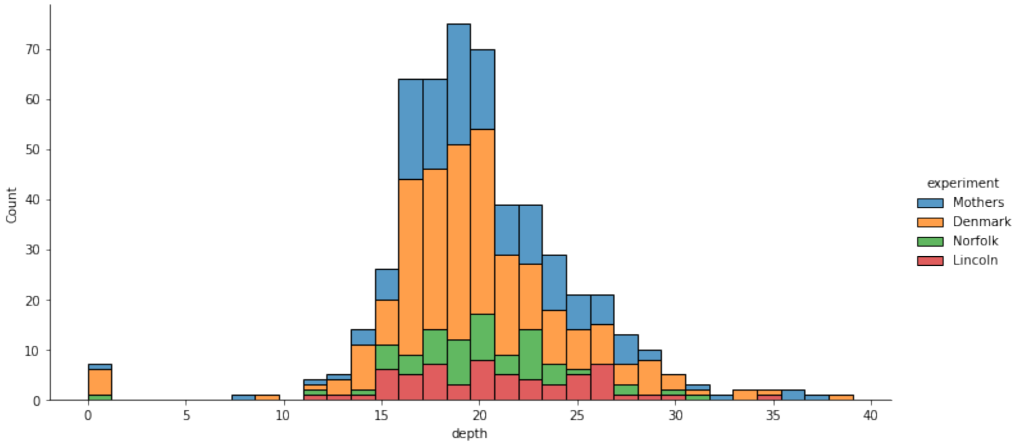


Figure S2 - Sequencing depth of the 520 re-sequenced sea beets.

The amount of data generated per sample was divided by the mean genome size of the 11 assemblies generated in the Chapter II (i.e. 714.182 Mb). The samples are coloured depending on where they have been collected: in one of the 3 large-scale association experiments (Norfolk, Lincolnshire or Denmark), or directly in their natural environment (Mothers).

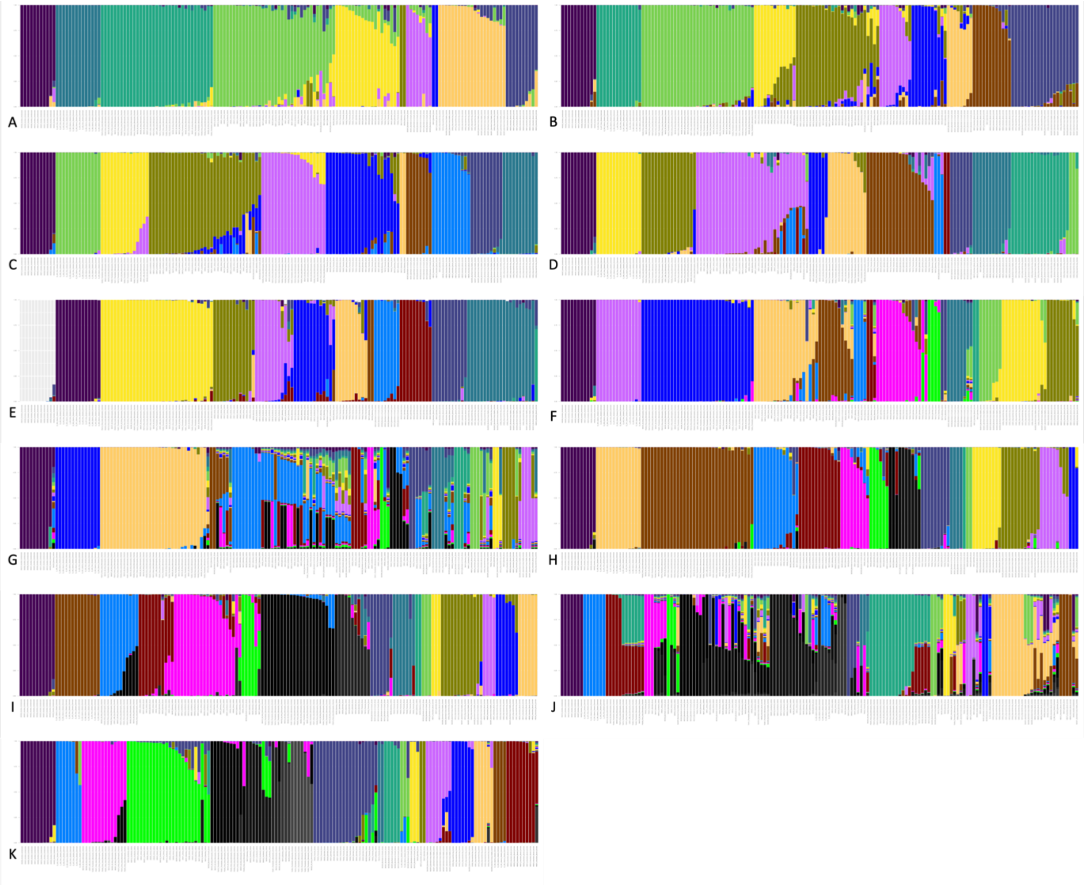


Figure S3 - Populations generated by PopCluster for *K*=10 (A) to *K*=20 (K).

Genome	Total number of shared *k*-mers in the reads	Estimated sequencing depth according to the estimated sea beet genome size	Estimated sequencing depth according to the estimated sugar beet genome size
Gb_Norfolk_426	22,172,131,108	39.1	29.9
Gb_Essex_038	27,711,071,730	48.9	37.4
Gb_Norfolk_095	31,726,002,724	56.0	42.8
Es_Catalonia_378	35,350,821,827	62.3	47.7
Gb_Merseyside_109	18,548,113,238	32.7	25.0
Gb_Humber_260	25,002,360,407	44.1	33.7
Dk_Sjælland_406	27,246,219,489	48.1	36.8
Fr_Bretagne_309	25,167,093,224	44.4	34.0
Gb_Merseyside_206	14,718,855,992	26.0	19.9
Gb_Suffolk_251	39,220,129,763	69.2	52.9
Gb_Essex_167	23,412,666,150	41.3	31.6

Table S1 - Genome sequencing depth estimation based on published beet genome size estimations.

Sea beet estimated genome size: 567 Mbp; sugar beet estimated genome size: 741 Mbp[16].

Number of genes	233,208
Number of genes in orthogroups	232,623
Number of unassigned genes	585
Percentage of genes in orthogroups	99.7
Percentage of unassigned genes	0.3
Number of orthogroups	19,454
Number of genome-specific orthogroups	265
Number of genes in genome-specific orthogroups	3,119
Percentage of genes in genome-specific orthogroups	1.3
Mean orthogroup size	12
Median orthogroup size	11
Number of orthogroups with all genomes present	17,736
Number of single-copy orthogroups	16,250

Table S2 - OrthoFinder statistics on annotated genes from the eleven sea beet assemblies.

Number of genes	2,157
Number of genes in orthogroups	2,152
Number of unassigned genes	5
Percentage of genes in orthogroups	99.8
Percentage of unassigned genes	0.2
Number of orthogroups	215
Number of genome-specific orthogroups	0
Number of genes in genome-specific orthogroups	0
Percentage of genes in genome-specific orthogroups	0
Mean orthogroup size	10
Median orthogroup size	11
Number of orthogroups with all genomes present	101
Number of single-copy orthogroups	58

Table S3 - OrthoFinder statistics on NLR loci extracted from the eleven sea beet genomes.

Genome	NLR loci	Core NLRome	Soft-Core NLRome	Shell NLRome	Cloud NLRome
Gb_Essex_167	192	123	143	48	1
Gb_Suffolk_251	194	131	149	45	0
Gb_Merseyside_206	186	121	143	43	0
Fr_Bretagne_309	207	126	148	59	0
Dk_Sjaelland_406	201	131	148	53	0
Gb_Humber_260	185	120	138	45	2
Gb_Merseyside_109	195	127	148	46	1
Es_Catalonia_378	205	138	163	42	0
Gb_Norfolk_095	201	127	144	57	0
Gb_Essex_038	196	132	151	44	1
Gb_Norfolk_426	195	129	148	47	0
Total	2,157	1,405	1,623	529	5

Table S4 - Number of NLR loci identified in the eleven genomes.

Genome	Core genome	Soft-core genome	Shell genome	Cloud genome	Total number of genes
Gb_Essex_167	19,738	20,336	600	219	21,155
Gb_Suffolk_251	19,616	20,193	559	382	21,134
Gb_Merseyside_206	19,762	20,309	567	256	21,132
Fr_Bretagne_309	19,632	20,318	548	266	21,132
Dk_Sjaelland_406	19,751	20,324	599	227	21,150
Gb_Humber_260	19,611	20,165	715	301	21,181
Gb_Merseyside_109	19,640	20,215	568	370	21,153
Es_Catalonia_378	19,761	20,293	567	275	21,135
Gb_Norfolk_095	19,699	20,241	610	330	21,181
Gb_Essex_038	19,808	20,351	556	201	21,108
Gb_Norfolk_426	19,693	20,254	616	292	21,162
Total	216,711	222,999	6,505	3,119	232,623

Table S5 - Number of genes annotated in the eleven sea beet genomes.

Genome name	Danish rust resistance	Are there sisters resistant to English rust?	Pigmentation	Additional phenotypes
Gb_Norfolk_426	Yes	Yes	Petioles	/
Gb_Essex_038	Yes	Yes	/	Stalk, leaf waxiness
Gb_Norfolk_095	Yes	Yes	Edges	/
Es_Catalonia_378	Yes	No	Petioles	/
Gb_Merseyside_109	Yes	Yes	Petioles	Leaf waxiness
Gb_Humber_260	No	Yes	Petioles	Trichomes, leaf waxiness
Dk_Sjælland_406	No	Yes	/	Leaf waxiness
Fr_Bretagne_309	No	Yes	Petioles	/
Gb_Merseyside_206	Yes	Yes	Petioles + Veins + Edges	Leaf waxiness
Gb_Suffolk_251	Yes	Yes	Petioles + Veins + Edges	Leaf waxiness
Gb_Essex_167	Yes	Yes	/	Trichomes, leaf waxiness, bolting

Table S6 - Phenotypes of the 11 genomes selected from the Danish large-scale association study (Chapter III), for whole genome PacBio HiFi sequencing.

Orthogroup	Gb_Essex_038	Gb_Norfolk_095	Gb_Merseyside_109	Gb_Essex_167	Gb_Merseyside_206	Gb_Suffolk_251	Gb_Humber_260	Fr_Bretagne_309	Es_Catalonia_378	Dk_Sjælland_406	Gb_Norfolk_426	Total
OG0000001	4	4	2	4	3	3	4	3	5	6	3	41
OG0000043	1	1	1	1	1	1	1	1	2	1	1	12
OG0000048	1	1	1	1	1	1	1	1	1	2	1	12
OG0000057	1	1	1	1	1	1	1	1	1	1	1	11
OG0000099	1	1	1	1	1	1	1	1	1	1	1	11
OG0000122	1	1	1	1	1	1	1	1	0	1	1	10
OG0000123	1	1	1	1	1	1	1	1	0	1	1	10
OG0000124	1	1	1	1	1	1	1	1	0	1	1	10
OG0000165	1	1	0	0	1	0	0	0	0	1	1	5

Table S7 - Number of NLR genes per orthogroup in the 11 sea beet assemblies.

Country	Site/Agricultural provider	Number of plants subject to rust inoculation		
		Norfolk experiment	Lincolnshire experiment	Danish experiment
NA	KWS	24	22	48
NA	BBRO	2	2	4
England (east coast)	Benfleet	48	46	47
	Burnham-on-Crouch	20	20	20
	Burnham-Overy-Staithe	38	36	36
	Cley-next-the-Sea	38	38	38
	Heacham_Beach	36	36	36
	Humber	36	35	36
	Bawdsey	28	28	28
	Orford	40	37	37
	Southwold	47	48	42
	Thurrock	41	39	40
England (west coast)	Little_Eye	8	7	8
	Thurstaston	42	41	42
	W_Kirby	67	66	65
	Park_Gate	32	32	32
France	PAL	7	5	6
	ROS	5	5	4
Spain	CRE	1	0	0
	MUN	1	1	1
Denmark	Faro	9	9	9
	Glaeno	10	9	10

Table S8 - Number of plants involved in the three large-scale rust inoculation trials

Sample name	Total data generated (Gb)	Sequencing depth	Coverage	Experiment	Sampling location
Ba001	13.05	18.27	99.84%	nature	Suffolk
Ba002	11.54	16.15	99.87%	nature	Suffolk
Ba005	14.88	20.84	99.83%	nature	Suffolk
Ba006	13.45	18.84	99.86%	nature	Suffolk
Ba007	12.19	17.06	99.82%	nature	Suffolk
Ba008	15.01	21.02	99.85%	nature	Suffolk
Ba009	12.63	17.68	99.86%	nature	Suffolk
Ba010	12.77	17.89	99.84%	nature	Suffolk
Ba011	12.81	17.94	99.86%	nature	Suffolk
Ba012	12.55	17.57	99.86%	nature	Suffolk
Ba013	12.79	17.91	99.85%	nature	Suffolk
Ba014	14.90	20.87	99.86%	nature	Suffolk

Ba015	13.91	19.48	99.83%	nature	Suffolk
BA016	18.91	26.47	99.85%	nature	Suffolk
BA258	12.92	18.09	99.82%	nature	Suffolk
BC001	17.73	24.82	99.85%	nature	Essex
BC002	20.90	29.26	99.84%	nature	Essex
BC003	25.06	35.09	99.85%	nature	Essex
BC006	20.10	28.15	99.86%	nature	Essex
BC007	19.02	26.64	99.85%	nature	Essex
BC008	20.68	28.96	99.85%	nature	Essex
BC009	23.80	33.32	99.86%	nature	Essex
BC012	18.61	26.06	99.85%	nature	Essex
BC013	15.96	22.35	99.87%	nature	Essex
BC015	18.33	25.66	99.86%	nature	Essex
BN001	9.44	13.21	99.85%	nature	Essex
BN002	10.38	14.53	99.86%	nature	Essex
BN003	11.99	16.79	99.85%	nature	Essex
BN004	13.75	19.25	99.84%	nature	Essex
BN005	11.75	16.45	99.86%	nature	Essex
BN006	10.89	15.24	99.83%	nature	Essex
BN008	13.11	18.36	99.84%	nature	Essex
BN009	12.43	17.40	99.86%	nature	Essex
BN011	12.40	17.36	99.86%	nature	Essex
BN012	12.10	16.94	99.83%	nature	Essex
BN015	14.10	19.74	99.86%	nature	Essex
BoS001	20.15	28.22	99.89%	nature	Norfolk
BoS002	13.90	19.47	99.90%	nature	Norfolk
BoS003	13.58	19.02	99.89%	nature	Norfolk
BoS005	14.30	20.03	99.89%	nature	Norfolk
BoS006	16.00	22.41	99.89%	nature	Norfolk
BoS007	14.50	20.30	99.89%	nature	Norfolk
BoS008	14.07	19.70	99.90%	nature	Norfolk
BoS011	14.36	20.10	99.91%	nature	Norfolk
BoS012	0.17	0.24	99.87%	nature	Norfolk
BoS014	14.17	19.84	99.87%	nature	Norfolk
Cl001	17.03	23.85	99.87%	nature	Norfolk
Cl003	13.40	18.76	99.85%	nature	Norfolk
Cl004	0.64	0.90	99.85%	nature	Norfolk
Cl005	13.44	18.81	99.88%	nature	Norfolk
Cl006	13.27	18.58	99.87%	nature	Norfolk
Cl007	12.77	17.89	99.86%	nature	Norfolk
Cl008	14.36	20.10	99.87%	nature	Norfolk
Cl009	11.67	16.35	99.88%	nature	Norfolk

Cl010	12.37	17.33	99.89%	nature	Norfolk
Cl012	14.26	19.97	99.84%	nature	Norfolk
Cl015	11.95	16.73	99.87%	nature	Norfolk
D001A	11.60	16.25	99.85%	denmark	Suffolk
D001B	20.07	28.10	99.84%	denmark	Suffolk
D001D	13.27	18.58	99.83%	denmark	Suffolk
D002A	12.49	17.49	99.85%	denmark	Suffolk
D008C	13.26	18.57	99.85%	denmark	Suffolk
D008D	11.53	16.14	99.85%	denmark	Suffolk
D013A	17.08	23.91	99.85%	denmark	Suffolk
D013B	13.05	18.27	99.84%	denmark	Suffolk
D013C	14.56	20.38	99.84%	denmark	Suffolk
D013D	11.98	16.77	99.84%	denmark	Suffolk
D015A	13.90	19.46	99.84%	denmark	Suffolk
D015B	15.91	22.27	99.83%	denmark	Suffolk
D016A	12.18	17.06	99.84%	denmark	Suffolk
D016B	17.72	24.82	99.82%	denmark	Suffolk
D016C	11.72	16.42	99.82%	denmark	Suffolk
D016D	19.77	27.68	99.84%	denmark	Suffolk
D038A	13.15	18.41	99.84%	denmark	Essex
D038B	12.90	18.07	99.85%	denmark	Essex
D038C	13.07	18.30	99.84%	denmark	Essex
D040B	12.70	17.78	99.85%	denmark	Essex
D040C	17.67	24.74	99.87%	denmark	Essex
D040D	9.88	13.84	99.79%	denmark	Essex
D042B	15.03	21.04	99.85%	denmark	Essex
D042C	15.61	21.85	99.84%	denmark	Essex
D042D	13.28	18.60	99.83%	denmark	Essex
D047B	13.32	18.65	99.85%	denmark	Norfolk
D047D	15.23	21.32	99.86%	denmark	Norfolk
D050A	15.08	21.11	99.86%	denmark	Essex
D050D	16.67	23.34	99.86%	denmark	Essex
D055A	15.70	21.99	99.83%	denmark	Suffolk
D061A	12.34	17.27	99.84%	denmark	Humberside
D061C	11.98	16.77	99.82%	denmark	Humberside
D065B	19.24	26.93	99.85%	denmark	Essex
D065D	20.62	28.88	99.85%	denmark	Essex
D068B	12.45	17.43	99.80%	denmark	Humberside
D068C	15.57	21.80	99.81%	denmark	Humberside
D068D	18.27	25.58	99.83%	denmark	Humberside
D071D	14.58	20.42	99.84%	denmark	Suffolk
D074A	14.71	20.59	99.83%	denmark	Humberside

D074B	13.37	18.72	99.81%	denmark	Humberside
D074C	15.58	21.81	99.81%	denmark	Humberside
D076B	0.17	0.24	99.68%	denmark	Suffolk
D076D	19.67	27.55	99.86%	denmark	Suffolk
D077A	9.31	13.04	99.81%	denmark	Suffolk
D077B	17.00	23.81	99.82%	denmark	Suffolk
D077D	21.12	29.57	99.83%	denmark	Suffolk
D080B	13.38	18.74	99.82%	denmark	Merseyside_Cheshire
D080C	12.75	17.86	99.79%	denmark	Merseyside_Cheshire
D080D	11.96	16.74	99.83%	denmark	Merseyside_Cheshire
D081A	12.92	18.09	99.88%	denmark	Norfolk
D081B	10.50	14.70	99.87%	denmark	Norfolk
D081C	11.89	16.65	99.89%	denmark	Norfolk
D083C	11.44	16.01	99.90%	denmark	Essex
D083D	14.06	19.69	99.82%	denmark	Essex
D088C	14.05	19.67	99.83%	denmark	Essex
D089AB	14.44	20.21	99.84%	denmark	Essex
D092A	15.83	22.17	99.83%	denmark	Essex
D092B	13.44	18.81	99.81%	denmark	Essex
D092D	11.14	15.59	99.84%	denmark	Essex
D095C	14.71	20.59	99.88%	denmark	Norfolk
D097A	14.63	20.48	99.84%	denmark	Essex
D097C	27.89	39.05	99.87%	denmark	Essex
D101C	20.20	28.28	99.83%	denmark	Suffolk
D103B	14.36	20.11	99.85%	denmark	Norfolk
D103D	18.74	26.24	99.86%	denmark	Norfolk
D105B	18.32	25.65	99.81%	denmark	Norfolk
D109A	15.46	21.64	99.83%	denmark	Merseyside_Cheshire
D112B	24.93	34.91	99.85%	denmark	Essex
D112C	14.02	19.64	99.80%	denmark	Essex
D112D	11.94	16.71	99.83%	denmark	Essex
D114B	14.26	19.97	99.78%	denmark	Humberside
D117B	14.56	20.39	99.79%	denmark	Suffolk
D117D	12.58	17.62	99.82%	denmark	Suffolk
D120B	16.45	23.04	99.83%	denmark	Humberside
D120C	15.13	21.18	99.82%	denmark	Humberside
D120D	0.22	0.31	99.71%	denmark	Humberside
D124B	12.40	17.36	99.83%	denmark	Essex
D124D	14.66	20.52	99.83%	denmark	Essex
D132A	11.55	16.17	99.83%	denmark	Suffolk
D132B	21.51	30.12	99.85%	denmark	Suffolk
D132C	17.04	23.86	99.84%	denmark	Suffolk

D132D	19.06	26.69	99.85%	denmark	Suffolk
D134C	0.64	0.89	99.72%	denmark	Essex
D135A	12.05	16.87	99.84%	denmark	Essex
D135C	14.59	20.43	99.88%	denmark	Essex
D138C	10.26	14.36	99.83%	denmark	Essex
D138D	14.10	19.74	99.83%	denmark	Essex
D145B	20.81	29.13	99.89%	denmark	Norfolk
D145C	15.22	21.31	99.88%	denmark	Norfolk
D145D	13.64	19.10	99.87%	denmark	Norfolk
D146A	12.50	17.50	99.84%	denmark	Essex
D146D	12.16	17.03	99.84%	denmark	Essex
D149D	15.56	21.79	99.83%	denmark	Suffolk
D151A	14.53	20.35	99.86%	denmark	Humberside
D151B	14.87	20.83	99.80%	denmark	Humberside
D155C	24.22	33.91	99.89%	denmark	Essex
D156B	13.44	18.82	99.85%	denmark	Norfolk
D156C	11.99	16.79	99.86%	denmark	Norfolk
D159D	18.26	25.57	99.86%	denmark	Essex
D161A	13.24	18.54	99.84%	denmark	Essex
D161D	13.70	19.19	99.85%	denmark	Essex
D162C	15.50	21.71	99.84%	denmark	Essex
D163D	15.28	21.39	99.83%	denmark	Humberside
D165B	16.53	23.14	99.83%	denmark	Essex
D165C	17.42	24.39	99.84%	denmark	Essex
D165D	14.65	20.52	99.85%	denmark	Essex
D166A	13.22	18.52	99.83%	denmark	Humberside
D167C	16.98	23.77	99.84%	denmark	Essex
D167D	16.15	22.61	99.85%	denmark	Essex
D172A	18.72	26.22	99.80%	denmark	Merseyside_Cheshire
D172C	9.44	13.22	99.82%	denmark	Merseyside_Cheshire
D172D	12.17	17.04	99.80%	denmark	Merseyside_Cheshire
D175D	18.83	26.36	99.82%	denmark	Merseyside_Cheshire
D179A	9.89	13.85	99.84%	denmark	Merseyside_Cheshire
D180D	17.63	24.69	99.83%	denmark	Merseyside_Cheshire
D183B	14.58	20.42	99.77%	denmark	Merseyside_Cheshire
D183C	12.78	17.89	99.75%	denmark	Merseyside_Cheshire
D183D	16.89	23.65	99.74%	denmark	Merseyside_Cheshire
D185B	13.03	18.25	99.82%	denmark	Merseyside_Cheshire
D186B	21.00	29.41	99.85%	denmark	Merseyside_Cheshire
D187A	14.07	19.70	99.81%	denmark	Merseyside_Cheshire
D187C	18.66	26.12	99.86%	denmark	Merseyside_Cheshire
D187D	16.51	23.12	99.83%	denmark	Merseyside_Cheshire

D188A	16.54	23.15	99.83%	denmark	Merseyside_Cheshire
D188C	15.61	21.86	99.83%	denmark	Merseyside_Cheshire
D189A	15.91	22.27	99.80%	denmark	Merseyside_Cheshire
D189C	20.34	28.48	99.84%	denmark	Merseyside_Cheshire
D192C	11.48	16.08	99.80%	denmark	Merseyside_Cheshire
D193D	13.38	18.73	99.81%	denmark	Merseyside_Cheshire
D194D	11.82	16.55	99.84%	denmark	Merseyside_Cheshire
D195B	13.12	18.37	99.81%	denmark	Merseyside_Cheshire
D196D	13.27	18.57	99.79%	denmark	Merseyside_Cheshire
D198C	15.40	21.56	99.82%	denmark	Merseyside_Cheshire
D201C	10.41	14.57	99.75%	denmark	Merseyside_Cheshire
D204A	13.04	18.25	99.81%	denmark	Merseyside_Cheshire
D205A	15.67	21.94	99.84%	denmark	Merseyside_Cheshire
D206B	15.01	21.02	99.83%	denmark	Merseyside_Cheshire
D206D	14.16	19.82	99.84%	denmark	Merseyside_Cheshire
D207B	14.01	19.62	99.84%	denmark	Merseyside_Cheshire
D218A	13.27	18.58	99.81%	denmark	Merseyside_Cheshire
D218D	12.99	18.19	99.81%	denmark	Merseyside_Cheshire
D219B	11.18	15.66	99.81%	denmark	Merseyside_Cheshire
D219C	14.91	20.88	99.81%	denmark	Merseyside_Cheshire
D220C	0.22	0.31	99.82%	denmark	Merseyside_Cheshire
D222C	14.64	20.50	99.84%	denmark	Merseyside_Cheshire
D223A	18.38	25.74	99.82%	denmark	Merseyside_Cheshire
D223B	13.88	19.43	99.82%	denmark	Merseyside_Cheshire
D223D	18.09	25.32	99.84%	denmark	Merseyside_Cheshire
D226B	11.34	15.88	99.81%	denmark	Merseyside_Cheshire
D226D	13.39	18.75	99.78%	denmark	Merseyside_Cheshire
D227B	20.05	28.08	99.82%	denmark	Merseyside_Cheshire
D227D	11.49	16.09	99.80%	denmark	Merseyside_Cheshire
D230A	13.12	18.37	99.87%	denmark	Norfolk
D230B	12.13	16.98	99.83%	denmark	Norfolk
D232A	14.06	19.69	99.86%	denmark	Norfolk
D232B	14.37	20.13	99.84%	denmark	Norfolk
D233A	13.40	18.76	99.87%	denmark	Norfolk
D233B	14.46	20.24	99.89%	denmark	Norfolk
D233C	18.15	25.41	99.87%	denmark	Norfolk
D234A	15.78	22.09	99.88%	denmark	Norfolk
D234B	15.65	21.91	99.87%	denmark	Norfolk
D236B	13.11	18.35	99.87%	denmark	Norfolk
D239C	12.15	17.01	99.86%	denmark	Norfolk
D240A	13.41	18.78	99.86%	denmark	Norfolk
D240B	12.13	16.99	99.83%	denmark	Norfolk

D240D	11.48	16.08	99.82%	denmark	Norfolk
D241A	15.64	21.89	99.84%	denmark	Norfolk
D246A	14.02	19.63	99.88%	denmark	Norfolk
D247A	19.29	27.01	99.83%	denmark	Suffolk
D248C	13.09	18.33	99.85%	denmark	Suffolk
D249A	19.24	26.95	99.84%	denmark	Suffolk
D249B	15.33	21.46	99.83%	denmark	Suffolk
D249C	0.01	0.01	99.68%	denmark	Suffolk
D250B	12.81	17.94	99.83%	denmark	Suffolk
D250D	13.21	18.50	99.83%	denmark	Suffolk
D251B	14.83	20.77	99.84%	denmark	Suffolk
D253B	13.55	18.98	99.84%	denmark	Suffolk
D254B	15.91	22.28	99.84%	denmark	Suffolk
D255A	13.02	18.22	99.79%	denmark	Suffolk
D255B	20.07	28.10	99.85%	denmark	Suffolk
D255C	12.99	18.18	99.83%	denmark	Suffolk
D257B	19.42	27.19	99.83%	denmark	Suffolk
D257C	10.21	14.29	99.81%	denmark	Suffolk
D258B	12.39	17.35	99.85%	denmark	Suffolk
D260A	13.87	19.42	99.80%	denmark	Humberside
D260C	12.03	16.84	99.84%	denmark	Humberside
D260D	12.79	17.92	99.83%	denmark	Humberside
D281	12.78	17.90	99.79%	denmark	Brittany_France
D282	22.06	30.89	99.82%	denmark	Brittany_France
D288	15.01	21.01	99.81%	denmark	Brittany_France
D297	12.62	17.67	99.80%	denmark	Brittany_France
D300	12.53	17.54	99.79%	denmark	Brittany_France
D301	18.34	25.68	99.82%	denmark	Brittany_France
D378	12.17	17.04	99.81%	denmark	Mediterranean_Spain
D384	12.77	17.88	99.80%	denmark	Zealand_Denmark
D388	13.70	19.19	99.77%	denmark	Zealand_Denmark
D389	16.09	22.53	99.80%	denmark	Zealand_Denmark
D391	15.09	21.13	99.79%	denmark	Zealand_Denmark
D392	14.33	20.06	99.75%	denmark	Zealand_Denmark
D393	16.04	22.46	99.79%	denmark	Zealand_Denmark
D394	10.61	14.85	99.78%	denmark	Zealand_Denmark
D395	13.42	18.80	99.81%	denmark	Zealand_Denmark
D400	12.78	17.90	99.79%	denmark	Zealand_Denmark
D404	14.86	20.81	99.79%	denmark	Zealand_Denmark
D407	14.31	20.03	99.78%	denmark	Zealand_Denmark
D408	13.58	19.02	99.79%	denmark	Zealand_Denmark
D409	13.89	19.45	99.79%	denmark	Zealand_Denmark

D412B	10.86	15.21	99.83%	denmark	Essex
D413B	11.59	16.23	99.83%	denmark	Essex
D413D	13.46	18.85	99.82%	denmark	Essex
D416D	15.88	22.23	99.83%	denmark	Essex
D417B	12.47	17.46	99.85%	denmark	Norfolk
D417C	12.91	18.08	99.86%	denmark	Norfolk
D418B	8.69	12.17	99.84%	denmark	Norfolk
D418C	10.92	15.29	99.83%	denmark	Norfolk
D418D	12.19	17.06	99.86%	denmark	Norfolk
D419A	23.68	33.15	99.88%	denmark	Norfolk
D419B	16.15	22.61	99.86%	denmark	Norfolk
D419D	14.27	19.98	99.87%	denmark	Norfolk
D421A	17.68	24.76	99.88%	denmark	Norfolk
D423D	6.22	8.71	99.84%	denmark	Norfolk
D424A	0.05	0.08	99.73%	denmark	Norfolk
D424C	15.08	21.11	99.89%	denmark	Norfolk
D426A	13.44	18.81	99.99%	denmark	Norfolk
D426B	20.19	28.28	99.94%	denmark	Norfolk
D426D	14.35	20.09	99.91%	denmark	Norfolk
D428A	12.08	16.91	99.87%	denmark	Norfolk
D428B	15.69	21.97	99.89%	denmark	Norfolk
D428C	18.95	26.53	99.90%	denmark	Norfolk
D428D	18.88	26.43	99.90%	denmark	Norfolk
D429A	12.34	17.28	99.88%	denmark	Norfolk
D429B	14.43	20.21	99.86%	denmark	Norfolk
D429C	10.87	15.22	99.86%	denmark	Norfolk
D440A	13.73	19.22	99.88%	denmark	Norfolk
D440B	12.38	17.33	99.86%	denmark	Norfolk
D440C	13.78	19.30	99.87%	denmark	Norfolk
D440D	20.59	28.83	99.90%	denmark	Norfolk
D442A	13.41	18.78	99.79%	denmark	Merseyside_Cheshire
D442D	9.90	13.86	99.80%	denmark	Merseyside_Cheshire
D444A	13.25	18.55	99.85%	denmark	Suffolk
D445A	13.70	19.19	99.78%	denmark	Suffolk
D446C	12.43	17.40	99.80%	denmark	Merseyside_Cheshire
D446D	18.52	25.93	99.79%	denmark	Merseyside_Cheshire
D447A	12.64	17.69	99.81%	denmark	Merseyside_Cheshire
D447B	9.96	13.94	99.80%	denmark	Merseyside_Cheshire
D447C	15.43	21.60	99.84%	denmark	Merseyside_Cheshire
D447D	11.68	16.35	99.83%	denmark	Merseyside_Cheshire
D448A	16.00	22.40	99.83%	denmark	Merseyside_Cheshire
D448B	13.40	18.76	99.81%	denmark	Merseyside_Cheshire

D449A	14.74	20.64	99.81%	denmark	Merseyside_Cheshire
HB001	17.65	24.72	99.89%	nature	Norfolk
HB002	12.03	16.85	99.86%	nature	Norfolk
HB003	9.33	13.06	99.88%	nature	Norfolk
HB004	12.72	17.81	99.88%	nature	Norfolk
HB005	13.73	19.22	99.88%	nature	Norfolk
HB006	13.28	18.60	99.87%	nature	Norfolk
HB008	10.29	14.41	99.88%	nature	Norfolk
HB010	14.06	19.69	99.88%	nature	Norfolk
HB012	12.48	17.48	99.88%	nature	Norfolk
HB014	11.18	15.65	99.88%	nature	Norfolk
HB015	12.52	17.53	99.87%	nature	Norfolk
HN001	12.93	18.11	99.81%	nature	Humberside
HN002	15.88	22.24	99.81%	nature	Humberside
HN003	27.97	39.16	99.84%	nature	Humberside
HN005	13.47	18.86	99.81%	nature	Humberside
HN006	13.47	18.86	99.83%	nature	Humberside
HN007	14.35	20.09	99.81%	nature	Humberside
HN008	14.67	20.54	99.81%	nature	Humberside
HN009	19.45	27.23	99.85%	nature	Humberside
HN010	15.88	22.24	99.81%	nature	Humberside
HN011	16.12	22.57	99.86%	nature	Humberside
HN013	20.70	28.98	99.83%	nature	Humberside
L040B	13.67	19.14	99.87%	lincolnshire	Essex
L040C	10.89	15.25	99.86%	lincolnshire	Essex
L050B	13.34	18.67	99.85%	lincolnshire	Essex
L074B	11.75	16.46	99.82%	lincolnshire	Humberside
L076C	11.27	15.78	99.89%	lincolnshire	Suffolk
L076D	14.77	20.68	99.85%	lincolnshire	Suffolk
L079B	11.00	15.41	99.84%	lincolnshire	Essex
L081C	14.40	20.16	99.88%	lincolnshire	Norfolk
L081D	10.89	15.25	99.89%	lincolnshire	Norfolk
L095C	10.96	15.35	99.87%	lincolnshire	Norfolk
L101D	12.30	17.23	99.80%	lincolnshire	Suffolk
L103D	14.31	20.03	99.86%	lincolnshire	Norfolk
L107B	15.33	21.47	99.82%	lincolnshire	Suffolk
L112A	16.26	22.77	99.85%	lincolnshire	Essex
L112B	18.53	25.95	99.84%	lincolnshire	Essex
L114A	17.46	24.45	99.83%	lincolnshire	Humberside
L120C	13.38	18.74	99.84%	lincolnshire	Humberside
L145B	12.88	18.04	99.87%	lincolnshire	Norfolk
L151A	12.02	16.84	99.80%	lincolnshire	Humberside

L151B	11.79	16.51	99.83%	lincolnshire	Humberside
L156C	20.58	28.82	99.87%	lincolnshire	Norfolk
L163A	15.52	21.74	99.78%	lincolnshire	Humberside
L163B	12.32	17.25	99.82%	lincolnshire	Humberside
L166B	9.22	12.90	99.85%	lincolnshire	Humberside
L166C	12.69	17.77	99.82%	lincolnshire	Humberside
L172B	11.01	15.42	99.82%	lincolnshire	Merseyside_Cheshire
L175D	8.36	11.71	99.82%	lincolnshire	Merseyside_Cheshire
L180B	12.01	16.81	99.82%	lincolnshire	Merseyside_Cheshire
L186A	14.95	20.93	99.81%	lincolnshire	Merseyside_Cheshire
L186B	11.58	16.21	99.80%	lincolnshire	Merseyside_Cheshire
L186C	14.40	20.16	99.81%	lincolnshire	Merseyside_Cheshire
L187B	14.76	20.67	99.82%	lincolnshire	Merseyside_Cheshire
L187C	18.58	26.02	99.77%	lincolnshire	Merseyside_Cheshire
L187D	12.90	18.06	99.83%	lincolnshire	Merseyside_Cheshire
L188C	10.89	15.24	99.79%	lincolnshire	Merseyside_Cheshire
L194C	10.86	15.21	99.80%	lincolnshire	Merseyside_Cheshire
L198B	15.76	22.07	99.83%	lincolnshire	Merseyside_Cheshire
L202A	21.01	29.42	99.83%	lincolnshire	Merseyside_Cheshire
L202C	14.43	20.20	99.85%	lincolnshire	Merseyside_Cheshire
L220A	11.78	16.49	99.82%	lincolnshire	Merseyside_Cheshire
L226B	14.99	20.98	99.83%	lincolnshire	Merseyside_Cheshire
L239A	11.02	15.43	99.86%	lincolnshire	Norfolk
L240B	12.30	17.22	99.82%	lincolnshire	Norfolk
L240C	14.83	20.77	99.86%	lincolnshire	Norfolk
L241B	9.22	12.92	99.87%	lincolnshire	Norfolk
L260A	18.79	26.31	99.82%	lincolnshire	Humberside
L260C	14.28	20.00	99.84%	lincolnshire	Humberside
L378	12.31	17.24	99.85%	lincolnshire	Mediterranean_Spain
L404	16.24	22.75	99.81%	lincolnshire	Zealand_Denmark
L406	14.76	20.67	99.80%	lincolnshire	Zealand_Denmark
L428A	11.25	15.76	99.88%	lincolnshire	Norfolk
L428B	10.96	15.35	99.88%	lincolnshire	Norfolk
L440A	11.74	16.44	99.88%	lincolnshire	Norfolk
L440C	14.98	20.98	99.90%	lincolnshire	Norfolk
L442C	14.92	20.89	99.80%	lincolnshire	Merseyside_Cheshire
L443A	13.39	18.74	99.82%	lincolnshire	Merseyside_Cheshire
L443C	12.05	16.87	99.78%	lincolnshire	Merseyside_Cheshire
L443D	12.72	17.81	99.81%	lincolnshire	Merseyside_Cheshire
L446D	11.05	15.48	99.83%	lincolnshire	Merseyside_Cheshire
L447A	11.97	16.77	99.83%	lincolnshire	Merseyside_Cheshire
LE001	13.08	18.31	99.82%	nature	Merseyside_Cheshire

LEREL	11.46	16.05	99.85%	nature	Merseyside_Cheshire
N015A	13.15	18.41	99.86%	norfolk	Suffolk
N015B	13.12	18.37	99.84%	norfolk	Suffolk
N042A	14.05	19.68	99.85%	norfolk	Essex
N042B	8.38	11.73	99.85%	norfolk	Essex
N055C	22.21	31.09	99.83%	norfolk	Suffolk
N061B	14.90	20.87	99.80%	norfolk	Humberside
N061C	16.56	23.18	99.71%	norfolk	Humberside
N068B	13.17	18.44	99.81%	norfolk	Humberside
N071D	15.13	21.19	99.83%	norfolk	Suffolk
N074B	11.83	16.56	99.86%	norfolk	Humberside
N080A	14.44	20.22	99.82%	norfolk	Merseyside_Cheshire
N080B	15.96	22.35	99.84%	norfolk	Merseyside_Cheshire
N080C	12.37	17.33	99.90%	norfolk	Merseyside_Cheshire
N109A	13.12	18.37	99.81%	norfolk	Merseyside_Cheshire
N112A	15.43	21.60	99.82%	norfolk	Essex
N112D	13.46	18.84	99.84%	norfolk	Essex
N117D	16.31	22.84	99.81%	norfolk	Suffolk
N120B	21.06	29.49	99.82%	norfolk	Humberside
N120C	24.31	34.03	99.86%	norfolk	Humberside
N120D	14.72	20.61	99.91%	norfolk	Humberside
N151A	15.59	21.83	99.82%	norfolk	Humberside
N151D	14.85	20.80	99.81%	norfolk	Humberside
N160C	17.49	24.49	99.83%	norfolk	Merseyside_Cheshire
N161B	13.45	18.83	99.84%	norfolk	Essex
N161C	14.06	19.69	99.84%	norfolk	Essex
N162A	18.64	26.10	99.86%	norfolk	Essex
N162D	19.13	26.79	99.81%	norfolk	Essex
N163B	17.56	24.59	99.81%	norfolk	Humberside
N180A	14.08	19.71	99.83%	norfolk	Merseyside_Cheshire
N186A	16.61	23.26	99.84%	norfolk	Merseyside_Cheshire
N186C	16.19	22.67	99.84%	norfolk	Merseyside_Cheshire
N187C	11.63	16.28	99.76%	norfolk	Merseyside_Cheshire
N187D	15.26	21.37	99.80%	norfolk	Merseyside_Cheshire
N188A	14.99	20.99	99.81%	norfolk	Merseyside_Cheshire
N198B	18.63	26.09	99.84%	norfolk	Merseyside_Cheshire
N204A	21.24	29.74	99.83%	norfolk	Merseyside_Cheshire
N223A	16.99	23.79	99.83%	norfolk	Merseyside_Cheshire
N223B	18.88	26.43	99.83%	norfolk	Merseyside_Cheshire
N223C	13.43	18.81	99.84%	norfolk	Merseyside_Cheshire
N223D	18.35	25.70	99.82%	norfolk	Merseyside_Cheshire
N227A	14.83	20.76	99.80%	norfolk	Merseyside_Cheshire

N232A	20.74	29.04	99.85%	norfolk	Norfolk
N232B	13.71	19.19	99.84%	norfolk	Norfolk
N255A	16.32	22.86	99.84%	norfolk	Suffolk
N418B	21.13	29.58	99.87%	norfolk	Norfolk
N418D	13.58	19.02	99.87%	norfolk	Norfolk
N419A	14.66	20.53	99.87%	norfolk	Norfolk
N419D	12.15	17.01	99.87%	norfolk	Norfolk
N422A	15.86	22.21	99.88%	norfolk	Norfolk
N422B	15.26	21.37	99.90%	norfolk	Norfolk
N423A	0.05	0.08	99.88%	norfolk	Norfolk
N423B	15.63	21.89	99.85%	norfolk	Norfolk
N442C	0.01	0.01	99.83%	norfolk	Merseyside_Cheshire
N442D	15.11	21.15	99.83%	norfolk	Merseyside_Cheshire
N446A	15.74	22.04	99.82%	norfolk	Merseyside_Cheshire
N446C	21.61	30.26	99.81%	norfolk	Merseyside_Cheshire
N447A	13.44	18.81	99.82%	norfolk	Merseyside_Cheshire
N447D	14.08	19.72	99.84%	norfolk	Merseyside_Cheshire
N449A	15.07	21.10	99.82%	norfolk	Merseyside_Cheshire
N449D	13.34	18.67	99.82%	norfolk	Merseyside_Cheshire
OR001	11.57	16.20	99.83%	nature	Suffolk
OR002	12.05	16.88	99.85%	nature	Suffolk
OR003	14.44	20.22	99.85%	nature	Suffolk
OR005	12.93	18.11	99.82%	nature	Suffolk
OR006	12.61	17.66	99.83%	nature	Suffolk
OR007	10.53	14.74	99.85%	nature	Suffolk
OR008	15.05	21.08	99.88%	nature	Suffolk
OR009	15.50	21.70	99.85%	nature	Suffolk
OR011	14.63	20.48	99.86%	nature	Suffolk
OR012	14.59	20.43	99.84%	nature	Suffolk
OR013	12.21	17.09	99.85%	nature	Suffolk
OR014	14.41	20.17	99.85%	nature	Suffolk
PG001	11.99	16.79	99.83%	nature	Merseyside_Cheshire
PG002	13.80	19.32	99.83%	nature	Merseyside_Cheshire
PG003	16.19	22.67	99.80%	nature	Merseyside_Cheshire
PG005	15.17	21.24	99.80%	nature	Merseyside_Cheshire
PG006	13.40	18.76	99.78%	nature	Merseyside_Cheshire
PG007	13.29	18.61	99.81%	nature	Merseyside_Cheshire
PG009	8.69	12.17	99.85%	nature	Merseyside_Cheshire
PG010	12.00	16.80	99.78%	nature	Merseyside_Cheshire
PG011	10.91	15.28	99.82%	nature	Merseyside_Cheshire
PG012	12.99	18.19	99.78%	nature	Merseyside_Cheshire
PG014	9.90	13.86	99.82%	nature	Merseyside_Cheshire

PG015	9.93	13.90	99.81%	nature	Merseyside_Cheshire
SW001	17.11	23.96	99.85%	nature	Suffolk
SW002	15.59	21.84	99.81%	nature	Suffolk
SW006	17.07	23.90	99.85%	nature	Suffolk
SW009	15.11	21.15	99.84%	nature	Suffolk
SW010	17.74	24.85	99.86%	nature	Suffolk
SW011	16.05	22.47	99.84%	nature	Suffolk
SW012	18.45	25.83	99.83%	nature	Suffolk
SW014	19.33	27.07	99.84%	nature	Suffolk
SW015	16.95	23.73	99.85%	nature	Suffolk
SW016	15.95	22.33	99.83%	nature	Suffolk
SW018	16.72	23.41	99.84%	nature	Suffolk
SW021	17.78	24.89	99.84%	nature	Suffolk
TA001	14.08	19.71	99.79%	nature	Merseyside_Cheshire
TA002	14.98	20.97	99.82%	nature	Merseyside_Cheshire
TA003	11.90	16.67	99.85%	nature	Merseyside_Cheshire
TA004	14.93	20.90	99.83%	nature	Merseyside_Cheshire
TA005	12.15	17.01	99.71%	nature	Merseyside_Cheshire
TA006	14.64	20.50	99.83%	nature	Merseyside_Cheshire
TA007	10.61	14.86	99.84%	nature	Merseyside_Cheshire
TA008	10.21	14.29	99.83%	nature	Merseyside_Cheshire
TA010	11.12	15.57	99.84%	nature	Merseyside_Cheshire
TA011	12.76	17.87	99.81%	nature	Merseyside_Cheshire
TA013	13.64	19.09	99.81%	nature	Merseyside_Cheshire
TH002	11.59	16.23	99.85%	nature	Essex
TH003	11.48	16.08	99.85%	nature	Essex
TH004	12.14	17.00	99.85%	nature	Essex
TH005	12.74	17.84	99.85%	nature	Essex
TH006	6.16	8.62	99.83%	nature	Essex
TH007	10.88	15.23	99.83%	nature	Essex
TH008	14.31	20.03	99.85%	nature	Essex
TH009	11.55	16.17	99.82%	nature	Essex
TH014	12.80	17.93	99.83%	nature	Essex
TH015	11.35	15.89	99.86%	nature	Essex
Wk001	15.41	21.58	99.82%	nature	Merseyside_Cheshire
Wk003	12.18	17.05	99.81%	nature	Merseyside_Cheshire
Wk006	14.57	20.40	99.83%	nature	Merseyside_Cheshire
WK007	12.17	17.03	99.83%	nature	Merseyside_Cheshire
Wk008	14.48	20.28	99.83%	nature	Merseyside_Cheshire
Wk012	12.14	17.00	99.81%	nature	Merseyside_Cheshire
Wk014	13.39	18.75	99.82%	nature	Merseyside_Cheshire
Wk015	13.27	18.58	99.78%	nature	Merseyside_Cheshire

Wk016	13.21	18.50	99.81%	nature	Merseyside_Cheshire
Wk018	12.62	17.67	99.80%	nature	Merseyside_Cheshire
Wk019	13.43	18.80	99.83%	nature	Merseyside_Cheshire
Wk020	11.88	16.63	99.84%	nature	Merseyside_Cheshire

Table S9 - Sequencing data generated for the 520 sea beets part of the association studies.

The depth is calculated as the total amount of data generated through sequencing divided by the mean genome size of the 11 assemblies generated in the Chapter II (i.e. 714.182 Mb). The coverage corresponds to the percentage of genome covered at least one time when the reads are mapped to the Gb_Norfolk_426.

	Max distance = 100 kb		Max distance = 1 kb	
	Min r^2	Max r^2	Min r^2	Max r^2
Humberside	0.230176	0.403456	0.2773	0.456674
Merseyside/Cheshire	0.093179	0.239304	0.135827	0.286283
Norfolk	0.0849467	0.228722	0.133979	0.266157
Suffolk/Essex	0.0522637	0.167262	0.0934712	0.211748

Table S10 - Minimal and maximal r^2 values measured in LD analyses.

NLR orthogroup	Humberside	Merseyside	Norfolk	Suffolk/Essex	Association output	Class
OG0000124	0.00360066	0.00332935	0.00321203	0.00385288	High-mean	NBARC-LRR
OG0000122	0.00418805	0.00438264	0.00504981	0.00554522	High-mean	CNL
OG0000123	0.00520677	0.0084891	0.00856914	0.0074381	High-mean	CNL
OG0000048	0.0276388	0.0151277	0.0218917	0.0243438	High-score	CNL
OG0000043	0.00161126	0.00957335	0.0115617	0.00919506	High-score	CNL

Table S11 - π values of the five associated orthogroups in the four English populations.

NLR orthogroup	Humberside-Merseyside/Cheshire	Humberside-Norfolk	Humberside-Suffolk/Essex	Merseyside/Cheshire-Norfolk	Merseyside/Cheshire-Suffolk/Essex	Norfolk-Suffolk/Essex	Association output
OG0000124	0.0766724	0.0379032	0.0640597	0.0327105	0.0285576	0.0167171	High-mean
OG0000122	0.104368	0.0422933	0.0365337	0.0444015	0.0390271	0.0133849	High-mean
OG0000123	0.102161	0.0649783	0.0235038	0.0425734	0.0707188	0.0529458	High-mean
OG0000048	0.253334	0.0896839	0.0424885	0.247815	0.170246	0.0598287	High-score
OG0000043	0.0730697	0.110987	0.0243551	0.0747696	0.045936	0.046636	High-score

Table S12 - F_{ST} values of the five associated orthogroups for pairwise comparisons of four English populations.