CHARACTERIZATION OF RESISTANCE GENE DIVERSITY AND STRUCTURAL VARIATION IN *BETA VULGARIS*

By

Andrew Joseph Funk

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Plant Breeding, Genetics, and Biotechnology--Crop and Soil Sciences--Doctor of Philosophy

2020

ABSTRACT

CHARACTERIZATION OF RESISTANCE GENE DIVERSITY AND STRUCTURAL VARIATION IN B*ETA VULGARIS*

By

Andrew Joseph Funk

Plants are under threat from bacteria, viruses, fungi, and nematodes in their environment. Plants deploy a sophisticated network of defense responses to avoid and defeat these pathogens, and in response pathogens counteract plant defenses through their own suite of biochemical weapons and signaling molecules. These battles are pervasive in both natural settings and agriculture. The goal of this dissertation research is to provide insight into genetic variation present in diverse populations of *Beta vulgaris* and identify patterns of disease resistance (R) gene variation.

Nucleotide-binding (NB-ARC), leucine-rich-repeat genes (NLRs) account for 60.8% of R genes molecularly characterized from plants. NLRs exist as large gene families prone to tandem duplication and transposition, with high sequence diversity among crops and their wild relatives. I used the conserved NB-ARC domain to build a *B.vulgaris*-specific hidden Markov model (HMM). The HMM identified 231 tentative NB-ARC loci in a highly contiguous genome assembly of sugar beet, revealing diverged and truncated NB-ARC signatures as well as full-length sequences. The putative NB-ARC-associated proteins contained NLR resistance gene domains, including Toll/interleukin-1 receptor (TIR), coiled-coil (CC), and leucine-rich repeat (LRR), as well as other integrated domains.

HMM-based domain detection was extended to 23 populations encompassing four crop types of *B. vulgaris.* Whole-genome sequences were generated by pooling 25 individuals per population, then sequencing each population in a single bulk reaction using 2x150 bp chemistry.

These reads were assembled *de novo* to efficiently capture population-wide genetic variation. The nucleic-acid-based NB-ARC HMM was used to scan *de novo* contigs and infer genetic variation within and between populations, which identified an average of 139.5 NB-ARC domains per population.

The pooled population sequencing strategy was expanded to 71 populations total. Short reads were used in a targeted reassembly pipeline to detect structural variation in each of the 71 populations. This method identified 4,995,443 indels with lengths under 1 kb. These indels were analyzed for chromosome position, length in bp, and frequency across populations, and revealed non-random patterns of indel variation. Half of the indels were detected in five or more populations, suggesting that indel assembly from pooled population sequences is reproducible. Furthermore, indels were sufficient to differentiate populations by crop type, supporting the conclusion that the data modeled genetic differences originating in historical crop development. Divergence in the population-wide distribution of seven- and eight-bp indels led to identification of an enriched sequence motif, suggesting possible biological function of the sequence such as a TE target site duplication or transcription factor binding site.

This work presents the first detailed view of NLR family composition in a member of the *Caryophyllales* and demonstrates an additional nucleic-acid-based method for resistance gene prediction in non-model plant species. Pooled population sequencing was used to access novel variation in breeding populations of *B. vulgaris* and identify structural variants that reflected underlying genotypic relationships. Future work will build on resistance gene modeling, pooled population sequencing, and detection of genetic variation to aid breeding for disease resistance.

For Anna

ACKNOWLEDGEMENTS

First, thanks to my family. My parents Joe and Donna have always supported me and encouraged me that I could do anything I set my mind to. My sister Marjie has been a good example of diligence, patience, and kindness that persists today. Together they provided stability and love that I am eternally thankful for.

Thanks to the members of my committee Dave Douches, Linda Hanson, and Ryan Warner for their questions, insights, and patience. Kevin Folta gave me tremendous encouragement that I had a decent mind and something to offer science. Payam Mershahi and Shin Han Shiu demonstrated science as a noble pursuit. Dave Douches helped me find a way forward with my degree when I otherwise might have left. There are countless others at MSU who helped me understand science over the years, so to that community as a whole, thank you.

Special thanks to the FAST teaching community that helped me find something that I enjoy and might be good at. Jon Stoltzfus was invaluable as my mentor for undergraduate education, and Diane Ebert-May was excellent in her instruction and example of scientific teaching. Without them I never would have gotten my first job as a professor.

Thanks to Anna, my soon-to-be wife, for her loyalty and encouragement throughout this process. Without her, the road would have felt longer. Thanks to the friends who helped keep me sane and journeyed with me through all manner of fantastical worlds and adventures. Those roads will go on forever.

Above all, thanks to God, who made the world and everything in it, who gave Jesus as an example, who fixes broken things and shows us how to love.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

Plants are sessile and need to cope with their environment to survive. There is no fight-or-flight for plants, the decision is always fight. Pathogens are a constant threat to plant productivity, attacking both above and below ground, leading to the development of myriad strategies for plants to avoid or defeat pathogens. These strategies include systems for direct perception of pathogens, indirect perception, local and systemic signaling, and self-inflicted cell death via the hypersensitive response (Jones & Dangl 2006). These mechanisms can be active against general classes of pathogens as well as specific species, races, or strains (Kourelis & Van Der Hoorn 2018).

Attacks by pathogens create selection pressures on plant populations that reward novel variation able to mitigate the effects of the attacker. Once a plant develops a defense against a certain pathogen, it creates new pressure selecting for variants of the pathogen able to counteract the defense response (Brown 2015). This cycle means defense responses are rarely solved, rather they are in a constant state of flux as selection pressure oscillates between host and pathogen. This in turn encourages cycles of birth and death of novel genetic variation, as genetic drift, mutation, and recombination create a maelstrom of genetic components with unforeseen effects on fitness (Zhang et al. 2014; Shao et al. 2014). This interplay is evidenced by the large families of some resistance (R) genes, which can number from hundreds to thousands depending on the specific plant species under consideration (Monteiro & Nishimura 2018).

NLR AND RECEPTOR-LIKE KINASE RESISTANCE GENES

The most well-studied R genes produce intracellular proteins containing nucleotide binding (NB) and leucine-rich repeat (LRR) domains, collectively known as NB-LRR (NLR)

(Kourelis & Van Der Hoorn 2018). These NLR genes are often divided into two groups based on their N-terminal signaling domain, which often consists of a coiled-coil (CC) domain or Toll/Interleukin-1-like receptor (TIR) domain (Jones & Dangl 2006). NLRs are found in mammals as well as plants, although they appear to have distinct evolutionary origins (Maekawa et al. 2011). NLRs comprise 61% (191/314) of the R genes cloned from plants to date (Kourelis & Van Der Hoorn 2018). The second-most studied class encodes membrane-bound receptor-like kinases (RLKs) and associated receptor-like proteins (RLPs), collectively representing an additional 19% (60/314) of cloned R genes. Unfortunately, the pathogen elicitor, mechanism of resistance, or both, remains unknown for 70% (177/251) of cloned mediators of resistance (Kourelis & Van Der Hoorn 2018).

How genomic features contribute to the development of novel disease resistance is a core question of this dissertation research. Gene duplication is a fundamental process generating new variants of NLRs and RLKs (Leister 2004; Lauer et al. 2018). The mechanisms of gene duplication are varied, including whole-genome and segmental duplications, tandem duplication, transposition, non-allelic homologous recombination, and replication-based mechanisms related to altered DNA polymerase template recognition (Carvalho & Lupski 2016). These mechanisms can be facilitated by homologous sequences in close physical proximity, whether on the same chromosome or on different chromosomes brought together during DNA replication and repair. Homologous sequences could include repetitive sequences such as transposable elements, other low-complexity repetitive sequences, or homologous genes sharing high sequence identity. Interactions between similar sequences can lead to gene duplications, deletions, and recombinations that shuffle genes, promoters, exons, and domains between loci, leading to novel organizations of coding regions (Shao et al. 2014; Bailey et al. 2018). NLRs and RLKs are often

found clustered in the genome and interspersed with repetitive sequences (Shiu & Bleecker 2003; Pan et al. 2000). These repetitive genomic landscapes create difficulty when resolving the identity and location of R genes.

NLRs and RLKs contain conserved sequences that are required for their core function and diverse sequences required for specific resistance interactions. This fact provides an opportunity to detect candidate resistance genes based on the presence of conserved domains, and then investigate novel resistance mechanisms by characterizing accessory domains. Strong conservation occurs in the NB domain of NLRs, which are constrained by the need to interact with ATP and ADP (van der Biezen & Jones 1998). The Serine/Threonine kinase domain of RLKs is similarly constrained due to the need for functioning kinase activity to propagate signals to downstream partners (Shiu & Bleecker 2001). It is not universally true that NB and kinase domains be functional to participate in the defense response network: functional complementation by heteromeric protein-protein interactions allows perception and signaling domains to be supplied by separate entities (Franck et al. 2018; Jones et al. 2016). However, the majority of cloned NLRs and RLKs contain conserved NB and kinase domains and thus this conservation remains a viable guideline for identifying resistance genes in a genomic context. Unfortunately, the high sequence similarity between some family members can confound attempts to resolve similar loci using short-read sequencing techniques, which remain the most common and cost-effective approach for whole-genome sequencing and assembly. Alternative methods are needed for cost-effective genomic analysis of resistance genes.

Duplicated and recombined R gene variants can provide candidates for new disease resistance, yet it is more likely their effects confer negative or neutral fitness rather than positive (Zhang et al. 2014; Leister 2004). Low fitness releases loci from selection pressure, leading to

the accumulation of mutations and indels through genetic drift (Keightley et al. 1998). For a time, these loci, even if non-functional, could act as a reservoir for genetic variation accessed through recombination. Eventually these sequences will become pseudogenes and be washed away into the genomic landscape (Freeling 2009). Alternatively, some loci confer high fitness to their host and are maintained through natural selection (Keightley et al. 1998). These loci propagate through lineages and populations.

STRUCTURAL VARIATION AND GENETIC DIVERSITY

In eukaryotes, genetic information does not exist in a vacuum, rather, it is carried along through the concerted action of biochemical processes that convey chromosomes from one generation to the next. Physical properties of chromosomes and the associated replicative machinery influence how information is maintained or altered over time (Lynch et al. 2016). Recombination is a fundamental source of organismal diversity, able to generate nearly infinite combinations of the thousands of gene and regulatory sequences found within populations (Ma et al. 2009). On a more basic level, sequence variation can arise either through single nucleotide polymorphisms (SNPs) or through structural changes to multiple nucleotides at once (Hodgkinson & Eyre-Walker 2011). Sequence differences between stretches of two or more nucleotides are termed structural variants and are often divided into short insertion/deletion variants <1,000 bp (indels) and longer structural variation >1,000 bp (SVs) (Wala et al. 2018).

The most common genotyping method in modern genetics is assessment of bi-allelic SNPs. However, there is reason to believe SNPs are not always in linkage disequilibrium with important genomic features. Pan-genomic investigation of maize revealed that 90% of the maize genome contained structural variants in at least one accession, and 70% of the genome had variations in 10 or more accessions (Chia et al. 2012). However, approximately 20% of structural

variants in the maize genome were not associated with SNPs (Chia et al. 2012). In humans, the number of structural variants without linked SNPs was slightly higher at 23% (Sudmant et al. 2015). One explanation for the lack of linkage between SNPs and structural variants is that generation of structural variation could occur more rapidly than SNP variation. This is an intriguing proposition given the rapid changes in selection pressure dynamics between pathogens and plant defenses.

Various lines of evidence support the hypothesis that structural variation is disproportionately responsible for phenotypic variation compared to SNPs. Indels were recently associated with ~80% of *de novo* gene evolution events in rice (Zhang et al. 2019), implicating indels as a major source of novel genes. Studies have shown that structural variation is disproportionately enriched near SNPs statistically associated with phenotypic variation in humans and maize (Sudmant et al. 2015; Chia et al. 2012). In a meta-genomic association mapping study of rice, 42% of trait-associated SNP loci were absent from the reference genome, suggesting those SNPs reside on indels (Yao et al. 2015). Increased resolution of structural variation should improve our understanding of relationships between genome biology and R gene diversity.

BETA VULGARIS

Research presented in this dissertation focuses on the crop plant *Beta vulgaris* spp. *vulgaris*. These plants are members of order Caryophyllales, situated at the base of the eudicot clade after divergence from monocots (Dohm et al. 2014). *B. vulgaris* spp. *vulgaris* has nine chromosomes with an estimated genome size between 569 and 758 MB and 42 to 63% repetitive sequences (Dohm et al. 2014; Arumuganathan & Earle 1991). The most economically important crop type is sugar beet, cultivated for processing of refined sugar from root extracts (Cooke &

Scott 1993). Table beet and chard are grown for direct human consumption, and the fourth crop type, fodder beet, is used as feed for livestock (Cooke & Scott 1993). *B. vulgaris* is naturally biennial and outcrossing, although modern plant breeders have adopted genetic self-compatibility to facilitate development of advanced cultivars (McGrath & Panella 2018).

Sugar beet was thought to contain only CC-NLR resistance genes until recently (Tian et al. 2004; Funk et al. 2018; Dohm et al. 2014). The predominance of CC-NLR genes groups *B. vulgaris* more closely with early plant lineages: *Selaginella moellendorffii, Brachypodium distachyon*, and the monocots maize, rice, and sorghum all have CC-NLRs but no TIR-NLRs. This is in contrast to eudicots such as Solanaceous species, *Arabidopsis,* and soybean, which all contain both CC and TIR NLRs (Jacob et al. 2013). How R genes evolved to fill different roles in different plant lineages is an ongoing subject of research.

In summary, NLRs and RLKs are key classes of resistance genes which exist as large gene families containing a continuum of functional and non-functional elements. These elements contain highly conserved as well as diverged domains and are often physically clustered together on chromosomes. The combination of sequence similarity and physical proximity creates difficulties when attempting to resolve similar loci. Still, given the central role of these types of genes in disease resistance, resolving each NLR and RLK in a species should provide increased resolution to determine candidate genes for deployment in crop production. Beyond the practical applications, characterizing the genomic functions that lead to novel disease resistance could lead to further insights into the mechanisms of disease resistance evolution more broadly. These insights could be applied to wild relatives to increase the search space for novel disease resistance genes, and ultimately could inform attempts to design disease resistance genes *de novo,* contributing to a new approach for mitigating the effects of pathogens in agriculture.

CHAPTER TWO

NUCLEOTIDE-BINDING RESISTANCE GENE SIGNATURES IN SUGAR BEET

INTRODUCTION

A wide range of organisms threaten the productivity of beets (*Beta vulgaris* L.), including bacteria, fungi, viruses, and nematodes (De Lucchi et al., 2017; Haverson et al., 2009; Leucker et al., 2016; Stevanato et al., 2014b, 2015; Webb et al., 2016). Sugar beet growers expend significant time and money protecting crops from pathogens, whether through chemical treatments, use of resistant varieties, or pursuing cultural practices. Deploying genetic resistance is a proven way to protect crop production while reducing costs and environmental impacts of agriculture (reviewed in Boyd et al., 2013). Because of this, disease resistance is one of the more important traits pursued by plant breeders. Complexities of the beet breeding system (e.g. self-incompatibility, biennial lifecycle) have generally precluded deep understanding of the inheritance and genetics of most disease resistance traits in sugar beet.

The NB-ARC protein domain is a nucleotide-binding (**NB**) domain initially found in human ***A****paf1*, plant **R** genes, and *Caenorhabditis elegans* ***C****ed4* (van der Biezen & Jones 1998). The majority (60.8%) of *R* genes cloned to date contain NB-ARC and leucine-rich repeat (LRR) motifs, which have been implicated in gene-for-gene resistance to pathogens (Jones & Dangl 2006; McHale et al. 2006; Maekawa et al. 2011; Kourelis & Van Der Hoorn 2018). NB-ARC-LRR (NLR) genes often exist as large, homologous gene families with conserved NB domains along with variable signaling, perception, and aggregation domains (Duxbury et al. 2016; Steinbrenner et al. 2015; Casey et al. 2016; Wang et al. 2015). Two common signaling domains are coiled-coil (CC) and Toll-interleukin1-receptor-like (TIR) at the N-terminus, and LRR domains are common at the C-terminus. NLR exist in smaller families in mammalian systems

(e.g. ~25 genes in humans) compared to hundreds of copies in plants (Sarris et al. 2016; Maekawa et al. 2011). Historically these genes have been difficult to identify via nucleotide sequence searches (Zhang et al. 2014). A recurring strategy in resistance gene identification has been to use the conserved NB-ARC domain as a predictor of putative resistance genes (Christopoulou et al. 2015; Jupe et al. 2012; Meyers et al. 2003; Monosi et al. 2004). Target-capture methods to enrich NLR-like sequences from genomic DNA, followed by motif-based annotation of CC, TIR, NB-ARC, and LRR domains is a promising additional means to identify potential R genes (Jupe et al. 2013; Steuernagel et al. 2015).

Beyond the classical gene-for-gene hypothesis of disease resistance (Flor 1942), experimental evidence shows that higher-order NLR complexes play an important role in many host-pathogen interactions (Belkhadir et al. 2004; Loutre et al. 2009; Césari et al. 2014; Sinapidou et al. 2004; Yuan et al. 2011; Huh et al. 2017). Multiple schemes for pathogen detection by NLRs have been established, including direct recognition of pathogen components, indirect recognition of pathogen effectors, and integration of target domains into host NLRs (reviewed in Jones *et al.*, 2016; Li *et al.*, 2015; Kourelis and van der Hoorn, 2018). Pairs of interacting NLRs are common, and the classic CC/TIR-NB-ARC-LRR domain organization is present across multiple interacting proteins, such that domains missing in one protein may be complemented by a partner (Bonardi et al. 2011; Hayashi et al. 2010; Nishimura et al. 2017). Based on these cited studies, determining the NLR complement in *B. vulgaris* should provide the basis for identification and deployment of new resistance genes for crop protection.

One of the most widely-deployed traits in the sugar beet industry, and one whose genetic basis is better known, is resistance to the root disease rhizomania, which is caused by *Beet necrotic yellow vein virus* (BNYVV) (Biancardi et al. 2002; Biancardi & Tamada 2016).

Rhizomania ('crazy root') is found worldwide, with symptoms of deformed hairy roots leading to reduced sugar yield. The only current control is genetic resistance conferred by either of two loci, *Rz1* and/or *Rz2*. These loci segregate as dominant, monogenic traits located between 5 and 25 cM from one another on Chromosome 3 (Barzen et al. 1997, 1992; Lewellen 1991; Scholten et al. 1999). The molecular basis of *Rz1* resistance is unknown. The nucleotide sequence of *Rz2* shows a classic CC-NLR located on Chromosome 3 (Capistrano-Gossmann et al. 2017). Other published sources of resistance also co-locate to the chromosomal region (Gidner et al. 2005; Grimmer et al. 2008, 2007).

Here, an 850 bp hidden Markov model (HMM) of a canonical beet NB-ARC domain was constructed from predicted mRNA nucleotide sequences and deployed to identify tentative NB-ARC sequences in a new sugar beet genome assembly (McGrath et al. unpublished). This genome was assembled from PacBio long reads, then scaffolded with BioNano optical mapping and Hi-C proximity-guided assembly. The resulting EL10 genome provides an opportunity to investigate duplicated and repetitive sequences that were previously difficult to resolve and place in their genomic context. The HMM scan revealed 231 tentative NB-ARC loci with homology to the NB-ARC model in this genome sequence.

RESULTS

Identification of NLR sequences in EL10

An HMM of *B. vulgaris* NB-ARC domains was developed using nucleotide sequences to allow direct interrogation of the genome assembly. A preliminary EL10 predicted protein set contained 185 proteins with NB-ARC domains. Transcript sequences corresponding to these predicted domains were filtered for lengths above 400 bp and Expectation values (e-values) below $1 \times e^{-10}$ then realigned and converted into an HMM of 850 bp. Probing the EL10 genome

with this nucleotide-based HMM identified 250 sequences with similarity to the NB-ARC domain model. Domain lengths ranged from 54 to 933 bp and included 183 of the 185 domains in the predicted protein set. Two domains from the predicted protein set that were not identified by nucleotide HMM had e-values of 3.1 e^{-6} and 3.3e^{-11}, suggesting divergence from both the model in the Pfam database as well as the HMM generated from the bulk of NB-ARCs in *B. vulgaris*.

The 250 initial domains were named Bv.nbarc.0001 through Bv.nbarc.0250 (Table S2.1). Some domains appeared to be incomplete fragments that could be joined together with a nearby fragment to form a single domain. In these cases, distances between fragments ranged from 39 bp to 7.2 kb. A BLAST search of the incomplete domains against the predicted gene set revealed that pairs of domains aligned to the same transcript, supporting the hypothesis that these fragments should be merged (Table S2.2). These fragments were joined with adjacent domains to form 18 distinct NB-ARC domains for subsequent analysis, including one set of three domains. This resulted in 231 tentative NB-ARC-like domains identified in the EL10 genome assembly, 48 of which were not present in the predicted proteins (Table S2.1).

Phylogenetic analysis and Subfamily classification of NB-ARC domains

NB-ARC sequences were arranged in a phylogenetic analysis to predict evolutionary relationships. The 231 nucleic acid HMM matches included 21 with e-values above 1e^{-10} and lengths less than 400 bp. These matches were withheld from the initial analysis due to difficulty incorporating partial sequences into the alignment and phylogeny. The remaining 210 high-confidence matches revealed distinct clades with bootstrap support between 50 and 100 (Figure 2.1, Figure 2.S1). NB-ARC subfamilies, as revealed by the phylogeny, were defined as sequences sharing at least 60% pairwise identity, resulting in 27 subfamilies identified in total,

named Sf. 1-27 (Figure 2.1). The 21 low-confidence HMM sequences were added to the initial tree using the evolutionary placement algorithm (EPA) of RAxML (Figure 2.1), resulting in 204 sequences assigned to 27 subfamilies, and 27 independent sequences (Table S2.1). The largest subfamily contained 22 loci, and six of the 27 subfamilies were composed of two loci. Domains with lengths less than 400 bp and/or above an HMM e-value of $1e^{-10}$ were distributed throughout the phylogeny.

Physical organization of NB-ARC domains and cluster analysis

Relationships between sequence identity and physical location of the NB-ARC domains were examined to assess gene family diversification. Genomic locations of NB-ARCs were classified as clusters or singletons (Figure 2.2). Clusters were defined as loci within 200 kb of their nearest neighbor. A total of 183 loci were found in 47 clusters, whereas 48 loci were singletons. There were NB-ARCs on each of the nine chromosomes, with 61.9% (143 of 231) occurring on Chromosomes 2, 3, and 7. Depending on whether the clustered loci came from same or different Subfamilies, 32 clusters were homogeneous and 15 were heterogeneous. Links between Subfamilies (Figure 2.3) were used to juxtapose the phylogenetic relationships and physical locations of domains in the genome. Some Subfamilies were clustered at one or two positions (e.g. Sf. 1, 4, 17, and 21), while others were distributed across five or more locations (e.g., Sf. 7, 8, 12, and 25).

Domain composition of proteins associated with genomic NB-ARC sequences

Predicted proteins associated with predicted NB-ARC domains were assayed for domain composition to determine the number and type of NLRs in the EL10 assembly. Preliminary gene annotations of the EL10 whole genome assembly overlapped 208 of the 231 NB-ARC predictions. The overlapping proteins were analyzed using InterProScan to determine their *cis*-

linked domain content (Supplementary file S1). Whole-protein domain organization indicated CC-NB-LRR was the most common domain arrangement (76 instances) followed by NB-LRR (69 instances). The most abundant domains observed were NB-ARC (183 predictions) and P-loop NTPase (180), which is a component of the NB-ARC domain (Table 2.1). Other abundant domains were leucine-rich repeat domains from several classes, including IPR032675 (155 predictions), IPR001611 (61), IPR003591 (20), IPR006553 (1), and IPR013210 (1). Seven proteins contained predicted AAA+ ATP hydrolase *cis*-domains, which overlapped predicted NB-ARC domains, however AAA+ e-values were low (0.029 to 4.4e-6), and may be spurious due to weak similarity between the AAA+ domain and the NB-ARC (Martin & Lupas 2013). An additional 38 non-CC, non-NB, or non-LRR *cis*-domains were found across 24 proteins, which may be additional components of disease resistance (Table 2.1). A summary of the CC, TIR, NB, and LRR organizations of the NB-ARC associated proteins is provided in Table 2.2.

Partial NB-ARC sequences detected by the HMM analysis

Seven HMM matches had lengths less than 100 bp and e-values between 6.5e-01 and 4.7e-10 yet were associated with predicted proteins (Table 2.3). Inspection of these proteins revealed one partial NB-ARC domain detected by Pfam (Bv.nbarc.0098), with a length of 70 amino acids and e-value of $1.6e^{-8}$. Other proteins contained LRRs with unusual domains; VQ (a short valine and glutamine-containing motif of unknown function), RNA-recognition motif, winged helix-turn-helix DNA binding domain, and tetra- or pentatricopeptide repeats (Table 2.3). Four HMM matches were grouped in Sf. 8, one with Sf. 22, while two had no Subfamily association. Three full-length NB-ARC HMM matches were also part of Sf. 8, located on Chromosomes 3, 4, and 7 (Figure 2.2). A BLASTx search using these domains identified three homologous proteins in RefBeet 1.2.2 predicted proteins, and these also contained CC, LRR,

VQ, and B3 DNA-binding *cis*-domains, but not NB-ARC (Table S2.3). The HMM which was designed to detect NB-ARC domains also detected a short homologous sequence that appeared to be present in predicted LRR-containing proteins and perhaps could represent diverged NB-ARC domains.

Cross-validation of R gene domain organization

Each of the 231 nucleotide domain sequences was queried against the RefBeet protein set to cross-validate class assignments (Table S2.1). Domain organizations from the two protein sets were in agreement for 118 loci. Disagreements were resolved using the union of the both sets, resulting in 204 NB-ARC-containing protein models in EL10 and RefBeet (Table 2.2). A total of 180 EL10 NB-ARC loci had similarity to 148 RefBeet proteins, using a minimum 90% identity threshold. Twenty-one NB-ARC HMMs matched EL10 genome without a *de novo* predicted proteins, compared to 48 matches in the RefBeet 1.2.2 protein set. When data were combined, 12 NB-ARC HMM predictions were found without a predicted protein observed in either genome.

TIR detected in *B. vulgaris*

InterProScan identified five proteins with TIR domains (IPR000157) on five chromosomes (Table 2.4). Pfam identified two of these domains as TIR (PF01582) and two others as TIR_2 (PF13676), and the fifth TIR domain detected by InterProScan module SUPERFAMILY was not detected by Pfam. One predicted protein, EL10Ac4g07495, contained an NB-ARC domain that was not identified by the nucleic acid HMM. This NB-ARC domain was incomplete, matching 136 amino acids of the Pfam model with an e-value of 3.1e-6, suggesting divergence compared to the rest of the NLR family in beets. The protein EL10Ac6g13736 overlapping Bv.nbarc.0184 contained TIR, NB-ARC, and Leucine-rich repeat

domains (Table 2.4). To our knowledge, protein EL10Ac6g13736 is the first indication of TIR-NLR presence in *B. vulgaris* (Tian et al. 2004).

Orthology

Each of the 231 NB-ARC HMM matches were used as a BLASTx query against the curated UniProt Swiss-Prot database to investigate the relationship between *B. vulgaris* NB-ARC domains and those from other species (Table S2.4). The single best-scoring match (by e-value) was retained as a classifier of the domains phylogenetic Subfamilies. Fifteen sequences had no matches, using an e-value cutoff of 1e-5. The best-scoring matches were placed in the *B. vulgaris* phylogeny using RAxML EPA as before (Figure 2.6). Phylogenetic Sf. 1 through 7 were predominantly associated with 16 different proteins annotated from *A. thaliana*, Sf. 8 through 17 were associated with a mix of annotated *A. thaliana* and *Solanum bulbocastanum* proteins, while Sf. 18 through 27 were associated with *Solanum bulbocastanum*. The closest homolog of Bv.nbarc.0184 was *Nicotiana glutinosa* resistance protein *N*, a TIR-NB-LRR (TNL) class resistance gene, further supporting the assignment of Bv.nbarc.0184 to the TNL class of resistance loci.

Characterization of NB-ARC surrounding *Rz2*

The *Rz2* gene sequence (Capistrano-Gossmann et al. 2017) overlapped Bv.nbarc.0121 at 9.3 MB on Chromosome 3 (Figure 2.4). Previous estimates of the genetic distance between *Rz1* and *Rz2* resistance loci range from 5 to 20 centimorgans (Barzen et al. 1997; Stevanato et al. 2015). The 20 MB surrounding the *Rz2* gene contained 25 other NB-ARC-like sequences (Figure 2.4). Twenty-one of these were assigned to seven phylogenetic Subfamilies, and four additional *trans*-domains lacked a Subfamily association. Twenty-four loci were contained within eight physical clusters (i.e. domains located within 200 kb of each other), while two of the loci

(including *Rz2*) were not clustered (i.e. singletons). *Rz1* likely resides within this 20 MB interval, however no obvious candidate was detected.

Description of the *Rz2* locus and phylogenetic subfamily

The *Rz2* gene in EL10 showed approximately 8 kb of Gypsy and Helitron elements inserted, as reported by Capistrano-Gossmann et al., (2017). The Bv.nbarc.0121 domain associated with *Rz2* was part of Sf. 26, which had 17 members unevenly distributed across five chromosomes. Sf. 26 was comprised of two large clades, two small clades of two domains each, and one basal domain (Figure 2.1). The basal member of its clade, Bv.nbarc.0194 domain, appeared to be a singleton on Chromosome 7, and the six remaining domains associated with the Bv.nbarc.0194 subclade formed a homogeneous physical cluster on Chromosome 9 (Figure 2.2). The other large clade within Sf. 26 was similarly dispersed, with the Bv.nbarc.0047/0048 domain on Chromosome 2 and the other four members of Sf.26 located within 2 MB of each other on Chromosome 3 (Figure 2.2).

DISCUSSION

NLR genes in plants exist as large families with capacity for diversification and differentiation. Diversification may be the result of co-evolution of plants and pathogens adapting to molecular changes in their adversaries. The diversity that makes disease resistance possible also confounds understanding the underlying genetics. Here we have taken steps towards identifying NLR resistance gene analogues in the EL10 genome using a novel nucleotide-based model search, and clarifying NLR gene family number and context in the agronomically important *Rz* resistance region.

NB domains of NLR genes are highly conserved across plant and animal species (Sarris et al. 2016; Seo et al. 2016; Urbach & Ausubel 2017). Identifying these domains through

homology-based approaches is an established method of NLR gene detection (Meyers et al. 2003; Monosi et al. 2004; Zhou et al. 2004). We extended NB-ARC domain detection beyond predicted protein-coding genes by using a nucleic acid based search model to search for more distantly-related and perhaps ancestral sequences. Adding search strategies based on homology to known proteins at the nucleic acid level might reveal diverged or inactivated sequences that are not recognized by current gene prediction algorithms. An expanded search was accomplished using a nucleic acid-based HMM applied directly to genomic sequence, which identified 48 putative NB-ARC signatures not present in the predicted protein set.

The 5' ends of NLR resistance genes are often thought to be associated with signaling, aggregation, or other processes downstream of pathogen perception. Two common domains at this position are CC and TIR, with numerous species containing both CC and TIR resistance genes (Arabidopsis, various Solanaceae, lettuce), while some species had been thought to contain only CC genes (e.g. monocots) (Christie et al. 2016; Christopoulou et al. 2015; Van Ghelder & Esmenjaud 2016; Meyers et al. 2003; Monosi et al. 2004; Pan et al. 2000; Shao et al. 2014; Zhou et al. 2004). This paradigm was revised due to the discovery of a second class of TIR domain persisting as a small family in all flowering plants, called TIR_2 (Nandety et al. 2013; Sarris et al. 2016). Until now, it was thought that *B. vulgaris* did not contain any TIR-type NLRs (Tian et al. 2004). In the current study, TIR, TIR-NB, and TIR-NB-LRR proteins were detected in EL10. Both TIR (PF01582) and TIR_2 (PF13676) domains were found, extending the evolutionary placement of these domains to the *Caryophyllales*. Further genomic analysis of diverse *B. vulgaris* accessions should help resolve the prevalence and diversity of TNL resistance gene analogues in sugar beet and advance efforts to identify additional molecular sources of disease resistance in this species.

There were 38 non-CC-TIR-NB-LRR domain types detected in NLR proteins from EL10 (Table 2.1). Kroj *et al.* (2016) identified 94 non-CC/TIR/NB-ARC/LRR domains across 31 genomes as hypothetical integrated decoys (Cesari et al. 2014; van der Hoorn & Kamoun 2008). Three of the 94 integrated domains were present in more than one species, one of which was also found in an EL10 NLR (kinase IPR011009). The *Hs1pro-1* gene, a historically important sequence for plant nematode resistance, contained N-terminal leucine-rich repeats, a transmembrane domain, but no CC, TIR, or NB-ARC (Cai et al. 1997). Similarly, our analysis identified putative proteins interspersed in a large cluster of NLRs on Chromosome 3 containing LRR domains but not full NB-ARCs. Possible defense-related domains in the EL10 NB-ARC-associated proteins included Calmodulin-binding-protein 60 (CBP60, found in positive regulators of plant immunity), various second messenger system components such as phosphatases and kinases, an RNAse H (IPR012337), and a poorly-characterized phloem protein-2-like domain (IPR025886) that was suggested to function in defense against phloem sucking insects (Dinant et al. 2003; Kehr 2006). Regulators of chromatin condensation domains in EL10 NLRs (IPR000408 and IPR009091) could also be involved in defense responses (Latrasse et al. 2017). The single EL10 NLR-associated concavalin-A-like lectin/glucanase domain could be associated with decoy function related to perception of pathogen carbohydrate signatures (Rüdiger & Gabius 2002).

The disease rhizomania is a major concern for sugar beet production worldwide (Biancardi & Tamada 2016). Two sources of resistance have been commercially deployed, both derived from the sea beet, *B. vulgaris* spp. *maritima*, a wild relative of cultivated *B. vulgaris* (Biancardi et al. 2002; Lewellen 1991). Two loci on Chromosome 3 confer resistance, *Rz1* and *Rz2*, and additional loci have been described (Grimmer et al. 2008; Litwiniec et al. 2015;

Stevanato et al. 2015). Generation of molecular markers across the *Rz* region has been ongoing for over two decades (Barzen et al. 1992, 1997; Francis & Luterbacher 2003; Gidner et al. 2005; Grimmer et al. 2008; Stevanato & Trebbi 2012; Stevanato et al. 2015; Scholten et al. 1999). Long after being reported as a separate source of resistance, a specific CC-NB-LRR gene associated with *Rz2* was identified (Capistrano-Gossmann et al. 2017). The molecular basis of *Rz1* resistance is still poorly understood (Biancardi and Tamada, 2016; Lewellen, 1991).

Seven phylogenetic Subfamilies of NB-ARCs have members on the arm of Chromosome 3 associated with rhizomania resistance (Figure 2.4). Multiple tandem duplication events appear to have generated three distinct interspersed Subfamilies in this 20 MB region. For instance, Sf. 26 had separate clusters at 11 and 22 MB, one of which included the *Rz2*-associated domain Bv.nbarc.0121 (Figure 2.4). These two clusters formed well-supported independent subclades in the phylogenetic analysis of the *Rz* region, supporting the hypothesis that these domains arose from localized tandem duplications rather than large-scale segmental duplication. Another clade within Sf. 26 appears to be the result of a progenitor sequence, represented by Bv.nbarc.0194, being translocated from Chromosome 7 to Chromosome 9 and subsequently duplicating into six new copies (Figures 1, 2, 3). The chromosome context of disease resistance genes could be a key component of local duplications and could play a critical role in the generation of new disease resistance (Bornemann & Varrelmann 2013; Bornemann et al. 2015; Liu et al. 2005; Pferdmenges et al. 2009).

The design of novel resistance genes has been demonstrated with some success in *Arabidopsis*, where a target of NLR surveillance was modified to interact with a different pathogen, causing new resistance specificity (Kim et al. 2016). NLRs transferred between closely-related species, as well as between monocots and dicots, have successfully conferred

disease resistance (Narusaka et al. 2013; Jacob et al. 2013). New strategies are emerging capable of reducing fitness costs of engineered resistance (Xu et al. 2017). The ability to design resistance genes and edit them into plant genomes is quickly becoming reality, and will be facilitated by contiguous genome assemblies able to resolve duplicated sequences and repetitive elements such as the present study.

Methods to identify functional resistance genes in diverse accessions of crops and their wild relatives are of continuing importance to agricultural sustainability. Nucleotide-based HMMs provide a tool for screening new genome assemblies without the need for computational prediction of transcripts or proteins. The NB-ARC HMM generated here was able to detect fragmented NB-ARC sequences, which is an important feature given that *de novo* assemblies are often composed of contigs of gene- or exon-sized fragments. A single reference genome does not capture the sequence diversity of a species (Hirsch et al. 2014; Horton et al. 2012; Lam et al. 2010; Schnable et al. 2012; Zhang et al. 2012; Hardigan et al. 2015). Fragmented NLR loci detected in one genome could be the site of full-length resistance alleles in other accessions. This is the case for the *Rz2* locus, which is interrupted by an ~8 kb transposable element in some accessions (Capistrano-Gossmann et al. 2017), making detection by homology-based methods more difficult. HMM searches could help generate molecular markers able to track resistance gene variation in populations and across gene pools, with application to a wide variety of diseases.

METHODS

A five-generation inbred of the sugar beet genome named ‘EL10’ was assembled using several approaches (McGrath et al., in preparation). Briefly, one inbred plant was chosen for

Illumina sequencing, optical mapping (Tang et al. 2015), PacBio RSII sequencing (using P6-C4 chemistry), and Hi-C scaffolding (van Berkum et al. 2010; Burton et al. 2013), and was largely able to reduce the number of scaffolds to the same number of chromosomes in beet (e.g., n = x = 9). Scaffolds were polished and gap-filled using a combination of approaches (PBJelly, Arrow, Pilon following Bickhart et al., 2017), and EL10 scaffolds showed high concordance with genetic maps and the RefBeet genome sequence (Dohm et al. 2014). This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession PCNB00000000. The version used in this paper is version PCNB01000000 (https://www.ncbi.nlm.nih.gov/assembly/GCA_002917755.1). The full genome sequence, gene annotations, and predicted NB-ARC domains are available for download and to browse at the Comparative Genomics portal (https://genomevolution.org/coge/GenomeInfo.pl?gid=37197).

Building the *Beta vulgaris* NB-ARC model

Predicted protein sets were generated with the MAKER pipeline (Law et al. 2015) and analyzed with InterProScan v. 5.19-57.0 (Philip Jones et al. 2014) for NB-ARC (IPR002182) domains of 200 amino acids or more. The nucleic acid sequences underlying the NB-ARC protein domains were extracted, and a consensus alignment was derived from the nucleic-acid-based NB-ARC domains using MAFFT v 7.215 with the --leavegappyregion option (Katoh & Standley 2013). The nucleotide alignment was converted into an HMM using the hmmbuild function of HMMER v. 3.1b2 (hmmer.org) (File S2.1), and the resulting nucleic acid HMM was used to query the EL10 genomic sequence using the nhmmer function of HMMER v. 3.1b2.

Phylogenetic analysis

Preliminary phylogenetic analysis was restricted to genomic NB-ARC sequences with e-values lower than 1e-10. A consensus alignment was created as described, and a phylogenetic

tree was constructed using RAxML v. 8.0.6 (Stamatakis 2014). The '-f a' function was used to conduct a rapid Bootstrap analysis and search for the best-scoring maximum likelihood tree, in a single run using 1,000 bootstrap replicates. The model of substitution was GTRGAMMA. This process was repeated five times using different random number seeds, retaining the highest scoring tree. Placement of incomplete NB-ARC sequences onto the full tree was done with the evolutionary placement algorithm (EPA) of RAxML (Berger et al. 2011) using the '-f v' command. Bootstrap values from the initial tree were combined with the EPA output using the labelled_tree function of Genesis software v. 0.19.0 (Czech & Stamatakis 2017). Trees were visualized with FigTree v1.4.2.

Association of NB-ARC sequences with other domains

Custom python scripts (available from the authors) were used to assess overlap between predicted NB-ARC domains and preliminary gene predictions in the EL10 genomic sequence. The preliminary gene predictions were analyzed with InterProScan v. 5.19-57.0, which incorporates a suite of feature detection applications including SUPERFAMILY, PANTHER, Gene3D, Hamap, Coils, ProSiteProfiles, ProSitePatterns, TIGRFAM, PRINTS, Pfam, and ProDom (Philip Jones et al. 2014). NB-ARC loci predicted from EL10 were used to extract cognate gene models, and these models were classified according to their combinations of CC, TIR, NB-ARC, and LRR domains determined by the InterProScan analysis. The HMM-based NB-ARC sequences were also used as translated protein queries (BLASTx search v.2.2.3) against the predicted proteome of RefBeet v1.2.2 (Dohm et al., 2014). Predicted proteins with at least 90% sequence identity were used to determine homology between EL10 and RefBeet sequences. RefBeet protein domains were identified using the same InterProScan method used for EL10. EL10 and RefBeet annotations were merged to create a consensus assignment of

domains for each putative NB-ARC locus. Identification of domains other than CC, TIR, NB-ARC, and LRR was based solely on the predicted protein set in the EL10 genome.

ACKNOWLEDGEMENTS

We thank Kevin Childs, Jie Wang, Jane Pulman, and Tiffany Liu for assistance with genome annotation. Ashley Weizcorek provided stimulating discussion and critique of the project. Pat Edger contributed valuable perspective on phylogenetics.

APPENDICES

APPENDIX A

Main Tables and Figures

Table 2.1. Domains identified in predicted NB-ARC-associated transcripts in *B. vulgaris*. Proteins overlapping the HMM-based NB-ARC domains are shown in the first column as "HMM-associated". Proteins with NB-ARC domains in the predicted protein set are shown as "predicted proteins only". Count is the number of proteins with one or more of the listed domains.

IPR code	Name	HMM-associated proteins	Predicted proteins only
IPR002182	NB-ARC	183	185
IPR027417	P-loop_NTPase	180	179
IPR032675	Leucine-rich repeat domain	155	146
IPR001611	Leu-rich_rpt	61	59
IPR003591	Leu-rich_rpt_typical-subtyp	20	19
IPR003593	AAA+_ATPase	7	8
IPR000504	RRM_dom	3	1
IPR012677	RNA recognition motif domain	3	1
IPR029058	AB_hydrolase	3	3
IPR002885	Pentatricopeptide repeat	2	0
IPR004910	Yippee/Mis18/Cereblon	2	0
IPR000073	AB_hydrolase_1	1	1
IPR000157	TIR_dom	1	2
IPR000232	HSF_DNA-bd	1	0
IPR000408	Reg_chr_condens	1	1
IPR000719	Prot_kinase_dom	1	1
IPR000760	Inositol_monophosphatase-like	1	0
IPR001878	Znf_CCHC	1	1
IPR004696	Tpt_PEP_transl	1	0
IPR004853	Sugar_P_trans_dom	1	0
IPR005814	Aminotrans_3	1	1
IPR006050	DNA_photolyase_N	1	1
IPR006239	Bisphos_HAL2	1	0
IPR006553	Leu-rich_rpt_Cys-con_subtyp	1	1
IPR008271	Ser/Thr_kinase_AS	1	1
IPR008808	Powdery_mildew-R_dom	1	1
IPR008889	VQ	1	0
IPR009091	RCC1/BLIP-II	1	1
IPR011009	Kinase-like_dom	1	1
IPR011990	TPR-like_helical_dom	1	0
IPR011991	WHTH_DNA-bd_dom	1	0
IPR012337	RNaseH-like_dom	1	1
IPR012416	CBP60	1	1
IPR013210	LRR_N_plant-typ	1	1
IPR013320	ConA-like_dom	1	1
IPR014729	Rossmann-like_a/b/a_fold	1	1
IPR015422	PyrdxlP-dep_Trfase_sub2	1	1
IPR015424	PyrdxlP-dep_Trfase	1	1
IPR018000	Neurotransmitter_ion_chnl_CS	1	1
IPR020550	Inositol_monophosphatase_CS	1	0
IPR020583	Inositol_monoP_metal-BS	1	0
IPR022739	Polyphenol_oxidase_cen	1	1
IPR023566	PPIase_FKBP	1	0
IPR025886	PP2-like	1	1
IPR027725	HSF_fam	1	0
IPR029472	UBN2_3	1	0

Table 2.2. NLR domain organizations in the *B. vulgaris* EL10 genome based on *de novo* transcript prediction and 90% identity with publicly available transcripts. The combined number was derived from the union of the domain predictions.

Domain organization	Composition using only EL10 transcripts	Composition using only RefBeet blastx	Combined number
CNL	76	72	97
CN	10	21	16
COIL	3	0	0
TNL	1	1	1
TN	0	0	0
TIR	0	0	0
NLR	69	54	64
NB-ARC	30	26	26
CL	3	5	5
LRR	4	4	6
Protein missing CC/TIR/NB/LRR	14	0	4
No predicted protein	21	48	12

Table 2.3. Partial NB-ARC sequences detected by the HMM in *B. vulgaris*.

Domain	Chromosome	HMM start	HMM stop	HMM Phylogenetic subfamily	Predicted Protein Domains
Bv.nbarc.0069	Chr3	13724032	13724086	22	HSF_DNA-bd, WHTH_DNA-bd_dom, HSF_fam
Bv.nbarc.0098	Chr3	51533578	51533673	0	NB-ARC, Leucine-rich Leu-rich_rpt, VQ
Bv.nbarc.0118	Chr3	51975010	51975105	0	Leu-rich_rpt, Leu-rich_rpt_typical-subtyp, Leucine-rich
Bv.nbarc.0101	Chr3	51556392	51556488	8	RRM_dom, Pentatricopeptide_repeat, Nucleotide-bd_a/b_plait_sf Leucine-rich
Bv.nbarc.0104	Chr3	51652191	51652287	8	-
Bv.nbarc.0109	Chr3	51713080	51713176	8	Leucine-rich
Bv.nbarc.0116	Chr3	51925255	51925351	8	RRM_dom, Pentatricopeptide_repeat, Nucleotide-bd_a/b_plait_sf Leu-rich_rpt, Leucine-rich

Table 2.4. Domain composition of predicted TIR-containing proteins in *B. vulgaris*.

Protein ID	IPR Code	InterPro Domain Name	Chromosome	Genomic Position
EL10Ac1g01449	IPR000157	TIR_dom	1	22,858,717
EL10Ac4g07495	IPR000157	TIR_2_dom	4	98,329
	IPR002182	NB-ARC		
	IPR003593	AAA+_ATPase		
	IPR027417	P-loop_NTPase		
EL10Ac5g12796	IPR000157	TIR_dom_unknown	5	55,158,173
	IPR003593	AAA+_ATPase		
	IPR027417	P-loop_NTPase		
EL10Ac6g13736	IPR000157	TIR_dom	6	9,786,798
	IPR001611	Leu-rich_rpt		
	IPR002182	NB-ARC		
	IPR003591	Leu-rich_rpt_typical-subtyp		
	IPR003593	AAA+_ATPase		
	IPR027417	P-loop_NTPase		
	IPR032675	Leucine-rich repeat domain		
EL10Ac8g20309	IPR000157	TIR_2_dom	8	52,781,311

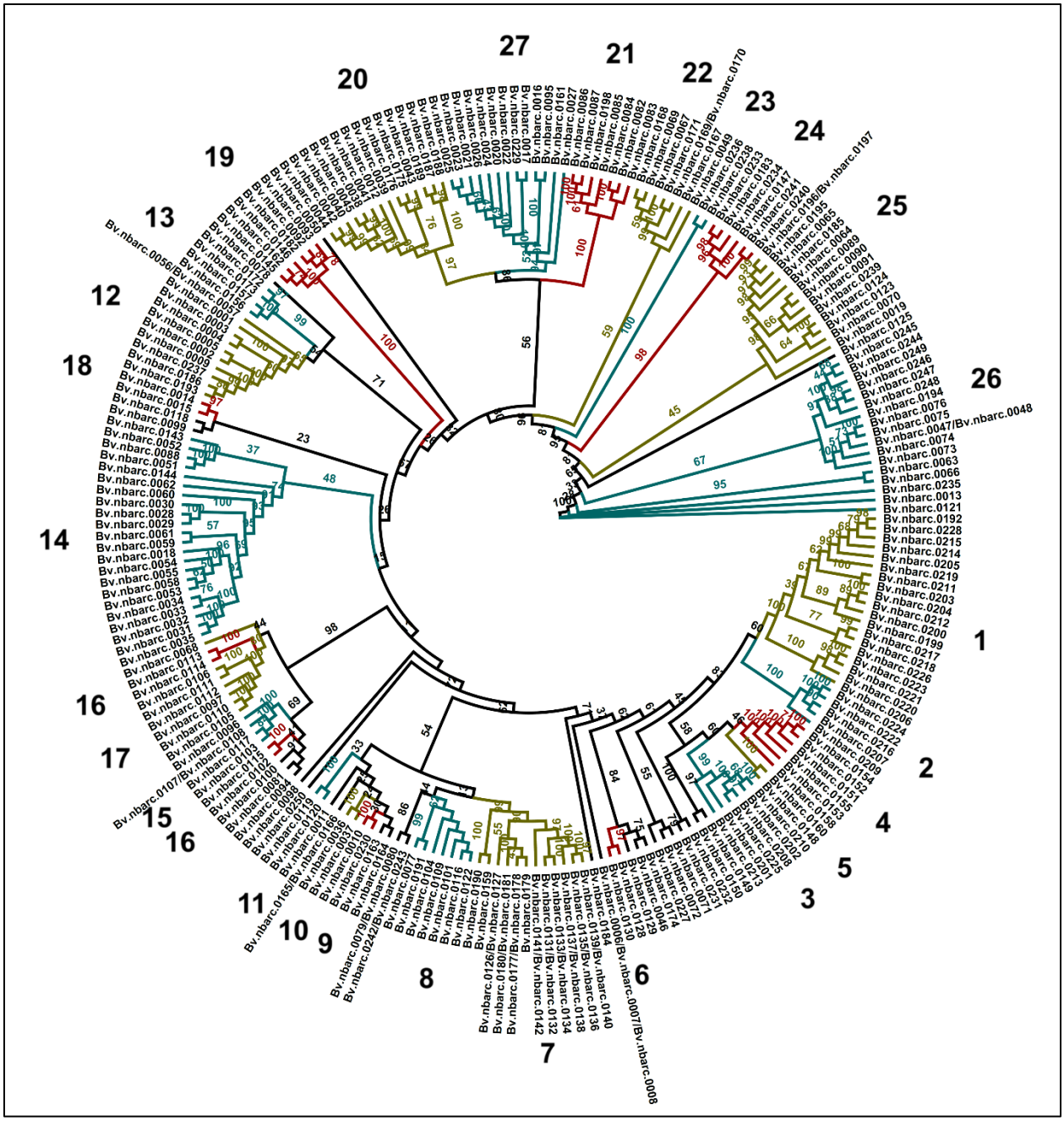

Figure 2.1. Phylogenetic relationships of putative NB-ARC domains in the EL10 genome of *B. vulgaris*. Clades with at least 60% pairwise identity are numbered and colored to distinguish between nearby Subfamilies. Bootstrap values of 1,000 replications are shown at branch points.

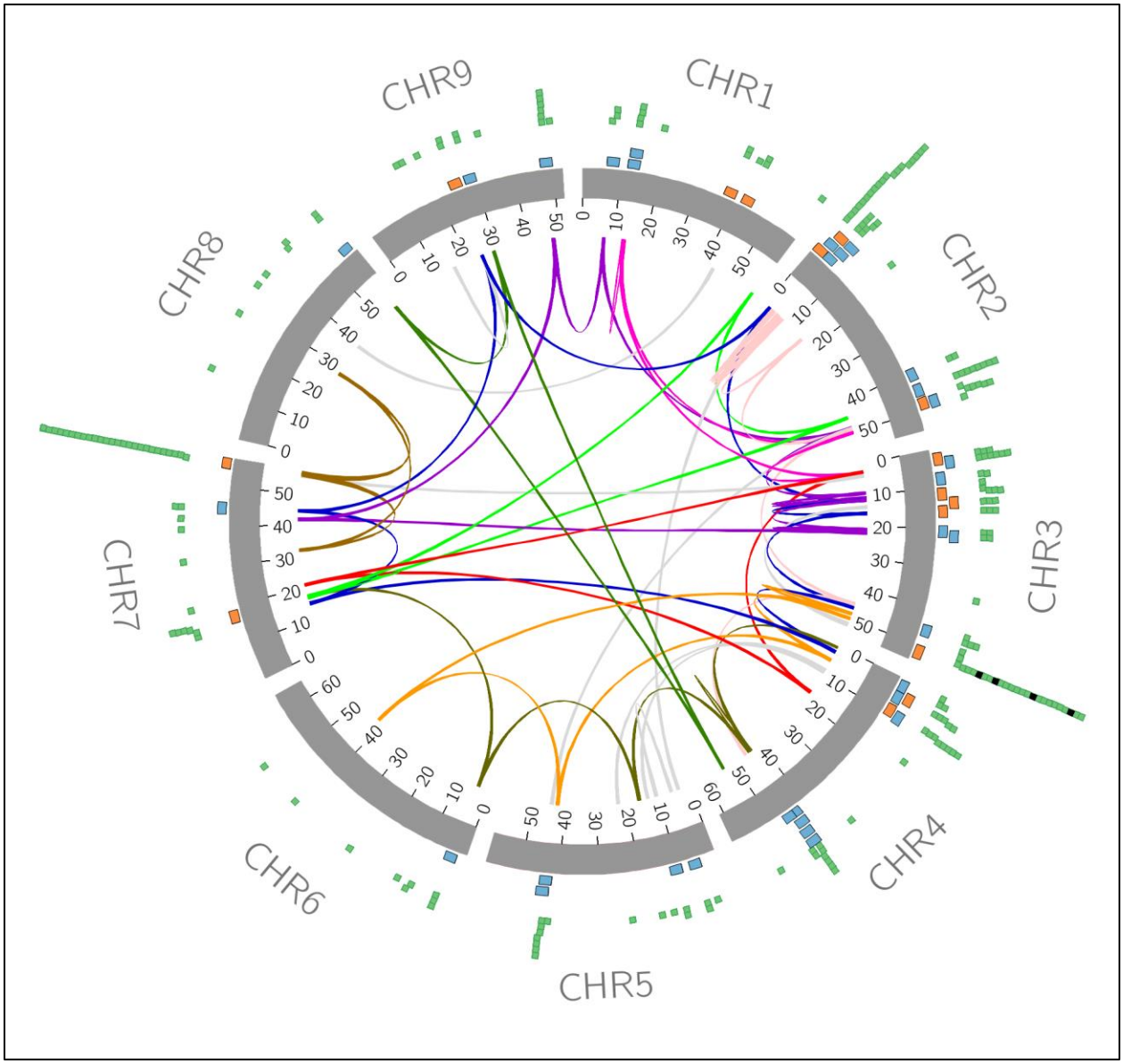

Figure 2.2. Location of NB-ARC domains and their phylogenetic associations in the *B. vulgaris* EL10 genome. The symbols depict chromosomes (grey rectangles), NB-ARC domains (green squares), winged helix-turn-helix domains (black squares), and physical clusters from the same Subfamily (blue rectangles) or different Subfamilies (orange rectangles). The phylogenetic relationships of NB-ARC domains are depicted as colored links, with each color representing a different Subfamily. Subfamilies with only two locations are shown with gray lines.

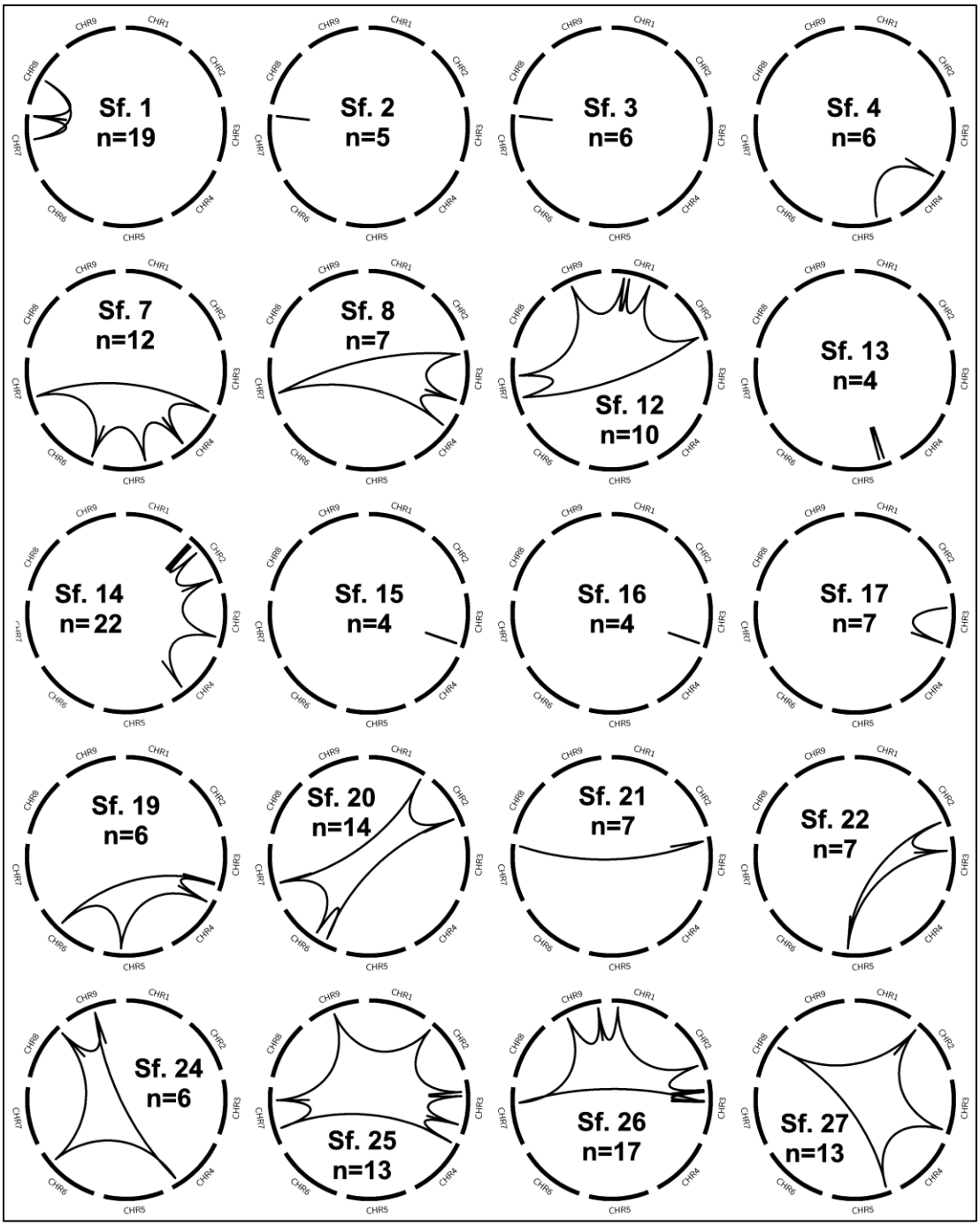

Figure 2.3. Physical distribution of phylogenetic Subfamilies (Sf.) across the *B. vulgaris* EL10 genome. Chromosomes are ordered and labeled as in Figure 2. Subfamilies with at least four members are shown in their physical locations in the EL10 genome. Lines are drawn linking adjacent family members within each plot. Multiple clustered domains form a straight line along the radial axis, e.g. Sf. 2, 3, and the portion of Sf. 4 on chromosome 4. The number of Subfamily members is denoted within each plot.

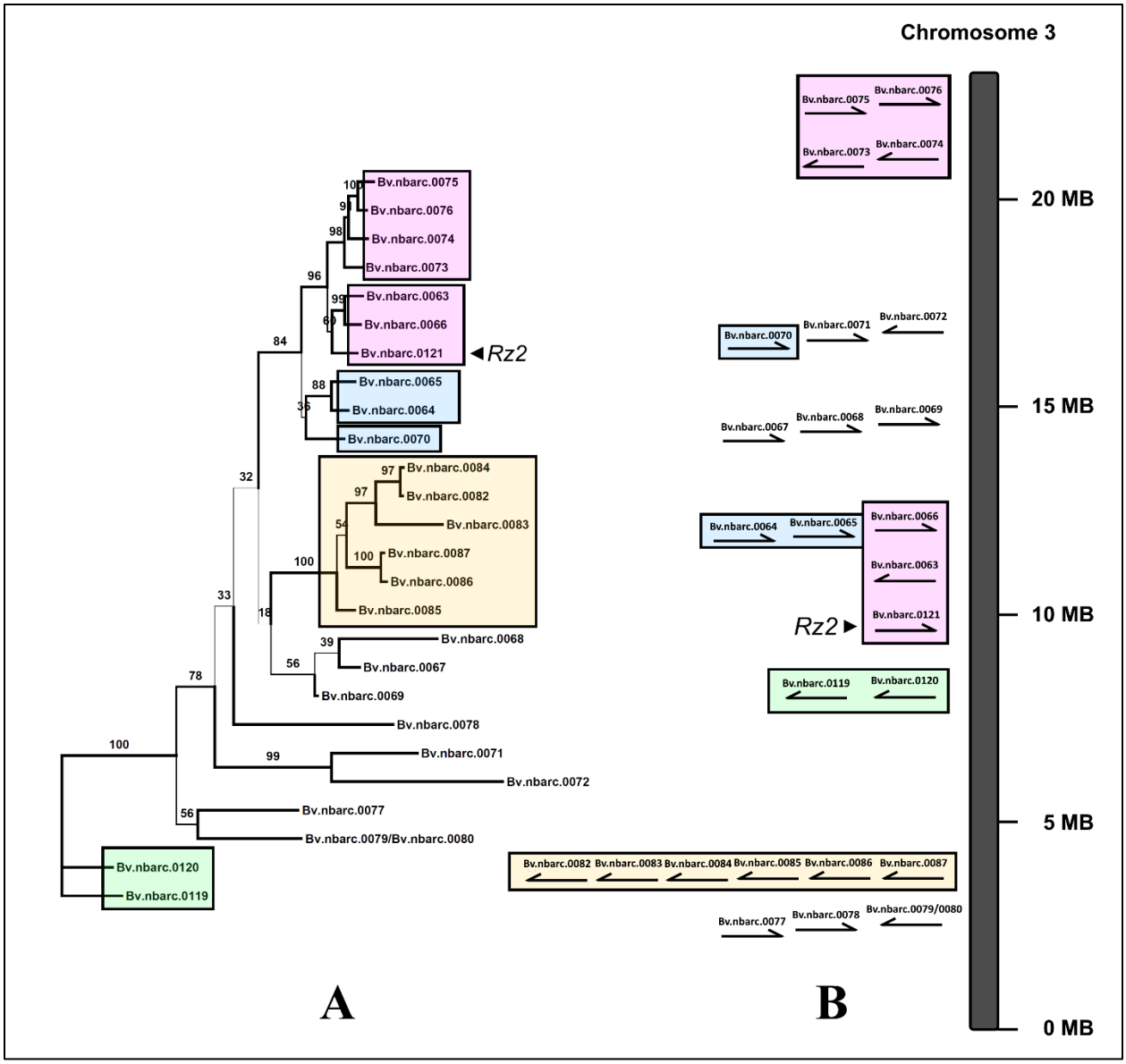

Figure 2.4. Phylogenetic relationship and physical location of putative NB-ARC domains in the *Rz* region of Chromosome 3 from *B. vulgaris*. A) Phylogeny of the 26 domains between 0 MB and 25 MB on Chromosome 3. Colored boxes represent four different Subfamily assignments based on the whole genome phylogeny (purple = Sf. 26, green = Sf. 11, blue = Sf. 25, yellow = Sf. 21). B) Physical locations of domains and their orientation. Colored boxes correspond to the Subfamily identity of the domains from panel A. Domains without boxes do not have a Subfamily member in the region and/or were not part of any Subfamily. The location of the *Rz2* gene is shown overlapping Bv.nbarc.0121.

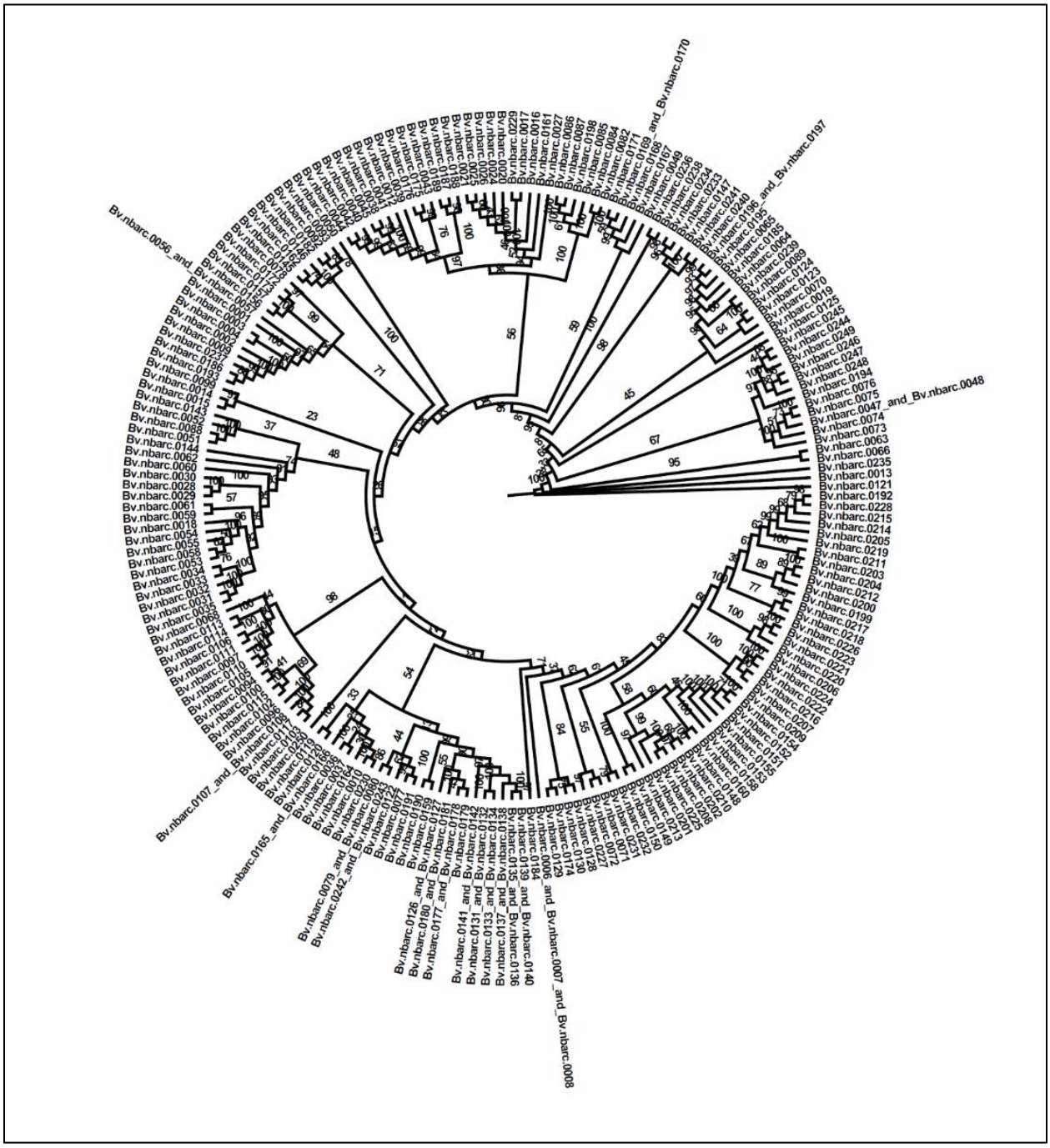

Figure 2.5. Initial tree of 210 *B. vulgaris* NB-ARC domains with initial lengths greater than 400 bp and e-values less than 1e-10.

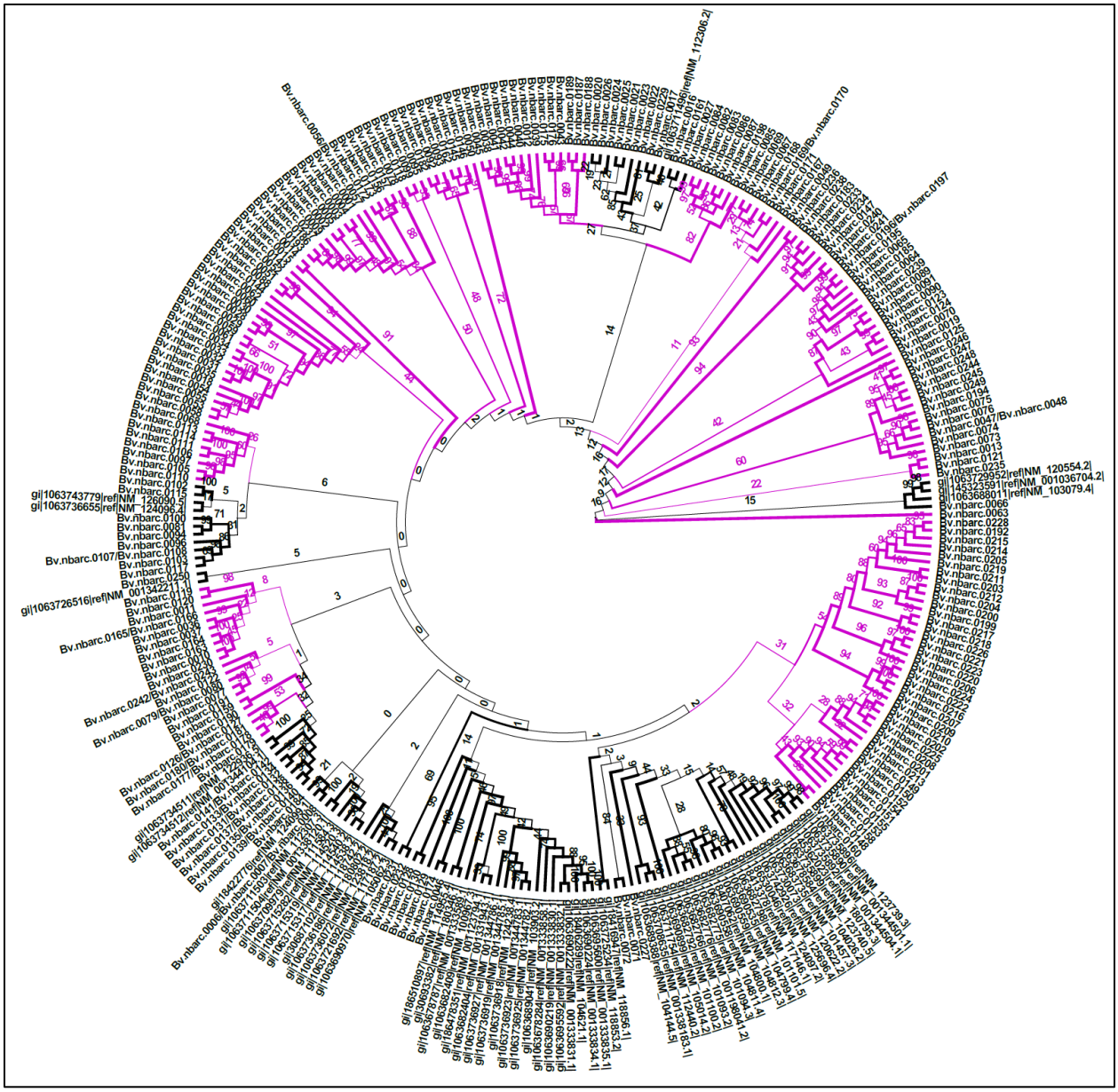

Figure 2.6. SwissProt homolog NB-ARC domains placed onto the *B. vulgaris* NB-ARC phylogeny. Clades without SwissProt domains are highlighted in magenta. Line thickness is proportional to the bootstrap support value of that branch.

APPENDIX B

Supplementary Tables, Figures, and Data

All supplementary tables and files available at

https://onlinelibrary.wiley.com/doi/full/10.1111/tpj.13977

Table S2.1. Putative NB-ARC domains and predicted transcripts in the EL10 genome. Phylogenetic Subfamily assignment was based on minimum 60% pairwise identity with other sequences in the clade.

Table S2.2. Domains concatenated into longer NB-ARC predictions.

Table S2.3. RefBeet proteins homologous to EL10 partial NB-ARC domains.

Table S2.4. Predicted NB-ARC domains with orthologous protein relationships.

File S2.1. Annotation file for EL10 predicted proteins overlapping HMM-derived genomic NB-ARC, from InterProScan

CHAPTER THREE

DETECTING SIGNATURES OF DISEASE RESISTANCE GENES IN WHOLE-POPULATION *DE NOVO* GENOME ASSEMBLIES

INTRODUCTION

Improved disease resistance is a perennial target for plant breeders. Tremendous effort is put into screening germplasm and wild accessions in the hope of identifying novel genetic variation to counteract pressure from pathogens (Capistrano-Gossmann et al. 2017; Young 1996; Poland & Rutkoski 2016). New sources of resistance are incorporated into elite germplasm and deployed in the field, which benefits growers by increasing yield and/or reducing the need for costly inputs (Boyd et al. 2013). Unfortunately, successful disease resistance often acts as a source of selection pressure, leading to the proliferation of pathogens able to defeat plant genetic defenses (reviewed in Brown 2015). For example, the spread of resistance-breaking pathogens has been observed in sugar beet, as strains of *Beet necrotic yellow vein virus* (BNYVV) begin to overcome genetic resistance conferred by *Rz* resistance loci (Liu & Lewellen 2007; Bornemann et al. 2015; Broccanello et al. 2018). Dynamic pathogen adaptation and spread will only increase in the future, as global connectivity and climate change further disrupt agriculture world-wide (Cilas et al. 2016).

Existing breeding populations balance genetic diversity for environmental adaptation and disease resistance against a core set of agronomic traits necessary for crop production (Tester & Langridge 2010). This allows breeders to target ongoing challenges while maintaining a baseline set of crop characteristics. One concern when breeding for disease resistance is whether resistance found in separate populations has the same genetic basis, or whether combining the two resistances could provide synergistic protection from pathogens (J. D. G. Jones et al. 2014).

These questions can be addressed with classic breeding techniques, but each new generation of plants requires expenditure of time and resources. Furthermore, genetic mapping strategies could struggle to differentiate one locus with multiple alleles versus two or more closely linked genes (Lipka et al. 2015). Defining members of common resistance gene families could provide information to help design crosses and breeding experiments, subsequently speeding the rate of crop improvement.

The two most prevalent disease resistance proteins are nucleotide-binding leucine-rich-repeat (NLR) and receptor-like kinase (RLK) (Kourelis & Van Der Hoorn 2018). Numerous reports have detailed the organization of NLR families in reference genome assemblies such as Arabidopsis, *Brassica rapa*, tomato, potato, pepper, lettuce, papaya, poplar, and peach (Jupe et al. 2012; Seo et al. 2016; Christopoulou et al. 2015; Andolfo et al. 2014; Van Ghelder & Esmenjaud 2016; Porter et al. 2009; Mun et al. 2009; Kohler et al. 2008; Meyers et al. 2003). In contrast, the receptor-like kinase superfamily encompasses diverse structures and functions compared to NLRs, precluding straightforward genome-wide characterization (Shiu & Bleecker 2003). The serine/threonine (Ser/Thr) subfamily of kinases is known to be involved in defense responses, providing guidance for parsing the diverse kinase superfamily of genes (Zhou et al. 1995; Gómez-Gómez & Boller 2000; Song et al. 1995).

NLRs and RLKs both contain domains critical for their function: the NB-ARC domain for NLRs and the kinase domain for RLKs (van der Biezen & Jones 1998; Shiu & Bleecker 2001). NLRs have been identified by parsing predicted proteins for R gene domains (Steuernagel et al. 2015), exome capture in the form of resistance gene enrichment (Jupe et al. 2013), and NB-ARC modeling from genomic sequences (Funk et al. 2018). RLK detection has been based primarily on protein prediction from reference genomes (Vij et al. 2008; Shiu et al. 2004).

Current knowledge of NLRs and RLKs has provided insights regarding evolution of these gene families, but expanding current approaches for R gene detection could aid understanding and deployment of disease resistance in plants.

In this report, I sought to apply model-based protein domain detection to two new areas. First, I tested the viability of using hidden Markov models (HMMs) to identify Ser/Thr kinase domains in a reference-quality sequence of sugar beet. Second, I asked whether *de novo* assembly of whole-population pooled sequences of *B. vulgaris* could differentiate paralogous NB-ARC domains such that they could be parsed using the HMM-based modeling scheme of Funk *et al.* (2018). Despite challenges with intron/exon structure, I identified 598 Ser/Thr kinase domains in the EL10 genome, with 146 (23%) existing in physical clusters on chromosomes. I constructed *de novo* assemblies of 23 populations using pooled reads to capture population-wide diversity in a single sequencing library. I assessed the completeness of the assemblies by quantifying single-copy conserved orthologs, finding 78-94% (median 90%) of query genes in each population despite 8-30% (median 17%) fragmented genes. Assembled contigs were mapped to the EL10 reference, with average 1-to-1 mapping identity covering 351-425 MB of the 540 MB genome with >97% average identity. Searching assembled contigs with an NB-ARC-derived HMM identified between 114 and 142 full-length domains per population, in line with the numbers found in the EL10 reference genome. I constructed phylogenetic trees for two different NB-ARC loci and observed distinct subclades suggesting allelic or copy-number variation. These results indicate HMM-based strategies can be used to define kinase diversity in genomic samples, and *de novo* assembly of pooled population sequences is a viable strategy for detecting kinase and NB-ARC diversity in whole populations. Subsequent analysis should add to improved understanding of R gene diversity and function.

RESULTS

Kinase detection in the EL10 reference genome

My first goal was to use HMM modeling to identify kinase domains in the EL10 genome assembly. This was accomplished with a similar strategy used to identify NB-ARC domains previously (Chapter 2, Funk *et al*. 2018). The predicted InterPro domains for kinase superfamily (IPR011009) and Ser/Thr kinase (IPR000719) were aligned separately and used to generate distinct HMMs of 577 and 576 bp, respectively (Philip Jones et al. 2014). The EL10 genome was scanned using each model, resulting in 1477 superfamily matches and 1590 Ser/Thr matches with maximum domain e-values of 1 (Figure 3.1). The median lengths of HMM matches were less than 60% of the full model length (309 and 349 bp for the Ser/Thr and superfamily domains, respectively) indicating that many of the matches are incomplete kinase domains. A 400 bp threshold was implemented to avoid double-counting domains split by introns or indels, resulting in 884 kinase domains total (Table 3.1). This included 598 Ser/Thr kinase domains and an additional 286 superfamily domains without an overlapping Ser/Thr kinase. Domains were unevenly distributed among chromosomes, with an average of 64.7 and standard deviation of 12.5 (Table 3.1).

I next asked if the putative Ser/Thr kinase domains were physically clustered in the genome in a similar manner as NB-ARC domains. I defined a cluster as two domains separated by 200 kb or fewer (following Christopoulou *et al*. (2015)), which resulted in 15 clusters encompassing 24% (146/598) of predicted Ser/Thr kinase domains (Table 3.2). There were clusters on all nine chromosomes. Chromosomes 1 and 5 had the fewest clustered Ser/Thr kinase domains, in line with those chromosomes also having the fewest total domains. Conversely,

Chromosome 7 was above average for domain count but had the fewest domains in clusters. The largest number of total domains and the largest cluster was found on Chromosome 4.

De novo assembly of pooled whole-population sequences

Mapping short sequencing reads to a reference genome could miss novel variation present in the sample but absent in the reference. To address this concern, I examined *de novo* whole-genome assemblies of the pooled populations. Total assembly lengths for the 23 populations ranged from 398 to 573 Mb, with contig N50 between 3.7 and 12.9 kb (Table 3.3). The single plant sample C869_US and the inbred population W357B generated the largest N50s of 11.9 and 12.9 kb, respectively, suggesting that decreased heterozygosity improved *de novo* assembly. The length of the C869_US assembly was 436 MB compared to the EL10 reference length of 540 MB. Given that these two assemblies were derived from DNA isolated from the same plant, the reduced length of the C869_US assembly suggests ~20% of the reference genome is unable to be assembled from short reads.

If the *de novo* assemblies are coherent representations of genomes, the assembled contigs should map with some specificity to the EL10 reference genome. I aligned the *de novo* assembled contigs with lengths greater than 200 bp to the reference genome to determine exact 1-to-1 alignments and length of total coverage (Table 3.4). The average identity of mapped contigs ranged from 96.8 to 99.56%. The total length of reference bases covered by each assembly was between 351 and 420 MB. The C869_US sample represented the highest reference base coverage as well as highest percent identity. The percent identity of mapped contigs was partitioned by crop type: the C869 reference population and other sugar beet samples had the highest percent identity scores, while table beet, fodder beet, and chard (leaf beet) samples were

lower (Table 3.4). Length of reference coverage did not follow the same crop-type trend, with two chards and one fodder sample exceeding the average coverage length of sugar samples.

To ascertain how well the gene space of each population was assembled, the *de novo* assemblies were searched for the presence of conserved orthologous genes identified using Benchmarking Universal Single-copy Conserved Orthologs (BUSCO) (Simão et al. 2015). My first query used a set of 303 genes thought to be conserved in all eukaryotes (Waterhouse et al. 2018). My samples contained 55-83% of complete conserved orthologs, with an additional 5-20% of orthologs identified as gene fragments (Table 3.5). In comparison, the EL10 reference genome contained complete copies of 87% of the queried genes, with an additional 1% as fragments. An increase in the portion of fragmented genes was seen between the genetically related samples of the EL10 reference genome, C869_US single plant, and C86925 pooled population. To test whether the table beet populations had more complete assemblies due to their presumed inbreeding (Paul Galewski, personal communication), I compared the means of the table beet and sugar beet pooled populations complete + fragmented gene sets, but there was no significant difference (t-test, $p < 0.3$). Parsing the table of 303 conserved genes revealed that some orthologs were missing from all *B. vulgaris* samples including the EL10 reference genome (Figure 3.2). It is possible that some genes considered universally conserved reside in intractable portions of the beet genome, which were unassembled even using long-read technology. Alternately, some genes tagged as "universally conserved" could be authentically absent from *B. vulgaris*.

To further investigate gene representation in the *de novo* assemblies, I used a set of 1,375 genes thought to be universal and single-copy across all embryophyta (Waterhouse et al. 2018). The EL10 reference genome contained 96% (1318/1375) of the queried genes. The most

complete *de novo* assemblies were C869_US (single plant) and W357B at 84% and 83%, respectively (Table 3.6). This is in line with the results from the eukaryote gene set, reinforcing the observation that these samples had the most complete gene space assembly. The remaining *de novo* assemblies contained 45% to 76% of the embryophyta query as complete gene models. While the complete gene model count was relatively low, there was a concurrent increase in fragmented gene models. These fragmented genes could be due to poor assembly of introns leading to splitting of exons between multiple contigs. Adding fragmented genes to complete genes increased the range of detection to 77-94% of the 1,375 embryophyta genes, with an average of 89% and standard deviation of 3.8% (Table 3.6). I compared mean numbers of complete + fragmented BUSCO genes between table and sugar beet due to possible differences in crop type heterozygosity. I found that table beets had a slight but significant increase in total BUSCO genes detected, 91.1% vs 88.4%, respectively (t-test, $p<0.015$). These results suggest the assemblies captured the majority of the gene space, sufficient to detect individual protein domains such as NB-ARCs.

NB-ARC detection in *de novo* assemblies

My next goal was to assess NB-ARC variation in each population. I used the nucleotide-based NB-ARC strategy described previously to scan the *de novo* assemblies (Chapter 2, Funk *et al*. 2018), finding from 333 to 1278 domains per population (Figure 3.3). Many of these domains were only a fraction of the full 850 bp domain model, so I applied a series of size and statistical filters (based on e-value) to create higher-confidence sets, which led to convergence of domain numbers across samples (Figure 3.3).

The EL10 reference genome, C869_US reference single plant, and C869_25 pooled population samples were all acquired from the same inbred population, providing a comparison

of how the assembly method and presumed heterozygosity influenced NB-ARC detection. Before filtering, the EL10 reference genome contained 250 domains, the fewest of the three genetically related samples. This is in comparison to the C869_US single plant assembly (343 domains) and the C96825 pooled population sample (559 domains). Filtering out short domains with statistical scores less than 1e-10 reduced the number of predicted domains, affecting the pooled samples disproportionately to the EL10 reference genome (Figure 3.3). The three related samples converged at the 650 bp filter, with 168 domains in each *de novo* assembly and 169 domains in the reference genome. This indicates that the NB-ARC domains over 650 bp in the reference genome were recovered in the pooled *de novo* assemblies. Similar convergence was seen across all samples, resulting in the 650 bp threshold being chosen to provide a conservative set of NB-ARC domains for further analysis.

Comparison of *de novo* and reference NB-ARC domains

I next wanted to classify the *de novo* domains and group them into homologous sequences. To accomplish this, I aligned predicted domains from the *de novo* assemblies to the reference genome. I counted the number of domains that mapped to each of the 231 reference loci and compared this number to the total domains found in each assembly (Figure 3.4). Non-sugar genotypes had the fewest matches to EL10 reference domains. The C869_US sugar beet, which is genetically related to the reference genome plant, had the most matches at 142. There was no correlation between overall number of domains and number of reference domains matched (Figure 3.4). This could be indicative of allelic variation within samples, locus-specific duplication, or novel NB-ARC domains present in some lineages.

I counted the number of homologous sequences mapping to each of the 231 reference NB-ARC domains to test if specific reference domains were over- or under-represented. I found

that thirty-seven reference domains below 650 bp in length had no representatives in the 23 *de novo* assemblies, owing largely to the fact that short reference domains would have their matches filtered by the quality control steps in the *de novo* analysis. Eighty-eight reference loci had multiple matches per assembled sample (Table 3.7). All of the reference NB-ARC domains greater than 650 bp had at least one homologous domain detected in the *de novo* assemblies, with the exception of 13 reference domains that were assembled from multiple concatenated fragments each fewer than 650 bp (Funk et al. 2018). Individual populations contained on average 139.5 of the 165 reference domains over 650 bp, with each reference domain found in 78% of populations (median 18, stdev 7.15). Only 23% of domains (43/165) had representatives in every population, confirming that different accessions and crop types harbor subsets of pan-genomic diversity.

Thirty-seven reference loci shorter than 650 bp had no assembled domains map to their position. This suggests that these domains were also below 650 bp in the assemblies and therefore removed by the 650 bp filter. In contrast, twenty-two reference domains had more than 23 mapped domains, indicating multiple alleles or duplicated domains in some populations. The reference locus with the most assembled domains was Bv.nbarc.0077, which had 65 separate sequences across all 23 assemblies. The NB-ARC domain associated with the *Rz2* resistance gene, Bv.nbarc.0121, was overlapped by 21 domains from 20 assemblies (Table 3.7) (Chapter 2, Funk et al. 2018).

Phylogenetic analysis of the most abundant NB-ARC domain Bv.nbarc.0077

If the assembled NB-ARC domains are genuine representations of sequences present in the populations, there should be phylogenetic relationships apparent between domains. To test this hypothesis, I constructed phylogenetic trees of domains mapping to the most-represented

domain Bv.nbarc.0077 as well as the Bv.nbarc.0121 domain associated with the *Rz2* resistance locus.

There were 65 predicted domains that mapped to Bv.nbarc.0077, which resolved into five well-supported clades (Figure 3.5). The two basal clades contained single domains from 17 different assemblies, with representatives of each crop type present in both clades. The two large terminal clades contained representatives from all 23 assemblies. The C869_25 assembly had a domain in one basal clade as well as both terminal clades. In contrast, the C869_US single plant assembly was only in one terminal clade, which could indicate successful assembly of the reference domain in both C869 samples as well as assembly of alleles or close paralogs in the more heterozygous C869_25 pooled sample.

Other assemblies also generated domains in a basal clade and both terminal clades, such as the four chard populations LUC, Vulcan, RHU, and FGSC, as well as sugar beets SP7322, GP9, GP10, SR98, and L19 (Figure 3.5). Some assemblies were only present in terminal clades, such as W357B (table), SR102 (sugar), EL50 (sugar), and EL51 (sugar). The strong bootstrap support for each clade, combined with patterns of presence/absence between clades, suggests that domain assembly was consistent between samples and that the phylogenetic relationships between domains are likely based on underlying genetic differences between populations rather than random artifacts. The Bv.nbarc.0077 domain could represent a highly polymorphic locus that cannot not be represented in a traditional haploid reference sequence.

Phylogenetic analysis of the *Rz2* CNL-associated NB-ARC domain Bv.nbarc.0121

My next goal was to investigate genetic diversity of the known resistance locus *Rz2*. I constructed another phylogenetic tree from the 21 domains that mapped to the *Rz2*-associated NB-ARC domain Bv.nbarc.0121 (Figure 3.6). This tree clearly linked the domain from the

C869_US single plant with the reference allele Bv.nbarc.0121, which remained separate from the largest two clades. Two chard varieties, Vulcan and RHU, formed outgroups to the tree, a third chard, TGSC, was grouped in the main clade, and the fourth chard, LUC, did not contain a domain mapping to this locus. Unlike the often-duplicated Bv.nbarc.0077 domains, only one assembly (table beet DDRT) had multiple Bv.nbarc.0121 domains. The two DDRT domains and another table beet, W357B, formed a well-supported clade clearly separated from the other assemblies, resulting in four total clades with bootstrap support of 96 or greater (Figure 3.6). The clade containing C86925 included chard, fodder, and table beet populations but no other sugar beet material from the East Lansing breeding program, which could coincide with segregation of an *Rz2* allele in the C86925 population that differs from the allele in the EL10 reference genome.

DISCUSSION

The goal of this study was to characterize genetic diversity in *B. vulgaris* with a focus on NB-ARC and Ser/Thr kinase protein domains. I extended the use of nucleic-acid-based HMMs to detection of kinase domains, identifying 598 putative Ser/Thr kinase domains and an additional 289 non-Ser/Thr kinase domains in the EL10 reference genome. To begin to characterize genetic diversity of disease resistance-related genes across *B. vulgaris*, I analyzed *de novo* assemblies of pooled sequences from 23 diverse populations. I scanned the assemblies for NB-ARC domains, finding on average 167 putative full-length domains per population. This allowed preliminary assessment of diversity at the most abundant locus, CNL Bv.nbarc.0077, as well as the *Rz2* resistance locus.

The nucleotide-based HMM is a powerful tool that can detect complete, fragmented, and diverged sequences that may or may not be part of coding sequences (Funk et al. 2018). This sensitivity is a benefit when attempting to identify novel genetic features and infer evolutionary

processes, but sensitivity creates challenges due to the gradient between functional sequences, non-functional duplications, pseudogenes, and spurious genomic noise (Krattinger & Keller 2016). It is possible that the presence of introns in the genome results in multiple model matches for what should be a single domain. This is less of a concern for NB-ARC domains, which are only rarely interrupted by introns (Funk et al. 2018). However, kinase domains are frequently split across multiple exons, as noted by previous literature (Shiu & Bleecker 2001) and their smaller domain fragments in EL10 (Figure 3.1). Despite these challenges, counting kinase domains that were clearly more than 50% of the HMM length led to 598 RLK domains in beet (Table 3.1). This is in agreement with numbers of RLKs in rice (646) and Arabidopsis (615) (Shiu et al. 2004; Shiu & Bleecker 2003), which supports the notion that the kinase HMM search was specific and complete. For NB-ARCs in the *de* novo populations, using a filter of 650 bp (~75%) recovered every reference allele in at least one assembly. This finding confirms that every reference NB-ARC locus was able to be assembled in the pooled population sequences. These two results taken together imply that conservative filtering of nucleotide HMM searches can be used to define a core set of non-redundant loci.

Identification of R gene loci across populations is not the same as defining R gene diversity. The methods employed here build upon the foundation that nucleotide HMMs have the sensitivity and specificity to capture target sequences in a high-quality reference genome (Funk et al. 2018). The extension of those methods to *de novo* assemblies from short reads, combined with the heterogeneity of pooled population sequencing, adds significant uncertainty. Evidence that the pooled short-read assemblies are reliable is crucial before making inferences about the R gene sequences found within. The evidence I used to ascertain assembly validity was based on two analyses: 1) the proportion of recovered genes in the eukaryote and embryophyta

orthologous gene sets, and 2) what proportion of the reference genome was covered by 1-to-1 unique mapping contigs.

Benchmarking Universal Single-copy Conserved Orthologs (BUSCO) is predicated on the notion that some core genes are preserved as single copies across the genomes of many species (Waterhouse et al. 2018). I assayed two sets of genes: a group of 303 genes for all eukaryotes, and a group of 1,375 plant-focused genes targeted at embryophyta (Tables 3.6 and 3.7, respectively). The number of fragmented eukaryotic BUSCOs was noticeably higher in the assemblies versus the reference, yet the total number of missing genes was comparable between all samples. This is remarkable considering the different methodology used to generate the genomes. Examining putative orthologous genes revealed that 24 of the "missing" genes were in fact missing from all *Beta* samples, including the reference (red lines, Figure 3.2). These missing genes could be in areas of the genome resistant to assembly, even with long-read sequencing technologies. Alternately, the missing orthologs could have diverged sufficiently to be outside the range of detection using universal gene models. Additional analysis might be able to identify these genes in the EL10 genome based on relaxed thresholds of homology. Finally, it's possible that what constitutes a "universal" gene is poorly understood, and different organisms could have satisfied core requirements for life using independent mechanisms. The embryophyta set more clearly differentiated the assembly methods, with more fragmented and missing genes in the pooled *de novo* assemblies vs the reference (Table 3.6). The fact that table beet populations had a small but significant increase in total complete + fragmented genes (91.1% vs 88.4%) could reflect differences in assembly quality based on the lower heterozygosity in table beet populations (Paul Galewski, in press). Overall the BUSCO analysis provides confidence that the assemblies capture true genetic features of the populations.

The other level of validation for each assembly was assessing the length of sequence that had unique mapping locations in the reference. The MUMmer analysis asked what portion of contigs had a single unique mapping location in the reference genome. The assemblies were between 65% and 80% of the 540 MB reference (median 73%). These are reasonable scores considering that repetitive sequences are not likely to be assembled into contigs over 200 bp using only short reads, and whatever did assemble is unlikely to be mapped uniquely. These lengths should cover the majority of gene space in the genome, providing access to unique population sequences.

Once I generated a baseline level of plausibility of the assemblies, I wanted to look at genetic diversity of individual loci. I began with phylogenetic analysis of the most abundant locus Bv.nbarc.0077 (Figure 3.5). Three clades comprising 74% (48/65) of the domains were characterized by extremely short branch lengths and high bootstrap support, which is what I would expect from sequences coming from recent shared ancestors. The fact that these three clades and the two basal outgroups were so clearly resolved lends additional weight to the idea that whatever is happening during pooled population sequencing is highly reproducible. The *de novo* domains shared enough identity to Bv.nbarc.0077 that they didn't inadvertently map to other NB-ARC loci in the reference genome, and yet three clades and two outgroups were consistently assembled into separate sequences and not chimeras. The precise topology of the phylogenetic tree could be characteristic of copy-number variation or multiple alleles. Distinguishing these two possibilities could be resolved by targeted cloning or sequencing and then extrapolated to other candidate loci in the future. The same molecular characterization, in combination with additional phenotyping, could also clarify whether genetic variation at a given locus correlates with phenotypic differences.

One of the most-studied genes in beets is *Rz2*, which confers partial resistance to the rhizomania disease caused by *Beet Nectotic Yellow Vein Virus* (Lewellen 1991). I applied the same phylogenetic analysis used for Bv.nbarc.0077 to the predicted domains that mapped to the *Rz2* locus Bv.nbarc.0121 (Figure 3.6). The resulting tree revealed two clades with extremely short branch lengths similar to the tree for Bv.nbarc.0077. The other central clade subdivision was exclusively sugar beet, although the intra-clade branch lengths were relatively long, indicating more variation within this group. The EL10 domain Bv.nbarc.0121 (Chapter 2, Funk et al. 2018) was added to the analysis and formed a distinct clade with the domain from the C869_US single plant. However, it is somewhat surprising that only one domain was identified from the related pooled population C86925. This could indicate that the EL10 reference allele for Bv.nbarc.0121 is at low frequency in the C86925 population, but the exact frequency is unknown. The germplasm release of C869 reported segregation for *Rz1* rhizomania resistance but did not mention *Rz2* (Lewellen 2004). If there was a minor allele of *Rz2* assembled from the C86925 pooled sequences, the associated domain was below the 650 bp threshold, the 1e-10 statistical threshold, or both. The large central clade had one subdivision with domains of high sequence similarity comprised of table beets, one chard, and, strangely, also the C86925 pooled sugar beet population (Figure 3.6). To my knowledge, the East Lansing sugar beet breeding program does not contain functional *Rz2* alleles, although they have not been targeted for detection in the system (J.M. McGrath, personal communication). The existing *Rz2* alleles could be subject to genetic drift heading toward pseudogenization. If this is the case, the alleles would have loosened selection pressure and begin to accumulate random lineage-specific mutations. This could be one explanation for the looser shape of the sugar beet clade. If that is true, then the SR102 domain forming an outgroup to the other sugar beets could be indicative of a unique

functional domain or simply a different history of drift. The other major feature of the tree is the extreme outlier formed by the domain from BBTB. The low bootstrap score means placement on the tree was nearly random, which is a feature shared with the other outliers Vulcan, RHU, and WT. It could be worth pursuing why two chard domains and two table beet domains are so different from everything else observed, and also from each other.

In summary, *de novo* assembly of pooled sequencing data is a promising strategy to efficiently characterize genetic diversity in existing populations. Both NB-ARC and kinase domains appear to be tractable targets of HMM scans, which together represent over 80% of known disease resistance genes in plants (Kourelis & Van Der Hoorn 2018). Further work is needed to fine-tune assembly parameters and validate the domains derived from *de novo* assemblies, but the preliminary results presented here provide a basis for further exploration of pan-genomic disease resistance diversity. Knowledge of relationships between disease resistance sequences could inform crossing decisions, candidate gene analysis, and help stack durable disease resistance for improved crop varieties.

METHODS

Domain HMM generation and scanning

Kinase superfamily domain IPR011009 and Ser/Thr kinase domain IPR000719 were identified in the InterProScan results of the preliminary protein prediction for the EL10 genome. The nucleotide coding sequences of domains with e-values below 1e-20 were extracted from the associated transcripts and aligned with MAFFT v. 6.716 with parameters --linsi --leavegappyregion. An HMM was generated from the alignment using the hmmbuild function of HMMER v 3.1 (Wheeler & Eddy 2013) with default parameters. The NB-ARC HMM was from

Funk et al. (2018). The nucleic acid HMM was used to query the EL10 v1.2 reference genome and *de novo* assemblies using the nhmmer function of HMMER with default parameters.

Population sampling, Illumina sequencing, and assembly

Seeds from twenty-three populations of *B. vulgaris* were soaked in 0.3% hydrogen peroxide for 24 hours to facilitate germination. These seeds were germinated in potting mix in plastic containers (24" W x 16 " D x 12" H) in the greenhouse under 16h day/8h night light cycle and temperatures between 65 and 80 degrees. Tissue was harvested at the two-leaf seedling stage, approximately two weeks after germination. For each population, leaf tissue was collected from twenty-five individuals and placed into a 50 mL conical polypropylene centrifuge tube (VWR, Radnor, PA) to create population-specific pools of tissue. These tissues were lyophilized and ground to a powder using a bead beater. For each population, DNA was isolated from 20 mg of lyophilized tissue using NucleoSpin Plant II kit (Macherey-Nagel, Duren, Germany). Libraries were prepared by the Michigan State Research Technology Support Facility (East Lansing, MI, USA) using Illumina TruSeq kits (Illumina, San Diegoi, CA, USA). Libraries were sequenced to 80x depth using HiSeq 2500 2 x 125bp paired-end chemistry. Illumina adapters were trimmed from the raw reads using Trimmomatic v.0.36 (Bolger et al. 2014) with parameters ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36. The raw reads were assembled by Paul Galewski using ABySS 2.0 with default parameters (Jackman et al. 2017).

Assessment of conserved orthologous gene assembly

Conserved orthologous genes were assessed with BUSCO v. 2.0.1 using default parameters (Simão et al. 2015). The two query gene sets were "eukaryota odb9" and "embryophyta odb10" available at https://busco.ezlab.org/.

Mapping NB-ARC domains to the reference genome

Predicted domains were mapped to the EL10 reference genome v. 1.2 (McGrath et al., in prep) using GMAP v. 2019-2-26 and default parameters (Wu & Nacu 2010). The single best mapping site was retained.

Phylogenetic analysis

Predicted NB-ARC domains were extracted from the genome assemblies and aligned with MAFFT v 7.215 with the –leavegappyregion option (Katoh & Standley 2013). The phylogenetic trees were constructed using RAxML v. 8.0.6 (Stamatakis 2014). The '-f a' function was used to conduct a rapid bootstrap analysis and search for the best-scoring maximum likelihood tree, in a single run using 1000 bootstrap replicates. The model of substitution was GTRGAMMA (Shapiro et al. 2006).

Whole genome alignment

Assembled genomes were aligned to the EL10 reference genome v. 1.2 using the nucmer module of MUMmer v. 4.0.0 (Marçais et al. 2018). Alignments were filtered using the delta-filter function with parameters '-i 80 -l 200' specifying minimum 80% identity and 200 bp query length, respectively.

ACKNOWLEDGEMENTS

Thanks to Paul Galweski for running the ABySS assembler and the Michigan State High-Performance Computing Center for technical support.

APPENDIX

Table 3.1. Kinase domains at least 400 bp long in the EL10 sugar beet reference genome.

Chromosome	non-Ser/Thr	Ser/Thr	Total
Chr1	25	44	69
Chr2	25	64	89
Chr3	37	74	111
Chr4	32	81	113
Chr5	26	46	72
Chr6	31	71	102
Chr7	35	68	103
Chr8	38	72	110
Chr9	28	62	90
Scaffold_0001	0	1	1
Scaffold_0002	2	2	4
Scaffold_0003	1	1	2
Scaffold_0004	0	1	1
Scaffold_0005	2	3	5
Scaffold_0006	4	5	9
Scaffold_0007	0	2	2
Scaffold_0009	0	1	1
Sum	286	598	884

Table 3.2. **Clusters of Ser/Thr kinase domains in the *B. vulgaris* EL10 reference genome.** Shading was applied to differentiate chromosomes.

Chr	Start	End	# Domains
Chr1	4,673,171	4,931,655	6
Chr2	676,787	717,896	10
Chr2	42,446,420	42,783,557	6
Chr2	49,593,854	49,724,247	5
Chr3	15,179,793	15,948,666	18
Chr4	56,105,830	56,849,766	25
Chr5	3,827,192	3,879,711	5
Chr6	6,921,781	7,994,442	11
Chr6	51,773,326	51,818,532	5
Chr6	60,070,057	60,411,923	5
Chr7	3,434,216	3,463,633	6
Chr8	43,625,244	43,932,126	8
Chr8	44,680,944	45,218,207	11
Chr8	49,914,043	50,133,051	9
Chr9	27,272,234	27,830,146	11

Table 3.3. Metrics for *de novo* assembly of pooled populations of *B. vulgaris*. The three genetically-related samples are the reference genome EL10, the C869_US sequenced from the same DNA as EL10, and the pooled C86925 population from which the EL10 plant was obtained.

Sample	Type	Number of contigs	Contigs > 500 bp	L50	N75	N50	Max contig (bp)	Total length (bp)
FGSC	chard	2,298,060	124,314	18,984	2,164	5,676	67,200	403,600,000
LUC	chard	6,836,362	225,936	41,148	1,628	3,720	63,564	567,800,000
RHU	chard	4,750,885	179,962	30,258	2,088	4,926	122,425	563,800,000
Vulcan	chard	2,427,916	126,944	19,070	2,551	6,586	85,653	477,700,000
MAM	fodder	4,228,318	155,723	24,956	2,130	5,252	85,199	503,600,000
WGF	fodder	4,916,463	184,342	30,885	1,998	4,746	66,355	557,500,000
C869_US	sugar	1,935,503	76,330	10,333	4,434	11,913	145,217	436,400,000
C86925	sugar	2,391,073	132,171	19,585	2,345	6,173	76,035	460,300,000
EL50	sugar	2,245,505	126,530	18,536	2,460	6,544	78,725	459,200,000
EL51	sugar	2,152,524	120,037	17,066	2,609	7,124	86,382	458,700,000
GP10	sugar	2,292,629	126,058	18,426	2,393	6,426	81,576	447,000,000
GP9	sugar	4,214,181	209,937	35,771	1,781	4,212	70,685	573,000,000
L19	sugar	5,077,152	184,733	32,147	2,071	4,791	89,113	571,500,000
SP7322	sugar	2,528,915	141,392	21,098	2,268	5,940	89,352	480,900,000
SR102	sugar	2,265,602	121,068	17,657	2,468	6,632	76,705	437,200,000
SR98	sugar	2,216,075	123,023	17,564	2,595	7,022	155,679	466,900,000
BBTB	table	1,914,427	105,592	15,214	3,125	8,311	81,701	470,200,000
Crosby	table	2,044,428	106,914	15,561	2,729	7,297	89,693	418,100,000
DDRT	table	2,017,523	104,925	15,012	2,847	7,666	86,975	425,800,000
RQ	table	2,237,760	112,369	16,175	2,759	7,456	85,116	451,000,000
TGSC	table	1,673,213	87,369	12,951	3,345	8,654	81,509	397,700,000
W357B	table	1,842,781	68,689	9,641	5,081	12,926	125,442	430,400,000
WT	table	2,048,310	106,104	15,353	2,914	7,785	86,534	442,000,000

Table 3.4. MUMmer alignment of *de novo* assembled contigs of 23 populations of *B. vulgaris*. Contigs were aligned to the sugar beet EL10 reference genome.

Sample	Type	Contigs mapping 1-to-1	Total Length (bp)	Average Length	Average Identity (%)
FGSC	chard	270,562	351,154,809	1297.87	96.80
LUC	chard	399,032	401,205,993	1005.45	96.81
RHU	chard	312,417	413,221,578	1322.66	97.29
Vulcan	chard	245,835	389,675,512	1585.11	97.20
MAM	fodder	308,195	397,319,242	1289.18	97.60
WGF	fodder	313,406	413,347,133	1318.89	97.45
C869_US	sugar	160,063	429,968,206	2686.24	99.56
C86925	sugar	227,457	396,959,171	1745.21	98.05
EL50	sugar	228,473	390,535,774	1709.33	97.59
EL51	sugar	225,898	391,786,636	1734.35	97.55
GP10	sugar	242,188	386,702,655	1596.7	97.69
GP9	sugar	291,881	425,770,026	1458.71	97.81
L19	sugar	325,107	421,951,081	1297.88	97.95
SP7322	sugar	246,520	398,507,205	1616.53	97.65
SR102	sugar	242,387	383,183,748	1580.88	97.63
SR98	sugar	229,338	395,140,302	1722.96	97.68
BBTB	table	213,995	393,654,653	1839.55	97.21
Crosby	table	236,422	367,810,632	1555.74	97.20
DDRT	table	228,462	372,737,989	1631.51	97.20
RQ	table	230,843	384,641,950	1666.25	97.29
TGSC	table	216,869	361,589,180	1667.32	97.19
W357B	table	190,163	384,835,309	2023.71	97.28
WT	table	228,039	379,865,789	1665.79	97.26

Table 3.5. Query of *de novo* assemblies using the eukaryota gene set of Benchmarking Universal Single-copy Conserved Orthologs (BUSCO) (Waterhouse et al. 2018).

Sample	Complete	Complete - single-copy	Complete - duplicated	Fragmented	Missing	Complete + Fragmented
FGSC	209 (69%)	191	18	39	55	82%
LUC	168 (55%)	131	37	62	73	76%
RHU	206 (68%)	165	41	48	49	84%
Vulcan	226 (74%)	194	32	30	47	84%
MAM	209 (69%)	174	35	45	49	84%
WGF	194 (64%)	150	44	55	54	82%
BBTB	234 (77%)	206	28	34	35	88%
Crosby	224 (73%)	200	24	38	41	86%
DDRT	228 (75%)	208	20	31	44	85%
RQ	219 (72%)	196	23	40	44	85%
TGSC	224 (73%)	202	22	31	48	84%
W357B	245 (80%)	217	28	22	36	88%
WT	232 (76%)	203	29	31	40	87%
C86925	213 (70%)	184	29	42	48	84%
C869_US	253 (83%)	227	26	15	35	88%
EL50	232 (76%)	204	28	27	44	85%
EL51	229 (75%)	199	30	32	42	86%
GP10	226 (74%)	200	26	37	40	87%
GP9	194 (64%)	163	31	48	61	80%
L19	197 (65%)	152	45	46	60	80%
SP7322	226 (74%)	193	33	32	45	85%
SR102	223 (73%)	192	31	39	41	86%
SR98	235 (77%)	205	30	30	38	87%
EL10	264 (87%)	240	24	4	35	88%

Total BUSCO groups searched: 303

Horizontal lines separate crop type groups top to bottom: chard, fodder, table, and sugar

Table 3.6. Query of *de novo* assemblies using the embryophyta gene set of Benchmarking Universal Single-copy Conserved Orthologs (BUSCO) (Waterhouse et al. 2018).

Sample	Complete	Complete - single-copy	Complete - duplicated	Fragmented	Missing	Complete + Fragmented
FGSC	904 (66%)	860	44	307	164	88%
LUC	622 (45%)	529	93	435	318	77%
RHU	879 (64%)	743	136	305	191	86%
Vulcan	994 (72%)	926	68	248	133	90%
MAM	910 (66%)	822	88	282	183	87%
WGF	818 (59%)	702	116	340	217	84%
BBTB	1049 (76%)	991	58	200	126	91%
Crosby	1015 (74%)	970	45	232	128	91%
DDRT	1019 (74%)	987	32	225	131	90%
RQ	1032 (75%)	993	39	222	121	91%
TGSC	1030 (75%)	1011	19	211	134	90%
W357B	1139 (83%)	1116	23	149	87	94%
WT	1051 (76%)	1008	43	192	132	90%
C86925	998 (73%)	950	48	237	140	90%
C869_US	1152 (84%)	1114	38	111	112	92%
EL50	954 (69%)	917	37	260	161	88%
EL51	1019 (74%)	967	52	221	135	90%
GP10	975 (71%)	937	38	268	132	90%
GP9	846 (62%)	732	114	333	196	86%
L19	823 (60%)	682	141	322	230	83%
SP7322	964 (70%)	900	64	269	142	90%
SR102	975 (71%)	934	41	249	151	89%
SR98	1010 (73%)	946	64	211	154	89%
EL10	1318 (96%)	1287	31	12	45	97%

Total BUSCO groups searched: 1375

Horizontal lines separate crop type groups: chard, fodder, table, and sugar

Table 3.7. Predicted domains mapping to NB-ARC loci in the sugar beet EL10 reference genome. Phylogenetic group assignment derived from Chapter 2 and Funk et al. 2018.

	Count of predicted domains by crop type						Location in EL10			
Domain	chard	sugar	table	fodder	Number of populations with domain	Number of predicted domains	Chr	Position	Length (bp)	Phylogenetic group
Bv.nbarc.0077	17	26	16	6	23	65	Chr3	2,526,326	882	8
Bv.nbarc.0196/ Bv.nbarc.0197	15	20	14	7	23	56	Chr7	44,685,699	861	25
Bv.nbarc.0063	11	23	11	5	23	50	Chr3	10,751,095	833	26
Bv.nbarc.0182	14	18	12	5	23	49	Chr6	42,007,184	834	19
Bv.nbarc.0012	12	19	11	6	23	48	Chr1	57,718,800	862	20
Bv.nbarc.0168	14	12	9	6	23	41	Chr5	44,161,416	854	22
Bv.nbarc.0129	9	16	6	5	23	36	Chr4	4,301,909	792	0
Bv.nbarc.0072	8	16	6	5	23	35	Chr3	15,832,082	782	0
Bv.nbarc.0068	9	15	6	3	23	33	Chr3	13,708,377	832	17
Bv.nbarc.0085	8	16	7	2	23	33	Chr3	3,793,573	849	21
Bv.nbarc.0236	8	11	9	2	23	30	Chr9	15,266,297	820	23
Bv.nbarc.0232	6	12	8	2	23	28	Chr8	46,866,939	786	0
Bv.nbarc.0191	5	13	6	4	23	28	Chr7	21,093,593	870	8
Bv.nbarc.0204	5	13	7	2	23	27	Chr7	56,275,021	657	1
Bv.nbarc.0202	6	12	6	2	23	26	Chr7	56,225,603	810	3
Bv.nbarc.0050	7	10	6	2	23	25	Chr2	48,382,459	859	0
Bv.nbarc.0231	6	10	6	2	23	24	Chr8	46,206,033	841	0
Bv.nbarc.0014	5	10	6	3	23	24	Chr1	7,707,563	898	18
Bv.nbarc.0158	5	10	6	3	23	24	Chr5	13,528,818	828	4
Bv.nbarc.0060	5	11	6	2	23	24	Chr2	6,210,542	851	14
Bv.nbarc.0089	5	11	6	2	23	24	Chr3	46,107,603	858	25
Bv.nbarc.0161	5	11	6	2	23	24	Chr5	3,284,084	836	27
Bv.nbarc.0015	5	10	6	2	23	23	Chr1	7,714,137	877	18
Bv.nbarc.0010	5	10	6	2	23	23	Chr1	42,983,211	783	9
Bv.nbarc.0016	5	10	6	2	23	23	Chr2	1,530,537	851	27
Bv.nbarc.0062	5	10	6	2	23	23	Chr2	7,037,682	845	14
Bv.nbarc.0119	5	10	6	2	23	23	Chr3	7,472,443	763	11
Bv.nbarc.0064	5	10	6	2	23	23	Chr3	11,368,213	863	25
Bv.nbarc.0070	5	10	6	2	23	23	Chr3	15,746,962	861	25
Bv.nbarc.0071	5	10	6	2	23	23	Chr3	15,792,701	796	0
Bv.nbarc.0097	5	10	6	2	23	23	Chr3	51,527,853	886	17
Bv.nbarc.0122	5	10	6	2	23	23	Chr4	16,926,661	823	8
Bv.nbarc.0171	5	10	6	2	23	23	Chr5	44,172,397	933	22
Bv.nbarc.0184	5	10	6	2	23	23	Chr6	9,791,224	716	0
Bv.nbarc.0198	5	10	6	2	23	23	Chr7	55,820,419	879	21
Bv.nbarc.0199	5	10	6	2	23	23	Chr7	56,133,936	648	1
Bv.nbarc.0200	5	10	6	2	23	23	Chr7	56,148,648	654	1
Bv.nbarc.0201	5	10	6	2	23	23	Chr7	56,167,061	827	3
Bv.nbarc.0213	5	10	6	2	23	23	Chr7	56,504,438	817	3
Bv.nbarc.0227	5	10	6	2	23	23	Chr8	14,165,838	808	0
Bv.nbarc.0229	5	10	6	2	23	23	Chr8	35,544,883	787	27
Bv.nbarc.0238	5	10	6	2	23	23	Chr9	24,745,296	817	23
Bv.nbarc.0250	5	10	6	2	23	23	Chr9	49,904,363	779	0

Table 3.7 (cont'd)

Bv.nbarc.0034	14	13	15	3	22	45	Chr2	3,563,556	829	14
Bv.nbarc.0128	10	15	8	3	22	36	Chr4	4,254,666	789	6
Bv.nbarc.0130	6	11	8	3	22	28	Chr4	4,321,501	785	6
Bv.nbarc.0033	6	13	6	2	22	27	Chr2	3,552,119	836	14
Bv.nbarc.0074	7	10	6	3	22	26	Chr3	20,840,767	854	26
Bv.nbarc.0036	6	11	6	2	22	25	Chr2	40,643,651	794	10
Bv.nbarc.0210	5	12	6	1	22	24	Chr7	56,460,251	767	3
Bv.nbarc.0002	4	10	7	2	22	23	Chr1	12,964,204	784	12
Bv.nbarc.0239	4	10	6	3	22	23	Chr9	24,880,509	864	25
Bv.nbarc.0120	5	9	6	2	22	22	Chr3	7,617,311	739	11
Bv.nbarc.0088	5	9	6	2	22	22	Chr3	44,825,379	797	14
Bv.nbarc.0013	4	10	6	2	22	22	Chr1	6,746,597	807	26
Bv.nbarc.0018	4	10	6	2	22	22	Chr2	15,909,900	820	14
Bv.nbarc.0037	4	10	6	2	22	22	Chr2	40,679,823	778	10
Bv.nbarc.0049	4	10	6	2	22	22	Chr2	48,372,030	811	22
Bv.nbarc.0208	4	10	6	2	22	22	Chr7	56,426,426	783	3
Bv.nbarc.0209	4	10	6	2	22	22	Chr7	56,451,912	815	2
Bv.nbarc.0052	7	9	6	3	21	25	Chr2	48,680,966	799	14
Bv.nbarc.0194	7	10	5	2	21	24	Chr7	42,024,505	860	26
Bv.nbarc.0172	6	10	6	2	21	24	Chr5	5,665,746	782	13
Bv.nbarc.0147	5	10	6	2	21	23	Chr4	54,002,101	845	24
Bv.nbarc.0043	6	9	5	2	21	22	Chr2	44,826,286	857	20
Bv.nbarc.0093	6	9	5	2	21	22	Chr3	49,686,639	795	19
Bv.nbarc.0164	5	10	6	1	21	22	Chr5	43,595,309	803	0
Bv.nbarc.0051	5	10	5	2	21	22	Chr2	48,673,998	804	14
Bv.nbarc.0027	5	9	6	1	21	21	Chr2	1,825,884	843	27
Bv.nbarc.0100	5	9	6	1	21	21	Chr3	51,550,251	760	0
Bv.nbarc.0206	4	9	6	2	21	21	Chr7	56,347,144	695	1
Bv.nbarc.0203	5	10	4	2	21	21	Chr7	56,241,312	637	1
Bv.nbarc.0207	4	10	5	2	21	21	Chr7	56,372,326	852	2
Bv.nbarc.0185	3	10	6	2	21	21	Chr7	14,975,714	863	25
Bv.nbarc.0065	5	9	6	3	20	23	Chr3	11,375,854	868	25
Bv.nbarc.0076	3	11	6	2	20	22	Chr3	21,988,362	860	26
Bv.nbarc.0121	4	9	6	2	20	21	Chr3	9,281,663	856	26
Bv.nbarc.0225	4	8	6	2	20	20	Chr7	56,796,048	772	3
Bv.nbarc.0056/ Bv.nbarc.0057	4	9	5	2	20	20	Chr2	49,832,616	848	12
Bv.nbarc.0111	3	9	6	2	20	20	Chr3	51,721,900	804	17
Bv.nbarc.0173	4	10	5	1	20	20	Chr5	5,673,545	779	13
Bv.nbarc.0075	2	10	6	2	20	20	Chr3	21,956,199	848	26
Bv.nbarc.0084	6	9	14	3	19	32	Chr3	3,786,095	846	21
Bv.nbarc.0040	5	11	6	2	19	24	Chr2	44,794,361	860	20
Bv.nbarc.0078	5	12	4	1	19	22	Chr3	2,663,692	686	0
Bv.nbarc.0211	4	8	7	2	19	21	Chr7	56,476,724	635	1
Bv.nbarc.0092	4	8	6	1	19	19	Chr3	48,160,569	750	19
Bv.nbarc.0102	4	8	6	1	19	19	Chr3	51,568,329	797	16
Bv.nbarc.0150	2	9	6	2	19	19	Chr4	8,635,156	624	0
Bv.nbarc.0096	9	9	5	3	18	26	Chr3	51,504,455	835	15
Bv.nbarc.0114	6	10	6	3	18	25	Chr3	51,766,681	753	16
Bv.nbarc.0219	4	8	5	1	18	18	Chr7	56,640,134	648	1

Table 3.7 (cont'd)

Bv.nbarc.0094	5	10	1	2	18	18	Chr3	51,177,270	755	0
Bv.nbarc.0003	8	13	5	1	17	27	Chr1	13,221,944	774	12
Bv.nbarc.0035	5	15	7	0	17	27	Chr2	3,586,743	836	14
Bv.nbarc.0032	2	9	8	5	17	24	Chr2	3,532,153	834	14
Bv.nbarc.0044	3	8	7	3	17	21	Chr2	44,837,375	858	20
Bv.nbarc.0157	3	7	6	3	17	19	Chr5	10,691,853	767	13
Bv.nbarc.0103	4	7	6	1	17	18	Chr3	51,637,898	718	15
Bv.nbarc.0234	2	11	3	2	17	18	Chr8	55,840,666	840	24
Bv.nbarc.0212	4	6	6	1	17	17	Chr7	56,484,802	658	1
Bv.nbarc.0061	4	10	1	2	17	17	Chr2	6,916,353	826	14
Bv.nbarc.0030	9	8	13	0	16	30	Chr2	2,756,956	832	14
Bv.nbarc.0156	4	16	2	2	16	24	Chr5	10,616,231	768	13
Bv.nbarc.0233	6	13	3	0	16	22	Chr8	55,834,354	837	24
Bv.nbarc.0087	6	11	2	1	16	20	Chr3	3,812,887	864	21
Bv.nbarc.0154	4	7	5	1	16	17	Chr4	9,339,104	825	4
Bv.nbarc.0110	1	8	5	2	16	16	Chr3	51,718,521	884	17
Bv.nbarc.0187	1	9	10	2	15	22	Chr7	16,860,855	857	20
Bv.nbarc.0086	1	9	5	3	15	18	Chr3	3,801,295	865	21
Bv.nbarc.0153	4	8	1	2	15	15	Chr4	9,329,735	830	4
Bv.nbarc.0073	2	10	1	2	15	15	Chr3	20,833,240	859	26
Bv.nbarc.0042	1	10	2	2	15	15	Chr2	44,814,665	857	20
Bv.nbarc.0082	1	12	5	1	14	19	Chr3	3,773,883	848	21
Bv.nbarc.0240	3	6	6	3	14	18	Chr9	28,727,870	832	24
Bv.nbarc.0188	4	5	6	2	14	17	Chr7	17,120,791	853	20
Bv.nbarc.0151	2	7	5	1	14	15	Chr4	9,272,315	821	4
Bv.nbarc.0059	4	8	0	2	14	14	Chr2	6,112,630	831	14
Bv.nbarc.0117	2	10	0	2	14	14	Chr3	51,935,689	772	15
Bv.nbarc.0189	3	8	7	0	13	18	Chr7	17,151,726	857	20
Bv.nbarc.0054	1	11	5	1	13	18	Chr2	4,934,197	829	14
Bv.nbarc.0045	6	6	3	1	13	16	Chr2	44,870,240	859	20
Bv.nbarc.0143	3	3	6	1	13	13	Chr4	45,618,316	268	14
Bv.nbarc.0218	5	6	4	1	12	16	Chr7	56,612,457	652	1
Bv.nbarc.0024	5	6	4	0	12	15	Chr2	1,710,379	775	27
Bv.nbarc.0001	0	11	3	0	12	14	Chr1	12,930,202	768	12
Bv.nbarc.0041	1	9	2	1	12	13	Chr2	44,803,073	813	20
Bv.nbarc.0021	4	4	4	0	12	12	Chr2	1,660,825	756	27
Bv.nbarc.0216	3	5	2	2	12	12	Chr7	56,547,668	797	2
Bv.nbarc.0190	2	5	5	0	12	12	Chr7	17,285,303	890	7
Bv.nbarc.0023	4	6	0	2	12	12	Chr2	1,704,997	148	27
Bv.nbarc.0123	3	6	2	1	12	12	Chr4	1,882,767	857	25
Bv.nbarc.0026	2	6	4	0	12	12	Chr2	1,819,277	888	27
Bv.nbarc.0247	2	7	3	0	12	12	Chr9	48,251,017	859	26
Bv.nbarc.0031	4	5	4	1	11	14	Chr2	3,516,321	834	14
Bv.nbarc.0148	2	7	2	2	11	13	Chr4	8,470,174	849	5
Bv.nbarc.0105	5	4	1	1	11	11	Chr3	51,658,816	884	17
Bv.nbarc.0144	2	7	1	1	11	11	Chr4	45,642,671	581	14
Bv.nbarc.0205	2	4	3	1	10	10	Chr7	56,344,391	658	1
Bv.nbarc.0169/ Bv.nbarc.0170	1	5	4	0	10	10	Chr5	44,166,728	897	22
Bv.nbarc.0214	5	3	2	1	9	11	Chr7	56,520,764	672	1

Table 3.7 (cont'd)

Bv.nbarc.0004	1	4	6	0	9	11	Chr1	13,237,558	777	12
Bv.nbarc.0244	2	3	4	0	9	9	Chr9	48,153,337	859	26
Bv.nbarc.0167	2	7	0	0	9	9	Chr5	44,132,111	861	22
Bv.nbarc.0017	2	6	0	1	8	9	Chr2	1,582,676	792	27
Bv.nbarc.0224	2	3	1	2	8	8	Chr7	56,746,673	570	2
Bv.nbarc.0155	0	4	4	0	8	8	Chr4	9,391,065	827	4
Bv.nbarc.0113	0	4	3	1	7	8	Chr3	51,734,884	752	16
Bv.nbarc.0019	2	6	0	0	7	8	Chr2	1,596,627	858	25
Bv.nbarc.0145	1	3	3	0	7	7	Chr4	4,937,580	795	19
Bv.nbarc.0246	1	3	3	0	7	7	Chr9	48,242,437	858	26
Bv.nbarc.0107/										
Bv.nbarc.0108	4	1	4	1	6	10	Chr3	51,701,706	569	15
Bv.nbarc.0160	0	3	5	0	6	8	Chr5	22,999,237	831	5
Bv.nbarc.0223	0	2	4	1	6	7	Chr7	56,722,969	688	1
Bv.nbarc.0055	1	3	3	0	6	7	Chr2	4,943,690	836	14
Bv.nbarc.0124	0	2	4	0	6	6	Chr4	1,913,572	854	25
Bv.nbarc.0058	1	3	2	0	6	6	Chr2	5,000,636	831	14
Bv.nbarc.0249	2	4	0	0	6	6	Chr9	48,287,349	856	26
Bv.nbarc.0020	2	4	0	0	5	6	Chr2	1,650,948	742	27
Bv.nbarc.0038	0	1	4	0	5	5	Chr2	44,780,392	859	20
Bv.nbarc.0241	0	7	0	0	4	7	Chr9	28,754,759	841	24
Bv.nbarc.0152	1	2	0	2	4	5	Chr4	9,305,033	823	4
Bv.nbarc.0221	3	1	0	0	4	4	Chr7	56,659,535	688	1
Bv.nbarc.0248	2	2	0	0	4	4	Chr9	48,263,009	856	26
Bv.nbarc.0215	0	2	2	0	4	4	Chr7	56,539,263	673	1
Bv.nbarc.0222	0	4	0	0	4	4	Chr7	56,684,938	779	2
Bv.nbarc.0125	2	1	0	0	3	3	Chr4	34,558,483	322	0
Bv.nbarc.0053	1	1	1	0	3	3	Chr2	4,918,682	839	14
Bv.nbarc.0046	1	2	0	0	3	3	Chr2	47,369,433	263	0
Bv.nbarc.0146	0	2	1	0	3	3	Chr4	5,050,630	795	19
Bv.nbarc.0162	0	3	0	0	3	3	Chr5	42,252,844	801	19
Bv.nbarc.0106	2	0	0	0	2	2	Chr3	51,682,457	389	17
Bv.nbarc.0245	1	0	1	0	2	2	Chr9	48,209,642	859	26
Bv.nbarc.0066	1	1	0	0	2	2	Chr3	11,491,939	841	26
Bv.nbarc.0029	0	0	2	0	1	2	Chr2	2,742,904	835	14
Bv.nbarc.0047/										
Bv.nbarc.0048	1	0	0	0	1	1	Chr2	48,308,428	583	26
Bv.nbarc.0217	0	0	1	0	1	1	Chr7	56,590,628	651	1
Bv.nbarc.0011	0	1	0	0	1	1	Chr1	42,989,247	383	11
Bv.nbarc.0028	0	1	0	0	1	1	Chr2	2,715,384	830	14
Bv.nbarc.0115	0	1	0	0	1	1	Chr3	51,902,160	797	16
Bv.nbarc.0174	0	1	0	0	1	1	Chr6	10,519,056	304	0
Bv.nbarc.0005	0	0	0	0	0	0	Chr1	18,669,315	145	12
Bv.nbarc.0006/										
Bv.nbarc.0007/										
Bv.nbarc.0008	0	0	0	0	0	0	Chr1	38,224,584	748	14
Bv.nbarc.0009	0	0	0	0	0	0	Chr1	41,307,114	343	12
Bv.nbarc.0022	0	0	0	0	0	0	Chr2	1,676,563	374	27
Bv.nbarc.0025	0	0	0	0	0	0	Chr2	1,766,028	585	27
Bv.nbarc.0039	0	0	0	0	0	0	Chr2	44,792,516	309	20

Table 3.7 (cont'd)

Bv.nbarc.0079/										
Bv.nbarc.0080	0	0	0	0	0	0	Chr3	2,670,498	826	0
Bv.nbarc.0083	0	0	0	0	0	0	Chr3	3,782,072	253	21
Bv.nbarc.0067	0	0	0	0	0	0	Chr3	13,703,588	85	22
Bv.nbarc.0069	0	0	0	0	0	0	Chr3	13,724,032	54	22
Bv.nbarc.0081	0	0	0	0	0	0	Chr3	36,387,225	147	0
Bv.nbarc.0090	0	0	0	0	0	0	Chr3	46,120,105	86	25
Bv.nbarc.0091	0	0	0	0	0	0	Chr3	46,120,587	181	25
Bv.nbarc.0095	0	0	0	0	0	0	Chr3	51,471,887	87	27
Bv.nbarc.0098	0	0	0	0	0	0	Chr3	51,533,578	95	0
Bv.nbarc.0099	0	0	0	0	0	0	Chr3	51,536,211	204	0
Bv.nbarc.0101	0	0	0	0	0	0	Chr3	51,556,392	96	8
Bv.nbarc.0104	0	0	0	0	0	0	Chr3	51,652,191	96	8
Bv.nbarc.0109	0	0	0	0	0	0	Chr3	51,713,080	96	8
Bv.nbarc.0112	0	0	0	0	0	0	Chr3	51,727,265	203	17
Bv.nbarc.0116	0	0	0	0	0	0	Chr3	51,925,255	96	8
Bv.nbarc.0118	0	0	0	0	0	0	Chr3	51,975,010	95	0
Bv.nbarc.0126/										
Bv.nbarc.0127	0	0	0	0	0	0	Chr4	362,526	701	7
Bv.nbarc.0149	0	0	0	0	0	0	Chr4	8,632,594	556	0
Bv.nbarc.0131/										
Bv.nbarc.0132	0	0	0	0	0	0	Chr4	43,642,613	838	7
Bv.nbarc.0133/										
Bv.nbarc.0134	0	0	0	0	0	0	Chr4	43,763,563	839	7
Bv.nbarc.0135/										
Bv.nbarc.0136	0	0	0	0	0	0	Chr4	44,096,241	835	7
Bv.nbarc.0137/										
Bv.nbarc.0138	0	0	0	0	0	0	Chr4	44,112,203	835	7
Bv.nbarc.0139/										
Bv.nbarc.0140	0	0	0	0	0	0	Chr4	44,135,021	833	7
Bv.nbarc.0141/										
Bv.nbarc.0142	0	0	0	0	0	0	Chr4	44,186,624	827	7
Bv.nbarc.0159	0	0	0	0	0	0	Chr5	16,216,648	419	7
Bv.nbarc.0163	0	0	0	0	0	0	Chr5	43,574,643	85	0
Bv.nbarc.0165/										
Bv.nbarc.0166	0	0	0	0	0	0	Chr5	43,624,210	807	0
Bv.nbarc.0177/										
Bv.nbarc.0178	0	0	0	0	0	0	Chr6	3,706,524	730	7
Bv.nbarc.0179	0	0	0	0	0	0	Chr6	3,727,205	226	7
Bv.nbarc.0180/										
Bv.nbarc.0181	0	0	0	0	0	0	Chr6	3,737,398	770	7
Bv.nbarc.0175	0	0	0	0	0	0	Chr6	13,174,097	308	20
Bv.nbarc.0176	0	0	0	0	0	0	Chr6	25,821,110	308	20
Bv.nbarc.0183	0	0	0	0	0	0	Chr6	52,432,576	102	24
Bv.nbarc.0186	0	0	0	0	0	0	Chr7	15,986,651	135	12
Bv.nbarc.0192	0	0	0	0	0	0	Chr7	32,143,206	596	1
Bv.nbarc.0193	0	0	0	0	0	0	Chr7	39,496,010	133	12
Bv.nbarc.0195	0	0	0	0	0	0	Chr7	44,659,105	174	25
Bv.nbarc.0220	0	0	0	0	0	0	Chr7	56,642,615	634	1

Table 3.7 (cont'd)

Bv.nbarc.0226	0	0	0	0	0	0	Chr7	56,821,347	621	1
Bv.nbarc.0228	0	0	0	0	0	0	Chr8	27,969,672	595	1
Bv.nbarc.0230	0	0	0	0	0	0	Chr8	38,571,249	429	9
Bv.nbarc.0235	0	0	0	0	0	0	Chr9	13,926,539	346	26
Bv.nbarc.0237	0	0	0	0	0	0	Chr9	18,991,181	135	12
Bv.nbarc.0242/ Bv.nbarc.0243	0	0	0	0	0	0	Chr9	33,474,866	761	0

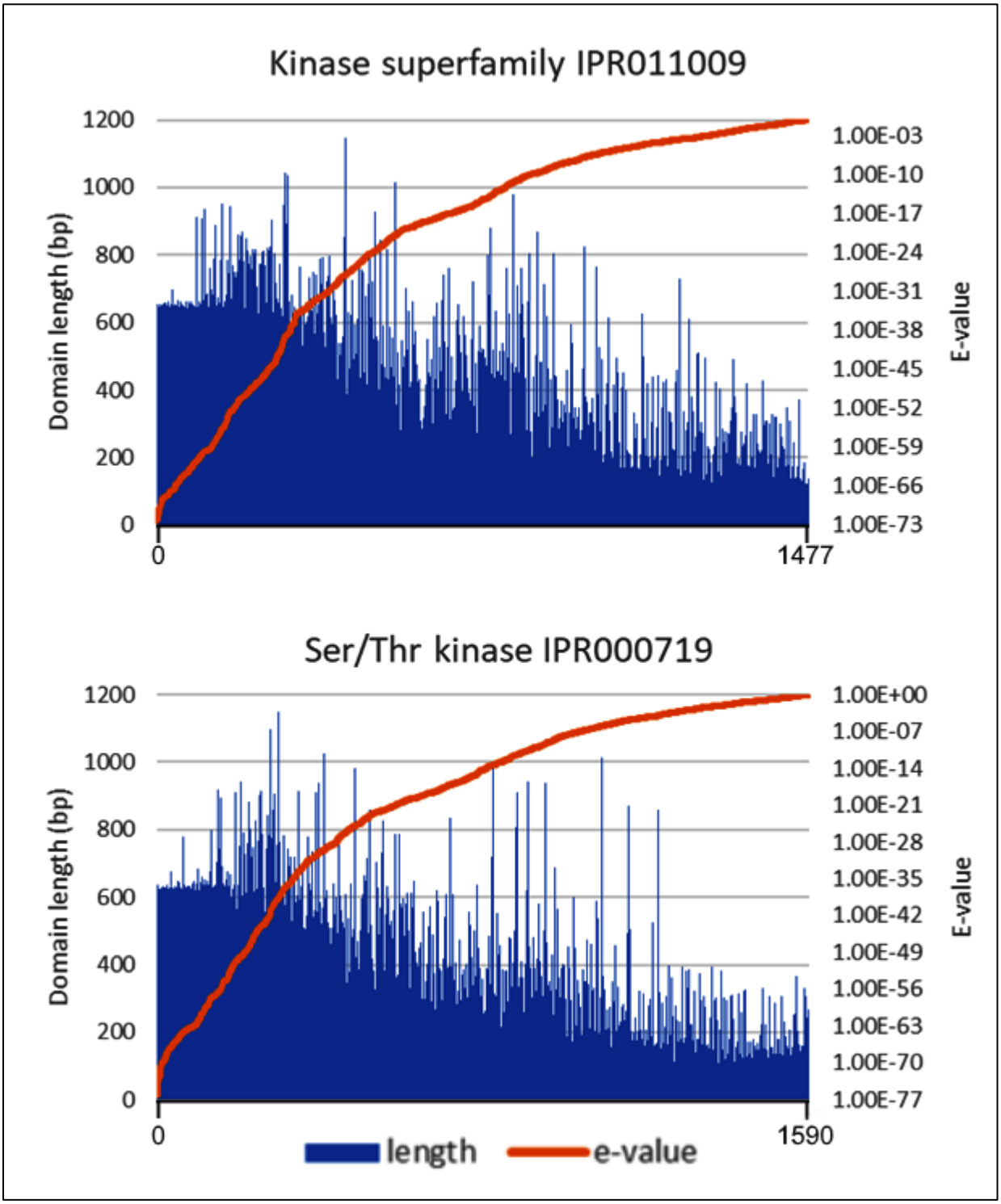

Figure 3.1. Total kinase domains detected in the sugar beet EL10 reference genome by each of two hidden Markov models (HMMs). There were 1,477 matches for the kinase superfamily domain and 1,590 matches for the Ser/Thr kinase domain. Each domain (x-axis) is defined by length in bp and e-value of HMM match.

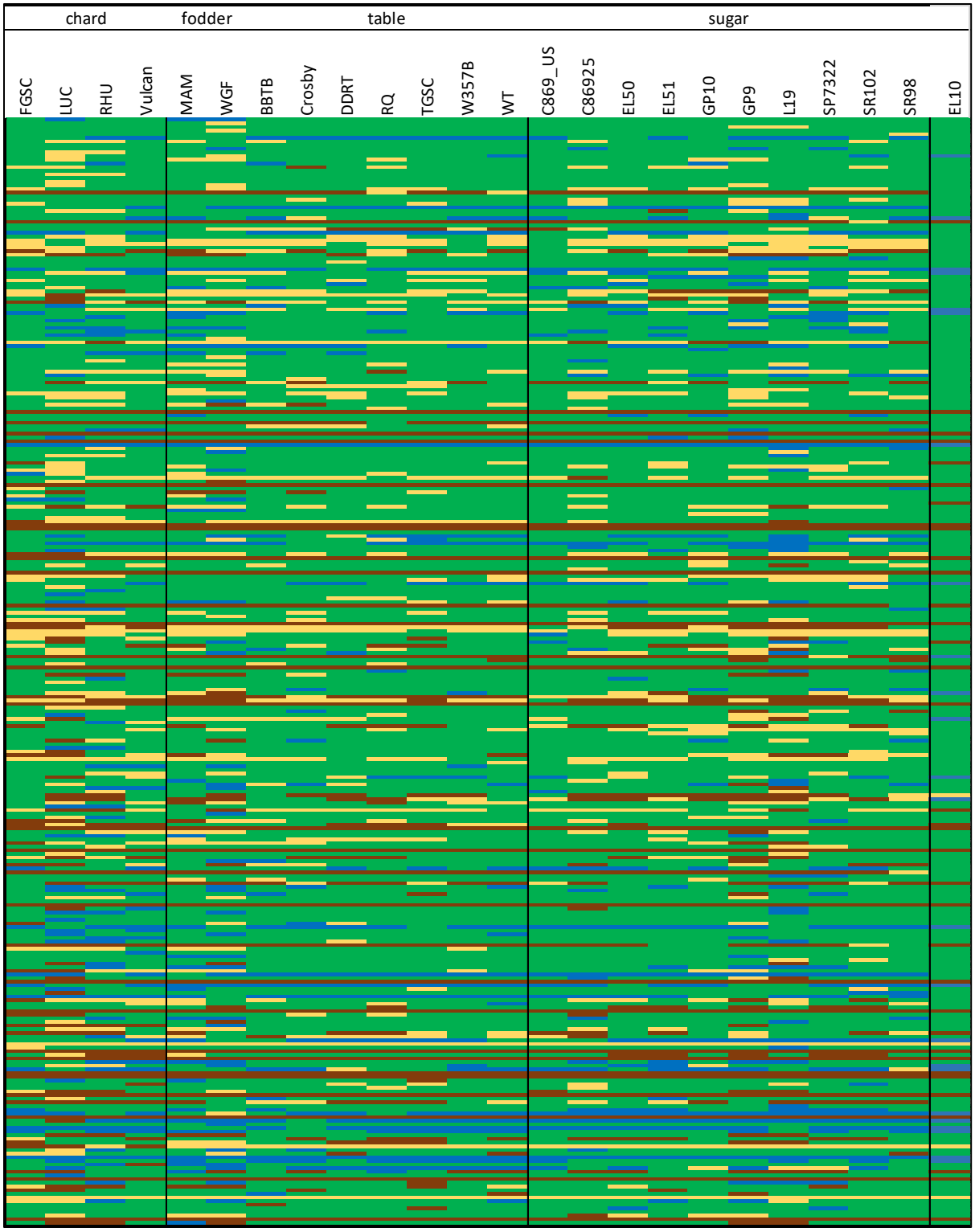

Figure 3.2. Single-copy universal conserved orthologs in *de novo* assemblies of *B. vulgaris*. Assemblies and EL10 reference genome are depicted in columns according to crop type (chard, fodder, table beet, sugar beet, EL10 reference). Each row represents a unique conserved gene. Cells are shaded according to the result of each specific gene in each assembly sample. Green: complete single-copy, blue: complete duplicated, yellow: fragmented, red: missing.

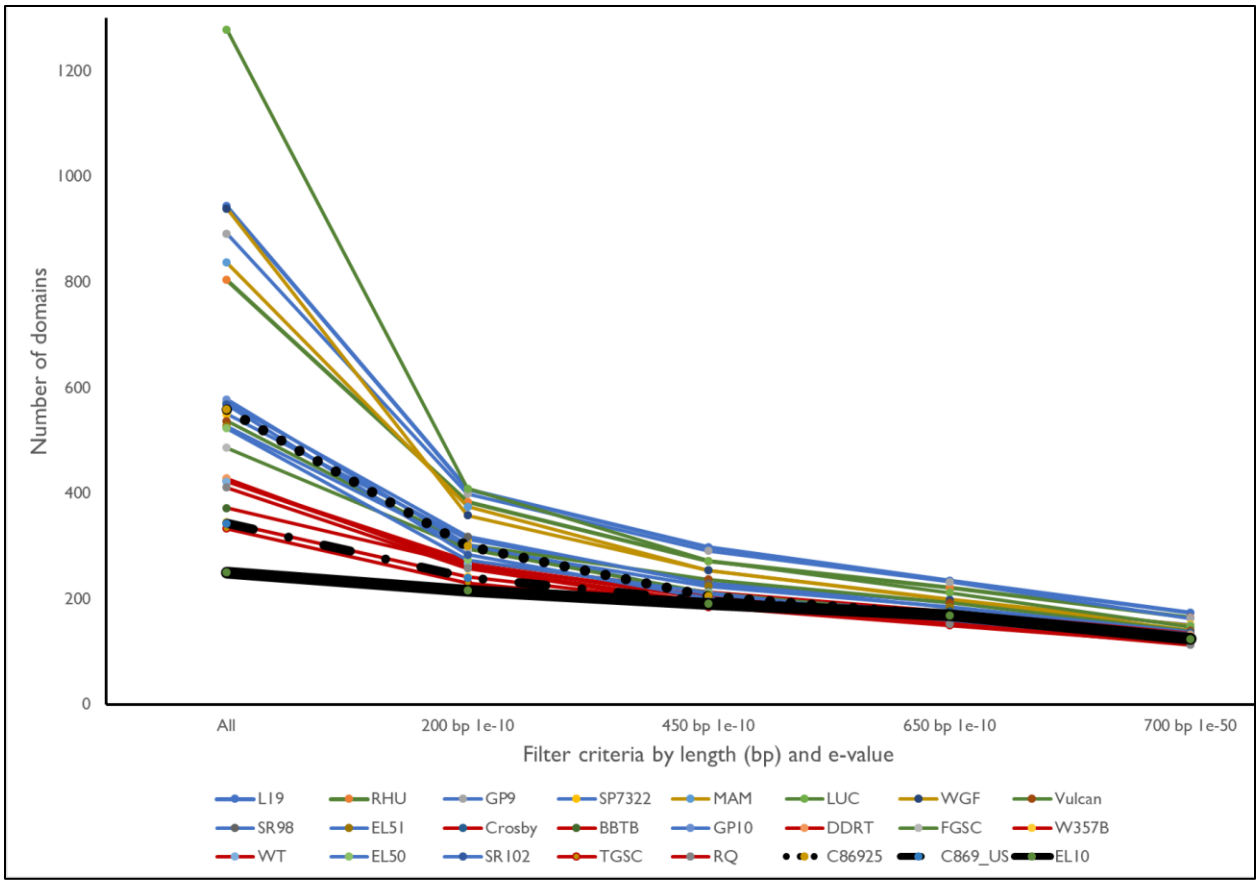

Figure 3.3. NB-ARC domains detected in *de novo* assemblies of *B. vulgaris*. Each sample is color coded according to its crop type. Blue: sugar, green: chard, yellow: fodder, red: table, black: reference. The EL10 reference genome is solid black, the single reference plant assembled from short reads (C869_US) is dot-dash black, and the assembly from pooled reads of the reference population (C86925) is dotted black. All domains detected at default HMM e-value of 1 are shown in the first column labeled All. Applying filters based on domain length in bp and domain e-value are shown in subsequent columns: 200 bp 1e-10, 450 bp 1e-10, 650 bp 1e-10, and 700 bp 1e-50.

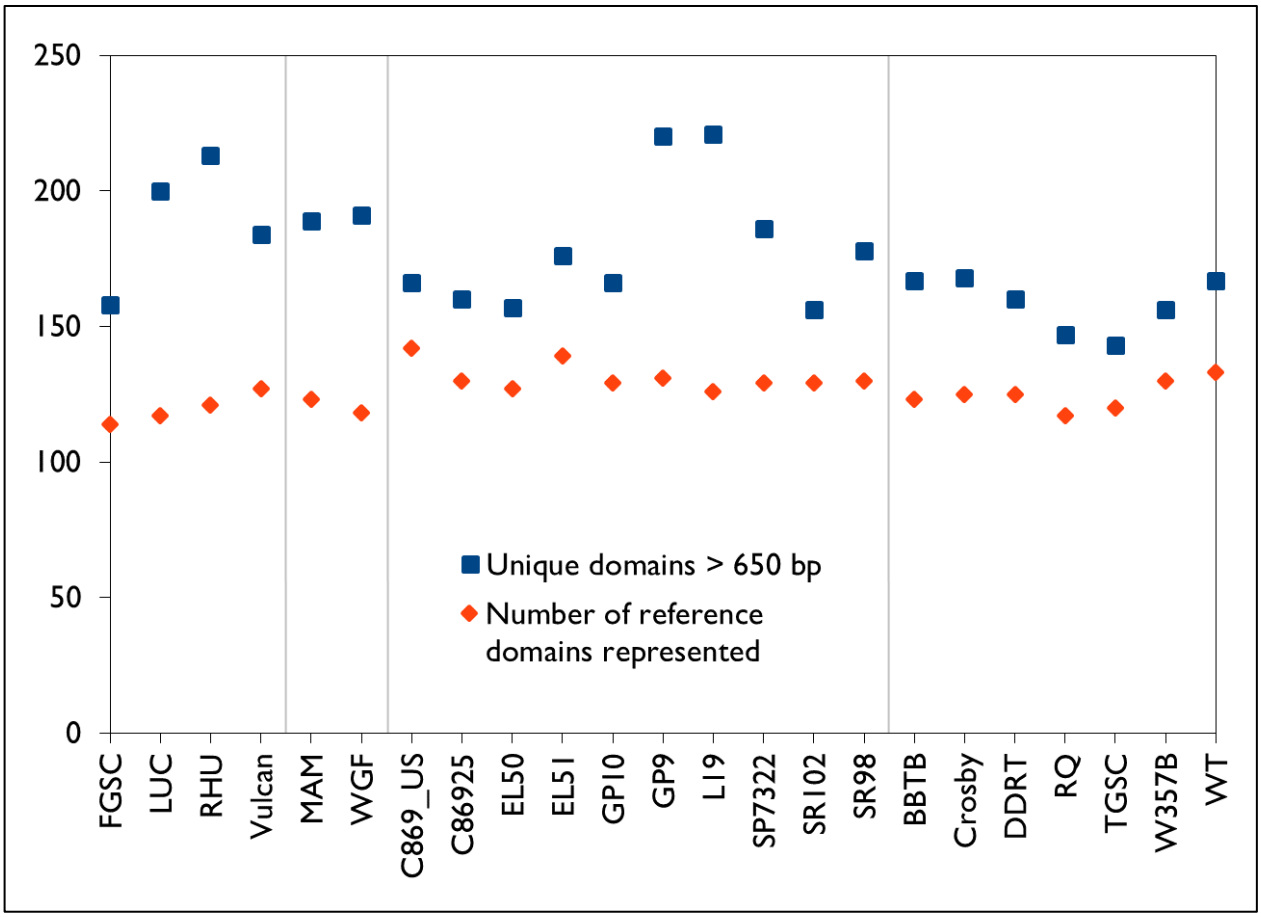

Figure 3.4. Number of NB-ARC domains per *de novo* assembly versus number of reference domains covered after alignment to the sugar beet EL10 reference genome. Populations are grouped left to right by crop type: chard, fodder, sugar, and table beet.

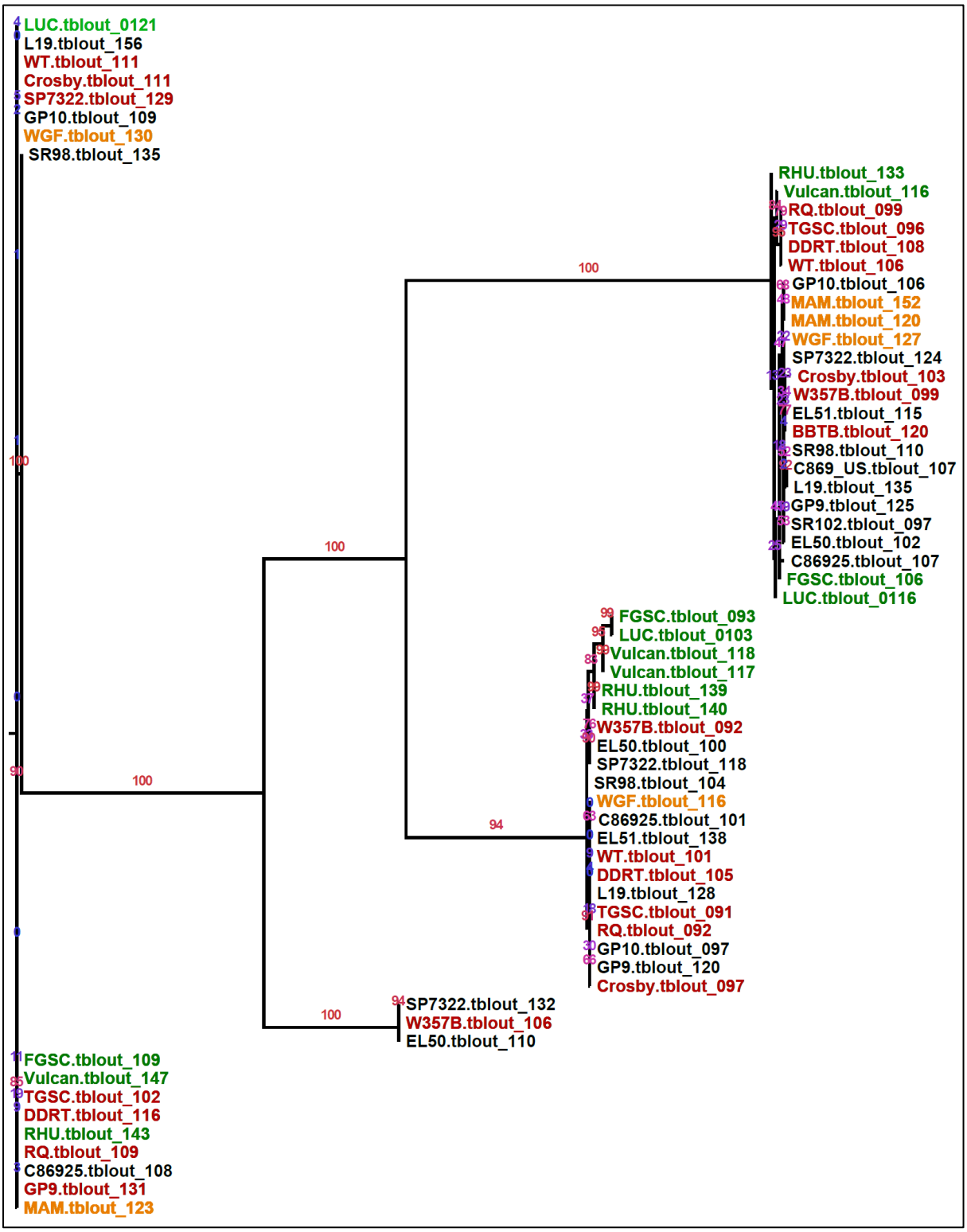

Figure 3.5. Phylogeny of NB-ARC domains mapping to Bv.nbarc.0077 in the *B. vulgaris* EL10 reference genome. Domains were extracted from 23 *de novo* assemblies of *B. vulgaris*. Populations are colored by crop type: green is chard, red is table beet, gold is fodder beet, and black is sugar beet. Bootstrap scores of 1000 replications are shown at branch points.

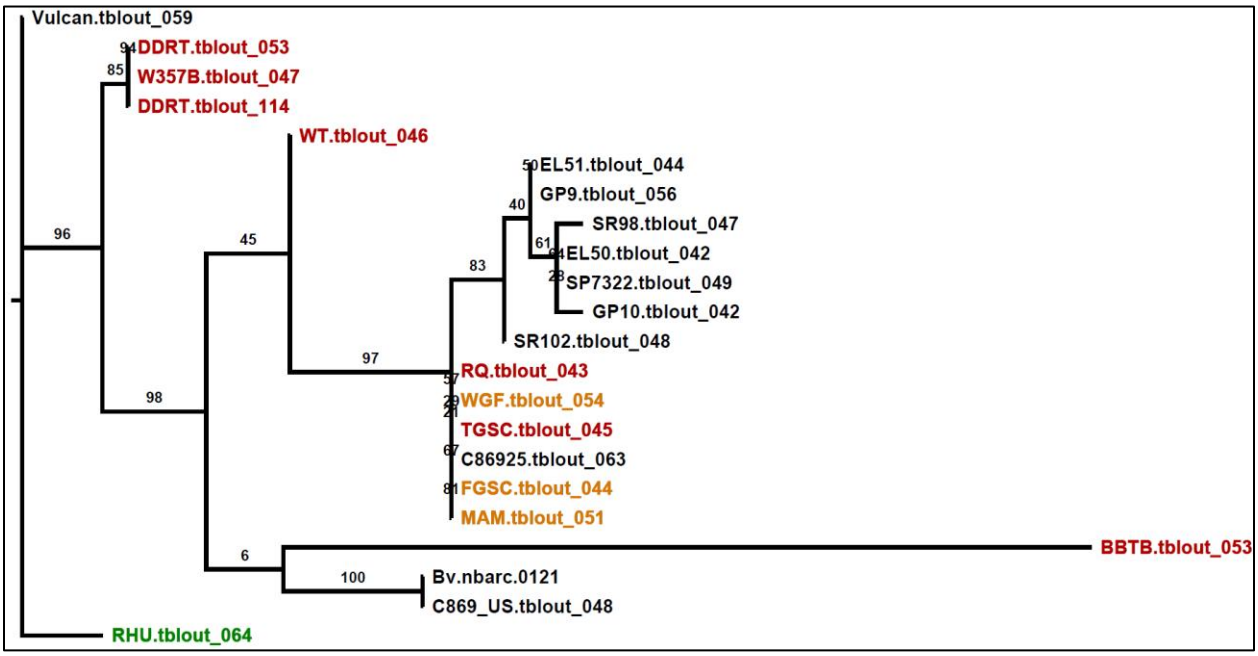

Figure 3.6. Phylogeny of NB-ARC domains mapping to Bv.nbarc.0121 in the *B. vulgaris* EL10 reference genome. Domains were extracted from 23 *de novo* assemblies of *B. vulgaris.* Populations are colored by crop type: green is chard, red is table beet, gold is fodder beet, and black is sugar beet. Bootstrap scores of 1000 replications are shown at branch points.

CHAPTER FOUR

WHOLE-POPULATION STRUCTURAL VARIANT DETECTION USING POOLED POPULATION SEQUENCING

INTRODUCTION

Genetic variation is the raw material fueling selection, creating phenotypic differences between individuals conditioned by genotype-by-environment interactions (Hammer et al. 2006). Knowledge of genetic variation within and between populations can be leveraged to improve agricultural crops as well as help expand basic understanding of genome biology. Advances in population-specific genotyping technology have accelerated the speed at which breeders can make selections for crop improvement (Elshire et al. 2011; Wijnen & Keurentjes 2014; Huang & Han 2014). The availability and quality of reference genomes across a range of organisms has been increasing at an accelerating rate (Armstrong et al. 2019), providing a basis for comparison of genetic variation within and between whole species. Unfortunately, a single reference genome is incapable of capturing pan-genomic variation, highlighting the need for additional sequence data to more fully characterize population- and species-wide genetic resources.

Single-nucleotide polymorphism (SNP) genotyping has emerged over the past decades as the most common tool to measure genetic variation (Rafalski 2002; Davey et al. 2011; Lipka et al. 2015). Combining short-read shotgun sequencing with reduced-representation library construction allowed the development of genotyping-by-sequencing (GBS), leading to cost-effective SNP genotyping of each individual within a population without the need for a reference genome (Davey et al. 2011). Statistical associations between SNPs and phenotypes could identify SNPs comprising the causal basis of the trait. Alternatively, SNPs could be in linkage

disequilibrium with the loci underlying the phenotype, providing rough genetic associations without illuminating the molecular genetic cause.

In addition to SNPs, other types of genetic variation can influence phenotypic expression. One example of non-SNP-based genetic variation is the presence of structural variation, including sequence insertions, deletions, translocations, and inversions (Wu et al. 2017; Krumm et al. 2012; Abyzov et al. 2015; Miles et al. 2016). Insertions and deletions (indels) are often treated as short features in the literature, while longer variants are referred to as structural variants (SVs). In this work I will use the term "indel" to define insertions and deletions shorter than 1,000 bp, and SVs as an inclusive term encompassing all forms of sequence changes longer than 1 kb. Indels can be deduced by whole-genome shotgun sequencing but are not directly detected using current genotyping strategies, while indels and SVs longer than the length of the sequencing reads require dedicated detection pipelines and analysis (Sudmant et al. 2015; Alkan et al. 2011; Kosugi et al. 2019). Re-sequencing projects have attempted to fill gaps in knowledge of structural variation, but to date resequencing diversity panels has been expensive, time-consuming, and focused on a few major animal and plant species (Hamilton & Buell 2012; Salgotra et al. 2014). Because of the additional resources required to identify indels compared to SNPs, many non-model systems are still awaiting descriptions of pan-genomic indel variation. Additional work is needed to understand the extent of indels and SVs in breeding populations, and how they function during population development across multiple generations.

Pan-genomic structural variation is a critical component of adaptation and selection, as revealed by existing SV detection in plants and animals. The first efforts in human genomics illuminated wide-spread structural differences between genome assemblies, which has led to comparisons of ever-increasing numbers of genomic sequences (Iafrate et al. 2004; Khaja et al.

2006; Kidd et al. 2008; Genome of the Netherlands Consortium 2014). More recently, plant resequencing efforts have identified SVs in a number of model plant species such as rice, maize, soybean, poplar, Chlamydomonas, and Arabidopsis (Pinosio et al. 2016; Flowers et al. 2015; Lam et al. 2010; Xu et al. 2012; Cao et al. 2011; Chia et al. 2012). While these studies are major contributions to understanding of pan-genomic variation, they remain missing from the literature of non-model systems.

Structural variation has been associated with a range of biological phenomena. Generation of novel coding sequences in rice was preferentially driven by indels that modified the reading frames of transcribed non-coding sequences (Zhang et al. 2019). Small deletions have been implicated in genome size differences between *A. lyrata* and *A. thaliana* (Hu et al. 2011). Fungal effector genes were found to associate with regions of genomic instability prone to acquisition of novel SVs (Plissonneau et al. 2017). Copy-number variation, a form of SV, was enriched in stress-responsive gene clusters in diploid potato (Hardigan et al. 2015; Pham et al. 2017, 2019), and mapping maize presence/absence variants revealed an enrichment in significant GWAS associations compared to SNPs (Lu et al. 2015). Indels fewer than 20 bp were implicated in 7% and 17% of human mutations causing heritable diseases, respectively (Ball et al. 2005). Taken together, indels and SVs are emerging as important components of the genetic underpinnings of phenotypic variation.

My work focuses on the crop species *Beta vulgaris*, which encompasses the root crops sugar beet, table beet, and fodder beet, as well as the leaf vegetable chard. *B. vulgaris* is a predominantly outcrossing species, 2n = 18, with the population generally representing the unit of genetic improvement (McGrath & Panella 2018). Cultivar development is based on increasing the frequency of desirable alleles in populations rather than selecting upon individual plant

genotypes (McGrath & Panella 2018). An open question concerns the number of novel variants generated during each cycle of population development, and to what extent novel indels and SVs contribute to genetic diversity. Defining the rate of indel generation within populations could expand our understanding of genome biology, adaptation, and the targets of selection.

In this paper I analyzed the genetic variation of *B. vulgaris* found in short-read sequence data of pooled populations. For each of 71 populations, 25 individuals were pooled into a single library. These libraries were sequenced to produce ~80 reads covering each base pair in the genome. I mapped these short-read data to the EL10 reference genome (McGrath et al., in prep) and identified total indels and SVs using the targeted reassembly pipeline SvABA. This revealed 4,999,533 indels across the 71 populations. The genotypic data contained in the indels were used to cluster populations, which clearly distinguished each of the four crop types and identified sub-groupings of table beet and chard. The density of short and long indels was calculated across chromosomes to compare the spatial arrangement of these features. I observed enrichment of longer indels towards the end of chromosome arms, while 1-bp indels were more evenly distributed across chromosomes. Pan-genomic hotspots of indel fluctuation, as well as population-specific sites of variation, each contribute to genome variation across *B. vulgaris* and illustrate the fluid nature of structural variation between populations.

RESULTS

Sequencing whole-population samples and mapping to the EL10 reference genome

My first goal was to efficiently acquire the genomic sequence of whole populations using short-read shotgun sequencing. Leaf tissue was collected from 25 two-week-old seedlings of each of 71 populations and pooled prior to DNA isolation (Table 4.1). Each pooled population sample was sequenced to a target depth of 80x using the Illumina HiSeq 2500 or NovaSeq

platforms. The final number of reads mapped from each population ranged from 278 to 862 million, with median depth of coverage between 49x and 165x (Table 4.2). The EL10 reference genome contains nine chromosome-size pseudomolecules representing ~95% of the total assembly. These sequences were used for all subsequent analysis.

Detecting structural variation

I used the targeted reassembly program SvABA (Wala et al. 2018) to detect structural variants based on the read mapping data for each population. Two classes of variant were detected, indels with lengths less than 1 kb and SVs greater than 1 kb (Table 4.2). Indels were determined by mapping a reassembled contig to a single site, while SVs required splitting the assembled contig between two mapping sites (see Methods). Allele frequency was calculated for each variant within each population using the number of split reads that mapped to the variant breakpoint divided by the total number of split reads plus non-variant reads. By default, features with low alternate allele frequency were assigned the homozygous reference genotype and flagged as false positives due to the expected 2n nature of the sample. However, in the pooled samples these could be true features with low minor allele frequencies. I searched the raw data for sites that were flagged as homozygous reference but had minor allele frequencies between 0.05 and 0.2. These sites were added to the analysis.

I implemented a system to distinguish indels that shared the same genomic coordinate yet differed in their variant sequence. Each indel was assigned a name that incorporated the chromosome position as well as the complete nucleotide sequence of the variant (e.g. two 10-bp insertions at the same site were counted as distinct if their insert sequences were different). Given these criteria, there were 2,636,112 insertions and 2,359,331 deletions (4,995,443 total) in the quality-filtered set, which is significantly distorted in favor of insertions (52.77% insertions,

47.23% deletions, exact binomial test, $p < 2.2e\text{-}16$). This set of features was termed sequence-preserved indels (SPI) to highlight the maintenance of unique sequence identities. Because the relative importance of insertion sequence versus insertion position was unknown, I also calculated the number of unique insertion positions regardless of the identity of the inserted sequence, called position-preserved indels (PPI). This reduced the calculated number of insertions to 1,988,677, indicating that ~24.5% of insertions were at the same genomic coordinate as another feature but contained variable insert sequences. By this metric, the number of genomic positions with deletions was significantly higher than the positions with insertions (52.46% deletion, 47.54% insertion, $p = 2.2e^{-16}$, exact binomial test). Subsequent analysis used the SPI set to maintain the maximum information obtained from SvABA targeted reassembly.

Distribution of indels across populations

I calculated the frequency of SPI (sizes less than 1kb) across populations and determined how often a given feature was found in more than one population. The number of SPI per population ranged from 432,544 to 1,281,298 (median 997,661 +/- 181,345), while the number of SVs (1kb or greater) ranged from 7,439 to 24,286 (median 20,947 +/- 7,548) (Table 4.2). Approximately 29% of SPI were found in only one population, while 50% of SPI were identified in five or more populations (Figure 4.1). This indicated that indel detection was reproducible between populations despite an increase in expected heterozygosity and confounding genetic variation.

The C869_US sample had a 30-fold reduction in SPI vs the pooled population mean (33,297 vs 982,607, respectively) and an 18-fold reduction in number of SVs (1,116 vs 20,811). C869_US reads were derived from the same plant that was used to generate the EL10 reference genome rather than pooled population reads. Reduction in numbers of indels and SVs is

consistent with shared similarity between C869_US and the reference genome, as well as decreased allelic diversity of the single C869_US plant compared to pooled whole-population samples.

Whole population indel genotypes resolved genetic relationships between populations

Indels should be able to act as molecular markers that illuminate the genetic variation within and between populations (Gabur et al. 2019). Populations were clustered using SPI to test whether populations could be differentiated into crop type groups. The C869_US single plant sample was fundamentally different than the pooled samples due to reduced heterozygosity and increased genetic relatedness to the reference genome. Therefore, this sample was omitted from further analysis, resulting in 71 populations for cluster analysis. I chose to use allele frequency rather than presence/absence for each SPI. Determining allele frequency of SPIs was accomplished using a custom formula rather than the default SvABA method (see Methods). The results of hierarchical clustering indicated up to six partitions existed within the populations (Figure 4.2). To further understand the genetic relationships between populations and crop types, I applied a k-means algorithm to resolve additional cluster relationships.

I began k-means clustering of populations based on SPI data. My initial cluster number was set to four based on the historical development of four crop types included in my sequencing data, which was supported by the hierarchical results. The analysis clearly separated sugar beet, fodder beet, table beet, and chard into distinct clusters (Figure 4.3). Crop-type-specific variation would have become fixed in these lineages and maintained by breeding for specific end-use qualities. Variation within chard and table beets would also become fixed during cultivar development and could be sufficient to create the subdivisions seen in my clustering analysis. The sugar beet populations are expected to be more heterozygous, and thus genetic diversity

between populations should have lower allele frequencies which are insufficient to create subdivisions in this crop type.

Distribution of indels by length

I examined the size distribution of SPI to see if there were trends related to chromosome position or spread among populations (Figure 4.4). Shorter indels were more common than longer indels, with 1-bp deletions or insertions accounting for 39% of total features (1,961,495 of 4,995,443). The number of detected SPI was inversely correlated with length, with 99% of indels shorter than 100 bp. To better understand the relationship between SPI size and distribution across populations, the data was partitioned according to two metrics: 1 bp versus longer, and population-specific versus those found in multiple populations. SPI of 2 bp or greater were significantly more likely to be population-specific than SPI of 1 bp (37% vs 24%, respectively, Pearson's Chi-square test, $p < 3e\text{-}16$).

Indel densities on chromosomes are correlated with length and population distribution

The next question was whether SPI density varied according to chromosome position. Predicted chromosomes were divided into 100kb sliding windows with 50kb overlap, and the number of indels of each type were calculated for each window (Figures 4.5-4.13). Absolute indel counts were normalized to the chromosome-wide average and used to calculate the relative density of SPI per window. Therefore, each window received a density score for each of four subcategories of SPI: 1 bp, 2 bp or longer, population-specific, and shared between 2 or more populations. SPI shared between populations had significantly higher variance of density scores versus population-specific SPIs (0.146 vs 0.089, respectively, $p < 2.2e\text{-}16$, F-test of two variances), indicating that SPI shared between populations were unevenly distributed in windows compared to population-specific variation. Chromosome regions with SPI density at least three

standard deviations above the mean are shown on the chromosome plots as clusters. There was no clear pattern of correlation between clusters of short SPI, long SPI, or gene density. Variation shared between populations was localized towards telomeres, with peaks visible on every chromosome (Figures 4.5-4.13). Interior chromosome regions were characterized by lower variance, with increased density of 1 bp indels and reduction of indels shared between populations. Windows of reduced variation in all SPI categories were seen on each chromosome, such as Chromosome 3 at 47 MB, Chromosome 4 at 40 MB, and Chromosome 9 at 17 MB (Figures 4.7, 4.8, and 4.13, respectively). Total loss of variation was observed at 18 MB on Chromosome 2 (Figure 4.6). Lower variance is consistent with the hypothetical location of centromeres, which could result in reduced recombination rates and therefore reduced structural variability. This is especially notable on Chromosome 9, where a known inversion in the reference assembly places a telomere in the interior of the chromosome at 35 MB (Figure 4.13) (McGrath et al., in prep).

SPI greater than 1 bp showed increased variance in their density within windows when compared to 1 bp (0.38 vs 0.30 respectively, $p < 2.2e-16$, F-test of two variances). This difference in variance indicated 1 bp SPI were more evenly distributed across chromosomes, while those 2 bp or greater exhibited differing density at different chromosome positions. I subset the data to focus on the size range covering 99% of features (-100 bp to 100 bp) and plotted SPI lengths for each population (Figure 4.14). Shorter SPI were more abundant than longer SPI, which was a consistent trend across all populations. The C869_US single plant showed a similar distribution of SPI sizes as other populations despite the reduction in absolute numbers.

Although there was a broad correlation between SPI length and number of SPI detected, there was some enrichment of SPI at specific sizes (Figure 4.14). As reported earlier in this paper, SPI were more likely to be to be population-specific than shared between populations. The next question was whether short, population-specific SPI had some unifying characteristics that were contributing to their abundance. To answer this question, the short, population-specific SPI were examined in more detail.

I isolated all the SPI between -20 and 20 bp and asked what their distributions were across populations. Different sizes of indels had different distributions, most notably the proportion of population-specific SPI changed depending on length (Figure 4.15). Isolating the SPI found in only one population revealed differences in the proportion of insertions (25-36%) versus deletions (15-20%) (Figure 4.16). This indicated deletions were more likely to be shared between two or more populations compared to sequence-preserved insertions. Closer inspection revealed that deletions and insertions had a significant reduction in the number of seven or eight bp features that were population-specific, meaning some 7 and 8 bp SPIs were more broadly distributed among populations than other short SPIs (Figure 4.16).

Identification of enriched motifs in 7 and 8 bp SPI

The observation that 7 and 8 bp SPI were more likely to be shared among populations compared to other short SPI suggested some of these sequences could be biologically relevant. I isolated the sequences of all 326,175 SPIs with lengths 7 or 8 bp and searched for enriched motifs using the DREME module of MEME software suite (Bailey 2011). I identified two motif models, HYAYAA and TYAYRA, that were statistically enriched versus the null distribution (e-values < 1e-50, Figure 4.17). The SPI containing HYAYAA and TYAYRA motifs were shared among more populations than the null 7 and 8 bp sequences ($p < 0.004$, Mann Whitney Wilcox

test, Figure 4.18). The two motifs were scanned versus JASPAR (Fornes et al. 2019) and Arabidopsis DAP-seq databases of transcription factor binding sites (Bartlett et al. 2017) but did not generate statistically significant matches (data not shown).

DISCUSSION

Characterization of indels in the *B. vulgaris* pan-genome supports the hypothesis that indel length, chromosome position, and interpopulation distribution are related attributes of structural variation. Single bp SPIs were distributed more evenly across chromosomes, while longer SPIs were relatively enriched towards telomeres. Population-specific SPIs were spread more evenly across chromosomes while variation shared between populations occurred in more dense windows on chromosome arms. This could be the result of increased recombination rate towards the ends of chromosome arms (Garcia-Diaz & Kunkel 2006). If recombination is enriched at certain genomic coordinates, and recombination is a source of indels greater than one base pair, then meiotic recombination over evolutionary time might create hotspots of indel variation. If this variation was biologically relevant, then selection could act on these sites, preferentially spreading functional alleles to subsequent populations (Halldorsson et al. 2019). Conversely, increased repetitive sequences in centromeres could lead to DNA replication-based errors (Carvalho & Lupski 2016). Strand slippage by DNA polymerase due to repetitive sequence or homopolymer runs would lead to short indels, with preference for deletions over insertions (McCulloch & Kunkel 2008; Tran et al. 1997). This is the same pattern seen in the interiors of chromosomes shown in the current report, including increased density of 1 bp indels and reduced density of genes (Figures 4.5-4.13).

Allele frequencies of the complete set of *B. vulgaris* SPIs were sufficient to resolve the relationships between 71 diverse breeding populations representing combinations of four crop

types. This suggests the reported SPIs represent real genetic differences between the populations that reflect their history of breeding and cultivation. We expect that the table populations are more inbred than the sugar and fodder populations, owing to their history of selection for specific consumer preferences and propagation as named cultivars (Paul Galewski and Mitch McGrath, personal communication). Selection of differentiating features within those groups could have led to fixation for alternate alleles, allowing fine-grained separation of within-group differences. This contrasts the heterozygous sugar beet populations undergoing selection for improved agronomic traits (McGrath & Panella 2018). The background genetics should be more muddled in the sugar populations and new variation often is incorporated to maintain genetic diversity (Panella & Lewellen 2007), which hinders the resolution of between-population differences. Additional experiments could be used to test the limits of population discrimination, provide some indication of diversity between subpopulations, and help determine whether phenotypic traits of populations are derived from shared or novel variation.

SPIs over 1 bp appeared to correlate with SPIs shared between populations. This is interesting considering the higher chance for genetic drift at sequence-preserved insertions, which should theoretically lead to higher numbers of unique insertions not shared between populations. One explanation for the unexpected increase in shared SPI is that they have recent origins in shared ancestors of the populations observed in this report. It also is possible that SPI greater than 1 bp have an outsized effect on phenotypic variation, as seen in human disease genetics (Ball et al. 2005). This could lead to their preservation across the genome. Alternatively, the increased numbers of 1 bp indels vs those 2 bp or greater could create a more even distribution of the abundant 1 bp features, effectively masking any enrichment of biologically relevant sequences in this subset. Regardless, the localization of shared variation provides a focal

point to ask what functional differences exist in those regions, why those features are shared between populations, and how recombination and replication-based mechanisms contribute to existing structural variation.

The current project was conceived in response to numerous hypotheses about R gene evolution, including the role of DNA replication-based mechanisms, recombination-based processes, and "hotspots" of recombination in generating novel diversity in the form of structural variation (McDowell & Simon 2006; Mondragon-Palomino & Gaut 2005; Christie et al. 2016; Carvalho & Lupski 2016). Structural variants, including small indels, have long been known to play a role in functional variation in humans (Ball et al. 2005; Abyzov et al. 2015), microorganisms (Miles et al. 2016), and plants (Yan et al. 2010; Schnable et al. 2009; Chia et al. 2012). The initial goal was to detect NLR genes with indels and copy-number variation, including novel sequences absent from the reference genome. In this way I hoped to integrate knowledge of genetic diversity with R gene evolution at the sequence level. It quickly became apparent that cataloging variation in the pooled population sequences was necessary before targeted analysis of R genes would be possible. Therefore, the main achievement of my current work is to advance the ability to identify structural variation in a large pool of short sequencing reads and differentiate features within and between populations. Such variant detection has the benefit of accessing sequences absent from a reference genome via targeted reassembly of discordant reads localized to genomic windows.

Non-uniform distributions of indels could indicate biological function (Schatz et al. 2014; Conrad & Hurles 2007). There were two enriched motifs in the 7 and 8 bp SPI sequences, HYAYAA and TYAYRA, that were more widely distributed among populations than the other 7 and 8 bp sequences (Figure 4.18). It is notable that any sequence was discernable against a

background likely dominated by relaxed selection and genetic drift (Zmienko et al. 2016; Zhang et al. 2014; Hardigan et al. 2015). Some of the enriched motifs could have been generated randomly by drift. However, the difference in population distributions between enriched motifs and the other 7 and 8 bp SPIs suggest that some mechanism is propagating or maintaining the enriched HYAYAA and TYAYRA motifs across populations. These motifs were scanned against known transcription factor binding sites but no putative targets were found. Transposon families such as *copia*, L1, R1, IS, Alu, P element, Tn10, and LINE could have short target site duplications in the 3-8 bp range (Flasch et al. 2019; Kojima & Fujiwara 2003; Tatout et al. 1998; Linheiro & Bergman 2012; Dewannieux et al. 2003; Liao et al. 2012; Dunsmuir et al. 1980; Halling & Kleckner 1982). Alternatively, transcription factor binding sites or other regulatory elements have variable frequency across populations. Localization of the enriched motifs and correlation with other genomic features could help shed light on their origins and biological functions, if any. Further investigation of enriched sequences at other sizes could provide insight into transposon activity, regulatory elements, and genome size variation in non-model plant systems.

One key question is how effective indel detection was in whole-population samples. Heterozygosity has historically created difficulties with *de novo* assembly from short reads (Hirsch & Buell 2013), and it stands to reason that the data presented in the current report contain a non-zero amount of artifacts based on incorrect read mapping or assembly. To mitigate errors, SvABA includes numerous error-correction and validation steps to reduce false-positive variant calls (Wala et al. 2018). One benefit of the SvABA program is that it reduces sequence complexity by assembling reads in windows, rather than the full read set. This allows improved assembly of alleles without confounding homologous sequences from other parts of the genome.

Multiple indels mapped to the same base pair position but had variable lengths, and some indels were found with the same length but different insert sequences. This indicates that SvABA was able to distinguish highly similar sequences and assemble them into contigs supporting multiple alleles, which would be difficult with traditional *de novo* assembly (Flowers et al. 2015; Alkan et al. 2011; Bickhart et al. 2017).

The data presented in the current report supports the possibility that population-wide inference of structural variation can be achieved using pooled population sequencing. A logical next step to further investigate the rate of false positive indel calls would be to validate such alleles by Sanger sequencing. PCR could be used to validate amplicon sizes using agarose gels or capillary gel electrophoresis, depending on indel length. Such validation could lead to additional applications for indel detection in plant breeding, such as improved genetic markers, as well as improved basic understanding of chromosome structure and function.

METHODS

Population sampling and Illumina sequencing

Seventy-one populations of *B. vulgaris* were germinated in plastic containers (24" W x 16 " D x 12" H) in the greenhouse and harvested at the two-leaf seedling stage, approximately two weeks after germination. For each population, leaf tissue was collected from 25 individuals and placed into a 50 mL conical polypropylene centrifuge tube (VWR, Radnor, PA) to create population-specific pools of tissue. These tissues were lyophilized and ground to a powder. For each population, DNA was isolated from 20 mg of lyophilized tissue using the NucleoSpin Plant II kit (Macherey-Nagel, Duren, Gernamy). Libraries were prepared by the Michigan State Research Technology Support Facility (East Lansing, MI, USA) using Illumina TruSeq kits (Illumina, San Diegoi, CA, USA). Libraries were sequenced to 80x depth using HiSeq 2500 2 x

125bp paired-end chemistry.

Read QC and mapping

Illumina adapters were trimmed from the raw reads using Trimmomatic v.0.36 (Bolger et al. 2014) with parameters ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36. For each population, reads with both pairs intact were interleaved using BBMAP's reformat.sh function. Reads were split into smaller files for parallel mapping using the fastq-splitter.pl script from Kirill Kryukov (http://kirill-kryukov.com/study/tools/fastq-splitter/). Reads were mapped to the EL10 reference genome v1.2 (Funk et al. 2018) using BBMap v.38.22, with parameters local=t interleaved=t k=13. Mapping coverage statistics per-chromosome and in 100kb bins were calculated with the BBMap pileup.sh script, with parameters binsize=100000 32bit=t. Output BAM files for each population were sorted, merged, and indexed with samtools v.1.9 (Li et al. 2009) prior to indel detection with SvABA.

Variant detection with SvABA

The sorted, indexed BAM files were analyzed individually with SvABA v. 134 (Wala et al. 2018)using the --germline parameter (equivalent to -I, -L 5, NM >= 3) to produce VCF files of filtered and unfiltered indels and structural variants (SVs). Briefly, SvABA generates 25kb sliding windows across the genome and identifies reads that map discordantly to the reference sequence (i.e. soft-clipped ends, split mapping sites, and incorrect distances between reads compared to the insert size). For each 25kb window, SvABA performs *de novo* assembly of the discordant reads together with the full set of unmapped reads. Two quality-control steps are performed: first, raw reads are mapped back to d*e novo* contigs generating LOD scores, and second the contigs themselves are mapped back to the reference. High-scoring variants are

collected into a filtered variant VCF file. SvABA differentiates indels and structural variants based on how the *de novo* supporting contigaligns to the reference genome. Any contig that maps to the reference at a single site is considered an indel, while contigs with split mapping sites are termed structural variants (SVs). The filtered VCFs were parsed to remove inter-chromosomal variants and then converted to BED format using a custom python script.

Allele frequency of indels

Allele frequencies between 0.05 and 0.2 are flagged as NONVAR, or non-variant, by default SvABA parameters and filtered out as a form of quality control. This is because SvABA was designed to call diploid samples, where allele frequencies should trend toward 0, 0.5, or 1. Therefore, any feature below 0.2 allele frequency is assumed to be technical error by the software. The pooled population samples could include up to 50 allele measurements (2n x 25 = 50) with legitimate results between 0.05 and 0.2. Because of this, the NONVAR features were retrieved from the unfiltered indel files and added to the analysis. Allele frequency was calculated from the resulting VCF data using a custom formula: reads that mapped across a break point (spanning reads, SR) were divided by paired reads that bracketed the interval (depth of coverage, DP). This more conservative adaptation was implemented to control for variation in paired-end insert size as well as indel length. The allele frequency of some variants was therefore lower than the default 5% threshold for SvABA, but they were retained in the analysis due to the already conservative nature of the custom allele frequency calculation.

Cluster analysis

Clusters of genes, indels and SVs were detected using Clusterscan (Volpe et al. 2018) with the *clustermean* script, window size 200kb, step size 100kb, k=3 (minimum threshold of three standard deviations to start a cluster) and e=2 (minimum threshold of two standard

deviations to extend a cluster). Computation was performed in a Linux CentOS 7 environment on the Michigan State High-Performance Computing Cluster (Michigan State University, East Lansing, MI).

Motif enrichment analysis and similarity scanning

The DREME module of the MEME software suite (Bailey 2011)was used to scan for enriched motifs in 7 and 8 bp indel sequences. The e-value cutoff for motif addition during model generation was 1e-5, while the e-value cutoff for final model inclusion was 1e-50. To generate e-values, DREME randomly shuffles query dinucleotides to create a null dataset of equivalent shape to the query. The top one scoring motif for analysis was used as a query in the Tomtom module of MEME, applied against the JASPAR CORE non-redundant database, JASPAR plant non-redundant database, and the Arabidopsis DAP-seq database (Gupta et al. 2007).

ACKNOWGLEDGEMENTS

Thanks to MSU RTSF for sequencing support, Admera Health for sequencing support, Paul Galewski for productive discussions, Christina Azodi and Beth Johnson in the Shiu lab for discussion of transcription factor binding sites.

APPENDIX

Table 4.1 Germplasm accession numbers for *B. vulgaris* accessions used in analysis.

Sample ID	Accession ID	Identifier	Crop type
08storage	EL-A024967	EL-A024967	sugar beet
5Estorage	EL-A15-00005	EL-A15-00005	sugar beet
BBTB	EL-A15-01112	Bulls Blood Table Beet	table beet
BDL	-	Bionda Di Lyon	Swiss chard
BYE	-	Yellow Eckendorf	fodder beet
C869_US	EL-A015027	C869 cms single plant	sugar beet
C86925	EL-A015027	C869 cms population	sugar beet
Chiog	-	Chioggia	table beet
Crosby	EL-A15-01115	Crosby's Egyptian Table beet	table beet
cylindra	-	Cylindra	table beet
DDRT	EL-A15-01114	Detroit Dark Red table	table beet
EL0204	EL-A012858	PI 632750	sugar beet
EL50_2	EL-A021482	EL50/2	sugar beet
EL51	EL-A12-00030	EL51	sugar beet
EL52	EL-A012200	PI 628274	sugar beet
EL53	EL-A013523	PI 641927	sugar beet
EL54	EL-A021483	PI 654357	sugar beet
EL55	EL-A013698	PI 655304	sugar beet
EL56	EL-A022799	PI 663211	sugar beet
EL57	EL-A022809	PI 663212	sugar beet
EL57-Rbulk	EL-A022809	PI 663212-Rhizoctonia resistant_bulk	sugar beet
EL57-Sbulk	EL-A022809	PI 663212-Rhizoctonia susceptible_bulk	sugar beet
EL58	EL-A022775	PI 664913	sugar beet
EL59	EL-A029768	PI 664914	sugar beet
EL60	EL-A021740	PI 664915	sugar beet
EL61	EL-A029769	PI 664916	sugar beet
EL63	EL-A027007	PI 664918	sugar beet
EL64	EL-A022776	PI 664919	sugar beet
EL65	EL-A027017	PI 664920	sugar beet
EL66	EL-A027143	PI 664921	sugar beet
EW	-	Early Wonder	table beet
FCxELcerc	EL-A16-00016	EL-A16-00016	sugar beet
FE	-	Flat of Egypt	table beet
FGSC	EL-A15-01116	Fordhook Giant	chard (leaf beet)
FK	-	Fuer Kugel	table beet
FP	-	Flamingo Pink	Swiss chard

Table 4.1 (cont'd)

Sample ID	Accession ID	Identifier	
GB	-	Geante Blanche	fodder beet
GP10	EL-A1402164	East Lansing Breeding Population GP10	sugar beet
Gp7&8	EL-A1402159	EL-A1402159	sugar beet
GP9	EL-A1402163	East Lansing Breeding Population GP09	sugar beet
L19	EL-A010101	L19 (PI 590690)	sugar beet
LUC	EL-A011917	Lucellus Chard	chard (leaf beet)
MAM	EL-A011928	Mammoth Red Fodder	fodder beet
MYC	-	Mangel Yellow Cylindrical	fodder beet
NNSGp3	EL-A1402161	EL-A1402161	-
Ocbordo	-	Okragly Ciemnoczerwony (Bordo)	table beet
PS	-	Perpetual Spinach	Swiss chard
RHU	EL-A011929	Rhubarb Swiss Chard	chard (leaf beet)
RQ	EL-A15-01113	Ruby Queen Table Beet	table beet
SF'A'	EL-A029686	EL-A029686	sugar beet
SF'B3'	EL-A12-00002	EL-A12-00002	sugar beet
SP7322	EL-A015030	SP22 (PI 615525) -> SP7322 (EL-A015030)	sugar beet
SR100	EL-A027152	PI 664923	sugar beet
SR101	EL-A024969	PI 664924	sugar beet
SR102	EL-A15-00006	SR102 (PI 675153)	sugar beet
SR80	EL-A012187	PI 607898	sugar beet
SR87	EL-A012148	PI 607899	sugar beet
SR93	EL-A012191	PI 598075	sugar beet
SR94	EL-A012172	PI 598076	sugar beet
SR95	EL-A012168	PI 603947	sugar beet
SR96	EL-A012189	PI 628272	sugar beet
SR97	EL-A012174	PI 628273	sugar beet
SR98_2	EL-A027006	SR98/2 (PI 659754)	sugar beet
SR99	EL-A024983	PI 664922	sugar beet
SST	-	Shiraz Tall Top	table beet
TGSC	EL-A15-01117	Touch Stone Gold Table Beet	table beet
VDT	-	Verde De Taglio	Swiss chard
Vulcan	EL-A1501111	Vulcan Swiss Chard	chard (leaf beet)
W357B	EL-A01406766	Wisconsin Table Beet Breeding Line	table beet
WGF	EL-A027193	Wintergold Fodder	fodder beet
WT	EL-A15-01116	Albino Table Beet	table beet
ZF	-	Zentuar	fodder beet

Table 4.2. Statistics for Illumina shotgun sequencing reads mapped to the *B. vulgaris* EL10 reference genome. Structural variants (SVs) were termed intra-chromosome when both sides of the variant mapped to the same chromosome. Inter-chromosome indicated split reads mapped to two different chromosomes.

Sample	Reads	Mapped reads (%)	Reference bases covered (%)	Median chromosome depth	Indels	Intra-chr SVs	Inter-chr SVs
08storage	379,011,848	95.85	99.96	91.78	1217031	24407	766
5Estorage	404,352,655	96.20	99.91	99.00	924002	19040	622
BBTB	474,227,044	96.29	99.59	86.44	542182	10028	251
BDL	422,236,198	95.34	99.92	91.89	1204030	23574	724
BYE	500,454,416	95.58	99.9	106.33	997661	20191	655
C869_US	862,135,388	98.77	99.90	157.44	774542	15984	407
C86925	385,124,972	96.70	99.95	76.56	33297	1078	38
Chiog	592,096,557	95.72	99.92	124.67	952119	19384	566
Crosby	295,162,179	96.31	99.81	54.00	828502	16839	498
cylindra	382,791,420	95.36	99.88	81.44	866460	17609	516
DDRT	438,787,401	96.31	99.87	80.89	832124	16371	432
EL0204	419,106,476	96.28	99.94	102.67	1200285	24231	750
EL50	358,746,795	96.00	99.78	73.33	709740	14148	483
EL51	379,768,734	96.27	99.91	80.67	859119	17403	539
EL52	388,016,405	96.11	99.92	94.11	981231	19951	653
EL53	454,850,298	96.02	99.93	112.22	1117386	22930	718
EL54_Mfert	361,766,312	95.99	99.91	90.33	1073424	21335	651
EL55	436,362,543	96.27	99.93	110.11	1082019	22240	732
EL56	418,036,257	96.43	99.94	101.11	1025395	21104	695
EL57	405,071,537	96.18	99.94	96.22	1129170	69500	2180
EL57_Rbulk	419,683,600	96.16	99.96	102.22	1105271	22594	676
EL57_Sbulk	443,305,221	95.96	99.96	103.22	1191583	24181	745
EL58	436,121,695	96.30	99.94	107.89	1014702	20797	661
EL59	411,120,257	96.38	99.95	98.67	1094632	21755	736
EL60	539,052,732	96.51	99.96	128.44	1092651	21885	730
EL61	455,519,183	96.41	99.94	108.11	1133693	22738	773
EL63	608,452,097	96.56	99.96	143.33	1209092	24404	824
EL64	447,448,008	96.32	99.94	102.44	1134698	22741	790
EL65	439,923,770	96.35	99.93	99.44	966258	19900	641
EL66	484,777,705	96.39	99.94	115.11	1220103	24800	820
EW	397,680,654	95.05	99.92	82.33	1080541	21419	638
FCxELcerc	457,701,480	96.77	99.8	110.44	523521	11051	364
FE	403,847,044	95.52	99.88	82.56	873607	17709	558
FGSC	368,321,206	95.99	99.94	69.44	1179237	23120	668
FK	396,924,783	95.31	99.85	81.44	863991	17558	516
FP	477,384,690	95.12	99.82	99.11	812932	16526	551

Table 4.2 (cont'd)

Sample	Reads	Mapped reads (%)	Reference bases covered (%)	Median chromosome depth	Indels	Intra-chr SVs	Inter-chr SVs
GB	407,246,658	95.27	99.93	92.56	1035618	20972	745
GP10	460,618,947	96.39	99.95	91.33	985583	19906	653
GP7-8	388,893,984	96.23	99.94	93.89	1060533	21398	695
GP9	789,080,200	96.15	99.99	164.78	1159698	24504	791
L19	726,725,464	96.80	99.94	127.22	850865	17139	547
LUC	579,807,347	95.90	99.95	111.89	1281298	24661	701
MAM	377,793,987	96.40	99.94	72.56	1027276	20380	667
MYC	461,414,216	95.59	99.92	96.56	1067441	21576	702
NNSGp3	410,884,134	96.28	99.94	103.89	1045171	21254	685
Ocbordo	415,105,062	95.47	99.91	88.56	1046539	20959	626
PS	509,165,452	95.50	99.92	107.67	1131384	22223	618
RHU	508,724,580	96.29	99.92	94.33	976160	18643	613
RQ	343,178,133	96.38	99.79	67.00	699626	13711	384
SF'A'	543,782,638	96.39	99.96	126.67	1250137	25643	797
SF'B3'	396,118,015	96.43	99.94	98.33	1086641	21821	684
SP7322	509,227,577	96.11	99.94	104.22	932364	18652	653
SR100	420,547,282	95.81	99.93	103.33	999649	20553	670
SR101	404,335,856	95.65	99.95	98.67	1089581	21878	724
SR102	359,343,504	96.24	99.96	75.78	959661	19519	616
SR80	450,325,001	96.12	99.9	106.78	767659	15866	506
SR87	404,487,388	95.91	99.91	98.78	969827	19902	655
SR93	452,175,105	96.36	99.89	106.89	898768	18513	631
SR94	451,484,416	96.07	99.93	110.44	969758	20019	658
SR95	414,151,935	96.10	99.93	102.22	911182	19036	600
SR96	472,970,372	95.88	99.94	117.33	1037649	21434	731
SR97	397,628,554	95.92	99.92	98.00	898347	18742	594
SR98	401,900,530	96.11	99.92	85.00	877031	17851	513
SR99	443,692,514	95.79	99.93	110.78	993526	20641	662
SST	487,200,853	95.65	99.86	99.78	827163	16916	464
TGSC	278,431,865	96.52	99.51	49.11	575434	11215	324
VDT	398,033,591	95.01	99.95	88.78	1341613	25912	820
Vulcan	510,112,562	96.36	99.89	93.11	930707	18053	607
W357B	381,184,038	96.22	99.30	69.89	432544	7606	234
WGF	517,713,429	96.62	99.90	94.33	972214	19147	634
WT	310,986,817	96.35	99.72	56.11	738611	14382	350
ZF	393,244,039	95.55	99.93	86.11	1124919	22507	742

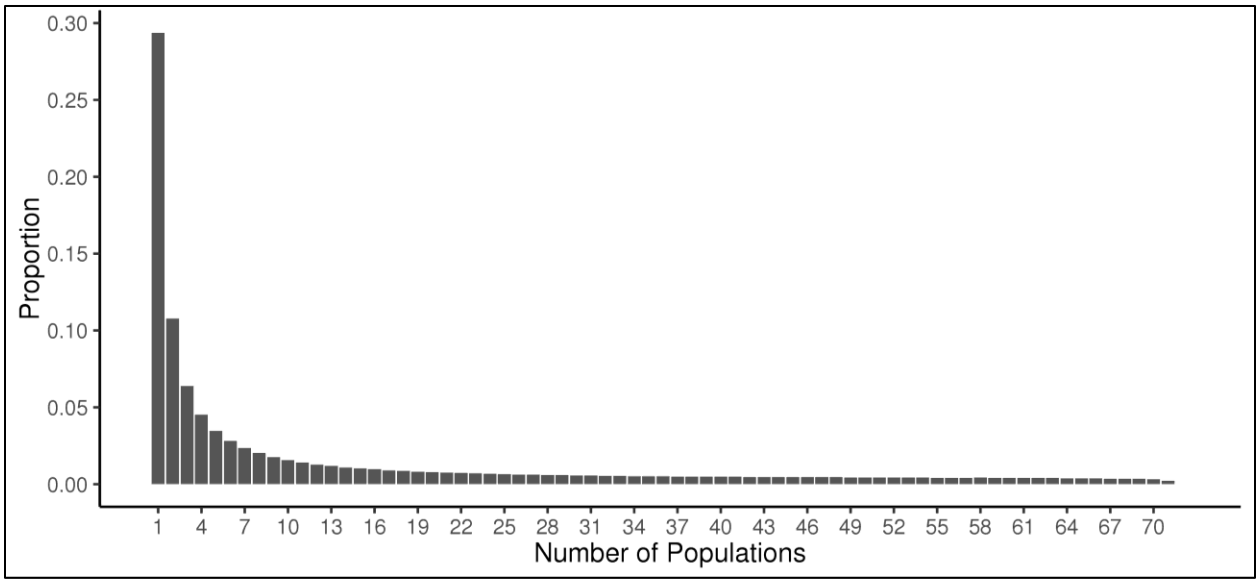

Figure 4.1. Distribution of sequence-preserved indels (SPIs) among populations of *B. vulgaris*. There were 4,995,443 distinct SPIs detected across all populations. The 72 population total includes the C869 single plant sample that generated the reference genome.

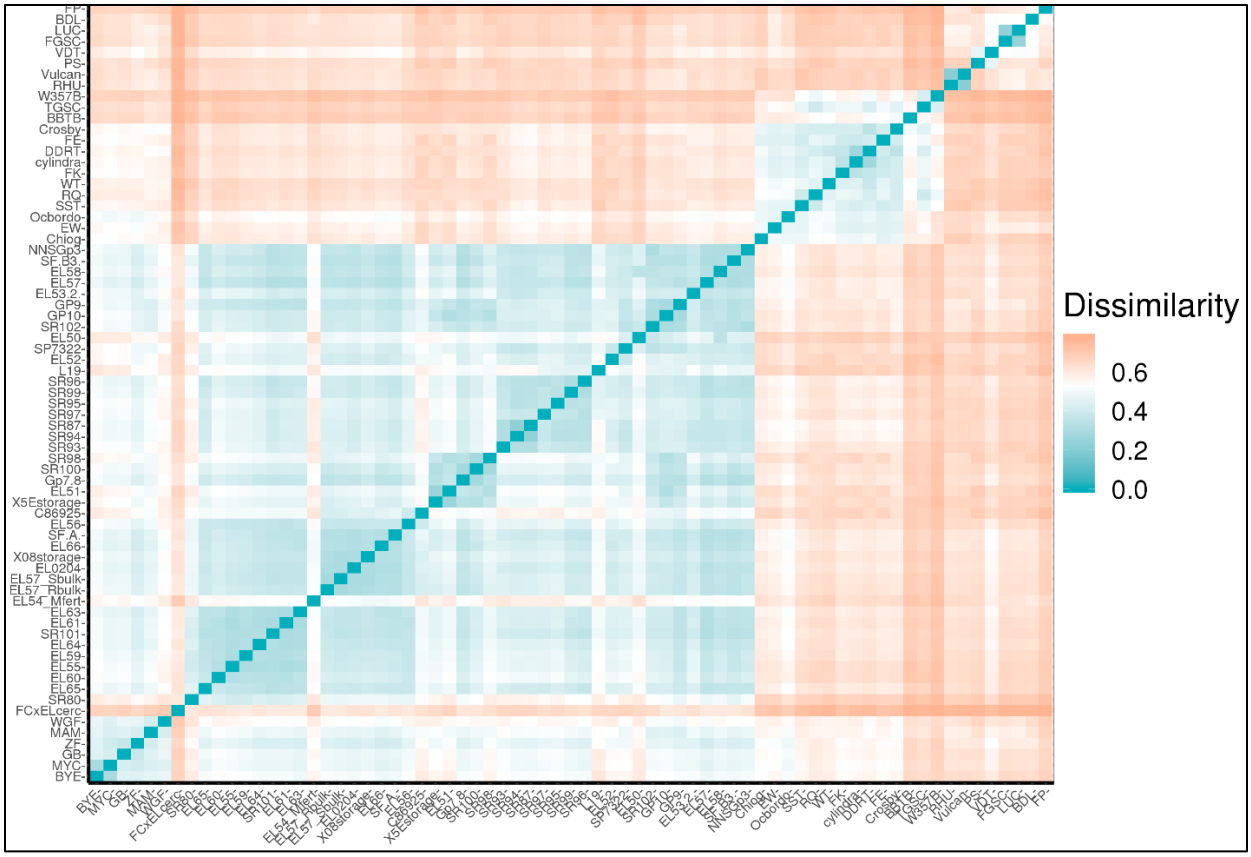

Figure 4.2: Pearson correlation of 71 *B. vulgaris* populations based on sequence-preserved indel (SPI) allele frequency. More similar populations are in blue, while more dissimilar populations are in tan. The C869 single plant sample was excluded from this analysis because of its identity as the reference genome.

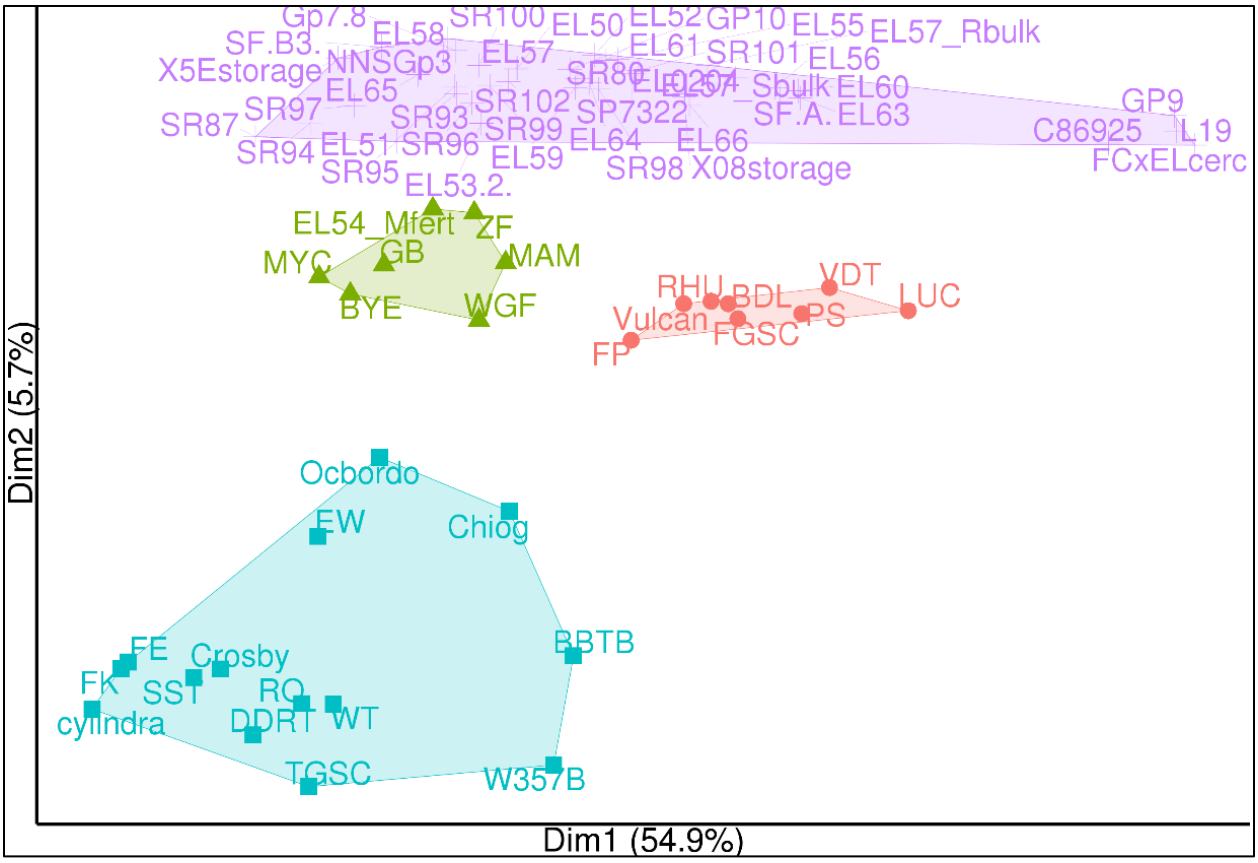

Figure 4.3: K-means clustering of 71 populations of *B. vlugaris* based on sequence-preserved indel (SPI) allele frequency. Cluster input k=4. Colors correlate to the four k-means clusters generated by the clustering algorithm.

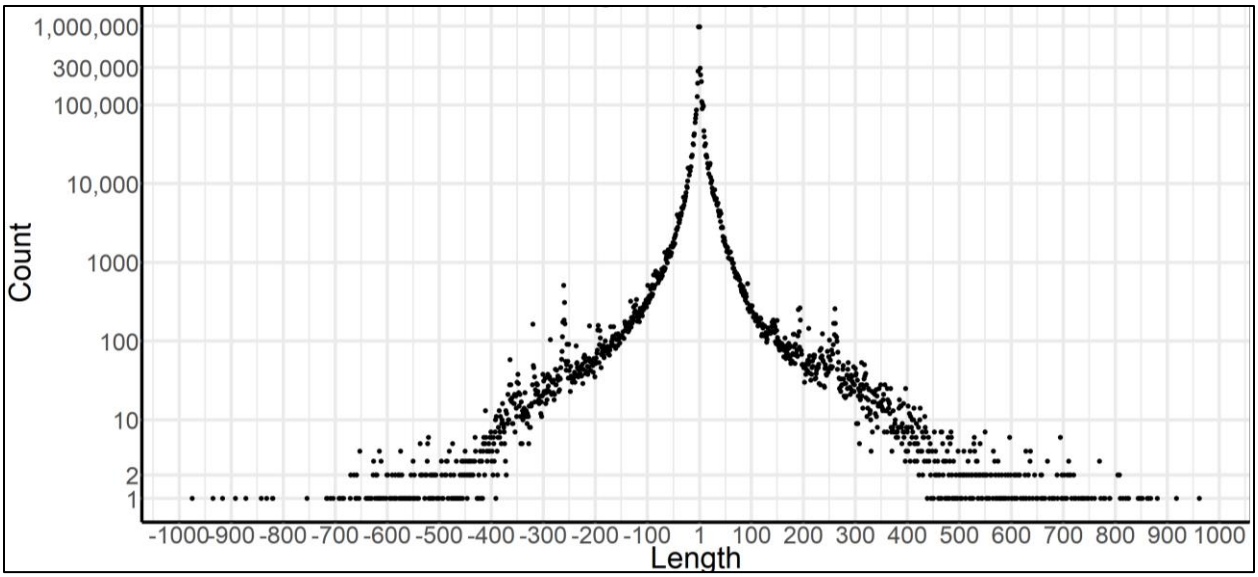

Figure 4.4: The number of unique sequence-preserved indels (SPIs) in the *B. vulgaris* pan-genome according to feature length in base pairs. Deletions are negative, insertions are positive relative to the *B. vulgaris* EL10.1 reference genome.

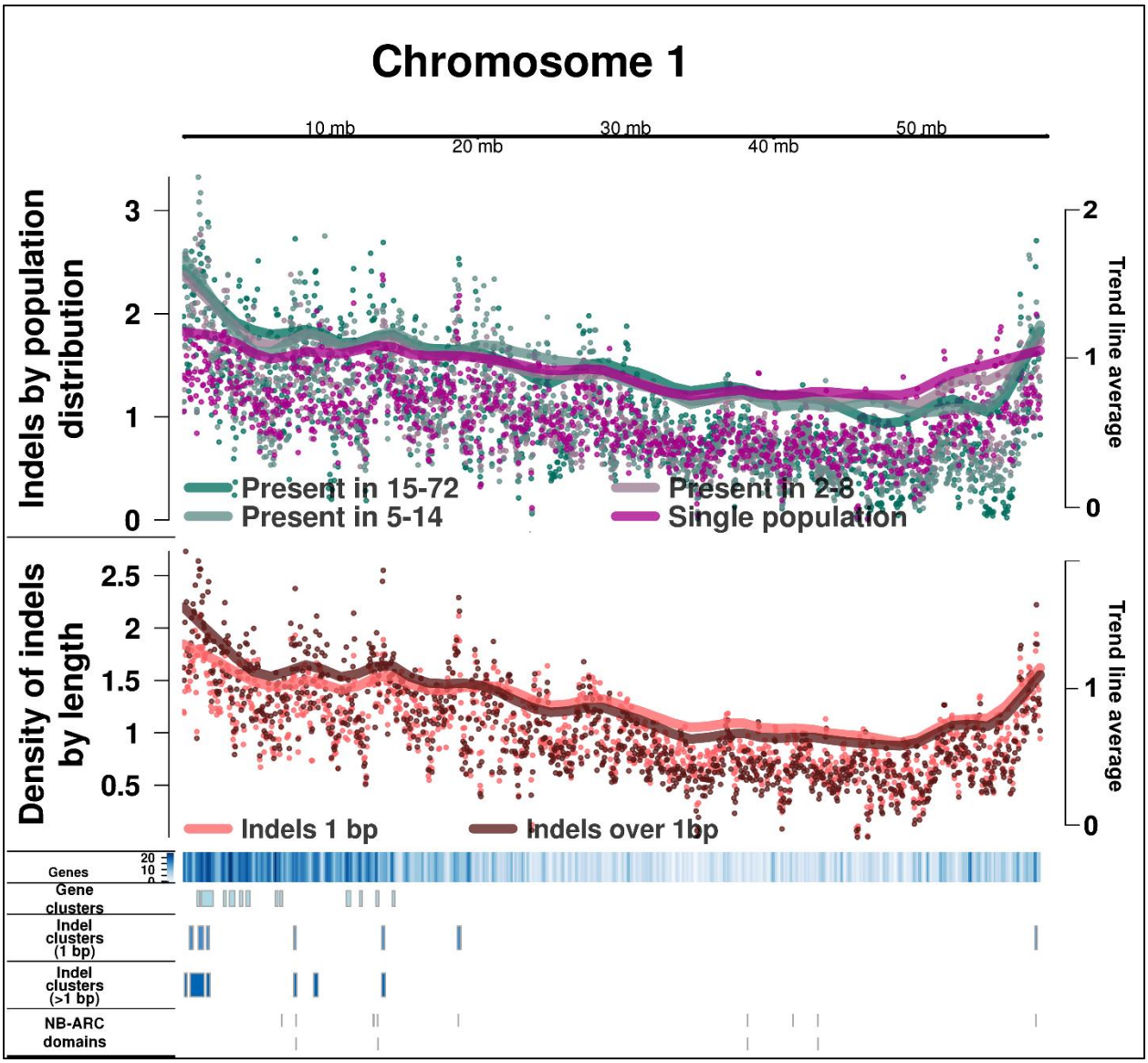

Figure 4.5. *B. vulgaris* chromosome 1 sequence-preserved indels (SPI) by length and population distribution. Points are normalized density of indels in 100kb sliding windows (scale left). Trend lines are calculated with Loess smoothing (scale right). Putative genes were derived from the EL10 annotation. Threshold for clusters of genes and indels was set at three standard deviations above the mean. NB-ARC domains were derived from Funk et al (2018).

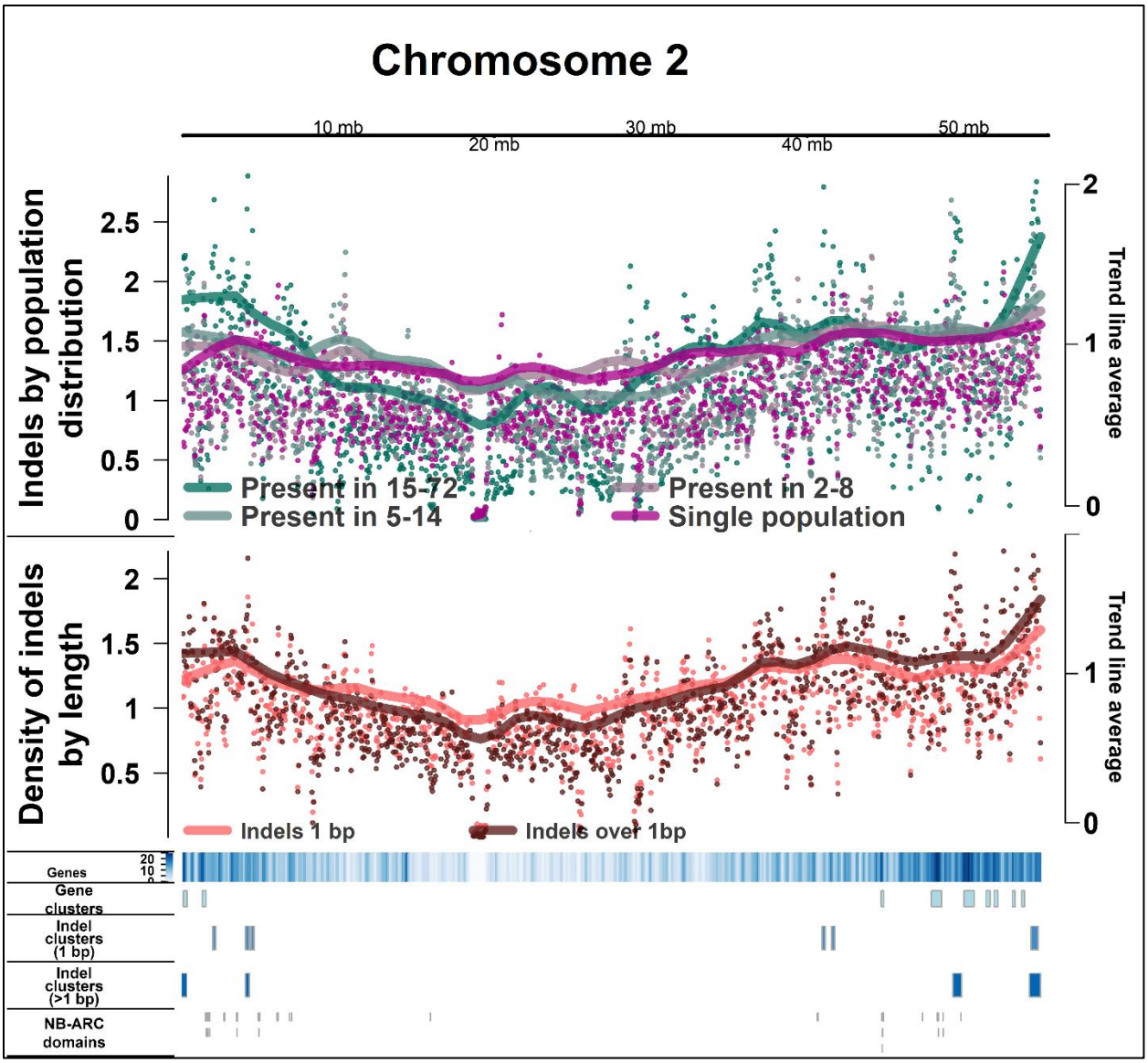

Figure 4.6. *B. vulgaris* chromosome 2 sequence-preserved indels (SPI) by length and population distribution. Points are normalized density of indels in 100kb sliding windows (scale left). Trend lines are calculated with Loess smoothing (scale right). Putative genes were derived from the EL10 annotation. Threshold for clusters of genes and indels was set at three standard deviations above the mean. NB-ARC domains were derived from Funk et al (2018).

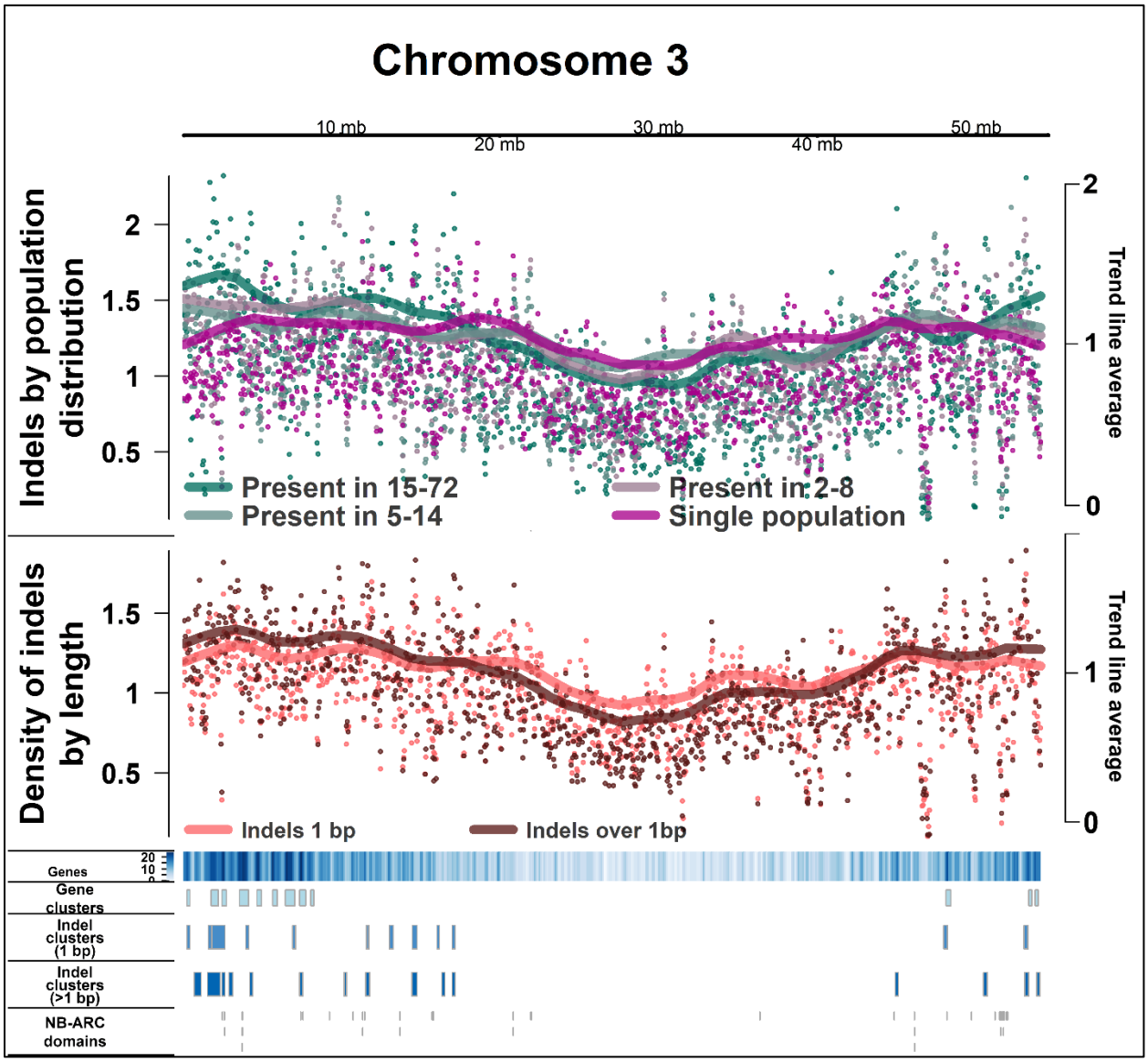

Figure 4.7. *B. vulgaris* chromosome 3 sequence-preserved indels (SPI) by length and population distribution. Points are normalized density of indels in 100kb sliding windows (scale left). Trend lines are calculated with Loess smoothing (scale right). Putative genes were derived from the EL10 annotation. Threshold for clusters of genes and indels was set at three standard deviations above the mean. NB-ARC domains were derived from Funk et al (2018).

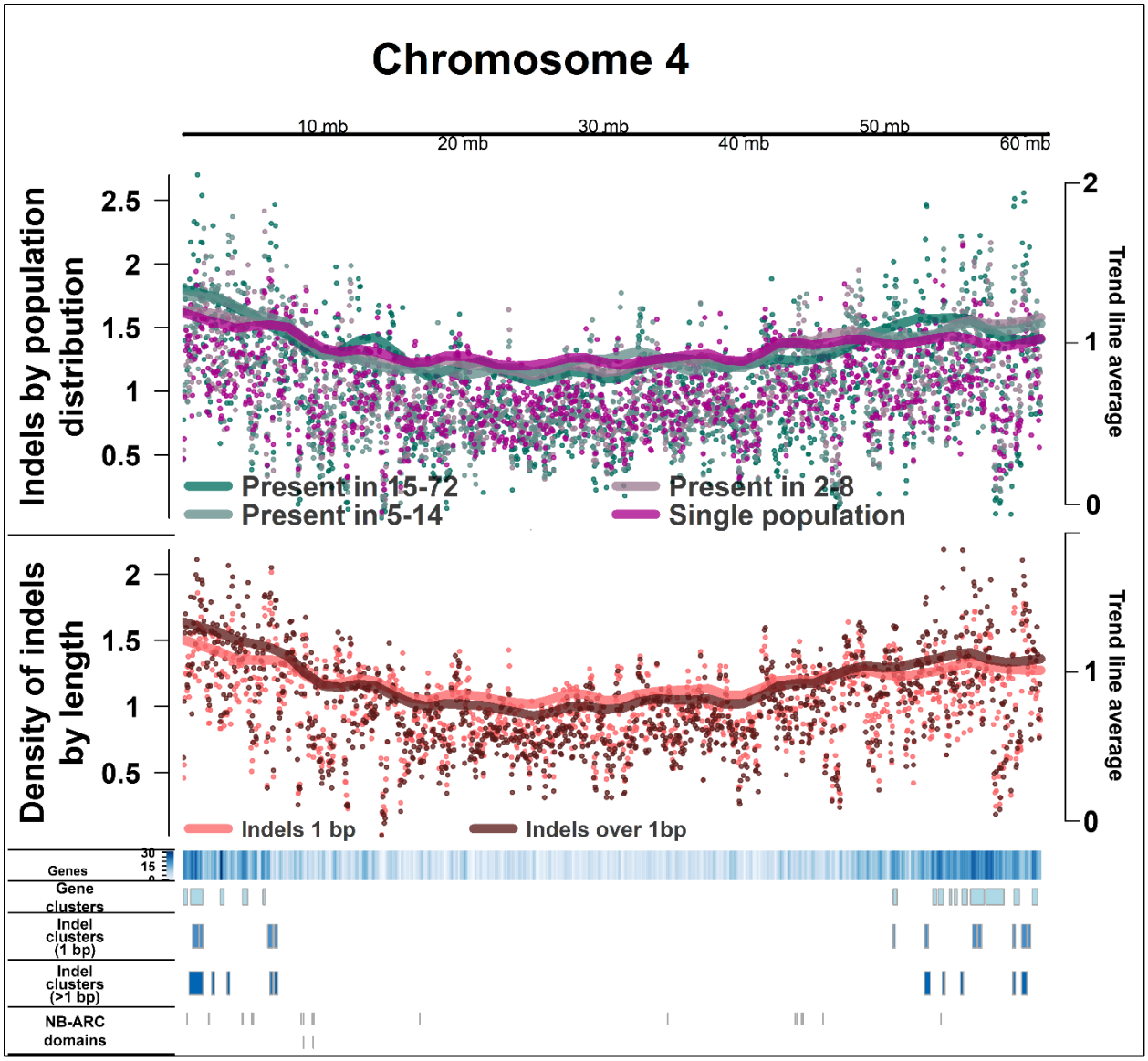

Figure 4.8. *B. vulgaris* chromosome 4 sequence-preserved indels (SPI) by length and population distribution. Points are normalized density of indels in 100kb sliding windows (scale left). Trend lines are calculated with Loess smoothing (scale right). Putative genes were derived from the EL10 annotation. Threshold for clusters of genes and indels was set at three standard deviations above the mean. NB-ARC domains were derived from Funk et al (2018).

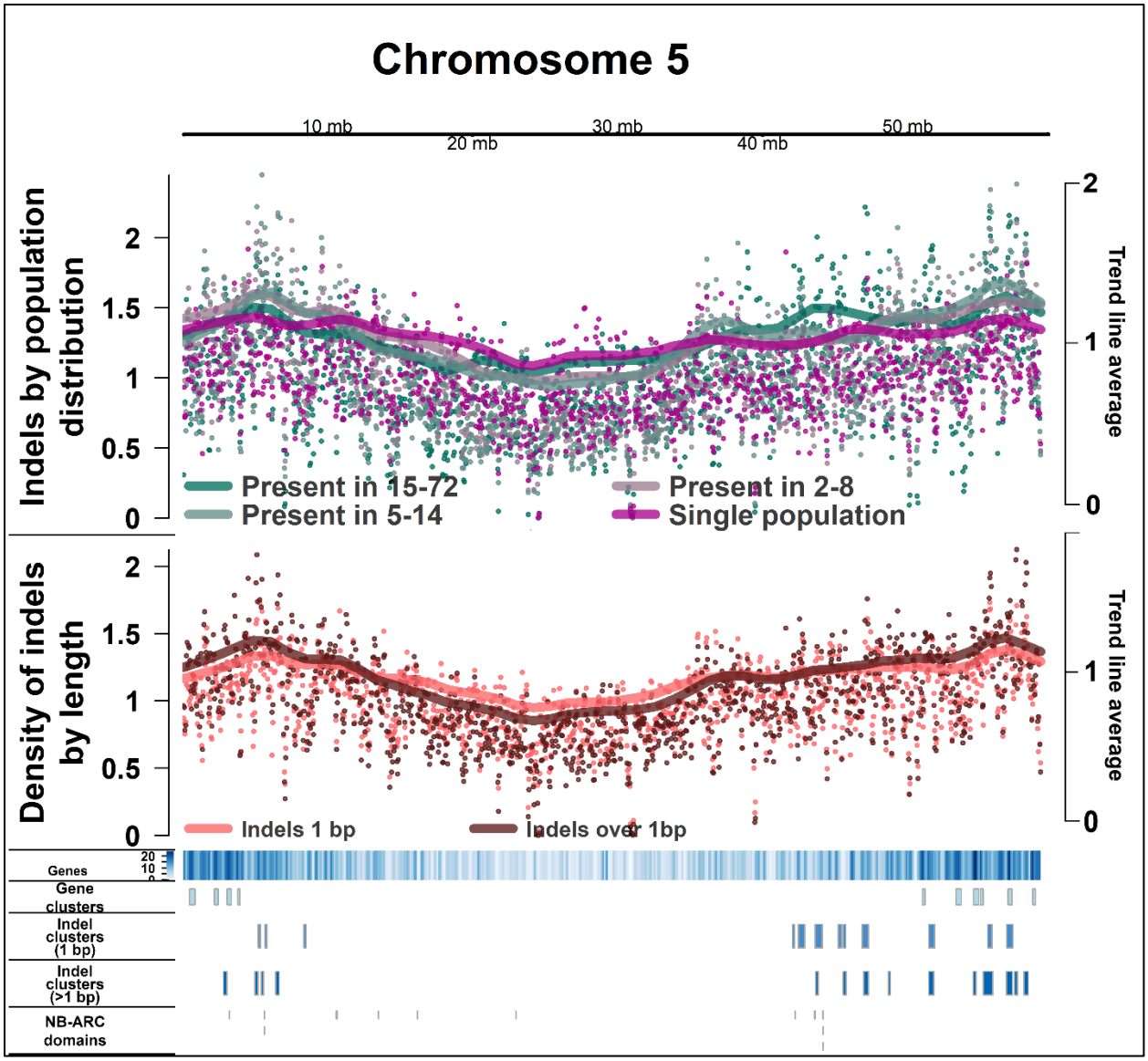

Figure 4.9. *B. vulgaris* chromosome 5 sequence-preserved indels (SPI) by length and population distribution. Points are normalized density of indels in 100kb sliding windows (scale left). Trend lines are calculated with Loess smoothing (scale right). Putative genes were derived from the EL10 annotation. Threshold for clusters of genes and indels was set at three standard deviations above the mean. NB-ARC domains were derived from Funk et al (2018).

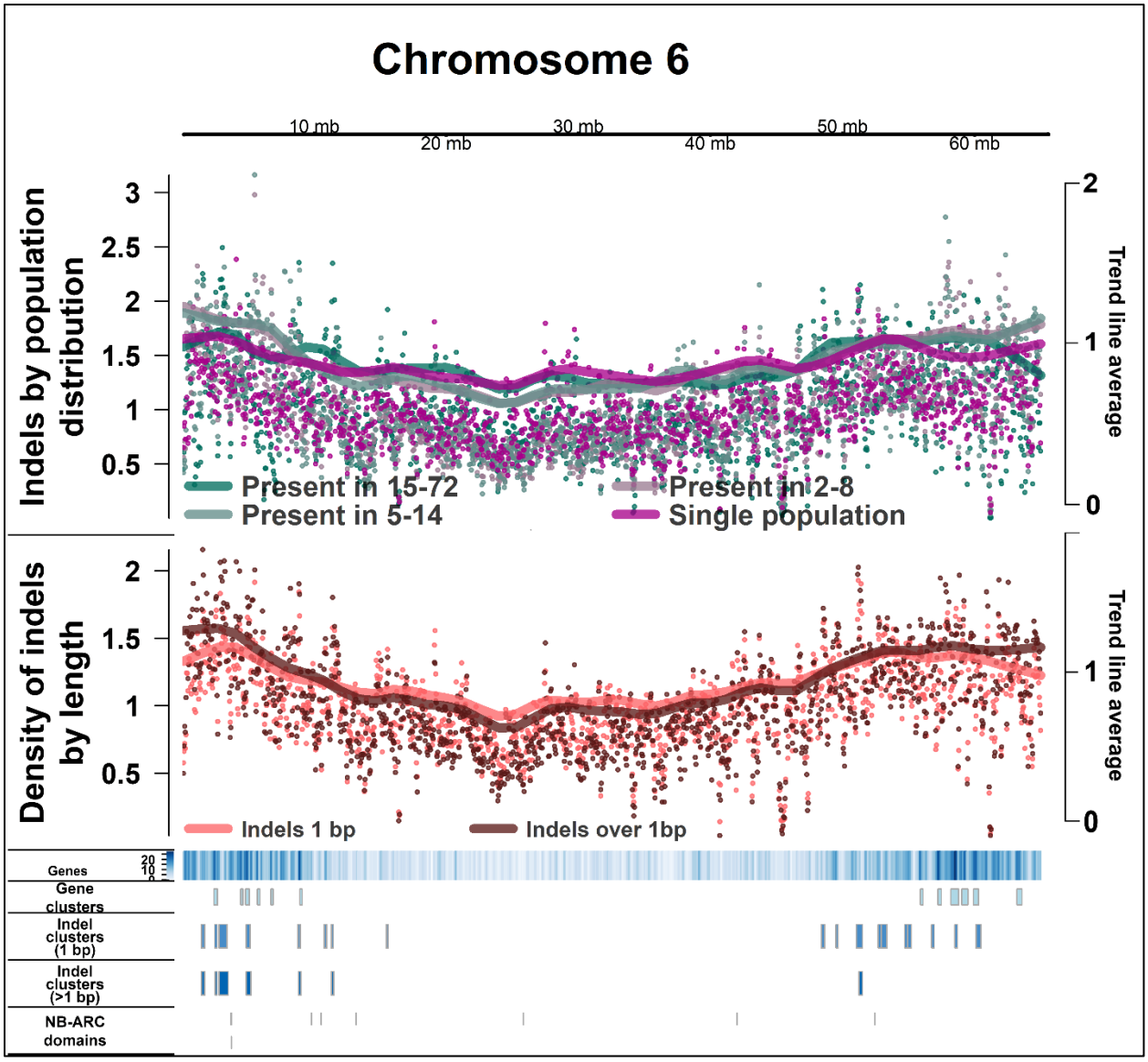

Figure 4.10. *B. vulgaris* chromosome 6 sequence-preserved indels (SPI) by length and population distribution. Points are normalized density of indels in 100kb sliding windows (scale left). Trend lines are calculated with Loess smoothing (scale right). Putative genes were derived from the EL10 annotation. Threshold for clusters of genes and indels was set at three standard deviations above the mean. NB-ARC domains were derived from Funk et al (2018).

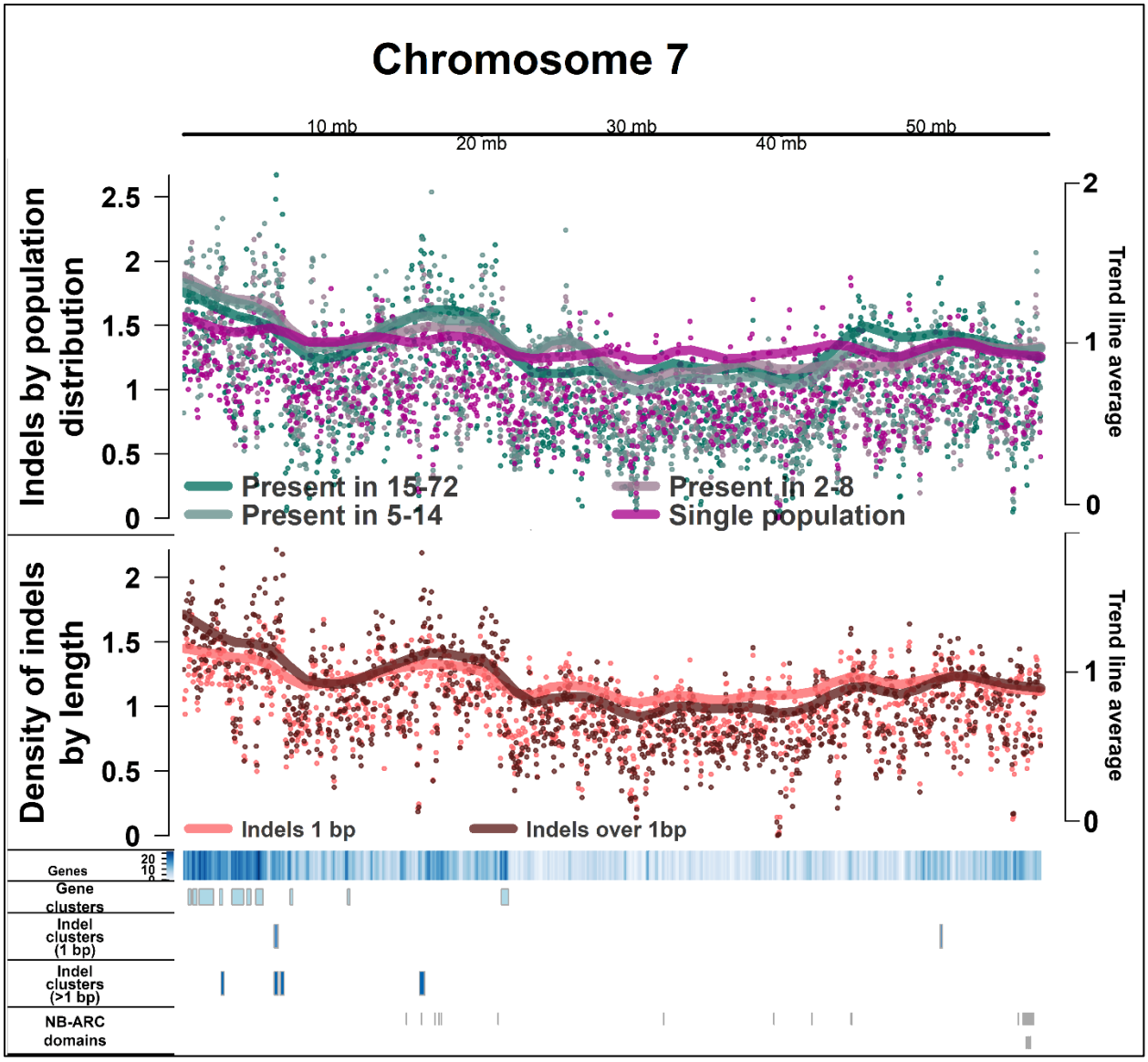

Figure 4.11. *B. vulgaris* chromosome 7 sequence-preserved indels (SPI) by length and population distribution. Points are normalized density of indels in 100kb sliding windows (scale left). Trend lines are calculated with Loess smoothing (scale right). Putative genes were derived from the EL10 annotation. Threshold for clusters of genes and indels was set at three standard deviations above the mean. NB-ARC domains were derived from Funk et al (2018).

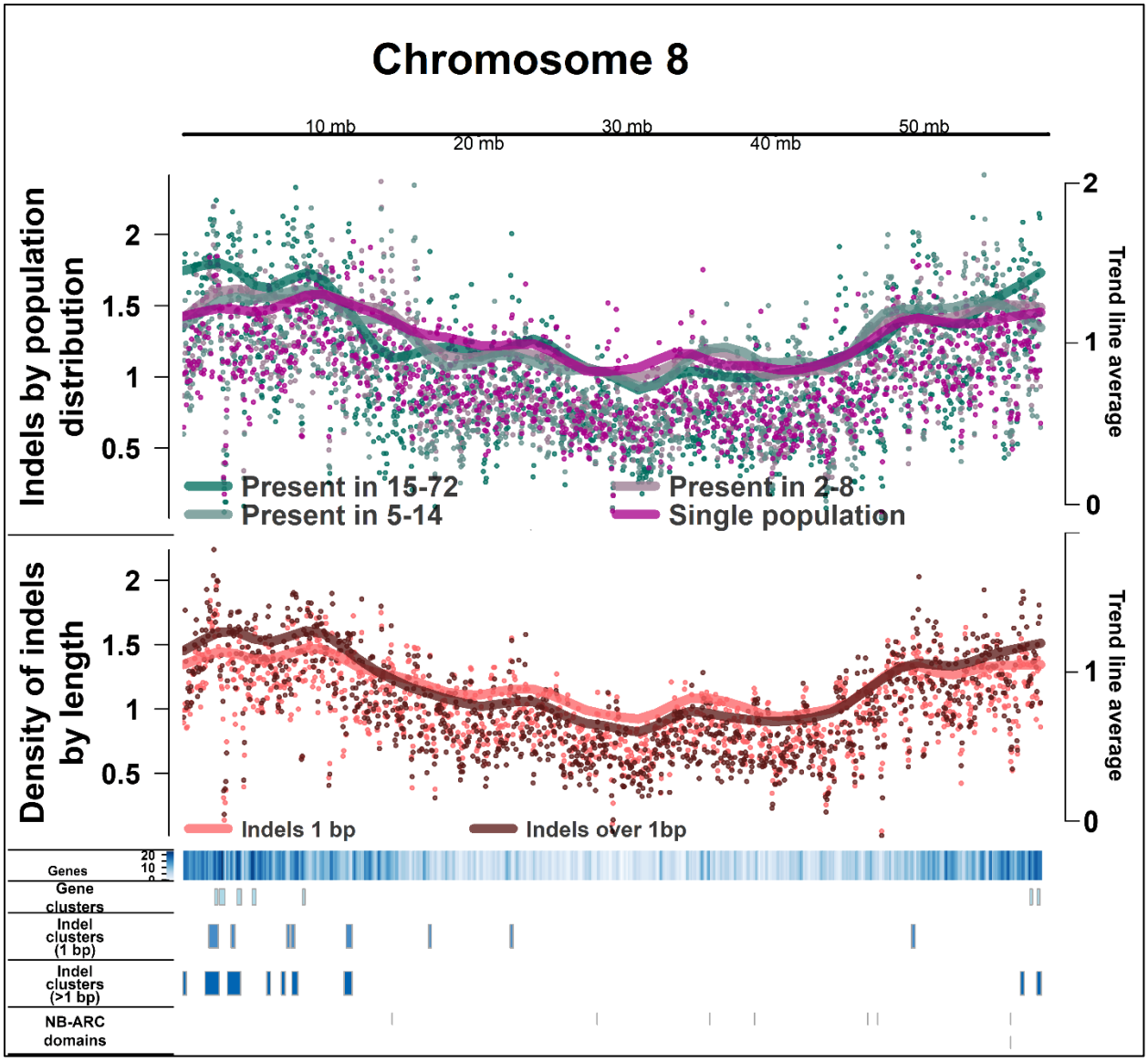

Figure 4.12. *B. vulgaris* chromosome 8 sequence-preserved indels (SPI) by length and population distribution. Points are normalized density of indels in 100kb sliding windows (scale left). Trend lines are calculated with Loess smoothing (scale right). Putative genes were derived from the EL10 annotation. Threshold for clusters of genes and indels was set at three standard deviations above the mean. NB-ARC domains were derived from Funk et al (2018).

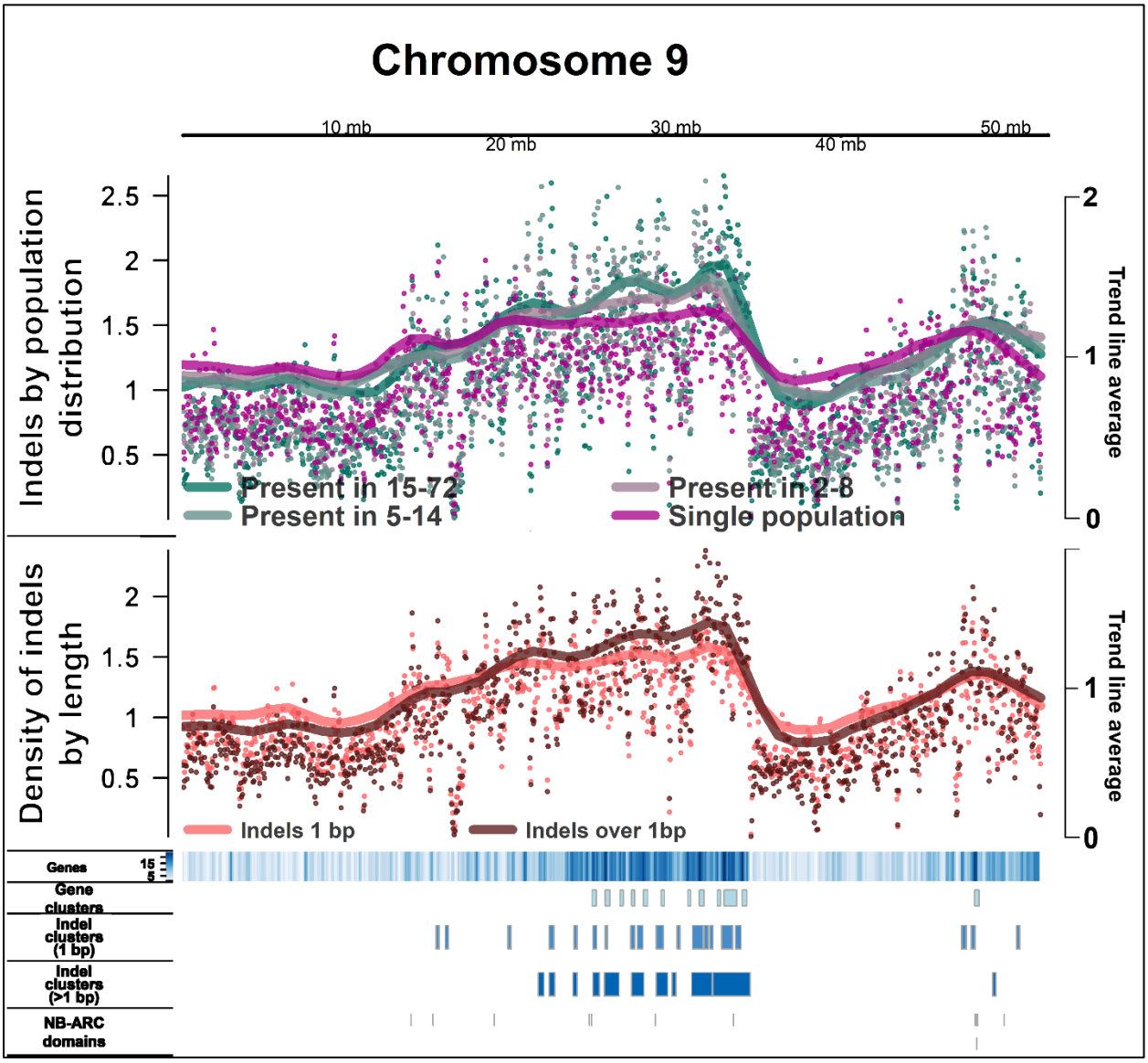

Figure 4.13. *B. vulgaris* chromosome 9 sequence-preserved indels (SPI) by length and population distribution. Points are normalized density of indels in 100kb sliding windows (scale left). Trend lines are calculated with Loess smoothing (scale right). Putative genes were derived from the EL10 annotation. Threshold for clusters of genes and indels was set at three standard deviations above the mean. NB-ARC domains were derived from Funk et al (2018).

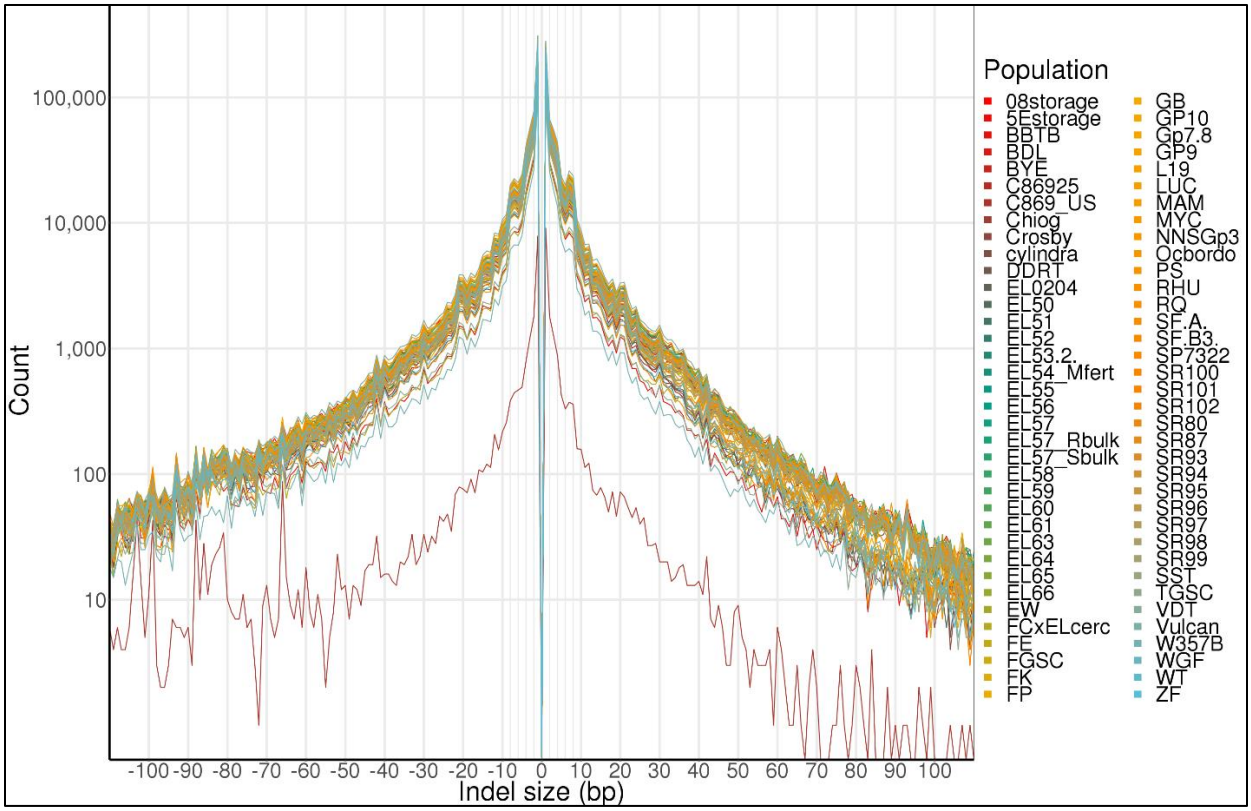

Figure 4.14: Number of sequence-preserved indels (SPI) of given lengths per *B. vulgaris* population. The reduced indel count in the C869 single plant reference sample (lower red line) was clearly distinguished from the pooled population samples. Colors are used to delineate different populations with no reference to other population qualities.

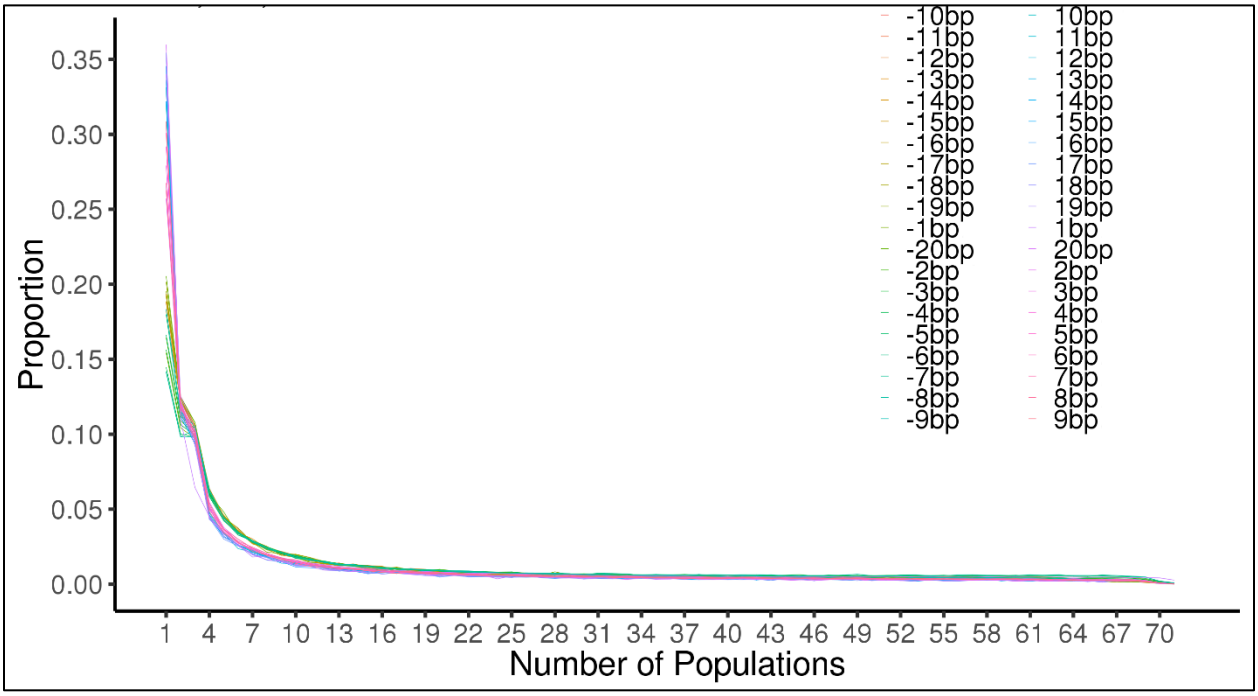

Figure 4.15: Distribution of sequence-preserved indels (SPI) 20 bp or fewer across 71 *B. vulgaris* populations. There were 4,995,443 distinct SPIs detected across all populations. The number of SPI appearing in only one population varied dependent on SPI length, with longer SPI more likely to be found in only one population.

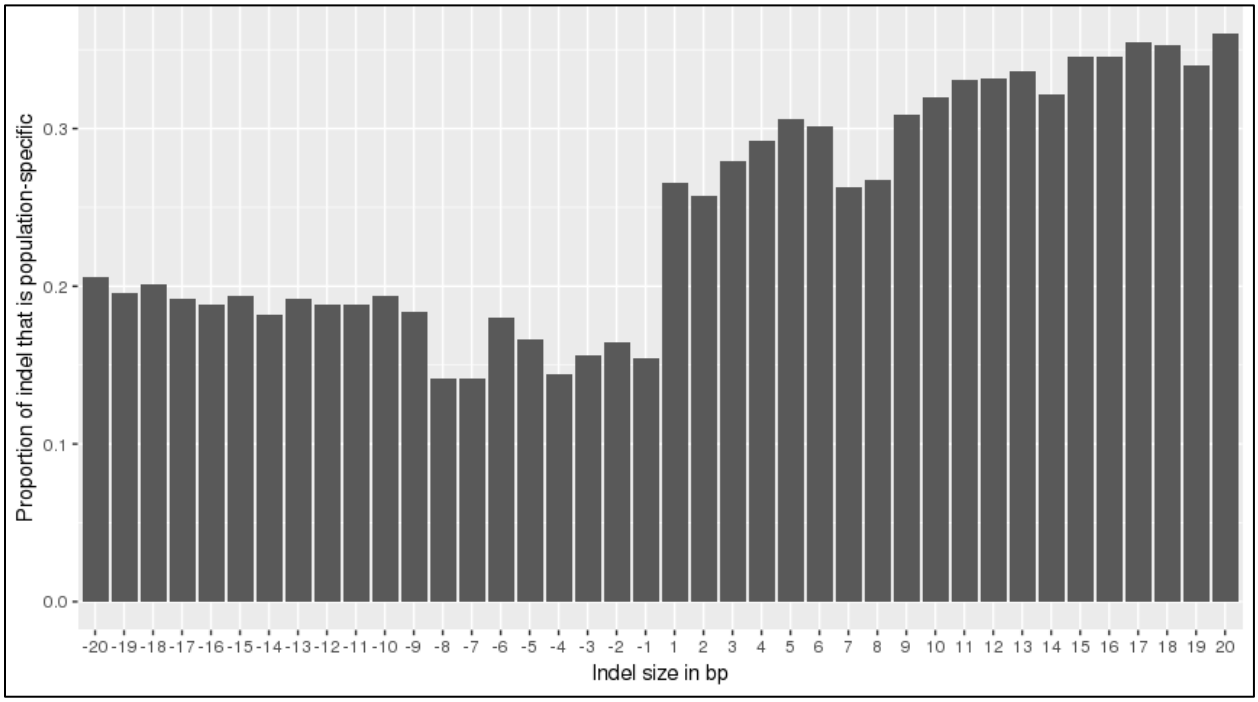

Figure 4.16: Proportion of 20 bp or shorter sequence-preserved indels (SPI) that are found in only one population of *B. vulgaris*. More SPI insertions (positive numbers) than deletions (negative numbers) were population-specific.

Figure 4.17: Sequence logos of the two most enriched motifs in the 7 or 8 bp sequence-preserved indels (SPI) sequences in populations of *B. vulgaris*. HYAYAA (left) and TYAYRA (right).

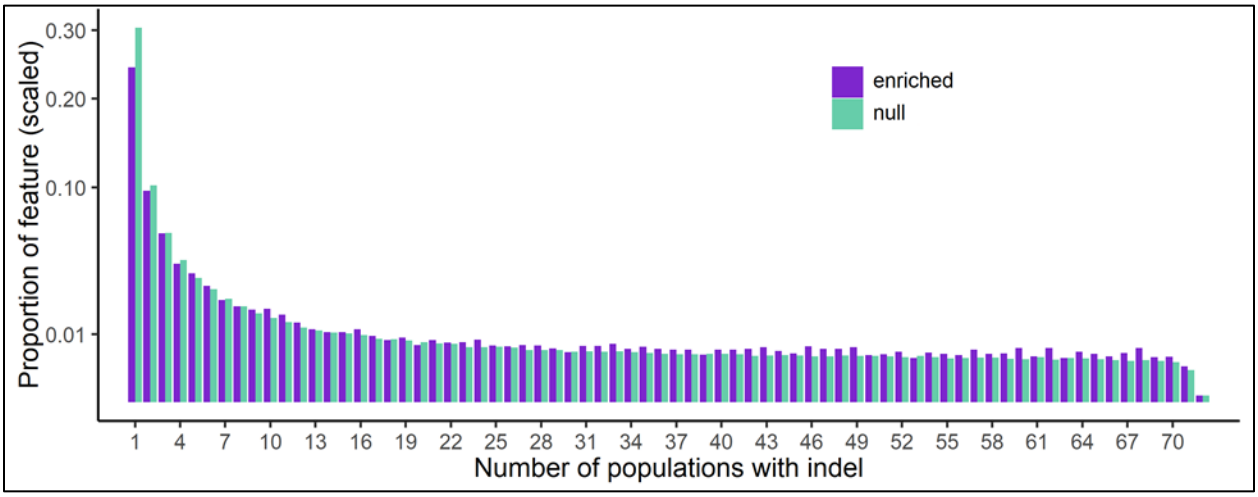

Figure 4.18: Distribution of 7 and 8 bp sequence-preserved indels (SPI) across populations of *B. vulgaris* partitioned by enriched sequence identity. SPIs containing HYAYAA or TYAYRA motifs were considered enriched (purple) versus the null distribution of all other 7-8 bp SPIs (green). SPIs containing either of the two enriched motifs were more likely to be found in multiple populations compared to other SPIs of similar sizes (significant difference between distributions, Mann-Whitney U test, $P < 0.004$).

CHAPTER FIVE

CONCLUDING REMARKS

The results presented in this dissertation follow my thought progression as I explored R gene diversity in beets over the past four years. Early questions began with "what do you do with a reference genome?" and expanded to questions of R gene family diversity, patterns of family expansion and subdivision, the mechanisms that generate genetic variation at the chromosome level, and how to extrapolate that to the pan-genome of a species.

The earliest question that piqued my interest was how to find resistance genes that I didn't know to look for. Protein prediction has been a shaky proposition, using transitive annotation to extrapolate protein sequence functions from a small set of experimentally validated sequences. I wondered if there could be genes in beet that were different enough from other model systems to be missed by current annotation pipelines. When the EL10 reference genome achieved chromosome-size scaffolds, I began to explore hidden Markov models to build a beet-specific NB-ARC domain model (Chapter 2). As that effort progressed, I became intimately familiar with the existing gene annotations and made the decision to avoid the abstraction of protein prediction in favor of the raw genomic sequence. That decision paid dividends as I was able to detect full-length as well as partial domains. I saw an array of domain length and completeness that suggested there was an unannotated graveyard of NB-ARC pseudogenes, hinting at the long evolutionary history of this gene family.

After successfully using nucleotide-based NB-ARC HMMs on the reference genome, I asked what domains existed in other populations of beets (Chapter 3). Scanning the *de novo* assemblies with the HMM was relatively straight-forward. The core set of domains detected in the various crop type assemblies indicated interesting subdivisions of resistance genes across the

B. vulgaris pan-genome. However, as I reflected on the assembly process and results, I became less confident in the assembly itself, and therefore felt cautious in drawing any strong conclusions from the data. I realized that the nature of pooled population sequences created additional complexity for the assembly algorithms, and we didn't have the data to confirm or deny conflicting hypotheses. Would multiple assembly programs generate similar results? Could we differentiate alleles and close paralogs, or were the short reads from similar sequences combined into a sort of genome soup? I backed away from interpreting NB-ARC domain data and focused on validating the assemblies. I believe the initial results are encouraging, both from the HMM and the assembly validation. However, more work is needed to establish the veracity of *de novo* assembly of pooled population sequences. I came to a paper late in the project that had pursued a similar assembly strategy in rice (Yao et al. 2015). Their methods and results reinforce the difficulty of working with distributed, dispensable genome sequences. However, if the methods could be fine-tuned and validated, the concept of *de novo* pooled population assembly could provide opportunities to investigate species evolution, domestication, breeding, and targeted trait analysis.

As I considered the sequences of R gene domains, I began to think about sources of genetic diversity in nature. A central concept in disease resistance literature is that diversity arises from gene duplication followed by sub-functionalization and neo-functionalization (Jones & Dangl 2006; Jones et al. 2016). Investigating mechanisms of gene duplication implicated errors in recombination, which led me to consider structural variation as an evolutionary process more broadly. Structural variation, whether small indels or larger chromosome rearrangements, has been increasingly recognized as a key component of genetic variation both mammals and plants (Carvalho & Lupski 2016; Thind et al. 2018). If recombination leads to structural

variation, and structural variation leads to R gene evolution, then we could ask whether structural variation in a genome actually corresponds with R gene diversity in agreement with current scientific understanding.

At that point I decided to try and characterize indels using pooled population sequences (Chapter 4). This was the largest data set I had worked with, leading to millions of indel predictions derived from terabytes of short-read sequencing data. Even though the program I used to predict indels was not intended for meta-genomic analysis, my efforts to compare indels across populations led to confidence that what I was seeing was grounded in biological activity rather than technical artifacts. Applying the SvABA targeted resequencing strategy to bulk segregant data sets generated a number of candidate genes with plausible biological function. As with the *de novo* whole-genome assemblies, validating the authenticity of predicted genomic features is a crucial next step. I believe the strength of the results presented in Chapter 4 is sufficient to warrant further efforts to confirm and extend those lines of research.

In summary, attempts to develop a nucleotide-based form of R gene detection led to questions of R gene pan-genomic diversity and signatures of R gene evolution in beets. My hope is not that I have solved anything, but that I have made tangible, well-reasoned progress to improve our understanding of both basic and applied biology. There are clear opportunities for additional research, which can be developed into stand-alone projects suitable for a variety of interests and skill levels. I look forward to my future efforts building on the foundation laid out in this dissertation.

Thanks for your time and attention.

LITERATURE CITED

LITERATURE CITED

Abyzov A, Li S, Kim DR, Mohiyuddin M, Stütz AM, Parrish NF, Mu XJ, Clark W, Chen K, Hurles M, Korbel JO, Lam HYK, Lee C, Gerstein MB (2015) Analysis of deletion breakpoints from 1,092 humans reveals details of mutation mechanisms. Nature Communications 6:7256

Alkan C, Coe BP, Eichler EE (2011) Genome structural variation discovery and genotyping. Nature Reviews Genetics 12:363–376

Andolfo G, Jupe F, Witek K, Etherington GJ, Ercolano MR, Jones JDG (2014) Defining the full tomato NB-LRR resistance gene repertoire using genomic and cDNA RenSeq. BMC Plant Biology 14:120–132

Armstrong J, Fiddes IT, Diekhans M, Paten B (2019) Whole-Genome Alignment and Comparative Annotation. Annual Review of Animal Biosciences 7:41–64

Arumuganathan K, Earle ED (1991) Nuclear DNA content of some important plant species. Plant Molecular Biology Reporter 9:415

Bailey PC, Schudoma C, Jackson W, Baggs E, Dagdas G, Haerty W, Moscou M, Krasileva K V. (2018) Dominant integration locus drives continuous diversification of plant immune receptors with exogenous domain fusions. Genome Biology 19:23

Bailey TL (2011) DREME: Motif discovery in transcription factor ChIP-seq data. Bioinformatics 27:1653–1659

Ball E V., Stenson PD, Abeysinghe SS, Krawczak M, Cooper DN, Chuzhanova NA (2005) Microdeletions and microinsertions causing human genetic disease: Common mechanisms of mutagenesis and the role of local DNA sequence complexity. Human Mutation 26:205–213

Bartlett A, O'Malley RC, Huang SSC, Galli M, Nery JR, Gallavotti A, Ecker JR (2017) Mapping genome-wide transcription-factor binding sites using DAP-seq. Nature Protocols 12:1659–1672

Barzen E, Mechelke W, Ritter E, Seitzer JF, Salamini F (1992) RFLP markers for sugar beet breeding: chromosomal linkage maps and location of major genes for rhizomania resistance, monogermy and hypocotyl colour. The Plant Journal 2:601–611

Barzen E, Stahl R, Fuchs E, Borchardt DC, Salamini F (1997) Development of coupling-repulsion-phase SCAR markers diagnostic for the sugar beet Rr1 allele conferring resistance to rhizomania. Molecular Breeding 3:231–238

Belkhadir Y, Subramaniam R, Dangl JL (2004) Plant disease resistance protein signaling: NBS-LRR proteins and their partners. Current Opinion in Plant Biology 7:391–399

Berger SA, Krompass D, Stamatakis A (2011) Performance, accuracy, and web server for evolutionary placement of short sequence reads under maximum likelihood. Systematic Biology 60:291–302

van Berkum NL, Lieberman-Aiden E, Williams L, Imakaev M, Gnirke A, Mirny LA, Dekker J, Lander ES (2010) Hi-C: A Method to Study the Three-dimensional Architecture of Genomes. Journal of Visualized Experiments 1–7

Biancardi E, Lewellen RT, De Biaggi M, Erichsen AW, Stevanato P (2002) The origin of rhizomania resistance in sugar beet. Euphytica 127:383–397

Biancardi E, Tamada T (2016) Rhizomania. Springer International Publishing, Cham

Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, Chan S, Lee J, Lam ET, Liachko I, Sullivan ST, Burton JN, Huson HJ, Nystrom JC, Kelley CM, Hutchison JL, Zhou Y, Sun J, Crisà A, Ponce De León FA, et al. (2017) Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. Nature Genetics 49:643–650

van der Biezen E a, Jones JDG (1998) The NB-ARC domain: a novel signalling motif shared by plant resistance gene products and regulators of cell death in animals. Current Biology 8:R226–R228

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120

Bonardi V, Tang S, Stallmann A, Roberts M, Cherkis K, Dangl JL (2011) Expanded functions for a family of plant intracellular immune receptors beyond specific recognition of pathogen effectors. Proceedings of the National Academy of Sciences 108:16463–16468

Bornemann K, Hanse B, Varrelmann M, Stevens M (2015) Occurrence of resistance-breaking strains of *Beet necrotic yellow vein virus* in sugar beet in northwestern Europe and identification of a new variant of the viral pathogenicity factor P25. Plant Pathology 64:25–34

Bornemann K, Varrelmann M (2013) Effect of sugar beet genotype on the *Beet necrotic yellow vein virus* P25 pathogenicity factor and evidence for a fitness penalty in resistance-breaking strains. Molecular Plant Pathology 14:356–364

Boyd LA, Ridout C, O'Sullivan DM, Leach JE, Leung H (2013) Plant-pathogen interactions: Disease resistance in modern agriculture. Trends in Genetics 29:233–240

Broccanello C, McGrath JMM, Panella L, Richardson K, Funk A, Chiodi C, Biscarini F, Barone V, Baglieri A, Squartini A, Concheri G, Stevanato P (2018) A SNP mutation affects rhizomania-virus content of sugar beets grown on resistance-breaking soils. Euphytica 214:1–8

Brown JKM (2015) Durable Resistance of Crops to Disease: A Darwinian Perspective. Annual Review of Phytopathology 53:513–539

Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J (2013) Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nature Biotechnology 31:1119–1125

Cai D, Kleine M, Kifle S, Harloff HJ, Sandal NN, Marcker KA, Klein-Lankhorst RM, Salentijn EM, Lange W, Stiekema WJ, Wyss U, Grundler FM, Jung C (1997) Positional cloning of a gene for nematode resistance in sugar beet. Science (New York, N.Y.) 275:832–4

Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J, Koenig D, Lanz C, Stegle O, Lippert C, Wang X, Ott F, Müller J, Alonso-Blanco C, Borgwardt K, Schmid KJ, Weigel D (2011) Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nature Genetics 43:956–965

Capistrano-Gossmann GG, Ries D, Holtgräwe D, Minoche A, Kraft T, Frerichmann SLM, Rosleff Soerensen T, Dohm JC, González I, Schilhabel M, Varrelmann M, Tschoep H, Uphoff H, Schütze K, Borchardt D, Toerjek O, Mechelke W, Lein JC, Schechert AW, et al. (2017) Crop wild relative populations of Beta vulgaris allow direct mapping of agronomically important genes. Nature communications 8:15708

Carvalho CMB, Lupski JR (2016) Mechanisms underlying structural variant formation in genomic disorders. Nature Reviews Genetics 17:224–238

Casey LW, Lavrencic P, Bentham AR, Cesari S, Ericsson DJ, Croll T, Turk D, Anderson PA, Mark AE, Dodds PN, Mobli M, Kobe B, Williams SJ (2016) The CC domain structure from the wheat stem rust resistance protein Sr33 challenges paradigms for dimerization in plant NLR proteins. Proceedings of the National Academy of Sciences of the United States of America 113:12856–12861

Césari S, Kanzaki H, Fujiwara T, Bernoux M, Chalvon V, Kawano Y, Shimamoto K, Dodds P, Terauchi R, Kroj T (2014) The NB-LRR proteins RGA4 and RGA5 interact functionally and physically to confer disease resistance. The EMBO journal 33:1–19

Cesari S, Bernoux M, Moncuquet P, Kroj T, Dodds PN (2014) A novel conserved mechanism for plant NLR protein pairs: the 'integrated decoy' hypothesis. Frontiers in plant science 5:606

Chia JM, Song C, Bradbury PJ, Costich D, De Leon N, Doebley J, Elshire RJ, Gaut B, Geller L, Glaubitz JC, Gore M, Guill KE, Holland J, Hufford MB, Lai J, Li M, Liu X, Lu Y, McCombie R, et al. (2012) Maize HapMap2 identifies extant variation from a genome in flux. Nature Genetics 44:803–807

Christie N, Tobias PA, Naidoo S, Külheim C (2016) The Eucalyptus grandis NBS-LRR gene family: Physical clustering and expression hotspots. Frontiers in Plant Science 6:1238

Christopoulou M, Wo SRC, Kozik A, McHale LK, Truco MJ, Wroblewski T, Michelmore RW (2015) Genome-wide architecture of disease resistance genes in lettuce. G3: Genes, Genomes, Genetics 5:2655–2669

Cilas C, Goebel F-R, Babin R, Avelino J (2016) Tropical crop pests and diseases in a climate change setting— a few examples. In: Climate Change and Agriculture Worldwide. Torquebiau, E, editor. Springer Netherlands, Dordrecht pp. 73–82.

Conrad DF, Hurles ME (2007) The population genetics of structural variation. Nature Genetics 39:S30–S36

Cooke DA, Scott RK (1993) The Sugar Beet Crop. Springer Netherlands, Dordrecht

Czech L, Stamatakis A (2017) Genesis.

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature Reviews Genetics 12:499–510

Dewannieux M, Esnault C, Heidmann T (2003) LINE-mediated retrotransposition of marked Alu sequences. Nature Genetics 35:41–48

Dinant S, Clark AM, Zhu Y, Vilaine F, Palauqui J-C, Kusiak C, Thompson GA (2003) Diversity of the superfamily of phloem lectins (phloem protein 2) in angiosperms. Plant physiology 131:114–28

Dohm JC, Minoche AE, Holtgräwe D, Capella-Gutiérrez S, Zakrzewski F, Tafer H, Rupp O, Sörensen TR, Stracke R, Reinhardt R, Goesmann A, Kraft T, Schulz B, Stadler PF, Schmidt T, Gabaldón T, Lehrach H, Weisshaar B, Himmelbauer H (2014) The genome of the recently domesticated crop plant sugar beet (Beta vulgaris). Nature 505:546–549

Dunsmuir P, Brorein WJ, Simon MA, Rubin GM (1980) Insertion of the *Drosophila* transposable element copia generates a 5 base pair duplication. Cell 21:575–579

Duxbury Z, Ma Y, Furzer OJ, Huh SU, Cevik V, Jones JDG, Sarris PF (2016) Pathogen perception by NLRs in plants and animals: Parallel worlds. BioEssays 769–781

Elshire RJ, Glaubitz JC, Sun Q, Poland J a., Kawamoto K, Buckler ES, Mitchell SE (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE 6:1–10

Flasch DA, Macia Á, Sánchez L, Ljungman M, Heras SR, García-Pérez JL, Wilson TE, Moran J V. (2019) Genome-wide de novo L1 retrotransposition connects endonuclease activity with replication. Cell 177:837-851.e28

Flor HH (1942) Inheritance of pathogenicity in Melampsora lini. Phytopathology 32:653–659

Flowers JM, Hazzouri KM, Pham GM, Rosas U, Bahmani T, Khraiwesh B, Nelson DR, Jijakli K, Abdrabu R, Harris EH, Lefebvre PA, Hom EFY, Salehi-Ashtiani K, Purugganan MD (2015) Whole-genome resequencing reveals extensive natural variation in the model green alga chlamydomonas reinhardtii. Plant Cell 27:2353–2369

Fornes O, Castro-Mondragon JA, Khan A, van der Lee R, Zhang X, Richmond PA, Modi BP, Correard S, Gheorghe M, Baranašić D, Santana-Garcia W, Tan G, Chèneby J, Ballester B, Parcy F, Sandelin A, Lenhard B, Wasserman WW, Mathelier A (2019) JASPAR 2020: update of the open-access database of transcription factor binding profiles. Nucleic Acids Research

Francis SA, Luterbacher MC (2003) Identification and exploitation of novel disease resistance genes in sugar beet. Pest management science 59:225–30

Franck CM, Westermann J, Boisson-Dernier A (2018) Plant malectin-like receptor kinases: from cell wall integrity to immunity and beyond. Annual Review of Plant Biology 69:301–328

Freeling M (2009) Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. Annual review of plant biology 60:433–453

Funk A, Galewski P, McGrath JMM (2018) Nucleotide-binding resistance gene signatures in sugar beet, insights from a new reference genome. The Plant Journal 95:659–671

Gabur I, Chawla HS, Snowdon RJ, Parkin IAP (2019) Connecting genome structural variation with complex traits in crop plants. Theoretical and Applied Genetics 132:733–750

Garcia-Diaz M, Kunkel TA (2006) Mechanism of a genetic glissando*: structural biology of indel mutations. Trends in Biochemical Sciences 31:206–214

Genome of the Netherlands Consortium (2014) Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nature genetics 46:818–25

Van Ghelder C, Esmenjaud D (2016) TNL genes in peach: Insights into the post-LRR domain. BMC Genomics 17:317

Gidner S, Lennefors B-L, Nilsson N-O, Bensefelt J, Johansson E, Gyllenspetz U, Kraft T (2005) QTL mapping of BNYVV resistance from the WB41 source in sugar beet. Genome 48:279–85

Gómez-Gómez L, Boller T (2000) FLS2: An LRR Receptor-like kinase involved in the perception of the bacterial elicitor flagellin in Arabidopsis. Molecular Cell 5:1003–1011

Grimmer MK, Trybush S, Hanley S, Francis S a., Karp a., Asher MJC (2007) An anchored linkage map for sugar beet based on AFLP, SNP and RAPD markers and QTL mapping of a new source of resistance to Beet necrotic yellow vein virus. Theoretical and Applied Genetics 114:1151–1160

Grimmer MK, Kraft T, Francis SA, Asher MJC (2008) QTL mapping of BNYVV resistance from the WB258 source in sugar beet. Plant Breeding 127:650–652

Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS (2007) Quantifying similarity between motifs. Genome Biology 8:R24

Halldorsson B V., Palsson G, Stefansson OA, Jonsson H, Hardarson MT, Eggertsson HP, Gunnarsson B, Oddsson A, Halldorsson GH, Zink F, Gudjonsson SA, Frigge ML, Thorleifsson G, Sigurdsson A, Stacey SN, Sulem P, Masson G, Helgason A, Gudbjartsson DF, et al. (2019) Human genetics: Characterizing mutagenic effects of recombination through a sequence-level genetic map. Science 363:eaau1043

Halling SM, Kleckner N (1982) A symmetrical six-base-pair target site sequence determines Tn10 insertion specificity. Cell 28:155–163

Hamilton JP, Buell CR (2012) Advances in plant genome sequencing. The Plant journal : for cell and molecular biology 70:177–90

Hammer G, Cooper M, Tardieu F, Welch S, Walsh B, van Eeuwijk F, Chapman S, Podlich D (2006) Models for navigating biological complexity in breeding improved crop plants. Trends in Plant Science 11:587–593

Hardigan MA, Crisovan E, Hamilton JP, Kim J, Laimbeer P, Leisner CP, Manrique-Carpintero NC, Newton L, Pham GM, Vaillancourt B, Yang X, Zeng Z, Douches DS, Jiang J, Veilleux RE, Buella CR (2015) Genome reduction uncovers a large dispensable genome and adaptive role for copy number variation in asexually propagated *Solanum Tuberosum*. Plant Cell 28:388–405

Hayashi N, Inoue H, Kato T, Funao T, Shirota M, Shimizu T, Kanamori H, Yamane H, Hayano-Saito Y, Matsumoto T, Yano M, Takatsuji H (2010) Durable panicle blast-resistance gene Pb1 encodes an atypical CC-NBS-LRR protein and was generated by acquiring a promoter through local genome duplication. Plant Journal 64:498–510

Hirsch CN, Foerster JM, Johnson JM, Sekhon RS, Muttoni G, Vaillancourt B, Peñagaricano F, Lindquist E, Pedraza MA, Barry K, de Leon N, Kaeppler SM, Buell CR (2014) Insights into the maize pan-genome and pan-transcriptome. The Plant cell 26:121–35

Hirsch CN, Buell CR (2013) Tapping the promise of genomics in species with complex, nonmodel genomes. Annual review of plant biology 64:89–110

Hodgkinson A, Eyre-Walker A (2011) Variation in the mutation rate across mammalian genomes. Nature reviews. Genetics 12:756–66

van der Hoorn RAL, Kamoun S (2008) From Guard to Decoy: A New Model for Perception of Plant Pathogen Effectors. the Plant Cell Online 20:2009–2017

Horton MW, Hancock AM, Huang YS, Toomajian C, Atwell S, Auton A, Muliyati NW, Platt A, Sperone FG, Vilhjálmsson BJ, Nordborg M, Borevitz JO, Bergelson J (2012) Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel. Nature genetics 44:212–6

Hu TT, Pattyn P, Bakker EG, Cao J, Cheng J, Clark RM, Fahlgren N, Fawcett J a, Grimwood J, Gundlach H, Haberer G, Hollister JD, Ossowski S, Ottilar RP, Salamov AA, Schneeberger K, Spannagl M, Wang X, Yang L, et al. (2011) The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. Nature Genetics 43:476–481

Huang X, Han B (2014) Natural variations and genome-wide association studies in crop plants. Annual review of plant biology 65:531–51

Huh SU, Cevik V, Ding P, Duxbury Z, Ma Y, Tomlinson L, Sarris PF, Jones JDG (2017) Protein-protein interactions in the RPS4/RRS1 immune receptor complex. PLoS Pathogens 13:1–22

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C (2004) Detection of large-scale variation in the human genome. Nature Genetics 36:949–951

Jackman SD, Vandervalk BP, Mohamadi H, Chu J, Yeo S, Hammond SA, Jahesh G, Khan H, Coombe L, Warren RL, Birol I (2017) ABySS 2 . 0 : Resource-Efficient Assembly of Large Genomes using a Bloom Filter Effect of Bloom Filter False Positive Rate. Genome Research 27:768–777

Jacob F, Vernaldi S, Maekawa T (2013) Evolution and conservation of plant NLR functions. Frontiers in Immunology 4:1–16

Jones J. D. G., Perkins S, Foster S, Tomlinson L, Verweij W, Jupe F, Witek K, Dorling S, Cooke D, Smoker M (2014) Elevating crop disease resistance with cloned genes. Philosophical Transactions of the Royal Society B: Biological Sciences 369:20130087–20130087

Jones JDG, Dangl JL (2006) The plant immune system. Nature 444:323–329

Jones JDG, Vance RE, Dangl JL (2016) Intracellular innate immune surveillance devices in plants and animals. Science 354:6395–6395

Jones Philip, Binns D, Chang H, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong S, Lopez R, Hunter S (2014) InterProScan 5: genome-scale protein function classification. Bioinformatics (Oxford, England) 30:1236–40

Jupe F, Pritchard L, Etherington GJ, MacKenzie K, Cock PJ, Wright F, Sharma SK, Bolser D, Bryan GJ, Jones JD, Hein I (2012) Identification and localisation of the NB-LRR gene family within the potato genome. BMC Genomics 13:75

Jupe F, Witek K, Verweij W, ??liwka J, Pritchard L, Etherington GJ, Maclean D, Cock PJ, Leggett RM, Bryan GJ, Cardle L, Hein I, Jones JDG (2013) 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 76:530–544

Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Molecular Biology and Evolution 30:772–780

Kehr J (2006) Phloem sap proteins: Their identities and potential roles in the interaction between plants and phloem-feeding insects. Journal of Experimental Botany 57:767–774

Keightley PD, Caballero A, García-Dorado A (1998) Population genetics: Surviving under mutation pressure. Current Biology 8:235–237

Khaja R, Zhang J, MacDonald JR, He Y, Joseph-George AM, Wei J, Rafiq MA, Qian C, Shago M, Pantano L, Aburatani H, Jones K, Redon R, Hurles M, Armengol L, Estivill X, Mural RJ, Lee C, Scherer SW, et al. (2006) Genome assembly comparison identifies structural variants in the human genome. Nature Genetics 38:1413–1418

Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F, Haugen E, Zerr T, Yamada NA, Tsang P, Newman TL, Tüzün E, Cheng Z, Ebling HM, Tusneem N, et al. (2008) Mapping and sequencing of structural variation from eight human genomes. Nature 453:56–64

Kim SH, Qi D, Ashfield T, Helm M, Innes RW (2016) Using decoys to expand the recognition specificity of a plant disease resistance protein. Science 351:684–687

Kohler A, Rinaldi C, Duplessis S, Baucher M, Geelen D, Duchaussoy F, Meyers BC, Boerjan W, Martin F (2008) Genome-wide identification of NBS resistance genes in Populus trichocarpa. Plant Molecular Biology 66:619–636

Kojima KK, Fujiwara H (2003) Evolution of target specificity in R1 clade non-LTR retrotransposons. Molecular Biology and Evolution 20:351–361

Kosugi S, Momozawa Y, Liu X, Terao C, Kubo M, Kamatani Y (2019) Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. Genome Biology 20:8–11

Kourelis J, Van Der Hoorn RAL (2018) Defended to the nines: 25 years of resistance gene cloning identifies nine mechanisms for R protein function. Plant Cell 30:285–299

Krattinger SG, Keller B (2016) Molecular genetics and evolution of disease resistance in cereals. New Phytologist 212:320-332l

Krumm N, Sudmant P, Ko A (2012) Copy number variation detection and genotyping from exome sequence data. Genome research 22:1525–1532

Lam H-M, Xu X, Liu X, Chen W, Yang G, Wong F-L, Li M-W, He W, Qin N, Wang B, Li J, Jian M, Wang Jian, Shao G, Wang Jun, Sun SS-M, Zhang G (2010) Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nature Genetics 42:1053–1059

Latrasse D, Jégu T, Li H, de Zelicourt A, Raynaud C, Legras S, Gust A, Samajova O, Veluchamy A, Rayapuram N, Ramirez-Prado JS, Kulikova O, Colcombet J, Bigeard J, Genot B, Bisseling T, Benhamed M, Hirt H (2017) MAPK-triggered chromatin reprogramming by histone deacetylase in plant innate immunity. Genome Biology 18:131

Lauer S, Avecilla G, Spealman P, Sethia G, Brandt N, Levy SF, Gresham D (2018) Single-cell copy number variant detection reveals the dynamics and diversity of adaptation.

Law M, Childs KL, Campbell MS, Stein JC, Olson AJ, Holt C, Panchy N, Lei J, Jiao D, Andorf CM, Lawrence CJ, Ware D, Shiu S-H, Sun Y, Jiang N, Yandell M (2015) Automated update, revision, and quality control of the maize genome annotations using MAKER-P improves the B73 RefGen_v3 gene models and identifies new genes. Plant physiology 167:25–39

Leister D (2004) Tandem and segmental gene duplication and recombination in the evolution of plant disease resistance genes. Trends in Genetics 20:116–122

Leucker M, Mahlein A-K, Steiner U, Oerke E-C (2016) Improvement of lesion phenotyping in Cercospora beticola – sugar beet interaction by hyperspectral imaging. Phytopathology 106:177–184

Lewellen R (1991) Registration of rhizomania-resistant germplasm of Beta vulgaris. Crop Science 31:291–293

Lewellen RT (2004) Registration of rhizomania resistant, monogerm populations C869 and C869CMS sugarbeet. Crop Science 44:357

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079

Li X, Kapos P, Zhang Y (2015) NLRs in plants. Current Opinion in Immunology 32:114–121

Liao G -c., Rehm EJ, Rubin GM (2012) Insertion site preferences of the P transposable element in Drosophila melanogaster. Proceedings of the National Academy of Sciences 97:3347–3351

Linheiro RS, Bergman CM (2012) Whole genome resequencing reveals natural target site preferences of transposable elements in *Drosophila* melanogaster. PLoS ONE 7

Lipka AE, Kandianis CB, Hudson ME, Yu J, Drnevich J, Bradbury PJ, Gore M a (2015) From association to prediction: statistical methods for the dissection and selection of complex traits in plants. Current Opinion in Plant Biology 24:110–118

Litwiniec A, Gośka M, Choińska B, Kużdowicz K, Łukanowski A, Skibowska B (2015) Evaluation of rhizomania-resistance segregating sequences and overall genetic diversity pattern among selected accessions of Beta and Patellifolia. Potential implications of breeding for genetic bottlenecks in terms of rhizomania resistance. Euphytica

Liu H-Y, Sears JL, Lewellen RT (2005) Occurrence of Resistance-Breaking Beet necrotic yellow vein virus of Sugar Beet. Plant Disease 89:464–468

Liu HY, Lewellen RT (2007) Distribution and molecular characterization of resistance-breaking isolates of Beet necrotic yellow vein virus in the United States. Plant Disease 91:847–851

Loutre C, Wicker T, Travella S, Galli P, Scofield S, Fahima T, Feuillet C, Keller B (2009) Two different CC-NBS-LRR genes are required for Lr10-mediated leaf rust resistance in tetraploid and hexaploid wheat. Plant Journal 60:1043–1054

Lu F, Romay MC, Glaubitz JC, Bradbury PJ, Elshire RJ, Wang T, Li Yu, Li Yongxiang, Semagn K, Zhang X, Hernandez AG, Mikel MA, Soifer I, Barad O, Buckler ES (2015) High-resolution genetic mapping of maize pan-genome sequence anchors. Nature Communications 6:6914

De Lucchi C, Stevanato P, Hanson L, McGrath M, Panella L, De Biaggi M, Broccanello C, Bertaggia M, Sella L, Concheri G (2017) Molecular markers for improving control of soil-borne pathogen Fusarium oxysporum in sugar beet. Euphytica 213:71

Lynch M, Ackerman MS, Gout JF, Long H, Sung W, Thomas WK, Foster PL (2016) Genetic drift, selection and the evolution of the mutation rate. Nature Reviews Genetics 17:704–714

Ma Y, Szostkiewicz I, Korte A, Moes D, Yang Y, Christmann A, Grill E (2009) Regulators of PP2C phosphatase activity function as abscisic acid sensors. Science (New York, N.Y.) 324:1064–8

Maekawa T, Kufer TA, Schulze-Lefert P (2011) NLR functions in plant and animal immune systems: so far and yet so close. Nature Immunology 12:817–826

Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A (2018) MUMmer4: A fast and versatile genome alignment system. PLoS Computational Biology 14:e1005944

Martin J, Lupas AN (2013) AAA-ATPases. In: Encyclopedia of Biological Chemistry. Elsevier pp. 1–6.

McCulloch SD, Kunkel TA (2008) The fidelity of DNA synthesis by eukaryotic replicative and translesion synthesis polymerases. Cell Research 18:148–161

McDowell JM, Simon SA (2006) Recent insights into R gene evolution. Molecular Plant Pathology 7:437–448

McGrath JM, Panella L (2018) Plant Breeding Reviews. Wiley

McHale L, Tan X, Koehl P, Michelmore RW (2006) Plant NBS-LRR proteins: adaptable guards. Genome biology 7:212

Meyers BC, Kozik A, Griego A, Kuang H, Michelmore RW (2003) Genome-wide analysis of NBS-LRR-encoding genes in Arabidopsis. Plant Cell 15:809–834

Miles A, Iqbal Z, Vauterin P, Pearson R, Campino S, Theron M, Gould K, Mead D, Drury E, O'Brien J, Rubio VR, Macinnis B, Mwangi J, Samarakoon U, Ranford-Cartwright L, Ferdig M, Hayton K, Su XZ, Wellems T, et al. (2016) Indels, structural variation, and recombination drive genomic diversity in Plasmodium falciparum. Genome Research 26:1288–1299

Mondragon-Palomino M, Gaut BS (2005) Gene conversion and the evolution of three leucine-rich repeat gene families in *Arabidopsis thaliana*. Molecular Biology and Evolution 22:2444–2456

Monosi B, Wisser RJ, Pennill L, Hulbert SH (2004) Full-genome analysis of resistance gene homologues in rice. Theoretical and Applied Genetics 109:1434–1447

Monteiro F, Nishimura MT (2018) Structural, functional, and genomic diversity of plant NLR proteins: an evolved resource for rational engineering of plant immunity. Annual Review of Phytopathology 56:243–267

Mun JH, Yu HJ, Park S, Park BS (2009) Genome-wide identification of NBS-encoding resistance genes in Brassica rapa. Molecular Genetics and Genomics 282:617–631

Nandety RS, Caplan JL, Cavanaugh K, Perroud B, Wroblewski T, Michelmore RW, Meyers BC (2013) The role of TIR-NBS and TIR-X proteins in plant basal defense responses. Plant physiology 162:1459–72

Narusaka M, Kubo Y, Hatakeyama K, Imamura J, Ezura H, Nanasato Y, Tabei Y, Takano Y, Shirasu K, Narusaka Y (2013) Interfamily Transfer of Dual NB-LRR Genes Confers Resistance to Multiple Pathogens. PLoS ONE 8:6–13

Nishimura MT, Anderson RG, Cherkis KA, Law TF, Liu QL, Machius M, Nimchuk ZL, Yang L, Chung E-H, El Kasmi F, Hyunh M, Osborne Nishimura E, Sondek JE, Dangl JL (2017) TIR-only protein RBA1 recognizes a pathogen effector to regulate cell death in *Arabidopsis*. Proceedings of the National Academy of Sciences 114:E2053–E2062

Pan Q, Liu YS, Budai-Hadrian O, Sela M, Carmel-Goren L, Zamir D, Fluhr R (2000) Comparative genetics of nucleotide binding site-leucine rich repeat resistance gene homologues in the genomes of two dicotyledons: Tomato and arabidopsis. Genetics 155:309–322

Panella L, Lewellen RT (2007) Broadening the genetic base of sugar beet: Introgression from wild relatives. Euphytica 154:383–400

Pferdmenges F, Korf H, Varrelmann M (2009) Identification of rhizomania-infected soil in Europe able to overcome Rz1 resistance in sugar beet and comparison with other resistance-breaking soils from different geographic origins. European Journal of Plant Pathology 124:31–43

Pham GM, Newton L, Wiegert-Rininger K, Vaillancourt B, Douches DS, Buell CR (2017) Extensive genome heterogeneity leads to preferential allele expression and copy number-dependent expression in cultivated potato. The Plant Journal 92:624–637

Pham GM, Braz GT, Conway M, Crisovan E, Hamilton JP, Laimbeer FPE, Manrique-Carpintero N, Newton L, Douches DS, Jiang J, Veilleux RE, Buell CR (2019) Genome-wide Inference of Somatic Translocation Events During Potato Dihaploid Production. The Plant Genome 12:180079

Pinosio S, Giacomello S, Faivre-Rampant P, Taylor G, Jorge V, Le Paslier MC, Zaina G, Bastien C, Cattonaro F, Marroni F, Morgante M (2016) Characterization of the poplar pan-genome by genome-wide identification of structural variation. Molecular Biology and Evolution 33:2706–2719

Plissonneau C, Benevenuto J, Mohd-Assaad N, Fouché S, Hartmann FE, Croll D (2017) Using population and comparative genomics to understand the genetic basis of effector-driven fungal pathogen evolution. Frontiers in Plant Science 8:119

Poland J, Rutkoski J (2016) Advances and challenges in genomic selection for disease resistance. Annual Review of Phytopathology 54:79–98

Porter BW, Paidi M, Ming R, Alam M, Nishijima WT, Zhu YJ (2009) Genome-wide analysis of Carica papaya reveals a small NBS resistance gene family. Molecular Genetics and Genomics 281:609–626

Rafalski JA (2002) Novel genetic mapping tools in plants : SNPs and LD-based approaches. 162:329–333

Rüdiger H, Gabius HJ (2002) Plant lectins: Occurrence, biochemistry, functions and applications. Glycoconjugate Journal 18:589–613

Salgotra RK, Gupta BB, Stewart CN (2014) From genomics to functional markers in the era of next-generation sequencing. Biotechnology letters 36:417–26

Sarris PF, Cevik V, Dagdas G, Jones JDG, Krasileva K V. (2016) Comparative analysis of plant immune receptor architectures uncovers host proteins likely targeted by pathogens. BMC Biology 14:8

Schatz MC, Maron LG, Stein JC, Wences AH, Gurtowski J, Biggers E, Lee H, Kramer M, Antoniou E, Ghiban E, Wright MH, Chia J, Ware D, McCouch SR, McCombie WR (2014) Whole genome de novo assemblies of three divergent strains of rice, Oryza sativa, document novel gene space of aus and indica. Genome Biology 15:506

Schnable JC, Freeling M, Lyons E (2012) Genome-wide analysis of syntenic gene deletion in the grasses. Genome biology and evolution 4:265–77

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlinson C, Strong C, Delehaunty K, Fronick C, Courtney B, et al. (2009) The B73 maize genome: Complexity, diversity, and dynamics. Science 326:1112–1115

Scholten OE, De Bock TSM, Klein-Lankhorst RM, Lange W (1999) Inheritance of resistance to beet necrotic yellow vein virus in Beta vulgaris conferred by a second gene for resistance. Theoretical and Applied Genetics 99:740–746

Seo E, Kim S, Yeom SI, Choi D (2016) Genome-wide comparative analyses reveal the dynamic evolution of nucleotide-binding leucine-rich repeat gene family among solanaceae plants. Frontiers in Plant Science 7:1205–1217

Shao ZQ, Zhang YM, Hang YY, Xue JY, Zhou GC, Wu P, Wu XY, Wu XZ, Wang Q, Wang B, Chen JQ (2014) Long-term evolution of nucleotide-binding site-leucine-rich repeat genes: Understanding gained from and beyond the legume family. Plant Physiology 166:217–234

Shapiro B, Rambaut A, Drummond AJ (2006) Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. Molecular Biology and Evolution 23:7–9

Shiu S, Karlowski W, Pan R (2004) Comparative analysis of the receptor-like kinase family in Arabidopsis and rice. The Plant Cell … 16:1220–1234

Shiu SH, Bleecker AB (2003) Expansion of the receptor-like kinase/Pelle gene family and receptor-like proteins in Arabidopsis. Plant Physiology 132:530–543

Shiu SH, Bleecker AB (2001) Plant receptor-like kinase gene family: diversity, function, and signaling. Signal Transduction Knowledge Environment 2001:1–13

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM (2015) BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210–3212

Sinapidou E, Williams K, Nott L, Bahkt S, T??r M, Crute I, Bittner-Eddy P, Beynon J (2004) Two TIR:NB:LRR genes are required to specify resistance to Peronospora parasitica isolate Cala2 in Arabidopsis. Plant Journal 38:898–909

Song WY, Wang GL, Chen LL, Kim HS, Pi LY, Holsten T, Gardner J, Wang B, Zhai WX, Zhu LH, Fauquet C, Ronald P (1995) A receptor kinase-like protein encoded by the rice disease resistance gene, Xa21. Science (New York, N.Y.) 270:1804–6

Stamatakis A (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313

Steinbrenner AD, Goritschnig S, Staskawicz BJ (2015) Recognition and Activation Domains Contribute to Allele-Specific Responses of an Arabidopsis NLR Receptor to an Oomycete Effector Protein. PLoS Pathogens 11:1–19

Steuernagel B, Jupe F, Witek K, Jones JDG, Wulff BBH (2015) NLR-parser: Rapid annotation of plant NLR complements. Bioinformatics 31:1665–1667

Stevanato P, Trebbi D, Panella L, Richardson K, Broccanello C, Pakish L, Fenwick AL, Saccomani M (2014) Identification and Validation of a SNP Marker Linked to the Gene HsBvm-1 for Nematode Resistance in Sugar Beet. Plant Molecular Biology Reporter 33:474–479

Stevanato P, De Biaggi M, Broccanello C, Biancardi E, Saccomani M (2015) Molecular genotyping of "Rizor" and "Holly" rhizomania resistances in sugar beet. Euphytica

Stevanato P, Trebbi D (2012) Identification of SNP markers linked to the Rz1 gene in sugar beet. International sugar Journal 114:2010–2013

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MHY, Konkel MK, Malhotra A, Stütz AM, Shi X, Casale FP, Chen J, Hormozdiari F, Dayama G, Chen K, et al. (2015) An integrated map of structural variation in 2,504 human genomes. Nature 526:75–81

Tang H, Lyons E, Town CD (2015) Optical mapping in plant comparative genomics. GigaScience 4:3

Tatout C, Lavie L, Deragon JM (1998) Similar target site selection occurs in integration of plant and mammalian retroposons. Journal of Molecular Evolution 47:463–470

Tester M, Langridge P (2010) Breeding technologies to increase crop production in a changing world. Science (New York, N.Y.) 327:818–22

Thind AK, Wicker T, Müller T, Ackermann PM, Steuernagel B, Wulff BBH, Spannagl M, Twardziok SO, Felder M, Lux T, Mayer KFX, Keller B, Krattinger SG (2018) Chromosome-scale comparative sequence analysis unravels molecular mechanisms of genome dynamics between two wheat cultivars. Genome Biology 19:1–16

Tian Y, Fan L, Thurau T, Jung C, Cai D (2004) The absence of TIR-type resistance gene analogues in the sugar beet (Beta vulgaris L.) genome. Journal of Molecular Evolution 58:40–53

Tran HT, Keen JD, Kricker M, Resnick MA, Gordenin DA (1997) Hypermutability of homonucleotide runs in mismatch repair and DNA polymerase proofreading yeast mutants. Molecular and Cellular Biology 17:2859–2865

Urbach JM, Ausubel FM (2017) The NBS-LRR architectures of plant R-proteins and metazoan NLRs evolved in independent events. Proceedings of the National Academy of Sciences 114:1063–1068

Vij S, Giri J, Dansana PK, Kapoor S, Tyagi AK (2008) The receptor-like cytoplasmic kinase (OsRLCK) gene family in rice: Organization, phylogenetic relationship, and expression during development and stress. Molecular Plant 1:732–750

Volpe M, Miralto M, Gustincich S, Sanges R (2018) ClusterScan: Simple and generalistic identification of genomic clusters. Bioinformatics 34:3921–3923

Wala JA, Bandopadhayay P, Greenwald NF, O'Rourke R, Sharpe T, Stewart C, Schumacher S, Li Y, Weischenfeldt J, Yao X, Nusbaum C, Campbell P, Getz G, Meyerson M, Zhang C-Z, Imielinski M, Beroukhim R (2018) SvABA: genome-wide detection of structural variants and indels by local assembly. Genome Research 28:581–591

Wang GF, Ji J, EI-Kasmi F, Dangl JL, Johal G, Balint-Kurti PJ (2015) Molecular and Functional Analyses of a Maize Autoactive NB-LRR Protein Identify Precise Structural Requirements for Activity. PLoS Pathogens 11:1–32

Waterhouse RM, Seppey M, Simao FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva E V., Zdobnov EM (2018) BUSCO applications from quality assessments to gene prediction and phylogenomics. Molecular Biology and Evolution 35:543–548

Webb KM, Freeman C, Broeckling CD (2016) Metabolome profiling to understand the defense response of sugar beet (Beta vulgaris) to Rhizoctonia solani AG 2-2 IIIB. Physiological and Molecular Plant Pathology 94:108–117

Wheeler TJ, Eddy SR (2013) nhmmer: DNA homology search with profile HMMs. Bioinformatics 29:2487–2489

Wijnen CL, Keurentjes JJB (2014) Genetic resources for quantitative trait analysis: Novelty and efficiency in design from an Arabidopsis perspective. Current Opinion in Plant Biology 18:103–109

Wu M, Chen T, Jiang R (2017) Leveraging multiple genomic data to prioritize disease-causing indels from exome sequencing data. Scientific Reports 7:1–12

Wu TD, Nacu S (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics 26:873–881

Xu G, Yuan M, Ai C, Liu L, Zhuang E, Karapetyan S, Wang S, Dong X (2017) uORF-mediated translation allows engineered plant disease resistance without fitness costs. Nature 545:491–494

Xu X, Liu X, Ge S, Jensen JD, Hu F, Li X, Dong Y, Gutenkunst RN, Fang L, Huang L, Li J, He W, Zhang G, Zheng X, Zhang F, Li Y, Yu C, Kristiansen K, Zhang X, et al. (2012) Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. Nature Biotechnology 30:105–111

Yan J, Kandianis CB, Harjes CE, Bai L, Kim EH, Yang X, Skinner DJ, Fu Z, Mitchell S, Li Q, Fernandez MGS, Zaharieva M, Babu R, Fu Y, Palacios N, Li J, Dellapenna D, Brutnell T, Buckler ES, et al. (2010) Rare genetic variation at Zea mays crtRB1 increases B-carotene in maize grain. Nature Genetics 42:322–327

Yao W, Li G, Zhao H, Wang G, Lian X, Xie W (2015) Exploring the rice dispensable genome using a metagenome-like assembly strategy. Genome Biology 16:1–20

Young ND (1996) Qtl Mapping and Quantitative Disease Resistance in Plants. Annual Review of Phytopathology 34:479–501

Yuan B, Zhai C, Wang W, Zeng X, Xu X, Hu H, Lin F, Wang L, Pan Q (2011) The Pik-p resistance to Magnaporthe oryzae in rice is mediated by a pair of closely linked CC-NBS-LRR genes. Theoretical and Applied Genetics 122:1017–1028

Zhang G, Liu X, Quan Z, Cheng S, Xu X, Pan S, Xie M, Zeng P, Yue Z, Wang W, Tao Y, Bian C, Han C, Xia Q, Peng X, Cao R, Yang X, Zhan D, Hu J, et al. (2012) Genome sequence of foxtail millet (Setaria italica) provides insights into grass evolution and biofuel potential. Nature Biotechnology 30:549–554

Zhang L, Ren Y, Yang T, Li G, Chen J, Gschwend AR, Yu Y, Hou G, Zi J, Zhou R, Wen B, Zhang J, Chougule K, Wang M, Copetti D, Peng Z, Zhang C, Zhang Y, Ouyang Y, et al. (2019) Rapid evolution of protein diversity by de novo origination in Oryza. Nature Ecology and Evolution 3:679–690

Zhang R, Murat F, Pont C, Langin T, Salse J (2014) Paleo-evolutionary plasticity of plant disease resistance genes. BMC Genomics 15:187

Zhou J, Loh YT, Bressan RA, Martin GB (1995) The tomato gene Pti1 encodes a serine/threonine kinase that is phosphorylated by Pto and is involved in the hypersensitive response. Cell 83:925–935

Zhou T, Wang Y, Chen JQ, Araki H, Jing Z, Jiang K, Shen J, Tian D (2004) Genome-wide identification of NBS genes in japonica rice reveals significant expansion of divergent non-TIR NBS-LRR genes. Molecular Genetics and Genomics 271:402–415

Zmienko A, Samelak-Czajka A, Kozlowski P, Szymanska M, Figlerowicz M (2016) *Arabidopsis thaliana* population analysis reveals high plasticity of the genomic region spanning MSH2, AT3G18530 and AT3G18535 genes and provides evidence for NAHR-driven recurrent CNV events occurring in this location. BMC Genomics 17:1–16