Contents

Chapter 1

Introduction

It defines the objectives and the importance of the research. It focus on the the application of Next Generation Sequencing to molecular biology, wheat genetics and ultimately to breeding programs. It also mentions the current status of the wheat reference genome and other resources (genetic maps, markers) the need of tools to query them effectively.

1.1 Wheat Breeding

An overview of how breeding is carried on currently, the different sources of genetic diversity and the relevance of fixing agriculturally important traits.

1.2 Wheat Genetics

The section describes alleles an the concept of gene, both as a locus in the genome (Quantitative Trait Locus, QTL) and an specific transcript (central dogma of molecular biology). Finally, it discuses traditional Mendelian inheritance and the effect of polyploidy.

1.3 Wheat Genomics

A description of the current status of the wheat genome (Mayer et al. (2014), Chapman et al. (2015)), the different available assemblies and and approaches to sort the scaffolds (Genome Zipper, the various genetic maps).

1.4 Sequencing

The Human Genome Project used Sanger sequencing Lander et al. (2001). This technology is the current gold standard in terms of quality of the sequence. It evolved from electrophoresis gels where the bands represented bases to a fully automated technique. However, the throughput is limited and doing genome wide analysis has prohibitive costs. In the second half of the 2000s high-throughput sequencing technologies emerged which had reduced the cost of sequencing. The main principle of the second-generation sequencing is to produce clusters of clones (i.e. ePCR), fix them in a plate and then add bases with a fluorescent marker. The reaction happens in parallel in millions of clusters at the same time. With each cycle, a picture is taken, showing the fluorescence of each base. Then, image processing algorithms find where in the image the clusters are and the bases are called. At this scale, the volume and complexity of the information is not trivial to manipulate, hence computing is required.

According to the objectives of the experiment and the quality and volume of the available DNA, the library can be prepared on fragments of different sizes, the classification of the available sequencing for the fragments is the followingMyllykangas et al. (2012); Metzker (2010); Shendure and Ji (2008); Hutchison (2007):

Single end When the fragments are short, it is possible to just sequence from the 5'-end the read.

Read Pairs When the sample consists fragments of up to 500bp, it is possible to read the 5' end up to the read length were the quality starts to drop, the molecule can be turned upside down, reverse complemented and sequence backwards. It is not required, but ideally, the fragments sequenced with read pairs should be selected to have an homogenous size. The reads are in opposite orientation relative to each other.

Overlapping Read Pairs are a variation to read pairs, where the size of the fragment is shorter than two times the read length. This allow an alignment between the two fragments to get an longer read with the limitations of the instrument.

Mate pairs are used to get reads separated at distances between 1kbp and 5kbp. To achieve this, the molecule is circularised and the point were the two ends of the fragment were joint a biotin marker is inserted. Then, the molecule is fragmented again and the fragments containing the biotin are sequenced in the same fashion that read pairs. The resulting reads have the same orientation.

There are several types of experiments that can be analysed with hight throughput sequencing, accordingly, different protocols for the sample preparation exist. The following is a short list of some of them

Whole genome shotgun When a sample is prepared for WGS, the DNA is extracted and chopped in fragments and sequenced. The reads obtained are, in principle, randomly distributed across the whole genome

RNA-Seq . Instead of sequencing DNA, mRNA is captured and sequenced. The fragments are not amplified in any way, to enable a portrait of the gene expression levels.

ChIP-SEQ . Chromatin Immunopresipitation is used to find relationships between proteins and DNA sequence. It is useful to find transcription factors and replication-related proteins.

Amplicon sequencing . Used primary to do barcoding of species. A known gene is amplified (i.e. 16S) with the intention of characterising the species present in the sample.

Metagenomic capture From a mixed sample (soil, root, animal fluids) all the DNA is extracted and sequenced, this gives a snapshot of the microbial community in the sample

RAD-seq Restriction site associated DNA markers are useful to do population analysis. The technique focus on sequencing regions around restriction sites and the variations around them can be used to genotype individuals.

Exon capture The DNA is extracted and baits are used to attract the regions with motif common around exons. This allows to sequence only the genes and regions near them.

The different sequencing technologies available as of 2013 have different yields, advantages and disadvantages, as described bellow:

Illumina Each fragment is amplified using bridge amplification over and over in the same place in the plate to form clusters. After the clusters are formed, a last cycle of amplification is carried on with the bases being added to the template, with the intervention of a polymerase, have a fluorescent marker which make the cluster glow depending on the added base. It adds one base per cycle. With a read length between 75bp and 250bp is currently the most widely adopted platform. The As a de facto standard, many tools exist to cope bioinformatically with the biases of the machine. The run takes 4 or 9 days, depending on days, depending if one or two reads are generated for each fragment. It produces up to 35 gigabases per run.

SOLiD The preparation of the fragments is similar to Illumina, however, when adding the bases they are added in pairs. This technique is called sequencing by ligation as it use a DNA Ligase, as opposed

to a polymerase, to determine the transition between bases. The resulting sequence is not in base space, but in colour space, which represents the transition state between bases. This technique is robust for finding SNPs when you have a good reference where to align the reads. However, the number of tools available and the research done to analyse sequences in colour space is low compared to the tools using base space. The runs take between one and two weeks to complete, with a yield of up to 50 gig abases per run. The read length can be up to 50 bases

Roche/454 The fragments are cloned in beads, which then fall in wells in the slide. The sequencing is done by adding nucleotides in a determined order. The next nucleotides to be added in the reaction contain a fluorescent marker. The bases are not added one by one, but all the bases that are the same are added together. The amount of glow on each well can tell how many times a base is added. As the glow is not a discrete number, when a long homopolymer appear (above 5 bases) the likelihood of having a wrong count of the homopolymer is increased. The average read length varies between 300 and 700bp. A run usually takes half a day, but it only yleds 0.45 gigabases. The cost of the reagents is relatively expensive, but if the experiment requires longer reads it is a good option.

PacBio Opposed to all the previous technologies, Pacific Biosciences has developed a sequencing technology where the molecules doesn't need to be PCR amplified before the sequencing. The glass slide used contains wells with a depth of 100nm where a polymerase lays at the bottom. The nucleotides to be added have a fluorescent marked that is freed when the polymerase adds the nucleotide, releasing a light signal, which then can be captured from the bottom of the glass. The error rate for this technology is still high (about 10% of the bases are miscalled), however reading several times the same molecule reduce the error rate. The main advantage is that the reads can be over 1kbp.

OpGen Additionally, high-througput optical mapping technologies, like OpGen, are becoming accessible. The maps are done by fixing single molecules of DNA are held on a slide. Then, restriction enzimes targeted to specific digestion sites cut the fragment and fluorescent

markers are added to the ends of the fragments. Finally, the fragments are visualised and the size of the molecules is measured by the distance between fluorescent points in the slide. This is done with several fragments at the same time. Then, the distances between restriction sizes can be compared across all the fragments to generate a consensus. Finally, if you have contigs from other technologies, it is possible to complement the information and get better assemblies. Even without the contigs, the data can be used to compare translocations within strains of different bacteria or homologous species at a chromosome level.

ION Torrent (Do some research on newer sequencing things)

1.5 Sequence analysis

This section discusses the criteria to decide analysis done after sequencing, when to do re-alignments or *de novo* assemblies, how to do SNP calling in diploid and polyploid organisims and the bulk frequency ratios.

DNA sequence alone is not alone to enough to understand the biology behind, a context is required. There are databases like Ensembl and NCBI that act as repositories of the known public sequences.

From the computational point of view, the problem can be viewed as a string matching. The Smith-WatermanSmith and Waterman (1981) and Needleman-WunschNeedleman and Wunsch (1970) algorithms are the gold standard interns of accuracy looking for similarity between sequences. However, the execution time for both of them is prohibitive to run in massive databases. The algorithm execution time is O(mn), as it requires calculating a matrix of size mn where m is the target sequence and n is the query sequence. To scale this to a manageable problem algorithms like BLAST index the references and use heuristics to make the search more manageable, with some penalty in the accuracy. This alignments tools are useful for long stretches of DNA (like cDNA or contigs)Altschul et al. (1990).

When looking at a protein level, where the sequences may be only loosely similar, Hidden Markov Models (HMM) are used to search for protein families. This can be useful to annotate putative proteins and

their functions. HMMs require a training dataset, where proteins are previously annotated and the reference is a model encoding the characteristics of a family, with associated probabilities. Hence, this technique is something between a sequences aligner and a classifierEddy (2004).

When analysing high-throughput sequencing, having millions of short sequences make unfeasible to try to align the data to every possible reference. However, one can take in advantage the fact that you know which organism you are looking for and, if available, use a genomic reference. For this, tools like MAQ, BWA, Bowtie, among others, provide indexed search. Once you have your reads aligned to a reference you can do more analysis, depending on the biological question being asked and the type of sequencing carried on. Fortunately, most of the Short-Read sequence alignment produce similar outputs and the SAM format is becoming a de facto standard. This is allowing to make more modularised downstream analysis where you can test different aligners with different settings and pick the algorithm that better fits your experimentLiu and Schmidt (2012); Li and Durbin (2009); Li et al. (2009).

1.5.1 Ambiguity Codes

Make a table with the ambiguity codes and why they are useful.

1.5.2 RNA-Seq

One way to narrow down which genes are involved in certain trait or response to the environment is to focus on studying only the expressed genes. One of the techniques involving high-throughput sequencing is RNA-Seq. This technique captures the messenger RNA in the tissue being studied and sequenced. The premise is that you will find a gene more expressed if it is being used by the organism. Some proteins with a vital role for the cell are always expressed (i.e. RuBisCO for carbon fixation in plantsGM (2000)). On the simplest of the experiments you would need two datasets to compare, one with the gene being looked expressed and one where it is not. The expression can come from different environmental conditions, development stage or different genotypes.Mortazavi et al. (2008)

Depending on how much *a priori* information of the analysed organism is available different bioinformatic approaches can be used.

Transcriptome alignment The reads are aligned to a database of known cDNA. Ideally, alternative splicing sequences are available, so a simple alignment should work (i.e. BWA, bowtie).

Genomic alignment The reads are aligned to the genome. The splice junctions, introns and axons need to be accounted, so simple alignment doesn't work. Regular alignments are used, but the reads may be trimmed at fixed sizes to allow discontinuous alignments using regular tools (i.e. Stampy, tophat/cufflinkns)

***De Novo* transcriptome assembly** If a reference of the organism is not available, it is possible to generate a draft transcriptome with the RNA-Seq reads with traditional assemblers (velvet, abyss) or with specialised assembler tools like Trinity.

Once you have the alignments it is possible to evaluate the relative expression of the genes in the sample calculating the Reads per Kilobase per Million mapped reads (RPKM) or the Transcripts per Million (TPM). This normalises the expression by the amount of sequenced data and can be used to find which genes change in expression volume across different samples.

1.6 Wheat online resources

A compilation of the currently available resource for whet genetics and genomics. MAS wheat, CeralsDB, Ensembl, etc.

Chapter 2

PolyMarker: A fast polyploid primer design pipeline

One of the main challenges of working with polyploid species is the design of genome specific molecular markers. This is particularly true when targeting conserved homoeologue regions, where a primer could bind to a pair,or triplet, of identical sequences. For that reason, designing primers for polyploids require to include bases that are specific to the target, in addition to the physicochemical properties of the primer. The traditional methodology to find primer candidates include a blast search and a local alignment, select the primer candidates manually, and finally, validate the primers with a tool, like `Primer3` (Rozen and Skaletsky, 2000). To reduce the time invested in designed primers I have developed PolyMarker (Ramirez-Gonzalez et al., 2015a), a pipeline to automate the primer design for polyploid organisms.

2.1 Pipeline

PolyMarker is an automated pipeline that takes as input a list of SNPs and a reference file and produces a list of primer triplets for SNP genotyping. The list of SNPs is first converted to a `FASTA` file with ambiguity codes (Cornish-Bowden, 1985) The template sequences are aligned with `exonerate` (Slater and Birney, 2005) to find the homoeologous regions to the target sequence. Then, the alignment between homoeologues is refined using `MAFFT` (Katoh and Standley, 2013). A list of candidate variations is produced and used as input for `Primer3` (Rozen and Skaletsky, 2000). Finally, the output of `Primer3` is parsed to find the best

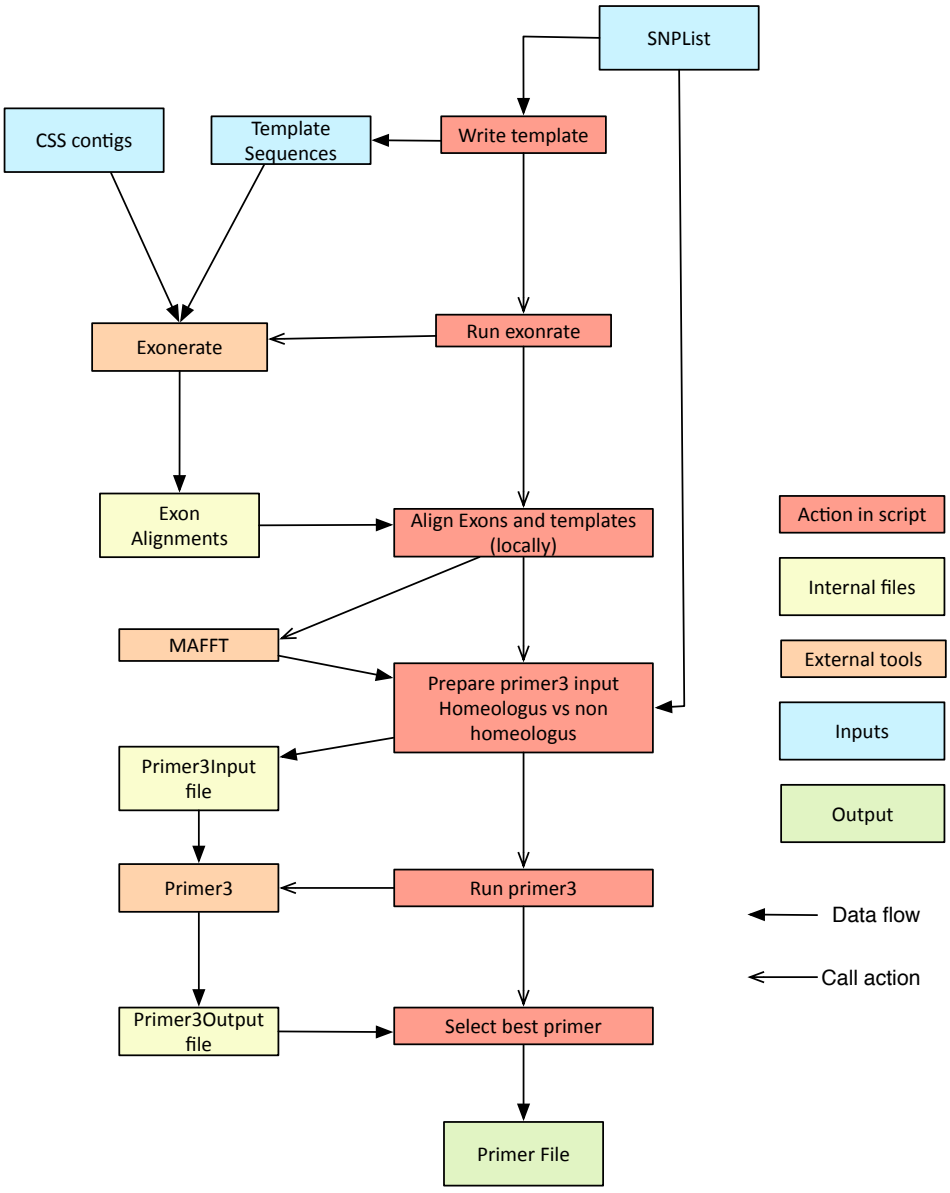


Figure 2.1: Steps and tools called by PolyMarker. The colour of the boxes represent: the step is an action inside the script(red); actions of the script(orange); temporary files(yellow); inputs(blue) and; outpus(green)

primer pair that contains a the targeted SNP and a base that is specific to the target genome (Figure 2.1). The pipeline is written as a Ruby script, using parsers and wrappers from BioRuby (Goto et al., 2010) and bio-samtools (Etherington et al., 2015; Ramirez-Gonzalez et al., 2012). The software is open source and released as a biogem (Bonnal et al., 2012), `bio-polyploid-tools`, the source code is available in github: `https://github.com/TGAC/bioruby-polyploid-tools`.

The PolyMarker input consist on SNP list with: unique name for the marker, the target chromosome and the sequence for the marker. The alternative alleles are surrounded by square brackets within the sequence. PolyMarker can take a list of several markers and design them in batch (Figure 2.2a). A `FASTA` file is produced with all the template sequences, with the alternative alleles substituted by the IUAPC ambiguity codes (Cornish-Bowden, 1985). The flanking sequence surrounding the SNP is limited by default to 100bp to reduce the search time and avoid missing regions that diverge near the SNP, as when the variation is near an intron-exon junction.

The template sequences are aligned to the reference using `exonerate` (Slater and Birney 2005; Figure 2.2b). The alignment is refined with the `--model est2genome` option, to allow the search of sequences coming from transcripts, a common source of SNPs (Allen et al., 2011). The exonerate output is formatted with the `--ryo` (roll your own format) to get an output easy to parse. All the hits that contain the SNP are extracted from the reference with a flanking sequence that extend out of the hit, by defualt, to 100bp on each side of the SNP, Figure 2.2c. The size of the flanking sequence can be set to different sizes to allow the design of different types of primers. Different homoeologues may contain small indels, Figure 2.2d. To enable a comparasion base-per-base, a local alignment with `MAFFT` (Katoh and Standley, 2013) is produced, Figure 2.2e.

PolyMarker searches across each base in the local alignment to identify the variations across homoeologues and the target marker. A mask is produced to highlight the bases with a variations, Figure 2.2f, on the following categories:

Specific — Homoeologous polymorphism which is only present in the target genome (upper case).

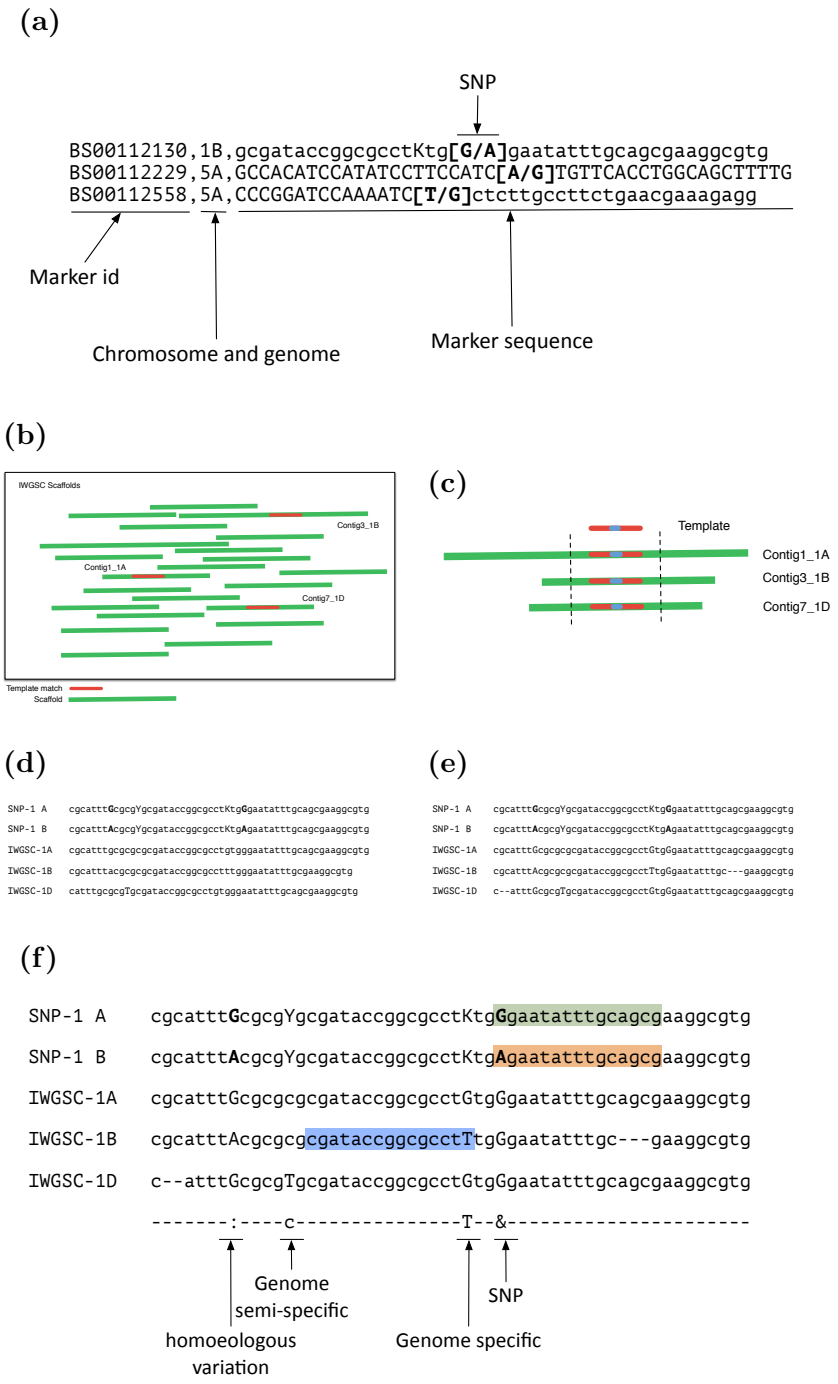


Figure 2.2: Alignments done by PolyMarker.(a) input. The alternative alleles are sorrounded by brackets. (b) Global search of templates in the reference contigs. (c) Selected regions around the SNP on every chromosome. (d) Sequence of found regions around the SNP. (e) Local alignment on regions around the SNP detects indels. (f) Alignment with mask and primer candidates.

Semi-specific — Homoeologous polymorphism which is found in 2 of the 3 genomes, hence it discriminates against one of the off-target genomes or when not all the homoeologous sequences were found (lower case).

Non-specific — No variation is found across homoeologues (-).

Homoeologous — The target SNP is present across different chromosomes, so candidate SNP markers on this category are not expected to be reliably identify the allele (:).

Non-homoeologous — The target SNP is not present across chromosomes, so it can be used to identify an allele (&).

PolyMarker was designed to produce SNP assays for KASP genotyping (LGC Genomics, 2013), which requires a common primer and two allele-specific primers. The common primer is selected to start on a position from a: Specific; Semi-specific or; Non-specific, on that priority. This means that the common primer will be as specific as possible in the region. For the allele-specific primers, the starting position of the primer is on the base with the SNP. To ensure that the stability of the candidate primers will be met, the putative starting positions are tested with `Primer3` (Rozen and Skaletsky, 2000).

PolyMarker was designed and validated with the markers described in section 3.7. For wheat, PolyMarker uses the contigs from Mayer et al. (2014), as deposited in Ensembl. As new releases of the wheat genome are made available, different parsers to assign the chromosome to each sequence can be added with little effort to PolyMarker.

2.2 Applications of PolyMarker

PolyMarker is not restricted to wheat or to KASP assays, the source code is flexible and can be extended for other types of analysis. On each of the following projects, PolyMarker has been adapted to design primers in species where KASP hasn't been used before, the primers are used for regular PCR amplification, or the use of KASP is not the conventional SNP calling.

2.2.1 KASP assays for public sets of SNPs

PolyMarker was used to design KASP assays for the 81,587 markers from (Wang et al., 2014), available on the PolyMarker website and in CeralsDB (Wilkinson et al., 2012). Of those markers, 40,267 where designed using the target chromosome using the genetic map published by the genetic map. Genes without a genetic position were aligned to scaffolds sorted by chromosome from the International Wheat Genome Sequencing Consortium (Mayer et al., 2014) with BLAT (Kent, 2002) and the best hit was selected as putative location. 97.5% of the assays where designed and 76% of them are semi-specific or specific, thereby improving their expected performance with respect to randomly designed primers (Table A.1). A subset of the designed assay was used to genotype a mapping population to find resistance to Fusarium head blight (Burt et al., 2015).

2.2.2 SNPs in a mutant population

PolyMarker was used to design primers to validate SNPs in a Targeted Induced Local Lesions in Genomes (TILLING) population, an approach to identify the function of genes by mutating them. To calibrate the SNP calling, KASP assays were designed to get the mutations from M_2, M_3 and, M_5 mutants (King et al., 2015). Then primers were designed for the whole mutant population, consisting of 1,200 Cadenza (Hexaploid) and 1,535 Kronos (Tetraploid) wheat lines (Krasileva et al., submitted 2016). Genome-specific primers 172 and 80 SNP assays on 19 and 8 M_4 Cadenza and Kronos lines respectively. Of those, 71(85.5%) Kronos and 147(88.8%) of the Cadenza primers where valid assays (Tables A.4 and A.5).

2.2.3 Deletions on a mutant population

On some of the TILLING mutant lines long deletions were detected (Krasileva et al., submitted 2016). To validate the deletions is possible to use KASP assays to produce primers that amplify homoeologues. PolyMarker was modified to search for variations across homoeologues to select a common primer that will amplify two genomes (Figure 2.3a, b). On lines without the targeted deletion, the amplification correspond to an homozygous assay (Figure 2.3c). When a deletion is present the results of the assay look like an homozygous sample, with the intensity

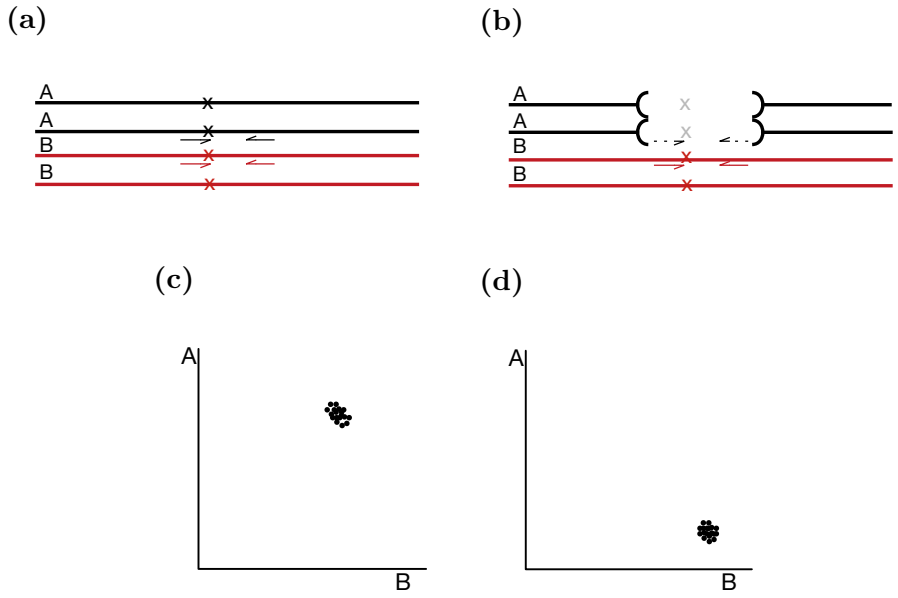


Figure 2.3: KASP assays to validate homozygous deletions. (a) Primer positions for wildtype. (b) Primer positions on homozygous deletion on M_4 (c Heterozygous amplification on wildtype, including both homoeologues. (d) Homozygous amplification on deletion line, only the non-deletad homoeologue is amplified.

of the assay towards the the conserved homoeologue (Figure 2.3d). A set of KASP assays for the the deletions and mutations located on the same chromosome where designed to validate 11 homozygous deletions on M_4 plants. In all cases the segregation of the mutations was as expected, except for a predicted heterozygous mutation that was called as homozygous. Also, all the KASP assays that contained a deletion were called homozygous, as expected. To ensure that the calls didn't come from a single cluster, 4 wildtype plants were genotyped and the markers for deletions where called as heterozygous. An example of a validated deletion, with the calls for each individual is shown on Table A.3.

2.2.4 PolyMarker public web service

To make PolyMarker accessible to the community, a web server that allow the submission of SNPs was developed. The web interface consists on two virtual machines, one with a web facing interface that stores the queries, and a dedicated node to submit jobs to an HPC cluster. The on-line interface further simplifies the design of KASP assays, a process

that used to take a couple of weeks now is done in a couple of hours. Since the release of the public service in July 2014 until August 2016, 1,739 requests to PolyMarker have been done.

2.2.5 Genotyping of *Puccinia striiformis* f. sp. *tritici* isolates.

In Hubbard et al. (2015), *Puccinia striiformis* f. sp. *tritici* (PST) isolates were sequenced and assigned to clusters, according to their genotype. The clusters are useful to monitor the changes in the pathogen population, which can be used to predict if certain wheat lines will be resistant to the isolates in the field. PolyMarker was used to design primers for PST, using the assembly PST-130 (Cantu et al., 2011). Out of 15 assays 11 can be used to identify to which cluster of isolates a sample is likely to belong, Supplemental Table A.2.

2.3 Discussion

PolyMarker is a tool that was born as part of the validation of the SNPs found in Chapter 3. Originally, the primer design was ought to be done manually, a slow, error-prone and, repetitive process. The steps require the use of several bioinformatics tools, but once I figured out the steps I decided to automate the process. Since designing genome-specific primers is a common task in wheat research and breeding, the community showed interest on the tool and I decided to refine it and make it open source. PolyMarker has been used successfully in several projects and it even allowed the novel use of KASP assays to validate long deletions in polyploids.

The current web interface of PolyMarker is limited to KASP assays, however the command line version is more flexible and has been used to design primers for PCR amplicons, capillary sequencing and on other organisms. The ideas behind PolyMarker had been taken by other projects like the scripts described in Ma et al. (2015) and the corresponding web interface, GSP (Wang et al., 2016). As new references of wheat come available, PolyMarker should be updated to work with pseudomolecules and the web interface updated accordingly.

Chapter 3

Genetic map of *Yr15* with RNA-Seq

Wheat breeding programs aim to improve the wheat lines available for production. One of the traits desired in an elite line is the resistance to pathogens, such as *Puccinia striiformis* f. sp. *tritici*, the fungi responsible of yellow rust. A source of resistance genes is are introgressions from other species, such as *Triticum diccocides*. In the University of Sydney a collection of Near Isogenic Lines (NILs) with introgressions to several Yellow Rust resistance genes on a susceptible background were developed (Wellings and McIntosh, 1998). On this chapter the NIL for the *Yr15* locus is used to produce a mapping population to improve diagnostic markers.

Line selection can be done with molecular markers that can be used to test if certain allele is present in a line, without the need to do a phenotype. To find which regions are linked to a trait the use of F_2 mapping populations is a common practice. The population is produced by crossing two homozygous parents (P_1 and P_2) with different alleles, A/A (dominant) and a/a (recessive). When the trait is dominant and has a mendelian segregation, the F_1 population show the dominant trait, as it has a copy of each allele (A/a). The F_1 is then self-crossed to and the population segregates with a ration 1:2:1, dominant:heterozygous:recessive respectively. This generates a population with a phenotype ratio of 3:1 (dominant:recessive), since the effect of the recessive allele is masked by the dominant gene (Van Ooijen and Jansen 2013; Figure 3.1).

Bulk Segregant Analysis (BSA) consists on pooling the DNA of individuals with contrasting phenotypes (Michelmore et al., 1991) on a

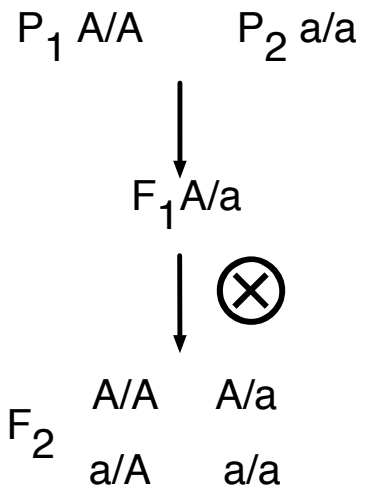


Figure 3.1: The cross of two homozygous parents, P_1 and P_2, with a dominant and a recessive allele of a gene produce an heterozygous F_1. The F_1 crossed with itself produce a segregating F_2 population with a 1:2:1 ratio (A/A:A/a:a/a). The upper and lower cases represent dominant and recessive alleles

segregating population. The bulks show as heterozygous except for the region that is linked to the trait of interest. This approach can be used to identify SNPs using High Throughput Sequencing, such as: exome capture (Hodges et al., 2007), RNA-Seq (Pickrell et al., 2010), whole genome resquencing (Schneeberger et al., 2009), among others.

To Call for SNPs from RNA-Seq a reference transcriptome is used as target when aligning the reads. The Bulk Frequency Ratio (BFR) methodology can work on organisms has more than one pseudo genome and that the genes are not necessarily fully characterised independently among homoeologues or paralogues, you can have in a single reference collapsing similar regions. The UniGenes database, from NCBI, contains the genes of each species, with all the variations of each gene automatically collapsed and represented as with the longest cDNA (Pontius et al., 2002). The UCW genes described in Krasileva et al. (2013) contains 94,177 models from tetraploid and hexaploid wheat, assembled and phased to separate different homoeologues. Both gene sets are complement each other, however, the UCW gene models should provide an improved alignment, since the different homoeologues aren't merged in a single model, a possible side effect of the UniGene pipeline.

Homoeologous variants, as exemplified by the G>T variant at position 181; K in consensus (Figure 3.2), will produce the same ambiguity code for both parental consensus sequences and can therefore be excluded. Real allelic SNPs between the parental genotypes, exemplified by the G>A variant at position 184; R in consensus, are distinguished by the presence in one, but not the other parental consensus sequence. The

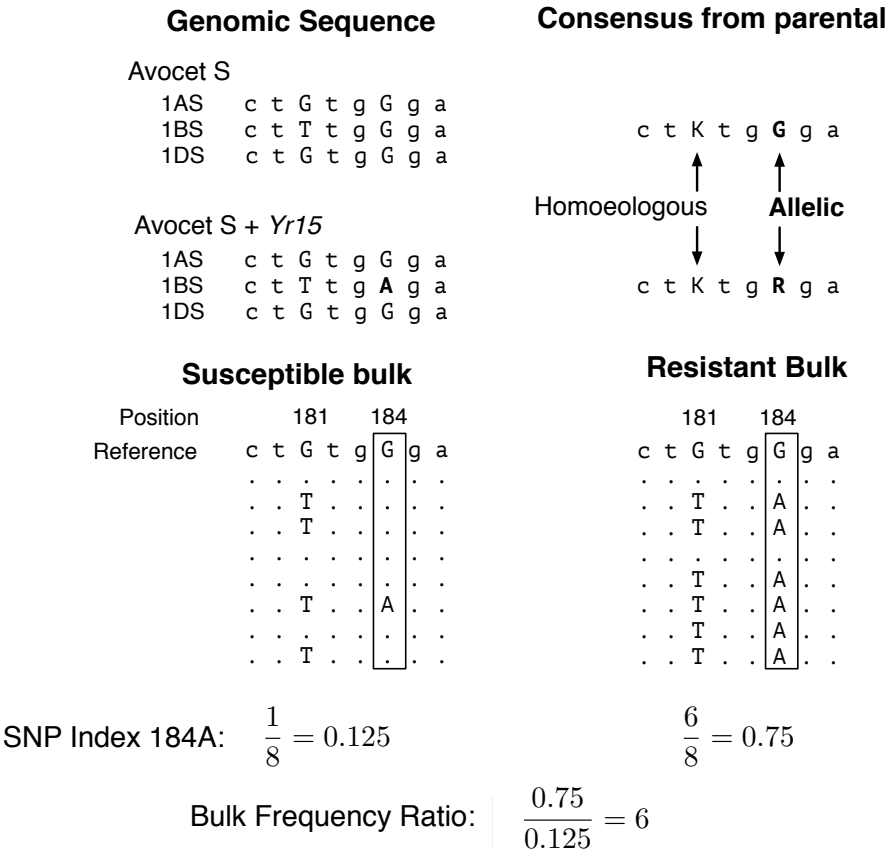


Figure 3.2: Illustration of a non-informative homoeologous SNP (G181T) present in both parental lines, and an informative allelic SNP (G184A), only present in the resistant progenitor Avocet S + Yr15. The consensus sequences from the parental genotypes include this information in the form of ambiguity codes (K and R, respectively). In the bulks, the individual reads align across the reference sequence, with matches indicated by dots, and polymorphisms at positions 181 and 184 indicated by the corresponding nucleotide variants at those positions. The SNP index is calculated as the frequency of the informative allelic SNP in each bulk. The Bulk Frequency Ratio is the quotient of the resistant and susceptible bulk SNP Indexes. Figure previously published in Ramirez-Gonzalez et al. (2015b).

allelic SNPs are then examined further with the alignments of the bulks to identified the SNPs that are enriched on the resistant plants. The SNP index is the proportion of times an alternative allele is observed over the coverage at certain, in the example the the susceptible bulk has an SNP index of $1/8 = 0.125$ and $6/8 = 0.75$ for the resistant bulk (Takagi et al., 2013). traditional The BFR are then calculated by dividing the SNP Index of sample containing the target phenotype (resistance) over the sample without the trait (susceptible), on the example is $0.75/0.125 = 6$. A high BFR suggests that the SNP is linked to the target trait (Trick et al., 2012).

Finally, the best candidate SNPs where selected to produce a genetic map which lead to a triplet of markers diagnostic to the target locus.

The steps described in this chapter were first published in Ramirez-Gonzalez et al. (2015b) and the results of this chapter are published in Ramirez-Gonzalez et al. (2015c).

3.1 Mapping population

The population was developed by crossing the resistant line Avocet + *Yr15* (*Yr15*) (Wellings and McIntosh, 1998), Figure 3.3a, to the susceptible line Avocet S (AVS), Figure 3.3b. *Yr15* is a NIL of a 6th generation Back-cross (BC) and the AVS background is highly succeptible to yellow rust, hence the resistance is coffered by the *Yr15* locus. F_2 seeds from tree independent F_1 plants where sown and tissue was collected, before the fungal inoculation to avoid the effect of the response on the gene expression. The plants were challenged at the three leaf stage as it is know that *Yr15* confers resistance in seedlings (Gerechter-Amitai et al., 1989). The expected segregation on an F_2 population is 3:1 (resistant:susceptible), since *Yr15* is a dominant gene. From the 232 plants in the F_2 population that germinated, 187 were resistant and 45 were susceptible, which deviates slightly from the expected ratio ($\chi^2 = 0.049$). Segregation distortion has been shown for the same *Yr15* donnor (Randhawa et al., 2009), however the decresed number of succeptible plants can be explained by escapes in the virulence essays (i.e. plants scored as resistant without the *Yr15* locus). For this study we extracted DNA from individual plants in the F_2 population and we bulked RNA on 6 different bulks: 3 resistant and, 3 succeptible (Figure 3.3c).

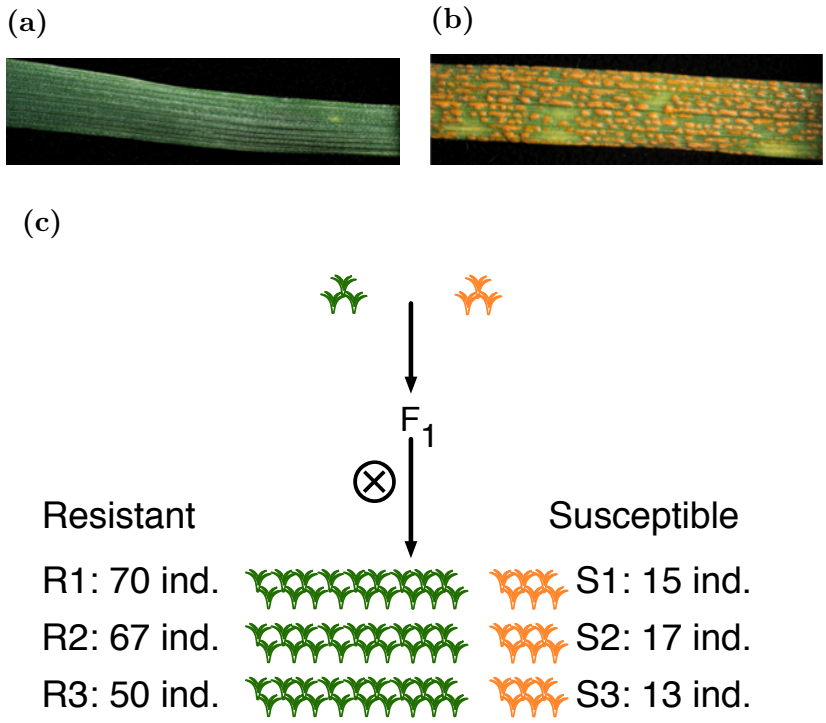


Figure 3.3: Response of (a) Avocet + *Yr15* and (b) Avocet when inoculated with *Puccinia striiformis* f. sp. *tritici* at the three leaf stage. (c) The phenotype of the F_2 population was used to produce 6 bulks, 3 resistant and 2 susceptible. The RNA was pooled in bulks accordingly. Adapted from (Ramirez-Gonzalez et al., 2015c)

Table 3.1: Arrangement and number of sequenced base pairs per sample.

Library	name	Bar code	Lane	Reads ($\times 10^8$ bp)
LIB1715	Bulk R1	ATCACG	1	0.77
LIB1716	Bulk R2	TAGCTT	1	1.20
LIB1717	Bulk R3	ACTTGA	2	0.96
LIB1718	Bulk S1	GGCTAC	2	1.64
LIB1719	Bulk S2	CGTACG	2	1.49
LIB1720	Bulk S3	GTGGCC	1	1.88
LIB1721	AvocetS	N/A	3	4.13
LIB1722	AvocetS + *Yr15*	N/A	4	3.99

3.2 Sequencing and mapping

RNA-Seq was used to avoid sequencing the non-coding regions and reduce the search space. The sequencing of the bulks and the parents were done on a single Illumina Hi-Seq2000 each. The bulks were multiplexed and sequenced on a third of a lane each, as shown on Table 3.1. To ensure that the quality of the sequencencing was good, `fastqc-0.10` (Babraham Bioinformatics, 2012) was run with its default parameters in each one of the fastq files. The GC content was around 52% in all the samples (Appendx B.2), which is expected as the sample should be of coding regions, and for wheat the reported GC content in genes is around 55%. The quality of the reads is fairly consistent, in general dropping after the base 80 across the samples (Appendix B.1).

When the analysis was started, the draft genome and the corresponding annotation where not not release yet, hence gene models where used. All the samples where aligned to the Unigenes v60 (56,954 genes) and the gene models from UCW (Krasileva et al., 2013) using `BWA 0.5.9` (Li and Durbin, 2009). The alignment provided showed that a few genes were overly expressed, however we still have have 22,107 and 36,808 genes, on the Unigenes and the UCW gene set respectivley, with a coverage greater than 20x in the progenitor with *Yr15*. Both gene sets performed similarly in terms of the percentage of genes with reads and percentage of aligned reads. For AVS and *Yr15*, the percentage of genes with a coverage of at least $20x$ is 45% and 39% respetively across both references (Figure 3.4a). Since each individual bulk has a lower coverage, the susceptible and resistant reads were merged *in silico* as: (i) susceptible bulks 1 with 2 (S1 + S2) and resistant bulks 1 with 2 (R1 + R2) and (ii) all the

Table 3.2: Number of genes with a coverage over 20x, 10x and at least one read (>0x).

		Bulks						Bulk mixes				Progenitors	
Coverage	Reference	R1	R2	R3	S1	S2	S3	R1+R2	S1+S2	R1+R2+R3	S1+S2+S3	*Yr15*	AVS
20x	UCW	16,434	27,871	27,223	32,287	28,669	34,898	33,968	41,019	40,985	47,507	36,808	42,248
		17%	30%	29%	34%	30%	37%	36%	44%	44%	50%	39%	45%
	UniGene v60	9,643	16,182	15,222	19,549	17,397	20,567	20,219	25,270	24,598	29,052	22,107	25,842
		17%	28%	27%	34%	31%	36%	36%	44%	43%	51%	39%	45%
10x	UCW	27,371	38,282	37,777	42,658	38,999	44,610	43,266	49,473	49,182	54,781	46,356	50,760
		29%	41%	40%	45%	41%	47%	46%	53%	52%	58%	49%	54%
	UniGene v60	16,201	22,948	22,130	26,200	24,130	26,914	26,318	30,579	29,857	33,557	28,044	31,095
		28%	40%	39%	46%	42%	47%	46%	54%	52%	59%	49%	55%
>0x	UCW	68,302	72,484	72,957	74,694	73,290	75,201	74,397	77,093	76,715	78,796	76,275	77,080
		73%	77%	77%	79%	78%	80%	79%	82%	81%	84%	81%	82%
	UniGene v60	40,717	42,489	42,595	43,625	43,059	43,748	43,393	44,655	44,364	45,392	43,732	44,596”
		71%	75%	75%	77%	76%	77%	76%	78%	78%	80%	77%	78%

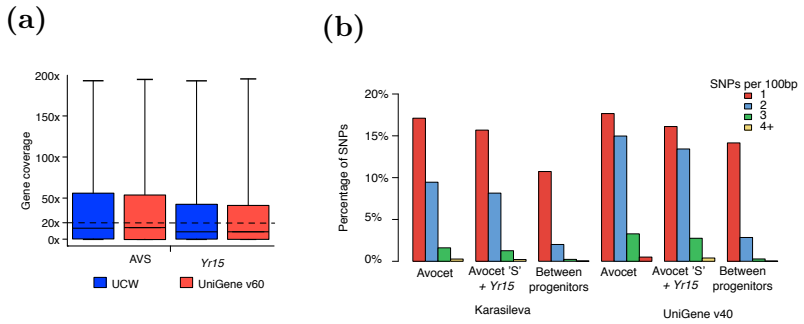


Figure 3.4: (a) Box plot distribution of the gene coverage of the parent reads (AVS and *Yr15*) across the UCW (blue) and the UniGene (red) gene models. The dashed line represents the 209 minimum coverage required for SNP calling. The full line represents the average coverage across all gene models. (b) Percentage of genes exhibiting SNPs across references. The number of SNPs between the parent reads and the corresponding references was calculated (per 100 bp, rounded). The between-parents category corresponds to putative SNPs when comparing the consensus sequence between AVS and *Yr15* Adapted from Ramirez-Gonzalez et al. (2015c)

susceptible (S1 + S2 + S3) and resistant bulks (R1 + R2 + R3). The merged samples increased the percentage of genes with coverage over 20x to 44% and 50% in the resistant and susceptible bulks (Table 3.2), which is close to the coverage from the progenitors.

3.3 SNP Calling

The SNP calling was done on positions with a coverage of at least $20x$ on the progenitor lines against the gene reference. The AVS progenitor had roughly 3% more genes with polymorphisms than *Yr15*, consistent with the difference in coverage, suggesting that with a higher coverage we could recover more SNPs from *Yr15*. The UniGenes have a higher number of SNPs because the UCW gene models have a higher number of monomorphic genes when compared to the UniGenes. (Figure 3.4b; Table 3.3). The difference in the number of relative monomorphic SNPs between references can be explained by the fact that the UniGenes have homoeologues can be represented as a single sequence, as opposed to the UCW set which are homoeologue-specific, improving the mapping to the correct homoeologue in the genes from the UCW set over the UniGenes.

Both gene sets were done from varieties different to AVS and are likely to be incomplete, hence we set a low threshold of at least 20% of the

Table 3.3: Count of SNPs per 100 bp on genes with at least 20x coverage.

SNPs per 100bp	UCW			UniGene v60		
	AVS	AVS+ *Yr15*	Between progenitors	AVS	AVS+ *Yr15*	Between progenitors
0	67, 389 71.6%	70,338 74.7%	81,921 87.0%	36,210 63.6%	38,339 67.3%	47,097 82.7%
1	16,111 17.1%	14,770 15.7%	10,107 10.7%	10,058 17.7%	9,175 16.1%	8,061 14.2%
2	8,904 9.5%	7,676 8.2%	1,893 2.0%	8,529 15.0%	7,648 13.4%	1,621 2.9%
3	1,517 1.6%	1,192 1.3%	215 0.2%	1,870 3.3%	1,568 2.8%	59 0.3%
4+	253 0.3%	198 0.2%	38 0.0%	287 0.5%	224 0.4%	16 0.0%

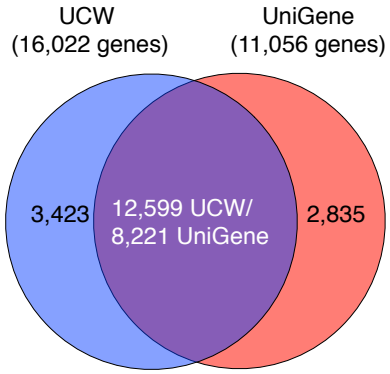


Figure 3.5: Gene models with putative SNPs in common between the UCW and UniGenes reference. The intersection represnts the genes that are common in both sets. Adapted from Ramirez-Gonzalez et al. (2015c)

observed nucleotides on any position to call an SNP. To represent cases were more than one consensus base is called we use International Union of Pure and Applied Chemistry (IUPAC) codes (Cornish-Bowden (1985); Section 1.5.1; Figure 3.2). To focus the analysis on informative SNPs, the common varietal SNPs and variations between homoeologues were removed by finding the cases when the consensus call on both progenitors is the same. The SNPs that are unique to a single parental were examined in detail. There are 66,426 putative SNPs across 16,022 (17%) UCW genes and 52,262 SNPs on 11,056 UniGenes (19.4%; Figure 3.5).

The high number of genes with SNPs was unexpected as a BC6 NIL used for an F_2 mapping population expects to have $< 1\%$ of the genetic background segregating. The both sets of gene models were aligned with BLAT (Kent, 2002) to the Chinese Spring Chromosome arm survey sequence (CSS; Mayer et al. 2014); the alignment resulted on 80,031 (85.0%) UCW gene models and 41,118 (72.2%) UniGenes assigned to a

Table 3.4: Number of genes with SNPs assigned to the wheat chromosome arm CSS scaffolds (Mayer et al., 2014) using the best hit from BLAT (Kent, 2002)

Wheat Chromosome Arm	UCW (94,177)	UniGene v60 (56,954)	Total (151,131)
1AL	3,251 (3.45%)	1,404 (2.47%)	4,655 (3.08%)
1AS	1,366 (1.45%)	560 (0.98%)	1,926 (1.27%)
1BL	2,610 (2.77%)	1,280 (2.25%)	3,890 (2.57%)
1BS	1,487 (1.58%)	693 (1.22%)	2,180 (1.44%)
1DL	997 (1.06%)	1,057 (1.86%)	2,054 (1.36%)
1DS	753 (0.80%)	687 (1.21%)	1,440 (0.95%)
2AL	3,491 (3.71%)	1,460 (2.56%)	4,951 (3.28%)
2AS	2,305 (2.45%)	974 (1.71%)	3,279 (2.17%)
2BL	3,658 (3.88%)	1,546 (2.71%)	5,204 (3.44%)
2BS	2,790 (2.96%)	1,139 (2.00%)	3,929 (2.60%)
2DL	1,098 (1.17%)	1,069 (1.88%)	2,167 (1.43%)
2DS	796 (0.85%)	833 (1.46%)	1,629 (1.08%)
3AL	2,135 (2.27%)	978 (1.72%)	3,113 (2.06%)
3AS	1,543 (1.64%)	718 (1.26%)	2,261 (1.50%)
3B	6,559 (6.96%)	2,839 (4.98%)	9,398 (6.22%)
3DL	915 (0.97%)	938 (1.65%)	1,853 (1.23%)
3DS	412 (0.44%)	450 (0.79%)	862 (0.57%)
4AL	3,393 (3.60%)	1,335 (2.34%)	4,728 (3.13%)
4AS	2,011 (2.14%)	817 (1.43%)	2,828 (1.87%)
4BL	2,119 (2.25%)	898 (1.58%)	3,017 (2.00%)
4BS	1,946 (2.07%)	892 (1.57%)	2,838 (1.88%)
4DL	1,069 (1.14%)	945 (1.66%)	2,014 (1.33%)
4DS	800 (0.85%)	699 (1.23%)	1,499 (0.99%)
5AL	2,640 (2.80%)	1,132 (1.99%)	3,772 (2.50%)
5AS	963 (1.02%)	407 (0.71%)	1,370 (0.91%)
5BL	5,324 (5.65%)	1,943 (3.41%)	7,267 (4.81%)
5BS	1,360 (1.44%)	591 (1.04%)	1,951 (1.29%)
5DL	2,067 (2.19%)	1,688 (2.96%)	3,755 (2.48%)
5DS	620 (0.66%)	614 (1.08%)	1,234 (0.82%)
6AL	2,397 (2.55%)	896 (1.57%)	3,293 (2.18%)
6AS	2,285 (2.43%)	936 (1.64%)	3,221 (2.13%)
6BL	1,564 (1.66%)	820 (1.44%)	2,384 (1.58%)
6BS	1,308 (1.39%)	731 (1.28%)	2,039 (1.35%)
6DL	1,399 (1.49%)	1,050 (1.84%)	2,449 (1.62%)
6DS	870 (0.92%)	680 (1.19%)	1,550 (1.03%)
7AL	1,918 (2.04%)	849 (1.49%)	2,767 (1.83%)
7AS	1,717 (1.82%)	764 (1.34%)	2,481 (1.64%)
7BL	1,592 (1.69%)	776 (1.36%)	2,368 (1.57%)
7BS	1,239 (1.32%)	713 (1.25%)	1,952 (1.29%)
7DL	2,040 (2.17%)	1,301 (2.28%)	3,341 (2.21%)
7DS	1,224 (1.30%)	1,016 (1.78%)	2,240 (1.48%)
Assigned	80,031 (84.98%)	41,118 (72.20%)	121,149 (80.16%)

Table 3.5: Total number of SNPs scored in parents, individual bulks and in silico merged bulks.

Gene set	$\frac{R1}{S1}$	$\frac{R2}{S2}$	$\frac{R3}{S3}$	$\frac{R1+R2}{S1+S2}$	$\frac{R1+R2+R3}{S1+S2+S3}$	SNPs in parents
UCW	16,269 24.49%	29,703 44.72%	31,891 48.01%	44,224 66.58%	64,522 97.13%	66,426
UniGene v60	15,261 29.20%	25,143 48.11%	24,548 46.97%	35,698 68.31%	49,738 95.17%	52,262

chromosome arm (Table 3.4). The SNPs found in the mapped genes are evenly distributed across all the chromosomes (Figure 3.7a), suggesting that the Avocet S (JIC, UK) used as parent in the F_2 is different to the Avocet S used for the *Yr15* NIL development (University of Sydney, Australia).

To confirm that the Avocet S seed stocks from JIC are distinct to the stocks in Sydney DNA from both stocks was procured and compared with the iSelect 90k wheat SNP chip. Between two independent Avocet S seeds from JIC only 58 out of 71,972 (0.08%) valid assays were polymorphic. Nonetheless, ther are over 5,000 ($> 7.5\%$) assays with polymorphisms between JIC-Avocet S and Avocet S from Sydney. The different was not expected originally, but considering that the Avocet S seeds are coming from different stocks and the fact that in both countries commercial varieties with the same name had been released, it is not surprising.

3.4 Bulk Frequency Ratios

The objective was find the SNPs enriched on each bulk and hence linked to the phenotype, variations from *Yr15* to resistance and from AVS to susceptibility in the segregating population. Across individual bulks, it was possible to score between 15,261 (24.5%) to 31,891(48.0%) SNPs across both reference sets. On the *in silico* mixes over 95% of SNPs where scored (Table 3.5), suggesting that the coverage of individual bulks is not enough to score all the SNPs. The scoring was done with the Bulk Frequency Ratio (Trick et al. 2012;Figure 3.2; Section 3.8.2), which has a value that increases as the *Yr15* allele is observed more times relatively to the AVS allele.

When increasing the minimum BFR threshold, enrichment of SNPs was observed in the short arm of the group 1 chromosomes (1S). Without

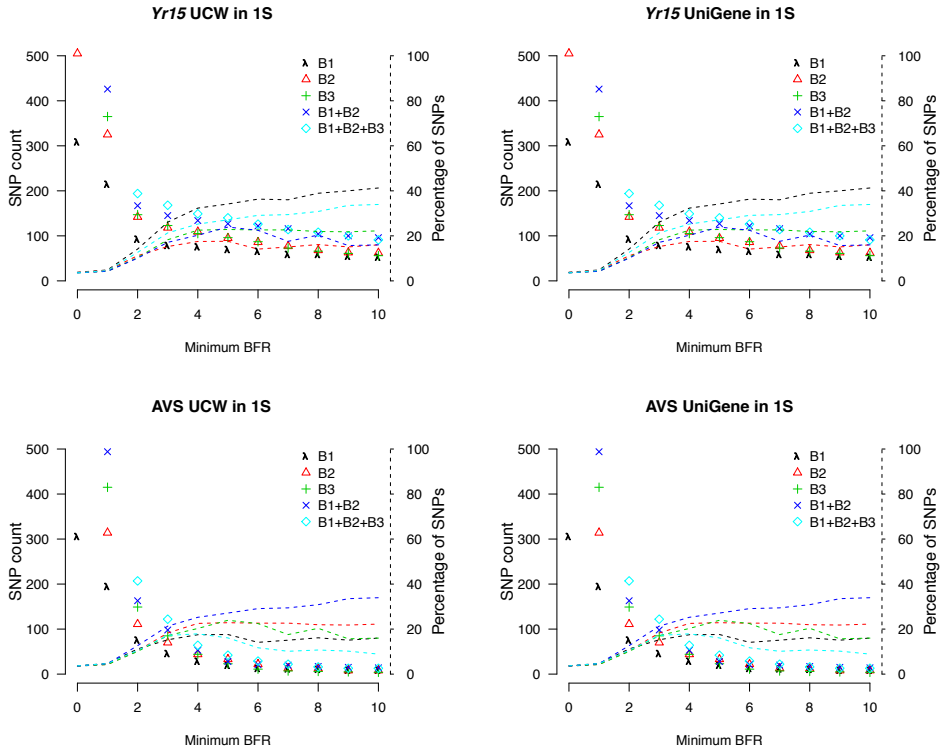


Figure 3.6: Effect of BFR threshold on the number of SNPs across the short arm of chromosome group 1. Figure previously published in Ramirez-Gonzalez et al. (2015c)

taking in account the BFR, 3.6% of the SNPs are located in the 1S group, similar to the number of SNPs located in other groups 3.4. However, when increasing the threshold (between $BFR > 5$ and $BFR > 7$) the relative number of SNPs in group 1S increases. After $BFR > 7$ the gains in relative enrichment only improves marginally, but the number of called SNPs is reduced (Table 3.6; Figure 3.6). For that reason, SNPs with a $BFR > 6$ were selected for further validation. The method described by Trick et al. (2012) was extended by including cases where there is a complete lack of coverage in one of the samples ($BFR = \infty$), which is an ideal case where the linkage between the SNP and the phenotype is perfect. A total of 1,582 SNPs across 1,173 genes had a $BFR > 6$.

3.5 *In silico* mapping

Mapping of the gene models to the IWGSC CSS Mayer et al. (2014) reference and the location of the SNPs using the genetic map from Wang et al. (2014).

Table 3.6: SNPs in chromosome group 1S vs total number of SNPs with a minimum BFR from 0 to 10.

Min BFR	Gene Set	R1/S1 Yr15	R1/S1 AVS	R2/S2 Yr15	R2/S2 AVS	R3/S3 Yr15	R3/S3 AVS	S1+2/ R1+2 Yr15	S1+2/ R1+2 AVS	S1+S2+S3/ R1+R2+R3 Yr15	S1+S2+S3/ R1+R2+R3 AVS
0	UCW	308/8,049 (3.83%)	305/8,220 (3.71%)	505/14,121 (3.58%)	556/15,582 (3.57%)	532/14,875 (3.58%)	623/17,016 (3.66%)	670/18,760 (3.57%)	885/25,464 (3.48%)	860/24,026 (3.58%)	1,505/40,496 (3.72%)
	UniGene v60	307/7,823 (3.92%)	299/7,438 (4.02%)	428/12,409 (3.45%)	421/12,734 (3.31%)	427/12,050 (3.54%)	415/12,498 (3.32%)	536/15,672 (3.42%)	595/20,026 (2.97%)	712/19,358 (3.68%)	901/30,380 (2.97%)
1	UCW	214/4,415 (4.85%)	194/4,108 (4.72%)	325/7,603 (4.27%)	314/7,374 (4.26%)	365/7,920 (4.61%)	415/8,850 (4.69%)	426/10,122 (4.21%)	494/12,185 (4.05%)	539/13,037 (4.13%)	842/19,466 (4.33%)
	UniGene v60	207/4,474 (4.63%)	194/3,630 (5.34%)	269/6,649 (4.05%)	269/6,193 (4.34%)	279/6,511 (4.29%)	272/6,436 (4.23%)	329/8,704 (3.78%)	369/9,343 (3.95%)	446/10,860 (4.11%)	541/14,226 (3.80%)
2	UCW	92/651 (14.13%)	75/671 (11.18%)	142/1,372 (10.35%)	111/1,101 (10.08%)	147/1,162 (12.65%)	149/1,411 (10.56%)	167/1,324 (12.61%)	163/1,478 (11.03%)	194/1,370 (14.16%)	207/1,765 (11.73%)
	UniGene v60	77/568 (13.56%)	58/527 (11.01%)	101/1,017 (9.93%)	81/720 (11.25%)	105/775 (13.55%)	84/867 (9.69%)	122/991 (12.31%)	116/973 (11.92%)	145/1,030 (14.08%)	132/1,210 (10.91%)
3	UCW	78/299 (26.09%)	45/295 (15.25%)	118/646 (18.27%)	70/409 (17.11%)	123/577 (21.32%)	85/494 (17.21%)	145/673 (21.55%)	98/563 (17.41%)	168/768 (21.88%)	122/665 (18.35%)
	UniGene v60	65/254 (25.59%)	26/186 (13.98%)	87/499 (17.43%)	54/294 (18.37%)	93/379 (24.54%)	48/315 (15.24%)	107/525 (20.38%)	66/379 (17.41%)	133/617 (21.56%)	78/489 (15.95%)
4	UCW	75/232 (32.33%)	28/160 (17.50%)	109/484 (22.52%)	44/217 (20.28%)	105/416 (25.24%)	44/246 (17.89%)	134/539 (24.86%)	53/277 (19.13%)	149/640 (23.28%)	64/323 (19.81%)
	UniGene v60	63/192 (32.81%)	17/104 (16.35%)	83/390 (21.28%)	29/155 (18.71%)	82/288 (28.47%)	29/173 (16.76%)	104/431 (24.13%)	40/214 (18.69%)	127/519 (24.47%)	29/266 (10.90%)
5	UCW	69/202 (34.16%)	19/108 (17.59%)	95/416 (22.84%)	33/138 (23.91%)	96/354 (27.12%)	23/143 (16.08%)	127/477 (26.62%)	28/175 (16.00%)	140/580 (24.14%)	42/222 (18.92%)
	UniGene v60	58/163 (35.58%)	11/70 (15.71%)	76/337 (22.55%)	14/102 (13.73%)	70/228 (30.70%)	20/112 (17.86%)	100/389 (25.71%)	23/146 (15.75%)	118/469 (25.16%)	21/178 (11.80%)
6	UCW	65/179 (36.31%)	12/85 (14.12%)	86/380 (22.63%)	22/98 (22.45%)	87/299 (29.10%)	11/94 (11.70%)	122/429 (28.44%)	21/130 (16.15%)	126/514 (24.51%)	29/165 (17.58%)
	UniGene v60	57/151 (37.75%)	7/48 (14.58%)	73/300 (24.33%)	6/71 (8.45%)	65/191 (34.03%)	13/84 (15.48%)	98/358 (27.37%)	20/122 (16.39%)	115/439 (26.20%)	16/143 (11.19%)
7	UCW	58/161 (36.02%)	11/73 (15.07%)	77/340 (22.65%)	13/74 (17.57%)	73/248 (29.44%)	7/69 (10.14%)	116/393 (29.52%)	20/111 (18.02%)	114/468 (24.36%)	22/143 (15.38%)
	UniGene v60	56/132 (42.42%)	4/37 (10.81%)	68/273 (24.91%)	5/58 (8.62%)	60/171 (35.09%)	9/64 (14.06%)	94/334 (28.14%)	18/103 (17.48%)	113/412 (27.43%)	16/124 (12.90%)
8	UCW	58/149 (38.93%)	10/62 (16.13%)	68/310 (21.94%)	12/59 (20.34%)	66/214 (30.84%)	6/56 (10.71%)	104/359 (28.97%)	17/102 (16.67%)	108/429 (25.17%)	16/119 (13.45%)
	UniGene v60	55/126 (43.65%)	3/33 (9.09%)	64/255 (25.10%)	5/50 (10.00%)	55/150 (36.67%)	9/55 (16.36%)	91/313 (29.07%)	14/89 (15.73%)	105/376 (27.93%)	15/108 (13.89%)
9	UCW	54/135 (40.00%)	8/53 (15.09%)	63/289 (21.80%)	8/51 (15.69%)	61/182 (33.52%)	5/49 (10.20%)	100/331 (30.21%)	15/91 (16.48%)	100/387 (25.84%)	13/106 (12.26%)
	UniGene v60	53/117 (45.30%)	1/30 (3.33%)	62/244 (25.41%)	5/46 (10.87%)	50/136 (36.76%)	9/48 (18.75%)	88/291 (30.24%)	13/83 (15.66%)	97/345 (28.12%)	12/99 (12.12%)
10	UCW	52/126 (41.27%)	8/50 (16.00%)	62/279 (22.22%)	8/50 (16.00%)	56/165 (33.94%)	4/45 (8.89%)	96/309 (31.07%)	14/82 (17.07%)	91/355 (25.63%)	13/100 (13.00%)
	UniGene v60	50/105 (47.62%)	1/28 (3.57%)	60/226 (26.55%)	5/39 (12.82%)	43/119 (36.13%)	7/45 (15.56%)	85/272 (31.25%)	13/82 (15.85%)	92/318 (28.93%)	12/97 (12.37%)

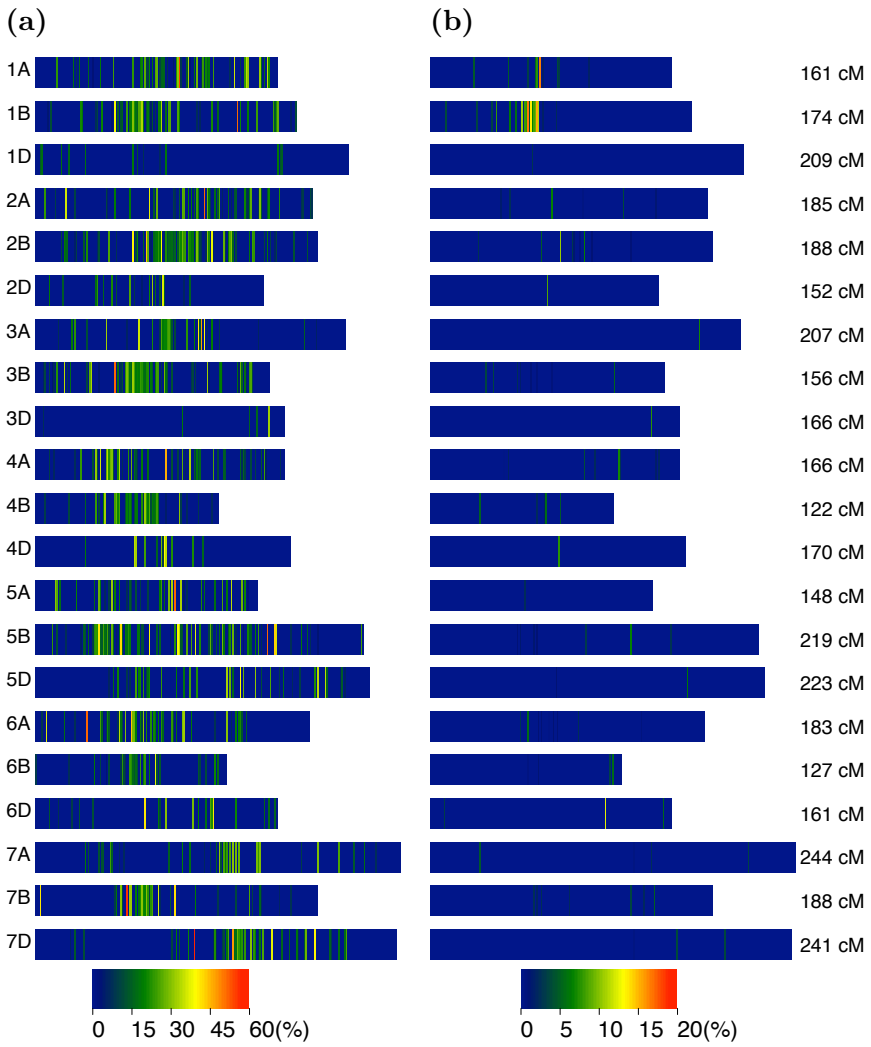


Figure 3.7: Genetic location of genes with SNPs between AVS and Yr15. The colour scale indicates the percentage of genes with SNPs per centi-Morgan (cM) across the 21 wheat chromosomes. The location of the genes was determined by the best alignment to the CSS scaffolds, and the location of these was determined by their position on a genetic map (Wang et al., 2014) (a). All the SNPS between progenitors. Note the lack of enrichment across any individual chromosome. (b) SNPs with BFR>6. Note the enrichment in Chromosome 1B

3.6 Assay selection

. The selection criteria to decide which SNPs where selected to produce the genetic map: BFR>6, in the short arm of chromosome group 1 and from the *Yr15* progenitor.

3.7 Genetic map

The three versions of the genetic map: With a subset of the F_2 population

3.8 Bioinformatic methods

3.8.1 Alignment reads to gene models

The raw output from the Illumina HiSeq 2000 was processed with Casava v1.8 (Illumina, 2011). Lanes 1 and 2, containing multiplexed bulks (Table 3.1) was demultiplexed with a tolerance of 1 mismatch in the barcode. Lanes 3 and 4 contained the parental sequences without a barcode. The FastQ files where left compressed and in chunks of 40,000, as teh default for the BCL conversion pipeline from Casava to allow parallel processing in a cluster environment. The quality of the sequencing lanes was assessed with FastQC v0.10.1 (Babraham Bioinformatics, 2012). The RNA-Seq reads were aligned with BWA 0.5.9 (Li and Durbin, 2009) to the wheat UniGene database v60 (Pontius et al., 2002) and to the UCW gene models (Krasileva et al., 2013), including the *T. turgidum* and complementary ORFs (MAS Wheat, 2013). The alignments where sorted and stored as single BAM files to have random access (Li et al., 2009).

3.8.2 Bulk Frequency Ratios

Listing 3.1: Method to find best BLAT alignment

```
def self.each_best_hit(text = '')
  emptyHit = Bio::Blat::Report::Hit.new
  emptyHit.score = 0
  best_aln = Hash.new(emptyHit)
  self.each_hit(text) do |hit|
    current_score = hit.score
```

```
    if current_score > best_aln[current_name].score
      best_aln[current_name] = hit
    end
  end
  best_aln.each_value { |val| yield  val }
end
```

Listing 3.2: Extensions to Bio::Blat::Report::Hit to get the percentage of coverage

```
class Bio::Blat::Report::Hit
  def covered
    match + mismatch
  end
  def query_percentage_covered
    covered * 100.0 / query_len.to_f
  end
  def target_percentage_covered
    covered * 100.0 / target_len.to_f
  end
end
```

3.9 Discussion

Remarks on how this techinque can be used to do fine-mapping and that if I were to start the project now I would use exome capture or Ren-Seq.

The references have changed since we started

There are new annotations, now we don't necessarily need to use unigenes anymore.

Importance of genotyping everything used in an experiment

Mention other people using a similar strategy since this was published.

We can use different techniques now (exome capture, ren-seq)

The markers are now used by our collaborators.

Chapter 4

Gene expression (expVIP)

4.1 Expression experiments (Introduction)

Describe the list of previously published expression experiments and how they can potentially be used as a framework for new experiments.

4.2 Database design

Description of how the database was designed and the flexibility given by having the factors and units as variables

4.3 Analysis pipeline

Implementation of the pipeline, from running kallisto to load the data in the database

4.4 Graphical interface

How the expression can be displayed filtered, and sorted

4.5 Conclusions

The use of previously published studies is a valuable resource. Also, mention that despite the fact that there are several expression/gene browsers, none of them allow comparisons between species and don't consider polyploids.

Chapter 5

Conclusions and final remarks

This section wraps up by showing the relationship and importance of a comprehensive approach to data analysis, from the field, genetics, molecular biology and genomics. I will also remark how the technology and the resources have changed in the last 4 years. As at the references used at beginning where superseded during the PhD.

Appendix A

Supplemental tables

Table A.1: Count of KASP assays designed for the 40,267 SNP markers located in the genetic map from Wang et al. (2014). 4,228 assays did not align to the target chromosome. Not designed: Primer3 could not find viable primers flanking the SNP.

	Homoeologous variant	Varietal SNP	Percentage
Non-specific	1,765	5,857	21.15%
Semi-specific	7,942	6,907	41.20%
Specific	6,813	5,957	35.43%
Not designed	242	556	2.21%
Total	16,762	19,277	36,039

A.1 PolyMarker supplemental tables.

Table A.2: PolyMarker used to genotype PST

Assay	Contig	Position	X	Y	Cluster I isolates		Cluster II isolates		Cluster III isolates			Cluster IV isolates	
					13/26	13/123	CL1	T-13/3	13/09	13/23	13/182	13/36	13/40
1	PST130_14470	268	C	T	X:Y	X:Y	X:X	X:X	X:X	X:X	X:X	X:X	X:X
2	PST130_8160	11876	C	T	Y:Y	Y:Y	X:Y	X:Y	X:Y	X:Y	X:Y	X:Y	X:Y
3	PST130_14628	1712	A	C	X:Y	-	X:X	X:X	X:X	X:X	X:X	X:X	X:X
4	PST130_14898	503	G	A	X:X	X:X	X:Y	X:Y	X:Y	X:Y	-	X:Y	X:Y
5	PST130_28344	2372	A	G	Y:Y	Y:Y	X:Y	X:Y	Y:Y	Y:Y	Y:Y	Y:Y	Y:Y
6	PST130_7634	3463	A	C	Y:Y	Y:Y	X:Y	X:Y	Y:Y	Y:Y	Y:Y	Y:Y	Y:Y
7	PST130_7629	11699	G	A	Y:Y	Y:Y	X:Y	X:Y	Y:Y	Y:Y	Y:Y	Y:Y	Y:Y
8	PST130_10943	2979	C	T	X:Y	X:Y	X:Y	X:Y	X:X	X:X	X:X	X:Y	X:Y
9	PST130_10126	6216	G	T	Y:Y	Y:Y	X:X	X:X	X:X	X:X	-	Y:Y	Y:Y
10	PST130_22010	172	C	T	Y:Y	Y:Y	Y:Y	Y:Y	X:Y	X:Y	-	X:Y	X:Y
11	PST130_16961	1098	C	T	X:X	X:X	X:Y	X:Y	Y:Y	Y:Y	Y:Y	X:Y	X:Y
12	PST130_6915	2710	A	T	Y:Y	Y:Y	Y:Y	Y:Y	Y:Y	X:Y	X:Y	Y:Y	Y:Y
13	PST130_12479	1428	C	T	X:X	X:X	Y:Y	Y:Y	X:X	X:X	X:X	Y:Y	X:X
14	PST130_7634	3883	C	G	X:X	X:X	X:Y	X:Y	X:X	X:X	X:Y	X:Y	X:X
15	PST130_14470	456	T	C	Y:Y	Y:Y	X:Y	X:Y	Y:Y	Y:Y	X:Y	Y:Y	Y:Y

Table A.3: Validation of homozygous deletions on line Cadenza0423.

Marker	Deletion	chr	cM	1	2	3	4	5	6	7	8	9	10	11	12	C	C	C	C	Result
5BS_2297308_Cadenza0423_12664_C12664T	-	5B	4.551	X	X	-	X	X	X	X	X	X	X	-	X	Y	Y	Y	Y	HOM Mutation
5BL_10812849_Cadenza0423_5664_G5664T	-	5B	38.769	X	X	-	X	X	X	X	X	X	X	-	X	Y	Y	Y	Y	HOM Mutation
5BL_10825062_Cadenza0423_7917_G7917A	-	5B	38.769	X	X	-	X	X	X	X	X	X	X	-	X	Y	Y	Y	Y	HOM Mutation
IWGSC_CSS_5BL_scaff_10847976:27068-27231	+	5B	38.769	X	X	-	X	X	X	X	X	X	X	-	X	H	H	H	H	Hom Deletion
IWGSC_CSS_5BL_scaff_10847976:28118-28674	+	5B	38.769	X	X	-	X	X	X	X	X	X	X	-	X	H	H	H	H	Hom Deletion
IWGSC_CSS_5BL_scaff_10865441:15863-15946	+	5B	38.769	X	X	-	X	X	X	X	X	X	X	-	X	H	H	H	H	Hom Deletion
5BL_10837222_Cadenza0423_4616_G4616A	-	5B	39.905	X	X	-	X	X	X	X	X	X	X	-	X	Y	Y	Y	Y	HOM Mutation
5BL_10891320_Cadenza0423_18847_C18847T	-	5B	45.594	Y	Y	-	Y	H	X	X	Y	H	Y	-	H	Y	Y	Y	Y	HET Mutation

Table A.4: Validation of mutations on M_4 on Cadenza

IWGSC contig	Line	Pos	WT	Mut	Predicted	M_4	Primer 1 (Cadenza)	Primer 2 (mutant)	Common Primer
IWGSC_CSS_3B_scaff_10445294	Cadenza1772	6019	C	T	het	het	caggatAgtGggactgtcaaaG	caggatAgtGggactgtcaaaA	ggagacGGctGtggacatT
IWGSC_CSS_3DL_scaff_6955403	Cadenza1772	2418	C	T	het*	hom	tcagCggattgtcgggatG	tcagCggattgtcgggatA	tgtcCatgaaTcttgtccacG
IWGSC_CSS_4AL_scaff_7106846	Cadenza1772	11277	G	A	hom	hom	tgggatccatgcctacactG	tgggatccatgcctacactA	gatggtGgatttgccgctA
IWGSC_CSS_4AS_scaff_5991335	Cadenza1772	15710	G	A	hom	hom	ctggccctgcgctgctaC	ctggccctgcgctgctaT	gtggaaGttcagaaggaccaG
IWGSC_CSS_4BS_scaff_4956646	Cadenza1772	252	G	A	het*	hom	gcaggttgacttcccggaG	gcaggttgacttcccggaA	tGaggtacgaGcTaaagAaagC
IWGSC_CSS_4DS_scaff_1715962	Cadenza1772	1225	G	A	hom	hom	cagctgtggTatctcaactgG	cagctgtggTatctcaactgA	CcCtGaaACACcGtttggaT
IWGSC_CSS_5AL_scaff_2763407	Cadenza1772	2119	G	A	hom	hom	gcgacGaacctcgagatctG	gcgacGaacctcgagatctA	gaTggcaAtcgtCgtgcA
IWGSC_CSS_5AS_scaff_1548786	Cadenza1772	12625	C	T	het	het	Ataggcacattgctagactgag	Ataggcacattgctagactgaa	ggattgggtgttgcacgC
IWGSC_CSS_5BL_scaff_10849226	Cadenza1772	2289	C	T	het*	hom	cctgacatcattgttcacgatC	cctgacatcattgttcacgatT	cactccgaggtgtccatgaT
IWGSC_CSS_5BS_scaff_2270737	Cadenza1772	2262	G	A	hom	—	attcCTgtgttgtggCaaatgaG	attcCTgtgttgtggCaaatgaA	taaGcacaaAccctccagctgG
IWGSC_CSS_1AL_scaff_3022915	Cadenza1661	891	C	T	hom	hom	ccacagtgagactcctattgaCG	ccacagtgagactcctattgaCA	atgtctgattcGtcGtagtcC
IWGSC_CSS_1AS_scaff_3297240	Cadenza1661	1970	C	T	het	het	catcccgccGtttcctcC	catcccgccGtttcctcT	gctcgccgatgaagagcT
IWGSC_CSS_1BL_scaff_3828996	Cadenza1661	1340	G	A	hom	hom	agccggatgttagtgttaacC	agccggatgttagtgttaacT	agcagcttgTcgcgttaaC
IWGSC_CSS_1DS_scaff_1884529	Cadenza1661	10575	G	A	hom	hom	aCagatacaAttgtcatgcaggC	aCagatacaAttgtcatgcaggT	acctgggTTgtccaatacttC
IWGSC_CSS_2AL_scaff_6318370	Cadenza1661	19142	C	T	het	—	cgtggcCgaatCtcGacG	cgtggcCgaatCtcGacA	ttcttgtgggagccgggC
IWGSC_CSS_2AS_scaff_5213460	Cadenza1661	1358	G	A	hom	hom	gtcacgaaCccgctcagG	gtcacgaaCccgctcagA	aggaaagagaggaaaagaGcG
IWGSC_CSS_2BS_scaff_5179331	Cadenza1661	5604	G	A	het	het	actctcgtcaagaactgatacaG	actctcgtcaagaactgatacaA	gcaGagaatgttcttgcaacT
IWGSC_CSS_2DS_scaff_5341235	Cadenza1661	4673	G	A	het	het	ggtgaggatctcggagctG	ggtgaggatctcggagctA	gcgcggtcgtacgagttG
IWGSC_CSS_3AL_scaff_4250995	Cadenza1661	7046	G	A	hom	hom	cCaagaaacgggtggtccaG	cCaagaaacgggtggtccaA	ctgcagctgtcccatcatcgT
IWGSC_CSS_3B_scaff_10404421	Cadenza1661	4303	G	A	het	het	ccttcgtcgaCaggacctG	ccttcgtcgaCaggacctA	GCcagtactCacAtgctctC
IWGSC_CSS_5DL_scaff_2390496	Cadenza1538	2125	C	T	hom	het	gcagttttatcctcagtagtcttgG	gcagttttatcctcagtagtcttgA	ttctgagaaTgtaatgtgcGatG
IWGSC_CSS_6AL_scaff_5753680	Cadenza1538	3920	C	T	hom	hom	tgctccaaatttgagcacaaTaaC	tgctccaaatttgagcacaaTaaT	aaatgcaaggggtaagtttttgT
IWGSC_CSS_6AS_scaff_4425792	Cadenza1538	4307	G	A	hom	het	agatgcttgtCggGccaG	agatgcttgtCggGccaA	gctgaagcaacgcgatcaaT
IWGSC_CSS_6BS_scaff_3003630	Cadenza1538	6933	C	T	het	het	ggcagtaatgtggtgctgagC	ggcagtaatgtggtgctgagT	tTgaCttctggtttggtggcA
IWGSC_CSS_6DL_scaff_3246988	Cadenza1538	9186	G	A	het	het	gctaaagaagagcttgagagaattC	gctaaagaagagcttgagagaattT	aatttctgaagagaggtgttgtatG
IWGSC_CSS_7AL_scaff_4480114	Cadenza1538	3446	C	T	het	—	gatatctcccacacggcgG	gatatctcccacacggcgA	tgagccactcttgcagtttT
IWGSC_CSS_7AS_scaff_4193541	Cadenza1538	8359	C	T	hom	het	agcaattctttggctatcaattagC	agcaattctttggctatcaattagT	tcatctGtcttaactctactgctG
IWGSC_CSS_7BL_scaff_6721572	Cadenza1538	9223	C	T	het	het	gctCagggaggaagacaagaaG	gctCagggaggaagacaagaaA	tgctatgaagaattccgacctC
IWGSC_CSS_7BS_scaff_3152545	Cadenza1538	3960	G	A	hom	—	tcagcaaaatcacctgcCgC	tcagcaaaatcacctgcCgT	gCtgccccatcatcgtttaT
IWGSC_CSS_7DS_scaff_3963838	Cadenza1538	2913	G	A	het	het	tCgttgcaagcCttTtgtgC	tCgttgcaagcCttTtgtgT	agaGttaTcaagcTactgtcacA
IWGSC_CSS_1AL_scaff_3903380	Cadenza1469	6193	G	A	hom	hom	ctcttcAgagatgaacgcgG	ctcttcAgagatgaacgcgA	tcGtGagatgGtggtttGTtA
IWGSC_CSS_1AS_scaff_3287728	Cadenza1469	3817	C	T	het*	hom	ccgaccaAttcactaaccgG	ccgaccaAttcactaaccgA	accctctttcccAgacatgaT
IWGSC_CSS_1BL_scaff_3815304	Cadenza1469	513	G	A	hom	hom	aacatttgcctTaCcaaaacGC	aacatttgcctTaCcaaaacGT	acacagcaagttataatgCAAgC
IWGSC_CSS_1DL_scaff_2266648	Cadenza1469	5926	C	T	het	het	caacatgagacacaacaccttC	caacatgagacacaacaccttT	gtcaacgcgtgaggattgtC
IWGSC_CSS_1DS_scaff_1906671	Cadenza1469	3697	C	T	hom	hom	tggTGtagacacttggcgaG	tggTGtagacacttggcgaA	catggcgaccaccAcctG
IWGSC_CSS_2AL_scaff_6337088	Cadenza1469	7334	G	A	het*	hom	acaatgccAagttgacaggttG	acaatgccAagttgacaggttA	gggagtgttggttCagaacaT

IWGSC contig	Line	Pos	WT	Mut	Predicted	M_4	Primer 1 (Cadenza)	Primer 2 (mutant)	Common Primer
IWGSC_CSS_2BL_scaff_7972799	Cadenza1469	8995	C	T	het	hom	gTgCtcctcGgcatccttC	gTgCtcctcGgcatccttT	gatccgGgcaaactacgTG
IWGSC_CSS_2DL_scaff_9832343	Cadenza1469	3262	G	A	het	het	TtgtctaAcagcacCGcagG	TtgtctaAcagcacCGcagA	agatctcggtcagcctttcT
IWGSC_CSS_2DS_scaff_5327939	Cadenza1469	3889	G	A	het	het	ttttTgccttatgtgactctagtaC	ttttTgccttatgtgactctagtaT	gaggccatcacagatagcG
IWGSC_CSS_3B_scaff_10395219	Cadenza1469	1292	G	A	hom	—	aggtgcttgtgcttgctgG	aggtgcttgtgcttgctgA	cctcttctgggggctttataC
IWGSC_CSS_3B_scaff_10592217	Cadenza0580	2994	C	T	het	—	acagcagtatcaagcccctC	acagcagtatcaagcccctT	tgatactgttgTggCggagG
IWGSC_CSS_3DS_scaff_2596771	Cadenza0580	1037	G	A	het	het	tggttatgCAcaggataatCagG	tggttatgCAcaggataatCagA	tggcaaatgtgatgtcattaggT
IWGSC_CSS_4AL_scaff_7093953	Cadenza0580	9881	C	T	hom	hom	GacaggaagccggtaacaC	GacaggaagccggtaacaT	ctccAgcaggcatgggaT
IWGSC_CSS_4BL_scaff_7037448	Cadenza0580	1837	C	T	hom	hom	CgttgaaaaGctgcaagaacttaaC	CgttgaaaaGctgcaagaacttaaT	cagttcttccTtCaGagcagataT
IWGSC_CSS_4BS_scaff_4929479	Cadenza0580	10668	G	A	hom	—	tggattttcccgcactgttC	tggattttcccgcactgttT	gtaaacaaggcatttcaagagtcA
IWGSC_CSS_4DL_scaff_14359838	Cadenza0580	1408	G	A	hom	—	gCtcAttcagggatTGTcCtaTatG	gCtcAttcagggatTGTcCtaTatA	tgaCagaacagttggtcatacT
IWGSC_CSS_4DS_scaff_2276484	Cadenza0580	8034	G	A	hom	hom	gccgtggttgatggAgaG	gccgtggttgatggAgaA	cgtccagattactgatacttgcA
IWGSC_CSS_5AL_scaff_2756579	Cadenza0580	5278	G	A	het	het	tgaatggatttttcgtcccgttC	tgaatggatttttcgtcccgttT	ggAAtCCTATgCAgaAgAaaCTG
IWGSC_CSS_5BL_scaff_10787208	Cadenza0580	10627	G	A	het	—	gcctctcacatgcggagaC	gcctctcacatgcggagaT	acgatgtcAggtggGcgT
IWGSC_CSS_5BS_scaff_2282179	Cadenza0580	5267	G	A	het	—	tgatgggctacgacgtgC	tgatgggctacgacgtgT	tcggcgcccttgaaAtcC
IWGSC_CSS_5DL_scaff_4498073	Cadenza0423	4937	C	T	hom	hom	gcaccctctggttggtcatC	gcaccctctggttggtcatT	tgagcagcaAagcagccG
IWGSC_CSS_5DS_scaff_2738970	Cadenza0423	2319	C	T	het	—	cgtgaggtgggtgatttgC	cgtgaggtgggtgatttgT	tggaactagttacactgcagtTC
IWGSC_CSS_6AL_scaff_5757109	Cadenza0423	2788	G	A	hom	hom	caggaGcctggcaaataaaGG	caggaGcctggcaaataaaGA	ctttcGcagtctcttagtttcG
IWGSC_CSS_6AS_scaff_4387871	Cadenza0423	2543	G	A	hom	hom	gcatgctaacaggcgaaaagG	gcatgctaacaggcgaaaagA	ctcatgctcctgatcttaaggtT
IWGSC_CSS_6BL_scaff_4271391	Cadenza0423	4660	C	T	hom	hom	tacgtgcatgatgtggtagtcgtaC	tacgtgcatgatgtggtagtcgtaT	gtttgaagtgcatcagatgTaccA
IWGSC_CSS_6DS_scaff_1880206	Cadenza0423	9159	G	A	het	het	ctgCgaaggctccacaaG	ctgCgaaggctccacaaA	ggatgagaagtttgcattgctC
IWGSC_CSS_7AS_scaff_4227506	Cadenza0423	952	G	A	het	—	ccatgtgtttccaatgttagagC	ccatgtgtttccaatgttagagT	tgccctagctggtatgcT
IWGSC_CSS_7BL_scaff_6681782	Cadenza0423	1486	C	T	hom	hom	agtaagCGtgacagcaatggG	agtaagCGtgacagcaatggA	AtgtctTtgGtggaagtacatcA
IWGSC_CSS_7BS_scaff_3160328	Cadenza0423	7801	C	T	het	het	tgttaaatGatacagCctgcagC	tgttaaatGatacagCctgcagT	tggaatggtgCgttgttttT
IWGSC_CSS_7DS_scaff_407428	Cadenza0423	2051	G	A	het	het	gtcGCgccatcctgacaG	gtcGCgccatcctgacaA	actcatcAggtcagcccaA
IWGSC_CSS_3AL_scaff_442479	Cadenza0364	3198	C	T	het	het	gagtcaTtaagttggtaagattggC	gagtcaTtaagttggtaagattggT	GCaGaTaaCaacaggatcacG
IWGSC_CSS_3AL_scaff_4447942	Cadenza0364	11917	G	A	het	het	gtcataaagattgctcctgtgaaG	gtcataaagattgctcctgtgaaA	ctcGgatgtgggaggaagA
IWGSC_CSS_3AS_scaff_1557483	Cadenza0364	2547	C	T	het	het	aaagtcacatcatgcttaccataaG	aaagtcacatcatgcttaccataaA	cgaaatccaacgcctcatcA
IWGSC_CSS_3AS_scaff_2648747	Cadenza0364	2688	G	A	het	het	tggAagcAcaaggggccC	tggAagcAcaaggggccT	GccgccgatggagactcG
IWGSC_CSS_3AS_scaff_3304956	Cadenza0364	1017	G	A	het	het	gtcccttgcacacagctttG	gtcccttgcacacagctttA	cctgctggactacaacttcaaT
IWGSC_CSS_3AS_scaff_3321091	Cadenza0364	4585	C	T	het	het	caagaatgATgctgatgttggaG	caagaatgATgctgatgttggaA	acatgctgaatcgccgaatC
IWGSC_CSS_3AS_scaff_3371333	Cadenza0364	538	G	A	het	het	gggaaaCgAgAcgagcgG	gggaaaCgAgAcgagcgA	ccgtgccttcctcacccT
IWGSC_CSS_3AS_scaff_3371815	Cadenza0364	1061	C	T	het	het	atccccacggcacagagG	atccccacggcacagagA	aAttggcccttggtgattcC
IWGSC_CSS_3AS_scaff_3440912	Cadenza0364	4498	G	A	het	het	ccgtaaaactttctgtgcttgC	ccgtaaaactttctgtgcttgT	atActgacaaactacatgatgtgC
IWGSC_CSS_3B_scaff_10343586	Cadenza0364	2242	G	A	het	—	ggttcTgTcctctcttccactG	ggttcTgTcctctcttccactA	tgtgttgaacccgcaagcA
IWGSC_CSS_3AL_scaff_442479	Cadenza0364	3198	C	T	het	het	gagtcaTtaagttggtaagattggC	gagtcaTtaagttggtaagattggT	GCaGaTaaCaacaggatcacG
IWGSC_CSS_3AL_scaff_4447942	Cadenza0364	11917	G	A	het	het	gtcataaagattgctcctgtgaaG	gtcataaagattgctcctgtgaaA	ctcGgatgtgggaggaagA
IWGSC_CSS_3AS_scaff_1557483	Cadenza0364	2547	C	T	het	het	aaagtcacatcatgcttaccataaG	aaagtcacatcatgcttaccataaA	cgaaatccaacgcctcatcA
IWGSC_CSS_3AS_scaff_2648747	Cadenza0364	2688	G	A	het	het	tggAagcAcaaggggccC	tggAagcAcaaggggccT	GccgccgatggagactcG
IWGSC_CSS_3AS_scaff_3304956	Cadenza0364	1017	G	A	het	het	gtcccttgcacacagctttG	gtcccttgcacacagctttA	cctgctggactacaacttcaaT
IWGSC_CSS_3AS_scaff_3321091	Cadenza0364	4585	C	T	het	het	caagaatgATgctgatgttggaG	caagaatgATgctgatgttggaA	acatgctgaatcgccgaatC

IWGSC contig	Line	Pos	WT	Mut	Predicted	M_4	Primer 1 (Cadenza)	Primer 2 (mutant)	Common Primer
IWGSC_CSS_3AS_scaff_3371333	Cadenza0364	538	G	A	het	het	gggaaaCgAgAcgagcgG	gggaaaCgAgAcgagcgA	ccgtgccttcctcacccT
IWGSC_CSS_3AS_scaff_3371815	Cadenza0364	1061	C	T	het	het	atccccacggcacagagG	atccccacggcacagagA	aAttggcccttggtgattcC
IWGSC_CSS_3AS_scaff_3440912	Cadenza0364	4498	G	A	het	het	ccgtaaaactttctgtgcttgC	ccgtaaaactttctgtgcttgT	atActgacaaactacatgatgtgC
IWGSC_CSS_3B_scaff_10343586	Cadenza0364	2242	G	A	het	—	ggttcTgTcctctcttccactG	ggttcTgTcctctcttccactA	tgtgttgaacccgcaagcA
IWGSC_CSS_5DL_scaff_242342	Cadenza0281	2433	C	T	hom	hom	catggCgacggtGtcctG	catggCgacggtGtcctA	aAccctcatTTtggCTACTtCT
IWGSC_CSS_5DL_scaff_4538822	Cadenza0281	1208	G	A	hom	—	acgtcagaacaaccgtttgaC	acgtcagaacaaccgtttgaT	ttaaattggttggcgccacC
IWGSC_CSS_6AL_scaff_5813297	Cadenza0281	4532	C	T	hom	—	gggagagggacgtctcgG	gggagagggacgtctcgA	ttcttctgccaacgattccG
IWGSC_CSS_6AS_scaff_4378990	Cadenza0281	6748	C	T	hom	hom	cccaggttctgcttcttttcC	cccaggttctgcttcttttcT	caagtatcaagaaaatgaagggTgT
IWGSC_CSS_6BL_scaff_4360781	Cadenza0281	5426	C	T	het	het	aCtactcaaatggcttGgtgtaG	aCtactcaaatggcttGgtgtaA	tcagtccaacatgTcaagagatT
IWGSC_CSS_7AL_scaff_4488310	Cadenza0281	3808	G	A	hom	hom	gttctcttgtagtagcagccG	gttctcttgtagtagcagccA	ggcgctttcttcggcctA
IWGSC_CSS_7BL_scaff_6696509	Cadenza0281	9232	G	A	het	het	gctctaggGgtggcaaAagG	gctctaggGgtggcaaAagA	ggcttGaGgtcGcagtgT
IWGSC_CSS_7BS_scaff_3143575	Cadenza0281	1866	C	T	het	het	agatgttgagagggcgcttC	agatgttgagagggcgcttT	gcttggAtggtggcaagtT
IWGSC_CSS_7DL_scaff_3346250	Cadenza0281	1663	G	A	het	het	acgtgcagcaacatcctaaC	acgtgcagcaacatcctaaT	TttcccaccaggcccaagA
IWGSC_CSS_7DS_scaff_3933917	Cadenza0281	1243	C	T	het	het	tgCtgagcCttTcaccttgC	tgCtgagcCttTcaccttgT	agaggtttggttccatcGG
IWGSC_CSS_3B_scaff_10626860	Cadenza0148	7847	G	A	het	het	gcagctctgggaaggagG	gcagctctgggaaggagA	gttaatgtacCTcctagcctcG
IWGSC_CSS_3DL_scaff_6915683	Cadenza0148	6904	C	T	het	het	cgtcaaCctgtgggcaattG	cgtcaaCctgtgggcaattA	tcatgctcataatgTcatagggT
IWGSC_CSS_4AS_scaff_5929057	Cadenza0148	4238	G	A	hom	hom	gcgcaacgtagCacctacC	gcgcaacgtagCacctacT	ttatctggtgaagtgacaggttCA
IWGSC_CSS_4AS_scaff_5950625	Cadenza0148	10590	C	T	het	het	agaTattCaaaTcggtggAttggC	agaTattCaaaTcggtggAttggT	cctgCtcccctcacgtcC
IWGSC_CSS_4AS_scaff_5967119	Cadenza0148	11626	C	T	hom	hom	cgtGgacaccccgagctG	cgtGgacaccccgagctA	gacgacgcactgcacgaC
IWGSC_CSS_4DL_scaff_14455742	Cadenza0148	1946	C	T	hom	hom	gCctgagggagatcgcgC	gCctgagggagatcgcgT	aaccgGtAaCTGtGgGcA
IWGSC_CSS_4DS_scaff_2318993	Cadenza0148	4000	C	T	hom	hom	tccagtttgacacagattgaatggG	tccagtttgacacagattgaatggA	tgagaTtctgtttcctttcacAttG
IWGSC_CSS_5AL_scaff_2750707	Cadenza0148	4603	G	A	het	het	ccttggtgctagccatttcaagTaG	ccttggtgctagccatttcaagTaA	ccaggaTgcAgtgcaatatttcaaG
IWGSC_CSS_5BL_scaff_10794137	Cadenza0148	9235	C	T	hom	hom	gaagctgcttctgcgttG	gaagctgcttctgcgttA	agtatcccttccatataagcagtG
IWGSC_CSS_5BS_scaff_1646558	Cadenza0148	2916	C	T	het	het	gccGtacactcacctAtcctttG	gccGtacactcacctAtcctttA	gcaaTgtccacttAtcatcccT
IWGSC_CSS_1AL_scaff_3883106	Cadenza0110	27536	C	T	het	het	accttccatcactggctgG	accttccatcactggctgA	gtgaagaacaacaggttgaagC
IWGSC_CSS_1BL_scaff_3812829	Cadenza0110	10770	G	A	het*	hom	cccccactccattccagG	cccccactccattccagA	gGatgttgttctgtgctggaA
IWGSC_CSS_1DL_scaff_2266648	Cadenza0110	6156	G	A	het	het	actgcgtggttatgggacC	actgcgtggttatgggacT	ccccatcactgaacacaacA
IWGSC_CSS_1DS_scaff_1889435	Cadenza0110	8826	C	T	hom	hom	aaccatgaattactcggacagG	aaccatgaattactcggacagA	gccctgaagaattgtatcaaaacaG
IWGSC_CSS_2AS_scaff_5268634	Cadenza0110	4636	G	A	het	het	gatccatgtgattggcatgtttG	gatccatgtgattggcatgtttA	TgctgtTggatatgcagttacT
IWGSC_CSS_2BL_scaff_7965110	Cadenza0110	15801	C	T	hom	hom	cattgaagcAtacacAattgcAtaC	cattgaagcAtacacAattgcAtaT	gccagagtatccagataaggTttA
IWGSC_CSS_2DL_scaff_9852812	Cadenza0110	13788	G	A	hom	hom	atttttgtatggtctcaatcttcgC	atttttgtatggtctcaatcttcgT	gaacgtTcattcttgtacttgcT
IWGSC_CSS_2DS_scaff_5371379	Cadenza0110	2166	C	T	hom	hom	agacacaaaactagtGatgcgC	agacacaaaactagtGatgcgT	gctgctgagaatgttTtgtatttG
IWGSC_CSS_3AL_scaff_4384278	Cadenza0110	1276	C	T	het	het	agcTgaactgccccTgtaG	agcTgaactgccccTgtaA	agggacctCgGtggatgaA
IWGSC_CSS_3AS_scaff_3340122	Cadenza0110	1467	C	T	hom	hom	attcctAgtgttgtcggaacatG	attcctAgtgttgtcggaacatA	gagaagactagaaagttttcAgcaT
IWGSC_CSS_5DL_scaff_4554222	Cadenza2103	6528	C	T	het*	hom	gctgccctacaaagaaacaaaattG	gctgccctacaaagaaacaaaattA	aTcccaactatCGaTtttgtcataC
IWGSC_CSS_6AL_scaff_5833640	Cadenza2103	7346	C	T	hom	hom	aagaaaagccacaatggtttctC	aagaaaagccacaatggtttctT	aCTctgTcagtgtttcccagC
IWGSC_CSS_6AS_scaff_4429974	Cadenza2103	3867	G	A	hom	hom	GagatgaAtttattgagcatgtggC	GagatgaAtttattgagcatgtggT	ggttccggctgcataagT
IWGSC_CSS_6DL_scaff_3307626	Cadenza2103	4970	C	T	hom	hom	tgcagatgttgtcctgtgtaG	tgcagatgttgtcctgtgtaA	ctaggaaggtgattttgtactGtC
IWGSC_CSS_6DS_scaff_2059604	Cadenza2103	5224	G	A	het	—	gctcaatgcatgcTgagtgG	gctcaatgcatgcTgagtgA	tgtcaagtattattttcctgctctG
IWGSC_CSS_7AL_scaff_4552322	Cadenza2103	1412	C	T	het	het	gcaaaggcTgatactccaacaG	gcaaaggcTgatactccaacaA	ggcAAGccAgtataaaagtaaGC

IWGSC contig	Line	Pos	WT	Mut	Predicted	M_4	Primer 1 (Cadenza)	Primer 2 (mutant)	Common Primer
IWGSC_CSS_7BS_scaff_3147455	Cadenza2103	4607	G	A	het	—	gcaccttaggatgtgagTtatgC	gcaccttaggatgtgagTtatgT	gcatgtagggtttatttgactgttA
IWGSC_CSS_7DL_scaff_3382467	Cadenza2103	3473	C	T	hom	—	GGTtctgCaGTTCATAActcatC	GGTtctgCaGTTCATAActcatT	attgaatcaactgatacGaaGactC
IWGSC_CSS_3B_scaff_10457010	Cadenza0277	10599	G	A	het	het	aaccttggccgcagaacaC	aaccttggccgcagaacaT	actggctgcacgagaggG
IWGSC_CSS_3B_scaff_10593852	Cadenza0277	10124	C	T	het	het	tgacaggggacgctatacaG	tgacaggggacgctatacaA	gtctaaCTtACattAcccatcagC
IWGSC_CSS_3DS_scaff_2583390	Cadenza0277	663	G	A	hom	hom	actgcactcatacaatActtCtgC	actgcactcatacaatActtCtgT	tcCacctggacagcaagtG
IWGSC_CSS_4AL_scaff_7093953	Cadenza0277	10004	C	T	hom	hom	ccttgtattcaatggaTtgTtttgG	ccttgtattcaatggaTtgTtttgA	ttccccaaaTaaaaaggaagagC
IWGSC_CSS_4AL_scaff_7176064	Cadenza0277	6220	C	T	het	het	gtgccgtaTtcCgcctgG	gtgccgtaTtcCgcctgA	atgttcgaggggatgggG
IWGSC_CSS_4DL_scaff_14122349	Cadenza0277	1010	C	T	hom	hom	gtcgctgctgCttgtgaG	gtcgctgctgCttgtgaA	ggaacaggcccaaggagG
IWGSC_CSS_5AL_scaff_2736916	Cadenza0277	4296	G	A	het	het	aagaactATgAaaGtaacacacgaC	aagaactATgAaaGtaacacacgaT	ttcGcTttTaagGcAttCtcG
IWGSC_CSS_5BL_scaff_10883744	Cadenza0277	2080	C	T	hom	hom	gcctctttCtgttTagcctcaG	gcctctttCtgttTagcctcaA	cgacaaggttcgtgatTgcA
IWGSC_CSS_1AL_scaff_3932013	Cadenza0548	11765	C	T	hom	hom	accgccaaCccaagacaG	accgccaaCccaagacaA	cccattaGccgTgcAacG
IWGSC_CSS_1BS_scaff_3417505	Cadenza0548	373	C	T	het	het	gtggtgaggaGGgtgGaG	gtggtgaggaGGgtgGaA	tggtcgGccagttgttgA
IWGSC_CSS_2AS_scaff_5305619	Cadenza0548	2786	C	T	hom	hom	atacagatgccctAAgtggTtC	atacagatgccctAAgtggTtT	ggaagacaAtGctccaggtaC
IWGSC_CSS_2AS_scaff_5306489	Cadenza0548	46953	T	G	het	wt	aggttccatgtccatagaagGT	aggttccatgtccatagaagGG	aggctaTAgactcctgtACAgT
IWGSC_CSS_2BL_scaff_7984123	Cadenza0548	11660	G	A	het	het	cattgtggcatagtaatcagtacaG	cattgtggcatagtaatcagtacaA	aatacattgaggaatcaaagccC
IWGSC_CSS_2DL_scaff_9907477	Cadenza0548	1363	C	T	hom	hom	tgcctccctttgccagaaC	tgcctccctttgccagaaT	ggcaaacctgatgtggcatC
IWGSC_CSS_2DS_scaff_5330886	Cadenza0548	5449	G	A	hom	hom	gcatgtccatttatactgaaCgtG	gcatgtccatttatactgaaCgtA	catgctgcttcttctggacC
IWGSC_CSS_3AL_scaff_4449951	Cadenza0548	633	C	T	het	het	tccaaacctaacagtctaacactaG	tccaaacctaacagtctaacactaA	gtctgcagTGCaatgtgC
IWGSC_CSS_3B_scaff_10479889	Cadenza0097	3339	C	T	hom	—	ttgTttctGgagaagatgcCG	ttgTttctGgagaagatgcCA	ggtgctcattcaAcGgcA
IWGSC_CSS_3B_scaff_10562262	Cadenza0097	7819	C	T	het	het	agaggggtgctatccatAttgG	agaggggtgctatccatAttgA	agcgatgccaaggcttcC
IWGSC_CSS_4AL_scaff_7040796	Cadenza0097	10772	G	A	hom	hom	acacaacattgccaccagaG	acacaacattgccaccagaA	CAatCgattgcttgctTctcC
IWGSC_CSS_4AL_scaff_7063488	Cadenza0097	6360	C	T	het	het	gcctctcacCttAatttgaagctgC	gcctctcacCttAatttgaagctgT	aggcagtggagtatgtgaagttT
IWGSC_CSS_4AL_scaff_7091701	Cadenza0097	5050	G	A	het	het	catgagcatctgggaggaaaatG	catgagcatctgggaggaaaatA	agcaagggaAtaatgaacggaaA
IWGSC_CSS_4DS_scaff_1845841	Cadenza0097	7110	G	A	hom	hom	aatgTAgctccccatacCgG	aatgTAgctccccatacCgA	actgaaacTgcaatcgtTtatggA
IWGSC_CSS_5AL_scaff_2767581	Cadenza0097	3737	G	A	het	het	gagaggtcctcactAtcggC	gagaggtcctcactAtcggT	cgTcatcacaaatattgctggG
IWGSC_CSS_5BL_scaff_10784643	Cadenza0097	1568	C	T	hom	hom	agaaaTAcatggatggatggaCG	agaaaTAcatggatggatggaCA	catctcCCttccaCgGaaaG
IWGSC_CSS_1AL_scaff_3952258	Cadenza2092	8107	C	T	het	—	tgagtagagaaattgacagtgtgG	tgagtagagaaattgacagtgtgA	tgccaccattgacatgagaG
IWGSC_CSS_1BL_scaff_3858008	Cadenza2092	10278	G	A	hom	hom	tttgagcaggcaggatcgC	tttgagcaggcaggatcgT	actcacggcctatatcActattC
IWGSC_CSS_1DL_scaff_2265172	Cadenza2092	9094	C	T	hom	hom	tgcaTGTcatttgttcttatcagC	tgcaTGTcatttgttcttatcagT	agtgtccaacttccGttcatC
IWGSC_CSS_2AL_scaff_6435867	Cadenza2092	16201	G	A	hom	hom	tttctgTaccttaacgtcaattgaC	tttctgTaccttaacgtcaattgaT	gtgaggatgatgaggtaagacC
IWGSC_CSS_2AL_scaff_6439430	Cadenza2092	25101	C	T	het	—	caagaaagggCagCtCagC	caagaaagggCagCtCagT	tcGttAcTctttcActggtgaA
IWGSC_CSS_2DL_scaff_9760848	Cadenza2092	4733	C	T	het	het	gcaccatgggtctcaggtaC	gcaccatgggtctcaggtaT	tcagtcagtttGCTCtgTCTG
IWGSC_CSS_3AL_scaff_4407012	Cadenza2092	2785	C	T	hom	hom	acatatAgtgttctcatccaccatC	acatatAgtgttctcatccaccatT	acctctctcatgttaataggtttgT
IWGSC_CSS_3AS_scaff_3441108	Cadenza2092	541	G	A	het	het	GtgatgaccttgagacGgaG	GtgatgaccttgagacGgaA	aggcaTgacaaCgcgcaA
IWGSC_CSS_3B_scaff_10449827	Cadenza1551	4779	G	A	hom	hom	ggcaaggtcaagaaacGgtC	ggcaaggtcaagaaacGgtT	aCagaGtgggttagaggcaG
IWGSC_CSS_3B_scaff_10550638	Cadenza1551	3250	C	T	het	het	ctccttcacttgttgcggC	ctccttcacttgttgcggT	gcaacAtTttgatactgcaaagG
IWGSC_CSS_3DL_scaff_6945816	Cadenza1551	589	C	T	hom	hom	agcatctcacctgcaaCaataC	agcatctcacctgcaaCaataT	TgtgcccTctgaAtattttcaTG
IWGSC_CSS_3DL_scaff_6954177	Cadenza1551	3508	C	T	het	het	tgtagcatcacattaactttcctG	tgtagcatcacattaactttcctA	gcttggtataaaccCttacgacA
IWGSC_CSS_4AS_scaff_5938272	Cadenza1551	19080	G	A	hom	hom	agAcCccgAtcgccatgG	agAcCccgAtcgccatgA	GggAgatAcaggtaaaActcTtcG
IWGSC_CSS_4AS_scaff_5977594	Cadenza1551	11092	C	T	het	het	gccttgattcggaacaacaaaC	gccttgattcggaacaacaaaT	gcgtctctcagtcctgcA

IWGSC contig	Line	Pos	WT	Mut	Predicted	M_4	Primer 1 (Cadenza)	Primer 2 (mutant)	Common Primer
IWGSC_CSS_5AL_scaff_2671035	Cadenza1551	5859	C	T	het	het	cggtgatattTttagacttcgacgC	cggtgatattTttagacttcgacgT	ggcagttcagcGacccatT
IWGSC_CSS_5BL_scaff_10889480	Cadenza1551	2530	G	A	hom	hom	gagcttaactcgcagatggaG	gagcttaactcgcagatggaA	tccatgCAacGccttggT
IWGSC_CSS_3B_scaff_10528396	Cadenza2088	8059	G	A	hom	—	cttttccgtccgtaagcaataG	cttttccgtccgtaagcaataA	gtgcactgttcaggcctgA
IWGSC_CSS_3B_scaff_10637573	Cadenza2088	16815	G	A	het	het	agcaagcttaccGgtctgC	agcaagcttaccGgtctgT	cgagcAactacgagcagctT
IWGSC_CSS_4AL_scaff_7086469	Cadenza2088	6697	G	A	het	het	gccgtctacttcaacgcG	gccgtctacttcaacgcA	ccaGaggcttgtTGcattttT
IWGSC_CSS_4AL_scaff_7126302	Cadenza2088	3627	G	A	hom	hom	gttcaaaaacaagtggctAatttgC	gttcaaaaacaagtggctAatttgT	cacaaggatatgaagcTcttctagA
IWGSC_CSS_4BL_scaff_7041808	Cadenza2088	10234	G	A	hom	hom	tcaatggatgagggtgcttC	tcaatggatgagggtgcttT	ccatagcagcatcagccacA
IWGSC_CSS_5AL_scaff_2794167	Cadenza2088	13162	G	A	het	—	agtattcaggacaagcatCttCaG	agtattcaggacaagcatCttCaA	caatgaaacctctcgaagaaGaG
IWGSC_CSS_5BL_scaff_10889232	Cadenza2088	3885	G	A	het	het	cTcaaccacaatgggcaAatC	cTcaaccacaatgggcaAatT	tccttcatcaatcatcaattgttgG
IWGSC_CSS_5BS_scaff_2267405	Cadenza2088	11113	C	T	hom	hom	ctttgatgatcctaggcctctTG	ctttgatgatcctaggcctctTA	tgatttggtCtggttAgagtttGA
IWGSC_CSS_3B_scaff_10475354	Cadenza1409	2203	G	A	hom	hom	agCgaacaagagGtcaaacG	agCgaacaagagGtcaaacA	ctgaaacacaCtagaCAattAccG
IWGSC_CSS_3B_scaff_10674115	Cadenza1409	4555	C	T	het	het	gcttcagtgcatgccttcaG	gcttcagtgcatgccttcaA	cttcacacccGagataatGtattG
IWGSC_CSS_4AL_scaff_7153568	Cadenza1409	13073	C	T	hom	hom	tccgaccgAtcaaccttgG	tccgaccgAtcaaccttgA	gaccggaactcctcggcC
IWGSC_CSS_4DL_scaff_14314966	Cadenza1409	2010	G	A	het	hom	gtaggtcccctcctCAggG	gtaggtcccctcctCAggA	cggcgTcacaAgttgCcT
IWGSC_CSS_4DS_scaff_2324074	Cadenza1409	7606	G	A	het	het	tGcatgaaaatgtgtGcaGaG	tGcatgaaaatgtgtGcaGaA	gggtaAgttcAaaactGaagtgaaG
IWGSC_CSS_5AS_scaff_1517889	Cadenza1409	3561	G	A	het	het	tctcgacatcttcccgtgtaC	tctcgacatcttcccgtgtaT	gtgcctggaacattgcttatttA
IWGSC_CSS_5AS_scaff_1523866	Cadenza1409	8054	G	A	hom	—	ggtgatctaccgccaGgaC	ggtgatctaccgccaGgaT	tcctgcagCcTctcctcA
IWGSC_CSS_5BL_scaff_10917655	Cadenza1409	19073	G	A	hom	hom	caaatgacatgcaaaagaagttgC	caaatgacatgcaaaagaagttgT	cgcttcatcactacaAaatatgtcT
IWGSC_CSS_1AL_scaff_3886649	Cadenza1599	5204	C	T	het	het	tgatgccaaccacaatGcC	tgatgccaaccacaatGcT	ggactgactgctgaccatatttaG
IWGSC_CSS_1BL_scaff_3810267	Cadenza1599	6634	C	T	hom	hom	ccCaggaaatgagcacctC	ccCaggaaatgagcacctT	cgcaggcgaagatgtgaTtG
IWGSC_CSS_1DL_scaff_2291677	Cadenza1599	12856	C	T	hom	hom	GgtagacaagtcgccgaG	GgtagacaagtcgccgaA	cctcctccttcaacGCcG
IWGSC_CSS_2AL_scaff_6354492	Cadenza1599	7566	G	A	het	het	gGagaatgcaCAgtAacTtctgG	gGagaatgcaCAgtAacTtctgA	ttccgaagaaccacaTccTG
IWGSC_CSS_2AS_scaff_5282937	Cadenza1599	9736	G	A	het	het	gctgtagattttatagctgctatgC	gctgtagattttatagctgctatgT	cacCagaattgttCactgatttTC
IWGSC_CSS_2BL_scaff_7952427	Cadenza1599	19249	G	A	hom	hom	cgTccctCcctagcacgaC	cgTccctCcctagcacgaT	aTcactccattagcgcgAG
IWGSC_CSS_2DL_scaff_9897981	Cadenza1599	5627	C	T	het	het	cttggtgctTgattgcttactC	cttggtgctTgattgcttactT	gTttgctCtctctgatctTtgtG
IWGSC_CSS_3AL_scaff_4446105	Cadenza1599	1765	G	A	hom	—	aaatgctttcctaCcgctagtG	aaatgctttcctaCcgctagtA	ttctAgaggcaatagctTatatgcT

Table A.5: Validation of mutations on M_4 on Kronos

IWGSC contig	Line	Pos	WT	Mut	Predicted	M_4	Primer 1 (Kronos)	Primer 2 (mutant)	Common Primer
IWGSC_CSS_1AS_scaff_3284790	Kronos3085	7449	G	A	Het	Het	ccacaccttgagcctcgC	ccacaccttgagcctcgT	gtgattttgccaggggagA
IWGSC_CSS_1BL_scaff_3897513	Kronos3085	1515	C	T	Het	Het	gcttccactGggtcctgC	gcttccactGggtcctgT	acAaggactgcttcagaGaC
IWGSC_CSS_2AL_scaff_6434745	Kronos3085	3424	C	T	Het	Het	cctcGgttttgcaaatttctatgC	cctcGgttttgcaaatttctatgT	gGCaaTggcataacaacagatA
IWGSC_CSS_3AS_scaff_3408995	Kronos3085	732	C	T	Het	Het	aggccatttcgaattccgC	aggccatttcgaattccgT	ggTgttaTccagAacctgagTG
IWGSC_CSS_3B_scaff_10708748	Kronos3085	2675	G	A	Het	Het	gttgcatgcttcacccagG	gttgcatgcttcacccagA	gtaacaatctgagttcgtagcaC
IWGSC_CSS_4AL_scaff_7132733	Kronos3085	1799	C	T	Hom	Hom	cacccgtgagtgaccctC	cacccgtgagtgaccctT	aCcGcctaGaaagaaagcttC

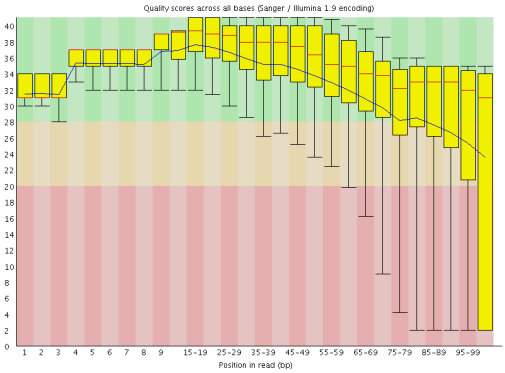
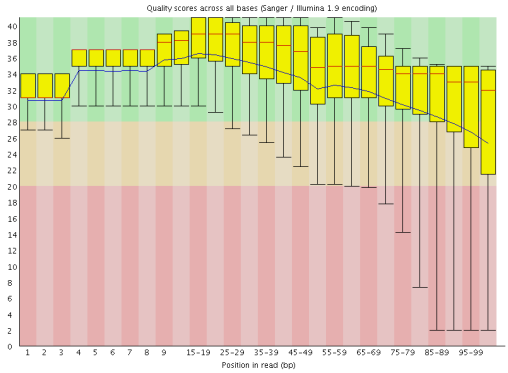
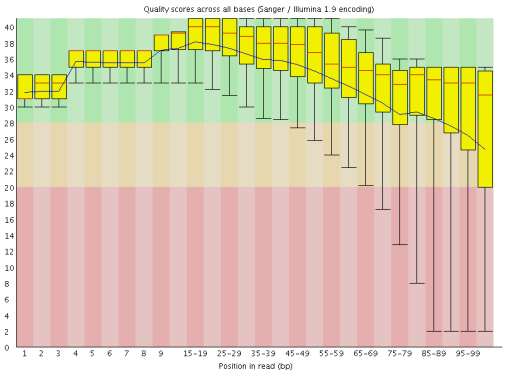
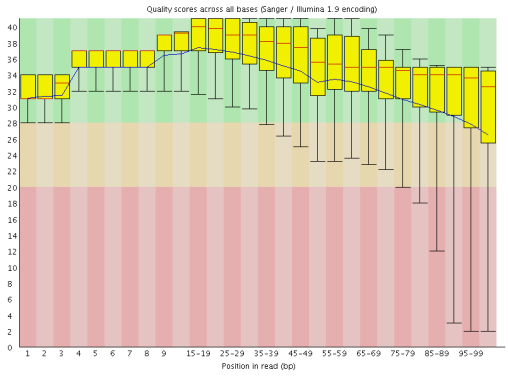
IWGSC contig	Line	Pos	WT	Mut	Predicted	M_4	Primer 1 (Kronos)	Primer 2 (mutant)	Common Primer
IWGSC_CSS_5AS_scaff_1534693	Kronos3085	4605	C	T	Het	Het	cagcttcctggccctcAtC	cagcttcctggccctcAtT	gtaCctcacgAgtcaTgagAG
IWGSC_CSS_6AS_scaff_4361911	Kronos3085	8857	G	A	Het	Het	tcacgaaagacgacttcaacctcC	tcacgaaagacgacttcaacctcT	catgaggtgctgcatctccatcA
IWGSC_CSS_6BS_scaff_3008326	Kronos3085	1528	G	A	Het	Het	ccatgttgtactggtggtgC	ccatgttgtactggtggtgT	ggaagcatggCaagtgcA
IWGSC_CSS_7AS_scaff_4214385	Kronos3085	27835	C	T	Hom	Hom	cgtaccttcgttgggaaagG	cgtaccttcgttgggaaagA	ctcttggtcagctgtataagacT
IWGSC_CSS_1AL_scaff_3929964	Kronos3191	1336	C	T	Het	Het	tttcggccatacctgacatC	tttcggccatacctgacatT	attgcctccagttcttgcaG
IWGSC_CSS_1BL_scaff_3899789	Kronos3191	7925	C	T	Het	Het	actctcacTggcagcagC	actctcacTggcagcagT	caacgtggtgcccatcGtA
IWGSC_CSS_2AL_scaff_6426728	Kronos3191	1481	G	A	Hom	Hom	gaaActgccgcagctCgC	gaaActgccgcagctCgT	ccaGcaGctcgtgagaaA
IWGSC_CSS_2BL_scaff_7960273	Kronos3191	690	C	T	Hom	Hom	gccattcatccttaggcgC	gccattcatccttaggcgT	acatgcaattgctgatgactG
IWGSC_CSS_3AS_scaff_3286603	Kronos3191	2975	G	A	Het*	Hom	ccgtgtggtttgttgtggG	ccgtgtggtttgttgtggA	gaaaggaacgtgTcaTgcaG
IWGSC_CSS_5AL_scaff_2694249	Kronos3191	2399	C	T	Het	Het	gccttccagatagagccGC	gccttccagatagagccGT	cgccacatcgacattcctG
IWGSC_CSS_5BL_scaff_10923577	Kronos3191	3713	C	T	Het	Het	gtggattgcctgagcttgC	gtggattgcctgagcttgT	tggtggccttcttgggaC
IWGSC_CSS_6AL_scaff_5823017	Kronos3191	13225	C	T	Hom	Hom	ccctttcgagcctctggaG	ccctttcgagcctctggaA	ttcgagaaggcccatcgA
IWGSC_CSS_6BS_scaff_2955394	Kronos3191	1622	C	T	Het*	Hom	gtggagatgaaggtctagcaaG	gtggagatgaaggtctagcaaA	gatactcgTgcaatgggtgT
IWGSC_CSS_7BL_scaff_6739382	Kronos3191	12261	G	A	Hom	Hom	gagacaagctttgaattgctcC	gagacaagctttgaattgctcT	CgagtgacctTcatttcccG
IWGSC_CSS_1AS_scaff_3276389	Kronos3288	9720	C	T	Hom	Hom	aCcaGcaggaccAatgtctC	aCcaGcaggaccAatgtctT	atgatgcaacctcagccaT
IWGSC_CSS_2AL_scaff_6367515	Kronos3288	6976	G	A	Het	Het	caggtcgagTgtctccgG	caggtcgagTgtctccgA	ggggtgatCtggaagggC
IWGSC_CSS_2AL_scaff_6422019	Kronos3288	4523	G	A	Het	Het	cgctaggtccctgcatagG	cgctaggtccctgcatagA	acgcAcgctaagccgtaC
IWGSC_CSS_3AL_scaff_4284850	Kronos3288	7901	C	T	Hom	Hom	tggctttggacaacatcgG	tggctttggacaacatcgA	tgtcAgcatcgacagccaG
IWGSC_CSS_4AS_scaff_5962359	Kronos3288	13049	G	A	Het	Hom	ccatcaagaagtacgagttcgaC	ccatcaagaagtacgagttcgaT	accatgcccagcttgtcA
IWGSC_CSS_6AL_scaff_5778773	Kronos3288	6853	G	A	Het	Het	gagtgaccttcccgtctttC	gagtgaccttcccgtctttT	ggagaacagctactcggcT
IWGSC_CSS_6AS_scaff_4392100	Kronos3288	3434	C	T	Het	Het	atggaagcacaggtgaccG	atggaagcacaggtgaccA	ggAagcgaaagtgaacaaacA
IWGSC_CSS_7BL_scaff_6744240	Kronos3288	9772	G	A	Het	Het	agctgttcttctcctacttcaaG	agctgttcttctcctacttcaaA	caggtcgttcttgagctcC
IWGSC_CSS_1AL_scaff_3887185	Kronos3413	9708	C	T	Hom	Hom	gcacgcctttatcgaggtaaaG	gcacgcctttatcgaggtaaaA	AgaaacagcagagcgcaA
IWGSC_CSS_2BS_scaff_3381362	Kronos3413	5160	C	T	Het*	Hom	caacttctgggctgtagtgtG	caacttctgggctgtagtgtA	tgAgaattctgacGcaaaagaC
IWGSC_CSS_3AS_scaff_3296605	Kronos3413	6154	G	A	Het	Het	ctggtcacgggctctagC	ctggtcacgggctctagT	cagcactgagagacatggaC
IWGSC_CSS_3B_scaff_10693516	Kronos3413	12632	C	T	Het	Het	ctaggcttggacaaacaggC	ctaggcttggacaaacaggT	agcttgcatctatgggcatT
IWGSC_CSS_5AS_scaff_1547699	Kronos3413	2686	G	A	Het	Het	gCtacaaccttcaccaatcgC	gCtacaaccttcaccaatcgT	gacggctttgaagtgtcatC
IWGSC_CSS_5BL_scaff_10856077	Kronos3413	5853	G	A	Het	Het	agagcttcaccccatgctC	agagcttcaccccatgctT	acgCacatttAatagctgaagC
IWGSC_CSS_6AL_scaff_5750718	Kronos3413	11046	G	A	Hom	Hom	cacgcTtcccgacttcttataG	cacgcTtcccgacttcttataA	AgacgatgtgatcaggattcaG
IWGSC_CSS_7AL_scaff_4433177	Kronos3413	3511	C	T	Het	Het	GaTgctccGtcaggctgG	GaTgctccGtcaggctgA	cactactggacaagctcttgG
IWGSC_CSS_7BL_scaff_6742567	Kronos3413	667	C	T	Het	Het	gttgcttgcgtggcagaC	gttgcttgcgtggcagaT	cattttgcaccgtgtgtcTG
IWGSC_CSS_1AL_scaff_3976389	Kronos3935	10941	C	T	Hom	Hom	ggtgaggagatcggCgatG	ggtgaggagatcggCgatA	cagtcatctacatgagaggtcaG
IWGSC_CSS_1BL_scaff_3873362	Kronos3935	1392	G	A	Het	Het	cagatctgaagcctaGcacatG	cagatctgaagcctaGcacatA	actaccagaatcagcacaaaaAC
IWGSC_CSS_2BL_scaff_7882382	Kronos3935	2721	C	T	Het	Het	gcaagctaagatgtaccgtagC	gcaagctaagatgtaccgtagT	gccacagtaggagaaagactT
IWGSC_CSS_3AL_scaff_4242376	Kronos3935	2410	C	T	Het	Het	agaacccaaaacccgTacttaG	agaacccaaaacccgTacttaA	gtagGgtCcatcCtaaagcttG
IWGSC_CSS_3B_scaff_10485067	Kronos3935	3349	C	T	Hom	Hom	gcttgagcaactactccaactG	gcttgagcaactactccaactA	gcaatttcctttaTccgcagT
IWGSC_CSS_4AS_scaff_5984153	Kronos3935	6006	G	A	Het	Het	agCaggtctggccaagttG	agCaggtctggccaagttA	cgaatGtatgaGtaggcgcT
IWGSC_CSS_4BL_scaff_7019402	Kronos3935	9081	C	T	Het	Het	tgcaatcatgtagtgagctgG	tgcaatcatgtagtgagctgA	agcatgatccctagaaCcataC
IWGSC_CSS_5BL_scaff_10842786	Kronos3935	3304	G	A	Het	Het	tggttcccGaagcctgaaC	tggttcccGaagcctgaaT	cgcatacttgaaacaTGagcAC
IWGSC_CSS_6BS_scaff_3045205	Kronos3935	2293	G	A	Het	Het	aaggaccaagcccaaactctcG	aaggaccaagcccaaactctcA	agtgatcaagcccaatgtcgcA

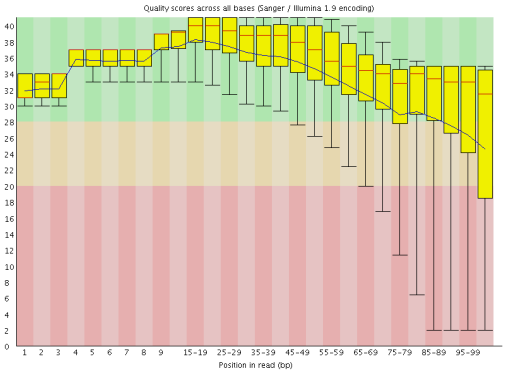
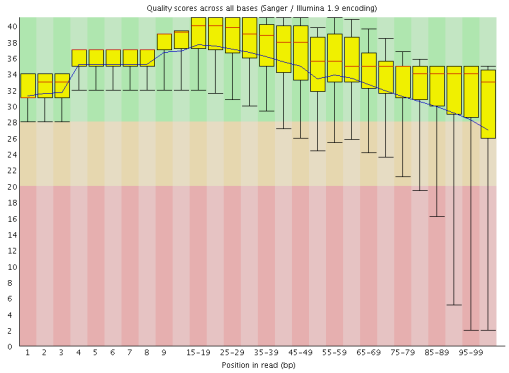
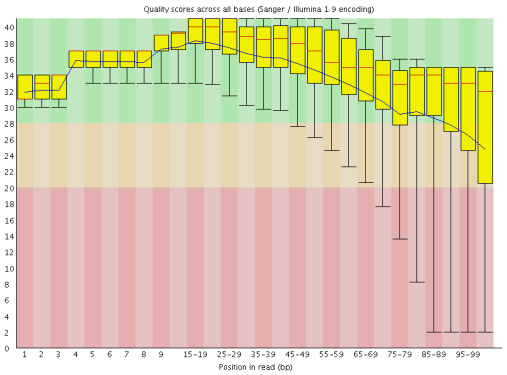
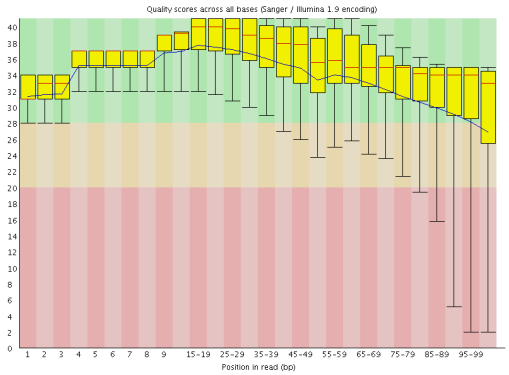
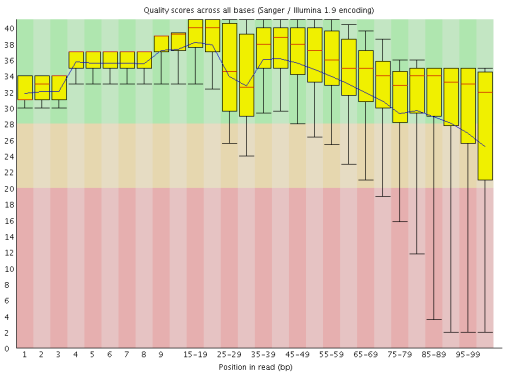
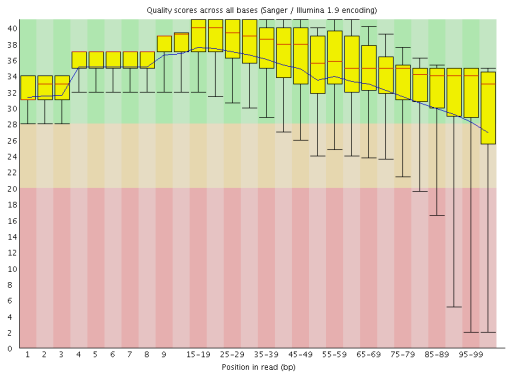
IWGSC contig	Line	Pos	WT	Mut	Predicted	M_4	Primer 1 (Kronos)	Primer 2 (mutant)	Common Primer
IWGSC_CSS_7AL_scaff_4555249	Kronos3935	4487	C	T	Het	Het	cAgtgctcgagatggcgC	cAgtgctcgagatggcgT	cCttgcaaccctcctgatT
IWGSC_CSS_1BL_scaff_3918498	Kronos4240	6096	G	A	Het	Het	ttgcatgccccaagaagaG	ttgcatgccccaagaagaA	tgggcgaactggtaatgtgG
IWGSC_CSS_2BS_scaff_5131713	Kronos4240	5900	G	A	Het	Het	cctttatcgaggaaagagacacC	cctttatcgaggaaagagacacT	caccattgtagggttccttTttC
IWGSC_CSS_5AL_scaff_2769540	Kronos4240	9626	C	T	Het	Het	tgCagtgtgggaaacggaG	tgCagtgtgggaaacggaA	catgagtGagatcttcctgcT
IWGSC_CSS_5BL_scaff_10871091	Kronos4240	7062	G	A	Het	Het	gccaaggAaccataacctgC	gccaaggAaccataacctgT	GgactcttggcAaccggA
IWGSC_CSS_6AL_scaff_5800333	Kronos4240	2360	G	A	Het	Het	cgacaggattgtgagCgC	cgacaggattgtgagCgT	tcagatgctgcaagattcatcT
IWGSC_CSS_7BL_scaff_6716931	Kronos4240	2613	G	A	Het	Het	gGtgGgtattTgcttggtgaG	gGtgGgtattTgcttggtgaA	tgGtggactcgacaGtGtA
IWGSC_CSS_2BL_scaff_8029221	Kronos4346	2860	G	A	Het	Het	tgcttccgctcttgctcC	tgcttccgctcttgctcT	atTtgcatTCgAtcgggcC
IWGSC_CSS_3B_scaff_10460714	Kronos4346	14359	C	T	Hom	Hom	ctaccttgccatgcgacatG	ctaccttgccatgcgacatA	agcaccccagtctttgacG
IWGSC_CSS_4AS_scaff_5989735	Kronos4346	6404	G	A	Hom	Hom	acgcatgctaacatcagcC	acgcatgctaacatcagcT	actcaagataccaCcgcacG
IWGSC_CSS_5BL_scaff_7648030	Kronos4346	6893	C	T	Het	Het	taccctttcctactggcagG	taccctttcctactggcagA	ttttcagaggaacacaggtatcA
IWGSC_CSS_6AL_scaff_5755840	Kronos4346	778	C	T	Het	Het	atcgagtaagctgtcacCgC	atcgagtaagctgtcacCgT	acctgcatgtcaCatccaC
IWGSC_CSS_6BS_scaff_2972151	Kronos4346	7876	G	A	Hom	Hom	gcagcaatgtcActgtttgG	gcagcaatgtcActgtttgA	gcttggactgggcatttatG
IWGSC_CSS_7AL_scaff_4542983	Kronos4346	18700	G	A	Het	Het	gcagggctAccggatacC	gcagggctAccggatacT	catctgccGgttaaacatgC
IWGSC_CSS_7BS_scaff_3098098	Kronos4346	5183	C	T	Het	Het	gCgatatggtacttgcaatgaG	gCgatatggtacttgcaatgaA	ttacattgcttataGTttgCcgG
IWGSC_CSS_1AS_scaff_3259804	Kronos4485	219	C	T	Het	Het	gtcggcacaacccctgC	gtcggcacaacccctgT	gcttctttaaggagggcgA
IWGSC_CSS_2AL_scaff_6315418	Kronos4485	10490	G	A	Hom	Hom	gcccctctcaaCcttctcagC	gcccctctcaaCcttctcagT	ttcagacgctCgaggaatttccC
IWGSC_CSS_2BS_scaff_5181092	Kronos4485	3742	G	A	Het	Het	TggccagcacacctgcaG	TggccagcacacctgcaA	tggacgatgagTgatggAaaT
IWGSC_CSS_3B_scaff_10425015	Kronos4485	2372	C	T	Het	Het	gctactgaagttggCtcGG	gctactgaagttggCtcGA	cttcacatccttgggggTtC
IWGSC_CSS_3B_scaff_10775915	Kronos4485	4701	C	T	Het	Het	ccaagggctgcagagagG	ccaagggctgcagagagA	agacctcacgatGtcctcC
IWGSC_CSS_5AL_scaff_2754304	Kronos4485	2301	G	A	Het	Het	taacccTgccatcgcccG	taacccTgccatcgcccA	cattgGccagccaTgacT
IWGSC_CSS_5BL_scaff_10919959	Kronos4485	1867	C	T	Hom	Hom	gatgccctttgtggagaagG	gatgccctttgtggagaagA	tcttgttcccgaaacatgtcA
IWGSC_CSS_7AS_scaff_4245431	Kronos4485	3402	G	A	Hom	Hom	aaggcgcctggtgtttcC	aaggcgcctggtgtttcT	agtaagtggaAcagctaagatcaT
IWGSC_CSS_7BL_scaff_6667357	Kronos4485	641	C	T	Het	Het	gatcAgctgctcattcgagG	gatcAgctgctcattcgagA	ttccctgtcaattgatgccC

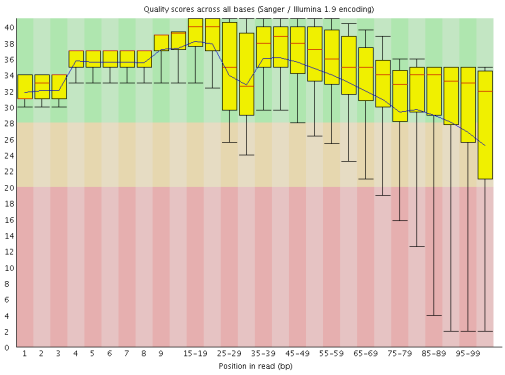
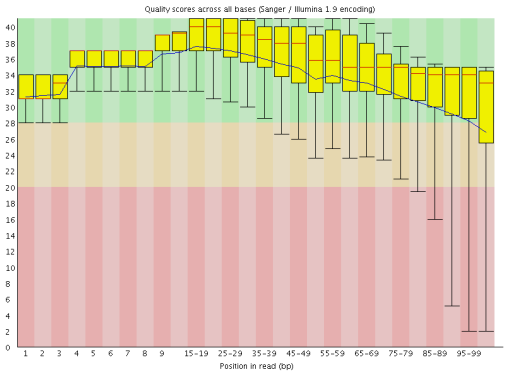
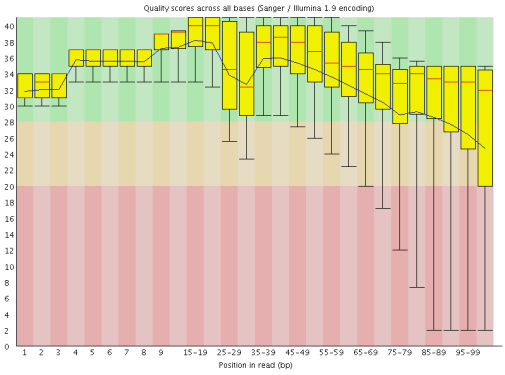
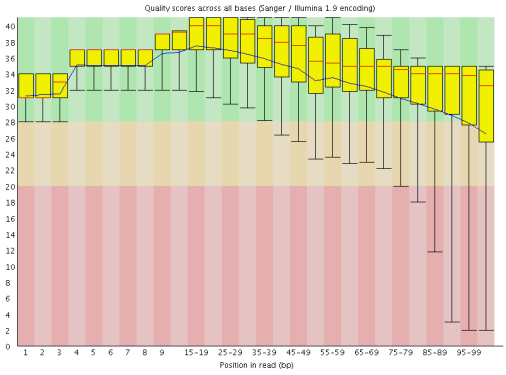
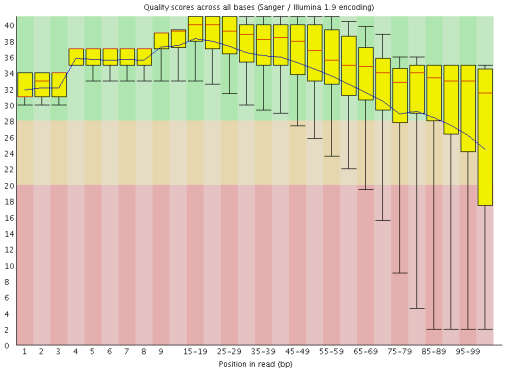
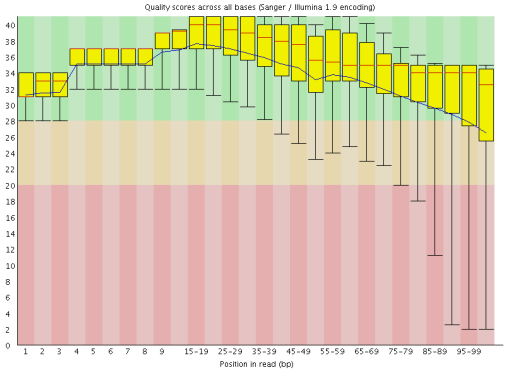
Appendix B

Quality Control

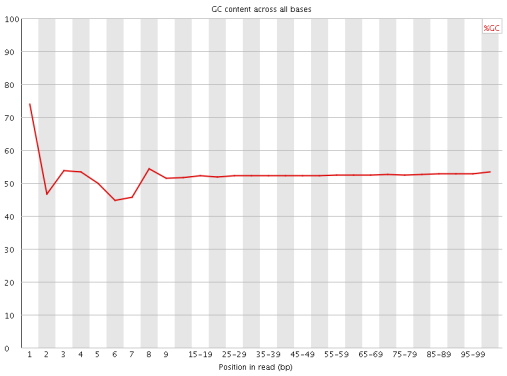
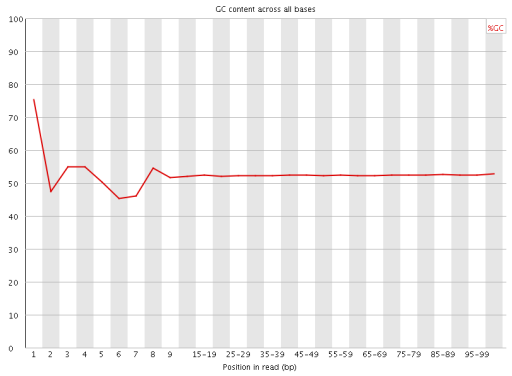
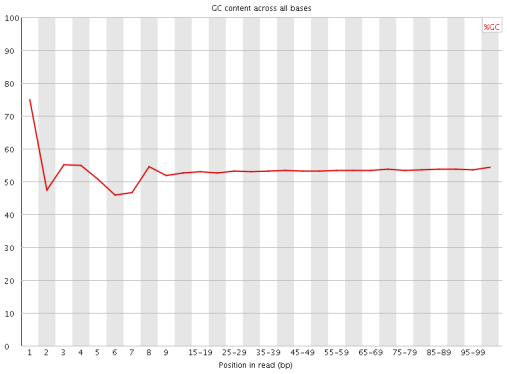
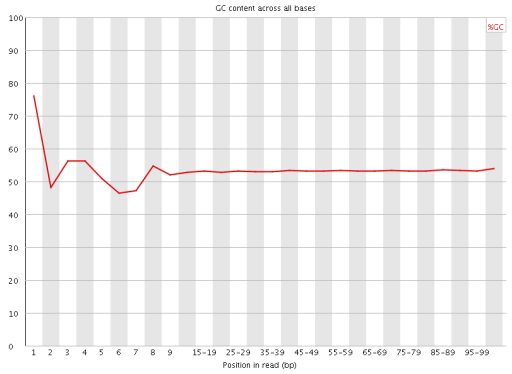
B.1 Sequence read quality

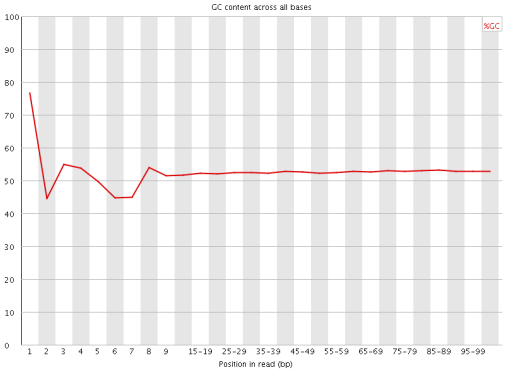
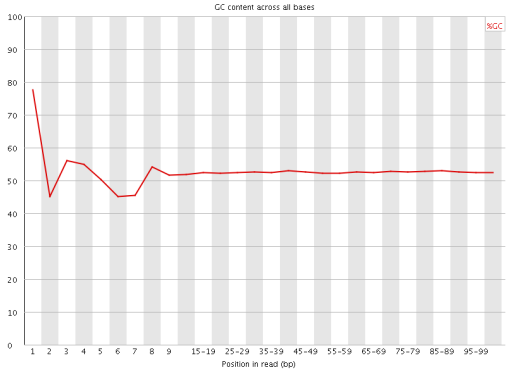
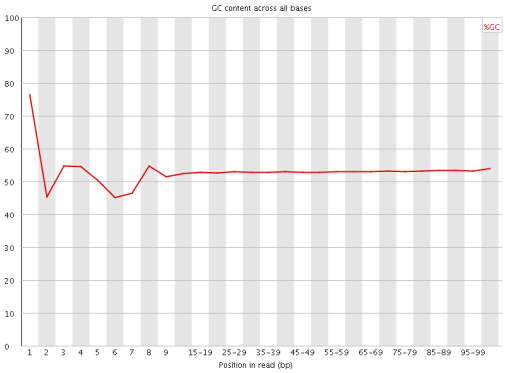
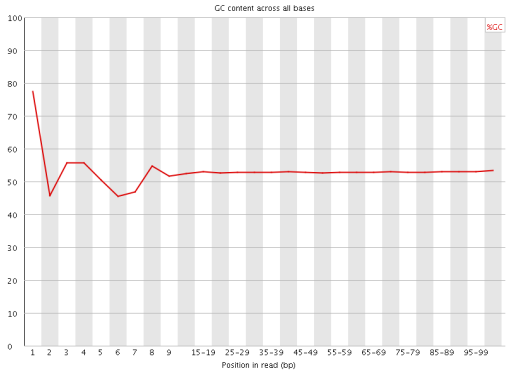
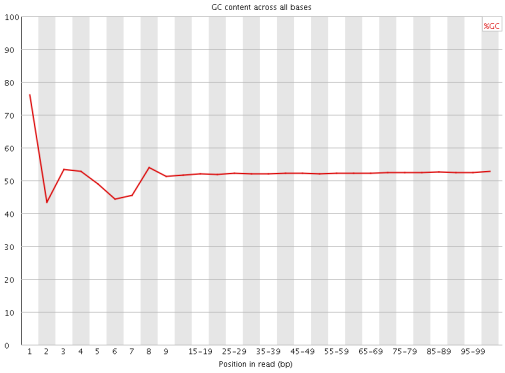
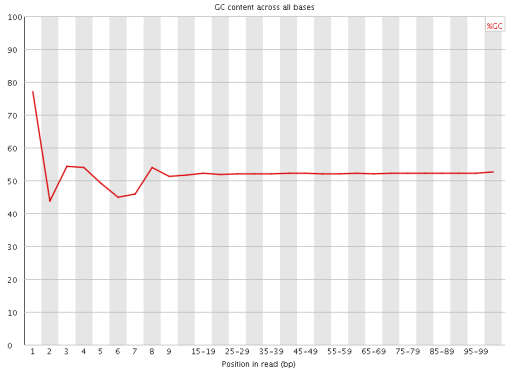
Sample	Read 1	Read 2
LIB1721 AvocetS		
LIB1722 AvocetS(Yr15)		

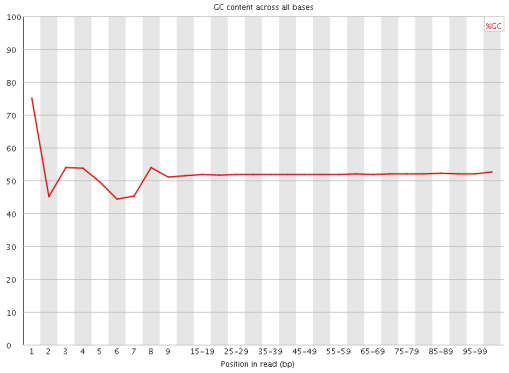
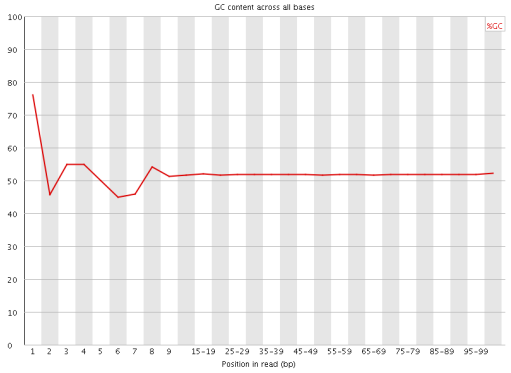
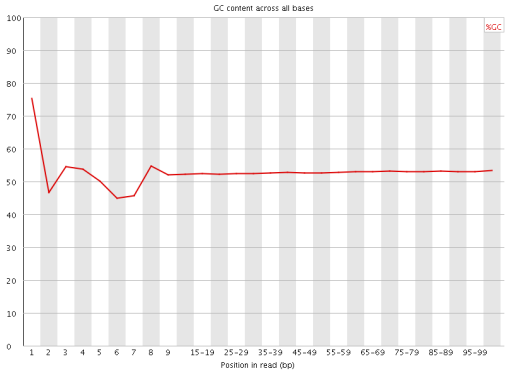
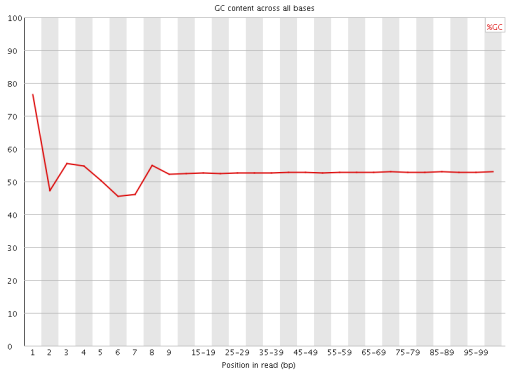
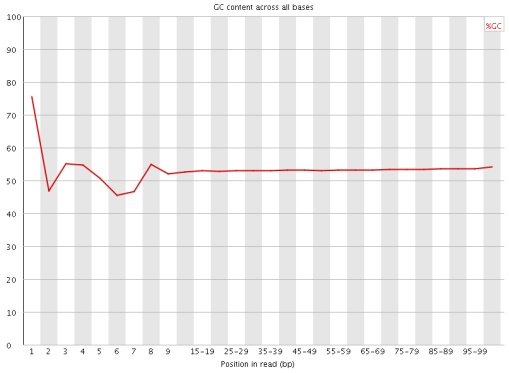
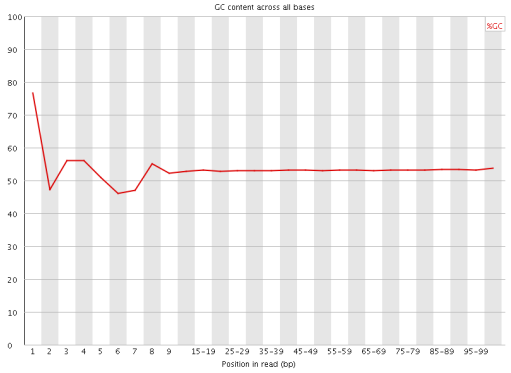
Sample	Read 1	Read 2
LIB1715 Bulk R1		
LIB1716 Bulk R2		
LIB1717 Bulk R3		

Sample	Read 1	Read 2
LIB1718 Bulk S1		
LIB1719 Bulk S2		
LIB1720 Bulk S3		

B.2 Sequence GC content

Sample	Read 1	Read 2
LIB1721 AvocetS		
LIB1722 AvocetS(Yr15)		

Sample	Read 1	Read 2
LIB1715 Bulk R1		
LIB1716 Bulk R2		
LIB1717 Bulk R3		

Sample	Read 1	Read 2
LIB1718 Bulk S1		
LIB1719 Bulk S2		
LIB1720 Bulk S3		

Bibliography

Allen, A. M., Barker, G. L. a., Berry, S. T., et al. Transcript-specific, single-nucleotide polymorphism discovery and linkage analysis in hexaploid bread wheat (Triticum aestivum L.). *Plant biotechnology journal*, 9(9):1086–99, December 2011. ISSN 1467-7652. doi: 10.1111/j.1467-7652.2011.00628.x.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. Basic local alignment search tool. *Journal of molecular biology*, 215 (3):403–10, October 1990. ISSN 0022-2836. doi: 10.1016/S0022-2836(05)80360-2.

Babraham Bioinformatics. FastQC A Quality Control tool for High Throughput Sequence Data, 2012. URL `http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`.

Bonnal, R. J. P., Aerts, J., Githinji, G., et al. Biogem: An effective tool-based approach for scaling up open source software development in bioinformatics. *Bioinformatics*, 28(7):1035–1037, April 2012. ISSN 13674803. doi: 10.1093/bioinformatics/bts080.

Burt, C., Steed, A., Gosman, N., et al. Mapping a type 1 fhb resistance on chromosome 4as of triticum macha and deployment in combination with two type 2 resistances. *Theoretical and Applied Genetics*, 128(9): 1725–1738, 2015. ISSN 1432-2242. doi: 10.1007/s00122-015-2542-9.

Cantu, D., Govindarajulu, M., Kozik, A., et al. Next generation sequencing provides rapid access to the genome of *Puccinia striiformis* f. sp. *tritici*, the causal agent of wheat stripe rust. *PLoS ONE*, 6(8):1–8, 08 2011. doi: 10.1371/journal.pone.0024230.

Chapman, J. a., Mascher, M., Buluç, A., et al. A whole-genome shotgun approach for assembling and anchoring the hexaploid bread wheat

genome. *Genome Biology*, 16(1):1–17, 2015. ISSN 1465-6906. doi: 10.1186/s13059-015-0582-8.

Cornish-Bowden, A. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic acids research*, 13(9):3021–30, May 1985. ISSN 0305-1048.

Eddy, S. R. What is a hidden Markov model? *Nature biotechnology*, 22 (10):1315–6, October 2004. ISSN 1087-0156. doi: 10.1038/nbt1004-1315.

Etherington, G. J., Ramirez-Gonzalez, R. H., and MacLean, D. bio-samtools 2: a package for analysis and visualization of sequence and alignment data with SAMtools in Ruby. *Bioinformatics*, pages 1–2, 2015. ISSN 1367-4803. doi: 10.1093/bioinformatics/btv178.

Gerechter-Amitai, Z. K., van Silfhout, C. H., Grama, A., and Kleitman, F. Yr15 — a new gene for resistance to puccinia striiformis in triticum dicoccoides sel. g-25. *Euphytica*, 43(1):187–190, 1989. ISSN 1573-5060. doi: 10.1007/BF00037912.

GM, C. Chloroplasts and Other Plastids. In *The Cell: A Molecular Approach*. Sinauer Associates, 2000. URL `http://www.ncbi.nlm.nih.gov/books/NBK9905/?redirect-on-error=__HOME__`.

Goto, N., Prins, P., Nakao, M., et al. BioRuby: bioinformatics software for the Ruby programming language. *Bioinformatics (Oxford, England)*, 26(20):2617–9, October 2010. ISSN 1367-4811. doi: 10.1093/bioinformatics/btq475.

Hodges, E., Xuan, Z., Balija, V., et al. Genome-wide in situ exon capture for selective resequencing. *Nat Genet*, 39(12):1522–1527, Dec 2007. ISSN 1061-4036. doi: 10.1038/ng.2007.42.

Hubbard, A., Lewis, C. M., Yoshida, K., et al. Field pathogenomics reveals the emergence of a diverse wheat yellow rust population. *Genome Biology*, 16(1):1–15, 2015. ISSN 1465-6906. doi: 10.1186/s13059-015-0590-8.

Hutchison, C. a. DNA sequencing: bench to bedside and beyond. *Nucleic acids research*, 35(18):6227–37, January 2007. ISSN 1362-4962. doi: 10.1093/nar/gkm688.

Illumina. *CASAVA v1.8.2 User guide*. Revc edition, 2011. URL http://support.illumina.com/sequencing/sequencing_software/casava/documentation

Katoh, K. and Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, 30(4):772–80, April 2013. ISSN 1537-1719. doi: 10.1093/molbev/mst010.

Kent, W. J. BLAT—The BLAST-Like Alignment Tool. *Genome Research*, 12(4):656–664, March 2002. ISSN 1088-9051. doi: 10.1101/gr.229202.

King, R., Bird, N., Ramirez-Gonzalez, R., et al. Mutation scanning in wheat by exon capture and next-generation sequencing. *PLoS ONE*, 10(9):1–18, 09 2015. doi: 10.1371/journal.pone.0137549.

Krasileva, K., Vasquez-Gross, H., Howell1, T., et al. Uncovering hidden variation in young polyploid wheat genomes. submitted 2016.

Krasileva, K. V., Buffalo, V., Bailey, P., et al. Separating homeologs by phasing in the tetraploid wheat transcriptome. *Genome biology*, 14(6): R66, June 2013. ISSN 1465-6914. doi: 10.1186/gb-2013-14-6-r66.

Lander, E. S., Linton, L. M., Birren, B., et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, February 2001. ISSN 0028-0836. doi: 10.1038/35057062.

LGC Genomics. http://www.lgcgroup.com/services/genotyping/, 2013. URL http://www.lgcgroup.com/services/genotyping/.

Li, H. and Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 25 (14):1754–60, July 2009. ISSN 1367-4811. doi: 10.1093/bioinformatics/btp324.

Li, H., Handsaker, B., Wysoker, A., et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16):2078–9, August 2009. ISSN 1367-4811. doi: 10.1093/bioinformatics/btp352.

Liu, Y. and Schmidt, B. Long read alignment based on maximal exact match seeds. *Bioinformatics (Oxford, England)*, 28(18):i318–i324, September 2012. ISSN 1367-4811. doi: 10.1093/bioinformatics/bts414.

Ma, J., Stiller, J., Zheng, Z., et al. A high-throughput pipeline for detecting locus-specific polymorphism in hexaploid wheat (triticum aestivum l.). *Plant Methods*, 11(1), aug 2015. doi: 10.1186/s13007-015-0082-6.

MAS Wheat. Mas wheat transcriptome. supplemental file 17., 2013. URL `http://maswheat.ucdavis.edu/Transcriptome/index.htm`.

Mayer, K. F. X., Rogers, J., Dole el, J., et al. A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. *Science*, 345(6194):1251788–1251788, July 2014. ISSN 0036-8075. doi: 10.1126/science.1251788.

Metzker, M. L. Sequencing technologies - the next generation. *Nature reviews. Genetics*, 11(1):31–46, January 2010. ISSN 1471-0064. doi: 10.1038/nrg2626.

Michelmore, R. W., Paran, I., and Kesseli, R. V. Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. *Proceedings of the National Academy of Sciences*, 88(21):9828–9832, November 1991. ISSN 0027-8424. doi: 10.1073/pnas.88.21.9828.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, 5(7):621–8, July 2008. ISSN 1548-7105. doi: 10.1038/nmeth.1226.

Myllykangas, S., Buenrostro, J., and Ji, H. P. *Bioinformatics for High Throughput Sequencing*. Springer New York, New York, NY, 2012. ISBN 978-1-4614-0781-2. doi: 10.1007/978-1-4614-0782-9. URL `http://www.springerlink.com/index/10.1007/978-1-4614-0782-9`.

Needleman, S. B. and Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453, March 1970. ISSN 00222836. doi: 10.1016/0022-2836(70)90057-4.

Pickrell, J. K., Marioni, J. C., Pai, A. A., et al. Understanding mechanisms underlying human gene expression variation with rna sequencing. *Nature*, 464(7289):768–772, Apr 2010. ISSN 0028-0836. doi: 10.1038/nature08872.

Pontius, J., Wagner, L., and Schuler, G. UniGene: A Unified View of the Transcriptome. In *The NCBI Handbook [Internet]*, chapter 21. National Center for Biotechnology Information (US), October 2002. URL http://www.ncbi.nlm.nih.gov/books/NBK21083/.

Ramirez-Gonzalez, R. H., Uauy, C., and Caccamo, M. PolyMarker: A fast polyploid primer design pipeline. *Bioinformatics*, pages 2–3, 2015a. ISSN 1367-4803. doi: 10.1093/bioinformatics/btv069.

Ramirez-Gonzalez, R. H., Bonnal, R., Caccamo, M., and Maclean, D. Bio-samtools: Ruby bindings for SAMtools, a library for accessing BAM files containing high-throughput sequence alignments. *Source code for biology and medicine*, 7(1):6, January 2012. ISSN 1751-0473. doi: 10.1186/1751-0473-7-6.

Ramirez-Gonzalez, R. H., Segovia, V., Bird, N., Caccamo, M., and Uauy, C. *Next Generation Sequencing Enabled Genetics in Hexaploid Wheat*, pages 201–209. Springer Japan, Tokyo, 2015b. ISBN 978-4-431-55675-6. doi: 10.1007/978-4-431-55675-6_22. URL http://dx.doi.org/10.1007/978-4-431-55675-6_22.

Ramirez-Gonzalez, R. H., Segovia, V., Bird, N., et al. Rna-seq bulked segregant analysis enables the identification of high-resolution genetic markers for breeding in hexaploid wheat. *Plant Biotechnology Journal*, 13(5):613–624, 2015c. ISSN 1467-7652. doi: 10.1111/pbi.12281.

Randhawa, H. S., Mutti, J. S., Kidwell, K., et al. Rapid and targeted introgression of genes into popular wheat cultivars using marker-assisted background selection. *PloS one*, 4(6):e5752, January 2009. ISSN 1932-6203. doi: 10.1371/journal.pone.0005752.

Rozen, S. and Skaletsky, H. Primer3 on the WWW for general users and for biologist programmers. *Methods in molecular biology (Clifton, N.J.)*, 132:365–86, January 2000. ISSN 1064-3745.

Schneeberger, K., Ossowski, S., Lanz, C., et al. Shoremap: simultaneous mapping and mutation identification by deep sequencing. *Nat Meth*, 6 (8):550–551, Aug 2009. ISSN 1548-7091. doi: 10.1038/nmeth0809-550.

Shendure, J. and Ji, H. Next-generation DNA sequencing. *Nature biotechnology*, 26(10):1135–45, October 2008. ISSN 1546-1696. doi: 10.1038/nbt1486.

Slater, G. S. C. and Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC bioinformatics*, 6:31, January 2005. ISSN 1471-2105. doi: 10.1186/1471-2105-6-31.

Smith, T. and Waterman, M. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197, March 1981. ISSN 00222836. doi: 10.1016/0022-2836(81)90087-5.

Takagi, H., Uemura, A., Yaegashi, H., et al. MutMap-Gap: whole-genome resequencing of mutant F2 progeny bulk combined with de novo assembly of gap regions identifies the rice blast resistance gene Pii. *The New phytologist*, 200(1):276–83, October 2013. ISSN 1469-8137. doi: 10.1111/nph.12369.

Trick, M., Adamski, N., Mugford, S. G., et al. Combining SNP discovery from next-generation sequencing data with bulked segregant analysis (BSA) to fine-map genes in polyploid wheat. *BMC plant biology*, 12 (1):14, January 2012. ISSN 1471-2229. doi: 10.1186/1471-2229-12-14.

Van Ooijen, J. and Jansen, J. *Estimation of recombination frequencies; Genetic Mapping in Experimental Populations*, pages 73–133. Cambridge University Press, Cambridge, 2013. ISBN 978-1-107-0132-16.

Wang, S., Wong, D., Forrest, K., et al. Characterization of polyploid wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array. *Plant biotechnology journal*, 12(6):787–796, March 2014. ISSN 1467-7652. doi: 10.1111/pbi.12183.

Wang, Y., Tiwari, V. K., Rawat, N., et al. GSP: a web-based platform for designing genome-specific primers in polyploids. *Bioinformatics*, page btw134, mar 2016. doi: 10.1093/bioinformatics/btw134.

Wellings, C. and McIntosh, R. A. Host-pathogen studies of wheat stripe rust in australia. In Slinkard, A., editor, *Proceedings 9th International Wheat Genetics Symposium*, pages 336–338. University of Saskatchewan, Saskatoon, SK, Canada, 1998.

Wilkinson, P. a., Winfield, M. O., Barker, G. L. a., et al. CerealsDB 2.0: an integrated resource for plant breeders and scientists. *BMC bioinformatics*, 13(1):219, January 2012. ISSN 1471-2105. doi: 10.1186/1471-2105-13-219.