1 Thesis Outline

The proposed the following outline for my PhD thesis:

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.

2. **Literature Review** It describes the current status of the wheat genome, genetics and other resources.

2.1. **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.

2.2. **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.

2.3. **Wheat Genomics**. A description of the current status of the wheat genome (**?**, **?**), the different available assemblies and and approaches to sort the scaffolds (Genome Zipper, the various genetic maps).

2.4. **Sequencing** The importance of the selection of the library preparation and the sequencing platforms available. A brief summary of RNA-Seq, Exome capture, ChIP-Seq, Whole Genome Shotgun, etc. and on which cases are more suitable for different experiments.

2.5. **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.

2.6. **Wheat online resources**. A compilation of the currently available resource for whet genetics and genomics. MAS wheat, CeralsDB, Ensembl, etc.

3. **Genetic mapping of *Yr15***. This section describes in detail than the paper of **?**

3.1. ***Yr15*** Breeding importance of *Yr15* and original source (an introgression of *T. diccocoides*).

3.2. **Segregating population and resistance essays**. A description of the starting material and how the population was generated.

3.3. **Sequencing and mapping** RNA-Seq and the decision to call SNPs on gene models rather than the whole reference. Details of the mapping against the Wheat UniGenes **?** and the UCW. **?** gene models.

3.4. **SNP Calling**. `Ruby` implementation of the methodology described by **?**.

3.5. **Bulk Frequency Ratios** Results of the simple SNP calls from the progenitors and how the score of the Bulk Frequency Ratios(BFR) improve the location of the SNPs.

3.6. ***In silico* mapping**Mapping of the gene models to the IWGSC CSS **?** reference and the location of the SNPs using the genetic map from **?**.

3.7. **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.

3.8. **Genetic map** The three versions of the genetic map: With a subset of the F_2 population

3.9. **Assembly of the transcriptome** A comparison between the known unigenes and the transcript from the progenitors. Since *Yr15* comes from an introgression with *T. diccocoides*, some novel transcripts can be extracted

4. **PolyMarker** A fast polyploid primer design pipeline

4.1. **Manual process of primer design** Explain how the SNP markers are designed without the tool.

4.2. **Global alignment** Search of the contigs with the sequence in the CSS reference and the importance of being able to distinguish between homoeologous regions.

4.3. **Local alignment** Once the region with the primer has been selected, make a local alignment. This section discusses why the local alignment is needed.

4.4. **Primer design tools** In this section, the principles of *in silico* primer design are discussed, and why not simply selecting a genomic variation is enough (thermal stability, primers folding on themselves)

4.5. **Primer selection algorithms** Different algorithms to select the "best primer". For KASP markers, the product should be as short as possible with the mutation in the first three bases. Other types of PCR require different priorities. A comparison with qPCR primers may be included in this section as well.

4.6. **Designed markers** Details of the generated primers for the 80k iSelect chip and the 820k axiom chip. This section also include counts on how many are genome specific, semi-specific and non specific. Also an analysis of how many are repeated or map to more than one chromosome perfectly.

5. **Data integration**

5.1. **Data sources** A summary of the different sources to be integrated. So far, the following are included:

5.1.1. **Assemblies** IWGSC Chinese Spring sequence survey, Chapman WGS assembly

5.1.2. **Genetic maps** POPSeq genetic map, 42k markers map, Chapman genetic map.

5.1.3. **Genetic markers** 80k iSelect, 820k axiom

5.1.4. **Mutations** Cadenza and Kronos mutations from Uauy's group.

5.1.5. **Wheat lines** Mutant lines from above and lines genotyped with the 820k chip

5.1.6. **Gene annotations** UniGenes, UCW gene models, MIPS annotation from IWGSC CSS.

5.1.7. **Syntenic regions** GeneZipper and alignment to genes in relative species, using the Esembl! databases.

5.2. **Database design** In this section the relationships between all the data sources is discussed and different ways to query the data to get markers in linked regions, mutations common across different lines.

5.3. **Graphical interface** The amount of data and the complex relationships between them require an easy to use interface. This section discusses the considerations on how to make the data available.

6. **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.