Contents

1 Expression quantitative trait loci mapping in Brassica rapa.

1.0.1 Key words: Agriculture, field, expression traits, genomics, quantitative genetics

2 Introduction

In the face of the change in climate we as human beings need to think about how we will grow food to feed 9 billion people. In order to feed these people we need to understand Clinton attics.Traits that are important for plant growth reproduction and development of been studied for a long time. These crates are quantitative and neuter because they involve many genes. A way to map the genotype to the phenotype is by using quantitative genetics. We can use quantitative genetics to understand complex traits By statistically associating the genotype to the phenotype. Two important phenotypes are plant leaf Lane and flowering time. Leaf length is important for how the plant captures energy and how this energy is converted to harvestable yield. Flowering time is an important trait because it determines when in developmental time there will be a large transition. In a previous paper we developed a new genetic map for this population. The new genetic map provided two major advantages. Firstly, it was much denser than the previous map allowing for more refined QTL architecture. Secondly, it provided us with the genomic anchors for each of the molecular markers to convert genetic distance into physical distance along each of the chromosomes. This is very powerful because the genome has been sequenced and the genomic coordinates of each of the genes have been roughly placed on the genome. Therefore if the QTL can be placed in context of physical markers, causitive genes underlying the QTL can theoretically be identified. There are many caveats to this assumption, but advances in statistical modeling provide ways to combine these different pieces of biological information into one coherent model of the underlying process.

One challenging problem facing quantitative genetic studies is finding the actual causative genes. If the assumption held that differences in gene expression underlying a QTL could help to determine what genes are involved in the phenotype then quantifying a genes expression. Messenger RNA is an intermediate form of biological information and is often used as a molecular proxy for cellular processes. The simplist use of mRNA is making comparisons between two groups of treatments to see if gene expression is different on a gene by gene basis. However, with advances in sequencing technology, it is now possible to quantify all of the genes that are being expressed in a given sample using RNA-seq. If this rich source of biological data is quantified on all the individuals in a genetic mapping population, the gene expression values themselves can be treated as molecular traits. This then allows the genetic basis of gene expression to be determined. If there are quantitative genetic loci that are controlling gene expression in a region of the genome that is also controlling a physiological phenotype, these two processes can be correlated with one another to find potential connections.

The genes, proteins, and metabolites in developmental pathways do not act in isolation, but rather in a network. Quantifying any form of biological information within the network provides snapshot views of what is occuring at that level of biological organization. This can be combined with computational modeling in order to infer what might be occuring at a different level of biological organization. Network modeling is a powerful approach towards the goal of combining different pieces of biological information. If network modeling is further combined with other techniques, such as quantitative genetics, the possibilities increase even further.

For this paper, we combined different forms of biological information to narrow down gene candidates for consistant physioligcal and development QTL. We will step through our thought process for collecting and integrating increasingly complex data sets towards finding the potential causative genes underlying these QTL. We will first describe an experiment just involving the parents of the genetic mapping population and then move on to quantifying gene expression across the entire population. At each step we will constantly ask what new form of information does this increasingly complex data give us? Then discuss the limitations of each form of information. We will finish by giving suggestions as to what genes are likely candidates for the QTL of interest.

3 Methods

3.1 What materials did you use?

Leaf and silique tissue from parents in the field. We decided to use the meristem tissue for RNA-seq as it usually gives the best information on what individual plants are "planning" on doing developmentally. We collected this tissue and flash froze it liquid nitrogen. We extracted the RNA and prepared libraries for RNA-sequencing following Ravi's protocol. The sequencing was preformed at the Berkeley Sequencing Facility. The raw reads were quality scored and mapped to the Chifu genome v1.5 . Counts of uniquely mapped reads were generated for each sample.

3.2 Population of the study

This study focused on the IMB211xR500 cross.

3.3 What is the design of your research?

The field site was the Wyoming experimental station. It was a randomized block design. Each genotype in the population has tissue collected from it.

4 Results

4.1 What are your most significant results?

Overview of cis and trans eqtl we get many more cis than trans and get the distinct cis diagonal band (Fig 1) There are 8907 cis eQTL for LOD scores > 4 (Fig 2) There are 3749 Trans eQTL for LOD scores > 4 (Fig 3) GO enrichment for metabolism related genes in the Cis eQTL set (Supp table 1) GO enrichment for defense genes and photosynthesis for the Trans set (Supp table 2)

eQTL hotspots on chromosomes A01, A02, A06, A09, A10 (Fig 4) Hotspots on A02 and A06 line up with flowering time genes (Fig 4) Other hotspots do not. (Fig 4)

Trans hotspots line up on LF1 portion of the genome (Fig 5)

Flowering QTL on A03, A07, A10. Leaf Length QTL on A01, A03, A06, A10.

4.2 What are your supporting results?

Hotspot on 6 effects many genes but signal is more diffuse across the chromosome. Many more flowering genes on this chromosome than the others?

cis effect eqtl with lod over 100: Largest cis-effect eQTL (chromosome A02) is protein of unknown function (LOD 287; Fig 1) Second largest cis-effect (A06) is AT3G49640 (E=9e-177) | FAD binding / catalytic/ tRNA dihydrouridine synthase (LOD 186; Fig 1) Third largest cis effect (A01) starch synthase - AT3G01180 (E=1e-058) AtSS2 | AtSS2 (starch synthase 2); transferase, transferring (LOD 160; Fig 1)

trans effect eqtl with lod over 100: Largest trans-effect eQTL are mostly proteins of unknown function. Is it possible that these are just misplaced in the genome and are actually cis effect. Might be fixed with new mapping?

5 Discussion

Provide a groundwork for combining datasets What can parental gene expression data tell us? What does eQTL data add? What does modeling multi-level datasets allow us to do?

This paper is a follow-up to two previous publications Brock2010 and Covington2016. These papers examined the genetic architecture of the important traits of flowering time and leaf length. Brock2010 quantified these important traits and mapped them using a first version of a genetic map produced by the creators of the mapping population . Covington2016 followed up on both of these papers by creating a new genetic map out of molecular markers derived from RNA-seq data. One of the main conclusions from Covington2016 was that the new genetic map could refine genetic architechture for traits already mapped and provided known genomic locations of the molecular markers used to create the map. These three papers opened up the possibilities of this papers to follow-up and try to estimate genes that are involved in genetic architecture of the trait rather than just the genomic regions. This lays the ground work for combining data sets across biological scales.

Traits that are segregating in a population have a molecular mechanistic underpinning. In the case of a recombinant in-bred line population, there are only two allelic states. This is an advantage because there is less concern about how heterozygous individuals at a locus are manifesting the trait being studied. One of the molecular mechanisms that can explain a trait of interest is differences in gene expression between the two allelic states. Or, to put it another way, the causative gene for the physiological phenotype could be manifest by a difference in how that gene is expressed. If a region of the genome confering the difference in phenotype is known, a simple question is if there are any differences in gene expression between the parents of the population in that region. If there are differences between many genes, how then to choose? Do any of the genes make sense according to what is known about the trait *and* are differentially expressed between the parents? We blah blah blah... need to answer this question first with the data.

What does eQTL data add to the picture? The limitation of using only parental gene expression data is that there may be more subtle differences in expression of the causative gene. Quantifying gene expression in all the individuals in the population is the next step. This allowed us to ask what genes are differentially expressed between the parents that are also differentially expressed in the population between the allelic states. The assumption here is that if there are no differentially expressed genes between the parents for a given genomic location, then either the ability to detect differences is not high enough, or gene expression is not the regulatory level. To add information to this story, we measured gene expression in the population in meristematic tissue pre-flowering. This allowed us to treat gene expression as a quantitative trait. We could then determine if the gene expression differences had the same genetic pattern as the physiological traits of interest. If the pattern is the same, then we used probabilistic modeling to determine if the variance in the population for the physiological trait could be explained by the differences in gene expression. Connecting all of these components was the ultimate goal of doing this large experiment in the first place.

Compare this to other work that has been done in B. rapa. Hotspot on 6 effects many genes but signal is more diffuse across the chromosome. Also, hotspot in Hammond et al. (2011). Wonder if this is the same hotspot? They showed enrichment for P related genes. Test this. Artifact?

If there is sequence level variation in the genes that could explain the other variation observed in the phenotypes, then we should be able to pick up on that variation. We can connect the pieces of information on a per gene basis, or chromosome block basis. Then we can compare those differences to the Provean results. The provean results show whether a SNP is related to an amino acid change, and if that amino acid change is related to differences in function for the protein.

6 Figures

6.1 Figure 1

Figure 1: Whole genome differential gene expression between R500 and IMB211. Displayed is the t-statistic of the difference in expression across X number of tissue types (open symbols) and X number of environments (closed symbols).

6.2 Figure 2

Figure 2: Whole genome expression quantitative trait loci (QTL). This plot displays local (cis) and distal (trans) gene expression. QTL are considered cis if they occur within X distance of the gene's physical position, or trans otherwise. Data points are false colored to represent the Likelihood Odds Ratio (LOD) significance score from black to blue.

6.3 Figure 3

Figure 3: Genome wide *cis* effect expression QTL distribution and significance score. Overlayed are red tick marks for genes differentially expressed in the parental dataset and displaying a significant cis eQTL.

6.4 Figure 4

Figure 4: Genome wide *trans* effect expression QTL distribution and significance score. Blue boxes surround trans eQTL hotspot peaks determined through permutation tests. Red ticks denote genomic location of flowering time genes.

7 References

Hammond, J. P., S. Mayes, H. C. Bowen, N. S. Graham, R. M. Hayden, C. G. Love, W. P. Spracklen, et al. 2011. "Regulatory Hotspots Are Associated with Plant Gene Expression Under Varying Soil Phosphorus Supply in Brassica Rapa." *PLANT PHYSIOLOGY* 156 (3) (July): 1230–1241. doi:10.1104/pp.111.175612. http://www.plantphysiol.org/cgi/doi/10.1104/pp.111.175612.