Part of the transPLANT European Plant Genomics Infrastructure

Part 4

3rd transPLANT Training Workshop - October 2014
Exploiting and understanding Solanaceous genomes

Protein function prediction with BMRF

Sven Warris
Aalt-Jan van Dijk

Plant Research International, WUR

PLANT RESEARCH INTERNATIONAL
WAGENINGEN UR

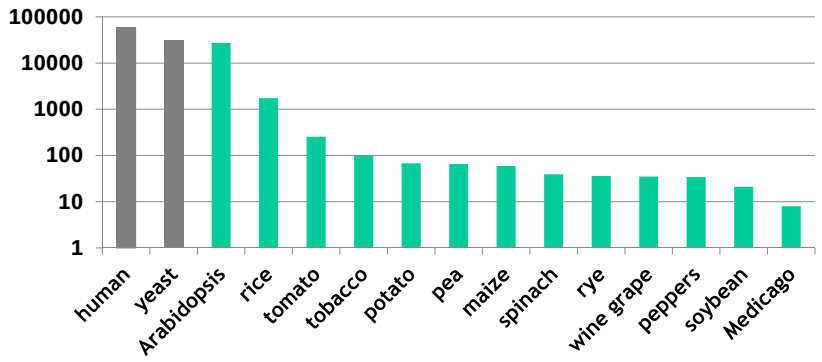
Connecting traits to gene functions

Aalt-Jan van Dijk & Sven Warris
Applied Bioinformatics, PRI, Wageningen UR
aaltjan.vandijk@wur.nl

transPLANT

- Joachim Bargsten
- Jan-Peter Nap
- Gabino Sanchez-Perez

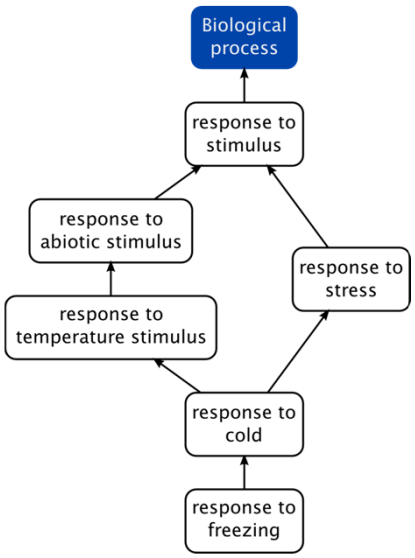
Experimental function annotation in plants

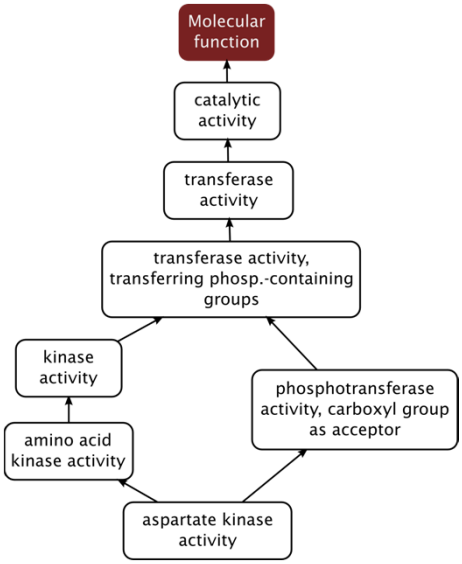


- **How can we identify gene functions at large scale?**

PLANT RESEARCH INTERNATIONAL
WAGENINGEN UR

Gene function as defined by Gene Ontology

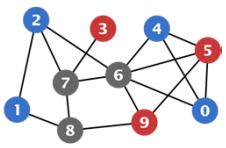




PLANT RESEARCH INTERNATIONAL WAGENINGEN UR

Gene function prediction

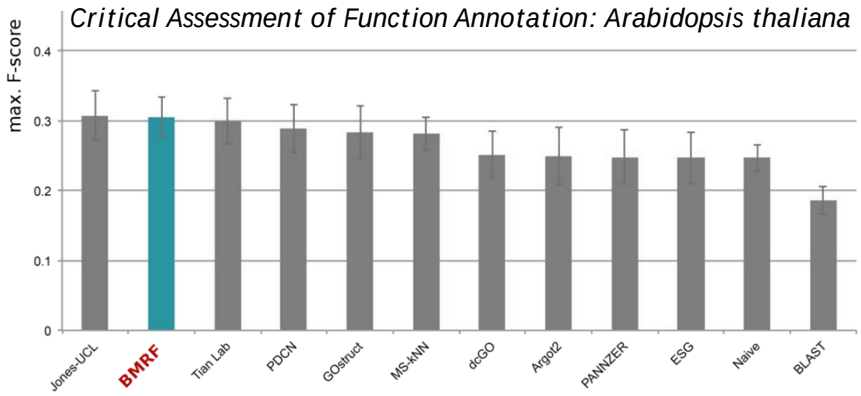
- Molecular function
  - Direct connection to sequence (e.g. domains)
  - Prediction using sequence similarity is fine
- Biological process
  - Indirect connection to sequence
  - Dependent on biological context
  - Our method (BMRF) combines sequence-similarity and co-expression networks



PLANT RESEARCH INTERNATIONAL WAGENINGEN UR

Kourmpetis et al. (2011). Plant Phys.

Prediction performance (I): CAFA

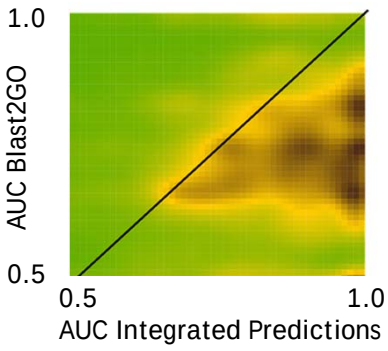


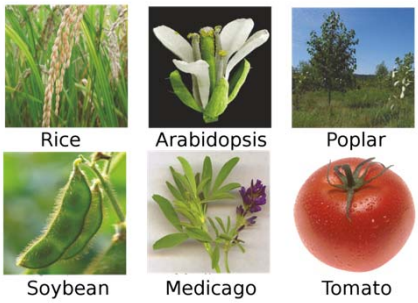
- BMRF: Place 1 in human; 10 and 11 in mouse and rat

PLANT RESEARCH INTERNATIONAL WAGENINGEN UR

Nature Methods 2013

Prediction performance (II): rice





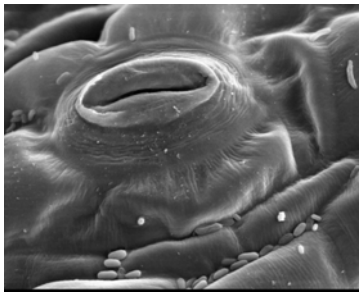
http://www.ab.wur.nl/bmrf

PLANT RESEARCH INTERNATIONAL WAGENINGEN UR

Bargsten et al. Curr Plant Biol, in press

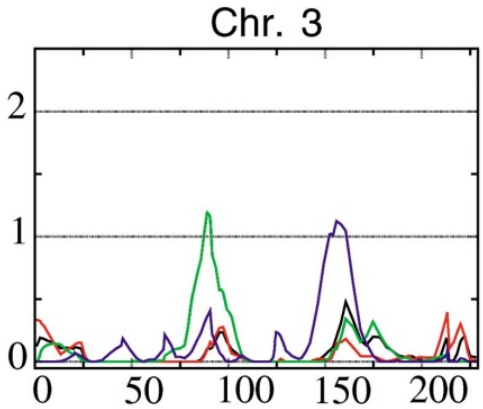
Case study

- Trait: stomatal density
  - Gene function prediction
  - Connecting QTL data with predicted gene functions
- Handout page 67-73
- Start: Questions #1 and #2



PLANT RESEARCH INTERNATIONAL
WAGENINGEN UR

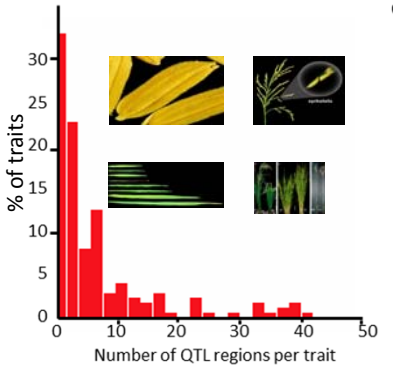
Quantitative Trait Loci (QTL) mapping



PLANT RESEARCH INTERNATIONAL
WAGENINGEN UR

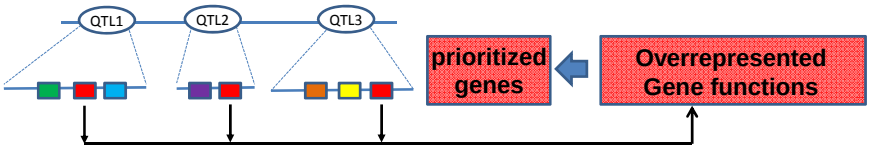
Datasets

	#traits	#genes per region
Rice	179	140+/-121
Tomato	13	148+/-119



PLANT RESEARCH INTERNATIONAL
WAGENINGEN UR

Overrepresentation analysis



PLANT RESEARCH INTERNATIONAL
WAGENINGEN UR

Linking macroscopic traits with gene functions

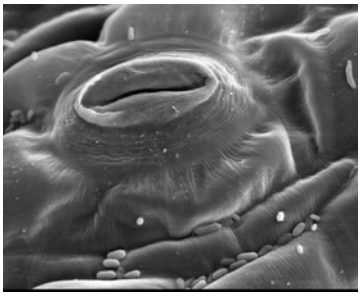
- Rice example: flowering time
  - Selected processes include "regulation of flowering"

- Tomato example: stomatal density
  - To be explored in case study

PLANT RESEARCH INTERNATIONAL
WAGENINGEN UR

Case study

- Trait: stomatal density
  - Gene function prediction
  - Connecting QTL data with predicted gene functions



- Questions #3 - #5
  - NB for Q3, use **http://www.applied-bioinformatics.nl/Q3bmrf.xlsx**
- Bonus questions: #6,#7

PLANT RESEARCH INTERNATIONAL
WAGENINGEN UR

Protein function prediction with BMRF

Sven Warris, Aalt-Jan Van Dijk

Many proteins in crop genomes have only tentative functions. In this tutorial, we will explore novel resources for protein function. In particular, we will focus on using such predicted gene functions for gene prioritization in Quantitative Trait Locus (QTL) studies.

Background

The de facto standard to capture function annotation today is the Gene Ontology (GO), in particular, the Molecular Function (MF) and Biological Process (BP) sub-ontologies [1]. MF describes activities, such as catalytic or binding activities, that occur at the molecular level, whereas BP describes a series of events accomplished by one or more ordered assemblies of molecular function. Compared to MF, terms in the BP ontology are generally associated with more conceptual and abstract levels of function. The prediction of BP terms can depend on the cellular and organismal context. Therefore, BP terms tend to be poorly predicted by methods based on sequence similarity only, such as BLAST [2].

We developed a method to predict biological processes based on sequence- and network-information [3-4]. This method, Bayesian Markov Random Field (BMRF), was applied to several crop species [5]. In the Critical Assessment of Function Annotation (CAFA) challenge, a worldwide comparison of gene function prediction methods, our method was among the best performing methods for several species [2].

In this tutorial, we will explore the use of our set of predicted protein functions for the prioritization of candidate genes in QTL regions. QTLs indicate genome regions influencing trait variation. These regions often contain tens to hundreds of genes, and we developed a bioinformatics approach to predict which of these genes are most likely to be causal genes for the trait-of-interest. In line with the focus of the training workshop, we will focus on tomato data. However, because the prioritization method was recently developed mainly using rice data, we will also use some rice data as examples.

Case study: stomatal density

As a biological case study, we will focus on the trait ‘stomatal density’ (Figure 1). Leaf stomata are the principal means of gas exchange in vascular plants. Stomata are small pores, that are opened or closed under the control of a pair of guard cells. When open, stomata allow CO_2 to enter the leaf for synthesis of glucose, and also allow for water, H_2O, and free oxygen, O_2, to escape. Plants may exert control over their gas exchange rates by varying stomata density in new leaves when they are produced. The higher the stomata density, the more CO_2 can be taken up, and the more water can be released.

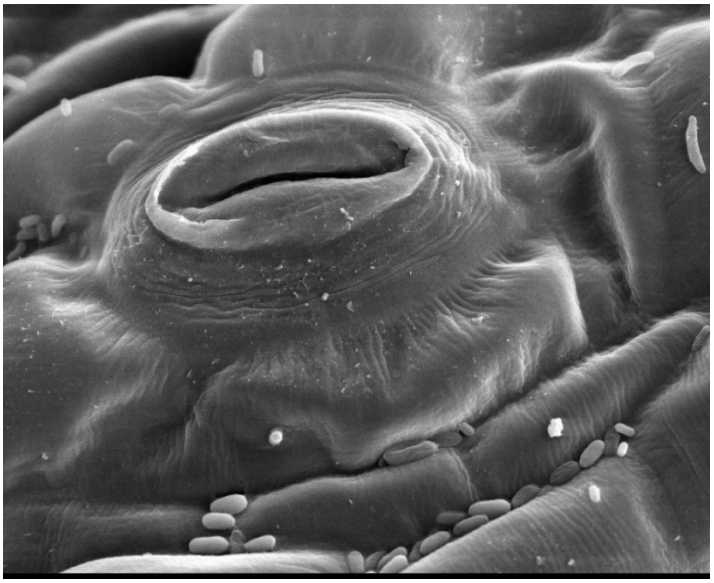


Figure 1. A stoma on the underside of a tomato plant at 300x magnification–it is closed or nearly closed. (http://www.suzymeyer.com/category/blog/)

Protein function prediction tool

As explained above, our method for function prediction uses sequence- and network-information. This method was applied to several crop species, using available co-expression networks. Resulting function predictions are available via http://www.ab.wur.nl/bmrf (Figure 2).

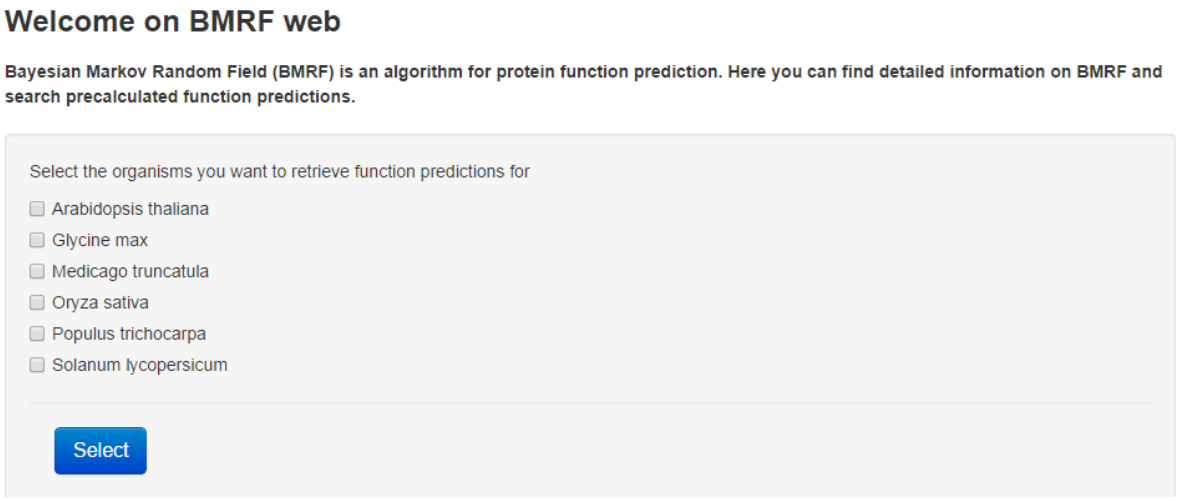


Figure 2. Screenshot of BMRF function prediction tool.

Question (1). Choose and Select rice ("oryza sativa") in the BMRF webtool (http://www.ab.wur.nl/bmrf), followed by searching with the keyword "stomatal" to obtain rice genes potentially involved in stomata functioning. For tomato, our function prediction has been applied so far to only a limited number of genes. Hence, searching for "stomatal" for tomato does not give much less results for tomato than for rice, currently.

Note that alternative existing annotations such as the ITAG annotation do not contain any annotation to such Gene Ontology terms; check this at http://solgenomics.net/organism/Solanum_lycopersicum/genome.

Have a look at the exact Gene Ontology terms that are obtained for rice and for tomato in the BMRF webtool– this can most easily be done by downloading the results in excel format. Do you observe the term "stomatal density" (the trait that we are interested in)? If not, what could be the reason?

Question (2). Recently, a tomato ERF transcription factor (Solyc10g006130) was found involved in stomatal density regulation in tomato [6]. Have a look at the predicted gene functions for this gene in the BMRF tool. Do you think that the highest-ranking term makes sense, given the role of stomata described above?

In addition, you might notice that the predictions for this gene do not include a term such as "transcription factor". Think about why such terms are not returned by the BMRF tool.

Integration of gene function prediction with QTL data

As mentioned above, QTL regions can contain many genes, and statistical analysis of predicted gene functions in those regions can help to prioritize candidate genes. Results from such prioritization are presented in the following webtool: http://dev1.ab.wurnet.nl:3335/ (Figure 3). This contains results on combining function predictions for rice with a large set of QTL data from the gramene database. We will not deal with these rice predictions in this tutorial, but in the Appendix an example of the rice data is presented. In addition, results on tomato QTLs from a set of introgression lines [7] are available. After some filtering this tomato set contains 23 traits; for ten of these there are currently results on prioritization of candidate genes in our webtool. These traits mainly describe properties of leaves. In this tutorial, as mentioned above, we will focus on the trait stomatal density, more in particular "Adaxial cotyledon absolute stomatal density".

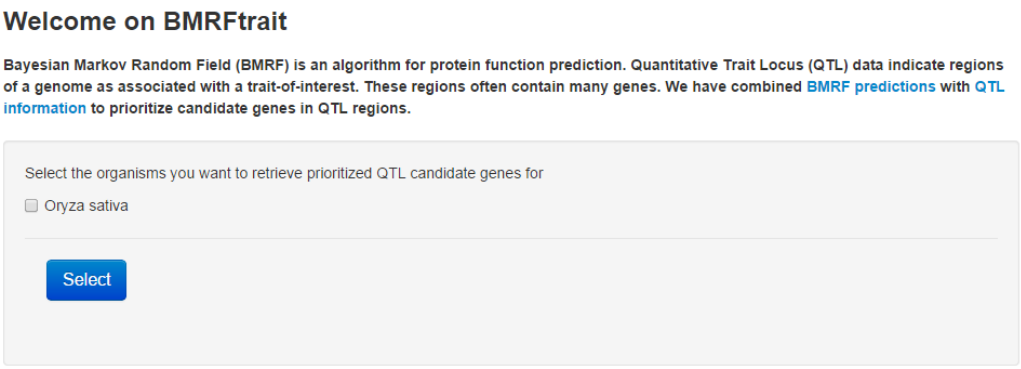


Figure 3. BMRFtrait (http://dev1.ab.wurnet.nl:3335/): integration of gene function prediction with QTL data in order to prioritize candidate genes.

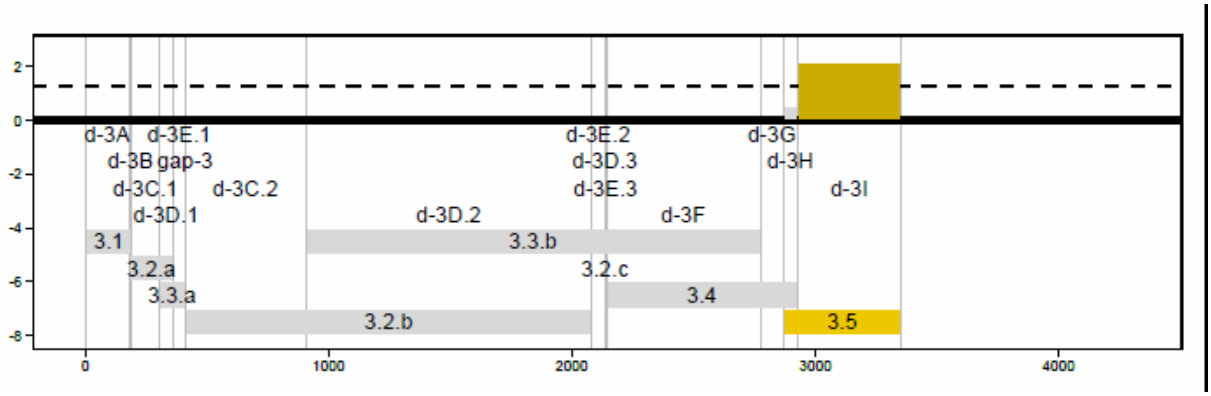


Figure 4. Part of the tomato introgression line (IL) mapping results for the trait stomatal density. Data are shown for chromosome 3. The bottom part of the graph is an IL map, the top part is a bin mapping graph; for both, the colored boxes indicate significant QTLs. The y-axis indicates -log10(p-value). For both the IL map and bin mapping results graph, the width of bins and ILs is proportional to the number of annotated genes that they harbor.

Question (3). Search for genes found as prioritized candidate genes for the trait "stomatal density" in tomato, using the BMRFtrait tool http://dev1.ab.wurnet.nl:3335/. You can use the abbreviation "CotStom" for this trait name, or search using the complete term "stomatal density". Subsequently, analyse which gene functions were predicted by BMRF for those genes, via http://www.ab.wur.nl/bmrf.

Do you find gene functions that look relevant for the trait "stomatal density"? One of the gene functions you should find is "response to wounding"; this is the gene function that was found statistically enriched in the QTL regions for this trait. Think about whether this gene function could be relevant indeed.

Question (4). The above mentioned ERF gene (Solyc10g006130) is not found back here as prioritized candidate gene for the trait "stomatal density". What could be the reason – does this indicate a failure of the prioritization approach, or could there be other reasons?

Question (5). Look back at the tomato genes you found in question #1 as potentially related to stomata. Why can most of these genes not be considered as causal candidate genes for the QTL region shown in Figure 4?

Appendix Example of rice QTL candidate gene prioritization for the trait 'days to heading'

To illustrate the value of our approach, we considered the rice trait 'days to heading' in depth. Days-to-heading, which is related to flowering time, is an important parameter for plant breeding and plays a key role in adaptation of rice to various environments.

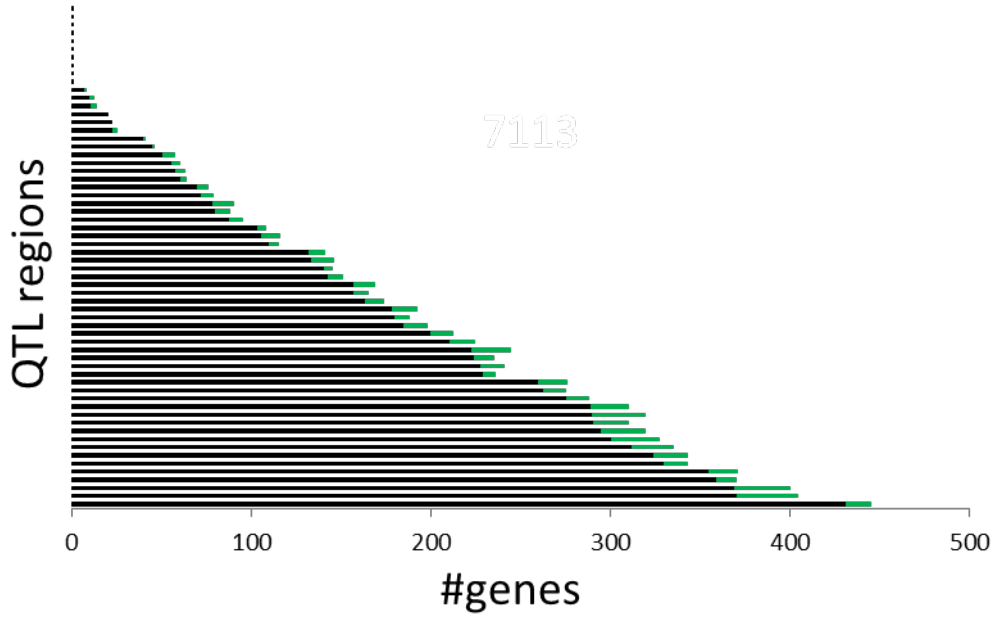


Figure 5. QTL candidate gene prioritization for rice days to heading. Number of genes (horizontal axis) present in various QTL regions (vertical axix) for this trait; green indicates sub-set of those genes that were selected as most likely causal candidate genes.

Figure 5 visualizes the number of genes prioritized, split up per QTL region. The most often occurring biological process predicted for this trait is 'regulation of multicellular organismal development'. This term, although quite general, is clearly relevant for the developmental trait days-to-heading. Two additional, relevant associated terms were 'cellular response to ethylene stimulus' and the obvious 'regulation of flower development'. We analyzed the genes associated with the term 'regulation of flower development' in somewhat more detail. Out of a total of 7113 genes in the rice QTL regions linked with the trait days-to-heading, 579 genes were prioritized. Of these, 79 genes were assigned to this BP term by our function annotation. Among these 79 genes some are so far only described as "unknown" by existing annotation. For example, gene LOC_Os04g54420 is annotated as containing a domain of unknown function (DUF618).

Question (6). Search for the gene LOC_Os04g54420 in the BMRF gene function prediction webtool (www.ab.wur.nl/bmrf). Depending on the settings, you will or will not obtain the exact term 'regulation of flower development'. Do you understand the reason?

To have a closer look at the prioritized genes for the trait 'days to heading' predicted to be involved in 'regulation of flower development', we focused on those genes that in the QTL region in which they occur were the only gene associated with this biological process. Given

the obvious relevance of the BP 'regulation of flower development' for the trait 'days to heading', the occurrence of a single gene annotated with that BP term in a QTL region for this trait makes that gene a prime candidate for further study. There are in total 11 of such genes. Some of these are indeed known to be involved in flower development. This includes two MADS genes, OsMADS34 (LOC_Os03g54170), involved in inflorescence and spikelet formation, and OsMADS18 (LOC_Os07g41370), involved in specifying floral determinacy and organ identity. Several other genes are however not characterized at all and should be considered new candidate genes involved in the regulation of flowering time. This includes a MYB protein (LOC_Os01g74020) and a GATA protein (LOC_Os10g40810).

Question (7). Do you observe a common molecular function for the proteins mentioned above? What could be the reason for that?

References

1. Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. Nat Genet 25: 25–29. doi:10.1038/75556

2.Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, et al. (2013) A large-scale evaluation of computational protein function prediction. Nat Methods 10: 221–227. doi:10.1038/nmeth.2340

3. Kourmpetis, Y.A., et al., Bayesian Markov Random Field analysis for protein function prediction based on network data. PLoS One, 2010. 5(2): p. e9293.
http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0009293

4. Kourmpetis, Y.A., et al., Genome-wide computational function prediction of Arabidopsis proteins by integration of multiple data sources. Plant Physiol, 2011. 155(1): p. 271-81.
http://www.plantphysiol.org/content/155/1/271

5. Bargsten et al., Current Plant Biology, *in press*
http://www.sciencedirect.com/science/article/pii/S2214662814000048

6. Kumar Upadhyay et al., J. Exp. Bot. (2013) 64 (11): 3237-3247
http://jxb.oxfordjournals.org/content/64/11/3237.full.pdf+html

7. Chitwood et al., Plant Cell (2013), 25, 2465-81
http://www.plantcell.org/content/25/7/2465.long