Research Project:  
 Plants use a large family of cell-surface receptors to sense environmental cues, most of which have no demonstrated biological function. One of the main aims of our lab is to assign function to these orphan receptors, using classical methods in addition to high-throughput screening and network analysis. Those receptors that have been extensively studied have taught us some general features of these sensing systems, including that the formation of signalling-competent units through physical interaction between multiple receptors is a critical step in signal transduction. As part of an international collaboration, we have recently completed a network analysis of the physical interactions between 200 receptor-family members. The proposed project will take that physical interaction data and combine it with gene expression data to build tissue-specific interaction networks. By looking only at those receptors that are expressed in a given plant tissue, we expect to find the interactions that are most important for detection of events specific to that tissue type. For example, the leaf-specific interaction network will assist us in identifying protein interactions required for foliar disease perception. This project will focus on constructing and analysing tissue-specific networks for a variety of plant tissues. Comparisons of these networks will allow for the identification of common network components, which we can imagine are the basal interactions required for plant survival. We will also identify the interactions and network components that are specific to each tissue, which we hypothesize will be key regulators of tissue-specific processes. The discoveries made in this project will then be layered with evolutionary data to further our understanding of how the receptor networks evolve to incorporate new receptors and functions. Depending on how quickly progress is made, this evolutionary study may also be undertaken.

Tools:

For the network building I have previously used the iGraph package as implemented in the programming language R. I think it would be easiest to continue using this package, and I think it is best if you become familiar with using Notebooks in RStudio as a good way to share code (https://bookdown.org/yihui/rmarkdown/notebook.html). If you have another preferred method though that is fine, just let me know. I expect you to keep well annotated code that will be left in a lab repository upon project completion. Probably best to include some kind of version control, maybe with Git (<http://r-bio.github.io/intro-git-rstudio/>) from the start. But again if you have another version control method you like that is fine. To get started there are a number of online tutorials (eg. <http://kateto.net/networks-r-igraph>) in addition to the package documents that can be found here: https://igraph.org/r/

Datasets:

The other thing that we need to consider from the start is the source of the RNAseq data that we will use to determine whether a given gene is expressed in a given tissue. This paper gives you the underlying data including the genes that we will include, and if you focus on the high quality interactions (567 bidirectional) that will be good to start for the network (<https://www.nature.com/articles/sdata201925>).

We will need to use existing data-sets to decide whether each of these genes is expressed in a given tissue. For the first pass I propose we try to construct a leaf expressed network and a root expressed network. You can start to think about how best to mine the existing resources, which should all be available here: <https://www.ncbi.nlm.nih.gov/geo/>

You will need to define a search for Arabidopsis RNAseq datasets. Ideally, we would want mature leaf tissue and whole root, though we may have to average across many different samples to achieve that and we won’t know until we start looking. Here is one possible set: <https://www.ncbi.nlm.nih.gov/pubmed/27840108>

So we will need to find any good sets (you can also check travadb.org but they don’t allow bulk downloads yet) and you should get comfortable with how to download and parse this data so we can isolate the RPKM for each gene tag. We will use RPKM or FPKM to determine expression, you can read background on RNAseq and RPKM or FPKM here: (<https://www.rna-seqblog.com/rpkm-fpkm-and-tpm-clearly-explained/> or <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4728800/> but just the quantification and differential gene expression sections). This will also give us some idea of the cut-off to be used to call a gene ‘expressed’ for our purposes (probably >1 RPKM I think).

Once we have the data and iGraph set up here is the underlying protein-protein interaction data: LINK RAW DATA. Then all you have to do is limit that data to only include the expressed genes and rebuild the network along with downstream analyses.

For general background on the LRR-RLKs you can read these, so you have some biological perspective. If we get beyond building these two networks then we can start looking for:

1. Other tissues
2. RNAseq data looking at what happens after pathogen infection or treatment with immune elicitors.
   1. Then layer that data onto the tissue specific networks to come up with a network over time analysis, which may be more challenging and also more rewarding
3. Start looking at the evolutionary pressures on these proteins and the network as a whole

The choice of route would be yours if we get there.

Obviously this is way more than I expect you to prepare before you arrive, but at least if I get this all down then you have a blueprint of the project in case you need some help and I am not immediately available.