**BBRO PROJECT REPORT FORM**

**Please note the details on page 2 will be used to formulate our Annual Report.**

|  |  |
| --- | --- |
| **Project Title:**  Discovering the source of sugar beet infection and re-infection by *Erysiphe betae* (amended to *Uromyces beticola*) | |
| **BBRO project no:** | **15/14** |
| **Project sponsor:** | **BBRO** |
| **Final report** | |
| **Project lead or student name:** | **Matt Clark** |
| **Project mentor or supervisors:** | **Mark Stevens** |
| **Report Date:** | **March 2017** |
| **Reporting period covered:**  **(e.g. 1/1/16 - 31/12/16)** | **01/07/15 – 31/03/17**  **(Sample collection began before project start date)** |
| **Timeline (e.g. Year 1 of 4)** | **Year 1 of 1** |
|  | |
| BBRO use only | Date assessed: |
| Assessors comments |  |
| Action required |  |

|  |  |
| --- | --- |
| **Project summary (no more than 300 words)** | |
| Wild plants may act as a reservoir for crop pathogens. We hypothesized that wild sea beet could harbour pathogens that attack sugar beet, if so wild and agricultural beets should share races of *Erysiphe betae* (mildew). However, *E. betae* infection was not present at sufficient levels and in consultation with BBRO we amended the project to use *Uromyces beticola* (rust) which has a more complex genome.  In July 2015 – December 2016 we sampled (~600) across Yorkshire, Lincolnshire, Nottinghamshire, Cambridgeshire, Norfolk, Suffolk and Essex. In 2016, we developed a new extraction and sequencing protocol designed to minimise the ratio of plant to pathogen DNA. We assembled a preliminary *U. beticola* genome and re-sequenced twenty isolates from both the wild and agricultural samples in 2016.  Results from preliminary sequencing highlighted that wild and agricultural *U. beticola* isolates share large parts of their genomes. However, these two populations are different. The regions of differentiation may be particularly important for agricultural pathogen success. Preliminary sequencing also highlighted numerous other microorganisms present on the leaves of wild and agricultural beets. Importantly these preliminary data gave us insight into how best to improve the extraction protocol to increase the amount of DNA present for sequencing.  Both our main objectives have been achieved. Our novel extraction and sequencing protocol “Peel&Seq” is being written up for publication. Our motivation underpinning this research was to identify the source of fungal foliar infection of the sugar beet crop. Data from preliminary sequencing shows us that wild hosts may share agricultural rust pathogens. We are now sequencing remaining isolates and aim to publish these analyses as well as use preliminary findings for a BBSRC fellowship application in May. | |
| **Main Objectives** | |
| 1. Quantify the diversity within UK *U. beticola* population 2. Identify whether wild and agricultural plants share races of *U. beticola*   Objectives are generated from:   1. Development of a DNA extraction protocol for *U. beticola* and *E. betae* 2. Construction of a draft genome of *U. beticola* 3. Re-sequence *U. beticola* isolates for population analyses | |
| **map-peel.pdf**  **Figure 1. Sample sites and DNA peel extraction.** | **Figure 2. Preliminary rust assembly shows regions of the genome with high similarity (close to zero) and other regions of higher divergence between wild and agricultural rust** |
| **Main outcomes and achievements** | |
| *E. betae* prevalence was very low in the 2015/16 sampling season. Therefore, we adapted the project for *U. beticola* which has a much larger and more complex genome. Due to increased costs of resequencing *U. beticola* samples extraction we used a reduced set of samples compared to the *E. betae* plan. Findings listed here are based on those preliminary data which confirmed our extraction protocols and our ability to sequence *U. beticola*. All isolates are now queued for sequencing and analyses.   1. New DNA extraction protocol applicable to many plant pathogens (Peel&Seq manuscript in prep). 2. *U. beticola* draft genome constructed to use with Peel&seq protocol and the proof of concept wild and agricultural pathogen study. 3. Preliminary data shows that rust diversity in the UK appears to be low. 4. Preliminary data shows differences between wild and agricultural beet rust pathogens is also low, suggesting that a single isolate could survive on both hosts. 5. Preliminary data shows that we cannot rule out the potential for rust on wild hosts to invade the agricultural crop. | |
| **Key messages for growers and industry** | |
| 1. Preliminary results show us that wild beets may harbour pathogens that are important to consider when trying to understand how crop pathogens evolve. 2. We are still not clear about how wild and agricultural pathogens interact with each other but they appear to be closely related to one another. 3. Of fundamental importance to continued reduction of pathogen levels on crops, is to better define the interaction between wild and agricultural pathogens. | |

|  |  |  |
| --- | --- | --- |
| **Section 1: To be completed by Project Lead:** | | |
| 1. Genome sequencing of *E. betae* and *U. beticola*; libraries prepared and queued for genome sequencing and DISCOVAR *de novo* genome assembly 2. 500 *E. betae* and *U. beticola* libraries prepared and queued for re-sequencing from which 40 will be selected for deep sequenced to address the wild and agricultural genetic diversity | | |
| **Milestones for current period** | | |
| **Note: mentors will be asked to comment on the status of this project (yellow column) using the scoring system shown below** | | |
| **Status - Mentor’s scoring system for interim reports.** | | |
| RED | “Major concern - escalate to the next level"  Slippage greater than 10% of remaining time or budget, or quality severely compromised. Corrective Action not in place, or not effective. Unlikely to deliver on time to budget or quality requirements. | |
| AMBER | "Minor concern – being actively managed”  Slippage less than 10% of remaining time or budget, or quality impact is minor. Remedial plan in place. | |
| GREEN | "Normal level of attention"  No material slippage. No additional attention needed | |
| Milestones | Comments + Any Action required | Status R/A/G |
| 1 | Complete |  |
| 2 | Complete |  |
| 3 | In progress (queued for sequencing) |  |
| 4 | Complete on preliminary data |  |
| 5 | Complete on preliminary data |  |
| 6 | Not yet complete and dependent on phase 2 QC (see below) |  |
| 7 | Divergence in preliminary is low and no races detected as yet |  |
| 8 | In progress (two papers: DNA extraction, Population genetics) |  |
|  |  |  |
|  |  |  |

|  |
| --- |
| **Summary of results (including figures and tables)**  ***Annual report****: please provide a 2 page summary of key findings from the reporting year.*  ***Final report:*** *please provide a summary of project findings and outcomes with relevant supporting data.* |
| We set out to sample, sequence and analyse fungal plant pathogens isolated directly from their host. This involves **sample collection** followed by **DNA extraction** followed by genome sequencing. From here we select a single isolate to be our reference genome and perform a **genome assembly** to which all other isolates are mapped and analysed using **population genetics algorithms**.  We proceeded in two phases. In the first phase, we implemented all elements of the proposed workflow, from DNA extraction, to population genetics on the preliminary data. The preliminary data were a set of 19 isolates we used to test the feasibility of the extraction protocol in data generation for population genetics. This two phase procedure allowed us to identify areas of improvement that could be implemented for the larger phase 2 experiment. Changed included a variation to our extraction protocol as well as adding a pre-screening step in which we identify the proportion of contaminant reads (introduced by other microorganisms) in each sample before they are deeply sequenced. Below we detail results chronologically with respect to phase 1 because this phase yields important preliminary biological results. We implement changes from phase 1 in phase 2 which we detail alongside phase 1 results.  **Sample collection**  Approximately 600 samples of both *E. betae* and *U. beticola* were collected between 2015 and 2016. These were isolated on both wild and agricultural hosts. The reduced prevalence of *E. betae* in the period of this project, was problematic, in discussions with the BBRO we switched to *U. beticola* as the focal species. The original estimated genome size of our focal species (*E. betae*) was ~120Mbp. Using preliminary data, we estimate the *U. beticola* genome to be ~400Mbp, moreover *U. beticola* is dikaryotic, thus we consider its genome as ~800Mbp for the purposes of sequencing. This switch in focal species meant that we had to adjust the number of samples in our population study downwards from 100 to 40. By implementing a pre-screen of 500 isolates these 40 should can be selected to provide the best “bang-for-buck” while covering UK diversity.  **DNA extraction**  DNA extraction is the critical first step for all of these analyses. DNA extractions from pathogens still growing on their host are hampered because they contain limited amounts of starting material, they also contain host DNA as well as the DNA of any other microorganism. Critically, we developed a peel extraction protocol (Fig. 1) and three assessment criteria to determine the quality of libraries in a pre-screen. These criteria were to assess the level of **-duplicate content**, **-host content**, and other **-microbial content** in the reads after sequencing. In out phase 1 pilot we sequenced 19 samples to determine the level of sequencing required per isolate.  ***-Duplicate content in reads after PCR amplification***  Genome sequencing and re-sequencing strategies typically require large amounts of starting material. Here, we must extract DNA from many tens of wild and agricultural samples and so we are not able to increase the quantity of DNA (*U. beticola* can’t be cultured contamination free). Therefore, we employed a modified low input method developed in house, this technique uses optimised PCR amplification of library molecules prior to sequencing. This allows us to directly peel the fungus from the surface of the leaf (minimising host DNA) and then to amplify the minute amounts of DNA (Fig. 1). A potential flaw with any PCR amplification library prep method is that if the template DNA is extremely low, the amplification process copies sequences that were generated in a previous round of amplification. Many duplicate representatives of the same original molecule cannot be used as independent evidence and must be filtered from the analysis. Thus we target a low level of duplicate content for our libraries.  ***-Host content in reads***  The peel extraction was developed to minimise the proportion of host to pathogen DNA. Previous DNA extraction methods involve taking a small punch of the leaf containing the pathogen. Our peel method involves painting a polymer solution onto the surface of the leaf, allowing it to dry, peeling and extracting DNA from the polymer peel. Here, a low ratio of host to pathogen read content is the target for our extraction protocol.  ***-Microbial content in reads***  As with any DNA extraction protocol from samples collected outside the lab there are biological contaminants. These contaminants are present in the form of bacteria, other fungi and oomycetes that also grow on the leaf. These are of lower consequence in genome re-sequencing projects because reads from other species don’t tend to map, but they use our sequencing budget. However, we also want to construct a genome to which we will map other samples. It is not possible to remove all biological contamination before DNA extraction, therefore, we determined the level of biological contamination in 19 isolates prior to selection of a single isolate for genome sequencing.    Figure 1. The polymer peel removes fungi and associated microbes from the surface of the leaf with minimal leaf contamination. Peels are then physically disrupted and DNA extracted using the CTAB protocol (ref).  **Extraction efficacy results**  Read duplicate content in our phase 1 extractions was high (10 – 85%). We determined that this was due to contaminants inhbiting enzyme activity during library preparation. After modification to the protocol (manuscript in prep) duplicates were reduced to 5-7% (Fig. 2).    Figure 2. In phase 1 DNA extractions we improved our extraction protocol to reduce duplicate content.  Biological contamination from the plant was largely removed using the peel method. However, here those data reduce our ability to observe pure *U. beticola* material. We used a program called Megan to quantify biological contamination in our sample (Fig. 3). Megan uses online databases to identify the numbers of reads that are associated with particular genomes. In that database *Phakopsora pachyrhizi* and *Puccinia graminis* are the closest basidiomycetes from within the pucciniales and we use the presence of reads mapping to those genomes as evidence for the presence of *U. beticola*.      Figure 3. Megan plot of the reads from two *U. beticola* isolates. The top plot is from an isolate with very little biological contamination and the bottom plot is from another isolate with much more biological contamination.  It is not possible to further remove all biological contamination in the extraction protocol therefore we used phase 1 sample data to identify and bioinformatically filter contamination content from the sequence read data. For phase 2 we implemented a pre-screen in which we use low level sequence 500 isolates on a single Illumina HiSeq2500 lane in order to identify those isolates to be used for genome re-sequencing. It is important to state that very little is known about the host plant microbiome so this too is an important output from these experiments. Our 500-sample pre-screen will reveal important insights on the association of microbes in the wild and agricultural settings. It is hoped that we may also use these data to drive future research questions.  In re-sequencing experiments, data is lost to biological contamination because it is not mapped to the genome of your focal specis (i.e. *U. beticola*) and so its presence is of limited consequence for genome re-sequencing. The presence of this biological contamination has a larger impact on *de novo* genome assembly. Below we detail the filtering steps implemented in our phase 1 **Genome Assembly**, but the presence of this contamination meant that we changed our assembly protocol for phase 2. Instead of using the peel method to extract DNA for our assembly, we cultured our isolate and collected spores to provide a purer sample for our phase 2 genome. Our peel extraction method provided us with 0.1 – 30 ng DNA. Our spore collection method provided us with 640 ng DNA (still a low amount by most standards) and allowed us to use the DISCOVAR genome sequencing and assembly protocol.  **Genome Assembly (phase 1)**  It is clear from the sequencing results of 19 *U. beticola* isolates that a genome generated from a leaf peel would require a lot of read filtering to remove other biological contaminants. As with any filtering protocol, filtering also removes some content from the focal species. For the purposes of exploring the population genetic signals we generated a heavily filtered phase 1 genome to test the levels of diversity in the 19 isolates we have sequenced so far.  ***-Read filtering and phase 1 genome assembly***  Megan was used to identify those isolates with the lowest level of other biological contaminants (Fig. 3). The two isolates with the greatest proportion of reads that map to a *U. beticola* relative were taken through to the next stage of filtering. We identified the top ten most prevalent contaminant species (Table 1). We mapped our reads to the genomes of the contaminants and retained all those reads that did not map for the next stage of genome assembly. Kmers are another way to look at read content. Here, reads are 150 bp long and we used a sliding window of 31bp to pull out all the windows of base pairs (kmers) from both samples. We identified those kmers that were present in both samples (i.e. those that are from the focal species) and retained all reads that were made up 65% of these kmers. At this stage then we have removed common contaminant species and retained all reads with 65% identity. We next used Abyss to assemble the reads from one sample into a genome.  Table 1. Top ten species found to have reads represented in 19 leaf peel DNA extractions   |  |  | | --- | --- | | **Species name** | **Common name / disambiguation** | | *Beta vulgaris subsp. vulgaris* | Beet host | | *Homo sapiens* | Human | | *Pseudomonas syringae* | g-proteobacteria | | *Pseudomonas fluorescens* | g-proteobacteria | | *Pseudomonas rhizosphaerae* | g-proteobacteria | | *Sphingomonas taxi* | a-proteobacteria | | *Clavibacter michiganensis* subsp. *sepedonicus* | high GC Gram+ | | *Curtobacterium* sp. | high GC Gram+ | | *Hymenobacter sp.* | CFB group bacteria | | *Kineococcus radiotolerans* | high GC Gram+ |   The level of filtering produced a highly-fragmented phase 1 assembly which we filtered further to remove unitigs (highly contiguous sequences of assembled reads) that were shorter than 1Kbp. We than blasted the remaining unitigs and retained those with an e-value hit to a basidiomycete that was less than 1x10-5 (Fig. 4). The resulting content was 38Mbp of short (N50 = 2.2Kbp) unitigs which represents fragments covering around 10% of the genome. This filtering pipeline is stringent with much of the *U. beticola* content being removed, however, this increases our confidence that results based on this portion of the genome represent the pattern observed in *U. beticola*.  Analysis of core eukaryotic genes (CEGs), present in all eukaryotes, shows that 35.5% of these genes are confirmed as complete, and 67% are partially present in our phase 1 genome. This also suggests that while we have ~10% of the genome, we have captured 40-60% of the genes – due to under representation of repeats (a well-known assembly artefact). We use this phase 1 genome to begin to understand the signal of genetic diversity using population genetic which does not rely on a highly contiguous genome.  ../../../Library/Application%20Support/com.evernote.Evernote/accounts/www.evernote.com/64352486/external-edits/163FE65E-17F0-4197-9D84-69E4A42B34FC/60D22091-C1C2-4197-8857-785145F7B37A.pdf  Figure 4. An example of the representation unitigs to the blast database for our phase 1 genome. After read filtering 96% of the unitigs have their closest match to a pucciniales species.  ***-Population genetics***  The overarching goal of this project is to understand the evolution of pathogens towards novel forms of host resistance and fungicidal treatment. We test for evidence of differentiation between pathogens that live on both wild and agricultural plants. We used the data from 19 isolates (10 agricultural, 9 wild) sequenced to a low level to get an assessment of genetic diversity in wild and agricultural populations. After assembling our marker quality genome, we map the reads from our 19 isolates to that genome and then measure the genetic differentiation between those groups of isolates.  We found 336,149 SNPs present in 19 isolates mapped to 38Mbp of the genome. *FST* is a number that describes the difference between two populations while also accounting for the amount of genetic diversity within them. Overall *FST* is very low, close to zero between the wild and agricultural populations (Fig. 5). This means that a large proportion of the diversity present in these populations is shared and further investigation could reveal that they are part of the same race. If we look at the pattern of *FST* on each of the scaffolds of our marker genome we see that most of them are distributed around zero with a few having higher values (Fig. 5). This means that wild and agricultural pathogens share a lot of polymorphism between hosts. Importantly though, there are regions of the genome of differentiation and these regions could be important for specificity to either crop or wild host. Without a gene annotation, it is too early to comment on specific genes this at this stage. One observation that is consistent with our expectation is that genes are shared across the wild – agricultural boundary and scaffolds which contain essential CEGs have the lowest levels of genetic differentiation. Phase 2 sequencing and genome assembly is underway and will provide better genomes, and more statistical power to detect pathogen migration. However, our understanding so far is that genetic information is transferred between pathogens living on wild and agricultural hosts which has implications for the future of any crop protection strategy.  **../../../Library/Application%20Support/com.evernote.Evernote/accounts/www.evernote.com/64352486/external-edits/9329D2F3-0354-4403-A0C5-5C372CA47A01/1fst.pdf**  Figure 5. The level of genetic differentiation (*FST*) between wild and agricultural pathogens is low. Low *FST* values indicate that much of the genome content is shared between wild and agricultural pathogens (left panel). However, The *FST* values of parts of the genome show considerable divergence (right panel). *FST* error generated using jack-knife replication. |
| **Annual report: Key issues to be addressed next year:** |
| This project was designed as a pump priming project that would address preliminary questions and produce data for ongoing analyses for future endeavours. Here, we identify an interesting genetic relationship between wild and agricultural pathogens. The data generated here will allow us to continue to explore these phenomenon in greater detail. Moreover, it will be the basis for future projects submitted to the research councils. We would like to continue to work closely with the BBRO to develop our ideas towards industry needs. Therefore, over the coming year we will publish our peel DNA extraction protocol. Using data presently queued for sequencing we will generate a more reliable genome (both *E. betae* & *U. beticola*) and re-map the reads from 40 high quality re-sequenced *U. beiicola* isolates in our phase 2 protocol. We will use preliminary data for fellowship applications (BBSRC and Royal Society) as well as in an application for a joint EI BBRO PhD studentship under the BBSRC iCASE scheme.  One of the future steps will be understanding how crop pathogens adapt to the treatments applied against them. |
| **Publication of results to date/planned publications**: |
| DNA extractions from pathogens that cannot be grown in isolation of their host will contain both pathogen and host material. Mixing host and pathogen DNA in an extraction makes sequencing infeasible for many crop pathogen systems. Our work on a DNA extraction protocol that can be performed on obligate biotrophic pathogens for genome re-sequencing will be a welcome addition to the plant pathogen community. We are currently sequencing ~500 isolates to a low level in order to determine the efficacy of the approach which we aim to submit in 2017.  In preliminary (phase 1) data we observe that the level of genetic diversity in UK *U. beticola* is low and that there is a large amount of sharing of genetic variation between wild and agricultural populations. This is an important discovery for sugar beet but also more broadly for the plant pathogen community. These findings are part of ongoing analyses and already form the bases of a fellowship proposal to be submitted in May and a PhD proposal (submitted). We could publish our high-quality isolates and DISCOVAR genome around 2017/2018. However, these analyses will be much stronger with a genome annotation. This will allow us to identify those genes that are shared between wild and agricultural populations and, perhaps more importantly, those genes that are not shared, perhaps because they are detrimental to the success of the pathogen in agriculture. |
| **Section 2: To be completed by project mentor** |
| **Is the project on track to meet the stated objectives? (please comment in relation to milestones and the status score awarded in section 1).** |
|  |
| **Please comment on any proposed changes to milestones.** |
|  |
| **Are conclusions scientifically robust? (please comment on data analysis/interpretation)** |
|  |
| **For final reports only:** |
| **How would you rate the project against the following criteria (please give a score out of 10, with 10 being highest)**  1 ) The project met its original objectives:  2) Contribution to scientific knowledge:  3) Direct relevance to growers: |