Phytoplasma Genome Sequencing Initiative (PGSI) Annotation School II, The John Innes Centre (**JIC**), Norwich, UK, April 29 – May 03, 2013.

Background information:

Saskia Hogenhout (**SH**, JIC, Norwich UK), Xavier Foissac (**XF**, INRA Bordeaux, France) and Michael Kube (**MK**, Humboldt University, Berlin, Germany) of working group 4 (WG4) organized the Phytoplasma Genome Sequencing Initiative (**PGSI**) with the goal to disseminate knowledge about how to carry out annotation of phytoplasma genome sequences.

PGSI was organized as follows:

PGSI participants submitted (CsCl-enriched) phytoplasma DNA samples to The Genome Analysis Centre (**TGAC**, Norwich, UK). TGAC conducted quality controls of submitted DNA samples, generated the libraries and conducted the sequencing. TGAC transferred the sequence reads to MK, who separated the phytoplasma reads from those of the (contaminating) plant/insect host and assembled the phytoplasma reads into contigs. Michael transferred the assembled contigs to XF, who uploaded the contigs into the iANT annotation database in collaboration with Sebastien Carrere (INRA Toulouse, France).

At the time of the school, TGAC sequenced 18 phytoplasma samples in three rounds of 6 samples and commenced library construction of a 4th round of 6 samples. MK assembled the reads of the 18 samples into contigs. Of these, contigs of 14 phytoplasma samples were uploaded into iANT for annotation. In addition, SH provided Maize bushy stunt phytoplasma genome sequence data for uploading into iANT. Thus, the PGSI school participants had access to sequence data of 15 phytoplasma genomes.

The goal of the school was to teach PGSI participants how to use iANT for annotation, so that the participants can annotate their own phytoplasma genomes using iANT.

School minutes:

Monday, April 29, 2013

Participants (13 trainees, 4 trainers, I guest speaker (Kirsten McLay, TGAC) and 5 other participants of the Hogenhout lab) arrived at the John Innes Centre at ca 4 PM. Zigmunds Orlovskis and Roland Wouters (SH lab) took the participants on a ca. 45-min tour of the JIC campus. The annotation school started at 5.30 PM at the Chris Lamb Training facility. The lecture theatre has been designed for bioinformatics workshops. Each participant had his/her own computer and was able to follow instructions of lecturers on the big computer screen.

At 5.30 PM SH opened the PGSI school with an introduction and an overview of the school program. She also informed the audience about the PGSI objectives and progress (see background information). It was agreed which genomes will be annotated during the school.

At 6.00 PM reception with snacks and drinks started.

Starting at 7.30 PM, all PGSI school attendees had dinner in The Mad Moose restaurant in Norwich

Tuesday, April 30, 2013

PGSI school participants arrived at the Chris Lamb training suite at 9 AM.

At 9.30 AM Kirsten McLay (TGAC, Norwich, UK) presented a lecture on library construction and sequencing methods. She also provided an overview of library and sequencing results of the 18 phytoplasmas samples and progress made with the 4th round of 6 samples.

At 10.00 AM MK presented a lecture on sequence assembly methods and parameters. He showed the assembly results of 18 phytoplasma genomes.

At 11.00 AM XF showed the status of the 15 phytoplasma genomes that are available for annotation in iANT, and presented a lecture on theoretical basis of annotation.

After lunch, we had breakout sessions with small groups to discuss interesting aspects of phytoplasmas genomes and how we may proceed to publish genome sequences. Opinions of each group were discussed with all participants.

All PGSI school attendees had dinner at Thai Lanna restaurant in Norwich. Matt Dickinson arrived in Norwich and joined for dinner.

Wednesday, May 1, 2013

PGSI school participants arrived at the Chris Lamb training suite at 9 AM.

From 9.15 – 11.30 AM, XF showed the iANT2.0 and Narcisse annotation tools, and conducted a real-time annotation of a CDS.

At 10.45 AM, Matt Dickinson presented a lecture on LAMP: “Development of rapid in-field loop mediated isothermal amplification (LAMP) assays for phytoplasmas”. The rationale for this seminar was to learn more about phytoplasma diagnostics and how the phytoplasma genome sequences may help towards improving phytoplasma detection specificity.

The PGSI school participants got together for a group photo at 12.15 PM. All participants received a copy of the photo.

After lunch break, the PGSI participants were divided into 4 annotation consortia. Consortium 1 was in charge of annotating the genomes of the 16SrI group (MBSP and CYP), consortium 2 the Napier grass phytoplasma genome, consortium 3 the coconut lethal yellows phytoplasma genome, and consortium 4 the *Ca*. Phytoplasma pyri and *Ca*. Phytoplasma pronorum genomes.

At 7 PM the pizza-baking event started in the Chris Lamb Training suite Blue Skies room and garden of the John Innes Centre

Thursday, May 2, 2013

PGSI school participants arrived at the Chris Lamb training suite at 9 AM.

The annotation jamboree continued at 9.15 AM and took until 5 PM. Jamboree was intercepted with 20-min coffee/tea breaks at 10.30 AM and 3.30 PM and a one-hour lunch at 12.30 PM.

AT 5 PM, we discussed how to proceed with the genome annotations and reached the agreement that we will finish the annotations of genomes by the end of May 2013, 2013. We also agreed on a list of phytoplasma genomes that will be taken forward for comparative genome analyses. As well, some of the genomes will be further sequenced using mate-pair and/or PacBio sequencing. This will help to close sequence and physical gaps and may lead to completion of some of the genome projects. For comparative genome analyses, the phytoplasma genome sequences will be uploaded into Molligen. We will assess which genes are conserved among members of the Class Mollicutes, identify genes that are uniquely present in phytoplasmas among the mollicutes and identify genes that are unique to some of the phytoplasma subgroups. The results of these studies will be written up for publication.

At 7 PM, we had dinner at the Spice Lounge (Indian restaurant), Norwich.

Friday, May 3, 2013

PGSI school participants arrived at the Chris Lamb training suite at 9 AM.

The annotation jamboree continued at 9.15 AM.

Between 10.30 AM and noon, most participants left to catch their trains and flights.

Abstracts

Phytoplasma Genome Sequencing Initiative (PGSI) school II objectives and progress

Saskia A. Hogenhout

The John Innes Centre, Norwich, NR4 7UH, United Kingdom

The goal of PGSI is to work with the EU-COST FA0807 phytoplasma community to sequence more phytoplasma genomes at about 2 Gb of 100-bp reads per phytoplasma DNA sample. The call for phytoplasma samples was sent out to all FA0807 members on July 2, 2011. Approximately 20 laboratories responded and submitted phytoplasma DNA samples to the Hogenhout lab at the John Innes Centre. These were checked for quality and quantity and then forwarded to The Genome Analysis Centre (TGAC) for a second quality assessment and library construction and sequencing. At the time of the PGSI school II, TGAC sequenced 18 samples in three rounds of 6 and constructed libraries for 6 samples for a 4th round of 6. Contigs of 14 phytoplasma samples were uploaded into iANT for annotation. In addition, SH provided Maize bushy stunt phytoplasma genome sequence data for uploading into iANT. Of these, 14 samples were assembled into contigs and four phytoplasma genomes were uploaded for annotation into iANT. In addition to these, the two contigs of the Maize bushy stunt phytoplasma genome was uploaded into iANT. Thus, the PGSI school participants had access to sequence data of 15 phytoplasma genomes. The goals of the school are: (i) Learn how to annotate and use iANT for phytoplasma genome annotation; (ii) Understand annotation rules; (iii) Ensure annotations are consistent and harmonized; (iv) Agree on revisions of existing annotations as necessary; (v) Learn strategies for comparing phytoplasma and other mollicute genomes; (vi) Discuss the next phase of PGSI; and (vii) Decide on topics for publications.

Phytoplasma sequence data processing and selection

Michael Kube

Humboldt University of Berlin, 14195 Berlin, Germany

Shotgun sequencing of two batches of six samples each was performed by TGAC sequencing unit (www.tgac.ac.uk). Paired–end reads with a read-length of 100 bases were generated by the sequencing by synthesis approach (illumina).
Initial metagenome analysis is based on incorporating all reads without preselection. Read numbers ranging from of 2 x 8 to 2 x 23 million reads have to be handled for *de novo* assemblies of the first batch. Second batch data shows increased read numbers up to 2 x 90 million reads/sample. The challenge of assembling these high read numbers was considered for present and on-going analysis. Advantages and disadvantages of different assemblers were highlighted.

First batch of data was assembled via CLC genomics workbench (http://www.clcbio.com) after an initial quality trimming. Analysis was performed using standard parameters with a few exceptions. The similarity index for matches during the assembly was increased from 0.8 to 0.9 and the paired-end information was used for guidance.
For taxonomical assignment contigs were extracted and compared via BLASTX against NCBI's NRPROT database (www.ncbi.nlm.nih.gov). BLASTX results were analysed via MEGAN (Huson et al., 2007) taking into account the usage of contigs instead of read data. Contigs were selected in MEGAN assigned to the taxonomical levels Tenericutes, *Acholeplasmataceae* or '*Candidatus* Phytoplasma' depending on the enrichment and input material. Selected contigs were provided for processing via iANT and annotation training.

Theoretical bases for bacterial genome annotation

Xavier Foissac
INRA, Villenave d'Ornon, F33140, France

Genome sequence annotation consists in giving a meaning to a nucleotide sequence upon two annotation steps: (1) syntaxic annotation, which consists in describing structural elements such as structural RNAs or coding sequences (CDS), (2) functional annotation upon which a a function is assign to the structural element. Ribosomal RNAs are usually predicted using using sofwares trained on a rRNA database (Ex: RNAmmer) and tRNAs can be detected by probabilistic models (Ex tRNA-ScanSE). CDS predictions are achieved by programs that build a model based on Hidden Markov Chains after being trained on a dataset of known CDS (Ex: Genemark, Glimmer or FrameD). The predicted protein sequences are compared to protein databases to look for similarity. A function can be given to the CDS if an ortholog with known function is found in the database, with a minimum identity of 20-30% over at least 80 % of the proteins length. Additional search for functional protein domains (interPro hit), transmembrane domains (TMHMM) can help further in the annotation. In some complex cases, phylogenetic analysis of the protein family must be performed in order to ascertain protein function. To avoid imprecise or false descriptions Gene Onthology provides a controlled vocabulary of terms (www.geneontology.org).

iANT 2.0 an integrated platform for genome semi-automatic annotation

Xavier Foissac
INRA, Villenave d'Ornon, F33140, France

The annotation platform iANT was initially developed for the annotation of *Ralstonia solanacearum* and *Sinorhizobium melliloti* genomes by LIPM at INRA, Toulouse. It implements automatic prediction of structural RNAs and coding sequences with frameD (CDS). Predicted proteins are compared to Uniprot and reference genomes and searched for protein domains (interPro) and transmembrane segments. Protein domains hits and similarity alignments are presented in a CDS comprehensive page with several links to Uniprot and interPro or other iANT annotation platforms. Expert annotation consist in the attribution of a description for the proteic product, a gene name if applicable, a functional class, an EC number for metabolic activities and Gene Ontology terms. All annotations can be retrieved after searching the annotation fields and sequences can be compared to the iANT database of the annotated genome.

Development of rapid in-field loop mediated isothermal amplification (LAMP) assays for

phytoplasmas

M. Dickinson
School of Biosciences, University of Nottingham, Sutton Bonington Campus, Loughborough, LE12 5RD, UK
Email: matthew.dickinson@nottingham.ac.uk

Abstract
Loop-mediated isothermal amplification (LAMP) is an isothermal amplification technique that can be undertaken with minimal equipment to obtain amplification of target DNA within 30 minutes. We have developed primers for a range of assays for specific 16Sr phytoplasma groups. These assays have been combined with a real-time isothermal amplification system and the OptiGene GenieII portable lightweight detection machine to develop a rapid in-field diagnostic test for plant diseases. When combined with a 2-minute DNA extraction method from plant material, including leaves and coconut trunk borings, the method can be used to detect pathogens in the field within 30 minutes of sampling. Use of the system in remote locations has been piloted for detection of the Cape St Paul wilt phytoplasma disease of coconuts in Ghana. At the workshop we will discuss and demonstrate the use of real-time LAMP for detection of phytoplasmas in plants, along with the scope for designing new primers based on the results from the on-going phytoplasma genome annotation workshop.

Guest lecture:

A guest lecture was presented by Dr. Kirsten McLay (The Genome Analysis Centre (TGAC), Norwich, NR4 7UH, UK)

Dr. McLay presented an overview of TGAC goals and projects. TGAC handles the library construction, sequencing operations and bioinformatics of various projects. Library construction includes quality control of the incoming samples, construction of different library types and library profiling. A good quality DNA sample has to be of good quality (RNA free, high molecular weight and free of contaminants such as ethanol, phenol and proteins) and quantity (appropriate quantification essential; Qubit 2.0 Fluorometer is preferred above NanoDrop ND-100). TGAC constructs Illumina, 454 and SOLiD libraries. Sequencing equipment at TGAC includes: three types of Illumina sequencers, including the HiSeq 2000, and Life Technologies, Roche 454 and PacBio RS sequencers. Overviews of the Illumina, Roche 454 and Pacific Biosciences technologies are available in a recent review (Nature Reviews Genetics 11, 31-46, January 2010). Illumina generates paired-end reads about 100 bp of in total 160 Gb in approximately 12 days and mate-pair reads of >3-kb fragments, Roche 454 generates read lengths of up to 400 bp of in total 400 Mb in about 10 hours, SOLiD generates reads of 75 bp in lengths of in total 200 Gb in 10 days, and finally PacBio RS generates read lengths of 2 – 10 kb of in total 40 Mb in about one hour.

TGAC processed three sets of 6 phytoplasma DNA samples at the time of the PGSI school. They successfully constructed 18 libraries. For library construction, the phytoplasma DNA was sheared at 300 bp with Covaris S2 instrument, the DNA was blunt-ended, A-tailed and ligated to adapters. The ligated fragments were size selected at about 450 bp. Quality analysis of the libraries revealed insert sizes of in between 439 bp to 719 bp. Each set of 6 samples was pooled in one lane of the Illumina HiSeq for 100-bp paired-end sequencing. Between 8.1 million to 50.4 million reads per sample were generated.

There may be several strategies for sequencing the phytoplasma genomes to completion. These are Roche 454 shotgun and Roche shotgun + Illumina Mate pair sequencing of 2-5 kb

fragments. For the latter, TGAC has developed protocols for 2-5 kb Illumina Mate pair library construction and for multiplexing mate-pair libraries for Illumina sequencing. TGAC has validated protocols for 2-5 kb Illumina Mate pair library construction. This method requires very high quality high molecular weight DNA and minimally 30 µg of DNA. Other strategies are a combination of Illumina (500-600 bp fragments) and PacBio (2 – 10 kb fragments) and a combination of various Illumina Paired and insert libraries (about 180-kb inserts, 500-600 bp inserts and 1 kb inserts). Newer technologies that may be used are 2 x 150bp/2 x 250 bp Illumina chemistries and the Oxford Nanopore (AGBT).

Acknowledgements: Sophie Janacek and Chris Watkins (Project Management), Melanie Febrer (Library Construction), Kirsten McLay (Sequencing Operations) and Nizar Drou (Bioinformatics).

Photos:

Group photo:


KEEP OUT