Sequencing the Genome of the Queensland Fruit Fly

Stuart Gilchrist
University of New South Wales (UNSW)

Project Number: CT10033

CT10033

This report is published by Horticulture Australia Ltd to pass on information concerning horticultural research and development undertaken for the citrus industry.

The research contained in this report was funded by Horticulture Australia Ltd with the financial support of the citrus industry.

All expressions of opinion are not to be regarded as expressing the opinion of Horticulture Australia Ltd or any authority of the Australian Government.

The Company and the Australian Government accept no responsibility for any of the opinions or the accuracy of the information contained in this report and readers should rely upon their own enquiries in making decisions concerning their own interests.

ISBN 0 7341 2750 2

Published and distributed by:
Horticulture Australia Ltd
Level 7
179 Elizabeth Street
Sydney NSW 2000
Telephone: (02) 8295 2300
Fax: (02) 8295 2399

Horticulture Australia

HORTICULTURE AUSTRALIA LTD

FINAL REPORT

PROJECT CT10033

Sequencing the Genome of the Queensland Fruit Fly

Dr. A. Stuart Gilchrist
School of Biological Earth and Environmental Sciences
University of New South Wales Sydney NSW 2052

UNSW
THE UNIVERSITY OF NEW SOUTH WALES

Know-how for Horticulture™

Horticulture Australia Project Number CT10033

Program/Project Leader

Dr A. Stuart Gilchrist

Principle Investigator

Dr. A. Stuart Gilchrist,

Address: School of Biological Earth and Environmental Sciences, University of New South Wales, Sydney NSW 2052, Australia

E-mail: a.gilchrist@unsw.edu.au

Telephone: +61 2 9385 8294

Mob: +0428 307 493

Collaborative Investigators

Dr. John Sved

Dr. Marianne Frommer

Professor Marc Wilkins

Dr. Nandan Despande

Dr. Deborah Shearman

Purpose of Report

This Report details progress on the construction of the complete genomic DNA sequence of the Queensland fruit fly, *Bactrocera tryoni*.

Funding:

The funding was provided by the citrus industry levy and matching funds from the Australian Government, facilitated through HAL.

Date of Report: November 2011

Media summary

Queensland fruit fly (Q-fly) is the most widespread pest of horticulture in Australia. To control Q-fly, industry uses essentially the same pheromone lures and spray techniques that were used in the 1960s. We need new research tools if we are to develop new and effective biosecurity tests, better lures and targeted chemical control.

To propel Q-fly research into the 21st Century, Horticulture Australia Ltd. provided start-up funding for researchers at the University of NSW to start sequencing the complete genome of the Q-fly. Using state-of-the-art sequencing technology, a first assembly of the genome has already been produced. That assembly of over 400 million base pairs of DNA has been shown to be largely complete.

At present, that genome assembly is being refined and further completed. Researchers are currently working on cataloguing all the Q-fly genes and are investigating distinctive features of the Q-fly genome. Once this work is complete in 2012, the genome will be made publicly available on the Internet. Researchers Australia-wide will be able to look for genetic regions involved in any aspect of Q-fly biology including, for example, lure responses. However, the first outcome is likely to be vastly improved biosecurity tests that can rapidly identify different local and exotic species at the larval stage.

Technical Summary

There have been few fundamental advances in Queensland fruit fly (Q-fly), *Bactrocera tryoni*, research in the past few decades. In order to produce new practical products and to engender new research opportunities, this project aimed to produce a partial sequence of the Q-fly genome.

The project exceeded that aim. We have successfully produced a first assembly of the genome using three paired-end and two mate-pair sets of Illumina next-generation sequencing data. Over 280 million reads were assembled using ABySS, SOAPdenovo and SSPACE software on the Intersect supercomputing cluster. A set of scaffolds with an N50 of ~7kb was produced and surveyed for two sets of core eukaryotic genes. The first were the cytoplasmic ribosomal protein set and the second a reduced set of core eukaryotic orthogonal genes. Both sets were at least 92% present with little evidence of over-representation. The assembly was also screened for repetitive DNA elements. Those findings were consistent with the expected classes of elements found in other Dipteran flies, being dominated by *Mariner* elements. The assembly therefore represents a largely complete Q-fly genome.

To build on the outcomes of this project, the assembly reported here is currently being improved by transcriptome sequencing, further mate pair sequencing, characterization of repetitive elements and annotation. The resulting genome will be made publicly available in 2012 via a web-based genome browser hosted at the University of NSW.

Introduction

The research background

Queensland fruit fly (*Bactrocera tryoni;* Q-fly) is the major horticultural pest in eastern Australia infesting 82 host fruits (Anon. 1996) including most commercially grown horticultural crops. The species originated from the rainforests of tropical northeastern Australia. However, the current distribution of Q-fly extends from northern Queensland to Victoria in a broad coastal zone and Q-fly also occurs in the Northern Territory and northern Western Australia. Accordingly, it is a widespread concern to Australian horticulture.

Q-fly is both a direct problem because of the damage it causes to pre-market produce, and an indirect problem due to its effects on trade. Both inter-State and international trade is hampered by quarantine requirements designed to prevent the introduction or spread of Q-fly.

In addition to this local problem, Australian horticulture also faces the threat of invasion by exotic species of fruit fly, which are present throughout much of the Pacific and southeast Asia. We do not need to speculate on the costs associated with this threat: the eradication of papaya fruit fly from Queensland in 1997 cost $37 million.

Given that it is a significant pest, recent decades have seen very few technological advances in Q-fly control. The protein bait sprays used are the same as those used since the 1950s and 1960s. Cuelure is still the only effective male lure and has been in routine use since the 1960s (Monro & Richards 1969), used even in new products such as Amulet®. The active ingredients used in traps designed to attract females are the same chemicals identified 30 years ago (Bateman & Morton 1981, Morton & Bateman 1981; i.e. ammonia generating compounds and/or volatile protein components). The Sterile Insect Release campaign for Q-fly has implemented some technical advances, but still relies on the strain management practices and techniques developed over 30 years ago (but see HAL Final Report CT06017).

What use is a genome project?

This genome project will have benefits in the following areas:

1. Biosecurity

Biosecurity is an increasing concern for the citrus industry in Australia. Techniques for rapid unambiguous identification of suspect larvae are vital for early responses to possible exotic incursions. Using traditional techniques, it is difficult to identify tephritid fruit fly larvae with any degree of confidence. Therefore, entomologists have usually had to wait 1-2 weeks for the adult to emerge before they could identify the species. Experience in other countries has shown that DNA-based tests are feasible (e.g. Armstrong *et al.* 1997, Yu *et al.* 2004), but they are specific to the pest species encountered in those countries. To date, no Australian-specific DNA-based test has yet been developed for routine use.

The sequence information will be used to develop **DNA-based identification systems** for biosecurity. With whole genome sequences, we will have a huge range of testable loci compared to past efforts, which have been limited to a few genes or loci. These systems will allow identification of larval stages of fruit fly. In addition, old methods of DNA fingerprinting of flies to track outbreaks will be considerably improved with the use of newly identified DNA-based techniques (such as RAD-mapping) that are considerably faster and cheaper than DNA microsatellite loci.

2. Better Sterile Insect Technique (SIT)

There are SIT campaigns against different species of fruit flies in different countries world-wide. One common aim that has already been achieved for some species (e.g. medfly) or is in the pipeline for other species, is the construction of male-only strains. These are highly desirable since they double the production capacity of the mass rearing facilities and make releases more effective by removing females from the releases.

Using the genome sequence, we will be able to locate genes necessary for the **construction male-only strains** of Q-fly for use in SIT programs. To date, all known major Q-fly sex-determination genes have been identified by our group of researchers located at UNSW. The whole genome sequence will allow the remaining genes to be identified, which will in turn provide the materials for the construction male-only strains of Q-fly using species-specific genes. Experience in the closely related medfly show that this approach can be highly successful (Gomulski *et a*l. 2008).

3. New areas of research

The sequence information will be used to search for **loci involved in taste and smell**, as has been done for model species of fly. Identification of those genes will provide a new way to search for Q-fly attractants and repellents. Also, as with every piece of basic research, there is the possibility that entirely unexpected area of research will open up. With a complete genome sequence, Australian researchers will be able to attract far **more Federal government funding to Q-fly research**. Our trading partners are already completing genome sequences of their fruit fly pests (medfly and Oriental fruit fly, *B. dorsalis*) and most areas of biological research now involve DNA sequence data.

4. Results publicly available

Ownership of the data will remain in Australia and be freely available to Australian researchers. Currently, the genomes of the medfly, *Ceratitis capitata*, and the oriental fruit fly, *Bactrocera dorsalis*, are being sequenced in the US (Peterson *et al*. 2009). It is likely that overseas groups would eventually sequence Q-fly due to its significance in international trade. However, this may mean that our access to the sequence data would be severely limited. For example, an explicit condition placed on one proposal to sequence

the Q-fly at an overseas sequencing centre was that our access to the data was to be for a single project only. We are committed to making the genomic sequence freely available to prevent any corporation or individual being able to register patents that may result in producers having to pay more for products using that information.

Cost-effectiveness

The project to sequence the human genome cost $1 per base pair of DNA (a total of $3,000,000,000). However, in this project, we have leveraged the Horticulture Australia contribution to the project with funds from other sources via the Science Leverage Fund (UNSW). By using the latest large-scale technologies, we are producing a near-complete genome sequence at a cost of only $14,000 to the Australian horticultural industry. That represents a cost of 0.003 *cents* per base pair.

Materials and methods

DNA isolation

DNA was isolated from the *bent-wings* strain of Q-fly that is maintained at UNSW. This strain was used since it is highly inbred, an important requirement when sequencing necessarily involved extracting DNA from a number of individuals. Isolation of the Q-fly DNA was performed as described in Shearman and Frommer (1998) using only the heads to avoid co-isolation of bacterial DNA from the gut of the flies.

Genome Sequencing

The results reported here were generated from sequencing undertaken at The Ramaciotti Centre for Gene Function Analysis at the University of NSW using Illumina GAIIx sequencing system. Three paired-end libraries with inserts of ~310 base pairs (bp) were sequenced using 102bp technology. Multiple runs are necessary because the small genome fragments are generated randomly. Many of the fragments overlap and so these regions will be sequenced multiple times, while for other genome regions, no fragment will be isolated and those regions will, within any single run, remain unsequenced. As with all genome sequences, many-fold coverage is required to complete the genomic sequence and to eliminate errors in base calling and assembly. The contiguous DNA sequence ("contigs") were assembled from the 100bp reads in collaboration with members of The NSW Systems Biology Initiative using the 256-processor computing cluster at the Intersect High Performance Computing Centre. The programs used for this purpose were AbySS and SOAPdenovo.

Two mate-pair libraries with 5kb inserts was also prepared and sequenced for the purpose of linking contigs into longer stretches of sequence ("scaffolds"). Again in collaboration with The NSW Systems Biology Initiative, the scaffolds were assembled using programs AbySS, SOAPdenovo and SSPACE. The commonly used statistic for assemblies is N50, which indicates the length of the contig that is the median of all assembled contigs. 50% of the sequence will be in contigs larger than N50.

Alignments were made using Bowtie. Basic searches were performed with stand-alone versions of the BLAST suite of programs. Repetitive DNA screens were performed using the RepeatMasker, PILER and RepeatModeler programs. Annotation was performed using the MAKER program with the *Drosophila melanogaster* transcript libraries as reference expression libraries.

Results

The three sets of 102bp paired-end runs produced a total of 170 million reads. This represents approximately 30x coverage of the Q-fly genome (assuming a genome size around 500 million bp). Various k-mer (overlap) values to use for the assembly were tested and the best assembly used a k-mer value of 65bp. N50 value for the initial assembly was 1276bp representing 652,000 contigs. The total genome size estimated from the total length of assembled contigs was 480 Mbp. Scaffolding using the mate pair reads (112 million 36bp reads) increased the N50 to over 7278bp. The total genome size of the scaffolds is closer to 700 Mbp, and it is likely that the presence of many repetitive sequences (see below) was the most probable reason for this over-estimate (i.e. over-representation of certain sequences). The overall G+C content of the scaffolds was 37%.

The completeness of a *de novo* genomic sequencing project is difficult to assess since, by definition, the detailed structure of the particular genome is unknown. However, there are certain sets of genes that are known to be well conserved among all animals. The assembly can be examined for the presence of these genes and the number of copies of each. For a good assembly, each gene should be present in only one copy but maybe spread over more than one contig.

The first set of genes used to test our assembly were the cytoplasmic ribosomal protein (CRP) genes. The set consisted of 89 CRPs based on the *Drosophila melanogaster* set (http://ribosome.miyazaki-med.ac.jp/info.html). Using a blastn search with an e-value cut-off of 1e-10, the CRP hits on scaffolds were as follows:

6 genes hit 0 scaffolds;

65 genes hit 1 scaffold;

14 genes hit 2 scaffolds;

1 gene hit 3 scaffolds;

2 genes hit 4 scaffolds; and

1 gene hit 14 scaffolds.

Due to cross-matching, there is not a complete one-to-one correspondence. Also, a small, but unknown, number of *D. melanogaster* CRPs will not be present in *B. tryoni*. However, the observation that 65 CRPs hit only one scaffold and 14 hit only 2 scaffolds indicates that the assembly is reasonably complete.

A second assessment involved a larger set of genes: a core selection of 458 genes from the much larger Eukaryotic Orthogonal Group (KOG; http://korflab.ucdavis.edu/Datasets/cegma/). A blastn search found scaffolds matching 423 of the 458 genes. Figure 1 shows the similarity of the *B. tryoni* KOG homologue to the *D. melanogaster* KOG gene (% identity) as a function of the coverage of the *D. melanogaster* KOG gene (coverage >1 indicates that the *B. tryoni* homologue contains insertions). Those results show that our assembly contains most of the KOG genes. The observation that coverage mean is approximately 50% was due to a number of factors. The search was relatively stringent, based on DNA, not protein sequences, and it is expected that there will be some divergence at the DNA level in all proteins. Second, not all scaffolds will be complete.

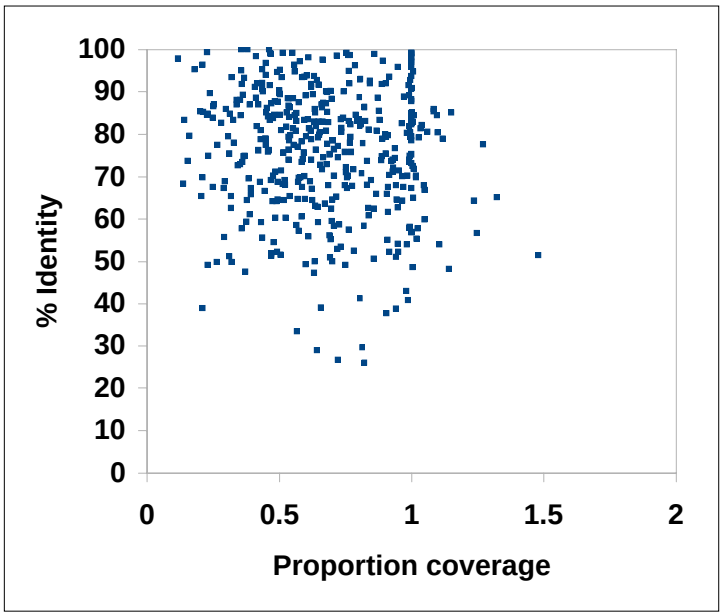

Figure 1. Result of a search of the *B. tryoni* assembly for a core set of 458 KOG genes. Most of the genes in the *B. tryoni* assembly had high identity with and/or coverage of the KOG genes (see text for details).

Repetitive sequences

The assembly was screened for repetitive DNA using RepeatMasker with the *D. melanogaster* repeat libraries. The results are shown in Table 1. The spectrum of repeats matched that expected for a Dipteran insect: no SINE or Alu elements and a large number of Tc1 (*mariner*) elements. Earlier work in our lab had suggested that *mariner* elements would be the most common in the A-fly genome (Green & Frommer 2001) and this turned out to be the case. There was a range of other retroelements with the high occurrence of R1 elements being promising for phylogenetic studies. As common with other non-model species, the use of model species repeat libraries identified only a proportion of the expected repetitive DNA. Identification of the remainder required the use of *de novo* predictor programs. These programs identified a further 3223 repetitive sequence families that are currently being further characterised.

Retroelements		N
	SINEs:	0
	Penelope	3
	LINEs:	19690
	L2/CR1/Rex	502
	R1/LOA/Jockey	19101
	R2/R4/NeSL	54
	RTE/Bov-B	30
	LTR	16702
	BEL/Pao	4688
	Ty1/Copia	939
	Gypsy/DIRS1	10021
	Retroviral	0
DNA transposons		
	hobo-Activator	387
	Tc1-IS630-Pogo	40385
	PiggyBac	174
	Other	4798
Rolling-circles		0
Interspersed repeats		
Small RNA		161
Satellites:		5
Simple repeats		11851
Low complexity		85

Discussion

The purpose of this project was to begin the construction of the complete DNA sequence of the Queensland fruit fly. Our results show that we produced a library of sequence scaffolds that contain most of the genetic loci expected to be present in a typical fly. Furthermore, those genetic loci were not over-represented in the library of scaffolds. Together, these results suggest that out library is both relatively complete and contains little redundancy. Therefore, we have exceeded the project objective, which was to produce a partial sequence of the Q-fly genome. It is often expected that genome projects produce one long sequence for each chromosome. However, this is rarely the case. For example, the published genome of the Golden Delicious apple (742 Mbp total) was published with an assembly with an N50 of 16kb (Velasco *et al.* 2010). Thus our assembly is already competitive with other similar projects.

Furthermore, this has been achieved at relatively low cost to Horticulture Australia Ltd stakeholders. The cost of genome projects like ours is typically split 90:10 as follows: the actual sequencing of the DNA usually represents only about 10% of the total project cost. The other 90% of costs involves DNA preparation prior to sequencing and analysis and annotation of the DNA sequence data. In this project, the Horticulture Australia Ltd stakeholders contribution has been used exclusively to meet that 10% direct cost (as per the initial proposal). The other 90% has been met from other sources, principally the University of NSW.

Future work

The main issue for the existing genome sequence is the need to increase average length of the scaffolds and to annotate the genome. Work is currently under way to increase scaffold length. We have prepared longer insert libraries for paired-end sequencing (500bp) as well as new mate-pair libraries. In combination with existing data that matches known genes to particular chromosomes (Zhao *et al.* 2003), we should be able to significantly increase the size of the scaffolds and anchor them to each of the six chromosomes.

Second, at the time of writing, we have sequenced, but not analysed, transcriptome data from all life stages of Q-fly. This data should provide us with the sequence of all genes that are expressed in Q-fly. Initially, that information will assist in increasing the quality of our assembly (by increasing scaffold lengths). More importantly, it will provide the basis for investigating the biology of Q-fly. We and other researchers will be able to investigate the genes involved in lure responses, chemical interactions and insecticide susceptibility. Of particular interest in our laboratory will be the genes involved in sex-determination pathways. Identifying all the genes involved will be crucial in the development of male-only strains. Male-only strains are used around the world in fruit fly factories to double the

efficiency of those factories. With a Q-fly male-only strain, the output of the Fruit Fly Production Facility at Camden ,NSW could be doubled which would increase the area of SIT campaigns.

The third aspect of future work is the identification of repetitive DNA elements and small RNAs in Q-fly. At first glance, these may seem to have little practical importance. However, it is becoming increasingly apparent that these elements are important in a range of processes in gene expression and evolution (e.g. Specchia *et al*. 2010). One immediate spin-off is that it is likely that these elements will provide the basis for the vastly improved biosecurity tests foreshadowed above.

Technology Transfer

The results of this project are still in the development stage and will not be published in the scientific literature until 2012. However, to make our findings accessible to both researchers and the general public, we will be developing a web-based genome browser. Examples already exist for *Drosophila* (www.flybase.org) and the apple (www.rosaceae.org). We are currently using genome annotation software that specifically produces output for that type of genome browser and have negotiated with the UNSW to host the site. Also, we have and will continue to prepare applications for federal government funding for research into various aspects of Q-fly biology and genetics.

It is envisaged that the main users of the short-term products of this project will be the State Departments of Primary Industries since they administer the current control programs and species identification protocols. The Fruit Fly Research Laboratory has previously been successful in transferring our DNA fingerprinting techniques to the NSW DPI. In the longer term, the annotated genome sequence will be made available to all researchers and other interested parties. We envisage this will be largely complete by the end of 2012. The genome sequence will serve as major infrastructure to stimulate Q-fly research by other research groups. The Fruit Fly Research Laboratory will actively incorporate the results into existing research projects including the production of male-only strains of Q-fly

Recommendations

N/A

Acknowledgements

We would like to thank Assoc. Prof. Bill Sherwin and the staff of the Department of Biological, Earth and Environmental Sciences at UNSW for their support and assistance to the Q-fly genome project.

References

Armstrong, K. F., C. M. Cameron, and E. R. Frampton (1997). Fruit fly (Diptera: Tephritidae) species identification: A rapid molecular diagnostic technique for quarantine application. Bulletin of Entomological Research 87:111-118.

Bateman, M. A. and T. C. Morton (1981). The importance of ammonia in proteinaceous attractants for fruit flies (Family: Tephritidae). *Australian Journal of Agricultural Research* 32: 883-903.

Gomulski, L. M., G. Dimopoulos, Z. Y. Xi, M. B. Soares, M. F. Bonaldo, A. R. Malacrida and G. Gasperi (2008). Gene discovery in an invasive tephritid model pest species, the Mediterranean fruit fly, *Ceratitis capitata*. *Bmc Genomics* 9.

Green CL, Frommer M (2001). The genome of the Queensland fruit fly *Bactrocera tryoni* contains multiple representatives of the *mariner* family of transposable elements. Insect Mol Biol 10:371-386.

Marygold S.J., Roote J., Reuter G., Lambertsson A., Ashburner M., Millburn G., Harrison, P., Yu Z., Kenmochi, N. Kaufman, T.C., Leevers, S.J. & Cook, K.R. (2007) The ribosomal protein genes and Minute loci of Drosophila melanogaster. Genome Biology 2007, 8:R216.

Monro, J. and N. I. Richards (1969). Traps, male lures and a warning system for Queensland fruit fly *Dacus (Bactrocera) tryoni* (Frogg.) (Diptera: Tephritidae) *Australian Journal of Agricultural Research* 20: 325-338.

Morton, T. C. and M. A. Bateman (1981). Chemical studies in the proteinaceous attractants for fruit flies, including the identification of volatile constituents. *Australian Journal of Agricultural Research* 32: 905-916.

Peterson, B. K., E. E. Hare, V. N. Iyer, S. Storage, L. Conner, D. R. Papaj, R. Kurashima, E. Jang, and M. B. Eisen (2009). Big genomes facilitate the comparative identification of regulatory elements. PLoS ONE 4:e4688.

Shearman, D. C. A. and M. Frommer (1998). The Bactrocera tryoni homologue of the Drosophila melanogaster sex-determination gene doublesex. Insect Molecular Biology 7: 355-366.

Specchia, V., Piacentini, L., Tritto, P., Fanti, L., D'Alessandro, R., Palumbo, G., Pimpinelli, S., Bozzetti, M.P. (2010). Hsp90 prevents phenotypic variation by suppressing the mutagenic activity of transposons. Nature 463(7281): 662--665.

Velasco R, Zharkikh A, Affourtit J, Dhingra A, Cestaro A, Kalyanaraman A, Fontana P, Bhatnagar SK, Troggio M, Pruss D et al. (2010) The genome of the domesticated apple (Malus x domestica Borkh.). Nat Genet. 42:833-9.

Yu, D. J., G. M. Zhang, Z. L. Chen, R. J. Zhang, and W. Y. Yin. 2004. Rapid identification of Bactrocera latifrons (Dipt., Tephritidae) by real-time PCR using SYBR Green chemistry. Journal of Applied Entomology 128:670-676.

Zhao JT, Frommer M, Sved JA, Gillies CB. (2003) Genetic and molecular markers of the Queensland fruit fly, Bactrocera tryoni. J Hered. 94:416-420.