

This document describes the features on a first-generation oligonucleotide microarray developed for the honey bee genome. Funding for this project was provided by USDA-National Research Initiative grant AG2004-36504-14277 (G.E. Robinson, PI, M. Band, J.D. Evans, G. deGrandi Hoffman, K.P. White, Co-PIs) “**Honey Bee Applied Genomics and Development of a Whole-Genome Array**”. More information on the array and connected research questions available elsewhere at Beespace and via Gene Robinson ([generobi@life.uiuc.edu](mailto:generobi@life.uiuc.edu)).

Included in this folder are two **subfolders**:

1) **Array\_fasta**: Fasta text files for the input and oligo sequences:

*ArraySetcombinedfin.txt* = Compilation of all input sequences

*OfficialGeneSet\_Array.txt* = Sequences from the honey bee genome ‘Official Gene Set’ (see HGSC (2006) Insights into social insects from the genome of the honey bee *Apis mellifera*, *Nature* Oct. 26) used in defining the array.

*Add-ons\_Array.txt* Additional sequences from the honey bee genome and the genomes of various parasites and pathogens included on the array (described below).

*Oligoset13440.txt* = Oligo sequences for these inputs.

‘*Add-on files*’ = subfolder containing FASTA files for all non-OGS sequences in the array, separated by category

2) **Array\_Analysis**: Flat files contrasting sources for the honey bee genome-array

*ArraySetSummary.xls* = Connections between input sequences and oligo ID’s. Shows sequence category (as below), scaffold location on assembly 4.0, oligos that are largely identical in sequence (column Oligosynon) and source sequences that are overlapping or redundant, including splice variants(column ArraySetSynon). Last column has input DNA sequences.

*ArraySetSummaryBrief.xls* = same as above but without DNA sequences.

*PeptsetvsDmel43ed.xls* = BLAT (Jim Kent, UCSC) alignments between the input sequences and the *Drosophila melanogaster* peptide set (version 4.3, 2006). These matches need not reflect orthology and are given mainly as a cross-check for input sequences. Details on BLAT criteria and caveats available below and from JDE, [evansj@ba.ars.usda.gov](mailto:evansj@ba.ars.usda.gov)

*Arraylookup3.xls* = Microsoft Excel workbook with all worksheets for above analysis

Statistics from the Array set:

---

**Input sequences:** A total of 13,145 sequences were used to design oligos, a set primarily from the Honey Bee Genome Sequencing Consortium ‘Official Gene Set’ (circa 11/2005) but also including genomic components and honey bee pathogens as detailed in Table 1 below.

Non-OGS sequences include

- 1) Variable exons from the antimicrobial peptide apidaecin (Genbank and Evans, J.D., *unpublished*)
- 2) Variable exons from the IG-family gene Dscam
- 3) miRNA precursor candidates from the bee genome (Weaver, D.B., et al., *submitted*).
- 4) non-OGS EST’s from a subtractive library biased toward larval genes upregulated with exposure to the bacterial pathogen *Paenibacillus larvae*, Evans, J.D., *unpublished*. RNA was derived from 1<sup>st</sup>-instar honey bee larvae challenged with bacteria as described in Evans and Pettis, 2005, *Evolution*, **59(10)**, 2270-2274).
- 5) Non-OGS EST’s from the Univ. Illinois bee brain EST project (Whitfield, C. W., Band, M. R., Bonaldo, M. F., Kumar, C. G., Liu, L., Pardinas, J. R., Robertson, H. M., Bento Soares, M. & Robinson, G. E., 2002, *Genome Research*, **12**, 555-566.).
- 6) Representative genes from viral, fungal, bacterial, and microsporidian pathogens of honey bees (all in Genbank, ID’s in fasta file).

Table 1.

Category	Total
OGS	10620
Apidaecin exons	11
Dscam exons	81
miRNA	59
JDE_EST	81
UI_EST	2271
Pathogen	22
Grand Total	13145

**Oligo Design:** Long oligos for the array were developed by Debashis Rana and Gos Micklem (<http://www.gen.cam.ac.uk/Research/micklem.htm>) at Cambridge University, using a modified version of OligoArray 2.1 in an iterative process to identify unique sequences (60-69mers) from each of the described (above) bee-related genes and gene fragments. The set of oligos was selected to have as tight a melting temperature distribution as possible, and to avoid repetitive sequences and other anomalies. A total of 12,915 unique oligos were generated (see below for redundancies) representing all but three of the 13,145 source sequences. Of those three (the pathogen gene PIDNAK, the EST sequence JDEA05\_1Def3, and the candidate miRNA precursor Hcmir13a), the EST and miRNA were represented by 98% identical oligos in the array. Reverse-strand oligos were added for 525 predictions, focusing on EST reads and transcripts predicted for bee pathogens (Table 2). As such the final set contains 13,440 oligos (sequences in Array\_fasta/Oligoset13440.txt). The design process was similar to that of the INDAC long oligo set designed for the fruit fly *Drosophila melanogaster* and available at:

<http://www.flymine.org/release-5.0/aspect.do?name=INDAC> and <http://www.flychip.org.uk/services/core/FL002>.

All but three of the 13,145 source sequences were successfully represented in the oligo set. Of those three (the pathogen gene PIDNAk, the EST sequence JDEA05\_1Def3, and the candidate miRNA precursor HCmir13a), the EST and miRNA were represented by 98% identical oligos in the array. A total of 12,915 unique oligos were generated (see below for redundancies). Reverse-strand oligos were added for 525 predictions, focusing on EST reads and transcripts predicted for bee pathogens (Table 2). As such the final set contains 13,440 oligos (sequences in Array\_fasta/Oligoset13440.txt).

Table 2. Reverse-strand oligos by Sequence category. Sources shown in ArraySetSummaryBrief.xls.

EST	415
miRNA	57
OfficialGene	31
Pathogen	22

**Oligo and Sequence Redundancy:** Distinctly numbered oligos had the same or similar sequences 69 times (>59/69 nt alignment, < 2 mismatches). Different source sequences matched identical oligos (>59/69, < 2 mismatch) 100 times, 44 of which were not genes with predicted splice variants (which were redundant in OGS). 18 were gene calls with splice variants for which oligos matched each variant. 639 source sequences showed matches at the sequence level but did not have identical oligo matches. Of these 524 reflect either splice variants or shared exons (e.g., Dscam exons vs. an entire proposed transcript). 115 are not indicated as splice variants and these appear to be redundant sequences in the source files, either from multiple predictions in OGS or from unrecognized similarity between EST's and other EST's or OGS. Most redundancies were single pairs, although one oligo sequence was similar across 6 distinctly called oligos.

**Other Comparisons:** Oligos were placed onto Honey bee Assembly 4 using BLAT (<http://www.kentinformatics.com/>) alignments. The 'Array\_BLAT' subfolder in ArrayAnalysis shows the statistics for these alignments, along with cases where oligos aligned to multiple places in Assembly 4. Summarize 'best hit' alignments are in ArraySetSummary.xls, and ArraySetSummaryBrief.xls. Source peptide sequences (e.g., from the OGS) were aligned with the *Drosophila melanogaster* protein set 4.3, also using BLAT. Best matches are again given in the summary files and in better detail in the Array\_BLAT subfolder. BLAT alignments NEED NOT reflect orthology, and are perhaps a bit greedy that way, but these matches do tend to be accurate, probably on par with BLASTP best matches.