

UIUC Honey bee oligo 13K v1
May 14, 2007

Array Development – text by Jay D. Evans, Gene E. Robinson and Gos Micklem.

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”.

The developmental files can be accessed at

http://www.biotech.uiuc.edu/centers/Keck/Functional_genomics/Honey%20Bee%20Oligo.htm

Input sequences:

A total of 13,145 sequences were used to design oligos, as detailed below:

- 1) A set primarily from the Honey Bee Genome Sequencing Consortium ‘Official Gene Set’ (circa 11/2005) (N = 10620).
- 2) Variable exons from the antimicrobial peptide apidaecin (Genbank and Evans, J.D., unpublished) (N = 11).
- 3) Variable exons from the IG-family gene Dscam (N = 81).
- 4) miRNA precursor candidates from the bee genome (Weaver, D.B., et al., submitted) (N = 81).
- 5) 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biased toward larval genes upregulated with exposure to the bacterial pathogen *Paenibacillus* larvae, Evans, J.D., unpublished.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 1st-instar honey bee larvae challenged with bacteria as described in Evans and Pettis, 2005, *Evolution*, 59(10), 2270-2274 (N=81).
- 6) 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.) (N=2271).
- 7) Representative genes from viral, fungal, bacterial, and microsporidian pathogens of honey bees (all in Genbank, ID’s in fasta file) (N=22).

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

(EST – 415; miRNA – 57; OGS – 31, and pathogen – 22). 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>.

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.

Printing the Array – text by Mark Band and Al Bari

Oligos were synthesized by Invitrogen (San Diego, CA) and aliquoted into 384 well plates in Sodium Phosphate buffer. Final concentration of the oligos was 20 uM (micromolar). Arrays were printed on Corning UltraGAPS slides using an Omingrid 100 printing robot (Genomic Solutions, Ann Arbor Michigan) with Arrayit SMP2.5 capillary pins. Following printing slides were stored in vacuum bags purged with Argon gas.

Creating the Array Design File (ADF) – text by Amro Zayed

To ensure that the ADF contains the latest annotation, we first blasted all Honey bee oligos against: 1) The prerelease updated version of OGS v2 (<http://racerx00.tamu.edu/downloadFASTA.html> - circa 4/9/2007), 2) NCBI's gene predictions (Ref RNA) for the Honey bee (circa 4/18/2007), 3) All honey bee derived EST sequences on NCBI (circa 4/25/2007), and 4) The latest assembly of the Honey bee genome (AMEL 4.0, circa 4/25/2007). We used blastn with no filtering and we initially retained all matches with an evalue smaller than $1e^{-3}$. We then removed blast hits that had an alignment length that is > 4 bp less than the oligo length and/or had an alignment identity less than 95%.

Except for the control groups and the Pathogen set, we assigned an oligo's "Reporter Name", regardless of which set it was originally designed from, based on the best blast match to the above mentioned databases, assigning priority in the following order: prerelease OGS v2, NCBI's gene predictions, EST sequence names, and AMEL 4.0

assembly location. In cases where an oligo matched more than one sequence in the prerelease OGS v2 or NCBI's gene predictions at the same evalue, we assigned the "Reporter Name" and corresponding database accession ID to the gene with the highest numerical value, but included the list of equally matching genes in the "Reporter Comment" field. If an oligo matched to both prerelease OGS v2 and NCBI's gene prediction, we assigned the "Reporter Name" as the OGS v2 gene name followed by the definition line from NCBI's gene prediction, in addition to assigning both accession numbers to the oligo. When an oligo matched a prerelease OGS v2 gene, we added its Drosophila ortholog as computed by C. Elsik (http://racex00.tamu.edu/bee_resources.html) in the "Reporter Comment". We also used blastp to query the Honey bee gene against Drosophila melanogaster v.5.1 genes (<http://flybase.bio.indiana.edu/> circa 5/1/2007), and the best match was also reported in the "Reporter Comment".

For oligos that did not match entries in any of the above mentioned databases, we retained the original annotation information used to design the oligo. Similarly, we retained all the annotation information from the design files for the control sequences and for the Pathogen set.