

Drosophila Tiling 2.0R array design and library files

Description and overview

The Drosophila 2.0R tiling array is a single 49-format, 5 micron feature-size design that contains 25-mer probes in perfect match/mismatch pairs, with average center base spacing of approximately 38 bases, and designed to be identical to the plus strand of the genome sequence. The array is based on the BDGP version 5 genome sequence, which, compared to previous versions, includes improved assembly and incorporates extra heterochromatic regions that include many known and predicted genes. The tiling strategy allows genome-wide repetitive elements to be represented on the array, although these are masked out in the standard (default) analysis library files (bmap files). Unmasked library files are available upon request through Affymetrix Technical Support. Reverse array synthesis is available on request on a per wafer basis, to interrogate the opposite strand.

Resource dates and version numbers

The Drosophila 2.0R tiling array is based on the Berkeley Drosophila Genome Project (BDGP) genome release 5 (<http://www.fruitfly.org>). The standard (default) analysis library file (bmap file) uses a combination of Flybase 5.2 annotations (<http://flybase.net>) and University of Santa Cruz (UCSC) RepeatMasker tracks (August 2007) (<http://www.genome.ucsc.edu>) to remove probes from repetitive regions.

Content

The BDGP 5 genome includes improved euchromatin assembly, guided by restriction fragment digest fingerprints. Also included are improved heterochromatin assemblies, both for the long chromosome arms and as separate assemblies. For more information see:

<http://www.fruitfly.org/sequence/release5genomic.shtml>

<http://www.fruitfly.org/sequence/README.RELEASE5>

Heterochromatin assembly was provided by the Drosophila Heterochromatin Genome Project (<http://www.dhgp.org>)

The array represents the following sequences from BDGP:

na_armX.dmel.RELEASE5
na_arm2L.dmel.RELEASE5
na_arm2R.dmel.RELEASE5
na_arm3L.dmel.RELEASE5
na_arm3R.dmel.RELEASE5
na_arm4.dmel.RELEASE5
na_XHet.dmel.RELEASE5
na_YHet.dmel.RELEASE5
na_2LHet.dmel.RELEASE5
na_2RHet.dmel.RELEASE5
na_3LHet.dmel.RELEASE5
na_3RHet.dmel.RELEASE5

Drosophila Tiling 2.0R array design and library files

Revision Date: 09-26-2007

Revision Version: 1.0

na_armU.dmel.RELEASE5

The sequence "na_armUextra.dmel.RELEASE5" was not included, per BDGP advice, since it contains lower quality and redundant data.

The Drosophila mitochondrion (GenBank accession NC_0017090) was also included.

A set of microRNAs were also tiled (as probes <25 bases when necessary). See Appendix 1.

Tiling strategy.

The tiling strategy was designed to tile the whole genome, including repeats and heterochromatic regions, on a single 49-format, 5 micron feature size array. Probes were scored for linear hybridization dose-response characteristics, according to the Affymetrix probe model algorithm, and then down-weighted according to:

- the number of times the 25-mer occurs in the genome
- the number of times the central 23-mer, 21-mer, 19-mer 17-mer and 15-mers occur in the genome

And up-weighted according to:

- the number of times the 25-mer occurs with a different base at the central position (i.e. SNP-like probes).

Then, the strategy proceeded to walk through the genome, selecting the highest scoring probe every 38 +/-10 base window. If a new window contained a probe that had previously been selected for another window, then that probe was chosen again, regardless of score. In this way, repetitive regions could be tiled in an economical fashion. The 38 base spacing was chosen empirically, such that the available array space was essentially 100% utilized.

Implications of the tiling strategy upon analysis.

The tiling strategy allows repetitive and heterochromatic regions to be tiled economically, since one probe can represent the same region in many copies of the same repeat. In this way, a small set of probes can essentially represent a whole family of closely related repetitive elements.

However, since the strategy starts at a single point and walks in 38 base windows through the genome, the probes that are chosen are influenced by that path and are not the absolute minimal set. For example, a new probe selection window may only partly overlap a previously chosen probe, in which case that probe is not considered for the current window. Thus, if one were to align all probes on the array to a single member of a gene family that had high sequence similarity amongst the members, one would find that the tiling density for that single member was > 38 bases.

If one were to require maximum discrimination between repetitive elements, one would select probes that were most unique (in the genome) for each window. In this way, identical copies of repeats would tend to have the same probes chosen (window phase issues aside), and yet diverged and older elements would tend to have more unique probes wherever possible. We did not do this, however, because a test run showed that this strategy overflowed the available array space. Thus, the strategy that was implemented actually does the opposite of this; namely it tiles repeated regions in a way

Drosophila Tiling 2.0R array design and library files

Revision Date: 09-26-2007

Revision Version: 1.0

that makes the array less able to discriminate between closely-related copies of repeated elements. For some elements that are sufficiently diverged, more unique probes would tend to be selected, simply because those elements are less related to other elements and probes. Thus, it remains possible that some individual repeats may be distinguishable from the rest of the population, although these might be older, mutated and therefore inactive repeats. Since the strategy had no knowledge of genic vs non-genic content, there are implications for being able to discriminate between closely related genes. The strategy may choose a more similar set, whereas a less similar set would actually be required if one wished to tell the two very similar copies apart. Thus, for certain specific genes, the version 1.0 array, with 35 base spacing and a different tiling strategy, may be more discriminatory.

A design-time analysis of all probes tiled vs. all *D. melanogaster* Refseq mRNAs showed that every Refseq was covered. The tiling strategy impacts the extent to which each Refseq is covered. Single copy genes are typically covered by 38-base spaced (on average) unique probes. Genes that are from repeated families may appear to be covered to higher density if every probe is mapped to every location (however, the probes were not mapped this exhaustively in the supplied library files – see later). Some genes appear to be hit by relatively few unique probes, if the gene contains a large number of repeating units (the tiling strategy will tend to pick the same probes over and over).

Differences between Drosophila Tiling 2.0 and Drosophila Tiling 1.0 Arrays

- The 2.0 array has approx. 38 base spacing (vs. 35 on the 1.0 array).
- The 2.0 array is based on the genome version 5 (vs. version 3) and consequently has more heterochromatin on the long chromosome arms and as separate contigs.
- The 1.0 array was RepeatMasked before probe selection, whereas the 2.0 array was not.
- The tiling strategy for the 2.0 array allows heterochromatic and repetitive-element probes to be tiled, at the cost of reducing discrimination between closely related sequences.
- **Note:** Due to the different content, design strategy and analysis library file content, direct comparisons of results between the Drosophila Tiling 1.0R Array and the Drosophila Tiling 2.0R Array are not recommended.

Analysis library files (bpmmap files)

Library files were created to represent the actual probe that was selected for each window in the tiling path. Since the tiling strategy re-used previously-selected probes from previous windows, some physical probes on the array map to more than one location in the genome, and the library files reflect this redundancy. The library files do not map every probe to every possible perfect match location in the genome; only to those locations where that probe was selected in the tiling path. In this way, the overall probe density is maintained fairly constant across the genome. If one were to map all probes to all locations, then repeated regions would have a higher probe density than single copy regions, due to the effect of window phase and polymorphisms during the probe selection procedure.

Drosophila Tiling 2.0R array design and library files

Revision Date: 09-26-2007

Revision Version: 1.0

Library files use the standard UCSC chromosome nomenclature (e.g. chr2L) rather than the BDGP/Flybase nomenclature (e.g. na_arm2L) for compatibility with the Affymetrix TAS analysis software and the UCSC genome browser.

The standard (default) library file is available from the support materials section of the Drosophila 2.0 product web page or the library file webpage:

<http://www.affymetrix.com/support/technical/libraryfilesmain.affx>

Designed for the user who is not interested in studying repetitive elements specifically, the standard library file masks out all probes that fall within repeat regions. Two sources of repeat mapping were combined to create this file. The first file was supplied courtesy of FlyBase (version 5.2), and includes hits to a slightly expanded (vs RepBase[tm]) set of transposable elements. The second file was taken from the UCSC genome track, and includes masking for low complexity regions, simple tandem repeats, heterochromatic repeats, centromeric and telomeric repeats, and the distal part of Arm4, which was not masked in the Flybase (version 5.2) files. Note that this standard library file may still contain probes that map to the genome multiple times (either as perfect matches or otherwise), which may be derived from repeated gene families and other sequences that were not repeat-masked.

The second library file, available upon request through Affymetrix Technical Support, simply contains all probes in the intended tiling path, and does not mask out any probes based on repeat annotations.

Appendix1.

```
>bantam
TGAGATCATTGAAAGCTGAT
>dme-let-7
TGAGGTAGTAGGTTGTATAGTA
>dme-miR-1
TGGAATGTAAAGAAGTATGGAG
>dme-miR-10
ACCCTGTAGATCCGAATTTGTT
>dme-miR-100
AACCCGTAAATCCGAACCTTGTG
>dme-miR-11
CATCACAGTCTGAGTTCTTGCT
>dme-miR-12
TGAGTATTACATCAGGTAAGTGG
>dme-miR-124
TAAGGCACGCGGTGAATGCCAA
>dme-miR-125
TCCCTGAGACCCTAACTTGTGA
>dme-miR-133
TTGGTCCCCTTCAACCAGCTGT
>dme-miR-13a
TATCACAGCCATTTTGATGAGT
>dme-miR-13b
TATCACAGCCATTTTGACGAGT
>dme-miR-14
TCAGTCTTTTTCTCTCTCTAT
>dme-miR-184
TGGACGGAGAACTGATAAGGGC
>dme-miR-210
```

Drosophila Tiling 2.0R array design and library files

Revision Date: 09-26-2007

Revision Version: 1.0

```
TTGTGCGTGTGACAGCGGCTAT
>dme-miR-219
TGATTGTCCAAACGCAATTCTTG
>dme-miR-263a
AATGGCACTGGAAGAATTCACG
>dme-miR-263b
CTTGGCACTGGGAGAATTCACA
>dme-miR-274
TTTTGTGACCGACACTAACGGGTAAT
>dme-miR-275
TCAGGTACCTGAAGTAGCGCGC
>dme-miR-276a
TAGGAACTTCATACCGTGCTCT
>dme-miR-276b
TAGGAACTTAATACCGTGCTCT
>dme-miR-277
TAAATGCACTATCTGGTACGAC
>dme-miR-278
TCGGTGGGACTTTTCGTCCGTTT
>dme-miR-279
TGACTAGATCCACACTCATTAA
>dme-miR-280
TGTATTTACGTTGCATATGAAATGATA
>dme-miR-281
TGTCATGGAATTGCTCTCTTTGT
>dme-miR-282
AATCTAGCCTCTACTAGGCTTTGTCTGT
>dme-miR-283
TAAATATCAGCTGGTAATTCTG
>dme-miR-284
TGAAGTCAGCAACTTGATTCCAGCAATTG
>dme-miR-285
TAGCACCATTTCGAAATCAGTGC
>dme-miR-286
TGACTAGACCGAACACTCGTGC
>dme-miR-287
TGTGTTGAAAATCGTTTGCAC
>dme-miR-288
TTTCATGTTCGATTTTCATTTTCATG
>dme-miR-289
TAAATATTTAAGTGGAGCCTGCGACT
>dme-miR-2a
TATCACAGCCAGCTTTGATGAG
>dme-miR-2b
TATCACAGCCAGCTTTGAGGAG
>dme-miR-2c
TATCACAGCCAGCTTTGATGGG
>dme-miR-3
TCACTGGGCAAAGTGTGTCTCA
>dme-miR-303
TTTAGGTTTTCACAGGAAACTGG
>dme-miR-304
TAATCTCAATTTGTAAATGTGA
>dme-miR-305
ATTGTAATTCATCAGGTGCTCT
>dme-miR-306
TCAGGTACTTAGTGACTCTCAA
```

Drosophila Tiling 2.0R array design and library files

Revision Date: 09-26-2007

Revision Version: 1.0

```
>dme-miR-307
TCACAACCTCCTTGAGTGAGCG
>dme-miR-308
AATCACAGGATTATACTGTGAG
>dme-miR-309
GCACTGGGTAAAGTTTGTCTTA
>dme-miR-310
TATTGCACACTTCCCAGCCTTT
>dme-miR-311
TATTGCACATTACCGGCCTGA
>dme-miR-312
TATTGCACTTGAGACGGCCTGA
>dme-miR-313
TATTGCACTTTTCACAGCCCGA
>dme-miR-314
TATTGAGCCAATAAGTTCGGC
>dme-miR-315
TTTTGATTGTTGCTCAGAAAGC
>dme-miR-316
TGTCTTTTTCCGCTTACTGGCG
>dme-miR-317
TGAACACAGCTGGTGGTATCCA
>dme-miR-318
TCACTGGGCTTTGTTTATCTCA
>dme-miR-31a
TGGCAAGATGTCGGCATAGCTG
>dme-miR-31b
TGGCAAGATGTCGGAATAGCTG
>dme-miR-33
AGGTGCATTGTAGTCGCATTGT
>dme-miR-34
TGGCAGTGTGGTTAGCTGGTTG
>dme-miR-4
ATAAAGCTAGACAACCATTGAA
>dme-miR-5
AAAGGAACGATCGTTGTGATAT
>dme-miR-6
TATCACAGTGGCTGTTCTTTTT
>dme-miR-7
TGGAAGACTAGTGATTTTGTG
>dme-miR-79
TAAAGCTAGATTACCAAAGCAT
>dme-miR-8
TAATACTGTCAGGTAAAGATGT
>dme-miR-87
TTGAGCAAAATTTTCAGGTGTGT
>dme-miR-92a
CATTGCACTTGTCCCAGCCTAT
>dme-miR-92b
AATTGCACTAGTCCCAGCCTGC
>dme-miR-9a
TCTTTGGTTATCTAGCTGTATG
>dme-miR-9b
TCTTTGGTGATTTTAGCTGTAT
>dme-miR-9c
TCTTTGGTATTCTAGCTGTAGA
>dme-miR-iab-4-3p
```

Drosophila Tiling 2.0R array design and library files

Revision Date: 09-26-2007

Revision Version: 1.0

CGGTATACCTTCAGTATACGTAAC

>dme-miR-iab-4-5p

ACGTATACTGAATGTATCCTGA