Quality Assessment of the Affymetrix U133A&B probe sets by target sequence mapping and expression data analysis

Yuri L. Orlov et al., August 2007
Topic

• careful checking of probe sequences of microarray
• verification of probe sequences against reference data in DBs => identification of unreliable probes
• remove data of problematic probes before analysis
• reevaluation of three corrected cancer data sets
• improvements in selection of differentially expressed genes, clustering and construction of co-regulatory expression networks expected
U133A & B

This research work is based on 2 specific microarrays

- Affymetrix Human Genome U133A Array
- Affymetrix Human Genome U133B Array
- Official annotation files for both arrays from Affymetrix, currently release 30 from 15/11/09
  http://www.affymetrix.com/support/technical/annotationfilesmain.affx?highlight=true&rootCategoryId=

- idea for this work seems to come from the Micro Array Quality Control project (MAQC) initiated by the FDA
Human Genome U133A GeneChip® Array

(1) Probe Array

(2) Probe Set

The Human Genome U133A GeneChip® array represents more than 22,000 full length genes and EST clusters.

(3) Probe Pair

Each Perfect Match (PM) and MisMatch (MM) spots are associated by pairs.

(4) Spot

Each spot contains $\sim 4 \times 10^8$ copies of a specific probe which is complementary to the genetic material of interest. One probe is a single stranded, fluorescently labelled oligonucleotide (25 mers).
## Critical Specifications for GeneChip® Human Genome Products

<table>
<thead>
<tr>
<th></th>
<th><strong>Cartridge Format</strong></th>
<th><strong>Plate Format</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Human Genome</strong></td>
<td><strong>Human Genome</strong></td>
</tr>
<tr>
<td></td>
<td><strong>U133 Plus 2.0</strong></td>
<td><strong>U133A 2.0</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Array</strong></td>
<td><strong>Array Plate</strong></td>
</tr>
<tr>
<td>Number of transcripts</td>
<td>~47,400</td>
<td>~18,400</td>
</tr>
<tr>
<td>Number of genes</td>
<td>&gt;38,500</td>
<td>&gt;14,500</td>
</tr>
<tr>
<td>Number of probe sets</td>
<td>&gt;54,000</td>
<td>&gt;22,000</td>
</tr>
<tr>
<td>Feature size</td>
<td>11 µm</td>
<td>8 µm</td>
</tr>
<tr>
<td>Oligonucleotide probe length</td>
<td>25-mer</td>
<td>25-mer</td>
</tr>
<tr>
<td>Probe pairs/sequence</td>
<td>11</td>
<td>11</td>
</tr>
</tbody>
</table>

Control sequences included:
- Hybridization controls: `bioB, bioC, bioD, cre, dap, lys, phe, thr`
- Poly-A controls: 100 probe sets
- Normalization control set: GAPDH, beta-Actin, ISGF-3 (STAT1)
- Housekeeping/Control genes: GAPDH, beta-Actin, ISGF-3 (STAT1)

Detection sensitivity: 1:100,000*

*As measured by detection of pre-labeled transcripts derived from human cDNA clones in a complex human background.*
Probe Set Naming I

• ..._at: (anti-sense target) detects antisense strand of given gene, these are unique probes

• _a_at: (gene family probeset) recognize multiple transcripts of same gene
  _s_at: (identical probeset) recognize multiple transcripts from different genes
  _x_at: (mixed probe set) cross-hyb. with other sequences used for design

• a, s, x derived from gene cluster + gene family info
Probe Set Naming II

**Legend**

- **G**: a set of sequences belonging to the same gene family
- **S**: a sequence
- **PS**: a probe set
- **P**: a probe in a probe set
Verifications

- Verification of probe quality - as annotated! (assuming identity of probe and annotation)
- Get official sequence information for the probes from Affymetrix annotation files
- Verify unique sequence to gene mapping
- Verify that sequence maps into human genome
- Verify correct strand orientation
- Consideration of repeated exon elements
BLAT Analysis

- BLAT sequence search at 90% similarity level
- check of overlap with exonic regions in RefSeq and mRNA / spliced EST variants in hg17 and hg18
- mapping of probe sets to gene sequence blocks based on initial target sequences
- results of analysis (chromosomal coordinates, orientation, overlapping details with exons / repeats...) stored in local DB against probe set ID
Used Reference Data

- NCBI Human Genome (hg17 and hg18)
- RefSeq
- NCBI GEO GSE4922 breast cancer data sets
- NCBI GEO GDS1962 brain cancer data sets
- NCBI GEO GGSE586 lung cancer data sets
Used Tooling

- For sequence verification:
  - BLAT (Difference to BLAST)
  - UCSC Genome Browser
  - Software developed at GIS (Genom Institute of Singapore) and BII (Bioinformatics Inst. Singapore)

- For statistical evaluations:
  - SAM 3.1 (Statistical Analysis of Microarrays)
  - Statistica 6
  - StatXact 6
Probe Quality Criteria

1. sequence with unique locus in human genome
2. perfect match of transcript
3. correspond to sequence of transcribed strand at this locus including correct strand orientation
4. no overlapping with any other non-gene sequence
5. correspond to mature RNA (only exons included)
Problematic Sequences

<table>
<thead>
<tr>
<th># locations (Hg18)</th>
<th>Tag1</th>
<th>Tag2</th>
<th>Tag3</th>
<th>Tag4</th>
<th>Tag5</th>
<th>Tag6+</th>
<th>Tag0</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Affymetrix IDs</td>
<td>42708</td>
<td>450</td>
<td>129</td>
<td>67</td>
<td>42</td>
<td>84</td>
<td>1212</td>
<td>44692</td>
</tr>
<tr>
<td>%</td>
<td>95.56</td>
<td>1.0</td>
<td>0.28</td>
<td>0.14</td>
<td>0.09</td>
<td>0.18</td>
<td>2.71</td>
<td>100</td>
</tr>
</tbody>
</table>

Sequences in U133A and U133B

- Tag0 - not found in human genome
- Tag1 - found exactly once => correct sequence
- Tag2+ - found at multiple loci
Tag0 and Tag2+

- **Tag0:**
  45% “xenosequence/non-human”  
  (cow, mouse, pathogens, rat ...)  
  17% classified as low-accuracy

- **Tag2+:**
  81737\_at found at 22 different locations  
  213089\_at found at more than 11 locations

- Probedesign based on Genbank sequences without verification of mapping to human genome
Repeats in Tag I

<table>
<thead>
<tr>
<th>Set of genome repeats</th>
<th>Repeat class</th>
<th># in target sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple repeats</td>
<td>Simple repeat, Low complexity</td>
<td>3233</td>
</tr>
<tr>
<td>Short transposons (&lt;300 bp)</td>
<td>DNA, SINE/Alu, SINE/MIR</td>
<td>4347</td>
</tr>
<tr>
<td>Long transposes (&gt;300 bp)</td>
<td>LINE/CR1, LINE/L1, LTR/ERV1/ERVK/ERVL/MaLR</td>
<td>5420</td>
</tr>
<tr>
<td>Non-transposons and satellites</td>
<td>Other, RNA, rRNA, Satellite, scRNA, snRNA, srpRNA</td>
<td>80</td>
</tr>
</tbody>
</table>

- lead to cross hybridization and wrong detection of expressed genes
Inverse Sequences

- inversely oriented if it matches the opposite strand of RefSeq, mRNA or EST related gene
- match to NAST (natural antisense transcript) 30%

<table>
<thead>
<tr>
<th>Sets</th>
<th>Correct orientation</th>
<th>Misoriented to intended transcript in as verified by</th>
<th>Total in Tag1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>Match to NAST</td>
<td>Manual curation</td>
</tr>
<tr>
<td># Target seq</td>
<td>41898</td>
<td>13260</td>
<td>370</td>
</tr>
<tr>
<td>%</td>
<td>93.74</td>
<td>29.66</td>
<td>0.82</td>
</tr>
</tbody>
</table>
# Total Picture

<table>
<thead>
<tr>
<th>Target sequences groups</th>
<th>Non-redundant # of probesets</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total # of non-Tag1 sequences, including:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total # of non-Tag1 sequences, including:</td>
<td>1984</td>
<td>4.43</td>
</tr>
<tr>
<td>Tag0</td>
<td>1212</td>
<td>2.71</td>
</tr>
<tr>
<td>Tag2+</td>
<td>772</td>
<td>1.72</td>
</tr>
<tr>
<td><strong>Total # misoriented target sequences, including:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RefSeq IDs (by blocks)</td>
<td>138</td>
<td>0.3</td>
</tr>
<tr>
<td>mRNA GenBank (by blocks)</td>
<td>302</td>
<td>0.67</td>
</tr>
<tr>
<td>Manual curation protocol</td>
<td>370</td>
<td>0.82</td>
</tr>
<tr>
<td><strong>Total # of target sequences overlapped with repeats, including:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overlap 80-100% of target sequence length</td>
<td>761</td>
<td>1.7</td>
</tr>
<tr>
<td>Overlap 60-80%</td>
<td>936</td>
<td>2.09</td>
</tr>
<tr>
<td>Overlap 40-60%</td>
<td>1690</td>
<td>3.78</td>
</tr>
<tr>
<td><strong>Total # of useful Tag1 target sequences, including:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overlaps with observed transcripts in opposite strand</td>
<td>13260</td>
<td>29.66</td>
</tr>
<tr>
<td>Misoriented to ESTs in Tag1</td>
<td>487</td>
<td>1.08</td>
</tr>
<tr>
<td>Target sequences with 20-40% of repeats</td>
<td>2409</td>
<td>5.39</td>
</tr>
<tr>
<td>Target sequences with &lt;20% of repeats</td>
<td>1210</td>
<td>2.7</td>
</tr>
<tr>
<td>TOTAL # of Affymetrix target sequences</td>
<td>44692</td>
<td>100</td>
</tr>
</tbody>
</table>
Expression Levels

CV per sequence category
Mean per sequence category
Mean for inverted sequences
Repeats in Sequences

• Data for G1 and G3 grade breast cancer which show expression for at least 4000 probe sets

• Through overlaps with repeats - less specific binding, this means: portion of these probe sets in statistically significant expressed genes decreases => decreased recognition of cancer signals

• Only relevant for longer repeat overlaps

• No impact for simple repeats and low complexity sequences
Cross-Hybridization I

- Kendall $\tau$ rank correlation between probe set values
- Comparison of problematic groups with randomly selected groups
- Same correlation behavior for whole array and randomly selected groups
- Higher positive correlation for problematic target sequence groups compared to control groups
  $\Rightarrow$ can lead to spurious correlation
Cross-Hybridization II

- Ratio of positive to negative correlation values was about 1 for random groups.
- For probe sets with repeat coverage the ratio increases with length of overlap (trend not visible for same size subgroup of normal probe-sets (Tag 1)).
Comparison U113A&B

<table>
<thead>
<tr>
<th></th>
<th># Probesets</th>
<th># Correct probesets (passed QC)</th>
<th>% of correct probesets (passed QC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A and B</td>
<td>100</td>
<td>98</td>
<td>98.0</td>
</tr>
<tr>
<td>Service probesets</td>
<td>68</td>
<td>N.A.</td>
<td>N.A.</td>
</tr>
<tr>
<td>Array U133A</td>
<td>22115</td>
<td>19753</td>
<td>89.3</td>
</tr>
<tr>
<td>Array U133B</td>
<td>22477</td>
<td>18660</td>
<td>83.0</td>
</tr>
</tbody>
</table>

- Passed QC means: tag1, correct orientation on chromosome and repeat coverage is less than 40%

- **U113A:**
  - more probes with annotation quality
  - higher signal intensity and lower noise
• lung cancer cell lines:
  - higher expr. values on A
  - more specific expr. on A
  - more noise on B
  - slightly higher expr. values for QC probesets
• 5-aza treated samples (11a.) vs. untreated control samples (10a.): (5-aza: higher methylation and expr. increase for many genes)
  - techn. differences larger than biological differences
  - effect of QC filtering similar to biological variation
SUMMARY I

• Careful analysis reveals probe sequence problems: non-human, multi-locus, misoriented, non-specific sequences

• Only 86% well designed sequences should be used for data analysis

• U113A has higher number of correct probesets and performs better - related to expression level and noise
SUMMARY II

- Unreliable sequences lower average expression level and add noise
- False correlation for probesets with higher repeat coverage
- Prerequisite for microarray data analysis: Check sequences with most recent annotation files against current references
SUMMARY III

• Excellent research work presented in paper

• No of unreliable probe sets can and will increase with new releases of annotation files and progress in gene definition

• 4 G probes: with 4 or more guanine bases

• chimeric RNA - transcribed from 2 genes ...

=> stay alert when using microarrays