Shi_NBT_RA14862A_SuppNotes

**Supplementary Notes: README for Probe Mapping Files**

August 3, 2006

On March 8, 2006 the following query on the NCBI web site human[orgn] AND "biomol mrna"[Properties] AND "srcdb refseq"[Properties] NOT "srcdb refseq model"[Properties]returned a list of curated human mRNA's. Model accessions (XMs) were excluded and only 24,000 curated mRNAs were kept for the mapping. After some preliminary considerations, we felt that 157 NMs were left out from the RefSeq release so we added them by hand. These directories are available upon request.

The supplemental material for this MAQC publication contains four files used in RefSeq based probe mapping and data generation.

1) Shi_NBT_RA14862A_SuppTable2.txt (Supplementary Table 2)
The complete mapping table with perfectly matching, high quality, gene specific probes (and 80% of probes within a probeset for AFX).

This master mapping table lists all perfect matches on 23,971 NMs ordered by chromosome, GeneID and gi. In this table, there is usually more than one NM per gene (on average, 4 NMs per 3 genes) and more than one probe per NM. Multiple hits were "; " delimited. In few extraordinary cases, and as most of the mapping was done before and around Christmass 2005, extra bonus was provided for so called "magnificent" probes. Magnificent probes perfectly matched their targets more than once and that number is reported in the parentheses. For example, Agilent probe A_24_P169864 (10) is the most extreme example – it hits NM_001007542.1 on gene FLJ40453 exactly 10 times.

Note that we did not provide probes that are not gene-specific. However, a list of 529 probes ("grade 3") that were not included in the mapping table because we believed those probes would very likely cross-hybridize is available upon request.

2) Shi_NBT_RA14862A_SuppTable3.txt (Supplementary Table 3)
A detailed table of the most 3' probes of the aforementioned 23,971 NMs, along with some additional information like UTR lengths and distance from the 3' end.

3) Shi_NBT_RA14862A_SuppTable4.txt (Supplementary Table 4)
The master mapping table. "All" refers to 18,114 genes perfectly matched through at least one NM by at least one platform on the MAQC project.

The gene specific probes were chosen from the NMs_to_3_UTR_probes_23971 table according to two rules:

A) If there is a TAQ, all probes that hit an NM along with the particular TAQ get chosen.
B) If there is not a TAQ probe, we choose NM targeted by the majority of manufacturers. In case of ties, randomly pick a transcript as the gene representative.

As stated in the manuscript, this choice was not without pitfalls, but it provided the least common denominator for the cross-platform comparison.

Note that according to the NCBI annotation, there were 2 pairs of genes that had same name, yet different GeneID (SKIP - 51763 and 80309; PRG2 - 5553 and 79948) and two pairs of genes with the same GeneID (29072 - SETD2 and HYPB; 83857 - ARG99 and TMTC1).

4) Shi_NBT_RA14862A_SuppTable5.txt (Supplementary Table 5)
The most important 1-1 mapping table, where probes were mapped to a unique gene.

This file contains unique probes that (in case of AFX - 80% of probes in a probeset) perfectly match, a single NM per gene. The number "12,091" represents the number of genes probed by each of the big six oligo platforms (ABI, AFX, AG1, GEH, ILM and NCI). For the same genes, probes on the alternative platforms (TAQ, QGN, GEX and EPP) are provided.

This file was derived from the file one_to_one_all.txt

In all tables, rows are printed in ascending order according to:
1) chromosome,
2) gid,
3) gi

Columns in the master mapping table and one-to-one tables (zero offset array and excel column labels):

| | |
|---|---|
| 0 | A ... gene |
| 1 | B ... gid |
| 2 | C ... accession |
| 3 | D ... gi |
| 4 | E ... ABI |
| 5 | F ... AFX |
| 6 | G ... AGL |
| 7 | H ... GEH |
| 8 | I ... ILM |
| 9 | J ... NCI |
| 10 | K ... TAQ |
| 11 | L ... QGN |
| 12 | M ... GEX |
| 13 | N ... EPP |