# CodeLink GE Healthcare Analysis

Damir Herman, UAMS

Wednesday, January 5, 2010

I analyzed probes that Jennifer sent me on December 22, 2010. I mapped the probes with standalone blast to three different databases that should contain all sequences deposited in the public domain (that I know of)

1) RefSeq database, downloaded in October 2010. Used in mRNA-seq read mapping.

2) Nucleotide sequence database, with entries from GenBank, EMBL and DDBJ, downloaded from <ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nt.gz>

3) Patent nucleotide sequence database, from USPTO and EU/Japan Patent Agency via EMBL/DDBJ, downloaded from <ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/patnt.gz>

There are 53,485 probes on the CodeLink array. I allowed that probes have at most 2 mismatches and alignment length is at least 28 (out of 30) bp. That is pretty stringent, and we work under assumption that probes do not cross hybridize.

Here is the mapping summary; targets denote how many sequences probes match under the described criteria.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Database** | **Sequences** | **Nucleotides** | **Probes** | **Targets** |
| 1 | RefSeq | 30,037 | 99,232,963 | 20,039 | 26,758 |
| 2 | Nt | 14,470,088 | 35,009,545,628 | 50,482 | 412,389 |
| 3 | Patnt | 19,085,735 | 10,090,389,286 | 42,435 | 709,675 |

The mapping details are provided in taxID\_CodeLink\_map\_results.xls. The columns are:

ProbeID, RefSeq hits, nt hits, patnt hits, taxa of nt hits, taxa of patnt hits.

The good news is that with exception with few E. Coli probes (most likely controls), all probes are human specific and it is pretty unlikely that they would be picking RNA from bacteria.

I suggest you use this mapping table in data interpretation.

Now I can finish the mRNA-seq mapping.