

Supplementary material

nuID: A Novel Identifier for Oligos, Ideal for Oligonucleotide-based Microarrays

Pan Du¹, Warren A. Kibbe¹ and Simon M. Lin^{1*}

¹Robert H. Lurie Comprehensive Cancer Center, Northwestern University, Chicago, IL, 60611, USA

Gene Identifier v.s. Oligonucleotide Identifier

Oligonucleotides (often simply referred to as oligos or oligomers), which are typically between 25 to 75 bases long, have been extensively used by DNA microarray manufacturers to detect the expression of different genes. An oligonucleotide sequence that is located on a specific bead or position on a microarray is usually referred to as the probe sequence, or just as the probe. Current microarrays by Affymetrix, Agilent, Illumina and others have tens of thousands or even hundreds of thousands of unique sequence probes in an array. The identification of the genes that hybridize to probe is a critical step in interpreting the results of microarrays.

Reporting results of the microarray at the probe-level is preferred because it requires less inference and enables researchers to reanalyze results with the latest mappings of probes to annotation resources such as RefSeq (Pruitt et al., 2005). Moreover, the detailed probe information facilitates aggregating the results across microarray platforms (Mecham, et al., 2004). For these reasons major commercial vendors now release the exact sequence of each of their probes.

In comparison with the universal identifiers for genes that are available through GenBank/EMBL/DDBJ, a universal oligonucleotide identifier has not been established. This has resulted in some confusion that oligonucleotide probes are uniquely identifying genes. Although a good probe sequence should uniquely hybridize with only the RNA from a single gene, this is very difficult to achieve in practice across a genome due to the presence of gene families, conserved domains, and other sequences that, if included in the probe, result in cross hybridization with RNA species from different genes. To make these distinctions clearer and enable better comparisons across microarrays, we have devised the nuID identifier.

Examples of Blackbox and Whitebox Identifiers

In general, there are blackbox or whitebox IDs. A blackbox ID does not reveal information about the object being tracked, and thus the ID is meaningless when outside of the context of a data source. To solve this problem, the Life Science Identifier (LSID) was devised (Clark, et al., 2004) to concatenate an identifier together with its database context using the syntax of urn:lsid:<authority>:<lsid_namespace>:<identifier>:<version>. In many cases, the resultant long string is cumbersome to use and

for oligonucleotides does not allow us to easily identify degenerate names (the subset of names that are pointing to a single, identical sequence). An alternative solution is to use an interpretable identifier (whitebox) based on properties of the object, such as a gene symbol. However, gene symbols can be ambiguous; for instance, the symbol NAP1 has been used to identify at least five different genes (Weeber, et al., 2003). We were seeking a solution to a simpler problem, that of finding a common identification and representation mechanism for oligomers across manufacturers.

Blackbox ID:

As an example, 3576 is used to represent the interleukin 8 gene in the Entrez database.

However, the same identifier, 3576, represents a completely different gene in the GeneFarm database.

Whitebox ID:

For example, the gene symbol IL8 can be used in the previous example to identify interleukin 8 across different databases. IL8 is not enough, however to track versioning information, nor other associated information typically stored in biological databases. As a further limitation, gene symbols can be ambiguous; for instance, the symbol NAP1 has been used to identify at least five different genes.

Examples of an LSID

For example, the Illumina 50-mer probe “ri|E030045A12| PX00206L14| AK053222| 1725-S” on the Mouse-6_v1 chip, is represented by the LSID as a 73-character string of “urn:lsid:illumina.com:Mouse-6_v1:ri|E030045A12| PX00206L14| AK053222| 1725-S”. In many cases, the resultant long string is cumbersome to use and for oligonucleotides does not allow us to easily identify degenerate names (the subset of names that are pointing to a single, identical sequence).

Examples of nuIDs

nuID can be applied to index all DNA microarrays based on oligonucleotides. We list some examples of both Affymetrix and Illumina in Table 1.

The practical application of nuIDs to an annotation pipeline is quite straightforward, and given the proliferation of gigabit/second interconnects and the efficiency of modern computer architectures the overhead cost of using nuIDs as the primary identifier for a sequence versus an 8 or 10 digit number is trivial.

Table 1 . Examples of nuIDs.

Array Type	Manufacturer's Proprietary Identifier	Nucleotide Sequence	nuID
Affymetrix Human	206064_s_at_probe1	TGTATATGTCTGGTTTTCTTACCCC	a7M7ev98VQ
Illumina Human	GI_23097300-A	GCTTCACTCGCTTCCCAGGGGCTCCG TTCACCAACTACATGAGCTACACG	cn0dn1Sqdb0UHE4nEY
Illumina Mouse	TRBV23_AE000664_T_cell_receptor_beta_variable_23_106-S	GACCCTTCGAAAGTGAAAGAACACAG TCATGTTATATGGTATAGTCATGGT	9hX2C4CBEtO8zrMtOs

nuID2 to Increase the Error Detection Power

The error detection power of nuID ($N=21$) is about 95.2%, which should satisfy most applications of nuID. If an application has a more stringent error checking requirement, one more digit of checking code can be added. We have named this identifier nuID2 ($N=1344$) to distinguish it from our standard encoding scheme. With nuID2, the error detection power reaches 99.93%.

Using nuID as a Universal ID to Annotate the Illumina Microarray

Constructing and managing identifiers on microarrays is a challenging problem. The design of the Illumina Mouse microarray, for example, agglomerated 13 sequence database sources by taking identifiers from each of those sources and appending them with design notes. However, the “TargetID” used to identify the probe in reporting hybridization results is not consistent between different versions of chips.

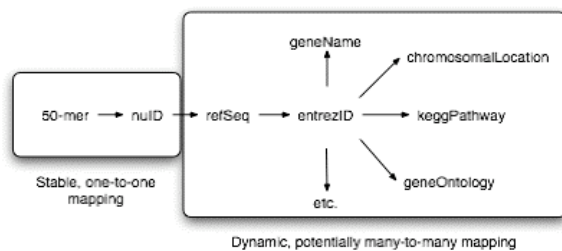


Figure 2 Annotate the Illumina microarray using nuID.

All the Illumina microarrays use 50-mers. Thus, it is possible to have a single annotation database for all different versions of the microarrays. However, we observed that the target-identifier-to-50-mer mapping is not unique across different versions of Illumina microarray. We also observed similar instances within the same chip. We have used the nuID universal identifier to resolve this issue, and it provides the additional benefits of being stable over time (since it is only dependent on the sequence) and incorporates error

checking. We have used the nuID to refSeq mapping provided by the manufacturer and the AnnBuilder package in Bioconductor to annotate the probes, as shown in Figure 2. By incorporating the nuIDs naming scheme for the 50-mers, we are able to build one annotation database for three different versions of the human chips, and for two different versions of the mouse chips.

Possible Future Extensions

By incorporating the nuID naming scheme into the probe annotation workflow, build a generic annotation pipeline that is independent of manufacturer and makes the maintenance of annotations independent of the manufacturer.

Another further application, that has not escaped our notice, is that the encoded string allows us to quickly and easily identify whether the oligos are identical or frameshifted from each other using standard bitstring comparison routines.

References

Clark, T., Martin, S. and Liefeld, T. (2004) Globally distributed object identification for biological knowledgebases, *Brief Bioinform*, 5, 59-70.

Weeber, M., Schijvenaars, B.J., Van Mulligen, E.M., Mons, B., Jelier, R., Van Der Eijk, C.C. and Kors, J.A. (2003) Ambiguity of human gene symbols in LocusLink and MEDLINE: creating an inventory and a disambiguation test collection, *AMIA Annu Symp Proc*, 704-708.

Zeeberg, B.R., Riss, J., Kane, D.W., Bussey, K.J., Uchio, E., Linehan, W.M., Barrett, J.C. and Weinstein, J.N. (2004) Mistaken identifiers: gene name errors can be introduced inadvertently when using Excel in bioinformatics, *BMC Bioinformatics*, 5, 80.

NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Pruitt KD, Tatusova, T, Maglott DR
Nucleic Acids Res 2005 Jan 1;33(1):D501-D504