doi:10.1093/nar/gkp969

PLANdbAffy: probe-level annotation database for Affymetrix expression microarrays

Ramil N. Nurtdinov[1,*], Mikhail O. Vasiliev[2,3], Anna S. Ershova[3,4], Ilia S. Lossev[5] and Anna S. Karyagina[3,4,6]

[1]Departament of Bioengineering and Bioinformatics, M.V. Lomonosov Moscow State University, Vorbyevy Gory 1-73, Moscow 119992, Russia, [2]Moscow Institute of Physics and Technology, Institutskii per. 9, Dolgoprudny, Moscow Region 141700, [3]Gamaleya Institute of Epidemiology and Microbiology Russian Academy of Medical Sciences, Gamaleya Street, 18, Moscow 123098, [4]Institute of Agricultural Biotechnology, Russian Academy of Agricultural Sciences, Timiryazevskay Street 42, Moscow 127550, Russia, [5]Parascript LLC, Winchester Circle 6899, Ste. 200, Boulder, CO 80301, USA and [6]A.N. Belozersky Institute of Physical and Chemical Biology, Moscow State University, Leninskie Gory 1-40, Moscow 119991, Russia

Received August 14, 2009; Revised October 12, 2009; Accepted October 14, 2009

ABSTRACT

Standard Affymetrix technology evaluates gene expression by measuring the intensity of mRNA hybridization with a panel of the 25-mer oligonucleotide probes, and summarizing the probe signal intensities by a robust average method. However, in many cases, signal intensity of the probe does not correlate with gene expression. This could be due to the hybridization of the probe to a transcript of another gene, mapping of the probe to an intron, alternative splicing, single nucleotide polymorphisms and other reasons. We have developed a database, PLANdbAffy (available at http://affymetrix2.bioinf.fbb.msu.ru), that contains the results of the alignment of probe sequences from five Affymetrix expression microarrays to the human genome. We have determined the probes matching the transcript-coding regions in the correct orientation. For each such probe alignment region, we determined the mRNA and EST sequences that contain the probe sequence. In the textual part of the database interface we summarize the data on the sequences that cover the probe alignment region and SNPs that are located inside it. The graphical part of our database interface is implemented as custom tracks to the UCSC genome browser that allows one to utilize all the data that are offered by UCSC browser.

INTRODUCTION

Affymetrix 3′ Gene as well as Exon and Gene level microarrays are widely used in gene expression studies. HG-U133A, HG-U133B and HG-U133 Plus 2.0 arrays consist of probe sets developed for each annotated human gene. A probe set is typically a set of 11 25-mer oligonucleotide probes, with a small number of probe sets consisting of more or less than 11 probes.

The majority of Human Exon 1.0 probe sets consist of four probes; these probes are developed to target all known and predicted human exons. The Human Gene 1.0 chip is based on Human Exon 1.0 data combining together all highly expressed probes that are confirmed by the transcriptome data for a particular gene. Thus the number of probes in a probe set depends on the transcript length.

The Affymetrix probes and probe sets remained unchanged for the past several years but our knowledge of their genome and transcriptome context has improved with every paper in this field.

The first annotation of Affymetrix probes was provided by Affymetrix staff in NetAffx database (1). This database contains information about transcripts that are recognized by the corresponding probe sets and fixes the problem of the absence of representative sequences in new versions of UniGene (2).

A careful analysis of HG-U133A probes was done by Gautier and colleagues (3). The authors aligned probe sequences with RefSeq (4) mRNAs, and found some discrepancies for 64% of the HG-U133A probes.

Using a similar approach, Harbig and colleagues (5) showed that ∼37% of the probes of the HG-U133 Plus

[*]To whom correspondence should be addressed. Tel: +7 495 939 14 59; Fax: +7 495 939 31 58; Email: ramil@bioinf.fbb.msu.ru
Present address:
Ramil N. Nurtdinov, INSERM U563, Pole des Neurosciences, University of Toulouse, BP 3048, CHU Purpan, Toulouse 3, 31024, France.

2.0 array should be redefined and more than 5000 probe sets detect multiple transcripts. Similar analyses for different expression arrays (6–10) brought similar results.

Non-specific hybridization is another big problem of microarray experiments. Several papers showed that the rule that a perfect match probe has a high signal level and a mismatch probe has a low signal level does not work in many cases (11–13).

In a subsequent paper (14), Zhang and colleagues developed a model of molecular interaction on short oligonucleotide arrays and applied it in their next work (15). It was shown that a significant amount of probes could give high signal level by a non-specific hybridization with short 10–16-nucleotide fragments.

Alternative splicing is another source of the inconsistency in microarray experiments. Recent articles showed that up to 93% of human intron-containing genes undergo alternative splicing (16,17) and up to 90% of the genome sequence is transcribed (18). An additional source of the inconsistency is the presence of single nucleotide polymorphisms (SNPs) within probe alignment positions.

There are several publicly available databases that contain annotation of the Affymetrix data: the official NetAffx (1), GeneAnnot (19), ADAPT (20) and X:Map (21) databases.

GeneAnnot and ADAPT align probe sequences to the RefSeq and Ensembl mRNAs, NetAffx additionally considers GenBank (22) and UniGene (2) mRNAs. The main problem of the common approach used by these three databases arises when a particular probe, in addition to the original position, recognizes another transcribed region that is absent in the considering mRNA sequences. This results in the incomplete probe set (probe) annotation.

The X:Map and presented here PLANdbAffy databases fix the above shortcoming. The authors of X:Map have aligned probe sequences with the genome and also took into account the ESTs. Unfortunately, X:Map contains data only for exon-level arrays, leaving other widely used arrays (HG-U133A&B and Human Gene 1.0) uncovered.

The interface of X:Map is based on Google Maps API covering the whole chromosome. To obtain the EST transcription state of a particular probe one has to calculate the ESTs manually. This is rather difficult, and becomes much more laborious for the exon-junction probes and probes that are close to splicing sites. Also the X:Map database uses only the Ensembl genome annotation and Ensembl EST accessions, which brings difficulties to the NCBI-oriented users.

Our PLANdbAffy database considers five widely used Affymetrix human microarrays: HG-U133A, HG-U133B, HG-U133 Plus 2.0, Human Exon 1.0 and Human Gene 1.0. Database provides user with information on all alignment places of the individual Affymetrix probes with the genome considering alignments with up to two mismatches, and also support each probe alignment region with all known to-date transcriptome data. Unlike the above databases (except NetAffx), PLANdbAffy also contains data on SNPs. Graphical information about each probe alignment region and gene is implemented as custom tracks to UCSC genome browser. After moving to the UCSC site it becomes possible to utilize the whole set of data and tools provided by the UCSC browser.

DATABASE CONSTRUCTION AND STRUCTURE

Data source

The files containing information about Affymetrix microarrays were downloaded from the official Affymetrix site (http://www.affymetrix.com/products_services/index.affx). For this analysis we selected three 3′ Gene arrays, Affymetrix HG-U133A, HG-U133B and HG-U133 Plus 2.0, and two Exon&Gene level arrays, Human Gene 1.0 and Human Exon 1.0.

The NCBI36 (hg18) genome assembly was download from UCSC ftp site. Also, we have downloaded EST and mRNA exon–intron structures (http://hgdownload.cse.ucsc.edu/goldenPath/hg18/database/ all_mrna.txt.gz and all_est.txt.gz files) that were obtained by Blat (23) alignment of the corresponding sequences with the genome. We used the NCBI annotation of the genome sequences. Refseq (4) and Unigene (2) were used to assign mRNA and EST sequences to the genes.

We used dbSNP (24) build 130 as a source of SNPs, the human readable text files were downloaded from the ftp site (ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606/ASN1_flat/) and parsed.

Database development

Each probe and probe set within five chips under consideration was assigned a unique ID. It was done because some probe sets in different chips have the same identification numbers. The Affimetrix numbers were also stored and could be used to search the database.

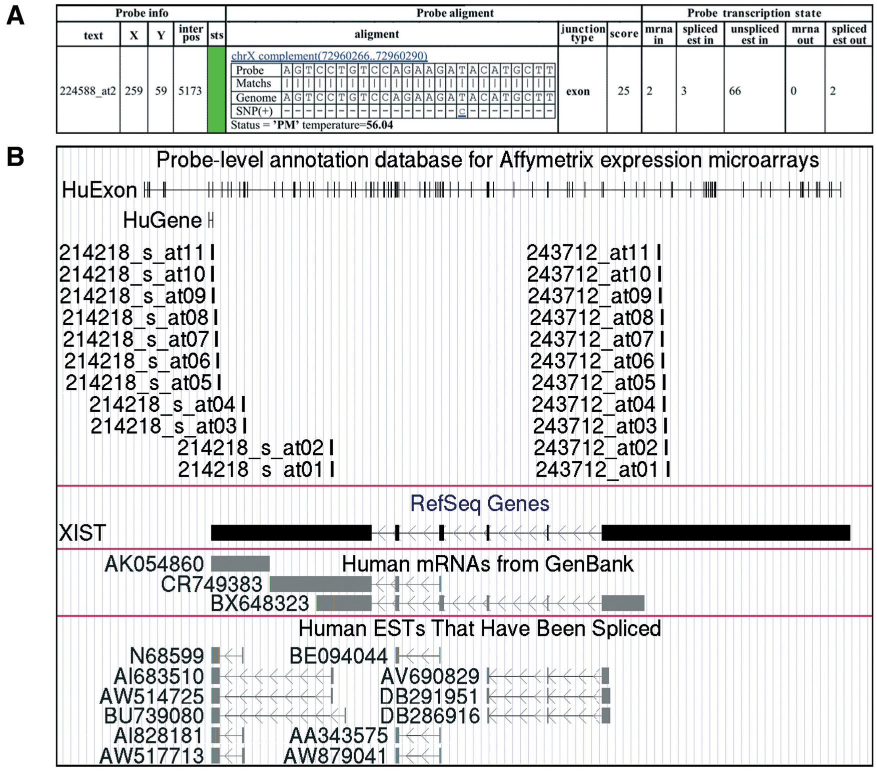
Probe sequences were mapped to the genome using Blat (23). We allowed alignments with no more than two mismatches and required 40- and more nucleotide introns for potential exon-junction probes. The hits found ('probe alignment regions') were stored and subjected to further analysis.

We assigned a probe to a particular gene ('the probe match the gene') if the probe alignment region intersected with the annotated gene region and was in the correct orientation. We also took into account possible mistakes in the gene annotation extending the 3′-end of each gene by 1000 nucleotides.

We annotated each probe alignment region using the mRNA and EST alignments provided by UCSC, considering only the sequences that were present in UniGene (219 build) for corresponding genes.

For each probe alignment region, we have calculated the number of mRNA and EST that either support (mrna_in, spliced_est_in, unspliced_est_in fields) or do not support (mrna_out, spliced_est_out fields) occurrence of the probe alignment region in an exon (see the database web site for further explanation).

To present the quality of a probe we divided all probes into four classes, and assigned a color to each class (Figure 1). Green probes (the best ones) are the probes meeting three conditions. First, the probe is



Probe info					Probe aligment			Probe transcription state				
text	X	Y	inter pos	sts	aligment	junction type	score	mrna in	spliced est in	unspliced est in	mrna out	spliced est out
224588_at2	259	59	5173	🟩	chrX complement(72960266..72960290) Probe A G T C C T G T C C A G A A G A T A C A T G C T T Matchs | | | | | | | | | | | | | | | | | | | | | | | | | Genome A G T C C T G T C C A G A A G A T A C A T G C T T SNP(+) - - - - - - - - - - - - - - - - c - - - - - - - - Status = 'PM' temperature=56.04	exon	25	2	3	66	0	2

Figure 1. Examples of textual (**A**) and graphical (**B**) interface of the PLANdbAffy database. The textual interface of the database consists of three sections. The first section (left five columns) contains the original information about the probes from Affymetrix and probe status (highlighted by green), the second section (6–8th columns) describes the probe alignment and the last section (rightmost five columns) describes the numbers of the ESTs and mRNAs either supporting or not supporting the occurrence of the probe in an exon (see 'Database development' section). An example of graphical interface was manually processed to reduce the image size.

aligned to the target gene without mismatches. Second, there are no matches of the probe to other genes. Third, there are no perfect alignments of the probe to non-coding regions. Unlike the green probes, a yellow probe has a perfect match to uncoding region. The yellow probes still have a perfect match to the target gene and no matches with other genes. The red probes are the probes that have a perfect match to the target gene and at least one alignment to other genes with no more than one mismatch. Finally, a black probe is aligned to the target gene with at least one mismatch.

Human-readable text files of dbSNP contain information about the SNP genome position, orientation and allelic variants. We have selected SNPs that were located in probe alignment regions and mapped them onto the probe sequences.

Database interface

The database is available at http://affymetrix2.bioinf.fbb.msu.ru. Text files containing the information about the mapping and annotation of the good (green) probes can be downloaded from web site.

The title page contains several search boxes. One may either search the database with an Affymetrix probe set identifier, or get all probes for the particular gene using the gene search boxes. The EntrezGene, HUGO and Ensembl identifiers, the gene symbol, a word or a phrase in the gene name can be used. It must be noted, however, that since our database is based on the RefSeq annotations, some of the Ensembl and HUGO identifiers could be missed.

Querying a probe set or a gene one could see the textual part of the database interface (Figure 1A). The textual part of the interface consists of probe information section, probe alignment section and transcription state section. The probe information section has four fields presenting the probe name ('text'), the probe position on the chip ('X', 'Y', 'inter pos') and the color representation of the probe's quality status ('sts'), see Database development section.

The probe alignment section contains information about the probe alignment and its mismatches. Positions of SNPs within the probe alignment region are marked and supported by the links to their descriptions in dbSNP.

For each probe, the information about EST and mRNA sequences that cover the probe alignment region is available at the transcription state section. The explanation of the corresponding fields for the exon and exon-junction probes is given in the 'Database development' section.

Each gene and each probe alignment region are supported by the graphical part of the database interface. It is organized as custom tracks to UCSC genome browser (Figure 1B), that allows one to utilize all information that is offered by UCSC browser for the corresponding region.

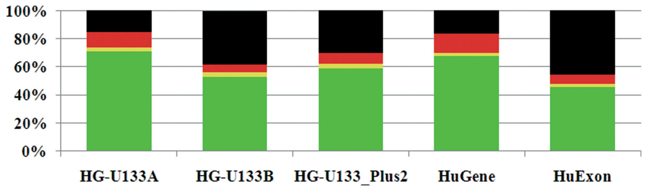


Figure 2. Frequencies of probes with different alignment and cross-hybridization state for all five considered Affymetrix arrays. See colours' definitions in the text.

Table 1. Genome and transcriptome annotation for good (green) probes

	HG-U133A	HG-U133B	HG-U133_Plus2	HuGene	HuEx
Exon	142 427	73 148	236 427	436 548	915 886
Exon/intron	19 567	20 198	50 154	81 189	288 217
Intron	12 374	38 680	70 605	67 619	1 271 110
SNP	19 249	11 380	34 724	70 525	274 208
Total genes	16 480	11 697	23 255	24 010	30 915

Data analysis

In the Figure 2, we present frequencies of each type of probes for all five arrays. Among the 3′ gene arrays HG-U133A has the highest frequency (70%) of good (green) probes. HG-U133B array has ~53% of good probes and HG-U133 Plus 2.0 array that was designed basically by combining the HG-U133A and HG-U133B arrays data is located in between and has 59% of good probes.

Table 1 contains summary information about the transcriptome annotation for good (green) probes. Probe is marked as 'exon' if it is confirmed by more than 90% of mRNA and EST sequences that cover this region. It is marked as 'intron' if it is confirmed by <10% of the sequences, whereas the probes that are in between are marked as 'exon/intron' ones.

The HG-U133A array contains the lowest amount of intron and exon/intron probes (18.3%). Considerably greater amount of such probes was observed for HG-U133B (44.6%) and HG-U133 Plus 2.0 (33.8%) arrays.

As Human Exon 1.0 chip was designed to recognize all potential transcribed segments, it contains the greatest amount of the intron and exon/intron probes (63.0%). The Human Gene 1.0 array has a similar to the HG-U133A array level of the intron and exon/intron probes (25.4%). All five arrays have almost an equal amount of SNPs (8–12%) in probe align region of good (green) probes.

Similar results were described in different research and database papers. Zhang and colleagues (7) have shown that HG-U133A array contains 12.1 and 8.0% of non-specific and mistargeted probes, respectively. GeneAnnot database summary (19) reports that ~16% of HG-U133A array probe sets recognize multiple genes. ADAPT database summary (20) reports ~23.1% of HG-U133 Plus 2.0 array probe sets, which match more than one RefSeq transcript.

X:Map database publication (21) contains detailed statistics for Human Exon 1.0 chip. The authors observed 9% of multitarget probe sets and 45% of intergenic probe sets. Very similar values were observed in PLANdbAffy database: 9.1% of multitarget (red and yellow) probes and 45.2% of intergenic (black) probes. X:Map annotates 21 and 23% of all studied probe sets as exon and intron ones respectively, and the similar values is observed in PLANdbAffy (Table 1).

DATABASE USAGE

The database can be used for interpretation of results of gene expression experiments, and also to perform the delicate analysis of expression in certain areas of genome. For example, it is a common situation that different probe sets of one gene demonstrate quite different expression values and it is not clear what is true. Careful analysis of the genomic probe alignment regions can help to explain the difference. It may appear due to some discrepancies in microarray design, the probe can be aligned into the spliced region of a gene, existence of SNPs in probe align regions may cause the decrease of probe signal intensity. In contrast, much more often observed cross-hybridization of a probe will increase the probe signal.

PLANdbAffy textual summary page of particular probe set or gene contains the information on transcription, cross-hybridization and SNP status for each probe. From this page one can move to UCSC Genome Browser and see the considered Affymetrix probes as a custom track. This browser contains different annotations for corresponding genome regions, e.g. mapping and sequencing annotation, phenotype and disease annotation, gene, protein, mRNA and EST annotation, etc. This information allows one to perform a qualitative analysis of microarray results and may suit as a good starting point for additional molecular studies.

FUTURE PLANS

We are planning to move our data from hg18 to hg19 version of human genome and update it twice a year by the new mRNA and EST alignments. We also are planning to perform this analysis for the mouse and rat exon-level arrays.

ACKNOWLEDGEMENTS

The authors thank Alexander Favorov and Alexander Poliakov for valuable technical advises and Mikhail Roytberg for helpful discussion.

FUNDING

Open Access charges were waived by Oxford University Press.

Conflict of interest statement. None declared.

REFERENCES

1. Liu,G., Loraine,A.E., Shigeta,R., Cline,M., Cheng,J., Valmeekam,V., Sun,S., Kulp,D. and Siani-Rose,M.A. (2003) NetAffx: Affymetrix probesets and annotations. *Nucleic Acids Res.*, **31**, 82–86.
2. Sayers,E.W., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S. *et al.* (2009) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **37**, D5–D15.
3. Gautier,L., Møller,M., Friis-Hansen,L. and Knudsen,S. (2004) Alternative mapping of probes to genes for Affymetrix chips. *BMC Bioinformatics*, **5**, 111.
4. Pruitt,K.D., Tatusova,T., Klimke,W. and Maglott,D.R. (2009) NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res.*, **37**, D32–D36.
5. Harbig,J., Sprinkle,R. and Enkemann,S.A. (2005) A sequence-based identification of the genes detected by probesets on the Affymetrix U133 plus 2.0 array. *Nucleic Acids Res.*, **33**, e31.
6. Dai,M., Wang,P., Boyd,A.D., Kostov,G., Athey,B., Jones,E.G., Bunney,W.E., Myers,R.M., Speed,T.P., Akil,H. *et al.* (2005) Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res.*, **33**, e175.
7. Zhang,J., Finney,R.P., Clifford,R.J., Derr,L.K. and Buetow,K.H. (2005) Detecting false expression signals in high-density oligonucleotide arrays by an in silico approach. *Genomics*, **85**, 297–308.
8. Okoniewski,M.J. and Miller,C.J. (2006) Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. *BMC Bioinformatics*, **7**, 276.
9. Yu,H., Wang,F., Tu,K., Xie,L., Li,Y.Y. and Li,Y.X. (2007) Transcript-level annotation of Affymetrix probesets improves the interpretation of gene expression data. *BMC Bioinformatics*, **8**, 194.
10. Orlov,Y.L., Zhou,J., Lipovich,L., Shahab,A. and Kuznetsov,V.A. (2007) Quality assessment of the Affymetrix U133A&B probesets by target sequence mapping and expression data analysis. *In Silico Biol.*, **7**, 241–260.
11. Lemon,W.J., Palatini,J.J., Krahe,R. and Wright,F.A. (2002) Theoretical and experimental comparisons of gene expression indexes for oligonucleotide arrays. *Bioinformatics*, **18**, 1470–1476.
12. Zhou,Y. and Abagyan,R. (2002) Match-only integral distribution (MOID) algorithm for high-density oligonucleotide array analysis. *BMC Bioinformatics*, **3**, 3.
13. Irizarry,R.A., Hobbs,B., Collin,F., Beazer-Barclay,Y.D., Antonellis,K.J., Scherf,U. and Speed,T.P. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.
14. Zhang,L., Miles,M.F. and Aldape,K.D. (2003) A model of molecular interactions on short oligonucleotide microarrays. *Nat Biotechnol.*, **21**, 818–821.
15. Wu,C., Carta,R. and Zhang,L. (2005) Sequence dependence of cross-hybridization on short oligo microarrays. *Nucleic Acids Res.*, **33**, e84.
16. Wang,E.T., Sandberg,R., Luo,S., Khrebtukova,I., Zhang,L., Mayr,C., Kingsmore,S.F., Schroth,G.P. and Burge,C.B. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470–476.
17. Pan,Q., Shai,O., Lee,L.J., Frey,B.J. and Blencowe,B.J. (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet.*, **40**, 1413–1415.
18. Johnson,J.M., Edwards,S., Shoemaker,D. and Schadt,E.E. (2005) Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *Trends Genet.*, **21**, 93–102.
19. Chalifa-Caspi,V., Yanai,I., Ophir,R., Rosen,N., Shmoish,M., Benjamin-Rodrig,H., Shklar,M., Stein,T.I., Shmueli,O., Safran,M. *et al.* (2004) GeneAnnot: comprehensive two-way linking between oligonucleotide array probesets and GeneCards genes. *Bioinformatics*, **20**, 1457–1458.
20. Leong,H.S., Yates,T., Wilson,C. and Miller,C.J. (2005) ADAPT: a database of affymetrix probesets and transcripts. *Bioinformatics.*, **21**, 2552–2553.
21. Yates,T., Okoniewski,M.J. and Miller,C.J. (2008) X:Map: annotation and visualization of genome structure for Affymetrix exon array analysis. *Nucleic Acids Res.*, **36**, D780–D786.
22. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2009) GenBank. *Nucleic Acids Res.*, **37**, D26–D31.
23. Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
24. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.