

# **Sleuthing With the Affymetrix NetAffx™ Website**

Identifying and Examining Probe Sets and Their Genomic Context

# Table of Contents

<b>SLEUTHING WITH THE AFFYMETRIX NETAFFX™ WEBSITE.....</b>	<b>1</b>
<b>TABLE OF CONTENTS .....</b>	<b>2</b>
<b>INTRODUCTION .....</b>	<b>3</b>
<b>RESOURCES.....</b>	<b>4</b>
NETAFFX™ PROBE SET ANNOTATION - FULL RECORD PAGE .....	4
NETAFFX™ TOOLS .....	8
<i>Probe Match Tool</i> .....	8
<i>BLAST Tool</i> .....	9
<i>Probe Set Display Tool</i> .....	9
PUBLIC DOMAIN BIOINFORMATICS DATABASES.....	10
<i>Entrez Gene and UniGene Annotations</i> .....	10
<i>Ensembl</i> .....	12
PUBLIC DOMAIN BIOINFORMATICS TOOLS .....	12
<i>NCBI Blast Suite</i> .....	12
<i>UCSC BLAT Tool</i> .....	14
<i>Integrated Genome Browser (IGB)</i> .....	15
<b>CASE STUDIES .....</b>	<b>17</b>
CASE 1: RESOLVING DMD GENE ISOFORMS.....	17
<i>Customer Question</i> .....	17
CASE 2: ERRONEOUS GENBANK MRNA SEQUENCE FOR RAP1A.....	22
<i>Customer Question</i> .....	22
<b>REFERENCES .....</b>	<b>26</b>

## Introduction

Identification of transcripts measured by probe sets is confounded by two main factors:

- The transcriptome's intricate network of multiple isoforms and overlapping sense and antisense transcripts (1).
- The incomplete, sometimes erroneous, and constantly evolving mRNA sequence record in the public domain.

For this reason, investigating the transcripts detected by individual probe sets often requires a broad and detailed knowledge of various bioinformatics tools and databases.

This document describes how Affymetrix NetAffx™ annotations can be used with other bioinformatics tools and databases to better understand biological functions of transcripts. It also helps the user to interpret the current, but often incomplete, understanding of a locus. The first half of this document outlines the NetAffx Full Record Page and the bioinformatics tools that are used to interpret, verify, or extend the NetAffx annotations. The second half describes case study examples of unusual probe sets and how they can be interpreted using the described resources.

# Resources

## NetAffx™ Probe Set Annotation - Full Record Page

The **Full Record** page for each probe set (NetAffx → Query → Search → Details → Full Record) provides a comprehensive collection of the latest annotations and resources. This page is usually derived through the results table of a search and has seven sections as shown in Figure 1.

**Probe Set and Array name**

Probe Set ID: 1415707\_at  
 GeneChip Array: Mouse Genome 430 2.0 Array  
 Organism: Mouse  
 Name: Mouse

**Evidence supporting Probe Set Design**

Transcript ID: Mm.202841.1  
 Probe Design Information: Consensus sequence  
 Representative Probe ID: Mm.202841.1[CG]  
 Target: chr2:2520483-2521807[CG]\_UCSC  
 Description: This cluster is supported by a rRNA and contains 216 sequence(s)  
 Probe Selection Criteria: This Probe Selection Region is supported by a EST Stack and contains 95 sequence(s)

**Latest evidence measured by the probe-set**

Annotation Description: This probe set was annotated using the Matching Probe based pipeline to a Ensembl Gene identifier using 1 transcript(s).  
 Annotation Grade: This is a grade A annotation.  
 Annotation Cluster ID: NM\_175300(1)

**Consensus alignment with the Genome**

Assembly: May 2004 ENCB 3.0  
 Alignment: chr2:2520483-2521807[CG]\_UCSC  
 Position: chr2:2520483-2521807[CG]\_UCSC  
 Coverage: 98.68  
 Cytosine: 6A3  
 \* You can now view alignments using the [Genomic Browser](#). Note that you must [click](#) before clicking on any of the "GOF" links above.

**Functional Annotations**

Gene Title: anaphase promoting complex subunit 2  
 Gene Symbol: Anapc2  
 Chromosomal Location: 2A3  
 RefSeq ID: Mm.291624[CG]\_FULLLENGTH  
 RefSeq: ENEMUS00000029865 [Ensembl]  
 RefSeq Gene: 91917 [Ensembl]  
 SwissProt: Q8R2U7 [EMBL], Q8R2U7 [EMBL], Q8R2U7 [EMBL]  
 RefSeq Protein: Mm.175300.1 [EMBL]  
 RefSeq: Mm.175300.1 [EMBL]

**Probe Design and Genome References**

ID	Title	Organism
AF011315.1 [EMBL]	E1 ubiquitin ligase, putative	af
U000001.1 [EMBL]	mon1a	dm
DR000001.1 [EMBL]	mon1a	dm
H01131.1 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.2 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.3 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.4 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.5 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.6 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.7 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.8 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.9 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.10 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.11 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.12 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.13 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.14 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.15 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.16 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.17 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.18 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.19 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.20 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.21 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.22 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.23 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.24 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.25 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.26 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.27 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.28 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.29 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.30 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.31 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.32 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.33 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.34 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.35 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.36 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.37 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.38 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.39 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.40 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.41 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.42 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.43 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.44 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.45 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.46 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.47 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.48 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.49 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.50 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.51 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.52 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.53 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.54 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.55 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.56 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.57 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.58 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.59 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.60 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.61 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.62 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.63 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.64 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.65 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.66 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.67 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.68 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.69 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.70 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.71 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.72 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.73 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.74 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.75 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.76 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.77 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.78 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.79 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.80 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.81 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.82 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.83 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.84 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.85 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.86 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.87 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.88 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.89 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.90 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.91 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.92 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.93 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.94 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.95 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.96 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.97 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.98 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.99 [EMBL]	anaphase promoting complex subunit 2	hs
H01131.100 [EMBL]	anaphase promoting complex subunit 2	hs

**Design Sequences**

BLAST Search: [BLAST Search](#)

Probe Sequence (5' to 3')	Probe X	Probe Y	Probe Intensity	Disambiguation
GAGTACTCTCTCTGGGATACA	612	537	2351	Arbansense
GGGATACATACAGGGCATTTAC	48	881	2367	Arbansense
TUTTAACCAACTAGAGAGCCCTC	885	793	2385	Arbansense
AGAGCCCTCTCTGGAGGATTA	878	125	2400	Arbansense
GTACTACATCATCTCTCTCTTC	158	819	2418	Arbansense
ATGTTTAAAGTCTGGCCCTGCG	530	81	2438	Arbansense
TCATCTATTCTGAGGTCCTACCG	882	887	2528	Arbansense
AGATTCCATCATCTCTCTCTTC	878	137	2558	Arbansense
GACTACAGATTCAGCTCTCTGAA	601	521	2592	Arbansense
TGTGCAACCCCTAGATGACAGCTT	772	783	2619	Arbansense
GGAGAGCTCTATTTCATCTCTCT	798	721	2626	Arbansense

Figure 1. NetAffx™ probe set annotation Full Record page

The **GeneChip Array Information** (Figure 2) marks the probe set name and the array it belongs to.

GeneChip Array Information	
Probe Set ID	1007_s_at
GeneChip Array	Human Genome U133 Plus 2.0 Array
Organism	Human
Common Name	

**Figure 2. GeneChip Array Information.**

The **Probe Design Information** section (Figure 3) displays the annotations generated at the time the array was designed. These annotations summarize the quality of the sequences that were used to design the corresponding probe set along with other details such as the exemplar/consensus design information. This information is useful when no additional information is displayed in the main record. Since this information is collected only for the original design date, it is important to remember that it can also be outdated, and the most current information about the probe set is found elsewhere on the page.

Probe Design Information	
Transcript ID(Array Design)	Hs2.324473.2
Sequence Type	Consensus sequence
Representative Public ID	NM_138957 <a href="#">NCBI</a>
Archival UniGene Cluster	Hs.324473 <a href="#">NCBI</a>
Target Description	gb:NM_138957.1 /DB_XREF=gi:20986530 /GEN=MAPK1 /TID=Hs2.324473.2 /CNT=74 /FEA=FLmRNA /TIER=FL+Stack /STK=16 /LL=5594 /UG=Hs.324473 /DEF=Homo sapiens mitogen-activated protein kinase 1 (MAPK1), transcript variant 2, mRNA. /PROD=mitogen-activated protein kinase 1 /FL=gb:NM_138957.1 gb:BC017832.1
Cluster Evidence	This cluster is supported by a Full-length mRNA and contains 74 sequence(s).
Probe Selection Region Evidence	This Probe Selection Region is supported by a Full-length mRNA and EST Stack and contains 16 sequence(s).

**Figure 3. Probe Design Information. The Cluster Evidence and Probe Selection Region Evidence fields describe the quality of the transcript sequence cluster that was used to design the probe set. The Probe Selection Region in this example is well supported both by a full length mRNA and a stack of EST sequences.**

The Cluster Evidence and Probe Selection Region Evidence fields (Figure 3) under Probe Design Information provide details of the quality of the subcluster sequences and the nature of the evidence supporting the probe set. The Cluster Evidence field indicates the best sequence evidence in the subcluster. Probe Selection Region Evidence indicates the sequence evidence supporting the precise region that was tiled on the array, and whether or not the target region is supported by mRNA alone, or mRNA/EST stack, or just EST stack.

The **Annotation Method Description** section provides the current known transcripts associated with the probe set and summarizes the evidence for that transcript assignment (Figure 4). The **Annotation Description** field in particular describes the transcript assignment method and the source that was used to gather gene level annotations for the assigned transcript. The **Annotation Notes** field lists the transcripts that potentially cross-hybridize with the probe set.

Annotation Method Description				
<b>Annotation Description</b>	This probe set was annotated using the Matching Probes based pipeline to a Entrez Gene identifier using 2 transcript(s).			
<b>Annotation Grade</b>	This is a grade A annotation.			
<b>Annotation Transcript Cluster (# of Matching Probes)</b>	NM_002931(16), Z14000(16)			
<b>Transcript Assignments</b>	<b>Representative Transcript</b>	<b>Description</b>	<b>Matching Probes</b>	<b>Related Probesets by Grade</b>
	Z14000_NCBJ	H.sapiens RING1 gene.	16/16	<u>A</u>
	NM_002931_NCBJ	Homo sapiens ring finger protein 1 (RING1), mRNA.	16/16	<u>A</u>
	ENST00000333656	cdna:known chromosome:NCBI35:6:33284269:33288468:1 gene:ENSG00000112477	16/16	<u>A</u>
ENSESTT00000096503		16/16	<u>A</u>	
<b>Annotation Notes</b>	<b>Cross Hybridizing mRNA</b>	<b>Matched Probes</b>	<b>Nature of Assignment</b>	<b>Related Probesets by Grade</b>
	GENSCAN00000060456	7/16	Cross Hyb Matching Probes	<u>A</u> <u>B</u>

**Figure 4. Annotation Method Description sample transcript. The transcript grade and the evidence underlying the grade assignment are provided. Cross-hybridization data is provided in the Annotation Notes section.**

In the **Transcript Assignments** field (Figure 4), NetAffx uses a battery of methods (quality assessments and grading techniques) to catalog transcripts measured by probe sets. The quality of each probe set is documented by listing genomic alignment, cross-hybridization and hybridization with reverse complements of known mRNA sequences. The **Related Probesets by Grade** column in the Transcript Assignments field provides a convenient method for identifying other probe sets assigned to the mRNA sequences in the current record. For more information about transcript assignments, see the whitepaper *Transcript Assignment for NetAffx™ Annotations* at [www.affymetrix.com](http://www.affymetrix.com).

The **Genomic Alignment of Consensus/Exemplar Sequence** section (Figure 5) details the genomic coordinates of the probe set for the most recent genome build available.

Genomic Alignment of Consensus/Exemplar Sequence					
<b>Assembly</b>	May 2004 (NCBI 35)				
<b>Alignment(s)</b>	<b>Position</b>	<b>View using IGB</b>	<b>Identity</b>	<b>Coverage</b>	<b>Cytoband</b>
	chr3:197927538-197950475(+) UCSC	<a href="#">IGB</a> *	71.99	98.23	q29
* You can now view alignments using the <a href="#">Integrated Genome Browser (IGB)</a> . Note that you must <a href="#">start IGB</a> before clicking on any of the "IGB" links above.					

**Figure 5. Genomic Alignment of Consensus/Exemplar Sequence.**

The **Public Domain and Genome References** section and the **Functional Annotations** section (Figure 6) are condensed views of known functional data including gene ontology, protein domain similarity, and ortholog information. As a reference for poorly characterized probe sets, the transcript is BLASTed (see *NetAffx™ Tools* below) against the non-redundant protein database.

Public Domain and Genome References			
Gene Title	phosphatidylinositol glycan, class X		
Gene Symbol	PIGX <a href="#">HGNC</a>		
Chromosomal Location	3q29		
UniGene ID Build 187 (02 Oct 2005)	Hs.223296 <a href="#">NCBI</a> (FULL LENGTH)		
Ensembl	ENSG00000163964 <a href="#">Ensembl</a>		
Entrez Gene	54965 <a href="#">Entrez gene</a>		
SwissProt	Q8TBF5 <a href="#">EMBL-EBI</a> Q9NWX2 <a href="#">EMBL-EBI</a>		
RefSeq Protein ID	NP_060331.1 <a href="#">NCBI</a>		
RefSeq	<a href="#">RefSeq Transcript ID</a>	<a href="#">RefSeq Title</a>	
	NM_017861	<a href="#">NCBI</a>	
Functional Annotations			
Ortholog	<a href="#">ID</a>	<a href="#">Title</a>	<a href="#">Organism</a>
	<a href="#">CANINE_2:CFA1589.1.A1_AT</a>	similar to GPI-mannosyltransferase subunit	cfa
	<a href="#">CANINE_2:CFA1589.1.A1_S_AT</a>	similar to GPI-mannosyltransferase subunit	cfa
	<a href="#">CANINE_2:CFAAFFX.20189.1.S1_AT</a>	similar to GPI-mannosyltransferase subunit	cfa
	<a href="#">CHICKEN:GGA12404.4.S1_A_AT</a>	phosphatidylinositol glycan, class X	gga
	<a href="#">CHICKEN:GGA12404.3.S1_AT</a>	Phosphatidylinositol glycan, class X	gga
	<a href="#">MOE430A:1425134_A_AT</a>	phosphatidylinositol glycan, class X	mm
	<a href="#">MOUSE430_2:1425134_A_AT</a>	phosphatidylinositol glycan, class X	mm
	<a href="#">MOUSE430A_2:1425134_A_AT</a>	phosphatidylinositol glycan, class X	mm
	<a href="#">MOE430A:1455284_X_AT</a>	phosphatidylinositol glycan, class X	mm
	<a href="#">MOUSE430_2:1455284_X_AT</a>	phosphatidylinositol glycan, class X	mm
	<a href="#">MOUSE430A_2:1455284_X_AT</a>	phosphatidylinositol glycan, class X	mm
	<a href="#">MG-U74AV2:160444_AT</a>	phosphatidylinositol glycan, class X	mm
	<a href="#">MG-U74BV2:164533_F_AT</a>	phosphatidylinositol glycan, class X	mm
	<a href="#">MU19KSUBB:TC27313_AT</a>	Phosphatidylinositol glycan, class X (Pigx), mRNA	mm
<a href="#">RG-U34C:RC_A1172192_AT</a>	similar to hypothetical protein FLJ20522 (predicted)	rn	
<a href="#">RAE230A:1374571_AT</a>	similar to hypothetical protein FLJ20522 (predicted)	rn	
<a href="#">RAT230_2:1374571_AT</a>	similar to hypothetical protein FLJ20522 (predicted)	rn	
Gene Ontology	<a href="#">GO Molecular Function (view graph)</a>		
	<a href="#">ID</a>	<a href="#">Description</a>	<a href="#">Evidence</a>
	16740	transferase activity	inferred from electronic annotation <a href="#">QuickGO</a> <a href="#">AmiGO</a>
16757	transferase activity, transferring glycosyl groups	inferred from electronic annotation <a href="#">QuickGO</a> <a href="#">AmiGO</a>	
Protein Similarities	<a href="#">Method</a>	<a href="#">ID</a>	<a href="#">Description</a>
	blast	<a href="#">NP_060331</a>	GPI-mannosyltransferase subunit [Homo sapiens] gb AAH22542.1  GPI-mannosyltransferase subunit [Homo sapiens]
	blast	<a href="#">BAA91233.1</a>	unnamed protein product [Homo sapiens]
	blast	<a href="#">XP_535778</a>	PREDICTED: similar to GPI-mannosyltransferase subunit [Canis familiaris]
Protein Domains	<a href="#">Trans Membrane</a>		
	<a href="#">ID</a>	<a href="#">Number Of Domains</a>	<a href="#">Domain Boundaries</a>
	NP_060331.1	1	188-210

Figure 6. Public Domain and Genome References.

## NetAffx™ Tools

NetAffx provides the following tools to help the user understand the rationale behind the probe selection process and to precisely identify probe set(s) that detect sequence(s) of interest:

- Probe Match Tool
- BLAST Tool
- Probe Set Display Tool

### Probe Match Tool

The Probe Match tool (**NetAffx** → **Expression** → **Probe Match**) is useful for determining precisely whether or not a gene or nucleotide sequence of interest is represented on an array. It provides the alignments of the probes against the input sequence (Figure 7) and helps identify the precise region of the transcript that hybridizes with the probes.

Search results for: HG-U133\_Plus\_2  
[Return to all results.](#)

Query: 'gi 44955884 ref NM\_203377.1 Homo sapiens myoglobin MB, transcript variant 2, mRNA' (1170 bases)  
 Matched Probes For Probe Set: HG-U133-PLUS 204179\_at

Note: you may also download this table as a tab-separated-value [spreadsheet file](#), or this entire page as a [text file](#).

Serial Order	Query Start	Query Stop	Probe Location in Affymetrix Consensus Exemplar	Probe Sequence (5'-3')
1	585	609	495	TGTTCCGGAAAGGACATGGCCTCCAA
2	667	691	577	GGGCCCCGGGTTCAAGAGAGAGCGG
3	692	716	602	GGTCTGATCTCGTGTAGCCATATAG
5	866	890	775	CTCACTGTGTTCTGCATGGTTTGG
6	909	933	818	TCTTCTAAATCCCAACCGAACTTCT
7	943	967	852	CAAACCTGGCTGTAACCCCAAAATCCA
8	962	986	871	AAATCCAAGCCATTAACCTACACCTGA
9	1005	1029	914	TAATCACTGGCCCTTGAAGACAGC
10	1024	1048	933	GACAGCAGAAATGTCCTTTGCAATG
11	1105	1129	1014	TGTGTGTGCTCCTCAGGTATGGCA

**Alignment**

```

Query      1  GAGCATGTTGGCCTGGTCCTTTGCTAGGTAAGTGTAGAGCAGGTGAGAGAGTGAGGGGGAA 60
Query     61  GGACTCCAATATTAGACCAAGTTCTTAGCCATGAAGCAGAGACTCTGAAGCCAGACTACCTG 120
Query    121  GGTCCCAATCTGGGCTTGGTATTTCTCGCTGTGTGACTCTGGACTGCGCCATGGGGCT 180
  
```

**Figure 7. Sample results from the Probe Match tool. Alignment of the myoglobin mRNA with the probes from the U133 Plus 2.0 array.**

For the best results, it is important to obtain the most complete mRNA sequence for the gene of interest. The sequences in the Reference Sequence (RefSeq) collection at NCBI (<http://www.ncbi.nlm.nih.gov/RefSeq/>) are usually the best choice. If there is no RefSeq sequence, then the UniGene cluster, annotated as containing “complete CDS” (Coding Sequence) for the gene of interest, is used.



## BLAST Tool

If there are no hits with the Probe Match tool, the next step is to use the NCBI BLAST (Basic Local Alignment Search Tool) against the consensus/target database (**NetAffx** → **Expression** → **BLAST** ). Since Probe Match only finds perfect matches, BLAST (Figure 8) helps to identify probes with mismatches. While the Probe Match tool looks for perfect matches between the probes and the input sequence, the BLAST tool is more liberal and allows for gaps and mismatches in the alignment. Moreover, the BLAST tool does not provide searches against the probes. It only allows comparison with the consensus and target sequences.

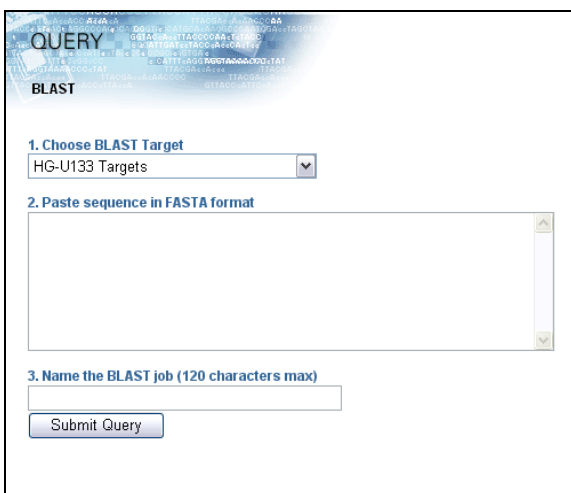
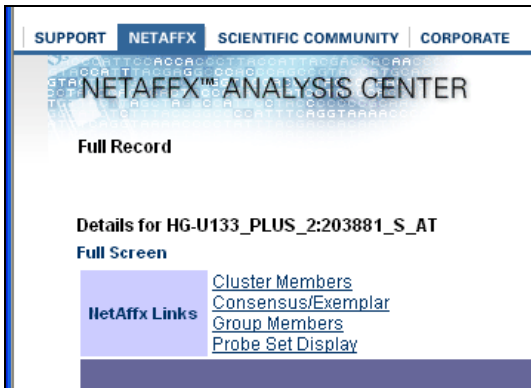


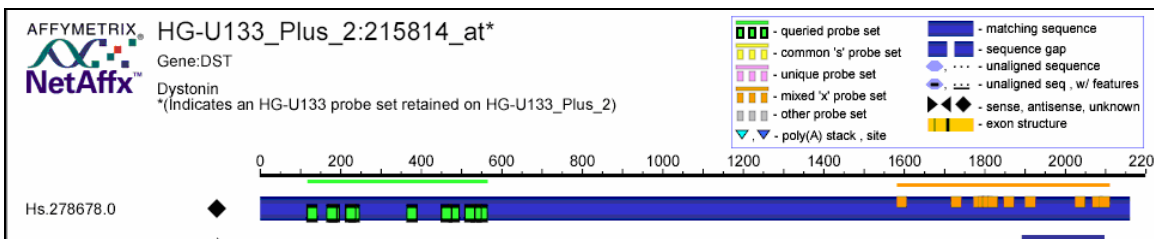
Figure 8. BLAST Tool

## Probe Set Display Tool

The **Probe Set Display** tool, a link (Figure 9) at the top of the Full Record page (**NetAffx** → **Expression** → **Query** → **Search** → **Details** → **Full Record** → **Probe Set Display**), provides a graphical display of the relationship between consensus, target and probe sequences on Expression Arrays. It provides a visual description of the unique versus non-unique (**\_s\_at**, **\_x\_at**, and **\_a\_at**) probe sets and helps understand probes that are designed against the antisense transcript. During sequence selection, an attempt is made to determine the orientation of the resulting consensus/exemplar sequence using a variety of evidence like EST read direction, CDS orientation, consensus splice sites, and polyA site/signal. Occasionally, it is not possible to determine the orientation of the transcript. In those cases, Affymetrix tiles probes against both strands of the consensus sequence. Unknown orientation is indicated by a solid black diamond in the probe set display (Figure 10).



**Figure 9. NetAffx link to Probe Set Display.**



**Figure 10. Probe Set Display.** Graphical display of a transcript sequence (blue bar) whose orientation is unknown (indicated by a solid black diamond to the left of the blue bar). Two probe sets (for both the forward and reverse strand) are tiled for this transcript to ensure that this gene is represented on the array.

## Public Domain Bioinformatics Databases

The transcript record assignments are not the end of the story. The NetAffx transcript assignment process identifies and catalogs the mRNA sequences detected by the probe sets. This is followed by the mapping of gene centric annotations for the assigned mRNA from a large collection of public domain annotation databases. The following is a brief description of some of these databases that catalog gene centric annotations for mRNA sequences.

### Entrez Gene and UniGene Annotations

**Entrez Gene** and **UniGene** databases on the NCBI website provide gene level annotations for a large group of organisms (3, 4). Entrez Gene (Figure 11) includes only well curated mRNAs for a given locus; therefore, the curation is stable but not comprehensive. The UniGene database is comprehensive and includes most of the mRNAs for each locus, but it is dynamic and changes significantly with each new version. Therefore, NetAffx uses UniGene only for mRNAs that are not included in the Entrez Gene database.

If you want to validate the NetAffx probeset-to-gene association, you have to validate both the probeset-to-mRNA association, mapped by the NetAffx annotation pipeline, and the mRNA-to-gene name association, provided by public domain databases like UniGene and Entrez Gene. To further illustrate this point, consider the probe set *215611\_at* on the U133 Plus 2.0 array. This probe set has a Grade A assignment to the GenBank<sup>®</sup> mRNA *AK022018*. NetAffx obtained gene centric annotations for this transcript from Entrez Gene (Figure 11). *AK022018* is annotated as TCF2 by Entrez Gene, and NetAffx reflects this association. However, the UCSC genomic alignment display for this probe set indicates that this transcript aligns with the intronic region of TCF2 RefSeq sequence (Figure 12). Although Entrez Gene, and therefore NetAffx, documents the mRNA *AK022018* as TCF2 mRNA, the genomic alignment contradicts this information. Using more than one database to confirm the annotations is necessary to produce the most accurate identification.

Annotation Method Description											
Annotation Description	This probe set was annotated using the Matching Probes based pipeline to a <u>Entrez Gene Identifier</u> using 1 transcript(s).										
Annotation Grade	This is a grade A annotation.										
Annotation Transcript Cluster (# of Matching Probes)	AK022018(11)										
Transcript Assignments	<table border="1"> <thead> <tr> <th>Representative Transcript</th> <th>Description</th> <th>Matching Probes</th> <th>Related Probesets by Grade</th> </tr> </thead> <tbody> <tr> <td>AK022018 <a href="#">NCBI</a></td> <td>Homo sapiens cDNA FLJ11956 fis, clone HEMBB1000893.11/11</td> <td></td> <td>A</td> </tr> </tbody> </table>	Representative Transcript	Description	Matching Probes	Related Probesets by Grade	AK022018 <a href="#">NCBI</a>	Homo sapiens cDNA FLJ11956 fis, clone HEMBB1000893.11/11		A		
Representative Transcript	Description	Matching Probes	Related Probesets by Grade								
AK022018 <a href="#">NCBI</a>	Homo sapiens cDNA FLJ11956 fis, clone HEMBB1000893.11/11		A								
Annotation Notes	There are no noteworthy cross hybridizing mRNAs found for this probe set.										
Genomic Alignment of Consensus/Exemplar Sequence											
Assembly	May 2004 (NCBI 35)										
Alignment(s)	<table border="1"> <thead> <tr> <th>Position</th> <th>View using IGB</th> <th>Identity</th> <th>Coverage</th> <th>Cytoband</th> </tr> </thead> <tbody> <tr> <td>chr15:55151662-55153850(+) <a href="#">UCSC</a></td> <td><a href="#">IGB</a>*</td> <td>72.65</td> <td>99.32</td> <td>q21.3</td> </tr> </tbody> </table>	Position	View using IGB	Identity	Coverage	Cytoband	chr15:55151662-55153850(+) <a href="#">UCSC</a>	<a href="#">IGB</a> *	72.65	99.32	q21.3
Position	View using IGB	Identity	Coverage	Cytoband							
chr15:55151662-55153850(+) <a href="#">UCSC</a>	<a href="#">IGB</a> *	72.65	99.32	q21.3							
* You can now view alignments using the Integrated Genome Browser (IGB). Note that you must start IGB before clicking on any											

Figure 11. NetAffx<sup>™</sup> transcript assignment and gene level annotations for probe set *215611\_at*.

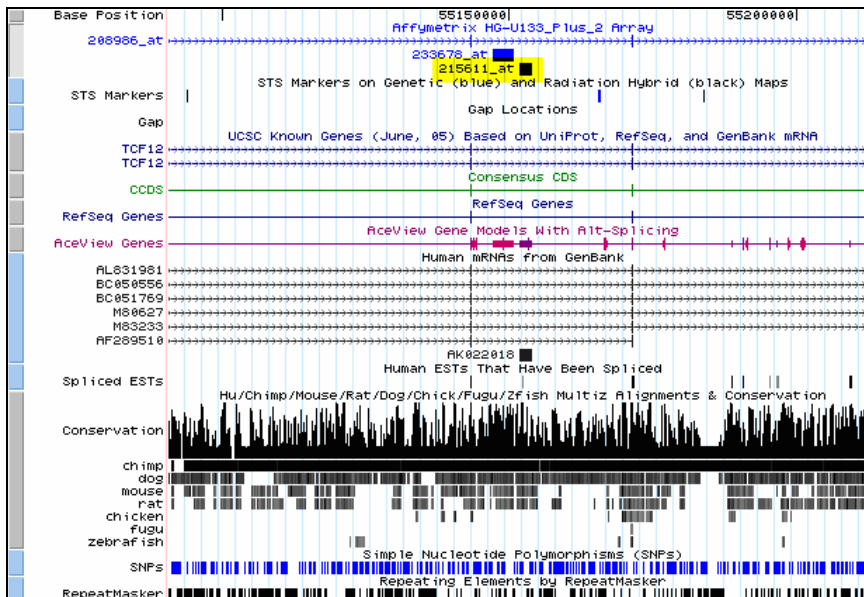


Figure 12. UCSC genome browser display of the alignment of the consensus sequence for the probe set *215611\_at*. It is evident from this display that the consensus and the corresponding mRNA align in the intronic region of the RefSeq for TCF2.

## Ensembl

**Ensembl** contains a comprehensive collection of annotations and transcripts (5). The Ensembl summary pages are available at the gene and transcript levels with excellent graphical representation. The NetAffx details view (**NetAffx** → **Query** → **Search** → **Details**) provides links to relevant gene level views on the Ensembl website.

## Public Domain Bioinformatics Tools

Bioinformatics tools are available in the public domain and are used to further explore probe set annotations provided in NetAffx.

### NCBI BLAST Suite

The [NCBI BLAST Suite](#) (Figure 13) is a collection of sequence comparison tools based on the BLAST algorithm (6). NCBI blastn (nucleotide-nucleotide BLAST), by default, searches the non-redundant (**nr**) database (Figure 14), a collection of all the mRNA sequences. You may have to select the EST database to explore some probe sets. NCBI only removes redundant sequence submissions to create the **nr** database. It does not remove sequences based on sequence identity threshold.

**NCBI -> BLAST** Latest news: 6 December 2005 : BLAST 2.2.13 released

**About**

- Getting started
- News
- FAQs

**More info**

- NAR 2004
- NCBI Handbook
- The Statistics of Sequence Similarity Scores

**Software**

- Downloads
- Developer info

**Other resources**

- References
- NCBI Contributors
- Mailing list
- Contact us

The **Basic Local Alignment Search Tool (BLAST)** finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

<b>Nucleotide</b> <ul style="list-style-type: none"><li>Quickly search for highly similar sequences (megablast)</li><li>Quickly search for divergent sequences (discontiguous megablast)</li><li>Nucleotide-nucleotide BLAST (blastn)</li><li>Search for short, nearly exact matches</li><li>Search trace archives with megablast or discontiguous megablast</li></ul>	<b>Protein</b> <ul style="list-style-type: none"><li>Protein-protein BLAST (blastp)</li><li>Position-specific iterated and pattern-hit initiated BLAST (PSI- and PHI-BLAST)</li><li>Search for short, nearly exact matches</li><li>Search the conserved domain database (rpsblast)</li><li>Protein homology by domain architecture (cdart)</li></ul>
<b>Translated</b> <ul style="list-style-type: none"><li>Translated query vs. protein database (blastx)</li><li>Protein query vs. translated database (tblastn)</li><li>Translated query vs. translated database (tblastx)</li></ul>	<b>Genomes</b> <ul style="list-style-type: none"><li>Human, mouse, rat, chimp <b>NEW</b>, cow, pig, dog, sheep, cat</li><li>Chicken, puffer fish, zebrafish</li><li>Environmental samples</li><li>Protozoa</li><li>Insects, nematodes, plants, fungi, microbial genomes, other eukaryotic genomes</li></ul>
<b>Special</b> <ul style="list-style-type: none"><li>Search for gene expression data (GEO BLAST)</li><li>Align two sequences (tbl2seq)</li><li>Screen for vector contamination (VecScreen)</li><li>Immunoglobulin BLAST (IgBlast)</li><li>SNP BLAST</li></ul>	<b>Meta</b> <ul style="list-style-type: none"><li>Retrieve results</li></ul>

Figure 13. NCBI BLAST Suite.

The **BLAST Database Content** page (NCBI → BLAST → About → Getting Started) displays its searchable, protein sequence databases.

**2. BLAST Database Content**

A BLAST search has four components: query, database, program, and search purpose/goal. To discuss effective BLAST program selection, we first need to know what databases are available and what sequences these databases contain. In this section, we will first take a look at the common BLAST databases. According to their content, they are grouped into nucleotide and protein databases. These databases and their detailed compositions are listed in the two tables below.

NCBI also provides specialized BLAST databases such as the vector screening database, variety of genome databases for different organisms, and trace databases. The contents for the three important model organisms, i.e., human, mouse, and rat, are described in Table 2.3. For other organisms, the content of their genome blast pages will be listed when these special BLAST pages are discussed.

Table 2.1 Content of Protein Sequence Databases	
Database <sup>1</sup>	Content Description
<b>nr</b>	Non-redundant GenBank CDS translations + PDB + SwissProt + PIR + PRF, excluding those in env_nr.
refseq	Protein sequences from <a href="#">NCBI Reference Sequence project</a> .
swissprot	Last major release of the SWISS-PROT protein sequence database (no incremental updates).
pat	Proteins from the Patent division of GenBank.
month	All new or revised GenBank CDS translations + PDB + SwissProt + PIR + PRF released in the last 30 days.
pdb	Sequences derived from the 3-dimensional structure records from the Protein Data Bank.
env_nr	Non-redundant CDS translations from env_nt entries.
Smart v4.0 <sup>2</sup>	663 PSSMs from Smart, no longer actively maintained.
Pfam v11.0 <sup>2</sup>	7255 PSSMs from Pfam, not the latest.
COG v1.00 <sup>2</sup>	4873 PSSMs from NCBI COG set.
kOG v1.00 <sup>2</sup>	4825 PSSMs from NCBI kOG set (eukaryotic COG equivalent).
<b>DDD v2.05 <sup>2</sup></b>	11399 PSSMs from NCBI curated cd set.

NOTE:  
<sup>1</sup> default database is in bold.  
<sup>2</sup> These databases are searchable only from rpsblast page, actual version may vary.

**Figure 14. Databases available on BLAST .**

The [BLAST 2 SEQUENCES](#) tool (Figure 15) can be used to align two nucleotide sequences and is useful for making a precise comparison between consensus/exemplar/target sequences and mRNA sequences.

**Figure 15. BLAST 2 SEQUENCES Tool.**

## UCSC BLAT Tool

The [UCSC BLAT Search Genome](http://www.genome.ucsc.edu) tool (www.genome.ucsc.edu) (Figure 16) can be used to align sequences of interest to the genome (7). It currently supports several organisms and integrates a wide variety of genome annotations. The results can be viewed in the genome browser in the context of other genome annotations.



The screenshot shows the UCSC BLAT Search Genome tool interface. At the top, there is a navigation bar with links: Home, Genomes, Tables, Gene Sorter, PCR, FAQ, and Help. Below this is the title "Human BLAT Search" and "BLAT Search Genome". The main form has five dropdown menus: "Genome:" (set to "Human"), "Assembly:" (set to "May 2004"), "Query type:" (set to "BLAT's guess"), "Sort output:" (set to "query.score"), and "Output type:" (set to "hyperlink"). Below these is a large text input area for the query sequence. At the bottom of the input area are three buttons: "submit", "I'm feeling lucky", and "clear". Below the input area is a paragraph of instructions: "Paste in a query sequence to find its location in the the genome. Multiple sequences may be searched if separated by lines starting with '>' followed by the sequence name." Below this is a "File Upload" section with the text: "Rather than pasting a sequence, you can choose to upload a text file containing the sequence." This section includes an "Upload sequence:" label, a text input field, a "Browse..." button, and a "submit file" button. At the bottom, there is a note: "Only DNA sequences of 25,000 or fewer bases and protein or translated sequence of 10000 or fewer letters will be processed. Up to 25 sequences can be submitted at the same time. The total limit for multiple sequence submissions is 50,000 bases or 25,000 letters." and a link: "For locating PCR primers, use [In-Silico PCR](#) for best results instead of BLAT."

**Figure 16. UCSC BLAT Search Genome tool.**

If you want to obtain a comprehensive, visual representation of a wide variety of features in the genomic context of a probe set or a gene of interest (8), you can use the UCSC Genome Browser (Figure 17). Affymetrix provides links from NetAffx to the UCSC genome browser for several arrays with genome alignments. The display shows the consensus sequences as a custom track. On the UCSC browser page, turn ON the GenBank/RefSeq/KnownGene mRNA and EST tracks to look at all the transcriptional evidence in that genomic region. One of the limitations here is that the display does not indicate whether or not the consensus sequence alignment is complete. This is due to the fact that sometimes the consensus sequence only partially aligns to the genome, and the most relevant target region of the consensus sequence may not align to the genome at all.

Figure 17. UCSC Genome Browser.

### Integrated Genome Browser (IGB)

The [Integrated Genome Browser](#) (IGB) is an Affymetrix desktop application that is used to visualize and explore genomic annotations from various data sources. Annotations from any publicly Distributed Annotations System (DAS) server(s), including UCSC and Ensembl, can be loaded and explored in IGB (Figure 18).

Figure 18. IGB on the Affymetrix web site

The IGB display can be accessed from links within the NetAffx probe set Full Record page under the **Genomic Alignment of Consensus/Exemplar Sequence** section (Figure 19). IGB displays the alignment of the consensus sequences and the 25-mers to the genome and therefore provides a very precise view of the relationship between the probe set and the latest transcriptional evidence (mRNA and EST). You may also load additional annotation tracks, such as **RefSeq** and **Known Genes**, from the UCSC DAS server. The IGB display color codes to differentiate between **\_at**, **\_s\_at**, and **\_x\_at** probe sets. In future versions of NetAffx, for designs other than Expression Arrays, IGB will replace the Probe Set Display tool (see Figures 9 and 10).

Genomic Alignment of Consensus/Exemplar Sequence					
Assembly	May 2004 (NCBI 35)				
	Position	View using IGB	Identity	Coverage	Cytoband
Alignment(s)	chr1:158307503-158309435(+) <a href="#">UCSC</a>	<a href="#">IGB</a> *	99.59	99.59	q23.3
	* You can now view alignments using the <a href="#">Integrated Genome Browser (IGB)</a> . Note that you must <a href="#">start IGB</a> before clicking on any of the "IGB" links above.				

**Figure 19. IGB links on the NetAffx Full Record page.**



# CASE STUDIES

## Case 1: Resolving DMD Gene Isoforms

Alternative splicing changes the mRNA sequence in several ways. At its simplest level, an exon can be removed (exon skipping), lengthened or shortened (alternative 5' or 3' splicing). These changes in the mRNA sequence may or may not result in changes in the 3' region of the mRNA sequence. Since the IVT assay is 3' biased, it may not be possible to tile probes that can resolve all variants or pick unique probe sets for each of the variants. This results in non-unique probe sets that hybridize to multiple variants and cause redundancy where multiple probe sets detect the same transcript.

### Customer Question

Three probe sets match the DMD gene. What does this mean?

On the U133A chip there are 3 probe sets:

- *203881\_s\_at* had a probe match score of 11/11
- *207660\_at* had a probe match score of 11/11
- *208086\_s\_at* had a probe match score of 10/11

In NetAffx, it appears that both of the *\_s* probe sets are further downstream (closer to the 3' end). When the array is designed with the transcripts available at the time, the *\_s* set represents more than one transcript, whereas the *\_at* is associated with a single transcript. Therefore, we expect that all the probe sets will measure DMD; however, the *\_s* probe sets tend to measure more than one variant.

The previous information is a guideline relating to the time of design. To perform a follow-up on the results, it is important to look at what is currently known about the probe sets and the transcript record. The following workflow further explores the question of the DMD probe sets:

1. Determine the quality of the probe set transcript assignment:
  - A. NetAffx was searched with three probe set IDs, and the preconfigured **Annotation Method** view was used (Figure 20) for annotation. The data clearly indicates that all three probe sets have the highest quality of grade assignments (Grade A), and therefore, the majority of the probes in these probe sets align with one or more of the mRNA sequences for this locus. The table also provides the **Annotation Transcript Count** field, which provides a count of all the mRNA sequences that are assigned to the corresponding probe set.
  - B. The information illustrates that both of the *\_s\_at* probe sets detect all 18 mRNA sequence isoforms, while the unique probe set specifically detects one of the isoforms.

Results 1-3 of 3.  
[Export](#) | [GO Browser](#) | [Show Orthologs](#)  
[Full Screen](#) | \*Annotation Method\* [v] 50 [v] Remove Checked [v] Sa

Details	Probe Set ID	GeneChip Array	Annotation Method	Annotation Grade	Annotation Database	Annotation Total Mapped	Annotation Transcript Count
<input type="checkbox"/> <a href="#">Details</a>	203881_s_at	Human Genome U133A Array	Matching Probes	A	Entrez Gene	1	18
<input type="checkbox"/> <a href="#">Details</a>	207660_at	Human Genome U133A Array	Matching Probes	A	Entrez Gene	1	1
<input type="checkbox"/> <a href="#">Details</a>	208086_s_at	Human Genome U133A Array	Matching Probes	A	Entrez Gene	1	18

**Figure 20.** NetAffx™ table view indicating the quality of the transcript assignment and the number of transcripts assigned to each probe set.

2. Look for additional clues:

Detailed annotation reports for each of the probe sets were then scanned. Some of the relevant sections are mentioned here.

- A. The **Transcript Assignments** field (Figure 21) in the Annotation Method Description (Figure 4) section on the Full Record page lists all the mRNA sequences that are assigned to the corresponding probe set .

Transcript Assignments	Representative Transcript	Description	Matching Probes	Related Probesets by Grade
	NM_004019 <a href="#">NCBI</a>	Homo sapiens dystrophin (muscular dystrophy, Duchenne and Becker types) (DMD), transcript variant Dp40, mRNA.	11/11	A

**Figure 21.** Transcript Assignments” section lists all the mRNA sequences assigned to the probe-set.

- B. The **Splice Variants** field in the Annotation Method Description section on the Full Record page lists all splice variants for the DMD locus as documented by the RefSeq database and the corresponding probe sets that detect these variants (Figure 22). After scanning these fields for all three probe sets, it is clear that there are 18 different isoforms of the DMD gene as documented by RefSeq. Probe sets *203881\_s\_at* and *208086\_s\_at* measure 17 isoforms while probe set *207660\_at* specifically measures the 18<sup>th</sup> isoform.

Entrez	Total Refseq Isoforms	Total Detected By This Probeset	Refseq	Probesets
1756	18	1	NM_000109	<a href="#">203881 s at(11), 208086 s at(11)</a>
			NM_004006	<a href="#">203881 s at(11), 208086 s at(11)</a>
			NM_004007	<a href="#">203881 s at(11), 208086 s at(11)</a>
			NM_004009	<a href="#">203881 s at(11), 208086 s at(11)</a>
			NM_004010	<a href="#">203881 s at(11), 208086 s at(11)</a>
			NM_004011	<a href="#">203881 s at(11), 208086 s at(11)</a>
			NM_004012	<a href="#">203881 s at(11), 208086 s at(11)</a>
			NM_004013	<a href="#">203881 s at(11), 208086 s at(11)</a>
			NM_004014	<a href="#">203881 s at(11), 208086 s at(11)</a>
			NM_004015	<a href="#">203881 s at(11), 208086 s at(11)</a>
			NM_004016	<a href="#">203881 s at(11), 208086 s at(11)</a>
			NM_004017	<a href="#">203881 s at(11), 208086 s at(11)</a>
			NM_004018	<a href="#">203881 s at(11), 208086 s at(11)</a>
			NM_004019	<a href="#">207660 at(11)</a>
			NM_004020	<a href="#">203881 s at(11), 208086 s at(11)</a>
			NM_004021	<a href="#">203881 s at(11), 208086 s at(11)</a>
			NM_004022	<a href="#">203881 s at(11), 208086 s at(11)</a>
			NM_004023	<a href="#">203881 s at(11), 208086 s at(11)</a>

**Figure 22.** The Splice Variants field lists all the RefSeq isoform documents for a particular locus and the probe sets that detect each isoform.

3. Perform genomic visualization for further confirmation:

Use the genome browser to visualize the probe sets and clarify the specificity of the individual probe set for each isoform. For this you must first identify whether or not these probe sets have a valid consensus-to-genomic alignment.

A. A custom view was created to check whether or not the probe sets have genomic alignments (Figure 23).

This indicates that the consensi of all three probe sets align well with the human genome sequence.

Results 1-3 of 3.  
[Export](#) | [GO Browser](#) | [Show Orthologs](#)  
[Full Screen](#) | genome | 50 | [Remove Checked](#) | [Save Current List](#)

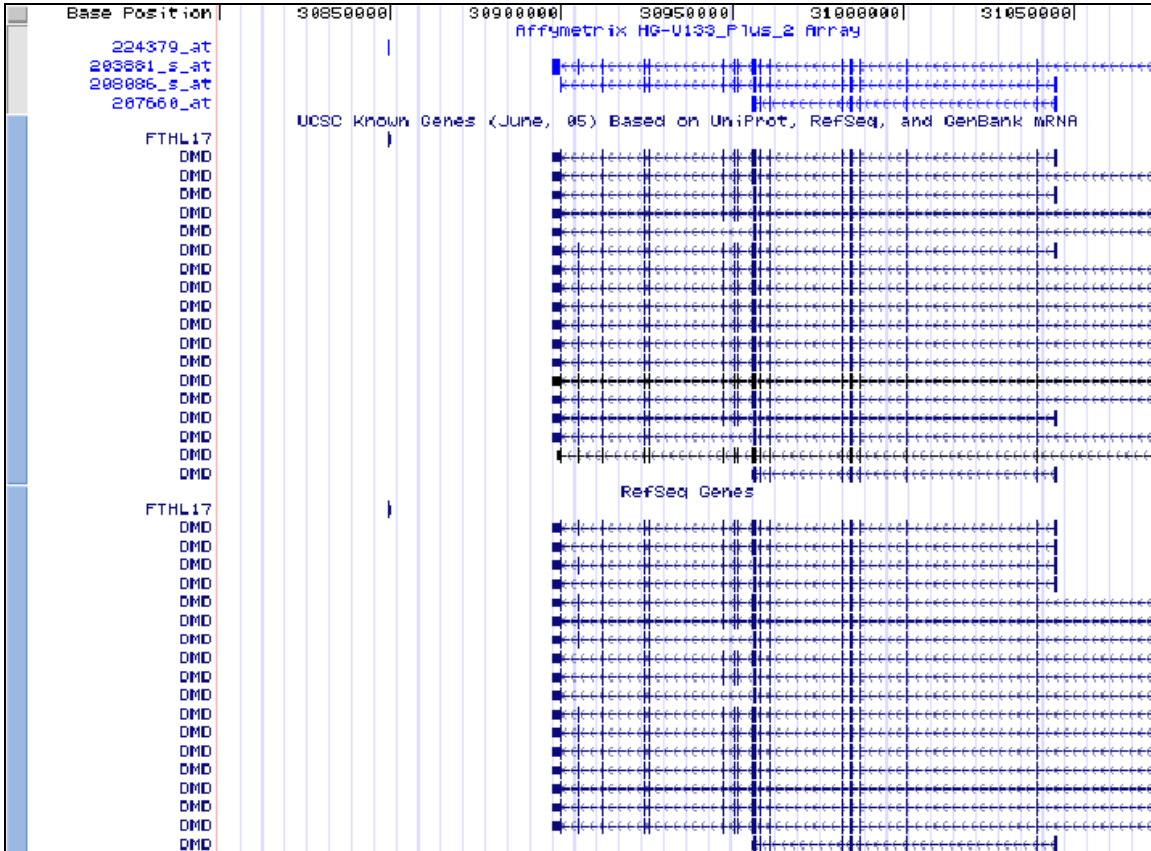
Details	Probe Set ID	Consensus/Exemplar Alignment Chromosome	Consensus/Exemplar Alignment Chromosome Start	Consensus/Exemplar Alignment Chromosome Stop	Consensus/Exemplar Alignment Coverage	Consensus/Exemplar Alignment Identity
<input type="checkbox"/> Details	203881_s_at	chrX	30897001	32906202	99.92	99.25
<input type="checkbox"/> Details	207660_at	chrX	30955968	31044681	95.47	95.47
<input type="checkbox"/> Details	208086_s_at	chrX	30899417	31044655	100.0	100.0

**Figure 23.** Custom View to check genomic alignment of the DMD probe sets.

The UCSC browser displays consensus alignment in a custom track and contains a comprehensive collection of genome annotations. On the other hand, IGB displays the alignments of the individual 25-mers along with the consensus sequence alignments.

- B. For one of the probe sets, the **UCSC** link, in the Genomic Alignment of Consensus/Exemplar Sequence section on the Full Record page, was followed.

The UCSC genome browser displays the region of the genome that encodes the DMD gene with the consensi alignments in a custom track (Figure 24). It is clear that the 3' region (the region where probes were designed) of two of the consensi align with 17 out of 18 RefSeq alignments, while the third consensus sequence targets one specific isoform.



**Figure 24.** The UCSC genome browser display for DMD gene.

The IGB display in Figure 25 shows that *207660\_at* is unique to a specific, shorter isoform of DMD represented by *NM\_004019*. In Figure, the other two probe sets are equivalent and detect all 17 isoforms equally well.



Figure 25. IGB display shows the probes (pink bars in the bottom) on the consensus that are specific for a particular RefSeq splice variant.

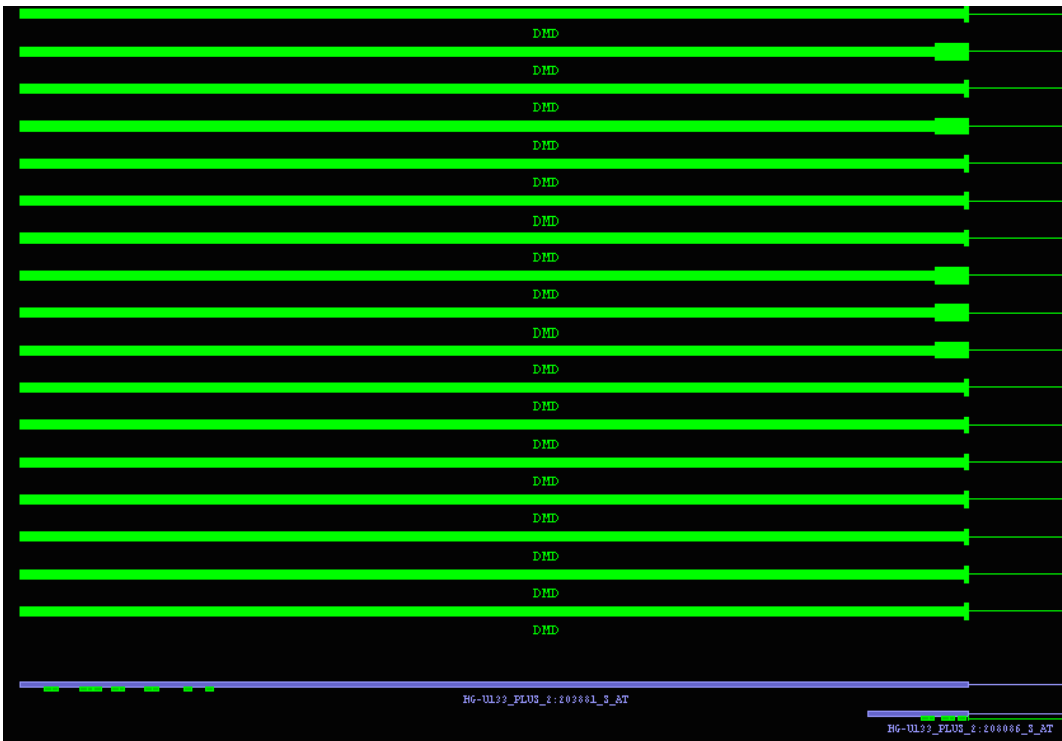


Figure 26. IGB display of two non-unique (s\_at) probe sets showing probes aligning with 17/18 RefSeq isoforms for the DMD gene.

**Conclusion:**

Based on the above findings, we can conclude that DMD is present in the experimental sample, but we cannot say which of the 17 isoforms is present. We can also conclude that the shortest isoform is absent, since *207660\_at* does not show a positive signal.

**Case 2: Erroneous GenBank mRNA Sequence for RAP1A**

One of the main objectives of our sequence selection process is to provide accurate representation of the mRNA and EST sequences in the public domain. The ability of the probe set to accurately measure a given transcript reflects quality of the transcript sequence information that is available at the time of array design. The mRNA and EST sequence information is limited, absent, or even erroneous in some cases. Moreover, the sequence information is constantly evolving – more so for some organisms than others – and therefore the quality of a probe set also changes with changing transcript information in the public domain.

**Customer Question**

The target sequence for probe set *1555339\_at* on the HGU133 Plus 2.0 array does not represent RAP1A transcript as NetAffx claims; however, the consensus had an overlap of ~500 bases. Does this probe set detect RAP1A, and which sequence is more reliable for the design of primers for PCR?

The following workflow explores the customer’s question:

1. Determine the quality of the transcript assignment.
  - A. The detailed NetAffx annotation report for this probe set indicates that it has a high quality transcript assignment (Grade A) to GenBank mRNA sequence AB051846. The **Transcript Assignments** section in NetAffx indicates that all 11 probes in the probe set align perfectly with this mRNA sequence. The **Annotation Description** field indicates that this mRNA is annotated as RAP1A by Entrez Gene (Figure 27).

Annotation Method Description				
<b>Annotation Description</b>	This probe set was annotated using the Matching Probes based pipeline to a Entrez Gene identifier using 1 transcript(s).			
<b>Annotation Grade</b>	This is a grade A annotation.			
<b>Annotation Transcript Cluster (# of Matching Probes)</b>	AB051846(11)			
<b>Transcript Assignments</b>	<b>Representative Transcript</b>	<b>Description</b>	<b>Matching Probes</b>	<b>Related Probesets by Grade</b>
	AB051846 <a href="#">NCBI</a>	Homo sapiens mRNA for Raichu404X, complete cds.	11/11	<u>A</u>
<b>Annotation Notes</b>	There are no noteworthy cross hybridizing mRNAs found for this probe set.			

**Figure 27. NetAffx™ transcript assignment report for 1555339\_at.**

B. Therefore, it is clear that the probe set accurately and precisely represents a GenBank mRNA. The fact that the gene-centric annotations are mapped from Entrez Gene provides further confidence that this probe set most likely detects RAPIA gene.

2. Determine the design evidence supporting this probe set.

Cluster evidence indicates how many sequence records support the probe set. The **Cluster Evidence** field in the **Probe Design Information** section in Figure 28 clearly indicates that the design evidence for this probe set is weak with only one mRNA sequence supporting this probe set. A well documented cluster may have many ESTs and a group of several mRNA full length sequences supporting it.

Probe Design Information	
Transcript ID(Array Design)	Hs2Affx.1.213
Sequence Type	Consensus sequence
Representative Public ID	AB051846 <a href="#">NCBI</a>
Target Description	gb:AB051846.1 /DB_XREF=gi:14595131 /GEN=Raichu404X /TID=Hs2Affx.1.213 /CNT=1 /FEA=FLmRNA /TIER=FL /STK=1 /NOTE=sequence(s) not in UniGene /DEF=Homo sapiens mRNA for Raichu404X, complete cds. /PROD=Raichu404X /FL=gb:AB051846.1
Cluster Evidence	This cluster is supported by a Full-length mRNA and contains 1 sequence(s).
Probe Selection Region Evidence	This Probe Selection Region is supported by a Full-length mRNA and contains 1 sequence(s).

**Figure 28. Design evidence supporting the probe set 1555339\_at. The evidence is weak with only one mRNA supporting the probe set.**

3. Explore current mRNA evidence in the public domain for RAPIA.

A. Since the probe set seems to accurately represent the mRNA sequence, the next step would be to check how the probe set compares with additional mRNA sequence evidence for RAPIA. The NetAffx Entrez Gene link (labeled NCBI) was followed.

Note: The NCBI link takes the user to a report that uses Entrez Gene, NCBI's database for gene-specific information.

B. The Entrez Gene report (Figure 29) indicates that other high quality mRNA sequences (RefSeqs) are available for this gene.

NCBI Reference Sequences (RefSeq)	
<b>mRNA Sequence</b> <a href="#">NM_001010935</a> <b>Transcriptional Variant</b> Transcript Variant: This variant (1) represents the longer transcript. Both variants 1 and 2 encode the same protein. <b>Source Sequence</b> <a href="#">BC014086</a> , <a href="#">BI460853</a> , <a href="#">BU150941</a> <b>Product</b> <a href="#">NP_001010935</a> RAP1A, member of RAS oncogene family <b>Conserved Domains</b> (1) <a href="#">summary</a> <a href="#">cd00154: RAB, Rab subfamily of small GTPases</a> Location: 19 - 165 Blast Score: 319	
<b>mRNA Sequence</b> <a href="#">NM_002884</a> <b>Transcriptional Variant</b> Transcript Variant: This variant (2) differs in the 5' UTR, compared to variant 1. Both variants 1 and 2 encode the same protein. <b>Source Sequence</b> <a href="#">BC014086</a> , <a href="#">BI460853</a> , <a href="#">BU150941</a> <b>Product</b> <a href="#">NP_002875</a> RAP1A, member of RAS oncogene family <b>Consensus CDS (CCDS)</b> <a href="#">CCDS840.1</a> <b>Conserved Domains</b> (1) <a href="#">summary</a> <a href="#">cd00154: RAB, Rab subfamily of small GTPases</a> Location: 19 - 165 Blast Score: 319	
Related Sequences	
<b>Nucleotide</b>	<b>Protein</b>
<b>Genomic</b> <a href="#">A08691</a>	<a href="#">CAA00804</a>
<b>Genomic</b> <a href="#">AL049557</a>	<a href="#">CAB55685</a>
	<a href="#">CAI22712</a>
<b>mRNA</b> <a href="#">AB051846</a>	<a href="#">BAB61868</a>
<b>mRNA</b> <a href="#">AF493912</a>	<a href="#">AAM12626</a>
<b>mRNA</b> <a href="#">BC014086</a>	<a href="#">AAH14086</a>
<b>mRNA</b> <a href="#">BT019666</a>	<a href="#">AAV38472</a>
<b>mRNA</b> <a href="#">CR597469</a>	None
<b>mRNA</b> <a href="#">CR623933</a>	None
<b>mRNA</b> <a href="#">M22995</a>	<a href="#">AAA36150</a>
<b>mRNA</b> <a href="#">X12533</a>	<a href="#">CAA31051</a>
<b>mRNA</b> None	<a href="#">P62834</a>

**Figure 29. Entrez Gene report indicates the mRNA sequences for the RAP1A locus.**

4. How does the consensus compare with the best mRNA evidence (RefSeq) for RAP1A?
  - A. Using the NCBI [align 2 sequences](#) tool, you can determine the relationship of the consensus sequence with each of the RefSeq sequences for RAP1A (Figure 30a).
  - B. The data clearly indicate that the target or 3' region of the consensus does not align with either of the RefSeq sequences for RAP1A.



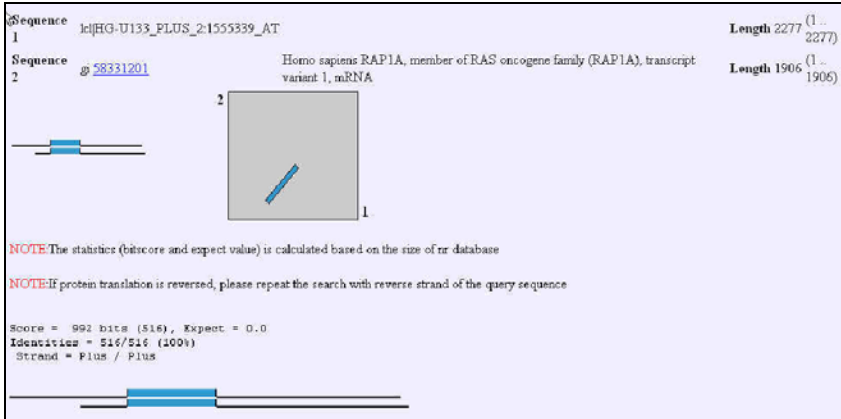


Figure 30a. Alignment of the consensus sequence with NM\_001010935, variant 1 of RAP1A. The thick blue bars indicate the region significance similarity.

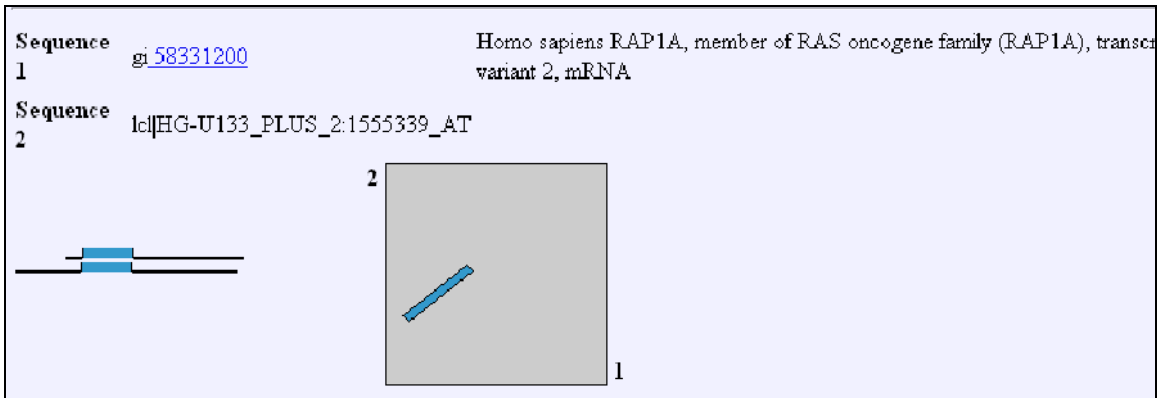


Figure 30b. Alignment of the consensus sequence with NM\_002884, variant 2 of RAP1A.

## Conclusion

The results indicate that the mRNA sequence used to design this probe set is erroneous and therefore this probe set does not actually measure RAP1A.

## REFERENCES

1. Kapranov P, Drenkow J, Cheng J, Long J, Helt G, Dike S, Gingeras TR. Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. *Genome Res.* **15**(7):987-97 (2005).
2. Liu G, Loraine AE, Shigeta R, Cline M, Cheng J, Valmeekam V, Sun S, Kulp D, Siani-Rose MA. NetAffx: Affymetrix probesets and annotations. *Nucleic Acids Res.* **31**(1):82-6 (2003).
3. Donna Maglott, Jim Ostell, Kim D. Pruitt, and Tatiana Tatusova. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.* **33**(Database Issue): D54–D58 (2005).
4. Pontius JU, Wagner L, Schuler GD. UniGene: a unified view of the transcriptome. In: *The NCBI Handbook*. Bethesda (MD): National Center for Biotechnology Information; 2003. [[Full Text](#)] [[PDF](#)]
5. Ewan Birney et al. An Overview of Ensembl. *Genome Res.* **14**(5):925-928 (2004).
6. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. "Basic local alignment search tool." *J. Mol. Biol.* 215:403-410 (1990).
7. Kent, W.J. BLAT - The BLAST-Like Alignment Tool. *Genome Res.* **12**(4), 656-664 (2002).
8. Kent, W.J., Sugnet, C. W., Furey, T. S., Roskin, K.M., Pringle, T. H., Zahler, A. M., and Haussler, D. The Human Genome Browser at UCSC. *Genome Res.* **12**(6), 996-1006 (2002).