Research Core Unit Transcriptomics

Crude probe characterization workflow

Version: 22.08.2012

Introduction

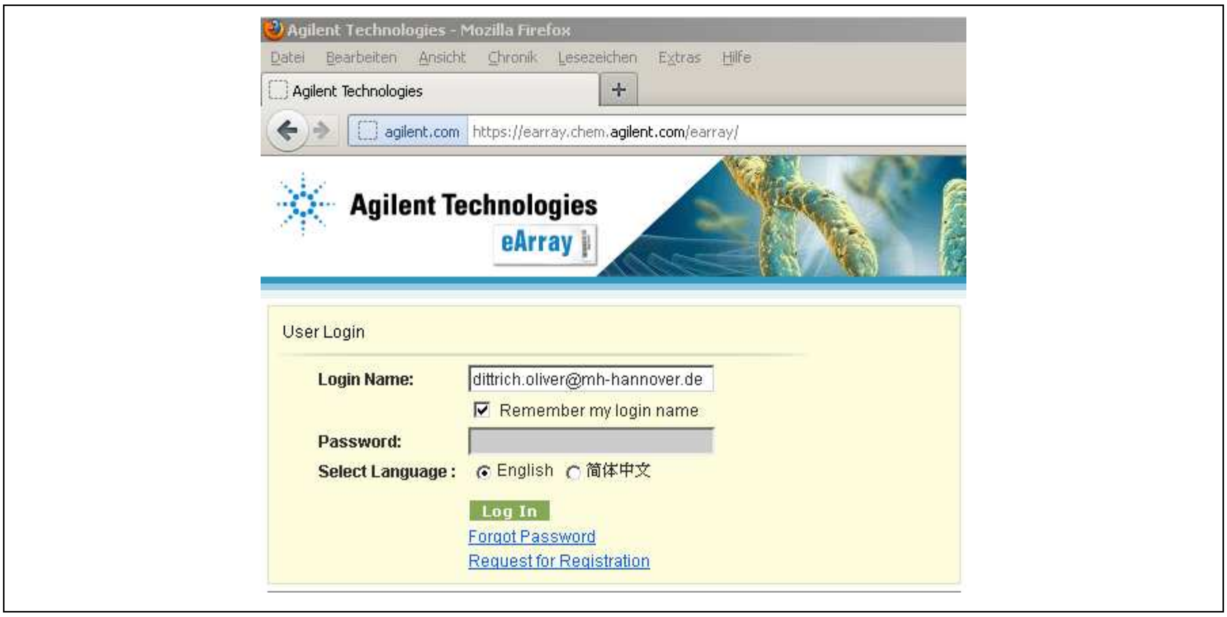
The crude probe characterization workflow is performed by us to allow for a rapid discrimination between probes that solely serve as technical controls and probes, directed against endogenous transcripts. Furthermore, the degree of functional characterization and the annotation state of the respective transcripts (genes) is considered and used to assign a specific attribute to all probes, according to the following 3 possibilities*:

- "1) gene symbol"
- "2) poorly characterized"
- "3) control"

The source of annotation data for Agilent's microarray probes

Annotation data is provided at Agilent's eArray portal (Figure 1). Regularly updated annotation files (Figure 2) are downloaded by us from this site and serve as basis for the annotation block within our standard Excel result files and for the crude probe characterization workflow (as described below).

* Please also note paragraph „ Important explanatory information " at the end of this manual.



Agilent Technologies - Mozilla Firefox
Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe
Agilent Technologies
agilent.com https://earray.chem.agilent.com/earray/
Agilent Technologies
eArray

User Login

Login Name: dittrich.oliver@mh-hannover.de

☑ Remember my login name

Password:

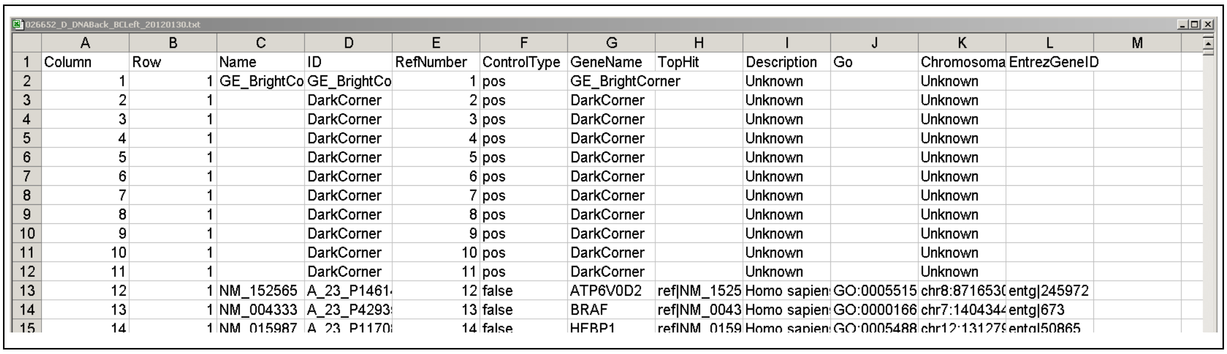
Select Language : ● English ○ 简体中文

Log In

Forgot Password

Request for Registration

Figure 1: The eArray login page.

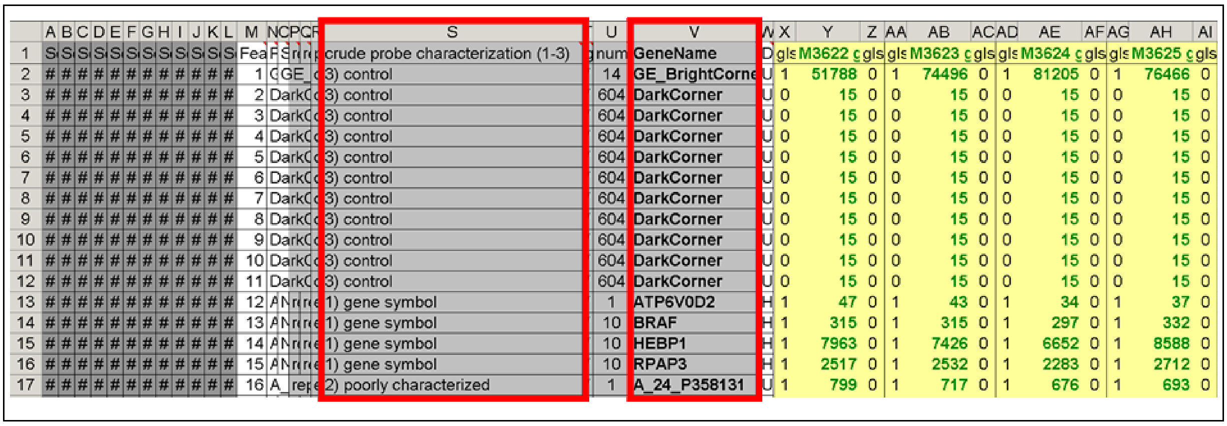


026652_D_DNABack_BCLeft_20120130.txt

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Column	Row	Name	ID	RefNumber	ControlType	GeneName	TopHit	Description	Go	Chromosoma	EntrezGeneID	
2	1	1	GE_BrightCo	GE_BrightCo	1	pos	GE_BrightCorner		Unknown		Unknown		
3	2	1		DarkCorner	2	pos	DarkCorner		Unknown		Unknown		
4	3	1		DarkCorner	3	pos	DarkCorner		Unknown		Unknown		
5	4	1		DarkCorner	4	pos	DarkCorner		Unknown		Unknown		
6	5	1		DarkCorner	5	pos	DarkCorner		Unknown		Unknown		
7	6	1		DarkCorner	6	pos	DarkCorner		Unknown		Unknown		
8	7	1		DarkCorner	7	pos	DarkCorner		Unknown		Unknown		
9	8	1		DarkCorner	8	pos	DarkCorner		Unknown		Unknown		
10	9	1		DarkCorner	9	pos	DarkCorner		Unknown		Unknown		
11	10	1		DarkCorner	10	pos	DarkCorner		Unknown		Unknown		
12	11	1		DarkCorner	11	pos	DarkCorner		Unknown		Unknown		
13	12	1	NM_152565	A_23_P1461	12	false	ATP6V0D2	ref|NM_1525	Homo sapien	GO:0005515	chr8:871653	entg|245972	
14	13	1	NM_004333	A_23_P4293	13	false	BRAF	ref|NM_0043	Homo sapien	GO:0000166	chr7:140434	entg|673	
15	14	1	NM_015987	A_23_P1170	14	false	HEBP1	ref|NM_0159	Homo sapien	GO:0005488	chr12:13127	entg|50865	

Figure 2: An example of an Agilent annotation file (AMADID 026652) available at the eArray portal. Such files serve as basis for the annotation block within our standard Excel result files and for the crude probe characterization workflow (Table 1).

As depicted in Figure 3, one of three possible attributes is assigned to each probe and is integrated into our standard Excel result file within column: "crude probe characterization (1-3)".



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI
1	S	S	S	S	S	S	S	S	S	S	S	S	Fea	F	S	r	r	p	crude probe characterization (1-3)	g	num	GeneName	D	gls	M3622 g	gls	gls	M3623 g	gls	gls	M3624 g	gls	gls	M3625 g	gls
2	#	#	#	#	#	#	#	#	#	#	#	#	1	C	GE_			c	3) control		14	GE_BrightCorne	U	1	51788	0	1	74496	0	1	81205	0	1	76466	0
3	#	#	#	#	#	#	#	#	#	#	#	#	2	DarkC				c	3) control		604	DarkCorner	U	0	15	0	0	15	0	0	15	0	0	15	0
4	#	#	#	#	#	#	#	#	#	#	#	#	3	DarkC				c	3) control		604	DarkCorner	U	0	15	0	0	15	0	0	15	0	0	15	0
5	#	#	#	#	#	#	#	#	#	#	#	#	4	DarkC				c	3) control		604	DarkCorner	U	0	15	0	0	15	0	0	15	0	0	15	0
6	#	#	#	#	#	#	#	#	#	#	#	#	5	DarkC				c	3) control		604	DarkCorner	U	0	15	0	0	15	0	0	15	0	0	15	0
7	#	#	#	#	#	#	#	#	#	#	#	#	6	DarkC				c	3) control		604	DarkCorner	U	0	15	0	0	15	0	0	15	0	0	15	0
8	#	#	#	#	#	#	#	#	#	#	#	#	7	DarkC				c	3) control		604	DarkCorner	U	0	15	0	0	15	0	0	15	0	0	15	0
9	#	#	#	#	#	#	#	#	#	#	#	#	8	DarkC				c	3) control		604	DarkCorner	U	0	15	0	0	15	0	0	15	0	0	15	0
10	#	#	#	#	#	#	#	#	#	#	#	#	9	DarkC				c	3) control		604	DarkCorner	U	0	15	0	0	15	0	0	15	0	0	15	0
11	#	#	#	#	#	#	#	#	#	#	#	#	10	DarkC				c	3) control		604	DarkCorner	U	0	15	0	0	15	0	0	15	0	0	15	0
12	#	#	#	#	#	#	#	#	#	#	#	#	11	DarkC				c	3) control		604	DarkCorner	U	0	15	0	0	15	0	0	15	0	0	15	0
13	#	#	#	#	#	#	#	#	#	#	#	#	12	A	N	r	r	e	1) gene symbol		1	ATP6V0D2	H	1	47	0	1	43	0	1	34	0	1	37	0
14	#	#	#	#	#	#	#	#	#	#	#	#	13	A	N	r	r	e	1) gene symbol		10	BRAF	H	1	315	0	1	315	0	1	297	0	1	332	0
15	#	#	#	#	#	#	#	#	#	#	#	#	14	A	N	r	r	e	1) gene symbol		10	HEBP1	H	1	7963	0	1	7426	0	1	6652	0	1	8588	0
16	#	#	#	#	#	#	#	#	#	#	#	#	15	A	N	r	r	e	1) gene symbol		10	RPAP3	H	1	2517	0	1	2532	0	1	2283	0	1	2712	0
17	#	#	#	#	#	#	#	#	#	#	#	#	16	A_		rep			2) poorly characterized		1	A_24_P358131	U	1	799	0	1	717	0	1	676	0	1	693	0

Figure 3: A part of the standard result file for single-color studies. The "crude probe characterization (1-3)" and the "GeneName" columns are highlighted in red.

Crude probe characterization workflows

The following workflow is used to assign the described attributes to all probes of Agilent's mRNA expression microarrays utilized by us. The column names refer to the annotation file as exemplified in Figure 2.

Table 1: The crude probe characterization workflow for human microarrays of design types AMADID 014850 and 026652, for murine microarrays of design types AMADID 014868 and 026655, and for rat microarrays of design types AMADID 014879 and 028282.

Step	Column to check in annotation file	Search pattern	Probe characterization
1	set all probes to		2) poorly characterized
2	Name	"NM_*"	1) gene symbol
3	Description	"*hypothetical*" or "*predicted*" (case insensitive), "*RIKEN*" or "*FLJ#*" or "*Mus musculus expressed sequence*" or "*Mus musculus cDNA sequence*"	2) poorly characterized
4	GeneName	"KIAA*" or "LOC*" or "hCG_*" or "RP11-*" or "RP1-*" or "RP3-*" or "RP6-*" or "MGC*" or "RGD*", "CXorf#*" or "CX#orf#*" or "CYorf#*" or "CY#orf#*" or "C#orf#*" or "C##orf#*" or "CXORF#*" or "CX#ORF#*" or "CYORF#*" or "CY#ORF#*" or "C#ORF#*" or "C##ORF#*"	2) poorly characterized
5	ControlType	not "false"	3) control
6	if one replicate probe of a gene is characterized as "1) gene symbol" then all other replicates are also set to		1) gene symbol

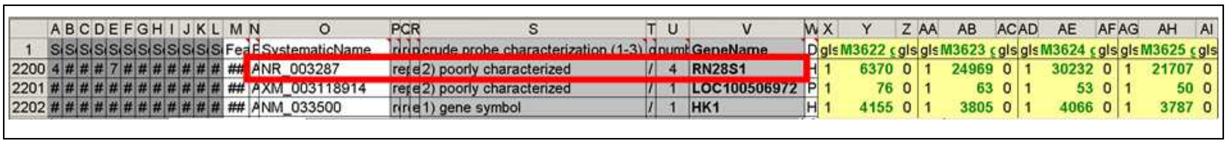
# - exactly one number is expected (0-9)

* - zero or more letters, numbers or any other character are allowed

Important explanatory information

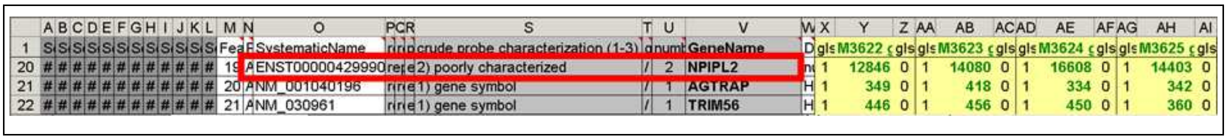
Please note, that the discrimination between "1) gene symbol" and "2) poorly characterized" in its present form is not completely conclusive in a semantical sense. According to our workflow, we decided to allocate the attribute "1) gene symbol" only for protein-coding transcripts (genes). Thus, all non-coding transcripts (e.g. those with an "NR_*" RefSeq accession entry) receive the attribute "2) poorly characterized", irrespective of their actual characterization/annotation status (see Figure 4 for an example).

On the other hand, there are some examples for transcripts (genes) for which no RefSeq accession entry of the type: "NM_*", but only an ENSEMBLE database entry exists: accession entry of the type "ENST*" (human), "ENSMUST*" (mouse) or "ENSRNOT*" (rat). Anyhow, even some of these transcripts (genes) could have a gene symbol officially allocated (Figure 5). In such a situation we decided to assign the respective probes with the attribute "2) poorly characterized", unless, additional probes, directed against the same transcript (gene) and possessing an "NM_*" accession entry are present on the microarray (see Table 1, steps 2 and 6).



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI
1	S	S	S	S	S	S	S	S	S	S	S	S	Fea	F	SystematicName	r	r	r	crude probe characterization (1-3)	d	numb	GeneName	D	gls	M3622	gls	gls	M3623	gls	gls	M3624	gls	gls	M3625	gls
2200	4	#	#	#	7	#	#	#	#	#	#	#	##	A	NR_003287	re	p	e	2) poorly characterized	/	4	RN28S1	H	1	6370	0	1	24969	0	1	30232	0	1	21707	0
2201	#	#	#	#	#	#	#	#	#	#	#	#	##	A	XM_003118914	re	p	e	2) poorly characterized	/	1	LOC100506972	P	1	76	0	1	63	0	1	53	0	1	50	0
2202	#	#	#	#	#	#	#	#	#	#	#	#	##	A	NM_033500	r	n	e	1) gene symbol	/	1	HK1	H	1	4155	0	1	3805	0	1	4066	0	1	3787	0

Figure 4: An example for RN28S1 (Homo sapiens RNA, 28S ribosomal 1 (RN28S1), ribosomal RNA [NR_003287]) as a transcript that is assigned by us as "2) poorly characterized" even though it is a well-characterized transcript but not a protein-coding transcript.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI
1	S	S	S	S	S	S	S	S	S	S	S	S	Fea	F	SystematicName	r	r	r	crude probe characterization (1-3)	d	numb	GeneName	D	gls	M3622	gls	gls	M3623	gls	gls	M3624	gls	gls	M3625	gls
20	#	#	#	#	#	#	#	#	#	#	#	#	19	A	ENST00000429990	re	p	e	2) poorly characterized	/	2	NPIPL2	n	1	12846	0	1	14080	0	1	16608	0	1	14403	0
21	#	#	#	#	#	#	#	#	#	#	#	#	20	A	NM_001040196	r	n	e	1) gene symbol	/	1	AGTRAP	H	1	349	0	1	418	0	1	334	0	1	342	0
22	#	#	#	#	#	#	#	#	#	#	#	#	21	A	NM_030961	r	n	e	1) gene symbol	/	1	TRIM56	H	1	446	0	1	456	0	1	450	0	1	360	0

Figure 5: An example for NPIPL2 (nuclear pore complex interacting protein-like 2 [Source:HGNC Symbol;Acc:34409] [ENST00000429990]) as a transcript that is assigned by us as "2) poorly characterized" even though it possesses a gene symbol (NPIPL2). However, this transcript has no RefSeq accession entry of the type: "NM_*".

Kontakt:

Heike Schneider
Technische Assistentin
Gebäude I3, Ebene 1, Raum 1080
Tel: +49 (0)511-532 2830
E-Mail: Schneider.Heike@mh-hannover.de

Torsten Glomb
Diplom-Bioinformatiker
Gebäude I3, Ebene 1, Raum 1300
Tel: +49 (0)511-532 2869
E-Mail: Glomb.Torsten@mh-hannover.de

Dr. rer. nat Oliver Dittrich-Breiholz
Leiter der Zentralen Forschungseinrichtung Transcriptomics
Gebäude I3, Ebene 1, Raum 1270
Tel: +49 (0)511-532 5814
E-Mail: Dittrich.Oliver@mh-hannover.de