Additional File 1

A novel approach to sequence validating protein expression clones with automated decision making

Elena Taycher¹, Andreas Rolfs¹, Yanhui Hu¹, Dongmei Zuo^{1,2}, Stephanie E. Mohr², Janice Williamson¹, Joshua LaBaer^{1,2§}

¹ Harvard Institute of Proteomics, Harvard Medical School, 320 Charles St., Cambridge, MA 02141

² DF/HCC DNA Resource Core, Harvard Medical School, 320 Charles St., Cambridge, MA 02141

Contents

Additional File 1, Figure 1. Isolate Ranking Report	2
Additional File 1, Figure 2. Request for Approval of Specific Primers	3
Additional File 1, Figure 3. Decision Tool Execution	4
Additional File 1, Figure 4. Create New Set of Parameters for Clone Ranking	5
Additional File 1, Figure 5. Online example of Gap Mapper Result	6
Additional File 1, Figure 6. Online example of Low Confidence Region Finder Result	s7
Additional File 1, Figure 7. Parameter Settings for Sequencing Primer Design	8

3 Back • 🕑 • 💌 🧯	Search	To Fa	avorites	9				>							1
dress 🕘 http://128.103.32.16	7:8080/ACE/Seq_GetItem	.do?forw Plate 1	ardName= Results	108.title=	Plate+Resu	ts&CONTA	INER_BAR	CODE=HS	50017108s	how_action	n=IR		*	Go 🔁	Links
Analysis settings Trace files Process	nalysis settings Plate Results rocess														
Keports View	Plate L Plate D Plate T Cloning	abel: D: ype: g Strat	HS 13 SE egy: <u>7</u>	3G001 92 QUEN	710 ICING_C	'ONTAI	NER								
						Q	uality R	eport							
		1	2	3	4	5	6	7	8	9	10	11	12		
	A	1	2	<u>17</u>	<u>25</u>	<u>33</u>	<u>41</u>	<u>49</u>	<u>57</u>	<u>65</u>	<u>73</u>	<u>81</u>	89		
	в	2	<u>10</u>	<u>18</u>	<u>26</u>	<u>34</u>	<u>42</u>	<u>50</u>	<u>58</u>	<u>66</u>	<u>74</u>	<u>82</u>	90		
		3	<u>11</u>	<u>19</u>	<u>27</u>	<u>35</u>	<u>43</u>	<u>51</u>	<u>59</u>	<u>67</u>	<u>75</u>	<u>83</u>	91		
	D F	4	<u>12</u>	<u>20</u>	<u>28</u>	<u>36</u>	<u>44</u>	<u>52</u>	<u>60</u>	<u>68</u>	<u>76</u>	<u>84</u>	92		
	F	2	13	21	<u>29</u> 20	<u>31</u> 20	<u>45</u>	<u>23</u>	<u>61</u>	<u>69</u> 70	<u></u>	82 02	93		
	G	2	14	22	21	<u>20</u> 20	40	<u></u>	<u>04</u> 62	70	70	00 07	94		
	н	8	16	<u>25</u> 24	<u>- 21</u> 32	<u>29</u> 40	48	<u></u> 56	<u>65</u>	<u>71</u> 72	<u>19</u> 80	88	95 96		
	7.53	-													
						C	olor Scl	iema							
			в	est Iso	olate in th	e group	:	10.000							
	Second best isolate in the group:														
			T	hird be	est isolate	e in the ;	group:			Contin	iue				
	Isolate that can not be used :														
	Isolate with no data : Wrong ORF:														
			С	ontrol											
			E	mpty s	ample:										

Additional File 1, Figure 1. Isolate Ranking Report.

Screenshot of ACE showing plate results of Isolate Ranker with color-coded clone rank. Identical border style (color and line style) is used to show wells that represent multiple isolates (clones) of the same gene. The color code is indicated below the plate map. As shown here, all isolates for the gene can be placed together, but it is not required. Clicking on any well number will open up a window with detailed information about the clone including the end reads, clone sequence and their alignments with the expected reference sequence.

🗿 <bean:message key="bec.name"></bean:message> : title - Microsoft Internet Explorer 📃 🗗															
Eil	e <u>E</u> dit	<u>V</u> iew F <u>a</u> vorites <u>T</u> o	ols <u>H</u> elp												1
😋 Back 🝷 💿 🕤 🙁 😰 🏠 🔎 Search 🌟 Favorites 🧭 🚱 🖉 🖉 👘 👘 🛄															
Address 🕘 http://128.103.32.167:8080/ACE/RunProcess.do 💌 🎅 Go													Links		
Google - 🕑 🏀 Search Web 👻 🚳 🔁 82 blocked 📲 AutoFill 🛛 🛃 Options 🥒															
settings Request for Approve Internal Primers														^	
	Analy	sis settings	1	reduceror there and transit											
	Trace	files													
	Proce	SS													
	•Rea	ad manipulation													
	•Eva •Inte	iluate clones ernal primer design	с	lone	e Id:	11	19382								
	and	order			Change	Status									
	sequ	oad clones lences		Priz	mer3 Sp	ecification: 87									
	•Set	final clone status		Keb	erence	Sequence Id: <u>42</u>	:140	T	Destature.	O-to-to-to-	Charles	Coloritation Trees	C.1		
	•Vie	w process results			Ivame	Sequence		Im	Position	Urientation	Status	Submission Type	Submitter		
	Report	rts			Fl	GCTCAAGACCATGACC	GAT 5	8.603	264	Forward	Designed	Gene specific, calculated	System		
	VIEW				F2	ATGCGATGCTCTCTCA	TCA S	57.939	622	Forward	Designed	Gene specific, calculated	System		
			-	Change Status Primer3 Specification: 88 Reference Sequence Id: 42746											
					Name	Sequence		Tm	Position	Orientation	Status	Submission Type	Submitter		
					Fl	ATCCTCTTCCTCCTGA	ccc :	58.028	313	Forward	Designed	Gene specific, calculated	System		
					F2	GGCCTTTATCACCATC	CAG 5	57.947	654	Forward	Designed	Gene specific, calculated	System		
Change Status Primer3 Specification: <u>26</u> Reference Sequence Id: <u>42746</u>															
					Name	Sequence		Tm	Position	Orientation	ı Status	Submission Type	Submitter		
				~	Fl	CTCCATCATTTGTTTC	GTGG	58.973	195	Forward	Approved	Gene specific, calculated	System		~
۲													🥑 Inter	net	

Additional File 1, Figure 2. Request for Approval of Specific Primers.

Screenshot shows the interface for selecting primers for the vendor order. ACE allows the user to design several sets of gene-specific primers for the clone using different Primer Designer specification. Primers from different sets can be put on the same vendor order. Upon placing the order ACE selects all primers approved for the specified clones and puts them in 96-well plate format. The user has an option to specify range of wells in which primers should be placed (preserving empty wells for controls if desired). ACE provides instructions for re-arraying the clone templates so that they match the vendor order and tracks all primers ordered for a particular clone.

🗿 ACE : title - Microsoft Inf	ternet Explorer	
Eile Edit View Favorites	Tools Help	4
Ġ Back 🝷 🐑 🕤 🔀	💈 🏠 🔎 Search 🤺 Favorites 🚱 🔗 🎍 🖥	2 · 📙 🖓 .
Address 🍓 http://128.103.32.16	57:8080/ACE/SelectProcess.do?forwardName=84	💌 🋃 Go 🛛 Links
Google G-	Los geht's! 🗄 🍘 🖉 🤔 👻 Lesezeichen 🕶 🔊 Kein	e Popups 🛛 😽 Rechtschreibprüfung 👻 🍙 Senden an 🕶 🏼 🖉 Einstellungen 🕶
VIEW	Select item type:	
	© Plate Labels	O Project Name Human
	Clone ID	
	Enter all items:	
	HS	G001710
	Process specification	
	Clone accentance criteria:	
		Orint results for all groups in one file
	Output format:	O Print results for each group into separate file
	User Comments:	
	Select fields to be reported	
	Clone storage information:	Clone raw data information
	Clone ID	✓ Is forward end read uploaded?
	Clone final status	Is reverse end read uploaded?
	V Plate label	Number of ordered internal primers
	Sample type	Number of submitted internal reads
	Well	
	Reference sequence information	n Clone assembly information
	Reference sequence ID	Clone sequence ID
	Reference sequence CDS start	Clone sequence assembly attempt status
	Reference CDS stop	Clone sequence analysis status
	Reference CDS length	Clone sequence CDS start
	Reference cequence text	Clone sequence CDS start
	Reference CDS	
	Gene symoor	Using sequence CDO High quality discremencies (sequenced by type)
	St linker as man	I am mulity discrepancies (separated by type)
	 D'inker sequence 2'inker se m 	 Low quality discrepancies (separated by type) Detailed discourse an annual is
	S inker sequence	Detailed discrepancy report
	Species specific ID	✓ Number of gaps in most recent assembly
	All available identifiers	Vumber of low quality discrepancies
🗐 Done		🥥 Internet

Additional File 1, Figure 3. Decision Tool Execution.

Screenshot of user interface that initiates execution of Decision Tool which sorts clones submitted for the analysis into appropriate categories (details in the manuscript). User can define a set of clones for analysis by providing: (1) plate labels; (2) unique clone ID(s); (3) specifying a project. Users invoke predefined acceptance criteria for the analysis by indicating the name of the saved set by selecting it from a pull down list of available sets (see text and Supplemental Figure 3). The output can be directed to separate files (useful for managing groups of clones for downstream steps) or to a single file (to obtain an overall project summary). Users can also specify which information they would like to include in the Decision Tool Report (e.g., clone sequence, clone CDS, discrepancy descriptions). An example of a decision tool report is included in the supplementary material.

🚰 Untitled Document - Micro	soft Internet Explorer			- 7 🔀
<u>Eile E</u> dit <u>V</u> iew F <u>a</u> vorites <u>T</u> o	ols <u>H</u> elp			1
🌀 Back 🔹 🕥 👘 🛃 🛃] 🏠 🔎 Search 🤺 Favorites 🚱 🎯 🖓 📓 🐨 🦳 🦓			
Address 🙆 http://128.103.32.167:	8080/ACE/page_main.jsp?forwardName=-3004&page_location=Home%20>%20Analysis%205	ettings%20>%20Clo	ne%20Ranking&title=Create% 🔽 🛃 Go	Links »
Cloning project	Home > Analysis Settings > Clone Ranking			~
Analysis settings	Create New Set of Parameter for Clone 1	Ranking		
Trace files Process				
Reports		View Mine View All		
View	C A Margar			
	Set Name			
	Penalties for discrepancy in the insert region			
	Process conservative & non-conservative substitutions togethe missense substitutions, system will ignore values for conservative a	er (System will us and non-conserv:	se the penalty values for ative substitutions)	
	⊙Process conservative & non-conservative substitutions separa	tely (System will	use the penalty values for	
	conservative & non-conservative substitutions; system will ignore	values for misser	nse substitutions)	
	Base Confidence Silent substitution	High	Low	
	Missense substitution	20	10	
	Conservative substitution	20	10	
	Non-conservative substitution	50	30	
	Frameshift	1000	1000	
	In-frame deletion	1000	1000	
	In-frame insertion	1000	1000	
	Truncation	1000	1000	
	No translation (e.g., no ATG)	1000	1000	
	Post-elongation (e.g., no stop codon)	1000	1000	
	Penalties for discremancies in the linker region			
	Base Confidence	High	Low	
	5' delation (magnition	1000	F00	
	3' substitution	50	30	
	3' deletion/insertion	1000	500	
	5 GROUP AND AND AND A	1000		
	Penalties for discrepancies introduced by ambiguous bases			
	Show			
	Des Carelana	777-3	T	
	Start coden substitution	nign	Low	
	Stop coden substitution	0	0	
	Substituttion cds region	0	0	
	Frameshift insertion	0	0	
	Inframe insertion	0	0	
	Substitution in 5' linker region	0	0	
	Insertion in 5' linker region	0	0	-
	Substitution in 3' linker region	0	0	
	Insertion in 3' linker region	0	0	~
8			🥥 Internet	

Additional File 1, Figure 4. Create New Set of Parameters for Clone Ranking.

Screenshot of interface to create a set of parameters for isolate ranking. This is particularly relevant for workflows in which multiple isolates for each gene are selected with the intent to pick one "best candidate". For each combination <discrepancy type, confidence> (where "confidence" can be "low" or "high"), the user specifies the penalty per discrepancy combination.

@] .	<bean:message key="bec.r</th><th>ame"></bean:message> : title	e - Microsofi	t Interne	t Explo	rer								ъX	
Eile	<u>E</u> dit <u>V</u> iew F <u>a</u> vorites <u>T</u> o	ols <u>H</u> elp												
G	Back 🔹 🕑 🗉 🗾 💋	🏠 🔎 s	iearch tr	avorites	0	2· 🎍	W -	_ &	\$					
A <u>d</u> d	ress 🙋 http://128.103.32.167;	3080/ACE/RunPro	ocess.do										💌 🔁 Go	Links
Go	ogle -	💌 👸 Sea	arch Web 🔻	🌮 🖓	82 block	ked 📲 AutoFi	i 🛛 🔁 🤆	Options	D					
	Cloning project settings Analysis settings	Home > Pr Vie	rocess > Vie w Late	ew Resul st Co	ts > Vı ntig	ew Latest Co Collect	ontig Co ion	lection						^
	Trace files													
	Process													
	View													_
		CI	one ID:			120016								
]	Reference	Sequen	ce ID:	42786								
			Contig Name	Contig Type	Con II	ttig CDS O Start	CDS Stop	LQR define	d A	lignment	Discrepa	ncy Report		
			Gap 1	Gap	1546	7 -1000	844							
			Contig 1	Contig	<u>1546</u>	<u>5</u> 845	1154	Yes		Alignment	Discrepa	incy Report		
			Gap 2	Gap	1546	8 1155	2462							
			Contig 2	Contig	<u>1546</u>	<u>6</u> 2463	3295	Yes		Alignment	No discrepan	cies		
			Gap 3	Gap	1546	9 3296	-1000							
		CI	one ID:			134915								
]	Reference	Sequen	ce ID:	42988								
			Contig Name	; Co Ty	ntig /pe	Contig ID	CD Stai	S (rt S	CDS Stop	LQR defined	Alignment	Discrepancy Report		
			Contig 1	Cor	ntig 3	6250		-49	35	Yes	Alignment	No discrepancies		
			Gap 1	Gat	3	6252		36	526					~
e	one											🥥 Ir	nternet	

Additional File 1, Figure 5. Online example of Gap Mapper Result.

Screenshot shows contigs and gaps defined for a clone by Gap Mapper. Contigs are characterized by the position on the clone reference CDS and their sequences are analyzed by Discrepancy Finder to allow assessment of clone quality. LQR indicates Low Quality Region.



Additional File 1, Figure 6. Online example of Low Confidence Region Finder Results.

Screenshot shows example of low confidence regions for clone sequence. Low confidence regions are defined for a sequence by applying a 'sliding window' algorithm (see manuscript for details), each region is characterized by position on clone reference CDS. Not viewable because it is at the bottom of the screen is the color code indicating the confidence score.

<u>File E</u> dit <u>V</u> iew F <u>a</u> vorites <u>T</u> ools	Help											
🚱 Back 🝷 🕥 🕤 😫 🛃	🏠 🔎 Search 👷 Favorites	🔗 🍰 🖬 -	_ &									
Address 🕘 http://128.103.32.167:808	0/ACE/page_main.jsp?forwardName=	-3005&page_location=Home%20)>%20Analysis%2	05ettings%20>%20Pr	imer%20Designer&title=Create® 🔽 🛃 Go	Links »						
Cloning project settings	Home > Analysis Settings > Primer Designer Create New Set of Parameter for Primer Design											
Trace files												
Process Reports View	<u>View Mine</u> <u>View All</u> <u>Delete</u>											
	Set Name											
	Primer Design Parameters											
	Primer Length (bp)	Min: 18	Optima	al: 21	Max: 27							
	Primer Tm (C)	Min: 57	Optima	al: 60	Max: 63							
	Primer GC%	Min: 30	Optim	al: 50	Max: 70							
	Sequencing Parameters Check here if J Universal Primer is checked Check here if J Distance between 5p UJ (For a left primer, primer start) a right primer, primer start) Estimated high quality o	NO 5p Universal Primer) NO 3p Universal Primer niversal Primer and STA rtposition is the position of t niversal Primer and STC position is the position of the read length (ERL)	100 100 400	rd PCR primer is used when No 5p bases bases bases								
	Window size for testing	primers		50	bases							
	Distance between sequered length	encing primer and start	of high quality	50	bases							
	Number of strands to so	odneuce 2	 Single Strand (Coding strand, forward primers only) Single Strand (Compliment to coding strand, reverse primers only) Both Strands (Both forward and reverse primers until meet in middle) Both Strands (Both forward and reverse primers, double coverage) 									
Done					Internet	×						
_												

Additional File 1, Figure 7. Parameter Settings for Sequencing Primer Design.

Screenshot shows interface that allows users to create a set of parameters for primer design which include: (1) sequence related parameters (e.g., Tm, GC content); (2) sequence processing parameters (e.g., length of reliable part of sequencing reads); (3) type of desired coverage (e.g., single forward, single reverse, double coverage).