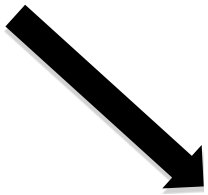
Report from the User Advisory Group (UAG)

Biocreative III Workshop
September 13th-15th, Bethesda

Andrew Chatr-aryamontri
Wellcome Trust Centre for Cell Biology, University of Edinburgh

How to transform killing ideas into killing tools



The user advisory group: a step towards the future

A critical aspect of the BioCreative III and IV evaluations is and will be the active involvement of the end users to guide development and evaluation of useful tools and standards.

Each end user or curator will be asked to normalize a set of articles. The task will produce a list of the genes/proteins in the article, including both the gene/protein name and the associated unique identifier (EntrezGene for genes, UniProt identifier for proteins).

Users will also provide feedback on what features they found useful and what additional features they would like to have. Features might include species for the gene/protein, links to mentions of genes in the article, or attributes of the gene such as chromosome location.

Some chronology

- The user advisory group started regular monthly meetings on December 2009
- Over the 9 months guidelines of the task were defined
- At the end August the tools were delivered and tested
- On the 8th of September the results were delivered

Defining the standards

The entry questionnaire

Please answer the following questions based on curation process in your institution. Your response will help us define annotation guidelines for manually annotating gene identifiers in full-length articles.

1. How would you identify and distinguish the primary genes that are main topic of an article vs. secondary genes simply mentioned “in passing”? In other words, what are the criteria for selecting and annotating genes from full-length articles?

2. Does any of the BMC (http://www.biomedcentral.com/browse/journals/) or PLoS journals (http://www.plos.org/journals/) often publish papers selected by your group for gene indexing? If so, please list those specific journals.

3. Would you annotate genes mentioned in the paper supplementary material?

4. Would you annotate genes mentioned in the Method/Material Section of a paper?

A lot of questions!

- What does interactive task mean?
- What are the groups developing the systems?
- What if the information in the paper is not present in the database?
- Literature corpus: How is this corpus to be selected?
- What papers?
- What journals?

Summarizing…….

1. What is relevant for curation?

2. How to distinguish primary/secondary genes

1) What is relevant for curation?

- Each database, or resource have their own focus:

  -Curation of genes for a given organism (MOD)

  -Curation of protein-protein interaction

  -Curation of phenotypes

  -GO annotation

  -Curation of pathways

What is a common interest among these?

-Identifying relevant articles for a given gene/protein
-Finding experimental data for a given gene/protein
-Identifying species

2) How to distinguish primary/secondary genes

First concept based on frequency of gene mention should be weighted considering specific sections in the article that they appear

Title/Abstract/Results

Introduction/Discussion

Exercise: 2 articles, normalization

PMID: 19513100		
Gene (species)	Entrez ID	Popularity
gata1 (human)	2623	9
gata1 (mouse)	14460	9
e2f2 (mouse)	242705	9
fog-1 (mouse)	22761	9
fog-1 (human)	161882	9
pRB (mouse)	19645	9
pRB (human)	5925	5
CD71 (mouse)	22042	4
c-kit (mouse)	16590	4
ter119 (mouse)	104231	4
pcna (mouse)	18538	3
p107 (mouse)	18148	3
beta-actin (mouse)	11461	3
eGFP (B. cereus)	8382257	1

PMID: 19014439		
Gene (species)	Entrez ID	Popularity
Prp40 (yeast)	853857	9
Snu71 (yeast)	852896	9
Luc7 (yeast)	851471	9
ypr152c (yeast)	856275	5
DBP2	855611	2
ECM33	852370	2
Clf1	850808	1
CA150	10915	1

What genes are primary/central?

1- Any gene that appeared in an experiment

2- Genes associated to experimental evidence and with biological significance in the context of the article

Gene/Protein Annotation Guidelines (v1.2)
Jan 25, 2010

What to annotate and normalize:

Find gene/protein mentions in the full-length article and map them to standard database identifiers (Entrez Gene IDs)

Entrez Gene Ids are required. (UniProt Ids or Model Organism Database Ids are optional).

Annotate all genes mentioned in the article including those in passing genes that may be only mentioned once in the article. However, no need to rank or group genes for this assignment.

When there is no explicit mention of organism about a gene in surrounding text, try to use the article context to help determine its species. Some helpful cues include details in the method/material section such as cell lines, organism-specific gene nomenclature conventions, etc.

Also use your domain knowledge for determining which organism a gene belongs to when no explicit species information is given in the text. If there is absolutely no clue about the species, or in situations where the authors use one gene as a representative of its homologs, do not annotate the gene.

When cell lines of different known organisms are used to study one gene/protein of a single species, determine and use the gene's organism for annotation regardless of multiple species in the cell lines.

What NOT to annotate:

Do not use references

Do not use and annotate supplementary material

Do not annotate gene/proteins mentioned **only** in the Method/Material section. Use this section for help annotation is allowed (e.g. help identify species information).

Do not annotate protein complex (e.g. TFTC complex) unless its members are explicitly given (NFKB-IKB complex)

Do not annotate protein family (e.g. cytokines; ring-h2 finger proteins) because no unique Entrez Gene ids can be assigned to them.

Do not annotate gene/protein with general species information (e.g. mammalian p53) for the same reason above.

Setting system specifications

- Data set: full text articles in XML from the PubMed Central Open Access collection
- Web-based interface
- Indexing Task: Given a PMCID, provide Gene list (normalized) and ranking (based on centrality)
- Retrieval: Retrieve documents for which a given gene is central.

Testing the systems

- All systems should be tested with same articles for direct comparison
- More than 1 tester per system (2-3)
- 2 articles per tester
- Articles tested are new to UAG so we can compare time of manual vs. system-assisted curation

Testing the systems – feedback from curators

- Comment on interface

  Is it easy to use?

  What features do you like and dislike?

- Comment on system performance:

  Did system assist on curation?

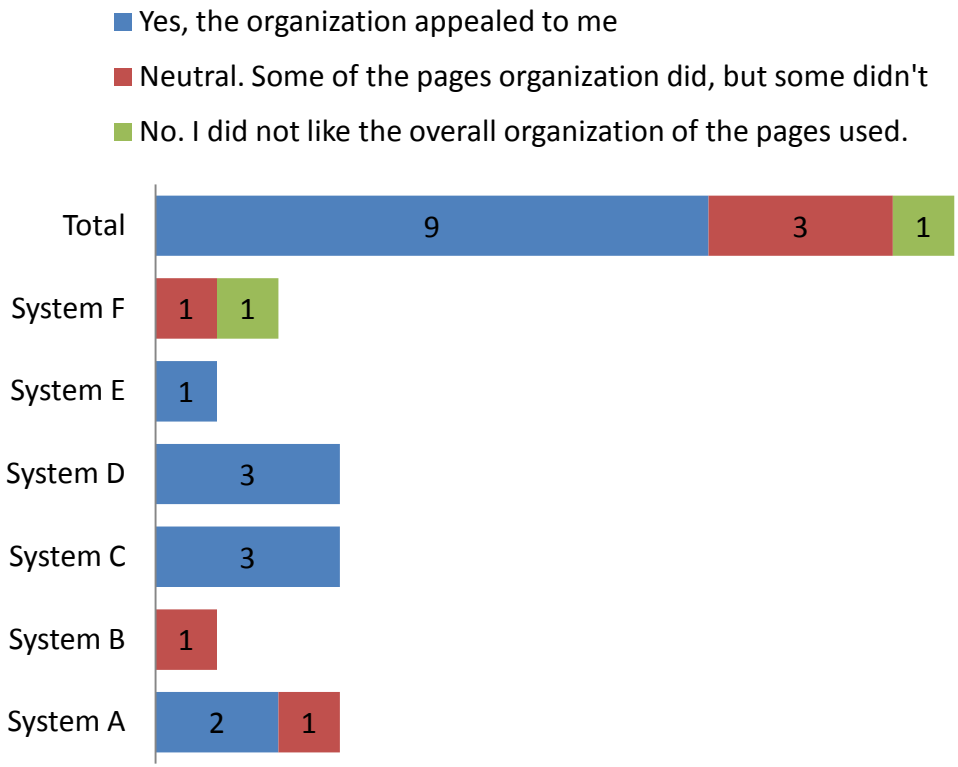
  Did system find primary genes and ranked them high?

  Is this faster than manual curation?

  Did system retrieve articles relevant for the selected gene?

About the Interfaces

1. As you operated the system interface, did the overall organization of the web pages appeal to you?



In general the UAG members like the organization of the interfaces

3. What aspects/features about the interface appealed to you the most?

System A	System B	System C	System D	System E	System F
Sorting by column	Allow user to validate or suggest gene names, and species	Intuitive	Intuitive and clear	Intuitive and clear	Simple, easy to use
Easy access to DBs	Easy access to DBs	Easy access to DBs	Multiple panel format with information linked	Nice table layout	Easy access to DBs
keyword in context colouring	Different color highlighting based on term	Nice layout of summary table with genes and sp info	Gene/protein distinction	keyword in context colouring	Link to highlighted text

- Intuitive
- Easy access to Databases (Entrez and UniProt)
- Different color highlighting based on term/context
- Summary table layout

4. What aspects/features would you like to see added to this interface?

System A	System B	System C	System D	System E	System F
Allow user input to validate or suggest gene names	Work with full text	Provide an option of highlighting a particular gene/protein (organism specific);	See figures and tables published with the paper.	Allows user input to correct species or gene names	I would like to see how the 'centrality' is calculated, and a value.
Gene-species pair, colouring in context;	Attempt to link gene with species and show source of assignment	Show/highlight terms which result in species assignment to gene mentions	It would be useful to deselect all the entry associated with a particular origin species.	That GeneID be linked to DB from table	Search options in the search results
Higher recall	Rank genes based on centrality	Gene mention counts but no attempt to assign primary genes	Show name next to the Entrez ID	Able to export table	Customizable table download
			Save selected gene/proteins while editing them on the panel		

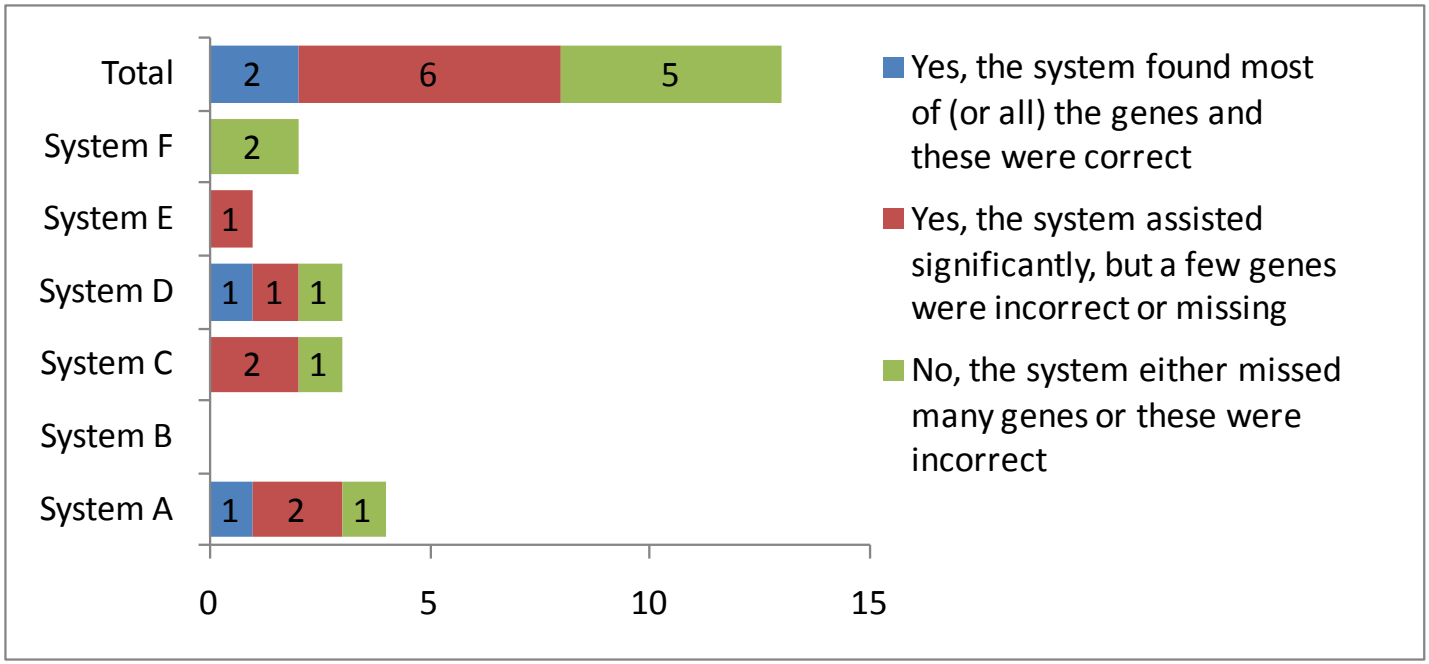
- **User Validation: able to add/delete species, and gene names followed by on the fly gene normalization and ranking**
- Gene-species pair coloring in context

5. List any aspects/features that did not appeal to you

System A	System B	System C	System D	System E	System F
Keyword in context colouring was not useful	Only links to Uniprot	Display web-browser dependent	System evaluated with a higher score human genes	Less biased toward human genes	Ordering did not make sense to me
Popularity was of no use	Does not allow an easy way to deal with PMCIDs	Reference not numbered	System failed in the identification of Drosophila genes although the term "Drosophila" was present in the text. Also has identified Rna codon (eg. GAA) as gene but failed in identifying HIV-1 gag protein.	High error rate	High error rate
Lack of sorting by position in text flow	Not very intuitive	Identified gene not highlighted throughout text	Would like higher recall	system not stable	Cannot sort alphabetically
Full text display is messy (suggest separating Introduction, Results etc, make it easier to read), also provide a link to Pubmed record	It would be useful to get the output as a table and not only as tagged text	Search is case sensitive	Ranking not very accurate	Ranking not very accurate	Lack of help documentation/instructions

Biased of some systems towards a given organism set

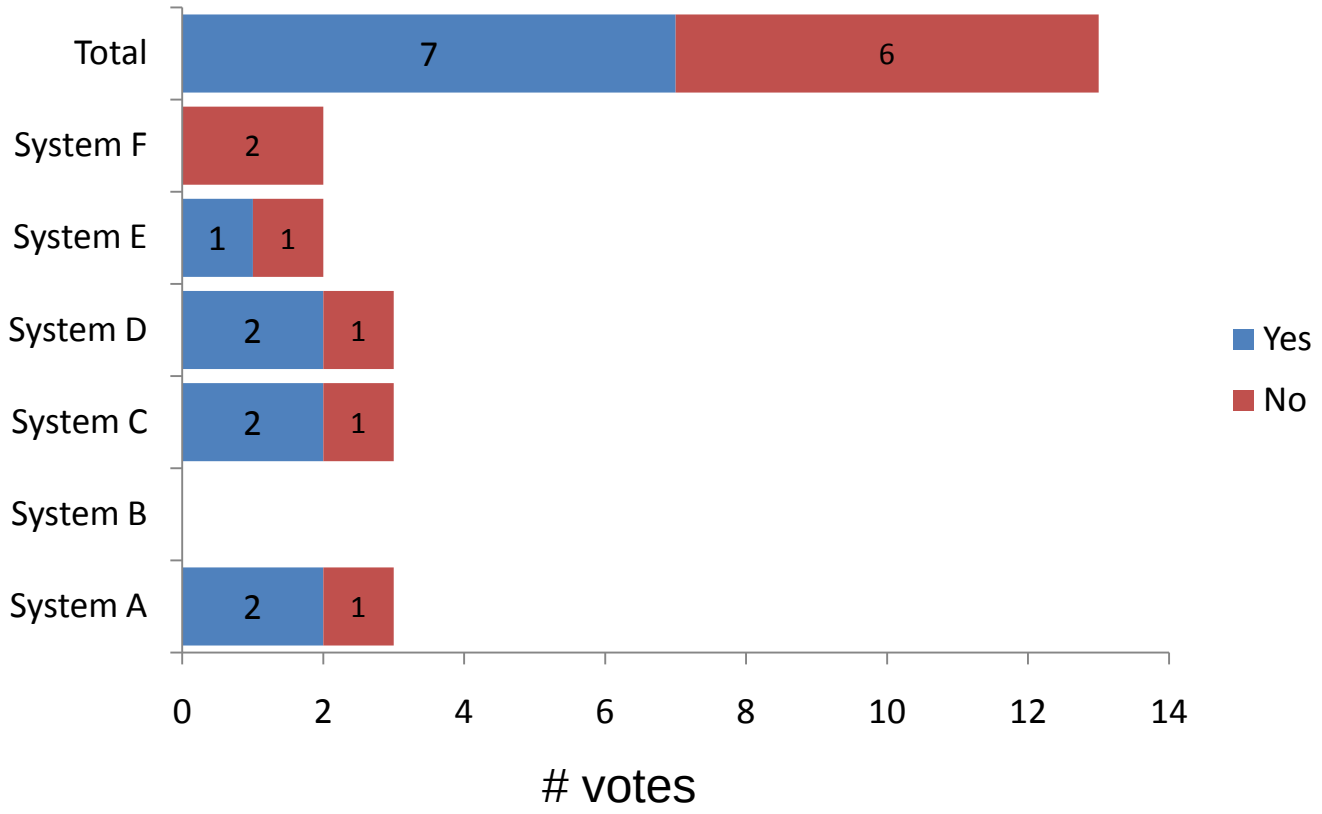
6.Did the system help you with the gene normalization task?



Reduced time spent searching for Entrez ID

The answers could differ based on specific article curated

7. Is the gene ranking correct? (are the top ranked genes primary?)



In some cases genes were ranked OK but the species were not assigned correctly

Results for PMCID:2275796

Easy case: All curators agree in the two genes and in the assignment of primary

		Manual		System A	System C	System D
Gene names	**Species**	**Curator 1 (Sr)**	**Curator 9 (Jr)**	**Curator 7 (Sr)**	**Curator 2 (Jr)**	**Curator 10 (Jr)**
glut9/SLC2A9	human	Y	Y	Y	Y	Y
WDR1/API1	human	Y	Y	Y	Y	Y
	Time (min)	15	27	7	20	48

Primary

Systems?: All systems found the primary, however for some there was ambiguity SLC2A9 and SLC2A6 were both ranked high (share GLUT9 in name, only SLC2A9 is correct.

System Raw Output					
System A	**System B**	**System C**	**System D**	**System E**	**System F**
Y	Y	Y	Y	Y	Y
Y	missed	Y	Y	Y	missed

Primary

		Manual		System A	System C	System D	System Raw Output					
		Curator 1	**Curator 9**	**Curator 7**	**Curator 2**	**Curator 10**	**System A**	**System B**	**System C**	**System D**	**System E**	**System F**
	Total genes	2	2	2	2	2	6	3	4	44	4	15
	FP	0	0	0	0	0	4	2	2	42	2	14
	FN	0	0	0	0	0	0	1	0	0	0	1
	TP	2	2	2	2	2	2	1	2	2	2	1
Precision = TP/TP+FP		1	1	1	1	1	0.33	0.33	0.50	0.05	0.50	0.07
Recall = TP/TP+FN		1	1	1	1	1	1	0.25	1	1	1	0.5

High number of false positives are related to normalizing things that are not genes: CAD, BIM and MI these are acronyms shared with clinical terminology.
<u>Suggestion:</u> Important to consider contextual information.

Case of a new gene name PMCID: 2764847

Not so easy.... A new gene is described AtHSB, the name is introduced for the first time along with its identifier: At5g06410

> “As the name Jac1 in Arabidopsis has been assigned to another protein we named At5g06410 AtHscB”

Curator result (Curator 4, 5 and 8 did not curated full set of genes, but primary)

			Manual		System A		System C		System D		System F	
Gene ID	Gene names	Species	Curator 1	Curator 2	Curator 11	Curator 9	Curator 6	Curator 3	Curator 8*	Curator 12	Curator 4 *	Curator 5*
830529	AtHscB, At5g06410	arabidopsis	Y	Y	missed	Y	Y	Y	missed	Y	Y	Y
852866	Jac1	yeast	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
828316	AtlscU1	arabidopsis	Y	Y	Y	Y	Y	Y	missed	Y	Y	Y
829947	AtHscA1, At4g37910	arabidopsis	Y	Y	missed	Y	Y	Y	missed	Y	Y	Y
851084	Ssq1	yeast	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
837590	SufA	arabidopsis	Y	Y	missed	Y	Y	Y		missed		
825814	SufB	arabidopsis	Y	Y	missed	Y	Y	Y		missed		
820236	SufC	arabidopsis	Y	Y	missed	Y	Y	Y		missed		
840144	SufD	arabidopsis	Y	Y	missed	Y	Y	Y		missed		
837370	SufS	arabidopsis	Y	Y	Y	Y	Y	Y		missed		
828756	SufE	arabidopsis	Y	Y	Y	Y	Y	Y		Y		
822033	HCF101	arabidopsis	Y	Y	Y	Y	y	Y		Y		Y
842789	APO1	arabidopsis	Y	Y	Y	Y	Y	Y		Y		
835939	Sta 1	arabidopsis	Y	Y	Y	Y	missed	Y		Y		
835169	AtNBP35	arabidopsis	Y	Y	missed	missed	missed	Y		Y		Y
852789	NBP35	yeast	Y	missed	Y	missed	y	Y		Y		Y
854814	cfd1	yeast	Y	missed	Y	missed	missed	Y		Y		
855968	ISU	yeast	Y	Y	Y	missed	missed	Y		Y		
946995	HscB	Ecoli	Y	Y	missed	missed	missed	Y		Y		
43847	Jac1	arabidopsis	Y	missed	Y	missed	y	missed		Y		
821316	AtlscU2	arabidopsis	Y	Y	missed	Y	Y	Y		Y	Y	
825719	AtlscU3	arabidopsis	Y	y	missed	Y	Y	Y		missed	Y	
830818	HscA2, AT5g09590	arabidopsis	Y	Y	missed	Y	Y	Y		Y	Y	
818077	atNAP1	arabidopsis	Y	missed	missed	missed	missed	Y		Y		
843182	Cut1	arabidopsis	Y	missed	missed	missed	y	missed		Y		
856692	URA3	yeast	Y	missed	Y	missed	Y	missed	Y	missed		Y
853503	SSC1	yeast	Y	missed	missed	missed	missed	Y	Y	missed		Y
947002	IscU	E. Coli	Y	missed	missed	missed	missed	Y		missed		
944885	HSCA	E. Coli	missed	missed	Y	missed	Y	missed		missed		

Number of genes annotated varied among curators

Now the systems...

			Raw Output					
Gene ID	**Gene names**	**Species**	**System A**	**System B**	**System C**	**System D**	**System E**	**System F**
830529	AtHscB, At5g06	arabidopsis	missed	N/A	Y (freq low)	missed	Y	missed
852866	Jac1	yeast	Y		Y	Y	missed	Y
828316	AtIscU1	arabidopsis	missed		missed	missed	missed	missed
829947	AtHscA1, At4g3	arabidopsis	missed		missed	missed	missed	missed
851084	Ssq1	yeast	Y		Y	Y	missed	Y
837590	SufA	arabidopsis	missed		N (ECOLI)	missed	missed	missed
825814	SufB	arabidopsis	missed		N (ECOLI)	missed	missed	missed
820236	SufC	arabidopsis	missed		N (ECOLI)	missed	missed	missed
840144	SufD	arabidopsis	missed		missed	missed	missed	missed
837370	SufS	arabidopsis	Y		missed	missed	missed	missed
828756	SufE	arabidopsis	missed		missed	Y	missed	missed
822033	HCF101	arabidopsis	Y		Y	Y	missed	Y
842789	APO1	arabidopsis	Y		missed	Y	missed	missed
835939	Sta 1	arabidopsis	missed		missed	Y	missed	missed
835169	AtNBP35	arabidopsis	missed		missed	missed	missed	Y
852789	NBP35	yeast	missed		Y	Y	missed	Y
854814	cfd1	yeast	missed		missed	Y	missed	missed
855968	ISU	yeast	missed		missed	Y	missed	missed
946995	HscB	Ecoli	missed		missed	missed	missed	missed
43847	Jac1	arabidopsis	Y		missed	missed	Y	missed
821316	AtIscU2	arabidopsis	missed		missed	missed	missed	missed
825719	AtIscU3	arabidopsis	missed		missed	missed	missed	missed
830818	HscA2, AT5g09	arabidopsis	missed		missed	missed	missed	missed
818077	atNAP1	arabidopsis	missed		missed	missed	missed	missed
843182	Cut1	arabidopsis	missed		Y	Y	missed	missed
856692	URA3	yeast	Y		Y	missed	missed	Y
853503	SSC1	yeast	Y		missed	missed	missed	Y
947002	IscU	E. Coli	missed		missed	missed	missed	missed
944885	HSCA	E. Coli	missed		missed	missed	missed	missed

*Curators 8, 4 and 5 only listed the primary genes that is why are not counted on the excercise below.

	Curator						
Total gene mentioned in	1 (Sr)	2 (Jr)	3 (Sr)	6 (Sr)	9 (Jr)	11 (Jr)	12 (Sr)
29	28	20	28	26	20	17	22
FP	0	0	2	4	0	3	2
FN	1	9	3	7	9	15	7
TP	28	20	26	22	20	14	22
Precision	1.00	1.00	0.93	0.85	1.00	0.82	0.92
Recall	0.97	0.69	0.90	0.76	0.69	0.48	0.76

	System (raw result)					
	A	B	C	D	E	F
Total # genes in article	54	N/A	22	65	9	23
FP	46		14	58	7	16
FN	21		21	19	27	22
TP	8		8	10	2	7
Precision	0.15		0.36	0.15	0.22	0.30
Recall	0.28		0.28	0.34	0.07	0.24

- Genes may be correct, but many cases wrong organism SufA in Ecoli instead of Arabidopsis
- Also link to human/mouse when non mention about these organism
- Other cases system would link [FE-S] to FES human

What about primary genes?

No so good news:

Taking majority votes: 5 primary genes

	Curators	System raw output					
		System A	System B	System C	System D	System E	System F
# Primary	5	2	N/A	2	2	1	2
Incorrect primary		1				1	2
Primary with low frequency (due to synonym not detected)				1			

- All systems missed the Arabidopsis genes (Curators could link these by using the identifiers described in the article (AtHscB->At5g06410; AtHscA1-> At4g37910)
- All systems except 1 detected the yeast genes as primary
- 2 systems ranked high incorrect genes (either species HSB human, or FE-S linked to human FES)
- Only one system detected 1 Arabidopsis gene, it detected the identifier but could with low frequency because it could not make the link to HscB. Adding HscB as synonym should have helped

-Case with many species

PMCID: 2680910

Results after Curation

			Manual		Systems		
Entrez Gene I	Gene names	Species	Curator 1 (Sr)	Curator 9 (Jr)	Curator 7 (Sr)	Curator 2 (Jr)	Curator 10 (Jr)
10015	ALIX	human	Y	Y	Y	Y	Y
57630	POSH	human	Y	Y	Y	Y	Y
36990	POSH	Drosophila	Y	Y	Y	Y	Y
43330	ALIX	Drosophila	Y	Y	Y	Y	Y
128866	CHMP4B	human	Y	Y	Y	Y	missed
155030	Gag	HIV-1	Y	Y	missed	Y	Y
39659	TAK-1	Drosophila	Y	Y	Y	Y	Y
3355106	ALG-2	Drosophila	Y	Y	Y	Y	Y
7323	UbcH5c	human	Y	Y	Y	Y	missed
1489984	p9	EIAV	Y	Y	Y	Y	missed
137492	HCRP1/Vps37A	human	Y	Y	Y	Y	missed
7251	Tsg101	human	Y	Y	Y	Y	missed
155030	p6	HIV-1	Y	missed	Y	Y	missed
7334	Ubc13	human	Y	missed	Y	Y	missed
		Time	60	180	25	75	83
		Total genes	14	19	13	26	10
		FP	0	5	0	0	3
		FN	0	2	1	0	7
		TP	14	12	13	14	7
	Precision = TP/(TP+FP)		1.00	0.71	1.00	1.00	0.70
	Recall = TP/(TP+FN)		1.00	0.71	1.00	1.00	0.70

In this case it looks like the systems sped up curation

The systems raw results...

Entrez Gene ID	Gene names	Species	System A	System C	System D	System E	System F
10015	ALIX	human	Y	Y	Y	Y	Y
57630	POSH	human	Y	y	Y	Y	Y
36990	POSH	Drosophila	Y	Y	missed	Y	Y
43330	ALIX	Drosophila	Y	Y	missed	missed	Y
128866	CHMP4B	human	Y	missed	Y	missed	Y
155030	Gag	HIV-1	Y	missed	missed	Y	missed
39659	TAK-1	Drosophila	missed	Y	missed	missed	Y
3355106	ALG-2	Drosophila	missed	missed	missed	Y	missed
7323	UbcH5c	human	missed	y	Y	missed	missed
1489984	p9	EIAV	missed	missed	missed	missed	missed
137492	HCRP1/Vps37A	human	Y	missed	Y	missed	missed
7251	Tsg101	human	Y	missed	Y	missed	missed
155030	p6	HIV-1	missed	missed	missed	missed	missed
7334	Ubc13	human	Y	Y	Y	missed	missed
		Total genes	90	22	120	9	52
		FP	81	15	113	4	46
		FN	5	7	7	8	8
		TP	9	7	7	5	6
	Precision = TP/(TP+FP)		0.10	0.32	0.06	0.56	0.12
	Recall = TP/(TP+FN)		0.10	0.32	0.06	0.56	0.12

- The systems identified all the human primary genes (red) but only 2 identified the HIV protein.
- They also identified the Eukaryotic genes but the species assignment were incorrect
- Cases of many false positives due to ambiguity in species assignment.

Multiple names for a given gene: MP20, MP18, MP17 are all synonyms and use indistinctively throughout the text

PMCID:48140

			Manual		System A	System C		System D	
Gene ID	**Gene names**	**Species**	**Curator 1**	**Curator 2**	**Curator 11**	**Curator 6**	**Curator 3**	**Curator 8**	**Curator 12**
no ID	MP20/MP17 /MP18	sheep	Y	Y	Y	missed	missed	Y	assigned everything to human
no ID	galectin-3	sheep	Y	Y	Y	missed	Y	Y	
100294602	MIP	sheep	Y	Y	missed	missed	Y	missed	
no ID	connexin 46	sheep	Y	missed	missed	missed	missed	missed	
100170231	connexin 50	sheep	Y	missed	missed	missed	missed	missed	
233187	MP20	mouse	missed	Y	missed	Y	Y	missed	
114903	MP20/MP17 /MP18	rat	Y	Y	Y	Y	Y	missed	
83781	galectin-3	rat	Y	Y	Y	Y	Y	missed	
233187	To3	mouse	Y	missed	missed	missed	Y	missed	
280859	MIP	bovine	missed	missed	missed	missed	missed	missed	
18858	PMP22	mouse	missed	missed	Y	Y	Y		
Time			20	30	45	20	60		20
Total genes	10		8	7	3	8	8	2	7
FP			0	1	0	0	2	0	7
FN			3	5	6	7	3	8	0
TP			8	6	5	4	7	2	0
Precision			1.00	0.86	1.00	1.00	0.78	1.00	0.00
Recall			0.73	0.55	0.45	0.36	0.70	0.20	

The systems:

Gene ID	Gene names	Species		System A	System B	System C	System D	System E	System F
no ID	MP20/MP17 /MP18	sheep		missed	N/A	missed	missed	missed	missed
no ID	galectin-3	sheep		missed		missed	missed	missed	missed
100294602	MIP	sheep		missed		missed	missed	missed	missed
no ID	connexin 46	sheep		missed		missed	missed	missed	missed
100170231	connexin 50	sheep		missed		missed	missed	missed	missed
233187	MP20	mouse		Y		Y	missed	missed	Y
114903	MP20/MP17 /MP18	rat		Y		Y	missed	Y	Y
83781	galectin-3	rat		Y		Y	missed	Y	missed
233187	To3	mouse		Y		missed	missed	missed	missed
280859	MIP	bovine		Y		Y	missed		missed
18858	PMP22	mouse		missed			Y		
		Time		47		13	56	4	21
		Total # Genes		41		7	55	2	19
		FP		36		3	10	7	8
		FN		6		6	1	2	2
		TP							
			Precision	0.13		0.46	0.02	0.50	0.10
			Recall	0.14		0.67	0.09	0.22	0.20

All systems got right the primary gene names but not the species assignment.
One of the systems failed to suggest rat as an organims, only suggested human and mouse
Maybe curator 12 got biased by the system??

In this case it would have help if the user can add an organism, an eliminate others

Easy case, but article about prokaryotic genes.

PMCID:102584

ABSTRACT

The mismatch repair pathway in *Escherichia coli* has been extensively studied *in vitro* as well as *in vivo*. The molecular mechanisms by which nucleotide cofactors regulate the whole process constitute an area of active debate. Here we demonstrate that nucleotide (ADP or ATP) binding to MutS mediates a switch in protein conformation. However, in MutS that is DNA bound, this switch ensues only with ATP and not with ADP and is similar, irrespective of whether it is bound to a homo- or a heteroduplex. The results envisage a minimal model of three conformational states of MutS as reflected in: (i) a specific and highly stable MutS–mismatch complex in the absence of a nucleotide; (ii) a specific but less stable complex in the presence of ATP hydrolysis; and (iii) an irreversibly dissociated complex in the presence of ATP binding (ATPγS). Such transitions are of relevance to the protein's function *in vivo* where it has to first recognize a mismatch, followed by a search for hemimethylated sites.

Mention of E. coli and MutS is spread all over the full text too

Results

Only manual curation vs. raw system results were compared

PMCID:102584

			System					
			A	B	C	D	E	F
Gene ID	Gene names	Gene Species ID	Correct?	n/A	Correct?	Correct?	Correct?	Correct?
4436	MSH2	human	missed		missed	y	Y	Y
2956	MSH6	human	missed		missed	missed	missed	Y
854063	MSH2	yeast	missed		missed	missed	missed	Y
851671	MSH6	yeast	missed		missed	missed	Y	Y
947206	MutS	Escherichia coli	missed		Y	missed	missed	missed
948691	MutL	Escherichia coli	missed		Y	missed	missed	missed
947299	MutH	Escherichia coli	missed		missed	missed	missed	missed
282217	DNAse I	bovine	missed		gene OK	gene OK	missed	gene OK but not specie
Total Genes	8		0		6	98	2	16
		FN			5	6	6	3
		FP			4	97	0	12
		TP			2	1	2	4
		Precision			0.33	0.01	1.00	0.25
		Recall			0.29	0.14	0.25	0.57

Only one system detects Muts as primary gene, other missed it. The system missed all the E.coli genes.

Gene not found maybe because of species considered by system???

Retrieval Task

- **Ambiguity**

Ambiguity was found among the systems,
example WASP ENTREZ:7454
retrieves highlighly ranked articles about N-WASP, or genes of the WASP family

"Human Subtelomeric WASH Genes Encode a New Subclass of the WASP Family"

- **Species assignment**

If query is ENTREZ:3965 (Tak-1 from Drosophila)
it should rank higher articles where Tak-1 is primary for this organism

This was not always the case: e.g.

2206340	17559674	not drosophila, picked up species from reference.
2766254	19893628	OK
2516935	18769721	only mentions Drosophila tak1 once, the rest is human
1373652	16451733	article mentions tak1 but not primary
514368	15302918	Incorrect it is about a different gene TAK1/TR4

The UAG team

Cecilia Arighi	Georgetown University Medical Center (Uniprot)
Gianni Cesareni	Tor Vergata University
Andrew Chatr-aryamontri	Wellcome Trust, University of Edinburgh
Pascale Gaudet	Dictibase, GO consortium
Michelle Giglio	University of Maryland (GO, PAMGO)
Ian Harrow	Pfizer
Eva Huala	TAIR
Pankaj Jaiswal	Oregon state University (Plant Ontology consortium)
Zhiyong Lu	NCBI
Lois Maltais	Mouse Genome Informatics group
Livia Perfetto	Tor Vergata University
Phoebe Roberts	Biogen
Carl Schmidt	Mouse Genome Informatics group
Paul Sternberg	Caltech
Luca Toldo	Merck
Jean-Francois Tomb	Dupont

BIOCREATIVE organizers